;

A Secret Agent? K-12 Data Science Learning Through the Lens of Agency

Joshua M. Rosenberg, University of Tennessee, Knoxville, jmrosenberg@utk.edu
Ryan Seth Jones, Middle Tennessee State University, ryan.jones@mtsu.edu

Note. Both authors contributed equally. This is a commissioned paper for the National Academy of Sciences, Engineering, and Medicine *Foundations of Data Science for Students in Grades K-12: A Workshop*. The authors wish to thank Conrad Borchers, Amanda Garner, Spencer Greenhalgh, Matthew Koehler, Jeanne McClure, and Sarah Narvaiz for feedback on an earlier version of this paper.

This paper was commissioned for the *Workshop on Foundations of Data Science for Students in Grades K-12*, whose work was supported by the Valhalla Foundation. Opinions and statements included in the paper are solely those of the individual authors, and are not necessarily adopted, endorsed, or verified as accurate by the *Workshop Planning Committee on Foundations of Data Science for Students in Grades K-12*, the Board on Science Education, the Board on Mathematical Sciences and Analytics, the Computer Science and Telecommunications Board, or the National Academy of Sciences, Engineering, and Medicine.

In this paper, we attempt to summarize what we know about student learning of data science across diverse communities of scholars in the emerging field that is data science education. A summary of what we know could have value as knowing about and being able to do data science is increasingly important to gain access to educational and professional opportunities (National Academies of Sciences, Engineering, and Mathematics, 2018). Further, since data scientists play an influential role in shaping the world around us (for example, in terms of privacy, profit, and fairness; D'Ignazio & Klein, 2020), it is important for a societally representative group to have a voice in the conversation. Though possibly valuable, summarizing what we know is challenging for several reasons.

First, data science is at least an interdisciplinary if not a transdisciplinary endeavor (National Academies of Sciences, Engineering, and Mathematics, 2018) that involves a range of competencies—like programming and using statistical thinking—that have been traditionally taught and learned separately (Cao, 2017; Jiang et al., 2022; H. Lee et al., 2022; Tierney, 2012). Second, because research on data science learning is developing in different disciplines, the research is motivated by different goals. For example, the special issue of the *Journal of Statistics and Data Science Education* (Horton & Hardin, 2021) was focused on reflecting on a key article on data science teaching and learning at the undergraduate level, Nolan and Temple Lang's (2010) call for statistics educators to introduce computing concepts and computational skills to students learning statistics. While this paper was clearly influential on authors doing research on data science learning at the undergraduate level, none of the authors of the articles in the special issues on data science education in the *Journal of the Learning Sciences* (Wilkerson & Polman, 2020) and the *British Journal of Educational Technology* (Jiang et al., 2022); one article in the special issue of the *Statistics Education Research Journal* (Biehler et al., 2022) cited Nolan and Temple Lang's (2010) work. This example highlights the distinctiveness within the various disciplines studying data science education, as the authors of these articles motivated their research based on their discipline-specific research—not the need to integrate computing into statistics that Nolan and Temple Lang (2010) highlighted. Last, data science practice is emergent and dynamic, changing quickly with new digital technologies and questions (Cao, 2017; National Academies of Sciences, Engineering, and Mathematics, 2018). It is hard to know what types of problems and goals data scientists will have in a few decades, which means the goals for data science education research have and will continue to change. The emergent and interdisciplinary nature of what falls under the data science and data science education umbrellas creates significant challenges in synthesizing what we know and are learning across different research and development agendas—as H. Lee et al. (2022) and others (e.g., V. Lee et al., 2022) point out, the field is different from most others in K-12 education in that it is still developing and open to being shaped.

Reflecting on these challenges to summarizing research about data science learning has led us to pull back a bit from the "What do we know about data science learning?" question we initially sought to answer, as any answer to that question may privilege a particular conceptual framing of the phenomenon of data science learning. Given the diverse conceptual frameworks and approaches, we instead chose to attempt a description of the current research that would illuminate the *dimensions* (areas of emphasis) along which the research is being carried out. The diverse communities engaged in this research have different foci, methods, and language for their data science education research, but in our review, we found that these often could be attributed to differences in how agency is prioritized by the communities. Because of this, we have chosen to review research on data science learning based on how the research prioritizes three dimensions: material agency, personal agency, and disciplinary agency. Last, we do not mean to suggest that these categories are mutually exclusive, or that these communities do not have variability in their prioritization of them. However, we hope that this framing will provide a language that supports synthesis across diverse approaches to understanding data science learning.

Material, Personal, and Disciplinary Agency in Data Science Education Research

Pickering's (1995) ideas about the performance of science are particularly helpful in understanding research about data science learning. Pickering's work comes from the discipline of

Science and Technology Studies—a community interested in the philosophy of science. He expands on prior research in this area by focusing not only on topics typically within the purview of sociological perspectives but also on what he referred to as the material aspects of how science works—or how science is *performed*. From this perspective, scientific work is a performance that emerges in real time without the knowledge of the final form that the result of the work will take. This means practices, methods, machines, explanations, and knowledge are built without knowing the final result or the implications of their development. This helps us think about the emergent fields of both data science and data science education by foregrounding various performances that are carried out through the agency of actors, what he refers to as a "dance of agency." We argue that Pickering's scholarship has something to say about data science and data science learning. Data science education is primarily focused on supporting students to learn to look at the world using machines and practices that extend the capacity of our human minds and bodies. We cannot see global migration patterns (Kahn, 2020) or movie ratings and revenue (Fergusson & Phannkuch, 2022), for example, without machines that humans build to "capture, seduce, download, recruit, enroll, or materialize" (Pickering, 1995, p. 7) the material agency that the world exerts on us. These machines, then, create a new form of material agency, as they generate new forms of data and questions we did not anticipate while building them. This is one reason why the field of data science is motivated by new problems—many of which were unanticipated until they emerged, created by innovative digital technologies and the worlds and *digital trace* data (Fischer et al., 2020; Welser et al., 2008) they create.

