Accelerating Data Science Research at the University of Delaware

White Paper v1.0
October 2017

University of Delaware Research Office
Contact: Dr. Anshuman Razdan
razdan@udel.edu

Executive Summary

Data science is a rapidly emerging interdisciplinary endeavor with the potential to impact virtually all aspects of human activity. This white paper—an outgrowth of the University of Delaware's Data Science Symposium—highlights national trends, key activities on campus and exciting opportunities for advancing data science research at UD for the benefit of students, faculty and staff.

The University of Delaware must act quickly and decisively in the area of data science. Top-tier public research institutions across the United States are investing heavily in the creation of data science institutes, research in data science, and educational programs in data science at both the undergraduate and graduate levels. A key recommendation in this document is that the University of Delaware create a data science institute responsible for catalyzing and coordinating emerging data science efforts at UD.

With a unique set of strengths and planned growth in faculty, undergraduate students and graduate students, UD is well positioned to capitalize on the data science revolution. The proposed institute will serve as a magnet for industrial interactions and fundraising, help ensure that our faculty remain on the cutting-edge of research across many disciplines, and help guarantee that our students are well-trained in this rapidly growing area, contributing to the labor force of the future. For students of all races, ethnic groups and orientations, data science knowledge and skills lead to intellectually rewarding careers and economic opportunity.

Recommendations

The following recommendations will strengthen and enhance UD's capabilities in research and education with respect to data science.

- Establish—and commit sustenance to—the UD Data Science Institute as a catalytic and coordinating force for data science efforts at UD.
- Identify a strong and broadly recognized institute director.
- Work collaboratively with academic departments to encourage the development of undergraduate and graduate programs in data science across all colleges, leveraging the resources of the Office of Undergraduate Research and the Office of Graduate and Professional Education.
- Engage in directed faculty hiring both in theoretical foundations and applied areas of data science. Pursue cluster hires with the aim to fill the gaps discussed in this paper at the University level rather than the college/department level.
- Foster industry engagement in UD data science research, education and entrepreneurship activities.
- Recognize key federal funding opportunities in data science and organize relevant teams to respond.
- Raise the profile of the UD Data Science Institute as a high-priority item in the University's development campaign scheduled to kick off in November 2017.
- Continue to build momentum for this critical area by organizing seminars and workshops on data science and hosting external speakers from a variety of disciplines.

Introduction

Data science, otherwise known as data-driven science, has emerged as an interdisciplinary field that encompasses statistics, computer science, mathematics, information sciences and numerous related fields. In leveraging tools from each of these disciplines, data science seeks to discover and/or extract scientific knowledge from data. Often the data of interest can be characterized as "big data," but that is not always the case.

In their 2017 contribution to the *Proceedings of the National Academy of Sciences*, David Blei and Padhraic Smyth argued that the effective combination of statistical, computational and human-oriented research perspectives represents the *essence* of data science. These researchers go on to note that "[d]ata 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 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 and Smyth, 2017).

Others have previously characterized data science more succinctly as a "concept to unify statistics, data analysis and their related methods" (Hayashi, 1998).

Such qualities provide data science with the potential to revolutionize research, governance and business activities in the coming decades. This realization, in part, led Turing Award winner Jim Grey to famously note that "almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, data-intensive science paradigm is emerging" (Edwards and Gaber, 2014:1). These qualities also underlie IBM's recent prediction that the demand for data scientists will "soar 28% by 2020" (Columbus, 2017).

As new and diverse datasets rapidly become available in nearly every aspect of life, data science has the potential to advance human understanding in all branches of science and the humanities and to address grand challenges facing society. It is now possible to combine disparate, dynamic and distributed datasets for tasks related to predicting the future behavior of complex systems, identifying precise medical treatments, maximizing smart energy usage and measuring the impacts of social and educational policies and interventions. Numerous federal granting agencies have identified the important role of data science in the 21st century.

