WHEN THE DATA DRIVE THE LEARNING

Traci Higgins[a], Andee Rubin[a], Jan Mokros[b], Jacob Sagrans, and Ada Ren-Mitchell[a]
[a]TERC, USA
[b]Science Education Solutions, USA
traci_higgins@terc.edu

What happens when a diverse group of youth ages 11 through 14 are introduced to data science using authentic, public, multivariate data in an out-of-school context assuming no special prerequisite knowledge? We designed three 10-hour Data Club modules in which real-world data and the questions students asked of such data drove the learning process. Each module was grounded in a topic that youth connected with at a personal level. Youth learned how to use a free online data platform that made it easy to rearrange, group, filter, and graph data. Within the progression of the module, we used youths' own questions, data moves, and data visualizations to engage them in critical inquiry and foster productive habits of mind for working with data. Our goal was for youth to emerge from the Data Clubs experience feeling empowered to interact with, ask questions of, and reason about and from data.

INTRODUCTION

We live in a world full of data. Data impact most human endeavors. An understanding of data and the reasoning used to make sense of and draw insights from data is necessary to "cope intelligently with the requirements of citizenship, employment, and family, and to be prepared for a healthy, happy and productive life" (Franklin et al., 2007). All youth should and can learn foundational concepts in data science (Usiskin, 2014).

With the rise of the open data movement, increased sharing of large datasets, broadening of the applications of data, and the emergence of "big data," our understanding of what it means to be data literate continues to evolve. Our ideas about statistics education are expanding to encompass a broader notion of data science that includes data wrangling, attending to metadata, working with a variety of data types, and using secondary data to investigate new questions.

How do we prepare learners for future encounters with data "in the wild"--the massive amounts of complex, multivariate, secondary data that can be found in the public sphere? This includes data of different types that can inspire a range of different investigations. It demands a level of critical data literacy not required by highly structured datasets found in textbooks (Weiland, 2019). In encountering public data, students need to learn to ask about the context surrounding the data--the who, what, when, where, and how of the data. They need to be able to pose their own questions, chart a series of "data moves" to make sense of the data (Erickson et al., 2019), notice interesting patterns, and be able to interpret such patterns, connecting findings back to the phenomena modeled by the data.

The Data Clubs project explores how youth interact with complex real-world public data when we allow the nuances of the data to drive the learning experience. The program is designed to run in an out-of-school learning context. This provides flexibility in selecting topics that cut across disciplinary silos and allows us to focus on data science ideas and practices as they emerge in the context of motivated inquiry, rather than being guided by a set of skills to be covered. Our research focuses on how youth use data tools, contextualized knowledge, and their developing understanding of data structures to interact with datasets, pose their own questions and make purposeful moves to extract insight from the data.

DATA CLUB DESIGN

We have designed three modules. We began by identifying topics that a focus group of middle schoolers responded to as personally relevant and meaningful. The topic needed to be of interest to students while also encompassing data that could serve as both a "window" and a "mirror" for youth (McIntosh & Style, 1999), connecting to their own lived experiences while also exposing them to variation in the experience of others (Rubin & Mokros, 2018). This would help them leverage background knowledge while also underscoring the usefulness of data in coming to better understand our world. The three modules are: Ticks and Lyme Disease, Teens and Time, and Injuries On and Off the Field.

For each topic, we sought public datasets that would be accessible to middle schoolers while also introducing some of the complexities of real-world datasets. We wanted datasets that, across the three modules, would provide opportunities for examining how youth work with different types of data (including time series, numerical, categorical, and geographical). We customized the datasets by limiting their size to keep them from becoming unwieldy and cumbersome, while still providing enough richness to inspire lots of question posing and a variety of investigations.

We used CODAP (Common Online Data Analysis Platform, developed by the Concord Consortium) in these modules because it is free, intuitive to use, and makes it easy for students to organize, group, filter, and graph data. It also supports more sophisticated data moves (Erickson et al., 2019) as the learner gains experience working with data. We wanted each module to engage youth as data detectives, using CODAP tools to interact with the data, ask questions of the data, and become familiar with patterns that could reveal new insights.

Our design was guided by a set of learning goals and experiences for youth, as follows:

- Appreciate the ubiquity of data and the potential for learning from data.
- Engage with data in ways that are intellectually and personally satisfying and lead to persistence in exploration.
- Be aware of the complexities of measurement and look at data through these complexities.
- Know which questions can be investigated using a dataset—and which can't.
- Employ graphing and analytical "moves" to investigate and make sense of data.
- Investigate relationships among variables by examining patterns created when comparing distributions and/or exploring covariation.
- Understand the case/attribute structure underlying different representations of the same data.
- Construct and make sense of data visualizations, both on and off the computer, using a variety of representational elements and extending beyond standard graphs.

