Data Scientists as Craft Workers: Theorizing Data Work

Research in Progress

Konstantin Hopf, University of Bamberg, konstantin.hopf@uni-bamberg.de
Mayur P. Joshi, The University of Manchester, mayur.joshi@manchester.ac.uk
Arisa Shollo, Copenhagen Business School, ash.digi@cbs.dk
Marta Stelmaszak, Portland State University, stmarta@pdx.edu

Background and Motivation

The diffusion of data science, analytics, and other data-intensive technologies in organizations (Agarwal & Dhar, 2014; Berente et al., 2021) led to a proliferation of data workers who develop and deploy them. Almost paradoxically to the expectations of the automation potential of these technologies, the increasing datafication of work exacerbates the need for data work in organizations (Jones, 2019; von Krogh, 2018). As a result, information systems (IS) researchers have started investigating data work, which includes activities around collecting, processing, analyzing, and generating insight from data in organizations. Data workers, thus, are an emerging breed of professionals with expertise in data (e.g., Pachidi et al., 2021; Parmiggiani et al., 2022). They draw on a range of technical skills around data science and machine learning (e.g., Aversa et al., 2018; Vaast & Pinsonneault, 2021), through more typical business intelligence and analytics (Koch et al., 2021; Shollo & Galliers, 2016), to more fundamental recording of various real-world phenomena in data (Cunha & Carugati, 2018; Waardenburg et al., 2022).

The datafication of many professions increases the proliferation of data work in various business areas. For example, account managers register sales differently to help later analyses (Pachidi et al., 2021), HR professionals change performance measurement standards (van den Broek et al., 2021), and police officers need to espouse new roles of algorithmic translators (Waardenburg et al., 2022). Data workers not only need technical expertise in statistics and computer science, but also need business and creative skills. As data science tools often grapple with a lack of ground truth (Lebovitz et al., 2021), especially while uncovering social phenomena, data workers need to embrace uncertainty using rules of thumb (Hill et al., 2016) and making complex choices (Patel et al., 2008) by drawing on their skills and attitude towards inventiveness, creativity, and experimentation (Avnoon, 2021; Parmiggiani et al., 2022). By contrast, data work is also portrayed as a source of efficacy, objectivity, and neutrality by mitigating human bias (Agarwal & Dhar, 2014; Davenport, 2018; Jones, 2019). Being considered as highly rational employees engaging in nearly mechanistic work they risk to be automated (Davenport, 2018).

Amid the apparent opposition of views on data work as a subjective art versus an objective science, in this study we set out to explore the possibility of examining it as a confluence of the two. We do so by drawing on the theory of craft that we discuss next.

Theoretical Framework

Craft as a theoretical framework allows bringing humanistic and mechanistic approaches to work together. The contemporary accounts of craft (as opposed to traditional craft), allow

for such a fusion, especially in the case of technical craft configurations (Kroezen et al., 2021), of which software engineering and work of technicians are good examples (Adler, 2015; Barley, 1996). Technical craft entails a distinct approach to work that is based on specific skills and attitudes that differ it from other, more mechanical approaches. Craft skills encompass the mastery of technique, whereby individuals have exceptional competency over their work, all-roundedness that allows them full control over the entire process, and embodied expertise that is tacit and contextual (Barley, 1996; Kroezen et al., 2021). Craft attitudes emphasize dedication, a full commitment to work and engagement in it for its own sake (Sennett, 2009), communality, that is attention to a shared occupational identity and purpose (Anteby, 2008), and exploration which requires tinkering and engaging with the complexity and ambiguity of tasks (Kroezen et al., 2021; Sennett, 2009).

Research Design and Methods

To better understand how data workers draw on their technical as well as artistic skills together in their everyday work, we conducted a qualitative study of data scientists as an extreme example of data workers. We interviewed 62 data scientists in 23 globally distributed organizations ranging from high-tech to traditional industries. Across the interviews, we found strong evidence of typical craft skills and attitude (Kroezen et al., 2021) among data scientists (e.g. mastery, abstract expertise, dedication, exploration).

