# Undergraduate Data Science and Diversity at Purdue University

Elizabett Hillery
Purdue University
eahillery@purdue.edu

Mark Daniel Ward,PhD
Purdue University
mdw@purdue.edu

Jenna Rickus, PhD
Purdue University
rickus@purdue.edu

Alex Younts
Purdue University
ay@purdue.edu

Preston Smith
Purdue University
psmith@purdue.edu

Eric Adams
Purdue University
ewa@purdue.edu

## ABSTRACT

The vision of Purdue University's Integrative Data Science Initiative is to be at the forefront of advancing data science-enabled research and education. Tightly coupling theory, discovery, and applications, while providing students with an integrated data science-fluent campus ecosystem, this initiative is designed for college graduates at Purdue University to have Big Data experiences in academics. Faculty and staff of the university assist students through curriculum, research, residential life and professional development.

In this paper, we present the framework for this initiative, which includes outlining course offerings, describing the residential living communities and workforce development on campus. We will finish by describing our initial successes.

## KEYWORDS

Workforce Development, High Performance Computing Education, Multi-disciplinary Education, Undergraduate Education

**ACM Reference Format:**
Elizabett Hillery, Mark Daniel Ward,PhD, Jenna Rickus, PhD, Alex Younts, Preston Smith, and Eric Adams. 2018. Undergraduate Data Science and Diversity at Purdue University. In *Proceedings of PEARC '19: PEARC Conference (PEARC '19).* ACM, New York, NY, USA, 6 pages.

## 1 INTRODUCTION

Purdue University has launched an ambitious Integrative Data Science Initiative (IDSI), which focuses on advancing the frontiers of research and the application of data science to pressing, socially relevant issues coupled with new campus-wide, transformational data science education initiative. The initiative is designed to build on and advance Purdue University's existing strengths and position the university as a leader at the forefront of advancing data science-enabled research and education.

In addition to the IDSI, several earlier efforts have been running which endeavored to teach undergraduate the skills they require to be productive members of their field which rely on computational science. These efforts and how they are aligning along with the IDSI are described below.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*PEARC '19, July 28-August 1 2019, Chicago, IL*
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

DOI: 10.1145/3332186.3332202

## 2 RESIDENTIAL LIFE

Purdue University formed a Living and Learning Community around data science, the Statistics Living-Learning Community or (STAT-LLC), and has become a premier example in the nation of combining college living with college learning. Students in the Learning Community are physically roomed together in university housing. This cohort of students also share portions of their course schedules with each other. Unlike other Learning Community's at Purdue, the Data Mine as it has become known as, brings data science seminars directly into the residence hall. Faculty teaching and mentoring the students have office space in the residence hall where they hold office hours and have space to work on projects together.

Bringing these elements together allow students to work cooperatively and collectively. This accelerates learning the core concepts we want to convey to students and supports them as they move to do undergraduate research.

### 2.1 Purdue Statistics Living-Learning Community - Data Mine Living and Learning Community

**Figure 1: Data Mine Living Community, Spring 2019**

Seminars are hosted in Hillenbrand dining food court. The space was fitted for presentations in a familiar environment and provides a large space for students to collaborate. Each Learning Community is 20 students and the seminars host up to 100 students. These Communities work through topics and activities on R, Python, using Jupyter notebooks, and introductory High Performance Computing(HPC) tools. These HPC tools include the Linux shell, batch

**Which gender(s) if any do you identify with?**

| Response | Total |
|---|---|
| Female | 57 |
| Males | 44 |
| Nonbinary | 1 |

**Ethnicity**

| Response | Total |
|---|---|
| Black - All Female | 6 |
| Hispanic | 2 |
| Deaf | 1 |
| Hard of Hearing | 1 |

**Table 1: Table of Demographics**

systems, database technologies, data parsing, and data visualization.

These seminars allow students to practice with their field's most modern tools. They work with real-world data provided by companies and other sources to extract real answers. The capability of working with "full strength" tools allows students to graduate from their undergraduate careers with defined focused and a known path forward.

## 2.2 Impact to date

The (STAT-LLC) program started in 2014 with a vision to revamp data analysis for undergraduates. The program has recruited 20 Sophomores each year since 2014. These sophomores sign up for a 12-month experience from August to August in the community. This first Learning Community has seen 102 students graduate and move forward in their academics lives.

