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

Hongyuan Xia
hx276@cornell.edu

Prasanna Tambe
tambe@wharton.upenn.edu

September 3, 2026

Abstract

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. At the industry level, this data depreciation measure ranges from 17% in Chemicals to 63% in Clothing Stores. We find that data investment is positively associated with market value, but this value is significantly discounted when data depreciates more rapidly. Exploiting variation from GDPR and data breach events, we provide evidence consistent with this interpretation. A back-of-the-envelope exercise suggests a $455 billion annual loss in market value due to data depreciation across U.S. public firms.

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; Veldkamp and Chung 2024)). Data can improve prediction, targeting, coordination, and learning within firms. Unlike physical assets, data is often non-rival: the same dataset can support multiple uses across products, functions, and periods without being depleted (Farboodi and Veldkamp 2023; Jones and Tonetti 2020). This property helps to explain why data can generate scale economies, reinforce competitive advantage, and increase returns to complementary investments in analytics, software, organizational capital, and artificial intelligence (Brynjolfsson et al. 2011; Brynjolfsson et al. 2012; McElheran and Brynjolfsson 2016). However, the economic value of data is not realized automatically. It depends on its quality, exclusivity, recency, legal usability, and whether it can be combined with other complementary assets (Bergemann et al. 2022; Hagiu and Wright 2023). These features make data central to production but difficult to value using conventional accounting measures. These measurement challenges have gained importance because of recent, growing investment in artificial intelligence, where performance often depends on access to large volumes of relevant and high-quality data (McElheran et al. 2024).

The economic treatment of data assets fits within growing levels of academic interest in intangible capital. A large literature in this area argues that modern firms derive increasing value from assets not fully captured on balance sheets, including software, R&D, organizational capital, brands, and other knowledge-based assets (Corrado et al. 2022; Corrado et al. 2009; Eisfeldt and Papanikolaou 2013; Hall et al. 2010; Haskel and Westlake 2017; Lev 2001; Tambe et al. 2020). Like these assets, databases are costly to create, difficult to observe, complementary with human and organizational capabilities, and often excluded from

standard capital measures (Bhutani et al. 2025).

This literature has focused on data accumulation and valuation, but there has been less attention to data depreciation. In striking constrast, the literature on R&D assets, another critically important form of knowledge stock, treats depreciation as central to constructing capital stocks and understanding growth, while also highlighting the difficulty of measuring depreciation directly (Griliches 1979; Hall et al. 2010; Ma 2025; Peters and Taylor 2017). A large and influential literature in this area spanning decades has focused on the economic properties of R&D investments that can impact their value, including non-rival and uncertain nature of the asset, and its depreciation.

Like R&D capital, data can lose relevance, accuracy, or predictive power. For instance, in machine learning, ideas related to depreciation have been discussed under the labels of "data drift" or "concept drift," but their economic implications for firms remain under-developed (Lu et al. 2018). Understanding the rate at which data depreciates is important for managers and policymakers because it determines the persistence of value. If data remains useful for long periods, data investment creates a durable asset that can support future production and growth. If assets depreciate rapidly, firms must continually reinvest simply to maintain their value. For instance, structures and buildings are commonly estimated to depreciate at annual rates of 2–4%, whereas R&D is often estimated to depreciate at 15–20%. These differences imply that investments in physical and knowledge capital have very different consequences for the persistence and time value of the resulting assets. Recent evidence suggests that data becomes less valuable over time, and in many settings may reduce model accuracy, highlighting the importance of recency and updating for data-intensive production (Valavi et al. 2022).

Estimating the rate of data depreciation poses several unique challenges. First, the stock of data held by a firm is difficult to observe and measure, as firms rarely disclose data inventories on their balance sheets. Without a reliable measure of the firm's data assets, 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 (similarly, the R&D literature has shown that R&D in different sectors depreciates at different rates (Li and Hall 2020), with high-tech R&D depreciating particularly rapidly). 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, integration, and governance. As a result, otherwise similar data may depreciate at very different rates depending on a firm's data management capabilities.

This paper develops a new approach to measuring the depreciation of a firm's data assets. Prior research documents the economic importance of data and firms' investments in data-related assets (Tambe 2014). We follow this literature in developing a proxy measure of a firm's data assets using investments in data-related labor. 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, updating, and integrating data assets. 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 proxy measures for investments in the labor needed to develop and maintain these assets, constructed using employee resume data from Revelio Labs. To develop a proxy measure for data depreciation, we classify data workers into (i) workers who maintain data assets and build new data pipelines or datasets (database workers), and (ii) workers who use data for analysis and decision-making (data scientists or data analysts). We use the ratio of database workers to total data workers as our depreciation proxy. Intuitively, we argue that when data depreciation is higher, firms devote a larger share of data labor to maintaining, updating, or rebuilding data assets to sustain their usefulness.

We validate our depreciation proxy in three ways. First, occupational descriptions and workers' observed skill profiles confirm that the two groups capture distinct data-related activities: database roles emphasize data infrastructure, integration, pipeline development, and maintenance, whereas data-science roles emphasize analysis, modeling, and decision-making. Second, three conceptually distinct alternative proxies—based on product-market fluidity, declining performance predictability, and data-skill obsolescence—produce qualitatively similar valuation discounts. Third, the cross-industry distribution of the proxy is consistent with economic intuition. The unstandardized measure ranges from 0.17 in Chemicals to 0.63 in Clothing Stores and is generally higher in industries with rapidly changing demand, products, or operations, such as retail and telecommunications, and lower in sectors with more stable information environments, such as chemicals, education, manufacturing, and hospitals.

We use these new measures to study the relationship between depreciation and firms' market value. If data depreciates slowly, a firm's data assets can continue to generate long-lived returns into the future. 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. Consistent with this logic, we find that the capital market assigns a positive and significant value to firms' data investment, and that this value is larger for investments in frontier data applications, particularly related to artificial intelligence, suggesting that not all data-related investments are valued equally. Our main result is that the value the market attributes to data assets diminishes in contexts where data depreciates rapidly. A one-standard-deviation increase in our data depreciation rate measure is associated with a 36.5 percent reduction in the market value attributed to data investment. For a firm with the median market value in our sample, this corresponds to approximately $131 million per year. Extrapolating this estimate across the 3,477 firms in our sample implies a sample-wide annual loss of roughly $455 billion in market value due to data depreciation.

Our findings contribute to three literatures. First, we contribute to research on the economics of data by providing cross-industry evidence on how capital markets value data

investment and how this value depends on depreciation (Bergemann et al. 2022; Farboodi and Veldkamp 2023; Hagiu and Wright 2023; Jones and Tonetti 2020). Second, we contribute to the literature on intangible capital by developing a firm-level measure of depreciation for a major intangible asset that is largely missing from standard accounting measures (Corrado et al. 2022; Corrado et al. 2009; Hall et al. 2010; Haskel and Westlake 2017; Lev 2001). Third, we contribute to the literature on IT value and data-driven decision-making by showing that variation in the durability of data assets is an important source of heterogeneity in how data and AI-related investments translate into market value (Brynjolfsson et al. 2011; Brynjolfsson et al. 2012; McElheran and Brynjolfsson 2016; Tambe 2014).

2 Theory and Measurement Approach

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 et al. 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.

A related case is personal data, whose value is closely shaped by institutional and technological conditions. Personal data enables firms to engage in targeted advertising, personalize products, price more effectively, and improve prediction, but its use also creates privacy costs and regulatory constraints (Goldfarb and Tucker 2011; Lei et al. 2026; Sun et al. 2024). Evidence from online advertising shows that restrictions on the use of personal data can materially reduce the effectiveness of targeting (Goldfarb and Tucker 2011). More broadly, privacy regulation changes the conditions under which firms can collect, retain, combine, and reuse personal data. The General Data Protection Regulation (GDPR), for instance, affected firms' use of data in advertising, analytics, and digital operations, with consequences for website traffic, revenue, technology use, and data-dependent business models (Aridor et al. 2023; Demirer et al. 2024; Johnson et al. 2022; Peukert et al. 2022). These findings

reinforce the view that the value of data depends not only on the information contained in existing data stocks, but also on the institutional conditions governing access, consent, linkage, and reuse.

