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 embraced 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 [9]. 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, experiential, and agile 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 leading 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" [18].

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 [10, 15, 21]. 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 [24, 19]. 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 [23, 7, 3, 28, 2].

Some research suggests that traditional approaches to tertiary education may only result in superficial learning, rather than a deep understanding of subject material [8]. 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. [27] 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 [12, 11]. [17] suggests a "mutiscience" approach to multidisciplinary science education, in which the diversity scientific disciplines is recognized and incorporated into the educational system. Project-based learning

Data for Good Exchange 2015, New York, NY

has been promoted at the institutional level for many years [14, 26].

We re-focused our attention to the education of data science. 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 an absence of scientific research or published literature on 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 BIDS 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 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** and the **experiential learning** educational model. We also consider the Data Science for Social Good model (DSSG; `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" [22]. 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 [6].

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 [25] might be used by an economist, Matlab [13] by the engineer, and R [1] 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)

The second key component was to make the projects truly experientially focused, meaning real projects with real clients. Under this model, students are no longer playing passive role in the educational process [4].

Research from Stanford [20] has validated that the flipped class experiential learning is a stronger learning strategy compared to the traditional approach of combining lectures and homework. Pushing students to do work without advance preparation seems to be effective in accelerating knowledge. This experiential approach means removing the accepted formate of lectures and tests entirely.

Experiential learning also emphasis project management, since classroom projects are have clear deadlines scheduled by the professor [16]. 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.

The experiential learning approach was also utilized because it was more realistic representation of how data science is done outside of classroom assignments. Many of the best graduate level courses like Lean Launchpad.

There are elements of cooperative learning in the too.

3. CASE STUDY: BIDS COLLABORATIVE

As detailed above, the central organizing principle of the collaborative was to organize teams of students around data science projects. In standard data science education, a common concern is the selection of domain-appropriate materials that will be of interest given a student's background. Thus, one of our primary concerns was to ensure projects that would be of interest to a wide variety of students, and we therefore collected a large, diverse set spanning 16 projects.

Based on lessons learned at the DSSG program in Chicago, we requested that project clients have data ready in hand that they were able to share with student teams. We also attempted to ensure that projects were clearly framed as an answerable question (though as we will discuss below, most projects were more exploratory, and this was not a bad thing). It seems clear in hindsight, however, that student enthusiasm for a project was driven almost entirely by the presence of the "client" to give a pitch to students during one of two informational sessions at the beginning of the semester. Given limited time and resources for managing the collaborative, most client proposals were not vetted for data availability. Ultimately, four projects received enough interest, along with a project from a team that had already formed prior to their participation in the collaborative (these projects are detailed below).

3.1 Project Selection and Team Formation

The BIDS Collaborative ran five projects during the course of the semester, with clients from industry, academia, non-profit, and government. Details of specific projects are relayed in separate sections below. Each project took a somewhat idiosyncratic course, though there were clear commonalities, both in terms of challenges and solutions. Teams consisting of four members from various disciplines were suggested. A particular concern was the need to balance teams so not everyone was technical.

Team formation was somewhat chaotic and arguably the most difficult part of the process from the facilitator perspective. Participants were largely attracted via two "mixers" at the beginning of the Spring 2015 semester, which consisted of a brief motivational presentation after which we asked students to meet with one another and form teams around one of the available projects. Students were given a large number of choices, and were encouraged to organize themselves via a google spreadsheet. While self-organization might work more efficiently via a system that could enforce a set of rules and policies, allowing all students to collaboratively edit a spreadsheet created numerous problems. The most dramatic difficulty occurred when a large number of student responses were inadvertently deleted from the spreadsheet. Ultimately, our efforts to create a "self-service" approach required facilitators in the collaborative (the authors of this paper) to engage in a very time-intensive process, organizing teams via extensive conversations in person and via email. While attendance at the mixers was quite strong, the collaborative ultimately retained thirteen individuals from the initial cohort, with an additional three joining a few weeks into the semester. One individual dropped off mid-semester (this individual was actually staff at Lawrence Berkeley National Labs, and not a student), though participation was otherwise stable after the first few weeks of the spring semester.

