Getting Messy with Authentic Data: Exploring the Potential of Using Data from Scientific Research to Support Student Data Literacy

Melissa K. Kjelvik* and Elizabeth H. Schultheis
BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI 48824; W.K. Kellogg Biological Station, Michigan State University, Hickory Corners, MI 49060

ABSTRACT

Data are becoming increasingly important in science and society, and thus data literacy is a vital asset to students as they prepare for careers in and outside science, technology, engineering, and mathematics and go on to lead productive lives. In this paper, we discuss why the strongest learning experiences surrounding data literacy may arise when students are given opportunities to work with authentic data from scientific research. First, we explore the overlap between the fields of quantitative reasoning, data science, and data literacy, specifically focusing on how data literacy results from practicing quantitative reasoning and data science in the context of authentic data. Next, we identify and describe features that influence the complexity of authentic data sets (selection, curation, scope, size, and messiness) and implications for data-literacy instruction. Finally, we discuss areas for future research with the aim of identifying the impact that authentic data may have on student learning. These include defining desired learning outcomes surrounding data use in the classroom and identification of teaching best practices when using data in the classroom to develop students' data-literacy abilities.

INTRODUCTION

Throughout K–12 and undergraduate education, a strong emphasis is placed on the development of language literacy to help students understand and navigate everyday life. Similarly, with the rapidly expanding role of data in society, educators need to consider the importance of literacy in the context of data (Mayes *et al.*, 2014; Wolff *et al.*, 2017; National Academies of Sciences, Engineering, and Medicine [NASEM], 2018). Data literacy involves the ability to understand and evaluate the information that can be obtained from data (Schield, 2004; Carlson *et al.*, 2011; Mandinach and Gummer, 2013). A data-literate student should possess the appropriate quantitative and analytical tools necessary to address a problem and the ability to apply these tools in context to analyze, interpret, and communicate findings from data (Gibson and Mourad, 2018).

Data literacy lies at the intersection between the fields of quantitative reasoning and data science, while grounding both in authentic context (Figure 1). Both quantitative reasoning and data science have distinguishing features, yet there is considerable overlap in learning outcomes surrounding the analysis and interpretation of data (Calzada Prado and Marzal, 2013). Both fields share important conceptual similarities, and thus instruction in either area is likely to support the other while increasing data literacy.

Quantitative reasoning is the ability to apply mathematical principles to everyday problems through critical thinking and sound logic (Steen, 2004; Piatek-Jimenez *et al.*, 2012; Boersma and Klyve, 2013; Vacher, 2014; Mayes *et al.*, 2014). In science, quantitative reasoning covers a variety of skills, including the ability to understand

Stephanie Gardner, *Monitoring Editor*
Submitted Feb 21, 2018; Revised Mar 11, 2019; Accepted Mar 22, 2019
CBE Life Sci Educ June 1, 2019 18:es2
DOI:10.1187/cbe.18-02-0023

Potential conflict of interest: Both authors, M.K.K. and E.H.S., are the creators of Data Nuggets and are involved in their production and research to determine their efficacy. No promotion of a particular product to the exclusion of other similar products has occurred.

*Address correspondence to: Melissa K. Kjelvik (kjelvikm@msu.edu).

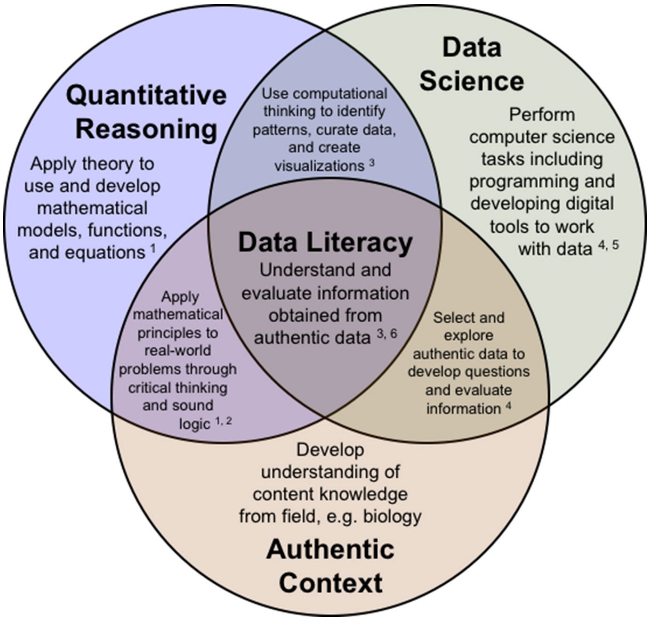

Quantitative Reasoning
Apply theory to use and develop mathematical models, functions, and equations 1
Data Science
Perform computer science tasks including programming and developing digital tools to work with data 4, 5
Use computational thinking to identify patterns, curate data, and create visualizations 3
Data Literacy
Understand and evaluate information obtained from authentic data 3, 6
Apply mathematical principles to real-world problems through critical thinking and sound logic 1, 2
Select and explore authentic data to develop questions and evaluate information 4
Develop understanding of content knowledge from field, e.g. biology
Authentic Context