The material agency of the natural world and the machines we build to capture it exert agency on humans, but humans also exert agency on the world in response. Humans exert *personal agency* in an attempt to capture material agency. These forms of agency "dance" in a dialectic of resistance and accommodation, as material agency resists being captured and personal agency accommodates the plans for capture and thus tunes these plans to where the world's agency is sufficiently pinned down. However, humans rarely act alone. Pickering also attends to the social aspect of scientific work by describing a new form of agency when communities shape practices, ideas, and machines. These kinds of *disciplinary agency* support humans to capture material agency but also constrain personal agency as practitioners are held to disciplinary norms and conventions in order to participate in particular communities. The agency of these community norms then engages in the dance with material and personal agency.

Although Pickering (1995) describes much more than these features of scientific work, we argue here that *material agency*, *personal agency*, and *disciplinary agency* help us see important dimensions of data science and data science education, as represented in Figure 1.

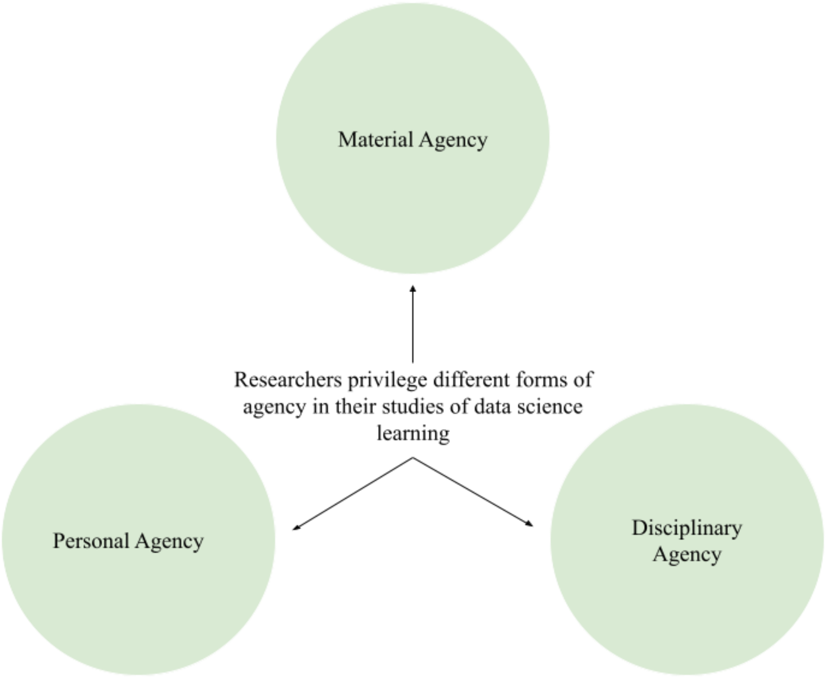
Material Agency
Researchers privilege different forms of agency in their studies of data science learning
Personal Agency
Disciplinary Agency

Figure 1. *Pickering's different forms of agency that we use to understand data science learning.*

Of course, following Pickering's (1995) line of reasoning, the phenomenon that is data science education that we are studying resists our attempts to study it: We acknowledge that most, if not all, of the prior research fits into multiple groups (material, personal, and disciplinary agency) we have used and that we have only reviewed a slice of the relevant research. Also, we note that the purpose of using agency as a frame is not to wedge research into a particular group; instead, it is to reveal the relative areas of emphasis and to highlight opportunities for future design, research, and development. Still, we maintain that this frame has pragmatic value in that we can consider research on data science learning from a wide enough perspective to talk across diverse communities.

Research on Data Science Learning That Emphasizes Material Agency

Material agency refers to the ways the natural world and human-engineered worlds are active in scientific work by creating phenomena that humans seek to understand, and by resisting human efforts to examine phenomena as we'd prefer. Both the natural world and the worlds that emerge from human-engineered tools and systems are constantly variable, which creates challenges for those seeking to understand it. After all, how can we answer a question about the world when the world is constantly changing? The statistics field grew in response to a variable world (Cobb & Moore, 1997). For example, Fisher developed statistical methods for making inferences about fertilizer treatments in agricultural research because outcomes from identical conditions were variable (Rodgers, 2010; Wild et al., 2018). As we suggested in the previous section, data science, too, has grown out of a need to manage new forms of variable data created by rapid growth in computing power (Cao, 2017; V. Lee & Wilkerson, 2018; H. Lee et al., 2022). Data science work, then, is fundamentally an effort to parse out and explain variation in ways that are meaningful for the questions at hand. Statistics education and science education communities have made productive student engagement with the variation and uncertainty generated from material agency a focus of both research and design.

The American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education grounds all statistical inquiry as a practice of making sense of variable data (Bargagliotti, 2020; Moore & Cobb, 1997). This is sometimes referred to as engaging students with the *context* of a data set (e.g., Rubin, 2020). This means that students should have opportunities to develop data creation practices to understand how choices about sample and measurement influence the variability in data (Hardy et al., 2020). When carefully designed, this type of engagement with variability can support students to better understand both practices related to data science and the phenomenon under investigation (Ford & Forman, 2006; Manz, 2015). This is especially true when students are positioned to parse variation that informs their question from variation due to random noise in the data (Hardy et al., 2020).

In addition to the material agency coming from the natural world, new digital technologies are creating types of material agency that influence data science work in new ways and that create new forms of variability to describe, explain, or predict. One example is the use of digital measurement tools (Lee & Wilkerson, 2018). Students can be supported to engage with these tools in ways that help them reflect on the agency of the tool and how to use it to accomplish their own goals (Hardy et al., 2020). When given opportunities to grapple with challenges associated with data creation, students are better positioned to manage material agency when cleaning and managing the "messy data" that results when they carry out investigation procedures (Rosenberg et al., 2020; Hammett & Dorsey, 2020; Schanzer et al., 2022; Kjelvik & Schultheis, 2019).

