Coding Like a Data Miner: A Sandbox Approach to Computing-Based Data Science for High School Student Learning

Justice T. Walker
Department of Education
University of Texas at El Paso
El Paso, Texas
jtwalker@utep.edu

Amanda Barany
Graduate School of Education
University of Pennsylvania
Philadelphia, Pennsylvania
amanda.barany@gmail.com

Alex Acquah
Department of Education
University of Texas at El Paso
El Paso, Texas
aacquah@miners.utep.edu

Sayed Mohsin Reza
Department of Computer Science
Pennsylvania State University
Harrisburg
Middletown, Pennsylvania
skr6024@psu.edu

Alan Barrera
Department of Education
University of Texas at El Paso
El Paso, Texas
abarrera6@miners.utep.edu

Karen Del Rio Guzman
Department of Education
University of Texas at El Paso
El Paso, Texas
kadelrioguzman@miners.utep.edu

Michael A. Johnson
Department of Learning Technologies
University of North Texas
Denton, Texas
michaeljohnson12cbc1@my.unt.edu

2023 IEEE Frontiers in Education Conference (FIE) | 979-8-3503-3642-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/FIE58773.2023.10343283

***Abstract*—Personal health tracking devices and internet-based digital platforms with the capacity to collect, aggregate, and store data at massive scales are examples of tools that have broadened priorities in computing to include data science. In response, there has been growing attention in research and practice emphasizing pre-college groups. This is partly because of the growing recognition—reflected in initiatives like CS4ALL, Code.org, Bootstrap: Data Science, Exploring Computer Science—that learning experiences before college are consequential in sustaining a robust pipeline of computer scientists and engineers.**

Despite these inroads, there is justifiable concern that existing efforts might not fully support learner development in the necessary conceptual, epistemological, and heuristic styles needed to productively parse and understand "big data." This is because computing-based curricula that include data science often involve data curated by others (rather than learners directly), which results in simulated versions of practice instead of engagement that is realistically discursive and messy. This is further complicated by the persistent shortage of K-12 computer science teachers in general and even fewer who can design and implement curricula that support authentic engagement with data science.

To address these issues, we leverage culturally relevant and constructionist perspectives in a sandbox (i.e., open-ended) science where tools like Scratch and electronic textiles (E-textiles) have had success expanding possibilities in computing to also include activities where learners can engage broadly along varied pursuits—and encounter challenges that spur computational thinking and problem-solving. The literature suggests that learning activities framed in this way encourage knowledge construction, practice literacies, and seriously impact learner attitudes, interest, and perceptions of growth in the field. This latter set of self-concept measures represents a few of many related key predictors of long-term field participation and persistence.

In this work-in-progress scholarship of discovery research, we co-develop, with youth and educators, "Coding Like a Data Miner" (CLDM)—a sandbox approach to computing-based data science wherein learners access a social media platform, Twitter, to mine, analyze, and understand quantitative and qualitative data sources. In this preliminary work, we assess affordances in co-developing a curriculum that leverages sandbox approaches to data science. Ultimately (and what will be presented in our final submission), we aim to study learning outcomes when high school students' access, analyze and make sense of "big data" sets of their own. We collaborated with high school teachers in a West Texas/Paso Del Norte region where computer science educators are exceptionally scarce and where there is an urgent and persistent need to support underrepresented learner access to burgeoning areas of computing. Using mixed-methodological approaches (e.g., quantitative analysis of learner pre- and post-survey responses along with qualitative assessments of semi-structured interview data), we address the following research questions: (1) What affordances exist using co-design approaches to develop sandbox data science for pre-college learners? (2) Which computational concepts do students learn when carrying out CLDM activities, (3) Which computational practices do high school students enact when mining, processing, and analyzing big data sets in CLDM? (4) How do learner knowledge and perceptions about data science shift after participating in CLDM? We use contemporary perspectives in computing education, constructionism, and equity to discuss how open-ended sandbox approaches to computing-based data science support learner computational thinking, practice literacies, and field perceptions.

