Data Science Summer Academy for Chicago Public School Students

Bamshad Mobasher
School of Computing
DePaul University
Chicago, Illinois, USA
mobasher@cs.depau.edu

Lucia Dettori
School of Computing
DePaul University
Chicago, Illinois, USA
lucia@cdm.depaul.edu

Daniela Raicu
School of Computing
DePaul University
Chicago, Illinois, USA
draicu@cdm.depaul.edu

Raffaella Settimi
School of Computing
DePaul University
Chicago, Illinois, USA
rsettimi@cdm.depaul.edu

Nasim Sonboli
School of Information Science
University of Colorado
Boulder, Colorado, USA
nasim.sonboli@colorado.edu

Monica Stettler
School of Computing
DePaul University
Chicago, Illinois, USA
mstettler@sbcglobal.net

ABSTRACT

In this article, we describe DePaul University's summer data science academy for Chicago Public School students which was, in part, funded through the SIGKDD Impact Award program in 2018. The goal of the academy was to increase awareness about data science among high school students. The program specifically aimed to broaden participation of underrepresented groups in computing by targeting economically disadvantaged, African American, Hispanic, and female students. A cohort of 15 high school students from the Chicago Public School system participated in this week-long lab-based data science program learning about a variety of data science methods and their applications, including data visualization, distance-based methods, classification, clustering and others. The group comprised of 75% African American and Hispanic students, 58.3% of whom were female.

Keywords

Data science, summer academy, Chicago, public schools.

1. INTRODUCTION

In the past few years, there has been a strong interest among academia and private organizations to support educational initiatives in computer science for K-12 students. Existing programs focus on a variety of topics from computational thinking to application development, game design and web development. However, very few K-12 programs focus specifically on data science, and there are limited opportunities for high school students to learn and understand this exciting field. As a result, not many students consider pursuing data science as a major in college or as a career, which is contributing to the pipeline issue in meeting increasing demand for professionals in the field.

The KDD Impact Award program [1] provided DePaul University School of Computing support to develop a week-long summer program [2] to introduce high school students to the fast-growing field of data science through a series of hands-on activities. The project developed new educational offerings to address the learning needs of current and future generations. The intent of this summer program was to excite students and help them build confidence in their problem-solving skills through data science. Specifically, the program had the following goals:

- Get high school students (particularly from underrepresented groups) excited about pursuing data science in college and careers.
- Teach basic data science concepts, tools & techniques with hands on applications for problems students can relate to.
- Introduce students to the wide variety of domain applications and career possibilities in data science through guest speakers.

The program curriculum and activities introduced students to the full data science cycle from data cleaning and visualization to model building and validation. The program taught students the power of computing for extracting information from data, encouraged them to think critically about the data around them, and to be informed and critical readers of quantitative arguments or claims found in news articles or social media discourse. Multiple guest speakers from industry presented various applications of data science in a variety of domains. Speakers included data science professionals from companies such as Microsoft, Facebook, BMW, SproutSocial, and others. In addition, the program served as the basis for developing pathways for students in the Chicago Public Schools (CPS) into university data science programs such as the B.S. in Data Science degree offered at DePaul University [3].

The KDD Impact Award provided support for the key personnel, including faculty and graduates students who developed curriculum and conducted the program. The Award also provided financial assistance for student participants in the form of a stipend. The College of Computing and Digital Media at DePaul University provided additional support including food, supplies, computer labs and technical support.

The program was conducted between June 25 and June 29, 2018. A cohort of 15 CPS students were selected from among 173 applicants to participate in the program. Of these students 13

completed the program. The pre and post surveys conducted by the program staff, as well as performance evaluation of students during hands on exercises and the final project indicated a significant increase in the knowledge of data sciences and its applications, as well as the use of specific data science methods.

2. COLLABORATION BETWEEN DEPAUL UNIVERSITY AND CHICAGO PUBLIC SCHOOLS

In addition to increasing the number of students interested in the field, the program specifically aimed at broadening participation of underrepresented groups in computing by targeting economically disadvantaged, African American, Hispanic, and female students. We recruited the cohort of students from high schools in the Chicago Public School District (CPS). CPS is the third largest school district in the country with 167 high schools varying from neighborhood to selective enrollment schools. Of the 371,000 total students, over 107,000 are high school students. Of the total student population, approximately 37% are African American and 47% are Hispanic, while 78% are economically disadvantaged and 50% are female.

