The Missing Value of Data

Ankit Bhutani Guillermo Ordonez Laura Veldkamp*

December 17, 2025

Abstract

Data assets are increasingly vital in modern economies, yet macroeconomic measurement is not well-adapted to capturing its value. Part of the problem is that data is an intangible asset: investments in data are missed in national accounts and losses due to depreciation are missed in firms' balance sheets. Another problem, unique to data, is that it constitutes means of payment in the modern economy: consumption bartered for data is also missed in national accounts. We propose an output-based approach to measure the missing value of data. We treat data as an asset, measure it with the quality of firms' revenue forecasts, and endogenously determine its depreciation. We then capitalize data value and explore what measured GDP would be if data were treated and transacted similarly to a physical asset. Our findings suggest that the aggregate value of data is about 4% of GDP, increasing in the last decade to around 6%.

1 Introduction

In the digital economy, data has become one of firms' most valuable and essential assets. Data is highly valued because it is the fuel for artificial intelligence (AI). Rapid progress in 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. Our objective is to provide a flexible approach to quantify the missing value of data on standard masures of GDP.

*ab4462@columbia.edu, ordonez@econ.upenn.edu, lv2405@columbia.edu. We thank participants at the 2025 SED and the Princeton Macro Finance Conference for helpful comments and suggestions. Keywords: AI, big data, GDP measurement, data barter.

Data plays three distinct roles in modern economics: 1) Data is a factor of production. It makes firms more productive by minimizing mistakes; it is monetized, often as advertising, or is sold directly. This activity is measured with standard approaches. 2) Data is an asset that requires costly investments and depreciates. This role of data is not typically captured and is similar to other intangible assets. 3) Data is bartered. Customers compensate firms, in part, with their data. This value, unique to data, distorts both national accounts and firm profits.

Understanding the aggregate value of data assets can help policymakers design privacy, tax and regulatory frameworks.Proper valuation of firms' data assets can improve financial market assessments and corporate valuation models. This is particularly relevant as data-driven business models become increasingly dominant, challenging conventional valuation techniques that were primarily designed for tangible capital.

At its core, AI is a prediction technology, fueled by data (Agrawal et al., 2018). Firms use data to predict demand, costs, inventories, advertising effectiveness, and competitors' strategies (Goldfarb and Tucker, 2019). 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. 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, either in the data or in our model. It captures both the broader productivity benefits of many types of data and the additional data that larger, more profitable firms generate.

A key contribution of the paper is to capitalize the value of data as an asset. That means we consider not just the value stream derived from an amount of data today, but

the flow of value derived from that data in the future, discounted appropriately. When we capitalize data value and study its evolution, we learn that two forces are leveling the data playing field: diminishing returns to data and depreciation.[1] 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 depreciation. Because their forecasts are so precise, changes in the state reduce their precision by more.

Our findings suggest that conventional measures of firm investment systematically underestimate the true value of internally generated data, particularly for large firms, but non-trivially for smaller firms as well. Since standard accounting practices often fail to recognize this data as an investment, the economic impact of future 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.

Our approach tackles a challenging reverse causality question: Do better predictions allow for higher profits? Or do higher profits induce better predictions because more profitable firms do more transactions that generate more data? We embed this two-way feedback in our setting and calibrate it to isolate the value of data when considering this reverse causality challenge.

We also validate our approach based on the output of data (forecasts) instead of input of data (labor cost of processing data) by conducting the following exercise. First, we show that increased employment of data workers is indeed correlated with forecast improvements. Second, we show that, in contrast to forecasts, which shows a significant impact on revenues,

[1]Data purchases may also level the playing field. We consider that separately at the end of the paper.

the use of data workers does not. This suggests that forecasts capture a more direct impact of data on revenues than the inputs to manage data. This result, we believe, rationalizes the use of forecast to capture the value of data in the economy.

The paper proceeds as follows: Section 2 describes the framework we construct for measurement. Section 3 discusses how to take the model to the data and the unique properties of data mismeasurement. Section 4 measures data stock and the sources of data and their dynamics. We also show how to capitalize the stock of data assets, per employee of the median firm and extrapolated to the whole economy. Section 5 validated pure output-based approach by comparing it with the more traditional input-based approach. Section 6 discusses future and work policy implications and concludes.

Related literature Many papers have explored the value of data barter for free digital goods (Nakamura et al., 2017; Brynjolfsson et al., 2025), modeled mechanisms through which data can affect output and welfare (Farboodi and Veldkamp, 2021; Asriyan and Kohlhas, 2024; Eeckhout and Veldkamp, 2023; Jones and Tonetti, 2020), and estimated the increase in firm value attributable to data and AI (Abis and Veldkamp, 2024; Eisfeldt et al., 2023). In contrast, this paper takes a national income accounting approach to simultaneously measure the value of data as a means of payment (data barter), as an asset of a firm that needs to be capitalized and that depreciates, as well as a means to increase firm productivity and output. The result is an improved understanding of the observed trends in output and profits, as well as a strategy to improve firm valuation and GDP measurement that closely follows current practice for tangible assets.

The most closely related papers preceding ours are as follows. Asriyan and Kohlhas (2024) 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, both in a model and in

data. Jones and Tonetti (2020) explore the optimal amount of data sales. At the end of our paper, we value data sales, which may be far from optimal. 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 data and purchased data, quantifying the value of data as a means of payment and capitalizing the data flow to value data as an asset.

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 (Crouzet et al., 2022; Eisfeldt et al., 2020) and proposed various methodologies to quantify the value of data, this paper introduces a novel approach: It recognizes that the data transferred from consumers to firms in the course of a transaction has value that should be counted as part of the transaction. This method contrasts with existing work because it acknowledges that not only is data a valuable asset, it is also a means of payment.

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 invest-

ments 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) provides a case study on Microsoft, demonstrating that the company's valuation is almost entirely driven by intangibles, likely including proprietary data.

2 A Framework for Data Measurement

At its heart, this is a forecasting model. It is not a production economy with equilibrium prices. It is simply a device to organize measurement of the relationship between forecasts and profits and to help us distinguish and value various types of information. Since firm interactions are not our object of study, we start by studying an individual firm i. Later, we discuss aggregation.

The production structure appears more complex that a standard production economy, because of the state θ_t appearing in output and costs, in linear and quadratic forms. Much of this is not essential for the model to function. Rather it makes the point that, even if data is used to predict various types of uncertain outcomes, we can still capture the value of these predictions through the relationship between revenue forecasts and profits.

Environment: 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

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

In each period t there is an uncertain "state", which we denote by θ_t, that captures characteristics of the good that all customers value at time t (such as design, availability, adaptability, etc). We can then define a *state-adjusted unit of the good* as $A_{i,t}$, which we assume

$$A_{i,t} = A_i(1 + b\theta_t) \tag{1}$$

A low state θ_t is akin to customers valuing a "smaller product."

The firm can charge each unit of the good either "in cash" or "in data", $P_{i,t} = P_{i,t}^{\$} + P_{i,t}^{D}$. The payment in cash is a standard monetary transaction. The payment in data is information that the firm obtains from selling the product and can be useful to improve forecasts but

not compensated monetarily. We discuss this decomposition of prices later. For now, the firm i's revenues can be written as

$$R_{i,t} = P_{i,t}A_{i,t}k_{i,t}^{\alpha} = P_{i,t}A_i(1 + b\theta_t)k_{i,t}^{\alpha}. \tag{2}$$

Since the model is for measurement, we want the model to have the flexibility for data to be used for multiple purposes. Therefore, we consider the possibility that the cost of production is also uncertain and data might be used to predict shocks that affect costs, such as inputs, production chains, etc. Cost also depends on the state θ_t.

We assume that the per-unit cost of production may depend on the price level, productivity, and a linear and quadratic term in the uncertain state θ_t:

$$c_{i,t} = P_{i,t}A_i(b\theta_t + \gamma(a_{i,t} - \theta_t)^2) + w \tag{3}$$

Note that this per-unit cost does not include the cost of capital rental. We count that separately below because it is accounted for differently, typically included as an overhead expense.

How might we interpret such a cost structure? The unit cost depends on the realized characteristics valued by customers, with a higher state θ_t generating more revenues but also more costly to produce (to fix ideas, think of the need to produce a larger product if θ_t is larger). Second, the cost depends on how well the firm targets such state. When firms take actions that are not well-aligned with the state θ_t, the costs of producing the adjusted unit of the good are higher (again, setting up to produce a small product may be costly if the right size of production is large). Finally, each unit of production represents a transaction that brings data into the firm. We denote by w the cost (in terms of data labor) to turn the data generated by a unit of production (and transaction) into information that is operational to reduce future forecast errors, by cleaning it, storing it, interpreting it, etc.

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

$$C_{i,t} = c_{i,t}k_{i,t}^{\alpha} + r_t k_{i,t}. \tag{4}$$

We assume that A_i, b, α, γ and w are known parameters. Since b is equivalent to normalizing units of variance and does not affect our estimates, we assume $b = 1$ henceforth. Finally, since we are interested in the price composition and not price levels, we normalize the price of the state-adjusted good by $P_{i,t} = 1$.

Multiplying (3) by $k_{i,t}^{\alpha}$ and subtracting from (2) yields the per-period, variable unit profits (before capital cost):

$$\pi_{i,t} = A_i(1 - \gamma(a_{i,t} - \theta_t)^2) - w$$

and total profits as

$$\Pi_{i,t} \equiv R_{i,t} - C_{i,t} = \pi_{i,t}k_{i,t}^{\alpha} - rk_{i,t} \tag{5}$$

Information sets In order for data to be useful for future periods, the state the data predicts must be persistent and not fully revealed at the end of each period. If either of these conditions fails, then data fully depreciates at the end of each period, which is not realistic or interesting for data valuation. Therefore, we assume that the state is a hidden Markov process (i.e. a Kalman filter system). It 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 a noisy realization θ_t of an AR(1) process $\widehat{\theta}_t$ with normally-distributed innovations:

$$\theta_t = \widehat{\theta}_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{6}$$

$$\widehat{\theta}_{t+1} \;=\; \rho\widehat{\theta}_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim N(0, \sigma_\epsilon^2), \tag{7}$$

for $0 < \rho < 1$. This structure implies that in each period t, the state is normally distributed: $\theta_t \sim N(0, \phi_t^{-1})$, where ϕ_t is a function of the parameters in (6) and (7).

