A Preliminary Report on Infusing Data-Enabled Active Learning in Undergraduate CS Mathematics and Statistics Courses

Carl Pettis[1], **Rajendran Swamidurai**[1], **and Ash Abebe**[2]
[1]Mathematics and Computer Science, Alabama State University, Montgomery, AL, USA
[2]Statistics, Auburn University, Auburn, AL, USA

Abstract - *This paper presents an experience in designing and implementing a big data based learning model for undergraduate computer science mathematics and statistics courses. Industry demands workers who can retrieve useful information from very complex, unstructured data. The mathematics courses offered for computing majors at universities help students develop the logical thinking and problem-solving skills while the statistics courses introduce students to methods of collection, organization, analysis, and interpretation of data; however, big data analytics requires new mathematical and statistical methods and algorithms developed specifically for use with big data. We describe the design and implementation of infusing big data analytics in existing computer science undergraduate mathematics and statistics courses.*

Keywords: Big data; big data analytics; active learning.

1 Introduction

Recent advances in technology, such as e-commerce, smart phones, and social networking, are generating new types of data on a scale never seen before—a phenomenon known as "big data." [1]. Industry demands workers that can retrieve useful information from very complex, unstructured data. The current undergraduate mathematics courses help students develop the logical thinking and problem-solving skills while the statistics courses introduce students to methods of collection, organization, analysis, and interpretation of data; however, big data analytics requires new mathematical and statistical methods and algorithms developed specifically for use with big data. We strongly believe that equipping students with such skills greatly improves their employability.

The U.S. Bureau of Labor Statistics (BLS), Occupational Outlook Handbook [2] highlights our claim. The report states, "The amount of digitally stored data will increase over the next decade as more people and companies conduct business online and use social media, smartphones, and other mobile devices. As a result, businesses will increasingly need analytics to analyze the large amount of information and data collected. Analyses will help companies improve their business processes, design and develop new products, and even advertise products to potential customers."

A recent survey of senior Fortune 500 and federal agency business and technology leaders by the Harvard Business Review [3] found that "85% of the organizations surveyed had funded Big Data initiatives underway or in the planning stage". The same survey reports that 70% of the respondents plan to hire data scientists, but nearly all report finding employees skilled in big data analytics as challenging to impossible. The nature of academic research is also transforming from model-driven to data driven. For instance, NASA is collaborating with Amazon Web Services Inc. (AWS) to make a large collection of NASA climate and Earth science satellite data publicly available to researchers in an effort to "grow an ecosystem of researchers and developers who can help us solve important environmental research problems" [8]. Higgs bosons were discovered recently by clever algorithms that mined terabytes of data for their signature. While STEM careers in academia and industry are increasingly requiring technical skills for dealing with the analysis of "big data", undergraduate courses in mathematics and statistics fall short of providing adequate training to students in data-driven methods that integrate theory and computation.

This paper presents an experience in infusing, teaching, and assessing big data modules in various undergraduate mathematics and statistics courses that immerses students in real-world big data practices through active learning. Our courses walked students through producing working solutions by having them perform a series of hands-on big data exercises developed specifically to apply cutting-edge industry techniques with each mathematics and statistics course module.

2 Related Works

Universities are waking up to the need for developing skills in big data analytics. Several universities now have graduate level courses focused on big data. Some have masters programs in data science; however, there is very little

evidence of big data concepts being integrated in undergraduate mathematics and statistics courses. Exceptions are the National Science Foundation (NSF) funded EXTREEMS-QED project at the College of William and Mary and a senior big data projects course offered by the Department of Mathematical Sciences of the University of Montana [9].

3 Big Data and Mathematics

3.1 Linear Algebra

Several of the methods used in big data analytics such as feature extraction, clustering, and classification involve the manipulation of large matrices. The important topics in big data analytics should have learn include understanding the relationships between matrix decomposition and principal components analysis for dimension reduction, application of eigenvectors (e.g. in Google's PageRank method), performing multiplication of very large matrices using block decomposition methods, measuring distance between objects represented as vectors (e.g. Jaccard, Hamming, cosine), and understanding the relationship between projections and least squares optimization for regression and clustering.

3.2 Discrete Mathematics

Linked data are usually represented by a graph (vertices and edges). Notions such as centrality, shortest path, and reachability can be derived from the graph using graph analytics. A widely used practical application of large graph analytics is the internet search engine. Topics such as visualizing big data as graphs (e.g. the World Wide Web), computation for strongly connected large graphs (e.g. PageRank for strongly connected graphs), matching in bipartite graphs (e.g. Internet advertising), and social networks and hubs are very important for big data analytics.

3.3 Differential Equations

Differential equations explain the underlying dynamics in spatiotemporal pattern formation and detection, disease modeling, image visualization, processing, and analysis, etc. Topics important for big data analytics include numerical solutions systems of differential equations, nonlinear differential equations and stability, and using observed data to refine solutions.

3.4 Probability and Statistics

Statistical methods make up the majority of methods employed for understanding big data and making inferences. Some topics to be considered for big data analytics are Markov processes and the Markov transition matrix (e.g. Web surfing), correlations in high dimensional data, the Bonferroni Principle, and Monte Carlo simulation.