CPS recently made Computer Science a high school graduation requirement. This was made possible by a collaboration between CPS and faculty at DePaul, Loyola, the University of Illinois at Chicago, and The Learning Partnership [4] who, through several NSF-funded grants, brought to CPS the Exploring Computer Science (ECS) curriculum [5]. The Partnership formalized into the Chicago Alliance for Equity in Computer Science (CAFECS) and a *CS for All* Department was created at CPS. The project utilized the Partnership to recruit teachers and students. The leadership of Chicago Chapter of the CS Teachers Association (CSTA) also helped deepen the connections between these stakeholders and the CS education community.

The School of Computing at the DePaul College of Computing and Digital Media (CDM) is uniquely positioned to offer this type of program because of its extensive offerings and faculty expertise in the field of Data Science. CDM has been offering a MS in Predictive Analytics (recently renamed MS in Data Science) degree [6] with concentrations in Computational Methods, Health Care, and Marketing since 2010 with a 5-year enrollment growth from 30 to over 300 students. In 2017, a BS in Data Science degree was launched.

CDM faculty includes nine full time members specializing in data science, and employs a number of data science professionals to teach specialized courses as adjunct faculty. Collaboration with industry is supported through the DePaul Center for Data Science [7], a joint venture including faculty from several colleges and disciplines across the University who are interested in applications of data science. The Center serves as a point of contact between industry and academia to foster partnerships for research and education. The Center supports projects in areas such as Web intelligence, social computing, recommender systems, biomedical informatics and healthcare, hospitality, marketing, imaging informatics, and more.

The program curriculum and activities were designed and taught by DePaul faculty associated with the Center for Data Science. Two graduate students provided support in organizing the logistics of the week, as well as assisting the students during the activities. In addition, a CPS high school math and computer science teacher acted as a consultant for the program and participated in the academy. The goal of having a CPS consultant was in part to develop data science expertise to the teacher could take back to high school classrooms and to further promote data science concepts and career opportunities in CPS classrooms.

3. SELECTION PROCESS AND PARTICIPANT DEMOGRAPHICS

CPS assisted us in broadly disseminating the program information and the inline application across the school districts in Chicago. The dissemination was done not only through the standard communication channels, but also by asking CPS teachers for referrals of students who might benefit from program. CPS high schools vary widely in terms of financial resources, support, and students learning outcomes. To ensure broad representation, the selections were made across a wide spectrum of schools in the district.

There were 173 applications for 15 spots in the academy. The program accepted applications from Chicago Public Schools high school students who had completed their sophomore year by July 1, 2018. Students had to have completed the "Exploring Computer Science" course or equivalent, and a course in geometry and/or algebra to be eligible.

While academic performance was one of the key criteria in the selection, gender and ethnic diversity were also taken under consideration to ensure we fulfilled the mission of the program. Our first step in the selection process was to remove those with a GPA of less than 3.0. A 3.5 GPA was a second level soft cut. We did accept two students with 3.0 GPA, as their essays were compelling, they came from disadvantaged backgrounds and we felt that the impact of participating would be significant for them. Compelling essays were key to determining the students' level of interest and commitment. CPS grades its schools on several factors related to academic performance. Those measures were factored into our selection to ensure diversity of school quality. The final selections were made with the goal of creating a mix of gender, ethnicity, and quality of high schools.

The 15 students represented the following high schools: Wells Community Academy High School, George Westinghouse College Prep, Mather High School, Whitney M. Young Magnet High School, John Marshall Metropolitan High School, Albert G. Lane Technical High School, Marie Sklodowska Curie Metropolitan High School, Friedrich W. von Steuben Metropolitan Science High School, and Nicholas Senn High School. In addition, high school math and computer science teacher Michael Kolody from Nicholas Senn High School assisted throughout the week. Of the 15 chosen, 13 completed the program.

The group comprised of 50% Hispanic, 25% African American, 16.7% Asian, 16.7% Caucasian and 8.3% Native American students. Gender broke down into 58.3% female and 75% were high school juniors.

Figure 1 depicts the demographic makeup of the final cohort of 15 students selected for participation in the summer academy.

