DATA CLUBS FOR MIDDLE SCHOOL YOUTH: ENGAGING YOUNG PEOPLE IN DATA SCIENCE

Andee Rubin[1] and Jan Mokros[2]
[1]TERC, Cambridge, Massachusetts, USA
[2]Maine Math and Science Alliance, Augusta, Maine, USA
Andee_Rubin@terc.edu

We describe a project to engage middle school students in out-of-school Data Clubs whose goal is to give them a "taste" of data science. We identify four criteria for materials we are developing, related to topics, datasets, tools, and activities. We discuss our justification for each, provide examples of how they might be realized, and consider the complexity of satisfying all four criteria simultaneously.

INTRODUCTION

What aspects of data science are accessible to middle school students, given appropriate technology, datasets, and materials? We describe a project designed to investigate this question by implementing out-of-school Data Clubs with community partners in both urban (Massachusetts) and rural (Maine) communities in the northeast United States. Youth will participate in 10-hour modules in an after-school or summer camp setting. We will recruit participants as much as possible from populations under-represented in STEM fields, including girls, youth of color, and economically disadvantaged and rural youth. Data Clubs will use both TinkerPlots and CODAP as data analysis tools, and a combination of small, self-collected and larger, web-based data sets. Each module will focus on a particular topic, selected with the advice of youth advisory focus groups. The project is developing three curriculum modules, a survey measure of "data science dispositions," and an interview assessment of students' ability to ask relevant questions of datasets.

Because the project is still in the development phase, we describe here our evolving criteria for the curriculum materials, illustrated by examples from one of the three modules. We have four inter-related criteria for our materials: topics, datasets, tools, and activities. All of these are in service of our over-arching goal: to give participants a "taste" of data science that helps them see how data can give them the power to answer questions they find relevant and interesting.

TOPIC CRITERIA

We are inspired by a perspective on curriculum dating back several decades and expounded upon in a chapter of the 1999 book *Educating Minds and Hearts* (McIntosh & Style, 1999). The general idea is that curriculum is a structure that provides "windows out into the experiences of others, as well as mirrors of the students' own reality." For students from under-represented groups, many mainstream curriculum activities skew toward being windows more than mirrors, focusing on experiences for which they are onlookers, rather than experiences in which they can see themselves. We want the topics for our modules to both reflect the lives of participants and expand their horizons by giving them insights into other people's lives. In order to assure that topics we are considering are "mirrors" for participants, we have been working with middle school advisors who weigh in on topics in terms of their relevance and interest to others like themselves.

To the "window" and "mirror" criteria, we have added a third, based on our own philosophy and on an awareness of middle school students' interests in "fairness" and concerns for those who are vulnerable. We want our modules to provide participants with a chance to take action in the pursuit of a cause they care about, if they are so moved.

Currently, we are pursuing four topics, in conjunction with our youth advisors:

- Animal shelters: How do animals get to them? How/how often do they get adopted? Which animals get adopted first or last? What happens to abandoned pets in natural disasters such as hurricanes?
- Media use: Does having a smart phone in the bedroom disturb sleep? How much screen time is too much? Are there relationships among gender, age, and media use?

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

- Disease spread and its relationship to climate change: Where are tick-borne diseases (such as Lyme disease and West Nile) most prevalent? How are they spreading? What relationship do these patterns have with temperature and precipitation patterns? How are disease symptoms related to a patient's age?
- Sports injuries: What sports are most likely to lead to injuries? Are there gender differences in types of injuries? What effects have changes in protective sports gear had on injury frequency and severity?

DATASET CRITERIA

There is much talk about "big data" and how it will change many aspects of our lives, from the workplace to the shopping mall to the doctor's office. Big data are often characterized with the 4 V's: (IBM, 2018): Volume (there are a lot of data points); Velocity (the data, especially when generated automatically, arrive quickly); Variety (data may include text, numbers, images, video, sound); and Veracity (data quality is sometimes marginal; a small percentage of incorrect data is to be expected). When we think about data science education, we sometimes assume that students should encounter big data sets immediately, in order to get an authentic "data science" experience. However, truly large data sets can be overwhelming to novices, who are likely to have trouble seeing the forest for the trees, as they get distracted by the many complexities that arise in cleaning, organizing and otherwise wrangling the data.

