Pushing Data Science Education into the Real World

Daniel Turek
Berkeley Institute for Data Science
University of California, Berkeley
190 Doe Library, Berkeley, CA
anthonysuen@berkeley.edu

Anthony Suen
Berkeley Institute for Data Science
University of California, Berkeley
190 Doe Library, Berkeley, CA
dturek@berkeley.edu

Dav Clark
Berkeley Institute for Data Science
University of California, Berkeley
190 Doe Library, Berkeley, CA
davclark@berkeley.edu

ABSTRACT

The discipline of data science has been viewed as an convergence of high-power computing, data visualization and analysis, and data-driven application domains over the past decade. Prominent research institutions and private sector industry have been quick to embrace the importance of data science, but the foundations for effective tertiary-level data science education are conspicuously absent. This is nothing new, however, as the university has a well-established tradition of developing its educational mission hand in hand with the development of novel methods for human understanding [6]. Thus it is natural that universities "figure out" data science hand in hand with the development of needed pedagogy. We consider the development of data science education with respect to recent trends in interdisciplinary and experiential education, along with agile and design thinking methodologies to understand how they could apply to data science educational programs. This historical perspective motivates us to consider what factors are necessary to drive effective data science education, which range from a complete end-to-end workflow, technological tools for development and team communications, and appropriate motivation and incentives. The first iteration of the *Berkeley Institute for Data Science (BIDS) Collaborative* started in the University of California, Berkeley in the Spring of 2015 is used as a case study. From this we draw lessons learned and form a hypothesis regarding the necessary ingredients for effective data science education at the tertiary level ? a topic that is presently understudied. This hypothesis will be tested and revised in subsequent iterations of the BIDS Collaborative as we continue our study of effective data science education, research, and social impact.

1. INTRODUCTION

The rapid advances in computational power and the ongoing "big data" craze, the discipline of data science has exploded onto the academic and business landscape. Master's programs in data science are now being offered at premier research institutions such as Stanford University and Columbia University, and centers for data science have recently opened their doors at the University of California, Berkeley, the University of Washington, and New York University. Led by the success of tech giants such as Google, Amazon and Facebook, the increasing availability of data is transforming industries ranging from medicine to media. The industrial sector is keeping pace by creating and actively recruiting for positions in data science. The profession of data scientist was even described by the Harvard Business Review as "the sexiest job of the 21st century" [15].

Despite this inundation of the term "data science," we still struggle to define what data science is, or to realise any boundaries as to what data science encompasses [7, 12, 18]. A common Venn diagram places data science squarely at the intersection of computer science, mathematical statistics, and scientific application domains. This perhaps most accurately depicts that data science is nebulous by nature, having ties to all areas of quantitative scientific research or computational data analysis, but falls short of providing an understanding of how this new scientific discipline will eventually settle into the scientific ecosystem. Fortunately, our aim is not to pin down the nature of data science itself, but instead to examine the practicalities and realities of data science education at the tertiary level.

There exists substantial literature regarding best practices and modern approaches to tertiary education. This has been a subject of interest since the first modern Universities appeared in Europe [21, 16]. Since then, the approach to higher education has evolved immensely, due to advances in technology, and also society's attitude towards higher education. Perhaps the single-most transformative influence on higher education has been the so-called digital revolution of the past decades, which has had a profound impact on the content and style of tertiary education [20, 4, 2, 25, 1].

Some research suggests that traditional approaches to tertiary education may only result in superficial learning, rather than a deep understanding of subject material [5]. Thus, the study of education itself is an area of prime interest. Many approaches have been suggested and studied over the past decades, in attempts to improve education at the tertiary level. [24] promotes the practice of peer-tutoring, while others have more recently endorsed "flipped classrooms" in which learning becomes more self-directed (as opposed to instructor-directed) and classrooms become a place for practice instead of lecture [9, 8]. [14] suggests a "mutiscience" approach to multidisciplinary science education, in which the diversity scientific disciplines is recognized and incorpo-

Data for Good Exchange 2015, New York, NY

rated into the educational system. Project-based learning has been promoted at the institutional level for many years [11, 23].

We re-focus our attention to the fledgling field of data science, but now in the context of education. In light of the academic and industry spotlight on data science, experiences and best practices for data science education should be a prime focus of research, just as it is for tertiary education in general. However, owing to its relatively recent mainstream debut, there is a noteworthy absence of scientific research or published literature on the subject of data science education. This fact motivates our present analysis of the history, current trends, and future prospects of data science education.

