2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #33834

Using Data Science to Create an Impact on a City Life and to Encourage Students from Underserved Communities to Get into STEM

Prof. Elena Filatova, City University of New York

Assistant Professor at CUNY, New York City College of Technology, Department of Computer Systems Technology. Director of the Bachelor of Science in Data Science program.

Dr. Deborah Hecht, Center for Advanced Study in Education

As Director of the Center for Advanced Study in Education, at the CUNY Graduate Center I am involved in a wide range of educational evaluations of funded and local projects. I also mentor graduate students interested in careers in evaluation.

Using Data Science to Create an Impact on a City Life and to Encourage Students from Underserved Communities to Get into STEM.

Abstract:

In this paper, we introduce a novel methodology for teaching Data Science courses at New York City College of Technology, CUNY (CityTech). This methodology has been designed to engage our diverse student body. CityTech is an urban, commuter, HSI (Hispanic Serving Institution) school with 34% Hispanic and 29% Black students. 61% of our students come from households with an income of less than $30,000. Thus, many students in our college come from the New York City communities that are underrepresented in the STEM fields and at the decision-making positions in the government (at the city level, state level, country level). However, our methodology flips the situation so that our students' living situation does not hold them back, but on the contrary, gives them an edge in their education. Our methodology uses case-based learning and diversity among our students who come from different city communities (location-wise, ethnicity-wise, income-wise) to enrich and drive the education experiences. We demonstrate that this combination can be the basis of a powerful teaching method that delivers STEM material and engaging students in the learning process.

To evaluate our novel methodology, we ran a pilot study within one introductory class designed specifically for the BS in Data Science program. In this pilot study, we taught data analysis utilizing data sets collected by the New York City agencies[1]. Our findings demonstrate that using real-life data sets encourages students to compare the results learned from data about their communities and their everyday experiences. We believe that using such a teaching approach can be a great start for igniting the interest in the field as well as in society-aware aspects of data analysis.

1. Introduction

Graduates with knowledge in the field of Data Science are currently in great demand in industry and research. This demand is much higher than the number of graduates with adequate training. A data scientist is expected to have training in mathematics, computing, and the domain for which the data should be analyzed. In Spring 2020, we launched a new baccalaureate-level program in Data Science at CUNY CityTech[2]. This program follows an important and current trend of preparing experts in Data Science.

The field of Data Science is relatively new. Thus, all the existing baccalaureate-level Data Science programs have only been launched recently, or only about to be launched. For example, New York University launched BS in Data in Fall 2019[3], New Jersey Institute of Technology plans to launch BS in Data Science in Fall 2021[4]. Moreover, since the field of Data Science is inherently interdisciplinary, the Data Science programs originate from various departments,

[1] https://opendata.cityofnewyork.us/
[2] http://www.citytech.cuny.edu/computer-systems/data-science-bs.aspx
[3] https://www.nyu.edu/about/news-publications/news/2019/september/nyu-launches-data-science-major.html
[4] https://news.njit.edu/njit-offer-new-undergraduate-degree-data-science-starting-fall-2021

including Business Information Systems, Computer Science, Mathematics, Statistics, Sociology, etc. Thus, given that the baccalaureate-level Data Science programs are new and run by departments with different backgrounds, the Data Science courses are being taught following a variety of pedagogical and methodological standards and approaches.

According to the Brookings "Middle-Class Mobility" report[5], CUNY CityTech is among the leaders "in lifting low-income students into the middle class." Our college is a commuter college in a city with a diverse population: most of our students come to classes from various areas of a diverse, big city. Thus, while thinking about the methodologies and best pedagogical practices for teaching Data Science we had two goals in mind:

1. Design the teaching methodology which is not only effective in teaching the Data Science skills but is also accessible and attractive to the students with a variety of education and life backgrounds;
2. Provide the Data Science students with in-demand skills and hands-on experience using real-life data.

We believe that both of these goals are equally important in providing the students from underserved communities with the skills that would allow them to find good employment options after graduation.