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 affect 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 et al. 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 The Literature on R&D Capital Stock

The literature on depreciating R&D capital stock finds broad agreement that knowledge capital, while a key driver of productivity and growth, depreciates substantially over time through technological obsolescence, competitive imitation, and shifting market conditions. Early macro and productivity studies—particularly those following the perpetual inventory approach pioneered by Griliches (1979)—treat R&D as an intangible capital stock that accumulates with investment but decays at a constant rate, typically assuming annual de-

preciation rates between 10%–25%. Empirical evidence suggests this depreciation is highly heterogeneous across industries: R&D in fast-moving sectors such as software, semiconductors, and biotech depreciates much more rapidly than in pharmaceuticals or industrial manufacturing, where knowledge remains valuable longer. More recent work emphasizes that depreciation is not merely physical "wear and tear" but reflects the declining economic usefulness of knowledge as technologies evolve, patents expire, and frontier innovation shifts. This perspective has become especially important in digital and AI-intensive sectors, where model architectures, datasets, and algorithmic innovations can lose value in months rather than years.

The central challenge in this literature is measurement. Unlike physical capital, R&D generates intangible assets that are difficult to observe directly, and estimated depreciation rates depend heavily on modeling assumptions. Most studies infer depreciation indirectly from market value, patent citations, productivity, or firm performance, each of which captures only part of the underlying knowledge stock. A major unresolved issue is whether depreciation should be modeled as a constant geometric rate or as endogenous to technological change, competition, and diffusion. Emerging evidence suggests R&D capital may follow nonlinear depreciation patterns, with slow early decay followed by rapid obsolescence when a new technological paradigm arrives. This is particularly salient for AI and data-driven innovation, where complementary assets—such as worker skills, proprietary data, and organizational capabilities—may depreciate on different timelines than formal R&D expenditures. As a result, a key frontier in the literature is moving from static depreciation parameters toward dynamic models that jointly capture innovation, diffusion, and obsolescence of intangible capital.

2.3 Conceptual Framework for Effective Data Depreciation

A large literature on intangible and R&D capital treats observed R&D expenditures as gross investment flows and uses perpetual-inventory methods to construct stocks of knowledge

capital. Because the service lives of individual R&D projects are rarely observable, standard implementations aggregate heterogeneous projects into a firm- or industry-level stock and apply a common geometric depreciation rate as a simplifying accounting assumption, although more recent work allows depreciation to vary across industries or over time (Griliches 1979; Hall et al. 2010; Li and Hall 2020; Peters and Taylor 2017). Related 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 et al. 2025).

Following this investment-accounting logic, we define *data depreciation* as the decline in the productive value of a firm's data assets over time. If data stocks and related expenditures were directly observable, we would distinguish three types of expenditure: (a) expansion investment, which increases the data stock; (b) replacement and maintenance investment, which preserves the usability of existing data assets; and (c) analytics expenditure, which extracts value from the data stock without maintaining it. When the effective data stock is held constant, replacement and maintenance investment corresponds to effective depreciation, and the depreciation rate is this amount divided by the beginning-of-period data stock. This replacement need reflects both the obsolescence of informational content and the organizational effort required to clean, update, store, integrate, and govern data.

This accounting exercise is especially difficult for data assets because neither data-related expenditures nor the underlying effective data stock is directly observable (Tambe et al. 2020). We therefore do not construct a literal data stock or estimate an effective depreciation rate by imposing an exogenous depreciation parameter. Instead, we infer an *effective depreciation proxy* from firms' observable allocation of data-related labor.

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 Appendix E.

Value creation from data. Data enhances firm performance by improving decision quality, 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. A detailed derivation can be found in Appendix E.

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 accu-

mulation. This two-part distinction maps the accounting decomposition above into observable labor inputs: database labor corresponds to expansion and replacement or maintenance activities that affect the data stock, whereas data-science labor corresponds to analytics expenditure that extracts value from the existing stock. 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

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

The model therefore provides a monotonic mapping between effective data depreciation and the share of data-related labor allocated to database work. Section 3.3 explains how we operationalize this mapping, compares the resulting proxy with an ideal investment-based measure, and discusses the assumptions and validation exercises underlying its interpretation.

3 Data and Research Design

3.1 Market Value Specification

Prior work has used market value regressions to uncover valuable but otherwise invisible (intangible) assets that contribute to a firm's productive capacity (Brynjolfsson et al. (2002)). This approach may be particularly useful for analyzing data investments, as firms often require time to adapt to new technologies, and the literature suggests that in the AI context, 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 are reflected in the firm's market valuation before they appear in productivity statistics. The market value regression used to estimate how depreciation affects the value of a firm's data assets is:

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

We distinguish among three conceptually different, though not mutually exclusive, digital

inputs. IT labor captures firms' broad digital infrastructure and general technical capabilities, data labor captures the specialized human capital used to build, maintain, and analyze firm data, and AI labor captures the algorithmic capabilities used to develop and deploy AI applications. In practice, data workers are part of the broader IT workforce and overlap to some extent with AI workers. Our baseline specifications therefore include all three measures jointly, so that the coefficient on data investment reflects its association with market value conditional on firms' IT and AI capabilities. In robustness tests, the results are unchanged when we redefine data labor to exclude AI workers, and placebo tests indicate that the effect of data depreciation is distinct from that of broader IT investment.

3.2 Data Sources

The empirical tests described above are informed by data sources on (a) firms' investments 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: Revelio Labs. 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 employment 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[1] 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 et al. (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.

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.[2] 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 how they are entered into

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

2. We focus on 2013–2023, the period for which our matched Revelio–Capital IQ panel provides the most complete and consistent coverage.

the multivariate regressions. Variables are standardized to aid interpretability.

3.3 Data Depreciation Proxy

Ideal measure and investment mapping. If data stocks and related expenditures were directly observable, we would distinguish three types of expenditure: (a) expansion investment, which increases the data stock; (b) replacement and maintenance investment, which preserves the usability of existing data assets; and (c) analytics expenditure, which extracts value from the data stock without maintaining it. Effective depreciation would equal the replacement and maintenance investment required to keep the effective data stock unchanged, and the effective depreciation rate would be this amount divided by the beginning-of-period data stock. In practice, neither the effective data stock nor these expenditure categories is directly observable.

Primary measure and interpretation. Our primary proxy for effective data depreciation is the ratio of database workers to the total number of data workers employed by a firm in a given year:

$$\text{Database Worker Share}_{it} = \frac{L_{it}^{db}}{L_{it}^{db} + L_{it}^{ds}}.$$

Database workers primarily collect, clean, update, integrate, and maintain data, so their labor maps most closely to replacement and maintenance investment. Data scientists primarily analyze existing data and therefore represent a complementary input used to extract value from the data stock.

As developed in Section 2.3, effective data depreciation, $\widetilde{\delta}_t = \delta_t/\eta_t$, reflects both information decay and the firm's capacity to maintain the usable value of its data. When effective depreciation is higher, firms must allocate a larger share of data labor to maintaining, updating, or rebuilding data assets, thereby increasing the database-worker share.

This interpretation relies on three assumptions. (1) Occupational categories provide an informative mapping from workers to tasks: database workers primarily maintain data assets,

whereas data scientists primarily analyze them, so occupational composition reflects firms' allocation of resources between maintenance and value extraction. (2) Under the model's constant-stock benchmark, expansion investment is set to zero and the effective data stock is held fixed. (3) Headcount shares approximate expenditure shares, despite differences in wages and unobserved nonlabor expenditures. Because firms hold multiple datasets that depreciate at different rates, the measure should be interpreted as an aggregate firm-level proxy rather than as a direct depreciation rate for any individual dataset.

Measure construction. We construct the measure using occupation data from Revelio Labs. Revelio provides approximately 1,500 classified occupations, each of which we manually review using job titles, job descriptions, and O*NET classifications. We classify data-related occupations into database workers, who are primarily responsible for building and maintaining data assets, and data scientists, who focus on analyzing data and extracting insights. For example, "database engineer" is classified as a database worker, while "data analyst" is classified as a data scientist.[3] We then aggregate worker counts to the firm-year level and calculate the database-worker share as the number of database workers divided by the total number of database workers and data scientists.

Measure validation. We first assess whether the occupational classification captures the intended distinction between data maintenance and analysis. O*NET descriptions provide an external comparison of the tasks performed by the two occupational groups. Database occupations emphasize database administration, integration, availability, and maintenance, whereas data-science occupations emphasize statistical analysis, modeling, prediction, and interpretation. Appendix Table A.6 provides complementary evidence from observed skill profiles: database workers are more associated with infrastructure and data-management skills, while data scientists are more associated with analytics and modeling skills. The cross-industry distribution in Figure 1 provides additional face validity, with higher measured depreciation in industries characterized by more rapidly changing demand, products, or

3. A full list of database-worker and data-scientist roles is provided in Appendix C.

operations.