Research has shown, however, that it is important to carefully consider the forms of variability that students encounter. Variation from measurement error alone has proven powerful and accessible for early experiences with data generation (Konold & Pollatsek, 2002; Konold & Lehrer, 2008). In contrast, variability caused by natural differences, production errors, and random fluctuations are more challenging to reason through. It is clear that the source of variability is highly consequential, but little is known about how students reason about measurement error when using opaque measurement tools such as digital probes. This is likely a consequential difference for students since visible measurement methods, such as using a ruler to measure the circumference of a tree trunk, provide accessible ways for students to reason about error due to measuring trees at different heights, reading the rulers with degrees of precision, or creating gaps or overlaps when iterating the ruler around the trunk. But, it is less clear what resources students can use to reason about sources of error when using digital probes, such as during investigations that involve recording the temperature or pH of a solution. The probes could be calibrated imprecisely, students might record data with a lag or at periodic intervals, or there might be slight differences between manufacturers of these instruments. These considerations may be less visible to students and more work is needed to understand how to support students well with these tools.

Although it is important to engage students with material agency, students are often not given these opportunities (Miller et al., 2018; Hardy et al., 2020). This happens when disciplinary agency is prioritized because students' responses to material agency rarely replicate disciplinary norms. So, if a teacher has a disciplinary norm as the primary goal of the investigation, they often restrict engagement with the challenges of material agency because they are likely to complicate students' conceptual interpretations of the investigation (Hardy et al., 2020). In addition, reasoning about variability is challenging for most people (Torok & Watson, 2000). Careful design, then, is needed to support teachers to engage students productively with material agency.

Summary of Research on Data Science Learning That Emphasizes Material Agency:

- It can be productive for students to describe sources of variability, including variability due to measurement, the phenomenon under investigation, and random noise, but careful design is needed because it can be tempting to avoid the challenges of engaging with the material agency.
- While there is value in productively engaging students with material agency in the context of complex (even "messy") data, the conditions under which data should be messy and when it should not need to be better understood.
- Variability can be described, explained, or predicted; there is a body of research on how students can productively describe and explain data.
- Material agency requires learners to engage with understanding the context through which data came to be. Given that data science draws on subject matter expertise that is relevant to a broad range of contexts, learning about the context of the data is challenging but important.

Prior Research on Data Science Learning That Emphasizes Personal Agency

In this context, personal agency refers to the efforts humans make to capture material agency in ways that help us control, observe, and explain it. As the natural and human-engineered worlds create variable phenomena, humans have developed strategies, representations, statistics, models, and computational systems to structure the variability in ways that allow us to make observations and inferences that can't be made by direct observation alone. However, these methods are always human-generated artifacts, which means that data science methods and claims are always theory-laden (Hardy et al., 2020). Methods are never neutral, and always represent the perspectives, goals, and values of the people and communities that develop them. Learning sciences communities and Statistics Education communities often focus on understanding the personal agency of learners in research on data science learning (e.g., Lee & Wilkerson, 2018).

When students are given personal agency to make decisions about their data science investigations, they can develop an understanding of and competency in using their agency to make claims about the world around them. One result is that students expand their notions of what counts as data to include information and artifacts from their own personal lives (Stornaiuolo, 2020). For learners, then, data is not information generated by someone else, but potentially information about themselves and their lives. Students can also develop competency in crafting "data stories" that are both empirically grounded and personally meaningful (Kahn, 2020; Stornaiuolo, 2020; Roberts & Lyons, 2020; Lee & Dubovi, 2020; Wilkerson & Laina, 2018; Wilkerson et al., 2021). These data stories position the students against the data and the wider society the data represent, and students often use their agency and resources around them to better understand their own history and community (Kahn, 2020; Roberts & Lyons, 2020). This can also help students understand that data and claims made with data are not objective but are always stories told by people from a particular perspective (Rubin, 2020). This can support them to critically interrogate the personal agency behind data science systems and claims that others have developed, sometimes referred to as critical data literacies (Stornaiuolo, 2020). This critical stance on data science positions students to be not only objects of investigation, but to be active agents in generating, interpreting, and critiquing data and decisions made with data.

Research has also shown that students’ personal agency can have strong epistemic congruence with disciplinary norms and values. By epistemic congruence, we mean that students' data science work rarely replicated disciplinary, but the motivation behind the students’ approaches and the ways they use the representations, statistics, models, and computational systems they develop strongly resembles the ways disciplinary tools are used. Student-generated data visualizations might not replicate the rules of a histogram or dot plot, but they can often use principles like scale, order, and frequency as tools to communicate something important about their data (Petrosino et al., 2003; Lehrer & Schauble, 2007; Konold, Higgins, Russell, & Khalil, 2015). This work can then support students to view data in terms of the aggregate shape created by representations, which is important for then thinking about descriptive statistics to index center and variability (Watson & Mortiz, 2000). Research has shown that students can invent innovative statistics that attend to distributional characteristics in meaningful ways and that with thoughtful design and support they use and reason about these statistical inventions in ways that are similar to professional statistical work (Lehrer & Kim, 2009; Jones et al., 2017). Students can then use these ideas and practices as epistemic tools to make inferences with data in innovative ways (Konold, 2012). And finally, research has shown that students can exert personal agency to create, compare, revise, and use probability models to make inferences and develop machine learning algorithms (Lehrer & English, 2018; Zimmerman-Neifeld et al., 2019).

