https://doi.org/10.48009/1_iis_2012_382-390
Issues in Information Systems
Volume 13, Issue 1, pp. 382-390, 2012

OUTLINE AND EXERCISES FOR A NOVEL INTRODUCTORY COURSE IN DATA SCIENCE AND VISUALIZATION

Megan Squire, Elon University, msquire@elon.edu

ABSTRACT

Data Science is an increasingly popular term for the deliberate, methodological study of the principles and techniques involved in the storage, management, mining, and visualization of large amounts of data, as used to solve problems in diverse domains. This paper provides a working definition of Data Science and examines the relationship between this emerging field and other, more familiar disciplines already established in the undergraduate curriculum. We then provide an operational framework ("The Six Steps") for an introductory course in Data Science and Visualization. We provide a comprehensive description of concrete, relevant example assignments that fit cleanly into this framework. We conclude with examples of how this course can achieve secondary objectives of delivering an opportunity for emphasis on Writing and Information Literacy.

Keywords: Data science, visualization, curriculum, undergraduate, education, data mining, data analysis, writing, information literacy.

INTRODUCTION

Data seems to have somehow recently captured the public imagination in the United States. *Moneyball*, a movie ostensibly about baseball but with its unlikely hero a Yale economics major with expertise in data mining and computer-driven quantitative modeling, was a surprise box office success. On his late night show, comedian Stephen Colbert routinely uses word clouds and other data visualization techniques to explain American politics, and delivers comedy bits that hinge upon consumer familiarity with targeted marketing databases, association rule mining, and predictive analytics [for example, 11]. One of the most popular sites in online dating, OkCupid, proudly hangs its success on the tagline: "we use math to get you dates". Even the *New York Times* is running articles hailing "The Age of Big Data" and quoting professors saying that statistics and visual analytics are "the sexiest subjects around" [17]. In fact, the *New York Times* (and other newspapers, for example *USA Today* and the *Guardian* in the UK) are now providing data sets and interactive visualizations side-by-side with their news articles, and are even producing APIs (application programming interfaces) for readers to download the data and produce their own analyses.

A *McKinsey* report [2] recently declared that there is a talent shortfall in the United States for people with Data Science skills. They claim a shortage of 190,000 people with what they called "deep analytical" skills, who are "typically experts in statistical methods and data-analysis technologies" which include programming, database management, and data-centric graphic design, in addition to computational statistics. [12] With demand outpacing supply, entry-level jobs in Data Science are currently starting with a $73,000 annual salary. [9]

This paper outlines a course that attempts to answer, at the introductory and non-majors level, some of this need for data-centric thinking and Data Science skills among undergraduate students. The next section proposes a definition of Data Science, and provides a sample course description for a course offered by a Department of Computing Sciences both as a science credit for non-majors seeking general education credits, and as an introductory course in the Information Science major [10].

WHAT IS DATA SCIENCE?

In any new and interdisciplinary field, sometimes the appropriate terminology is not immediately clear or available. For example, the introduction to this paper cites popular media that, in trying to describe Data Science, make confusing and varied references to data analytics, data mining, data visualization, data journalism, statistics, and something they are calling "Big Data". Similarly, a few years ago when the field of data mining was new, practitioners and theorists struggled [6] with similar naming incongruities (cf. knowledge discovery in databases,

machine learning, computational statistics, data analysis). Now the terms *data mining* and *knowledge discovery* are themselves being used as synonyms for Data Science.

The academic catalog description for the Data Science and Visualization course upon which this paper is based reads as follows:

> **Data Science and Visualization. CSC 111/ISC 111.**
> The Internet is full of rich data sources that anyone can use to answer questions and solve problems. How can we process this data to uncover interesting patterns? How can we visualize this data to reveal trends or to spur additional questions? This course teaches students how to access online data, write programs to analyze the data, and use visualization tools to describe the patterns we find in a compelling way. Students of any major are welcome. No prerequisites. Offered fall and spring.

