Is It AI or Data That Drives Market Power?*

Roxana Mihet† Kumar Rishabh‡ Orlando Gomes§
University of Lausanne University of Lausanne ISCAL

The technology revolution is transforming firm and industry dynamics, yet the roots of firm growth in the modern economy remain unclear. Is industry dynamism driven by compute capabilities (AI), access to data, or the interaction between them? We develop a dynamic model in which firms gain knowledge from raw data using AI, but face *"informational entropy"*: without sufficient AI, more raw data leads to information overload and has negative returns. The model predicts two key dynamics: (1) improvements in AI (compute) disproportionately benefit data-rich firms; and (2) access to processed data substitutes for compute, increasing industry dynamism. We confirm these predictions using novel data from 2000–2023 and two exogenous shocks: the 2006 launch of Amazon Web Services (AWS) and the 2017 introduction of transformer-based architectures. Our findings suggest that regulating data usability, not just AI models, is essential to preserving firm dynamism in the modern economy.

Keywords: Data economy, industry dynamics, Artificial Intelligence, value of data.

JEL-Codes: D8, O3, O4, G3, L1, L2, M1.

*The authors are grateful to Andreas Fuster, Luise Eisfeld, Thierry Foucault, Leonardo Gambacorta, Wei Jiang, Christian Peukert, Vatsala Shreeti, Antoine Uettwiller, Laura Veldkamp, Venky Venakteswaran, and Xavier Vives for their feedback. Roxana Mihet acknowledges generous funding from The Sandoz Family Foundation - Monique de Meuron Programme. The authors disclose no conflicts of interest. All errors are the authors' sole responsibility.

†Roxana Mihet. Address: Quartier UNIL-Chamberonne, Extranef 138, Lausanne, 1015 Vaud, Switzerland. Phone: (+41) 0216923674. Email: roxana.mihet@unil.ch. (Corresponding author)

‡Kumar Rishabh. Address: Bank for International Settlements and University of Lausanne, Centralbahnplatz 2, 4051 Basel, Switzerland. Email: kumar.rishabh@unil.ch.

§Orlando Manuel Da Costa Gomes. Address: Polytechnic Institute of Lisbon, Estr. de Benfica 529, 1549-020 Lisboa, Portugal. Email: omgomes@iscal.ipl.pt.

1 Introduction

The rise of Artificial intelligence (AI) has drawn considerable regulatory attention, with agencies such as the Federal Trade Commission and the Justice Department focusing on AI's potential to consolidate market power.[1] This approach overlooks a critical driver: the data that fuels AI systems. AI and data resemble cars and fuel; while AI algorithms and computing power harness insights (the car), it is the access to vast datasets (the fuel) that often creates the advantage. Current regulatory frameworks may miss an important dimension by focusing solely on AI, risking the reinforcement of market power not only through AI sophistication, but also through unequal data access.

In this paper we test a theory of how data and AI jointly shape firm growth, innovation, and competitive dynamics. We assume firms generate raw data as a by-product of economic activity and accumulate knowledge through learning from data, as in Farboodi et al., 2019. Raw data, however, is noisy and subject to informational entropy — adding more data can degrade signal quality unless firms possess sufficient AI capacity to process it.[2] In contrast, processed data — which has already been structured and distilled — allows firms with limited AI capabilities to bypass computational constraints.[3] The central theoretical insight is that while AI and data are complements in extracting value from raw information, processed data can act as a substitute for internal AI infrastructure. This aligns with the concept of knowledge distillation in AI, where a high-capacity model (teacher) generates structured outputs that a lower-capacity model (student) can efficiently learn from, bypassing the computational burden of processing raw data.[4]

[1] See, also, Athey, 2018; Agrawal et al., 2018, 2022; Mihet and Philippon, 2019.

[2] Dugast and Foucault (2025) show that an abundance of raw data can worsen decision-making in financial markets, and Acemoglu et al. (2022) highlight inefficiencies arising from excessive raw data in economic environments. More broadly, research on bounded rationality and information processing constraints suggests that firms lacking sufficient AI infrastructure struggle to filter and utilize large datasets effectively (Abis and Veldkamp (2024); Farboodi and Veldkamp (2021); Van Nieuwerburgh and Veldkamp (2010)).

[3] We do not attempt to disentangle whether access to processed data arises via licensing, market trade, or via mergers and acquisitions; our framework should be read as reduced-form effects of increased access to processed information.

[4] Throughout the paper, we use the terms processed data and distilled data interchangeably, to emphasize that processed data reflects structured, teacher-generated information (i.e., distilled knowledge in the sense

Our framework yields two testable insights. First, improvements in AI disproportionately benefit data-rich firms, reinforcing existing advantages and increasing concentration. We define this effect as *AI-data complementarity*. Second, access to processed data reduces the dependence on internal AI capabilities, enabling smaller or less AI-intensive firms to compete, thereby reducing concentration. We refer to this effect as *AI-data substitutability*. These forces jointly shape the evolution of firm heterogeneity, industry dynamics, and the distribution of market power in the modern economy.

Our framework connects to and extends a broader literature on firm learning under information frictions. Canonical models in macroeconomics and finance emphasize attention constraints or fixed signal capacities (Sims, 2003). These frameworks typically assume signals are clean, but limited in flow. In contrast, we model a friction that arises from the interaction of data abundance and limited processing capacity: firms must de-noise raw data to generate useful knowledge, and failing to do so leads to declining marginal informativeness—*informational entropy* (Shannon, 1948). In our theoretical setting, AI capabilities across firms are assumed to be heterogeneous, leading to differences among firms regarding their potential ability in distilling data into knowledge.

We test the theoretical insights using novel data and two plausibly exogenous technology shocks. First, the 2006 rollout of Amazon Web Services (AWS) dramatically lowered the cost of compute, expanding access to scalable AI infrastructure. Second, the 2017 introduction of transformer-based architectures, which underpin modern foundation models like BERT and GPT, revolutionized firms' ability to structure and learn from unstructured data. These shocks provide variation in the cost of compute and processed data quality, respectively. Using novel firm-level measures of AI and data intensity, constructed from patent activity and business-model text analysis, we estimate difference-in-differences models with continuous

used in machine learning). Also, we use "AI" to refer to computational capability—that is, a firm's ability to process information using hardware, software pipelines, or pretrained models. We do not use AI to mean generative or language models specifically, though these exemplify the compute-intensive trajectory we study. This broader framing is important: we are concerned with the infrastructure for learning from data, not the specific architecture used.

treatment intensity.

We find that the AWS shock disproportionately benefits data-intensive firms, consistent with AI–data complementarity. In contrast, the transformer shock disproportionately benefits low-AI firms, consistent with AI-data substitutability. These results hold across firm size, innovation output, and market share, and are robust to alternative specifications, including placebo and falsification tests.

Our findings carry important implications for how we understand the architecture of the AI economy. Market power may stem less from algorithmic (computing) superiority than from the capacity to process and scale data efficiently. Policies that promote access to processed data—via open-source tools, application programming interfaces (APIs), or shared model layers—can reduce barriers to entry and democratize innovation, even in compute-constrained environments. This aligns with calls in Athey (2018) to promote open-source licensing and facilitate the creation of processed-data-enabling firms, and Conti et al. (2025) which highlights the importance of open source engagement for fostering innovation and competitiveness among startups. It also complements recent discussions on data-driven market power and feedback loops in AI markets (Gans, 2024). At the same time, such access raises questions around intellectual property, data control, and the incentives for foundational model development. Our theoretical framework provides a foundation for examining and measuring these trade-offs. By distinguishing between raw and processed data and endogenizing the value of information, it explains why some firms grow faster, learn better, and concentrate market share in the modern economy, and how this can change when the cost of compute or processed data falls.

The remainder of the paper is organized as follows. Section 2 presents the theoretical framework, distinguishes raw from processed data, and explores the role of data trading. Section 3, tests the theoretical predictions and highlights the main empirical results. Section 4 discusses policy and market implications. Section 5 concludes. Additional results and discussion are presented in the Appendix.

2 Theoretical Framework

We develop a dynamic model of monopolistic competition in which firms differ in their ability to extract value from data, based on Farboodi et al. (2019).[5] Firms combine capital and knowledge to produce differentiated goods. Critically, knowledge is accumulated by learning from data, as in Farboodi et al. (2019), but only if firms can process that data effectively. Without sufficient AI processing capacity, additional data may degrade decision quality, due to *informational entropy*. In contrast, firms with access to processed data or advanced AI tools can efficiently transform raw data into usable knowledge.

2.1 A Model of the Data Economy with Informational Entropy

Environment. The economic environment matches Farboodi et al. (2019).[6] A monopolistically competitive industry has firms i producing $k_{i,t}^{\alpha}$ units of quality $A_{i,t}$. Quality depends on a technique choice $a_{i,t}$ relative to an unknown persistent component $\theta_{i,t}$ and an i.i.d. transitory shock, yielding a quadratic loss from the frontier $\bar{A}$. The persistent $\theta_{i,t}$ follows AR(1). Raw data arises as a by-product of scale,

$$n_{i,t} = n_0 \cdot k_{i,t}^{\alpha}. \tag{1}$$

All primitives are standard and listed in App. A.

To fix ideas, we first present the economy, the solution, and perform numerical exercises for the baseline economy where raw data can be processed into usable data with the help of AI compute. Then, we allow trade in raw and processed data and draw the implications for equilibrium firm valuations and data prices.

[5] While our framework adopts monopolistic competition for tractability and clarity, it naturally lends itself to extension into richer strategic environments. One potential avenue is to model a multi-stage game where high-AI firms choose whether to process and sell knowledge, anticipating the responses of low-AI firms. As a first step, however, our reduced-form approach captures the essential dynamics of knowledge accumulation, data trade, and firm heterogeneity.

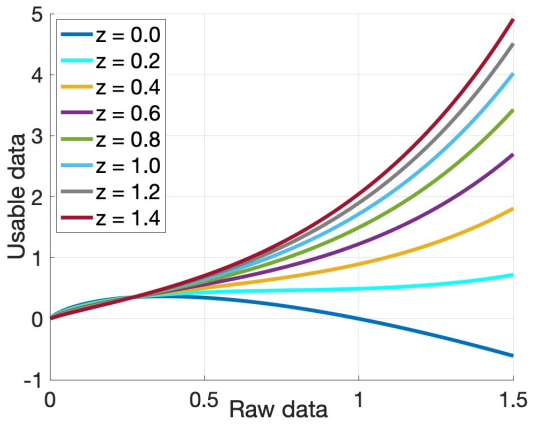
[6] Relative to Farboodi et al. (2019), our only change is the AI-dependent mapping from raw observations to usable signals, $\tilde{n}(n; z)$. This introduces (i) potential non-monotonicity of marginal informativeness in n at low z and (ii) compute–data complementarity, which together drive the different shapes our data processing function can acquire, as illustrated in the numerical exercises presented throughout the analysis.

Processed-data mapping. Let z_i be the firm's AI capability (compute). Processed data $\tilde{n}_{i,t}$ is the output of converting raw data $n_{i,t}$ into usable signals using compute z_i. We model

$$\tilde{n}_{i,t} = \underbrace{e^{-z_i} \cdot (-n_{i,t} \ln n_{i,t})}_{entropy-dominated effect} + \underbrace{(1 - e^{-z_i}) \cdot n_{i,t} e^{n_{i,t}}}_{efficient data processing} \tag{2}$$

which nests two regimes: (i) with low $z_i \approx 0$, the entropy term dominates and marginal informativeness falls at scale;[7] with higher z_i, the learning term dominates and marginal informativeness rises.

Figure 1: Processed data, $\tilde{n}_{i,t}$, for different AI levels as a function of raw data, $n_{i,t}$



Legend: Processed data $\tilde{n}$ as a function of raw data n,, for varying levels of AI intensity $z \in (0, 1.4)$. Firms with low z suffer from informational entropy; those with high z can extract increasing value from data.

Figure 1 illustrates how $\tilde{n}_{i,t}$ evolves with $n_{i,t}$ for different levels of compute z_i. When AI capability is low (low z_i), the relationship is non-monotonic: processed data initially increases with raw data, but then declines as informational entropy overwhelms the firm's ability to extract useful signals. As AI capability increases, the entropy constraint is gradually lifted, and the relationship becomes monotonic and convex. High-AI firms experience increasing returns to data processing, efficiently converting large volumes of raw data into structured, decision-relevant knowledge. This figure visualizes the core mechanism of the

[7]The function $-n \ln n$ originates from Shannon's entropy, which quantifies the expected information content of a probabilistic distribution. In our setting, it captures the idea that as the volume of raw data rises, the marginal informativeness of an extra unit may decline due to redundancy, noise, or lack of structure.

model: that AI intensity governs whether additional data amplifies or erodes a firm's ability to learn, and therefore underpins the theoretical predictions regarding complementarity and substitutability between AI and data.

Why these two basis terms? Intuitively, raw data is only valuable if a firm has the capacity to process it: to filter noise, identify structure, and extract meaningful information. When AI capability is low, this process is inefficient and may even backfire. Increasing data volume without sufficient processing power leads to redundancy or information overload—what we refer to as *informational entropy*.

A natural question arises: why can't a firm just extract a small, representative random subsample of its raw data and avoid information overload? If raw data were truly identically and independently distributed (i.i.d.) and structureless, such subsampling would be effective—and the entropy effect would be irrelevant. However, in reality, raw data is often rich in latent patterns and correlations that are not easily separable without advanced processing. Subsampling would risk discarding precisely those rare but predictive combinations. AI becomes essential not to process more data per se, but to detect structure within high-dimensional, noisy environments. This aligns with the notion that entropy effects only bite when there is something to be learned—and when the learning requires scale-sensitive signal extraction.

Robustness and alternatives. Our results are robust to replacing $n_{i,t}e^{n_{i,t}}$ by any strictly convex super-linear $g(n_{i,t})$ and $-n_{i,t}ln(n_{i,t})$ by any concave $h(n_{i,t})$ with $h'(n_{i,t}) \downarrow 0$. In App. *A* we show the same propositions hold for the class $\tilde{n}_{i,t} = (1-s(z_i))h(n_{i,t}) + s(z_i)g(n_{i,t})$, with examples $h(n_{i,t}) = n_{i,t}ln(1 + n_{i,t}^{-\phi})$ and $g(n_{i,t}) = n_{i,t}(1 - e^{\kappa n_{i,t}})^{-1}$. We keep the closed-form above in all following sections for transparency; alternatives are in App. *A*.

Empirical link. Eq. (2) formalizes how raw data becomes usable information through an AI-dependent aggregator. The entropy component captures diminishing marginal informativeness of raw scale when processing is constrained, while the learning component captures the convex gains unlocked by compute. This structure is consistent with recent

evidence: Demirer et al. (2024) document strong *complementarity* between data and compute—restricting access to either input lowers firm productivity—precisely the pattern implied when the entropy term dominates at low z. Although they do not model a processed-data market, their estimates provide an external anchor for the key premise of Eq. (2): raw data has limited standalone value without adequate processing capacity.

Beyond external validation, we provide direct empirical support for Eq. (2). In Appendix I we show that for low-AI-capability firms, higher data load depresses *idea yield on assets* and productivity, whereas for high-AI-capability firms the same increase in data load raises both (Figure 14). A further falsification test (App. J.2) examines *compute-insensitive* industries and finds that a positive computing-power shock (e.g., the AWS rollout) does not increase the marginal returns to data intensity; if anything, the post-shock slope turns negative. Together, these patterns line up with the two properties we impose on $\tilde{n}$: (i) complementarity, $\partial^2\tilde{n}/(\partial n\,\partial z) > 0$; and (ii) low-$z$ non-monotonic marginal informativeness in n.

Information, firm problem and optimal choices. Each processed data point $m \in [1, \tilde{n}_{i,t}]$ reveals a signal $s_{i,t,m} = \theta_{i,t} + \epsilon_{i,t,m}$, as in Farboodi et al. (2019), where $\epsilon_{i,t,m}$ is *i.i.d.* across firms, time, and signals. For tractability, we assume that all the shocks in the model are normally distributed: fundamental uncertainty is $\eta_{i,t} \sim N(\mu, \sigma_\theta^2)$, signal noise is $\epsilon_{i,t,m} \sim N(0, \sigma_\epsilon^2)$, and the unlearnable quality shock is $\epsilon_{a,i,t} \sim N(0, \sigma_a^2)$.

Firms optimize their production strategy choosing a sequence of production and quality decisions $k_{i,t}, a_{i,t}$ to maximize

$$\max_{k_{i,t},a_{i,t}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(A_{i,t}k_{i,t}^{\alpha} - rk_{i,t}\right) \tag{3}$$

Firms update beliefs about $\theta_{i,t}$ using Bayes' law. Each period, firms observe last period's revenues and processed data, and then choose capital level $k_{i,t}$ and production technique $a_{i,t}$.

Under Gaussian learning, the posterior precision $\Omega_{i,t} = \Sigma_{i,t}^{-1}$ is deterministic given $\tilde{n}_{i,t}$, and the technique optimally sets $a_{i,t}^* = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. The firm's dynamic problem collapses to

a single state variable $\Omega_{i,t}$, representing the stock of knowledge (i.e., the accumulated value of processed data), with usual co-state $p_{i,t}$ and single choice variable $k_{i,t}$. The Hamiltonian, Kalman updating, and the law of motion for $\Omega_{i,t}$ are standard and can be found in Appendix A, where we also give a closed-form derivative of $\tilde{n}_{i,t}$ used in the first order conditions.

Numerical exercises. Next, we conduct three baseline numerical exercises: (i) steady-state responses of firm size, profits, and knowledge to AI capability z; (ii) sensitivity to data intensity n_0 and key parameters; and (iii) a calibration of the firm-size distribution with counterfactual increases in AI compute.

AI changes steady-state firm size and profits. In App. A, Figs. 7–8, we study steady-state responses as AI capability z rises from 0 to 1.4. Profits and capital both increase with the steepest gains at low–mid z, then flatten as diminishing returns set in. The mechanism is the processed-data mapping: higher z increases the informational content of a given raw-data scale, but entropy and Bayesian updating bound gains at high z. The main difference from the model of Farboodi et al. (2019) is that, because z warps $n_{i,t}$ into $\tilde{n}_{i,t}$, profits and capital rise nonlinearly with compute, z: steep at low–mid z as entropy is overcome, then flatten, whereas in Farboodi et al. (2019), where each observation adds information directly, this would deliver nearly proportional (or at least monotone) gains without the low-z kink.

Sensitivity analysis. Profit–z curves are steepest for data-intensive firms (n_0 high), confirming that compute disproportionately benefits data-rich firms. Conversely, at low z the profit–n_0 relationship can be non-monotonic (entropy); as z rises, it turns convex and increasing, consistent with AI–data complementarity. These patterns guide our empirical tests. The details are in App. A, Fig. 9. Relative to Farboodi et al. (2019), with $\tilde{n}(n; z)$, the ranking by data intensity can *flip*: at low z, higher n_0 may *lower* value (entropy), but as z increases, the slope with respect to n_0 turns positive and convex (compute–data complementarity). Farboodi et al. (2019) delivers monotone gains in n regardless of z, so this cross-over does not arise there.

Calibrate economy to 2005 and turn on AI computing power. In a last quantitative exercise, we calibrate the economy to 2005 and turn on AI computing power. We match the 2005 Longitudinal Business Database (LBD) firm-size distribution by choosing values of $\bar{A}$ that align the steady-state size of a low-AI firm ($z = 0.15$) with the displayed size bins; we then raise z to 0.5 and 1.2 for all firms. The right tail thickens *non-homothetically*. Data-rich firms expand disproportionately, while low n_0 firms move little, an effect driven by complementarity, $\partial^2 \tilde{n}/(\partial n\, \partial z) > 0$. In a one-for-one mapping from n to precision (as in Farboodi et al. (2019)), the same shock would mainly scale the entire distribution without this differential tilt. This exercise is illustrative, not a quantitative fit; the full distributional plots and parameter list are in App. *A*, Fig. 10.

2.2 A Growth Model of the Data Economy with Data Trade

We now study two data markets that operate very differently once informational entropy is present. The *raw-data market* scales a firm's *raw* observations before they pass through the processed–data mapping $\tilde{n}(\cdot; z)$ in Eq. (2). By contrast, the *processed-data market* (distilled knowledge) adds units *directly* to usable signals, bypassing entropy. This is the key novelty relative to existing models of data trade (Farboodi and Veldkamp, 2021; Jones and Tonetti, 2021). This distinction generates the results that trading processed (distilled) information substitutes for internal compute, whereas trading raw data does not.

Intuitively, pushing more raw data points through a narrow processing pipe can deliver little (or even negative) incremental usable information at low compute capability. By contrast, each processed unit directly lifts the stock of usable signals one-for-one, effectively substituting for compute. This logic aligns with evidence that data and compute are strong complements in production: restricting either depresses productivity (Demirer et al., 2024).

Set–up. Let $\delta_{i,t}$ denote data units traded (positive if bought). Let $\iota \in (0, 1]$ capture partial rivalry/strategic loss for sellers. With baseline raw data $n_{i,t} = n_0 k_{i,t}^{\alpha}$ and AI capability z_i:

$$\underbrace{\tilde{n}_{i,t}^{\text{raw,buy}}}_{\text{buyer}} = \tilde{n}(n_{i,t} + \delta_{i,t}; z_i), \quad \underbrace{\tilde{n}_{i,t}^{\text{raw,sell}}}_{\text{seller}} = \tilde{n}(n_{i,t} - \iota\delta_{i,t}; z_i).$$

Processed-data trade shifts usable signals one-for-one:

$$\underbrace{\tilde{n}_{i,t}^{\text{proc,buy}}}_{\text{buyer}} = \tilde{n}(n_{i,t}; z_i) + \delta_{i,t}, \quad \underbrace{\tilde{n}_{i,t}^{\text{proc,sell}}}_{\text{seller}} = \tilde{n}(n_{i,t}; z_i) - \iota\delta_{i,t}.$$

Prices and quantities are pinned down by optimality condition (with H the Hamiltonian function) $\partial H/\partial\delta = 0$ for buyers and sellers together with market clearing; full Hamiltonians, first order conditions and closed forms are in Apps. B– C.

