Masthead Logo

Smith ScholarWorks

Mathematics and Statistics: Faculty Publications

Mathematics and Statistics

12-29-2015

A Data Science Course for Undergraduates: Thinking with Data

Benjamin Baumer
Smith College, bbaumer@smith.edu

Follow this and additional works at: https://scholarworks.smith.edu/mth_facpubs

Part of the Statistics and Probability Commons

Recommended Citation

Baumer, Benjamin, "A Data Science Course for Undergraduates: Thinking with Data" (2015). Mathematics and Statistics: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/mth_facpubs/25

This Article has been accepted for inclusion in Mathematics and Statistics: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu

A Data Science Course for Undergraduates: Thinking with Data

Firstname Lastname

Affiliation

Abstract

Data science is an emerging interdisciplinary field that combines elements of mathematics, statistics, computer science, and knowledge in a particular application domain for the purpose of extracting meaningful information from the increasingly sophisticated array of data available in many settings. These data tend to be non-traditional, in the sense that they are often live, large, complex, and/or messy. A first course in statistics at the undergraduate level typically introduces students to a variety of techniques to analyze small, neat, and clean data sets. However, whether they pursue more formal training in statistics or not, many of these students will end up working with data that are considerably more complex, and will need facility with statistical computing techniques. More importantly, these students require a framework for thinking structurally about data. We describe an undergraduate course in a liberal arts environment that provides students with the tools necessary to apply data science. The course emphasizes modern, practical, and useful skills that cover the full data analysis spectrum, from asking an interesting question to acquiring, managing, manipulating, processing, querying, analyzing, and visualizing data, as well communicating findings in written, graphical, and oral forms.

Keywords: data science, data wrangling, statistical computing, undergraduate curriculum, data visualization, machine learning, computational statistics

1. Introduction

The last decade has brought considerable attention to the field of statistics, as undergraduate enrollments have swollen across the country. Fueling the interest in statistics is the proliferation of data being generated by scientists, large Internet companies, and electronic devices of all shapes and sizes. There is widespread acknowledgement—coming naturally from scientists, but also from CEOs and government officials—that these data could be useful for informing decisions. Accordingly, the job market for people who can translate these data into actionable information is very strong, and there is evidence that demand for this type of labor far exceeds supply (Harris, Shetterley, Alter, and Schnell 2014). By all accounts, students are eager to develop their ability to analyze data, and are wisely investing in these skills.

But while this data onslaught has strengthened interest in statistics, it has also brought challenges. Modern data streams are importantly different than the data with which many statisticians, and in turn many statistics students, are accustomed to working. For example, the typical data set a student encounters in an introductory statistics course consists of a several dozen rows and three or four columns of non-collinear variables, collected from a simple random sample or a randomized trial. These are data that are likely to meet the conditions necessary for statistical inference in a multiple regression model. From a pedagogical point-of-view, this makes both the students and the instructor happy, because the data fit the model, and thus we can proceed to apply the techniques we have learned to draw meaningful conclusions. However, the data that many of our current students will be asked to analyze—especially if they go into government or industry—will not be so neat and tidy. Indeed, these data are not likely to come from an experiment—they are much more likely to be observational. Secondly, they will not likely come in a two-dimensional row-and-column format—they might be stored in a database, or a structured text document (e.g., XML), or come from more than one source with no obvious connecting identifier, or worse, have no structure at all (e.g., data scraped from the web). These data might not exist at a fixed moment in time, but rather be part of a live stream (e.g., Twitter). These

data might not even be numerical, but rather consist of text, images, or video. Finally, these data may consist of so many observations that many traditional inferential techniques might not make sense to use, or even be computationally feasible.

In 2009, Hal Varian, chief economist at Google, described *statistician* as the "sexy job in the next 10 years" (Lohr 2009). Yet by 2012, the Harvard Business Review used similar logic to declare *data scientist* as the "sexiest job of the 21st century" (Davenport and Patil 2012). Speaking at the 2013 Joint Statistical Meetings, Nate Silver—as always—helped us unravel what had happened. He noted that "data scientist is just a sexed up term for a statistician." If Silver is right, then the statistics curriculum needs to be updated to include topics that are currently more closely associated with data science than with statistics (e.g., data visualization, database querying, data wrangling, algorithmic concerns about computational techniques)[1]. Statisticians and data scientists share a common goal—namely, to use data appropriately to inform decision-making.

What we describe in this paper is a course at a liberal arts college in *data science* that is atypical within the current statistics curriculum. Nevertheless, what we present here is wholly consistent with the vision for the future of the undergraduate statistics curriculum articulated by Horton (2015) and the American Statistical Association Undergraduate Guidelines Workgroup (2014). The purpose of this course is to prepare students to work with these modern data streams as described above. Some of the topics covered in this course have historically been the purview of computer science. But while the course we describe indisputably contains elements of statistics and computer science, it just as indisputably belongs exclusively to neither discipline. Furthermore, it is not simply a collection of topics from existing courses in statistics and computer science, but rather an integrated presentation of something more holistic. Nevertheless, the course consists of a series of largely independent modules, each of which could be expanded into stand-alone curricular elements (e.g., a full-semester, half-semester, or interterm course). Thus, while some readers might view this paper as a specification for a single new offering, others might use it as a

[1]The Wikipedia defines "data science" as "the extraction of knowledge from data", whereas "statistics" is "the study of the collection, analysis, interpretation, presentation, and organization of data." Does writing an SQL query belong to both?

blueprint for a significant expansion of the existing statistics curriculum.

