Tools to Support Data Analysis and Data Science in K-12 Education

Daniel R. Pimentel
Nicholas J. Horton
Michelle Hoda Wilkerson

September 2, 2022

This paper was commissioned for the Workshop on Foundations of Data Science for Students in Grades K-12, whose work was supported by the Valhalla Foundation. Opinions and statements included in the paper are solely those of the individual authors, and are not necessarily adopted, endorsed, or verified as accurate by the Workshop Planning Committee on Foundations of Data Science for Students in Grades K-12, the Board on Science Education, the Board on Mathematical Sciences and Analytics, the Computer Science and Telecommunications Board, or the National Academy of Sciences, Engineering, and Medicine.

Tools to Support Data Analysis and Data Science in K-12 Education

Daniel R. Pimentel
Graduate School of Education, Stanford University
and
Nicholas J. Horton [*]
Department of Mathematics and Statistics, Amherst College
and
Michelle Hoda Wilkerson
School of Education, University of California, Berkeley

September 2, 2022

Abstract

There has been a proliferation of tools for teaching data analysis and data science at the middle and high school levels. While a few frameworks for systematically exploring the affordances and constraints of such tools exist, most work has only explored one or a few tools at once, or has not focused on K-12 usage. In this paper, we blend first-hand comparative analysis methods and supplemental literature review to conduct a systematic analysis of several common data analysis software packages in use at the K12 level. Using an adaptation of a framework proposed by McNamara (2019), we grouped the tools into related genres. Spreadsheets, while familiar and accessible to many, lacked many desirable features. Visual tools (e.g., CODAP, Social Explorer, iNZight) lower the barrier for data exploration, but may not easily support more advanced statistical tests. Scripting tools (e.g., Python, Pyret, R) provide great flexibility but with increased degree of difficulty. Looking across tools and genres, our analysis suggests that these genres boast complementary strengths depending on students' developmental and investigative needs. We make recommendations for the design and use of tools, notably highlighting the importance of working across different tool types as a part of data practice.

Keywords: data science education; data analytics; K12 education; data tools; spreadsheets

[*] nhorton@amherst.edu

1 Introduction

Over the past few decades, there has been a proliferation of digital tools for teaching data analysis at the middle and high school levels (Hancock et al. 1992, Konold 2007, McNamara 2019). Understanding and effectively using such tools is often cited as one of the more enabling–and difficult–aspects of teaching about data at any level. Issues of access, complexity, functionality, and format abound. While a few frameworks for systematically exploring the affordances and constraints of such tools exist, most explore only one or compare a few tools. This paper instead provides a high-level analysis of the existing landscape of tools being used in K-12 education. Our analysis integrates a review of the evidence base exploring student learning with these technologies, and a first-hand systematic comparative review of several of these tools' key features.

Addressing systemic inequities should be an important part of any conversation about STEM education (NASEM 2022*a*, Dewsbury & Brame 2019). Throughout our analysis, we analyzed tools with respect to dimensions of inclusion and diversity including disability and linguistic diversity. However, we note that tools can play a much broader role in maintaining or disrupting such inequity. Both the types of investigations that tools can support (Washington et al. 2015, Jones et al. 2020), as well as the structure and values embedded within the design of tools themselves (Bang et al. 2013, Paré 2021), have important implications for education. Some preliminary work has been done in data science and statistics (see for example Dana Center 2021, Liao 2022) but much more is needed. Future work on tools for K12 data science should prioritize diversity and inclusion, and their relationship to tools design, to reverse the growing divide that many have noted (Rawlings-Goss et al. 2018).

Section 2 provides an overview of our methods. We introduce the framework that will be adopted in Section 2.1 and describe the focal tools and genres in Section 2.2. A motivating example is introduced in Section 2.3 that is applied (see supplementary materials) to several tools. Section 3 describes our analysis and Section 4 provides closing thoughts and future directions.

2 Our Methodology

Given recent inquiry-based turns across the STEM education landscape including *Next Generation Science Standards (NGSS)* (2013), the GAISE II PreK12 report (Bargagliotti et al. 2020), and *Common Core State Standards Initiative (CCSS)* (2010), we limit our attention here to statistical computing and data analysis tools that are designed to support user-guided explorations of data. In other words, we focus on tools that include flexibility for teachers and students to upload their own datasets, select which visualizations and tools to generate, and otherwise pursue their own paths for investigation. We note that there are also many other useful tools used in Data Science Education for introducing and practicing key statistical methods in a more structured manner (we review these briefly in Section 4.1). Our goal is to highlight tools that are meant to reflect and support learning through the *practice* of data science as a fundamentally technologically-mediated investigative activity.

2.1 Framework for Analysis

McNamara (2019), building on Biehler (1997)'s earlier work, laid out a framework for modern statistical computing tools. The framework articulates a set of ten principles that characterize the needs of novice and professional users (see also McNamara 2016; Kaplan 2007; and Lee et al. 2022). It is meant to shape discussions about the development and use of various statistical computing tools. We began with this lens to examine features of existing K12 data science education tools because although it was primarily designed for software development, it also presents useful considerations for educators considering the affordances and challenges associated with various tools.