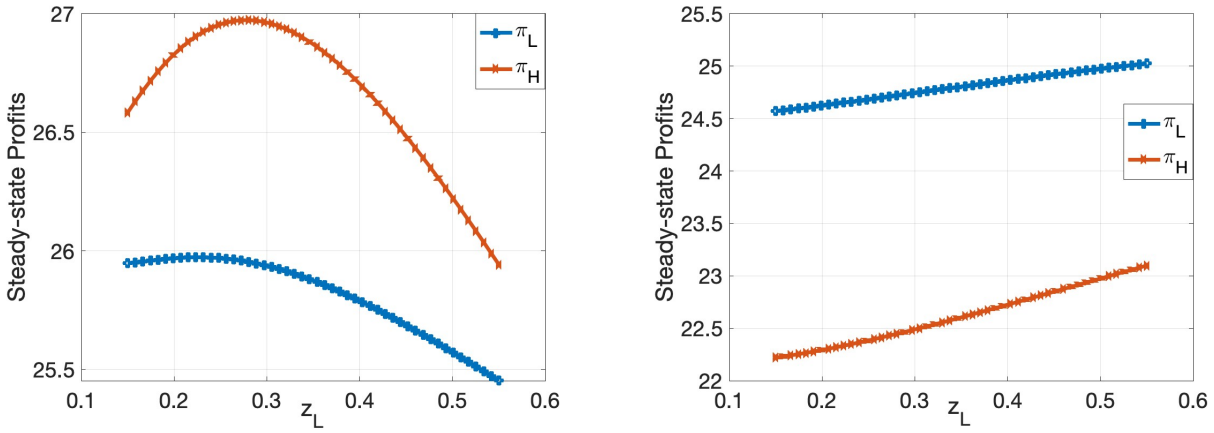
Intuition: why the two markets behave differently. (i) *Raw trade loads entropy.* For low-z buyers, adding raw volume can *lower* $\tilde{n}$ at the margin, so willingness to pay is small or even negative; the market can be thin or missing. (ii) *Processed trade bypasses entropy.* Each unit δ raises $\tilde{n}$ directly, so low-z buyers have strictly positive valuations. (iii) *Distributional effect.* Processed-data access acts as a substitute for compute and compresses the profit gap between high- and low-AI firms; raw-data trade does not.

Quantitative illustration. We compute steady-state profits for a high-AI firm (fixed z_H) and a low-AI firm as we vary z_L, comparing equilibria with (i) raw-data trade and (ii) processed-data trade. In steady state, low-AI firms gain little — and can even lose — from raw-data purchases when entropy bites, whereas processed-data access raises their profits and can close (or reverse) the high- vs. low-AI gap; raw-data prices are near/under zero while processed-data prices are strictly positive (Figs. 2–3).

Figure 2 Panel (a) shows that adding raw input to a constrained processor leaves low-AI firms on the flat (or negative) part of the $\tilde{n}$ curve; the high-AI firm, already on the increasing-returns side, benefits most. Panel (b) flips the ranking: processed units add usable signals

Figure 2: Profits for high-AI and low-AI firms with raw and processed data trade

(a) Profits with Raw Data Trade (b) Profits with Processed Data Trade



Legend: Simulated profits of high-AI and low-AI firms under two data market structures. The x-axis shows the AI intensity of low-AI firms, keeping the AI intensity of high-AI firms fixed, while the y-axis shows their steady-state profits. The left panel displays outcomes under a market for raw data; the right panel displays outcomes under a market for processed data. In the raw data regime, profits are increasing in AI, and high-AI firms maintain a persistent advantage. In contrast, under processed data trade, low-AI firms experience greater relative gains, and their profits may surpass those of high-AI firms when AI intensity is moderate, due to their improved ability to bypass informational entropy. Parameter values are: $\rho = 0.95$; $\sigma_a^2 = \sigma_\theta^2 = 0.25$; $\sigma_\epsilon^2 = 0.35$; $T = 10$; $\bar{P} = 1$; $\gamma = 0.7$; $\bar{A} = 10$; $m_H = 0.3$; $\iota = 0.6$; $r = 0.05$; $\alpha = 0.25$ $k_0 = 5$;. In all simulations, we keep $z_H = 4.5$ and we vary z_L from 0.15 to 0.55.

directly, letting low-AI firms overcome entropy and compress the profit gap. The object being traded—not just the quantity—matters for firm profitability.

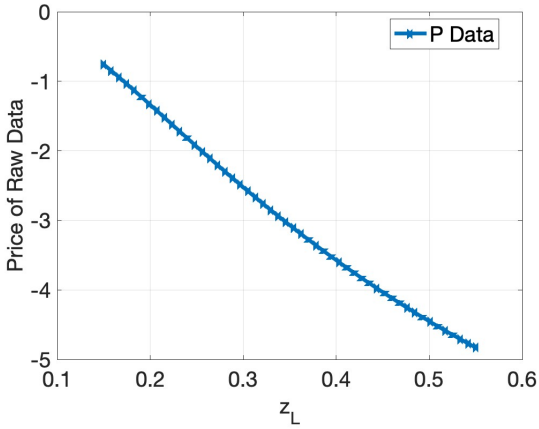
Figure 3 shows why: raw data is entropy-loaded for low-z buyers and often resembles a zero-priced byproduct, whereas processed data delivers distilled, usable signals. A real-world parallel is access to older/free ChatGPT-3–class systems versus paid frontier models (GPT-4 and -5-class): what is monetized is the *processed intelligence*, not raw tokens.

A notable implication of the model is that low-AI firms may become more profitable than high-AI counterparts, even when the latter produce and sell the processed data. Processed data acts as a substitute for compute, allowing low-AI firms to bypass entropy constraints and achieve effective learning without large AI investments, narrowing performance gaps and reshaping competitive dynamics.

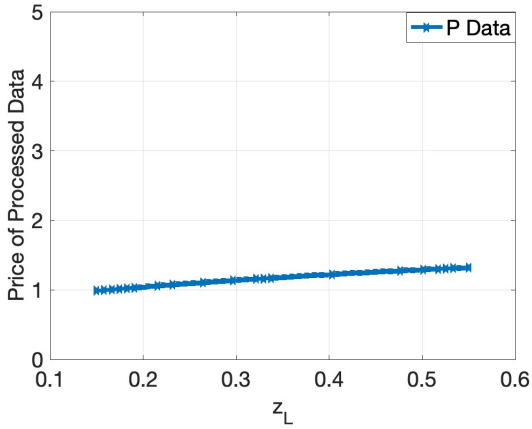
Finally, processed-data markets may be constrained institutionally: intermediaries may withhold analytics if they cannibalize other services or raise security risks (Smichowski et

Figure 3: Price of data with raw and processed data trade

(a) Price of data with raw data trade



(b) Price of data with processed data trade



Legend: Simulated data prices under two data market structures. The x-axis shows the AI intensity of low-AI firms, keeping the AI intensity of high-AI firms fixed, while the y-axis shows their steady-state profits. The left panel displays outcomes under a market for raw data; the right panel displays outcomes under a market for processed data. In the raw data regime, the price is negative and decreasing, suggesting missing markets for raw data or the fact that raw data is often shared for free. In the processed data regime, the price is strictly positive and increasing in the AI intensity of low-AI firms, suggesting processed data becomes even more valuable the higher the AI intensity of low AI firms is. Parameter values are: $\rho = 0.95$; $\sigma_a^2 = \sigma_\theta^2 = 0.25$; $\sigma_\epsilon^2 = 0.35$; $T = 10$; $\bar{P} = 1$; $\gamma = 0.7$; $\bar{A} = 10$; $m_H = 0.3$; $\iota = 0.6$; $r = 0.05$; $\alpha = 0.25$ $k_0 = 5$;. In all simulations, we keep $z_H = 4.5$ and we vary z_L from 0.15 to 0.55.

al., 2023). Our results should be interpreted in light of this friction.

Alternative mechanisms: M&A, data brokers. Firms can also acquire data—and the algorithms that process it—via mergers and takeovers rather than spot markets (Eisfeld, 2024; Gautier and Lamesch, 2020; Schoch, 2020). Relative to trade, M&A delivers lumpy (often exclusive) access to data assets, can fuse heterogeneous datasets, and frequently bundles trained models and talent. In our framework, this is closest to a one-off purchase of *processed* data with high exclusivity ($\iota \approx 1$): it bypasses entropy and can substitute for internal compute, but it also raises classic foreclosure concerns. We outline a short theoretical extension of this mechanism in App. E.

Competition authorities increasingly analyze M&A through a data lens. Google/Fitbit was cleared only with a "data silo" commitment to prevent Google from using Fitbit health data in ads; Microsoft/LinkedIn was cleared with API-access and bundling commitments

that recognized LinkedIn's data as a potential input to be foreclosed; the UK Competition Market Authority ordered Meta to divest Giphy, citing risks that control of data and integration could distort display advertising; Google/DoubleClick and Apple/Shazam were both scrutinized for data combination effects (cleared, but with data explicitly assessed). More recently, Amazon terminated its iRobot deal after EU objections that combined data and platform power could harm rivals.

M&A can replicate the positive effects of processed-data access for low-z firms (bypassing entropy), but because the access is exclusive, it can also widen gaps if the acquirer is already high-z. Our empirical analysis does not distinguish licensing/processed-data trade from data access obtained via M&A. The patterns we document are consistent with either channel–processed data can arrive through contracts or through consolidation. Disentangling these mechanisms (e.g., with deal-level data or industry split-samples) is an important direction for future work, but is beyond our scope here.

Beyond spot trade and M&A, firms routinely obtain data through intermediaries. Data brokers and cloud marketplaces license large, cleaned, and linked datasets, often non-exclusively, so the buyer receives *processed* information that slots directly into $\tilde{n}$ (a δ outside Eq. (2) with $\iota < 1$). Pricing is typically subscription or usage-based, consistent with strictly positive valuations even for low-z buyers (Liu et al., 2025). By contrast, some brokers sell more *raw-like* feeds (e.g., event-level telemetry), which behave like additional n *inside* the mapping and thus can be entropy-loaded at low z.

Data clean rooms and federated collaborations provide query access or aggregated analytics without transferring raw records. Economically, this is even closer to processed data: buyers acquire distilled statistics or model scores (pure δ outside the mapping), usually with strong non-exclusivity ($\iota \ll 1$) and contractual privacy constraints. Credit bureaus sit between these poles: partners may exchange semi-processed streams with partial rivalry (ι intermediate), and the effective object can be raw-like or processed depending on how much enrichment occurs before delivery.

In short, intermediated channels typically deliver *usable* (processed) signals rather than raw bytes, making them closer substitutes for compute and therefore especially valuable to low-z firms. Some U.S. states even maintain data-broker registries (reflecting the scale of this market), but our empirical analysis does not separate these channels from other ways firms gain access to processed information. In that sense, our analysis should be read as reduced-form effects of increased access to processed information and compute, regardless of the channel of access.

2.3 Empirically Testable Theoretical Predictions

Our theoretical framework yields several testable predictions about the relationship between firms' AI capabilities, data assets, and competitive outcomes. These predictions center on two key dimensions: the complementarity between AI and data, and the substitutability between AI capabilities and processed data.

First, our model predicts that AI computing power and data assets exhibit strong complementarities in environments with limited processed data trade. Firms with substantial data resources can extract more value from AI compute improvements than data-poor firms. Consequently, enhancements in AI compute should disproportionately benefit data-intensive firms, potentially worsening market concentration.

Theoretical Prediction 1 ***Compute favors data-intensive firms.*** *AI (compute) improvements disproportionately benefit data-rich firms, amplifying returns to data and potentially increasing market concentration.*

Second, our framework suggests that processed data and AI compute have a substitutive relationship. When processed tradable data improves in quality or the technologies to transform raw data into processed data become more efficient, the relative importance of AI computing power diminishes. Our theory predicts that shocks that improve the quality of tradable data allow even low-compute firms to extract meaningful economic value without

requiring sophisticated AI capabilities, effectively bypassing the AI-data complementarity constraint. This substitution effect implies that AI-intensive firms' competitive advantage should decrease in environments where the quality of processed data that can be traded improves. Consequently, market concentration should decline as firms with limited AI resources gain the ability to compete through better processed data quality.

Theoretical Prediction 2 ***Processed data quality improvements benefit low-compute firms.*** *When the quality of tradable processed data improves, firms can extract more economic value with lower AI (compute) capabilities, weakening the AI-data complementarity. This reduces the competitive advantage of AI-intensive firms and may lower market concentration.*

To empirically test these theoretical predictions, we need to identify settings where significant changes occurred in AI computing power and in processed data quality. Testing the complementarity hypothesis requires examining how data-intensive versus data-poor firms respond when AI computing improves. Similarly, testing the substitutability hypothesis requires analyzing how firms with different compute capabilities perform when tradable processed data becomes more qualitative and efficient. In the following section, we describe our empirical strategy that leverages technological advancements in cloud computing and natural language processing as identification opportunities to examine these predictions.

3 Empirical Investigation

3.1 Identification Strategy

We exploit two technological shocks that fundamentally altered firms' ability to leverage their resources: the introduction of Amazon Web Services (AWS) in 2006 and the development of the Transformer architecture in 2017.

The AWS shock dramatically reduced computing infrastructure costs, creating an ideal setting to test our complementarity hypothesis by observing how data-intensive firms responded when AI capabilities improved. Conversely, the Transformer shock lowered barriers to processing unstructured text data, providing the perfect environment to test our substitution hypothesis by examining how firms with different AI capabilities performed when data processing became more efficient.

It is worth emphasizing that the Transformer shock did not introduce new types of tradable data, but rather altered the quality of the information that could be shared or monetized. While firms could technically trade unstructured raw data before, its usefulness was limited for low-AI firms due to entropy effects. The breakthrough was in making distilled, high-value knowledge accessible—effectively enhancing the economic value of existing data markets. We thus interpret this as a technological innovation, not a change in market structure per se.

Our identification strategy leverages firms' varying exposures to these technological shocks based on their pre-existing AI computational capacity and data intensity. This cross-sectional variation in treatment intensity allows us to implement a difference-in-differences approach with continuous treatment variables. Our approach parallels methods in related studies: Ewens et al. (2018) document that sectors with stronger online presence experienced increased entrepreneurial entry and reduced initial capital requirements following AWS adoption, while Bonelli (2025) finds that venture capitalists exposed to cloud-impacted sectors adopted data technologies to manage the resulting surge in investment opportunities. Before presenting our empirical specifications, we explain these two shocks in detail below.

AWS Cloud Computing Shock (2006): Amazon Web Services (AWS), launched in 2006, represented a transformative technological shock that fundamentally altered computing infrastructure power, efficiency and accessibility across industries through cloud computing (Ewens et al., 2018). Originally developed for Amazon's internal needs, AWS was

unexpectedly opened to external developers, allowing businesses to rent world-class cloud computing infrastructure that *"even a kid in a college dorm room"*[8] could access by the hour (Athey, 2024).

Prior to 2006, handling large-scale computation required significant physical infrastructure investments, limiting such capabilities primarily to large, well-capitalized corporations. AWS's launch democratized access to high-performance computing by offering services like Elastic Compute Cloud (EC2) and Simple Storage Service (S3), eliminating the need for substantial upfront capital investment in servers and storage. This utility-style, pay-as-you-go model transformed computing from a fixed capital expense to a variable operational cost (Brand et al., 2024). Many firms possessed valuable data assets that remained unexploitable due to computational constraints before cloud computing. AWS potentially activated these latent data assets by making affordable computing power widely accessible—an exogenous shock to firms' ability to extract value from their existing data.

Transformers Shock (2017): The 2017 introduction of the Transformer architecture by Vaswani et al. (2017) marked a breakthrough in natural language processing. Originally developed to improve machine translation, the self-attention–based model significantly outperformed prior architectures on benchmark translation tasks. It soon became the foundation for state-of-the-art Natural Language Processing models and directly enabled the development of models such as BERT and GPT.

Transformer models fundamentally altered the economics of converting unstructured text into structured information. Before Transformers, natural language processing relied heavily on recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) architectures, which processed text sequentially word by word. These models struggled with long-range dependencies in text and faced computational bottlenecks that limited their scalability. Transformers addressed these limitations through their self-attention mechanism, which allowed the model to consider relationships between all words in a sequence simultaneously rather

[8] As described by Amazon at: https://aws.amazon.com/about-aws/our-origins

than sequentially (Vaswani et al., 2017). This architecture achieved superior results on translation benchmarks while enabling greater parallelization during training (Uszkoreit, 2017).

The Transformer architecture proved particularly valuable given that approximately 70-80% of enterprise data exists in unstructured formats (Lewington and Sartenaer, 2023), with organizations having accumulated vast repositories of text in emails, documents, and customer interactions. Prior to Transformers, extracting value from this data required extensive domain expertise, manually labeled training datasets for each specific task, and significant computational resources. The self-attention mechanism allowed models to learn broader language patterns from unlabeled data and transfer this knowledge to specific tasks with fewer labeled examples, dramatically reducing the expertise and resources needed to implement effective NLP solutions. This paradigm shift made previously inaccessible processed data suddenly available at unprecedented scale and cost-efficiency.

For our identification strategy, this shock to processed data availability created meaningful variation in treatment intensity across firms. Companies sitting on large repositories of unstructured text—but lacking efficient means to extract its value—suddenly gained improved access to powerful tools that converted their latent information assets into structured, decision-relevant data. This differential impact on firms' ability to transform raw text into processed data provides an ideal setting for our difference-in-differences approach. While the AWS shock primarily reduced the cost of computing infrastructure, the Transformer shock specifically lowered barriers to extracting structure from unstructured information.

3.2 Data and Construction of AI and Data Intensity Measures

AI intensity We construct a continuous measure of firms' AI computational capacity (*AI intensity*) that quantifies US public companies' high-performance computing capabilities for AI development and deployment. Using patent data as a proxy for these capabilities, we capture firms' investment in proprietary AI hardware and specialized computing technologies. Our measure combines three key dimensions: (1) the proportion of a firm's AI hardware

patents to its total patent stock, (2) the concentration of compute-specific patents in a firm's portfolio, and (3) the absolute count of compute-specific patents. This balanced approach identifies meaningful AI compute capabilities across firms of varying sizes and specializations. The final AI computational capacity score is a weighted combination of these three components, each converted to percentile ranks to ensure comparability. Detailed measurement methodology and data sources are provided in Appendix G.

Data Intensity: We construct a measure of firms' data intensity that captures their reliance on and utilization of large datasets in their operations and innovation. Unlike our AI computational capacity measure, which focuses on computing hardware capabilities, this approach identifies firms that leverage data extensively in their business activities.

This measure, developed in Rishabh et al. (2025), identifies data-intensive firms through a two-step process. First, we establish "base firms" that demonstrate significant innovation in broader AI applications (speech, vision, knowledge processing, and machine learning) with a moderate threshold (probability score $\geq 50\%$), as these firms necessarily build their innovations on substantial data resources. Second, we extend beyond these base firms by identifying companies whose business descriptions in 10-K filings show strong similarity to the base firms, using the Hoberg and Phillips (2016) textual similarity dataset.

For each firm-year, we calculate a data intensity score by combining three components: textual similarity to base data-intensive firms (80% weight), the firm's own share of general AI patents (10% weight), and its share of specialized AI patents in areas like machine learning and natural language processing (10% weight). This methodology enables us to identify data-intensive firms comprehensively, including those that may not actively patent their innovations but operate business models centered around data utilization. The resulting measure ranges from 0 to 1, with higher values indicating greater data intensity.

We validate our AI and data intensity measures against independent proxies capturing AI labor shares (Babina et al., 2024) and data-infrastructure spending (SWZD IT spend

data). For detailed validation results, see Appendix H.

Innovation Output and Financial Variables: We measure firms' innovation activity through their patent filings, focusing on the innovation's both scientific and economic impact. For scientific importance, we use citation-weighted patent counts, as more influential patents typically receive more citations (Kogan et al., 2017; Aghion et al., 2013). To address the truncation bias inherent in citation data—where recently filed patents have fewer citations simply due to less elapsed time—we normalize each patent's citation count by the average citations received by patents filed in the same year. This approach corrects for the right-censoring problem noted in innovation literature and the solutions proposed in the existing literature (Lerner and Seru, 2022; Hall et al., 2005). For economic significance, we construct value-weighted patent measures using market-based patent valuations that capture the wealth created for shareholders upon patent award announcements, scaled by firm assets. All patent data, including patent cites and value, are drawn from the comprehensive dataset maintained by the authors of Kogan et al. (2017).

We obtain firm level financial data from the CRSP-Compustat merged database. We construct control variables with size (log of market capitalization), growth opportunities (Tobin's Q), capital structure (leverage), profitability (Return On Assets), tangibility (Property, Plant, and Equipment expenditure to asset ratio), and innovation input (R&D expenditure to asset ratio). We use other firm financial variables such as log of assets, log of sales, log of R&D expenditure as out main outcome variables, besides the innovation output variables.

We measure firm concentration using the sales-based Herfindahl-Hirschman Index (HHI) from the Hoberg and Phillips data library. Unlike traditional industry concentration measures, this metric is firm-specific because it derives from the Text-based Network Industry Classification (TNIC) methodology, where each firm has its unique set of competitors based on textual analysis of business descriptions (Hoberg and Phillips, 2016).

To mitigate the influence of outliers, all numerical variables are winsorized at the 0.5th

and 99.5th percentiles of their respective distributions. Table 8 in Appendix *L* presents descriptive statistics for these measures.

3.3 Methodology

Our analysis spans 2000 to 2023, focusing on two major technological shocks: the AWS shock of 2006 and the Transformer shock of 2017. We employ a difference-in-differences approach with two continuous treatment intensities. For each technological shock, we include data spanning up to ten years before and after the event. Following the recommendations by Miller (2023) and Schmidheiny and Siegloch (2023), we employ a binning strategy. Specifically, we estimate distinct coefficients for the four years immediately preceding the shock (-4 to -1) and the five years immediately following it (0 to $+4$), with year immediately prior to the shock (-1) serving as the reference period. For observations falling outside these windows, we create two separate "end-cap" dummies that bin together the effects in periods five or more years before the shock and five or more years after it respectively. This approach allows us to precisely estimate immediate pre-trends and short-term effects while still leveraging the larger dataset for improved statistical power in our fixed effects estimation. The binning strategy is particularly valuable as it prevents distant observations from contaminating our reference period while avoiding sample truncation that could induce selection bias (Miller, 2023; Schmidheiny and Siegloch, 2023). This methodological choice strengthens our identification strategy by isolating the causal effect of the technological shock from potential confounding factors that may emerge over longer time horizons.

We implement this binning strategy in both our static and dynamic difference-in-difference analyses, where the static version takes the following form:

$$\begin{aligned} Y_{i,t} = {} & \beta_a(AI_Int_i \times Post_t) + \beta_d(Data_Int_i \times Post_t) + \\ & + \beta_{a+}(AI_Int_i \times Post_{5+}) + \beta_{d+}(Data_Int_i \times Post_{5+}) + X_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t} \end{aligned} \tag{4}$$

In this specification, $Y_{i,t}$ denotes the outcome of interest for firm i in year t. The variables

AI_Int_i and $Data_Int_i$ are the firm's AI and data intensity scores, respectively, as defined in Section 3.2. These are cross-sectional measures taken one year before each shock. Specifically, for the AWS analysis, we use values from 2005 (or 2004 if 2005 is missing), and for the Transformer analysis, we use values from 2016 (or 2015 if 2016 is missing). To ensure consistent scaling, both measures are converted into within-sample percentile ranks.