We also construct three supplementary measures that capture different components of effective depreciation. *Depreciation (Product Fluidity)* uses the product fluidity index of Hoberg et al. (2014), constructed from year-to-year changes in firms' 10-K product descriptions. Higher product fluidity indicates a more rapidly changing product environment in which previously collected data become obsolete more quickly, and therefore proxies for higher information decay, δ_t.

Following Liu et al. (2025), *Depreciation (ROA Predictability)* captures how quickly historical firm information loses its predictive value. We estimate a firm-level AR(1) model of ROA and define the measure as an indicator equal to one when the model's R^2 is below 10 percent. Lower predictability implies that past performance contains less information about current outcomes and therefore indicates higher information decay, δ_t.

Finally, *Depreciation (Worker Skill Obsolescence)* is a shift-share measure that combines a firm's exposure to particular data-related skills with the subsequent decline of those skills in the broader labor market. Firms that rely more heavily on rapidly becoming obsolete data skills face greater difficulty maintaining and updating their data systems. We therefore interpret higher skill obsolescence as lower organizational maintenance effectiveness, η_t, and hence higher effective depreciation. Appendix F provides additional construction details.

The database-worker ratio may also reflect factors other than data depreciation, including firms' IT infrastructure, outsourcing decisions, local labor markets, data-stock scale, and unobserved expectations. Section 4.3 addresses these observable confounders, while Section 4.5 uses GDPR exposure and data-breach events as instruments for data depreciation and data investment, respectively, to assess whether the estimated depreciation discount is driven by unobserved confounders.

3.4 Other Key Measures

Data investment. Occupation data from Revelio are used to measure firm-year data investment. For each firm-year, we measure data investment as the total number of workers classified as either database workers or data scientists. The resulting count captures the scale of the firm's internal data-related workforce and serves as our proxy for firm-year data investment. While the database-worker share distinguishes maintenance-oriented from analysis-oriented labor for the depreciation measure, the total number of workers across both groups captures the firm's overall investment in data capabilities.

This labor-based measure provides 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. As a robustness check, we also show that the results are similar when data workers are weighted by seniority.

Other variables. We also construct measures of firms' IT and AI investment using Revelio data. IT labor is measured as the number of workers in information-technology-related occupations, including roles related to software, systems, networking, infrastructure, and data management. AI investment is measured by counting workers with AI skills, where AI skills are defined as described in Tambe (2025). Because data-related workers can overlap with broader IT and AI categories, we treat IT labor as capturing general digital capabilities and data labor as capturing the data-specific component of these capabilities. In the empirical analysis, we control for IT and AI labor separately and conduct robustness checks that exclude AI-skilled workers from the data-worker category.

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 et al. (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

Figure 1 illustrates the cross-industry distribution of data depreciation. The pattern is economically intuitive. Industries with volatile demand or rapidly changing products and operations, such as retail, telecom, construction, utilities, clothing retail, and food services, 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 or longer-lived information environments, including chemicals, education, manufacturing, hospitals, and information services, display lower depreciation rates, suggesting that their data remain useful over longer horizons.[4] [5] 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 lower in traditional manufacturing and resource-based industries.

4. Industry practice provides suggestive validation: Clothing Stores rely on rapidly updated sell-through and demand data, whereas Chemicals use historical process data for predictive maintenance. See McKinsey on retail supply-chain agility and McKinsey on analytics in chemical manufacturing.

5. For comparison, BEA's industry-specific annual depreciation rates for R&D capital range from 0.11 in Chemicals to 0.18 in Transportation Equipment, whereas our measure ranges from 0.17 to 0.63 across industries. However, our measure captures the share of database workers among database workers and data scientists, rather than the share of data value that becomes obsolete annually. See BEA (2007).

Table 1 reports summary statistics for the main regression sample. The sample contains firm-year observations for U.S. public firms from 2013 to 2023. The main variables include market value, assets, PPE, employment, IT labor, data labor, AI labor, and several proxies for data depreciation. All continuous variables used in the regression analysis are standardized to facilitate interpretation across measures.

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 et al. (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. As a robustness check, we further show in Table A.4 that this moderating effect is specific to data investment and does not appear for general IT investment.

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 Data Depreciation versus Firms' Strategic Choices

Our database worker ratio reflects firms' endogenous allocation of data labor between maintaining data assets and applying them in analytics and decision-making. This allocation is intended to capture effective data depreciation, which combines information decay and firms' data management capacity. A key measurement concern, however, is that the ratio may also reflect factors other than depreciation. We focus on four plausible confounders.

First, firms differ in IT infrastructure, especially cloud adoption, which can change the amount of in-house labor needed to store, update, and process data. Although cloud adoption may itself reflect data management capacity, it may also capture a broader IT strategy that substitutes cloud-managed services for internal database work. Second, firms may outsource data-related tasks, making internal data labor an incomplete measure of the true mainte-

nance burden. Third, local labor market conditions may shape workforce composition: when database workers are more available or less costly relative to data scientists, firms may employ more of them even if underlying data depreciation is not higher. Fourth, the database worker ratio may reflect the scale and stage of a firm's data stock rather than its depreciation rate. In the model, maintenance labor depends on both effective depreciation and the existing data stock. Firms with larger accumulated data assets, or firms in an early data-stack build-out stage, may require more database labor even holding depreciation constant.

Table 3 directly examines these concerns. Column (1) relates the database worker ratio to cloud adoption, data outsourcing, local relative wages, and lagged data investment. Cloud adoption is negatively associated with the database worker ratio, suggesting that firms with more advanced cloud infrastructure rely less on in-house database labor for data maintenance. Data outsourcing is positively associated with the database worker ratio, while local relative wages do not significantly predict the measure. Lagged data investment is positively associated with the database worker ratio, consistent with the concern that firms with a larger prior scale of data-related activity may require more database labor even holding depreciation fixed. This result underscores the importance of accounting for data-stock scale when interpreting the labor-based depreciation proxy.

Columns (2)-(6) then add these factors to the market value regressions. Across specifications, the interaction between data investment and the database worker ratio remains negative and statistically significant, with broadly similar magnitudes. Overall, the evidence indicates that the depreciation discount is not explained solely by cloud adoption, data outsourcing, local labor market conditions, or differences in lagged data investment.

We further examine whether the results are driven by persistent firm-level differences. Table A.1 adds firm fixed effects, so that identification comes from within-firm changes over time. The interaction becomes smaller and exhibits larger variation around the estimate, indicating that an important part of the baseline variation is cross-sectional. However, the coefficient remains negative for the full sample and for non-IT firms, suggesting that the de-

preciation discount is not entirely explained by time-invariant firm characteristics or broad strategic differences across firms. Taken together, these results support the interpretation of the database worker ratio as an informative labor-allocation proxy for effective data depreciation.

4.4 Effect Heterogeneity and Robustness Tests

In this section, we explore heterogeneity in the results and conduct further robustness checks.

Heterogeneity. The average effects in Table 2 may obscure important heterogeneity across firms. Tables A.2 and A.3 examine heterogeneity by cloud intensity and data outsourcing, while Figure 2 presents additional split-sample results based on firm characteristics.

Table A.2 examines whether the role of data depreciation varies with firms' cloud adoption. Using Revelio Labs, we identify IT skills and manually classify them as either cloud-related or on-premises.[6] 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. Columns (1) and (2) restrict the sample to high-cloud firms, while Columns (3) and (4) restrict the sample to low-cloud firms. The interaction between data and depreciation is negative and statistically significant among high-cloud firms, but smaller and statistically insignificant among low-cloud firms. Column (5) further confirms this pattern in the pooled sample: the triple interaction between data, depreciation, and the high-cloud indicator is negative, although with larger variation around the estimate. Overall, these results suggest that capital markets place greater weight on data depreciation when firms rely more heavily on cloud-based data infrastructure.

Table A.3 examines heterogeneity by data task outsourcing, using supply chain information from FactSet Revere. The market assigns less value to data investment for firms that outsource data tasks. However, the interaction between data and depreciation remains nega-

6. Classified skills are listed in the Appendix.

tive in both subsamples and is more precisely estimated among non-outsourcing firms. This pattern suggests that the baseline results are mainly driven by in-house data investment and depreciation, while outsourced data activities are less strongly valued as internal data investment.

