Human-Centered Study of Data Science Work Practices

Michael Muller
IBM Research AI
Cambridge, MA, USA
michael_muller@us.ibm.com

Timothy George
Project Jupyter
San Luis Obispo, CA, USA
timsfilez@gmail.com

Bonnie E. John
Bloomberg LP
New York, NY, USA
bjohn11@bloomberg.net

Samir Passi
Cornell University
Ithaca, NY, USA
sp966@cornell.edu

Melanie Feinberg
University of North Carolina
Chapel Hill, NC, USA
mfeinber@unc.edu

Steven J. Jackson
Cornell University
Ithaca, NY, USA
sjj54@cornell.edu

Mary Beth Kery
Carnegie Mellon University
Pittsburgh, PA, USA
mkery@cs.cmu.edu

ACM Reference Format:
Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY. 8 pages. DOI: https://doi.org/10.1145/3290607.3299018

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3299018

ABSTRACT

With the rise of big data, there has been an increasing need to understand who is working in data science and how they are doing their work. HCI and CSCW researchers have begun to examine these questions. In this workshop, we invite researchers to share their observations, experiences, hypotheses, and insights, in the hopes of developing a taxonomy of work practices and open issues in the behavioral and social study of data science and data science workers.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Data science; Work practice.

INTRODUCTION

With the rise and frequent opacity of big data, there has been a growing need to understand who is working in data science and how they are doing their work. Researchers in HCI, CSCW, and critical data studies have begun to examine these questions. In this workshop, we invite researchers and practitioners to share their observations, experiences, hypotheses, and insights. Depending on participant interest, we hope to answer questions such as the following:

- How do data science workers approach their tasks? What are their strategies? What can we learn about their views of data and process [4, 9, 25, 28]?
- What (and who) is missed in our prevailing accounts and assumptions around data and data work? If "data is never raw" [10], who does the cooking, and how is this work performed, recognized, and accounted for?
- Data science tools are generally designed for one user at-a-time. Other complex tasks have benefited greatly from collaboration practices and collaboration technologies. Are there opportunities to bring some of these lessons to data science? What are the distinctive collaboration needs and constraints in data science [3, 26]?
- What methods are valid and tractable to assess the usefulness and usability of tools for data scientists in service of iterative design? How can learnability be assessed when domain experts, not programmers, need to learn new languages and tools simultaneously, the learning of which takes far longer than typical usability studies? How can efficiency for skilled users be assessed when the uses of the tools are so diverse that benchmark tasks have little face validity?

> Rules as constraints vs guidelines [25]
> Data as donnè vs. capta [4]
> Data-driven vs. measurement-driven [21][28][32]
> Features as emergent vs. features as constructed [27]
> Exploration vs. explanation [18][31]
> Solitary vs. collaborative data science activities [6][26]

Sidebar 1: Tensions in human-centered study of data science

- Who are the consumers of data sciene work and what are their needs? How can data science processes and tools address these needs, e.g., the level of transparency and comprehension consumers desire?

BACKGROUND

Extraordinary claims are made about the promises and current successes of data science [1, 8, 14, 15, 19, 22, 29, 33] While some of these claims are stated for the future [7], Agarwal and Dhar editorialized in 2014 that "This is powerful... we are in principle already there"[2]. Meanwhile to the dystopian extreme, scholars warn about the "mythology" of working with big data, that the quantitative nature of data gives a false illusion that all data-driven outcomes are objective, ethical, or true [5]. Further complicating these discussions, there is considerable diversity in the tools and methods [12], challenges [13], and job-roles [15, 19, 23] involved in data science. Detailed study will be necessarily partial and contextualized, adding depth of description and understanding, but potentially lacking a broader view.

HCI and Data Science

Several studies in HCI, CSCW, FAT*[1] , and critical data studies have begun to look at facets of these diverse topics.

Dealing with data, or data wrangling, has been estimated to take up 80-90% of the effort in a typical data science project [11, 16, 30]. Understanding how people approach their data is therefore important. Bilis contrasted two views of the analyst's relationship with data [4]. In one view, the analyst takes a relatively passive stance, and receives data as "given" by the environment ("donnÃl'"). In a second view, the analyst takes a more active role as s/he captures data ("capta"). Pine and Leboiron made a similar point, claiming that in some cases "human-computer interactions start before the data reaches the computer because various measurement interfaces are the invisible premise of data and databases" (emphasis in their original text) [28]. Feinberg describes the "design" of data [9], and Patel et al. similarly describe the creation of features for analysis [27]. Muller et al. documented the sometimes necessary processes of the creation of data, including the creation of grounded truth data [24].