#	Feature	Mapping to McNamara (2019)	Description	Relevant Articles
1	Accessibility	Accessibility [1]	Includes cost, simplicity of cloud-based tools, disability access, multilingual support	Marriott et al. (2021); Wu et al. (2021); Rosenberg et al. (2022)
2	Ease of entry	Easy entry for novice users [2]; Inherent documentation [8]	Clarity about how the tool works; includes consideration of students' conceptions of data and developmental appropriateness	Konold (2007); Lehrer et al. (2007); Biehler et al. (2012)
3	Data as a first-order object	Data as a first order persistent object [3]	Data as primary interest: hierarchical vs. tabular formats, viewing data; key to building "students conception of data"	Ben-Zvi & Sharett-Amir (2005); Konold et al. (2015); Haldar et al. (2018)
4	Data analysis cycle; Reproducible workflows	Support for a cycle of exploratory and confirmatory analysis [4]; Simple support for narrative, publishing, and reproducibility [9]	Iterative cycle of posing questions, exploring data, visualizing results, modeling, model assessment, and communicating results; reproducing data wrangling, analyses, and explorations	Biehler et al. (2012); Hancock et al. (1992); Wilkerson et al (2021); Wing (2019); Wise (2020); Lee et al. (2022)
5	Interactivity	Interactivity at every level [7]	Support for direct interaction with data, e.g., pinch, click-and-drag, brushing, hovering	Alexrod & Kahn (2019); McClain & Cobb (2001); Hammett & Dorsey (2020)
6	Flexible plot creation	Flexible plot creation [5]	Univariate, bivariate, and multivariate displays with ability to augment graphics in a variety of ways	Ben-Zvi (2000); Burrill (1996); Pratt (1995)
7	Inferential analysis	Support for randomization throughout [6]	Reasoning with samples and inferring beyond data; support for simulations and resampling; offering probabilistic or uncertain expressions of data	Henriques & Oliveira (2016); Makar & Rubin (2018); Watson (2008); Wild et al. (2011)
8	Non-standard data	-	Working with multiple forms of data such as spatial data, network data, etc.	Doering & Veletsianos (2008); Yoon (2011); SRTL-12 (2022)
9	Extensibility	Flexibility to build extensions [10]	(beyond our scope)	-

Table 1: Framework for considering K12 data science tools, adapted from McNamara's (2019) key attributes for a modern statistical computing tool.

Building on the work of McNamara (2019), we adapted the framework to account for considerations specific to the K12 education context. We expanded the attribute "accessibility" to include explicit attention to support for learners with disabilities, as well as language support for multilingual students. Regarding the attribute "ease of entry", we considered how tools might align with what is known about the development of students' conceptions of data (Konold et al. 2015, Rosenberg & Jones 2022, Lehrer et al. 2007). We broadened the attribute "randomization" to become "support for inferential reasoning," to account for the range of ways tools might support learners with developing and utilizing inference in their analyses (an important learning goal in K12 data science education). Some other changes included adding the attribute, "non-standard data," to account for interest in exploring multiple forms of data in K12 data science education, and combining related attributes such as "reproducibility" and "support for data analysis cycles" to emphasize how these attributes support one another. We did not consider extensibility in our analysis but note its importance in allowing for tools to be adapted toward new conceptual and pedagogical uses.

2.2 Focal Tools and Genres

We have identified a set of software tools that are commonly used in K-12 data oriented lessons, listed below. Our selection of tools was informed by existing literature (e.g., McNamara 2016, Rosenberg et al. 2022) as well as consultation with the field during the process of planning the September, 2022 workshop (NASEM 2022*b*) to elicit examples from practice.

The specific tools are further grouped into larger umbrella "genres" including Spreadsheets; Visual Interfaces; Scripting Languages; and Other Interfaces. We grouped tools into genres to reflect that specific tools may come and go with the passing of time, but are likely to share families of features that will remain more consistent. Considering tool genres allows us to reflect on those more general features that are present (or absent) across a wide range of tools.

The first genre, spreadsheets, are commonly used tools for data analysis in education contexts (McNamara 2016, Rosenberg et al. 2022). They allow for data to be arranged in the rows and columns of a grid, which can be transformed and used for calculations with built-in software functions. Many spreadsheet tools are included part of larger workplace productivity software packages (*Google Sheets* 2022, *Excel* 2022, *Apple Numbers* 2022), and thus are available for free or are already a part of many schools' existing software resources.

Visual tools allow users to flexibly construct graphics to analyze data, typically by using menus or drag-and-drop features. Many of the tools we highlight in this genre are designed specifically for education contexts (*CODAP* 2022, *DataClassroom* 2022, *Tuva* 2022, *iNZight* 2022, Van Wart et al. 2010) or for use by diverse audiences (*Social Explorer* 2022, *Gapminder* 2022, *Tableau* 2022). These tools provide graphical user interfaces that allow students to organize quantitative and qualitative data in the form of tables, graphs, and figures to explore patterns in datasets. Many visual tools also often come with sample datasets for students to explore, and most also support analysis with imported data.

Scripting languages (e.g., Python (2022), R (2021), and Julia (2022)) are commonly used by professional statisticians and data scientists to conduct statistical analyses. They are increasingly being used in education contexts and some languages (e.g., Pyret within Bootstrap, 2022) have been specifically developed for use in education (see for example Fleischer et al. 2022; *Data8* 2022; Pruim et al. 2017; and Thompson & Irgens 2022). These tools are characterized by a base language that allows the user to specify, customize, and automate data analyses using command-line functions (often in an integrated development environment, like RStudio). They also can be modified through libraries or packages that extend the functionality of the tool.

2.2.1 Other tools

Other tools represent a range of products including commonly used commercial products (e.g., Stata (2022) and SPSS (2022)), and integrated environments that support scripting with scaffolded workbook and template structures (e.g., Burckhardt et al. (2021), *Youcubed* (2022), and Colab (2022)).

Integrated environments are frameworks for interactive lessons that introduce students to computational scripts and actions step-by-step by modifying existing code, or writing code incrementally in response to prompts. Some follow the format of a workbook or interactive textbook, while others serve more as templates whose broader structure is meant to be modified and supplemented by students. Most integrated environments have been used with more advanced high school and college students.

Commercial tools are commonly used by professionals to conduct data analyses, and assume high levels of background knowledge on the part of users. These tools often are frequently used in industry or academic contexts and require paid subscriptions or licenses (a major deterrent to adoption according to Rosenberg et al. 2022).

2.3 Our Trial Investigation: Lobster Data

To explore how the selection of data analysis tools can fundamentally shape what is highlighted in an investigation, we conducted a comparative analysis using free, popular tools from three distinct genres: R (scripting), CODAP (visual), and Google Sheets (spreadsheets). (Supplementary materials are available at `https://nicholasjhorton.github.io/K12-Data-Tools`.)