Figure 2 further explores heterogeneity by firm age, size, 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, consistent with older firms relying more on accumulated data stocks whose value may be more sensitive to depreciation. The discount is also more negative for smaller firms than for larger firms. For firm growth, the coefficient is somewhat more negative among high-growth firms, 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. Although the differences across groups are not always statistically distinct, the direction of the heterogeneity is broadly consistent with our interpretation that the measure captures economically meaningful variation in the effective depreciation of firm data.

Robustness. Table A.4 reports several robustness checks. Columns (1) and (2) use one-year lagged data, IT, and AI investment to alleviate concerns about contemporaneous measurement and reverse timing. The interaction between data investment and depreciation remains negative. Columns (3) and (4) strengthen the fixed-effect structure by including industry-by-year fixed effects. The results remain negative under these more demanding specifications. Columns (5) and (6) use a weighted measure of data investment, where data workers are weighted by seniority, and the results remain similar. Columns (7) and (8) address the overlap between data labor and AI labor by reconstructing data investment after excluding workers with AI skills from the data-worker category. Across these specifications, the interaction between data investment and depreciation is consistently negative and, in most cases, statistically significant. Column (9) further interacts the data depreciation rate

with firms' IT investment. The coefficient is close to zero and statistically insignificant, suggesting that data depreciation does not significantly affect the value assigned to general IT investment. This placebo test reinforces that the moderating effect is specific to data-related assets rather than broader IT capital.

Table A.5 repeats the main analysis using firm sales as the dependent variable instead of market value. Prior work emphasizes that market value regressions can uncover the value of intangible assets that have not yet translated into realized productivity (Brynjolfsson et al. 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 sales regressions, whereas they are 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 before these gains are fully realized in current firm performance.

4.5 Estimation Using Instrument Variables

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 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 the estimated coefficients. In this section, we use instrumental variable analyses as suggestive evidence based on external shocks that affect either the effective depreciation of data or firms' incentives to invest in data.

4.5.1 GDPR and Data Depreciation

We first use the implementation of the General Data Protection Regulation (GDPR) as an instrument for data depreciation. GDPR was approved by the European Union in 2016 and became enforceable in 2018, giving firms time to adjust their data practices before formal enforcement. The regulation imposed stricter requirements on data collection, storage, retention, and reuse, including more stringent consent and compliance standards. In our framework, these rules can affect effective data depreciation through both information decay and data management capacity. Limits on retention and reuse may reduce the flexibility of existing data, while stronger governance requirements may also force firms to delete redundant or outdated data, standardize consent management, and improve the usability of the remaining data stock. Thus, GDPR may reduce measured effective depreciation if improvements in data management capacity dominate the loss of flexibility in data reuse.

Figure 3 compares the data depreciation rate for firms with and without GDPR exposure over time, where exposure is defined based on whether firms had sales in Europe before 2018. The two vertical lines mark 2016, when GDPR was approved, and 2018, when it became enforceable. Before 2016, the two groups follow similar downward trends. After 2016, and especially after enforcement in 2018, exposed firms exhibit a lower database worker ratio than non-exposed firms. This pattern is consistent with GDPR inducing exposed firms to rationalize data retention, reduce redundant data assets, and improve the manageability and legal usability of their remaining data stock. Although GDPR may affect firm value through channels such as compliance costs, advertising and targeting, consumer consent, and broader regulatory risk, its direct implications for data retention, reuse, and governance make it a useful regulatory shock for examining whether changes in firms' data environment are associated with the depreciation discount documented in our baseline results.

Table 4 reports how the moderating role of data depreciation varies with GDPR exposure and presents the IV specification. Columns (1)–(4) show that the interaction term $\log(Data) \times Depreciation$ is more negative for GDPR-exposed firms than for non-exposed firms, both

before and after 2018. This pattern suggests that depreciation is more strongly reflected in the valuation of data investment among firms more exposed to data regulation. Column (5) uses the interaction between European sales exposure and a post-2018 indicator as an instrument for data depreciation. The first stage is negative and statistically significant, consistent with Figure 3: after GDPR enforcement, exposed firms exhibit lower measured effective depreciation. In our framework, this may reflect improvements in data management capacity, as GDPR encouraged firms to standardize data governance and reduce redundant or difficult-to-maintain data assets. In the second stage, the interaction between data investment and depreciation becomes substantially more negative. We interpret this result as suggestive evidence that, when GDPR-related variation shifts firms' data environment, the market continues to discount data investment more strongly when measured data depreciation is higher.

4.5.2 Data Breach and Data Investment

We also use firm-level data breach events as an instrument for data investment. Data breaches, such as cyber attacks or unauthorized data disclosures, represent shocks that compromise the integrity and usability of existing data assets. Following a breach, firms may need to rebuild, upgrade, or reorganize their data infrastructure, leading to increased investment in data-related capabilities and systems. Although breaches may directly affect firm value through reputation damage, legal liability, customer churn, cybersecurity concerns, and investor beliefs about managerial quality, they also generate sharp shocks to firms' data systems. We therefore interpret the breach-based IV as capturing a defensive or remediation-driven margin of data investment.

Figure 4 provides descriptive support for this interpretation. Panel (a) shows a positive relationship between the number of breach events and data investment, consistent with breaches triggering subsequent rebuilding or upgrading of data infrastructure. Panel (b) shows little systematic relationship between breach events and data depreciation, suggesting

that breaches primarily shift firms' incentives to invest in data rather than directly altering the depreciation measure. This pattern motivates using breach events to examine whether the depreciation discount remains present when firms are pushed to adjust their data capabilities after shocks to their data systems.

Table 5 reports the corresponding regression results. Columns (1) and (2) split firms by whether they have ever experienced a breach. Data investment is positively associated with market value among breach firms, but the relationship is weaker among non-breach firms. Column (3) implements the IV specification, using the interaction between the number of breach events and an indicator for years after the first breach as an instrument for data investment. The first stage is positive and significant, indicating that breach shocks predict subsequent increases in data investment. In the second stage, the coefficient on $\log(Data)$ becomes negative, while the interaction between data investment and depreciation becomes substantially more negative. This pattern suggests that breach-induced data investment captures a defensive or remediation-driven margin of data accumulation rather than ordinary data investment. Importantly, even along this margin, the market discounts data investment more strongly when depreciation is higher. We therefore interpret the breach-based IV results as suggestive evidence that the depreciation discount remains present when firms are pushed to adjust their data capabilities after shocks to their data systems.

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.

The results have direct managerial implications. For managers, the main lesson is that data strategy should not be evaluated only by how much data a firm collects or how many data workers it hires, but also by how long those data assets remain usable. Firms with high data depreciation face a lower market return to data investment because a larger share of resources must be devoted to cleaning, updating, integrating, and rebuilding data rather than extracting value from it. This suggests that managers should treat data maintenance, governance, and curation as core investment decisions rather than back-office costs. In particular, investments in data infrastructure that reduce the need for repeated maintenance may increase the economic life of data assets and raise the value of subsequent investments in analytics and AI.

Although our main estimates are based on market valuation rather than aggregate production, they also speak to the potential economic relevance of data depreciation. A simple back-of-the-envelope exercise illustrates the economic magnitude of this estimate. A one-standard-deviation increase in our data depreciation proxy is associated with a 36.5 percent lower market value attributed to data investment. For a firm with the median market value in our sample, this corresponds to approximately $131 million in market value. Scaling this figure by the 3,477 firms in our sample implies a sample-wide magnitude of roughly $455 billion. This exercise highlights the economic relevance of data depreciation in market valuation.

Data depreciation may also matter for debates on competition and artificial intelligence. Discussions of data-driven market power often emphasize increasing returns to scale and network effects, while paying less attention to the fact that data may lose value over time. If data depreciates quickly, the advantages generated by data accumulation may be more short-lived than commonly assumed. Similarly, since data is a key input into AI models, rapid depreciation raises an important question of how firms can sustain a consistent supply of high-quality, up-to-date data for analytics and AI applications.

There are several limitations to this analysis that should be noted. The data offer 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. We hope this paper serves as a foundation for further research on data depreciation and its broader strategic and economic implications.

