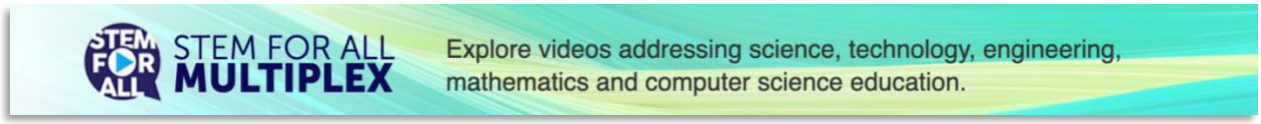

Leveraging Authentic Data Across STEM Curricula
Multiplex Monthly Theme
September 2020

> Data fluency is emerging as an essential part of modern education, not only for science, technology, engineering, and mathematics (STEM) in the U.S., but for every domain in a data-driven world. Our lives have become surrounded— often even governed—by data.... Data literacy was 20th century. Data fluency is now. (Concord Consortium, "Thinking and doing with data")

This Theme was highlighted by an expert panel of researchers and educators who have been centrally focused for many years on students' engagement with authentic data. (Note: The panel presentation was accompanied by a very active chat session, which is available, together with a collection of the resources suggested by the panel and chat participants.)

The panel included moderator Randy Kochevar, the director of EDC's Oceans of Data Institute (ODI); Chad Dorsey, President and CEO of the Concord Consortium, former teacher educator, curriculum developer, and science instructor; Margo Murphy, in her 35th year teaching and currently at Camden Hills Regional High School (Rockport, ME); Andee Rubin, a mathematician and computer scientist now at TERC, who has been combining expertise in technology design, math education, and artificial intelligence to improve math and science education both in and out of school for over 40 years; and Suyen Machado, a former classroom teacher, instructional coach, professional development facilitator, and administrator for LAUSD, now the Data Science Education Project Director at UCLA's Center X.

Even though the importance of data has long been known — and enshrined in standards for science and mathematics education for decades — these experts agreed that the present time is different in important ways from times past. As Randy Kochevar said in his blog, preparing for the expert panel, the ubiquity of data in every scholarly field of inquiry, from science to sociology, from art to archeology as well as astronomy has been evident for some time. Now, however, in order to navigate many aspects of daily life, people need to know something about how to think about data, and understand "the life of data" — how data are collected, curated, explored, and interpreted to represent and investigate some aspect of the world. At the time of this Theme, issues of public health with their ramifying links to income inequality, K-12 education, unemployment, and so on require a public that can make sense of the claims and counterclaims, the shifting evidence and interpretations, that bombard us — but are shaping our future. If we are to play any active role as participants — citizens — we need to be equipped to make judgments, ask questions, and make decisions.

So the world situation provides lots of motivation for learning, and data science education has evolved to the point that educators and students can be well placed to meet and understand the issues of our times. As Andee Rubin said in her presentation, data educators have known all along that "data can be inherently interesting to students." Data are not just numbers, but numbers with context. As several of the panelists suggested, students are easily engaged with the contextual matters of their lives and their world, so the challenge for education is to honor that engagement, and help students see how data fluency serves and even enhances their curiosity, and equips and develops their agency as learners and as members of society.

Finding and using authentic data — and addressing the barriers to its use

Andee Rubin pointed out that one thing that is different about the educational landscape is that there is much greater access to different kinds of data for learners, and teachers. Data are accessible from many fields, including census and other social science data, data from citizen science projects (such as Cornell's eBird data sets), and visual and numerical data from NASA, NOAA, and other government agencies. Chad Dorsey noted that educators are beginning to tune in, also, to research that examines text, art, or music for enlightening statistical patterns. And of course, students can collect data about themselves or matters that are important to them or their communities, to see what the data can reveal about themselves.

Data sets that were accessible in the past tended to be "simple" — small, and univariate. Now, however, authentic data sets, such as those enumerated above, tend to have many more variables, and be much more complex. Indeed, one of the major challenges for data science in schools is one that is often overlooked, in the excitement of having so much data available to work with — and that is bringing these data sets into a form that is usable for students, depending on their age and level of sophistication. Andee Rubin commented that these complex sets require operations that Erickson et al. call "data moves" -- reformatting, restructuring, or otherwise reshaping the data sets for use by students (this paper is available as one of the Resources for this Theme). This in turn requires the embracing of computational tools by teachers, curriculum developers, or other "curators" to clean and prepare the data so that students can work with them.