We aim to begin filling this void by providing a tangible case study of data science education, which was undertaken at the University of California, Berkeley, under the BIDS Collaborative. We consider the successes and failures of the first cohort to pass through the Collaborative, and the pain points which were encountered by the students and mentors, alike. We make practical recommendations for educational approaches to data science curricula, and formulate a hypothesis regarding the "best practices" of tertiary data science education. Study of our hypothesis will require subsequent experiential testing, which will be the subject of on-going and future research.

2. PARADIGMS OF DATA SCIENCE EDUCATION

To set the context for data science training and research methods, we briefly review several education and research paradigms that have become prominent in the past few decades. These include **interdisciplinary research**, **experiential learning**, and the **lean and agile practices** models of learning and research. We will also look at the Data Science for Social Good model (`http://dssg.io`) for training 40-50 graduate students in data science each summer since 2013. Finally, we will break down from these paradigms the theories we hope to practice in the BIDS Collaborative.

2.1 Interdisciplinary Research

Interdisciplinary or multidisciplinary research is "a format for conversation and connections that will lead to new knowledge" [19]. The word has been fashionable academic buzzword for decades now, but it has looking and solving issues from multiple angles has failed to fundamentally scale beyond the confines of certain research groups to change the way research is carried out. There are major obstacles like cultural, organizational, technical barriers that prevent such learning and research environments [3].

First, interdisciplinary training is not generally taught, whether you are an undergraduate or graduate student. Interdisciplinary is not a core metric for industry or academic career paths in order to be exposed to the values. This lack of training has to due in part to organizational structure with many departments providing little to no incentives for interdisciplinary collaborations. By making job prospects for those who have multidisciplinary focus difficult, it creates a vicious cycle that reinforces single disciplinary specialist work.

Siloed organizational structures also created silos around sets of tools – different software and methods are used to achieve similar goals via widely divergent means, potentially obscuring the fact that disparate groups are in fact grappling with the same underlying problems. Social scientists, physical sciences, and engineering use very different tools to tackle data. They use these different tools to run models that often have similar predictive goals. For example, Stata [22] might be used by an economist, Matlab[10] by the engineer, and R by the statistician. This divergences and specializations in tools creates an ever widening gap between major disciplines. All of this makes interdisciplinary collaboration among a diverse team rare to come by.

Observing these barriers to multidisciplinary data science, our mission was three fold. First, we implemented a framework for interdisciplinary learning beyond traditional academic lecture and coursework structures. Second, we cultivated an environment independent of the "rules" or established incentive structure of traditional academic departments. Our third goal was to show interdisciplinarity was possible with limited resources and incentives, and that data science tools can become unified.

This reinforces a need for leaders within multidisciplinary teams which might be easier to achieve in a graduate student and undergraduate student teams than among faculty due to lower barriers in terms of technology, different incentivizes, and greater openness to doing things in a different way. We hypothesized that multidisciplinary collaboration would be easier when the stakes are smaller than traditional academics and learning is driven within student peer groups.

2.2 Experiential Learning (Flipped Classroom Model)

The second key component is learning real projects with real clients. Research from Stanford [17] has validated that the flipped class experiential learning as a stronger learning strategy compared to homework and lectures. Pushing students to do work without first preparation seems to be effective in accelerating knowledge. This meant we were scraping the lecture and testing format entirely. Experiential learning also emphasis project management since classroom projects are have clear deadlines scheduled by the professor [13]. Given the diverse nature of each project, our limited staffing resources, and the need to manage relationships with clients, it became critical that students figure become project managers to push the team forward to success.

2.3 Lean and Agile Practices

Another method we wanted to adopt for our projects were Lean and Agile. Lean comes from lean manufacturing which seeks to deliver speed, quality, and alignment. Analytics in organizations is usually done using a waterfall method which require long development cycles. Lean has a great way to look at help accelerate operations, m operates as a whole.

3. LESSONS LEARNED

The first iteration of the BIDS Collaborative was an experimental undertaking, in which both students and facilitators were learning through the duration of the program. Subsequent to this, many important lessons for effective data science education became apparent. These lessons have been documented, and will serve as guiding principles for subsequent iterations of the Collaborative. We now discuss several of these "lessons learned," which will lead to our hypothesis for achieving effective data science education.