References

Aridor, Guy, Yeon-Koo Che, and Tobias Salz. 2023. "The Effect of Privacy Regulation on the Data Industry." *RAND Journal of Economics* 54 (4): 821–852.

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

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan. 2022. "Measuring the Value of Data and Data-Driven Prediction." *American Economic Review* 112 (11): 3448–3483.

Bhutani, Ananya, Guillermo Ordonez, and Laura Veldkamp. 2025. "The Missing Value of Data." *NBER Working Paper,* no. 33763.

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, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" *SSRN Electronic Journal.*

Brynjolfsson, Erik, Andrew McAfee, and Lorin Hitt. 2012. "Big Data: The Management Revolution." *Harvard Business Review* 90 (10): 60–68.

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

Corrado, Carol, Jonathan Haskel, Cecilia Jona-Lasinio, and Massimiliano Iommi. 2022. *Measuring Data as an Asset: Framework, Methods and Challenges.* OECD Digital Economy Papers 327. OECD.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55 (3): 661–685.

Demirer, Mert, Sonia Jaffe, Ravi Kirpalani, et al. 2024. *Data, Privacy Laws and Firm Production.* Working Paper. Federal Reserve Bank of Chicago.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2013. "Organization Capital and the Cross-Section of Expected Returns." *Journal of Finance* 68 (4): 1365–1406.

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

———. 2023. "Data and markets." *Annual Review of Economics* 15 (1): 23–40.

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

Goldfarb, Avi, and Catherine E Tucker. 2011. "Privacy regulation and online advertising." *Management science* 57 (1): 57–71.

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

Hagiu, Andrei, and Julian Wright. 2023. "Data-enabled learning, network effects, and competitive advantage." *The RAND Journal of Economics* 54 (4): 638–667.

Hall, Bronwyn H., Jacques Mairesse, and Pierre Mohnen. 2010. "Measuring the Returns to R&D." *Handbook of the Economics of Innovation* 2:1033–1082.

Haskel, Jonathan, and Stian Westlake. 2017. *Capitalism without Capital: The Rise of the Intangible Economy.* Princeton University Press.

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

Johnson, Garrett A., Scott K. Shriver, and Samuel G. Goldberg. 2022. "The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR." *Management Science* 68 (12): 8762–8787.

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

Lei, Xiaoxia, Yixing Chen, and Ananya Sen. 2026. "Trade-offs in leveraging external data capabilities: Evidence from a field experiment in an online search market." *Management Science* 72 (4): 2998–3015.

Lev, Baruch. 2001. *Intangibles: Management, Measurement, and Reporting.* Brookings Institution Press.

Li, Wendy CY, and Bronwyn H Hall. 2020. "Depreciation of business R&D capital." *Review of Income and Wealth* 66 (1): 161–180.

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, and Erik Brynjolfsson. 2016. "Data in Action: Data-Driven Decision Making in U.S. Manufacturing." *Rotman School of Management Working Paper.*

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.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. 2022. "Privacy and Digital Entrepreneurship: The Effect of GDPR on Venture Capital Investment." *Management Science* 68 (11): 8142–8163.

Sun, Tianshu, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu. 2024. "The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy." *Management Science* 70 (4): 2645–2660.

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.*

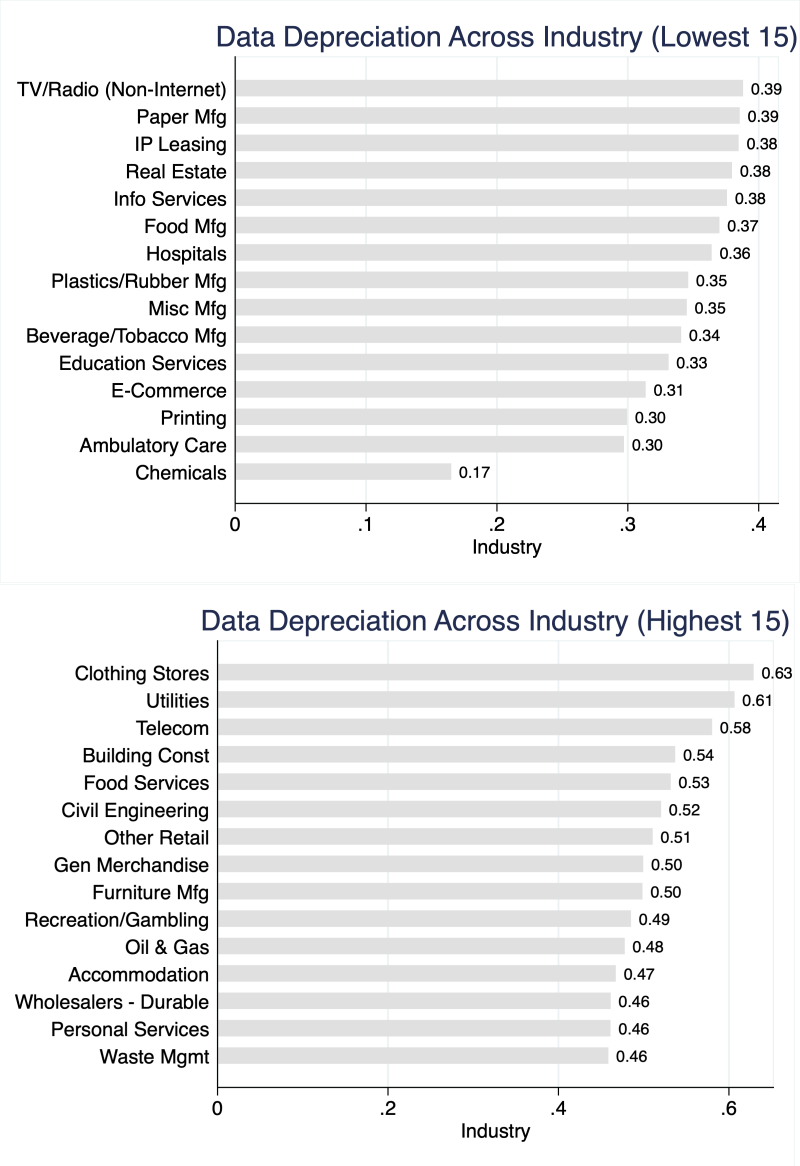
Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson. 2020. *Digital capital and superstar firms.* Technical report. National Bureau of Economic Research.

Valavi, Ehsan, Joel Hestness, and Marco Iansiti. 2022. "Time and the Value of Data." *arXiv preprint arXiv:2203.09118.*

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

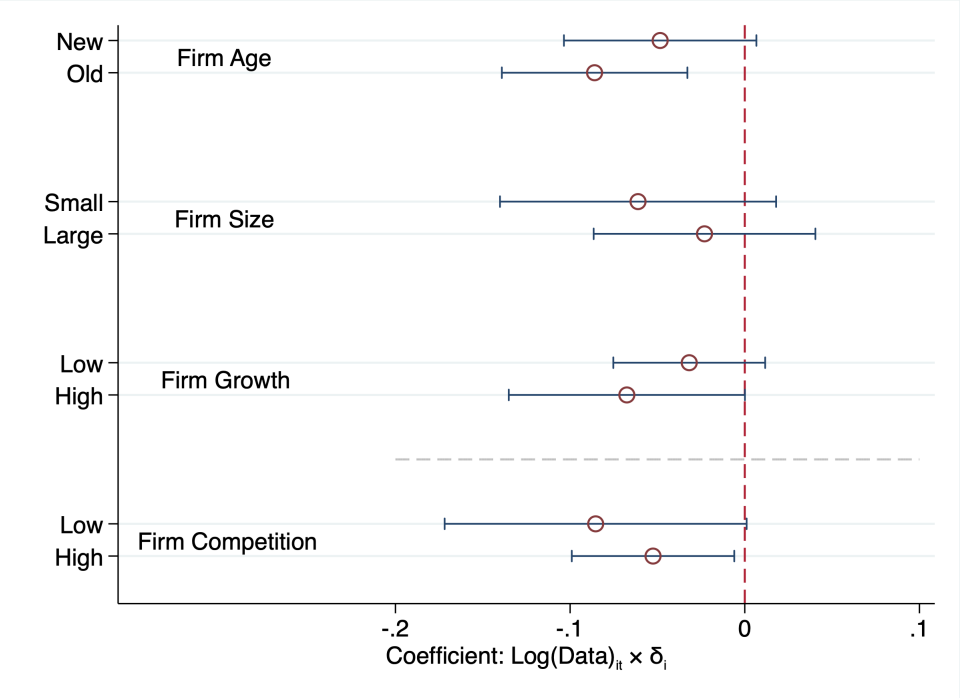
Figures

Figure 1: Industry Distribution of Data Depreciation (Higher Values = Faster Depreciation)



