McEligot, A.J., Cuajungco, M.P., Behseta, S., Chandler, L., Chauhan, H., Mitra, S., Rusmevichientong, P., Charles, S. / Californian Journal of Health Promotion 2018, Volume 16, Issue 1, Pages 1-5.

Big Data Science Training Program at a Minority Serving Institution: Processes and Initial Outcomes

Archana Jaiswal McEligot[1], Math P. Cuajungco[1], Sam Behseta[1], Laura Chandler[1], Harmanpreet Chauhan[1], Sinjini Mitra[1], Pimbucha Rusmevichientong[1], and Shana Charles[1]

[1] *California State University, Fullerton*

Keywords: Data Science, Big Data, Education, Training

Introduction

Big Data refers to data that are so large and complex, making it difficult or impossible to manage with traditional computing/software and hardware (Frost 2015; Bellazzi 2014). The volume of electronic data worldwide, just in the healthcare field, is overwhelming with an estimate of nearly 1 billion terabytes of data, and continuing to exponentially grow to zetabytes of data (Cottle, Hoover, Kanwal, Kohn, Strome & Treister, 2013). Not only is the volume of Big Data staggering, but the variety and veracity of the various forms and types of data (structured and unstructured) make it a critical area for research, training and education (Cottle et al., 2013).

Clearly the area of Big Data and the applied field of data science is rapidly growing, and faces organizational, analytic and other visualization challenges (Sinha 2009; Dinov 2016; Slobean 2015); therefore training diverse undergraduate students to appropriately manage, analyze, interpret and present data provides a key opportunity to uniquely tackle big data problems and provide real-life health solutions to improving health conditions (McEligot 2015; Canner 2017). Recently higher education institutions have initiated programs to train diverse undergraduate students in the complex field (McEligot 2015; Canner 2017). The primary aim of the present commentary is to describe and summarize undergraduate big data/data science-related research experiences via a program established for undergraduate underrepresented students at a minority serving institution. First, we describe programmatic and recruitment procedures. Thereafter, importantly, we discuss the program's multidisciplinary didactic and hands-on research training aspects related to data science, and the subsequent hypothesis generation and testing engaged by faculty-student, faculty-faculty teams utilizing large-open source datasets, as well as lab based techniques and research subsequently utilized to generate big data.

BD-3 REAP Program

At California State University, Fullerton (CSUF), the Big Data Discovery and Diversity-Research Education Advancement and Partnership (BD3-REAP) program provides enriching research experiences and opportunities through exploration and understanding big data sources, diversity, computation, and analytics in efforts to improve health. This program is a faculty mentored, yet student led Big Data science research experience in neuroimaging, genomics and epidemiologic data types. BD3-REAP provides opportunities for students to gain skills in managing, analyzing, and intelligently organizing and conveying biomedical/health information to scientific and local communities.

The primary aim of the project is to provide underrepresented students with in-depth, guided, research-based and hands-on experiences in the emerging field of Big Data science (BDs). This is done while incorporating BDs career advising and graduate pathways seminars to help narrow the achievement gap for underrepresented students. In addition, the program develops and implements a multidisciplinary, multi-institutional program to integrate BDs research

and education into CSUF curricula and research activities across two universities [(CSUF & University of Southern California (USC) – NIH designated Big Data to Knowledge (BD2K) Centers of Excellence (COE)] and three departments (Mathematics, Biology & Health Science). The multidisciplinary focus also contributes to an additional overarching BD3-REAP goal of faculty/faculty research collaborations across disciplines. Overall, the program prepares underrepresented students to participate and continue research and education in BDs, contributing to closing the gap in underrepresented students in bioscience and the BDs workforce.

