The Data Mine: Enabling Data Science Across the Curriculum

Ellen Gundlach and Mark Daniel Ward
The Data Mine, Purdue University, West Lafayette, IN

ABSTRACT
In this article, we describe a large-scale living learning community (LLC) for undergraduate students of any major or background. Our students are united by a desire to learn data science skills and to apply those skills in a specific academic discipline or a corporate partner project. We provide explanations of why an LLC is beneficial; the curriculum (motivated by Nolan and Temple Lang); resources required to coordinate such a community; lessons learned from the first year at a large scale; plans for an assessment and a shared resource repository; and plans for an even more accessible, differentiated learning environment in the future.

KEYWORDS
Data science; Introductory statistics; Learning communities; Massive datasets; Statistical computing; Undergraduate curriculum

1. Introduction
Nolan and Temple Lang's classic article on "Computing in the Statistics Curricula" encourages Statistics educators to help their students become more competent, using modern computational tools (Nolan and Temple Lang 2010). They believe that a course teaching "basic proficiency, problem solving, and familiarity with useful programming environments" should "be a core requirement for all undergraduate majors" (pp. 102–103), emphasizing that students should not only become skilled with the currently available computational tools but also know how to "keep abreast of new technologies as they evolve" (p. 97). Nolan and Temple Lang recommend three key considerations (p. 98):
1. Broaden statistical computing to encompass accessing and integrating large datasets, wrangling data, and producing interesting presentations of data.
2. Deepen computational reasoning and literacy, including vocabulary.
3. Compute with real data on real problems in the practice of statistics.

Nolan and Temple Lang want students to develop “into discerning, critical-thinking active participants of an emerging data-driven society” (p. 100). The National Academies of Science, Engineering, and Medicine (NASEM), in their comprehensive 2018 report, agree with this sentiment. They recommend: “To prepare their graduates for this new data-driven era, academic institutions should encourage the development of a basic understanding of data science in all undergraduates” (Recommendation 2.3, p. 22). They emphasize that students from all backgrounds, majors, and career plans would benefit and should have opportunities to learn data science at all levels. NASEM also emphasizes the multi- and interdisciplinary nature of data science, including its breadth of application.

TEConomy prepared a report on “Artificial Intelligence and Advanced Analytics in Indiana” (2020). They assess industry needs and university capabilities. TEConomy emphasizes that the changing landscape… may also require adding Digital literacy to the existing foundation [of education]. Digital technology and data pervade modern economic and societal activity and are at the core of most expanding job markets. A key subcomponent of this skill set involves Data Analytics—providing capacity to understand, process, manage and use sets of digital information (p. 7).

The national job market for people with Data Science and Analytics skills is rapidly expanding, with not enough qualified workers to fill the positions (Columbus 2017). Whether you are a farmer trying to figure out the best time to plant your corn, a research scientist who wants to use artificial intelligence to garner insight from many hours of video data, an engineer developing a self-driving car, or a consumer who benefits from innovations like these, data science is becoming a major part of everyone’s daily lives.

The initiative described in this article is a large-scale living learning community (LLC) for undergraduate students from any major. Our students are interested in learning data science skills and unique data science opportunities beyond the classroom. By mixing many undergraduate majors into one supportive residence hall, hundreds of students are integrating data science skills and thinking into their chosen majors. The result is a diverse undergraduate culture in the data sciences, including strong peer-to-peer learning and support. This article will explain why an LLC is a valuable way to teach data science for all, the history of our program, who our students are, the components of the program, the specialty cohorts, assessment, and next steps.
2. Why a Living Learning Community?

Chickering and Gamson (1987) listed seven best practices in undergraduate education, including active learning techniques, peer collaboration, and contact between students and faculty. Kuh (2008) updated the list of high-impact practices to include common intellectual experiences, learning communities, collaborative projects, undergraduate research, and internships. A learning community (LC) is a group of students who explore an area of interest by taking a class or doing other activities together with the guidance of faculty (Kuh 2008; Goodsell Love 2012). There are even learning communities for faculty (Cox and Richlin 2004).

There are several ways to coordinate an LC. LLCs are a model in which students in the LC share one or more floors in a residence hall, in addition to other academic and professional development activities. LLCs are a natural way for faculty and residential life staff to work together in support of student success. Faculty may not have (otherwise) collaborated with residential life. In our experience, resident life staff members have greatly valued genuine partnerships with faculty, and LLCs can be a feasible way to build such student-centered opportunities. The initiative that we describe is multi-faceted and is too comprehensive for one faculty member to build alone. We emphasize the opportunity for teams of faculty and residential life staff to work together on initiatives like this.

LLCs have been rated as especially effective for promoting positive academic outcomes, critical thinking, ability to apply knowledge to new settings, healthier behavioral choices, and connection to the university community. Faculty and staff from several areas of campus need to work closely, to coordinate the LLC according to known best practices (Kurotsuchi Inkelas and Weisman 2003; Brower and Kurotsuchi Inkelas 2010; Wawrzynski and Jessup-Anger 2010). With an LLC, students’ curricular and extra-curricular learning are supported “seamlessly” by peers and faculty (Kurotsuchi Inkelas and Weisman 2003). Characteristics of many of the most effective LLCs include (Brower and Kurotsuchi Inkelas 2010):

- Have a strong partnership between residential life and academics.
- Identify clear academic learning objectives, including at least one for-credit course taught specifically for LLC participants, study space in the residence hall, and additional academic opportunities such as internships and workshops.
- Create opportunities for learning wherever and whenever it occurs, including faculty and staff taking on a variety of mentoring roles, and faculty and T.A.s providing office hours in the residence hall.

