International Journal of Computer-Supported Collaborative Learning
https://doi.org/10.1007/s11412-020-09327-1

Data wrangling practices and collaborative interactions with aggregated data

Shiyan Jiang[1] • Jennifer Kahn[2]

Received: 18 February 2020 / Accepted: 19 August 2020

Abstract
Data visualization technologies are powerful tools for telling evidence-based narratives about oneself and the world. This paper contributes to the literature on data science education by examining the sociotechnical practices of data wrangling—strategies for selecting and managing large, aggregated datasets to produce a model and story. We examined the learning opportunities related to data wrangling practices by investigating youth's talk-in-interaction while assembling models and stories about family migration using interactive data visualization tools and large socioeconomic datasets. We first identified ten sociotechnical practices that characterize youth's interaction with tools and collaboration in data wrangling. We then suggest four categories of activities to describe patterns of learning related to the practices, including addressing missing data, understanding data aggregation, exploring social or historical events that constitute the formation of data patterns, and varying data visual encoding for storytelling. These practices and activities are important to understand for supporting future data science education opportunities that facilitate learning and discussion about scientific and socioeconomic issues. This study also sheds light on how the family migration modeling context positions the youth as having agency and authority over the data and contributes to the design of CSCL environments that tackle the challenges of data wrangling.

Keywords Data wrangling · Modeling · Storytelling · Family migration · Data visualization · Sociotechnical practices

✉ Shiyan Jiang
sjiang24@ncsu.edu

[1] Department of Teacher Education and Learning Sciences, North Carolina State University, Poe Hall, 208, 2310 Stinson Dr, Raleigh, NC 27695, USA

[2] University of Miami, Coral Gables, FL, USA

Introduction

As the interdisciplinary field of data science has grown, large-scale datasets and interactive data visualization tools have become increasingly open and accessible, creating opportunities for learning. Likewise, data science education is a growing new area of CSCL, in which youth interact with others, digital technologies, and complex datasets to support inquiry in multiple disciplines, such as science, mathematics, and social studies (Lee and Wilkerson 2018). Data science tools provide novel ways of exploring traditional disciplinary content and developing statistical and data literacy across formal and informal settings. While education research has focused on facilitating the learning of statistical reasoning (e.g., Aridor and Ben-Zvi 2018; Hancock et al. 1992; Konold et al. 2015; Moore 1990), inference (e.g., Makar et al. 2011; Makar and Rubin 2018), and modeling (Lehrer and English 2018), more research is needed to understand how to support the development of skilled, data-savvy citizens to ethically tackle society's biggest problems with larger, more complex datasets and interactive technologies (Venturini et al. 2015; Boyd and Crawford 2012).

The new energy around data science education reflects a *datafication* process in society in which data increasingly influences individual and public decision-making and interactions across all sectors of life (Pangrazio and Sefton-Green 2020). This includes the growing role of data and data visualization tools in the public news media to present arguments or stories about important scientific and social issues (e.g., https://www.nytimes.com/section/upshot). Data and data visualization tools are especially powerful for enriching narratives about the social and scientific world (Segel and Heer 2010) and providing evidence to challenge or confront misinformation (Dalton et al. 2016). While modeling and storytelling with data is, and arguably has been for some time, a key professional activity for STEM professionals (Kosara and Mackinlay 2013), assembling models and narratives is now also a means for participating in civic discourse for all individuals (Philip et al. 2013). The field also calls for innovative ways of understanding storytelling, such as viewing storytelling as practices of social data scientists who wrangled data and created models to explain social changes (Engel 2017). The role of data and visualizations in public conversations can be attributed to these technologies' capacities to support critical inquiry.

How exactly these tools can be used to support critical inquiry into social and scientific issues, particularly among youth, is still under study. One key finding around learning from the data science education literature thus far is that collaborative interactions around data and visualizations provide students with opportunities to negotiate meaning making of data. In one study, Wilkerson and Laina (2018) found that students shared their own experiences in places when they are reasoning about geo-referenced data and context in small groups. Participants in this study brought their experiences into the negotiation of issues in sampling and trustworthiness of data. The authors called for more work in exploring how instructors could serve as facilitators to identify and take advantage of the learning opportunities emerging from the process of storytelling with data. In another example, Radinsky et al. (2019) designed small-group activities to help students learn historical reasoning with data visualization tools. They showed that students co-constructed the meaning of data visualizations and argued that co-construction might contribute to in-depth learning of complex visual data and historical reasoning beyond data. These studies emphasized the need of understanding the process of meaning-making when students interact with others.

Opportunities for youth to engage in data science activities that support analyses of social and scientific issues from local and global perspectives are needed. However, few studies have closely investigated how youth engage with larger, multivariable data interfaces at the level of discursive and embodied interaction, and especially the role of instructors in these learning environments. In this paper, we examine the complex, sociotechnical practices of *data wrangling* (introduced in Kahn 2020; Jiang and Kahn 2019), the strategies for selecting and managing datasets, for youth assembling models and stories about family migration with public data using interactive visualization tools. We consider data wrangling as necessary for storytelling and modeling with data. An examination of the interactions among youth modelers, instructors, and data visualization tools is needed for understanding learning and engagement in data wrangling.

This paper presents a detailed, descriptive analysis of middle and high school youth engaging in storytelling activity with multivariable data visualization tools and large-scale datasets in a summer program at an urban public library. We examine their data wrangling work in a unique and meaningful context: Youth explored their personal *family geobiography* (Kahn 2020), or family migration history, in relation to socioeconomic data trends indicating the push and pull factors of migration. Our analysis of youth's data wrangling interactions addresses the following research questions:

1. What are the *practices* that describe participants' data wrangling interactions with data visualization tools?
2. How can collaborative interactions support learning for youth engaging in data wrangling practices?

Theoretical framework

This study follows an interactionist perspective (Greeno 1994) that views learning as occurring through participation situated in activity (Bandura 1986; Greeno and Engeström 2014; Lave and Wenger 1991). This perspective treats learning as a process comprised of both changes in action and knowledge development (Lave 1996). For our analysis of data wrangling, the theory of situated learning provides a framework for understanding how youth engage with datasets and data visualization technologies as well as with their peers within the social context of our instructional design. Likewise, we focused on participant interaction, including talk, gesture, and tool use, to understand participant knowledge development in data wrangling activities.

The theory of situated learning described is also a "practice-based theory of knowledge and action" (Goodwin 1994) in which learning is "a process of changing understanding in practice" (Lave 1996, p. 6). Practice theory traditionally has been used in social studies of science and technology in the description and analysis of multi-participant representational activity (e.g., Goldstein and Hall 2007; Goodwin and Goodwin 1996; Latour 1999; Noss and Hoyles 1996; Stevens and Hall 1998). Practice theory is also a good fit for the study of Computer-Supported Collaborative Learning (CSCL) environments, in which the goal is to understand how individuals, typically working together with digital interfaces, develop new understandings and knowledge over the course of their participation in the activity. Thus, practices involve engagement with people, tools, and content for the pursuit of knowledge development. In these practices, artifacts (in this study, family migration stories) are generated and evaluated via distinctive ways of talking, acting, reading, and composing. Drawing on a practice theory perspective, we view data wrangling as a representational, sociotechnical (Tuominen et al. 2005) activity that involves coordination of bodies and tools (Hall and Nemirovsky 2012). Accordingly, we are interested in understanding the representational practices associated with wrangling large datasets that make data wrangling a rich space for CSCL.

As in these studies of STEM representational practices, we find practice theory to be useful for understanding how data wrangling exemplifies the human effort and messiness that comprise scientific activity (Latour 1999). Our analysis of data wrangling seeks to unpack the particular kinds of responses that participants engage in to manage uncertainty in scientific activity (Star 1985). This includes the negotiation, interpretation, and explanation of model assumptions (Goldstein and Hall 2007) among co-participants in data wrangling activity.

