




## RESEARCH ARTICLE

# The next wave: We will all be data scientists

Margaret Betz<sup>1</sup> | Ellen Gundlach<sup>1</sup> | Elizabett Hillery<sup>2</sup> | Jenna Rickus<sup>3,4</sup> |  
Mark D. Ward<sup>1,5</sup> 

<sup>1</sup>The Data Mine, Purdue University, West Lafayette, Indiana

<sup>2</sup>Research Computing, Purdue University, West Lafayette, Indiana

<sup>3</sup>Office of the Provost, Purdue University, West Lafayette, Indiana

<sup>4</sup>Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, Indiana

<sup>5</sup>Department of Statistics, Purdue University, West Lafayette, Indiana

**Correspondence**

Mark D. Ward, The Data Mine, Purdue University, 1301 Third Street, West Lafayette, IN 47906.  
Email: mdw@purdue.edu

**Funding information**

Cummins Incorporated; Foundation for Food and Agriculture Research, Grant/Award Number: 534662; National Institute of Food and Agriculture; National Science Foundation, Grant/Award Numbers: 0939370, 1246818; Society of Actuaries

**Abstract**

In the next wave of educating future data scientists, we need to think of all undergraduate students, regardless of background or major, as future data scientists. We should train them in supportive, interdisciplinary environments. Starting from their first day at college, they should be given the opportunity to apply powerful tools to large data sets, using real-world problems. Partnerships with research computing, academic departments, research centers, companies, government, and nonprofits will all be necessary to fully prepare these students for the breadth of the data science workforce.

**KEYWORDS**

active learning, data science education, high-performance computing, learning community, undergraduate

## 1 | INTRODUCTION

*Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5* [5] warned of the looming STEM workforce shortage, particularly in the United States. At the time of that report's release, data science was on the rise. Now it is a crucial field that touches every industry and research area. Thus, the next wave of future data science needs to be open to all undergraduate students, regardless of background or major. Moreover, data science and statistics benefit from an interdisciplinary

approach with a broad range of perspectives, as illustrated in *Leadership and Women in Statistics* [4]. National Science Foundation (NSF) has launched the “Harnessing the Data Revolution” initiatives, and universities are building data science programs. With proper support from faculty, teaching assistants, and their peers, we know that undergraduate students starting on their first day of their first year of college—even with no background in computing—can work with massive data sets on research computing clusters. Donoho [1] warns against data science that is “motivated by commercial rather than

intellectual developments.” Instead, Donoho urges the need for “learning from data.” Nolan and Temple Lang [6] provide a framework for learning data science content and methods. The GAISE College Report [2] emphasizes principles for how to teach students about statistics. In response to these needs to create learning environments for students, in which students can learn data science tools in tandem with disciplinary expertise and domain knowledge, we have created The Data Mine program at Purdue University. We have created active learning seminars in which undergraduate students use real data and high-performance computing (HPC). Our undergraduate students complete meaningful projects as an on-ramp (from any major) to data science. Our paper describes two Purdue initiatives that introduce both data science and HPC skills to a wide variety of undergraduate students. We recognize that many colleges and universities have created new data science programs. In this short space, it would be impossible to give a comprehensive landscape of these new initiatives. Instead, we offer two models implemented at Purdue, which can be viewed as case studies. These are programs for undergraduate students that broaden the pipeline into statistics, the data sciences, and their applications across the full spectrum of disciplines. For this reason, the next wave of statistics and data scientists strongly depends on having innovative undergraduate programs.

## 2 | THE DATA MINE

The Data Mine consists of approximately 600 undergraduate students from diverse majors who all live together in a residence hall (Hillenbrand Hall). The students, regardless of background, and without prerequisites, take our weekly 1-credit hour active learning (peer-to-peer interactions; hands-on, no lectures) seminar about data science skills. It is taught during a meal in Hillenbrand’s dining court. The topics taught in this seminar were inspired by Nolan and Temple Lang [6], including Python, R, SQL, scraping and parsing, and data visualization. Students work with large, real data sets on a distributed UNIX cluster, whose nodes have 768 GB of RAM, as well as a dedicated 1 PB filesystem. The students quickly learn to think about their own laptops and desktops as simply a means of connecting to the UNIX cluster, rather than as a place to do computation locally.