Keywords*—*Computer Science Education, Data Science Education, Constructionism, Curriculum Design, Computer Science Learning

I. Introduction

Digital technologies—embedded with automated data collection capabilities—have transformed the nature of data—shifting it from small data sets that can be readily processed and analyzed using traditional methods to massive compilations of information that can only be handled using computing. Concomitant with these developments is also a change in knowledge and skills needed to with and understand these data alongside their real-world applications [1]. This has ignited calls for a reevaluation of how data science might fit in pre-college

education [2], including concerns about how I might deliver data science in ways that "count" as well as ways the field may be enacted in service of productive learning [3]. Computer science education (CSE) has shown viability to support pre-college data science education [4] because CSE provides opportunities for learners to conduct data science through authentic acts [5] whereby learners use computer code to harness data-generating technologies in order to access, collect and make sense of large amounts of information (e.g., Bootstrap: Data Science, Exploring Computer Science, etc.) that is tied to everyday contexts.

In this article, we preliminarily report on the participatory approach we took to involve both educators and learners in the process of co-designing a culturally relevant data science based CSE curriculum that accesses data drawn from Twitter (research phase one of this study to address research question one). The aim of this approach is to address the equity issues that emerge in curricular designs that are designed exclusively by others or are inherently restrictive in topic area and scope. Our stance is that a co-design approach in our West Texas/El Paso del Norte context—a binational region that is a social and cultural watershed where competing public needs, values, and priorities converge—will inform a growing body of literature that addresses the complex and often inter-related tensions that arise when diverse students direct the helm in data science-based CSE (where inquiries can be expansive). This context provides a frame to examine how key stakeholders respond to methods for supporting inclusive learning design. Co-design generally refers to the collaborative process of designing something with the active involvement and contributions of multiple stakeholders. Our co-design approach includes a collaborative process of designing a CSE curriculum with the active involvement and contributions of the following stakeholders: high school teachers, students, and university researchers.

In phase two of this study (to address research questions two to four), we will pilot the curriculum. Using interview data and pre/post survey responses we are conducting mixed-methodological exploratory analysis to address the following research questions: (1) What unique learning affordances and constraints exist using co-design approaches to develop sandbox data science for pre-college learners? (2) Which computational concepts do students learn when carrying out CLDM activities, (3) Which computational practices do high school students enact when mining, processing, and analyzing big data sets in CLDM? (4) How do learner knowledge and perceptions about data science shift after participating in CLDM? Preliminary findings of the first phase of our research project (co-designing a sandbox data science curriculum) suggest that participants had similar perspectives of benefits related to their online co-design and learning experiences. The most prominent reports were stated as experiences of educational freedom, pragmatism, and management versatility of the co-design process—these shaped learning designs and learner experiences when engaged in sandbox data science. In phase two of our project, we plan to address learning outcomes (see research questions 2-4). We discussed these findings in relation to literature centered on participatory co-design alongside equity approaches to data science based CSE curriculum development and implementation.

II. Background

A. Pre College Data Science Education

While the literature suggests data science is viable in any number of academic subjects, computer science classes have gained research and practitioner attention as a valuable venue for data science work, in part because the development of automated data collection technologies that generate "big data" [6] can be readily managed and processed in computing. Data science in CSE can be characterized as a field that leverages computer code to extract, wrangle (i.e., clean up), analyze and represent data for use in a wide range of fields. While data science based CSE is not a novel concept in university settings, there are significant efforts to examine how these areas might be deployed in pre-college contexts to support formal learning, interdisciplinary engagement, and student perceptions (e.g., public perception, etc.). These efforts have drawn rightful attention to learning designs and the ways they overcome or reify equity issues that persistently challenge computing education—including who gets to participate and how [7]. Along this frame, some have argued for a more humanistic stance [8] not only to support authentic and situated practice but also to highlight the inextricable links (and tensions) that exist between data and society. Recent attention in pre-college settings—due in part because middle and high school settings represent critical field pipeline points where learners are making decisions about longer-term academic and career interests—has raised concerns about how CSE (with or without data science) can responsibly and suitably support social and cultural student engagement and interests [9-10]—including when data sets are not defined by students themselves. This research addresses these issues by using culturally relevant and responsive pedagogies (CRPs) as a design principle to guide curricular products.