2. Background and Related Work

While many believe that to understand statistical theory, a solid foundation in mathematics is necessary, it seems clear that *computing* skills are required for one to become a functional, practicing statistician. In making this analogy Nolan and Temple Lang (2010) argue strongly for a larger presence for computing in the statistics curriculum. Citing their work, the American Statistical Association Undergraduate Guidelines Workgroup (2014) underscores the importance of computing skills (even using the words "data science") in the 2014 guidelines for undergraduate majors in statistical science. Here, by *computing*, we mean *statistical programming* in an environment such as R. It is important to recognize this as a distinct—and more valuable—skill than being able to perform statistical computations in a menu-and-click environment such as Minitab. Indeed, Nolan and Temple Lang (2010) go even further, advocating for the importance of teaching general command-line programs, such as `grep` (for regular expressions) and other common UNIX commands that really have nothing to do with statistics, *per se*, but are very useful for cleaning and manipulating documents of many types.

Although practicing statisticians seem to largely agree that the lion's share of the time spent on many projects is devoted to data cleaning and manipulation (or *data wrangling*, as it is often called (Kandel, Heer, Plaisant, Kennedy, van Ham, Riche, Weaver, Lee, Brodbeck, and Buono 2011)), the motivation for adding these skills to the statistics curriculum is not simply convenience, nor should a lack of skills or interest on the part of instructors stand in the way. Finzer (2013) describes a "data habit of mind...that grows out of working with data." (This is not to be confused with "statistical thinking" as articulated by Chance (2002), which contains no mention of computing.) In this case, a data habit of mind comes from experience working with data, and is manifest in people who start thinking about data formatting *before* data are collected (Zhu, Hernandez, Mueller, Dong, and Forman 2013), and have foresight about how data should be stored that is informed by how they

will be analyzed. Furthermore, while some might view *data management* as a perfunctory skill on intellectual par with *data entry*, there are others thinking more broadly about data.[2] Just as Wilkinson (2006) brings structure to graphics through "grammar," Wickham (2014) and Wickham and Francois (2015) bring structure to data manipulation through the five "verbs": *select, filter, mutate, arrange*, and *summarise*. These common single-table data manipulation operations are the practical descendents of theoretical work on data structures by computer scientists who developed notions of normal forms, relational algebras, and database management systems.

While the emphasis on computing within the statistics curriculum may be growing, it belongs to a larger, more gradual evolution in statistics education towards data analysis—with computers—and encourages us to reflect on shifting boundaries between statistics and computer science. Moore (1998)—viewing statistics as an ongoing quest to "reason about data, variation, and chance"—sees statistical thinking as a powerful anchor that can prevent statistics from being "overwhelmed by technology." Cobb (2011) argues for an increased emphasis on conceptual topics in statistics, but also sees the development of statistical theory as an anachronistic consequence of a lack of computing power (Cobb 2007). Moreover, while much of statistical theory is designed to make the strongest inference possible from what are historically scarce data, we are now often challenged trying to draw meaningful conclusions from an abundance of data.

Breiman (2001) articulates the distinction between "statistical data models" and "algorithmic models" that in many ways characterizes the relationship between statistics and machine learning, viewing the former as being far more limited than the latter. And while machine learning and data mining have traditionally been subfields of computer science, Finzer (2013) notes that data science does not have a natural home within traditional departments, belonging exclusively to neither mathematics, statistics, nor computer science. Indeed, in Cleveland (2001)'s seminal action plan for data science, he envisions data science

[2]Examples of poor data management abound, but one of the most common is failure to separate the actual data from their analysis. Microsoft Excel is a particular villain in this arena, where merged cells, rounding induced by formatted columns, and recomputed formulas can result in the ultimate disaster: losing the original recorded data!

as a "partnership" between statisticians (i.e., data analysts) and computer scientists.

3. The Course

In this paper we describe an experimental course—called "Data Science" (a.k.a., SDS 292) and now offered through the Statistical & Data Sciences Program at XXX College—that was offered in the fall of 2013 and again in the fall of 2014. In the first year, 18 students completed the course, as did another 24 in the following year. The prerequisites are an introductory statistics course and some programming experience. Existing courses at the University of California at Berkeley, as well as Macalester and St. Olaf Colleges, are the pedagogical cousins of SDS 292 (see Hardin, Hoerl, Horton, and Nolan (2015) for a comprehensive comparison).

SDS 292 is organized into a series of two-to-three week modules: data visualization, data manipulation/data wrangling, computational statistics, machine/statistical learning, and additional topics. In what follows we provide greater detail on each of these modules.

Learning Outcomes In Figure 1, we present a schematic of a modern statistical analysis process, from formulating a question to obtaining an answer. In the introductory statistics course, we teach a streamlined version of this process, wherein challenges with the data, computational methods, and visualization and presentation are typically elided. The entire process informs the material presented in the data science course. The goal is to produce students who have *confidence* and foundational skills—not necessarily expertise—to tackle each step in this modern data analysis cycle, both immediately and in their future careers.

