Training Students Concurrently in Data Science and Team Science: Results and Lessons Learned from Multi-institutional Interdisciplinary Student-led Research Teams 2012-2018

Brent Thomas Ladd[1] and Mark Daniel Ward[1]
[1]Purdue University, 610 Purdue Mall, West Lafayette, IN 47907

Abstract
Dedicated training was designed and offered annually to introduce diverse cohorts of students and early-career scientists to first principles and concepts from data analysis, while also working within interdisciplinary teams. Participants completed a pre-workshop online four-week Introduction to R course. The week-long workshop emphasized hands-on tutorials with techniques for data wrangling and visualization including data scraping, parsing, cleaning, and analysis while also fostering interdisciplinary team science. Diverse backgrounds and experience were prioritized during the selection of participants, along with disciplinary interests from the full spectrum of STEM disciplines and beyond. Teams were organized around real-world, data-driven research projects. Students from statistics, math, and computer science domains were matched with students from engineering, life sciences, and liberal arts. Multi-institutional interdisciplinary teams received funds for continuing collaborative research with the goal of co-publishing results. Outcomes demonstrate that participants gain tangible data science skills and knowledge. Further, the interdisciplinary team experiences result in successful long-term student collaborations across institutions and topic domains at the nexus of data science.

Key Words: Collaboration, Data Science, Diversity, Engaged Learning, Interdisciplinary, Multi-institutional, R

1. Introduction

Data science skills and the ability to work effectively in interdisciplinary teams are highly desired and sought in today's workforce. This paper presents key components and outcomes of a focused training combining data science with interdisciplinary student-led research teams during the time period of 2012-2018. This training was developed and organized through the Center for Science of Information (CSoI), a National Science and Technology Center fully funded by the National Science Foundation (NSF grant CCF-0939370. URL: http://soihub.org).

Since its inception in 2010, the CSoI has designed and implemented an *Information Frontiers Learning Initiative* (https://soihub.org/education/overview/) with goals focused on workforce development training of a diverse next-generation science community while creating a science of information curriculum for classroom and online learning. Thus, an annual engaged learning workshop was designed to introduce diverse cohorts of students to data science concepts and techniques while providing positive interdisciplinary research team experiences by fostering team science best practices.

1.1 Workshop Goal and Objectives

The workshop goal is to prepare diverse cohorts of students to engage in interdisciplinary team-based learning and problem solving by using knowledge of data science methods. The workshop objectives and strategies include:

- Bring together diverse cohorts of students annually to work and learn together
- Create an engaged learning environment where interdisciplinary discussions and team work can thrive
- Engage students in hands-on learning experiences in introductory data science methods and tools using real world data

1.2 Workshop Content and Expected Outcomes

The workshop content and expected outcomes were clearly defined and reiterated for potential students and postdocs as follows:

- All topics will be offered in team-oriented projects. The spirit of the workshop is to bring together students and postdocs from multiple fields and universities to lower barriers for understanding the language and approaches across multiple disciplines and data science.
- No training in computer science or expertise in any particular area is needed.
- The intended audience is students at all levels who have not yet delved into a data science experience, but want to begin working in this area.
- Students who complete the workshop will learn several technologies, including skills for data wrangling and data visualization.
- Participants will have a high level of interest to engage in interdisciplinary team work, and willing to bring their own unique expertise to contribute to peer-to-peer learning.
- We will utilize the R platform for data analysis and discuss strategies for reproducible research.
- Participants will learn how to use R to interact with SQL databases, how to scrape and parse XML code, techniques of data visualization, and use of LaTeX.
- A pre-workshop online course will entail four weeks of online tutorials and exercises (4 hr. / wk) introducing the R environment, and provide a foundation of understanding in preparation for the hands-on, in-person workshop, allowing participants to take a deeper dive into data science.
- Post-workshop activities include the opportunity to develop and submit a grant proposal supporting continued research team collaborations.

