Abstract: We examine how developers of data science curricula determine what makes a “good” dataset that enables 10–14-year-old students to engage in the data analysis cycle, especially posing their own questions about relationships between variables. We describe processes for curating existing datasets to address pedagogical goals for learning about data. How can data science educators construct and transform scientific datasets into ones appropriate for students with little data experience? We draw from our experiences working with several publicly available datasets, which were used with CODAP (the Common Online Data Analysis Platform).
In our work, we consider characteristics of good datasets for engaging students in data investigations, including but not limited to:
- Multivariate—there should be enough variables so students have agency in asking their own questions of the data;
- Variable types—includes both categorical and numerical variables so students develop an understanding of the affordances of different types of questions they can ask and patterns they might find (e.g., comparing groups, exploring correlation, etc.);
- Provides opportunities to learn—including enough meaty relationships between variables so something interesting can be discovered;
- Authentic and meaningful—scientifically and methodologically sound, but minimizing complexities that obscure meaning for students;
- Size/sampling—having enough cases to enable meaningful comparison and filtering, but not so many cases that it becomes overwhelming; and,
- Case/data structure—how the case is defined should make sense given the questions we want students to be able to explore with the data.
We use examples from multiple projects to elucidate these characteristics, including: 1) A dataset on injuries that enables students to ask questions about relationships between variables such as type of injury, body part hurt, and whether the injury was seen in the emergency department; 2) Datasets where students can explore trends over time in rates of Lyme disease in different US states and in relation to different climate variables; and, 3) Data from the Luquillo Long-Term Ecological Research (LTER) site in Puerto Rico that allows students to investigate the impact of hurricanes and droughts on the populations of three species of shrimp in the rainforest over the past 20 years.
Bios:

**Andee Rubin** is a mathematician and computer scientist 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. She has led projects to build and study software for elementary math education ([ink-12.terc.edu](http://ink-12.terc.edu)), support high school teachers in having students construct infographics ([science-infographics.org](http://science-infographics.org)), use log data to support learning in museums ([nysci.org/connected-worlds/](http://nysci.org/connected-worlds/)), consider how mathematics is used in making ([mathinthemaking.terc.edu](http://mathinthemaking.terc.edu)) and study how students and teachers use software tools to visualize data. She was a graduate student in the MIT Artificial Intelligence Laboratory and worked at Bolt Beranek and Newman before coming to TERC in 1990. She is an avid choral singer and has been a member of Coro Allegro for over 25 years.

**Dr. Jacob Sagrans** is Senior Research Associate at Tumblehome, where he works on out-of-school K-12 STEM education projects. Since 2019, he has worked on the NSF-funded Data Clubs project to develop, teach, and research three data-focused modules for middle school youth. He also is a key staff member of the COVID-Inspired Data Science Education through Epidemiology project and a new project introducing youth to data science and ecology through the lens of Atlantic puffins. In addition, Jacob is a consultant for Concord Consortium’s Common Online Data Analysis Platform (CODAP) Version 3.0 project and previously consulted for an ecology/data science education project at the Gulf of Maine Research Institute. Jacob holds a master’s degree in data analytics from Northeastern University’s Roux Institute, a PhD in music history from McGill University, and a BA from Grinnell College.

**Dr. Jan Mokros** is a developmental psychologist and Lead Researcher at Tumblehome, where she is the Principal Investigator for two NSF-funded projects. Her current projects focus on how youth learn about data in out-of-school settings. Jan’s work with data science education introduces youth to a range of scientific topics including Lyme disease, sports injuries, COVID-19, and puffins. She has conducted substantial research and has authored numerous publications on how children understand data and statistical ideas. Jan also has authored three books, including one for museum educators on incorporating math into exhibits and programs, one for parents on exploring math in everyday life, and one for elementary teachers on approaches to teaching math. She has been involved as a writer and researcher for the math curriculum developed by TERC, *Investigations in Number, Data and Space*.

**Dr. Traci Higgins** is a senior researcher in STEM education at TERC. She has over 20 years’ experience developing educational materials, processes, and models to support STEM learning and teaching both in-school and out-of-school, focusing on K-8 mathematics, data science K-12+, and interdisciplinary reasoning. Her work includes co-developing out-of-school curricula to introduce adolescents to data science using multivariate real-word data and studying how youth use technology to engage in data investigations driven by their own questions. Currently, she co-directs a project that examines how undergraduate teams leverage interdisciplinary knowledge in their work with complex real-world data during DataFest, an open-ended, weekend-long, hackathon-style challenge. She works from a student-centered, inquiry-oriented, culturally responsive stance, embracing the assets of learner diversity, and her research is grounded in the cognitive sciences. When not working, Traci plays basketball, hikes, rides and follows cycling, raises heritage chickens, gardens, hunts mushrooms, cooks with wild edibles, and reads voraciously.