Harvard Data Science Review • Issue 3.2, Spring 2021

Building Data Science Infrastructures and Infrastructural Data Science

Xiao-Li Meng[1,2]

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

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

Published on: Apr 30, 2021

DOI: https://doi.org/10.1162/99608f92.abfa0e70

If you have never heard of the term 'support vector machine (SVM),' you can easily find its meaning or even a tutorial via any search engine (which, by the way, might itself be supported by a support vector machine). Prior to that search, you might have guessed that the term has its origin in machine learning. But at the turn of the 21st century, such a guess was not quite so intuitive. A bit more than 15 years ago, a colleague in a university that shall remain nameless discovered this the hard way—by applying for an internal research grant on SVM. He was confident that his application would be approved, not only because internal grants usually have simpler approval processes, but also because the topic of machine leaning was rather cutting-edge back then, at least for the statistics community. To his surprise, however, his application was rejected swiftly on grounds that left him unsure as to whether he should laugh or cry: "Funding for equipment cannot exceed $5000." (I must confess, however, that it took me a while to understand why SVM, a rather intuitive method for classification, needs to be labelled as a 'machine.')

This is only one tiny example of the important role of research infrastructure in university settings. At minimum, inadequate grant administration or support can cause individual frustration. More importantly, it can delay the advancement of a field, particularly when the field in question poses challenges to existing infrastructures and demands the creation of new ones. Such is the case with the data science ecosystem, which, due to its inherently transdisciplinary nature, generates many synergistic endeavors that call for the development and institutionalization of new infrastructures in administrative, educational, and organizational domains, to name just a few.

This issue of *HDSR* takes a close look at the institutional effort made by University of California, especially its Berkeley campus, to build infrastructures and programs for data science research and education. It starts with a conversation with President Michael Drake of the University of Californian system, and with Jennifer Chayes, Associate Provost of the Division of Computing, Data Science, and Society and Dean of the School of Information, at UC Berkeley. The conversation covers topics on how data science fits into the University of California's broader mission of education, research and public service, and more broadly how data science impacts society and higher education in general. Two discussion articles, one of them by Chayes, detail Berkeley's visions and actions in building a university-wide entity for data science and computing. A discussion of Chayes' article by all three deans of its Donald Bren School of Information and Computer Sciences (founding dean Debra Richardson, former dean and present UC Irvine provost Hal Stern, and current dean Marios Papaefthymiou), summarizes UC Irvine's history, experiences, and efforts in establishing and sustaining infrastructures and programs for Computer Science, Informatics and Statistics, and now more broadly for Data Science.

Together with over a dozen discussions from other institutions, they provide a timely snapshot of current infrastructure building efforts in universities around the world. Refreshingly, some of the discussions remind us

that data science itself serves as a 'virtual infrastructure' for scientific research enterprise, global society, and—more broadly—the human ecosystem, a theme echoed by multiple other articles in this issue.

Most of the discussions of systems and programs in this issue, including this editorial, focus on academic and educational institutions. Judging from a host of articles featured in previous issues of *HDSR*, however, similar revolutions and evolutions are taking place in industrial and government sectors as well. The recent partnership of *HDSR* with USA for IOM (UN Migration Agency) and USA for UNHCR (UN Refugee Agency) to co-organize the upcoming World Migration & Displacement Symposium **(**May 11–13, 2021; free and open registration) further reminds us that the NGO sector is also in great need of data science infrastructures and infrastructural data science, a topic to which *HDSR* will turn in an upcoming special theme following the symposium.

Building Data Science Institutions—What's in the Name?

Universities around the globe are currently racing to build at least one data science entity on their campus. The race is intense because building organizational structures requires a large amount of time, resources, and leadership skills, and those who build early and better will have a higher chance of attracting more of the best talents. These talents include the obvious ones, such as students at all levels and faculty from all disciplines that contribute to data science. The competition is even more intense for rarer talents (who are by definition in short supply), such as visionary institution builders and leaders who are also internationally-renowned data scientists; highly skillful research and computing support staff who have a good understanding of the need and habits of students and faculty; and entrepreneurially-spirited human resources experts who see traversing and breaking tangled and protean bureaucratic skeins as a career dream rather than a nightmare.

