Enhancing Undergraduate Education and Curriculum through an Interdisciplinary and Quantitative Initiative to Broaden Participation in Big Data

Patricia Ordóñez Franco, Juan S. Ramírez-Lugo, Humberto Ortiz Zuazaga, Maria-Eglée Pérez Hernández, Luis Raul Pericchi, José E. García Arrarás. University of Puerto Rico Río Piedras, patricia.ordonez@upr.edu, juan.ramirez3@upr.edu, humberto.ortiz@upr.edu, maria.perez34@upr.edu, luis.pericchi@upr.edu jose.garcia36@upr.edu

Abstract*– Representation of Hispanics, especially Hispanic women, is notoriously low in data science programs in higher education and in the tech industry. The engagement of undergraduate students in research, often and early in their path towards degree completion, has been championed as one of the principal reforms necessary to increase the number of capable professionals in STEM. The benefits attributed to undergraduate research experiences have been reported to disproportionately benefit individuals from groups that have been historically underrepresented in STEM. The IDI-BD2K (Increasing Diversity in Interdisciplinary Big Data to Knowledge) Program funded by the NIH at the XXXXXXwas designed to bridge the increasing digital and data divide at the university. The college's population is 98% Hispanic, it is one of the top 20 producers of Hispanic PhDs in Science and Engineering and yet there is no formal data science program. There also exists a gender imbalance in computing at the College of Natural Sciences at the XXXX. Over 60% of the undergraduate students in Biology are women. However, the percentage of women in Computer Science hovers around 15%. The IDI-BD2K was created to address these concerns and increase the participation of Hispanics in interdisciplinary computational and quantitative research in XXXX. The Interdisciplinary and Quantitative Biology Research Experience for Undergraduates (IQ-Bio-REU) summer program forked off from the IDI-BD2K and was created to engage ten (10) underrepresented undergraduate students from the US and its territories in authentic research experiences in emerging fields of biology which integrate quantitative and computational approaches to projects ranging from molecular biosciences to bioinformatics to ecology to bridge the digital and data divide for Hispanics and women in computing. This paper documents the additions to curriculum as a result of the IDI-BD2K, the first summer of the IQ-Bio-REU and highlights the importance of mutually beneficial collaborations with top research institutions to make it possible.

Keywords*— Interdisciplinary Data Science, Computational Biomedical Research, Big Data, Undergraduate Research, Mutually Beneficial Collaborations

I. Introduction

Now more than ever, the social stratification of higher education in the United States and its territories is increasing the digital and data divide along socio-economic and racial lines (https://academicmatters.ca/higher-education-and-growing-inequality/. No where is this more apparent than in the lack of data science programs at majority minority institutions. To bridge the gap, a group of researchers from different disciplines of the University of Puerto Rico Río Piedras joined forces with biomedical data science researchers at top tier research institutions such as the University of Pittsburgh, Harvard University, and University of California Santa Cruz as well as the Center for Brains, Minds and Machinery at MIT and the DataUp Program at Georgia Tech.

Prior to the creation of the IDI-BD2K, the relationship that the XXXXX had with its collaborators was that of producing potential students for the top tier universities, leaving XXXX devoid of its most talented students. When data science arrived, top tier universities were building programs; however, universities in Puerto Rico were still focused on computer science. The economic crisis and hurricanes Irma and María left the university in a hiring freeze. Thus, faculty had to develop a mechanism to train themselves and their students. They found it in the transformation of their collaborations to mutually beneficial ones. Developing such a collaboration requires creating informal learning spaces to develop interuniversity networks among students and faculty.

I. II. Background

According to the National Center for Women in Technology 2019 Scorecard of the women working in the United States in technology in 2017, 63.2% were White,12.9% were African American, 19.7% were Asian/Pacific Islanders, and only 5.4% of them were Latina. Seeing as women occupied 26% of the occupations in Computer and Mathematical Occupations, we estimate Latinas occupy about 1% of them [1]. In the high-tech sector, the numbers for the underrepresented are even more disproportionate [2].

