Computing Community Consortium CS for Social Good White Paper

Using open-sourced data to nurture a civically engaged and computationally fluent generation

Monica Chan Teachers College, Columbia University
& Ipek Ensari Columbia University Data Science Institute

ABSTRACT

As we are living in an increasingly data-oriented and data-driven 21st century, there is a need to educate our next generation to be computationally fluent in data science and analytics, yet additionally be cognizant about the ever-changing climate of the society they live in by staying civically engaged. This white paper responds to the call for innovative computing research in the intersection of computer science, data science and computing education research. We propose suggestions for interdisciplinary partnerships to make community data available and open-sourced for the purpose of effective, interdisciplinary K-12 data science education, with a curriculum that emphasizes civic engagement and provides access to all.

INTRODUCTION

Over the past years, many institutions ranging from government to academia and corporate firms have used data science and data analytics to make decisions, predict phenomena, and affect human behavior. We are in an era where our toddlers articulate “Alexa, play Baby Shark”, and our children are growing up in a world where their social media platforms magically recommend their favorite artistes and online stores. These emerging technologies may be widespread now, but our education system is not yet up to speed with curriculum and assessment metrics to educate our youth about how these emerging technologies work.

PROBLEM STATEMENT

Data science is a field growing rapidly in all industries, but there is no standardized method, resource or platform to teach our students about the data science that drives the technology that we interact with. There exists recently developed educational frameworks such as the Next Generation Science Standards (NGSS) (Pellegrino et al., 2014) and K-12 Computer Science Framework (Grover & Pea, 2013; K-12 Computer Science Framework Steering Committee, 2016;

Yadav, Hong & Stephenson, 2016), and Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Franklin et a., 2005) that highlight scaffolding for K-12 STEM education, but these do not directly address data science for K-12 levels. Current practice is also limited, because data science is a new field to K-12 teachers, too. Several private high schools have begun offering elective courses in machine learning, artificial intelligence and advanced data analytics (The Nueva School, 2019), but these curricula are hardly democratized beyond the high school itself.

Another societal problem that has surfaced is that people are often caught in echo chambers, be it in their social media platforms or with the people they typically interact with. To build a more civically engaged, socially intelligent and politically aware generation of youth, we need to find areas in school curricula for youth to explore dominant and nondominant narratives of their neighborhood, society, and culture. As datasets on societal services and communities have become public and more data tools such as Google Toolbox Dataset Search become available, a wider population has been able to access these data sets online. However, there are still gaps between these public datasets and K-12 data science education: How can we facilitate and scaffold this connection? How could research in various fields of computer science such as human-computer interaction (HCI) and machine learning (AI) construct connections to facilitate K-12 data science learning? There are many education technology tools that emphasize STEAM and maker-centric learning that have leveraged advances in computer science research to improve user experience and personalize learning, thus how could we build on similar practices to construct tools for civically-oriented K-12 data science education?

In this white paper, we provide a literature review of seminal learning theories and past studies that have been conducted in K-12 data science education. Success would be to translate these ideas into a K-12 curriculum in the form of learning activities, pedagogical strategies and perhaps a platform to educate youth at various age groups about sub-fields of data science, from analytics to algorithms and machine learning. Our primary beneficiaries would be K-12 students and teachers, but also designers (curriculum, instructional and product) and engineers who may be involved in building the platform later.

LITERATURE REVIEW

Computational Thinking & Computational Fluency

Many scholars have defined *computational fluency* and *computational thinking* in various manners. *Computational fluency* highlights the development of youths' perspectives as computational creators and thinkers with the ability to problem-solve, understand computational concepts, and express themselves with digital technologies (Resnick et al., 2013). Brennan and Resnick (2012) summarize *computational thinking* as the set of processes that involve formulating

problem questions and creating solutions represented in a form via an information-processing agent (Cuny, Snyder & Wing, 2010). They further frame *computational thinking* and *computational fluency* along three dimensions: concepts (e.g. conditionals), practices (e.g. debugging), and perspectives (e.g. viewing oneself as the creator and collaborator) (2012; Resnick et al., 2013).