On the other hand, the data students generally encounter in school could be considered "small data." Often there are just a few questions, each of which has only a few possible answers (e.g. what's your favorite color?). Even if the individual questions are relatively open-ended and interesting, there are frequently no meaty relationships among the variables, so analysis begins and ends with looking at univariate distributions. Interacting with these kinds of datasets seldom gives students a glimpse of the power, relevance, and excitement that data science can yield.

We find ourselves in a situation analogous to Goldilocks and the Three Bears fairy tale: one chair is too big and another is too small. Goldilocks eventually finds a "just-right" chair, and in the Data Clubs project we will be searching for "just-right" datasets. As starting criteria, we want these datasets to be multivariate, contain multiple types of data (numerical, categorical, geospatial) and have within them rich relationships among variables, allowing for different students to investigate different possible patterns. We will pre-process the datasets we find so that they are accessible for middle school students; this is likely to involve choosing a subset of variables or cases from particularly large databases, as well as the standard "cleaning" steps.

TOOLS CRITERIA

Given the age of the youth in our program and the relatively short amount of time they will spend in a Data Club, any data analysis software we offer will have to have a minimal learning curve that allows participants to create data visualizations almost immediately. The two tools that we will use—Tinkerplots and CODAP-- have been designed specifically as data education tools, rather than as approximations to professional tools. As such, they make extensive use of direct manipulation in the construction of graphs, allow for rapid transitions from one type of representation to the next, and do not require coding knowledge (as does, for example, R). While they are simple to use, however, these two tools have considerable power and give novice users the ability to carry out relatively complex data analyses. In particular, both tools feature linked representations, in which data selected in one graph is highlighted in all other graphs on the screen, thus allowing a user to see complex relationships among variables. Both tools also support sophisticated filtering and subsetting by direct manipulation, rather than only through formulas (although selection by formula is also supported).

One of these tools is TinkerPlots (Konold & Miller, 2005), which was originally designed for use in middle school classrooms. Since its introduction over a decade ago, TinkerPlots has undergone a large number of classroom trials, including with students as young as fourth grade, and has been used in scores of research projects on statistical reasoning. This research has allowed statistics educators to understand different ways in which novice data analysts view data distributions (Konold et al, 2015) and to know how to support them in developing more sophisticated approaches.

The second tool we will use was developed more recently and builds on many of the same design characteristics. An additional feature of the Common Online Data Analysis Platform (CODAP) (Finzer, 2016) is that it makes it possible to view both geospatial and distributional data simultaneously and to see the connections between locations and other values. The figure below displays a particularly compelling example of this kind of analysis, in which we can see both the swimming routes of four different elephant seals off the coast of California and a scatterplot of their speed vs. their depth. These two graphs are linked; all of the points corresponding to one of the seals (#546) are highlighted in both graphs, so we can see its route and how its speed varied with depth.

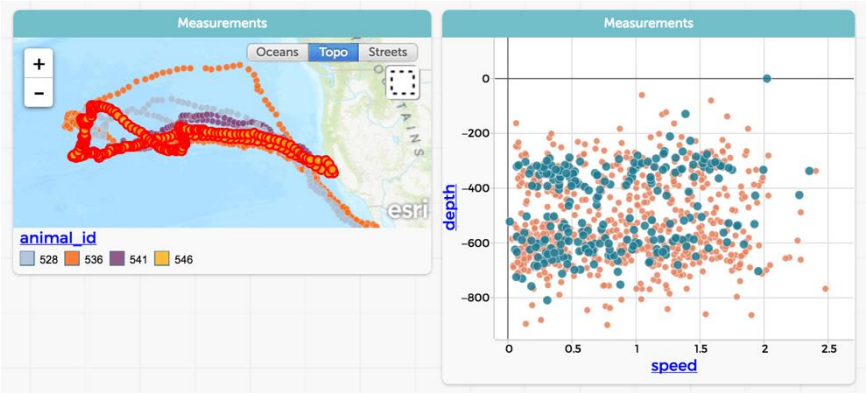

Figure 1. Swimming routes of four different elephant seals

ACTIVITY CRITERIA

Since our intent is to introduce middle school youth to data science, we might ask: what distinguishes "data science" from "statistics"? Professional data scientists use a variety of techniques, such as machine learning, that are not accessible to middle school students, but we still want participants in our activities to get a "taste" of data science. To that end, we are guided by Tim Erickson's "sniff test" (at the risk of mixing senses) for Data Science activities (Erickson, 2017). Erickson identifies three basic characteristics of data science activities: being awash in data, using data moves, and dealing with data properties.