Data collection: The interviews focused on what, how, and why data scientists do what they do. They followed a semi-structured guide, flexibly adjusted over time, driven by the accounts of the participants (Gioia et al., 2013). All the interviews were recorded, lasted on average 55.3 minutes, and were transcribed verbatim in their original language (English in 38, German in 14, and English combined with Hindi in 10 cases). This resulted in 57:08 hours of recordings and 1,005 pages of transcribed text.

Data analysis: We engaged in iterative data analysis and coding, drawing on grounded theory methods (Glaser & Strauss, 1967) with the help of software. We generated craft-related first order codes by labeling reported practices of data scientists. We compared and contrasted the labels and arrived at sub-practices in an axial coding step. In a selective coding step, we consolidated these to higher order themes (craft practices of data science work). Finally, we identified cross-cutting themes that demonstrated the iterative nature of practices and allowed us to theorize and conceptualize a model of data-based craft.

Preliminary Findings

We uncovered that data workers not only engage in crafting under-defined and incomplete data science products for their business customers (such as dynamic prediction dashboards or reports), but also actively craft their tools by developing algorithmic models, as well as the material by generating and giving shape to data. In particular, we identified generating material for a specific purpose, making the material processable, and giving meaning to the material as the practices of (i) crafting the material (data); (re)searching for the right tool, tuning the tool, and trying out the tool as the practices of (ii) crafting the analytical tools; and envisioning, pitching, and cultivating the product as the practices of (iii) crafting data science products (see Table 1).

Table 1. Summary of findings on the practices of craft in data (science) work

Activities of data worker	Higher order themes: Practices in data science craft	Core categories: Sub-practices in data science craft	Reasons why data scientists need to engage in craft
Crafting the material: data science inputs	**Generating material for a specific purpose**, relies on drawing from embodied expertise within the organization, and mastery of data within the context	1) Identifying data sources 2) Creating data from other digital objects 3) Remoulding data from other data products	- Absence of specific, indicated datasets - Required data do not exist within the organization - Underlying data is pre-crafted for a different purpose
	Making the material processable, draws on the understanding of constantly changing processes through embodied expertise, and the mastery of data science	1) Initial pre-processing of data 2) Constant processing of data	- Dirty data, that is format and contents have to be cleaned up - Data drift that turns data less relevant and more obsolete
	Giving meaning to the material, relies on embodied experience acquired through immersed domain knowledge and engagement, and drawing on communality	1) Explicating semantics 2) Working in tandem to understand data 3) Getting hands dirty for embodied expertise	- Data as material have a semantic value that is not obvious from its characteristics - The organization is a constantly evolving and changing entity, thus meanings shift
Crafting the tools: data science models	**(Re)searching for the right tool**, requires mastery, communality and dedication to the role through continuing learning	1) Distributing tool mastery 2) Engaging in double communality	- Expanding number of potential models to be deployed forces specialization - Working with specific tools as a matter of occupational identity
	Tuning the tool, relies on constant experimentation and the mastery of data science skills	1) Experimenting with models 2) Tuning the tuners	- Define model parameters, data formats for every project - Models produce different results every run
	Trying out the tool, draws on experimentation in real-world conditions, as well as communality with other departments	1) Evaluating performance 2) Conducting real-world experiments	- Plethora of performance metrics available - Evaluation on real-world data is needed - Validating causality in detected correlations is required
Crafting the products: data science outputs	**Envisioning the product**, relies on all-roundedness and communality with other organizational members	1) Transforming business to data-driven problems 2) Exploring business area data	- Organizations are struggling with exploiting available data - Data are ambiguous - Clients are not clear on what they want
	Pitching the product, emphasizes the dedication of data scientists to their craft as communicated to customers	1) Storytelling 2) Timing the products 3) Educating customers	- Digital and intangible nature of data science products makes them elusive to customers
	Cultivating the product, steeped into dedication, ensuring that products remain in good functioning	1) Competing against self 2) Refreshing the products	- Models decaying over time - Data drift that turns data products less relevant and more obsolete