Addressing grand-challenge problems surrounding data science requires cross-disciplinary collaborations. To capitalize on the data revolution, teams of data scientists and disciplinary researchers that can work AND think together must be forged. Past work has shown that data science lacks maximal impact without the knowledge, involvement and collaboration of those who have a deep understanding of a given grand-challenge problem. Likewise, disciplinary researchers who seek to address critical problems are better able to do so when data science experts are leveraged to support claims, decisions and new strategies.

National Landscape

The Data Science Working Group reviewed information from both public and private institutions that are members of the Association of American Universities (AAU). Top data science programs are underway at UC Berkeley, Brown, New York University, Columbia and Carnegie Mellon.

Out of 33 AAU public institutions, at least 28 (85%) have university-wide data science institutes or centers. Similarly, 23 of the 26 AAU private universities (89%) had such institutes or centers. It is important to note that the highest performing research universities in the AAU have devoted considerable resources to data science initiatives that are multidisciplinary and span multiple colleges. The resources for these initiatives have come from internal funding as well as from development funds from alumni and other donors. A list of AAU institutions and relevant centers is in Appendix A.

University of Delaware Landscape

UD Data Scientists

An attempt was made to identify data scientists through faculty records in UD Academe. Keyword analysis of publication and sponsored research activity for more than 1,000 faculty was used to provide a rough approximation of whether or not a UD faculty member's research program incorporated aspects of data science. The following keywords and their text variations were used: *data, computing, computational, informatic, data mining, modeling, simulation, analytic, analysis, information, database, text mining, machine learning, statistic, omic*. The keyword-matched results from publications and grants were combined and filtered, resulting in a list of 166 faculty considered likely to be data scientists.

A total of 13,622 journal publications from 956 faculty members and 840 conference proceedings from 159 faculty over the last five years (2012–2017) were analyzed. The keyword search revealed that about half of all faculty had at least one match to the titles of their papers or the names of the journals or conference proceedings. A filter for each faculty's records was applied based on the number of matches (at least five matching papers) and the percent of matches (at least 20% of papers matching the data science-relevant keywords). As a result of this analysis, 136 faculty were identified as likely data scientists from the publications.

From sponsored research data over the past 10 years (2007–2017), encompassing 6,201 awards from 1,080 faculty and using the keywords cited above, 259 faculty were identified as potential data scientists based on their award titles. Using the filtering step noted previously for consistency, only 32 faculty were identified as data scientists based on sponsored research alone, which is a low number, but perhaps reflective of the relatively recent emphasis on data science within sponsored research initiatives.

UD Data Science Symposium

On May 12, 2017, UD's Data Science Symposium brought together faculty from across campus with the goal to understand common interests, capacities and aspirations for data science research at UD. Organized by the Research Office, the symposium included external speakers from Harvard, Georgia Tech and the National Science Foundation, remarks from President Assanis, faculty research presentations, breakout sessions and graduate student poster presentations. One hundred and twenty-five UD researchers and faculty registered, and 98 attended. Valuable information was gathered, providing the foundation for the recommendations of this white paper.

UD Strengths in Data Science. The symposium and subsequent analysis of feedback from its attendees identified a number of areas of institutional strength. With respect to the *theoretical underpinnings of data science,* UD holds comparative advantages over competitor institutes in data science theory and computation:

- The Department of Mathematical Sciences is illustrative of these comparative advantages and boasts several recent data science hires in these areas.
- UD's core strengths in theoretical and computational data science are further underpinned by the Department of Computer and Information Sciences, the Department of Electrical and Computer Engineering and related engineering programs. These departments and programs provide institutional leadership and rigorous training in data science programming and theory; including at the undergraduate level through undergraduate programming competitions such the IBM-sponsored Hackathon.
- Similarly, the Department of Applied Economics and Statistics has a core of early career faculty who have training in and professional experience in big data. This department has longstanding collaborative working relationships with institutes and centers across

campus such as the Center for Bioinformatics and Computational Biology, the Institute for Financial Services Analytics and the Delaware Health Sciences Alliance.