Using each tool, we loaded the same CSV (comma separated value) dataset focused on the recent historical migration patterns of the American Lobster (*Homarus americanus*) off the coast of Northern East Coast of the United States. These data are part of a large curated repository of data about aquatic species (`https://oceanadapt.rutgers.edu`). We filtered the larger dataset to focus only on the mean location (centroid by latitude, longitude, and depth, weighted by biomass) of this species of lobster. The dataset reports these mean locations for different regions/seasons between the years of 1970 and 2020. It is structured such that each row includes one observation record including `year`, `region`, `latitude`, `longitude`, and `depth`.

It is well-established that American lobster populations have been moving north in recent years (Pinsky et al. 2013). Exploring such movement is an example of an activity we might envision in schools as part of a mathematics or science class. We selected this

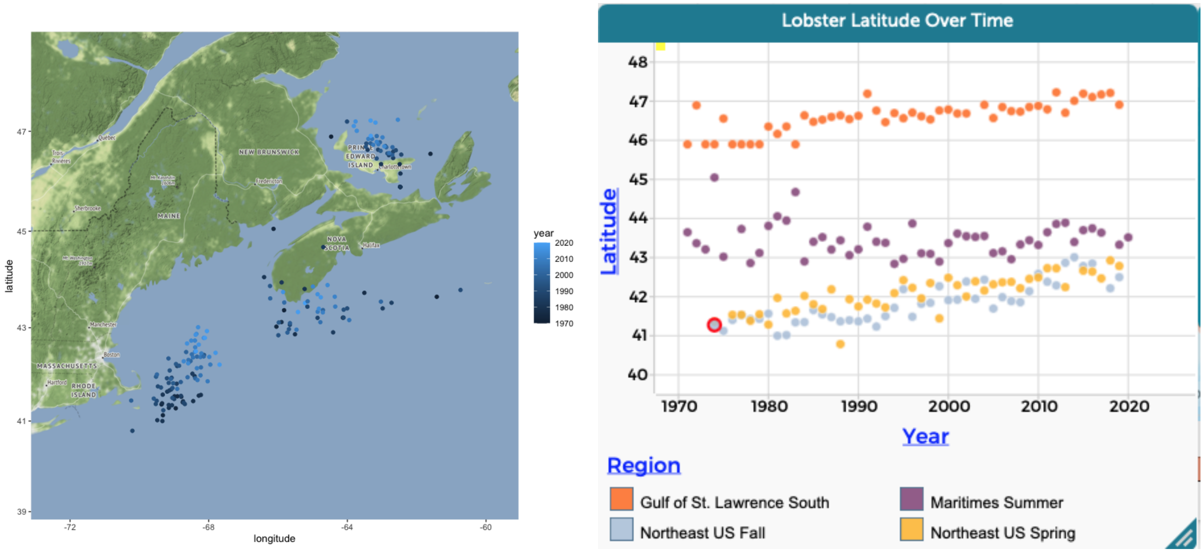

Figure 1: Left: Map of groups of lobster between 1970 and 2020, colored by year. Right: Scatterplot of latitude of lobster populations over time, colored by region.

examination for the vignettes because it reflects a popular 'alternative' data type (geographic data); and there is a clear pattern to observe amidst relatively messy, multivariate data. This example is also deeply embedded in a specific context that connects to other disciplines, and can bring in connections to students' everyday lives through multiple entry points including the media, marketplace, fishing as a hobby and profession, and so on. More generally, we note that data science explorations across the curriculum are increasingly leveraging larger, multivariate datasets (Lee & Wilkerson 2018, Radinsky et al. 2014, Bargagliotti et al. 2020) and that they often feature alternative data types including text data, networks, and image data (Baumer et al. 2021, *AI4K12* 2022). These vignettes minimize "data moves" and wrangling (Erickson et al. 2019) and instead demonstrate other aspects of data visualization and exploration.

Figure 1 features output from our investigations, including a map of the populations over time (using R) and with a linked scatterplot of latitude versus year (using CODAP; colors indicate distinct regions). Both displays provide a visual sense for the northward movement for the American lobster populations. The supplemental materials exemplify differential modes of engagement with the dataset and emergent trends across tool genres. In Google Sheets, it is straightforward to generate a plot of lobster latitude over time. However, distinguishing the different regions by color requires considerable additional work including coding. It is comparable simply to plot distinct regions in CODAP (Figure 1 right) by dragging and dropping the variables to be visualized directly onto the plot. Creating this same plot in R requires coding knowledge, however such knowledge allows the user to more robustly visualize lobster populations' northward movement over time (as featured in Figure 1 left) and could further support simultaneous making regional distinctions by hue.

3 Analysis

In the following section we describe the relative strengths and weaknesses of the primary three genres (Spreadsheets, Visual Tools, and Scripting Languages) in more detail, using our adapted set of attributes in Table 1 as a lens to characterize each genre. We acknowledge the importance and usefulness of other tools in K12 education, however for brevity we focus our analysis specifically on spreadsheets, visual tools, and scripting languages in this section.

3.1 Spreadsheets

When it comes to 'accessibility', spreadsheets are commonly available in K12 educational contexts. Tools like Google Sheets are cloud-based and typically licensed at the district level. Similarly, while Excel requires a license for use, districts typically purchase access as a part of Microsoft Office. These programs typically can be used with screen readers and Excel includes an "Accessibility Checker" which detects issues that arise as users work in a spreadsheet. Spreadsheets also have strengths in 'ease of entry' compared to other tools like scripting languages and commercial software because they allow users to view and organize data in cells rather easily. They also allow for easy data entry and management, which might support students as they learn to consider and collect data. For example, in elementary school, students might begin to record information from groups of interest, and learn that a variable can measure the same characteristic across many individuals (Bargagliotti et al. 2020). In middle school, students might learn about different structures for organizing collected data (Bargagliotti et al. 2020). Spreadsheets can be useful tools for supporting these emerging ideas.

