"Data comes from the real world": A Constructionist Approach to Mainstreaming K12 Data Science Education

"Data comes from the real world": A Constructionist Approach to Mainstreaming K12 Data Science Education

Prerna Ravi
Massachusetts Institute of Technology
Cambridge, Massachusetts
prernar@mit.edu

Robert Parks
Massachusetts Institute of Technology
Cambridge, Massachusetts
rparks@mit.edu

John Masla
Massachusetts Institute of Technology
Cambridge, Massachusetts
j_masla@education.mit.edu

Hal Abelson
Massachusetts Institute of Technology
Cambridge, Massachusetts
hal@mit.edu

Cynthia Breazeal
Massachusetts Institute of Technology
Cambridge, Massachusetts
cynthiab@media.mit.edu

Abstract

Data science is emerging as a crucial 21st-century competence, influencing professional practices from citing evidence when advocating for social change to developing artificial intelligence (AI) models. For middle and high school students, data science can put formerly decontextualized subjects into real-world scenarios. Many existing curricula, however, lack authenticity and personal relevance for students. A critique of data science courseware cites the lack of "author proximity," in which students do not contribute to the data's production or see their personal experiences reflected in the data. This paper introduces a novel data science curriculum to scaffold middle and high school students in undertaking real-world data science practices. Through project-based learning modules, the curriculum engages students in investigating solutions to community-based problems through visualization and analysis of live sensor data and public data sets. Materials include formative assessments to help educators (especially those from non-math and computing backgrounds) measure their students' abilities to identify statistical patterns, critically evaluate data biases, and make predictions. As we pilot and co-design with teachers, we will look closely at whether the curriculum's resources can successfully support non-technical practitioners engaging in an integrated curriculum.

CCS Concepts

• **Applied computing → Interactive learning environments**; • **Social and professional topics → Computational thinking**.

Keywords

K12 Data science, Sensors, Computational action, Project-based learning, Participatory data collection

ACM Reference Format:
Prerna Ravi, Robert Parks, John Masla, Hal Abelson, and Cynthia Breazeal. 2024. "Data comes from the real world": A Constructionist Approach to Mainstreaming K12 Data Science Education. In *Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 1 (SIGCSE Virtual 2024), December 5–8, 2024, Virtual Event, NC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3649165.3703623

SIGCSE Virtual 2024, December 5–8, 2024, Virtual Event, NC, USA

ACM ISBN 979-8-4007-0598-4/24/12
https://doi.org/10.1145/3649165.3703623

1 Introduction

Data science is emerging as a crucial 21st-century competence, influencing professional practices from citing evidence when advocating for social change to developing artificial intelligence (AI) models. By March 2024, ten states offered data science to students in grades 6-12, with an additional fifteen piloting curricula or setting standards [13]. Understanding the nuances of data science can form a foundation for navigating the capabilities of AI; data science and AI share competencies in understanding how personal data is used to train models and critically examining data with "skepticism and interpretation" [33]. School administrators typically motivate data science as a means for job readiness, social impact, and improved math outcomes [44]. However, according to a recent survey of high schoolers, the chief reasons to study data science are the abundance of data available and intellectual proficiency with data, with employment prospects a distant third place [27]. For middle and high school students, data science can put formerly decontextualized subjects such as math and statistics into real-world scenarios. Many existing curricula, however, lack authenticity and personal relevance for students. A critique of existing data science courseware cites the lack of "author proximity": students do not contribute to the data's production or see their personal experiences reflected in it [30]. An additional challenge is integrating data science as a formal subject into schools and supporting teachers with professional development and assessments [21].

This paper introduces a novel curriculum to scaffold middle and high school students in undertaking real-world data science practices. We intend to study how MIT App Inventor's mobile data science toolkit [12] could allow learners to visualize and analyze both sensor data and public data sets. Through project-based learning modules aligned with the Big Ideas in K-10 Data Science [39], the curriculum employs participatory data collection, allowing students to lead investigations on topics of personal interest, to foster higher authorship proximity to their data [2, 8, 28]. These modules also include adaptable formative assessments to help teachers (especially those from non-math and computing backgrounds) measure students' abilities to identify statistical patterns, critically evaluate data biases, and make predictions.