The variety of data science organizational infrastructures is a tell-tale sign of their rapid organic evolution. It is the result of many local optimization strategies that were executed under constraints of time, resources, and politics, and without a global objective function (to borrow a term from data science). David Madigan, Provost at Northeastern University, highlights this variety in his discussion of Chayes's article, and notes that, just as Berkeley is building a *College of Data Science*, the University of Virginia announced in 2019 a plan to establish a *School of Data Science*, with a physical infrastructure scheduled to be built between 2021-2024. In 2015, Fudan University (my alma mater) started to build simultaneously a *School of Data Science* and an *Institute of Big Data*. Columbia and Northeastern, the two institutions at which Madigan has served in leadership roles, have established respectively a *Data Science Institute* and *The Institute for Experiential AI.*

Other institutions, such as Cornell and Yale, have chosen to rename their statistics departments as *Department of Statistics and Data Science* or have established new departments devoted to branches of data science; see, for example, the *Department of Biomedical Data Science* at Stanford. As I summarized in my first editorial for *HDSR*, a consensus has increasingly emerged that the academic department is not a possible or at least not optimal unit to host data science in its entirety. These renamed departments or domain-specific data science

departments have more focused scopes and missions than an all-encompassing 'Department of Data Science,' and hence they are less at risk of outpacing their intellectual and other resources. They serve the timely purpose of meeting (at least partly) immediate research and education needs, even as their host institutions explore and develop longer-term and more holistic infrastructures.

The urgency of putting data science on the map—and on the campus—has pushed more institutions to adopt organizational units that require for their construction the least resources and the fewest political entanglements with existing infrastructures. This has led to the creation of many more data science institutes, or entities with various other creative or clever designations, than it has to the creation of data science schools or departments. For example, at Harvard, one finds the Harvard Data Science Initiative (HDSI), which was designated as an 'initiative' to signal its exploratory nature, and thus to reduce potential concerns regarding the perpetual competition for limited resources, among other considerations. Not incidentally, this designation also encourages the initiative to build sustainable plans and funding resources. (Finally, it has the benefit of promising continuity with respect to the HDSI acronym when and if the 'initiative' becomes an 'institute.') It is worth emphasizing that data science initiatives are by no means built only at research universities. Liberal arts colleges are pushing their frontiers as well, as evidenced by the data science initiative at Colby College.

As another example, at MIT, we find a *College of Computing*, whereas all its other multi-department units are designated as *schools*. To many readers, the distinction between colleges and schools must be either confusing or too academic. Well, it is. However, the hitherto non-existent nature of the 'college' designation at MIT helped to accelerate the establishment of its Schwarzman College of Computing since it had fewer institutional hoops to jump over than yet another 'school' amongst many schools would have had to surmount. More importantly, the designation allows the new entity to break free from disciplinary divisions as traditionally signified by schools' respective identities and missions, making it easier for the new college to interact with every existing school. The grand aim here is to enable the new college to achieve a broader impact, in terms of research, education, and societal outreach than that of the traditional schools, as emphasized by MIT President Rafael Reif in his conversation with *HDSR* in issue 2.3.

During the same conversation, President Reif explained why MIT faculty (not the administration) decided, after a long and intense discussion, to name the new entity the College of Computing, instead of the College of Data Science: "their conclusion was that 'computing' was much more general and much more inclined to not change in decades to come." The same discussion or debate at another institution, however, may lead to a different conclusion. Indeed, Chayes' emphasis on Berkeley's infrastructural attempt to address the challenges created by "Data Science and Computing" signifies the interwoven nature of the two terms and an acknowledgement that they do not subsume each other. This acknowledgment is important not only for building broader alliances but also for anticipating larger infrastructural needs, especially for those technological advancements that are essential for computing but are generally not considered to be an integrated part of data science (e.g., computer hardware). Of course, some of my statistics colleagues might

ask, "What about statistics?" I tried in vain to think of any topic in statistics that does not belong to data science (but not the other way around). If you can think of any such topic, I'd be very interested to know of it.

Proper naming or framing is also a matter of insightfully recognizing the vicissitudes and scope of a given intellectual landscape and its horizon of broader societal impact. In his discussion of Chayes' article, Munther Dahleh, the Director of MIT's *Institute for Data, Systems, and Society* (IDSS, a part of the Schwarzman College of Computing), provides a deeply perceptive explanation of the term 'Systems' as it figures in the institute's name. I must confess that I had complained to others about what I took to be the slightly exotic nature of the IDSS's name, and expressed the wish that at least one of the S's would stand for 'Statistics.' Reading Dahleh's discussion made me realize that my complaint reflected my ignorance and disciplinary blind spots. Dahleh emphasizes that whereas data are useful for understanding a system, they are only a part of the picture. He effectively demonstrates this point by posing the (rhetorical) question: "is it possible to understand, predict, or mitigate systemic risk or cascaded failures of the financial system entirely from observations of past data?"