2. Participant Population, Recruitment, and Diversity

The workshop attracts a cross-section of participants who are interested in learning to use R and data science skills in general, as well as in experiencing interdisciplinary team collaborations. Funding is provided through our NSF grant for students to travel and attend the workshop. Due to the hands-on nature of the workshop and overall facilitation and availability of funds, participants numbers were kept to no more than 30 (in 2017 we held two separate concurrent workshops, one for undergraduates, and one for graduate students and postdocs).

The workshop was advertised widely, and students were recruited from the CSoI membership, as well as through contacts with programs from around the U.S. 149 students participated in trainings from 2012-2018* including advanced undergraduate, graduate, and post-doc levels. *(*the workshop was not offered in 2013 due to hosting the national NASIT summer school)*.

Student diversity in the traditional STEM areas of CSoI is low (e.g. see Ladd & Brown, 2019). NSF charged the CSoI to focus on increasing female participants as the core of our diversity goals. In addition to achieving gender balance, diverse student backgrounds and experience, and institutional breadth were priorities for recruitment. Students are recruited in light of the overall learning objectives and diversity goals of the training.

During the first four years the workshop was offered, only in 2014 did we approach gender balance (A larger percentage of participants were from the life sciences), whereas we were successful in achieving gender balance in 2017 and 2018 (see Figure 1). We believe this was likely due to allowing a higher percentage of undergraduates participating – most of who were female students – and the fact that we reached critical mass with larger numbers of applicants allowed us the freedom to be selective in inviting participants. The overall gender make-up of all workshops combined is 37.6% female and 62.4% male. Following the workshop, the gender ratio of funded project team members (see section 5.3) begins to approach gender balance (Figure 2).

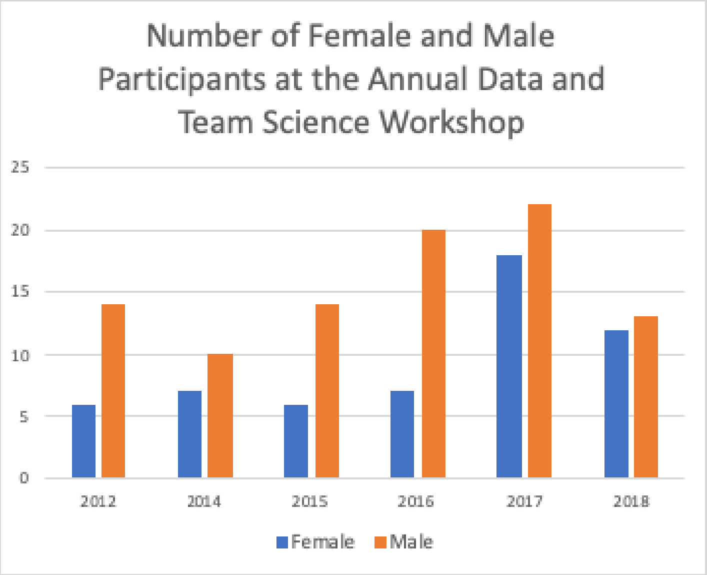


Figure 1. Number of female and male workshop participants per annual workshop (the workshop was not held in 2013)