Our study also builds on a collection of data science education studies that have found that interactive, multivariable data visualization tools support learning with data across a wide range of disciplines (i.e., STEM, social studies), primarily in secondary settings (Lee and Wilkerson 2018; Rubel et al. 2017). Collectively, these studies have directed the discussion of data visualization tools and datasets towards understanding how these technologies can enrich disciplinary learning and how youth learn to engage in data inquiry practices within disciplinary settings. For example, Radinsky et al. (2014) found that middle school and college social studies students were able to make observations, inferences, and explain data limitations as part of their inquiry project exploring national historical migration trends in the social studies classroom. Students used a version of Social Explorer, one of the georeferenced data tools used in this study. In another example, Krumhansl et al. (2013) found that working with complex, real scientific datasets was compelling for high school students' scientific inquiry and learning despite challenges with navigating large oceanic databases and multiple data representational tools.

This growing body of literature collectively calls for more opportunities for youth to develop critical data literacy by using visualization tools to ask/answer social and scientific questions (Engel 2017; Philip et al. 2016; Philip et al. 2013) and make inferences about data in personally and culturally meaningful ways (Börner 2019; Rubel et al. 2017; Wilkerson and Polman 2020). For example, Polman and Hope (2014) found that high school students' creation of scientific data infographics facilitated meaningful connections between their home or community experiences and their disciplinary learning. In another example, Rubel et al. (2017), in their study of urban youth exploring alternative financial institutions in their communities, found that digital GIS maps facilitated opportunities to critically reason about the relationship between variables affecting spatial justice. Furthermore, Radinsky (2020) illustrated diverse patterns of narrating data when students presented migration trends and demonstrated that students learned to reason about historical events and social justice. We extend this research by investigating the kind of personal learning prompted by our instructional design, the exploration of the family geobiography in relation to broader socioeconomic data trends. Our study situated data modeling and storytelling activities in a personally meaningful context and elicited complex thinking about the relationships between individual experiences and society-level issues and trends. Additionally, in line with the study design, we viewed learners' identities and histories to be central in the study of learning processes and considered present, past, and future relations among social actors and data technologies in our analysis of learning.

With the expansion of data visualization technologies, data science education research has focused increasing attention to understanding possibilities for collaboration and multiparty interactions in data science activity, such as between students and peers, instructors, and family members. For instance, Ingulfsen et al. (2018) argued that it was a challenge for teachers to facilitate productive collaborative reasoning about data. In another study, Lee and Dubovi (2020) found that teachers and parents could serve as external resources to help students with Type 1 diabetes make their own healthcare decisions based on diabetes-related data. Roberts and Lyons (2017) developed a coding framework to characterize learning in conversations among museum goers using interactive data visualizations in an exhibit. A focus of the current study is the microanalysis of students' interactions with both data visualization technologies and others, which will contribute to the design of effective peer and teacher support in data wrangling.

Moreover, the current analysis seeks to deepen the field's knowledge of learners' understanding of "the relationship between data, context, and uncertainty" (Wilkerson and Laina 2018, p. 1224). To do this, we examine the complex, interdisciplinary practices involved in working with multivariable data visualization tools and unrestricted databases, where there are hundreds of datasets and variables to choose from. Our analysis contributes to the design of CSCL environments that tackle the challenges of data wrangling, including how to support students in addressing missing data (similar to what Radinsky et al. 2014 called data limitations), understanding data aggregation and related statistical concepts and measures (a struggle for students in Rubel et al. 2016), learning about historical events and context behind trends (Boyd and Crawford 2012), and visually encoding data to tell a compelling story (Cairo 2019).

Methods

Context and participants

We report on a single iteration of a design-based research program (Cobb et al. 2003) that explored how young adults learn to tell stories and build models about social and scientific issues with aggregated data. Previous iterations found that making personal connections to large-scale phenomena represented by large, aggregated data could be productive for generating critical perspectives (Kahn and Hall 2016). In turn, this design iteration sought to explore the benefit of embedding storytelling and modeling with data in a personal context and to better understand the role of complex data interfaces in relating personal experiences to the larger phenomena being represented by the data. Family migration was chosen as a topic because the sharing of family histories benefits youth development (Fivush et al. 2011), and migration continues to be a pressing issue globally, nationally, and across communities. We hoped that this study would offer insight into how to facilitate learning and productive dialogue around such a timely matter.

In the study, middle and high school youth ($N = 17$; self-identified as 6 male; 11 female; 13 African-American; 3 White; 1 Asian; Mean hours of attendance = 13; sample included 6 sibling pairs) created what we called *family data storylines* to explore reasons for personal family mobility as well as national and global migration in a free summer workshop at a city public library. The library staff recruited youth through personal connections, community partners, and public listservs. Participants were between ages 10–16, mostly identifying as multigenerational African-American.

The decision to design for learning with data visualizations at the local public library was intentional. For many youth, libraries have become central nodes in their interest-driven learning ecologies (Barron 2006; Barron et al. 2014). The city library had been a longstanding partner of the research team. Also, the library held archival materials important to the research design, including an oral history collection from immigrant residents, which is publicly available in the Library of Congress through the national organization, StoryCorps®, and was one of the inspirations for the study. As part of the project, youth contributed oral histories to the city's archives.

The rationale for the project was to understand how youth engage with data visualization technologies in our designed media learning environment. The primary learning goal for the instructional design was for youth to learn how to tell stories with data that connected personal (my family) and aggregate (families like mine) experiences. As part of this activity, we hoped that youth would learn to manipulate and interpret geospatial datasets in order to build data visualizations that could be aligned with their stories and would learn something new about their family or community histories.

Youth represented family decision-making and social conditions with online modeling and mapping tools and related the lives of their family members and ancestors to their own experiences and futures. The three weekly workshop sessions (2 days per week, 5 h per day) culminated in a public community exhibit of youth's family data storylines. The research and instructional team, who facilitated the program implementation, was comprised of a graduate researcher, a senior university faculty member, two graduate research assistants who were both Masters students in education, and four library staff.

Assembling family data storylines involved the following tasks: First, participants chose a side of the family to focus on. Second, participants chose one of two interactive web-based data tools accessed via their laptop's Internet browser: Social Explorer (figures in Excerpts 3 and 5) or Gapminder (figures in Excerpts 1, 2, and 4). Both tools were selected because they afford (Gibson 1986) interactivity, as opposed to static data displays, and provide access to a wide range (hundreds) of social and scientific datasets. Social Explorer is a historical thematic mapping tool that uses US demographic data. It accesses hundreds of variables from Census and other demographic datasets that go as far back as 1790. Data can be encoded as multiple visualization types—as dots, with colorful shading, with bubble size—and at national, regional, and local scales. In Social Explorer, users can also create different side-by-side temporal and spatial comparisons. Gapminder is a multivariable data tool for graphing public global socioeconomic data. It has five possible quantities or variables that can be selected. Y-axis, x-axis, color, and bubble size each have over 500 health and wealth indicators or measures to choose from. Timescales of datasets vary: Some start as early as 1800, and others only have one year of available data. In Gapminder, users can also alternate between logarithmic and linear scales, and select particular country bubbles to leave data trails over time.

Participants selected variables from each tool's available datasets, like education level or household income in the US Census. Participants captured screenshots of models or maps and inserted them into a Microsoft PowerPoint, accompanied by slides or texts that explained data selections and what participants learned from the data.

Data collection and analysis

We video and audio recorded all activities and recorded participants' work on laptop computers with screen capture software. Our qualitative analysis took an interactionist approach (Greeno 1994; Azevedo and Mann 2018) towards engagements between individual agents and the learning environment. This approach, with a focus on both bodily interactions with the physical environment and tools as well as conceptual practices and perceptions of participants, supports the study of enactive and multimodal, embodied learning.

We used interaction analysis methods (Jordan and Henderson 1995) to understand participants' data wrangling strategies. In particular, we paid close attention to multimodal talk-in-interaction, that is, the sequence of turns of talk (Schegloff 1997), tool use, gesture, and body movement that was publicly visible in the video record. The interaction analysis helped to clarify the contexts in which students made sense of data and to show data wrangling practices related to student learning in their talk-in-interaction. To address the first research question about data wrangling practices, we primarily focused on participants' interactions with tools. For our second research question, our interaction analyses focused on episodes that revealed collaboration among participants and/or instructors as well as learning opportunities around critical inquiry with the data.

