An Undergraduate Data Science Program

Jim Albert and Maria Rizzo

Abstract

Bowling Green State University has created a new undergraduate data science program within a department of mathematics and statistics. This program is a synthesis of courses in mathematics, statistics and computer science designed to prepare the students for opportunities in data science. We describe the new data science and statistical learning courses and discuss challenges in the development of this program.

1 Introduction

1.1 The Modern Statistician

Brown and Kass (2009) describe the challenges of meeting the statistical needs of science, technology, business, and government. Given the exponential increase in the collection of "big data," there is a tremendous demand for people who have the ability to perform exploratory statistical analyses and draw inferential conclusions from these data. But the statistics curricula at universities has been very slow in reacting and changing to meet the new demands of the modern statistician. The curriculum for most degrees in statistics focuses on learning existing methodologies for handling different types of data and models, say time series, categorical, ANOVA, and regression models. Brown and Kass argue that this type of training prepares students to be short-term consultants able to answer well-defined statistical questions from a practitioner. This short-term consultant typically matches one of the statistical methods he/she has learned to the problem and learns little about the nuances of the applied problem. However, modern statistical methodology is deeper and broader than in the past, and a modern statistician really only has strong expertise in a few well-developed methodology subdomains. The demands for statisticians have changed. It is more common currently for a statistician to work together with people from other fields in interdisciplinary research. In this new collaborative role, the statistician must understand the scientific or technology application very well, and be able to apply flexible problem-solving strategies to the statistical problems in this research.

Brown and Kass argue that the primary goals of statistical training should be to help students develop statistical thinking and to encourage work on cross-disciplinary projects. Statistical thinking can be defined as the use of probabilistic descriptions of variability in inductive reasoning and analysis of procedures for data collection, prediction, and inference. By placing a value on cross-disciplinary activities, the modern statistics department is decreasing the

importance of particular theoretical and methodology courses and encouraging quantitatively oriented students from other applied departments to learn statistical thinking.

Nate Silver has received much recent attention in the media due to his success in predicting presidential elections on his `fivethirtyeight.com` blog. Mr. Silver was recently asked "For aspiring applied statisticians, what do you think are the best and hottest new skills to learn and add to one's resume?" Silver responded by saying "the most important thing is just to lessen the amount of book-learnin' that you do and start to play around with some data sets instead." This comment reinforces the belief that statistics training must include substantial experience exploring data.

1.2 Teaching Statistical Computing

Since the discipline of statistics has gone through dramatic changes, Nolan and Lang (2010) describe how the curricula and education culture of statistics must change. Nolan and Lang describe three key components of the education of the modern statistician. First, the meaning of statistical computing must be broadened so that the statistician can access and integrate large amounts of data, use data mining methods to explore complex data, and produce interesting statistical presentations using new technologies. Second, statisticians need a deeper understanding of computational reasoning. This computational reasoning includes knowledge of fundamental concepts of programming languages, and the ability to express statistical methods by writing computational algorithms. Last, statistical computing must be taught in the context of statistical practice. Students should work on computational problems arising from the acquisition, statistical analysis, and reporting of interesting and relevant data.

1.3 New Training for the Modern Statistician

Traditionally, students get introduced to statistics through a one-year sequence in probability and mathematical statistics. One obtains a foundation in univariate and multivariate distributions, and sampling results such as the Central Limit Theorem. Inferential methods are introduced by properties such as unbiasedness, mean square error, size and power, and one learns the standard inferential procedures based on the assumptions of normality and equal variances.

This type of statistics training was appropriate at a time when standard inferential procedures such as regression and ANOVA were commonly applied in the practice of statistics. But with the common occurrence of large datasets with many irregularities such as non-normal population shapes, missing data, and multivariate response and covariates, many of the standard statistical inferential procedures are not suitable. When faced with these large datasets, the modern statistician may have to devise innovative statistical methods to address the new descriptive and inferential problems. To successfully develop new exploratory and inferential methods, we believe that the statistician needs new types of training at the university level.

The following is a list of topics not currently being taught or emphasized in the traditional statistics curriculum that should be learned by the modern statistician.

