ISSN: 2473-3857
v3 n4330

Data Analytics Workshop Series for Non-Computing Major First-Generation-College-Bound Students

Sam Chung
School of Information Systems and Applied Technologies
Southern Illinois University
Carbondale, IL 62901, USA

Abstract

The purpose of this research is to propose how we can encourage non-computing major first-generation-college-bound students to be actively involved in learning data analytics. Non-computing major students have limited opportunities to take a data analytics related course. The computing major programs have the resource limit for offering none-major electives. The first-generation-college-bound students need more mentoring for their future directions. For this purpose, we challenge the following three goals: 1) What practices can we use with underpinning scientific evidences? 2) How can we enhance engagement of students with very limited interventions from an instructor? And, 3) how can we motivate the participation of the non-computing major students? To accomplish these goals, we start with developing a series of summer workshops that is Evidence-Based Practices (EBP)-guided, student-driven, and applied. Based upon the EBPs, we develop a series of student-driven summer workshops, not a regular elective course, with an emphasis on application of data analytics to the main areas in which the students are interested. We conclude that the EBP first helped us to develop these workshop series for applied data analytics with underpinning scientific evidences. Second, these workshops using active learning methods allowed the students to have their strong engagement with very limited interventions from an instructor. Third, these workshops motivated the participation of the non-computing major students by influencing the students to seek how data analytics can be applied to their domain. Last, the outcome of their team project allowed them to experience undergraduate research.

Keywords: Data Analytics, Non-Computing Major, First-Generation-College-Bound, Evidence-Based Practice, Active Learning Methods, Experiential Learning.

1. INTRODUCTION

In the era of Big Data, Data Science becomes an emerging subject across disciplines. Because of the multidisciplinary and interdisciplinary subject, it is important for students to learn concepts of data analytics, earn skills of data analysis, and then apply the skills to their real word problems regardless of their major (Wymbs, 2016; Wright, 2016).

Although students in a computing major such as Computer Science, Information Technology, and Information Systems can take a data analytics course as a regular class, most of non-computing major students have limited opportunities to take a data analytics related course. Since many computing programs have to teach pre-existing courses, it is not easy for them to open a data analytics course for non-major students. In addition, underrepresented students such as first-generation-college-bound students or socioeconomically disadvantaged students have limited resources of understanding their demands for their advanced study of professional careers in emerging areas (National Academic of Sciences (NAS), 2011; The Executive Office of the President, 2014).

To solve the constraints of non-computing major students, the resource limit of the computing major programs, the lack of mentoring of the first-generation-college-bound students, we propose an approach that focuses on first-

generation-college-bound student, not instructor but student-driven and non-computing majors. However, we encounter three challenges: 1) What scientific evidence-based practices can we use? 2) How can we enhance engagement of students with very limited interventions from an instructor? And, 3) how can we motivate the participation of the non-computing major students?

For those three challenges, we start with identifying best practices that have sound scientific evidences to make a course very friendly for the students who are non-computing major and first-generation-college-bound. Based upon the identified Evidence-Based Practices (EBP), we first choose the workshop series format for a small group during summer (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). Second, to encourage the students drive the workshop series, we infuse active learning methods and a classroom assessment technique into each workshop module. Third, to be attractive to non-computing majors, we allow students to apply data analytics to their real problems and present their peer-reviewed research outcomes at a local conference. Fourth, we assess the workshop series in terms of formative and summative evaluation. Then, we discuss what we learned through the series of summer workshops. Last, we conclude that we could encourage non-computing major first-generation-college-bound students to be actively involved in learning data analytics.

