This is a preprint version of the introduction to Wilkerson, M. H. & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning [Special Issue]. *Journal of the Learning Sciences, 29*(1).

Situating Data Science: Exploring How Relationships to Data Shape Learning

Michelle Hoda Wilkerson[1] and Joseph L. Polman[2]

[1]University of California, Berkeley and [2]University of Colorado Boulder

Abstract

The emerging field of Data Science has had a large impact on science and society. This has led to over a decade of calls to establish a corresponding field of Data Science Education. There is still a need, however, to more deeply conceptualize what a field of Data Science Education might entail in terms of scope, responsibility, and execution. This special issue explores how one distinguishing feature of Data Science—its focus on data collected from social and environmental contexts within which learners often find themselves deeply embedded—suggests serious implications for learning and education. The learning sciences is uniquely positioned to investigate how such contextual embeddings impact learners' engagement with data including along conceptual, experiential, communal, racialized, spatial, and political dimensions. This special issue demonstrates the richly layered relationships learners build with data and reveals them to be not merely utilitarian mechanisms for learning about data, but a critical part of navigating data as social text and understanding Data Science as a discipline. Together, the contributions offer a vision of how the learning sciences can contribute to a more expansive, agentive and socially aware Data Science Education.

The emerging field of Data Science has had a large impact on science and society. This has led to over a decade of calls to establish a corresponding field of Data Science Education (Berman et al., 2016; Cleveland, 2001; Finzer, 2013). Data Science, the argument goes, prepares students for high-paying jobs, fuels scientific advancement, and provides communities with new tools for expression and empowerment. There is still a need, however, to more deeply conceptualize what a field of Data Science Education might entail—in terms of scope, responsibility, and execution. In particular, it is important to understand what makes learning Data Science sufficiently different from mathematics, computer science, or statistics that it requires new approaches to research and instructional design, and to explore the theoretical and practical implications of these differences for constructing an ethical and effective Data Science Education.

This special issue explores how one distinguishing feature of Data Science, the relational nature of data involved, suggests serious implications for learning and education. In contrast to data that are constructed to answer a particular question, Data Science is concerned with data collected in an incidental or automated manner from extensive social and environmental contexts (Donoho, 2015)—contexts within which learners often find themselves deeply embedded. The learning sciences is uniquely positioned to investigate how such contextual embeddings impact learners' engagement with data including along conceptual, experiential, social, racialized, spatial, and political dimensions (Philip, Olivares-Pasillas, & Rocha, 2016; Rubel, Lim, Hall-Wieckert, & Sullivan, 2016). The authors in this issue explore how learners' situatedness relative to data, to the contemporary and emerging field of data science as well as other disciplinary domains, and to the social histories interwoven with data necessitate new lines of research, new theoretical and methodological development, and new approaches to educational design and practice.

Current Conversations in Data Science and Data Science Education

Colloquially, the term data science refers to the use of computational tools and methods to collect, process, analyze, store, and visualize large quantities of data. It is broadly associated with the emergence of a growing number of data visualizations, open data repositories, and infographics intended for public consumption (McGhee, 2010), as well as shifts in how professional disciplines, from the sciences to the arts, make use of data and computing (Hey, Tansley, & Tolle, 2009). Data Science as a formal field of study, however, is still not well defined; the diversity of perspectives regarding its identity has led some to avoid defining it altogether (Cassel & Topi, 2015). Instead of attempting to create our own definition, we highlight some distinguishing characteristics that emerge as points of agreement in discussions of Data Science and Data Science Education.

One of the most commonly cited characteristics of Data Science is that it is concerned with a new class of data that are not only big in the traditional sense of scale, but "…pervasive, tacit, and often collected without a specific [or explicit] intent" (Cassel & Topi, 2015, p. 10). Social network and clickstream data are recorded from popular websites such as Facebook. Large-scale civic data about legislation, policy opinions, and municipal demographics are captured within historical census and voting data; weather, climate, and air or water quality benchmark data are collected multiple times per day from satellites, tide gauges, and other automated devices. Such datasets are encompassing in scope, and capture details about ourselves, our behavior, and our place in a shared world.

Such data can then be repurposed, or exploited, by students and practitioners alike for a variety of purposes (Donoho, 2015; Wilkerson & Laina, 2018). Individual preferences can be inferred from website interaction data to target advertisements, and predictive climate models can be constructed from environmental data to aid in policy development and city planning. This process requires not only computational and statistical knowledge, but also deep knowledge of both the target domain and the original context in which the data were collected. As a result, Data Science is intensely interdisciplinary. New tools and statistical techniques are constantly in development that incorporate and innovate on methods from multiple disciplines to visualize, store, organize, process, and interpret data.