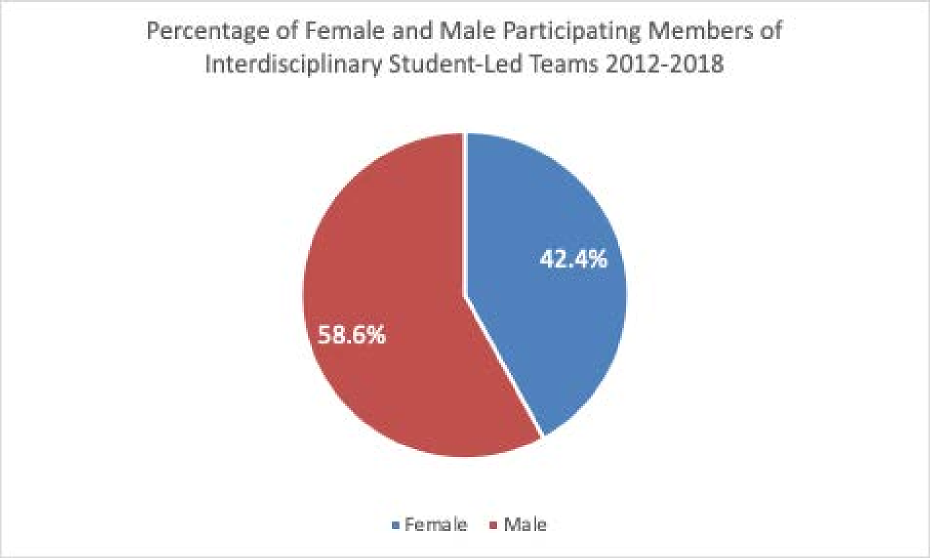


Figure 2. Percentage of female and male participants of the post-workshop funded interdisciplinary science team population from 2012-2018.

Other aspects of diversity important to the innovative functioning of the research teams and participant learning are breadth of domain areas, and an inclusion of a broad spectrum of universities and colleges. Table 1 below in column 1 displays the domain areas that participants represented, and column 2 displays the names of the universities that participants represented. There were 22 distinct departments represented across 25 universities during 2012 - 2018.

3. Evaluation and Survey Method

A survey instrument was designed to capture participant feedback, experience, and perceptions about the knowledge and skills they gained as a result of the training. Experts from Purdue's department of Engineering Education and the Center for Teaching Excellence were consulted in the design of the survey. Although the survey also captured feedback related to workshop logistics and specific instructor ratings for improving the workshop year to year, here we present the results specific only to student perceptions about the data science and interdisciplinary team work skills and knowledge gained.

A four-point Likert scale was employed to elicit participant feedback on a positive to negative scale of Strongly Agree, Agree, Disagree, Strongly Disagree. Open comments following each question also provided opportunity for students to anonymously give additional and detailed feedback. Students were invited by email following the workshop to complete the survey via a weblink. The survey was anonymous and optional to participate. Participants received two reminders via email to complete the optional survey. There was a 57% response rate for the feedback survey (n=85 out of 149 total participants). The survey and results represent the workshops taking place from 2014 through 2018. The pilot year of 2012 participants did not receive this survey, and we did not offer the workshop in 2013 due to hosting a national summer school.

The primary questions asked on the survey related to skills and knowledge gained are:

Survey Questions Related to Skills and Knowledge Gained:
I gained an improved understanding to approaching a data research problem within an interdisciplinary team.
Overall, I gained skills and knowledge I can put to use in my own research and courses.
I gained useful experience with data science tools and methods I can apply to projects and research.
I started some level of professional connections with peers through the workshop.
I improved my ability to explain to others concepts and methods that I use in my own field of study.

Table 1. Disciplines and Universities Represented in the Workshop Trainings 2012 – 2018.

Disciplines Represented (22)		Universities Represented (25)
Agronomy		Boston University
Anthropology		Bryn Mawr College
Behavior and Brain Science		Carnegie Mellon University
Biological Engineering		Eastern Kentucky University
Biology		Howard University
Chemical Engineering		Johns Hopkins University
Civil Engineering		M.I.T.
Computational Biology		Princeton University
Computer Engineering		Purdue University
Computer Science		Rutgers University
Ecological Science and Engineering		Stanford University
Educational Psychology		Southern Illinois University
Electrical and Computer Engineering		Texas A&M
Electrical Engineering		University of Alaska - Fairbanks
Environmental Engineering		University of Arizona
Forestry and Natural Resources		University of Birmingham (UK)
Geology		University of California, Berkeley
Languages		University of California, San Diego
Math		University of Florida
Medical		University of Hawaii at Manoa
Physics		University of Illinois at Urbana-Champaign
Sociology		University of Notre Dame
Statistics		University of Pennsylvania
		University of Pittsburgh
		University of Texas at Dallas