- UD has substantial research capacity at several theoretical frontiers in data science. These include positivity preservers, non-commutative harmonic analysis, topological data analysis and phylogenetic regularization. Theoretical advances by investigators at UD have already yielded innovative approaches to seemingly intractable challenges such as the reconstruction of our paleoclimate and our understanding of the function of neural circuits in the hippocampus. A data science institute will create institutional synergies by bringing together theoretical insight and innovation, as well as domain expertise.

With respect to *data science applications*, UD exhibits comparative advantages in these areas (in alphabetical order):

- **Astronomy and Physics**: UD has built strengths in high-performance computing and the analysis of large datasets. On the theoretical side, UD has expertise in numerical simulations of subjects ranging from atoms and material science, to biomolecular physics and planet formation. UD's Bartol Research Institute has a long history as a leading center in particle astrophysics and space physics, and faculty are engaged in the next generation of international collaborations that will generate petabyte-scale datasets, such as the upgraded IceCube Neutrino Observatory, Cherenkov Telescope Array, Large Synoptical Survey Telescope and Parker Solar Probe. Deeper engagement in interdisciplinary data science efforts is the natural next step to bring these core disciplinary strengths to a wide range of scientific research problems and the education and training of scientists.
- **Bioinformatics, Health and Life Sciences**: UD has significant computing infrastructure and scientific expertise in bioinformatics and computational biomedicine, highly regarded graduate programs (master's, Ph.D., certificate) in bioinformatics and systems biology, and active engagement in regional, national and international biomedical informatics research networks including the NIH Big Data to Knowledge (BD2K) program. Delaware is an ideal population health laboratory—with 950,000 people socially and demographically representative of the U.S. and highly linked clinical informatics systems—presenting unique opportunities for UD and partner health institutions to develop a national model for big data in precision medicine.
- **Civil Infrastructure/Transportation Systems**: The UD Department of Civil and Environmental Engineering, in conjunction with the Delaware Department of Transportation, has collected a broad array of transportation/infrastructure data over the years, including bridge deterioration data, highway performance data and traffic flow data using GPS. In recent years, the Delaware Transportation Center, which is part of the College of Engineering, has been the beneficiary of several federal transportation grants and initiatives.
- **Education**: The past 15 years have seen a revolution in education research with an explicit focus on understanding "what works" to improve students' outcomes, and UD's College of Education and Human Development (CEHD) has been a leader in the effort to conduct rigorous large-scale research studies to develop and test new interventions. External funding for CEHD research over the previous five years has passed $35 million. The Center for Research in Education and Social Policy currently is conducting large-scale surveys and randomized field trials involving thousands of study participants in schools and communities across the nation.

- **Environment/Climate/GIS**: Delaware's environment is arguably the most comprehensively monitored of any state, allowing for innovation in the use of high-frequency, real-time environmental sensors and subsequent development of unique decision support and data visualization systems. The University is a major contributor to this effort with assets including the Center for Environmental Monitoring and Analysis, the Delaware Geological Survey and the Delaware Environmental Institute, to name only a few. In addition, strong, long-term partnerships are in place between environmental researchers at the University and state agencies, allowing for a smooth transfer of technology for the benefit of the public.
- **Financial Data**: The banking industry has long been a strength in Delaware's economy. It was therefore natural for the University of Delaware to establish the Institute for Financial Services Analytics, a joint program between the Lerner College of Business and Economics and the College of Engineering with collaboration between the University and JPMorgan Chase. The first of its kind institute offers an interdisciplinary Ph.D. in financial services analytics with research focusing on the theory of data science as well as its application in risk management, security, enhanced customer services and business operations.
- **Statistics**: An essential component of any data science curriculum is statistics. UD's Statistics Program has experience training M.S. and undergraduate students in statistical analysis, data management and statistical computing. UD is able to contribute to both theoretical and applied aspects of data science. Our early career core faculty specialize in developing new methodologies in big data, genomics, dealing with excessive zeros and recognizing false positives. The Statistics Program maintains an internship program with many of Delaware's credit card banks and has hundreds of alumni working in the banking industry.
- **Politics and Policy**: The state of Delaware affords UD unprecedented opportunities for engagement and collaboration with state, local and national government. These qualities—alongside the synergies and expertise found within UD's School of Public Policy and Administration, Department of Political Science and International Relations, Biden Institute and Center for Political Communication—underscore UD's potential as a national leader in policy-oriented data science and fostering real-world data science impacts.