FIGURE 1. Venn diagram illustrating overlap between the fields of quantitative reasoning and data science, both within and outside authentic contexts. Data literacy lies at the intersection of these two fields when both are explored in an authentic context. Citations reference definitions of the fields, including discussions of overlap between the fields, found in the existing literature. Citations listed in the diagram: 1) Mayes *et al.*, 2014; 2) Steen, 2004; Piatek-Jimenez *et al.*, 2012; Boersma and Klyve, 2013; Vacher, 2014; 3) Calzada Prado and Marzal, 2013; 4) Finzer, 2013; 5) Baumer, 2015; 6) Schield, 2004; Carlson *et al.*, 2011; Mandinach and Gummer, 2013; Gibson and Mourad, 2018.

numerical information found in graphs, tables, equations, and descriptive statistics and to express coherent and logical thinking about quantitative information (Mayes *et al.*, 2014). In addition to specific skills, quantitative reasoning encompasses a learner's emotional responses to quantitative information, including their attitudes, interest, and beliefs (Aikens and Dolan, 2014).

Data science education is a more recent movement with motivation and goals similar to quantitative reasoning. Data science is an interdisciplinary field that embraces sophisticated analytical programming to draw out patterns and useful information from the vast abundance of data available today (Baumer, 2015). Proficiency in data science relies on an understanding of the disciplinary context surrounding a data set, knowledge of math and statistical concepts, and possession of computer science skills (Conway, 2010; Finzer, 2013).

We predict that the strongest learning experiences surrounding data literacy will arise when students have the opportunity to engage in quantitative reasoning and data science while exploring data sets from scientific research. In this paper, we highlight learning opportunities in data literacy that can result from the use of data in secondary and postsecondary classrooms. To facilitate the use of data by educators and students, we identify and describe features that influence the complexity of data sets—scope, selection, curation, size, and messiness. Finally, we describe the resource and training needs of educators and identify areas for future research.

AUTHENTIC DATA IN THE SCIENCE CURRICULUM

Authentic data are true, quantitative or qualitative information, collected from real-life phenomena. Authentic data contrast with inauthentic data, which may be generated to demonstrate a particular pattern or result from manipulation of data to force a specific result or interpretation. Authentic data can be collected using a variety of methods, including the use of measurement tools and automated sensors, or generated through models and simulations. These data sets can be collected by anyone, including scientists, students, and citizen scientists. For the purposes of this article, we will focus on authentic data resulting from scientific observations and investigations.

Scientists rely on many forms of authentic data, including data they collect themselves, data from collaborators, and data archived in online repositories where the scientist may have no direct connection to the individual or sensor that collected the data (Kastens *et al.*, 2015). These sources fall into two general categories: firsthand data collected by the researcher themselves and secondhand data from a variety of external sources (National Research Council [NRC], 1996; Palincsar and Magnusson, 2001; Magnusson *et al.*, 2004). Similarly for students, authentic data in the classroom may come in many forms, including student-collected data from inquiry projects, searches of online data repositories, figures and tables in textbooks, or scientific publications (Hug and McNeill, 2008; Kerlin *et al.*, 2010).

Using Authentic Data to Improve Data Literacy

Because data from scientific research are attached to the context from which they were collected, the use of these authentic data sets in the classroom has the unique potential to develop student data literacy and draw out connections between quantitative reasoning and data science (Figure 1). Learning mathematics in the context of authentic data from scientific investigations reinforces the importance of math for answering questions and may more actively engage students in both math and science (Sorgo, 2010; American Association for the Advancement of Science [AAAS], 2011). In addition, experiences working with authentic data have the potential to engage students in a broader suite of science practices and improve critical thinking (Kerlin *et al.*, 2010; Gould *et al.*, 2014; Holmes *et al.*, 2015), particularly in the areas of analyzing and interpreting data, using mathematics and computational thinking, and engaging in argument from evidence (NRC, 2012; NGSS Lead States, 2013). In fact, student data literacy has been shown to improve when given opportunities to interact with authentic data (Duschl, 1990; Gould *et al.*, 2014; Kastens *et al.*, 2015). Therefore, it is crucial that instructors not overlook the context of a data set as they help students develop their data-literacy abilities. Intentional focus should be placed on exploring authentic situations and the mathematical ideas involved in solving or investigating them (Piatek-Jimenez *et al.*, 2012; Common Core State Standards Initiative [CCSSI], 2014).

In addition to improving data literacy and engaging students in science practices, the use of authentic data in the classroom has the potential to be more interesting and engaging for students (Schultheis and Kjelvik, 2015). When working without an understanding of context, the navigation of a database, management of a data set, or interpretation of output from an online visualization platform may appear more daunting. Inauthentic experiences may fail to engage students, as context is removed and students are asked to explore patterns or trends without meaning. In addition, students may find results from inauthentic data more difficult to interpret than those from authentic data (Piatek-Jimenez *et al.*, 2012; CCSSI, 2014), and students' abilities to transfer new skills to novel contexts in and outside of the classroom may be reduced (Borges-Rey, 2017). In contrast, authentic data transform a typical lesson on data analysis or interpretation by providing real-world context and making connections to disciplinary content, and students report feeling an increased emotional connection to data when they are better able to recognize practical application and relevance (Langen *et al.*, 2014; Wolff *et al.*, 2017). Connecting science to a learners' experience makes content more accessible (Stoddart *et al.*, 2010) and increases student interest in the material (Hulleman and Harackiewicz, 2009). By encouraging students to make connections between the data and their everyday lives, authentic data have the potential to give real-world relevance to data-literacy instruction and tap into students' natural curiosity about their world (Doering and Veletsianos, 2007).

