Harvard Data Science Review • Issue 3.2, Spring 2021

Leadership in Data Science: Lessons Learned From Time Invested in Helping to Build the Field

Patrick J. Wolfe¹

¹Department of Statistics and Computer Science, College of Science, Purdue University, West Lafayette, Indiana, United States of America

Published on: Jun 30, 2021

DOI: <https://doi.org/10.1162/99608f92.5bf277ba>

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

Congratulations are in order both to UC Berkeley and to my good friend Jennifer Chayes in thoughtfully forging stunning intellectual raw materials into a world-class data science effort (“[Data Science and Computing at UC Berkeley](#),” this issue). Bearing in mind that of course free advice is worth precisely what you pay for it, I shall nevertheless attempt by this discussion to offer up a few time-tested observations on leadership in data science, which might hopefully also prove useful to the broader community. These reflections are gleaned from many years of efforts—some objectively successful and others less so—invested in everything from subject-matter teaching and multi-investigator, multidisciplinary grants, through to academic centers, professional society initiatives, and national institutes. (And yes, in case, dear reader, that you are curious, I certainly do have the bruises to prove it!)

A few words about my background: My current ‘day jobs’ include looking after the mathematical, computing, and natural sciences at Purdue University as dean, helping the new National Science Foundation-funded Institute for Mathematical and Statistical Innovation as its board chair, and serving as a trustee and non-executive director of the national Alan Turing Institute in London. Notwithstanding these affiliations, this article is written in a purely personal capacity, and all opinions, along with any omissions or errors, are entirely my own.

The central leadership challenge of data science, as I would describe it, lies in creating a whole that is greater than simply the sum of its parts. Before you dismiss this as a truism, hear me out. Academic departments and institutions have amazingly long time constants. By design and necessity—not to mention a direct tradition dating back well over a thousand years—we self-organize within our universities to codify and impart the knowledge we judge necessary to constitute a degree in a given discipline. But the broad data and computational sciences—meaning, as Chayes writes, the technical core of data science plus a wide spectrum of disciplines across which it interacts—have arisen very quickly as a function of unforeseen advances: the rapidly lowering costs of high-volume data collection, storage, and transmission technologies; the commoditization of computing and its appearance as an on-demand utility resource; and a globally connected community of researchers constantly pushing the edge of the technical envelope to uncover new ways to make sense of such heretofore-unseen volumes of newly available raw data through rigorous and defensible means.

As experts, we know that many of the fundamental algorithms and principles that drive recent technical advances in data science date back decades, if not longer—some simply needed to be brought to bear on vastly greater quantities of sufficiently high-quality data in order to demonstrate their empirical successes. At the same time, with disciplines such as statistics coming of age in an era when data sets were typically small, costly, and difficult to obtain, the founding intellectual questions that researchers sought to ask and answer were fundamentally different from those we study today. As I said, every one of us working in the field knows this history well—but it is worth bearing this carefully in mind when discussing the future of data science with colleagues and constituents who don’t spend their lives working with Greek symbols!

The factors cited in the last two paragraphs have combined to cause data science to emerge through a rapid, messy process as an incredibly powerful and challenging area that stands to alter our economies, societies, and lives in unexpected and fundamental ways. In this sense, it is as far from simply necessitating the founding of a new academic department as anything could be—and that is why good leadership is so immensely critical to realizing the full potential of advances that data science may hold for us now and in future.

Thence a few observations for those considering working on behalf of our global community of researchers and scholars to help lead in this arena, which I’ve tried to distill into three guiding principles for your consideration:

1. Represent and uphold rigorous scholarship; be relentlessly skeptical of facile conclusions.

The future of any area of intellectual inquiry is only ever as rich as its underlying scholarship. Core to the future of data science is a healthy technical ecosystem that not only spans from the very theoretical to the concrete and applied, but more crucially engages a fully functioning feedback loop whereby the ‘pull’ of practical problems motivates new theory and methods, which in turn connect back to a ‘push’ from theory that suggests new avenues of application in the world at large. We are incredibly fortunate that so many open data science challenges lie so close to problems of great import and interest in today’s world—this, and the opportunity to help shape tomorrow’s world, is not an opportunity that every researcher or scholar is lucky enough to have. We ought well to remember to cherish this opportunity and to approach it with humility and respect.

