Preparing All Students for the Data-driven World

Christo Dichev, Darina Dicheva
Department of Computer Science
Winston Salem State University
601 S. Martin Luther King Jr. Drive
Winston Salem, NC 27110
dichevc@wssu.edu
dichevad@wssu.edu

Lillian Cassel, Don Goelman
Department of Computing Sciences
Villanova University
800 Lancaster Avenue
Villanova, PA 19085
lillian.cassel@villanova.edu
don.goelman@villanova.edu

Michael Posner
Department of Mathematics and Statistics
Villanova University
800 Lancaster Avenue
Villanova, PA 19085
michael.posner@villanova.edu

ABSTRACT
As Data Science gains in importance in industry, government and society, the creation of appropriate courses and teaching activities is necessary for building up a competent workforce. In this paper, we describe our experience of the development of a low-level undergraduate introductory Data Science course. The course is without prerequisites and multidisciplinary in terms of targeted skills, involving expertise in computer science and statistics and incorporating different domain areas, such as astronomy and social networks. We describe the general design decisions, the adopted learner-centered approach and the stepwise development of the course. The first version of the course is taught at Winston-Salem State University (WSSU) in Spring 2016 and the second iteration will occur at Villanova in Fall 2016.

Categories and Subject Descriptors
K.3.2 [**Computers and Education**]: Computer and Information Science Education – *Computer science education, Information systems education; G.3* [**Probability and Statistics**]*: Statistical computing, Statistical software.*

Keywords
Data Science, General Education Course, Flipped Learning.

1. INTRODUCTION
The pervasiveness of data in the modern digital world is rapidly and fundamentally changing the way institutions and organizations operate, and supports increasing levels of evidence-based solutions. Data Science is the discipline that develops models, algorithms, processes, and frameworks for interfacing with data to extract understanding of system behaviors and to support decision making, especially in the face of rapidly changing situations. The Chronicle of Higher Education also identified data-driven computational science as one of "five emerging areas of study," [6] and it is now widely viewed as the third basic methodology in scientific inquiry, complementing theory and experiment. The Fourth Paradigm of Science: Data-intensive Scientific Discovery identifies the key role of data analysis in modern scientific discovery [10].

According to a McKinsey Institute study [11], there will be a shortage of nearly 200,000 people with deep analytical expertise and another 1.5 million managers with basic analytics skills by 2018. This need motivated our work on developing an introductory Data Science course intended to increase students' understanding of the role that data play as well as to increase their ability to use the technologies associated with data acquisition, mining, analysis, and visualization.

The reason for proposing Data Science as an entry-level course is to make it accessible for a broad range of students. Students graduating with a traditional bachelor's degree in biology, chemistry, mathematics, physics or astronomy generally do not have the required computational background necessary to participate as productive members of interdisciplinary teams, which frequently require computational and data science expertise. With this in mind, the skills gained in the Data Science course can best be seen as an enabler to applications of another domain. Natural science students, for instance, will learn how data-intensive computations can be brought to bear on the scientific problems. Computer science students, knowledgeable in databases and programming, will learn how to refine techniques in light of the special implementation challenges that Big Data involves or how to organize data, synthesize it, look at it critically, mine, and analyze it using appropriate statistical techniques. Mathematics or Statistics students, having studied modeling from a theoretical or applied perspective, can employ Data Science as a practical solution to problems and recognize its importance in extracting and assembling usable data. Students in social sciences and business, as well as nearly any other field that captures and uses data, will benefit from this material as well. Future journalists, policy makers, thought leaders, business managers, etc. who will commission or interact with the results of analysis of data on an ongoing basis will be empowered to better communicate or make better decisions with a cursory knowledge of data science techniques.

Beyond the obvious need in scientific domains, there is an increasing need to educate a technically literate populace and work force [11], even in those areas not directly under the umbrella of data science. This concern implies the importance of course curricula that would be accessible without formal prerequisites. Rather than competing with existing data analytical courses, we aim with the proposed Data Science course to enhance and expand the learning experience to a wider range of students starting at a lower level. Thus the course is envisioned as multidisciplinary in terms of targeted skills, involving expertise in computer science and statistics and incorporating different domain areas, such as astronomy and social networks.

