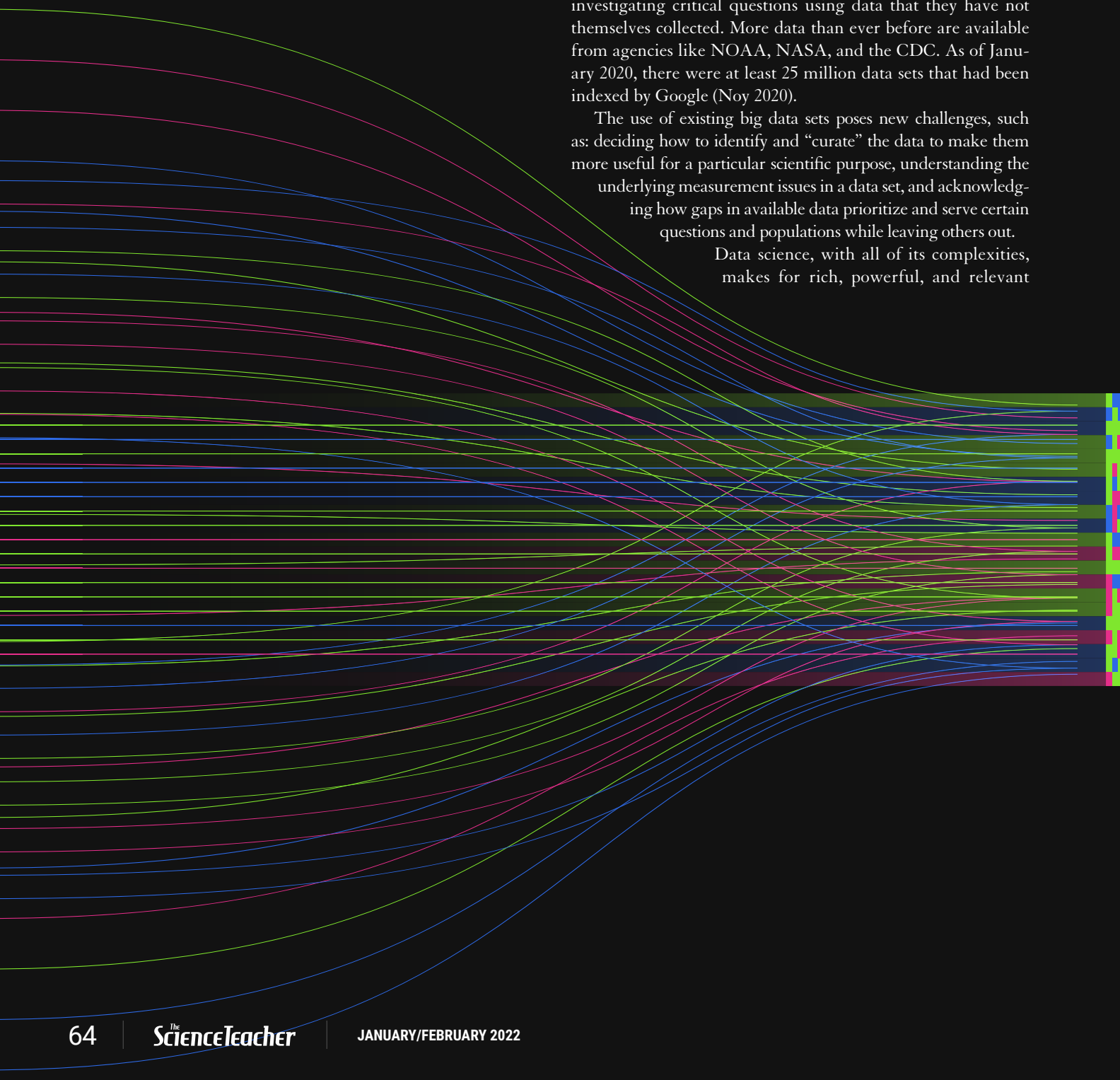
Data Science Meets

JACOB SAGRANS, JANICE MOKROS, CHRISTINE VOYER, AND MEGGIE HARVEY

The use of large, open-source data sets is ubiquitous in scientific research. Scientists—ranging from meteorologists to chemists to epidemiologists—are researching and investigating critical questions using data that they have not themselves collected. More data than ever before are available from agencies like NOAA, NASA, and the CDC. As of January 2020, there were at least 25 million data sets that had been indexed by Google (Noy 2020).

