DATA AND DATA SCIENCE: MAKING SENSE OF THE WORLD

Gail Burrill, Curtis Brown, and Thomas Dick
Michigan State University, Texas Instruments Education Technology, and Oregon State University
burrill@msu.edu

Data and data science are becoming increasingly important, and initiatives related to data science at the school level are emerging in countries across the world. This paper considers what data science means, considers its connections to statistics and critical literacies, and suggests a framework for introducing students to the data science process. The framework can provide a structure to inform the development of materials for better understanding of data science and its role in making sense of the world. Several examples aligned with the framework are given and are appropriate for use in the secondary curriculum to maximize learning opportunities for all students. The discussion ends by considering some implications for the school mathematics curriculum.

BACKGROUND

Investigating phenomena in the world can engage students in building understanding of statistical and mathematical concepts, provide them with the tools to make data-based decisions, and enable them to solve problems related to a variety of contexts such as health, economics, opportunity, and access. Although the term data science has become popular and data science programs are being established in a plethora of institutions of higher learning, many teachers are not sure what data science actually is and how it is related to topics such as statistics or quantitative literacy. An internet search brings many different definitions of data science, for example: "an interdisciplinary field that embraces sophisticated analytical programming to draw out patterns and useful information from the vast abundance of data available today (Baumer, 2015)" (Kjelvik & Schultheis, 2019, p. 2); or "a field that gives insights from structured and unstructured data, using different scientific methods and algorithms, and consequently helps in generating insights, making predictions and devising data driven solutions" (Bansel, 2020, section 1).

A simple definition for data science at the school level might be "the science of learning from data" (Donaho, 2017, p. 247), which involves knowledge of the disciplinary context surrounding a data set, knowledge of mathematical and statistical concepts, and possession of computer science skills (Engel, 2017; Finzer, 2013). Adapting frameworks from that of Mojica and colleagues (2021) and the International Data Science at Schools Project (IDSSP, 2019), we propose the following cycle for introducing students to the data science process:

- *Identify the problem*, which includes understanding the context, relevant vocabulary, and why the problem is important and for whom.
- *Consider and gather data*, either primary (generated in the absence of original data using class surveys, observational studies, or simulations) or secondary (databases from sources such as the web, governments, science, or businesses, typically gathered for a specific purpose).
- *Process data*, which involves data moves such as cleaning data, accounting for missing values or outliers; quantifying categorical data; manipulating and transforming data that have different time frames, scales, units, magnitude, or location; summarizing raw data; or synthesizing multiple data sets (Kjelvik & Schultheis, 2019; Wickham, 2014). Depending on the question of interest, the data may have to be rearranged by category or time.
- *Explore and visualize data*, which goes beyond calculating standard statistical summary measures and traditional graphical displays with an emphasis on multivariate representations that might involve using color; changing shape and size; using dynamic iterations, morphing over time; using heat maps; or using classification and regression trees.
- *Consider models* involves the use of mathematics to represent, analyze, make predictions, and provide insights into real-world phenomena and the recognition that this mathematical relationship in all likelihood involves variability.
- *Communicate and propose action* refers back to the original problem and the need to describe the analysis and potential findings in a way that is understandable to a lay audience.

This framework has considerable overlap with frameworks such as the *Guidelines for Assessment and Instruction in Statistics Education II* (GAISE II) investigative process: formulate

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

statistical investigative questions; collect/consider the data; analyze the data; interpret the results (Bargagliotti et al., 2020) or the PPDAC investigative cycle (Wild & Pfannkuch, 1999): problem, plan, data, analysis, and conclusion. Of note is that the original GAISE cycle called only for the collection of data as does PPDAC, which implies a focus on primary data as opposed to using data already collected by someone else, often for a targeted goal. GAISE II has revised the process to include "consider data." These statistical frameworks involve a focus on the design of experiments and hypotheses tests, but in data science, this focus is insufficient (Gould, 2022). According to Bock (2021), "The end-goal of statistical analysis is often to draw a conclusion about what causes what, based on the quantification of uncertainty. By contrast, the end-goal of data science analysis is more often to do with a specific database or predictive model" (para. 9).