Summary of Research on Data Science Learning That Emphasizes Personal Agency:

- Learners possess epistemic—knowledge-related—assets that can serve researchers, designers, and developers of data science learning when these assets are recognized in instruction and instructional design.
- Though students are unlikely to reinvent all of data science during the course of a class, there is significant evidence that with careful design students can be positioned to see data as personally meaningful, community coherent data stories that communicate important findings to target audiences, and develop data science approaches that resemble disciplinary approaches in meaningful ways.
- It is important to scaffold and sequence students’ learning about data in such a way that permits learners to not only exert personal agency but also make progress toward generating answers to the questions on which they were working.

Prior Research on Data Science Learning That Emphasizes Disciplinary Agency

Disciplinary agency concerns the norms and practices that individuals in a discipline adopt—and explicitly and implicitly coerce others to adopt. Of course, such forms of agency are malleable and can change over time, as is the case with the emerging domain that is data science. Statisticians and computer scientists have been prominent in writing about data science learning from a perspective that highlights the norms and practices of their disciplines. This work is taking place in communities that are likely familiar to statisticians and computer scientists, especially those who carry out research in post-secondary settings.

Indeed, one of the most important papers that highlights disciplinary agency is Nolan and Temple Lang’s (2010) article calling for a greater role for computation in the statistics curriculum, as this work was not primarily motivated by issues related to personal or material, but rather by changes in the statistics (and burgeoning data science) discipline. Ahead of its time, this article calls for greater breadth and depth regarding students’ use of computational methods when they are learning statistics—as well as the importance of using computational methods in the context of working with data. Disciplinary norms are privileged by considering the six topics Nolan and Temple Lang (2010) recommended for integration into the undergraduate curriculum, including scientific computing with data (i.e., programming the steps to be taken in an analysis), computational statistics, and the use of integrated development environments (applications for editing, writing, and debugging code). Several of these are contiguous with traditional statistics topics; others—such as advanced computing, including the use of distributed/high-performance computing systems—are less often covered in traditional statistics education.

Many articles highlight the importance of programming to the statistics and data science disciplines (Çetinkaya-Rundel et al., 2022; Çetinkaya-Rundel & Ellison, 2021; Dogucu & Çetinkaya-Rundel, 2021; Fergusson & Pfannkuch, 2020; Hardin et al., 2015; Heiznman, 2020; Kim & Hardin, 2021; Kim & Henke, 2021). Also, an emphasis on programming has been an emphasis for researchers who approach data science through the lens of computing education and the role of data in the work of computer scientists (Dryer et al., 2018; Schanzer et al., 2022). This emphasis on programming is not limited to these articles; Nolan and Temple Lang (2010) pulled no punches when writing about their importance, claiming that, "computational literacy and programming are as fundamental to statistical practice and research as mathematics." (p. 96). Providing further evidence for the centrality of programming, Schwab-McCoy et al. (2021) conducted a survey of introductory data science course instructors and found that RStudio (a commonly-used integrated development environment for R) and Jupyter notebooks (a type of document for data scientists to use Python) were the most commonly used software for introductory, college-level data science courses. Mirroring this finding, the most commonly used programming languages were R and then Python, respectively. Further, there is a clear synergy between this work and that of computer scientists, especially those taking a *data-centric* approach (e.g., Schanzer et al., 2022; Krishnamurti & Fisler, 2020). Notably, two prominent K-12 data science curricula, the high school-focused *Introduction to Data Science* (Gould et al., 2018) and the middle and high school-focused *Bootstrap:Data Science* (Schanzer et al., 2022) both emphasize programming.

In short, programming—especially in R at the undergraduate level (Schwab-McCoy et al., 2021)—is a disciplinary tool that is having a strong influence on data science. This research teaches us that novices can learn to code in the context of a one-semester course. Admittedly, this research is primarily at the undergraduate level, but the students in such courses often bring similar degrees of preparation as students at the high school level—especially older high school students. Further, the research around the two K-12 data science curricula also emphasizes programming—and there is some evidence for how even young students can learn to program and find programming to be valuable to them (Heinzman 2022; Schanzer et al., 2022). We do note that far from all research on data science learning (even at the undergraduate level) emphasizes programming. Many scholars have called for an approach that emphasizes data literacy at least at the very outset of students learning data science. These studies either provide web-based tools for students to gradually become familiar with code (Burckhardt et al., 2021) or do not use programming at all (Mike & Hazzan, 2022).

Another element of this research is on undergraduate course contexts and particular designs that engage students with the specific technologies data scientists use in practice; what Horton and Hardin (2021) refer to as *creative structures*. These include deliberate choices about aspects of the course beyond the programming language that is (often) used: the course website, nature of assignments and projects, and means for teachers and students of giving and receiving feedback, among others. As others have pointed out, this work often takes the form of case studies (for examples of case studies, see Schwab-McCoy et al., 2021). Çetinkaya-Rundel and Ellison's (2021) paper is emblematic of this approach. Other papers extend this sophisticated infrastructure to emphasize reproducibility through the use of the server tool *Docker* (Çetinkaya-Rundel & Rundel, 2018) and other technically sophisticated technologies and programming tools (Burckhardt et al., 2021; Dogucu & Çetinkaya-Rundel, 2022; Kim & Henke, 2021). Thus, a second focus is on creative, thoughtfully designed course structures, especially at the undergraduate level. This work suggests that in order for students to be successful, instructors need to consider the instructional design of their courses; how the course goals, instruction, assessments, opportunities for practice and help, and even the technologies used to manage the submission of assignments align and work together to support learners to know about and do data science.

