At the Intersection of Self and Society: Learning, Storytelling, and Modeling With Big Data

By

Jennifer Beth Kahn

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In

Learning, Teaching, and Diversity

December 16, 2017

Nashville, Tennessee

Approved:

Rogers Hall, PhD

Richard Lehrer, PhD

Thomas Philip, PhD

Andrew Hostetler, PhD

To my parents, for their constant support and love

and

To Hans Rosling, who gave us a new way of seeing the world and talking about our shared, human experiences.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation NSF Grant #1341882 "Building Learning in Urban Extended Spaces" and the Vanderbilt Peabody Dean's Fellowship.

I would like to thank my advisor Rogers Hall and my committee members, Richard Lehrer, Andrew Hostetler, and Thomas Philip for their guidance, insight, and feedback; Andrew Hostetler and Emily Shahan for inviting us into their classrooms; our gracious library partners, in particular Andrea Blackman and Amber Williams, for offering a vibrant space and community for our designs for learning; the many members (alumni and current) of the Space Learning and Mobility research team (Rogers Hall, Andrew Hostetler, Kevin Leander, David Owens, Tyler Holette, Nathan Phillips, Katie Headrick Taylor, Alvin Pearman, Benjamin Shapiro, Lauren Vogelstein) for their collaboration and for providing new ideas along the way; Carol Eid and Haley Perkins for their contributions to research study design and implementation; Jackie Griffin for her constant encouragement; and the faculty, staff, and my fellow graduate students in the Department of Teaching and Learning for their support and companionship. Last, I would also like to thank my husband, Juan Jose Thorne, for his patience, love, and willingness to engage with my work.

TABLE OF CONTENTS

Appendix

List of Tables

Table Page

Chapter 3

Chapter 4

Chapter 5

List of Figures

Figure Page

Chapter 1

Chapter 2

Chapter 3

Chapter 4

CHAPTER I

INTRODUCTION

How we experience and encounter information about the social world is changing. The recent availability of large-scale datasets, also known as “big data,” and digital visualization tools has ushered in new ways of reporting and making social, economic, political, and scientific arguments. From cable television newscasters covering election-cycle poll data to news websites providing analyses and assertions about climate change and the environment, narratives told about society with big data visualizations impact our public conversations. Storytelling and modeling with big data not only constitute a new *cultural activity* (Engeström & Sannino, 2010) but also serve as prevalent forms of participation in civic and public life (Philip, Schuler-Brown, & Way, 2013). In this dissertation, I investigate how both youth and young adults learn to engage in the interdisciplinary representational practices that support becoming modelers, storytellers, and consumers of stories told with big data. Consequently, these practices deepen understandings of the self in relation to society.

There is a substantial history of research on science, technology, engineering, and mathematics (STEM) modeling practices in the learning sciences, the field of education research that approaches learning as theoretical and design sciences. That research has typically focused on young students’ representations of self-collected datasets and supported the growth of children’s understanding of statistical concepts, such as distribution and uncertainty; inducted children into statistical practices, such as constructing and visualizing data; and employed these concepts and practices as conceptual tools for developing understandings of natural systems, such as ecosystems and population growth (e.g., Lehrer & Schauble, 2004, 2017; Lehrer, Schauble, Carpenter, & Penner, 2000; Manz, 2012). This dissertation project departs from that body of work and embraces a more recent effort to understand what the youth (and adult) learning possibilities are within the interdisciplinary field of data science.

At the heart of this new area of research and design is the recent availability of big data and digital visualization and analysis tools. Public access to both socioeconomic and scientific datasets and interactive data visualization tools has grown significantly over the last decade. In particular, the *open data* movement has paved the way for public access to (formerly) proprietary datasets; state governments, nongovernmental organizations, and research institutions alike have made their data publicly available for downloading via the Internet by professional and lay users. Undoubtedly, major obstacles still exist for utility of public big data for nonprofessional citizens, such as unequal Internet access, lack of specialized technical training and support, as well as nonstandardized data formats. However, the rise of open tools and big data has already benefited science and social science research by enabling kinds of modeling that were previously too expensive (boyd & Crawford, 2012; Venturini, Jensen, & Latour, 2015). Likewise, despite the current challenges for lay publics, we are hopeful that as open data technologies continue to develop and become ubiquitous, the opportunities to engage in new forms of learning with big data will grow for everyone.

Technical advances in data science afford data aggregation that facilitates novel ways of drawing comparisons across people and shared human experiences. While the comparative method is not new to social science analytics (Ragin, 1987/2014), the availability of big data and tools enables new kinds of complexity in comparative analyses. Big data and digital visualization tools support comparisons that span temporal, spatial, and social *scales* (see Figure 1). For instance, Gapminder motion charts (Al-Aziz, Christou, & Dinov, 2010) can compare the health and wealth developmental trajectories of WEIRD (Western Educated Industrial Rich Developed) nations over the first half of the 20th century to that of BRIC (Brazil, Russia, India, China) nations during the latter half of the 20tt century (Chapter 3). Alternatively, new GIS tools can compare local experiences, like increasing unemployment rates within a county, to national economic trends over the same time period, or reveal internal variation among census tracts (neighborhoods) that is not otherwise visible in state-level averages (Chapter 4). Access to open big data and modeling tools that permit *scaling*—the conceptual movement between different times, places, and social life—afford new kinds of comparisons of socioeconomic, political, and environmental conditions. As a result, not only can we discover new patterns and changes in trends, but we can also bring new insights to our understanding of the relation between the self and society. Scaling between one's local experiences—past, present, or imagined—to broader social conditions, as represented by the big data, invites a consideration of the socioeconomic and political forces that push and pull on one's circumstances and decisions and, conversely, a reflection on how one might affect the larger social tide.

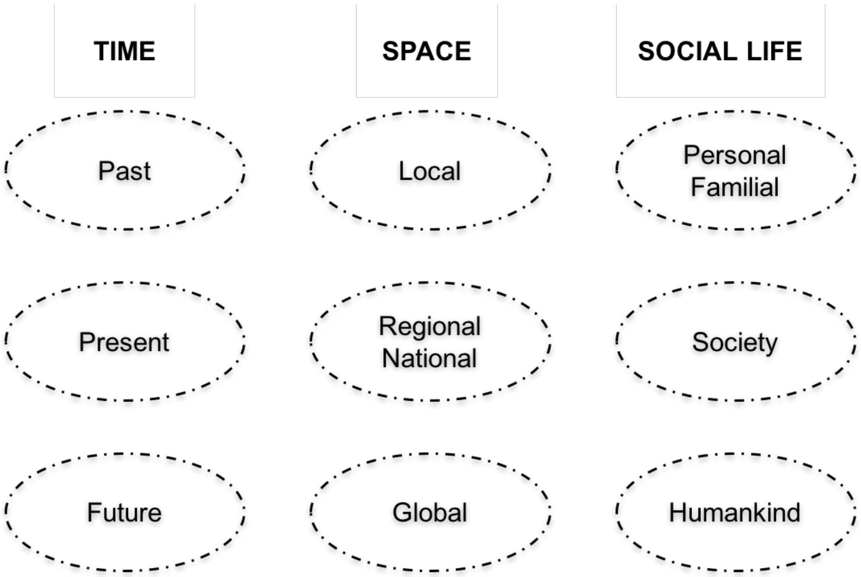


Figure 1. Scaling entails the conceptual movements or shifts across and between different scales of time (past, present, future), space (local, regional/national, global), and organizations of social life (personal/familial, society, and humankind). Big data and novel visualization tools permit comparisons that involve temporal, spatial, and social scaling.

This dissertation presents the case that scaling practices are integral to critical storytelling and modeling with big data. The paired focus on storytelling and modeling follows a rich tradition of recognizing narrative as important to both data visualization (e.g. infographics; Tufte, 2001) and data modeling. Data modeling describes the exploration of the relations between representing and represented worlds (Gravemeijer, 1994; Hall, 2000), and scientists use narratives to describe those relations. For instance, ethnographic social studies of science and technology (STS) have found that scientists engage in compelling narrative performances with models as part of their professional practice (e.g., Goldstein & Hall, 2007; Hall, Wright, & Wieckert, 2007; Ochs, Gonzales, & Jacoby, 1996). In work meetings, scientists and analysts bring stories of field experiences to “ground-truth” data models in order to support or dispute model interpretations (Goldstein & Hall, 2007; Pickles, 1995, 2006).

However, dynamic, digital data visualization tools afford novel ways of seeing, examining, and critiquing trends and changes. In particular, big data tools that permit temporal, spatial, and social scaling support new, powerful forms of telling stories about society (Becker, 2007; Phillips, 2013). Likewise, storytelling with dynamic data visualizations has become a basic component of data science (Busch, 2014; Kosara & Mackinlay, 2013; Segel & Heer, 2010). For example, the proliferation of companies that offer digital statistical platforms for telling stories with data and the growth of data journalism, a data-driven branch of media that incorporates quantitative information and data visualizations into reporting (e.g., theguardian.com/data), reflect the important role of new tools for expressing narratives in data science.

Furthermore, joint-attention to storytelling and modeling with big data revisits and reenergizes the discussion of narrative that has preoccupied learning theorists, psychologists, and sociolinguists for several decades. Narratives are a ubiquitous and conventional means for interpreting, organizing, representing, and constructing knowledge of our own human experiences and a broader social reality (Bruner, 1991). Ochs and Capps (1996) describe the unique, meditational function of narratives as follows: “The power to interface self and society renders narrative a medium of socialization par excellence. Through narrative we come to know what it means to be a human being.” (Ochs & Capps, 1996, p. 31). What Ochs and Capps are referring to is how narratives can *position* speakers and their audiences in relation to themselves, their local communities, and the larger society, in terms of their personal experiences and the ways in which they make sense of the world (Bamberg, 1997). Telling stories is an “interactive activity” that supports the “formation of local identities” (Bamberg, 1997, p. 336); narratives become resources for understanding one’s community membership and relationship to the social context. Narratives, for instance, are the medium for amplifying voices and leveraging the lived experiences of marginalized and underrepresented communities (Milner & Howard, 2013;

Solorzano & Yosso, 2001).

In turn, big data and dynamic modeling tools have the power to animate, enrich, and deepen narratives. Big data and visualization tools support the integration of micro and macro levels of analysis (Cicourel, 1981) and consequently expand the kinds of stories that can be told with data to include descriptions of both local and global social processes. Interactive data visualizations also facilitate new ways of examining model alternatives, which makes narratives told about the world with big data both more challengeable and robust. Moreover, large-scale datasets and data visualization tools change the interactive landscape for storytelling about society and subsequently the stories that can be told about oneself, one's family, or one's community.

What Comes Next

This manuscript reports on a series of design studies that explore learning through storytelling and modeling activities that "get personal" with large-scale data sets. The papers that comprise this dissertation reflect several theoretical commitments: First, learning is culturally, historically situated in activity (Lave & Wenger, 1991; Wertsch, 1998). Our personal and shared cultural and political histories shape how we, as individuals, participate in such activities. Second, participation in data modeling and storytelling activities is distributed and embodied (Hutchins, 1995; Wilson, 2002); it entails the coordination of multiple bodies, tools, and artifacts across contexts. Third, this project takes an *interactionist perspective* (Azevedo & Mann, 2017). According to this approach, close studies and analyses of interaction in activity can reveal participants' *knowledge-in-use*, such as their knowledge of data tools, datasets, statistics, global history, or family history (Hall & Stevens, 2016). The empirical analyses in this dissertation pay close attention to discourse and negotiation, coordinated uses of data technologies, as well as the organization of the body as sources for evidence of local knowledge and learning in context. This approach dates back to early work in the learning sciences that saw the study of cognition and knowledge like an “outdoor psychology” (Geertz, 1983; Hall & Stevens, 2016), in which one can systematically analyze, interpret, and describe knowledge in representational activities (Hutchins, 1995; Lave, 1988). To engage in an outdoor psychology also implies moving outside classrooms and labs into communities to understand and study sociotechnical activity and knowledge construction. Likewise, the current project is a part of a broader youth learning and data science movement that embraces this effort to take an elite activity currently practiced by graduate-level people in finance, machine learning, science, and journalism “outdoors” to youth and families.

Collectively, these theoretical commitments motivated the decisions to conduct design-based research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) and to employ interpretative analytic methods. We (the research team) looked to studies of professionals engaged in representational practices “in the wild” (Hutchins, 1995)—in their professional contexts—to inform our designs of learning environments and experimental teaching. We located, studied, and described the knowledge-in-use in comparative analysis and argumentation activities with datasets and tools among professional (e.g., a world-famous epidemiologist, Hans Rosling, in Chapter 3) and nonprofessional STEM storytellers (e.g., young adults in training to become secondary teachers of math or social studies in Chapter 3 and youth who are telling stories about their families in Chapter 4).

From these investigations, we developed detailed case studies. As Flyvbjerg (2006) writes, the case study methodology offers means for examining learning at all stages of an activity, from beginners or newcomers to experts. Case studies support the “thick” descriptions

(Geertz, 1973) of context-bound local learning and participation in technoscience practices that we desired. Moreover, Flyvbjerg (2006) argues that the “force of example” (p. 288) is underestimated in the pursuit of developing scientific and social-scientific theory. Accordingly, in assembling a vibrant collection of exemplars, this project sought to establish foundations of a grounded theory of understanding oneself in relation to society. Inferences from the comparison of cases advanced into grounded theoretical categories (Corbin & Strauss, 2008; Charmaz, 2008; Charmaz, 2006; Yin, 2000) that describe the conceptual practices involved in storytelling and modeling with big data. Our descriptions of these categories examine the comparative logics that participants employed and the ways in which their comparisons traversed temporal, spatial, and social scales. The hope is that these theoretical categories will “humbly” (Cobb et al., 2003) describe data-science learning processes and, more ambitiously, broaden our understanding of data modeling as an activity for exploring personal and shared human experiences.

Paper 1 (Chapter 2 “Big Data and New Designs for Learning”) frames data science education and technology (DSET) efforts within a literature review that describes the history of modeling with big data as reported in STS research. The paper addresses Philip et al.’s (2013) and Philip, Olivares-Pasillas, & Rocha’s (2016) call for both increasing critical data science literacy among youth and leveraging data science as a means for civic and democratic engagement. The chapter reviews existing empirical research studies that have embraced data science as an elastic field that lives within and across disciplines; such studies described youth and young adults in modeling activities that mostly depart from modeling in the STEM education literature to date. Their study designs leveraged big data technologies to support interdisciplinary inquiry and took steps towards broadening participation in STEM practice as well as bringing data science and issues of culture and equity together. I also describe considerations for designs for learning, including the potential “cost of entry” for youth to participate in modeling and telling stories of comparative analyses using large-scale datasets; I review key conceptual practices as well as persistent conceptual and material challenges for modeling with big data, such as the challenges associated with understanding statistics and data quality. The review includes illustrative examples from my own design research program as well.

Papers 2 and 3 (Chapter 3 “Storytelling With Big Data: Multivariable Modeling of Global Health and Wealth” and Chapter 4 “Getting Personal With Big Data: The Assembly of Family Data Storylines” respectively) offer empirical examples of DSET efforts that support youth and adult use of data science tools and public datasets in understanding social, economic, and political issues. Both papers draw on a corpus of observational and design studies (Cobb et al., 2003) of experimental teaching (diSessa & Cobb, 2004) and dive deeply into the interaction in each setting to understand participants’ modeling activity with big data.

Paper 2 describes the comparative modeling practices in stories concerning global development with an interactive, big data visualization tool, called Gapminder, across professional and student case studies. The paper provides a close analysis of Hans Rosling, a public health statistician and a professional storyteller and modeler with big data, and describes a hybrid third space (Gutierrez, 2008) in which relative newcomers (students) engaged in forms of storytelling and modeling with data across two design studies. Hans’ use of his tool Gapminder in public talks has changed how many people think about statistics and the “developing world.” Gapminder is a dynamic, digital, statistical visualization tool that models multivariable data through time (Al-Aziz, Christou, & Dinov, 2010). It is free to use on Gapminder.org, with open, large, global socioeconomic datasets. In the design studies, the research team asked classes of preservice secondary mathematics and social studies teachers (separately), with Hans’ public performances as a template, to use Gapminder to create models and tell stories about global development that also reflect their own interests and values. In the analysis comparing video records of Hans' public performances and of preservice teachers in each class, we discovered that connecting personal experiences and aggregate trends described in the model can support telling stories about society. In particular, we found that (some) preservice teachers and Hans personalized development in ways that implicated themselves and their peers or audiences. We describe these examples as cases of *counter-modeling*, in which participants used the data model to tell stories that challenged or critiqued dominant or conventional social narratives.

The discovery of rich examples of Hans and preservice teachers counter-modeling by getting personal with big data raised new questions: Can we create DSET learning environments that inspire connections between the self and society? How could we design experimental teaching activities that invite getting personal as a form of counter-modeling? We thus came to see getting personal with big data as an *ontological innovation* (diSessa & Cobb, 2009) in our design studies—a new construct for understanding storytelling and modeling activity with big data. Subsequently, we wanted to further investigate how scaling personal histories to larger social, economic, and historical issues described by big data can support critical storytelling and counter-modeling about society. For this iteration, we chose a context for design that we felt was unambiguously personal (our "design alternative"; diSessa & Cobb, 2009). We conducted workshops in the city public library in which teenage youth were asked to tell stories of their family mobility histories or *geobiographies* ("What moves *my* family?") in the context of exploring broader reasons for family migration nationally and globally ("What moves families?"). Students used online data visualization tools (Socialexplorer.com and Gapminder.org) to assemble *family data storylines* and told stories that became part of the library's public history archives. Family data storylines positioned personal family mobility in relation to national and global demographic and social trends. Likewise, the design of this iteration explicitly framed the activity around a relationship between the individual and the aggregate.

Notably, we found that youth participants built comparisons that were richly complex in scaling, like those of comparative methodologists in the social sciences. Furthermore, family storytelling was a distributed achievement across family members present and non-present in the workshop. Additionally, participants engaged in laborious *data wrangling* in order to bring the family story into alignment with the big data.

Our analysis also considers to what extent our design efforts elicited counter-modeling: While our instructional design invited critical and social-justice perspectives, the final artifacts—the family data storyline that traced familial movement across states or countries and measured varied socioeconomic characteristics—generally lacked critiques of society with regards to race or equity. Nonetheless, our case examples demonstrate that our design for getting personal with big data positioned youth to take a value-based position or moral stance towards socioeconomic change and their family's and community experiences. Participants examined social and economic forces that they never seriously considered before in terms of *"my* history." Participants' comparisons often revealed trends that did not align with broader historical stories or assumptions about family decisions, and this misalignment subsequently generated questions that required new, meaningful conversations with parents and grandparents. In this way, our design, in which being critical was elective but honored, seeded the practices for using quantitative big data and data modeling tools to tell powerful counter-narratives.

In conclusion, this collection of papers reflects the learning sciences field’s commitments to theories of learning, teaching, and design that contribute to an equitable and socially just democracy (Bang & Vossoughi, 2016). Particularly in the “age of ‘alternative facts,’” (Ingraham, 2017), to be able to both “read and write the world” (Gutstein, 2006) with data is increasingly important for examining and challenging power and injustice (Philip, Jurow, Vossoughi, Bang, & Zavala, 2017). Using big data to understand the self and family in relation to society can evoke powerful and important dialogue between local and global perspectives. Getting personal with big data is a framework for storytelling and modeling that can serve as a foundation to support counter-modeling in future design work. Optimistically, one of the utilities of this approach to modeling with big data lies in its potential to ultimately effect social and cultural change.

REFERENCES

Al-Aziz, J., Christou, N. and Dinov, I.D. (2010). SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *Journal of Statistics Education, 18*(3), 1-29.

Azevedo, F. S., & Mann, M. J. (2017). Seeing in the Dark: Embodied Cognition in Amateur Astronomy Practice. *Journal of the Learning Sciences*, (Accepted).

Bamberg, M. G. (1997). Positioning between structure and performance. *Journal of Narrative and Life History*, *7*(1-4), 335-342.

Bang, M., & Vossoughi, S. (2016). Participatory design research and educational justice: Studying learning and relations within social change making.

Becker, H. S. (2007). *Telling about society*. University of Chicago Press.

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.

Bruner, J. (1991). The narrative construction of reality. *Critical inquiry*, *18*(1), 1-21.

Busch, L. (2011). *Standards: Recipes for reality*. MIT Press.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative research.* London: Sage Publications Ltd,

Charmaz, K. (2008). Constructionism and the grounded theory method. *Handbook of constructionist research*, 397-412.

Cicourel, A. V. (1981). Notes on the integration of micro-and macro-levels of analysis. In K. Knorr-Cetina, & A. V. Cicourel (Eds.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies* (pp. 51-80). NY, NY: Routledge.

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

Corbin, J.M., & Strauss, A.L. (2008). *Basics of qualitative re- search: Techniques and procedures for developing grounded theory* (3rd ed.). Los Angeles, CA: Sag

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *The Journal of the Learning Sciences*, *13*(1), 77-103.

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational research review*, *5*(1), 1-24.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219-245.

Geertz, C. (1973). *The interpretation of cultures.* New York: Basic Books, Inc.

Geertz, C. (1983). *Local knowledge: Further essays in interpretive anthropology* (Vol. 5110). New York: Basic Books, Inc.

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, and R. Lesh (Eds.), *Foundations for the future* (pp. 57-76). Erlbaum.

Gutiérrez, K. D. (2008). Developing a sociocritical literacy in the third space. *Reading Research Quarterly*, *43*(2), 148-164.

Gravemeijer, K. E. P. (1994). *Developing realist mathematics education.* Utrecht, The Netherlands: CDBeta Press.

Gutstein, E. (2006). *Reading and writing the world with mathematics: Toward a pedagogy for social justice*. New York, NY: Routledge.

Hall, R. (2000). Work at the interface between representing and represented worlds in middle school mathematics design projects. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 675–680). Mahwah, NJ: Erlbaum.

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

Hutchins, E. (1995). *Cognition in the wild*. MIT press.

Ingraham, C. (February, 2017). Remembering Hans Rosling, the visualization pioneer who made data dance. *The Washington Post*. Retrieved from http://www.washingtonpost.com

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

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge University Press.

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

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal*, *41*(3), 635-679.

Lehrer, R., & Schauble, L. (2017). Children's conceptions of sampling in local ecosystem investigations. Science Education, 1-17. wileyonlinelibrary.com/journal/sce

Lehrer, R., Schauble, L., Carpenter, S., & Penner, D. (2000). The inter-related development of inscriptions and conceptual understanding. *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design*, 325-360.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education*, *96*(6), 1071-1105.

Milner IV, H. R., & Howard, T. C. (2013). Counter-narrative as method: race, policy and research for teacher education. *Race Ethnicity and Education*, *16*(4), 536-561.

Ochs, E., & Capps, L. (1996). Narrating the self. *Annual review of anthropology*, 19-43.

Ochs, E., Gonzales, P., & Jacoby, S. (1996). “When I come down I’m in the domain state”: Grammar and graphic representation in the interpretive activity of physicists. In E. Ochs, E. Schegloff, & S. Thomson (Eds.), *Interaction and grammar* (pp. 328-369). Cambridge: Cambridge University Press.

Philip, T. M., Jurow, A. S., Vossoughi, S., Bang, M., & Zavala, M. (2017). The learning sciences in a new era of US nationalism. *Cognition and Instruction*, *35*(2), 91-102.

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.

Phillips, N. C. (2013). *Investigating adolescents’ interpretations and productions of thematic maps and map argument performances in the media* (Doctoral dissertation). Retrieved from etd.library.vanderbilt.edu.

Pickles, J. (Ed.). (1995). *Ground truth: The social implications of geographic information systems*. Guilford Press.

Pickles, J. (2006). Ground Truth 1995–2005. *Transactions in GIS*, *10*(5), 763-772.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.

Ragin, C. C. (1987/2014). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press.

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

Solórzano, D. G., & Yosso, T. J. (2002). Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative inquiry*, *8*(1), 23-44.

Tufte, E. R. (2001). *The visual display of quantitative information*. *Second edition.* Cheshire, Connecticut: Graphics Press.

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.

Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, *9*(4), 625-636.

Yin, R. K. (2000). Case study evaluations: A decade of progress?. In D.L. Sluffkbeam, G.E Madaus and T. Kellaghan (Eds.), *Evaluation models* (pp. 185-193). Springer Netherlands.

CHAPTER II

BIG DATA AND NEW DESIGNS FOR LEARNING

Introduction

Tarek comes up to the front of classroom and pulls up a map of Syria on the big projection screen from The New York Times website. He shares with the class that "something pretty big" is happening in Syria right now, something that is all over news, which his classmates might not know about: a Civil War that has torn Syria apart for nearly five years. Motioning with an oversized pointer (a repurposed pool cleaning pole) towards the center of the map, Tarek talks about the (at that time, potential) destruction of ancient ruins by the Islamic State (ISIS) in Palmyra. Moving over to the computer, he then opens a second webpage from the New York Times with two black and white images that show light captured by satellites in Syria at night, one from early in the war (2012) with scattered light and one taken two years later that is nearly entirely dark. He explains that the reason for the change in light is many of Syria's residents have been forced to flee their homes, or they live in places that have been engulfed by the war. Tarek pulls up a third map of Syria, this one overlaid with colored shading. Tarek points with the pool pole to different colored areas on the map and explains that each represents the presence of various fighting forces (e.g., pro-government forces, Kurdish forces, ISIS) and the lands they each control. He then scrolls down to an image of a mother and her children huddled together with all their belongings ready to board a bus. He reads to the class that 7.6 million people have been displaced inside Syria and 3.9 million more, half of them children, have been pushed to seek refuge in other countries. Tarek describes the experience of the displaced families in his own words: "Their lives are being changed because of this war, and like they're *having to flee their homes, and like all the things they're familiar with and having to go into something they don't know about."*

Tarek's presentation then shifts. He orients away from the map, plants the pool pole down, faces the class, and states that this war has affected him and his family because his family was born in Syria. He continues, "The war is kind of crashing down on my parents because Syria was their home growing up." He then goes on to say that a few months ago, a bomb was dropped near his grandparents' home. Luckily, they were not there, but it made his mom very upset, because that is where she lived for so long. Most of his family, he thinks, is okay. He tells the class that his mom's sister lives in Damascus, and sometimes the power will go off for hours or days, sometimes they will not have water, and it is not safe to go out on the street because there are "bodies in the street." He says that his dad's family is mostly safe because they live in mountains, and more of the war has been concentrated in cities like Damascus, where his mom's sister and his grandparents live. He tells the class that "there's some pretty bad thing happening there," and it is sad for him because he has been to Damascus. He remembers from his visits that Syria is a very beautiful place, especially Damascus at night, when the entire city's lights glow. He adds that he has not been to Syria for the last five years, so he has not seen many of his cousins or extended family recently. He concludes by reminding his classmates that this war is not only an important news story but also a personal one: "It's really kind of affecting my family as well."

Tarek's re-contextualization of satellite photographs, maps of armed forces, and UN statistics within the story of his personal and his family's experiences are powerful. His performed story spans scales of time and space, from his parents' childhood home to his home today in the US. In turn, his storytelling challenges and collapses the global distances separating *the war and refugee crises in the Middle East from our local, daily routines in the U.S. Indeed, the contrast in his narrative between the lights that used to light up Damascus during his family trips to the current darkness, as captured by satellites, reveals a kind of sense-making that is deeply personal. Through his story, Tarek both assembles and reveals an understanding of his identity in the world and alerts his classmates to a far-reaching, global conflict that is intimately consequential for him and his family.*

The vignette above describes an adolescent learner's initial engagement with maps and models made with large-scale aggregate data and the powerful stories that can be told with them. Were Tarek's exploration of the data used to model the war in Syria extended into a longer project beyond a couple hours in the computer lab, I argue that his meaningful, personal interests could support and strengthen critical approaches to the data and models. Moreover, I view Tarek's personal engagement and self-interest as akin to the passion that field scientists, social science researchers, and technologists bring to data modeling environments, often expressed in the stories they tell about the models, and therefore an important aspect of professional data modeling practice (Goldstein & Hall, 2007). Thus, for Tarek, relating his private, personal experiences to large-scale state-provided data (collected from DOD satellites, government intelligence and public policy institutes, UN agencies) can be understood as a promising form of powerful inquiry around data models describing complicated issues of public interest.

Manuscript Digest

This manuscript describes the emergence of new opportunities for learning with large-scale datasets (LSDS, which I used interchangeably with "big data"), models made with LSDS, and new modeling tools (Figure 1). Reviewing literature mostly from sociology of science and technology (STS) and *data science* fields (the field of using LSDS for social and scientific research and analysis), I first discuss the increasing availability of LSDS and data modeling tools. I describe the kinds of datasets that qualify as LSDS, how they are produced, and their attributes. I consider the relationship between the recent release of state provided and procured data and participatory democracy as well as government accountability. I also identify whom the data stakeholders and users are and provide examples of representational tools and models that use LSDS and are now available to the general public. I describe the potential learning opportunities with these tools and data as new forms for telling about society (Becker, 2007) as well.

I follow this background with a review of what we currently know about mathematical, scientific, and statistical data modeling in the science, technology, engineering, and mathematics (STEM) education literature and from ethnographic and observational studies of scientific and technology professionals, also drawing from STS literature. How are data modeling contexts typically scaled—temporally, spatially, and socially in schools? In what ways (or not) do instructional designs invite personal and critical relationships with the data? How do school data modeling contexts compare with professional science data modeling?

In the last section of the paper, I describe how access to LSDS and modeling technologies permit new designs for learning and opportunities for developing critical perspectives toward, and exploration into, data models and the social, economic, and scientific phenomena they describe. Specifically, I review research in learning sciences that demonstrates how youth facility and inquiry in rich modeling environments with LSDS can support meaningful engagement and critical reflection around local and global social issues, including equity and social justice. In light of this wedge of empirical research, I consider how phenomena described by models and big data become objects of critical (and potentially consequential; see Hall & Jurow, 2015) inquiry for learners when they connect their personal histories and experiences to the aggregate data. Within this final part of the review, I discuss risks and implications for *getting personal* and critical with models created using LSDS and conclude with an example of experimental teaching from a design study that engaged youth in personal and critical inquiry with data modeling practices using LSDS. The objective of this paper is to push forward the education field's understanding of how practices of storytelling, modeling, and inquiry with LSDS can open doors towards engaged citizenship and agency for social change.

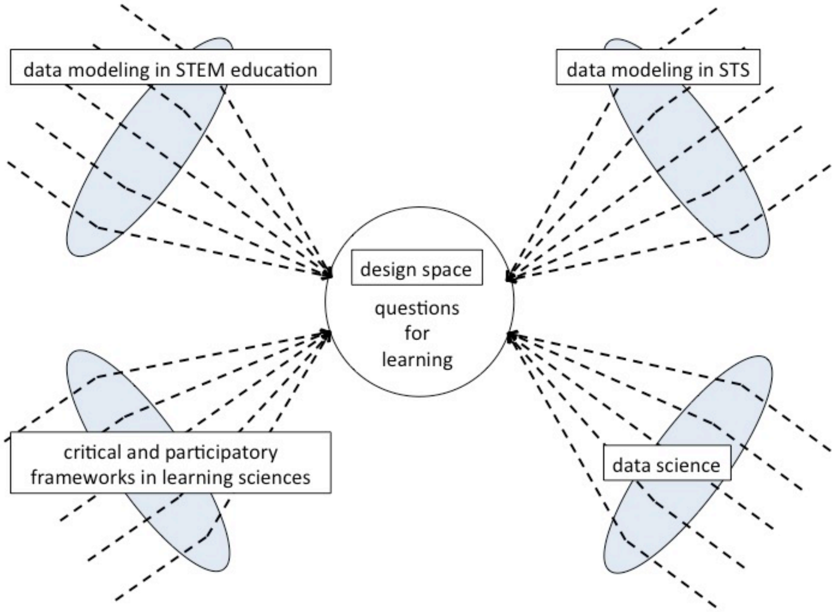

data modeling in STEM education
data modeling in STS
design space
questions
for
learning
critical and participatory
frameworks in learning sciences
data science

Figure 1. Each literature functions as a focusing lens for designs for learning. Findings and concepts from each literature reviewed focus on a design space for asking and answering central questions about how to design learning opportunities with LSDS. Questions include: What is the nature of modeling practices with big data? How does one "get personal" with big data? What kind of critical perspectives are needed in modeling big data? What are the challenges for reasoning about a population as compared to using inferential methods for reasoning about population parameters from samples?

Ubiquity and Public Access to Large-Scale Data and Modeling Tools

In the following section, I review literature published in primarily data science and STS fields in order to describe the recent growth and availability of LSDS and several distinctive attributes and qualities of LSDS, including common applications and uses. I contextualize increased public access to LSDS within a broader social movement that endorses open sharing of technology and software and the historical infrastructure that enables the production of LSDS and related tools. I also name data stakeholders—who are collecting (or scraping), producing, analyzing, processing, and interpreting the data—and give examples of the models and tools that make LSDS meaningful. This background is important to understand for designs of learning environments that incorporate LSDS.

Rapid Rise of Large-Scale Data and Modeling Tools

"All science is fast becoming what is called data science."

--Dr. Bill Howe, Data Science Fellow at University of Washington's eScience Institute (Venkatraman, 2013)

Large-scale data and data visualization tools have transformed the ways in which we interact with the world (Cukier & Mayer-Schoenberger, 2013; Mayer-Schönberger & Cukier, 2013). LSDS refer to quantitative datasets that are sufficiently large that they require a computer for processing (i.e., data preparation, selection, cleaning), analysis (e.g., data mining, the application of algorithms and statistical techniques for extracting patterns from data; Fayyad, Piatetsky-Shapiro, & Smyth, 1996), and storage (Busch, 2014). These datasets can amount to *petabytes* of digital information, equivalent to 10^{15} bytes of data, or 1,000 terabytes or 1,000,000 gigabytes of digital information. LSDS cover every sector of human and social interaction, including the environment, technology use, education, crime, migration, weather, and economics while spanning spatial (local, regional, national, global) and temporal scales (historical, current, and predictive). The field of data science thus "represents an interdisciplinary point of convergence driven by socio-technical innovation, the opportunities for inquiry facilitated by it and the technical challenges posed by them" (Carrigan, 2014, p. 1).

In general, there are two types of LSDS that describe the social world: data that is gathered by state governments, international agencies, and research institutions to describe social trends or phenomena (Busch, 2014) and user-generated records of individual's experiences, online activity, and consumer transactions, also called *digital traces* (Karanasois et al., 2013; Rogers, 2013; Venturini, Jensen, & Latour, 2015). Increasingly, there are some LSDS that are shared and coveted by both public and private worlds respectively, like big data in health informatics (e.g., tracking steps with a commercial pedometer or fitness band). This paper primarily focuses on the first type of LSDS, which is used in its aggregate form to describe society en masse, like US Census data.

State and local governments, international agencies, research institutions, and businesses all collect and aggregate large-scale data about citizens and consumers and are increasingly using data science methods to inform decision-making (Davidian & Louis, 2012). While quantitative data, to some extent, has always motivated policy and industry decisions (Kraemer, Dickhoven, Tierney, & King, 1987; Porter, 1996), the volume and variety of data that can be aggregated and stored easily today is unprecedented. This "data deluge" (The Economist, 2010) can be attributed to advances in digital technology, Internet (Web 2.0) development, remote "cloud" storage, and innovation in data visualization, mining, and knowledge discovery processes (Fayyad et al., 1996).

Counter to impressions that there exists social and scientific "raw" data that is naturally found, all data is constructed (Dalton & Thatcher, 2014; Strasser, 2012). Rather, various data stakeholders and users (e.g., companies, governments, NGOs, professionals) negotiate LSDS, its classification systems, and standards across contexts (Bowker & Star, 2000; Busch, 2011; Busch, 2014; Latour, 1987). Indeed, the production of LSDS is a complexly distributed process in which the people who collect the data in the field are different than those who analyze it in the laboratory (Strasser, 2012). The labor and social and bureaucratic resources that are mobilized to aggregate the data, the value systems embedded in the layers of negotiated standardization in laboratories, and the data that have been discarded (e.g., errors or anomalies) are all typically invisible in data models (Busch, 2011; Busch, 2014; Latour, 1999; Star, 1983; Strasser, 2012). Unresolved moral questions and the satisficing of classifications that is inherent in the design of all information systems may also be obscured from data analysts and users (Bowker & Star, 2000). In turn, certain aspects of the phenomena under study may be lost during case construction and aggregation (Busch, 2014).

Moreover, the practices of large-scale data collection and aggregation have both epistemological and ontological consequences for data use: These often lengthy, years-long data processes determine what counts as the phenomena under study and what gets “simplified out” through standardization, as well as classification and categorization decisions, which dictate what knowledge is necessary to understand the phenomena (Busch, 2011; Busch, 2014). Following these premises, data do not only describe worldly phenomena but they are also cultural, technological, political, and scholarly phenomena to be understood and used with care and a critical stance themselves (boyd & Crawford, 2012; Dalton & Thatcher, 2015; Johnson, 2014; Wilson, 2015, Winner, 1980).

Open Government Data

While much of LSDS remains proprietary, the development of cyber and digital infrastructure has supported the availability of some aggregate data for public inspection and interpretation. Over the past decade, organizations and government bodies (e.g., World Health Organization, U.S. State Department) have released to the Internet what was formerly considered private data to improve government transparency, efficiency, and effectiveness (Buckland, 2011; Johansson, 2012; Mayer-Schönberger & Cukier, 2013). Consequently, there has been a surge in public access to LSDS (Hammerman, 2009).

These *open government data* initiatives partly comprise an *open data movement*, a larger, ongoing "big data revolution" that is transforming 21st century society (Cukier & Mayer-Schoenberger, 2013; Gurstein, 2011; Mayer-Schönberger & Cukier, 2013). The goal of the open data movement is as follows:

> The overall intention is to make local, regional and national data (and particularly publicly acquired data) available in a form that allows for direct manipulation using software tools as for example, for the purposes of cross tabulation, visualization, mapping and so on. The underlying idea is that public (and other) data, whether collected directly as part of a census collection or indirectly as a secondary output of other activities (crime or accident statistics, for example) should be available in electronic form and accessible via the Web. (Gurstein, 2011, p. 1)

Proponents of the open data movement argue that the value of data—from assessing and improving democratic governance, policy, and accountability to increasing efficiencies in industry and social sectors—can be better leveraged with increased public and private sector access (Donovan, 2012; Robinson, Yu, Zeller, & Felten, 2009; Johnson, 2014). In turn, governments and international NGOs have elected to become primarily the "custodians" of hundreds of datasets of social and economic indicators and have left the private and nonprofit sectors and individual citizens (with the necessary means) to operationalize the data (Mayer-Schönberger & Cukier, 2013, p. 128).

Data professionals Mayer-Schönberger and Cukier (2013) credit President Obama for leading the open data movement nationally and internationally. Unlike presidents and precedents before him, President Obama called for government agencies to release national data from nearly every federal department or agency. He created Data.gov, which showcases national data, statistics, and reports that can be downloaded on topics like expenditures on construction, traffic deaths, climate, and adolescent obesity. Other nations (e.g., Britain's Open Data institute), U.S. states (e.g., https://data.ny.gov/), city governments (e.g., Chicago's Data Portal), and global organizations and not-for-profits (e.g., European Union and World Bank) now similarly pursue "open-data strategies" (Gurstein, 2011; Mayer-Schönberger & Cukier, 2013). The move to release the data makes the products of state-invested time and money in technology infrastructure and data collection available to the public that underwrites the expense. These steps to increase public access to large-scale data broaden the government's previous commitments—such as those stated in Freedom of Information Act—to disclose information and records (Cukier & Mayer-Schoenberger, 2013). However, local municipalities and city governments, like Chicago and New York, have had trouble keeping up with the demand to release departmental data reports, in part because most governments lack legislation to ensure compliance (Wisnieski, 2015).

Mayer-Schönberger and Cukier (2013) also attribute increased public access to data to pressure from professional communities that advocate for the release of government data in order to improve government accountability and transparency (e.g., Code for America and Sunlight Foundation in the US, Open Knowledge Foundation in the UK, Gapminder Foundation in

Sweden; Johansson, 2012). Activist support from professional groups for the continued release of data, open-systems, and tools from governments follows a related, successful social initiative for open-source software and code. Hess (2005) calls the free and open-source software (FOSS) movement a *technology and product-oriented movement* in which private-sector firms align with civil society advocates to pursue a certain kind of technology in order to achieve social change. The FOSS movement grew from university computer scientists' and programmers' desires to improve and refine software programs in the 1980s by sharing code and fixes. These interests developed into a social movement to challenge established hardware companies that were dominating consumer markets with costly proprietary software. Eventually, as the FOSS movement and the availability of its software products grew, the major hardware companies adopted variations of alternative open-source software, and consumers received better (fewer bugs) and cheaper software (Hess, 2005). Similarly, the push and pull for open data among government and commercial sectors is politically, socially, and economically consequential for citizens.

Data and Democratization

Open government data is neither politically neutral nor a guarantor of democratization (Coleman, 2004; Hacklay, 2013; Johnson, 2014; Winner, 1997; Winner 1980). Critical technologists, data scientists, anthropologists, sociologists who study the relationship between society and technology recognize that all technology, including open data and software, is value-laden and socially constructed (Porter, 2012). Indeed, the programmers, hackers, and private sector entrepreneurs that lead the FOSS movement have their own political positioning. Their advocacy for a free, unregulated digital technology market is rooted in cultural liberalism ideas that support protections for freedom of expression and intellectual property and more radical libertarian perspectives (Coleman, 2004; Hacklay, 2013; Johnson, 2014; Winner, 1997). Coleman (2004) notes that the larger Internet-using public subsequently has adopted these liberal and libertarian sentiments and discourse around open data and software. However, the majority of lay citizens (and especially youth) still have only limited access to large open datasets, whether they be comprised of digital traces or state-collected social and demographic data, as will be subsequently discussed.

Significant amounts LSDS remains privately-owned; these data, mostly transaction data (i.e., digital traces), are expensive to produce and store and are considered a highly valuable commodity for private sector Internet and technology companies like Google (Cukier & Mayer-Schoenberger, 2013; Wilson, 2015). As for open and free government data and software, public access to LSDS and opportunities for purposeful data use vary considerably, and unequal access to LSDS can create power asymmetries between governments, private sector companies, professionals and technical elite, and lay citizens (Dalton & Thatcher, 2015). For instance, collection and formatting decisions affects the utility of the data for nonprofessional users (Gurstein, 2011). In order for the data to be used more broadly by professional users and the lay public, data management and structures need to be in place so that data is available in cleaned, standardized or "interoperable" formats (e.g., CSV, HTML; Buckland, 2011; Cukier & Mayer-Schoenberger, 2013).

While public datasets are occasionally accompanied by free or open software and code in formats familiar to the professional computer science community for downloading and visualizing the data (e.g., application programming interfaces or APIs), using these data tools and datasets meaningfully requires a specific skillset and knowledgebase. Consequently, beyond unequal access to the Internet (which is still a problem, Gurstein, 2011[1]), a *data science fluency divide* exists among big data users and potential users. Hacklay (2013), a critical technology theorist of *neogeography*, a field of geography that uses open source and free digital geographic data and software in the creation of maps, argues that much of purportedly open data is only user-friendly for technical elites from select professional communities. He pokes a few holes in neogeography's claims that increased access to data and technology will lead to increased participation in civic and social engagement and social change, particularly for marginalized communities. He suggests the using open data to initiate social change requires becoming professionally and technically skilled in "hacking," coding, or programming; most lay citizens will only be able leverage geographic LSDS in a limited way without the specialized training and advanced levels of education and technical support. And those with a computational background—the technical resources to leverage available datasets for analysis—tend to be adult, white males (Daileda, 2016; National Center for Education Statistics, 2015).

In turn, the open data movement has benefited private upper and middle class individuals—those already empowered—rather than those with lower socioeconomic statuses (Johnson, 2014; Donovan, 2012). The more powerful and privileged social groups typically design data and tools in a process that excludes others and predetermines rules for how others will use the technologies (Klein & Kleinman, 2002; Philip, Schuler-Brown, & Way, 2013). While increased access to data and tools is important and necessary, it is not sufficient for social and political change (Donovan, 2012). As Gurstein (2011) writes, "In the absence of financial resources to interpret the [open] data and then develop advocacy actions based on the data

[1] A UN agency, the International Telecommunications Union, reported in 2016 that 47.1% of the world's population consists of Internet users, although 84% of the world population have access via mobile broadband network. They noted that the 3.9 billion people not using the Internet are disproportionally low income, less educated, female, elderly, and rural.

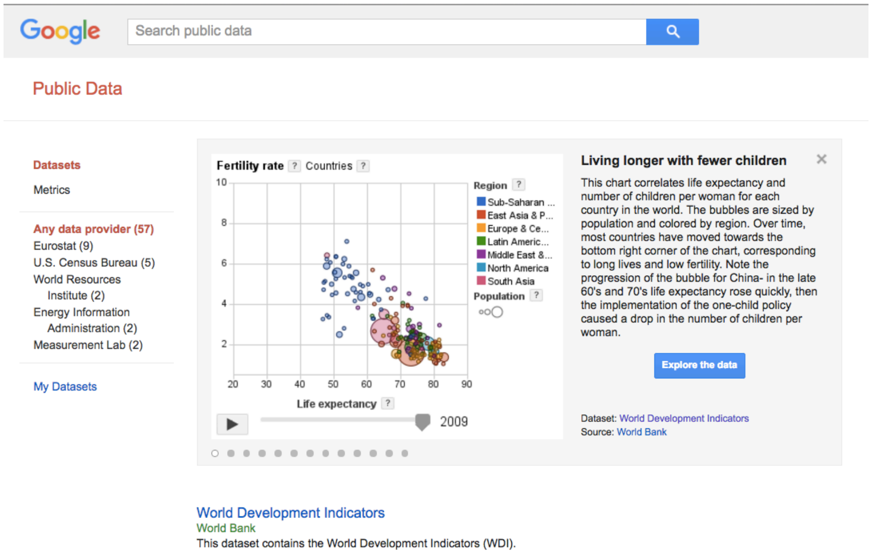
[advocacy and governance], the poor and marginalized would be unable to use their data access in any meaningful way" (p. 4). Indeed, if open data efforts and projects are not deliberatively critical and take no pains to bridge the data science fluency divide, only those with educational and economic resources will be able leverage new open data, in ways that may or may not benefit the greater social good (e.g., using big data to target advertisements of consumer goods to particular demographics to for commercial profit; see ESRI Tapestry, Joiner, 2014).

At the same time, the open release of data presents risks for data privacy and protection of human subjects. The collection of big data in general, both by state governments and by private companies that glean transactional data from digital devices and online activity, can compromise private information that leaves individuals vulnerable to fraud (Johnson, 2014), to infringement on privacy rights (see when big data becomes "big brother" in Cukier & Mayer-Schoenberger, 2013), and exploitation by individuals in positions of greater political and economic power (e.g., Bhoomi case of land possession by wealthy citizens from members in lower socioeconomic castes; see Donovan, 2012; Johnson 2014).

Opportunities for Learning: Modeling With LSDS

Despite cautions over whether open data will expand democracy, the rise of open-source software, open tools, and open large-scale data has benefited science and social science research by enabling new kinds of modeling that were previously too expensive (boyd & Crawford, 2012; Venturini et al., 2015). The scale and distributions of computers, cloud data storage, and web-based statistical and visualization tools (e.g., Tableau, Kosara & Mackinlay, 2013; see Chen & Zhang, 2014 for a full discussion of new tools and methods) have grown to match the massive data sets they compute (Anderson, 2008), giving birth to a broad class of novel modeling tools that permit nearly instant, dynamic, and often interactive visualization to support varied kinds of analyses across disciplines (Al-Aziz, Christou, & Dinov, 2010; Chen, Chiang, & Storey, 2012; Fayyad et al., 1996; Kosara & Mackinlay, 2013, Rosling, Ronnlund, & Rosling, 2005).

For instance, data mining tools assist users with extracting specific information from large data sets, often in science and business contexts (e.g., database marketing to identify populations that adhere to particular consumption and purchasing patterns; Fayyad et al., 1996). Social network analysis tools identify trends in across human and material relations (Latour, 2005; e.g., following online conversations and controversies; Venturini et al., 2015). Time series tools display historical data that can be used for historical or predictive analyses (e.g., motion charts for tracking socio-economic data trends of individual nations over time; Al-Aziz et al., 2010; Johansson, 2012; Kosara & Mackinlay, 2013; Rosling et al., 2005; Figure 2). Spatial analysis and modeling tools create quantitative thematic maps that display layers of data with geographic information (e.g., using GIS to address natural resource management; Duncan, 2006). Wearable electronics and mobile personal devices produce data on daily physical and physiological activity to offer individual health and mobility analyses (e.g., the "Quantified-Self"; Lee, 2013; Nafus, 2016). Graphical and statistical tools support exploratory multivariate analysis (e.g., analysis of how car data varies for acceleration, horsepower, and cylinders; Wongsuphasawat et al., 2016). While many of these tools are privately owned and sold, Internet megaliths (e.g., Google), private companies (e.g., ESRI, Social Explorer), not-for-profit organizations (e.g., Gapminder, DataUSA), and state agencies (e.g., U.S. Census) also host some of these mapping and statistical graphical tools freely online.



Living longer with fewer children

This chart correlates life expectancy and number of children per woman for each country in the world. The bubbles are sized by population and colored by region. Over time, most countries have moved towards the bottom right corner of the chart, corresponding to long lives and low fertility. Note the progression of the bubble for China- in the late 60's and 70's life expectancy rose quickly, then the implementation of the one-child policy caused a drop in the number of children per woman.

Dataset: World Development Indicators
Source: World Bank

World Development Indicators
World Bank
This dataset contains the World Development Indicators (WDI).

Figure 2. Image of Google's online public data directory that offers a motion chart tool. The motion interactive multivariate modeling tool that displays socioeconomic data for select nations overtime.

Modeling describes the exploration of the relations between representing and represented worlds (Gravemeijer, 1994; Hall, 2000). Modeling tools using LSDS serve as instruments for both reasoning about quantitative information and abstractly representing worldly phenomena. Likewise, the models—the inscriptions and representations that such tools and data produce—are subject to judgments of their perceived fit with the world they describe (Giere, 1990; Kuhn, 2012; Manz, 2014). Models made with LSDS are used to ask and answer questions and tell narratives about the world (Duncan, 2006; Kosara & Mackinlay, 2013): Models built with LSDS can be predictive (Giere, 1990), descriptive, historical, explanatory, depth or breadth-oriented (Wongsuphasawat et al., 2016). In the cases of LSDS, the data would be incoherent without effective modeling and visualization tools. Without strong modeling and visualization, "we find ourselves drowning in a sea of data" (Latour, 1999, p. 39).[2]

[2] Busch (2014) and Porter (1996) point out that "distance" (from observations of worldly phenomena described) offers data modelers a more objective perspective. Other critical theorists

In particular, the proliferation of big data and modeling tools has given rise to new representational forms for telling about society (Becker, 2007; Kosara & Mackinlay, 2013; Tufte, 2001). Storytelling and modeling with open, large-scale data has become a new professional, *cultural activity* (Engeström, & Sannino, 2010). Models made with LSDS are ubiquitous in the media, professional fields, and state policy arenas, often found in service of political arguments or narratives aimed at adults and youth alike (Phillips, 2013). However, the openness of publicly published models varies and subsequently so does the depth to which audiences can question, examine, and challenge narratives with a given model. Additional commentary on the importance of *modeling depth* will be provided in the last section of the paper.

With the rise of open data and dynamic data visualizations, the possibilities for storytelling and modeling among ordinary people, including youth, are growing. These possibilities bring new opportunities for learning: Educators and learning scientists can design environments to support participation in storytelling and modeling with big data with these new kinds of tools. Engaging youth and young adults in time series and spatial analysis and modeling practices with quantitative thematic maps and GIS tools has been studied in science and mathematics education (Edelson, Gordin, & Pea, 1999; Enyedy & Mukhopadhyay, 2007; Gordin, Polman, & Pea, 1994; Radinsky, 2008); these studies have found that dynamic data interfaces can promote theory-building and scientific and mathematical sense making. These empirical studies constitute a foundation for future designs for learning in which the capacity of newer digital tools to support comparisons between scales (e.g., local, regional, national) and animate LSDS provokes questions and permits pursuit of those questions. For instance,

of technology (boyd & Crawford, 2012) argue any such claim of big data's objectivity is just pretense.

exploration with the Gapminder, a multivariable modeling tool that displays global socioeconomic data for countries over time (Figure 2), regularly produces dynamic country-bubble trajectories that trigger a "What was that?" response, which, in turn, becomes the starting point for inquiry (Johansson, 2012, p. 195; Kahn, 2014).

