Firm-level Data Assets and Inorganic Growth

Abstract

We examine whether the data revolution promotes acquisition activities at the firm level. We develop firm-year data assets measures using a combination of economic modelling and a large language model. Since 2010, data assets have expanded and increasingly concentrated in large firms. Firms with greater data assets are more likely to be a bidder, with the effect significantly stronger among large firms. Evidence supports a defensive "eat or be eaten" motive: data-driven M&A activity is stronger in industries with a higher concentration of small- and medium-sized data-intensive firms and is often value-destroying. Data assets improve the information environment and reduce financing costs are also channels for data-asset-induced mergers. Our findings are robust to alternative proxies for data assets and remain consistent under instrumental variable estimations. Overall, the data era has intensified defensive and inorganic growth, accruing disproportionately to large firms.

Keywords: Data asset; Mergers; Acquisitions; Defensive acquisitions; Machine Learning; *FinBERT*
JEL Classification Codes: G34, O33

1. Introduction

The rise of the big data economy is reshaping competition, reallocating resources, and redefining firm boundaries. Yet we know little about how data assets affect corporate decisions, partly because data assets are absent from balance sheets and difficult to measure consistently (Abis and Veldkamp, 2024; Veldkamp, 2023). We address this gap by developing firm-level measures of data assets that combine an economic model with a large language model to provide evidence at the firm level. Specifically, we examine whether data assets influence acquisitions activities, a central corporate response to technological and regulatory shocks (Gort, 1969; Harford, 2005; Mitchell and Mulherin, 1996).

Major technological shifts alter firm boundaries and market structure, trigger large-scale reallocation, and induce creative destruction (Aghion and Howitt, 1992). The data economy constitutes a new economy-wide technological revolution that fits this regime-shift framework. Data and algorithms transform production, competition, and growth with economy-wide effects, with magnitudes comparable to the industrial revolution (Abis and Veldkamp, 2024; Farboodi and Veldkamp, 2021; Jones and Tonetti, 2020). Mergers are a key channel for rapid expansion in response to shocks. Early theory (e.g., Gort, 1969; Mitchell and Mulherin, 1996) predicts intensified merger activity following shocks, and empirical evidence confirms this pattern for regulatory, economic, and capital shocks (e.g., Harford, 2005). These insights imply that data assets, as fuel for the current technological shock, may drive another structural regime shift and prompt M&A. We therefore expect data assets to raise a firm's likelihood of becoming a bidder. In addition, the literature predicts a size effect: large firms accumulate more data, extract more value from it, and expand further (Veldkamp, 2023). Small firms, by contrast, struggle to keep pace with the required investments (Lashkari et al., 2024), creating an inherent tendency toward scale and concentration. Begenau et al. (2018) show that big data in finance disproportionately lowers the cost of capital for large firms, enabling greater investment and faster growth relative to small firms. Abis and Veldkamp (2024) find that the returns to big-data investment are stronger for larger firms, further increasing industry concentration. Moreover, facing constraints on organic growth in day-to-day operations (Akcigit and Kerr, 2018; Babina et al., 2024), large firms may rely on M&A to expand inorganically. We therefore expect data-asset-induced M&A activity to be concentrated among large firms.

A firm's data assets are digital information the firm owns or can lawfully access and analyze to inform capital allocation, operations, and forecasting (Goldfarb and Tucker, 2019; Veldkamp, 2023; Veldkamp and Chung, 2019). Despite economic importance, measurement of data assets is challenging because data are intangible and nonrival; quality is heterogeneous; units and prices are not standardized; and standard accounting practices are silent (Bhutani et al., 2025; Jones and Tonetti, 2020; Veldkamp, 2023).

This paper construct firm-level measures of the price and quantity of U.S. public firms' data assets by combining economic modeling with finance-specialized machine learning (*FinBERT*). First, prior work shows that the quantities of a firm's capital, especially intangible capital, can be inferred from security values (Hall, 2001; Tambe et al., 2020). Hall (2001) develops a model to back out the price and quantity of total capital from market value, and Tambe et al. (2020) apply this framework to attribute market value to IT investment and estimate IT capital. We extend the model to measure firm-level data capital. Second, as the model requires a root to attribute market value to data assets, we leverage *FinBERT*[1] to classify data-asset-related sentences in 10-Ks and construct a disclosure-based root.[2] Our fine-tuned *FinBERT* achieves an AUC of 97.80% and 94.00% accuracy in detecting data-asset-related sentences (AUC is preferred under class imbalance). We then construct a firm-level disclosure measure for all U.S. public firms. A standard *FinBERT* application could stop here, using the frequency of data-asset-related sentences (*Data_freq*) as a proxy for data assets. However, disclosure may not map one-for-one into the level or value of data assets, given the potential big-data wash (firms may exaggerate data disclosure without undertaking investments to build data assets) (Barrios et al., 2024; Li, 2026) and the heterogeneous quality and user-specific pricing of data assets (Veldkamp, 2023). Our economic model addresses these concerns by incorporates market assessments that exploit equity markets' ability to distinguish big-data washing from genuine investment (Barrios et al., 2024; Li, 2026) and maps disclosure into separate price and quantity estimates of data capital

[1] *FinBERT* builds on BERT and is further trained on financial text (e.g., 10-K filings and earnings call transcripts). Huang et al. (2023) show that *FinBERT* outperforms BERT in detecting sentiment and ESG sentences in financial contexts. See Section 3 for details.

[2] Tambe et al. (2020) use IT-related labor as the accumulation root. A natural root for data assets is data-related labor, but outsourcing data collection or analysis can depress internal labor without reducing the underlying data capital, biasing estimates downward (Tambe et al., 2020). We still use labor-based roots in robustness checks. In addition, see Section 3 for why we use annual reports rather than conference calls.

to capture heterogeneous data-asset pricing. Using the *FinBERT*-based root to attribute market value to data assets, we back out the price *(Q)* and quantity *(K)* of data assets: Q borrows the notion of "Tobin's Q" and K is the "book value". See Section 3 for details.

To validate our firm-level data-asset measures, we conduct several tests. First, *Data_freq* and K rise sharply after 2009, consistent with the emergence of the big-data economy and advances in machine learning and AI. Second, *Data_freq* and K follow expected industry patterns, with higher levels in telecommunications, business equipment, finance, and health care, and lower levels in utilities, durable goods, and nondurable goods. Third, the price of data assets *(Q)* varies over time: it increases after the publication of the Transformer architecture in 2017 and the advent of Chat-GPT in 2022, and it declines following privacy shocks, including the proposal (2012) and enactment (2018) of the General Data Protection Regulation (GDPR) and the disclosure of the Cambridge Analytica data misuse scandal in 2018. Fourth, *Data_freq* and K are strongly related to big-data-related labor (employees with data or AI skills). Fifth, in line with literature (Abis and Veldkamp, 2024; Begenau et al., 2018; Veldkamp, 2023), data assets become increasingly concentrated among large firms.

We document two main results. First, consistent with shock-induced merger predictions, firms with greater data assets are more likely to become acquirers in the subsequent year. A one standard deviation increase in lagged data capital (*K_l1*) from its mean raises the next-year likelihood of becoming an acquirer by 1.77% (coefficients across measures are about 1-2%), representing 13.02% of the sample average likelihood 13.59%. Second, consistent with the view that the data revolution disproportionately benefits large firms, the baseline effect is concentrated among larger firms (either market capitalization or total assets), while firms of average size do not exhibit the baseline effect. Thus, data assets primarily facilitate M&A-driven growth among large firms, contributing to the rise of superstar firms. We further address endogeneity from reverse causality and omitted variables using two-stage least squares with instruments based on big-data-related human capital.

We examine explanations why data-asset induce merger activities. First, the main interpretation is defensive acquisitions. Gorton et al. (2009) develop a model in which defensive motives operate in a shock-induced merger environment: anticipating merger waves after regime shifts, managers undertake acquisitions to "eat" others before being eaten, creating a race to expand through M&A,

often with unprofitable deals. We therefore expect data-asset-induced M&A to display the defensive features highlighted by Gorton et al. (2009) and Harford (2005). We find patterns consistent with defensive motives: the baseline effect strengthens in industries with many small and medium data firms and weakens where large data firms dominate; data-asset-induced deals exhibit lower bidder announcement returns, lower deal synergies, and higher premiums.[3] Second, data assets can promote mergers by improving the information environment, liquidity, and the cost of capital. On the one hand, theory and evidence show that data assets enhance the information environment, increase liquidity, and lower the cost of equity (Baginski and Rakow, 2012; Lai et al., 2024). On the other hand, merger activity can be triggered directly by improvements in the information environment (Ferreira et al., 2010; Nain and Yao, 2013) and in financing conditions (Almeida et al., 2011; Harford, 2005; Huang et al., 2024). A direct link is provided by Abis and Veldkamp (2024), who show that firms' data acquisition improves the aggregate information environment and reduces the cost of equity, thereby increasing real investment. We therefore expect data assets to improve information and financing conditions, raising acquisition likelihood. The results support this expectation.

The paper contributes to the literature in several areas. First, we contribute to the measurement of firm-level data assets (e.g., Goldstein et al., 2021, 2024; Veldkamp, 2023). We introduce novel and comprehensive firm-level measures of data capital that explicitly treat data as an asset that accumulates and depreciates over time. Prior attempts are relatively narrow or indirect: Abis and Veldkamp (2024) measure data assets in the financial sector; Birch et al. (2021) examine a small set of large technology firms; Ma et al. (2025) and Tambe et al. (2020) measure firms' IT capital; and Schoch (2020) proxies data-rich firms using similarity to well-known data firms. To the best of our knowledge, our measures are the first comprehensive firm-level data-asset measures covering US public firms with 10-K disclosures.[4] Our methodological contributions are tri-fold. 1) We advance textual analysis in finance (e.g., Florackis et al., 2023; Hoberg and Phillips, 2016; Loughran and Mcdonald, 2011). Unlike keyword or dictionary approaches used to quantify policy uncertainty (Baker et al., 2016), geopolitical risk (Caldara and Iacoviello, 2022), and data-related content (Xiao et al., 2025; Xiao and

[3] We rule out two alternative mechanisms, the "positioning" motive and the misvaluation hypothesis in Section 6.

[4] Bhutani et al. (2025) provide a comprehensive measure based on voluntary management forecasts, covering necessarily smaller sample than ours due to the limited availability of such forecasts.

Liu, 2025), our fine-tuned *FinBERT* mitigates false positives from generic phrases (e.g., "financial data"). 2) Disclosure-based measures may overstate firms' data assets, as firms can exaggerate data-related language without making corresponding investments in data capital (big-data washing) (Barrios et al., 2024; Li, 2026). These measures also ignore heterogeneity in the pricing and quality of data, which are valued in a user-specific rather than uniform manner (Veldkamp, 2023). To address these concerns, our economic model incorporates market assessments, consistent with evidence that equity markets can distinguish big-data washing from genuine investment (Barrios et al., 2024; Li, 2026), and further maps disclosure into separate estimates of the price and quantity of data capital to capture the heterogeneous pricing of data assets. 3) Relative to labor-based proxies (either used alone or in our economic model) our disclosure-based root is less prone to downward bias from outsourcing data collection or analysis (Babina et al., 2024; Tambe et al., 2020), which reduces internal headcount without lowering underlying data capital. Our integrated "economics + LLM" method can be leveraged for other similar capital and are readily usable by researchers, analysts, investors, and regulators to study the role of these capital in corporate behavior and market structure.

Second, we contribute to the literature on technological shocks and resource reallocation. Prior work shows that regulatory, economic, and capital-market shocks trigger merger waves and firm reshuffling (e.g., Harford, 2005; Mitchell and Mulherin, 1996), yet we know little about how data assets shape acquisition decisions. Existing studies are mainly theoretical, case-based, or survey-based (Brynjolfsson and McElheran, 2016; Chen et al., 2022; Tambe, 2014) due to the absence of clear and comprehensive measures of data assets. Some papers take early steps to measure data assets narrowly or indirectly and to explore the relation between the big-data economy and mergers. Leveraging our new measures, we provide possibly the first comprehensive firm-level evidence that data revolution can induce resource reallocation through M&A activities.

Third, we link our findings to the literature on innovation-driven creative destruction and defensive mergers (e.g., Aghion and Howitt, 1992; Gorton et al., 2009). Our results supports data revolution can cause creative destruction by showing that data-asset-induced acquisitions are predominantly defensive and value-destroying. Results are difficult to reconcile with pure synergy or misvaluation explanations and instead point to a defensive motive: data-rich firms use M&A as an

inorganic response to data-driven technological change, at the expense of shareholder value.

Fourth, we identify an information and financing channel through which data assets foster M&A. A growing literature shows that data and analytics improve firms' information environment, increase liquidity, and lower the cost of equity (e.g., Abis and Veldkamp, 2024; Baginski and Rakow, 2012; Lai et al., 2024). At the same time, better information and easier financing spur merger activity (e.g., Almeida et al., 2011; Ferreira et al., 2010; Nain and Yao, 2013). Our evidence provides micro-level support for the mechanism highlighted by Abis and Veldkamp (2024): data acquisition enhances the aggregate information environment, lowers the cost of equity, and feeds into real corporate decisions.

Finally, we contribute to the literature on the rising of superstar firms (e.g., Autor et al., 2020; Corhay et al., 2025; Firooz et al., 2025). Prior work shows that larger firms adopt more and benefit more from IT and AI technologies (Babina et al., 2024; Tambe et al., 2020). Theory further predicts that larger firms accumulate more data and benefit disproportionately from the data revolution (Abis and Veldkamp, 2024; Begenau et al., 2018; Veldkamp, 2023). Our results suggest that the data revolution is an important driver of the superstar phenomenon by promoting rapid, inorganic M&A-led growth. Although we do not directly test whether these firms become superstars, our evidence provides a micro-foundation for how data assets and M&A jointly contribute to the rise of large, dominant firms in the data economy.

Section 2 reviews the basic data. Section 3 develop the data assets measures. Sections 4 reports how firms data asset promoting merger probability. Section 5 reports causal identification results. Section 6 presents mechanism analyses. Sections 7 reports how data assets affect deal-level outcomes. Section 8 presents robustness and other tests. Section 9 concludes.

2. Data and sample

2.1. Compustat sample

Our sample comprises the universe of Compustat firm-years with available 10-K filings on EDGAR from 1993 to 2024[5]. Results are robust when restricting the sample since 1996, when EDGAR coverage is complete. Following Edmans et al. (2012), we exclude acquirers and targets in financial

[5]We thank Bill McDonald for making the 10-Ks texts publicly available at https://sraf.nd.edu/sec-edgar-data

(SIC 6000-6999) and utility (SIC 4000-4949) industries. We further merge in stock price data from the Center for Research in Security Prices (CRSP). The resulting dataset contains 149,527 firm-years with non-missing 10-K filings since 1993 and 45,583 firm-years with non-missing data capital K, since 2009, and prior to removing observations with missing firm-level controls.

2.2. Merger and acquisition sample

The Securities Data Corporation (SDC) Mergers and Acquisitions database is the source of the M&A deals. The origin sample of deals are announced from 1 Jan 1980 to 31 Dec 2024; the acquirers are publicly traded firms and have stock price data from CRSP and accounting data from Compustat.[6] We require the acquirer to start with less than 50% of the target firm shares six months before the announcement date and aim to control higher than 50% of the target firm shares after the deal. We keep deals with the acquirer being publicly listed, and the target can be classified as private, public, or subsidiary. We exclude deals classified as restructuring, liquidation, bankruptcy, reverse takeovers, leveraged buyouts, going private transactions, government-owned involvement, spin-offs, recapitalizations, repurchases, exchange offers, and divestitures. We also exclude acquirers and targets in the financial (SIC 6000–6999) and utility (SIC 4000–4949) industries.