2 Background

2.1 Data Sources and Learning Impact

Much of the scholarship on recent data science curricula for schools categorizes courseware according to the provenance of its data with implications on learning goals, student engagement, and opportunities for critical inquiry [16, 17, 30]. Datasets can originate from learner-collected, fictional, or publicly available data, allowing multiple opportunities to build learner competencies and drive motivations [17]. Collecting sensor data can help students engage meaningfully in data practices, explore statistical patterns, and make inferences based on their knowledge of the data context [16, 30, 31, 35, 40]. With publicly available data, students can experience how data is used in the workforce and scientific practices [11, 30]. Rubin calls for students to develop the skills of "data journalists," understanding and interpreting others' data by becoming familiar with the domain, the measurement process, potential biases, and scientific limitations in producing that data [43]. In the case of fictional or public data, however, researchers warn that materials disconnected from contexts will fail to engage students and fail to meet crucial learning goals such as drawing conclusions and making predictions based on real-world connections [21]. An additional constraint is the cost of materials: sensor-based data science curricula have recently launched [17, 20, 29, 48, 53], but rely on expensive consumer wearables or "probeware" — sensors made for the education market that must stay tethered to a computer.

2.2 Equity and Constructionism in Data Science

Prior work has established the need for datasets that actively engage students from historically underrepresented communities. High engagement and task persistence are linked to student work on personally meaningful topics, a core idea of constructionism [6, 41]. Student-led work prompts youths' conceptions of data and its limitations when creating meaningful data artifacts within a social context [6, 18]. Recognizing that some data collection methods can be biased toward specific research goals or ideological agendas is essential for critically reflecting on the data's origins [24]. Additionally, Jiang et. al. suggest that data science practiced across disciplines validates multiple forms of participation and supports epistemological pluralism [23, 46]. Cultivating data literacy for people in non-technical fields also forms an avenue for increasing equity in learning activities [8]. This approach ensures that activities make sense within broader social contexts, empowering students to use data to advocate for change [14, 47]. By allowing learners to decide what questions to ask with data and whether the necessary data has been collected, students can better engage with real-world activities, bringing their lived experience and prior knowledge to the classroom [9, 49]. Dangol & Dasgupta, however, underscore the need for more research on supporting teachers in implementing constructionist approaches to teach data literacy [6].

3 Curriculum

The curriculum utilizes a hybrid approach with both learner generated and public data, allowing students to engage in sensemaking about short- and long-term trends. The curriculum provides flexibility for different learning goals, using low-dimensional data to introduce concepts, messy datasets to demonstrate issues of bias, and personally relevant datasets to deepen engagement [8, 33, 45]. By employing an abstracted, block-based programming environment within MIT App Inventor's data science toolkit [12], the curriculum can lower barriers for non-technical students and facilitate professional development for teachers. Enabling students to create custom mobile apps for data collection and analysis also improves accessibility in low-resource contexts where mobile phones are more prevalent [42]. The curriculum incorporates learner-generated environmental data collected with micro:bit sensors, an approach not widely used in previous work. We provide opportunities for students to explore the capabilities and limitations of sensors, which are essential for understanding how AI devices gather data and interact with the world [33]. The curriculum has a significant focus on data cleaning, an activity that occupies up to 80% of a data scientist's time [15, 28], but is often missing from contemporary student resources (except in YouCubed and scant other materials) [37, 52]. Data cleaning activities are context-dependent, inviting students to "dig into the circumstances surrounding data collection" [43] to identify and address data anomalies and uncertainties [4, 28].

