Learning With Data: Designing for Community Introspection and Exploration

Sayamindu Dasgupta
MIT Media Lab
Cambridge, MA 02142, USA
sayamindu@media.mit.edu

Benjamin Mako Hill
University of Washington
Seattle, WA 98195, USA
makohill@uw.edu

Abstract
In this position paper, we present ongoing research that aims to engage young people in the analysis of data about their own participation and learning in a large online community. We describe a new system we have designed to let users of the Scratch online community analyze and reflect on their own participation, which, we argue, represents a new model of and a new pathway to data science. We conclude by connecting back to some of the core questions in human-centered data science and describing some of the challenges and possibilities with the approach we have outlined.

Author Keywords
data science education; human-centered data science; block-based programming

ACM Classification Keywords
K.3.2 [Computers and Education]: Computer and Information Science Education—*Information systems education*

Introduction
Much of the emerging discourse around human aspects of data science has focused on implications of human-derived datasets (e.g., datasets from online communities or urban activities) and the practices and processes of professional researchers and scientists who try to extract meaning from

them. In this position paper, we draw attention to the potential of providing the human data-sources, i.e., the participants in online communities, with opportunities to engage in analysis of data about themselves, as well as to some of the challenges raised by this approach. We describe an ongoing research project to provide members of the Scratch online community with the ability to analyze and visualize data about their activity on the website. We argue that this work has the potential to empower Scratch's community of youth programmers to understand and reflect on their own participation and learning. We also suggest that it may provide new pathways to learning core data-science concepts. Following a description of a new system, we reflect on broader questions from the agenda of human-centered data science, i.e., how do we leverage data science to study, evaluate, and improve the design of systems that enable end-user data science?

Background & Motivation

Scratch is a visual, block-based[1], programming language and community designed for children and youth aged 8-16 [5]. The Scratch online community is a venue for Scratch users to share projects, comment on each other's work, follow other users, and engage in a variety of other forms of social interaction. As of late 2015, the online community hosts more than 12 million publicly shared projects, more than 9 million registered users, and more than 62 million comments. The Scratch community has been a rich source of data for both qualitative and quantitative researchers interested in a wide range of questions, from examining the leadership roles played by young people [6] to understanding the trade-offs between generativity and originality in remixing [3].

[1] Scratch programming primitives are called blocks because they are represented by on-screen visual blocks that snap together (see Figure 1).

Although data has always been public on the Scratch website, the need for tools, application programming interfaces (APIs), and programming languages outside of Scratch has effectively placed analysis of data from Scratch in the domain of professional researches and data scientists. That said, the Scratch dataset contains information of deep personal interest to Scratch users. After all, the Scratch dataset consists of the digital footprints created by Scratch users and their friends. Providing programmatic access to data about Scratch within the Scratch programming environment would enable users to conduct their own data analysis and visualization in a familiar setting, in any way they choose, and in a personally relevant context. In the following section, we describe a system to enable such access.

A New System: Community Data Blocks

To allow programmatic access to the Scratch dataset from Scratch, we have designed a series of new blocks within Scratch, called *community data blocks*, that encapsulates queries to the Scratch back-end database. It should be noted that all data accessible through these blocks are public data visible on the Scratch website. The blocks are capable of queries of both social interaction data and code-metadata.

Users can use these new blocks for a wide variety of purposes. For example, a Scratch user could use the blocks to find out the most common type of blocks she has used or discover which blocks she has never tried. Some sample Scratch scripts using these blocks are shown in Figure 1. The first script makes a list of all projects shared by a user that use blocks in the "pen" category. The second script identifies all followers of a Scratch user who are from India. The third script finds all the projects bookmarked by followers of a user that manipulate or play sound. These blocks have been implemented and are currently being al-

