Democratizing Data: Measuring and Decomposing the Growth of Firms' Data Assets

Guillermo Ordonez and Laura Veldkamp*

March 19, 2025

Abstract

Data assets are increasingly vital in modern economies, yet quantifying their value remains elusive. This difficulty not only distorts our measurements of investment and GDP but also obscures the evolution and cross-section of firms' asset values. Using a novel framework, we estimating the aggregate value of firms' data assets and their distribution across firms, exploiting the evolution of firms' forecasting accuracy. Recognizing that firms rely on proprietary and external data sources, we adjust our estimates to account for the value of both internally generated and purchased data. Unlike prior approaches, we treat data both as an asset and an input in production. By capitalizing the data flow value and incorporating depreciation, we provide an asset valuation of data. Our findings suggest that the aggregate value of data has increased substantially, but less than suggested by the explosion of data storage, and that most of this increase came from mid-sized firms' data values catching up with those of the largest firms.

In the digital economy, data has become one of firms' most valuable and essential assets. The speed of growth of digital technologies leads to the conjecture that the value of data has grown immensely. It is also thought to primarily benefit large firms, exacerbating problems with market competition. Despite its growing economic importance, measuring the aggregate value of firms' data and its distribution remains elusive. Data is more highly valued than ever because it is the fuel for artificial intelligence, which is, at its core, a prediction technology (Agrawal et al., 2018). Firms use data to predict demand, costs, inventories, advertising effectiveness, and competitors' strategies (Goldfarb and Tucker, 2019). Our objective is to quantify the value of data assets both in the aggregate and the cross-section of firms.

*Preliminary and incomplete. Please do not cite or circulate. ordonez@econ.upenn.edu, lv2405@columbia.edu. We thank Ankit Bhutani for his excellent research assistance.

The argument for why large firms benefit more from data rests partly on large firms' ability to generate more data internally from their large volume of customer transactions Farboodi and Veldkamp (2021). However, new data platforms are arising to help smaller firms access larger pools of data.[1] Are these data-sharing mechanisms leveling the playing field? Is the value of data growing at the speed of data itself? To decompose the source of changes in firms' data values, we decompose data value into components derived from firms' internally generated data from their data acquired from external sources. This distinction is important because it speaks to the changing competitive landscape and because the value of purchased data is already accounted for on firms' balance sheets and national income accounts.

Since data is the fuel for prediction technologies, we put firms' predictions at the center of our analysis: Firms use data, in part, to forecast their future revenue. Therefore, we quantify the value of data based on improvements in firms' forecasting accuracy. We then leverage estimates of the relationship between firms' forecast errors and their profits to infer the value of data (Asriyan and Kohlhas, 2024). This approach does not imply that data is only used for revenue forecasting. Our approach recognizes that data is an input into many firm predictions. Firms with more or better data can predict many things well – including revenue. The observed correlation between revenue forecast errors and profits is not causal. It captures the benefits of many accurate forecasts a firm with good data can make. The correlation between revenue forecast errors and firm profits captures the broader productivity benefits of data.

To separate purchased data from internally generated data, we use income disclosures from data platforms to determine the value of platform data to firms and, hence, how much firms are willing to pay to improve their forecasts. This information allows us to assess the value of data derived from platforms relative to the value of internal data, which we find is

[1]Examples abound, and include Square, which helps small firms manage and share data, Visa, which sells sector and regionally tailored reports derived from their transaction data, or Amazon, which provides insight reports for the retailers on their platform.

still an order of magnitude more valuable for the average firm.

Unlike previous approaches that emphasize the consumer value of free digital goods, our method treats data as a productive asset, similar to capital or labor. We measure not only the immediate flow value of data but also capitalize it to derive its long-term asset value. This approach acknowledges that data collected today continues to provide value over time, albeit with depreciation. Properly accounting for depreciation is critical, as data can lose relevance due to changes in market conditions, consumer behavior, or firm-specific dynamics. We derive a theory-based depreciation rate for data, measure that rate and use it to capitalize the value of data assets.

When we capitalize data value and study its evolution, we learn that two other forces, besides the sale of platform data, are leveling the data playing field: diminishing returns to data and depreciation. The first effect comes from recognizing that data fuels predictions. It reduces prediction errors. However, expected squared prediction errors are bounded below by zero. This naturally gives data diminishing returns. The fact that the tenth data point reduces forecast errors much more than the millionth is a force enabling medium-size firms to catch up with larger ones. The second comes from the particular depreciation rate of data. We find that firms with large data sets suffer more from data deprecation. Because their forecasts are so precise, changes in the state increase their forecast errors proportionately more. We find that these two forces contribute to the catch-up we observe in mid-sized firms' prediction errors relative to those of large firms.

Estimating the aggregate value of firms' data is crucial for several reasons. First, data has become a key input in production and innovation. Yet, it is often misclassified or excluded from traditional capital investment measures, leading to an incomplete understanding of economic growth and productivity. Proper valuation would enable more accurate national accounting and GDP measurement. Second, understanding the value of data assets can inform firm-level and policy decisions regarding data governance, investment taxation, and

regulation. Recognizing the economic significance of data helps policymakers design frameworks that promote competition, innovation, and responsible data-sharing practices. Third, an aggregate valuation of data assets can improve financial market assessments and corporate valuation models, ensuring that investors account for firms' intangible assets. This is particularly relevant as data-driven business models become increasingly dominant, challenging conventional valuation techniques that were primarily designed for tangible capital.

Our findings suggest that conventional measures of firm investment systematically underestimate the true value of internally generated data. Since standard accounting practices often fail to recognize this data as an investment, the economic impact of data-driven decision-making remains obscured in firm valuations and national income accounts. By providing a market-based approach to valuing data, our study contributes to ongoing debates on the role of intangible assets in economic growth and firm productivity.

The paper proceeds as follows: Section 1 describes the model that serves as our framework for measurement. Section 2.2 starts with a simplified, static, single-firm version of that model and describes our methodology for measuring data's contribution to firms' forecasting accuracy and the flow value of that data to firms. Section 2.3 adjusts data flow value for data purchased from a data platform or vendor. What is left is the value of firms' internally generated data, which is not typically counted as an investment. Section 3 presents the methodology to capitalize that flow value as an asset. The key step here is to derive a discount rate for data. Section 4 applies this methodology to heterogeneous firms and presents preliminary evidence of data value heterogeneity. Section 5 discusses policy implications and concludes.

Related literature The economic value of data has been a growing focus in recent research, with multiple strands of literature contributing to our understanding of its impact at the firm and macroeconomic levels. While prior studies have examined the role of intangible

assets broadly and proposed various methodologies to quantify the value of data, this paper introduces a novel approach: using the revenues of data platforms to infer the aggregate value of data assets to firms. This method contrasts with existing work in several key ways.

A closely related paper is Asriyan and Kohlhas (2024), who measure firms' revenue forecast errors and show that these errors are reasonable measures of firms' data. They correlate with firms' profits and productivity. Furthermore, they establish a relationship between forecast errors and firm size. In related theory work, Jones and Tonetti (2020) share our focus on data sharing. Begenau et al. (2018) model the data of large and small firms, but with a focus on their financial data. . We build on this work by distinguishing between internally generated (or "own") data and purchased data, capitalizing the data flow, establishing that the size-data relationship is flattening over time, and decomposing the sources of this changing relationship.

