Harvard Data Science Review • Issue 1.2, Fall 2019

Highlights of the Inaugural Data Science Leadership Summit

Jeannette M. Wing^{1,2} David Banks³

¹Data Science Institute, Columbia Institute, New York City, New York, United States of America,

²Department of Computer Science, The Fu Foundation School of Engineering and Applied Science, Columbia Institute, New York City, New York, United States of America,

³Statistical and Applied Mathematical Sciences Institute, Department of Statistical Science, Trinity College of Arts and Sciences, Duke University, Durham, North Carolina, United States of America

Published on: Nov 01, 2019

DOI: <https://doi.org/10.1162/99608f92.e45fcb79>

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

ABSTRACT

On March 26, 2018, [the Data Science Institute at Columbia University](#) hosted the inaugural Data Science Leadership Summit. It drew together a cross-section of academic researchers and administrators to discuss how university practices could best adapt to the emerging field of data science. The summit was co-funded by the National Science Foundation, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation. This interview with Jeannette Wing, the Avaneessians Director of the Data Science Institute at Columbia University and the author of [The Data Life Cycle](#), discusses the motivation of the summit and summarizes its key findings. The interview was conducted by another experienced leader in data science, David Banks, the Director of [the Statistical and Applied Mathematical Sciences Institute \(SAMSI\)](#).

Keywords: data science institute, education units, inter-disciplinary, multi-disciplinary, university administration

David Banks (DB): How did you get involved in issues related to university management of the emerging field of data science?

Jeannette Wing (JW): Even before I joined Columbia University as the director of the Data Science Institute in July 2017, I had the idea of getting together my counterparts—academic leaders of data science entities. I knew of a handful of universities which had data science institutes, but by the time I joined, another handful had sprung up. It seemed like good timing to get all of us together to share best practices.

In running the Data Science Institute at Columbia, which is a university-level and university-wide institute, I have the incredible and exciting opportunity to collaborate across the university, but I also face challenges by operating in the context of a large, decentralized campus. I was curious how people at other universities handled both the unique opportunities and challenges.

DB: What motivated you to organize the Data Science Leadership Summit? Why was it necessary?

JW: Data science is a new, emerging field and as such I wanted us in academia to start building a community. Academia has the rare opportunity now, just as the field is forming, to help shape what the field might look like in the future. Also, industry is moving very quickly in data science, hiring or retraining people to be data scientists. What academia says about data science can and should influence industry, and vice versa.

I was inspired by what the [Computing Research Association](#) (CRA) has done in building a community for computer science (CS). CRA started in 1972 and over the years has introduced many programs and activities. Two stand out, the two that motivated me to organize the Data Science Leadership Summit. First, CRA holds bi-annual meetings in Snowbird, Utah for the heads of computer science departments and schools as well as

industrial research labs to discuss common challenges faced by participants, all relevant to the vibrancy of the field. Discussions include the rise and fall of CS enrollments, federal funding opportunities and challenges, and policy issues affected by and that affect computing technology. Second, since 1974, CRA has conducted the annual Taulbee survey to track numbers for the field. It is the principal source of information on enrollment, production, and employment of computer scientists in North America. It provides academic salary and demographic data (including a breakdown by gender and ethnicity). I had both the Snowbird meeting and the Taulbee survey in mind when deciding to organize the first Academic Data Science Leadership Summit.

DB: What were the goals for the summit?

JW: The goals of the summit were to initiate the formation of an academic community for data science; to share best practices among academic leaders who face similar challenges and opportunities; and to take collective responsibility in the broader effort to prepare next-generation data scientists to contribute in the best interests of society.

This meeting was intended to be the inaugural meeting of a regular series. Also, given the existence of prior workshops and reports on data science, the intention was to minimize repeating what had been said before, but at the same time, provide for all attendees a common level of understanding of the state of data science in academia. One of the major outcomes of this meeting is the realization that many of the universities represented at the summit are just now working through the challenges in establishing a data science effort on the participants' respective campuses. At the same time, participants recognized the tremendous opportunity to help shape the field of data science and to respond to the overwhelming excitement for data science in academia and industry.