This definition of Data Science implies a statistical piece, a programming and software development piece, a database management piece, a graphical and visual design piece, and even a journalistic or rhetorical piece ("interesting", "compelling"). It is an interdisciplinary endeavor, but the description above frames the course in terms of everyday problems and is written in plain language.

Contrast to Related Fields

Students taking this Data Science course are mostly first- and second-year students, so they are in all likelihood not familiar with many of these other fields such as databases and software development (especially true for the non-majors), though they all will have probably taken an introductory statistics course. Thus, statistics may be the one complementary field to Data Science that will need the most differentiation for them.

To embrace and extend the relationship between statistics and Data Science, it is helpful to remind students that much of statistics is based around the notion of sampling. Sampling is used as a problem solving strategy when there is a scarcity of resources, for example when there is scarce data or scarce computing power: Given a pattern we discern in a sample of only *n* people, can we generalize this pattern to a larger population, and with what certainty? Given only *x* CPU cycles or *y* bytes of hard drive space, can we choose a representative group of data points to build a model, and have this model work with new data points later?

Data Science, in contrast, is predicated on abundance. It is a strategy that works well when there is an abundance of computing power and an abundance of data. Students already know that terabytes of data storage are available at any big box mart for less than a basket of groceries will cost. In Data Science, we solve a data-centric problem by writing programs (or using computational approaches) to collect data, we use databases (or similar technologies) to store it, we write programs and use software to find patterns in the data, we use basic graphic design skills and programming to visualize the patterns we found, and we use techniques and principles of human computer interaction (HCI) to allow others to interact with those visualizations.

> "Sampling is dead," said Abhishek Mehta, a big data lead at a large U.S. bank institution. "When banks stored petabytes of information on magnetic tape, it was impossible for them to develop appropriate models to measure risk without resorting to sampling techniques. Today we can run analysis on upwards of 50 petabytes of data to more accurately calculate risk." [16]

This remainder of this paper outlines a foundational structure ("The Six Steps") for teaching this broad, interdisciplinary subject at the introductory level in a way that will maximize flexibility as the discipline grows and changes. Special attention is paid to providing examples in the form of daily exercises and assignments that can reinforce the concepts, techniques, and tools of Data Science to undergraduate students.

THE FORMAT AND OUTCOMES OF THE COURSE

This section discusses the student learning objectives for the introductory course in Data Science and Visualization, and explains how these can be taught in a systematic way through a "Six Steps" process.

The goals for this course in Data Science and Visualization are as follows:

- Students will explain and perform tasks appropriate to each of the Six Steps of data science:
  1. Problem Formulation,
  2. Data Collection,
  3. Data Storage and Cleaning,
  4. Data Transformation and Mining,
  5. Data Visualization, and
  6. Problem Resolution

- Students will select and apply a variety of computational tools and methods to perform these Steps.
- Students will critique and evaluate the alternative methods and outputs of each of these Steps, given multiple formats, contexts, and purposes.

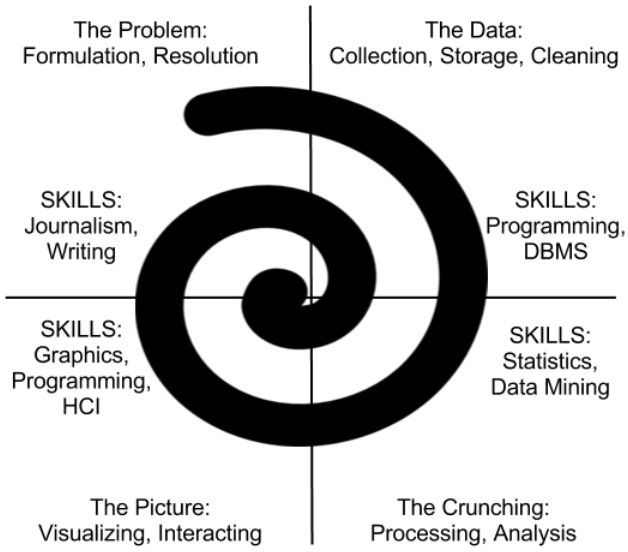

Figure 1. A spiral model for iterating through the Six Steps of Data Science

These broad course goals are met with specific course objectives that match each of the Six Steps (plus one more step for iteration):

