Data Wrangling Practices and Process in Modeling Family Migration Narratives with Big Data Visualization Technologies

Shiyan Jiang, Carnegie Mellon University, shiyanj@cs.cmu.edu
Jennifer B. Kahn, University of Miami, jkahnthorne@miami.edu

Abstract: Big data technologies are powerful tools for telling evidence-based narratives about oneself and the world. In this paper, we examine the sociotechnical practices of data wrangling—strategies for selecting and managing datasets to produce a model and story in a big data interface—for youth assembling models and stories about family migration using interactive data visualization tools. Through interaction analysis of video data, we identified ten data wrangling practices and developed a conceptual model of the data wrangling process that contains four interrelated recursive stages. These data wrangling practices and the process of data wrangling 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 positioned the youth as having agency and authority over big data.

Keywords: data wrangling, modeling, storytelling, family migration, data science education

Introduction

As the interdisciplinary field of data science has grown, large-scale datasets, also known as big data, and interactive data visualization tools have become increasingly open and accessible, creating opportunities for learning. Assembling models and narratives with big data is an important STEM practice for many professions (Kosara & Mackinlay, 2013) and for participating in civic discourse (Philip, Schuler-Brown, & Way, 2013). The role of big data and visualizations in public conversations can be attributed to these technologies' capacities to support critical inquiry (boyd & Crawford, 2012). These tools are especially powerful for enriching narratives about the social and scientific world and providing evidence to challenge or confront misinformation.

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 big data interfaces at the level of discursive and embodied interaction. In this paper, we examine the sociotechnical practices of data wrangling (introduced in Kahn, under review), the strategies for selecting and managing datasets, for youth assembling models and stories about family migration with public big data using interactive visualization tools. We consider data wrangling as necessary for storytelling and modeling with big data, and an examination of the active interactions between youth modelers and visualization tools is necessary for understanding data wrangling. Our analysis addresses the following research questions: *1) What are the* ***enactive practices*** *that describe participants' data wrangling interactions with big data visualization tools? 2) What is the* ***conceptual process*** *of data wrangling to assemble family data storylines?*

Theoretical framework

First, this research follows an interactionist perspective (Greeno, 1994) that views learning as occurring through participation situated in activity (Bandura, 1986; Greeno & Engestrom, 2014; Lave & Wenger, 1991). Second, we approach family storytelling with big data as a representational, sociotechnical activity that involves coordination of bodies and tools (Hall & Nemirovsky, 2012). Third, learners' capacities, dispositions, and histories are central in our learning environment design and study of learning processes.

Furthermore, our study builds on a collection of data science education studies that have found that interactive, multivariable data visualization tools support learning across STEM and social studies disciplines (e.g., Philip, Olivares-Pasillas, & Rocha, 2016; Polman & Hope, 2014; Radinsky, Hospelhorn, Melendez, Riel, & Washington, 2014; Rubel, Hall-Wieckert, & Lim, 2017) and understandings of statistical concepts (e.g., Engel, 2017; Harrison, Yang, Franconeri, & Chan, 2014). These studies also call for more opportunities for youth to develop critical data literacy by using visualization tools to ask/answer questions and make inferences about data in personally and culturally meaningful ways (Borner, Peppler, Kennedy, Uzzo, & Heimlich, 2017).

Methods

Context and participants

We report on a single iteration of a design-based research program focused on exploring how youth and young adults learn to tell stories and build models about social and scientific issues with big data. Previous iterations found that making personal connections to large-scale phenomena represented by big data could be productive for generating critical perspectives. In turn, this design iteration sought to explore the benefit of embedding storytelling and modeling with big data in a personal context and to better understand the role of big data interfaces in relating personal experiences to a larger scale—the scale of the phenomena being represented by the big data. Conceptually, our design was thus intended to support learning across temporal, spatial and social scales (Hall & Leander, 2010). We chose family migration because the sharing of family histories benefits (Fivush, Bohanek, & Zaman, 2011), and migration continues to be a pressing issue globally, nationally, and in our communities. We hoped that this study would offer insight into how to facilitate learning and productive dialogue around such a timely matter.

In the current 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 *family data storylines* to explore reasons for personal family mobility (What moved my family?) as well as national and global migration (What moves families?) in a free summer workshop at a city public library. Youth represented family decision-making and social conditions with online modeling and mapping tools and related the lives of their ancestors to their own experiences and futures. The three weekly workshop sessions (2 days per week, 5 hours per day) culminated in a public community exhibit.