LLCs are an effective way to offer a supportive environment for undergraduate students. LLCs provide a safe, welcoming place in which students can learn new skills and apply them to projects in new settings. This is especially impactful in the data sciences, which are just getting integrated into the curriculum, and are new for everybody. Data science is collaborative and team-oriented by its very nature. Having students live together while they explore this new field provides a supportive environment. Students can inspire and aid each other during their immersive work in the data sciences. A student wrote:

Since most of the floor’s residents are in the Data Mine [defined below], we would all ask one another for help on our common assignments. Doing so helped us to break down the doors that we hid behind. Living in this community helped me to form healthy and resilient relationships with those around me, thus making me feel comfortable enough to call Purdue my home.

While the residential component does have added benefits, nonresidential colleges (e.g., commuter or community college) have also implemented learning communities. Instead of having a living component, faculty and staff can collaborate with students in designated communal areas on campus, for example, in an academic building located near student study spaces and faculty offices.

3. Early Versions of the Course and Living Learning Community

Ward attended faculty workshops on “Integrating Computing into the Statistics Curricula” in 2008 and 2009 at University of California, Berkeley, led by Nolan and Temple Lang (https://www.stat.berkeley.edu/~statcur/Workshop2/ and https://www.stat.berkeley.edu/~statcur/CaseStudiesWorkshop/Overview.html). These workshops enabled faculty participants to learn how to teach hands-on data wrangling skills to undergraduate students. Hands-on, in this context, means the students spend most of their time working with the data themselves, not merely watching instructors. The ideas from the workshops are discussed in Nolan and Temple Lang’s 2010 article. In 2009, Ward began teaching a 3-credit hour active-learning data science course, “Introduction to Computing With Data.” This course covered topics including R, SQL, bash shell, XML scraping and parsing, and data visualization. Ward learned from Nolan and Temple Lang that B.S. and M.S. “students who enter the workforce spend much of their efforts retrieving, filtering, and cleaning data and doing initial exploratory data analysis…working with different data technologies and having general programming skills” (Nolan and Temple Lang 2010, p. 99).

During 2014–2020, Ward was the Principal Investigator for the $1.5 million NSF grant called “Sophomore Transitions: Bridges into a Statistics Major and Big Data Research Experiences via Learning Communities.” Each year, 20 sophomores lived together in a residence hall, learned data science skills in the active-learning course (3 credits), took the key courses in probability and mathematical statistics (3 credits each), worked on large, data-driven research problems with a faculty mentor, and participated in a professional development seminar (1 credit per semester). The students received a 12-month research stipend. This experience, called the Statistics Living Learning Community (STAT-LLC), was not limited to Statistics students. The selection process was holistic. The STAT-LLC was open to all undergraduate sophomores, from any major program in the university. Of the 102 sophomores who benefitted from this program, 57 are female, 1 is nonbinary, 6 are Black (all female), 2 are Hispanic, 1 is Deaf, and 1 is Hard of Hearing. All students are U.S. citizens. In the final year, 70% of the cohort was female. The students produced (to date) 160 journal article...
publications and conference presentations and posters. Gokalp Yavuz and Ward (2018) provided more details about the STAT-LLC model.

In 2018–2019, Ward expanded on this model, and began offering a simplified but larger LLC program, with 100 students per year. It featured a new active learning Data Science seminar, offered for 1 credit in the fall and the spring, with the expectation that students would take both semesters of the seminar. This program began as a pilot called “The Data Mine.” It accommodated students of all levels (not just sophomores) who were interested in learning data science in tandem with their majors. After the successful pilot, in 2019–2020 we offered an even larger Data Mine, with over 600 students, and with additional features. We discuss the 2019–2020 model in the remainder of this article.

4. Students

Undergraduate students from all majors, all years, and all levels of experience are invited to participate in The Data Mine. There are no prerequisites, but the student must express an interest in learning new data science skills and live in the residence hall assigned to the program. We encourage students to participate in both the fall and spring semesters. At the start of the Fall 2019 semester, 657 students were in The Data Mine, including 1 Deaf and 1 Blind student. Half of the students in The Data Mine in Fall 2019 were in their first year of college. A demographics comparison between the Fall 2019 undergraduate students in The Data Mine versus all undergraduate students at Purdue University, College of Science undergraduate students, and College of Engineering undergraduate students is shown in Table 1. The goal of The Data Mine is for students to “Understand Data and Make It Mine.” The minimum requirements to be considered a member of The Data Mine are to live in the specified residence hall and take the 1-credit hour Data Mine seminar in the fall and spring semesters. In this section, we describe the coursework and requirements. The majority of the students also take an additional specialty cohort component (typically another 3 credits). Those cohorts will be explained in Section 6. Undergraduate students are able to participate in The Data Mine for as many years as desired, and can start at any point in their academic career.

