Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 197 (2015) 1443 – 1446

Procedia
Social and Behavioral Sciences

7th World Conference on Educational Sciences, (WCES-2015), 05-07 February 2015, Novotel Athens Convention Center, Athens, Greece

Preparing undergraduate students majoring in Computer Science and Mathematics with Data Science perspectives and awareness in the age of Big Data

Kanyarat Bussaban[a*] , Phanu Waraporn[a]

[a]*Suan Sunandha Rajabhat University, 1 U-TongNok Road, Wachira, Dusit,Bangkok, THAILAND 10300*

Abstract

Undergraduate students majoring in Computer Science and Mathematics are entering the workforce not only as programmers and mathematicians but also as data and business intelligent analysts. These job profiles require students to effectively utilize databases and data warehouses technologies, summarize data from external sources including the Internet and provide solutions to complicate, dynamic and ever-changing problems. These areas of hard skills have not been integrated as a major component of undergraduate programs in mathematics and computer science. This paper is aimed at showing how to motivate the significance of mastering data science proficiency as well as depicting examples and resources for lecturers in implementing data science in computer sciences and mathematics curriculum. Two case studies from Computer Science and Informatics Mathematics Programs at Faculty of Science and Technology, Suan Sunandha Rajabhat University in Bangkok, Thailand are presented.

Peer-review under responsibility of Academic World Education and Research Center.

Keywords: Undergraduate Science Curriculum, Data Science, Data Analyst

1. Introduction

The growing number of bachelors in science and technology, particularly computer science and mathematics, conferred has increased in the last several years and the trend is continuing. This signifies strong workforces that can be deployed in both public and private sectors. Surprisingly, this is not the valid as the skills trained are not fully maximized by most employers as they are not looking for those to work in solitude but rather trans-disciplinary. In

* Kanyarat Bussaban. Tel.: +662-160-1159; fax: +662-160-1146
E-mail address: kanyarat.bu@ssru.ac.th

Peer-review under responsibility of Academic World Education and Research Center.
doi:10.1016/j.sbspro.2015.07.092

order to address this impending issue of hugely available but lack in quantitatively trained-skills human resources (Manyika, et al., 2011; Zorn, et al., 2014), changes in curriculum are apparent and indeed mandatory. This will ensure that undergraduates are ready for the practical data analysis works ahead using the real world data. With the emergent rise of Big Data era brings about even more an immediate need for handling the existing and future revised curriculum so that benefits are for both students and entrepreneurs. Various leading associations stress the need to attract, train and retrain present and future batch of undergraduates (Rudin, 2014 ;Fox and Hendler, 2014).

These are challenges facing faculty members to properly improve the process of preparing students to the outside world. This paper identifies major areas that must be attended to through several approaches being integrated to the curriculum without changing it prior to its round of revision (every five year, as stipulated by Commission on Higher Education law in Thailand).

DATA SCIENCE – WHAT IS IT?

Cleveland (Cleveland, 2001) coined this term and described this discipline as generally involving the mixture of statistics and large-scale computing (Greenhouse, 2013). Typically, data science personnel per Wills (2014) should be inquisitive, be able to make use of data from multiple sources in order to detect either the trend or anomaly. His or her work involves cycle being generally surrounded by (1) problem understanding, (2) statement of problem development, (3) acquisition of high quality data, (4) data generation process, (5) domain specialization and (6) modelling ability.