2. BACKGROUND

Evidence-Based Practice (EBP)

To select a correct format for non-computing major first-generation-college-bound students, we adopt Evidence-Based Practice (EBP), which came from Medicine. According to Dr. David L. Sackett, who is known as one of the fathers of Evidence-Based Medicine (EBM), EBM means clinical practices should be conducted based upon the best available external scientific clinical evidences (Sackett, et al., 1996; Straus, Glasziou, Richardson, Haynes, & Sackett, 2011). Software Engineering research community adopted EBM and proposed Evidence-Based Software Engineering (EBSE) (Kitchenham, Dyba, & Jorgensen, 2004; Dyba, Kitchenham, & Jorgensen, 2005).

Active Learning Methods – FC, JiTT, & PI

An active learning method allows each student to be more directly involved in his or her leaning process. Each student is a learner as connection-maker, content producer, and sharer. We use three active learning methods: Flipped Classroom (FC), Just-in-Time Teaching (JiTT), and Peer Instruction (PI).

FC is an active learning method that reverses typical lecture in classroom to homework activities in classroom (Bergmann & Sams, 2012; Rutherfoord & Rutherfoord, 2013). Students are required to study lecture materials at home by watching short videos and solving short quizzes before the class session while they are devoted to discussions, exercises, or projects in class. JiTT is an active learning method that allows an instructor to adjust lecture materials within a short time before a lecture begins (Novak, Patterson, Gavrin, & Christian, 1999; Gurka, 2012; Martinez, 2012). PI is an active learning method that encourages student to help other students by discussing an answer of a given concept test (Mazur, 1997; Mazur and Watkins 2010; Simon and Cutts, 2012). The main point of a concept test is to confirm whether the students in class truly understood one important concept that the student should take away from the class.

Classroom Assessment Techniques (CAT's)

CAT's are strategies that allow instructors to conduct formative assessment in order to assess how well students are learning key concepts during class time (Angelo and Cross, 1993). Among CAT's, we use the muddiest point that allows students to have their own chance to briefly describe what part of the lesson or the assignment in class was most confusing to them.

3. RELATED WORK

Data Science becomes an emerging subject across disciplines for both undergraduate and graduate degree programs. Data Analytics has been one of important Body of Knowledge (BoK) of Data Science. A Data Science degree program was proposed and implemented for an interdisciplinary undergraduate degree (Anderson, Bowring, McCauley, Pothering, & Starr, 2014). An interdisciplinary data analytics track and its minor for undergraduate students were proposed (Wymbs, 2016). A Master's of Science degree in Data Analytics was also proposed for business major students (Jafar, Babb, & Abdullat, 2016). A data analytic centric MS Degree was proposed for Information Sciences and Technologies students (Kang, Holden, & Yu, 2014; Kang, Holden, & Yu, 2015). An online graduate program in Information Security and Analytics was proposed (Kumar, 2014).

However, most efforts above focused on a degree program with regular courses for either undergraduate or graduate students. The degree program considered multidisciplinary and interdisciplinary subjects. Some degree programs emphasized applied data analytics with advanced skills (Kang, Holden, & Yu, 2014; Kang, Holden, & Yu, 2015). All degree programs stay with a specific major such as Computer Science, Information Technology, and Information Systems. We could not find any explanations of teaching data analytics for either non-computing majors or underrepresented students.

4. RESEARCH METHOD

EBP in Data Analytics – Workshops

We apply EBP to data analytics education to identify unique values and preferences (Dyba, Kitchenham, & Jorgensen, 2005; Straus, Glasziou, Richardson, Haynes, & Sackett, 2011). As scientific evidences, our literature surveys show that underrepresented students including first-generation-college-bound students need proven and intensive interventions in Science, Technology, Engineering, and Mathematics (STEM) (NAS, 2011). Bettinger and Baker (2014) found that the students who received mentoring kept higher retention.

Based upon NAS recommendations (2011), we propose a series of summer workshops, which are not an elective course for the student participants. Instead of a regular semester, we target a summer semester. We include engagement in networking, peer-to-peer support, study groups, and social activities. Due to the small group setting, we constantly provide them with mentoring.