The key selection of sophomores was deliberate to combat the "sophomore slump". We recognized this group of students as the most likely beneficiaries of additional mentorship.

The 80 students from the first 4 cohorts of the STAT-LLC, with their research mentors, so far have 135 journal papers published either through conference presentations or posters. This tremendous research output is far beyond what was expected for the students and their research mentors.

Among the 20 sophomores from the first cohort of the STAT-LLC, 10 are pursing graduate degree, including 2 while working; 9 are working and 2 of these 9 are planning for graduate school, and 1 is completing their studies.

Beside simply the number of students successfully retained and graduated through the program, a primary focus has been on diversity. Diversity comes in many forms and diversity in demographics is certainly found in the Data Mine. Through direct outreach and intentional action to seek out diverse students, the STAT-LLC has expanded to include a positive representation of the whole student body as seen in table 1. Additionally, during the 2019-2020 academic year, we expect 70% of the students in this data science program to be female.

| Learning Community | Students Enrolled |
|---|---|
| Actuarial Science | 50 |
| Agriculture | 27 |
| Biological Science | 27 |
| Chemistry | 27 |
| 6 Corporate Partners | 150 |
| Critical Data Studies | 27 |
| Data Visualization | 27 |
| Earth, Atmospheric, and Planetary Science | 23 |
| Human Development and Family Studies | 23 |
| Institutional Assessment, Diversity and Student Success | 23 |
| Krannert School of Management | 50 |
| Nursing | 27 |
| Pharmacy | 23 |
| Philosophy | 23 |
| Physics | 27 |
| Psychology | 27 |
| Regenstrief Center for Healthcare Engineering | 27 |
| Statistics | 23 |
| Vertically Integrated Projects | 50 |
| Writing and the Data Sciences | 23 |

**Table 2: Data Mine Student enrollment for 2019-20.**

## 2.3 Future Work

From 2014's initial wild success focusing on statistics majors and expansion of that program, in the fall of 2019, additional Data Mine Learning Communities will be formed. The current enrollment stands at 725 undergraduate students. These students will not come just from statistics or fundamental sciences but from disciplines all around the university. Table 2 lists the communities and the current enrollment numbers.

These Learning Communities will help to make a large university feel smaller and more inviting to a diverse set of students. These students will feel more supported to branch out and explore the world of research.

The same ingredients that made the STAT-LLC successful (Learning Communities colocated in the residence halls, seminars focusing on demonstrated skills, and working with real world data) will be put into these new Data Mine learning communities.

The seminars will host students from a mix of learning communities so that teams of students can develop multi-disciplinary expertise and incorporate data science tools and methodologies into both their coursework and their research.

The students will get hands on experience in:

- Training and research in domain-specific areas
- Open Source software
- First-hand experience analyzing big, complex data
- Statistics and analytics
- Conference talks, posters and publications
- Teamwork and leadership

## 2.4 Shared Resources

A core component developed in parallel with the STAT-LLC and now the Data Mine is the infrastructure necessary to provide hands on experience with data sets both big and large and from the real world. With the dramatic scaling up of the Data Mine initiative, some discussion of the investment in computing resources is important.

Supporting just the STAT-LLC was a big undertaking in 2014. These students require interactive R sessions and access to all the standard data analysis tools. Additionally, as real world data was introduced and was adopted by the students into real research projects, the computing resources needed to grow. To respond to this need, the STAT-LLC along with Purdue University's Research Computing Center set out to design and build the Scholar cluster.

The Scholar cluster consists of 8 frontends running interactive sessions and each frontend includes at least 512GB of memory. The cluster also includes a full complement of modern supercomputing technologies: petabyte-scale parallel file system for scratch, Hadoop for data warehousing, Infiniband for MPI codes, a batch system for non-interactive work, and dedicated compute nodes ([1]). Research Computing runs this cluster identically to the rest of the Purdue supercomputers except for one very important difference. Instead of user authorizations based on research lab membership, faculty simply enter their course number into the Center's web portal to have their course roster automatically populated onto Scholar. The cluster and users authorized for the cluster are based automatically on the academic calendar providing for automatic clean-up of past courses([3]).

An amazing result has been that the Scholar cluster has seen dramatic usage growth (from a single course to 51 courses registered and online for the Spring of 2019 academic semester). The STAT-LLC kick-started the use of HPC in the classrooms for Purdue.

