Growing Data, Slowing Productivity

Guillermo Ordonez and Laura Veldkamp*

July 24, 2025

Abstract

Data assets are increasingly vital in modern economies, yet macroeconomic measurement is not well-adapted to capturing the value of a data economy. Part of the problem is that data is an invisible, intangible asset. Investments in this asset are missed in national accounts and losses due to depreciation are missing in firms' balance sheets. The other half of the problem, unique to data, is that data is a means of payment in the modern economy. Consumption is under-valued when we count only its monetary price and not the value of the data bartered as part of the transaction. Using evidence on the quality of firms' revenue forecasts to measure data, new measures of data depreciation and asset pricing techniques to capitalize its asset value, we explore what measured GDP would be if data were treated similarly to a physical asset. Our findings suggest that the aggregate value of data has increased substantially and is currently between 5-8% of GDP. Data value is of the right magnitude to explain the productivity slowdown puzzle.

In the digital economy, data has become one of firms' most valuable and essential assets. The speed of growth of digital technologies leads to the conjecture that the value of data has grown immensely. It is also thought to primarily benefit large firms, exacerbating problems with market competition. Despite its growing economic importance, measuring the aggregate value of firms' data and its distribution remains elusive. Data is more highly valued than ever because it is the fuel for artificial intelligence (AI). Missing the value of data leads economists to underestimate output and productivity growth. Our objective is to quantify the value of missing GDP and determine how much of the measured productivity slowdown and rise in profits can be attributed to data.

*Preliminary. Please do not cite or circulate. ordonez@econ.upenn.edu, lv2405@columbia.edu. We thank Ankit Bhutani for his excellent research assistance and participants at the 2025 SED for helpful comments and suggestions. Keywords: AI, big data, GDP measurement, data barter.

Data plays three distinct roles in modern economics that show up in national accounts and firms' profits: 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 that promote competition, innovation, and responsible data-sharing practices. Improved 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 (Asriyan and Kohlhas, 2024). This approach does not imply that data is only used for revenue forecasting. Our approach recognizes that data is an input into many firm predictions. Firms with more or better data can predict many things well – including revenue. The observed correlation between revenue forecast errors and profits is not causal, either in the data or in our model. It captures the benefits of many accurate forecasts a firm with good data can make. The correlation between revenue forecast errors and firm profits captures both the broader productivity benefits of data and

the improvement in forecasting that results from firms' growth.

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. (Data purchases may also level the playing field. We consider that separately at the end of the paper.) The first effect comes from recognizing that data fuels predictions. It reduces prediction errors. However, expected squared prediction errors are bounded below by zero. This naturally gives data diminishing returns. The fact that the tenth data point reduces forecast errors much more than the millionth is a force enabling medium-size firms to catch up with larger ones. The second comes from the particular depreciation rate of data. We find that firms with large data sets suffer more from data deprecation. Because their forecasts are so precise, changes in the state increase their forecast errors proportionately more. We find that these two forces contribute to the catch-up we observe in mid-sized firms' prediction errors relative to those of large firms.

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.

The paper proceeds as follows: Section 1 describes the model that serves as our framework for measurement. Section 2 discusses how to take the model to the data and the unique properties of data mismeasurement. Section 3 measures the different sources of data and

their dynamics. Section 4 uses these measures and present the value of new data and the stock of data in the economy, with the implications for GDP and firm value mismeasurements. Section 5 applies this methodology to heterogeneous firms and presents preliminary evidence of the role of firm size. 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., 2019b), 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, 2021; 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, productivity 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 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. Similar to the analysis at the end of our paper, Jones and Tonetti (2020) explore the optimal amount of data sales, while we attempt to value measured 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 (or "own") 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 investments are expensed rather than capitalized, leading to an underestimating GDP. He suggests imputing the value of free digital services based on consumer willingness to pay, an approach later extended by Brynjolfsson et al. (2019a), who estimate the consumer surplus generated by digital goods. Their findings suggest that traditional GDP metrics significantly underestimate welfare gains in data-driven economies. However, these consumer-centric approaches do not address the firm-side valuation of data assets. Nakamura et al. (2017) advocate for

