Big Data Analytics - With an Infusion of Statistics for the Modern Student

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.

Dr. Cadavious M. Jones, Alabama State University

Dr. Cadavious M. Jones is an Associate Professor of Mathematics at Alabama State University. He received his BS in 2006 and MS in 2008 from Alabama State University, and PhD in Mathematics from Auburn University in 2014. He is a contributor to the Australian Maths Trust, and member of the MASAMU international research group for mathematics.

Dr. Carl Pettis, Alabama State University

Carl S. Pettis, Ph.D. Professor of Mathematics Department of Mathematics and Computer Science Alabama State University

Administrative role:

Interim Provost Office of Academic Affairs Alabama State University

Dr. Uma Kannan, Alabama State University

Dr. Uma Kannan is Assistant Professor of Computer Information Systems in the College of Business Administration at Alabama State University, where she has taught since 2017. She received her Ph.D. degree in Cybersecurity from Auburn University in 2017. She specialized in Cybersecurity, particularly on the prediction and modelling of insidious cyber-attack patterns on host network layers. She also actively involved in core computing courses teaching and project development since 1992 in universities and companies.

Big Data Analytics: with an infusion of statistics for the modern student

1. Introduction

Recent technological advancements in various fields such as e-commerce, smart phones, and social media generate huge volumes of data on a scale never seen before [1]. New data are generated every second. For example, every second on average 40,000 search queries are performed on Google; 520,834 messages are sent by Facebook users; and 5 hours of video are uploaded to YouTube [2]. The digital data generation and storage volume has been growing exponentially [1, 3] and the growth is observed in every sector including but not limited to government, healthcare, banking, manufacturing, retail, transportation, and education [3]. There are about 4.4 zettabytes (1 zettabyte is equivalent to 1021 bytes or 270 bytes) of data in the World. 90% of this data were created in the past two years alone. By 2020 the data volume is expected to be 44 zettabytes (or 44 trillion gigabytes) [2, 4]. We are continuously relying on data to tell us things about the world [1].

According to the U.S. Bureau of Labor Statistics (BLS), Occupational Outlook Handbook 2018 [5] this large increase in available data from the Internet will open new jobs up to 34 percent in the area of big data analytics from 2016 to 2026. McKinsey Global Institute's May 2011 [3] research report indicated that the demand for big data analytical talent could reach up to 490,000 positions in 2018. The same report also indicated that 50 to 60 percent more qualified data analytic professionals would be needed by 2018. Harvard Business Review [4] indicates that 70 percent of the organizations that they surveyed are finding it difficult to hire data scientists. The hiring scale for big data jobs is 73; this high score indicates the amount of difficulty in finding the right candidates for that job [6]. In 2011, the EU (European Union) public sector and US health-care sector planned to use data and analytics to improve their services and reduce errors. As on today, only 10 to 20 percent of the opportunities planned in 2011 have been realized by both the sectors due to the shortage of technical talent. [7]

In recent years, employment for mathematics related occupations increased by almost 4 percent, yet over the same period of time, the number of degrees conferred in math, statistics, and engineering declined by 2 percent [3]. A recent survey from Harvard Business Review indicated that big data initiatives are underway in 85 percent of the companies they surveyed. These organizations also indicated that they planned to fill 91 percent of their data science jobs with new graduates [6]. Though the private sector requires at least a master’s degree in mathematics or statistics for data analytics jobs, the government typically requires only a bachelor’s degree [5] for similar job postings. Moreover, it is impractical to fill this huge demand for big data analytics via a candidate pool that consist solely of candidates who have graduate degrees in mathematics-related fields.

The current undergraduate mathematics courses help students develop their logical thinking and problem-solving skills. The statistics courses introduce students to methods of data collection, organization, analysis, and interpretation. Big data analytics requires three key talents from STEM graduates [3]: 1) they must have deep analytical talent (possess statistics and machine

learning skills); 2) they must be data-savvy managers and analysts (have the ability to develop the right questions for analysis, interpret and challenge the results, and make the right decisions); and 3) they need the support of technology (to develop, implement, and maintain big data software and hardware tools).

This paper presents our experience with infusing, teaching, and assessing big data modules in undergraduate statistics and probability courses that have immersed 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 statistics and probability course module.

2. Big Data and Statistics & Probability

Statistics is a discipline that is concerned on designing experiments and other data collection, summarizing information to aid understanding, drawing conclusions from data, and estimating the present or predicting the future [8]. Statistics has long been the avenue for answering important research questions and statistics has evolved with the advent of computers. The growth in computational statistics research has allowed statisticians to fit more complex models than ever before and improving inference using bootstrap and cross-validation methods. [9] Computational statistics methods such as Markov processes and the Markov transition matrix (e.g. Web surfing), correlations in high dimensional data, the Bonferroni Principle, and Monte Carlo simulation are employed for understanding big data and making inferences.

