ISSN: 2167-1435
v31 n3027

# Teaching "Big Data" in a Business School: Insights from an Undergraduate Course in Big Data Analytics

Stephen E. Hill
hills@uncw.edu

Douglas M. Kline
klined@uncw.edu

Information Systems and Operations Management Department
University of North Carolina Wilmington
Wilmington, NC 28403-5611

## Abstract

In this article, we describe the development of an undergraduate business course in "Big Data Analytics." We outline the preparation necessary for the development of such a course and describe our experiences with the initial delivery this course in a business school at a regional university located in the southeastern United States. In particular, we focus on challenges that an instructor may face as well as hurdles that students may need to overcome to be successful in the class.

**Keywords:** Big data, Education, Undergraduate, Analytics

### 1. INTRODUCTION

Few analytics-related topics have garnered recent attention to the degree that "big data" has. What is "big data"? The Opentracker website lists 32 definitions of "big data" from a variety of sources ("Definitions of Big Data", 2014). These definitions range from "Any amount of data that's too big to be handled by one computer" (as cited in Butler, 2012) to "The definition of big data? Who cares? It's what you're doing with it" (Franks, 2013). This considerable degree of debate and ambiguity about the definition of the term "big data" has done little to diminish the interest in this topic.

Google Trends (a website for the visualization of the frequency of Google search queries and something of a big data tool itself) indicates (see Figure 1) that queries for the term "big data" have increased over ten-fold since the beginning of 2011 ("Google Trends", 2014). This suggests that the big data discussion has shifted outside of the realm of niche academic and practitioner outlets and gained considerable attention among the general public.

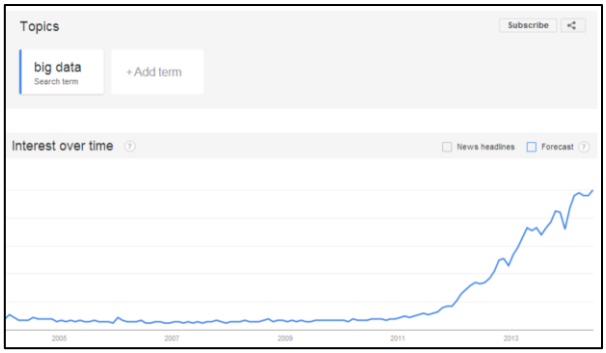


**Figure 1. Google Trends and "Big Data"**

An illustration of this shift occurred when the widely-circulated Wall Street Journal recently (March 24, 2014) dedicated an entire section of its publication to big data. This section included articles describing topics such as the privacy implications of big data (Dwoskin, 2014), various consumer-oriented big data applications (Kolodny, 2014), and unusual correlations in data that have been uncovered by corporate data analysts (Gage, 2014).

Of course big data has not always been mentioned in a positive light. The recent revelations of the data collection practices of the United States' National Security Agency (NSA) (Walsh, 2013) have affected public sentiment about big data and led to media articles with strong negative connotations (Gerstein and Simon, 2014). The Target "pregnant teen" anecdote (Duhigg, 2012) has also been widely used an example of big data's infiltration into the lives of members of the public by corporate entities.

Despite the negative press that big data has received, there has been in an emphasis in popular press articles on the need for more employees with training in big data-related skills. For example, a 2012 article co-authored by analytics and big data thought leader Thomas Davenport in the Harvard Business Review describes the job of "data scientist" as the "sexiest job of the 21st-century" (Davenport and Patil, 2012). A more recent Wall Street journal noted an increase in the demand for analytics professionals (Waller, 2014). It is clear that, as education professionals, we must be able to help students obtain the experience and expertise that they need to fill this need. This article presents the details of an attempt to teach a course on big data analytics to business school undergraduates.

## 2. CLASS PREPARATION

In this section of the article we describe the selection of topics, software tools, and textbook for the course.