Data is a specific example of an intangible asset with its own mis-measurement literature. Crouzet and Eberly (2018) and Corrado et al. (2016) examine intangible assets as a broad class that includes brand equity, personal relationships, goodwill, and patents. Hulten and Hao (2008) explore how the mismeasurement of intangibles distorts economic indicators and propose methodologies to impute the value of software and R&D, an approach extended in Hulten and Nakamura (2020). However, data is distinct from these other assets because it is generated as a by-product of economic activity. As such, it is a good candidate for explaining the strategic advantages of large firms that generate lots of data by participating in a large number of transactions. At the same time, data depreciates differently from other assets. This is at the heart of the trade-off we explore and is unique to data.

Many papers have noted the mis-measurement of GDP that arises from the digital economy and proposed remedies. Nakamura (2010) highlights that many data-related investments are expensed rather than capitalized, leading to an underestimating GDP. He suggests imputing the value of free digital services based on consumer willingness to pay, an approach later

extended by Brynjolfsson et al. (2019), who estimate the consumer surplus generated by digital goods. Their findings suggest that traditional GDP metrics significantly underestimate welfare gains in data-driven economies. However, these consumer-centric approaches do not address the firm-side valuation of data assets. Nakamura et al. (2017) advocate for developing satellite accounts to systematically track data-related investments, an approach that complements traditional national accounts. While such methodologies have been applied in pilot studies, they remain limited in scope and are not yet widely implemented. Syverson (2017) investigates whether digital services could explain the productivity slowdown but finds that the discrepancy would need to be implausibly large to account for observed trends. Finally, internal presentations by the Bureau of Economic Analysis have proposed data measures based on the cost of recording, organizing, and maintaining databases Calderón and Rassier (2022). However, since customers often provide data as a by-product of economic activity, a cost approach is likely to miss much of data's value.

A distinct body of work explores how data boosts firms profits. Brynjolfsson and McAfee (2014) document significant productivity enhancements at the firm level due to data utilization. Their findings underscore the firm-level importance of data but do not provide an economy-wide valuation. Tambe et al. (2021) infer the value of data assets by analyzing hiring patterns and identifying the skills of data-complementary workers, offering an indirect approach to measuring data's impact on firms. Hulten (2010) provide a case study on Microsoft, demonstrating that the company's valuation is almost entirely driven by intangibles, likely including proprietary data.

1 A Framework for Data Measurement

At its heart, this is a forecasting model. It is not a production economy with equilibrium prices. The model is simply a device to organize measurements of forecasts and profits and help us to distinguish and value various types of information. We consider one firm, which

is one among many. Since firm interactions are not our object of study, we can consider a representative firm. Later, we add back multiple firms and consider how to aggregate them.

Costs and Revenues: Time is discrete and infinite. There is a continuum of competitive firms indexed by i. Each firm produces $y_{i,t}$ units of a final good using capital, according to the following production function

$$y_{i,t} = k_{i,t}^{\alpha}.$$

Each unit of production is characterized by an uncertain "state", which we denote by θ_t. The state captures, for instance, characteristics of the good that all customers value at time t (design, availability, adaptability, etc). We can then define a state-adjusted unit of the good as A_t, which only depends on the period (a market characteristic) and not on the firm. We assume

$$A_t = A + b\theta_t$$

Normalizing the price of the good by $P_t = 1$, we can write the total revenues of the firm as

$$R_{i,t} = P_t A_t k_{i,t}^{\alpha} = (A + b\theta_t) k_{i,t}^{\alpha}.$$

The cost of production also depends on the realized state, as the firm needs to adapt production to satisfy the characteristics valued by costumers at period t. We assume the cost function per *state-adjusted unit of production* net of rental capital is,

$$c_{i,t} = b\theta_t - \gamma(a_{i,t} - \theta_t)^2$$

First, the unit cost depends on the realized characteristics valued by costumers, with a higher θ_t generating more revenues but also more costly to produce. Second, the cost also depends on how well the firm targets the state. When firms take actions that are not well-aligned with the profit state θ_t, its unit costs are higher.

Assuming the rental rate of capital is r_t, total costs are

$$C_{i,t} = c_{i,t}k_{i,t}^{\alpha} - r_t k_{i,t} = (b\theta_t - \gamma(a_{i,t} - \theta_t)^2)k_{i,t}^{\alpha} - r_t k_{i,t}.$$

We assume that A, b, α and γ are known parameters. Since b is equivalent to normalizing units of variance and does not affect our estimates, we assume $b = 1$ henceforth.

Now we can define per-period unit profits net of capital cost as

$$\pi_{i,t} = A - \gamma(a_{i,t} - \theta_t)^2$$

and total profits as

$$\Pi_{i,t} = \pi_{i,t}k_{i,t}^{\alpha} - rk_{i,t} \tag{1}$$

Information sets The state is the combination of a "fundamental" component, $\widehat{\theta}_t$, which is forecastable, and a "transitory" component η_t, which is nonforecastable. More precisely, the state at time t is determined by

$$\theta_t = \widehat{\theta}_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{2}$$

Assume $\widehat{\theta}_t$ comes from an AR(1) process with normally-distributed innovations:

$$\widehat{\theta}_{t+1} \;=\; \rho\widehat{\theta}_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim N(0, \sigma_\epsilon^2), \tag{3}$$

for $0 < \rho < 1$. This structure implies that in each period t, the state is normally distributed: $\theta_t \sim N(0, \phi^{-1})$, where ϕ is a function of the parameters in (2) and (3).

We assume that each firm has an information set $\mathcal{I}_t$, which contains data that helps the firm forecast $\hat{\theta}_t$, the forecastable part of the profit state. When combined optimally with an algorithm, this data can be used as a signal s_t about θ_t that has normally distributed signal

noise: $s_t = \hat{\theta}_t + \xi_t$, where $\xi_t \sim N(0, \tau^{-1})$.

In addition, we allow for the firm to have data from a data seller or platform. That data also has normally-distributed noise. By Bayes law, it can, therefore, be aggregated into one signal that provides data of precision ψ to the average firm.

Given that η_t is a nonforecastable transitory shock, forecasting revenue is the same as forecasting the fundamental component plus some mean-zero noise. Define the stock of data to be $\Omega_t := V[\hat{\theta}_t|\mathcal{I}_t]$. We define Ω with the inverse because this lends itself to interpreting Ω_t as the stock of data. A lower variance or more accurate estimate means we have more data about $\widehat{\theta}_t$.

Solution The firm chooses a sequence of state-contingent production to maximize

$$\sum_{t=0}^{\infty}(1+r)^{-t}\mathbb{E}\left[(A - \alpha\left(a_{i,t} - \theta_t\right)^2)k_{i,t}^{\alpha} - rk_{i,t}|\mathcal{I}_t\right]. \tag{4}$$

The firm uses its prior about θ_t and its data, represented by s_t to take the expectation of the objective function (4). The firm's first-order condition dictates that the action that maximizes the expectation of that objective is $a_t = E(\theta_t|\mathcal{I}_t)$. Thus, the objective consists of minimizing the expected squared forecast errors $\left(E(\theta_t|\mathcal{I}_t) - \theta_t\right)^2$ about a state θ_t. The expected squared forecast error is, by definition, a conditional variance.

$$V(\theta_t|w_t) = \frac{1}{\phi + \tau + \psi}, \tag{5}$$

which depends on prior variance ϕ and own-firm data with precision τ.