Small projects are due each week. This assignment structure enables students to feel a sense of palpable achievement. Even on the first day of class, when the students have no background in the data sciences at all, we give them a project that utilizes New York City yellow taxi cab data (<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>), which contains hundreds of

gigabytes of data. We continue this trend of using large, publicly available data, such as data from federal election campaign donations, domestic airline flights, AirBnB, Amazon reviews, grocery store retail, databases (e.g., for baseball statistics), etc. We learned this methodology from Nolan and Temple Lang [7].

Undergraduate students can join The Data Mine at any point in their academic careers. They can stay involved for up to 4 years, with leadership opportunities as the students progress. Advanced versions of the seminar for second year (and beyond) Data Mine students are taught simultaneously with the introductory version so that newer students can learn from more experienced ones.

In addition to the weekly seminar, students learn data science skills in the context of a field they choose. Students are organized into 20 learning communities. Nineteen of these are sponsored by academic departments—from every college in the Purdue university system—and offer coursework; research experiences; seminars; and meetings with visitors, alumni, and industrial professionals. These learning communities are in areas with disparate topics, including:

Actuarial Science; Agriculture; Analyzing Digital Gaming and Culture; Biology; Critical Data Studies; Data Visualization; Discoveries in Chemistry; Earth, Atmospheric, and Planetary Science; Electrical and Computer Engineering; Healthcare Engineering; Human Development and Family Studies; Institutional Assessment, Diversity, and Student Success; Management; Pharmacy/Drug Discovery; Philosophy; Physics; Political Science; Psychology; and Statistics.

The 20th learning community in The Data Mine is a Corporate Partners program with 10 external corporate partners and 150 students. (We are planning for as many 300 students in the Corporate Partners program in 2020–2021.) These companies provide mentoring; access to proprietary algorithms and models; data to be analyzed; and visits, both at Purdue and onsite at the companies. Students work in teams to support each other professionally, technically, and emotionally. They teach each other new tools and create professional presentations together. For a corporate partner, working with a group of students for a full academic year may be an unfamiliar model, but this time frame has myriad benefits. It gives students time to grow in maturity, to fully understand their project, and to have uncertainties and make mistakes. The mentors quickly grow to respect and value the abilities and dedication of the students.

Many faculty and teaching assistant office hours happen in the residence hall, including on Sunday afternoons and evenings, which is when many students seek help. Peer support arises naturally, since the students live together, giving and receiving assistance with each other.

Many of the students in the community have no prior coding experience. Next year, we plan to include even more evening help sessions in the first few weeks of the semester, to bring people up to speed on fundamental concepts, such as variables, vectors, functions, loops, etc. Our goal is to empower the students to think like data scientists, including how to learn new skills, because the field is evolving rapidly. The tools they use today may not be the tools they will use in 5 years or if they switch workplaces or domains of application.

### 3 | HPC SEMINAR

The High-Performance Computing Seminar introduces undergraduates from any major, regardless of previous experience, to advanced topics in HPC clusters, operating systems, and the cluster batch-operating systems, in a two-credit hour class. Topics covered focus on aspects of the design, implementation, and use of HPC systems at the system level. The course integrates parallel computing instruction with different scientific domains (such as weather prediction, astrophysics, and fluid and molecular dynamics), with a combination of lectures from domain science faculty and hands-on labs, where students lead the discussions on the tools and assignment. Students work with the Linux Shell, High-Performance Linpack, installing and using scientific applications in a HPC environment, cluster design, analyzing big data with HPC, Python, R Studio, and JupyterHub. In addition to individual technical projects, students also summarize guest lecturer work and create a final team project summarizing their work for the semester.

The learning outcomes for the class are:

- A strong understanding of scientific workflow (including effective communication).
- Familiarity of building and using scientific applications.
- Basics of parallel computing, such as difference between multinode parallelism and node level parallelism.
- Overview of state-of-art computing architectures (e.g., accelerators).
- Performance characteristics (strong and weak scaling) and their connection with the architecture choices.
- Bottlenecks in HPC (e.g., communication and data movement) and strategies to minimize them

The fall 2019 course included students from a broad range of majors, including: Actuarial Science, Applied Statistics, Computer Engineering, Computer Information and Technology, Computer Science, Data Science, Economics, Electrical Engineering, Exploratory Studies,

Mathematics, Mechanical Engineering, Neurobiology, Physiology, Psychological Sciences, and Statistics.

A lack of scientific subject expertise (e.g., in meteorology) can be a difficulty for students when learning applications to HPC, so instructors may need to consider modifying some projects to allow students to become more familiar with what the output looks like and means, before asking them to start with creating a visualization. Also, many of the students have little to no experience in Linux, so Linux knowledge may need to become a prerequisite; otherwise, significant time should be dedicated to providing this knowledge at the beginning of the semester.