An additional challenge at the beginning of the Spring 2015 semester was the lack of leadership in the various teams. Teams could best be described as loose assemblies of individuals working relatively independently on related topics, and progress was slow. Recognizing this lack of cohesion, we suggested that each team choose a lead – a process that generally consisted of one person volunteering to take on this role. At least one individual expressed reluctance to engage in a management role, as their interest was largely in hands-on experience *doing data science*. Moreover, after this transition, it was often difficult for these reluctant managers to lead their teams. At this stage, team leads served a gatekeeper function in allowing students into their team. For these students, the authority of the team lead appeared to be more established. To be clear, "authority" here was very gently exercised, and primarily consisted of working with facilitators to be clear about intended work, and progress achieved.

For the upcoming Fall 2015 session, we intend to utilize these lesson and focus on identifying clear team leads for a set of projects. At this point, facilitators can work with team leads to select remaining team members. This will serve to simplify and distribute the process of team formation, while also clearly establishing a leadership role for the team lead from the beginning.

3.2 Projects

While both the facilitators and participating teams were divided on the issue at first, all teams switched to development on GitHub in a matter of weeks. Teams formed in mid-February, and all teams were committing to GitHub by March (Figure 1). While a training on collaborative software development with GitHub was provided in the context of "The Hacker Within" meeting in BIDS, few members of the collaborative were able to attend. The differing skill-levels of the participants further problematized training in these skills. As such, the start was a bit rocky, and again consisted primarily of smaller-scale coaching from facilitators and mentors. Relatedly, the usage of BIDS space was ad hoc, with some members of the BIDS community finding this usage disruptive. This was addressed by identifying a single weekday where collaborative members were particularly encouraged to attend and "take over" the space, and to be particularly conscientious (for example, using private breakout rooms) outside of this time.

In our upcoming session, we have thus adopted a clear plan that we will establish immediately at the beginning of the semester. Students will be encouraged to come to a practicum on a weekday that will include an initial half-hour of orientation to technical tooling, including project management and documentation.

Strategic Sourcing

Director of the University of California's strategic sourcing unit had a pre-existing connection with the Berkeley D-Lab (`http://dlab.berkeley.edu/`), primarily via a previous analyst's use of D-Lab python training and consulting. This work led to a conference talk at SciPy 2014 discussing how straightforward scientific python scripts were able to accelerate previously spreadsheet-based analyses from taking approximately a week to a matter of minutes. A Collaborative facilitator assisted strategic sourcing in this work, and was therefore well aware of the opportunities available to save the university system time and money. From the perspective of an academic institution, this project illustrates an exciting double-win, with potential benefits to administration *and* student training.

We learned from this attempt that it is incredibly hard for a team of graduate students to obtain actionable insights while working only a few hours per week for a semester. Useful progress was made, however, in identifying basic workflows and determining which questions appear to be answerable. In particular, clustering techniques showed promise in identifying non-obvious partnerships where collective purchasing might provide savings.

Another key insight that five hours was insufficient to maintain consistent progress, as much time is spent each

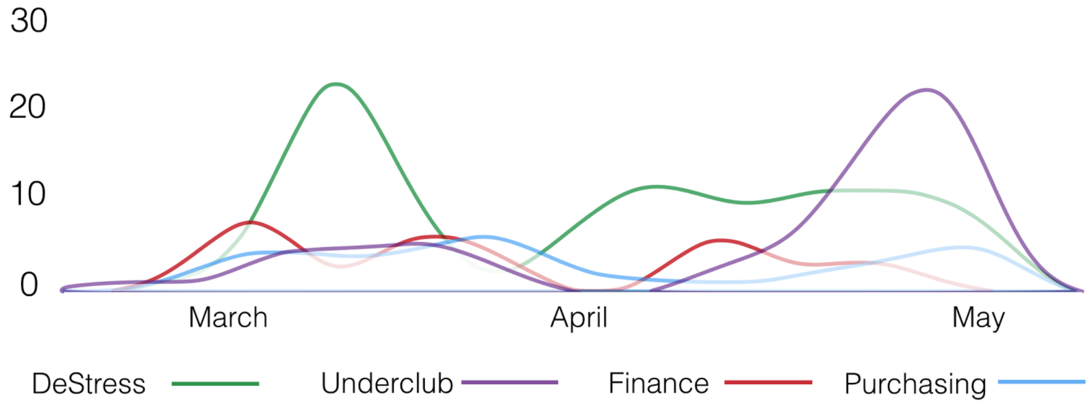

Figure 1: GitHub commit histories for each BIDS Collaborative team between March and May 2015.

week keep track of what happened. Ten hours seems to be required at a minimum.

