Infusion of Big Data Concepts Across the Undergraduate Computer Science Mathematics and Statistics Curriculum

Dr. Carl Pettis, Alabama State University

Dr. Carl S. Pettis is a Professor of Mathematics at Alabama State University. He received his BS degree in 2001 and his MS degree in 2003 both from Alabama State University in Mathematics. Dr. Pettis received his PhD in Mathematics from Auburn University in 2006. He currently serves as the Interim Associate Provost for the Office of Academic Affairs.

Dr. Rajendran Swamidurai, Alabama State University

Dr. Rajendran Swamidurai is an Associate Professor of Computer Science at Alabama State University. He received his BE in 1992 and ME in 1998 from the University of Madras, and PhD in Computer Science and Software Engineering from Auburn University in 2009. He is an IEEE senior Member.

Prof. Ash Abebe, Auburn University

Ash Abebe is a professor of statistics at Auburn University. He received a B.Sc. in statistics from Addis Ababa University, Ethiopia, in 1995 and a Ph.D. in statistics from Western Michigan University in 2002. His research focus is on developing non-parametric statistical methods for analyzing complex data, especially those derived from spatio-temporal processes.

Dr. David Shannon, Auburn University

Dr. Shannon has a Ph.D. in Educational Research and Evaluation Methodology and Statistics from the University of Virginia and is currently the Humana-Sherman-Germany Distinguished Professor at AU. He teachers courses in research methods and program evaluation.

Infusion of Big Data Concepts Across the Undergraduate Computer Science Mathematics and Statistics Curriculum

1. Introduction

Stored digital data volume is growing exponentially [1]. Today, there are about 4.4 zettabytes (1 zettabyte is equivalent to 10^{21} bytes) of data in the World and it is expected to be about 44 zettabytes by 2020 [2, 3]. Society increasingly relies on such data to tell us things about the world [1]. Recent advances in technology, such as e-commerce, smart phones, and social networking, are the main reason behind this exponential data growth [1]. This large volume of structured and unstructured data is known as "big data" [1, 4]. Data is generated every rapidly. For example, in just one second, users are performing 40,000 search queries on Google, sending 520,834 messages on Facebook, and uploading 5 hours of video on YouTube on average [2].

The large increase in data opens up doors for new types of data analytics called big data analytics and new job opportunities [5]. The U.S. Bureau of Labor Statistics (BLS), Occupational Outlook Handbook 2018 [5] project that this large growth in data will create 34 percent more jobs from 2016 to 2026. The BLS's [6] report states that, "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 mathematicians 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] reports that 70% of the respondents plan to hire data scientists. McKinsey Global Institute's May 2011 [7] research report indicates that the demand for big data analytical talent could reach up to 490,000 positions (50 to 60 percent more data analytics jobs) in 2018.

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.

A recent survey from Harvard Business Review [9] indicates that 85 percent of the organizations that they surveyed planned to fill 91 percent of their data science jobs with new graduates [9]. Though the private sector asks at least a master's degree in mathematics or statistics for data analytics jobs, the government sector requires only a bachelor's degree [4]. Moreover, it is impractical to fill this huge demand for big data analytics through only from graduate degree holders in mathematics-related fields. The Harvard Business Review [9] reports that 70 percent of the surveyed organizations they described finding big data talents challenging to impossible. The hiring scale for big data jobs is 73; this high score indicates the amount of difficult in finding right candidates for that job [10]. In order to address this serious problem, Alabama State University, with the support of Auburn University, has begun infuse big data analytics in various undergraduate mathematics and statistics courses. Our big data course modules guide 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. We strongly believe that equipping students with such skills greatly improves their employability. This paper presents our three years' experience in adapting and integrating big data concepts across the computer science undergraduate mathematics and statistics curriculum.

2. Big Data and Mathematics

Linear algebra concepts such as manipulation of large matrices, matrix decomposition, and eigenvectors are extensively used in big data analytics. Manipulations of large matrices are used in feature extraction, clustering, and classification. Matrix decomposition is used in principal components analysis for dimension reduction. Similarly, application of eigenvectors used in Google's PageRank method.

