# INTEGRATING TECHNOLOGY AND NARRATIVE TO ENGAGE YOUNG ADOLESCENTS WITH COVID DATA

Pendred Noyce, Janice Mokros, Laura Martin, and Jacob Sagrans
Tumblehome, Inc and Science Education Solutions
penny@tumblehomelearning.com

*The COVID-19 pandemic has offered opportunities to immerse high needs 10–14-year-olds in real-life data that matter. We developed and tested an out-of-school program to increase youth understanding of data science through epidemiology. The 15–20-hour Data Detectives Clubs are structured around an adventure novel, The Case of the COVID Crisis, which introduces readers to epidemics across space and time. Chapters are accompanied by activities, discussion, and exploration of infection and vaccine data using CODAP, the Common Online Data Analysis Platform. To date, around 600 youth have experienced the program, focusing primarily on time-series data. Youth learned about noisy data, comparing graphs, and matching rates to cumulative numbers. These clubs demonstrate the feasibility of integrating data science with epidemiology outside of school.*

## INTRODUCTION

As the COVID-19 pandemic started its spread across the world in the spring of 2020, we began to discuss the opportunity it offered for getting young students interested in data science. Here was a fascinating disease, growing exponentially, that increasingly threatened to disrupt their lives. We decided to create a program combining a time-travel adventure book, *The Case of the COVID Crisis* (Noyce, 2022), with a series of activities introducing ways of examining and understanding data. The Museum of Science, Boston, helped us pilot a remote book club that proved surprisingly popular; subsequently, we obtained National Science Foundation grant funding to adapt and expand the program through after-school and summer camp clubs across the United States. By the spring of 2021, we were partnering with The Concord Consortium to make use of CODAP, the Common Online Data Analysis Platform, a free online tool for exploring/visualizing data (The Concord Consortium, n.d.).

From the start, the CIDSEE project (COVID-Inspired Data Science Education through Epidemiology) had multiple ambitious goals: supporting literacy, introducing data science and data careers, teaching the basic science of viruses and vaccines, and helping young people face the social and emotional challenges of the pandemic. The book, curriculum, and club facilitator guide have been continually updated so that we are now using a third edition of each. Previous communications (Martin et al., 2022; Mokros et al., 2021) have detailed the gains participating students made in their confidence in looking at data, their science engagement and interest, and their career knowledge. This paper focuses on what we are learning about teaching data science to children ages 10–14 in informal settings; specifically, that with a story, activities, and a graphing tool, students can read, understand, and compare time-series graphs.

## SETTING AND TECHNOLOGY AFFORDANCES

Implementation of the full CIDSEE curriculum began in the summer of 2021 in 17 afterschool clubs in five states. At this point, over one year later, close to 600 children have participated. Clubs are led by youth workers, about half of whom have a science background, and all of whom received four to six hours of training in the program materials. The clubs are all affiliated with Imagine Science, a nonprofit STEM education organization that provides logistical support, professional development, and evaluation to afterschool and camp programs that commit to providing at least 12 and usually 15–20 hours of science programming to underserved youth ages 10–14. Imagine Science programs serve primarily minority and low-income youth, and almost a third of them are from homes where a language other than English is spoken.

Facilitators and youth programs freely elected to use the CIDSEE curriculum. Some students chose this particular after-school or camp activity, but many of them signed up for "science" without knowing what it would involve. Nevertheless, facilitator and student focus groups confirm that the Data Detectives Clubs have been popular among program facilitators and children alike. Although the informal learning setting has brought with it some challenges, such as technology glitches, facilitators

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

who are youth leaders rather than credentialled STEM teachers, a lack of formal assessment of youth, and logistical difficulties, these challenges are outweighed by the flexibility, enthusiasm, physical activity, and fun that an informal context can provide.

Because its basic moves are easy to learn, CODAP is well adapted to use in an informal environment. The platform allows users to convert tables of data into graphs by dragging and dropping column titles onto axes. Graphs can incorporate categorical and numeric data and can be scaled, moved, annotated, or color coded when different sets of data (e.g., infections in Louisiana vs. North Dakota) are displayed on one graph. For the purpose of this program, a number of special features were added:

- A NetLogo simulation of infection spread embedded in CODAP, allowing a time-series graph of daily infections to appear as the simulation runs.
- CODAP portals to access constantly updated data from the U.S. Centers for Disease Control and Prevention (CDC) for infection and vaccination levels, organized by state and county.
- Accommodations to decrease student confusion about how to use CODAP, for example, having CODAP automatically place date on the horizontal axis as students examined time-series data.
- Brief video tutorials on how to use CODAP in the context of each lesson.