4. Data Science Training

4.1 Data Science Training Environment

There are no pre-requisites for the workshop. The large majority of participants have no specific data science training, though many students have some level of experience in at least one programming language. In 2016 the team realized that it would be advantageous if participants were able to install the R software and learn to navigate the basic syntax and methods of using R. Thus, workshops from 2016, 2017, and 2018 had participants complete pre-workshop tutorials, including a four-week online course, Introduction to R for Data Science (see Ward & Ladd). Students are then engaged for 36 direct training hours during the face-to-face workshop with an intensive series of hands-on examples using R with tools and techniques for data scraping, parsing, cleaning, and analysis during the workshop. Additional training sessions include SQL databases, data visualization tools and techniques, using LaTeX, and techniques for working with large datasets, including machine learning in some years of the workshop.

One of the key aspects of the training and lessons learned is that the instructors are available to consult with the students and teams the entire span of the workshop. We also involve two teaching assistants during the intensive workshop to aid in the learning process and assist individual students on the fly during tutorials. There are many crucial moments of learning and exchange that occur following the morning tutorial trainings, due directly to the fact that instructors and teaching assistants are available 9-5pm each day. We did not purposely schedule evenings, although teams are encouraged to dine together and work on their projects during this non-programmed time. Students report that this is a valuable time of learning and exchange and helps reinforce what was learned during the tutorials.

The student learning process may have been further enhanced by reinforcing lessons learned in data science methods within interdisciplinary project teams. The combination of learning data science techniques and tools within the context of team science projects led to tangible career professional development outcomes as detailed in the results below.

4.2 Data Science Training Outcomes

Nearly all participants responded positively (99%, n=85; Strongly Agree 65%, Agree 34%) that the training allowed them to gain general knowledge and skills they can put to use in their own research and courses (Figure 3).

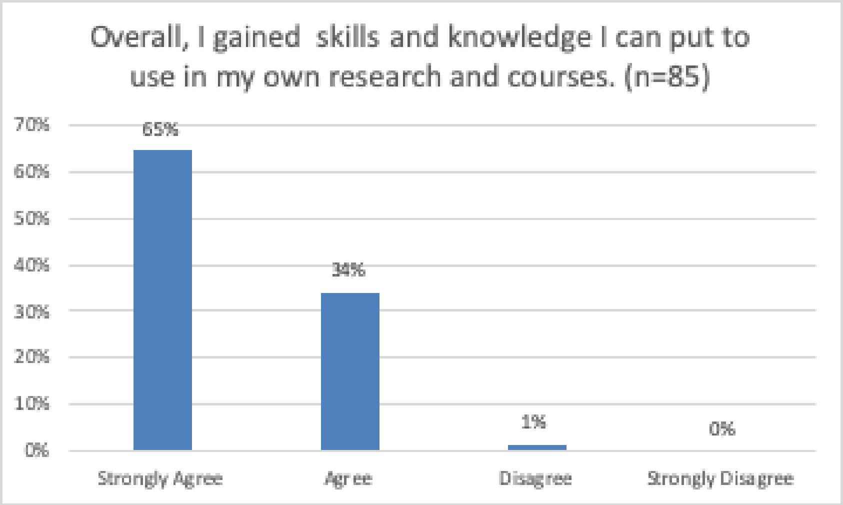


Figure 3. Participant perception of skills and knowledge gained they can put to use in their own research and courses, as a result of the training.

Participants also perceived their training experience positively (93%, n=85; 62% Strongly Agree, Agree 31%) in terms of gaining useful experience specifically with data science tools and methods that they can apply to projects and research (Figure 4).