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. Discrete mathematics 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), and matching in bipartite graphs (e.g. internet advertising) are widely used in big data analytics.

Differential equations explain the underlying dynamics in spatiotemporal pattern formation and detection, disease modeling, image visualization, processing, and analysis, etc. Differential equations topics such as numerical solutions systems of differential equations, nonlinear differential equations and stability, and using observed data to refine solutions are widely used in big data analytics.

Statistical methods make up the majority of methods employed for understanding big data and making inferences. Some example statistical topics used in 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.

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. Geometry topics such as data visualization and recovering low dimensional structures from high dimensional data are widely used in big data analytics.

3. Integrating Big Data Analytics in Existing CS Mathematics Courses

We have introduced big data modules in the following six Alabama State University undergraduate Mathematics and Statistics courses during the spring 2016, fall 2016, and spring 2017 semesters.

- MAT 251 Introduction to Linear Algebra
- MAT 256 Discrete Mathematics
- MAT 375 Elementary Differential Equations
- MAT 472 Probability and Statistics I
- MAT 473 Probability and Statistics II
- MAT 484 Modern Geometry

The big data concepts were infused into each of these courses in two parts: the theoretical and conceptual ideas behind the big data concept under discussion were introduced in the first part of the module; whereas, the hands-on experimentations were introduced in the second part of the module. The students are advised to use both R and Python general-purpose programming languages to complete their projects. The students can also use MALAB programing to perform their project.

3.1. Introduction to Linear Algebra

The following big data lectures and lab modules were infused to the existing linear algebra course:

Lecture: The instructor presented the class with a concept of "Big Data" that best suits linear algebraic viewpoint. The topics covered include linear equations and matrices, eigenvalues, eigenvector, and singular value decomposition.

Hands-on activities: 1) classify data sets into categories that describe the shape of the data distribution. For this lab activity students were encouraged to use the practical big data in linear equations about business problems, tax problems, economic planner's models, input-output matrix for economic producing transportation problems, and interpret data analytics problems, 2) explore some real data using eigenvalues, and eigenvectors to explore and discover information from the real data. For this lab, students were encouraged to solve the practical problems such as economic development problems, analysis of situations as diverse as land problems, applications in structural engineering, control theory problem, vibration analysis problem, electric circuits problem, and advanced dynamic problem and so on.

3.2. Discrete Mathematics

The following big data lectures and lab modules were infused to the existing discrete mathematics course:

Lecture: Introduction to algorithms, Introduction to Matrix Multiplication, and Introduction to analysis of algorithms.

Hands-on activities: For the hands-on activities the students were grouped into pairs. The activities are: 1) find a specific "key card" in a deck of random cards; while one student searches, the other records the algorithm in plain language, 2) calculate a product of 2 NxN matrices and write down the steps in plain language, and 3) compute the complexity of their matrix multiplication algorithm created on day 2.

Assignments: 1) Rewrite the algorithm they wrote during the hands-on activity with Pseudocode and implement it in a high-level programming language, 2) Rewrite the algorithm they wrote during the hands-on activity with Pseudocode and implement it in a high-level programming language, and 3) Complexity analysis of PageRank Algorithm – The Mathematics of Google Search.

3.3. Elementary Differential Equations

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

Lecture: The instructor presented the class with a concept of "Big Data" that best suits elementary differential equational viewpoint. The students were introduced to an open-source differential equational program.

Lab-1: In this lab, the main contents include first order differential equations. Students classified data sets into categories that describe the shape of the data distribution. Both simple and complicated examples were used in the lab. For simple examples, students were asked to compare the results of simulation experiments with the corresponding solutions obtained using formula. Students were encouraged to use the practical big data for differential equations such as simple chemical conversion problems, growth of population problems, price of commodities models, Newton's law of cooling problems, and many physical problems.