Specifically, we content logged 90 h of video and developed analytic memos around what appeared to be—through continuous or constant comparisons (Glaser 1965) of individual participants—conceptual and technical practices to describe participants' data wrangling. We considered discourse, gesture, and tool use together in order to understand participants' practices. As an example, in the practice of pursuing data surprises (described in the findings section), one participant was talking with emphasis ("NO") while watching a Gapminder animation in which the life expectancy for nations dropped dramatically. The comparison of participants also focused on critical reflections on the data (i.e., data quality, data stakeholders), social history, and family migration. We reviewed and compared participant records until no new data wrangling practices emerge (Strauss and Corbin 1998).

After identifying data wrangling practices (RQ1), we flagged episodes of data wrangling within the analytics memos for microanalysis using interaction analysis methods. In order to better understand the role of collaborators in data wrangling, we specifically looked for episodes in which participants not only engaged in data wrangling practices but also interacted with others, including peers, instructors, and family members. We had weekly meetings to discuss these episodes to understand youth learning around the capacities of data tools or meaning of the data, including broader social events, personal histories, or the connections between them. We then sought to describe students' activities in these data wrangling episodes. The categories of activities were generated through iterative analysis (Strauss and Corbin 1998) of the analytics memos. We reviewed the memos and discussed common learning opportunities and challenges in these activities. We also challenged each other's perspectives on findings and interpretations throughout the analysis. In the presentation of categories of activities, we selected excerpts to show the common patterns of data wrangling and to highlight voices from a range of participants. We often recreated maps or models in Gapminder or Social Explorer to better understand participant data explorations as well.

We recognize that learning also came in moments when youth were working individually; however, we are interested in youth interaction with others in order to inform our understanding of this instructional activity as a CSCL design.

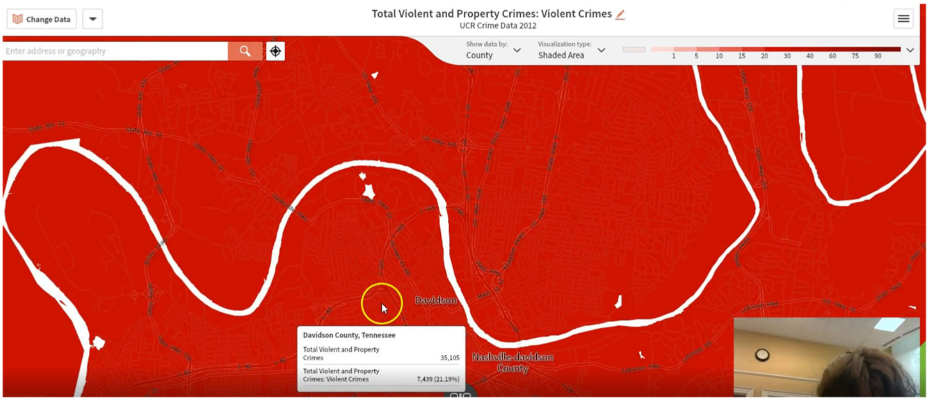
Findings

In the following, we present the data wrangling practices that describe youth's interaction with data visualization technologies and others and highlight data wrangling activities that involved collaborative learning about themselves, their families, data, and society.

> RQ1: What are the practices that describe participants' data wrangling interactions with data visualization tools and others?

We identified ten sociotechnical practices of data wrangling for youth: 1) filtering data; 2) selecting indicators; 3) data visual encoding; 4) interpreting data points; 5) identifying data patterns; 6) pursuing data surprises; 7) reasoning about data relationships; 8) countering data; 9) approximating data; and 10) making data predictions. These practices tended to be in the service of building comparisons to tell a story about both family history and broader socioeconomic trends (e.g., a comparison of origin place A to destination B that shows better economic conditions in B). While the practices occurred at any time in data wrangling, they contributed differently to storytelling with data. For instance, data visual encoding could happen at the beginning of the workshop when participants tried to understand how data was encoded and could also happen at the end of the workshop when they leveraged encoding options for presenting salient differences between original place A and destination B. Below, we introduce and define each practice and provide several brief illustrations.

In filtering data, participants either selected a country bubble and all others faded in Gapminder, or located specific locations in a map in Social Explorer using the zoom feature (Fig. 1). This practice entails highlighting specific data points or locations through a change of data visualization without changing the dataset being used to build data visualization. Making such selections was challenging in an open data interface. Personal connections tended to drive participant selections from the beginning, such as when youth selected a country in which their family members lived or zoomed into their own neighborhoods on a map. These personal selections helped participants become familiar with the data interface.


Total Violent and Property Crimes: Violent Crimes
UCR Crime Data 2012

Show data by: County
Visualization type: Shaded Area

Davidson County, Tennessee

Total Violent and Property Crimes	35,105
Total Violent and Property Crimes: Violent Crimes	7,439 (21.19%)

Fig. 1 A participant is viewing the violent crime rate of her neighborhood in Social Explorer

Selecting indicators or variables involves participants' changing indicators by following specific constraints. This practice changed the dataset in use to build the data visualization. While the constraints guiding indicator selections varied among participants, the most common constraints included a) finding a dataset based on a predeveloped storyline, such as one participant's selection of *Mean Years in School for Men 25 Years and Older* in Gapminder to explore whether her father moved to the U.S for better educational opportunities; b) following a scale, such as one participant's successive selections of *Household Income Less Than $10,000, Less Than $50,000, and Less Than $200,000 in Social Explorer*; and c) identifying oneself or people that participants were familiar with in the data, such as one participant (13 years old, female) selecting *Female Population Aged 10–14 Years* in Gapminder. Students would select and explore indicators representing themselves when developing the narrative. This is consistent with Radinsky (2020)'s finding that participants tended to narrate themselves into data-represented worlds.

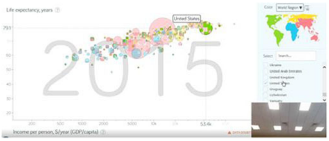
Data visual encoding is the practice of interpreting or understanding the relationship between two representations in which data was mapped into visual structures (e.g., color legends). For instance, in a choropleth map, areas are shaded in proportion to the measurement of variables (e.g., dark colors represent higher population density while light colors represent lower population density). Data visual encoding usually occurred when participants were trying to understand how the data was encoded in different visual representations, such as when one participant said, "Sure are a lot of crime rates" after hovering her mouse over a dark red area in Social Explorer, or when participants changed the details of visual encoding to leverage the data visualization for storytelling purposes. For example, in order to support the claim of moving due to safety concerns, one participant changed visual encoding from "one dot represents the value of 10,000" to "one dot represents the value of 5,000" in Social Explorer to make the differences in the number of crimes between two states more visible. As another example, guided by a researcher–instructor, one participant changed the x- and y-axis scales from linear to logarithmic in Gapminder to explore potentials of each visualization in explaining the family story. Although the participant did not understand the statistical concept of transforming linear to logarithmic scale, she was aware of the change in visualization. Furthermore, as reflected in these examples, a persistent challenge across interfaces for youth was differentiating between data encoding for rates or percentages and counts.

Interpreting data points refers to the practice of interpreting a single data point at one time. This practice was marked by highlighting with the cursor tooltip or selecting (filtering) a particular bubble in Gapminder or discrete geographic area in Social Explorer and by providing verbal explanations of the meaning of the numbers in the tooltip in discourse, as illustrated below by Francine[1] (Excerpt 1). She shared her understanding of the model to her tablemate by focusing on a specific data point. Francine was looking at a Gapminder model of life expectancy (y-axis) and income (x-axis) in 2015 for all nations (this is the default model that is generated when one opens Gapminder).

[1] All participant names are pseudonyms.

Excerpt 1[2]

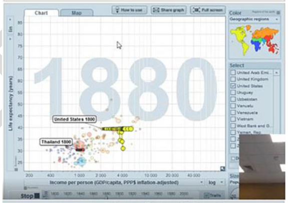
1 Francine (*hovering mouse over US bubble in Gapminder, turning to her tablemate*): See, right here. It's United States. It's about how much money they earn, 53.4, and how long they live, 79.1.



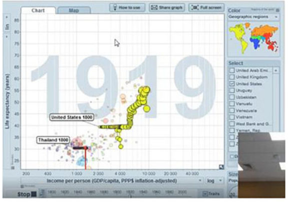
Identifying data patterns represents the practice of describing a data trend, such as the trend of a variable for a single geographic area (i.e., census tract, county, state, nation) over time (e.g., 1920–1970) or the trend of a variable for multiple areas at one point in time (e.g., a regional trend in the year 2000). For instance, in the excerpt below (Excerpt 2), participant Sage, whose mother came from Thailand and father was born in the US, described a data pattern for life expectancy while watching an animation in Gapminder in which she compared the two countries from 1800 to 2015 on life expectancy (y-axis) and income (x-axis).