For each module, a progression of activities was developed to provide a balance of time spent on the computer using CODAP and time spent working with contextual information and data offline. Each module centered on the introduction of at least two public datasets. In learning about each dataset, youth began with activities that engaged them in thinking about individual cases and the meaning of the attributes. When working with survey data, students first surveyed each other using a subset of the questions from the actual instrument used to gather the public data they would be exploring. CODAP was used to visualize data, examine and describe distributions, and explore questions involving comparisons or relationships between attributes. Data analysis was grounded in visualization and pattern finding, requiring minimal formal mathematics. Activities were developed to introduce youth to data moves that would be especially productive for making sense of the type of data highlighted in the module (geographical, dates, numeric, ordinal, categorical with few attribute values, and categorical with many attribute values) and the type of questions youth were curious about (such as group comparisons, relationships between attributes, time series, and geographical patterns).

RESEARCH METHODOLOGY AND PARTICIPANTS

We used design-based research to iteratively develop three 10-hour "Data Club" modules. After initial development, each module went through at least two implementations and two rounds of revision. The participants for each implementation included between 6 and 20 middle school youth involved in camps, afterschool, or community outreach programs serving diverse, under-resourced, and/or rural populations. Permission for participation was obtained from both youth and their guardians. Clubs have taken place on a school-sponsored farm in rural Maine, an afterschool program offered by Girls, Inc., a summer camp offered by the YMCA, and as afterschool programs. The series was offered free of charge, facilitated by Data Clubs staff, and co-facilitated by local program staff at most locations. We collected pre and post data from a "Data Dispositions" survey, copies of work and recordings of post-interviews with participants, and field notes and video-recordings of sessions.

RESULTS

The specific data concepts, practices, and skills youth developed over the course of each module, above and beyond our general goals for learning, were shaped by the datasets explored and the questions youth posed. In this section, we highlight some of the module-specific ways that youth deepened their understanding of data through their work with different datasets.

Ticks and Lyme Disease Module

Lyme disease is a salient problem in the US Northeast. Many youth know someone who has been affected. News stories about ticks and Lyme disease are common. Scientists are researching how environmental and weather attributes impact the populations of ticks and the spread of Lyme disease. There is much that is not known. The topic fit our design criteria—youth could draw on their personal knowledge of people who had been impacted by ticks (mirror) and learn how it impacted other regions of the country and where it might be spreading to (window).

We made use of data from the US Centers for Disease Control and Prevention (CDC), which has state-level data on the rates and number of cases of Lyme disease going back more than a decade. Relevant US state-level data on environment and weather were procured from the USDA's Forest Inventory Analysis and NOAA. We created a series of curated datasets with states as cases that progressed from basic information about presence of deer ticks, rates of infection, and location (including geographical data that allowed students to map rates of Lyme disease); to infection rates by year (for each state over 12 years); to a complex set of attributes that could hypothetically be related to the spread of deer ticks and Lyme disease (percent forest cover, average summer high temperature, average summer low temperature, average winter low temperature, summer moisture level, and winter precipitation) (see Harvey, Mokros, Sagrans, & Voyer, 2020).

Although youth worked with a variety of data types, a topic-based theme carried across these datasets. Throughout the Data Club experience, students were continually tasked with finding ways to organize the data that would help them learn something about rates of Lyme disease. In one such activity, youth worked with data on the rate of infection by year for each state (see Figure 1). Their question was: "In which three states will the rate of Lyme disease be highest next year?" Youth were not told a method to apply. Would they look for the three states with the highest rates in the most recent year? Would they look for states with the highest average rates over the 12-year period? Or would they consider states with upward trends over time and use those to make predictions about the rate for next year?