3.1. Day One

The first class provides an important opportunity to hook students into data science. Since most students do not have a firm grasp of what data science is, and in particular, how it differs from statistics, the diagram in Fig. 1 can help draw these distinctions. The goal

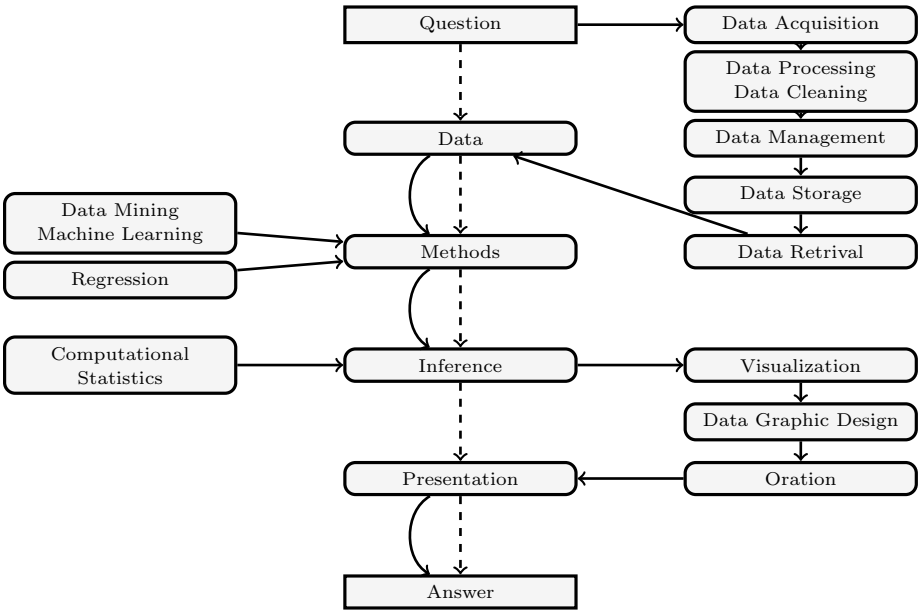

Figure 1: Schematic of the modern statistical analysis process. The introductory statistics course (and in many cases, the undergraduate statistics curriculum) emphasizes the central column. In this data science course, we provide instruction into the bubbles to the left and right.

is to illustrate the richness and vibrance of data science, and emphasize its breadth by highlighting the different skills necessary for each task. Students should be sure within the first five minutes of the semester that there is something interesting and useful for them to learn in the course.

Next, we engage students immediately by exposing them to a recent, relevant example of data science. Students are asked to read a provocative paper by DiGrazia, McKelvey, Bollen, and Rojas (2013) and a rather ambitious editorial in *The Washington Post* written by Rojas (2013), a sociologist, in which he claims that Twitter will put political pollsters out of work. (See Finger and Dutta (2014) for a counterpoint.)

This represents a typical data science research project, in that:

- The data being analyzed were scraped from the Internet, not collected from a survey or clinical trial. Typical statistical assumptions about random sampling or random assignment are clearly not met.
- The research question was addressed by combining *domain knowledge* (i.e., knowledge

of how Congressional races work) with a data source (Twitter) that had no obvious relevance to one another.

- A *large* amount of data (500 million tweets!) was collected (although only 500,000 tweets were analyzed)—so large that the data itself was a challenge to manage. In this case, the data were big enough that the Center for Complex Networks and Systems Research at Indiana University was enlisted to assist.
- The project was undertaken by a team of researchers from multiple fields (i.e., sociology, computing) working in different departments who brought complementary skills to bear on the problem—a paradigm that many consider to be optimal (Patil 2011).

Students are then asked to pair up and critically review the paper. The major findings reported by the authors stem from the interpretation of two scatterplots and two multiple regression models, both of which are accessible to students who have had an introductory statistics course. There are several potential weaknesses in both the plots presented in the paper (Linkins 2013; Gelman 2013), and the interpretation of the coefficients in the multiple regression model, which some students will identify. The exercise serves to refresh students' memories about statistical thinking, encourages them to think critically about the display of data, and illustrates the potential hazards of drawing conclusions from data in the absence of a statistician. Instructors could also use this discussion as a segue to topics in experimental design, or introduce the ASA's Ethical Guidelines for Statistical Practice (Committee on Professional Ethics 1999).

Finally, students are asked directly how they would go about replicating this study. That is, they are asked to identify all of the steps necessary to conduct this study, from collecting the data to writing a report, and to think about whether they could accomplish this with their current skills and knowledge. While students are able to generate many of the steps as a broad outline, most are unfamiliar with the practical considerations necessary. For example, students recognize that the data must be downloaded from Twitter, but few have any idea how to do that. This leads to the concept of an API (application programming interface), which is provided by Twitter (and can be used in several environments, notably R

and Python). Moreover, most students do not recognize the potential difficulties of storing 500 million tweets. How big is a tweet? Where and how could you store them? Spatial concerns also arise: does it matter in which Congressional district the person who tweeted was? Most students in the class have experience with R, and thus are comfortable building a regression model and overlaying it on a scatterplot. But few have considered anything beyond the default plotting options. How do you add annotations to the plot to make it more understandable? What principles of data graphic design would help to determine which annotations are necessary or appropriate?

Students are then advised that this course will give them the tools necessary to carry out a similar study. This will involve improving their skills with programming, data management, data visualization, and statistical computing. The goal is to leave students feeling *energized*, but open to exploring their newly-acquired, more complex understanding of data science.

3.2. Data Visualization

From the first day of class, students are reminded that statistical work is of limited value unless it can be communicated to non-statisticians (Swires-Hennessy 2014). More specifically, most data scientists working in government or industry (as opposed to those in academia) will work for a boss who generally possesses less technical knowledge than the employee. A perfect, but complicated, statistical model may not be persuasive to non-statisticians if it cannot be communicated clearly. Data graphics provide a mechanism for illustrating relationships among data, but most students have never been exposed to structured ideas about how to create effective data graphics.

In SDS 292, the first two weeks of class are devoted to data visualization. This serves two purposes: 1) it is an engaging hook for a science course; and 2) it gives students with weaker programming backgrounds a chance to get comfortable in R.