Data Science in K-12

Gould and the Mobilize Team at the University of California Los Angeles (UCLA) had led an NSF-funded project that introduced data science and statistics to high school students in the Los Angeles Unified School District (Gould et al., 2018; Gould et al., 2016). The Mobilize team had designed a yearlong curriculum named Introduction to Data Science for the LAUSD, and also received California's approval to partially satisfy the mathematics requirement for admission to the University of California and California State University. Curriculum design relied heavily on GAISE (Franklin et al., 2005), and emphasized the Data Cycle as a four-step investigation process (ask questions, consider data, analyze data, interpret data) through "campaigns" the Mobilize team designed (topics included food, time use, stress and water use) and a collaborative classroom-designed campaign using open source data.

Gould and colleagues' study indicated that there was positive interest amongst schools and students, but acknowledged that their evaluation methods failed to capture the computational and statistical thinking aspect they had hoped to develop (Gould et al., 2016). This study demonstrated a strong effort in implementing a high school data science curriculum across a large school district, and we hope that our work and suggestions will build on it. This brings us back to our problem statement - there are currently no standardized guidelines on K-12 data science education, hence the lack of measurability of learning. We also desire to strengthen the civic engagement portion that is apparent in Gould and colleagues' study through their use of "campaigns" and open source data, and will present two in-progress pilot studies to illustrate.

CURRENT WORK

To address the gaps presented in our problem statement, we present two pilot studies at Columbia University that we conducted to demonstrate evidence to our ideas on using public datasets to educate K-12 learners about data science and analytics in a civic-oriented manner.

Pilot Study 1: Dog Data Scientists (Columbia Data Science Institute)

This was a project led by the Data Science Institute at Columbia University and undertaken by an interdisciplinary team of researchers from various schools at Columbia University. This project was showcased at the annual Columbia Alumni Association STEM Day, an event that 600+ visitors attended on one day. The goal of this day was to introduce children to fields of STEM in an interactive, engaging way to encourage curiosity and explore questions relevant to their

everyday lives through STEM. The Dog Data Scientists project was created to introduce the field of data science to children between the ages of 8 and 14. The data was sourced from NYC Open Data, where data on a variety of metrics related to the city of New York are freely available for use. This project extracted data on dogs to create interactive dashboards and maps that the children could explore to answer a set of research questions. The activity steps were guided by research methods procedures (i.e., visual exploration, analysis, inference and ultimately dissemination of data). The dashboards included bar, line and pie charts, which the children could scroll through and extract information (e.g., frequencies), aggregate or disaggregate layers of information, to find the correct answer to their research question (e.g., "Which zip code had the most number of dog bites in 2015? How does that compare to the national average?", "Looking at this bar chart, can you tell which dog name was the most popular overall?"). The interactive map encouraged the children to apply their geospatial ability to make inferences related to a variety of information on dogs of the city. After exploring the interactive dashboards to gain insights and make inferences, children had to generate a statement of their findings to disseminate with the community, for which the project's Twitter account was used (view Appendix for screenshots of students' work).

Pilot Study 2: *Make with Data* Project (Teachers College)

Teachers College and Lamont-Doherty Earth Observatory at Columbia University are currently running a research pilot that aims to introduce data science and analytics to high school students and high school teachers from the five boroughs. This research project pairs high school students and teachers with professional data scientists, and students and teachers embark on a yearlong community-oriented project that uses open source data from NYC Open Data. The goals are to have students and teachers engage in their communities through investigating data sets and formulating research questions to drive their investigations, and to strengthen students' sense of ownership of specific technical skills required for success in STEM majors they might intend to pursue in higher education. For teachers, the goals are to have them expand their understanding of student capacity and improve their personal capacity for data to increase comfort with hands-on science. We highlight a flat hierarchy between students and teachers, in that both parties are collaboratively driving the projects they chose and working as teammates, as compared to a more typical dynamic where teachers decide on a project and students follow the instructions.