We also require research experiences, participation in conferences, and presentation of research. Students should apply what they learned in data analytics to their application to experience applied data analytics. In addition, since faculty resources are very limited during summer semester, we educate and train two IT major senior students before summer semester through an independent study. Then, we have the students lead the workshop as their credits for their second independent study course.

Curriculum – Active Learning Methods

In order to enhance engagement of students, we infuse three active learning methods and the muddiest point into each teaching module. Each teaching module employs the same cycle of active learning method sequence: 1) JiTT with the muddiest point and FC out of classroom, 2) JiTT and PI in classroom, and 3) FC in classroom.

Before starting next class, students were required to study their reading assignment for next workshop. We designed this reading assignment for two purposes: engagement and JiTT. Because the purpose of this assignment was not evaluation but engagement, we allowed unlimited trials for maximum five multiple-choice quizzes. To prepare for taking the quiz, we published lecture-related materials such as links to video clips, articles, and slides, etc. to the Canvas Learning Management System (LMS).

To check whether the students actually conducted their reading assignment and to know what the students could not understand clearly from the given teaching materials, we required a reading assignment quiz. The quiz consists of two types of questions: one for multiple-choice quizzes to check whether the students studied or not before next class, the other for an essay quiz to receive feedback from the students. We designed the last essay question to know the muddiest point of the given reading materials.

Before starting each workshop one day earlier, the instructor (one of two leading students) evaluated the submitted reading assignment and identified what part the students answered correctly or incorrectly from the multiple-choice quizzes and what they really got confused or wanted to learn from the muddiest point question. Then, the instructor adjusted his topics that he will cover in class. After explaining the confusing or interesting topics, the instructor gave a concept test to the students in class. A concept test consists of a quiz with multiple choices. Each student selected an answer and then discussed his or her answer with another classmate who answered differently. Then, the instructor explained the right answer to the students.

After conducting a concept test in the middle of class, the instructor conducted classroom activities for lab assignment. Students practiced R programming with an Integrated Development Environment (IDE), RStudio. Then, they started their labs in class and could continue to finish it out of class. For next class, the instructor requested the students to start the reading assignment for next workshop again after class.

Evidenced Approach – Team Project

At the end of the workshop series, students were required to apply what they learned to a problem

ISSN: 2473-3857
v3 n4330

domain of their major trough a team project. Before starting their team project, the students studied a sample case during Weeks 9 and 10. Then, they could apply what they learned to their problem trough a team project. If necessary, we allowed them to finish after week 10.

5. RESULTS

Teaching Modules

We developed the workshop series of data analytics in Spring 2015 with two IT major students. Then, by using the developed workshop series, the IT students conducted a 10-week workshop by tutoring seven non-major students for their second independent study course. We used first nine chapters of Jared Lander's book for R programing with RStudio (Lander, 2013). Table 1 shows the topics of 10 teaching modules. Figure 1 shows that the 10 teaching modules were uploaded onto the Canvas LMS.

Table 1. Teaching Modules for Workshops

W	Teaching Module	Resource
1	Active Learning Methods Getting R	Syllabus Ch. 1
2	R Computing Environment	Ch. 2
3	R Packages	Ch. 3
4	Basic Math Functions	Ch. 4
5	Data Structure	Ch. 5
6	Reading Data & Visualization	Ch. 6 & 7
7	Function Definition & Call	Ch. 8
8	Control Structure	Ch. 9
9	Clustering	Online
10	Classification	Online