The use of existing big data sets poses new challenges, such as: deciding how to identify and "curate" the data to make them more useful for a particular scientific purpose, understanding the underlying measurement issues in a data set, and acknowledging how gaps in available data prioritize and serve certain questions and populations while leaving others out.

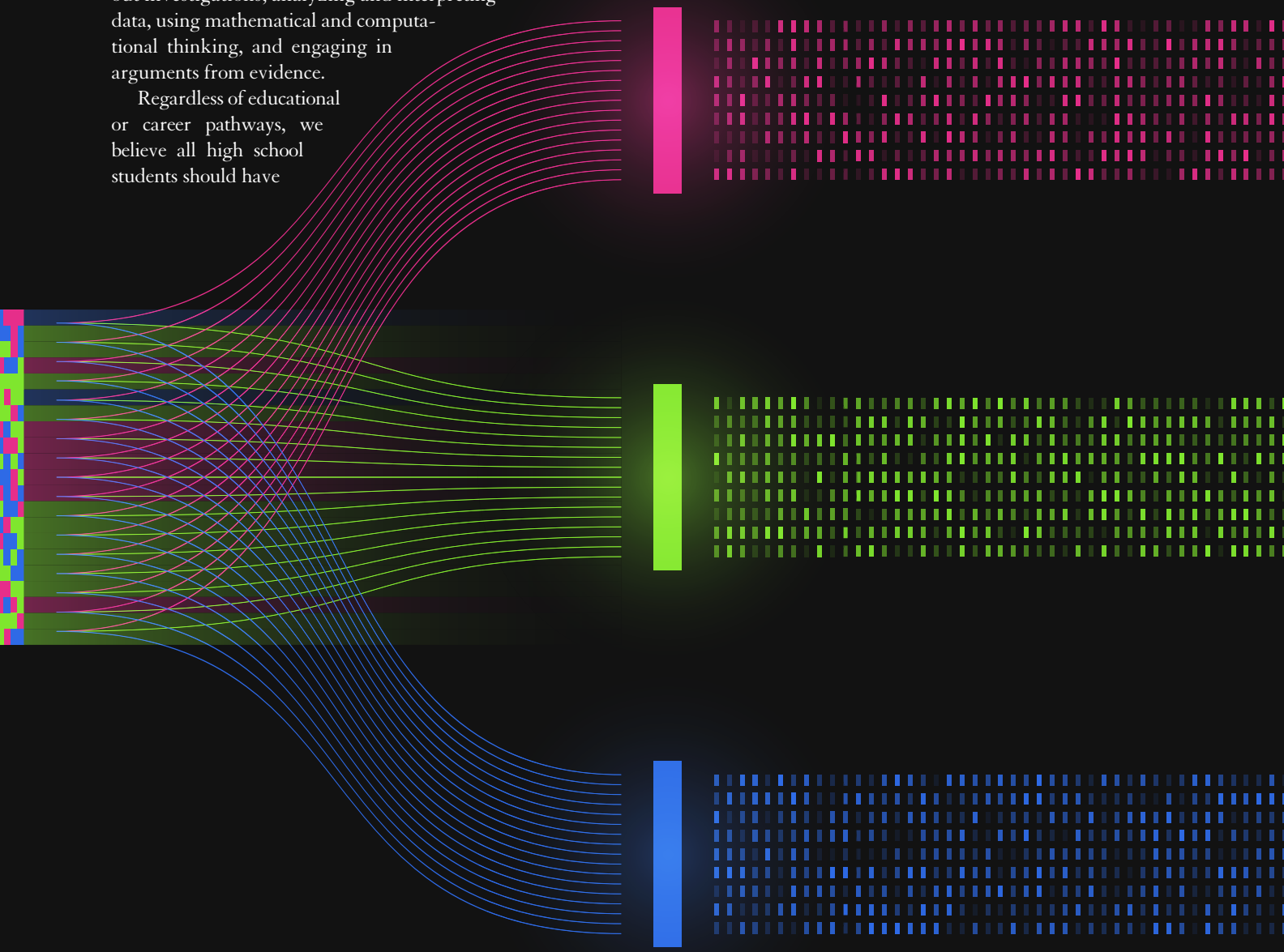
Data science, with all of its complexities, makes for rich, powerful, and relevant