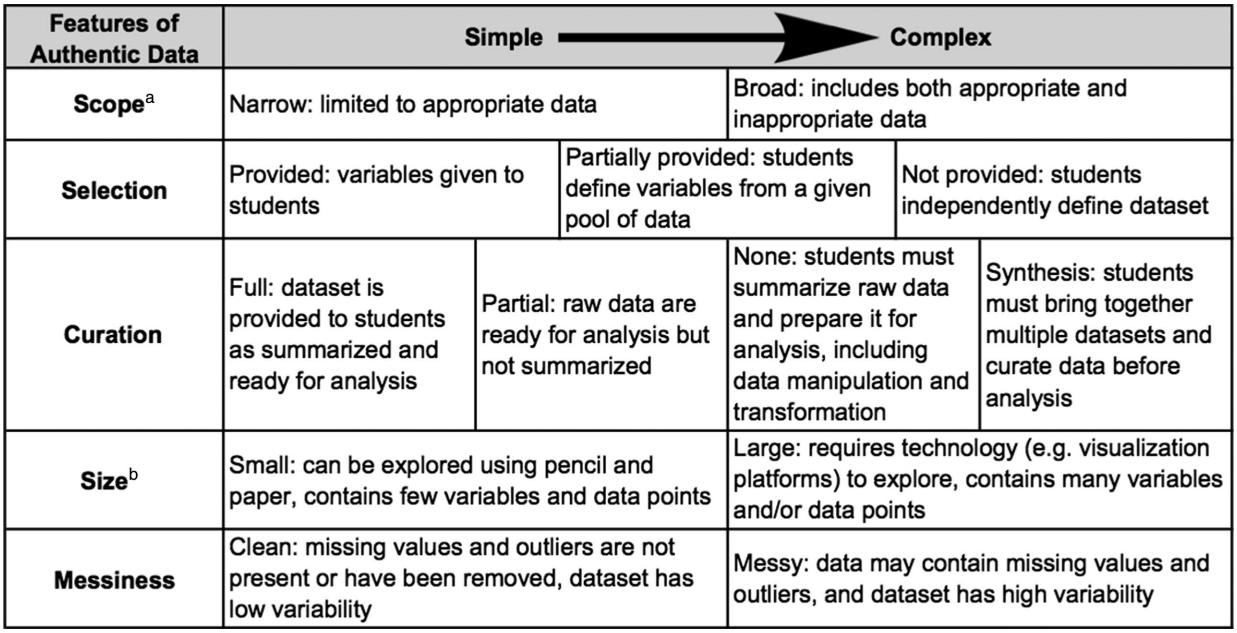
Features of Data Complexity

Data complexity has been shown to influence student learning and classroom discourse surrounding data. For example, in a study using earth science data, Kerlin and colleagues (2010) found that, while students had an easier time interpreting data from textbook graphs, those who used raw, complex data were more likely to identify patterns in data, make predictions, and evaluate the arguments made by fellow classmates. Additionally, compared with the use of graphs from textbooks, classroom discourse expanded when students worked with raw data (Kerlin *et al.*, 2010).

Authentic data sets vary in their complexity in several ways that impact how they are used in the classroom and the learning opportunities afforded. Here, we identify and define several features of authentic data that influence complexity: scope, selection, curation, size, and messiness (Table 1). We identified these features by reflecting on our own experiences working with data, through discussions at conferences and working groups, and through a review of the literature. To define each feature, we reference the existing literature, and if a discussion of complexity is available for a feature, we identify those citations in Table 1. In addition, we describe the unique learning opportunities surrounding data literacy that each feature engenders and make suggestions for instruction.

Scope. The scope of a data set is determined by the breadth of the information contained within it. Scope is determined by the number of variables represented and the amount of information contained within each variable. Simple data sets consist of few variables and contain only appropriate information for the scientific question being asked. Complex data sets may contain many variables and will provide both appropriate and inappropriate data (Berland and McNeill, 2010). Instructors can use data sets that are narrow in scope to provide experience identifying dependent and independent variables and the relationships between them. Moving to data sets that are broader in scope, students will face the additional challenge of determining which variables are necessary to address a scientific question. In addition, data sets with more variables will provide the opportunity for open-ended investigations that may lead to unanticipated research questions.