By bringing data to the fore, spreadsheets also support notions of 'data as a first-order, persistent object'. By making it easy to view data directly, spreadsheets might support students in reasoning about case values by examining individual points, an approach commonly used by students in elementary school (Konold et al. 2015). Students are also be able to use simple plots like histograms and bar graphs, which can be used to support reasoning about data as classifiers (i.e., cases with similar values). In addition to recording data, spreadsheets are also a useful tool for introducing students to data moves (Erickson et al. 2019). For example, spreadsheets allow for the creation of filters to only view select cases and to group cases with similar values. Students can also learn to use simple commands to create summary statistics. However, because spreadsheets allow for data to be changed, modified, or deleted directly, original data can also be lost more easily compared to other tools, which often only use copies of original data (McNamara 2016, Broman & Woo 2018). In terms of 'interactivity', spreadsheets have some strengths in that graphics created in them are responsive to changes in associated data values, updating automatically.

Although they are commonly used, spreadsheets tend to fall short in relation to a number of attributes. Support for 'data analysis cycles and reproducibility' is limited, especially at the high school level, the built-in commands provided by the software can constrain lines of inquiry during exploration. Spreadsheets are also limited when it comes to 'flexible plot creation'. They do allow for the creation of common plots and graphics (e.g., scatterplots and histograms), however only a limited number of analyses and plot types are available. Notably, Excel does not have a function for creating dot plots, a visual commonly used to help students reason about distributions in earlier grades (Konold 2007). There is also a lack of general purpose multivariate displays in spreadsheets. Spreadsheets also do not support reproducible work flows, meaning that students need to record their analysis process using some other means (e.g., summarizing steps taken in a written or digital notebook). When it comes to supporting 'inferential reasoning', spreadsheets do have some built-in features to perform statistical tests, (e.g., t-tests), however the tools don't describe to learners how these tests are being performed (or why one would use them). Spreadsheets can be configured to run re-sampling protocols, which might be useful at the high school level, however this can be a rather complex process in a spreadsheet compared to other tool genres. Finally, spreadsheets provide little support for non-standard data. For example, capacity for mapping spatial data is often limited or not supported at all. As a result of these shortcomings, relying solely on spreadsheets as a data analysis tool is likely to lead to a frustrating experience for those learning and engaging in data science.

3.2 Visual

Most of the tools in the visual genre are cloud-based, which allow for data to be uploaded and saved into web-based programs. This is a strength in terms of accessibility. However, they do come with a range of pricing options. CODAP and iNZight (Wild et al. 2021) are freely available while commercial products like Tuva, DataClassroom, and Tableau have free versions but require licenses to use the tools to their full potential. Some tools, like Tableau, offer students one year of free access. Ease of entry is another strength of this genre as many of the tools are specifically designed for educational purposes. CODAP, for example, builds on research conducted with other visual data tools like Tinkerplots and Fathom to enhance usability by students (Mojica et al. 2019). These visual tools often include accessible documentation and options such as text-to-speech, adjustments for color vision, and language translation support. Students can easily upload and view data in the form of tables in these tools. Compared to the spreadsheets genre, however, these tools are often less amenable to data input and collection.

Tools in the visual genre have primary strengths in representing data as a first-order, persistent object, flexible plot creation, and interactivity. Accordingly, they are great for engaging students in exploratory data analysis and interpretation of results. For example, CODAP and Tuva allow students to flexibly create and move between scatter plots by dragging variables to axes. Interactive features include hot-linking data points across multiple formats (tables, scatter plots, bar charts, etc.) or allowing students to "scrub" back and forth between maps or time series graphs allow students to explore patterns. This can support students' exploratory data analysis, and help them reason about data from different lenses (e.g., examining data as case values in a table, as classifiers in a dot plot, and in aggregate by overlaying a box plot onto a dot plot) (Konold et al. 2015). Students can use these tools to calculate summary statistics and aggregate values. Another strength is in supporting nonstandard data, in particular geospatial data through mapping.

While visual tools have strengths in supporting data analysis, they are limited in their support for reproducibility. Due to the interactive nature of these tools (e.g., use of click-and-drag, pinching, swiping, etc.), moves made are often not recorded. This genre is also varied in its support for inferential reasoning. Some tools like CODAP and many public visualization tools like Social Explorer have limited support for sampling, randomization, or statistical testing (e.g., t-tests, ANOVAs, etc.). This means that these tools may reach an upper limit as students learn to engage in inferential statistics, run simulations, etc. Other visual tools, however, like DataClassroom, excel with inferential reasoning, including by providing visual scaffolds for students so that they can learn how to interpret their results.

3.3 Scripting

Scripting languages like R, Python, and Pyret are freely available and can be either desktop or cloud-based, allowing for relative 'accessibility' in educational contexts. Scripting languages can be made accessible to those with disabilities with the support of add-on packages (e.g., BrailleR for those with visual impairment Godfrey 2013; Godfrey & Loots 2015; see also Schanzer et al. 2020) to increase functionality for a wide range of users. The use of notebooks (e.g., Jupyter for Python, *Data8* 2022, Kim & Henke 2021) and other tools, like RMarkdown (Baumer et al. 2014) and Quarto, allow for the incorporation of narrative and audience-friendly publication of analyses. Scripting tools also often allow for commands and analyses to be saved easily, making reproducibility a strength. Packages such as `ggplot2` in R and `matplotlib` in Python make it easy to create a range of graphics and plots. The flexibility of scripting tools allows for both exploratory and confirmatory analysis cycles to be conducted although exploratory analyses may not be as interactive as with the visual tool genre.

Of all of the tool genres, scripting languages provide the most functionality. However they often have relatively steep learning curves, making 'ease of entry for novices' a key limitation. Scripting languages also tend to have less of a focus on interactivity, prioritizing text-based commands. While this is the case, users can often employ graphical user interfaces (e.g., Commander & Reducer) or integrated development environments (e.g., RStudio & PyCharm) to increase interactivity of this tool genre with menus, clickable icons, and other features. Various approaches (e.g. data.tables for Data 8 Python and mosaic/formula interface for R [Pruim et al 2017]) can help to minimize barriers for newer users. Additionally, scripting languages tend to hide data unless they are explicitly pulled up by a user, making 'data as a first-order object' a less prominent feature. These tools also tend to assume that users have well-developed conceptions of data organization and structure. They require a substantial amount of background knowledge and may be most appropriate for older high school students and those transitioning to post-secondary data work.

