Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

HOLLOW DATASETS: ALGORITHMIC CALCULABILITY IN DATA CURATION

Alejandro Alvarado Rojas
University of Southern California

Introduction

Digital platforms like Kaggle and Hugging Face are integral to the scientific exploration of data. They serve as infrastructure for data scientists to curate, process, analyze, and share datasets, develop models, and build professional networks. In facilitating these activities, these platforms contribute to the *platformization* of data science, transforming the social and material conditions of data science work (Poell et al., 2019). An important aspect of this transformation is data curation. Data curation systematically manages data resources to make data "fit-for-use, archive-ready, and accessible over the long-term" (Thomer et al., 2022, p. 2), often to improve data quality and inform decision-making (Leese, 2023). However, data curation remains largely invisible, undermining its epistemic value and reinforcing alienated forms of labor (Irani, 2015). While labor issues in data curation are concerning the data science enterprise, the transformation of this process within data science platforms remains understudied.

This research asks how data curation is changing with increasing reliance on data science platforms. It examines the sociotechnical organization of data curation as a component of data science work and its implications for assessing data quality. Drawing from the literature in Science and Technology Studies (STS) and Critical Data Studies, I conceptualize data science platforms as calculative infrastructures. By employing a technographic analysis of Kaggle—a leading data science platform where users curate datasets and host model competitions, this research traces the algorithmic curation of datasets.

Data Platforms As Calculative Infrastructures

A calculative infrastructure is a network of devices, practices, and narratives that enable forms of calculation (Kurunmäki & Miller, 2013). Calculation distinguishes "between things or states of the world, and by imagining and estimating courses of action associated with those things or with states as well as their consequences" (Callon & Muniesa, 2005, p. 1231). It informs decision-making through the use of both numerical

Suggested Citation (APA): Alvarado Rojas, A. (2025, October). *Hollow Datasets: Algorithmic Calculability in Data Curation.* Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

indicators and judgments about their significance for action (Cochoy, 2019). Calculative infrastructures render legible the objects and objectives of calculation by transforming qualities into quantities, supporting “quantified information to travel and be consumed in diverse social and organizational contexts” (Reilley & Scheytt, 2019, p. 46). Social media platforms exemplify this logic by measuring, ranking, and ordering social interactions based on relative social and economic worth (van Dijck, 2013; Mau, 2019).

As calculative infrastructures, data platforms for data science employ a similar logic to order interactions with datasets and models. While research has examined the effects of social interactivity metrics on dataset applications or their material-economic dynamics of communication practices (Twyman et al., 2023; Bounegru, 2023), the processes enabling data curation platform spaces remain underexplored. Data platforms increasingly incorporate social media-style logics that shape how datasets are created, represented, and shared. In other words, data curation consists of the chain of calculations that transform data quality into quantitative indicators to inform decision-making (Herzog et al., 2017). However, calculating data quality spans a network of practices, instruments, and interpretations requiring justification and legitimation (Ruppert & Scheel, 2021). Within data science platforms, these justifications are embedded in an algorithmic rationality that prioritizes “efficiency as an epistemological standard in the service of a new form of scientific inquiry” (Lowrie, 2017, p. 3).

Technographic Analysis of Data Curation and Kaggle’s Usability Rating

This research investigates data curation in Kaggle. Kaggle provides Data Cards that contain key metadata about the datasets, such as title, description, content tags, and column names, among others. In May 2019, Kaggle released the Usability Rating feature, which ranks dataset quality on a scale from 0 to 10. While the calculation of this score is based on aspects of the Data Cards, the algorithmic operations behind it are partially opaque. As such, the Usability Rating is well-suited for infrastructural analysis (Bowker, 1994).

Using a technographic approach (Bucher, 2016), I investigate how Kaggle’s Usability Rating functions as a calculative device, organizing data curation practices and foregrounding understandings about data quality. Specifically, I investigate its performative role in defining data quality as the object of calculation and its objectives in rationalizing algorithmic curation. My technographic analysis directly engages in data curation by curating a dataset, reviewing community discourse on the Usability Rating, and identifying platform features that coordinate data curation calculations.

Making Data Curation Calculable and Hollow Datasets

By making data curation calculable, Kaggle’s Usability Rating illustrates how algorithmically generated indicators of data quality overlook the epistemic context of dataset creation and maintenance. Data quality shifts from being problem-driven to process-driven, where calculated acts of documentation become the primary measure

of quality. This transformation conflates the “epistemic value of data” (Leese, 2023, p. 6) with algorithmic metadata documentation, reducing reflexivity in data curation to quantitative referents of effort rather than concerns with quality itself. This results in what I term *hollow datasets:* datasets with algorithmically generated quality metrics that overlook meaningful contextual or qualitative information.