TABLE 1. Features of authentic data that can be used to characterize data-centric classroom activities



<table>
<tr><th>Features of Authentic Data</th><th colspan="12">Simple → Complex</th></tr>
<tr><td><b>Scope</b>[a]</td><td colspan="6">Narrow: limited to appropriate data</td><td colspan="6">Broad: includes both appropriate and inappropriate data</td></tr>
<tr><td><b>Selection</b></td><td colspan="4">Provided: variables given to students</td><td colspan="4">Partially provided: students define variables from a given pool of data</td><td colspan="4">Not provided: students independently define dataset</td></tr>
<tr><td><b>Curation</b></td><td colspan="3">Full: dataset is provided to students as summarized and ready for analysis</td><td colspan="3">Partial: raw data are ready for analysis but not summarized</td><td colspan="3">None: students must summarize raw data and prepare it for analysis, including data manipulation and transformation</td><td colspan="3">Synthesis: students must bring together multiple datasets and curate data before analysis</td></tr>
<tr><td><b>Size</b>[b]</td><td colspan="6">Small: can be explored using pencil and paper, contains few variables and data points</td><td colspan="6">Large: requires technology (e.g. visualization platforms) to explore, contains many variables and/or data points</td></tr>
<tr><td><b>Messiness</b></td><td colspan="6">Clean: missing values and outliers are not present or have been removed, dataset has low variability</td><td colspan="6">Messy: data may contain missing values and outliers, and dataset has high variability</td></tr>
</table>

Categories have been given for each feature, placed on a scale from simple to complex, based on the difficulty of the interaction for students. While features are represented in the table as discrete categories, they should instead be thought of along a continuum.

[a] From Berland and McNeill (2010).

[b] Modified from Berland and McNeill (2010) and Kastens *et al.* (2015).

Selection. Data selection involves making decisions as to which data are necessary to address a particular scientific question. The data-selection process will differ depending on whether students are working with first- or secondhand data, and instructors can use both types of data sets to provide students with the broadest experiences with this data feature. In activities that use secondhand data, students face challenges associated with data discovery, including finding and navigating online databases, locating appropriate data, evaluating the information they find, and interpreting variables and associated metadata files (Carlson *et al.*, 2011; Calzada Prado and Marzal, 2013; Langen *et al.*, 2014). In activities relying on firsthand data collection, students must decide which data they need to collect and the methods and tools required to do so. This process requires knowledge of the study system and how to design protocols for data collection. These abilities are captured by the quantification act, a component of quantitative reasoning that involves the process of assigning mathematical properties to an object such that it becomes data (Mayes *et al.*, 2014). The quantification act requires an understanding of appropriate units of measure and the proportional relationship of those units (Mayes *et al.*, 2014).

By giving students opportunities to choose which data are needed for a particular investigation, instructors are allowing students to take ownership of a lesson and determine the direction of their learning (Gould *et al.*, 2014). For secondhand data, instructors can build students' abilities in data selection by beginning with experiences in which students are provided with only the data necessary to answer a particular scientific question, followed by instances in which students must select the appropriate data from a larger pool (Table 1). Students can begin by selecting variables from a provided data set and then use open searches to identify data repositories and search independently. For firsthand data, complexity can increase as instructors remove protocols, or instructions, for data collection. By starting with guided inquiry and moving to open inquiry, students will take on increasing responsibility for defining the data they collect or the methods they use.

Curation. Data curation involves the cleaning and preparation of data sets for visualization and analysis. Instructors will often provide students with well-organized, ready-to-use data sets to ease analysis and interpretation (Grimshaw, 2015). However, when curation is left to the student, several learning opportunities are afforded, specifically in areas of data science and literacy (Carlson *et al.*, 2011). Also referred to as handling data (Calzada Prado and Marzal, 2013), curation involves several processes such as tidying up data, summarizing raw data, or synthesizing multiple data sets. Tidying up data involves manipulations and transformations to format a data set for analysis and visualization (Wickham, 2014). For example, in a tidy data set, the data have been organized such that each column represents a different variable and each row an observation of that variable (Wickham, 2014).

Statistics may be used when summarizing data to condense raw data down to measures such as sums, means, or measures of variability. Summarized data maintain authenticity so long as the curation process is done in a genuine way that reveals true patterns in data and does not obfuscate the truth. Thus, curation involves an understanding of ethical issues surrounding data and how data can be used transparently and honestly (Calzada Prado and Marzal, 2013). Providing students with opportunities to see and discuss how ethics play a role in data curation may lead them to be more critical consumers of data in their everyday lives.

Finally, synthesis of multiple data sets into one data set may be necessary depending on the research question. Synthesis requires organization of data so that they can be merged together. This includes working with data sets that may have asynchronous collection time frames, a mismatch of scale or location across variables, or missing data points or proper documentation. For larger data sets, curation involves familiarity with spreadsheets and coding software. By modeling the process of data synthesis, and providing students with opportunities to synthesize data themselves, instructors can help students develop this important data science skill that is necessary for working with data collected from multiple sources.

Size. The size of a data set is defined by the number of data points it contains (Berland and McNeill, 2010). Large data sets contain 1) many attributes, captured by the number of columns and 2) many cases, captured in rows. The size of a data set determines whether students can work with it by hand, using paper and pencil, or whether they must work with the data set digitally (Krumhansl *et al.*, 2012; Kastens *et al.*, 2015). Instructors can use small data sets to provide an entry point for students to have their first experiences working with data while practicing basic data-literacy skills, such as sketching simple graphs or using data as evidence to support a claim. Then, instructors can transition students to larger data sets to create opportunities for students to develop comfort and familiarity with digital tools, such as data visualization platforms or statistical programs. To create a link between these two data forms, firsthand student data can be nested within larger data sets, potentially through the use of citizen science databases and programs (Kastens *et al.*, 2015).

Messiness. Messiness is an important, yet often overlooked, feature of authentic data. The messiness of data sets is determined by the presence of variability, outliers, missing values, and unexpected trends. Messy data sets contain variability, which reflects both natural variation as well as variation from systematic error and precision of data-collection methods (Gould *et al.*, 2014). Additionally, messy data sets may be incomplete or may have missing values due to events that took place during a study.