There are several barriers to students' access to good quality data sets. First, it can be time-consuming to find and evaluate data sets relevant to the curriculum. Moreover, it is easier to find good data in some fields than others; biology was mentioned as an example. With equity of access in mind, a basic issue is cost — the data sets need to be low-cost or freely available. But as mentioned above, finding the data is just the beginning. Promising data sets still need curation and customization for use in the classroom. (And I note that one can select a promising data set, begin the curation work, and discover that it is not really suitable after all.)

As Suyen Machado pointed out, there are well-known advantages to student-collected data. For one thing, the students are able to understand in detail the life of the data they are working with — how it was collected, under what conditions, what the potential sources of error were in the data collection, the challenges of taking reliable data, etc. The real world is messy and data as collected tends to reflect that messiness and variability; the richness of context available to students for the data they collect can help them understand how to work with the complexity. If students and teachers don't have the tools and techniques necessary for evaluating the quality of their data, student misconceptions can be perpetuated, as they struggle to make sense of the complexity and uncertainty they encounter.

There are some tools that are designed to support student agency and curiosity, by allowing flexible, accessible, yet powerful ways to represent data sets in various forms, and to "wander through" the data sets, and play with them to understand relationships, patterns, and variations. Panelists Machado and mentioned TuvaLabs (https://tuvalabs.com), and Concord Consortium's CODAP (https://codap.concord.org) as examples of environments that support student data work of increasing sophistication.

More and more, teachers are helped in their incorporation of data into their curriculum by curriculum materials that incorporate data and data analysis into the design, often in partnership with data scientists. Such curricula can be educative for the teachers as well as the students, if the materials follow "best practices" in curriculum design, helping the teacher understand the pedagogy and scaffolding that the design team has envisioned and supported.

Learning goals for data fluency across the curriculum

In addressing the moderator's questions, "What learning outcomes are we trying to support with data-intensive STEM curricula and how do we measure them? What have we learned about how students approach & use data?", the experts on the panel focused on several key qualities of data fluency in relation to student curiosity and agency. Margo Murphy noted that in the Next Generation Science Standards (NGSS) many of the science practices defined by the standards require some engagement with data and its interpretation and communication. Even more notable, as Chad Dorsey said, is that almost half (46% by one count) of the performance expectations in the NGSS involve working with data in some way.

The panelists suggested that while much has been learned about pedagogy and curriculum design for authentic data work, many of the assessments do not go below the surface of student data science, focusing still too often on the mechanics of computation and graphing. Assessments do not yet, for the most part, understand what students know and can do with respect to messy data, data visualization, data as it is used in different subject areas, or the life of data. The panelists agreed that if our teaching, learning and assessment don't get past the superficial, we can't reach the deeper layers of data science, which often are the most compelling and motivating for learners (and teachers) — the ethics of data collection and use and how equity relates to the use of data; the nature and bias in algorithms; issues of privacy and data security.

In reflecting on learning goals, Andee Rubin noted that there is much to learn about how to help students move from 'getting **the** answer' from a data set, to 'getting **my** answer' — curiosity and agency again. How do we scaffold that transition? One strategy that has been used in Rubin's Data Clubs is to resist close analysis at first, and just learn to evaluate whether a particular set of data can or cannot be used to answer the question on the students' minds. The way that you view data shapes how teachers and students will approach data inquiry

Good teacher professional development is a critical barrier right now

Margo Murphy emphasized, in connection with most aspects of the subject touched on during the panel, teacher learning is critical to future progress. As the incorporation of data and data science becomes more compelling across the curriculum, a key task must be to help teachers build the skills they need — both the technical skills, and the interpretive and pedagogical skills needed to work with students & data.

As with so many other areas in STEM education, the teachers need to be learners and in a certain way, practitioners of the discipline they teach, moving beyond the mechanics of *graphing* data for example to the *visualization* of data, with the question-driven attitude that this entails. As Suyen Machado said, "If we expect students to wrestle with data, teachers have to be able to do that, also." They need to experience and understand how, as Chad Dorsey noted, data are collected through a purposeful interaction between an instrument or tool and the world — the phenomenon under study. This is at the core of what science is about. Critique and interpretation of data often drives improvements in instrumentation as well as in theory — and thus opens the door to new questions. Moreover, as Dorsey notes, teachers need to feel uncomfortable with uncertainty. If students collect data, it'll be different every year! The COVID-19 pandemic has offered important opportunities for going deeper with authentic data — from many different disciplinary vantage points. With respect to teacher PD, Murphy noted that the widespread use of Zoom or other "remote learning" systems may facilitate new avenues for broad-scale teacher learning about authentic data, and this possibility was affirmed by some of the commenters in the chat accompanying the presentation.