Lab-2: In this lab, the main contents include high order linear and non-linear differential equations, including homogeneous and non-homogeneous equations. Many datasets are publically available from sites such as data.gov. Students need to explore and discover information from the real data. Students also were encouraged to solve the practical problems such as vibration and resonance problems, including damped and undamped vibrations, Newton's laws of motion and gravity problems, Kepler's laws for planetary motion problems, and so on.

3.4. Probability and Statistics

The following big data lectures and lab modules were infused to the existing probabilities and statistics courses:

Lecture: The instructor presented the class with a formal definition of "Big Data" that best fits a statistical viewpoint. The topics covered are, sampling methods, measures of central tendency, symmetric and skewed data, and graphic visualization techniques such as histogram, boxplot, and scatterplot to advanced graphics such as PCA projection plots, trellis plots, maps, etc.

Lab-1: In this lab, the main contents include sampling methods, measures of central tendency, and symmetric and skewed data. Students classified data sets into categories that describe the shape of the data distribution. Both simple and complicated examples were used in the lab. For simple examples, students were asked to compare the results of simulation experiments with the corresponding analytical solutions obtained using hand calculation.

Lab-2: In this lab, the main contents include 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. Students need to explore some real data using graphics to explore and discover information from the real data.

Take-home project: Students were used some simulation examples relevant to the real world. Topics for recommendation include (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.

3.5. Modern Geometry

The following big data lectures and lab modules were infused to the existing modern geometry course:

Lecture: Intuitive introduction to topology and homology, bar codes, intuitive discussion of big data in general, dealing with pictures on computers, projective transformations and vision, representing a given 3-D scene in 3-D coordinates, discussion about Mobius transformations, inversions, and conformal mappings, general discussion of the problem of reconstructing a scene in 3-dimensions from flat pictures, and mathematical discussion of the reconstruction problem in terms of projective geometry.

Assignments: 1) Simple calculation of homology using linear algebra over the rational numbers, 2) Given a set of data points and radii visually determine which cycles persist, 3) write a mathematical program to invert a picture and change the colors, 4) 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, 5) Take the picture stored in the computer and transform it by Mobius transformations mappings.

Project: 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. Reconstruct the 3-D coordinates of each point.

4. Results

During the spring 2016, fall 2016, and spring 2017 semesters, Alabama State University faculty developed 31 big data modules and implemented them into the existing mathematics and statistics courses and evaluated its effectiveness through pre- and post-tests. In addition, students in all offered classes were asked to complete a survey pertaining to their coursework, confidence in using big data modules in their classes, and strategies they use to learn in their math classes.

4.1.Student Knowledge

Students in each class completed pre- and post-tests to examine changes over the duration of the module implementation. In each class, there were students that failed to complete the pre, post, or both tests. Overall, scores on the pre-tests averaged just 49.35% while averaging 78.06% on the post-tests. These results are summarized in the table-1 below.

TABLE I. STUDENT PERFORMANCE ON MODULE TESTS

Course	Computer Science (CSC) Majors				Mathematics (MAT) Majors				Other (OTH) Majors			
	Pre Avg.	Post Avg.	Pre Std.	Post Std.	Pre Avg.	Post Avg.	Pre Std.	Post Std.	Pre Avg.	Post Avg.	Pre Std.	Post Std.
MAT251	67	69.3	4.2	2.3	39.4	77.5	19.7	11.3	48.1	68.6	20.6	13.8
MAT256	54.2	91.7	29.2	12.9	25.3	80	10.2	20.9	33	87	18.4	12.8
MAT472	25	100	0.0	0.0	30	98.7	7.7	3.0	25.5	63.9	6.8	30.0
MAT473	25	62.5	0.0	17.7	57.5	73	35.5	33.1	37.5	77.8	13.2	29.2
MAT484	60	0	0.0	0.0	66	74.3	2.8	4.9	62.4	68.2	1.5	2.0
All Courses	46.2	80.9	6.7	6.6	43.6	80.7	15.2	14.6	41.3	73.1	12.1	17.6

The pre- and post-test average and standard deviation scores for computer science, mathematics and other majors in each course are shown in figure 1 and 2. Figure 3 and 4 shows the overall pre- and post-test average and standard deviation scores for computer science, mathematics and other majors.