3.1 Teaching and Learning Materials

This curriculum includes educator guides, student resources, and assessment modules for teaching data science practices aligned with the Big 10 Ideas of Data Science [39]. The materials, openly accessible on appinventor.mit.edu, feature structured activities, teaching slides and scripts, and tool guides. Each project team (3-4 students) needs one laptop, an iOS/Android phone/tablet supplied by the student or school, and a micro:bit sensor ($17-22 each). The target audience includes K-12 middle and high school educators (including curriculum designers, formal and informal teachers, and school districts) and students. The lessons described below show an example of using environmental data, split across two modules. Each lesson takes 60-90 minutes to complete.

3.1.1 ***Module 1:*** This module aims to educate students on the fundamentals of collecting, analyzing, and visualizing sensor data curated from the environment. It provides students with a framework needed to plan investigations for community challenges using IoT sensors and to prepare them to share the evidence obtained.

Lesson 1: Hands-on experiments with sensors: This lesson teaches students to connect sensors to a mobile device and visualize the data. It starts with unplugged activities demonstrating sensor functionality, followed by instructions to connect sensors to students' custom mobile apps (created with App Inventor) via BluetoothLE for real-time data visualization, and concludes with a game for identifying sensor types as they correlate data outputs with changes in the physical environment.

Lesson 2: Brainstorming sensor use cases: Students identify sensors ubiquitous in their environment and imagine their creative uses, enhancing their understanding of how sensor technology gathers important data. The lesson includes interactive activities such as mapping a typical school day with sensor applications and using Slow Reveal Graphs (an instructional routine to promote sensemaking of environmental visualizations) [25].

Lesson 3: Project ideation: Students form teams to pursue their project ideas, focusing on local environmental issues in their community. They brainstorm themes (e.g., air quality, water, sanitation, etc.) using card games, vote on their favorite ideas, and formulate research questions. After scouting sensor locations around their school surroundings, they test for proper sensor placement, data quality, and collection timelines to answer their research question, then set up sensors to save data automatically to Google Sheets.

Lesson 4: Building data applications: Students use their apps to import datasets, experiment with different graph types and scales, and apply these principles to their real-time sensor time series data, enhancing their understanding of data visualizations.

Lesson 5: Visualizing final sensor data: Students visualize their group's collected sensor data to analyze trends relevant to their original research question. They customize their visualizations and present their final projects to the community, reflecting on their accomplishments, challenges, and directions for future inquiries.

3.1.2 ***Module 2:*** In this module, students select a public dataset related to their Module 1 sensor data. They review long-term curated datasets and their contexts, exploring visualizations, predictions, and inference through coding activities and scaffolded discussions.

Lesson 6: Visualizing a data set: Students discuss trends in the sensor data gathered, linking them to broader environmental and climate change issues. They validate their small data collection by selecting curated long-term public datasets for further analysis. Students review spreadsheet features, identify unusual data points, and program their team's app to visualize and explore possible correlations between data series.

Lesson 7: Modeling data: Students start with an unplugged activity to understand the concept of a line of best fit by visually fitting lines to sample data points. Teachers use guided prompts to discuss the value of models for trends, predictions, and confidence levels. Student teams then add a line of best fit to their app visualizations and discuss non-linear models, the slope-intercept form, and the correlation coefficient, tying these to their sensor data and potential long-term data collection.

Lesson 8: Cleaning data: Teachers use lesson prompts to discuss the relevance of anomalies. Student teams distinguish between in-context and out-of-context anomalies in their public data graphs, code their apps to detect and remove selective anomalies, and evaluate the updated trend line. They then apply these concepts to their sensor data, comparing emerging trends against the public dataset.

Lesson 9: Predictions and AI analysis with data: Students identify trends in their public data, use the slope to predict future values, and extend their graph's domain in the app. They program a generative AI chatbot within App Inventor to provide additional context and analysis. Students examine confounding variables (location, human judgment, and organizational ethics) in their personal and public datasets, recognizing how these can skew results.