Just as power and responsibility go hand in hand, data modelers, users, and analysts, both lay and professional, should proceed with caution with LSDS (boyd & Crawford, 2012; Wilson, 2015). Models made with large-scale data are often statistical, but data are neither populations nor samples as typically defined in statistics textbooks (Busch, 2014). While "N" appears to, and in some case actually does, approach all of a population (e.g., U.S. Census data; Cukier & Mayer-Schoenberger, 2013), the size of LSDS does not mean that the models display or tell universal "Truths." Rather, new modeling tools support model competition by giving public users new access to tools that they can use to try to model alternative conjectures and assemble comparisons. Access to tools that can produce competitive models is important in an age when automated data analysis and Artificial Intelligence (AI) systems generate and publish some of the stories that accompany data models. These stories can be summative (e.g., Narrative Science Inc. uses an automated intelligence system to translate company billing records and investment portfolios into narratives) or oversimplifications that promote stereotypes (e.g., ESRI Tapestry classifies populations in residential areas into discrete consumer profiles to sell to companies; Joiner, 2014).

Finally, social science researchers warn that data science puts qualitative analysis methods and theory development at risk in favor of "dust bowl empiricism," which describes the "arid" accumulation and application of data without a strong grounding in theory that potentially leads to counting spurious patterns as meaningful relationships (Suddaby, 2014, p. 452). This risk of spurious relationships and findings is not new to social sciences research. However, with such large amounts of data, social scientists may mistake the noise for the signal in the data (Silver, 2012) or mistake the form for the substance of social science research: While the data is big, LSDS are not sufficient to answer big questions about society, such as questions about how to solve systemic social problems related to racism or gaps in economic welfare (Uprichard, 2013). Big data can elide and simplify qualities—such as how disparities in income and social services are experienced—that other methods (qualitative methods) can and should help to discover. Furthermore, big data models do not reliably predict how social or scientific systems advance (Derman, 2012; Silver, 2012, Uprichard, 2013). Silver (2012) gives the examples of the failure of financial professionals and terror analysts to predict and prevent the 2008 stock market crash and the 9/11 terrorist attacks, respectively, as evidence. Consequently, our reliance on large-scale data alone could forfeit the kind of theoretical and conceptual grappling that is still necessary to address complex social and social-scientific issues. This is a design problem for learning to engage in inquiry with LSDS, unless we make an effort in our designs for learning to ask complex questions about social and scientific relations that push learners to think about problems and experiences in relation to the data trends.

Summary

The 21st-century rise and spread of large-scale data has transformed data modeling across fields and institutions: in social science, policy and politics, business, and scientific arenas. More of this data than ever before are publicly available, as are software and modeling tools for analysis and data visualization. Peering through the data science lens (see Figure 1), we see both the potential benefits and pitfalls for research, governance, and ultimately designing for learning:

- On the upside, data visualization and modeling tools offer novel, interactive, dynamic ways for answering and asking questions about the world and informing our reports about society.
- Increased transparency of state records, for instance, can also be used to hold governments accountable for discretionary decisions, such as the use of monetary funds, or the unjust enforcement of laws and to put pressure on political actors to bring more resources to underserved populations.
- At the same time, there are significant challenges to using data for social change efforts. Major disparities and obstacles to equal data and tool access, including overcoming a data science fluency divide, and substantial privacy and security threats persist for users.
- Our designs for learning can capitalize on new tools and data as novel means for reasoning and telling stories about the social world. However, we need to still consider and be able to identify when data stories create social realities that benefit some at the cost of others. We must develop an awareness of the ways that big data mask the complexity of social problems and neglect the methods and theories that enrich and deepen the meaning of data. It is a catch-22: Modeling provides simplification that is absolutely necessary to avoid drowning in a sea of data (post collection) or in a sea of phenomena (prior to collection); on the other hand, modeling without awareness of simplification risks making decisions that damage those made "invisible" by data. Without critical awareness of these issues and intentions to address them, data can and will reify unbalanced power relations among stakeholders.

Presently, the discussion among data scientists and critical technology theorists generally lacks attention to the ways in which large-scale data can transform learning and education as well as to ideas about where youth fit in relation to the big data. This will be partly addressed in the next section, in which I will review how large-scale data and critical perspectives in data modeling are currently treated both in education and STS literatures. The risks and challenges of LSDS will be returned to in the context of designs for learning in the last section of the paper.

Data Modeling in School and Professional Contexts

In the following section, I review studies of youth and professionals engaged in data modeling practices, and I describe what such practices entail. I first review design studies (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) of mathematical, scientific and statistical modeling in schools from the learning sciences and STEM education research. I identify important conceptual aspects and practices in (primarily) school-based data modeling environments in relation to opportunities afforded by large-scale data and tools, highlighting exceptions to these trends as necessary, and I address what it means to take a critical stance in these learning environments (i.e., critical stance as in critique of measures and methods or as in political and ideological interpretations). Data modeling in both primary and secondary educational settings primarily describes gathering data as a sample to model a phenomenon in a population. I then review observational and ethnographic studies of data modeling in STEM professional fields in order to characterize important conceptual practices among data-modeling professionals that typically differ from school data modeling environments.

Data Modeling in School Contexts

An important portion of the literature on data modeling in STEM and statistics education research encourages students' statistical reasoning about variation in distributions of measures in order to support inferences about measurement processes and populations. In this work, students often work with classroom-scaled data (i.e., data children themselves collect)—a process I call data modeling from "soup to nuts." I also describe how many studies attempt to cultivate or develop disciplinary engagement (Engle & Conant, 2002) by using fictitious problem contexts intended to resemble students' cultural or out-of-school experiences (Wager, 2012). On the other hand, I describe how students' personal experiences are positioned as bias that interferes with "good," objective data modeling practice. I conclude this section with a comment on how rarely statistical reasoning is situated within critical discussions of broader social, economic, and scientific issues across these data modeling environments.

Data modeling from soup to nuts. Empirical work in the elementary and middle school STEM modeling literature suggests that children's relationship to the phenomena being modeled becomes meaningful for children in part because the construction of data models often stays close to the data collection process. For example, youth participants engage in rounds of hands-on scientific investigation of natural phenomena inside classrooms (e.g., teacher arm-span; Lehrer, Kim, & Schauble, 2007) or nearby outside spaces (e.g., fast plant growth as models of organismic and population growth, measures of ecosystem in a local pond or school backyard; Lehrer & Schauble, 2004; Lehrer, Schauble, Carpenter, & Penner, 2000; Manz, 2012). These studies aim to support students' (a) understanding of data as constructed, not given, often by positioning children as deciders about what and how to measure; (b) understanding of variability (distribution) as emerging from, not inherent in, individual cases; (c) developing conceptual tools for visualizing and summarizing sample distributions; and (d) making informal inferences about processes by comparing sample distributions, either over time, between two processes, or both.

In contrast with the distributed processes of collecting and analyzing LSDS, the sample sizes of data, especially if collected by participants, are typically manageable without serious computer processing. Repeated individual measurements and collective class measurements are subsequently compiled into children-designed data displays that are created digitally (e.g., Tinkerplots software; Lehrer et al., 2007) and by hand (e.g., markers and poster paper; Lehrer et al., 2007; Lehrer & Schauble, 2004). Student-invented notations, inscriptions and statistics simultaneously support mathematical thinking about visualizing and measuring variability, and about sources of variability, such as variation attributed to sampling error or due to natural variation in the population (see Konold & Lehrer, 2008; Lehrer & Kim, 2009; Lehrer, Kim & Jones, 2011). Instruction and designed activities (Konold & Lehrer, 2008; Lehrer et al., 2007; Lehrer & Schauble, 2004;) subsequently encourage students' statistical reasoning about variation in sampling distributions in order to support inferences about processes ranging from measurement and production to the those generating natural variation and populations. These studies focus on developing a sense of sampling variability that arises from chance and then of modeling chance as well as determined effects, as in modeling shifts in variability due to optical illusions, changes in measurement or production methods, or changes in conditions of growth. Other studies (e.g., Bakker, 2004; Garfield & Ben-Zvi, 2007) concentrate on developing students' understanding and facility with descriptive statistics: applying measures of center—namely mean, median, and mode—and informal inference (e.g., Gould, Davis, Patel, & Esftandiari, 2010) to characterize a sample, typically from K-1 onward.

Generally speaking, across the STEM education literature on data modeling, the timescales of the phenomena being modeled are short. Typically, data collection and observations are contained within classroom hours and unit lessons within a single school year. For instance, Lehrer, Schauble et al. (2000) report on two classroom studies, the first with second graders working to determine combinations of Lego cars and "racetracks"" (inclined planes) so that the cars will cross a fixed distance at a desired speed (fast or slow), and the second with third-graders investigating factors that affect the growth rates of Wisconsin “Fast Plants” (i.e., students were asked to describe the conditions lead to faster or more productive growth). In the first study, the phenomena under investigation takes merely seconds, and in the latter case, students were given the opportunity to invent their own inscriptions for modeling daily plant growth (height) across two 40-day cycles.

In Lehrer and Romberg (1996), the phenomena under study had the potential for data analysis over a more extended timescale. The researchers studied fifth-graders who developed survey questionnaires to compare social and scientific aspects of their own lives (their daily routines and habits, likes and dislikes) to that of American colonists. However, while students made qualitative comparisons across hundreds of years with their constructed datasets, comparable historical data was not easily accessible in the context of the study, and the differences were not represented within a data display showing change over large periods of time.

Additionally, in these modeling environments described above, the social production of the data models and measures are confined to the classroom community, its practices, and norms. In turn, the scale of individual agency is similarly bounded to students’ decisions for what measures to use and how measurements are taken. This, of course, makes sense when the purpose of the design is to develop a community of data modelers who learn to makes choices about data including what to measure, how to construct the data (how much, from whom, when), how to organize and use the data (table structures, computational tools), how to represent it, and how to make inferences about the data in light of chance or uncertainty.

In contrast, when starting with state-collected, large-scale open data, the data and measures are already constructed. Data modelers, in labs or in our designed learning environments, generally do not face decisions for how to collect or measure large-scale social, environmental, or economic processes. (In the education world, the Concord Consortium's Common Online Data Analysis Platform [CODAP] project is trying to change this; the CODAP online tool allows public users to visualize their own large-scale data CSV, JSON, or TXT files.) However, modelers with big data are in the position to examine social activities that have much larger scales in time, space, and social life and ask/answer questions about other places and times. Whereas close proximity to the phenomena under study yields disciplinary meaning-making for many of the students in the studies just reviewed, as we will see in the cases of STEM professionals, making modeling with large-scale data meaningfully requires alternative kinds of modeling practices: namely, moving between personally-scaled activity and geographically and historically scale aggregated phenomena (as opposed to the classroom-level aggregate of cases described in the studies above).

Modeling a made-up world. In middle and high school mathematics, science, and statistics learning, the instructional approach has generally been to encourage facility with models at a distance; data sets are no longer collected first hand, but the data are still samples as opposed to census data or datasets reporting for entire populations (e.g., Cobb, McClain, & Gravemeijer, 2003; Lesh, Cramer, Doerr, Post, & Zawojewski, 2003; Rubin, Hammerman, & Konold, 2006; Sandoval & Millwood, 2005). In these environments, learners occasionally still have choices about the units and measures of provided quantities. For example, in the Volleyball Problem described by Lesh et al., 2003, graduate students decide how to measure jumping distances and running times as either ranks (ordinal numbers), as point scores (cardinal numbers), or as directed distances (vectors) and determine how to create a quality-rating measuring athletic performance potential. However, more typically learners are given the datasets, and the modeling context provides limited insight into processes of data construction or sources of variability.[3]

A significant chunk of the secondary education literature also advocates for instructional activity focused on statistical facility around reasoning and argumentation with data models and problems imitating "real world" applications (e.g., mushroom brush factories; Rubin et al., 2006). For instance, some authors suggest that giving adolescents "simulation of life problems" supports the development of quantitative reasoning as well as "real life learning and problem-solving," such as the task of evaluating imaginary sports players abilities (i.e., comparing how far players can jump, how fast they can run) to make a successful team (Lesh et al., 2003, p. 43).[4] A portion of this literature has focused thinking relationally about how variables—usually bivariate data—or quantities change continuously (e.g., rates of change problems about radii of water ripples in Musgrave & Thompson, 2014; dimensions of a box increasing in size in Moore & Carlson, 2012; ambulance arrival times in Cobb, McClain, & Gravemeijer, 2003; the time it takes for a headache pill to take effect in Konold & Lehrer, 2008). Typically, the tasks resemble fictive word problems; the use of computer or digital tools for modeling is less common (e.g., Cobb, McClain, & Gravemeijer, 2003; Konold & Lehrer, 2008). We also see this in the elementary education literature, where researchers provide both real and fabricated datasets of

[3] With open govenermnt data tools (e.g., gapminder.org, socialexplorer.com), the data are already usually averages, direct measures or "best estimates" within an area unit (the case) at a point in time, and the question for users is how values for the units in question (e.g., countries, census tracks) vary together (covariation) over time.

[4] Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, and Woloshin (2007) would argue these kinds of mathematical or statistical tasks do not constitute real-world problem solving," as compared to using at statistics to inform health decisions about breast cancer treatments, preventive health measures, and risks.

“naturally occurring” phenomena such as bird wingspan or car battery life (Bakker & Gravemeijer, 2004; Lehrer & Schauble, 2004).

However, the data modeling tasks rarely index to a real passage of time or meaningful sense of place (i.e., they do not “take place” anywhere, see Ehret & Hollett, 2013; Leander, Phillips, & Taylor, 2010) that would support comparative historical or spatial analysis (Manz, 2012 study in the wild school backyard and the children’s comparative ecology of a local river, pond, forest, or prairie in Lehrer & Schauble, 2012 are such exceptions). Occasionally, students have been asked to reason about a single variable with respect to time in years on the x-axis (e.g., average monthly temperatures in Fujii & Iitaka, 2013; CO_2 emissions in Cobb, McClain, & Gravemeijer, 2003). Nonetheless, discussions of variation and trends for multiple variables over large temporal scales and/or geographic scales are not characteristic of these modeling environments.

Rather, designs typically imitate situations that researchers feel will be accessible from the lived experience of students. Bakker (2004) describes this pedagogical approach towards statistics and mathematical modeling as a theory of “Realistic Mathematics Education” (RME):

> The central principle of RME is that mathematics should always be meaningful to students. The term ‘realistic’ stresses that problem situations should be ‘experientially real’ for students. This does not necessarily mean that the problem situations are always encountered in daily life. Students can experience an abstract mathematical problem as real when the mathematics of that problem is meaningful to them. (Bakker, 2004, p. 5)

Such studies attempt to cultivate students’ interests in mathematics by using context for problems or projects that invoke students’ cultural or out-of-school experiences (Wager, 2012).

At best, these purportedly *realizable* contexts (Wake, 2014) encourage student interest in statistical reasoning, challenge students mathematically, and thus support productive disciplinary engagement (Engle & Conant, 2002). However, the contexts are generally only tenuously connected with youth's lives and experiences beyond the classroom or school (studies of local water quality would be another exception to this), or such connections between personal or family experiences and the data modeling activity are not highlighted in the empirical descriptions of the studies. Many school-based contexts for data analysis and exploration are not consequential to an individual's past, present, or future activity (Hall & Jurow, 2015; see also Gigerenzer et al., 2008 for a discussion of consequential health contexts for statistical literacy and problems-solving), or at least not consequential beyond academic achievement and matriculation into a higher-tracked mathematics class or college.

For instance, Cobb, McClain, and Gravemeijer (2003) state that it was important to use datasets that have a history with students (SAT test data and student expenditures, AIDS and treatment data, ambulance time data, car emissions data, and alcohol consumption), although it is unclear how the students were involved in the data generation process and if they chose the subject matter (not documented in the report). The authors write that students shared personal narratives in conversations of data creation at the beginning of the seventh-grade experiment, but discussions of students' experiences do not persist into the reported second phase of experiment in which the data is modeled at the end of the year. In this study, as in others, middle-school youth are not reported as having a personal stake in correlation graphs about SAT scores or alcohol consumption. Conversely, I conjecture that if the data were personally meaningful for the participants' lives and a part of an instructional design that drew explicit attention to the relationship between the individual and the social aggregate, the data and analysis could inspire student views of themselves as historical and social actors (Gutierrez & Jurow, 2016; Gutierrez & Vossoughi, 2010), as agents in enacting broader social and cultural change as represented by the data.

Modeling activities that do not connect deeply with students' experiences stand at risk of becoming compulsory. For example, Sandoval and Millwood (2005) designed activities in which high school biology students engaged two week-long, computer-supported investigations of two real cases of natural selection: one that asked students to explain the survival of some Galápagos island finches and not others after a natural catastrophe and a second problem that asked students to explain how the bacteria that causes tuberculosis developed resistance to antibiotics. Students were asked to create arguments through the coordination of claims and evidence using data tables and models. The authors report that many students engaged in answer finding likely to satisfy the teacher instead of interpreting the data to create meaningful explanations.

Lehrer and Romberg (1996) also present a departure from modeling a made-up world to a certain extent. Their fifth-grade students in their study created data models from surveys they invented about their present-day daily routines after learning about lifestyles in Colonial America. Yet, while the design privileged youth interests and experiences, whether the students felt that the data reflected themselves as historical, social actors was unclear. The focus of the class was to model and summarize trends across classmates, and the design lacked critical perspectives towards colonization in relation to children's daily routines.[5] Notably, the social studies curricula similarly treated the topic of colonization as unproblematic.

[5] While navigating conversations for younger children around larger social processes from critical perspectives has challenges, studies have demonstrated that young children can participate in activities and discussions related to power relations and society-level social issues. For instance, Turner, Gutiérrez, Simic-Muller, and Díez-Palomar (2009) productively engaged Latino/a elementary-age youth in an afterschool club in critical mathematics tasks around

Indeed, there are few reports in the data modeling education literature of designs that position youth engagement in building, evaluating, and refining models in relation to larger social and scientific issues. Lehrer and Schauble (2012) guided children's ecology investigations into a local water retention pond, a live, local setting for examining constantly changing social and material processes driven by large-scale natural activity (i.e., of the wider, evolving ecosystem) and less-natural human activity (i.e., construction and development of housing developments). The study led to a partnership between the local school district, the local research university, and the city public works department to establish the pond as an official outdoor laboratory. In this case, the outcome of an extended inquiry project and modeling activities situated in a dynamic environment resulted in civic change and children's continued contribution to city planning and engineering. However, the interaction between ecological and larger social processes and negotiations with the city, while welcomed and subsequently facilitated by the research and instructional team, were not part of the initial design, which raises the question how could researchers design for children or young adults to be convincing actors with data models in meetings with city and community stakeholders? (For a design that successfully arranged a youth meeting with city planners around bike lanes, see Taylor & Hall, 2013).

Motivating learners to identify and describe social problems through modeling activity involves connecting students' local or cultural experiences to aggregate data in rich disciplinary ways (Bang & Medin, 2010; Enyedy & Mukhopadhyay, 2007; Philip et al., 2013; Rubel, Hall-Wieckert, & Lim, 2017; Rubel, Lim, Hall-Wieckert, & Sullivan, 2016). Yet it is challenging to

immigration that invited disciplinary learning and drew on youth's community-based knowledge, like calculating the time it would take for border crossings between Mexico and the US. In Lehrer & Romberg's study (1996), the design could have supported a critical conversation around the diversity of experiences and hardships of various populations in colonial America (e.g., displaced indigenous communities vs. colonialists) as well as among American youth today.

incorporate students’ experiences productively in data modeling environments (Enyedy & Mukhopadhyay, 2007) and to design rich STEM disciplinary experiences that are also culturally relevant (Bang & Medin, 2010; Lamb, Polman, Newman, & Smith, 2014). However, as will be discussed in further detail in the final section of the manuscript, the recent availability of LSDS and open source modeling tools makes it possible to engage with the very data and tools that are used to inform policy that governs youth and their families' lives and that they and their families are (or might be) reading about in the news. The modeling tools and datasets currently in schools, on the other hand, typically do not describe scientific and socio-scientific phenomena that affect the lives of youth and their families outside of schools, which mostly eliminates opportunities for students to contextualize their personal experiences within broader social and scientific issues during data modeling activities.

Bias as contamination. One of the struggles for design research that strives to be personally meaningful or culturally relevant and also meet conventional STEM disciplinary standards is that under the latter framework, students’ individual experiences are positioned as bias that interferes with proper data modeling practice and critical interpretation of evidence (Kuhn, 1989; Kuhn et al., 1988). Kuhn and colleagues (1988) describe *belief bias* as individual “weakness” (p. 102) and a “significant deficienc[y]” (p. 112) of both lay (including youth) and professional modelers. In their research, Kuhn (1989) and Kuhn et al. (1988) found the failure to coordinate and differentiate theory and evidence to be prominent among both youth and adults. For instance, when modelers’ theoretical beliefs are not supported by the evidence before them, they are likely to “use a variety of devices to bring them into alignment: either adjusting the theory—typically prior to acknowledging the evidence—or ‘adjusting’ the evidence by ignoring it or by attending to it in a selective, distorting manner” (Kuhn, 1989, p. 687). Approaches that take up this position towards bias in modeling activity assume that objectivity is obtainable as long as one is able to separate their experiences from their data interpretations.

Even the literature that strives to be culturally relevant tends to revert to the perspective that student's personal experiences will interfere with productive mathematical activity significant without instructional intervention. In a community mapping project (Enyedy & Mukhopadhyay, 2007), the research team explicitly tried "to honor" students' local knowledge by supporting quantitative thematic mapping of local-area school resources and demographics using GIS statistical software (p. 167). Researchers were dismayed to find that their students tended to use maps of the socio-economic conditions of their cities and neighborhoods only "to confirm one's preexisting conclusions" (p. 159). The authors argue that the students' deep personal connections and familiarity with the places and social inequities being described and modeled prevented them from critically examining their quantitative maps, adequately using evidence to support claims, and discovering alternative or new interpretations of the data. They suggest that the teachers who facilitated the GIS and data analysis sessions contributed to the problem. In the discourse examples provided, students noticed alternative or surprising findings in the quantitative data, but the teachers quickly brushed these announcements aside, halting any further mathematical inquiry.

The tension in designed learning environments that Enyedy & Mukhopadhyay (2007) describe between mathematical inquiry and culturally relevant pedagogy is rooted in a broader conversation in the education field of what counts as mathematical or STEM activity. While many scholars advocate for designing STEM learning opportunities that are personally relevant, there is a concern that the mathematics (or science) may become unidentifiable in culturally rich, interdisciplinary, and multi-modal contexts (Ma, 2016). Cobb and Moore (1997) avoid entering these murky waters altogether by differentiating mathematics from "data analysis." While Cobb and Moore argue for the importance of "realistic" contexts (e.g., cigarette smoking, eruptions of Old Faithful) to statistical thinking and problem solving, they suggest the work of data analysis does not qualify as mathematics. They write,

> *In mathematics, context obscures structure.* Like mathematicians, data analysts also look for patterns, but ultimately, in data analysis, whether the patterns have meaning, and whether they have any value, depends on how the threads of those patterns interweave the complementary threads of the story line. *In data analysis, context provides meaning.* (p. 803, their emphasis)

While the value of data is revealed when it is used to tell a meaningful storyline about patterns and trends (Kosara & Mackinlay, 2013), detaching mathematics from context challenges theories of mathematics and statistics as living in the world and woven into our daily practices (Lave, Murtaugh, & De la Rocha, 1984; Saxe, 1988).

Summary. Using the STEM education literature as focusing lens, we can identify the points at which designs for learning with large-scale data will build on or depart from current reports and perspectives:

- In elementary education, we find studies of iterative mini-modeling cycles in which children pose questions, collect data and/or create measurements, construct data inscriptions, generate models of processes that include chance, and make inferences about the classroom phenomena under investigation to generate new questions. These reports demonstrate how designs can support the development of important modeling practices like reasoning about trends and generalizing about data samples. These studies are important for thinking about learning with LSDS since many of the challenges of interpreting and using statistics persist even when geographic and historical scale of the phenomena under study is much larger.

- In secondary education, students engage with data at a distance. In some cases, there is an effort to tie content and contexts to student interests and experiences but, for the most part, there is little evidence that the students take up conceptual tools to change the way that they view their interests and experiences. This may be because the tools and data necessary to supporting other kinds of modeling and analysis (e.g., spatial or historical) were previously not as readily available.
- Generally speaking, lines of inquiry in these learning environments tend not to extend into critical examinations of broader social issues that researchers fear could sidetrack learners from disciplinary content and traditional academic achievement, that the project would motivated by the ideology of the researchers as opposed to the interests of the students (Brantlinger, 2013; 2014), or that the data will confirm experiences and avoid asking and exploring questions (Enyedy & Mukhopadhyay, 2007). In turn, designers for learning environments with LSDS should devote time to investigating what would be meaningful big data for youth and what kinds of dynamic modeling interfaces will motivate and support asking questions.
- There are some scholars who argue that mathematics and statistics can and should be used to model and examine power relations and critique social injustice despite its challenges (Enyedy & Mukhopadhyay, 2007; Gutstein, 2006; Philip, Olivares-Pasillas, & Rocha, 2016; Philip et al., 2013; Rubel et al., 2017, 2016), although their research is still often tasked with producing canonical evidence of STEM learning to satisfy the question, “Where’s the math?!” (N. Enyedy, personal communication, April 13, 2016). Some of

this STEM modeling work that takes up critical pedagogy goals will be revisited in the final section of the paper when I review recent learning sciences research that engages youth in data modeling and mapping in creative and personal ways.

- As we will see in the subsequent section, the suggestion that mathematical and scientific inquiry, whether through modeling or mapping, can be and should be unbiased—as in the myth of objectivity—is problematic if we consider reports of professional STEM practice that demonstrate that data, no matter how big or small, are socially constructed and value laden with human meaning (Latour, 1987; Porter, 2012). Of course, standardized processes and tools help professional communities reach some level of consensus about interpretations of data to produce mathematical, scientific, and social-scientific advancements, but participation in collective critique is still a social practice.

Data Modeling in STEM Professional Fields: Important Conceptual Practices

Ethnographic and observational studies of professional scientists, researchers, and analysts engaged in data modeling reveal that professional data and model creation and interpretation involve messy, complex, representational practices. In STS, models are constructed representations that not only support prediction and explanation but also generate questions and facilitate problem solving and decision-making. Their meanings are often contested and constructed through extensive negotiation; professionals in STEM fields spend significant time sitting around model displays, making claims and proposing evidence in order to make sense of the data models. Furthermore, *professional visions*, the organized ways of viewing and construing events and inscriptions that reflect the historical traditions and interests of a particular social or disciplinary world (Goodwin, 1994), and epistemic beliefs influence how stakeholders perceive and interpret the data, the measures, and the model (Goldstein & Hall, 2007; Goodwin, 1994; Hall & Horn, 2012; Stevens & Hall, 1998).

Distributed data modeling. The cognitive and representational practices of professional STEM data modeling work are complex and distributed (Giere, 2002; Hall, Wright, & Wieckert, 2007; Hutchins, 1995; Hutchins, 2006). Data models in professional STEM settings are very costly and complicated to collect and produce (Bowker & Star, 2000; Hess, 2005; Latour, 1999; Latour & Woolgar, 2013). Indeed, the time it takes to produce scientific data and analyze models is often long—data modeling projects extend over years, sometimes without a conclusive ending for decades (Goldstein & Hall, 2007)—and data collection, storage (i.e., shared data inventories on servers), modeling, and analysis often occurs across multiple locations and even global networks (Knorr Cetina, 1999).

Strasser (2012) writes that three aspects typify the distributed modeling process of data-driven research today. First, data collection and analysis often involves a collaboration of researchers and analysts from different backgrounds and from different stakeholder communities; often the researchers who collect or produce the data are in different fields than the researchers who analyze the data. For instance, in biomedical and health fields, funders, researchers, data providers, clinicians, data scientists, and librarians all negotiate data decisions (Margolis et al., 2014). Second, the data that professional modelers engage with often are large-scale and require digital tools for statistical analysis, which leaves plenty of room for interpretation, negotiation, and persuasion (Strasser, 2012). We find examples of Hall et al.'s (2007) description of various meetings between statistical consultants and entomologists, flu case researchers, and ecologists (all separate cases); they found that across cases, meetings involved complicated discussions through which consensus was assembled among meeting participants and new routines for future work are established. When the consultants advocated for particular statistical techniques, using forms of narrative talk, new understandings, concepts, and representational forms were created collectively, and the various scientific projects advanced. Third, Strasser (2012) points out that many scientists work with data without the experience of either producing or collecting the data (in this aspect, one could argue that modeling studies with car battery datasets resembles "authentic" professional data-driven scientific practice). However, conversations in laboratory meetings around models still index the field environment, whether in reference to pastime field experiences or hypothetical future trips (Goldstein & Hall, 2007; Hall et al., 2007; Noss, Bakker, Hoyles, & Kent, 2007).

For the professional modelers, coordinating field experiences with measures and models is important for integrating micro and macro levels of analysis (Cicourel, 1981). For example, in Goldstein and Hall's (2007) studies of regulatory and field biologists looking at models of lizard habitats, scientists relate their field experiences of soil quality and animal sightings in the desert to verify or challenge the data model of lizard habitats; this process of introducing field experiences to evaluate a model is called "ground-truthing" the model (Pickles, 1995, 2006). (Citizens do this as well in conversations about future development around maps of their neighborhoods with professional urban planners; see Taylor & Hall, 2013). Similarly, in Hall et al. (2007), both the entomologists and the field ecologist, in conversation with a statistical consultant, have to determine when to use statistical techniques (to identify new termite groups and classes of biodiversity, respectively) in relationship their field data (jars of bugs to be collected or sampling in the stream). In Noss et al. (2007), Jim, a technician at an industrial factory that produces plastic film, indexes his experience on the shop floor onto series of graphical models that display historical data describing the tension of the plastic tape and revolutions per minute of the machines in order to solve problems in the production line. In all of these examples, shifting between micro and macro data is an important but complicated interpretative practice for asking and answer questions with models (Cicourel, 1981). Furthermore, ground truthing serves one way to settle correspondence problems between model structures and outcomes in the world being modeled.

Modeling with passion. Representational practices, professional and disciplined perceptions, and epistemic beliefs always influence how stakeholders perceive and interpret the measures, the resulting data, and the models created using the data (Goldstein & Hall, 2007; Goodwin, 1994; Hall & Horn, 2012; Stevens & Hall, 1998). In these contexts, participants assemble and coordinate aspects of the model that are specific to their histories of participation and learning in a particular field (*disciplined perception;* Stevens & Hall, 1998). What they notice in the models indexes the particular interests of their professional or social group (professional vision; Goodwin, 1994). However, if that professional vision is not shared across data modeling participants, data modeling conversations can quickly become contentious (Hall, Stevens, & Torralba, 2002; Hall et al., 2007). Indeed, conflicts often arise in collaborative modeling in ways that reflect different professional communities, especially when careers and funding are on the line. In turn, an individual's descriptions of real and imagined field experiences and their efforts to make that scene accessible to an audience become very important to convincing other lab meeting participants of what one is seeing in the data.

For example, Goldstein and Hall (2007) describe how conflicting professional viewpoints of the local scientists, state regulators, and local land managers could not be resolved to move forward with an agreed upon model interpretation in a meeting to review an existing model of viable habitat for an endangered lizard species. In one highlighted scenario, the local field biologist ground-truths the measures in the model of the land preserve (a measure of light intensity/reflectance that also indicates soil density and likelihood of lizard habitat) by suggesting that they do not reflect human activity (off-road vehicles) that he has witnessed while out in the field. While traversing off-road vehicles stir up sand to create the sand conditions that the lizards like (lizards prefer sand that is loose or not compacted), he suggests the vehicles' presence is destructive and reduces the likelihood of lizards in that area. Subsequently, a regulatory biologist deepens the uncertainty around the reflectance measures by asking about how wind can affect the sand quality and whether are not the reflectance layer in the model captures these longer natural processes (i.e., the wind sources). In this case, the regulatory and field biologists treat the same data about species-habitat relations in ways that reflect different forms of disciplined perception and professional vision (Goodwin, 1994; Stevens & Hall, 1998). Consequently, the meeting participants assume different orientations towards time, space, and agency as revealed in their talk that leads to trouble in the interaction and reaching a consensus around the model (Goldstein & Hall, 2007). At one point, when asked by a land manager whether the plan area for the current land use development project would become lizard habitat without human activity, the local biologist suggests that they consider what the habitat will be like in thousands of years in a time and place where lizards have outlasted humans.

This emphasis on the social ways of seeing and understanding models underscores that collaborative data modeling practices are always social, subjective, negotiated, and messy. From start to finish, data modeling is a human endeavor. As Latour describe in *Pandora's Hope* (1999), sweat and labor goes into transforming samples from the field into data, maps, models, and reports in the lab that can then support canonical descriptions of worldly phenomena; indeed, he points that scientific facts are also artifacts of human construction and fabrication. Moreover, as exemplified above, despite any distance between the field (where the phenomena being modeled lives) and lab (where the modeling takes place), discussions of modeling assumptions become conversations about personal matters and moral and ethical stances. Duncan (2006), in her analysis of meetings of professional scientists and other stakeholders conducting a regional assessment of the ecological (vegetation, wildlife) and socioeconomic (land use, policy outcomes) conditions of a coastal landscape area with GIS models, similarly found that participants' data analysis processes reflected their social values and goals. In another example, Hall et al. (2007) describe the statistical consultant and the respective scientists as adopting moral positions in the decision over whether or not to use a statistical tool.

However, this understanding of data modeling as social and interactive is spurned today in a society that endeavors to produce what Porter (2012) calls "thin" descriptions of the world. Twenty years ago, Porter (1996) traced the history of society's ever-growing reliance of quantitative expertise in public and bureaucratic decision-making. He suggested that by adhering sanctioned, mathematical rules, quantitative information afforded increased public trust. More recently, Porter (2012) has retraced some of this history to suggest that the pendulum has swung too far in the name objectivity; efforts to make scientific knowledge impersonal (Porter, 1996) have resulted in a form of "technical" science that is "motiveless and detached from material and social circumstances" (Porter, 2012, p. 215). Porter (2012) writes that prior to the introduction of positivism in the 19th century, scientific and technical work used to be understood and accepted as not value free but depending on and reproducing a Christian moral order. Then, in the 20th century, Porter laments that the sciences and social sciences no longer desired what Geertz (1973) calls *thick descriptions* of social and cultural life—accounts that seek to reveal the "depth and complexity" of human interaction and its cultural products—in favor of objective analysis (Porter, 2012, p. 211).

Paradoxically, Porter notes that the embrace of scientific objectivity was an "answer to a moral demand for impartiality and fairness" among the public (Porter, 1996, p. 8), and significant "moral cultivation" and "preaching" has since gone into branding scientists and social scientists as disinterested in, and separated from, the social world (Porter, 2012, p. 209). Indeed, human morals and values have never been divorced from scientific practice (Law, 2004). As we see in the examples above, "differences of politics, culture, and values have not only persisted, but proliferated" in scientific and social-scientific work (Porter, 2012, p. 211).

Storytelling with data. STEM professionals operate diverse ways of "seeing" data—of seeing statistics as measures and evaluating models—through particular forms of discourse, gesture, and tool use (representation) in order to tell stories about society and the world.[6] Becker (2007) has chronicled how different professions (e.g., photojournalist, statisticians, cartographers, novelists) tell about society. He writes that the reports that they produce are artifacts constructed in their professional, organized activities that only become meaningful through communication and interpretation. Neither data nor models speak for themselves; rather, "someone speaks for them, interpreting their meaning" (Becker, 2007, p.14). Accordingly, storytelling provides essential and expressive means for building relationships between representing (i.e., model display) and represented worlds (Gravemeijer, 1994; Hall, 2000; Latour, 1999). In professional STEM modeling contexts "in the wild" (Hutchins, 1995), stories describe forms of agency, influence, and changes through time (i.e., covarying relationships

[6] Duncan (2006) has labeled this practice of assembling stories in collaborative data modeling contexts as *story-making*; in her research on collaborative modeling of coastal landscapes with GIS maps, she found that story-making is an initial step for data analysis.

between quantities) that are inaccessible except through specific courses of talk that narrate and action that animates the model (Goldstein & Hall, 2007; Goodwin, 1994; Hall & Horn, 2012; Hall et al., 2007; Hall et al., 2002; Stevens & Hall, 1998).

The stories that scientists and analysts tell with data models depend on particular forms noun and verb predication in utterances, gesture, and tool use. Who or what modelers animate as actors in their stories often reveals their underlying assumptions about the represented or modeled world. For example, Ochs, Gonzales, and Jacoby (1996) describe how laboratory physicists animate different worlds in talk to construct models that describe their physics experiments. The physicists met to determine how to configure their experiment (temperature, strength of magnetic field) to attain certain domain states for a physical system. Ochs et al. demonstrate that they utilize linguistic and semiotic resources (static graphic representations of experiments) to convincingly present their findings and ideas in meetings. The physicists, in the performed construction and reconstruction of the graphic representations to their colleagues, assume multiple perspectives through the use of different grammar types. Through their use of pronouns and gestures with indeterminate referents, the physicists animate themselves and their bodies as both the agents conducting the experiment and the modeled physical entity (an atomic system). Alternatively, the physicists also animate the physics system as having human agency and cognitive power. Ochs et al. suggests that the ambiguity of referents permits assigning different degrees of agency and causality in the experiments to either the physicists or the physics itself. When physicists project themselves into the model as physical systems moving through different domain states, they empathetically assume the perspective of the phenomena being modeled to enact physical events (changes in domain states) that are otherwise impossible for them to experience.

Hall et al. (2007) and Goldstein and Hall (2007) also undertook a complex analysis of multimodal, intertextual modeling in their ethnographic studies of scientists at work. Hall et al. (2007) found that scientists (entomologists and field ecologists) and statisticians assembled story or narrative structures to support the decision of whether or not to use statistical cluster analysis methods and algorithms for insect classification and modeling stream biodiversity, respectively. The entomologists, field ecologists, and statisticians used particular object-referents and verb predication to animate both technical concepts and people as protagonists in their discussions of the role of statistical tools in facilitating the discovery of bug clusters and species classification. In Goldstein and Hall (2007), participants also use narrative forms in their talk, as in the example when local biologists tells the story of a potential future apocalyptic time where "people die out" and lizards hang on (p. 72). Across these modeling settings, participants coordinated distinct scales and orientations towards past and present time, space, and agency in their talk and gesture in order to make the represented worlds present for their colleagues in compelling ways (Goodwin, 1994; Hall & Leander, 2010).

Summary. The studies highlighted describe professional modeling as multimodal, intertextual activity; namely, these studies underscore how talk, gesture, and visualizations together construct relations between scientists and other analytic professionals and their objects of inquiry. Making sense of data models often involves coordinating multiple technologies and data outputs with professional peers, reasoning about multivariate relations between aggregate data and personal field experiences, and animating and narrating of graphical space and its referents. Several aspects of professional modeling practice from this literature are important to highlight for thinking about designs of collaborative modeling opportunities with large-scale data for youth:

- Modeling brings stakeholders with different histories of participation in distinct socio-technical practices and different forms of disciplined perception together, sometimes with great difficulty, to determine what the models describe and to choose the next step in the modeling process.
- Storytelling can facilitate collaborative decision-making and analysis (in the face of disciplinary differences) by helping audiences or co-narrators or co-modelers connect to the data: "To provide a deeper connection, the story aims to get the reader or viewer closer to the data, or to at least find out how it relates to them" (Kosara & Mackinlay, 2013, p. 46).
- In particular, storytelling in scientific modeling contexts commonly involves forms of ground-truthing, in which professional modelers bring their field experiences into contact with aggregate data in order to make phenomena described with aggregate data real and present for other participants. These narrations reflect participants' professional values and epistemic and moral stances. In turn, how can we design for youth participation in practices of social science inquiry that embraces "messiness," is deliberate about values, that recognizes the shaping influence of method, and challenges the hegemony and "singularity" of "objective" social research methods (Law, 2004)?
- Ground-truthing as a representational practice entails shifting between micro (personal) and macro (aggregate) scales in ways that support and deepen sense-making of models and their structures, such as when a researcher puts forward a field experience to challenge a quantity or measure in the model. The introduction of personal experiences under these circumstances appears to differ from a reductive understanding of bias (i.e., when bias prevents critical interpretation of data). Modeling activity with big data in that

relates personal experience to the aggregate similarly can similarly leverage lived experiences and local knowledge to deepen understandings of the modeled phenomena.

Designing for Learning with LSDS

Relatively few studies in the education literature have carefully or seriously considered how to bring youth into critical inquiry, data modeling, and learning with publicly available LSDS and interactive data visualization tools (Wilkerson-Jerde & Laina, 2015). One explanation for this is that, previously, selections of standardized data and tools were too expensive for individual, lay users to acquire for building models (Venturini et al., 2015). Presently, while there has been a surge in data and tools that are freely accessible online, the majority of current users are adult, male, white professionals with computer science backgrounds. Indeed, work needs to be done to bring in new kinds of users, like youth, into the data movement.

Accordingly, in this section, I review new learning sciences scholarship that endeavors to design for youth learning with LSDS. These studies have posed and started to answer questions such as what is involved in closing the data science fluency divide and how do we engage and support learners in identifying, modeling, and describing social problems with big data. With references to this work, I describe important considerations—potential design principles—for designing for learning with LSDS. First, I consider what it means to design so that participants assume critical perspectives with large-scale data—perspectives that include both critiques of measures and methods, as well as political and ideological interpretations. Second, building on designs for culturally relevant pedagogy, I offer suggestions for supporting getting personal with data as a modeling practice (Kahn & Hall, 2016). Third, I review some of the risks and challenges for these design efforts. Lastly, I introduce a case study from a multi-year design-based research project (Cobb, Confrey et al., 2003) for critical storytelling and modeling with open big data (Kahn & Hall, in preparation). Open data and tools are changing the terrain for how scientists and social scientists tell stories about society (Becker, 2007) and are fostering new forms of public scholarship of social and scientific issues across local and global scales. The design space I describe with critical data storytelling opens that territory to youth.

Getting Critical With Big Data

The idea of assuming a critical perspective with data is not new to the data modeling community in education studies. However, critical perspectives in data modeling have generally been described as questions related to data collection processes and sampling. For instance, in Lehrer and Schauble's (2004) study, students evaluated their own invented displays using criteria for communicability and design attributes (see also Lehrer, Schauble et al., 2000). The creation of measures leads to evaluating fit and misfit between representations and target phenomena (which one might argue is a form of ground-truthing)—in their case, Fast Plant growth as a model for other plant species—and opens the way for learning concepts of variability and distribution. In other studies, tasks support critical statistical reasoning with data models through practices of asking questions about measure qualities and attributes, creating models to support the questions, and developing arguments, explanations, and inferences. Critical appraisal of a model's measures or the relationship between errors and variation, while important, differs from critiquing social injustices and inequities that are produced historically by mostly WEIRD (Western, Educated, Industrialized, Rich and Democratic) social structures. I suggest that these two versions of "critical" are both valuable and complementary for data modeling opportunities for youth.

Recently, several authors have introduced frameworks for the latter kind of being critical—designing instruction that explicitly encourages social and political criticism. These studies build on sociocultural theories of learning (Lave & Wenger, 1991; Rogoff, 1990; Wertsch, 1998) and critical pedagogy and philosophy (Freire, 2001; Giroux, 2001). The latter includes critical race theory (Ladson-Billings & Tate, 1995), theories of culturally responsive pedagogy (Gay, 2010; Howard, 2010; Ladson-Billings, 1995; Moll & Gonzalez, 2004), and critical pedagogies of place (Gruenewald, 2003).

For instance, Gutstein's (2003; 2006) critical mathematical literacies framework argues that mathematics and statistics should be used to examine sociopolitical and cultural-historical contexts and challenge existing power relations. Enyedy and Mukhopadhyay (2007) adopted this critical mathematics framework in their community-mapping project for high school teens. They designed their instructional activities to support students' use of statistics using GIS software to evaluate claims about inequitable distribution of resources and funding across schools and districts in the Los Angeles area. They wanted students to consider the quantitative data in relation to individual student cases at their own schools and the qualities of the data distribution across the LA region. The research team hoped mathematical critiques of inequity would serve as a way "to contribute to the public discourse and be part of the solution" (Enyedy & Mukhopadhyay, 2007, p. 143).

More recently, Philip et al., (2013) advanced a framework specifically for *Big Data literacy*. They suggest that in light of the rise and ubiquity of large-scale data (both government and digital traces), youth should be able to: (a) be proficient with analytical data tools to engage in the "formulation of questions and the generation, collection, representation, visualization, analysis, interpretation, and communication of data" (p. 115); (b) use big data to serve their own interests; (c) use data to address societal and community issues; (d) recognize that no data are objective, and all data rest on assumptions about the world; and (e) leverage the discourse associated with big data technology to support causes they care about while also recognizing its limitations in meeting the needs of the most marginalized groups.

Central to Philips et al.'s (2013) argument is that big data is now, and will continue to be, essential to civic and democratic participation, and thus youth interaction with big data should be embedded in conversations about equity and social justice. The researchers want youth to use big data with their eyes open towards potential infringements on civil liberties, the effects of contemporary and historical power relations, and the opportunities for this new technology to invite cross-cultural relationships. While this approach to design-based research in data science takes a risk that overarching social justice issues will not interest youth (Brantlinger, 2013; 2014), Philips et al.'s (2013) framework ensures that there is room for critical conversations around topics such as equity and race that are arguably always present in social science data, particularly when students want to have those conversations (Philip et al., 2016). Additionally, as for all instructional design, disinterest is avoidable if designers honor what participants find meaningful about the data or the data activity.

There is a set of empirical work on *critical data literacy* practices that has made headway towards getting critical with big data, both in terms of critiquing of measures and interpreting of social relations. In Polman and Hope (2014), high school students created data visualizations and published science news stories on topics of personal and local interest to develop scientific data literacy. They framed their work as part of an effort to educate youth who "can critically engage with science information for personal decisions about issues such as health and public policy related to science (e.g., fracking, climate change, smoking)" (p. 316), albeit without further defining critical engagement. Nonetheless, the authors present cases of youth who participated in scientific inquiry practices that positioned them to productively mediate between their local community or family experiences and the scientific community through cycles of research, revision, and publication of their news articles and infographics.

For instance, one student focused on the relationships and tensions between Chinese cultural healing practices, such as those practiced by her family, and Western medical research. Another published an article promoting awareness of a rare disease that affected a family member, and her article subsequently garnered attention from a national advocacy organization representing families affected by the disease. A third student reported on teen pregnancy in relation to receiving sexual education in her school community, which was an important issue among her peers and the high school community. The authors also describe a student who reported on a local apartment building's problem with mold; subsequently, the high school media and local community newspaper both recognized his article, which resulted in a summer job opportunity in science journalism. While Polman and Hope do not frame student critical stances towards science as critiques of broader social power relations, the design prompted youth in each of the cases to negotiate family and community knowledge in relation to general scientific and environmental research and to explore personal and societal impacts of their topic of interest.

In separate study on data literacy, Wilkerson-Jerde and Laina (2015) describe instructional activities for exploring data visualization literacy in which middle school youth built digital data representations with city public data. The research design prompted youth to reflect on social and scientific matters affecting their city, such as racial demographics (e.g., To what extend has the city diversified over time? Is the city as diverse as it claims to be?) and public land use (e.g., how much land is devoted to public services?). Although the authors do not frame student outcomes in terms of opportunities for critical data literacy, their design created possibilities for discussion and learning around critical perspectives with big data, in both senses of the term “critical.” For instance, the class could have discussed what does it mean to use data to tell multiple stories or support competing claims (such as those regarding diversity), what to do when the available data does not align with the story one wants to tell, or the status of minority populations in the city as demographics shift or stay the same over time.

Rubel et al.’s (2016) design-based research study, anchored mathematics, such as comparisons of ratios and rates, to issues of spatial justice with zoomable GIS maps of city data on financial institutions, median household incomes, lottery revenue, and spending. Students raised important questions of equity and spatial justice (i.e., “Is the lottery fair?”) with maps, multimedia, and accompanying narratives that were critical of the lottery and pointed to related social, systemic problems. Accordingly, the more participatory nature of Rubel et al.’s (2017, 2016) projects reveals the power of learner-led exploration, reflection, discovery from observations and making (Collins, Brown, Newman, 1989; Dewey, 1933; Engle & Conant, 2002; Hall, 1996; Lehrer, Carpenter, Schauble, & Putz, 2000; Reiser, 2004). Participants in Rubel et al.’s (2017, 2016) studies learned to direct the data towards asking questions and telling stories about what they discovered. This kind of active, analytic engagement with big data is necessary in order to make data meaningful for users, at least in part because the “value” of LSDS is “latent and requires innovative analysis to unleash” (Mayer-Schönberger & Cukier, 2013, p. 128).

Counter-stories, counter-models, and counterpublics. For most of the students in Rubel et al.’s (2016) study, a critical narrative was the outcome of an extended inquiry into the lottery, income distribution, and spending in their neighborhood and their own family practices. The critical stories the teenage youth told that spanned persona/familial, street, neighborhood, and city scales were *counter-narratives* or *counter-stories*: the stories of marginalized people’s experiences that challenge dominant discourses, narratives, and racial privilege of the majority

(Solórzano & Yosso, 2002). From a critical race theory perspective, counter-stories serve as tools for examining and opposing master narratives of racial oppression and of white privilege (Howard, 2008) that promote claims about the cultural deficits of historically-marginalized people (Solórzano & Yosso, 2002). In turn, telling counter-stories work against these claims by strengthening the social, political, and cultural power and voice of minority communities. For instance, in Rubel et al.'s (2016) study, Hispanic and Black youth from a low-income, urban neighborhood, developed stories that revealed the complex role of the lottery in their community. Drawing on both digital data investigations and interviews with community residents, youth told stories that critiqued the lottery system as taking advantage of their community members and simultaneously recognized the bodega or corner store, where the lottery tickets were sold, as community hallmarks for "sustenance and hope" (p. 20).

When individuals assemble and tell counter-narratives over large-scale data models, they engage in *counter-modeling* (Kahn & Hall, 2016). Counter-modeling describes changing model assumptions to show patterns that support an alternative explanation or counter-narrative. Given open data and tools, one could remix or make models that reflect a different set of assumptions or bring new social problems into view, challenging a normative and majoritarian storyline. In counter-modeling, a critical stance towards to data collection processes, measures, qualities or model attributes can support a question that drives the creation of a new model, which, in turn, advances a social critique about the world. Counter-modeling entails understanding models and associated causal inferences (stories) that are then subsequently contested.

For instance, in Kahn & Hall (2016), undergraduate and graduate preservice mathematics teachers assembled a model with Gapminder comparing China and the US on CO_2 emissions to answer the question which country produces the most carbon pollution. Students demonstrated that, in line with the popular narrative in the news media, China was the world's worst polluter on a yearly measure of total CO_2 emissions, but by looking at a measure of CO_2 per person in their model, they countered that the US could be viewed as the leader in carbon emissions and climate change; the students also suggested that their consumptions of goods produced in Chinese factories (like their computers) accounted for some of China's carbon emissions, which was not reflected in the distribution of global CO_2 in the model. Counter-modeling follows the lead of other learning sciences social design experiments (Gutierrez & Vossoughi, 2010) that have sought to give new representational tools and agency to youth in order to encourage their voices and support social change (e.g., see Taylor & Hall, 2013 and Van Wart, Lanoutte, & Parikh, 2016 for design studies on youth *counter-mapping*).

Engaging youth in generative and compelling counter-storytelling, counter-mapping, or counter-modeling may demand positioning youth as members of a scientific and social-scientific *counterpublic* (Hess, 2011). A counterpublic is an alternative network of organizations and citizens across fields of interest that advocates, on behalf of a greater public interest, a different perspective on scientific issues than the official, dominant perspective and is ready to act in local political and social conflicts. Participation in a counterpublic network that maintains a critical consciousness (Enyedy & Mukhopadhyay, 2007; Ladson-Billings, 1995) and engages in counter-modeling with large-scale data could potentially motivate *counter-data actions* (Dalton & Thatcher, 2014) to respond to issues of social justice and equity. For example, the Invisible Institute in Chicago is a collaboration of journalists, social activists, university law faculty, artists, and community residents that requested the release and then published data models tracking filed complaints of police misconduct (https://cpdb.co/findings). The reports attracted both local and national government and news media attention; a month after the data was published online, the city's mayor fired the police superintendent.

Getting Personal With Big Data

One way to support critical inquiry in modeling with large-scale data is to design learning environments that support the development of a relationship between youth identities and the story told about the model—what we have termed *getting personal* with the big data (Kahn & Hall, 2016). Personal interest, meaning, and agency have often been black boxed (e.g., in battery life) in data modeling activity. With open LSDS, pursuing personal interests into curated data sets is now possible for youth and adults. Getting personal with big data is both a practice of modeling with data and engaging in social analysis. It builds on Cicourel's (1981) theory of integrating micro-and macro analysis levels for studying and telling about society (Becker, 2007). Cicourel argues that neither micro- nor macro structures are self-contained units of analysis; rather, macro social facts emerge from the routine practices and encounters of everyday life. In turn, interpreting and relating individual, personal experiences to models of aggregate data seems both basic and generative for critically making sense of the complexity of social life.