### Topic, Textbook, and Tools Selection

The selection of topics for this course was driven by several questions. First, what topics would be timely and give the students a competitive advantage when seeking employment? Second, what topics could feasibly be covered given the potentially limited statistical and programming experience of the students? Third, what topics could be covered with significant depth in the time allotted for the course?

To answer the first question we needed to adopt a definition of "big data." As noted in the Introduction of this article, there are numerous definitions of this term. We ultimately chose to craft a definition of big data that could loosely be stated as: Big data is data that is too large to comfortably store, manage, and/or analyze in Excel. Although, ideally, we would cover Hadoop and other similar topics, we chose to narrow the focus of our course to the analysis of data that fits our loose definition of big data. Issues such as big data infrastructure would be left to future courses.

We then began to consider the second question: How could we feasibly teach big data analytics concepts and techniques to undergraduate students with limited statistical and programming background? We suspected that class enrollment in the initial offering of the course would be limited. The course was offered (due to several factors outside of the instructor's control) at 8:00 AM in a summer session lasting for approximately five weeks (125 minutes per class period, meeting four days per week). Students tend to be repelled by the thought of an 8:00 AM class in the summer. Also some students may be reluctant to take a brand new course offering, particularly a course that is an elective.

We expected that the course would attract two types of students and we attempted to develop a "typical" profile of these two types of students. Let's call the two students Student A and Student B. Both students (A and B) would have had a minimum of a single course in introductory business statistics, a single course in business calculus, a single course in information systems (including coverage of Microsoft Access and Excel), and a minimum of a single course in operations management. Because the big data analytics course was offered as an operations management course, it was also likely that both types of students would be operations management majors.

Student A would be a highly motivated student with an acute interest in analytics and quantitatively-oriented topics. This student may have previously taken an Excel-focused analytics class and may have sought out programming experience (either self-taught or via

coursework). Student B would be a student in search of an additional elective, but with no particular interest in analytics. In particular, such a student may be seeking to fulfill requirements to graduate at the end of the summer term or at the end of the fall semester.

Given these two expected student types, we ultimately settled on a set of topics that would serve as a partial review of concepts taught in the introductory statistics course and exploration of new analytics and big data-related concepts. Table 1 gives a list of the proposed topics for the course.

| Topic | Notes |
|---|---|
| Introduction to Big Data | Why big data and this class are important and how it may benefit the students |
| Finding, Getting, and Cleaning Data | Show difficulty and importance of finding and cleaning datasets |
| Visualization | New topic for most students |
| Linear Regression (Simple and Multiple) | Partially covered in previous statistics course |
| Logistic Regression | New topic for most students |
| Classification and Regression Trees and Random Forests | New topic for students |
| Clustering | New topic for students |
| Text Mining | New topic for students |
| Building a Recommendation System | New topic for students |

**Table 1. Proposed Topic List for Course**

**Textbook Selection**

The process of selecting an appropriate textbook for this course was difficult. Candidate texts were either too technical for the undergraduate students that would take the class or too high-level. For example, an advanced textbook by Rajaraman and Ullman (2011) (legally available for download from http://i.stanford.edu/~ullman/mmds.html) that is used in a "Mining of Massive Datasets" course in Stanford's Computer Science program is likely too advanced for undergraduate business students. On the other end of the spectrum would be books such as the excellent, but insufficiently technical "Big Data at Work: Dispelling the Myths, Uncovering the Opportunities" book from Thomas Davenport (2014). We ultimately decided against using a textbook for the class. In place of a textbook we identified a list of books that we categorized as optional readings for those seeking additional information about a variety of big data-related topics. The list is available from the Goodreads social reading site ("Popular Big Data Books", 2014).