Science Teaching

learning in science classrooms. Despite the potential for learning, data science is not well represented in today's science classroom, and resources to support teachers and students in this work are just beginning to emerge. The practices that are the essence of data science are also the essence of the science practices promulgated by the *NGSS* (NGSS Lead States 2013). Examining existing data sets provides ideal opportunities for asking good questions, planning and carrying out investigations, analyzing and interpreting data, using mathematical and computational thinking, and engaging in arguments from evidence.

Regardless of educational or career pathways, we believe all high school students should have experience examining and posing questions about data. Data science is not only critical to the work of professional scientists but also instrumental in building an informed citizenry. Understanding data about COVID-19, for example, leads to better assessment of health risks that are changing on a daily basis. Data science also helps students understand and address social and racial disparities in science, health, the environment, and policy.



An excellent example of this is discussed in the book *Data Feminism*: Tennis star Serena Williams gathered data about racial disparities in pregnancy outcomes after she experienced childbirth complications and discovered that black women are three times more likely than white women to die of complications from pregnancy and childbirth (D'Ignazio and Klein 2020).

To contribute to the growing effort to bring data science into classrooms, we have been implementing the NSF-funded "Data Clubs" project to examine using data sets on topics such as ticks and Lyme disease, COVID-19, and sports and leisure injuries. Much of this work takes place with youth in out-of-school settings. In addition to developing modules for youth, we worked with a group of 18 high school science and computer science teachers from Maine, New Hampshire, and Massachusetts who participated in a virtual 15-hour workshop series on data science education over the summer of 2020. The goal of the workshop was to introduce teachers to real and complex data sets, models for scaffolding learning, and tools for working with those data sets. In this article we share some of the key findings from this effort.

How do we support data science learning?

According to Erickson (2021), author of a pioneering text in the field, data science involves two basic propositions: "1) A data science problem often begins with a feeling of being awash in data; 2) Data science uses data moves to manipulate data." Being awash means navigating through a sea of turbulent data by making wise choices that enable one to focus on their questions and chart the course of their own research. Data moves are defined as actions one can perform on a data set to alter its content, values, structure, or representations to make sense of the data. For example, a student may wish to examine only a subset of data, which means filtering the data, or to combine variables to make them easier to understand or have them reveal something new.

Web-based tools like the Common Online Data Analysis Platform (CODAP), which we have used in our project, or others, such as Tuva, make it easier for students to ask their own questions about a data set and then engage in data moves to address these questions. After learning a few moves, and without any coding skills, students can begin to dig into their data, a feat that would be much harder to do when plotting data by hand.

The goal of our project's teacher professional learning experience was to provide resources and introduce new tools and skills so that teachers could begin to engage students in data science. A significant part of this work was engaging teachers as data investigators, modeling experiences that they might bring to students to promote posing and exploring questions with data, looking at patterns, and making interpretations.

A focus of the teacher workshop was investigating current and meaningful questions using open-source data sets. We focused on a different data set each day, covering various scientific and social science topics. In the context of these data sets, it is essential to start by building background knowledge of the topic itself. This could include both pooling existing knowledge of students or teachers (or both) and doing additional research to uncover what is known and what gaps in knowledge exist. This groundwork supports investigators as they begin to meaningfully explore the data based on what they already know about the topic.

