Harvard Data Science Review • Issue 2.3, Summer 2020

The Future of Data Science

Sach Mukherjee[1] **Sylvia Richardson**[2]

[1]Statistics and Machine Learning, German Center for Neurodegenerative Diseases, Bonn, Germany,
[2]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, England, United Kingdom

Published on: Sep 30, 2020

DOI: https://doi.org/10.1162/99608f92.f09d7aa3

Dramatic advances in machine learning and statistics and their interfaces with science, industry, and policy have ushered in a ‘data era.’ Data science has emerged as a term to capture the broad range of concepts, methods, and tools involved in this transformation. We warmly commend Xuming He and Xihong Lin and Jeannette Wing on their thought-provoking ideas on data science and the scientific and societal challenges therein. We would like to take the opportunity to touch on some additional points that we hope will complement the stimulating discussion in the two articles.

A general theme in both articles is the need to go beyond the current boundaries of the fields of statistics and machine learning in various ways. The authors draw attention to a number of issues, ranging from ethics and privacy to the need for new methods addressing specific questions (e.g., around the data life-cycle or heterogeneous data). These questions are highly relevant to contemporary applications, but have nevertheless remained incompletely studied.

The points that follow are areas where we think more work is needed. Points 1 to 4 are of a more technical nature, related to the research agenda in data science per se, while points 5 and 6 relate to the way data science is organized and its relationship to other fields.

1. New notions of optimality, combining statistical, computational, decision-theoretic and design criteria.

He and Lin rightly note that classical statistical notions of optimality do not take account of computational considerations. This has motivated recent efforts to develop new notions of optimality that merge statistical and computational considerations (and allow for trade-offs between the two). Furthermore, as noted by Wing, the intersection of data science and computer hardware is of key importance. We see a need to expand this program to include decision-theoretic and design considerations. This would involve two key shifts:

First, toward considering the impact of the estimator or learner on the real-world problem. This would move beyond generic loss functions toward eliciting or learning more relevant notions of loss in real-world contexts, as well as improving the understanding of their impact on inference. Here, what we have in mind goes further than, say, the use of weighted losses in supervised learning toward quantification of the effects of analyses on downstream scientific and policy goals. This requires both new theory and methods as well as detailed domain-specific work.

Second, a shift from thinking about data analysis and experimental design separately toward a unified view. Elements of these ideas already exist in various subareas, including reinforcement learning, active learning, retrospective designed sampling of big data, sequential experimental design, and more, but a satisfactory synthesis remains lacking.

In short, such a program would seek to ask, ‘What is the optimal acquisition/estimation sequence with respect to substantive, real-world goals?’ This requires including aspects such as study design, data collection, measurement technology, and interpretation/assessment of results into the theoretical framework. Over time, this would unify data science with aspects of scientific or policy-related decision-making and strategy, thus broadening our theoretical view to include more of the scientific or policy workflow. We think such a broadening is overdue and would benefit both data science and the fields to which it contributes.

2. Model checking and selection in the large data context.

Here we agree with both articles in their emphasis on issues of heterogeneity arising in large data. One aspect that we would further emphasize is the issue of checking or selecting between models themselves. For supervised learning this can be addressed using more or less standard notions of predictive risk, at least in problems that cleanly fit the classical predictive schema. But for the wide range of analyses involving elements of exploration, discovery, causality, and mechanism that do *not* quite fit this mould, this is not enough. Currently, we have a number of information and statistical criteria that are principled and well understood, but in our view, these share the limitation that they are not sufficiently task- or question-oriented. As a result, the optimal model with respect to such criteria may often be suboptimal with respect to real-world utility. Progress here is closely related to point 1, since we do not think a fully generic model assessment criterion can be stated in an entirely task- or problem-independent manner. This issue is most prominent in areas like unsupervised learning and exploratory analysis, but arises to a greater or lesser extent in many data scientific settings.

Moreover, it is often assumed that the goal is to search for a unique model to fit the data. In our view, the idea of a ‘true’ model, while certainly useful as a mathematical notion in studying properties such as consistency, is at odds with the fact that in contemporary data science applications the true underlying data-generating system is typically so complicated as to be essentially (and at times possibly fundamentally) inaccessible. In fact, the power of data science approaches often lies in their ability to cope with such systems *without ever needing access to the true generative model*. Hence, rather than the notion of a true model, we think it makes sense to emphasize good performance in the broad sense of point 1. Indeed, rather than looking to fit and choose a single model it is often useful to exploit heterogeneity by searching for ‘meta-models’ or ‘ensemble features’ that relate to robustified criteria.

3. Biases and distribution shifts.

Both articles highlight issues of data integration. We fully agree that the question of jointly analyzing multiple data types is an important one. A standard approach to such problems involves positing a latent model of a shared underlying process that in turn gives rise to the various data types. However, for complex high-dimensional data, building models of this kind that are both data efficient and truly capable of yielding new insights remains challenging.

In addition, we would like to draw attention to a different aspect of data integration. Data may be collected in different ways that vary with respect to sampling and other biases or with respect to causal properties (e.g., whether interventional or observational). There remains a need for new approaches to integrate across such distributional/design regimes. Here, there has been relevant recent work (in biostatistics, econometrics and machine learning, among other areas) but many questions remain open. A topical recent example comes from COVID surveillance and associated research, including estimation of the space–time disease burden. Here, the key information comes from testing, which is typically done on a needs basis and is subject to a variety of selection biases. To allow valid inference about the underlying epidemic pattern, there is a need to combine such biased data with designed testing studies (e.g., based on random sampling). Similarly, in causal inference, methods have been developed that allow analysis of combined interventional and observational data.