1. Problem Formulation: Students will accurately and thoroughly describe data-centric problems using both statements and questions.
2. Data Collection: Students will evaluate and apply appropriate Internet-based search strategies to collect data needed to answer the problems they have formulated.
3. Data Storage, Cleaning and Curation: Students will select and create appropriate electronic structures to store the data they have collected. Students will annotate (curate) their stored data using metadata. Students will thoroughly describe the processes used to clean, store, and curate the data.

4. Data Transformation and Mining: Students will select and apply procedures, models, algorithms, and formulas to explore, transform, or summarize the cleaned data. Students will thoroughly describe the processes used to transform and mine the data.
5. Data Representation and Visualization: Students will select and apply appropriate visualization techniques to represent the data, find patterns in the data, or allow users to interact with the data. Students will explain how principles of human cognition and visual perception influenced their visual design choices.
6. Drawing Conclusions: Students will draw accurate and thorough conclusions about what the collected, stored, cleaned, curated, mined, transformed, and visualized data may mean for the problem as initially formulated. Students will identify and propose solutions to any limitations in their work, including those that occur in collection, cleaning, storage, transformation, or visualization and affect the conclusions presented. Students will express these conclusions and limitations in written and oral form.
7. Iteration: Students will recognize when and how to repeat these steps to clarify, to refine, and to more effectively use data to address the problem.

In further explaining the Six Steps of Data Science to students, Figure 1 may be helpful, especially in describing the spiral model (continuous refinement and iteration) through the Steps, and to describe which other skills and toolsets they can draw on for the course. (This also serves to inspire talented students to take follow-on courses.)

USEFUL ASSIGNMENTS, ORGANIZED BY TOPIC OR TECHNIQUE

With the Six Steps clearly defined as a framework for learning, the daily classroom routine for this course consists of demonstration of a new technique, discussion of the theoretical principles to support this technique, followed by a problem-based activity that allows the students to implement the Six Steps and integrate this new technique. (Our class is taught in a computer lab.)

Table 1 shows some real-world problems that have been used to teach Data Science. These examples are notable for their variety of visualization and analysis types, and their relevance to the current news cycle. As the course evolves and changes, the assignments used should easily track along with the interesting news of the day.

OTHER AREAS OF EMPHASIS

Data Science and Writing

Organizing each day's work around the Six Steps allows for students to submit a written report (and associated data artifacts like spreadsheets, maps, etc) in each class period. This can amount to a substantial amount of writing each day, which allows for significant feedback opportunities on writing to persuade, writing to inform, writing using visual media, and/or how to write about technical procedures to a non-technical audience. Figure 2 shows the average amount of writing done for each in-class daily assignment and homework assignment by each student for the Data Science and Visualization course offered in Fall 2011 and Spring 2012.

Averages for in-class daily assignments were around 400 words, while averages for a homework assignment were around 900 words. Student prompts for in-class assignments always include the students writing a report that covers each step of the Data Science process. Here is the text of a sample in-class assignment that followed a description of Edward Tufte's "slopegraphs"[22, 8]. Slopegraphs are a technique for showing a single data set that has been ranked two different ways. Students were asked to update a 2008 ranking of baseball teams using a slopegraph.

1. Update Ben Fry's baseball slopegraph of team salaries versus wins/losses using 2011 data. Collect your own data, clean it, and store it using one of the tools we have learned so far.
2. Use Google drawing tool to hand-draw and color a slopegraph of this data. (From Google Docs, choose "Create" and then choose "Drawing", then enter the data as textboxes. Draw the lines with the line tool.) Save your drawing as a PDF.
3. Write up your 6 steps, and upload your artifacts (cleaned data spreadsheet and PDF drawing).