The initial goal of curriculum design was to position students as data scientists engaging authentically with data science skills, knowledge, identities, values and epistemologies in ways that are self-directed, collaborative, and multi-modal. This aligns with recent calls for data science-based computing education that leverages humanistic approaches [8] to connect CS processes to real-world issues and contexts. Through co-design with youth and educators, a series of typical steps taken as part of data science praxis were identified and modeled through guided activities centered around Twitter data that could be aligned with student and teacher needs and interests. The first step is characterized as Data Gathering, which includes identifying the appropriate tools (e.g., Application Processing Interfaces), choosing Twitter hashtags or keywords related to a topic of interest, and using those hashtags to mine a sample of data from Twitter. The second step is Data Pre-processing, which involves checking, cleaning, organizing, and analyzing sampled Twitter data. Data Analysis was designed with flexibility in terms of the complexity level of statistical analyses introduced to best meet the needs of different learning contexts. Finally, students using the curriculum could engage in Data Visualization, using existing tools to create different models of patterns (e.g., pie chart, word cloud) in their chosen datasets. Curricular design guides students through each of these four steps repeatedly with different degrees of scaffolding to support their authentic use of data mining tools and procedures and the

exploration of questions and data they find personally relevant and meaningful from a humanistic perspective. The full curriculum can be viewed at https://www.cs.utep.edu/DataMiner/explore_curriculum.html.

As students repeatedly engage with our sandbox data science activities, the arc of learning progresses from structured guided practice to more open-ended free inquiry. The aim of this design model was to provide foundational information and skill development—that we will eventually assess—so that students with an emerging understanding of data science and computer science practices might have access to needed supports, all while gradually shifting students to practices of greater autonomy in terms of data collection, pre–processing, data analysis, and data visualization decisions related to their topics of interest. Building on the success of existing research on constructionist approaches in CS education [11-12], four inquiry-based modules also leveraged constructionist design (where learners build their own code to show what they've learned) to emphasize real-world, project-based activities intended to empower learners to use their emerging awareness of CS practices to further develop their knowledge and skills.

B. SandBox Data Science: A Culturally Responsive Approach

Culturally relevant [13] and responsive [14] pedagogies (CRPs) are asset-based perspectives that have had transformative impacts on STEM education to constructively center diverse learners [15-16], and the social and cultural assets they bring to learning experiences. This occurs because learning experiences start with topics that are relevant to students and affirm their lived, social, cultural, and linguistic experiences. CRPs have also had notable impacts in shaping discourse, research and practice in CSE [17] and while these inroads represent important steps toward mitigating persistent equity issues in CSE as a whole—far less progress has been achieved in burgeoning areas of CSE, such as in areas of data science. This research is an early step in developing CRPs for data science CE and assessing the associated learning outcomes that result using this approach. We accomplish this by leveraging application programming interfaces (APIs) that make it possible to access rich collections of data at large scales. Combined with public access to data sets drawn from social media platforms (e.g., Twitter, YouTube, Facebook, Instagram, etc.)—the potential to empower learners as producers of knowledge, rather than consumers, is vast since data can be accessed and like a sandbox explored by learners themselves, and on their own terms. This work explores this possibility.

