When Data Fades: Data Investment, Depreciation, and Firm Market Value

Hongyuan Xia
hx276@cornell.edu

Prasanna Tambe
tambe@wharton.upenn.edu

March 22, 2026

Abstract

This paper studies how the depreciation of firms' data assets affects their market value. Despite growing recognition that data is a valuable intangible asset, only a small body of empirical work has quantified its financial value and very little prior work has examined how the depreciation of these assets moderates that value. We develop a novel firm-level measure of data depreciation based on the allocation of data labor between maintenance and analysis, using employee resume data merged with financial data on U.S. public firms from 2013 to 2023. We find that data investment is positively associated with market value, but this value is significantly discounted when data depreciates more rapidly. These effects are stronger for cloud-intensive firms and older firms. Exploiting variation from GDPR and data breach events, we provide evidence consistent with a causal interpretation. Overall, our results highlight that the value of data depends not only on its accumulation but also on its longevity.

Keywords: data value, IT management, human capital, IT intangibles

"The world's most valuable resource is no longer oil, but data."

1 Introduction

Production increasingly depends on data (Brynjolfsson and McElheran (2016) and Veldkamp and Chung (2024)). Many firms, particularly the most valuable U.S. companies, derive a substantial share of their market value from the data they have accumulated.[1] This trend has become especially pronounced since 2012, when deep learning—highly dependent on large volumes of high-quality data—began generating significant commercial value (McElheran et al. (2024)).

Despite widespread discussion of the importance of data accumulation, much less attention, both in academia and in industry, has been paid to data depreciation. Like physical capital, data can lose relevance, accuracy, or predictive power over time, a process we refer to as data depreciation. While related ideas appear in the machine learning literature under the notion of "data drift" or "concept drift," the economic implications of data depreciation for firms have received limited attention (Lu et al. 2018). The rate at which data depreciates is economically important because it determines how persistent the value of data investments is.[2] Similar to other intangible assets such as R&D or innovation capital, depreciation governs how quickly past investments lose their productive value and therefore shapes the process of creative destruction (Aghion and Howitt 1992; Hall 2007; Peters and Taylor 2017). A firm may accumulate large amounts of data, but if that data depreciates rapidly, the economic

1. For example, an article from the MIT Center for Transportation & Logistics notes that for firms like Google and Amazon, data has become more valuable than physical assets or even people. Available at: https://medium.com/mitsupplychain/what-is-a-companys-most-valuable-asset-not-people-4e5b854f713d.

2. A well-known example comes from Zillow's home price algorithm, "Zestimate," which relied on historical housing data. When housing market conditions changed rapidly during the COVID-19 pandemic, the data underlying the model quickly became outdated, contributing to systematic mispricing in Zillow's automated home-buying program (Zillow Offers), which the company shut down in 2021 after incurring substantial losses. See https://www.wired.com/story/zillow-ibuyer-real-estate/.

value of these data investments will be correspondingly lower. To study this phenomenon, this paper develops a proxy for the rate of data depreciation and examines how it affects firms' ability to capture value from data.

Estimating the rate of data depreciation is challenging for several reasons. First, the stock of data held by a firm is inherently difficult to observe and measure, as firms rarely disclose their data inventories. Without a reliable measure of the underlying data stock, it is difficult to track the value of data over time and estimate how quickly it depreciates. Second, there is substantial heterogeneity in the types of data firms collect—ranging from customer behavior data to operational and sensor data—and these data are likely to depreciate at different rates. This heterogeneity makes it difficult to apply traditional approaches that assume a uniform depreciation rate for intangible assets such as R&D. Third, data depreciation depends not only on the intrinsic properties of the data itself but also on how firms maintain and manage their data assets. Firms differ in their investments in data cleaning, updating, and integration. As a result, even similar types of data may depreciate at very different rates depending on a firm's data management capabilities.

This paper develops a new approach to measure data depreciation. Measuring data depreciation first requires a measure of firms' data assets. Recent research documents the economic importance of data and firms' investments in data-related assets (Tambe 2014). Following this literature, we proxy firms' data assets using their investment in data-related labor. Building on this construction of data investment, we define data depreciation as the extent to which firms must reinvest, over time, to maintain a given level of data assets. This definition reflects two sources of effective data depreciation. First, data may lose relevance or predictive power as underlying conditions change. Second, firms must devote resources to maintaining, cleaning, and updating data assets. As a result, the reinvestment required to sustain a given level of data assets captures both the natural decay of information and firms' data management effort. To formalize this intuition, we develop a stylized model in which firms allocate data workers to both data maintenance and data analysis. The model implies

that the relative allocation of labor toward data maintenance reflects the effective rate of data depreciation faced by the firm.

We empirically measure data depreciation using employee resume data from Revelio Labs. Specifically, we measure firms' data investment by observing the number of workers hired to maintain or utilize data assets. To proxy data depreciation, we classify data workers into (i) database workers, who maintain data assets and build new data pipelines or datasets, and (ii) data analysts, who use data for analysis and decision-making. We then proxy data depreciation using the ratio of database workers to total data workers. Intuitively, when data depreciation is higher, firms must devote a larger share of data labor to maintaining, cleaning, updating, or rebuilding data assets in order to sustain their usefulness. This may occur because existing data lose relevance or predictive power as underlying conditions change, but it may also arise because firms' data infrastructure or management practices require greater effort to maintain usable data assets. In either case, a higher share of database workers reflects a greater reinvestment associated with sustaining the firm's data assets.

We provide several pieces of evidence to validate this measure. First, we document the cross-industry distribution of estimated depreciation rates: industries such as fashion and apparel exhibit higher depreciation, whereas industries such as chemical manufacturing exhibit lower depreciation, consistent with the idea that rapidly changing consumer-facing information becomes obsolete more quickly than relatively stable technical or scientific data. Second, we benchmark our measure against three alternative proxies that capture either information decay or maintenance efforts, and show that our baseline measure is strongly aligned with these alternatives. Third, we exploit the implementation of the General Data Protection Regulation (GDPR), which strengthens the protection of personal data and privacy and may accelerate the effective depreciation of previously collected data. Consistent with this prediction, firms with greater exposure to GDPR experience significantly faster increases in measured data depreciation after 2018.

We next examine the relationship between data depreciation and firms' market value. If

data depreciates slowly, a firm's data assets, once created, can continue to generate long-lived returns into the future. By contrast, if data depreciates rapidly, firms must continuously reinvest to maintain the value of their data assets, leading investors to place a lower market value on those assets.

We find that the capital market assigns a positive and significant value to firms' data investment. The magnitude is larger for investments in frontier data applications, particularly artificial intelligence, suggesting that not all data-related investments are valued equally. Investments that support the deployment of cutting-edge technologies generate higher returns. Our main result is that the market value attributed to data diminishes when a firm's data depreciates rapidly, and this relationship holds consistently across alternative measures of data depreciation. A one standard deviation increase in data depreciation is associated with a 36.5 percent reduction in the market value attributed to data investment. Our heterogeneity analysis further shows that both the value of data and the impact of depreciation are more pronounced for firms that deploy their IT infrastructure in the cloud. Finally, the results remain robust across a range of alternative model specifications.

Although we find that capital markets assign positive value to firms' data investment and discount that value when data depreciation is high, this relationship may be subject to endogeneity concerns. Both the data investment and depreciation measures are inferred from firms' data labor allocation, which may also reflect managerial practices or organizational capabilities. If better-managed firms both allocate data labor differently and generate higher firm value, our estimates may capture these unobserved factors. To address this concern, we exploit two empirical settings that provide exogenous variation in data investment and depreciation. The first is the implementation of the General Data Protection Regulation (GDPR), which requires firms to delete personal data when it is no longer necessary and therefore accelerates the effective depreciation of existing data assets. The second uses data breach events as an instrument for firms' data investment, as such events force firms to increase investment in data infrastructure and security. Across both settings, the instrumental-variable

estimates remain consistent with our baseline results.

Our findings contribute to several academic literatures. There has been growing interest in understanding the role of data in firms, markets, and the broader economy. Prior research emphasizes how data accumulation enhances decision-making, competitive advantage, and market power at the firm level (Brynjolfsson and McElheran 2016; Goldfarb and Tucker 2019), drives network effects and concentration at the industry level (Eeckhout and Veldkamp 2022; Farboodi et al. 2022; Jones and Tonetti 2020), and shapes productivity and innovation at the macroeconomic level (Acemoglu et al. 2022; Arrieta-Ibarra et al. 2018). Much of this work treats data as a durable asset and focuses on data accumulation within specific industries. Recent theoretical work, however, suggests that data depreciates over time as conditions change (Farboodi and Veldkamp 2021), yet empirical evidence on this phenomenon remains scarce. This paper helps fill this gap by providing cross-industry evidence on the market value of data and highlighting the role of data depreciation. In doing so, it offers a more complete picture of how capital markets evaluate both the accumulation and obsolescence of data assets.

Second, this paper proposes a novel, theoretically and empirically grounded firm-level measure of data depreciation. A large literature on intangible and R&D capital incorporates depreciation as a necessary parameter for constructing capital stocks but provides limited direct measurement of depreciation itself (Griliches 1979; Hall 2007; Peters and Taylor 2017). Recent work measures technological obsolescence through patent-based indicators of innovation turnover (e.g., Ma (2025)) and conceptualizes data depreciation as declining information relevance in changing environments (Liu, Ma, and Veldkamp 2025). However, as discussed above, due to the complementary nature of data investment, its depreciation depends not only on information decay but also on firms' maintenance and management of data assets. This paper addresses this gap by providing a new measure that captures both dimensions of data depreciation.

Third, this paper contributes to the broader literature on IT value. Data is a founda-

tional input for artificial intelligence, which represents the frontier of IT innovation. Prior work shows that IT investments, including those in data infrastructure, can generate productivity gains and are capitalized into firm market values (Brynjolfsson, Hitt, and Yang 2002; Tambe 2014; Tambe and Hitt 2012). These gains are not uniformly distributed across firms (Tambe et al. 2020). One distinct feature of data, compared with other IT assets, is that it is intangible and may depreciate relatively quickly. This paper demonstrates that variation in data depreciation is an important source of heterogeneity in how IT-related productivity gains translate into market value.

2 Theoretical Background

2.1 Data Depreciation and Firms' Value Capture from Data

Data is increasingly recognized as a critical economic asset, often among the most valuable resources a firm can possess in the modern economy (Veldkamp and Chung (2024)). Unlike traditional tangible assets, data is intangible and derives value from the insights and predictive power it enables. By improving forecasting and supporting evidence-based decision-making, data enhances productivity and innovation. Greater data accumulation allows firms to train more effective algorithms and extract sharper insights, leading to better strategic and operational choices. Prior studies show that data-driven decision-making improves firm performance through enhanced managerial decisions, product personalization, and stronger competitive positioning—especially in data-intensive industries (Babina et al. (2024), Brynjolfsson and McElheran (2016), and Tambe (2014)). In short, firms that invest in acquiring and analyzing data build valuable knowledge assets and capabilities, which should, in theory, be reflected in higher firm performance and valuation.

However, the value of data as an asset has several characteristics that distinguish it from conventional assets, including other forms of intangible capital. First, data is non-rivalrous. The same dataset can be used by multiple processes or even multiple firms simultaneously

without being depleted (Jones and Tonetti 2020). This property allows data to generate increasing returns and network effects—for example, more users generating more data can improve services in a reinforcing cycle. In contrast, many traditional intangible assets such as R&D knowledge or human capital are tied to specific technologies, teams, or products, and therefore cannot be scaled across applications as easily as data.