We believe that a critical place to start each data exploration is by evaluating the structure of the data set, including identifying the cases and attributes (or variables) that are included. In the

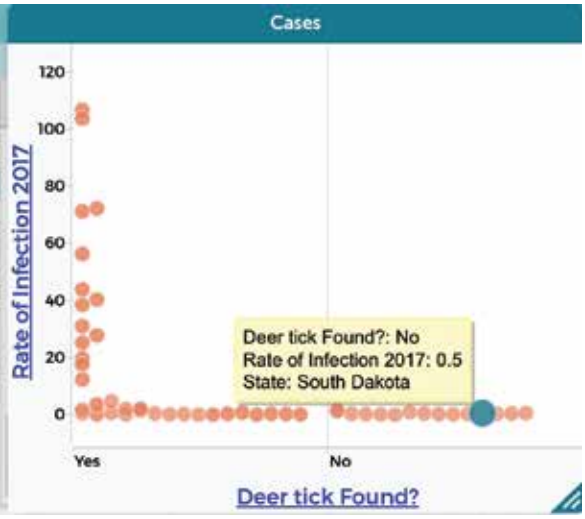
FIGURE 1

Ticks and Lyme disease data set.

Lyme_dataset1_121218 2

Cases (51 cases)

index	State	Boundary	Rate of ...ion 2017	Percent ...st cover	latitude	longi-tude	Deer tick Found?
40	South C...		0.3	63.8	33.84	-81.16	Yes
41	South D...		0.5	3.1	43.97	-99.9	No
42	Tenness...		0.2	52.9	35.52	-86.58	Yes
43	Texas		0.1	7	31.97	-99.9	Yes
44	Utah		0.3	8.9	39.32	-111.09	No
45	Vermont		103.6	75.7	44.56	-72.58	Yes
46	Virginia		12.3	60.7	37.43	-78.66	Yes
47	Washin...		0.4	40.7	47.75	-120.74	No
48	West Vir...		27.7	77.2	38.6	-80.45	Yes
49	Wiscon...		31	45.2	43.78	-88.79	Yes
50	Wyomi...		0.5	9.2	43.08	-107.29	No
51	U.S. Inci...		9.1				

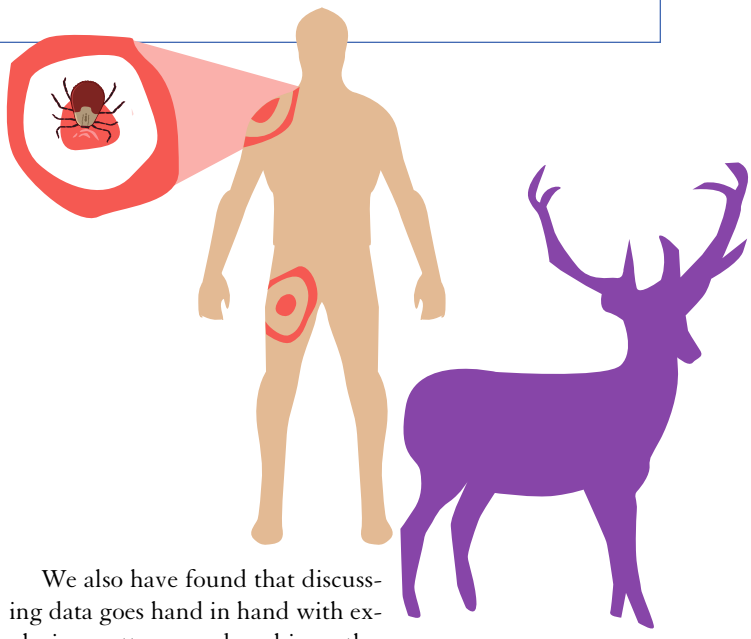


simple data set on ticks and Lyme disease discussed below, each case is a state and the attributes include presence/absence of deer ticks in the state as well as the rate of Lyme disease in that state.

After considering how the data are structured, the next critical step is to consider the affordances and limitations of the data, paying particular attention to measurement issues inherent in the data and that emerge because of choices made during data production. For example, in the ticks and Lyme disease data set (see Figure 1), cases of Lyme are reported to public health authorities and tallied in the state of diagnosis, even if a person might have contracted the disease while out of state.