- Being "awash" in data: In contrast to those statistics class datasets in which there are two variables, a simple question and an agreed-upon formula or test to use, data science activities present the user with a confusing amount of data, with no obvious direction. The task is to find some signal(s) in the data, but they may be hard to discern and different analysts may come up with different ones.
- Using data moves: Manipulating the data to find meaningful subsets is critical to analysis. For example, there may be a relationship between an animal's brain size and body size, but it may be a different relationship in mammals and reptiles. A particular elephant seal may have a different feeding pattern than the others – and the cause may not even be known. Finding this individual difference may set off a search for reasons or other related phenomena. Data moves include filtering, subsetting, summarizing, transposing, and merging; determining what are relevant data moves and how best to introduce students to them is an ongoing topic of research.
- Dealing with data properties: Data sets that support data science investigations should have multiple variables, more than a few dozen cases, and possibly represent data drawn from a variety of sources. They may represent data that was automatically gathered via sensors or may be a byproduct of data collected for other purposes (e.g. transit data from an automatic toll system). These characteristics make an inquiry more complex, as there may be substantial questions about how the data were obtained and whether the fit the new purpose to which they are being applied.

We will also include two other kinds of activities, given our topic criteria and the age group with which we will be working. Early in each module, we will incorporate a data-collection activity, in which participants get the experience of collecting their own data. There are two

reasons for this. First, in pursuit of seeing Data Clubs modules as "mirrors," we want participants to work with data that describe their experience in some way. For example, in a module on social media use, we might use data from a Pew survey: Before looking at the Pew data, participants would answer the same set of questions for themselves, and possibly collect data from a friend or sibling as well. Later, participants' own data could be integrated into the larger dataset. Second, we want to counteract the tendency people have to accept data as "true," rather than being critical consumers who understand that the data collection process produces a model of reality that is necessarily incomplete.

Inspired by a recent publication *Dear Data* (Lupi & Posavec, 2016) we will also try out activities in which participants can create their own idiosyncratic, hand-drawn data representations. While computer-generated representations of data are "neat," relatively simple to create and easily modifiable, there is something charming and personal about hand-made representations that we hypothesize will appeal to middle-school students, particularly girls. Having ownership of one's data, as well as being able to creatively represent it, may prove to be an important starting point in learning about data science.

INTERACTIONS

Satisfying all of the criteria described above in a single 10-hour experience with Data Clubs may be impossible. We have already encountered conflicts between topics that are interesting to potential participants and the availability of appropriate datasets. For example, Common Sense Media carries out some of the best surveys of media use among tweens, teens and their parents and publishes intriguing Fact Sheets with their findings. Their sample sizes are around 1000, and they even publish their survey questions and encourage other researchers to use them. However, they do not yet make their raw data available for others to use. On the other hand, there are many open datasets on city and country websites about traffic violations and parking tickets, but the students we work with probably wouldn't find those relevant. The project will determine ways of negotiating and optimizing the criteria, and provide those working in the newly-emerging field of data science education with guidance about designing curricula.

CONCLUSION

The Data Clubs project responds to the increasing importance of data science in both the workplace and everyday life and to the lack of opportunities for middle school students, especially those from under-represented groups, to have engaging experiences with data. The project seeks to bring data science to participants who might otherwise not encounter it, to encourage them to seek more opportunities to engage with data – either in-school or out-of-school – and to view data science as inviting and empowering. Achieving this goal requires focused and innovative materials that take into account the capabilities and interests of middle-school youth and at the same time take advantage of the proliferation of available datasets and data visualization tools. We presented here four criteria for materials design and look forward to sharing the resulting modules with the international statistics education community.

REFERENCES

Erickson, T. (2017). https://bestcase.wordpress.com/2017/02/21/smelling-like-data-science/

Finzer, W. (2016). Common online data analysis platform (CODAP). Emeryville, CA: Concord Consortium. https://codap.concord.org/

IBM (2016). http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. Educational Studies in Mathematics, *88*(3), 305-325. https://doi.org/10.1007/s10649-013-9529-8

Konold, C. & Miller, C. (2005). *TinkerPlots : dynamic data exploration.* Emeryville, CA: *Key Curriculum Press* and Amherst, MA: *University of Massachusetts.* Currently published by *Learn Troop.* http://www.tinkerplots.com/

Lupi, G. & Posavec. S. (2016). *Dear Data.* New York: Princeton Architectural Press.

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.