Harvard Data Science Review • 1.1

Data Science: An Artificial Ecosystem

Xiao-Li Meng

Published on: Jun 28, 2019

Updated on: Oct 04, 2019

DOI: [10.1162/99608f92.ba20f892](https://doi.org/10.1162/99608f92.ba20f892)

Type the word ‘scientist’ into your favorite browser and search for images. Most likely you will see photos of actual scientists from various fields. Now repeat the search, using ‘data scientist.’ You will see far fewer photos, but many animated figures standing by or pointing to various lists of must-have skills that read like tiger parents’ assignments for their children.

As we advance deeper into the digital age, our societal demands for data scientists naturally rise in both quantity and quality. Most of us have some knowledge about other kinds of scientists (granted that such knowledge can be quite flawed), but we are much less clear about who data scientists are and what they do. Indeed, what exactly is *data science* (DS)? As you may have guessed, the answer depends on whom you ask. Some say DS is CS (computer science). Others think DS is simply S (statistics). You may even run into someone who declares DS is just hyped-up BS (and I don’t mean “Bayesian statistics”).

The central mission of the *Harvard Data Science Review* is to help define and shape what DS is or should be. The inaugural issue of *HDSR* features articles by prominent researchers and educators from the humanities, engineering, sciences, and social sciences, as well as by leaders in government and industry. I hope you will agree that all of their articles properly belong in the realm of data science.

When something is too vast to enumerate, an effective mathematical strategy is to describe it by its complement—that is, what it is not. I will now list five things that DS is not, in approximately the decreasing order of their narrowness in conceptualizing DS, to address some common sources of confusion or even misconception about DS.

First, DS is not just machine learning or just statistics. MIT financial economist [Andrew Lo and his team](#) achieve unprecedented scale and accuracy in predicting drug approval because of a powerful integration of machine learning and statistical imputation, which permits the extraction of valid information from much more data than previous studies using machine learning alone. Similarly, in order to generate more reliable predictions concerning the question of disputed song authorship between John Lennon and Paul McCartney, my Harvard colleague [Mark Glickman and his co-authors](#) go beyond the traditional statistical models for resolving authorship disputes, and use various ideas and methods from both machine learning and statistics to achieve their notable success.

Second, DS is not all about prediction. Those who think it is or it should be may wish to consult the article by WarnerMedia Applied Analytics’ Chief Scientist [Nathan Sanders](#). His article offers insights from an industrial perspective on balancing prediction and inference, and showcases how such a balanced perspective benefits business decision making.

Third, DS is not only about data analysis. Columbia computer scientist [Jeannette Wing’s](#) data life cycle diagram makes it clear that data analysis is only one process in the long journey from data

generation to decision and action. I might add that even the data generation process itself can entail several steps, such as goal setting, questionnaire or experimental design, and field testing.

Fourth, DS is not a discipline that sits merely within STEM (Science, Technology, Engineering, and Mathematics) fields. University of Exeter’s philosopher [Sabina Leonelli](#) reminds us that the concept of data is a subject of philosophical research because there is no such thing as objective raw data. Questions of what to collect, how to collect and measure, and how to process data all involve human motivation, judgments, and preferences; each of these questions directly impact data collection, analysis, and the interpretation of results. At the other end of the data life cycle, UCLA’s information scholar [Christine Borgman](#) points out the “after lives” of data. Data, especially those that are complex and large, often get reused to address new problems or to make deeper probing. It is therefore crucial to have appropriate mechanisms to store and curate data, with proper provenance to inform future users and investigators. All of these issues go beyond what is taught traditionally in STEM fields, and yet they are vital for the validity and applicability of DS, especially in terms of ethical ramifications and long-term impact.

Fifth, and most critically, DS is not even a single discipline by itself. Increasingly, there is a general recognition that because DS has evolved in such a diverse way, it is unwise to use a list of must-have skills to conceptualize it as a single discipline. Like science, social science, or humanities, DS is best understood as a collection of disciplines with complementary foundations, perspectives, approaches, and aims, but with a shared grand mission. That is, to use digital technologies and information of any kind to advance human society as a harmonious, responsible, and vital ecosystem.