These shifts in how, and why, data are constructed and used have led to the development of new programs, primarily in higher education, that focus on helping students learn to "think with data" (Baumer, 2015; Hardin et al., 2015) and learn computational methods for working with and communicating about data (Nolan & Temple Lang, 2010). Others focus on data science competencies as part of the civic and information literacies needed to navigate a data-rich world (e.g., Bergstrom & West's, 2017, "calling bullshit" course and initiative; Grawe, 2011). Recommendations across these efforts (e.g., De Veaux, et al. 2017) highlight two core pedagogical commitments. The first is that educators must not only attend to technical skill (e.g., programming languages such as R or Python, statistical methods and machine learning) but also flexibility to learn and develop novel tools and methods for working with data. The second is that Data Science must be grounded in consequential investigations in which learners pose questions, obtain data, and communicate findings within meaningful disciplinary contexts. Some go so far as to suggest that Data Science courses should always be offered concurrent to, and even embedded within, content courses in relevant disciplines[1]. While these pedagogical commitments are important contributions, they often do not address the challenges and possibilities that repurposable, exploitable, and contextually embedded data present for learners.

The Need for Learning Sciences Perspectives on Data Science Education

At the same time, emerging literature from the learning sciences highlights a number of ways in which skills- and technology-driven approaches to Data Science Education often fall short. One early such challenge was put forth by Philip, Schuler-Brown, and Way (2013), who surfaced tensions between the purported goals of Data Science Education efforts to support equity, power, and democratic participation on one hand and the status of Data Science as a discourse of power developed to advance economic and national interests on the other. These authors suggested that developing student proficiency with tools and techniques for working with data should be only one strand of a much broader Data Science Education project. Another goal must be to develop students' identities as agentive data practitioners who recognize the historical and political dimensions of data as social texts, and of Data Science as a disciplinary discourse.

Indeed, the role of data as historical and political has become especially apparent in recent efforts to introduce Data Science Education at pre-collegiate levels (e.g., Gould, Machado, Ong, Johnson, & Molyneux, 2016). A major reason for this is that whereas undergraduate programs situate data science

[1] See UC Berkeley's "Data 8 Connector Program" for one example; http://data8.org/connector/

within students' *disciplinarily* relevant areas of study (e.g., ecology, genetics, demography), precollegiate efforts have sought to do the same through the use of *socially* relevant datasets (e.g., popular movies or music, students' local communities). In one case, students resisted units that leveraged spatial data about the lottery and alternative banking institutions to teach statistics—a curricular approach which the researchers reflectively noted pathologized communities of color (Rubel, Hall-Wieckert, & Lim, 2016). In a study to explore how new features of the Scratch programming language could introduce young learners to Data Science, children warned that allowing code that takes others' user statistics as input could create exclusionary programs that only "popular" programmers could access (Hautea, Dasgupta, & Hill, 2017). In other cases, high school students' reasoning about racial and economic dimensions of spatial datasets were dismissed (Philip et al., 2016); or not well motivated given students' familiarity with the neighborhoods of study (Enyedy & Mukhopadhyay, 2007). All of these instances demonstrate how learners find themselves embedded within, and impacted by, the environments, histories, and social narratives woven into the datasets they investigate.

That context impacts learning is no surprise to learning scientists, whose work often leverages well-developed theoretical and methodological tools for understanding how experiences, tools and materials, environments, communities, and social positioning impact learning (Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991). It is for these reasons we dedicate this issue to situated perspectives toward learning Data Science. Some progress has already been made, for example, in exploring the role of embodiment and mobility in learners' reasoning with and about data, leading to the development of successful interventions that engage learners deeply with data analysis cycles about self and movement relative to history and health in formal and informal settings (Kahn, 2017; Lee, 2013; Taylor & Hall, 2013). We use the term *situated* in its broadest sense, to refer to a collection of approaches integrating learning, context, cognition, and participation (Roth & Jornet, 2013). We seek a path forward that enables data science educators at precollegiate and undergraduate levels alike to respond to a dramatically expanded, personalized, and politicized reality of data by enacting a correspondingly expansive, consequential, and responsible Data Science Education.