Students read the classic text of Tufte (1983) in its entirety, as well as excerpts from Yau (2013). The former provides a wonderfully cantankerous account of what *not* to do when

creating data graphics, as well as thoughtful analyses of how data graphics should be constructed. Students take delight in critiquing data graphics that they find online through the lens crafted by Tufte. The latter text, along with Yau (2011), provides many examples of interesting data visualizations that can be used in the beginning of class to inspire students to think broadly about what can be done with data (e.g., *data art*). Moreover, it provides a well-structured taxonomy for composing data graphics that gives students an orientation into data graphic design. For example, a data graphic that uses color as a visual cue in a Cartesian coordinate system is what we commonly call a "heat map". Students are also exposed to the hierarchy of visual perception that stems from work by Cleveland (2001).

Homework questions from this part of the course focus on demonstrating understanding by critiquing data graphics found "in the wild," an exercise that builds *confidence* (i.e., "Geez, I already know more about data visualization than this guy...") and encourages critical thinking. Computational assignments introduce students to some of the less trivial aspects of annotating data graphics in R (e.g., adding textual annotations and manipulating colors, scales, legends, etc.). We discuss additional topics in data visualization in Section 3.6.

3.3. Data Manipulation/Data Wrangling

As noted earlier, it is a common refrain among statisticians that "cleaning and manipulating the data" takes up an overwhelming majority of the time spent on a statistical project. In the introductory class, we do everything we can to shield students from this reality, exposing them only to carefully curated data sets. By contrast, in SDS 292 students are expected to master a variety of common data manipulation techniques. The term *data management* has a boring, IT connotation, but there is a growing acknowledgement that such *data wrangling*, or *data manipulation* skills are not only valuable, but in fact belong to a broader intellectual discipline (Wickham 2014; Horton, Baumer, and Wickham 2015). One of the primary goals of SDS 292 is to develop students' capacity to "think with data" (Nolan and Temple Lang 2010), in both a practical and theoretical sense.

Over the next three weeks, students are given rapid instruction in data manipulation in R

and SQL. In the spirit of the data manipulation "verbs" advocated by Wickham and Francois (2015), students learn how to perform the most fundamental data operations in both R and SQL, and are asked to think about their connection.

- *select*: subset variables (`SELECT` in SQL, `select()` in R (`dplyr`))
- *filter*: subset rows (`WHERE, HAVING` in SQL, `filter()` in R)
- *mutate*: add new columns (`... AS ...` in SQL, `mutate()` in R)
- *summarise*: reduce to a single row (`GROUP BY` in SQL, `summarise(group_by())` in R)
- *arrange*: re-order the rows (`ORDER BY` in SQL, `arrange()` in R)

By the end, students are able to see that an SQL query containing

```
SELECT ... FROM a JOIN b WHERE ... GROUP BY ... HAVING ... ORDER BY ...
```

is equivalent to a chain of R commands involving

```
a %>%
  select(...) %>%
  filter(...) %>%
  inner_join(b, ...) %>%
  group_by(...) %>%
  summarise(...) %>%
  filter(...) %>%
  arrange(...)
```

A summary of analogous R and SQL syntax is shown in Table 1.

Moreover, students learn to determine for themselves, based on the attributes of the data (most notably size), which tool is more appropriate for the type of analysis they wish to

Concept	SQL	R (dplyr)
Filter by rows & columns	`SELECT col1, col2 FROM` a `WHERE col3 = 'x'`	`select(filter(`a`, col3 == "x"), col1, col2)`
Aggregate by rows	`SELECT id, sum(col1) as total FROM` a `GROUP BY id`	`summarise(group_by(`a`, id), total = sum(col1))`
Merge two tables	`SELECT * FROM` a `JOIN` b `ON a.id = b.id`	`inner_join(x=`a`, y=`b`, by="id"))`

Table 1: Conceptually analogous SQL and R commands. Suppose a and b are SQL tables or R `data.frame`s

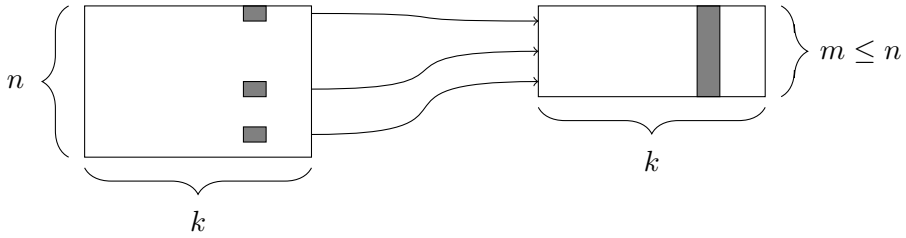

Figure 2: The filter operation

perform. They learn that R stores data in memory, so that the size of the data with which you wish to work is limited by the amount of memory available to the computer, whereas SQL stores data on disk, and is thus much better suited for storing large amounts of data. However, students learn to appreciate the virtually limitless array of operations that can be performed on data in R, whereas the number of useful computational functions in SQL is limited. Thus, students learn to make choices about software in the context of hardware—and data.

Care must be taken to make sure that what students are learning at this stage of the course is not purely progamming syntax (although that is a desired side effect). Rather, they are learning more generally about operations that can be performed on data, in two languages. To reinforce this, students are asked to think about a physical representation of what these operations do. For example, Figure 2 illustrates conceptually what happens when row filtering is performed on a `data.frame` in R or a table in SQL. Less trivially, Figure 3 illustrates the useful `gather` operation in R.