Second, the value of data is highly context-dependent and time-sensitive. As market conditions, consumer preferences, or technologies evolve, previously collected data may become less relevant or less predictive. This phenomenon is often related to data depreciation: data lose their usefulness as the environment changes. For example, consumer behavior data from several years ago may lose value if preferences shift, and predictive models may become less accurate as real-world patterns evolve. Recent theoretical work models data as a stock that grows with new information inflows but depreciates as older data become obsolete (Farboodi and Veldkamp 2021; Liu, Ma, and Veldkamp 2025). Compared with other intangible assets—such as patents or software—that often retain their usefulness for longer periods, the economic value of data may be more sensitive to changes in the external environment.

Third, maintaining usable data requires ongoing organizational effort. Data must be cleaned, updated, integrated, and stored in functioning data systems in order to remain valuable. Firms differ substantially in their data infrastructure and management practices, which affects how much effort is required to keep data assets usable over time. As a result, the effective depreciation of data may reflect not only information decay but also the cost of maintaining and managing data assets. This feature further distinguishes data from many other intangible assets, whose depreciation primarily reflects technological obsolescence rather than continuous maintenance requirements.

Third, maintaining usable data requires ongoing organizational effort. Data must be cleaned, updated, integrated, and stored in functioning data systems in order to remain valuable. Firms differ substantially in their data infrastructure and data management practices, which affects how much effort is required to keep data assets usable over time. This

feature is closely related to the broader literature on IT complementarities, which emphasizes that the value of digital technologies depends critically on complementary organizational investments and capabilities (Brynjolfsson and Hitt 2000; Forman, Goldfarb, and Greenstein 2012). As a result, the effective depreciation of data may reflect not only information decay but also the cost of maintaining and managing data assets. This feature further distinguishes data from many other intangible assets, whose depreciation primarily reflects technological obsolescence rather than continuous maintenance requirements.

Taken together, these features imply that the economic value of data depends not only on how much data firms accumulate but also on how quickly those data depreciate. The rate of data depreciation is shaped by two forces: external environmental change and firms' internal data management. As a result, the useful life of data depends both on how quickly new information arrives and on how effectively firms maintain and update their data assets.

2.2 How to Measure Data Depreciation

A large literature on intangible and R&D capital incorporates depreciation as a necessary parameter for constructing capital stocks, but provides limited direct measurement of depreciation itself (Griliches 1979; Hall 2007; Peters and Taylor 2017). Recent work measures technological obsolescence through patent-based indicators of innovation turnover (e.g., Ma (2025)) and conceptualizes data depreciation as declining information relevance in changing environments (Liu, Ma, and Veldkamp 2025).

Building on this literature, we define *data depreciation* as the decline in the productive value of a firm's data assets over time, captured by the reinvestment required to maintain a given level of data value. Conceptually, this decline arises from two distinct but related forces. First, the informational content of data may become obsolete as environments, technologies, or user behavior change. Second, maintaining the usability of data requires ongoing organizational investments in cleaning, updating, storage, and integration; without such efforts, data gradually loses operational value even if its raw informational content remains

unchanged. These two forces imply that data capital differs fundamentally from traditional R&D capital: it lacks a well-defined observable investment flow (Tambe et al. 2020) and depreciates both through information decay and through limits in firms' data management capacity. In practice, neither component is directly observable.

We therefore develop a simple framework that identifies an *effective* data depreciation rate from firms' observable responses—specifically, changes in labor composition—rather than imposing exogenous depreciation parameters to construct capital stocks.

Data dynamics and effective depreciation. Firms enter period t with a predetermined stock of effective data D_t, summarizing usable information accumulated from past operations. Unlike physical capital, data loses value both because past information becomes less relevant as environments evolve and because maintaining usable data requires continuous organizational effort. We capture these features with the law of motion

$$D_{t+1} = (1 - \delta_t)D_t + \eta_t L_t^{db}, \tag{1}$$

where δ_t reflects information decay due to environmental change, and η_t represents the firm's organizational capacity to preserve and maintain the productive value of its data stock.

Equation (1) implies that sustaining a usable data stock requires offsetting both information decay and maintenance frictions. We therefore define *effective data depreciation* as

$$\tilde{\delta}_t \equiv \frac{\delta_t}{\eta_t}, \tag{2}$$

which measures the maintenance-adjusted rate at which existing data must be replaced to keep the effective stock constant. Conceptually, effective data aggregates past observations with declining relevance as environments evolve; a formal micro-foundation based on information decay is provided in the Appendix D.

Value creation from data. Data enhances firm performance by improving decision qual-

ity, targeting, and operational efficiency. After optimizing over conventional inputs such as physical capital and standard labor under competitive factor markets, period profit can be written in reduced form as

$$\pi_t(D_t) = \tilde{A}_t D_t^{\vartheta}, \qquad \vartheta > 0, \tag{3}$$

where $\tilde{A}_t$ captures productivity and factor price terms. This reduced-form representation summarizes the idea that a larger effective data stock raises firm value. Detailed derivation can be found in Appendix D.

Data scientists versus database workers. To generate value from data, firms must process and interpret it. We distinguish two types of data-related labor. Data scientists L_t^{ds} analyze existing data and affect contemporaneous value extraction, while database workers L_t^{db} collect, clean, integrate, and maintain data and therefore determine future data accumulation. We model complementarity between data and data scientists as

$$\pi_t(D_t, L_t^{ds}) = \tilde{A}_t D_t^{\vartheta} (L_t^{ds})^{\kappa} - w_{ds} L_t^{ds} - w_{db} L_t^{db}, \qquad 0 < \kappa < 1. \tag{4}$$

This specification captures that extracting value from data requires analytical labor, while maintaining the data stock requires database labor.

Maintenance requirement and labor allocation. To illustrate the implications of effective depreciation, consider maintaining a constant effective data stock, $D_{t+1} = D_t = \bar{D}$. From equation (1), the maintenance condition becomes

$$L^{db} = \tilde{\delta}_t \bar{D}, \tag{5}$$

where $\tilde{\delta}_t = \delta_t / \eta_t$ summarizes the maintenance-adjusted depreciation rate. Substituting (5) into (4), the firm chooses data science labor to solve

$$\max_{L^{ds} \geq 0} \; \tilde{A}_t \bar{D}^{\vartheta} (L^{ds})^{\kappa} - w_{ds} L^{ds} - w_{db} \tilde{\delta}_t \bar{D}. \tag{6}$$

The optimal level of data science labor is therefore

$$L^{ds*} = \left(\frac{\kappa \tilde{A}_t}{w_{ds}}\right)^{\frac{1}{1-\kappa}} \bar{D}^{\frac{\vartheta}{1-\kappa}}. \tag{7}$$

Empirical implication. The model implies that effective data depreciation $\tilde{\delta}_t$ determines the organizational effort required to maintain usable data. Combining equations (5) and (7) shows that the share of database labor in total data-related labor satisfies

$$\frac{L^{db}}{L^{db} + L^{ds*}} = \frac{\tilde{\delta}_t \bar{D}}{\tilde{\delta}_t \bar{D} + L^{ds*}} \propto \tilde{\delta}_t. \tag{8}$$

We therefore use the *database labor share* as an empirical proxy for effective data depreciation. This measure captures both information decay driven by environmental change (through δ_t) and organizational maintenance frictions (through η_t), and is directly observable in labor force composition data. We also briefly discuss alternative proxies for δ_t and η_t, and compare them to the *database labor share* measure to demonstrate its distinctiveness and robustness.

3 Data and Research Design

This section describes the empirical tests used to: (i) estimate the market value assigned to data investment, and (ii) examine whether financial markets assign less value to firms with higher data depreciation rates.

Prior work has used market value regressions to uncover valuable but otherwise invisible (intangible) assets that contribute to a firm's productive capacity (Brynjolfsson, Hitt, and Yang (2002)). This approach is particularly useful for analyzing data investments, as firms often require time to adapt to new technologies, and the literature suggests that many firms are not yet consistently realizing value from such investments. Investors, however, assign value based on the expected future stream of benefits an asset will generate, so these assets may be reflected in the firm's market valuation before they appear in productivity statistics.

The market value regression used to test whether the market assigns value to data investment and whether this value is discounted when the data depreciation rate δ_i is high—is:

$$\begin{aligned} Log(MV)_{it} =& Log(Assets)_{it} + Log(PPE)_{it} + Log(Employment)_{it} \\ &+ Log(IT)_{it} + Log(Data)_{it} + Log(AI)_{it} + \\ &+ \delta_i + Log(Data)_{it} \times \delta_i + \gamma_{it} + \epsilon_{it} \end{aligned} \tag{9}$$

3.1 Data Sources

The empirical tests described above are informed by data sources on (a) how firms invest in data and (b) how data depreciates within firms. These are supplemented with financial data from public firms to assess how technological and workforce changes are connected to the value investors assign to firms.

Corporate employment database. The corporate employment data were provided by Revelio Labs, a workforce intelligence company. Their databases are constructed from a variety of online sources and are similar in informational content to data posted on professional networking platforms such as LinkedIn. They cover a large share of white-collar work in the U.S., including both public and private firms, but the sample used in this analysis is limited to public firms so that observations can be linked to financial market data. The data are over-sampled in management, business, and technology occupations, and under-sampled in fields such as agriculture and manufacturing, consistent with the greater use of online professional platforms in knowledge-intensive occupations (Tambe (2025)).

This database is used to generate measures of annual firm-occupation-skill employment activity from 2013 through 2023. This panel provides information on the types of technical skills present and the number of data workers employed by each firm. The records for each employer include CUSIP identifier codes, which allow them to be merged with external financial databases such as Compustat-Capital IQ.

Supplementary data: Compustat. Our analyses incorporate firms' financial and em-

ployment information from Compustat–Capital IQ, accessed through the WRDS data service. We use data on firms' assets, market value, sales, and employment. In addition, we use Compustat segment data to construct measures of firms' sales across geographic regions.

Supplementary data: Data Breach Database. We collect firm-level data breach events from the Privacy Rights Clearinghouse (PRC), an organization that tracks breaches investigated by U.S. state Attorneys General and the U.S. Department of Health and Human Services. The PRC chronology of data breaches[3] contains 42,465 breach events from 2013 to 2023, of which 21,961 are classified as external cyberattacks (e.g., malware, ransomware, and network intrusions). We link these data to our main sample using firm names.

Supplementary data: FactSet Revere. We use FactSet Revere to obtain firm-level supply chain relationships, including customer–supplier linkages. This information allows us to identify whether a firm outsources data-related tasks by observing transactions involving data-related products and services within its supply chain network.

Supplementary data: Product Market Fluidity Data. Hoberg, Phillips, and Prabhala (2014) provide firm-level product market fluidity measures constructed from textual analysis of 10-K product descriptions. This measure captures the extent to which a firm's product offerings change from year to year.

3.2 Sample Construction and Key Measures

Sample construction. The Revelio workforce sample, which forms the core of this analysis, includes firm-occupation-year-skill counts for the years 2013 to 2023. To merge these data with financial information from Capital IQ, the sample is limited to public firms, resulting in 20,926 firm-year observations. Table 1 reports summary statistics for all observations in the sample. The statistics are presented in logs, consistent with their use in the multivariate regressions. Variables are also standardized to aid interpretability.