Books and materials were provided to each club free of charge, along with a stipend to offset facilitator training time, but facilitators received no direct compensation. Facilitators initially received four hours of training for the CIDSEE clubs; later, that was increased to six hours to allow more and deeper exploration of CODAP.

Beyond these technology supports, it is important to note that the informal afterschool/camp setting, often led by a facilitator well known to the students, provided a low-stress place for students to explore. Other features of the program helped to create a motivating environment. Each club had one or two career visits, virtual or in person, from data professionals such as epidemiologists, frontline health workers, or researchers. The adventure storyline, supported by discussion questions and brief podcasts of the characters debating, tied the various activities together and allowed the young participants to see the data in context. Finally, discussing the situations and dilemmas faced by fictional characters encouraged youth to address their own worries and sometimes sadness in a supportive, safe environment.

## DATA-SPECIFIC GOALS OF THE PROJECT

Among the goals of the CIDSEE project are three drawn from the *Guidelines for Assessment and Instruction in Statistics Education II* (GAISE II) report (Bargagliotti et al., 2020):

- "Pose statistical investigative questions that require looking at a variable over time" (p. 44).
- "Understand that data can be used to make comparisons between different groups at one point in time and the same group over time" (p. 44).
- "Use statistical evidence from analyses to answer statistical investigative questions and communicate results with comprehensive answers and some teacher guidance" (p. 45).

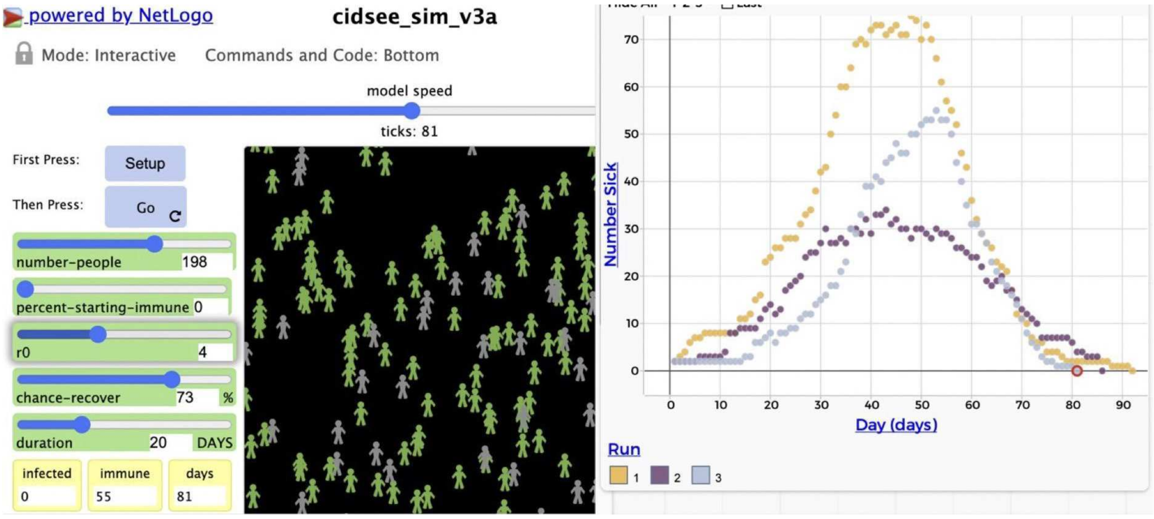
We added a fourth goal related to the epidemiology focus of this project:

- Recognize that epidemiologists collect data to inform the public and help people make data-based decisions.

## DESCRIPTION OF DATA CURRICULUM

As evident above, a major emphasis of the Data Detectives curriculum was time-series data, and therefore we will address the activities using the classification levels provided by Passmore (2018). We began with smaller, sometimes historical, datasets and quickly moved to large contemporary ones. Note that Sessions 4 and 6 are not described below because they focused on virology rather than data science.

- Session 1**:** Youth investigate the effect of R-naught (in epidemiology, the average number of new, non-immune people who can be infected by any one infected person) on the speed of spread of a hypothetical infection. This lesson involves a NetLogo simulation embedded in CODAP in which agents (stick-figure people) change color as they become infected and then immune. Simultaneously, a graph is generated showing how the number of people infected rises and then falls to zero. The purpose of this lesson is to give students a sense of what the shape they see on a time-series graph means. In addition, by running several simulations with the same starting conditions, youth begin to develop an intuitive sense of variability and uncertainty in data (see Figure 1).


powered by NetLogo

cidsee_sim_v3a

Mode: Interactive

Commands and Code: Bottom

model speed

ticks: 81

First Press: Setup

Then Press: Go

number-people 198

percent-starting-immune 0

r0 4

chance-recover 73 %

duration 20 DAYS

infected 0

immune 55

days 81

Figure 1. Variability in simulated data, as shown in sessions 1 and 9

- Session 2**:** In considering real data gathered in Pittsburgh during the 1918 flu, students make note of inflection points and speculate about why the data appear so jagged rather than smooth. After first looking at graphs of infections and deaths over time, they combine the two into a single graph with two variables and reason about the differences. By contrasting 1918 to today, they also reflect on what it means to be able to collate data quickly from all over the world (see Figure 2).