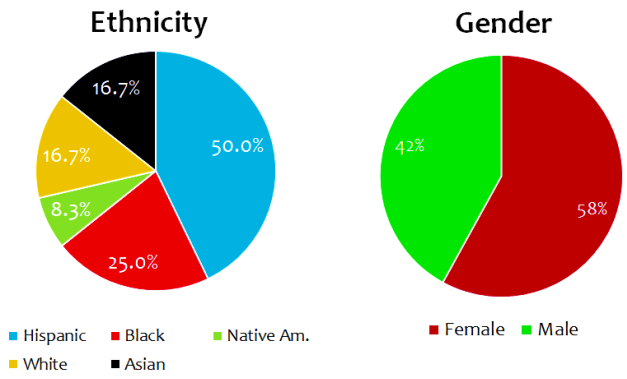

Figure 1. Demographics of Participants. Note that Economic data was not gathered, however 78% of the CPS population are from economically underserved areas.

4. SUMMER ACADEMY CURRICULUM AND ACTIVITIES

Students engaged in hands-on activities taught in a computer lab setting by full-time DePaul faculty members with real world experience. Projects were drawn from real-life applications around topics that are of interests to the students.

The students were exposed to a variety of tools commonly used for data exploration and analysis. The goal was not to teach coding skills, but to gain familiarity with how different tools serve different needs. Simplicity of use was key so that students could run various types of analyses. The tools used were Matlab, SPSS, Tableau and Jupyter with Python.

Slack was used as the main communication tool for the program. Staff used Slack for announcements, reminders and to share documents and datasets. Students were able to communicate with each other and ask questions to staff at any time. The pre and post surveys of students were also conducted via Slack.

Datasets used were from: Divy (a bicycle sharing system in the City of Chicago and two adjacent suburbs operated by Motivate for the Chicago Department of Transportation.), Human Activity Recognition dataset, ATT face recognition dataset (for image classification) and Spotify song feature data set. These data sets were used for hands on exercises associated with individual modules (see below for details). The Spotify data set was used throughout the week-long program as the basis for a final project and presentation that brought together various concepts such as classification, clustering and distance-based methods. For the purpose of the final project and presentation, students were organized into 5 groups of up to three students.

We invited undergraduate, Master's and PhD students to share their research work with the group during lunch time and other social periods. Topics covered were: recommender systems, movement behavior analysis of roundworms, using machine learning for suicide prevention, and computer-aided medical diagnosis. Furthermore, guest speakers from industry presented various applications of data science in a variety of domains. Speakers included data science professionals from companies such as Microsoft, Facebook, BMW, Sprout Social, and others. Guest speaker presentations were typically scheduled during the lunch periods.

Table 1: Schedule of Activities - June 25-29, 2018

Mon	• Opening remarks, introductions, entrance survey, ice breaker activity • The ABCs and Applications of Data Science (DS), Machine Learning, and AI • Careers and type of jobs in DS, ML, and AI • Research project presentations by undergraduate, Master's and PhD students at DePaul • Overview of the overarching project for the week: Analysis of Spotify data set • Industry Speaker: Jeff Hamilton, Head of Consumer Insights BMW Group
Tue	• Data Summarization and Visualization module (using SPSS & Tableau) • Lab activity: Data summarization and visualization with Chicago Divvy bikes usage during the past year • Group project activity: Characterize the Spotify data set • Industry speaker: Becky Jacob, Data Scientist, Sprout Social
Wed	• Model building through classification using decision trees (using Matlab & SPSS) • Lab Activity: extracting pixel data for face recognition used in image classification • Group project activity: build classifier to classify Spotify songs into existing categories based on song features • Industry Speaker: Adam Hecktman, Microsoft's Director of Technology & Civic Innovation for Chicago
Thu	• Model building through distance-based methods – (using Python, pandas, scikit-learn, Jupyter Notebook) • Classification techniques using K-Nearest-Neighbor lecture and hands-on lab session • Jupyter Notebooks, Python, pandas, scikit-learn • Lab activity: Using KNN for song recommending similar songs • Group project activity: performing cluster analysis and visualization on Spotify song data • Guest Speaker: Jon Gemmell, Assistant Professor, DePaul University – Artificial Intelligence & Machine Learning
Fri	• Students' final presentation • Post Survey • Industry speaker: Amy Foran, Marketing Science Partner, Retail Vertical, Facebook • Reception (including parents and other DePaul Faculty)

Table 1 provides the detailed schedule of content modules and other activities for the program, including the names of guest speakers. All the materials used in the program including PowerPoint slides, data sets, and relevant code are available at the Github repository [8] for the Data Science Academy.