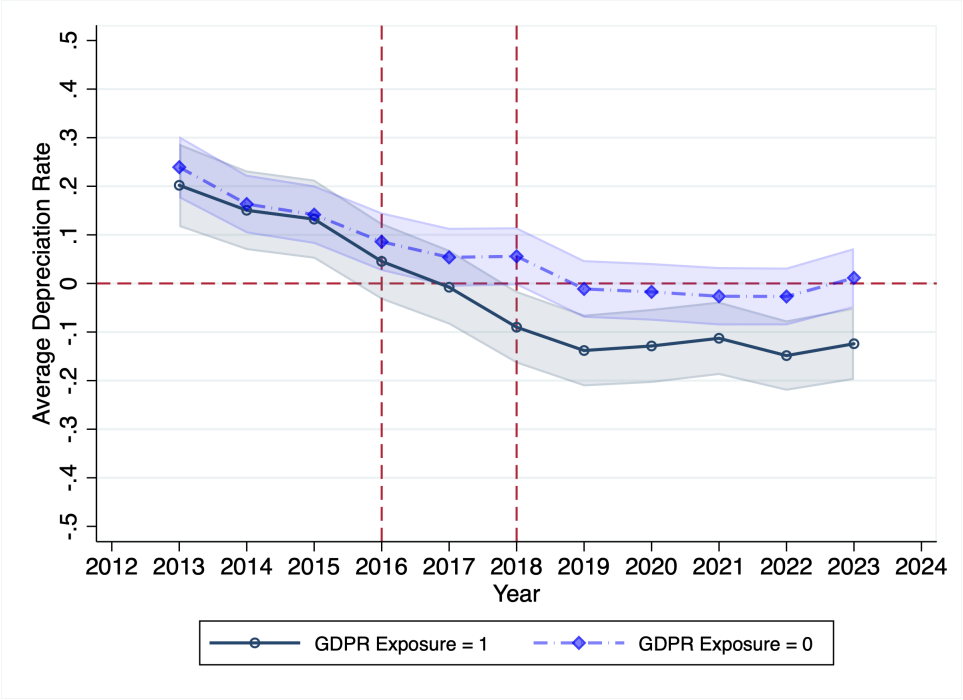
Notes. This figure reports the distribution of average firm-level data depreciation rates across three-digit NAICS industries in our sample. Industries with fewer than 50 firms are excluded. Data depreciation is measured as the share of database workers among database workers and data scientists. The measure shown in this figure is not standardized.

Figure 2: Heterogeneity of Data Depreciation Discount



Notes. This figure reports the coefficients of $Log(Data)_{it} \times \delta_i$ from equation 9 for the split-sample analysis. The figure 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.

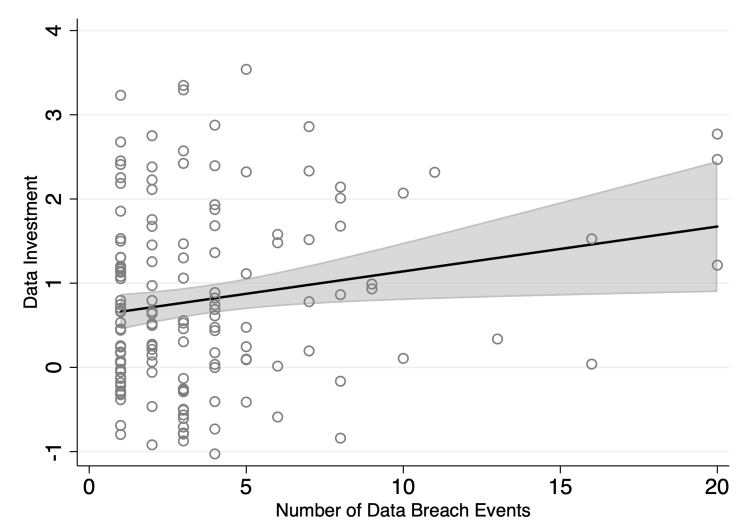
Figure 3: GDPR and Data Depreciation



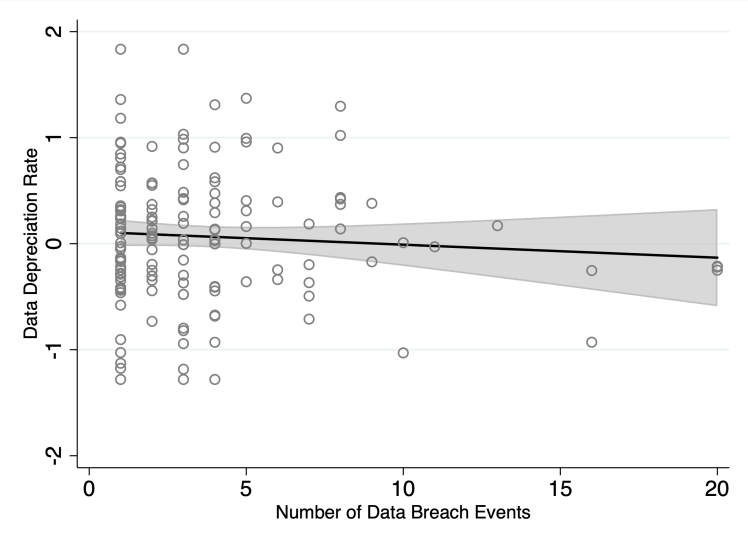
Notes. This figure shows the average (standardized) data depreciation rate of firms with and without exposure to GDPR over time. *Depreciation (Database Worker Ratio)* measures data depreciation as the share of database workers among database workers and data scientists. *GDPR Exposure* is defined as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

Figure 4: Data Investment, Data Depreciation, and Data Breach

(a) Data Investment and Data Breach



(b) Data Depreciation and Data Breach



Notes. This figure reports the relationship 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 five 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 provided by Hoberg et al. 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)
Depreciation (Database Worker Ratio)				-0.116***			
				(0.037)			
Log(Data) × Depreciation (Database Worker Ratio)				-0.061**			
				(0.027)			
Depreciation (Product Fluidity)					-0.006		
					(0.022)		
Log(Data) × Depreciation (Product Fluidity)					-0.037*		
					(0.021)		
Depreciation (Worker Skill Obsolescence)						-0.041***	
						(0.014)	
Log(Data) × Depreciation (Worker Skill Obsolescence)						-0.028**	
						(0.014)	
Depreciation (ROA Predictability)							-0.023
							(0.039)
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 et al. 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: Data Depreciation or Firms' Strategic Choices

	(1)	(2)	(3)	(4)	(5)	(6)
	Depreciation (Database Worker Ratio)	Log(Market Value)				
Log(Assets)	-0.010	0.400***	0.410***	0.403***	0.404***	0.389***
	(0.016)	(0.062)	(0.065)	(0.063)	(0.064)	(0.060)
Log(PPE)	0.087**	1.206***	1.122***	1.129***	1.119***	1.200***
	(0.039)	(0.096)	(0.107)	(0.103)	(0.108)	(0.097)
Log(Employment)	-0.042	0.353***	0.331***	0.357***	0.341***	0.380***
	(0.035)	(0.065)	(0.065)	(0.066)	(0.067)	(0.066)
Log(IT)	0.356***	0.120***	0.017	0.074*	0.040	0.177***
	(0.089)	(0.040)	(0.054)	(0.044)	(0.054)	(0.039)
Log(Data)	-0.411***	0.073	0.166	0.150	0.707***	0.609***
	(0.058)	(0.080)	(0.107)	(0.094)	(0.138)	(0.105)
Log(AI)	-0.112**	0.213***	0.286***	0.223***	0.294***	0.177***
	(0.056)	(0.061)	(0.062)	(0.065)	(0.058)	(0.061)
IT cloud index	-0.206***	0.289***				0.260***
	(0.040)	(0.029)				(0.031)
Outsourcing Data	0.064**		0.033			0.003
	(0.026)		(0.050)			(0.047)
Local Relative Wage	0.375			-3.339***		-2.535***
	(0.310)			(0.404)		(0.351)
Log(Data) - 1-Year Lag	0.184***				-0.580***	-0.579***
	(0.050)				(0.070)	(0.070)
Depreciation (Database Worker Ratio)		-0.084**	-0.116***	-0.108***	-0.109***	-0.074**
		(0.034)	(0.037)	(0.035)	(0.036)	(0.032)
Log(Data) × Depreciation (Database Worker Ratio)		-0.049*	-0.062**	-0.057**	-0.058**	-0.044*
		(0.028)	(0.027)	(0.026)	(0.025)	(0.025)
Constant	-0.494	8.402***	8.372***	12.084***	8.351***	11.193***
	(0.344)	(0.003)	(0.006)	(0.447)	(0.005)	(0.386)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.220	0.805	0.799	0.803	0.802	0.810
Observations	20926	20926	20926	20926	20926	20926