A final finding from research highlighting disciplinary agency concerns scaffolds for learning new and highly valued analytic and modeling techniques in data science disciplines (Horton & Hardin, 2021), such as machine learning. For example, Fergusson and Pfannkuch (2022) show how the core tenets of machine learning can be taught to K-12 aged students when designed and taught in a particular manner; namely, using a particular ("informal") approach that emphasized visualizations, a potentially relevant data set (movie ratings), and a browser-based environment for students to run R code. Other papers emphasize machine learning (e.g., Jiang et al., 2022; Zimmermann-Niefield et al., 2019) and even artificial intelligence (Druga & Ko, 2021) as well as modern approaches to inferential modeling (Kim et al., 2021), including Bayesian approaches (Erickson, 2017; Kazak, 2015; Rosenberg et al., 2022; Warren, 2020), developing statistical software (Reinhart & Genovese, 2021), web scraping using social media data (Boehm & Hanlon, 2021; Dogucu & Çetinkaya-Rundel, 2021), and using git and GitHub (Adams et al., 2021; Beckman et al., 2021; Curtis et al., 2020; Kim & Henke, 2021;). This work shows that learners can develop the capacity to use new analytic and programming tools with deliberately-designed courses.

Summary of Research on Data Science Learning That Emphasizes Disciplinary Agency:

- Research that prioritizes disciplinary agency has emphasized the roles of computing and highly valued procedures and tools within data science disciplines (e.g., programming and scientific computing with data, the use of integrated development environments, and version control tools).
- Many studies at the undergraduate level, and a few at the K-12 level, have chosen particular programming languages amenable to teaching beginning programmers, especially R and python.
- Students can learn particular, technically-sophisticated, often recently-emerged skills, like machine learning, inferential statistical methods, and statistical software engineering, when supported through careful instructional design and support from software tools and teachers.

How Can We Generate a Deeper Understanding of Data Science Learning?

Data science education is a new field, and summing up what is known about data science learning requires a landscape perspective. To take a wide-angled perspective, we considered a framework that includes three forms of agency that have a bearing on how a field of science may unfold—one based on Pickering's (1995) description of the forms of material, personal, and disciplinary agency.

Using this framework, we first identified and described research that highlights the role of material agency, or how the world affords and constrains particular forms of data and types of data analysis. In many ways, this form of agency is likely familiar to statistics education researchers who have long emphasized ideas around variability, context, and case and aggregate (Rubin, 2020)—indeed a lot of this work has been carried out by scholars in statistics education, with contributions by scholars in the learning sciences, computer science education, and other disciplines. But, new technologies invite new forms of data (e.g., digital trace data) and new methods (e.g., machine learning and text analysis methods). In this way, we can see how the present-day world "resists" (in a way that can be thought of as agentic) traditional attempts to be understood. This means innovation on the part of data scientists—and teachers and learners—is needed. Following material agency, we considered research predominantly by scholars in the learning sciences that prioritized the role of learners' personal agency, describing research that showed how careful design and a strident focus on supporting the goals and ideas that learners have can lead them to meaningful questions, answers, or solutions to problems. This research shows us that learners' motivations and resources matter with respect to learning data science; even young learners know about the world and can use data and data science methods to make sense of the world. Finally, personal agency—in a field of study—can develop into the norms and practices of a discipline. Such has been the case with even the emerging field of data science, where there now exist many courses and curricula, reform documents, and books and articles in journals. Statistics and data science educators and computer scientists have been at the forefront of this research, conducting case studies, carrying out survey research, and developing and assessing curricula that reflect what data scientists do in a form that is intended to be accessible and useful to learners at the K-12 and undergraduate level. This work privileges computer programming and other computational skills to learn new data science techniques.

Following this summary, we make several recommendations for how to extend the state of what we know about data science learning. We frame these recommendations humbly and in the spirit of questions to work to answer together.

To what ends can we orient our work? There are two freely-available curricula focused on data science learning (*Introduction to Data Science* and *Bootstrap:Data Science*) worth celebrating, but the data science education community might find value in considering general “models” for how research on data science learning could be oriented. The general question we ask invites answers inspired by the work of scholars in other (K-12, especially) educational research disciplines. One answer that we think is useful is the model provided by learning analytics research, which has the potential to help to organize the work of interdisciplinary and collaborative research groups. Learning progressions are developmental, empirically grounded, and interpretable and usable to educators (Arnold et al., 2018). Thus, learning progressions also knit a developmental view of learning with specific contexts and strategies for supporting that learning. Since learning progressions are empirically grounded, they can be well attuned to learners’ ideas and challenges, providing a means to focus on personal agency and how it interacts with material agency. Since they are developmental, learning progressions can have an eye on important disciplinary goals, and communicate the ways students might engage with material agency to challenge and grow their thinking towards these goals. For instance, if designers and educators wish for students to think about how data can be used for a purpose other than what it was originally intended, a research program organized around learning progressions might develop empirical evidence of students’ initial thinking about secondary data sources, how their thinking shifts and changes as they work to create new questions while exploring the data, and how they think with secondary data sources about how others generated the measures. Other education fields, including statistics (Arnold et al., 2018), science (Alonzo & Gotwals, 2012; Schwarz et al., 2009) and mathematics (Fonger et al., 2018), and, recently, computing (Rich et al., 2018) have worked to develop such learning progressions. What, for instance, might a learning progression for machine learning or predictive analytics be like? Or a learning progression for students’ ideas about the nature and role of data? Or how do ideas about programming develop alongside ideas about exploring data? Last, how might we assess student learning along a progression (e.g., with an assessment like the *Levels of Conceptual Understanding in Statistics;* Whitaker et al., 2015)? As noted earlier, we think learning progressions are an example of one innovation we can borrow from other fields.

What is the role of systems-level organizations and supports? We began this paper with a reference to the important work of Nolan and Temple Lang (2010) which is essential to statistics and data science educators at the undergraduate level, but, through the present, not referenced very much by researchers focused on data science learning at the K-12 level. We think this one example highlights how there is little dialogue between data science education researchers at the undergraduate and K-12 levels (in addition to data science educators in industry—a substantial group we have not yet mentioned; Kross & Guo, 2019). This example also suggests that there may be other, similar disconnects, such as those between scholars at the K-12 level who approach data science education from backgrounds in computing education or the learning sciences—or any of the other sub-disciplines that make up educational research. In sum, we think that research in data science education may be developing more rapidly than one might expect, but also in more siloed ways. If this is the case, what are the roles of systems-level organizations and supports—such as (additional) journals, conferences, and funding sources—to support sustained growth in research about learning data science? How might teachers be supported to teach data science? And how might schools and districts deploy the necessary infrastructure to support data science learning?