Based on these findings, we conceptualize a model of data work (see Figure 1) that explains the work of data scientists as well as other increasingly data-intensive occupations, such as economists, engineers, scientists, and data workers in general (Dougherty & Dunne, 2012). The model highlights that data work is a creative combination of technical and human approaches to work, where data workers not only craft the products (analytical models), but also the tools (algorithms), and the material (data). Data workers not only need technical expertise to understand the nature of data, advanced algorithms, and statistical models, they also need domain knowledge to understand the complex social phenomena and creative skills to generate novel and relevant insights through the models they build. As such, data work is underpinned by the malleable nature of data as its material, the autonomy of the tools, and the permanent incompleteness of the crafted products, shifting the data workers' focus from only crafting final products to crafting all elements.

Potential Contributions

While our findings are grounded in the empirical context of data science, the model of data work offers a theoretical understanding of the emerging type of work that uses data as material (i.e., data work). For example, economists who now work with constantly changing and flowing big data find their existing stable, fixed tools and metrics becoming increasingly obsolete (Kansas, 2021). Similar developments happen across other long-standing occupations such as medical professionals (Lebovitz et al., 2021), police officers (Waardenburg et al., 2022), or chip designers (Zhang et al., 2021, p. 1192). As such, we are planning to develop our contribution toward a better understanding of the changing nature of work when it involves data, thus building on and expanding current IS studies of data work. We based our theory on the case of data scientists, but further empirical studies are needed to extend the arguments to other occupations whose work becomes increasingly underpinned by data but are not necessarily embedded in organizational contexts.

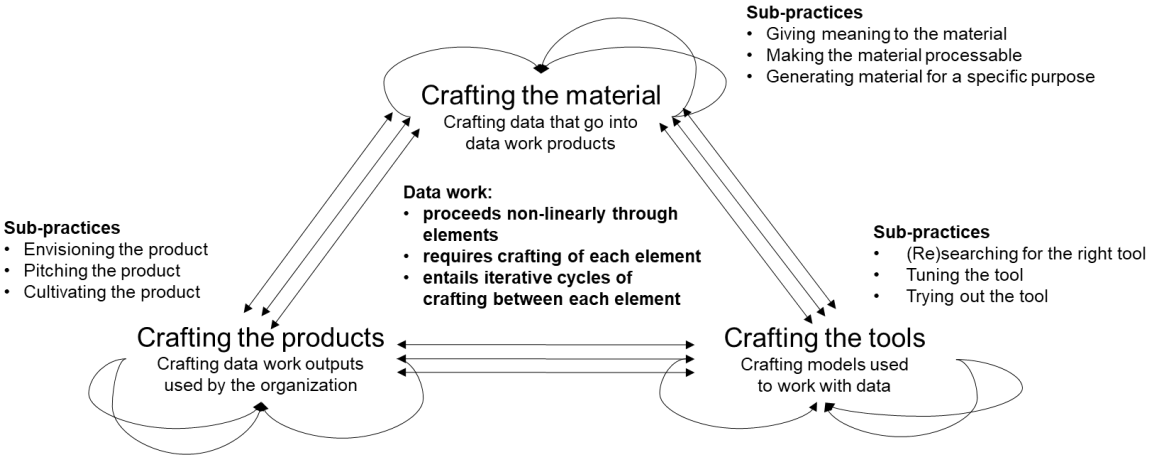

Figure 1. A model of data work based on craft

References

Adler, P. S. (2015). Community and Innovation: From Tönnies to Marx. *Organization Studies*, *36*(4), 445–471. https://doi.org/10.1177/0170840614561566

Agarwal, R., & Dhar, V. (2014). Editorial - Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546

Anteby, M. (2008). Identity Incentives as an Engaging Form of Control: Revisiting Leniencies in an Aeronautic Plant. *Organization Science*, *19*(2), 202–220. https://doi.org/10.1287/orsc.1070.0343