The XXXXXXis a top research institution, categorized as a doctoral university with high research activity by The Carnegie Classification of Higher Institutions [3]. It is also one of the top producers of Hispanic PhDs in Science and Engineering in the US [4], yet it lags behind in computational science. Perhaps as a result of the financial and political crisis and natural disasters it has faced in the last decade. At the College of Natural Sciences at the XXXX, over 60% of the undergraduate students in Biology are women. However, the percentage of women in Computer Science hovers around 15%. The college's population is 98% Hispanic.

II. Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K)

In 2015, an interdisciplinary group of faculty members from Computer Science, Mathematics, and Biology joined forces to create an informal training program for both faculty and students to develop computational biomedical research on campus with the ulterior motive of increasing diversity in computing. These trainings lead to the creating and updating courses in data science, statistical methods, and big data in the departments of Computer Science, Mathematics and Biology

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.581
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

18th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering, Integration, and Alliances for a Sustainable Development" "Hemispheric Cooperation for Competitiveness and Prosperity on a Knowledge-Based Economy", July 27-31, 2020, Virtual Edition.

respectively, as well as to the ongoing development of a data science program. It has also enriched collaborations with doctoral universities in the mainland of Very High Research Activity that have strong computational biology and biomedical data science programs.

The project has led to the creation of or adoption of three new curricular courses: Introduction to Data Science, a new elective course offered in Computer Science in Python; Statistical Methods, a revised course based on the online MOOC by Rafael Irizarry at Harvard University; and Big Data in Biology Seminar, an elective course in Biology. We expect these courses will become the core introductory courses for our future Interdisciplinary Data Science program.

A. *Introduction to Data Science*

Data Science is an interdisciplinary field that requires statistics, computing and domain expertise to solve complex problems by extracting knowledge from large data repositories. This course is an introductory, applied, collaborative programming elective that brings students from different disciplines together, including biology and other natural sciences, statistics, computer science, and other areas, to work on local, culturally relevant projects using real datasets. By the end of this course, students are expected to clean and manipulate data, use relevant data science libraries to analyze and visualize data to derive conclusions. The students will develop critical and statistical thinking skills and become familiar with the tools necessary for collaborative and interdisciplinary data science projects that they may encounter in summer internships and future research experiences.

B. *Statistical Methods*

This is an intermediate course focusing on statistical concepts and programming competences in R for the analysis of Big Data. The course is designed as a "reverse flip", using as a basis the first three courses of the MOOC "Data Analysis in Life Sciences" designed by Prof. Rafael Irizarry (Biostatistics, Harvard University) for HarvardX on http://www.edx.org.
The course reviews some basic concepts on Exploratory Data Analysis and Statistical Inference, and then focuses on Linear Models. Later, the course covers some topics which are relevant in the analysis of large datasets produced by Molecular Biology studies as microarrays, including for example methods for controlling error rates in a multiple testing setting.

C. *Seminar Big Data in Biology*

This course seeks to expose students to the fascinating and important field of the generation, management, visualization and analysis of large complex data. It also addresses the development of human resources and computational capabilities to fulfill the needs associated with big data. The course consists of bringing data scientists from the United States and Europe to give workshops to students of biology in a wide array of topics in biomedical data science.

The final result of these courses was a program that we are still developing for Interdisciplinary Data Science. The original outline is in Table 1. The BBD I has become the Introduction to Data Science Course. BBD2 is the Statistical Methods course and the final Capstone course has not been developed but may be modeled after the Seminar course in Big Data in Biology. The goal is to create a Data Science program that can adapt to the specialization of the student's preferences, so that a student from a natural science could do complete all three courses and be able to program in R and Python, but a Computer Science student with an interest in data science may take advanced programming courses in Machine Learning, Computer Vision or Visualization, and a Mathematics student might focus more on Statistics, Probability, and Machine and Statistical Learning.