Table **2** presents the sample of deals. SDC provides information on the target's public status (public, private, or subsidiary), as well as whether the offer is diversified, a tender offer, friendly, or includes an Earnout contract. SDC also reports details on the payment method (full cash or full stock), relative deal size, number of bidders, and toehold positions. The full-period sample comprises 43,249 deals as shown in Panel A. The post-2009 sample with non-missing deal-level controls includes 5,444 deals as shown in Panel B, and the sample that also excludes observations with missing firm-level controls contains 3,858 deals. The cumulative abnormal return (CAR) for acquirers and targets is computed using a standard event study window [-2, +2], with the announcement date as day zero. Abnormal returns are measured using the widely adopted market-adjusted return approach, following Fuller et al. (2002). See Appendix Table **A1** for complete variable definitions.

[6] We adhere to the commonly accepted practice of maintaining a minimum six-month gap between the fiscal year-end data from Compustat and the share data from CRSP to ensure the financial report data are publicly known (see e.g., Alford et al., 1994; Fama and French, 1992).

2.3. Summary statistics

Table 1 presents summary statistics, including means, standard deviations, medians, and extreme values for the variables used in the analysis. All continuous variables are winsorized at the 1% and 99% level. Panel A reports results for the Compustat sample. The mean *M_A* (the annual probability of being a bidder) is 0.14% with a standard deviation of 0.34%. The mean *Data_freq* (the share of data-asset-related sentences in Items 1 and 1A of firms' 10-K filings) is 2.39%. The mean K (the model-generated quantity of data capital) is $170 million. The mean Q (the model-generated price of data capital) is 1.68. Panel B reports results for the SDC mergers and acquisitions sample. On average, deals are value creating in our sample, with mean CAR and synergy of 0.92% and 3.30%.

3. Measuring firm-year level data assets

Our objective is to develop firm-level measures of data assets. We proceed in four steps: defining data assets, specifying an economic model, using a large language model to proxy data disclosure, and inverting the model to recover data capital.

3.1. What is data asset?

Definition. We define a firm's data assets, following Veldkamp (2023) and Veldkamp and Chung (2019), as digital information the firm owns or can lawfully access and analyze to inform capital allocation, daily operations, and forecasts of revenues and earnings. Firms analyze data assets through analytics ranging from conventional statistical methods to modern big data technologies, including artificial intelligence and machine learning. **Sources and accumulation.** Much of the data economy is a byproduct of routine activity. Firms build data assets through transactions with customers and the digital footprints those interactions generate, such as browsing histories and search logs.[7] **Economic characteristics.** Data assets share key features with other intangibles such as technology or patents. They are inexpensive to reproduce, yet storage, curation, and analysis can be costly, and their value can erode through obsolescence, leakage, and regulatory shocks. Economically,

[7] According to Veldkamp (2023), transactions reveal preferences, willingness to pay, and other attributes. Additional signals arise from traffic patterns, social media content, and server logs, which can be monetized. Firms also expand their data holdings via acquisition, licensing, or sharing arrangements with third parties, including other firms, commercial data providers, and public sources.

data are nonrival (Jones and Tonetti, 2020), which limits their usefulness as collateral relative to physical capital or more generic intangible assets. **Measurement challenges.** Despite their growing centrality, data assets remain hard to measure. Standard macroeconomic accounts are ill suited to valuing them because data are invisible and lack natural units (Bhutani et al., 2025). The quality of data is heterogeneous, and pricing is user specific rather than uniform (Veldkamp, 2023).

3.2. Theoretical framework and economic modeling

The first key component of our data-asset measurement is the economic model. Prior work shows that the quantities of a firm's capital, including intangible capital, can be inferred from its market value (Hall, 2001; Tambe et al., 2020). Hall (2001) develops a capital accumulation framework that recovers the price and quantity of total capital from market value under a assumed zero-rent economy: 1) competitive markets, 2) constant returns to scale, and 3) long-run full factor adjustment. Tambe et al. (2020) apply this model to attribute market value to IT investments and estimate firm-level IT capital. To the best of our knowledge, these methods have not been used to measure data capital. We therefore extend this framework to proxy firm-level data assets by the quantity of data capital.

We have two equations of the capital accumulation model (Hall, 2001; Tambe et al., 2020):

$$V_t/K_t = Q_t \tag{1}$$

$$Q_t = 1 + c'(i_t//K_{t-1}) = 1 + \alpha \left[\frac{K_t - (1-\delta)K_{t-1}}{K_{t-1}}\right] \tag{2}$$

Equation 1 defines average Q (Tobin's Q) as the ratio of the market value attributed to data assets, V, to the dollar value of data capital, K. Equation 2 defines marginal Q as one plus the marginal adjustment cost as a function of investment growth, measured by the current year's investment, i_t, relative to the previous year's capital, K_{t-1}. The constant 1 captures the numeraire purchase price of a unit of capital. The adjustment-cost function is assumed to be convex, $c'(x) = \alpha x$, where $\alpha > 0$ is a constant. Investment is defined as $i_t = K_t - (1-\delta)K_{t-1}$, the change in the capital stock over the year, where δ is the annual depreciation rate. Under the zero-rent economy assumptions, average and marginal Q coincide, so Equation 1 equals Equation 2 (Hall, 2001; Hayashi, 1982; Tambe et al., 2020). This equality implies a unique solution for K and Q in the interval

$\min\left[(1-\delta)K_{t-1},\ V_t\right],\ \max\left[(1-\delta)K_{t-1},\ V_t\right]$.[8] Given assumed values for the remaining parameters, α, δ, and the initial K_{t-1},[9] we need only the V_t to solve the equation. We estimate the following yearly hedonic regression to recover the V_t of data assets:

$$V_i = \beta_0 + \beta_1 Root_i + \beta_2 PPENT + \beta_3 OtherAssets_i + FEs + u_i \tag{3}$$

where the dependent variable V_t denotes the firm's total market capitalization. $PPENT_i$ is property, plant, and equipment, and $OtherAssets_i$ is other physical assets. The specification includes industry but not year fixed effects as it is estimated as a yearly cross-sectional regression, with each yearly regression uses a three-year window $[t-1, t+1]$. These controls and fixed effects are not strictly required (Tambe et al., 2020). The estimation yields a time series of the data-capital-related coefficient $\widehat{\beta}_1$ and a firm-year Panel of predicted data-asset values, $\widehat{V}_t = \widehat{\beta}_1 \times Root_i$. An underlying assumption of this step is that the market recognizes the existence of data assets, that is, $\widehat{\beta}_1$ (the estimated coefficient on $Root_i$) is significantly positive. This assumption does not require correct market pricing of firm-level data assets; it only requires that the market does not ignore their existence, in the sense that data assets carry positive market value.

3.3. A large language mode: ***FinBERT***

After adopting the capital accumulation model, we must choose a root variable that links firms' market value to their data assets. Tambe et al. (2020) use IT-related labor as the economic input for IT capital, and big-data-related labor is a natural analogue for data assets. However, a labor-based root may understate data capital because firms can outsource data collection and analysis to third parties, which reduces internal labor and the implied price of data capital without necessarily reducing its quantity. We therefore use labor only as a robustness check. Instead, we select a disclosure-based root that more comprehensively captures the collection and use of data assets from internal and external sources. We measure this disclosure using annual reports. Following (Loughran and McDonald, 2024), annual reports are 10-K, 10-K405, 10KSB, and 10KSB40 filings, excluding amendments. Quarterly reports are unaudited and rarely describe business fundamentals, and conference calls focus on transitory topics such as labor shortages (Harford et al., 2024) or sentiment

[8] See the derivation of this equality in Section IA.1 and the proof of the unique root in Section IA.2.

[9] Results are not sensitive to these parameter assumptions, see Section *3.1* for the corresponding robustness checks.

Cao et al. (2023) rather than structural attributes such as having data assets. Data assets are a persistent feature of the business, are more likely to appear in routine legal disclosure, and are unlikely to change sharply over time, which supports the use of annual reports over call transcripts.

We identify data-asset disclosure by counting sentences that indicate the firm has data assets in Item 1s of the 10-K, where the firm describes its business; results are robust to including Item 7. Sentence classification uses *FinBERT* (Huang et al., 2023), which assigns one to data-asset-related sentences and zero otherwise. These sentences state that the firm's business involves data assets, including internal accumulation and external acquisition via public data, purchases, cooperation, or outsourcing. This meaning-based approach improves on simple keyword counts, which may misclassify statements such as "below is our financial data, revenue is ..." as data-asset-related.

We first fine-tune *FinBERT* and then apply it to the full sample. Fine-tuning requires a labeled training set, which we construct by using a deliberately broad list of data-asset-related keywords to retrieve potentially relevant sentences. The goal is to reduce false negatives by capturing a wide set of candidates and letting *FinBERT* separate true data-asset sentences from noise. The keyword list is assembled with the help of Chat-GPT and DeepSeek; standard embedding-based expansions such as *Word2Vec* perform poorly in this context (see Internet Appendix Table IA1 for the keyword list and example sentences). We randomly sample 4,000 sentences and manually label them; we have 968 data-asset-related. We hold out 10% of observations as a test set, use 10% of the remainder for validation, and train on the rest with a learning rate of 5e-4 and five epochs.

The fine-tuned *FinBERT* achieves an AUC of 98% and an accuracy of 94% in detecting data-asset-related sentences (Figure 1 and Internet Appendix Table IA2), where AUC is a better metric given the unbalanced sample with many non-data-asset sentences. We then construct $Data_freq_{i,t}$, a firm-year data disclosure for U.S. public firms from 1993 to 2024:

$$Data_freq_{i,t} = \frac{DASentences_{i,t}}{TotalSentences_{i,t}}, \tag{4}$$

where $DASentences_{i,t}$ is the number of data-asset-related sentences in Item 1s of firm i's 10-K in year t, and $TotalSentences_{i,t}$ is the total number of sentences in the same section.

In a standard *FinBERT* application, $Data_freq_{i,t}$ would serve directly as a proxy for data assets.

In our setting, however, it is not obvious that more frequent disclosure translates mechanically into more quantity or value of data for at least two reasons. First, firms may exaggerate data disclosure without undertaking investments to build data assets (data washing) (Barrios et al., 2024; Li, 2026). Second, data quality and value are highly heterogeneous, and data are priced in a user-specific rather than uniform manner (Veldkamp, 2023), which standard disclosure does not capture.

3.4. Joint estimation of data assets

Our economic model addresses these concerns by incorporating the market's assessment of each firm's data-asset disclosure,[10] aggregating investors' views, and separating the heterogeneous price of data assets from their quantity. We use the *FinBERT*-based disclosure measure as the root to identify the market value of data assets and to back out their price (Q) and quantity (K). Here, Q is Tobin's Q, capturing the price of data assets, and K is the book value, capturing their quantity.

The procedure is as follows. First, we estimate the yearly regressions in Equation 3 and obtain a time series of the data-capital-related coefficient $\widehat{\beta}_1$ and a firm-year Panel of predicted $\widehat{V}_t$ for data assets, $\widehat{V}_t = \widehat{\beta}_1 \times Root_i$. Figure 2 reports these results with 95% confidence intervals. Panel a shows that the market assigns a positive $\widehat{\beta}_1$ to data disclosure only after 2009, consistent with the rise of the big data economy, so we restrict our data-capital from 2009 onward. Panels b and c plot $\widehat{\beta}_1$ and $\widehat{V}_t$ since 2009 and indicate that the market attributes increasing value to data assets over time. Second, we accumulate data capital each year by solving the equality between the average-Q and marginal-Q conditions, using assumed values for α and δ, and an initial K_{t-1}. Following Tambe et al. (2020), we set $\alpha = 3$, $\delta = 0.1$, and $K_{t-1} = 0$ for each firm (see Section *3.1* for robustness to alternative settings). This procedure yields firm-year estimates of K (the quantity or "book value") and Q (the price, Tobin's Q) for data assets.

[10] Recent evidence suggests that equity markets can distinguish big-data washing from genuine investment (Barrios et al., 2024; Li, 2026).

3.5. Validation

3.5.1. Variance decomposition

Following Harford et al. (2024) and Hassan et al. (2019), we decompose the variance of our data-asset measures (*Data_freq*, *Data_freq* since 2009, K, and Q). Internet Appendix Table IA3 reports the incremental adjusted-R^2 from adding year, industry, and year-industry fixed effects. These effects explain less than 40% of the variation, leaving over 60% at the firm level. For *Data_freq*, *Data_freq* since 2009, and K, this firm-level variation is almost entirely between firms, whereas most of the variation in Q is within firms, consistent with heterogeneous pricing of data assets.

3.5.2. The distribution of data assets measures

Figure 3 plots heatmaps of *Data_freq*, K, and Q by year, industry, and year-industry cell. Data assets (*Data_freq* and K) accumulate over time and rise sharply after 2009, 2017, and 2022. They are also concentrated in industries where data use is expected to be intensive, such as telecommunications, business equipment, finance, and health care, and are lower in utilities and in durable and nondurable goods. The price measure Q is volatile: it increases after the introduction of the Transformer architecture in 2017 and the advent of Chat-GPT in 2022, and it declines around major privacy shocks, including the proposal (2012) and enactment (2018) of the General Data Protection Regulation and the Cambridge Analytica scandal in 2018.

3.5.3. The real examples of data assets measures

We also provide firm-level examples to illustrate the data asset measures in Figure 4, Panels (a), (b), and (c), respectively. First, the yearly trajectories of these measures for each firm broadly track the corresponding full-sample averages. Second, the accumulation of data capital is gradual: for instance, a late-sample surge in data disclosure by Johnson & Johnson that exceeds Merck & Co. (two companies in the health-care industry) does not immediately raise Johnson & Johnson's stock of data capital K above that of Merck & Co., instead it is reflected in a higher price Q. This gap between disclosure and data assets may also be interpreted as potential data washing, whereby firms engage in cheap talk rather than substantive investment (Barrios et al., 2024; Li, 2026). Third, in

some cases data disclosure and data capital move together. Microsoft exhibits both higher data disclosure and a larger stock of data capital than Apple, with data capital of about 1 billion and 600 million, respectively. Likewise, Bank of America has substantially higher disclosure, capital, and price than Goldman Sachs, consistent with a retail bank controlling a much larger volume of customer information than an investment bank.

3.5.4. The firm-level big-data input: labor

We further validate our measure by showing that the accumulation of data capital depends primarily on big-data inputs, measured by labor with data and AI skills. Table 5, column (2), shows that a firm's AI-skilled labor and data-skilled job postings are both significantly and positively associated with the quantity of data capital K. A one standard deviation increase in these labor inputs is associated with 5.46% and 6.38% of a standard deviation of K (228.20 million), corresponding to increases of 12.46 and 14.56 million, respectively.

4. Data asset and firms' acquisitiveness

4.1. Data-asset-induced mergers and acquisitions

To test the impact of data assets on a firm's acquisitiveness, we run the following linear probability model (LPM) regressions:[11]

$$M_A = \beta_0 + \beta_1 X + \beta_2 Controls + u \tag{5}$$

where M_A is a dummy equal to one if a firm announces an M&A deal as an acquirer in a given year. X denotes one of three alternative measures, indicated under column numbers: the fine-tuned FinBERT-based data asset disclosure measure ($Data_freq$), the model-implied data capital (K), and the model-implied price of data assets (Q). All variables are lagged by one year, though results are robust to contemporaneous specifications.