Contributions of the Special Issue

The questions that frame the contributions of this collection are: *In what ways are reasoning with data influenced by learners' situatedness—relative to data, data contexts, and data science as a field?* And, given that, *how can educators build on learners' existing relationships with data to develop a more ethical and effective Data Science Education?* In reading across this issue's contributions, we find insights to these questions with respect to variety of such relationships with data: including as learners *locate self* within data and *observe self* through data "traces" and "catchments"; *project self* and family into aggregated social and historical data; *manipulate and generate* data in interaction with the techno-material world; and exercise *ownership and critique* of data, its implications, and its limitations as a discourse.

This special issue includes a collection of original empirical reports that detail learners' work with data about their communities, families, and experiences across museum, after-school, school, and home contexts. Data Science is often described in emerging policy and consensus documents as a highly novel set of tools and practices, or as an eclectic combination of fields with which learners are likely to have had little experience with outside of formal, collegiate instruction. Across the articles in this special issue, however, we find a number of extant opportunities for learners, especially youth, to develop extensive and rich experience with data.

Lee and Dubovi (2020/this issue) for example, demonstrate how deep and complex engagements with data can emerge from necessary everyday routines. They describe how families dealing with type 1 diabetes exercise agency in making sense of and acting on data that is generated by monitoring devices and repeated measurements. Their use of the concept of "data catchment," the temporary storage and eventual transfer of data across people and spaces, is a unique theoretical contribution that extends the language of data flow and deluge. The notion of catchment emphasizes how data is "caught" and moved through the distributed cognition of families coping with chronic disease. Considering how data flows across catchments and is used by various parties to make sense and take consequential action is an idea with wide-ranging implications.

Exploring families' work with health data highlights one data relationship that is emerging as a rich site for learning and theorization, the use of "digital traces" (recorded on medical devices, transit passes, and documented through clicks and keystrokes; Latour, 2007), whether they be deliberately or incidentally captured in everyday life. Other research featured in this issue extends the notion of "traces" beyond the explicit capture of personal information, to the *projection* of self into existing datasets as another meaningful relationship with data. For example, Roberts and Lyons (2020/this issue) examine the sorts of first-person "actor perspective-taking" that learners take on when they imagine themselves represented visuo-spatially inside datasets, within the context of an interactive map museum exhibit based on census data. Designers of other immersive experiences are likely to find their notions of *orientation*, *role-play*, and *projection* as ways to relate self to data particularly useful.

Kahn (2020/this issue) extends the notion of perspective taking to explore co-construction of "family geobiographies" that span datasets, spaces, and generations. Similar to Roberts and Lyons' personalization strategies to encourage learners to locate themselves in data, Kahn engaged youth in using "big data" including census, income, and housing datasets to investigate their own family's intergenerational migrations. Youth's engagements with these datasets were layered with deeper sociohistorical context as other family members evaluated, re-interpreted, and contested their migration stories. Indeed, all three of Lee & Dubovi, Roberts & Lyons, and Kahn's offer new theoretical and design insights into how leveraging multiple learners' different relationships with the same data can support the co-construction of robust perspectival understandings (Greeno & Van De Sande, 2007).

Of course, perspective does not only affect the interpretation of data but also its construction. Several other papers in this collection explore learners' role in the design and construction of data across formal and informal contexts. Stornaiuolo (2020/this issue) explores how engaging high school learners in the intentional construction and analysis of data about their everyday activities helped them develop agentive orientations toward data. A particularly intriguing finding that emerged from students' construction and analysis of highly personal data was the negotiation of which data and patterns they chose to make public, versus what they kept private. This exploration reveals how relational intersections, for instance between self as data producer and self as object of study, provide rich opportunities for learning that extend across major issues in the field of Data Science, including data privacy and visual storytelling.

Hardy, Dixon, and Hsi (2020/this issue) also engage learners in data construction, in their case building on insights from science studies which have demonstrated how data is “produced” rather than merely found or discovered. They demonstrate how high school students engaged in data production with contemporary technologies like sensors gain a deeper understanding not only of the data and patterns related to a given scientific phenomenon, but also of the material complexity of scientific data more broadly. Harris, Dixon, Bird, and Ballard (2020) also connect to this theme by exploring how learners involved in environmental community and citizen science initiatives benefit from being positioned as active producers of data. Their paper also highlights the importance of the social systems in which data collection is embedded for providing opportunities for not just learning about data, but also developing agentive orientations toward data.