Data investment. Occupation data from Revelio are used to measure firm-year data in-

3. Available at: https://privacyrights.org/data-breaches

vestment. A key challenge is classifying occupations as data-relevant and further distinguishing between data scientists and database workers. Recent studies using large-scale workforce data have employed manual mappings. For instance, Goldfarb, Taska, and Teodoridis (2023) select a cluster of skills related to machine learning technologies for their analysis. Deming and Kahn (2018) curate words and phrases in the Lightcast data associated with different job skills, including cognitive, social, character, and computer categories.

This paper adopts a similar approach based on the occupation categorizations provided by the data source. Revelio offers 1,500 classified occupations, and we manually reviewed each one using job titles, O*NET classifications, and job descriptions. For example, "database engineer" is classified as a database worker, while "data analyst" is classified as a data scientist. [4] After identifying data-related roles, we calculate the number of data workers employed by each firm in a given year. This count serves as the proxy for firm-year-level data investment.

This labor-based measures provide a useful proxy for intangible IT investments because many data capabilities are embodied in specialized human capital rather than directly observable capital expenditures (Tambe 2014). In particular, investments in data infrastructure often require complementary organizational capabilities and technical expertise, making workforce composition a practical way to observe firms' commitment to data-related activities.

Data depreciation - database labor share. We construct several measures to proxy the firm-level data depreciation rate. As discussed in the model, effective data depreciation $\tilde{\delta}_t = \delta_t/\eta_t$ is jointly determined by information decay (δ_t) and the firm's organizational capacity to maintain the usable value of data (η_t). Intuitively, this ratio governs how frequently firms must refresh and reprocess their data in order to sustain its productive value.

Our primary proxy for data depreciation is the ratio of database workers to the total number of data workers employed by a firm in a given year. We distinguish between two types

4. A full list of data scientist roles and database worker roles can be found at Appendix C

of data-related occupations: database workers, who are primarily responsible for collecting, cleaning, and maintaining data, and data scientists, who focus on analyzing and extracting insights from data. The intuition is straightforward. When effective data depreciation is low (low δ_t or high η_t), existing datasets remain usable for longer, allowing firms to devote relatively more resources to analysis rather than data preparation. In contrast, when effective data depreciation is high (high δ_t or low η_t), firms must continuously acquire, clean, and integrate new data to maintain its usefulness. This increases the demand for database workers relative to data scientists. Therefore, a higher ratio of database workers to total data workers serves as a proxy for a higher effective data depreciation rate.

Data depreciation - supplementary measures. We also construct three supplementary measures that map to the two components of effective data depreciation, $\tilde{\delta}_t = \delta_t/\eta_t$. The first two measures proxy for information decay (δ_t), while the third captures organizational maintenance capacity (η_t). More details on their construction are provided in Appendix E.

First, we use the product fluidity index from Hoberg, Phillips, and Prabhala (2014), constructed from textual analysis of firms' 10-K product descriptions. This measure captures how rapidly a firm's product space evolves over time. We interpret higher product fluidity as reflecting faster environmental and demand shifts, which reduce the relevance of previously collected data. Accordingly, this measure serves as a proxy for higher information decay (δ_t).

Second, following Liu, Ma, and Veldkamp (2025), we measure the persistence of firm performance using an AR(1) specification of return on assets (ROA). Lower predictive power of past performance indicates that historical information loses relevance more quickly. We therefore interpret lower persistence as indicating higher information decay (δ_t).

Finally, we construct a skill-based shift-share measure that captures how quickly the data-related skills employed by a firm become obsolete in the broader labor market. This index reflects the pace of technological and methodological change in data practices. Because faster skill obsolescence increases the need for continual data cleaning, updating, and integration, we interpret this measure as capturing lower organizational maintenance effectiveness (lower

η_t), and hence higher effective data depreciation.

Other variables. We also calculate firms' AI investment using Revelio data by counting how many workers have AI skills, where AI skills are defined as described in Tambe (2025). The Compustat-Capital IQ data are used to construct employer-year measures for total market value, employment, industry classification, the value of property, plant, and equipment (PPE), and other assets. As discussed earlier, the use of Capital IQ financial data necessitates limiting the sample to public firms. Industry variables are retrieved at the three-digit NAICS (North American Industry Classification System) level. Total market value is computed following the approach used in prior research on intangible assets and firm valuation (e.g., see the Appendix of Brynjolfsson, Hitt, and Yang (2002) for details on variable construction). It is calculated as the sum of the value of equity at the end of the fiscal year, the value of preferred stock, and total debt, representing the total worth of a firm as assessed by financial markets. "Assets" are computed as total assets minus PPE.

4 Results

4.1 Summary Statistics

Figures 1 illustrate the cross-industry distribution of data depreciation. The pattern is economically intuitive. Industries facing more volatile demand and rapidly changing products or operations, such as retail, telecom, and construction, exhibit higher depreciation rates, consistent with data becoming obsolete more quickly and requiring more frequent updating and integration. In contrast, sectors with more stable production processes and slower-changing information environments, including chemicals and education, display substantially lower depreciation rates, suggesting that their data remain useful for longer periods. More broadly, industries such as utilities, clothing retail, and food services (where demand and pricing conditions change frequently) also exhibit relatively high depreciation, whereas sectors such as manufacturing, hospitals, and information services (where operational or scientific data

remain relevant for longer horizons) tend to display lower depreciation rates. Figure A.1 shows a related pattern for data investment. It is concentrated in information- and service-oriented sectors such as broadcasting, finance, and professional services, where production relies heavily on data processing and analytics, and is comparatively lower in traditional manufacturing and resource-based industries.

Figure 2 plots the average data depreciation over time using different proxies and reveals several patterns. The primary measure based on the database worker ratio exhibits a clear and monotonic downward trend over the sample period. The two proxies intended to capture information decay, product fluidity and ROA predictability, rise through the middle of the sample and then flatten or partially reverse in later years, suggesting that information-decay pressures intensify during the mid-sample period. In contrast, the skill-obsolescence-based proxy, which is best interpreted as capturing *maintenance frictions* (i.e., a proxy for $1 - \eta_t$ rather than η_t), displays a modest downward trend in earlier years.

Taken together, these patterns are broadly consistent with a setting in which information-decay pressures increase over time while firms' ability to preserve and extract value from existing data improves. In our framework, the database-worker ratio reflects firms' endogenous allocation of data labor between maintaining data assets and applying them in analytics and decision-making. As data analytics capabilities expand, a larger share of data workers may be engaged in interpreting and utilizing existing datasets, which allows firms to sustain the usefulness of their data assets for longer horizons, thus effectively raising η_t. The resulting decline in the database-worker-ratio measure can therefore be interpreted as reflecting improvements in effective data maintenance and utilization that offset rising information-decay pressures. Because the measure embeds both forces through firms' labor allocation decisions, its time trend need not map one-for-one to any individual proxy series, but instead summarizes their joint effect.

4.2 Main Results

Table 2 reports estimates from the market value regression on an eleven-year panel of public firm investments (2013–2023). All columns include year and 3-digit industry fixed effects. Column (1) reports results for market value regressed on measures of IT investment, assets, capital (PPE), and employment, with all variables entered in logs. Consistent with prior work (Brynjolfsson, Hitt, and Yang (2002)), financial markets assign significant value to general IT capital (t = 11.41). In Column (2), after adding data investment to the regression, the coefficient on general IT capital drops to 0.120 and becomes statistically insignificant, suggesting that the market returns to IT are primarily driven by frontier investments—here represented by data investment (t = 3.16). Column (3) introduces AI investment, and the coefficient on data investment falls from 0.329 to 0.207, indicating that a significant portion of the market value attributed to data investment is likely realized through its role as an input into AI applications.

Columns (4) to (7) examine how data depreciation moderates the market value assigned to data investment. Our primary measure of data depreciation is the database worker ratio, which captures the extent to which firms must continually refresh and manage their data, while the remaining three measures serve as supplementary proxies based on information decay and maintenance frictions. The results indicate that firms with higher data depreciation receive lower market valuations for their data investments. In Column (4), using the database worker ratio as the main measure, a one standard deviation increase in data depreciation is associated with a 36.5 percent reduction in the market value attributed to data investment. Estimates using the alternative proxies are smaller but consistently negative, ranging from 15 percent to 36.5 percent, providing supportive evidence for the moderating role of data depreciation.

These results suggest that data investment contributes significantly to firm valuation, reinforcing the view that data functions as a valuable intangible asset. However, the value of these assets is not uniform across firms. It depends critically on two dimensions of data

depreciation: the rate at which the informational content of data becomes obsolete due to environmental change, and the firm's organizational capacity to maintain, update, and integrate data so as to preserve its usable value. Firms therefore benefit more from data investments when the informational environment evolves more slowly and when they possess stronger data management capabilities that mitigate effective depreciation. This underscores the importance of not only investing in data capabilities but also managing both the external sources of information decay and the internal processes that sustain the longevity and reusability of data.

4.3 Effect Heterogeneity and Robustness Tests

The average effects in Table 2 may obscure important heterogeneity. Figure 3 presents split-sample regression results based on firms' cloud infrastructure and other statistics.

Panel (a) of Figure 3 divides firms into high-cloud and low-cloud groups based on their cloud skill intensity. Cloud technologies provide more scalable and flexible infrastructure for storing, updating, and processing data, and thus are expected to enhance firms' ability to maintain and refresh their data. Using the Revelio database, we identify IT skills and manually classify them as either cloud-related or on-premises.[5] We then compute the share of IT workers with cloud-related skills at the firm level and classify firms above the median as high-cloud firms. The estimated discount associated with data depreciation is more negative for high-cloud firms, indicating that capital markets place greater weight on data depreciation when firms rely more heavily on cloud-based data infrastructure.

Panel (b) of Figure 3 further explores heterogeneity by firm characteristics, splitting the sample by firm age, size (measured by assets), sales growth, and industry competition, with "old," "large," and "high" defined as above the median in each case. The depreciation discount is more negative for older firms than for newer firms, although the difference is not precisely estimated; this is consistent with older firms relying more on accumulated data

5. Classified skills are listed in Appendix.

stocks whose value is more sensitive to depreciation. The discount is also more negative for smaller firms than for larger firms, with limited statistical separation. For firm growth, the coefficient is somewhat more negative among high-growth firms, again without strong statistical distinction, consistent with faster-changing business environments increasing the effective depreciation of data. Finally, the discount is more negative in high-competition industries than in low-competition industries, though the difference is imprecisely estimated. Overall, the direction of heterogeneity is broadly consistent with our interpretation that the measure captures economically meaningful variation in the effective depreciation of firm data.

The results are robust across different specifications, as shown in Table A.1. Columns (1) and (2) incorporate firm fixed effects; the discount in market value associated with data depreciation becomes smaller but remains negative, indicating that a substantial portion of the variation is driven by cross-firm differences rather than within-firm changes over time. Columns (3) and (4) exclude IT firms with NAICS codes beginning with "51" or "54." The results remain consistent with the main findings, and the coefficients on data investment become statistically significant, suggesting that non-IT firms may be less likely to operate at the technological frontier of AI use, making the value of data investments more salient to investors. Columns (5) and (6) exclude workers with AI skills from the data worker category and yield similar results, indicating that the main findings are not driven by the inclusion of AI-specific occupations. Column (7) interacts the data depreciation rate with firms' IT investment. We find that a firm's data depreciation rate does not significantly affect the value assigned to general IT investment, serving as a placebo test and reinforcing that the moderating effect is specific to data-related assets rather than broader IT capital.