5. Basic Components of The Data Mine

The 600+ students are divided into five sections, for scheduling purposes. They attend a weekly 1-credit hour Data Mine seminar in the residence hall's large dining court, typically during a meal. The students work together on the analysis of real-world datasets using a high-performance computing cluster. The curriculum for the year is similar to the 15-week semester course for undergraduates outlined in Nolan and Temple Lang (2010): Basic proficiency, problem solving, and familiarity with useful programming environments, including an introduction to R, Python, SQL databases, visualization and graphics, bash shell, regular expressions, Google maps, and scraping and parsing HTML and XML. Although it is not a Computer Science course, the students also learn about functions, data structures, and some concepts about programming. We do not cover statistical methods, beyond basic exploratory data analysis and graphing, in the first-year course. This curriculum combines elements of the Park City Math Institute's Introduction to Data Science I and II (without the modeling or statistical inference), Algorithms and Software Foundations, and Data Curation (De Veaux et al. 2017) courses and is similar to the introductory courses from various schools described in Hardin et al. (2015). The topics cover the outcomes recommended by NASEM (2018, p. 21):

- Combine many existing programs or codes into a “workflow” that will accomplish some important task;
- “Ingest,” “clean,” and then “wrangle” data into reliable and useful forms;
- Think about how a data processing workflow might be affected by data issues;

5.1. Data Mine Seminar Curriculum

The 600+ students are divided into five sections, for scheduling purposes. They attend a weekly 1-credit hour Data Mine seminar in the residence hall's large dining court, typically during a meal. The students work together on the analysis of real-world datasets using a high-performance computing cluster. The curriculum for the year is similar to the 15-week semester course for undergraduates outlined in Nolan and Temple Lang (2010): Basic proficiency, problem solving, and familiarity with useful programming environments, including an introduction to R, Python, SQL databases, visualization and graphics, bash shell, regular expressions, Google maps, and scraping and parsing HTML and XML. Although it is not a Computer Science course, the students also learn about functions, data structures, and some concepts about programming. We do not cover statistical methods, beyond basic exploratory data analysis and graphing, in the first-year course. This curriculum combines elements of the Park City Math Institute's Introduction to Data Science I and II (without the modeling or statistical inference), Algorithms and Software Foundations, and Data Curation (De Veaux et al. 2017) courses and is similar to the introductory courses from various schools described in Hardin et al. (2015). The topics cover the outcomes recommended by NASEM (2018, p. 21):

- Combine many existing programs or codes into a “workflow” that will accomplish some important task;
- “Ingest,” “clean,” and then “wrangle” data into reliable and useful forms;
- Think about how a data processing workflow might be affected by data issues;

Table 1. Comparison of demographics for Fall 2019 Data Mine students versus all undergraduate students at Purdue University, College of Science undergraduate students, and College of Engineering.

|                      | The Data Mine | All undergraduates | College of Science | College of Engineering |
|----------------------|--------------|--------------------|-------------------|----------------------|
| Female               | 35%          | 42%                | 38%               | 25%                  |
| First generation     | 12% (49% female) | 21%             | 16%               | 11%                  |
| International        | 18%          | 14%                | 21%               | 18%                  |
| White, non-Hispanic  | 47%          | 63%                | 51%               | 56%                  |
| Asian American, Pacific Islander | 41% | 9%     | 17%               | 12%                  |
| Hispanic, Latinx     | 5%           | 6%                 | 3%                | 5%                   |
| Multiple ethnicities | 5%           | 4%                 | 3%                | 4%                   |
| African American, Black | 3%       | 3%                 | 2%                | 2%                   |
| Unknown              | 2%           | 2%                 | 2%                | 2%                   |
| Total # students     | 657          | 33,646             | 4565              | 9413                 |
• Question the formulation and establishment of sound analytical methods; and
• Communicate effectively about properties of computer codes, task workflows, databases, and data issues.

As recommended by Chance and Peck (2015), we have established learning outcomes for the seminar:

1. Students will discover data science and professional development opportunities to prepare for a career.
2. Students will explain the difference between research computing and basic personal computing data science capabilities to know which system is appropriate for a data science project.
3. Students will design efficient search strategies to acquire new data science skills.
4. Students will devise the most appropriate data science strategy to answer a research question.
5. Students will apply data science techniques to answer a research question about a big dataset.

The first learning outcome is assessed through the Outside Event reflections (see Section 5.3). Learning outcomes 2–5 are assessed through the weekly seminar projects, described below.

The datasets used are chosen from domains and contexts that are of interest to our broad range of students. We have used data from the American Statistical Association's Data Expo competitions (http://stat-computing.org/dataexpo/) (Nolan and Temple Lang 2010; GAISE 2016), AirBnB, Google Maps, 84.51 grocery store analytics, New York City taxi cab rides, Presidential election campaign contributions, Amazon music reviews, financial loans, Goodreads, movie reviews from Rotten Tomatoes and IMDB, NASA, crowd-sourced reviews, music databases, dialogs, and amusement parks.

One of the goals of our program is to enable students to learn new technologies and skills. Nolan and Temple Lang (2010) claim that

Technology will continually evolve, which is why it is crucial to teach students the art of learning new technologies so that they can understand, evaluate, and compare them as they emerge. More important than the details of the specific technologies, we should teach students to learn how to learn about new technologies on their own, for example, to cull information from on-line documentation, tutorials, and resources and to identify important concepts (p. 100).