The powered nature of the social systems in which data engagements unfold play a large role in Van Wart, Lanouette & Parikh’s (2020/this issue) retrospective account of two community-based data science activities. These authors challenge the normative “scripts” that often accompany data science education efforts, such as that data can lead to actionable recommendations, or that data lend more credibility and validity to claims than other forms of evidence or experience. They demonstrate how these scripts silence learners’ perspectives—both by privileging data over experience, and by ignoring that decisionmakers and others in power may still easily dismiss findings. This work provides important lessons for getting beyond naive expectations that young people, particularly from historically marginalized communities, will or should simply adopt educators’ or data scientists’ norms and goals.

The special issue concludes with two commentaries, one largely focused on contextualizing the articles by looking back, and the other by looking at present day trends and toward the future. In the first commentary, Andee Rubin (2020/this issue) grounds current learning sciences research on a half-century of scholarship on reasoning with data. She reviews five critical aspects of working with data that have emerged from that literature: how context shapes the meaning of data for different constituencies; issues of variability and distribution within datasets; how cases are aggregated and summarized; how data is visualized; and what sorts of inferences, generalizations, conclusions and actions people make on the basis of data analysis. By positioning the current issue's papers relative to these themes, Rubin shows how the current work surfaces new issues (e.g., how the personal, temporal, and spatial nature of much "big data" complicates notions of variability and aggregation) among these well-established themes.

The special issue is concluded by a second commentary from Alyssa Wise (2020/this issue), who explores how each paper maps to data-science practice trajectory including the generation, storage, transformation, representation, and interpretation of data. Wise's analysis pushes us to consider how these experiences can be brought together across time and nature of data engagement so that they form an integrated approach that reflects the needs and nature of data science practice. She also encourages us as learning scientists to consider the design and learning of full *data science systems*—trajectories of tools, discourse practices, and materials needed to move, transform, and make appropriate use of data.

Together, this collection of papers demonstrates the richly layered relationships learners build with data. They also offer a vision of how the learning sciences can contribute to a more expansive, agentive and socially aware Data Science Education. This is in no small part because these relational layers reveal themselves to be not merely utilitarian mechanisms for learning about data, but a critical part of navigating data as social text and understanding Data Science as a discipline. We hope the issue serves as a learner-experience centered collection of studies that, in its contributions to theory and design, pushes toward a more ethical and effective Data Science Education.

Acknowledgements

Many of the ideas discussed in this introduction emerged from collective discussion at the 2016 Youth, Learning, and Data Science Summit. This small workshop was funded by National Science Foundation Cyberlearning capacity building grant IIS-1541676. We are grateful for comments on this introduction and related conversations provided by the contributors, anonymous reviewers, editors, and our generous colleagues.

References

Ainley, J., Gould, R., & Pratt, D. (2015). Learning to reason from samples: commentary from the perspectives of task design and the emergence of "big data." *Educational Studies in Mathematics*, *88*, 405–412. https://doi.org/10.1007/s10649-015-9592-4

Baumer, B. S. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician*, *69*, 334–342. https://doi.org/10.1080/00031305.2015.1081105

Bell, P., Van Horne, K., & Cheng, B. H. (2017). Special Issue: Designing Learning Environments for Equitable Disciplinary Identification. *Journal of the Learning Sciences*, *26*, 367–375. https://doi.org/10.1080/10508406.2017.1336021

Bergstrom, C. T., & West, J. (2017). Calling bullshit in the age of big data. University of Washington. Retrieved from http://callingbullshit.org/syllabus.html

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. (2016). *Realizing the potential of data science: Final report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group*.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. https://doi.org/10.2307/1176008

Cassel, B., & Topi, H. (2015). *Strengthening data science education through collaboration: Workshop report 7-27-2016*. Arlington, VA.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26. https://doi.org/10.1002/sam.11239

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruviluamala, N., Uhlig, P., Washington, T. M., Wesley, C. L., White, D., & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*(1), 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Donoho, D. (2015). 50 Years of Data Science. *R Software*, 41.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2), 1–9. Retrieved from http://escholarship.org/uc/item/7gv0q9dc

Gould, R., Machado, S., Ong, C., Johnson, T., & Molyneux, J. (2016). Teaching Data Science To Secondary Students: the Mobilize Introduction To Data Science Curriculum. *Iase-Web.Org*, (July). Retrieved from http://iase-web.org/documents/papers/rt2016/Gould.pdf

Grawe, N. D. (2011). The potential for teaching quantitative reasoning across the curriculum: Empirical evidence. *International Journal for the Scholarship of Teaching and Learning*, *5*(1), 1–12.