Data science has connections to other emerging topics, in particular computational thinking. Weintrop et al. (2016) define computational thinking using a taxonomy consisting of four categories of practices: data, modeling and simulation, computational problem-solving, and systems thinking. The first two components, data practices and modeling and simulation practices, are central to data science: "Data science is a synthesis of statistics and computational thinking" (Gould, 2022). However, while statistical literacy and quantitative reasoning skills (Burrill, 2020) are important for a literate citizenry, data science goes beyond the idea of literacy and pushes for actually solving problems using a combination of mathematics, statistics, and computational thinking. The connection between mathematics and data science varies depending on the context. At the secondary level such knowledge could involve proportional reasoning, percentages, frequencies, and relative frequencies; regression and mathematical functions such as trigonometric, exponential, or logarithmic; or calculus concepts such as rate of change, accumulation, sequences and series, vectors, and polar coordinates, depending on the context and student background.

APPROACH

Engaging students in meaningful experiences with data should be done in light of the research about developing student understanding, which suggests those experiences should be characterized by a set of principles (Burrill & Dick, 2022a): informal before formal, delayed definitions, scaffolded steps to automaticity, reasoning from visualizations, and communicating and connecting. In particular, when introducing data science at the school level, students might begin with given data and spend time interrogating the data, "noticing and wondering," and identifying any patterns or unusual observations they see. The activities should focus on understanding the words and the context before involving mathematical formulas (Rumsey, 2002) and engage students in simple representations or procedures before those that are more complicated but may be more informative. As the data become more complex, coding becomes useful to manage the data, allowing students to filter by some criteria, use an algorithm to transform the data, or create different visual representations. However, initial experiences with coding should begin by writing codes in meaningful words without mathematical formulas and symbols (Gould, 2022).

A variety of activities have been designed according to the framework and have been used in high school teacher professional development settings and by many teachers in their classrooms, typically with secondary students of varying mathematical backgrounds. These examples include using data science to address pressing societal challenges that often disproportionally affect minoritized groups, such as income gaps or the recent pandemic (Burrill & Dick, 2022a, 2022b). Two examples and how they align with the framework are described below.

EXAMPLES

Example 1: Optimizing Resources Using Census Data (Texas Instruments, 2021b)

Climate change is creating more and more dangerous situations for people across the world. In this activity, students investigate how to optimize resources to maximize support for people in danger of flooding from Hurricane Sandy, one of the most powerful hurricanes of all time. Hurricane Sandy hit the eastern part of the United States on October 2, 2012, caused 70 billion dollars damage, and was responsible for the deaths of at least 75 Americans. Authorities in New York City had to prepare for the likely storm surge and consider how to maximize the allocation of limited resources, in particular a limited number of buses and respirators (*identify the question*). Students investigate data about Staten Island, a part of New York City and one of the hardest hit areas, to answer the question.

The US Census Bureau divides a geographic area into tracts (Figure 1), often following natural boundaries such as major roads or rivers. Students initially look at census tracts on the eastern shore of Staten Island because they seemed most likely to be inundated by a storm surge. The *data were gathered* from a variety of sources, including the U.S. census, New York state, and the U.S. Geologic Survey, which produces topographic maps. For each tract, the data included an estimated average elevation level, the number of blocks, the population, the number and percentage of the population older than 65 and older than 75, and the number and percentage of those identified as white (*process the data*). Some information might be puzzling (see the zeros in Figure 2, row 9). Students have to research information about the predicted storm surge level and decide which tracts are likely to be flooded, then balance that information with the characteristics of the population in each of those tracts to decide where to put the limited number of generators and where to send the 50 busses they had available to move people to less dangerous locations.