We denote the firm's information set at time t as $\mathcal{I}_t$. This set contains prior information about the distribution of θ_t, new data generated from current transactions, old undepreciated data, and, potentially, data acquired from outside parties. These sources of information help the firm to forecast the state $\hat{\theta}_t$. In other words, when combined optimally with an algorithm, this data can be used as a signal s_t about θ_t.

2.1 Solution

Definition: A solution to the model is a sequence of state-contingent production plans, given by capital renting $k_{i,t}$, and actions $a_{i,t}$ every period to maximize:

$$\sum_{t=0}^{\infty}(1+r)^{-t}\mathbb{E}\left[(A_i(1-\gamma\left(a_{i,t}-\theta_t\right)^2)-w)k_{i,t}^{\alpha}-rk_{i,t}|\mathcal{I}_t\right], \tag{8}$$

where beliefs are formed using Bayes' law, in each period t, given information sets $\mathcal{I}_t$.

Since firms know their capital decision and other parameters, forecasting total revenues involves forecasting the state θ_t. First, we need to derive prior beliefs and variances for the dynamic model state using Bayes law. Given the linear revenue function, those are simply linear transforms of the mean and variance of θ_t.

Stock of Data and its Components: The prior mean and variance of the state, which determine the expectation and variance of profits per unit of production, are $E[\theta_t|\mathcal{I}_t] = E[\hat{\theta}_t|\mathcal{I}_t]$ and $Var[\theta_t|\mathcal{I}_t] = Var[\widehat{\theta}_t|\mathcal{I}_t] + \sigma_\eta^2$. The expected squared forecast error is, by definition, a conditional variance. We define the stock of data in period t as the inverse of the conditional variance, so

$$\Omega_t \equiv Var^{-1}(\widehat{\theta}_t|\mathcal{I}_t), \tag{9}$$

This definition lends itself to be interpreted as stock of data because a lower variance, or more accurate estimates, means a firm has more data about $\widehat{\theta}_t$. By the definition of the

conditional variance, the stock of data can be decomposed in the four main components previously discussed,

$$\Omega_t = \phi_t + \tau_t(k_{t-1}) + \psi_t + (1 - \zeta_{t-1})(\Omega_{t-1} - \phi_{t-1}) \tag{10}$$

The first component is the *prior precision*, ϕ_t, which captures the quality of public data about the state and its dynamics. The second component is the precision of new data obtained from the firm's period transactions, τ_t, which depends on the production scale hence the number of transactions, k_{t-1}. The third component is the precision of data purchased from external parties, ψ_t. The fourth component is the undepreciated past data stock. We assume in this specification that public data fully depreciates, while the private data (this is $\Omega_{t-1} - \phi_{t-1}$) depreciates at an endogenous rate ζ_{t-1}.

Data Depreciation: Despite the properties that capital shares with data that comes from a flow of transactions and is accumulated into a stock, we cannot assume an exogenous depreciation rate. To determine what is the depreciation of data, we first determine the forecast and variance of tomorrow's state, based on today's information set. Taking the mean and then the variance of both sides of (7) once combined with (6), we get

$$E[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho E[\hat{\theta}_t|\mathcal{I}_t], \tag{11}$$

$$Var[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] + \sigma_\epsilon^2. \tag{12}$$

This last conditional variance of the forecast of future state $\hat{\theta}_{t+1}$ is not a measure of the volatility of $\hat{\theta}_{t+1}$. Rather, it is a measure of uncertainty: the expected squared forecast error: $Var[\hat{\theta}_{t+1}|\mathcal{I}_t] := E[(\hat{\theta}_{t+1} - E[\hat{\theta}]_{t+1}|\mathcal{I}_t])^2|\mathcal{I}_t]$. In Bayesian terms, 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 all data sources 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, and back out θ_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.[2] Note that observing 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,

$$\Omega_{t+1} \equiv Var^{-1}(\hat{\theta}_{t+1}|\mathcal{I}_{t+1}) = \left[\rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] \ + \sigma_\epsilon^2\right]^{-1} \ + \sigma_\eta^{-2} + \sigma_s^{-2}. \tag{13}$$

But the firm's profits do not depend on the conditional variance, or expected squared forecast error about $\hat{\theta}_t$, but about θ_t,

$$Var(\theta_{t+1}|\mathcal{I}_{t+1}) = \left[\left[\rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] \ + \sigma_\epsilon^2\right]^{-1} \ + \sigma_\eta^{-2} + \sigma_s^{-2}\right]^{-1} + \sigma_\eta^2. \tag{14}$$

Rewriting equation (14) in terms of stock of data, we get

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

which maps the time-t stock of data Ω_t, into the time-$t+1$ stock of data, Ω_{t+1}. This 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

[2] If new data were correlated with old data, we could 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.

stock. This implies that $(1-\zeta_t)\Omega_t = \left[\rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] \ + \sigma_\epsilon^2\right]^{-1}$. Hence, the depreciation rate is

$$\zeta_t = 1 - (\rho^2 + \sigma_\epsilon^2 \Omega_t)^{-1}. \tag{15}$$

This shows that the properties of data endogenizes its depreciation rate. If the stock of data is large, data will depreciate more quickly.

2.2 A Recursive Representation

Conditional on the amount of data, Ω_{it}, the firm chooses the best production strategy, a_{it} and the optimal scale of production k_{it}. It is clear that the action that maximizes the objective is $a_{i,t} = E(\theta_t|\mathcal{I}_t)$ each period. Thus, the objective consists of minimizing the expected squared forecast errors $(E(\theta_t|\mathcal{I}_t) - \theta_t)^2$ about a state θ_t. Since the expected squared forecast error is, by definition, a conditional variance, and we have defined data stock as the inverse of the conditional variance, we can write the *expected per-period profit* from the objective (8) as an explicit function of data stock as $\left[(A_i(1-\gamma\Omega_{it}^{-1}) - w)k_{i,t}^{\alpha} - rk_{i,t}|\mathcal{I}_t\right]$.

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

$$\mathbb{V}_i(\Omega_{i,t}^{-1}) = \max_{k_{i,t}} \Big\{ \underbrace{\left(A_i(1-\gamma\,\Omega_{i,t}^{-1}) - w\right) k_{i,t}^{\alpha}}_{\equiv\ \pi_{i,t}(k_{i,t},\Omega_{i,t}^{-1})} - r\,k_{i,t} + \beta\,\mathbb{E}_t\big[\,\mathbb{V}_i(\Omega_{i,t+1}^{-1})\,\big] \Big\}. \tag{16}$$

subject to equation (10) at t

$$\Omega_{i,t+1} = \phi_{t+1} + \tau_{i,t+1}(k_{i,t}) + \psi_{i,t} + (1 - \zeta_{i,t}(\Omega_{i,t}))(\Omega_{i,t} - \phi_t) \tag{17}$$

where $\tau_{i,t+1}(k_{i,t})$ captures in general how new data is generated from the scale and operation of the firm, and $\zeta_{i,t}(\Omega_{i,t})$ is from equation (15). We characterize next the solution of this general setting and discuss its properties. In the next sections we calibrate it and solve it numerically. We solve for the deterministic steady state in Appendix A, and show it can be

solved in closed-form when shutting down the reverse causality that $\tau_{i,t+1}(k_{i,t})$ generates.

Optimal Choices of Data and Scale By the envelope theorem applied to (16), fixing the value of $k_{i,t}$ at the argmax value.

$$\mathbb{V}'(\Omega_{i,t}^{-1}) = \frac{\partial \pi_{i,t}}{\partial \Omega_{i,t}^{-1}} + \beta\, \mathbb{E}_t \left[\mathbb{V}'(\Omega_{i,t+1}^{-1})\, \frac{\partial \Omega_{i,t+1}^{-1}}{\partial \Omega_{i,t}^{-1}} \right]. \tag{18}$$

From the period payoffs,

$$\frac{\partial \pi_{i,t}}{\partial \Omega_{i,t}^{-1}} = -\,\gamma\, A_i\, k_{i,t}^{\alpha}.$$

Using the chain rule to determine the derivative of precision over time

$$\frac{\partial \Omega_{i,t+1}^{-1}}{\partial \Omega_{i,t}^{-1}} = \frac{\partial \Omega_{i,t+1}^{-1}}{\partial \Omega_{i,t+1}} \cdot \frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} \cdot \frac{\partial \Omega_{i,t}}{\partial \Omega_{i,t}^{-1}} = \left(-\,\Omega_{i,t+1}^{-2} \right) \cdot \frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} \cdot \left(-\,\Omega_{i,t}^{2} \right) = \frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} \left(\frac{\Omega_{i,t}}{\Omega_{i,t+1}} \right)^2.$$

Differentiate (17) with respect to $\Omega_{i,t}$:

$$\frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} = 1 - \zeta_{i,t}(\Omega_{i,t}) - \left(\Omega_{i,t} - \phi_{i,t} \right) \zeta_{i,t}'(\Omega_{i,t}) \quad \equiv \quad \Gamma_{i,t}(\Omega_{i,t}), \tag{19}$$

so that

$$\frac{\partial \Omega_{i,t+1}^{-1}}{\partial \Omega_{i,t}^{-1}} = \Gamma_{i,t} \left(\frac{\Omega_{i,t}}{\Omega_{i,t+1}} \right)^2.$$

Hence, we can rewrite the envelope condition in (18) as,

$$\mathbb{V}'(\Omega_{i,t}^{-1}) = -\,\gamma\, A_i\, k_{i,t}^{\alpha} + \beta\, \mathbb{E}_t \left[\mathbb{V}'(\Omega_{i,t+1}^{-1})\, \Gamma_{i,t} \left(\frac{\Omega_{i,t}}{\Omega_{i,t+1}} \right)^2 \right]. \tag{20}$$

This condition determines the evolution of the value function given the argmax $k_{i,t}$. It shows that more data affect value both statically, by improving production by reducing forecast errors, but also dynamically, by the extent to which some of that data remains in the future

and improve future forecasts and profits. Notice that full depreciation of data would imply $\Gamma_{i,t} = 0$ from equation (19), and no dynamic benefit of data.