Student Selection

BD3-REAP carried out several procedures related to student engagement, outreach and selection. For recruitment, we developed brochures, fliers and outreached to students via e-mail, telephone, one-on-one interactions, inter-department correspondence and in-person class visitations in the Health Science (HESC), Mathematics (MATH) and Biology (BIO) departments. For the first year, we outreached and presented to a total of 13 CSUF freshman and sophomore classes [4 Health Science (144), 6 Math (n = 120), 3 Biology (n = 124)] with 58%, 25% and 18%, respectively, completing interest forms. We followed-up via e-mail with interested students and a total of 25 completed an application consisting of academic performance, a narrative on Big Data science interest, as well as a personal statement. BD3-REAP faculty rank ordered the 25 applicants, selecting the top 15 to interview and based on the in-person interviews and assessments, six students (referred to as BD3 Scholars) were selected to enter the program. Post-selection, several meetings were coordinated, including an introductory "Meet & Greet" to familiarize the students with their cohort, faculty and program staff, and a more in-depth small-group research meeting to preliminarily discuss faculty research (both at USC-BD2K COE and CSUF), and potential student-student and faculty-student pairings related to their respective data science research interests.

Research Areas and Subsequent Research Partnerships

During the first year of the program, the BD3 Scholars met regularly with their CSUF respective research mentors to perform and implement research fundamentals for predominantly two purposes: 1) To gain an initial formalized exposure to Big Data science (BDs) research concepts and methodologies, conducting specific research projects 2) To prepare for their BD2K-COE summer research experience at USC. For three weeks at the beginning of the Spring semester, students engaged in visitations/rotations with the three CSUF faculty to learn about their research and select a mentor and research area of interest. After the rotations, students were matched with Dr. Cuajungco's laboratory and with Dr. McEligot, while Dr. Behseta provided statistical support to students for both projects. Students also interacted with and were provided computational and statistical support, as well as exposure to and tools/skills to analyze large open-source datasets by other collaborating faculty and consultants (Drs. Rusmevichientong, Charles and Pogoda). Students also met regularly with the BD3-REAP adviser, Dr. Laura Chandler who provided guidance and input on program requirements and BDs career/graduate school opportunities.

For the initial rotations, Dr. Cuajungco hosted BD3 Scholars in his research laboratory in order for the cohort to be familiar with the process of how Big Data in the field of genomics are created using Next Generation Sequencing (NGS) techniques such as RNA sequencing (RNA-Seq). For four weeks, the scholars learned how to culture and passage human embryonic kidney (HEK)-293 cells. They were also trained to extract RNA, and perform reverse transcription of RNA into cDNA. Three scholars (Amber Cornelius, Jonathan Chacon, Silvia Orozco) joined the Cuajungco laboratory to learn and perform RNA-Seq data analysis. The RNA-Seq data were generated from three brain RNA samples of a mouse model for Mucolipidosis type IV (MLIV) and three wild-type littermate control mice. The scholars were subsequently trained on using RNA-Seq tools available in the Galaxy website

(www.UseGalaxy.org) and Lasergene version 15 Genomics suite for NGS data analysis (DNAStar). Dr. McEligot in collaboration with Dr. Rusmevichientong's, and with Dr. Charles rotations consisted of becoming more familiar with epidemiologic data types, specifically the large on-going open-source National Health and Nutrition Examination Study (NHANES), as well as the California Health Interview Survey (CHIS). Students were introduced to NHANES study design, data collection procedures, the various data types and variables. Subsequently, several students (Galilea Patricio, Emma Navajas, Stephen Gonzalez and Shaina St. Cruise) selected to explore the NHANES dataset. Interweaved with these two research areas, Dr. Behseta conveyed the ideas behind sound statistical approaches for BDs analyses. As such, Dr. Behseta has helped them with R programming, and has guided them initially and preliminarily on statistical modeling of their data analytical projects, and subsequent presentations.

BD3 Scholars research experience was also uniquely augmented and tremendously enriched via their summer research experiences at the BD2K COE at USC where students engaged in exploring various domains of big data from genomics to brain imaging, exploring and visualizing brain imaging datasets, as well as investigating potential demographic, proteomics and genomics variables associated with brain health outcomes via datasets such as the Alzheimer's Disease Neuroimaging Initiate (ADNI) dataset. The critical need and nature of big data science training and the USC BD2K COE research opportunities have been published previously (Toga 2015)

Brief Overview of Articles and BD3 Scholar Research Involvemement

Three of the research articles (McEligot et al., Pogoda et al., Rusmevichientong et al.) in the present Special Issue focus on the large NHANES open-source dataset. Key learning outcomes and goals via these faculty/student collaborations include gaining a more in-depth understanding and ability to characterize the NHANES study design, complexities of accessing large, publically-available volumes of data, challenges of locating and accessing the variety of > 50 data types (demographic, laboratory, dietary, etc.), and subsequently merging, analyzing and presenting, and preparing these data for publication in collaboration with their mentors.