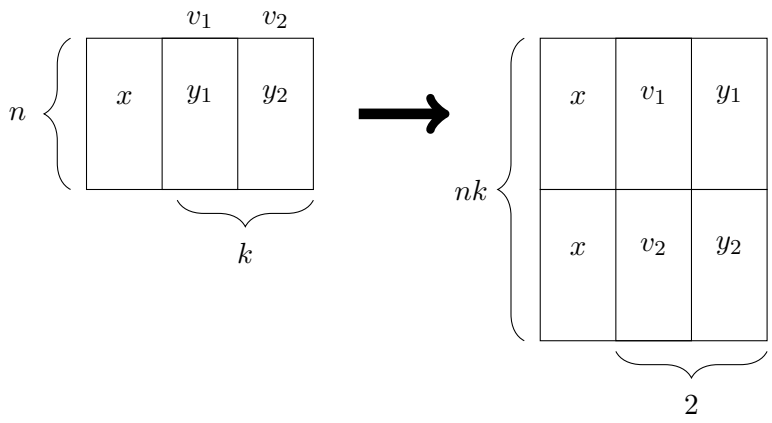

Figure 3: The gather operation

3.4. Computational Statistics

Now that students have the intellectual and practical tools to work with data and visualize them, the third part of the course provides students with computational statistical methods for analyzing data in the interest of answering a statistical question. There are two major objectives for this section of the course:

1. Developing facility constructing interval estimates using resampling techniques (e.g., the bootstrap). Understanding the nature of variation in observational data and the benefit of presenting interval estimates over point estimates.

2. Developing the capacity to fit and assess regression models, beginning with simple linear regression (which all students should have already seen in their intro course), but continuing to include multiple and logistic regression, and a few techniques for automated feature selection.

The first objective underlines the *statistical* elements of the course, encouraging students to put observations in relevant context by demonstrating an understanding of variation in their data. The second objective, while not a substitute for a semester course in regression analysis, helps reinforce a practical understanding of regression, and sets the stage for the subsequent machine learning portion of the course.

3.5. Machine Learning

Two weeks are devoted to introductory topics in machine learning. Some instructors may find that this portion of the course overlaps too heavily with existing offerings in computer science or applied statistics. Others might argue that these topics will not be of interest to students who are primarily interested in the communication and visualization side of data science. However, a brief introduction to machine learning gives students a functional framework for testing algorithmic models. Assignments force them to grapple with the limitations of large data sets, and pursue statistical techniques that are beyond introductory.

In order to appreciate machine learning, one must recognize the differences between the mindset of the data miner and the statistician. Breiman (2001) distinguishes two types of models f for y, the response variable, and $\mathbf{x}$, a vector of explanatory variables. One might consider a *data model* f such that $y \sim f(\mathbf{x})$. If it can be determined that f is a reasonable approximation of the real-world process by which y was generated from $\mathbf{x}$, then we can proceed to make inferences about f. The goal is to learn about that unknown real process, and the conceit is that f is a meaningful reflection of it. Alternatively, one might construct an *algorithmic model* f, such that $y \sim f(\mathbf{x})$, and use f to predict unobserved values of y. If it can be determined that f does in fact do a good job of predicting values of y, one might not care to learn much about f. In the former case, since we want to learn about f, a simpler model may be preferred. Conversely, in the latter case, since we want to predict new values of y, we may be indifferent to model complexity (other than concerns about overfitting and scalability).

These are *very* different perspectives to take towards learning from data, so after reinforcing the former perspective that students learned in their introductory course, SDS 292 students are exposed to the latter point-of-view. These ideas are further explored in a class discussion about Chris Anderson's famous article on *The End of Theory* (Anderson 2008), in which he argues that the abundance of data and computing power will eliminate the need for scientific modeling.

The notions of cross-validation and the "confusion" matrix frame the machine learning unit (ROC curves are also presented as an evaluation technique). The goal is typically to predict the outcome of a binary response variable. Once students understand that these predictions can be evaluated using a confusion matrix, and that models can be tested via cross-validation schemes, the rest of the unit is spent learning classification techniques. The following techniques are presented, mainly at a conceptual and practical level: decision/classification trees, random forests, k-nearest neighbor, näive Bayes, artificial neural networks, and ensemble methods.

One of the most satisfying aspects of this unit is that students can tackle a massive data set. Past instances of the KDD Cup (http://www.sigkdd.org/kddcup/index.php) are an excellent source for such examples. We explore data from the 2008 KDD Cup on breast cancer. Each of the n observations contains digitized data from an X-Ray image of a breast. Each observation corresponds to a small area of a particular breast, which may or may not depict a malignant tumor—this provides the binary response variable. In addition to a handful of well-defined variables ((x, y)-location, etc.), each observation has 117 nameless attributes, about which no information is provided. Knowing nothing about what these variables mean, students recognize the need to employ machine learning techniques to sift through them and find relationships. The size of the data and number of variables make manual exploration of the data impractical.

Students are asked to take part in a multi-stage machine learning "exam" (Cohen and Henle 1995) on this breast cancer data. In the first stage, students are given several days to work alone and try to find the best logistic regression model that fits the data. In the second stage, students form groups of three, discuss the strengths and weaknesses of their respective models, and then build a classifier, using any means available to them, that best fits the data. (These classifiers are ultimately evaluated on as yet unseen data.) The third stage of the exam is a traditional in-class exam.

3.6. Additional Topics

As outlined above, data visualization, data manipulation, computational statistics, and machine learning are the four pillars of this data science course. However, additional content can be layered in at the instructor's discretion. We list a few such topics below. Greater detail is provided in our supplementary materials.