Now we can obtain the FOC that determines $k_{i,t}$

$$\frac{\partial \pi_{i,t}}{\partial k_{i,t}} - r + \beta\, \mathbb{E}_t \left[\mathbb{V}'(\Omega_{i,t+1}^{-1})\, \frac{\partial \Omega_{i,t+1}^{-1}}{\partial k_{i,t}} \right] = 0. \tag{21}$$

where

$$\frac{\partial \pi_{i,t}}{\partial k_{i,t}} = \alpha \Big(A_i \big(1 - \gamma\, \Omega_{i,t}^{-1}\big) - w \Big)\, k_{i,t}^{\alpha-1},$$
$$\frac{\partial \Omega_{i,t+1}^{-1}}{\partial k_{i,t}} = \frac{\partial \Omega_{i,t+1}^{-1}}{\partial \Omega_{i,t+1}} \cdot \frac{\partial \Omega_{i,t+1}}{\partial k_{i,t}} = \big(-\Omega_{i,t+1}^{-2} \big)\, \tau'_{i,t+1}(k_{i,t}).$$

Hence, the general Euler equation (21) can be written as:

$$\alpha \big(A_i \big(1 - \gamma\, \Omega_{i,t}^{-1}\big) - w \big)\, k_{i,t}^{\alpha-1} - r \;-\; \beta\, \mathbb{E}_t \big[\, \mathbb{V}'(\Omega_{i,t+1}^{-1})\, \Omega_{i,t+1}^{-2}\, \tau'_{i,t+1}(k_{i,t}) \big] = 0. \tag{22}$$

This last equation constitutes a generalized Euler equation. The scale of production, $k_{i,t}$ not only equalizes the marginal product of capital with its cost, but also considers the future value of capital in generating data and reducing forecast errors that turn into future values (the third term).

3 From Model to (Mis)measurement

Even though we normalize the total price of the good to one, its composition will be endogenous and determined by the value of data that the firm obtains from selling a unit of good. If selling the unit generates data that is very valuable to reduce forecast errors and target good production more effectively, the data value $P_{i,t}^D$ would be high, and the firm would

charge less in cash, this is

$$P_{i,t}^{\$} = 1 - P_{i,t}^{D}$$

This construct assumes that customers can fully capture the value of the data the transaction they initiate provides. We will entertain later the possibility that the firm does not fully compensate consumers for data. At the other extreme, for instance, if the firm does not compensate customers from the value of data they provide, the firm is able to charge fully for its good, i.e., $P_{i,t}^{\$} = 1$. In our benchmark we maintain the assumption that firms do not have market power and compensate customers fully for their data. Later we will discuss alternative realistic assumptions on the firm's market power.

3.1 GDP Accounting

How does data value in the model correspond to measured macro aggregates? How does it show up in the mismeasurement of GDP?

3.1.1 Expenditure Approach

Under the expenditure approach, GDP in period t is defined as the sum of all expenditures:

$$\text{GDP}_{t|\text{exp}} = C_t + I_t + G_t + NX_t,$$

where C_t is total Consumption, I_t is Investment, G_t and $NX_t = X_t - M_t$ are Government Expenditures and Net Export, respectively. These figures are measured by their market monetary prices. Our economy produces final goods and data (a type of capital not explicitly measured). Define real output and the value of data as

$$Y_t = \sum_i \pi_{i,t} k_{i,t}^{\alpha} \tag{23}$$

$$D_t = \sum_i P^D_{i,t} \pi_{i,t} k^{\alpha}_{i,t} \tag{24}$$

Consider the following thought experiment: Some food, which is worth \$1, is exchanged for data, which is also worth \$1, as a pure barter trade. Imagine both are final goods. No money changes hands. What should this exchange add to GDP? If the food were sold first and the buyer paid \$1 and then the data were sold as a separate transaction, then the total price times quantity of both transactions would be \$2. That is the total value exchanged. Since the sequencing or purchases should not matter, the barter trade, ideally, should also be counted as twice the value of the good exchanged.

This is, however, not how GDP is measured in practice, but instead, using the monetary prices of final goods. The only good exchanged in the economy is the final good, for a market value of $\sum_i P^{\$}_{i,t} \pi_{i,t} k^{\alpha}_{i,t} = Y_t - D_t$. Investment on data, on the other hand, is not recorded as a transaction for its full value D_t. Hence the mismeasured value of data in GDP is $2D_t$. If customers were not compensated for the value of the data they provide, the value of the good would correctly be measured by its monetary value, and GDP would be mismeasured only by missing data investment, D_t. The correct formulation of GDP *inclusive of data*, is:

$$\text{GDP}^{\text{corrected}}_{t|\text{exp}} = (C_t + D_t) + (I_t + D_t) + G_t + NX_t,$$

In few words, GDP is undermeasured for $2D_t$ (value of generated data), and for two motives. One is *consumption bartering*, under which the monetary value of consumption is D_t less than the true value. The other is *intangible investment*, as the investment in data with value D_t is not recorded as generated in the economy. This is an important distinction. If consumers were not compensated by data (no bartering), then only one D_t would be mismeasured, as standard with intangibles.

Adjusting for market power: The formulation above assumes that to obtain data worth D_t, the firm must give a discount of D_t to consumers. However, in the presence of

market power, the firm may be able to acquire the same data while offering a smaller discount. Suppose a firm charges a markup μ over marginal cost when selling goods, i.e., $P = \mu MC$, where P is price, MC is marginal cost and $\mu \in [1, \infty]$. The markup of $\mu = 1$ is the case of perfect competition (our benchmark) and higher values of μ reflect greater market power. We assume that the firm can obtain data at the same markdown — specifically, by giving a discount of D_t/μ instead of D_t. Incorporating this into the expenditure-side measurement, the correct GDP inclusive of data becomes

$$\text{GDP}^{\text{corrected}}_{t|\text{exp}} = \left(C_t + \frac{D_t}{\mu}\right) + (I_t + D_t) + G_t + NX_t.$$

Note that only the consumption component is adjusted because the firm still receives data worth D_t; market power affects only how much discount the firm must provide to acquire it, not the actual value of the data it received. To see this clearly, let's return to the food example discussed above. Without market power, the firm trades \$1 worth of discount for \$1 worth of data. With market power, the firm obtains \$1 worth of data by discounting the price by only $1/\mu$, so the consumer pays $1 - 1/\mu$. The amount of GDP that goes unmeasured is therefore

$$2 - \left(1 - \frac{1}{\mu}\right) = 1 + \frac{1}{\mu}.$$

In general, when the value of data is D_t, the under-measured amount of GDP, after accounting for market power, equals

$$\left(1 + \frac{1}{\mu}\right) D_t.$$

When $\mu = 1$, the firm has to fully compensate consumers for their data, and the missing value of data corresponds to both intangible capital and consumption bartering. At the other extreme, when $\mu = \infty$, the firm does not compensate consumers for their data, and the missing value of data corresponds to just intangible capital.

3.1.2 Income Approach

Under the traditional income approach, GDP is defined as the sum of all factor payments plus allowances for capital consumption:

$$\text{GDP}_{t|\text{inc}} = w_t L_t + r_t K_t + \Pi_t + \delta_{t,K} K_t,$$

where $w_t L_t$ denotes labor compensation (both on labor in firms not modeled here, and to data workers explicit in our setting)[3], $r_t K_t$ rental return to reproducible capital, Π_t pre-depreciation corporate profits, and $\delta_{t,K} K_t$ depreciation of fixed capital. Because firms report profits net of depreciation, adding back $\delta_{t,K} K_t$ simply yields gross profits.

The correct formulation of GDP using the income approach, *inclusive of data*, is:

$$\text{GDP}^{\text{corrected}}_{t|\text{inc}} = w_t L_t + r_t K_t + D_t + (\Pi_t + D_t - \delta_{t,D} DS_{t-1}) + \delta_{t,K} K_t + \delta_{t,D} DS_{t-1},$$

where $\delta_{t,D} DS_{t-1}$ is the depreciation of total data stock in $t-1$. Notice that since data depreciation is both added and subtracted, it does not contribute to mismeasurement.

The first time the value of data D_t enters into a correctly measured GDP is through the correct value of the firms. The true profit of firms are adjusted for two reasons. First, they should be larger because of bartering firms fail to record the value of their sales are indeed higher by D_t. Second, firms should deduct the depreciation of their data. The second time D_t enters is as payments for data to consumers, who generate data by transacting and sell that data to firms.

One might ask: If the firm needs to pay customers for data in the unbundled transaction, why isn't this cost of data purchase subtracted from their profits? Think of this as if this is a firm buying a physical asset. In the income approach to GDP, the purchase cost of an

[3]What we denote w_t is the wage per hour of labor in the data, and the *payment to data labor per unit of output* in the model. We ensure that the total labor payments are compatible when we calibrate the model.

asset is not subtracted from revenues. Instead, it is amortized, meaning that its depreciation value is subtracted from gross profits in the future to arrive at net profits and then added back. The idea is that purchasing an asset is not a loss of value. It is a transfer of a valuable asset from one party to another, in a way that does not affect GDP. Depreciation makes a firm less valuable. But it does not imply that less value of goods and services was produced. Thus, neither show up in the national income accounting approach to GDP.

As with the expenditure approach, GDP is undermeasured for $2D_t$ (value of generated data), and also for two motives. One is *consumption bartering*, under which firms values are underestimated. The other is *intangible investment*, as consumers should have been compensated for selling their generated data. This is that important distinction arising again: If consumers were not compensated by data (no bartering), then only one D_t would be mismeasured, as standard with intangibles.

Similar arguments as in the expenditure approach suggest that with market power, the under-measured value of GDP would be $(1 + 1/\mu)D_t$.

3.2 Discussions

Firms' Valuations The previous discussion shows clearly that also firms values are mismeasured. However, the extent of bartering is critical in determining whether firms are over- or under- valued. When there is bartering, firms' values are undermeasured by $D_t - \delta_{t,D} DS_{t-1}$, where D_t is the value of new data and DS_{t-1} the value of data stock in the previous period. This is an issue only outside the steady state, since in steady state $D_t = \delta_{t,D} DS_{t-1}$. Firms are undervalued when data grows, and overvalued when it declines. If there is no bartering, however, firms are always overvalued, since firms report more profits than what they should if also deducting the depreciation of data $\delta_{t,D} DS_{t-1}$. In other words, as long as the firm monetize the value of data (by capturing their value from consumers without compensation), profits are overvalued because they should also report data depreciation.