With our topics selected and no appropriate textbook identified, we then selected software tools for use in the class. We selected the free and open source software R (R Core Development Team, 2014) to serve as our primary software tool in conjunction with the free and open source RStudio integrated development environment (RStudio, 2014). R is widely used and has been recognized as important tool for data analysis (Vance, 2009). Note that both R and RStudio are available for Windows, Mac OS X, and Linux operating systems.

| Package | Usage |
|---|---|
| datasets | Set of datasets available from within R |
| ggplot2 | Enhanced graphics package for data visualization |
| GGally | Helper package for ggplot2 that allows combination of plots |
| Hexbin | Creation of hexagonal bins for visualization |
| caTools | Variety of uses including splitting of datasets into training and testing sets |
| ROCR | Creation of Receiver Operating Characteristic curves |
| rpart | Classification and Regression Trees |
| rpart.plot | Helper package to enable better visualization of trees |
| tm | Package for text mining |
| SnowballC | Package for stemming in text mining |

**Table 2. R Packages for Course**

R is a powerful statistical programming language with a significant online support community and a variety of free, add-on packages available to perform advanced statistical analyses and

visualize data. Table 2 provides a listing of a sampling of the R packages that were utilized in the course.

R is easily able to handle large datasets that contain millions of rows of data (Bracht, 2013) and gives the students good experience with the challenges of dealing with larger datasets. It should be noted that R has a relatively steep learning curve and can be intimidating to students at times. We will describe the impact of this reality in Section 4 of this article.

**Finding Datasets**

We are fortunate to live in a time when data is widely and often freely available, particularly for use in educational settings. Despite this, it was difficult, at times, to identify ideal datasets for use in the course. Table 3 contains a set of online locations that we frequently utilized to find datasets. Table 4 lists several of the specific datasets that were utilized in the course and the sources for the datasets. R also includes a large number of datasets (most are relatively small) that are available to users upon installation of R. A number of the datasets listed in Table 4 are useful for multiple concepts. For example, the Titanic dataset was used during lessons on logistic regression and classification and regression trees.

| Source | Location |
|---|---|
| Kaggle | kaggle.com/competitions |
| Lahman's Baseball Databases | seanlahman.com/baseball-archive/statistics/ |
| Stanford SNAP | snap.stanford.edu |
| UC Irvine Machine Learning Repository | archive.ics.uci.edu/ml/ |
| Vanderbilt Biostatistics Datasets | biostat.mc.vanderbilt.edu/wiki/Main/DataSets |
| Reddit Datasets (Links to datasets) | reddit.com/r/datasets |
| KDnuggets (Links to datasets) | kdnuggets.com/datasets/index.html |

**Table 3. Dataset Sources**

| Dataset | Source |
|---|---|
| Lahman 2013 Baseball Dataset | Lahman (see Table 3) |
| Historical Annual Return Data for S&P 500, Treasury Bills, and Treasury Bonds | pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html |
| Motor Trend Car Road Tests Data | Built-in R Data Frame |
| Power Consumption Dataset | UCI Repository (see Table 3) |
| Lending Club Loan Data for 2012-2013 | lendingclub.com/info/download-data.action |
| Bike Sharing Dataset | UCI Repository (see Table 3) |
| Titanic Survival Dataset | Vanderbilt Biostatistics Datasets (see Table 3) |
| Framingham Heart Study Dataset | biolincc.nhlbi.nih.gov/teaching/ |

**Table 4. Selection of Datasets Used**

## 3. CLASS DELIVERY

In this section of the article we describe the students that enrolled in the course, the planned delivery implementation for the course, and the actual implementation after adjustment. Early in the course (on third day of class, after class enrollment had stabilized due students adding or dropping the course) a survey of the students (number of enrolled students = 12) was administered. This survey gauged the level of statistical and programming experience of the students and largely confirmed our initial assumptions that we would be teaching two general types of students in the course (as noted in Section 3 of this article).

| **Gender** | % |
|---|---|
| Male | 67 |
| Female | 33 |
| **Classification** | % |
| Junior | 8 |
| Senior | 92 |
| **Age Range** | % |
| 21 – 23 | 83 |
| 24 – 26 | 0 |
| 27 + | 17 |

**Table 5. Basic Student Demographics**

Table 5 includes the demographics of student respondents. The overwhelming majority of

students in the class had senior class standing and were of an age typical of this class standing. A majority of students of the class were male (in line with the ratio of male to female students in our academic department and in line with typical ratios from science, technology, engineering, and mathematics fields (see Beede et al., (2011)).