- **Exploratory data analysis**. The traditional statistics curriculum does not emphasize exploratory methods. Specifically, the modern student needs to understand the fundamental aspects of looking for patterns in data, including the importance of reexpressing variables, the need to look at residuals from a fitted model, and the use of resistant procedures insensitive to small changes in the data. Although these concepts are described by Tukey (1975) in the context of simple data structures, all of the exploratory principles play important roles in understanding high dimensional data.

- **Statistical graphics**. Patterns in high-dimensional data are often best communicated by effective graphical displays rather than tables. Students should understand the fundamental principles in creating effective statistical graphics. Also they should have knowledge of the basic graphical methods for working with categorical and measurements data and experience in creating graphs using a modern statistical computing language to achieve particular goals.

- **Statistical modeling**. The traditional statistics curriculum focuses on the general linear model which assumes that the mean response is a linear function of the predictors and the errors are normally distributed. The modern student needs to think of modeling from a more general perspective. For example, generalized linear models extend the normal linear model to handle responses that are proportions, counts, or survival times. The student needs exposure to general nonparametric regression models helpful for representing nonlinear patterns between response and covariates.

- **Data mining**. "Big datasets" typically have a large number of observations or cases collected on a large number of variables which can be numeric or categorical or ordinal in nature. Methodologies for extracting patterns from these large datasets are collectively known as data mining. The modern student needs familiarity with a variety of data mining methods such as principal components, cluster analysis, variable-selection algorithms for large regression problems, classification and ensemble methods, and visualization of high-dimensional data.

- **Simulation-based inference**. With the advent of high-speed computing, bootstrap and other randomization methods are straight-forward to implement by simulation on a computer. These methods provide general inference strategies when the standard inferential methods are inappropriate. The students should be familiar with the fundamentals of Markov chain Monte Carlo methods that are useful for fitting a large class of statistical models with potentially hundreds of parameters.

- **Introduction to programming**. The modern statistician needs a solid background in basic programming concepts. Given the ready availability of "big data", the statistician will need to write scripts to download and manage data from the Internet. It is unlikely that this type of raw data can be analyzed using traditional statistics methods that are included in standard statistical packages, and the statistician will typically need to write special scripts in his/her pre-processing and analysis.

- **Statistical programming using R.** The modern statistician needs to work in a flexible computing environment. In this environment, one needs the availability of a variety of statistical methodologies to explore and draw inferences from data. This environment should support the construction of statistical graphs using familiar graphical methods. In addition, this environment should support a scripting language so the statistician can design and implement new methodologies and new graphs to communicate the results of statistical studies. The popular R statistical system is currently the most popular interactive environment for statisticians and it is important that the student gets experience using R in much of the coursework.

1.4 Data Science Curriculum

Hardin et al (2015) describe a curriculum shift to include more data science topics into the undergraduate statistics program. To motivate the importance of data science proficiency, they provide case studies from four institutions describing different approaches to teaching data science. The case studies vary in terms of audience and scope, but they introduce strategies for introducing "computing with data" curricula into a statistics or computer science major. This article is part of a special issue in *The American Statistician* on "Statistics and the Undergraduate Curriculum". In this special issue, Horton and Hardin (2015), Chamandy et al (2015), Nolan and Lang (2015), and Baumer (2015) describe special implementations of data science material.

Donoho (2015) provides a general overview of the discipline of data science, describing a broad academic field of data science that he calls "Greater Data Science" or GDS. The six divisions of GDS are data exploration and preparation, data representation and transformation, computing with data, data modeling, data visualization and presentation, and science about data science. In Data Exploration and Preparation, one explores the data in various ways to discover anomalies and artifacts, find basic patterns, and expose unusual features. In Data Representation and Transformation, one finds suitable reexpressions that help in finding patterns. Simple examples of these expressions are square roots and logs of count and amount data, and logits for proportion data. The Computing with Data unit applies R or some other language to manipulate and summarize the data – this work could be called data analysis together with computational work (such as working in clusters and cloud computing) so that the data analysis operations are performed efficiently. The Data Visualization and Presentation unit involves more than basic graphs – how can one visualize the

data efficiently when there are many variables to explore? The Data Modeling division includes the methods of statistical and machine learning and the Science of Data Science studies the relevant data science topics in the workspace.