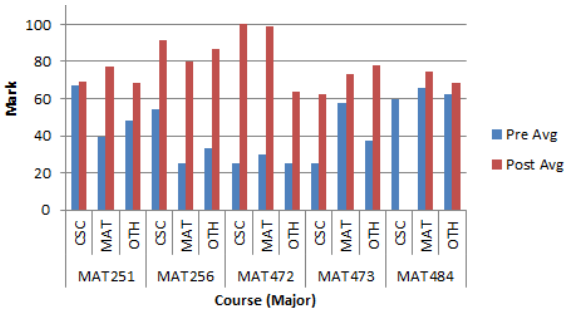

Fig.1. Pre-Test and Post-Test Averages

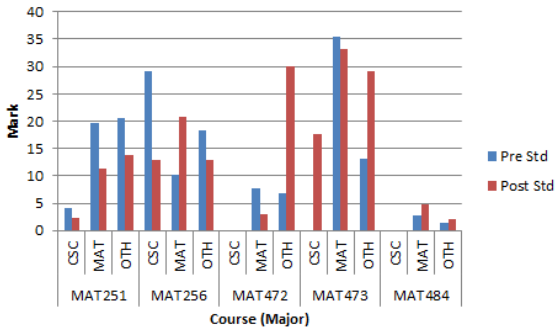

Fig.2. Pre-Test and Post-Test Standard Deviations

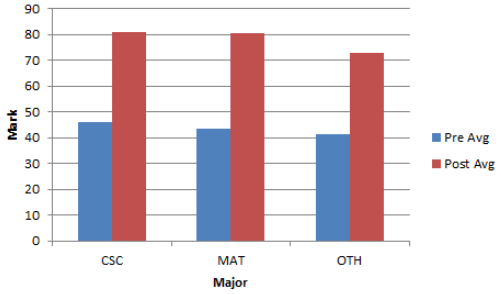

Fig.3. Overall Pre-Test and Post-Test Averages

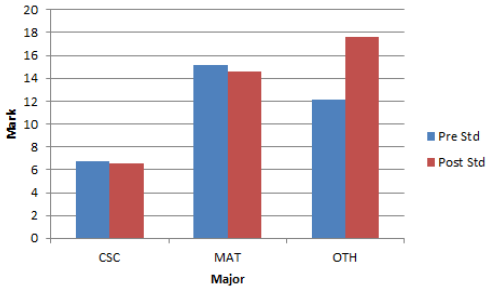

Fig.4. Overall Pre-Test and Post-Test Standard Deviations

4.2.Matched Pre-Post Student Knowledge

To better examine gains made by students after using these modules, the analysis was limited to those students with complete pre- and post-test data. A total of 52 students had completed both the pre- and post-test. Scores for this matched sample increased from pre-test (M=22.11, SD=10.7) to post-test (M=69.23, SD=33.4). Using a paired-samples t-test, changes from pre-test to post-test were statistically significant (t=10.76, p<0.001). These results are summarized in the table-2 below.

TABLE II. MATCHED PRE-POST STUDENT KNOWLEDGE

Course	Computer Science (CSC) Majors				Mathematics (MAT) Majors				Other (OTH) Majors			
	Pre Avg.	Post Avg.	Pre Std.	Post Std.	Pre Avg.	Post Avg.	Pre Std.	Post Std.	Pre Avg.	Post Avg.	Pre Std.	Post Std.
MAT251	64.0	68.0	0.0	0.0	39.4	79.0	19.7	11.9	29.7	73.7	27.6	14.6
MAT256	54.2	91.7	29.2	12.9	25.3	81.3	10.2	23.9	32.6	87.5	20.5	12.8
MAT472	25.0	100.0	0.0	0.0	28.0	98.7	6.7	3.0	26.0	66.9	7.1	28.4
MAT473	25.0	62.5	0.0	17.7	57.5	73.0	35.5	33.1	36.1	77.8	13.2	29.2
MAT484	0.0	0.0	0.0	0.0	66.0	75.5	2.8	6.4	61.7	68.7	1.5	1.2
All Courses	42.1	80.6	7.3	7.6	43.2	81.5	15.0	15.7	37.2	74.9	14.0	17.2

