Harvard Data Science Review • Issue 3.2, Spring 2021

Data Science in Domains: Interaction of Physical, Social, and Institutional Systems

Munther A. Dahleh^{1,2}

¹Institute for Data Systems and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America,

²Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Published on: Jun 30, 2021

DOI: <https://doi.org/10.1162/99608f92.3e19b17c>

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

I enjoyed reading Jennifer Chayes's article ("[Data Science and Computing at UC Berkeley](#)," this issue) discussing the mission of Computing, Data Science, and Society (CDSS) at UC Berkeley. I see many parallels between this effort and MIT's initiatives in computing and data science. In 2015, MIT launched the Institute for Data, Systems, and Society (IDSS) to integrate MIT's activities in data-driven research with critical societal challenges. And in 2019, the Institute launched the Schwarzman College of Computing (SCC) to integrate computing in education and research across all of MIT.

As we contemplate the preparation of the future data science learner, I want to highlight a core principle of IDSS and how it affected our decision to set up two of our flagship graduate education programs. In particular, I want to draw attention to the systems piece of IDSS—a dimension that seems to be unique to us and one that can significantly affect the overall success of data science impact.

As machine learning became a standard and accessible tool, a concomitant danger emerged from its usage without attention to particular domain knowledge. This concern is deeper than just understanding the context of the observed data. For example, is it possible to understand, predict, or mitigate systemic risk or cascaded failures of the financial system entirely from observations of past data? This is just one example of a rare event that is heavily dependent on the financial protocols, underlying government policies, and various incentives put into place to motivate people's behaviors. The latter is what I would term the 'system.' The financial system comprises the interactions among financial protocols, social behaviors, and institutional policies. Data may be available on some parts of such a system (and different data are available to different decision makers), and it is difficult to predict rare events without understanding this interaction.

Another example of great societal impact is combating systemic racism. Effective research certainly will explore and collect relevant data across domains where racism has been prominent. The research, however, may lead to incorrect conclusions and self-fulfilling cycles of inequity if it does not take into consideration the ways such data are collected, and if it does not identify explicit or implicit ways that racial bias may creep into the data collection process. A proposed policy or intervention must have a causal effect on the desired outcome. In the context of scientific development, for instance, it would seem a daunting exercise to develop a vaccine for hepatitis C solely by experimentation and without a deeper dive into functions and mechanisms inside the liver. In smart infrastructures such as transportation or energy, the physical and economic systems aspects are quite prominent in understanding the data collected from these large and distributed applications.

The model my IDSS colleagues began to formulate at MIT in 2015 grew out of the convergence of three rising pools of data: data from scientific, economic, and engineering processes (common in physical systems and well documented in academia and industry); data from institutional sources (global in scope, aggregated from the outcomes of various organizational interventions, and reliant on existing mechanisms); and, finally, data from social interactions (individual in scale but newly available in mass quantities). Most societal challenges that interest us, in domains such as finance, energy, urbanization, social networks, and personal and public health,

typically hinge on the interactions among these three heterogeneous components. Without explicit understanding of the interactions between these components, data alone may result in vastly erroneous conclusions. Causality is certainly one dimension that suffers in the absence of such systemic dependencies.

This way of thinking has led us to structure our academic program around what we refer to as The Triangle: systems, human and social interactions, and institutional policies. In the context of a particular domain, we observe aspects of these three pieces through heterogeneous data sets. Such data sets can be noisy—different resolution levels, different time scales, sparse, high dimensional, or aggregate. We need new statistical theories with which to assess causality, prediction, or systemic effects. A certain level of abstraction is necessary to get any quantifiable or testable outcomes. Such an abstraction takes us back to modeling, which requires domain expertise at the heart of the analysis.

The IDSS Social and Engineering Systems (SES) Ph.D. program (incidentally based on data but not explicitly stated) brings these pieces together through three foundational pillars of education: information sciences (probabilistic modeling, information theory, optimization and decision theory), statistics, and humanities and social sciences. Students acquire in-depth expertise in all three dimensions as they study particular domains, enabling them to address questions related to assessment, design, and ethics. While clearly quite ambitious, one can imagine how powerful each learner will be once they join in with the rest of the workforce to solve complex societal challenges.

Finally, I want to reemphasize a point mentioned in Chayes's article, namely, the importance of quantifying the uncertainty of results using data science. Indeed, developing algorithmic solutions is key to implementation, but the assessment of the reliability of the results can prevent us from making grave and regrettable mistakes once we implement recommendations from such approaches. In response, IDSS created the Interdisciplinary Ph.D. Program in Statistics (IDPS) to partner with other Ph.D. programs in domains that deem data science to be critical to their research. IDPS provides education in advanced and contemporary statistics topics involving high dimensions, networks, complexity, online platforms, causality, and experimental design. It is anchored on four pillars: probabilistic modeling, statistics, computation (e.g., machine learning and various algorithmic solutions in artificial intelligence), and application of data science (including social and ethical issues). Through interdisciplinary collaborations, students from various domains can use and develop rigorous statistical tools to address their challenging research problems. Almost all IDSS's SES students enroll in IDPS. Ten other Ph.D. programs at MIT have already partnered with IDPS.

I contend that data science is most effective when it is fully integrated with domains. The expert in biology, urban planning, supply chain, chemical processes, pollution, or gender bias, for example, must be equipped with a rigorous understanding of data science and computational thinking to enable a systematic approach to research, and vice versa: The statistician and data scientist must be deeply grounded in domain knowledge in order to have a meaningful positive societal impact. This integration is central to the mission of IDSS.

Disclosure Statement

Munther A. Dahleh has no financial or non-financial disclosures to share for this article.

©2021 Munther A. Dahleh. This article 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 article.