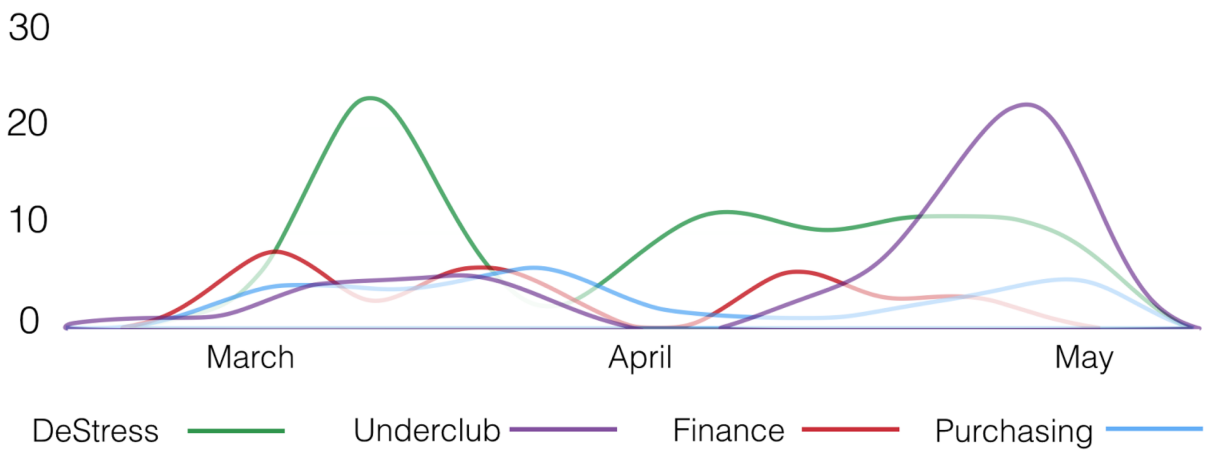

Figure 1: CAPTION

Using a model where students decide among various predetermined capstone research projects, it is important to present options representing a broad variety of disciplines. For example, having projects relating to physical sciences, social sciences, technology, health, environment, commerce, among many other possibilities. By presenting this diverse range of project options, we were able to leverage students' innate interests in particular areas of study; thus, students did not feel shoehorned into a research areas of little or no interest to them.

It was critical that the logistics of each project were fully in order in advance of being presented to students. Projects must be well-posed, and have a well-defined goal which students could work towards. This also included having a responsive and interested point of contact, representing the client or the underlying organization, who could answer questions as necessary. And most importantly, the relevant data must be available in advance, such that students could get a sense of what the project would entail, and could begin work immediately. This consideration encompasses any legal releases, non-disclosure agreements, etc, pertaining to data access. Fundamentally, for projects to begin with an enthusiastic start, students must have adequate access to the necessary data and a contact point representing the underlying organization.

Project and team components needed to be organized in the appropriate order, as each step necessarily followed the previous. This process began with identifying clients with data science research projects of the appropriate scale. The relevant data must be already available, as this was critical to generate student interest. Next, we were able to identify student team leads, who had the motivation and skills to lead a research team, and interact with the client. The team leads played a critical role interacting between these two parties, and having this team structure in place from the beginning was important for the overall organization of each team. Finally, with all these parts in place, we were able to assemble teams for each project based on students' interests and skill sets. Only by following this order of client, data, team lead, team members, did each step flow smoothly from those previous.

Along these same lines, we observed that regular interaction between the client and team lead was necessary. The team lead served as the bridge, for processing and presenting the client needs to the team of student researchers. We learned it would not be possible the facilitators of the BIDS Collaborative to fill this role for several reasons. First, it was impractical to micro-manage each project at this level. And second, requiring this direct interaction between clients and team leads formed the reality of students organizing and accomplishing a real-world project. This effectively maintained a sense of responsibility and self accomplishment for team leaders and team members, alike.

As projects got underway, we observed that creating well-defined intermediate goals was helpful to maintain forward momentum within each team. These took for the form of "milestones" which would otherwise exist in a data science research workflow (e.g., data cleaning, or preliminary visualization analyses), but formally assigning dates for these tasks ensured that each team was progressing forward. In addition, this helped the Collaborative facilitators passively monitor the progress of individual teams, and provide additional help when necessary. Larger milestones were also created including mid-semester presentations, and a final capstone event where projects and results were presented to the clients. The existence of these formal milestones kept all research teams on forward-moving (and similar) timelines.

Finally, we noticed the importance of introducing (and promoting the use of) team collaborative tools from the very beginning. The most successful project teams immediately adopted GitHub for all project code, and also Slack for team communications. It appeared that the sooner team members adopted these tools into their research workflows, the sooner meaningful progress began. We believe it is critical to introduce these tools to students not familiar with them immediately, and promote, if not mandate their usage for teamwork and communications.

4. CLOSING THOUGHTS

A number of open questions remain from the first iteration of the Collaborative. We first discuss a few of these questions, before presenting our conclusions and hypothesis

for future data science education.