Table 3 shows that firms with greater data assets are more likely to become acquirers in the subsequent year. Panels A, B, and C present results without controls or fixed effects; with controls

[11] Employing logistic, probit, or linear probability regressions with/without fixed effects yields quantitatively and qualitatively similar results. In all regressions, the indicated "Controls" means the same set of control variables; "YearEffect" means the year fixed effect; and "IndustryEffect" means the Fama–French 48 industry groups of the company.

and year and industry fixed effects; and with controls and year and firm fixed effects, respectively. The significance of all data asset measures in Panels A and B, but not in Panel C, suggests that the effect of data assets is primarily driven by between-firm variation rather than within-firm variation. This pattern aligns with prior evidence on the rise of superstar firms (Autor et al., 2020; De Loecker et al., 2020) and mirrors dynamics observed in shock-driven industry transformations (Gorton et al., 2009; Harford, 2005).[12] All variables are standardized to facilitate economic interpretation. The main results, shown in Column (2) of Panel B (with full controls, and year and industry fixed effects), indicate that a one standard deviation increase in lagged data capital (K_l1) from its mean raises the likelihood of becoming an acquirer in the following year by 1.77% (coefficients across all measures are around 1–2%), representing 13.02% of the sample average likelihood 13.59%.

4.2. Size effect

After establishing that data assets increase a firm's acquisitiveness, we examine how this relation varies with firm size. Figure 5 plots the annual distribution of our data-asset measures across size groups. Results show that, although the implied prices of data assets are similar across size groups, the levels of data assets (*Data_freq* and K) are increasingly concentrated among large market-cap firms. We further directly test whether the baseline effect asymmetrically affects large firms:

$$M_A = \beta_0 + ... + \beta_3(X \times Size_Proxy) + \beta_4 Controls + u \tag{6}$$

where *Size_Proxy* refers to either the firm's standardized market capitalization or total assets. Table 4 shows that the baseline effect of data assets is disproportionately concentrated among larger firms, as indicated by market capitalization in Panel A and total assets in Panel B. The interaction term is consistently positive and significant, suggesting that a one standard deviation increase in firm size amplifies the baseline effect by approximately 1%. In column (2) of Panel A, the coefficient on our main variable of interest, K, becomes statistically insignificant, indicating that firms with average market capitalization do not exhibit the baseline effect. In column (2) of Panel B, the coefficient on the interaction term is about twice the magnitude of the coefficient on K, implying that relatively

[12] The only exception is the contemporaneous data asset price measure (Q), which remains significant, albeit with a smaller coefficient in Panel C. This likely reflects the greater relevance of data asset valuation during the deal year at the firm level.

small firms do not experience the baseline effect. Results indicate that data assets primarily facilitate M&A-driven growth among large firms, contributing to the rise of super-sized firms. As large firms face constraints on organic growth, including rising costs of innovation (Akcigit and Kerr, 2018; Babina et al., 2024). Our evidence shows that data assets facilitate their inorganic expansion.

5. Endogeneity and Identification

Having established a statistical association between data assets and merger activity, we next present causal evidence. Two identification threats warrant attention: reverse causality, whereby potential acquirers invest more in data assets; and omitted variables, whereby unobserved firm characteristics affect both data accumulation and merger activity.

We provide evidence of causal identifications by using instrumental variable two-stage least squares (IV-2SLS) regressions in this section. Specifically, we use three instruments: 1) *AI_exposure_2010*, 2) *njob_data*, and 3) *njob_AI*.

The first instrumental variable, *AI_exposure_2010*, captures a firm's historical link to AI-intensive universities as of 2010. The long-term value of data assets relies on its interaction with technological progress (i.e., AI) (Bhutani et al., 2025; Farboodi and Veldkamp, 2021), and firms benefit disproportionately from AI due to their capacity to leverage data assets (Babina et al., 2024). In addition, according to Babina et al. (2024), university expansion is unlikely to directly affect firms' acquisitiveness and is plausibly exogenous to their post-2010 AI or data investment decisions, given that commercial interest in AI only began to accelerate around 2012. In the same spirit, we use firms' AI-related employees and data-related job postings as the second and third instruments, respectively. Related workers is the necessary condition to realize the benefits data assets (Abis and Veldkamp, 2024). Firms hire such labor with the primary intent of enabling data collection, storage, and analysis. The more worker, the more valuable is the data stock.[13] Table **5** presents the 2SLS results to

[13]We thank the authors of Babina et al. (2024) for providing all three IVs.

identify the causal interpretation of data assets:[14]

$$M_A = \beta_0 + \beta_1 \widehat{K} + \beta_2 Controls + u \tag{7}$$

Columns (1) and (2) report the first-stage results, while columns (3) and (4) present the second-stage estimates. Using either the first ex-ante instrument or all three instruments jointly, the coefficient on $\widehat{K}$ in columns (3) and (4) remains significantly positive. The results support the causal effect of K, validated by three identification tests: under-identification, over-identification, and weak identification. Kleibergen-Paap-LM-statistics reject the null hypothesis of under-identification at the 1% level, while the Sargan-Hansen-J-statistics fail to reject the null hypothesis of over-identification at the 10% level. The Kleibergen-Paap F statistics are 137.6167 or 35.0588 for regressions using only the first IV or all three IVs, respectively, exceeding the corresponding critical values of 16.38, 13.91, and 22.30 from Stock and Yogo (2005) and indicating strong instruments. The first-stage results in columns (1) and (2) align with expectations: greater labor availability and inputs are associated with more data assets. Overall, the findings support a causal interpretation.

6. Mechanism

Building on prior literature, we identify potential channels through which data assets may increase firms' acquisitiveness and test whether these channels serve as underlying mechanisms. Specifically, we examine two mechanisms: (1) industry transformation in the data age: "eat or be eaten", (2) improved information environment, liquidity, and the cost of capital.

Our findings indicate that in industries populated by many small and medium-sized data-intensive competitors, firms face heightened survival threats and respond with defensive acquisitions to avoid becoming targets. Improvements in the information environment, market liquidity, and cost of equity from data assets provide a complementary channel to the shock-induced mergers.

[14] To preserve the ex-ante nature of the first instrument, the sample prior to 2010 is excluded. The results remain robust when using the full sample or when additionally excluding 2010. Moreover, untabulated results confirm robustness when K is instrumented separately by each of the three IVs.

6.1. Eat or be eaten: a defensive motive of M&A

Major technological shifts reshape the economy by altering firm boundaries and market structure. They trigger large-scale asset reallocation and induce creative destruction (Aghion and Howitt, 1992). Mergers are a key channel for firms to respond to technological and regulatory shocks through rapid, inorganic expansion (Gort, 1969; Harford, 2005; Mitchell and Mulherin, 1996). Building on this, Gorton et al. (2009) develop a model in which defensive motives are embedded in a shock-induced merger environment. Managers, in anticipating merger waves following regime shifts, engage in defensive acquisitions. They are actively "eating" other firms to avoid being eaten, resulting in a race to expand through M&A. They show that firms are more (less) likely to make defensive acquisitions when their own industries have more medium-sized (large) firms. Similarly, Hagiu and Wright (2023) develop a model in which firms' learning from their data assets leads to "killer" M&A, defined as aggressive and preemptive acquisitions undertaken before competitors can do.

The data economy constitutes a new, economy-wide technological revolution that naturally fits within this regime-shift framework. Data is transforming production functions, competitive dynamics, and firm growth (Abis and Veldkamp, 2024; Farboodi and Veldkamp, 2021; Jones and Tonetti, 2020) and the magnitude of changed production is comparable to industrial revolution (Abis and Veldkamp, 2024). Taken together, these arguments suggest that data assets, as the fuel of the current technological shock, are driving another structural regime shift and prompting firms' defensive M&A in anticipation of data-driven creative destruction. We therefore expect data-asset-induced M&A activity to exhibit the shock-induced defensive characteristics highlighted by Gorton et al. (2009) and Harford (2005). Specifically, 1) the baseline effect of data assets is stronger (weaker) in industries with a high concentration of small and medium-sized (large) data firms, where takeover threats are more (less) salient; and 2) such deals are more likely to exhibit value-destroying features, including lower bidder announcement returns and synergies, and higher offer premiums.[15] Table **6** presents

[15]We also rule out other two potential interpretations or mechanisms underlying our baseline effect. First, Gorton et al. (2009) also identify a "positioning" motive, where firms acquire others not to grow independently, but to better position themselves for future acquisition by large firms. In this case, deals are value-creating once the option value of future resale is considered, and the more likely to emerge in industries with the presence of large firms. Our findings contradict both predictions, thereby ruling out the positioning motive as a driver of our results. Second, we consider the misvaluation hypothesis, which posits that acquirers exploit overvalued equity to finance acquisitions. According to Harford (2005), deals motivated by misvaluation should disproportionately involve stock payment, deviating from

results about how data assets distribution within industry affect the baseline effect:

$$M_A = \beta_0 + ... + \beta_3(K \times pct_Data_firms) + \beta_4 Controls + u \tag{8}$$

Following the approach in Gorton et al. (2009), *pct_Data_firms* refers to one of *pct_K_lt_5*, *pct_K_5_30*, or *pct_K_gt_30*, which indicate the percentage of firms whose data capital K falls within [0%, 5%), [5%, 30%], and (30%, 100%] of the maximum data capital value in each industry, respectively. The significant positive and significant negative coefficients in the first two columns and the last column indicate that data-induced acquisitions are stronger (weaker) when there are more small and medium-sized (large) data firms, consistent with the defensive motive. The second part of expectations that data-induced acquisitions are value destroying is also supported and discussed detail in Section 7.

Overall, the defensive strategy mechanism is likely affirmed. Data assets are prompting firms' defensive M&A in anticipation of data-driven creative destruction.

6.2. Information environment, liquidity, and cost of capital

Data assets can also promote merger activity by improving firms' information environment, liquidity, and cost of capital. On the one hand, data assets enhance firms' ability to communicate internal information to the market, thereby improving the information environment, increasing liquidity, and lowering the cost of equity (Baginski and Rakow, 2012; Lai et al., 2024). These effects are more pronounced for complex firms, where disclosure is more necessary (Farzamfar et al., 2024), and when disclosure is more quantitative (Campbell et al., 2025). Empirical studies also document that better information environments are associated with lower cost of capital and greater real investment (Billett et al., 2017; Derrien and Kecskés, 2013).[16] A more direct link is provided by Abis and Veldkamp (2024), who show that firms' data acquisition improves the aggregate information environment, reduces the cost of equity, and increases real investment. On the other hand, liquidity and the cost of capital are essential transmission channels through which economic shocks lead to capital reallocation (i.e., mergers and acquisitions in our context) (Eisfeldt and Rampini, 2006; Harford, 2005). In

the even likelihood of cash or stock predicted under neoclassical shock-driven mergers. Untabulated results ruling out the misvaluation channel by showing that data assets do not increase but insignificantly decrease the likelihood of stock payment (neither fully or partially).

[16] See also Begenau et al. (2018); Farboodi et al. (2022); Farboodi and Veldkamp (2020); Zhu (2019) for how the development of big data enhances firm-level informativeness and market liquidity.

addition, merger activity may also be directly triggered by improvements in the information environment (Ferreira et al., 2010; Nain and Yao, 2013) and financing conditions (Almeida et al., 2011; Harford, 2005; Huang et al., 2024). We therefore expect that data assets improve firms' information environment and financing conditions, increasing the likelihood of engaging in acquisitions. Empirically, we test whether firms with more data assets: (1) are more likely to issue earnings guidance; (2) attract greater institutional ownership and analyst coverage; (3) have more informative (i.e., less noisy) stock prices; and (4) experience higher liquidity and lower cost of equity. We then examine whether these improvements lead to a higher probability of acquisition.

We employ two approaches. First, we separately estimate: 1) whether data assets improve the information environment and cost of capital, as in Equation 9; and 2) whether an improved information environment and lower cost of capital increase the likelihood of becoming an acquirer, as in Equation 10. Second, following the two-step approach of Di Giuli and Laux (2022), we test whether the portion of the improved information environment and cost of capital attributable to data assets leads to a higher probability of acquisition. Specifically, we treat the main variable of interests (data assets) as an "instrument" of the mechanism variables to assess their effect on acquisition likelihood, as shown in Equation 11 and using Equation 9 as the first stage:

$$Mecha = \beta_0 + \beta_1 K + \beta_2 Controls + u \tag{9}$$

$$M_A = \beta_0 + \beta_1 Mecha + \beta_2 Controls + u \tag{10}$$

$$M_A = \beta_0 + \beta_2 \widehat{Mecha} + \beta_3 Controls + u \tag{11}$$

Table 7 reports the results of Equation 9, Equation 10, and Equation 11 in Panels A, B, and C, respectively. Results are consistent with expectations. Panel A shows that firms with greater data assets are more likely to issue earnings guidance, attract more institutional ownership and analyst coverage, experience lower price noise, have higher stock liquidity, and exhibit a lower cost of equity (but not debt). Panel B indicates that these improvements significantly increase firms' acquisitiveness. Panel C further confirms this mechanism by showing that improvements in the information environment and financing conditions, driven by data assets, lead to a higher probability of acquisition.

Overall, the data assets improved information environment and cost of equity is also likely affirmed to facilitate data-induced firm reallocation. Data assets enable earlier equity financing, thereby promoting merger activity.

7. Data capital and deal-level outcomes

We also explored how data capital affect deal-level outcomes to verify the defensive motive of data-assets-induced mergers. Specifically, Gorton et al. (2009) illustrate that defensive deals are: more likely to be unprofitable and value-destroying, especially for large firms. In addition, as large firms experiences poor merger performance and the data-asset-induced acquisitiveness concentrates in large firms, we expect that the data-asset-induced deals are more likely to be value-destroying. Table 8 presents results about how data assets affect deal-level outcomes:

$$SDC_Ys = \beta_0 + \beta_2 K + \beta_3 Controls + u \tag{12}$$

where SDC_Ys denote deal-level outcomes. In each specification, when an outcome would otherwise appear as a control, we exclude the corresponding dependent variable from the control set to present perfect predictions. First, consistent with the baseline effect, acquirers with greater data assets are more likely to be frequent acquirers and to have more deals in the prior 3-year or 5-year windows (Panel A, columns 1 to 3). Second, consistent with the defensive strategy in Gorton et al. (2009), data assets are associated with value-destroying deals and relatively smaller targets, evidenced by lower bidder announcement returns and lower deal synergies (Panel A, columns 4 and 5), higher premiums (Panel B, column 1), a lower likelihood of private targets (Panel B, column 2), and a lower relative size (Panel B, column 4).[17] Third, Harford (2005) argues that shock-induced mergers should not rely asymmetrically on stock financing, unlike misvaluation-driven deals. Untabulated tests support this prediction, as data assets do not affect the choice between stock and cash (even insignificantly negative to stock payment). Fourth, data-asset-induced deals are less likely to be diversifying (Panel B, column 5), indicating that data assets facilitate matching within the same

[17] In untabulated tests, instrumenting premiums with firm-level data assets following the method of Baker et al. (2012), the data-asset-driven component of premiums is significantly negatively related to bidder CAR. We interpret this evidence cautiously about the magnitude of coefficients, given a relatively weak first stage. The negative sign is consistent with the view that data-asset-induced defensive mergers reflect inorganic expansion in response to regime shifts through an overpayment channel.

industry (2-digit SIC). Fifth, data acquirers are more likely to include earnout provisions (Panel B, column 3), probably reflecting either data assets allowing enhanced monitoring or data-related deals are with greater evaluation difficulty (Schoch, 2020).