Changing Data Technology There are two types of technological change with regards to data. One collects more data points or extracts more data precision and the other makes the same amount of forecast error more valuable. Both are captured by our measurement strategy. When technology enables more data collection, more data refinement, or more precise forecasting from the same data, as is the case with AI, then this all shows up as a lower forecast error. We interpret this as more data. Even if the number of data points is the same and more value is being extracted from it, counting it as more data captures the additional revenue this data generates for the firm. Data should be valued more highly when it leads to more accurate forecasts and our method captures the increase in forecast accuracy. When technological progress allows firms to profit more from a given forecast accuracy, this is captured by an increase in the coefficient in the estimation of the relationship between forecast error variance and profits.

4 Quantifying the Sources and Dynamics of Firms' Data

To measure the value of data we need to tackle several challenges. First, to measure the stock of data available for a representative firm in different periods. Second, to decompose such stock between public data, new private data, purchased data and undepreciated old private data. Third, to translate the amount of data, its evolution and its components into dollars, to determine GDP mismeasurements. This last task is contaminated by a reverse causality problem. Is more data and better forecasts what leads to better outcomes, or more production and better outcomes is what lead to more data?

We proceed in three steps. First, we describe the data we use. Then, we use the evolution of forecast errors to infer how the stock of data has changed for a median firm in our data. From the process of fundamentals, we back out the public component, and from the endogenous computation of depreciation from (15), we separate between new private data and undepreciated private data. In this exercise we abstract from purchased data. In

Appendix E we propose a strategy to assess its importance, and show its precision is below that of public data and private data that we measure here. Once we have the evolution of data stock and its components, we proceed to value it. We calibrate our model to the heterogeneity across firms and to the relation between forecast errors and revenues to estimate the value of data for a public firm, controlling for the reverse causality problem. Finally, we extrapolate this estimation to the whole economy.

4.1 Data Description

Our data is based on US public firms. In terms of firms' fundamentals, we construct use Compustat North-America for the years 2003-2022. Following the convention in finance literature, we exclude utilities and financial firms from the data. We also exclude the firm-years with missing and negative values of revenue, employment, and capital. 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 2002 dollars.

To compute forecast errors, a critical input to assess the stock of data, we use the data from I/B/E/S Guidance and compute the sales forecast accuracy of US public firms. 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 February 19, 2025. We use the "Detail" table which contains firms' generated financial forecasts (e.g., annual or quarterly sales or earnings). We keep only US firms (observations with "usfirm"=1), available from 2002 to 2021. We focus our attention on the forecasts of annual revenue as these constitute the biggest share of annual forecasts made by the firms.[4]

[4]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 use the forecast made by a firm in the first quarter of the financial year, which is not affected by subsequent realized quarterly data. When the sales forecast is expressed as a range, we take the mid-point.

After merging with Compustat data and retaining only firm-year observations for which we have one year ahead revenue forecasts, we are left with 12,525 firm-year observations. We compare firms' sales guidance with realized sales data (retrieved from Compustat North America) and define the squared relative forecast error as in Equation (25). We winsorize the relative forecast error at the one percent level on the right tail.

4.2 Stock of Data: Dynamics and Components

We have defined the stock of data as the inverse of the conditional variance, which can be measured as the *expected squared forecast errors*. Using firms' forecasts of future sales revenue, and the realized sales, we compute squared forecast errors, relative to total sales,

$$\text{Squared Relative Forecast Error} = \left| \frac{\text{Forecasted Sales} - \text{Actual Sales}}{(\text{Forecasted Sales} + \text{Actual Sales})/2} \right|^2. \tag{25}$$

The median squared relative forecast error in the whole sample is $\sigma^2_{post} = 0.0023$. The inverse constitutes our measure of *data stock*, which is $\bar{\Omega} = 437$ on average for the whole sample. The dynamics of this measure is plotted in blue in Figure 1.

To decompose the *public data component*, we estimate the prior variance σ^2_ϵ 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, we construct a proxy for the state: $\theta = \frac{\text{Revenue}}{\text{Total Assets}}$ i.e. revenue normalized by total assets of the firm. We then estimate the following regression:

$$\theta_t = \mu + \rho\theta_{t-1} + \epsilon_t$$

Note that as many firms do not provide guidance every year, they may not appear in our sample every year. However, we can still calculate the variable (θ_{t-1}) as long as we can obtain the revenue and total assets estimate for the previous year from the Compustat data. Next, to calculate the prior variance, we calculate the predicted value of our state variable

using the estimated AR-1 coefficients. Denote the predicted state as $\hat{\theta}_t$. We calculate the prior forecast error FE_{prior} in a similar way as FE above

$$FE_{prior} = \left| \frac{\hat{\theta} - \theta}{(\hat{\theta} + \theta)/2} \right|^2$$

The median prior squared relative error in our sample is $\sigma_\epsilon^2 = 0.007$. The inverse constitute our measure of *the average stock of public data*, and it is in average $\bar{\phi} = 143$. The dynamics of this measure is plotted in orange in Figure 1.

This analysis immediately implies that the amount of private data, both new and undepreciated old data, is the residual. Hence *the average stock of private data* is $\bar{\Omega} - \bar{\phi} = 294$. This is a first insight: *The precision of private data is twice the precision of public information based on prior realizations.* We decompose now private data between new data and undepreciated old data.

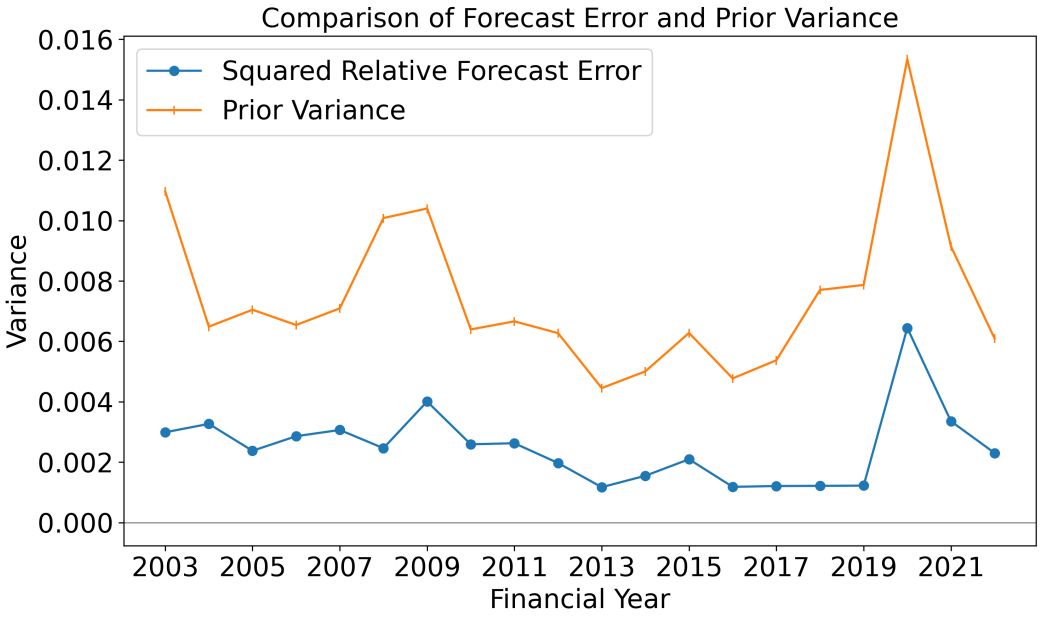


Figure 1: Forecast Errors and Precision of Own Data.

By applying the recursive structure of data stock (10) and the endogenous depreciation rate (15), we can obtain the evolution of the different sources of data precision. The details

of this construction are in Appendix C.

The key question in this dynamic decomposition is how quickly 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, we will derive the depreciation rate from our model and estimate it with firm revenue data. The model implies that data obtained in one period has value for multiple future periods, as long as $\rho > 0$. We find that the persistence coefficient is estimated to be $\rho = 0.95$, which we can now use in equation (15) to back out yearly depreciation.

Figure 2 decomposes the data stock into components that come from prior knowledge, which we call public data ϕ_t, newly obtained data from own operations τ_t and the depreciated data stock carried over from the previous period. We find that the majority of the precision comes from the new data inflows. Notice that two years stand out. First it seems there is a structural change around 2012. At that point the stock of data roughly doubled. This period coincides with the advent of cloud computing, big data, and mobile dominance. The year 2020 also looks quite different. This reflects the higher uncertainty in the COVID-19 period. It shows up as almost no new data flow. The fact that earnings variance in this period was high shows up as a lower value of prior information or public data. One might think of this period as one with a high depreciation shock to previous knowledge as the business environment changed rapidly.

4.3 Value of Data

To capture the mismeasurement of GDP we need to assign monetary value to the present discounted value of any data generated in a given year, hence the relevance of identifying the main components of the stock of data, as we did in the previous section. This involves expressing the marginal value of an improvement in forecast errors that come from new data, correcting for reverse causality. To make progress on this task, we calibrate the model. Here

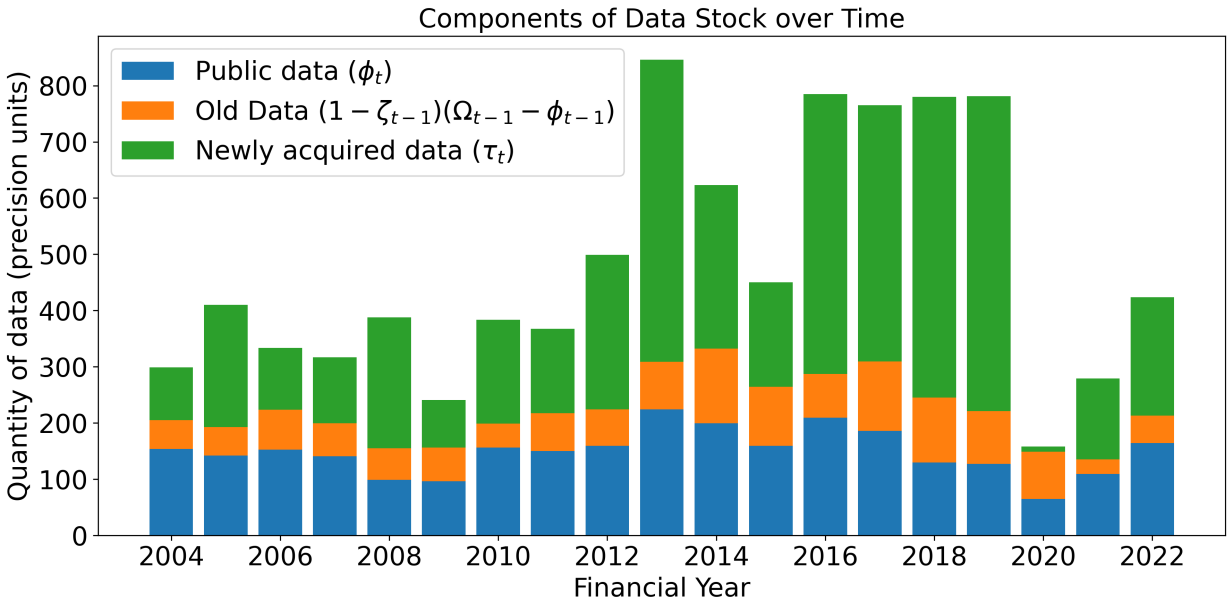


Figure 2: Data Precision: Public, Old and New Data (median firm).

we explain the main elements and results. The full simulation procedure, solution method, and robustness checks are provided in Appendix B.