How good this expectation is taken determines the marginal return of investing in capital and, hence, the choice of capital and total production. Here we shut down the potential spiral that more production also generates more data.

Belief updating: Applying the Kalman Filter Since firms know their capital decision, forecasting total revenues is the same as forecasting per unit revenues, for which only the state θ_t is uncertain. First, we need to derive prior beliefs and variances for the dynamic model state using Bayes law. Given the linear revenue function, those are the same as the mean and variance of the revenue, up to a constant. Then, we update beliefs with the new data firms observe.

The prior mean and variance of revenue per unit of production are $A + E[\theta_t|\mathcal{I}_t] = A + E[\hat{\theta}_t|\mathcal{I}_t]$ and $V[\theta_t|\mathcal{I}_t] := \Omega_t^{-1} + \sigma_\eta^2$, where $\mathcal{I}_t$ represents whatever information set the firm has at the start of time t.

Next, determine today's forecast and variance of tomorrow's state. Taking the mean and then the variance of both sides of (3) once combined with (2), we get

$$E[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho E[\hat{\theta}_t|\mathcal{I}_t], \tag{6}$$

$$V[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho^2\Omega_t^{-1} + \sigma_\epsilon^2. \tag{7}$$

Equation (7) tells us the conditional variance of the forecast of future state $\hat{\theta}_{t+1}$, conditional on the information known at the start of time t. A conditional variance is not a measure of the volatility of $\hat{\theta}_{t+1}$. Rather, it is a measure of uncertainty. It is the expected squared forecast error: $V[\hat{\theta}_{t+1}|\mathcal{I}_t] := E[(\hat{\theta}_{t+1} - E[\hat{\theta}]_{t+1}|\mathcal{I}_t])^2|\mathcal{I}_t]$. It reveals how inaccurate our forecasts are. In Bayesian language, this is a prior variance of θ_{t+1}.

If new data is used to forecast $\hat{\theta}_{t+1}$ with normally-distributed noise, then Bayes' Law says that we can combine that data and represent it as a signal about tomorrow's state $s_t = \hat{\theta}_{t+1} + e_{st}$, with $e_{st} \sim N(0, \sigma_s^2)$. The firm will also observe time-t revenue at the end of period t. In other words, $\mathcal{I}_{t+1} = \{\mathcal{I}_t, \theta_t, s_t\}$. The information the agent will have to make decisions tomorrow is the information available today, plus the new revenue and the new data observed at the end of period t. This construct assumes the data is informative about θ_{t+1} but has noise that is independent of the previous noise in the stock of data. If new

data were correlated with old data, we could want to count only the part of the new data that constituted independent information. In a simple case, this might be accomplished by regressing new data on old data and using the residual to form s_t.

Observing the revenue θ_t is a signal about the unobserved state $\hat{\theta}_t$, with signal noise variance equal to σ_ϵ^2.

According to Bayes law for normal variables, the posterior precision of the estimate of $\hat{\theta}_{t+1}$ is the sum of the prior precision and the precision of both the new data and the revenue signal. In other words,

$$V(\hat{\theta}_{t+1}|\mathcal{I}_{t+1})^{-1} = \left[\rho^2\Omega_t^{-1} \ + \sigma_\epsilon^2\right]^{-1} \ + \sigma_\eta^{-2} + \sigma_s^{-2} := \Omega_{t+1}. \tag{8}$$

Recall from (4) that the firm's objective depends on the conditional variance, or expected squared forecast error, not of the state $\hat{\theta}$, but of the revenue θ_t, which is the state, plus transitory, unforecastable noise. Thus firm revenue depends on

$$V(\theta_{t+1}|\mathcal{I}_{t+1}) = \left[\left[\rho^2\Omega_t^{-1} \ + \sigma_\epsilon^2\right]^{-1} \ + \sigma_\eta^{-2} + \sigma_s^{-2}\right]^{-1} + \sigma_\eta^2 = \Omega_{t+1}^{-1} + \sigma_\eta^2. \tag{9}$$

We can invert this expression to get the firm's stock of data, as a function of of σ_η^2, which we can estimate and the firm's revenue forecast variance $V(\theta_{t+1}|\mathcal{I}_{t+1})$:

$$\Omega_{t+1}^{-1} = V(\theta_{t+1}|\mathcal{I}_{t+1}) - \sigma_\eta^2$$

Deriving Data Depreciation Equation (8) maps the time-t stock of data Ω_t, into time-$t+1$ stock of data. That law of motion says that we take the stock Ω_t, depreciate it by transforming it into $(\rho^2\Omega_t^{-1} + \sigma_\epsilon^2)^{-1}$ and then add on the flow of new data. New data is like the new investment in the data stock. This implies that for data, the depreciation rate is

$$\zeta_t = 1 - (\rho^2 + \sigma_\epsilon^2\Omega_t)^{-1}. \tag{10}$$

This depreciation rate teaches us that if the AR(1) process for the unobserved state is highly volatile (high σ_ϵ), then the stock of data will depreciate quickly. Also, we learn that large stocks of data depreciate at a faster rate than small stocks.

Capitalizing the flow value of data We can now write the *expected per-period profit* from the objective (4) as an explicit function of data stock as

$$\mathbb{E}\left[(A-\gamma\left(a_{i,t}-\theta_t\right)^2)k_{i,t}^{\alpha}-rk_{i,t}|\mathcal{I}_t\right]=\pi(\Omega_{i,t})k_{i,t}^{\alpha}-rk_{i,t} \tag{11}$$

where we define $\pi(\Omega_{i,t})=\mathbb{E}_{\Omega_{i,t}}(\pi_{i,t})$.

The optimal sequence of capital investment $\{k_{i,t}\}$ then solves the following recursive problem:

$$\mathbb{V}(\Omega_{i,t})=\max_{k_{i,t}}\pi(\Omega_{i,t})k_{i,t}^{\alpha}-rk_{i,t}+(1+r)^{-1}\mathbb{V}(\Omega_{t+1}). \tag{12}$$

This specification collapses into a deterministic problem with choice variable $k_{i,t}$ and state variable $\Omega_{i,t}$. To value a firm's data, consider a firm that does not add any data to its stock in any future period. The value of that firm is the value of its current data. For a firm that adds no new data, the future data stock is the depreciated current data stock: $\Omega_{t+1}=(1-\zeta)\Omega_t$.

Approximation Given a stock of data $\Omega_{i,t}$, the optimal capital choice in period t is given by $\alpha\pi(\Omega_{i,t})k^{\alpha-1}=r$. Hence,

$$k_{i,t}^{*}=\left[\frac{\alpha\pi(\Omega_{i,t})}{r}\right]^{\frac{1}{1-\alpha}}$$

If we approximate the profit function linearly a function of $\Omega_{i,t}$, this is $\pi(\Omega_{i,t})\approx\pi\Omega_{i,t}$.

Estimating the Depreciation Rate of Data In our model, firms predict a variable proportional to revenue. Therefore, we use the dynamics of firm revenue to estimate the data depreciation rate. In our 2002-2022 sample of Compustat firms, we estimate (3) for all firms and years on firms' revenues that are detrended using an HP filter. (See step 3 above

for details.)