Excerpt 2

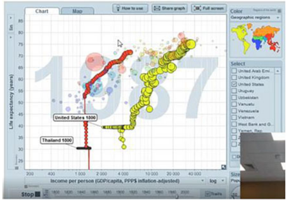
1 Sage (*explaining what she's seeing to a research-instructor*): It's going down a little bit for the United States. (*In the model, the US and Thailand are selected, and time plays from 1800 through 2015.*)



2 Sage (*seeing a decrease in life expectancy in [especially] Thailand and the US in 1919*): Oh, NO!



3 Sage: Life expectancy is going up quite a bit in both countries now.



Pursuing data surprises involves participants' actively looking for or noticing outlier data, typically represented by a divergent bubble trail in Gapminder or a very light or dark region in Social Explorer. For instance, after noticing the drop in Turn 2 of Excerpt 2, Sage went to search online for possible reasons for a dramatic decrease in life expectancy in Thailand in 1919 (was the 1918 influenza pandemic). This example indicates the close relationship between the practices of identifying data patterns and pursuing data surprises.

[2] Transcript conventions: CAPITALS indicate emphasis; (*Observer notes*) indicate significant gesture;? indicates rising intonation;! indicates exclamations; [indicates overlapping talk;, or. indicates pauses less than a half-second … indicates pauses longer than a half-second

Reasoning about data relationships refers to when participants described relationships between two or more indicators or covariation to explain family migration. In most cases, participants had difficulties relating two variables and tended to examine indicators (e.g., household income, educational attainment) independently.

Countering data involves sharing a personal experience or account that did not align with the data. For example, participant Naimah assembled a Social Explorer comparison showing an increase in the Black population in the North and a decrease in Black population in the South between 1920 and 1970. In her explanation of her comparison, Naimah noted that Social Explorer did not include how "moving from South Carolina affected [her] family's connections, and how it was a milestone in [her] history." We view this kind of critique as a challenge to what the Census takers deem as important for understanding social history. This practice shows that participants had agency and authority in questioning data, which challenges the contention that personal experience serves to confirm data for youth (Enyedy and Mukhopadhyay 2007).

Approximating data describes when participants made selections that were approximations for the data they were looking for when there was no data for the exact year, category, indicator, country or place that they desired. For example, for Naimah, survey years 1920 and 1970 roughly corresponded with the birth of her paternal great-grandparents in South Carolina and their decision to leave (1920) and her grandparents' return to the South (Alabama) from Illinois (1970). These approximations could be understood as emergent fixes to manage local uncertainty (Star 1985) in data wrangling.

Making data predictions represents the practice of explaining a prediction about data trends prior to engaging with the interface. Making data predictions was followed by performing corresponding changes in data interfaces to examine the actual data trend and reconcile conflicts between predicted and actual data trends. For instance, Skylar proposed that three data indicators (education, housing, and population) could have explained why her grandmother never moved. In her conversation with Ines (researcher–instructor), she then considered why her choices might not be good ones for demonstrating positive change over time in Canton, MS in her story, "...as it [education] might stay the same as the schools are the same. The houses might be old, but still the same. The population might be less. I know people have left Canton." Skylar's comments indicate that she was making temporal comparisons. After making these three data predictions, she selected the three indicators subsequently to get the actual result in Social Explorer. We only observed this practice in Skylar's data wrangling.

While the frequency of these practices depended on our instructional design and instructors' scaffolding, we believe that they are not unique to our learning contexts. For instance, the practice of data visual encoding occurs when students interpret the relationship between two representations in any type of data visualization. We also believe that we can target specific practices and investigate ways to support these practices that work across learning contexts and data interfaces. For instance, reasoning about data relationships was not a common practice as participants tended to interpret indicators separately. We could support this practice by presenting closely-related indicators (e.g., household income and unemployment) and prompt students to interpret the relationship.

Furthermore, while we describe the above data wrangling practices distinctly, typically, one data wrangling practice went hand-in-hand with others, such as interpreting the meaning of a dark red county in Social Explorer involved both the practices of data visual encoding and interpreting data points. These practices individually might not signal a rich moment for learning, but youth engaged in multiple practices in any given data wrangling interaction in order to solve data wrangling challenges or to investigate parts of their stories or discoveries in the data in the assembly of their family data storylines. Enacting multiple practices likewise appears to be central for youth learning in data wrangling. This will be demonstrated in our data wrangling illustrations subsequently.

> RQ2: How can collaborative interactions support learning for youth engaging in data wrangling practices?

From our interaction analysis, two general thematic categories of activities in which youth engaged in data wrangling practices with others emerged: 1) reasoning about missing data and 2) reasoning about aggregated data trends. We further refined these categories to four salient subcategories of activities—one around missing data and three around aggregated data. Across data wrangling interactions, participants 1) constructed solutions to handle missing data, a common source of trouble for youth with both data tools; 2) connected personal and local experiences to aggregated data; 3) explored data surprises through the investigation of social and historical events; and 4) varied visual encoding to leverage visualizations for storytelling. In these activities, we found that members of the research team in particular, who facilitated the instructional activities, played important roles as guides and co-inquirers with youth. The support from instructors moved participants through and beyond the data wrangling challenges. We illustrate these findings with excerpts from youth's data wrangling interactions with peers and instructors that show how youth learned about the capacities of the data tools or the meaning of the aggregated social data in telling stories about family migration.

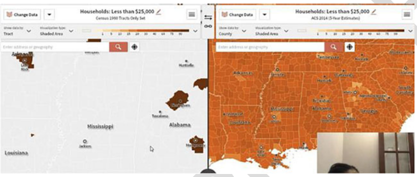
Constructing solutions for missing data

Missing data was a local contingency that sometimes required a complex response. Participants typically approached the issue of missing data as a problem and consequently solicited researchers–instructors for help; researchers–instructors would then sit down with participants and co-investigate a solution. The following episode captures a moment of addressing missing data through co-investigation with a research–instructor on Carter's third day of the workshop. Carter started by exploring the story of her maternal grandparents' migrations from Mississippi to Illinois for better economic opportunities. She then examined her parent's migrations from Illinois to Tennessee for college, which she compared to her own college aspirations. In the subsequent data wrangling episode (Excerpt 3), Carter was working to assemble comparisons of where her grandparents grew up in Mississippi, where her mom grew up in Illinois, and where she grew up in Tennessee, but in 1960, the year approximately when her grandparents moved north.

In the process of determining an appropriate household income dataset from 1960 for her comparison, Carter encountered missing data in Mississippi, which she confused for not having any households within the data criteria. She then wrestled with different income rates due to inflation over 50 years as well as different income breaks between Census survey years. Carter called on the graduate research assistant, Ines, for help, who also struggled to make sense of the data.

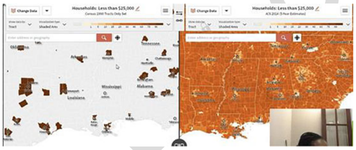
Excerpt 3

1 Carter: Oh, there we go, cool. (*Carter selects Household Income Less than $25,000 for 2014 on the right side of her screen. Data populates by county, for all counties. Household Income Less than $25,000 in 1960 is selected on the left side by census tract, but only a few tracts display data. Both maps are centered on Mississippi, but at different zoom levels. The darkest red on the scale, as on the 1960 side, indicates that over 90% of the population had a household income less than $25,000. The orange color, as on the 2014 side, represents 20–60% of the population.*) OH, my goodness gracious, that's a lot ... Is less than 25,000 a LOT, like now or back then.
2 Ines (*seating next to Carter, offscreen*): Wait, what?
3 Carter: Like in 1960 was 25,000 a lot?
4 Ines: ... for a house?
5 Carter: Hmm-hmm.
6 Ines: Um.
7 Carter: Well a household, (*hovering mouse over Mississippi on the left in 1960*) cause Mississippi didn't have ANY that have less than 25,000.