Through the first few months of the pilot, we have discovered that students and teachers choose research questions that are very personal and meaningful, typically on topics such as income inequality, mental health, and crime around their neighborhood. Although pivots and modifications happened during the process, this personalization motivates them to learn the technical skills and content knowledge about data science required to find the answers. During final presentations in June 2019 (the end of the pilot), students and teachers presented their

findings in a coherent manner, but also highlighted technical challenges they faced, missing data and skepticism regarding data ethics and issues in data collection.

SUGGESTIONS

Approaching from the HCI perspective

Given the target demographic, Dog Data Scientists was created while keeping in mind that the target demographic included children of diverse abilities and previous skill levels. As a first introduction to data science, our dashboards included language that at least a 2nd-grade student could understand (since age 8 was the lower limit). We categorized inquiry questions prepared based on difficulty level to engage the appropriate age level, and excluded any terms that would be jargon. Importantly, the project aimed to introduce the idea of generating scientific questions and inference through a dataset that children could understand and relate to at some capacity (i.e., dogs in a geographic area they live in and are familiar with), and to surface the "human aspect" of data science (i.e., as opposed to data science being an "abstract" concept that lacks applicability to day-to-day activities). As such, from a human-computer interaction perspective, this project aims to maximize effective engagement with the interactive dashboards and *autonomously generate* as many meaningful inferences from visualization by keeping the interaction language and interaction flow as simple and straightforward as possible.

Similar to Dog Data Scientists, the Make with Data project could extend in a more HCI-oriented manner, such as examining the students' and teachers' user experience of the various data analytics software they used, or by implementing a different methodology of studying the students' and teachers' learning process, for example through a more quantitative lens using eye-tracking as compared to qualitative case studies.

Approaching from the AI / ML perspective

While the Dog Data Scientists and Make with Data projects did not have direct AI/ML components, add-on components that integrate AI and MI can easily be accomplished. A future direction for Dog Data Scientists could involve adding in a clustering model into the dashboards that relies on unsupervised learning to investigate whether certain parts of the city have similar or dissimilar characteristics with regards to the data on dogs from these areas. Such an activity would not only generate information with potentially meaningful implications, but also be done in a way that keeps children of this age group engaged.

IMPLICATIONS

Dog Data Scientists is unique in that it combines uses of open source data and software to create an interactive STEM activity for children while also adhering to the methodological steps of data science and research (i.e., data extraction-visualization-analysis-inference-dissemination). Due to these characteristics, the Data Science Institute in collaboration with Teachers College is currently looking into implementing the approach this project takes into a K-12 STEM education curriculum and holding hands-on training for K-12 teachers on how to teach data science in their classrooms. For the *Make with Data* project, 2019-2020 will be the second year of the pilot, where Teachers College and Lamont-Doherty Earth Observatory plan to help the teachers who participated this year implement data science modules in the subjects they teach at their schools.

There are certainly broader implications regarding implementing a K-12 data science curriculum and building a learning platform for data science. This movement would involve city and state governments, schools, and partnerships between nonprofit and for-profit organizations at various levels of research, design and implementation. An entirely new pedagogical experience would be created.

CONCLUSION

In response to the call for innovative computing research in the intersection of computer science, data science and computing education research, we demonstrate through two pilot projects that interdisciplinary partnerships can create opportunities for effective, interdisciplinary K-12 data science education that integrates civic engagement and is available to all regardless of socioeconomic status, income or academic backgrounds. Data Science as a field on its own is domain-agnostic and has applicability and use across a wide range of domains. As such, introducting data science as part of STEM education in a hands-on approach to children of school age can more effectively engage them to gain interest and computational fluency in STEM fields, while deepening their understanding of their communities and society.