The algorithmic logic enabled by platform design contributes to the generation of hollow datasets. Digital platforms often operate as empty shells by acting as “rigid, yet purposely under-determined mediating structure, which intensively responsibilize workers for the quality and content of their work” (Huber & Pierce, 2023, p. 1). Similarly, Kaggle’s Usability Rating is a calculative device that structures curatorial tasks by instructing users on how to populate metadata. However, the values associated with quality and dataset content remain subject to interpretation by both curators and users. These subjective assessments can be overlooked, despite constituting key aspects of data curation (Leese, 2023; Thomer et al., 2022). Here, algorithmic rationality reinforces the technocratic values of efficiency and control that are central to data science practice (Lowrie, 2017). While the precise workings of the Usability Rating are publicly unknown, its purpose is to streamline data curation and facilitate dataset searchability. While efficiency itself is not inherently problematic, justifying algorithmic data curation on this basis ignores the social relationships and affective labor involved in dataset creation (Poirier, 2021).

This research introduces the notion of hollow datasets to characterize algorithmic calculability in data curation as a platform-based phenomenon. In my technographic approach, I also recognize that this analysis is primarily an interpretative approach. Future research could center on the perspectives of developers and users of the Usability Rating to understand how such a metric informs other interpretations of data quality. Additionally, other calculations in algorithmic data curation could be compared across platforms or particular data quality standards. By illustrating one way in which algorithmic data curation produces hollow datasets, this research surfaces the need to critique and build on the conditions that justify data quality as a sociotechnical process.

References

Bounegru, L. (2023). The platformisation of software development: Connective coding and platform vernaculars on GitHub. *Convergence*, 13548565231205867. https://doi.org/10.1177/13548565231205867

Bowker, G. (1994). Information mythology and infrastructure. In L. Bud-Frierman (Ed.), *Information acumen: The understanding and use of knowledge in modern business* (pp. 231–247). Taylor & Francis.

Bucher, T. (2016). Neither Black Nor Box: Ways of Knowing Algorithms. In *Innovative Methods in Media and Communication Research* (pp. 81–98). https://doi.org/10.1007/978-3-319-40700-5_5

Callon, M., & Muniesa, F. (2005). Peripheral Vision: Economic Markets as Calculative Collective Devices. *Organization Studies*, *26*(8), 1229–1250. https://doi.org/10.1177/0170840605056393

Cochoy, F. (2019). The Cultivation of Market Behaviors and Economic Decisions: Calculation, Qualculation, and Calqulation Revisited. In F. F. Wherry & I. Woodward (Eds.), *The Oxford Handbook of Consumption* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190695583.013.13

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). What is Data Quality and Why Should We Care? In T. N. Herzog, F. J. Scheuren, & W. E. Winkler (Eds.), *Data Quality and Record Linkage Techniques* (pp. 7–15). Springer. https://doi.org/10.1007/0-387-69505-2_2

Huber, L., & Pierce, C. (2023). Navigating the empty shell: The role of articulation work in platform structures. *Journal of Computer-Mediated Communication*, *28*(4), zmad004. https://doi.org/10.1093/jcmc/zmad004

Irani, L. (2015). The cultural work of microwork. *New Media & Society*, *17*(5), 720–739. https://doi.org/10.1177/1461444813511926

Kurunmäki, L., & Miller, P. (2013). Calculating failure: The making of a calculative infrastructure for forgiving and forecasting failure. *Business History*. https://www.tandfonline.com/doi/abs/10.1080/00076791.2013.838036

Leese, M. (2023). Data curation: A conceptual framework for the study of data quality. *CURATE Working Paper*, *2*. https://doi.org/10.3929/ethz-b-000597540

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, *4*(1), 2053951717700925. https://doi.org/10.1177/2053951717700925

Mau, S. (2019). *The Metric Society: On the Quantification of the Social*. Wiley.
Poell, T., Nieborg, D., & Dijck, J. van. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://policyreview.info/concepts/platformisation

Poirier, L. (2021). Reading datasets: Strategies for interpreting the politics of data signification. *Big Data & Society*, *8*(2), 20539517211029322. https://doi.org/10.1177/20539517211029322

Reilley, J., & Scheytt, T. (2019). A Calculative Infrastructure in the Making: The Emergence of a Multi-Layered Complex for Governing Healthcare. In M. Kornberger, G. C. Bowker, J. Elyachar, A. Mennicken, P. Miller, J. Randa Nucho, & N. Pollock (Eds.), *Thinking Infrastructures* (Vol. 62, pp. 43–68). Emerald Publishing Limited. https://doi.org/10.1108/S0733-558X20190000062004

Ruppert, E., & Scheel, S. (2021). *Data Practices: Making Up a European People*. MIT Press.

Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW2), 414:1-414:29. https://doi.org/10.1145/3555139

Twyman, M., Murić, G., & Zheng, W. (2023). Positioning in a collaboration network and performance in competitions: A case study of Kaggle. *Journal of Computer-Mediated Communication*, *28*(4), zmad024. https://doi.org/10.1093/jcmc/zmad024

Van Dijck, J. van. (2013). *The Culture of Connectivity: A Critical History of Social Media*. OUP USA.