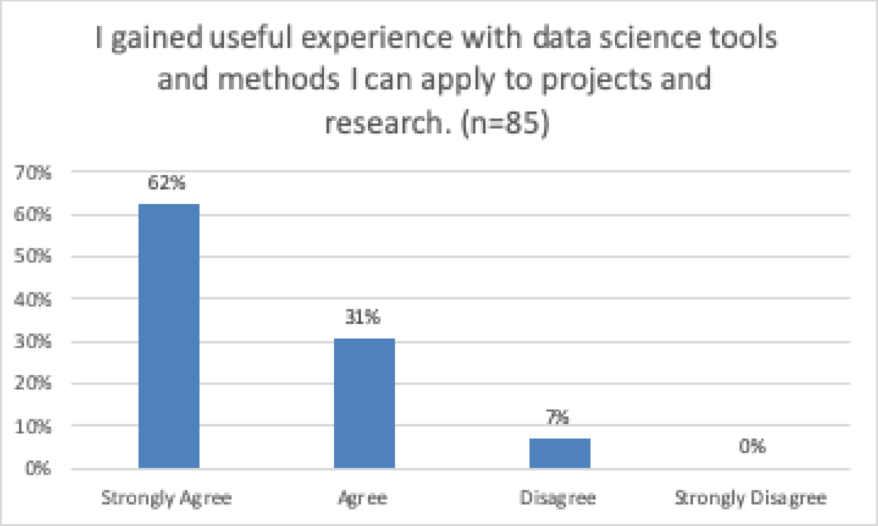


Figure 4. Participant perceptions that they gained useful experience specifically with data science tools and methods they can apply to projects and research.

5. Interdisciplinary Student Teams

5.1 Student Team Learning Environment

Results from the literature on team-based learning and the science of team science were used to inform the organization of the workshop and preparation of participants (Jones et. al 1994, Lightner et. al. 2007, Michaelson et. al. 2009, Ladd, 2019, SciTS Toolkit).

Active graduate student-led research projects are the focus of team collaborations during the workshop. Each annual workshop involves 5-7 teams of 3-6 students in each team. Teams are organized and facilitated using best practices in team science to work on real world research project data that calls for interdisciplinary collaborations. Members of teams are selected to comprise broad interdisciplinary perspectives, with students and postdocs from multiple institutions, gender and racial diversity, as well as a mix of graduate, undergraduate, and postdocs. During the four-week pre-workshop period students read about and are prepared in best practices for successful team science, and team leaders are prepped for organizing their projects and data for team input, while also preparing to present and describe their research and data for an interdisciplinary audience.

During the workshop, teams meet and work on their research projects in the afternoons and evenings following morning training sessions. As discussed above, team members spend a great deal of time together discussing their projects and the various approaches and potential methods. The space created for this experience emphasizes the creative wisdom that each student brings to the process. They are not only allowed, but encouraged, to explore new questions and ways of thinking. The workshop week culminates in team project presentations where each team presents to peers and guests their overarching problem/topic, the methods they have used to analyze data, results gained thus far, and any plans for future collaborations to continue the project. They field questions from the audience which provides additional insights. The presentations are filmed and team members receive access to the videos for their own professional development purposes.

5.2 Workshop Team Outcomes

Specific to the workshop training itself, three survey questions elicited participant perceptions of outcomes related to interdisciplinary team experiences, and resulting skills or knowledge gained.

The first result is the combining of data science problems within the context of interdisciplinary teams. Participants perceived a large positive response (98%, Strongly Agree 69%, Agree 28%) to improving their understanding of how to approach a data problem within an interdisciplinary team (Figure 5).

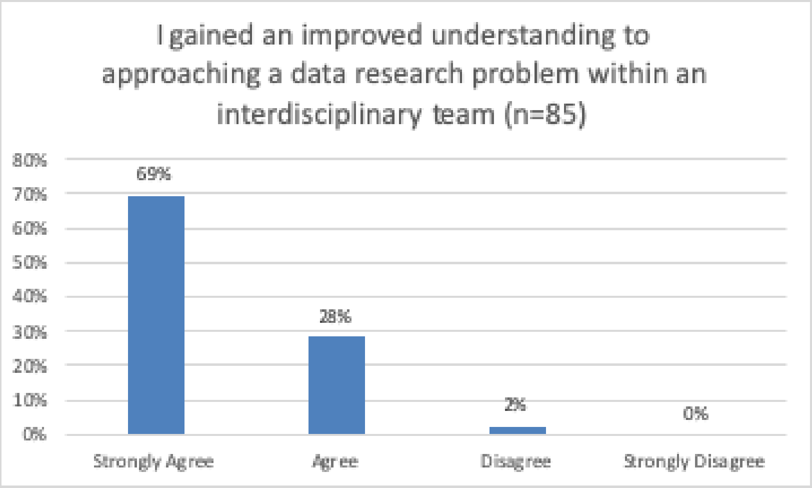


Figure 5. Participant perception of improving their understanding to approaching a data research problem within an interdisciplinary team.

