Accounting for the Spatial and Temporal Dimensions of Data Science Processes

Samir Passi
Department of Information Science
Cornell University
Ithaca, NY, 14853, USA
sp966@cornell.edu

ABSTRACT

My research involves the analysis of the ongoing forms of *in-situ* work of data science practitioners to unpack the human, organizational, and ethical dimensions of the process of building data science systems. In my work, I approach data science as a social, situated, and collaborative practice. In this paper, I briefly describe my experiences with studying data science *processes* in corporate settings, highlighting the spatial and temporal challenge of the work of mapping processes.

INTRODUCTION

This short paper has two parts. In the first section *Research Description*, I briefly describe my own research to situate how, when, and where the notion of 'process' features in my work. In the second section *Spatial and Temporal Dimensions of Data Science in Applied Settings*, I describe ways in which I attempt to pragmatically, though partially, manage questions of space and time in my own work to map the relationship between multiple data science processes. I conclude by raising further methodological questions concerning the workshop's central theme—*the study of processes.*

ACM ISBN XXX-X-XXXX-XXXX-X/X/XX.
DOI: https://doi.org/XXXXXX

Box #1
Research Works

1. six-month long ethnographic research with a corporate data science team of a multi-billion-dollar technology company based on the US West Coast (I worked *as* a data scientist, while simultaneously conducting research),
2. three-month long qualitative research on Microsoft product teams (~30 interviews) to study how they scope and resolve the ethical dimensions of the AI-based systems that they develop,
3. four-month long ethnographic participant-observation study of a data science course on machine learning at a US university,
4. four-month long ethnographic participant-observation study of a data science course on natural language processing at a US university, and
5. participant-observation study of a series of three data science workshops on digital humanities at a US university

People and organizations names are anonymized to preserve the privacy anonymity and privacy of research participants as per Cornell Institutional Review Board (IRB) approved research protocols: 1406004779, 1410005026, 15080005788, and 1705007175. The research at Microsoft fell under Microsoft's AI and Ethics in Engineering and Research (AETHER) initiative—the results are currently not public.

RESEARCH DESCRIPTION

My doctoral dissertation examines and unpacks the human [1], organizational [2], and ethical [3] dimensions of data science practices. Towards this end, I have conducted a set of ethnographic and qualitive research studies to analyze the ongoing forms of *in-situ* work of data science practitioners in both corporate organizations and academic institutions (Box 1). The notion of 'process' has been central to my work particularly given that the practice of data science constitutes and is constitutive of a diverse set of practices. These range from the processes of problem formulation and data curation to the processes of analyzing algorithmic results and managing corporate data science projects. In this paper, I focus on my methodological experiences with researching data science processes in one specific corporate organization (#1 in Table 1).

SPATIAL AND TEMPORAL DIMENSIONS OF DATA SCIENCE IN APPLIED SETTINGS

Where is a process?

It might be natural—in fact, intuitive—to assume that the data science practice begins and ends with the work of data scientists. After all, they are the ones who interact with data, algorithms, and models—the necessary and crucial ingredients of data science. Seen this way, the answer to the question 'where is the data science process?' appears to lie in the activities, desks, movements, and lives of data scientists. During my first fieldwork in a technology company, however, I realized that such an understanding was, at best, incomplete (or, worse, incorrect). Applied data science work in corporate settings is exceptionally heterogeneous. A much wider set of actors, goals, and practices transect ongoing and everyday forms of corporate data science work. Project managers, product designers, business analysts, and corporate executives are as much a part of applied data science work as data scientists and software engineers. Faced with the complex and interactive nature of corporate data science work, the answer to the question 'where is the data science process' seems to be—*everywhere*. To effectively map data science processes, the researcher needs to find ways to account for the seeming omnipresence of such processes [4].

The specific way in which I managed this problem was by adhering to the research maxim—follow the actors [5]. Situating the ongoing practices of algorithmic work as the result of the everyday interactions between multiple groups of actors, not just data scientists, the research heuristic of 'follow the actors' provided me with an actionable way to identify, trace, and describe applied data science processes. My approach to actors was two-fold. First, I examined the everyday professional lives of scientists, managers, engineers, and analysts to understand not only their process of working on a specific project, but also how different professional groups perceived and participated in corporate data science projects in specific ways. Second, I examined people's articulations of and work on data, slides, algorithms, numbers, and models as these moved between desks, projects, rooms, servers, and groups. My own understanding of what is and is not a part of the ecology of data science actors impacted my research work to situate, map, and examine data science process. At this

workshop, I wish to make visible and further discuss aspects concerning the implications of the sociomaterial [6] nature of professional practices for the effective mapping and study of the processes that constitute and are constitutive of professional practices. *How do we, as researchers, engage with the sociomaterial ecology of the processes we study, and how our contingent engagement changes, or should change, our research goals, work, and insights?*

When is a process?