5. PROGRAM OUTCOMES

Overall, the students in the program showed a significant increase in interest in data science as well as knowledge about the use of different data science methods such as exploratory data analysis, data visualization, classification, and clustering. This was especially evident in the work students did in their final projects involving the detailed analysis of the Spotify song data set (including using clustering, characterization, and visualization).

In addition, in order to gauge the impact of the program on students, we conducted pre and post surveys of students. At the beginning of the week, only 34.6% of the students reported knowing what data science was. At the end of the week, all the students felt that they had a good understanding of data science. Their interest in pursuing data science as a career increased from 65.4% of students to 91.7%. All the students reported that they liked the camp and that they learned a lot. They also felt that the pace was just right for them. The industry speakers were by far their favorite part of the week, followed by data visualization. At the end of the program, more than 90% of the participants agreed (38%) or strongly agreed (58%) that they understood classification methods (KNN and decision trees) and how they should be used. About 88% said that they understood when clustering should be used. More than 80% stated that they knew how to use data visualization to tell a story.

The students commented that they appreciated the academic rigor of the program. They also liked having the Spotify dataset project that spanned the week where they could apply and practice what they learned each day. They also reported that it was challenging to stay focused for a 2-hour lecture. They suggested either having more breaks or breaking up the lecture with more hand-on activities.

The full survey results are available at the Github repository [8] for the summer academy.

6. CONCLUSIONS

Our goal was to expose high school students to data science and get them excited about it as a possible field of study and a career choice. The results of the program exceeded our expectations. We plan to use this experience as a stepping stone to developing additional programs with expanded reach. We also hope that the experience and the materials made available can help others in developing similar programs.

7. REFERENCES

[1] https://www.kdd.org/News/view/announcing-the-kdd-impact-program-recipients-for-2018

[2] https://ddssacademy.wordpress.com/

[3] https://www.cdm.depaul.edu/academics/Pages/BS-in-Data-Science.aspx

[4] http://www.thelearningpartnership.com/

[5] http://www.exploringcs.org/

[6] https://www.cdm.depaul.edu/academics/Pages/MS-in-Data-Science.aspx

[7] http://cds.cdm.depaul.edu/

[8] https://github.com/nasimsonboli/data_science_summer_academy

About the authors:

Bamshad Mobasher served as the PI overseeing the program. He is a Professor at DePaul School of Computing and co-founder of the Center for Data Science. He is also the director of the Center for Web intelligence where he conducts research and works with industry on problems related to machine learning, Web mining, recommender systems, information retrieval, and social computing. He has served in leadership positions on related journals and conferences, including SIGKDD.

Lucia Dettori served as the co-PI in charge of advertising, recruitment and selection of candidates in collaboration with CPS. She is an Associate Dean at the DePaul College of Computing and Digital Media. She is one of the founding members of the Chicago Alliance for Equity in Computer Science (CAFECS) and currently serves on its leadership team. Her main research focus is Computer Science Education with special emphasis on broadening participation.

Daniela Raicu served as senior personnel for the project. She is a Professor at DePaul School of Computing and the founding Director of Center for Data Science. She is also the founding co-director of the Medical Informatics (MedIX) and the Intelligent Multimedia Processing Laboratories. She has been a leader in promoting undergraduate research at the national levels. She has mentoring over 100 undergraduates through the NSF REU MedIX Program since 2005.

Raffaella Settimi served as senior personnel to oversee the logistics of the program and the design of the surveys. She is an Associate Dean at CDM, and co-founder of Center for Data Science. She has extensive experience teaching data mining courses at both the graduate and undergraduate level and has led several data science projects in collaboration with a variety of organizations both in the private and public sectors.

Monica Stettler served as a graduate assistant for the program. She is an Alumnus of the DePaul MS in Data Science Program and also has an MBA in Finance. She was in charge of developing and managing the application and selection processes and assisted with the program logistics.

Nasim Sonboli served as a graduate assistant for the program. She was PhD student at the School of Computing and a research assistant at the Center for Web Intelligence. She assisted with the development of content modules for the academy, mentored students in their lab activities, and helped with day-to-day logistics for the program. She is currently a PhD student in Information Science at the University of Colorado, Boulder.