It is important for statistics to be one of the key disciplines for Big Data, because: 1) statistics is fundamental to ensuring meaningful, accurate information is extracted from Big Data, 2) statistics brings sophisticated techniques and models to address the issues that might arise due to data quality, data volume, missing data, and uncertainty during prediction, forecasting and modeling, and 3) statisticians help translate the scientific question into a statistical question, which contains the data model, data structure, and the parameters we are trying to assess or predict. [10]

Probability distributions are the basics for many big data models. The big data models are based on 1) partition function or normalized probability density function such as Gibbs distributions, 2) unnormalized probability density function such as Markov chain Monte Carlo methods, or 3) approximate Bayesian computation (ABC) methods [11, 12].

Based on the sheer volume of data produced each year data mining has become a cornerstone of analytics. Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include but not limited to market segmentation, customer churn, fraud detection, direct marketing, interactive marketing, market basket analysis, and trend analysis. [13]

3. Infusing Big Data Analytics in Existing Undergraduate Statistics and Probability Courses

We introduced big data modules in the two undergraduate Statistics and Probability courses, *MAT 472 Probability and Statistics I* and *MAT 473 Probability and Statistics II*, during the spring 2016, fall 2016, spring 2017, fall 2017, and spring 2018 semesters.

Our undergraduate Statistics and Probability course, *MAT 472 Probability and Statistics I* and *MAT 473 Probability and Statistics II* began with this as the core idea of its lecture and modules. This allowed the students an opportunity to examine and process raw data using basic techniques already learned through the course or those closely related to topics outlined in the syllabi. Students explore variable selection using intuitive approaches such as correlations. The hope was to produce students, who were more capable of examining large unfiltered data sets in order to provide meaning solutions to questions of concern.

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 experimentation was 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 programming to perform their project, and depending on the level of programming experience each student has, Excel to a lesser extent.

We began by examining the curricula for both *MAT 472 Probability and Statistics I* and *MAT 473 Probability and Statistics II* to determine what concepts the students would naturally be introduced to throughout the course and their relationship to big data analytics. The students received a pretest prior to any discussion pertaining to big data analytics in each course.

Since this is an introductory course and its extension, students were beginning to understand key concepts of enumeration and the applications of both discrete and continuous distributions. We focused on filtering through raw data sets in order to remove erroneous or unnecessary information as related to the questions of concern. Once students had a filtered data set, they would employ data sampling techniques and descriptive statistics to summarize several samples of the population to infer a more accurate solution(s) pertaining to the questions of concern.

In the second course, we continued with data mining as its core idea, but students had a greater understanding of how simple actions can produce enormous amounts of data and how to filter the raw data. By the time, a student has reached the second set of modules and lectures they have an understanding of simple regression and maximum likelihood. During this, lecture and relating modules students learned Monte Carlo methods and how this technique relies on repeated random sampling. Students also filter through raw data sets to produce histograms to find solutions for questions of concern. To conclude each one week lecture and module a post test and survey were given to gauge the students' growth.

In order to understand as completely as possible the student's competency, certain standards were considered. There were ten standards addressed to some degree by this project. The standards are: Students will be able to filter raw data sets, select reasonable sampling methods given the

data at hand, use descriptive statistic techniques; display data in a graphical manner; determine standard error, develop models; determine levels of accuracy needed; organize materials; interpret the data and draw a conclusion from the data; explain their thought process. The procedure for this is provided for each course in later subsections.

The criteria which identified indicators of good performance on the task and in class discussions were:

- Accuracy of removing erroneous or unnecessary information from data sets
- Accuracy of calculations
- Accuracy of model and graphs or charts
- Organization of calculations
- Clear explanations

With our criteria for success defined, we now examine each course individually.

3.1. MAT 472 Probability and Statistics I

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

During the lecture, students received a brief review of the topics previously discussed during the semester that would be necessary for an understanding of the labs. This further provided a practical connection to Big Data Analytics. A formal definition of “Big Data” that best fits a statistical viewpoint that was provided was an "n by p" structure, meaning n observations on p variables, thus large n and large p. Depending on the progression of the lecture a variety of topics covered, beginning with 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.

Each course received one lab and a discussion/results presentation period associated with the aspect of data mining we were to examine. The lab for the *MAT 472 Probability and Statistics I* course revolved around applying several of the sampling methods discussed during the lecture, on data previously filtered through by the students from its raw form. Students would then take these filtered data set samples and use descriptive statistics techniques that involved measures of central tendency, symmetric and skewed data analysis. Students then classified data sets into categories that described 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.

