WORKING PAPER – final version to be submitted to JSM Proceedings 2019

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 and Mark Daniel Ward, Purdue University, Center for Science of Information

Presented at the American Statistical Association Joint Statistical Meetings 2019, Denver, Colorado, July 29.

"I never teach my pupils. I only attempt to provide the conditions in which they can learn."

~Albert Einstein

Summary:
Our engaged learning model of training provides diverse students with immediately useful data science skills, while learning to collaborate in interdisciplinary, multi-institutional research teams that quickly progress to co-producing conference and journal publications.

Keywords: Collaboration, Data Science, Diversity, Engaged Learning, Interdisciplinary, R

Introduction:
This summary paper presents key components and outcomes of an annual training workshop combining data science and team science with diverse cohorts of students 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 Initiative 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.

Broader Impacts Training Goal:
With a goal of training the next generation of science of information scholars, an annual engaged learning summer workshop was designed to introduce diverse cohorts of students to data science techniques while providing positive interdisciplinary research team experiences by fostering team science best practices.

Student Population, Recruitment, and Diversity:
The workshop content and expected outcomes were clearly defined and reiterated for potential students and postdocs. The workshop attracts a cross-section of participants who are interested in learning R and data science skills in general, as well as in experiencing interdisciplinary team collaborations. Funding is provided for students to travel and attend the workshop. 149 students participated in trainings from 2012-2018* including advanced undergraduate, graduate, and post-doc levels with 25 universities and 22 distinct departments represented (*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 asked the CSoI to focus on increasing female participants as the core of our diversity goals. In addition to achieving gender balance, diverse student backgrounds, experience, and institutional breadth have been priorities for recruitment. Students are recruited in light of the overall learning objectives and diversity goals of the training. The first four years the workshop was offered, only in 2014 did we approach gender balance (A large 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 to participate – 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 the workshops is 37.6% female and 62.4% male. Following the workshop, the gender ratio of funded project team members begins to approach balance (Figure 2).

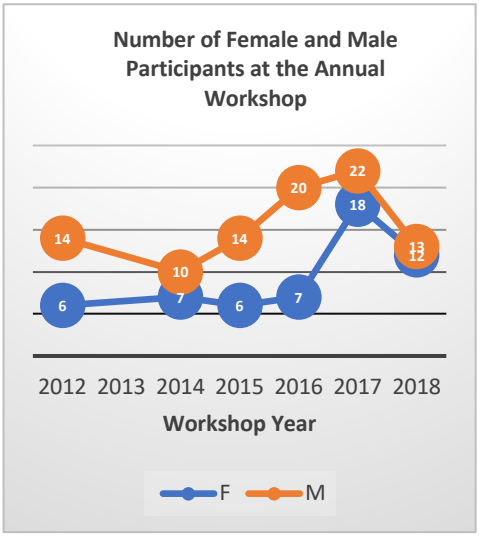


Figure 1. Number of Female and Male Participants per Annual Workshop.

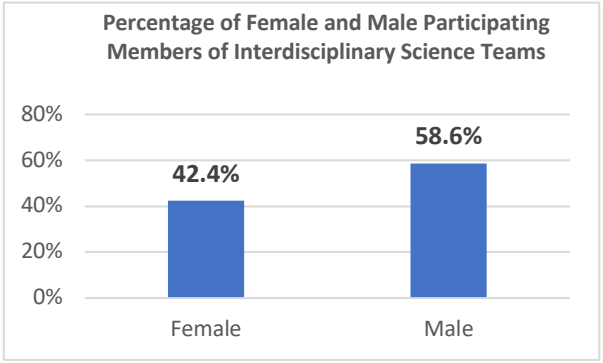


Figure 2. Percentage of Female and Male Participants of the Post-Workshop Funded Interdisciplinary Science Teams Population.