REFERENCES

Brennan, K., & Resnick, M. (2012, April). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American Educational Research Association, Vancouver, Canada* (Vol. 1, p. 25).

Cuny, J., Snyder, L., & Wing, J.M. (2010). Demystifying computational thinking for noncomputer scientists. Unpublished manuscript in progress, referenced in http://www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Perry, M., & Scheaffer, R. (2005, August). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*(Rep.). Retrieved June 28, 2019, from American Statistical Association website: https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf

Gould, R., Moncado-Machado, S., Johnson, T., Molyneux, J., & Trusela, L. (2016). Teaching Data Science to Secondary Students: The Mobilize Introduction to Data Science Curriculum. In *International Association for Statistical Education*. Retrieved June 19, 2019, from https://iase-web.org/documents/papers/rt2016/Gould.pdf

Gould, R., Moncado-Machado, S., Johnson, T., Molyneux, J., & Trusela, L. (2018). Mobilize: A Data Science Curriculum For 16-year-old Students. In *International Conference on Teaching Statistics*. Retrieved June 19, 2019, from https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9B1.pdf

Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. *Educational researcher*, *42*(1), 38-43.

K-12 Computer Science Framework Steering Committee. (2016). K-12 computer science framework.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing Assessments for the Next Generation Science Standards*. National Academies Press. 500 Fifth Street NW, Washington, DC 20001.

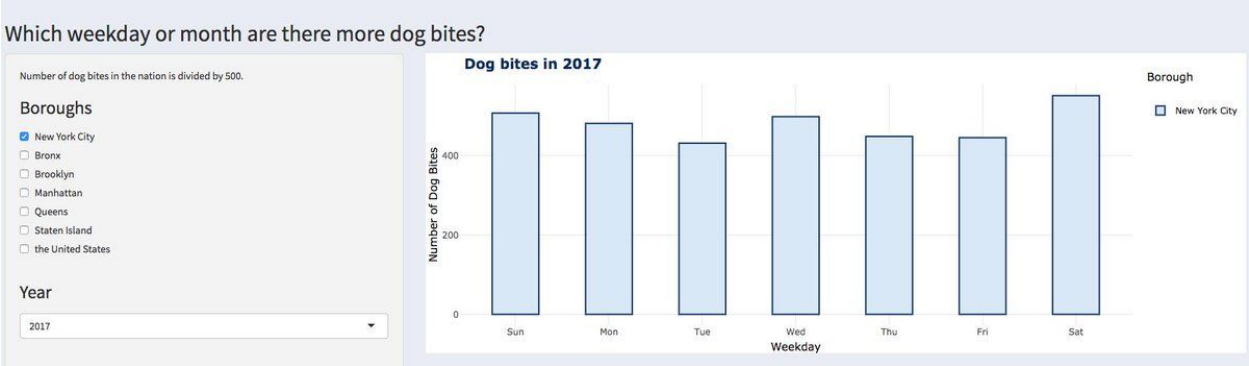
Resnick, M., Ito, M., Gasser, U., Rusk, N., & Schmidt, P. (2013). Coding for all: interest-driven trajectories to computational fluency. *Proposal to the National Science Foundation*.

The Nueva School. (n.d.). Upper School Design Engineering & Computer Science. Retrieved June 19, 2019, from https://www.nuevaschool.org/student-experience/upper-school/design-engineering-computer-science

Yadav, A., Hong, H., & Stephenson, C. (2016). Computational thinking for all: pedagogical approaches to embedding 21st century problem solving in K-12 classrooms. *TechTrends*, *60*(6), 565-568.