4 Discussion

Twenty-five years ago, Biehler (1997) noted the need for tools and computing environments that would allow secondary and postsecondary students to learn statistics in the way that statisticians practice. Adopting and modifying McNamara (2019)‘s framework allowed us to examine features of various K12 tool genres, and map these features to known developmental issues in the learning and doing of data analysis. One can imagine a trajectory where students are introduced to data work using spreadsheet tools as they learn to develop statistical investigative questions and collect and manage relevant data. As students' conceptions of data continue to develop in elementary and middle school, they might then be introduced to visual tools, which allow for more flexibility in creating and working across representations including tables, plots, and visuals. These features support students as they learn to think about data in the aggregate. Alternatively, teachers may want to simultaneously take advantage of the affordances of both spreadsheets and visual tools by having students in early grades collect and organize data in spreadsheets and then upload them into visual software for further exploration and analysis. Students may continue to work with visual tools in high school, perhaps transitioning toward those with capabilities for inferential reasoning (e.g., Data Classroom). Once students have sufficient background in working with data analysis cycles (e.g., asking questions, collecting data, analyzing data, and interpreting results), they may transition toward scripting languages to conduct more advanced data work and prepare for engagement with these tools in post-secondary environments.

As learners transition into post-secondary settings, they will need to be supported with increasing sophistication and complexity presented in data workflows and tools. Educators might support students with some aspects of this transition by engaging students in comparative data analysis using multiple tools simultaneously to help them transition from familiar tools to new, more complex tools (such as learning to conduct the same analyses in CODAP and R, for example). In addition, students will likely need to learn about rigorous standards for reproducibility often employed in post-secondary settings and how to leverage tools that support this feature (e.g. Colab and other sophisticated cloud-based computing tools). Students will also need an understanding of different workflows and version control so that they are prepared to engage with tools that explicitly support collaborative work (e.g., GitHub). In post-secondary work, students will also need to be able to work with multivariate data and tools that support advanced analyses with these types of data.

4.1 Other Considerations

No discussion of educational tools should be done in isolation. Tools are only as good as the instructional systems, curricula, and learning experiences in which they are embedded (Philip & Garcia 2013, Roschelle et al. 2010). This paper was developed as part of a collection of materials and panels, and we look forward to further discussion around these issues. We also very briefly note particularly important points of contact between data analysis tools like the ones we discuss here and other aspects of curriculum and instruction.

First, this paper and the framework it presents focus squarely on features of software tools that support data analysis and associated statistical methods as these may be practiced across the K12 curriculum. We note, however, that several tools make additional connections to other potential learning goals. Scripting tools such as Bootstrap, R, and Python, for example, can additionally support the development of computational thinking and programming (e.g., Schanzer et al. 2022); spreadsheets and commercial packages can support familiarity with business practices and applications (e.g., Kazak et al. 2021); and the availability of certain visual representations including maps, graphs, box plots, and so on can connect to content in history, science, and mathematics (Radinsky et al. 2014, Lee & Wilkerson 2018).

Another important consideration for data work in K12 contexts is the availability of engaging and appropriate datasets for students. Some tools we discuss above (e.g., GapMinder, Bootstrap, R/RStudio, iNZight, and CODAP) come with pre-installed educational datasets. Educators should consider how features of these datasets align with their learning goals for students. Careful selection and framing of datasets can also support socially relevant and justice-oriented investigations that benefit from students' already-rich knowledge of place and practice (though educators should avoid false promises about the power of data; see Van Wart et al. (2020) and Rubel et al. (2016)). There have also been ongoing efforts to create repositories of educational datasets, though challenges remain in findings ways to keep such datasets current, well-documented, and usable (Wallis et al. 2006). Rubin & Mokros (2018)'s consideration of what are the characteristics of "Goldilocks data"–data that is not too big and not too small, but just right for student engagement–is still an area for research.

A number of digital tools exist that are not necessarily intended to support full cycles of data exploration, but can nevertheless provide important learning opportunities for students in this space. This includes tools that provide hints to students as they work on specific skill-building problems (e.g., *ASSISTments* 2022, *STATS4STEM* 2022), visualize or otherwise interact with specific statistical methods (PhET (2022)), or for guiding students through structured activities that introduce key computational and/or conceptual content to students (*JupyterHub* 2022, Kim & Henke 2021, *CourseKata* 2022, Tucker et al. 2022).

4.2 Recommendations

The framework and analyses we present here can serve as further guideposts for the development of K12 data science tools. For example, future development might consider how tools with existing strengths can be more intentionally engineered to support additional valuable features for learning. While an attractive aspect of visual tools is their interactive nature, one can imagine allowing "record" mode, which would capture interactive moves (e.g., drags, clicks, etc.) made to analyze the data in a reproducible manner (Nolan & Temple Lang 2007).

Our analysis, consistent with the guidance of Biehler (1997) and McNamara (2019), also suggests the importance of movement across and between tools depending on student needs and conceptual development. One tool does not, and should not, fit all. Rather, we should prepare students and teachers to work within a rich ecosystem of complementary tools for teaching and learning about data.

McNamara (2016) articulated that most tools can be categorized as those used for learning to work with data (e.g., DataClassroom, CODAP, Bootstrap, and Tuva) and those used for doing statistics and data work (e.g., R and Python). In characterizing these tools this way, McNamara (2016) recommends that future tools be designed to "bridge the gap" by providing supports for new users while also keeping advanced users from "experiencing out" of the tool (e.g., block-based languages that are easy to learn but also support advanced learners in using scripting languages) (p. 382). We agree with this recommendation, particularly as K12 data science education seeks to help learners apply data analyses to solve authentic problems (e.g., Bargagliotti et al. 2020). Tools should support students in seeing themselves not only as "learners" of data science, but as "doers" as well. As designers continue to develop and refine K12 data science tools, we note Biehler (1997)'s recommendation that these tools be prototyped and studied within laboratory and classroom settings to better understand how design features support student learning.