developing satellite accounts to systematically track data-related investments, an approach that complements traditional national accounts. While such methodologies have been applied in pilot studies, they remain limited in scope and are not yet widely implemented. Syverson (2017) investigates whether digital services could explain the productivity slowdown but finds that the discrepancy would need to be implausibly large to account for observed trends. Finally, internal presentations by the Bureau of Economic Analysis have proposed data measures based on the cost of recording, organizing, and maintaining databases Calderón and Rassier (2022). However, since customers often provide data as a by-product of economic activity, a cost approach is likely to miss much of data's value.

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

1 A Framework for Data Measurement

At its heart, this is a forecasting model. It is not a production economy with equilibrium prices. It is simply a device to organize measurements of forecasts and profits to help us distinguish and value various types of information. Since firm interactions are not our object of study, we study an individual firm i. Later, we add back multiple firms and consider how to aggregate them.

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

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

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

$$A_t = A + b\theta_t \tag{1}$$

Importantly, the firm can charge per unit of the good "in cash" or "in data", $P_t = P_{i,t}^{\$} + P_{i,t}^{D}$. The payment in cash is the 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 normalize the price of the good by $P_t = 1$. Hence we can write the monetary revenues of the firm as

$$R_{i,t} = P_t A_t k_{i,t}^{\alpha} = P_t(A + b\theta_t)k_{i,t}^{\alpha}. \tag{2}$$

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

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

First, the unit cost depends on the realized characteristics valued by costumers, with a higher θ_t generating more revenues but also more costly to produce. Second, the cost also depends

on how well the firm targets the state. When firms take actions that are not well-aligned with the profit state θ_t, its unit costs are higher.

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

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

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

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

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

and total profits as

$$\Pi_{i,t} = \pi_{i,t}k_{i,t}^{\alpha} - rk_{i,t} \tag{3}$$

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

$$\theta_t = \widehat{\theta}_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2) \tag{4}$$

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

$$\widehat{\theta}_{t+1} \;=\; \rho\widehat{\theta}_t + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim N(0, \sigma_\epsilon^2), \tag{5}$$

for $0 < \rho < 1$. This structure implies that in each period t, the state is normally distributed: $\theta_t \sim N(0, \phi^{-1})$, where ϕ is a function of the parameters in (4) and (5).

We assume that each firm has an information set $\mathcal{I}_t$, which contains data that helps the

firm forecast $\hat{\theta}_t$, the forecastable part of the profit state. When combined optimally with an algorithm, this data can be used as a signal s_t about θ_t that has normally distributed signal noise: $s_t = \hat{\theta}_t + \xi_t$, where $\xi_t \sim N(0, \tau^{-1})$.

In addition, we allow for the firm to have data from a data seller or platform. That data also has normally-distributed noise. By Bayes law, it can, therefore, be aggregated into one signal that provides data of precision ψ to the average firm. Given that η_t is a nonforecastable transitory shock, forecasting revenue is the same as forecasting the fundamental component plus some mean-zero noise.

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

$$\sum_{t=0}^{\infty}(1+r)^{-t}\mathbb{E}\left[(A-\gamma\left(a_{i,t}-\theta_t\right)^2)k_{i,t}^{\alpha}-rk_{i,t}|\mathcal{I}_t\right]. \tag{6}$$

The firm uses its prior about θ_t and its data, represented by s_t to take the expectation of the objective function (6). The firm's first-order condition dictates that the action that maximizes the expectation of that objective is $a_{i,t} = E(\theta_t|\mathcal{I}_t)$. Thus, the objective consists of minimizing the expected squared forecast errors $\left(E(\theta_t|\mathcal{I}_t) - \theta_t\right)^2$ about a state θ_t. The expected squared forecast error is, by definition, a conditional variance.

$$Var(\theta_t|\mathcal{I}_t) = \frac{1}{\phi+\tau+\psi}, \tag{7}$$

which depends on prior variance ϕ, data generated as a by-product of their activity with precision τ, and data purchased in the market with precision ψ.