8. Robust checks and other tests

8.1. The effectiveness of other possible data asset measures

We replace K with three groups of alternative data-asset measures. First, we use our LLM-generated *Data_freq*, which in the baseline regressions increases firms' acquisitiveness. Second, we employ the share data-related job posting from Babina et al. (2024).[18] In addition, we use data-job-posting count as the root to infer data-related market value and construct data capital *K_labor* by our model, following Tambe et al. (2020). Third, motivated by the ideal that main effect of data assets is to resolve risk of users (Abis and Veldkamp, 2024; Bhutani et al., 2025; Farboodi et al., 2024), we use the negative of the absolute management forecast errors to proxy for data assets, based on both earnings and sales forecasts. Appendix Table **A2** reports results replacing K in Equation 5 for the post-2009 sample (Panel A) and for the full sample with non-missing corresponding data measures (Panel B). Across Panels, the estimates confirm the baseline: stronger data assets raise the next-year probability of becoming a bidder, with comparable magnitudes. A one standard deviation increase in a data-asset proxy lifts the probability of being a bidder in the next year by around 1% to 2%.

8.2. Data assets and firms' targetiveness

An extensive literature shows that firms with emerging technologies or capital attract acquirers (Chen et al., 2022; Faria, 2008; Huang and Xie, 2023). It is also a natural conclusion from the "eat or being eaten" framework of (Gorton et al., 2009) that all firms, especially data-firms, pursue data targets. We run the following regressions:

$$M_A_target = \beta_0 + ... + \beta_1 X + \beta_2 Controls + u \tag{13}$$

$$M_A_target = \beta_0 + ... + \beta_3 (X \times Year_2009) + \beta_4 Controls + u \tag{14}$$

[18] We thank the authors of Babina et al. (2024) for sharing their data. The coverage for data-related labor is 2007 and 2010–2018. Consequently, regressions using labor counts have a slightly smaller sample from 2009 onward than the full-period baseline. In addition, the model-generated *K_labor*, rooted in data-labor, begins in 2010 to ensure continuity.

where *M_A_target* is a dummy equals one if a firm announces an M&A deal as an acquirer in a given year. Panel A in Appendix Table **A3** reports results for the post-2009 sample as Equation 13, showing an insignificant effect of data assets on the likelihood of becoming a target. However, interpretations are tri-fold. First, data assets do not affect whether a firm is acquired. Second, data-intensive targets are primarily small firms or start-ups that are mainly private and therefore absent from our public-firm data-asset measures. Third, data capital has become increasingly concentrated in large firms (through organic accumulation or by large firms acquiring data-rich small firms). As large firms are difficult to acquire for various constraints (Gorton et al., 2009), the attraction of data assets may persist but rarely translates into being a target after this concentration. We first rule out the "small-private-target" interpretation and then distinguish between the "no-effect" and "post-2009 concentration" interpretations.

The evidence contradicts the second interpretation. Untabulated results shows that data acquirers are more likely to match with targets that hold more data assets in public-public deals, although this subsample is small because public-public domestic deals with non-missing K after 2009 are rare. Data-asset acquirers are also less likely to purchase private targets. Moreover, data-induced acquisitiveness is stronger in industries with many small and mid-sized public data firms, indicating that acquirers value smaller public data firms. Taken together, these patterns rule out the second interpretation. Furthermore, interacting firm size with data assets yields insignificant effects on becoming a target, further ruling out the possibility that the target effect is confined to small public firms.

We further distinguish between the first and third interpretations. Figure 5 shows that *Data_freq* and K begin to concentrate in large firms after 2009. We therefore test whether the data-target relation exists before 2009, when data capital had not yet been asymmetrically concentrated in large firms. We estimate Equation 14 by interacting a post-2009 year dummy with data assets. A significantly negative interaction supports the third interpretation; an insignificant coefficient supports the first; a significantly positive coefficient is unlikely given the insignificant post-2009 effect. This test requires measures observed before 2009, leaving *Data_freq* and two management-forecast-based measures. Panel B supports the third interpretation. In the first three columns, none of the measures affects target likelihood in the full sample. In the last three columns, the interaction

terms are negative, indicating that data assets increase target probability before 2009, and that this effect is fully offset after 2009 (although the statistically insignificance in the last column).

Overall, data assets increase the likelihood of being targeted only before data capital becomes concentrated in large firms. The economic magnitude is also significant: a one standard deviation increase in data assets raises the pre-2009 probability of being target by about 0.16%, representing 23.88% of the sample average of being target likelihood 0.67%.

8.3. Further robust checks

We show that our results are robust to alternative model parameter choices in the Internet Appendix *3.1*. The model-generated data capital K is not overly sensitive to the adjustment-cost parameter α or the initial capital K_0. More importantly, changes in parameter values affect only the level of K, not its between-firm or within-firm variation.

9. Conclusion

We study how firm-level data assets shape mergers and acquisitions in the big data economy. We begin by developing new measures of the price and quantity of data capital for U.S. public firms, combining an economic model with a finance-specialized large language model. Our fine-tuned *FinBERT* classifier isolates data-asset-related disclosure in 10-K filings, and the economic model maps this disclosure into separate estimates of data-asset prices and quantities, rather than relying solely on textual frequency or data-related labor inputs.

Using these measures, we provide firm-level evidence that the data revolution reallocates corporate control through M&A. Consistent with shock-induced merger theories, we show that data-rich firms are significantly more likely to become bidders in the subsequent year, and that this effect is concentrated among large firms. Thus, data assets primarily facilitate rapid, inorganic growth for large incumbents rather than for firms of average size. We further document that data-asset-induced deals exhibit the hallmarks of defensive mergers: the effect of data assets on bidder incidence is stronger in industries populated by many small and medium data firms, and acquisitions undertaken by data-rich bidders are, on average, value-destroying. In addition, we show that data assets improve

the information environment, liquidity, and the cost of equity, and that these improvements are associated with a higher likelihood of becoming a bidder, highlighting an information-and-financing channel through which data assets foster M&A.

Ultimately, our findings indicate that data assets are a central mechanism through which the big data economy reshapes firm boundaries and market structure. By treating data as a form of capital that accumulates and depreciates, our measurement framework provides a reusable toolkit for researchers, analysts, investors, and regulators interested in the role of data and similar intangible assets in corporate behavior. At the same time, the evidence that data-rich, large firms use M&A to respond to technological change highlights a "big get bigger" dynamic that may underpin the rise of dominant firms in data-intensive industries. We do not directly test whether data-rich acquirers become superstar firms or assess the welfare implications of data-driven consolidation. Examining these questions across other corporate policies, private firms, and international settings remains an important avenue for future research on the interaction between data, corporate control, and long-run economic dynamism. Investigating how data assets shape industry- and macro-level merger waves is another promising direction. Moreover, the defensive and value-destroying nature of the M&A response suggests that technological innovation in the big data economy can foster inefficient within-market consolidation. These patterns call for regulators and antitrust authorities to pay close attention to potentially value-destroying, defensive M&A by data-rich incumbents.

References

Abis, S., Veldkamp, L., 2024. The Changing Economics of Knowledge Production. The Review of Financial Studies 37, 89–118.

Aghion, P., Howitt, P., 1992. A Model of Growth Through Creative Destruction. Econometrica 60, 323–351.

Akcigit, U., Kerr, W. R., 2018. Growth through Heterogeneous Innovations. Journal of Political Economy 126, 1374–1443.

Alford, A. W., Jones, J. J., Zmijewski, M. E., 1994. Extensions and violations of the statutory SEC form 10-K filing requirements. Journal of Accounting and Economics 17, 229–254.

Almeida, H., Campello, M., Hackbarth, D., 2011. Liquidity mergers. Journal of Financial Economics 102, 526–558.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., Van Reenen, J., 2020. The Fall of the Labor Share and the Rise of Superstar Firms. The Quarterly Journal of Economics 135, 645–709.

Babina, T., Fedyk, A., He, A., Hodson, J., 2024. Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics 151, 103745.

Baginski, S. P., Rakow, K. C., 2012. Management earnings forecast disclosure policy and the cost of equity capital. Review of Accounting Studies 17, 279–321.

Baker, M., Pan, X., Wurgler, J., 2012. The effect of reference point prices on mergers and acquisitions. Journal of Financial Economics 106, 49–71.

Baker, S. R., Bloom, N., Davis, S. J., 2016. Measuring Economic Policy Uncertainty. Quarterly Journal of Economics 131, 1593–1636.

Barrios, J. M., Campbell, J. L., Johnson, R. G., Liu, C., 2024. Signals or Smoke? The Determinants and Informativeness of Corporate Artificial Intelligence (AI) Disclosures</p>.

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

Bhutani, A., Ordonez, G., Veldkamp, L., 2025. The Missing Value of Data.

Billett, M. T., Garfinkel, J. A., Yu, M., 2017. The effect of asymmetric information on product market outcomes. Journal of Financial Economics 123, 357–376.

Birch, K., Cochrane, D., Ward, C., 2021. Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. Big Data & Society 8, 20539517211017308.

Brogaard, J., Nguyen, T. H., Putnins, T. J., Wu, E., 2022. What Moves Stock Prices? The Roles of News, Noise, and Information. The Review of Financial Studies 35, 4341–4386.

Brynjolfsson, E., McElheran, K., 2016. The Rapid Adoption of Data-Driven Decision-Making. The American Economic Review 106, 133–139.

Caldara, D., Iacoviello, M., 2022. Measuring Geopolitical Risk. American Economic Review 112, 1194–1225.

Campbell, J. L., Zheng, X., Zhou, D., 2025. Number of numbers: Does a greater proportion of quantitative textual disclosure reduce information risk? Journal of Corporate Finance 94, 102813.

Cao, S., Jiang, W., Yang, B., Zhang, A. L., 2023. How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. The Review of Financial Studies 36, 3603–3642.

Chen, Z., Choe, C., Cong, J., Matsushima, N., 2022. Data-driven mergers and personalization. The RAND Journal of Economics 53, 3–31.

Corhay, A., Kung, H., Schmid, L., 2025. Q: Risk, rents, or growth? Journal of Financial Economics 165, 103990.

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

Derrien, F., Kecskés, A., 2013. The Real Effects of Financial Shocks: Evidence from Exogenous Changes in Analyst Coverage. The Journal of Finance 68, 1407–1440.

Di Giuli, A., Laux, P. A., 2022. The effect of media-linked directors on financing and external governance. Journal of Financial Economics 145, 103–131.

Edmans, A., Goldstein, I., Jiang, W., 2012. The Real Effects of Financial Markets: The Impact of Prices on Takeovers. The Journal of Finance 67, 933–971.

Eisfeldt, A. L., Rampini, A. A., 2006. Capital reallocation and liquidity. Journal of Monetary Economics 53, 369–399.

Fama, E. F., French, K. R., 1992. The Cross-Section of Expected Stock Returns. The Journal of Finance 47, 427–465.

Farboodi, M., Matray, A., Veldkamp, L., Venkateswaran, V., 2022. Where Has All the Data Gone? The Review of Financial Studies 35, 3101–3138.

Farboodi, M., Singal, D., Veldkamp, L., Venkateswaran, V., 2024. Valuing Financial Data. The Review of Financial Studies p. hhae034.

Farboodi, M., Veldkamp, L., 2020. Long-Run Growth of Financial Data Technology. American Economic Review 110, 2485–2523.

Farboodi, M., Veldkamp, L., 2021. A Growth Model of the Data Economy.

Faria, A. L., 2008. Mergers and the market for organization capital. Journal of Economic Theory 138, 71–100.

Farzamfar, A., Foroughi, P., Hamisheh Bahar, H., Ng, L., 2024. Illuminating the murk: The effect of business complexity on voluntary disclosure. Journal of Corporate Finance 87, 102612.

Ferreira, M. A., Massa, M., Matos, P., 2010. Shareholders at the Gate? Institutional Investors and Cross-Border Mergers and Acquisitions. The Review of Financial Studies 23, 601–644.

Firooz, H., Liu, Z., Wang, Y., 2025. Automation and the rise of superstar firms. Journal of Monetary Economics 151, 103733.

Florackis, C., Louca, C., Michaely, R., Weber, M., 2023. Cybersecurity Risk. The Review of Financial Studies 36, 351–407.

Fuller, K., Netter, J., Stegemoller, M., 2002. What Do Returns to Acquiring Firms Tell Us? Evidence from Firms That Make Many Acquisitions. The Journal of Finance 57, 1763–1793.

Goldfarb, A., Tucker, C., 2019. Digital Economics. Journal of Economic Literature 57, 3–43.

Goldstein, I., Spatt, C. S., Ye, M., 2021. Big Data in Finance. The Review of Financial Studies 34, 3213–3225.

Goldstein, I., Spatt, C. S., Ye, M., 2024. The Next Chapter of Big Data in Finance. The Review of Financial Studies p. hhae083.

Gort, M., 1969. An Economic Disturbance Theory of Mergers. The Quarterly Journal of Economics 83, 624–642.

Gorton, G., Kahl, M., Rosen, R. J., 2009. Eat or Be Eaten: A Theory of Mergers and Firm Size. The Journal of Finance 64, 1291–1344.

Hagiu, A., Wright, J., 2023. Data-enabled learning, network effects, and competitive advantage. The RAND Journal of Economics 54, 638–667.

Hall, R. E., 2001. The Stock Market and Capital Accumulation. American Economic Review 91, 1185–1202.

Harford, J., 2005. What drives merger waves? Journal of Financial Economics 77, 529–560.

Harford, J., He, Q., Qiu, B., 2024. Firm-Level Labor-Shortage Exposure.

Hassan, T. A., Hollander, S., van Lent, L., Tahoun, A., 2019. Firm-Level Political Risk: Measurement and Effects. The Quarterly Journal of Economics 134, 2135–2202.

Hayashi, F., 1982. Tobin's Marginal q and Average q: A Neoclassical Interpretation. Econometrica 50, 213–224.

Hoberg, G., Phillips, G. M., 2016. Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy .

Huang, A. H., Wang, H., Yang, Y., 2023. FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research 40, 806–841.

Huang, J., Xie, T., 2023. Technology centrality, bilateral knowledge spillovers and mergers and acquisitions. Journal of Corporate Finance 79, 102366.

Huang, S., Maharjan, J., Nanda, V., 2024. Liquid stock as an acquisition currency. Journal of Corporate Finance 85, 102562.

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

Lai, S., Lin, C., Ma, X., 2024. RegTech Adoption and the Cost of Capital. Management Science 70, 309–331.

Lashkari, D., Bauer, A., Boussard, J., 2024. Information Technology and Returns to Scale. American Economic Review 114, 1769–1815.

Lee, C. M. C., So, E. C., Wang, C. C. Y., 2021. Evaluating Firm-Level Expected-Return Proxies: Implications for Estimating Treatment Effects. The Review of Financial Studies 34, 1907–1951.

Li, B., 2026. AI Washing. In: AFA poster session.

Loughran, T., Mcdonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2024. Measuring Firm Complexity. Journal of Financial and Quantitative Analysis 59, 2487–2514.

Ma, W., Ouimet, P., Simintzi, E., 2025. Mergers and acquisitions, technological change, and inequality. Journal of Financial Economics 172, 104136.

Mitchell, M. L., Mulherin, J. H., 1996. The impact of industry shocks on takeover and restructuring activity. Journal of Financial Economics 41, 193–229.