- Spatial Analysis: creating appropriate and meaningful graphical displays for data that contain geographic coordinates
- Text Mining & Regular Expressions: learning how to use regular expressions to produce data from large text documents
- Data Expo: exposing students to the questions and challenges that people outside the classroom face with their own data
- Network Science: developing methods for data that exist in a network setting (i.e., on a graph)
- Big Data: illustrating the next frontier for working with data that are truly large scale

4. Computing

Practical, functional programming and computational abilities are essential for a data scientist, and as such no attempt is made to shield students from the burden of writing their own code. Copious examples are given, and detailed lecture notes containing annotated computations in R are disseminated each class. Lectures jump between illustrating concepts on the blackboard and writing code on the computer projected overhead, and students are expected to bring their laptops to class each day and participate actively. While it is true that many of the students struggle with the programming aspect of the course, even those that do express enthusiasm and satisfaction as they become more comfortable. Newly-focused on becoming data scientists, several students will go on to take subsequent courses on data structures or algorithms offered by the computer science department.

In this course, programming occurs exclusively in R and SQL. Others may assert that

Python is also necessary, and future incarnations of this course may include more Python. In my view these are the three must-have languages for data science. [3]

4.1. A Note to Prospective Instructors

Several people familiar with this course have asked about the skills required to teach it. From my point-of-view the most important thing is to have the same willingness to learn new things that you ask of your students. In terms of the content, a deep knowledge of all subjects is not required, although comfort and troubleshooting ability with R is necessary. Students are willing to accept a certain amount of frustration that goes hand-in-hand with learning a new programming language, but when they encounter roadblocks that seem immovable, that frustration can mutate into helplessness. The instructor must provide support mechanisms to avoid this—student teaching assistants and office hours can be especially helpful.

Even without prior knowledge, enough of the material on data visualization and machine learning can be absorbed in a relatively short period of time by reading a few of the books cited. SQL has many subtleties—but most are not likely to come up in this course, and the basics are not difficult to learn, even via online tutorials and self-study. Here again, some experience and practice are important.

For students, prior programming experience is essential. Experience with R is not required, and in my experience, computer science majors with weaker statistical backgrounds usually fare better than students with stronger statistical backgrounds but less programming experience. This is a demanding course that requires most students to spend a substantial

[3] SQL is a mature technology that is widely-used, but useful for a specific purpose. R is a flexible, extensible platform that is specifically designed for statistical computing, and represents the current state-of-the-art. Python has become something of a *lingua franca*, capable of performing many of the data analysis operations otherwise done in R, but also being a full-fledged general purpose programming language with lots of supporting packages and documentation.

At XXX, all introductory computer science students learn Python, and all introductory statistics students in the statistical and data sciences program learn R. However, it is not clear yet how large the intersection of these two groups is. It is probably easier for those who know Python to learn R than it is for those who know R to learn Python, and thus the decision was made in this instance to avoid Python and focus on R. Other instructors may make different choices without disruption.

amount of time working through assignments. However, even students who struggle are so convinced that what they are learning is useful that there are few serious complaints. Nevertheless one could certainly experiment with slowing down the pace of the course.

5. Assignments

Reading assignments in SDS 292 are culled from a variety of textbooks and articles available for free online. Non-trivial sections are assigned from a number of texts (James, Witten, Hastie, and Tibshirani 2013; Tan, Steinbach, and Kumar 2006; Rajaraman and Ullman 2011; Murrell 2010) [4].

Concepts from the readings are developed further during the lecture periods in conjuction with implementations demonstrated in R. Homework, consisting of conceptual questions requiring written responses as well as computational questions requiring coding in R, is due approximately every two weeks. Two exams are given—both of which have in-class and take-home components. The first exam is given after the first two modules, and focuses on data visualization and data manipulation principles demonstrated in written form. The second exam unfolds over two weeks, and focuses on the challenging breast cancer classification problem discussed above. An open-ended project (described below) brings the semester to a close. More details on these assignments, including sample questions, are presented in our supplementary materials.

Project The culmination of the course is an open-ended term project that students complete in groups of three. Only three conditions are given:

1. Your project must be centered around data
2. Your project must tell us something
3. To get an A, you must show something beyond what we've done in class

[4] Please see our supplementary materials for more information.

Just like in other statistics courses, the project is segmented so that each group submits a proposal that has to be approved before the group proceeds (Halvorsen and Moore 2001). The final deliverable is a 10-minute in-class presentation as well as a written "blog post" crafted in R Markdown (Allaire, Horner, Marti, and Porte 2015).

Examples of successful projects are presented in the supplementary materials.

6. Epilogue

The feedback that I have received on this course—through informal and formal evaluations—has been nearly universally positive. In particular, the 42 students (mostly from XXX but also including five students from three nearby colleges) seemed convinced that they learned "useful things." More specific feedback is available in the supplementary materials.

Several of these students were able to channel these useful skills into their careers almost immediately. Internships and job offers followed in the spring for a handful of students: two students spent their summers at NIST (one of whom later accepted a full-time job offer from MIT's Lincoln Laboratory; the other is headed to the Ph.D. program in statistics at Berkeley and is a trainee in the NSF-funded "Environment and Society: Data Science for the 21st Century" research program), one student landed a job as a research analyst at the nonprofit research organization MDRC, and three students have joined the new Data Science Development Program at MassMutual. External validation also came during the ASA Five College DataFest, the local version of the national data analysis competition (Gould, Baumer, Çetinkaya Rundel, and Bray 2014). ASA DataFest is an open-ended data analysis competition that challenges students working in teams of up to five to develop insights from a difficult data set. In each of the last two years, a team of five students from XXX—four of whom had taken this course—won the Best In Show prize (one student was a member of both teams). In the first year in particular, skills developed in the course helped these students perform data manipulation tasks with considerably less difficulty than other groups. For example, each observation in this particular data set included a *date* field, but the values were encoded as strings of text. Most groups struggled to work sensibly with these data, as

familiar workflows were infeasible (e.g., the data was too large to open in Excel, so "Format Cells..." was not a viable solution). The winning group was able to quickly tokenize these strings in R, and—having cleared this hurdle—had more time to spend on their analysis and interpretation.