3.2 Assessments

Practitioners of project-based learning have noted its difficulty in assessing a wide range of cognitive, interpersonal, and intrapersonal competencies involved [5, 19]. While standardized tests focus on lower-order thinking skills, our curriculum targets higher-order thinking, such as conceptual statistical understanding, outlined in GAISE II [1]. We also aim to foster positive attitudes toward data science, including perceived competence, enjoyment, and value, drawing from the Intrinsic Motivation Inventory [32]. To measure conceptual growth, we integrate open-response questions related to the four Big Ideas in K-10 Data Science: (1) formulate statistical investigative questions, (2) collect/consider data, (3) analyze data, (4) interpret and communicate data [39]. We use short, formative "exit tickets" at the end of each lesson for student reflection and self-assessment of skills covered, guiding them through statistical reasoning over time [1, 5, 7, 26]. We also base some of our questions on the LOCUS project's assessments, aligning with Common Core and GAISE II standards [38]. This approach helps track learning trajectories, informs teacher instruction, and provides consistent, daily feedback to reinforce student learning [36].

4 Discussion and Future Work

In this paper, we present a novel data science curriculum enabling students to become data readers, communicators, and makers [50, 51]. This is unlike typical sensor-based laboratory investigations in which students carry out procedures without acting as agents in producing and using data [17]. We support a scientific data collection process that serves students' personal, cultural, or sociopolitical goals [28] to mimic real-world practices [10, 23]. While we acknowledge that some students may not initially show interest in environmental data [21], engagement can increase when they reflect on direct community impacts such as heat islands and flooding. Linking broad issues like climate change to students' experiences can enhance resonance [30, 34]. Integrating data science with commonly taught subjects broadens its utility and opens interdisciplinary possibilities, making it more relevant to teachers and students [22]. While students may implicitly engage with ethical considerations when contextualizing and cleaning data, the curriculum currently lacks specific support for broader ethical discussions, including data use in AI. Furthermore, while the curriculum touches on the data pipeline, it does not yet include machine learning activities, which needs further exploration [33].

Several tensions highlight opportunities for future work, such as balancing student-driven data collection with the need for teacher preparation and classroom time. Finding manageable open data for students is challenging, but curriculum scaffolding can assist with data cleaning and preparation [3]. We must consider the size and messiness of curated data to maintain authentic experiences without overwhelming students or teachers. We will continue testing ways to support problem-definition routines that connect students to their interests and community issues. Lastly, there is a tension between using authentic tools connected to professional practice (such as Python and R) and more accessible tools for computational thinking (such as MIT App Inventor). While we build on D'Ignazio's advocacy for data literacy pathways in non-technical fields [8], our goal is fostering computational thinking, cross-cutting conceptual understanding, habits of mind, and processes, rather than job preparation. Our future efforts will focus on co-designing the curriculum with teachers and investigating various student data collection methods from different disciplines (e.g., community surveys, social media, personal data logs) to ensure that the resources can support non-technical practitioners.

References

[1] Anna Bargagliotti. 2020. Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II). American Statistical Association.

[2] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. 2006. Participatory sensing. *Presented at World Sensor Web Workshop, ACM SenSy, Boulder, CO* (2006).

[3] Andrea Bussani and Cinzia Comici. 2023. Thermal tide detection: A case study to introduce open data analysis in high school. *The Physics Teacher* 61, 1 (2023), 68–70.

[4] Clark A Chinn and William F Brewer. 1993. The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of educational research* 63, 1 (1993), 1–49.

[5] Barbara Condliffe. 2017. Project-Based Learning: A Literature Review. Working Paper. *MDRC* (2017).

[6] Aayushi Dangol and Sayamindu Dasgupta. 2023. Constructionist approaches to critical data literacy: A review. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 112–123.

[7] Linda Darling-Hammond, Brigid Barron, P David Pearson, Alan H Schoenfeld, Elizabeth K Stage, Timothy D Zimmerman, Gina N Cervetti, and Jennifer L Tilson. 2015. *Powerful learning: What we know about teaching for understanding*. John Wiley & Sons.