5 Acknowledgements

Danny Pimentel is first author of this paper. Nicholas Horton and Michelle Hoda Wilkerson contributed equally as supporting authors. This paper was commissioned by the National Academies of Science in anticipation of the Foundations of Data Science for Students in Grades K-12 Workshop to be held September 13-14 in Washington, DC and online. Some of the material discussed in this paper was developed under National Science Foundation grant IIS-1900606.

References

AI4K12 (2022).
URL: *https://ai4k12.org*

Apple Numbers (2022).
URL: *https://www.apple.com/numbers/*

ASSISTments (2022).
URL: *https://new.assistments.org*

Axelrod, D. B. & Kahn, J. (2019), Intergenerational family storytelling and modeling with large-scale data sets, *in* 'Proceedings of the 18th ACM International Conference on Interaction Design and Children', pp. 352–360.

Bang, M., Marin, A., Faber, L. & Suzukovich III, E. S. (2013), 'Repatriating indigenous technologies in an urban indian community', *Urban Education* **48**(5), 705–733.

Bargagliotti, A., Franklin, C. A., 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., Alexandria, VA. https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports.

Baumer, B., Çetinkaya-Rundel, M., Bray, A., Loi, L. & Horton, N. J. (2014), 'R markdown: Integrating a reproducible analysis tool into introductory statistics', *Technology Innovations in Statistics Education* **8**(1).
URL: *https://escholarship.org/uc/item/90b2f5xh*

Baumer, B. S., Kaplan, D. T. & Horton, N. J. (2021), *Modern Data Science with R*, 2nd edn, Chapman and Hall/CRC Press: Boca Raton.
URL: *https://mdsr-book.github.io/mdsr2e*

Ben-Zvi, D. (2000), 'Toward understanding the role of technological tools in statistical learning', *Mathematical Thinking and Learning* **2**(1-2), 127–155.

Ben-Zvi, D. & Sharett-Amir, Y. (2005), How do primary school students begin to reason about distributions, *in* 'Reasoning about distribution: A collection of current research studies. Proceedings of the fourth international research forum on statistical reasoning, thinking, and literacy (SRTL-4), University of Auckland, New Zealand', pp. 2–7.

Biehler, R. (1997), 'Software for learning and for doing statistics', *International Statistical Review* **65**(2), 167–189. https://doi.org/10.1111/j.1751-5823.1997.tb00399.x.

Biehler, R., Ben-Zvi, D., Bakker, A. & Makar, K. (2012), Technology for enhancing statistical reasoning at the school level, *in* 'Third International Handbook of Mathematics Education', Springer, pp. 643–689.

Bootstrap (2022).
URL: *http://bootstrapworld.org*

Broman, K. W. & Woo, K. H. (2018), 'Data organization in spreadsheets', *The American Statistician* **72**(1), 2–10.
URL: *https://doi.org/10.1080/00031305.2017.1375989*

Burckhardt, P., Nugent, R. & Genovese, C. R. (2021), 'Teaching statistical concepts and modern data analysis with a computing-integrated learning environment', *Journal of Statistics and Data Science Education* **29**(sup1), S61–S73.
URL: *https://doi.org/10.1080/10691898.2020.1854637*

Burrill, G. (1996), 'Graphing calculators and their potential for teaching and learning statistics', *Research on the Role of Technology in Teaching and Learning Statistics* pp. 24–37. https://chance.dartmouth.edu/teaching_aids/IASE/IASE.book.pdf.

CODAP (2022).
URL: *https://codap.concord.org*

Common Core State Standards Initiative (CCSS) (2010).
URL: *http://www.corestandards.org*

CourseKata (2022).
URL: *https://coursekata.org*

Dana Center (2021), 'Data science course framework'.
URL: *https://www.utdanacenter.org/sites/default/files/2021-05/data_science_course_framework_2021_final.pdf*

Data8 (2022).
URL: *http://data8.org*

DataClassroom (2022).
URL: *https://about.dataclassroom.com*

Dewsbury, B. & Brame, C. J. (2019), 'Inclusive teaching', *CBE—Life Sciences Education* **18**(2), 1–5.
URL: *https://www.lifescied.org/doi/epdf/10.1187/cbe.19-01-0021*

Doering, A. & Veletsianos, G. (2008), 'An investigation of the use of real-time, authentic geospatial data in the K–12 classroom', *Journal of Geography* **106**(6), 217–225.

Erickson, T., Wilkerson, M., Finzer, W. & Reichsman, F. (2019), 'Data moves', *Technology Innovations in Statistics Education* **12**(1).
URL: *https://doi.org/10.5070/T5121038001*

Excel (2022).
URL: *https://www.microsoft.com/en-us/microsoft-365/excel*

Fleischer, Y., Biehler, R. & Schulte, C. (2022), 'Teaching and learning data-driven machine learning with educationally designed Jupyter notebooks', *Statistics Education Research Journal* **21**(2).
URL: *https://doi.org/10.52041/serj.v21i2.61*

Gapminder (2022).
URL: *https://gapminder.org*

Godfrey, A. J. R. (2013), 'Statistical software from a blind person's perspective', *R Journal* **5**(1).
URL: *https://doi.org/10.32614/RJ-2013-007*

Godfrey, A. J. R. & Loots, M. T. (2015), 'Advice from blind teachers on how to teach statistics to blind students', *Journal of Statistics Education* **23**(3).
URL: *https://doi.org/10.1080/10691898.2015.11889746*

Google (2022), 'Colaboratory'.
URL: *https://colab.research.google.com*

Google Sheets (2022).
URL: *https://www.google.com/sheets/about/*

Haldar, L. C., Wong, N., Heller, J. I. & Konold, C. (2018), 'Students making sense of multi-level data', *Technology Innovations in Statistics Education* **11**(1).