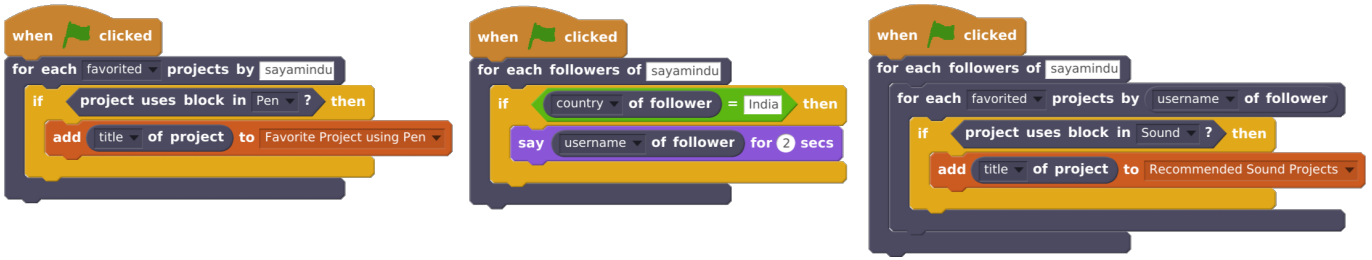

Figure 1: Scratch scripts to query the Scratch dataset. The first script makes a list of all projects shared by a user that use blocks (Scratch primitives) in the "pen" category. The second script identifies all followers of a Scratch user who are from India. The third script finds all the projects bookmarked by followers of a user that manipulate or play sound.

pha tested by a small group of Scratch administrators and researchers. We plan to deploy these blocks in early 2016.

Studying the Effects of Community Data Blocks

While the system described above aims to enable youth programmers to engage in data-science, it also provides an opportunity for researchers to understand how users engage with the programming blocks and learn. Toward this end, we are designing both quantitative and qualitative studies to understand how Scratch users employ the blocks and how their use of the blocks shapes their other activities. One major set of challenges in this regard stem from the unstructured and informal nature of Scratch. As there is no defined lesson plan or expected output in Scratch, the trajectories of Scratch users over time vary enormously and learning happens along many dimensions. Quantitative measures of learning in informal learning environments such as Scratch are still being developed and evaluated [7, 2]. We plan to leverage previous measures and build on them to evaluate changes in the learning outcomes of Scratch users who program with these new blocks.

Certainly, insights into important concepts including users' (mis)conceptions about computational thinking concepts, their understanding of functional possibilities of data and computational tools, and their mental models of computational systems are often best understood through qualitative research, and we plan to interview users and analyze their projects qualitatively. For example, to study a related system in Scratch that enables users to persistently store data in a cloud-based infrastructure, we asked individual Scratch users to draw on paper how they thought the system worked [1]. These drawings gave us insights into the mental models the young programmers had about the system in question, and we plan to utilize and build upon similar strategies in the evaluation of the community data blocks.

Conclusion

We believe that designing for the ability to analyze and visualize one's own social data is a promising way to engage end-users in data science. As a community focused on programming, Scratch seems particularly well suited to explore these possibilities; however it is not unique. For example,

one of the most popular features in the Wolfram Alpha computation system was a tool that allowed Facebook users to visualize, analyze, and cluster their own personal social network data [4]. We hope that the system we have created and the analyses that we have proposed to evaluate the system, will help demonstrate that data science has potential not just for scientists and researchers but also for a much wider audience. We hope that our work will contribute to the exploration of what a more democratized model of data science might constitute, as well as a more informed understanding of its effects, promise, and limitations.

Acknowledgments

Financial support for this work was provided by the National Science Foundation (grants DRL-1417663 and DRL-1417952). We would like to express our gratitude to Mitchel Resnick, Natalie Rusk, and Hal Abelson for their guidance and feedback.

REFERENCES

1. Sayamindu Dasgupta. 2012. *Learning with data : a toolkit to democratize the computational exploration of data*. Thesis. Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/78203

2. Sayamindu Dasgupta, William Hale, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016 (forthcoming). Remixing as a Pathway to Computational Thinking. In *Proceedings of the 19th Conference on Computer Supported Cooperative Work & Social Computing (CSCW '16)*. ACM.

3. Benjamin Mako Hill and Andrés Monroy-Hernández. 2013. The Remixing Dilemma The Trade-Off Between Generativity and Originality. *American Behavioral Scientist* 57, 5 (May 2013), 643–663. DOI: http://dx.doi.org/10.1177/0002764212469359

4. Wolfram Research. 2015. Wolfram Research News » Wolfram|Alpha Personal Analytics for Facebook: Last Chance to Analyze Your Friend Network. (April 2015).

5. Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, and Yasmin Kafai. 2009. Scratch: Programming for All. *Commun. ACM* 52, 11 (Nov. 2009), 60–67. DOI: http://dx.doi.org/10.1145/1592761.1592779

6. Ricarose Roque, Natalie Rusk, and Amos Blanton. 2013. Youth Roles and Leadership in an Online Creative Community. In *Computer Supported Collaborative Learning Conference Proceedings*, Vol. 1.

7. Christopher Scaffidi and Christopher Chambers. 2011. Skill Progression Demonstrated by Users in the Scratch Animation Environment. *International Journal of Human-Computer Interaction* 28, 6 (June 2011), 383–398. DOI: http://dx.doi.org/10.1080/10447318.2011.595621