When evaluating sweeping data science conclusions or new techniques, bear in mind the old adages that, first, if it seems too good to be true, it most likely is; and second, that there is very little truly new under the sun. This latter point in particular is cause for celebration; we do much more to advance the field by correctly identifying a reinvented technique without malice than by adding noise to the field overall in the rush of accompanying self-promotion that can quite naturally accompany the trumpeting of a hopefully novel development. The excitement of data science also makes it especially important to separate opinion from fact. One useful advantage here that a statistician’s lens can provide is that of a professional skeptic—though this can sometimes cause the forest to be missed for the trees. This is why as a data science leader you will ideally want not only a strong working knowledge yourself of mathematics, statistics, and computer science but also a leadership team of trusted experts who can represent accurately the latest developments in these fields and enable you to synthesize them effectively.

2. Build and exercise the skills necessary to be a trusted advisor and good partner to decision makers within and across sectors; remain intellectually humble and curious.

To be taken seriously by leaders and decision makers from across sectors—public, private, and otherwise—requires developing and maintaining a broad contextual awareness of, understanding of, and

intellectual respect for all the various avenues and arenas to which data science might be brought to bear. This awareness will help you to find the resonances necessary to engage with those whose expertise, skills, and experience lie elsewhere. When this understanding complements genuine subject matter expertise, you are ideally placed to deliver advice that will be trusted and to build strong partnerships that will endure for the long term.

3. Cultivate, engage with, and show the utmost respect for domain expertise; and recognize that connecting with global grand challenges and societal questions is crucial.

Quite simply—and I have written about and advocated for this point on many occasions and in many venues before now—data science done right requires domain expertise. The lack of respect of domain expertise that at times shows itself in pockets of our community speaks in part to the natural tendency of all of us inventing hammers to sometimes see only nails. But perhaps the most significant structural risk to the future of data science is that we would somehow pass on this tendency to our students. This would render the future of data science incredibly intellectually impoverished and reduce it to nothing more than a technological flash in the pan.

To go several steps further, I would urge us all: Don't just reject the technical hubris and fallacious reasoning of 'N equals all' arguments and their brethren; place priority on helping to train the next generation of data scientists to be holistic in their understanding of problems, able to engage effectively with experts from other disciplines and non-experts from other walks of life, and cognizant of the social, ethical, and societal responsibilities inherent in our chosen line of work. As data science algorithms are more and more often repurposed from their original low-risk contexts to tackle questions of societal import where the cost of reaching erroneous conclusions can be enormously high, we can ill afford to ignore the broader context surrounding the techniques we develop and the ends to which they are put to use.

Training our next generation of data scientists holistically is of course a tall order, as is the matter of balancing a healthy portfolio of research and development efforts ranging from the very theoretical to the quite concrete in a way that engages with top-flight scholars across a range of academic domains. But at a time when our academic colleagues are compelled to pen essays such as "Data Is Not the Enemy of the Humanities" (Sinykin, 2021), we see just how important it is to set the right tone and expectations as leaders when engaging with scholars across various disciplines. Done right, this can provide the lifeblood that will keep data science fresh, innovative, and relevant in a way that ensures a future whole that is truly and determinedly greater than the sum of its constituent parts.

These reflections follow from a set of practical experiences and educations that I can promise you, dear reader, have been hard won. They are surely incomplete and at best scratch the surface of what is needed to propel this

exciting and critically important emerging field forward—yet I hope that they may prove to be of at least some good use.

Please indulge me if I close with a brief anecdote that might help remind us not to take our own individual disciplines too terribly seriously as we work together: Several years ago when I had the chance to help put some substantial data science investments into motion in the midst of a very competitive environment, internal arguments about the primacy of various disciplines naturally arose, and the remark was made that surely, statistics was the very beating heart of data science... Without missing a beat, one esteemed colleague shot back that, yes, of course that was all well and good, but mind you it was patently obvious that computer science was its soul!

... Well may the dialog continue, and with good leadership be to the benefit of this enormously important and enjoyable area...

Disclosure Statement

Patrick J. Wolfe has no financial or non-financial disclosures to share for this article.

References

Sinykin, D. (2021, April 29). *Data is not the enemy of the humanities*. The Chronicle of Higher Education. <https://www.chronicle.com/article/data-is-not-the-enemy-of-the-humanities>

©2021 Patrick J. Wolfe. 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.