We examine data task outsourcing to address a potential measurement concern: firms may outsource data-related activities rather than perform them in-house, which could affect our measure of data depreciation. Table A.2 compares firms that outsource data tasks to those that do not, using supply chain information from FactSet Revere. We find that the main results are robust across both groups. In particular, the interaction between data

and depreciation remains negative and statistically significant among non-outsourcing firms, while it is smaller and less precisely estimated among outsourcing firms.

Table A.3 repeats the main analysis using firm sales (productivity) as the dependent variable instead of market value. Prior work has emphasized that market value regressions can uncover the value of intangible assets that have not yet translated into realized productivity (Brynjolfsson, Hitt, and Yang 2002). By contrast, sales-based regressions capture the contemporaneous production payoff from these investments rather than investors' expectations about their future benefits. Several coefficients differ notably from the market value specifications. In particular, the coefficients on $\log(Data)$ and $\log(AI)$ become smaller or negative in the productivity regressions, whereas they were generally positive in the market value regressions. This divergence is consistent with the idea that data and AI investments often require complementary organizational changes and learning before improving realized output. Investors, however, may anticipate longer-term productivity gains and capitalize them into market value even before they are fully realized in current firm performance.

4.4 Causal Estimation

Estimating the market value assigned to data depreciation and data investment raises several endogeneity concerns. First, firms' data investment and data management practices are likely jointly determined with their expected growth opportunities and unobserved managerial capabilities, leading to reverse causality and omitted variable bias. Firms that anticipate higher future returns may invest more in data and simultaneously adopt better data maintenance practices. In addition, measurement error in both data depreciation and data investment may attenuate estimated coefficients. In this section, we use instrumental variable approach that exploits plausibly exogenous shocks affecting either the effective depreciation of data or firms' incentives to invest in data.

We first use the implementation of the General Data Protection Regulation (GDPR) as an instrument for data depreciation. GDPR, enacted by the European Union in 2018,

imposes stricter requirements on data collection, storage, and usage, including limitations on data retention and more stringent consent and compliance standards. These regulatory constraints can accelerate the effective obsolescence of firm data by restricting how long data can be stored and how flexibly they can be reused across contexts, thereby increasing the effective depreciation rate of data. Importantly, GDPR represents a regulatory shock that is largely external to individual firms' investment decisions, making it a plausible source of exogenous variation in data depreciation.

Second, to instrument for data investment, we use firm-level data breach events. Data breaches, such as cyber attacks unauthorized data disclosures, represent unexpected shocks that compromise the integrity and usability of existing data assets. Following a breach, firms often need to rebuild, upgrade, or reorganize their data infrastructure, which can lead to increased investment in data-related capabilities and systems. Because breaches are typically unanticipated and driven by external security vulnerabilities rather than firms' contemporaneous investment choices, they provide plausibly exogenous variation in firms' incentives to invest in data, allowing us to identify the causal impact of data investment on firm valuation.

4.4.1 GDPR and Data Depreciation

Figure 4 shows the data depreciation rate for firms with and without exposure to GDPR over time. We define GDPR exposure based on firms' sales in Europe prior to 2018, measured as the share of sales to European Union countries in total firm sales. Before 2018, the two groups exhibit broadly similar downward trends in data depreciation. Following the implementation of GDPR in 2018, the series begin to diverge, with firms exposed to GDPR displaying a relatively higher depreciation rate than non-exposed firms. This pattern is consistent with the interpretation that GDPR imposes stricter constraints on data storage, retention, and reuse, thereby shortening the effective useful life of firm data.

Table 3 reports regression results examining how the moderating role of data depreciation

varies with GDPR exposure and using an instrumental variable strategy. Focusing on the interaction term $\log(Data) \times Depreciation$, Columns (1)–(4) show that the coefficient is more negative for GDPR-exposed firms than for non-exposed firms, both before and after 2018. This pattern suggests that when firms are subject to stricter data regulations, the market discounts data investment more strongly when data depreciate faster, consistent with the idea that regulatory constraints increase the effective obsolescence of data assets.

Column (5) presents the IV results using the interaction between European sales exposure and a post-2018 indicator as an instrument for data depreciation. The first-stage coefficient is large and negative (-0.941, significant at the 1% level), indicating that GDPR exposure after 2018 is strongly associated with higher effective data depreciation as measured by the database worker ratio. In the second stage, the interaction term $\log(Data) \times Depreciation$ becomes substantially more negative (-0.265, significant at the 1% level), and the magnitude is considerably larger than in the OLS estimates. This implies that a one-unit increase in the data depreciation rate leads to a markedly stronger discount in the market value associated with data investment when instrumented by GDPR exposure. The larger IV magnitude suggests that measurement error or attenuation bias in the OLS specifications may understate the true moderating effect of data depreciation on firm valuation.

4.4.2 Data Breach and Data Investment

Figure 5 summarizes how firms' data investment and data depreciation vary with data breach events. Panel (a) shows a clear positive relationship between the number of breach events and firms' data investment (measured by the firm-level mean of $\log(Data)$), consistent with the idea that breaches trigger subsequent rebuilding and upgrading of data infrastructure. Panel (b) shows little systematic relationship between breach events and the data depreciation rate, with only a mild negative slope. This pattern suggests that breach events primarily shift firms' incentives to invest in data rather than directly altering the underlying rate of data depreciation, supporting their use as an instrument for data investment.

Table 4 reports the corresponding regression results. In Columns (1) and (2), which split firms by whether they have ever experienced a breach, the coefficient on $\log(Data)$ is positive and significant for breach firms (Column (1)) but smaller and statistically insignificant for non-breach firms (Column (2)). This indicates that, in the raw correlations, data investment is positively associated with firm valuation, particularly among firms that have experienced breaches and subsequently expanded their data capabilities.

Column (3) implements the IV specification, using the interaction between the number of breach events and an indicator for being after the first breach year as an instrument for data investment. The first stage is strong, with a positive and significant coefficient (0.014) and an F-statistic of 18.62, indicating that breach shocks substantially increase subsequent data investment. In the second stage, the coefficient on $\log(Data)$ becomes negative and significant, while the interaction term $\log(Data) \times Depreciation$ becomes substantially more negative (-0.331). The sign reversal in $\log(Data)$ is consistent with breach-driven investment capturing a different margin of data accumulation: investment undertaken in response to breaches may reflect costly rebuilding or compliance-driven spending that does not immediately translate into higher market value. At the same time, the larger negative interaction coefficient implies that when data depreciate faster, the market discounts the value of breach-induced data investment more strongly. Together, the IV results suggest that once we isolate plausibly exogenous variation in data investment, faster data depreciation significantly reduces the valuation payoff to data assets.

5 Discussion and Conclusion

This paper provides evidence using employee resume data to estimate how depreciation affects the market value of data investment. It yields several findings. First, data is a valuable intangible asset—firms that invest more in data, especially when combined with complementary cloud and AI technologies, tend to have higher market valuations. Second, the market

value attributed to data investment declines when data depreciates rapidly, highlighting the importance of data longevity in realizing its full economic value.

There are several limitations to this analysis that should be noted. The data offers limited visibility into the exact volume of data assets or how firms implement and utilize those assets in practice. The analysis is also confined to the specific relationship between data-related human capital investment and firms' market value. Moreover, the data depreciation measures used in this paper are proxies—they reflect relative depreciation rates across firms but do not capture depreciation directly. While the results are robust across several alternative measures, future work could explore more direct approaches to quantifying both data value and depreciation.

Beyond these limitations, this study opens up several promising directions for future research and policy discussion. Debates around the role of data in market power and competition often emphasize increasing returns to scale and network effects, while overlooking the role of depreciation. Yet, if data depreciates quickly, these dynamics may be more short-lived or fragile than commonly assumed. In addition, since data is a key input into AI models, a high rate of depreciation raises concerns about how to sustain a consistent supply of high-quality data. We hope this paper serves as a foundation for further exploration into data depreciation and its broader strategic and economic implications.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022. "Too much data: Prices and inefficiencies in data markets." *American Economic Journal: Microeconomics* 14 (4): 218–256.

Aghion, Philippe, and Peter Howitt. 1992. "A model of growth through creative destruction." *Econometrica* 60 (2): 323–351.

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E Glen Weyl. 2018. "Should we treat data as labor? Moving beyond "free"." *AEA Papers and Proceedings* 108:38–42.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics* 151:103745.

Brynjolfsson, Erik, and Lorin M Hitt. 2000. "Beyond computation: Information technology, organizational transformation and business performance." *Journal of Economic perspectives* 14 (4): 23–48.

Brynjolfsson, Erik, Lorin M Hitt, and Shinkyu Yang. 2002. "Intangible assets: Computers and organizational capital." *Brookings papers on economic activity* 2002 (1): 137–181.

Brynjolfsson, Erik, and Kristina McElheran. 2016. "The rapid adoption of data-driven decision-making." *American Economic Review* 106 (5): 133–139.

Deming, David, and Lisa B Kahn. 2018. "Skill requirements across firms and labor markets: Evidence from job postings for professionals." *Journal of Labor Economics* 36 (S1): S337–S369.

Eeckhout, Jan, and Laura Veldkamp. 2022. "Data and market power."

Farboodi, Maryam, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran. 2022. *Valuing financial data.* Technical report. National Bureau of Economic Research.

Farboodi, Maryam, and Laura Veldkamp. 2021. *A model of the data economy.* Technical report. National Bureau of Economic Research Cambridge, MA, USA.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2012. "The Internet and local wages: A puzzle." *American Economic Review* 102 (1): 556–575.

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis. 2023. "Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings." *Research Policy* 52 (1): 104653.

Goldfarb, Avi, and Catherine Tucker. 2019. "Digital economics." *Journal of economic literature* 57 (1): 3–43.

Griliches, Zvi. 1979. "Issues in assessing the contribution of research and development to productivity growth." *The bell journal of economics,* 92–116.

Hall, Bronwyn H. 2007. "Measuring the returns to R&D: The depreciation problem."

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala. 2014. "Product market threats, payouts, and financial flexibility." *The Journal of Finance* 69 (1): 293–324.

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

Liu, Ernest, Song Ma, and Laura Veldkamp. 2025. "Data sales and data dilution." *Journal of Financial Economics* 169:104053.

Lu, Jie, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. "Learning under concept drift: A review." *IEEE transactions on knowledge and data engineering* 31 (12): 2346–2363.

Ma, Song. 2025. "Technological Obsolescence." *Review of Financial Studies.*

McElheran, Kristina, J Frank Li, Erik Brynjolfsson, Zachary Kroff, Emin Dinlersoz, Lucia Foster, and Nikolas Zolas. 2024. "AI adoption in America: Who, what, and where." *Journal of Economics & Management Strategy* 33 (2): 375–415.

Peters, Ryan H, and Lucian A Taylor. 2017. "Intangible capital and the investment-q relation." *Journal of Financial Economics* 123 (2): 251–272.

Tambe, Prasanna. 2014. "Big data investment, skills, and firm value." *Management science* 60 (6): 1452–1469.

