Comment on: Is It AI or Data That Drives Market Power?

Miao Ben Zhang[a],

[a] *USC Marshall School of Business, Los Angeles, CA 90089, USA*

1. Introduction

Mihet, Rishabh, and Gomes (2025) offer a substantive contribution to our understanding of the mechanisms driving market power in the age of artificial intelligence (AI) and big data. The paper expands the horizon in three ways, as I see it. First, the paper takes a deep dive into the difference between "raw" and "processed" data, highlighting the cost of possessing the raw data and the ability of firm-level AI capability to turn raw data into processed data. This distinction among raw data, AI capability, and processed data links the literature on information entropy and the value of data, as I discuss below. A second contribution is the modeling of a secondary market for trading processed data among firms, which is highly relevant for policy discussion on how to facilitate and regulate data sharing platforms, such as via API or data vendors. Finally, the paper makes an effort to empirically test the model's implications for how improvements in firms' ability to obtain raw versus processed data offer opposite predictions on industry concentration. Below, I comment on each contribution by benchmarking it against the current literature.

2. Raw Data, AI Capability, and Processed Data

A burgeoning literature recognizes data as a central input in modern firm production, innovation, and competition (Farboodi and Veldkamp, 2022; Veldkamp and Chung, 2024; Farboodi et al., 2019; Mihet and Philippon, 2019). A workhorse model for data in firm production

Email address: miao.zhang@marshall.usc.edu (Miao Ben Zhang)

treats data as contributing to the prediction of firms' next-period productivity, as pioneered by Farboodi and Veldkamp (2022).[1]

On the other hand, prior literature has also examined the costs of utilizing information. In particular, information-theoretic constraints, such as entropy, long studied in macroeconomics and finance (Shannon, 1948; Sims, 2003; Veldkamp, 2011), suggest that the marginal value of unprocessed raw data may even turn negative in the absence of adequate processing capabilities. Hence, firms' capability to process raw data, such as their skilled labor (Abis and Veldkamp (2024)), can play an important role in the costs and benefits of using data.

A central innovation of this paper is to explicitly model the distinct roles of raw data, AI capability, and processed data in a firm's production. Raw data is modeled as a by-product of firms' production and linearly related to firm size, following the literature (Jones and Tonetti, 2020; Veldkamp and Chung, 2024). Processed data are modeled as the abundance of signals, which improves the precision of firms' prediction of their next period productivity (Farboodi and Veldkamp (2022)). The key contribution is the paper's explicit model of the relation between processed data and raw data, which accounts for both the information entropy effect that reduces the precision of signals for firms with low AI capability, and the positive effect where firms' AI capability (z_i) turns raw data ($n_{i,t}$) into abundant useful signals, i.e., processed data ($\tilde{n}_{i,t}$):

$$\tilde{n}_{i,t} = \underbrace{e^{-z_i} \cdot (-n_{i,t} \ln n_{i,t})}_{\text{Information Entropy Effect}} + \underbrace{\left(1 - e^{-z_i}\right) \cdot n_{i,t} e^{n_{i,t}}}_{\text{AI Capability Effect}}. \tag{1}$$

This formula offers a nice insight. However, given the central importance of this distinction to the paper, it would be helpful if the authors could provide more support for the distinct impact of raw and processed data on firms.

Suggestions: I wonder if the authors can provide more motivating facts regarding the two effects in the above equation. In particular, in your empirical sample of publicly-traded firms,

[1] See Veldkamp and Chung (2024) for an excellent review of this literature.

which you have measures of each firm's data-intensity and AI-intensity, do you see low-AI firms have lower productivity, such as TFP (Imrohoroglu and Tüzel (2012)), than high-AI firms while controlling for similar levels of data-intensity. More broadly, some anecdotal facts specifically about the "negative" effects of information entropy with respect to having too much data but no processing capability would be particularly helpful for motivating this equation.

The authors should also discuss existing measures of firms' data processing ability, such as firms' possession of human capital related to data engineering, annotation, cleaning, or ML pipeline automation (Abis and Veldkamp (2024)).[2] For instance, how do changes in the competition among firms in the talent market (Chen et al. (2023)) affect a firm's ability to transform raw data into processed data and impact its future productivity and growth? In summary, more supporting evidence for equation (1) above would make the paper's core contribution more convincing.

3. The Market for Trading Raw and Processed Data

A particularly timely feature of the paper is its analysis of processed data as a tradable commodity. The results in Section 2.2, supported by simulation and empirical analysis, show that facilitating processed data markets (e.g., via APIs or structured knowledge sharing) can democratize innovation and reduce market concentration, enabling even low-AI firms to compete. This insight aligns with recent work by (Gans, 2018, 2024; Conti et al., 2023, 2024; Athey, 2019) on the emergence of secondary markets for structured data and foundation model outputs. These developments raise new questions about market design, property rights, and competition policy. Who controls access to processed data, and what regulatory or market mechanisms best promote entry and innovation? As noted by Carballa Smichowski et al. (2023), platforms may have incentives to restrict processed data access, which could create new bottlenecks or reinforce incumbent power. Policy responses will need to consider

[2]See Veldkamp (2023) for a review of various existing measures of a firm's data assets.

the trade-offs between open access and proprietary data rights.