Our BS in Data Science program has been designed following the best practices identified after attending an NSF-sponsored workshop (De Veaux, et al., 2017). In our teaching methodology, we combine the Case-Based Learning approach (CBL), Project-Based Learning approach (PBL), and the life experiences of our students. To create cases for the class discussions and projects for the students to work on we use the NYC Open Data platform. On this platform, the city collects various data sets about the city life and the work of different city departments.

In this paper, we describe our novel teaching methodology of teaching Data Science and report the results of the pilot study on how students relate to this teaching methodology.

2. Using Case-Based and Project-Based Learning for Teaching Data Science.

In Fall 2020 we taught for the first time a course on information and data management. This course is the first in the series of the courses designed specifically for our BS in Data Science program. We used feedback from this course to learn about the effectiveness of our CBL approach towards teaching Data Science.

According to the Case-based Learning page of the Yale Poorvu Center for Teaching and Learning[6], retrieved July 24th 2020, "Case-based learning (CBL) is an established approach used across disciplines where students apply their knowledge to real-world scenarios, promoting higher levels of cognition." We use this resource from Yale University as well as resources from the Educational Technology Research and Teaching Unit at the University of Geneva[7] to prepare our Data Science classes.

[5] https://www.brookings.edu/research/opportunity-engines-middle-class-mobility-in-higher-education/
[6] https://poorvucenter.yale.edu/faculty-resources/strategies-teaching/case-based-learning
[7] http://edutechwiki.unige.ch/en/Case-based_learning

CBL has a long history as a successful pedagogy technique in medical, law, and business schools. Recently, the CBL method has become an increasingly popular trend in teaching sciences [1]. This method involves guided inquiry and is grounded in constructivism whereby students form new meanings by interacting with their knowledge and the environment [2]. Examples of teaching using CBL for different disciplines include:

- Organic Chemistry and Your Cellphone: Organic Light-Emitting Diodes[8];
- Light on Physics: F-Number and Exposure Time[9];
- Public Health: Credible Voice: WHO-Beijing and the SARS Crisis[10];
- Engineering (Mustoe & Croft, 1999).

In STEM education CBL is often combined with Project-Based or Problem Based Learning (PBL)[11]. According to Jane David, of ASCD's Educational Leadership journal [3]: "The core idea of project-based learning is that real-world problems capture students' interest and provoke serious thinking as the students acquire and apply new knowledge in a problem-solving context. Advocates assert that project-based learning helps prepare students for the thinking and collaboration skills required in the workplace."

We believe that combining case-based and problem-based learning in Data Science can be a powerful approach for both delivering STEM material and engaging students in the learning process. Moreover, we expect that engaging our students in the learning process will our Data Science program with the task of closing the education gap for the underrepresented groups of students from diverse urban communities.

3. Pilot Study

For our pilot study, we chose the first class in the sequence of classes designed specifically for the Data Science program. This class covers topics related to the digital infrastructure, acquisition, organization, management, and curation of data, and uses IPython notebooks as the programming environment. The students' final project includes an application of the technical topics discussed in class.

Throughout the semester, the course material was introduced using real-life examples. The primary source of data sets used in our class was the NYC Open Data platform.
The data sets from this platform were used to introduce Python dictionaries, nested structures, APIs, libraries (Numpy, Pandas, Requests, etc.), Python arrays, data series, data frames, etc.

The NYC Open Data platform is a central place where different departments and agencies that operate in New York post their data sets. We believe that using such real-life data sets encourages our students to compare the results of what they learn about their communities from the data and their everyday experiences. Such comparisons can be a great start for igniting interest in the field.

---

[8] https://sciencecases.lib.buffalo.edu/cs/files/oled.pdf
[9] https://sciencecases.lib.buffalo.edu/cs/files/optics.pdf
[10] https://casestudies.ccnmtl.columbia.edu/case/crediblevoice/
[11] https://serc.carleton.edu/introgeo/icbl/resource.html

3.1 Case-Based Learning Using Real-Life Data Sets

The case studies are used to describe the topics of regular expressions, APIs, Python data frames, etc. The data sets are downloaded from the NYC Open Data where the city stores the data sets about the city life collected by different city departments and agencies.