We foster an "active learning" environment, that is, our students spend the majority of class time with hands-on work on data science problems, bolstered by faculty, teaching assistants (TA), and peer support. Active learning is a best practice for student learning, according to GAISE (2016). Students are provided with examples to introduce concepts, but they are also encouraged to talk to each other and search the internet for help in working through the projects. This methodology models what they will do in the workplace and also when learning other technologies in the future. Students in The Data Mine all live and have class together, regardless of their number of years of involvement in this program. In this way, students in their first year of participation have the ability to learn from students who have participated in previous years, and more advanced students have the opportunity to explain data science concepts. This peer support is a hallmark of The Data Mine.

At the start of each 50-min weekly class, the instructor gives a five-minute overview of the projects for that week, including any common challenges the students might experience. (There is one project for each level of students; that is, one project for students who are in their first year of participation, one project for students in their second year of participation, etc.) Then the instructor and TAs walk around to help students while they work and talk to each other. Each project has (roughly) three questions about a large dataset. Our projects are submitted using GitHub Classroom (Fiksel et al. 2019). Our projects are posted in our book of examples (https://thedatamine.github.io/the-examples-book/). Students are allowed to drop their two lowest project grades. For additional flexibility, at the end of the semester, we offer several optional (all new) make-up projects, so that students have the opportunity to replace their grades from low scores or missing scores during the semester. Attendance (pre-COVID-19) is required at the weekly seminars, with students allowed to have two “free” seminars for illness or conflicts. Students and TAs pick up their name tags at the beginning of each seminar, and drop off the name tags in a box at the end of class. This allows attendance to be taken accurately, but more importantly, it enables everybody to more readily learn each other's names.

Professional staff for the first year of The Data Mine include a Professor of Statistics who serves as Director, a Managing Director, a Corporate Partners Senior Manager, and an Instruction Specialist/Senior Data Scientist. There are also 47 TAs, a mix of undergraduate and graduate students, from departments all over campus. The TAs hold office hours in one of the residence hall's lounges; they answer questions from students on the online discussion board Piazza (https://piazza.com/); they give feedback to the faculty and staff about the weekly projects (before distributing the projects to students); and they grade the weekly projects. (The students, TAs, faculty, and staff had 3506 total contributions to discussions on Piazza during the 2019–2020 academic year.) The TAs also learn how to become data science educators, which is recommended by NASEM (2018, p. 77). In particular, the TAs work closely with the students and with the faculty mentors to learn the nuances of the data wrangling. This data wrangling constitutes the majority of the work on many projects; the TAs learn what methods work effectively with helping the students learn the subtle aspects of data science, usually for the first time.

5.2. Computing Infrastructure Requirements

The students use a cluster of UNIX machines, which have 512 GB or 768 GB front ends, some with GPUs, and are supported by a 1-petabyte storage array. The students work with large datasets throughout their participation in The Data Mine. For instance, they are introduced to approximately half a terabyte of data on the very first day of the experience, regardless of their background or previous experiences. Students are strongly encouraged to bring laptops, to connect to the computing cluster. This enables them to work on the projects during class and in office hours. Instructors have a few spare laptops available
for loaning to students who are missing their own laptops. It is helpful to have multiple convenient outlets for students to charge their laptops during class.

5.3. **Outside Events and Reflections**

In addition to the weekly projects and seminar attendance, Data Mine students are required to attend four “outside events.” These are special presentations or guest speakers from alumni, visitors, artists, scientists, engineers, and practitioners throughout industry. Our presenters share testimonials about how they use data science or scientific exploration in their work, professional development, diversity initiatives, hackathons, and in many other interesting ways. In the fall semester alone, 49 potential outside events were advertised, usually occurring directly in the residence hall or occasionally (if the event was co-sponsored or had a different sponsor) elsewhere on campus. After attending an event, the students write a one-page reflection that includes what they learned and what new ideas they want to explore, as a result of attending the event. Through reading these reflections, we are able to find new guest speakers who reflect the students’ interests. We can also connect students to opportunities uniquely suited to their professional development and their interests. The students generally describe these events as “eye opening,” for a multitude of reasons. Students are also surprised by the broad possibilities for careers in data science. After hearing from data science practitioners, students often discuss the need to develop “soft skills.” Examples of student reflections include:

- We need communication and storytelling skills in addition to the coding, statistics, and mathematics if we want to truly maximize our potential success as data scientists.
- Too often “data mining” can feel like an intrusive force used by corporations to maximize profits, so seeing the potential for the use of data science for the public good is fascinating and exciting.
- I used to think of data from an antiseptic, corporate perspective, primarily dealing with inanimate objects and numbers. After [this speaker’s] presentation, I’ve begun to see how data science can apply to biology as well as various liberal arts studies that are widely used in politics and effect a more broad public [sic] in the United States and globally. After having attended the presentation, I would certainly like to explore how data science can be used in politics, and especially how data science’s use in biology can then affect politics and alter the rhetoric that shapes policy, examples possibly being in women’s rights, gender, and more.
- Although this is my final writeup for the Data Mine this semester, I definitely don’t plan on making this the final outside event that I attend. I’ve really enjoyed the presentations and speakers that I’ve been able to listen to, and I’d like to thank everyone in the Data Mine that is responsible for organizing these events and scheduling these speakers. It’s such a unique way to learn interesting things outside of the classroom and connect with well-established professionals from many different fields, and it’s just another one of those things that makes me really love all that Purdue has to offer.