How much weight should be given to the different dimensions of knowing and doing data science? This is a broad question that relates to how data science is defined. Here, we point to the work of other scholars on definitional questions—especially V. Lee et al.(2022)—and instead focus narrowly on the computational (and programming-related) parts of data science. While computing is generally considered to be a core part of what sets data science apart from statistics (Breiman, 2001; Donoho, 2017; National Academies of Sciences, Engineering, and Medicine, 2018), its role in data science learning is not entirely settled. Namely, some data science curricula—even at the undergraduate level—de-emphasize programming (Burckhardt et al., 2021; Mike & Hazzan, 2022). This is the case at the K-12 level, too; few papers in either the aforementioned *Journal of the Learning Sciences* or the *British Journal of Educational Technology* special issues involved programming (though many used computers to access, process, visualize, and model data), though the high school (*Introduction to Data Science*) and middle and high school (*Bootstrap: Data Science*) curricula both emphasize programming. We think the question about the role of programming is an important one on its own, especially as computer science courses become more common at the K-12 level. Also, we think that a more general topic to consider is the weight given to the different concepts (the "knowing") and practices or skills (the "doing") that comprise data science. Even if students at the K-12 level do not program to wrangle or process data, they might when modeling data, for instance (e.g., Wilkerson-Jerde & Wilensky, 2015; Wise, 2020). Also, given that many definitions of data science include a discipline-specific component (cf. V. Lee et al., 2022), a related question concerns what about learning data science generally applies across disciplines, and what is specific to the topics, data, methods, and phenomena that characterize specific disciplines, a question asked by data science education scholars (Finzer, 2013; Jiang et al., 2022). As teachers in different subject areas may have different degrees of experience teaching with data (Rosenberg et al., 2022), efforts to support data science learning across disciplines may need to carefully consider what discipline-specific supports, tools, and resources are needed for teachers and learners alike.

Whose voice is highlighted in scholarship on data science learning? The question of who has a voice—and whose voices are valued—clearly overlaps with questions about not only diversity but also questions about what is societally just. This question also pertains to which disciplinary backgrounds are privileged or discounted. We think these different dimensions of who has a voice cannot be easily disentangled, and so we address them together. While there is doubtless value in coming to similar findings despite using different research approaches, there is also value in different approaches to researching data science learning. We have found it beneficial to consider essential research by scholars not only with backgrounds in computer science and statistics education, for example, but also in the arts (Matuk et al., 2022), humanities (Vance et al., 2022), social studies (Shreiner & Guzdial, 2022), and everyday life and experiences (Gebre, 2022; Radinsky & Tabak, 2022; Vacca et al., 2022). Other scholars have argued for the importance of a broadly humanistic approach to data science education, one that recognizes that questions about who uses data overlap with questions about who has value and power (V. Lee et al., 2021; Philip et al., 2016). As the field develops, it is important to continuously reflect on who shapes what we know about data science learning.

Conclusion

In some ways, reviewing research on data science learning is easier now than it was at any point in the past: We were guided by the work of the contributors and the editors of no fewer than four special issues on the topic across different disciplines as well as several related reports and review papers. At the same time, the developing research across diverse communities creates new challenges in summarizing where we are and what we know. We believe attention to how material, personal, and material agency are prioritized in scholarship about data science learning can support synthetic conversations across communities–and possibly collaborations across different disciplines. Future progress depends on bringing together the strengths of the research across different fields, and attention to how students are supported to meaningfully use data to explore a variable and changing world. We are hopeful that such progress can lead to a more complete portrait of data science learning and opportunities in the future for students at the K-12 level to do and share the results of their ambitious work with data.

References

Adams, B., Baller, D., Jonas, B., Joseph, A.-C., & Cummiskey, K. (2021). Computational Skills for Multivariable Thinking in Introductory Statistics. *Journal of Statistics and Data Science Education, 29,* S123–S131. https://doi.org/10.1080/10691898.2020.1852139

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions.* Springer Science & Business Media.

Arnold, P., Confrey, J., Jones, R. S., Lee, H.L., & Pfannkuch, M. (2018). Statistics learning trajectories. In D. Ben-Zvi., K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education*. Springer, Cham.

Bargagliotti, A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II* (GAISE II). American Statistical Association.

Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with git and GitHub as a learning objective in statistics and data science courses. *Journal of Statistics and Data Science Education, 29,* S132–S144. https://doi.org/10.1080/10691898.2020.1848485

Boehm, F. J., & Hanlon, B. M. (2021). What is happening on Twitter? A framework for student research projects with tweets. *Journal of Statistics and Data Science Education, 29,* S95–S102. https://doi.org/10.1080/10691898.2020.1848486

Breiman, L. (2001). Statistical Modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Burckhardt, P., Nugent, R., & Genovese, C. R. (2021). Teaching statistical concepts and modern data analysis with a computing-integrated learning environment. *Journal of Statistics and Data Science Education, 29,* S61–S73. https://doi.org/10.1080/10691898.2020.1854637

Cao, L. (2017). Data science: A comprehensive overview. ACM Computing Surveys, 50(3), 1–42. https://doi.org/10.1145/3076253

Çetinkaya-Rundel, M., Dogucu, M., & Rummerfield, W. (2022). The 5Ws and 1H of term projects in the introductory data science classroom. *Statistics Education Research Journal, 21*(2), 4–4. https://doi.org/10.52041/serj.v21i2.37

Çetinkaya-Rundel, M., & Ellison, V. (2020). A fresh look at introductory data science. *Journal of Statistics Education, 29*(1), 1–11. https://doi.org/10.1080/10691898.2020.1804497

Çetinkaya-Rundel, M., & Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician, 72*(1), 58-65.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. T*he American Mathematical Monthly, 104*(9), 801-823.