The COVID-19 pandemic as a source of authentic data

Randy Kochevar described at the outset of the panel, and in his blog, how the COVID event has produced an abundance of publicly available data that is constantly changing. The panel discussed several points about this remarkable moment. It has offered students a chance to see trends and patterns develop — and to understand how real-world data can be messy in several different ways. What are the sources of error in the data being reported? What are the difficulties in getting reliable data? What if two different agencies report different data from the same area? How do we reconcile conflicting interpretations? This is exploratory, investigatory data analysis in real time, with very high stakes for public health, politics, economics, and social welfare. As Rubin said, the emerging, turbulent picture of COVID and its consequences "took the mask off the adult world," and enabled students to see data science at work. Moreover, data and discussions of data quality and interpretation have come into almost every home in the country, as people watch this unforeseen event unfold. Suyen Machado spoke of how teachers have found the Data Cycle idea (see https://hdsr.mitpress.mit.edu/pub/577rq08d/release/3 for a good brief explanation) can help students understand the life of the data that are emerging about the pandemic and its effects.

And as the epidemic has ground on, Dorsey and Machado point out, it is not the only urgent crisis that can hardly be understood without some engagement with data that is just as messy and challenging to understand, evaluate, and interpret — such as the turmoil in national and international weather patterns; the wildfires in the west and north (and the Amazon and Africa and Asia...); unemployment and income inequality; and the issues of racial justice, which are deeply intertwined in all of these other crises — these are medical or natural events, but they are also social events. Such vital challenges are gripping for students, and they quickly see that in engaging with the data, they gain insight into the problems, their implications, and possible solutions. Now as never before is a time to collaborate with, and learn from, many disciplines that are collecting data about different elements of the world we live in, and working to make sense of them. Students also are in the business of making sense of their world, and they can easily come to see how they can make good, purposeful use of data science in that endeavor.

Recommendations for Policy-makers and administrators

- In order to leverage the use of authentic data across the STEM curriculum, policy makers and administrators need to recognize the need for infrastructure and resources to support teachers' learning and experimentation. With so many resources and data tools now Web-enabled, equitable access for all students is necessary. Teachers should be supported in the acquisition of curriculum materials that incorporate and support work with authentic data sets.
- Time for teacher collaboration and discussion will be essential to a successful incorporation of data science into the teaching of STEM disciplines. A "design-based implementation" approach, in which feasible, concrete goals are set based on a solid conceptual framework and conjectures about the outcomes of each innovation, can help structure this effort. It can encourage regular collection of information, and reflection on progress and problems in a timely fashion.
- Authentic data often cross disciplinary boundaries, and because they have to do with the world, they are likely to include social, economic, and ethical elements of interest to the students' communities. This means that the leveraging of authentic data for a data-fluent education may well connect the curriculum with the community, and suggest new patterns of collaboration across disciplines.

Recommendations for researchers

- The expert panel mentioned several areas in which research is needed. For example, when students work with data using different computational or analytic tools, how does the tool shape the way they make sense of the data? How, for example, do the constraints of a specific tool facilitate students' acquisition of key disciplinary ideas (e.g. statistical distributions and variance, or even something so basic as the nature of variables)?
- Dorsey spoke of the lack of research on how students group and categorize data at different ages? How do students learn about the structure of a data set? How do they understand sampling?
- The changes in pedagogy needed to leverage authentic data across the curriculum, and the challenges and strategies for teacher learning are other areas in need of research.
- Finally, the school-cultural changes suggested above ought to support new patterns of teacher collaboration, foster teacher learning, incorporate discipline-crossing ideas and questions, and support the development of a culture of data fluency in schools, with resultant effects on student learning. This complex model is a fertile field for research that problematizes and examines every link in the chain.

Recommendations for Teacher Leaders

All the panelists emphasize the importance of teacher PD, and especially teachers' learning to experience working with data as inquiry. By having time to exercise their own curiosity and agency, it is argued their pedagogy around data will gain flexibility and authenticity. Teacher leaders can exert leadership by doing this work themselves, and supporting and facilitating teacher collaboration and communication about their experiments with authentic data.

Written by:
Brian Drayton,
Co-Director for the Center for School Reform at TERC.