These kinds of approaches complement integration of the kind mentioned in both articles by emphasizing joint analysis over different sampling/acquisition/causal regimes (but perhaps only one data type). An essential point is that these kinds of distributional/design issues, if not properly dealt with, can lead to misleading, if not entirely incorrect, results from otherwise sophisticated analyses. This is due to the fact that they affect not only parameter learning but also the very risk estimation and model assessment steps that are usually relied upon as empirical checks. This is an area that is also closely connected to machine learning research on distribution shifts and on the robustness of highly flexible models such as deep neural networks under such shifts.

4. Fair and trustworthy data science.

We strongly agree that questions around the fairness and trustworthiness of data scientific procedures are crucial. We would like to add the perspective that we see these two aspects as distinct in important ways. This is due to the fact that an algorithm might be reliable/trustworthy/interpretable (e.g., giving robust results, based on a solid, well-understood optimization with good empirical performance and interpretable model parameters) *but nonetheless be profoundly unethical*.

A key observation in this context is that ethical issues and fairness involve normative questions that can only really be addressed by society and politics. In other words, it is a social and political question as to whether certain algorithmic behavior is desirable or not and this cannot be left to data scientists to decide upon. The implications of this observation for the data science research agenda are twofold. First, there is a need for ways to mathematically encode ethical constraints and impose them on learning schemes. Second, there is a need for new approaches that enable a broad range of stakeholders to query and interface with complex algorithms to in turn allow an informed debate about their ethical properties. The latter is challenging for complex, high-dimensional models where there may not exist a suitable human-interpretable reduction.

In addition, we will need new ways of guaranteeing certain mandatory ethical requirements from the outset and of continuously monitoring algorithmic behavior so as to rapidly flag any concerns (including unexpected ones). New engagement with the social sciences and with fields such as philosophy, law, education, economics, and politics will be needed and specialists in those fields and their interface with data science will be essential to developing the new approaches needed to study and ensure fairness.

Such efforts are already underway, but a step change in activity is now needed to allow the relevant methods and structures to be built up in a timely fashion. The time sensitivity here stems from the risk that a profusion of analyses in socially important areas that do not pass muster in an ethical sense will, in addition to the immediate harm done, likely result in a backlash against algorithms and data science in general that would hinder progress in the many areas where data science could be harnessed for real and truly fair social benefit.

5. A deeper embedding of data science into other fields.

We think it important to draw a distinction between 'shallow' and 'deep' embeddings of data science into other fields. The distinction lies in whether data science is used mainly for analysis without changing the field itself very greatly ('shallow') or whether it changes the conceptual basis or epistemology within the field itself ('deep'). We stress that the term 'shallow' should not be understood as carrying any value judgment; indeed, much outstanding data scientific work is 'shallow' by the foregoing definition.

To make the notion of 'deep' embedding clearer we draw a parallel to the relationship between mathematics and physics on the one hand, and fields such as engineering and chemistry on the other. During the 20th century, elements of physics and mathematics became deeply embedded into many fields. Little by little, this went beyond the application of mathematical and physical methods to otherwise predefined problems to the point where the fields themselves changed in a rather substantive manner. Today, in these fields (and many others) physical and mathematical ideas are prominent at every level, conceptual, didactic, and practical. As a rough indicator of the depth of this integration, we can see that any standard textbook in engineering or chemistry today has from the outset a conceptual framework and intellectual toolkit that has clear physical and mathematical underpinnings.

In contrast, if we look at many fields where data science is prominently used today, the integration remains on the shallow end of the spectrum, and a standard textbook in the specific domain (law, biology, medicine, education…) may involve little or no data science per se. A concrete example is biomedicine, where despite the routine use of statistical and machine learning tools in analyzing data, and the prominence of statistical decision processes and data-driven discovery, data scientific concepts do not yet play a central *conceptual* role, nor are they yet a truly core, essential part of the training.

In our view, the lack of deeper integration limits the rate of progress not only in the fields themselves but also in fundamental data science, because it focuses the latter on only one part of the scientific process—namely, the analysis of acquired data—to the exclusion of others (conceptualization, study design, assessment of scientific value). This in turn biases both theoretical views within data science (see point 1 above) as well as the methodological research agenda.

6. Better organizing/sharing of knowledge.

The rapid growth of data science has been exhilarating, but it has left knowledge gaps between the multiple contributing fields. As a result, it is often the case that important work on a specific topic is well known in one subarea but almost unknown in another. For example, issues around nonrandom missingness and censoring have been studied deeply in statistics and biostatistics but are less well known in other areas of data science, while modern stochastic gradient methods (to take just one example) are a standard tool in machine learning but less widely used in statistics. This issue has structural roots relating to the fact that most data scientists have an identity (usually the primary one) in one of the contributing areas (statistics, machine learning, computer vision…). This diversity is, in our view, a hugely positive feature of data science, but nevertheless, as the field grows, we should arguably pay more attention to establishing structures and fora that improve dissemination of knowledge *between* the contributing fields. The goal of such efforts should be to avoid as much as possible duplication and reinventing of the wheel and furthermore ensure that a breakthrough of broad relevance made in any one area reaches the others as rapidly as possible.

The continuation of this debate is essential. We would like to encourage a broad spectrum of data scientists to participate and enrich it, building on the insights gained from their specific fields of expertise to advance shared conceptual and strategic thinking on the future of data science.

Acknowledgments

We cannot do full justice here to the breadth of literature that has influenced our thoughts on these topics and in this commentary we have therefore eschewed references in the interests of space and fairness. However, we would like to warmly acknowledge many colleagues and collaborators over the years (in statistics and machine learning, as well as in other fields, including biology, medicine, physics, economics, and philosophy) whose ideas have influenced our thinking.

Disclosure Statement

Sach Mukherjee and Sylvia Richardson have no financial or non-financial disclosures to share for this article.