Student responses to the Six Steps for assignment ranged from just under 200 words to over 600 words. Students spent the majority of their words explaining where they got their data, how the organized it in a spreadsheet to show ranked teams, and how they drew the visualizations. In a later homework assignment, students could optionally recall this assignment to produce slopegraphs on a different problem: airports ranked in terms of tarmac delays versus passenger throughput.

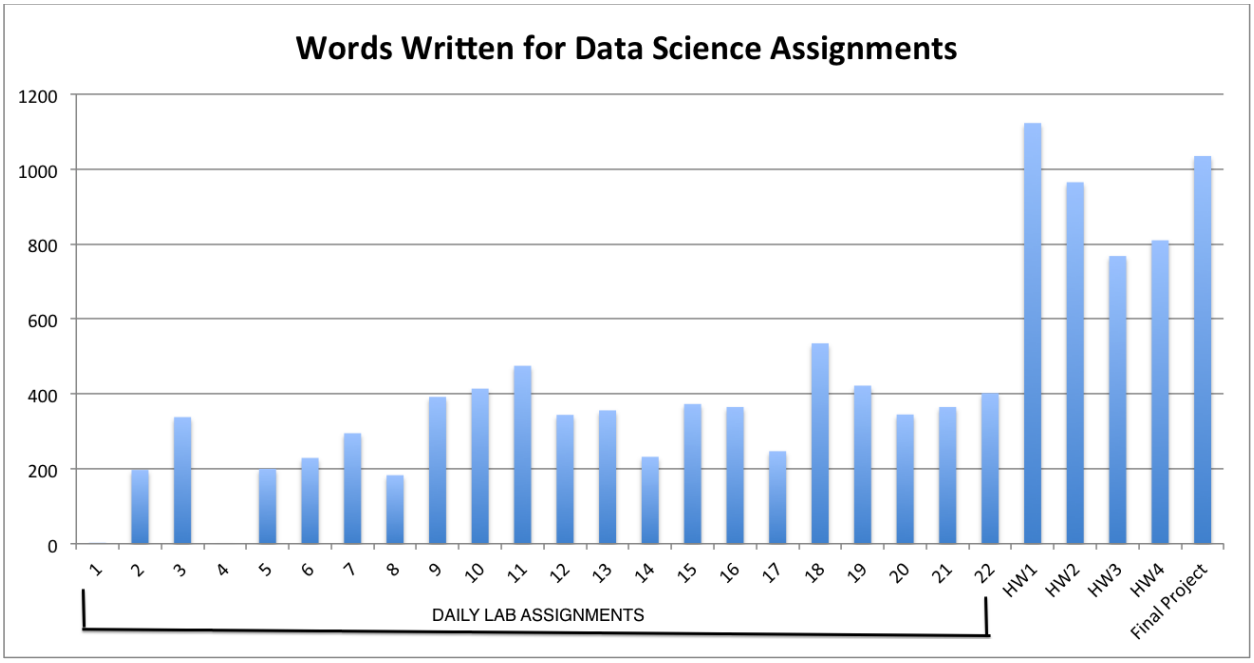

Figure 2. Writing in Data Science

By providing consistent feedback on the student's application of the Six Steps to a new problem each day, students have ample opportunity to refine their skills in both writing and Data Science. We find that rubrics are helpful here, and can be easily constructed to give different weight to whichever of the Six Steps are emphasized in this assignment. An example rubric follows.

This sample rubric can be used with every laboratory activity and every homework assignment, with the instructor altering the point values in each section as appropriate to the task at hand. (Some assignments will have more emphasis on data collection, others will have more emphasis on data cleaning, etc.) Additional guidance may also be given to the students in each section as scaffolding for constructing arguments appropriate to a particular assignment. For example, for the assignment called "Amazon.com and the Hot, Hot Summer", additional guidance for the *Problem Formulation* section read, "Student clearly refers to both claims made by Amazon, and describes how each of these will be supported or refuted." Additional guidance for the *Data Collection* section read, "Student clearly describes how this data relates to the two claims made by Amazon."