7. Discussion

It is clear that the popularity of *data science* has brought both opportunities and challenges to the statistics profession. While statisticians are openly grappling with questions about the relationship of our field to data science (Davidian 2013a,b; Franck 2013; Bartlett 2013; Horton 2015; Wasserstein 2015), there appears to be less conflict among computer scientists, who (rightly or wrongly) distinguish data science from statistics on the basis of the heterogeneity and lack of structure of the data with which data scientists, as opposed to statisticians, work (Dhar 2013). As *Big Data* (which is clearly related to—but too often conflated with—data science) is often associated with computer science, computer scientists tend to have an inclusive attitude towards data science.

A popular joke is that, "a data scientist is a statistician who lives in San Francisco," but Hadley Wickham (2012), a Ph.D. statistician, floated a more cynical take on Twitter: "a data scientist is a statistician who is useful." Statisticians are the guardians of statistical inference, and it is our responsibility to educate practitioners about using models appropriately, and the hazards of ignoring model assumptions when making inferences. But many model assumptions are only truly met under idealized conditions, and thus, as Box (1979) eloquently argued, one must think carefully about when statistical inferences are valid. When they are not, statisticians are caught in the awkward position, as Wickham suggests, of always saying "no." This position can be dissatisfying.

If data science represents the new reality for data analysis, then there is a real risk to the field of statistics if we fail to embrace it. The damage could come on two fronts: first, we lose data science and all of the students who are interested in it to computer science; and second, the world will become populated by data analysts who don't fully understand or appreciate

the importance of statistics. While the former blow would be damaging, the latter could be catastrophic—and not just for our profession. Conversely, while the potential that data science is a fad certainly exists, it seems less likely each day. It is hard to imagine waking up to a future in which decision-makers are not interested in what data (however they may have been collected and however they may be structured) can offer them.

Data science courses like the one described in this paper provide a mechanism to develop students' abilities to work with modern data, and these skills are quickly transitioning from desirable to necessary.

References

Allaire J, Horner J, Marti V, Porte N (2015). `markdown`*: Markdown rendering for R*. R package version 0.7.7, http://CRAN.R-project.org/package=markdown.

American Statistical Association Undergraduate Guidelines Workgroup (2014). *2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science*. http://www.amstat.org/education/curriculumguidelines.cfm, last accessed: 2015-05-19.

Anderson C (2008). "The End of Theory." *Wired*. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory, last accessed: 2015-05-19.

Bartlett R (2013). "We Are Data Science." *AMSTAT News*. http://magazine.amstat.org/blog/2013/10/01/we-are-data-science/, last accessed: 2015-05-19.

Box GE (1979). "Some problems of statistics and everyday life." *Journal of the American Statistical Association*, **74**(365), 1–4. http://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481600.

Breiman L (2001). "Statistical modeling: The two cultures." *Statistical Science*, **16**(3), 199–215. http://www.jstor.org/stable/2676681.

Chance BL (2002). "Components of Statistical Thinking and Implications for Instruction

and Assessment." *Journal of Statistics Education*, **10**(3). http://www.amstat.org/publications/jse/v10n3/chance.html.

Cleveland WS (2001). "Data science: an action plan for expanding the technical areas of the field of statistics." *International Statistical Review*, **69**(1), 21–26. http://www.jstor.org/stable/1403527.

Cobb GW (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?" *Technology Innovations in Statistics Education*, **1**(1), 1–15. http://escholarship.org/uc/item/6hb3k0nz.

Cobb GW (2011). "Teaching statistics: Some important tensions." *Chilean Journal of Statistics*, **2**(1), 31–62. http://chjs.deuv.cl/Vol2N1/ChJS-02-01-03.pdf.

Cohen D, Henle J (1995). "The Pyramid Exam." *Undergraduate Mathematics Education Trends*, **7**(3), 2,15.

Committee on Professional Ethics (1999). *Ethical Guidelines for Statistical Practice*. http://www.amstat.org/about/ethicalguidelines.cfm, last accessed: 2015-05-19.

Davenport TH, Patil D (2012). "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1, last accessed: 2015-05-19.

Davidian M (2013a). "Aren't We Data Science?" *AMSTAT News*. http://magazine.amstat.org/blog/2013/07/01/datascience/, last accessed: 2015-05-19.

Davidian M (2013b). "The ASA and Big Data." *AMSTAT News*. http://magazine.amstat.org/blog/2013/06/01/the-asa-and-big-data/, last accessed: 2015-05-19.

Dhar V (2013). "Data Science and Prediction." *Communications of the ACM*, **56**(12), 64–73. http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext.

DiGrazia J, McKelvey K, Bollen J, Rojas F (2013). "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior." *Social Science Research Network*. http://ssrn.com/abstract=2235423.

Finger L, Dutta S (2014). *Ask, Measure, Learn: Using Social Media Analytics to Understand and Influence Customer Behavior*. O'Reilly Media, Inc., Sebastopol, CA.

Finzer W (2013). "The Data Science Education Dilemma." *Technology Innovations in Statistics Education*, **7**(2), 1–9. http://escholarship.org/uc/item/7gv0q9dc.pdf.