Text Mining for Stress

This was one of two projects that was actually driven by faculty involvement. What differentiated these projects from standard faculty-driven research was the inclusive call for participation, and the engagement in a collaborative open-source development framework. In this case, we pursued a project using Prof. John Canny's BIDMach system ([5]; `http://bid2.berkeley.edu/bid-data-project/`) – a performant GPU-accelerated system for machine learning in the Scala language. A domain focus on determining stress and major life events using large-scale machine learning was chosen by one of Prof. Canny's graduate students, Pablo Paredes.

The project was enabled in part by providing a commodity workstation (with GPU) that was already available in the D-Lab, with the addition 1TB of hard disk storage. Thus, while this project was pushing the limits of academic machine learning, the resources for this project would be readily accessible to modestly funded labs. This project maintains robust activity, and the participants are working towards the publication of a handful of papers this summer.

Analyzing Financial Market Data with Spark

This project was the second of two that was driven by a faculty member, in this case, Prof. Justin McCrary, faculty director of the D-Lab. Prof. McCrary has been working to develop efficient workflows to take advantage of the UC Berkeley campus compute cluster, Savio. To this end, we worked with the Berkeley Research Computing team that manages the cluster to enable a modern Spark-based workflow. Unlike other projects, this initiative had large startup costs, including software installation challenges, and integrating with the traditional HPC Scheduler (Slurm). As such, much of the semester was spent getting to proof-of-concept workflows using spark to analyze a subset of the data. Despite modest progress on the domain questions, this project was likely of the highest value to campus, as it improved the ability to take advantage of the impressive parallel capabilities that are now provided to all senior campus researchers as a "birthright" – particularly for those users who may need something other than a traditional HPC workflow.

Underclub

While our intention was to primarily recruit clients from outside the university, Underclub was the only client that approximated this intent. Indeed, even they had a pre-existing affiliation with the university via the Haas school of business (the founder was a Haas graduate, and Underclub remains connected to the school). Much as with the strategic sourcing project above, participants struggled to achieve actionable insights, though again promising directions were established. It is clear that more guidance is needed to efficiently connect analysis with potential business-relevant actions.

4. 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.

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.

5. 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.

5.1 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.

5.2 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 DSSG program, 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.

In addition, the first iteration of the Collabora brought a variety of practical considerations for effective data science education to light. Perhaps foremost is the importance of having real-world projects, which represent interested clients and are backed by accessible data, ready at hand. Forethought about the relationships and communication lines between clients, projects team leads, and team members was also proved surprisingly important for the smooth operation of the research teams. Finally, the early introduction of the appropriate technical tools (and training for these tools, if necessary) was also necessary for effective team dynamics.

We hypothesize that effective multidisciplinary data science education must address the complexities which are fundamental to both technical research and human team dynamics. This requires imposing a structured hierarchcy for client-team dynamics, augmented with workshops and consulting services to provide the resources necessary for productive research, and most importantly, we note the value of a capstone collaborative project which provides genuinely experiential learning, which we feel is most beneficial for effective data science education.

Stepping back, we have uncovered certain best practices in terms of creating multidisciplinary teams to solve real life challenges. Although these lessons arose from a semester-long program for experiential data science education, they can be applied equally well to experiential learning in other disciplines. We believe our general conclusions may benefit project-based educational programs throughout university systems as a whole.

6. REFERENCES

[1] R: A language and environment for statistical computing.

[2] 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.

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

[4] C. Beard. *The experiential learning toolkit: Blending practice with concepts*. Kogan Page Publishers.

[5] J. Canny and H. Zhao. Bidmach: Large-scale learning with zero memory allocation. In *BigLearning, NIPS Workshop*.

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

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

[8] 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.

[9] 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.

[10] C. Hayashi. What is data science ? fundamental concepts and a heuristic example. In P. E. C. Hayashi, P. K. Yajima, P. H.-H. Bock, P. N. Ohsumi, P. Y. Tanaka, and P. Y. Baba, editors, *Data Science, Classification, and Related Methods*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 40–51. Springer Japan.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[28] 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.