We find that the persistence coefficient is estimated to be $\rho = 0.34$. The variance of the innovations was estimated to be $\sigma_\epsilon^2 = 0.08$.

To determine the depreciation rate, we also need an estimate of the stock of data. This stock of data is defined as $\Omega_t = V[\theta_{t+1}|\mathcal{I}_t]^{-1}$. In our forecasting model, firms have three sources of information, with precisions ϕ. τ and ψ. Using Bayes law for normal variables, the conditional precision of the firm's forecast is thus $\Omega_t = \phi + \tau + \psi$. All three parameters were calibrated in the previous section. We estimated that $\Omega_t = \phi + \tau + \psi = 22.22$.[2]

Thus, the depreciation rate estimate is

$$\zeta = 1 - (0.34^2 + 0.08 \cdot 22.22)^{-1} = 0.47 \tag{13}$$

If we approximate the depreciation rate as a linear rate (constant zeta) and the profit function as linear in data ($\Pi(\Omega_t) = \pi\Omega_t$), given that the flow value of data in dollars was estimated to $\Pi = 21.21$ million 2022 dollars per firm, then the per unit of data is $\pi = 21.21/(12.40 + 9.82) = 0.95$ the value function will also be linear and solvable analytically:[3]

$$V(\Omega_t) = \nu\Omega_t \qquad \text{where } \nu = \frac{\pi}{1 - \beta(1-\zeta)}. \tag{14}$$

Substituting in values with $\nu = 0.95/(1 - 0.96(0.53)) = 1.94$ yields a capital value

$$V(\Omega_t) = 1.94 \cdot 22.22 = 43.18$$

To put this number in context, the capitalized value of data for the average firm represents around 14% of a year's revenue.

For the discount factor, we use a constant rate rather than a stochastic discount factor

[2]This is also simply the inverse of the forecast error from step 1 above. Should we be using the entire stock of data the firm has to depreciate? Or only the firm's private stock of data?

[3]This expression comes from solving $\nu\Omega_t = \pi\Omega_t + \beta\nu(1-\zeta)\Omega_t$ for ν.

because our framework has no covariance of data value with any aggregate risk. Therefore, we simply discount by the riskless rate: β is the rate of time preference. Future work could introduce stochastic valuation that would realistically discount the value of data for risk.

2 Measuring The Flow Value of One Firm's Data

A central object of our study is the distribution of data values across firms. However, we build up our understanding of where data value comes from by starting with a one-firm analysis for a median firm. This first set of measurement exercises are very simple, but demonstrate the essence of our measurement approach and show that the measures are reasonable, in a transparent way. First, we describe the data we use. Then, we consider the flow value of one firm's data. We measure the value for an average public U.S. firm. Flow value means the additional profit per year that a firm with this data earns. The, we decompose this data value into the value of data the firm generates internally and the value of purchased data.

2.1 Data Description

Measuring Prior Variance and Persistence of Revenues. We construct the dataset by using Compustat North-America annual fundamentals data for the years 2002-2022. We restrict attention to firms incorporated in the US, listed on one of the three major exchanges (AMEX, NYSE, Nasdaq). Following the convention in finance literature, we exclude utilities and financial firms from the data. We also exclude the firm-years with missing values of revenue, profits, or total assets. This leaves us with 65,988 firm-year observations. We adjust all nominal figures for inflation by using the CPIAUCSL series and report all figures in terms of December 2015 dollars.

To estimate the prior variance of revenue, we first create a new variable sales-ratio by dividing the revenue of the firm by it's trend revenue, obtained by using an HP filter. We winsorize this variable at 1% level. We then run an AR(1) regression on this variable by

pooling all firm-year observations. The RMSE reported by this regression gives an estimate of the prior standard-deviation.

Measure of the Squared Relative Forecast Error. We use the data from I/B/E/S Guidance and compute the sales forecast accuracy of US public firms changes over time. Guidance refers to any forward-looking statement made by a company to offer insights into its future financial performance. I/B/E/S Guidance extracts quantitative company expectations from press releases and transcripts of corporate events.

We retrieved the annual sales guidance from I/B/E/S Guidance via the WRDS platform on April 20, 2023. We use both the "Detail" and "Identifiers" tables. The "Detail" table contains all kinds of financial estimates (e.g., annual or quarterly sales or earnings) made by firms. The "Identifiers" table contains identifying information, including "ticker" and "cusip." We merge the two tables and only keep US firms (observations with "usfirm"=1). Our sample runs from 2002 to 2021. We chose 2002 as the starting date because the sales guidance data only became available for more than 10 firms after 2001. Each year, the sample size varies from 50 to 550 firms. After 2021, the number of firms with valid sales guidance dropped below 10 again.

Firms revise and update their annual sales guidance over the course of the fiscal year as realized quarterly sales data comes in. Since our analysis focuses on firms' ability to forecast annual sales, we only keep the initial estimates, which are not affected by realized quarterly data. When the sales forecast is expressed as a range, we take the mid-point (the lower and upper bound average). To adjust for inflation, we use the Consumer Price Index for All Urban Consumers (CPIAUCSL) monthly series downloaded from the FRED[4] and take the average in each year. We express all the sales numbers in 2002 dollars. We compare firms' sales guidance with realized sales data (retrieved from Compustat North America) and define the squared relative forecast error as in Equation 17. We winsorize the relative

[4]https://fred.stlouisfed.org/series/CPIAUCSL

forecast error at the one percent level on the right tail. In each year, we take the average across firms.

2.2 Measuring the Value of Own Data

We begin with a setting where no firms purchase information ($\psi = 0$) and data is one-period lived because there is no persistence in the state ($\rho = 0$). Since there is no persistence, the state and observation equations (2) and (3) are redundant. Therefore, we set $\sigma_\eta^2 = 0$, which means that forecasting the hidden fundamental state $\hat{\theta}$ is the same as forecasting the profit-relevant state θ_t. This allows us to understand the simple static model in isolation before adding back these richer features.

In this simple setting, the objective consists of minimizing the expected squared forecast errors $\left(E(\theta_t|\mathcal{I}_t) - \theta_t\right)^2$ about a state θ_t, which turns out to be directly forecast the firm's revenues. The expected squared forecast error is, by definition, a conditional variance.

$$V(\theta_t|w_t) = \frac{1}{\phi + \tau} \tag{15}$$

where all firms have the same prior variance ϕ and the same own-firm data w_t, with precision τ. Assuming that firms make their forecast purely using their own data, how important is the private flow of data relative to the prior information? What is the value of firms' own data relative to their revenue and in terms of dollars?

Since, revenue is $A + b\theta_t$, this expected squared forecast error for the state implies an expected squared revenue forecast error of

$$V(\theta_t|w_t) = \frac{b^2}{\phi + \tau} \tag{16}$$

We proceed to obtain τ in three steps. First, we use results from Asriyan and Kohlhas (2024) to estimate expected forecast errors. They study firms' forecasts of their future sales

revenue. They define the mean squared relative forecast error to be the average over time and over firms of[5]

$$\text{Squared Relative Forecast Error} = \left| \frac{\text{Forecasted Sales} - \text{Actual Sales}}{(\text{Forecasted Sales} + \text{Actual Sales})/2} \right|^2. \tag{17}$$

The average squared relative forecast error between in the sample 2002 and 2021 is 0.045.[6] In our model, all firms use prior information and private data to achieve a forecast error of the state θ_t given by (15).