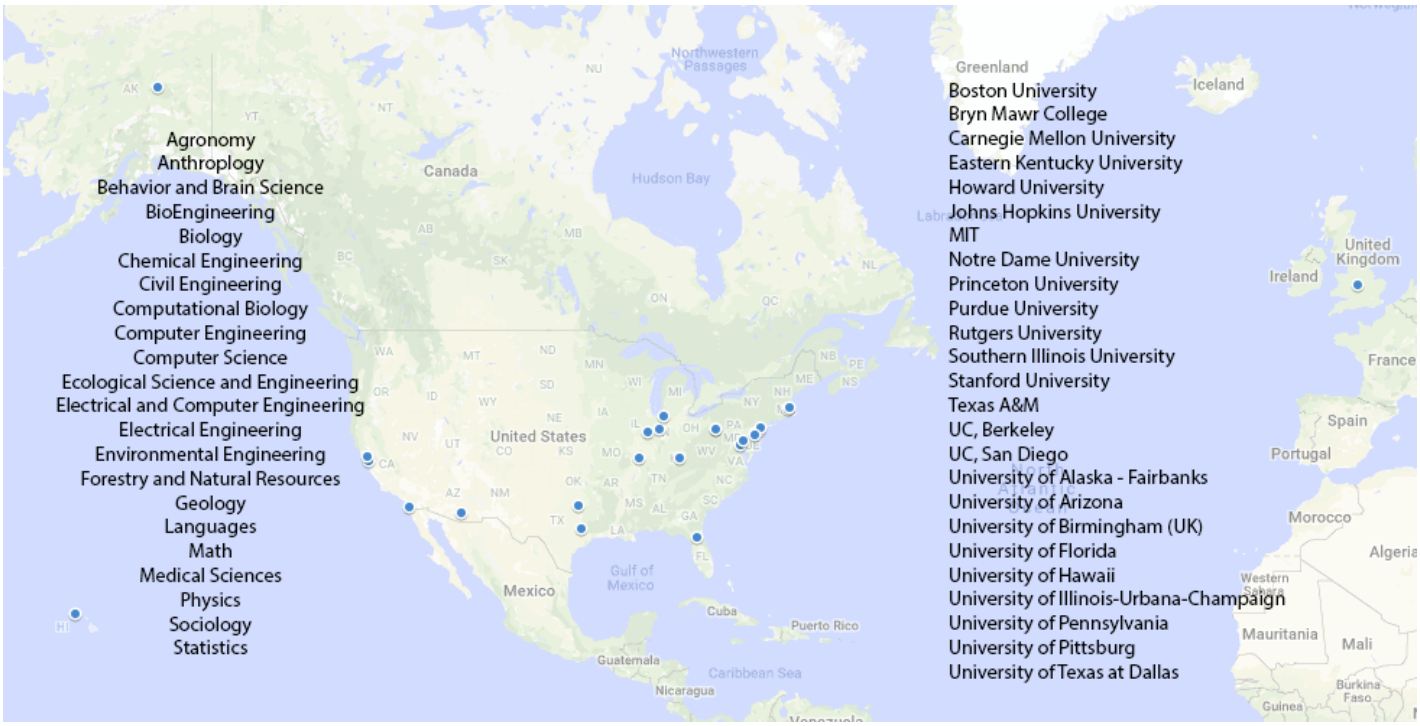


Figure 3. Disciplinary and Institutional Breadth of the Student Research Team Members

Data Science Training and Outcomes:

There were no pre-requisites for the workshop. The large majority of participants have not had specific data science training, though many students have some level of experience in at least one programming language. Participants complete pre-workshop tutorials, including a four-week online course, Introduction to R for Data Science. 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 and 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 Table 1.

Table 1. Mean Ratings of the Professional Development Learning Objective Outcomes from the Annual Data Science & Team Science Research Workshops (2012-2018*) using a 4-point Likert scale.

Student Evaluations Reflected Against Learning Objectives:	Mean/4.0
I received useful feedback to my own research by my interactions with peers and faculty in the workshop	3.81
I gained an improved interdisciplinary understanding to approaching a research problem	3.66
My overall experience of working in a multi-institutional interdisciplinary team during the workshop	3.66
Overall, I learned specific skills I can put to use in my own research/courses	3.63
I started some level of professional connections with peers through the workshop	3.60
I improved my ability to explain my research to others as a result of interactions during the workshop.	3.53

n=85 anonymous responses. 4-point Likert scale 1=poor, 2=fair, 3=good, 4=excellent
*the workshop was not offered in 2013 due to hosting a national level summer school

Interdisciplinary Student Team Training and Outcomes:
Active graduate student-led research projects are the focus of team collaborations. 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.