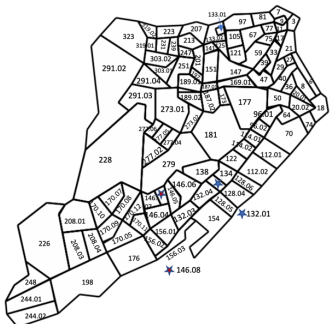

Figure 1. Staten Island census tracts

	A tract	B blocks	C ele_ft	D percgr...	E num_...	F perc_
4	198	129	16	14.3	936	
5	176	135	30	17.4	818	
6	156.01	80	69	18	1159	
7	156.02	51	39	20.7	693	
8	156.03	57	13	12.5	594	
9	154	8	10	0	0	
10	132.03	58	79	22	1190	
11	132.04	85	23	17.8	855	

percgreater65

	A tract	B blocks	C elev_ft	D num_...	E perc_...	F num_
13	134.	66.	20.	327.	7.9	83
14	128.05	30.	10.	207.	8.	4
15	128.06	46.	7.	560.	10.	10
16	128.04	33.	10.	184.	4.6	5
17	122.	58.	23.	223.	8.4	6
18	114.02	38.	23.	284.	8.1	5
19	114.01	30.	10.	170.	5.8	4
20	96.02	45.	20.	233.	7.1	5

num_over75

Figure 2. Census tract elevation and population data, coloring assumptions

A simple way to *explore and visualize* the data is to use color. Suppose the assumptions are any tract with an elevation under 16 feet is likely to flood, and the authorities are concerned about people over the age of 75, who are likely to need help. In the graph on the right in Figure 2, the rows highlighted in green represent the tracts with an elevation under 16 feet, and the orange text indicates tracts with more than 200 people over the age of 75. The intersection of the two colors, as in tracts 128.05 and 128.06, are those tracts that satisfy both constraints and would be good locations for the generators. A more nuanced analysis can be done with a program that allows students to use sliders to change the assumptions (Figure 3), with a given color showing the intersection of the assumptions (*consider models*). Students can check their recommendations with the actual data (*propose action*) and reflect on the statement "Turns out, you can really do well with 50 buses if you have the right data" (Subramanian, 2020). As a follow up from a social justice perspective, they might look for associations between elevation or density and the percentage of the population that is non-white.

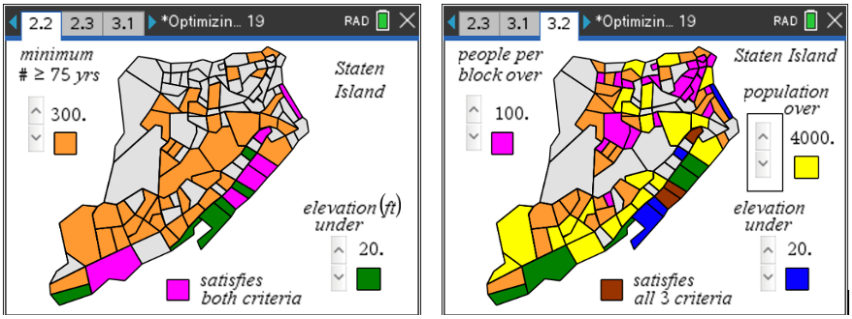

Figure 3. Identifying tracts by age and elevation and by age, elevation, and population density

Example 2: Optimal Location for a Food Truck in Washington DC (Texas Instruments, 2021a)

Students investigate where to locate a food truck in Washington DC to maximize their profit (*identify the problem).* Possible sites are eleven points of interest, chosen either because of nearby tourist attractions or because of the density of businesses and office buildings in the area (Figure 4). They *consider data* such as population density, average daily foot traffic, number of metro fares per day, number of permanent residents in the area, and average income and balance these with the cost of parking a truck in a given location, the number of possible competitors at a given site, the cost of two different storage sites, and the distance from a storage site to possible locations. The data are from a variety of sources, and students might examine some of the assumptions made in the collection process. Students have to decide which variables are important to consider and need to reconcile information from different time spans (per week versus per month). Because the data have different units (distance, cost, rate) and magnitudes (tens to thousands), students have to transform the data into some standard form (*process the data).* This could be done in a variety of ways depending on the mathematical background of the students; for example, simply ranking each variable for each site from best (1) to worst (11), finding the proportion of the best value per variable for each site, or calculating z-scores (*explore and visualize the data).*