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 (Figure 1) or Gapminder (Figure 2). Both tools were selected because they afford (Gibson, 1979/1986) interactivity, as opposed to static data displays. 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, dynamic graphing tool that uses 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. Third, 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.

Learning environment design

Our learning environment design aligns with a 4E learning model. First, our design recognizes STEM modeling as distributed and embodied activities (Hall & Nemirovsky, 2012; Hutchins, 1995). We thus included activities beyond interacting with computers. For example, once or twice in each workshop session, the day started with a "four walls" game that set up comparisons between experiences of the youth and that of their ancestors, related to circumstances surrounding family migration (i.e., family size, educational attainment, occupation, neighborhood racial diversity). We asked questions such as: "How far along in school did your family member go?" (Wall 1: Grade school, Wall 2: High school, Wall 3: College and beyond, Wall 4: No formal schooling), which was followed by "How far along do YOU expect to go?" In Sessions 2 and 3, we performed a walking-scale timeline that, like the four walls game, asked participants to stand in for their family in historical time. These embodied comparison activities asked the participants to consider the historical social conditions that motivated or forced their families to move and that possibly could be explored with the data tools.

Second, we consider our design to be an example of extended learning with big data. Learning was distributed across workshop activities and family members. Youth assembled their storylines with siblings who were in the workshop and with parents, via phone calls, text messages, visits during the workshop day, and athome conversations over intervening nights and days, in which pieces of family story were filled in or changed or corrected by family members.

Third, our design embedded the sociotechnical practices of data science in a personal topic (family migration) and in a common, cultural, collaborative activity (family storytelling).

Finally, as described earlier, the focus for the current analysis is on enactive learning, or how the youth participants interacted with designed activities and context for storytelling and modeling with big 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 & 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 & Henderson, 1995) to understand participants' data wrangling strategies. As we reviewed participant trajectories (Jiang, 2018) across activities, we 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 then selected episodes for microanalysis, in which we analyzed gesture, discourse, and tool use as sources of knowledge and learning. We often recreated maps or models in Gapminder or Social Explorer to better understand participant data explorations. 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 categories of data wrangling practices emerge (Strauss & Corbin, 1994). After identifying the categories of practices, we reanalyzed the selected episodes for microanalysis to trace the trajectories of practices and developed themes around the process of data wrangling.

Findings

We define data wrangling as the practices that manage multiple datasets and measures in order to connect narratives with aggregate data. Data wrangling in an open data interface appears to involve the application of technical and statistical knowledge as well as an individual's social values and sense of identity. For instance, many of the participants' interactions with the data tools involved locating places that held personal meaning for them, whether in their family histories or own experiences, or selecting datasets that they identified with. In previous iterations, we have called this "getting personal with big data" (Kahn & Hall, 2016). Consequently, in our instructional context, data wrangling served as the means for participants to establish self–society relations (Kahn, under review).

RQ1: What are the enactive practices that describe participants' data wrangling interactions with big data visualization tools?

We identified ten sociotechnical practices that comprise data wrangling for youth: 1) filtering data; 2) selecting indicators or variables; 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). Below, we introduce and define each practice and provide several brief illustrations.

In **filtering data**, participants either select a country bubble and all others fade in Gapminder or locate specific locations in a map in Social Explorer using the zoom feature (e.g., Figure 1). This practice results in highlighting specific data points or locations through a change of data visualization, but there is no change to the dataset being used to build the data visualization. Making such selections is challenging in an open data interface, when there are hundreds of datasets and variables to choose from. Consequently, 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 neighborhoods on a map.