Nain, A., Yao, T., 2013. Mutual fund skill and the performance of corporate acquirers. Journal of Financial Economics 110, 437–456.

Schoch, D. S., 2020. Mergers and Acquisitions in the Data Economy.

Stock, J. H., Yogo, M., 2005. Testing for Weak Instruments in Linear IV Regression.

Tambe, P., 2014. Big Data Investment, Skills, and Firm Value. Management Science 60, 1452–1469.

Tambe, P., Hitt, L., Rock, D., Brynjolfsson, E., 2020. Digital capital and superstar firms.

Veldkamp, L., 2023. Valuing Data as an Asset. Review of Finance 27, 1545–1562.

Veldkamp, L., Chung, C., 2019. Data and the Aggregate Economy. Journal of Economic Literature .

Xiao, S., Wu, K., Xiao, X., Lin, C., 2025. How do data assets affect firm investment? International Review of Financial Analysis 106, 104526.

Xiao, Y., Liu, X., 2025. Data assetization and firms' investment efficiency: evidence from China. China Economic Journal 0, 1–33.

Zhu, C., 2019. Big Data as a Governance Mechanism. The Review of Financial Studies 32, 2021–2061.

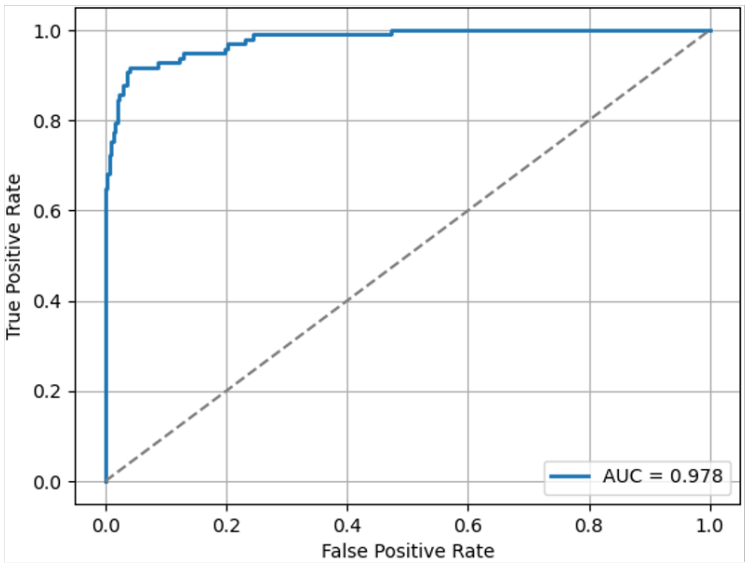
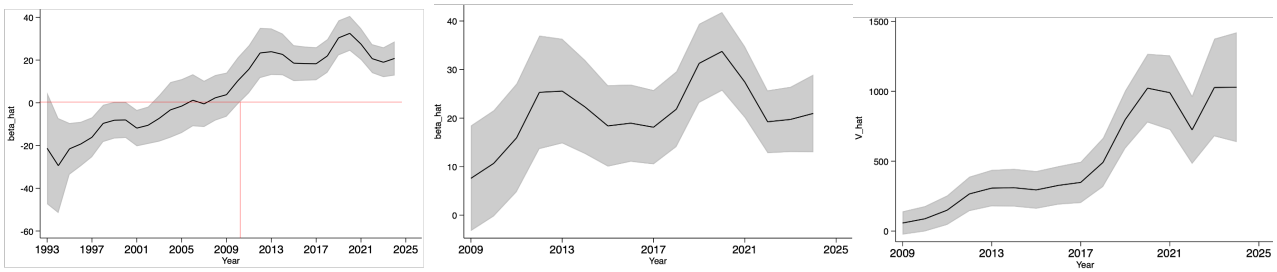
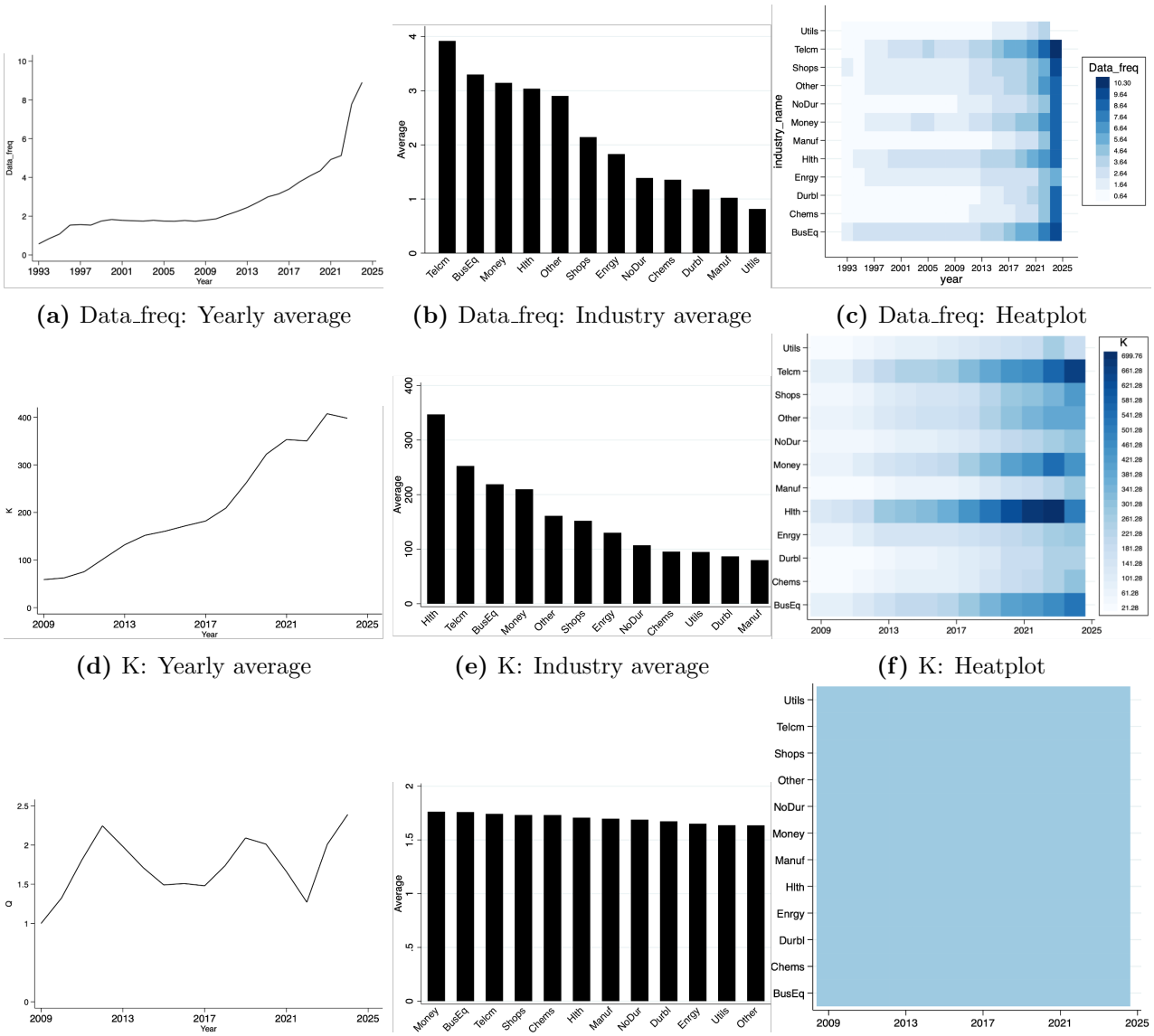


Figure 1: ROC curve for *FinBERT* data asset sentence classification. This figure plots the receiver operating characteristic (ROC) curve of the fine-tuned *FinBERT* classifier, trained to identify whether a sentence indicates the presence of firm-level data assets. The curve illustrates how the model's true positive rate and false positive rate vary across different classification thresholds in the test set. The area under the curve (AUC), ranging from 0 to 1, reflects the model's overall ability to distinguish between positive (data-asset-related) and negative sentences. A higher AUC indicates better performance. In settings with highly imbalanced data, where most sentences are unrelated to data assets (i.e., labeled as 0), the AUC is more informative than Accuracy.



(a) $\widehat{\beta}_{\text{Data_freq}}$: Full sample **(b)** $\widehat{\beta}_{\text{Data_freq}}$: Since 2009 **(c)** $\widehat{V}_{\text{Data_freq}}$: Since 2009

Figure 2: Predicted yearly *Data_freq* coefficients and associated market value. This figure plots estimates and 95% confidence intervals from yearly regressions of firm market capitalization on *Data_freq*. Panels (a) and (b) report the yearly estimated coefficient $\widehat{\beta}_{\text{Data_freq}}$ for the full sample and for the post-2009 period, respectively. Panel (c) shows the estimated yearly average market capitalization associated with *Data_freq*, denoted as $\widehat{V}_{\text{Data_freq}}$, since 2009.

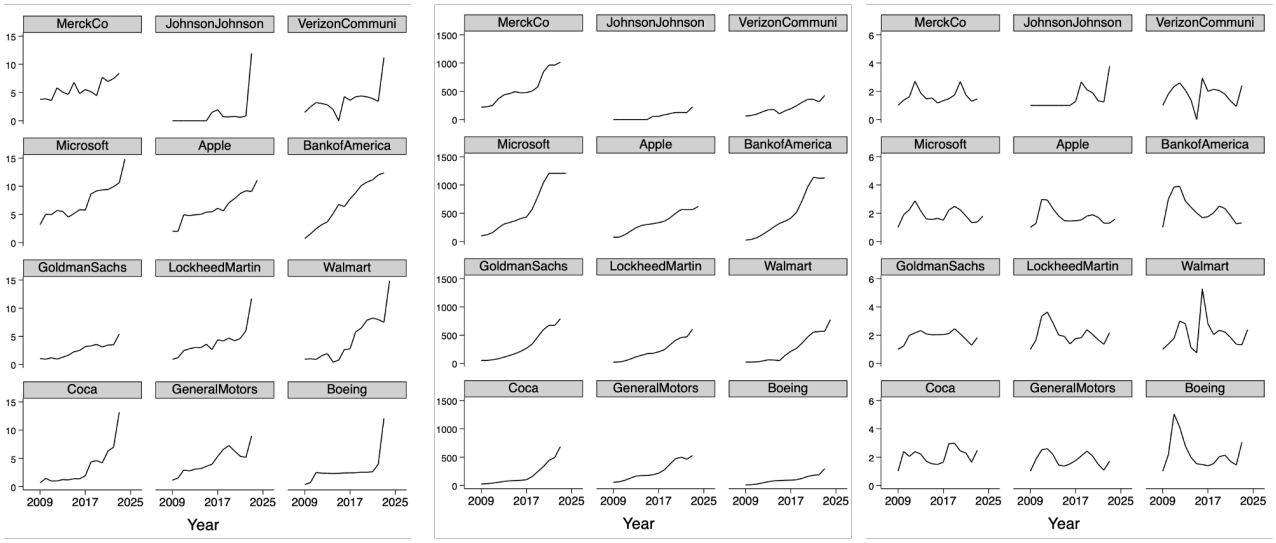


(a) Data_freq: Yearly average **(b)** Data_freq: Industry average **(c)** Data_freq: Heatplot

(d) K: Yearly average **(e)** K: Industry average **(f)** K: Heatplot

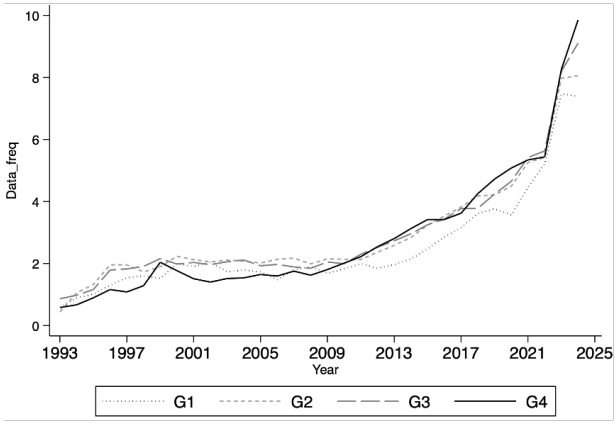
(g) Q: Yearly average **(h)** Q: Industry average **(i)** Q: Heatplot

Figure 3: Distribution of *Data_freq*, K, and Q. This figure plots the distributions of *Data_freq*, Data capital quantity (K), and Data capital market pricing (Q). For each variable, the first, second, and third subfigures in each row display the yearly average, the Fama–French 12 industry average, and the year–industry average, respectively.

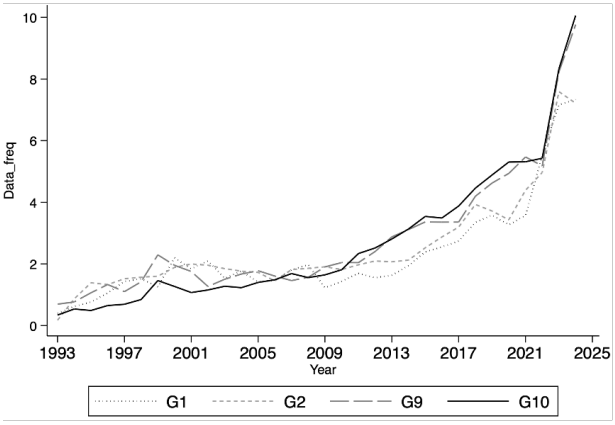


(a) Data_freq: Yearly average **(b)** Data_freq: Industry average **(c)** Data_freq: Heatplot

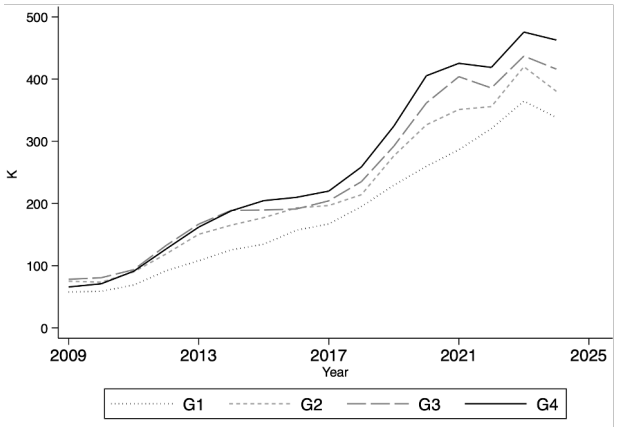
Figure 4: Example firms' *Data_freq*, K, and Q. This figure plots the yearly values of *Data_freq*, data capital quantity (K), and data capital market valuation (Q) for a set of example firms. These firms are drawn from various industries, including healthcare, telecommunications, business equipment, finance, defense, wholesale, and manufacturing.