Second, we estimate the prior variance $1/\phi$ by estimating the residual variance of firm revenues after removing a trend and the part of revenues that are predicted by the prior year's revenue. To do this, estimate a trend for sales (revenue) of each firm in Compustat, using an HP-Filter. Then, divide sales by trend sales for each firm. This puts each firm in comparable units of percent deviation from its trend. Next, pool the data across firms and estimate an AR(1) on the sales ratio. The mean-squared error (MSE) for the AR(1), estimated across all firms/periods is $0.284^2 = 0.08$. This gives us the prior information precision, $\phi = 1/0.08 = 12.40$.

Finally, from equating this model forecast error in (16) (normalizing $b = 1$) to the average value 0.045, we obtain the precision of own information as follows.

$$0.045 = \frac{1}{12.4 + \tau} \quad \Longrightarrow \quad \tau = 9.82$$

This implies that the precision of private data accrued in the year is of the same order of magnitude as the precision of prior information.

What is the value of private data in terms of revenues? A key finding in Asriyan and Kohlhas (2024) is that a 1% point decrease in squared relative revenue forecast error (as

[5]This is not exactly how AK define the error. They use log differences, instead of the percentage differences. But we know these are good approximations of each other and the percentage change representation works with our simple, linear-normal model structure.

[6]We provide details on our computation of this error in Appendix 2.1.

defined in 17) (e.g., 0.03 to 0.02) is associated with a 1.9% increase in profits. Assuming that the impact of forecast in profits comes from affecting revenues and not costs, this also captures the effect on revenues. We denote this semi-elasticity of revenues to forecast errors as α. The medium firm in the sample has \$391m in revenues 2022, which implies an additional \$7.4m in revenues for a 1% point decrease in squared relative revenue forecast error.

We can now compute π, defined as the flow value or extra annual profit a firm receives because of its own data. This is,

$$\pi \equiv \alpha \left(\frac{1}{\phi} - \frac{1}{\phi + \tau} \right) = 1.9 \left(0.08 - 0.045 \right) = 0.0665 \tag{18}$$

The units here are a fraction of the firm's revenue. The median firm in our sample earned \$319m in 2022 in revenue. Multiplying this by 319 yields $\Pi = 319 \cdot 0.0665 = \$21.21$ mil. For the average firm, this implies that around 21 million dollars per year of flow value should be assigned to the firm's own data.

Evolution of Data Flow Values Next, we examine how the flow value of a firm's data has changed over time. While this is still a very rough estimate, examining the changes in estimated data value will give us a sense of whether the results are sensible.

We repeat the flow value estimation described above for each year from 2002 - 2021. We hold fixed all parameters. Holding all parameters fixed is surely not correct. However, it makes for a simple first pass. We add more realism, step-by-step, in order to understand the role each assumption plays in shaping our estimates. In each year, what is changing is the average firm's forecast error and the revenue of the median firm.

Figure 1 shows the evolution of the mean squared relative forecast errors in different years (this is $V(\theta|w_t)$)) and the implied precision of average firm (τ from the equation (15), holding fixed our estimate of the prior precision ϕ). There is a clear negative, non-linear,

relation between these two objects.

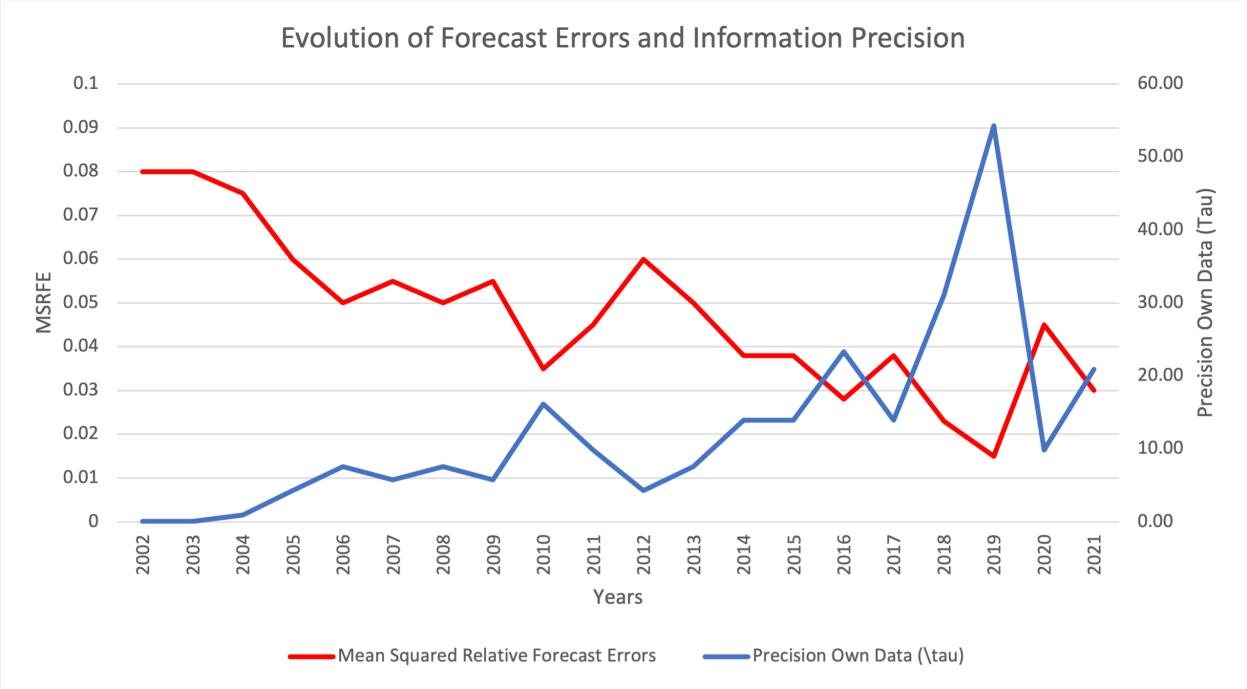


Figure 1: Forecast Errors and Precision of Own Data.

We use the series of observed average forecast errors to see what our framework teaches us about the flow value of this data. Figure 2 shows a gradual rise in the annual income firms derive from their data. While the solid line shows a gradual increase in the value of data that combines the value of data from forecasting better one dollar of revenue and the increase in the revenue of the median firm, the dashed line shows the value of data maintaining the revenue of the median firm fixed at the 2002 level. This decomposition shows that half of the increase in the value of data can be attributed to better data and half to a larger scale in the use of such data.

The gradual growth might be surprising, given the exponential explosion of data storage, transmission, and cloud computing. However, keep in mind that data used for forecasting has diminishing returns. The tenth data point reduces forecast errors much more than the ten millionth. The diminishing returns and exponential growth resulted from a gradual rise in annual profits derived from data.