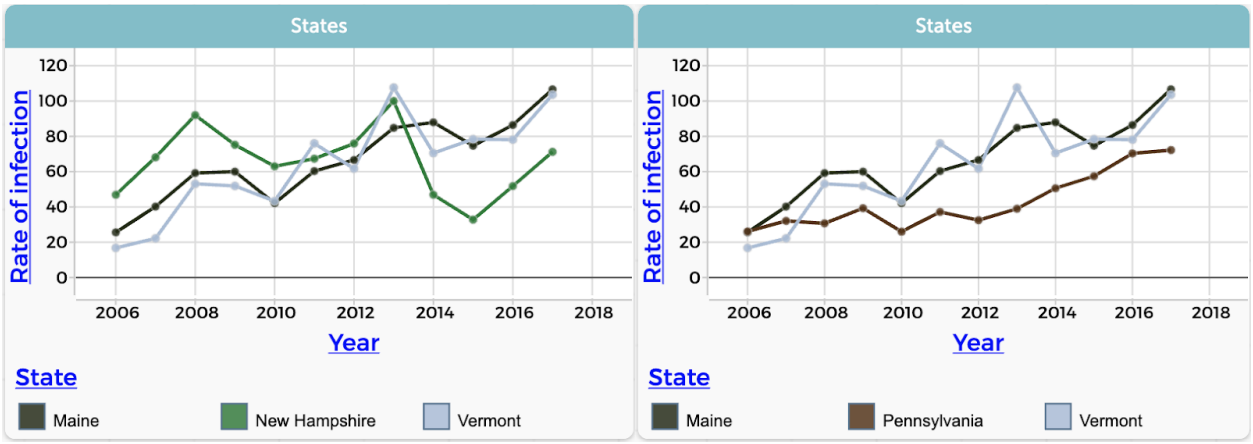

Figure 1. Graphs produced in CODAP exploring the question: "In which three states will the rate of Lyme disease be highest next year?"

Youth easily used CODAP to plot infection rate as a function of year. Many color-coded the states and created a legend. Initially, most youths' graphs included every state. Youth then looked for states that either had high rates of infection in general or that showed an upward trend over time. Some youth looked for both. They used a filtering move to focus on states of interest. As seen in the graph above, one boy selected New Hampshire as his third state, arguing that most years the rate was above 60 per 100,00 people. Conversely, a girl selected Pennsylvania as her third state arguing that if you imagined the pattern continuing it would be at least 70 per 100,000 people by next year.

Youth noticed that trends could include some variability from year to year, but that over time, for many states, a general pattern emerged. Several states seemed to "win" on account of generally high rates of infection and an upward trend, but there was some ambiguity in predicting the three states most

likely to have the highest rates of infection next year. As the discussion wound down, students wanted to know, "who's right?" When the facilitator shrugged, and said, "We can't be sure, it's still a year away" there was a moment of surprise as students came to terms with the fact that the data provide useful insight, but that there is a level of uncertainty when drawing inferences that go beyond the data at hand.

Sometimes the questions raised by youth involve complex statistical ideas that are critical to data literacy, such as nuanced interpretation of correlations. During the final project, one boy decided to investigate the relationship between average winter low for states and their rate of Lyme disease. The boy had created a graph and was pondering a pattern that seemed to suggest that states with balmier winter weather have lower rates of Lyme disease. Do deer ticks do better in below freezing temperatures? As the youth puzzled about this, the facilitator asked, "What else do we know about states that tend to have warmer winters?" After mentioning a few things he knew about warm states, the boy speculated that those states were in the southern part of the country where it is hotter. The facilitator suggested, "Can you use the data to test this, to see if states with balmier winters also have hotter summers?" She left him to his work, and he created a series of graphs (see Figure 2).