We also envisage that the developed learning materials and modules could be integrated into other curricular settings. Thus, a course in sociology might use one or two modules to enable students to analyze a dataset as part of the course project, without

ADMI 2016, March 31–April 3, 2016, Winston Salem, USA.

having the time or ability to use all of the material for a full data science course. One of the challenges of the project is to make the modules as independent as possible, and to make a clear statement of all dependencies.

The paper is organized as follows. The next section sheds light on the related work on developing Data Science courses. Section 3 presents our approach to the design of a Data Science course. Section 4 describes the version of the course developed and being taught at Winston Salem State University and preliminary plans for a related course to be offered at Villanova University. Section 5 concludes the paper, reflecting on the experience thus far and the remaining work to be accomplished.

2. BACKGROUND AND RELATED WORK

Introducing Data Science in the university curriculum is a recent trend reflecting the rapidly increasing market need of specialists in 'big data'. Several national study groups have issued reports on the urgency of establishing scientific and educational programs to face the data deluge challenges. Some examples include the NSF Atkins Report of Revolutionizing Science and Engineering Through Cyberinfrastructure [6], the Computing Research Association Report on Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda [16], the NSF report on Cyberinfrastructure Vision for 21st Century Discovery [3], and the ASA statement on the role of statistics in Data Science [18]. As a response, the first Data Science courses appeared.

In a review conducted in November 2013 we found that there were few courses, and fewer programs, offered on Data Science, and most were exclusively at the graduate level. Notable exceptions were bachelor's degrees in Data Science offered at the College of Charleston and at the University of Northern Kentucky. Examples of courses at the undergraduate level included the Data Science course offered by Coursera (Bill Howe, University of Washington), by the College of Charleston, and by Elon University. The College of Charleston offers a Major in Data Science. There were a few Data Science courses offered as continuing education courses (e.g. by Harvard Extension School and the Professional & Continuing Education School of University of Washington), a few Certificate Programs in Data Science, consisting of 3 to 4 courses (University of Washington, Columbia University, Stanford University, Syracuse University, Indiana University), and a limited number of Masters Programs (UC Berkeley School of Information, Columbia University, Northwestern University, North Carolina State University, NYU, Stanford, University of Florida, and Rensselaer Polytechnic Institute, for example).

Two years later, we witness an explosion in the number of Data Science courses offered at different levels. A Google search with the key phrase "Data Science course" returned 29,500 results. The majority of the courses are still offered as master or certificate courses, or undergraduate courses that are part of Computer Science, Statistics, Mathematics or Data Science majors. All the courses, however, even the undergraduate and certificate ones that do not explicitly require prerequisite courses, expect some level of computational and/or mathematical skills. For example, a Data Science course in a Computer Science department, even when it does not require programming skills in a particular language, would presume computer literacy, knowledge of basic computing concepts, such as variables, data types, assignment and flow control statements, etc., as well as algorithmic thinking. Differently, we are developing course materials that neither require nor presume any specific knowledge and skills, that is, an introductory undergraduate course that is self-contained and without prerequisites. This will allow offering the course not only to Computer Science and Statistics/Mathematics majors, but also as a general education course. The latter is a step towards filling an important gap - "the need of many more college graduates, including those with non-STEM degrees, to be proficient in the application of analytics in their field, which provides an opportunity to take a fresh look at the skills of liberal arts graduates and the value of liberal arts degrees" [2].

In the past few years, a few articles related to undergraduate Data Science courses also appeared. What is noticeable is that their authors seem to tackle the problem from a single perspective – either statistics or computer science (e.g. programming). Among those that target the introduction of computing concepts in statistic courses are [1, 2, 12, 14]. Horton and colleagues, for example, propose the integration of data management skills, which they call "precursors to data science" in introductory and second courses in statistics [12]. They suggest leaving more sophisticated data management and manipulations for the second course. The papers in the other group address the infusion of statistical concepts in computing sciences; for example, in databases, machine learning, computer vision, and big data, as well as the use of computers in data modeling, prediction and analysis [4, 7, 9, 13]. For example, Olaf and Sanft proposed a statistics-infused Introduction to Computer Science course, where the focus is on teaching students to write small amounts of code in the context of work in other fields [9]. In [13] Bill Howe and fellow leaders in the database education community point out that the data management challenges – the acquisition, cleaning, integration, manipulation, and sharing of data, their flexibly and scalability – are the real bottleneck for data scientists, and advocate the need for updating the traditional "Intro to Databases" course and its transformation to an "Intro to Data Science" course.