Because of these qualities, messy data create several learning opportunities that may not be available in other classroom activities such as interpreting textbook figures (Kerlin *et al.*, 2010), polished graphs from published studies (Harsh and Schmitt-Harsh, 2016), or inauthentic data sets designed to demonstrate a specific concept. Instructors can provide messy data sets to create opportunities for critical thinking during data exploration. For example, students typically fail to recognize that variability is a common feature of data (Lawson, 1995) and interpret small amounts of variation between treatments as meaningful, without consideration to whether the variation is from experimental error or represents a true difference (Germann and Aram, 1996). However, when instructors provide students with messy data, students are able to explore sources of variability and become driven to explain unexpected patterns (Gould *et al.*, 2014).

Instructors can incorporate data-centric activities that highlight messy data in two forms: first- and secondhand data. Research has found that both sources lead to different learning experiences for students surrounding messiness, and the use of both first- and secondhand data can be complementary (Hug and McNeill, 2008). Firsthand data allow students to have direct experiences with data collection, leading to a better understanding of sources of variability, while secondhand data introduce increased complexity and more sources of variability beyond what can be collected in a classroom setting (Palincsar and Magnusson, 2001; Hug and McNeill, 2008; Langen *et al.*, 2014). Therefore, providing students with firsthand data experiences followed by more complex secondhand data sets might be an effective way to build their abilities in dealing with messy data (Kastens *et al.*, 2015).

Scaffolding Data Complexity

Use of complex, authentic data can be challenging for students of all ages, especially novices with few data or research experiences to draw from. However, students do not need to jump into experiences with complex data all at once. The use of simple data sets may be a good starting point for instruction, yet relying solely on simple data will quickly become repetitive. Instead, using data sets that vary in complexity may engage students in a broader array of scientific practices and provide increased diversity of learning opportunities.

We hypothesize that intentionally scaffolding the features of data complexity, by using incrementally more complex data sets over time, may facilitate the development of students' data-literacy abilities. Scaffolding, or providing instructional support, has been shown to be an effective strategy used to assist students in building complex skills (NRC, 2000). We suggest faded scaffolding, a modified version of scaffolding in which supports are gradually removed (McNeill *et al.*, 2006), as a strategy to help students build their abilities using and interpreting data. Over time, the complexity of data-intensive activities can be gradually increased to continue to challenge the students, while not moving beyond their current problem-solving abilities. Instructors can use formative assessments to monitor their students' abilities and select appropriate data sets to build complexity over time (Table 1).

While the five features are correlated to some degree, it is possible to isolate them to some degree as well. For example, instructors may choose to start with a small data set that is limited in scope when first increasing complexity in other areas, such as messiness. As students become more proficient in their understanding of variability and unexpected results, instructors may choose to increase complexity in another area, such as selection or size.

DISCUSSION

The integration of data into contemporary K–16 education is gaining attention, and today data literacy is broadly recognized as an important aspect of training students for modern careers and developing a data-conscious citizenry (Finzer, 2013; Baumer, 2015; NASEM, 2018). This is reflected in data analysis and interpretation becoming more commonplace throughout formal and informal education (Konold *et al.*, 2000; Metz, 2008; Speth *et al.*, 2010; Calzada Prado and Marzal, 2013) and the promotion of data literacy throughout undergraduate and K–12 reform efforts (NRC, 2012; College Board, 2013; NGSS Lead States, 2013; ACT, 2014; American Statistical Association, 2014; AAAS, 2015). In addition, as more and more data are being collected electronically and stored in open repositories (Hug and McNeill, 2008; Kastens *et al.*, 2015), opportunities to use authentic data in the classroom are increasing every day.

Yet, despite its importance and ubiquity, the use of authentic data in the classroom is limited by the availability of high-quality classroom resources, appropriate instructor training, and research to determine effective teaching strategies for developing student data literacy (Picone *et al.*, 2007; Metz, 2008; Gould *et al.*, 2014; Schultheis and Kjelvik, 2015; Harsh and Schmitt-Harsh, 2016; Angra and Gardner, 2017). We are currently at the beginning of an exciting shift in science education, creating targeted areas for collaboration for science educators, curriculum developers, programmers, and education researchers.

Resource Development and Educator Training

To use data effectively with their students, educators will require tools and training. Despite the prevalence of authentic data, teachers admit they sometimes still create inauthentic data sets to demonstrate a particular data science concept, resulting in the use of data that are devoid of context and lacking complexity. Yet these same teachers recognize their students' excitement when they know they are working with authentic data (Schultheis and Kjelvik, 2015). Additionally, teachers report that there are several limitations to incorporating data-centric activities into their classrooms, including the time commitment required to fully delve into complex data sets, which becomes particularly daunting when students have little prior experience with data analysis and interpretation (Konold *et al.*, 2000).

These reports from teachers point to a need for resource development to facilitate the introduction of data into K–12 and undergraduate education while also teaching core content. In recent years, resource developers have been responding. Many promising classroom activities and digital learning environments are being actively developed to facilitate student interactions with data (see Supplemental Table 1A). As more resources are developed, a few key challenges for the field become apparent, including the creation of 1) opportunities for educators, curriculum developers, and researchers to work together to design the strongest resources applicable to address current needs; and 2) a central repository for high-quality, data-rich educational resources.