Franck C (2013). "Is Nate Silver a Statistician?" *AMSTAT News*. http://magazine.amstat.org/blog/2013/10/01/is-nate-silver/, last accessed: 2015-05-19.

Gelman A (2013). "The Tweets-Votes Curve." http://andrewgelman.com/2013/04/24/the-tweets-votes-curve/, last accessed: 2015-05-19.

Gould R, Baumer B, Çetinkaya Rundel M, Bray A (2014). "Big Data Goes to College." *AMSTAT News*. http://magazine.amstat.org/blog/2014/06/01/datafest/, last accessed: 2015-05-19.

Halvorsen KT, Moore TL (2001). "Motivating, monitoring, and evaluating student projects." *MAA Notes*, pp. 27–32.

Hardin J, Hoerl R, Horton NJ, Nolan D (2015). "Data Science in the Statistics Curricula: Preparing Students to 'Think with Data'." *The American Statistician*, **69**(4). http://arxiv.org/abs/1410.3127.

Harris JG, Shetterley N, Alter AE, Schnell K (2014). "It Takes Teams to Solve the Data Scientist Shortage." *The Wall Street Journal CIO Report (blog)*. http://blogs.wsj.com/cio/2014/02/14/it-takes-teams-to-solve-the-data-scientist-shortage/, last accessed: 2015-05-19.

Horton NJ (2015). "Challenges and opportunities for statistics and statistical education: looking back, looking forward." *The American Statistician*, **69**(2), 138–145. http://www.tandfonline.com/doi/full/10.1080/00031305.2015.1032435.

Horton NJ, Baumer BS, Wickham H (2015). "Setting the stage for data science: integration of data management skills in introductory and second courses in statistics." *Chance*, **28**(2). http://chance.amstat.org/2015/04/setting-the-stage/.

James G, Witten D, Hastie T, Tibshirani R (2013). *An introduction to statistical learning*. Springer. http://www-bcf.usc.edu/~gareth/ISL/.

Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Riche NH, Weaver C, Lee B, Brodbeck D, Buono P (2011). "Research directions in data wrangling: Visualizations and transformations for usable and credible data." *Information Visualization*, **10**(4), 271–288. http://research.microsoft.com/EN-US/UM/REDMOND/GROUPS/cue/infovis/.

Linkins J (2013). "Let's Calm Down About Twitter Being Able To Predict Elections, Guys." *The Huffington Post*. http://www.huffingtonpost.com/2013/08/14/twitter-predict-elections_n_3755326.html, last accessed: 2015-05-19.

Lohr S (2009). "For Today's Graduate, Just One Word: Statistics." *The New York Times*. http://www.nytimes.com/2009/08/06/technology/06stats.html, last accessed: 2015-05-19.

Moore DS (1998). "Statistics among the liberal arts." *Journal of the American Statistical Association*, **93**(444), 1253–1259. http://www.jstor.org/stable/2670040.

Murrell P (2010). *Introduction to Data Technologies*. Chapman and Hall/CRC. https://www.stat.auckland.ac.nz/~paul/ItDT/.

Nolan D, Temple Lang D (2010). "Computing in the statistics curricula." *The American Statistician*, **64**(2), 97–107. http://www.stat.berkeley.edu/users/statcur/Preprints/ComputingCurric3.pdf.

Patil D (2011). *Building data science teams*. O'Reilly Media, Inc.

Rajaraman A, Ullman JD (2011). *Mining of massive datasets*. Cambridge University Press. http://www.mmds.org/.

Rojas F (2013). "How Twitter can help predict an election." *The Washington Post*. http://www.washingtonpost.com/opinions/how-twitter-can-predict-an-election/2013/08/11/35ef885a-0108-11e3-96a8-d3b921c0924a_story.html, last accessed: 2015-05-19.

Swires-Hennessy E (2014). *Presenting Data: How to Communicate Your Message Effectively*. 1st edition. Wiley. http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118489594.html.

Tan PN, Steinbach M, Kumar V (2006). *Introduction to Data Mining*. 1st edition. Pearson Addison-Wesley. http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

Tufte ER (1983). *The Visual Display of Quantitative Information*. 2nd edition. Graphics Press.

Wasserstein R (2015). "Communicating the Power and Impact of Our Profession: A Heads Up for the Next Executive Directors of the ASA." *The American Statistician*, **69**(2), 96–99. http://www.tandfonline.com/doi/full/10.1080/00031305.2015.1031283.

Wickham H (2012). "my cynical definition: a data scientist is a statistician who is useful ;)." https://twitter.com/hadleywickham/status/263750846246969344, last accessed: 2015-05-19.

Wickham H (2014). "Tidy data." *The Journal of Statistical Software*, **59**(10), 1–23. http://vita.had.co.nz/papers/tidy-data.html.

Wickham H, Francois R (2015). *dplyr: a grammar of data manipulation*. R package version 0.4.2, http://CRAN.R-project.org/package=dplyr.

Wilkinson L (2006). *The grammar of graphics*. Springer.

Yau N (2011). *Visualize this: the Flowing Data guide to design, visualization, and statistics*. Wiley Publishing.

Yau N (2013). *Data points: visualization that means something*. John Wiley & Sons.

Zhu Y, Hernandez LM, Mueller P, Dong Y, Forman MR (2013). "Data Acquisition and Preprocessing in Studies on Humans: What is Not Taught in Statistics Classes?" *The American Statistician*, **67**(4), 235–241. http://dx.doi.org/10.1080/00031305.2013.842498.