Specifically, for the McEligot et al. faculty/student collaboration, BD3 Scholars investigated the association between dietary folate intake and circulating folate concentrations with depression. Utilizing the same large open dataset (NHANES), in the Pogoda et al. contribution, BD3 Scholars examined the role of caffeine and caffeine metabolites with depression. Via these collaboration students explored the complexities of not only dietary data collection, and nutrient identification (i.e. dietary folate/caffeine intakes with depression), but also gained understanding and appreciation of the nuances, differences and correlations between dietary intakes and corresponding biomarkers. Further, BD3 scholars were exposed to data analytic and computational techniques, including introductory SAS programing, stratification, dietary data distribution and log transformation, energy adjustment methodologies and depression outcomes categorization.

For the Rusmevichientong faculty/student collaboration, BD3 Scholars focused on investigating the role of soda consumption and overall diet quality and obesity using an open access national survey dataset from What We Eat In America, which is the dietary/nutrition data collection portion of NHANES. In order to successfully engage and execute the research question, students underwent various research processes and steps. BD3 Scholars completed key tasks/assignments, including downloading data, converting data from SAS format to SPSS format, merging the data, creating variables using SPSS syntax function, visualizing the data and analyzing the data. Each task was basically designed to build onto the previous assignment. For example, the student must have completed assignment #1 which aims to download the relevant data files and convert them to SPSS format in order to proceed to the second assignment which required the BD3 Scholars to

merge the files together. The students faced challenges, obstacles and errors almost for every task, but it was part of the learning process. In summary, throughout the semester students were continuously and rigorously trained on how to tackle Big Data step by step using the assignment based training and successfully initiated their own research question under faculty's guidance.

Another large open-source dataset, the California Health Interview Survey (CHIS), that collects detailed health-related data, including health outcomes, health behaviors, and access to care on approximately 20,000 households per year was also accessed and explored via a faculty/student team. In the Charles et. al., utilizing CHIS, faculty and BD3 Scholars compared California's adult self-reported rating of delays in seeking medical care. It is anticipated that those adults who were only insured for part of the year will have greater reported delays of medical care than those who were insured or uninsured for the entire year. BD3 Scholars gained research skills in crafting a conceptual framework and research question, creating a data analysis plan, identifying variables in the dataset, examining the variables through frequency analysis, variable consolidation and modification, and final data analysis using chi-square tests and logistic regression. Students each created their own research studies based on their own interests, two of which were subsequently accepted for presentation at the CSUF SCAR conference. One student-led project was also accepted for presentation at the 2017 Annual Meeting of the American Public Health Association in Atlanta, GA.

In Chacon and Cuajungco's paper on comparative de novo transcriptome assembly using two commercial software programs, the authors used Lasergene NGS v.15 and CLC Genomics Workbench v.10 to assemble Newt transcriptomic data, a non-referenced genome, in order to compare the output between the two softwares. The authors reported that both software packages satisfactorily produced de novo assembly data. However, they found that the gene ontology outputs differed markedly based on the differences in algorithm used for the annotation process. Chacon learned how to perform de novo assembly using both software and the open-source Trinity assembly program available on Galaxy hosted by Indiana University (www.galaxy.ncgas-trinity.indiana.edu). Chacon also used R coding to generate the gene ontology (GO) images and learned how to annotate gene contigs using third-party software called Blast2Go (www.blast2go.com). The authors concluded that commercial software platforms have streamlined workflows that make it easier for novice users, but that researchers must fully evaluate the software to determine whether it is the right package for their project.