Data Science and Information Literacy

Another interesting side benefit of the Data Science course is that is contains many elements of an information literacy program. For example, the American Association of College and Research Libraries (ACRL) has five standards that, when met, indicate that a college student is information literate [1]. They state that the information literate student...

Standard 1: ...determines the nature and extent of the information needed.
Standard 2: ...accesses needed information effectively and efficiently.

> Standard 3: ...evaluates information and its sources critically and incorporates selected information into his or her knowledge base and value system.
> Standard 4: ...individually or as a member of a group, uses information effectively to accomplish a specific purpose.
> Standard 5: ...understands many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally.

From the ACRL guidelines, each of the five standards has associated performance indicators, 22 in total. Each performance indicator has a number of learning outcomes. The standards, indicators, and outcomes are arranged purposefully to move from lower-order to higher-order thinking.

Here is an example of how to apply ACRL information literacy standards in the Data Science and Visualization course. Consider the following ACRL standard, and its associated performance indicator and outcome:

- Standard (3): The information literate student evaluates information and its sources critically.
    - Performance Indicator (2): The information literate student articulates and applies initial criteria for evaluating the information and its sources.
        - Outcome (b): Analyzes the structure or logic of supporting arguments or methods.

In examining the Wall Street Journal article (see assignment description in Table 1), the student will view the accompanying graph and question whether the x and y axes are arranged correctly. She may notice that the "bins" arranged on the x-axis of the graph seem to be chosen arbitrarily. She will then attempt to apply Outcome (c): "Recognizes prejudice, deception, or manipulation". She will then start her Data Science procedure: She will locate the original IRS data (available online), and she will download this data, clean it to remove irrelevant columns, and create her own plot (ideally experimenting with different sized bins for the X-axis). She will write a report of her Six Steps of data science, beginning by formulating the problem, then describing her collection and cleaning procedures, describing her visualization of the data (including how and why she deviated from the original), and rationalizing why her solution is better and how it solves the problem presented in the problem formulation section. In this way, ACRL standards dovetail perfectly with the Six Steps of Data Science.

CONCLUSIONS

This paper outlines a course in Data Science and Visualization. Novel contributions of this paper include the presentation of a Six-Step framework for students to work with Data Science problems. These Six Steps provide a solid framework for answering a wide variety of interesting, cross-disciplinary problems that involve data at their core. Students will learn to formulate a data-centric question, to collect data to answer the question, to apply techniques for data cleaning and curation, and to apply procedures for analyzing and transforming the data. Students will also learn numerous techniques for producing persuasive and informative data visualizations, and for showing how this Data Science process provides value in answering the question as initially formulated. Secondary objectives of the Data Science course taught using this framework include additional opportunities for writing instruction and delivery of information literacy content.

REFERENCES

1. Association of College and Research Libraries. (2000). Information Literacy Standards for Higher Education. Report available at http://www.ala.org/ala/mgrps/divs/acrl/standards/standards.pdf.
2. Bughin, J., Livingston, J., and Marwaha, S. (2011). Seizing the potential of big data. *McKinsey*. October 2011. http://www.mckinseyquarterly.com/Seizing_the_potential_of_big_data_2870
3. CNN. (2011). Texas drops last meal for death row inmates. September 23, 2011. http://www.cnn.com/2011/09/22/justice/texas-last-meal/index.html