Another skill that lends itself to working in an interdisciplinary team is the ability to explain to others who are not experts in your field the concepts and methods used in your specific field of research and study. The workshop training emphasized these kinds of discussion. It was necessary for students to learn how to bridge their own expertise and understanding with that of the peers on their team. Participants indicated a strong positive response that their ability to explain to others the concepts and methods they used in their own field of study had improved as a result of the workshop training (Figure 6).

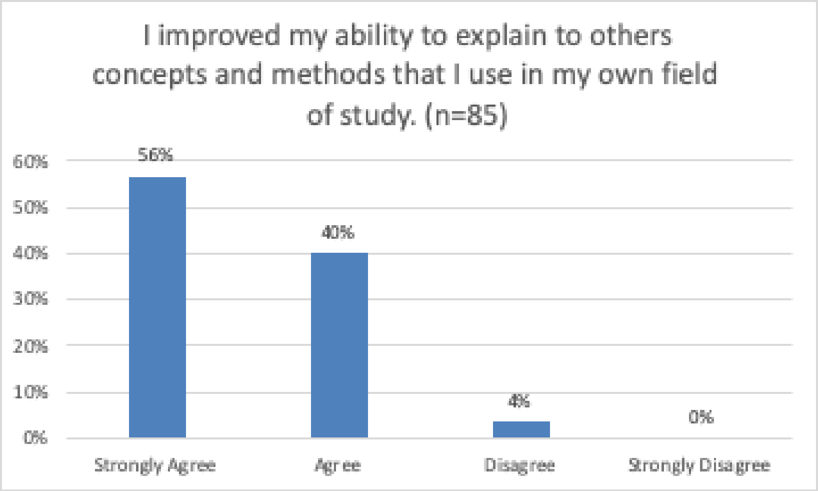


Figure 6. Participant perceptions of how the training influenced their ability to explain concepts and methods used in their field of study to others.

Although the workshop took place during a one-week period of time, the workshop environment was organized such that participants could bridge across disciplines as well as build comradery around a specific research problem, and get to know one another well enough to build professional connections. Participants gave a strongly positive response (97.7%, Strongly Agree 62%, Agree 35%) that they had gained these kinds of professional connections by the end of the workshop (Figure 7). This is also indicated as a result of the grant proposals that participants later wrote and submitted post-workshop (see section 5.3 below).

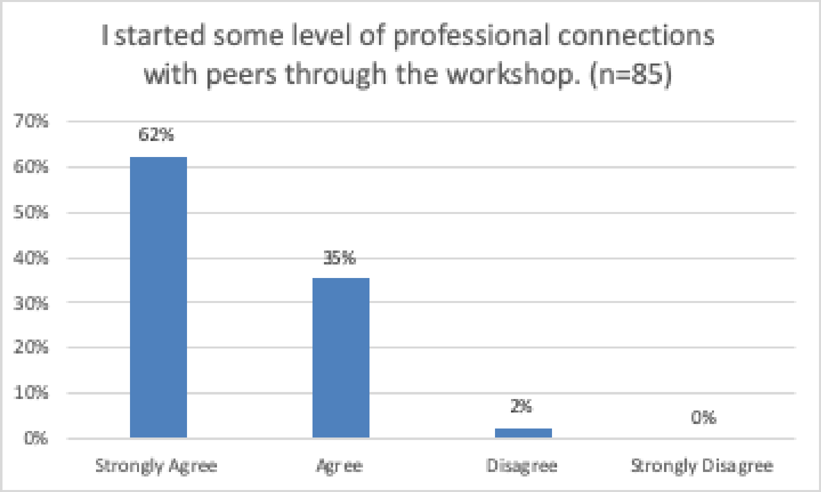


Figure 7. Participant perceptions that they started a level of professional connections with peers as a result of the training.