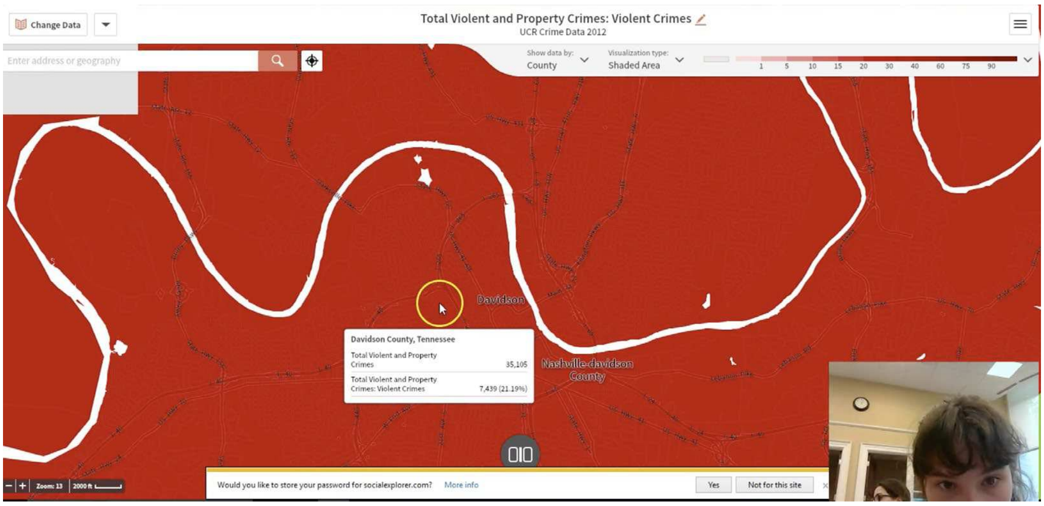


Figure 1. A participant is viewing the violent crime rate of her neighborhood in Social Explorer.

Selecting indicators or variables involves participants' changing default indicators or variables by following specific constraints. This practice changes the dataset being used 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; 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.

Data visual encoding is the practice of interpreting or understanding the relationship between two representations in which data is 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 a) understand how the data is 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 b) change 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, one participant changed the x- and y-axis scales from linear to logarithmic in Gapminder to explore potentials of each visualization in explaining family story. Although the participant did not understand the statistical concept, she was aware of the change in visualization. A persistent challenge across interfaces for youth was differentiating between data encoding for rates or percentages and counts.

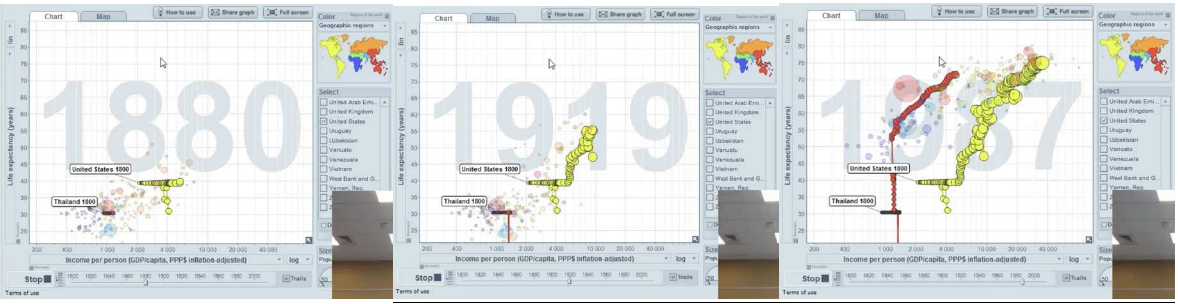
Interpreting data points refers to the practice of interpreting a single data point at one time. This practice was marked by highlighting with the cursor tooltip or selecting (filtering) a particular bubble in Gapminder or discrete geographic area in Social Explorer and by providing verbal explanations of the meaning of the numbers in the tooltip in discourse, as illustrated below.

Excerpt 1 [Week 1, Day 1]
Francine (*hovers mouse over US bubble in Gapminder, turns 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. (1)

Identifying data patterns represents the practice of describing a data trend, such as the trend of a variable for a single geographic area over time (e.g., for a state 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, 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.

Excerpt 2 [Week 1, Day 2]

Sage: It's going down a little bit for the United States (*Figure 2a*).
Sage (*while seeing a drop in life expectancy in Thailand and in the US in 1919 [Figure 2b]*): Oh, no!
Sage: Life expectancy is going up quite a bit in both countries now (*Figure 2c*).



(2a) (2b) (2c)

Figure 2a, 2b, 2c. In Sage's Gapminder model, the US and Thailand are selected, and time plays from 1800 through 2015. Life expectancy (y-axis) and income (x-axis) generally increased over time for the US and Thailand, except for a dramatic drop in life expectancy in both countries in 1919 (2b).