Model Implementation Each firm i is characterized by a permanent productivity draw A_i, a stock of data precision $\Omega_{i,t}^{-1}$, and a choice of capital $k_{i,t}$. Forecasted revenues are,

$$\text{Forecasted Revenue}_{i,t} = \Big[(1-\gamma\Omega_{i,t}^{-1})A_i - w\Big]k_{i,t}^{\alpha},$$

while realized revenues incorporate forecast error shocks,

$$\text{Actual Revenue}_{i,t} = \Big[(1-\gamma\psi_{i,t}^{2})A_i - w\Big]k_{i,t}^{\alpha},$$

with $\psi_{i,t} \sim \mathcal{N}(0, \Omega_{i,t}^{-1})$. In the model we can measure forecast errors as in the data (25).

The stock of data evolves deterministically according to equation (17), and we have to take a stand on the functional form of $\tau_{i,t+1}(k_{i,t})$: how new data depends on the previous period scale of production. Since we calibrate the model to the median firm in the sample,

we assume that

$$\tau_{i,t+1}(k_{i,t}) = \tau_{t+1} + \chi \ln\left(\frac{\text{Rev}_{i,t}(k_{i,t})}{\text{Rev}_{M,t}}\right) \tag{26}$$

where $\text{Rev}_{i,t}(k_{i,t}) = [(1-\gamma\psi_{i,t}^2)A_i - w]k_{i,t}^{\alpha}$ is Revenue of firm i, τ_{t+1} is the baseline new data obtained by the median firm (with revenue $\text{Rev}_{M,t}$), and χ captures the slope: how data change with changes in scale relative the median firm.

We obtain the policy functions that solve (16) using value function iteration (VFI) on discretized grids for productivity, precision, and capital. The productivity grid is constructed from quantiles of a log-normal distribution, the precision grid spans a geometric sequence covering the feasible range implied by the law of motion, and the capital grid spans a geometric sequence covering the feasible range implied by the optimal choice of capital from the Bellman equation (16) for all reasonable combinations of productivity and precision.

Calibration Strategy To calibrate the model, we *i*) fix some parameters outside the model and *ii*) simulate a panel of 1000 firms for 100 time periods and match model's moments with data moments.

First, on parameters that we calibrate through external measures, public data ϕ is set to $\bar{\phi} = 143$ to match prior precision, and baseline private data τ is set to 220, which jointly with ϕ and depreciation matches the posterior precision of the median firm in steady state. Notice this is consistent with our previously discussed private data in average, both new and undepreciated. Such estimation was $\bar{\Omega} - \bar{\phi} = 295$, with new data τ coming from the decomposition in Figure 2 of around 75% in average. The persistence of the state variable, $\rho = 0.95$, is estimated from an AR(1) regression of revenues normalized by assets. The capital share is set to $\alpha = 1/3$ and the rental rate of capital to $r = 0.04$, consistent with standard values in the literature (Table 1).

Second, on parameters that we calibrate through simulating the model, each firm draws a permanent productivity A_i from a log-normal distribution with mean $\mu_{\ln A}$ and standard

Table 1: Parameters Calibrated Outside the Simulation

Parameter	**Value**	**Strategy**
ϕ	143	Match Prior Precision
τ	220	Match Posterior Precision
ρ	0.95	Match Persistence of State in Data
α	1/3	Standard Capital Share in Literature
r	0.04	Standard Rental Rate of Capital

deviation $\sigma_{\ln A}$. The panel is then generated period by period: firms optimally choose capital based on the policy function, forecast errors are computed, revenue shocks are realized (implying forecast errors), and precision is updated according to the law of motion of data. We calibrate $\mu_{\ln A}$, $\sigma_{\ln A}$, γ, χ, and w to minimize the sum of squared percentage deviations between targeted empirical moments and the corresponding simulated moments.

Although all five parameters ($\mu_{\ln A}$, $\sigma_{\ln A}$, γ, χ, w) are calibrated jointly to match the five targeted moments, different moments load more heavily on specific parameters. The distributional moments of log revenue (mean 6.63 and standard deviation 1.72) are most sensitive to $\mu_{\ln A}$ and $\sigma_{\ln A}$. The cost parameter w is primarily disciplined by the empirical share of revenues paid to data workers for the median firm—0.16 percent, based on Occupational Employment and Wage Statistics (OEWS) data from the U.S. Bureau of Labor Statistics.

To obtain the moments that discipline γ and χ, we simulate a panel of 1000 firms for 100 time periods and run the following regressions on both simulated and empirical data:

1. To discipline how forecast errors translate into revenues (γ), we run the following *Revenue regression:*

$$\ln(Rev_{i,t}) = \beta_0 + \beta_1 \ln(FE_{i,t}) + \beta_2 \ln(emp_{i,t}) + \epsilon_{i,t}.$$

where $\ln(emp)$ is the log of number of employees which serves as a proxy for firm-size in the data. As the model does not have total employment explicitly, we proxy for firm-size in the simulated data by using $\ln(A_i)$. Table 2 shows the results, Column

(1) without any controls and Column (2) controlling for firm-size (log (number of workers)). Column (3) shows the results with Industry and Year fixed effects while also controlling for firm-size.

For our calibration, we use the result from column (3), which shows that a 1% decrease in Forecast errors is associated with a 0.016% increase in revenue. This is sizable. Given our estimate of 0.007 for prior variance and 0.002 for the posterior variance, a decline in forecast errors by 1% is akin to a decline of 71% from prior to posterior. Hence, the log revenue for the median firm goes up by approximately $0.016 \times \ln(1 - 0.71) \approx 0.02$, hence approximately 2%.

Table 2: Effect of Forecast Error on Firm Sales

	(1) No Controls	(2) Controlling for Employment	(3) Industry, Year FE + Employment
ln(FE)	-0.145***	-0.012***	-0.016***
	(0.006)	(0.003)	(0.004)
ln(Emp)		0.884***	0.925***
		(0.004)	(0.020)
Observations	12521	12521	12420
Fixed effects	None	None	Industry, Year
SE clustered at	None	None	Industry & Year (two-way)

Standard errors in parentheses

Standard Errors in columns (1) and (2) are heteroskedasticity-consistent (White 1980)

Standard Errors in column (3) clustered by year and industry identifier following Cameron, Gelbach, and Miller (2011).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2. To discipline how revenues translate into new data available next period (captured by χ), we run the following *Forecast error regression:*

$$\ln(FE_{i,t}) = \beta_0 + \beta_1 \ln(Rev_{i,t-1}) + \epsilon_{i,t}.$$

Table 3 shows the results. Column (1) does not use any controls whereas column (2) controls for industry and year fixed effects. For our calibration, we use the result

from column (2) which shows that a 1% increase in revenue is associated with a 0.41% decrease in forecast errors. The standard deviation of the lagged $\log(Rev)$ in our data is 1.77. Therefore, a one-standard-deviation increase in $\log(Rev)$ is associated with a change of 0.73 in $\log(FE)$. Since the standard deviation of $\log(FE)$ is 2.7, a one-standard-deviation increase in $\log(Rev)$ is associated with a 0.27 standard deviation decrease in $\log(FE)$ in the next year.

Table 3: Forecast Error vs. Lagged Revenue

	(1) No Controls	(2) Industry, Year FE
ln(Rev(t-1))	-0.338***	-0.406***
	(0.013)	(0.031)
Observations	12511	12412
Fixed effects	None	Industry, Year
SE clustered at	None	Industry & Year (two-way)

Standard errors in parentheses
Standard Errors in column (1) are heteroskedasticity-consistent (White 1980)
Standard Errors in column (2) are clustered by year and industry identifier following Cameron, Gelbach, and Miller (2011).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4 shows the fit of our calibration relative to the empirical moments. Table 5 reports the corresponding parameter values.

Table 4: Calibration Results (Simulation): Targeted vs. Model Moments

Moment	**Target**	**Model**
Mean log revenue	6.63	6.63
SD log revenue	1.72	1.72
$\beta_{\ln(FE)}$ in ln Rev	-0.016	-0.015
$\beta_{\log(\text{Rev}_{t-1})}$ in ln FE	-0.41	-0.38
Fraction of revenue paid to data workers	0.0016	0.0016

Given our previous estimation of the stock of data, and its components, for the median firm over 2003-2022, we can now use this calibration, which maps forecast errors into revenues controlling for reverse causality, to value such stock of data for the median firm in our sample.

Table 5: Parameter Values calibrated through simulation

Parameter	Value
$\mu_{\ln A}$	3.75
$\sigma_{\ln A}$	1.15
γ	4.3
χ	120
w	0.06

Normalizing by firm size, we obtain an estimate of the value of data per employee for the median firm. Figure 3 shows how this value has evolved over the past two decades. Despite the growth of data in the economy, the value of data per employee remains stable for most years around $2,000 a year, and shows a slight upward trend only after 2014.