In Turns 1 and 3, Carter selected household income and asked if $25,000 was considered a lot of money for household income in 1960. In particular, she *pursued data surprises*: There was no data or no population with a household income less than $25,000 in 1960 in the county of interest in Mississippi. Ines, in Turn 4, confused Carter's inquiry about household income for the value of a home, and Carter's repair in Turn 7, "well a household," did not clarify Ines's misunderstanding.

8 Ines: Uh, is that true?
9 Carter: Hmmhmm ... (*Ines leans in to look at Carter's computer screen, takes the mouse*) well they do now.
10 Ines: Well no because this is ... see this is what I don't like.
11 Carter: OH (*pointing to the screen*) because this one's county and that's one's.
12 Ines: EXACTLY so you have to do it by the same.
13 Carter: Okay. (*Carter takes mouse, changes the scale on the right side to show data by census tract.*)


Change Data
Households: Less than $25,000
Census 1960 Tracts Only Set
Tract
Shaded Area
Change Data
Households: Less than $25,000
ACS 2014 (5-Year Estimates)
Tract
Shaded Area
Tennessee
Arkansas
Mississippi
Alabama
Louisiana

In Turn 11, Carter recognized that the two sides of her comparison were showing data at different geographic scales (county vs. tract), and suspected that might be the source of the problem. Ines affirmed this, that she should be comparing data at the same grain size. Carter changed the scale on the right map to the census tract level, but her rectification did not solve her problems. The 1960 display still lacked data, so in Turn 14, Ines asked Carter to look for a different year.

14. Ines: Can you find a different year, like a year that this has like change data see if you can do like what happens if you click 1970 like what just (*Carter goes back to Change Data on the left side, clicks on the survey year to 1970, looks at their data categories, clicks on income*) what are you looking.
15. Carter: (*unintelligible*) The income like changed completely like I don't know see like it went from, is this is households? (*scrolling through Income indicator options left side*)

In Turn 15, Carter found that the household income breaks in the 1970 Census survey were not familiar; they did not match the income breaks for 1960. There was no "Less Than $25,000" dataset in 1970. Then Carter *approximated data* (i.e., changing the survey year to be 1970) to address the issue of missing data for 1960.

16. Ines: No.
17. Carter: No, it's not. I don't know, it like totally changed. What about this one? [YEAH (*clicking on survey year 1980 left side*)]
18. Ines: You're at income. There you're at households. So, you got to go to categories, housing, right? no? (*Carter clicks on survey year 1960 Housing category left side.*)
19. Carter: No cause like I was.
20. Ines: OHH that's per household.
21. Carter: Yeah.
22. Ines: Income at 25,000 dollars.

23. Carter: Hmm-hmm.
24. Ines: Got it.

When Carter went back to the 1960 data categories, Ines realized Carter was not interested in home value data but household wealth. Ines' confusion of household income and housing value was resolved at Turn 20.

In this elaborated excerpt, Carter encountered multiple challenges of data wrangling, including whether light color indicates data are missing or a threshold (in this case, no household income) and how to manage different levels of spatial and social aggregation (data reported by county versus census tract). Carter was trying to understand wealth at the household level and confused sparse data for low household income, which led to further trouble when the research-instructor, Ines, misunderstood what Carter was trying to do until the end of the episode. Ines still offered helpful guidance by encouraging Carter to find comparable and nonsparse data.

Ines and Carter's collaborative process of addressing missing data highlighted that co-inquiry involved back and forth exchanges between the participant and the instructor as they engaged with the data tools. First, Carter noticed the data surprise—the missing data. Then both Carter and Ines shared their initial understandings of what the missing data indicated and reasoned about why it occurred. Afterward, Ines suggested possible ways to address it, which Carter implemented. This trial-and-error approach continued until Carter was satisfied with the solution. Although the multivariable data interface could lead to misunderstanding in inter-subjectivity (as shown in this case), these co-inquiry exchanges might help youth gain in-depth understandings of the meaning behind data.

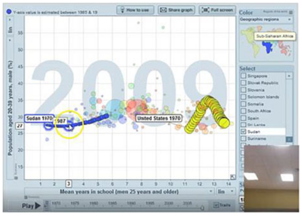
While missing data was common in data exploration, most of our participants, including Carter, successfully found and arranged data to make comparisons relevant for explaining why their family moved or stayed. When students could not initially find data to fit a story of meaning for them, they considered alternative stories and datasets. In Carter's case, when she encountered spotty data again in the workshop—the missing data for certain locations—she was guided by a researcher-instructor to use a near-by county for her comparisons (a data approximation based on commensuration; Kahn 2020).

Connecting personal and local experiences to aggregated data

In their reasoning about aggregated social data, we found that youth were aware that individual differences were neglected in aggregated data patterns by linking personal stories with global trends. Specifically, participants noticed that the aggregated pattern conflicted with their family experiences. The conflict provided unique learning opportunities for participants to understand that individual differences were lost in aggregation, and stories told with the aggregated data could be problematic for explaining their or their family's individual cases. Francine's data wrangling highlights this finding and shows that interpreting data at both personal and aggregate scales could contribute to fruitful discussions about data aggregation (Excerpt 4). To explore whether her father moved from Sudan to the U.S for better education, she selected the model with *Mean School Years for Men 25 Years and Older* (x-axis) and *Population Aged 20–39 Years Old* (y-axis) in Gapminder. She engaged in the practice of *filtering data*, selecting both the US (where her dad moved to) and Sudan (where her dad moved from), while building the model as these two indicators aligned with her father's experience.

Excerpt 4

1 Francine (*taking notes*): You know that in the 1980s, in Sudan, the average school year was around THREE. (*highlighting a bubble representing data for Sudan in 1987*)
2 Sage: Is it real?
3 Francine: Yes, according to this.
4 Sage: WOW.
5 Francine: My dad went to school EVERY year.



In Turn 1, Francine explained her understanding of the visualization through the practice of *interpreting data points*. To address Sage's question in Turn 2, Francine used the data representation as evidence to confirm that people had fewer educational opportunities in Sudan (Turn 3). Sage was surprised about the low number of average school years in Sudan when looking at the representation that Francine showed in Turn 4. Francine countered and challenged the data (*countering data*) with her personal story in Turn 5. Her dad's experience was different from the average described in the data. In this excerpt, the data visualization served as a mediator that supported communicating information clearly and effectively between Sage and Francine.

Similar to Francine's efforts to counter the scenario in which aggregated data patterns did not align with personal experiences in her data wrangling, Skylar also countered the data trend with a personal story. In her family data storyline, Skylar wrote that while her grandma owned a house like those represented in the data map she created, many people in her grandmother's neighborhood could not afford a home:

> I've passed by her [grandma] neighborhood and saw a lot of people in one home or four children packed into one bedroom. My Grandma is lucky to have a home for her and her family just like 64% of people in Canton (text from Skylar's family data storyline slides).

As shown in Skylar's textual narrative, Skylar summarized the aggregated data patterns in her model (i.e., only 64% of people owned a house) and compared it with her grandma's case (i.e., her grandma owned a house). Interpretations of data at different scales (i.e., comparing individual or familial cases to data aggregated at tract, county, state, or national levels) provided unique learning opportunities for students to learn and discuss data aggregation.

Exploring data surprises through the investigation of historical events

Pursuing data surprises drove youth's exploration of social and historical events that could explain changes in data patterns or the formation of certain data patterns. For example, Sage set up a comparison between the US, where her dad is from, and Thailand, where her mom is from, looking at Life Expectancy on the y-axis and Income Per Person on the x-axis in Gapminder. While playing the animation over time, she noticed a dramatic decrease in life expectancy and investigated possible events that might explain the decrease (see Excerpt 2). She described the pattern accurately while watching the animation play from 1800 to 2015 (see Turns 1 to 3 in Excerpt 2). Below, in the continued excerpt, guided by the research assistant Ines, she considered historical events that might explain the trends.

Excerpt 2 Continued

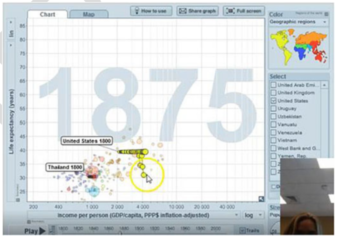
4. Sage: (*The animation ends at the year of 2015*) WOW, the life expectancy in Thailand is 75 years old versus 80 in the United States. In 2015, Thailand is around 75 and US is around 80. I want to see this data. I wonder why it (*referring to the dramatic decrease in life expectancy on the y-axis*) happens like this. Maybe there was a WAR.