Hammett, A. & Dorsey, C. (2020), 'Messy data, real science', *The Science Teacher* **87**(8), 40–49.

Hancock, C., Kaput, J. J. & Goldsmith, L. T. (1992), 'Authentic inquiry with data: Critical barriers to classroom implementation', *Educational Psychologist* **27**(3), 337–364.

Henriques, A. & Oliveira, H. (2016), 'Students' expressions of uncertainty in making informal inference when engaged in a statistical investigation using tinkerplots', *Statistics Education Research Journal* **15**(2), 62–80.

iNZight (2022).
URL: *https://inzight.nz/*

Jones, S. T., Thompson, J. & Worsley, M. (2020), 'Data in motion: Sports as a site for expansive learning', *Computer Science Education* **30**(3), 279–312.

Julia programming language (2022).
URL: *https://julialang.org*

JupyterHub (2022).
URL: *https://jupyter.org/hub*

Kaplan, D. (2007), 'Computing and introductory statistics', *Technology Innovations in Statistics Education* **1**(1).
URL: *https://doi.org/10.5070/T511000030*

Kazak, S., Fujita, T. & Turmo, M. P. (2021), 'Students' informal statistical inferences through data modeling with a large multivariate dataset', *Mathematical Thinking and Learning* pp. 1–21.

Kim, B. & Henke, G. (2021), 'Easy-to-use cloud computing for teaching data science', *Journal of Statistics and Data Science Education* **29**(sup1), S103–S111.
URL: *https://doi.org/10.1080/10691898.2020.1860726*

Konold, C. (2007), 'Designing a data analysis tool for learners', *Thinking with Data* pp. 267–291.

Konold, C., Higgins, T., Russell, S. J. & Khalil, K. (2015), 'Data seen through different lenses', *Educational Studies in Mathematics* **88**(3), 305–325.
URL: *https://link.springer.com/10.1007/s10649-013-9529-8*

Lee, H., Mojica, G., Thrasher, E. & Baumgartner, P. (2022), 'Investigating data like a data scientist: key practices and processes', *Statistics Education Research Journal* **21**(2).
URL: *https://doi.org/10.52041/serj.v21i2.41*

Lee, V. R. & Delaney, V. (2022), 'Identifying the content, lesson structure, and data use within pre-collegiate data science curricula', *Journal of Science Education and Technology* **31**(1), 81–98.

Lee, V. R. & Wilkerson, M. H. (2018), 'Data use by middle and secondary students in the digital age: A status report and future prospects'.

Lehrer, R., Kim, M.-j. & Schauble, L. (2007), 'Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability', *International Journal of Computers for Mathematical Learning* **12**(3), 195–216.

Liao, S.-M. (2022), 'SCRATCH to R: Toward an inclusive pedagogy in teaching coding', *Journal of Statistics and Data Science Education* pp. 1–18.
URL: *https://doi.org/10.1080/26939169.2022.2090467*

Makar, K. & Rubin, A. (2018), 'Learning about statistical inference', *International Handbook of Research in Statistics Education* pp. 261–294.

Marriott, K., Lee, B., Butler, M., Cutrell, E., Ellis, K., Goncu, C., Hearst, M., McCoy, K. & Szafir, D. A. (2021), 'Inclusive data visualization for people with disabilities: a call to action', *Interactions* **28**(3), 47–51.

McClain, K. & Cobb, P. (2001), 'Supporting students' ability to reason about data', *Educational Studies in Mathematics* **45**(1), 103–129.

McNamara, A. (2016), ‘On the state of computing in statistics education: Tools for learning and for doing’.
URL: *https://arxiv.org/abs/1610.00984*

McNamara, A. (2019), ‘Key attributes of a modern statistical computing tool’, *The American Statistician* **73**(4), 375–384.
URL: *https://doi.org/10.1080/00031305.2018.1482784*

Mojica, G. F., Azmy, C. N. & Lee, H. S. (2019), ‘Exploring data with CODAP’, *The Mathematics Teacher* **112**(6), 473–476.
URL: *https://pubs.nctm.org/view/journals/mt/112/6/article-p473.xml*

NASEM (2022*a*), ‘Equity in PreK-12 STEM education’.
URL: *https://www.nationalacademies.org/our-work/equity-in-prek-12-stem-education*

NASEM (2022*b*), ‘Foundations of data science for students in grades K-12: A workshop’.
URL: *https://www.nationalacademies.org/our-work/foundations-of-data-science-for-students-in-grades-k-12-a-workshop*

Next Generation Science Standards (NGSS) (2013).
URL: *https://www.nextgenscience.org*

Nolan, D. & Temple Lang, D. (2007), ‘Dynamic, interactive documents for teaching statistical practice’, *International Statistical Review* **75**(3), 295–321.
URL: *https://doi.org/10.1111/j.1751-5823.2007.00025.x*

Paré, D. (2021), A critical review and new directions for queering computing and computing education, *in* G. Noblit, ed., ‘Oxford Research Encyclopedia of Education’, Oxford University Press.

PhET Interactive Simulations (2022).
URL: *https://phet.colorado.edu*

Philip, T. & Garcia, A. (2013), ‘The importance of still teaching the igeneration: New technologies and the centrality of pedagogy’, *Harvard Educational Review* **83**(2), 300–319.

Pinsky, M. L., Worm, B., Fogarty, M. J., Sarmiento, J. L. & Levin, S. A. (2013), ‘Marine taxa track local climate velocities’, *Science* **341**(6151), 1239–1242.

Pratt, D. (1995), ‘Young children’s active and passive graphing’, *Journal of Computer Assisted Learning* **11**(3), 157–169.

Pruim, R., Kaplan, D. T. & Horton, N. J. (2017), ‘The mosaic Package: Helping Students to ‘Think with Data’ Using R’, *The R Journal* **9**(1), 77–102.
URL: *https://doi.org/10.32614/RJ-2017-024*

Python software foundation (2022), ‘Python’.
URL: *https://python.org*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Radinsky, J., Hospelhorn, E., Melendez, J. W., Riel, J. & Washington, S. (2014), 'Teaching american migrations with gis census webmaps: A modified "backwards design" approach in middle-school and college classrooms', *The Journal of Social Studies Research* **38**(3), 143–158.