Informatics Mathematics Program

Since mid of 2014, data science is integrated into various courses using R programming and SQL so that a practical foundation for students to compute with data, by participating in the entire data analysis cycle (from forming a statistical question, data acquisition, cleaning, transforming, modeling and interpretation). Also this introduces students to tools for data management, storage and manipulation that are common in data science, and students apply those tools to real scenarios. Students undertake practical analyses using real, large, messy datasets using this widely used modern computing tools and learn to think statistically in approaching all of the aspects of data analysis. While some of the topics covered in the course come from existing offerings in applied statistics and computer science, effort is made in Data Science to present a unified blend of this material, such that students recognize that both fields contribute to answering questions from data. The course itself can be thought of as having five components: data visualization (e.g., data graphics, elements of visual perception), data manipulation (e.g., SQL, merging, aggregating and iterating), computational statistics (e.g., confidence intervals via the bootstrap, simulation, regression, variable selection), data mining/machine learning (e.g., classification, cross-validation), and additional topics (e.g., text mining, mapping, regular expressions, network science).

In its first offering, the staff of Smith's GIS (geographical information systems) laboratory regularly attended the class. This facilitated incorporation of lessons on spatial data and mapping techniques into the curriculum. This topic was popular with the students, since the ability to generate data maps was perceived to be useful in terms of visualization and communication.

Computer Science Program

Starting of 2014, an introduction to programming with Python and R programming is offered to CSC1002 Programming Language I instead of Java. Here a wide variety of computer technologies relevant to storing, managing, and processing data are added apart from normal course outline. The course has two aims: to teach programming tools specific to the handling of data, and to teach and build confidence with general concepts of computer languages. Data Technologies also aims to build students' awareness of the range of tasks that a computer is capable of performing (in addition to providing concrete tools for performing specific tasks). Specific topics include: How to write computer code; publishing data on the World Wide Web (HTML); data description and semantic markup (XML); data storage (file formats, spreadsheets, databases); data management and summary (database queries, SQL); data processing (R).

2. Proposed alteration to existing curriculum or addendum

Area of focus as depicted in Table 1: Proposed specific contents to be emphasized is what was being covered during year 2014 and continuing. The results as to the number of enrolled students increases, number of satisfying students and the satisfaction level of entrepreneurs are quite early to report on. However, the preliminary assessment is found to be satisfactory since the advanced courses taken by these students currently assessed are very promising.

The extensive coverage of R Programming and Python as an open-sourced programming not only assist efficiently in this attempt but the widely adoption by rest of world proving from the freely available communities and libraries allowed for the detailed study that can be modified, manipulated and adapted to the local party very well.

Table 1. Proposed Specific Contents to be emphasized

Focus	Particular
Statistics	Simulation, Visualization and Implementation
Data Format	Arrays, Textual Data and Data Cleansing
Data Technologies	Relational Databases, Structured Query Language (SQL), Big Data and Cloud, Regular Expression (RE), eXtensible Markup Language (XML), Shell Scripting and Web Scraping
Programming	Structured Programming, Scripting Programming, High Performance and Scientific Computing and Efficiency
Business Process	Web Publishing and Version Control

3. Conclusion, Discussion and Future Works

There are surprisingly quite a number of positive responses on the integration and incorporation of data science into the teaching of computer science and mathematics both at the faculty and student levels as they perceived these attempts as relevant, exciting and practical. This new perspective of opportunity allows for more detailed coverage, for example, research methodology (statistics), problem solving and critical thinking and computing. An interesting but vital is that the complete data analytics cycle; formulating a statement of problem, acquiring data, pre-processing data, analysing the data and finally conveying the analytic results, can be repeated. The ability to perform data analysis according to Speed (1986) helps students achieving their quantitative skills practically for the employers since most of the assignments involve elaborating computing aptitude and insight of data analysis. Encouragement of early trained K-12 education definitely helps transit this process easily. But as the Commission on Higher Education (CHE) is not responsible for this level, this plan will be addressed separately in other venue.

In conclusion, the focus of this paper is to prepare the students with the myriad of data they will soon deal with in their future career and that is what Nolan & Temple Lang (2010) and Brown & Kass (2009) has asked for a thorough restructuring of the syllabus. This paper merely proposes what being currently tried out at our programs to

a community at large so that comments received can be used to further and better develop a more compact program that can modernize the science and technology curriculum, particularly the computer science and mathematic programs.

