Harvard Data Science Review • Issue 1.2, Fall 2019

Five Immersive 3D Surroundings of Data Science

Xiao-Li Meng^{1,2}

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

²Harvard Data Science Review, Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Nov 01, 2019

DOI: <https://doi.org/10.1162/99608f92.ab81d0a9>

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

“Why do you consider data science an ecosystem, not just a transdisciplinary field?” Several readers of the inaugural issue posed this question in response to my editorial. According to the Oxford English Dictionary, an *ecosystem* is “a complex network or interconnected system,” whereas the adjective *transdisciplinary* refers to “more than one branch of knowledge.” The first connotes a holistic evolving entity, and the second, an assemblage of interrelated but ultimately discrete parts. As co-editor Robert Lue put it, in commenting on this editorial, “Transdisciplinarity is often applied to multiple disciplines having their discrete perspectives used collectively to solve a problem (versus interdisciplinarity where the disciplines fuse). However, the discrete disciplines remain somewhat unchanged. In contrast, in an ecosystem the different species often change and evolve based on their networked interactions in the whole. This implies something very compelling with data science where the disciplines themselves evolve from their interactions in the ecosystem.”

Indeed, as a statistician, I am witnessing a profound evolution of statistics as a contributing discipline to data science, from research paradigms to curriculum designs—albeit the latter is evolving at a pace too slow to meet the demand. In general, the more deeply we appreciate the impetus, impacts, and imprints of data science, the more acutely we become aware of the inwardness, incompleteness, and inaccuracies that entail when data science is framed merely as an academic field of knowledge. Adding to the evidence presented in our first issue, the 13 articles in this second issue further demonstrate the ecosystemic nature of data science.

The immersive collaboration among academia, government, and industry that is depicted [in the interview about the Laboratory of Analytical Sciences at North Carolina State University](#) provides a glimpse into the making and living of data science as a complex ecosystem. The word “immersive” captures the extent to which data science has evolved in multiple dimensions. The articles in this issue remind us that there are at least five such immersive “3D surroundings” in the data science ecosystem. They are listed here in the approximate ordering from organizational and structural surroundings to individual and epistemological ones.

Academia, Government, and Industry

Data science is now recognized as a high priority by these three major sectors of the global society. [The interview on the Data Science Leadership Summit](#), led by Jeannette Wing, the Avaneessians Director of the Data Science Institute at Columbia University, illustrates how academic institutions are working collaboratively to address challenging issues in establishing data science research and education programs on campus. [The overview article](#) by Nancy Potok, the Chief Statistician of the United States, provides an example of how the government sector is increasing its use of data and data science for better governance and public good. Each of the 24 major US government agencies is now required by law to have a Chief Data Officer, a Chief Evaluation Officer, and a Senior Statistical Officer. 72 data science-centered leadership positions have thus been created in just one government.

This growth in demand is amplified by orders of magnitude once we turn our attention towards the burgeoning number of data science positions of in local, national, and international governmental organizations (e.g., the United Nations). These leadership roles can directly impact the lives of many millions, and even billions, of people. Do we have a sufficient number of qualified and willing individuals who are ready to assume these roles? Anyone who knows the answer undoubtedly will share the sense of urgency conveyed in [the article on data science training for government employees](#) by Kreuter, Ghani, and Lane, showcasing another collaboration between academic and government sectors.

The industrial sector faces a similar shortage of qualified data scientists, despite the fact that it tends to fare better in recruiting and retaining talent than the government sector. [The article](#) by Michael R. Berthold, the CEO and co-founder of KNIME, provides a reason: “Data Science is as much about knowing the tool as it is about having experience applying it to real-world problems, about having that ‘gut feeling’ that raises your eyebrows when the results are suspiciously positive (or just weird).” Developing that “gut feeling,” or “data acumen,” as emphasized by [the NAS report highlighted](#) in the first issue, takes far more time and practice than what is available from almost any of our current data science education and training programs.

This demand also illustrates the ecosystemic nature of data science, as it puts much greater pressure on educational institutions at all levels, most of which are struggling with their own shortages of teaching resources. For example, a good number of data science master degree programs in the United States rely heavily or even exclusively on adjunct faculty to complete their offerings. Adjunct faculty can and often do bring much-needed experience and acumen that are essential to the provision of practical training. However, I have been in higher education long enough to understand that the very term ‘adjunct’ suggests that these universities yet need to build efficient and sustainable infrastructures to support such training (if they do believe it is essential).