Two other articles high-light research activities via faculty/faculty collaborations (Charles & McEligot; Mitra & McEligot). The Charles & McEligot article addresses the importance of not only investigating overall expansions of health insurance in California, but also addressing continued ethnic/race disparities in health care access and utilization. While some disparities persist regardless of increases in health insurance, the overall picture has become more complex as California has moved to become a "majority-minority" state. In the article by Mitra & McEligot, both faculty, from two different colleges (Business & Health and Human Development) collaborated to assess curricula effectiveness via implementation of a Big Data Science Video tutorial. Collaborations involved not only development of the video (a nearly year-long process), but also regular meetings to assess student learning, address student concerns and implement appropriate changes related to the video tutorial. The primary goal was to investigate whether curricula implemented via multimedia can serve as an effective learning tool in introducing and creating interest in the field of Big Data and Data Science for underrepresented students.

Conclusion

In conjunction with their research experiences, students have gained foundational pedagogical training in Big Data science tenants and have learned to: define the complex field of data

science, identify open-source datasets, conduct literature reviews, distinguish study designs, identify various data storage tools, develop a codebook, review and comprehend data dictionary structure, access data, merge data, preliminarily navigate neuroimaging platforms, analyze via R and subsequently synthesize data. Initial formal classroom student assessment via tests, literature review write-up and mini-progress reports indicate that majority of the students excel at searching for peer-reviewed manuscripts via pub-med, have appropriate grammar and writing skills, preliminarily program in R, have a good understanding of neuroscience and can extract data, merge and synthesize the results. Also, investigations and findings from the present Data Science Training Special Issue demonstrate competencies in hands-on research experiences testing research hypotheses utilizing large open-source data sets, as well as data science techniques in relation to behavioral and brain health issues.

The Big Data Discovery and Diversity through Research Education Advancement and Partnerships (BD3-REAP) program has impacted human resource development via providing opportunities in Big Data science (BDs) research and teaching, improving underrepresented student skills and attitudes towards BDs, as well as developed and disseminated educational materials. The hallmark of the BD3-REAP program is to provide underrepresented students research and training experiences in the complex, multidisciplinary BDs field.

References

Bellazzi, R. (2014). Big data and biomedical informatics: a challenging opportunity. *Yearbook of Medical Informatics, 9(1),* 8.

Canner, J. E., McEligot, A. J., Pérez, M. E., Qian, L., & Zhang, X. (2017). Enhancing diversity in biomedical data science. *Ethnicity & Disease, 27(2),* 107.

Cottle, M., Hoover, W., Kanwal, S., Kohn, M., Strome, T., & Treister, N. (2013). Transforming health care through big data strategies for leveraging big data in the health care industry. New York: Institute for Health Technology Transformation; 2013.

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience, 5(1),* 12.

Frost, S. (2015). Drowning in big data? Reducing information technology complexities and costs for healthcare organizations. Retrieved on June 7th, 2018 from: http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf

McEligot, A. J., Behseta, S., Cuajungco, M. P., Van Horn, J. D., & Toga, A. W. (2015). Wrangling big data through diversity, research education and partnerships. *Californian Journal of Health Promotion, 13(3),* vi.

Scruggs, S. B., Watson, K., Su, A. I., Hermjakob, H., Yates, J. R., Lindsey, M. L., & Ping, P. (2015). Harnessing the heart of big data. *Circulation Research, 116(7),* 1115-1119.

Sinha, A., Hripcsak, G., & Markatou, M. (2009). Large datasets in biomedicine: a discussion of salient analytic issues. *Journal of the American Medical Informatics Association, 16(6),* 759-767.

Slobogean, G. P., Giannoudis, P. V., Frihagen, F., Forte, M. L., Morshed, S., & Bhandari, M. (2015). Bigger data, bigger problems. *Journal of Orthopedic Trauma, 29,* S43-S46.

Toga, A.W., Foster, I., Kesselman, C., Madduri, R., Chard, K., Deutsch, E.W., Price, N.D.,
Glusman, G., Heavner, B.D., Dinov, I.D., Ames, J., Van Horn, J., Kramer, R., & Hood, L.
(2015). Big biomedical data as the key resource for discovery science. *Journal of the American Medical Informatics Association, 22(6)*:1126-31.

Van Horn, J.D., Fierro, L., Kamdar, J., Gordon, J., Stewart, C., Bhattrai, A., Abe, S., Lei,