The next stage of the infusion was the take-home project. Students were provided with some simulation examples relevant to the real world. Topics for recommendation included (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 were amenable to plug-and-play experimentation were provided to facilitate the activity and reduce the effort of writing a program. Those who wished to write their own programs were strongly

encouraged to do so. In both cases, students were asked to examine and manipulate the python code provided.

During the discussion and review session, students would compare their results and discuss open-ended questions that related to the project and data mining. "Starter" questions were listed on the modules as, “Class questions for discussion.”

The raw data sets for the *MAT 472 Probability and Statistics I* lab and take-home project was gathered from the websites of local business, such as car companies and the stock markets among other sources to provide a base understating of data analysis. Students later examined any large data set of their choice using sites such as www.data.gov.

3.2. MAT 473 Probability and Statistics II

The structure of the *MAT 473 Probability and Statistics II* course’s infusion of the big data lecture and lab modules to utilizes the existing probabilities and statistics course material mirrored that of the *MAT 472 Probability and Statistics I* course.

Once again students received a brief review of the topics previously discussed during the semester that would be necessary for an understanding of the labs. However, students would also be reminded of the methods of sampling, measures of central tendency, and symmetric and skewed data, in addition to the techniques to filter raw data. Here the students were introduced to more graphic visualization techniques and the class of computational algorithms known as Monte Carlo methods. They were also instructed as to how such methods could be applied to the field of data mining (and big data analytics).

In the second lab for the *MAT 473 Probability and Statistics II* course, the main contents included graphical visualization for some real data. Given that many datasets are publically available from sites such as kaggle.com and data.gov, large data sets were easily applied to the techniques utilized. Once again, graphical visualization ranged from simple graphics such as histogram, boxplot, and scatterplot to advanced graphics such as PCA projection plots, trellis plots, maps, etc.

For the take-home project students were again proved some simulation examples relevant to the real world. In one example students were to play the game the game of darts using n number of darts, both by hand and by using a python program. Using this method of dart throwing they would create large amounts of data that could be used to calculate pi using a Monte Carlo simulation. The students continued to use Monte Carlo simulations for data analysis by using them to run simulation to play the game of battleship, for which they generated win/loss data, and displayed data using histograms

During the discussion and review session, students would compare their results and discuss open-ended questions that related to the project and data mining. "Starter" questions were listed on the modules as, “Class questions for discussion.”

The discussion and review session would also mirror the structure of that used in the *MAT 472 Probability and Statistics I* course.

4. Results

From the spring 2016 to spring 2018 semesters, University faculty developed big data modules and implemented them into the existing statistics and probability courses and evaluated its effectiveness through pre-, post-tests, projects and adhering to outlined criteria for success. In addition, students in all participating courses were asked to complete a survey pertaining to their coursework, confidence in using big data modules in their classes, and strategies they learned 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 30.8% while averaging 72.16% on the post-tests.

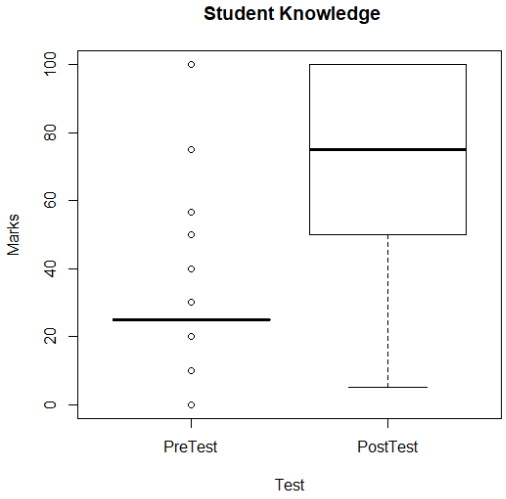

Figure-1: Box Plot for Unmatched Pre-Post Tests

The Student t-test results are shown figure 2 below:

Unpaired* t *test results

P value and statistical significance:

The two-tailed P value is less than 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

Confidence interval:

The mean of PreTest minus PostTest equals -41.3533

95% confidence interval of this difference: From -51.6175 to -31.0891

Intermediate values used in calculations:

t = 7.9885

df = 105

standard error of difference = 5.177

Review of data:

Group	PreTest	PostTest
Mean	30.8036	72.1569
SD	21.8419	31.2611
SEM	2.9188	4.3774
N	56	51

Figure-2: Student t-Test Results (Student Knowledge)

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 54 students had completed both the pre- and post-test.