2 The Data Science Specialization

To respond to the need for training for new types of skills of the modern statistician, we wrote a grant proposal to the National Science Foundation to develop a new undergraduate data science major at Bowling Green State University. This grant was funded in the Fall of 2013; the first year of the grant was devoted to the development of the new program and recruiting the first group of data science majors with scholarships provided by the grant.

This new program was formally included in the undergraduate catalog in the Fall of 2015 as a specialization of the mathematics major. As in all degree programs in mathematical sciences in the department, the program requires a foundation of three calculus courses and elementary linear algebra. Similar to the statistics major, the data science degree requires a calculus-based probability course, and a course in linear regression. In addition, the program requires a course in programming fundamentals taught in the computer science department. Table 1 displays a typical four-year plan for the required courses in mathematics, computer science, statistics, and data sciences in the program. Four new data science courses were developed as a part of the new program: MATH 2960 Seminar in Data Science, MATH 3430 Computing with Data, MATH 3440 Statistical Programming, and MATH 4440 Statistical Learning. The specific content of these courses will be described in Section 3.

Table 1: Four-year plan of required courses in the Specialization in Data Science.

Year 1	Calculus 1	Calculus 2
	Seminar in Data Science	Introduction to Programming
Year 2	Calculus III	Linear Algebra
	Computing with Data	Statistical Programming
Year 3	Probability and Statistics 1	Statistical Learning
	Regression Analysis	Data Science Elective 1
Year 4	Capstone Experience	Capstone Experience
	Data Science Elective 2	

3 The Data Science Curriculum

3.1 The Freshman Data Science Seminar

Although students are familiar with "big data" in some way, they are not very knowledgeable about the discipline of data science. The purpose of the one-hour seminar on data science MATH 2960 is to provide a broad overview of a number of data science applications. In a typical seminar, a person on or off campus is

invited to describe a particular data science application. The invited speakers represent a range of applications such as sports analytics, GPS and spatial data, data visualization, data mining methods, quality systems, forensic science, GIS systems, biological sciences, geography, and text mining in linguistics research. Each student is expected to attend each seminar and write a report on one of the data science applications.

3.2 The Sophomore Data Science Courses

3.2.1 Prerequisites and Pedagogy

In the sophomore-level sequence MATH 3430 and MATH 3440, the students have not yet had formal coursework in statistics. Thus the course material and activities must be designed to be accessible to students who have not been exposed to a typical one year course in mathematical statistics.

The two courses were taught for the first time in Fall of 2015 and Spring 2016. Rather than a single textbook, a reading list of several books has provided the references. The books were available through our library in e-book format, so that this was not a financial burden for the students. The software used was R with RStudio IDE, free for students to download to their personal laptops.

During the first week of the course, students were expected to get their software installed and to become familiar with the RStudio IDE for R. Most of the class work and assignments were set up in R Markdown files, and converted into reports using the knitr package. This approach allows the report and the R code to be blended seamlessly into a single HTML, PDF, or Word document. It provides self-documenting examples of each topic and exercise, and results in what is essentially a collection of chapters that together make up a textbook-like notebook for all of the course topics.

For a typical class, a lesson is set up for interactive student work starting with a Markdown (.Rmd) template that has been customized with background information and examples for the current lesson. As the instructor works through the class examples, students add the code to implement each step, and the instructor explains any new concepts or tricky implementation details along the way. Most lessons end with a few exercises for students to complete in class or after class. The class examples and exercises are then submitted online usually in the HTML format output of their finished report.

The topics selected for illustrating many of the concepts can vary, but the topics outlined below have been chosen to give data science students a good toolbox of skills covering some of the most useful techniques that they may be likely to need in future work.

3.2.2 MATH 3430: Computing with Data

The BGSU Fall 2016 catalog description for MATH 3430 is as follows:

> Computational methods for collecting, manipulating, exploring, and graphing data. Basic principles of exploratory data analysis and statistical graphical methods. Methods for downloading and organizing

large data sets. All of the computing methods will be illustrated using a high-level language such as R or Python. Prerequisite: CS 2010 and C or better in MATH 1310 or MATH 1340 and MATH 1350.