III. Methods

A. Phase One: Participants and Study Context

This ongoing project is designed to be implemented in two phases: the first involved curriculum development through co-design to examine the various affordances associated with sandbox-framed data science (to address research question one and reported in this work in progress). The second will be an implementation pilot to assess learning outcomes (to address research questions two through four). This exploratory project used a participatory design (i.e., curricular co-design) [18]. We collaborated with youth and educators from underrepresented groups to co-construct the curriculum in ways that reflect relevant needs and practices. Co-design sessions were carried out online using Zoom, an online communication platform. Sessions included a total of 14 participants (excluding study researchers) comprising five educators and nine youth, of which eight identified as male and six identified as female. Sessions were held on Zoom and were enhanced by a variety of features such as text-based synchronous chat forums, annotation tools, file and screen sharing functions and breakout rooms. The sessions were augmented with supplementary resources that included Google tools (e.g., co-laboratory, sheets, slides, and docs) to support participant collaboration. Sessions were held over the course of two weeks (consisting of five days per week and each week representing one phase of co-design). Session phases consisted of feedback and prototyping. Phase one involved participants conducting critiques of starter curriculum materials and participants were grouped homogeneously by age category (i.e., youth or adult) when aspects of sessions warranted topical expertise in specific areas (e.g., adults collaborated on standards alignments while youth worked on activity design)—all other session grouping were heterogeneous and consisted of both youth and adults. Phase two involved participants prototyping new versions of curriculum materials to reflect feedback generated in the prior phase. This process included four activities: (1) curriculum map and standard alignment critiques, (2) starter slide deck critiques, (3) activity design critiques, and (4) performance task critiques. Participant groups collaborated daily over the course of the week to discuss and generate critiques about these aspects of the starter curriculum, and this culminated in a final presentation wherein participants summarized their ideas. Phase two consisted of the same sessions (minus curriculum mapping) and participant groups collaborated daily over the course of the week to generate redesigned prototypes of the curriculum.

B. Phase One: Data Sources and Analysis

Authors #4 and #7 conducted semi-structured interviews with eight of fourteen participants at the end of each co-design session phase and asked questions about group work dynamics, co-design session structure, and curriculum in comparison to their prior educational experiences. Interviews were then transcribed using Otter.ai—a transcription software—and then edited by author #4 to ensure readability. Transcripts were labeled using pseudonyms and then uploaded to Dedoose—an online qualitative coding tool—and analyzed using inductive approaches [19]. Codes were developed iteratively (through collective dialogic engagement between authors) based on themes observed in interview transcripts. Authors #2 and #3 coded each of the interviews separately and reconciled any differences by reviewing the coding together. Any remaining disagreements were then reconciled (to 100% agreement) together with authors #1 and #2.

IV. Preliminary Results

We organize results around two descriptive themes that address the central exploratory research question: What did participants have to say about the process of co-constructing an equity-driven data science-based CSE curriculum? These themes include what participants had to say about (i) the co-design implementation structure and (ii) reflections on

experiences engaging in the online co-design process. Findings are reported in the next sections.

Understanding participant reactions to online co-design implementation structures was important to us because affordances provide good insights about the extent to which participants viewed the tools and arrangements designed in the process as restrictive, or not. In other words, it provides us with an initial opportunity to assess whether online co-design is a viable approach to enact equity-driven curriculum development. When examining reactions to implementation structures, we found that participants noted a wide range of technical features that supported their involvement (curricular task management practicality and versatility)—and mostly enabled through identified platform-specific features that supported resource sharing (e.g., screen shares, file sharing, etc.), varied interpersonal interactions (relationship building, collaboration, discussions, dis/engagement, etc.), and artifact construction. We observed that these affordances were supported across co-design session curriculum content areas (e.g., slide deck revisions, performance task designs, etc.). We also observed that participants mostly discussed these affordances in relation to their interactions with each other and across levels of relative expertise (e.g., experienced, inexperienced educator, youth, adult, etc.). For instance, when asked about the implementation structure, Sonja explained:

I liked it. Um, I think, with being on an online platform allowed for a lot more flexibility. And the way I know it was very convenient for me, because I don't think I would have been able to drive out somewhere, to be able to participate in it. And so because he was online, I was able to participate in the codesign, which I'm grateful for. And then I think it was easier to share resources. (Sonja Montana, 07/06/2022)

We see that Sonja points to several affordances that highlight how platform features made participation accessible in addition to features that enabled productivity (e.g., resource sharing). This is significant since equity-driven co-design can be understood as encompassing both issues related to developing curriculum but also in identifying ways to enlist diverse voices—which in this case involved overcoming logistical constraints. Davian also pointed to platform-specific features and interpersonal interactions as an affordance as he explained:

The fact that we could really easily, like share screens and kind of see the documents as they're being edited…it felt like there was more like streamline discussion, because for instance, right now, I'm the one talking you know, and that's pretty clear is like, I guess you can do that in person too, of course, you're supposed to do that. But it just felt like less cluttered where it was like: Okay, we're looking at one screen where we're hearing one person. (Davian Zubek, 07/08/2022)