For example, the data set about the city Restaurant Inspections contains "*every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias in an active status*" and is updated on daily basis. The case study that uses the city Restaurant Inspections data set is designed to introduce or review and practice the topics of APIs, Python Pandas and Numpy libraries; teach the students to upload data using the NYC Open Data API; demonstrate how regular expressions and Python can be used for data analysis. For example, during this case analysis, the students can find all the inspections where the VIOLATION DESCRIPTION field has the word *mice* or all the restaurants that have the word *pizza* in their names. Using this data set we can introduce the operations with arrays, vectors, and dataframes by counting the number of restaurants in a particular ZIP code. During the class sessions that deal with case studies, students are encouraged to formulate questions answers to which can be obtained given the data sets. For example, the students can learn what code violations are listed for their favorite eating establishments; how often inspections happen in particular ZIP codes; find out about the health code violations in the cafeteria in CityTech, and even learn (to the surprise of the students and the instructor alike) about a new pizza place in CityTech whose opening was delayed due to the COVID-19 pandemic.

Given the students' interests and background, it is possible to choose a dataset other than the Restaurant Inspections data set that allows to demonstrate the topics that are part of the class learning objectives.

3.2 Project-Based Learning Driven by Students

After the class material was delivered using traditional lectures and case-based learning, the students were asked to choose a data set that is of interest to them. This data set became the central point for the final class project. Within this project, the students were asked to apply all the skills they learned in class and to answer questions (using data) that they find interesting.

The students chose a variety of data sets: several years of data about the new cases of HIV/AIDS identified in New York grouped by race/ethnicity, neighborhood, and sex; NYPD Complaint Data containing information about the victims (age, race, sec), date and time of the crime, etc.; motor vehicle collisions; Housing Preservation and Development (HPD) data set on buildings, units, and projects that began after January 1, 2014 and are counted towards the Housing New York plan; New York Air Quality data set; etc.

After the completion of the project, the students wrote a report describing what they learned, what conclusions can be made based on this data. Student enthusiastically discussed their findings, and the semester concluded in a lively and productive discussion about the importance of data analysis for the understanding of the city day-to-day life, problems. At the same time, the

students learned about the importance of the broad and, at the same time, detailed view on the data so that the presented numbers tell a true story.

3.3 Evaluation of the Pilot Study

Table 1. Student agreement with statements about data analysis and data analytics (N=9)

Statement	Strongly Disagree (1)	Disagree (2)	Agree (3)	Strongly Agree (4)	Average[12]
Data can teach us about important social issues		1 (11%)	3 (33%)	5 (56%)	3.4
Data analysis is confusing[13]	2 (22%)	4 (44%)	2 (22%)	1 (11%)	2.2
Data analytics can be used for the good of society	1 (11%)	0	2 (22%)	6 (67%)	3.4
Community databases include important information	1 (11%)	0	2 (22%)	6 (67%)	3.4
The world is changing as new data sets become available	1 (11%)	0	2 (22%)	6 (67%)	3.4
I am interested in learning more about data analytics		1 (11%)	3 (33%)	5 (56%)	3.4
I am considering a career in data analytics		1 (11%)	3 (33%)	5 (56%)	3.4
Data analytics is an important career for society[14]	1 (13%)	0	2 (25%)	5 (62%)	3.4
I would like more hands-on experience with data[14]	1 (13%)	0	2 (25%)	5 (62%)	3.4

Fall 2020 was the first time we taught the first course on information and data management in the series of courses designed specifically for our BS in Data Science program. We applied a combination of CBL and PBL methodologies using real-life data sets. At the end of the semester, we asked all the students enrolled in the class to answer a set of questions about the class and their learning. The questions and the answers are presented in Table 1.

[12] Average based on: 1-strongly disagree, 2-disagree, 3-agree, 4- strongly agree
[13] Negatively worded statement

[14] Eight students responded to these statements

As seen in Table 1, the majority of students agreed or strongly agreed with most statements about data analytics. Students reported data analytics is important and interesting. They did not view data analytics as confusing.