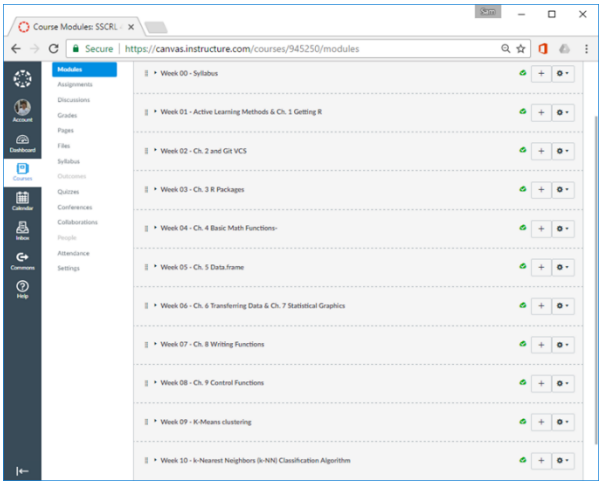

Figure 1: Canvas LMS for Summer Workshop Series

At week 1, students were informed how this workshop class will be conducted and how active learning methods will be integrated into each teaching module. During next seven weeks, students focused on how to use RStudio and R programming. We started with packages to explain how R can be easily extensible for other business domains. Then, we introduced foundation of R programming - objects, language-defined methods, basic data structures, visualization, user-defined methods, and control structures.

For data clustering and classification, we used online resources (Influxity, 2013; Jalayer Academy, 2015a; Jalayer Academy, 2015b) with data samples from University of California Irvine (UCI) Machine Learning Repository (UCI MLR, 2017). We introduce two well-known k-means clustering and k-nearest neighbor algorithms to the students and practiced the algorithms with Iris Data Set at weeks 9 and 10.

Weekly Workshop

Two IT major students took an independent study course for preparation in spring semester and led the workshop during summer. By using week 9 as an example, we (two students leading the workshop) explain how we conducted a weekly workshop. Table 2 shows which week we took an active learning method and where we exercised the method - out of class or in class.

Table 2. Active Learning Methods (CAT) & Teaching Materials (Where, Where)

Active Learning Methods (CAT)	Teaching Materials (When, Where)
FC (MP)	Course Material for W9 Reading Assignment Quiz for W9 (W8, out of class)
JiTT	Adjusted lectures for W9 (W9, out of class)
PI	Concept Test for W9 (W9, in class)
FC	Lab Assignment for W9 (W9, in class)
FC (MP)	Course Material for W10 Reading Assignment Quiz for W10 (W9, out of class)

ISSN: 2473-3857
v3 n4330

FC: Flipped Classroom, MP: the Muddiest Point; JiTT: Just-in-Time Teaching; PI: Peer Instruction

At Week 9, the students learned a clustering algorithm with the k-means R package. We held the workshop on Saturday from 9:30 AM to 12:30 PM. There was a reading assignment quiz at the end of Week 8 for the students to be engaged in Week 9 class. We provided the students with course materials and video clips. We required the students to take a reading assignment quiz including the muddiest point question. Before starting the workshop one day earlier, we checked the reading assignment quiz to know what concepts we need to explain in next class. After conducting the JiTT, we required the students to take a concept test quiz in class for PI. The students themselves discussed and answered the concept test first and we explained its answer later. Then, the students started hands-on labs in class and could continue the lab after class. We required the students to answer several questions at the end of each lab to confirm whether the students truly understood the lab – why you did this lab. We repeated this cycle throughout 10 weeks.

Three Team Projects

At the end of the workshop series, the students applied what they learned to their application domain trough a team project. Table 3 shows the three-team projects (total 7 non-computing major students). The cases show that the students could apply the Machine Learning approach to their problem domains - identification of automotive vehicles, on-time performance of airline operations, and smoking effect on newborn babies.

Table 3. Applied Data Analytics Projects

C	Majors	#	Title
1	Automotive Technology	2	Vehicle Clustering by Manufacturer Region Using the K-means Clustering Machine Learning Algorithm
2	Aviation Mgmt. & Flight	3	K-means Clustering of Airline On-Time Performance Statistics
3	Physiology & Biochemistry	2	Effect of Smoking on Newborn Weight and Length at Birth by Using the K-Means Clustering Algorithm

C: Case; #: the number of participants

In Case 1, the team sought to find how a machine learning approach using the k-means clustering algorithm can be applied to vehicle clustering. The vehicles were clustered into North America, Europe, and Asia regions in terms of engine displacement and Mile Per Gas (MPG). The team discovered that vehicles from the Asian region were surprisingly clustered quite differently from the actual data set because the vehicles were clustered as European vehicles rather than Asian (Figure 2).

