Harvard Data Science Review • Issue 4.1, Winter 2022

Kaleidoscopic Data Science

David C. Parkes^{1,2}

¹Paulson School of Engineering and Applied Sciences, Harvard University, Allston, Massachusetts, United States of America,

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

Published on: Jan 27, 2022

DOI: <https://doi.org/10.1162/99608f92.dbd5f147>

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

I have been delighted to work with my colleague and co-conspirator, Francesca Dominici, these past few months in taking forward the scholarly project that is *HDSR*. Our tagline is “A Telescopic, Microscopic, and Kaleidoscopic View of Data Science,” and I can’t think of a better way to illustrate the multifaceted world of data science than by reading through this current issue.

I would like to thank [Rita Ko](#), [Marie McAuliffe](#), and [Liberty Vittert](#) for their work in editing a special theme on migration, [with a series of articles](#) that reflect on the role of data science in helping to understand migration, helping those who are displaced, and combating disinformation. This special theme came about as a result of discussions that took place at the “[World Migration and Displacement Symposium: Data, Disinformation and Human Mobility](#),” a virtual event held in the spring of 2021 and organized in part by *HDSR*.

As befits our continuing struggles with the pandemic, this issue also features two articles on COVID-19. [The first](#) studies social distancing which, along with wearing masks, is one of the most effective nonpharmaceutical interventions in limiting COVID’s spread. The article reports an analysis of the results of a survey of 2,500 U.S. respondents in late March 2020, coming a couple of weeks after the CDC had announced strict guidelines on gatherings and social distancing and as the world was shutting down. The authors seek to identify causal relationships between factors such as financial security, information-seeking tendency, and social distancing behavior, and suggest how machine learning can be used to prioritize messages to particular groups. As they write, factors such as political affiliation and race, which were significant *predictors* of social distancing, were not found to be significant, direct *causal* drivers of social distancing. Although these results come with the usual caveats around causal inference based on observational data, I am impressed with the care that the research team has taken to validate their methodology on synthetic data sets on which they can access the ground truth. The data and R script that underpin this analysis are available on [Harvard Dataverse](#).

The [second COVID-19-related article](#) makes use of an agent-based simulation to assess the impact of different test and isolation schemes on the containment of COVID-19 outbreaks in a primary school setting in England. How to control infections during the pandemic while giving students as much in-person contact with teachers as possible has been a signal challenge of this disease. Just yesterday my sister shared the current reality on this in England, with a third of the kids in my nephew’s class out of school because of COVID. The simulation results in the article provide numerical evidence on the benefits of regular (e.g., once- or twice-weekly), whole-school cross-sectional testing. It’s fascinating to see this use of agent-based methods providing a computational tool to help inform policy. The authors have also developed a [Shiny app](#), designed to help public health analysts tailor simulations to their own scenarios.

We also have two Stepping Stones articles. [The first](#) reports on a wonderful development in machine learning education that has been led by my colleagues in Harvard Engineering, together with their collaborators at Google, edX, and Arduino. The authors describe their work in developing a massive open online course on

“Tiny Machine Learning” (TinyML). TinyML refers to the use of machine learning (ML) for low-power data analytics on small, battery-operated devices. The course is offered on [edX](#), and along with Google Colab notebooks, these educators worked with [Arduino](#), an open-source hardware and software company, to develop a low-cost, globally accessible TinyML kit. Ethical considerations in the design and deployment of ML-driven applications are covered through a tie-in with [Harvard’s Embedded EthiCS program](#). Embedded EthiCS had earlier been founded in a collaboration between computer science and philosophy at Harvard and with the leadership of Barbara Grosz and Alison Simmons. This TinyML course is responsive to challenges in expanding access to applied ML education, being online, requiring little data and affording simple model training, and being deployed on low-cost hardware. As of December 2021, over 60,000 students have enrolled—these students coming from more than 170 countries around the world. The typical path is for students to audit the course for free, and students can also ‘upgrade’ and pay to earn a professional certificate. This is an inspiring story, with this course empowering students from diverse backgrounds to develop complete, self-contained ML-driven applications.

The [second Stepping Stone article](#) reports on a fascinating experiment in which educators at four European universities (KTH Stockholm, KU Leuven, UPC Barcelona, and Grenoble INP) collaborated to develop a new data science curriculum for students who are pursuing studies in energy engineering. Data science is a key enabler for technologies that help decarbonize global energy use, yet the energy sector is struggling to attract and train enough data scientists. As we adopt more renewable energy sources, for example, we need to move from ‘demand following’ to ‘supply following,’ in handling the uncertainty in supply, with a role for data science and artificial intelligence in the prediction and control of consumer demand. The authors describe their experience in designing and delivering new data science content in this cross-university context. Over time, the instructors have introduced Kaggle-based group activities, open-ended assignments, and more theory. As with the TinyML course, Python programming and Jupyter Notebooks form an important part of the student experience. In overcoming institutional inertia, GitHub repositories and Slack have helped to bridge divides and accelerate the pace of collaboration.