DB: Who attended the summit, and how were they chosen?

JW: Sixty-five participants from 29 public and private universities and three funding organizations in the US attended. I worked with the funding agencies and others who run data science institutes to decide on whom to invite. Word spread quickly and many people wanted to attend. There was clearly pent-up energy for such a summit to happen. All academic participants are leaders of data science institutes, centers, or initiatives on their respective campuses and/or leaders of projects funded by NSF, Sloan, and/or Moore.

DB: How do these universities organize data science on their respective campuses, infrastructurally?

JW: [The summit report](#) provides a good answer to that, so I'll quote it directly [pp. 10-12].

The opening question of the summit was: "How does data science fit into a university?" It led to a lively and engaging discussion. The answers reflect the ongoing broader community discussions about whether data science is a new field or not, what fields (e.g., in addition to computer science and statistics) feed its

foundations, what fields can benefit from the application of data science, and perhaps most importantly, how the field will evolve—what will data science look like in 10 to 20 years?

This opening question led to many participants responding, “Here’s what we do at my university.” One person said data science is within the university’s College of Information and Computer Sciences; one person said data science is part of the newly renamed Statistics and Data Science department; and another said data science defines a new Division of Data Sciences. Some reported that data science is a free-standing entity or a joint effort drawing from multiple departments or schools on campus. Some universities have multiple entities (institutes, departments, colleges, etc.) that support different aspects of data science, for example, research or education. Depending on the university, these entities report to a dean, multiple deans, the provost, or even the president. How an entity got started and how it will be sustained also vary across universities and influence its structure.

There were four main types of models discussed: (a) creating a brand new academic unit, for example, a School of Data Science; (b) repurposing an existing entity, e.g., adding data science to the statistics department; (c) creating a new entity (institute/center/initiative) that is not tied to any one or academic unit; and (d) creating a new entity that is joint with multiple academic units across campus. Some argued that even if data science draws on computer science, statistics, and other fields as its foundations, there is still value to having a separate entity, for example, an institute, that draws on these foundational disciplines, but transcends disciplinary boundaries. Some expressed concern that creating or repurposing a new academic disciplinary unit could continue to reinforce disciplinary silos.

DB: I bet they also reported a lot of challenges.

JW: Absolutely. [The report continues:](#)

Regardless of model, the immediate questions a university faces include:

- Which faculty are part of the new or extended entity? How is membership decided?
- What is the governance structure of the data science entity? To whom does the entity report?
- How are new faculty lines in data science allotted and distributed (e.g., is there a split with other departments)? Who pays for these lines along with startup costs?
- What is the role of the data science entity in education, research and service to the university? What should the balance be? Here, service means both people and computational infrastructure.
- Research/technical staff can play a critical role to serve disciplines across campus, bringing their data science expertise to domain experts. Is there funding to support such staff? Are there career development plans and pathways for such staff? How can universities attract such staff, given that they are in high demand by industry as well?

- Can the data science entity hire its own tenure-track faculty? Research faculty and technical staff? Can it run its own academic programs?
- If faculty have an academic home in one department and membership in the data science entity, how does one incentivize faculty to contribute, for example teaching and service, to the data science entity, and reward them for their contributions? How is faculty recruiting, hiring, mentoring, promotion, and evaluation done especially if faculty are from different disciplines?
- What is the financial model for supporting the data science entity? Does the data science entity get any tuition or indirect cost recovery (on grants)?
- What is the contribution that existing schools make toward supporting the data science entity?
- What is the long-term sustainability plan for the data science entity?

From the variety of ways in which participants described their structures, participants realized early on that no one model fits all universities. Each university has its own traditions, culture, funding models, and politics. Participants felt there was no need (at least in the course of one day) to come to consensus on what is best for all universities, especially since the field is still evolving. Rather, summit participants can best help the community, and in particular, university administrators, by providing a list of questions, such as those above, that each university would have to face and that many of us have had to face or are facing, along with responses where possible, to how such issues are being addressed currently. Given that so many universities are just now planning some kind of data science effort, this summit's contribution, as documented in this report, would be useful, practical, and timely.