Students made connections to new research mentors, new majors, and new internship opportunities through these events. (Lists of current and past outside events can be found on the website for The Data Mine: https://datamine.purdue.edu/ in the “Education” tab.)

6. **Specialty Cohorts**

In addition to the weekly seminars, most of The Data Mine students chose to be members of a smaller LC within The Data Mine. Each one is a special cohort that involves some combination of academic research, coursework, and/or research with a corporate partner. Each academic LC focuses on a specific disciplinary area of study, as described in Section 6.1. The Corporate Partners cohort focuses on the way that data science is applied in a corporate setting, as described in Section 6.2. We encourage such pairings because “Effective application of data science to a domain requires knowledge of that domain. Grounding data science instruction in substantive contextual examples...will help ensure that data scientists develop the capacity to pose and answer questions with data. Reinforcing skills and capacities developed in data science courses in the context of a specific domain will help students see the entire data science process” (NASEM 2018, p. 29).

6.1. **Academic Cohorts**

For students interested in academic research or taking academic courses in their major, 19 academic cohorts are available, including at least one each from the colleges of agriculture, management, education, engineering, health and human sciences, liberal arts, pharmacy, science, and technology. These cohorts involve taking additional course credits. (For more details about these academic cohorts and the analogous coursework, see the website for The Data Mine https://datamine.purdue.edu/ in the “20 Learning Communities” tab.) We have found that these academic cohorts work best for students and their academic advisors if the additional coursework can fit into their major plan of study. Most of these courses have minimal (1–3) additional credit hours. No compensation is offered to the faculty, other than the opportunity to work with these interested students. Over 40 faculty members from these academic cohorts have office hours in the residence hall. Many of them serve as “faculty fellows” who work with the student resident assistants (RAs) to plan special social events in the residence hall or shared meals in the residence hall’s dining court.

Students enjoy seeing how widely applicable data science is. In particular, they repeatedly tell us that they enjoy living with students from the entire breadth of the university, all united by the common goal of learning more about the applications of the data sciences across disciplines. We are especially pleased to include nontraditional-STEM cohorts, keeping in mind that as data science begins to enter conversations in many disciplines, educators and administrators will have to consider the roles of the humanities, social sciences, and arts programs. There are also opportunities for developing programs for students in non-STEM fields, although there are risks that these become “data science-lite” programs that add limited
marketable or intellectual value to students (NASEM 2018, p. 66).

We are able to bring the same data science skills training to students from all disciplines through the standard weekly 1-credit seminar, but then students with varied interests could also see how data science plays a role in their major field of study. The whole Data Mine community benefits from guest speakers on ethics and bias that are hosted by these non-STEM cohorts. A Data Mine student wrote:

For any and every topic field, Data Science meshes well with any Interdisciplinary field. Psychology benefits from Data Science as much as Data Science can benefit from Psychology. For instance, when I first heard of the many cohorts in The Data Mine, I didn’t understand how valuable those cohorts will respond to Data Science, however, by moving beyond that closed mindset made me think of how much powerful change can happen in this world.

### 6.2. Corporate Partners Cohort

The Corporate Partners cohort has eleven external companies and three campus offices, mentoring approximately 150 students on academic year-long internships. These data science projects are similar to capstone projects (Lazar, Reeves, and Franklin 2011) or the Berkeley Explorations in Statistics Research summer workshops (Nolan and Temple Lang 2015), but they are stretched over an academic year. However, unlike a capstone project (which typically happens close to graduation), our students can work on these projects as soon as their first week in college. More advanced students—including those who have interned over the summer with the companies—serve as peer team leaders. Corporate mentors check in with the students weekly, either virtually or by visiting campus. The types of projects that the students work on include anomaly detection, databases, spatial/temporal visualizations, data wrangling, natural language processing, machine learning, image analysis, prediction, model building, UX/UI, survey design, etc. (A list of current corporate partners can be found here: https://datamine.purdue.edu/corporate/. ) Many of our projects have other faculty and/or staff involvement, in which they learn first-hand how to help students learn these skills. In our Corporate Partners projects, the faculty often learn (sometimes for the first time) how their academic data science skills translate into real-world practices in corporate settings.

Setting up the corporate projects in the beginning is quite involved, including legal agreements between the university and the company, agreements on who owns the intellectual property, nondisclosure agreements, and data privacy issues. We rely on the help and guidance of the university’s lawyers, development officers, and sponsored program services staff. Some companies, especially those who have defense-related contracts, require all students and TAs assigned to their team to be United States citizens. For this reason, citizenship documents (e.g., passports) need to be checked by Data Mine staff before data can be distributed. Some companies share data with Purdue, to be analyzed on Purdue’s computing cluster, but other companies provide their own laptops, with VPNs and remote access to the data, on the companies’ own internal computing servers or on cloud-based platforms.

