

Harvard Data Science Review • Issue 5.2, Spring 2023

Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop

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

The MIT Press

Published on: Apr 27, 2023

DOI: <https://doi.org/10.1162/99608f92.68a012eb>

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

The term human-in-the-loop (HITL) generally refers to the need for human interaction, intervention, and judgment to control or change the outcome of a process, and it is a practice that is being increasingly emphasized in machine learning, generative AI, and the like. For readers who have not enjoyed machine-assisted learning or who are comfortable with their human intelligences, it would not surprise me if some of you would wonder, upon hearing the term HITL, ‘Wait, when did we take humans out of the loop?’ The diversity of human minds guarantees that such a question would generate reactions ranging from being inspired to being despised. Regardless, however, the arrival of ChatGPT and other AI chatbots has made the mostly academic debate of machine intelligence versus human intelligence into a household conversation piece. It is therefore an apt time to turn such a rhetorical question into retrospection and reflection.

With my human thinking in the loop, I probably can take any issue of *HDSR* to remind ourselves of the roles we humans play in the data science (DS) ecosystem. Indeed, as I wrote in the inaugural editorial, titled “[Data Science: An Artificial Ecosystem](#),” “the term ‘artificial’ highlights both the fact that DS is a human construct and that it depends critically on computing advances” ([Meng, 2019](#)). (The same can and should be said about data engineering (DE), which can advance faster than DS, as we witnessed recently with the progression from GPT-n to GPT-(n+1), $n < 4$.) However, [the current issue](#) is particularly fitting for this contemplation because it is the first *HDSR* issue with a special theme dedicated to a beautiful mind that came with a beautiful heart, a term that yet to be applied to any machine. As such, I ask for readers’ indulgence—and my fellow data scientists’ forgiveness—for creating sibling rivalries to HITL.

Human in the Loop: From Developers to Deployers

As cliché as this may sound, humans are always in the loop for advancing data science or engineering at all fronts, literally and figuratively. Indeed, for those who don’t mind the label of anthropocentrist, the paraphrasing ‘DS/DE are of the people, by the people, and for the people’ might be a more arousing title for this editorial. Data are typically digital representations of some people’s mental activities or physical endeavors, and they are collected because some (other) people see the benefit to doing so. The benefits can be for the data subjects, the data collectors, or other people, which often include those yet to be born. To realize and optimize the benefit, yet more groups of people would be required for developing methods, implementing algorithms, interpreting and deploying the results, and so on, in order to reap the perceived benefits.

For the [Recreations in Randomness column](#), [Kavya Mehul Shah and Ammaar Ahmed Saeed’s \(2023\)](#) account on “[Innovating a Centuries-Old Sport: How Emerging Data Analytics Tools Are Redefining Cricket](#)” provides a concrete illustration of such human effort. For example, cricket highlights are short video clips that are collected because they can benefit players, coaches, and organizations. Traditionally the creation of these highlights “requires sports broadcasters and video editors to manually parse through hours of footage to identify exciting moments”—clearly a not very exciting process to say the least. Automated highlights generation is then a very welcome advance that requires both data science (e.g., semantic segmentation to

identify “exciting moments”) and data engineering (e.g., effective software for routine applications). Much progress has been made in the last 20 years, as reviewed in the article. Yet the 2 decades of progress also indicate the amount of human effort required to *automate* the creation process, not to mention human effort needed in translating highlights into the desired benefit.

As another indication of the human effort, when I initially opened the document file for this column article to prepare for this editorial, I was a bit surprised to see that the word count was about 4,600—far exceeding the normally expected 2,500 words for a column. In case a reader wonders why I was not aware of the length of a submission until the end stages, I am grateful to the interim Co-Editor-in-Chiefs, Francesca Dominici and David Parkes, whose dedication allowed me to take a real sabbatical (which incidentally gave me time to appreciate the benefit of sports). It turns out that the article has 65 references (which are not counted towards the 2,500 limit), which is an unusually extensive list even for a regular article. But it is fitting for this column since the list provides a chronological picture of human effort and progress in reshaping a half-millennium-old sport via data science and engineering. Indeed, it is almost certain that the list is still only a (non-random) sample.