Humanities, Social Sciences, and STEM

The inaugural issue of *HDSR* demonstrated the extent to which data science permeates these three major disciplinary clusters. The current issue reinforces this point with three more research articles, one in each of the three areas, demonstrating respectively: (1) how data science is creating new directions for well-established disciplines; (2) how well-established disciplinary approaches address emerging issues in data science; and (3) how the increased availability of data helps shed new light on long-standing problems of general societal interest.

Specifically, [Gregory Crane’s article](#) on language hacking provides a blueprint of the means by which machine-aided translations and annotations are opening up “a new intellectual space for the ancient discipline of philology.” Using well-known ethnographic and meta-analytic methods, the [article by Pasquetto, Borgman, and Wofford](#) addresses—from a social science perspective—a perennial question that has become increasingly

important because of open science: how do scientists find, reuse, and interpret data which they did not collect themselves?

Addressing a longstanding and complex problem in medical and health sciences, [Gibbons et al.'s article](#) provides a statistical tool for large-scale screening of increases and decreases in the suicide risk associated with over 900 medications, including all psychotropic medications, based on medical claims from over 150 million patients. Such methodologies and findings can help to reduce inappropriate prescribing practices that contribute to the alarming national and international suicide epidemic.

Spatial, Temporal, and Structural

Data share the physical and social dimensions of the objects that they intend to record and reveal. Christopher Phillips' [historical account on baseball statistics](#) reminds us, “Data are physical objects, subject to friction and corruption.” Another facet of data's temporal dimension is reflected by Stephen Stigler's warning that [data have a limited shelf life](#), a complementary view to Phillips' emphasis that “the stability and reliability of data is an accomplishment, not a natural state.”

Understanding these dimensions, therefore, is critical to the task of properly processing and analyzing data, and minimally to avoid misleading ourselves and generations to come. [Pasquetto et al.'s study](#) on the data creators' advantages demonstrates the truth of [Phillips' remark](#), “Data don't just emerge: they must be created,” while [Stigler's emphasis](#) on the “statistical worries about the effect of data selection on induction” should remind us of the detrimental or even fatal consequences of ignoring the temporal and social context of data.

Data are also created or collected in different spaces, literally and figuratively, and with different structures, such as with differential privacy protections, as reflected in the forthcoming 2020 census (a theme topic for *HDSR* in 2020). Peter Christen's [big picture-view of data linkage](#) details the complexity of integrating data sources derived from varied spaces, times, and structures, and states it clearly, “Data linkage is far from a completely solved problem, and the modern big data era is presenting various new challenges.” This is just one of many areas of data science where the problems present conceptual and technical challenges because of the great heterogeneity in the nature and structures of data.

Novice, Apprentice, and Expert

This trichotomy [described in Michael R. Berthold's article](#), from a skillset perspective, matches (in reverse order) the three student groups identified in [the first issue's editorial](#) by Harvard Provost Alan Garber. These include: first, the students who intend to become full-fledged experts in data science; second, those who want to use data science tools to advance fields of their primary interest; and third, those who need to gain a basic understanding of data science. Understanding the differences among these three skillsets and the needs of the corresponding learning groups is essential to effective pedagogical design and deliberation, as well as to

establishing and optimizing data science infrastructures in all societal sectors and, more broadly, our global community.

Compared to the rapid evolution of data science methods and technologies, it is not unfair to say that the research on effective data science curriculum and pedagogy for meeting these needs is mostly still in the Jurassic Period. Show me the curriculum from a data science master degree program, and I will tell you with replicable statistical confidence which species of dinosaur has put its weight down on it. This metaphor may sound harsh, but it is meant to remind ourselves that doing the best with old methods provides no assurance of meeting new demands, or even of our surviving the data science ecosystem. We all know that there is a huge hunger out there, and we all believe (rightly) that whatever we already teach can help to quench the thirst to some degree, perhaps by adding some additional courses from a neighboring department or two. This is a practical way to address urgent needs, but ultimately, our aims should be ecosystemic—that is, conducive to the flourishing of the entire data science ecosystem. Piecemeal fixes do not point toward or take the place of a much-needed curricular overhaul, aimed at addressing the different learning objectives and diverse backgrounds of students and trainees.