4.1 Open Questions

We explored possibilities for offering academic credit as motivation for students, but no students decided to pursue this option. The reason behind this remains uncertain, though it could be due to the lack of structure and the additional hurdles in enrollment. We believe one pathway would be creating a framework that plugs into an existing or a new project-oriented course. Even so, however, the option of receiving course credit did not appear to be a strong motivation for students.

Some fraction of students, however motivated, were not prepared with the technical background for jumping into a data science research project. The usefulness of periodic training sessions, and how these could be organized or delivered, remains an open question. Who would teach such sessions and exactly what material would be most beneficial for students is also unclear. This training could possibly link with existing workshop or training programs on campus, so certain students have the opportunity to get up to speed on the relevant tools.

Generally, the best approach to organizing, managing, and motivating teams remains unclear to the facilitators. One approach would be to micromanage to some degree, and organize regular weekly team meetings. However, this level of management is very time-consuming, and not always effective or appreciated by student groups. In addition, exactly how to motivate a strong commitment from team members is a difficult question. Fundamentally, we would like to rely on students' desires for real-world data science experience and education, but this will not always suffice. How all students can be effectively motivated remains open for discussion.

4.2 Conclusions

The BIDS Collaborative was a small educational experiment done in BIDS with a bare bones staffing and limited planning. We did not have the resources of the Data Science for Social Good but we were successful in motivating a group of students to complete client provided real-world data science problems over the course of an academic semester. The design and overall success of the BIDS Collaborative program shows promise in terms of scalability among the wider university.

Stepping back, we have uncovered certain best practices in terms of creating multidisciplinary teams to solve real life challenges.

[some statement about stepping back, and how our conclusions and lessons can be applied equally well to well to non-data-science projects, etc]

Hypothesis: [NEEDS TO BE WRITTEN] We hypothesize that?. Effective multidisciplinary data science education must deal with the complexity that is fundamental to novel data sources and / or methods. This requires a unified academic curriculum, augmented with workshops and consulting services that can help teams develop compatible approaches for working together, Most importantly, we have reported on the apparent value of a capstone collaborative project that provides an opportunity for a multidisciplinary groups to integrate their skills.

5. REFERENCES

[1] Y. Baek, J. Jung, and B. Kim. What makes teachers use technology in the classroom? exploring the factors affecting facilitation of technology with a korean sample. 50(1):224–234.

[2] W. Baker. Technology in the classroom: From theory to practice. 32(5).

[3] L. Eisenberg, T. C. Pellmar, and others. *Bridging disciplines in the brain, behavioral, and clinical sciences*. National Academies Press.

[4] D. P. Ely. Technology is the answer! but what was the question?.

[5] N. J. Entwistle. *The impact of teaching on learning outcomes in higher education: a literature review*. Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom, Universities' Staff Development Unit.

[6] M. Feingold. Tradition versus Novelty: Universities and Scientific Societies in the Early Modern Period. *Revolution and Continuity: Essays in the History and Philosophy of Early Modern Science*, pages 45–62, 1991.

[7] C. Hayashi. *What is Data Science ? Fundamental Concepts and a Heuristic Example*, pages 40–51. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan.

[8] C. F. Herreid and N. A. Schiller. Case studies and the flipped classroom. 42(5):62–66.

[9] M. Horn. The transformational potential of flipped classrooms. 13(3):78–79.

[10] M. W. Incorporation. MATLAB user manual version 7.1 r14.

[11] J. S. Krajcik and P. C. Blumenfeld. *Project-based learning*.

[12] M. Loukides. *What is data science?* O'Reilly Media, Inc.

[13] H. N. Mok. Teaching tip: The flipped classroom. 25(1):7.

[14] M. Ogawa. Science education in a multiscience perspective. 79(5):583–593.

[15] T. H. D. J. Patil. Data scientist: The sexiest job of the 21st century.

[16] O. Pedersen. *The First Universities: Studium Generale and the Origins of University Education in Europe*. Cambridge University Press.

[17] D. Plotnikoff. Classes should do hands-on exercises before reading and video, stanford researchers say.

[18] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. 1(1):51–59.

[19] A. F. Repko. *Interdisciplinary research: Process and theory*. Sage.

[20] N. Roberts and A. Ferris. Integrating technology into a teacher education program. 2(3):215–25.

[21] W. Rudy. *The universities of Europe, 1100-1914: a history*. Fairleigh Dickinson Univ Pr.

[22] Stata Corporation. *Stata Statistical Software Release 9*. Stata Press Publication.

[23] J. W. Thomas. A review of research on project-based learning.

[24] K. J. Topping. The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. 32(3):321–345.

[25] E. Wood, J. Mueller, T. Willoughby, J. Specht, and T. Deyoung. TeachersâĂŹ perceptions: barriers and supports to using technology in the classroom. 5(2):183–206.