Notes. This table reports how firms' strategies and data-stock scale affect the data depreciation proxy, the database worker ratio, and its effect on market value. The strategic-choice variables include cloud IT, data task outsourcing, and local database-worker relative wages. *IT Cloud Index* is the standardized share of IT workers with cloud-related skills. *Outsourcing Data* is a dummy variable equal to one if the firm purchases or licenses data-related products or services. *Local Relative Wage* is the ratio of the average wage of database workers to the average wage of all data analysts in a region. For firms operating in more than one region, we construct the measure as the employment-weighted average across regions. *Log(Data) - 1-Year Lag* is the one-year lag of the log number of data workers and is included to proxy for the scale and maturity of firms' accumulated data-related assets. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table 4: 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)
Depreciation (Database Worker Ratio)	-0.165***	-0.133***	-0.063	-0.084	-0.202*
	(0.052)	(0.026)	(0.064)	(0.052)	(0.110)
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 5: Data Breach, Data Investment, and Data Depreciation

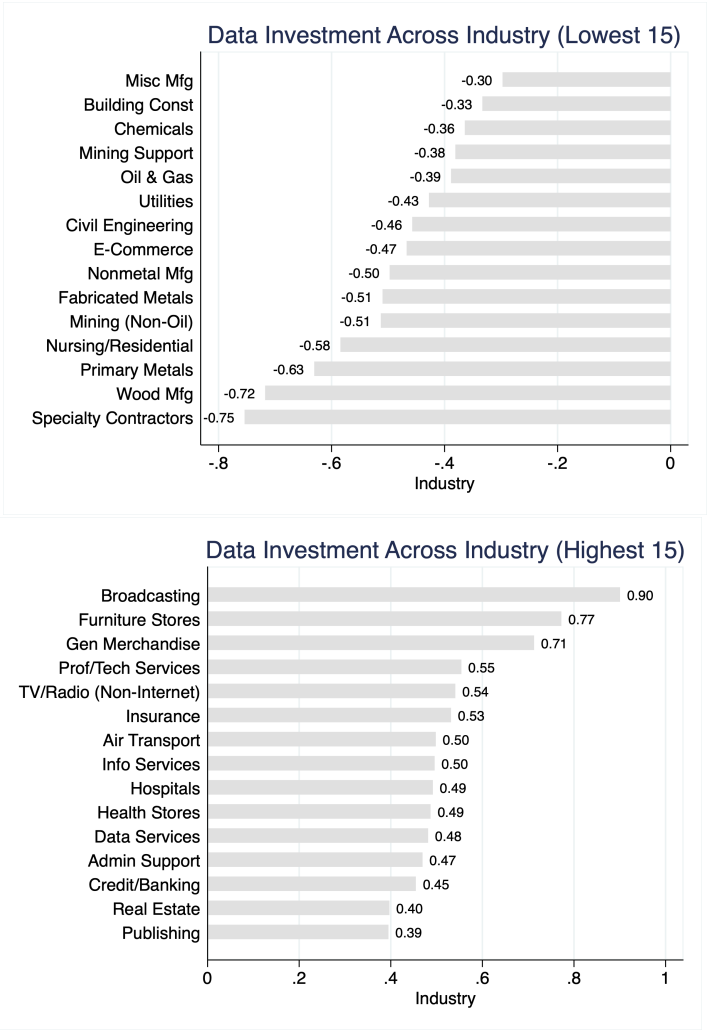
	(1)	(2)	(3)
	Log(Market Value)		
	Data Breach = 1	Data Breach = 0	
Model	OLS	OLS	IV
Log(Assets)	0.308***	0.435***	0.414***
	(0.099)	(0.070)	(0.011)
Log(PPE)	0.919***	1.122***	1.028***
	(0.191)	(0.109)	(0.032)
Log(Employment)	0.425**	0.342***	0.505***
	(0.192)	(0.068)	(0.059)
Log(IT)	-0.318	0.002	0.775**
	(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)
Depreciation (Database Worker Ratio)	-0.054	-0.100**	-0.364***
	(0.068)	(0.042)	(0.077)
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 3-digit NAICS industries in our sample. Industries 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: Firm Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Market Value)					
	All Firms		Exclude IT Firms		IT Firms	
Log(Assets)	0.410***	0.174***	0.373***	0.165***	2.052***	1.085***
	(0.065)	(0.034)	(0.060)	(0.033)	(0.183)	(0.106)
Log(PPE)	1.122***	0.332***	1.122***	0.391***	0.741***	-0.076*
	(0.107)	(0.055)	(0.113)	(0.054)	(0.052)	(0.031)
Log(Employment)	0.332***	0.952***	0.336***	0.923***	0.033	0.847***
	(0.066)	(0.082)	(0.075)	(0.091)	(0.069)	(0.072)
Log(IT)	0.017	-0.102	-0.017	-0.118	0.126	-0.095
	(0.054)	(0.081)	(0.068)	(0.086)	(0.136)	(0.501)
Log(AI)	0.288***	0.035	0.252***	0.021	0.321*	0.114*
	(0.061)	(0.025)	(0.067)	(0.022)	(0.136)	(0.038)
Log(Data)	0.167	0.189***	0.226**	0.171***	-0.116	0.219
	(0.108)	(0.062)	(0.112)	(0.063)	(0.064)	(0.244)
Depreciation (Database Worker Ratio)	-0.116***	0.029	-0.136***	0.050	-0.053	-0.103
	(0.037)	(0.042)	(0.041)	(0.050)	(0.043)	(0.158)
Log(Data) × Depreciation (Database Worker Ratio)	-0.061**	-0.011	-0.071**	-0.015^{+}	-0.066	0.017
	(0.027)	(0.008)	(0.034)	(0.009)	(0.040)	(0.026)
Constant	8.374***	7.275***	8.358***	7.373***	8.018***	6.666**
	(0.004)	(0.371)	(0.007)	(0.375)	(0.039)	(1.366)
Industry FE	Yes	No	Yes	No	Yes	No
Firm FE	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.799	0.956	0.802	0.954	0.861	0.968
Observations	20926	20616	18267	18002	2659	2614

Notes. This table reports the robustness results using firm fixed effects. Columns (1) and (2) include all firms. Columns (3) and (4) include non-IT firms, while columns (5) and (6) only include IT firms. IT firms are those with NAICS codes beginning with "51" or "54." $+p < 0.15$; $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table A.2: Firm Cloud Index, Data Value, and Data Depreciation

	(1)	(2)	(3)	(4)	(5)
	Log(Market Value)				
	High Cloud Index		Low Cloud Index		Interaction
Log(Assets)	0.976***	0.970***	0.289***	0.289***	0.403***
	(0.220)	(0.218)	(0.035)	(0.035)	(0.063)
Log(PPE)	0.831***	0.822***	1.483***	1.487***	1.146***
	(0.106)	(0.113)	(0.068)	(0.069)	(0.104)
Log(Employment)	0.386***	0.407***	0.258***	0.255***	0.383***
	(0.057)	(0.050)	(0.076)	(0.075)	(0.069)
Log(IT)	0.050	0.109**	0.067	0.082	0.069
	(0.049)	(0.049)	(0.063)	(0.063)	(0.047)
Log(Data)	0.183	0.123	0.019	0.012	0.090
	(0.122)	(0.100)	(0.069)	(0.068)	(0.088)
Log(Data) × High Cloud Index					0.073*
					(0.041)
Log(AI)	0.226***	0.214***	0.235***	0.231***	0.222***
	(0.083)	(0.070)	(0.059)	(0.060)	(0.055)
Depreciation (Database Worker Ratio)		-0.127**		-0.045	-0.020
		(0.051)		(0.030)	(0.032)
Log(Data) × Depreciation (Database Worker Ratio)		-0.071*		-0.020	-0.006
		(0.042)		(0.028)	(0.028)
∼ × High Cloud Index					-0.052
					(0.054)
Constant	8.442***	8.423***	8.041***	8.045***	8.205***
	(0.070)	(0.071)	(0.026)	(0.027)	(0.023)
Industry FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.812	0.814	0.825	0.826	0.803
Observations	10392	10392	10409	10409	20926