3.5 Modern Geometry

The use of geometry and topology is an emerging area of research in big data analytics. Currently, the methods are used for exploratory data analysis in high dimensional spaces. When exploring big data analytics in this area, one should learn the topics such as the geometry of data, visualization, and recovering low dimensional structures from high dimensional data.

4 Integrating Big Data Analytics in Existing CS Mathematics Courses

To facilitate active learning, the methods were included in two-part modules. The first part focused on theoretical and conceptual ideas behind the methods under discussion and the second part had hands-on experimentation using simulation experiments as well as real data. The initial set of courses in which we integrated big data analysis methods were chosen using two criteria: suitability of material for pedagogical integration of big data methods and impact on all computing majors. Instructors may eventually choose to expand the integration of methods to other mathematics courses in the future. The initial set of courses included:

- Introduction to Linear Algebra
- Differential Equations
- Probability and Statistics
- Modern Geometry

4.1 Introduction to Linear Algebra

Linear algebra concepts such as feature extraction, clustering, and classification involving the manipulation of large matrices are extensively used in big data analytics; therefore, this is a natural course to start introducing students to big data analytics.

Problem-solving is at the heart of computer science, whether it is games or working with data, we are trying to create tools to help us solve whole categories of problems. We have created a one-week big data module, which introduces the idea of an "algorithm" as a set of instructions used to solve a problem. This sets the context for our discussion of searching and matrix multiplication algorithms, which is used in Google PageRank.

The instructional unit was divided into the following three day lectures:

Day-1: 1) *Lecture:* Introduction to algorithms, 2) *Hands-on activity:* For the hands-on activity the students were grouped into pairs. Each group gets a deck of random number of play cards and is asked to find a specific "key card" in the deck. While one student searches, the other records the algorithm in plain language, and 3) *Assignment:* Rewrite the algorithm they wrote during the hands-on activity

with Pseudocode and implement it in a high-level programming language.

Day-2: 1) *Lecture:* Introduction to Matrix Multiplication, 2) *Hands-on activity:* Each group is to calculate a product of 2 NxN matrices and write down the steps in plain language, and 3) *Assignment:* Rewrite the algorithm they wrote during the hands-on activity with Pseudocode and implement it in a high-level programming language.

Day-3: 1) *Lecture:* Introduction to analysis of algorithms, 2) *Hands-on activity:* Each group is to compute the complexity of their matrix multiplication algorithm created on day 2, and 3) *Assignment:* Complexity analysis of PageRank Algorithm – The Mathematics of Google Search.

4.2 Differential Equations

Differential equations deal with applications making use of differentials. In order to introduce big data in this course, we felt it would be important to discuss the connection between data assimilation and big data. Data assimilation involves comparing a previous model of a state with newly obtained real observations and using this information to update the numerical model of the system.

The following three lectures were added to the existing differential equations course to infuse the big data concept:

- The connection between data assimilation and the discretization of the model state in the first lecture of the Big Data module. After a review of differential equations that cannot be solved by previously introduced methods, we have introduced numerical methods to approximate solutions of those equations. We used Euler's Method to solve linear differential equations.
- In the second lecture, we used the MatLab software to solve systems of ODEs (ordinary differential equations), effectively reinforcing the idea that discretization is a first step toward making a model or function suitable for numerical evaluation and implementation on a computer.
- In the third lecture, we introduced the basics of Monte Carlo simulation and as an exercise to teach the students how to draw a random number according to basic distributions using MatLab or another computer program.

4.3 Probability and Statistics

Statistical methods make up the majority of methods employed for understanding big data and making inferences.

We have created the following one-week module to enhance learning and expose students to big data:

- *Lecture:* The instructor presented the class with a formal definition of "Big Data" that best fits a statistical viewpoint. A brief review of the topics covered during the semester that are necessary for an understanding of the big data labs that follow was given. In addition, the students were introduced to the Python open-source statistical program. Python is a general-purpose programming language, and is more flexible and powerful than R, which is commonly used by statisticians for data analysis and modeling. Therefore, Python was selected as the instruction language.

- *First lab module:* In this lab, the main contents included random number generation as well as calculation of probabilities and expectations using Monte Carlo simulation. The lab used both simple and complicated examples. For simple examples, students were asked to compare the results of simulation experiments with the corresponding analytical solutions obtained using hand calculation.

- *Second lab module:* In this lab, the main contents included graphical visualization for some real data. Many datasets are publically available from sites such as kaggle.com and data.gov. Graphical visualization ranges from simple graphics such as histogram, boxplot, and scatterplot to advanced graphics such as PCA projection plots, trellis plots, maps, etc. were used. Students explored some real data using graphics to investigate and discover information from the real data.

- *Take-home project:* Students used simulation examples relevant to the real world including (a) gambling games, (b) biological evolution, (c) finance, (d) social network, (e) forensic science, etc. Depending on the students programming background, some template codes that are amenable to plug-and-play experimentation were provided to facilitate the activity and reduce the effort of writing a program. In this case, students were asked to examine and manipulate the python code.