Davian is highlighting another accessibility feature which in this case was the use of screen sharing tools (in Zoom) and tools that support synchronous collaboration (e.g., Miro board and Google Docs) that enabled him to work with others constructively as they provided critique to curricular starter materials and developed curricular revisions along the way. In this case, we observed that supporting robust interactions and information sharing translated, for Davian and others, enhanced interpersonal interactions, which in this case meant close collaborations to manage tasks and discussion. In this case, access to these tools enhanced the depth of engagement. We observed in several participant responses that enhanced accessibility provided inherent freedom to engage in diverse ways—to overcome access challenges, to design collaboration structures, and to engage deeply in the work of generating curriculum resources and we attribute these perspectives to platform-specific features inherent to co-design implementation structures.

Our explorations also took stock of what participants had to say about their experiences co-designing an equity-driven data science CE curriculum. This is important in our study because while we understand technical structures can shape what individuals can do in an online environment (in this case online co-design), how experiences unfold can be quite different—one can engage in collaboration supported by digital tools, but the quality of that engagement can vary significantly. In our assessment, we learned that participants frequently pointed to social aspects of their experiences—highlighting the development of interpersonal relationships that not only supported productivity (through versatile task management and pragmatic task execution strategies), but also feelings of legitimacy, efficacy, and—ultimately—freedom to engage and explore ideas freely. In other words, relationships in the co-design sessions were valuable not only in supporting group productivity but also in reaffirming each other's contributions. This was evident in Alfonso's response as he reflected aloud:

I think the dynamic was fine, you know, for a starter group. We obviously built a relationship over the days, but that's still not as deep or as profound of a bond that you would have with students over time as well. So, it was still effective and that's how our relationships are built just slowly. I think we were very effective in the time we had. We trusted each other and respected each other's opinions. (Alfonso Oliver, 07/08/2022)

Alfonso pointed directly to the connection he perceives between group relationships and respect for opinions. It is through this intellectual deference that participants across our interviews found valuable in affirming their legitimacy, efficacy, and freedom to shape the curriculum. He also explained to us why he thought his group in the session was successful, "you know, I guess societally, socially, people do take on roles naturally. However, they're usually allocated by rank. And I felt…[the] confidence factor for each participant was really successful [because] people could choose to take on roles. So, I think it was so effective (Alfonso Oliver, 07/08/2022)." Alfonso went on to explain that very often youth are designated roles in learning (which can be typical cooperative learning arrangements)—and then they enact those roles in service of work outcomes. Alfonso compared that to the structures he perceived in co-design which included opportunities for participants (and youth) themselves to choose their roles and involvement. Alfonso viewed this as very different and, in our observation, was what undergirded his perceptions of freedom in the process. Others went on to explain how feedback and diverse perspectives enhanced their work—and our examination of these accounts suggests that affirming mutual feelings of legitimacy and choice were key in spurring intellectual freedom and efficacy. Together these insights about implementation structures and participant

experiences suggest that co-design intervention designs and interpersonal social relationships are linked and should be considered together when seeking to develop equity-driven curricula using the voices, perspectives and cultural assets of the communities where these tools are meant to serve.

V. Preliminary Discussion

Our findings suggest that developing pre-college humanistic [5]) or culturally relevant/responsive (13-14] curricular design in nascent areas such as data science-based CE requires that co-design sessions [18] as a means of accomplishing this, should be structured in ways that promote access to key stakeholders—namely those themselves who represent or serve learners underrepresented in the field. In this research, this meant not only leveraging technologies that provided a means to overcome access barriers but also tools that enabled flexible forms of engagement. We learned from participant accounts that this can have meaningful impacts on productivity as participants could generate myriad approaches to evaluating and critiquing curricular materials and managing progress in generating curricular outputs that reflect the varying needs, values, and interests of the learners that learning designs are meant to support [17]. In data science-based CE, this is especially relevant since there are comparatively fewer benchmarks and existing frameworks on which to structure or drive participant engagement—and fewer case examples or starter materials from which to draw for building humanistic or culturally relevant/responsive curricula (e.g., [8]). Creating access to stakeholders enables the participation necessary to bring together the requisite intellectual diversity necessary to co-construct a curriculum that is sufficiently relevant to support learners who are typically not reached in designs with narrower and more generalized scopes. We learned, through our analysis, from participants that such an approach brought individuals together, and deepened their ability to engage constructively, freely and collaboratively. Phase two of this work-in-progress project will examine learning outcomes (research questions two to four).