Suggestion: The fundamental driver for this market is that processed data, unlike ideas or technologies, is less portable, as emphasized by Veldkamp and Chung (2024). Data cannot be easily obtained by poaching a skilled employee from one firm to another. Hence, firms can either obtain processed data through a platform in a centralized form or through the acquisition of the target firm. This paper has so far focused on the discussion of the platform market for firms to trade data, but is silent on the discussion of one firm acquiring another firm to access their data, which seems to be another crucial means by which processed data affects business concentration. I would encourage the authors to include a discussion on this alternative channel, which appears to be central to the paper's focus on market concentration.

4. Data and Market Concentration

Empirically, the authors construct innovative firm-level proxies for both AI and data intensity and exploit two exogenous technological shocks—the advent of AWS cloud computing and transformer-based architectures—to identify the causal effects of improvements in compute and processed data accessibility. Their evidence that compute improvements disproportionately benefit data-rich firms (Begenau et al. (2018)), while processed data access disproportionately benefits low-AI firms, is consistent with theoretical predictions. Thus, a key message from this analysis is that access to raw data tends to foster market concentration, whereas access to processed data tends to reduce market concentration.

This is a nice insight. However, I have some thoughts on the empirical measures and thus the inference of the results. The paper relies on firm-level Herfindahl-Hirschman Index (HHI) calculations based on time-varying, text-based industry definitions (Hoberg and Phillips, 2016). While this measure is perfect for identifying firms' product market competition, used in this particular context, I can see there are two competing ways to infer the results. In particular, a positive effect on this HHI can represent either genuine firm growth relative to its competitors or a reclassification of the firm into different (possibly more concentrated)

sectors. In other words, I find the HHI measure alone is not strong enough to support a claim regarding the effects on product market concentration.

Suggestion: I recommend that the authors complement their main analysis with industry-level concentration measures anchored to fixed (baseline) industry codes—for instance, using the FIC code of (Hoberg and Phillips, 2016) available at the Hoberg-Phillips Data Library. Specifically, the authors may construct concentration measures at the FIC-year level. Then, the authors can analyze how industries with high and low average (or median) AI-intensity and data-intensity respond to the two shocks in terms of their concentration measures. An example of such industry-level analysis is Gutiérrez and Philippon (2017) who examine industries' growth and their median Q. This would clarify whether the observed changes in concentration are due to within-industry dynamics or shifts in the firm's business focus, and would align their measurement with best practices in recent research on persistent market power (De Loecker et al., 2020). Such robustness checks would further strengthen the empirical foundation for their theoretical claims.

5. Conclusion

Overall, Mihet, Rishabh, and Gomes (2025) offer a substantial leap forward in our understanding of data, AI, and market power. Their explicit attention to the transformation from raw to processed data, the empirical operationalization of these concepts, and the analysis of processed data markets are all meaningful contributions. Future research can build on their work by developing more direct proxies for data-processing investments and by exploring the evolving structure and regulation of secondary processed data markets.

References

Abis, S. and Veldkamp, L. (2024). The Changing Economics of Knowledge Production. *The Review of Financial Studies*, 37(1):89–118.

Athey, S. (2019). 21. The Impact of Machine Learning on Economics. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, pages 507–552. University of Chicago Press.

Begenau, J., Farboodi, M., and Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, 97:71–87.

Carballa Smichowski, B., Lefouili, Y., Mantovani, A., and Reggiani, C. (2023). Data sharing or algorithm sharing?

Chen, A. J., Zhang, M. B., and Zhang, Z. (2023). Talent market competition and firm growth. *Available at SSRN 4597388*.

Conti, A., Gupta, V., Guzman, J., and Roche, M. P. (2023). Incentivizing Innovation in Open Source: Evidence from the GitHub Sponsors Program.

Conti, A., Peukert, C., and Roche, M. (2024). Beefing IT Up for Your Investor? Engagement with Open Source Communities, Innovation, and Startup Funding: Evidence from GitHub. *Organization Science*.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics*, 135(2):561–644.

Farboodi, M., Mihet, R., Philippon, T., and Veldkamp, L. (2019). Big Data and Firm Dynamics. *AEA Papers and Proceedings*, 109:38–42.

Farboodi, M. and Veldkamp, L. (2022). A model of the data economy. Technical report, National Bureau of Economic Research Cambridge, MA, USA.

Gans, J. (2018). Enhancing competition with data and identity portability. *The Hamilton Project*, pages 1–28.

Gans, J. S. (2024). Market Power in Artificial Intelligence.

Gutiérrez, G. and Philippon, T. (2017). Investmentless Growth: An Empirical Investigation. *Brookings Papers on Economic Activity*, pages 89–169.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Imrohoroglu, A. and Tüzel, S. (2012). Firm Level Productivity, Risk, and Return.

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9):2819–2858.

Mihet, R. and Philippon, T. (2019). The Economics of Big Data and Artificial Intelligence. In *Disruptive Innovation in Business and Finance in the Digital World*, volume 20, pages 29–43. Emerald Publishing Limited.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Veldkamp, L. (2011). *Information Choice in Macroeconomics and Finance*. Princeton University Press.

Veldkamp, L. (2023). Valuing data as an asset. *Review of Finance*, 27(5):1545–1562.

Veldkamp, L. and Chung, C. (2024). Data and the Aggregate Economy. *Journal of Economic Literature*, 62(2):458–484.