[8] Catherine D'Ignazio. 2017. Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal* 23, 1 (2017), 6–18.

[9] Catherine D'Ignazio and Rahul Bhargava. 2018. Creative data literacy: A constructionist approach to teaching information visualization. (2018).

[10] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism*. MIT press.

[11] Richard Duschl. 2008. Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education* 32, 1 (2008), 268–291.

[12] Hanya Elhashemy, Robert Parks, David YJ Kim, Evan Patton, and Harold Abelson. 2023. Empowering Learners with a Low-Barrier Mobile Data Science Toolkit. *MIT App Inventor* (2023).

[13] Data Science for Everyone. N/A. Data Science 4 Evreyone (DS4E). https://www.datascience4everyone.org/.

[14] Paulo Freire. 2020. Pedagogy of the oppressed. In *Toward a sociology of education*. Routledge, 374–386.

[15] Armand Ruiz Gabernet and J Limburn. 2017. Breaking the 80/20 rule: How data catalogs transform data scientists' productivity. *IBM Cloud Blog* (2017).

[16] Engida Gebre. 2022. Conceptions and perspectives of data literacy in secondary education. *British Journal of Educational Technology* 53, 5 (2022), 1080–1095.

[17] Lisa Hardy, Colin Dixon, and Sherry Hsi. 2022. From data collectors to data producers: Shifting students' relationship to data. In *Situating Data Science*. Routledge, 104–126.

[18] Samantha Hautea, Sayamindu Dasgupta, and Benjamin Mako Hill. 2017. Youth perspectives on critical data literacies. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 919–930.

[19] Margaret L Hilton and James W Pellegrino. 2012. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

[20] Ayaz Hyder and Andrew A May. 2020. Translational data analytics in exposure science and environmental health: A citizen science approach with high school students. *Environmental Health* 19 (2020), 1–12.

[21] Rotem Israel-Fishelson, Peter Moon, Rachel Tabak, and David Weintrop. 2024. Understanding the data in K-12 data science. *Harvard Data Science Review* 6, 2 (2024).

[22] Shiyan Jiang and Jennifer Kahn. 2020. Data wrangling practices and collaborative interactions with aggregated data. *International journal of computer-supported collaborative learning* 15, 3 (2020), 257–281.

[23] Shiyan Jiang, Victor R Lee, and Joshua M Rosenberg. 2022. Data science education across the disciplines: Underexamined opportunities for K-12 innovation. , 1073–1079 pages.

[24] Britney Johnson, Ben Rydal Shapiro, Betsy DiSalvo, Annabel Rothschild, and Carl DiSalvo. 2021. Exploring approaches to data literacy through a critical race theory perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[25] Jenna Laib. N/A. Slow Reveal Graphs: An Instructional Routine to Promote Sensemaking about Data. https://slowrevealgraphs.com/.

[26] John Larmer, John R Mergendoller, and Suzie Boss. 2015. Gold standard PBL: Essential project design elements. *Buck Institute for Education* 2 (2015).

[27] Victor R Lee. 2024. Humanistic Pre-Service Data Science Teacher Education Across the Disciplines. In *Improving Equity in Data Science*. Routledge, 112–132.

[28] Victor R Lee and Victoria Delaney. 2022. Identifying the content, lesson structure, and data use within pre-collegiate data science curricula. *Journal of Science Education and Technology* 31, 1 (2022), 81–98.

[29] Victor R Lee, Joel Drake, Ryan Cain, and Jeffrey Thayne. 2021. Remembering what produced the data: Individual and social reconstruction in the context of a quantified self elementary data and statistics unit. *Cognition and Instruction* 39, 4 (2021), 367–408.

[30] Victor R Lee, Michelle Hoda Wilkerson, and Kathryn Lanouette. 2021. A call for a humanistic stance toward K–12 data science education. *Educational Researcher* 50, 9 (2021), 664–672.