... and From Software Engineers to Data Science Managers

The sabbatical year also afforded me the opportunity to join an international collaboration on assessing living conditions globally in general, and in Africa in particular, by combining ground-level survey data with satellite image data via deep learning. The data involved are treasures not only for the mission of the collaboration, but also for data science and engineering education in general. They exhibit almost the full spectrum of complexity and defects as routinely encountered in practice, including selection bias, measurement errors, loss of information due to privacy protection, mismatches in spatial, temporal, and contextual resolutions, and incomplete documentation of data provenance, just to provide a partial list. The collaboration team is also multi-faceted, involving multiple disciplines (e.g., computer science, engineering, sociology, statistics) and members in all career stages from undergraduate to senior researchers with a variety of working styles and who are driven by different interests or incentives. This collaborative experience, which is still ongoing, deepened my appreciation of two kinds of humans in the loop, so to speak.

One kind of profession is software engineers, whose technical skills and teamwork (or lack thereof) typically govern the pace for progress. For complex projects, many methodological ideas need to be (stress) tested. These tests can take weeks or minutes, all depending on the engineers’ experiences and skills. For the project I am involved in, a clever implementation circumventing a map computation with mismatched resolutions sped up the test by several weeks. With those experiences still fresh, I literally thought I was daydreaming when I saw the article “[Software Engineering Practices in Academia: Promoting the 3Rs—Readability, Resilience, and Reuse](#)” by [Andrew Connolly, Joseph Hellerstein, Naomi Alterman, David Beck, Rob Fatland, Ed Lazowska, Vani Mandava, and Sarah Stone \(2023\)](#). The first section title says it all: “Good Data Science Requires High-Quality Software Engineering.” Its diagnosis of the major issues in academia is spot on, as academics tend to

“address their software development needs by hiring graduate students, often without any formal training in software engineering,” and the vast majority of the software projects are created in an often-inefficient ‘3S’ setting: single developer, single user, and single study. For any academics who are interested in improving the quality of their data science projects and of their life by reducing work inefficiency, the time invested on digesting this article is likely to be negligible compared to the time saved moving forward, because the article provides real-life successes, reasons, and recommendations—with specific recipes—for adopting the ‘3R’ principle.

The other kind of profession is data science managers. With a team of diverse expertise, interests, and work habits and schedules, and the nature of the project being multi-phased, having an all-around data science manager is another key for both the quality and speed of the project. For the collaboration I am involved with, things changed substantially once we had such a data science manager in place. A data science manager is more like a chief of staff to the intellectual leaders of a data science collaboration, which means that the manager needs to have sufficient technical, personal, and organizational skills to be effective. A crucial task for a data science manager is to help to identify, establish, and sustain a collaborative platform and working culture that is conducive to making the whole more than the sum of its parts, or at least not less. The latter can easily happen (and did happen even in my very limited experience) when there are miscommunications, misunderstandings, and mismatched expectations among different team members or groups. The manager needs to be a strategic thinker with sufficient know-how to recognize or anticipate weak links, devise solutions or preventive measures accordingly, and ultimately create an incentivizing infrastructure to enhance team efficiency and team members’ desire to further the collaboration.

With these reflections from my recent experiences, I was instantly attracted by another article in this issue, [“Managing Embedded Data Science Teams for Success: How Managers Can Navigate the Advantages and Challenges of Distributed Data Science,”](#) by [Marta Stelmaszak and Kelsey Kline \(2023\)](#). Through a case study, the authors summarize three advantages for the embedded structures, “1) agility in delivering quicker and better tailored analytics, 2) concentrated and cultivated data science expertise, and 3) emerging bottom-up solution ideas,” as well as three disadvantages: “1) growing away from the enterprise infrastructure, 2) duplicating data science efforts, and 3) weakening links with strategic impact.” Although the setup here is rather different from academic collaborations, the main conclusion of the article is the data science manager is the critical HITL to reap these advantages and curtail the disadvantages. To do so, the article argues that data science managers need to have aptitudes for all three managing directions: within a team (e.g., ability to balance workload for members with varying expertise), vertically in an organization (e.g., interpersonal skills to maintain team visibility at all levels of the organization), and horizontally across the organization (e.g., capability to develop networks with other data science teams).