Passi and Jackson described an ongoing tension over the use of algorithms as rules [25]. They propose that data science students learn to practice a kind of data vision that emphasizes the discretionary craftspersonship needed to appropriately handle data analysis. This notion of "crafts[person]ship" appears again in studies of how novices develop machine learning models: novices fail when they cannot intuitively relate how the code they write interplays with the data itself [27, 35]. This problem becomes especially important from an HCI perspective, in which domain knowledge is often crucial to understand how to analyze a topic and its data (e.g., [34]). Recording the outcomes and managing

[1] Fairness, Accountability, and Transparency

the diverse experimental analytic histories (i.e., provenance) of data and code are also challenging. Despite the promise of literate programming [20], people who engage in data science tend to scant their documentation, apparently because of a tension between dynamic, engaged exploration and time-consuming explanation [17, 18, 31].

Despite the paucity of colleague-oriented documentation in many data science projects, there is increasing evidence that data science workers nonetheless collaborate. In an ethnographic study, Passi and Jackson reported on diverse actors, with diverse motivations, working in and with corporate data science teams [26]. They highlighted issues of trust among heterogeneous teams of designers, managers, business analysts, and data scientists, which are often resolved in part through work that is simultaneously "calculative" (e.g., reliance on quantified metrics, statistical tests) and "collaborative" (e.g., negotiation and translation work). Chang et al. proposed collaborative commenting in JupyterLab notebooks [6]. Recent tools such as Co-Calc[2] or Colaboratory[3] provides real-time chat and multi-user storage in a notebook programming environment.

Tensions in Work with Data in Data Science

In summary, the field of data science presents us with multiple tensions which might be addressed through HCI research. Sidebar 1 provides a partial list of these tensions. Some of these challenges may be resolvable. Other challenges may take the form of enduring analytic dimensions that inform our research plans and outcomes.

WORKSHOP SCHEDULE

HCI engagement with data science is relatively new for most people. Therefore, we propose to spend a half-day to share knowledge and perspectives in this domain, through brief presentations of participants' position papers (which we will make available in advance, at the workshop website). The focus will be on the intersection of data science and HCI. Co-organizers will organize the presentations into "mini-sessions" based on common themes.

We will conclude the morning by organizing the afternoon session into small groups, formed through recognition of shared interests during the morning discussions. (Based on the accepted position papers, we will also have a default set of proposed topics, but we hope that participants will generate their own insights and choices). The afternoon will conclude with small-group reports, a plan for future meetings, plans for a workshop report, and dinner.

CALL FOR PARTICIPATION

We hope to gather the growing community of researchers and practitioners in HCI and allied fields who are building new insights, methods, and collaborative practices around data science. We invite

[2] https://cocalc.com/

[3] https://colab.research.google.com

reports of experiences, ethnographies, tools, methods, designs, and theories, as well as critiques of theories, practices, social implications, and design fictions. We seek work that helps us to:

- Contextualize and understand data science work practices - by individuals, and by groups and teams
- Characterize the work practices of data science workers, including programming, ideation, and collaboration
- Show how practices of data creation and aggregation inform the work of data science
- Understand the shared and unique design challenges of data science environments, including methods and tools for comprehending data, data wrangling, model building, debugging, collaborating and communicating results, especially to non-programmers
- Support the incorporation of diverse ethics and human values into data science work, e.g., in relation to algorithmic fairness or bias reduction
- Bring sociotechnical and organizational perspectives on data work to bear on data science education and practice
- Suggest methods of standardizing or coordinating data collection across organizational and industry boundaries
- Bridge the gap between the knowledge of data scientists and that of domain experts in various fields of application
- Widen the audience for data science beyond highly technically skilled programmers, to include UX designers, project managers, novice programmers, and other stakeholders into a data-driven project
- Help policymakers to build more effective, appropriate, and transparent rules around the complex domains of data science work

For details, please see the workshop website, https://husdatworkshop.github.io/.

Workshop website online 11 December 2018
Call published and distributed 12 December 2018
Submission deadline 12 February 2019
Notifications to authors 1 March 2019
Workshop date 4 May 2019

Sidebar 2: Important dates

POST-WORKSHOP PLANS

During the concluding session of the workshop, we will discuss content and formats for a workshop report that will be shared at the workshop's website (and as a poster at the conference, if possible). Author a workshop report outlining the key takeaways of the workshop and their potential for impact in industry, education, and the data science research community. We hope that outcomes of the workshop (both report and cross-cutting conversations among participants) will gather and extend the human-centered study of data work practices in HCI and allied communities (CSCW, ECSCW, and FAT), critical data studies, etc.), leading to further collaborations in research and additional empirical, theoretical, and practical refinements what will contribute to the ongoing evolution of data science research, education, and industrial practice.

ORGANIZERS

Michael Muller works as a researcher at IBM Research AI, where he studies data science work, and collaborates with data science workers to design future tools for data science.

Melanie Feinberg is an associate professor at the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill. She studies the practices by which data is made, and the characteristics of data as both design artifact and design material.

Timothy George works as a UI/UX designer for Project Jupyter, where he designs next generation data science tools. He also works to develop open standards, protocols and practices for data scientists.

Steven Jackson is an Associate Professor and Chair of Information Science at Cornell University. His work addresses questions of ethics, policy and practice in emerging computing fields.