Such systems thinking has led Dahleh and his colleagues to structure IDSS's academic programs around "The Triangle: systems, human and social interactions, and institutional policies." This framing echoes well Chayes' general insight that "data science properly sits at the nexus of STEM, and what I have begun to call the human-centered disciplines (HCD)—which includes social sciences, arts, and humanities, as well as the professional disciplines of public health, policy, social welfare, education, and law." After all, the data science ecosystem exists because of our common human desire and need to understand ourselves and the natural ecosystem that allows for our existence, and all the interactions within and in-between human and natural ecosystems. Doing data science without systemic thinking or treatment is no less fatal than treating a patient without an understanding of how diseases and medications may interact with each other, and how they can impact the patient's overall physical and mental well-being.

Data Science as a Virtual Infrastructure

Not surprisingly, this emphasis on ensuring that data science is built on systems theory and informed by its fundamental principles is also the central theme of the discussion by Joe Qin, Dean of (yet another) newly-established School of Data Science at the City University of Hong Kong. Like Dahleh, a renowned scholar in networked systems and robust control, Qin is a leading scholar in chemical systems and process control. Their respective backgrounds in systems engineering have clearly enabled them to bring a holistic system perspective into the task of developing and articulating data science framing and principles. Indeed, Qin writes, "While statistics and computing are two pillars of data science, I argue that system principles, which could be domain specific, should be another pillar of data science, since most data are generated from a system, be it natural or engineering systems." As much as I have tried to frame data science as an ecosystem, I find this 'trio-pillar' formulation to be rather refreshing and inspiring. The fact that I, being a senior (as in 'senior citizen') statistician, have little sense of what constitutes 'systems theory and principles,' is an example of the need for

explicating the role of systems thinking in data science research, education, and especially in our endeavors to increase data science's positive impact and reducing its negative consequences.

Qin's systems thinking provides another refreshing view concerning the essence of data science, which he considers to be provision of "*virtual instruments* to analyze data for scientific discoveries and engineering problem solving." This insight reminded me that, for the grand mission of sustaining a healthy human ecosystem, data science itself can be viewed as a *virtual* infrastructure. At the very least, data science helps to build virtual infrastructures for supporting and advancing research ranging from the physical sciences to the social sciences and, indeed, to physical infrastructures themselves as well.

The overview article by Matthew Schwartz on how "modern machine learning has been quietly revolutionizing particle physics" in the past five years provides a vivid testimony to this realization. It is also a striking example of how "data science adds a new pillar to the three existing ones of scientific research, i.e., theory, experiment, and computing, especially where first principles are not well established," as Qin reminds us in the aforementioned discussion. To many of us, physics is the scientific discipline in which one has the most luxury to invoke first principles. The community of particle physicists is particularly well known for its self-imposed high standards and its rigor with respect to achieving scientific reproducibility and replicability, as demonstrated by Thomas Junk and Louis Lyons in their article published under the rubric of the special theme of Reproducibility and Replicability in issue 2.4. The article by Junk and Lyons itself is scrutinized by a fellow particle physicist, Andrew Fowlie, in a Letter to Editor in this issue, which pointed out the need for more precise description of the discovery process in particle physics. Therefore, relying on mostly black-box or at least non-interpretable machine learning algorithms to discover physics laws was something virtually unthinkable a decade ago, and perhaps it is still unfathomable to many inside and outside physics. But the practice of such reliance is here now, whether we like or understand it. Schwartz contemplates the possibility that machine learning output may always outpace our ability to interpret, if we only try to "force the machines into our traditional worldview." He suggests that "we may humbly need to learn the machine's language, rather than ask the machines to speak ours." We can only imagine the infrastructural needs entailed in adopting and training on such a new language.

Even more nascent efforts in building research infrastructures via data science are taking place in social science, where first principles (at least those that have made into textbooks) are far scarcer than they are in physics. The most stimulating article by a multidisciplinary team of thought leaders—i.e., Tal Yarkoni (Psychology), Dean Eckles (Marketing), James Heathers (Health Science), Margaret Levenstein (Economics), Paul Smaldino (Cognitive and Information Sciences), and Julia Lane (Economics and Economic Statistics)—takes an in-depth look at relevant challenges and opportunities in "Enhancing and Accelerating Social Science Via Automation." It identifies five core infrastructural and methodological needs: (1) machine-readable standards, (2) data access platforms, (3) search and discoverability, (4) claim validation, and (5) insight generation. The last one is particularly thought-provoking. Is it at all possible to automate the process of

generating scientific insights? If such automation exists, even to a small degree, it could have very broad implications, e.g., pertaining to the acceleration of the process of gaining physical understanding in the context of Schwartz's article. The team is acutely aware of the "considerable skepticism" enticed by the mere suggestion of such an automation, even only as a possibility into the distant future. Nevertheless, they venture to suggest three possible directions of explorations. Not wanting to deprive readers of the joy of quenching your own curiosity, I will just say that my own skepticism was reduced by the three examples, but the degree of reduction depended on the pairing beverage at the time of my reading.