While UD has obvious strengths in data science, the UD Data Science Symposium also provided evidence to suggest that UD's current institutional strengths in data science are falling below their potential.

Symposium attendees widely identified low levels of collaboration and communication among UD data scientists as a key reason for these shortfalls. Many attendees pointed to the fragmented nature of data science research at UD and the lack of centralized leadership, which within the areas of theoretical and computational data science research has led to unnecessary repetition in algorithm and software development among data science groups. These same theoretical and computational data scientists also noted difficulties in connecting social science domain experts with computational experts in their data science research.

A similar concern was echoed by the social scientists in attendance at the symposium, who emphasized difficulties in identifying and accessing the necessary computational and software expertise for their applied data science projects. Additionally, gaps in undergraduate and graduate level data science academic programs make attracting future data scientists to UD a struggle. Infrastructure weaknesses in the areas of high-performance computing and data processing also present challenges for storage and maintenance of large data sets.

To address these concerns and weaknesses, virtually all groups of data science researchers in attendance strongly voiced the need for an institutional space to (a) spur interaction and collaboration among data science researchers at UD; (b) facilitate the sharing, management and development of data science algorithms, software and hardware; and (c) improve graduate student education in data science research.

Faculty Hiring Needs
The following hiring needs were identified during the breakout sessions at the Data Science Symposium. This summary should not be considered a comprehensive list of hiring needs in this area:

- Theoretical foundations of data science which bring together statistics, mathematics, signal processing, and computer science and engineering communities.
- Basic computational support and software and database development (core facility)
- GIS expertise and biostatisticians (application-focused expertise)
- Quantitative methods across social sciences, behavioral science and digital humanities
- Cybersecurity and machine learning
- Computational scientists in most natural science departments (Chemistry, Mathematics, Physics, Biology, Psychology, Linguistics)

Needed hiring falls into two broad categories. In the first category, "theoretical foundations," we see the need to grow our expertise in the three pillars of data science: statistics, mathematics and computer science. We emphasize that this expertise is not tied to *departments*, but rather to the nature of the work conducted. Faculty hired to support this category, while having demonstrated a clear attention to applications, will primarily be methodologists focused on building the underlying theory, tools and techniques necessary to extract meaning from large unstructured datasets.

In the second category, "applications," we see the need to grow our expertise in those faculty whose primary focus is a disciplinary one, but who bring the tools of data science to bear on these problems. Ideally, faculty hired in both areas will overlap significantly, speak a common language and together form the core of the proposed data science institute. We recommend that hiring in data science be accomplished both through department-based disciplinary searches and through university-wide cluster hire searches, with the latter of special importance. Cluster hiring in this area should focus on identifying deeply interdisciplinary researchers of the type who do not always fit naturally into department-based searches.

UD's Roadmap to Success in Data Science Research

Essential elements of UD's roadmap to success in data science include a University-wide institute that is appropriately resourced; faculty hiring coordinated at the University level; and the development of robust, interdisciplinary undergraduate and graduate programs that exemplify UD's commitment to workforce training through real-world learning and entrepreneurship.

Research Institute
A proposal has been submitted to the Unidel Foundation to establish the University of Delaware Institute for Data Science (UDIDS). The institute will serve as a nucleating unit, leading to an externally funded and self-sustaining data science initiative with national impact.