Notes. This table reports how the moderating role of data depreciation varies across firms' cloud index. 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 with high cloud index, defined as firms with share of IT workers with cloud-related skills above median. *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: Outsourcing Data Tasks, Data Value, and Data Depreciation

	(1)	(2)	(3)	(4)	(5)
	Log(Market Value)				
	Outsource Data Tasks = 1		Outsource Data Tasks = 0		Interaction
Log(Assets)	0.343***	0.343***	0.421***	0.417***	0.410***
	(0.103)	(0.103)	(0.066)	(0.065)	(0.065)
Log(PPE)	1.286***	1.273***	1.104***	1.106***	1.119***
	(0.123)	(0.139)	(0.102)	(0.106)	(0.107)
Log(Employment)	0.301**	0.313**	0.325***	0.335***	0.334***
	(0.116)	(0.130)	(0.058)	(0.064)	(0.066)
Log(IT)	-0.178	-0.079	-0.027	0.024	0.013
	(0.131)	(0.139)	(0.060)	(0.053)	(0.054)
Log(Data)	-0.041	-0.091	0.220*	0.181*	0.175
	(0.132)	(0.154)	(0.124)	(0.105)	(0.106)
Log(Data) × Outsource Data Tasks					-0.096***
					(0.033)
Log(AI)	0.529***	0.499***	0.297***	0.278***	0.291***
	(0.105)	(0.093)	(0.065)	(0.062)	(0.063)
Depreciation (Database Worker Ratio)		-0.036		-0.116***	-0.114***
		(0.051)		(0.041)	(0.041)
Log(Data) × Depreciation (Database Worker Ratio)		-0.087		-0.060**	-0.058*
		(0.058)		(0.030)	(0.030)
∼ × Outsource Data Tasks					-0.044
					(0.051)
Constant	8.469***	8.456***	8.383***	8.376***	8.373***
	(0.051)	(0.049)	(0.004)	(0.005)	(0.006)
Industry FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.867	0.868	0.782	0.783	0.800
Observations	1338	1338	19581	19581	20926

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.4: Robustness Check

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Log(Market Value)								
	1-Year Lag		Industry × Year FE		Weighted by Seniority		Data Workers Excl. AI Skills		Placebo
Log(Assets)	0.394***	0.391***	0.419***	0.416***	0.410***	0.406***	0.415***	0.411***	0.412***
	(0.064)	(0.063)	(0.068)	(0.067)	(0.065)	(0.064)	(0.067)	(0.065)	(0.066)
Log(PPE)	1.112***	1.114***	1.123***	1.124***	1.112***	1.114***	1.122***	1.122***	1.128***
	(0.105)	(0.110)	(0.101)	(0.106)	(0.105)	(0.109)	(0.101)	(0.107)	(0.102)
Log(Employment)	0.306***	0.316***	0.314***	0.325***	0.325***	0.335***	0.318***	0.330***	0.322***
	(0.067)	(0.075)	(0.059)	(0.066)	(0.060)	(0.068)	(0.057)	(0.065)	(0.061)
Log(IT)	-0.037	0.017	-0.043	0.009	-0.004	0.024	-0.032	0.022	-0.002
	(0.071)	(0.065)	(0.063)	(0.058)	(0.025)	(0.023)	(0.062)	(0.055)	(0.057)
Log(Data)	0.203	0.165	0.203	0.165	0.142**	0.120**	0.113	0.090	0.183
	(0.123)	(0.108)	(0.125)	(0.109)	(0.061)	(0.051)	(0.074)	(0.064)	(0.118)
Log(AI)	0.326***	0.307***	0.315***	0.297***	0.111***	0.101***	0.326***	0.304***	0.293***
	(0.059)	(0.057)	(0.062)	(0.059)	(0.027)	(0.026)	(0.056)	(0.054)	(0.062)
Depreciation (Database Worker Ratio)		-0.036		-0.109***		0.020		-0.005	-0.082***
		(0.051)		(0.036)		(0.034)		(0.026)	(0.024)
Log(Data) × Depreciation (Database Worker Ratio)		-0.049*		-0.055**		-0.036**		-0.039**	
		(0.026)		(0.026)		(0.014)		(0.018)	
Log(IT) × Depreciation (Database Worker Ratio)									-0.016
									(0.014)
Constant	8.410***	8.408***	8.378***	8.373***	7.552***	7.479***	8.048***	8.111***	8.378***
	(0.001)	(0.001)	(0.003)	(0.004)	(0.124)	(0.115)	(0.220)	(0.189)	(0.004)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	No	No	Yes	Yes	No	No	No	No	No
Adjusted R-squared	0.771	0.772	0.805	0.806	0.798	0.800	0.798	0.799	0.799
Observations	18592	18592	20926	20926	20926	20926	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 the data investment, IT investment, and AI investment in year t-1. Columns (3) and (4) add industry × year fixed effects. Columns (5) and (6) use data, IT, and AI investment weighted by seniority. Columns (7) and (8) exclude workers with AI skills from the data worker category. Column (9) interacts the data depreciation rate with firms' IT investment. $*p < 0.10$; $**p < 0.05$; $***p < 0.01$.

Table A.5: 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)
Depreciation (Database Worker Ratio)				-0.091***
				(0.030)
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 firm sales 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. Colum (4) examines how data depreciation moderates the productivity 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 Roles and Skills

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 learning 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.*

A full list of cloud IT skills: *amazon ec2, amazon s3, amazon web services, amazon web services (aws), aws, azure, cloud, cloud applications, cloud computing, cloud computing iaas, cloud security, cloud storage, cloud-based, cloud-computing, gcp, google cloud platform, microsoft azure, windows azure.*

A full list of on-prem IT skills: *emc storage, aix, ibm aix, ibm iseries, backup exec, blackberry enterprise server, cisco routers, cisco certified network professional, cisco ios, cisco certified network associate, cisco nexus, routing protocols, cisco, netapp, data center, it infrastructure management, dwdm, dynamic host configuration protocol (dhcp), netbackup, job control language (jcl), bmc remedy, backup solutions, hp-ux, ibm mainframe, itil v3 foundations certified, mpls vpn, juniper, juniper technologies, lan-wan, mainframe, mpls, vmware infrastructure, storage area networks, storage virtualization, vmware vsphere, virtualization, vmware, vmware esx, vsphere.*

D Skill Differences Across IT Workers, Data Scientists, and Database Workers

Table A.6 compares the most common skills across general IT workers, data scientists, and database workers. General IT workers are more concentrated in software-development skills, while data scientists report more analysis-oriented skills. In contrast, database workers are more concentrated in database, warehousing, and data-infrastructure skills.

Table A.6: Skill Differences Across IT Workers, Data Scientists, and Database Workers

Rank	General IT Workers	Data Scientists	Database Workers
1	structured query language	data analysis	databases
2	java	powerpoint	business intelligence
3	software development	microsoft word	pl/sql
4	javascript	research	data warehousing
5	agile methodologies	business intelligence	python
6	linux	analysis	etl
7	html	python	it infrastructure library
8	c++	public speaking	solution architecture
9	project management	social media	shell scripting
10	microsoft sql server	teamwork	disaster recovery
11	requirements analysis	training	data modeling
12	xml	analytics	it service management
13	c	matlab	requirements gathering
14	integration	strategy	microsoft word
15	management	sales	enterprise architecture
16	microsoft office	marketing	powerpoint
17	mysql	strategic planning	performance tuning
18	software development life cycle (sdlc)	process improvement	database design
19	c#	financial analysis	data migration
20	business analysis	r	database administration

E 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.

E.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 workers 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:

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

E.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.

E.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

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

E.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 $\tilde{\delta}_t$.

F 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).

F.1 Product Fluidity

The measure draws on the product fluidity index developed by Hoberg et al. (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.

F.2 Predictive Persistence of Performance

The second measure follows Liu et al. (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.

F.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 $\tilde{\delta}_t$.