D'ignazio, C., & Klein, L. F. (2020). *Data feminism.* MIT press.

Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web scraping in the statistics and data science curriculum: challenges and opportunities. *Journal of Statistics and Data Science Education, 29*(sup1), S112–S122. https://doi.org/10.1080/10691898.2020.1787116

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745-766.

Druga, S., & Ko, A. J. (2021). How do children's perceptions of machine intelligence change when training & coding smart programs? *In Interaction Design and Children*. June 24–30, 2021, https://doi.org/10.1145/3459990.3460712

Dryer, A., Walia, N., & Chattopadhyay, A. (2018). A middle-school module for introducing data-mining, big-data, ethics and privacy using RapidMiner and a Hollywood theme. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 753–758. https://doi.org/10.1145/3159450.3159553

Erickson, T. (2017). Beginning Bayes. *Teaching Statistics, 39*(1), 30-35.

Fergusson, A., & Pfannkuch, M. (2022). Introducing high school statistics teachers to predictive modelling and APIs using code-driven tools. *Statistics Education Research Journal, 21*(2), 8–8. https://doi.org/10.52041/serj.v21i2.49

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). https://doi.org/10.5070/T572013891

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education, 44*(1), 130-160.

Fonger, N. L., Stephens, A., Blanton, M., Isler, I., Knuth, E., & Gardiner, A. M. (2018). Developing a learning progression for curriculum, instruction, and student learning: An example from mathematics education. *Cognition and Instruction, 36*(1), 30-55.

Ford, M. J., & Forman, E. A. (2006). Chapter 1: Redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30*(1), 1-32.

Gebre, E. (2022). Conceptions and perspectives of data literacy in secondary education. *British Journal of Educational Technology,* 53(5), 1080-1095. https://doi.org/10.1111/bjet.13246

Gould, R., Machado, S., Johnson, T. A., & Molynoux, J. (2018). *Introduction to Data Science v 5.0.* UCLA Center X.

Hammett, A., & Dorsey, C. (2020). Messy data, real science. *The Science Teacher, 87*(8), 40-49.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "Think with data." *The American Statistician, 69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

Hardy, L., Dixon, C., & Hsi, S. (2020). From data collectors to data producers: Shifting students' relationship to data. *Journal of the Learning Sciences, 29*(1), 104–126. https://doi.org/10.1080/10508406.2019.1678164

Heinzman, E. (2022). "I love math only if it's coding": A case study of student experiences in an introduction to data science course. *Statistics Education Research Journal, 21*(2), 5–5. https://doi.org/10.52041/serj.v21i2.43

Horton, N. J., & Hardin, J. S. (2021). Integrating computing in the statistics and data science curriculum: creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education,* 29, S1–S3. https://doi.org/10.1080/10691898.2020.1870416

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. https://doi.org/10.1111/bjet.13258

Jiang, S., Nocera, A., Tatar, C., Yoder, M. M., Chao, J., Wiedemann, K., ... & Rosé, C. P. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. British *Journal of Educational Technology, 53*(5), 1114-1133.

Jones, R. S., Lehrer, R., & Kim, M.-J. (2017). Critiquing statistics in student and professional worlds. *Cognition and Instruction, 35*(4), 317-336.

Kahn, J. (2020). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences, 29*(1), 57–80. https://doi.org/10.1080/10508406.2019.1693377

Kazak, S. (2015). A Bayesian inspired approach to reasoning about uncertainty: 'How confident are you?'. In *CERME 9-Ninth Congress of the European Society for Research in Mathematics Education* (pp. 700-706).

Kim, A. Y., & Hardin, J. (2021). "Playing the whole game": A data collection and analysis exercise with Google Calendar. *Journal of Statistics and Data Science Education, 29*(sup1), S51–S60. https://doi.org/10.1080/10691898.2020.1799728

Kim, B., & Henke, G. (2021). Easy-to-use cloud computing for teaching data science. *Journal of Statistics and Data Science Education, 29*(sup1), S103–S111. https://doi.org/10.1080/10691898.2020.1860726

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE—Life Sciences Education, 18*(2), 1-8.

Krishnamurthi, S., & Fisler, K. (2020). Data-centricity: A challenge and opportunity for computing education. *Communications of the ACM, 63*(8), 24–26. https://doi.org/10.1145/3408056

Kross, S., & Guo, P. J. (2019, May). Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *CHI Conference on Human Factors in Computing Systems Proceedings* (CHI 2019). 1-14. https://doi.org/10.1145/3290605.3300493

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: key practices and processes. *Statistics Education Research Journal, 21*(2), 3–3. https://doi.org/10.52041/serj.v21i2.41

Lee, V. R., & Dubovi, I. (2020). At home with data: family engagements with data involved in type 1 diabetes management. *Journal of the Learning Sciences, 29*(1), 11–31. https://doi.org/10.1080/10508406.2019.1666011

Lee, V. R., Pimentel, D. R., Bhargava, R., & D'Ignazio, C. (2022). Taking data feminism to school: A synthesis and review of pre-collegiate data science education projects. *British Journal of Educational Technology, 53*(5), 1096-1113bjet.13251. https://doi.org/10.1111/bjet.13251

Lee, V. R., & Wilkerson, M. (2018). Data use by middle and secondary students in the digital age: A status report and future prospects. *Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12.* Washington, D.C.

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. https://doi.org/10.3102/0013189X211048810

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In *Thinking with data* (pp. 163-190). Psychology Press.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal, 21*(2), 116-133.

Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In *International handbook of research in statistics education* (pp. 229-260). Springer, Cham.