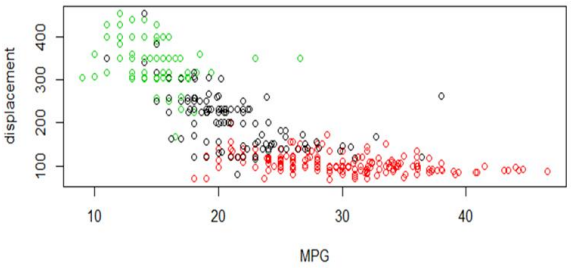

Figure 2. Scatterplot results of clustering engine displacement and MPG according to manufacturer regions (green for North America, black for Europe, and red for Asia) (Chung et al. 2017)

In Case 2, the team analyzed airline on-time performance statistics and found that the lowest clustering group reflecting the low on-time performance rate is mainly associated with winter and summer months (Figure 3).

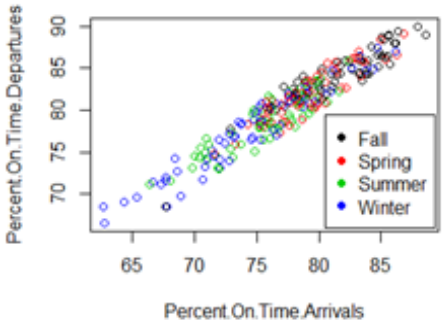

Figure 3. Scatterplot results of clustering flight departures and percentages of on-time arrivals according to four seasons (Chung et al. 2017)

In Case 3, the team analyzed the relationships of the height and weight of infants from non-smoking mothers and mothers with regular contact with cigarette tobacco. The team confirmed that there is a significant difference in the height and weight variables when maternal smoking is a contributing factor (Figure 4). More detail information of three case studies can be found in Chung et al. (2017).

ISSN: 2473-3857
v3 n4330

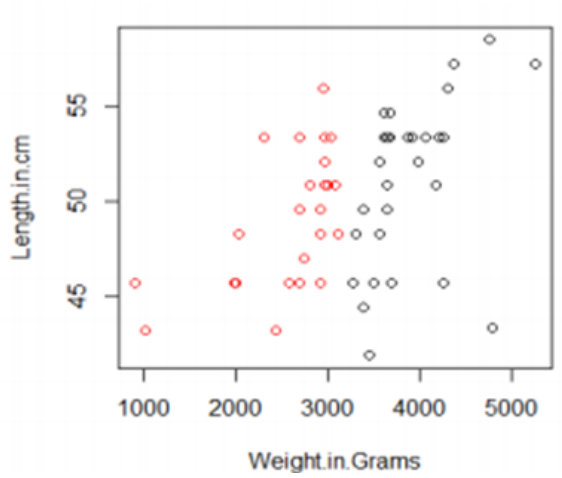

Figure 4. Scatterplot results of clustering length and weight of infants according to maternal smoking (red) and non-smoking (black) (Chung et al. 2017)

6. EVALUATION AND DISCUSSION

Formative Evaluation

Based upon our observation, the percentages of submitting both the reading assignment quiz before next class and the lab assignment quiz after class were not high because of their distraction due to their other priorities. However, they were strongly engaged in the concept test and lab in class if they did not miss their workshop. Later, we allowed them to finish both the reading assignment and the lab assignment quizzes in class. Then, the percentages increased.

Summative Evaluation

We required each student to be a part of a team project and apply what he or she learned about data analytics to their problem. All students successfully finished their term project. Evan after they finished the workshop series, they worked together since we strongly recommended submission of an abstract to a peer-reviewed student poster session at a local symposium. All three teams participated in submitting their abstract to a university-hosted symposium in Fall 2015. The symposium organizer accepted all of three abstracts after peer review and invited two teams for oral presentation and one team for poster presentation in November 2015 at the symposium.