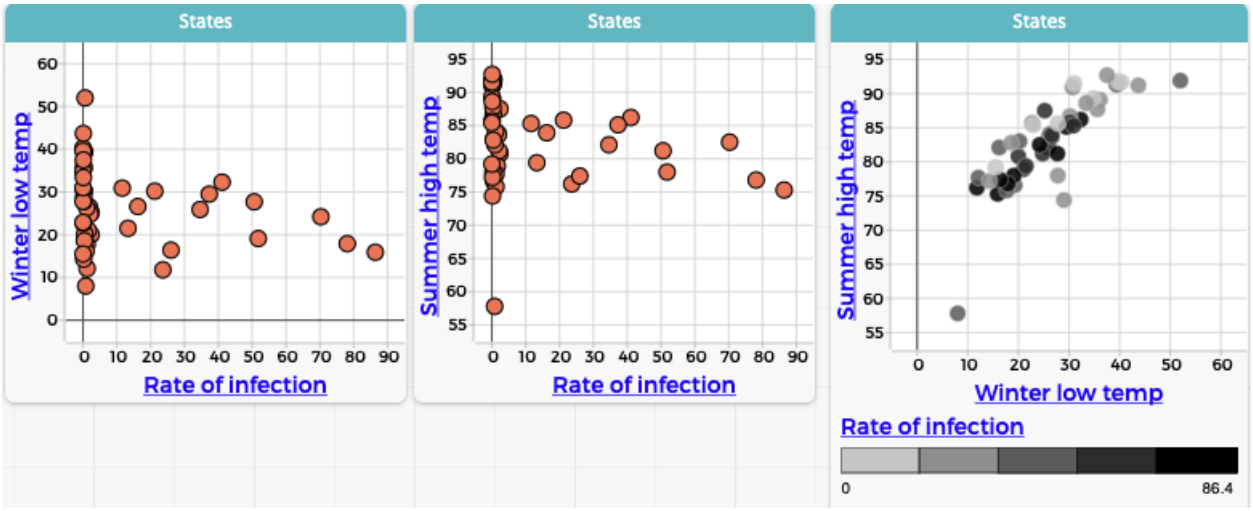

Figure 2. CODAP graphs exploring the relationship between rate of Lyme infection and temperature

During the whole group discussion, the boy volunteered to share his findings. He began by pointing to the second graph and asserting that when average summer highs get above 87 (Fahrenheit), rates of Lyme disease drop to zero. As youth discussed this, they concluded that ticks don't thrive where summers are really hot. The boy presenting then drew attention to the first graph. He pointed out that infection rates were near zero for average winter temperatures above freezing. Youth wondered if ticks primarily affect people when the weather is consistently below freezing in the winter. Did this even make sense? The facilitator drew attention to the last graph: "Let's see what the data say, are the states with the warmest winters also the states with the hottest summers?" Youth concluded by noting that the two variables are themselves correlated and that this investigation offered clues about differences in rates of infection across states but that the cause couldn't be directly determined from correlations.

Injuries On and Off the Field Module

Another topic that fits our design criteria is injuries occurring during recreational activities. Most youth have experienced adults admonishing them to wear helmets while bicycling, be careful on playground equipment especially when conditions are icy, not to dive in the shallow end of a pool, etc. Youth can easily conjure up memories of a time that they or someone they know was injured playing a sport or engaging in other free time activities. This is a topic that youth bring their own knowledge to, but that they can also mine to learn more about the experience of others.

To address this topic, we downloaded data from the 2018 National Health Interview Study. The files include information on injury episodes that occurred for a representative sample of individuals

across the United States over a three-month period. In curating this dataset, we filtered out all injuries except those that occurred during exercise/sports or leisure activity and cut the number of attributes to 11 (age, age group, gender, month of injury, main cause of injury, main body part hurt, other body part hurt, type of injury, admission to an Emergency Department, activity at time of injury, location at time of injury).

The data inspired many questions. Do people get hurt more often while involved with exercise/sports or leisure activities? (Leisure!) Is the pattern the same for males and females? (No!) All attributes but one in this dataset are categorical. This raises issues about the definition of the attributes, such as: how were the categories within each attribute decided upon? Before diving into the data, students survey each other about a past injury to gain a personal understanding of the attributes. One boy wondered whether getting hurt while racing your cousin in your backyard counts as "sports and exercise" or "leisure." As a group, they decided it should count under "leisure." They were learning to attend to how data were generated and the need to carefully define categories.

There are certain data moves that prove especially useful in exploring complex categorical data like those found in this dataset. Several attributes that youth were curious about had many values. For example, "Main body part hurt" included 29 categories for body parts. For "Type of injury" there were 10 categories. A girl wanted to investigate how the body part hurt was related to the type of injury. Her first graph was hard to make sense of because it crossed 29 categories along one axis with 10 categories along the other. However, she was able to use a filtering move in CODAP to select for just cases involving the wrist. It turns out that the most common type of injury to the wrist was a break or fracture. She then explored injury episodes involving the knee. The pattern was different. The most common injury was a sprain, strain, or twist. The filtering move gave youth the opportunity to dive deeper into the data and reveal additional findings. Instead of throwing up their hands when a graphing move created something that made no sense to them, youth were able to utilize a variety of data moves to reorganize and reexamine the data, ultimately drawing new insights from the data.

