International Society of the Learning Sciences

Storytelling *With* and *About* Data: Mapping the Terrain

Kyla A. Kemble, UC Berkeley, kyla_kemble@berkeley.edu
Michelle H. Wilkerson, UC Berkeley, mwilkers@berkeley.edu

Abstract: "Data storytelling" is described in a variety of ways in literature, and even within the same project what constitutes a "data story" can vary among learners. These different treatments are likely to support different engagements with data, and therefore different learning opportunities for students. Here, we describe preliminary efforts to characterize the variety of ways in which data stories may differ in their mode (e.g., story about work with data and story about the data's implications) and in their features (e.g., attention to data source; attention to history; case vs aggregate reasoning). To illustrate, we present an analysis of two data story artifacts produced by adolescents that participated in the same data storytelling workshop focused on health and the environment.

Introduction

A growing number of projects, curricula, and initiatives are focused on *data storytelling* (for a recent example from the learning sciences, see Matuk, et al. 2022). This focus builds on decades of work spanning the learning sciences (e.g., Hall, Wright, & Weikert, 2007), statistics education (e.g., Pfannkuch et al. 2010), and the information and data sciences (e.g., Segel & Heer, 2010) that emphasize the role of narration and storytelling in work involving data. Many such projects argue that by engaging in data storytelling, students practice clearly reasoning and communicating about data, and can also develop an understanding of how data may (or may not) be used to understand and advocate for themselves and their communities.

It is less clear, however, what could constitute a "data story" in these educational contexts. Looking across recent work in the learning sciences and allied fields suggests there is wide variability in how data storytelling is conceptualized in research, and in how it is presented to teachers and students in educational contexts. Some definitions emphasize a data story as highlighting key features or findings of a data analysis, such as the "data visualization and text narrative" described by Sarei and colleagues (2023). Others describe a data story as emphasizing the process of data analysis, making visible the practices and decisions that support sensemaking (Pfannkuch, Regan, Wild, & Horton, 2010). Yet others emphasize the ways in which a data story might illuminate how data are embedded within, and imperfectly reflect, more complex social dynamics and histories (Kahn, 2020), or how data counter-storytelling might offer ways for learners to dispute or reframe disparaging narratives about their own communities (Amato et al, 2022).

Even within the same educational contexts, the nature of a data story may vary across students and situations. Storniauolo (2020) describes how considerations such as privacy and aesthetics led students participating in the same arts-based data workshop to produce very different "data stories" as public artifacts. Similarly, Radinsky (2020) described how both students and educational administrators moved between what he called multiple modes of data narrative. These modes included: (1) narrating what was done analytically with data, (2) animating the data representation in order to highlight patterns and findings, (3) incorporating the animated data into broader socio historical narratives, and (4) narrating oneself into the integrated, data represented world. Each narrative mode, Radinsky argued, offered different learning opportunities and different ways of shaping others' access to and perception of the data and associated arguments.

Here, we describe ongoing work to map the terrain of data storytelling as it is practiced by students. Our driving question is: How can we characterize the diversity of ways in which data storytelling is taken up in both the processes and products of student work? We propose key dimensions along which we have observed data storytelling to vary in our preliminary analyses of student artifacts. We then present a closer analysis of two very different data stories written and reported by adolescent youths who participated in the same workshop offering. Our goal is not to advocate for a particular approach to data storytelling, but rather to understand how different approaches to data storytelling might support complementary learning goals.

Study Context: Writing data stories project

During Summer and Fall of 2021, our project offered a series of four approximately 15-hour out-of-school educational workshops focused on data storytelling for students in the middle grades. The workshops were advertised to students and parents as an opportunity to explore issues of environmental injustice, and students were supported in examining the historical and social foundations of environmental justice through local and global case studies. They were then encouraged to select a specific issue to explore further using a large, locally situated public dataset with geospatial data that included health, demographic, and environmental indicators. Students' final products from the workshop were "data stories" using this dataset.