The students take a 3-credit hour research course in the fall and spring to participate. For their grades in this course, in addition to their corporate mentor’s evaluation of their work, the students must complete biweekly and semester progress reports, submit a resume, and record a video of themselves answering 7 job interview questions related to their Data Mine experience (end of fall semester). Our students’ work in the Corporate Partners program culminates in a Symposium at the end of the academic year (their work can be viewed online at https://datamine.purdue.edu/symposium/welcome.html).

To be prepared for this experience, we require our corporate mentors to work with us on a syllabus that describes what is expected from the students, what the deliverables are (including timelines), and when/how to reach the corporate partner with questions. We assign a maximum of 25 students to each corporate partner, usually decomposed into teams of 5–10 students per project. To set up the teams, the students complete an online survey about their major, citizenship, skills, experiences, and interests. The students rank the descriptions of the projects from most interesting to least interesting, but we do not include the names of the companies in the descriptions. Instead, we want the students to think carefully about the work involved, rather than simply name recognition. We try to balance each team with more and less experienced people for the best mentoring and pipeline opportunities. When possible, we try to find a faculty member with domain knowledge on campus to be an additional resource to the students for each project.

Finding a common fall semester meeting time for all the students on a team is challenging, with the challenge beginning again at the start of the spring semester. Teams need to determine the best ways to communicate among themselves and with their mentors. Our residence hall has several rooms that accommodate groups of approximately 25 students, some with projectors, screens, or large monitors, for presentations. Project teams use combinations of Slack (https://slack.com/), GroupMe (https://groupme.com/en-US/), and E-mail for communication, and some like the project management application Trello (https://trello.com/en-US) for assigning tasks to individuals. We find the Meeting OWL (https://www.owllabs.com/meeting-owl) to be a great help in allowing the mentor to see and hear all the team members during the Microsoft Teams (https://www.microsoft.com/en-us/microsoft-365/microsoft-teams/group-chat-software), Skype (https://www.skype.com/en/), WebEx (https://www.webex.com/), or Zoom (https://www.zoom.us/) meetings. When corporate guests visit campus, they sometimes choose to stay in a guest apartment in the residence hall, so that interactions with the students can be frequent and convenient.

The corporate partners program has numerous benefits to students and to the companies. NASEM (2018) characterized industry partnerships with data science programs as “well matched to the needs of the rapidly evolving data science workforce” (p. 67). Such opportunities give students the opportunity to practice both applied technical skills and soft skills within corporate culture, such as the “ability to understand client needs, clear and comprehensive reporting, conflict resolution skills, well-structured technical writing without jargon, and effective
presentation skills” (p. 29). TEConomy, in their report to businesses and universities in the State of Indiana, recommends a more intentional approach to promote engagement between and among our corporations and universities...to attract, grow and retain more of the talent that is trained by our excellent research universities in those skills that are in the highest demand, ensure that our existing workforce is provided opportunities for upskilling with new and expanded continuing education programs, and anchor our corporate community to drive Indiana’s economy and improve our quality of life (TEConomy 2020, pp. A5–A6).

7. How Will We Measure Success?

Since The Data Mine is a relatively new program, formal assessment results are not yet available. There is a lack of validated, published assessments for Data Science Education (DSE), unlike the thoughtful research priorities and resources for Statistics Education (Pearl et al. 2012). In the future, we hope to be able to objectively assess how well students are mastering the learning outcomes beyond the grades they earn on their work (Chance and Peck 2015). A team of our Purdue colleagues are working on a Data Mine assessment about self-efficacy and identity in data science, similar to Hazari et al. (2010) for Physics Education and Godwin (2016) for Engineering Education, that will be piloted with The Data Mine students during the 2020–2021 academic year. It will include both quantitative measurements and interviews.

Lazar, Reeves, and Franklin (2011) assessed the success of their capstone course by considering the data analyses their students performed, external client satisfaction with the students’ work, the enthusiasm of the students, the students’ newly acquired soft skills, and graduate school plans of the students. We do have the students’ work on their Outside Events reflecting, seminar projects, and corporate partners projects to show their progress on the learning outcomes. We could eventually potentially measure student success by internships, student careers that involve data science, student enrollment in graduate school programs, or students completing the university’s Applications in Data Science Certificate. However, our program is still too new for any of those measurements. On the other hand, our visitors and our university leadership repeatedly emphasize the need for this new program. TEConomy (2020) characterizes The Data Mine as a “signature example of industry-facing, immersive talent pipeline program” (p. 55), so we are hoping for good outcomes for our students’ careers. “Purdue’s Data Mine is an example of a developing world class DSE program that is organized around industry engagement and immersive skills-building in data sciences that can serve as a model for other universities” (p. 86).

One measure of success is retention from the fall to the spring semester. Over half (54%) of the students who finished the fall semester in The Data Mine were still participating in the semester, with better retention of the first-year students (58%) than upper level students (44.11%). Students majoring in the colleges of pharmacy, science, exploratory studies (for undecided majors), and management are more likely to continue from fall to spring, as compared to students from other colleges.