Acknowledgment

This work was supported in part by a grant from the National Science Foundation (#2137708). We would also like to thank Crystal Fuentes, Jesus Oropeza, and Omar Badreddin for their contributions to curriculum development. Any opinions, findings, conclusions, and/or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Science Foundation or the University of Texas at El Paso.

References

[1] Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165-181.

[2] National Academies of Sciences, Engineering, and Medicine. (2018). Data science for undergraduates: Opportunities and options. Washington, DC: The National Academies Press.

[3] Jiang, S., Lee, V. R., & Rosenberg, J. M. (2022). Data science education across the disciplines: Underexamined opportunities for K-12 innovation. *British Journal of Educational Technology, 53*(5), 1073-1079.

[4] Krishnamurthi, S., Schanzer, E., Politz, J. G., Lerner, B. S., Fisler, K., & Dooman, S. (2020). Data science as a route to AI for middle-and high-school students. arXiv preprint arXiv:2005.01794. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. *Educational Researcher, 50*(9), 664-672.

[6] Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology, 50*(1), 101-113.

[7] Madkins, T. C., Howard, N. R., & Freed, N. (2020). Engaging equity pedagogies in computer science learning environments. *Journal of Computer Science Integration, 3*(2).

[8] Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. *Educational Researcher, 50*(9), 664-672.

[9] Blikstein, P., & Moghadam, S. H. (2018). *Pre-college computer science education: A survey of the field*. Google LLC. https://goo.gl/gmS1Vm.

[10] Tissenbaum, M. & Weintrop, D., Holbert, N., & Clegg, T. (2021). The case for alternative endpoints in computing education. *British Journal of Educational Technology, 52* (3), 1164-1177. https://doi.org/10.1111/bjet.13072.

[11] Fields, D. A., Kafai, Y. B., Morales‐Navarro, L., & Walker, J. T. (2021). Debugging by design: A constructionist approach to high school students' crafting and coding of electronic textiles as failure artefacts. British Journal of Educational Technology, 52(3), 1078-1092.

[12] Fields, D., Lui, D., Kafai, Y., Jayathirtha, G., Walker, J., & Shaw, M. (2021). Communicating about computational thinking: understanding affordances of portfolios for assessing high school students' computational thinking and participation practices. Computer Science Education, 31(2), 224-258.

[13] Ladson-Billings, G. (2008). Yes, but how do we do it?": Practicing culturally relevant pedagogy. *City kids, city schools: More reports from the front row*, 162-177.

[14] Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.

[15] Brown, B. A., Boda, P., Lemmi, C., & Monroe, X. (2019). Moving culturally relevant pedagogy from theory to practice: Exploring teachers' application of culturally relevant education in science and mathematics. *Urban Education, 54*(6), 775-803.

[16] Enyedy, N., & Mukhopadhyay, S. (2007). They Don't Show Nothing I Didn't Know: Emergent tensions between culturally relevant pedagogy and mathematics pedagogy. *Journal of the Learning Sciences, 2*, 139-174.

[17] Madkins, T. C., Martin, A., Ryoo, J., Scott, K. A., Goode, J., Scott, A., & McAlear, F. (2019, February). Culturally relevant computer science pedagogy: From theory to practice. In 2019 research on equity and sustained participation in engineering, computing, and technology (RESPECT) (pp. 1-4). IEEE. Mensah, F. M. (2021). Culturally relevant and culturally responsive. *Science and Children, 58*(4), 10- 13.

[18] Penuel, W. R. (2019). Co-design as infrastructuring with attention to power: Building collective capacity for equitable teaching and learning through design-based implementation research. In *Collaborative curriculum design for sustainable innovation and teacher learning* (pp. 387-401). Springer, Cham.

[19] Ravitch, S. M., & Carl, N. M. (2019). *Qualitative research: Bridging the conceptual, theoretical, and methodological*. Sage Publications.