Bonnie E. John is a Senior Interaction Designer at Bloomberg where she uses user-centered methods to design and evaluate tools for financial data scientists and collaborates with Project Jupyter.

Mary Beth Kery is a PhD student at the Human-Computer Interaction Institute at Carnegie Mellon University. Her research focuses on studying program-mer behavior and designing new kinds of programming tools to support exploratory data science work.

Samir Passi is a PhD candidate in the Department of Information Science at Cornell University. His research focuses on the forms of human work in data science learning, research, and practice. He studies such forms of work ethnographically in the context of academic as well as corporate data science.

REFERENCES

[1] 2018. Doing data science: A Kaggle walkthrough: Cleaning data. KDNuggets. Retrieved Febuary 2, 2015 from https://www.kdnuggets.com/2016/03/doing-data-science-kagglea-walkthrough-cleaning-data.html.

[2] Ritu Agarwal and Vasant Dhar. 2014. Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research* 25, 3 (2014), 443–448.

[3] Andrew Begel, Jan Bosch, and Margaret-Anne Storey. 2013. Social networking meets software development: Perspectives from github, msdn, stack exchange, and topcoder. *IEEE Software* 1 (2013), 52–66.

[4] Hélène Bilis. 2018. Mapping fiction: Social networks and the novel. In *Shifting (the) Boundaries Conference*. Wellesley College.

[5] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[6] Rose Chang, Meredith Granger, Alena Mueller, and Taka Shimokobe. 2018. Designing comments. In *JupyterCon 2018*. O'Reilly.

[7] Ciprian Dobre and Fatos Xhafa. 2014. Intelligent services for big data science. *Future Generation Computer Systems* 37 (2014), 267–281.

[8] Hugh Durrant-Whyte. 2015. Data, Knowledge and Discovery: Machine Learning meets Natural Science. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 7–7.

[9] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2952–2963.

[10] Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT Press.

[11] Michele Goetz. 2015. 3 ways data preparation tools help you get ahead of big data. Blog. Retrieved Sep 2, 2018 from https://go.forrester.com/blogs/15-02-17-3_ways_data_preparation_tools_help_you_get_ahead_of_big_data/.

[12] Bob Hayes. 2018. Most used data science tools and technologies in 2017 and what to expect for 2018. BusinessOverBroadway. Retrieved Sep 2, 2018 from http://businessoverbroadway.com/most-used-data-science-tools-and-technologies-in-2017-and-what-to-expect-for-2018.

[13] Bob Hayes. 2018. Top 10 challenges to practicing data science at work. BusinessOverBroadway. Retrieved Sep 2, 2018 from http://businessoverbroadway.com/top-10-challenges-to-practicing-data-science-at-work.

[14] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.

[15] Kaggle. 2017. A big picture view of the state of data science and machine learning. Kaggle ML and data science survey. Retrieved Febuary 2, 2015 from https://www.kaggle.com/kaggle/kaggle-survey-2017.

[16] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3363–3372.

[17] Mary Beth Kery, Amber Horvath, and Brad Myers. 2017. Variolite: supporting exploratory programming by data scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1265–1276.

[18] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 174.

[19] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 96–107.

[20] Donald Ervin Knuth. 1984. Literate programming. *Comput. J.* 27, 2 (1984), 97–111.

[21] Helena M Mentis, Ahmed Rahim, and Pierre Theodore. 2016. Crafting the Image in Surgical Telemedicine. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 744–755.

[22] Renée J Miller. 2017. The Future of Data Integration. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 3–3.

[23] Steven Miller. 2014. Collaborative approaches needed to close the big data skills gap. (2014).

[24] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkorsky, Jason Tsay, Q. Vera Laio, Casey Dugan, and Tom Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design and creation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA.

[25] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, 2436–2447.

[26] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 136 (Nov. 2018), 28 pages. https://doi.org/10.1145/3274405

[27] Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. 2008. Investigating Statistical Machine Learning As a Tool for Software Development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 667–676. https://doi.org/10.1145/1357054.1357160

[28] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3147–3156. https:

//doi.org/10.1145/2702123.2702298

[29] Krishna Rajan. 2013. *Informatics for materials science and engineering: Data-driven discovery for materials science and engineering.* Elsevier.

[30] Tye Rattenbury, Joseph M. Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. 2017. *Principles of data wrangling: Practical techniques for data preparation.* O'Reilly.

[31] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and Explanation in Computational Notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 32, 12 pages. https://doi.org/10.1145/3173574.3173606

[32] Alex S. Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-Place: Thinking Through the Relations Between Data and Community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 2863–2872. https://doi.org/10.1145/2702123.2702558

[33] Wil MP Van der Aalst. 2014. Data scientist: The engineer of the future. In *In Enterprise interoperability VI.* Springer, 13–26.

[34] Leland Wilkinson. 2005. *The grammar of graphics.* Springer.

[35] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18).* ACM, New York, NY, USA, 573–584. https://doi.org/10.1145/3196709.3196729