4. Elliott, C. The Navigator: Should tarmac delay rules become law? *Washington Post*. May 20, 2011. http://www.washingtonpost.com/lifestyle/travel/the-navigator-should-tarmac-delay-rules-become-law/2011/05/17/AFTqHy7G_story.html
5. Enron Email Data Set. (2009). Available at http://www.cs.cmu.edu/~enron/
6. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, Fall 2006. 37-54.
7. Fernandez, M. (2011).Texas death row kitchen cooks its last meal. New York Times (September 22, 2011). http://www.nytimes.com/2011/09/23/us/texas-death-row-kitchen-cooks-its-last-last-meal.html
8. Fry, B. (2008). *Visualizing Data*. Sebastopol, CA, USA: O'Reilly and Associates.
9. Hardy, Q. (2012). What are the odds that stats would be this popular? *New York Times (Bits).* January 26, 2012. http://bits.blogs.nytimes.com/2012/01/26/what-are-the-odds-that-stats-would-get-this-popular/
10. Heinrichs, L., Hutchings, D., Kleckner, M., Squire, M. (2011). Charting a new curriculum for a data-driven world. *Issues in Information Systems*, XII(2). 256-263.
11. Hill, K. (2012). How Target figured out a teen girl was pregnant before her father did. *Forbes*. February 16, 2012. http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
12. Lohr, S. (2011). The age of big data. *New York Times Sunday Review*. February 12, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
13. Mongabay Deforestation Statistics. (n.d.) Available at http://rainforests.mongabay.com/defor_index.htm
14. Politilines. http://politilines.periscopic.com/
15. Sachare, A. (n.d) 24-second clock revived the game. http://www.nba.com/history/24secondclock.html
16. Sims, D. (2011). Big data thwarts fraud. *O'Reilly Radar*. February 8, 2011. http://radar.oreilly.com/2011/02/big-data-fraud-protection-payment.html
17. Singer, N. (2011). When the data struts its stuff. *New York Times*. April 2, 2011. http://www.nytimes.com/2011/04/03/business/03stream.htm
18. Soper, S. (2011). Inside Amazon's warehouse. *The Morning Call*. September 18, 2011. http://www.mcall.com/news/local/mc-allentown-amazon-complaints-20110917,0,7937001,full.story
19. Spurlock, M. (2006). Don't Eat this Book. Berkley Trade.
20. Stynes, T. (2011). Amazon air conditions centers after criticism. *Marketwatch*. September 23, 2011. http://www.marketwatch.com/story/amazon-air-conditions-centers-after-criticism-2011-09-23
21. Sun-Sentinel. (2012). 'Irene' retired from list of hurricane names. Ft. Lauderdale Sun-Sentinel. April 13, 2012. http://bit.ly/HS0E6B
22. Tufte, E. (2001). *The Visual Display of Quantitative Information, Second Edition*. Graphics Press.
23. "Where the Tax Money Is". Editorial. *Wall St. Journal.* April 17, 2011. http://online.wsj.com/article/SB10001424052748704621304576267113524583554.html
24. Whitburn Project. (2008). The Whitburn project: 120 years of music chart history. Available at: http://waxy.org/2008/05/the_whitburn_project/
25. Yau, N. (2012). Spotlight on movie profitability. *Flowing Data Blog.* http://flowingdata.com/2012/03/02/spotlight-on-movie-profitability/
26. Zurer, R. (2011). Last meals. Wired. August, 2011. 68. Available online at http://katemacdonald.wordpress.com/2011/07/20/last-meals-feature-wired-magazine/