Table 6 illustrates the GPA, level of statistics, calculus, analytics, and programming experience of the students. Note that percentages in the programming portion of this table add to greater than 100% as one student had programming experience in two languages. The data in the table supports our pre-course notion that students taking the initial offering of the course would have limited preparation beyond basic statistics and, perhaps, a single course in basic analytics.

| **GPA** | % |
|---|---|
| 3.50 to 4.00 | 16 |
| 3.00 to 3.49 | 42 |
| 2.50 to 2.99 | 42 |
| **Statistics Experience** | % |
| Intro Statistics Only | 92 |
| Beyond Intro Statistics | 8 |
| **Calculus Experience** | % |
| Intro Calculus Only | 100 |
| Beyond Intro Calculus | 0 |
| **Analytics Experience** | % |
| None | 25 |
| Intro Analytics Only | 75 |
| **Programming Experience** | % |
| None | 50 |
| C/C++ | 8 |
| Java | 17 |
| C# | 8 |
| PHP | 8 |
| Visual Basic | 42 |
| Python | 8 |
| JavaScript | 8 |
| SQL | 25 |
| jQuery | 8 |

**Table 6. Student Academic Experience**

The class was taught in a computer lab with one computer for each student. R and RStudio were installed on the machines prior to the class. The initial plan for course delivery was to follow a model where the 125 minute class period would divided roughly in two. In the first half of each class period a topic would be introduced and demonstrated via one or more datasets in R. For example, logistic regression would be described and contrasted with linear regression. The Titanic dataset would be used to demonstrate logistic regression analysis in R. After the first half of the class period was concluded, the students would be given an assignment on the topic of the day. For logistic regression, the students were asked to analyze data from the Framingham Heart Study and identify an appropriate model to predict heart disease risk.

However, this plan proved to be unworkable. The students showed initial apprehension toward working in R (perhaps from a general concern about working in a command line-oriented environment). Several class periods that would have been split between content and assignment ended up being entirely dedicated to allowing students opportunities to work on their assignments. We opted to allow this to occur so that the students would have the opportunity to ask questions of the instructor in-person and to work with other students. Given the compressed schedule of the Summer term this allowance seemed acceptable. However, we were unable to cover all of the topics that we initially intended to cover.

## 4. CONCLUSIONS AND LESSONS LEARNED

This article describes the authors' experiences in preparing for and teaching such an undergraduate course in big data analytics at a business school at a regional university in the southeastern United States. Teaching such a course was a rewarding experience for the authors. Students appeared to be excited about the opportunity to enhance their knowledge of a "hot" topic and gain expertise in subject areas that they would not normally be exposed to in typical undergraduate business coursework. Student preparation for the class (as described in Table 6) was sufficient to allow for coverage of several complex topics using a variety of datasets.

However, teaching such a course was not without its challenges. In particular, instructors in such a course should be prepared to confront the following issues:

- Students may lack sufficient course preparation to take on highly technical topics. This may lead to a need to review or introduce basic concepts early in the big data analytics course.

- If R or a similar statistical programming language is used, students may be apprehensive regarding the use of a command line-based interface. To overcome this anxiety, the instructor should be prepared to spend significant class time introducing the programming environment and ensuring that students
- Teaching tasks may take longer than expected. The instructor should be prepared to allocate additional class time or provide significant time for guidance outside of class.

To better prepare students for a course in big data analytics, there may need to be significant changes made to the undergraduate business curriculum and course offerings. Ideally, each student taking a big data analytics course would have previously taken a basic course in business analytics. Although the authors' also offer a basic business analytics course, they chose not to make that course a prerequisite for the big data course. This decision may change for future semester offerings of big data analytics.