In order to get personal with big data, in addition to holding an active interest or stake in the phenomena being modeled, storytellers and modelers must actively *position* themselves in relationship to the data model as well as the narrative being told with the model (Bamberg, 1997). In order to re-contextualize personal experience in relation to aggregate, state-collected data, individuals must consider whether their experiences live within or outside of the model (i.e., asking where do I fit or not fit in the aggregate or to what the state says is "normal"). This move to establish oneself in relation to models that describe trends in scientific and social life with large-scale data situates individuals as agents who can enact broader social and cultural change by understanding processes represented in the data. Getting personal thus entails a shift in position that reorients oneself to the modeling activity at hand as a historical social actor and an agent for change. Designs for getting personal with big data build on theories of pedagogy that argue that learning environments should (a) be culturally relevant and useful to communities (Bang & Medin, 2010), (b) leverage participants' cultural funds of knowledge (Moll, Amanti, Neff, & Gonzalez, 1992; Moll & Gonzalez, 2004), and (c) be transformative for community members (Gutiérrez & Vossoughi, 2010).

Other learning sciences studies have started to move towards explore how relationships between personal experiences and large-scale spatial and social data can support learning and critical inquiry. Radinsky, Hospelhorn, Melendez, Riel, and Washington (2014) studied middle and college students' reasoning and inference processes using Social Explorer, a web-based dynamic mapping interface that accesses US Census datasets. Students engaged in historical inquiry projects to investigate African American and Latino migrations in their local neighborhood. The authors viewed students' lived experiences as "a meaningful source of knowledge for social inquiry" (p.152) in their research designs and encouraged students to look across data, texts, and experiences or memories of family and community history (e.g., interviews with community elders about migration). They found that recalling personal experiences prompted new inferences or questions about data trends describing population changes. For example, students drew on conversations with family members or interviews with neighborhood elders about migration experiences to draw relevant inferences regarding data patterns, such as changes in racial segregation as evidenced by losses or gains in population in neighborhoods over time.

Additionally, Polman et al. (2016) looked at the relationship between societal concerns and personal concerns in their studies of youth who create scientific data narratives based on personal interests in health or other science fields (e.g., effects of caffeine, earthquakes, high-school students sleeping habits). They are currently investigating how their designs support students in asking and answering questions that are important for supporting the coordination of aggregate and individually-scaled data, such as: Why would I care or others about the issue? What does this data mean in a broader context? How would it change my own lived experience?

Rubel et al. (2017) discovered that over the course of the project in which youth examined lottery-related socioeconomic data in their city, students engaged in "finding self," locating and exploring their homes, familiar routes for commutes, and other meaningful places that they frequented in the city in GIS maps and other spatial texts (i.e., youth also immediately located their homes when they encountered an oversized floor map of the city that the research team created). Wilkerson-Jerde and Laina (2015) explored how middle school youth coordinated mathematical, representational, and personal community knowledge in order to assemble data visualizations of public city infrastructure and demographic data; they similarly describe how participants "immediately situate *themselves* in the data – seeking those languages and ethnic identities that they felt reflected them and their families" (p. 5). Repeatedly across design studies, we have also found that getting personal with datasets is an entryway into large-scale databases (Kahn & Hall, 2016).

Scaling. Scaling in time, space, and social life is a key socio-technical practice for getting personal with big data and for making critical and meaningful sense of the phenomena being modeled (Hall & Leander, 2010). Dynamic big data modeling tools (e.g., Gapminder; Johansson, 2012) permit scaling in order to answer questions about temporal relations (i.e., between the past, present, and possibly future), geographic relations (i.e., between local, national, and global areas), and social activity (i.e., what is the relation between what I do and what collective society does). Previously, data and tools did not have the capacity to support this kind of conceptual and technical practice. Web 2.0 and the surge of available large-scale data have pushed the envelope for forms of spatial and temporal analysis and modeling.

Both the individual and the collective human story is made accessible by big data. Fully understanding any single social interaction always entails appealing to surrounding context at some larger, macro scale (Cicourel, 1981). Critical storytelling with LSDS thus builds on this premise that traversing micro and macro scales in time, space, and social life can foster relations between both individual lived experience and collective activity. As a result, new insight into issues about social distribution and spatial justice can be revealed.

We see examples of scaling with big data in the public news media: A recent New York Times article wove the personal experiences of African-American residents in Greensboro, North Carolina together with city data on police traffic stops to demonstrate the racial disparities in treatment by police officers and the disproportionate risk of arrests and fines that African-American drivers face (LaFraniere & Lehren, 2015). Embedding narratives of individual experience (that describe my body,[7] my family, my home, my youth during my childhood) within narratives of large-scale aggregate, human social participation spanning much broader histories (that describe many bodies and generations of people within and across nations) creates

[7] Lee (2013) and (Nafus, 2016) write about the Quantified Self movement as offering new opportunities for educational technology and learning. The Quantified Self refers to the troves of body-scaled data produced through wearable electronics and mobile personal devices that track daily activity (e.g., how many steps we take, how much we sleep, how many calories we eat) and now can be easily downloaded, visualized, and analyzed by consumers. While the data is undoubtedly very personal, there are many unknowns for how the data can be publicly shared and aggregated in secure ways.

opportunities for considering the ethics and politics of both one's own decisions and society's choices. At the same time, bringing aggregate data down to the personal scale suggests that individual agency matters, and our individual choices may align with or challenge the direction that the world is moving towards. This new terrain for learning is made possible by access to open data and visualization/modeling tools.

Risks and Challenges for Learning With LSDS

There are risks and challenges associated with using LSDS in designs for learning that foster personal and critical inquiry, but each challenge affords new possibilities as well.

Sample versus census. Hammerman, (2009) offers several conceptual challenges with big data for engaging in statistics and mathematical modeling. First, Hammerman (2009) and others (e.g., Enyedy & Mukhopadhyay, 2007; Konold & Higgins, 2003) note that generalizing and reasoning about aggregate characteristics—that certain features emerge in the aggregate that are not visible when looking at individual elements—is difficult for students. Hammerman (2009) points out that that multivariable data, such as data that are time series or show spatial proximity, can be challenging in their complexity and messiness. Second, LSDS are often census datasets and not samples, and statistical methods of exploring null hypotheses may no longer be apropos, especially with other data mining techniques available (Busch, 2014). Even when large-scale data are samples (which is rare), small differences in the measures students are comfortable with (e.g., mean, median, mode, range) may appear statistically significant but may not actually be practically significant and thus meaningful. Indeed, the sheer volume of available data and huge N sizes are unconventional compared to what exists in statistics and mathematics textbooks (Busch, 2014), which can be overwhelming and challenging for all students in activities (Kahn, Hall, & Phillips, 2014). For instance, what is the meaning of "statistically significant" when exploring relations or comparisons using a census (a population)? How is using descriptive statistics with populations different than with samples?

Part of the struggle with finding meaning in large-scale data comes from understanding the underlying "uncertainty" in the models. The nature of uncertainty in models made with large-scale data differs from the uncertainty that statistical models try to measure in their estimations of sample statistics. Unlike in many classroom studies (e.g., where one asks how certain am I that our group of class Fast Plants model the population of Fast Plants in the wild), one is typically not making inferences from samples about population parameters with large-scale data, even when describing trends (Busch, 2014). For instance, when modeling with a tool like Gapminder, a web-based multivariable modeling tool which access global demographic and socioeconomic datasets, provided from state governments and NGOs like the Work Bank and UN, or Social Explorer, the available data counts as the population for historical modeling.

Using these tools as designed—for comparisons to understand causal historical conditions—there is no way of modeling likelihood, probability, or chance as when one makes predictions about the future (which is possible with other types of tools). Rather, uncertainty with historical LSDS still represents the uncertainty around data quality processes—processes that are often not made explicit. Considerations for designing LSDS as censuses include how well the big data measures capture the phenomena in the wild, many of which are composite measures themselves; the sources of variability for large-scale data; the completeness of data if it is a census; and the trustworthiness of reporting. Furthermore, how would one know if the given aggregate data is "noisy"? Do the measures make unwarranted assumptions about the phenomena through simplifications that encode an interest or bias? These are important questions for critical data literacy and inquiry alike. We see this in play, for instance, when climate change proponents and naysayers offer differ opinions regarding the accuracy of climate change models that use historical times-series of global surface temperatures based on their assessments of the certainty of the measures and variables.

Unlike classroom-collected data (e.g., modeling the age of the artist for a group of self-portraits, measuring growth or plant “success,” etc.), large-scale data, like open government data, undergo processes of distillation, simplification, and selection to be turned into representations and models (Busch, 2014; Latour, 1999; Star, 1983). This kind of time and labor intensive data cleaning and model preparation is now a part professional technoscientific practice. On the other hand, many professional data users and analysts never partake in those processes and yet still manage to develop meaningful relationships and engage in sensemaking with the data without proximity to data collection and processing (Strasser, 2012). In thinking about designing instructional modeling environments with LSDS, how likely are youth to be able to develop conceptual understandings of aggregate characteristics, (multi-)variation, and distribution without access to the complicated processes of data and measure construction? This remains an open question. Potentially, this is the role of soup-to-nuts data modeling described earlier.

Bias. As noted earlier in the paper, there is a tendency for educators and social researchers in data modeling to dismiss individual stories as bias and irrelevant to “true” interpretations. This has been our practice since the development of average measures of human and social life and our conceptions of “normal” in the history of social science: Since the mid-19th century, “the individual person [became] synonymous with error, while the average person represented the true human being” (Rose, 2016, p. 5). Indeed, this rooted perspective towards individual experience as problematic in data modeling is hard to shake, perhaps because our design studies appear to give legitimacy to the fear that personally meaningful engagement will be unproductive in critical sense-making with large-scale data.

For instance, in Enyedy & Mukhopadhyay (2007), students' own experiences in school, from the researchers' perspective, inhibited them from discovering socioeconomic data and stories that describe experiences that differ from their own. Rubel et al. (2016) found that the time the project lived (conceptually) at the "personal/home" scale was less "productive" for certain participants (p. 20): Some students, who were themselves regular lottery players or had family members who regularly played lotto, defended their and others' participation and resisted the curricula premise that the lottery systematically takes advantage of residents in communities that face economic hardship. Wilkerson-Jerde and Laina (2010) also describe the mismatch of middle school youth's experiences of the city and city data as an impasse in students' mathematical understandings and progress in the activity.

Youth (and adults) are likely to produce evidence from data that supports their personal beliefs (Kuhn et al., 1988). Consequently, the question how to scaffold reflective and open stances towards data exploration to avoid deterministic causal view of the world and, at the same time, encourage students to consider values that motivate choices of problems, measurement, and investment in modeling (a critical perspective) remains. Enyedy and Mukhopadhyay (2007) give examples of moments where students noticed and introduced surprising tends in the data that could be pursued but were unheeded by teachers. Rubel et al. (2016) conjectured that less resistance or (more productive resistance) might have emerged if the topic of inquiry was student-owned or co-constructed as opposed to researcher-driven by an instructional design team that was neither comprised of community members nor shared the same ethnic and racial background as the students (the research team was white and Asian). Van Wart, Tsai, and

Parikh's participatory mapping study (2010) found their open-ended approach to data collection allowed for students to propose questions and pursue issues of interest in their neighborhoods—questions that the research team admitted they would not have considered or introduced. Another possibility is that the dynamism of the modeling tool could spark questions and generate open inquiry, as with Gapminder's nation bubble trails (Johansson, 2012; Kahn & Hall, 2016).

One possibility is that designs for learning with social-scientific LSDS could make the problem of confirmation bias part of the investigation. That is, studies can support learners in a critical assessment of personal fit (or exclusion) with the aggregate as one of the objectives of the data exploration (Figure 3; Kahn & Hall, in preparation); such a design would intentionally scaffold participants in evaluating personal fit to the aggregate data in their models. One would ask does his or her individual or local experiences correspond (or not) to society's experiences as represented by the LSDS? Participants would discover that either the aggregate is better than, worse than, or matches (in a positive or negative way) their own experiences.

Evaluating Personal Fit With the Aggregate			
		Aggregate data trends	
		Data +	Data -
My (family's, community's) story	Me +	All is swell	I beat the odds
	Me -	My life is a struggle	My challenges are normal

Figure 3. This matrix, adopted from Kahn & Hall (in preparation), represents a framework for evaluating personal fit with big data that could be incorporated as part of design for learning in order to make confirmation bias part of the object of learning.

Modeling depth. Historically, scientists and analysts have always faced what they considered to be a data overload, as new technologies in every era permit novel access to different kinds and greater degrees of data (Strasser, 2012). Today, dynamic, interactive data tools (e.g., motion chart; Al-Aziz et al., 2010; Hammerman, 2009; Rosling et al., 2005) make data trends, variation, and covariation visible in new ways to both public and professional users.

In turn, this access offers novel kinds of support for reasoning about and modeling scientific and social phenomena. However, the public availability and the *modeling depth* of the tools determine the extent to which learners can engage in critical inquiry with large-scale data models. Modeling depth describes the degree to which a model can serve as an exploratory tool for generating questions and discussion about model assumptions, stakeholder perspectives, quantities, and measures. The modeling depth of the object depends on its capacity to allow alternative views of the data, multiple choices of measures, and various selections of quantities that influence the relation of interest. A "deep" modeling tool might also give access to raw data, data sources, and information about data quality and measure validity. Modeling depth becomes visible in the contrast between digital modeling tools and static models, like news infographics. Moreover, the user interface for selecting and manipulating data strongly influences the depth of inquiry processes that are possible in telling stories with LSDS.

There is evidence of how access to, and the utilization of, tools with modeling depth can foster critical inquiry practices in data modeling environments in the learning sciences literature. Gordin et al. (1994) and Edelson et al. (1999) conducted studies with adolescents using multivariable climate and ecology visualizations. They found that the modeling interface's dynamism and interactivity, which permitted selections among indicators, representation types, and side-by-side comparisons supported adolescents' engagement in inquiry-based scientific practices (i.e., pursuing and refining questions, analyzing, communicating results, modeling, inscriptions) and deepened understandings of scientific phenomena and concepts.

Radinsky (2008), in his study of students' collaborative modeling of plate tectonics with GIS software that permitted zooming in and out and the manipulation of data layers, similarly found "the dynamic nature of [data visualization] tools places students in a position in which they are continually both reading and generating new inscriptions" (p. 5) and supported joint sense-making among participants about plate tectonics. Roberts, Lyons, and Radinsky's (2013) CoCensus project created an embodied interaction with GIS Maps using US census and American Community Survey data, in which physical body movements drove model/map choices and animation, like zooming and changing scales, in museum exhibits. CoCensus supports a new kind of meaningful sensemaking approach to the data interface, one that engages both personal and physical connections to data.

Conversely, other studies have shown that restrictions on model choices and a lack of engagement with the interactivity or dynamism of the data modeling tools can impact student participation and learning. Rubel et al. (2016) noted the availability of only select lottery datasets thwarted the discovery of certain narratives about the relationship between lottery and race. Wilkerson-Jerde and Laina (2015) found that students who did not animate models to manipulate data in their visualization missed opportunities to uncover new patterns and advance their mathematical learning. Similarly, in Kahn, Hall, & Phillips' (2014) design study of preservice secondary mathematics teachers modeling global development with motion charts, students who chose not to animate their motion charts were restricted in their capacity to reason about multivariable socioeconomic change over time across nations. Of course, the goals and intentions of the modelers influence the available learning opportunities regardless of a tool's modeling depth; a lack of motivation or buy-in for an activity can restrict one's willingness to engage with a tool, and in turn, his or her ability to reason about trends or patterns.

Reproducing inequities and unjust distributions of power. Access to LSDS can create and reinforce power asymmetries between companies, governments, and citizens (Dalton & Thatcher, 2014). To begin, the cost of producing datasets dictates who can collect and distribute data, which subsequently contributes to this power differential. Indeed, the manufacturing or collection, aggregation, classification, and organization of large-scale data require extensive resources and negotiations (Bowker & Star, 2000; Goldstein & Hall, 2007; Kraemer et al., 1987; Porter, 1996; Star, 1983). The labor and social resources that were mobilized to aggregate the data, the value systems embedded in the layers of standardization in laboratories, and the data that have been discarded are typically invisible in data models (Latour, 1999; Star, 1983; Strasser, 2012). Finally, the popular discourse around how big data can improve democracy, solve social problems, and empower marginalized populations (Hacklay, 2013; Johnson, 2014) "unintentionally but problematically reproduce[s] assumptions that increased knowledge or access through technological capabilities by itself could address social inequalities and injustices" (Philip et al., 2013, p. 118).

Consequently, establishing legitimate, authentic outlets to professional and public audiences, particularly those who are in positions to steer public discourse or realize community change, is a potentially important outcome of student critical modeling and inquiry projects that seek to support youth in civic and democratic participation. Several of the studies highlighted in the review have done this. In Polman and Hope (2014) and Polman et al.'s (2016) studies, students can submit their scientific data infographics and articles to be reviewed and published in a national scientific journal with an outside editor. The infographics and articles can then be accessed by teachers, students, and outside audiences across the country for discussion and learning of personal and public health and scientific issues. In four of the five cases described in Polman and Hope (2014), student participation in data science activities not only strengthened their academic and disciplinary identities but also produced new relations to various cultural and professional communities, both locally and nationally.

In Taylor and Hall (2013), youth participated in a bicycle building workshop and participatory mapping activities in their neighborhoods and then met with professional city planners to share recommendations for improving the local biking infrastructure (e.g., bike lanes). The city subsequently developed a bicycle map that included the bicycle routes that the students had presented to city planners as desirable, and regional planners subsequently included high school youth as part of the city's next formal planning effort. In Van Wart et al. (2010), teenage youth participants also met with city officials to share the digital map products and data collected on the things they would change in their neighborhoods. These studies encourage forms of consequential learning for youth in their communities through their participation in various conceptual and representational practices (Hall & Jurow, 2015).

Notably, the citizen science movement is one possible space to look to for future designs for learning that seek to leverage big data and modeling tools in order to promote democracy, serve local community interests, or benefit a larger public good. Citizen science data projects use dynamic data visualizations to pursue connections between personally meaningful experiences or places and larger phenomena described by aggregated large-scale datasets. In projects like the New York City Street Tree Map, individuals record personal or individual scientific data, like counting and logging tree species on their city block, which they share to a public map and database. Edelson (2012) argues that these citizens science projects, made possible by new GIS tools and technological infrastructure, hold powerful and rich opportunities youth learning and engaging individuals in STEM modeling.

Geobiographies and LSDS: A Case Study for Storytelling and Modeling Self and Society Relations

To end, I offer a brief illustration of a design for personal and critical learning with

LSDS. Kahn & Hall (in preparation) investigate how adolescent youth assemble models and narratives about personal and national or global migration with open public data. The design-based research program consisted of a series of experimental teaching activities in intensive two-day workshops (5 hours each day) over three weeks (three sessions), which were free to the public and housed in the city's main public library branch. The activities invited middle and high school youth to explore the reasons for migration and immigration (What moves families?) nationally and globally at the scale of their personal, multi-generational family *geobiography* or mobility history (What moved my family?). Participants ages 10–16 (n=17) attended 2–4 days of the workshops (between 10–20 hours); the majority (12/17) were comprised of sibling pairs and identified as multi-generational African Americans. Seven participants (3 sibling pairs, 1 individual teen) completed follow-up interviews with their parents 6–10 months later to discuss their workshop activities and projects.

The workshop framed instructional activities in a discussion about the ways in which the proliferation of large-scale data and models has opened up opportunities for powerful storytelling about society (Becker, 2007; Kosara & Mackinlay, 2013) and consistently positioned youth experiences in relation to their ancestors and social history. The participants were asked to assemble *family data storylines* (in Microsoft PowerPoint) that included models of national or global big data made with dynamic, digital modeling tools (Social Explorer or Gapminder) to describe migration at a broader scale and to construct narratives that index the data models to personal familial migration stories and experiences. As a cornerstone for their projects, youth had the opportunity to tell and record an oral history describing their personal family mobility history, which are now housed in the library's permanent public archive. Participants' family data storylines were publicly displayed in a digital exhibit open to the community and family members at the library at the end of the three weeks.

Qualitative analyses found that students engaged in meaningful, personal storytelling and modeling practices to build and refine the family data storyline over the course of the workshop. In order to answer the question "What moves my family?" youth assumed a personal stance on cultural heritage and social life. We also found that scaling personal stories to the social aggregate is complex: Participants performed a range of spatial and historical comparative analyses with models of aggregate data. To do this, they engaged in messy *data wrangling*, laborious practices for managing multiple datasets and measures (Social Explorer and Gapminder each offer hundreds of indicator choices), in order to align the family story to the aggregate data. Participants employed various forms of data wrangling when they encountered choices among indicators, missing or spotty data, income rates at different social scales, or changing data categories over time.

For instance, one participant sought to compare the county where she grew up and lives now to the counties where her grandparents and her mom grew up on a measure of household income in 1960, close to when her grandparents moved from Mississippi to the North. Social Explorer did not populate 1960 survey data for the county where her grandparents lived in Mississippi along the Mississippi River, so she used what she considered to be an equivalent county, also along the Mississippi River in nearby Tennessee, for her comparison. Statisticians would call her decision to use a similar case with data as a proxy for the original with missing data "imputation"—predicting or filling in missing data values on the basis of what is known about relevant covariation between the two cases (Allison, 2002). In turn, we conjecture that data wrangling practices are necessary, take a huge amount of time, and carry in it whatever critical understandings of data, measures, and methods that youth have or develop during these activities.

Overall, critical perspectives towards the relationship between participants' experiences and larger social, economic, and historical issues or reflections on power were present in conversations but were generally tangential to final family data storylines; race relations in the South and North came up inconsistently, as did mentioning of family history in which family members were disenfranchised. While there are several aspects of the design that could account for this, which are discussed in further detail in Kahn & Hall (in preparation), it is important to note that the difficulty with incorporating social critiques in such activities may be, at least in part, attributed to the data itself: Despite their capacity for modeling social conditions, big data, including census data, mostly mute processes of discrimination and assume categories like race are static (Philip et al., 2016; Radinsky et al., 2014). This is a challenge that deserves further study and consideration for future designs for learning.

Importantly, we also found that the workshop activities placed youth in authoritative and agentic roles to ask and answer questions about society and their families. These conversations with family members, as evidenced in the follow-up interviews, led to data explorations and sharing stories about family history that described local injustices and demonstrated family members' success in challenging inequitable distributions of power. In this way, we believe that integrating personal experiences with large-scale, aggregate data that indicate broad social-scientific conditions can seed important critical and relational thinking about the self and society.

As in any design study, our design contained many tradeoffs. Students were with us a relatively short period of time; that meant that other aspects of critical data literacy were not addressed exhaustively and remained a persistent challenge, like the differences between counts and percentages. (Notably, Radinksy et al., 2014 describes the choropleth webmaps' display of proportion as a misleading aspect of the Social Explorer representational platform.) On the other hand, we designed a program that can be reproduced or "popped up" in this local library or a similar public community space. The decision to design for spaces for learning outside of school was intentional, due to the growing role of libraries in community responses to social change.

While the design could be adapted to work in schools, the current instructional climate in public education (e.g., the required adherence to curriculum standards, emphasis on testing, and disciplinary silos) is not entirely amenable to broadening what counts as acceptable data modeling activity. Additionally, the design was limited by and benefited from using free, open datasets and tools as opposed to preselected data loaded into researchers' proprietary software.

Teaching youth to tell stories with data was not without its challenges, as youth have the tendency to over rely on the perception that data is transparent and self-evident in their oral presentations of data displays (Enyedy & Mukhopadhyay, 2007; Sandoval & Millwood, 2005). However, we found that youth data explorations consistently generated new questions both to be pursued in the tool and with family members. Moreover, tracking family spatial trajectories in relation to national global events and patterns of migration encouraged telling narratives that carried reflections on the participants' relationships with the place in which they live and that expressed their values and vision for what they want that relationship to be in the future (Basso, 1984). Through storytelling and modeling with big data, youth have the opportunity to recontextualize the moral dimensions of personal life—such as those surrounding family decisions or their own future aspirations to stay or move—within models built from large-scale data. This activity, we believe, avoids thin descriptions of place that the data alone might produce (Porter, 2012) and instead yields a rich, "global sense of place" that captures people's multiple identities, their social relations, global and local connections, and their mobility histories (Massey, 1994).

Conclusion

This review described a new data-modeling world that is opening its doors to the public. Adults and youth alike can now have access to dynamic data visualization tools and large-scale data for modeling scientific and social phenomena. However, LSDS are political, social, and complex. Thus, big data require a critical approach when used to read and write the world (Gutstein, 2003). Further research is needed to develop taxonomies of the kinds of analyses that are possible with LSDS and new modeling tools as well as to advance concepts for understanding large-scale data that will support civic identities and democratic participation; educational designs that position youth as historical, social actors of change by integrating individual experiences with aggregate data offer one direction for this research. Moreover, if society's data modelers and storytellers hold onto a moral framework when they engage with big data, then all science will not become data science; instead, big data can help inform their thinking about, and understanding of, both shared and personal human experiences.

REFERENCES

Al-Aziz, J., Christou, N. and Dinov, I.D. (2010). SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *Journal of Statistics Education, 18*(3), 1-29.

Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 193-196.

Anderson, C. (2008, June 28). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine.* Retrieved from http://www.wired.com/

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools. Utrecht: CDß-Press.

Bakker, A., & Gravemeijer, K. P. (2004). Learning to reason about distribution. In D. Ben-Zvi and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147-168). Netherlands: Springer

Bamberg, M. G. (1997). Positioning between structure and performance. *Journal of Narrative and Life History*, *7*(1-4), 335-342.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*,*94*(6), 1008-1026.

Basso, K. H. (1984). "Stalking with Stories": Names, places, and moral narratives among the western Apache. In S. Plattner, & Bruner, E. M. (Eds.), *Text, play, and story: The construction and reconstruction of self and society (*pp. 41-52). American Ethnological Society.

Becker, H. S. (2007). *Telling about society*. University of Chicago Press.

Biermann, F., & Boas, I. (2010). Preparing for a warmer world: Towards a global governance system to protect climate refugees. *Global environmental politics*, *10*(1), 60-88.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.

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.

Brantlinger, A. (2013). Between politics and equations teaching critical mathematics in a remedial secondary classroom. *American Educational Research Journal*, *50*(5), 1050-1080.

Brantlinger, A. (2014). Critical mathematics discourse in a high school classroom: Examining patterns of student engagement and resistance. *Educational Studies in Mathematics*, *85*(2), 201-220.

Buckland, Michael K. (2011). Data Management as Bibliography. *Bulletin of the American Society for Information Science and Technology, 37*(6), 34–37.

Busch, L. (2011). *Standards: Recipes for reality*. Mit Press.

Busch, L. (2014). A dozen ways to get lost in translation: Inherent challenges in large-scale data sets. *International Journal of Communication*, *8*, 18.

Carrigan, M. (2014, August 9). The origins of methodological genocide: "All science is becoming data science" [Web log post]. Retrieved from http://sociologicalimagination.org/archives/15816

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, *36*(4).

Cicourel, A. V. (1981). Notes on the integration of micro-and macro-levels of analysis. In K. Knorr-Cetina, & A. V. Cicourel (Eds.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies* (pp. 51-80). NY, NY: Routledge.

Cobb, G. W., & Moore, D. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, 104, 801-823.

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

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*(1), 1-78.

Coleman, G. (2004). The political agnosticism of free and open source software and the inadvertent politics of contrast. *Anthropological Quarterly*,*77*(3), 507-519.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, *18*, 32-42.

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, *92*, 28-40.

Daileda, C. (2016, May 19). Tech is dominated by even more white dudes than the rest of the private sector. *Mashable*, *Inc*. Retrieved from: http://mashable.com/2016/05/19/diversity-report-silicon-valley-white-men/#aXiwvyehCuqw

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to "Big Data." *Society and Space open site*.

Davidian, M., & Louis, T. A. (2012). Why statistics?. *Science*, *336*(6077), 12-12.

Derman, E. (2012). Apologia Pro Vita Sua. *The Journal of Derivatives*, *20*(1), 35-37.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educational process.* Lexington, MA: Heath.

Donovan, K. P. (2012). Seeing like a slum: Towards open, deliberative development. *Georgetown Journal of International Affairs*, *13*, 97.

Duncan, S. L. (2006). Mapping whose reality? Geographic information systems (GIS) and "wild science." *Public Understanding of Science*, *15*(4), 411-434.

Economist, T. (2010, February 25). The data deluge. *The Economist. Special Supplement*. Retrieved from: http://www.economist.com/

Edelson, D. (2012). GIS, Education and Citizen Science. *Global Dialogue (Online)*, *14*(1), 41.

Edelson, D. C., Gordin, D. N., & Pea, R. D. Pea. (1999). Addressing the challenges of inquiry based learning through technology and curriculum design. *The Journal of the Learning Sciences 8(3-4), pp.391-450.*

Ehret, C., & Hollett, T. (2013). (Re) placing school: Middle school students' countermobilities while composing with ipods. *Journal of Adolescent & Adult Literacy*, *57*(2), 110-119.

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review*, *5*(1), 1-24.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, *20*(4), 399-483.

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.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Freire, P. (2001). *Pedagogy of the oppressed*. New York, NY: Continuum International Publishing Group Inc.

Fujii, T., & Iitaka, S. (Eds.). (2013). *Mathematics International, Grade 5 (English translation)*. Tokyo: Tokyo Shoseki.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372-396.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.

Geertz, C. (1973). *The interpretation of cultures: Selected essays* (Vol. 5019). Basic books.

Giere, R. N. (1990). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Giere, R. N. (2002). Discussion note: Distributed cognition in epistemic cultures. *Philosophy of Science*, *69*(4), 637-644.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53-96.

Giroux, H. A. (2001). *Theory and resistance in education: Towards a pedagogy for the opposition*. Westport, CT: Bergin & Garvey.

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

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606-633.

Gordin, D. N., Polman, J. L., & Pea, R. D. (1994). The Climate Visualizer: Sense-making through scientific visualization. *Journal of Science Education and Technology*, *3*(4), 203-226.

Gould, R., Davis, G., Patel, R., & Esfandiari, M. (2010). Enhancing conceptual understanding with data driven labs. In *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics, Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute*.

Gravemeijer, K. E. P. (1994). *Developing realist mathematics education.* Utrecht, The Netherlands: CDBeta Press.

Gruenewald, D. A. (2003). The best of both worlds: A critical pedagogy of place. *Educational Researcher*, *32*(4), 3-12.

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday* 16(2). http://firstmonday.org/

Gutiérrez, K. D., & Jurow, A. S. (2016). Social design experiments: Toward equity by design. *Journal of the Learning Sciences*, *25*(4), 565-598.

Gutiérrez, K. D., & Vossoughi, S. (2010). Lifting off the ground to return anew: Mediated praxis, transformative learning, and social design experiments. *Journal of Teacher Education*.

Gutstein, E. (2003). Teaching and learning mathematics for social justice in an urban, Latino school. *Journal for Research in Mathematics Education*, 37-73.

Gutstein, E. (2006). *Reading and writing the world with mathematics: Toward a pedagogy for social justice*. NY, NY: Routledge.

Hacklay, M. M. (2013). Neogeography and the delusion of democratisation. *Environment and Planning A*, *45*(1), 55-69.

Hall, R. (1996). Representation as shared activity: Situated cognition and Dewey's cartography of experience. *The Journal of the Learning Sciences*,*5*(3), 209-238.

Hall, R. (2000). Work at the interface between representing and represented worlds in middle school mathematics design projects. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 675–680). Mahwah, NJ: Erlbaum.

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

Hall, R., & Horn, I. (2012). Talk and conceptual change at work: Adequate representation and epistemic stance in a comparative analysis of statistical consulting and teacher workgroups. *Mind, Culture, and Activity*, *19*(3), 240-258.

Hall, R., & Jurow, A. S. (2015). Changing concepts in activity: Descriptive and design studies of consequential learning in conceptual practices. *Educational Psychologist*, *50*(3), 173-189.

Hall, R., Stevens, R., & Torralba, T. (2002). Disrupting representational infrastructure in conversations across disciplines. *Mind, Culture, and Activity*, *9*(3), 179-210.

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

Hammerman, J. K. (2009). Statistics education on the sly: Exploring large scientific data sets as an entrée to statistical ideas in secondary schools. *IASE/ISI Satellite*.

Hess, D. J. (2005). Technology-and product-oriented movements: Approximating social movement studies and science and technology studies. *Science, Technology & Human Values*, *30*(4), 515-535.

Hess, D. J. (2011). To tell the truth: on scientific counterpublics. *Public Understanding of Science*, *20*(5), 627-641.

Howard, T. (2008). Who really cares? The disenfranchisement of African American males in preK-12 schools: A critical race theory perspective. *The Teachers College Record*, *110*(5), 954-985.

Howard, T. C. (2010). *Why race and culture matter: Closing the achievement gap in American classrooms*. New York, NY: Teachers College Press.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT press.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 375-398.) NY, NY: Berg.

International Telecommunication Union. (2016). *Measuring the information society report 2016*. Retrieved from: http://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2016/MISR2016-w4.pdf

Johansson, V. (2012). A time and place for everything?: Social visualisation tools and critical literacies. The Swedish School of Library and Information Science: The University of Borås.

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*(4), 263-274.

Joiner, J. (2014, October 16). Who does big data think you are? *Esquire*. Retrieved from: http://www.esquire.com/news-politics/news/a30534/this-is-who-they-say-you-are/

Kahn, J. (2014). "What in the world?" Animated worlds in multivariable modeling with motion chart graph arguments. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. Lee, & L. D'Amico (Eds.), Learning and Becoming in Practice: Proceedings of the International Conference of 11th International Conference of the Learning Sciences (pp. 1649-1650). Boulder, CO: International Society of the Learning Sciences.

Kahn, J. B., Hall, R., & Phillips, N. (2014). *Dissecting, remixing, and making graph arguments using motion charts and public data about global wealth and health*. Paper presented at the meeting of the American Educational Research Association, Philadelphia, PA.

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..

Kahn, J., & Hall, R. (in preparation). Getting personal with big data: The assembly of family data storylines.

Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. *Journal of the American Society for Information Science and Technology*, *64*(12), 2452-2467.

Klein, H. K., & Kleinmann, D. L. (2002). The social construction of technology: Structural considerations. Science, Technology and Human Values, 27(1), 28–52.

Knorr Cetina, K. D. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge: Harvard University Press.

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W.G. Martin, & D.E. Schifter (Eds.), A research companion to principles and standards for school mathematics (pp.193-215). Reston, VA: National Council of Teachers of Mathematics (NCTM).

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. *Handbook of International Research in Mathematics Education*, *2*, 49-71.

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

Kraemer, K.L., Dickhoven, S., Tierney S.F., & King, J.L. (1987). *Datawars: The politics of modeling in federal policymaking*. Columbia University Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4), 674.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. Academic Press.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago: University of Chicago Press. (Original work published 1962).

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491.

Ladson-Billings, G., & Tate, B. (1995). Toward a critical race theory of education. *Teachers College Record, 97*, 47–67.

LaFraniere, S. & Lehren, A. W. (2015). The disproportionate risks of driving while black. *The New York Times.* Retrieved from: http://www.nytimes.com/2015/10/25/us/racial-disparity-traffic-stops-driving-black.html?_r=0

Lamb, G. R., Polman, J. L., Newman, A., & Smith, C. G. (2014). Science news infographics: Teaching students to gather, interpret, and present information graphically. *The Science Teacher*, *81*(3), 29.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge: Harvard University Press.

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

Latour, B. (2005). *Reassembling the social*. Oxford: Oxford University Press.

Latour, B., & Woolgar, S. (2013). *Laboratory life: The construction of scientific facts*. Princeton University Press.

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

Lave, J., Murtaugh, M., & de la Rocha, O. (1984). The dialectics of arithmetic in grocery shopping. In B. Rogoff and J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 67–94). Cambridge, UK: Cambridge University Press.

Law, J. (2004). *After method: Mess in social science research*. Routledge.

Leander, K. M., Phillips, N. C., & Taylor, K. H. (2010). The changing social spaces of learning: Mapping new mobilities. *Review of Research in Education*,*34*(1), 329-394.

Lee, V. R. (2013). The Quantified Self (QS) movement and some emerging opportunities for the educational technology field. *Educational Technology*, (November-December 2013), 39.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116-133.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction*, *14*(1), 69-108.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal*, *41*(3), 635-679.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education*, *96*(4), 701-724.

Lehrer, R., Carpenter, S., Schauble, L., & Putz, A. (2000). Designing classrooms that support inquiry. In J. Minstrell & E. Van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 80-99). Reston, VA: American Association for the Advancement of Science.

Lehrer, R., Kim, M. J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM*, *43*(5), 723-736.

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

Lehrer, R., Schauble, L., Carpenter, S. & Penner, D. (2000). The inter-related development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing, mathematizing, and communicating: Perspectives on discourse, tools, and instructional design* (pp. 325-360). Mahwah, NJ: Lawrence Erlbaum Publishers.

Lesh, R., Cramer, K., Doerr, H. M., Post, T., & Zawojewski, J. S. (2003). Model development sequences. In R. Lesh & H. M. Doerr (eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning and teaching.* (pp. 35-58). Mahwah, NJ: Lawrence Erlbaum Publishers.

Ma, J. (2016, April). *Discussant notes*. In R. Nemirovsky (Chair), *Broadening what counts as mathematics in mathematics education.* Symposium conducted at the American Education Research Association 2016 Annual Meeting, Washington D.C..

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education*, *96*(6), 1071-1105.

Manz, E. (2014). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, 0034654314558490.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., ... & Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*(6), 957-958.

Massey, D. (1994). *Space, place and gender*. Minneapolis: University of Minnesota Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Moll, L. C., & González, N. (2004). Engaging life: A funds of knowledge approach to multicultural education. *Handbook of Research on Multicultural Education*, *2*, 699-715.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into practice*, *31*(2), 132-141.

Moore, K. C., & Carlson, M. P. (2012). Students' images of problem contexts when solving applied problems. *The Journal of Mathematical Behavior, 31*, 48-59. doi: 10.1016/j.jmathb.2011.09.001

Musgrave, S., & Thompson, P. W. (2014). Function notation as idiom. In P. Liljedahl & C. C. Nicol (Eds.), *Proceedings of the 38th Meeting of the International Group for the Psychology of Mathematics Education*, (Vol 4, pp. 281-288). Vancouver, BC: PME. Retrieved from http://bit.ly/1p08TCG.

Nafus, D. (2016). *Quantified: Biosensing technologies in everyday life*. MIT Press.

National Center for Education Statistics. (2015). *Integrated Postsecondary Education Data System* [Data File]. Retrieved from https://datausa.io/profile/cip/11/#demographics.

Noss, R., Bakker, A., Hoyles, C., & Kent, P. (2007). Situating graphs as workplace knowledge. *Educational Studies in Mathematics*, *65*(3), 367-384.

Ochs, E., Gonzales, P. & Jacoby, S. (1996). "When I come down I'm in the domain state": Grammar and graphic representation in the interpretive activity of physicists. In E. Ochs, E. Schegloff & S. Thomson (Eds.), *Interaction and grammar*. Cambridge: 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.

Phillips, N. C. (2013). *Investigating adolescents' interpretations and productions of thematic maps and map argument performances in the media* (Doctoral dissertation). Retrieved from etd.library.vanderbilt.edu.

Pickles, J. (Ed.). (1995). *Ground truth: The social implications of geographic information systems*. Guilford Press.

Pickles, J. (2006). Ground Truth 1995–2005. *Transactions in GIS*, *10*(5), 763-772.

Polman, J. L, Gebre, E. H., Rubin, A., Hinojosa, L., Sommer, S., & Graville, C. (2016, April). *Organizing data journalism activity in school and community learning environments to contextualize science in life*. Poster session at the American Education Research Association 2016 Annual Meeting, Washington D.C..

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.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Porter, T. M. (2012). Thin description: Surface and depth in science and science studies. *Osiris*, *27*(1), 209-226.

Radinsky, J. (2008). Students' roles in group-work with visual data: A site of science learning. *Cognition and Instruction*, *26*(2), 145-194.

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

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, *13*(3), 273-304.

Roberts, J., Lyons, L., & Radinsky, J. (2013). Become One With the Data: Technological Support of Shared Exploration of Data in Informal Settings. In Anne Knowles (Chair), *From Visualizing to Understanding Historical Change: Using GIS Tools on the Web, in Class, and in Museums. Paper session conducted at the meeting of the Social Science History Association, Chicago, IL*.

Robinson, D. G., Yu, H., Zeller, W. P., & Felten, E. W. (2009). Government data and the invisible hand. *Yale Journal of Law & Technology*, *11*, 160.

Rogers, R. (2013). *Digital methods*. MIT press.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford University Press.

Rose, T. (2016, February 8). How the idea of a 'normal' person got invented. *The Atlantic.* Retrieved from: http://www.theatlantic.com/business/archive/2016/02/the-invention-of-the-normal-person/463365/

Rosling, H., Ronnlund, A.R. & Rosling, O. (2005). New software brings statistics beyond the eye. In E. Giovannini (Ed.), *Statistics, knowledge and policy: Key indicators to inform decision making*, (pp. 522-530). Organization for Economic Co-Operation and Development.

Rubel, L. H., Hall-Wieckert, M., & Lim, V. Y. (2017). Making Space for Place: The Role of Mapping Tools in Learning as Political Formation. *Journal of the Learning Sciences*, (just-accepted).

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.

Rubin, A., Hammerman, J., & Konold, C. (2006, July). Exploring informal inference with interactive visualization software. In *Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town, South Africa: International Association for Statistics Education.* Online: www. stat. auckland. ac. nz/~ iase/publications.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23-55.

Saxe, G. B. (1988). Candy selling and math learning. *Educational Researcher*, *17*(6), 14-21.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Solórzano, D. G., & Yosso, T. J. (2002). Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative Inquiry*, *8*(1), 23-44.

Star, S. L. (1983). Simplification in scientific work: An example from neuroscience research. *Social Studies of Science*, *13*(2), 205-228.

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

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 85-87.

Suddaby, R. (2014). Indigenous management theory: Why management theory is under attack (and what we can do to fix it). In J. A. Miles (Ed.), *New directions in management and organization theory* (pp. 457-468). Cambridge Scholars Publishing.

Taylor, K. H., & Hall, R. (2013). Counter-mapping the neighborhood on bicycles: Mobilizing youth to reimagine the city. *Technology, Knowledge and Learning*, *18*(1-2), 65-93.

Tufte, E. R. (2001). *The visual display of quantitative information*. *Second edition.* Cheshire, Connecticut: Graphics Press.

Turner, E. E., Gutiérrez, M. V., Simic-Muller, K., & Díez-Palomar, J. (2009). "Everything is math in the whole world": Integrating critical and community knowledge in authentic mathematical investigations with elementary Latina/o students. *Mathematical Thinking and Learning*, *11*(3), 136-157.

Uprichard, E. (2013). Focus: Big data, little questions? *Discover Society*, (1).

Van Wart, S. Lanoutte, K., & Parikh, T. (2016). *Local Ground: Supporting data-driven inquiry with youth*. Poster session at the American Education Research Association 2016 Annual Meeting, Washington D.C..

Van Wart, S., Tsai, K. J., & Parikh, T. (2010, December). Local ground: A paper-based toolkit for documenting local geo-spatial knowledge. In *Proceedings of the First ACM Symposium on Computing for Development* (p. 11). ACM.

Venkatraman, V. (2013, May 13). When all science becomes data science. *Science*. Retrieved from: http://www.sciencemag.org/careers/2013/05/when-all-science-becomes-data-science

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.

Wager, A. A. (2012). Incorporating out-of-school mathematics: From cultural context to embedded practice. *Journal of Mathematics Teacher Education*,*15*(1), 9-23.

Wake, G. (2014). Making sense of and with mathematics: The interface between academic mathematics and mathematics in practice. *Educational Studies in Mathematics*, *86*(2), 271-290.

Wertsch, J.V. (1998). *Mind as action*. Oxford University Press.

Wilkerson-Jerde, M., & Laina, V. (April, 2015). *Stories of our city: Coordinating youths' mathematical, representational, and community knowledge through data visualization*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wilson, M. W. (2015). Morgan Freeman is dead and other big data stories. *Cultural Geographies*, *22*(2), 345-349.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121-136.

Winner, L. (1997). Cyberlibertarian myths and the prospects for community. *ACM Sigcas Computers and Society*, *27*(3), 14-19.

Wisnieski, A. (2015, December 15). NYC’s open data law lacks teeth, lags deadlines. Retrieved from: http://citylimits.org/2015/12/15/nycs-open-data-law-lacks-teeth-lags-deadlines/

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., & Heer, J. (2016). Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *Visualization and Computer Graphics, IEEE Transactions on*, *22*(1), 649-658.

CHAPTER III

STORYTELLING WITH BIG DATA: MULTIVARIABLE MODELING OF GLOBAL HEALTH AND WEALTH

Introduction

"Does your mindset correspond to my dataset? If not, one or the other needs upgrading, [does]n't it?"

—Swedish global health expert and data statistician Hans Rosling giving TED Talk at the US State Department (Rosling, 2009)

The recent availability of large-scale data sets (LSDS), also known as "big data," and data visualization tools has provided opportunities for powerful storytelling about society (Kosara & Mackinlay, 2013). LSDS refer to quantitative datasets that are sufficiently large and require a computer for processing (i.e., data preparation, selection, and cleaning), analysis (e.g., data mining, the application of algorithms and statistical techniques for extracting patterns from data; Fayyad, Piatetsky-Shapiro, & Smyth, 1996), and storage (Busch, 2014). State and local governments, international agencies, research institutions, and businesses alike are collecting and aggregating LSDS about citizens and consumers, some of which are used to inform public policy and decision-making (Busch, 2014; Davidian & Louis, 2012). LSDS now commonly appear in public media and discourse, targeting youth and adults alike. Quantitative data, to some extent, has always motivated policy and industry decisions (Kraemer, Dickhoven, Tierney, & King, 1987; Porter, 1996) but not in such large volumes. In turn, large-scale data collection and aggregation raise important questions about the epistemological, ontological, political, and ethical consequences of data use for learning and telling about society (Carrigan, 2014;

Uprichard, 2013).

This manuscript presents part of a multiyear design-based research study (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) that engaged prospective secondary mathematics and social studies teachers (separately) in multivariable modeling practices using socioeconomic large-scale data. Our instructional goal was to capitalize on the availability of big data and interactive digital visualization tools to open possibilities in learning and teaching across math and social science subject-domains. We sought to expand upon school-valued disciplinary goals by familiarizing preservice teachers with new forms of data representation that are commonly found in public media and professional analytical fields (Kosara & Mackinlay, 2013) and serve as tools for democratic and civic participation (Gutstein, 2006; Philip, Schuler-Brown, & Way, 2013). Likewise, our experimental teaching engaged university students' in the animation of "messy" big data and invited critical perspectives towards models of global development.

We frame our findings as *ontological discoveries* that were consequential for participant learning (diSessa & Cobb, 2004) and thus informed our design choices for subsequent iterations. Namely, across iterations, we found that to assemble models with open socioeconomic big data was challenging but rich for learning. Our design invited novel modeling practices that involved complex forms of temporal, spatial, and social *scaling* (Hall & Leander, 2010). In certain cases, participants became *data stakeholders*—they perceived the data as consequential for them and themselves as consequential to the data. In such cases, when participants related their personal experiences to patterns found in the aggregate data, compelling counter-narratives, stories that challenge normative power relations or dominant sociocultural and socioeconomic discourses (Solórzano & Yosso, 2002) emerged.

This report centers on a close analysis of case studies drawn from an observational study of a professional big data storyteller and public health professor, Hans Rosling, and from two (separate) design studies of "newcomers"—undergraduate and graduate preservice secondary mathematics and social studies teachers—to storytelling and modeling with big data. Both Hans and the preservice teachers told stories about the world with a multivariable, dynamic modeling tool that accesses open, global socioeconomic big data. We identify complex modeling practices for assembling models and telling stories with big data across cases.

The Rise of Open Big Data and Visualization Tools

In general, there are two types of LSDS: user-generated records of an individual's experiences, online activity, and transactions, also called *digital traces* (Karanasois et al., 2013; Rogers, 2013; Venturini, Jensen, & Latour, 2015), and data that is gathered by state governments to describe social trends or phenomena (Busch, 2014). While much of the data remains proprietary, the development of cyber and digital infrastructure has supported the availability of aggregate data for public inspection and interpretation. Over the past decade, organizations and Western government bodies (e.g., World Health Organization, U.S. State Department) have led this movement by releasing national data and reports on the Internet to expand government transparency, efficiency, and effectiveness (Cukier & Mayer-Schönberger, 2013). Consequently, there has been a surge in public access to LSDS (Hammerman & TERC, 2009), opening doors for new forms of democratic participation (Philip et al., 2013). The *open data movement* is thus transforming 21st century society (Cukier & Mayer-Schönberger, 2013) by making local, regional and national data freely available to citizens in digital forms that allow for manipulation and modeling (Gurstein, 2011).

However, open government datasets are neither politically neutral nor a guarantor of democratization (Hacklay, 2013; Johnson, 2014; Winner, 1980). While transparency is important, access alone is not sufficient for social and political change (Donovan, 2012). The majority of lay citizens (and especially youth) still have only limited access to Internet and open data in cleaned, standardized formats (e.g., CSV, HTML; Mayer-Schönberger & Cukier, 2013), and variable access to LSDS can create power asymmetries between companies, governments, and citizens (Dalton & Thatcher, 2015). Furthermore, using open data to initiate social change—to bring resources to underserved populations or to hold governments accountable for their decisions or uses of funds (e.g., the Trends case in Johansson, 2012)—requires becoming professionally and technically skilled in "hacking," coding, or programming. Most lay citizens are only able to leverage LSDS in a limited way without specialized training and advanced levels of education and technical support (Hacklay, 2013). Furthermore, those with a computational background tend to be white and male (boyd & Crawford, 2012). In turn, the open data movement has benefited private upper and middle-class individuals already empowered rather than historically marginalized populations or groups struggling socioeconomically (Johnson, 2014; Donovan, 2012).

Despite any challenges for leveraging big data for democratic ends, the rise of open source software, open tools, and open LSDS data has benefited science and social science research and enabled new kinds of modeling that were previously cost prohibited (boyd & Crawford, 2012; Venturini et al., 2015). The scale and distribution of computation, cloud storage and processing, and statistical tools have grown to match the massive data sets they compute (Anderson, 2008), giving birth to novel modeling tools that permit nearly instant, dynamic, and often interactive visualization to support analysis (Al-Aziz, Christou, & Dinov, 2010; Johansson, 2012). Today, visualizations made of large-scale data are ubiquitous in the media, professional fields, and state policy arenas, often used in service of political arguments or narratives aimed at adults and youth. These models take the form of quantitative thematic maps, time-series charts, relational graphs like scatter plots, (Tufte, 1983) or multivariate, digital amalgams of data visualization, like the motion chart (Figure 1). Internet megaliths (e.g., Google), media companies (e.g., New York Times), not-for-profit organizations (e.g., ESRI, Gapminder, Social Explorer), and public entities (e.g., U.S. Census) also host some of these mapping and statistical graphical tools without cost.

Models created with LSDS ask and answer questions and support telling narratives about phenomena in the world (Kosara & Mackinlay, 2013). Models built using LSDS can be predictive (Giere, 1990) but also historical and explanatory; models are always subject to judgments of their perceived fit with the world they describe (Giere, 1990; Kuhn, 1962/2012; Manz, 2014). In the cases of LSDS, the data would be incoherent without effective modeling and visualization tools. Without strong modeling and visualization, we are lost in a "sea of data" (Latour, 1999, p. 39).[8] Nonetheless, data modelers, users, and analysts should proceed with caution in using LSDS (boyd & Crawford, 2012; Wilson, 2015). Models of large-scale data are often statistical, but LSDS are neither populations nor samples as typically defined in statistics textbooks (Busch, 2014). While the "N" (sample size) appears to and in some case actually does approach all of the population (Cukier & Mayer-Schönberger, 2013), the size of LSDS does not mean that the models display universal "Truths."

Indeed, large-scale data do not only describe worldly phenomena but they are also cultural, technological, and scholarly phenomenon to be understood and used cautiously and critically (boyd & Crawford, 2012; Dalton & Thatcher, 2015; Johnson, 2014; Wilson, 2015).

[8] Some claim that the "distance" (from observations of worldly phenomena described) offers data modelers a more objective perspective, although many critical theorists of technology argue any such claim of LSDS' objectivity is just pretense (e.g., boyd & Crawford, 2012; Porter, 1996).

Thus there are consequences for learning too (boyd & Crawford, 2012), although research has yet to elaborate on the implications for education beyond a "functional," anti-Freirean (Gutstein, 2006) call on our educational system to produce data-savvy and statistically literate analysts for the 21st century workforce (Davidian & Louis, 2012). Moreover, in the context of this "data deluge" (The Economist, 2010), we seek to understand the ways in which LSDS can transform learning and education.