Table 1. Sample Assignments to teach the Six Steps

Name of Assignment	Description	References
Amazon.com and the Hot, Hot Summer	Amazon.com is accused of workplace safety violations in summer of 2011 in Pennsylvania warehouse. They claim that weather is to blame, not their policies. Students use data from online weather databases to prove or disprove the heat claims on the dates provided.	[18, 20]
Accuracy of McDonald's Food Testing	This is based on a popular book about fast food which claims that McDonald's food routinely tests higher-than-advertised in values for sodium, fat, and calories in several menu items. Students establish correctness threshold and test whether tested values deviate from that established threshold (over/under). Students use both whole servings and 100g servings.	[19]
Airline Tarmac Delays	Did the institution of a fine of $27,000 per passenger in April of 2010 for delays of longer than 3 hours have an affect on how long passenger aircraft sit on the tarmac at the 10 busiest US airports? Students collect data to show whether delays were affected, and if so, at which airports and for how long.	[4]
Trends in Popular Music	Students explore 50 years of Billboard music data for every "Hot 100" song, including beats per minute, number of weeks and position on the charts, genre, artist, label. They look for trends and visualize the patterns they find in this massive data set.	[24]
Movie Budgets Vs. Box Office Gross	Do larger budget movies make more money at the box office? Students use slopegraphs to discover money-losing movies and movies that were surprise (low-budget) hits.	[25]
Social networks in Enron email corpus	Students visualize the to-from (including cc: and bcc:) patterns in email correspondence between Enron employees as social networks.	[5]
Retired hurricane names	Names of particularly destructive Eastern Atlantic hurricanes are retired from use forever (for example Katrina and Andrew). Students use scatter plots to build visual profiles of the deadliest and costliest hurricanes, comparing destruction to hurricane category (1-5 on the Saffir-Simpson scale).	[21]
NBA shot clock and 3-point line	Students use tools for measuring one attribute before and after a single "moment in time", for example showing how the addition of the NBA shot clock or 3-point line affected game scores, and by how much (taking into account team and division).	[15]
Political candidate speech	Students use word clouds and text analysis tools to discover interesting or hidden patterns in candidate speeches.	[14]
Deforestation	Students compare and contrast the abilities of mapping tools, including map bubbles and heatmaps, to show global patterns of deforestation.	[13]
Last meals on Death Row	Was Texas justified in abolishing the last meal in 2011? How much does a last meal cost compared to the cost of feeding a prisoner normally? Students gather data on the content and cost of the last meals requested by prisoners on Death Row, the average cost to feed prisoners, the average cost of a meal for the "typical American".	[3, 7, 26]
Which tax bracket pays the most taxes?	Students read the Wall St. Journal editorial in April 2011 (and articles critical of it) about the distribution of tax filings between brackets. Students collect data on taxable income by tax bracket, then establish different "bin" sizes on X-axis and see how this affects the visualization.	[23]

Table 2. Sample Rubric for a 100-point Assignment in Data Science and Visualization

Item	Possible Score	Your Score
Problem Formulation: Student describes completely and accurately the problem domain (what is the question we are trying to answer, and what the impact might be if we were to answer this question. This section answers the question "why are we doing this?"). Student clearly refers to the case we are studying, and describes how data will support or refute the case.	5	
Data Collection: Student located high-quality data to help solve the problem described above. The original location(s) of this new data is described accurately and completely, with references given and a rationale for why this is a good source and which problem it is going to help solve. If this data came from a web site, what specific fields did you use? What form values did you fill in? (Screenshots are helpful. Make sure another person could follow your steps and reproduce exactly what you did and get the same numbers.)	20	
Data Storage/Cleaning/Curation: Student placed data into appropriate storage mechanism for the assignment (spreadsheet, text file, etc). Data was submitted with assignment in an easily-understandable, useful format. Data is labeled and organized. Data cleaning procedures were described completely and in sufficient detail that another person could easily end up with the same cleaned data or facts if they needed to repeat the study.	10	
Data Processing & Analysis: Student performs basic analysis procedures on the data (equations, formulas, processing) if necessary. Student describes completely and accurately not only what equations or formulas or procedures were used, but why these were necessary to answer the question formulated in step one (and why they are better than the alternatives if any).	10	
Data Visualization: Student designs and implements appropriate visualization(s) for the data. Visualization mechanisms can be charts, graphs, tables, or other, but the method(s) chosen must be accurate, complete, and appropriate for the problem being solved. Student describes why alternatives to this visualization method were rejected.	25	
Drawing Conclusions: Student describes completely and accurately the conclusions that can be drawn from these procedures (collection, storage, analysis, visualization), and clearly ties this conclusion to the problems formulated in step one. Student clearly sates limitations of the work. (Did we answer the original questions or would we have to do additional research? What data or analysis is still missing, if any?)	20	
Organization & Quality of Writing: Student displays excellent grammar, spelling, layout, organization skills. Writing is interesting and compelling. References are given in sufficient detail that reader can find the sources and replicate the work. References are high quality and are original sources if possible.	10	