Wrangling Big Data Through Diversity, Research Education and Partnerships

Archana Jaiswal McEligot[1], Sam Behseta[1], Math P. Cuajungco[1],
John D. Van Horn[2], and Arthur W. Toga[2]

[1]*California State University, Fullerton*
[2]*University of Southern California, Los Angeles*

Abstract

© 2015 Californian Journal of Health Promotion. All rights reserved.

Big Data Challenges

The advent of Big Data science (BDs) has generated enormous amounts, varieties, and sources of complex datasets that have vast potential for the creation of new knowledge, particularly in relation to primary and secondary disease prevention (Eaton et al., 2012); yet BDs also brings inherent challenges of utilization and value. A critical cross-cutting issue is the creation of a compelling and effective user experience that can empower biomedical researchers and trainees with limited information technology budgets access to powerful and intuitive tools designed to effectively address the challenges posed by the four dimensions of Big Data: (1) volume: the vast amount of data that is generated through source integration; (2) variety: the lack of standardization that is inherent in combining data from different resources; (3) velocity: the high rate at which data is constantly changing; and (4) veracity: the need for reliability measures and safeguards protecting the confidentiality of the individuals involved (Otero, Hersh, & Jai Ganesh, 2014). These challenges are particularly pronounced in neuroscience Big Data, as neuroimaging produces some of the largest and most complex data types (Van Horn & Toga, 2014; Turner & Van Horn, 2012; Bowman, Joshi, & Van Horn, 2012). Through advances in neuroimaging techniques, such as functional magnetic resonance image (fMRI) and positron emission tomography (PET), massive stores of high-resolution and high-dimensional brain images have been produced (Fan, Han, & Liu, 2014; Van Horn & Toga, 2014). With the emergence of publicly available databases and repositories, the vast nature of neuroimaging, genomics/proteomics, and epigenetics data enables integrative analysis combining information from many sources. This important challenge, as well as opportunity, in neuroscience of aggregated datasets from multiple sources can lead to systematic biases caused by experimental variations and result in outliers and missing values (Fan, Han, & Liu, 2014). While posing a significant challenge in neuroimaging and epigenetics Big Data utilization, the collaborative and profound nature of BDs processing tools/technology sets the stage for future development of advanced BDs computational and visualization methods combined with fundamental biological and environmental bases of disease, which could contribute to novel approaches in understanding disease etiology, personalized medicine, and other human health issues (Dinov et al., 2014).

Utilization of Big Data and Diversity

In addition to the development of systems and methods to most effectively address these BDs issues and best utilize vast stores of data, the informatics field must train professionals who will pioneer this work (Otero, Hersh, & Jai Ganesh, 2014). According to a report by the McKinsey Global Institute, the United States "faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze Big Data and

make decisions based on their findings" (Manyika et al., 2011). To effectively meet the BDs skills gap, bioscience programs must incorporate analytics and Big Data into curricula and provide research experiences that (1) integrate ongoing peer group collaboration, concentrated mentored research projects, and research ownership; and (2) applies foundational knowledge to analyzing and solving Big Data problems through the development of four key skills: (a) programming with data-oriented tools; (b) developing working knowledge of how to apply statistical tools and techniques; (c) developing higher-level bioscience domain knowledge; and (d) being able to understand community needs and articulate results to them (Corwin, Graham, & Dolan, 2015; Manyika et al., 2011).

The increasingly technological and data-driven environments that hold the key to solving critical bioscience questions herald the need for a diverse workforce reflective of community demographics and capable of managing, analyzing, and intelligently organizing and conveying biomedical/health information to scientific and local communities. While ethnic diversity within the U.S. population has changed between 2000-2010, with Hispanics growing by 43% and African Americans by 12.3% (Humes, Jones, & Ramirez, 2011), this demographic shift is far from adequately represented in the biomedical sciences, as blacks, Hispanics, and Native Americans account for only 7.1% of all employed biological/biomedical and life sciences workforce, including BDs (National Academy of Sciences, 2011; NSF, 2015). Such a shortage of diversity among data scientists limits not only the perspectives and insights brought into BDs fields, but also the ability of data science to address and communicate issues specific to historically underrepresented ethnic communities.

Big Data Programs and Partnerships:

The crux of developing a skilled and sustained diverse pipeline in Big Data science (BDs) and the biomedical field is to create research learning communities that integrate ongoing peer group collaboration, concentrated mentored research projects and research ownership (Corwin, Graham & Dolan, 2015). The emergence of BDs has generated enormous amounts, varieties, and sources of complex datasets (Eaton et al., 2012); yet BDs also brings inherent challenges of utilization and value. The increasingly technological and data-driven environments that hold the key to solving critical bioscience questions herald the need for diverse trained researchers reflective of community demographics and capable of managing, analyzing, and intelligently organizing and conveying the biomedical/health information to scientific and local communities.