It is therefore useful to have the umbrella term ‘data scientist,’ just as it is useful to have terms such as ‘scientist,’ ‘social scientist,’ or ‘humanist,’ but these latter terms do not come with the same great expectations as does the former. We do not expect a scientist to possess expert knowledge in all major scientific fields, such as astronomy, biology, chemistry, environmental science, and physics. Nor do we expect a humanist to know about all major civilizations and religions, or to read and write in all major languages (even if I know some amazing colleagues who do!).

The hiring dilemma for data scientists as described by the Chair of the Department of Statistical Sciences at the University of Toronto, [Radu Craiu](#), hints at structural complications that arise when the nature of DS is not properly framed. With its more than 4,500 (yes, with double zeros) undergraduate students—a figure intensely heart-warming to me as a statistician—Craiu’s department is flourishing. At the same time, this unprecedented figure should serve as a cautionary reminder to universities contemplating the creation of a Department of Data Science, instead of, more wisely, a School or a Division of Data Science, such as is currently being established at the [University of California, Berkeley](#).

There are good reasons why we almost never see a Department of Science or a Department of Humanities in major universities, but rather Schools or Divisions bearing these names. The issue is not just about creating sufficiently cohesive academic hiring units with shared research expertise and discourses. It is also about educational quality and pedagogical effectiveness, as well as the kind of builders, leaders and thinkers our education programs help to create for the future DS workforce. Just as we need primary care physicians and medical specialists, so do we need both generalists and specialists of DS in order to effectively address many challenges and opportunities unique to our digital society, from precision medicines, to cybersecurity, and to smart cities (all theme topics for future issues of *HDSR*). The training of either group cannot be done adequately in the context of narrowly-structured educational units, which tend to have the least resources (as well as the ability to raise them) to deal with emerging challenges or seize new opportunities.

The broad nature of DS is also evident from the report on “[Data Science for Undergraduates: Opportunities and Options](#),” issued by a committee of the National Academies of Sciences, Engineering, and Medicine (NAS). As a simple indication, committee members come from computer science, statistics, engineering, natural sciences, social sciences, and the humanities. In [his interview](#) of the committee’s co-chairs Laura Haas and Alfred Hero, *HDSR* co-editor on DS education, Rob Lue, asks thought-provoking questions, including how to use the encompassing nature of DS as an overarching theme to reimagine liberal arts education. The interview highlights the diverse nature of DS undergraduate education but with a concrete aim: to help students gain *data acumen*, in contrast to a laundry list of skills unattached to specific purposes or higher goals.

Harvard Provost [Alan Garber’s](#) editorial addresses a broader question: what does an educated citizen need to know about DS? Garber’s list is far more selective than many aforementioned lists of skills, due to the broader question that he addresses and his deep understanding of pedagogical goals and effectiveness. The column on short tutorials, *Diving into Data*, edited by eminent statistician and writer [David Hand](#), is part of *HDSR*’s effort to answer Garber’s call, with its first tutorial devoted to the understanding of the concept of statistical modelling.

It is extremely fortuitous that this editorial was conceived on my way to attend this year’s ceremony of the [Turing Award](#), given to Yoshua Bengio, Geoffrey Hinton, and Yann LeCun. The award is richly deserved, because their deep learning methods can extract patterns from chaos in ways that many once thought—and some still think—implausible. However, just as the Berkeley computer scientist and statistician [Michael Jordan](#) questions whether it is time to retire the Turing Test, the seemingly magical power of deep learning reminds us that we urgently need deep understanding and deep thinking in order to avoid getting ourselves into deep trouble.

Jordan’s provocative article on AI does exactly that. It triggers reflection and questions from 11 thinkers and leaders within and outside of academia. These exchanges, together with University of

Pennsylvania's historian of science [Stephanie Dick's](#) historical account of AI, make it clear that AI is evolving into an *artificial ecosystem*, rather than a single discipline with well-specified goals (e.g., creating machines that mimic the human brain). Natural ecosystems are susceptible to natural disasters. An artificial one can suffer from artificial disasters, all the more so when its inhabitants do not have a good understanding of its operating characteristics or internal principles. The study conducted by Oxford philosopher [Luciano Floridi and his co-author](#) on the principles of AI for society demonstrates the heightened awareness of the risk of such destructive events, and the efforts to reduce both their frequency and severity.