Rawlings-Goss, R., Cassel, L., Cragin, M., Cramer, C., Dingle, A., Friday-Stroud, S., Herron, A., Horton, N. J., R, I. T., Jordan, K., Ordonez, P., Rudis, M., Rwebangira, R., Schmitt, K., Smith, D. & Stephens, S. (2018), 'Keeping data science broad: Negotiating the digital and data divide among higher education institutions'.
URL: *https://southbigdatahub.org/resources/newsblog/keeping-data-science-broad-program*

Roschelle, J., Knudsen, J. & Hegedus, S. (2010), From new technological infrastructures to curricular activity systems: Advanced designs for teaching and learning, *in* 'Designs for Learning Environments of the Future', Springer, pp. 233–262.

Rosenberg, J. & Jones, R. S. (2022), NASEM workshop commissioned paper on student learning.

Rosenberg, J. M., Schultheis, E. H., Kjelvik, M., Reedy, A. & Sultana, O. (2022), 'Big data, big changes? the technologies and sources of data used in science classrooms', *British Journal of Educational Technology* **53**(5), 1179–1201. Publisher: John Wiley & Sons, Ltd.
URL: *https://doi.org/10.1111/bjet.13245*

Rubel, L. H., Hall-Wieckert, M. & Lim, V. Y. (2016), 'Teaching mathematics for spatial justice: Beyond a victory narrative', *Harvard Educational Review* **86**(4), 556–579.

Rubin, A. & Mokros, J. (2018), Data clubs for middle school youth: Engaging young people in data science, *in* 'Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute'. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9B2.pdf.

Schanzer, E., Bahram, S. & Krishnamurthi, S. (2020), Adapting student ides for blind programmers, *in* 'Koli Calling'20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research', pp. 1–5.

Schanzer, E., Pfenning, N., Denny, F., Dooman, S., Politz, J. G., Lerner, B. S., Fisler, K. & Krishnamurthi, S. (2022), Integrated data science for secondary schools: Design and assessment of a curriculum, *in* 'Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1', pp. 22–28.

Social Explorer (2022).
URL: *https://socialexplorer.com*

SPSS Statistics (IBM) (2022).
URL: *https://www.ibm.com/products/spss-statistics*

SRTL-12 Final Announcement: Rethinking learners' reasoning with non-traditional data (2022).
URL: *https://blogs.uni-paderborn.de/srtl/files/2022/01/SRTL12-Final-Announcement-updated-11-January-2022.pdf*

Stata (2022).
URL: *https://www.stata.com*

STATS4STEM (2022).
URL: *https://www.stats4stem.org*

Tableau (2022).
URL: *https://tableau.com*

Thompson, J. & Irgens, G. A. (2022), 'Data detectives: A data science program for middle grade learners', *Journal of Statistics and Data Science Education* **30**(1), 29–38.
URL: *https://doi.org/10.1080/26939169.2022.2034489*

Tucker, M. C., Shaw, S. T., Son, J. Y. & Stigler, J. W. (2022), 'Teaching statistics and data analysis with r', *Journal of Statistics and Data Science Education* pp. 1–15.
URL: *https://doi.org/10.1080/26939169.2022.2089410*

Tuva (2022).
URL: *https://tuvalabs.com*

Van Wart, S., Lanouette, K. & Parikh, T. S. (2020), 'Scripts and counterscripts in community-based data science: Participatory digital mapping and the pursuit of a third space', *Journal of the Learning Sciences* **29**(1), 127–153.
URL: *https://www.tandfonline.com/doi/full/10.1080/10508406.2019.1693378*

Van Wart, S., Tsai, K. J. & Parikh, T. (2010), Local ground: A paper-based toolkit for documenting local geo-spatial knowledge, *in* 'Proceedings of the First ACM Symposium on Computing for Development', pp. 1–10.

Wallis, J. C., Milojevic, S., Borgman, C. L. & Sandoval, W. A. (2006), 'The special case of scientific data sharing with education', *Proceedings of the American Society for Information Science and Technology* **43**(1), 1–13.

Washington, A. N., Burge, L., Mejias, M., Jean-Pierre, K. & Knox, Q. (2015), Bridging the divide: Developing culturally-responsive curriculum for K-12 computer science education, *in* 'Proceedings of the 46th ACM Technical Symposium on Computer Science Education', pp. 707–707.

Watson, J. M. (2008), 'Exploring beginning inference with novice grade 7 students.', *Statistics Education Research Journal* **7**(2).

Wild, C. J., Elliott, T. & Sporle, A. (2021), 'On democratizing data science: some iNZights to empowering the many', *Harvard Data Science Review* **3**(2). https://doi.org/10.1162/99608f92.85206ff9.

Wild, C., Pfannkuch, M., Regan, M. & Horton, N. J. (2011), 'Towards more accessible conceptions of statistical inference', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(2), 247–295.

Wilkerson, M., Finzer, W., Erickson, T. & Hernandez, D. (2021), Reflective data storytelling for youth: The CODAP story builder, *in* 'Interaction Design and Children', pp. 503–507.

Wing, J. M. (2019), 'The data life cycle', *Harvard Data Science Review* .

Wise, A. F. (2020), 'Educating data scientists and data literate citizens for a new generation of data', *Journal of the Learning Sciences* **29**(1), 165–181.

Wu, K., Petersen, E., Ahmad, T., Burlinson, D., Tanis, S. & Szafir, D. A. (2021), Understanding data accessibility for people with intellectual and developmental disabilities, *in* 'Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems', pp. 1–16.

Yoon, S. A. (2011), 'Using social network graphs as visualization tools to influence peer selection decision-making strategies to access information about complex socioscientific issues', *Journal of the Learning Sciences* **20**(4), 549–588.

Youcubed (2022).
URL: *https://hsdatascience.youcubed.org/curriculum*