As the Data Mine expands to 725 students and disciplines from across the university, Scholar has to grow. There will be potentially 100 students working at any given time together. This means open R sessions, Jupyter notebooks churning away, and graphical desktops running. This type of resource alone for the STAT-LLC student population was a big undertaking and now it will become a premier resource provided by Research Computing for all coursework that requires it.

The Scholar cluster and the close partnership between faculty and Research Computing were a critical component to develop. Purdue University now routinely powers not only classes but also outreach activities like workshops, competitions, hackathons, and extracurricular student activities. All year round, students are able to learn about advanced computing and practice not only for class but to solve real world challenges which is a critical feedback loop powered by cyberinfrastructure ([1]).

# 3 DELIVERING HIGH PERFORMANCE COMPUTING TO STUDENTS

The Purdue University Research Computing department has delivered advanced scientific computing resources to campus since 1968. From the beginning with the early supercomputers on campus to today, outreach and education have been core to the mission of the department.

## 3.1 Student Cluster Competition

*3.1.1 History of Purdue's Competition Involvement.* Purdue University took a team to the inaugural Cluster Challenge event in Reno, Nevada, at the Supercomputing 2007 conference ([5]). Since that first whirlwind of a competition, Purdue and Research Computing have been major advocates of the competition as a vehicle to expose undergraduate students to the world of High Performance Computing.

Purdue has fielded over 12 teams at Supercomputing(SC), International Super Computing(ISC), and Asia Supercomputing Cluster(ASC) Competition events. The effort has impacted the lives, educations, and careers of over 70 students directly. Team alumni have gone into HPC related fields, hyperscale businesses, and pursued numerous advanced degrees. As well, multiple staff members have taken key roles during off years to be part of the Cluster Competition committee. For two years, Research Computing staffs member Stephen Harrell was the Committee Chair for the overall Supercomputing Student Cluster Competition and guided the program through its 10th anniversary ([4]).

*3.1.2 Competition Background.* A cluster competition will bring together many teams of six undergraduate students. These teams will each bring a compute cluster of their own to the conference to use to complete the challenges. These clusters are not limited by fictional budgets or rules, but they are constrained to a predefined power budget. This equalizer levels the playing field and removes some barriers to entry for teams. Teams always partner with a hardware vendor to assist them in their mission. At the competition, students put their knowledge of HPC to the test to optimize science workloads during a 48-hour event. The team who completes the most work wins.

*3.1.3 Purdue Team Structure and Organization.* Every Purdue team that has competed worked together in a class structured around teaching and practicing HPC technologies as well as the science behind the applications of the year. These courses focus on the six core team members and alternates to prepare them for the competition. Forming a course provides additional incentives to the students to invest time in the endeavor and acts as a lightning rod to focus the year's activities. These courses focus heavily on cluster design, systems administration, and skills around application building and management. The competition is a work flow optimization problem and these skills are what the courses focus on providing to students.

A potential downside to the cluster competition is that only six team members get to travel for the competition and compete. Alternates are encouraged to apply as student volunteers, publish posters, or otherwise participate in the whole of the Supercomputing Conference.

## 3.2 High Performance Computing Seminar

**Figure 2: HPC Seminar Series Class**

After Supercomputing 2018 and Purdue's first all female team, recruiting efforts had paid off in a large way. The potential candidate pool for the Supercomputing 19 team had grown to almost 30 students and continued to be composed of a majority of female students. The Purdue Research Computing Center had wanted to start running year-around competition preparation classes, as the competition prep course was traditionally held during the Fall semester. With such a large number of interested students, the opportunity seemed perfect to launch a Spring course.

The focus of the new Spring seminar was to keep momentum building and open the doors to students of all skill levels to participate. The concept for the class was to excite students about the power of scientific computing and show them how that power could be applied to their field of study. While the Student Cluster Competition was used a lightning rod to attract students, those who decided not to compete were provided direction on how they could continue their studies in other classes. The course was broken down into three modules. The first module introduced core cluster design concepts including how to interact with the Scholar cluster and the Linux environment. It covered these topics:

- Introduction to Linux
- Historic supercomputer design and current technologies
- Using Purdue compute clusters
- Software environment management using Spack