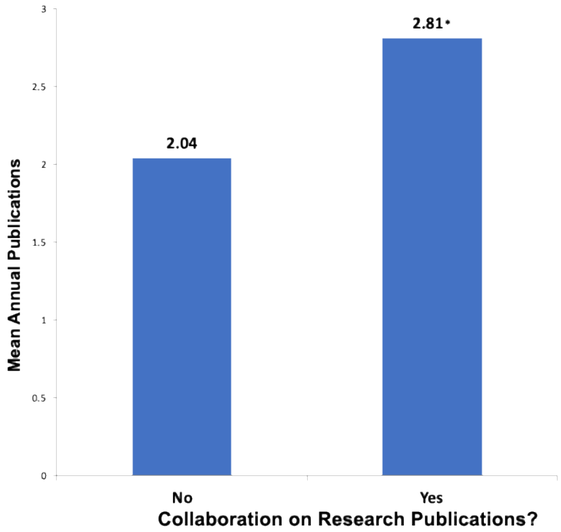
5.3 Post-Workshop Interdisciplinary Team Outcomes

Professional development in writing an NSF style grant is available following the workshop, with funding for teams to continue active collaborations for one to two years (although some teams have collaborated for longer). Student teams work through the process of bridging across disciplines and institutions to develop a mini-NSF style grant proposal. Teams receive feedback to improve their proposals. Depending on CSoI funding available and the quality of projects submitted, between one and four teams annually have been funded to continue research collaborations. Funding amounts are small and used strictly for team travel expenses and has ranged from 4-6K per team. Teams meet monthly using online meeting technology, and at least annually have a face-to-face working session for 2-3 days. Teams are responsible for submitting a six-month progress report, and an annual report. Most teams have co-presented results at one or more conferences within 18 months of working together. Many teams have co-published papers within 24 months of working together.

As of July, 2018, 18 multi-institutional interdisciplinary teams and a total of 66 team members have been funded through the CSoI Information Frontiers program for year-round collaborative research. These teams have collectively produced 25 published papers, and 44 conference posters. Many alumni of this program have remarked it was this interdisciplinary research team experience that gave them a distinct advantage in securing an academic or industry position following their Ph.D. or postdoctoral program.

5.4 Impact on the Science of Information Student Community

One of the influencing results of the training workshop is a ripple effect over time that helped foster a spirit of collaboration being infused across the larger population of the CSoI Science of Information student community membership. A robust General Linear Mixed Model analysis of our graduate students publishing research who collaborated with others in the community vs. those that did not collaborate revealed that collaborating graduate students were significantly more productive in publishing journal papers during the time period of 2012-2018 (2.81 vs. 2.04, p < .001, Figure 4), as well as producing higher numbers of conference posters/presentations (3.06 vs. 2.59, p = .07), with these results due primarily to the factor of collaboration itself. The preliminary results of these collaboration effects and influencing pathways have been reported elsewhere (Ladd, 2018), and will be further detailed in a future paper.



*n=256, F=11.89, p<0.001

Figure 8: Comparison between students who collaborated with others in our CSoI community vs. those that did not collaborate reveals that our collaborating graduate students are more productive in publishing journal papers.

6. Discussion

6.1 Lessons Learned

Specific lessons learned from conducting the training during the time period of 2012-2018 include:

- Training students in data science within the context of interdisciplinary teams working with real world data and problems is a powerful and effective engaged learning model incorporating team-based-learning philosophy.

- Creating a learning environment where students are fully supported and encouraged to ask new and difficult questions, and test risky hypotheses, while bridging across disciplines leads to knowledge and skill attainment and exchange not otherwise possible. The space created for team-based-learning emphasizes the creative wisdom that each student brings to the process.