Aversa, P., Cabantous, L., & Haefliger, S. (2018). When decision support systems fail: Insights for strategic information systems from Formula 1. *The Journal of Strategic Information Systems*, *27*(3), 221–236. https://doi.org/10.1016/j.jsis.2018.03.002

Avnoon, N. (2021). Data Scientists' Identity Work: Omnivorous Symbolic Boundaries in Skills Acquisition. *Work, Employment and Society*, *35*(2), 332–349. https://doi.org/10.1177/0950017020977306

Barley, S. R. (1996). Technicians in the Workplace: Ethnographic Evidence for Bringing Work into Organizational Studies. *Administrative Science Quarterly*, *41*(3), 404–441. https://doi.org/10.2307/2393937

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

Cunha, J., & Carugati, A. (2018). Transfiguration Work and the System of Transfiguration: How

Employees Represent and Misrepresent Their Work. *MIS Quarterly*, *42*(3), 873–894. https://doi.org/10.25300/MISQ/2018/13050

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73–80. https://doi.org/10.1080/2573234X.2018.1543535

Dougherty, D., & Dunne, D. D. (2012). Digital Science and Knowledge Boundaries in Complex Innovation. *Organization Science*, *23*(5), 1467–1484. https://doi.org/10.1287/orsc.1110.0700

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, *16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research* (4. paperback printing). Aldine Publishing.

Hill, C., Bellamy, R., Erickson, T., & Burnett, M. (2016). Trials and tribulations of developers of intelligent systems: A field study. *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 162–170. https://doi.org/10.1109/VLHCC.2016.7739680

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3–16. https://doi.org/10.1016/j.jsis.2018.10.005

Kansas, S. (2021, October 23). Enter third-wave economics. *The Economist*. https://www.economist.com/briefing/2021/10/23/enter-third-wave-economics

Koch, H., Chipidza, W., & Kayworth, T. R. (2021). Realizing value from shadow analytics: A case study. *The Journal of Strategic Information Systems*, *30*(2), 101668. https://doi.org/10.1016/j.jsis.2021.101668

Kroezen, J., Ravasi, D., Sasaki, I., Żebrowska, M., & Suddaby, R. (2021). Configurations of Craft: Alternative Models for Organizing Work. *Academy of Management Annals*. https://doi.org/10.5465/annals.2019.0145

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI Ground Truth Really 'True'? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, *45*(3), 1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, *32*(1), 18–41. https://doi.org/10.1287/orsc.2020.1377

Parmiggiani, E., Østerlie, T., & Almklov, P. G. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, *23*(1), 139–164. https://doi.org/10.17705/1jais.00718

Patel, K., Fogarty, J., Landay, J. A., & Harrison, B. (2008). Investigating statistical machine learning as a tool for software development. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, 667. https://doi.org/10.1145/1357054.1357160

Sennett, R. (2009). *The Craftsman*. Yale University Press.

Shollo, A., & Galliers, R. D. (2016). Towards an understanding of the role of business intelligence systems in organisational knowing. *Information Systems Journal*, *26*(4), 339–367. https://doi.org/10.1111/isj.12071

Vaast, E., & Pinsonneault, A. (2021). When Digital Technologies Enable and Threaten Occupational Identity: The Delicate Balancing Act of Data Scientists. *MIS Quarterly*, *45*(3), 1087–1112. https://doi.org/10.25300/MISQ/2021/16024

van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing Ai for Hiring. *MIS Quarterly*, *45*(3), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, *4*(4), 404–409. https://doi.org/10.5465/amd.2018.0084

Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2022). In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science*, *33*(1), 59–82. https://doi.org/10.1287/orsc.2021.1544

Zhang, Z., Lindberg, A., Lyytinen, K., & Yoo, Y. (2021). The unknowability of autonomous tools and the liminal experience of their use. *Information Systems Research*, *32*(4), 1192–1213. https://doi.org/10.1287/isre.2021.1022