Data Modeling in Schools and Professional STEM Contexts

Our design for learning with big data stems from an interest in modeling practices with novel representational forms. Likewise, this paper builds on and moves beyond current reports on data modeling and modeling practices in K–12 and Learning Sciences research. Our research draws from science, technology, engineering, and mathematics (STEM) education design-based research activities (Cobb, McClain, & Gravemeijer, 2003; Konold & Lehrer, 2008; Lehrer, Kim, & Schauble, 2007; Lehrer & Schauble, 2004) that encourage students' statistical reasoning about variation, distributions, and representations of data. In contrast with the distributed processes of collecting and analyzing LSDS (Strasser, 2012), the volume of the data collected in these studies is mostly manageable without computers, data mining tools, or cloud-based storage. Rather, the construction of data models typically stays close to the data collection process. Students often analyze classroom-scaled data such as the data that children themselves collect. For example, youth participants engage in rounds of hands-on scientific investigation of phenomena inside classrooms (e.g., teacher arm-span; Lehrer et al., 2007) or nearby outside spaces (e.g., plant growth in the classroom or school backyard; Lehrer & Schauble, 2004; Lehrer, Schauble, Carpenter, & Penner, 2000; Manz, 2012). These studies engage with processes of repeated measure and production and aim to support students' coordination between properties of class-aggregated data and individual cases in terms of understanding distributions. Other studies (e.g., Bakker, 2004; Cobb & Moore, 1997; Enyedy & Mukhopadhyay, 2007; Garfield & Ben-Zvi, 2007) concentrate on developing students understanding and facility with descriptive statistics: applying measures of center—namely mean, median, and mode—to characterize a sample.

In the secondary education literature, while the number of cases making up a sample can be larger, the data typically do not qualify as LSDS (e.g., salary data, Cobb et al., 2003), and data modeling contexts are not contingent on any real sense of time, space, and human experience. Instead, data-laden word problems often describe covariation in which change over time is momentary (e.g., speed of a car going through a stoplight) or nonexistent (e.g., dimensions of a box increasing in size). LSDS, in contrast, describe vast scales and spans of measured time and space—and social life.

A significant chunk of the secondary education literature also advocates for instructional activity focused on statistical facility around reasoning and argumentation with data models and problems imitating "real world" applications (e.g., mushroom brush factories; Rubin, Hammerman, & Konold, 2006). A portion of this literature has focused on covariation with simulated data, or thinking relationally about how variables or quantities change continuously (e.g., ambulance arrival times in Cobb et al., 2003; the time it takes for a headache pill to take effect in Konold & Lehrer, 2008). We also see this in the elementary education literature, where researchers provide both real and fabricated datasets of naturally occurring phenomena such as bird wingspan or car battery life (Bakker & Gravemeijer, 2004; Appendix A Lehrer & Schauble, 2004). At best, these contexts cultivate student interests in statistical reasoning and support productive disciplinary engagement (Engle & Conant, 2002). However, the topics are generally only tenuously connected with youth's lives and experiences beyond the classroom.[9] Modeling activities that do not connect deeply with students' experiences can lead students to become engaged in answer-finding instead of meaningfully interpreting the data to create explanations for personally-relevant questions (Sandoval & Millwood, 2005).

Conversely, if the data were personally meaningful for the participants' lives, the data and analysis could inform their views of themselves as historical and social actors (Gutiérrez & Jurow, 2016), perhaps as agents who can enact broader social and cultural change by understanding processes represented in the data. However, the social production of the data models and measures are typically confined to the classroom community, its practices, and norms. In turn, the scale of individual agency is similarly bounded by students' decisions for what measures to use and how measurements are taken. This, of course, makes sense when the purpose of the design is to develop a community of data modelers who learn to make choices about data (what to measure), how to construct it (how much, from whom, when), how to organize and represent it (table structures, computational tools), and how to make inferences about the data. In contrast, when starting with state-collected, large-scale open data, the data and measures are already constructed (boyd & Crawford, 2012). Data modelers, in labs or in our designed learning environments, do not face decisions for how to collect or measure large-scale social, environmental, or economic processes. Yet they are in the position to examine social activities that have much larger scales (in time, space, and social life) and ask/answer questions about other places and times.

[9] Lehrer & Romberg, 1996 present an exception to a certain extent in that students in their study created data models from surveys they invented about their daily routines after learning about lifestyles in Colonial America. Whether the students felt that the data reflected themselves as historical, social actors was unclear. The focus of the class was to model and summarize trends across classmates.**)**

Furthermore, many of opportunities for building, evaluating, and refining models in such studies are not critically contextualized in relation to broader phenomena and scientific and socio-scientific issues. Some learning sciences research has intended to support the extension of lines of disciplinary inquiry into critical examinations of broader social issues but have encountered obstacles to productively connecting students' local or cultural experiences to aggregate data in rich disciplinary ways (Bang & Medin, 2010; Enyedy & Mukhopadhyay, 2007; Philip et al., 2013; Philip, Olivares-Pasillas, & Rocha, 2016; Rubel, Lim, Hall-Wieckert, & Sullivan, 2016; Wager, 2012). In particular, one of the struggles for design research that strives to be personally meaningful or culturally relevant and also achieve conventional STEM disciplinary standards is that under the latter framework, students' individual experiences are positioned as bias that interferes with "good" data modeling practice and limits critical interpretations of evidence of covariation. Kuhn et al. (1988) describe belief bias as the failure to differentiate theory and evidence, an individual "weakness" (p. 102), a "significant deficiency" in the coordination or alignment of theory and evidence (p. 112), which they found to be prominent among both youth and adults. Approaches that take up this position assume that objectivity is attainable as long as one is able to separate their experiences from their data interpretations.

However, we distinguish our approach to model creation and data interpretation from educational research that treats modeling as separated from or actively scrubbed free of values, bias, or social life. Rather, our framing of storytelling as a fundamental inquiry practice in STEM modeling recognizes the social construction of models and data (Goldstein & Hall, 2007; Kraemer et al., 1987; Kuhn, 1962/2012; Porter, 1996; 2012). In practice, scientists and data analysts bring personal viewpoints and individual field experiences to support model interpretations (Goldstein & Hall, 2007; Noss & Hoyles, 1996). For instance, Goldstein and Hall (2007) found that regulatory and field biologists often treated the same data about species-habitat relations in ways that reflected very different forms of disciplined perception or professional vision (Goodwin, 1994; Stevens & Hall, 1998).

Representational practices, professional and disciplined perception, and epistemic beliefs always influence how stakeholders perceive and interpret the measures, the resulting data, and the models created using the data (Goldstein & Hall, 2007; Goodwin, 1994; Hall & Horn, 2012; Stevens & Hall, 1998). STEM professionals operate diverse ways of "seeing" data—of seeing statistics as measures and evaluating models—through particular forms of talk, gesture, and tool use in order to tell stories about the world. These stories describe forms of influence and changes through time (i.e., covarying relationships between quantities) that are inaccessible except through specific courses of talk and action that animate models (Goldstein & Hall, 2007; Goodwin, 1994; Hall & Horn, 2012; Hall, Wright, & Wieckert, 2007; Stevens & Hall, 1999; Hall, Stevens, & Torralba, 2002).

A Framework for Storytelling and Modeling With Big Data

Storytelling with big data is an interactive event between presenter and audience (Bamberg, 1997); how participants orient to each other and to their models can be systematically studied and analyzed across contexts (Hall, 1999). For instance, stories told with big data, like other narratives, have definable syntactic structures. Applying Labov's (1972) approach to narrative structure in personal storytelling to our context, presenters first introduce their story and familiarize audiences to model elements (the abstract and orientation). They then explain and evaluate the narrative's "complicating action" that produces patterns in the data, such as historical actors and events that influence society's progress, before arriving at a result or resolution and a summative “so what?” statement (the coda).

Like Bamberg (1997), Becker (2007) argues that reports about society are defined by the presenter–audience interaction: Reports “exist fully” only when an audience is “constructing for themselves a reality” of the world from what was presented to them (Becker, 2007, p. 25). Becker gives examples of different genres of reports that “tell about society,” produced by a variety of storytelling professionals. Filmmakers, journalists, artists, social scientists, and novelists each presents a particular kind of social analysis through unique media. Similarly, we view big data visualization tools as a new genre of representational artifact (Johansson, 2012) for telling about society. We believe the population of storytellers and modelers who can produce these kinds of reports is broad; in our design work, it includes professional statisticians, newscasters, families, youth, as well as mathematics teachers and social studies teachers.

Becker (2007) describes the process of producing reports to tell a story about society, which we can apply to assembling models with big data: Individuals first select what to show and hide in a model (selection). Second, they use materials and language to describe the aspects of reality they want to convey; this translation activity is what Labov (1972) calls as the transformation of experience into language or narrative structures, described above. In our design studies, a large part of the translation work, which includes standardizing and cleaning the data into a readable format, has already been completed by the creators of the tool (Johansson, 2012). Consequently, these data processes remain mostly invisible for users and excluded from participant stories. Furthermore, translations vary among presenters depending on their particular disciplinary lens (e.g., mathematics and social studies educators). Next, presenters of stories about society arrange the elements of their model to support their narrative. Finally, storytellers and modelers analyze and interpret the meaning of the trends, patterns, or outliers in the data.

Ochs, Taylor, Rudolph, and Smith (1992) describe this last part, the performance of stories, as theory-building activity. Stories, they argue, are personal theories for how the world works. Stories contain inciting events (akin to Labov's [1972] complicating actions) to convey their theory to explain influence or cause (Hall, 1999) and, like theories, can be challenged by audiences or other narrators. Like theory-building, storytelling with big data is a conceptual practice in which an individual uses the animation of data to impress upon an audience a different way of seeing the world on the basis of the model. An approach to storytelling with visualization tools and open big data as theory-building activity focuses our attention dynamic modeling interfaces as new mediums for animating or making present (Hall, 1999) socioeconomic influences that produce trends and as new technical means for challenging stories. That is, if an individual can access and interact with the same data, they can "produce alternative analyses" that oppose, confirm, or complicate the story being told (Johansson, p. 169). In turn, storytelling and modeling with big data is a form of inquiry that supports model competition and does not claim simple truths.

While big data modeling tools can enrich storytelling as theory-building, stories are also necessary for leveraging the power of big data visualizations (Kosara & Mackinlay, 2013). Kosara and Mackinlay (2013) argue that stories are how data become memorable and meaningful. Furthermore, they suggest that new dynamic, interactive data visualization technologies have "storytelling affordances" to support narrative structure, particularly in their ability to represent time.

In summary, we see the relationship between telling stories and modeling with big data visualization tools as follows: We view modeling practices as purposeful efforts to explore the relations between representing and represented worlds (Gravemeijer, 1994; Hall, 2000). Every model must be interpreted, and interpretations of models are stories told with data. Different stories can be told with the same model and data, and one can change the model's variables or quantities to show patterns that support an alternative explanation or narrative. We consider *counter-modeling* to be when a data model tells a story that reveals and critiques social inequities, state or political actors, or unjust power relations. Counter-modeling draws on the Critical Race Theory concept of counter-storytelling, which describes telling stories of marginalized people's experiences that challenge dominant discourses, narratives, and racial privilege of the majority (Solórzano & Yosso, 2002). Counterstories (or counternarratives) serve as tools for examining and opposing master narratives of racial oppression and of white privilege (Howard, 2008); some of these narratives have been developed using large-scale data and models that promote claims about the biological and cultural deficits of historically-marginalized people (Solórzano & Yosso, 2002). In turn, telling counterstories that work against these claims strengthens the social, political, and cultural power of minority communities.

In line with other Learning Sciences research that seeks to give new representational tools and agency to youth and encourage their voices (see Taylor & Hall, 2013 for a design study on youth counter-mapping), we hope that counter-modeling will empower preservice teachers and their future students to assume critical perspectives with an eye towards participating and supporting local and broader social justice efforts. We view STEM school-based learning activities as opportunities to promote social justice, equity, and democracy (Enyedy & Mukhopadhyay, 2007; Gutstein, 2006; Philip et al., 2013). Our instructional design values critique, moral dispositions, and discourse as worthwhile outcomes of learning in STEM classrooms and as foundational for contributions to public and civic life outside of school (Berkowitz, Althof, Jones, 2008).

However, we propose a new framework for supporting critical literacy and inquiry with large-scale data and models. We suggest that becoming critical with large-scale data involves exploring lines of interest and inquiry that lead to the development of personal relationships with models, particularly if the modeling interface permits deep investigation of quantities, measures, and stakeholder perspectives. Getting personal with big data in this way, we contend, invites shifts in the learner's epistemic and moral stance toward models and social, historical, and scientific phenomena described by models. These shifts can support telling counter-narratives about society.

Methods

Participants

We report from two cycles of design-based research (Cobb et al., 2003) oriented toward supporting preservice teachers in telling stories with models and new modeling tools using LSDS. Both rounds took place at a college of education in a private university in the southern United States. The first study was with undergraduate and graduate secondary mathematics teacher candidates in a Mathematics Literacies course in the fall of 2012. All 13 students in the class participated in the study, with a total of 12 licensure (preservice teachers) students. Of the 12, five were Master's students, four in the licensure program, and one enrolled in a nonlicensure Master's programs. Nine were undergraduates, all of whom were licensure students.

The second study took place with undergraduate and graduate secondary social science teacher candidates in a Human Geography course in the fall of 2014. Fourteen of 15 students in the class participated in the study, with a total of seven licensure students. Of the 14, 11 were Master's students, five of whom were licensure students; the other six Master's students, one of whom already had a teaching license, were enrolled in nonlicensure Master's programs supporting careers in informal and formal educational settings such as social service agencies, nonprofit or for-profit organizations, or in academic or policy research. There were three undergraduates: two were licensure students, and one was in a creative writing program focused on spatial thinking.

As in most preservice teaching programs (Dedeoglu & Lamme, 2011; Gay & Howard, 2000; Lowenstein, 2009), students across both classes were generally white and female.

Design of the Study

With the second iteration, we sought to understand how the experiences of preservice social studies teachers with modeling LSDS would compare to that of preservice mathematics teachers. While design rounds varied slightly, both iterations spanned three 3-hour classes over 4 weeks in the middle of the semester and consisted of three primary tasks: (a) the dissection of STEM arguments from the public media, (b) assembling models with Gapminder© motion charts, and (c) class performances of data stories with their Gapminder models.

We introduced the research activities for both classes with Gutstein's (2006) critical literacies framework, which argues that mathematics and statistics should be used to examine sociopolitical and cultural-historical contexts and challenge existing power relations. For the dissection activity (Class 1), we asked students to "forage" for statistical or mathematical arguments or stories that interested them from online media and to share them in class in small groups on a laptop computer. At the end of Class 1, we introduced Gapminder as a big data tool used to tell STEM arguments and stories in the public media. Students watched a professional online TED Talk performance (conference presentations on Technology, Entertainment, and Design, freely available online) of Hans Rosling, Gapminder's creator, using the model to tell stories about world development. We then asked students to remix (manipulate the model to tell a different story) one of the arguments from Hans' performance using the graphing tool for homework. For the assembling activity (Classes 2 and 3), we asked all students, in small groups or pairs on computers, to construct models with Gapminder motion charts around a topic of interest that they would use to tell a compelling story about global socioeconomic data in front of their classmates. The performance of stories was the last segment of the design cycle (Class 3). We video and audio-recorded all three-class periods of the design studies.

We chose the Gapminder motion chart for our designed activities for its *modeling depth*. Modeling depth describes the extent to which a model can serve as an exploratory tool for generating questions and discussion about model assumptions, stakeholder perspectives, quantities, and measures. The modeling depth of the object depends on its capacity to allow alternative views of the data, multiple choices of measures, and various selections of quantities that influence the relation of interest. A "deep" modeling tool might also give access to raw data, data sources, and information about data quality and measure validity. In addition, the user interface for selecting and manipulating data strongly influences the depth of inquiry processes in telling stories with LSDS.

Accordingly, the motion chart is a dynamic, interactive, digital, statistical visualization tool that models multivariable data through time (Al-Aziz et al., 2010; Hammerman & TERC, 2009; Rosling, Ronnlund, & Rosling, 2005; Johansson, 2012). Motion charts have been popularized in Technology, Entertainment, and Design (TED) Talks by Hans Rosling, and are free to use through Google™ (with public data and one's own data) and on Gapminder.org. Gapminder.org uses large, global, socioeconomic data sets released by state and nongovernment organizations. Gapminder motion charts let users select multiple variables, define timescales, alternate between logarithmic and linear scales, and highlight particular country actors (see

Figure 1). Gapminder allows comparison for different scales or levels of analysis, such as across geographic regions or between countries (Johansson, 2012). We also chose Gapminder because we felt it affords the interdisciplinary use of "academic language" for STEM and social studies teaching (e.g., variation, important critiques of measures).

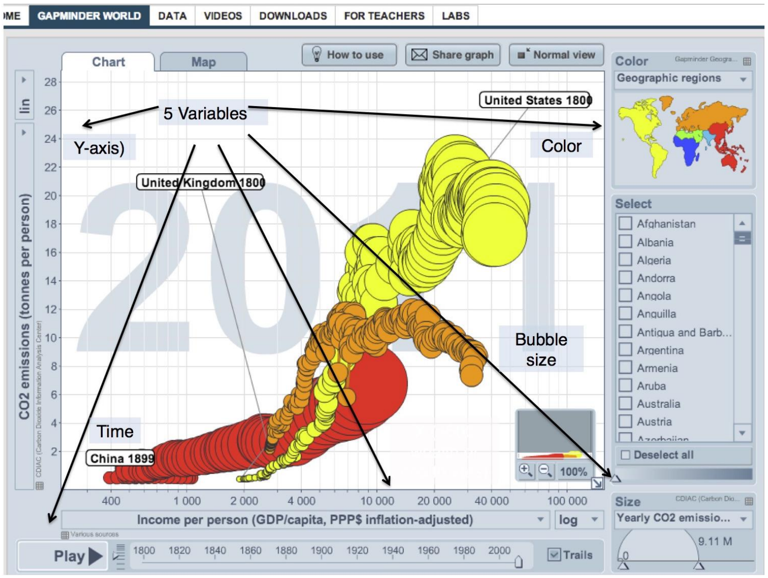


Figure 1. This is the Gapminder interface. Each model has five possible quantities or variables that can be selected. Y-axis, x-axis, color, bubble size each has over 500 health and wealth indicators (measures) to choose from. Timescales of datasets vary; some start as early as 1800, and others only have one year of available data.

From Johansson's (2012) ethnographic study of the Gapminder Foundation, we know that Hans Rosling and colleagues built the motion chart tool and assembled hundreds of LSDS in order to promote public access to global socioeconomic data collected by state governments and institutions and to ground both laypersons' and NGO professionals' questions and thinking about world development in data. The team behind Gapminder are firm believers that open data can support democratic transparency. They also sought to make information and statistics about the health and wealth of nations more compelling and manageable through the dynamic visualization tool. Hans demonstrates Gapminder's capacities by using the tool in public performances to let his "dataset change your mindset" (Rosling, 2009). While we acknowledge the possibility that very persuasive data visualizations could limit inquiry (Hans and colleagues were concerned about this too, Johansson, 2012), we felt that Hans' approach to modeling and storytelling with datasets "to change mindsets" encompasses our approach to telling stories with big data as theory-building activity for how the social world works.

The relation between telling stories and models built out of LSDS is bidirectional. Modelers can use LSDS to create a model that either confirms or disconfirms a particular mindset or story, as Hans does in his talks. On the other hand, model "behavior" during data exploration may produce stories as explanations of that behavior (an answer to the question, "Why did that happen?"). Johansson (2012) describes the use of Gapminder in this way by the Gapminder Foundation staff: When a Gapminder model plays over time, a country trail might elicit a response from a user—"'What was that?'"—which becomes a "start-point" for future questions (p. 185). In our designed activities, we also found this to be the case for student inquiry. In student pursuits of "interesting" models and stories for their final model performances*,* model behavior prompted questions (e.g., "What in the world?") that were then followed by efforts to learn more about a country's history or potential socioeconomic conditions or influences that could explain what they saw. While Hans' team was aware of the risk that nonprofessional users could misinterpret such dynamic model behavior, they viewed the potential benefits of public usership and open-data as overriding these concerns (Johansson, 2012).

In both design iterations, the research team and the course instructor codesigned activities, and course instructors situated activities in their appropriate disciplinary frameworks for mathematics and social studies teaching, respectively. In the Mathematics Literacies course, associated assignments and additional readings drew attention to mathematics "in the world" outside of school and teaching mathematical and statistical literacies (e.g., students read Gal's "Statistical Literacy: Meaning, Components, Responsibilities," 2004; Cobb & Moore's "Mathematics, Statistics, and Teaching," 1997). The Mathematics Literacies course instructor encouraged students to ask critical questions whenever they encountered mathematical, graphical, and statistical representations and models, such as: What does the data show and hide? What interesting stories are being told? What are the ways (i.e., highlighting, selection of variables, measures choices) in which an author presents an argument with statistics?

In the Human Geography class, the Gapminder unit was intended to motivate the Human Geography students to think about storytelling with data at a global scale, and the assigned readings shared this focus (e.g., students read Kosara & Mackinlay's "Storytelling: The next step for visualization," 2013). The research activities also culminated in a graded paper in which students used multiple data representations to tell historical, spatial narratives about places at local, regional, and global scales (also described for students as "macro" and "micro" stories). The Human Geography instructor pushed students to explore how representational forms influence stories told. Indeed, students from each class approached the activities with well-developed disciplinary perspectives and academic-disciplinary language. In both iterations, this became evident in the class discussions and questions following each Gapminder performance. While we address disciplinary differences here as they become relevant for the current analysis, a separate analysis is warranted to more fully examine how these disciplined perceptions (Stevens & Hall, 1998) among future teachers shaped their participation in the activities.

Qualitative Analysis

We sought to understand how participants told stories and assemble models about the social world with big data. In particular, we were interested in understanding when do participants engage in counter-modeling and how does scaling support telling critical stories about patterns in the aggregate data. Our research question for this analysis was as follows:

- How do participants tell stories about society with models made using LSDS?

To answer this question, we developed a comparative analysis of cases to identify and describe storytelling and modeling practices using the Gapminder LSDS and visualization tools.

We first analyzed a collection of Hans Rosling's video-recorded public modeling performances with Gapminder as a case study of professional STEM modeling with LSDS (Kahn, Hall, & Phillips, 2014). We were particularly interested in how Hans shifted between personal and aggregate (state) scales of time, space, and social life in making STEM arguments about world health and economic development, using his wildly popular public performances as a form of storytelling and modeling with big data by professionals "in the wild" (Hutchins, 1995). We conducted close analyses of three of Hans Rosling's most popular public TED Talks (each with millions of views). We selected these talks not only for their popularly, but also because they occurred at the time when Gapminder became available for public use; during this time, Hans also garnered international attention from state actors, developed a large following on the Internet, and won commitment from the UN to support access to open data and modeling tools. We coded Hans first TED talk in 2006 where he introduced Gapminder to the world ("The Best Stats You've Ever Seen," Rosling, 2006), his 2007 TED Talk when Gapminder had just became a completely online resource ("New Insights on Poverty," Rosling, 2007), and his 2009 address to the U.S. State Department ("Let My Dataset Change Your Mindset," Rosling, 2009), which was the talk assigned to the two classes.

For comparison, we turned to the preservice teachers' modeling performances hoping to see if the students animated their personal experiences in relation to a global data set in similar ways. More technically, for the purposes of design, we wanted to understand if and how particular forms of scaling supported the assembly of counter-models with Gapminder.

We drew from two student groups—one from each iteration—for selective case study comparing (Flyvbjerg, 2006). All three cases (including Hans and the two student groups) served as *paradigmatic* cases—exemplars for persuasive and complicated storytelling and modeling practices with big data. Furthermore, we selected the two student cases because they also share a dimension that makes them uniquely comparable across settings. Both groups presented counter-models around the same topic: the relationship between countries' economies and carbon dioxide (CO2) production. Both groups compared China and the United States on measures of economic wealth and CO2 emissions and demonstrated that the story of who produces more of the world's CO2 depends on how carbon emissions are measured (i.e., in yearly totals, a measure which China currently leads, or per person, a measure that the U.S. currently leads). In addition, both groups critiqued Gapminder's CO2 emission measures for not accounting for their (and their fellow Americans') consumption of imported Chinese goods. The focal Mathematics Literacies group consisted of Master's licensure students, Nathan, Nicole, and Yuri. The Human Geography course group consisted of Master's licensure students, Luke and Daphne, and the one undergraduate noneducation major, Carly.

Multiple interaction analysis sessions (Derry et al., 2010; Jordan & Henderson, 1995) were devoted to reviewing video recordings of multiple public presentations by Hans Rosling and of our focal cases across the study activities. While we only report here on students' final performances (Class 3), our video analysis followed the focal student groups across their dissection of foraged STEM arguments from public media sources, their exploration of Gapminder.org and construction of a model using those data and tools, and their performances of stories and models with Gapminder resources (Classes 1, 2, & 3). Analyzing how the two student groups collaboratively developed their stories and models about global development gave us deeper insight into their stories and models (Kahn, 2014).

For instance, both focal student groups had histories of pursuing environmental and economic narratives across activities. In the Mathematics Literacies class, Nathan and Nicole foraged for data stories related to the environment: They introduced a National Weather service chart presenting how quickly frostbite onsets at various temperatures and a report on energy use from the American Petroleum Institute, respectively. In the Human Geography class, Luke shared a website with multiple graphs displaying carbon emissions and a map of global daily income. Carly was an environmental studies and a creative writing double-major; she found two TED talks on the use of big data visualizations in the service of arguments on global differences of government spending. These students subsequently pursued their interests in the environment and global economics in the designed tasks with Gapminder. This was not the case for all participants, such as students who explored themes or interests in the foraging activity that were not as amenable to exploration with the Gapminder tool and datasets (e.g., data stories on sports, consumption of red meat, election data).

As in other studies of multimodal, intertextual modeling (e.g., Goldstein & Hall, 2007; Hall et al., 2007; Ochs, Gonzales, & Jacoby, 1996), we considered participants' noun and verb use within utterances, body movement and gestures, and use of the Gapminder tool as sources for evidence of modeling, storytelling, and learning. Across the entire data corpus of Gapminder performances (three Hans performances, six Mathematics Literacies student performances, six Human Geography student performances, 10–20 min per performance), we looked at how participants used the Gapminder tool (i.e., playing time) to model the world. We considered the kinds of comparisons and the forms of temporal, spatial, and social scaling participants performed to ask and answer the questions about society with their model. For storytelling, we examined whom or what participants animated as historical actors in global development (e.g., political crises, governments, individual citizens) and whether their storylines challenged dominant narratives. To understand what students learned through the experience of telling stories with big data, we looked for students' reflections on how a Gapminder model can be manipulated to tell different stories. We also looked to see if they used their models to support critical perspectives (counter-modeling) and if they related models made with aggregate data to individual (possibly personal) experiences. In our content logs, we noted pronoun and subject referents (e.g., "us" as the United States or "us" as students), the animation of graphical space (Ochs et al., 1996), and the spatial (local, global) and temporal (past, present, future) scales invoked that indicate participants' perspectives toward relations among data, themselves, models, and the world.

In the following section, we identify and describe modeling practices and substantiate these analytical categories with examples from both Hans and the two focal student groups: *horse racing*, *time jumping*, and *getting personal*. We also developed a coding framework in order to determine the frequency of these kinds of comparison practices and counter-modeling across the larger corpus of performances in each class and in Hans' talks (Appendix A). We discuss our corpus-level findings from this coding work as well

Findings: Ontological Discoveries

We describe modeling practices with Gapminder resources that were common across Hans and students' Gapminder performances in both iterations: *horse racing*, *time jumping*, and *getting personal*. We suggest that each of these practices involves particular spatial, temporal, and social scaling of global phenomena in relation to personal experience, and these practices shape the kinds of critical stories that can be told with the models based on LSDS. We describe these novel forms of modeling with LSDS as ontological discoveries or innovations (diSessa & Cobb, 2004) made possible in our design research. They are analytic categories that we have developed and refined over the course of the design-based research rounds that help us "to see order, pattern, and regularity" in the complex setting of data modeling with LSDS (diSessa & Cobb, 2004, p. 84). These discovered practices may provide a taxonomy for further design work that is grounded with case study examples (Hans; focal students). For each practice, we give descriptive excerpts of multimodal discourse to demonstrate how our design cultivated students' learning to model and tell critical stories with LSDS.

Our coding revealed that these three modeling practices were common among Hans and less so among students. Horse racing, time jumping, and getting personal with big data peppered Hans talks. Furthermore, in each performance, Hans engages in counter-modeling, using his datasets to critique a Western perspective towards splitting the world into developed and developing. He animates his data to convince viewers to recognize convergence of nations on health and wealth measures by showing how gaps over the last 60 years have been replaced by continuums and, at the same time, wide variability exists among nations within regions (e.g., there no "one" Africa) and even within their own national borders.

For the students, the prevalence of both the conceptual comparative practices under study and social critiques was inconsistent. However, counter models that presented critical perspectives towards powerful state or political actors appeared more frequently in the Human Geography class than the Mathematical Literacies class. This may because the instructor in the Human Geography course had established class norms for sharing justice-oriented, critical perspectives on geospatial phenomena and for valuing such perspectives as form of civic engagement (based on class observations, conversations with the instructors, and the class syllabus). Alternatively, while the Mathematical Literacies course centered mathematical and statistical argumentation and literacies in the world outside of school, there was not a strong foundation for political or social critique.

Table 1

Storytelling and Modeling Across Han and All Student Performances

			Instances of Modeling Practices Across Each Performance				
	# of Models	# of Stories	Horse Racing	Time Jumping	Getting Personal	Counter-modeling (counter to powerful state actors)	"Counter-modeling" (counter to Gapminder/ study design)
Hans 2006	5	9	2	2	6	Yes	
Hans 2007	3	7	2	2	5	Yes	
Hans 2009	5	12	4	6	4	Yes	
ML 1*	1	1	1		1	Yes	
ML 2	1	2					
ML 3	2	2					Yes
ML 4	1	2					
ML 5	2	2		1			
ML 6	3	6					
HG 1	1	8				Yes	
HG 2	3	3			1	Yes	
HG 3	1	1			1	Yes	
HG 4	5	5					
HG 5**	2	3		1	1	Yes	
HG 6	2	2					Yes

Note. A new model was identified when the axes or indicators, or modeling interfaced changed. A new story was identified by a new model, when new country actors became a focus, or when scales transformed (i.e., from linear to logarithmic scale or vice versa) with the same model.
*Mathematics Literacies Group Nathan, Nicole, and Yuri
**Human Geography Group Carly, Luke, and Daphne

We attribute this variation to several aspects related our study design that are worthwhile to note. Hans typically tells stories with multiple models during a TED performance, and TED talks are developed and practiced over a period of 6 months (based on the TEDx website tutorial). Our student performances generally used fewer models, although most told multiple stories, and our students had a class period and some homework time to develop the model for their performance. Additionally, many of Hans' models involve operations on data that are not possible with the public version of Gapminder. Hans and his team (from a Reddit AMA) use a great deal of Flash animation that permits splitting of nations and regions into quintiles, transforming motion chart bubbles into pieces of a global distribution (a non-Gapminder interface), or showing nations on the same model over different two periods of time. A few student groups brought in models from other sources to index to their Gapminder motion charts, like our case students Carly, Luke, and Daphne, but these *mashups* (Phillips, 2013) were relatively rare among our class participants.

Additionally, searching in the Gapminder database for variables that would be dynamic and interesting was a bit like wandering in the desert searching for water, even for university students. Gapminder has over 500 "indicators" to choose from, each of which can be displayed in one of its four variable quantities in the graphical space (x-axis; y-axis, bubble-size, and bubble color; see Figure 1). Time runs below the x-axis. Users can select countries to compare over time (by "playing" time), if they desire, on these indicators; countries can be selected to leave a bubble imprint for each year that there is data, creating a "trail" or trajectory through time (Robinson, Yu, Zeller, & Felton, 2009)

Students in both the Mathematics Literacies and Human Geography classes struggled to determine which sets of covarying quantities (indicators or variables) felt relevant for them to make a compelling story (Kahn, 2014). Students had to respond to the many novelties of the Gapminder tool. It was challenging for them to decide between different display and scale options (x- and y-axes, bubble size, linear or logarithmic scale). While building models, they repeatedly animated the dynamic simulation (Figure 1), looking for unexpected country "paths" in the graphical display, like the appearance of outliers undergoing dramatic change (backwards or vertical movement on the x- and y-axes, respectively) or unanticipated stasis (consistent movement along either axis). In particular, how pairs of indicators covaried (e.g., showed positive or negative correlation) captivated their attention. As noted earlier, faced with a surprising or interesting graphical pattern, students frequently turned to Wikipedia to contextualize the visible paths (trails) of country bubbles in terms of historical events and to develop storylines.

In response to these *data wrangling* challenges, two student groups (ML 3 and HG 6) chose to simplify the multivariable messiness by selecting only one indicator for the y-axis, placing time back on the x-axis, and keeping the other variables (bubble size and color) constant. ML 3 also challenged the validity of Hans' data by looking at Gapminder's measure of Data Quality of their indicator under investigation (income per person). HG 6 centered on the misleading nature of data visualizations by presenting a false story about global geopolitical influence to explain why the two countries diverge economically and socially overtime. While this kind of "counter-modeling" was not an intended result of our design, we recognize these particular modeling performances as deliberate, critical efforts in which the students developed models that challenged or countered both Gapminder and our instructional design. Hans, an internationally renowned epidemiologist who has been called the "Mike Jagger" of statistics (a mispronunciation of "Mick Jagger" that Hans uses intentionally), never simplified his models in this way in the corpus of public performances we analyzed. In fact, Rosling argues that taking time off of the x-axis was a major breakthrough in his field of public health (Blum, 2006). In a related practice to simplify the complexity of the tool, other students captured and presented static motion charts without playing time, possibly holding Gapminder “steady” to look for functional relationships. The result of their efforts resembled what is typically displayed with a regression scatter-plot.

While the LSDS and visualization were initially overwhelming for some students, it is important to stress how the modeling depth of the Gapminder data interface enabled students’ exploration, inquiry, and assembly of critical models and stories about society. Although we only focus on modeling performances in this paper, Gapminder was the primary resource for students’ modeling practices across their class activity in both study iterations. For instance, students could manipulate the world and, as students acknowledged, alter the stories told by changing or transforming x- and y-axis indicators between logarithmic and linear scales, which has the effect of shifting the bubbles around in two-dimensional space (a log scale compresses very high values and make differences between low values more visible, while a linear scale reveals changes in very high values).

Johansson (2012) suggests that the Gapminder tool “discipline[s]” its users by limiting the kinds of inferences made, questions asked, and stories told (Hutchins, 1995, as cited in Johansson, 2012, p. 117). However, we found that Gapminder’s multivariable dynamic capacities supported a wide variation of stories and questions. In their final modeling performances, student participants not only compared national trends for health and wealth indicators over time, as our focal cases did, but they also used to the tool to focus on outliers, to consider models with ambiguous relational patterns, to introduce issues of data quality, or to challenge the meaning of axis transformations. We conjecture that these other kinds of stories that participants presented partly account for the lack of time jumping, getting personal, and horse racing across student groups as exhibited in Table 1.

Gapminder's modeling depth stood in contrast to the students' foraged stories and models from the public news media, which were often "finished" infographics and shielded access to data sources (Gapminder database listed data sources for each indicator). While not all of students' questions about Gapminder's data trends or concerns about data quality were answered, as Nicole said in a post-semester interview, with Gapminder, "it's easier to ask questions and investigate deeper for yourself" (Interview, December 7, 2012). For example, in both classes, each student performance was followed by rich, whole class discussions that involved returning to the Gapminder tool to explore new questions that were raised by the modeling performance. This kind of interaction with the audience does not happen within a tightly scripted Han's TED Talk, but we know (from Johansson, 2012) that Rosling and colleagues engaged with Gapminder models in ways that led to discoveries.

Horse Racing

Horse racing describes the practice of comparing two nations' trajectories (trails) of development over time in the model as competing entities in a race for socioeconomic development on a global stage. Students and Hans horse raced countries in a comparative modeling practice to tell stories about nations undergoing continuous change over time, changes that were typically influenced by specific historical events. In the following excerpt[10] (Excerpt

[10] Transcript conventions for excerpts are as follows: CAPITALS indicate emphasis; (*Observer notes*) indicate significant gesture; ? indicates rising intonation; ! indicates exclamations; [

1), taken from a TED Talk made to the U.S. State Department in 2009,[11] Hans races his native Sweden's child mortality rate and income per person against that of Bangladesh, Egypt, and Brazil. He plays Bangladesh, Egypt, and Brazil each from 1990–2006, while he runs Sweden for consecutive 16-year periods, starting in 1916. Hans is using a different, earlier version of the motion chart display that our students used (see Figure 2); Hans is able to "start" time for different nations at different years simultaneously on the same graph, which the students could not do with their version of Gapminder. Gesture is omitted from Hans' excerpt because he is not visible in the video record as the motion chart plays.[12]

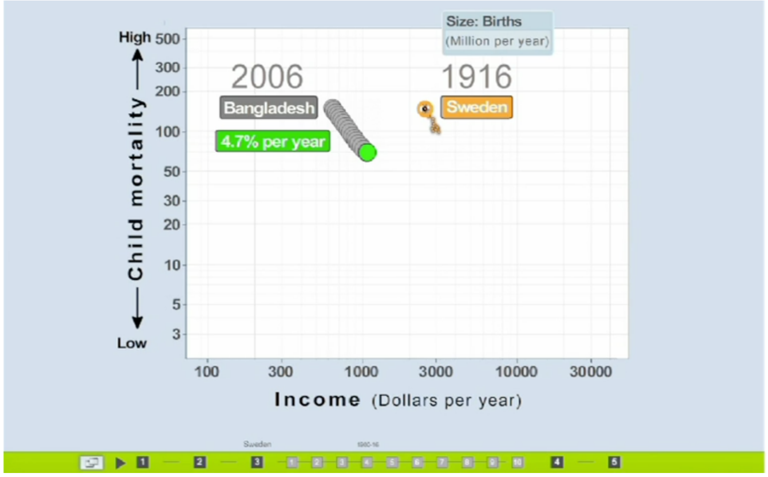


Figure 2. See Excerpt 1. Motion chart display shows income in dollars per year on x-axis and child mortality from low to high on y- axis; both axes appear to use a logarithmic scale. There is one yellow bubble for Sweden in 1900 with label, and one grey bubble for Bangladesh in 1990 with label. Hans plays the motion chart for a 16-year period as he speaks. Country trails are on to create paths through time.

Excerpt 1

1 Hans: 1900 Sweden was there. SAME child mortality as Bangladesh had 1990 though they had lower . income . they started very well (*speaking pace speeds up*) they used the aid well they vaccinated the kids they get better water AND they reduce child mortality with an amazing 4.7%

indicates overlapping talk; = indicates latched talk; ::: indicates elongated vowels; . indicates pauses less than a half-second … indicates pauses longer than a half-second.

[11] Hans Rosling's full talk can be found here: https://www.ted.com/talks/hans_rosling_at_state.

per year they beated Sweden . I run Sweden the SAME sixteen-year period (*Gapminder display changes. Bangladesh disappears and Egypt bubble appears at 1990. The Sweden bubble remains 1916 in yellow. Hans plays graph for another 16-year period. Trails are on.*)

2 Hans: second round . it's Sweden 1916 against=

3 Audience: (*rising laughter*)

4 Hans: -=Egypt 1990 . here we go once again (*pace of voice has speed up*) U.S. aid is part of the reason here they get safe water they get food for the poor and they get malaria eradicated 5.5% they are FASTer than millennium development goal rate

5 Hans: and third chance for Sweden against Brazil here

6 Audience: (*laughter*)

(*Gapminder display changes. Egypt disappears and Brazil bubble appears at 1990. The Sweden bubble remains 1932 in yellow. Hans plays graph for another 16-year period. Trails are on.*)

7 Hans: and . Brazil here has an aMAZing social improvement over the last . 16 years and they go faster than Sweden . this means that the world is converging . the middle-income countries the emerging economies they are catching up .

Hans assumes the role of a sportscaster as he describes the various nations' progress in world development with heightened cadence and tempo, as if describing a horse race or competitive sporting event instead of a statistical argument. Hans artfully positions the audience as rooting for one side or the other. The audience, comprised of U.S. State Department employees, is left rooting against Hans' native Sweden in favor of nations where U.S. aid, Hans suggests, facilitated development. He describes the pace of development in Bangladesh, Egypt, and Brazil as "faster" than Sweden's rate of development to reach the same rate or "finish line" in terms of child mortality, granted Sweden reached that milestone half a century earlier. Hans' comparisons assume that nations around the world follow the same trajectory and Western *telos* of development and progress (Kahn, Hall, & Phillips, 2014).

Students in the Mathematics Literacies class and in the Human Geography class watched this particular video of Hans "racing" countries prior to their performance, and their comparative argument structures likely drew from their viewing and analysis of his performance. Students describe countries as catching up to one another on various indicators (emissions, income, population) over time. The Mathematics Literacies group horse raced countries to establish contrasting cases between the United States or Western highly developed nations (e.g., Organization for Economic Co-operation and Development [OECD] members) and emerging economies (e.g., Brazil, Russia, India, China [BRIC]) to confront popular criticisms directed towards the latter group as the worst contributors to global CO^2 emissions.

Preservice mathematics teachers Nathan, Nicole, and Yuri horse raced China, the United States, and the United Kingdom to demonstrate how culpability in global climate change is diffuse and complex as a result of historical and present power and capitalistic relations (see Figures 3a and 3b). As time plays from 1800, Nicole's cadence speeds up as they tell the story of how rising CO^2 is associated with historical periods of industrialization for each of the nations. In Excerpt 2, at the turn of the century, the students describe the United States, which was lagging behind the United Kingdom in income and carbon emissions, as "catching up and surpassing the United Kingdom" on these measures.

Excerpt 2

(*Nicole stands with a pool pole on the left. Yuri stands by the computer on the far right. Nathan, right of the screen presses plays. The pace of play is about 2 years per second; Gapminder shows Income per person (x) by CO^2 emissions (y) in metric tonnes person; initially only two countries are shown, the U.S. in yellow and UK in orange. Color indicates geographic region. Both are at about $2000 to $3000 with 1 to 2 tons of CO^2. China arrives on display in 1900.*)

1 Yuri: (*Gapminder is at 1864*) so this is the U.S. civil war, and then after that, BAM, it goes up . cause it's the . U.S. industrial revolution . (*Nicole is pointing with the pool pole*)

2 Nicole: there's that there … (*Gapminder is at 1900*) so you can see (*pace of voice has sped up*) the

United States is kind of catching up and surPASssing the United Kingdom pretty quickly . right here (*points to place where country trails intersect*)

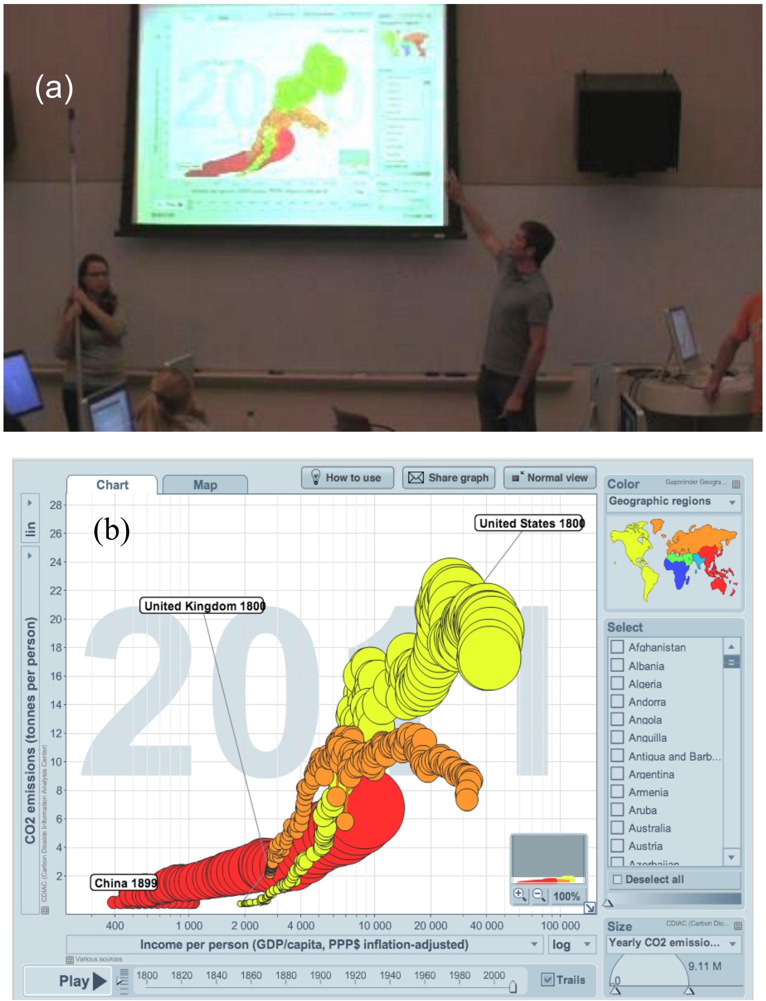


Figure 3a and 3b. Nicole, Nathan, and Yuri perform their story (a). Their final Gapminder model (b) displays income per person (GDP/capita) on the x-axis, CO^2 emissions per person on the y-axis, and bubble size represents yearly total CO^2 emissions. China (red), the United States (yellow), and the United Kingdom (orange) are displayed. Time played (marked by the scale at the bottom) from 1800 until 2011.

As the race unfolds, the United States has finished in the leading position in CO^2 per person emissions (the indicator on the y-axis; see Figure 3b). But then the students point out that the answer to the question of who is responsible for global CO^2 emissions depends on how CO^2 is measured. In Excerpt 3, Nathan explains that CO^2 emissions per person (y-axis) and a country's yearly CO^2 emissions (bubble diameter) offer different answers to the question of who is responsible for the world's carbon pollution.

Excerpt 3

1 Nathan: so um looking at this graph (*he is looking at the graph, body faced toward the projection then turns to audience*) . I guess you could ask the question , which country is producing the most? . and it really depends on how you look at it (*gaze returns to projection*) . because . if you're looking at it per person? . (*on "DEF" R hand raises as if marking a height, faces audience on "U.S."*) then DEFinitely the U.S. is gonna be . your . uh . is gonna produce more (*R arm extends and reaches towards bottom of projection and gaze follows*; *looks to audience on "emissions" and returns gaze to projection*) . but if you're looking at just total CO^2 emissions. then it's gonna be China (*points to self on China*) . and we actually in our research we found that . um . China: produces a quarter of the world's CO^2 and the United States produces um about twenty percent . so together they account for like HALF of (*on "all of the world's" makes a round motion like locomotion train and drops hands*) all of the world's co2 emissions

2 Audience: oh:::

3 Nathan: so

4 Nicole: but . we were we were discussing earlier . like-that might not be a FAIR . statement? because we're talking that China's economy? they . produce a lot of exports . that go all over to different places in the world . and so they're producing (*arm extends slightly towards audience on "apple computers"*) apple computers or whatever that (*points to self on "we're"*) we're buying here . so are those emissions . really . THEIR::: emissions or if we're buying their products . from the factories that they're making . should we be allotted some of those emissions that are happening .

In the excerpt above, responsibility depends on who is "credited" with emissions, the producers of these products (BRIC nations) or people, like themselves, who consume the products. They critically point out that Gapminder CO^2 emissions are not measured in this way in the model: China's CO^2 emissions, on either indicator, do not reflect U.S. imports, which they suggest might be driving Chinese emissions. This narrative counters conventional data stories found in the media and politics. While conventional understanding might frame highly populated, developing nations as being responsible for current world pollution, Nathan, Nicole, and Yuri develop a mathematical argument for countering this way of telling the story with data

For both Hans and the students, horse racing entails spatial scaling across vast stretches of geography in order to compare, contrast, and critique global measures of CO^2. From the Middle East to China, Europe, and the Americas, Gapminder serves as an international arena for watching assorted nations compete in industrial, social, and economic events; individual nations rise and fall as they face, stumble, and possibly overcome political, economic, and social crises. To continue the horse racing metaphor, this modeling practice contracts both time and space so that national narratives unfold on a "level" playing field, prompting the exploration of possible global relationships between nations' historical trajectories (e.g., the relationship between U.S. consumption habits and China's CO^2 emissions).

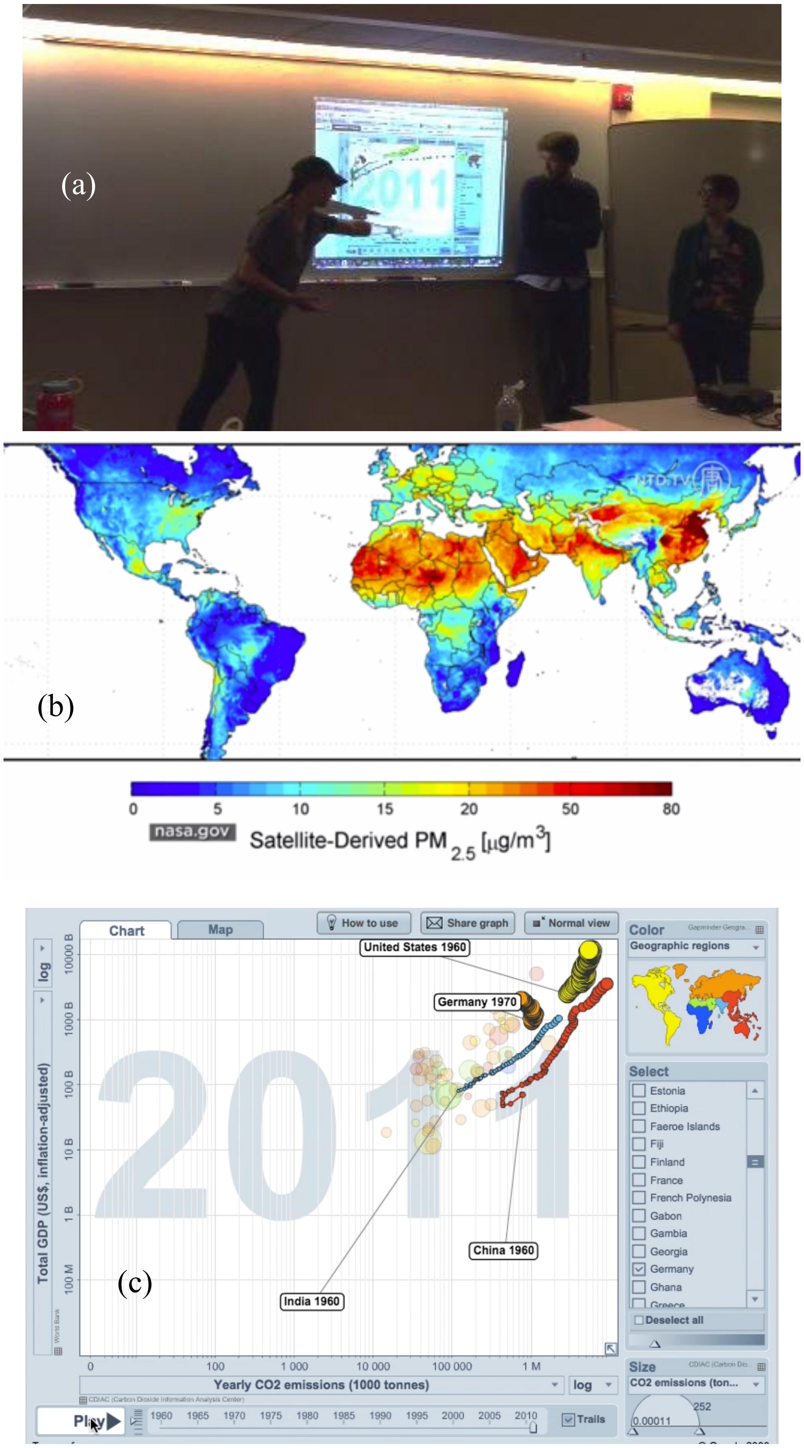
Time Jumping

Time jumping refers to the comparative modeling practice of moving between varied temporal scales in order to support a narrative that contrasts nations over discrete time periods. Time jumping interrupts and reorders continuous variation in measured variables. Time jumping, as particular kind of modeling practice, differs greatly with, for instance, a multiple regression analysis, which depends on continuous multivariable variation over time in order to explain variance. Hans and the focal students in Human Geography time jump in order to highlight the differences or similarities between two countries' advances or declines on a specific indicator. For example, Hans in Excerpt 1 above, compares development in Bangladesh 1990–2006 to Sweden 1916–1932, Egypt 1990–2006 to Sweden 1932–1948, and Brazil 1990–2006 to Sweden

1948-1966. He compares BRIC millennial development with OECD 20th century development to demonstrate that the pace of industrialization (income) and expansion in health care services (reduction in child mortality) has grown faster in the past two decades for these nations than it did for Sweden over a half century.

Human Geography students Carly, Luke, and Daphne (Figure 4a) compare China, the United States, India, and Germany from 1960 until 2011. Like Nathan, Nicole, and Yuri, they also tell a story about the rise of the world's carbon emissions (yearly CO^2 emissions is on the x-axis; per person emissions is the bubble diameter) as related to industrialization, which is represented by each nation's total GDP (y-axis). Luke, Daphne, and Carly begin their performance with an announcement announcing China's commitment to reduce emissions 45% by 2020. They then show a YouTube video entitled "China's Real Pollution Problem," which shows a news-style interview about China's pollution. In the video, the interviewee uses a choropleth map produced by NASA that displays global amounts of small, harmful particle production (Figure 4b) to demonstrate China's pollution as significant and problematic. The students pull up the NASA map and suggest that the scale misrepresents the U.S. as a low polluter even though U.S. produces a significant number of harmful particles. Carly then goes a step further to suggest that Western colonization is partly responsible for the distribution of dark red (larger amounts of particles) in Africa, the Middle East, and Asia today.