A large number of UD faculty work with large or complex datasets, representing untapped potential that this initiative will address. These investigators currently work in units and centers scattered across the seven colleges with no cohesion or platforms for mutual support. UDIDS will provide the organization and resources to leverage UD's intellectual capital to its greatest

potential. Such an institute will move issues related to the nature of large or complex datasets to the forefront where they are currently secondary to the disciplinary questions being investigated. Also, UDIDS will provide a focal point for the key disciplines involved in the theoretical foundations of data science. Finally, UDIDS will make our faculty more competitive for federal funding opportunities directly or indirectly involved in data science.

UDIDS will be cross-disciplinary with two specific aims: To establish and advance research in data science and related disciplines at UD; and to develop and execute longer-term strategic plans toward excellence in data science. While the institute will focus on research, like other UD research institutes, it will serve to foster the development of new multidisciplinary data science degree programs, as well as strengthen related academic programs at UD, enhancing our ability to meet the workforce development needs of the state and region.

Academic Programs

Although the Data Science Working Group did not expressly address issues of academic programs at UD, it is important to note that initiatives like NSF's Transdisciplinary Research in Principles of Data Science (TRIPODS), with its support for collaborative institutes, require a workforce development program.

Also, during the UD Data Science Symposium, it was noted that there is a need for courses and faculty to teach them to meet the needs of graduate students and undergraduates who are interested in data science. Further, attendees expressed interest in having a central listing of data science courses available to students.

Data science does not enjoy an agreed-upon definition, nor are there widely accepted frameworks for programs and curricula. Nonetheless, many diverse industries are investing heavily in data science in the hopes of gaining commercial advantage from the vast amounts of diverse data available. Students are interested in the field. While UD has not yet created any general data science programs, we offer data-heavy specialized programs like the Ph.D. in Financial Services Analytics. An undergraduate 4+1 program in GIS Science and Environmental Data Analytics also soon will be proposed.

To bring some clarity to the field as an academic subject, the National Academies of Science, Engineering and Medicine have convened three roundtables (one, two and three) on post-secondary data science education. They also hosted a workshop series on envisioning an undergraduate data science education. A final report is expected in a year and the interim report can be found here.

The University of Delaware offers a range of graduate and undergraduate courses in statistics, machine learning, data structures, data mining and topological data science, as well as domain-specific courses. Thus, UD is well-positioned to offer programs and certificates at all levels. We recommend that UD convene a committee to coordinate, support and expand courses and programs in data science, cognizant of national and international best practices and UD's specific strengths, resources and needs.

Partnership with Industry

An important component of the path forward is to engage with industry. Data analytics has become a staple for consumer and social media-driven industry. Foundations of data analytics such as machine learning, artificial intelligence (AI) and data mining are almost discipline/ application agnostic. Thus, engaging with industry in Delaware, the greater Mid-Atlantic region and nationally, is extremely important. Industry can provide research support and direction, influence and inform academic programming, and collaborate in workforce development and entrepreneurship opportunities for our students.

The Institute for Financial Services Analytics in the Lerner College of Business and Economics has partnered with JPMorgan Chase and provides an excellent model for industrial participation. The UD Data Science Institute also should collaborate with the Office of Economic Innovation and Partnerships (OEIP) and the Horn Program to aggressively explore this avenue.

Diversity in Data Science

Data Scientists are critically needed. The United States will need 1.5 million more data professionals, and 140,000–190,000 more professionals with data analytic skills by 2020 (Berman and Bourne, 2015). Unfortunately, the number of women and underrepresented minorities in data science is relatively low. Efforts are underway to improve inclusive representation in data science, including national initiatives such as Yes We Code, Girl Smarts and TechGirlz, which empower millions of underserved students to explore their talent in STEM fields. As one example at UD, Professor Lori Pollock leads the NSF-supported Partner 4 Computer Science (Partner4CS) to increase participation in computer science among Delaware K–12 teachers and students. There is evidence that those students who have access to computers early on are more apt to consider seeking a career that includes computing.