The vital role of managing is furthered demonstrated in [“To Deploy Machine Learning, You Must Manage Operational Change—Here Is How UPS Got It Right”](#) by [Eric Siegel \(2023\)](#). The article starts with a telling

story of how a trash collection company dumped a university generated route-optimizing plan that could cut its number of trash tracks by 50%. I won't spoil the rest of the plot, which could easily be adapted into a script for a Hollywood movie on machine learning or AI (perhaps with the help of GPT-n), but only to repeat the metaphor Siegel used. The "rocket science" of machine learning for commercial use is easy. The hard part is getting the end product or process launched, which requires changing humans' habits. That's not rocket science; that's beyond rocket science.

Human Behind the Loop: With Beautiful Minds and Beautiful Hearts

Much of the research of and discussions about HITL focus on the interaction between humans and machines. But as we just discussed, translating and transforming the advances in data science and engineering into actual value for human societies is a matter of interactions among humans. Arguably another kind of human interaction is even more important for the data science enterprise, namely the professional and interpersonal interactions among members of the data science ecosystem. A shining example is provided by the [special theme in this issue celebrating the life and work of Sir David Cox](#), a giant in statistics and in science in general.

In his long and extremely distinguished career, David served as a global model in everything he did. He was an extremely deep, broad, and influential scholar, a much-loved advisor of over 60 doctoral students, an exceedingly effective editor for *Biometrika* for 25 years, and a well-respected academic and professional leader (e.g., serving as Warden of Nuffield College, and as president of multiple professional societies). The 20 articles in the special theme document many interactions with David in a myriad of ways and across many decades (and countries). Yet they reflect a common theme: the lasting impression and impact of the interactions. I am deeply grateful to the co-editors of this special theme, Sylvia Richardson and Nanny Wermuth, for organizing and introducing ([Richardson & Wermuth, 2023](#)) such a rich and heartwarming collection, and to all contributors for sharing their personal stories with all of us, to which I add my own.

My first interaction with David took place when I was a PhD student in late 1980s, where I co-authored an article that was submitted to *Biometrika*. At that time, the review processes for most statistical journals were notoriously slow; it was rather common for the review process to take more than 6 months for the first round. *Biometrika* was an exception, despite having the smallest editorial board among all top-tier journals. Even with that expectation, I was rather surprised when we heard from David within 2 weeks of submission, and it was not a quick rejection letter. Rather, David had provided detailed comments and asked us to do a revision before he sent it out for a full review. I do not recall the details of David's comments, but I remembered distinctively being very impressed by the promptness and insightfulness of the comments, which led to a much-improved revision, and ultimately my first publication in the tier-one journals in statistics.

Years later I had an opportunity to meet with David in person, when I could no longer hold my curiosity and hence asked how he could handle so many articles personally, and yet with such efficiency. The answer he

gave can be found in [A.C. Davison's \(2023\)](#) reminisces: David read every submission on the daily commuting Tube ride. But of course, that is only half of the story since reading is not the same as coming up with insightful and persuasive comments. For that, few could compete with David's encyclopedic knowledge, deep insights, and rich intuition. Like [Davison \(2023\)](#) who tried to “emulate his quick turnaround of submissions and constructive suggestions to authors,” I too tried to follow David's model in my own editorial work, only to find how challenging that is—it is certainly not nearly as easy a ride as David's daily Tube ride.