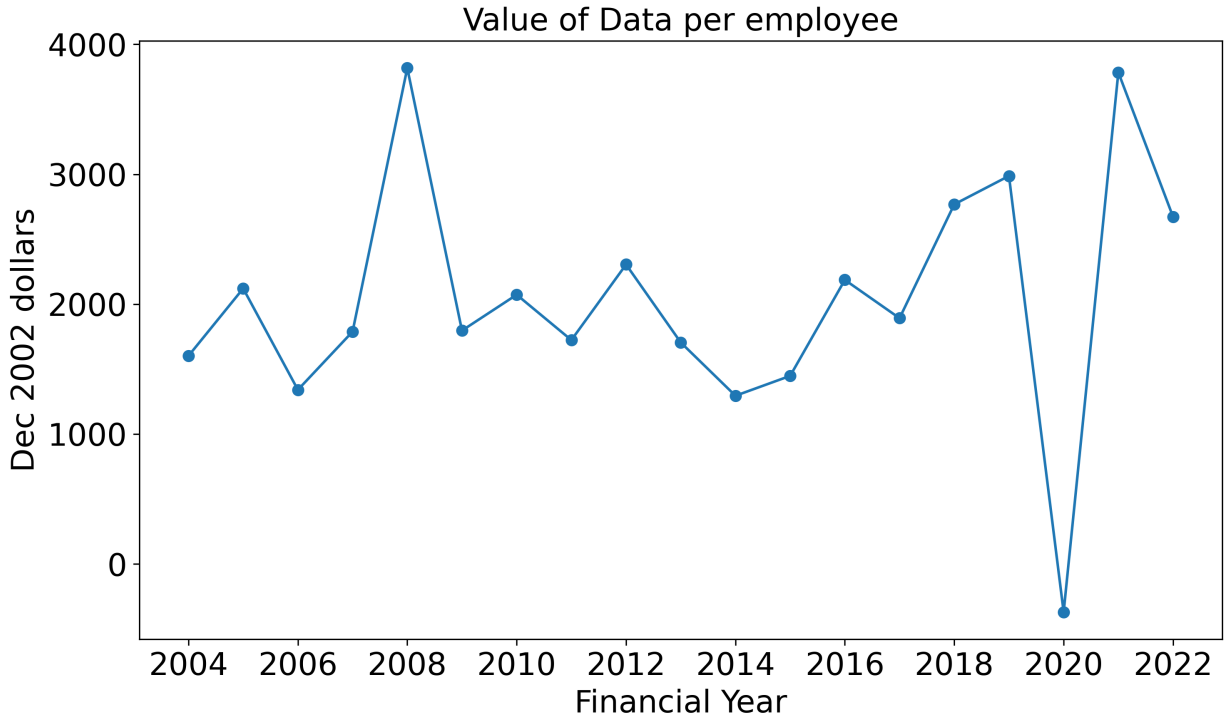


Figure 3: Estimate of the value of data per employee for the median firm in our sample

4.4 Scaling Up: From the Median firm to GDP

Here we scale up the model-implied value of data for the median firm to the whole economy, and obtain three estimates of the associated GDP mismeasurement. We provide details in

Appendix D.

The first projection is a clear lower bound. It assumes that only the firms in our sample possess valuable data. Under this assumption, missing GDP equals $2D_t$ for each firm, where D_t is the value of data in period t, and as explained it duplicated given a double mismeasurement, as intangible capital and as bartered consumption. We therefore take twice the value of data per employee for the median firm in our sample and multiply it by the total employment of all *firms in the sample*. Under this approach, the implied missing GDP averages about 0.3% of measured GDP.

The second projection uses the same logic but scales the calculation to the entire economy, assuming all firms obtain and value data in the same way as public listed firms. Specifically, we take twice the value of data per employee for the median firm and multiply it by total U.S. employment. Under this approach, the implied missing GDP averages about 4% of measured GDP, rising from 2014 to 2022 to about 6%.

Both of these projections assume that firms must fully compensate consumers for the data they obtain. However, if firms have market power, they may only need to offer a discount of D_t/μ rather than D_t to acquire data worth D_t. In this case, the amount of GDP missing from national accounts reduces to $(1 + 1/\mu)D_t$, given the lower impact of missing bartering. We set $\mu = 1.26$, following Edmond et al. (2023), who show that after 1990 the cost-weighted average markup for U.S. public firms remained within the range 1.2–1.26. Because our sample covers 2003–2022, this estimate applies directly to our setting. This yields our third projection, which adjusts the missing GDP estimate for market power, and reduces the estimated missing GDP merely by 0.5% in average.

The estimates of missing GDP from all three methods are plotted in Figure 4.

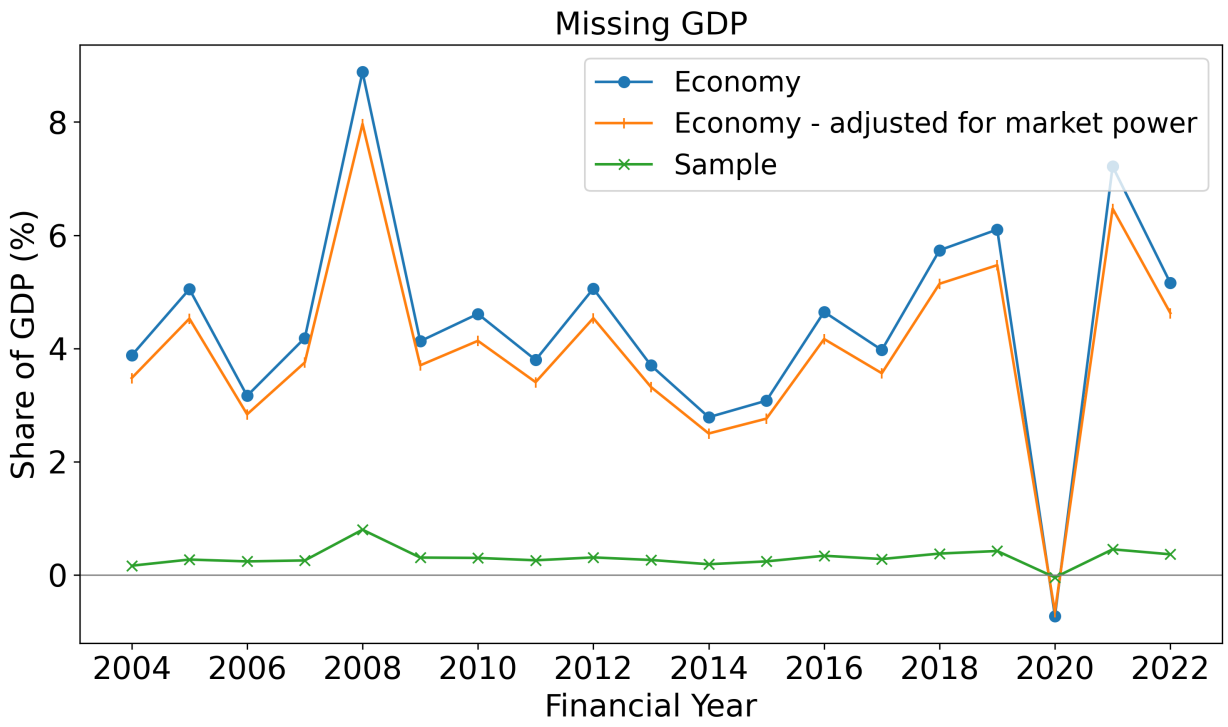


Figure 4: Estimates of missing data GDP measure with three methods.
The first series scales the median firm's data value to the employment in our sample, the second extends this scaling to the entire U.S. workforce, and the third adjusts the implied missing GDP for market power using $(1 + 1/\mu)D_t$ with $\mu = 1.26$.

5 Validation of Forecast-Based Approach

In this paper we value data in the economy based on an *output approach*: we explore the evolution and impact of forecast errors on outcomes, the output of data. To validate our exercise here we explore the relation between forecast errors and the labor that in principle makes data valuable, such as data workers: an *input approach*. This is the most explored approach, and examples include Abis and Veldkamp (2024), Crouzet et al. (2022), and Calderón and Rassier (2022).

To obtain a proxy for the number of workers working with data, we use the Occupational Employment and Wage Statistics (OEWS) program of the U.S. Bureau of Labor Statistics (BLS). The OEWS reports employment shares and wages for detailed occupations within each NAICS industry, with consistent coverage starting in 2012. We focus on the occupa-

tional category Database and Network Administrators and Architects[5] as our proxy for Data Workers. For each firm–year observation, we assign the employment share of Data Workers in the firm's four-digit NAICS industry, defined as the percentage of total workers in that industry classified as Data workers. If four-digit industry data are unavailable, we use the corresponding three-digit industry average; if still unavailable, we fall back to the two-digit industry. Observations that remain unmatched are assigned an employment share of zero.[6]

Using this proxy, we regress our measure of squared forecast errors (FE) on the employment share of Data workers (DataEmp) using the following specification:

$$FE_{i,t} = \alpha + \beta \times DataEmp_{i,t} + \gamma \times emp_{i,t} + \delta_i + \xi_t + \epsilon_{i,t} \tag{27}$$

where $FE_{i,t}$ denotes the squared forecast error for firm i in year t, $DataEmp_{i,t}$ is the employment share of Data workers in the industry of firm i, $emp_{i,t}$ is the log of firm i's total employment, and δ_i and ξ_t represent firm and year fixed effects.

Table 6 reports the results. We find a statistically and economically significant negative relationship between the share of Data Workers and the magnitude of forecast errors. Column (1) presents the specification without controlling for firm size, while Column (2) includes log employment. The estimate in Column (2) implies that a one–percentage-point increase in the employment share of Data Workers reduces squared forecast error by 0.005 percentage points. To gauge economic significance, note that the standard deviations of FE and DataEmp are 0.019 and 1.829, respectively. Thus, a one–standard-deviation increase in Data worker employment share is associated with a 0.48–standard-deviation decline in forecast errors.

Columns (3) and (4) repeat the analysis using the wage-bill share of Data workers (DataWage), defined as the percentage of the total wage bill paid to Data workers. Column (4) shows that DataWage is also significantly negatively associated with squared forecast

[5]We use this broader category rather than more granular codes to avoid breaks across years when detailed occupation definitions change or suffer from sampling limitations.

[6]Results are robust to excluding these unmatched observations.

errors after controlling for firm size and including firm and year fixed effects.

Table 6: Forecast Errors regressed on employment of Data workers

	(1) FE	(2) FE	(3) FE	(4) FE
DataEmp	-0.004*	-0.005**		
	(0.002)	(0.002)		
DataWage			-0.003	-0.004**
			(0.002)	(0.002)
emp		-0.015**		-0.015**
		(0.006)		(0.006)
Constant	0.022***	0.045***	0.022***	0.046***
	(0.002)	(0.009)	(0.003)	(0.009)
Observations	6431	6431	6431	6431

Standard errors in parentheses

Standard errors are two-way clustered by year and firm identifier following Cameron, Gelbach, and Miller (2011).