4.4 Modern Geometry

The following are sketches of a one-week module consisting of outlines for three lectures and assignments for the big data topics in modern geometry.

Lecture-1: *A1-Intuitive introduction to topology and homology:* Define topology in terms of continuous and continuously invertible mappings and illustrate by examples of topologically equivalent spaces. Approach homology in terms of the number of holes of different dimensional spaces and give a formal definition of simplicial homology and a computation. *Assignments:* Visually classify different spaces first topological and then in terms of topology. Then do a

simple calculation of homology using linear algebra over the rational numbers.

A2-Bar codes: Introduce the idea of looking, for finite sets of points, of the geometric set where we expand the points to disks of radius 'r'. Define the image of homology classes under a continuous mapping and give examples of this. Study how its topology and homology changes as we change 'r'. Visually identify which cycles persists. *Assignments:* Given a set of data points and radii visually determine which cycles persist.

A3-Intuitive discussion of big data in general: Give a definition of big data and tell how it is used. Discuss how bar codes in particular might be used to study it. *Assignment:* Write a short report on some aspect of big data.

Lecture-2: *B1-Dealing with pictures on computers:* Show students how to upload pictures from computers and how to write programs to modify these pictures by mathematical transformations. *Assignment:* Take a picture with a cell phone and upload it to a computer. Then write a mathematical program to invert it. Lastly, write a mathematical program to change the colors.

B2-Projective transformations and vision: Review projective geometry and projective transformations and tell what geometric properties are preserved by projection and what change. Also tell how to represent a given 3-D scene, given in 3-dimensional coordinates, as it would be viewed from any angle. *Assignment:* Write out the formulas or a program to show how a simple 3-D scene (without overlaps) would be seen or displayed on a screen as viewed from a given angle.

B3-Experimentation with computers: Discuss some other types of transformations such as Mobius transformations, inversions, and conformal mappings. Relate this to the theory of map projections. *Assignment:* Take the picture stored in the computer and transform it by these mappings.

Lecture-3: *C1-General discussion of the problem of reconstructing a scene in 3-dimensions from flat pictures*: Give examples from the Internet where several different views of the same scene are available. Archaeologists use this to reconstruct what ancient buildings may have been like, such as temples. Real estate agents and companies can use it to give a 3-D model of a house they are selling. Tourist bureaus might want to give a visual tour of a city. Architects might want to have a 3-D model of what they are planning. Next discuss Magic Eye pictures (random dot stereograms), in which students can focus at the right distance on a picture of apparently random sets and see a 3-D scene. *Assignment*: Find an example of this 3-D reconstruction on the Internet.

C2-Mathematical discussion of the reconstruction problem in terms of projective geometry: Discuss aspects of the situation such as how a computer might divide the scene into objects (this ties in somewhat to the first module and topology). How could the computer tell that it is looking at the same objects when pictures are taken from different angles?

C3-In class assignments: Suppose Mathematically you are given a set of points in two flat pictures with a labelling of which point corresponds to which and necessary information about the points of view. Students would be asked to reconstruct the 3-D coordinates of each point.

5 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1436871. We are thankful for the discussion and contribution to the learning modules provided by the participants of the Big Data Analytics Workshop held at Alabama State University (ASU) on November 13, 2015.

6 Conclusions

We have infused one-week big data modules into three of our existing core undergraduate mathematics and statistics course and evaluated its effectiveness through pre- and post-tests. The modules were taught using examples that were worked through interactively during class. The students then worked on a programming assignment that incorporated the new instructional concept into concepts previously taught. This allowed the instructors to evaluate the students on their performance and to give feedback as to how they might improve. We feel the courses were a moderate success, but indicated there was room for improvement. A formal evaluation of these course modules is underway; we are quite optimistic that the big data based learning model will prove to be an effective approach to reinvigorating mathematics and statistics education for undergraduate computer science students.

7 References

[1] Sara Royster (2013), Working with big data, Occupational Outlook Quarterly, 57, 3, 2-10

[2] Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook, 2014-15 Edition, Mathematicians, on the Internet at http://www.bls.gov/ooh/math/mathematicians.htm

[3] http://blogs.hbr.org/2012/11/the-big-data-talent-gap-no-pan/

[4] Labor Force Characteristics by Race and Ethnicity, 2012, BLS Reports, October 2013, http://www.bls.gov/cps/cpsrace2012.pdf

[5] http://www.aps.org/programs/education/statistics/aamajors.cfm

[6] Hankerson, Darrel; Harris, Greg A. ; Johnson, Peter D., Jr. Introduction to information theory and data compression. Second edition. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[7] http://blogs.hbr.org/2012/11/the-big-data-talent-gap-no-pan/

[8] http://www.nasa.gov/press/2013/november/nasa-brings-earth-science-big-data-to-the-cloud-withamazon-web-services

[9] http://cas.umt.edu/math/

[10] http://www.ibmbigdatahub.com/blog/addressing-big-data-skills-gap

[11] http://www.data.gov

[12] http://aws.amazon.com/datasets

[13] http://figshare.com

[14] http://www.kdnuggets.com/datasets/index.html