Harvard Data Science Review • Issue 3.2, Spring 2021

Data Science Education With Domain Knowledge and System Principles

S. Joe Qin¹

¹School of Data Science and Hong Kong Institute for Data Science, Centre for Systems Informatics Engineering, City University of Hong Kong, Kowloon, Hong Kong

Published on: Jun 30, 2021

DOI: <https://doi.org/10.1162/99608f92.6f037267>

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

The world's attention to data science, big data, and machine learning has been triggered by successful applications in recommender systems, business analytics, natural language processing, computer vision, image processing, autonomous systems and processes, social media and networks, and so on. A revolutionary impact on science discoveries is that data science adds a new pillar to the three existing ones of scientific research, that is, theory, experiment, and computing, especially where first principles are not well established. For example, it is not straightforward to design new material or molecules with desired functional properties from chemistry principles. Massive experimental data could be analyzed to reveal the mapping from molecular structures to properties. Another example is to reveal causal relations among research findings to accelerate scientific discoveries (Gates et al., 2019). On the other hand, industries are seeing the next revolution (i.e., Industry 4.0) that will unleash values in massive data from real-time operations, production processes, services, and municipal operations (Qin & Chiang, 2019). Data literacy for the next generation workforce is a necessity, which is best achieved via education.

The institutional approach adopted at City University of Hong Kong (CityU) was to establish the first standalone School of Data Science in Asia in mid-2018. Concurrently, the Hong Kong Institute for Data Science (HKIDS) was established to oversee transdisciplinary research activities and to reach out for societal impact. The curricular approach is to adopt the framework of data science plus a domain at the undergraduate level. In the meantime, general education and minor options are available for all college students. Two bachelor's programs were created, with one focusing on data science and another one on data and systems engineering. The master's-level curriculum follows the same philosophy, where the domain knowledge integration is highlighted by a term project. The Ph.D.-level education is similar to those of many institutions, with data science training via core courses and domain specialization led by their advisors.

The article "[Data Science and Computing at UC Berkeley](#)," by Professor Jennifer Chayes (this issue) addresses many critical aspects in data science education and research that are encountered by institutions globally. I would say that every point addressed in the article is worth considering when establishing or revising a data science curriculum. The article is also concise in illustrating the critical aspects, so there is no need to summarize here.

In this discussion, I would like to offer some complementary views to the already rich set of answers in the article. It is convenient to point out that both Berkeley's Division of Data Science and Information and CityU's School of Data Science were announced in 2018, with inaugural deans Professor Chayes and myself taking offices in January 2020 at the respective institutions. Both Berkeley and CityU are located next to the innovation hubs at the San Francisco Bay Area and the Greater Bay Area of Shenzhen, respectively. Both bay areas have witnessed tremendous growth of large companies in big data, e-commerce, and artificial intelligence (AI). The demands for well-trained next-generation talents in data science are strong in both areas.

What Is Data Science?

Simply put, data science provides *virtual instruments* to analyze data for scientific discoveries and engineering problem solving. As data get massive, heterogeneous, and high-dimensional, simple curve fitting will not get the job done. Data science provides scientific ways to analyze, visualize, interpret, predict, infer, and even make decisions or take actions on the system under study. These purposes are well elucidated in Chayes's article within various domains. Instruments in a narrow sense can be Galileo's telescope, which helped him see what others could not. In automation and control, instruments include actuators and controllers that manipulate the system. Reinforcement learning, for example, works as a controller or decision maker to actively learn from data to optimize an objective.

I fully agree with Chayes's view in that data science innovations should be human centered. Along with this view, they should be instrumental for human beings to achieve their goals with meaningful and ethical purposes. A great utility of data science is to assist the human to acquire knowledge from massive data. Therefore, human beings must be in the loop to interpret, understand, and acquire the knowledge. Interpretability of data analytic outcomes is a necessity. Another utility of machine learning and data science is to develop autonomous systems, but one should resist the temptation to bypass the human prematurely to avoid mishaps like the Boeing 737 Max airplanes with the MCAS software (*cf.* https://en.wikipedia.org/wiki/Maneuvering_Characteristics_Augmentation_System).

The 'Age of Interaction' points to an opportunity to enhance the data science discipline with domain knowledge. In control and cyber-physical systems, it is obvious that one needs to understand the external system in order to predict, change, and manipulate it. 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. Wiener's *Cybernetics: Or Control and Communication in the Animal and the Machine* (1948) was for this purpose, although control, decision-making, and interactions represent different levels of interventions to the system.