We believe the results in Table 1 show that a Data Science program that combines CBL and PBL using real-life data about the city where the students live is likely to "get the students' attention" and may encourage them to study Data Science for their future careers.

Table 2. Student ratings of importance of various tasks involved in data analytics (N=10)

Statement	**Not at all important (1)**	**A little important (2)**	**Somewhat important (3)**	**Important (4)**	**Very important (5)**	**Average[15]**
Communicate the results clearly				3 (30%)	7 (70%)	4.7
Be creative when presenting data		1 (11%)		3 (30%)	4 (40%)	4.0
Be able to problem solve			1 (11%)	5 (50%)	4 (40%)	4.3
Try different approaches when analyzing data				5 (50%)	5 (50%)	4.5
Use visualizations when sharing results			1 (11%)	4 (40%)	5 (50%)	4.4
Understand the context in which the data were collected[2]				4 (44%)	5 (56%)	4.6
Pay attention to ethics[2]				6 (67%)	3 (33%)	4.3
Know about a variety of data analytic approaches[2]				4 (44%)	5 (56%)	4.6
Present data in ways others easy understand is easy[2]				4 (44%)	5 (56%)	4.6
Make data interpretable to the public[16]				6 (67%)	3 (33%)	4.3

Students were also asked to rate the importance of 10 data analytic tasks. Table 2 shows students reported most data analytic tasks are important or very important. Seventy percent of the students reported that communicating the results was very important, the most strongly endorsed statement. Trying different approaches, understanding the context of data collection, knowing a

[15] Average based on: 1-not at all important, 2-a little important, 3-somewhat important, 4- important, 5- very important

[16] Nine students responded to these statements

variety of data analytic approaches, and presenting data in a way that others would understand, were also reviewed as very important by at least 50% of the students. No student rated any task as not at all important, and only one task, "be creative when presenting data" was rated as of little importance by one student.

4. Conclusions

In Fall 2020 we taught for the first time a Data Science class using CBL, PBL and real-life data sets about the city where the students live. Our findings suggest that the launching of the BS in Data Science program is timely as many students are interested in the field. We also believe that our methodology for teaching Data Science will encourage and facilitate students from underrepresented communities to get into the field. The use of real-life data sets and case studies may inspire students to get into the field of Data Science as they can compare their everyday life experiences to the results obtained from data analysis. According to the tables presented in Section 3, we see that the students agree that (1) data can teach about important social issues, (2) community databases include important information; (3) they want to pursue careers in data science.

Although the pilot study only examined student feedback following the completion of the course, in future work, we want to ask the same set of questions at the beginning of the class. This will allow us to examine changes in perceptions about data analysis before and after they take the class. For this pilot study, however, the results are encouraging and suggest that after students complete the first course on information and data management, they value data science and recognize the importance of tasks engaged in by Data Scientists.

These conclusions are also supported by the fact that several students continued working on their class projects and presented their work at a college student conference.

5. References

[1] C. Herreid, "Case Studies in Science: A Novel Method of Science Education.," *Journal of Research in Science Teaching,* pp. 221-229, 1994.

[2] V. Lee, "What is inquiry-guided learning?," *New Directions for Teaching and Learning,* 2012.

[3] J. L. David, "What Research Says About Project-Based Learning," *Educational Leadership,* pp. 80-82, 2008.

[4] R. D. De Veaux, M. Agarwal, M. Averett, B. S. Baumer, A. Bray, T. C. Bressoud, L. Bryant, L. Z. Cheng, A. Francis, R. Gould, A. Y. Kim, M. Kretchmar, Q. Lu, A. Moskol, D. Nolan, R. Pelayo, S. Raleigh , R. Sethi, M. Sondjaja, N. Tiruviluamala, P. X. Uhlig, T. M. Washington, C. L. Wesley, D. White and P. Ye, "Curriculum Guidelines for Undergraduate Programs in Data Science," *Annual Review of Statistics and Its Application,* pp. 15-30, 2017.

[5] L. Mustoe and A. C. Croft, "Motivating Engineering Students by Using Modern Case Studies," *European Journal of Engineering Education,* pp. 469-476, 1999.