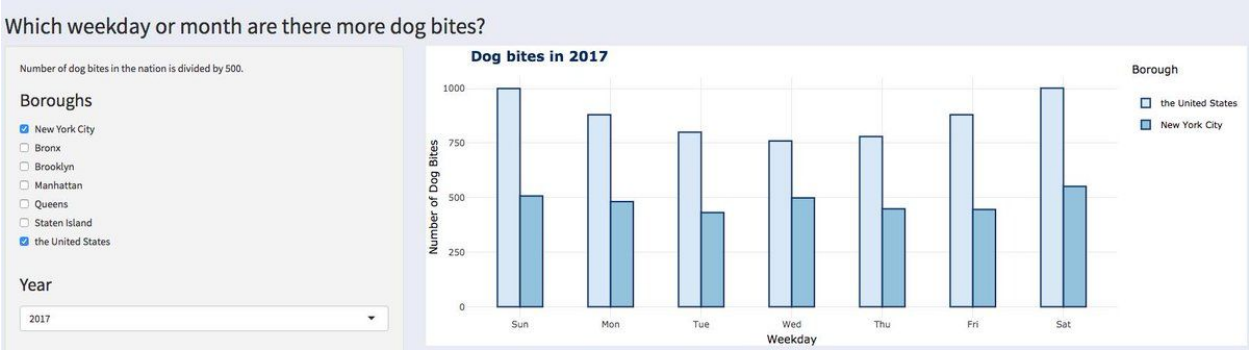
Appendix

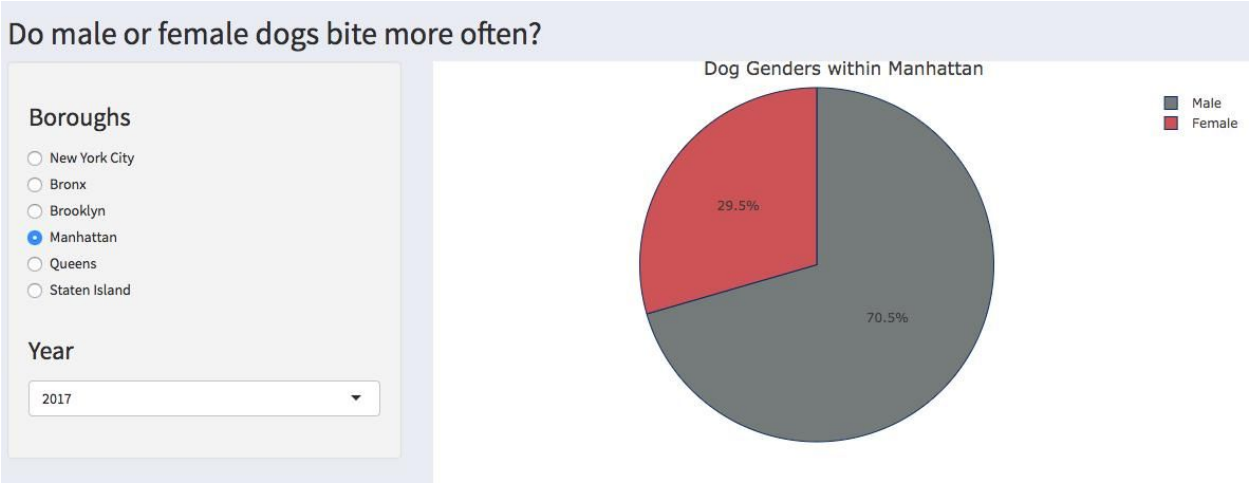
Dog Data Scientists screenshots:



An example of an insight generated by a child: "Dogs aged 1 bite approximately 218 more bites than dogs aged 17, probably due to having more energy." Here is to youth and another great

finding unearthed by (name removed)! Thank you for your help! #CAASTEMDAY #nycopendata @DSI_Columbia @Columbia





Another insight by another kid (no figure though, for some reason): "Yorkshire terriers are the most popular breeds in NYC and yet they are among the least biters." Who would have thought, thank you for unearthing this, Ivy!! Great finding from a budding data scientist. #CAASTEMDAY #columbia @DSI_Columbia #nycopendata