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Next, we examine whether the employment of Data workers, while strongly associated with lower squared forecast errors, also has a significant effect on firm revenue. Column (1) of Table 7 reports results from the following specification:

$$Rev_{i,t} = \alpha + \beta \times DataEmp_{i,t} + \gamma \times emp_{i,t} + \delta_i + \xi_t + \epsilon_{i,t} \tag{28}$$

where $Rev_{i,t}$ is the log of revenue for firm i in year t, and the remaining terms are defined as in Equation 27. To gauge economic significance, note that the standard deviations of log revenue and DataEmp are 1.679 and 1.829, respectively. A one–standard-deviation increase in the employment share of Data workers is associated with only a 0.04–standard-deviation change in log revenue. Likewise, given that the standard deviation of the wage-bill share is 2.113, a one–standard-deviation increase in this measure is also associated with just a 0.04–standard-deviation change in log revenue. These results suggest that while the presence of Data workers is clearly linked to reduced forecast errors, it does not translate into a

large changes in revenue. While statistically significant, the magnitude of the effect of Data workers on revenues pale in comparison to that induced by improved forecasts.

Table 7: Log revenue regressed on employment of data workers

	(1) log (Rev)	(2) log (Rev)
DataEmp	-0.039**	
	(0.015)	
DataWage		-0.032**
		(0.010)
emp	0.757***	0.756***
	(0.033)	(0.033)
Constant	5.797***	5.799***
	(0.058)	(0.056)
Observations	6431	6431

Standard errors in parentheses

Standard errors are two-way clustered by year and firm identifier following Cameron, Gelbach, and Miller (2011).

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Lastly, Table 8 reports results using the employment share (ESM) and wage-bill share (WSM) of Mathematical Professions, instead of Data workers. None of the measures shows a strong relationship with squared forecast errors. This placebo test reinforces our conclusion that the labor input most relevant for reducing forecast errors is captured by the employment of Data workers in the category *Database and Network Administrators and Architects*, but in general it is difficult to identify which inputs are the ones that map into data outputs.

We take these results as implying that, while the input that firms use to manage data clearly affect the output of such data (in terms fo reducing forecast errors), is not so clearly related to the revenue implications of such improvement. This result gives us confidence that in order to value data it is more natural and precise to directly study the relevance of data through predictions, than to study the indirect use of data through processing inputs.

Table 8: Forecast Errors regressed on employment of Maths occupation workers

	(1) FE	(2) FE	(3) FE	(4) FE
ESM	0.008	0.011		
	(0.008)	(0.008)		
WSM			0.009	0.012*
			(0.006)	(0.006)
emp		-0.015**		-0.015**
		(0.006)		(0.006)
Constant	0.015***	0.036***	0.014***	0.036***
	(0.002)	(0.009)	(0.002)	(0.009)
Observations	6431	6431	6431	6431

Standard errors in parentheses

Standard errors are two-way clustered by year and firm identifier following Cameron, Gelbach, and Miller (2011).

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6 Conclusion and Next Steps

This paper presents a novel framework for computing the value of firms' data that is missing in standard measures of GDP. Our approach recognizes data as both a productive asset and a means of payment. By quantifying data's value through its impact on firms' forecasting accuracy, we demonstrate that traditional measures of investment underestimate firms' true asset base. We argue that our *output-based* approach to measure the value of data complements and adds to the more standard *input-based* approach advanced in the literature.

Our approach allows us to answer several questions about the role of data in modern economies. First, the extent of reverse causality. Does data induce firms' growth or does firms' growth induce data? We show that data is indeed quite powerful on improving outcomes. Second, the extent to which data is obtained in day-to-day operations, undepreciated old data or purchased from external parties. We show that the most important source of data is the one obtained from day-to day operations. Finally, future work could explore the degree of heterogeneity in the value of data across firms.

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.

An important consideration to tackle in future work is accounting for market power. When equating the value of the data barter discount to the value firms extract from data, we assume perfect competition. This assumption delivers all the data surplus to the customers. While perfect competition is a natural starting point, it surely over-estimates the size of the data barter. We have explored robustness using estimations of product market power as applying for data market power. A fruitful research agenda involves measuring and disciplining the extent to which firms can extract rents from customers when buying data, which is likely quite different than the extent to which they can extract rents from selling products.

References

Abis, Simona and Laura Veldkamp, "The Changing Economics of Knowledge Production," *The Review of Financial Studies*, None 2024, *37* (1), 89–118.

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.

_ , _ , W. Erwin Diewert, Felix Eggers, and Kevin J. Fox, "GDP-B: Accounting for the Value of New and Free Goods," *American Economic Journal: Macroeconomics*, October 2025, *17* (4), 312–44.

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," EIB Working Papers 2016/08, European Investment Bank (EIB) 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.

_ , Janice C Eberly, Andrea L Eisfeldt, and Dimitris Papanikolaou, "The Economics of Intangible Capital," Technical Report 3 August 2022.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How costly are markups?," *Journal of Political Economy*, 2023, *131* (7), 1619–1675.

Eeckhout, Jan and Laura Veldkamp, "Data and markups: A macro-finance perspective," 2023.

Eisfeldt, Andrea L, Edward Kim, and Dimitris Papanikolaou, "Intangible value," Technical Report, National Bureau of Economic Research 2020.

_ , **Gregor Schubert, and Miao Ben Zhang**, "Generative AI and firm values," Technical Report, National Bureau of Economic Research 2023.

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.

Online Appendix

A Steady State

Assume a deterministic steady state with time-invariant primitives $\{\phi, \psi, \tau(\cdot), \zeta(\cdot)\}$ $\Omega_{t+1} = \Omega_t = \Omega$, $k_{t+1} = k_t = k$. The steady-state state equation follows from (17) with $\Omega_{t+1} = \Omega_t = \Omega$:

$$\Omega \;=\; \phi + \tau(k) + \psi + \big(1 - \zeta(\Omega)\big)\big(\Omega - \phi\big). \tag{29}$$

Equations (19) and (20) collapse to

$$V'(\Omega^{-1}) = -\,\gamma\, A\, k^{\alpha} + \beta\, \Gamma(\Omega)\, V'(\Omega^{-1}), \qquad \Gamma(\Omega) \equiv 1 - \zeta(\Omega) - (\Omega - \phi)\, \zeta'(\Omega),$$

so

$$V'(\Omega^{-1}) \;=\; \frac{-\,\gamma\, A\, k^{\alpha}}{1 - \beta\, \Gamma(\Omega)}\,.$$

So, the steady-state Euler equation is:

$$\alpha\Big(A\Big(1 - \frac{\gamma}{\Omega}\Big) - w\Big)\, k^{\alpha-1} - r \;+\; \frac{\beta\, \gamma\, A\, k^{\alpha}\, \tau'(k)}{\big(1 - \beta\, \Gamma(\Omega)\big)\, \Omega^2} = 0. \tag{30}$$

Equations (29) and (30) jointly determine steady state levels of scale k^*, and data Ω^*.

A.1 Steady-State System when depreciation is endogenous.

As we discussed, depreciation depends on the stock of data as follows

$$(1 - \zeta(\Omega)) = (\rho^2 + \sigma_\epsilon^2 \Omega)^{-1}$$

Hence, equation (29) of the state steady state becomes quadratic on the stock of data

$$\Omega \;=\; \phi + \tau(k) + \psi \;+\; \frac{\Omega - \phi}{\rho^2 + \sigma_\epsilon^2\, \Omega}. \tag{31}$$

Similarly, the definition $\Gamma(\Omega)$ in equation (19), whcih enters into the steady state Euler in equation (30) becomes

$$\Gamma(\Omega) = 1 - \zeta(\Omega) - (\Omega - \phi)\, \zeta'(\Omega) = \frac{1}{\rho^2 + \sigma_\epsilon^2 \Omega} - \frac{\sigma_\epsilon^2 (\Omega - \phi)}{(\rho^2 + \sigma_\epsilon^2 \Omega)^2} = \frac{\rho^2 + \sigma_\epsilon^2 \phi}{\big(\rho^2 + \sigma_\epsilon^2 \Omega\big)^2}. \tag{32}$$

In this case, the steady state is also the solution for scale and data, k^* and Ω^* that jointly solve the modified state steady state (31) and the steady state (30) with the modified expression (32).

A.2 Steady State without reverse causality

When there is no reverse causality, this is when scale depends on data (k depends on Ω) but data does not depend on scale (Ω does not depend on k), the deterministic steady state solution is analytic, even when depreciation is endogenous.

In our setting, this is the same as assuming $\tau(k) = \tau$. Define the total exogenous signals as $S \equiv \phi + \tau + \psi$. The state fixed point is then fiven by

$$\Omega \;=\; S \;+\; \frac{\Omega - \phi}{\rho^2 + \sigma_\epsilon^2\,\Omega}.$$

Equivalently, the quadratic in Ω:

$$\sigma_\epsilon^2\,\Omega^2 + \Big(\rho^2 - 1 - \sigma_\epsilon^2 S\Big)\Omega + \phi - \rho^2 S \;=\; 0.$$

Hence, in closed-form:

$$\Omega^* = \frac{-B \;\pm\; \sqrt{B^2 - 4\,\sigma_\epsilon^2\,C}}{2\,\sigma_\epsilon^2}, \qquad B \equiv \rho^2 - 1 - \sigma_\epsilon^2 S, \quad C \equiv \phi - \rho^2 S. \tag{33}$$

which is well-defined if the root is economically admissible: $\Omega^* > 0$ and $\rho^2 + \sigma_\epsilon^2\Omega^* > 0$.

Since $\tau'(k) = 0$, the continuation term vanishes and the steady-state FOC reduces to

$$\alpha\Big(A\Big(1 - \frac{\gamma}{\Omega^*}\Big) - w\Big)\;(k^*)^{\alpha-1} \;=\; r,$$

Hence,

$$k^* \;=\; \left[\frac{\alpha\big(A(1 - \gamma/\Omega^*) - w\big)}{r}\right]^{\frac{1}{1-\alpha}}. \tag{34}$$

B Model Simulation and Calibration

This appendix describes the simulation framework used to discipline the relationship between firm-level revenues, forecast errors, and data accumulation. The objective is to generate model-implied moments that can be compared to their empirical counterparts, thereby guiding the calibration of key parameters.