The students then present their Gapminder model to examine the issue of industrialization and carbon emissions. In their performance (Excerpt 4), they change the x-axis scale for yearly CO^2 emissions from linear (Figure 4c) to logarithmic (Figure 4d) in order to highlight China's very recent and substantial growth in CO^2 emissions. The variation in the rates of change of yearly CO^2 across the selected nations was less visible when the scale was logarithmic.


(a)
2011
(b)
NTDTV
0
5
10
15
20
50
80
nasa.gov
Satellite-Derived PM$_{2.5}$ [μg/m^3]
(c)
Chart
Map
How to use
Share graph
Normal view
Color
Geographic regions
Select
Estonia
Ethiopia
Faeroe Islands
Fiji
Finland
France
French Polynesia
Gabon
Gambia
Georgia
Germany
Ghana
Greece
Deselect all
Size
CO2 emissions (ton...
252
0.00011
United States 1960
Germany 1970
China 1960
India 1960
2011
log
Total GDP (US$, inflation-adjusted)
10000 B
1000 B
100 B
10 B
1 B
100 M
World Bank
0
100
1 000
10 000
100 000
1 M
Yearly CO2 emissions (1000 tonnes)
log
CDIAC (Carbon Dioxide Information Analysis Center)
Play
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
Trails

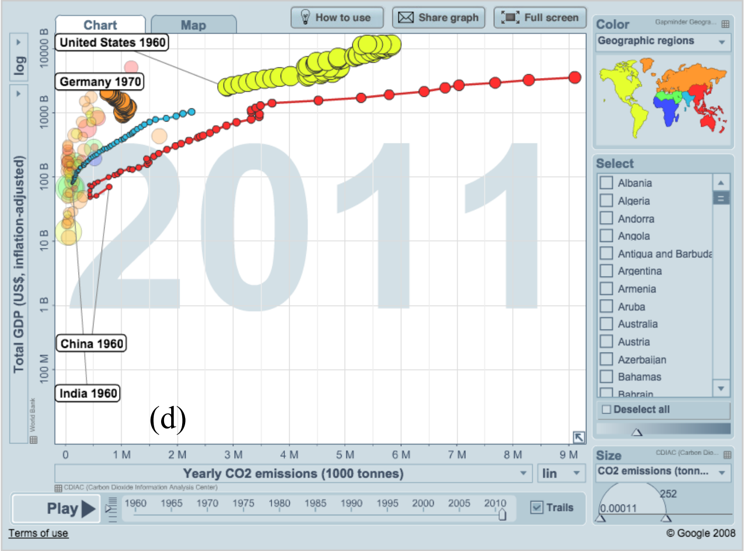


Figure 4a, 4b, 4c, and 4d. Carly, Luke, and Daphne perform their story to classmates (a). They begin their performance with this NASA image (b) of harmful CO2 particles which they later index to their model. Their model (c) displays yearly CO^{2} emissions on the x-axis with a linear scale, total GDP on the y-axis with a linear scale, and bubble diameter is CO^{2} emissions per person. China (red), United States (yellow), India (teal), and Germany (orange) are selected. Time played from 1960 until 2011. For their final model (d), the x-axis has changed to a logarithmic scale.

Excerpt 4

(Daphne is at the computer, manipulating the Gapminder interface.)

1 Daphne: so so at first we had nothing to talk about because this seemed kind of too cut and dry:: and not not anything that we didn't know before . but then we changed it be linear? (*changes x - axis to linear, data bubbles slides left*) . so we changed the emissions to be linear and um then this was what we got... first of all we're WAY ahead with our emissions even at the beginning (*moves time scale back to 1961. Gapminder plays again from 1961 with x-axis in linear scale*)

In Excerpt 4, Daphne tells the class audience that the U.S. ("we're") was "way ahead" (farther along the x-axis) from "the beginning" (1961, where their model starts, see Figure 4b). This time jump contrasts U.S. carbon emissions in the 1960s with past and present carbon emission rates of the other selected nations. The implication is that the United States' was always a leader in CO^{2} production, at least as compared to the other nations in the 20th century. We see more time jumping in Excerpt 5, which occurs a few moments later, after Gapminder has played again from 1961 with the x-axis using a linear scale (Figure 4c).

Excerpt 5

1 Luke: just go to two thousand twelve I think that's good (*points to display*)

2 Daphne: 2000 is when this goes crazy (*moving time scale back and forth between 2000 and 2011. leaves graph at 2002*) so that's 2000

3 Luke: uh yeah go all the way to the future=

4 Audience: (*laughs*)

5 Luke: =to the present ... um . so we see here like this (*R pinky finger points to china's trail and traces trail*) HU::Ge increase in china . um what's particularly interesting to me (*points to self*) anyway is there is, china is so far to the right (*R hand goes vertical, parallel with y axis, moves up and down*) on emissions? but the per person? (*R hand points to bubble size*) is a lot smaller than the U.S. (*R hand points to U.S. trail*), so the U.S. per person emits three times the amount . as the average person in China . and we were were trying to figure out . we figured out (*R hand points to bubble size*) emission per person is simply just all the emissions divided by the population.

Like Hans, the students in Excerpts 4 and 5 must "start" the world and then "stop" time to issue observations and explanations. In their comparisons of ratios and transformations of model axes between logarithmic and linear scales, students time jump between the past and the present (Luke says, "yeah go all the way to the future=the present" Excerpt 5). The students use the interactive capacity of Gapminder for inquiry: they manipulate their model to go "back to the future" in order to juxtapose present day BRIC emissions alongside present and past OECD emissions. Subsequently, they demonstrate that 1960 U.S. per person emissions were greater than 2011 India per person emissions. These time jumps invite critical examination of measures, which develops into a mathematical argument: They explain that how CO^2 emissions are calculated determines which nation produces more CO^2 (i.e., while China leads in yearly total emissions, an American "emits three times the amount as the average person in China"). The argument in turn sustains their counter-narrative: While China and India's economies have practically "caught up" to that of the U.S., their citizens still have far to go to match the size of American citizens' carbon footprints (and their contributions to climate change).

The students subsequently index their new story with the transformed model to the map representation that they introduced in the beginning of their performance. Pointing to both the ending point for the U.S. bubble trail in 2011 as the blue in the NASA map, Carly and Luke present that using U.S. as a baseline for comparing global emission standards is problematic, as the U.S. produces a tremendous amount of CO2 emissions. In turn, the students contend that the popularly espoused American criticism towards China and India, even if warranted, is hypocritical. As their Gapminder model indicates, the U.S. has not been very successful in reducing carbon emissions, and most of the U.S.'s socioeconomic achievements occurred historically without emissions restrictions. Yet, the U.S. demands that China and India adhere to a different standard.

Getting Personal

Getting personal with big data describes the final modeling practice that we identified across Hans and both of the student groups' Gapminder performances. Getting personal entails finding oneself in (or outside of) the model or aggregate data, which subsequently demands *positioning* oneself in relationship to the data model and the narrative being told. Goffman (1981) described positioning or framing as a participant's *footing*. Footing, Goffman writes, refers to:

> …the alignment we take up to ourselves and the others present as expressed in the way we manage the production or reception of an utterance. A change in our footing is another way of talking about a change in our frame for events. (p. 128)

Accordingly, this “aligning,” to the data model and the narrative being told, as expressed through utterances, is one way to see “getting personal.” This active framing is similar to what Bamberg (1997) calls *narrative positioning* and engages what Wortham (2000) describes as the *interactional positioning* of the audience. In shifting alignments, we can examine how Hans and our students, as both storytellers and modelers, express a relationship between themselves (their identities) and the reported events (the story told about the model) and between themselves and the audience. For instance, in the Human Geography class, students made implicit and explicit identity references to themselves—as American women, as “social studies teachers,” or as liberal millennials—and noted how those identities shaped the models they created and the stories they told. As our examples from both Hans and the focal groups show, getting personal in this way supports the integration of macro, aggregate trends with the micro, individual experience, which can lead to a critical stance (a way of looking at society, and seeing oneself in that relation) that invites counter-narratives and counter-modeling.

As demonstrated in the first two parts of the analysis, students and Hans built models to support structured comparisons that engaged temporal and spatial scaling practices: They compared various countries’ trajectories over time (horse racing) and compared countries at different points or periods in time (time jumping). The students and Hans also got personal with aggregate data in their modeling performance. Getting personal with aggregate data involves the scaling of social life in order to describe social processes, their moral implications, and their consequences across individual and aggregate scales. It entails moving outside Gapminder in order to relate life experience—personal scales in time, space, and social relations—to trends described by the aggregate data available in Gapminder. In Excerpt 6 (Figure 5), Hans (Rosling, 2007) does this when he aligns his own family history to current millennial stages of economic and social development for nation-states.[13]

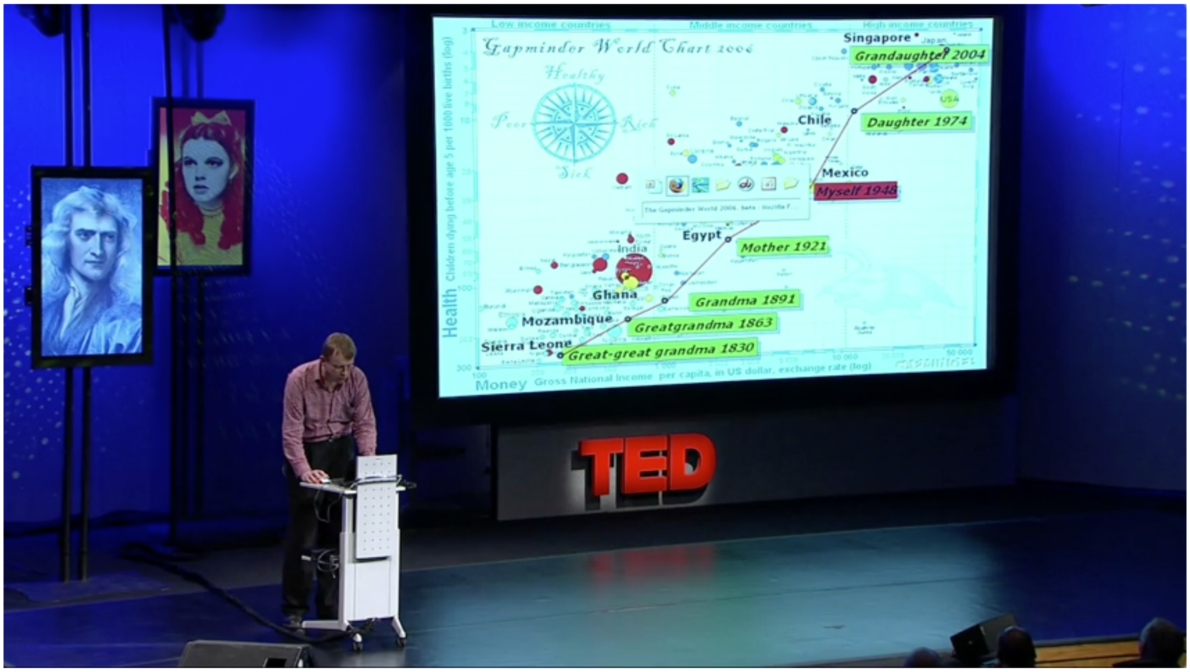


Figure 5. See Excerpt 6. Gapminder motion chart shows all of the nations in 2006. X-axis shows Money, measured by the gross national income per capita in U.S. dollars on a logarithmic scale. The y-axis shows Health, measured by child deaths per 1000 births on a logarithmic scale, with fewer deaths at the top of the y-axis and more deaths closer to the origin. As Hans mentions each member in his family, they appear as labels (highlighted in green) next to points on a line graph connecting nations with different values for Money and Health.

Excerpt 6

1 Hans: and let me show you my own sort of family . history eh we made these graphs here . and this is the same thing money down there and health you know? and this is my family ... this is Sweden 1830 when my great-great-grandma was born … eh .... Sweden was like Sierra Leone today …. and this is when great-grandma was born 1863 eh . and Sweden was like Mozambique and this is when my grandma was born 1891 she took care of me as a child so I'm not talking about statistic now now it's oral HISTORY in my family … that's when I believe statistics when it's grandma-verified statistics you know …

2 Audience: (*laughter)*

[13] Full Hans talk can be found here: https://www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty/transcript?language=en#t-520612

3 Hans: I think U think it's the BEST way of verifying historical statistics=Sweden was like Ghana it's interesting to see the eNORmous diversity within sub-Saharan AFRICA . I told you last year I'll tell you again my mother was born in Egypt and I? who am I? I'm the Mexican in the family uh? . and my daughter she was born in Chile and the granddaughter was born in Singapore now the healthiest country on this earth it bypassed Sweden about two to three years ago with better child survival huh but they're very small, you know? they're so close to the hospital we can never beat them out in these forests.

4 Audience: (*laughter*)

5 Hans: but homage to Singapore ... Singapore is the best one … now … this looks also like a very good story.

Hans explains increases in national income and decreased child mortality over matrilineal generations. He locates himself and his family in the data model. To do so, he scales time, space, and social life, the latter represented by the wealth and wealth of nations, in complicated ways. He says, “who am I? I'm the Mexican in the family uh? and my daughter she was born in Chile.” Hans’ alignment of himself and his family, his “grandma-verified statistics,” also rests on an assumption that the life circumstances of developing nations are not that different from our own, historically speaking, and that all nations want the same conditions of economic and social life as the OECD countries. Hans’ and the students’ comparative work smoothly positions the audience as willing followers of his logic. (Note: This was not an assigned video for the students in either research iteration).

Both focal student groups got personal with the LSDS to tell a counter-narrative and engage in counter-modeling. As we see in Excerpt 3 and below in Excerpt 7, the focal students got personal with the data by positioning themselves as the United States, a powerful actor in the model and in global CO^2 production. As in Ochs et al.’s (1996) analysis of hybrid predicate structures in physicists modeling states in which pronouns (e.g., I) denote both the physicist's perspective and the perspective of the physical matter in a graphic model, the students' grammar reflect a hybrid identity. The shifts in pronoun use (e.g., in Excerpt 3 Turn 4 we in "we were discussing earlier" to "what we're buying here" or I in "I think" to "we account" below in Excerpt 7) establish the speakers both as students engaged in modeling and as the U.S., a nation of consumers displayed in the data representation. This hybrid identity is further developed with their references to their own consumption of Apple personal computing products (e.g., "apple computers or whatever that we're buying here" in Excerpt 3 and "iPhones in our pockets" in Excerpt 7), objects which they enjoy in the United States but are manufactured in China and India. Getting personal thus supports counter-narratives that challenge popular public media arguments framing nations like China and India as the worst perpetrators of carbon emissions and climate change. The focal students tell this counter-narrative by assembling and animating a counter-model about the rapid increase in American and BRIC CO^2 emissions in the late 20th century.

In Excerpt 7 (see Figure 4c), Luke, Daphne, and Carly have just demonstrated that U.S. standards of emissions, if used to judge the rest of the world, are problematic since the U.S. have very high total yearly emissions compared to other nations (are very far along the x-axis) and lead the per person emissions measure (bubble diameter size).

Excerpt 7

1 Carly: but on top of that we are all people who probably have amazon accounts and eat food on a daily basis that shipped far across the country and have IPhones in our pockets that are made in china and are supported technologically in India um and our number . our country number . doesn't account for that so what is our virtual impact? which I think if you look at the united states um (*gestures towards self on "we"*) we account individually for a lot of China's emissions . just through our consumption and the economic exchange that goes on there. so something that might be interesting to look at was imports and exports? to a country

> in relation to this um . so just kind of thinking of what we buy and how we buy it uh all the transportation [and how often we eat red meat because there are cows

In this Excerpt 7, and in Excerpt 3, the students' stories span scales of time (i.e., long periods of industrial development), geographic space (i.e., from U.S. to China), and social life (consumer habits and technology production). This scaling, supported with the use of first-person pronouns, reflects a change in stance from themselves as students and observers of the modeled phenomena to themselves as the United States, an actor in the model, and assigns social responsibility for a global phenomenon represented by the model (i.e., climate change) within the consumption habits of people sitting together in the audience (as well as the presenters). Like Hans, who draws on his family relations, the students cannot express the alignment between personal experience and the aggregate they want, so they step outside the Gapminder model to appeal directly to patterns of consumption visible in the classroom. This move gets personal in a way that transforms a critique of the aggregate data (what is hidden in the way CO^2 is measured in the Gapminder model) into a critique of their and their peer's personal choices and thus positions themselves and their peers as consequential to the data and the story told with the data model.

In positioning themselves and their peers in relation to the aggregate data, the focal case students assume moral stances to make sense and meaning of the aggregate data (Goodwin, 2007a; Philip et al., 2016). Goodwin (2007) argues that moral stances are central to cognitive and instructional activity. In looking at interactions in daily family activity (i.e., parent and child homework routines), Goodwin illustrates how family members establish themselves as moral and social actors in the task at hand. They embody, represent, or communicate in some visible way what it means to be "a moral member of the community" of focus, which in Goodwin's case is the family (p. 71). In another example of moral stance in interaction closer to the present discussion, Philip et al. (2016) describe when a student takes up a moral stance towards Computer Science classroom discussions of a big data visualization: unlike her teachers or peers, she takes a position that race should be part of their conversation about the data. In our study, the students take up a moral stance towards what it means to be an ethical global citizen in terms of individual contributions (such as from technology and food purchases) to global emissions and climate change.

Furthermore, these personal and *moral stances* are enacted through what Wortham (2003) calls *participant examples*, in which, “participants in the speech event get cast as characters in an example” (p. 194), as evidenced in their use of dual pronoun and subject referents (e.g., “us” or “we” as the U.S. or as students) in their model animations and stories. In his research, Wortham found that in a high school class reading of Cicero's letter to Atticus where Cicero must decide what to do to about the plot to overthrow Caesar, the teacher prompted the student reading the scene to not only speak to what Cicero should do but also what he (the student) should do. As the interaction develops, the character’s and the student’s social identities are conflated, and the student (for Cicero/himself) must respond to the expectations of teacher and his peers. In our project, the students similarly casted themselves and their peers as nation-state actors in global development processes faced with a moral dilemma: What is their responsibility in reducing global emissions as individual consumers of products, like iPhones and computers, that are made in Chinese and Indian factories that significantly contribute to their respective measures of carbon emissions?

Relating personally-scaled experiences to the models and asking ethical or moral (and sometimes political) questions departed from preservice teachers’ general questions about the validity of measurements and what aggregate data measures show and hide in the models. By bringing a story about global data down to the personal scale, aligning individual human experience with aggregate data, the focal students assumed new epistemic stances towards what counts as good measures of carbon pollution and adopted moral stances towards who is responsible for carbon emissions and climate change. Carly's statement (Excerpt 7) that "we account individually for a lot of China's emissions" denotes this shift in an epistemic stance towards the global emissions data displayed. Carly gives an ethical critique that complicates the validity of the data measures and thus changes what can be known about global carbon emissions in relation to development from the model. Indeed, by critically engaging with global LSDS at a personal scale—at the level of individual purchasing power and daily experience—students complicated the accuracy of CO^2 measures and the ethics of censuring BRIC nations for industrial growth and pollution.

Discussion

Our design-based research is developing activities in which getting personal with LSDS can lead to critical practices of counter-narrative and counter-modeling. Using a learning sciences perspective on creating new cultural activities for youth learning, we looked to existing or "real world" practices of storytelling with data in professional and everyday settings that are important for civic participation to inform our design. Our analysis assumed that meanings of inscriptions, data visualizations, models, and representations are assembled in interaction and through practices (Goodwin, 2004).

Gapminder's modeling depth was indispensable for asking questions and supporting critical inquiry. Across our designed activities in both iterations, students examined trends and relationships among socioeconomic variables and selected historical actors for their stories. In their performances, focal students positioned listening audiences to change their mindsets towards world development and social responsibility, using hybrid referents and personal pronouns and the Gapminder tool itself to traverse spatial, temporal, and social scales. As Radinsky (2008) found in his study of middle school students investigating plate tectonics with GIS, the visual data interface only became meaningful to students when they assembled resources to make phenomena not represented by the model (in their case, earthquakes and earth crust and earth plates) present in their discourse. In our study, Hans and students animated nation-states and historical events and made their experiences outside of the model (consumption habits) present to make their stories meaningful, relevant, and consequential for their peers.

Defenders of standards-based STEM education might argue that asking students to tell stories and bringing individual experience to bear on interpretations of aggregate data risks inviting bias that might interfere with "good" data modeling practice or "objective analysis" and limit critical interpretation of covariation evidence (Kuhn et al., 1988). Some data scientists share a similar perspective. They suggest that when the amount of data is so large, the data "speak for themselves" (Anderson, 2009), and theoretical or moral frameworks are no longer relevant or necessary. However, without moral structure or values—the personal in LSDS—we are left unequipped "to meaningfully interrogate the social systems and structures that make up the social world" (Uprichard, 2013, p. 4). Similarly, as Cicourel (1981) points out, only describing the macro-structures of society is incomplete, because the aggregate data are not immune from the nuances of individual social interactions and context. Rather, such data emerge out of local events in everyday life. Moreover, our data use practices have ethical and real political implications, particularly for marginalized populations (e.g., Hans speaking to the U.S. State Department). Moral frameworks necessarily are involved to counter-model, to tell counter-narratives, and to engage in counter-data actions (Dalton & Thatcher, 2014).

What is the generative structure of practices of storytelling and modeling with big data that we have observed or created in these studies? The prior sections introduced the modeling practices of a professional epidemiologist, a group of preservice mathematics teachers, and a group of preservice social studies teachers. Each case study performance was dense with related forms of storytelling and modeling, that is, instances of horse racing, time jumping, and getting personal. With this current analysis, rather than focusing on the frequency at which each modeling practice occurs, we considered the types of processes that our design invited (although, if pressed, our analysis suggested that getting personal was rare, especially among the preservice mathematics teachers).

First, whenever "motion trails" are used to create a visual comparison of countries, some kind of "race" is produced with rich information about the rise, fall, retreat or backsliding, and catching up experienced by countries as social units of analysis. This may be a typical form of comparative discourse when data are visualized using motion charts, and it is particularly well afforded by the use of motion trails (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008). Second, time jumping is an internally diverse set of processes for storytelling and modeling with LSDS. The simplest form seems to be comparing two points in time (discretizing the passage of time) to tell a story about the fate of a nation of interest or about how two (or a small collection) of countries compare over time. The important role of storytelling as an interpretive resource in data modeling (Kosara & Mackinlay, 2013) is most clear here. A more complex variation of the time jumping form is when sequences or paths of development are compared, but they happen over different intervals of common, chronological time (*chronos*) and are expressed in universal developmental time (a particular kind of *kairos*; Erickson, 2004). The idea that BRIC nations will repeat the path of OECD industrialization, with rising wealth and health measured as child mortality, but with catastrophic consequences for the Earth's atmosphere, would be an example of this more complex time jumping.

It is in a particularly spectacular series of examples in Hans' performances of time jumping (e.g., his matriline) that we also discover and can begin to analyze processes through which storytellers and modelers get personal with LSDS. Moments of getting personal, in the sense of embedding oneself or personal experiences at the scale of everyday life or the life of one's family into the model, appear to arise in response to a question of whether an official interpretation (story) about a data model accurately describe one's own lived experience. It is here that moral stance (Goodwin, 2007) and epistemic critique (at least regarding what has been or could be measured, and whether others forms of horse racing and time jumping are plausible) emerge in storytelling and modeling performances. Getting personal with LSDS may not be so much a process or a strategy as a stance in epistemic or moral terms. This stands in line with other reports that have found that learning about how to work differently (to change their professional practice) involves taking up new epistemic and moral stances (Hall & Horn, 2012).

Finally, getting personal with data may be a form of what Erickson (2004) called “wiggle room” in moments of social interaction that may perform forms of resistance that lead to social change (Johnson & Amador, 2011). Erickson argued that in every interaction, individuals have opportunities to make political and moral choices. One has to decide, “Do you swim against the currents of your world” or “Do you go downstream with the currents?” (Johnson & Amador, 2011, p. 100). Bringing aggregate data down to the personal scale suggests that individual agency matters in the sea of data, and our individual choices may align with or challenge the direction that the world is moving towards.

Design Plans and Conjectures for the Future

We argue that our design thus supported new forms of modeling practice, critical inquiry, and discourse around stories told with models of large-scale data. Our future design plans include continuing to bring conceptual practices of getting personal with LSDS into view and designing learning environments in which youth can have these experiences. Esmonde's (2007) narrative of when classroom, world, and personal *figured worlds* (the set of norms and expected roles for a given activity system or setting; Holland, Lachiotte, Skinner, & Cain, 2001) collide around issues of social and economic justice offers some insight for understanding getting personal as an object of design. The fair sharing of cookies (a figured world of doing middle school math) met the distribution of global wealth and developmental aid (a figured world of global wealth) in a participatory simulation in which students stood in for countries and cookies symbolized country wealth (a hybrid figured world). In the hybrid, intermediary figured world, students were poised to take a moral stance towards what is an equitable or fair distribution of wealth and aid in relation to personal financial circumstances (a figured world of family wealth and debt). Likewise, we might think of getting personal as a product of the collision between figured worlds that our current design invited—that is, an encounter between the world of data science and socioeconomic modeling and with the world of personal experiences and purchasing decisions. In future work, we could design more intentionally in order for contact between these kinds of figured worlds to occur.

Personally-scaled counter-narratives emerge as participants engage in scaling practices (time, space, and social life), from local to global and back again, across macro and micro data structures. We conjecture that counter-narratives LSDS are always possible, but without the depth afforded by visualization tools like motion charts, it is difficult or impossible to explore assumptions in a found model and to produce counter-models. Moreover, critical and counter-narratives do not emerge if the learning design does not either permit it or deliberatively encourage it (Donovan, 2012).

Conclusion

In summary, our designed instruction guided preservice secondary mathematics and social studies teachers to become critical authors, consumers, and producers of data, models, and stories. Like Hans, students engaged in modeling practices that entailed conceptual movement across scales of time, geography, economic and social activity, and moral accountability. The interactive nature of Gapminder motion charts facilitated their complex comparisons, descriptions of covariation, and their animation of global development. Storytelling with data, as a new perspective on modeling practices, thus offers essential and expressive means for building relationships between representing and represented worlds (Gravemeijer, 1994; Hall, 2000). Moreover, students' personal identities and interests supported their critical inquiry with socioeconomic issues in the designed activities. Positioning themselves as historical actors in relationship to the data supported assembly of counter-models that integrated macro and micro data. In turn, students were able to tell counter-narratives to challenge dominant discourses and ideas about society and world development. Through the modeling practices and storytelling performances, students developed personal relationships with the big data that made their stories consequential and relevant for themselves and their peers.

Finally, our analysis will continue to pursue a comparison of the two iterations of prospective mathematics and social studies teachers, respectively, and an understanding of tensions between these disciplines in school and in the world. As the prominence of aggregated data in public and professional spheres continues to grow, it has become more important for teachers and their students across disciplines to develop critical literacy practices with big data.

Making data open for public use and increasing transparency and access to data tools that permit authoring and manipulation are essential for individuals to be able to examine assumptions of models, pursue questions about measures, and consider their own relationships with the stories told with data. Public STEM learning opportunities with open data and technologies are thus necessary to sustain diverse public participation in democratic processes and scholarship.

REFERENCES

Al-Aziz, J., Christou, N., & Dinov, I.D. (2010). SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. Journal of Statistics Education, 18(3), 1-29.

Anderson, C. (2008, June 28). The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine. Retrieved from http://www.wired.com/

Bakker, A. (2004). Design research in statistics education: On symbolizing and computer tools.

Bakker, A., & Gravemeijer, K. P. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. B. Garfield (Eds.), The challenge of developing statistical literacy, reasoning and thinking (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bamberg, M. G. (1997). Positioning between structure and performance. Journal of Narrative and Life History, 7(1-4), 335-342.

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. Science Education,94(6), 1008-1026.

Becker, H. S. (2007). Telling about society. University of Chicago Press.

Berkowitz, M. W., Althof, W., & Jones, S. (2008). Educating for civic character. In D. Hess & P.G. Avery (Eds.), The Sage handbook of education for citizenship and democracy (pp. 401-409). Sage Publications Ltd.

Blum, A. (2006, February 21). Graphing the development gap. Bloomberg Business. Retrieved from http://www.bloomberg.com/

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.

Busch, L. (2014). A dozen ways to get lost in translation: Inherent challenges in large-scale data sets. International Journal of Communication, 8, 18.

Carrigan, M. (2014, August 9). The origins of methodological genocide: “All science is becoming data science” [Web log post]. Retrieved from http://sociologicalimagination.org/archives/15816

Cicourel, A. V. (1981). Notes on the integration of micro-and macro-levels of analysis. In K. Knorr-Cetina & A. V. Cicourel (Eds.), Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies (pp. 51-80). NN, NY: Routledge.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. The American

Mathematical Monthly, 104(9), 801-823.

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

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. Cognition and Instruction, 21(1), 1-78.

Cukier, K., & Mayer- Schönberger, V. (2013). The rise of big data: How it's changing the way we think about the world. Foreign Affairs, 92, 28-40.

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial 'big data' and geodemographics. Available at SSRN 2544638.

Davidian, M., & Louis, T. A. (2012). Why statistics?. Science, 336(6077), 12-12.

Dedeoglu, H., & Lamme, L. L. (2011). Selected demographics, attitudes, and beliefs about diversity of preservice teachers. Education and Urban Society,43(4), 468-485.

Derry, S. J., Pea, R., Barron, B., Engle, R., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J., Sherin, M., Sherin, B (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. Journal of the Learning Sciences, 19, 1-51.

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. The Journal of the Learning Sciences, 13(1), 77-103.

Donovan, K. P. (2012). Seeing like a slum: Towards open, deliberative development. Georgetown Journal of International Affairs, 13, 97.

Economist, T. (2010, February 25). The data deluge. The Economist. Special Supplement. Retrieved from http://www.economist.com/

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. Cognition and Instruction, 20(4), 399-483.

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.

Erickson, F. (2004). Talk and social theory: Ecologies of speaking and listening in every life. Cambridge, UK: Polity Press.

Esmonde, I. (2014). "Nobody's rich and nobody's poor… it sounds good, but it's actually not": Affluent students learning mathematics and social justice. Journal of the Learning Sciences, 23(3), 348-391.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. Qualitative Inquiry, 12(2), 219-245.

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi & J. B. Garfield (Eds.), The challenge of developing statistical literacy, reasoning and thinking (pp. 47-78). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. International Statistical Review, 75(3), 372-396.

Gay, G., & Howard, T. C. (2000). Multicultural teacher education for the 21st century. The Teacher Educator, 36(1), 1-16.

Giere, R. N. (1990). Explaining science: A cognitive approach. Chicago: University of Chicago Press.

Goffman, E. (1981). Forms of talk. University of Pennsylvania Press.

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). Erlbaum.

Goodwin, C. (1994). Professional vision. American Anthropologist, 96(3), 606-633.

Goodwin, C. (2007). Participation, stance and affect in the organization of activities. Discourse & Society, 18(1), 53–73.

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. Journal for Research in Mathematics Education, 443-471.

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? First Monday 16(2). http://firstmonday.org/ htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3316/2764. January 27, 2016.

Gutstein, E. (2006). Reading and writing the world with mathematics: Toward a pedagogy for social justice. New York, NY: Routledge.

Hacklay, M. M. (2013). Neogeography and the delusion of democratisation. Environment and Planning A, 45(1), 55-69.

Hall, R. (1999). The organization and development of discursive practices for “having a theory”. Discourse Processes, 27(2), 187-218.

Hall, R. (2000). Work at the interface between representing and represented worlds in middle school mathematics design projects. In L. R. Gleitman & A. K. Joshi (Eds.), Proceedings of

Twenty-Second Annual Conference of the Cognitive Science Society (pp. 675–680). Mahwah, NJ: Erlbaum.

Hall, R., & Horn, I. S. (2012). Talk and conceptual change at work: Adequate representation and epistemic stance in a comparative analysis of statistical consulting and teacher workgroups. Mind, Culture, and Activity,19(3), 240-258.

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

Hall, R., Stevens, R., & Torralba, T. (2002). Disrupting representational infrastructure in conversations across disciplines. Mind, Culture, and Activity, 9(3), 179-210.

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

Hammerman, J. K., & D TERC, C. (2009). Statistics education on the sly: Exploring large scientific data sets as an entrée to statistical ideas in secondary schools. IASE/ISI Satellite.

Holland, D., Lachiotte, W., Jr., Skinner, D., & Cain, C. (2001). Identity and agency in cultural worlds. Cambridge, MA: Harvard University Press.

Howard, T. C. (2008). "Who really cares?" The disenfranchisement of African American males in PreK-12 schools: A critical race theory perspective. Teachers College Record, 110, 954–985.

Hutchins, E. (1995). Cognition in the wild. MIT press.

Johansson, V. (2012). A time and place for everything?: Social visualisation tools and critical literacies. The Swedish School of Library and Information Science: The University of Borås.

Johnson, J. A. (2014). From open data to information justice. Ethics and Information Technology, 16(4), 263-274.

Johnson, S. J., & Amador, L. (2011). A pioneer in the use of video for the study of human social interaction: A talk with Frederick Erickson. Crossroads of Language, Interaction and Culture, 8(1).

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

Kahn, J. (2014). "What in the world?" Animated worlds in multivariable modeling with motion chart graph arguments. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. Lee, & L. D'Amico (Eds.), Learning and Becoming in Practice: Proceedings of the International Conference of 11th International Conference of the Learning Sciences (pp. 1649-1650). Boulder, CO: International Society of the Learning Sciences.

Kahn, J. B., Hall, R., & Phillips, N. (2014). Dissecting, remixing, and making graph arguments using motion charts and public data about global wealth and health. In the Power of Education Research for Innovation in Practice and Policy: Proceedings of the American Education Research Association 2014 Annual Meeting. Philadelphia, PA: American Education Research Association.

Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. Journal of the American Society for Information Science and Technology, 64(12), 2452-2467.

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. In L. English (Ed.), Handbook of international research in mathematics education (2nd edition) (pp. 49-72). New York: Routledge.

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

Kraemer, K.L., Dickhoven, S., Tierney S.F., & King, J.L. (1987). Datawars: The politics of modeling in federal policymaking. Columbia University Press.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. Academic Press.

Kuhn, T. S. (2012). The structure of scientific revolutions. Chicago: University of Chicago Press. (Original work published 1962).

Labov, W. (1972). The transformation of experience in narrative syntax. Language in the inner city, 354, 96.

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

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. Cognition and Instruction, 14(1), 69-108.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. American Educational Research Journal, 41(3), 635-679.

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

Lehrer, R., Schauble, L., Carpenter, S., & Penner, D. (2000). The interrelated development of inscriptions and conceptual understanding. In P. Cobb, E. Yackel, & K. McClain (Eds.), Symbolizing and communicating in mathematics classrooms (pp. 325-360). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lowenstein, K. L. (2009). The work of multicultural teacher education: Reconceptualizing white teacher candidates as learners. Review of Educational Research, 79(1), 163-196.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. Science Education, 96(6), 1071-1105.

Manz, E. (2014). Representing student argumentation as functionally emergent from scientific activity. Review of Educational Research, 0034654314558490.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

Noss, R., & Hoyles, C. (1996). The visibility of meanings: Modelling the mathematics of banking. International Journal of Computers for Mathematical Learning, 1(1), 3-31.

Ochs, E., Gonzales, P., & Jacoby, S. (1996). “When I come down I’m in the domain state”: Grammar and graphic representation in the interpretive activity of physicists. In E. Ochs, E. Schegloff, & S. Thomson (Eds.), Interaction and grammar (pp. 328-369). Cambridge: Cambridge University Press.

Ochs, E., Taylor, C., Rudolph, D., & Smith, R. (1992). Storytelling as a theory-building activity. Discourse Processes, 15(1), 37-72.

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.

Phillips, N. C. (2013). Investigating adolescents’ interpretations and productions of thematic maps and map argument performances in the media (Doctoral dissertation). Retrieved from etd.library.vanderbilt.edu.

Porter, T. M. (1996). Trust in numbers: The pursuit of objectivity in science and public life. Princeton UP.

Porter, T. M. (2012). Thin description: Surface and depth in science and science studies. Osiris, 27(1), 209-226.

Radinsky, J. (2008). Students' roles in group-work with visual data: A site of science learning. Cognition and Instruction, 26(2), 145-194.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. Visualization and Computer Graphics, IEEE Transactions on, 14(6), 1325-1332.

Robinson, D. G., Yu, H., Zeller, W. P., & Felten, E. W. (2009). Government data and the invisible hand. Yale Journal of Law & Technology, 11, 160.

Rogers, R. (2013). Digital methods. MIT press.

Rosling, H. (2006, June). Hans Rosling: The Best stats you've ever seen [Video file]. Retrieved from https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Rosling, H. (2007, March). Hans Rosling: New insights on poverty [Video file]. Retrieved from https://www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty

Rosling, H. (2009, August). Hans Rosling: Let my dataset change your mindset [Video file]. Retrieved from https://www.ted.com/talks/hans_rosling_at_state

Rosling, H., Ronnlund, A.R., & Rosling, O. (2005). New software brings statistics beyond the eye. In E. Giovannini (Ed.), Statistics, knowledge and policy: Key indicators to inform decision making, (pp. 522-530). Organization for Economic Co-Operation and Development.

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.

Rubin, A., Hammerman, J., & Konold, C. (2006, July). Exploring informal inference with interactive visualization software. In Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town, South Africa: International Association for Statistics Education. Online: www. stat. auckland. ac. nz/~ iase/publications.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. Cognition and Instruction, 23(1), 23-55.

Solórzano, D. G., & Yosso, T. J. (2002). Critical race methodology: Counter-storytelling as an analytical framework for education research. Qualitative Inquiry, 8(1), 23-44.

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

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1), 85-87.

Taylor, K. H., & Hall, R. (2013). Counter-mapping the neighborhood on bicycles: Mobilizing youth to reimagine the city. Technology, Knowledge and Learning, 18(1-2), 65-93.

Tufte, E. R. (1983). The visual display of quantitative information. Graphics Press, Cheshire, Connecticut.

Uprichard, E. (2013). Focus: Big data, little questions? Discover Society, 1. http://discoversociety.org/2013/10/01/focus-big-data-little-questions/.

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.

Wager, A. A. (2012). Incorporating out-of-school mathematics: from cultural context to embedded practice. Journal of Mathematics Teacher Education,15(1), 9-23.

Wilson, M. W. (2015). Morgan Freeman is dead and other big data stories. Cultural Geographies, 22(2), 345-349.

Winner, L. (1980). Do artifacts have politics? Daedalus, 121-136.

Wortham, S. (2000). Interactional positioning and narrative self-construction. Narrative Inquiry, 10(1), 157-184.

Wortham, S. (2003). Accomplishing Identity in Participant-Denoting Discourse. Journal of Linguistic Anthropology, 13(2), 189-210.

CHAPTER IV

GETTING PERSONAL WITH BIG DATA: THE ASSEMBLY OF FAMILY DATA STORYLINES

Introduction

Our paper reports on a design-based research study in which teenage youth built and told *family data storylines* at the city public library. Youth curated personal family mobility history while exploring public, large-scale socioeconomic datasets using online, data visualization tools. We, the research team, examine how digital data visualization tools and large-scale datasets, also known as "big data," mediate transdisciplinary, collaborative learning and youth's development into storytellers of both family history and broader socioeconomic trends.

The design and study of activities with publicly available, open big data and data visualization tools is a timely area for education and learning sciences research. The ubiquity of storytelling with big data in public conversations in the news media around important socioeconomic, scientific, political, and social issues (Segel & Heer, 2010) indicates that storytelling and modeling with big data are important STEM practices for participating in political discourse (Philip, Olivares-Pasillas, & Rocha, 2016; Philip, Schuler-Brown, & Way, 2013). Designing learning environments that engage youth and adults in storytelling with socioeconomic big data and new tools is important for more sophisticated, critical, and relevant modeling of socioeconomic phenomena as well as civic engagement.

The pairing of storytelling and modeling not only recognizes the role of models in the development of theory (Koschmann, 2011; Star & Griesemer, 1989) but also takes up an approach to narrative activity as theory-building (Ochs, Taylor, Rudolph, & Smith, 1992).

Stories are like theories in that they are explanatory in nature and vulnerable to challenges; versions of experience or observation can be questioned and other speakers can offer alternative causal explanations. Moreover, narratives, such as narratives about family history, report on past events and circumstances that inspired a decision, an action, or change and, consequently, establish links between personal and social worlds. Stories describe interpersonal relationships and community memberships; in this way, stories have "the power to interface self and society (Ochs & Capps, 1996, p. 31).

In the present study, youth were tasked with developing theories for why their families moved, using big data and dynamic visualization tools. Youth storytellers positioned themselves and their families in relation to past, present, and imagined worlds represented by the data. Big data and dynamic modeling tools offered new modalities for theory-building—that is, for both explaining and challenging stories told about the self and society. If one can interact with the data, one can interrogate a story and investigate, verify, or elaborate the narrative.

This paper reports on a single design iteration of a multiyear design-based research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) project focused on storytelling and modeling with open socioeconomic big data. The overarching object of inquiry of the broader project is to understand how youth and adults learn to engage in the interdisciplinary representational practices that support becoming modelers, storytellers, and consumers of stories told with big data. The investigation described in this paper focuses on how learning and modeling with big data can deepen the relationship between the self and society and, in turn, support critical understandings of that relationship. Our research builds on the earlier design iterations that found that *getting personal with big data* can support critical modeling and storytelling about the social world (Kahn & Hall, 2016). That is, relating one's personal experiences to the social phenomena described by the aggregate data entails conceptual shifts in spatial, temporal, and social scales. Subsequently, shifts in scale permit critical reflections on social, economic, and scientific phenomena that examine relations of power and equity among individuals and social entities (Gutstein, 2006; Philip et al., 2016).

The present study considers the relationship between the individual and society as central to engaging in critical, representational practices with big data. Unlike the previous design iterations, this work explores storytelling and modeling with big data in a context that is explicitly personal: the family *geobiography*, which is the multigenerational history that traces how one's family came to be where they live today. Our research question for the current investigation is as follows: How does scaling personal experiences into aggregate data facilitate learning to critically model and tell stories about oneself, one's family, and society?

Data Science: Relevant for Youth Learning

Big data comprise a sociotechnical phenomenon that has advanced the interdisciplinary fields of data science, analysis, and visualization (Busch, 2014). While data has always informed scientific and social-scientific practice (Porter, 1996), the nature of tools and data has changed. The vast amount of scientific and social-scientific data that can easily be computed, stored, analyzed, and shared is unprecedented (Busch, 2014; Hammerman, 2009), and the data visualization and analysis tools available offer novel forms of representation and interactivity (Al-Aziz, Christou, & Dinov, 2010; Segel & Heer, 2010). With the growth of big data and digital tools, the cost of social scientific modeling with big data has decreased and generated new ways of analyzing and tackling socioeconomic and scientific problems (Venturini, Jensen, & Latour, 2015). Big data increasingly inform public and private discourse and decision-making (Davidian & Louis, 2012), and data science has likewise become as an interdisciplinary professional field.

Storytelling and modeling with big data has emerged as a new form for telling about society (Becker, 2007) across disciplinary fields. Telling stories with big data tools as a way to model social and scientific phenomena is a ubiquitous practice across commercial, scientific, and, notably, media sectors (Segel & Heer, 2010). From cable television newscasters covering election cycle poll data to TED Talks to news websites, public conversations in the news media about socioeconomic and scientific issues increasingly involve narratives told about society with big data visualizations (Kosara & Mackinlay, 2013; Segel & Heer, 2010).

We argue that storytelling and modeling with big data is now more widespread and conventionalized and thus constitutes a new *cultural activity* (Engeström & Sannino, 2010) that can be observed, learned, and studied. Storytelling and modeling with big data also constitutes a prevalent form of participation in civic and democratic discourse (Philip et al., 2013). As big data and visualization software becomes increasingly open and accessible to professional and public lay users, opportunities for "critically-informed [data] use" (Dalton & Thatcher, 2014) are materializing. Open data, released by city, state, and federal governments, NGOs, and research institutions has led to new forms of accountability for city, state and federal governments and institutions (Cukier & Mayer-Schönberger, 2013). Undoubtedly, the perils of power asymmetry and lack of data privacy are real (Johnson, 2014). Big data, like other technologies, is no equalizer (Selwyn, Gorard, & Williams, 2001), designed by a typically white, male workforce who predefine the rules for others' use (Daileda, 2016; National Center for Education Statistics, 2015; Klein & Kleinman, 2002; Philip et al., 2013). Open data, in practice, often serves to "empower the empowered" (Donovan, 2012, p. 99). Nonetheless, open and free government data and software could lead to important discoveries about communities, such as the distribution of public services, law enforcement, and investment infrastructure and development (Gurstein, 2011; Johansson, 2012). The advance of open big data has increased usership and access to data and tools that support asking and answering new questions about the social world (boyd & Crawford, 2012).

Currently, there is a small but growing area of developmental and learning sciences research that is exploring both the utility and design possibilities for youth learning and development with big data and digital visualization tools (e.g., Enyedy & Mukhopadhyay, 2007; Erickson, 2012; Philip et al., 2016; Rubel, Hall-Wieckert, & Lim, 2017; Rubel, Lim, Hall-Wieckert, & Sullivan, 2016; Wilkerson-Jerde & Laina, 2015; Williams, Deahl, Rubel, & Lim, 2015). Our paper contributes to this collection of work. We are interested in the critical consumption of stories told with big data—or *critical data literacy*. By critical, we draw on Gutstein's (2006) definition of critical literacy, which calls for "examin[ing] one's own and others' lives in relationship to sociopolitical and cultural-historical context" in terms of distributions of power, equity, and social justice (p. 5). We also agree with Philip et al. (2016) that data literacy must entail attention to structures of race and ideology embedded in the data. Particularly because data can mute processes of discrimination and support the (false) assumption that categories like race are static, stable designations (Autry, 2017), critical perspectives and reflections towards big data are important for indexing the relationship between individual experiences and larger social, economic, and historical issues described by the data.

A Framework: Theoretical Commitments Towards Learning

In line with Philip et al.'s (2013) framework for "Learning About Big Data for Democratic Participation," we offer here a series of commitments to theories and approaches to studying learning with technology. First, this research follows an approach to learning as socially, culturally, and historically situated in activity (Lave & Wenger, 1991; Wertsch, 1998).

Our personal and shared cultural and political histories shape how we, as individuals, learn and become participants in activities. Second, participation in data modeling and family storytelling activities is distributed and embodied (Hall & Nemirovsky, 2012; Hutchins, 1995; 2006; Wilson, 2002); it entails the coordination of multiple bodies, tools, and artifacts across contexts. Additionally, we take a perspective that considers the present, past, and future relations among social actors, technologies, and the learning environment (Bang & Vossoughi, 2016; DiGiacomo & Gutiérrez, 2016; Gutiérrez & Jurow, 2016). This kind of relational perspective is important in pursuing designs for learning that are maximally inclusive for linguistically, racially, and culturally diverse learners and that contribute to an equitable and socially just democracy.

We also approach family storytelling with big data as a representational, sociotechnical activity (Stevens & Hall, 1998; Goodwin, 1994), albeit also an interdisciplinary activity. Our focus on knowledge-in-use in interaction (Hall & Stevens, 2016) considers learning from the learner's point of view (Stevens, 2010). This approach assumes that we can systematically and rigorously describe knowledge, learning, and participation in local, everyday activities (Geertz, 1983) by paying attention to discourse, gesture, and tool use in context. Azevedo and Mann (2017) describe this as an *interactionist perspective,* in which the social organization of bodies and materials is central for cognition, learning, and participation in activities.

We are also committed to honoring learners' capacities, dispositions, and histories as well as to positioning youth with agency to create and share new knowledge (Vossoughi, Hooper, & Escudé, 2016). We aimed to design opportunities for youth to pursue social and scientific inquiry with big data through both consumption and authorship, or what Gutstein (2006) calls "reading and writing the world." We approach storytelling and modeling with big data as a context for making or producing new knowledge and *inscribing* (Taylor, 2017) society. This framing also draws on constructionist theory that posits that learning and knowledge-building occur through the production or making of sharable artifacts (Peppler, 2013).

This work responds to the call to engage in design-based research that speaks more clearly to distributions of equity and power in terms of the technology and the research context (Gutiérrez & Jurow, 2016). Our study considers how authority and power feature in public large-scale datasets and how this affects youth interaction with the data. Every open socioeconomic dataset has qualities that obscure both aspects of the data and aspects of the phenomena the data is measuring or quantifying (Busch, 2014). Leonelli, Rappert, and Davies (2017) suggest users must contend with *data shadows*, which describe the ambiguous and stealthy nature of big data visualizations in both revealing and hiding particular kinds of data and stories. Data stakeholders, whose ontological structures are embedded in datasets and tools, as well as the *residual categories* of lived experiences that do not fit neatly into data classifications systems (Bowker & Star, 2000) elicit data shadows.

Lastly, our activities locate both participants and researchers at what Tuck, Smith, Guess, Benjamin, and Jones (2014) and Radinsky (personal communication, June 11, 2017) describe as a “threshold” for transforming data into stories and stories into data. They warn that at this threshold, as participant stories become data for social science researchers, the origins and intent of those stories can be lost. In our study, the participants, like young social scientists, are tasked with aligning family stories with big data and big data with family stories, and the research team is subsequently tasked with producing a narrative of this process. We consider the data–story threshold to be a challenging and productive conceptual place for both participants and researchers to study the relations between personal and shared human experiences.

Context

Getting Personal With Big Data: The Self and Society

This study is part of a larger design-based research program (Cobb et al., 2003) that is focused on how youth and adults learn to engage in the interdisciplinary representational practices that support becoming modelers, storytellers, and consumers of stories told with big data. Our previous design iterations (Kahn & Hall, 2016) found that connecting personal experiences to aggregate trends described in the model can support telling stories about society that counter, challenge, or critique dominant or conventional social narratives. In turn, the idea that getting personal with big data was important to critical modeling and storytelling about the social world was something we wanted to examine more deeply.

The focus on the relationship between the self and society is not new to the social sciences. For instance, Cicourel's (1981) theory of integrating micro and macro analysis levels for sociology studies suggests that neither micro nor macro structures are self-contained units of analysis; rather, macro social issues should be viewed as aggregated micro events or everyday, routine exchanges. At the same time, fully understanding any single social interaction always entails considering the larger, macro context. In turn, relating individual, personal experiences to models of aggregate data seems both commonsense and generative for making critical and meaningful sense of social life being modeled by big data. Getting personal with big data is thus both a practice of modeling with data and engaging in social analysis.

Our workshop intended to support participants in making comparisons across times and places with the data tools, which we see as central to critical storytelling and modeling with big data. Ragin’s (1987/2014) description of the case-oriented comparative method in social science goal guided this design. According to Ragin, social scientists engaged in this research tradition have a two-fold task: historical interpretation and casual analysis of social phenomena. To do this, social scientists must assess “causal complexity;” that is, they must evaluate and untangle combinations of conditions that produce large-scale change. Like social scientists, in asking youth to determine and tell why their family moved, they had to consider the conditions of cases —the cities, states, and nations in which they or family members lived—within a wider context—the broader social history at the time. On the other hand, not only did participants work to identify similarities and differences at the macro social or society level (e.g., comparison of cities, states, and nations), but they also considered the outcomes of social phenomena at a personal scale to explain why their family members or ancestors moved.

We call the conceptual shifting and traversing between personal, local experience and the society-level phenomena *scaling* (Hall & Leander, 2010). New dynamic spatial analysis and modeling tools support conceptual movements in time (i.e., past, present, and future), in space (i.e., local, national, and global bounded geography), and in social activity (i.e., between what I do and what collective society does) (see Figure 1). We argue that scaling practices are integral to critically storytelling and modeling with big data. Access to open big data and modeling tools makes it possible to embed narratives of individual experience that describe my body (e.g., Lee, 2013), my family, my home, or my childhood within narratives of aggregate, human social participation spanning histories that describe many bodies and generations of people within and across nations. In turn, this creates opportunities for considering the ethics and politics of both one’s own decisions and society’s choices.