After instructors have identified data sets and materials, they are not always confident in their abilities to lead data-focused discussions with their students (Konold *et al.*, 2000). Historically, training of preservice teachers has not included opportunities to work with complex data, resulting in teachers who are uncomfortable scaffolding their students' experiences graphing and interpreting data from inquiry (Bowen and Roth, 2005). Without training, the unique learning opportunities that arise from the use of authentic data may not be apparent to instructors. Therefore, in addition to resources, instructors will require training to adequately prepare them to teach with data, including the digital skills necessary to interact with digital data (Claro *et al.*, 2018).

Professional development offerings are a way to highlight potential benefits and guide instructors toward best practices when using authentic data in the classroom (Cooper *et al.*, 2015), and workshops should be offered to support instructors as they incorporate data literacy into their curricula (Wilson, 2013). In addition to professional development, there are professional and peer-mentoring networks to support instructors as they seek to increase their use of data in the classroom (Bonner *et al.*, 2017). At the university level, data-literacy programs are under development, and libraries are being considered as a key player in helping students and instructors build comfort and ability to use data (Calzada Prado and Marzal, 2013; Carlson *et al.*, 2015). Additionally, field stations and marine laboratories can facilitate direct communication between the science community and educators (Struminger *et al.*, 2018), providing opportunities for scientists to share their data and the stories behind their research with students and the public (Schultheis and Kjelvik, 2015).

Call for Research to Determine Authentic Data Best Practices

Just as Schield (2004) argued for further research to determine how information literacy, statistical literacy, and data literacy are related, researchers need to decipher key similarities and differences between quantitative reasoning, data science, and data literacy. Specifically, there appears to be considerable overlap in desired learning outcomes across fields, such as developing students' ability to interpret and use data in their everyday lives (Calzada Prado and Marzal, 2013). Similar to the work that has taken place in citizen science (Phillips *et al.*, 2018), research is necessary to define desired learning outcomes surrounding data use in the classroom, which will inform how these interrelated fields can be taught synergistically to improve student performance.

Once learning outcomes have been identified, future research can determine the best practices for improving data literacy for science, technology, engineering, and mathematics (STEM) students. Discussions within the science education community are ongoing and have resulted in suggested approaches for developing data-literacy skills. One idea put forth involves providing students with repeated practice using data, while prioritizing discussion of why they are collecting the data and how the data can be used to answer questions they are investigating (Gibson and Mourad, 2018). Additionally, ensuring these data experiences involve a diversity of data types and remain connected to the authentic context of the data could be crucial in developing student data literacy (Schultheis and Kjelvik, 2015). Yet few studies have tested the efficacy of educational materials using authentic data (Aikens and Dolan, 2014), and we are aware of no studies that compare the effects of authentic and inauthentic data on student learning outcomes. To address this need, future research should isolate the impact of data authenticity on student data literacy and interest and engagement in STEM by manipulating the use of authentic and inauthentic data resources in the classroom.

Additionally, although there are hypothetical advantages and disadvantages for working with paper or digital data activities, we are not aware of any research that documents how data-literacy learning opportunities differ between paper or digital media. In language literacy, research has demonstrated that students experience paper and digital media very differently, and the use of both might be complementary (Singer and Alexander, 2017a,b). Similarly, future research on data literacy could study the effects of paper versus digital data experiences on the learning opportunities afforded and the depth of student discourse when using data.

CONCLUSION

The challenges brought about by data's increasing role in science and society are being met by a shift in science education. Through experiences collecting, analyzing, and interpreting authentic data, students have the potential to become better critical thinkers, increase their scientific content knowledge, and understand the value of data for comprehending the natural world (Mourad *et al.*, 2012; Langen *et al.*, 2014). Despite this progress, research is necessary to define learning outcomes and best practices when using data in the classroom. Additionally, curricular development and increased accessibility of resources is necessary, as instructors lack access to high-quality, data-intensive resources and the training necessary to effectively implement such resources. Though these challenges may seem daunting, we are currently at the beginning of an exciting new phase in science education, one that comes with the opportunity to increase the data competencies of the next generation of scientists, workers, and citizens.

ACCESSING MATERIALS

Websites for data-intensive classroom resources and other materials cited in this paper can be found in Supplemental Table 1A.

ACKNOWLEDGMENTS

We give special thanks to Julie Morris, Ariel Cintron-Arias, Laurel Hartley, Kristin Jenkins, Robert Mayes, Louise Mead, Paul Strode, Molly Stuhlsatz, Gordon Uno, and Jeremy Wojdak for thought-provoking conversations at the Data Nuggets National Institute for Math and Biology Synthesis (NIMBioS) working group (National Science Foundation [NSF] DBI 1300426). We also thank the organizers of the 2017 Data Science Education Technology conference, along with all the participants for sharing great insights into the state of the field. Thank you to Cheryl Hach, May Lee, Paul Strode, and the anonymous reviewers who provided useful feedback on earlier versions of this article. Many ideas expressed in this article are related to our experiences developing Data Nuggets at Michigan State University. We received funding and support for Data Nuggets from the Kellogg Biological Station (KBS) GK–12 Project (NSF DGE 0947896), KBS Long-Term Ecological Research Program K–12 Partnership (NSF DEB 1027253), and BEACON Center for the Study of Evolution in Action (NSF DBI 0939454). Data Nuggets are currently funded by NSF DRK-12 1503211. We also thank teachers Marcia Angle, Cheryl Hach, Ellie Hodges, and Kristy Campbell for their valuable insights and work identifying key areas needed in curricular development. This is KBS Publication #2008.