$Post_t$ is a dummy equal to one for the four years following the shock year (2006–2010 for AWS; 2017–2021 for the Transformer). The notation $Post_{5+}$ is a dummy that takes value 1 for observations at least five years after the shock. This binning captures the effects of the shocks in the distant post-shock periods without discarding those observations. $X_{i,t}$ are time-varying controls that include the log of market capitalization, Tobin's Q, Leverage, Returns on Assets, R&D to asset ratio. α_i is firm fixed effects, and γ_t is year fixed effects. Standard errors are clustered at the firm level.

To assess the dynamic effects of the shocks and formally test for parallel pre-trends, we also estimate an event study specification that incorporates our binning strategy for distant time periods:

$$Y_{i,t} = \sum_{l \in \mathcal{L}} \Big(\beta_{a,l} \, \text{AI_Int}_i \; + \; \beta_{d,l} \, \text{Data_Int}_i \Big) \, \text{Year}_t(l) \; + \; \alpha_i + \gamma_t + X_{i,t} \, \delta \; + \; \varepsilon_{i,t}, \tag{5}$$

where, $\mathcal{L} = \{-5+, -4, ..., 0, ..., 4, 5+\}$. and, $Year_t(l)$ is a dummy variable such that,

$$Year_t(l) = \begin{cases} 1 \quad \text{if } (t - \text{shock_year}) = l \quad \text{for } l \in \{-4, -3, -2, 0, 1, 2, 3, 4\}, 0 \text{ otherwise} \\ 1 \quad \text{if } (t - \text{shock_year}) \leq -5 \quad \text{for } l = -5+, 0 \text{ otherwise} \\ 1 \quad \text{if } (t - \text{shock_year}) \geq 5 \quad \text{for } l = 5+, 0 \text{ otherwise} \end{cases}$$

For innovation measures such as citation-weighted patent counts, we employ Poisson pseudo-maximum likelihood (PPML) estimation rather than OLS.[9] This approach addresses

[9]Specifically, for innovation measures, we estimate the PPML equivalent of static specification 4 with the conditional expectation:

$$\begin{aligned} \mathbb{E}[Y_{i,t} | \text{AI_Int}_i, \text{Data_Int}_i, X_{i,t}, \alpha_i, \gamma_t] = \exp(&\beta_a \text{AI_Int}_i \times \text{Post}_t + \beta_d \text{Data_Int}_i \times \text{Post}_t \\ &+ \beta_{a+} \text{AI_Int}_i \times \text{Post}_{5+} + \beta_{d+} \text{Data_Int}_i \times \text{Post}_{5+} + X_{i,t} + \alpha_i + \gamma_t) \end{aligned}$$

the highly skewed nature of patent data, which often contains many zeros and extreme values. While methods like log transformations have been used for such data, they introduce significant limitations, including the exclusion of zero observations and inconsistent parameter estimates (Cohn et al., 2022). PPML has emerged as the preferred method for analyzing innovation data (Azoulay et al., 2019; Aghion et al., 2013), as it accommodates fixed effects, while handling both count and continuous non-negative variables (Silva and Tenreyro, 2006; Cohn et al., 2022). Our choice aligns with recent methodological recommendations in the innovation literature (Cohn et al., 2022).

Mapping Theoretical Predictions with Regressions: Our theoretical framework yields specific predictions about how firms with different AI and data capabilities should respond to each technological shock. For the AWS shock, which reduced the cost of AI computing infrastructure, our theory predicts that more data-intensive firms should benefit disproportionately due to the complementarity between AI and data. This implies it would predict $\beta_d > 0$ in the AWS period, particularly for firm performance and innovation measures. The prediction for β_a is less clear, as firms with high AI capabilities but limited data may face diminishing returns when computing becomes more efficient.

For the Transformer shock, which lowered barriers to processing unstructured data, our theory predicts a substitution effect between AI capabilities and processed data. As processed data becomes more efficient, the competitive advantage of AI-intensive firms should diminish with low AI firms catching up, suggesting $\beta_a < 0$ for the Transformer period. Our theoretical framework does not yield an unambiguous prediction for β_d in this context, as the effect depends on how data-intensive firms adapt to the new data processing capabilities.

Identification Assumptions: Our empirical strategy relies on two key assumptions. First, we assume parallel pre-trends across firms with varying AI and data intensities, ensuring that less intensive firms provide valid counterfactuals for more intensive firms absent the

We apply a similar PPML approach to the dynamic specification in equation 5.

shocks. We test this using event studies to verify pre-shock interaction coefficients are statistically indistinguishable from zero. Second, we assume no concurrent systematic changes differentially affected firms based on their AI and data intensities. While time and firm fixed effects control for common trends and time-invariant characteristics, this is unlikely to be a serious concern given the unexpected nature of both technological shocks. In Appendix J, we conduct placebo tests using artificial shock dates and perform falsification tests on firms theoretically unaffected by these shocks to further isolate their causal effects.

3.4 Empirical Results

3.4.1 AWS Shock

Table 1 and Figure 4 present our difference-in-differences estimates of how firms with varying AI and data intensities responded to the AWS shock of 2006.

We test our first theoretical prediction by looking at the coefficient on Data Intensity (β_d) in the post-AWS period. Consistent with our theoretical predictions, more data-intensive firms gained significant advantages against less data-intensive firms, as evident from the positive coefficient β_d (coefficient for DI Intensity $\times$ Post), across multiple dimensions following AWS. In the log-sales regression ($\beta_d = 0.1901$), moving from the 25th to the 75th percentile in data intensity (a shift of 0.50 on a 0–1 scale) is associated with about 10 percent higher sales in the post-shock period. Given that average annual sales growth in our sample is 6.3 percent, this corresponds roughly to gaining one and a half years of typical firm growth. In the log-assets regression ($\beta_d = 0.0662$), the same shift corresponds to roughly 3.4 percent larger assets post-shock. With average annual asset growth of 5.3 percent in our sample, this gain is equivalent to about two-thirds of a year's typical asset expansion. These differences persist throughout the post-shock years, suggesting data-intensive firms achieved meaningful and lasting increases in scale following the AWS introduction.

Innovation metrics show a similar pattern. Firms with higher data intensity increased patenting significantly following AWS. The citation-weighted patent regression indicates that

Table 1: Impact of AWS Shock on Firm Outcomes

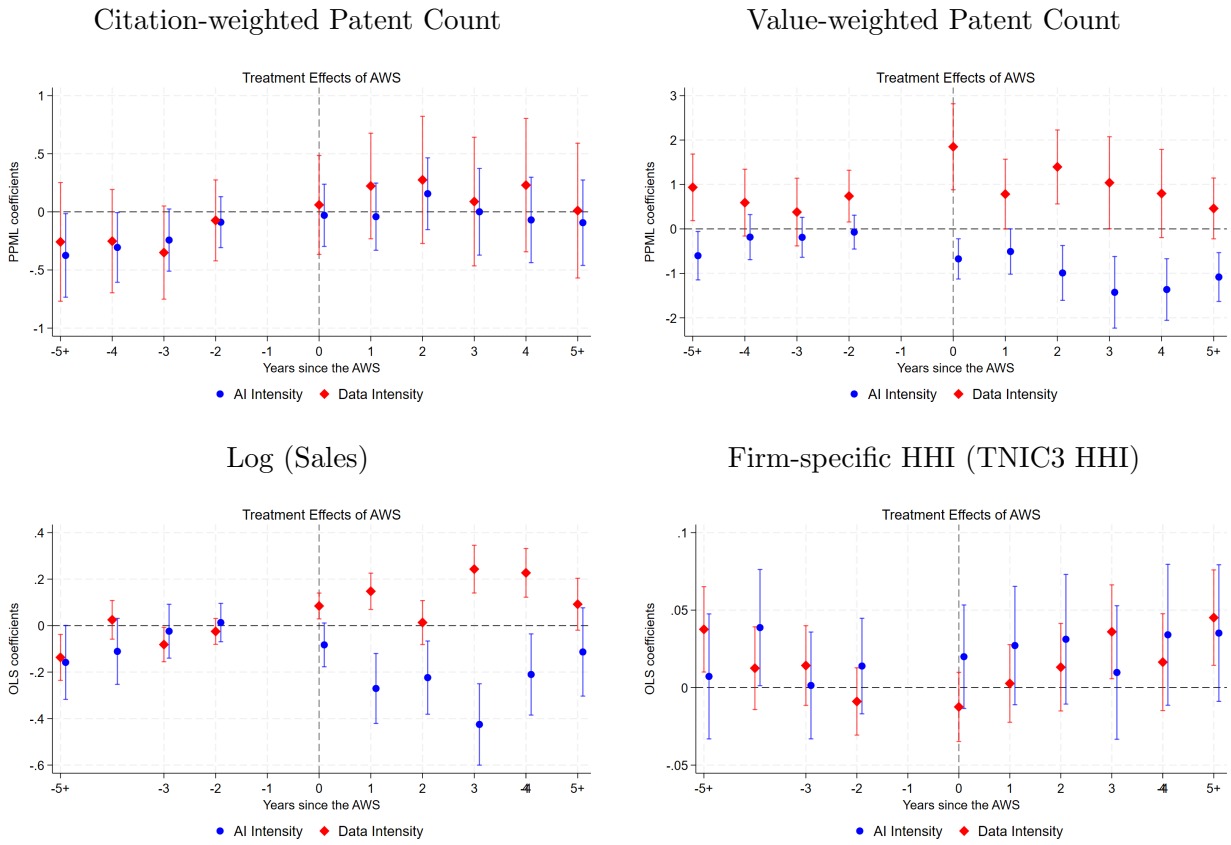
	Log (Assets) (1)	Log (Sales) (2)	HHI (TNIC3) (3)	Log (R&D Exp) (4)	Cit-wtd Count (5)	Value-wtd Count (6)
AI intensity × Post	-0.1783***	-0.1612**	0.0128	-0.0546	0.2252*	-0.6901***
	(0.0505)	(0.0674)	(0.0156)	(0.0642)	(0.1301)	(0.2314)
Data intensity × Post	0.0662**	0.1901***	-0.0050	-0.1112*	0.3644*	0.5915**
	(0.0309)	(0.0395)	(0.0101)	(0.0661)	(0.2190)	(0.2889)
AI intensity × Post_{5+}	-0.1845**	-0.0371	0.0236	-0.1360	0.1294	-0.8054***
	(0.0794)	(0.0983)	(0.0206)	(0.0997)	(0.1809)	(0.2830)
Data intensity × Post_{5+}	0.0593	0.1430**	0.0299**	-0.0326	0.2047	-0.1200
	(0.0472)	(0.0578)	(0.0144)	(0.0927)	(0.2799)	(0.3067)
Number of Observations	58530	57888	57166	24100	29293	29293
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Control	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. See notes for Figure 4 and Section 3.2 for more details on the variables. Citation-weighted (cit-wtd) patent counts weight each patent by its forward citations, adjusted for filing vintage. Value-weighted counts use market-based valuations capturing shareholder wealth effects at patent award from Kogan et al. (2017), scaled by assets. AI intensity and data intensity are cross-sectional measures defined in Section 3.2. The former captures a firm's high-performance computing capacity for AI development, combining percentile ranks of patent stock, hardware concentration, and compute capacity. The latter reflects reliance on large-scale data, identified through innovation in core AI areas and business-description similarity. Both scores range from 0 to 1 and are converted into within-sample percentile ranks. For the AWS analysis, cross-sectional intensities are taken from 2005 (or 2004 if 2005 is missing). $Post_t$ is a dummy equal to one for the four years (2006–2010) following the shock year (2006). $Post_{5+}$ equals one for all observations at least five years after the shock. This binning captures longer-run effects without discarding later observations, ensuring sufficient sample size in distant post-shock periods. Controls include log(market capitalization), Tobin's Q, leverage, Return on Assets, and asset tangibility. Patent regressions are estimated using Poisson Pseudo Likelihood Estimation.

a one-unit increase in data intensity is associated with about 44% higher patent output per year post-AWS. A realistic shift, such as moving from the 25th to the 75th percentile of data intensity, corresponds roughly to a 20% higher patent output. Although the year-to-year estimates are noisy, the average difference post-treatment is statistically significant at the 10% level. The value-weighted patent measure, capturing economic relevance, shows an even larger and immediate effect (about 81% per unit increase, or roughly 35% for an interquartile shift), sustained throughout the post-shock period. These results indicate that higher data intensity led firms to produce innovations that were both scientifically valuable and economically relevant.

Market concentration measures (TNIC3 HHI) show more modest coefficients, though firms with higher data intensity exhibit increasing concentration in the longer term (Post_{5+} coefficient of 0.0299), consistent with competitive advantages accumulating over time. Importantly, when we re-estimate concentration using the Fixed Industry Classification, which holds industry boundaries constant over time, the post-AWS effect is statistically flat (see

Figure 4: AWS Shock: Impact by AI and Data Intensities



Notes: This figure shows the dynamic effects, with 95% confidence intervals, of the AWS shock (2006—year 0 in the figure) on firm outcomes by AI and data intensity levels, as estimated using Equation 5. See the notes for Table 1 and Section 3.2 for more details on the variables.

Appendix K). The contrast suggests that the TNIC uptick largely reflects leading data-rich firms re-focusing into narrower product niches rather than gaining additional market power within a stable industry definition.

These results confirm our first theoretical prediction: when computing costs decrease, the returns to data increase. The AWS shock reduced the cost barrier to processing large datasets, enabling firms with higher data intensity to extract more value from their existing data assets. The cloud computing revolution created a complementarity between efficient computing and data assets, giving data-rich firms a competitive edge.

Turning to the results for AI intensity, firms with higher AI intensity experienced performance declines compared to their lower AI intensity counterparts, with a negative coefficient

(−0.1612) representing lower post-shock sales by about 16%. This indicates AWS led to low-AI firms catching up with high-AI intensity firms. However, the catching up of low-AI firms was not generalized as we observe conflicting results in terms of patenting. While high-AI firms show negative effects on value-weighted patents, they show positive effects on citation-weighted patents, suggesting a shift in innovation strategy rather than a comprehensive decline.

Overall, our results regarding AI intensity suggest that availability of cheaper computing might have helped low-AI firms catch up with high-AI firms in some dimensions because the AWS shock democratized access to computing infrastructure. This reduced the competitive advantage previously held by firms that had made large investments in proprietary computing resources, while simultaneously enhancing the value of data assets that could now be processed more cost-effectively.

3.4.2 Transformer Shock

To test our second theoretical prediction we switch our attention to the Transformer shock to study how processed data quality improvements affect the competitive landscape between firms with different *AI capabilities*. Table 2 and Figure 5 present our analysis.

The results strongly support this prediction through the negative coefficient on AI Intensity, β_a (coefficient for AI Intensity × Post). Our results indicate that lower-AI firms closed the performance gap against high-AI firms by approximately 37% in sales and 24% in total assets, after the Transformer shock. The event study shows this catch-up effect accelerating over time, all with parallel pre-trends that strengthen our causal interpretation.

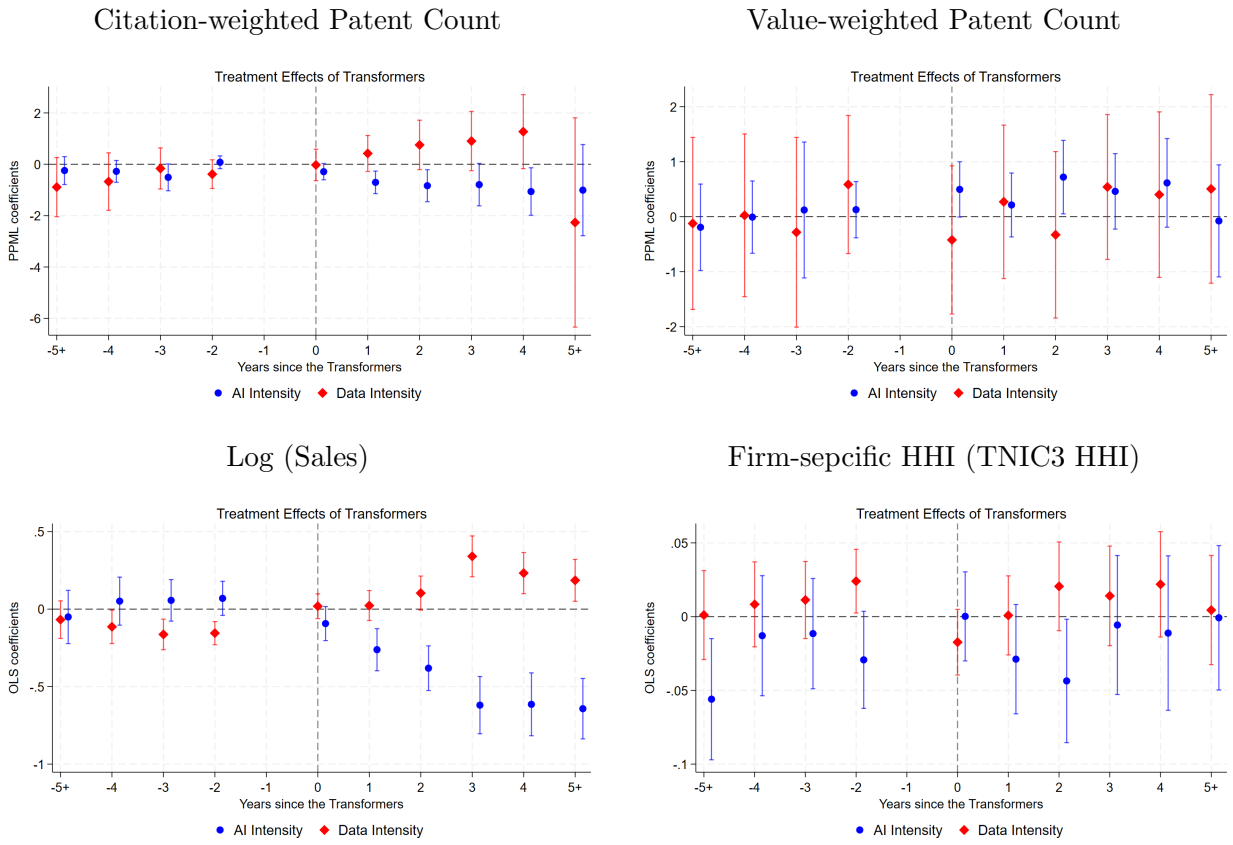
The Transformer shock equalised the quantity of inventive output but not its market value. Column (5) of Table 2 shows low-AI firms reduced the citation-weighted patent gap by roughly 40 percent ($e^{-0.5130} - 1 \approx -0.40$). However, the value-weighted measure in Column (6) shows a premium of around 70 percent for high-AI firms ($e^{0.5290} - 1 \approx 0.70$). Transformers thus democratised the volume of innovation, yet high-AI firms maintained val-

Table 2: Impact of Transformer Shock on Firm Outcomes

	Log (Assets) (1)	Log (Sales) (2)	HHI (TNIC3) (3)	Log (R&D Exp) (4)	Cit-wtd Count (5)	Value-wtd Count (6)
AI intensity × Post	-0.2417***	-0.3688***	0.0191	-0.2125**	-0.5130**	0.5290*
	(0.0525)	(0.0691)	(0.0157)	(0.0865)	(0.2288)	(0.2984)
Data intensity × Post	0.0574	0.2182***	0.0005	0.1224	1.1837**	-0.0468
	(0.0366)	(0.0493)	(0.0117)	(0.0842)	(0.5393)	(0.4397)
AI intensity × Post_{5+}	-0.3637***	-0.6256***	0.0362*	-0.5085***	-0.8184	-0.0312
	(0.0763)	(0.1003)	(0.0219)	(0.1226)	(0.8625)	(0.4762)
Data intensity × Post_{5+}	0.1459***	0.2653***	-0.0023	0.2627**	-1.6675	0.4855
	(0.0525)	(0.0702)	(0.0178)	(0.1160)	(2.1059)	(0.6670)
Number of Observations	48648	47280	46292	20389	22896	22896
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. See the notes for Table 1 and Section 3.2 for more details on the variables.

Figure 5: Transformer Shock: Impact by AI and Data Intensities



Notes: This figure shows the dynamic effects, with 95% confidence intervals, of the Transformer shock (2017—year 0 in the figure) on firm outcomes by AI and data intensity levels, as estimated using Equation 5. See the notes for Table 1 and Section 3.2 for more details on the variables.

uation advantages. Two plausible mechanisms explain this divergence: (i) reduced fixed costs allowed low-AI firms to leverage previously unused data for patenting, closing the innovation count gap; (ii) high-AI firms capitalised on complementary assets such as advanced computing infrastructure, data integration, and distribution channels, increasing the expected returns from their patents. Investor valuation likely reflects expectations that incumbents will sustain a competitive advantage, amplified by short-term market enthusiasm for advanced AI technology.

Market concentration effects show an initial decrease in concentration for high-AI firms in the immediate post-shock period, consistent with competitive catch-up by lower-AI firms. However, the longer-term coefficient (Post_{5+} of 0.0362) hints at a potential trend reversal, although we place less confidence in these estimates due to potential confounding factors over longer time horizons. Meanwhile, data-intensive firms leveraged the newly accessible processing capabilities, showing 22% higher sales and significantly increased citation-weighted patent output. This indicates that data assets became more valuable when paired with the newly democratized NLP capabilities. Market concentration effects again diverge once we hold industry boundaries fixed (Appendix K. While TNIC3 HHI hints at a long-run uptick ($\text{Post}_{5+} = 0.0362$), the FIC-based HHI (see Appendix K) actually declines in data-rich and low-AI industries, indicating that processed-data access widens innovation yet disperses market share inside a stable market definition.

These findings confirm our second theoretical prediction: when processed data becomes widely accessible, it substitutes for advanced AI capabilities, enabling lower-AI firms to catch up to high-AI counterparts in operational and innovation terms. The Transformer architecture provided a technological shortcut, allowing firms without extensive AI investments to utilize previously inaccessible unstructured data, narrowing the innovation gap. However, our theoretical framework did not explicitly anticipate incumbents' sustained valuation advantages. Investors' persistent premium on high-AI firms' innovations underscores the ongoing importance of complementary assets such as advanced computing infrastructure,

integrated data systems, and established distribution networks.

Our findings across the two technological shocks reveal asymmetric effects on how firms leverage resources for competitive advantage. The AWS shock primarily benefited data-intensive firms, reinforcing the data-feedback loop and creating additional advantages for data-rich incumbents. In contrast, the Transformer shock reduced the relative importance of AI computing power, enabling firms with less sophisticated AI capabilities to compete more effectively through easier access to higher quality processed data. These results suggest that regulatory approaches focused on data access may be more effective than AI-centered regulation for promoting market competition.