We explicitly framed the data story as a way for students to share their motivations for exploring a question with data, their processes and decisions regarding how to analyze the data, and to use the data to support arguments and "calls to action" to address unjust circumstances in their lives and communities. The workshops were held online using the Zoom video conference service at a time when many schools in the United States were still disrupted due to the global COVID-19 pandemic. Upon reviewing the stories students constructed during these workshops, we found that even though we defined and modeled a particular type of data storytelling, the stories constructed by participants varied in focus and composition.

Data sources

The preliminary analyses reported here are based on the final data story artifacts that students constructed during these workshops, as well as video recordings of their verbal narrations of these stories during the last workshop session. Data stories were primarily constructed using Google Slides and often featured combinations of text, images, screenshots (e.g., of news articles), maps and scatter plots or histograms. Future work will also include analysis of videos of students' consultations with workshop facilitators, check-ins with peers, and other data sources that will provide more insight into the development of these stories over time.

Analysis

Analysis is ongoing but based on preliminary open coding of students' data stories and informed by literature in data science and statistics education (e.g., GAISE II, 2020; Rubin, 2020), we developed an initial coding scheme described in Table 1. For this analysis, we developed two main categories of codes which focus on articulating considerations *about* the social and contextual nature of the data, and articulating statistical methods conducted *with* data to make inferences and conclusions.

Table 1

Features of Stories About and With Data

Story *About* Data	A-SEL	Describes process of data selection.
	A-SOU	Describes data source(s).
	A-HIS	Describes context/history of the data.
	A-MOV	Describes transformations or representations applied to data.
	A-RV	Describes reliability/validity of data.
	A-ALT	Describes consideration of alternative/null patterns or findings.
	A-MOT	Describes one's motivation for analyzing the data.
Story *With* Data	W-CAS	Identifies specific cases as they manifest within the data.
	W-AGG	Identifies patterns or trends as they manifest within the data.
	W-SUP	Describes how data lends support for an argument.
	W-REV	Describes how data suggests refinement of argument.
	W-REJ	Describes how data suggests rejection of an argument.
	W-HYP	Describes how well data generalizes to population/new situations.
	W-CON	Describes final conclusions from data.

Results

We highlight two data stories constructed by students from the same workshop session. We selected these stories because they highlight complementary approaches to data storytelling, even as they leverage similar representations of the same dataset. These two data stories were also selected because of the density of features we were able to illustrate in a single screenshot of each. Our analysis, however, describes features identified in the full finalized stories, including slides that are not featured.

Student A

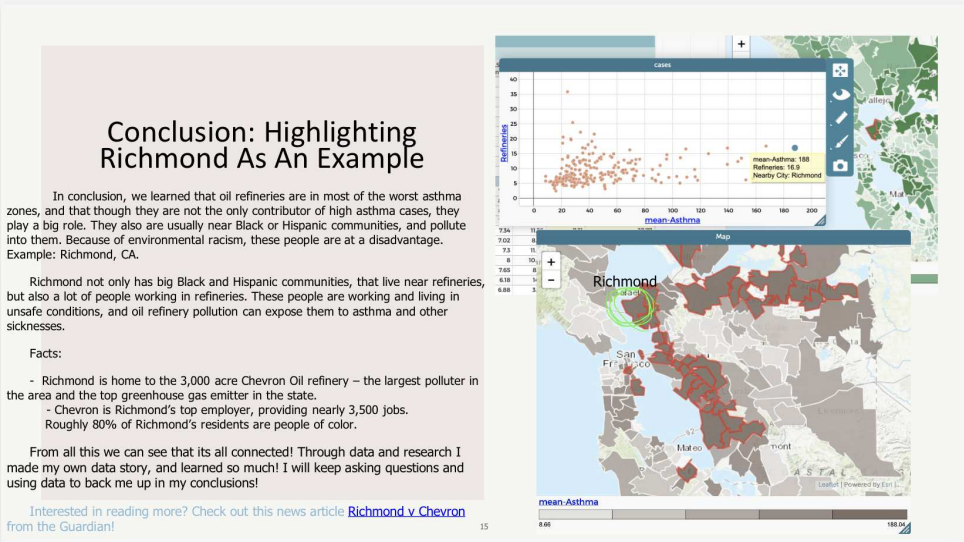
In the first data story (see Figure 1), Student A chose to explore the impacts of oil refineries and environmental racism on rates of asthma in the San Francisco Bay Area. In regards to storytelling *about* data, the student made clear that she used the dataset provided during the workshop, which included information about asthma rates (A-SOU; A-SEL). She shared that examining this dataset led her to formulate her research question focused on the location of oil refineries and the rates of asthma. Since oil refineries were not an existing part of the dataset, she sought this information and described how she added it herself (A-MOV; A-SEL). The student explained that she has seen oil refineries where she lives and has heard her parents discuss the negative impacts of oil spills or refinery leaks. In combination with her previous knowledge and seeing the Asthma dataset the student expressed