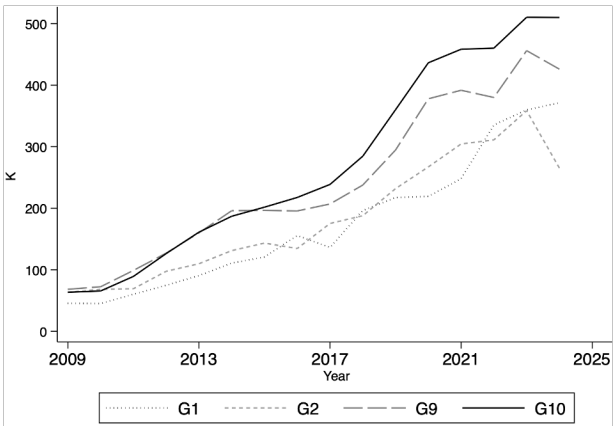
(a) Data_freq: 4 groups



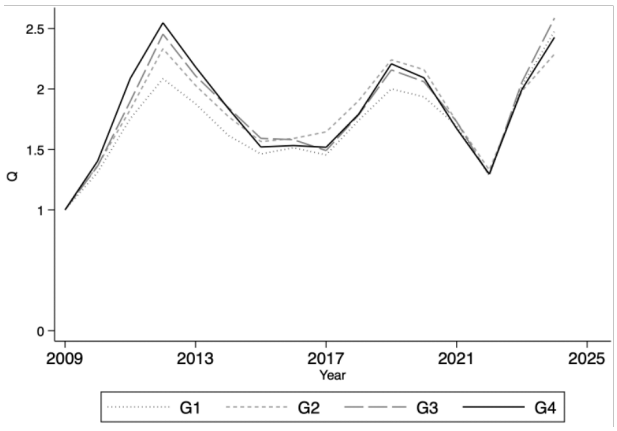
(b) Data_freq: 10 groups



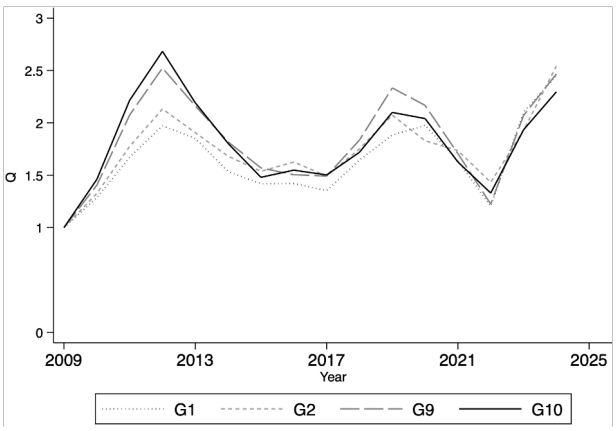
(c) K: 4 groups



(d) K: 10 groups



(e) Q: 4 groups



(f) Q: 10 groups

Figure 5: Distribution of *Data_freq*, K, and Q by market value groups. This figure plots the distributions of *Data_freq*, Data capital quantity (K), and Data capital market pricing (Q) by market value groups. The first and second subfigures in each row display firms grouped into four and ten categories, respectively, based on their market value.

Table 1: Summary statistics

This table presents variables' summary statistics (*N*, *mean*, *standard deviation*, *10th percentile*, *25th percentile*, *median*, *75th percentile*. All variables are defined in the Appendix Table **A1**. Panel A presents the Compustat data sample, and Panel B presents the SDC deal sample.

Panel A: Summary statistics: Compustat sample									
VarName	N	Mean	SD	P10	Q1	Median	Q3	P90	Winsorized
M_A	149527	0.14	0.34	0.00	0.00	0.00	0.00	1.00	N
Data_freq	149527	2.39	3.07	0.00	0.00	1.28	3.33	6.52	Y
K	45583	170.43	228.20	0.00	24.37	86.56	219.73	439.84	Y
Q	45583	1.68	0.88	1.00	1.00	1.53	2.07	2.71	Y
Control variables									
Size	89871	2.64	7.47	0.02	0.07	0.32	1.43	5.43	Y
ROA	142136	-0.12	0.88	-0.43	-0.02	0.08	0.14	0.21	Y
Leverage	147143	0.33	0.54	0.00	0.04	0.22	0.42	0.66	Y
Tobin_Q	84602	2.19	1.97	0.86	1.09	1.53	2.47	4.24	Y
Sales_growth	137843	0.20	0.84	-0.19	-0.02	0.04	0.20	0.52	Y
Profit_margin	135605	-1.22	7.02	-0.79	0.01	0.10	0.20	0.36	Y
M/B	89814	3.19	5.35	0.51	1.07	1.98	3.73	7.18	Y
Volatility	87708	0.56	0.34	0.24	0.33	0.47	0.70	0.99	Y
Past_return	87708	0.14	0.64	-0.51	-0.24	0.05	0.35	0.79	Y
Panel B: Summary statistics: SDC sample									
VarName	N	Mean	SD	P10	Q1	Median	Q3	P90	Winsorized
FrequentAcquirer	5444	0.25	0.43	0.00	0.00	0.00	1.00	1.00	N
#deal_3yr	5438	3.81	4.05	1.00	1.00	3.00	5.00	8.00	N
#deal_5yr	5438	5.27	6.32	1.00	2.00	3.00	6.00	11.00	N
CAR	5441	0.92	6.76	-5.60	-1.93	0.45	3.31	7.93	Y
Synergy	699	3.30	7.17	-4.13	-0.51	2.33	6.82	12.41	Y
Premium	959	48.73	50.60	6.35	20.76	37.30	62.65	104.82	Y
Control variables									
Rel_size	5444	0.20	0.47	0.00	0.02	0.05	0.18	0.49	Y
Earnout	5444	0.14	0.35	0.00	0.00	0.00	0.00	1.00	N
Tender	5444	0.05	0.22	0.00	0.00	0.00	0.00	0.00	N
Full_Stock	5444	0.05	0.22	0.00	0.00	0.00	0.00	0.00	N
Full_Cash	5444	0.61	0.49	0.00	0.00	1.00	1.00	1.00	N
Private	5444	0.77	0.42	0.00	1.00	1.00	1.00	1.00	N
Subsidiary	5444	0.02	0.15	0.00	0.00	0.00	0.00	0.00	N
Diversified	5444	0.43	0.50	0.00	0.00	0.00	1.00	1.00	N
Hostile	5444	0.00	0.04	0.00	0.00	0.00	0.00	0.00	N
#bidder	5444	0.02	0.12	0.00	0.00	0.00	0.00	0.00	N
Toehold	5444	0.02	0.14	0.00	0.00	0.00	0.00	0.00	N
ROA	5332	-0.00	0.19	-0.15	-0.01	0.04	0.08	0.13	Y
Tobins_Q	5441	2.35	1.74	1.01	1.30	1.83	2.71	4.25	Y
Leverage	5416	0.23	0.19	0.00	0.05	0.21	0.34	0.49	Y
FCF	5323	0.02	0.19	-0.12	0.02	0.07	0.10	0.15	Y
M/B	5441	4.18	5.47	0.89	1.51	2.67	4.63	8.87	Y
Runup	5444	11.57	41.59	-33.65	-7.46	13.20	33.45	56.41	Y
BidderSize	5444	17.03	48.51	0.10	0.42	1.59	7.33	36.49	Y

Table 2: M&A Sample distribution

This table presents the annual distribution of M&A announcements between 1 January 1980 and 31 December 2024. Panel A reports the full sample, while Panel B restricts to observations with non-missing deal control variables and includes only deals announced since 2009.

Year	Total	Public	Private	Subsidiary	Diversified	Cash	Stock	Completed	Withdrawn	Friendly	Earnout	Tender
Panel A: Full sample												
1980	12	3	6	3	10	2	4	11	1	11	0	2
1981	22	10	11	1	17	2	4	18	4	15	0	9
1982	10	3	6	1	9	1	1	8	2	7	0	3
1983	18	9	6	3	11	2	3	13	4	15	0	6
1984	42	22	17	3	30	8	6	33	5	40	0	9
1985	234	93	130	9	138	61	37	156	25	229	5	32
1986	361	100	236	23	342	97	45	280	33	353	2	32
1987	416	119	267	29	335	72	50	315	47	404	4	25
1988	393	132	238	21	211	90	27	297	45	379	3	44
1989	638	144	458	34	325	102	74	439	60	634	14	44
1990	626	114	473	38	303	86	52	495	53	626	20	29
1991	674	110	525	36	308	83	92	500	41	673	22	36
1992	881	83	751	44	379	124	137	694	59	875	20	25
1993	962	116	818	25	436	123	140	745	70	961	29	37
1994	1243	149	1029	60	593	187	203	1022	69	1232	30	39
1995	1495	203	1237	49	688	201	244	1219	67	1482	31	47
1996	1971	189	1680	98	886	247	329	1639	79	1960	29	35
1997	2500	210	2169	117	1104	277	339	2143	86	2492	56	56
1998	2848	251	2460	130	1143	336	304	2402	69	2845	77	55
1999	2268	251	1955	56	1043	279	326	1976	53	2257	56	68
2000	1849	221	1545	76	876	260	352	1581	60	1846	57	53
2001	1014	168	803	40	470	179	129	881	27	1013	38	44
2002	924	116	761	42	394	214	60	826	34	923	45	39
2003	844	121	690	28	336	208	50	770	18	841	61	37
2004	1071	104	922	43	412	318	50	966	17	1067	70	26
2005	1120	118	969	31	472	334	33	1032	16	1117	68	25
2006	1273	149	1075	45	554	375	26	1128	29	1272	81	26
2007	1242	135	1075	28	526	351	23	1118	20	1241	76	42
2008	1106	118	954	34	424	279	17	971	37	1102	54	41
2009	698	98	570	25	277	166	20	610	20	697	49	26
2010	917	107	764	44	391	210	7	839	15	916	50	27
2011	1049	69	933	45	427	256	12	952	17	1047	63	16
2012	1020	80	917	21	423	242	11	936	13	1019	59	22
2013	890	58	802	27	354	200	14	809	12	890	42	18
2014	1102	77	984	35	464	253	11	988	14	1101	59	27
2015	1054	117	921	13	488	269	19	919	17	1054	58	32
2016	938	75	845	17	439	238	9	806	11	936	36	21
2017	905	78	813	14	444	235	16	758	16	905	35	16
2018	976	106	849	18	466	259	22	816	11	975	54	16
2019	804	62	721	19	390	208	19	676	9	804	51	8
2020	831	69	742	16	440	209	33	661	15	831	52	18
2021	1343	88	1235	19	690	331	33	1059	21	1343	87	12
2022	1027	86	921	17	530	226	37	822	15	1027	54	12
2023	827	83	732	11	429	183	35	671	19	827	47	16
2024	811	75	720	15	396	190	32	624	12	810	37	6
Total	**43249**	**4889**	**36735**	**1503**	**19823**	**8573**	**3487**	**36624**	**1367**	**43094**	**1781**	**1259**
Panel B: Sample without missing deal controls and since 2009												
2009	298	70	217	8	121	148	18	273	14	297	49	25
2010	356	87	255	13	139	193	4	327	13	355	46	27
2011	374	54	303	17	146	208	8	345	12	372	61	16
2012	370	73	293	3	156	212	10	343	10	369	57	21
2013	314	55	247	11	121	174	12	291	6	314	41	18
2014	401	66	321	13	140	219	9	371	12	400	53	27
2015	385	92	286	5	148	253	15	336	15	385	52	31
2016	303	63	236	3	126	215	7	267	9	301	35	21
2017	318	68	248	2	147	221	15	274	13	318	32	15
2018	355	82	263	7	160	248	20	305	9	354	52	16
2019	280	52	222	6	119	190	16	236	7	280	44	8
2020	317	61	250	6	153	192	29	256	13	317	49	18
2021	501	76	416	8	252	306	26	396	19	501	72	12
2022	325	65	249	8	166	204	28	263	10	325	48	10
2023	262	76	180	5	128	172	29	215	15	262	43	16
2024	285	61	218	5	144	173	26	215	9	284	37	6
Total	**5444**	**1101**	**4204**	**120**	**2366**	**3328**	**272**	**4713**	**186**	**5434**	**771**	**287**

Table 3: data assets and being bidder in t0

This table shows the effect of a firm's data assets on the likelihood of becoming a bidder in the following year using the following linear probability model (LPM):

$$M_A = \beta_0 + \beta_1 X + \beta_2 Controls + u$$

X is indicated under column numbers. Control variables are without lags in the columns of Q. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics with robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: without any controls and fixed effects				
Y: being bidder in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0389***	0.0311***	0.0148***	0.0172***
	(26.6587)	(15.7713)	(8.6856)	(10.6012)
Controls* & FEs	N	N	N	N
N	69408	43060	43060	46611
Adjusted-R2	0.0131	0.0072	0.0016	0.0023
Panel B: with controls and fixed effects (year and industry)				
Y: being bidder in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0221***	0.0177***	0.0102***	0.0130***
	(7.4887)	(5.4575)	(3.6409)	(4.6749)
Size_l1	0.0077***	0.0074***	0.0075***	0.0079***
	(25.0524)	(22.6335)	(22.8549)	(25.0233)
ROA_l1	0.0408***	0.0394***	0.0353***	0.0248**
	(5.3261)	(3.2641)	(2.9742)	(2.2431)
Leverage_l1	0.0149*	0.0253***	0.0268***	0.0651***
	(1.8854)	(2.6136)	(2.7736)	(6.8099)
Tobin_Q_l1	-0.0081***	-0.0125***	-0.0111***	-0.0137***
	(-5.2959)	(-7.0972)	(-6.3932)	(-8.3153)
Sales_growth_l1	0.0026	0.0091**	0.0098**	0.0110***
	(0.9689)	(2.1794)	(2.3347)	(2.7246)
Profit_margin_l1	0.0010***	0.0005	0.0004	0.0006*
	(4.6629)	(1.4989)	(1.1371)	(1.7116)
M/B_l1	0.0007	0.0008	0.0010*	0.0016***
	(1.5843)	(1.5369)	(1.7634)	(3.0196)
Volatility_l1	-0.1883***	-0.2314***	-0.2316***	-0.2340***
	(-23.1502)	(-22.4619)	(-22.5325)	(-24.6463)
Past_return_l1	0.0438***	0.0476***	0.0464***	0.0192***
	(10.9492)	(10.4161)	(10.1940)	(4.6309)
Constant	0.2845***	0.3080***	0.3057***	0.3023***
	(43.1641)	(39.2458)	(39.0239)	(41.5142)
Controls*	Y	Y	Y	Y
Year & Industry Effect	Y	Y	Y	Y
N	29500	23051	23051	24562
Adjusted-R2	0.1003	0.0994	0.0986	0.1059
Panel C: with controls and fixed effects (year and firm)				
Y: being bidder in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0108	0.0076	0.0008	0.0086**
	(1.5314)	(1.2852)	(0.2472)	(2.7624)
Controls*	Y	Y	Y	Y
Year & Firm Effect	Y	Y	Y	Y
N	28999	22824	22824	24306
Adjusted-R2	0.2713	0.2704	0.2703	0.2666

Table 4: Size effect of Data asset

This table shows how firm size affects the effect of data assets on the likelihood of becoming a bidder in the following year using the following model:

$$M_A = \beta_0 + ... + \beta_3(X \times Size_Proxy) + \beta_4 Controls + u$$

Control variables are without lags in the columns of Q. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: Log(Mktcap) as the Size proxy				
Y: being bidder in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0136***	0.0030	0.0029	0.0068**
	(4.3144)	(0.8229)	(0.9819)	(2.5109)
Log(MktCap)_std	0.0973***	0.1062***	0.1051***	0.0996***
	(26.7989)	(24.2486)	(24.5908)	(25.0972)
Proxy_l1_std x Log(MktCap)_std	0.0091***	0.0091***	0.0055**	0.0071***
	(3.5878)	(3.3021)	(2.0647)	(2.9060)
Controls*	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y
N	24978	18963	18963	24562
Adjusted-R2	0.1168	0.1167	0.1162	0.1247
Panel B: Log(Asset) as the Size proxy				
Y: being bidder in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0175***	0.0066**	0.0036	0.0068**
	(6.1553)	(2.1016)	(1.3497)	(2.5736)
Log(Asset)_std	0.1282***	0.1332***	0.1329***	0.1275***
	(26.3428)	(23.7687)	(23.9431)	(22.7904)
Proxy_l1_std x Log(Asset)_std	0.0081**	0.0129***	0.0063*	0.0055
	(2.3184)	(3.5168)	(1.7897)	(1.5443)
Controls*	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y
N	29494	23046	23046	24562
Adjusted-R2	0.1171	0.1166	0.1159	0.1210

