Connecting Research and Creating Frameworks:
A Report from the Youth, Learning, and Data Science Summit

Michelle Wilkerson & Kathryn Lanouette; University of California, Berkeley

Reasoning with and about data is an important competence in today's world. Data-driven visualizations, open data repositories, and infographics are being produced for public use (McGhee, 2010). Professional disciplines, from the sciences to the arts, are integrating computation, dynamic visualization, and "big data" into their everyday work (Hart, 2010; Hey, Tansley & Tolle, 2009; "Dealing with Data", 2011). And, public advocacy groups and local communities are examining the social, cultural, and ethical dimensions of data use - as a tool that can expose and address inequity, but also reproduce it (e.g., School of Data, schoolofdata.org; Data & Society, datasociety.net).

These shifts in the role of data in citizens' and practitioners' day-to-day lives have provoked calls for intentional research and community building efforts around K-12 Data Science Education, both within and beyond the educational research community (Finzer & Parvate, 2008; *New York Times*, 2010; 2014). On August 11-12, 2016, a group of researchers active in this emerging field across disciplines held the "Youth, Learning, and Data Science Summit" at the University of California, Berkeley. The meeting, funded through a capacity building grant from the NSF's Cyberlearning and Future Technologies Program (NSF IIS-1541676; Wilkerson, Lee, Parikh & Polman, 2015) explored the questions: *What makes data science education different?* And, *What does the general public need to know right now in order to be truly data literate?* Here, we review major themes emerging from that meeting. We concentrate on three fundamentally new aspects of data use that are transforming what needs to be taught and learnt about data: (1) so-called 'open' datasets are larger and more accessible than ever before; (2) new, powerful tools and techniques for working with and analyzing data emerge constantly; and (3) the social, connected nature of modern life leaves "data traces" about learners, their communities, and our world.

What is Data Science?

Data Science, as a field and as a concept, is still new and nebulous. There is general consensus that data science is emerging from the integration of a number of established fields including mathematics and statistics, computational science, and the natural and social sciences. A few commonly cited descriptions position data scientists as practitioners at the intersection of computation, statistics, and domain expertise (Conway, 2010; Friendly, 2008), and emphasize the role of computation in changing what it means to collect, manipulate, and work with data. For example, privacy, storage, archiving, visualization, and machine learning are core issues in scholarly dialogues about data science ("Dealing with Data", 2011; National Science Foundation, 2012). Discussions of data science also typically emphasize the exploratory (rather than only confirmatory) nature of work with data - a notion popularized by Tukey (1977), and now considered to be central to contemporary data-related work.

Research on Learners' Reasoning With and About Data

Though Data Science is new, decades of research have already shed light how young people think and learn about data, and have highlighted pedagogies and tools that help them engage in data-rich inquiry. Statistics Education Research, for example, examines how to support learners in developing understandings of statistical inference, measures of mean, spread, distribution, and trend in datasets, and connections between specific *cases* of data, the broader patterns that aggregations of those data reflect, and the contexts which those data and patterns describe. In science and mathematics education, researchers have examined data modeling as a way to engage learners in exploring the structures and relationships that underlie scientific and mathematical situations phenomena, such as chance and probability (Konold, et al., 2011) or natural variation in biology (Lehrer & Schauble, 2012). This work suggests it is important for learners to generate their own data, and work on iterative modeling and representation of that data as new findings and questions emerge. Also related are studies investigating "quantitative literacy" (Mathematical Association of America, 2008) or "information literacy" (Johnston & Webber, 2003). This work is primarily focused on undergraduate and adult learners, but emphasizes the importance of learners' reasoning by using mathematical ideas *in context* for the problems, policies, and situations in which they are most relevant.

Data Science Education: Differences, Challenges and Opportunities

Thus far, however, extant research on learning with data has mostly focused on students' meaning making with small datasets, collected by students themselves or designed especially for classroom use. Or, it has focused on whether learners correctly interpret and use public or pedagogical statistics and visualizations, with little attention to exploratory or critical examinations of data. Advances in data science, however, have introduced new dimensions to how students and citizens can expect to work with data in the coming decades: Datasets are larger, and more accessible, than ever before. New methods and tools for exploring and analyzing those data are emerging at a dizzying rate. And "data traces" of people's day to day activity are captured in social networks, personal logging devices, and environmental sensors. If this proposal is accepted, we will examine and offer examples of how these developments can challenge and extend extant literature on statistical reasoning, thinking, and learning.

Unprecedented Quantities of and Access to Data. Educators in statistics, mathematics, and the sciences have emphasized the importance of making inferences about a population from sample data, learning to relate a dataset to the context in which it was collected, or learning to construct and communicate about one's own dataset. Given the unprecedented size and access of contemporary data, however, these ideas take on new meaning. 'Open data' initiatives have released large data sets about public services, the environment, financial and social networks. This information can be useful for answering any number of questions beyond the original purview of data collection, but are also imperfect and incomplete for those purposes. Learners need to manipulate (clean, create subsamples, re-organize) these data to speak to their own interests. This may also involve piecing together datasets that describe the same phenomena at different levels of analysis or using different frameworks. Or, it may involve extending existing datasets to test new hypotheses about patterns found within them, or to uncover new relationships. Learners also need to understand how publically available data were collected, and how the interests or questions that drive the original data collection effort may influence its organization, selection, and presentation.