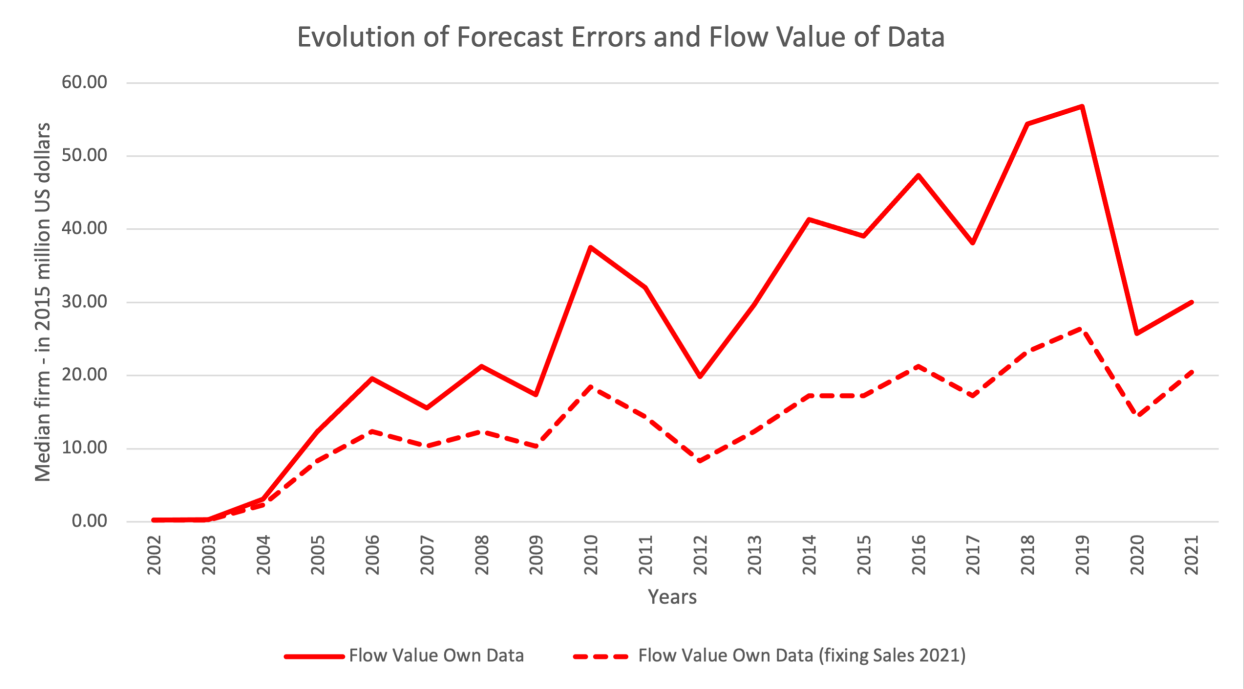


Figure 2: Annual Income Derived by the Average Firm from its Own Data.

2.3 Measuring the Value of Purchased Data

While the previous valuation is not explicitly assigned to own data in the GDP figures, there is a market for data that explicitly assigns value to data on the national accounts when measuring GDP. Several platforms collect and sell data directly or with previous processing and advise orientation. This ranges from general data (such as Snowflake, Google or Amazon), to financial data (Bloomberg, Refinitiv, or Quandl), consumer data (Eagle, SafeGraph or Experian), healthcare (IQVIA, HealthVerity or Komodo), real estate (Zillow, CoreLogic) or labor and tech (LinkedIn, Crunchbase or PitchBook).

While it is tempting to measure the GDP value of data directly from the GDP of these platforms, an important question is how large this is compared to the value of firms own data, estimated above. To estimate this relative value, we now consider a version of our model that is still static in nature ($\rho = 0$ and $\sigma_\eta = 0$), but where firms buy data from platforms. We assume a platform that provides data of precision ψ to the average firm.

That data generates additional annual revenue for the firm, given by

$$\alpha \left(\frac{1}{\phi + \tau} - \frac{1}{\phi + \tau + \psi} \right) \tag{19}$$

where again $\alpha = 1.9$ is the parameter that maps precision units into units of additional revenues per year.

Note that ψ incorporates how many firm outcomes are being averaged by the platform, whether it adds noise to the reports, and even what fraction of firms buy the platform's data. As an average precision, this is lower if a smaller fraction of firms get no external data.

For this exercise we use information from a platform in particular, Visa Inc. Visa is major player in the global payments industry, processing substantial transaction volumes annually. According to its 2024 10-K earnings statement, "Visa's total payments and cash volume was $16 trillion, and the Visa network processed 234 billion total transactions — 639 million transactions every day." This volume is larger than the main competitors like Mastercard (handling 171 billion transactions a year) or American Express (with $1.5 trillion in total payments).

This volume of transactions uniquely positions Visa as an intermediary of data and data analytics. According to its 2024 10-K earnings statement, "we have continued to expand our services beyond Visa transactions to no-Visa transactions and non-payment services.....For example, in Advisory Services alone, we delivered more than 3,000 consulting engagements during 2024, up nearly 50 percent from last year, and we estimate that we helped clients realize over $5 billion in incremental revenue as a result."

This last statement implies that by selling analytics based on unique data from transactions across the board, Visa generates $1.67m additional revenues per firm, which constitutes the value of Visa's data to those firms. We equate this figure to (19), the additional profit generated by platform revenue, above what the firm would have earned using its own data. As before, we express everything in percentages of firm revenue of the median firm in our

sample ($319m in 2022).

$$1.67/319 = 1.9\left(\frac{1}{12.40+9.82} - \frac{1}{12.40+9.82+\psi}\right) \quad \Longrightarrow \quad \psi = 1.45.$$

Substituting our parameter estimates into (20) yields

$$\alpha\left(\frac{1}{\phi} - \frac{1}{\phi+\tau}\right) = 1.9\,(0.045 - 0.042) = 0.0057 \tag{20}$$

The units of this answer are as a fraction of firm revenue. Multiplying the estimated fraction of revenue but the average firm's revenue ($0.0057 * 319m$) reveals that the nonmeasured value of the data is $1.67 million, for the average firm.

Combining this result with the one in the previous section, the nonmeasured data value in GDP figures is an order of magnitude larger than the one assigned to data. More specifically, while data sold by platforms (measured as such in GDP) imply an increase of $1.67 million per average firm, data generated within firms (not explicitly assigned to data in GDP figures) imply an increase of $21.21 million per average firm, a factor or 12.7.

3 From Flow to Stock: Capitalizing the Value of Data

In order to value data as a multi-period-lived asset, we now consider the full dynamic model. We now assume that the state has persistence $\rho > 0$, which makes data from past periods relevant for forecasting today's state. This generalization of the model implies that data obtained in one period has value for multiple future periods. It is this multi-period valuation that we compute next.

The key question in this dynamic value measurement is determining how much data depreciates. Data about yesterday's state is less relevant to today's state because the state is changing. That decline in relevance is the source of data depreciation. While previous studies have guessed a depreciation rate, such as (Calderón and Rassier, 2022), we will derive

the depreciation rate from our model and estimate it with firm revenue data.

3.1 Evolution of Data Stock Values

How does the value of the average firm's data stock evolve over time? To answer this, we repeat our one-period data valuation exercise in the years from 2002-2021; as before, we hold parameters fixed.

We do not add these valuations up. In each period, the forecast errors inform us about the total data firms have at that time. These are stocks, not flows. So, we back out the firm's data from forecast errors in each period. Then, we consider the flow of value the firm would reap from that data stock. In the next period, we consider the firm's next stock of data and capitalize that specific stock of data.

Figure 3 shows the evolution of the stock of data an average firm holds in each year, measured as $\Omega_t = (\phi + \tau)^{-1}$, and the corresponding depreciation ζ_t from equation (10). It is clear that a larger stock of data depreciates faster.