Table 5: Instrumental variable two-stage least squares regressions

This table reports the 2SLS results using *AI_exposure_2010*, *njob_data* and *njob_AI* as instrumental variables, the second stage:

$$M_A = \beta_0 + \beta_2 \widehat{K_l1} + \beta_3 Controls + u$$

The first stage is *K_l1* on the instrument(s):

$$K_l1 = \beta_0 + \beta_1 Instrument\ variable(s) + \beta_2 Controls + u$$

The sample prior to 2010 is excluded to preserve the ex-ante nature of the first instrument. The results remain robust when using the full sample and when excluding 2010. Robustness is also confirmed when *K* is instrumented separately with each of the three instruments. See the variable definitions in the Appendix Table **A1**. Control variables are those in Panel B Table **3**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

year >= 2010	(1) 1st-stage	(2) 1st-stage	(3) 2nd-stage	(4) 2nd-stage
$\widehat{K_l1_std}$			0.1628***	0.3485***
			(4.4487)	(5.4047)
AI_exposure_2010	0.0863***	0.0443***		
	(11.9020)	(3.8953)		
njob_AI_l1_std		0.0546***		
		(3.7464)		
njob_data_l1_std		0.0638***		
		(5.0474)		
Under-identification: Kleibergen-Paap LM	-	-	142.0659	111.2262
Under-identification: P-value	-	-	(0.0000)	(0.0000)
Weak-identification: Kleibergen-Paap F	-	-	141.6570	33.7556
Over-identification: Hansen J	-	-	-	1.8351
Over-identification: P-value	-	-	-	(0.3995)
Controls	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y
N	20784	8434	20784	8434

Table 6: Eat or be eaten

This table shows how industry data assets concentration affects the effect of data assets on the likelihood of becoming a bidder in the following year using the following model:

$$M_A = \beta_0 + ... + \beta_3(K \times pct_Data_firms) + \beta_4 Controls + u$$

where *pct_Data_firms* is indicated under column numbers and refers to one of *pct_K_lt_5*, *pct_K_5_30*, or *pct_K_gt_30*, indicating the percentage of firms whose data capital K falls within [0%, 5%), [5%, 30%], and (30%, 100%] of the largest data capital value within each industry, respectively. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Y: being bidder in t0	(1) pct_K_lt_5	(2) pct_K_5_30	(3) pct_K_gt_30
K_l1_std	0.0318***	0.0214***	0.0291***
	(5.7177)	(6.2087)	(6.6031)
K_l1_std x Proxy_std	0.0145***	0.0069***	-0.0080***
	(3.1840)	(3.5579)	(-4.0290)
Controls	Y	Y	Y
YearEffect	Y	Y	Y
IndustryEffect	Y	Y	Y
N	23043	23043	23043
Adjusted-R2	0.0998	0.0999	0.1000

Table 7: data assets improves information environment, liquidity, and cost of capital

This table shows how the baseline effect of a firm's data assets works through improving the firm's information environment, liquidity, and cost of capital using the following models:

$$Mecha = \beta_0 + \beta_1 K + \beta_2 Controls + u$$

$$M_A = \beta_0 + \beta_1 Mecha + \beta_2 Controls + u$$

$$M_A = \beta_0 + \beta_2 \widehat{Mecha} + \beta_3 Controls + u$$

where *Mecha* represents measures of a firm's information environment (including a dummy for whether management provides earnings guidance, institutional ownership, analyst coverage, and noise share in the stock price); liquidity (Amihud illiquidity and bid–ask spread); and cost of capital (cost of equity and cost of debt). Results from the first, second, and third equations are reported in Panels A, B and C, respectively. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: Mecha on X								
Y: Mecha_std:	(1) Guidance	(2) Institution%	(3) #Analysts	(4) Noise%	(5) Amihud	(6) Spread	(7) CoE	(8) CoD
K_l1_std	0.0944***	0.1426***	0.0916***	-0.0653***	-0.0592***	-0.0191***	-0.1462***	-0.0006
	(7.9716)	(22.1346)	(13.0203)	(-10.6977)	(-9.0799)	(-4.9133)	(-18.1457)	(-0.3230)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y	Y	Y	Y
N	23051	16696	16110	18953	18959	18959	14907	21490
Adjusted-R2	0.1965	0.2913	0.5425	0.1011	0.0573	0.6221	0.3823	0.3716
Panel B: Y on Mecha								
Y: being bidder in t_0	(1) Guidance	(2) Institution%	(3) #Analysts	(4) Noise%	(5) Amihud	(6) Spread	(7) CoE	(8) CoD
Mecha_std	0.0254***	0.0600***	0.0390***	-0.0210***	-0.0114***	-0.0597***	-0.0506***	-0.0239***
	(15.1185)	(19.9076)	(10.6836)	(-8.2776)	(-7.1057)	(-13.2496)	(-18.3429)	(-3.0335)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y	Y	Y	Y
N	31239	22418	22477	26156	26165	26165	19732	29038
Adjusted-R2	0.1034	0.1109	0.0977	0.0948	0.0938	0.0989	0.0988	0.0969
Panel C: Y on $\widehat{Mecha}$ instrumented by K								
Y: being bidder in t_0	(1) Guidance	(2) Institution%	(3) #Analysts	(4) Noise%	(5) Amihud	(6) Spread	(7) CoE	(8) CoD
$\widehat{Mecha_std}$	0.1873***	0.1500***	0.1658***	-0.2694***	-0.2978***	-0.9248***	-0.1511***	-28.3982
	(4.7286)	(5.3933)	(3.7035)	(-4.4627)	(-4.3249)	(-3.6164)	(-4.9049)	(-0.3226)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y	Y	Y	Y
N	23051	16696	16110	18953	18959	18959	14907	21490
1st-stage F	63.5472	489.9391	169.5273	114.4412	82.4450	24.1408	329.2646	0.1043

Table 8: data assets and deal-level outcomes

This table shows the effect of a firm's data assets on deal-level outcomes using the following ordinary least squares (OLS), LPM, or Poisson models:

$$SDC_Ys = \beta_0 + \beta_2 K + \beta_3 Controls + u$$

Regressions are estimated using OLS for continuous dependent variables, LPM for binary dependent variables, and Poisson models for count-type dependent variables. *SDC_Ys* are indicated under the column numbers. Control variables include a series of deal and acquirer characteristics. In each column, the corresponding dependent variable is excluded from the set of controls if it is otherwise included as a control variable. Results are robust without control variables. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: Frequent acquirer and market reactions					
Y:	(1) FrequentAcquirer	(2) #deal_3yr	(3) #deal_5yr	(4) CAR	(5) Synergy
K_std	0.0169*	0.0517***	0.0649***	-0.3389**	-0.6982*
	(1.7606)	(2.8347)	(3.3703)	(-2.0402)	(-1.7211)
Controls*	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y
IndustryEffect	Y	N	Y	Y	Y
N	3858	3857	3857	3858	547
Adjusted/Pseudo-R2	0.1890	0.2134	0.2854	0.0285	0.1853
Panel B: Offer premium and deal characteristics					
Y:	(1) Premium	(2) Private	(3) Earnout	(4) Rel_size	(5) Diversified
K_std	5.8658*	-0.0149*	0.0305***	-0.0203***	-0.0635***
	(1.8792)	(-1.7168)	(3.5267)	(-3.2654)	(-5.7703)
Controls*	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y
N	733	3858	3858	3858	3858
Adjusted-R2	0.1272	0.3658	0.0831	0.1767	0.1389

Appendix Table A1: Variables, Definitions, and Sources

Variable	Definition	Source
Compustat sample		
M_A	Dummy variable equals to one if a firm become a bidder within the year.	SDC
Data_freq	The number of data-asset-related sentences in Item 1s of firm i's 10-K in year t over the total number of sentences in the same section.	Bill McDonald's website & Authors' Estimations
K	Model-implied quantity of data capital.	Authors' Estimations
Q	Model-implied quantity of data capital.	Authors' Estimations
Size	Total market capitalization, measured in billion.	CRSP
ROA	Return on assets = EBITDA / total assets.	Compustat
Leverage	(Long debt + short debt) / total assets.	Compustat
Tobin's Q	(Total assets + market cap + value of common stock - deferred taxes) / total assets.	CRSP & Compustat
Sales_growth	$(Sales_{i,t}$ / $Sales_{i,t-1})$ - 1	Compustat
Profit_margin	EBITDA / sales	Compustat
M/B	Market-to-book ratio = market cap / book equity. Market equity is at the fiscal year-end.	CRSP & Compustat
Volatility	Realized volatility over the past fiscal year.	CRSP
Past_return	Realized return over the past fiscal year.	CRSP
pct_Data_firms	pct_K_lt_5, pct_K_5_30, and pct_K_gt_30, indicating the percentage of firms whose data capital K falls within [0%, 5%), [5%, 30%], and (30%, 100%] of the largest data capital value within each industry.	Compustat & Authors' Estimations
Guidance	Dummy = 1 if the firm management provide a voluntary EPS forecast.	I/B/E/S
Institutional%	The percentage of shares held by institutions to the total number of shares outstanding before the announcement.	Thomson Reuters 13f Holdings database
#analyst	The number of analysts following the firm in the past year.	I/B/E/S
noise%	Noise share in price Brogaard et al. (2022).	CRSP & Authors' Estimations
Amihud	Amihud illiquidity: absolute daily returns / daily dollar trading volume.	CRSP
Spread	(bid - ask) / mid point of bid and ask, using the yearly average over the past fiscal year.	CRSP
CoE	Cost of equity: Proxies from Lee et al. (2021).	Authors' Website
CoD	Cost of debt: Total interest expenses / total assets × 100.	Compustat
SDC sample		
FrequentAcquirer	Dummy variable equals to one if a firm has >= 5 deals in a 3-year period.	SDC
#deal_3yr	The number of deals within the 3-year period before the deal announcement day.	SDC
#deal_5yr	The number of deals within the 5-year period before the deal announcement day.	SDC
CAR	Market-adjusted return of the bidder for the 5-day centered on the announcement date of the current deal.	CRSP & Authors' Estimations
Synergy	The combined bidder and target CAR weighted by market value 2 days before the announcement date.	CRSP & Authors' Estimations
Premium	Percentage difference between the offer price and the target price (30 days before the deal announcement date).	SDC
Relative Size	Deal value divided by the acquirer's pre-acquisition market value.	SDC
Earnout	Dummy = 1 if the deal includes an earnout, otherwise 0.	SDC
Tender	Dummy = 1 if the deal is a tender offer.	SDC
Full_Stock	Dummy = 1 if the deal is 100% paid by stock.	SDC
Full_Cash	Dummy = 1 if the deal is 100% paid by cash.	SDC
Private	Dummy = 1 if the target is a private firm.	SDC
Subsidiary	Dummy = 1 if the target is a subsidiary.	SDC

Table Continued

Appendix Table **A1**: **Variables, Definitions, and Sources (Continued)**

Variable	Definition	Source
Diversified	Dummy = 1 if the acquirer and target have different two-digit SIC codes.	SDC
Hostile	Dummy = 1 if the acquirer's attitude is hostile.	SDC
#bidders	The number of bidders for the target.	SDC
Toehold	The percentage of target shares held by the acquirer six months before the acquisition.	SDC
ROA	Return on assets = EBITDA / total assets.	Compustat
Tobin's Q	(Total assets + market cap + value of common stock - deferred taxes) / total assets.	CRSP & Compustat
Leverage	(Long debt + short debt) / total assets.	Compustat
FCF	Free cash flow over the total asset, where the free cash flow is income before extraordinary items + depreciation and amortization - common dividends - preferred dividends.	Compustat
M/B	Market-to-book ratio = market cap / book equity. Market equity is at the fiscal year-end.	CRSP & Compustat
Runup%	Cumulative log return from 365 days to 2 trading days before the announcement.	CRSP
BidderSize	Total market capitalization, measured in billion.	CRSP

Appendix Table A2: Other Data asset measures

This table shows the effect of a firm's data asset on the likelihood of becoming a bidder in the following year using other proxies of data asset in the following model:

$$M_A = \beta_0 + \beta_1 Data_proxy + \beta_2 Controls + u$$

Results using sample since 2009 and full period sample are reported in Panels A and B, respectively. *Data_proxy* is indicated under the column numbers and corresponds to the following measures: (1) *Data_freq*; (2) the share of data skill–related job postings at the firm level; (3) model-implied data capital based on the number of such postings; (4) the data effect derived from management EPS forecasts; and (5) the data effect derived from management sales forecasts. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics with robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: keep sample since 2009					
Y: being bidder in t_0	(1) Data_freq	(2) share_data	(3) K_labor	(4) EPS_Forecast	(5) Sale_Forecast
Proxy_l1_std	0.0221***	0.0184***	0.0162**	0.0229***	0.0100*
	(7.4887)	(4.2483)	(2.1982)	(5.0806)	(1.9177)
Controls	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y
N	29500	12046	7546	9030	11441
Adjusted-R2	0.1003	0.0948	0.0995	0.0915	0.1010
Panel B: keep full period sample with non-missing data measures					
Y: being bidder in t_0	(1) Data_freq	(2) share_data	(3) K_labor	(4) EPS_Forecast	(5) Sale_Forecast
Proxy_l1_std	0.0154***	0.0167***	0.0162**	0.0284***	0.0091**
	(7.6974)	(4.1086)	(2.1982)	(8.9112)	(2.2345)
Controls	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y
N	62255	13182	7546	16114	15358
Adjusted-R2	0.0881	0.0941	0.0995	0.0841	0.0981

Appendix Table A3: Data asset and being target in t0

This table shows the effect of a firm's data asset on the likelihood of becoming a target in the following year using the following model:

$$M_A_target = \beta_0 + \beta_1 X + \beta_2 Controls + u$$

$$M_A_target = \beta_0 + ... + \beta_3 (X \times Year_2009) + \beta_4 Controls + u$$

where *Year_2009* is a dummy variable equal to one for years 2009 and onward. X is indicated under column numbers and refers to either previous used Xs or one of two management forecast–based data asset measures. Results from the first equation (sample since 2009) and the second equation (full period sample) are reported in Panels A and B, respectively. See the variable definitions in the Appendix Table **A1**. Numbers in parentheses are t-statistics based on robust standard errors. Results are robust to clustering by firmID or year & firmID. ***, **, and * denote the significance levels of 1%, 5%, and 10%, respectively.

Panel A: keep sample since 2009				
Y: being target in t_0	(1) Data_freq_l1	(2) K_l1	(3) Q_l1	(4) Q
Proxy_std	0.0002 (0.2666)	0.0009 (1.1650)	-0.0003 (-0.4859)	-0.0004 (-0.6187)
Controls*	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y
N	29483	23053	23053	24562
Adjusted-R2	0.0007	0.0013	0.0012	0.0020

Panel B: keep full period sample						
Y: being target in t_0	(1) Data_freq	(2) EPS_Forecast	(3) Sale_Forecast	(4) Data_freq	(5) EPS_Forecast	(6) Sale_Forecast
Proxy_l1_std	0.0006 (1.1961)	-0.0006 (-0.6471)	-0.0005 (-0.4316)	0.0017* (1.9002)	0.0016*** (2.5773)	0.0002 (0.1664)
Proxy_l1_std x Year_2009				-0.0018* (-1.7905)	-0.0045*** (-2.6219)	-0.0011 (-0.5498)
Controls	Y	Y	Y	Y	Y	Y
YearEffect	Y	Y	Y	Y	Y	Y
IndustryEffect	Y	Y	Y	Y	Y	Y
N	62250	16114	15358	62250	16114	15358
Adjusted-R2	0.0016	-0.0002	0.0019	0.0016	0.0003	0.0018

Internet Appendix

This Internet Appendix contains derivations of the unique root in the model, and additional tables for robustness checks and subsample analyses.