California State University, Fullerton (CSUF) enrolls a diverse student population that is 36% Hispanic, 21% Asian and Pacific Islander, 25% non-Hispanic white, and 2% African American. *Hispanic Outlook in Higher Education* ranks CSUF first in California and fifth in the nation among colleges and universities awarding bachelor's degrees to Hispanics. The colleges of Natural Science and Mathematics (NSM), and Health and Human Development (HHD) conduct epigenetics, neuroscience and analytical research, and provide didactic and research training to underrepresented students in these areas (McEligot et al., 2014; Cuajungco et al., 2014; Behseta et al., 2011). However, CSUF has yet to develop a comprehensive/focused Big Data bioscience program, integrating BDs didactic competencies with in-depth Big Data research experiences for their underrepresented students and faculty. The University of Southern California (USC), home to the Big Data for Discovery Science (BDDS), a National Institute of Health (NIH) Big Data to Knowledge (BD2K) Center of Excellence, is focused on the design and development of BDs analysis tools, creating a framework for interactive integration and exploration of multi-omic neuroimaging Big Data in order to discover new insights, organize new knowledge, and form hypotheses (Van Horn & Toga 2014). A critical cross-cutting issue is the creation of a compelling and effective user experience that can empower biomedical researchers and trainees with limited information technology budgets access to powerful and intuitive tools. CSUF's existing epigenetics, neuroscience and statistical expertise and diverse student engagement

matches well with USC's BDDS neuroimaging capabilities and goal to empower the next generation BDs users. Thus, CSUF, in collaboration with USC proposes the *Big Data Discovery and Diversity through Research Education Advancement and Partnerships* (BD^3-REAP) program, which will develop new neuroimaging and epigenetics BDs curricula at CSUF, while providing mentored, student-owned research experiences for diverse undergraduate CSUF students and faculty. By integrating CSUF's strong history in training underrepresented students, as well as existing epigenetics, neuroscience and statistical expertise with USC's BDDS neuroimaging and Big Data Center of Excellence, we aim to establish an innovative program on BDs comprehension, computation, and analysis related to neuroimaging, genomics/proteomics and epigenetics that incorporates BDs concepts into classroom learning while emphasizing multi-faceted undergraduate research experiences for predominantly underrepresented students. Moreover, the program will work closely with the BD2K effort's Training Coordinating Center (TCC), also housed at USC, to ensure the integration of BD^3-REAP with broader NIH center programs relevant to training in big data biomedicine nationwide. Therefore, through comprehensive classroom-to-research-intensive training at CSUF in partnership with a complementary and integrated research experience at USC's BDDS center and other BD2K-affiliated centers, BD^3-REAP will matriculate 18 ($\geq$ 18 yrs; 9 males & 9 females) students from predominantly underrepresented backgrounds by 2020 in an effort to prepare them for higher education and futures in the increasingly technical and competitive field of BDs, while concurrently creating role models for future underrepresented students interested in pursuing BDs research.

Acknowledgments

We would like to thank Harmanpreet Kaur Chauhan for formatting and citation assistance.

References

Behseta, S. & Chenouri, S. (2011). Comparison of two population curves with an application in neuronal data analysis. *Statistics in Medicine, 30*(12), 1441-1454.

Bowman, I., Joshi, S.H., & Van Horn, J.D. (2012). Visual systems for interactive exploration and mining of large-scale neuroimaging data archives. *Front Neuroinform, 6*, 11.

Corwin, L., Graham, M.J., & Dolan, E.L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE–Life Sciences Education, 14*(1), 1-13.

Cuajungco, M.P., Basilio, L.C., Silva, J., Hart, T., Tringali, J., Chen, C.C., Biel, M., & Grimm, C. (2014). Cellular zinc levels are modulated by TRMPL1-TMEM163 interaction. *Traffic, 15*(11), 1247-1265.

Dinov, I.D., Petrosyan, P., Liu, Z., Eggert, P., Hobel, S., Vespa, P., Woo Moon, S., Van Horn, J.D., Franco, J., & Toga, A.W. (2014). High-throughput neuroimaging-genetics computational infrastructure. *Frontiers in Neuroinformatics, 8*, 41.

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., & Zikopoulos, P. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. New York: McGraw-Hill.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review, 1*(2), 293-314.

Humes, K.R., Jones, N.A., & Ramirez, R.R. (2011). *Overview of race and Hispanic origin: 2010*. U.S. Census Bureau. Accessed April 6, 2015, from http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A.H. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Accessed April 5, 2015, from http://bigdatawg.nist.gov/MGI_big_data_full_report.pdf

McEligot, A.J., Chandler, L., Tran, N., Pillazar, L., & Steinberg, F. (2014). Cultural context is associated with intent to pursue nutrition careers in a diverse population. *The Journal of the Federation of American Societies for Experimental Biology, 28*(Suppl. 1), 118-124.

National Academy of Sciences, National Academy of Engineering and Institute of Medicine. (2011). Expanding Underrepresented Minority Participation: America's Science and Technology Talent at the Crossroads. Washington, DC: National Academies Press; www.nap.edu.

Otero, P., Hersh, W., & Jai Ganesh, A.U. (2014). Big data: Are biomedical and health informatics training programs ready? *International Medical Informatics Association Yearbook of Medical Informatics, 9*(1), 177-181.

Turner, J.A. & Van Horn, J.D. (2012). Electronic data capture, representation, and applications for neuroimaging. *Frontiers in Neuroinformatics, 6*, 16.

Van Horn, J.D. & Toga A.W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behavior, 8*(2), 323-331.

Author Information

*Archana Jaiswal McEligot
California State University, Fullerton
Department of Health Science
800 North State College Blvd.
Fullerton, CA 92834
Email: amceligot@fullerton.edu
Phone: 657-278-3822

* corresponding author