An interesting approach is presented by Gil [7], where the goal is to develop an open and modular course for data science and big data analytics that is accessible to non-programmers. The course is designed to cover major concepts that are useful to understand the benefits of parallel and distributed programming not relying on a programming background but through the use of a semantic workflow system. However, it is intended as a higher-level Data Science course for non-CS students (it introduces the concepts of parallel computing, semantic metadata and semantic workflow, provenance, data stewardship, etc.).

The targeted audience, together with our learner-centered approach to the design of the course distinguishes us from the few existing Data Science undergraduate courses.

3. OUR PROJECT

3.1 The goals for our project

This project is designed with a primary goal to define and support an introduction to data science for undergraduate students with materials for a flipped classroom approach. To accomplish that goal, we proposed the following specific objectives:

- Define the base content for an introduction to data science that does not rely on previous courses.
- Using results of educational research on the benefits and limitations of the flipped classroom approach, prepare materials for the introduction to data science in flipped

classroom mode that align with proposed course learning goals.

- Evaluate the transferability of learning materials and modules for integration into other curricular settings.
- Support the use of student-centered and student-led learning in diverse classroom settings.
- Document the challenges associated with the development and application processes involved in the design of new flipped classroom courseware
- Share the developed learning resources through direct contact with individual faculty, faculty workshops, and through posting at the NSF NSDL Computing Pathway, Ensemble (www.computingportal.org) and the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) (www.causeweb.org).

The project is motivated by the belief that data science is a pervasive component of a wide variety of fields and that students from nearly all majors benefit from some level of understanding about what is implied by the term, what is required to get results, and how those results can be used.

3.2 General Design Decisions

3.2.1 A learner-centered approach: flipped learning model

The amount of factual information in an introductory Data Science course is low relative to introductory courses in other disciplines, but the amount of assimilation of those basic ideas that is required is relatively high. It is difficult to learn to program and apply statistical principles to data by merely watching a set of lectures. The difficulty lies not only in mining, analyzing and applying the statistical methods to real data, but also in interpreting the results. Students in the traditional classroom are largely left to figure out such techniques on their own. Therefore the traditional teaching models seem not quite effective for courses such as Data Science.

The problem originates from the fact that in the traditional classroom model the information transfer takes place in class, with assimilation of that information taking place outside the class. The information transfer -- attending lectures and taking notes -- is far simpler than assimilation and therefore less needful of help from an instructor. However, in a traditional classroom this is the time when the instructor is most fully available. What we see as critical in the Data Science context is the ability of students to apply techniques and knowledge to existing problems. These are often the types of tasks relegated to homework that students complete outside of class and on their own. Yet they are also the tasks that most benefit from instructor support. We believe that by scaffolding the process of analyzing and problem solving through in-class activities, students would gain more experience with applied skills and would better assimilate the data science material.

With this in mind we decided to flip the Data Science class with an intention to shift from passive to active learning and to focus on the higher order thinking skills such as analysis, synthesis and evaluation. The transfer and memorization of raw information can happen, often better, when students work at their own pace outside of class. On the other hand, the assimilation and application of the concepts that are covered in a typical introductory Data Science course require much more complex thinking and reflection; these activities are best done, or at least best begun, in the presence of the instructor and peers. Facilitating deep understanding of the Data Science material and enabling students to apply knowledge or skills in new contexts can be difficult to accomplish when working outside of class in isolation. Therefore we plan to use the in-class time on application and exploration rather than on delivery of information. We will do that by offloading the relevant passive lecture content to homework outside the classroom, creating additional time in the classroom for active and higher level learning.

3.2.2 Balancing the course topics between computer science and statistics

Our project team necessarily includes participants from computer science and from statistics. Naturally, that leads to diverse opinions about the relative importance of topics as well as which tools are best to accomplish the learning and tasks that we will ask of students. Available options include a variety of tools that provide results without the students having to do detailed implementation work. Some useful tools we have identified are Weka [17] and Watson Analytics, which provides access to machine learning and visualization, respectively, without student programming. When programming appears, statisticians prefer to use R, while computer scientists prefer Python. Our project team will develop modules for each choice, allowing those who use them to have the one that best suits their needs.

Projects help provide student engagement. For example, an early project plan involved use of the Twitter API to retrieve data that the students would find interesting, then provide a collection of tasks to be performed on the data. The Twitter data provides both numeric and textual data and a good variety of potential projects. Identifying the topic themes of the collection, for example, requires some thought about how themes can be recognized and then some kind of programming to compare the data to the theme indicators. Numeric characteristics include the number of tweets in a topic area, the identification of the most often re-tweeted message, features of frequency both by topic and by individual, and more. That project has proven useful both in exercising project goals and in engaging student interest.

3.2.3 Stepwise development: two versions of the course

Development of the course content, and the associated modules, evolves in an iterative fashion. The first version is taught at Winston-Salem State University (WSSU) in Spring 2016, in a more traditional (not flipped) format. The second iteration will occur at Villanova in Fall 2016 as a collaborative undertaking between Computer Science and Statistics faculty and delivered in the flipped format. The Villanova course will build on the lessons learned on the first offering at WSSU.

4. THE DATA SCIENCE COURSE VERSIONS

This section describes the current state and offering of the two courses: the WSSU Data Science course and the Villanova Data Science course.

4.1 The WSSU Data Science Course

The course development process was guided by two general observations:

- Data science is emerging as an academic discipline, defined not by a mere amalgamation of interdisciplinary fields but as a body of knowledge.
- The ability to use computational tools to collect, organize, visualize, and analyze data is a valuable skill both inside and outside of computer science and statistics.

Thus the driving insight behind this new course is that whether we are talking about 'data science', data analytics or 'big data', the trends in addressing problems using both computation and data will grow and will be of interest to a larger and more diverse group of students. Pragmatically, such students will benefit from a data science course adapted to their needs. As computation and data become more necessary in a variety of contexts, this course is aimed at supporting a growing and diverse body of non-majors, while still providing potential majors a beneficial start. Through such a course students should acquire skills, language and tools for making sense of data and visualizing it as evidence to support their reasoning. The primary challenge with an introductory course such as this one is to make it appealing to students with diverse interests and backgrounds. The success of any course structure shaping the course content for non-homogeneous audience lies in the answer of this challenge. One of the strategies that can help such a data science class to become more appealing is to make it more relatable to students. This implies that the course should not only teach fundamental data science literacy topics but also demonstrate the power and value of data science and how it relates to various practical problems.

Selecting the appropriate content for general education courses is an important part of attracting and retaining students. In this context, we tried to apply the lessons learned from introductory computing courses. A problem with many of these courses is that they focus too much on the syntax and the mechanics of programming and not enough on the motivation and the practical use of computing to solve problems in a variety of disciplines. As a result, traditional introductory computer science courses have had little success in engaging non-computer science majors. The alternative is to incorporate project and case studies motivating the taught concepts by showing students how to use data science as a tool to solve real life problems and thus make the course more practical and relevant.

From the perspective of the intended audience, the purpose of the course is two-fold: satisfying students seeking to acquire some data science skills that could be used in the future either in their field of study or their career, and adjusting the level to satisfy students seeking to fulfill the Information Literacy area of knowledge requirements for WSSU general education courses.

Effectively addressing the needs of the intended audience required us to reexamine many of our initial assumptions about the course. For example, which computing, statistical, and visualization topics are most beneficial to this group of students? In this aspect one important course development decision was the choice to use two languages, Python and R, in different course offerings, and compare the results. The motivation was that Python is open source widely used among data scientists, who may use Python for their initial data collection and formatting, even if they plan to use R in later stages of analysis. Python has a plethora of associated teaching materials for introductory students, including a well-written Python books for a general audience, such as by Allen Downey [5] and Charles Severance [15] and a well-designed beginners' code visualization tool by Philip Guo [8]. On the other hand, R is one of the dominant programming languages in use by statisticians. R is open-source, and it features an extensive list of well-documented packages for use in a broad variety of data-analysis activities. In addition, RStudio is a very useful front-end interface to R. In the early part of our experiment, Python is the language tool for the "Introduction to Data science" course offered at WSSU, while both R and Python will be used in the Data Science course offered at Villanova.

