Collaboration *as* Participation: The Many Faces in a Corporate Data Science Project

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

Abstract

Critical data studies research has made visible the 'design-use gap'—users (as people most affected by data science systems) often do not have a say in the system's design. Much discussion thus focuses on the role and place of user participation in data science practices. In this piece, however, I focus on already-existing forms of participatory work in corporate data science practices. Corporate projects are highly participatory in nature (though not in the way we often define and expect participation). These projects necessitate diverse forms of work on the part of multiple personnel such as data scientists, project managers, business analysts, product managers, and business executives. Unpacking the collaborative work in corporate data science projects as forms of participation provides us with a different perspective on the design-use gap, helping us focus on *different* forms of participation in corporate data science practices.

Author Keywords

Collaboration; Participation; Data Science.

ACM Classification Keywords

H.m. [Information Systems]: Miscellaneous

Introduction

This paper comprises two sections. In the first section, I describe a set of three brief ethnographic vignettes to make visible the collaboration-as-participation work in corporate data science projects. In the second section, I explore how unpacking such forms of collaboration-as-participation work help us to reconsider the boundaries between users, designers, and stakeholders in corporate data science practices and projects.

DeepNetwork Inc.

Founded in the nineties, DeepNetwork is a multi-billion-dollar US-based ecommerce and new media organization.

The organization owns 100+ companies in domains such as health, legal, and automotive. Many of these are multi-million-dollar companies with several thousand clients each.

The organization has a central data science team based on the US West Coast that works with multiple businesses across different domains and states.

I worked as a data scientist at DeepNetwork between June-November 2017 while simultaneously conducting ethnographic fieldwork with their core data science team.

I. The Many Faces in Corporate Data Science

Vignette 1: The First Meeting

Afternoon. June 20, 2017. Eleven DeepNetwork[1] company personnel (including me) convened in a conference room on the 9th floor of an office building somewhere in California. These were (from my left): Daniel *(data science project manager),* Jeff *(director of the content team)*, Cliff *(data engineer),* Cathy *(data engineer),* Martin *(director of data science)*, Brad *(project manager),* Alex, Kent, Max, and David (data scientists). The agenda of the meeting was to discuss the development of an in-house data science tool to provide "better" content to users on several websites owned by DeepNetwork—a form of recommender system to, for example, provide better related articles to the content that users were interacting with. The hope was to also explore ways to personalize content delivery for each user (e.g., by considering content that users had previously interacted with).

At first, there seemed to be no users present in the room. People who visited the websites, however, were not the *only* users of the proposed data science tool. The tool was to be *used* by the content team as a feature on several DeepNetwork subsidiaries with the explicit aim of increasing business revenue by maximizing aspects such as clickthrough rates and user engagement. The content team (as users) were not only visibly present in the room, but also involved in discussions on what the tool should do and why. What do we mean by 'related' content? How do we ascertain which related content is 'better'? What is the baseline that we wish to 'improve' on? What kind of content do we feel is 'relevant' to specific categories of content? The meeting revolved around many such questions—answers to most of which were negotiated between the members of the two teams.

Vignette 2: Data Scientists on their Desks

Evening. July 10, 2017. David and I had been assigned to work on the recommender project. Our desks were next to each other. I had my headphones on and was working with the data from the Self-Project company's website—a DeepNetwork subsidiary that provided people with information on all kinds of do-it-yourself projects. Self-Project had agreed to be the first one to try out the recommender tool for A/B testing. The more time I spent with the data, the more questions I had. What were these 'category tags' assigned to web posts? Were these automatically generated or manually assigned? Could we use them as inputs to our recommender models? If we wanted more 'clicks,' should we also prioritize popularity of articles in addition to their relatedness? Should our system have a clickbait component? If websites belonged to different domains (e.g., health or home) should we build a one-size-fits-all system or model each for the specific type of content? Should we provide related content *in the same way* for an article on blood pressure as we did for an article on gardening? The lack of answers to these questions inhibited my ability to work on the project.

I turned to David to inquire about these questions. We discussed for 15 minutes, coming up with strategies to deal with the questions. We both went to Daniel – our project manager – to get his feedback. The three of us spent the next half an hour in a small conference room. Daniel argued that our decisions needed to align with

[1] All company and personnel names have been replaced with pseudonyms to preserve research participant anonymity.

project priorities and requirements. He took down a list of our suggestions and emailed it to the members of the data science and content teams. The next day we heard back from the content team. They agreed with some of our suggestions but disagreed with others. A week later – July 17, 2017 – we met again with the content team to further discuss our options. Answers to some questions were given by us, answers to some by them, but most answers were negotiated outcomes.