General comments

This course begins with a general introduction to the use of a scripting language in data analysis. One introduces different data types (numerical, character, logical, etc.) and containers with a focus on data frames. Manipulations with data frame are introduced using the `filter`, `arrange`, `select`, `mutate`, and `summarize` verbs that are part of the `dplyr` package. Once data is arranged in data frames, then one introduces basic exploratory methods for summarizing and graphing one and two variables.

The course introduces different methods for reading and writing datasets. One begins with creating comma, separated values (csv), reading these files in R, and writing csv files to a webserver. Later, one introduces reading streams of textual data, and then data available as html tables and xml files. Since textual data is usually dirty, this motivates the use of regular expressions to perform substitutions and other manipulations on character strings.

Since visualization is an important aspect of data science, graphical data analysis is an important component of the course. Basic graphical methods for one and two variables are introduced together with the use of more sophisticated systems (ggplot2) to graph data in panels according to values of a third variable.

This course demonstrates computing with data in the context of interesting applications. The students work in small groups on a graphing activity using the rich PitchFX database of variables collected on pitches thrown in Major League Baseball. The students explore patterns of words of Twitter conversations using search tags on subjects of interest. In a final project, the students work with a larger dataset with many variables and suggested explorations. Sample "large" datasets were running times of three years of the Boston Marathon, data about on-time performance for a large number of airline flights, data about every shot for every player in golf tournaments in the 2014 PGA season, trip records for a large number of trips completed by taxis in New York City, data on victim based crime in the Baltimore city area, and data on a large number of trips made by people using the Pronto bike-sharing system in Seattle.

Although there was no general text for the course, portions of different books were used for specific topics. O'Neil and Schutt () and Nolan and Lang () were used to introduce data science, and a few chapters from Baumer, Kaplan, and Horton () were used for specific data science topics.

The general order of topics in this first data science course is presented below.

Weekly Topics of MATH 3430 Computing with Data

1. Introduction to data science.
2. Vectors, main variable types, and operations in R
3. Data frames in R

4. Basic data analysis methods (histograms, frequency tables, barplots, scatterplots, comparing groups by dotplots and boxplots)
5. Basic data wrangling using the R dplyr package
6. Statistical graphics using the base graphics system in R
7. Creating csv files and uploading the files to a web server
8. Data visualization, Part 2 (ggplot2 package in R)
9. Graphing project using baseball data from the PitchFX system.
10. Textual data — Exploring words in a Presidential inaugural speech
11. Working with regular expressions
12. Reading in HTML tables and data from XML pages
13. Mining Twitter data
14. Working with mapping data (basic mapping functions in R and spatial visualization with ggmap and ggplot2)
15. Exploring a larger dataset project.

3.2.3 MATH 3440: Statistical Programming

The BGSU Fall 2016 catalog description for MATH 3440 is as follows:

> Applying a statistics programming language to facilitate the exploration and visualization of data. Basic objects such as data frames, matrices, tables, and lists, and how to perform manipulations with these objects. Writing functions with looping and conditional structures. Use functions to perform simulation-based statistical algorithms. Understand and develop object-oriented programming. Develop manipulations with character data. Prerequisite: MATH 3430.

This data science course builds on the foundation of the Computing with Data course. After some review of basic syntax of the R system, the course begins with a discussion of writing functions, commands for iteration, and helpful summary functions for matrics and arrays. When programming, it is important to write efficient code, and methods for benchmarking and comparing algorithms are described. Database methods are introduced by constructing a SQL backend with the `dplyr` package. There is an overview of computational aspects of regression and smoothing statistical methods. The use of R to implement resampling methods such as the bootstrap are described. The course concludes with an introduction to time series methods, working with financial data, and integrating C++ code with R code.