———. 2025. "Reskilling the Workforce for AI: Domain Expertise and Algorithmic Literacy." *Management Science.*

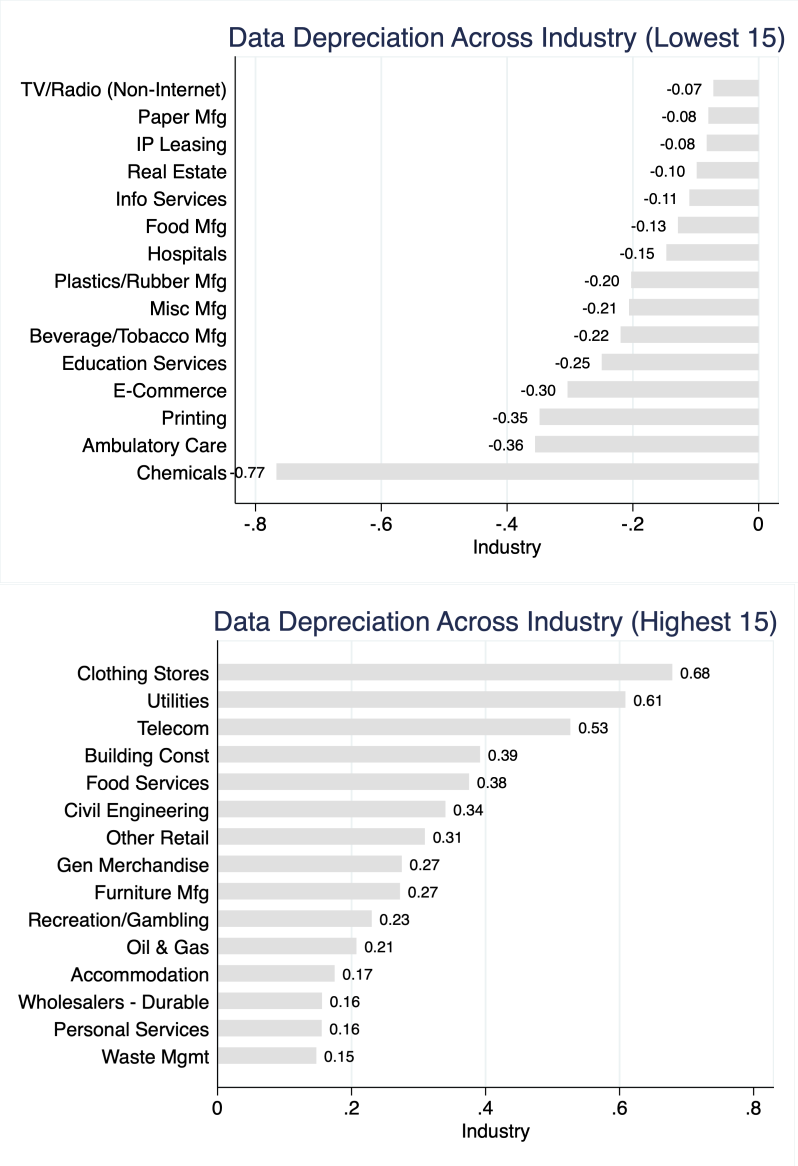
Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson. 2020. *Digital capital and superstar firms.* Technical report. National Bureau of Economic Research.

Tambe, Prasanna, and Lorin M Hitt. 2012. "The productivity of information technology investments: New evidence from IT labor data." *Information systems research* 23 (3-part-1): 599–617.

Veldkamp, Laura, and Cindy Chung. 2024. "Data and the aggregate economy." *Journal of Economic Literature* 62 (2): 458–484.

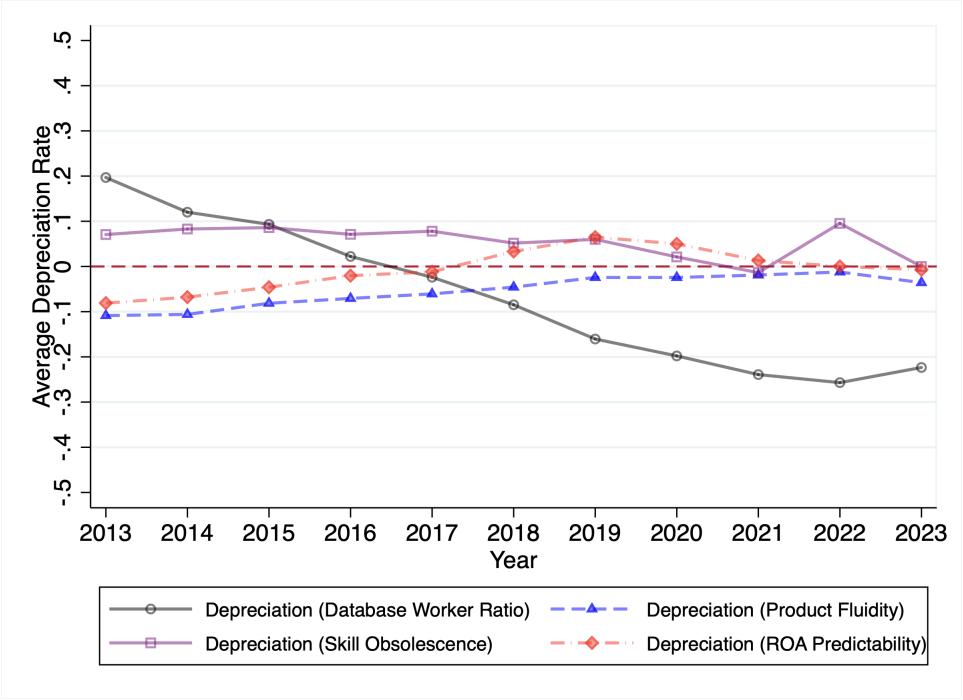
Figures

Figure 1: Industry Distribution of Data Depreciation



Notes. This figure reports the distribution of average firm data depreciation rate across NAICS 3 digit industries in our sample. Industry with fewer than 50 firms are dropped. Data depreciation is measured using the share of database workers among database workers and data scientists.

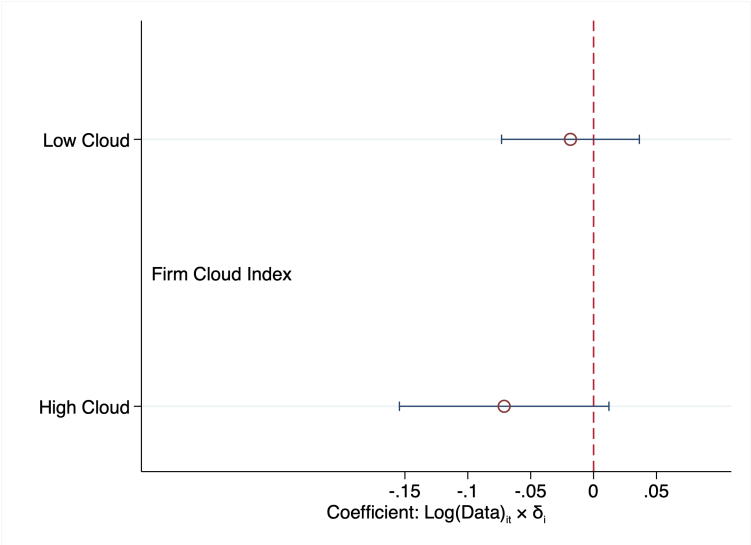
Figure 2: Time Trend of Average Data Depreciation Rates (Different Measures)



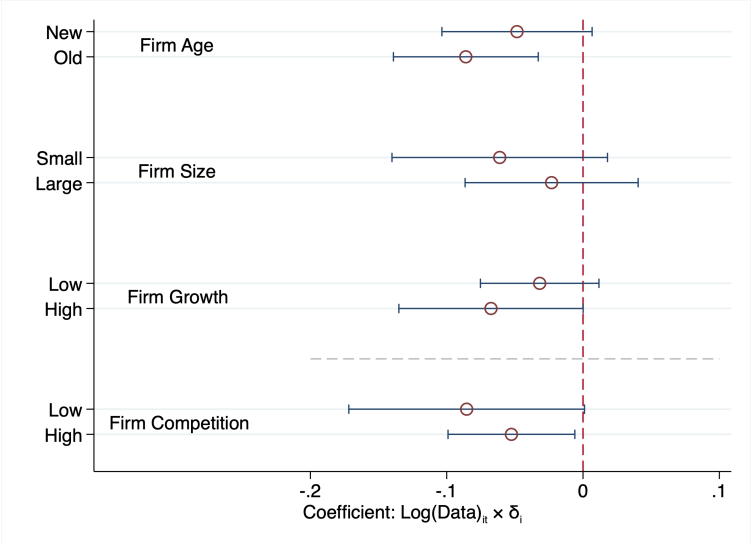
Notes. This figure shows the average data depreciation rate of all firms over time using different measures. *Depreciation (Database Worker Ratio)* measures data depreciation using the share of database workers among database workers and data scientists. *Depreciation (Product Fluidity)* proxies data depreciation using the product fluidity index derived from textual analysis of firms' 10-K product descriptions, which reflects year-to-year changes in product offerings. *Depreciation (Skill Obsolescence)* infers data depreciation by tracking the extent to which a firm's data-related skills become obsolete in the broader labor market after two years. *Depreciation (ROA Predictability)* is a dummy variable that equals 1 if the R^2 of an autoregressive model of return on assets (ROA) is smaller than 10%. All these measures are standardized.

Figure 3: Heterogeneity of Data Depreciation Discount

(a) High Cloud or Not

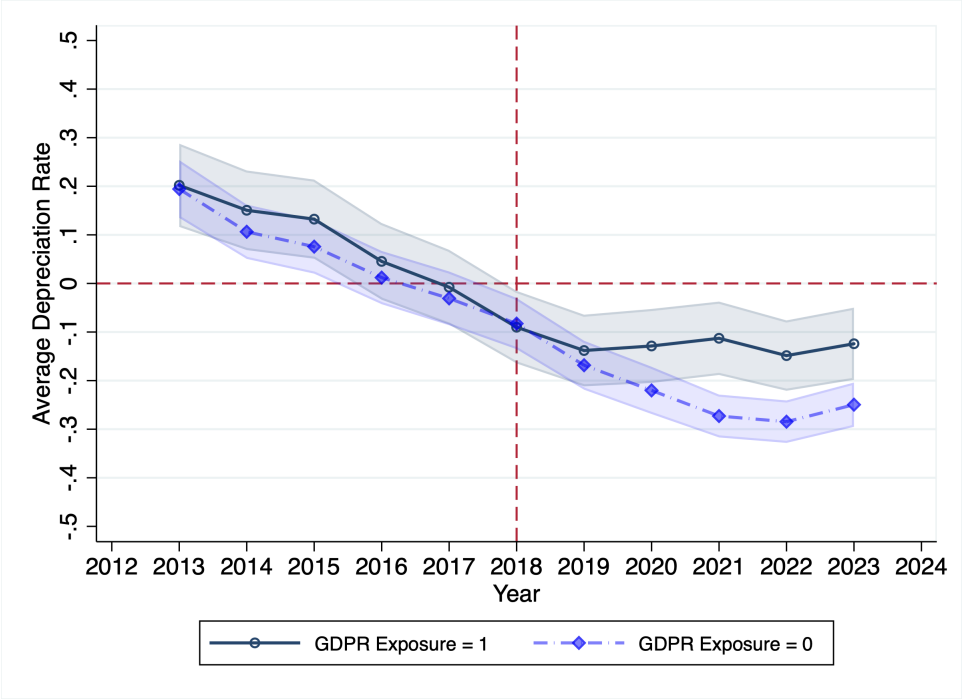


(b) Firm Statistics



Notes. This figure reports the coefficients of $Log(Data)_{it} \times \delta_i$ from equation 9 for the split-sample analysis. Panel (a) splits the sample based on the firm cloud index, defined as the share of IT workers with cloud-related skills; "high" is defined as above the median. Panel (b) splits the sample by firm age, firm size (measured by assets), firm sales growth, and firm competition (measured using the industry HHI index), with "old", "large", and "high" defined as above the median in each case.