To put this retention rate in context, Purdue has 90 LLCs, but the majority (57%) have programming only in the fall semester with others doing less formal programming in the spring (such as monthly meals), and 71% of LLCs accept only first-year students. Therefore, The Data Mine staff was disappointed but not at all surprised to lose some students, especially upper-level students, for the spring semester. In general, LLCs at Purdue tend to have higher retention rates when the course requirements are minimal (1 credit hour, for example) or if the coursework directly ties into a student’s major plan of study. However, this time, The Data Mine 1-credit seminar, the Corporate Partners 3-credit course, and some of the academic cohort classes, do not yet count toward a major. The Data Mine coursework has been recently approved to count toward the Applications in Data Science Certificate. We celebrate students staying in The Data Mine for two, three, or even all four years of their undergraduate studies. We do not require any advanced coursework (in particular, no advanced mathematics or statistics courses) for this continued involvement. Instead, as the projects become more complicated, we trust that the students learn the necessary domain knowledge from our own projects themselves.

Like Lazar, Reeves, and Franklin (2011), we can also measure success through the feedback of our students. In addition to the student testimonials given earlier in this article, one student stated, at the end of the fall semester, that “I wanted to take the opportunity to sincerely thank the Data Mine for these amazing opportunities and events. I have really enjoyed my experience with the Data Mine and I have learned so much about data science, but more importantly, about where my generation can solve the problems of yesterday, today, and tomorrow.” In a non-blinded survey of 84 Corporate Partners cohort students at the end of the spring 2020 semester, 84.5% rated their year in The Data Mine as a 4 or a 5 (on a scale of 1–5), with an average rating of 4.2. When asked for their biggest challenges, students mentioned balancing their workload for The Data Mine with their other coursework, the number of credit hours required, time management, and communication issues with peers. Students do feel comfortable to give critical and constructive feedback to The Data Mine faculty and staff, and we sincerely value this feedback.

At the end of the spring 2020 semester, the Corporate Partner mentors were asked to provide feedback, and we received ratings from 16. On a scale of 1–5, the average rating was 4.25, and nobody rated their experience lower than a 3. All but one of the 2019–2020 Corporate Partners returned in 2020–2021. (The one Corporate Partner that did not return in 2020–2021 involved in emergency medicine, and they have become extremely busy with the COVID-19 situation.) The number of Corporate Partners, projects, and students involved doubled in 2020–2021 due to the success of the program in 2019–2020. Typical challenges experienced by Corporate Partners mentors included needing a better understanding of the students’ backgrounds so that the work can be planned accordingly and wanting clearer roles for the mentor and team. (Due to the Corporate Partner feedback, in 2020–2021 we are requiring Agile training and structure, Trello boards, and twice-weekly team meetings.) Corporate Partners said, “We hope this program will help us build a talent pipeline for future interns and full-time hires.” The most common response from mentors for the best part of
working with The Data Mine was the students. A typical quote expressing this idea is "The students, hands down! They were engaged, curious, excited, it was awesome to get to work with them and build off of their energy!"

8. Next Steps

Because this was the first year of offering The Data Mine at this scale and level of complexity, we have room for improvement. A goal for the future is to increase the participation of female and underrepresented minorities. As emphasized in NASEM (2018, p. 64): "Data science would particularly benefit from broad participation by underrepresented minorities because of the many applications to problems of interest to diverse populations." We have plans to partner with other LLCs for women and underrepresented minorities next year, and to work with a summer high school program to increase diversity in recruiting.

As mentioned by several of the authors in Hardin et al. (2015), it can be a challenge to meet all students at the appropriate level, when their backgrounds are so diverse. Our goal for 2020–2021 is to create a better on-ramp for students with little coding background. Some students have no coding experience whatsoever and need the basics of the coding thought process explained, while others have already been coding for several years. We have a sample project that incorporates questions using R, SQL, Python, and UNIX available in the beginning of the fall semester to help students decide which level of the seminar course is most appropriate for them so that everyone has an opportunity to be challenged appropriately. We are also working on a plan for "alumni" of The Data Mine, who have lived in the residence hall and participated in the program for at least a year, but want to move into off-campus housing. In 2020–2021, we will switch to using almost all undergraduate TAs, like the "near-peers" used by Nolan and Temple Lang in their summer workshops (2015), yielding more leadership opportunities for our students. Most of these undergraduate TAs and corporate partners team leaders will be current or former Data Mine students.

In 2019–2020, our projects were submitted using GitHub Classroom. In 2020–2021, our students are submitting their projects in Gradescope (https://www.gradescope.com/). This switch was made because Gradescope streamlines the grading and regrading process while interacting well with our course management system, Brightspace. In 2019–2020, we allowed our students to drop their two lowest project grades, and we had a series of make-up projects at the end of the semester, for additional flexibility. In 2020–2021, impacted by COVID-19, we are only using the students’ top 10 out of 15 project grades as the project portion of their overall assessment. In 2019–2020, we required attendance at our weekly seminars, but during COVID-19, we are being flexible about online, synchronous learning, and we are recording all seminars, so that (for instance) students can watch the seminars again and again if needed, after they occur, to digest the material at their convenience.

Our Corporate Partners program was our biggest success from the first year. Due to popular demand from the companies and the students, our Corporate Partners cohort will expand from 150 to approximately 300 students, and will have several new corporate partnerships for the 2020–2021 academic year. We did learn from the inaugural year that students who have not had project-based learning experiences can use more guidance on project and time management, and they also want more social connections with their teammates and corporate mentors. Several of the teams were able to visit the corporate headquarters during the year to meet other data science professionals and tour the facilities. Most of the students who had this opportunity rated it as the highlight of the year. Some mentors who visited campus also planned for relaxed social activities with the students, such as pizza and bowling outings, and students appreciated these and found the mentors less intimidating. We believe that corporate-academic partnerships will be an important part of the future of undergraduate data science education.