In Turn 4, Sage interpreted Thailand and US data points for life expectancy in 2015. She expressed curiosity around an earlier drop in life expectancy in the animation. When Ines approached the table, Sage asked her about the surprise in the animation.

5. Sage: Was it around World War II?
6. Ines: You might want to Google events in the year when the dip appears.

In Turn 5, Sage hypothesized that the dramatic decrease could be caused by World War II. Following Ines' suggestion in Turn 6, Sage searched "Thailand war" and opened a Wikipedia page entitled "List of wars involving Thailand" and went back and forth between Wikipedia and Gapminder to highlight the country bubble at the year when the dip happened.

7 Ines: Try to select "US" again and restart the animation.
8 Sage: I wonder what time would that DIP be.
9 Ines: That is 1884 (*Ines hovers over the lowest bubble on the chart*).
10 Sage (*playing animation*): The life expectancy dropped around 31. (*The model shows that the life expectancy dropped.*)



In Turn 7, Ines changed the direction of exploration from dip in Thailand to the US, when there were two "dips" in the visualization. In Turns 8–10, Sage and Ines found the year when the decrease of life expectancy occurred in the US. Then Sage Googled "United States war" to explore the cause of the "dip" in the US bubble trail and found out that it could have been caused by the American Civil War. This was followed by a conversation between Ines and Sage about the impact of the Civil War on the data. The collaborative inquiry of "dip" pattern continued with her exploration of life expectancy trends in Canada and many more countries. When encountering the same pattern, she would search for information online to look for events that could explain the dip and ask for clarification from peers, instructors, and parents. Sage's data wrangling illustrates how the practice of pursuing data surprises helped participants to move beyond describing data patterns. Participants consistently asked questions when encountering data surprises and engaged in investigating social and historical events (typically through their web browser, such as Google maps and Wikipedia) that could explain changes in data, especially when there was guidance from an instructor.

Varying visual encoding to leverage visualizations for storytelling

Participants varied data visual encoding to highlight data patterns for storytelling, such as showing that a specific population historically faced socioeconomic hardship and presenting the level of safety in certain neighborhoods. Stephen, for example, had a very clear storyline:

His family moved from South Florida to Tennessee for safety concerns. Thus, he focused on choosing a visual representation to support that story.

Stephen opened a map showing total violent and property crimes using bubble map as the visualization type and hovered his mouse over Florida. Then he zoomed into the map to closely look at the bubbles in South Florida and hovered "Palm Beach" (Fig. 2a). Subsequently, he zoomed out the map and changed the visualization type to dot density map. In the dot density map, since there was only one tiny dot over Florida (Fig. 2b), he changed the visualization type back to a bubble map. Stephen changed the visual representation multiple times (i.e., dot density map, bubble map, and shaded area or choropleth map) in order to make a visualization that could illustrate the story that his family moved for safety concerns.

In addition to changing the visual representation, he adjusted the scale legend to make the tiny dot in Fig. 2b more visible in South Florida. By default, the scale legend set one dot to represent a value of 25,000 total violent and property crimes. He changed the scale from 25, 000 to 50, and the map showed smaller dots. Stephen spoke in a low voice to himself, "one dot represents the value of 50" and continued to try different configurations for the scale. He then shared his data wrangling challenge with the research assistant Ines when she came over to his table to check in (see Excerpt 5).

Excerpt 5

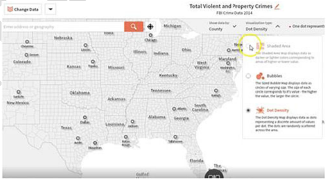
1 Stephen: I am using Social Explorer, and explored safety.

2 Ines: Is there a category called safety in Social Explorer?

3 Stephen: No, it's the total number of violent crimes in the place where my parents came from…I need to make my dot bigger (*referring to the size of dots in the model*), it's so SMALL.

4 Ines: We have a dot map here. More dots show a greater number of crimes.

5 Stephen: I want to have bigger dots.



From Turns 1 to 5, we can see that Stephen had a clear idea about what he intended to show, such as showing that it is not safe in Florida, where his family moved from. That could be the reason why he preferred bigger dots on the map in Turns 3 and 5. Confused by the representation of bubble maps, he misinterpreted that two states would be different when having different sizes of dots in a dot map. In fact, more dots represent more crimes in a dot map while larger bubbles represent more crimes in a bubble map. Still, his data wrangling

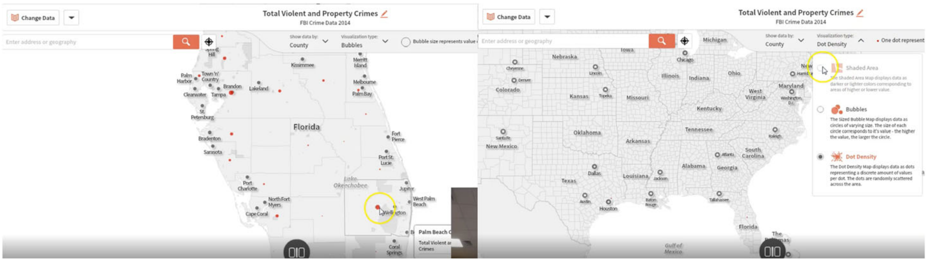


Fig. 2 **a**, **b**. This screenshot captured Stephen's data wrangling. 2a is a bubble map with Stephen hovering a county in South Florida while 2b is a dot map showing only one dot in South Florida. There are few dots in the dot map because the scale legend is set to one dot representing 25,000 total violent and property crimes

showed the process of selecting a visualization that could clearly communicate the message of South Florida being a state with a larger number of violent crimes. In his interaction with Ines, he emphasized that his family moved to Tennessee for safety concerns and intended to show a salient difference between the two places. Ines, instead, drew his attention to understand the encoding of the data representation (Turn 4).

Participants, including Stephen, intended to support their story with effective data visual encoding. They explored a multitude of encoding options and made decisions based on their narrative. Effective visual encoding could be a way to prepare youth to generate and evaluate data-based evidence and explanations for narratives. At the same time, interactions with instructors around data visual encoding appear to be important for students to better understand measures and tell stories with appropriate or fair data visualizations.

These excerpts from youth's data wrangling interactions show that participants often needed guidance to move beyond various challenges that emerged when engaging in data wrangling practices (approximating data, pursing data surprises, countering data, and data visual encoding). The data visualization tools afforded the identification of outliers, and youth were inclined to share their understanding of the outliers with others (e.g., missing data in Carter's case and dip in Sage's case). Peers, and especially instructors, led to in-depth meaning making of these outliers by offering strategies to address missing data and investigate historical or social events that contributed to data trends. While the visualization tools also afforded an examination of data aggregation by showing data trends at different scales and representations, instructional support is needed to guide youth to critically evaluate data visualizations when making encoding decisions for compelling storytelling.

Discussion and implications

Our study illustrates what data wrangling looks like in an informal learning setting. We described how students engaged in various data wrangling practices and collaborative inquiry in the process of assembling models and stories about family migration with interactive data visualization tools. In the following sections, we highlight how engaging in data wrangling practices provides opportunities and challenges for learning about the social meaning of data at both personal and broader scales. We focus on this aspect of learning in data wrangling because more research is needed to support individuals in developing an understanding of the relationship between data, including underlying model assumptions and uncertainty in datasets, and context (Wilkerson and Laina 2018). Based on our microanalysis of youth's interactions with data tools and others, we offer suggestions for future research for data wrangling instructional designs in supporting collaborative inquiry in CSCL settings.

Learning about the social meaning of data

Our analysis of data wrangling practices revealed that these sociotechnical practices can contribute to learning about social meanings of data, which is an essential component of data science education (Boyd and Crawford 2012). For instance, both the practices of *identifying data patterns* and *pursuing data surprises* facilitated students' exploration of social and historical events (e.g., Sage). Presenting youth with visible data trends could help them understand and discuss the socioeconomic context and significance of data patterns. It also shows the possibilities of integrating data wrangling practices into disciplines in social studies, such as history, or broadening STEM inquiry with data to include learning about the larger scientific context. More research is needed to examine opportunities and challenges in adopting these data wrangling practices in different (inter)disciplinary contexts.