After getting the lay of the land, the next step is to explore by dragging and dropping attributes onto the graph in CODAP. In the data set in Figure 1, one can address questions like: "What is the relationship between Lyme disease and geographical location?" or "What is the relationship between the presence of deer ticks in the state and the rate of Lyme disease?" While the latter seems like an easy question, there is Lyme disease in places where there are no deer ticks, as can be seen in Figure 1 (see also the data in CODAP at *http://bit.ly/lyme17*). Based on our evaluation of the data set, this could result from the way the data are collected and reported.

With data tools, one can ask questions in an iterative fashion, quickly making new graphs to examine relationships between attributes. One can see graphs of whatever relationships one wants to examine. To help build facility with CODAP, teachers and students use the project-designed CODAP Challenge Cards. These cards are designed to support making data moves such as selecting, summarizing, grouping, and filtering in the context of any data set.



We also have found that discussing data goes hand in hand with exploring patterns and making other data moves. Discussions often arise spontaneously as new patterns are found, and they help investigators articulate their thinking, identify additional elements of the data that they had not yet considered, and figure out explanations for the relationships they found in a collaborative forum, just as professional scientists do. Having the chance to challenge one another's thinking enhances collective understanding of the data and what can be gleaned from it. In many cases, discussion arises about definitions of the variables, the meaning and handling of missing data, and the ways in which a particular approach to measuring influences findings. (For example, consider how cases of Lyme disease are assigned to states, as discussed above.)

Bringing data into the classroom

Teachers came to the 15-hour workshop with a wide range of experiences with data science. A few had worked with advanced science students on coding, but most had not used large data sets before, and none had used CODAP. Evaluation results indicate that teachers with a broad range of experience found these data investigations were relevant and accessible. More importantly, teachers came away from the workshop with solid plans for engaging their students as data investigators and incorporating data science into their teaching. In their planning documents and in individual follow-up interviews, teachers indicated that they were doing the following:

- Six teachers planned to use COVID-19 data sets introduced during the workshop. In one case, the teacher incorporated this data into an existing unit on infectious disease, while the others added this content to their courses and described it as being engaging to students. One teacher noted the relevant connection between this data and her students' study of exponential growth in algebra.
- Four teachers planned to have students look at larger data sets as part of their work in science fairs. At least two teachers planned to incorporate engineering data as part of students' projects.
- At least five teachers identified intriguing data sets not touched upon during the workshop, primarily data sets related to their perceptions of students' interest. These include data on Sumo wrestlers, seal strandings, crime, and moose hunting. Teachers indicated that they were drawn to data that had the highest relevance to their students as well as a good connection to course content.

In addition to surveying teachers about their plans, we subsequently followed two teachers as they implemented data projects. These case studies are summarized below.

Jake Bogar, who works with ninth-grade pre-engineering students, used data sets to get his students interested in how to improve engineering design, based on the results of previous design experiments. Jake reasoned that existing data provide a foundation for doing one's own design work. He used a data set on bridge failures (Naser 2019), which focused on the question of whether past failures can help identify bridges that are vulnerable to extreme events. Jake's examination of the data set revealed that there were a great number of fire failures in the last 20 years on steel, composite, and concrete bridges—not just on wooden bridges.

Jake introduced the bridge data after he reviewed CODAP with his students. After students examined the data, he posed these questions:

- How does the frequency and location of bridge failures change over time?
- What is the most common mode of bridge failure?
- How could this data set be used to predict bridge failures?

Students addressed these questions by making a series of graphs in CODAP. Jake reported that students were especially interested in the type of bridge failure as it related to other variables. Like Jake, students were surprised by the high incidence of recent bridge fires. This led to a discussion of the interaction between year, type of failure, and bridge composition. Students were able to explore the relationship between failure and other bridge characteristics, and to see that failures may have more than one cause. Students also began differentiating causality from correlation.

Bringing data science to high school classrooms will transform science courses but will take time. Teachers in our workshop compared introducing data science to the process of introducing engineering design, noting that it took several years to feel comfortable with the integration of engineering. The same could happen with data science.

A second teacher, Leslie Marquis, teaches biology in a high school in northern Maine. The class where she did her data project comprised 10 students, most of whom were boys who were struggling learners. Her goal was to teach a curricular unit about ecosystems, making use of data about moose, because this species is familiar to her students. Leslie knew about public data sets on moose harvests, collected by the Maine Department of Inland Fisheries and Wildlife. She was in an ideal position to undertake this project, due to the convergence of her curricular goal (teaching about the relationship between hunting permits and moose population), the informal knowledge of her students (which Leslie said was "part folklore, part based on experience"), and public data sets.