You will have surely heard about quantum computing but, like me, may have some confusion about what is happening and where the research frontiers lie. Yazhen Wang’s article “[When Quantum Computation Meets Data Science: Making Data Science Quantum](#)” is a tour de force in scientific exposition. We learn about the superposition of states in a quantum system, with the bit in a quantum state (a ‘qubit’) having a simultaneous occurrence as a one and zero (illustrated through [Schrödinger’s cat](#), a hypothetical cat that may be simultaneously alive and dead). We learn about the exponential power of a qubit system, we visit quantum gates and quantum circuits, and the concepts of ‘quantum entanglement’ and ‘quantum teleportation.’ Wang builds through quantum algorithms such as Shor’s factoring algorithm toward today’s work in building quantum computers and demonstrating ‘quantum supremacy,’ which refers to outperforming a classical computer on a tough computational task. This is where Wang advocates for a *quantum data science*, by which he refers to the need for data science to address the inherent stochasticity of quantum physics that makes

quantum computation random. As he writes, “quantum supremacy claims should heavily rely on sound statistical analysis and justification, including quantum and classical computing experiment design, data collection and analysis, assumption validation, and model assessment.” The article concludes with a “big data poem” and I regret that my colleague Xiao-Li Meng is not here to share his appreciation of its accompanying Chinese version. Sometimes we have the chance to include short discussion pieces to accompany an article, and I am very grateful to [Anurag Anshu](#), [Maria Kieferova](#), [Susanne Yelin](#), [Xun Gao](#), [Richard Gill](#), [Yaakov Weinstein](#), and [Nana Liu](#) for their willingness to expand on this discussion of the frontiers of quantum data science. I hope you find this content as interesting and provocative as I have.

Do also take a look at this issue’s column articles. In [Effective Policy Learning](#), we read about a collaboration between states in the Midwest in building a shared data infrastructure. In [Recreations in Randomness](#), we dive into the (nonquantum!) entanglement of data science and online dating platforms. [Liesel Sharabi](#) provides a definitive study of computerized dating, tying this all the way back to “[Operation Match](#)” in 1965 at Harvard University. Today, yours truly is faculty advisor to Harvard’s [Datamatch](#), which started in 1994 and has since expanded to more than 30 colleges, with over 40,000 students participating last year. I’m not quite sure what this says about Harvard, but [OkCupid](#), widely recognized as having brought ‘scientific matching’ to online dating through its lengthy questionnaires, is itself a product of four Harvard mathematicians. I know you will be reading *HDSR* on Valentine’s Day, and as you do, you can learn all about this and more in Sharabi’s fascinating article. Did you know that Tinder uses the Elo rating system, first developed for chess, to infer the desirability of users? Did you know that Hinge, an app where U.S. Secretary of Transportation Pete Buttigieg met his husband, Chasten Glezman, uses Gale-Shapley’s stable matching, which was first described in a 1962 [American Mathematical Monthly article](#)? Sharabi also gets into the important topic of the ethics of algorithmic matching, considering the safety and equity of online dating apps.

Lastly, it has been a great privilege to work with [Radu Craiu](#), Professor and Chair of Statistical Sciences at the University of Toronto, to curate a series of short articles for [Bits and Bites](#). Beyond helping with *HDSR*, another hat that I wear is that of co-faculty chair, along with [Karim Lakhani](#), of the [Harvard Business Analytics Program](#). Back in November 2020, we launched an essay writing competition on the topic of “The Future of Work, Business Leadership and Managerial Decision Making in an AI-Driven World,” inviting submissions from our community. As we write [here](#), these essays examined a diverse and important range of issues, providing viewpoints whose insights lay the groundwork for how businesses should think about AI in the days to come. Radu and I have selected six of these essays ([Tom Casey](#), [Ben Dooley](#), [Nini Hu](#), [Holger Pietzsch](#), [Raidford Smith](#), and [Maryse Williams](#)) for this issue of *HDSR*. In the spirit of data science everywhere, behind the scenes we were also running a randomized experiment into contest design! This experiment was led by [Shreyas Sekar](#), then a postdoctoral fellow at the [Laboratory for Innovation Science](#) at Harvard. Innovation contests have been employed by a variety of organizations (e.g., NASA, Starbucks, Walmart). A fundamental challenge is to get solutions that are both high quality and diverse, and research has shown that providing participants with more information in the form of sample solutions can lead to higher-quality but less-diverse

solutions. In this competition, which was conducted under Harvard’s Institutional Review Board, we randomly provided some participants with sample essays (“information”), some with summaries of sample essays (“partial information”), and some with no prompts (“no information”). As expected, full information was found to improve quality but reduce diversity relative to no information. Surprisingly, partial information was found to adversely affect *both* quality and diversity relative to no information. In this case, it seems a little knowledge can lead to poor results compared to either full or no knowledge. I think I can close here, with this call for full knowledge connecting us nicely back to [Yazhen Wang’s big data poem!](#)

Disclosure Statement

David C. Parkes has no financial or non-financial disclosures to share for this editorial.

©2022 David C. Parkes. 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.