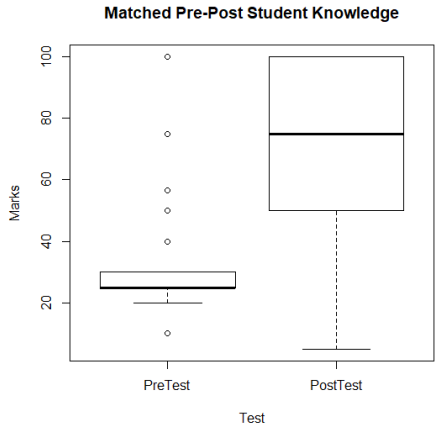

Figure-3: Box Plot for Matched Pre-Post Tests

The Student t-test results are shown figure 4 below:

Unpaired* t *test results

P value and statistical significance:

The two-tailed P value is less than 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

Confidence interval:

The mean of PreTest minus PostTest equals -39.5985

95% confidence interval of this difference: From -49.0669 to -30.1302

Intermediate values used in calculations:

t = 8.2916

df = 106

standard error of difference = 4.776

Review of data:

Group	PreTest	PostTest
Mean	33.6420	73.2406
SD	20.3252	28.6094
SEM	2.7659	3.8932
N	54	54

Figure-4: Student t-Test Results (Matched Pre-Post Student Knowledge)

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 term 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. A more detailed summary of items from these scales are shown in Table 1.

Measurement Scale	Items	Reliability	Mean (SD)
Academic Efficacy[a]	5	.864	4.16 (.8)
Goals in Math[b]	10	.920	4.26 (1.13)
Preferred Learning Situations[c]	11	.869	5.42 (1.56)
Learning Strategies used in class (general)[d]	15	.890	5.35 (1.43)
MSLQ- Motivation - Task Value[e]	6	.909	5.71 (1.21)
MSLQ – Critical Thinking[e]	5	.888	5.05 (1.46)
MSLQ – Self-Regulation[e]	11	.821	5.049 (1.45)
MSLQ – Time and Student Environment Management[e]	8	.491	4.87 (1.61)
a=5-point scale (1=Strongly Disagree…5=Strongly Agree) b=7-point scale (1=Not at all important …7=Extremely important) c=7-point scale (1=Strongly Disagree…7=Strongly Agree) d=7-point scale (1=Very Seldom…7=Very Often) e=7-point scale (1=Not True of Me…7=Very True of Me)			

5. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers 1436871 and 1247525. 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 on Aug 10, 2015 and November 13, 2015.

6. Conclusions

We have created one-week big data modules and infused them into existing undergraduate statistics and probability courses over a period of three 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.

An issue that arouse due to our students only being trained, in their basic computing courses, in C++ but not python. Their computer programming knowledge did not allow for all students to easily write or alter simple code for analyzing data. Another issue was the availability of software, students did not have access to MatLab, and Excel for example, on their person computers. Our solution was to use to open source software such as Open Office/Libre Office, and octave. We also used already programmed applications of finished code in python or excel) to perform Monte Carlo simulations on phenomena (basketball free throws, darts to estimate pi) by entering some parameters.

References

1. Sara Royster, “Working with big data,” Occupational Outlook Quarterly, 57, 3, 2-10, 2013
2. Bernard Marr, “Big Data: 20 Mind-Boggling Facts Everyone Must Read,” Forbes, September 30, 2015, https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#322cdf0017b1
3. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, June 2011
4. Ralph Jacobson, “2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?,” IBM, April 24, 2013, https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
5. Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook, Mathematicians and Statisticians, on the Internet at https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm (visited January 30, 2018)
6. Paul Barth and Randy Bean, “There’s No Panacea for the Big Data Talent Gap,” Harvard Business Review, November 29, 2012, https://hbr.org/2012/11/the-big-data-talent-gap-no-pan
7. Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy, “The age of analytics: Competing in a data-driven world,” McKinsey Global Institute, December 2016
8. Data Science, Statistics, Mathematics and Applied Mathematics, Operations research, and Astronomy @ Unisa: A complete guide to preparing yourself for career opportunities, University of South Africa
9. Ross Sparks, Adrien Ickowicz and Hans J. Lenz, “An Insight on Big Data Analytics,” Springer International Publishing Switzerland 2016, Japkowicz and J. Stefanowski (eds.), Big Data Analysis: New Algorithms for a New Society, Studies in Big Data 16, DOI 10.1007/978-3-319-26989-4_2
10. Roger Peng, “Statistics and Big Data,” The American Statistical Association
11. Blum, M. G. and Tran, V. C., “HIV with contact tracing: a case study in approximate Bayesian computation,” Biostatistics, 11(4) 644–660, 2010
12. Franke, B., Plante, J. F., Roscher, R., Lee, E.A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., and Reid, N, “Statistical Inference, Learning and Models in Big Data,” International Statistical Review, 84: 371–389, 2016, doi: 10.1111/insr.12176.
13. Doug Alexander, “Data Mining,” University of Texas, https://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/