B.1 Value Function Iteration

Firms choose capital to maximize the present discounted value of revenues net of capital costs, taking into account the evolution of precision. The Bellman equation is:

$$V_i(\Omega_{i,t}^{-1}) = \max_{k_{i,t}} \left\{ \big[(1 - \gamma\Omega_{i,t}^{-1})A_i - w\big] k_{i,t}^{\alpha} - r k_{i,t} + \beta\,\mathbb{E}[V_i(\Omega_{i,t+1}^{-1})] \right\}.$$

The solution proceeds as follows:

1. **Grids.**

   - The productivity grid for A_i is constructed using quantiles of a log-normal distribution with mean $\mu_{\ln A}$ and variance $\sigma^2_{\ln A}$. We use a grid with 50 points.
   - The precision grid for Ω^{-1} is defined over a geometric sequence with 200 points covering the range 10^{-6} to 0.1 allowing for a wide range of forecast errors to be generated.
   - The capital grid for k is constructed using 80 points over a geometric sequence. The end points of the range were chosen through hit and trial to ensure that the optimal choice of k implied by VFI doesn't go out of the range.

2. **Flow Payoffs.** For each point on the grids (A, Ω^{-1}, k), flow revenues net of capital costs are computed.

3. **Continuation Values.** Given current revenues, the law of motion yields the next-period precision. The continuation value is then obtained by interpolating the value function $V(\cdot)$ between adjacent points on the Ω^{-1} grid.

4. **Optimal Capital Choice.** For each state (A, Ω^{-1}), the capital choice k^* is obtained numerically by maximizing the sum of current flow payoffs and discounted continuation values over the capital grid.

5. **Iteration.** The value function is updated iteratively until the supremum norm between successive iterations falls below a pre-specified tolerance.

B.2 Simulation of a Firm Panel

After solving for the value and policy functions, a panel of 1000 firms is simulated and is allowed to run for 100 periods.

1. **Initialization.** Each firm draws a permanent productivity A_i from the calibrated log-normal distribution. Initial precision is set at the median value of the Ω^{-1} grid.

2. **Evolution.** For each period and firm:

   - The policy function yields the optimal capital given $(A_i, \Omega^{-1}_{i,t})$.
   - An error shock is drawn with variance $\Omega^{-1}_{i,t}$.
   - Revenues and forecast errors are recorded.
   - Precision is updated deterministically using the law of motion.

3. **Output.** The simulated panel contains firm-year observations of revenues, forecast errors, capital, and precision, along with the underlying productivity draw.

B.3 Calibration

Our model has 10 parameters in total as described in section 2. We calibrate 5 of them outside the simulation and the remaining 5 are calibrated to match 5 targeted moments that we observe in the data.

We start with the procedure to calibrate of the parameters chosen outside the simulation. Rental rate of capital (r) and capital share (α) are chosen as 0.04 and 1/3 respectively. These values are used commonly in the literature. Next, we calibrate the persistence of the state (ρ). To obtain these, we create a new variable (θ) which serves as the proxy for the state: $\theta = \frac{\text{Revenue}}{\text{Total Assets}}$ i.e. revenue normalized by total assets of the firm. We then estimate the following regression:

$$\theta_t = \mu + \rho\theta_{t-1} + \epsilon_t$$

Note that as many firms do not provide guidance every year, they may not appear in our sample every year. However, we can still calculate the variable (θ_{t-1}) as long as we can obtain the revenue and total assets estimate for the previous year from the Compustat data. This regression estimates $\rho = 0.95$ as the persistence of the state.

Next, we calibrate ϕ, the prior precision (or equivalently precision of public information) and τ, the precision of new data for the median firm. We calibrate the model in steady state i.e. instead of time-varying values of public (ϕ_t) and private data of median firm (τ_t), we assume these to be at some steady state values ϕ and τ. To get an estimate for prior precision, we calculate the predicted value of our state variable using the estimated AR-1 coefficients. Denote the predicted state as $\hat{\theta}_t$. We calculate the prior forecast error FE_{prior} in a similar way as FE above

$$FE_{prior} = \Big| \frac{\hat{\theta} - \theta}{(\hat{\theta} + \theta)/2} \Big|^2$$

ϕ is chosen to match the prior precision for the median firm estimated from the data which in our panel is 0.007. Therefore, $\phi = 1/0.007 = 143$. We calibrate the median firm's data Ω to match it's forecast precision. In our sample, the median forecast error is 2.29×10^{-3} which we interpret as the posterior variance in revenue for a median firm conditional on all information available to the firm. Therefore, $\Omega_M = 1/(2.29 \times 10^{-3}) = 437$. Based on these, the median depreciation rate is estimated using the expression for depreciation in the steady state.

$$\zeta = 1 - \left(\rho^2 + \frac{\Omega}{\phi}\right)^{-1}$$

This gives the rate of depreciation $\zeta = 0.75$ for the median firm. Finally, using the data accumulation expression for a median firm in steady, we back out the expression for τ as

$$\zeta(\Omega - \phi) = \tau$$

This yields τ as 220. This completes the calibration of the five parameters $(r, \alpha, \rho, \phi, \tau)$ chosen outside the model.

With these parameters in place, we now calibrate the remaining parameters $(\gamma, \chi, \mu_{\ln A}, \sigma^2_{\ln A}, w)$ to match the moments from model simulated data with that obtained from our sample data. Specifically, we target the following moments

1. β_1 from this regression

$$\Pi_{i,t} = \beta_0 + \beta_1 \times \ln(FE_{i,t}) + \gamma \times \ln A_i + \epsilon_{i,t}$$

where $\Pi_{i,t}$ is log of revenue. From data: $\beta_1 = -0.016$

2. β_1 from this regression

$$\ln(FE_{i,t}) = \beta_0 + \beta_1 \times \Pi_{i,t-1} + \epsilon_{i,t}$$

From data: $\beta_1 = -0.41$

3. Mean and standard deviation of log of revenue (in millions of Dec 2002 USD):
   (a) Mean = 6.69
   (b) standard deviation: 1.72

4. Wages paid to data workers by the median firm as a fraction of it's revenue which we estimate as 0.16% from Occupational Employment and Wage Statistics (OEWS) data.

Note that we run the same regressions in our simulated panel as we run in the data. The calibration minimizes a loss function defined as the sum of squared percentage deviations between model-implied and empirical moments.

These moments and results are summarized in Tables 1-5 in the manuscript.

C Measuring the Evolution of Data Stock

Note that just as we have FE and FE_{prior} for each firm-year observation in our data, we can aggregate it at any level. For calculating the value of data, we calculate the median values across firms for every year. This gives us estimates for prior and posterior variance for every year.

Period 1 (Year 2003):

- Assumption: $FE_0 = 1/\phi_1, \Omega_0 = \phi_1$.
- FE_1 = median forecast error in 2003, $1/\phi_1$ is our estimate of prior variance for 2003 (for example through variance of quarterly revenues in the 3 year window around 2003)
- $\Omega_1 = \phi_1 + \tau_1$ and $\frac{1}{\Omega_1} - \frac{1}{\Omega_0} = FE_1 - FE_0$. We can calculate τ_1 from this
- By our assumption, $\Omega_0 = \frac{1}{FE_0} = \phi_1$, so $1/\Omega_0$ and FE_0 cancel out from the expression above.

Period t (Years 2004 - 2022):

- Stock in period t is $\Omega_t = \phi_t + (1-\zeta_{t-1})(\Omega_{t-1} - \phi_{t-1}) + \tau_t$ where $1-\zeta_{t-1} = (\rho^2 + \sigma^2_{\epsilon.t-1}\Omega_{t-1})$
- Compute τ_t using above expression and $\frac{1}{\Omega_t} - \frac{1}{\Omega_{t-1}} = FE_t - FE_{t-1}$

D Extrapolation of GDP (Mis)measurement

As discussed in the paper, GDP mis-measurement at the firm level is given by $2 \times D_t$. Note that because many firms do not provide guidance consistently every year, we cannot calculate the value of data at the firm level for all firms. To get around this, we calculate the value of data per employee for the median firm in our sample and use the employment count at different levels of aggregation to extrapolate the value of data (and GDP mis-measurement).

$$\text{Missing-GDP} = 2 \times \text{Value of data per employee} \times \text{Total number of workers}$$

To get the number of workers in the economy, we use the employment level series (CE160V) provided by the Federal Reserve Bank of St. Louis[7].

The estimate above does not take into account the fact that firms with market may not have to compensate the consumer fully for the data. To adjust this estimate for a firm's market power, we assume that if a firm charges a markup of μ over it's marginal cost in the goods market, it can apply a symmetric markdown μ when giving discount to the consumer for data. Therefore, the missing-GDP after accounting for market power takes the firm

$$\text{Missing-GDP} = \left(1 + \frac{1}{\mu}\right) \times \text{Value of data per employee} \times \text{Total number of workers}$$

Edmond et al. (2023) show that between after 1990, the cost-weighted average markup has stayed roughly stable between the range 1.2 - 1.26. To get a conservative estimate of missing GDP, we use the maximum value of 1.26 for μ in our calculations.

E Data Purchase

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).

From the model, we obtain the following estimate for the median firm in our data

$$\frac{\partial V_M(\Omega^{-1})}{\partial \Omega^{-1}}\Big|_{\Omega=\Omega_M} = -\$5941$$

which tells that that a unit decrease in Forecast errors is associated with an increase of \$5.941 billion in terms of December 2024 dollars.

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

[7]See https://fred.stlouisfed.org/series/CE16OV

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. If we interpret this amount as the present discounted value of all future additional revenue, then we can use this to estimate the value of data bought by the median firm as follows:

$$1.67 = \frac{\partial V_M(\Omega^{-1})}{\partial \Omega^{-1}}\Big|_{\Omega_M} \times \Delta\Omega^{-1} = 5941 \left(\frac{1}{\Omega_M} - \frac{1}{\Omega_M + \psi} \right)$$

This reduces to

$$\frac{1}{\Omega_M + \psi} = \frac{1}{\Omega_M} - \frac{1.67}{5941} = 0.002$$

This implies that $\psi = 61$ which is much smaller than the precision of public data $\phi = 143$ and the precision of newly acquired $\tau = 220$.