Every process has a unique temporal rhythm—a life of its own. The holistic study of processes requires a researcher to account for their beginnings and ends—the work of carving out the object of study. My own research was no different. To study the corporate practice of data science, I had to demarcate and categorize the different processes making up this practice. The important thing to note here is that I wasn't researching *one* process. I was examining a practice—i.e., a collection of processes bound by projects. The challenge was that during the six-months of fieldwork, the data science team worked on several corporate projects. Data science projects have varied lifespans—some last longer than others. Some projects were already underway when I started fieldwork, others began during my time. Some projects ended in different ways during my fieldwork (e.g., either the data science team stopped working on models were successfully deployed), others remained midway when I left. (I went back to the field-site twice after finishing my fieldwork to conduct further interviews to fill in the blanks in fieldnotes.) Sometimes I observed the ~~same~~ similar process (e.g., model selection) across different projects—the nature of processes remained recognizably similar but practically different across projects. At other times, I observed only one instance of a process (e.g., bespoke data collection). The nature of the projects that the team worked on was beyond my control. Even post-fieldwork, categorizing processes within the collected ethnographic and interview data for analysis and writing was not a straightforward task.

The way in which I partially managed this problem—both during and after fieldwork—was by shifting my analytic point-of-view. Attempting to map processes from start to finish, at least in my case, seemed problematic and futile. I focused instead on how actors articulated and perceived processes in different contexts and situations. The answer to the question 'when is a process?' became less about the temporality of a process and more about the (a)rhythmic ways in which actors brought up and described different processes at specific points in time during data science projects. I focused less on *when* processes are (in a project) and more on *when* processes do (things in a project). For example, concerning the process of data collection, I followed how this process was not only described and understood differently in data science and business meetings, but also used to both rationalize and problematize different kinds of algorithmic results. While far from perfect, taking this pragmatic perspective on the temporal doings, not natures, of processes helped me to move forward in my study of data science processes.

At this workshop, I want to contribute to and learn from other participants how they manage and account for the temporalities of the processes they study. *How do we, as researchers, account for and manage the varied, often problematic, temporalities of processes?*

AUTHOR BIO

Samir Passi is a PhD candidate in the Department of Information Science at Cornell University where he works with Steve Jackson, Phoebe Sengers, Solon Barocas, and David Mimno. He is also a member of the *Artificial Intelligence, Policy, and Practice (AIPP)* initiative and of the *Culturally Embedded Computing (CEMCOM)* research group at Cornell University. His current work lies at the intersection of critical data studies, computer-supported cooperative work (CSCW), fairness, accountability, and transparency in machine learning (FAT/ML), and science and technology studies (STS).

ACKNOWLEDGMENTS

My research has been funded by the National Science Foundation (Grant #: CHS-1526155), the Intel Science & Technology Center for Social Computing (ISTC), and Cornell University.

REFERENCES

[1] S. Passi and S. J. Jackson, "Data Vision: Learning to See Through Algorithmic Abstraction," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, 2017, pp. 2436–2447.

[2] S. Passi and S. J. Jackson, "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects," *Proc. ACM Human-Computer Interact.*, vol. 2, no. CSCW, pp. 1–28, Nov. 2018.

[3] S. Passi and S. Barocas, "Problem Formulation and Fairness," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 2019, pp. 39–48.

[4] L. A. Suchman, *Human-Machine Reconfigurations: Plans and Situated Actions*, 2nd ed. New York: Cambridge University Press, 2007.

[5] B. Latour, *Reassembling the Social: An Introduction to Actor-Network Theory*. New York: Oxford University Press, 2005.

[6] W. J. Orlikowski, "Sociomaterial Practices: Exploring Technology at Work," *Organ. Stud.*, vol. 28, no. 9, pp. 1435–1448, Sep. 2007.