Greeno, J. G., & Van De Sande, C. (2007). Perspectival understanding of conceptions and conceptual growth in interaction. *Educational Psychologist*, *42*, 9–23. https://doi.org/10.1080/00461520709336915

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In *Transfer of learning from a modern multidisciplinary perspective* (pp. 89–120). https://doi.org/10.1002/car.1158

Hand, V., Penuel, W. R., & Gutiérrez, K. D. (2013). (Re)Framing educational possibility: Attending to power and equity in shaping access to and within learning opportunities. *Human Development*, *55*(5–6), 250–268. https://doi.org/10.1159/000345313

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B. S., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., & Ward, M. D. (2015). Data science in statistics curricula : Preparing students to “think with data.” *The American Statistician*, *69*, 343–353. https://doi.org/10.1080/00031305.2015.1077729

Hardy, L. Dixon, C. & Hsi, S. (2020/this issue). From data collectors to data producers: Shifting students' relationship to data. *Journal of the Learning Sciences, 29*.

Harris, E., Dixon, C. G. H., Bird, E. B., and Ballard, H. L. (2020, forthcoming). For science and self: Youth interactions with data in community and citizen science. *Journal of the Learning Sciences*.

Hautea, S., Dasgupta, S., & Hill, B. M. (2017). Youth Perspectives on Critical Data Literacies, 919–930. https://doi.org/10.1145/3025453.3025823

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm. *Data-Intensive Scientific Discovery. Microsoft Research*, 287. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Kahn, J. B. (2017). *At the intersection of self and society: Learning, storytelling, and modeling with big data*. Vanderbilt University.

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

Latour, B. (2007). Beware, your imagination leaves digital traces. *Times Higher Literary Supplement*, *6*(4), 2007. Retrieved from http://bruno-latour.fr/sites/default/files/P-129-THES-GB.pdf

Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. *Learning in Doing*, *95*, 138. https://doi.org/10.2307/2804509

Lee, V. R. (2013). The Quantified Self (QS) movement and some emerging opportunities for the educational technology field. *Educational Technology*, *53*(6), 39–42. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1015&context=victor_lee

Lee V. R., & Dubovi, I. (2020/this issue). At home with data: Family engagements with data involved in type 1 diabetes management. *Journal of the Learning Sciences, 29*. doi: 10.1080/10508406.2019.1666011

McGhee, G. (2010). *Journalism in the age of data*. Retrieved from datajournalism.stanford.edu

Nolan, D., & Temple Lang, D. (2010). Computing in the Statistics Curricula. *The American Statistician*, *64*, 97–107. https://doi.org/10.1198/tast.2010.09132

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 microcontestations. *Cognition and Instruction*, *34*, 361–388. https://doi.org/10.1080/07370008.2016.1210418

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

Polman, J. L., & Miller, D. (2010). Changing stories: trajectories of identification among african american youth in a science outreach apprenticeship. *American Educational Research Journal*, *47*, 879–918. https://doi.org/10.3102/0002831210367513

Roberts, J. & Lyons, L. (2020/this issue). Examining spontaneous perspective taking and fluid self-to-data relationships in informal open-ended data exploration, *Journal of the Learning Sciences, 29*. doi: 10.1080/10508406.2019.1651317

Roth, W.-M., & Jornet, A. (2013). Situated cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 463–478. https://doi.org/10.1002/wcs.1242

Rubel, L. H., Hall-Wieckert, M., & Lim, V. Y. (2016). Teaching mathematics for spatial justice: Beyond a victory narrative. *Harvard Educational Review*, *86*, 556–579.

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*, 1–26. https://doi.org/10.1080/07370008.2015.1118691

Rubin, A. (2020/this issue). Learning to reason with data: How did we get here and what do we know? *Journal of the Learning Sciences, 29*.

Stornaiuolo, A. (2020/this issue). Authoring data stories: Adolescents developing critical data literacies. *Journal of the Learning Sciences, 29*.

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.

Van Wart, S., Lanouette, K., & Parikh, T. (2020/this issue). Third spaces for data science education using participatory digital mapping. *Journal of the Learning Sciences, 29*.

Wilkerson, M. H., & Laina. (2018). Middle school students' reasoning about data and context through storytelling with repurposed local data. *ZDM Mathematics Education, 50*(7), 1223-1235. https://doi.org/10.1007/S11251-017-9430-4.

Wise, A. (2020/this issue). Educating data scientists and data-literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*.