New Tools and Techniques for Working With Data. Educational efforts have also often concentrated on a small, but powerful, set of statistical tools (such as measures of mean and distribution) and techniques (such as examining descriptive statistics or plotting data using scatterplots and boxplots) as ways to make initial inquiries into data. There is now, however, a proliferation of new tools and techniques for exploring data: network visualization and analysis, interactive and multidimensional data representations, and free, open access data analysis tools are now common not only in the natural and social sciences but also in journalism, activism, and digital art. These tools and techniques make finding basic statistics, such as means, distributions, and regressions, and exploring data using a wide variety of known and novel representations trivial. At the same time, they require special competencies in themselves to be learnt, and are prone to be misused or misunderstood. And while existing research suggests exploratory tools (such as Tinkerplots) can be useful, classroom teachers cite a tension between the need for learners to spend times with tools that are useful and provide insight into statistical ideas and patterns, versus tools that might have less pedagogical value, but are ubiquitous (such as Excel or emerging data science tools like Tableau).

"Data Traces" and Activity, Agency and Activism. Finally, data are available and used in more personal, and personally relevant, ways that ever before. Learners can examine their own health, activity, social network, and achievement data. Looking beyond merely personal data, it is more apparent than ever that *all data are social*: constructed by humans, for human interests, and interpreted through the lens of our own cultures and experiences. Understanding these social aspects of data construction and interpretation offer learners new ways to become careful stewards of their own information and privacy, as well as agents and advocates empowered to use data to improve the world around them. They also offer students with new contexts and motivations to learn data science in ways that are rich and highly connected to situations of special relevance to their own lives.

Discussion

These changes introduce new issues, opportunities, and challenges on the mathematical and statistical ideas at the heart of work with data. They also introduce questions: Might too much attention to these new aspects of context be short changing epistemic and conceptual opportunities through simplification or limited engagement in full dimensions of data science work? If accepted, we will present detailed examples from the data science and education fields for each theme: the addition of publically available data to existing textbook exercises on regression and the case of the Flint, Michigan water crisis in the US for the first theme; the case of exploratory data environments in journalism for the second; and, the cases of machine learning, safety mapping, and international examples of learners' uses of data to highlight the need for change in their own communities. We will use these examples as a context to discuss challenges and opportunities for educators.

Connections to SRTL-10 Theme

While this report is necessarily broad, we do see meaningful connections to the conference theme. These connections especially lie in the notions of *innovation*, and *connecting data to context*. The report and synthesis will fundamentally examine how *innovations* in how data are collected, stored, analyzed, and shared are changing what is needed in K-12 education. And, it will explore how the ways in which data are connected to context - in increasingly personal, permanent, and persistent ways - are changing what it means to think and learn with data.

References

Conway, D. (2010). The data science venn diagram. *Dataists [Web page].* Retrieved February 9, 2012.

Dealing with Data [Special Issue]. (2011). *Science, 331*(6018).

Finzer, W. & Parvate, V. (2008). Who will teach them about data?: The responsibility of mathematics and statistics educators to support the integration of data analysis across all subjects. In Proceedings of the International Congress on Mathematical Education (ICME 11), Morelia, Mexico.

Friendly, M. (2008). A brief history of data visualization. In C. Chen, W. Härdle, & A. Unwin (eds.), *Handbook of data visualization* (pp. 15-56). Springer Berlin Heidelberg.

Hart, H. (2010). Decode exhibition points way to data-driven art. *Wired Magazine*. Retrieved March 23, 2015, from http://www.wired.com/2010/01/decode-exhibition-points-way-to-data-based-future-art/.

Hey, T. Tansley, S., & Tolle, K. M. (Eds.). (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft Research.

Johnston, B. & Webber, S. (2003). Information Literacy in Higher Education: A review and case study. *Studies in Higher Education, 28*(3), 335-352, doi: 10.1080/03075070309295

Johnston, B., & Webber, S. (2003). Information literacy in higher education: a review and case study. *Studies in Higher Education*, *28*(3), 335-352.

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., Finzer, W., Horton, N. J., & Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning*,*13*(1-2), 68-86.

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

Madison, B. L., & Steen, L. A. (Eds.). (2008). *Calculation vs. context: Quantitative literacy and its implications for teacher education.* Mathematical Association of America.

McGhee, G. (2009/2010): Journalism in the Age of Data. A video report on data visualization as a storytelling medium. http://datajournalism.stanford.edu/ (retrieved July 10, 2013).

National Science Foundation (2012, March 29). NSF leads federal efforts in big data. Press release 12-060. Accessed from https://www.nsf.gov/news/news_summ.jsp?cntn_id=123607.

Schulten, K. (2010, August 23). Teaching with infographics: Places to start. *The New York Times*, accessed from http://learning.blogs.nytimes.com/2010/08/23/teaching-with-infographics-places-to-start

The Learning Network (2014, August 26). Reader idea: Telling stories with data. *The New York Times,* accessed from http://learning.blogs.nytimes.com/2014/08/26/reader-idea-telling-stories-with-data/

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company: Reading, MA.

Wilkerson, M., Lee, V., Parikh, T., & Polman, J. (2015). CAP: Data Science, Learning and Youth: Connecting Research and Creating Frameworks. Unpublished grant proposal.