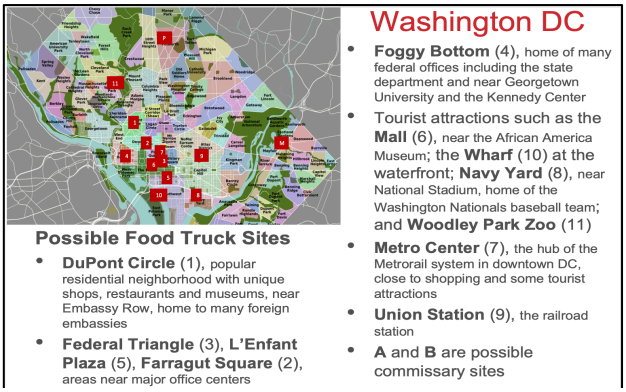

Figure 4. Possible food truck locations in Washington DC

Students might create representations (*consider models)* such as bar graphs or interactive dot plots, and an algebraic model might be the sum of the rankings or weighted *z*-scores such as $T = 0.5H + 0.25I + 2P + 1.5T + 5M + 10V - 5C$ (see Figure 5). Making a decision is complicated by the fact that several possible locations are subject to a lottery with about a 40% chance of getting a spot. After making and justifying their *recommendation* for a site, students are asked to follow up by reading an article describing how successful food trucks choose the best locations and comment on how the article does or does not support their reasoning or interview a local "Truckr" on optimal locations.

GAISE II (Bargagliotti et al., 2020) describes basic skills that should be considered in preparing students to engage in data-based activities. The next section describes several prerequisite skills implicit in the examples above, based on the literature and observations of students working with the tasks.

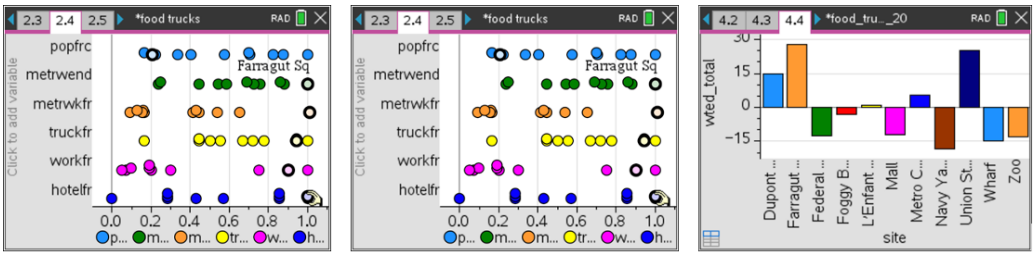

Figure 5. Comparing Farragut Square by fraction of best, *z*-scores, and weighted total

IMPLICATIONS FOR SCHOOL MATHEMATICS

Using Tables

Tables are complicated (Gould, 2022), including how to structure them with enough information to convey what the entries represent: labels, units, order of cases (alphabetized, by magnitude, etc.). Tables also must be easily reproduced in a spreadsheet, which necessitates understanding the software and how labels will be displayed in graphs (given a limited number of characters, what labels will be displayed and will they convey the information). Teachers often want students to fill information in designated cells in empty tables, but this does not let students grapple with what those values mean and how they should be considered in their work. Experiences like those described in GAISE II (Bargagliotti et al., 2020) can help students develop these skills.

Interrogating the Data

Students need early opportunities to collect potentially messy data such as measuring foot length first without specifying procedures (shoes on or off?) or using an ill formed question like how many pets do you have (are fish one pet?). This allows them to develop an understanding of both the data collection process and a sense of why it is important to interrogate secondary data with questions such as: "Who collected these data? When? Where? How? Why?" (Rubin, 2021). Such experiences will also help students manage missing data values and recognize outliers, consider reasons for their presence, and make decisions about how to work with them.

Learning to Communicate

Communicating results using mathematics and words is a skill students need to develop over time. Often students describe only the outcome in words with no mathematics or statistical evidence to support their statement or give a series of mathematical steps without explanatory words (Burrill & Dick, 2022b). They rarely describe why their work makes sense. Early lessons in communicating their thinking in solving mathematical problems and in working with data can build the skills students need to make and justify their reasoning about a problem in data science.