When looking at numeric data, youth are often drawn to measures of the mean or median. Those terms make no sense when applied to categorical data. However, when investigating differences between groups of unequal sizes, youth sometimes sought out the percent tool in CODAP. They discovered that there is not just one way to find percentages. They needed to think about whether they wanted to find percentages by column or row. This pushed youth to consider what they were implicitly considering as the "whole." Did they want to know for each age group, whether injuries occurred more often with sports/exercise versus leisure? Or did they want to know, of all the injuries that occurred with exercise/sports, was there a specific age group that accounted for most of the injuries? In our experience, students ask the first question, which requires treating each age group, or the columns, as the "whole." The first graph below (Figure 3, left) allows students to compare activity at time of injury within each age group. The second graph does not (see Figure 3, right). Some middle schoolers are just beginning to develop the reasoning to make sense of this choice. The data in this module provided opportunities for them to explore what percentages mean and why the "whole" they chose needs to be connected to the question they are asking.

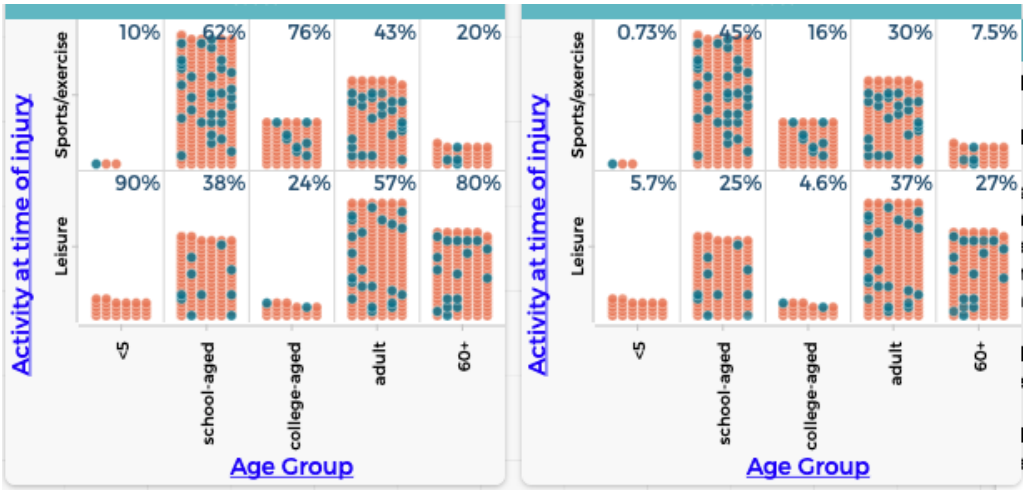

Figure 3. CODAP graphs of age group vs. activity at time of injury with percentages by column/age group (left) vs. percentages by row/activity type (right)

CONCLUSION

While participating in Data Clubs, youth became deeply engaged in working with data, asked questions of the data, and used data moves to make sense of the data as they investigated their own questions. The lessons students learned came from the data itself, including how to: interrogate data asking questions of who, what, when, where, and how; generate a juicy question that can be investigated with the data at hand; create and make sense of a variety of data visualizations; not accept a confusing data visualization but instead employ some data moves to dig deeper; visually compare groups in the aggregate; identify patterns of change suggestive of a trend over time; notice features of scatterplots that suggest a relationship between attributes; point to features of the data in justifying a claim; and have confidence that they can find meaning and insight in data.

REFERENCES

Concord Consortium. (2021). *Common Online Data Analysis Platform (CODAP)*. http://codap.concord.org

Engle, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44-49.

Erickson, T, Wilkerseon, M., Finzer, W., & Reichsman, F. (2019). Data moves. *Technology Innovations in Statistics Education, 12*(1). https://escholarship.org/uc/item/0mg8m7g6

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association.

Harvey, M., Mokros, J., Sagrans, J., & Voyer, C. (2020). What makes them tick? Middle school data science explorations of ticks and Lyme disease. *Connected Science Learning, 2*(3).

McIntosh, P. & Style, E. (1999). Social, emotional, and political learning. In Cohen J. (Ed), *Educating Minds and Hearts: Social Emotional Learning and the Passage into Adolescence*. Series on Social Emotional Learning. New York: Teachers College Press.

Rubin, A. & Mokros, J. (2018). *Data Clubs for Middle School Youth: Engaging Young People in Data Science*. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10), Kyoto, Japan, July 8–3.

Usiskin, Z. (2014). *On the relationships between statistics and other subjects in the K-12 curriculum.* Paper presented at the 9th International Conference on Teaching Statistics. Flagstaff, Arizona.

Weiland, T. (2019). The contextualized situations constructed for the use of statistics by school mathematics textbooks. *Statistical Education Research Journal 18*(2), 18-34.