### 4 | CLOSING THOUGHTS

In addition to teaching data science and HPC skills to a large number of students, we would like for these student cohorts to be as diverse as possible. In a 5-year National Science Foundation funded pilot project (2014–2020), which led to the creation of The Data Mine in 2018, more than 50% of the 100 students were female. In the last year of the pilot, we had 70% female student participants, 20% Black females, and one Deaf female participant [3]. Now that we are operating at a much larger scale of 600+ students, we still strongly emphasize a diverse and inclusive environment. In the 2019–2020 academic year, The Data Mine had 35% female participants, and we are working to improve this percentage, as well as our numbers of first-generation and underrepresented minority students, through targeted recruiting and partnerships. In the HPC class in the Spring 2019 semester, 16 out of the 28 students identified as female. In 2018, Purdue sent an all-female team (“the Ada Six,” in honor of Ada Lovelace) to the Supercomputing Conference (undergraduate) Student Cluster Competition (<https://www.studentclustercompetition.us/>). Purdue’s 2019 team had 50% female students. It is imperative that students from all backgrounds should work with the full spectrum of their peers, across disciplinary boundaries, to enrich the field of data science.

With a supportive environment and opportunities to learn data science and HPC skills beginning early in an undergraduate student’s plan of study, we have seen our students perform work on par with students in master’s and Ph.D. programs at internships and conferences. Even first-year students who have had experiences in The Data Mine or the HPC class are receiving paid internship offers. With the training and experience our students are learning, companies no longer need only candidates with graduate degrees. The active learning environment is extremely effective in providing experience in numerous computing tools. It allows the students to complete the courses with a high-quality portfolio of projects.

TEconomy Partners, an economic consulting firm, stated that “Purdue’s Data Mine is an example of a developing world class DSE [data sciences environment] program that is organized around industry engagement and immersive skills-building in data sciences that can serve as a model for other universities.” [8] We believe that our approach of enabling undergraduate students in every field to benefit from data science thinking is an appropriate model for training the next generation of data scientists.

### ACKNOWLEDGMENTS

M.D. Ward’s research is supported by National Science Foundation (NSF) Grant DMS-1246818; the NSF Science & Technology Center for Science of Information Grant CCF-0939370; the Foundation for Food and Agriculture Research (FFAR) Grant 534662; the National Institute of Food and Agriculture (NIFA) of the United States Department of Agriculture (USDA); the Society of Actuaries; and Cummins Inc. The authors also wish to thank their many partners at Purdue who have supported The Data Mine.

### CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

### ORCID

Mark D. Ward  <https://orcid.org/0000-0001-8795-746X>

### REFERENCES

1. D. Donoho, 50 years of data science, *J. Comput. Graph. Stat.* 26 (2017), 745–766.
2. GAISE College Report ASA Revision Committee, *Guidelines for assessment and instruction in statistics education college report*, American Statistical Association, Alexandria, VA 2016, [https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege\\_Full.pdf](https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf).
3. F. Gokalp Yavuz and M. D. Ward, Fostering undergraduate data science, *Am. Stat.* 74 (2018), 8–16. <https://doi.org/10.1080/00031305.2017.1407360>.
4. A. L. Golbeck, I. Olkin, and Y. Gel, *Leadership and women in statistics*, Chapman and Hall/CRC, Boca Raton, FL, 2015.
5. National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *Rising above the gathering storm, revisited: Rapidly approaching category 5*, The National Academies Press, Washington, DC, 2010.
6. D. Nolan and D. Temple Lang, Computing in the statistics curricula, *Am. Stat.* 64 (2010), no. 2, 97–107.
7. D. Nolan and D. Temple Lang, *Data science in R: A case studies approach to computational reasoning and problem solving*, Chapman and Hall/CRC, Boca Raton, FL, 2015.
8. TEconomy Partners LLC, 2020, January: *Artificial intelligence and advanced analytics in Indiana: An initial discussion of industry needs and university capabilities*, available at <https://www.biointellex.com/artificial-intelligence-and-advanced-analytics-in-indiana/>.

**How to cite this article:** Betz M, Gundlach E, Hillery E, Rickus J, Ward MD. The next wave: We will all be data scientists. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2020;1–4. <https://doi.org/10.1002/sam.11476>