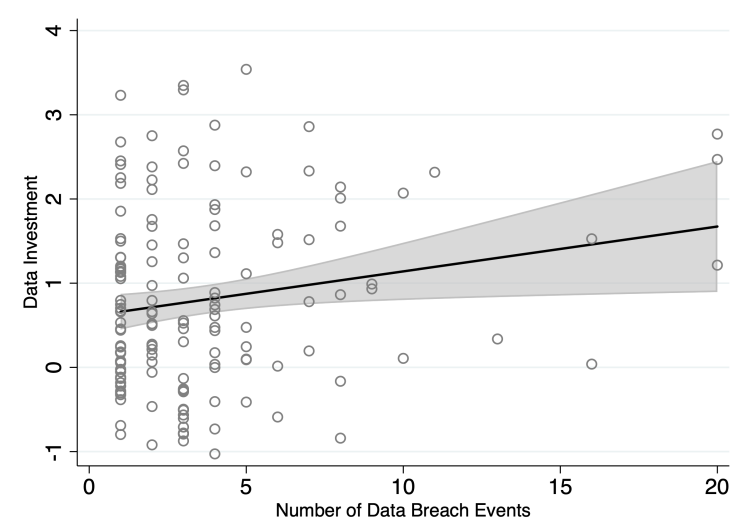
Figure 4: GDPR and Data Depreciation



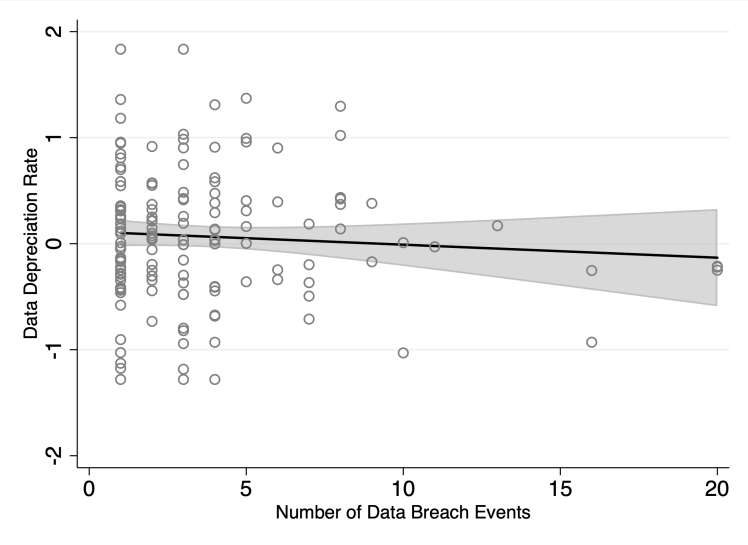
Notes. This figure shows the average data depreciation rate of firms with and without exposure to GDPR over time, using different measures. *Depreciation (Database Worker Ratio)* measures data depreciation as the share of database workers among database workers and data scientists. *GDPR Exposure* is defined as whether the firm had sales in Europe before 2018. Shaded areas represent 95% confidence intervals around the mean.

Figure 5: Data Investment, Data Depreciation, and Data Breach

(a) Data Investment and Data Breach



(b) Data Depreciation and Data Breach



Notes. This figure reports the distribution of firm data investment and depreciation rates by the number of data breach events. Data investment is proxied by the firm-level mean of log(Data), the log number of data workers employed by the firm across years. Data depreciation is measured using *Depreciation (Database Worker Ratio)*, defined as the share of database workers among database workers and data scientists. Data breach events are counted for each firm using data from the Privacy Rights Clearinghouse.

Tables

Table 1: Summary Statistics

Variable	Units	Mean	Std. Dev.	N
Log(Market Value)	Millions (USD)	8.388	1.99	20926
Log(Assets)	Millions (USD)	0.002	1.02	20926
Log(PPE)	Millions (USD)	-0.003	1.01	20926
Log(Employment)	Thousands (Employees)	-0.014	1.00	20926
Log(IT)	Thousands (Role count)	-0.002	1.01	20926
Log(Data)	Thousands (Role count)	0.031	1.01	20926
Log(AI)	Thousands (Role count)	0.030	1.01	20926
Depreciation (Database Worker Ratio)	Standardized Value	-0.078	0.96	20926
Depreciation (Product Fluidity)	Standardized Value	-0.051	0.81	20926
Depreciation (Worker Skill Obsolescence)	Standardized Value	0.060	0.87	15299
Depreciation (ROA Predictability)	Indicator	0.218	0.41	20926

Notes. This table reports summary statistics for firms in the regression panel constructed from the workforce data. The data source for the first four rows (Market Value, Assets, PPE, Employment, ROA Predictability) is the Capital IQ database. The measures (IT, Data, AI, Database Worker Ratio, Worker Skill Obsolescence) are constructed from the Revelio workforce database. *Depreciation (Database Worker Ratio)* measures data depreciation using the share of database workers among database workers and data scientists. *Depreciation (Product Fluidity)* proxies data depreciation using the product fluidity index derived from textual analysis of firms' 10-K product descriptions, which reflects year-to-year changes in product offerings. The data is share by Hoberg, Phillips, and Prabhala 2014. *Depreciation (Skill Obsolescence)* infers data depreciation by tracking the extent to which a firm's data-related skills become obsolete in the broader labor market after two years. *Depreciation (ROA Predictability)* is a dummy variable that equals 1 if the R^2 of an autoregressive model of return on assets (ROA) is smaller than 10%. All these measures are standardized.

Table 2: Data, Data Depreciation, and Firm Value

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log(Market Value)						
Log(Assets)	0.428***	0.420***	0.414***	0.410***	0.414***	0.374***	0.414***
	(0.071)	(0.067)	(0.067)	(0.065)	(0.067)	(0.056)	(0.066)
Log(PPE)	1.087***	1.111***	1.122***	1.122***	1.125***	1.195***	1.123***
	(0.133)	(0.104)	(0.102)	(0.107)	(0.101)	(0.073)	(0.102)
Log(Employment)	0.345***	0.303***	0.320***	0.332***	0.317***	0.271***	0.319***
	(0.073)	(0.063)	(0.058)	(0.066)	(0.058)	(0.061)	(0.058)
Log(IT)	0.422***	0.120	-0.038	0.017	-0.035	-0.019	-0.040
	(0.037)	(0.096)	(0.060)	(0.054)	(0.060)	(0.074)	(0.059)
Log(Data)		0.329***	0.207	0.167	0.206	0.154	0.219*
		(0.104)	(0.125)	(0.108)	(0.125)	(0.114)	(0.125)
Log(AI)			0.307***	0.288***	0.309***	0.345***	0.307***
			(0.064)	(0.061)	(0.066)	(0.047)	(0.064)
Log(Data) × Depreciation (Database Worker Ratio)				-0.061**			
				(0.027)			
Log(Data) × Depreciation (Product Fluidity)					-0.037*		
					(0.021)		
Log(Data) × Depreciation (Worker Skill Obsolescence)						-0.028**	
						(0.014)	
Log(Data) × Depreciation (ROA Predictability)							-0.045*
							(0.026)
Constant	8.396***	8.385***	8.380***	8.374***	8.381***	8.321***	8.385***
	(0.001)	(0.002)	(0.003)	(0.004)	(0.003)	(0.008)	(0.010)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.788	0.792	0.798	0.799	0.798	0.800	0.798
Observations	20926	20926	20926	20926	20926	15299	20926

Notes. This table reports regressions examining how data investment and data depreciation relate to firms' market value using a firm panel from 2013 to 2023. Each observation is at the firm-year level. Column (1) reports results from regressing market value on measures of IT investment, assets, capital (PPE), and employment, with all variables entered in logs. Column (2) adds data investment to the regression, and Column (3) introduces AI investment. Columns (4) to (7) examine how data depreciation moderates the market value assigned to data investment. *Depreciation (Database Worker Ratio)* measures data depreciation as the share of database workers among database workers and data scientists. *Depreciation (Product Fluidity)* proxies data depreciation using the product fluidity index derived from textual analysis of firms' 10-K product descriptions, which reflects year-to-year changes in product offerings; the data are shared by Hoberg, Phillips, and Prabhala 2014. *Depreciation (Skill Obsolescence)* infers data depreciation by tracking the extent to which a firm's data-related skills become obsolete in the broader labor market after two years. *Depreciation (ROA Predictability)* is a dummy variable equal to 1 if the R^2 of an autoregressive model of return on assets (ROA) is smaller than 10%. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table 3: GDPR and Data Depreciation

	(1)	(2)	(3)	(4)	(5)
	Log(Market Value)				
	GDPR Exposure = 1		GDPR Exposure = 0		
	Year < 2018	Year ≥ 2018	Year < 2018	Year ≥ 2018	
Model	OLS	OLS	OLS	OLS	IV
Log(Assets)	0.525***	0.482***	0.370***	0.330***	0.356***
	(0.109)	(0.096)	(0.062)	(0.056)	(0.009)
Log(PPE)	1.154***	1.323***	1.199***	1.339***	1.223***
	(0.119)	(0.134)	(0.121)	(0.109)	(0.028)
Log(Employment)	0.208**	0.176	0.295***	0.287***	0.297***
	(0.081)	(0.135)	(0.088)	(0.084)	(0.022)
Log(IT)	-0.006	-0.063	0.022	-0.024	0.075
	(0.098)	(0.115)	(0.075)	(0.120)	(0.064)
Log(AI)	0.247**	0.474***	0.219***	0.325***	0.286***
	(0.100)	(0.123)	(0.055)	(0.056)	(0.025)
Log(Data)	0.235	0.053	0.191*	0.127	0.115***
	(0.178)	(0.112)	(0.101)	(0.103)	(0.043)
Log(Data) × Depreciation (Database Worker Ratio)	-0.093***	-0.075	-0.013	-0.023	-0.265***
	(0.026)	(0.055)	(0.051)	(0.038)	(0.093)
Industry FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
F-Statistics					49.83
First Stage					-0.941***
					(0.096)
Adjusted R-squared	0.854	0.837	0.792	0.805	0.783
Observations	2783	2793	4795	4689	15063

Notes. This table reports how the moderating role of data depreciation varies across GDPR exposure and instrumental variable (IV) estimation. Each observation is at the firm-year level, and the specification follows Column (4) of Table 2. Columns (1) and (2) consider only firms exposed to GDPR, defined as having positive sales in Europe before 2018. Column (1) uses observations for these firms before 2018, and Column (2) uses observations after 2018. Columns (3) and (4) examine firms without exposure to GDPR. Column (5) implements an instrumental variable estimation, where the interaction between the ratio of European sales and an indicator for $Year > 2018$ is used as an instrument for data depreciation. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table 4: Data Breach and Data Depreciation