- Infused diversity at the combined levels of disciplines, institutions, academics, and student demographics coalesce to foster broad insights and exchanges among participants.

- The depth of learning is enhanced and supported when instructors make themselves available to directly assist students and facilitate teams throughout the entire intensive training period. Having teaching assistants available to assist during hands-on tutorials is effective to help individual students on-the-fly when they get stuck, allowing the faculty instructor to focus on the tutorial material.

- Incorporating a pre-workshop online course can be an effective and efficient method for students to learn basic and foundational concepts of installing and using R for data analysis, and allows a much deeper dive into a subsequent hands-on learning during an intensive summer workshop. Including an open discussion platform allows peer-to-peer learning to occur at a distance.

- Small amounts of funding encourage students to apply their knowledge to real world problems across disciplines and institutions and engage in long-term research collaborations.

- We have found that diverse cohorts of advanced undergraduates, graduate students and postdocs are absolutely capable of successful interdisciplinary science co-producing solutions and findings to challenging problems and questions and sharing results through conferences and journal publications.

6.2 Conclusion

Collectively, these results demonstrate that providing focused data science training with full access to instructors during a short period of time (four-week online course, one week in-person workshop) within interdisciplinary teams, combined with small amounts of funding for continued collaborations can lead to tangible data science skills and highly successful student research outcomes.

We think this is especially the case when including participants across institutions and topic domains at the nexus of data science, and organizing an inclusive environment where diverse cohorts work together. It is the use of data science training that brings together a broad spectrum of students who are eager to learn and work together on real world problems that deepens the learning experience.

Concurrent support and training of students in both data science and active team-based learning is a successful engaged learning model for professional development training of the next generation of scientists for interdisciplinary research in industry and academia.

Acknowledgements

We thank the National Science Foundation for their support through NSF grant CCF-0939370, and the Computer Science Department for offering workshop space and technology support.

References

Jones, B., Valdez, G., Nowakowski, J., & Rasmussen, C. 1994. Indicators of Meaningful, Engaged Learning, from. Designing Learning and Technology for Educational Reform. Oak Brook, IL: North Central Regional Educational Laboratory https://www.learner.org/workshops/socialstudies/pdf/session6/6.MeaningfulLearning.pdf

Ladd, B.T. 2016. Best Practices Guide for Formation of Interdisciplinary Science Teams. Available at the CSoI website: https://soihub.org/site/assets/files/6656/basics-of-successful-formation-of-science-teams-1.pdf

Ladd, B.T. 2018. Case Study of Interdisciplinary Student Research Teams: Factors, Outcomes, and Lessons Learned. Science of Team Science National Conference, Galveston, TX, May 23, 2018. https://www.teamsciencetoolkit.cancer.gov/Public/TSResourceTool.aspx?tid=1&rid=4738

Ladd, B.T. and Brown, R.E. 2019. Broader Impacts of the Information Frontiers Integrated Education and Diversity Program. National Alliance for Broader Impacts Summit, May 1, 2019. https://soihub.org/resources/posters/broadening-participation-in-the-science-of-information/

Lightner, S., Bober, M.J. & Willi, C. (2007) Team-Based Activities to Promote Engaged Learning, College Teaching, 55:1, 5-18, DOI: 10.3200/CTCH.55.1.5-18

Michaelson, L., Sweet, M. & Parmalee, D. 2009. Essential Elements of Team-Based Learning, Chapter 1 from Team-Based Learning: Small Group Learning's Next Big Step. New Directions in Teaching and Learning, 7-27.

SciTS Toolkit. Available at the Science of Team Science website: https://www.teamsciencetoolkit.cancer.gov/Public/Home.aspx

Ward, M.D. and Ladd, B.T. 2016. Introduction to R for Data Science online course. FutureLearn Platform: https://www.futurelearn.com/courses/data-science, Science of Information YouTube Channel: https://www.soihub.org/resources/learning-hub-main/course-modules/introduction-to-r-for-data-science/