Year	Terms	Science majors	Comp. Sci. majors	Math majors
1	S1	QUIM3001	CCOM3030	CCOM3030
	S2	BIOL3101	CCOM3033	CCOM3033
2	S1	MATH3026	CCOM 3034 QUIM3001	CCOM3034
	S2	CCOM3030	BIOL3101 (CCOM4027)	(MATH4031)
Summer		Cohort activity- One week workshop		
3			**BBD 1**	
	S1		MATH5001	MATH5001 QUIM3001
	Winter	Cohort activity		
			BBD 2	
	S2		(CCOM5050)	MATH5002 BIOL3101
Summer		10 week internship at BD2K Centers		
4	S1	**Capstone I**		
	Winter	Cohort activity		
	S2	**Capstone II**		

Table 1: Timeline of courses for students from Science programs, Computer Science and Mathematics.

In Table 1, the black courses represent required courses for the major and the red courses would be the required courses from another major that the student would have to take to acquire a data science certification in the their own discipline. The blue activities would be required of everyone and would be considered the interdisciplinary data science courses.

None of this program or the development of these courses would have been developed without inspiration and collaboration from our partnering BD2K Centers: University of Pittsburgh, Harvard University and University of California Santa Cruz plus the Center for Brains, Minds and Machinery that has invited our students every year to their Quantitative Biology Workshop as part of their winter cohort experience. For more information about this innovative program that uses hackathons and informal learning to create these mutually beneficial collaborations, see our previous publication [ref LACCEI 2019].

III. Interdisciplinary and Quantitative Biology Summer Research Experience for Undergraduates (IQ BIO REU)

These collaborations lead to the creation of an innovative and resourceful summer REU program at XXXXXto bring underrepresented students in STEM from the US and all of XXXXX to our campus and train them to become successful researchers in interdisciplinary and quantitative biological

research. Our collaborations with high research universities allowed us to bring postdoctoral and graduate students from our collaborating institutions to Puerto Rico and train our REU students as well as our summer students and faculty.

Undergraduate research experiences (UREs) have long been considered one of the principal reforms that are necessary to meet the workforce demands of the 21st century [5-9]. UREs have been reported to disproportionately benefit individuals from groups that have been historically underrepresented in science, a growing segment of the population whose enrollment in STEM programs is increasing, but that abandons STEM degrees at a high rate [10–11].

IV. Program Description

The Interdisciplinary and Quantitative Biology Research Experience for Undergraduates (IQ-Bio-REU) was created to engage ten (10) undergraduate students in authentic research experiences in emerging fields of biology which integrate quantitative and computational approaches to projects ranging from molecular biosciences to ecology. Throughout the course of an intensive nine-week period students were immersed in a wide array of high-impact practices including participation in a mentored research project in partnership with faculty members from the College of Natural Sciences at the XXXXX training and practice for fluency in computational skills and data analysis and opportunities for professional development. Ultimately, our program aims to channel more students into careers at the vanguard of science and achieve the goal of promoting participation by historically underrepresented groups within scientific disciplines, thus enhancing diversity within STEM.

A. Participants

The first cohort to participate in IQ-Bio-REU was of 10 students, from which only 2 identified as computer science majors as seen in Table 1. The remaining 8 students identified as majoring in biology (6), environmental sciences (1) and psychology (1). 5 of the participants were rising juniors, 6 of them were female. 7 out of the 10 participants identified as Hispanic/Latinx, 1 as African-American, and 1 as Native American. 2 identified as non-traditional students returning to college after an extended leave from school to work in industry and teach as summarized in Table 2.