CONCLUSION

We described what data science at the secondary level might be, offered a framework to structure data-centered activities in which students conduct open ended investigations beginning with real problems with real data, and illustrated this with several examples. Early results indicate students are highly engaged ("working bell to bell") and not afraid to experiment with a variety of approaches ("did you try …"); after finding one possible answer, they keep looking for others (Burrill & Dick, 2022b). They explore different questions depending on their interests ("I am Hispanic so I looked at data on Hispanics"). The results also make visible the need to provide support for teachers as they work with their students in unfamiliar territory (Gould, 2022).

Students should be able to make sense of the world, make data-based decisions, think critically about data used as evidence in the social and natural sciences, and use data to solve problems related to a wide variety of contexts. The framework can provide a roadmap for using the data science process to give students experiences that will prepare them to reach these goals.

REFERENCES

Bansal, S. (2020, July 19). What is data science? Roles, skills & courses. *Analytix Labs*. https://www.analytixlabs.co.in/blog/what-is-data-science/

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

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician, 69*(4), 334–342. https://doi.org/10.1080/00031305.2015.1081105

Bock, T. (2021). Statistics vs data science: What's the difference? *DisplayR*. https://www.displayr.com/statistics-vs-data-science-whats-the-difference/

Burrill, G. (2020). Statistical literacy and quantitative reasoning: Rethinking the curriculum. In P. Arnold, (Ed.), *Proceedings of the roundtable conference of the International Association for Statistical Education (IASE).* ISI/IASE. https://iase-web.org/documents/papers/rt2020/IASE2020%20Roundtable%2019_BURRILL.pdf?1610923749

Burrill, G., & Dick, T. (2022a). Connecting mathematics to the world: Engaging students with data science. In J. Morska & A. Rogerson (Eds.), *Building on the past to prepare for the future: Proceedings of the 16th International Conference of the Mathematics Education for the Future Project* (pp.90–94). https://doi.org/10.37626/GA9783959872188.0.017

Burrill, G., & Dick, T. (2022b, May 20). *Connecting to the world through data and data science* [Workshop session]. 2022 Electronic Conference on Teaching Statistics (eCOTS).

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. http://doi.org/10.1080/10618600.2017.1384734

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16(*1), 44–49. https://doi.org/10.52041/serj.v16i1.213

Finzer W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). https://doi.org/10.5070/T572013891

Gould, R. (2022, March 9). *Data science education and statistics education: Why the distinction is needed* [Webinar]. International Association for Statistical Education. https://iase-web.org/Webinars.php?p=220309_2200

International Data Science School Project (IDSSP). (2019). *Curriculum frameworks for introductory data science*. http://idssp.org/files/IDSSP_Frameworks_1.0.pdf

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE-Life Sciences Education, 18*(es2), 1–8. https://doi.org/10.1187/cbe.18-02-0023

Mojica, G., Lee, H., Thrasher, E., Vaskalis, Z., & Ray, G. (2021). Making data science practices explicit in a data investigation process: A framework to guide reasoning about data. In R. Helenius & E. Falck (Ed.), *Proceedings of the satellite conference of the International Association for Statistical Education (IASE).* ISI/IASE. https://doi.org/10.52041/iase.dyjku

Texas Instruments. (2021a). *Optimal locations for a food truck*. Texas Instruments Education Technology. https://education.ti.com/en/timathnspired/us/mathematical-modeling/data-science

Texas Instruments. (2021b). *Optimizing resources using census data*. Texas Instruments Education Technology. https://education.ti.com/en/timathnspired/us/mathematical-modeling/data-science

Rubin, A. (2021). What to consider when we consider data. *Teaching Statistics, 43*(1), S23–S33. https://doi.org/10.1111/test.12275

Rumsey, D. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3). https://doi.org/10.1080/10691898.2002.11910678

Subramanian, S. (2020). *Data disappeared*. Highline Huffpost. https://highline.huffingtonpost.com/article/disappearing-data/#

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25,* 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59(*10), 1–23. https://doi.org/10.18637/jss.v059.i10

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x