At the inaugural Women in Data Science conference held at Stanford earlier this year, 10 simple rules for increasing diversity in data science were presented. (Please see Appendix B.) We aim to follow these guiding principles in our efforts to enhance the representation of women *and* other underrepresented groups within data science at UD. Doing so will help broaden our recruitment pipeline and enable UD to effectively communicate that women and underrepresented minorities should consider data science as a career path. Along these lines, our graduate-level diversity efforts will be complemented by revisions in UD's undergraduate curriculum to ensure broad and diverse participation in data science at both the undergraduate and graduate levels.

Bolstering UD's potential to build greater diversity in data science is UD ADVANCE, which is aimed at increasing opportunities for UD's women faculty. Supported by a five-year Institutional Transformation grant from NSF through summer 2019, UD ADVANCE is helping to energize our campus and creating new tools to forge a more inclusive community that will benefit the University as it creates possibilities for every individual.

The UD Data Science Institute will work to uphold the mission of inclusive excellence set forth by the University's existing diversity initiatives. Such programs and opportunities increase the potential for underrepresented groups to participate fully in the next digital revolution.

Timeline

- Spring 2017: Early discussions, Data Science Symposium (May 12)
- Summer 2017: Unidel Foundation application submitted for establishing UD Institute for Data Science. Ranked high internally, awaiting Unidel Foundation approval.
- Fall 2017: Development of (this) white paper, reviewed by Vice President for Research, Scholarship and Innovation and University leadership.
- Spring 2018: Establish UD Data Science Institute. Set up governance and search for a director.
- Faculty hiring: New faculty hires start in the 2018 academic year and beyond. As of mid-October 2017, Provost has given approvals for cluster hires in data science, CIS has announced faculty search with big data as one of the areas, and other searches may be underway that call for data science/analytics expertise.
- Academic programming in data science: This area needs further attention. While there is growing interest among UD faculty to start undergraduate and graduate programs, the

academic process is more complex than launching a research initiative. As of mid-October 2017, there is a discussion underway about establishing a graduate (MS) program in data science at UD. These discussions are at a very preliminary stage.

Current Funding Climate

Data science and related fields (high performance computing, machine learning, etc.) are of great interest to federal funding agencies, industry and philanthropic organizations. At least $100 million in federal funding opportunities are "open" in these areas at any given instance. Almost all federal agencies have programs related to data sciences (NSF's *Convergence* and *TRIPODS* and NIH's *Big Data to Knowledge* are just a few leading ones) with goals to establish national centers of excellence. Related disciplines in computer science and engineering also are driving these needs, and other fields of scholarship—from agriculture to digital humanities to medicine—have demonstrated varying degrees of interest.

In response to these trends, government agencies are now beginning to place data science at the forefront of their funding initiatives. For example, in its efforts to highlight the value of scientific convergence, NSF has identified *Harnessing the Data Revolution for 21st Century Science and Engineering* as one of four areas of potential for catalyzing new research directions and advancing scientific discovery and innovation. DARPA'S recent Data Driven Discovery of Models (D3M) program and the European Space Agency's Copernicus Masters Competition are both reflective of similar initiatives.

Likewise, as biomedical tools and technologies rapidly improve, researchers in the biomedical arena are increasingly producing and analyzing a rapidly expanding amount of complex biological data. In response, NIH launched the Big Data to Knowledge (BD2K) program to facilitate the broad use of biomedical big data, develop and disseminate analysis methods and software, enhance training relevant for large-scale data analysis, and establish centers of excellence for biomedical big data. The BD2K program also has supported initial efforts toward making data sets "FAIR"—Findable, Accessible, Interoperable and Reusable.

Data-driven consumer profiles are being used in all parts of the industry. Philanthropic institutions are not far behind in support, from scholarships to policy and social implications perspectives. The Flatirons Institute (Simons Foundation) and Taner Halicioglu's gift of $75 million to his alma mater, UC San Diego, to establish a data science institute, are two such examples.