The second module introduced building applications and focused on weather forecasting with the Weather Research and Forecasting code (WRF). WRF was chosen for its wide applicability to everyday life. Dr. Michael Baldwin took the students on a deep dive through the history of weather forecasting and how weather and computing inter-played between themselves in the formative days. Dr. Baldwin also offered an exciting look at the leading edge of storm forecasting. This module covered these topics:

- History of Weather and Computing
- Building WRF and post-processing tools
- Scaling study of WRF on Scholar

Finally, the third module focused on the computational fluid dynamics application OpenFOAM. Dr. Carlo Scalo introduced the students to a brief overview of fluid dynamics and covered what his lab is studying in the area of aeronautics. Students were challenged to understand enough of the science, software, and supercomputing to simulate a simple system of their own creation. Most students focused on water flows over different shapes while also studying the impact on simulation time to produce results.

**Which gender(s) if any do you identify with?**

| Response | Total |
|---|---|
| Female | 16 |
| Male | 12 |

**Linux experience before attending class**

| Response | Total |
|---|---|
| 1-2 Years | 9 |
| 0-12 Months | 11 |
| None | 5 |

**Table 3: STAT490 Student Survey**

A supplementary module was offered at the same time as the second module for students from computer science and technology backgrounds which covered Linux system administration and cluster management. This material had the students building a virtual compute cluster on Scholar up to and including running a real scientific application. Students who took this grueling track jumped back into the third module to cover OpenFOAM.

A primary goal of this class was to reach beyond the Computer Science students by deliberately engaging with female students, students with disabilities, and student minorities. A secondary diversity goal was to include different majors and skill levels to create a diverse cluster competition team that worked well together. On that regard, our success can be found from survey results in table 3 and 4.

Purdue University will once again throw its hat into the ring to take a student team to the Student Cluster Competition at Supercomputing 19. This team will be one of the most diverse and best prepared teams ever fielded in the many years of our competing in the arena.

## 3.3 The Future: Combining Efforts with the Data Mine

In the fall of 2019, our goal is to offer this material to every student in the Data Mine Living and Learning Community. This will align the Cluster Competition effort and Research Computing's education program goals with the broader University initiatives, and it will bring a closer partnership between the faculty and Research Computing staff. The goal is to produce students with real world HPC skills so they can immediately participate in scientific research no matter what their degree programs may be in.

# 4 PROFESSIONAL DEVELOPMENT

Purdue University's Integrative Data Science Initiative (IDSI) goal is to work with major sectors of the global economy with such industries as aerospace, agriculture, automotive/mobility, defense, energy, financial services, healthcare delivery, information technology, manufacturing, pharmaceuticals, retail, social media, &

| Major | Students in Major |
|---|---|
| Actuarial Science | 2 |
| Applied Statistics | 1 |
| Computer Engineering | 1 |
| Computer Information and Technology | 4 |
| Computer Science | 9 |
| Cybersecurity | 1 |
| Data Science | 1 |
| Economics | 1 |
| Electrical Engineering | 1 |
| Exploratory Studies | 1 |
| *High School Student (Fall FYE) | 1 |
| Mathematics | 5 |
| Mechanical Engineering | 1 |
| Neurobiology | 1 |
| Physiology | 1 |
| Psychological Sciences | 1 |
| Statistics | 3 |

**Table 4: Majors Represented in Spring HPC Class**

telecommunications. The development and rapid application of advanced data science and analytic methods is a key focus area of the initiative.

### 4.1 Where the Data Mine meets the Real World

Research Computing at Purdue has been able to hire undergraduate students through National Science Foundation (NSF) grants to assist with Advanced Cyberinfrastructure and Cybersecurity projects. This collaboration between the student and staff/researcher has been found to be highly beneficial to the student, to research activities at the university, and has served as a foundation for workforce development in Cyberinfrastructure/Cybersecurity. One current NSF grant includes two Cybersecurity undergraduate students assisting to create a framework for handling controlled unclassified information (CUI) data. These students were chosen based on their aptitude and interests in security methodologies and their desire to learn more within these areas. For example, one of the students has a keen interest in social engineering and physical penetration testing and jumped at the chance to become involved with the project. As the grant includes tasks to train current systems/security staff on the handling of CUI data, the students have been assigned with creating a tabletop exercise (TTE) to simulate security threats. This TTE not only includes a spearfishing campaign, but also provides exercises in white hat security (which coincidentally aligned with the students interests and goals). Research Computing at Purdue has been quite effective over the years with our student program; the development and training strategies we have used have enabled multiple students to become successful full-time employees within our department. [2].