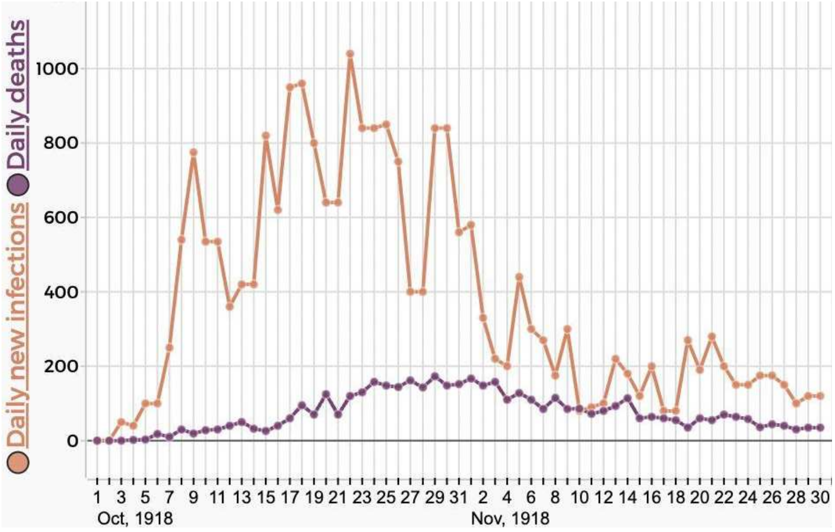


Figure 2. CODAP graph showing flu infection and death rates in Pittsburgh in Fall 1918

- Session 3: Students gain more experience with interpreting the shape of a graph as they examine real data from the Diamond Princess, one of the first cruise ships to suffer a wide outbreak of COVID-19. Because they act as decision-makers responding to the outbreak, students learn about data smoothing at a basic level.
- Session 5: In considering smallpox eradication, students are introduced to visualizing data in color-coded maps that change over time. Students explore different countries in this challenging, data-rich visualization that includes three "dimensions": geography, smallpox case levels, and time.

- Session 7**:** Students examine graphs of daily COVID infections from four mystery countries and match them to graphs of cumulative infection from the same four countries. Then they match their paired graphs with headlines or other statements about the course of the pandemic in the countries.
- Session 8: In the context of considering “fairness,” students compare rates of infection for different racial groups in two states.
- Session 9: Returning to the NetLogo simulation, students experiment with what percentage of people need to be immune (vaccinated) for an outbreak to end quickly with different initial R-naughts. Multiple runs reinforce the idea of variability in data (see Figure 1).
- Session 10: Students compare the rise over time of vaccination levels in different countries and then, for a final project, in different counties in the United States.

## LEVELS OF STATISTICAL REASONING ADDRESSED

In comparing program activities to Passmore’s (2018) framework of types and levels of reasoning in time-series data, we would classify the session activities as follows:

- Level 1: Vertical reasoning—examining specific points on the graph, such as inflection points—is used in sessions 2 and 3.
- Level 2: Horizontal reasoning—interpreting the shape of the entire graph—appears in most sessions, including 1, 2, 3, 7, and 9.
- Level 3: Procedural reasoning—performing calculations using time-series data—is touched on lightly in session 9, where students calculate average results from several different runs of a simulation.
- Level 4: Extended procedural reasoning—using calculations to predict beyond the current—did not appear in club sessions.
- Level 5: Interpretive reasoning—interpreting key features of a time-series graph through integration of statistical and contextual knowledge—is called on, though without calculations, in Sessions 2, 5, 7, 9, and 10. The latter three sessions can also be said to touch on Level 6, interrogative reasoning, if that can be interpreted to include comparison of similar or related time-series graphs *without calculation.*

## OBSERVATIONS OF STUDENT REASONING

The afterschool/camp setting and Imagine Science’s commitment to providing an enjoyable, informal learning experience precluded formal student assessment during the sessions themselves. Moreover, our primary research focus has been on measuring self-reported changes in participants’ STEM engagement and identity, knowledge of STEM careers, and attitudes toward data science over the course of the program (Martin et al., 2022; Mokros et al., 2021). Nevertheless, in order to discern what students have learned about data, samples of student work and videos of final projects were gathered from nine clubs serving about 130 youth. We developed rubrics and counted key features in graphs and final presentations, and we noted whether students correctly matched graphs with descriptions of the pandemic’s course in different places. In addition, the project team observed most sessions of two clubs in the northeastern US, checking for student understanding. Student and facilitator focus groups provided additional information about which activities students found engaging or overwhelming. Preliminary observations are described below.