interest in wanting to learn more (A-MOT). The student considered the impacts of environmental racism and the societal disadvantages for Black and Brown communities that majorly comprise these areas (A-HIS). In this data story there is no evidence of reliability or validity considerations (A-RV), or alternate hypotheses (A-ALT).

Regarding storytelling *with* data, Student A used two maps of the San Francisco Bay Area, one showing where high rates of asthma exist and the other showing where specific oil refineries are located, to support the claim that higher rates of asthma exist in areas with oil refineries (W-CAS). To showcase these patterns to the audience, the student digitally highlighted and drew on the maps (W-AGG; W-SUP). In addition, the student included and digitally annotated several maps throughout their presentation which added to their evidence in support of their claim. The student concluded their presentation with key takeaways and demonstrated an interest in wanting to continue to learn more (W-CON). In this data story there is no evidence of revision (W-REV) or rejection (W-REJ) of their argument, or of inference (W-HYP).

Figure 1

Concluding Slide of the First Data Story



Student B

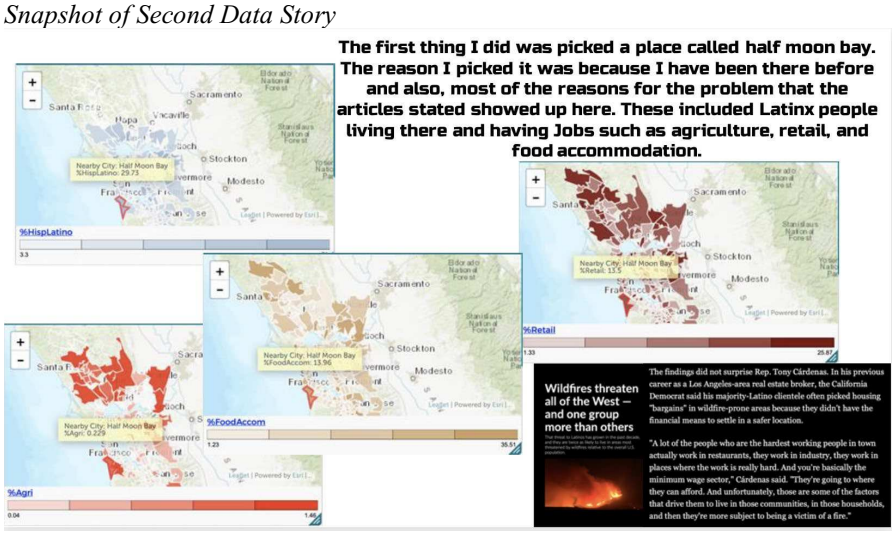
In the second data story (see Figure 2) Student B chose to explore if wildfires in the San Francisco Bay Area impacted one ethnic population, Latinx, more than others. As for storytelling *about* data, the student started by describing their motivations for selecting the research topic (A-MOT). They began their investigation by focusing on Half Moon Bay, a city that they were familiar with and that they were able to connect to historical and residential patterns they had researched (A-SEL; A-HIS). Unlike the first data story, there was no discussion of the data source (A-SOU), how data were transformed or represented (A-MOV). Like the first story, there was no discussion of the reliability or validity of the data (A-RV), or alternate explanations or hypotheses (A-ALT).