What We Learned

First, we answer what practices that have underpinning scientific evidences we could use. We chose the summer, student-driven, and applied workshop format for non-computing major first-generation-college-bound students. The proven practices such as active learning, learning by doing, and research experiences were integrated into the teaching modules.

Second, we answer how we could enhance engagement of students with very limited interventions from an instructor. While the intervention from a faculty member was minimized, the interventions among the students were maximized. Two leading students, who were IT major, were educated and trained first through their independent study for one semester (academic learning). Then, they led the workshop series with limited interventions from the instructor for other non-computing major students (experiential learning). They could also learn how the integration of active learning methods into teaching modules could improve student engagement and learning.

Third, we answer how we could motivate the participation of the non-computing major students. The non-computing major students could learn data analytics by applying what they learned to their own data analytics problem. The summer workshop series could support networking, peer-to-peer support, study groups, and social activities. Due to the small group setting, we constantly provided them with mentoring. The inclusion of undergraduate research experiences based upon the outcomes of their team projects helped them to learn how creative thinking, technical writing, and professional presentations are important.

7. CONCLUSION & FUTURE WORK

We propose how we can encourage non-computing major first-generation-college-bound students to be actively involved in learning data analytics. EBP guided us to choose the teaching format, summer workshop, to implement each teaching module with three active learning methods, one CAT, and one team project. The team project allowed the students apply data analytics to their own business domain and experience research and creative activities through a student poster with their abstract. This approach could enhance engagement of the students and motivate the participation of the students.

A series of summer workshops for data analytics that are student-driven brought several benefits to either computing-major students who led the workshops or the non-computing major first-generation-college-bound who participated in the

workshops. The leading students had both academic learning through their preparation of the workshops and experiential learning through the workshop series. The non-computing major first-generation-college-bound students could experience hands-on learning and learning by doing through the applied team projects

For next steps, we will expand the workshop series to more diverse groups of students. The current population is limited to male and one ethnic group. Also, we will analyze how these workshop series could help retention, pathways, and graduation of the students participated since we have kept connections with all of nine participant students (2 IT major and 7 non-IT major) since October 2015.

8. REFERENCES

Anderson, A., Bowring, J. McCauley, R., Pothering, G., & Starr, C. (2014). An undergraduate degree in data science: curriculum and a decade of implementation experience, *Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE '14)*. Atlanta, GA. March 5-8, 2014. 145-150.

Angelo, T. A., & Cross, K. P. (1993). Classroom assessment techniques: A handbook for college teachers. *San Francisco: Jossey-Bass*.

Bergmann, J., & Sams, A. (2012). *Flip Your Classroom. Reach Every Student in Every Class Every Day.* Washington, DC: International Society for Technology in Education.

Bettinger, E.P., & Baker, R. (2014). The Effects of Student Coaching: An Evaluation of a Randomized Experiment in Student Advising. *Educational Evaluation and Policy Analysis*, 36(1), 3-19.

Chung, S., Oh, T. J., Kim, J. Kang. A. (2017). Data Analytics Workshop with First-Generation-College-Bound, Underrepresented, and Non-Computing Major Students. *Consortium for Computing Sciences in Colleges Midwest Conference (CCSC MW 2017)*, September 22-23, 2017. Grand Rapids, MI.

Dyba, T., Kitchenham, B. A., & Jorgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE software*, *22*(1), 58-65.

Gurka. J. S. (2012). JiTT IN CS 1 AND CS 2. *Journal of Computing Sciences in Colleges*, Volume 28, Issue 2, December 2012, p. 81-86.