Placebo and Falsification Tests: To validate that our baseline estimates genuinely reflect technological shocks, we implement two robustness tests, described in detail in Appendix J. First, shock-timing placebos assign artificial event years earlier than the actual shocks—2002 or 2003 instead of the actual 2006 AWS event, and 2012 or 2013 instead of the 2017 Transformer event. We restrict analysis to periods strictly preceding the true shocks. These tests confirm that the baseline results vanish entirely in these placebo settings. Second, exposure falsification tests differentiate between firms theoretically sensitive and insensitive to the shocks. For AWS, baseline effects persist strongly in compute-sensitive firms but disappear in compute-insensitive firms. Similarly, Transformer effects remain significant among text-heavy B2C firms yet vanish for text-light B2B firms. These detailed falsifications thus provide rigorous evidence supporting the causal interpretation of our findings.

4 Policy and Market Implications

The global shortage of advanced semiconductor chips has emerged as a critical bottleneck in the development and deployment of artificial intelligence (AI) systems. High-performance computing (HPC) chips, essential for training and running large AI models, are increasingly scarce, driven by surging demand, geopolitical tensions, and supply chain disruptions. This

chip penury constrains the computing power available to firms, creating a substantial barrier to scaling AI capabilities and exacerbating inequality between AI-intensive and non-AI-intensive firms. In this context, the limited availability of advanced chips amplifies the importance of optimizing data processing and trading strategies and leveraging processed (i.e., distilled) data to mitigate the computational burden.

Our findings highlight the crucial role of data access and computing power in shaping firm growth and competitive dynamics in the AI-driven economy. In an AI-constrained environment, where computing power is limited, firms are unable to efficiently utilize raw data, leading to diminishing returns from data accumulation. This entropy effect underscores the importance of processed data trade as a key mechanism to sustain competition and promote firm growth.

Policymakers aiming to foster innovation must therefore focus on regulating data access alongside ensuring fairness in AI model development. Facilitating the exchange of processed data between firms can mitigate the adverse effects of raw data entropy and support smaller or less AI-intensive firms in leveraging data-driven insights. This aligns with calls from Athey (2024) to promote open-source licensing and create data-enabling firms that can act as intermediaries in processing and distributing data, enhancing competitive dynamics by making processed data more accessible. Moreover, fostering open-source engagement is essential to spur innovation and competition, as emphasized by Conti et al. (2025), which shows that startups participating in open-source communities are more likely to secure funding and develop novel technologies, leveraging collaborative innovation to overcome resource constraints. However, Conti et al. (2025) also cautions that competitive environments might reduce the benefits of open-source engagement, pointing to the need for nuanced policy approaches that account for industry structure and firm heterogeneity.

Our results also consistent with Gans (2024), which examines how the combination of data concentration and AI capabilities creates persistent market power, and argues that regulating data portability and fostering data markets could mitigate these effects by en-

abling competition across firm boundaries. These recommendations align with our findings, suggesting that processed data trade not only enhances competition but also reduces concentration risks associated with entrenched AI-driven market power.

Regulatory strategies must also account for potential geopolitical risks. Open-source processed data, while essential for domestic competition, could inadvertently empower foreign competitors to build on U.S.-originated innovations using techniques like knowledge distillation. This dynamic poses challenges to maintaining technological leadership and safeguarding economic and strategic interests.

Balancing these objectives demands policies that encourage open innovation but safeguard intellectual property. Policymakers should create frameworks enabling domestic firms to utilize processed data without exposing it to foreign appropriation. Regulators must account for internal market competition and external geopolitical pressures to manage trade-offs in AI and data regulation within a globalized economy.

5 Conclusion

We develop a dynamic framework to examine how AI capabilities and data access interact to shape firm productivity, innovation, and market structure. We distinguish between raw and processed data, introducing informational entropy as a friction that limits the value of data without sufficient AI capacity. Our central theoretical insight is that while AI and data are complements in extracting value from raw inputs, processed data can act as a substitute for compute, enabling less AI-intensive firms to compete more effectively.

We formalize these dynamics in a structural model and test them using two exogenous technological shocks: the rollout of AWS, which lowered the cost of compute, and the introduction of transformer-based models, which reduced the cost of processing unstructured data. The empirical evidence supports our predictions. Compute availability amplified the returns to data, reinforcing advantages for data-rich firms. In contrast, access to processed data

narrowed performance gaps by enabling low-AI firms to bypass computational constraints.

Our findings imply that market power in the AI economy is not solely determined by computational sophistication, but also by access to usable, structured data. Policies that broaden access to processed data—through APIs, open models, or shared layers—could reduce concentration and promote innovation. By shifting focus from AI models to the data they depend on, we provide a new lens for understanding the economics of AI-driven growth and its policy implications.

References

Abis, Simona and Laura Veldkamp, "The Changing Economics of Knowledge Production," *The Review of Financial Studies*, 2024, *37* (1), 89–118.

Acemoglu, Daron, Aryan Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar, "Too Much Data: Prices and Inefficiencies in Data Markets," *American Economic Journal: Microeconomics*, 2022, *14* (3), 1–36.

Aghion, Philippe, John Van Reenen, and Luigi Zingales, "Innovation and Institutional Ownership," *American Economic Review*, 2013, *103* (1), 277–304.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Boston, MA: Harvard Business Review Press, 2018.

_ , _ , and _ , *Power and Prediction: The Disruptive Economics of Artificial Intelligence*, Boston, MA: Harvard Business Review Press, 2022.

Athey, Susan, "The Impact of Machine Learning on Economics," in Ajay Agrawal, Joshua S Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2018, pp. 507–547.

_ , "Artificial Intelligence & Machine Learning in Finance," https://www.youtube.com/watch?v=gVCwiMqwkWA March 2024. YouTube Video.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin, "Does Science Advance One Funeral At a Time?," *American Economic Review*, 2019, *109* (8), 2889–2920.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen, "The Distinct Effects of Information Technology and Communication Technology on Firm Organization," *Management Science*, 2014, *60* (12), 2859–2885.

Bonelli, Maxime, "Data-driven Investors," *Review of Financial Studies*, 2025. Forthcoming.

Brand, James M, Mert Demirer, Connor Finucane, and Avner A Kreps, "Firm Productivity and Learning in the Digital Economy: Evidence from Cloud Computing," Working Paper 32938, National Bureau of Economic Research September 2024.

Chen, AJ, Miao Ben Zhang, and Zhao Zhang, "Talent Market Competition and Firm Growth," *Available at SSRN 4597388*, 2023.

Cohn, Jonathan B, Zack Liu, and Malcolm I Wardlaw, "Count (and Count-like) Data in Finance," *Journal of Financial Economics*, 2022, *146* (2), 529–551.

Conti, Annamaria, Christian Peukert, and Maria Roche, "Beefing IT Up for Your Investor? Engagement with Open Source Communities, Innovation, and Startup Funding: Evidence from GitHub," *Organization Science*, 2025, *36* (4), 1221–1642.

Demirer, Mert, Diego J Jiménez Hernández, Dean Li, and Sida Peng, "Data, Privacy Laws and Firm Production: Evidence from the GDPR," Working Paper 32146, National Bureau of Economic Research February 2024.

Dugast, Jérôme and Thierry Foucault, "Data Abundance and Asset Pricing," *Journal of Finance*, 2025, *80* (1), 211–258.

Eisfeld, Luise, "Entry and Acquisitions in Software Markets," Working Paper, University of Lausanne, HEC Lausanne and Swiss Finance Institute Nov 2024.

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf, "Cost of Experimentation and the Evolution of Venture Capital," *Journal of Financial Economics*, 2018, *128* (3), 422–442.

Farboodi, Maryam and Laura Veldkamp, "A Growth Model of the Data Economy," 2021, (28427).

_ , **Roxana Mihet, Thomas Philippon, and Laura Veldkamp**, "Big Data and Firm Dynamics," *AER Papers and Proceedings*, May 2019, *109*, 38–42.

Gans, Joshua S, "Market Power in Artificial Intelligence," Working Paper 32270, National Bureau of Economic Research March 2024.

Gautier, Axel and Joe Lamesch, "Mergers in the Digital Economy," Working Paper 8056, CESifo 2020.

Hall, Bronwyn H, Adam Jaffe, and Manuel Trajtenberg, "Market Value and Patent Citations," *RAND Journal of Economics*, 2005, *36* (1), 16–38.

Helfand, Jessica, Akbar Sadeghi, and David Talan, "Employment Dynamics: Small and Large Firms Over the Business Cycle," *Employment Dynamics: Monthly Labor Review*, March 2007.

Hoberg, Gerard and Gordon M. Phillips, "Scope, Scale and Concentration: The 21st Century Firm," *Forthcoming in the Journal of Finance*, 2024.

_ **and Gordon Phillips**, "Text-based Network Industries and Endogenous Product Fifferentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

İmrohoroğlu, Ayşe and Şelale Tüzel, "Firm-level Productivity, Risk, and Return," *Management science*, 2014, *60* (8), 2073–2090.

Jones, Charles I and Christopher Tonetti, "Data as an Input in Production and Growth," *Econometrica*, 2021, *89* (6), 2525–2564.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological Innovation, Resource Allocation, and Growth," *The Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

Lerner, Josh and Amit Seru, "The Use and Misuse of Patent Data: Issues for Finance and Beyond," *The Review of Financial Studies*, 2022, *35* (6), 2667–2704.

Lewington, Daniel and Laura Sartenaer, "NLP in Financial Services: Summary of Research Findings," Industry Report, LSEG Labs 2023. Accessed: 2025-03-24.

Liu, Ernest, Song Ma, and Laura Veldkamp, "Data Sales and Data Dilution," *Journal of Financial Economics*, 2025, *169*, 104053.

Mihet, Roxana and Thomas Philippon, "The Economics of Big Data and Artificial Intelligence," *International Finance Review*, 2019, *20* (3), 29–43.

Miller, Douglas L, "An Introductory Guide to Event Study Models," *Journal of Economic Perspectives*, 2023, *37* (2), 203–230.

Nieuwerburgh, S. Van and L. Veldkamp, "Information Acquisition and Under-Diversification," *American Economic Review*, 2010, *100* (2), 159–164.

Pairolero, Nicholas A, Alexander V Giczy, Gerard Torres, Tisa Islam Erana, Mark A Finlayson, and Andrew A Toole, "The artificial intelligence patent dataset (AIPD) 2023 update," *The Journal of Technology Transfer*, 2025, pp. 1–24.

Rishabh, Kumar, Roxana Mihet, and Julian Jang-Jaccard, "Cyberrisk and AI Firms," 2025. Swiss Finance Institute Research Paper n^{o} 25-39.

Schmidheiny, Kurt and Sebastian Siegloch, "On Event Studies and Distributed-lags in Two-way Fixed Effects Models: Identification, Equivalence, and Generalization," *Journal of Applied Econometrics*, 2023, *38* (5), 695–713.

Schoch, Daniela S., "Mergers and Acquisitions in the Data Economy," Working Paper 3686247, SSRN 2020.

Shannon, Claude E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948, *27* (3), 379–423.

Silva, JMC Santos and Silvana Tenreyro, "The Log of Gravity," *The Review of Economics and Statistics*, 2006, *88* (4), 641–658.

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Smichowski, Carballa Bruno, Yassine Lefouili, Andrea Mantovani, and Carlo Reggiani, "Data Sharing or Algorithm Sharing?," Working Paper 23-08, NET Institute September 2023. Available at SSRN: https://ssrn.com/abstract=4641819.

Tuzel, Selale and Miao Ben Zhang, "Economic Stimulus at the Expense of Routine-task Jobs," *The Journal of Finance*, 2021, *76* (6), 3347–3399.

Uszkoreit, Jakob, "Transformer: A Novel Neural Network Architecture for Language Understanding," https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/ August 2017. Google Research Blog, Accessed: 2025-03-24.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is All You Need," *Advances in neural information processing systems*, 2017, *30*.

Veldkamp, Laura, "Valuing Data as an Asset," *Review of Finance*, 2023, *27* (5), 1545–1562.

Zhang, Miao Ben, "Labor-technology Substitution: Implications for Asset Pricing," *The Journal of Finance*, 2019, *74* (4), 1793–1839.

Appendices

A Baseline Theoretical Model

Setup. We consider a competitive industry. Time is discrete and infinite. There is a continuum of firms indexed by i. Firm i uses $k_{i,t}$ units of capital to produce $k_{i,t}^{\alpha}$ units of goods of quality $A_{i,t}$. Let P_t denote the equilibrium price of quality-adjusted goods. The inverse demand function and the industry supply are:

$$P_t = \bar{P} Y_t^{-\gamma}, \tag{6}$$

$$Y_t = \int_i A_{i,t} k_{i,t}^{\alpha} di. \tag{7}$$

Note that firms' outputs are perfect substitutes and firms take the industry price P_t as given.

Quality depends on a firm's choice of a production technique $a_{i,t}$. In each period, and for each firm, there is one optimal technique $\theta_{i,t}$. Deviating from that optimum incurs a quadratic loss:

$$A_{i,t} = \bar{A} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2, \tag{8}$$

where $\epsilon_{a,i,t}$ is an *i.i.d.* productivity shock. The role of ϵ_a is to prevent firms from inferring θ_{it} at the end of each period. It makes the accumulation of past data a valuable asset. If a firm knew the current value of $\theta_{i,t}$, it could maximize quality by setting $a_{i,t} = \theta_{i,t}$. In our model, however, $\theta_{i,t}$ is unknown and follows an AR(1) process: $\theta_{i,t} = \bar{\theta} + \rho(\theta_{i,t-1} - \bar{\theta}) + \eta_{i,t}$ where $\eta_{i,t}$ is *i.i.d.* across time and firms.

Raw and Processed (Usable Data). Raw data is a by-product of economic activity. We assume that the number of raw data points observed by firm i a time t depends on their t production $k_{i,t}^{\alpha}$, with n_0 representing a parameter for the efficiency with which firms accumulate raw data as a function of its production volume.

$$n_{i,t} = n_0 * k_{i,t}^{\alpha}. \tag{9}$$

Raw data can be used to accumulate knowledge, but it is subject to entropy (large amounts of data contribute negatively to knowledge accumulation, rather than contributing positively). For example, without the proper AI tools, a firm can make wrong predictions when confronted with large quantities of data. In this perspective, processed data corresponds to:

$$\tilde{n}_{i,t} = -n_{i,t} * ln(n_{i,t}). \tag{10}$$

After a given threshold, collecting additional raw data with no artificial intelligence (AI) to transform into into processed knowledge becomes counterproductive – it lowers the stock of knowledge instead of increasing it.

Artificial Intelligence (AI), denoted by firm-specific capabilities $z_i \geq 0$, changes this scenery, transforming the processed data equation into:

$$\tilde{n}_{i,t} = e^{-z_i}(-n_{i,t}ln(n_{i,t})) + (1 - e^{-z_i})n_{i,t}e^{n_{i,t}} \tag{11}$$

where z_i is a firm-specific parameter that governs AI-intensity (i.e., computing power). If $z_i = 0$, then we have the aforementioned scenario where additional raw data is counterproductive. As z_i rises above zero, the entropy effect is progressively overcome and the value of the flow of processed data, $\tilde{n}_{i,t}$, increases, as large amounts of raw data are accumulated.

Each processed data point $m \in [1, \tilde{n}_{i,t}]$ reveals a signal, $s_{i,t,m}$ given by

$$s_{i,t,m} = \theta_{i,t} + \epsilon_{i,t,m}, \tag{12}$$

where $\epsilon_{i,t,m}$ is $i.i.d.$ across firms, time, and signals. For tractability, we assume that all the shocks in the model are normally distributed: fundamental uncertainty is $\eta_{i,t} \sim N(\mu, \sigma_\theta^2)$, signal noise is $\epsilon_{i,t,m} \sim N(0, \sigma_\epsilon^2)$, and productivity shock is $\epsilon_{a,i,t} \sim N(0, \sigma_a^2)$.

Closed-form derivatives and sign properties The first and second derivatives of Eq.(11) are:

$$\frac{\partial \tilde{n}}{\partial n}(n,z) = -\,e^{-z}(\ln n + 1) \;+\; (1-e^{-z})\,e^{n}(1+n), \tag{13}$$

$$\frac{\partial^2 \tilde{n}}{\partial n^2}(n,z) = -\,\frac{e^{-z}}{n} \;+\; (1-e^{-z})\,e^{n}(2+n), \tag{14}$$

$$\frac{\partial \tilde{n}}{\partial z}(n,z) = e^{-z}\,n\ln n \;+\; e^{-z}\,ne^{n}, \tag{15}$$

$$\frac{\partial^2 \tilde{n}}{\partial n\,\partial z}(n,z) = e^{-z}\Big[\;\ln n + 1 + e^{n}(1+n)\;\Big]. \tag{16}$$

Lemma A.1 (Complementarity threshold) *Define $\underline{n} \in (0,1)$ as the unique root of*

$$\Psi(n) \;\equiv\; e^{n}(1+n) \;+\; \ln n \;+\; 1 \;=\; 0.$$

Then for all $n \geq \underline{n}$ and all $z \geq 0$, the cross-partial is strictly positive: $\frac{\partial^2 \tilde{n}}{\partial n\,\partial z}(n,z) > 0$.

Proof: By (16), $\frac{\partial^2 \tilde{n}}{\partial n\,\partial z} = e^{-z}\Psi(n)$. Since $e^{-z} > 0$, the sign equals the sign of $\Psi(n)$. Note that $\Psi(1) = e(2) + \ln 1 + 1 = 2e + 1 > 0$ and $\lim_{n\downarrow 0}\Psi(n) = -\infty$. Continuity implies a unique root $\underline{n} \in (0,1)$; for all $n \geq \underline{n}$, $\Psi(n) > 0$. QED.

Remark. Numerically, $\underline{n} \approx 0.108$.[10]

Proposition A.1 (Low-z non-monotonicity in raw data) *For each $z \geq 0$, define $F(n;z) = \frac{\partial \tilde{n}}{\partial n}(n,z)$ from (13). (a) At $z = 0$, $F(n;0) = -(\ln n + 1)$, which is positive for $n < e^{-1}$ and negative for $n > e^{-1}$. (b) There exists $\bar{z} > 0$ such that for all $0 \leq z < \bar{z}$ there is a unique threshold $n^{(z)\in(e^{-1},1]}$ with $F(n;z) > 0$ for $n < n^{\star}(z)$ and $F(n;z) < 0$ for $n > n^{\star}(z)$. For $z \geq \bar{z}$, $F(n;z) \geq 0$ for all $n \in (0,1]$.*

Proof: Part (a) follows directly from $F(n;0) = -(\ln n + 1)$ and the identity $\ln n + 1 = 0 \iff n = e^{-1}$. For (b), observe that $F(\cdot;z)$ is continuous in n and strictly increasing for sufficiently large n because $\frac{\partial^2 \tilde{n}}{\partial n^2} = (1-e^{-z})e^{n}(2+n) - e^{-z}/n$ eventually becomes positive. At $n = 1$, $F(1;z) = -e^{-z} + 2e(1-e^{-z}) = 2e - (1+2e)e^{-z}$. Define $\bar{z}$ by $F(1;\bar{z}) = 0$, i.e.

$$e^{-\bar{z}} = \frac{2e}{1+2e} \quad \Rightarrow \quad \bar{z} = \ln\!\Big(\frac{1+2e}{2e}\Big) \approx 0.168.$$

[10] Solve $e^{n}(1+n) + \ln n + 1 = 0$. A bisection/Newton step yields $\underline{n} \approx 0.1077$.

For $z < \bar{z}$, we have $F(1; z) < 0$ while $F(n; z) \to +\infty$ as $n \downarrow 0$ (because $-e^{-z}(\ln n + 1) \to +\infty$). By the intermediate value theorem and the fact that $F(\cdot; z)$ crosses zero only once (its derivative $\partial F/\partial n$ changes sign at most once on $(0, 1]$), there exists a unique $n^{(z)\in(e^{-1},1]}$ with the stated sign pattern. If $z \geq \bar{z}$, then $F(1; z) \geq 0$ and $F(n; z) \geq 0$ for all $n \in (0, 1]$ by monotonicity. QED.

Generalizations Consider the generalized aggregator

$$\tilde{n}(n, z) \;=\; (1 - s(z))\, h(n) \;+\; s(z)\, g(n), \tag{17}$$

with the following conditions:

1. $s : \mathbb{R}_+ \to [0, 1)$ is C^1, strictly increasing, $s(0) = 0$, $s'(z) > 0$.
2. $h : (0, 1] \to \mathbb{R}_+$ is C^2, concave, with $h'(n)$ strictly decreasing and $\lim_{n\downarrow 0} h'(n) = +\infty$.
3. $g : (0, 1] \to \mathbb{R}_+$ is C^2, strictly convex, with $g'(n)$ strictly increasing and unbounded as $n \uparrow 1$.

Then

$$\frac{\partial \tilde{n}}{\partial n} = (1 - s)h'(n) + s\, g'(n), \qquad \frac{\partial^2 \tilde{n}}{\partial n\, \partial z} = s'(z)\big[g'(n) - h'(n)\big].$$

Thus the cross-partial is positive whenever $g'(n) > h'(n)$. By continuity and the stated limits, there exists a unique $\underline{n} \in (0, 1)$ such that $g'(n) \geq h'(n)$ for all $n \geq \underline{n}$, which delivers complementarity on $[\underline{n}, 1]$. Moreover, for any fixed $z \in [0, \bar{z})$ small enough, $\partial \tilde{n}/\partial n$ is positive for low n and becomes negative for high n, yielding the low-z non-monotonicity in Proposition A.1. Examples include

$$h(n) = n \ln(1 + n^{-\phi}),\ \phi > 0, \qquad g(n) = \frac{n}{1 - e^{-\kappa n}},\ \kappa > 0,$$

which replicate the same sign properties as Eq.(11).

Firm Problem and Solution. Firms optimize their production strategy choosing a sequence of production and quality decisions $k_{i,t}, a_{i,t}$ to maximize

$$\max_{k_{i,t},a_{i,t}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(A_{i,t} k_{i,t}^{\alpha} - r k_{i,t} \right) \tag{18}$$

Firms update beliefs about $\theta_{i,t}$ using Bayes' law. Each period, firms observe last period's revenues and processed data, and then choose capital level $k_{i,t}$ and production technique $a_{i,t}$. The information set of firm i when it chooses $a_{i,t}$ is $\mathcal{I}_{i,t} = [\{A_{i,\tau}\}_{\tau=0}^{t-1}; \{\{s_{i,\tau,m}\}_{m=1}^{\tilde{n}_{i,\tau}}\}_{\tau=0}^{t}]$.