Weekly Topics of MATH 3440 Statistical Programming

1. Getting started: RStudio, knitr, reproducible research, Help system
2. Basic syntax of R, useful R functions to know
3. Writing functions; arguments, return values, loops, vectorization
4. Working with arrays, matrices; sweep, apply
5. Tables, histograms, binning data
6. Simulating data from a given model
7. Simulation experiments
8. Benchmarking code and comparisons
9. Computational efficiency: efficient vs non-efficient methods
10. Working with larger data sets: dplyr and SQL
11. Simple linear regression, residuals, prediction intervals
12. Non-linear models: polynomial, exponential, etc.
13. Multiple regression models
14. Comparing the fit of different models
15. Nonparametric regression: lowess, loess, supsmu, etc.
16. Predictions from smoothers
17. The CDF, ECDF, and resampling
18. Introduction to ordinary bootstrap
19. Bootstrap estimation of bias and standard error
20. Bootstrap confidence intervals, using boot and boot.ci
21. Introduction to time series data analysis
22. Working with dates and the R Date object
23. Financial data: import from web, analyze, plot
24. Financial data: comparing two or more series
25. Use Rcpp to execute C++ code directly from an R script

By the end of the course, one has progressed from writing simple R code to organizing complicated programming tasks and using sophisticated programming techniques. Students have learned to make their functions re-usable and as general as possible, using default arguments, the `...` argument, and various types of return objects. They learned how to take advantage of existing R functions and investigate their output, for example using the return value from `hist` to write a function to bin data. Along the way students learned to study how certain R functions are implemented, study the structure of complex objects and the return values of functions. They learned how to find source code and available methods using utilities like `getAnywhere` and `methods`.

To prepare students for working with big data, we have also focused on the idea of computational efficiency, how to measure it and several strategies to improve the execution time. The Rcpp package is introduced to speed operations in compiled C++ code when necessary. Also helpful tools such as `dplyr` and SQL are introduced for working with large data sets.

3.3 MATH 4440: Statistical Learning

The statistical learning course was created as an integral component of the data science program. The prerequisite for the course is the calculus-based probability course MATH 4410. As the catalog description indicates, the course is a survey of statistical methods for supervised and unsupervised learning tasks including linear regression, classification, nonparametric regression, tree-based methods, classification methods, and principal component analysis. The text by James et al (2013) seems suitable for our vision of this statistical learning course since it focuses less on the mathematical details and more on the applications of the methods. One attractive feature of this text is the inclusion of labs in each chapter where the student gets experience working with the book data using the R system.

3.4 Applying Data Science

3.4.1 Capstone Experience

One important component of the data science program is a capstone experience where the student works on a large dataset from a particular application. There are several ways for a student to complete the capstone project.

1. (Project) The student can work on an interesting dataset with a sponsor from a faculty member or administrator on campus. One current student is currently working with crime data with a professor associated with the Bureau of Criminal Investigation crime laboratory on the BGSU campus. Another student is working an a project with the Office of Admissions on campus. Several students are working on sports data projects with the Bowling Green High School football team, the BGSU women's basketball team and the BGSU ice hocky team.

2. (REU) Another way to satisfy the capstone project is to participate in one of the many Research Experiences for Undergraduates (REU) sponsored by different Universities that are focused on applications of data science.

To complete the Capstone Experience requirement, the student is required to write a paper that summarizes the data science exploration and findings and present the results of the work. BGSU currently has an initiative to encourage more research among undergraduate students and there currently is a symposium where students have the opportunity to present their research work.

3.4.2 Datafest

Datafest, sponsored by the American Statistical Association, is an event where teams of undergraduates work over the weekend to explore and find meaning in a large, rich, and complex dataset. BGSU was fortunate to be able to participate in a Datafest sponsored by Miami University (Ohio) during a weekend in April 2016. Six of the seven students were sophomores in the data science program. In this activity, the students had the challenge of reading in the 3 gigabyte dataset into their laptop, learning about the complex dataset, and formulating interesting questions to explore. They performed a variety of different data science tasks to summarize and visualize the data to address the research questions and prepared a presentation of the key findings during the last day of the event.

Datafest was a good experience, both from the viewpoints of the students and faculty who participated. In Datafest, the students saw the relevance of the data science skils they learned in working and learning about a completely new dataset. They experienced the challenges of working in a group on a common task over an extended period. It was fun for the students — they enjoyed the competitive aspects of the activity. From the BGSU faculty perspective, the Datafest event was a great way of applying data science on an interesting dataset, and this can be used as a recruiting and motivating tool for future data science students.