First, we observed wide variation in students’ basic facility with computers. Some had difficulty entering a URL or using a mouse, whereas others wanted to keep exploring with CODAP even at home or during other program activities. Early student products showed that in an open exploratory environment, not all students recognized that for a time-series graph, time should occupy the horizontal axis. In general, students in grades 4 and 5 (ages 9–11) struggled more than older students did with the data content of the program.

Middle school students easily understood the significance of peaks and valleys in a daily graph of new infections (Passmore’s Level 1). Virtually all students understood the link between a NetLogo simulation and a simultaneously constructed graph; this appears to be a successful tool for helping students understand the concept of a time-series graph. This understanding was carried forward by many students into the more complicated smallpox eradication maps. A late addition to the curriculum, experienced by only a few clubs, was an exercise sonifying a time-series COVID infections graph using voice or a kazoo; again, this exercise demonstrated that students were able to follow the graph horizontally (Passmore's Level 2).

In this informal setting, with students at different mathematical levels, we did not demand that students perform calculations (Passmore's Levels 3 and 4); however, in order to complete activities on racial disparities in session 8, students had to understand the concept of looking at infection *rates* rather than numbers of infections, and they had to compare rates in two different states. Most students we observed navigated these activities well, and many reasoned about how to account for the differences seen. Later, in session 9, they were asked to average data from several NetLogo simulation runs at each of several different vaccination rates in order to reason about what percentage of the population needs to be vaccinated to end an epidemic. Students supported their reasoning with the averaged data.

Finally, many of the sessions involved students comparing more than one graph, noting key features (Passmore's Level 5) and how they differed between graphs, and speculating about reasons for the differences (Level 6). It should be noted that many students made insightful inferences without calculation. One of the sessions that was most informative for us was session 7. For the "matching countries" activity, 56% of student groups (usually two to four students working together) successfully matched daily rates with cumulative rates, thus demonstrating an intuitive understanding of the relation of the first derivative to the integral. About the same percentage of groups successfully matched a verbal, journalistic description of a country's pandemic course with the correct graph of daily infections over time.

Final projects asked students to develop a public health message based on their analysis of local pandemic data. In spite of their work with data, in only 15 of 23 group projects submitted did students actually tie their messaging to a graph they had created. This may have been a problem with facilitation, or it may be that students were already so swamped with public health messaging that they had a hard time stepping back to base an original recommendation on the data they found for themselves.

During the next stage of the project, we will be delving more into student understanding by interviewing individual students about how they make sense of project data, and of time-series graphs in particular. Insights gained will allow us to further tweak the curriculum and facilitator training.

## CONCLUSION

Over the course of the COVID pandemic, we are finding that middle school youth can gain an understanding of time-series data in a context that matters to them. Their understanding derives from experience and exploration rather than highly didactic instruction and calculations. Neither epidemiology nor data science are currently included in US curriculum standards. Nevertheless, our next challenge will be to adapt this active and interdisciplinary unit—which mixes literacy, mathematics, computers, data, geography, and history—for use in schools. We believe that developing an intuitive understanding of how time-series graphs work will give students a foundation of motivation and confidence as they continue their studies in data and statistics, and even as they follow the daily news.

## REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II). A framework for statistics and data science education*. American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf

Martin, L., Mokros, J., Noyce, P., Sagrans, J., & Deol-Johnson, N. (2022). *Data Detectives Clubs: A collaborative approach to data science through epidemiology.* Manuscript submitted for publication.

Mokros, J., Sagrans, J. & Noyce, P. (2021). Data science for youth in the time of COVID. In R. Helenius & E. Falck (Eds.), *Statistics education in the era of data science. Proceedings of the satellite conference of the International Association for Statistical Education (IASE)*. ISI/IASE. https://doi.org/10.52041/iase.fjrqt

Noyce, P. (2022). *The case of the Covid crisis* (3rd ed.). Tumblehome, Inc.

Passmore, R. (2018). Time series—its place in the secondary school curriculum. 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*. ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2H3.pdf?1531364244

The Concord Consortium. (n.d). *Common Online Data Analysis Platform* (Version 2.0) [Computer Software]. https://codap.concord.org/