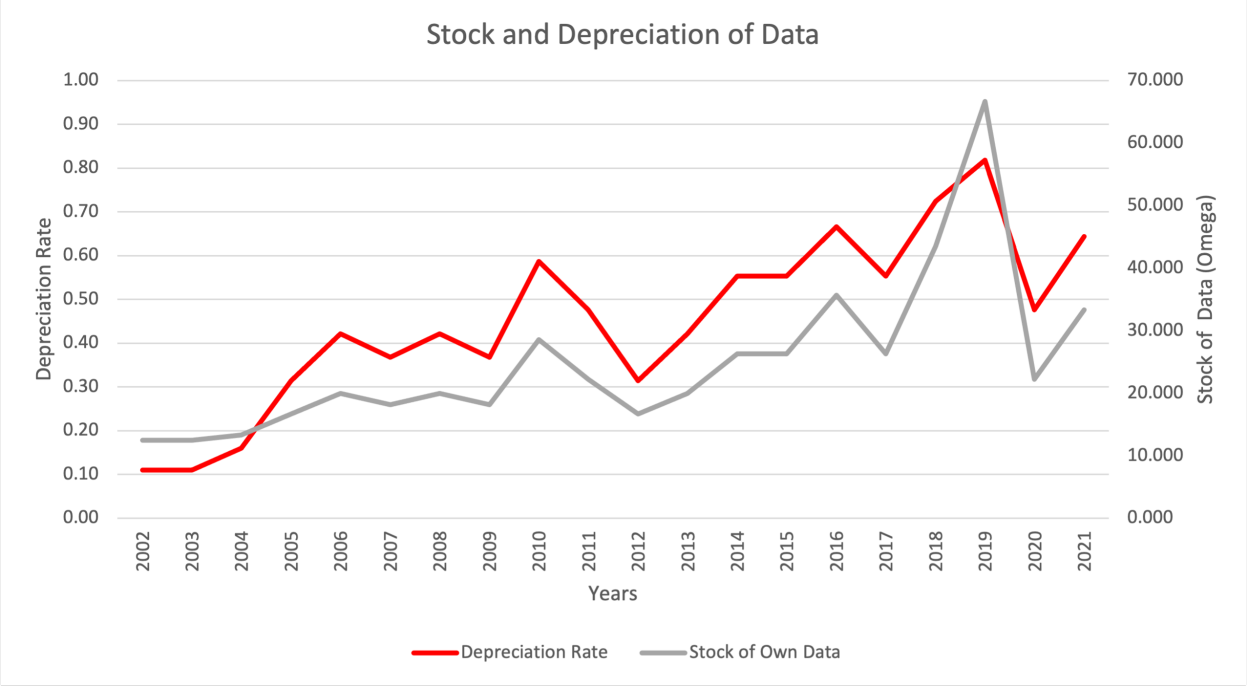


Figure 3: Stock of Data (Ω_t) and Depreciation Rate (ζ)

Figure 3 shows a steady rise in the value of firms' data assets that mirrors the growth in data flow value. Interestingly, the increase is even slower than the increase in the data flow. In dashed, we have plotted the asset value of data fixing the 2002 revenue of the median firm. The asset value increases because the flow value of data increases, but the growth is slower because data depreciation also increases as the stock of data grows. In solid, we also allow the revenue of the median firm to increase. This captures the higher flow value of data from the solid line in Figure 2.

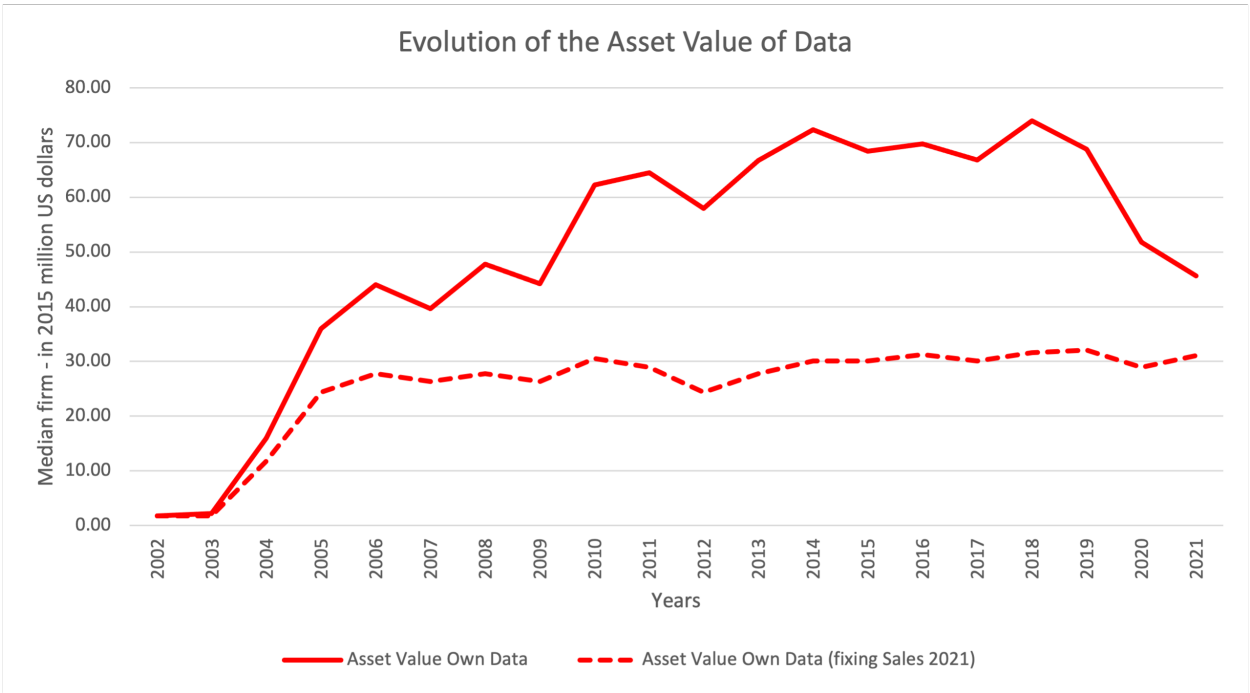


Figure 4: Capitalized Value of the Average Firm's Data Assets.

4 Aggregating the Data of Heterogeneous Firms

The previous analysis is based on taking averages across firms that, in principle, are quite heterogeneous regarding economic activity, technology, size, and dimensions of uncertainty.

4.1 Evidence of Data Democratization

To grasp the extent of heterogeneity in forecast errors and their evolution over time, we split firms by employment quintiles in 2005; we computed their forecast errors relative to other firm sizes. More specifically, we computed squared errors of log-revenues divided by total intangible assets per bin. In Figure 5, we plot these forecast errors minus the average forecast errors in the sample size (for 2005, the mean of that measure is about 0.0005 and is about 0.0002 for 2022). We report the log difference between the average scaled measure for an employment bin and the overall average (which is 1 by construction). Since employment in 2022 is larger than in 2005, we included all the firms in 2022 with more employment than the 2005 maximum in the last bin. Hence, a somewhat higher number of firms in the last bin (shown in brackets).