[31] Richard Lehrer and Lyn English. 2018. Introducing children to modeling variability. *International handbook of research in statistics education* (2018), 229–260.

[32] Eow Yee Leng, Roselan Baki, Rosnaini Mahmud, et al. 2010. Stability of the Intrinsic Motivation Inventory (IMI) for the use of Malaysian form one students in ICT literacy class. *Eurasia Journal of Mathematics, Science and Technology Education* 6, 3 (2010), 215–226.

[33] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[34] M Lisette Lopez, Colette Roberto, Edward Rivero, Michelle Hoda Wilkerson, Michael Bakal, and Kris Gutiérrez. 2021. *Curricular reorganization in the third space: A case of consequential reasoning around data*.

[35] Katie Makar and Andee Rubin. 2009. A framework for thinking about informal statistical inference. *Statistics education research journal* 8, 1 (2009), 82–105.

[36] Audrey Martínez-Gudapakkam, Karen Mutch-Jones, and Jennifer Hicks. 2017. Formative Assessment Practices to Support Students who Struggle in Science. *Science and Children* 55, 2 (2017), 88.

[37] Peter F Moon, Rotem Israel-Fishelson, Rachel Tabak, and David Weintrop. 2023. The tools being used to introduce youth to data science. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 150–159.

[38] Levels of Conceptual Understanding in Statistics (LOCUS). 2013. Filling in the A in the GAISE Framework with Assessments for Statistics in the Common Core State Standards. https://locus.statisticseducation.org/.

[39] Stanford Graduate School of Education. N/A. Big Ideas for K10 Data Science. https://www.youcubed.org/data-big-ideas/#grade_8-10.

[40] Maxine Pfannkuch, Dani Ben-Zvi, and Stephanie Budgett. 2018. Innovations in statistical modeling to connect data, chance and context. *ZDM* 50 (2018), 1113–1123.

[41] Thomas M Philip, Sarah Schuler-Brown, and Winmar Way. 2013. A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning* 18 (2013), 103–120.

[42] Prerna Ravi, Azra Ismail, and Neha Kumar. 2021. The pandemic shift to remote learning under resource constraints. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[43] Andee Rubin. 2022. Learning to reason with data: How did we get here and what do we know? In *Situating Data Science*. Routledge, 154–164.

[44] Emmanuel Schanzer, Nancy Pfenning, Flannery Denny, Sam Dooman, Joe Gibbs Politz, Benjamin S Lerner, Kathi Fisler, and Shriram Krishnamurthi. 2022. Integrated data science for secondary schools: Design and assessment of a curriculum. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*. 22–28.

[45] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R Cooperstock. 2019. Can you teach me to machine learn?. In *Proceedings of the 50th ACM technical symposium on computer science education*. 948–954.

[46] Sherry Turkle and Seymour Papert. 1992. Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior* 11, 1 (1992), 3–33.

[47] Alan Tygel and Rosana Kirsch. 2015. Contributions of Paulo Freire for a critical data literacy. In *Proceedings of web science 2015 workshop on data literacy*. 318–34.

[48] Sarah Van Wart, Kathryn Lanouette, and Tapan S Parikh. 2022. Scripts and counterscripts in community-based data science: Participatory digital mapping and the pursuit of a third space. In *Situating Data Science*. Routledge, 127–153.

[49] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[50] Alyssa Friend Wise. 2022. Educating data scientists and data literate citizens for a new generation of data. In *Situating Data Science*. Routledge, 165–181.

[51] Annika Wolff, Daniel Gooch, Jose J Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2016. Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics* 12, 3 (2016).

[52] YouCubed. 2022. Explorations in data science. YouCubed High School Data Science Course. https://hsdatascience.youcubed.org/.

[53] James Yurkovich, Waseem Abbas, and Steve Yurkovich. 2023. Transcending the classroom using digital wearables. (2023).