REFERENCES

ACT. (2014). *ACT college and career readiness standards: Science*. Retrieved January 1, 2018, from www.act.org/content/dam/act/unsecured/documents/CCRS-ScienceStandards.pdf

Aikens, M. L., & Dolan, E. L. (2014). Teaching quantitative biology: Goals, assessments, and resources. *Molecular Biology of the Cell*, *25*, 3478–3481.

American Association for the Advancement of Science (AAAS). (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC. Retrieved December 13, 2017, from www.visionandchange.org/

AAAS. (2015). *Vision and change in undergraduate biology education: Chronicling change, inspiring the future*. Washington, DC. Retrieved January 1, 2018, from www.visionandchange.org/

American Statistical Association. (2014). *Curriculum guidelines for undergraduate programs in statistical science*. Retrieved December 15, 2017, from www.amstat.org/education/curriculumguidelines.cfm

Angra, A., & Gardner, S. M. (2017). Reflecting on graphs: Attributes of graph choice and construction practices in biology. *CBE—Life Sciences Education*, *16*(3), ar53.

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *American Statistician*, *69*(4), 334–342.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765–793.

Boersma, S., & Klyve, D. (2013). Measuring habits of mind: Toward a promptless instrument for assessing quantitative literacy. *Numeracy: Advancing Education in Quantitative Literacy*, *6*(1), ar6.

Bonner, K. M., Fleming-Davies, A. E., Grayson, K. L., Hale, A. N., Wu, X. B., & Donovan, S. (2017). Bringing research data to the ecology classroom through a QUBES faculty mentoring network. *Teaching Issues and Experiments in Ecology*, *13*. Retrieved April 11, 2018, from http://tiee.esa.org/vol/v13/issues/commentary.html

Borges-Rey, E. L. (2017). Data literacy and citizenship: Understanding “big data” to boost teaching and learning in science and mathematics. In Ramírez-Montoya, M.-S. (Ed.), *Handbook of research on driving STEM learning with educational technologies* (pp. 65–79). Hershey, PA: IGI Global.

Bowen, G. M., & Roth, W. M. (2005). Data and graph interpretation practices among preservice science teachers. *Journal of Research in Science Teaching*, *42*(10), 1063–1088.

Calzada Prado, J., & Marzal, M. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, *63*(2), 123–134.

Carlson, J. R., Fosmire, M., Miller, C., & Sapp Nelson, M. R. (2011). Determining data information literacy needs: A study of students and research faculty. *Libraries Faculty and Staff Scholarship and Research*, Paper 23.

Carlson, J. R., Sapp Nelson, M. R., Johnston, L. R., & Koshoffer, A. (2015). Developing data literacy programs: Working with faculty, graduate students and undergraduates. *Bulletin of the Association for Information Science and Technology*, *41*(6), 14–17.

Claro, M., Salinas, A., Cabello-Hutt, T., San Martín, E., Preiss, D. D., Valenzuela, S., & Jara, I. (2018). Teaching in a digital environment (TIDE): Defining and measuring teachers’ capacity to develop students’ digital information and communication skills. *Computers & Education*, *121*, 162–174.

College Board. (2013). *AP Biology: Course and exam description* (rev. ed., effective Fall 2012). Retrieved December 4, 2017, from www.collegeboard.org

Common Core State Standards Initiative (CCSSI). (2014). *Common Core State Standards for Mathematics*. Retrieved January 2, 2018, from www.corestandards.org

Conway, D. (2010). *The data science Venn diagram*. Retrieved December 13, 2017, from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., … Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, *350*(6258), 281–282.

Doering, A., & Veletsianos, G. (2007). An investigation of the use of real-time, authentic geospatial data in the K–12 classroom. *Journal of Geography*, *106*, 217–225.

Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. New York: Teachers College Press.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2). Retrieved April 15, 2017, from https://escholarship.org/uc/item/7gv0q9dc

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching*, *33*(7), 773–798.

Gibson, J. P., & Mourad, T. (2018). The growing importance of data literacy in life science education. *American Journal of Botany*, *105*(12), 1–4.

Gould, R., Sunbury, S., & Dussault, M. (2014). In praise of messy data. *Science Teacher*, *81*(8), 31.

Grimshaw, S. D. (2015). A framework for infusing authentic data experiences within statistics courses. *American Statistician*, *69*(4), 307–314.

Harsh, J. A., & Schmitt-Harsh, M. (2016). Instructional strategies to develop graphing skills in the college science classroom. *American Biology Teacher*, *78*(1), 49–56.

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences USA*, *112*(36), 11199–11204.

Hug, B., & McNeill, K. L. (2008). Use of first-hand and second-hand data in science: Does data type influence classroom conversations? *International Journal of Science Education*, *30*, 1725–1751.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410–1412.