In the world of leading scholars, there must be many minds that are as beautiful as David's intellectual mind, to borrow the Hollywoodized phrase ‘beautiful mind.’ But having a beautiful mind does not necessarily make one a great educator, and indeed in some academic circles, a negative correlation between the two has been suggested. As reported by my colleague [Joe Blitzstein \(2023\)](#) in the [Bits and Bites column](#) of this issue, David's beautiful mind for research came with a beautiful heart for education. About a decade ago, Joe and I co-taught a course on reading Cox's many contributions, and we invited David to come to provide a unique opportunity to students to meet an intellectual giant in person. He agreed very promptly, but ultimately a health issue (and his doctors) prevented him from making the international trip. He kindly offered to communicate with students in writing, which the students of course were thrilled to have the chance to directly ask arguably the world's most preeminent statistician alive then. They hence asked many tough questions, which were summarized and sent to David. We of course only expected him to answer some of the questions, knowing how busy he was—he was ageless in terms of professional activities, and consequently in extremely high demand. To our great surprise, David provided detailed responses to essentially all questions. I was therefore so pleased when I was reminded by Joe that he had kept this precious document. I immediately invited Joe to annotate it so we could share with the world this hidden set of treasures of Cox's insights and wisdom. It is a great testimony to David's beautiful heart for education and his care for future generations, a repeating theme that is also vivid in the 20 articles celebrating Sir David Cox, a true model for all (data) scientists.

Human Above the Loop: Oversight and Insight

A key public concern of ChatGPT and the like is that AI will replace humans in many ways, especially regarding jobs. There are indeed jobs that will be done by AI—I certainly have used ChatGPT increasingly so for checking my Chinglish instead of sending my draft to a human editor. However, there is one kind of human function that will never be replaced by machine or AI because of its very nature: human's oversight, whether over machines, processes, or people themselves. [Pavle Avramović's \(2023\) “Digital Transformation of Financial Regulators and the Emergence of Supervisory Technologies \(SupTech\): A Case Study of the U.K. Financial Conduct Authority”](#) provides ample reasons why human oversight is always needed for financial markets, especially with their adoption of advanced technologies. For example, human oversight is necessary to ensure the functionality of the technologies as intended, curtail their unintended consequences, and prevent their abusive and malicious use by some people. To do this well, human oversight itself should utilize technologies, as long as we understand how these technologies fulfill our supervision and regulatory goals.

Avramović's article provides a rather revealing case study to demonstrate how this can be done but cast through a theoretical lens by examining institutional factors and dynamic capabilities. It also provides an overview of the emerging field of SupTech, as well as an outlook for the field based on the lessons learned for research, regulators, and industry.

Indeed, learning from lessons is another human endeavor that cannot be replaced by machine or AI, for a very simple reason: humans need to internalize the lessons for themselves in order to actualize the meaning of *learning*, which is very different from its meaning as in the phrase 'machine learning.' Machine learning is an algorithmic process of revealing patterns and organizing knowledge (hopefully). Learning from a lesson requires a changing of behavior together with a changing of the tendency of having such behavior to ensure that either change is sustainable. The need of both changes to actualize the learning is perhaps best illustrated by the joke or reality for some: 'Quitting smoking is easy—I have done that many times.' Or in the case of ChatGPT, it will apologize profusely when told that its answer is wrong, but that does not necessarily mean that it would change its answer. Even when it does correct its answer in a particular case, it will repeat the same kind mistakes when prompted with a similar but differently worded instruction, as demonstrated by its persistent ability to 'hallucinate' (e.g., by making up many plausible-looking but hopelessly wrong biosketches.)

Granted, sometimes we humans do no better than ChatGPT in this regard, or do not want to do better, when it comes to changing behaviors because of negative consequences. But even in that regard, humans are still one step ahead of ChatGPT, at least for now, as we often rationalize our behaviors by changing our evaluative metrics, perhaps subconsciously. However, when our humanity and humility are at their best, we not only actualize the lessons learned, we turn them into insights to generate principles for success. The article "[Frontline Data Science: Lessons Learned From a Pandemic](#)" by [Emily A. Beck](#), [Hannah Tavalire](#), and [Jake Searcy \(2023\)](#) does exactly that. The COVID-19 pandemic highlights both the importance and challenges of using data science for disaster response, with lack of accessibility, training, and flexibility being particularly problematic. Beck et al.'s article outlines the authors' experience building a COVID-19 Monitoring and Assessment Program and provides lessons learned and principles for success for future data scientists wishing to contribute to emergency response and disaster recovery efforts. It provides a particularly useful perspective because neither Beck nor Searcy started with the pertinent experience or training in epidemiology or health care, one being a geneticist and the other a particle physicist. Like many data scientists, we all had and have the strong desire to do something, as we know we can and should contribute to the fight against the pandemic, but we didn't or don't know how. The article's delineation from getting started to informing future data scientists therefore is especially welcome and helpful.