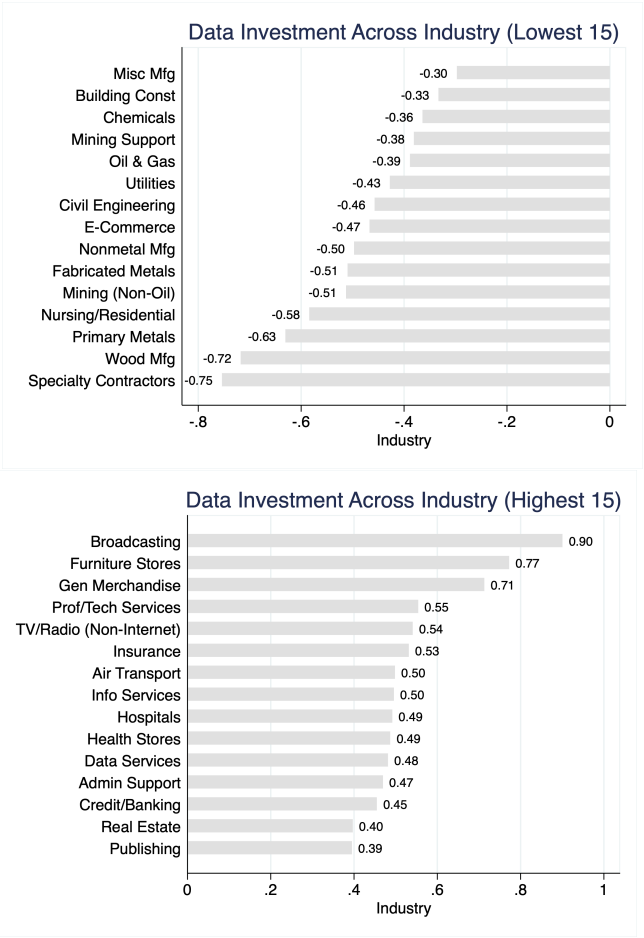
	(1)	(2)	(3)
	Log(Market Value)		
	Data Breach = 1	Data Breach = 0	
Model	OLS	OLS	IV
Log(Assets)	0.308***	0.435***	0.414***
	(0.099)	(0.070)	(0.011)
Log(PPE)	0.919***	1.122***	1.028***
	(0.191)	(0.109)	(0.032)
Log(Employment)	0.425**	0.342***	0.505***
	(0.192)	(0.068)	(0.059)
Log(IT)	-0.318	0.002	0.775**
	(0.215)	(0.056)	(0.301)
Log(AI)	0.495***	0.282***	0.485***
	(0.102)	(0.066)	(0.101)
Log(Data)	0.409***	0.180	-0.849**
	(0.120)	(0.110)	(0.422)
Log(Data) × Depreciation (Database Worker Ratio)	-0.004	-0.053*	-0.331***
	(0.081)	(0.030)	(0.101)
Industry FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
F-Statistics			18.62
First Stage			0.014***
			(0.002)
Adjusted R-squared	0.913	0.789	0.730
Observations	863	18485	20926

*Notes.*This table reports how the moderating role of data depreciation varies across data breach events and instrumental variable (IV) estimation. Each observation is at the firm-year level, and the specification follows Column (4) of Table 2. Column (1) considers only firms that have experienced at least one data breach event. Column (2) examines firms that have never experienced any data breach events. Column (3) implements an instrumental variable estimation, where the interaction between the number of a firm's data breach events and an indicator for Year $>$ the first year of a data breach event is used as an instrument for data investment. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Online Appendix

A Appendix Figures

Figure A.1: Industry Distribution of Data Investment



Notes. This figure reports the distribution of average firm data investment across NAICS 3 digit industries in our sample. Industry with fewer than 50 firms are dropped. Data investment is measured as the number of workers whose primary roles are data scientists or database workers employed by a firm in a given year.

B Appendix Tables

Table A.1: Robustness Check

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log(Market Value)						
	Firm FEs		Exclude IT Firms		Other Definition		Placebo
Log(Assets)	0.178***	0.178***	0.375***	0.371***	0.415***	0.411***	0.412***
	(0.034)	(0.034)	(0.064)	(0.062)	(0.067)	(0.065)	(0.066)
Log(PPE)	0.356***	0.358***	1.110***	1.115***	1.122***	1.122***	1.128***
	(0.054)	(0.054)	(0.113)	(0.120)	(0.101)	(0.107)	(0.102)
Log(Employment)	0.975***	0.975***	0.315***	0.329***	0.318***	0.330***	0.322***
	(0.080)	(0.080)	(0.071)	(0.081)	(0.057)	(0.065)	(0.061)
Log(IT)	0.133**	0.139**	-0.079	-0.038	-0.032	0.022	-0.002
	(0.058)	(0.058)	(0.074)	(0.068)	(0.062)	(0.055)	(0.057)
Log(Data)	0.032	0.022	0.301**	0.261**			0.183
	(0.031)	(0.031)	(0.140)	(0.117)			(0.118)
Log(AI)	0.056**	0.056**	0.258***	0.246***	0.326***	0.304***	0.293***
	(0.026)	(0.025)	(0.078)	(0.074)	(0.056)	(0.054)	(0.062)
Log(Data) × Depreciation (Database Worker Ratio)		-0.009		-0.066*			
		(0.014)		(0.039)			
Log(Data) - Exclude AI Workers					0.113	0.090	
					(0.074)	(0.064)	
Log(Data) - Exclude AI Workers × Depreciation (Database Worker Ratio)						-0.039**	
						(0.018)	
Log(IT) × Depreciation (Database Worker Ratio)							-0.016
							(0.014)
Constant	8.405***	8.404***	8.383***	8.371***	8.048***	8.111***	8.378***
	(0.002)	(0.002)	(0.012)	(0.010)	(0.220)	(0.189)	(0.004)
Industry FE	No	No	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.955	0.955	0.797	0.799	0.798	0.799	0.799
Observations	20616	20616	16524	16524	20926	20926	20926

Notes. This table reports robustness check results. Each observation is at the firm-year level, and the specification follows Columns (3) and (4) of Table 2. Columns (1) and (2) use firm fixed effects instead of industry fixed effects. Columns (3) and (4) exclude IT firms with NAICS codes beginning with "51" or "54." Columns (5) and (6) exclude workers with AI skills from the data worker category. Column (7) interacts the data depreciation rate with firms' IT investment. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table A.2: Robustness Check - Outsource Data Tasks

	(1)	(2)	(3)	(4)
	Log(Market Value)			
	Outsource Data Tasks = 1		Outsource Data Tasks = 0	
Log(Assets)	0.343***	0.343***	0.421***	0.417***
	(0.103)	(0.103)	(0.066)	(0.065)
Log(PPE)	1.286***	1.273***	1.104***	1.106***
	(0.123)	(0.139)	(0.102)	(0.106)
Log(Employment)	0.301**	0.313**	0.325***	0.335***
	(0.116)	(0.130)	(0.058)	(0.064)
Log(IT)	-0.178	-0.079	-0.027	0.024
	(0.131)	(0.139)	(0.060)	(0.053)
Log(Data)	-0.041	-0.091	0.220*	0.181*
	(0.132)	(0.154)	(0.124)	(0.105)
Log(AI)	0.529***	0.499***	0.297***	0.278***
	(0.105)	(0.093)	(0.065)	(0.062)
Log(Data) × Depreciation (Database Worker Ratio)		-0.087		-0.060**
		(0.058)		(0.030)
Constant	8.469***	8.456***	8.383***	8.376***
	(0.051)	(0.049)	(0.004)	(0.005)
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.867	0.868	0.782	0.783
Observations	1338	1338	19581	19581

Notes. This table reports how the moderating role of data depreciation varies across data task outsourcing. Each observation is at the firm-year level, and the specification follows Columns (3) and (4) of Table 2. Columns (1) and (2) restrict the sample to firms that outsource data tasks, defined as those purchasing or licensing data-related products or services. *Depreciation (Database Worker Ratio)* measures data depreciation as the share of database workers among database workers and data scientists. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table A.3: Data, Data Depreciation, and Firm Productivity

	(1)	(2)	(3)	(4)
	Firm Sales (Productivity)			
Log(Assets)	0.231***	0.236***	0.240***	0.235***
	(0.027)	(0.028)	(0.030)	(0.030)
Log(PPE)	1.377***	1.360***	1.353***	1.320***
	(0.169)	(0.143)	(0.137)	(0.111)
Log(Employment)	0.501***	0.530***	0.518***	0.553***
	(0.084)	(0.062)	(0.063)	(0.041)
Log(IT)	0.337***	0.545**	0.657**	0.704***
	(0.084)	(0.220)	(0.257)	(0.257)
Log(Data)		-0.226	-0.140	-0.180
		(0.152)	(0.120)	(0.145)
Log(AI)			-0.217**	-0.229***
			(0.090)	(0.081)
Log(Data) × Depreciation (Database Worker Ratio)				-0.164**
				(0.072)
Industry FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.845	0.846	0.848	0.850
Observations	20926	20926	20926	20926

Notes. This table reports regressions examining how data investment and data depreciation relate to firms' productivity (as measured by total sales) using a firm panel from 2013 to 2023. Each observation is at the firm-year level. Column (1) reports results from regressing market value on measures of IT investment, assets, capital (PPE), and employment, with all variables entered in logs. Column (2) adds data investment to the regression, and Column (3) introduces AI investment. Columns (4) examines how data depreciation moderates the market value assigned to data investment. *Depreciation (Database Worker Ratio)* measures data depreciation as the share of database workers among database workers and data scientists. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

C Classification of Skills and Roles

A full list of data scientist roles: *analytics, business analytics, business data analyst, business intelligence, business intelligence analyst, business intelligence consultant, clinical data, data, data analyst, data analytics, data science, data scientist, information analyst, machine engineer, reporting analyst, statistical programmer.*

A full list of database worker roles: *data architect, data center, data engineer, database, database administrator, database analyst, database developer, database engineer, dba, etl developer, oracle database administrator, oracle dba, oracle developer, sql server dba*

D A Model of Data Depreciation

This appendix develops a simple model to discipline the measurement of data depreciation and to motivate the empirical proxies used in the paper. The model follows the measurement logic of the intangible capital literature, while adapting it to data, an intangible asset whose economic value depends critically on information persistence and organizational maintenance. We provide detailed derivations and state all assumptions explicitly.

D.1 Timeline, Set-up, and Interpretation

Timeline within period t. Each period proceeds as follows:

1. The firm enters period t with a predetermined stock of effective data D_t.
2. The firm observes current conditions (including the realized state s_t) and chooses contemporaneous inputs for production and value extraction.
3. The firm chooses maintenance effort (database labor) that produces data investment I_t, which together with depreciation determines D_{t+1}.

For the purposes of a toy model, we treat the choices of contemporaneous inputs and data labor within the same period; what matters is that D_t is predetermined for contemporaneous production, while maintenance affects the transition to D_{t+1}.