The state variables of the recursive problem are the prior mean and variance of beliefs about $\theta_{i,t-1}$, and, last period's revenues, and the set of signals. Taking a first order condition with respect to the technique choice, we find that the optimal technique is $a_{i,t}^* = E_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. According to Bayes' law, the posterior variance of beliefs is $E_i[(E_i[\theta_{i,t}|I_{i,t}] - \theta_{i,t})^2] := \Sigma_{i,t}$. Thus, expected productivity is $E_i[A_{i,t}] = \bar{A} - \Sigma_{i,t} - \sigma_a^2$. We can then write a Hamiltonian equation for expected firm value.

Lemma A.2 *The optimal sequence of capital investment choices $\{k_{i,t}\}$ solves the following firm Hamiltonian problem:*

$$\begin{aligned} H(\Omega_{i,t}, k_{i,t}, p_{i,t}) = & \left(\bar{A} - \Omega_{i,t}^{-1} - \sigma_a^2\right) k_{i,t}^{\alpha} - r k_{i,t} + \\ & + \beta p_{i,t+1} \left[\left[\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2 \right]^{-1} + \tilde{n}_{i,t} \sigma_\epsilon^{-2} - \Omega_{i,t} \right] \end{aligned} \tag{19}$$

where $p_{i,t}$ is the shadow-price or co-state variable associated with knowledge, processed data is $\tilde{n}_{i,t} = e^{-z_i}(-n_{i,t}\ln(n_{i,t})) + (1 - e^{-z_i}) n_{i,t} e^{n_{i,t}}$ and the knowledge of firm i at time t is

$$\Omega_{i,t+1} = \left[\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2 \right]^{-1} + \tilde{n}_{i,t} \sigma_\epsilon^{-2} \tag{20}$$

This result greatly simplifies the problem by collapsing it to a deterministic problem with only one state variable, $\Omega_{i,t} = \Sigma_{i,t}^{-1}$.[11] The reason we can do this is that productivity

[11] For any $k_{i,t} \forall\, t$, the optimal choice of technique is always the same: $a_{i,t}^* = E_i[\theta_{i,t}|I_{i,t}]$. The way $a_{i,t}$ enters into expected productivity $A_{i,t}$ is through $E[(E[\theta_{i,t}|I_{i,t}] - \theta_{i,t})^2]$, which is the conditional variance $\Sigma_{i,t}$. We can replace the entire sequence of $a_{i,t}^*$ with the sequence of variances, which is deterministic here because of normality. The only randomness in this model comes from the signals and their realizations, but they never

$A_{i,t}$ depends on the conditional variance of $\theta_{i,t}$ and because the information structure is similar to a Kalman filtering system, where the sequence of conditional variances is generally deterministic. This Kalman system has a 2-by-1 observation equation. At each date t, data is about $\theta_{i,t}$ and output is informative about $\theta_{i,t-1}$. Let $\bar{s}_t$ be the average data realization. Note that last period's output reveals productivity $A_{i,t-1}$, which, in turn, reveals $\tilde{y}_{i,t-1} := (\theta_{i,t} + \epsilon_{a,i,t})$.

The conditional mean and variance evolve as

$$\Sigma_{i,t} = \left[\left[\rho^2(\Sigma_{i,t-1}^{-1} + \sigma_a^{-2})^{-1}+\right]^{-1} + \tilde{n}_{i,t}\sigma_\epsilon^{-2}\right]^{-1} \tag{21}$$

$$\mu_{i,t} = (1 - \tilde{n}_{i,t}\sigma_\epsilon^{-2}\Sigma_{i,t})\left[\bar{\theta}(1-\rho) + \rho\left((1 - M_{i,t})\mu_{i,t-1} + M_{i,t}\tilde{y}_{i,t-1}\right)\right] + \tilde{n}_{i,t}\sigma_\epsilon^{-2}\Sigma_{i,t}\bar{s}_t \tag{22}$$

where $M_{i,t} = \sigma_a^{-2}(\Sigma_{i,t-1} + \sigma_a^{-2})^{-1}$.

Here, $k_{i,t}$ is a deterministic function of the prior knowledge $\Omega_{i,t-1}$ and the processed data flow $\tilde{n}_{i,t}$. The updated knowledge $\Omega_{i,t}$ is a deterministic function of $\tilde{n}_{i,t}$, and past precisions (knowledge). Processed data $\tilde{n}_{i,t}$ depends on $k_{i,t}$. There is one random state variable, $\mu_{i,t}$. But it is only relevant for choosing the optimal technique $a_{i,t}$.

The solution to the firm's Hamiltonian comes from the first-order conditions with respect to capital and to knowledge,

$$\frac{\partial H}{\partial k_{i,t}} = 0 \implies \alpha\left[\bar{A} - \left(\Omega_{i,t}^{-1} + \sigma_a^2\right)\right] k_i^{-(1-\alpha)} - r + \beta p_{i,t+1}\frac{d\tilde{n}_{i,t}}{dk_{i,t}}\sigma_\epsilon^{-2} = 0 \tag{23}$$

$$\beta p_{i,t+1} - p_{i,t} = -\frac{\partial H}{\partial \Omega_{i,t}} \implies \left[\rho + \frac{\sigma_\theta^2}{\rho(\Omega_{i,t} + \sigma_a^{-2})}\right]^{-2} \beta p_{i,t+1} = p_{i,t} - \Omega_{i,t}^{-2}k_{i,t}^{\alpha} \tag{24}$$

The derivative of the processed data generation function with respect to capital is given

affect the conditional variance (as the mean and variance for normal variables are independent).Thus, $\Sigma_{i,t}$ is a sufficient statistic for $n_{i,t}$ and $\Sigma_{i,t-1}$. Moreover, we can also drop the mean $\mu_{i,t}$ as a state as it only matters for determining $a_{i,t}$ and does not affect anything else.

by

$$\frac{d\tilde{n}_{i,t}}{dk_{i,t}} = -\alpha n_0 k_{i,t}^{-(1-\alpha)} \left[e^{-z_i} \left(1 + \ln(n_0 k_{i,t}^{\alpha})\right) - (1 - e^{-z_i}) \left(1 + n_0 k_{i,t}^{\alpha}\right) e^{n_0 k_{i,t}^{\alpha}} \right] \tag{25}$$

Substituting expressions above and simplifying yields

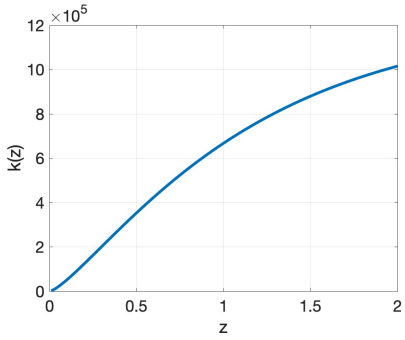
$$r k_{i,t}^{1-\alpha} = \alpha A_{i,t} - \alpha\beta p_{i,t+1} \sigma_{\epsilon}^{-2} n_0 e^{-z_i} \left[1 + \ln(n_{i,t}) + (1 - e^{z_i})(1 + n_{i,t}) e^{n_{i,t}}\right] \tag{26}$$

where $n_{i,t} = n_0 * k_{i,t}^{\alpha}$ and $A_{i,t} = \left[\bar{A} - \left(\Omega_{i,t}^{-1} + \sigma_a^2\right)\right]$ and

$$p_{i,t} = \Omega_{i,t}^{-2} k_{i,t}^{\alpha} + \beta \cdot p_{i,t+1} \left[\rho + \sigma_{\Theta}^2 \left(\Omega_{i,t} + \sigma_a^{-2}\right)/\rho\right]^{-2} \tag{27}$$

Equation 26 plays a key role in our analysis. In particular, we see that $k_{i,t}$ increases with z_i. Firms are forward looking. They know that by being active today they will generate data that will help them in the future.

Figure 6: Capital, k as a function of AI-intensity z



Parameter values are: $\alpha = 0.25$; $r = 0.05$; $A_{i,t} = 1.0$; $\beta = 0.95$; $p_{i,t} = 1$; $\sigma_{\epsilon}^2 = 0.2$; $n_0 = 0.35$; $n_{i,t} = 5$.

How AI and Data Change Steady State Firm Size

Our main experiment is to study firm size when we vary data-intensity, n_0, and AI-intensity, z. The steady-state is given by the first order conditions with respect to capital and knowledge and the law of motion for posterior variance. The system of 8 equations below defines the steady-state, with 8 endogenous variables: $n, \tilde{n}, k, \Sigma, \Omega, A, \Pi, p$ and parameters: $n_0, \alpha, z, \rho, \sigma_a^2, \sigma_\theta^2, \sigma_\epsilon^2, \bar{A}, r$

$$n = n_0 \cdot k^\alpha \tag{28}$$

$$\tilde{n} = e^{-z}\left(-n \ln n\right) + \left(1 - e^{-z}\right) n e^n \tag{29}$$

$$\Sigma^{(+1)} = \left[\left(\rho^2 \cdot \left(\Sigma^{-1} - \frac{1}{\sigma_a^2}\right)^{-1} + \sigma_\theta^2\right)^{-1} + \tilde{n} \cdot \frac{1}{\sigma_\epsilon^2}\right]^{-1} \tag{30}$$

$$\Omega = \Sigma^{-1} \tag{31}$$

$$A = \bar{A} - \Sigma - \sigma_A^2 \tag{32}$$

$$\Pi = A \cdot k^\alpha - r \cdot k \tag{33}$$

$$r \cdot k^{1-\alpha} = \alpha A - \beta \cdot p^{(+1)} \cdot \frac{1}{\sigma_\epsilon^2} \cdot \alpha n_0 \left[e^{-z}(1 + \ln n) - \left(1 - e^{-z}\right)(1+n)e^n\right] \tag{34}$$

$$p = \Sigma^2 \cdot k^\alpha + \beta \cdot p^{(+1)} \left[\rho + \frac{\sigma_T^2 \left(\Sigma^{-1} + \frac{1}{\sigma_A^2}\right)}{\rho}\right] \tag{35}$$

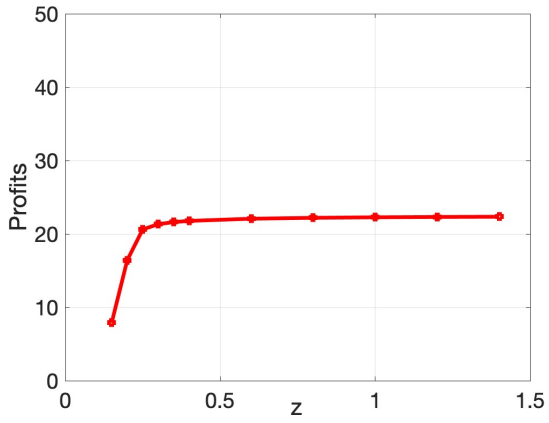
Note: numerically, we choose $\bar{P}$ to normalize $P = 1$ in the steady-state.

Quantitative experiments

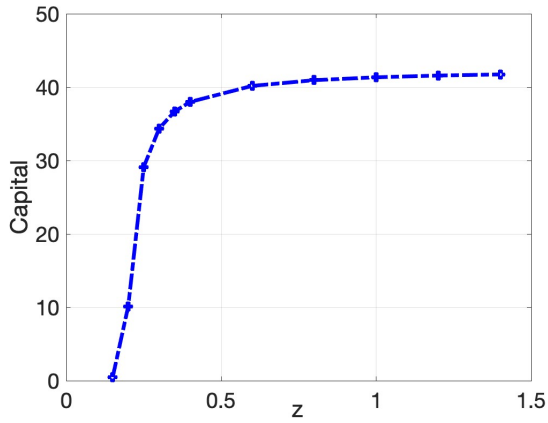
(1) AI changes steady-state firm size and profits Our first numerical experiment investigates how advancements in firms' AI capabilities (i.e., computing power) influence firms' profits and productivity levels. We increase firms' AI processing capabilities from $z_i = 0$ to $z_i = 1.4$ to examine how firms adapt in steady-state. To ensure interpretability and relevance, we select parameter values that yield realistic outcomes.

Figure 7: Steady-state profits and capital vs. z; steep gains at low–mid z, flattening at high z.

a) Steady-state profits



b) Steady-state capital

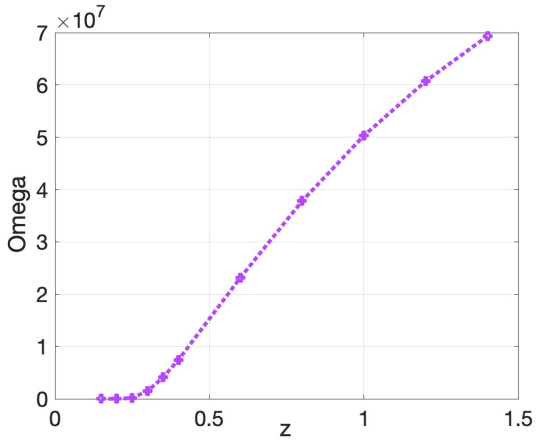


Legend: The left hand-side figure plots profits in steady-state, $A_{ss}k_{ss}^{\alpha} - rk_{ss}$, and the right hand-side figures plots capital in steady-state, k_{ss} for different AI computing-power levels, $z \in (0, 1.4)$. Higher AI capability increases firm profits and capital until diminishing returns set in. Parameter values are: $\rho = 0.95$; $\sigma_a^2 = \sigma_\theta^2 = 0.25$; $\sigma_\epsilon^2 = 0.35$; $T = 10$; $\bar{P} = 1$; $\gamma = 0.7$; $\bar{A} = 10$; $\alpha = 0.25$; $n_0 = 0.35$; $r = 0.05$; $k_0 = 5$;. Differences from Farboodi et al. (2019): The curvature and cross-overs come from the AI-dependent processed-data mapping $\tilde{n}(n; z)$; without it, the same exercises lack the low-z drag and the differential tilt toward data-rich firms.

As shown in Figure 7, higher AI capabilities imply both higher profits and higher capital in steady-state. The gradient is largest for lower AI-intensity firms, implying that these firms benefit the most from advances in computing power. It is also worth noticing that while firms' profits and capital investments increase steeply for low values of computing power, z, there is an upper barrier to the maximum firm size and profitability that can be attained. Our model reveals that firms' profits and capital peak at a certain level and do not grow indefinitely, reflecting a saturation effect driven by the interaction between data volume and AI processing capacity. This outcome arises from the entropy effect: as firms accumulate larger volumes of raw data, the marginal utility of additional data decreases unless sufficient AI capabilities are available to process it efficiently. This phenomenon inherently limits the returns to data accumulation, particularly for low-AI firms that struggle to extract meaningful insights from vast, unstructured datasets. Even high-AI firms face diminishing returns when the computational burden of processing additional data grows exponentially,

leading to a steady-state equilibrium where capital and profits stabilize. This finding aligns with Farboodi and Veldkamp (2021), which also highlights diminishing marginal returns to data and the costs of data processing and learning. In their framework, firms collect and analyze data to gain competitive advantages, but the incremental knowledge gained from additional data decreases as the volume of accumulated data rises. Our results mirror this insight, demonstrating that the burden of processing raw data can outweigh its potential benefits, particularly when firms lack adequate AI infrastructure.

Figure 8: Steady-state knowledge Ω vs. z; compute overcomes entropy at low z.



Legend: AI intensity determines whether raw data leads to knowledge gains or overload; higher AI enables more effective data distillation and higher knowledge, Ω. Parameter values are as before. Differences from Farboodi et al. (2019): The curvature and cross-overs come from the AI-dependent processed-data mapping $\tilde{n}(n; z)$; without it, the same exercises lack the low-z drag and the differential tilt toward data-rich firms.

Figure 8 plots the steady-state knowledge stock, Ω, for different levels of AI computing-intensity, z. In the low to moderate AI range, increasing AI capacity significantly improves the ability to extract meaningful information from raw data. This phase corresponds to the overcoming of the entropy effect, where AI tools enable firms to efficiently sift through large volumes of data, turning noisy and unstructured information into actionable knowledge. During this phase, the marginal productivity of AI is high, as each additional unit of AI investment yields substantial improvements in knowledge accumulation. But, beyond a certain threshold of AI capacity, the gradient starts to decrease, indicating diminishing returns to AI investment. This phenomenon can be interpreted as reaching a saturation

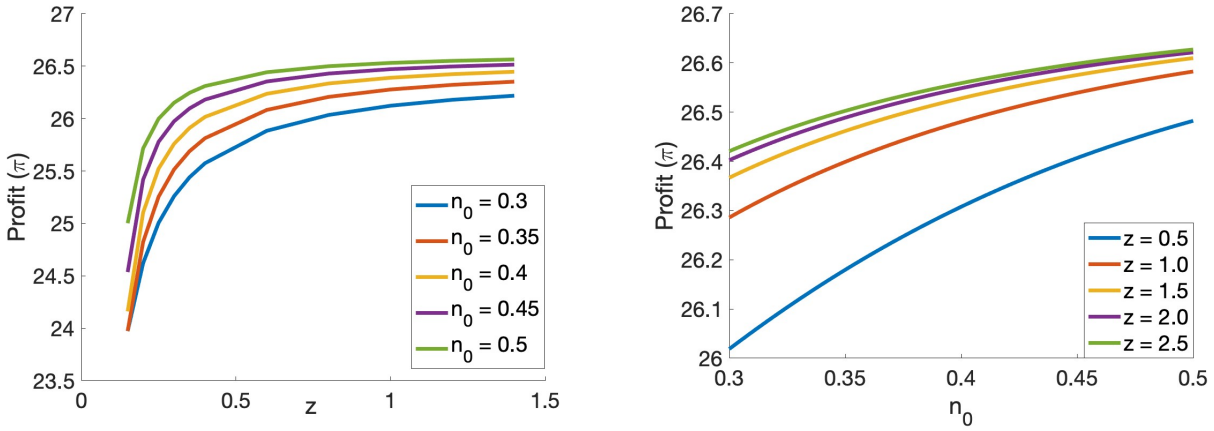
point where the AI system has become sufficiently capable of processing data efficiently, and further improvements in AI capability do not yield proportionally higher knowledge gains. Essentially, at high levels of AI, most of the low-hanging fruit of data processing efficiency has been captured, and the marginal benefit from adding more AI becomes limited.

(2) Sensitivity analysis. The sensitivity analyses presented in Figure 9 indicate how firm profits vary as a function of the data-intensity (n_0) and the AI-intensity (z) parameters.

Figure 9: Sensitivity: profit vs. z across n_0, and profit vs. n_0 across z; complementarity emerges with higher z.

(a) Profit as a function of AI intensity (z) for different data-intensity levels (n_0)

(b) Profit as a function of raw data intensity (n_0) for different AI intensity levels (z)



Legend: Simulated relationship between firm profits and AI intensity under different levels of raw data capability. The x-axis shows firm-level AI intensity, and the y-axis shows steady-state profits. Each curve represents a different level of raw data intensity. When AI intensity is low, profits are non-monotonic in data due to informational entropy. As AI intensity increases, the relationship becomes convex and increasing, reflecting complementarity between AI and data. Parameter values as before. Differences from Farboodi et al. (2019): The curvature and cross-overs come from the AI-dependent processed-data mapping $\tilde{n}(n; z)$; without it, the same exercises lack the low-z drag and the differential tilt toward data-rich firms.

Panel (a) shows that firms with the highest data intensity (n_0) firms benefit the most from improvements in AI. When AI intensity is low, it is these data-intensive firms that experience the steepest profit growth as AI improves. This indicates that AI advancements are most impactful when paired with extensive raw data accumulation. However, as AI intensity reaches higher levels, the curve flattens, suggesting that further increases yield smaller profit gains.

Panel (b) conveys that the synergy between data accumulation and AI intensity is crucial for maximizing profits. AI investments are most beneficial when they complement a substantial data foundation. Firms that are already data-intensive can further boost profits with AI improvements, but the marginal benefit decreases as the data accumulation becomes extremely efficient. On the other hand, firms with low data intensity see limited gains from increasing AI intensity, emphasizing the need to focus on building a robust data foundation before heavily investing in AI.

This finding highlights a key insight: while AI intensity is vital for profitability, it is most beneficial for firms that already possess significant data intensity. Firms with higher AI capacity and efficient capital utilization can better manage data-driven productivity improvements, reinforcing the positive relationship between AI intensity and firm performance. Therefore, data-intensive firms are exceptionally well-positioned to capitalize on AI improvements, a theoretical prediction we empirically test in Section 3.

(3) Calibrate to 2005 and turn on AI Compute Using the 2005 Longitudinal Business Database on firm sizes and their market share (Helfand et al. (2007)), we run the following experiment: we choose eight levels of $\bar{A}$ that match the steady-state size of the firm with little AI ($z_i = 0.15$) to the average size of the firm in each of the nine firm-size categories for 2005. We then increase the AI parameter z_i to 0.5 and 1.2 for all firms in the economy.

Figure 10 shows the new firm size distribution (blue-line is the calibrated to 2005 data) has more large firms. The largest firms, that generate a lot raw data, now take advantage of the improved AI capability to get substantially larger.

Although it is intuitive that larger firms with higher AI capabilities would experience the most pronounced growth, a counteracting force exists. Bayes' law implies diminishing returns to AI: as AI capabilities increase, each additional unit contributes less to the precision of forecasts. This principle works against increasing returns to scale. Nonetheless, our findings suggest that the scale effect predominates. While precise quantitative conclusions would

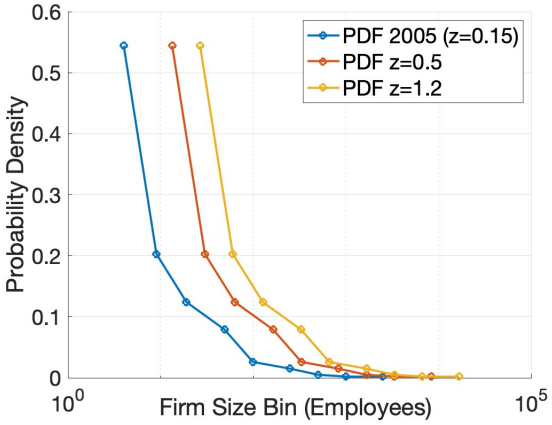


Figure 10: 2005 calibration and counterfactual $z = \{0.5, 1.2\}$; thicker right tail with higher compute.

Legend: Simulated firm size distributions under different AI regimes. The x-axis shows firm size, and the y-axis shows the probability density function. The figure compares the distributional effects of different information environments: raw data with no, low AI, and high AI. Higher compute (AI) makes large firms substantially larger. Parameters: As before. Differences from Farboodi et al. (2019): The curvature and cross-overs come from the AI-dependent processed-data mapping $\tilde{n}(n; z)$; without it, the same exercises lack the low-z drag and the differential tilt toward data-rich firms.

require more nuanced calibration, this exercise highlights how the model can be employed to measure the AI-driven transformation of firm size distributions and emphasizes the potential magnitude of AI's impact on firm size.