4 Challenges/Opportunities

4.1 Progress Report

At the time of this paper, there are currently 20 students in the new data science specialization and the two sophomore-level data science courses have been offered and seven students participated in the Datafest at the end of the second course.

Our general impression after seeing the students' performance at Datafest and other similar projects is that already as sophomores, these students have developed quite sophisticated skills in programming and handling data. These students are able to handle data science tasks at ease that would be impossible to expect from a senior in our traditional statistics major. The students gave nice presentations of their findings and they appear to feel quite accomplished that they can implement the methods on new data.

4.2 Teaching Topics in Data Science

On the data science skills side, it seemed challenging to teach SQL (databases) and Rcpp since some of the instructions did not work on both Windows and Macintosh platforms. Teaching databases would be easier if all students accessed a database that was available on a server.

From a statistical side, we found it was difficult teaching the bootstrap. We think this is understandable because the bootstrap relies on a relatively abstract foundation and the students did not any formal statistical training.

4.3 Challenges

Recruitment

Recruitment - we have no trouble recruiting for actuarial science, so it more a matter of data science being new. People want a bankable education, I guess, and they know actuaries have good jobs. I think that these two directions require different abilities and interests. Not everyone is cut out to work on computers a lot. Even computer science majors are in decline. Another thing is that so far, undergraduates have not been trying to be data scientists. The data scientists of the world have been just pulled from other kinds of jobs and had to figure stuff out on their own. So to me one challenge is that it is early in the game.

Industry Connections

I don't know how to increase connections with industry, but I think that with the incoming students we should encourage them to sign up for BG's internship program. That way at least we make the business community aware. Plus they might get an internship and possibly a job out of it.

Faculty

Other faculty? I think our colleagues expect us to do it all, and in our spare time do it for a new PhD program. If anyone is going to JSM there would probably be a workshop. Maybe we could fund that workshop fee from the grant.

References

[1] Albert, Jim and Marchi, Max (2013), *Analyzing Baseball Data with R*, Chapman and Hall.

[2] Albert, Jim and Rizzo, Maria (2012), *R by Example*, Springer.

[3] Baumer, Ben (2015), "A Data Science Course for Undergraduates: Thinking With Data," *The American Statistician,* 69, 4, 334-342.

[4] Baumer, B., Kaplan, D., and and Horton, N. (2016). *Modern Data Science with R.* CRC Press.

[5] Brown, Emery and Kass, Rob (2009), "What is Statistics", *The American Statistician*, Vol. 63, 2, 105-110.

[6] Chamandy, Nicholas, Muralidharan Omkar and Stefan Wager (2015), "Teaching Statistics at Google-Scale," *The American Statistician,* 69, 4, 283-291.

[7] Cleveland, William (1994), *Elements of Graphing Data*, Hobart Press.

[8] Donoho, David (2015), "50 Years of Data Science," preprint.

[9] Gentle, James (2004), *The American Statistician*, Vol. 58, 1, 2-5.

[10] Hardin, D., Hoerl, R., Horton, N., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. and M. D. Ward (2015), "Data Science in Statistics Curricula: Preparing Students to "Think with Data"," *The American Statistician,* 69, 4, 343-353.

[11] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition , Springer.

[12] Horton, Nicholas and Hardin, Johanna (2015), "Teaching the Next Generation of Statistics Students to "Think With Data": Special Issue on Statistics and the Undergraduate Curriculum," *The American Statistician,* 69, 4, 259-265.

[13] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), "An Introduction to Statistical Learning," . New York: Springer.

[14] Nolan, Deborah and Lang, Duncan (2010), "Computing in the Statistics Curricula", *The American Statistician*, Vol. 62, 2, pp. 97-107.

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

[16] Nolan, D., and Lang, D. T. (2015), "Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Analysis," *The American Statistician,* 69, 4, 292-299.

[17] O'Neill, C. and R. Schutt (2013). *Doing Data Science: Straight Talk from the Frontline.* O'Reilly Media.

[18] Rizzo, Maria L. (2008), *Statistical Computing with R*, Chapman and Hall.

[19] Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley.