Big Data science education: A case study of a project-focused introductory course

Jeffrey Saltz, Robert Heckman
jsaltz@syr.edu, rheckman@syr.edu

Syracuse University, USA

Abstract. This paper reports on a case study of a project-focused introduction to big data science course. The pedagogy of the course leveraged boundary theory, where students were positioned to be at the boundary between a client's desire to understand their data and the academic class. The results of the case study demonstrate that using live clients within a team-based, project-focused course provides a useful platform in which to teach an introduction to data science course to graduate students across a range of backgrounds. While more work needs to be done to compare different possible pedagogies for teaching an introduction to data science course, the results of this study indicate that one successful approach is a project-focused class that puts students at the boundary between the academic context of the course and solving a real-world problem for their client.

Keywords: Data science, big data, project-based learning, graduate education

Introduction

Data Science is an emerging field that combines expertise across a range of domains, including computer science, data management, effective communication, and statistics. Big Data is a related field, often thought of as a subset of data science, in that data science applies to large and small data sets and covers the end-to-end process of collecting, analyzing and communicating the results of the analysis.

Data scientists play important roles in the four A's of data: data architecture, data acquisition, data analysis and data archiving. A data scientist provides input to system architects on how data needs to be routed and organized to support analysis, visualization, and presentation of the data to the appropriate people. Data acquisition, from a data scientist's standpoint, is a critical phase in which data is obtained in the right format, linked to other relevant data and screened to ensure appropriateness for analysis. Analysis is where the data scientists are often most heavily involved: summarization of the data, using portions of data (samples) to make inferences about the larger context and visualization of the data by presenting it in tables, graphs, and even animations. Finally, the data scientist must become involved in the archiving of the data. Preservation of collected data in a form that makes it highly reusable - what some people might think of as "data curation", is a often hard to do because it is so difficult to know how the data might be used in the future.

With the increasing ability to collect, store and analyze an ever-growing diversity of data that is being generated with increasing frequency, the field of data science is growing rapidly. The need for Data Science skills has been rapidly growing for many years, and this has lead to a rise in the number of data science programs (O'Neil, 2014). However, little has been written about how to best educate data scientists, especially for an introductory data science course, a course that could be useful to students that plan to specialize in data science as well as other students that need some data science knowledge as part of their broader education.

The purpose of this case study is to explore the viability of doing real-world projects within an introduction to data science course. While it has been well documented that there is tremendous value in having students do real-world projects, the viability of doing such real world projects has not been explored within a data science context. Hence, this paper seeks to address the following research questions:

RQ 1: Could students, with diverse educational backgrounds, succeed in a project-focused introduction to data science class?

RQ 1a: Could students with more advanced technical knowledge (ex. advanced statistics, visualization) learn a significant amount during the semester while other students, with minimal background, were able to learn the foundational concepts of applied data science?

RQ 1b: Could students learn enough during the semester to work on real-world problems and deliver real value to their clients (or would the problem need to be simplified into a "toy example")?

RQ 1c: Would clients find the student projects of value?

Previous Research

Two areas of research are relevant to the development and analysis of a data science course. While there is not much published on teaching data science or big data, we first explore this area of research. An additional area to explore is on experiential learning and boundary theory, which helps explain the value of doing real consulting projects.

Data Science Courses

There are many data science courses, and in fact, a growing number of data science programs to educate future data scientists (O'Neil, 2014). Adding to this demand is that fact that many disciplines, such as statistics, realize the importance of including data science concepts (Hardin, 2014). However, there has been little research reported on how to teach best an introduction to data science course (Mellody, 2014). Perhaps the closest topic is the description of the data science processes required for one to do data science. For example, Jagadish (2014) described a process that includes acquisition, information extraction and cleaning, data integration, modeling, analysis, interpretation, and deployment. Guo (2013) approached the problem from a slightly different perspective and provided a Data Science Workflow framework. However, even these have focused more on describing what data science was, and not how a team should do a data science project (Saltz, 2015).

There are also several books on data science, written by faculty members, which can be used as possible textbooks for an introduction to data science. These books, such as Stanton's freeware book (Stanton, 2013), typically cover the technical details about how to do data science, not what is the best pedagogy to be used to teach the course. In fact, there has been very little reported on the effectiveness of how to teach a face-to-face introduction to data science course.

Boundary Learning via Projects

There has been significant research demonstrating the value in having students do real-world projects. For example, within computer science, it has been well documented that there is much to gain by using real-world, live projects (Grisham et al., 2006; Tan & Phillips, 2005; Saltz, Serva & Heckman, 2013). Within a data science context, we define a live, real-world project as those that serve a real client, with a real problem, where students work with

real datasets (perhaps masked to maintain confidentiality). As compared to more general computer science courses, there has been little documented in terms of the viability of such projects within a data science course. In fact, the use of live real-world projects has always come with a variety of challenges including supporting the project once it is complete, providing secure and safe access to large data sets, adjusting live client expectations, and limiting the scope of a project to the context of a single term and within the learning outcomes.

As one way to understand the learning opportunity provided by our real-world client engagements, we turn to a learning theory that focuses on the learning potential created by the boundaries between and within organizations – in our situation, the boundary between school and the client's organization. Within an academic context, this type of boundary spanning can be created via the establishment of an industry-academic collaboration (Kingma, 2011). Specifically, scholars have observed that boundaries carry learning potential and that boundary crossing and boundary spanning are activities that are conductive to the construction of new knowledge (Heckman, Oserlund & Saltz, 2015).

This idea has been derived from the rapidly growing literature on knowledge sharing across boundaries in organizational and community contexts. Some researchers have examined how learning can leverage boundaries. For example, Akkerman and Bakker define a boundary as a "sociocultural difference leading to a discontinuity in action or interaction" (Akkerman & Bakker, 2011, p. 133). Such situations create opportunities for learning because existing ways to describe knowledge often are not effective in the new situation. The simultaneous coexistence of difference and similarity allows for the creation of useful artifacts that carry meaning on either side of a boundary and can be shared between them, and this artifact plays a pivotal role in learning (Heckman, Oserlund & Saltz, 2015).

Researchers have begun to explore how industry-academia boundaries can be leveraged to facilitate professional growth. Harreveld and Singh (2009) noted the importance of how to think about learning when crossing the boundary between work and school. However, as Eraut (2004) notes, the intentional facilitation of boundary spanning between school and work has largely been absent, and for that reason, the full learning potential of boundary-spanning remains unrealized (Heckman, Oserlund & Saltz, 2015).

Specifically, the way people in a work context do their work, talk about, and conceive problems differ from an academic context. At the boundary between academic work and "work work", participants do not simply convert and transfer their knowledge into a different form understandable to people in a different group. In other words, their knowledge is not merely a mental (cognitive) capacity or resource to transfer but something that must be truly internalized. An emphasis on boundary spanning calls for a practice-based approach to learning. In this approach, knowledge is not treated as a thing or possession stuffed into people's brains but an ability rooted in social practices. Only through the genuine engagement in organizational practices do participants learn how to act knowledgably in a given context (Suchman, 2007). Learning involves changes in how people are present in a broad array of arrangements, some mental, some material, and some organizational. Learning takes place not only as legitimate peripheral participation in one community but as mastery of relationships to many communities.

Applying Boundary Learning to a Data Science Course

The use of boundary spanning is particularly useful for data science, since data science is not just about how to do a certain technique, but rather, about the end-to-end process of analyzing data. Within a project-focused class, the boundaries could be between the organization (the "client") and the faculty member teaching the course. For example, as

students work on their project, they will get advice from the faculty member (on one side of the boundary) as well as requested actions from the client (on the other side of the boundary). This tension, created because the requests are sometimes conflicting, enables students to be able to internalize and think through the issues – and most importantly, integrate their learning to determine the best path forward.

Methods

This paper reports on a case study of an introduction to data science course, called "applied data science." The face-to-face course provides an overview of data science, as well as the opportunity to use a popular open-source data science tool, the "R" open-source statistical analysis and visualization system. R is reckoned by many to be the most popular choice among data analysts worldwide; having knowledge and skill in using it is considered a valuable and marketable job skill for most data scientists.

A mixed-methods action research approach was used to evaluate the perceived usefulness of the project-focused course. Mixed methods allow for a triangulation of data, helping to explain the data more fully by providing insight into complex human behavior (Cohen, Manion, & Morrison, 2011).

Data Capture and Analysis

This study used student surveys, as well as interviews, which were conducted with some of the students in order to obtain a better understanding of student experiences. In addition, to gain insight into the amount of knowledge acquired by the students, feedback on the project's usefulness was collected from the clients via semi-structured interviews. Finally, faculty observations were also used to evaluate the effectiveness of the class.

As this analysis included only one class, with a correspondingly small number of surveys and interviews, the results were analyzed by comparing answers from the interviews and surveys.

Course Overview

The course, Applied Data Science, was an introduced to the fundamentals of data science. The focus of the course was to enable the students to understand the full life cycle of how to do data science (i.e. it was not just about how to implement an advanced analytics technique). The course included applied examples of data collection, processing, transformation, management, and analysis. Students also became knowledgeable about the broader issues relating to leveraging both data analytics and data visualization to answer real-world data challenges. Most importantly, the course was different from most of the introduction to data science courses in its focus on experiential learning. In other words, even though this was each student's first data science course, the students did a "capstone like" project with a real client.

Course Participants

Since this was a pilot course (the University has offered this course for many years, but this was the first time it was being offered in a team-based project focused manner), the course was limited to 16 students, with 15 graduate students and one undergraduate student. The academic background of the students varied greatly. For example, one graduate student was in a linguistics program, with this course being the first "technical data" course the student had taken. Another example of a student with minimal background was the one

undergraduate student who was majoring in information management and had minimal preparation for a data science course. At the other end of the spectrum, one of the students had significant experience in statistics, and other one had extensive knowledge on visualization.

Course Goals – Learning Objectives

First and foremost, the primary goal of this course was to enable a broad base of students to get a foundational understanding of "how to do" data science. In other words, there were no perquisites for the course and students with any background should have been able to take and benefit from the course. This meant that the course should be of value and interest to students with statistics, analytical and/or a programming background, as well as students with minimal experience in any related data science domain. In addition to providing the foundational concepts of applied data science, there was a second goal of increasing the student's ability to understand the value of data science to industry and not-for-profit organizations. Finally, the third goal was to increase the students' desire to "do data science". Specifically, with respect to learning outcomes, at the end of the course, students were expected to be able to:

- Identify a problem and the data needed for addressing the problem
- Perform basic computational scripting using R and other optional tools. R is one of the *de facto* tools commonly used to do data science (via programming and scripting). Hence, ensuring students did R programming provided the technical depth to the course.
- Transform data through processing, linking, aggregation, summarization, and searching
- Organize and manage data at various stages of a project lifecycle
- Determine appropriate techniques for analyzing data
- Be effective in communicating the results to decision makers and other non-technical people that potentially were not experts in data science (and hence, students needed to be able to explain the results "in plain English").

Structure of class

As previously noted, the students in the course had a range of backgrounds, including some of them with a background in statistics, visualization, or programming, while others in less technical fields such as linguistics and information management. Hence, students were placed into teams, with an attempt to mix experience and expertise (so each team had a range of experience/expertise). The face-to-face class met weekly throughout the semester. Approximately half the course was structured to enable students to gain the knowledge required for the client project. In effect, students were given the necessary foundational knowledge within the first seven weeks of the course, and then for the next seven weeks, they focused on solving their client's data challenge. As such, an essential component of the class was that students worked in teams on real data problems posed by an organization. It is important to note that students continued to learn new data science concepts during the second part of the semester, but in a less structured manner. The results of the data analysis were presented to the client organizations at the end of the semester.

Two student teams were assigned to each client (two large organizations, one public sector, and one private sector). This enabled the sharing of ideas across teams as well as helping to ensure the client would get something useful at the end of the semester. The student projects

were modeled like a consulting engagement. Each client provided data for a real problem, as well as one or more domain experts to help explain the data, the problem to be solved and the business context (ex. challenges, standard metrics used). Students talked with the domain experts to understand the data and desired analysis.

Students followed a data science process methodology that has been found to be used in industry (Saltz & Shamshurin, 2015). This process was explained to the students at the beginning of the course (Figure 1). Briefly, the methodology focuses on the end-to-end process needed to do a data science project. The *preparation* phase focuses on the initial work that needs to be done to understand the domain of interest as well as collect the data that might be useful for analysis. A key focus of this phase is to enable a broad understanding, across all interested people working on the project, of the goal of the effort and the data that would be useful within the analysis of the project. Next, the *analysis* phase is an iterative process that provides for an incrementally better understanding of the data. The skills required to do the analysis typically include a combination of computer programming, statistics and the use of information visualization techniques. Finally, the last phase is *dissemination*, which focuses on enabling the results to be released to a broader community. Of particular importance was the focus on iterative analysis that was shared across student teams (to get feedback and suggestions) and, as previously mentioned, the focus on explaining the results via a final presentation and report to the client.

Findings

As previously mentioned, the evaluation of this case study was driven via feedback from the clients who provided the projects, survey results from the students who took the course and faculty observations via field notes and reflection after course completion.

Project Usefulness

Feedback from clients was one method used to understand the level of knowledge students were able to obtain from the course, in that the students had to use the knowledge in doing their projects. Hence, a key point of evaluation of student outcomes was the final project, specifically, would the students have enough knowledge such that the projects be of value to the organizations that provided the data. Based on comments from the clients for the course projects, the answer to that question was clearly yes. In fact, the organizations valued the results so much that they explicitly asked for the "code" that was used to analyze the data (so that others, within their organization, could do additional analysis).

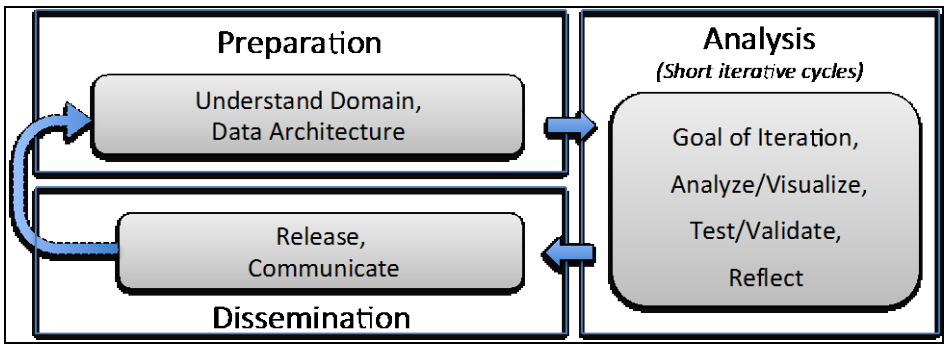

Figure 1. Data science life cycle

Student Surveys

A key focus of our quantitative analysis was the student surveys. In those surveys, 100% of the students agreed with the statement "this course stimulated critical thinking." In addition, 100% of the students also agreed that the course "provided new viewpoints" of insight. Finally, 92% of the students felt that the course "provided an intellectual challenge." These findings confirm that the students were engaged and used higher level thinking to work through their challenges.

From a qualitative perspective, students reported that they enjoyed the class, even though they thought they worked harder in this class than most other classes. A comment typical of student thinking was that "I gained a lot out of the experience and am very grateful for it." An example of a student's realization of their critical thinking was that the student noted that the professor "challenged students to think outside the textbook." This critical, higher-level thinking was due to the boundary condition created for the class project.

The fast pace of learning the material at the start of the course was a challenge for some students. However, at the same time, many students wanted to get more of the information mastered prior to the start of the project. For example, one student noted, "since the second half of the semester was spent solely on our final projects, we should have spent more time in the first half working through the book." Balancing these two contradicting needs was one of the key challenges in creating such a team-based project focused introductory course.

Faculty Observations

The faculty reported, and students confirmed, that students voluntarily spent more time on this course, as compared to a traditional/standard class. In addition, there was an increase in the number of questions students asked to the faculty – both during class and outside class, via an online discussion forum. Taken together, this shows increased motivation to learn, which translated into improved student learning.

The faculty also noticed improved "softer" skills, which are often difficult to quantify. Examples of such skills included students internalizing the importance of how to communicate the results of the analysis to technical people as well as less technical managers/decision makers. Skills related to this include the ability for students to tell a client that the desired analysis is not possible with the existing data and enhanced presentation techniques (more focus on creating a coherent presentation). In addition, students also gained "client engagement skills", which included the ability to ask better questions to understand client data and the opportunity to determine what analysis might be of value to the client.

Finally, by working with real-world clients and real-world data, students were able to work through real-world data challenges such as trying to figure out the actual meaning of data attributes, how to handle missing data, attempting to determine if a data sample is bad data or just an outlier, and techniques to analyze large datasets (100GB) versus smaller, "more standard" datasets.

Discussion

Doing real-world client projects had a positive impact on the class. It increased student motivation during the class as well as their interest in the field by enabling them to work on real-world challenges during the semester. The project enabled students to deliver value to organizations during their first-semester 'intro to data science' course.

However, the structure of the class, where the first half was focused on "learning the content" and the second half was focused on "doing the project" should be refined such that some of the content would be taught in the second half of the semester. However, this change will imply either more work for students or less polished results to the client (likely a combination of both).

Often, the challenge of teaching a real-world, project-based course is finding organizations willing to participate and share their data and knowledge. However, from the perspective of an organization providing the data, this experience was of significant value. The analysis provided was "free", but in reality, the project also required support time from each organization (such as to supply the data and explain the data fields). In addition to the insight provided by the students about the data supplied for the project, the organizations developed a student pipeline for hiring. In fact, one organization has already used that pipeline in their hiring of data scientists.

Finally, the concern about students with less technical background being able to keep up with the other students was, at least, in this case study, not realized. In fact, some of those students were often the most helpful in ensuring that the results were effectively communicated with minimal data science jargon, perhaps because it was easier for them to remember when they did not know the data science jargon.

Implications and Limitations

The results of this case study demonstrated the positive impact of doing a project-based introduction to data science course. The group nature of the projects enabled students with diverse educational backgrounds to succeed in this introduction to data science class. Students with minimal relevant background and students with skills such as advanced statistics or visualization reported a positive experience and perceived learning, which was corroborated via faculty feedback.

Hence research question 1 (could students, with diverse educational backgrounds, succeed in a project-focused introduction to data science class?) and research questions 1a (could students with a more advanced background learn a significant amount during the semester while other students, with minimal background, also learn the foundational concepts of applied data science) were both found to have a positive answer. Furthermore, research question 1b (could students learn enough during the semester to work on real-world problems and deliver real value to their clients) and the closely related research question 1c (would clients find the students projects of value) were also clearly answered positively. Specifically, there was no in need to "dumb down" the problem (or data) to a toy example and the results were deemed of value to the clients who provided the data.

While this case study did generate positive outcomes for project-focused courses, more work needs to be done to compare better the results this project-focused course with other potential pedagogies. For example, one might compare case studies or Kaggle-like competitions with this project-focused course.

Also, this course framework will continue to be refined. Specifically, there are two important next steps resulting from this case study. First, the course will be improved, based on student and faculty feedback. Specific refinements include (a) starting the project a bit earlier in the semester and (b) continuing with "data science content" later into the semester, thereby reducing, but not eliminating, the amount of self-learning that needs to occur for students to be successful in the class. Second, more sections of this course need to be offered

and monitored, to determine if the course works well with additional students and other faculty teaching the class.

Conclusion

In summary, this case study demonstrated that an introductory project-based course in data science can provide a meaningful set of skills, and that such a course might be appropriate for students across many disciplines, where data science is not the primary focus. In effect, the course provided a set of tools and knowledge that could be used within domains such as physics, engineering and the many other fields where significant data is available to be analyzed.

References

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, 81(2), 132-169.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education*. Milton Park. Abingdon, Oxon, [England]: Routledge.

Eraut, M. (2004). Transfer of knowledge between education and workplace settings. In H. Rainbird, A. Fuller & A. Munro (eds.), *Workplace Learning in Context* (pp. 53 - 73). London: Psychology Press.

Grisham, P., Krasner, H., & Perry, D., (2006). Data engineering education with real-world projects, *ACM SIGCSE Bulletin*, 38(2), 64-68.

Guo, P., (2013), Data science workflow: Overview and challenges. *Communications of the ACM.* Retrieved 30 October 2013, from http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext.

Hardin, J., Hoerl, R., Horton, N. J., & Nolan, D. (2014). Data Science in the Statistics Curricula: Preparing Students to "Think with Data". *arXiv preprint arXiv:1410.3127*. Retrieved 30 October 2015, from http://arxiv.org/ftp/arxiv/papers/1410/1410.3127.pdf.

Harreveld, B., & Singh, M. (2009). Contextualising learning at the education-training-work interface. *Education and Training*, 51(2), 92-107.

Heckman, R., Østerlund, C. S., & Saltz, J. (2015). Blended learning at the boundary: Designing a new internship. *Online Learning*, 19(3), 1-17.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., & Shahabi, C., (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7), (86-94).

Kingma, B. (2011). *Academic entrepreneurship and community engagement*. Northampton, Maine, USA: Edward Elgar Publishing.

Mellody, M. (2014). Training students to extract value from big data. Summary of a Workshop. *The National Academies Press, Washington DC*. Retrieved 30 October 2014, from http://www.nap.edu/openbook.php?record_id=18981.

O'Neil, M. (2014). As data proliferate, so do data-related graduate programs. *The Chronicle of Higher Education*. Retrieved 30 October 2014, from http://m.chronicle.com/article/As-Data-Proliferate-So-Do/144363.

Saltz, J., & Shamshurin, I. (2015). Exploring the process of doing data science via an ethnographic study of a media advertising company. In H. Ho, B. C. Ooi, M. J Zaki, X. Hu, L. Haas, V. Kumar, S. Rachuri, S. Yu, M. Hui-I Hsiao, J. Li, F. Luo, S. Pyne & K. Ogan (eds.), *IEEE International Conference on Big Data* (pp. 2098-2105). USA: IEEE.

Saltz, J. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In H. Ho, B. C. Ooi, M. J Zaki, X. Hu, L. Haas, V. Kumar, S. Rachuri, S. Yu, M. Hui-I Hsiao, J. Li, F. Luo, S. Pyne & K. Ogan (eds.), *IEEE International Conference on Big Data* (pp. 2066-2071). USA: IEEE.

Saltz, J., Serva, M., & Heckman, R. (2013). The GET immersion experience: A new model for leveraging the synergies between industry and academia. *Journal of Information Systems Education*, 24(2), 121.

Smeby, J. C., & Vågan, A. (2008). Recontextualising professional knowledge–newly qualified nurses and physicians. *Journal of Education and Work*, 21(2), 159-173.

Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge: Cambridge University Press (2nd edition).

Stanton, J. (2013). *Introduction to Data Science (Version 3)*. Retrieved 30 October 2015, from https://itunes.apple.com/us/book/introduction-to-data-science/id529088127?mt=13#.

Tan, J., & Phillips, J. (2005). Real-world project management in the academic environment. *Journal of Computing Sciences in Colleges*, 20(5), 200-213.

To cite this article: Saltz, J., & Heckman, R. (2016). Big Data science education: A case study of a Project-Focused Introductory Course. *Themes in Science and Technology Education*, 8(2), 85-94.

URL: http://earthlab.uoi.gr/theste