B Raw data trade

Let $\delta_{i,t}$ be the amount of raw data traded by firm i, after producing in date t. We assume that the firm that buys $\delta_{i,t} > 0$ units of data adds up the data it produced and the data it purchased, $n_0 k_{i,t}^{\alpha} + \delta_{i,t}$, to its stock of raw data. If firm i sells an amount $\delta_{i,t} > 0$ of raw data, given its partial non-rivalry, the firm keeps $n_{i,t} - \iota\delta_{i,t}$. This loss of raw data is a proxy for the loss of market power that comes from sharing one's own raw data. It can also represent a transaction cost of trading data, proportional to the data value.

In all simulations it is high AI-intense firms (high $z_i = z_H$) that sell raw data, and low AI-intense firms (low $z_i = z_L << z_H$) buying the raw data.

This volume of data can be transformed into usable data given the corresponding technology, for the firm that buys and the firm that sells, respectively is:

$$\tilde{n}_{i,t} = e^{-z_i}(-(n_0 k_{i,t}^{\alpha} + \delta_{i,t}) \ln(n_0 k_{i,t}^{\alpha} + \delta_{i,t})) + (1 - e^{-z_i})(n_0 k_{i,t}^{\alpha} + \delta_{i,t}) e^{n_0 k_{i,t}^{\alpha} + \delta_{i,t}} \tag{36}$$

$$\tilde{n}_{i,t} = e^{-z_i}(-(n_0 k_{i,t}^{\alpha} - \iota\delta_{i,t}) \ln(n_0 k_{i,t}^{\alpha} - \iota\delta_{i,t})) + (1 - e^{-z_i})(n_0 k_{i,t}^{\alpha} - \iota\delta_{i,t}) e^{n_0 k_{i,t}^{\alpha} - \iota\delta_{i,t}}. \tag{37}$$

Given the price of raw data π, and the quantity of raw data traded $\delta_{i,t}$, the objective function, for the firm that buys and the firm that sells, respectively, is:

$$V_0 = E_0 \sum_{t=0}^{\infty} \beta^t (A_{i,t} k_{i,t}^{\alpha} - r k_{i,t} - \pi \delta_{i,t}) \tag{38}$$

$$V_0 = E_0 \sum_{t=0}^{\infty} \beta^t (A_{i,t} k_{i,t}^{\alpha} - r k_{i,t} + \pi \delta_{i,t}). \tag{39}$$

The law of motion for knowledge is the same, where $\tilde{n}$ for the raw data buyer is given by (36), and $\tilde{n}$ for the raw data seller by (37):

$$\Omega_{i,t+1} = \left[\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \tilde{n}_{i,t} \sigma_\varepsilon^{-2} \tag{40}$$

For the firm that sells data, the Hamiltonian now becomes:

$$\begin{aligned} H(\Omega_{i,t}, k_{i,t}, p_{i,t}, \delta_{i,t}) = {} & [A - (\Omega_{i,t}^{-1} + \sigma_a^2)] k_{i,t}^{\alpha} - r k_{i,t} + \pi \delta_{i,t} \\ & + \beta p_{i,t+1} \left\{ [\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2]^{-1} + \tilde{n}_{i,t} \sigma_\varepsilon^{-2} - \Omega_{i,t} \right\} \end{aligned} \tag{41}$$

The solution to the raw data seller's Hamiltonian comes from the first-order conditions with respect to capital, to knowledge, and to data,

$$\frac{\partial H}{\partial k_{i,t}} = 0 \implies \alpha \left[\bar{A} - \left(\Omega_{i,t}^{-1} + \sigma_a^2\right)\right] k_i^{-(1-\alpha)} - r + \beta p_{i,t+1} \frac{d\tilde{n}_{i,t}}{dk_{i,t}} \sigma_\epsilon^{-2} = 0 \tag{42}$$

$$\beta p_{i,t+1} - p_{i,t} = -\frac{\partial H}{\partial \Omega_{i,t}} \implies \left[\rho + \frac{\sigma_\theta^2}{\rho(\Omega_{i,t} + \sigma_a^{-2})}\right]^{-2} \beta p_{i,t+1} = p_{i,t} - \Omega_{i,t}^{-2} k_{i,t}^{\alpha} \tag{43}$$

$$\frac{\partial H}{\partial \delta_{i,t}} = 0 \implies \pi + \beta p_{i,t+1} \frac{d\tilde{n}_{i,t}}{d\delta_{i,t}} \sigma_\epsilon^{-2} = 0 \tag{44}$$

The derivative of $\tilde{n}$ with respect to data is given by

$$\frac{d\tilde{n}_{i,t}}{d\delta_{i,t}} = \iota \left[e^{-z_i} \left(1 + \ln(n_0 k_{i,t}^\alpha - \iota\delta_{i,t})\right) - (1 - e^{-z_i}) \left(1 + n_0 k_{i,t}^\alpha - \iota\delta_{i,t}\right) e^{n_0 k_{i,t}^\alpha - \iota\delta_{i,t}} \right]. \tag{45}$$

The derivative of the processed data generation function with respect to capital is given by

$$\frac{d\tilde{n}_{i,t}}{dk_{i,t}} = -\alpha n_0 k_{i,t}^{-(1-\alpha)} \left[e^{-z_i} \left(1 + \ln(n_0 k_{i,t}^\alpha - \iota\delta_{i,t})\right) - (1 - e^{-z_i}) \left(1 + n_0 k_{i,t}^\alpha - \iota\delta_{i,t}\right) e^{n_0 k_{i,t}^\alpha - \iota\delta_{i,t}} \right] \tag{46}$$

Raw data seller: Substituting expressions above and simplifying yields for the firm that sells raw data, $i = H$

$$r k_{H,t}^{(1-\alpha)} = \alpha A_{H,t} - \alpha\beta p_{H,t+1} \sigma_\epsilon^{-2} n_0 e^{-z_i} \left[e^{-z_H} \left(1 + \ln(n_{H,t})\right) - (1 - e^{-z_H}) \left(1 + n_{H,t}\right) e^{n_{H,t}} \right] \tag{47}$$

where $n_{H,t} = n_0 * k_{H,t}^\alpha - \iota\delta_{H,t}$ and $A_{H,t} = \left[\bar{A} - \left(\Omega_{H,t}^{-1} + \sigma_a^2\right)\right]$ and

$$\pi = \beta p_{H,t+1} \iota e^{-z_H} \left[\left(e^{z_H} - 1\right) \left(1 + n_{h,t}\right) e^{n_{h,t}} - \left(1 + \ln(n_{h,t})\right) \right] \sigma_\epsilon^{-2} \tag{48}$$

$$\text{and } p_{h,t} = \Omega_{h,t}^{-2} k_{h,t}^\alpha + \beta \cdot p_{h,t+1} \left[\rho + \sigma_\Theta^2 \left(\Omega_{h,t} + \sigma_a^{-2}\right) / \rho \right]^{-2} \tag{49}$$

The law of motion for the high AI raw data seller is:

$$\Omega_{H,t+1} = \left[\rho^2 (\Omega_{H,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2 \right]^{-1} + \tilde{n}_{H,t} \sigma_\varepsilon^{-2} \tag{50}$$

with

$$\tilde{n}_{H,t} = e^{-z_H} (-(n_0 k_{H,t}^\alpha - \iota\delta_{H,t}) \ln(n_0 k_{H,t}^\alpha - \iota\delta_{H,t})) + (1 - e^{-z_H})(n_0 k_{H,t}^\alpha - \iota\delta_{H,t}) e^{n_0 k_{H,t}^\alpha - \iota\delta_{H,t}}. \tag{51}$$

For the firm that buys raw data, the Hamiltonian now becomes:

$$H(\Omega_{i,t}, k_{i,t}, p_{i,t}, \delta_{i,t}) = [A - (\Omega_{i,t}^{-1} + \sigma_a^2)]k_{i,t}^{\alpha} - rk_{i,t} - \pi\delta_{i,t} \\ + \beta p_{i,t+1}\left\{[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2]^{-1} + \tilde{n}_{i,t}\sigma_\varepsilon^{-2} - \Omega_{i,t}\right\} \tag{52}$$

Raw data buyer: The solution to the raw data buyer's Hamiltonian comes from the first-order conditions with respect to capital, to knowledge, and to data, which after simplifying, for the firm that buys raw data, $i = L$, yields:

$$rk_{L,t}^{(1-\alpha)} = \alpha A_{L,t} - \alpha\beta p_{L,t+1}\sigma_\epsilon^{-2} n_0 e^{-z_L}\left[e^{-z_L}\left(1 + \ln(n_{L,t})\right) - \left(1 - e^{-z_L}\right)\left(1 + n_{L,t}\right)e^{n_{L,t}}\right] \tag{53}$$

where $n_{L,t} = n_0 * k_{L,t}^{\alpha} + \delta_{L,t}$ and $A_{L,t} = \left[\bar{A} - \left(\Omega_{L,t}^{-1} + \sigma_a^2\right)\right]$ and

$$\pi + \beta p_{L,t+1}\sigma_\epsilon^{-2}\left[e^{-z_L}\left(1 + \ln(n_{L,t})\right) - \left(1 - e^{-z_L}\right)\left(1 + n_{L,t}\right)e^{n_{L,t}}\right] = 0 \tag{54}$$

$$\text{and } p_{L,t} = \Omega_{L,t}^{-2}k_{L,t}^{\alpha} + \beta \cdot p_{L,t+1}\left[\rho + \sigma_\Theta^2\left(\Omega_{L,t} + \sigma_a^{-2}\right)/\rho\right]^{-2} \tag{55}$$

The law of motion for the high AI raw data seller is:

$$\Omega_{L,t+1} = \left[\rho^2(\Omega_{L,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \tilde{n}_{L,t}\sigma_\varepsilon^{-2} \tag{56}$$

with

$$\tilde{n}_{L,t} = e^{-z_L}(-(n_0 k_{L,t}^{\alpha} + \delta_{L,t})\ln(n_0 k_{L,t}^{\alpha} + \delta_{L,t})) + (1 - e^{-z_L})(n_0 k_{L,t}^{\alpha} + \delta_{L,t})e^{n_0 k_{L,t}^{\alpha} + \delta_{L,t}} \tag{57}$$

The last condition is that the market for raw data trade clears. Assuming a proportion m_H of sellers, and $m_L = (1 - m_H)$ of buyers of raw data, the raw data market clearing condition is:

$$m_H * \delta_{H,t} = m_L * \delta_{L,t} \tag{58}$$

Equilibrium equations The system below with 19 endogenous variables and 19 equations defines the equilibrium:

$$\Sigma_L, \Omega_L, A_L, \tilde{n}_L, n_L, k_L, \Pi_L, \text{data}_L, p_L, \Sigma_H, \Omega_H, A_H, \tilde{n}_H, n_H, k_H, \Pi_H, \text{data}_H, p_H, \text{ and } p_{\text{data}}(\pi).$$

Parameters include: $n_0 = 0.35$, $\alpha = 0.25$, $\rho = 0.95$, $\beta = 0.95$, $m_H = 0.3$, $\sigma_A^2 = 0.25$, $\sigma_T^2 = 0.25$, $\sigma_\epsilon^2 = 0.35$, $\bar{A} = 10$, $r = 0.05$, $\iota = 0.6$, $z_L = 0.2$, and $z_H = 1.5$.

For the L firm buying data:

$$n_L = n_0 \cdot k_L^\alpha + \text{data}_L \tag{1}$$

$$\tilde{n}_L = e^{-z_L}\left(-n_L \ln n_L\right) + \left(1 - e^{-z_L}\right) n_L e^{n_L} \tag{2}$$

$$\Sigma_L^{(+1)} = \left[\left(\rho^2\left(\Sigma_L^{-1} - \frac{1}{\sigma_A^2}\right)^{-1} + \sigma_T^2\right)^{-1} + \tilde{n}_L \cdot \frac{1}{\sigma_\epsilon^2}\right]^{-1} \tag{3}$$

$$\Omega_L = \Sigma_L^{-1} \tag{4}$$

$$A_L = \bar{A} - \Sigma_L - \sigma_A^2 \tag{5}$$

$$\pi_L = A_L \cdot k_L^\alpha - r \cdot k_L - p_{\text{data}} \cdot \text{data}_L \tag{6}$$

$$0 = -\beta \cdot p^{(+1)} \cdot \frac{1}{\sigma_\epsilon^2}\left[e^{-z_L}(1 + \ln n_L) - (1 - e^{-z_L})(1 + n_L)e^{n_L}\right] - p_{\text{data}} \tag{7}$$

$$r \cdot k_L^{1-\alpha} = \alpha A_L - \beta \cdot p^{(+1)} \cdot \frac{\alpha n_0}{\sigma_\epsilon^2}\left[e^{-z_L}(1 + \ln n_L) - (1 - e^{-z_L})(1 + n_L)e^{n_L}\right] \tag{8}$$

$$p_L = \Sigma_L^2 \cdot k_L^\alpha + \beta \cdot p_L^{(+1)}\left[\rho + \frac{\sigma_T^2\left(\Sigma_L^{-1} + \frac{1}{\sigma_A^2}\right)}{\rho}\right] \tag{9}$$

For the H firm selling data:

$$n_H = n_0 \cdot k_H^\alpha - \iota \cdot \text{data}_H \tag{10}$$

$$\tilde{n}_H = e^{-z_H}\left(-n_H \ln n_H\right) + \left(1 - e^{-z_H}\right) n_H e^{n_H} \tag{11}$$

$$\Sigma_H^{(+1)} = \left[\left(\rho^2\left(\Sigma_H^{-1} - \frac{1}{\sigma_A^2}\right)^{-1} + \sigma_T^2\right)^{-1} + \tilde{n}_H \cdot \frac{1}{\sigma_\epsilon^2}\right]^{-1} \tag{12}$$

$$\Omega_H = \Sigma_H^{-1} \tag{13}$$

$$A_H = \bar{A} - \Sigma_H - \sigma_A^2 \tag{14}$$

$$\pi_H = A_H \cdot k_H^{\alpha} - r \cdot k_H + p_{\text{data}} \cdot \text{data}_H \tag{15}$$

$$0 = \iota \cdot \beta \cdot p^{(+1)} \cdot \frac{1}{\sigma_\epsilon^2} \left[e^{-z_H}(1 + \ln n_H) - (1 - e^{-z_H})(1 + n_H)e^{n_H} \right] + p_{\text{data}} \tag{16}$$

$$r \cdot k_H^{1-\alpha} = \alpha A_H - \beta \cdot p^{(+1)} \cdot \frac{\alpha n_0}{\sigma_\epsilon^2} \left[e^{-z_H}(1 + \ln n_H) - (1 - e^{-z_H})(1 + n_H)e^{n_H} \right] \tag{17}$$

$$p_H = \Sigma_H^2 \cdot k_H^{\alpha} + \beta \cdot p_H^{(+1)} \left[\rho + \frac{\sigma_T^2 \left(\Sigma_H^{-1} + \frac{1}{\sigma_A^2} \right)}{\rho} \right] \tag{18}$$

And the market clearing for data condition is:

$$m_H \cdot \text{data}_H = (1 - m_H) \cdot \text{data}_L \tag{19}$$

C Processed data trading

In the case of processed data, a firm that buys processed data adds $\delta_{i,t}$ to its own processed data stock, while a firm that sells processed data loses a fraction $\iota\delta_{i,t}$ of its processed data. Therefore, the after-trade processed data stock, for the firm that buys and the firm that sells, respectively is:

$$\tilde{n}_{i,t} = e^{-z_i}(-(n_0 k_{i,t}^{\alpha}) \ln(n_0 k_{i,t}^{\alpha})) + (1 - e^{-z_i})(n_0 k_{i,t}^{\alpha}) e^{n_0 k_{i,t}^{\alpha}} + \delta_{i,t} \tag{20}$$

$$\tilde{n}_{i,t} = e^{-z_i}(-(n_0 k_{i,t}^{\alpha}) \ln(n_0 k_{i,t}^{\alpha})) + (1 - e^{-z_i})(n_0 k_{i,t}^{\alpha}) e^{n_0 k_{i,t}^{\alpha}} - \iota\delta_{i,t}. \tag{21}$$

The objective function, law of motion for knowledge, and the Hamiltonian functions remain the same as in the setting with raw data trade. Only the solution to the Hamiltonian problem (i.e., the first-order conditions for capital and data) changes.

For the firm that sells data, the solution to the first-order conditions with respect to capital, data, and knowledge become:

$$r k_{i,t}^{(1-\alpha)} = \alpha A_{i,t} - \alpha\beta p_{i,t+1} \sigma_\epsilon^{-2} n_0 e^{-z_i} \left[e^{-z_i} \left(1 + \ln(n_{i,t})\right) - \left(1 - e^{-z_i}\right) \left(1 + n_{i,t}\right) e^{n_{i,t}} \right] \tag{22}$$

where $n_{i,t} = n_0 * k_{i,t}^{\alpha}$ and $A_{i,t} = \left[\bar{A} - \left(\Omega_{i,t}^{-1} + \sigma_a^2\right)\right]$ and

$$\pi = \beta p_{i,t+1} \iota \sigma_{\epsilon}^{-2} \tag{23}$$

$$\text{and } p_{i,t} = \Omega_{i,t}^{-2} k_{i,t}^{\alpha} + \beta \cdot p_{i,t+1} \left[\rho + \sigma_{\Theta}^2 \left(\Omega_{i,t} + \sigma_a^{-2}\right)/\rho\right]^{-2}. \tag{24}$$

For the firm that buys data, the solution to the first-order conditions with respect to capital, data, and knowledge become:

$$r k_{i,t}^{(1-\alpha)} = \alpha A_{i,t} - \alpha\beta p_{i,t+1} \sigma_{\epsilon}^{-2} n_0 e^{-z_i} \left[e^{-z_i}\left(1 + \ln(n_{i,t})\right) - \left(1 - e^{-z_i}\right)\left(1 + n_{i,t}\right) e^{n_{i,t}}\right] \tag{25}$$

where $n_{i,t} = n_0 * k_{i,t}^{\alpha}$ and $A_{i,t} = \left[\bar{A} - \left(\Omega_{i,t}^{-1} + \sigma_a^2\right)\right]$ and

$$\pi + \beta p_{i,t+1} \sigma_{\epsilon}^{-2} = 0 \tag{26}$$

$$\text{and } p_{i,t} = \Omega_{i,t}^{-2} k_{i,t}^{\alpha} + \beta \cdot p_{i,t+1} \left[\rho + \sigma_{\Theta}^2 \left(\Omega_{i,t} + \sigma_a^{-2}\right)/\rho\right]^{-2}. \tag{27}$$

The last condition again is that the market for processed data trade clears. Assuming a proportion $m_{Sellers}$ of sellers, and $m_{Buyers} = (1 - m_{Sellers})$ of buyers of processed data, the processed data market clearing condition is:

$$m_{Sellers} * \delta_{seller,t} = m_{Buyers} * \delta_{buyer,t} \tag{28}$$

Equilibrium equations The system below with 19 endogenous variables and 19 equations defines the equilibrium:

$$\Sigma_L, \Omega_L, A_L, \tilde{n}_L, n_L, k_L, \Pi_L, \text{data}_L, p_L, \Sigma_H, \Omega_H, A_H, \tilde{n}_H, n_H, k_H, \Pi_H, \text{data}_H, p_H, \text{ and } p_{\text{data}}(\pi).$$

Parameters include: $n_0 = 0.35$, $\alpha = 0.25$, $\rho = 0.95$, $\beta = 0.95$, $m_H = 0.3$, $\sigma_A^2 = 0.25$, $\sigma_T^2 = 0.25$, $\sigma_{\epsilon}^2 = 0.35$, $\bar{A} = 10$, $r = 0.05$, $\iota = 0.6$, $z_L = 0.2$, and $z_H = 1.5$.