A second multidisciplinary team—consisting of Harvey Miller (Geography), Kelly Clifton (Civil and Environmental Engineering), Gulsah Akar (City and Regional Planning), Kristin Tufte (Computer Science), Sathya Gopalakrishnan (Agricultural, Environmental and Development Economics), John MacArthur (Transportation), Elena Irwin (Agricultural, Environmental and Development Economics), Rajiv Ramnath (Computer Science and Engineering), and Jonathan Stiles (Geography)—put data science squarely at the core of another enormous (physical and virtual) infrastructural building task: that of urban sustainability. Their article provides a broad overview and in-depth discussion of the concept of *Urban Sustainability Observatories* (USO) as primarily data-science enabled vehicles designed "to generate new scientific insights and design effective policies to meet sustainability goals for cities." USO goes beyond current urban data observatories and cyberinfrastructure, because it "treats the city as a complex system best understood one event or intervention at a time, and treats sustainability as a crucial but conflicted societal challenge that requires new forms of scientific, policy and community collaborations." It is another shining example of the systems thinking that we need to address simultaneously the global challenges of urban revolution, mobility revolution, and data revolution, as summarized in this truly thought-provoking article (so much so that it took my mind away from choosing a pairing beverage for reading it).

Two more articles in this issue touch upon a critical component of data science in general and, more particularly, in its capacity as virtual infrastructure: the sources of data. By examining the 'data exhaust' of daily production records kept by a Louisiana cotton plantation around 1860, historian Caitlin Rosenthal reminds us that "Scholars who have relied on slaveholders' data have tended to answer slaveholders' questions." The modern term *data exhaust* refers to "trails of information that individual leave behind as they move through the digital world," such as cookies, log files, temporary files, etc.—the digital equivalent of paper trails. Such trails, overlooked by analysts who only value the face/numerical value data, may contain critical information to reveal hidden or not so hidden biases in the data. Such biases can invalidate the entire study, in either a technical sense (such that it entails statistically biased estimates) or a substantive sense (such that it is technically correct but amounts to misinformation). Like the investigation of paper trails, the investigation of data exhaust takes time and an inquisitive mindset. Given our increasingly-rushed digital world, the diminution of our attention span, and the influence of misinformation upon our judgments, Rosenthal's call for "slow data, meaningful data" should be taken to heart by all who proudly call ourselves data scientists. This is the case even, and in fact especially, *in times of pan(dem)ic*, to borrow a term from the

title of philosopher Sabina Leonelli's discussion article in the last issue, which forcefully argued that "fast data science need not be rushed."

There are of course no more telling signatures of our rushed digital world than the various social media platforms that reflect and reinforce the habits of our hasty lifestyles, most notably Twitter (at least for my generation). Whereas each tweet is short, the cumulative tweet data are astronomically large (currently about 10,000 tweets per second, as of April 2021), and most critically complex. They also contain rich information, for example, that pertains to the COVID pandemic's impact on social discourse.

The process of properly extracting such information cannot be rushed, both in the literal sense because the information is in the longitudinal dynamics of the data, and in the sense of carefully developing new methodologies to meet the relevant challenges. This is exactly the approach taken by a team of data scientists from University of Michigan: Yu Wang (Statistics), Conrad Hougen (EECS), Brandon Oselio and Walter Dempsey (Biostatistics), and Alfred Hero (Statistics and EECS). Their article proposes a scalable framework for analyzing temporal and spatial dynamics by leveraging dimensionality reduction methods from computational geometry. By applying the framework to Twitter data subsampled from 2020, the article demonstrates both the potential of their framework and the usefulness of social media data—when analyzed properly—for enhancing and deepening our understanding of the interaction between human and natural ecosystems, and their impact on each other.