Table 1. Introduction to Data Science Course Topics

(* denotes the number of 75 min. classes)

Topics	C*	Details
Introduction and motivation	1	
Introduction Python	9	Variables, Expressions, Statements, Conditionals, Functions, Iteration, Strings, Lists, Dictionaries, Word frequency analysis, Random numbers, Word histogram, Most common words, Markov analysis, Files, Databases, Classes and objects
Data Preprocessing	2	The source of our data: The National Survey of Family Growth
Basic Statistics	6	Types of Data, Collecting Data, Biased Samples, Design of Experiments, Confounding, Controlling Effects of Variables, Describing, Exploring and Comparing Data, Normal Probability Distributions, Estimates and Sample Sizes, Hypothesis Testing, Inferences from Samples, Correlation and Regression
In class project	3	
Machine Learning	2	Classification learning, Association learning, Numerical prediction, Clustering, Discretization, Supervised and unsupervised ML
Classification	2	Classification problem, Numerical vs. nominal attributes, ARFF file format
K-Nearest Neighbors	2	KNN algorithm, similarity, accuracy
The Naïve Bayes' Classifier	1	Document classification, text preprocessing, The NaiveBayes Multinominal classifier
Decision Trees	1	Decision trees, The C4.5 classifier, Confusion matrix
Clustering	1	The KMeans algorithm, Distance metrics

The introduction to Data Science is currently taught in the Department of Computer Science at WSSU (Spring 2016). The Python sections of the course are similar to many other introductory programing courses, although truncated. Our intent was to provide enough Python background to allow students to

take some data, and write their own Python code to clean it up. We thus introduced the most basic Python data structures and control features. Outside of Python, we also intend to introduce briefly basic statistics concepts, data preprocessing, relevant machine learning and data mining topics (see Table 1). The course topics indicate that the course is aimed at familiarizing students with tools that would allow them to obtain data from different sources, and make them aware of a range of platforms such as Weka that could be used to transform, organize, and analyze the data. The proposed topics also indicate that introductory Data Science courses face a unique challenge in transmitting a great deal of necessary background information while also teaching higher-level thinking.

While course topics define content to be taught, instructional strategies determine the manner in which teaching and learning activities are designed in order to facilitate the achievement of learning goals. For the WSSU "Introduction to Data Science" course we have adopted a hybrid flipped classroom model where a significant part of the class time is used to discuss cases and work through problems together, while lectures are available for review at home. This approach translates in instruction through brief class lectures and discussions, and lab projects completed during class hours. The laboratory projects are group (2-3 students) work, where the students apply programmatically or use tools incorporating concepts from the corresponding lectures. In such a hybrid approach, students benefit from both the traditional face-to-face instruction, and a self-directed, self-paced learning experience.

There are 11 students currently enrolled in the WSSU Data Science course. Table 2 shows basic course demographic information. It also illustrates that the enrolled students are drawn from a wide variety of disciplines across campus.

Table 2. Course Demographics

Student Level	
Freshman	36%
Sophomore	36%
Junior	18%
Senior	9%

Student Major	
Computer Science	36%
Psychology	18%
Social Work	9%
Exercise Science	9%
Early Education	9%
Undecided	9%

At the start of the course, we administered a pretest, designed to assess the students' backgrounds and preparedness for the course. The pretest was given to all students. It contained 25 questions designed to assess the students' previous experience and knowledge level. The questions were designed to cover two general areas: computation and statistics. The results of the pretest are summarized in Table 3.

Table 3. Pre-test Scores

Questions	**Class Average Score**
All questions	62.72
Computation-related questions	59.55
Statistics-related questions	65.77

The pretest included 9 additional questions intended to assess student attitudes about the usefulness, relevance and worth of Data Science in personal and professional life. An interesting result from this part of the pretest is that it indicates a strongly positive attitude towards Data Science.

4.2 The Villanova Data Science Course

Villanova will offer a version of this course for the first time in Fall 2016. The course will be co-taught by project team members from the Department of Computer Science and the Department of Mathematics and Statistics, and will be listed with both department designations. Villanova is highly supportive of interdisciplinary learning experiences for undergraduates and this course will provide an example of collaboration between the participating departments, but also will involve input and contributions from faculty from other departments, representing the domains to which the lessons will apply. The course will not be open to computer science or mathematics majors, emphasizing that it is accessible to a broader audience. Students from all other departments will be welcome. The makeup of the class enrollment will determine the domain content for the course.

As stated previously, the Villanova course will differ from the WSSU course in two main aspects. First, the course will be offered in a flipped format. Second, the course will focus less on programming and more on using existing tools that reduce the burden on the students of prerequisite computing skills.

5. CONCLUSION

From our review of the literature and discussion with colleagues, the need for increased Data Science literacy and skills is clear. Students in the sciences, both natural and social, as well as in the arts and humanities, will benefit from an awareness of how the massive amounts of data are being processed and how knowledge is extracted from it. Designing this course using a flipped classroom model will better allow teachers to access and use these material in a variety of courses and tailor a course to their students, their environments, and their individual skills and knowledge.