Vignette 3: Project Status Update Meeting

David and I created a set of four models that provided content related to articles on Self-Project's websites. Director of data science Martin, director of content Jeff, and Self-Project's director of technology Wu had decided that the evaluation of the models would be done by Katie. Katie was a senior content manager working at Self-Project and was considered a domain expert on the kind of content provided on Self-Project's website. David and I chose a small set of random Self-Project articles to use as test-cases while Katie hand-picked a few test cases that she believed were difficult or challenging. There were in total 25 articles. We sent the results for the 25 articles from all four of our models to Katie.

Afternoon. July 31, 2017. Katie finished her evaluation. In addition to data science and content team members, and Katie, other members of Self-Project's business team were also present in the meeting: Mihae *(project manager),* Nathan *(project manager),* and Wu *(director of technology).* Katie declared model #4 as the 'winner' and said that it 'was the best 21 out of 25 times.' Model #3 was a not-so-close second, model #2 wasn't 'that good,' and model #1 was, according to Katie, 'horrible.' The project managers from each team were convinced that the data scientists should continue working on models #3 and #4 but stop working on models #1 and #2. It was also decided in the meeting that the models should not only be autonomous (i.e., as and when new articles get added to the website, the models should self-update the list of related articles), but also be customizable (i.e., engineers should be able to tweak model parameters for different websites).

II. Collaboration-as-Participation

My aim with these three brief ethnographic vignettes is to make two points. **First,** in academic and research contexts, data science often appears as the exclusive domain of data scientists. Corporate data science projects, however, are inherently heterogeneous and collaborative in nature comprising a diversity of professional roles and aspirations.[2] Different experts collaborate and intervene at different times in such projects, directing the project and shaping the data science system design in specific and significant ways. Focusing on and unpacking such forms of collaboration work in corporate data science practices enables us to nuance our understanding of data science system "designers," "users," and "stakeholders."

In common formulations of the design-use gap, the term 'designers' is used to refer to data scientists, 'users' for people who end up using the system (or ones most affected by it), and 'stakeholders' for people

[2] See our CSCW 2018 paper for more details on the collaborative nature of corporate data science practices. Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 2, CSCW, Article 136 *(November 2018).* ACM, New York, NY. 28 pages. https://doi.org/10.1145/3274405

who aren’t data scientists but have social, economic, or political stakes in the data science system. Through these three brief vignettes, we begin to see how the boundaries between these three roles are porous at best. The content team was *also* a user of the data science tool though in a different way compared to people who visit the websites. The business team was yet another kind of user—one who wanted to maximize the revenue generated from their content. The business team wasn’t just a by-standing stakeholder but an active participant in system evaluation (Katie) and design (e.g., they wanted the tool to have certain functionalities to match existing business practices).

The data science team members were definitely not the only designers. This was true not only at the beginning of the project (i.e., what is usually referred to as the ‘business understanding’ phase) but even in the everyday work of data scientists and that of system development and evaluation.

Second, focusing on different corporate personnel’s role in and impact on the design and development of data science systems, we being to see how the work of different personnel (often referred to as ‘collaboration work’) is, in fact, a kind of participation work—experts taking on different roles (e.g., designers, users, and stakeholders) at different points in time to intervene in specific ways in the data science system’s design, development, and evaluation. The content and business teams – as particular kinds of users – shaped the working and functionality of the data science tool.

In this workshop session, I wish to take seriously such forms of collaboration-as-participation work and discuss with workshop participations a concrete set of questions that arise when we unpack the heterogeneous nature of corporate data science practices and projects:

Q.1 If we – as researchers – start considering business teams also as users of data science tools, how does it impact our understanding of “design,” “participation,” and “use”?

Q.2 What forms of participation do we desire in the collaboration between data scientists (as designers) and business teams (as users)?

Q.3 Are there specific parts of the data science design process in which we may actually *not want* the participation of business-teams-as-users? [E.g., business teams may prioritize revenue maximization at the expense of computational ideals or social values]

Q.4 Given this multifaceted view of collaboration-as-participation work in corporate data science practices and projects, what are some sites for design/research interventions? How can we – as researchers – leverage existing forms of such collaboration-as-participation work to inform and shape data science systems?

I hope that the workshop will provide me with the opportunity to not only discuss such questions, but also inform my own understanding and help me to further nuance these rough first steps towards the concept of collaboration-as-participation.

Acknowledgements

The funding for this research was provided by National Science Foundation grant CHS-1526155.