The matched pre- and post-test average and standard deviation scores for computer science, mathematics and other majors in each course are shown in figure 5 and 6. Figure 7 and 8 shows

the overall matched pre- and post-test average and standard deviation scores for computer science, mathematics and other majors.

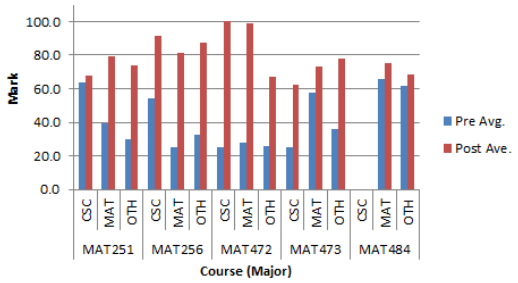

Fig.5. Matched Pre-Test and Post-Test Averages

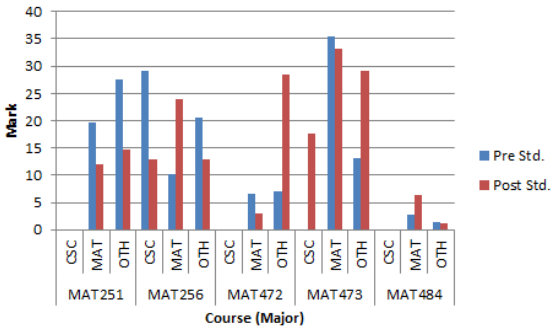

Fig.6. Matched Pre-Test and Post-Test Standard Deviations

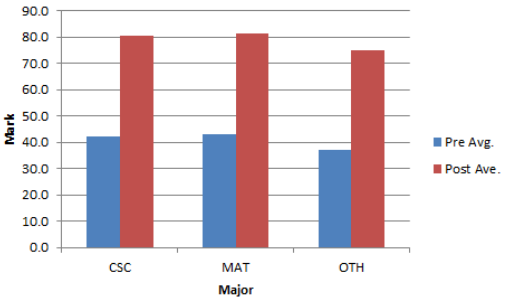

Fig.7. Overall Matched Pre-Test and Post-Test Averages

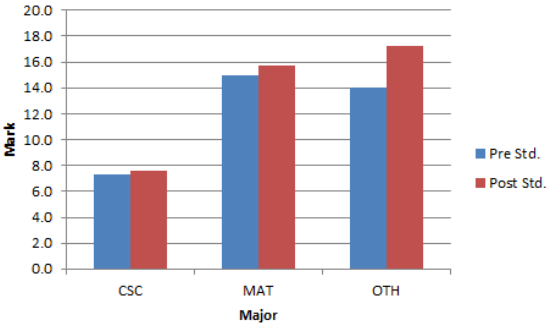

Fig.8. Overall Matched Pre-Test and Post-Test Standard Deviations

4.3.Confidence in Using Big Data Modules in Class

In spring 2016, nearly 80% of the overall survey respondents were either juniors or seniors and nearly 30% were enrolled as computer science majors. The sample was balanced in terms of gender (52.9% female), but offered little diversity in terms of race, ethnicity or disability. In Fall 2016, nearly 95% of the overall survey respondents were either juniors or seniors and over 38% were enrolled as computer science majors. The sample offered little diversity in terms of race, ethnicity or disability and over 32% were female. In spring 2017, nearly 95% of the overall survey respondents were either juniors or seniors and nearly 28% were enrolled as computer science majors. The sample had a larger number of males (53.1%), with majority of participants identifying as Black (87.5%) and primarily not identifying with Hispanic or Latino ethnicity (90.6%).