Table 2: Demographic Characteristics of 2019 Cohort

Participant Characteristics	Number
Total	10
Non-CS Major	8
Hispanic/Latinx	7
African American	1
Native American	1
Female	6
Rising Junior or earlier	5
No prior research experience	4
Non-traditional returning students	2

B. Software Carpentry

Critical to the success of student projects and program goals is ensuring that all participants gain an adequate level of fluency in computational and quantitative skills. To make certain that all students have the knowledge and competency to progress towards fluency in data analysis and successfully perform tasks related to individual projects, the first week of the proposed activities were devoted to an immersive Software Carpentry workshop covering the full lifecycle of data-driven research. Software Carpentry workshops are explicitly designed to target an audience of learners with little to no prior computational experience [12]. The workshop was structured as five half-day sessions, allowing students time to digest the material, and apply new knowledge in the lab. Training focused on core data skills for efficient, shareable and reproducible research practices. Topics covered in the workshop included the UNIX shell, git and version control, R programming, data analysis and visualization with R. The original workshops were sponsored by our grant and produced one certified Data Carpentry instructor. Thanks to a collaboration with the South Big Data Hub and its DataUp program, we were able to send three more people to Georgia Tech for a Train the Trainer workshop and triple the number of certified Data Carpentry Instructors.

C. Research

Each student was assigned a research project in the laboratory of the participant faculty. Under the guidance of a faculty member, with support from other faculty and members of the laboratory, participating students were fully immersed in all aspects of the experimental process, including experimental design, data collection, troubleshooting, data analysis, working group discussions such as lab meetings and the communication of their results to peers.

D. Replicathon

Interpreting information extracted from massive and complex datasets requires sophisticated statistical methodology as one can easily be fooled by patterns arising by chance or systematic errors that are hard to detect. A workshop presented by members of Rafael Irizarry's laboratory (Harvard BD2K) focused on the development of statistical tools that helped students better interpret their data. In the Replicathon, students were presented with two different conclusions from one dataset and they replicated several statistical analyses to present and defend their conclusions to the judges. The students worked in R and R markdown. This event was open to all students on campus.

E. Machine Learning Workshop

Postdoctoral and graduate students from our collaborators at the Center for Brain, Minds and Machines at MIT prepared two days of workshops to expose students to the field of neuroscience, statistical modeling, and machine learning in computational neuroscience. The students worked in Python in Jupyter notebooks.

F. Hackathon

Through experiences of the NIH-funded training grant Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K) we have identified hackathons as an ideal environment to teach science in an interdisciplinary fashion. Thus, for the IQ-Bio-REU we organized a mini-hackathon at the end of the summer, first to help the students of the REU to solidify their understanding of the research carried out throughout the summer and second to get others, i.e. graduate students, faculty and other undergraduates, motivated to explore the field of data science. REU members lead an interdisciplinary team, presented their research in code using Jupyter or R markdown, received feedback on their code and research from the local community of mentors and our collaborators of the REU and finally presented their research to other scientists in the community.

The hackathon served as a venue for REU participants to get immediate feedback and coaching on their summer work, engage with faculty and students from different disciplines, build leadership and public speaking skills, and refine critical thinking and research skills needed to succeed in science today through a more interactive environment than the typical poster session. Several of our mentors from the Machine Learning Workshop and the Replicathon returned for the event to help mentor the students.

V. Program Outcomes

A. Professional Workshops

From June 3-7, 2019 we ran a 5 half-day workshop Software Carpentry Introduction to R programming workshop, plus professional workshops to help students understand what is research for the other half of the day including topics such as imposter syndrome, science communication, lab safety, and proposal writing to name a few. The Software Carpentry workshop included a pre- and post- workshop surveys to see participants attitudes and self-reported skill levels. The surveys indicated that participants gained confidence in their data analysis skills. In response to "I can write a small program/script/macro to solve a problem in my own work" the pre-workshop survey results showed 46% of learners strongly disagreed with that statement, 23% agreed, and 0% strongly agreed. Post-workshop numbers changed to 14% strongly disagree, 43% agree, and 29% strongly agree.

B. Undergraduate Research Student Self Assessment (URSSA)

The Undergraduate Research Student Self-Assessment (URSSA) is a self-report survey instrument that is widely used for the evaluation of research experiences. It is intended as a retrospective self-report of perceived gains in understanding, skills and shifts in attitudes of students after participating in a research experience. This instrument as been repeatedly tested and verified for item reliability and validity [13]. It is available online (https://salgsite.net), for free and it is the required survey instrument for all NSF Biology REUs. The survey instrument was administered to participants on the final day of the REU experience.