Conclusion

The UD Data Science Symposium—and a review of data science research on campus—reveal immense potential for future growth. UD data scientists are excited about developing this area and are firmly committed to cross-disciplinarity and diversity in doing so.

The symposium also helped to reveal UD's considerable strengths in data science research. These include theoretical and computing "nodes of excellence" within the Department of Mathematical Sciences, Department of Computer and Information Sciences, and Department of Electrical and Computer Engineering. In terms of data science applications, UD likewise has current strengths, and centers, in biotechnology, bioinformatics and systems biology, environmental sciences, astronomy, policy, education, financial analytics and research ethics. We believe these existing strengths can serve as a strong foundation for collaborative data science at UD, with an anticipated focus on both data science theory and applications. Such a foundation, in turn, has the potential to transform UD into an indispensable national leader in theoretical and applied data science research.

However, to achieve these goals, the University of Delaware Institute for Data Science (UDIDS) is critically needed. This institute will provide a physical space for face-to-face collaborations and interactions, as well as the requisite software and hardware for computationally and data-intensive collaborative research endeavors. The faculty hiring needs that have been identified are targeted to directly support such an institute, build upon UD's current strengths *and* connect what is currently a fairly disparate data science community. Together, these "next steps" will allow UD to fully tap the potential of its current 120+ data scientists, and will put UD on the path to becoming a national leader in both research and real-world impact in data science.

Contributors

The Research Office would like to thank the following who contributed to the development of this white paper (in alphabetical order):

- Gonzalo Arce
- Ben Bagozzi
- Tracey Bryant
- John Gizis
- Dan Leathers
- Henry May
- Nii Attoh-Okine
- John Pelesko
- Anshuman Razdan
- Lou Rossi
- Eric Wommack
- Cathy Wu

References

Berman F. D., and P. E. Bourne. 2015. “Let’s make Gender Diversity in Data Science a Priority Right from the Start.” *PLoS Biol* 13(7): e1002206. https://doi.org/10.1371/journal.pbio.1002206

Blei, D. M., and P. Smyth. 2017. “Science and Data Science.” *PNAS* 114(33): 8689–8692.

Columbus, L. (5/17/2017). “IBM Predicts Demand for Data Science Will Soar 28% by 2020.” *Forbes.* https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#3fd8b5eb7e3b Accessed on 9/12/2017.

Edwards, K., and M. M. Gaber. 2014. *Astronomy and Big Data: A Data Clustering Approach to Identifying Uncertain Galaxy Morphology*. Springer International Publishing, Switzerland.

Hayashi, C. 1998. "What Is Data Science? Fundamental Concepts and a Heuristic Example." In Data Science, Classification, and Related Methods: Studies in Classification, Data Analysis, and Knowledge Organization, pp. 40–51. Eds: C. Hayashi, K. Yajima, H. Bock, N. Ohsumi, Y. Tanaka, and Y. Baba. Springer International Publishing, Japan.

Appendix A

AAU Data Science Universities

Public:	Website
Georgia Institute of Technology	http://bigdata.gatech.edu/
Indiana University	http://www.soic.indiana.edu/faculty-research/research/center-data-search-informatics.html
Iowa State University	
Michigan State University	https://icer.msu.edu/
The Ohio State University	https://www.osc.edu/
The Pennsylvania State University	http://discovery-informatics.ist.psu.edu/
Purdue University	https://engineering.purdue.edu/PURVAC/
Rutgers University-New Brunswick	http://idsla.newark.rutgers.edu/
Stoney Brook University	http://www.iacs.stonybrook.edu/
Texas A&M University	http://isc.tamu.edu/
University at Buffalo	http://www.buffalo.edu/genomics.html
The University of Arizona	
University of California- Berkeley	https://bids.berkeley.edu/
University of California- Irvine	http://datascience.uci.edu/
University of California- Los Angeles	http://scai.cs.ucla.edu/
University of California, San Diego	http://www.sdsc.edu/
University of California, Santa Barbara	http://www.cs.ucsb.edu/research/computational-science-and-engineering
University of Colorado, Boulder	