For the firm buying processed data:

$$n_L = n_0 \cdot k_L^{\alpha} \tag{1}$$

$$\tilde{n}_L = e^{-z_L}\left(-n_L \ln n_L\right) + \left(1 - e^{-z_L}\right) n_L e^{n_L} + \text{data}_L \tag{2}$$

$$\Sigma_L^{(+1)} = \left[\left(\rho^2\left(\Sigma_L^{-1} - \frac{1}{\sigma_A^2}\right)^{-1} + \sigma_T^2\right)^{-1} + \tilde{n}_L \cdot \frac{1}{\sigma_\epsilon^2}\right]^{-1} \tag{3}$$

$$\Omega_L = \Sigma_L^{-1} \tag{4}$$

$$A_L = \bar{A} - \Sigma_L - \sigma_A^2 \tag{5}$$

$$\pi_L = A_L \cdot k_L^\alpha - r \cdot k_L - p_{\text{data}} \cdot \text{data}_L \tag{6}$$

$$p_{\text{data}} + \beta \cdot p^{(+1)} \cdot \frac{1}{\sigma_\epsilon^2} = 0 \tag{7}$$

$$r \cdot k_L^{1-\alpha} = \alpha A_L - \beta \cdot p^{(+1)} \cdot \frac{\alpha n_0}{\sigma_\epsilon^2}\left[e^{-z_L}(1 + \ln n_L) - \left(1 - e^{-z_L}\right)(1 + n_L)e^{n_L}\right] \tag{8}$$

$$p_L = \Sigma_L^2 \cdot k_L^\alpha + \beta \cdot p_L^{(+1)}\left[\rho + \frac{\sigma_T^2\left(\Sigma_L^{-1} + \frac{1}{\sigma_A^2}\right)}{\rho}\right] \tag{9}$$

And for the firm selling processed data:

$$n_H = n_0 \cdot k_H^\alpha \tag{10}$$

$$\tilde{n}_H = e^{-z_H}\left(-n_H \ln n_H\right) + \left(1 - e^{-z_H}\right) n_H e^{n_H} - \iota \cdot \text{data}_H \tag{11}$$

$$\Sigma_H^{(+1)} = \left[\left(\rho^2\left(\Sigma_H^{-1} - \frac{1}{\sigma_A^2}\right)^{-1} + \sigma_T^2\right)^{-1} + \tilde{n}_H \cdot \frac{1}{\sigma_\epsilon^2}\right]^{-1} \tag{12}$$

$$\Omega_H = \Sigma_H^{-1} \tag{13}$$

$$A_H = \bar{A} - \Sigma_H - \sigma_A^2 \tag{14}$$

$$\Pi_H = A_H \cdot k_H^\alpha - r \cdot k_H + p_{\text{data}} \cdot \text{data}_H \tag{15}$$

$$p_{\text{data}} - \iota \cdot \beta \cdot p^{(+1)} \cdot \frac{1}{\sigma_\epsilon^2} = 0 \tag{16}$$

$$r \cdot k_H^{1-\alpha} = \alpha A_H - \beta \cdot p^{(+1)} \cdot \frac{\alpha n_0}{\sigma_\epsilon^2}\left[e^{-z_H}(1 + \ln n_H) - \left(1 - e^{-z_H}\right)(1 + n_H)e^{n_H}\right] \tag{17}$$

$$p_H = \Sigma_H^2 \cdot k_H^\alpha + \beta \cdot p_H^{(+1)}\left[\rho + \frac{\sigma_T^2\left(\Sigma_H^{-1} + \frac{1}{\sigma_A^2}\right)}{\rho}\right] \tag{18}$$

The processed data market clearing condition:

$$m_H \cdot \text{data}_H = (1 - m_H) \cdot \text{data}_L \tag{19}$$

D Trade in Raw Data and Processed Data

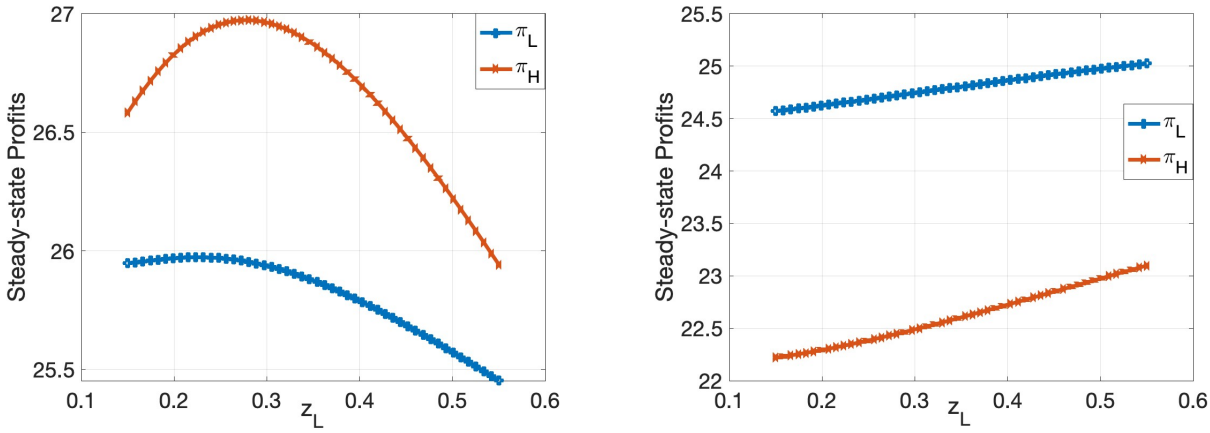
In an environment characterized by a penury of AI resources—such as a shortage of advanced computing chips—firms with high AI capabilities may find it challenging to further scale their data processing and analytical power, while competitors with low AI capabilities may struggle to catch up. This creates an opportunity for a market where firms can trade raw data or processed data with each other. Raw data trading allows firms with abundant data but limited AI capabilities to acquire additional data from other firms, potentially enhancing their own data intensity. Alternatively, trading processed data—already enriched by the advanced AI capabilities of high-tech firms—can enable low-AI firms to bypass the need for substantial AI investment while still benefiting from refined, value-added insights.

We aim to investigate whether trading in either raw or processed data can mitigate the disadvantage of low-AI firms and foster more inclusive growth by leveraging data-sharing synergies. This perspective resonates with Gans (2024), who argues that the operation of data markets—enabling cross-firm ata trade—is key for whether AI markets become concentrated or remain competitive. Understanding these mechanisms is crucial for designing policies and market structures that support equitable growth in an AI-constrained world.

Equilibrium and Implications To understand the importance of raw data trade versus processed data trade (i.e., knowledge distillation), we perform the following experiment. We compute profits in steady-state for a high-AI and a low-AI firm in an economy where raw data trade is possible, but it is subject to entropy, and separately, in an economy where processed data trade is also possible, but it does not suffer from entropy concerns. We assume that the high AI firm's computing capability stays fixed at a constant high level, and we vary the AI-capabilities of a low-AI firm. We recompute the equilibrium steady-state outcomes at each value of AI-intensity for the low-AI firm.

Figure 11: Profits for high-AI and low-AI firms with raw and processed data trade

(a) Profits with Raw Data Trade (b) Profits with Processed Data Trade



Legend: Simulated profits of high-AI and low-AI firms under two data market structures. The x-axis shows the AI intensity of low-AI firms, keeping the AI intensity of high-AI firms fixed, while the y-axis shows their steady-state profits. The left panel displays outcomes under a market for raw data; the right panel displays outcomes under a market for processed data. In the raw data regime, profits are increasing in AI, and high-AI firms maintain a persistent advantage. In contrast, under processed data trade, low-AI firms experience greater relative gains, and their profits may surpass those of high-AI firms when AI intensity is moderate, due to their improved ability to bypass informational entropy. Parameter values are: $\rho = 0.95$; $\sigma_a^2 = \sigma_\theta^2 = 0.25$; $\sigma_\epsilon^2 = 0.35$; $T = 10$; $\bar{P} = 1$; $\gamma = 0.7$; $\bar{A} = 10$; $m_H = 0.3$; $\iota = 0.6$; $r = 0.05$; $\alpha = 0.25$ $k_0 = 5$;. In all simulations, we keep $z_H = 4.5$ and we vary z_L from 0.15 to 0.55.

Figure 11 simulates the model and plots the steady-state profits of a high-AI firm and a low-AI firm for different values of AI in an economy with raw data trade (Panel (a)), and in an economy with processed data trade (Panel (b)). The figure underscores that facilitating trade in processed data, rather than raw data, can better support competitive market dynamics, allowing low-AI firms a chance to grow as well.

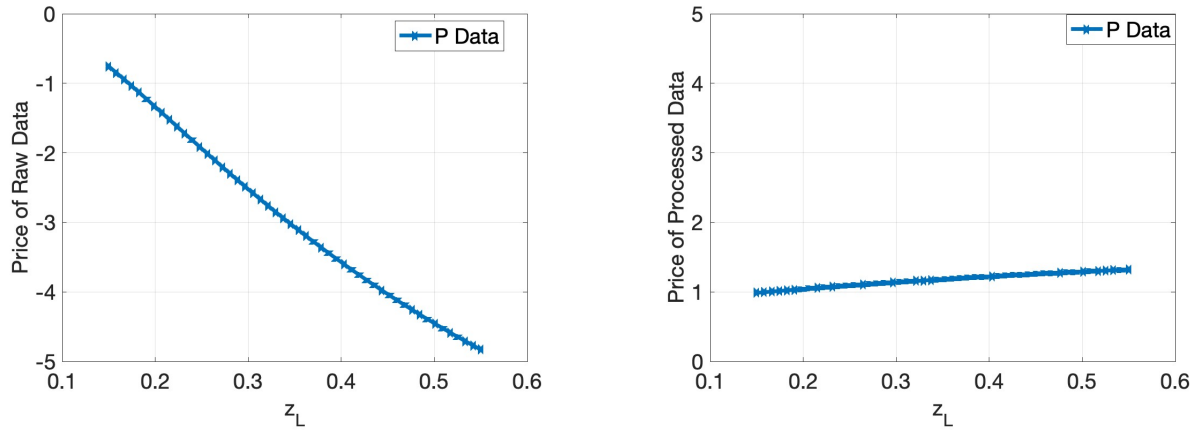
Panel (a) shows that the profits of a low-AI firm are always lower than the profits of a high-AI firm in a setting with raw data trade. At low levels of AI, there is actually a divergence in valuations between high-AI and low-AI firms as raw data is subject to entropy, meaning that accumulating raw data without sufficient AI capabilities leads to diminishing returns or even negative impacts on knowledge accumulation and profitability. The lowest-AI firms, in particular, suffer from this entropy effect as they cannot efficiently process or extract valuable information from the bulk of raw data. This explains why, in the raw data trade scenario, low-AI firms exhibit little to no growth or even a decline in market share

despite increased raw data availability. Their inability to process data effectively renders the additional raw data a burden rather than an asset.

Panel (b) shows that the processed data trade scenario significantly reduces or even eliminates the entropy problem. When data is processed or distilled, it becomes structured and of better quality (i.e. more usable) to even low-AI firms, which can avoid the computing constraint by using processed data as a substitute for computational power. This results in improved market share and profitability for low-AI firms that can even exceed that of high-AI firms, as they can leverage insights without the computational burden of processing raw data. In this scenario, AI-intensive firms can even lose their competitive advantage and be taken-over by low-AI firms that use distilled knowledge (i.e., processed data) to grow.

Figure 12: Price of data with raw and processed data trade

(a) Price of data with raw data trade (b) Price of data with processed data trade



Legend: Simulated data prices under two data market structures. The x-axis shows the AI intensity of low-AI firms, keeping the AI intensity of high-AI firms fixed, while the y-axis shows their steady-state profits. The left panel displays outcomes under a market for raw data; the right panel displays outcomes under a market for processed data. In the raw data regime, the price is negative and decreasing, suggesting missing markets for raw data or the fact that raw data is often shared for free. In the processed data regime, the price is strictly positive and increasing in the AI intensity of low-AI firms, suggesting processed data becomes even more valuable the higher the AI intensity of low AI firms is. Parameter values are: $\rho = 0.95$; $\sigma_a^2 = \sigma_\theta^2 = 0.25$; $\sigma_\epsilon^2 = 0.35$; $T = 10$; $\bar{P} = 1$; $\gamma = 0.7$; $\bar{A} = 10$; $m_H = 0.3$; $\iota = 0.6$; $r = 0.05$; $\alpha = 0.25$ $k_0 = 5$;. In all simulations, we keep $z_H = 4.5$ and we vary z_L from 0.15 to 0.55.

Figure 12 illustrates equilibrium data prices under two market regimes: raw data trade (left) and processed data trade (right). In the raw data market, prices are low and potentially volatile. This reflects the informational entropy effect: raw data, in its unstructured form,

has limited value unless paired with substantial processing capacity. Low-AI firms, lacking the computational ability to extract useful signals, place little value on raw data. The result is a thin market with flat or even declining prices. This dynamic mirrors real-world cases—such as free apps and social media platforms—where firms treat raw data as a zero-priced byproduct of user activity, rather than a priced asset.

By contrast, when processed data is traded, prices are higher and more stable. The transformation of raw inputs into structured, low-entropy information makes the data broadly valuable, including to low-AI firms that lack compute but can still leverage insights. In this setting, processed data functions as a productivity-enhancing input, and its price reflects both its usability and strategic value. Real-world analogs include subscription-based access to large foundation models, such as OpenAI's GPTPro, which sells access to structured outputs rather than raw training data.

A notable implication of the model is that low-AI firms may become more profitable than their high-AI counterparts, even when the latter produce and sell the processed data. This seemingly paradoxical outcome arises because processed data acts as a substitute for compute: it allows low-AI firms to bypass entropy constraints and achieve effective learning without investing in AI infrastructure. In effect, high-AI firms enable less capable firms to compete more effectively—sometimes eroding their own strategic advantage. The processed data market thus becomes a vehicle for redistribution, narrowing the performance gap between firms and reshaping competitive dynamics.

Our model assumes that processed data can be traded freely between firms, conditional on AI capabilities. However, in practice, many data intermediaries are platforms that face trade-offs when deciding whether to provide raw data or distilled insights. Smichowski et al. (2023) shows that platforms may avoid offering processed analytics if it cannibalizes high-value services or raises security risks. This adds a strategic and institutional layer to our model—processed data might be socially optimal, but withheld unless incentive-aligned. Our results should be interpreted in light of this potential friction.

E Alternative Mechanism

Firms can also acquire data—and the algorithms that process it—via mergers and takeovers rather than spot markets.

Set–up. Let firm i (the acquirer) contemplate acquiring target κ at time τ. The deal delivers a *lumpy* jump in usable signals and is (largely) exclusive:

$$\begin{aligned} \tilde{n}_{i,\tau^+} &= \tilde{n}_{i,\tau^-} \;+\; \phi\,\Delta_\kappa, \qquad \Delta_\kappa \equiv \tilde{n}(n_\kappa; z_\kappa)\,, \\ \tilde{n}_{\kappa,\tau^+} &= (1-\upsilon)\,\tilde{n}_{\kappa,\tau^-}, \qquad \upsilon \in [0,1], \end{aligned} \tag{20}$$

where $\phi \in (0,1]$ captures transfer efficiency/overlap (data cleaning, schema alignment) and υ captures exclusivity/foreclosure (with $\upsilon = 1$ a full data silo). The acquisition incurs a fixed cost $F > 0$ and an integration lag $L \geq 0$ during which the jump is not productive.[12]

Mapping to the baseline. In the flow-trade formulation in Section 2.2, small purchases add δ_t units *outside* Eq. (2). M&A is the *impulse* analogue: a single jump Δ_κ at time τ with exclusivity ($\iota \to 1$) and a fixed cost F.

Value and acquisition rule. Let p_t denote the shadow value of a marginal unit of processed information (the co-state on $\tilde{n}$, as in the baseline Appendix). For a small jump, the value gain is locally

$$\Delta V_i \;\approx\; e^{-rL}\,\phi\,\Delta_\kappa\,p_\tau \;-\; F,$$

so a sufficient condition for acquisition is

$$e^{-rL}\,\phi\,\Delta_\kappa\,p_\tau \;\geq\; F \tag{21}$$

(with equality defining the marginal deal). For finite Δ_κ, the exact gain integrates the policy path after the jump; (21) is the envelope condition that guides screening.

[12] Think of L as the time needed to integrate pipelines, legal permissions, and teams; ϕ subsumes de-duplication and overlap with the acquirer's stock.

Comparative statics.

- *Selection by AI capability.* Under our processed–data substitutability (Appendix A), p_τ is *decreasing* in z; hence, holding $(F, \phi, L, \Delta_\kappa)$ fixed, (21) is easier to satisfy for low-z acquirers. If F falls sufficiently with z (e.g., integration is cheaper for high-z firms), selection can flip.

- *Exclusivity and foreclosure.* Higher υ raises private gains (more exclusive access) but lowers rivals' $\tilde{n}$ (if the target exits or siloes data), shifting industry concentration.

- *Welfare.* A planner comparing trade vs. M&A weighs F and L against (i) duplicative fixed costs avoided by licensing and (ii) the foreclosure externality via υ.

Testable implications (for future work). This pasimonious extension delivers several testable predictions: (i) Deal probability is higher when $e^{-rL}\phi\Delta_\kappa p_\tau/F$ is high, predicting more data-motivated M&A where acquirers are compute-constrained, but integration is feasible. (ii) Post-deal, acquirer outcomes mimic our processed-trade experiment (profits rise more for low-z acquirers); rival performance depends on υ (larger drops when data is siloed). (iii) Licensing (processed trade) and M&A are substitutes when F is high or L is long; complements when exclusivity (υ) is pivotal.

These predictions are not separately identified in our current empirical analysis; we state them to guide future work that can exploit deal-level data or sectoral case studies.

Numerical simulation guidelines. Numerically, we could implement M&A as a dated impulse: at $t = \tau$ set $\tilde{n}_i \leftarrow \tilde{n}_i + \phi\Delta_\kappa$ and, if modeling foreclosure, $\tilde{n}_\kappa \leftarrow (1-\upsilon)\tilde{n}_\kappa$; then discount the jump by e^{-rL} in the objective or hold the jump latent for L periods. Report (ϕ, υ, L, F) alongside processed-trade parameters for transparency.

F 2005 Longitudinal Business Database

Table 3: Firm Size Bins, Average Firm Size, and Market Shares for 2005

Firm Size Bin (Employees)	Average Firm Size	Market Share (2005) (%)
1 to 4	2.10	54.40000
5 to 9	6.50	20.30000
10 to 19	13.50	12.40000
20 to 49	29.00	7.90000
50 to 99	68.00	2.60000
100 to 249	150.00	1.50000
250 to 499	350.00	0.50000
500 to 999	700.00	0.20000
1,000 to 2,499	1,500.00	0.20000

G Measure of AI intensity: Details

We construct a continuous measure of firms' AI computational capacity (*AI intensity*) that quantifies US public companies' high-performance computing (HPC) capabilities for AI development and deployment. Due to limited direct data on HPC facilities, we use patent information as a proxy for these capabilities. The idea behind using patents data is to capture firms' investment in proprietary AI hardware and specialized computing technologies. To construct this measure we combine data from two sources. The KPSS patent dataset built in Kogan et al. (2017) and annually updated on patents issued to the US public firms. We link each patent in KPSS dataset to USPTO AI patent dataset updated in Pairolero et al. (2025).

Our AI computational capacity score combines three key dimensions:

1. *AI-Hardware Patent Stock Ratio (PSR)*: The proportion of a firm's cumulative AI hardware patents to its total patent stock since 1990. We identify AI hardware patents using the AI Patent Dataset (AIPD) from Pairolero et al. (2025). We classify patents as AI hardware if they receive a probability score of being AI hardware of at least 90

percent in AIPD. This high threshold ensures we capture patents specifically focused on hardware architectures for AI applications.

2. *Compute-Specific Concentration (CSC)*: The ratio of compute-specific CPC-coded patents to total patent stock since 1990. We identify compute-specific patents using a targeted set of Cooperative Patent Classification (CPC) codes that precisely capture specialized computing hardware (see Table 4 below for the complete list of codes). The idea behind looking at the compute-specific patents is to go beyond broad AI hardware that may include general-purpose processing components. Even though those other AI hardware patents identified under PSR form an important part of AI computing, we want to make sure that we capture innovations specifically designed for parallel processing, specialized neural network accelerators, and high-throughput computation architectures.

3. *Absolute Capacity Indicator (ACI)*: The raw count of compute-specific CPC-coded patents accumulated since 1990, using the same specific CPC codes as in the CSC measure.

The idea behind considering both relative and absolute measures is to capture both specialized AI computing firms with high concentration ratios and large diversified technology firms with significant absolute patent counts. This balanced approach identifies firms with meaningful AI compute capabilities across the size spectrum.

For each component, we calculate its percentile rank across the full panel sample. This conversion to percentiles ensures each measure is comparable on the same scale. The final AI computational capacity score (*AI intensity*) for firm i at time t is:

$$\text{AI intensity}_{i,t} = 0.4 \times \text{Percentile}(\text{PSR}_{i,t}) + 0.3 \times \text{Percentile}(\text{CSC}_{i,t}) + 0.3 \times \text{Percentile}(\text{ACI}_{i,t})$$

This weighted approach balances relative concentration with absolute capacity, avoiding bias toward either small specialized firms or large generalists. By using percentile ranks

rather than raw values, each component contributes on a standardized 0-1 scale, with higher values indicating greater AI computational capacity.

We identify compute-specific patents using a targeted set of Cooperative Patent Classification (CPC) codes that precisely capture specialized computing hardware. The list in Table 4 represents a manually selected subset from a broader taxonomy of AI hardware-related CPCs compiled by the United States Patent and Trademark Office (USPTO) in their study Pairolero et al. (2025).

Table 4: CPC Codes Related to Specialized Compute Hardware

CPC Code	**Description**
G06F9/00	Arrangements for program control, e.g. control units (program control for peripheral devices G06F 13/10)
G06T1/20	[*Within* General purpose image data processing] Processor architectures; Processor configuration, e.g. pipelining
H04N19/42	[*Within* Methods or arrangements for coding, decoding, compressing or decompressing digital video signals] characterised by implementation details or hardware specially adapted for video compression or decompression, e.g. dedicated software implementation (H04N 19/635 takes precedence)
H04N19/43	[*Within* Methods or arrangements for coding, decoding, compressing or decompressing digital video signals] Hardware specially adapted for motion estimation or compensation
G06T1/60	[*Within* General purpose image data processing] Memory management
G06N3/06	[*Within* Computing arrangements based on biological models] Physical realisation, i.e. hardware implementation of neural networks, neurons or parts of neurons
G06N7/04	[*Within* Computing arrangements based on specific mathematical models] Physical realization

H Validating AI Intensity and Data Intensity Measures

We validate our AI intensity and data intensity measures by comparing them to independent external proxies. This ensures the constructs we propose have economic relevance.

Our AI intensity measure captures firms' AI compute capacity using compute-hardware patent stocks. These patents reflect firms' technological capability to deploy AI at scale. This contrasts with Babina et al. (2024), who measure AI intensity through employment shares of AI-skilled workers using Cognism resume data. The approach by Babina et al. (2024) explicitly measures human capital, whereas ours emphasizes technological capacity. Both approaches thus capture overlapping yet distinct dimensions of AI deployment.

To validate our AI intensity, we classify firms into high (top 10%) and low AI compute intensity groups based on our AI compute intensity measure. Firms in the of high compute AI group exhibit AI-skilled employment shares approximately 0.14 percentage points higher than firms in the remaining 90%. Given the low baseline share of 0.05 percent, this increase is economically substantial, roughly three times higher (Figure 13 left panel). This gap remains robust after controlling for industry-year fixed effects.[13] The pattern fits the complementary inputs view: compute capacity and AI labor move together but capture different inputs to the same production process (Veldkamp, 2023; Abis and Veldkamp, 2024).[14]

Our data intensity measure captures firms' operational reliance on data, determined by textual similarity in 10-K filings to data-intensive business models and supported by general AI application patents. To validate this, we use external data from SWZD, a detailed firm-level IT spending dataset aggregated and linked with Compustat through firm name matching.[15] We compute firm-level averages by summing expenditures on storage, servers,

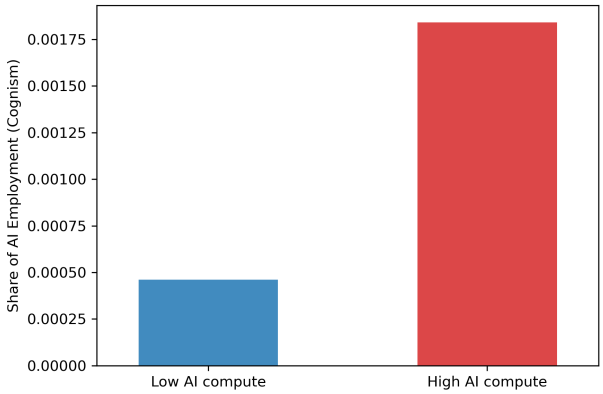
[13] Detailed regression results are available upon request.

[14] Another reason for observing a lower correlation could be that AI-labor shares also depend on local talent conditions; tight markets can lower observed AI-labor shares through higher departures and slower hiring, even when compute capacity is high (Chen et al., 2023).