In particular, the lessons from the article have broader implications than just dealing with pandemics or disasters. For example, for anyone who truly cares about data science done right and has gotten their hands dirty, the main lesson of the article will resonate extremely well (replacing "testing" with 'study' in general),

“The most important lesson learned ... was that the data science team needed to be involved in the design of every aspect of testing for ensuring accuracy and quality assurance of data pipelines” (Beck et al., 2023). The article makes it clear that this necessity is mainly created by humans. Human errors are unavoidable, from typos on patients’ data to system glitches preventing prompt reporting to state agencies. But “data scientists can rapidly craft quality assurance and compliance solutions to catch and fix errors.” Together, the authors’ experience and their four major lessons and five principles for success provide an exceedingly informative and concrete demonstration that data science is of the people, by the people, and for the people.

Disclosure Statement

Xiao-Li Meng has no financial or non-financial disclosures to share for this editorial.

References

- Avramović, P. (2023). Digital transformation of financial regulators and the emergence of supervisory technologies (SupTech): A case study of the U.K. financial conduct authority. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.7a329be7>
- Beck, E. A., Tavalire, H., & Searcy, J. (2023). Frontline data science: Lessons learned from a pandemic. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.eea782bf>
- Connolly, A., Hellerstein, J., Alterman, N., Beck, D., Fatland, R., Lazowska, E., Mandava, V., & Stone, S. (2023). Software engineering practices in academia: Promoting the 3Rs—readability, resilience, and reuse. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.018bf012>
- Davison, A. C. (2023). Some reminiscences of David Cox. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.85038493>
- Meng, X.-L. (2019). Data Science: An artificial ecosystem. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.ba20f892>
- Richardson, S., & Wermuth, N. (2023). Remembering David Cox. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.35281d57>
- Siegel, E. (2023). To deploy machine learning, you must manage operational change—Here is how UPS got it right. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.73fc83a3>
- Shah, K. M., & Saeed, A. A. (2023). Innovating a centuries-old sport: How emerging data analytics tools are redefining cricket. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.1ea13423>

Stelmaszak, M., & Kline, K. (n.d.). Managing embedded data science teams for success: How managers can navigate the advantages and challenges of distributed data science. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.1f068331>

©2023 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.

References

- Avramović, P. (2023). Digital Transformation of Financial Regulators and the Emergence of Supervisory Technologies (SupTech): A Case Study of the U.K. Financial Conduct Authority. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.7a329be7> ↵
- Beck, E. A., Tavalire, H., & Searcy, J. (2023). Frontline Data Science: Lessons Learned From a Pandemic. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.eea782bf> ↵
- Blitzstein, J. K. (2023). Sir David R. Cox: A Beautiful Mind With a Beautiful Heart. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.74fb2783> ↵
- Connolly, A., Hellerstein, J., Alterman, N., Beck, D., Fatland, R., Lazowska, E., Mandava, V., & Stone, S. (2023). Software Engineering Practices in Academia: Promoting the 3Rs—Readability, Resilience, and Reuse. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.018bf012> ↵
- Davison, A. C. (2023). Some Reminiscences of David Cox. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.85038493> ↵
- Meng, X.-L. (2019). Data Science: An Artificial Ecosystem. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.ba20f892> ↵
- Richardson, S., & Wermuth, N. (2023). Remembering David Cox. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.35281d57> ↵
- Shah, K. M., & Saeed, A. A. (2023). Innovating a Centuries-Old Sport: How Emerging Data Analytics Tools Are Redefining Cricket. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.1ea13423> ↵
- Siegel, E. (2023). To Deploy Machine Learning, You Must Manage Operational Change—Here Is How UPS Got It Right. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.73fc83a3> ↵
- Stelmaszak, M., & Kline, K. (2023). Managing Embedded Data Science Teams for Success: How Managers Can Navigate the Advantages and Challenges of Distributed Data Science. *Harvard Data Science Review*, 5(2). <https://doi.org/10.1162/99608f92.1f068331> ↵