Kastens, K. A., Krumhansel, R., & Baker, I. (2015). Thinking big—Transitioning your students from working with small, student-collected data sets towards “big data.” *Science Teacher*, *82*(5), 25–31.

Kerlin, S. C., McDonald, S. P., & Kelly, G. J. (2010). Complexity of secondary scientific data sources and students’ argumentative discourse. *International Journal of Science Education*, *32*(9), 1207–1225.

Konold, C. E., Coulter, R., & Feldman, A. (2000). Engaging students with data. *Learning & Leading with Technology*, *28*(3), 50–55.

Krumhansl, R., Foster, J., Busey, A., Baker, I., & DeLisi, J. (2012). *Visualizing oceans of data: Designing educational interfaces*. Waltham, MA: Education Development Center.

Langen, T. A., Mourad, T., Grant, B. W., Gram, W. K., Abraham, B. J., Fenrnandez, D. S., … Hampton, S. E. (2014). Using large public datasets in the undergraduate ecology classroom. *Frontiers in Ecology and the Environment*, *12*, 362–363.

Lawson, A. E. (1995). *Science teaching and the development of thinking*. Belmont, CA: Wadsworth.

Magnusson, S. J., Palincsar, A. S., Hapgood, S., & Lomangino, A. (2004). How should learning be structured in inquiry-based science instruction? Investigating the interplay of 1st- and 2nd-hand investigations. In *Proceedings of the 6th international conference on learning sciences* (pp. 318–325). Alpharetta, GA: International Society of the Learning Sciences.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, *42*(1), 30–37.

Mayes, R. L., Forrester, J. H., Christus, J. S., Peterson, F. I., Bonilla, R., & Yestness, N. (2014). Quantitative reasoning in environmental science: A learning progression. *International Journal of Science Education*, *36*(4), 635–658.

McNeill, K. L., Lizotte, D. J., Krajcik, J. S., & Marx, R. W. (2006). Supporting students’ construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153–191.

Metz, A. M. (2008). Teaching statistics in biology: Using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. *CBE—Life Sciences Education*, *7*(3), 317–326.

Mourad, T., Grant, B. W., & Gram, W. K. (2012). Engaging undergraduate students in ecological investigations using large, public datasets. *Teaching Issues and Experiments in Ecology*, *8*.

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. Washington, DC: National Academies Press. https://doi.org/10.17226/25104

National Research Council (NRC). (1996). *National science education standards*. Washington, DC: National Academies Press.

NRC. (2000). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academies Press.

NRC. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning. In Carver, S., & Klahr, D. (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 151–193). Mahwah, NJ: Erlbaum.

Phillips, T., Porticella, N., Constas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science: Theory and Practice*, *3*(2), 1–19.

Piatek-Jimenez, K., Marcinek, T., Phelps, C. M., & Dias, A. (2012). Helping students become quantitatively literate. *Mathematics Teacher*, *105*(9), 692–696.

Picone, C., Rhodes, J., Hyatt, L., & Parshall, T. (2007). Assessing gains in undergraduate students' abilities to analyze graphical data. *Teaching Issues and Experiments in Ecology*, *5*, 1–54.

Schield, M. (2004). Information literacy, statistical literacy and data literacy. *IASSIST Quarterly*, *28*(2), 6–11.

Schultheis, E. H., & Kjelvik, M. K. (2015). Data Nuggets: Bringing real data into the classroom to unearth students' quantitative and inquiry skills. *American Biology Teacher*, *77*(1), 19–29.

Singer, L. M., & Alexander, P. A. (2017a). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *Journal of Experimental Education*, *85*(1), 155–172.

Singer, L. M., & Alexander, P. A. (2017b). Reading on paper and digitally: What the past decades of empirical research reveal. *Review of Educational Research*, *87*(6), 1007–1041.

Sorgo, A. (2010). Connecting biology and mathematics: First prepare the teachers. *CBE—Life Sciences Education*, *9*, 196–200.

Speth, E. B., Momsen, J. L., Moyerbrailean, G. A., Ebert-May, D., Long, T. M., Wyse, S., & Linton, D. (2010). 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. *CBE—Life Sciences Education*, *9*(3), 323–332.

Steen, L. A. (2004). *Achieving quantitative literacy: An urgent challenge for higher education*. Washington, DC: Mathematical Association of America.

Stoddart, T., Solis, J., Tolbert, S., & Bravo, M. (2010). A framework for the effective science teaching of English language learners in elementary schools. In Sunal, D. W., Sunal, C. S., & Wright, E. L. (Eds.), *Teaching science with Hispanic ELLs in K–16 classrooms* (pp. 151–181). Charlotte, NC: Information Age Publishing, Inc.

Struminger, R., Zarestky, J., Short, R. A., & Lawing, A. M. (2018). A framework for informal STEM education outreach at field stations. *BioScience*, *68*(12), 969–978.

Vacher, H. L. (2014). Looking at the multiple meanings of numeracy, quantitative literacy, and quantitative reasoning. *Numeracy*, *7*(2), ar1.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23.

Wilson, S. M. (2013). Professional development for science teachers. *Science*, *340*(6130), 310–313.

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2017). Creating an understanding of data literacy for a data-driven society. *Journal of Community Informatics*, *12*(3), 9–26.