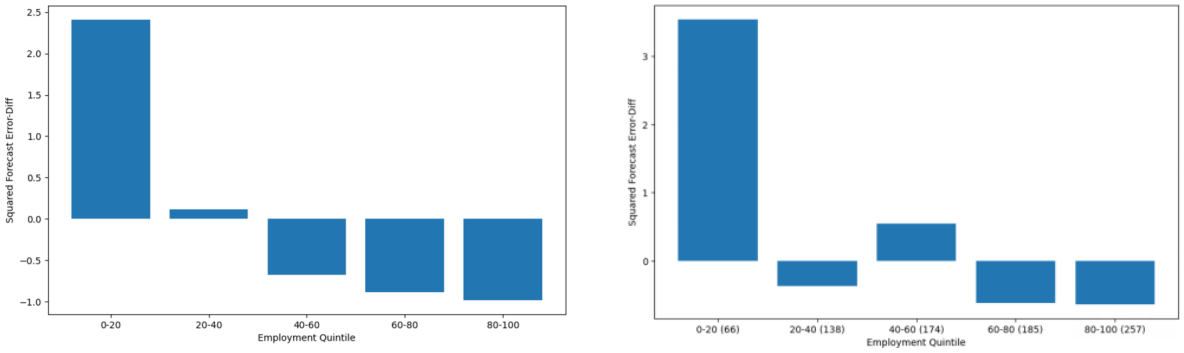


Figure 5: Squared Forecast Errors by Employment Size (2005 and 2022)

Notice that although the forecast errors have declined between these two years, the distribution has also changed dramatically. In 2005, there was a clear monotonic negative relation between size and forecast errors. The largest firms were making fewer forecast errors than the next bin, with the smallest firms making much more mistakes in forecasting than the rest. While this relation persisted in 2022, the differences among the upper four bins dissipated, while the smallest firms were left behind much more.

These results suggest that the differences in data available to forecast were determined by size over the whole distribution in 2005 but not in 2022, except for the very small firms.

Indeed, the differences in forecast errors for firms other than the smallest became very similar, pointing towards the dissipation of the largest firms' advantage in gathering data.

4.2 Dissecting Sources of Data Democratization

Why did the forecast errors of mid-sized firms fall more than those of large firms? There are three forces depressing the value of large firms' data. One possibility that explains this trend is the rise of platforms and the more extensive selling and purchasing of data. By purchasing data, buying reports or paying for data services, mid-sized firms can draw on larger data sets, for a fee. A second possible explanation is that data had diminishing returns. Data is valuable because it reduces forecast errors. Those are bounded below by zero. The third explanation is that large data sets depreciate more quickly than small ones (Figure 3). When forecast errors are near zero, the additional noise from the state innovation is a large proportional increase in the conditional variance of the forecast. This shows us as a large drop in precision from period to period, which is a large depreciation rate. Our model teases out and quantifies these three channels.

4.3 Evidence on Firms' Platform Data Purchases

Based on the case-studies reported by Visa, their typical customer is a medium size Fin-tech company or a large credit union. Typical examples include firms like Remitly, University of Virginia credit union, and Bankcard USA. These firms typically have revenue in the range of a few hundred million USD. They use data and consulting services from Visa to understand their customers and competition better. Visa reports that lenders and payment companies use these insights for risk management activities like fraud prevention while others use them to identify opportunities to increase customer base. These examples support our finding that it is primarily mid-size companies acquiring data. These companies purchases of relevant data substitute for their inability to generate that data internally, at scale.

5 Conclusion

This paper presents a novel framework for valuing firms' data, recognizing it as a productive asset rather than an unmeasured input. By quantifying data's value through its impact on firms' forecasting accuracy, we demonstrate that traditional measures of investment significantly underestimate firms' true asset base. This mis-measurement affects mid-size and large firms differently and evolves differently over time for such firms. Mid-size firms appear to be catching up.

Our findings distinguish internally generated data from data purchased from a platform. This allows us to assess to what extent data purchases account for the catch-up or "democratization" of data, and to what extent this is simply diminishing returns or high depreciation of large-firm data. These findings provide a new perspective on firm dynamics and the evolution of firm competition.

Addressing these measurement gaps is essential for policymakers, economists, and corporate decision-makers. More accurate valuation of data assets can improve financial market assessments, guide policy on data governance, and refine taxation frameworks for intangible investments. By incorporating a rigorous, market-based approach to data valuation, this research contributes to ongoing debates about the role of intangible assets in economic growth and firm productivity. Future work should explore refinements to our methodology, including sector-specific variations in data value, firm-level heterogeneity, and alternative approaches to estimating data depreciation. Ultimately, recognizing and properly accounting for firms' data assets will provide a more comprehensive and accurate understanding of the digital economy's impact on productivity, investment, and innovation.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press, 2018.

Asriyan, Vladimir and Alexandre Kohlhas, "The Macroeconomics of Firm Forecasts," 2024. Working Paper, University of Oxford.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp, "Big data in finance and the growth of large firms," *Journal of Monetary Economics*, 2018, *97*, 71–87.

Brynjolfsson, Erik and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W W Norton & Co, 2014.

_ , Avinash Collis, and Felix Eggers, "Using Massive Online Choice Experiments to Measure Changes in Well-Being," *Proceedings of the National Academy of Sciences (PNAS)*, March 2019, *116* (15), 7250–7255.

Calderón, José Bayoán Santiago and Dylan G. Rassier, "Valuing Stocks and Flows of Data Assets for the U.S. Business Sector," Presentation at the BEA Advisory Committee Meeting May 2022.

Corrado, Carol, Jonathan Haskel, Cecilia Jona-Lasinio, and Massimiliano Iommi, "Intangible investment in the EU and US before and since the Great Recession and its contribution to productivity growth," Technical Report 2016.

Crouzet, Nicolas and Janice Eberly, "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles," in "Jackson Hole Economic Policy Symposium" 2018, pp. 87–149.

Farboodi, Maryam and Laura Veldkamp, "A model of the data economy. No. w28427," 2021.

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

Hulten, Charles and Leonard I. Nakamura, "Expanded GDP for Welfare Measurement in the 21st Century," in "Measuring and Accounting for Innovation in the Twenty-First Century" NBER Chapters, National Bureau of Economic Research, Inc, January 2020, pp. 19–59.

Hulten, Charles R., "Decoding Microsoft: Intangible Capital as a Source of Company Growth," NBER Working Papers 15799, National Bureau of Economic Research, Inc March 2010.

_ and Xiaohui Hao, "What is a Company Really Worth? Intangible Capital and the "Market to Book Value" Puzzle," *Review of Income and Wealth*, 2008, *54* (3), 379–404.

Jones, Charles and Chris Tonetti, "Nonrivalry and the Economics of Data," *American Economic Review*, September 2020, *110* (9), 2819–58.

Nakamura, Leonard, "INTANGIBLE ASSETS AND NATIONAL INCOME ACCOUNTING," *Review of Income and Wealth*, 2010, *56* (s1), S135–S155.

Nakamura, Leonard I., Jon Samuels, and Rachel Soloveichik, "Measuring the "Free" Digital Economy Within the GDP and Productivity Accounts," Working Papers 17-37, Federal Reserve Bank of Philadelphia October 2017.

Syverson, Chad, "Challenges to Mismeasurement Explanations for the US Productivity Slowdown," *Journal of Economic Perspectives*, May 2017, *31* (2), 165–86.

Tambe, Prasanna, Lorin M. Hitt, Daniel Rock, and Erik Brynjolfsson, "Digital Capital and Superstar Firms," 2021.