DB: There are nine recommendations in your report, and some seem a little abstract. Could you summarize them and elaborate on why the participants feel they are needed?

JW: Five recommendations speak directly to the main purpose of the meeting: to build a community of academic data science leaders. This leadership community would convene regularly to share best practices, to provide guidance to university administrators, to monitor the health and growth of the field (e.g., through an annual survey), and to initiate and oversee shared activities (e.g., events, data sharing) as the field evolves. Since the first summit, a second summit was held and a third is being planned.

One recommendation speaks directly to master's level education in data science. Given hundreds of master's programs in data science sprouting up left and right around the world, participants agreed that we should help define what it means to be a data scientist through a set of standards. Standards would help potential employers and PhD programs know what to expect of someone coming out with an MS in data science degree and they would help calibrate current and future programs. We recommended working with industry, professional societies, and the National Academies of Science, Engineering, and Medicine to define these standards. There is a new effort by industry called "Industry Initiative for Analytics and Data Science Standards" toward similar goals, especially given the proliferation of titles. (I'm on IIADSS's advisory board.) Summit participants also discussed undergraduate and doctoral levels of education. Many reports, including a study by the National

Academies, which HDSR has already featured in its [inaugural issue](#), and academic papers address undergraduate education for data science, and participants felt there was no need to repeat their recommendations. For doctoral education, participants felt it was premature to make recommendations.

Two recommendations speak to ethics: the importance of defining a code of ethics for data science and the recommendation to train data scientists in data ethics. The summit report cites many ongoing community efforts by non-profits, industry, and academia to define a code and cites a growing list of courses on ethics and technology.

Finally, the last recommendation speaks to the need for industry and academia to explore new ways to bring data scientists to data held by industry in order to allow academics to test their models and analyses on industry data. Since industry collects and owns a lot of data, especially about people, they need to protect the privacy of their customers. At the same time, researchers in academia are hamstrung by not having access to real-world, large-scale data. This issue is similar to the problem decades ago when academic researchers in software engineering did not have access to large-scale software systems; open source changed that.

DB: Do you see a process for deciding what should be included in a typical data science curriculum? Should it include things like probability, Spark environments, maximum likelihood inference, database management systems?

JW: I hope the momentum built up from the academic data science leadership summits will crystallize into a process. It should involve both academia and industry. It should involve computer science and statistics and other related disciplines. The joint ACM-IMS [Association for Computing Machinery and Institute of Mathematical Statistics, respectively] [Interdisciplinary Summit on the Foundations of Data Science](#), held in June 2019, is a step in the right direction: to get the computer science and the statistics communities together at the level of their respective professional societies, building on what is happening among relevant departments per university.

The MS in Data Science program at Columbia sets a good bar: it's rigorous (three hard-core computer science courses plus three hard-core statistics courses), practical (a team-based capstone and ethics course), and flexible (three electives). The computer science courses are in algorithms, machine learning, and computer systems, each targeted for data science students. The statistics courses are in probability, statistical inference, and data exploration and visualization. All the topics you mention in your question—and much, much more—are covered by Columbia's curriculum.

Of course, each university is different, serving a different population, and any data science curriculum should be easy to tailor to the strengths and interests of a university's students and faculty.

DB: What challenges does data science face, both in general and in terms of how to manage it in an academic setting?

JW: Section 2 of the [summit report](#) addresses this question, so let me summarize it here.

As with any new field, the immediate challenge data science faces is defining the field. As with any scientific discipline, one can ask what are its theoretical foundations, methods used to discover new knowledge, the kinds of artifacts it studies and produces, the types of problems it solves, and, ultimately, its potential impact on humanity and society.

Data science presents a particular challenge to universities, organized by discipline, because data science is inherently multidisciplinary in two ways: depth and breadth.

In terms of depth, the technical foundations of data science draw on computer science and statistics, but are also informed by other areas of study, such as biostatistics, electrical engineering, mathematics, and operations research.

DB: I gather not everyone would agree on that, because some consider data science as computer science or statistics, dressed up.