Leslie's potential questions included:

- How is the size of the moose harvest related to the time periods during which hunting is allowed?
- How has the number of permits changed over time, and what attributes relate to this?
- What is the distribution of the ages and gender of harvested moose?

Leslie ultimately decided to focus on age and gender of harvested moose because it was a good starting point for relatively inexperienced students. Leslie reported that students were surprised about the distribution of moose ages at time of harvest, and they immediately offered possible explanations. Students noticed the wide range of ages (from 6 months to 17.5 years) and the modal age at harvest of 2.5 years. They speculated that younger moose were more frequently killed because they were inexperienced, and easily drawn to the fake moose calls of hunters.

Leslie felt she could have done more with the data, but that she was already stretching her students' thinking. She had accomplished her goals of helping students see the value of data in studying ecosystems and of integrating this data with their informal knowledge of hunting.

More broadly, two additional findings emerged. First, nearly all of the teachers who participated in the workshop said that a motivator for students in exploring data sets was the *immediacy* of getting results, which in turn freed up time for interpretation of findings and asking next questions. One teacher said: "Using data this way is more like sports. It's fast-moving with quick feedback." Second, many teachers were challenged by having too many data sets. Whatever topic they chose to explore, they found hundreds of data sets and they did not have time to find data sets that were just right for their students in terms of size and complexity.

Bringing data science to high school classrooms will transform science courses but will take time. Teachers in our workshop compared introducing data science to the process of introducing engineering design, noting that it took several years to feel comfortable with the integration of engineering. The same could happen with data science. These teachers also felt that the transformation would enable students to gain a deeper understanding of science practices and of the nature of the work that scientists do. We are in the midst of an exciting, timely, and critical evolution of STEM teaching and learning. ■

ACKNOWLEDGMENTS

We would like to thank the 18 teachers who participated in our workshop and follow-up interviews/work, especially Jake Bogar and Leslie Marquis, who participated in extended teacher case studies. We would also like to thank Andrew Allyn, Tim Erickson, Traci Higgins, Leigh Peake, Ada Ren, Andee Rubin, and Tracey Wright for their work on the professional development workshop and the broader project. This project is funded by the National Science Foundation, grant nos. DRL-1742255 and 1917653. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Common Core State Standards in Mathematics. 2020. *http://www.corestandards.org/Math/*

Concord Consortium. 2022. Common Online Data Analysis Platform (CODAP). *https://codap.concord.org/*

Data Clubs. 2022. *https://www.terc.edu/dataclubs/*

D'Ignazio, C., and L. Klein. 2020. *Data feminism*. Cambridge, MA: MIT Press. *https://data-feminism.mitpress.mit.edu/*

Erickson, T. 2021. *Awash in data. https://codap.xyz/awash/*

Maine Department of Inland Fisheries and Wildlife. 2021. Harvest information. *https://www.maine.gov/ifw/hunting-trapping/harvest-information.html*

Naser, M.Z. 2019. Can past failures help identify vulnerable bridges to extreme events? A biomimetical machine learning Approach. *Engineering with Computers. https://doi.org/10.1007/s00366-019-00874-2*

NGSS Lead States. 2013. *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press. *https://www.nextgenscience.org/next-generation-science-standards*

Noy, N. 2020. Discovering millions of data sets on the web. The Keyword *https://blog.google/products/search/discovering-millions-data sets-web/*

Society for Science and the Public. 2020. Research at home: Large data sets. *https://www.societyforscience.org/research-at-home/large-data-sets/*

Jacob Sagrans (*jsagrans@scieds.com*) is is Senior Research Associate and **Janice Mokros** is Senior Research Scientist at Science Education Solutions, Los Alamos, NM; **Christine Voyer** is Science Education Senior Program Manager and **Meggie Harvey** is Science Curriculum Specialist at Gulf of Maine Research Institute, Portland, ME.