Information environment. Let s_t denote a latent state relevant for the firm (demand, preferences, business processes). The state evolves as:

$$s_{t+1} = \rho s_t + \mu_{t+1}, \qquad |\rho| < 1, \qquad \mathbb{E}[\mu_{t+1}] = 0, \quad \mathbb{E}[\mu_{t+1}^2] = \sigma_\mu^2. \tag{10}$$

Production function. The firm produces output using physical capital K_t, standard labor L_t, and an effective stock of data D_t:

$$Y_t = A_t K_t^{\alpha} L_t^{\beta} D_t^{\gamma}, \qquad 0 < \alpha, \beta, \gamma < 1, \qquad \alpha + \beta < 1. \tag{11}$$

The firm is a price taker. The rental rate of physical capital is $r > 0$ and the wage for standard labor is $w > 0$. D_t is predetermined within period t for contemporaneous production. It is chosen through past maintenance and depreciates between t and $t+1$.

Effective data and raw data. Let x_t (where $x_t = s_t + \varepsilon_t$) denote the *raw* data contribution collected in period t (e.g., new observations, newly labeled records). Define effective data as:

$$D_t \equiv \sum_{k=0}^{\infty} \rho^k x_{t-k}. \tag{12}$$

Older data is less informative for the current environment when $|\rho| < 1$.

Data labor: data scientists versus database worker We distinguish two types of data-related labor:

- **Data science labor** L_t^{ds}: analyzes existing data and affects contemporaneous value extraction.
- **Database labor** L_t^{db}: collects, cleans, integrates, and maintains data and affects the future data stock.

The effective data stock evolves as:

$$D_{t+1} = (1 - \delta_t) D_t + I_t. \tag{13}$$

Data investment is produced by database labor:

$$I_t = \eta_t L_t^{db}. \tag{14}$$

D.2 From Cobb–Douglas to a Reduced-Form Profit Function

Given D_t, the firm chooses K_t and L_t to maximize period-t profit:

$$\pi_t(K_t, L_t; D_t) = A_t K_t^{\alpha} L_t^{\beta} D_t^{\gamma} - rK_t - wL_t. \tag{15}$$

We assume the optimal choices satisfy $K_t^* > 0$ and $L_t^* > 0$. Under Cobb–Douglas and competitive factor markets this holds for standard parameter values.

Differentiating (15) yields:

$$\frac{\partial \pi_t}{\partial K_t} = \alpha A_t K_t^{\alpha-1} L_t^{\beta} D_t^{\gamma} - r = 0, \tag{16}$$

$$\frac{\partial \pi_t}{\partial L_t} = \beta A_t K_t^{\alpha} L_t^{\beta-1} D_t^{\gamma} - w = 0. \tag{17}$$

Profit Maximization. Divide (16) by (17), the optimal ratio satisfies:

$$\frac{L_t}{K_t} = m \equiv \frac{\beta}{\alpha} \cdot \frac{r}{w}, \qquad \text{so } L_t = mK_t. \tag{18}$$

Substitute $L_t = mK_t$ into (16), we will have K_t^* and L_t^*, and K_t^* and L_t^* each scale with $D_t^{\frac{\gamma}{1-\alpha-\beta}}$, then we have

$$(K_t^*)^{\alpha}(L_t^*)^{\beta} \propto \left(D_t^{\frac{\gamma}{1-\alpha-\beta}}\right)^{\alpha+\beta} = D_t^{\frac{\gamma(\alpha+\beta)}{1-\alpha-\beta}}.$$

Thus

$$Y_t^* \propto D_t^{\gamma + \frac{\gamma(\alpha+\beta)}{1-\alpha-\beta}} = D_t^{\frac{\gamma}{1-\alpha-\beta}}. \tag{19}$$

Under Cobb–Douglas and competitive factor markets, optimal factor payments $rK_t^* + wL_t^*$

are proportional to output, so optimal profit inherits the same scaling in D_t. Therefore we write:

$$\pi_t^*(D_t) = \tilde{A}_t D_t^{\vartheta}, \qquad \vartheta \equiv \frac{\gamma}{1-\alpha-\beta} > 0, \tag{20}$$

where $\tilde{A}_t$ collects constants involving $(A_t, r, w, \alpha, \beta)$.

Complementarity between Data and Data Scientist. To generate value from data, firms must process and interpret it. We model this with a simple multiplicative form:

$$\pi_t(D_t, L_t^{ds}) = \tilde{A}_t D_t^{\vartheta} (L_t^{ds})^{\kappa} - w_{ds} L_t^{ds} - w_{db} L_t^{db}, \qquad 0 < \kappa < 1. \tag{21}$$

Equation (21) is a reduced form of a Cobb–Douglas environment in which conventional inputs have been optimized out and data must be processed by analytical labor to generate economic value.

D.3 How to Estimate Data Depreciation

State persistence Using (12),

$$D_{t+1} = \sum_{k=0}^{\infty} \rho^k x_{t+1-k} = x_{t+1} + \sum_{k=1}^{\infty} \rho^k x_{t+1-k}.$$

Re-index the sum with $j = k - 1$:

$$\sum_{k=1}^{\infty} \rho^k x_{t+1-k} = \sum_{j=0}^{\infty} \rho^{j+1} x_{t-j} = \rho \sum_{j=0}^{\infty} \rho^j x_{t-j} = \rho D_t.$$

Hence:

$$D_{t+1} = \rho D_t + x_{t+1}. \tag{22}$$

Comparing (22) with (13), we obtain the structural mapping:

$$1 - \delta_t \equiv \rho \qquad \Rightarrow \qquad \delta_t = 1 - \rho. \tag{23}$$

Lower persistence ρ implies faster information decay and higher data depreciation δ_t. This is the microfoundation for proxies that capture environmental instability (e.g., product fluidity) or outcome persistence (e.g., ROA predictability).

Processing efficiency Let θ_t denote the data-processing technology (tools, pipelines, standards, schemas). The productivity of database labor depends on how compatible current technology is with existing data infrastructure. We model this reduced-form as:

$$\eta_t = \eta(\theta_t) \tag{24}$$

A faster evolution of θ_t that reduces compatibility and thus lowers η_t:

$$\frac{\partial \eta_t}{\partial \theta_t} < 0. \tag{25}$$

When skills and tools evolve rapidly, a given amount of database labor generates fewer units of reusable effective data (lower η_t), because more effort must be devoted to reformatting, rebuilding pipelines, and ensuring compatibility.

Labor allocation To illustrate the maintenance implications of depreciation, consider maintaining a constant effective data stock $D_{t+1} = D_t = \bar{D}$.

From the data accumulation equation (13) and the investment technology (14), maintaining $D_{t+1} = D_t = \bar{D}$ implies the maintenance condition

$$L^{db} = \frac{\delta_t}{\eta_t}\bar{D}. \tag{26}$$

Substituting (26) into the contemporaneous profit function 21 yields the firm's problem

$$\max_{L^{ds}\geq 0} \ \tilde{A}_t \bar{D}^{\vartheta}(L^{ds})^{\kappa} - w_{ds}L^{ds} - w_{db}\frac{\delta_t}{\eta_t}\bar{D}, \qquad 0 < \kappa < 1. \tag{27}$$

Since the last term does not depend on L^{ds}, the first-order condition is

$$\kappa \tilde{A}_t \bar{D}^{\vartheta} (L^{ds})^{\kappa-1} = w_{ds}, \tag{28}$$

which implies the optimal level of data science labor

$$L^{ds*} = \left(\frac{\kappa \tilde{A}_t}{w_{ds}} \right)^{\frac{1}{1-\kappa}} \bar{D}^{\frac{\vartheta}{1-\kappa}}. \tag{29}$$

D.4 Structural Parameters and Empirical Measures

The model yields three structural primitives related to effective data depreciation:

- **State persistence** ρ, information decay.
- **Processing efficiency** η_t, maintenance effectiveness.
- **Effective Depreciation** $\tilde{\delta}_t = \delta_t / \eta_t$, maintenance-adjusted information decay rate..

These primitives map into the empirical measures used in the paper:

- **Database labor share** captures variation in $\tilde{\delta}_t$ through $\frac{L^{db}}{L^{db}+L^{ds*}} = \frac{\frac{\delta_t}{\eta_t}\bar{D}}{\frac{\delta_t}{\eta_t}\bar{D}+L^{ds*}} \propto \tilde{\delta}$
- **Product fluidity** captures lower ρ (less persistent environments), implying higher δ_t.
- **Outcome predictability** (e.g., ROA persistence or AR(1) R^2) provides an outcome-based proxy for ρ and thus for δ_t.
- **Skill obsolescence** captures declines in η_t due to rapid changes in data-processing technologies, increasing effective depreciation $\tilde{\delta}_t$.

E Alternative Measures of Data Depreciation

This appendix describes the construction of three supplementary proxies for firm-level data depreciation. These measures are designed to separately capture the two conceptual components of effective data depreciation, $\tilde{\delta}_t = \delta_t/\eta_t$: (i) information decay due to environmental change (δ_t) and (ii) the firm's organizational capacity to maintain and update the usable value of data (η_t).

E.1 Product Fluidity

The measure draws on the product fluidity index developed by Hoberg, Phillips, and Prabhala (2014), which is constructed from textual analysis of 10-K product descriptions and captures the extent to which a firm's product offerings change from year to year.

We interpret higher product fluidity as indicating a more dynamic and volatile product environment, where market demands, technologies, and competitive positioning evolve rapidly. In such settings, previously collected data become obsolete more quickly because they reflect outdated products, customer needs, or market conditions. Therefore, this measure primarily proxies for higher information decay (δ_t), holding constant the firm's internal data management capability.

E.2 Predictive Persistence of Performance

The second measure follows Liu, Ma, and Veldkamp (2025) and captures the extent to which historical information remains predictive of future business performance. The underlying idea is that if past information continues to explain future outcomes well, the informational content of data depreciates slowly.

To operationalize this, we estimate an AR(1) model for each firm:

$$\mathrm{ROA}_{i,t} = \alpha_i + \rho\,\mathrm{ROA}_{i,t-1} + \varepsilon_{i,t}.$$

The R^2 from this regression reflects the predictive power of past performance for current outcomes. A lower R^2 indicates that historical information quickly loses relevance, suggesting a higher rate of information decay (δ_t). This measure therefore captures the speed at which past firm-level data become obsolete due to changing economic conditions.

E.3 Skill-Based Shift–Share Measure

The third measure is based on a shift–share approach using data worker skills. For each firm-year, we first identify the composition of data-related skills among its employees (the "share"). Skills are defined based on standardized skill tags associated with data-related occupations, such as database management systems, data engineering tools, and analytics platforms. The share component reflects the firm's internal stock of data management capabilities embodied in its workforce.

We then combine this with the rate of change in the usage of the same skills across all other firms over a given time window (the "shift"). Specifically, for each skill s, we compute the external rate of decline or replacement in its usage in the overall labor market. This captures how rapidly particular data tools, technologies, or practices are being superseded by newer ones.

The firm-level index is constructed as a weighted average of these skill-level obsolescence rates, where the weights correspond to the firm's own skill composition. Intuitively, the measure reflects how exposed a firm is to externally driven technological change in data management practices. Firms whose workforce relies more heavily on rapidly declining skills face greater challenges in maintaining and updating their data infrastructure, implying lower organizational maintenance effectiveness (η_t). Consequently, this measure maps primarily to the denominator of effective data depreciation, with higher values indicating lower η_t and thus higher $\tilde{\delta}_t$.