Furthermore, requiring a minimum of one programming course would likely reduce student anxiety about the command line interface and would potentially open the door for the use of other analytics tools (e.g., the pandas analysis library using Python and Hadoop using Java). Student exposure to additional statistical knowledge beyond topics covered in an introductory statistics course would also be beneficial.

Despite the challenges of teaching a big data analytics course on the undergraduate level in a business school, it is certainly possible to prepare and deliver a rewarding course. However, the instructor must be willing to be flexible and accommodating to the skill level of the students that are likely to enroll in the course. We look forward to the evolutionary process of enhancing our course for future semesters.

## 5. REFERENCES

Beede, D., Julian, T., Langdon, D., McKittrick, G., Khan, B., & Doms, M. (2011). *Women in STEM: A Gender Gap to Innovation* (Economics and Statistics Administration Issue Brief #04-11). Retrieved from http://files.eric.ed.gov/fulltext/ED523766.pdf

Bracht, O. (2013, November 7). Five ways to handle Big Data in R. [Weblog]. Retrieved from http://www.r-bloggers.com/five-ways-to-handle-big-data-in-r/.

Butler, B. (2012). Defining 'Big Data' Depends on Who's Doing the Defining. *Network World*. Retrieved from http://www.networkworld.com/article/2188435/data-center/defining--big-data--depends-on-who-s-doing-the-defining.html.

Davenport, T. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Cambridge, MA: Harvard Business Review Press.

Davenport, T. & D. Patil (2012). Data Scientist: The Sexiest Job of the 21st Century*, Harvard Business Review,* 90(10),70-76.

"Definitions of Big Data" (2014). Retrieved May 4, 2014, from http://www.opentracker.net/article/definitions-big-data.

Duhigg, C. (2012, February 16). How Companies Learn Your Secrets. *The New York Times*. Retrieved from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0.

Dwoskin, E. (2014, March 23). Give Me Back My Online Privacy. *Wall Street Journal.* Retrieved from http://online.wsj.com/news/articles/SB10001424052702304704504579432823496404570.

Franks, B. (2013, June 26). Cutting Through the Hype: What You Need to Know about Big Data, *Vimeo*. Retrieved from http://vimeo.com/95193715.

Gage, D. (2014, March 23). Big Data Uncovers Some Weird Correlations, *Wall Street Journal.* Retrieved from http://online.wsj.com/news/articles/SB10001424052702303369904579423132072969654.

ISSN: 2167-1435
v31 n3027

Gerstein, J. & Simon, S. (2014, May 14). Who watches the watchers? Big Data goes unchecked, *Politico*, Retrieved from http://www.politico.com/story/2014/05/big-data-beyond-the-nsa-106653.html.

Kolodny, L. (2014, March 23). How Consumers Can Use Big Data, *Wall Street Journal.* Retrieved from http://online.wsj.com/news/articles/SB10001424052702303369904579423383203100794.

"Popular Big Data Books" (2014). Retrieved June 8, 2014, from http://www.goodreads.com/shelf/show/big-data.

R Core Devlopment Team. (2014). R: A language and environment for statistical computing. (Version 3.1) [Computer software]. Available from http://www.R-project.org/.

Rajaraman, A. & Ullman, J. (2011). Mining of Massive Datasets, Cambridge, UK: Cambridge University Press.

RStudio (2014). RStudio: Integrated development environment for R (Version 0.98.507) [Computer software]. Available from http://www.rstudio.com/.

Vance, A. (2009, January 6) Data Analysts Captivated by R's Power. *The New York Times*. Retrieved from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all.

Waller, N. (2014, April 9) Get Familiar With Big Data Now—or Face 'Permanent Pink Slip', *Wall Street Journal.* Retrieved from http://online.wsj.com/news/articles/SB10001424052702304819004579489541746990638.

Walsh, B. (2013, June 24) The NSA's Big Data Problem. *Time*. Retrieved from http://content.time.com/time/magazine/article/0,9171,2145481,00.html.