Influxity. (2013) How to Perform K-Means Clustering in R Statistical Computing, Retrieved June 15, 2017. https://www.youtube.com/watch?v=sAtnX3UJyN0

Jafar, M. J., Babb, J., & Abdullat, A. (2016). Emergence of Data Analytics in the Information Systems Curriculum. In *Proceedings of the EDSIG Conference ISSN* (Vol. 2473, p. 3857).

Jalayer Academy. (2015a). R - kNN - k nearest neighbor (part 1), Retrieved June 15, 2017. https://www.youtube.com/watch?v=GtgJEVxl7DY

Jalayer Academy. (2015b). R - kNN - k nearest neighbor (part 2), Retrieved June 15, 2017. https://www.youtube.com/watch?v=DkLNb0CXw84

Kang, J. W., Holden, E. P., & Yu, Q. (2014, October). Design of an analytic centric MS degree in information sciences and technologies. In *Proceedings of the 15th Annual Conference on Information technology education* (pp. 147-152). ACM.

Kang, J. W., Holden, E. P., & Yu, Q. (2015, September). Pillars of Analytics Applied in MS Degree in Information Sciences and Technologies. In *Proceedings of the 16th Annual Conference on Information Technology Education* (pp. 83-88). ACM.

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004, May). Evidence-based software engineering. In *Proceedings of the 26th international conference on software engineering* (pp. 273-281). IEEE Computer Society.

Kumar, S. A. (2014, October). Designing a graduate program in information security and analytics: master's program in information security and analytics (MISA). In *Proceedings of the 15th Annual Conference on Information technology education* (pp. 141-146). ACM.

Lander, J. (2014). R for Everyone: Advanced Analytics and Graphics. Upper Saddle River, New Jersey: Addison-Wesley.

Martinez, A. (2012). Using JiTT in a Database Course, the Proceedings of the 43rd ACM Technical Symposium on Computer Science Education (SIGCSE'12), p. 367-372.

ISSN: 2473-3857
v3 n4330

Mazur, E. (1997). Peer Instruction: A User's Manual. Prentice Hall.

Mazur, E. & Watkins, J. (2010). Just-in-Time Teaching and Peer Instruction. Just-in-Time Teaching, edited by Simkins, S. and Maier, M., Stylus, Publishing, LLC, Sterling, VA.

National Academic of Sciences (NAS). (2011). Expanding Underrepresented Minority Participation. The National Academic Press. Washington, D.C.

Novak, G., Patterson, E., Gavrin, A., & Christian, W. (1999). *Just-in-Time Teaching: Blending Active Learning and Web Technology*. Prentice Hall.

Rutherfoord, R. H., & Rutherfoord, J. K. (2013, October). Flipping the classroom: Is it for you? In *Proceedings of the 14th annual ACM SIGITE conference on Information technology education* (pp. 19-22). ACM.

Sackett, D. L., Rosenberg, W., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ* 312, 71072.

Simon B., & Cutts, Q. (2012). How to implement a peer instruction designed CS principles course. *ACM Inroads*, Volume 3, Issue 2, June 2012. pp. 72-74.

Straus S. E., Glasziou, P., Richardson, W. S., Haynes, R. B., & Sackett, D. (2011). Evidence-Based Medicine: How to Practice and Teach It, 4th edition. Churchill Livingstone, Edinburgh, 2011, p.1

The Executive Office of the President. (2014). Increasing College Opportunity for Low-Income Students. The White House. January 2014.

UCI MLR. (2017). The UC Irvine Machine Learning Repository, Retrieved March 17, 2017. http://archive.ics.uci.edu/ml/

Wright, A. A. (2016). Jobs of the Future Will Require Data Analysis. *Society for Human Resource Management*. Retrieved March 17, 2017 from https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/jobs-of-the-future-will-require-data-analysis.aspx

Wymbs, C. (2016). Managing the Innovation Process: Infusing Data Analytics into the Undergraduate Business Curriculum (Lessons Learned and Next Steps). *Journal of Information Systems Education,* 27(1), 61-74.