Regarding storytelling *with* data, the student explained how they modified their argument over time, as they realized the data they had initially selected did not support their initial claim (W-REV, W-REJ). They thought Half Moon Bay would have experienced many fires, because this location had a High Latinx population and jobs in agriculture, retail and food accommodation, which meant a higher likelihood for wildfire as described by a news article they read. This student shared their experience with hypothesis rejection and revision (W-HYP). The student then decided to explore a broader region of the Bay Area and realized that a factor that they hadn't considered, proximity to water, could contribute to the lower rates of fire in Half Moon Bay. The student used data to show these trends, inland vs. coastal (W-AGG) and support their claim (W-SUP). In this data story there is also evidence for showing claims through data (W-CAS), but there is no evidence for conclusions (W-CON).

Cross-case comparison

Student A emphasized their story *about* data: emphasizing the process of sourcing, transforming, and visualizing their dataset as an entry into their investigation. Student B instead emphasized their personal motivations and familiarity with a specific local case, Half Moon Bay, as an entry. However, Student B described a detailed process of revision, reconsideration, and new discovery *with* data after an initially surprising finding. Further analysis will explore how elements of data storytelling may impact students' learning and development of agency.

Figure 2
Snapshot of Second Data Story



Conclusions

Thus far, our findings suggest a large variety of types of data storytelling, each which are likely to support different sets of relationships, skills, and competencies with data. Consistent with Radinsky's (2020) modes of narrative, we see students moving between storytelling *about* data including how it is constructed, transformed, and connected to context (Modes 1 & 3), and storytelling *with* data to highlight its statistical and social implications (Modes 2 & 4). However, we also observe more subtle yet important variations to these modes. For example, we observe some students focus on telling the story of (meta)data itself – its reliability, validity, appropriateness for a question or context – while others take data for granted and instead focus on implications for action. As this work continues, we hope to identify concentrations of foci that can help validate and expand an understanding of data narrative modes and their respective opportunities for learning.

References

Amato, A., Matuk, C., DesPortes, K., Silander, M., Tes, M., Vacca, R., Woods, P. J. (2022). Postcards and photo walks: Telling community data stories through photography. In *Proc. 16th Intl. Conf. Lrn Sci,* pp. 1493-1496.

[GAISE II] 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 (GAISE) report II. American Statistical Association and National Council of Teachers of Mathematics.

Hall, R., Wright, K., & Wieckert, K. (2007). Interactive and historical processes of distributing statistical concepts through work organization. *Mind, Culture, and Activity*, *14*(1-2), 103-127.

Matuk, C., Amato, A., Davidesco, I., Rubel, L., Stornaiuolo, A., Bumbacher, E., ... & Woods, P. J. (2022). Data Storytelling in the Classroom. In *Proc. 16th Intl. Conf. Lrn Sci,* pp. 1779-1786.

Pfannkuch, M., Regan, M., Wild, C., & Horton, N. J. (2010). Telling data stories: Essential dialogues for comparative reasoning. *Journal of Statistics Education, 18*(1).

Rubin, A. (2020). Learning to Reason with Data: How Did We Get Here and What Do We Know? *Journal of the Learning Sciences*, *29*(1), 154-164.

Sanei, H., Khan, J. B., Yalcinkaya, R., Jiang, S., & Wang, C. (2023). Examining how students code with socioscientific data to tell stories about climate change. *Journal of Science Education and Technology*

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics, 16*(6), 1139-1148.

Storniauolo, A. (2020) Authoring data stories in a media makerspace: Adolescents developing critical data literacies. *Journal of the Learning Sciences, 29*(1), 81-103.

Radinsky, J. (2020). Mobilities of data narratives. *Cognition and Instruction, 38*(3), 374-406.

Acknowledgments

This work is funded by the National Science Foundation IIS-1900606. Any findings or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.