These great challenges come with great opportunities. [Robert Lue’s thought-provoking proposal](#) identifies one grand opportunity to use data science education as a general platform to foster truly inclusive learning. Lue’s main point, which reflects clearly the ecosystemic nature of data science, is that, “for the first time it is highly likely that every student will be able to connect what they care about with an approach or set of analytical tools from this emergent field.” The philosophical and practical considerations of data science, which I will discuss next, also encourage students—or anyone for that matter—to engage in and contribute to the data science ecosystem without being intimidated by the analytical sophistication of data science. One thinks in this context of the oenological ecosystem, which relies substantially on many connoisseurs, whose knowledge depth of fermentation or edaphology is no match for the depth of the wine they enjoy.

Data science, therefore, has the potential to provide the most versatile platform the world has ever seen for pedagogical engagement for learners of all ages, backgrounds, and interests. Future issues of *HDSR* will feature a series of articles to demonstrate and explore such opportunities, addressing challenges and providing content for data science education and training from classrooms to boardrooms, and every playfield in between.

Philosophical, Analytical, and Practical

Until very recently, data science courses and training programs focused almost exclusively on analytical and technical skills, such as programming, optimization, web scraping, and data pre-processing, analysis, and visualization. One of the most rapidly-developing areas of data science is the so-called ‘politics of algorithms,’ addressing issues such as algorithm accountability, auditability, bias, fairness, interpretability, and transparency. [The question posed by Cynthia Rudin and Joanna Radin](#)—“Why are we using black box models in AI when we don’t need to?”—is one of many such questions being raised and examined. The inherently philosophical

issues in this and similar areas are typically of broad societal relevance, especially with respect to morality and ethics. Our pedagogical content therefore needs to be adjusted to ensure that philosophical thinking, broadly construed, is a pillar, not a façade, of data science education.

In addition to the philosophical and analytical pillars, we need to establish a third pillar for data science education and training: making practical choices. The vast majority of real-life problems do not come with an analytically tractable, or even definable, optimal solution, regardless of how much data one has. Many of them involve difficult compromises, for instance, between data utility and data privacy. Sensible compromises rarely come out directly from the numerical output of our data science procedures. We need to help our students and trainees gain a deep appreciation of practical contexts as well as the short- and long-term consequences of our choices. We need to facilitate their development in the ability of conducting *principled corner-cutting*—that is, the ability to approximate with progressively fewer sacrifices as the resource constraints become increasingly relaxed.

Contrary to common misperception of the impracticality of philosophical contemplation, much of the philosophical thinking and practical considerations in the data science realm deal with the same trade-offs, dilemmas, conundrums, and so on—that is, with all complexities that render analytical methods insufficient in delivering satisfactory answers on their own. Philosophical thinking addresses the challenges from a broader, more abstract perspective, and provides the necessary guidelines, especially of an ethical nature, to guide practical choices in specific contexts. Therefore, data science training that is supported by only the analytical pillar is becoming increasingly inadequate, and indeed dangerous, for maintaining a healthy equilibrium of the data science ecosystem, and ultimately the human ecosystem.

There are likely other 3D surroundings, and more may emerge as the data science ecosystem evolves. But these five should suffice to remind us that there is always a data science dimension in which each of us may be immersed. With that in mind, I cannot resist inviting each of you, as I did in my first editorial, to help *HDSR* live up to its motto: *Everything Data Science and Data Science for Everyone*.

Acknowledgments and Notes on a Sequential Publication Plan

Like the first editorial, this one benefited from many individuals' comments and editing. I thank all authors of the second issue for providing sanity checks and moral support. Among them, Christine Borgman, Robert Gibbons, Christopher Phillips, and Stephen Stigler offered especially constructive suggestions. I also thank Suzanne Smith, Radu Craiu, and Robin Gong for their tireless editing of my Chinglish and for making the editorial more succinct and engaging. All tunnel visions, wishful thoughts, and oversights are mine, even though I really don't want them.

For those readers who access to the site before the end of November 2019, you may see (increasingly fewer) articles marked as "Forthcoming." Responding to a reader's suggestion of pacing the publication releasing, as

well as to optimize our production process, we are experimenting a sequential releasing plan, about one article per week after the initial launching of the issue with a majority of articles. We greatly appreciate readers' feedback on this plan and other suggestions, as our ultimate goal is to enhance the readers' experience and reader-author interaction via PubPub's commenting platform. Please send your feedback to datasciencereview@harvard.edu. THANK YOU!

©2019 Xiao-Li Meng. This editorial 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 editorial.