Harvard Data Science Review • Issue 2.3, Summer 2020

Training and Funding Pipelines for Data Science: the Need for a Common Core of Transdisciplinary Principles

Shivani Agarwal

Published on: Oct 07, 2020

In their articles in this HDSR issue, Jeannette Wing, Xuming He, and Xihong Lin have eloquently laid out several research challenges for data science in the years to come. In order to support the development of a conducive research environment for effectively tackling such challenges, we need to ensure the development of two fundamental ingredients:

1. An effective training pipeline for transdisciplinary training of young researchers in data science; and
2. An effective funding pipeline to support such transdisciplinary training and research.

In what follows, I discuss some important considerations for developing these ingredients, focusing in particular on the need to develop both a strong core of *principles* for data science, and a broad arsenal of *tools* for applying data science methods in diverse domains (much as has been done previously in mathematics, statistics, and computer science, except that the emerging field of data science brings some unexpected and unique considerations as outlined below).

1. Data Science Education: Need for Integrative Education With a Common Core of Transdisciplinary Principles

Data science builds on multiple disciplines including computer science, mathematics, and statistics. Historically, in each of these individual disciplines, there has been a progression of ideas from a core of fundamental *principles* to a broad range of *tools* that can then be applied to solve problems in a variety of domains. In each case, advances in both the principles and the tools—and crucially, in their interaction—are important for the overall progression of the discipline, and training programs in each of these fields are designed to reflect this.

For example, in mathematics, there is a gradual progression of ideas from a core of 'pure' mathematics to 'applied' mathematics, the latter then being used to solve a variety of problems in physics, engineering, biology, finance, and a whole host of other domains. Training (and research) programs in mathematics usually include elements of both, although some focus exclusively on the 'pure' or 'applied' side, depending on needs and interests.

Similarly, in statistics, there is a gradual progression of ideas from a core of 'theoretical' statistics to 'applied' statistics. Again, training (and research) programs in statistics usually include elements of both, although some focus exclusively on the 'theoretical' or 'applied' side.

In computer science too, there is a core of computer science 'theory,' which facilitates implementation of computer science 'systems,' which are then deployed to solve a variety of problems in various

domains. Again, training (and research) programs in computer science usually include elements of both, although some focus exclusively on the 'theory' or 'systems' side.

For data science to evolve into a truly mature discipline, we need to similarly identify and develop both a core of data science 'principles,' and a broad arsenal of tools that can be used in the 'practice' of data science; both of these should then be incorporated in educational and research programs. Two challenges in doing so—which also distinguish the development of data science from that of the three parent disciplines above—are the following:

- **Practice has preceded principles.** In each of the three disciplines above—mathematics, statistics, and computer science—the development of theoretical principles has often preceded that of applied tools (this is of course an oversimplification—in reality, the development of each informs the other—but overall, the trend has been largely that of principles being developed before applications). In data science, on the other hand, there has been a tremendous push from application domains that need various kinds of data analyses to be performed, and this has resulted in the rapid development and dissemination of a variety of applied tools, sometimes applied in an ad hoc manner, before a common core of agreed-upon principles has been developed. While this has helped to jumpstart the field in the short term, in the longer term, it is crucial to develop a strong foundation of fundamental principles for data science that training programs for future data scientists should be expected to include, and that can guide the further development of the field.
- **Transdisciplinarity.** An additional challenge—which has perhaps also contributed to the delay in developing a common set of core principles—is that the core principles of data science must be inherently transdisciplinary. In particular, the core principles of data science will need to draw on—and suitably integrate—a variety of principles from computer science, statistics, mathematics, and electrical engineering (from the latter, notably principles from information theory).

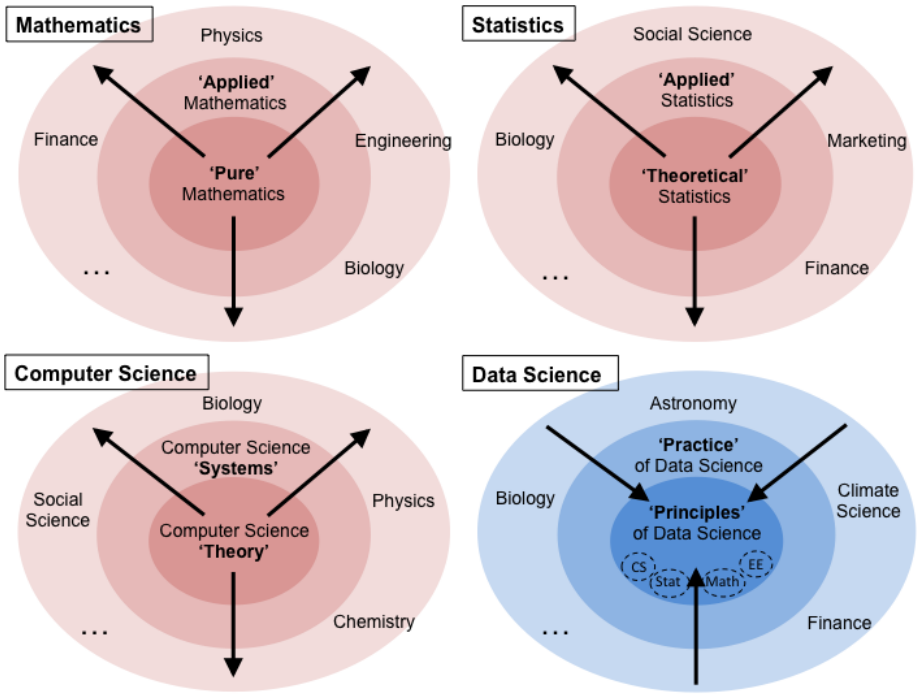

Figure 1. In each of mathematics, statistics, and computer science, there has been a progression of ideas from a core of fundamental *principles* to a suite of applied *tools*. In data science too, there is a need to develop a common core of fundamental principles. Unlike its parent disciplines, however, data science comes with two additional considerations: (a) In many cases, the practice of data science has begun before the development of common principles, and so the principles of data science must factor in observations from current practice; (b) The principles of data science must be inherently transdisciplinary, incorporating elements from mathematics, statistics, computer science, and electrical engineering.

2. Data Science Funding: Clarion Call to Long-Term Thinkers!

As with the development of any field, funding and institutional commitments—from universities, government agencies, industry, nonprofit foundations, and private donors—will play an important role in shaping the future development of data science. While applied tools are easy to appreciate and to provide support for—and these should certainly continue to be supported—for the long-term health of the field, it is imperative that we also remember to build a strong foundation for it by supporting the development of a core set of transdisciplinary principles. In recent years, the National Science Foundation (NSF) has recognized this need and provided some support for it through its Transdisciplinary Research in Principles of Data Science (TRIPODS) program, but many more such efforts are needed.

This will not be a short-term exercise; given the challenges involved, it will require universities and other stakeholders to plan for 10-years-plus commitments. But the rewards associated with such commitments will be huge: in addition to helping to put the field of data science on a strong footing and directly advancing its long-term promise to society, those who make such commitments early on will be well-poised to emerge as leaders in the field in the future.

In summary, the development of a strong core of transdisciplinary principles for data science is a long-term exercise that needs much thought, care, and support; nevertheless, it is crucial for the maturation of the field, for the effective training of future generations of data scientists, and for making progress on a variety of important research challenges in the field, including those highlighted in this issue of HDSR.

Acknowledgements

Thanks to David Parkes for helpful comments on a draft of this discussion.