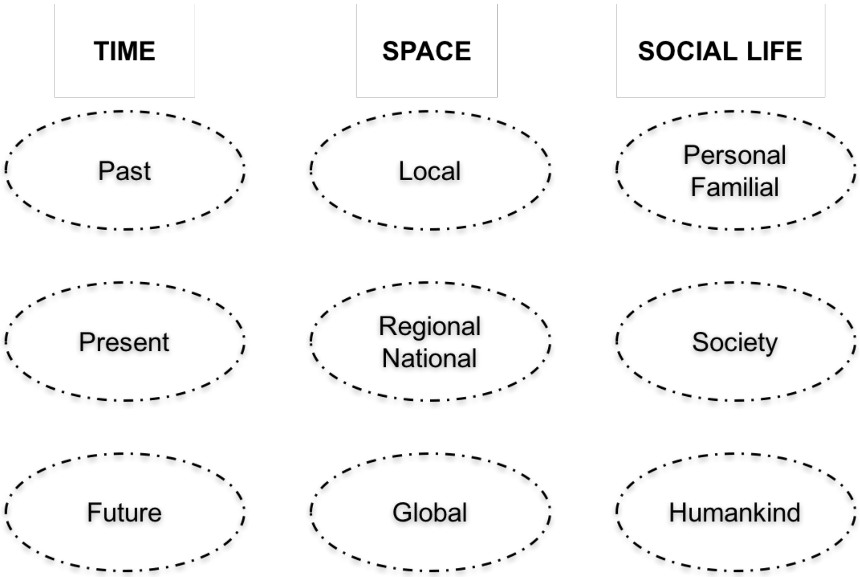


Figure 1. Scaling entails the conceptual movements or shifts across and between different scales of time (past, present, future), space (local, regional/national, global), and organizations of social life (personal/familial, society, and humankind). Big data and novel visualization tools permit comparisons that involve temporal, spatial, and social scaling.

A handful of learning sciences studies have explored how relationships between personal experiences and big data with dynamic, digital visualization tools can support learning and critical inquiry. For example, Rubel et al. (2016) and Rubel et al. (2017) looked at how narratives of students' personal experiences and their neighbors compared with the broader stories told by official GIS maps of demographic and socioeconomic data with regards to the local distribution and role of alternative financial institutions (e.g., pawn shops). The interactivity and "zoomability" of the GIS tool also supported students in noticing patterns in inequities in their neighborhood and across the city. Wilkerson-Jerde and Laina (2015) studied how middle school youth coordinated mathematical, representational, and personal community knowledge in order to assemble data visualizations of public city data and city demographic composition with Google's suite of tools. The data science activities prompted youth to raise questions about social and scientific matters affecting their city, such as racial diversity and public land use. Polman and Hope (2014) and Polman et al. (2016) looked at the relationship between societal concerns and personal concerns in their studies of youth who create scientific data narratives based on personal interests in health or other science fields (e.g., effects of caffeine or smoking, climate change, high-school students sleeping habits). These research designs placed youth's local experiences, daily interactions, and interests in conversation with a society-scaled perspective of the phenomena the students were exploring with (mostly big) data.

We also see getting personal with big data as related to the practice of "ground truthing" data models (Goldstein & Hall, 2007; Pickles, 1995, 2006; Taylor & Hall, 2013). Ground truthing is the practice of using lived experiences to evaluate and challenge official, typically remotely gathered (e.g., satellite) representations of data. For example, ground truthing describes when scientists index pastime field experiences to data models in ways that evaluate model assumptions and missing data as well inform decision-making (Goldstein & Hall, 2007). Taylor & Hall (2013) and Taylor (2017) present a case of youth residents engaged in ground truthing with critical and political ends: Youth used GPS track data of their everyday mobility experiences to challenge and ultimately change official city maps that did not take into consideration their mobility constraints and transportation needs.

The study described in this paper shares the quality of bringing the personal into contact with "official" information or data and similarly sought to elevate cultural and local experiences in credibility and authenticity. Just as ground truthing can assume the critical form of *counter-mapping* (Taylor & Hall, 2013; Taylor, 2017), getting personal with big data can support *counter-storytelling* (Kahn & Hall, 2016; Nazario, 2006; Solórzano & Yosso, 2002) and *counter-modeling* (Kahn, Hall, & Pearman, 2016) in order to critique dominant or conventional narratives, which could potentially lead to *counter-data actions* (Dalton & Thatcher, 2014). Kahn & Hall (2016) present cases of undergraduate and graduate students who, as storytellers and modelers, actively *positioned* themselves and their peers (Bamberg, 1997) in relationship to the narrative they produced about society with the big data. In their models of nation-state carbon-dioxide emissions, the students assumed a critical, moral-based orientation towards the data by positioning themselves and their peers as culpable in the production of harmful greenhouse gasses. The students further suggested that their contributions were not accurately represented in the model.

For the study reported here, we also designed with the potential for supporting critical perspectives and moral stances towards the data. We think about this critical positioning process in terms of a conceptual matrix (Figure 2), in which one's experiences or trajectory can match, exceed, or contrast the big data trends, which are moving in a value-assigned direction—positively or negatively—as interpreted by the modeler. We conjectured that participants would determine if their family life story fits or (perhaps unjustly) contrasts with the aggregate social change demonstrated by the data over time and would position themselves and their families' socioeconomic trajectories somewhere in this matrix. That is, we imagined that youth would attempt to align their and their families' experiences with social data trends, and the results would be one of the following: Both my life and the aggregate have improved (+, +), we did socioeconomically better than the aggregate (+, -), both my family and the aggregate have struggled (-, -), or my family experienced obstacles that the aggregate did not (-, +). Subsequently, we looked to see if any counter-narratives or counter-models emerged in the exploration of the aggregate and themselves, and if so, we considered what were they counter to and from whose perspective. As part of our descriptive, analytic report, we assess to what extent our design-based research program elicited these kinds of critical responses and address why youth did or did not take a critically-oriented approach.

Critical Positioning			
My (family’s) life story	Aggregate data trends		
		Data +	Data -
	Me +	All is swell	I beat the odds
	Me -	My life is a struggle	My challenges are normal

Figure 2. This conceptual matrix represents the critical positioning process that was used to inform our design and data analysis. We use critical positioning to describe to what extent students presented their personal or familial experiences as corresponding to or departing from the conditions of the aggregate.

Family Storytelling and the Geobiography

For this design study iteration, we chose a context that we felt was unambiguously personal: the family geobiography or family mobility history. Using geobiography as a concept for design and teaching extends back to our earliest work. The previous study iterations were grounded in ongoing observational studies of professional storytellers and modelers with big data (Kahn, Hall, & Phillips, 2014). We watched Hans Rosling, a public health professor and statistician who created the Gapminder data tool, perform “grandma-verified statistics” (Rosling, 2007), in which he aligned the birth of his maternal relatives in Sweden to the conditions of nation states today as represented by global socioeconomic big data. Using his Gapminder tool, with GDP on the x-axis and life expectancy on the y-axis, he compared the health and wealth of nation states today to Sweden in the years when his great-grandmother, grandmother, mother, daughter and granddaughter were born. For example, Hans animated his model of global development as follows:

> And this is my family. This is Sweden, 1830, when my great-great-grandma was born. Sweden was like Sierra Leone today. And this is when great-grandma was born, 1863. And Sweden was like Mozambique…and I -- who am I? I'm the Mexican in the family (Rosling, 2007).

In analyzing Hans' performance, we discovered that the family geobiography could be scaled to large-scale data (Kahn & Hall, 2016) and conjectured that this might provide for a captivating way for youth to engage these data.

We subsequently piloted this use of the geobiography concept in experimental teaching in two one-day programs for middle school youth and in a residential summer program for high school youth at a public library. In these mini design iterations, youth were asked to trace and present their family mobility histories using Google maps. In these cases, we found the family geobiography to be personal, accessible, and meaningful for youth participants.[14] Migration and relocation comprise significant events in a person's lifespan; moving from place to place can be stressful and emotional (Chow & Healey, 2008). Relocating or moving typically leads to transitions in terms of relationships, community memberships, networks, and ways of seeing the self (Schlossberg, 1981).

Tracing and sharing our family mobility histories also was a compelling ice-breaker for the teens. Asking about one's geographic origins is routine in conventional conversations with new acquaintances (Svennevig, 2000), and sometimes these exchanges are fraught with racial overtones (micro-aggressions; Sue et al., 2007) that challenge identities and cultural heritage. The geobiography as both an expression of identity and a site for racial contestations in social interactions was not lost on our participants or their families, as we found in the family follow-up

[14] For a reference, see the opening excerpt of Chapter 2 of this dissertation, which was taken from a couple of short design experiments housed within a human geography summer residential course for rising 9th and 10th graders. Activities included youth discovering their own geobiographies, finding stories told with data in the news media that they felt personally connected to, and exploring local city neighborhoods using US Census data to support spatial narratives about those communities (Kahn, Hall, & Pearman, 2016). The high school residency program took place in Summer 2015; the design iteration of focus in this paper took place in Summer 2016.

interview with two of the participants, Naimah and Isis[15], and their parents, Zuri (Mom) and Akinjide (Dad). Prior to Excerpt 1 below, Researcher 1 had asked Zuri (Mom), “How does your daughters’ lives compare to yours when you were their age?” In response, Zuri described her parents’ choice to become members of the African-American Islamic party in the 1960s as an effort “to look backwards to reclaim what was lost.” Zuri subsequently explained, as captured in the excerpt, that this family history is why Zuri and her siblings, and now Zuri’s daughters, have African cultural names, which regularly leads to difficult encounters with racial biases, stereotypes, and assumptions.[16]

Excerpt 1
[Family follow-up interview]

1 Zuri: ((*looking at daughters*)) and so you all I think are an extension of that in that you may meet people who will say is that your real name? because there's still this idea that

2 Akinjide: tell me where you're really from

3 Zuri: yeah you all tell me all the time people say where you from? right?

4 Naimah: I'll be like Oaktown, and they're like [no, where you from?

5 Zuri: [no no no where you from? and so you are prepared for in a solidified way this identity shift that you all know that YOU celebrate Kwanzaa when we got to Kwanzaa community events you've mentioned before you feel those are your people like ((*impersonating one of the daughters, voice shifts*)) everyone here has an African name, everybody here has natural hair, remember talking about that?

6 Isis: ((*nodding*)) hmhmm

7 Zuri: at the village church, oh we're not weird, that's what I think one of you said, we're not weird here, we feel good here, so:: but I think at the same time, what you might have less of, what I hope you have less of than I did growing up, is feeling that you are sort of on the outside of what's considered cool

8 Akinjide: yeah, marginalized

[15] Participant, family, and library staff names are pseudonyms, as are the city and county name. Participants and families chose their own pseudonyms to help us stay truer their identities; this is the case unless noted otherwise.

[16] Transcript conventions for excerpts are as follows: CAPITALS indicate emphasis; ((*Observer notes*)) indicate significant gesture or expression; ? indicates rising intonation; ! indicates exclamations; [indicates overlapping talk; = indicates latched talk; ::: indicates elongated vowels; periods . or commas , indicates pauses less than a second; longer pauses are indicated with observer notes; […] indicates redacted talk.

9 Zuri: yeah, like so, yeah I hope that for you all, I don't really know

The “really” in “tell me where you're really from” challenges the family’s structuring of names and identities within a deliberate African history; at the same time, Zuri countered the marginalization that her daughters experience in everyday conversations by reminding her daughters that there is a community that they belong to that celebrates Kwanzaa and embraces African cultural heritage like them. We include this illustration upfront to demonstrate how the geobiography, as a context for storytelling and data exploration, (a) affectively resonated with adults and youth, (b) sometimes invoked very serious reflections on society’s prejudices, and (c) simultaneously revealed the family’s efforts to counter those prejudices.

We also chose the family geobiography because youth are knowledgeable and active participants in family storytelling activities, whether during dinner table conversations, road trips, or family reunions at holidays. The sharing of family stories and histories is an important site for youth learning, development, and wellbeing, particularly for adolescents (Duke, Lazarus, & Fivush, 2008; Fivush, Bohanek, & Zaman, 2011; Linde, 1993; Taumoepeau & Reese, 2013). We are also interested in family storytelling as an intergenerational activity because intergenerational figurations are central to adolescent youth learning ecologies (Barron, Martin, Takeuchi, & Fithian, 2009).

Finally, we felt that family migration was a relevant, timely issue to focus on in light of an ever-growing number of displaced persons in the world (UNHCR, 2015), national elections that have produced volatile discourse around immigration around the world, and changing demographics in US cities and neighborhoods (“A Census Time Machine,” 2017). In particular, we hoped that our research would contribute to understanding the role of data in public conversations regarding migration. We conjectured that narratives of familial migration, if told in relation to big data, would evoke powerful and important dialogue between local and global perspectives.

Methods

Data Collection

We conducted a design-based research study (Cobb et al., 2003) to investigate our research question: *How does scaling personal experiences into aggregate data facilitate learning to critically model and tell stories about oneself and society?* Our focus on the relationship between local, personal experiences and aggregate trends developed from two earlier design study iterations (Kahn & Hall, 2016) that found that getting personal with big data in modeling activity—shifting across scales of time, space, and social life in discourse and model animation—can facilitate critical perspectives towards the social, economic, and historical issues described by the big data. In the first two studies, preservice secondary social studies and mathematics teachers were asked to tell stories and assemble models with open public data about global socioeconomic development. For the study described here, instead of asking participants to model global health and wealth, we implemented a *design alternative* (diSessa & Cobb, 2004) to explore getting personal with big data as a new theoretical category of representational, sociotechnical work. That is, we chose a personal context for modeling with big data—the family geobiography—to direct attention to the individual–aggregate relationship. We believed that a focus on family migration history in relation to national and global socioeconomic and demographic data would prompt scaling between individual or local experiences and society-level phenomena as a part of modeling activity.

Setting. The study took place in a free summer workshop at a public library in a mid-sized Southern city. The workshop was entitled “Storytelling With Big Data: Tracing My Family to Oaktown." The program was sponsored in partnership with the library's special collections staff and was designed as a "pop-up" environment (comparable to a mobile maker-space; Krishnan, 2015; Peppler, Halverson, & Kafai, 2016) that could be reproduced in similar kinds of spaces or libraries. There were three sessions over the course of 3 consecutive weeks; each lasted 2 days (Tuesday, Thursday) for 5 hours each day (9:30 AM-2:30 PM). The library provided laptop computers for each participant with Internet access. The library also provided software access (e.g., a professional version of the data tool, Social Explorer, and the ubiquitous Microsoft PowerPoint). If participants did not finish their projects, they were welcome to return the following session. The 3 weeks culminated in a one-day community exhibit at the library in which participant projects were put on display; parents attended the event, and the exhibit was open to the public.

Our decision to design for learning at the local public library was intentional. First, we recognize the growing role of libraries in community responses to social change and as a place for family and community gathering as well as informal learning (Rogoff, Callanan, Gutiérrez, & Erickson, 2016). For many youth, libraries have become central nodes in their interest-driven learning ecologies (Barron, 2006; Barron, Gomez, Martin, & Pinkard, 2014). Second, our city library, and specifically the special collections division, had been a longstanding partner for our research lab; they are committed to creating opportunities for youth to develop and deepen relations to the city and to giving voice and recognition to underrepresented populations.

The special collections archivists have recorded hundreds of oral histories in order to preserve a city public history that is both inclusive and representative of the city's increasingly diverse demographics. Their collection includes a set of oral histories from residents of Oaktown who were born in other countries, 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 our project, we wanted youth to contribute their stories to the city oral history archives. We also wanted to design a program that could be easily replicated or "popped up" in another library or a similar public community space. While our design could be adapted to work in schools, the current instructional climate in public education (e.g., the required adherence to curriculum standards, emphasis on testing, and disciplinary silos) is not entirely amenable to broadening what counts as acceptable data modeling activity.

Participants. Library staff recruited middle school and high school youth through personal connections, community partners, and public listservs. Participants consisted of 17 diverse middle and high school youth, between ages 10–16, mostly identifying as multi-generational African-American. Out of these teens and tweens, 12 participants composed 6 different sibling pairs. These sibling arrangements were not a result of intentional recruitment. Rather, summertime is when there are opportunities for intergenerational and familial learning. The workshop was conducted in weekly sessions, and participants attended the workshop for single or multiple sessions (see Appendix B). The workshop was free, and the research team provided lunch.

While our earlier design iterations involved university undergraduate and graduate students, we chose to work with adolescents to better understand how youth learn to engage in data science activities. While youth develop capacities to organize narratives by ages 9–10, youth struggle with structuring expository and argumentative discourse and text until high school (Berman & Nir-Sagiv, 2007; Uccelli, Phillips Galloway, Barr, Meneses, & Dobbs, 2015). Likewise, we also seek to learn how adolescent youth organize or restructure family narratives as they incorporate analyses of historical, demographic, and socioeconomic big data into their stories. We are interested in understanding the conceptual and representational challenges that youth might encounter in storytelling and modeling activities and what kinds of socioeconomic and demographic data would be appealing or interesting to youth.

The research team consisted of Researcher 1 (lead author, identifying as White, European origin, Jewish, female), Researcher 2 (a university faculty identifying as White, European origin, male), and two research assistants who were both Masters students in education, one of whom was in a certification program to become a secondary social studies teacher (identifying as White, Hispanic origin, female), and the other who was a Fulbright scholar from the Middle East (identifying as White, Middle Eastern origin, Female). Library staff that supported workshop activities on various days consisted of four women (two identifying as African American and two identifying as White, European origin). We identify the research team and library staff's ethnicities/races and genders here in order to first acknowledge the possibility that certain stories were not told (or models assembled) in the workshop because of differences between participants' racial or ethnic identities and those of the research team, and second, to recognize that this project context was compelling for the research team as well in terms of their own reflections on family ancestry and points of origin.

Instructional design. In order to learn more about how getting personal with big data could foster youth learning to model and tell stories about society, our study engaged in experimental teaching activities around an explicit relation between aggregate trends and personal, individual experience. The workshop goal for each participant was to assemble a *family data storyline* to explore the relation between national and global migration trends (the answer to the question of "What moves families?") and personal experiences (the answer to the question of "What moved my family?").

A pre-assignment was emailed to enrolled participants the week prior to their participation in the workshop. The pre-assignment asked participants to talk with their families about their family mobility histories—where have they lived since they were born, where did their family come from, and how did their family get to where they live now? We asked participants to trace back to their great-grandparents' generation if possible and invited participants to include earlier ancestors if they wished. We also asked participants to create a list of places and people in historical order, kind of like a place-based family tree, and we encouraged participants to bring to the workshop images or other artifacts, physical or digital, to use when telling their stories. Participants who did not complete the pre-assignment were given time to do so during the first day of the session; they were encouraged to call and text parents and relatives using phones and to talk to their parents during the intermediary day before they returned for the second day of the workshop. We hoped families would discuss why generations of their family moved, what was going on in history when they moved, and if there were consequences of moving for their family (i.e., improvements or challenges). We returned to these questions consistently throughout each workshop session.

Workshop activities conceptually moved participants back and forth between thinking about their families' experiences and the social aggregate. We introduced each session with a presentation of examples of family data storylines made by the research team. We then engaged participants in a "What Moves Families?" brainstorm in which the youth participants generated broader reasons for migration (Figure 3). We also gave tutorials on using two open data tools on the first day of each weekly session.

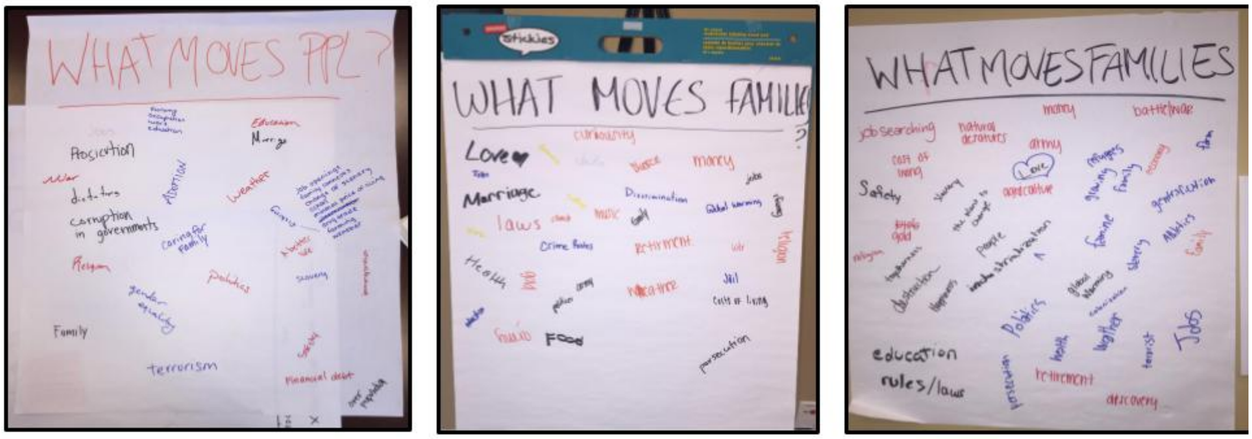


Figure 3. “What moves families?” brainstorm results from Sessions 1, 2, and 3 (left to right). Youth participants collectively generated and recorded broader reasons for why families move. This activity led each session, before youth engaged with the data visualization tools.

Once or twice in each workshop session, we started the day with a “four walls” game that set up comparisons between familial experiences and demographic or socioeconomic categories. Teens, researchers, and librarians participated. Questions leading these games asked the participants and adults to think about the historical social conditions that motivated or forced their families to move and that possibly could be explored with our data tools. Our questions set up comparison structures between nations or states as well as between ancestors and current or projected future experiences of the youth. We asked questions such as: “My family moved to the United States for?” (Wall 1: Education; Wall 2: Jobs/work; Wall 3: Marriage, Wall 4: Other); “I believe the economy of the US is better than that of my family's country of origin” (Wall 1: True, Wall 2: False); “How far along in school did your family member go?” (Wall 1: Grade school, Wall 2: High school, Wall 3: College, Wall 4: No formal schooling), which was followed by “How far along do YOU expect to go?” or “How many siblings did your family member have?” and “How many siblings do YOU have?” These questions set up intergenerational comparisons of society-level socioeconomic indicators (family size, educational attainment, occupation, neighborhood racial diversity) between family members and the participants. We conducted both class-level discussions and sharing with wall-mates after each question.

Since each weekly session served as a mini-design cycle iteration within the larger program, some workshop activities varied slightly between sessions. In our first activity in Session 1 Day 1, participants and researchers calculated the distance from where their ancestors came from to Oaktown using a website that calculates flight distance in miles between places; we each placed ourselves using a sticker on a space/time graph, with the x-axis representing the year of our earliest known family member or great-grandparent in the US, and the y-axis representing the distance between where the ancestor lived and Oaktown. In Session 1, we also asked participants to forage for or find big data stories in online news media as homework and to share and discuss their found articles with the group on Day 2. In Sessions 2 and 3, we performed a walking-scale timeline that, like the four walls game, asked participants to stand in for the family in historical time; we repeated this embodied activity with parents in the family and community exhibit (Figure 4). Additionally, in Session 2, we had a “big data challenge” activity in which participants had to take a stance on whether someone should move to Oaktown or not and support their stance with data in Social Explorer. Finally, a few questions differed in the four walls game across sessions.

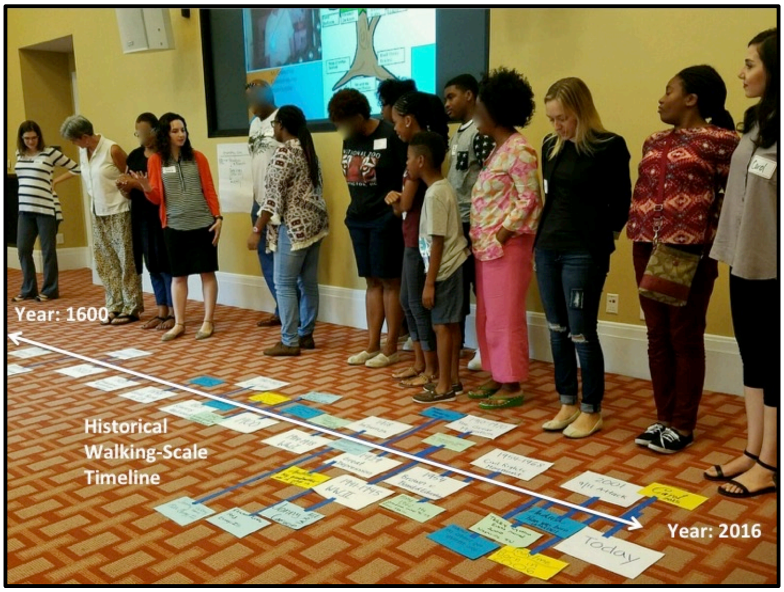


Figure 4. Walking-scale timeline of historical ancestry with parents, teens, library staff, and research team members during the public community exhibit.

Students also contributed oral histories to the city library's Special Collections public archives about their family history and individual experiences. Typically, this took place during the first afternoon of each weekly session. They were recorded "StoryCorps® style," in which oral histories are conducted as conversations between friends, colleagues, or family rather than interviews. One of the library staff introduced the oral history tradition to the participants, which she described as "history from your mouth." Participants also watched two official StoryCorps® animations as models and then practiced asking/answering each other follow-up questions. Participants conducted interviews with peers in the workshop close in age (siblings were separated); we provided them a set of guiding questions to ask (see Appendix D), which participants mostly followed, and encouraged participants to ask follow-up questions. They asked each other questions like: How did your parents' childhoods differ from your own? What was the major thing that moved your family? Do you remember any classic family stories? Each youth interviewed a peer for approximately 10 minutes. Library staff and researchers emphasized that these oral histories would be included in the city's public history archive forever. The library staff made each oral history available for the participants on the second day of each session; we gave participants time to listen to their records and support for making selections (using a media splicing tool) for their presentation if they desired. Consequently, this durable archival object became a base for building their maps or models.

The remainder of each workshop day was unstructured and dedicated to exploration and assembly of the family data storyline. This involved: (a) choosing a side of the family to focus on; (b) choosing one of two web-based data tools accessed via the laptop's Internet browser—Social Explorer, a historical thematic mapping tool that accesses US demographic data, or Gapminder, a multivariable dynamic graphing tool that uses public global socioeconomic data—both described in further detail below; and (c) selecting indicators or variables from each tool's available datasets, like education level or household income in the US Census. Participants then captured final screenshots of models or maps and inserted them into a Microsoft PowerPoint. Models or maps were often accompanied by slides or text that explained participants' data selections and what they learned from the data. Participants also added slides that introduced themselves; provided additional information about their family mobility histories, including family trees, anecdotes, and images; listed their future aspirations (e.g., what they wanted to do when they were adults); and raised new questions for future inquiry.

These family data storylines were refined over workshop activities. The research team was consistently available for one-on-one assistance and was deliberate about asking participants questions as well as providing suggestions for next steps. In terms of tangible scaffolds, we also provided a matrix (Figure 5) to participants in each session in the form of a worksheet. This matrix served as a focusing device to help explore temporal, spatial, and social comparisons in the data as well as align family stories to aggregate social history. This matrix, which juxtaposes dimensions of time, place, and social life for the self in relation to society, was a central, working concept that informed the design of the instructional and learning environment. The worksheet also listed questions that we intended to deepen participants' understanding of their family story in relation to the historical circumstances or trends the data might be representing (Appendix C). Participants shared an update on an unfinished project's status or presented a finished projected at the end of each day.

	Time	Place	Social Life
My family story			
Data			

Figure 5. This matrix served as a scaffold for participants to help them with aligning their family story to big data selections.

Inviting critical perspectives. The researcher team introduced critical questions and encouraged threads related to historical and current issues of power, race, and equity in order to promote critical stances, which we view as important for learning with big data. We also wanted to better understand the extent to which counter-narratives can accompany more analytically sophisticated ways of visualizing big data. Oaktown's history and the larger, national history of slavery, racism, segregation, and Civil Rights were unavoidable and welcomed as pieces of the instructional design. The research team and library staff members also encouraged participants to think about who was counted in the data (e.g., Native Americans were not fully counted as part of the population until 1890; Pew Research Center, 2015) and whether social conditions differed by race (e.g. unemployment in the 1940s). We did this in part by positioning kids in historical time. For example, when Researcher 2 and one of the African-American participants (Carter) looked together at 1960's income data for Oaktown in Social Explorer, Researcher 2 gave some of Oaktown's history in a way that located both him and the participant in the past, with different social and civil rights:

Excerpt 2
[Week 3 Day 2]

1 Researcher 2: so let's look at this for a second. We are in 1960. so this is, downtown is still segregated, if you went down to buy food you couldn't sit down and eat it and I could...

Similarly, when asking the participants to pick a family member to think of during the four walls game or in the walking-scale timeline, we were also asking youth to physically stand in and speak for their family members, who potentially had survived or witnessed discrimination and prejudice. Most of the kids choose parents or grandparents, but some youth tried standing in for earlier relatives. Students who did not know a country of origin other than the United States were advised to consider the US state where their family members lived as an alternative for country of origin. Questions about occupation raised additional queries like what constitutes "work," particularly for participants thinking of enslaved ancestors, with some participants clearly stating that slavery was not work and others feeling less sure. As another illustrative example, in Weeks 2 and 3, in the four walls game, we asked how far did your family member/ancestor have to go in order to find a neighbor who did not share his or her race. Answer choices were Wall 1: less than a block, Wall 2: a couple blocks, Wall 3: 1-2 miles, or Wall 4: not in walking distance. Teens struggled on where to go, and the researcher team emphasized that everyone was making a guess. Some of the adults shared and brought up things like a family member living in segregation or conversely living in integrated spaces because of their profession. Hannah, one of the African-American participants who "stood in" for her paternal grandmother, left Wall 4 (not in walking distance) only to come back to it. Becky, a (White) library staff person, asked her about her decision-making.

Excerpt 3
[Week 2, Day 2]

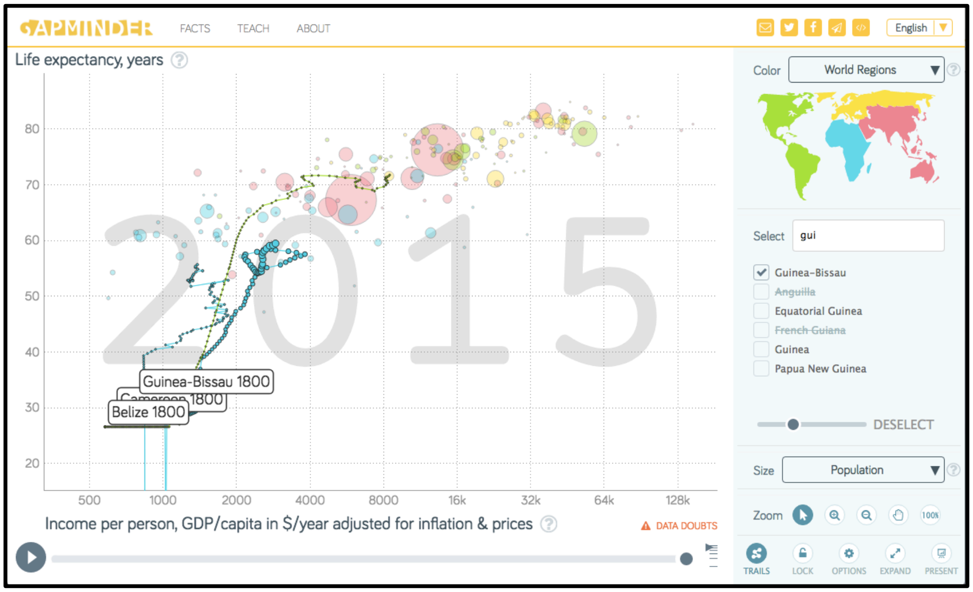
1 Becky: Hannah you were waffling. why did you decide to come back [to wall 4]?

2 Hannah: um because we went to Chicago a couple weeks ago, and we would have to go a long way to see someone who didn't look like [us] ((*trails off*)

We include these illustrative examples here to demonstrate that a goal of the instruction design was to foster critical perspectives that could lead to stories or models that would be grounded in experiences of people of color and challenge dominant social ideologies of equality and opportunity. Furthermore, we felt that the four-wall games and walking-scale timeline would be rich resources for reasoning about historical socioeconomic conditions in assembling family data storylines. Rubel et al. (2017) found that embodied activity on floor-sized representations support disciplinary and conceptual reasoning across other representational forms. In their study, activity with an oversized floor map supported proportional reasoning about the distribution of financial resources across the city from a spatial justice perspective; youth's embodied interactions on the floor map became a knowledge source for subsequently understanding interactive GIS maps that used social and economic datasets like city poverty rates.

Data tools. Students could use either the Gapminder motion chart (Figure 6; Rosling, Ronnlund, & Rosling, 2005) or Social Explorer (Figure 7), which both use open big data and are found in public media STEM argument performances. The Gapminder motion chart is a dynamic, interactive, digital, statistical visualization tool that models multivariable data through time and is freely available on Gapminder.org. Gapminder.org uses large, global, socioeconomic data sets released by state and nongovernment organizations. Countries leave trails of bubbles for each year of data that can be used to look at trends overtime. Social Explorer is an interactive, dynamic web-based mapping tool that accesses national demographic data, mostly from the US Census Bureau. Limited data is available freely online on socialexplorer.com, but free access to the full professional version was provided by the public library.

We chose Gapminder and Social Explorer in our designed activities because we consider both of these tools as case studies of interactive, multivariable digital, visualizations tools. Their interactivity supports questions and investigations about model assumptions and stakeholders, and they afford cross- and interdisciplinary uses for STEM and social studies learning (Kahn & Hall, 2016; Radinsky, 2008a; Radinsky, Hospelhorn, Melendez, Riel, & Washington, 2014). The majority of teens assembled maps made in Social Explorer for their family data storylines.



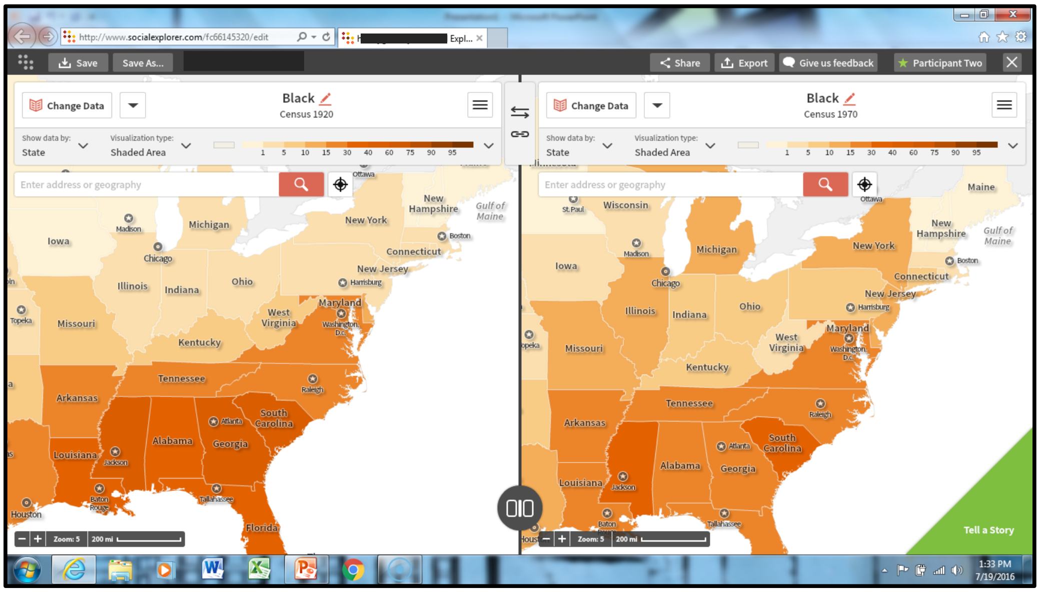
Color: World Regions

Select: gui

☑ Guinea-Bissau
□ ~~Anguilla~~
□ Equatorial Guinea
□ ~~French Guiana~~
□ Guinea
□ Papua New Guinea

Size: Population

Figure 6. This is the Gapminder interface. Each model has five possible quantities or variables that can be selected. Y-axis, x-axis, color, bubble size each has 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. Users can also alternate between logarithmic and linear scales, and highlight particular country actors.



Show data by: State
Visualization type: Shaded Area
Black
Census 1920

Show data by: State
Visualization type: Shaded Area
Black
Census 1970

Figure 7. This is the Social Explorer interface. Social Explorer can be used for different side-by-side temporal and spatial comparisons of socioeconomic data at national, regional, and local scales. Social Explorer accesses hundreds of variables from Census and other demographic datasets that go as far back to 1790. Data can be displayed in multiple visualization types: as dots, with shading, with bubble size. The side-by-side view permits comparisons of different places and/or datasets. The swipe view permits a user to move back and forth (swiping left to right) between two times (i.e., data survey years) and datasets, covering the screen entirely with one dataset or the other, in a single geographic location.

Additionally, participants had unlimited access to an Internet browser, which they used to access Gapminder; Social Explorer; Google (including Google Maps) or Bing search engines, which they used to search people, events, and places or access media (mostly images); as well Ancestry.com. Ancestry.com is a crowd-sourced tool for finding historical records of individuals, using names, birth and death dates, and places where relatives live; the library provided a full subscription to Ancestry.com. Engagement with Ancestry.com varied across participants. We view Ancestry.com and GIS tools like Google Earth or Google Maps that participants used in the workshop as powerful, online big data interfaces. While interaction with such technologies was included in the analysis, we did not consider them to be primary applications for understanding storytelling and modeling with big data.

Family exhibit. The three weekly sessions of the workshop culminated in single community exhibit that was open to family and the general public on a Sunday afternoon. Four sets of parents attended. All participants' projects were on display, and computer stations were set up with Social Explorer, Gapminder, and Ancestry.com for use by visitors and family groups (Figure 8). We also asked families and teens, librarians and staff to participate in a walking scale timeline (Figure 4).

Figure 8. Carter (right, age 13) and her younger sister Hannah (left, age 10) participated in the second and third weeks of the workshop and then brought their parents to the community exhibit, when this photo was taken.

Follow-up interviews. The lead author completed semi-structured follow-up interviews with participants and their parents to further ascertain how modeling with public, socioeconomic data affects one's understanding of the family, how youths' accounts of family history and broader social history compare to those of their parents, and how parents and children engaged in data exploration together. Discussions were facilitated around recreated data comparisons that youth participants made during the workshop and audio clips from their oral histories in the library. Families were given access to the data tools to engage in additional exploration together. Four interviews were conducted with seven participants and their parents (three of the interviews were with sibling pairs) 6–9 months after the summer program. Interviews lasted for an hour each per family and took place at the public library or at Researcher 1 and 2's university. (See Appendix E for the family interview protocol).

Analysis

We made video and audio records during all workshop activities. All computers had screen capture software operating on them as well; we also collected screen capture data during the family exhibit and the follow-up interviews. We collected participant artifacts, including their final PowerPoint family data storylines, completed worksheets, and their oral history recordings. Informed consent was obtained from parents and youth participants in accordance with our university institutional review board.

We took two-pronged approach to our analysis. Because much of the work of assembling the geobiography involved sibling exchanges and familial contributions, we followed families—sibling pairs or individuals—as our unit of analysis through the study activities. While we did not intentionally recruit siblings, we quickly realized that this was a unique opportunity to study family learning, since siblings are usually separated from each other during a typical school day. However, while we approached the records of workshop activities with a focus on families and siblings, each individual participant produced his or her own data storyline. Subsequently, individual participants also served as cases of storytelling and modeling with big data in our study, nested within a larger family narrative of participation. Likewise, we studied individual work always in relation to that of their siblings.

The analysis for this paper focuses on three pairs of siblings: Naimah (age 13) and Isis (age 10); Carter (age 13) and Hannah (age 10); and Daniel (age 12) and Brigitte (age 10). We selected these three sibling pairs as case studies of family units engaged in storytelling and modeling with big data. We chose these three cases for *maximum variation* (Flyvbjerg, 2006) in terms of their response to the designed activity. That is, we selected these pairs because of the diversity in their family histories and the variation in terms of how many days and which weeks they participated in the workshop (see Table 1). Our selection was also what Flyvbjerg (2006) calls an informed-oriented selection: From our fieldnotes, our review of their finished family data storylines, and completed follow-up interviews with each pair, we expected these cases to contain rich examples of scaling and assembling comparisons in telling family data storylines. Furthermore, we selected and compared cases in order to develop a grounded theory of storytelling and modelling with big data (Strauss & Corbin, 1994). Theoretical categories that reveal patterns of interaction (i.e., interaction between sibling participants or other family members, with the research and instructional team, with data tools) were produced though continuous or constant comparisons (Glaser, 1965; Glaser & Strauss, 1977) of data across cases of participants with the emerging conceptual categories.

Table 1

Attendance Record of Focal Participants

			Week 1		Week 2		Week 3			
Name	Age	Race	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2	Exhibit with parents	Follow-up interview
Naimah S.	13	African-American	x	x	x					x
Isis S.	10	African-American	x	x	x					x
Carter F.	13	African-American			x	x	x	x	x	x
Hannah F.	10	African-American			x	x	x	x	x	x
Daniel T.	12	White				x	x			x
Brigitte T.	10	White				x	x			x

We traced the focal sibling pairs together through the corpus of materials (instead of, for instance, following one participant at a time through all records). To organize our data, we divided their participation into several distinct activity structures within the study: framing workshop activities (e.g., four corners game), workshop activities involving data wrangling, the oral history recordings, the final family data storyline and accompanying presentations, the family exhibit, and the final interview. Across each record of activity for each pair, we pursued following questions: What does it mean for youth speak for the family or tell the family story? How do youth align the family story to the data? What is the nature and logic of their spatial, temporal, and social comparisons? In what ways were participants critical towards historical and present distributions and relations of power?

We used methods of interaction analysis (Derry et al. 2010; Jordan & Henderson, 1995) when analyzing the video records (primarily screen captures) of focal pairs. To answer our leading questions, we content logged video records (including screen captures) of focal pairs' participation throughout the workshop, in the community exhibit, and in the follow up interviews. We paid particular attention to activities within the course of their participation that, from our fieldnotes and observations, we felt would be "hotspots" or rich episodes for interaction with data and the production of family data storylines. In a more systematic, second pass through the recordings, analytic memos were developed around what appeared to be productive conceptual categories to describe participants' sociotechnical activity—namely, conceptual comparisons and comparative logics, the use of the data tools for getting personal, and instances of telling the family story. We classified or coded these practices into grounded theoretical categories (Glaser & Strauss, 1977).

We identified episodes of interest to develop detailed transcription and engage in microanalysis. For these selected episodes of interest in the record, we paid close attention to multimodal talk-in-interaction, that is, the sequence of turns of talk (Schegloff, 1991), tool use, gesture, and body movement that is publically visible in the record. Our interaction analysis efforts honored youth participants' perspectives towards spatial analysis and modeling, as evident in their own accounts of what they were learning (Stevens, 2010). We sought to uncover local or context-dependent knowledge-in use in their activities (Geertz, 1983; Hall & Stevens, 2016), such as knowledge of family ancestry, broader social events or histories, or the capacities of data tools or datasets. We often recreated participant maps or models in Social Explorer or Gapminder interface in order to closely trace their data explorations.

Like Radinsky et al. (2014), who designed activities for middle, high school, and college students to explore African-American and Latino migrations with Social Explorer, we took note of participants' verbal observations, inferences, and questions regarding data quality.

Additionally, we had a specific interest in understanding participants' multimodal comparative practices with the data tool (it appears that the version of Social Explorer used in Radinsky et al. [2014] did not permit the same kinds of side-by-side comparisons as our version). Furthermore, our focus on a getting personal with the big data in these cases represents a novel approach to studying modeling with big data modeling interfaces.

We triangulated video analyses with fieldnotes and participant artifacts. We repeated our analytic efforts (content logs, memos, interactional analysis of selected episodes) until our exhaustive study of these six participants, in addition to our review of the larger body of family data storyline projects, did not lead to new categories of representational practices of getting personal with big data (Charmaz, 2006).

Subsequently, we describe these categories of representational and sociotechnical practices, their properties, and the relations between them through grounded examples drawn from the three family cases in two parts below. We first describe how we came to understand *telling the family stor*y *as a distributed achievement* across family members and workshop activities for the three sibling pairs. We introduce each of participant's data stories within the broader family case of assembling and sharing the family geobiography. The analysis in this section stays primarily at the level of the family. The second section examines how our design activities supported participants in *becoming comparative social scientists.* In particular, we describe two aspects of participants engagement with the big data interfaces which we argue share resemblance with professional comparative social science practice: participants' *data wrangling to make contact with society,* or the set of sociotechnical practices to manage and select data from multiple survey years and hundreds of variables or indicators, and participants' *comparative logics* in their models, or their rationales and strategies for selecting data to assemble their comparisons in Social Explorer or Gapminder. In this latter section, we go more deeply into individual accounts to understand what youth made in the workshop, how and why they made their comparisons and family data storylines, as well as what their models and storylines meant to them.

Lastly, while our design invited a critical perspective, ultimately, we felt that such perspectives were mostly tangential in final storylines. Issues of race were present in conversations through each workshop day, but critical reflections were not strongly stated (or stated at all) in their final products. We return to this more seriously in the discussion and offer possible reasons for why the participants told the stories and assembled the models they did, ponder the value or utility of designing for counter-models, and the implications for future design work.

Findings Part I: Telling the Family Story as a Distributed Achievement

Through no effort of our own, telling the family story became a family endeavor and a distributed achievement. As youth began to develop family data storylines, they negotiated, debated and assembled the story with siblings who were in the workshop and parents—via phone calls, text messages, parent visits during the workshop day, and at-home conversations over intervening nights and days, in which pieces of family story were filled in, possibly changed or corrected by family members. Siblings sat side-by-side each other in the workshop at the same table but with individual laptops. Sibling pairs typically divided the family, with one or the other taking the maternal and paternal sides respectively; that sibling became the "expert" on his or her side of the family. (Similarly, unattached participants in the workshop had to choose just one side of the family to focus on.)

Participants initially had a broad or vague version of the family story (e.g., the family moved for job opportunities, for safety, and for a better quality of life). These stories became more refined over the course of the workshop as participants spoke with and learned information from family members and as a response to designed prompts and scaffolding (i.e., research team continued to ask questions to extend participants investigation with the data tools). Assembling the family data storyline required knowledge of details including names, dates, places, and broader historical, social circumstances, such as specific global and national wars, civil movements, or economic depressions and recessions. Participants entered the workshop with different amounts of this information about the family story; to supplement and help organize this material, we provided a family tree worksheet for teens to complete. Regardless of access to prior documentation of the family history, all participants reached out to parents and family members with inquiries during the session and searched for names, events, and places on Ancestry.com, Google, and Wikipedia; in addition, adults (researchers, library staff, parents when present) would offer details about historical events (wars, the influenza epidemic, Civil Rights Movement, the Great Depression, the Great Migration) as well.

Talking to family members did not diminish the agency of the teen as the storyteller. Rather, our instructional design positioned the youth participants as the official speakers for the family. This was a new role for the teens, as family storytelling is a domain typically dominated by parents or grandparents, such as in car trips and at family reunions, which was reported in the family follow-up interviews. There were multiple and repeated opportunities to tell pieces of the family story across the workshop to peers, to the research team and library staff and to family members. The oral histories with peers were particularly rich with family and personal stories, as were the family follow-up interviews. These moments of family storytelling exposed differences in accounts of events among family members, including generational differences between youth and parents' experiences. Youth drew on these perceived intergenerational differences throughout the family data storyline assembly process.

In order to help uncover what it means for youth to tell the family story, we introduce each family data storyline, using the matrix that served as conceptual scaffold for participants to align their family geobiographies to participants' data selections in either Social Explorer or Gapminder. We present the story of what moved each family by weaving together the two storylines produced by each sibling in the matrix. We give descriptive examples from each family across the data records to illustrate how telling the family story was—and perhaps is always—a distributed achievement.

Naimah and Isis' Family Data Storylines

		Time	Place	Social Life
Naimah	Family Story	1920–2016	South Carolina California Evanston, Illinois Mont Kisco, New York Tuskegee, Alabama Jackson, Mississippi Syracuse, New York Oaktown, Tennessee Belize	Paternal great-grandparents moved from South Carolina to California to Illinois for better jobs as part of the Great Migration. Paternal grandparents moved from Upstate New York and Illinois to Alabama and Mississippi for work opportunities. Father and mother met in Oaktown and moved to upstate New York, where Naimah and Isis were born, and then to Belize, before returning to Oaktown.
	Data	1920 1970	East Coast (data displayed at county level)	Black population increased in the North and decreased in the South after the Great Migration.
Isis	Family Story	1950	Pike Road, Alabama Syracuse, New York	Maternal great-grandmother moved from Alabama to upstate New York for "money jobs and a better life" as part of Great Migration. Her grandmother faced challenges in school because Syracuse was predominantly white. Her grandfather was born in upstate New York.
	Data	Isis did not include any Social Explorer data in her final family storyline.		

Figure 9. Naimah and Isis' family data storylines. This matrix presents the family story and the socioeconomic datasets participants used in their map comparisons for their final family data storylines. The Time column corresponds to the timespan covered in participants' final family data storylines the Family Story row and corresponds to survey years (US Census Surveys that occur every decade as well as American Community Surveys, which is an ongoing, annual demographic survey that began in 2005) of the data used for the Data row. The Place column identifies the named places in the text of their family data storylines for the Family Story row and the geographic areas of focus in their maps and models in the Data row. The Social Life column offers the narrative that the participants told in their text or in their presentations of their final family data storylines in the Family Story row, and the narrative, typically describing a change over time, that participants demonstrated with their data models or maps, in the Data row.

Naimah and Isis (Figure 9) entered the workshop having completed the homework assignment; they brought with them a record of names of family members, the places where they lived, and the dates when they lived in those places going four generations back, through their paternal and maternal-great-grandparents. On the first day of the workshop, Naimah shared that there was a gap in their knowledge of their family history from post-slavery until the birth of their great-grandparents; she reported that her parents had done investigative work to trace their ancestry with DNA swabs to African tribes in Guinea-Bissau and Cameroon. Guinea-Bissau and Cameroon, as their ancestral origins, stayed present for Naimah and Isis throughout their participation even though they are not in their final storylines. For instance, in the oral history recording session, when Isis' peer asked her where her family is from, she first responded with the tribe names in Guinea-Bissau and Cameroon; the librarian facilitating the oral history then prompted Isis to share where her parents lived before Oaktown. In her answer, Isis took up her ancestral family origins as part of her present personal history.

Naimah and Isis' family story began with their maternal (Isis' side) and paternal (Naimah's side) great-grandparents moving northwards as part of the Great Migration, primarily for better jobs. Isis also wrote about her paternal grandfather. For Naimah, the storyline continued to her grandparents, parents, and her and her sisters. Isis and Naimah's parents met in Oaktown and started their family there. In 2010, they moved to Belize for a year and a half, so their daughters could experience living in a predominantly African-heritage culture (based on family follow-up interview). In the oral history and family interview, both parents and teens reflected on this time in Belize as important, unique, and special for their daughter's sense of self.

In terms of data, Naimah focused on finding the black population in both the South and the North. She included two comparisons to demonstrate an increase in the Black population in the North and a decrease in Black population in the South between 1920 and 1970 (Figure 7). These survey years roughly corresponded with the birth of Naimah's paternal great-grandparents in South Carolina and their decision to leave South Carolina (1920) and her grandparents' return to the South (Alabama) from Illinois (1970); Naimah also noted that these survey years corresponded with the estimated start and end of the Great Migration as well. (Interestingly, the percentage of African Americans in these areas out of total population was the same in South Carolina in these years, although Naimah did not discuss this.) While Isis used Social Explorer in the workshop and examined a variety of datasets, she ultimately did not include data in her final project. Her challenges to align the story she wanted to tell about her family's celebration of Kwanzaa will be described in further detail in the latter half of the Findings section of this paper.

Naimah and Isis came from a family tradition of storytelling: Naimah shared in her oral history that her maternal grandparents and her mother were trained professional storytellers. In the family follow-up interview, both parents said that beyond storytelling as a family professional practice, their family tells stories all the time: Zuri (Mom) described the family as "operat[ing] in a lot of narrative," and Akinjide (Dad) added, "the stories are always flowing."

Based on the oral histories and the interviews, the stories about the family geobiography that Naimah and Isis shared have different temporal and spatial scales than the stories of their parents. In the example below, taken from the oral history recorded on the first day of the workshop, Isis story juxtaposed a very personal, affective experience (e.g., not going to school the day her grandmother died) with events over varied spatial and temporal scales (e.g., moving back to the United States, living in her grandmother's house for two years and then moving, once again, locally to their own house within Oaktown). Becky was the archivist and reference librarian who moderated the recording session.

Excerpt 4
[Week 1, Day 1]

1 Becky: How did your parents get to Oaktown? Did we already talk about that?

2 Isis: um because my grandma=we was in Belize, and my dad, he took a plane to go to Oaktown, because my cause my grandma had cancer, so when my grandma died, we just didn't go to school that day and flew to Oaktown with all of our stuff, and then we we just end up living in her house for like two years, and then we just got our own house

In one turn, Isis shared both specific memories of experiencing or being in places as well as her family's path from home to home (what Ingold, 2011, might call a *point-to-point traversal*) from Central America to the United States. This youth-perspective compression of time and space in family migration, in which Isis foregrounds the day her grandmother passed and the day they moved, contrasted with the temporal and spatial scale of moving for parents. In Excerpt 5 below from the family follow-up interview with Naimah, Isis, their mom and dad, Researcher 1 asked how did their family come to Oaktown. As part of that answer, the family described their moves from Oaktown to upstate New York then to Belize in Central America, and back to Oaktown. Isis offered her version of the story first.