In general, participants perceived that they had significant gains in thinking and working like a scientist, skills and more positive attitudes towards research after participating in the program. Notably when asked to gauge their skills in "analyzing data from patterns" on a 5 point scale, where 1=no gain and 5=great gain, the median score (M) was 4 (Standard Deviation [SD]=1.12). For "Understanding the connections among scientific disciplines", M=4.9 (SD=0.32) with 90% of participants reporting "great gain". Specifically with regards to gaining skills "working with computers", M=4.4 (SD=0.84, 60% great gain). Taken together, these results suggest that students embraced the interdisciplinary nature of the program and perceived that the program had been useful for the development of general computational skills in an interdisciplinary setting.

In addition to the quantitative data, qualitative data from the URSSA survey about reflections on changes in attitudes and behaviors further suggests that the first iteration of the IQ-Bio-REU was successful in engaging more students in computer science research. Table 2 shows sample quotes of reflections from participants regarding the impact of the program on their career choices. Similarly, in Table 3 when asked to expound about other gains from the program, participants also talked about gains in computational skills and integration into the scientific community.

Table2: How did your research experience influence your thinking about future career and graduate school plans?

"I wanted this experience to decide if I was going to continue on Computational Biology and I am very satisfied to know that I do feel confident about it."

"Working here made me realize that I may be more interested in sociology than biology, but has helped me to feel that computer science is accessible. It is possible that I will work in the field of sociology and incorporated the interest in computer science that I have gained here."

"It makes me feel that I belong in research."

Table 3: Did you make other gains from doing research that we didn't mention? If so, please briefly describe these.

"I learned computer science, and programming. When I came here I didn't know anything about it, and my research mentor was an excellent help in all of that. "

"Inspiration, experience in other fields, global networking, computational skills, first hack-a-thon. "

"It really helped me that they addressed imposter syndrome here. It helped me to feel like I belong in the community of scientists. It also helped me that they addressed professional development, networking, and gave us a lot of access to scientists in informal settings which gave us the ability to ask questions and learn about the structure of the scientific world."

C. Unexpected Outcomes

As seen in the quotes above, the experiences of IQ-Bio-REU triggered changes in attitudes beyond skill acquisition. In an unsolicited email months after the summer REU had concluded, one of our Hispanic REU students from the US stated that "XXXX was honestly such a transformative experience, I left so prideful of being Latino. Really, really proud! I wear my XXX lanyard everywhere here at [university]. I may not be XXX, but we share so much in common! I think that's what's getting my through the hard times here is that I'm not just pursuing my goals for myself, but for my family and wider community." Others in XXXX from predominantly undergraduate colleges have continued their research after the summer and become a part of the research community on both on the XXXX campus and at their home institutions.

The most unexpected outcome was the bond that the trainers that came from our collaborators on the mainland made with our students and program. Several were moved to return for the Hackathon and others were determined to help the university after seeing the lack of computational science on campus to the point of writing a proposal to develop curriculum for the university. It was clear that it had been an eye-opening experience for them to witnessing the digital and data divide that exists between our university and their own.

On the night before the REU students had to turn in final research reports, they lost power in the dorm and some woke to flooded rooms. We were able to find them housing shortly afterwards. The experience was one they will never forget nor take for granted. We suspect that this experience in particular will make them more resilient researchers and investigators. One thing is clear is that they created a bond with each other that will not be broken, and they will support each other no matter what career path they choose because of their experience here.