Acknowledgements

The authors wish to gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

References

Bailer, J., Hoerl, R., Madigan, D., Montaquila, J., & Wright, T. (2012). *Report of the ASA Workgroup on Master's Degrees.* Alexandria, VA: American Statistical Association. Retrieved from http://magazine.amstat.org/wp-content/uploads/2013an/masterworkgroup.pdf

Brown, E., & Kass, R. (2009). What is statistics? *The American Statistician, 63*, 105-110.

Chance, B., & Rossman, A. (2014). *Investigating Statistical Concepts, Applications, and Methods.* San Luis Obispo, CA: http://www.rossmanchance.com/iscam2/. Retrieved from http://www.rossmanchance.com/iscam2/

Cleveland, W. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistics Review, 69*, 21-26.

Conway, D. (2010 йил 30-9). *The Data Science Venn Diagram*. Retrieved from drewconway.com: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Cuny, J., Snyder, L., & Wing, J. M. (2010). *Demystifying Computational Thinking for Non-Computer Scientists.*

Fox, P., & Hendler, J. (2014, June). The Science of Data Science. *Big Data, 2*(2).doi:10.1089/big.2014.0011

Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review, 78*(2), 297-315.

Gould, R., Baumer, B., Mine, C.-R., & Bray, A. (2014). Big Data Goes to College. *AMSTAT News*, http://magazine.amstat.org/blog/2014/06/01/datafest/.

Greenhouse, J. B. (2013, 7 26). Statistical Thinking: the bedrock of data science. *The Huffington Post*. Retrieved 7 29, 2014, from http://www.huffingtonpost.com/american-statistical-association/statistical-thinking-the-bedrock-of-data-science_b_3651121.html

Horton, N. J., Baumer, B. S., & Wickham, H. (2014). Teaching Precursors to Data Science in Introductory and Second Courses in Statistics. *International Conference on Teaching Statistics 9.*

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Johnstone, I., & Roberts, F. (2014). *Data Science at NSF*. NSF. Retrieved from http://www.nsf.gov/attachments/130849/public/Stodden-StatsNSF.pdf

Madigan, D. (2014). *Statistics and Science: A Report of the London Workshop on the Future of the Statistical Sciences.* Retrieved from http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Madigan, D. (2014). *Statistics and Science: A Report of the London Workshop on the Future of the Statistical Sciences.* Retrieved from http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data:The next frontier for innovation, competition, and productivity.* McKinsey & Compnay. Retrieved from

Murrell, P. (2009). *Introduction to Data Technologies.* Chapman Hall. From https://www.stat.auckland.ac.nz/~paul/ItDT/

National Academy of Sciences, N. A. (2010). *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5.* Washington, DC: The National Academies Press.

Nolan, D., & Temple Lang, D. (2010). Computing in the Statistics Curricula. *The American Statistician, 64*, 97-107.

Nolan, D., & Temple Lang, D. (2014). *Explorations in Statistics Research: A model for undergraduate co-curricular exposure to modern research problems.*

Nolan, D., & Temple Lang, D. (2015). *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving.* CRC Press.

Rudin, C. (2014). *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society.* American Statistical Association. Retrieved from http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf

Speed, T. (1986). Questions, Answers, and Statistics. *Conference Proceedings of the International Conference on Teaching Statistics 2* (pp. 18-28). IASE. Retrieved from http://iase-web.org/Conference_Proceedings.php?p=ICOTS_2_1986

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* New York: Springer. From http://had.co.nz/ggplot2/book

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, submitted.

Zorn, P., Bailer, J., Braddy, L., Carpenter, J., Jaco, W., & Turner, P. (2014). *The INGenIOuS Project: Mathematics, Statistics, and Preparing the 21st Century Workforce.* The Mathematical Association of America. Retrieved from http://www.maa.org/programs/faculty-and-departments/ingenious