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, while noticing the drop in Figure 2b, Sage went to search online for possible reasons for a dramatic decrease in life expectancy in Thailand in 1919. This example indicates the close relationship between the practices of identifying data patterns and pursuing data surprises.

Reasoning about data relationships refers to when participants describe relationships between two or more variables 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 does not align with the data. For example, Francine explained that her father went to school every year while noticing that the average number of years of schooling for men in Sudan, according to Gapminder, was only 3 years in the 1980s. Another participant, Naimah, who 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, 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 big data, which challenges the contention that personal experience serves to confirm data for youth (Enyedy & Mukhopadhyay, 2007).

Approximating data describes when participants make selections that are approximations for the data they were looking for when there is no data for the exact year, category, indicator, country or place that they desire. 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 can be understood as emergent fixes to manage uncertainty (Star, 1985) or missing data in data wrangling.

Making data predictions represents the practice of imagining what the data could show in the future. In our analysis, this practice did not occur while participants interacted with Gapminder and Social Explorer, although participants considered their own futures in our activities. For instance, in her family data storyline, Hannah compared the percentage of persons 25 years and over with some college experience or more between Chicago and Oakville in 1990 to examine whether her father moved to Oakville for college. She subsequently included a slide in her family storyline describing where she wanted to attend college in the future.

RQ2: What is the conceptual process of data wrangling to assemble family data storylines?

We developed a conceptual model (Figure 3) of the data wrangling process that contains four interrelated recursive stages: 1) *finding my family and me in big data*, 2) *understanding my family and me in relation to the society*, 3) *challenging big data with authority and agency*, and 4) *constructing a big data model to share my family story*. During the stage of *finding my family and me in big data*, participants selected data and indicators related to themselves. They described how historical events affected family migration in the stage of *understanding my family and me in relation to the society*. In the stage of *challenging big data with authority and agency*, participants critically analyzed data when the data was not consistent with their initial understandings of family stories and actively explored alternative indicators that might contribute to family migration. They selected indicators and visualizations that aligned with the family story in the stage of *constructing a big data model to share my family story*. Participants tended to start with finding data related to themselves in this process (*finding my family and me in big data*), which is a "first move" in data wrangling, and ended with constructing models for assembling their family data storylines (*constructing a big data model to share my family story*), which they presented to their peers and were displayed in a public exhibit.

The four stages are recursive, and participants moved back and forth between these stages. Although participants' conceptual movements between these stages indicate a variety of paths in the process of data wrangling, some movements were more challenging for the youth. For instance, Sage engaged in comparing different variables (e.g., income) between Thailand and the U.S in the stage of *finding my family and me in big data*. Her mother moved from Thailand to the U.S for love, which is an abstract concept that she had difficulties in finding variables to represent in the stage of *constructing a big data model to share my family story*. In her case, the movement between these two stages rarely occurred after several failures in finding appropriate variables to represent the abstract concept.

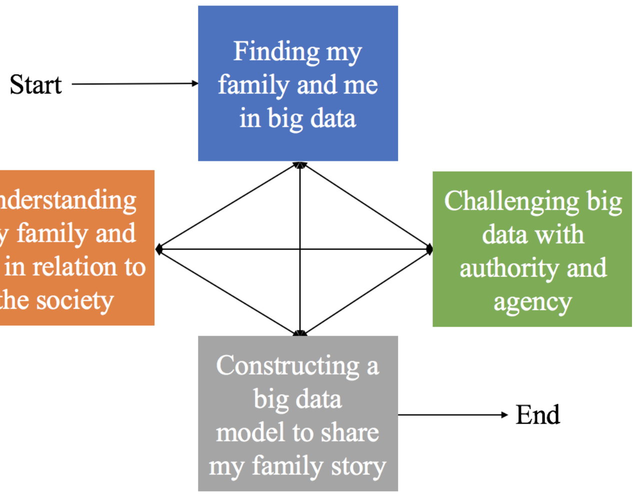


Figure 3. A conceptual model of data wrangling to assemble family data storylines.

Discussion and implication

Our study illustrates what data wrangling looks like in an informal learning setting and how students engaged with various sociotechnical practices in the process. Participants engaged in learning about their families to learning about society through data wrangling. Typically, one practice went hand-in-hand with other practices (e.g., interpreting the meaning of a dark red county involves both the practices of visual data encoding and interpreting data points) in interrelated recursive stages.