Excerpt 5
[Family follow-up interview]

1 Researcher 1: how did you get from Oaktown to NY to Belize to here?

2 Isis: we were just looking for places to live, ((*Naimah leans back and brings her Left hand to her forehead and both parents laugh*)) so obviously they ((*points with L hand to parents*)) searched or looked for Belize they ((*unintelligible*)) and then we just drove to it and rented a house there and lived there for a year. We went to Ladyville school.

In Excerpt 5, Isis gave an account of the decision to move to Belize ("we were just looking for places to live"), in which the process of finding house is abbreviated (with "obviously," Isis may have been indicating that she did not want to tell either what she felt was a known part of the story or a story that was not hers to tell), while the road trip to Belize and her school are in the foreground. Unlike Isis' condensed orientation towards time and place both in the oral history and in Excerpt 5, her parents foregrounded reasons for moving in their version of the story. For Mom and Dad (Zuri and Akinjide), the decision to move to Belize was very purposeful, tied to the parents' past experiences traveling to Africa and their desires for shaping their daughters' identities. The move itself took years of planning and execution.

Excerpt 5 continued

[… *((Two minutes later in the record*))]

3 Zuri: SO we wanted to do two things. one was to infu::se a sense of adventure because we had personal love of adventure into the girls so we wanted them to have experiences throughout their lives of being more global citizens ((*unintelligible*)) and we also wanted them to live somewhere, that had a predominantly African presence, both politically, socially, visually, so we knew THAT was important *[…* ((*40 seconds in the record*))] So it took us was it a year and a half?

4 Akinjide: yeah about that

5 Zuri: yep about that to two years to start to figure out how we were going to make that transition, to then sell our home, to start selling possessions, to move into an apartment, to clear all debt, to save money, to then lastly tell family [*…20 seconds in the record*] We drove all the Belize. Isis was right.

[… ((*Another two minutes later in the record*))]

6 Akinjide: We wanted to have that type of experience for the girls, so they would not have an understanding of society based on what somebody's telling them they should be, and what exists, and that they actually are the majority and not the minority, and so when someone asks them who you are you can tell them who you are, have you been, yes you've been, and you have a more global experience, um, support that. Because that I feel will really anchor them. It anchored me.

Zuri and Akinjide presented a longer history of decision-making, interactions, and exchanges that characterized their move to Belize; they saw their family's time in Belize as social bedrock for their daughters' present and future senses of identity. We interpret their sharing of their decision to have their daughters experience a place with "a predominantly African presence both politically socially, visually" (Zuri, Turn 3), where "they actually are the majority and not the minority" (Akinjide, Turn 6), as a moment of critical storytelling that targets the relation between the individual, or their daughters, and society. The family geobiography itself became a counternarrative: The parents shared their intentional approach to their children's upbringing that challenged American society's norms for majority–minority relations.

In the workshop, however, Naimah and Isis' versions of the family geobiography served as the official story on record. The girls developed and demonstrated ownership of this story, as indicated by how they positioned themselves in relation to the data and to their ancestral history, as revealed through their language and grammar. During her presentation to the class of her work at the end of Week 1, Day 2, Researcher 1 asked Naimah what the data in her family data storyline left out.

Excerpt 6
[Week 1, Day 2]

1 Researcher 1: what were you going to say about what the data doesn't include?

2 Naimah: ((*reading from her notes*)) the data doesn't include why the population increased in the north, it also doesn't include how from the south, moving from South Carolina affected my family's connections, and how it was a milestone in my history

Her use of a possessive pronoun in her description of her family's historical movement as "a milestone in *my* history" indicated how her family story, her grandparents move from the South to North, featured as a significant part of her own history. We view this as a critique (and as a brief example of critical storytelling) that challenges what the census takers deem as important for understanding social history (i.e., "demographic facts" irrelevant of family relations and connections).

In general, assembling the family geobiography was a labor-intensive task, and Naimah and Isis turned to each other for support. Throughout the workshop, Naimah and Isis, like other siblings, consistently corrected each other, offered competing accounts of the same event—some of which were more light-hearted or somber, or co-remembered names, dates, and memories. The conversations about family history moved quickly between serious and silly. On the third day they attended the workshop, their final day, Naimah and Isis, seated next to each other, had an exchange about how their maternal grandfather died. Isis was writing a PowerPoint slide about her grandfather.

Excerpt 7
[Week 2, Day 1]

1 Isis: you know what's his name died from asthma

2 Naimah: who? You can't die from asthma

3 Isis: yes you can die from asthma

4 Naimah: you can die from an asthma attack but not from asthma

5 Isis: well he had an asthma attack

6 Naimah: who's what's his name ((*looks up from her computer at Isis to her right*))

7 Isis: [Grandfather's English name]

8 Naimah: no he had a heart attack

9 Isis: he fell down the stairs, he did fall down the stairs and ((*unintelligible*)) and then he had a heart attack then or an asthma THEN he di::ed ((*sings*)) don't tell me about my grandfather, don't tell me about my daddy ((*Naimah stares at her in silence*))

In Excerpt 7, Naimah tried to correct Isis' version of the story. Naimah's challenge to Isis' narrative explanation reflects a quality of stories that makes narratives like theories (Ochs et al., 1992): Other co-narrators can offer alternative versions of the story. Ochs et al. (1992) give the example of a family retelling the story of the mother mistakenly feeding one of her children a hot chili pepper; in reconstituting the event, while the mother and child co-narrate the story, the child's version of what happened is consistently subject to the mother's challenges. Applying Ochs et al.'s (1992) framework, Naimah's challenge in Turn 8 "no he had a heart attack" functioned like a *third turn* or *third position repair* in conversation sequences, in which a turn of talk addresses trouble or tries to clarify a misunderstanding in a previous turn of talk (Schegloff, 1991; Schegloff, Jefferson, & Sacks, 1977). Naimah's utterance addressed the trouble that Isis raised in her preceding turns: Isis attributes her paternal grandfather's passing to an asthma attack. However, Isis resisted Naimah's challenge playfully and continued to work on her grandfather's slide; she typed that her grandfather passed away from asthma. In thinking about the family story as a distributed accomplishment, these challenges (in the third person position) to events or their shared meaning indicate how siblings mutually elaborated and assembled different parts of the family story.

Ten minutes later, Isis called for Naimah's attention again. This time, her inquiry about her grandfather was more solemn.

Excerpt 7 continued

[…] ((*10 minutes later in the record*))

10 Isis: Naimah, do you remember him? ((*pause*, *no response*)) NAIMAH

11 Naimah: hmhmm

12 Isis: ((*speaking softly*)) do you remember him? ((*no response*))

As demonstrated in Excerpt 7, the nature of conversations shifted quickly in the workshop sessions, from lively contestations and banter about family history to more serious, sincere inquiries. Teens laughed and sang, listened to music, and told jokes. The balance between chitchat and focused inquiries could be very lopsided. Defining "time on task" in the workshop or productive disciplinary engagement (Engle & Conant, 2002) in terms of participant efforts to broadly align the family story to the data was initially challenging from our observations in the workshops. The leg work for producing stories that shifted scales of time, place, and social life was only revealed through our close analysis of the content logs that traced youth trajectories of participation. Teens regularly pulled out their smartphones but often to text questions to their parents about family history. A one-hour-long conversation about family and school life, in some cases without anyone using the data tool or PowerPoint, led to a powerful new insight in the storyline or data exploration. Alternatively, an hour of fastidious work on the PowerPoint might have been entirely devoted to adding animations to text and changing slide or font colors. Furthermore, sometimes, the most serious inquiry would involve no talk at all, such as when a youth was working through data wrangling with a data tool or developing text in a PowerPoint.

For instance, on Isis and Naimah's last day of the workshop (Week 2, Day 1), Isis conducted a Google Image search of herself by typing in her name into the search bar to find a picture of herself for her PowerPoint. As she scrolled through publicly available photos of her and her family members that her search generated, she came across an image—the confederate flag. She clicked on it and a few other images of the Confederate flag that appeared as "Related Images." A couple of the flags were captioned with "Southern Heritage," and one of the images had underneath it, "Racism, Segregation, Slavery, Treason" in small print. She clicks on another image with the Confederate flag and an automatic weapon, after which she promptly exited out of flag images and began clicking on images of Black women who appeared in her search. How Isis interpreted her search findings is hard to say, as she engaged in this interaction privately. One could imagine that this was not her first encounter with a Confederate flag, but it was unclear how she viewed those flags in relation to her search for herself, her ancestors' history as slaves, or her family's experiences today as an African-American family living in the U.S. South.

Carter and Hannah's Family Data Storylines

		Time	Place	Social Life
Carter	My Family Story	Mid-20th century–2016 and beyond	Gunnison, Mississippi Decatur, Illinois Oaktown, Tennessee	Maternal grandparents moved to Illinois for economic opportunity. Her mother moved to Tennessee for economic and education opportunities. Carter wants to go to an IV League for college and play viola professionally.
	Data	1940 1960 1990	Shelby County, Tennessee Macon County, Illinois Magnolia County, Tennessee (data displayed at county and tract levels)	There was a higher income rate in Macon County and Magnolia County than in Shelby County (a proxy for Bolivar County, Mississippi) in terms of percentage of households with less than $7,000 in annual income in 1960. Unemployment rates across Bolivar County, Macon County, and Magnolia County were similar in 1940, but unemployment in Magnolia County was lower than Macon County in 1990. Percent of population with some college education or more was higher in Magnolia County than in Macon County in 1990.
Hannah	My Family Story	1950–2016 and beyond	Gulfport, Mississippi Chicago, Illinois Oaktown, Tennessee	Paternal grandfather moved to Chicago from Mississippi because of job opportunities, lower cost of living, higher salaries, and it was safer [for African-Americans] at the time. Her father moved to Oaktown for college, met her mother, and stayed. Hannah wants to go to Duke or Princeton to become a teacher or musician.
	Data	1990 2008	Chicago, Cook, County, Illinois Oaktown, Magnolia County, Tennessee (data displayed at county or tract levels)	Percent of people with some college education is the same for Cook County and Magnolia County (close to 50%) in 1990. Magnolia County has more homes valued less than $50,000 than Cook County in 1990. Percent of married families with children living below the poverty rate in Chicago (zoomed into one of the census tracts) and in Magnolia County in 2008. Percent of employed Civilian 16 years and older in Cook County and in Magnolia County at the tract level in 1990.

Figure 10. Carter and Hannah's family data storylines. This matrix presents the family story and the socioeconomic datasets participants used in their map comparisons for their final family data storylines.

Like Naimah and Isis, Hannah and Carter told a multigenerational history of their family. They first traced their grandparents' migration from Mississippi (maternal side) and Alabama (paternal side) to Illinois in the mid-20th century for work. They followed this with the story of their parents' migrations from their respective cities in Illinois to Oaktown for college with accompanying data displays and a presentation their own college and career aspirations. Hannah and Carter were both very focused on using data to show the reasons that their family moved. They both primarily pursued economic indicators in their temporal and spatial comparisons using Social Explorer.

Hannah and Carter, like other participants, worked diligently to piece together the family geobiography. They sat beside each other each day (they attended the workshop for four days) and frequently asked each other questions about members in the family. They brought with them some information on the family history and made phone calls or sent text to their parents when they had difficulty remembering a name, date or place. Carter and Hannah spent considerable time on Ancestry.com, searching for family members and names, and shouting excitedly when they "found" a record of an ancestor or family member. These findings in Ancestry.com sometimes became a resource for telling about the family. For instance, one afternoon (Week 3, Day 1, their third day attending) of the workshop, Carter used an address listed in a record from Ancestry.com to find one of her father's childhood homes in Chicago using Google Earth, which prompted Hannah to talk about her father's economic hardships.

Hannah and Carter also demonstrated ownership of the family story by extending their family geobiography to include their current and future experiences. Carter, while working on the text of a PowerPoint slide of how her parents met, described her own birth as a dynamic and significant event in the family storyline following her parents' marriage (Excerpt 8).

Excerpt 8
[Week 2 Day 2]

1 Carter: ((*to Hannah*)) they got married in oh one, right? then I was born and BOOM! ((*Carter and one of her peers sitting at the table laugh*))[17]

Subsequently, in her comparisons that she assembles with the big data, Carter similarly reordered time and place in ways that reflect a youth-scaled perspective: Carter featured Tennessee in her historical comparisons, even when looking at a time (1940 or 1960) before her family members arrived in Tennessee (her parents moved to Tennessee in the late 1980s). The history of Tennessee as Carter's present home state overrides the temporal and spatial scales of other events from the family geobiography.

Hannah and Carter's family storytelling also revealed vast generational differences between their experiences and that of their family members. In the oral histories and family follow-up interviews, we asked teens how their parents' lives compared to their lives when they were their age. Across all three sibling pair cases, parents contributed alternative perspectives on childhood in the family storytelling interviews. Youth described stories at the personal scale that contrasted with their parents' experiences and desires when they were youth, as in Excerpt 9 below from Hannah and Carter's family follow-up interview with their mom Summer.

Excerpt 9
[Family follow-up interview]

1 Carter: they like went outside more, we don't really [… ((*5 seconds))*] um and like I guess, they were closer to their neighbors. [… ((*5 seconds))*] but I think it's because their community was different, it was like everyone was close and everyone was family, you can go over to that person's house without worrying about your child

2 Researcher 1: hmmhmm

[17] Carter also used the expression "Boom!" earlier in the record when she found an entry for mother in the Ancestry.com database. Researcher 2 in the record subsequently used the word "boom" to describe appearances of data on a map after making a selection in Social Explorer. The data populating a map in Social Explorer has this dynamic, powerful quality to it when it appears that seems to reflect, representationally, the significance of people coming into being in historical time or in a database.

3 Carter: but now it's like you don't know who that it is next door you have to like be careful, so we can't really like hang out with everyone you have to like be sure they're good

Unlike their mom and dad when they were growing up, Carter and Hannah do not live physically close to cousins; their family is located around the country. I then asked Hannah the same question about how she thinks her parents' experiences compare to her experiences when they were her age.

Excerpt 9 continued

4 Hannah: I feel like when my parents were 11, I feel like they had more responsibilities, because they didn't have as much as we have now, let's think of something ((*turns to mom*)) did you guys have outhouses?

5 Summer: no::: we did not, I don't know about your dad but I didn't have one ((*laughing*))

Their stories highlighted differences in daily and future mobility (e.g., how far a teen can travel independently by bike or walking and where the teen imagines they will live in the future), household and financial responsibilities, as well as the role of technologies like phones and social media, which were mostly echoed by Hannah and Carter's mom, with some elaborations or corrections. Today the girls commute to school in cars or buses (they go to two different magnet middle schools); they cannot ride their bikes around the neighborhood because cars drive too fast, and there are no sidewalks to walk on. Hannah also talked about how her dad endured economic hardships as a child that she does not experience now.

While our focal participants drew on perceived intergenerational differences in assembling their storylines in the workshops, in some cases, when telling the family story, participants found parallels with older generations that were meaningful. In their family data storylines, Carter and Hannah compared their parents' decision to go to a Historically Black College and University (HBCU) in Oaktown, where they met, with their own college aspirations to go university, although they imagine themselves attending a non-HBCU, elite private university out of state.

However, when participants aligned the family to the big data, the social distance collapsed among family members and generations as the youth, their siblings, and their older family members became a singular unit for comparison with the social history described by the data (e.g., Carter refers to “my people” to describe her ancestors who lived by Mississippi river in Excerpt 11). When participants compared their family’s experience with what the US Census reports about people like their families, the intergenerational differences and the internal complexity of the geobiography partly fade away. One conjecture for this is that when youth storytellers find the data to be too spotty or inadequate, they consequently abandon all the rich detail they achiev with a distributed geobiography. This could be the case for Isis, but our other focal participants did not do this. Rather, our interpretation is that it is always difficult to create a model with large-scale datasets about a particular topic, because the data were not specifically gathered for that topic. In turn, a participant’s model or map comparison was the result from an effort to use the data to fit the story, rather than the result from model competition (i.e., the more typical version of modeling), in which one produces and evaluates multiple models, each accompanied by an alternative story, explanation, or question, until one identifies the model and story with the “best fit” (Lehrer & Schauble, 2010).

Daniel and Brigitte's[18] Family Data Storylines

		Time	Place	Social Life
Daniel	My Family Story	1900-2016	Calumet County, Wisconsin	His maternal great grandfather stayed on the farm his whole life because of love and marriage, farming opportunities, and house value. His maternal great grandmother moved from a smaller farm to her husband's larger farm in a nearby town.
	Data	1920	Calumet County, Wisconsin	Compares of the number of smaller farms with 20-49 Acres to the number larger farms of 100-174 Acres in Calumet County in 1920.
Brigitte	My Family Story	1942 2003 2016	New York, New York Seoul, South Korea Durham, North Carolina Oaktown, Tennessee Madison, Wisconsin Atlanta, Georgia Brooklyn, New York	Paternal Grandparents born in New York City. Grandfather became a doctor and then served in the Vietnam War as a doctor, stationed in South Korea. He settled in Oaktown and raised her dad. Parents met in college in Wisconsin, moved to Atlanta for her dad to go graduate school, and the moved to New York for her dad's marketing job. They then moved to Oaktown because they felt it was safer (the 2003 Blackout was a scary experience for them) and less expensive to raise a family. If her parents had not moved to Oaktown, Brigitte would have lived in a smaller house and taken public transportation.
	Data	2003 2006 2010	Oaktown, Magnolia County, Tennessee Brooklyn, Kings County, New York	Percentage of violent crimes of total crimes in Magnolia and Kings County in 2010 (was lower in Magnolia). Owner-occupied Housing units valued less than $500,000 in Magnolia and Kings County in 2006 (was higher in Magnolia). Workers 16 year and over that commute with a car, truck, or van in Magnolia and. Kings County in 2006 (was higher in Magnolia). Workers 16 year and over that commute via public transportation Magnolia vs. Kings County in 2006 (was higher in Kings). Satellite image of nighttime lights in New York during the 2003 Blackout.

Figure 11. Daniel's and Brigitte's family storylines. This matrix presents the family story and the socioeconomic datasets participants used in their map comparisons for their final family data storylines.

[18] Authors selected their pseudonyms.

Daniel and Brigitte attended the workshop the second day of Week 2 and the first day of Week 3. Unlike other participants, they came to the workshop with a stockpile of family documents and artifacts, mostly archived by prior generations—including printed emails from grandparents and parents, a book of family history, and a CD with their maternal ancestry listed for 200 years. There were considerably more artifacts with information about mom's side of the family. At the start of their first day in the workshop, Brigitte and Daniel, like other siblings, negotiated who would tell the maternal or paternal family story. In Brigitte and Daniel's case, their exchange and disagreement revolved around their available resources; they negotiated to determine who would report on the side with the most amount information. In Excerpt 10 below, Daniel (12) convinced his younger sister Brigitte (10) not to do his mom's side of the family, which he already had already began reviewing and had a stake in, by positioning the maternal side of the geobiography as the more difficult task, involving more work.

Excerpt 10
[Week 2, Day 2]

1 Brigitte: I want to do mom's side

2 Daniel: why?

3 Brigitte: because mom's side has more information

4 Daniel: okay then I'll just do these ((*R hand points to right half of the family tree document that lay on the table between them*)) because it's less people ((*smiling*))

5 Brigitte: hmmm never mind ((*shakes head and hand, pauses 2 seconds*)) wait I'm confused

6 Daniel: do you want to do more people with more information ((*R hand rests on left half of the family tree*)) or less people with less information ((*R hand rests on right half of the family tree*))?

7 Brigitte: This one. ((*R hand taps the right side of family tree document that represents "less," pauses 3 seconds*)) ((*whispers*)) Is that reverse psychology?

Daniel suggested (and subsequently Brigitte agreed) that with more information, it may be harder (i.e., take more time and effort) to produce a narrative that is both accurate and integrative of that information. Interestingly, from a data perspective, this conclusion runs against the assumption that the more data and information you have, the better off you are.

Their storylines diverged in their approaches to telling the family story. Daniel, whose storyline begins with his maternal great-great grandfather, investigated why his great-grandfather stayed on the same farm for his entire life in Wisconsin. He found data that showed the number of farms of different sizes by county in Wisconsin. Brigitte started her storyline with her paternal grandparents: Her grandfather was a surgeon in Seoul during the Vietnam War before settling in Oaktown where her father grew-up. She then told the story of how her parents met in college in Wisconsin and lived in Brooklyn, New York until they had her older brother, Daniel, at which point they moved to Oaktown, where she was born. Brigitte then went a step further and imagined what her life would be like today if her parents had stayed in Brooklyn. Her storyline subsequently developed both texture and depth for a single time period. She compared New York and Oaktown in 2006 and 2010, close in time to when her parents moved, on indicators that reflected differences in cost of living, transportation or mobility, and safety with the Social Explorer tool. She also incorporated other forms of data into her storyline, including an image from Google Street View of the street where family lived in Brooklyn and satellite image of nighttime lights in New York during the 2003 Blackout, which she described as a scary experience for her parents.

For Daniel and Brigitte, conversations with their mom, Kara[5], led to new insights in the stories they told; these conversations were captured on the record at the end of the workshop day when she stopped by to pick them up. In talking with his mom, Daniel discovered that his great-great-grandfather was able to acquire his farmland in the first place because the federal government had sold it cheaply to White farmers in the early 20th century after forcibly seizing territory from Native Americans. Similarly, a conversation with her mom led to a redrafting of Brigitte's storyline in her PowerPoint; Brigitte originally wrote that New York City's "busy" "lifestyle" was "unsuitable for a child." Kara insisted that was not case. Rather, there were many other reasons for why her parents moved, such as price of homes, cost of childcare, and to be close to family, and there were positive aspects about city living that they missed, like walkability and public transportation, which Kara also discussed in the family follow-up interview. Similar to the mother in Ochs et al.'s (1992) analysis of a family telling the "chili peppers" story, Kara challenged both Daniel and Brigitte's accounts of the events that set a decision to move or stay in motion for the family. While participants across all three sibling cases seemed to prefer a simple (A happened, and so B was necessary) explanatory story about mobility or staying put, we found that parents worked to make these simplifications more complicated (e.g., Naimah and Isis' parents' elaboration of the decision and process to move to Belize in Excerpt 5).

Brigitte and Daniel's storylines illustrate how inciting or initiating events (Ochs et al., 1992) feature in youth family storytelling. In general, narratives often focus on a central problematic event or circumstance (the inciting or initiating event) that results in either a decision or action and possibly resolution in the narrative. In Daniel's story, the government taking away territory from the Native Americans and reselling the land cheaply to White farmers was a central event in the narrative of why his great-grandfather never left the farm he lived on; the farmland today is worth a lot of money. The 2003 New York Black out was the initiating event in Brigitte's story of her parents' decision to move to Oaktown from New York. In her PowerPoint, she wrote "This happened about two years after 9/11, so my parents were scared that it could have been another terrorist attack. Everything turned out OK, but that was one of the main reasons of moving."

In summary, our analysis of telling the family story as a distributed achievement found similar patterns and practices across sibling cases, although each pair started with different amounts of information already in hand. All the youth appeared to have a simplified, youth-centric version of the story at the outset. However, parents or other sources of narrative material and explanation, including their siblings, complicated their stories, deepened their engagement, and added meaning to their activity. In turn, it is possible that the role of parents or other relatives in the distributed family story is to provide a learning opportunity for youth and that this is a benefit of family storytelling, irrespective of modeling with big data. Lastly, as indicated by the storyline matrices, the sibling participants handled time and socioeconomic influence in various ways, as will be further developed in the subsequent section.

Findings Part II: Becoming Comparative Social Scientists

The storylines described above came about through hours of *data wrangling to make contact with society*. Data wrangling, in storytelling and modeling with big data, refers to the set of sociotechnical practices to manage and select from hundreds of datasets to create comparisons that investigate the possible social conditions that contributed to their family's decisions to move. Though data wrangling, youth aligned the family story to the data and bring the self and society together.

We suggest that participants' forms of data wrangling bear resemblance to the methods and practices of social scientists and scientists engage in to manage uncertainty or local difficulties in their research efforts (Star, 1985). Star describes how scientists and doctors respond to uncertainty in their local work settings produced by both material challenges and political pressures; scientists, for instance, will alter their division of labor or change their data organization structures and their standardization practices to address daily problems. Law (2004) also frames local uncertainties as a problem of social scientists, what he calls the "practical contingencies" of messy social science research work (p. 22).

Participants wrangled data and refined their storylines until they have achieved what Becker (2002) calls a "specified generalization." Becker describes how photographs (like photographs for historical documentation or journalism) are always images of specific people, places, and things but, at the same time, embody a "generalized story" of what is possible in the world (p. 5). Like photographers who document people and events, our study participants worked until they produced what they perceived to be "enough parts of a long story"—enough data that they deemed relevant to their proposed reasons for their family moving or staying put (this was how participants viewed data wrangling) as well as enough text and images to convince an audience of their ancestors' experiences, as part of a more generalized (or aggregate) social history.

Participants' wrangled data by assembling spatial, temporal, and social comparisons in order to discover and reveal social and economic conditions of influence that could have motivated family members decisions to migrate or stay. Like professional comparative social scientists, participants developed *comparative logics* that drove their modeling. We present their rationales and strategies for selecting datasets in these comparisons, which we relate to traditional social science research methods and approaches to comparative case research. Notably, youth data wrangling and comparisons involved a variety of practices that leveraged the dynamism of Social Explorer to traverse temporal, spatial, and social scales.

Table 2 below provides a guide for looking across our case studies at the stories that the participants were trying to align to the big data, the data they were trying to wrangle or find, and their logic or rational for their comparisons. Minuses (-) indicate less desirable social and economic conditions in a given place or time (e.g., lower income rates, higher crime). Pluses (+) indicate positive or improved socioeconomic conditions (e.g., higher income rates, lower unemployment). Comparative practices that participants employed are listed below. We first offer illustrative examples of how youth wrangled data to align themselves with society, as represented by the aggregate data in Social Explorer. We then describe the focal participants' strategies for their comparisons in their family data storylines.

Table 2

Data Wrangling and Comparative Logics

Participant	Story being told/aligned to data	Data wrangled	Comparative logic and practices
Naimah	Family was part of great migration	Data that demonstrates Black population growth in the North and Black population decline in the South	(- → +) Approximation
Isis	Who celebrates Kwanzaa	Data that shows families like hers who celebrate Kwanzaa	N/A
Carter	How my family’s circumstances improved over time and place	Economic data that showed a difference between Mississippi, Illinois, and Tennessee	(- → +→++) Approximation Time-jumping Imputation Commensuration
Hannah	Why dad moved	Data that showed a difference between Chicago and Oaktown	(- → +) Approximation Variation through changing scale
Daniel	Why great-grandpa stayed	Data about farmland in Wisconsin	(+ → ++)
Brigitte	What my life would have been like if my parents had stayed in NY	Data that showed a difference between Brooklyn and Oaktown	(- → +) Hypothetical counterfactual

Data Wrangling to Make Contact With Society

Participants performed a range of comparisons that involved scaling time, space, and social life using the data tools. To do this, they engaged with *data wrangling*, practices that manage multiple datasets and measures in order to connect the family story with the aggregate data. Data wrangling is more than a matter of technical or statistical skill; it is a process of alignment, in which one’s social values and sense of identity is central.

In data wrangling activity, we see evidence of youth getting personal with big data in trying to locate, or inhabit the places that hold meaning for them, whether in their ancestral family history or their own personal, lived history, or in trying to find a population that one identifies with in the data. When participants first opened a dataset in Social Explorer, they searched Oaktown or the surrounding state. Participants also selected datasets that represent populations that they identify with, like the African-American or Black population, or the population under 15 years of age. Many of the participants searched for themselves in Ancestry.com or looked for their homes on Google Earth as well.

For example, Carter’s first move in exploring Gapminder for the first time was to change the x-axis indicator to the percentage of the female population aged 10–14 years; she kept life expectancy, which was preselected, on the y-axis. Carter then selected the US, the United Kingdom, Afghanistan, and Australia to look at over time. We learn from her oral history and from discussions in the workshop that Carter’s best friend from school is from Afghanistan and was spending the summer there; Australia and the United Kingdom are two places where Carter desires to visit or possibly live in the future.

Similarly, Isis (Figure 6), in her first Gapminder exploration, assembled a comparison between Cameroon, Guinea-Bissau, and Belize and played them out on income per person and life expectancy (the default x- and y-axes in Gapminder) from 1800 to present day. Naimah and Isis’ family had previously traced their ancestors to tribes in Cameroon and Guinea-Bissau before they were forced to go to the United States as slaves. Isis also tried to find New York, where her mother's family is from and where they also lived briefly before moving to Belize, and

Utah, a place where, based on her oral history, her family had recently visited for a vacation, in Gapminder. However, she was unable to add New York and Utah to her motion chart because they are places at the wrong spatial scale (state and not nation) for the data tool. Naimah, who was exploring Gapminder with Isis, subsequently added United States onto the model but took out Belize. These participants' first moves in the data was to select places of meaning across time, past, present, and future. Getting personal with big data through data wrangling seems to be a primary way that one enters the world of spatially organized big data or a demographic database.

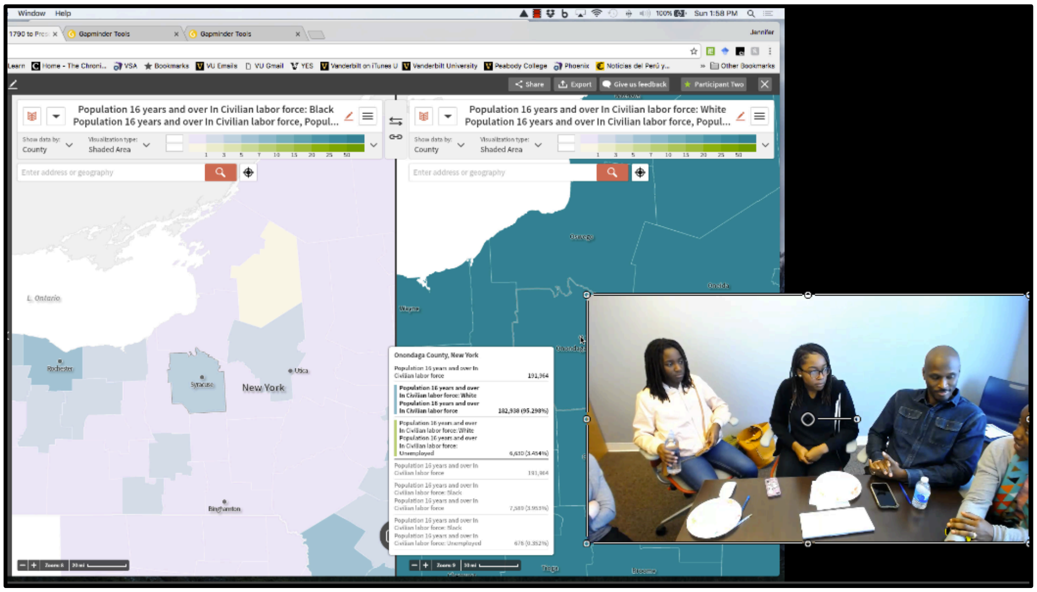
Below (Table 3) we have noted when one of the focal participants, upon entering a data interface (Social Explorer, Gapminder, or Ancestry.com), got personal with the big data. That is, we have reported when participants initial interaction with the data interface was to locate places that holds meaning or select population that one identifies with in the data without an instructional prompt from a research team member or librarian.

Table 3
Getting Personal With Big Data First Moves

Participant	Data wrangling: Getting personal first moves
Naimah	Looked up Black or African American Alone population and types in her street address.
	Compared Cameroon, Guinea-Bissau, and US in Gapminder.
Isis	Compared Cameroon, Guinea-Bissau, Belize, New York and Utah on Gapminder.
Carter	Compared female population in Australia, Afghanistan, United Kingdom, US in Gapminder.
Hannah	Looked at crime statistics in Magnolia County, Tennessee [in context of big data challenge activity]
Daniel	Searched for self on Ancestry.com
Brigitte	Looked up total population female in Social Explorer, highlighted Tennessee

Other learning sciences studies offer additional evidence for getting personal as a way to become familiar with large-scale databases. Rubel et al. (2017) discovered that over the course of a project in which youth examined lottery-related socioeconomic data with GIS maps in their city, students engaged in "finding self," locating their homes, routes for commutes, and other meaningful places that they frequented in the city in the GIS maps and other spatial texts (i.e., youth also immediately found their homes on the oversized floor map of the city). Wilkerson-Jerde and Laina (2015) similarly describe how middle school youth, in their assembly of data visualizations of public city infrastructure and demographic data, "immediately situate *themselves* in the data – seeking those languages and ethnic identities that they felt reflected them and their families" (p. 5). We have seen these ways of getting personal with big data repeatedly and consistently across our design iterations with youth and big data interfaces.

In general, wrangling big data for the family geobiography also entailed an effort to decide if family members were represented by the dataset. In Naimah and Isis' family follow-up interview, the family looked together at a map comparison that Researcher 1 created in Social Explorer (so the data was already partly wrangled) based on indicators (socioeconomic variables or conditions) that Isis considered during the workshop (Figure 12). The map showed Black (on left) and White (on right) population of 16 years and older in the civilian labor force for upstate New York in 1970. Collaboratively, the family all participated to dissect the map comparison. They concluded that the employment rate was higher for the White population in the county.



Population 16 years and over In Civilian labor force: Black
Population 16 years and over In Civilian labor force, Popul...

Show data by: County
Visualization type: Shaded Area

Population 16 years and over In Civilian labor force: White
Population 16 years and over In Civilian labor force, Popul...

Show data by: County
Visualization type: Shaded Area

Onondaga County, New York

Population 16 years and over In Civilian labor force	191,964
Population 16 years and over In Civilian labor force: White Population 16 years and over In Civilian labor force	182,938 (95.298%)
Population 16 years and over In Civilian labor force: White Population 16 years and over In Civilian labor force: Unemployed	6,630 (3.454%)
Population 16 years and over In Civilian labor force	191,964
Population 16 years and over In Civilian labor force: Black Population 16 years and over In Civilian labor force	7,589 (3.953%)
Population 16 years and over In Civilian labor force: Black Population 16 years and over In Civilian labor force: Unemployed	676 (0.352%)

Figure 12. This image was take from the family follow-up interview with Naimah, Isis, and their parents. The family is viewing a comparison in Social Explorer for the family follow-up interview, assembled by Researcher 1. The map shows the Black (on the left) and White (on the right) population of 16 years and older in the civilian labor force for upstate New York in 1970. Data is displayed at the county level. In the image of the interview, from left to right sat Researcher 1 (off screen), Isis, Naimah, Akinjide, and Zuri.

The family then engaged in getting personal work—to determine if Zuri's mom (Naimah and Isis' maternal grandmother) would have been within the "counted" population in the map, that is, if she was old enough to work in 1970 in Syracuse. This led to a moment of storytelling about their grandmother's experiences working and encountering a sexist (and likely racist) slight.

Excerpt 11
[Family follow-up interview]

1 Zuri: so if she moved to Syracuse in 1966 I guess she would have been like the 16 and up person. How did that impact her? I'd have to ask about. Did she feel like it was difficult? She does have this story she tells about working in the department store

2 Akinjide: oh gosh I know where you're going with this

3 Researcher 1: Have you heard this story? ((*Naimah and Isis shake their heads no*.))

4 Zuri: ooh my mother loved [to tell this story]. She is definitely the person to tell her life story. there were Afro wigs that came in style. and she:: I guess um the Styrofoam heads that you put the wigs on they were compacted, so she took a pick and picked them all out and made them like these big beautiful afros, and so they started to sell really fast, a::nd the owner of the store asked her to keep doing it. and she said well, I'm happy to but could there be some reflection in my pay in doing such work and her line was he said to her ((*voice shifts register, lowers*)) girls like you are a dime or dozen, so ((*returns to higher register, turns to daughters)*) which do you understand that term?

5 Naimah: mmhmhm [no]

6 Zuri: I can fill your spot anytime I can get you

7 Akinjide: ((*turns to girls*)) [You're replaceable

8 Researcher 1: ((*turns to girls*)) [You're replaceable yeah

9 Zuri: yeah you’re replaceable. She tells that story often. and I think it happened again like 15 years later when she did part time work in another department store. same kind of sentiment she went above and beyond and they let her know like pshh you know please I'm not going to pay you anymore. but I don't know if it was difficult. I know they moved there because it was easier

10 Akinjide: ((*looks to Mom*)): than Alabama, right?

11 Zuri: yeah they moved from Alabama I think for reasons in pursuit of economic freedoms

In Turn 4, Zuri shifted to addressing her daughters directly, and the story pivoted towards a teaching moment (what “girls like you are a dime or dozen” means). Zuri then returned back to a discussion of the family geobiography in relation to the data in Turn 9; she wondered aloud if life for her mother in upstate New York was challenging (“but I don’t know if it was difficult”), which could be an inference drawn from the data comparison. The process of aligning oneself or family members (in this excerpt, accomplished by Mom) to the dataset was a rich source of storytelling, teaching, and learning.

In our activities, data wrangling was both necessary for building comparisons and took a large amount of time. Participants had to make selections from among hundreds of indicators or variables. In Social Explorer, youth first chose a survey year to look at, then selected a data category (e.g., Race, Income, Employment, etc.), and then selected an available variable from within the category (e.g., in 1990, under Race, one could select the variable Persons: Black), and the data would populate on the map. Participants then could zoom or focus the map window on a particular location or two locations if the tool was set up to swipe or for a side-by-side comparison; the zoom level could also change the geographic scale of the data or the reporting area (data shown by state, county, or census tract), if the selected data was available by each of those reporting areas.

As we have seen in previous design iterations (Kahn, 2014), youth tried many different indicators and pairs of datasets before settling on a particular comparison to pursue and refine for their family data storyline. Bouts of trial and error were shaped by missing or spotty data, different indicator breakdowns for different years (e.g., income), or changes in data categories. Participants also had to discern between counts versus percentages; this was a persistent data literacy challenge across cases. In Social Explorer, shading was by far the most popular visualization type, in which the color darkened according to percent based on percentage; the color served as a heuristic for overcoming this persistent data literacy challenge.

The following episode captures a moment of “trouble” for Carter as she explored aggregate-scale economic indicators with Social Explorer. She encountered missing data in 1960 in Mississippi, which she confused for not having any households within the data criteria, and wrestled with different income rates due to inflation over 50 years as well as different income breaks between census survey years. In this episode, Carter called on the research assistant Ines for help, who also struggled to make sense of the data.

Excerpt 12
[Week 3, Day 1]

((*The left side shows Household Income Less than $25,000 for 1960 with census tracts available for only certain city areas. Census tracts are the statistical subdivisions or reporting districts determined by the previous decennial Census survey. The map is centered on Mississippi, which has no data. Ines is seated to the right of Carter, off screen, in Figure 13*))

1 Carter: oh there we go, cool ((*Carter selects Household Income Less than $25,000 for 2014, Data populates by county, for all counties; Visualization type: shading*)) OH my goodness gracious, that's a lot ((*pause 3 seconds*)) Is less than 25,000 a LOT, like now or back then.

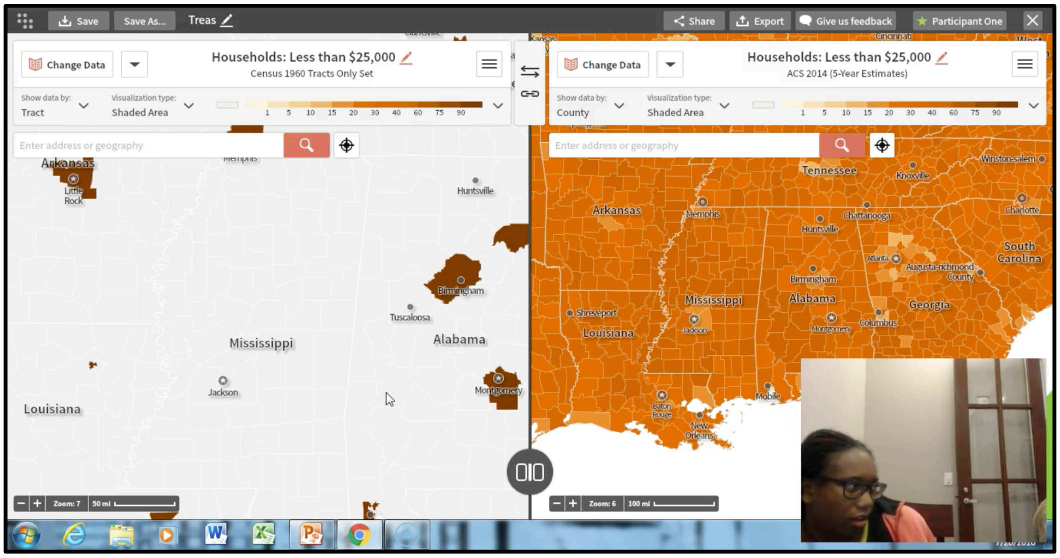


Figure 13. This screenshot captured Carter engaged in data wrangling. Carter pulled put up Household Income Less than $25,000 for 2014 on the right side, displayed by county. Data was available for all counties in 2014 (right side). On the left, data was shown by census tract in 1960, and only select tracts displayed data (left side). The shading scale was the same on both maps. The darkest red on the scale, as on the 1960 side, indicates 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. In Turn 2 in Excerpt 12, Carter asked if $25,000 was considered a lot of money for household income in 1960**.**

2 Ines: wait what?

3 Carter: like in 1960 was 25,000 a lot?

4 Ines: ((*pause two seconds*)) for a house?

5 Carter: hmmhmm

6 Ines: u::m

7 Carter: well a household, ((*mouse hovers over Mississippi on the left in 1960*)) cause Mississippi didn't have ANY that have less than $25,000

Ines, in Turn 4, confused household income for the value of a home, but Carter's repair in Turn 7 "well a household" does not clarify her misunderstanding.

8 Ines: uh is that true?

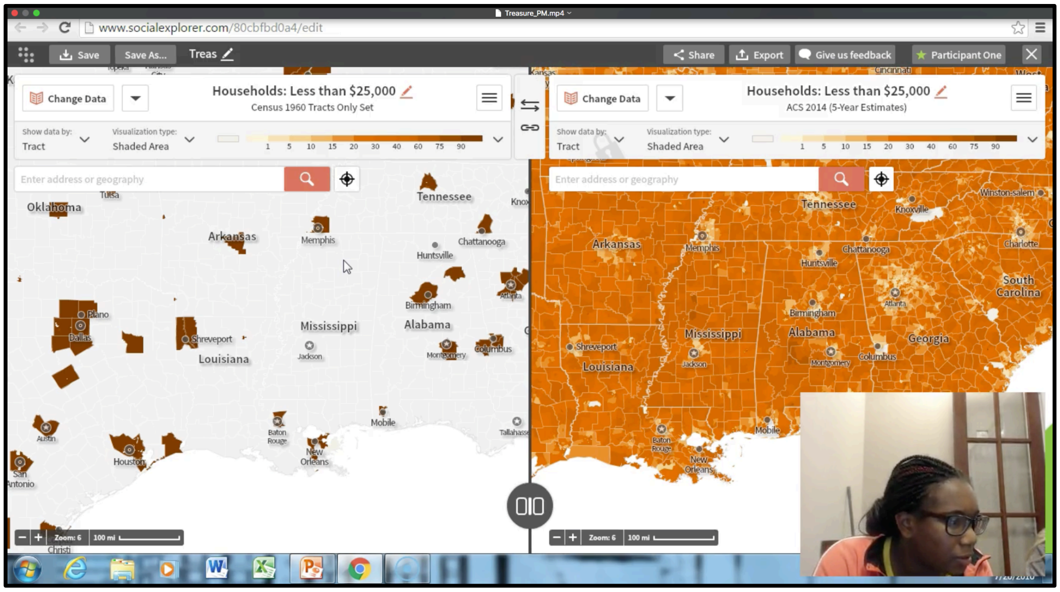
9 Carter: hmmhmm ((*pause three seconds*; *Ines leans into look at Carter's computer screen, takes the mouse*)) well they do now

10 Ines: well no because this is ((*pause three seconds*)) see this is what I don't like

11 Carter: OH:: ((*points 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 tract*))


www.socialexplorer.com/80cbfbd0a4/edit

Households: Less than $25,000
Census 1960 Tracts Only Set
Show data by: Tract
Visualization type: Shaded Area

Households: Less than $25,000
ACS 2014 (5-Year Estimates)
Show data by: Tract
Visualization type: Shaded Area

Figure 14. Carter's data wrangling continued. Carter had a lot of data available for 2014 but not so for 1960. Carter suggested that the problem could be that her map on the left displayed data by census tract, and her map on the right showed data by county, so she changed the right side to show data by tract.

Carter recognized that the two sides of her comparison are showing data at different geographic scales (county versus tract). Ines affirmed this, that she should be comparing data at the same zoom number or grain size. Carter changed the scale on the right map to the tract level, but places at the same zoom number or grain size, but her rectification does not solve her problems. The lack of data in 1960 led to a display with lots of missing data, so Ines asks 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*))

Carter finds that the income breaks in the 1970 Census survey are not familiar; they do not match the income breaks for 1960. There is no "Less Than $25,000" dataset in 1970.

16 Ines: [No

17 Carter: [No it's not. I don't know, it like totally changed. what about this one? [YEAH ((*clicks 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: hmmhmm

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.

Excerpt 12 is part of a longer line of model development for Carter, in which she eventually assembled a model that compared an estimate for average household income (less than $7,000) between Mississippi and Tennessee in 1960 for her family data storyline. In this excerpt, Carter encountered multiple challenges of data wrangling, including determining whether the indicator is a count (amount) or a rate, whether blank 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 assistant, 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 non-sparse data, which she then proceeded to wrangle in the workshop.

Producing variation by changing the aggregate scale. Participants also changed scale (grain size, using the zoom feature) to find variation as a form of data wrangling. For instance, in a comparison of the population age 25+ with some college experience or more between Chicago and Oaktown in 1990, which was around the time when her dad moved to Oaktown for college, Hannah noticed that at the census tract level, neighborhoods in the Southwest side of Chicago had lower rates of college attainment than in Magnolia County. However, at the county level, they were very similar. Similarly, Carter, in tracing her mother's move from Decatur to Oaktown for college, initially discovered the percent of population age 25+ with some college experience or more was nearly ten percent higher in Illinois than in Tennessee at the state level in 1990, but she found a 10% difference at the county level in Oaktown' s favor in 1990 (38% with some college experience or more in Macon County, where her mother grew up, compared to 49% in Magnolia County, where Carter grew up). Carter included the latter comparison of education data in her family data storyline. As a data wrangling strategy, changing scales better supported Carter's comparative logic that her each time her family moved, they found greater economic and educational opportunities (in the Critical Positioning matrix, my life and the aggregate have improved +, +).

Approximation as a response to uncertainty. Participants constantly engaged in acts of approximation; when the data did not have the exact year or category they were looking for, youth made selections that were approximations for the data they desired. Decisions to approximate were emergent responses or fixes to manage uncertainty (Star, 1985) or trouble such as an event falling in between census years, (e.g., choosing 1990 for a move that took place in the mid 1980s), a lack of specific details or information about family history (e.g., When did Dad actually move to Oaktown?), or a flexible approach to time (e.g., 1940s and 1960s were both adequate representations of "back then" when Carter's grandparents from the South to the North).

Inconclusive data wrangling. In certain cases, data wrangling went unresolved. In her oral history, Isis described her family celebrating of Kwanzaa; after listening to her oral history in Day 2, Isis wanted to find data to reflect the population in the US that celebrates Kwanzaa. However Social Explorer's religion dataset did not have any data on Kwanzaa, which is a cultural heritage holiday. Isis looked up the Black population in Social Explorer in Oaktown in 2014, but this data did not make it into her PowerPoint. Her data exploration did not end there; with another researcher that same day, she also looked at unemployment data across the US in 1930 and 1960, but this data did not make it into her final presentation either. She does not enter either Social Explorer or Gapminder interfaces when she comes back the following week to finish her family data storyline.

In summary, across sibling cases, episodes of data wrangling to make contact with society produced new insights. In Excerpt 11, Mom aligned a family story with the data, which led to episodes of storytelling and teaching. In Excerpt 12, Carter enlisted Ines' help with addressing sparse data and took steps towards assembling a cleaner comparison. In several cases, data wrangling strategies, such as approximation or changing the aggregate scale to elicit variation, served the comparisons sought by youth in workshop activities. Moreover, each of our study participants who produced a family data storyline (except Isis) successfully found and arranged data to make comparisons relevant for explaining why their family moved or stayed, although our present analysis suggests that some participants engaged in more data wrangling than others. An additional, future analysis solely devoted to more intensively uncovering the types of trouble participants faced in data wrangling and how they responded to that trouble is necessary.

Comparative Logics in Modeling and Mapping the Geobiography

In assembling their family data storylines, participants built complex comparisons that scaled time, space, and social life. As part of their efforts to understand why their parents or grandparents moved, participants compared various indicators between different decades and across cities and states to understand the social and economic conditions that influenced their family members' decisions to move or stay. The participant's pursuit of influence is similar to the tradition of comparative social scientists that engage in comparative case research in order to *assess causal complexity* (Ragin, 1987/2014)—the intersecting socioeconomic conditions produced historical social events. While the youth participants were not in the game of determining "cause," like the comparative social scientists and methodologists that Ragin describes, participants nonetheless pursued comparative practices with big data that have distinct, identifiable logics or rationales. We describe some of these comparative strategies below.

Ruling out "rival explanations." In general, participants favored comparisons that revealed differences as opposed to similarities over time or between places to tell stories of influence and change. Youth avoided showing the absence of influence, perhaps because they felt these comparisons would not be interesting. However, they found similarities across places, or undesirable differences (e.g., unemployment rates were higher in Mississippi than Illinois), participants were typically hesitant to include such map comparisons and looked for other indicators demonstrate improved socioeconomic conditions in the destination. For instance, Hannah, wanted to compare rates of college attainment for African-Americans (her own family) as an indicator of opportunity in Tennessee. However, when she compared the population with some college education or more in Cook County, Illinois and Magnolia County, Tennessee in 1990, she found that the percent of people with some college education or more was practically the same in both counties (47% and 49% respectively). While Hannah did not entirely reject higher college attainment as a potential factor in her father's move to Tennessee, this experience motivated her to look at other variables, such as employment and cost of housing, as alternative explanations for her father's decision to move. This approach seems similar to Yin's (2000) concept of ruling out "rival explanations," in which one shows that a possible cause does not account for the outcome in a given case.

Hypothetical counterfactual. In her family data storyline, Brigitte imagined what her life would have been like if her parents had not moved from Brooklyn, New York to Oaktown. She constructed a series of comparisons between Kings County, where her parents lived before she was born, and Magnolia County in 2006, an approximation of the year when her parents moved (there was no survey data for 2003). She compared property and violent crimes; owner-occupied housing with a value less than $500,000; workers over 16 who drive a car, truck, or van; and workers over 16 who take public transportation, using the shading visualization. She wrote in her PowerPoint:

> If my parents had stayed in New York, I would have had to take public transportation,

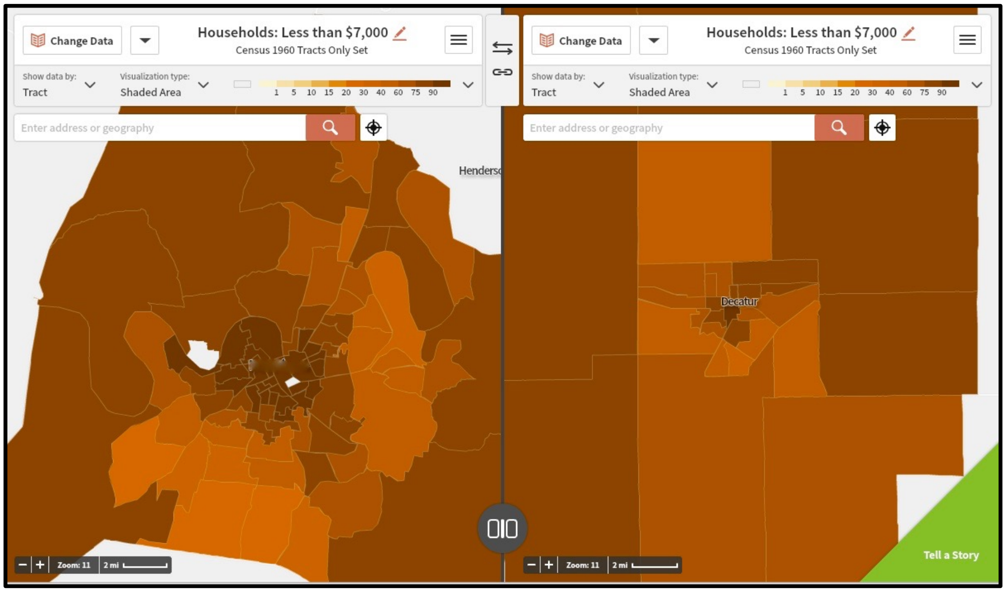
> and we probably wouldn't have had a house like we have now because it would cost too much to own a house in New York.