Data Science vs. First Principles

One frequent question in applying data science to various domains is whether data science and machine learning would replace existing engineering and science principles. The answer is no. Rather, data-driven machine learning complements first principles where they are lacking or unknown. Take airplane flying as an example, which is shown in the left panel of Figure 1. The airplane (i.e., system) is designed with engineering and science principles and made with every part known. However, when flying the airplane, the ambience it flies through is unknown and time-varying. It is through real-time sensor data and the pilot's judgment that the ambient condition is learned. The coexistence of knowns and unknowns in real-world systems is always there, which is illustrated with the Yin-and-Yang symbol in the right panel of Figure 1. The dark side represents the unknowns and the white side the knowns. Data science helps interpret data and assess the ambient condition in

a timely manner to assist the human operator, although the airplane system was designed with known principles. It would be unwise to forget the principles and rely purely on data, or vice versa. Data scientists should be trained with basic engineering and science principles so that they are more effective and not to reinvent the wheel.

Figure 1. Flying an airplane through a turbulent ambience (left panel) and the coexistence of knowns and unknowns in a system via the Yin-Yang symbol (right panel).

Note that the airplane system can change over time with drifting, aging, and even malfunctioning. The notion of the time variedness of machines and systems is well illustrated in Jordan (2019). There will be constant transfers at the boundary between the knowns and the unknowns. Machine intelligence from data analytics will make airplane flying more autonomous, although a proper balance and a fallback option are always necessary.

A Data Science Curricular Approach: DS + X

How should data science curricula reflect the need for domain knowledge and first principles? Technological advances in data science and AI are transforming traditional industries, manufacturing, and business operations. The immersion of data science education into science and engineering curricula is just at its beginning and fast developing from graduate to undergraduate programs. New schools, departments, minors, concentrations, and modules of data science have been established worldwide in the last few years.

However, due to the diverse disciplines of engineering and science and the diverse analytical tools in data science, a critical challenge is how to create a curriculum that will equip the student with the highly needed data science skills and a domain specialization. This approach of data science plus a domain, that is, DS + X, where X = a domain, was adopted at CityU's School of Data Science. Our core courses such as *Systems Dynamics and Control*, *Systems Modelling and Simulation*, and *Quality Technologies*, are nontraditional statistics or computing courses. Detailed curricular information at CityU can be found at <https://www.sdsc.cityu.edu.hk/programmes>.

The composite education approach of 'DS plus a domain' is like the two sides of a coin. Two undergraduate degree programs were established at CityU, with one focusing on data science and another one on data and

systems engineering. The curriculum design embraces many engineering and science domains, including energy, environment, smart manufacturing, smart cities, social media, FinTech, and health analytics. Conversely, students from other schools can take data science courses or a minor in data science. The master's-level curriculum follows the same philosophy, with the domain education highlighted by a semester-long project. The Ph.D.-level education is analogous to that of many global institutions, with data science training via core courses and domain specialization by the advisors. It is anticipated that data scientists with a domain specialization will be in high demand.

The last question is also the first one: What exactly is data science? Is it a discipline? Data science being an applied science is convincingly illustrated in Jordan (2019). At this stage it is probably healthy to keep an open mind and state what it is not, as was done in Meng (2019). The CityU's composite talent education with DS + X encourages cross fertilization and in-depth collaboration. The institutional approach at CityU's School of Data Science is not to build internal divisions in the foreseeable future. Faculty colleagues understand that heterogeneity and gaps between theorists and domain specialists are healthy at the current stage of the discipline. Fundamental theory and rigorous development will be an eternal goal, but historically, domain applications often precede theory development. The steam engine was put into wide use after James Watt created the governor for the steam engine in 1776, which brought the first Industrial Revolution. If one had to wait for the rigorous theory of J.C. Maxwell's famous paper "On Governors" in 1868, the Industrial Revolution would have been delayed by nearly a century.

Acknowledgments

I would like to thank Profs. Zijun Zhang and Xiang Zhou, Program Leaders at CityU's School of Data Science for their input and discussions.

Disclosure Statement

S. Joe Qin has no financial or non-financial disclosures to share for this article.

References

- Chayes, J. (2021). Data science and computing at UC Berkeley. *Harvard Data Science Review*, 3(2). <https://doi.org/10.1162/99608f92.12c8533a>
- Gates, A. J., Ke, Q., Varol, O., & Barabási, A. L. (2019). Nature's reach: Narrow work has broad impact. *Nature*, 575(7781), 32–34. <https://doi.org/10.1038/d41586-019-03308-7>
- Jordan, M. I. (2019). Artificial intelligence—The revolution hasn't happened yet. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.f06c6e61>

Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1).
<https://doi.org/10.1162/99608f92.ba20f892>

Qin, S. J., & Chiang, L. H. (2019). Advances and opportunities in machine learning for process data analytics. *Computers and Chemical Engineering*, 126, 465–473. <https://doi.org/10.1016/j.compchemeng.2019.04.003>

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. Hermann & Cie & MIT Press.

©2021 S. Joe Qin. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.