VI. Discussion

In 2015, an interdisciplinary group of faculty members from Computer Science, Mathematics, and Biology joined forces to create an informal training program for both faculty and students to develop computational biomedical research on campus with the ulterior motive of increasing diversity in computing. These trainings lead to the creating and updating of courses in data science, statistical methods, and big data in the departments of Computer Science, Mathematics and Biology, respectively, as well as to the ongoing development of computational research in the natural sciences and of a potential interdisciplinary data science program. It has also enriched collaborations with doctoral universities in the XXX of Very High Research Activity that have strong computational biology, public health, and biomedical data science programs. These collaborations lead to the creation of an innovative and resourceful summer REU program XXXX to bring underrepresented students from the US and all of XXXX to our campus and train them to become successful researchers in interdisciplinary and quantitative biological research. Our collaborations with high research universities allowed us to bring postdoctoral and graduate students from our collaborating institutions to XXXX to train our students and faculty as well as our send students and faculty to those institutions to be trained. These mutually beneficial collaborations also have an indelible mark on the lives of the post-doctoral and graduate students from these high research institutions which will make them better faculty and professionals in the future.

The need for mutual collaborations among top tier research universities and other research institutions that are experiencing the data and digital divide is crucial to bridge this divide. Developing such collaboration requires creating informal learning spaces to develop interuniversity networks among students and faculty. Top research institutions need to recognize that if they are taking intellectual human resources from a community that is economically challenged, they need go above and create mutually beneficial collaborations with the university in that community to help fill the void that has been left so that the student that they educate from that community may return to a university that is not stuck on the other side of the digital and data divide.

Acknowledgments

We would like to thank our outstanding collaborators at our partnering institutions: Rafael Irizarry, our former alumnus and co-PI Rafael Irizarry at Harvard University; David Boone at the University of Pittsburgh; Benedict Paton and Zia Isola at the University of California Santa Cruz; Mandana Sassanfar at the Center for Brains, Minds and Machinery; and Renata Rawlings-Goss of the South Big Data Hub at Georgia Tech. This research is funded by NSF Award #1852259.

References

[1] DuBow, W. & Pruitt, A.S. (2019 Update) NCWIT Scorecard: The Status of Women in Technology. Boulder, CO: NCWIT, www.ncwit.org/scorecard

[2] Rayome, A.D. 5 Eye-opening statistics about Minorities in Tech, https://www.techrepublic.com/article/5-eye-opening-statistics-about-minorities-in-tech/

[3] The Carnegie Classification of Institutions of Higher Education (n.d.). About Carnegie Classification. Retrieved (date optional) from http://carnegieclassifications.iu.edu/ .

[4] Top 20 doctorate-granting institutions ranked by number of minority U.S. citizen and permanent resident doctorate recipients, by ethnicity and race of recipient: 5-year total, 2013–17, NSF Science and Engineering Doctorate Awards https://ncses.nsf.gov/pubs/nsf19301/data Table 9.

[5] Seymour, Elaine, Anne Barrie Hunter, Sandra L. Laursen, and Tracee Deantoni. 2004. "Establishing the Benefits of Research Experiences for Undergraduates in the Sciences: First Findings from a Three-Year Study." Science

Education 88 (4): 493–534. https://doi.org/10.1002/sce.10131.
[6] American Association for the Advancement of Science. Vision and Change in Undergraduate Biology Education: A Call to Action. Washington, DC; 2011
[7] PCAST, Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics.2012, President's Council of Advisors on Science and Technology: Washington, DC.
[8] Linn, M. C., E. Palmer, A. Baranger, E. Gerard, and E. Stone. 2015. "Undergraduate Research Experiences: Impacts and Opportunities." Science 347 (6222): 1261757–1261757. https://doi.org/10.1126/science.1261757.
[9] Gentile, James, Kerry Brenner, Amy Stephens, and Life Studies. 2017. "Undergraduate Research Experiences for STEM Students." Edited by James Gentile, Kerry Brenner, Amy Stephens, and Life Studies. Washington, DC. https://doi.org/10.17226/24622.
[10] NSB, National Science Board Science and Engineering Indicators. 2014, National Science Foundation
[11] NSF, Women, Minorities, and Persons with Disabilities in Science and Engineering. 2013, National Science Foundation.
[12] Wilson G. Software Carpentry: lessons learned [version 2; peer review: 3 approved]. F1000 Research 2016, 3: 62 https://doi.org/10.12688/f1000research.3-62.v2