Our experiences, thus far, have led to a few findings. Distilling data science course into a limited number of topics with no prerequisite knowledge is challenging. This is especially true when done in an interdisciplinary manner. We hope that our collaboration between Computer Science and Statistics provides an improved product, both in terms of student learning outcomes and acceptance across the disciplines. There are also a variety of choices to be made around the tools that can be used to perform the data acquisition, analysis, and visualization. The initial experience of teaching the WSSU course indicates that making an introductory Data Science course more attractive and interesting for diverse student population requires a more practical and example-based teaching involving concepts relevant to the students mind.

6. ACKNOWLEDGMENTS

This work is supported by the projects NSF DUE-1432257 and 1432438: IUSE Collaborative Research: Data Computing for All: Developing an Introductory Data Science Course in Flipped Format (09/01/2014-08/31/2017). Dr. Gabriele Bauer supports the work as project evaluator. Graduate Assistant Manaswini Arcot provides support for our presence in computingportal.org and assists in analysis of data from class surveys.

7. REFERENCES

[1] Baumer, B. 2015. A Data Science Course for Undergraduates: Thinking with Data. *The American Statistician*, 69, 4 (October 2015), 9, http://arxiv.org/pdf/1503.05570v1.pdf.

[2] Cárdenas-Naviaa, I. and Fitzgerald, B. K. 2015. The Broad Application of Data Science and Analytics: Essential Tools for the Liberal Arts Graduate. *Change: The Magazine of Higher Learning*, 47, 4 (2015), 8.

[3] Cyberinfrastructure Vision for 21st Century Discovery. National Science Foundation, 2007, http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

[4] Demchenko, Y., Gruengard, E. and Klous, S. 2014. Instructional Model for Building Effective Big Data Curricula for Online and Campus Education. In *Proceedings of the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science* (CloudCom), (15 - 18 December, 2014). IEEE.

[5] Downey, A. B. 2016. *Think Python*. O'Reilly, http://www.greenteapress.com/thinkpython2/index.html.

[6] Fischer, K. and Glen, D. 2009. 5 College Majors on the Rise, http://chronicle.com/article/5-College-Majors-On-the-Rise/48207/.

[7] Gil, Y. 2014. Teaching Parallelism without Programming: A Data Science Curriculum for Non-CS Students. In *Proceedings of the 2014 Workshop on Education for High Performance Computing* (EduHPC).

[8] *PythonTutor* http://pythontutor.com/.

[9] Hall-Holt, O. A. and Sanft, K. R. 2015. Statistics-infused Introduction to Computer Science. In *Proceedings of the Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. ACM.

[10] Hey, T., Tansley, S. and Tolle, K. 2009. The Fourth Paradigm. Data-Intensive Scientific Discovery. *Microsoft Research*, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf .

[11] Holdren, J. P. and Lander, E. 2012. Engage to Excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. *Executive Office of the President of the United States*.

[12] Horton, N. J., Baumer, B. S. and Wickham, H. 2015. Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *Chance* (2015), http://arXiv:1502.00318v1.

[13] Howe, B., Franklin, M. J., Freire, J., Frew, J., Kraska, T. and Ramakrishnan, R. 2014. Should we all be teaching "intro to data science" instead of "intro to databases"? In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 2014. ACM

[14] Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P.,, Peng, R., Roback, P., Temple Lang, D., and Ward, M. D. 2015. Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*, 69, 4 (2015), 11. http://arxiv.org/ftp/arxiv/papers/1410/1410.3127.pdf.

[15] Severance, C. 2013. Python for informatics: Exploring information. *CreateSpace Independent Publishing Platform*, 2013, http://www.py4inf.com/book.php.

[16] Shaaron Ainsworth, Margaret Honey, W. Lewis Johnson, Kenneth Koedinger, Brandon Muramatsu, Roy Pea, Mimi Recker and Weimar, S. 2005. Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda. *Computing Research Association.* http://www.cra.org/reports/cyberinfrastructure.pdf.

[17] Witten, I. H., Frank, E. and Hall, M. A. 201. *Data Mining: Practical Machine Learning Tools and Techniques*, http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

[18] ASA statement on the role of statistics in Data Science. 2015. Available at: http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/.