Brigitte's comparisons constitute a hypothetical counterfactual. In social science research method, a counterfactual conditional statement establishes that if a false Condition A had actually occurred, then Consequence B would have happened (Lewis, 1973; Little, 2004). Consequence B describes "how the world *would have been* if the antecedent [A] had obtained" (Little, 2004 p. 206). In Brigitte's case, Condition A is her parent's staying in Brooklyn, and the consequence is what she describes and grounds in the big data.

The hypothetical counterfactual is evocative of a (less sophisticated) version of Mill's method of difference that Ragin (1987/2014) describes as a "thought experiment." In the thought experiment method one "contrast[s] an empirical case with an imaginary case representing a theoretically pure instance of the phenomenon of interest…the divergence of the empirical case from the imaginary case in causes is the experimental or treatment variable" (p. 39). The imaginary case is the ideal or most typical case, and the goal is to specify how the real case diverges from the ideal case in order to identify the underlying causes. For Brigitte, the empirical case is her life as a youth in Oaktown, and the imaginary case is an alternate reality in which she grew up in New York City. The causes for the divergence experiences in terms of housing, daily mobility, and perceptions of safety are reflected in the data. The hypothetical counterfactual followed a line of reasoning that we introduced in our oral history prompts when we asked, "In what ways do you think your life would have been different had your family stayed in their state or country of origin?" Several other participants pursued a similar approach in their family data storylines to imagine and explore different futures as teenagers if their families had not moved to Oaktown.

Time-jumping. In Carter's comparison of Mississippi, Illinois, and Tennessee in 1960, time and space are reordered nonsequentially in terms of the timeline for her family's geobiography. In aligning the family story to the big data, Carter reordered time and space in a comparison of where her grandparents grew up, where her mom grew up, and where she grew up in 1960, when her grandparents migrated northwards. She compares unemployment rates in Decatur in 1940 to unemployment rates in Decatur in 1990 and unemployment rates in Oaktown in 1990 to demonstrate a progressive increase in improvement in terms of economic conditions. Tennessee is persistent in her comparisons, even in years before her family members have arrived there. Rather, Carter aligned places that were significant for her family at different historical times (long ago, the more recent past, and the present) and are meaningful to her now. We have called this kind of comparative practice *time-jumping* (Kahn & Hall, 2016). Time jumping describes the kinds of cross-time comparisons that interrupt and rearrange continuous variation in measured variables. Notably, Social Explorer and Gapminder, the data technologies available in the workshop, both afforded these kinds of tangled comparisons.

Imputation and commensuration. Carter also applied interesting comparative logics in her family data storyline to explore aggregate-level economic conditions for why her family moved. Carter compared the county where she grew up and lives now (Magnolia County, Tennessee) to the counties where her grandparents and her parents grew up in Mississippi and Illinois, respectively. She compared these places across 1960, the year when her maternal grandparents migrated North, from Mississippi to Illinois, on a measure of household income (Figure 15). However, Social Explorer did not populate data for the county where her grandparents lived in Mississippi so she used what she considered to be an equivalent county in Tennessee, along the Mississippi River, for her comparison, as illustrated in Excerpt 13, described below.



Households: Less than $7,000
Census 1960 Tracts Only Set

Show data by: Tract
Visualization type: Shaded Area

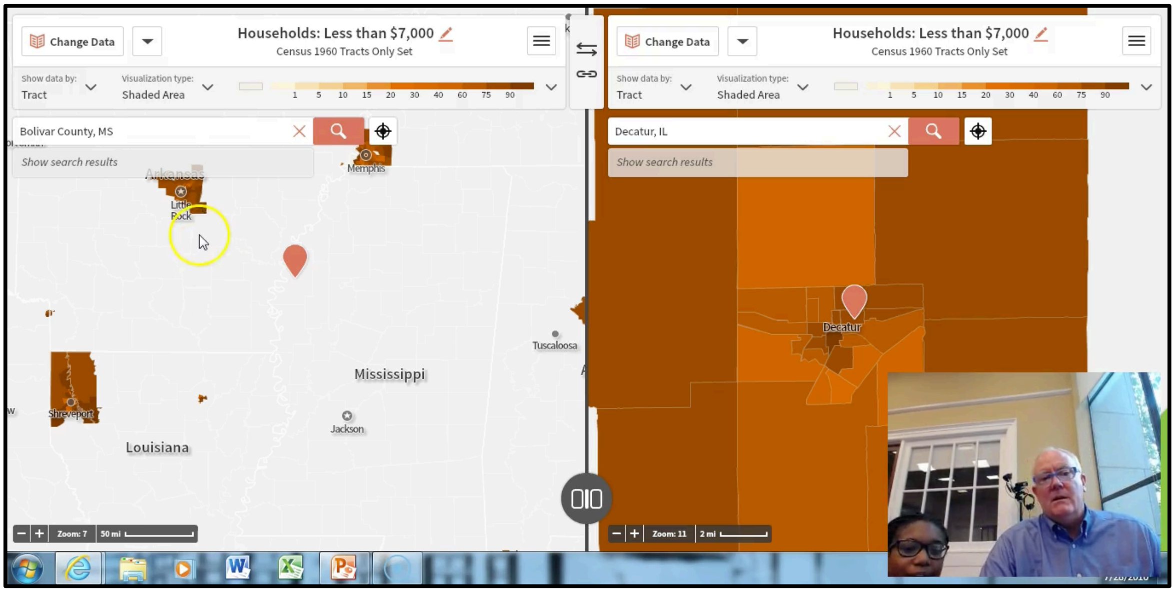
Households: Less than $7,000
Census 1960 Tracts Only Set

Show data by: Tract
Visualization type: Shaded Area

Figure 15. This was one of the map comparisons in Carter’s slides from her family data storyline. Carter compared the county where she grew up and lives now in Magnolia County, Tennessee (left) to where her mom grew up in Decatur, Macon County, Illinois (right) in 1960, the approximate year when her maternal grandparents migrated north from Mississippi. Data displayed is household income less than $7,000, close to the average income for a household in 1960 at the census tract scale.

Carter and Researcher 2 were seated next to each other. They were assembling comparisons together on Social Explorer on Researcher 2’s computer (to take screenshots of) and recording data points to insert into Carter’s PowerPoint. Researcher 2 was driving the mouse with Carter looking over his laptop. They have just completed a comparison of 1960 household income less than $7,000, with Magnolia County, Tennessee on the left and Macon County, Illinois on the right (Figure 15). They determined that $7,000 roughly equated to the national average household income for 1960. Data was displayed by census tract.

Researcher 2 then searched for Bolivar County, Mississippi, where Carter’s maternal grandparents lived before moving to Macon County, Illinois, to put on the on the left side of her comparison with Macon County, Illinois. Researcher 2 and Carter found there was no census tract data for Bolivar County, Mississippi.



Households: Less than $7,000
Census 1960 Tracts Only Set
Show data by: Tract
Visualization type: Shaded Area
Bolivar County, MS

Households: Less than $7,000
Census 1960 Tracts Only Set
Show data by: Tract
Visualization type: Shaded Area
Decatur, IL

Figure 16. Carter encountered missing data. There was no census tract data available for the Household Income data category in Bolivar County, Mississippi in 1960, where Carter’s maternal grandparents lived before moving to Macon County, Illinois.

Excerpt 13
[Week 3, Day 2]

1 Researcher 2: I don't think we're gonna have any data on Bolivar at the tract level, that's the problem with this thing

2 Researcher 2: oh, we might need to go ((*unintelligible*))

3 Researcher 2: We’ll zoom out to see if there's a city nearby that has track data ((*he zooms out and several clusters of data appear on the map*))

4 Carter: nope

5 Researcher 2: so little rock Arkansas does [have data] and ((*moving the frame of the map around*)) that's interesting. the data density at that time is kind of interesting like what they were=now wait a minute we didn't ((*pause, staring at screen*)) yeah no we're still in tract stats that's okay. So I mean you could you could say we::ll ((*rests fist on table, pauses for 4 seconds*)) I'd say that the closest most similar city would either be Little Rock Arkansas or Memphis Tennessee. Do you want to go to Memphis or you want to Little Rock? Bill Clinton or Elvis?

6 Carter: ((*laughs*)) ah Bill Clinton ((*softly*)) uh Elvis

7 Researcher 2: ELVIS ((*laughter from around the room; Researcher 2 zooms into Memphis*))

Carter observed that Memphis and Decatur areas look similar. Carter and Researcher 2 were next faced with the challenge of choosing individual tracts to compare (each tract has a unique percentage and count). Carter picked a census tract in Decatur randomly (the “eeny, meeny, miny, moe” method) that has 73.1% of households with an income of less than $7,000 in 1960 and records that in her PowerPoint. They had the same challenge for Shelby County, where Memphis is located, and there was great variation among the counties.

Excerpt 13 continued

[…] ((*transcript redacted for 3 minutes of talk))*

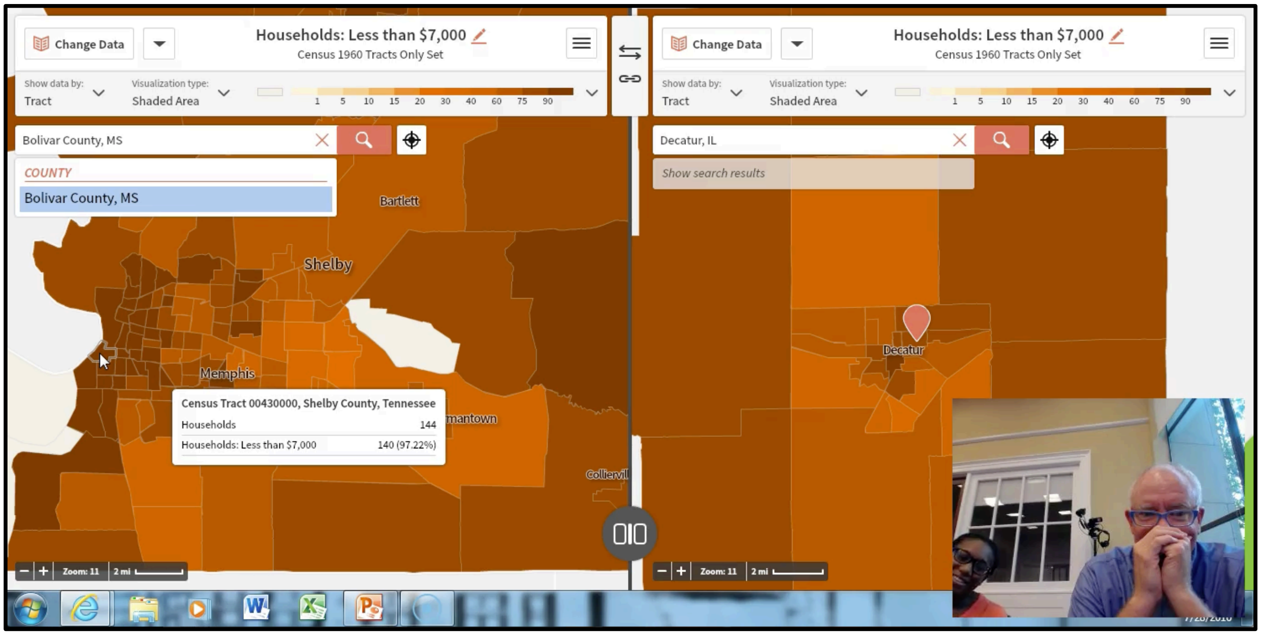
8 Researcher 2: a::nd it depends if you go down I think this is the Mississippi River ((*points out the Mississippi River in the map with L hand*)) and you go down near the river ((*moving cursor over darkly shaded tracts by the river to get measurements*)), and it's like [ninety-seven percent

9 Carter: [Yeah which that's where, well not where, MY people, because they lived down, they lived BY the river, so should I just do an eeny, meeny, miny, moe?

10 Researcher 2: well you could say if you think if your ancestors were living in Shelby County they would be near the River in which case here is one ((*highlighting a dark county near the Mississippi river with the cursor*)) for example that has almost three thousand and ninety-two point eight one percent of them have income below seven thousand dollars

11 Carter: mmkay

Carter subsequently recorded in her PowerPoint 92.81% for Shelby County “if my ancestors lived near the river” and noted that it is because they had “no data for MS.” In this excerpt, Carter and Researcher 2 recalled that Gunnison in Bolivar County, Mississippi, where Carter’s ancestors lived, like the dark census tracts in Memphis, Shelby County, Tennessee, sits by the Mississippi River. This informed their choice to use a data point in Shelby County, Tennessee as a proxy or equivalent (also could be considered an approximation) for Bolivar County, Mississippi. Statisticians would call their decision to use a similar case that does have data as a proxy for the original with missing data “imputation”—predicting or filling in missing data values on the basis of what is known about relevant covariation between the two cases (Allison, 2002). Researcher 2 and Carter used the shared geography of being by the Mississippi river to make Shelby County comparable to Macon and Magnolia counties in her family data storyline.



Households: Less than $7,000
Census 1960 Tracts Only Set
Show data by: Tract
Visualization type: Shaded Area

Census Tract 00430000, Shelby County, Tennessee
Households 144
Households: Less than $7,000 140 (97.22%)

Households: Less than $7,000
Census 1960 Tracts Only Set
Show data by: Tract
Visualization type: Shaded Area

Figure 17. Carter and Researcher 2 engaged in approximation, imputation, and commensuration. This map comparison was also in one of Carter's slides. The comparison shows data in Shelby County, Tennessee on the left, a proxy country for where her maternal grandparents grew up in Mississippi, and Macon County, Illinois is show on the right, on household income less than $7,000. Social Explorer did not populate data for the county where her grandparents lived in Mississippi so she used what she considered to be an equivalent county close-by along the river in Shelby County, Tennessee for her comparison.

Carter and Researcher 2 still needed to select a data point for comparison in Magnolia County (Figure 17). They pulled up data for Oaktown in 1960, households that make less than $7,000 again on the left side. Decatur was still on the right, and only census tract data for the county areas were visible. They looked both at tracts where the university that her parents attended is located and where she lives today in southeast Oaktown. They found that the tract with the university, which was and is located in a predominantly African American neighborhood, had a very high percentage of households that earn less than $7,000 (94.482%); the percentage in the tract where Carter's current home is located was lower (80.22%), making the contrast between the origin and final destination locations—Mississippi and Oaktown—greater, which better supported her storyline and excited Carter. Carter selected the tract that contained her home address today and used that data point as the 1960 Oaktown percentage for households with income less than $7,000, which she annotated with "(approximately where I live now)."

Time jumping and imputations are possible in Social Explorer because the data is commensurable or able to be measured by a common standard. Commensuration, or "the transformation of different qualities or measurements represented by different units into a common metric" (Espeland & Stevens, 1998, p. 314) is a form of data wrangling. Carter and Researcher 2 commensurated places on the basis of similar geography (down by the river), racial segregation and demographics (mostly African-American census tracts), and economic opportunity (low). Furthermore, they were able to do this because of the commensuration (and data wrangling) that has already been completed for Social Explorer users: Social Explorer and US Census Bureau datasets make cities and nations numerically comparable over time using percentages and statistics.

Comparative Logics and Critical Perspectives

As a result of the data wrangling process, each of the participants' family data storylines presented a particular stance toward the family's relationship to society, as represented by the aggregate data trends in Social Explorer. Returning to the Critical Positioning Matrix (Figure 2), all six cases produced family data storylines that sit in the upper left corner of the Critical Positioning Matrix ("All is swell; data+, my family+). In general, participants produced narratives of improvement over time; from their standpoints as children of the middle or upper-middle class, economically and socially, life got better for their families (Daniel, does not for instance, spend time on the outcome of the others in his story, like the Native Americans, whose land was taken away by the U.S. government). Their storylines traced their parents or grandparents moving for higher wages, better job opportunities, better education opportunities, and larger homes. This logic could differ for a more diverse sample of youth and family stories, which can be further investigated in future iterations.

Generally, storylines explored what historians would argue are the effects of the entrenched, institutional racism that motivated the Great Migration for households and families (lack of jobs, low salaries, desire for improved safety for Black families), but youth not did not talk about racism itself. A few participants (none of the six cases) looked at employment or income data specifically for African Americans. Isis, in what may have been the most explicit reference to racialized experiences, wrote about her grandmother having a hard time at school after moving as part of the Great Migration from Alabama to upstate New York, which she noted was predominantly white; Isis wrote her grandmother felt that the Northern, White students treated the teachers poorly.

However, in data wrangling, there were moments for reflection on the intersection of inequity and race, even if these considerations did not make it into the final product. These conversations typically involved looking at data within cities displayed by census tract. For instance, the persistence of low incomes for predominantly African American neighborhoods, compared to other census tracts, in Oaktown from the 1960s until today surprised Carter. During Naimah's presentation, the instructors prompted her to zoom into Evanston, Illinois, where her grandmother moved to from the South. Naimah's move to change the spatial scale from the county level to the census tract level revealed internal variation among neighborhoods in terms of the concentration of African Americans in particular areas. Similarly, Hannah, in a data wrangling session with Researcher 1, zoomed from county level to census tract level for college attainment in the Chicago area; again, changing spatial scales produced a discussion with a researcher about the internal variation and distribution of education attainment within and across Cook County (Chicago area). We view the practices of hiding local variation by selecting a larger spatial scale, or searching for variation (relevant to your question) at smaller scale or grain size as a valuable data wrangling practice. We also view changing scales to conceal or reveal variation as a new critical data literacy topic with big data and dynamic visualization tools.

For our focal participants, assembling family data storylines raised new questions for family members about the story they had just told: Carter wanted to know "How did my family start out in Mississippi in the first place?" Hannah wanted to know "Why did my family choose Chicago?" Both wanted to learn more about their great-great grandfather whose parents were from China. Daniel wanted to know more about "Why did my family never decide to move out of their original town?" While these were not critical questions, one of the utilities of the workshop design was that it motivated participants to have conversations with parents and grandparents about their families' choices and decisions.

Likewise, while youth participants did not dwell on the obstacles their families might have faced in the workshop, these challenges surfaced in the family interview discussions of the teens' work. For example, while examining a map comparison of the Black population in Cook County in 1920 and 1970, a comparison that Naimah had made during the workshop, Researcher 1 suggested that there was and is still de facto segregation in the North. Naimah and Isis' parents indexed this to what they call the "up-south theory," to describe when people in the North "ha[ve] racial or racist tendencies [like] so much of the south" (Akinjide [Dad]) and "that it's full of people who are Southern in ways of thinking" (Zuri [Mom]); they shared that they had experienced this when they lived in upstate New York.

Similarly, in our conversation about Carter's comparison of unemployment data in Bolivar County, Mississippi and Macon County, Illinois in the family follow-up interview, Hannah's and Carter's mom Summer indicated that the unemployment rate, which was higher in Decatur then Mississippi n the 1940s, did not reflect their family's experience; Summer said it was even harder for African-Americans in Mississippi to get a job that was not picking cotton or farm labor than in Decatur, where small businesses, restaurants, and the railroad would hire African Americans. In the family follow-up interview with Brigitte and Daniel, their mom Kara talked to them about Oaktown's low walkability and car dependency. Kara told the story that when she first moved to Oaktown from Brooklyn, New York, friends of hers who were seasoned Oaktown residents laughed at her when she tried to walk to the nearby pharmacy a mile down the road instead of taking the car. Families served as powerful sources of knowledge and learning in getting personal with big data.

Moreover, participant inquiries and stories were of great cultural significance and personal consequence or impact. For instance, Hannah compared Chicago and Oaktown on a number of indicators, including college attainment, house value, and number of families living in poverty; her father's family had faced economic hardships on the South Side of Chicago, and her paternal grandfather was a Chicago Police Department police officer who was killed in the line of duty. She understood her father's choice to live in Oaktown and raise his kids there meant that she and her sister would live without being surrounded by crime and poverty.

Discussion

We demonstrate in this paper that getting personal with big data is a primary way for entering the world of large-scale demographic and socioeconomic databases. Participants found the geobiography to be compelling site for storytelling and modeling with big data and dynamic visualization tools. Youth continuously refined their family data storylines, through data wrangling and through conversations with siblings, peers, and family members. Despite material and conceptual challenges that they encountered, participants successfully aligned their family and personal histories to the aggregate. They thoughtfully assembled comparisons and told stories that scaled between themselves or their families and society.

For our focal participants, assembling family data storyline raised new questions for family members about the story they had just told. One possibility is that despite participants' efforts to wrangle data that confirmed an expected storyline (e.g., economic conditions were better in the place my family moved to), the data produced a different or more ambiguous story (e.g., unemployment rates were the same across states). In some of these cases, one of the main reasons that families moved or stayed was to be near other family members, which was not captured in the data.

Participants' stories and comparisons presented tangled weaving of past, present, and future. Their storylines often traversed participants' ancestral history, their present lived experiences, and their imagined future, such as places they will be in 10 years. This alignment of the past, present, and future self is a product of a *hybrid third space* (Gutiérrez, 2008: Leander, 2001) that our design elicited, in which students' home and cultural practices and their emerging cultural identities came into contact with the sociotechnical practices for storytelling and modeling with big data. Our hybrid third space design invited participants to stand in for their ancestors, to relate their lives to their own lived experiences, and to become an embodied data point on in historical and present time and place.

The relationship between participant data exploration and critical stances towards social and cultural issues was fuzzy. Teens had the agency to tell the story they wanted to tell, and participants pursued issues of meaning and consequence for them. Furthermore, many of the participants' stories were grounded in their experiences as youth of color, who are underrepresented in the city's historical archives, and these stories are now a part of Oaktown's public history. However, though our design invited conversations of race and equity, there is not strong evidence that counter-narratives or counter-models that critique social injustice emerged in the exploration of the aggregate and themselves. We offer several thoughts on why this might have been:

- Who the data counted and excluded (Uprichard, 2013) was not always clear. For instance, Social Explorer does not differentiate between slaves and free people in their counts of non-white populations as the US Census did, nor does it present Native American populations as present in the either states or "empty" territory until 1890.
- "What moves families?" was a strong an invitation for getting personal with big data, but it did not carry an explicit social justice framework.
- Counter-modeling is not a spontaneous phenomenon. There is a lot to learn in order to be able to build a critique of society with large-scale datasets and do this in a way that would stand up to competitive argumentation.
- Teens had the agency to tell the story about the geobiography they desired.
- Participants may not have been comfortable with discussing issues of race or histories of racism with a majority White research team.

- The research/instructional team missed opportunities to model how racialized experiences that came up in our conversations could be connected to the aggregate scale with a data tool.

Nonetheless, we found that getting personal with big data opens pathways for those kinds of critical conversations with parents. Across parent-child interactions captured on tape, discussions about the history of why our family moved led to sharing of memories and teaching moments, as did family explorations of data together in the follow-up interviews. Like in Radinsky et al. (2014), participants' lived experiences, memories, or conversations with parents comprised "a meaningful source of knowledge for social inquiry" (p.152) for their engagement with the big data. Youth, through their interactions with siblings and parents, developed and expanded what Fivush, Bahanek, and Duke (2008) call "the intergenerational self." The intergenerational self—which, they argue, develops in early adolescence—refers to one's sense of self or personal history that is defined by family history. This sense of self develops through the kinds of family stories and reminiscing we found in the workshop and interviews. Fivush et al. (2008) suggest that these stories, by building connections among family members and by positioning youth in intergenerational contexts, "create meaning beyond the individual" that produces a sense of self in historical time and in relation to parents, grandparents, and other ancestors (p. 134). In turn, in future design iterations, we want to create more opportunities for families to engage with youth around the data and build-in occasions for youth to pursue follow up questions.

Moving forward, if future design iterations were to expand participation and contexts to include communities with different geobiographies and social histories, such as members of transnational or immigrant communities, we would need to be sensitive to youth for whom this information would be inaccessible or un-sharable, such as children in foster care or undocumented immigrants. Even in the present study, we found that a certain degree of depth in one's family history was inaccessible for some of our African-American participants as compared to the White European participants and researchers. Who has access to a well-documented family history is related to broader equity issues of which families have the rights to tell history and who gets to decide what is told and what counts as important for understanding the family.

In future iterations, we also might rethink how we approach youth with certain forms and aspects of data wrangling and strategies for comparisons: encountering missing or spotty data, the difference between percentages and counts, practices like changing scale to reveal variation, time-jumping, approximation, commensuration, or imputation. We could design so that challenges are made visible and discussed through the activities and comparative strategies or practices are explicitly addressed in pedagogy. For instance, we could discuss together moving around in time between survey years in order to find complete data for a desired comparison is tolerable, as is finding another region, in a similar state, that does have data for the year in question (commensurating regions).

Another possible way to support critical or counter-perspectives is to frame the critical positioning matrix (Figure 3) as a learning objective for participants. That is, we can openly support learners in a critical assessment of personal fit (or exclusion) with the aggregate (society) as one of the objectives of the data exploration. Such a design would intentionally scaffold participants in evaluating personal fit to the aggregate data in their models with explicit questions regarding how one's family compares to the data in order to prompt thinking about other populations. Participants would be in the position of asking and discovering if their family experiences correspond (or not) to society's (positive or negative) experiences as represented by the big data.

Our study and analysis had its limitations. We did not describe developmental differences between siblings, although differences in storylines and approaches could be attributable to developmental differences in storytelling, argument, and data fluency. A future analysis can track the functional structures of family data storylines across participants of different ages, including the degrees to which participants are able to organize both language and data to form a narrative account about the family in relation to a conjecture or argument about social history. Additionally, while our study highlights the value of intergenerational figurations for learning sciences design, a separate analysis that dives more deeply in the possibilities for family engagement and intergenerational learning around storytelling and modeling with big data is also warranted. Further research is needed to examine both the opportunities and limitations of big data as a technoscientific object and the design possibilities for youth learning.

Conclusion

What moves families is and will continue to be a relevant and meaningful topic for data science exploration for youth and adults. The demographics of our cities and states continue to shift[19], and conversations that bridge local and global perspectives are increasingly important. Our research suggests that storytelling and modeling with big data activities can operate within a

[19] Allen (2017) reported that Black millennials today are undertaking a reverse migration from expensive, northern cities, like New York, to increasingly liberal-minded, growing Southern cities, like Atlanta and Dallas. This account follows Carol Stack's (1996) ethnographic narrative of African American families returning to the South in the 1970s–1990s to "reclaim" the land, physically and spiritually. Notably, both of these migration paths diverge from the storylines of youth in our study, many of whom focused on their grandparents or great-grandparents decisions to move north and did not frame their own upbringing in Oaktown as a "return" to the South. At the same time, Allen, Stack, and our participants' storylines together illuminate the perpetual nature of family migrations, particularly in the US, over generations.

dynamic and diverse demographic, political, economic, social, and cultural landscape in order to inform understandings of both personal and shared human experiences. Moreover, continued research is needed to move towards a fuller grounded theory of understanding oneself, in terms of the geobiography and family history, in relation to society that does not yet exist in the youth learning and data science space. Continued support for open big datasets and interactive visualization tools that are central to supporting these kinds of activities is needed as well. Getting personal with big data offers a novel and promising approach to building relations between the self and society through modeling and storytelling—an approach that humanizes the technocratic practices of data science.

REFERENCES

Al-Aziz, J., Christou, N. and Dinov, I.D. (2010). SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *Journal of Statistics Education, 18*(3), 1-29.

Allen, R. (2017, July 8). Racism Is Everywhere, So Why Not Move South? Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/03/27/AR2009032701576.html

Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 193-196.

Azevedo, F. S., & Mann, M. J. (2017). Seeing in the Dark: Embodied Cognition in Amateur Astronomy Practice. *Journal of the Learning Sciences*, (Accepted).

Autry, R. (2017, November 5). How Racial Data Gets 'Cleaned' in the U.S. Census. Retrieved from https://www.theatlantic.com/technology/archive/2017/11/how-racial-data-gets-cleaned/541575/

Bamberg, M. G. (1997). Positioning between structure and performance. *Journal of Narrative and Life History*, *7*(1-4), 335-342.

Bang, M., & Vossoughi, S. (2016). Participatory design research and educational justice: Studying learning and relations within social change making.

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

Barron, B., Martin, C. K., Takeuchi, L., & Fithian, R. (2009). Parents as learning partners in the development of technological fluency.

Becker, H. S. (2002). Visual evidence: A Seventh Man, the specified generalization, and the work of the reader. *visual studies*, *17*(1), 3-11.

Becker, H. S. (2007). *Telling about society*. University of Chicago Press.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, *43*(2), 79-120.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.

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.

Busch, L. (2014). A dozen ways to get lost in translation: Inherent challenges in large-scale data sets. *International Journal of Communication*, *8*, 18.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative research. *London: SagePublications Ltd*.

Chow, K., & Healey, M. (2008). Place attachment and place identity: First-year undergraduates making the transition from home to university. *Journal of Environmental Psychology*, *28*(4), 362-372.

Cicourel, A. V. (1981). Notes on the integration of micro-and macro-levels of analysis. In K. Knorr-Cetina, & A. V. Cicourel (Eds.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies* (pp. 51-80). NY, NY: Routledge.

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

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, *92*, 28-40.

Daileda, C. (2016, May 19). Tech is dominated by even more white dudes than the rest of the private sector. *Mashable*, *Inc*. Retrieved from: http://mashable.com/2016/05/19/diversity-report-silicon-valley-white-men/#aXiwvyehCuqw

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to "Big Data." *Society and Space open site*.

Davidian, M., & Louis, T. A. (2012). Why statistics?. *Science*, *336*(6077), 12-12.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., ... & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences*, *19*(1), 3-53.

DiGiacomo, D. K., & Gutiérrez, K. D. (2016). Relational equity as a design tool within making and tinkering activities. *Mind, Culture, and Activity*, *23*(2), 141-153.

DiSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *The journal of the learning sciences*, *13*(1), 77-103.

Donovan, K. P. (2012). Seeing like a slum: Towards open, deliberative development. *Georgetown Journal of International Affairs*, *13*, 97.

Duke, M. P., Lazarus, A., & Fivush, R. (2008). Knowledge of family history as a clinically useful index of psychological well-being and prognosis: A brief report. *Psychotherapy: Theory, Research, Practice, Training*, *45*(2), 268.

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review*, *5*(1), 1-24.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, *20*(4), 399-483.

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.

Erickson, T. (Ed.), 2012. *Signs of change: History revealed in U.S. Census data*. CreateSpace Independent Publishing Platform

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, *24*(1), 313-343.

Fivush, R., Bohanek, J. G., & Duke, M. (2008). The intergenerational self: Subjective perspective and family history. In Sani, F. (Ed.), *Self continuity: Individual and collective perspectives* (pp. 131-143). NY, NY: Taylor and Francis.

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.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry*, *12*(2), 219-245.

Geertz, C. (1983). *Local knowledge: Further essays in interpretive anthropology* (Vol. 5110). Basic books.

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

Glaser, B. G., & Strauss, A. L. (1977). *The discovery of grounded theory: Strategies for qualitative research*. London: Aldine.

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

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606-633.

Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday* 16(2). http://firstmonday.org/

Gutiérrez, K. D. (2008). Developing a sociocritical literacy in the third space. *Reading Research Quarterly*, 43(2), 148-164.

Gutiérrez, K. D., & Jurow, A. S. (2016). Social design experiments: Toward equity by design. *Journal of the Learning Sciences*, *25*(4), 565-598.

Gutstein, E. (2006). *Reading and writing the world with mathematics: Toward a pedagogy for social justice*. NY, NY: Routledge.

Hall, R., & Leander, K. (2010, July). Scaling practices of spatial analysis and modeling. In R. Hall, R. (Chair), *Scaling practices of spatial analysis and modeling*. Symposium conducted at the 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.

Hall, R., & Stevens, R. (2016). Interaction analysis approaches to knowledge in use. In A. A. diSessa, Levin, M., & Brown, N. J. S. (Eds.), *Knowledge and interaction*: *A synthetic agenda for the Learning Sciences* (pp. 72-108). New York: Routledge.

Hammerman, J. K. (2009). Statistics education on the sly: Exploring large scientific data sets as an entrée to statistical ideas in secondary schools. *IASE/ISI Satellite*.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT press.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 375-398.) NY, NY: Berg.

Ingold, T. (2011). *Being alive: Essays on movement, knowledge and description*. Taylor & Francis.

Johansson, V. (2012). *A time and place for everything?: Social visualisation tools and critical literacies* (Doctoral dissertation). Retrieved from http://hdl.handle.net/2320/11462. (ISSN 1103-6990)

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*(4), 263-274.

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

Kahn, J. (2014). "What in the world?" Animated worlds in multivariable modeling with motion chart graph arguments. In J. L. Polman, E. A. Kyza, D. K. O'Neill, I. Tabak, W. R. Penuel, A. S. Jurow, K. O'Connor, T. Lee, & L. D'Amico (Eds.), *Learning and Becoming in Practice: Proceedings of the International Conference of 11th International Conference of the Learning Sciences* (pp. 1649-1650). Boulder, CO: International Society of the Learning Sciences.

Kahn, J. B., Hall, R., & Phillips, N. (2014). *Dissecting, remixing, and making graph arguments using motion charts and public data about global wealth and health*. Paper presented at the meeting of the American Educational Research Association, Philadelphia, PA.

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.

Kahn, J., Hall, R., & Pearman, F. A. (2016, April). Telling the city with big data. In R. Hall (Chair), *Re-inscribing the city in design studies of critical steam conceptual practice.* Symposium conducted at the American Education Research Association 2016 Annual Meeting, Washington D.C.. Available in the AERA Online Paper Repository.

Klein, H. K., & Kleinmann, D. L. (2002). The social construction of technology: Structural considerations. Science, Technology and Human Values, 27(1), 28–52.

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

Koschmann, T. (Ed.). (2011). *Theories of learning and studies of instructional practice* (Vol. 1). Springer Science & Business Media.

Krishnan, G. (2015). *Designing a Mobile Makerspace for Children's Hospital Patients: Enhancing Patients' Agency and Identity in Learning* (Doctoral dissertation). Retrieved from ProQuest Dissertations Publishing (10004920).

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

Law, J. (2004). *After method: Mess in social science research*. Routledge.

Leander, K. M. (2001). "This is our freedom bus going home right now": Producing and hybridizing space-time contexts in pedagogical discourse. *Journal of Literacy Research*, *33*(4), 637-679.

Lee, V. R. (2013). The Quantified Self (QS) movement and some emerging opportunities for the educational technology field. *Educational Technology*, (November-December 2013), 39.

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model?. In M.K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 9-22). Springer US.

Leonelli, S., Rappert, B., & Davies, G. (2017). Data shadows: Knowledge, openness, and absence. *Science, Technology, & Human Values*, *42*(2)*,* 191-202.

Lewis, David K. 1973. *Counterfactuals*. Cambridge: Harvard University Press.

Linde, C. (1993). *Life stories: The creation of coherence*. Oxford University Press.
Little, D. (2004). Counterfactuals. In M. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. 206-207). Sage Publications.

Nazario, S. (2006). *Enrique's Journey*. New York: Random House, LLC.

National Center for Education Statistics. (2015). *Integrated Postsecondary Education Data System* [Data File]. Retrieved from https://datausa.io/profile/cip/11/#demographics.

Ochs, E., & Capps, L. (1996). Narrating the self. *Annual review of anthropology*, *25*(1), 19-43.

Ochs, E., Taylor, C., Rudolph, D., & Smith, R. (1992). Storytelling as a theory-building activity. *Discourse processes*, *15*(1), 37-72.

Peppler, K. (2013). New Opportunities for Interest-Driven Arts Learning in a Digital Age. Wallace Foundation.

Peppler, K., Halverson, E., & Kafai, Y. B. (Eds.). (2016). *Makeology: Makerspaces as learning environments* (Vol. 1). Routledge.

Pew Research Center. (2015, June 210). What Census Calls Us: A Historical Timeline. Retrieved from http://www.pewsocialtrends.org/interactives/multiracial-timeline/

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.

Pickles, J. (2006). Ground Truth 1995–2005. *Transactions in GIS*, *10*(5), 763-772.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.

Pickles, J. (Ed.). (1995). *Ground truth: The social implications of geographic information systems*. Guilford Press.

Polman, J. L, Gebre, E. H., Rubin, A., Hinojosa, L., Sommer, S., & Graville, C. (2016, April). *Organizing data journalism activity in school and community learning environments to contextualize science in life*. Poster session at the American Education Research Association 2016 Annual Meeting, Washington D.C..

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.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Radinsky, J. (2008a). GIS for History: a GIS learning environment to teach historical reasoning. In M. Alibrandi, & A. Milson (Eds.), *Digital geography: Geo-spatial technologies in the social studies classroom* (pp. 99–117). Greenwich, CT: Information Age Publishing.

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

Ragin, C. C. (2014). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press. (Original work published 1987)

Rogoff, B., Callanan, M., Gutiérrez, K. D., & Erickson, F. (2016). The organization of informal learning. *Review of Research in Education*, *40*(1), 356-401.

Rosling, H. (2007, March). *Hans Rosling: New insights on poverty* [Video file]. Retrieved from https://www.ted.com/talks/hans_rosling_reveals_new_insights_on_poverty

Rosling, H., Ronnlund, A.R. & Rosling, O. (2005). New software brings statistics beyond the eye. In E. Giovannini (Ed.), *Statistics, knowledge and policy: Key indicators to inform decision making*, (pp. 522-530). Organization for Economic Co-Operation and Development.

Rubel, L. H., Hall-Wieckert, M., & Lim, V. Y. (2017). Making Space for Place: The Role of Mapping Tools in Learning as Political Formation. *Journal of the Learning Sciences*, (Accepted).

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.

Schegloff, E. A. (1991). Conversation analysis and socially shared cognition. *Perspectives on socially shared cognition*, *150*, 171.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 361-382.

Schlossberg, N. K. (1981). A model for analyzing human adaptation to transition. *The counseling psychologist*, *9*(2), 2-18.

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

Selwyn, N., Gorard, S., & Williams, S. (2001). Digital divide or digital opportunity? The role of technology in overcoming social exclusion in US education. *Educational Policy*, *15*(2), 258-277.

Solórzano, D. G., & Yosso, T. J. (2002). Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative Inquiry*, *8*(1), 23-44.

Stack, C. B. (1996). *Call to home: African-Americans reclaim the rural South*. Basic Books.

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

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, *19*(3), 387-420.

Stevens, R. (2010). Learning as a Members' Phenomenon: Toward an Ethnographically Adequate Science of Learning. *Yearbook of the National Society for the Study of Education*, *109*(1), 82-97.

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

Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of qualitative research*, *17*, 273-85.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist*, *62*(4), 271.

Svennevig, J. (2000). *Getting acquainted in conversation: a study of initial interactions* (Vol. 64). John Benjamins Publishing.

Taumoepeau, M., & Reese, E. (2013). Maternal reminiscing, elaborative talk, and children’s theory of mind: An intervention study. *First Language*, *33*(4), 388-410.

Taylor, K. H. (2017). Learning Along Lines: Locative Literacies for Reading and Writing the City. *Journal of the Learning Sciences*, 1-42.

Taylor, K. H., & Hall, R. (2013). Counter-mapping the neighborhood on bicycles: Mobilizing youth to reimagine the city. *Technology, Knowledge and Learning*, *18*(1-2), 65-93.

Tuck, E., Smith, M., Guess, A. M., Benjamin, T., & Jones, B. K. (2014). Geotheorizing Black/Land. *Departures in Critical Qualitative Research*, *3*(1), 52-74.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond Vocabulary: Exploring Cross-Disciplinary Academic-Language Proficiency and Its Association With Reading Comprehension. *Reading Research Quarterly*, *50*(3), 337-356.

UNHCR. (2015). *2014 annual global trends report of the UN Refugee Agency*. Retrieved from http://unhcr.org/556725e69.html

Uprichard, E. (2013). Focus: Big data, little questions? *Discover Society*, (1).

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.

Vossoughi, S., Hooper, P. K., & Escudé, M. (2016). Making through the lens of culture and power: Toward transformative visions for educational equity. *Harvard Educational Review*, *86*(2), 206-232.

Wager, A. A. (2012). Incorporating out-of-school mathematics: From cultural context to

Wertsch, J.V. (1998). *Mind as action*. Oxford University Press.

Wilkerson-Jerde, M., & Laina, V. (April, 2015). *Stories of our city: Coordinating youths' mathematical, representational, and community knowledge through data visualization*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Williams, S., Deahl, E., Rubel, L., & Lim, V. (2015). City Digits: Local Lotto: Developing Youth Data Literacy by Investigating the Lottery. *Journal of Digital and Media Literacy*.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625-636.

Yin, R. K. (2000). Rival explanations as an alternative to reforms as "experiments." In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy (pp. 241=268).* Thousand Oaks, CA: Sage Publications.

CHAPTER V

CONCLUSION

The three papers that comprise this dissertation collectively introduce a design space for learning with open, socioeconomic big data and data visualization tools. This body of research began with an observational study of novel forms of modeling with big data and tools that supported provocative ways of reporting and making social arguments—telling about society (Becker, 2007)—in the public media. Public presentations by Swedish public health professor, Hans Rosling, featured at the center of this initial work. The "discovery" of these modeling practices motivated experimental teaching across a multi-iteration, multi-year design based research program (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). We (the research team) conducted storytelling and modeling activities with big data, first with preservice secondary mathematics teachers and then with preservice secondary social studies teachers, as reported in Chapter 3. In both of these studies, we invited participants to use an open, online, data visualization tool for modeling global and national socioeconomic big data (Gapminder; Rosling, Ronnlund, & Rosling, 2005) to tell stories about global development. The comparative analysis of video records across Hans Rosling and preservice teachers found that connecting personal experiences to aggregate trends described in the model (*getting personal*) can support telling stories about society that counter, challenge, or critique dominant or conventional social narratives—what we call *counter-modeling*.

This finding prompted a couple short design experiments with middle and high school youth in which we explored the *geobiography* as a personal context for learning with big data (not reported in full in the dissertation; for an excerpt from this work, see introductory vignette in Chapter 2). The design experiments preceded the most recent design study iteration (Chapter 4) that asked teenage youth in the public library to create *family data storylines* about personal family mobility in relation to national census trends using Gapminder or Social Explorer, an online, visualization and mapping tool that displays datasets from the U.S. Census Bureau. Our study found that getting personal with big data is a primary way for entering the world of large-scale demographic and socioeconomic databases. Furthermore, assembling models and stories that relate family experiences to society entails complex representational practices. Participants wrangled data in order to produce comparisons that scaled time (i.e., past, present, future), space (i.e., neighborhoods, states, countries) and social life (i.e., my and my family's experiences, society's experiences, humankind's experiences) in order to build a coherent narrative to explain why their families moved. Their strategies for comparisons addressed uncertainty (Star, 1985) surrounding the data, knowledge of the family story, or historical events and leveraged the data tools' capacities for dynamically displaying data at different scales (i.e., grain-sizes).

Moreover, the overall object of inquiry for this dissertation was to explore how youth and adults learn to engage in the interdisciplinary representational practices that support becoming modelers, storytellers, and consumers of stories told with big data. In pursuit of this objective, the dissertation offers several novel contributions to the learning sciences and education research fields:

First, the dissertation establishes storytelling and modeling with big data as a new *cultural activity* (Engeström & Sannino, 2007). Chapter 2 describes how large-scale datasets and data visualizations tools permit sophisticated, critical, and relevant modeling of socioeconomic phenomena. As data and tools become available to the public, this data science activity space becomes increasingly accessible to youth and adults. Chapters 3 and 4 give examples of designs that illustrate new opportunities for rich, interdisciplinary learning with big data and data science technologies.

Second, this dissertation extends the framework for learning about big data by Philip, Schuler-Brown, and Way (2013). Rather, the current work investigates contexts for learning *with* big data *about* self and society relations. The dissertation presents getting personal with big data as a way to advance an understanding of individual experiences in relation to broader social, economic, and political conditions. The empirical examples described demonstrate that modeling with socioeconomic, big data and interactive data visualization tools can interface micro (individual) and macro (aggregate) phenomena in meaningful and critical ways. Furthermore, getting personal with big data establishes self–society relations as a target of future data science education and technology investigations.

The three papers also collectively address Philip et al.'s (2013) argument that critical data literacy and big data technologies are potential means for civic and democratic participation. Our design gave youth and young adults the agency to tell stories and build models about the social world, which in some cases, led to assertions about fairness and justice (preservice teachers in Chapter 3) or conversations around issues of equity and race (Chapter 4). However, how to support counter-modeling with big data still remains an open question. Nonetheless, the dissertation studies establish a deeper understanding of modeling with big data as socially, historically negotiated activity that entails assuming a value-based position or moral stance towards socioeconomic change. Additionally, these papers expand the population of learners with big data beyond young adults in STEM classrooms. In particular, Chapter 3 found that intergenerational family figurations were valuable to storytelling and modeling with big data activities and possibly key to seeding critical understandings about the relationship between personal experiences and larger social, economic, and historical issues.

Lastly, the dissertation reveals a diverse set of representational practices that support “learning and becoming” (Gutiérrez, & Jurow, 2016; Lave, 1991) modelers, storytellers, and consumers of stories told with big data. The qualitative analyses uncovered and described participants’ *data wrangling* practices to identify data that describe social conditions or influence, their logics for building comparisons in their models and maps, and the ways in which they animate the represented and representing worlds in their storytelling performances (Gravemeijer, 1994; Hall, 1999; Hall, 2000). In their stories and with their maps or models, learners (and Hans) traversed temporal, spatial, and social scales in service of building theories (Ochs, Taylor, Rudolph, & Smith, 1992) that explain the relationship between the individual and society. The findings in Table 1 look across the dissertation cases and summarize (a) the stories participants told about society, (b) how they positioned themselves and possibly others to that story, and (c) the conceptual practices they employed to express that relationship and narrative with a data model; all of these aspects are pieces of a grounded theory of learning to model and tell stories with big data about society.

Table 1

A Cross-Case Comparison of Storytelling and Modeling With Big Data

	Hans Rosling	Preservice Teachers	Library Youth
Narrative told with data about society	There is no developed and developing world.	US and OECD countries should share culpability for world's problems.	Why did my family move? Was it related to broader social forces?
Getting Personal	Health and wealth in bathroom and kitchen WEIRD vs. BRIC car purchases Hans' matriline	Peers' consumption habits	Youth and family mobility history
Conceptual Practices	Data wrangling (Johansson, 2012) Time-jumping Horse Racing	Data wrangling Time-jumping Horse Racing	Data wrangling Time-jumping Approximation Imputation Commensuration Hypothetical-counterfactuals Variation through changing scale

Moving forward beyond this dissertation, my research program will continue to ask/answer new questions about critical, representational, and conceptual practices with big data. In particular, social scaling with big data opens new, unique avenues in data modeling, and additional work is needed to better understand how to support data wrangling practices for future iterations. My research designs will also expand participation and contexts to include communities with different geobiographies and social histories. For instance, I would like to assemble family data storylines with participants from transnational or immigrant communities.

The goal would be to situate individuals from these communities as agents who can enact social and cultural change by understanding the processes represented in the data. This is how I envision this work as transformative for learners. Furthermore, I will continue to explore alternative data science tools that permit manipulation and visualization of a wide range of open socioeconomic or scientific datasets, such as datasets gathered and released online by city or state governments and other institutions, and that are ideal for youth and family learning.

We can now design for learning and participation "in the continuum between local exchanges and global trends" (Venturini, Jensen, & Latour, 2015, p. 2), a design space that we were not able to readily access before the public availability of large-scale datasets and visualization tools. We can seize the current opportunity for empowering youth and adults in becoming storytellers and modelers with big data. And, we can connect our personal histories with models made with big data to better understand ourselves as well as the social world and to tell powerful stories about society.

REFERENCES

Becker, H. S. (2007). *Telling about society*. University of Chicago Press.

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

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review*, *5*(1), 1-24.

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education*, 443-471.

Gutiérrez, K. D., & Jurow, A. S. (2016). Social design experiments: Toward equity by design. *Journal of the Learning Sciences*, *25*(4), 565-598.

Hall, R. (1999). The organization and development of discursive practices for "having a theory". *Discourse Processes*, *27*(2), 187-218.

Hall, R. (2000). Work at the interface between representing and represented worlds in middle school mathematics design projects. In L. R. Gleitman & A. K. Joshi (Eds.), Proceedings of Twenty-Second Annual Conference of the Cognitive Science Society (pp. 675–680). Mahwah, NJ: Erlbaum.

Lave, J. (1991). Situating learning in communities of practice. *Perspectives on Socially Shared Cognition*, *2*, 63-82.

Ochs, E., Taylor, C., Rudolph, D., & Smith, R. (1992). Storytelling as a theory-building activity. *Discourse Processes*, 15(1), 37-72.

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.

Rosling, H., Ronnlund, A.R. & Rosling, O. (2005). New software brings statistics beyond the eye. In E. Giovannini (Ed.), *Statistics, knowledge and policy: Key indicators to inform decision making*, (pp. 522-530). Organization for Economic Co-Operation and Development.

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

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.

Appendix A

Chapter II Coding Framework

[Time Stamp in Video Record] Model #, Story # (Story description)

Model variables
timeline =
x = (linear/log)
y = (linear/log)
bubble size =
color =

Horse Racing
- present/absent

then if present
- entities compared (in competition)
- scale used for comparison (type of quantity/quality)
- value expressed (winning means...)
- quality of enactment (event casting speech...)

Time Jumping
- present/absent

then if present
- time frame (how are points, intervals arranged)
- entities compared (within, across cases)
- what has changed (difference in quantities)
- value expressed (change means...)

Getting Personal
- present/absent

then if present
- entities compared (difference in social scale)
- nature of influence (weaker than cause, levels of analysis)
- positions offered audience (part of problem or solution, calls to action)

Counter-modeling
- present/absent

If present then
- Entities compared (data used as evidence)
- Counter to what (dominant social narrative)
- Story expressed (rival cases)

Appendix B

Chapter IV Family Data Storyline Workshops Participant Attendance

			Week 1		Week 2		Week 3		Sunday exhibit	Additional scheduled time outside workshop	Follow-up interview	Completed family data storyline
Participant	Age	Sibling Pair#	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2				
P1	13		x	x								
P2	13	SP1	x	x	x						x	x
P3	10	SP1	x	x	x						x	x
P4	13		x	x						x	x	x
P5	13		x	x						x		x
P6	12				x	x						
P7	14	SP2			x	x	x					x
P8	16	SP2			x	x	x					x
P9	10	SP3			x	x	x	x	x		x	x
P10	13	SP3			x	x	x	x	x		x	x
P11	13	SP4			x	x	x	x	x			x
P12	10	SP4					x	x	x			x
P13	14	SP4*			x	x						x
P14	12	SP5				x	x				x	x

P15	10	SP5				x	x				x	x
P16	13	SP6					x	x	x			x
P17	10	SP6					x	x	x			x

[*] Sibling Pair 4 was comprised of two siblings (P11 and P12) and one cousin (P13).

Appendix C

Chapter IV Day 1 Homework Assignment and Scaffold

Family Data Storyline Checklist:
For your **personal story**:

- □ What moved my family?
- □ What was better for your family after moving?
- □ What was challenging for your family after moving?
- □ What new question do you have about your family?

For your **data story**:

- □ What can we learn about the historical circumstances of my family or families like mine (or not like mine)?
- □ What new question do you have about your family?
- □ Are you making a fair comparison? Are you comparing counts or percentages/rates?
- □ Who is counted in the data? Who (or what) is not?

Appendix D

Chapter IV Oral History Guiding Questions

Questions/Prompts for Story Lines

(these are hints... tell the story that interests YOU)

- Where is your family from?

- What are some traditions that have been passed down in your family?

- Do you remember any classic family stories? When and where did this happen?

- How did your parents meet? Did they come from different places?

- How did your parents' childhoods differ from your own? Can you share any interesting stories about your parents growing up?

- What were your (or your family's) expectations when coming to Oaktown? America? Were they different from what you have experienced?

- What is the major thing that has moved your family? Or if you have family members that did not move, why did they stay?

- What were your (or your family's) expectations when coming to Oaktown?

- What was better for your family after moving?

- What was challenging for your family after moving?

- In what ways you think your life would have been different had your family stayed in their state or country of origin?

- If you could live anywhere in the world, where would it be?

- Where do you think you will be in 10 years? If not here, why move?

- Who is your best friend? How did they get to Oaktown?

- What school do you go to? What is your favorite thing about your school? What is your least favorite?

Appendix E

Chapter IV Family Interview Protocol

Prior to the interviews: Send PPTs and ask families to bring artifacts that will extend or deepen the family story in the interview.

Ask teens first, then parents about the family story:

1. How did your family come to live in Oaktown?
2. What is your artifact(s) and how is it related to your family storyline?
3. How are stories told in your family?

Pick a map comparison from their work and ask the teens:

4. How does this map comparison relate to your family?
5. Why did you choose it?
6. What is being compared?

Ask parents and youth:

7. Does the data represent your family story?
8. What does the data leave out?
9. Are there different data you would like to see for this story? (*Might want/need to ask, Do these data patterns differ by social or racial group?)
10. Looking at these data, do they change your understanding of your family story?

Select and play 30 second clips from oral history on themes of
(a) youth agency and their current life in the city and
(b) youth image of their future as an adult.
Ask youth, then parents:

11. How does your current life compare to what you think your parents experienced at your age?

Ask youth:

12. How do you think your future as an adult will be different from what your parents experience right now?