Acknowledgments

The authors thank the valuable contributions of fellow staff members Maggie Betz and Kevin Amstutz, as well as many partners at Purdue who have supported the creation of The Data Mine. We thank Carl Krieger and Jonathan Manz for suggesting several of the literature references in Section 2.

Funding

M.D. Ward’s research is supported by National Science Foundation (NSF) grants DMS-1246818, CCF-0939370, and OAC-2005632, by the Foundation for Food and Agriculture Research (FFAR) grant 534662, by the National Institute of Food and Agriculture (NIFA) grants 2019-67032-29077 and 89591, by the Society of Actuaries grant 19111857, and by Cummins Inc., through the Indiana Digital Crossroads 2020 Pilot.

ORCID

Ellen Gundlach http://orcid.org/0000-0002-5242-4441
Mark Daniel Ward http://orcid.org/0000-0001-8795-746X

References

Brower, A. M., and Kurotsuchi Inkelas, K. (2010), "Living-Learning Programs: One High-Impact Educational Practice We Now Know a Lot About," Liberal Education, 96, 36–43.
Chance, B., and Peck, R. (2015), "From Curriculum Guidelines to Learning Outcomes: Assessment at the Program Level," The American Statistician, 69, 409–416, DOI: 10.1080/00031305.2015.1077730.
Chickering, A. W., and Gamson, Z. F. (1987), "Seven Principles for Good Practice in Undergraduate Education," American Association of Higher Education Bulletin, 39, 3–7.
Columbus, L. (2017), "IBM Predicts Demand for Data Scientists Will Soar 28% by 2020," Forbes, available at https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6cc1f2bf7e3b.
Cox, M. D., and Richlin, L., eds. (2004), Building Faculty Learning Communities: New Directions for Teaching and Learning (No. 97), San Francisco, CA: Jossey-Bass.
De Veaux, R. D., Agarwal, M., Avrett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruvuliamala, N., Uhlig, P. X., Washington, T. M., Wesley, C. L., White, D., and Ye, P. (2017), "Curriculum Guidelines for Undergraduate Programs in Data Science," Annual Review of Statistics and Its Application, 4, 15–30.
Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), "Using GitHub Classroom to Teach Statistics," Journal of Statistics Education, 27, 110–119, DOI: 10.1080/10691898.2019.1617089.
GAISE College Report ASA Revision Committee (2016), "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," available at http://www.amstat.org/education/gaise.

Gokalp Yavuz, F., and Ward, M. D. (2018), "Fostering Undergraduate Data Science," The American Statistician, 74, 8–16, DOI: 10.1080/00031305.2017.1407360.

Goodsell Love, A. (2012), "The Growth and Current State of Learning Communities in Higher Education," New Directions for Teaching and Learning, 2012, 5–18.

Godwin, A. (2016), "The Development of a Measure of Engineering Identity," in ASEEE’s 123rd Annual Conference and Exposition, New Orleans, LA, June 26–29.

Hardin, J., Hoerl, R., Horton, N., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., and Ward, M. D. (2015), "Data Science in Statistics Curricula: Preparing Students to ‘Think with Data,’" The American Statistician, 69, 343–353, DOI: 10.1080/00031305.2015.1077729.

Hazari, Z., Sonnert, G., Sadler, P. M., and Shanahan, M.-C. (2010), "Connecting High School Physics Experiences, Outcome Expectations, Physics Identity, and Physics Career Choice: A Gender Study," Journal of Research in Science Teaching, 47, 978–1003.

Kah, G. (2008), "High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter," Association of American Colleges & Universities, available at https://www.aacu.org/leap/hips.

Kurotsuchi Inkelas, K., and Weisman, J. L. (2003), "Different by Design: An Examination of Student Outcomes Among Participants in Three Types of Living-Learning Programs," Journal of College Student Development, 44, 335–368.

Lazar, N. A., Reeves, J., and Franklin, C. (2011), "A Capstone Course for Undergraduate Statistics Majors," The American Statistician, 65, 183–189, DOI: 10.1198/tast.2011.10240.

National Academies of Sciences, Engineering, and Medicine (NASEM) (2018), Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press.

Nolan, D., and Temple Lang, D. (2010), "Computing in the Statistics Curricula," The American Statistician, 64, 97–107.

TEConomy Partners LLC for BioCrossroads (2020), "Artificial Intelligence and Advanced Analytics in Indiana: An Initial Discussion of Industry Needs and University Capabilities," available at https://biocrossroads.com/artificial-intelligence-and-advanced-analytics-in-indiana/. Also presented on November 12, 2019, at the Integrative Data Science Initiative Summit, Purdue University, West Lafayette, IN, available at https://www.teconomypartners.com/.

Wawrzynski, M. R., and Jessup-Anger, J. E. (2010), "From Expectations to Experiences: Using a Structural Typology to Understand First-Year Student Outcomes in Academically Based Living-Learning Communities," Journal of College Student Development, 51, 201–217.