IA.1. Proof: the equality of average Q and marginal Q

The goal is to show that the average Q equals the marginal Q, and to derive the final system of equations under our assumptions.

From Hall (2001), Hayashi (1982), and Tambe et al. (2020): under the assumption of a zero-rent economy and by applying Euler's Theorem for homogeneous functions of degree 1:

$$x\frac{\partial f(x)}{\partial x} = f(x) \Rightarrow K_t\frac{\partial V_t(K_t)}{\partial K_t} = V_t(K_t)$$

$$\begin{aligned}
\text{Then, } Q_{\text{average}} &= \frac{V_t}{K_t} \\
&= \frac{1}{K_t}V_t \\
&= \frac{1}{K_t}K_t\frac{\partial V_t(K_t)}{\partial K_t} \\
&= \frac{\partial V_t(K_t)}{\partial K_t} = Q_{\text{marginal}}
\end{aligned}$$

$$\Rightarrow Q_{\text{average}} = Q_{\text{marginal}}$$

IA.2. Proof: the unique root of the model

The goal is to show that the function equating average Q and marginal Q has a unique root(i.e., the function $x(K_t)$ = $Q_{marginal}$ - $Q_{average}$ has a unique roots). We establish this by showing that $x(K_t)$: 1) is monotonically increasing and 2) crosses zero within the domain between $\min[(1-\delta), K_{t-1}\ V_t]$, and $\max[(1-\delta)K_{t-1},\ V_t]$.

1): $x(K_t)$ is an increasing function if its first derivative is positive

$$\begin{aligned} x(K_t) &= 1 + \alpha \left[\frac{K_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] - \frac{V_t}{K_t} \\ &= \alpha \left[\frac{K_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 1 - \frac{V_t}{K_t} \\ x'(K_t) &= \frac{\alpha}{K_{t-1}} + \frac{V_t}{K_t^2} \end{aligned}$$

Given α, V_t, K_{t-1}, K_t^2 are all positive, then $x'(K_t) > 0$. Hence, $x(K_t)$ is an increasing function.

2) The function has crosses zero in the domain: $\{\min[(1-\delta)K_{t-1},\ V_t],\ \max[(1-\delta)K_{t-1},\ V_t]\}$:

if $(1-\delta)K_{t-1} < V_t : \frac{V_t}{(1-\delta)K_{t-1}} > 1, V_t - (1-\delta)K_{t-1} > 0$

$$\begin{aligned} x((1-\delta)K_{t-1}) &= \alpha \left[\frac{(1-\delta)K_{t-1} - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 1 - \frac{V_t}{(1-\delta)K_{t-1}} \\ &= \alpha \left[\frac{0}{K_{t-1}} \right] + 1 - \frac{V_t}{(1-\delta)K_{t-1}} \\ &= 1 - \frac{V_t}{(1-\delta)K_{t-1}} < 0 \\ x(V_t) &= \alpha \left[\frac{V_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 1 - \frac{V_t}{V_t} \\ &= \alpha \left[\frac{V_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 0 \\ &= \alpha \left[\frac{V_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] > 0 \end{aligned}$$

if $(1-\delta)K_{t-1} > V_t : \frac{V_t}{(1-\delta)K_{t-1}} < 1, V_t - (1-\delta)K_{t-1} < 0$

$$\begin{aligned} x((1-\delta)K_{t-1}) &= \alpha \left[\frac{(1-\delta)K_{t-1} - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 1 - \frac{V_t}{(1-\delta)K_{t-1}} \\ &= 1 - \frac{V_t}{(1-\delta)K_{t-1}} > 0 \\ x(V_t) &= \alpha \left[\frac{V_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] + 1 - \frac{V_t}{V_t} \\ &= \alpha \left[\frac{V_t - (1-\delta)K_{t-1}}{K_{t-1}} \right] < 0 \end{aligned}$$

In either case, the monotonically increasing function $x(K_t)$ has crosses the zero line from the low to high bound of the domain $\{\min[(1-\delta)K_{t-1},\ V_t],\ \max[(1-\delta)K_{t-1},\ V_t]\}$. Then there is a unique root K_t between $(1-\delta)K_{t-1}$ and V_t.

IA.3. Robust checks

3.1. Robust: models' parameters

According to Hall (2001) and Tambe et al. (2020), the model-generated data capital K is not overly sensitive to the parameter choices for the adjustment-cost parameter α, the initial capital K_0, and the depreciation parameter δ. More importantly, alternative parameter values primarily change the scale of K rather than the shape of its time path or its between- and within-firm variation.

Figure 6 presents the distributions of data capital quantity (K) and price (Q) under alternative parameter values. Panels (a) and (b) vary the adjustment-cost parameter α (1.5, 2.0, 3.0, 6.0, 8.0). Panels (c) and (d) vary the initial data capital K_0 (0, 40% of PPE, 60% of PPE, 80% of PPE, 100% of PPE). Panels (e) and (f) vary the depreciation rate δ (0, 5%, 10%, 40%, 70%). Consistent with Hall (2001) and Tambe et al. (2020), the resulting distributions either converge quickly or remain very similar patterns across parameter choices, especially for the data quantity K.

Internet Appendix Table IA1: Keyword List and sample sentences

Panel A: Data-asset-related keyword list	
Words	"data", "files", "registries", "platform", "cloud", "repository", "intangible", "digital", "analytic", artificial\s*intelligen, machine\s*learning, "security", "cybersecurity", "privacy", "retriev", "backup", "archiv", "query", "access", "record", "collect"

Panel B: Examples of data-asset-related sentences	
Code	Sentences
0	see note 17 quarterly financial data to our consolidated financial statements included elsewhere in this report, for further discussion.
0	if we are unable to source investments, access financing or manage future growth effectively, we may be unable to achieve our investment objective.
0	as of may 2, 1994, there were 4,942 holders of record of the common stock of the company.
1	privacy risks relating to our clients' actions or inactions on behalf of certain of our clients using some of our services, we collect and store information derived from the activities of website visitors and their devices
1	as we grow viewership, we are creating a large, valuable data base that we use to better understand what our viewers watch and how they engage with advertising.
1	with our existing product line, we have gained some limited experience in data mining and believe it offers another revenue source to be able to obtain information from consumers using the jabbermonkey site and sell such information to companies that would be able to use the information in their advertising or other business needs.
1	deterring data suppliers from supplying us data or customers from using our services, or changing consumer behavior adversely affecting our technology s market coverage.
1	we have reviewed publicly available data from those studies and incorporated the results of our analysis into the design of our apex study, but we could have misinterpreted the data or performed a flawed analysis.
1	we and any third-party collaborator who we engage to develop companion diagnostics will work cooperatively to generate the data required for submission with the pma application, and will remain in contact with the center for devices and radiological health, or cdrh, at the fda to ensure that any changes in requirements are incorporated into the development plans.
1	as a provider of security solutions designed to provide content security by protecting an organization s data and users, we may face cyber-attacks that attempt to penetrate our network security, including our data centers, to sabotage or otherwise disable our products and services, misappropriate our or our customers proprietary information, which may include personally identifiable information, or cause interruptions of our internal systems and services.
1	the remainder of the proceeds will be used to fund capital expenditures to support our data center expansion plans and to provide working capital.

Internet Appendix Table IA2: FinBERT performance

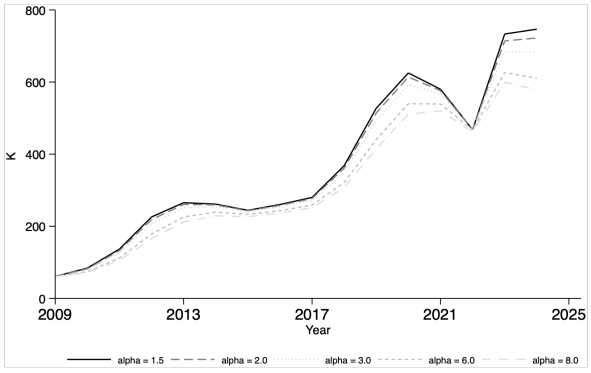
This table reports the prediction performance of the fine-tuned FinBERT model in classifying data-asset-related sentences within the testing sample. The testing sample has 400 manually labeled sentences, randomly drawn from a larger pool of 4,000 sentences. Among them, 97 are labeled as data-asset-related (1), and 303 as non-data-asset-related (0). For each sentence category, we report three standard performance metrics: precision, recall, and F1-score. For the full testing sample, we also report AUC, overall accuracy, macro average, and weighted average. AUC, ranging from 0 to 1, reflects the model's overall ability to distinguish between positive (data-asset-related) and negative sentences. A higher AUC indicates better performance. Overall accuracy is defined as the number of correctly classified sentences divided by the total number of sentences in the testing sample. The macro average is the unweighted mean across both categories and does not account for class imbalance. In contrast, the weighted average accounts for label imbalance by weighting each class metric by its support (i.e., sample size). Precision is defined as true positives divided by the sum of true positives and false positives. Recall is defined as true positives divided by the sum of true positives and false negatives. The F1-score is the harmonic mean of precision and recall, calculated as 2 × (precision × recall) / (precision + recall).

	Precision	**Recall**	**F1-score**	**#Sentence**
0	0.9729	0.9472	0.9599	303
1	0.8476	0.9175	0.8812	97
AUC			**0.9780**	400
Overall accuracy			**0.9400**	400
Macro Average	0.9103	0.9324	0.9205	400
Weighted Average	0.9425	0.9400	0.9408	400

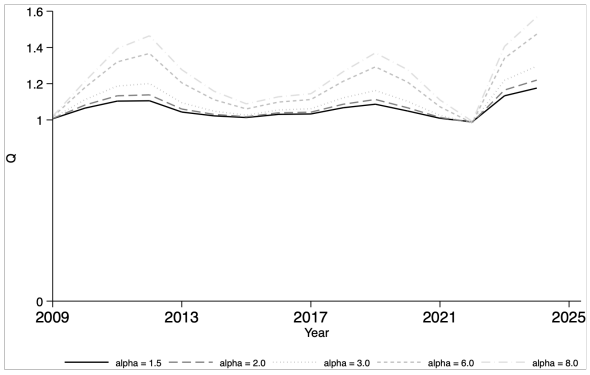
Internet Appendix Table IA3: Variance Decomposition of data asset

The table shows the results on the incremental Adjusted-R2 from a projection of the data asset measures on various sets of fixed effects.

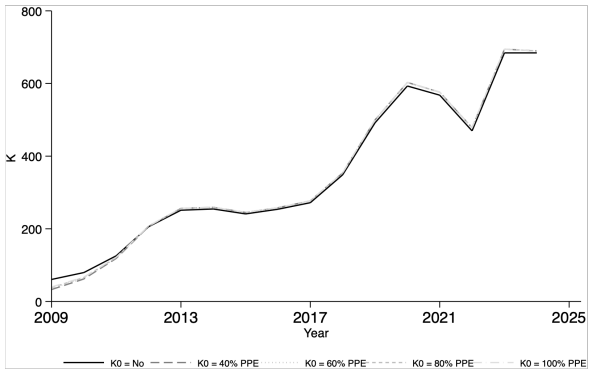
Var-decomp by Adj-R^2	Data_freq,1993	Data_freq,2009	K,2009	Q,2009
year	21.29%	21.61%	19.29%	16.57%
ff_48_in	11.39%	8.83%	16.27%	0.50%
ff_48_in x year extra	0.17%	-0.13%	2.59%	0.74%
Left to "Firm-level"	**67.15%**	**69.69%**	**61.85%**	**82.19%**
Permanent differences across firms within industries (Firm FE)	**47.15%**	**50.65%**	**45.18%**	13.04%
Variation over time in identity of firms within industries (Residual)	20.00%	19.04%	16.67%	**69.15%**



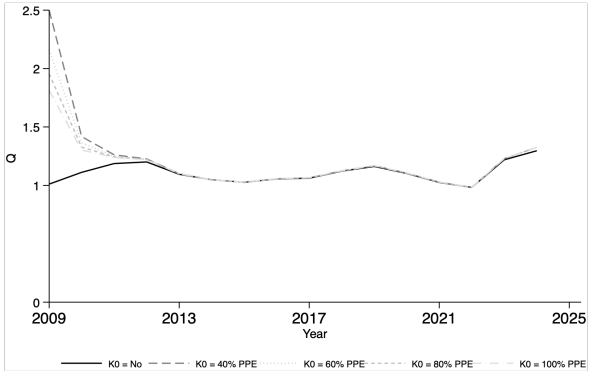
(a) K by Adjustment costs (α)



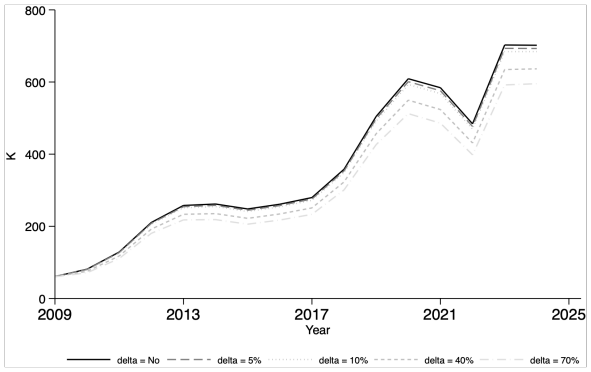
(b) Q by Adjustment costs (α)



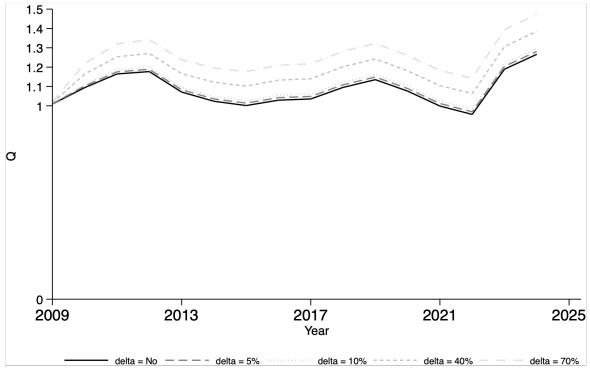
(c) K by Firms' initial data capital levels (K_0)



(d) Q by Firms' initial data capital levels (K_0)



(e) K by data capital depreciation rate (δ)



(f) Q by data capital depreciation rate (δ)

Figure 6: Distribution of K and Q after perturbing model parameters. This figure shows the distributions of data capital quantity (K) and price (Q) under alternative parameter values. Panels (a), (c), and (e) plot K, and panels (b), (d), and (f) plot Q. Panels (a) and (b) vary the adjustment cost parameter α (1.5, 2.0, 3.0, 6.0, 8.0). Panels (c) and (d) vary initial data capital K_0 (0, 40% of PPE, 60% of PPE, 80% of PPE, 100% of PPE). Panels (e) and (f) vary the depreciation rate δ (0, 5% of PPE, 10% of PPE, 40% of PPE, 70% of PPE).