Manz, E. (2015). Resistance and the development of scientific practice: Designing the mangle into science instruction. *Cognition and Instruction, 33*(2), 89-124.

Matuk, C., DesPortes, K., Amato, A., Vacca, R., Silander, M., Woods, P. J., & Tes, M. (2022). Tensions and synergies in arts-integrated data literacy instruction: Reflections on four classroom implementations. *British Journal of Educational Technology, 53*(5), 1159-1178. bjet.13257. https://doi.org/10.1111/bjet.13257

Mike, K., & Hazzan, O. (2022). Machine learning for non-majors: a white box approach. Statistics Education Research Journal, *21*(2), 10–10. https://doi.org/10.52041/serj.v21i2.45

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options.* National Academies Press.

Nolan, D., & Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician, 64*(2), 97–107. https://doi.org/10.1198/tast.2010.09132

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*(2-3), 131-156.

Philip, T. M., Olivares-Pasillas, M. C., & Rocha, J. (2016). Becoming racially literate about data and data-literate about race: Data visualizations in the classroom as a site of racial-ideological micro-contestations. *Cognition and Instruction, 34*(4), 361-388.

Pickering, A. (1995). *The mangle of practice: Time, agency, and science.* University of Chicago Press.

Radinsky, J., & Tabak, I. (2022). Data practices during COVID: Everyday sensemaking in a high-stakes information ecology. *British Journal of Educational Technology, 53*(5), 1221-1242.

Reinhart, A., & Genovese, C. R. (2021). Expanding the Scope of Statistical Computing: Training Statisticians to Be Software Engineers. *Journal of Statistics and Data Science Education, 29*(sup1), S7–S15. https://doi.org/10.1080/10691898.2020.1845109

Rich, K. M., Strickland, C., Binkowski, T. A., Moran, C., & Franklin, D. (2018). K--8 learning trajectories derived from research literature: sequence, repetition, conditionals. *ACM Inroads, 9*(1), 46-55.

Roberts, J., & Lyons, L. (2020). Examining spontaneous perspective taking and fluid self-to-data relationships in informal open-ended data exploration. *Journal of the Learning Sciences, 29*(1), 32–56. https://doi.org/10.1080/10508406.2019.165131

Rosenberg, J. M., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher, 87*(5). https://learningcenter.nsta.org/resource/?id=10.2505/4/tst20_087_05_30

Rosenberg, J. M., Kubsch, M., Wagenmakers, E.-J., & Dogucu, M. (2022). Making sense of uncertainty in the science classroom: A Bayesian approach. *Science & Education.* https://link.springer.com/article/10.1007/s11191-022-00341-3

Rosenberg, J. M., Schultheis, E., Kjelvik, M., Reedy, A., & Sultana, O. (2022). Big data, big changes? A survey of K-12 science teachers in the United States on which data sources and tools they use in the classroom. *British Journal of Educational Technology, 53*(5), 1179-1201. https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13245

Rubin, A. (2020). Learning to reason with data: how did we get here and what do we know? *Journal of the Learning Sciences, 29*(1), 154–164. https://doi.org/10.1080/10508406.2019.1705665

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J. G., Lerner, B. S., ... & Krishnamurthi, S. (2022, February). Integrated data science for secondary schools: design and assessment of a curriculum. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1* (pp. 22-28).

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education, 29*(sup1), S40–S50. https://doi.org/10.1080/10691898.2020.1851159

Shreiner, T. L., & Guzdial, M. (2022). The information won't just sink in: Helping teachers provide technology-assisted data literacy instruction in social studies. *British Journal of Educational Technology, 53*(5), 1134-1158. https://doi.org/10.1111/bjet.13255

Stornaiuolo, A. (2020). Authoring data stories in a media makerspace: Adolescents developing critical data literacies. *Journal of the Learning Sciences, 29*(1), 81–103. https://doi.org/10.1080/10508406.2019.1689365

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 12*(2), 147-169.

Vacca, R., DesPortes, K., Tes, M., Silander, M., Matuk, C., Amato, A., & Woods, P. J. (2022, April). " I happen to be one of 47.8%": Social-emotional and data reasoning in middle school students' comics about friendship. In *CHI Conference on Human Factors in Computing Systems (*pp. 1-18).

Vance, E. A., Glimp, D. R., Pieplow, N. D., Garrity, J. M., & Melbourne, B. A. (2022). Data science in 2020: computing, curricula, and challenges for the next 10 years. Integrating the humanities into data science education. *Statistics Education Research Journal, 21*(2), 9–9. https://doi.org/10.52041/serj.v21i2.42

Warren, A. R. (2020). Impact of Bayesian updating activities on student epistemologies. *Physical Review Physics Education Research, 16*(1), 010101.

Welser, H. T., Smith, M., Fisher, D., & Gleave, E. (2008). Distilling digital traces: Computational social science approaches to studying the internet. The *SAGE handbook of online research methods,* 116-141.

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The Levels of Conceptual Understanding in Statistics (LOCUS) project: Results of the pilot study. *Numeracy, 8*(2), Article 3. Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1175&context=numeracy

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is statistics?. In *International Handbook of Research in Statistics Education* (pp. 5-36). Springer, Cham.

Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM, 50*(7), 1223-1235.

Wilkerson, M., Finzer, W., Erickson, T., & Hernandez, D. (2021, June). Reflective data storytelling for youth: The CODAP story builder. In *Interaction Design and Children* (pp. 503-507).

Wilkerson-Jerde, M. H., & Wilensky, U. J. (2015). Patterns, probabilities, and people: Making sense of quantitative change in complex systems. *Journal of the Learning Sciences, 24*(2), 204-251.

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. https://doi.org/10.1080/10508406.2019.1705678

Zimmermann-Niefield, A., Turner, M., Murphy, B., Kane, S. K., & Shapiro, R. B. (2019, June). Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM international conference on interaction design and children* (pp. 121-132).