Building Data Science Connector Courses and Educational Infrastructures at All Levels

Back in 2015, I had the honor of serving on the visiting committee for the Department of Statistics at Berkeley. The committee was very pleased to hear about the department's aspiration (jointly with some other departments) to push for a school or college level entity of data science, which ultimately led to the current effort led by Chayes. I was particularly intrigued by the 'connector courses' for their now internationally-renowned Data 8: *Foundations of Data Science*, as featured in the second discussion article in this issue, by three members of the original team that conceived and taught this course: Ani Adhikari, John DeNero, and Michael Jordan. What intrigued me was not merely the idea of offering connector courses simultaneously with the core course, but the fact that the Berkeley team was actually able to pull it off. As we learn from this article, about dozen connector courses have been taught by a wide range of faculty members from science and engineer, social sciences, humanities, law, etc. After initially hearing about the connector courses, I was so excited that the moment I got back to my hotel, I wrote a long email to a good number of university leaders, as well as pedagogical innovators who I thought would be interested in experimenting something similar at Harvard. I urged them all to take a look at the Data 8 webpage.

Some readers might wonder what I was so excited about. What is the big fuss about having multiple courses from different departments—isn't that what a university is supposed to offer? But for those familiar with the

labyrinthine ways in which each department determines its own curriculum and the complexities of (typically insufficient) resource allocation, my excitement will come as no surprise. Merely to offer just one course across departments can take extensive arrangements and negotiations. That the Berkeley team was able to do so on the order of 10 is no small feat. I was not alone in my excitement. Indeed, several of my Harvard colleagues were also intrigued, especially after Adhikari's visit at the invitation of the late Robert Lue (*HDSR*'s founding co-editor for data science education) and myself in 2016. But five years later, we are still at the level of aspiration rather than accomplishment for exactly the reason that prompted my envy of Berkeley's success: we have not found the appropriate organizational structure and teaching infrastructures to support experimentation based on Berkeley's model. However, we did initiate a more traditional form of collaboration between computer science and statistics, offering a joint introduction level course on data science, for which we benefitted again from the Berkeley team's experience and advice, especially that of Adhikari.

More broadly speaking, the concept of 'teaching infrastructure' is not as well recognized as 'research infrastructure.' Departments often have a *research administrator* or *grant manager*, but one rarely hears about a *teaching administrator* or *course manager*. This is particularly unfortunate for data science education, which, given its wide-ranging nature—as vividly demonstrated in the case of Berkeley's success—will require designated and multidisciplinary-oriented teaching infrastructure to ensure continuity and sustain long term success. The emerging schools of data science and the like provides a real opportunity to establish such teaching infrastructures in cost effective ways. (I am well aware of the usual arguments regarding the financial incentives of having designated *grant managers* verses *course managers*.) The wide-ranging discussions of the article by Adhikari, DeNero, and Jordan—from the delicious "Recipes for Connector Courses from the Early-Career Board Kitchen" to the visionary "iNZights" for democratizing data science by a New Zealand team of educators—demonstrate clearly that we have far more innovative ideas of data science education than our current teaching infrastructures and resources can support and sustain at scale.

Indeed, this is even more vital at the pre-college level, as the column on "Engaging Young Learners With Data: Highlights From GAISE II, Level A" by Leticia Perez, Denise Spangler, and Christine Franklin reminds us. This is the second piece in the series of articles that overview and delineate the recommendations from the *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*, released recently by the American Statistical Association and the National Council of Teachers of Mathematics (of the United States). Pre-college education in statistics and data science suffer from a severe shortage of experienced teachers and teaching resources to implement such recommendations at scale. I certainly hope that in our current global effort to build data science infrastructures and programs on university campus, we do not develop short-sightedness and, further, short-change ourselves by failing to invest in helping pre-college educational communities to enhance their teaching training, resources and infrastructures. Even from a purely economic perspective, it is not a wise strategy to overload the top floors when the foundation is shaky.

Computational Humor is Not a Joke, but Would You Laugh If It Were Funny?

To balance out on that weighty thought, let me conclude yet another lengthy editorial with a lighter touch—or maybe not. If you have gotten a good laugh by reading my opening story about support vector machines, then you have just made yourself an example of the incongruity-resolution (IR) theory. If you have no idea what IR is, then it is a good reason to read "Computers Learning Humor Is No Joke" by Thomas Winters, even if you don't like jokes (who *does*?). For those who oversee funding research, if you ever wonder why tax money—or any money—should be spent on making jokes, well, this article is for you, too. I promised myself that this editorial should not exceed 4500 words (and that it would be done by 8 a.m.), so let me end it rudely by complaining that the artificial (intelligent?) jokes in Winters' article are not that funny. But for those who have feared that AI is taking over humanity, that should make you smile. You still have the last laugh and will stay smiling for the foreseeable future.

Disclosure Statement

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

Preview image by Jorge Guillen from Pixabay.