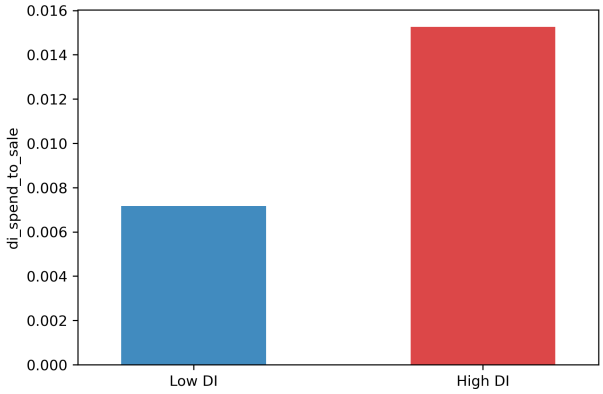
[15] Several studies have utilized this dataset for similar purposes (See Bloom et al. (2014); Zhang (2019); Tuzel and Zhang (2021)). SWZD data was historically collected through firm-level surveys before transi-

Figure 13: Validation of AI Compute and Data Intensity Measures

AI Compute Intensity and AI Employment Share



Data Intensity and Data-related Spending



Notes: Left panel plots the relationship between our AI compute intensity measure and AI employment shares from Babina et al. (2024). We construct high and low AI compute intensity groups by classifying firms in the top 10% of our AI compute intensity measure as "high" and the remaining firms as "low". AI-skilled workers data constructed by Babina et al. (2024) uses Cognism resume data and is available for the period 2007 and 2018. Right panel validates our data intensity measure against external data-related spending from SWZD, a detailed firm-level IT spending dataset aggregated and linked with Compustat through firm name matching. We compute firm-level data spending as the sum of expenditures on storage, servers, and hardware normalized by sales, averaged across 2010-2014. Firms are classified into high and low data intensity groups based on our measure, and we plot the average data spending ratios for each group. The SWZD sample is restricted to 2010-2014 when data collection relied on firm-level surveys rather than model-based estimates. For detailed discussion of the AI and data intensity measures, see Section 3.2.

and hardware normalized by sales, averaging across the years 2010 to 2014. Firms are classified into high and low data intensity groups based on our measure, and we analyze differences in average data spending ratios between these two groups.

Firms classified as high data intensity spend approximately 0.81 percentage points more on data infrastructure relative to sales than low data-intensity firms (Figure 13 right panel). This economically meaningful difference represents about a 113 percent increase over the baseline of low-intensity firms. Statistical precision is limited, however, due to sparse and uneven coverage in the SWZD data. Despite these constraints, the positive difference supports our measure's ability to identify firms investing more substantially in data infrastructure.

tioning to model-based estimates. To maintain accuracy, we restrict our analysis to years 2010–2014.

I Data Load, AI Capability, and Signal Quality

We seek empirical support for equation (2), which links processed signal quality to raw data volume and AI capability. By useful signals we mean information that raises the precision of decisions and selection. Higher precision allows firms to better identify and prioritize the most economically valuable ideas and innovations. Two observables help us capture this: a market-based innovation efficiency measure and a firm-level productivity measure.

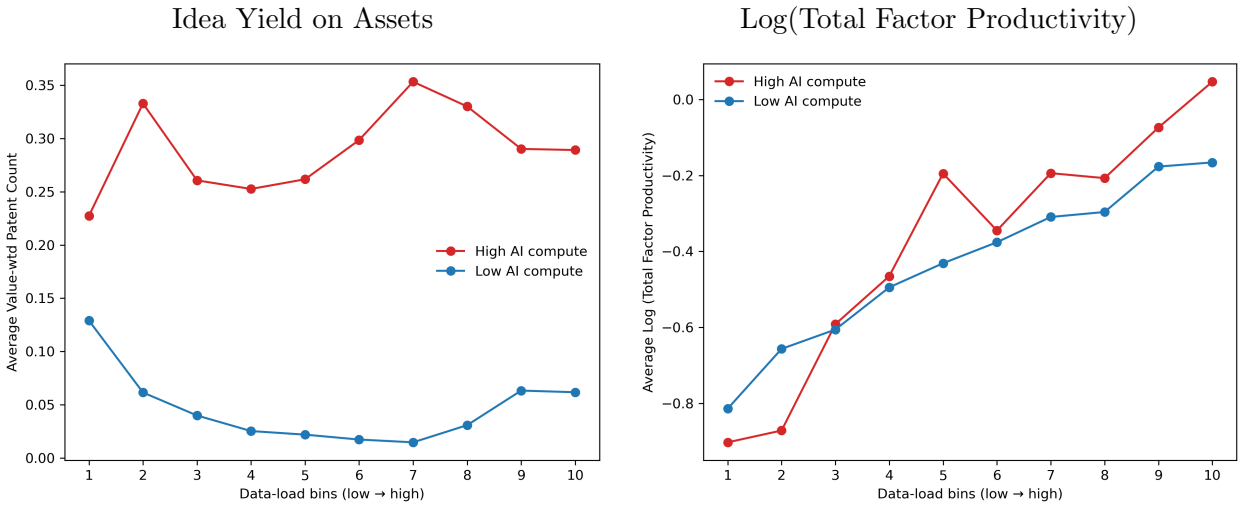
Our first measure is Idea Yield on Assets, defined as the total market value generated by all patents filed by a firm within a year (based on event-study returns at the time of the grant of the patent), divided by the firm's lagged total assets. This metric captures the market-assessed economic value generated by patented innovations relative to the firm's productive capital base. It rises when firms better convert raw data into valuable, patentable ideas using their existing assets. Firms with better processing capabilities (high AI) should more effectively filter, prioritize, and execute projects, resulting in higher idea yields. This measure isolates variation driven primarily by differences in information quality and processing efficiency rather than firm size.

Our second measure is Total Factor Productivity (TFP). We obtain firm-level TFP estimates from İmrohoroğlu and Tüzel (2014). TFP provides a complementary measure of processed signals by capturing efficiency improvements across a firm's entire production process. Firms with high-quality processed signals can allocate resources better, optimize production, and innovate more effectively.

We focus on the pre-Transformer period (2006–2016). We measure raw data volume by combining a firm's data intensity and employment. Specifically, our data-load measure is $R_{it} = \text{DI}_{it} \times ln(\text{Employment}_{it})$. Data intensity reflects the extent of a firm's reliance on data, while the number of employees scales the volume of raw observations the firm can generate. Thus, their product proxies firm-specific raw data volume, $n_{i,t}$. We sort firms into deciles based on R_{it} across the entire sample. Independently, we classify firms into high or low AI capability based on their AI intensity in the pre-period (2005), and keep these groups fixed

throughout the analysis. Within each data-load decile and AI capability group, we calculate the average values of value per idea and log(TFP).

Figure 14: Data Load, AI Capability and Signal Quality



Notes: This figure plots the average of the outcomes for different groups formed by intersection of data load deciles and AI capability types. Data load measure is $R_{it} = \text{DI}_{it} \log(\text{Employment}_{it})$. Idea Yield on Assets is proxied by total nominal market value generated by all patents filed by a firm within a year, divided by the firm's total assets. The data on market value of patents is from Kogan et al. (2017). Firm-level TFP estimates from the updated data from İmrohoroğlu and Tüzel (2014).

Figure 14 summarizes our key findings. High-AI firms consistently exhibit high Idea Yield on Assets across data-load bins, indicating sustained value generation even at greater data volumes. Low-AI firms, by contrast, experience a modest initial decline in idea yield, which stabilizes at higher data loads—consistent with diminishing returns from limited processing capability. This divergence matches predictions from equation (2): high AI capability preserves innovation productivity, whereas low capability increasingly succumbs to information entropy. Similarly, TFP increases with higher data load for both groups, but high-AI firms achieve consistently higher TFP and experience faster growth across nearly all data levels, highlighting that stronger AI capability allows firms to extract greater productivity from additional data.

Additional evidence from Appendix J.2 reinforces these findings. In compute-insensitive industries, where effective AI capability is inherently low, the marginal returns to data intensity become negative after the AWS shock. This aligns directly with the entropy effect

in equation (2), confirming that without adequate processing capability, additional raw data can reduce firm performance.

J Placebo and Falsification Tests

A credible identification strategy requires verifying that the observed effects are not artifacts of pre-existing trends or unrelated shocks. We thus implement placebo tests designed so that estimated impacts vanish when applied either to periods without genuine technological events or to subpopulations unaffected by these shocks. These placebo tests replicate our baseline specifications exactly, varying only the timing or exposure while preserving treatment intensities, fixed effects, controls, and clustering.

J.1 Shock-Timing Placebos

Timing placebos specifically verify whether estimated effects disappear when fake event years are deliberately assigned to periods preceding the actual technological shocks. For AWS, which actually occurred in 2006, we conduct placebo analyses by setting artificial event years as either 2002 or 2003, restricting the analysis window between 2000 and 2005. Similarly, for the Transformer shock occurring in 2017, placebo tests assign fake event years of either 2012 or 2013 within the window from 2011 to 2017.

These placebo event years are selected in alignment with theoretical predictions related directly to the genuine technological shocks. AWS, which significantly reduced computational costs, theoretically predicts a positive slope for the Data Intensity $\times$ Post interaction without a definitive prediction for AI Intensity $\times$ Post. Conversely, the Transformer shock primarily lowered text-processing barriers, theoretically implying a negative slope for AI Intensity $\times$ Post but leaving ambiguous predictions for Data Intensity $\times$ Post.

The placebo test results are given in the Table 5. In the AWS placebo tests, assigning event years of 2002 or 2003 sharply reduces the baseline Data Intensity $\times$ Post effect from

significant values of 0.066 (assets) and 0.364 (patents) in baseline (Table 1, to economically negligible magnitudes around 0.01 for assets and roughly 0.13 for patents, each losing statistical significance.

Table 5: Shock-Timing Placebo Tests

	AWS Placebo				Transformer Placebo			
	Event Year: 2002		Event Year: 2003		Event Year: 2012		Event Year: 2013	
	Log (Assets) (1)	Cit-wtd patent count (2)	Log (Assets) (3)	Cit-wtd patent count (4)	Log (Assets) (5)	Cit-wtd patent count (6)	Log (Assets) (7)	Cit-wtd patent count (8)
AI intensity X Post	-0.0568	0.2144*	-0.0673	0.2652**	-0.0474	0.0466	-0.0689*	0.0755
	(0.0439)	(0.1208)	(0.0430)	(0.1137)	(0.0391)	(0.1420)	(0.0393)	(0.1487)
Data intensity X Post	0.0093	0.1209	0.0099	0.1366	-0.0512*	0.5011	-0.0424	0.5447
	(0.0264)	(0.1902)	(0.0257)	(0.1861)	(0.0302)	(0.3076)	(0.0303)	(0.3744)
Number of Observations	24019	10899	24019	10899	18230	7405	18230	7405
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Placebo tests examine alternative shock timings: AWS shock uses 2003 and 2004 as alternative shock years (sample: 2000-2005); Transformer shock uses 2012 and 2013 as alternative shock years (sample: 2011-2017). Cross-sectional intensities from 2005 (2004 if unavailable). For each placebo test, $Post_t$ dummies are redefined to equal 1 for periods after the respective alternative shock year. See Table 1 notes and Section 3.2 for more details on variables.

For the Transformer placebo tests, artificially shifting event years to either 2012 or 2013 similarly diminishes the strong negative AI Intensity $\times$ Post effects in baseline (2) of -0.242 (assets) and -0.513 (patents) to trivial, statistically insignificant values around ± 0.05. The Data Intensity slopes either become imprecise or switch signs, consistent with theoretical uncertainty.

The evaporation of baseline differential slopes under these placebo scenarios strengthen our causal interpretation, affirming the authenticity of the identified AWS and Transformer effects on growth and innovation.

J.2 Exposure Falsification

Exposure falsification tests verify that observed effects appear only in sectors theoretically sensitive to technological shocks. For the AWS falsification test, we identify industries historically unaffected by scalable computing due to their reliance on physical capital and embedded control systems. We classify firms in Agriculture (NAICS 11), Mining/O&G (21), Utilities (22), Construction (23), and select Manufacturing sectors (31–32) as compute "insensitive". If these sectors exhibit the baseline data complementarity effect, it would imply a spurious relationship.

To perform this test, we extend the baseline AWS regression by adding a triple interaction (Data Intensity × Post × Compute-Insensitive). The results are presented in Table 6. In sectors sensitive to AWS, the baseline Data Intensity × Post slope remains robust, even slightly increasing to 0.109 for assets and remaining rougly the same for the patents, compared to the original baseline of 0.066 and 0.364 respectively. However, the triple interaction coefficients are strongly negative (-0.233 for assets, -0.701 for patents), causing implied slopes in compute-insensitive industries to become economically negligible and statistically indistinguishable from zero. This precisely matches theoretical expectations.

Although designed to probe our theoretical mechanism specificity, the compute-insensitive placebo also maps to the information–entropy effect in equation (2). In sectors where scalable compute is unproductive, effective capability z_i is low and the entropy term $e^{-z_i}\,[-n_{i,t}ln(n_{i,t})]$ can dominate, so additional raw data $n_{i,t}$ yields weak or negative marginal returns. Consistent with this, the negative Data Intensity × Post × Compute-Insensitive coefficients in Table 6 are strongly negative (0.233 for assets, 0.701 for patents), turning the otherwise positive effect of data intensity negative in these compute-insensitive industries. This provides complementary evidence of an overload effect where having more data, without productive processing capability, reduces firm performance.

For the Transformer falsification, we isolate firms unlikely to benefit due to minimal exposure to unstructured text. We first categorize firms into B2C/B2B groups using a large-language-model parser on each firm's pre-2017 10-K "Business Description" using the idea that public firms in the USA are required to provide accurate and extensive information about their business in their annual 10-K filings (Hoberg and Phillips, 2016). We then create a "LowText" dummy equal to one only for pure B2B firms outside known text-heavy sectors, specifically excluding Information (NAICS 51), Finance and Insurance (52), Computer and Electronic Product Manufacturing (334), and certain professional and administrative services (NAICS 5415, 5416, 5418, 5614). All other firms, including non-B2B firms and those in text-heavy sectors, receive a zero on the Text-Light dummy. The regression includes a triple

interaction (AI Intensity × Post × Text-Light).

Within the text-heavy B2C baseline group, the negative AI Intensity × Post effect notably strengthens to -0.265 for assets and -0.606 for patents, compared to original baseline estimates of -0.242 and -0.513 respectively. However, adding the triple interaction coefficients (with *positive* triple interaction for both assets and patents) neutralizes this negative slope among text-light B2B firms, resulting in statistically insignificant effects around -0.20.

Hence, for both AWS and Transformer shocks, differential effects emerge precisely in sectors where theoretical mechanisms predict their presence and vanish in sectors where they should be absent. These rigorous falsification tests reinforce the causal interpretation of our baseline AWS and Transformer estimates.

Table 6: Shock-Exposure Placebo Tests

	AWS Falsification		Transformer Falsification	
	Log (Assets) (1)	Cit-wtd patent count (2)	Log (Assets) (3)	Cit-wtd patent count (4)
Is AI Comp Insensitive X Post	0.0009	0.0141		
	(0.0632)	(0.2836)		
Is LowText X Post			0.0580	0.8785
			(0.0552)	(0.5889)
AI intensity X Post	-0.2117***	0.1235	-0.2651***	-0.6062*
	(0.0553)	(0.1400)	(0.0617)	(0.3107)
Data intensity X Post	0.1094***	0.3619	0.1123**	1.5599**
	(0.0346)	(0.2427)	(0.0458)	(0.6517)
AI intensity X Post X Is AI Comp Insensitive	0.1637	0.4944		
	(0.1348)	(0.3017)		
Data intensity X Post X Is AI Comp Insensitive	-0.2325***	-0.7014*		
	(0.0781)	(0.4070)		
AI intensity X Post X Is LowText			0.0586	0.4036
			(0.1181)	(0.4099)
Data intensity X Post X Is LowText			-0.1648**	-1.4004*
			(0.0787)	(0.8350)
Number of Observations	58530	29293	48648	22896
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Controls for Post 5+ and its Interactions	Yes	Yes	Yes	Yes

Standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. See Table 1 notes and Section 3.2 for more details on variables.

K Robustness with Fixed Industry Classification (FIC) HHI

We construct the Herfindahl-Hirschman Index (HHI) using the Fixed Industry Classification (FIC) provided by the Hoberg–Phillips Data Library. The FIC classification is superior to traditional industry classifications such as NAICS and SIC, as it allows each firm to simultaneously belong to multiple industry categories based on textual product descriptions from firms' 10-K filings. In particular, it represents a significant improvement over traditional methods by capturing firms' actual product-market scope more accurately.

In addition, compared to Text-based Network Industry Classification (TNIC) developed by Hoberg and Phillips (2016) and used by us in the baseline, which dynamically updates each firm's competitor set annually based on textual similarity, the FIC provides a stable and fixed definition of industries. Specifically, while TNIC redefines industries each year based on pairwise textual cosine similarities between firms, FIC fixes the set of industries ex-ante. Consequently, FIC allows us to clearly attribute changes in market concentration to within-industry dynamics rather than changes in industry definitions or firm classifications.

The FIC dataset is structured at the firm-FIC Industry-year level, with each firm in a given yeat possibly associated with multiple industries. Importantly, the data also provides a textual loading score derived from the fraction of each industry's vocabulary that appears in each firm-year's product description.

To construct the FIC HHI at the industry-year level, we follow the recipe described in Hoberg and Phillips (2024), and proceed as follows:

(i) Using the textual loadings (L_{ijt}), we allocate each firm's Compustat sales ($sales_{it}$) proportionally across industries $sales_{ijt}^{alloc} = L_{ijt} \times sales_{it}$.

(ii) Total industry sales for industry j in year t are computed as $sales_{jt}^{ind} = \sum_i sales_{ijt}^{alloc}$

(iii) Each firm's market share within industry j in year t is: $s_{ijt} = \frac{sales_{ijt}^{alloc}}{sales_{jt}^{ind}}$

(iv) Industry-year level HHI is the sum of squared market shares within each industry-year: $HHI_{jt}^{ind} = \sum_i (s_{ijt})^2$

Calculation of Industry Intensities, outcomes and Controls: We calculate industry-level AI and Data intensities by taking weighted averages of firm-level intensities measured on the eve of each shock, using textual loadings as weights. Specifically, for each industry-year, we compute the weighted average of AI and Data intensities where the weights are given by the textual loadings, L_{ijt}. Similarly, for time-varying industry-level variables (outcome or control), we apply the same textual loading-weighted averaging procedure for ratios (such as ROA and Tobin's Q) and a textual loading-weighted sum for level variables (such as total assets, market capitalization, citation-weighted patent count). Our regression specification at the industry-year panel remains similar to Equation 4, employing a continuous DID design.

Results of industry regressions: Table 7 reports continuous DID estimates of industry outcomes around the AWS (2006) and Transformer (2017) shocks. Two results matter.

Scale and innovation. The directions match our firm-level evidence. After AWS, data-intensive industries expand assets and citation-weighted patents, while AI-intensive industries do not. The Transformer shock flips the pattern: low-AI industries close the patenting gap, and data-rich industries keep innovating. These parallels reinforce our theoretical predictions about compute complementing data and processed data substituting for compute.

Concentration. Once industry boundaries are fixed, market-power effects are muted. The FIC-based HHI is statistically unchanged after AWS and declines modestly following the Transformer shock. This suggests that the scale and innovation shifts we document occur through parallel expansion and entry rather than outright consolidation. The modest TNIC-HHI uptick in our baseline firm regressions therefore appears to capture firms re-scoping toward narrower product niches, not an across-the-board tightening of market power. In

short, compute and processed-data accessibility reshape who grows and innovates, while leaving within-industry concentration largely intact.

Table 7: Impact of AWS Shock and Transformer Shock on Industry Outcomes

	AWS Shock			Transformer Shock		
	Log (Assets) (1)	Cit-wtd patent count (2)	FIC HHI (3)	Log (Assets) (4)	Cit-wtd patent count (5)	FIC HHI (6)
AI intensity X Post	-1.3547***	-1.7480***	0.0130	-0.2925	-2.3840***	-0.0665
	(0.4089)	(0.3126)	(0.0635)	(0.2792)	(0.3741)	(0.0532)
Data intensity X Post	0.9113***	1.6934***	-0.0461	-0.4852***	1.8438***	-0.0793**
	(0.1700)	(0.1961)	(0.0291)	(0.1508)	(0.2921)	(0.0320)
AI intensity X Post_{5+}	-0.0555	-2.4881***	-0.0956			
	(0.4207)	(0.5323)	(0.1370)			
Data intensity X Post_{5+}	0.6903***	2.2092***	-0.0098			
	(0.1995)	(0.3235)	(0.0576)			
Number of Observations	5100	5100	5100	4500	4500	4500
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. For details on the construction of FIC Industry HHI see the discussion in Appendix K. See notes for Figure 4 and Section 3.2 for more details on the variables. The regression for citation-weighted patent count is estimated using Poisson Pseudo Likelihood Estimation.

L Summary Statistics

Table 8: Descriptive Statistics

	N Obs	Mean	Std Dev	P25	P50	P75	P90
I. Financial Variables							
Log (Total Assets)	103022	6.39	2.22	4.83	6.41	7.87	9.22
Log (Market Capitalization)	114744	6.01	2.14	4.50	5.92	7.42	8.80
Log (Total Sales)	99986	5.60	2.40	4.05	5.67	7.21	8.58
Log (R&D Expenditure)	42722	2.90	1.99	1.62	2.96	4.15	5.35
R&D Expenditure to Asset	103022	0.07	0.23	0.00	0.00	0.05	0.19
Return on Assets	99133	0.01	1.70	0.01	0.07	0.13	0.20
Leverage	102641	0.24	0.34	0.03	0.17	0.37	0.56
Tobin's Q	102828	2.57	52.62	1.04	1.37	2.19	3.80
Asset Tangibility	99308	0.20	0.24	0.03	0.10	0.29	0.60
II. Market Concentration							
HHI (TNIC3)	98033	0.26	0.27	0.07	0.15	0.35	0.70
III. Innovation Output							
Cit-wtd Patent Count	114903	14.80	84.51	0.00	0.00	0.00	11.07
Value-wtd Patent Count	103022	0.07	0.40	0.00	0.00	0.00	0.12

The descriptive statistics are calculated over the entire sample covering 2000–2023. N refers to the total number of firm-year observations. p25-p90 refer to the 25th to 90th percentile values. Citation-weighted (cit-wtd) patent counts weight each patent by its forward citations, adjusted for filing vintage. Value-weighted counts use market-based valuations capturing shareholder wealth effects at patent award from Kogan et al. (2017), scaled by assets. Tobin's Q is defined as the ratio of the market value of equity plus the book value of debt to total assets. Leverage is total debt—comprising long-term debt and debt due within one year—expressed as a share of total assets. Return on Assets is operating income before depreciation divided by total assets. Asset tangibility is the ratio of net property, plant, and equipment to total assets.