Using a 5-point scale (1=little of no confidence…5=A great deal of confidence), students were asked to respond to 31 different potential big data modules/applications. These responses were requested prior to the implementation of modules in math coursework. In spring 2016, only 8 out of 26 modules (30.8%) received an average response of 3 or above, in fall 2016, only 2 out of 26 modules (6.5%) received an average response of 3 or above, and in fall 2017, 30 out of 31 modules (96.8%) received an average response of 3 or above.

4.4.Student Academic Efficacy, Motivation and Learning Strategies in Math Courses

Finally, students were asked to respond to survey items pertaining to their level of academic efficacy, motivation and goals in learning math, and strategies that they use and prefer to learn math.

- Academic Efficacy: Students were asked to respond to five items related to their academic efficacy as it pertains to the math class in which they were enrolled. Overall, students reported a great deal of confidence in their academic abilities with the average for each tem above 4 (on a 5-point scale). Students believed that they would learn if they tried, worked hard, and did not give up. They also believed that they could master the skills and figure out the most difficult class work.
- Goals in Math: While all goals were important to them, students believed that getting a good grade was most important. They also wanted to meet requirements for their degree, improve their ability to communicate math ideas to others, learn new ways of thinking and specific procedures for solving math problems.
- Preferred Learning Environments: When asked to indicate their perceptions of statements describing different learning environments, students reported the greatest agreement with "the instructor explains the solutions to problems" and "the assignments are similar to the examples considered in class." Students also indicated situations in which they compared their math knowledge to other students, studied their notes, explained ideas to others, worked in small groups, and got frequent feedback on their mathematical thinking. They were less supportive of having the class critique their solutions, exams that prove their skills and group presentations.
- General Learning Strategies Used by Students: In general, students reported using a variety of strategies in their math classes and not giving up when they get stuck. They most frequently reported finding their own ways of thinking and understanding and reviewing their work for mistakes or misconceptions. They also reported checking their understanding of what a problem is asking, studying on their own and using their intuition about what an answer should be.
- Motivation to learn Math - Task Value: Students reported high levels of task value, indicating their belief in the importance and utility of course content in their math classes. Their understanding of math is extremely important to them and their motivation to learn math is strong.
- Learning Strategy – Critical Thinking: In terms of learning math, students reported many strategies that require critical thinking. They reported developing their own ideas based on course content and evaluating the evidence before accepting a theory or conclusion. They also reported questioning what they read or hear in class and thinking or possible alternatives.
- Learning Strategy – Self- Regulation: Students reported using many effective self-regulation strategies in their math classes. In particular, they pay careful attention to concepts that they find confusing and focus of studying and reviewing these so they learn them.
- Learning Strategy – Time and Study Environment Management: Another positive strategy reported by students related to the management of their time and study environment. They reported attending class regularly, finding a place to study and keeping up with the weekly readings and assignments.

The reliability of these scales was generally supportive, with internal consistency estimates ranging from .491 to .926, with a median of .867. Perceptions were also very positive as overall scale means exceeded the scale midpoints.

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 two Big Data Analytics Workshop held at Alabama State University (ASU) on Aug 10, 2015 and November 13, 2015.

6. Conclusions

We have created about 31 one-week big data modules and infused them into seven existing core undergraduate mathematics and statistics courses over a period of two years. The modules were taught using examples that were worked through interactively during class. The students then worked on assignments that incorporated the new big data instructional concepts. We have evaluated the big data modules effectiveness through pre- and post-tests, and surveys. The paired-samples t-test results show that matched pre-post student knowledge is statistically significant. Regarding confidence in using big data modules in class, we had mixed results. Students' perception was very positive as overall scale means exceeded the scale midpoints. We feel the courses were a success, but indicated there was room for improvement.