### 4.2 Formalizing the Student Program

Research Computing's student program continues to support grants and research partners at Purdue University. In order to curate a well-rounded experience for the student, we have implemented a project assignment and tracking management tool (Asana) and have created a category within it devoted entirely to the student program. Each student has their own task list they are assigned to and are able to quickly check to see their current tasks. When a project calls for a task that a student can assist with, staff/researchers can add those tasks to the individual student project list.

Managing Research Computing's students in this collaborative manner allow us to find work more easily during a project's quiet phase, and then conversely assign more students to a single task when necessary. This cross-training method allows our student resources to be ready and highly available when needed, while providing more experience in all aspects of Cyberinfrastructure/Cybersecurity.

We also encourage our students to look at the overall task list and choose ones that closely align with their research interests. Providing this to our students have enabled them to successfully submit papers and posters to conference/workshops as they are more excited about their tasks and are more than willing and enable to showcase their work to others. We support this 100

We find this approach increases student satisfaction, improves their experience and provides more opportunities for them in the future. Also, this variety of skills and knowledge leads to more creative solutions to problems and directly impacts the scientific community at Purdue University and the research community as a whole. The collaboration with the Data Mine provides opportunities for students who might not have known about Research Computing at Purdue University and allows them to quickly gain experience within. Research Computing at Purdue benefits greatly from the access to a cadre of undergraduate students with advanced knowledge in research computing fundamentals and methodologies.

## 5 CONCLUSION

Whatever path students take after their educations, they will live in a data driven world. Our encompassing set of student programs, from learning communities to exciting opportunities, will provide Purdue students the tools they need to navigate tomorrow. As we look to build more diversified programs, we would like to take a closer look at students from different income levels, languages and first-generation students to see if there is further impact we can have on the HPC community.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. 1738981, 1827184 and 1840043.

M. D. Ward's research is supported by NSF Grant DMS-1246818, and by the NSF Science  Technology Center for Science of Information Grant CCF-0939370.

The authors would like to thank the many students who have persevered through early courses and provided feedback to guide the various programs talked about here to success. Additionally, it takes a whole village to keep these efforts staffed and fed throughout a semester. We thank our colleagues who have passed the mantle of these programs on for all their work.

A personal thanks to Claire Stirm for doing copy editing for this paper.

Finally, we would like to thank the following people for their involvement in the HPC Seminar, without their help, the class would not have been a success. Dr. Carlo Scalo, Dr. Alejandro Stachan,

Dr. Michael Baldwin, Dr. Xiao Zhu, Dr. Matthew Route, Stephen Lien Harrell, Chuck Schwarz, Geoffrey Lentner, Nicole Brewer and Prabhjyot Saluja

## REFERENCES

[1] Michael E Baldwin, Xiao Zhu, Preston M Smith, Stephen Lien Harrell, Robert Skeel, and Amiya Maji. 2016. Scholar: A campus hpc resource to enable computational literacy. In *2016 Workshop on Education for High-Performance Computing (EduHPC)*. IEEE, 25–31.

[2] M. E. Baldwin, X. Zhu, P. M. Smith, S. L. Harrell, R. Skeel, and A. Maji. 2016. Scholar: A Campus HPC Resource to Enable Computational Literacy. In *2016 Workshop on Education for High-Performance Computing (EduHPC)*. 25–31. https://doi.org/10.1109/EduHPC.2016.009

[3] Kevin D Colby, Daniel T Dietz, Preston M Smith, and Donna D Cumberland. 2014. Self-service queue and user management in shared clusters. In *Proceedings of the First International Workshop on HPC User Support Tools*. IEEE Press, 22–31.

[4] Stephen Lien Harrell, Hai Ah Nam, Verónica G Vergara Larrea, Kurt Keville, and Dan Kamalic. 2015. Student cluster competition: a multi-disciplinary undergraduate HPC educational tool. In *Proceedings of the Workshop on Education for High-Performance Computing*. ACM, 4.

[5] Stephen L Harrell, Preston M Smith, Doug Smith, Torsten Hoefler, Anna A Labutina, and Trinity Overmyer. 2011. Methods of creating student cluster competition teams. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*. ACM, 50.