In addition, our analysis revealed several conceptual challenges in data wrangling practices that need to be addressed for constructing stories that demonstrate the social meaning of data. One challenge is ensuring the understanding of statistical concepts, such as the difference between counts and rates, as well as the understanding of how to make a reasonable comparison, such as zooming two maps to the same level to set up comparisons. Another challenge is encouraging more data exploration when the current model does not correspond with family stories. In this case, students should be guided to discuss the reason about the lack of alignment, including missing data (Radinsky et al. 2014) and uncertainty in the data (Wilkerson and Laina 2018). Instructors can help participants engage in evaluating their model assumptions, which underlie decisions regarding approximations, and analyze different strategies to find alternative views of the data, or otherwise change the narrative associated with the data. In addition, students should be taught strategies to leverage visualizations for storytelling. For instance, students should be provided with opportunities to discuss how visualizations could be used in appropriate ways or to mislead the public, such as exaggerating the presentation of a real situation (Cairo 2019). Thus, identifying tools, materials, and activities to help students to critically interpret and create visualizations for storytelling with data is a fruitful area for future exploration.

Furthermore, this study addresses tensions of confirming data presented in Enyedy and Mukhopadhyay's (2007) study, in which youth used data to confirm pre-existing conclusions about the social or economic context without questioning the model. Our study demonstrates that the family migration context fostered the youth's authority and agency (Tchounikine 2019) over the data, and this produced both questions about and challenges to the data as well as the exploration of new models and stories. This finding has implications for designing learning activities with data in which youth have their own voices, even in collaborative settings, which was highlighted as an effective and important design principle for data science education by Wilkerson and Polman (2020). This builds on studies that have shown the importance of building a close relation between the self and digital artifacts (e.g., Jiang 2018; Philip et al. 2016). Additionally, while other instructional designs with these tools reduced the number of indicators available for exploration (e.g., Radinsky et al. 2014), we found that it was important to provide a rich and inclusive dataset to offer equitable learning opportunities in data wrangling. This finding suggests that future designs could pursue other personal contexts for data exploration, such as daily mobility datasets and personal healthcare data.

Our findings demonstrate that some practices were more related to tool use (e.g., filtering data) while others were more evidenced in knowledge sharing (e.g., countering data). For instance, filtering data was often about finding oneself in the data (Rubel et al. 2017) through various selections while youth tended to express a personal experience or account that did not align with the data in conversations. This shed light on the design of collaborative data wrangling activities for co-constructing meaning making of data: We should carefully design the learning experiences that can help youth find, express, and (probably) narrate the self when working with others.

The data visualization technologies, the presence of peers, and the design of family migration activity also afforded some specific practices. As an example, filtering data was usually afforded by both features of the technologies (e.g., zooming into neighborhoods where participants lived in Social Explorer) and the design of family migration activity (e.g., instructors encouraged participants to select a couple of places where family members lived). Also, there were scenarios in which participants selected countries (filtering data in Gapminder) that tablemates explored and shared. However, although the data wrangling practices and learning related to the practices were situated in a family narrative construction context in which youth used two interfaces (motion charts and web maps), we expect they would apply to other data wrangling contexts and learning with other data interfaces involving multimodal, multivariable dynamic representations (e.g., stacked graph and radar chart). Future studies can examine whether some practices (e.g., identifying data patterns) are more influenced by the tool affordances (e.g., bubble trail in Gapminder) while some practices (e.g., pursuing data surprises) are more universal across data wrangling activity and interfaces. It would also be worthwhile to investigate how other data interfaces afford learning around the social meaning of data.

Data wrangling as collaborative inquiry

While each youth assembled their own family data storyline on an individual laptop, participants frequently interacted with others, including tablemates, who were often their siblings, and instructors. This interaction often occurred, when youth discovered data that they were personally connected to or when they encountered *trouble* (Davis et al. 2015) in data wrangling. In the following, we highlight the nature of knowledge sharing in data wrangling and discuss future directions for supporting collaborative inquiry.

Our findings indicate that using personal contexts for data science activities could produce fruitful peer interactions. In addition to self-oriented interactions (i.e., sharing inferences or knowledge about their own data models or individual experiences), students should be guided to engage in what Stahl (2013) describes as *transactive*, peer-oriented interactions in which they provide constructive feedback to each other and "build on each other's reasoning through the shared use of computer media" (p. 145). In data wrangling activity, one strategy to support this could be to present students with conflicting results from the same dataset but on different scales. Discussion about the conflicts could lead to in-depth understanding of data aggregation. Future research is needed to investigate how to support reasoning between personal experiences and aggregated data trends in more collaborative learning settings, such as data modeling activities designed for intergenerational families or teams of peers.

The finding that trouble in interacting with data visualizations could lead to more or rich exchanges with others suggests that trouble in interaction is important to understand for supporting productive collaborative inquiry with digital interfaces (Davis et al. 2015). The trouble could come from the complexity of dealing with messy data (i.e., data with various indicators and datasets). Given that there were hundreds of indicators to choose from and some of them represented similar meanings, it is critical to help youth make sense of the differences and similarities between the social meaning of these indicators when constructing narratives. Collaborative inquiry with these indicators might make students' understanding and misinterpretations of these indicators visible to each other. Also, trouble could occur when youth encountered data that did not align with their initial understanding of why their families moved, which could lead to deeper inquiry with the appropriate support or feedback, as noted above. Missing data was another source of trouble in data wrangling, and instructors and participants' collaborative efforts to address missing data involved multiple data wrangling practices. More research is needed to understand the sources of trouble and how or when to support repairing trouble in collaborative data wrangling activity.

Also, this study calls attention to the complexity of instructor support in the context of data wrangling. This finding aligns with results in other CSCL studies (e.g., Ingulfsen et al. 2018), which demonstrate that instructors should carefully address various data-related challenges that students might encounter. In addition, our analysis offers further insights that instructors can be co-inquirers with youth. While instructors provided guidance for participants to resolve data wrangling trouble, youth maintained agency and authority over the narrative, perhaps because of its personal nature. Also, instructors often did not know initially how to solve a data wrangling challenge that a youth encountered. This positioned them as genuine co-inquirers to explore possible solutions alongside them. For example, data approximation was only one way to resolve missing data. Sometimes instructors were confused by the complexity of the interface, datasets, or names of indicators, as in Carter's case, which was only resolved after work to repair intersubjectivity (Schegloff 1997). This finding indicates that when designing collaborative learning environments, researchers or practitioners should carefully define the role of the instructor in data wrangling in order to preserve youth agency in their learning.

Moreover, the field of CSCL needs significant contributions of a variety of researchers who engage in designing effective instructional strategies and technological innovation to support peer interaction and group inquiry around the social meaning of data. In particular, in the case where the inquiry is of a personal nature, this study opens the door to needed conversations about how to leverage that youth tended to share inferences drawn from data related to “me” with others. Our study also illustrates that data wrangling could motivate rich discussions of social context, including decisions connected to one's current and future experiences.

Our suggestions for future research in supporting collaborative inquiry with aggregated data point to the limitations of our study. This study highlights the critical role of collaborators in helping students to assemble meaningful stories and models with aggregated data, but not all learning environments may be able to provide this level of support to students. Furthermore, our study provides insights into engaging students in knowledge sharing through designing personal relevant learning tasks with aggregated data. It is possible that these data wrangling practices might not emerge in contexts that have little connection with participants' lives. Finally, this study fills a gap by delineating data wrangling challenges that students might encounter in data wrangling practice and suggesting that instructors should preserve students' agency over narrative in collaborative inquiry. However, we expect that the challenges might be even more complex when students work in small groups to create data stories, and we encourage the CSCL field to further explore collaboration in data wrangling.

In conclusion, our findings regarding data wrangling practices and learning related to these practices should be the starting point in research on designs for learning and instruction that promote modeling narratives with complex data visualization technologies. Understanding data wrangling practices can provide direction and design guidance to move students towards meaningful interactions with instructors, peers, datasets, and data visualization tools. In particular, we should design data-related activities that students can draw personal connections to and support collaborative inquiry in these activities. We believe that data technologies, and particularly those that aggregate large-scale, complex, or big data, are valuable instructional tools for offering authentic learning opportunities and sharing with others about important social and scientific issues.