Like AI, DS is evolving into an artificial ecosystem; the term 'artificial' highlights both the fact that DS is a human construct and that it depends critically on computing advances. The extent of this dependence is "painlessly" demonstrated in the article on massive computing for DS by the Stanford statistician [David Donoho and his collaborators](#) from computer science and statistics. Epistemologically recognizing DS as an ecosystem should help to ameliorate the unproductive, ego-driven friction between disciplines that drives attempts to claim primary ownership of the DS enterprise.

In the same way that it takes the collaborative efforts of all nations to maintain the well-being of the world, so it will take all of our collective efforts to ensure the health and growth of the DS ecosystem. We need at least computer science, statistics, engineering and operation research, information and library sciences, law and philosophy, (applied) mathematics, social and behavioral sciences, history of science, and data visualization, not to mention countless areas of application from astronomy to zoology and back to agriculture, and the vital participation of industry, government, NGOs, and beyond. The ecosystem paradigm also reminds us that there is and will be climate change, for better or worse. What's cold today could be hot tomorrow, and what's cool tomorrow might be lukewarm the day after. Our educational structures, content, and deliveries therefore should be eco-friendly and weatherproof, not merely demand-driven.

Furthermore, it is not unreasonable to imagine that the term 'data science' itself may evolve, because the notion of 'science' is already inadequate to properly convey the breadth of DS as it is now. For example, few would argue that the digital humanities do not belong to DS, but digital humanities are neither about the science of data nor about using data to advance sciences, the two common identifiers of DS. However, if DS stands for 'data studies,' then surely the digital humanities are about studies of data and use data to advance studies. Of course the term 'data studies' is far less catchy than 'data science' and hence unlikely to catch on. Nevertheless, we should always be mindful of the volatile and complex nature of any ecosystem, especially during its formative stages.

Finally, the eco-systemic nature of DS levels the playing field for everyone. It requires that we all interact and contribute, whether you are a bright star who needs no introduction or a bright starter who seeks every introduction. Regardless of who you are, I am delighted to introduce you to the

Harvard Data Science Review, which aims to keep you informed, engaged, and intrigued by DS. Your feedback, via on-line comments, letters to editors, and perspective articles, will help *HDSR* live up to its motto: *Everything Data Science, and Data Science for Everyone*.

Disclaimers and Acknowledgements

The views expressed in this editorial are neither Harvard's nor entirely mine. I have benefitted from discussions and debates with countless stars and starters of DS, especially with many editorial members of *HDSR* around the world (e.g., the notion of DS as 'data studies' was conceived in Christine Borgman's kitchen and study room). Deep gratitude to Radu Craiu and Robin Gong for multiple rounds of correcting my Chinglish and more serious flaws, and to Suzanne Smith for making me start all over again just when I thought I could sit down to enjoy an issue of *Wine Economics*.

I am deeply grateful to hundreds of individuals for their collective efforts, which enable *HDSR* to be launched exactly one year from its conception. On the content side, I thank *HDSR*'s co-editors, associated editors, authors, and reviewers, for generating and ensuring the quantity and diversity of the articles in this issue and forthcoming. On the production side, I thank the MIT Press and Knowledge Futures Group and *HDSR*'s editorial office, as well as *HDSR*'s advisory board, for their tireless effort and strategic planning. I also thank Harvard's Provost office, Harvard's and MIT's Offices of General Counsel, and many friends of *HDSR* for their multi-dimensional and on-going support. Finally, I owe a big thanks to the Co-Directors of Harvard Data Science Initiative (HDSI), Francesca Dominici and David Parkes, for endorsing my spur-of-the-moment idea of learning from the models of *Harvard Business Review* and *Harvard Law Review* but expanding them with an education focus to create *HDSR*, and to the Executive Director of HDSI, Elizabeth Langdon-Gray, for going above and beyond to support the launching of *HDSR*.

This article is © 2019 by Xiao-Li Meng. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author identified above.