JW: Indeed! Summit participants debated whether data science is a new field, emerging from the convergence of existing fields, or the evolution of an existing field. Those who see it as a new, emerging field see data science as drawing on methods from many existing fields, for example, computer science, mathematics, operations research, and statistics. Others see data science as simply an evolution of statistics, for example, anticipated as early as 1962 by John Tukey ([Donoho, 2017](#); [Tukey, 1962](#)), or an evolution of computer science, for example, as probabilistic and statistical reasoning becomes as important as symbolic and logical reasoning in computing. Regardless of whether data science is ‘new’ or not, there was consensus that concepts and techniques from (at least) computer science and statistics are core to data science.

DB: As a statistician, I am certainly happy to hear about such a consensus, but let’s not forget operations research, a field that has generated many of the optimization methods that are now routinely used by machine learning and data science in general.

JW: I cannot agree more. In fact, at Columbia, because operations research is prominent in both the engineering and business schools, we build the foundations of data science from three pillars of strength: computer science, statistics, and operations research. Moreover, as I mentioned in my [opening remarks](#) at the aforementioned ACM-IMS joint summit, the foundations of data science draws on signal processing from electrical engineering, analysis from applied math, and optimization from operations research.

DB: We talked about data science as being multidisciplinary in terms of depth. How about breadth?

JW: In terms of breadth, data science is used in context, e.g., to explore a data set, to create models, and/or to discover and test hypotheses—in a given domain. Because all domains generate or collect data, all domains

have the potential to benefit from the analytical techniques in data science. Thus, one can say data science methods can be applied to all fields, professions, and sectors.

In their *PNAS* 2017 article, “Science and Data Science,” Blei and Smyth emphasize the importance of the domain in data science, where “data scientists and domain experts” collaborate:

Data science focuses on exploiting the modern deluge of data for prediction, exploration, understanding, and intervention. It emphasizes the value and necessity of approximation and simplification; it values effective communication of the results of a data analysis and of the understanding about the world and data that we glean from it; it prioritizes an understanding of the optimization algorithms and transparently managing the inevitable tradeoff between accuracy and speed; it promotes domain-specific analyses, where data scientists and domain experts work together to balance appropriate assumptions with computationally efficient methods ([Blei & Smyth, 2017](#)).

In his 2007 presentation to the National Academies’ Computer Science and Telecommunications Board, Jim Gray anticipated data science by arguing the centrality of data (“the fourth paradigm”) for driving new discovery in science, for example, astronomy and biology (Hey, Stewart, & Tolle, 2009). Ten years later, the community now recognizes that data science is applicable not just to science disciplines, but to *all* disciplines.

DB: With grand opportunities come grand challenges.

JW: True. Any field that emerges from existing fields will face certain challenges in terms of crossing boundaries: communication and culture, faculty hiring and promotion, faculty service, joint degree programs, and so on. Universities have handled such emerging fields in the past, helping to break down disciplinary boundaries. Computer science is one example. Computational biology is a more recent example.

It is the second multidisciplinary aspect of data science, however, that presents an unusual challenge for most universities. How does one embrace a field that has the potential to transform every other field on campus?

Thus, the biggest administrative challenge is trying to figure out where to ‘put’ data science. Since data is everywhere and everyone has data, it’s not easy to compartmentalize it and tuck it away within a traditional university structure that aligns with disciplinary boundaries.

DB: I know that too well from my own experiences with such administrative challenges. Speaking of which, both of us have some real ones waiting for us as Directors, and I am sure we both feel that we never have enough time to address them all, so let me conclude this interview by thanking you for your leadership, Jeannette! This is an important challenge for the future of many fields, and on behalf of many, I appreciate your work to address it.

Disclosure Statement

David Banks and Jeannette M. Wing have no financial or non-financial disclosures to share for this interview.

References

- Blei, D., & Smyth, P. (2017, June). Science and data science. *Proceedings of the National Academies of Sciences*, 114(33), 8689–8692. <https://doi.org/10.1073/pnas.1702076114>
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>
- Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft Research.
- Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematics Statistics*, 33(1), 1–67. <https://doi.org/10.1214/aoms/1177704711>
-

©2019 David Banks and Jeannette M. Wing. 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.