The study shows that the "first move" in data wrangling tends to be finding data related to oneself while youth navigate data visualization tools to assemble their family storylines. This finding supports the current understanding of the youth's interactions with big data while contributing new insights: Whereas research in data science and STEM education has evidenced the value and challenges of culturally relevant pedagogy (e.g., Enyedy & Mukhopadhyay, 2007), this study illuminates that getting personal with big data provides an entry point and access to interactions with big data technologies. 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 is important to provide a rich and inclusive dataset to offer equitable learning opportunities in data wrangling. This finding suggests that future designs could engage youth in exploring big data technologies with intentional scaffolds for getting personal with data.

Also, this study addresses tensions of confirming data presented in Enyedy and Mukhopadhyay's (2007) study, in which youth use data to confirm pre-existing conclusions without questioning the model. Our study demonstrates that family data storylines fostered the youth's authority and agency over big data, and this produced both questions about and challenges to the big data. This finding has implications for designing learning activities with big data in which youth have their own voices. The literature supports this design principle and studies have shown the importance of building a close relation between the self and digital artifacts (e.g., Philip et al., 2016). Future research can pursue other personal contexts for big data exploration, such as daily mobility datasets or personal health data.

In addition, our study presents several challenges that need to be addressed in data wrangling practices in future studies. The first 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. A second challenge is engaging youth in discussions of whether variables in a dataset can be used to measure abstract concepts (e.g., love). A third challenge is encouraging more data exploration when the model does not correspond with family stories. In one such case, data wrangling consequently went unresolved (i.e., no data was included in Isis's final family data storylines).

Although each youth assembled their own family data storyline on an individual laptop, participants frequently interacted with tablemates, who were often their siblings, when they discovered data that they were personally connected to, such as data representing the city where participants lived. Future research is needed to investigate how to support getting personal with data in more collaborative learning settings, such as data modeling activities designed for intergenerational families or teams of peers. In addition, this study shows that youth had more exchanges with each other when they encountered trouble in data wrangling, such as when they encountered data that did not align with their initial understandings of why their families moved. These exchanges, particularly between siblings, entailed negotiations of authority in interpreting data and telling the family history, such as Isis told her older sister Naimah, "I do not need you to tell me about my father's story." Much more needs to be learned about the effect of agency and authority on peer interactions with big data.

Our findings regarding data wrangling practices and process should be the starting point in research on designs for learning and instruction that promote modeling narratives with big data visualization technologies in different contexts. For example, 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. Although these practices were used with two interfaces (motion charts and web maps), we expect they would apply to other data interfaces involving multimodal, multivariable dynamic representations (e.g., stacked graph and radar chart). We believe that big data technologies are valuable instructional tools for offering learning opportunities and communication with others about important social and scientific issues.

Endnotes

(1) All names are pseudonyms.

References

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, NJ: Prentice-Hall.

Borner, K., Peppler, K., Kennedy, B., Uzzo, S., and Heimlich, J. (2017). "Data Visualization Literacy: Research and Tools That Advance Public Understanding of Scientific Data" *NSF - National Science Foundation* Available at https://www.nsf.gov/awardsearch/showAward?AWD_ID=1713567

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.

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*, *2011*(131), 45-57.

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

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

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* (pp. 128–147) (2nd ed.). London: Cambridge University Press.

Hall, R. & Leander, K. (2010, July). Scaling practices of spatial analysis and modeling. In Hall, R. (Symposium Chair). Scaling practices of spatial analysis and modeling. International Conference of the Learning Sciences.

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.

Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking Visualizations of Correlation Using Weber's Law. *IEEE Trans. Vis. Comput. Graph.*, *20*(12), 1943-1952.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, 19(3), 265-288.

Jiang, S. (2018). STEM+ L: Investigating Adolescents' Participation Trajectories in a Collaborative Multimodal Composing Environment. *ProQuest LLC*.

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

Kahn, J. (under review). Learning at the intersection of self and society: The family geobiography as a context for data science education. *Journal of the Learning Sciences Special Issue: Situating Data Science.*

Kahn, J., & Hall, R. (2016, April). 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.. Available in the AERA Online Paper Repository.

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

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

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.

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.

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., 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.

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.

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