University of Florida	
University of Illinois at Urbana-Champaign	http://datascience.cs.illinois.edu/page/research
The University of Iowa	https://uiowa.edu/datascience/
The University of Kansas	
University of Maryland at College Park	https://www.umiacs.umd.edu/
University of Michigan	http://midas.umich.edu/dsi/
University of Minnesota, Twin Cities	https://www.msi.umn.edu/
University of Missouri, Columbia	http://muii.missouri.edu/index.php
The University of North Carolina at Chapel Hill	https://research.ncsu.edu/dsi/about/
University of Oregon	
University of Pittsburgh	Center for Computational Biology and Bioinformatics
The University of Texas at Austin	https://www.ices.utexas.edu/
University of Virginia	https://dsi.virginia.edu/
University of Washington	http://escience.washington.edu/
The University of Wisconsin- Madison	https://research.wisc.edu/funding/uw2020/round-2-projects/transforming-data-science-at-uw-madison-and-beyond/
Private:	Website
Boston University	https://www.bu.edu/datascience/about-dsi/
Brandeis University	https://lts.brandeis.edu/research/datamanagement.html
Brown University	https://www.brown.edu/initiatives/data-science/
California Institute of Technology	http://cd3.caltech.edu/

Carnegie Mellon University	http://www.hcii.cmu.edu/
Case Western Reserve University	
Columbia University	http://datascience.columbia.edu/
Cornell University	https://www.cac.cornell.edu/
Duke University	http://childandfamilypolicy.duke.edu/research/nc-education-data-center/
Emory University	
Harvard University	http://iacs.seas.harvard.edu/
The Johns Hopkins University	http://idies.jhu.edu/
Massachusetts Institute of Technology	https://idss.mit.edu/
New York University	http://cds.nyu.edu/
Northwestern University	
Princeton University	http://www.princeton.edu/researchcomputing/
Rice University	http://research.computing.yale.edu/
Stanford University	https://sdsi.stanford.edu/
Tulane University	http://grdc.sphtm.tulane.edu/
The University of Chicago	https://www.ci.uchicago.edu/
University of Pennsylvania	https://pics.upenn.edu/
University of Rochester	http://www.sas.rochester.edu/dsc/research/index.html
University of Southern California	http://www.isi.edu/home
Vanderbilt University	http://www.isis.vanderbilt.edu/
Washington University in St. Louis	https://olin.wustl.edu/EN-US/Faculty-Research/research-centers/center-customer-analytics-big-data/Pages/default.aspx
Yale University	http://research.computing.yale.edu/

Canadian:	
McGill University	
University of Toronto	

Appendix B

Ten Simple Rules for increasing Gender Diversity in Data Science
From the 2017 Inaugural Women in Data Science Conference, Stanford

1. Foster a recruitment process that seeks out diverse candidate pools and engages in targeted and intentional outreach efforts that attract a diverse applicant pool.
2. Monitor and promote pay equity.
3. Develop organizational mechanisms for promoting diversity in which success of such efforts can be measured and rewarded.
4. Provide leadership opportunities for women, promote their efforts, and help women identify advancement opportunities.
5. Make diversity a strategic priority and expect those who work for and/or with us to do so as well.
6. Put women colleagues and students up for awards and recognitions. Share their work with colleagues. Provide mentorship that helps them navigate the pathways to success. Help clear those paths.
7. Raise awareness of diversity. If you are asked to present, be on a panel, or serve on a committee, ask if there are (other) women participating. If not, suggest names of women to invite. (Bonus points for not accepting the response that finding good women candidates is hard.) This approach is particularly effective when both men and women ask these questions.
8. Help create positive language and social expectations around data science: assume that women will be part of the process and part of the lea