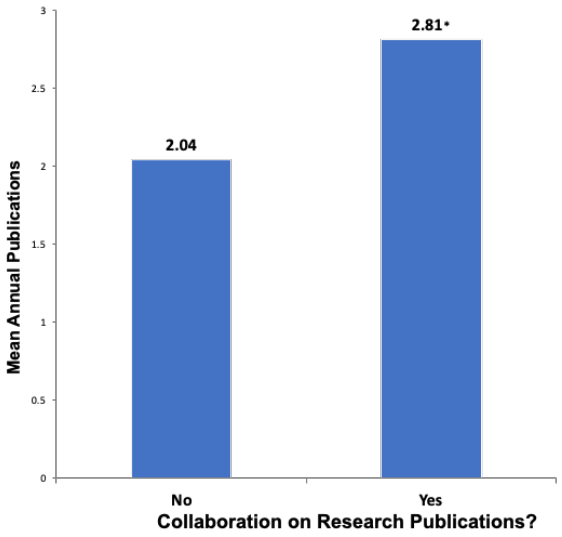
During the workshop, teams meet and work on their research projects in the afternoons and evenings following morning training sessions. As discussed above, team members spend a great deal of time together discussing their projects and the various approaches and potential methods. The space created for 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.

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.

To date, 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.

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 4: 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.

Lessons Learned and Conclusion:

Specific lessons learned 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/questions and sharing results through conferences and journal publications.

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. 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 science is a successful engaged learning model for professional development training of the next generation of scientists for interdisciplinary research in industry and academia.

Acknowledgement:
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.

About the Authors:

Brent Thomas Ladd has served as the Director of Education for Center for Science of Information based at Purdue University since its inception in 2010. He holds a M.S. degree in Ethology, with post-MS research and study in Human Dimensions of Natural Resources. His interdisciplinary efforts in developing research & education projects spans domains of agriculture, engineering, and science. Brent is available to answer questions about this project and assist with innovative education, diversity, and research projects. He can be reached via email: laddb @ purdue.edu , or brentladd1 @ gmail.com, and via his LinkedIn page: https://www.linkedin.com/in/brent-thomas-ladd/.

Mark Daniel Ward is Professor of Statistics, and of Mathematics (Courtesy) at Purdue University and serves as the Associate Director of both Center for Science of Information and of Purdue's Integrative Data Science Initiative. He holds a Ph.D. in Mathematics from Purdue University. His research has focused on probabilistic, combinatorial, and analytical techniques for the analysis of algorithms and data structures. Mark is a Fellow of the Purdue University Teaching Academy and is the inaugural Director of The Data Mine, a unique 800 student living learning community focused on teaching and learning tangible data science skills and knowledge across all domain areas at Purdue. His full CV and contact information are available at his Purdue website: http://www.stat.purdue.edu/~mdw/).

References & Resources:

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. 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. 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/

Redington, L. and Ladd, B.T. 2016. How to Grow Researchers: A Fresh Perspective on Graduate Student Collaboration. Available on the CSoI website: https://soihub.org/resources/articles/how-to-grow-researchers-a-fresh-perspective-on-graduate-student-collaboration/

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

Ward, M.D. Online Course: Introduction to R for Data Science. 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/

CSoI Annual Training Workshops. Available at the CSoI website: https://soihub.org/education/workshops/

Science of Team Science Toolkit: https://www.teamsciencetoolkit.cancer.gov/Public/Home.aspx

Student-led Interdisciplinary Research Teams. Available at the CSoI website: https://soihub.org/education/research-teams/

Indicators of Meaningful, Engaged Learning, from Jones, B., Valdez, G., Nowakowski, J., & Rasmussen, C. (1994). 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

Center for Engaged Learning, Elon University. An International Center for the Study of Engaged Learning: https://www.centerforengagedlearning.org/engaged-learning/

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