Acknowledgements This work was supported by the National Science Foundation under grant number 1341882. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aridor, K., & Ben-Zvi, D. (2018). Statistical modeling to promote students' aggregate reasoning with sample and sampling. *ZDM, 50*(7), 1165–1181.

Azevedo, F. S., & Mann, M. J. (2018). Seeing in the dark: Embodied cognition in amateur astronomy practice. *Journal of the Learning Sciences, 27*(1), 89–136.

Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive view*. Englewood Cliffs: Prentice-Hall.

Barron, B. (2006). Interest and self-sustained learning as catalysts of development: A learning ecology perspective. *Human Development, 49*(4), 193–224.

Barron, B., Gomez, K., Pinkard, N., & Martin, C. K. (2014). *The digital youth network: Cultivating digital media citizenship in urban communities*. Cambridge: MIT Press.

Börner, K. (2019). VIS keynote address: Data visualization literacy. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 1-1). IEEE.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Cairo, A. (2019). *How charts lie: Getting smarter about visual information*. New York: WW Norton & Company.

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951716648346.

Davis, P., Horn, M., Block, F., Phillips, B., Evans, E. M., Diamond, J., & Shen, C. (2015). "Whoa! We're going deep in the trees!": Patterns of collaboration around an interactive information visualization exhibit. *International Journal of Computer-Supported Collaborative Learning, 10*(1), 53–76.

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49.

Enyedy, N., & Mukhopadhyay, S. (2007). They don't show nothing I didn't know: Emergent tensions between culturally relevant pedagogy and mathematics pedagogy. *The Journal of the Learning Sciences, 16*(2), 139–174.

Fivush, R., Bohanek, J. G., & Zaman, W. (2011). Personal and intergenerational narratives in relation to adolescents' well-being. *New Directions for Child and Adolescent Development, 131*, 45–57.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale: Erlbaum (Original work published 1979).

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems, 12*(4), 436–445.

Goldstein, B. E., & Hall, R. (2007). Modeling without end: Conflict across organizational and disciplinary boundaries in habitat conservation planning. In J. Kaput, E. Hamilton, S. Zawojewski, & R. Lesh (Eds.), *Foundations for the future* (pp. 57–76). Mahwah: Erlbaum.

Goodwin, C. (1994). Professional vision. *American Anthropologist, New Series, 96*(3), 606–633 Wiley.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as a situated activity: Formulating planes. In Y. Engeström & D. Middleton (Eds.), *Cognition and communication at work* (pp. 61–95). Cambridge: Cambridge University Press.

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review, 101*, 336–342.

Greeno, J. G., & Engeström, Y. (2014). Learning in activity. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 128–147). London: Cambridge University Press.

Hall, R., & Nemirovsky, R. (2012). Introduction to the special issue: Modalities of body engagement in mathematical activity and learning. *Journal of the Learning Sciences, 21*(2), 207–215.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27*(3), 337–364.

Ingulfsen, L., Furberg, A., & Strømme, T. A. (2018). Students' engagement with real-time graphs in CSCL settings: Scrutinizing the role of teacher support. *International Journal of Computer-Supported Collaborative Learning, 13*(4), 365–390.

Jiang, S. (2018). *STEM+ L: Investigating Adolescents' participation trajectories in a collaborative multimodal composing environment* (Doctoral dissertation, University of Miami).

Jiang, S., & Kahn, J. B. (2019). Data wrangling practices and process in modeling family migration narratives with big data visualization technologies. In *13th International Conference on Computer Supported Collaborative Learning-A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings, CSCL 2019* (pp. 208-215). International Society of the Learning Sciences (ISLS).

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences, 4*(1), 39–103.

Kahn, J. (2020). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences, 29*(1), 57–80.

Kahn, J., & Hall, R. (2016). *Getting personal with big data: Stories with multivariable models about global health and wealth*. Paper presented at the American education research association 2016 annual meeting, Washington D.C.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*(3), 305–325.

Kosara, R., & Mackinlay, J. (2013). Storytelling: The next step for visualization. *Computer, 46*(5), 44–50.

Krumhansl, R., Busey, A., Krumhansl, K., Foster, J., & Peach, C. (2013). Visualizing oceans of data: Educational interface design. In *2013 OCEANS-San Diego* (pp. 1-8). IEEE.

Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge: Harvard University Press.

Lave, J. (1996). Teaching, as learning, in practice. *Mind, Culture, and Activity, 3*(3), 149–164.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lee, V. R., & Dubovi, I. (2020). At home with data: Family engagements with data involved in type 1 diabetes management. *Journal of the Learning Sciences, 29*(1), 11–31.

Lee, V. R., & Wilkerson, M. (2018). Data use by middle and secondary students in the digital age: A status report and future prospects. Commissioned paper for the National Academies of sciences, engineering, and medicine, board on science education, committee on science investigations and engineering Design for Grades 6–12. Washington, D.C.

Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In *International handbook of research in statistics education* (pp. 229–260). Springer, Cham.

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In *International handbook of research in statistics education* (pp. 261–294). Springer, Cham.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1–2), 152–173.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, D.C.: National Academy Press.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers* (Vol. 17). Dordrecht: Kluwer Academic Publishers.

Pangrazio, L., & Sefton-Green, J. (2020). The social utility of 'data literacy'. *Learning, Media and Technology, 45*(2), 208–220.

Philip, T. M., Schuler-Brown, S., & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning, 18*(3), 103–120.

Philip, T. M., Olivares-Pasillas, M. C., & Rocha, J. (2016). Becoming racially literate about data and data-literate about race: Data visualizations in the classroom as a site of racial-ideological micro-contestations. *Cognition and Instruction, 34*(4), 361–388.

Polman, J. L., & Hope, J. M. (2014). Science news stories as boundary objects affecting engagement with science. *Journal of Research in Science Teaching, 51*(3), 315–341.

Radinsky, J. (2020). Mobilities of data narratives. *Cognition and Instruction*, 1–33.

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. *Journal of Social Studies Research, 38*(3), 143–158.

Radinsky, J., Tabak, I., & Moore, M. (2019). Disciplinary task models for designing classroom orchestration: The case of data visualization for historical inquiry. *Proceedings of the 13th international conference of the computer supported collaborative learning (CSCL)*, Lyon, France.

Roberts, J., & Lyons, L. (2017). The value of learning talk: Applying a novel dialogue scoring method to inform interaction design in an open-ended, embodied museum exhibit. *International Journal of Computer-Supported Collaborative Learning, 12*(4), 343–376.

Rubel, L. H., Lim, V. Y., Hall-Wieckert, M., & Sullivan, M. (2016). Teaching mathematics for spatial justice: An investigation of the lottery. *Cognition and Instruction, 34*(1), 1–26.

Rubel, L. H., Hall-Wieckert, M., & Lim, V. Y. (2017). Making space for place: Mapping tools and practices to teach for spatial justice. *Journal of the Learning Sciences, 26*(4), 643–687.

Schegloff, E. A. (1997). Conversation analysis and socially shared cognition. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 150–171). Washington, DC: American Psychological Association.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1139–1148.

Stahl, G. (2013). Transactive discourse in CSCL. *International Journal of Computer-Supported Collaborative Learning, 8*(2), 145–147.

Star, S. L. (1985). Scientific work and uncertainty. *Social Studies of Science, 15*(3), 391–427.

Stevens, R., & Hall, R. (1998). Disciplined perception: Learning to see in technoscience. In M. Lampert & M. L. Blunk (Eds.), *Talking mathematics in school: Studies of teaching and learning* (pp. 107–149). Cambridge: University Press.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research. Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.

Tchounikine, P. (2019). Learners' agency and CSCL technologies: Towards an emancipatory perspective. *International Journal of Computer-Supported Collaborative Learning, 14*(2), 237–250.

Tuominen, K., Savolainen, R., & Talja, S. (2005). Information literacy as a sociotechnical practice. *The Library Quarterly, 75*(3), 329–345.

Venturini, T., Jensen, P., & Latour, B. (2015). Fill in the gap: A new alliance for social and natural sciences. *Journal of Artificial Societies and Social Simulation, 18*(2), 11.

Wilkerson, M. H., & Laina, V. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM, 50*(7), 1223–1235.

Wilkerson, M. H., & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences, 29*(1), 1–10.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.