How good this expectation is taken determines the marginal return of investing in capital and, hence, the choice of capital and total production. Here, we shut down the potential spiral that more production also generates more data; this is $s_{i,t}$ is not a function of $k_{i,t}$.

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

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

Next, determine today's forecast and variance of tomorrow's state. Taking the mean and then the variance of both sides of (5) once combined with (4), we get

$$E[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho E[\hat{\theta}_t|\mathcal{I}_t], \tag{8}$$

$$Var[\hat{\theta}_{t+1}|\mathcal{I}_t] = \rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] + \sigma_\epsilon^2. \tag{9}$$

Equation (9) tells us the conditional variance of the forecast of future state $\hat{\theta}_{t+1}$, conditional on the information known at the start of time t. A conditional variance is not a measure of the volatility of $\hat{\theta}_{t+1}$. Rather, it is a measure of uncertainty. It is the expected squared forecast error: $Var[\hat{\theta}_{t+1}|\mathcal{I}_t] := E[(\hat{\theta}_{t+1} - E[\hat{\theta}]_{t+1}|\mathcal{I}_t])^2|\mathcal{I}_t]$. It reveals how inaccurate our forecasts are. In Bayesian language, this is a prior variance of θ_{t+1}.

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

θ_{t+1} but has noise that is independent of the previous noise in the stock of data.[1]

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

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

$$Var(\hat{\theta}_{t+1}|\mathcal{I}_{t+1})^{-1} = \left[\rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] \ + \sigma_\epsilon^2\right]^{-1} + \sigma_\eta^{-2} + \sigma_s^{-2} := \Omega_{t+1}. \tag{10}$$

Recall from (6) that the firm's objective depends on the conditional variance, or expected squared forecast error, not of the state $\hat{\theta}$, but of the revenue θ_t, which is the state, plus transitory, unforecastable noise. Thus firm revenue depends on

$$Var(\theta_{t+1}|\mathcal{I}_{t+1}) = \left[\left[\rho^2 Var[\hat{\theta}_t|\mathcal{I}_t] \ + \sigma_\epsilon^2\right]^{-1} + \sigma_\eta^{-2} + \sigma_s^{-2}\right]^{-1} + \sigma_\eta^2 = \Omega_{t+1}^{-1} + \sigma_\eta^2. \tag{11}$$

Define the stock of data to be $\Omega_t := Var[\hat{\theta}_t|\mathcal{I}_t]^{-1}$. We define Ω with the inverse because this lends itself to interpreting Ω_t as the stock of data. A lower variance or more accurate estimate means we have more data about $\widehat{\theta_t}$. Identically, one period ahead we can define $\Omega_{t+1} := Var[\hat{\theta}_{t+1}|\mathcal{I}_{t+1}]^{-1}$, as defined in equation (10). Using this definition in equation (11), $Var(\theta_{t+1}|\mathcal{I}_{t+1}) = \Omega_{t+1}^{-1} + \sigma_\eta^2$.

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

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

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

2 From Model to Measurement

2.1 Data Barter

Even though we normalize the total price of the good to one, its composition will be endogenous and determined by the value for data to the firm from selling the unit of good.[2] If selling the unit generates data that is very valuable to reduce forecast errors and target good production more effectively, then it will charge less in cash, this is

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

2.2 GDP Accounting

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

2.2.1 Expenditure Approach

Under the traditional expenditure approach, GDP is defined as the sum of all expenditures:

$$\text{GDP}_{\text{exp}} = C + I + G + NX,$$

where C is total Consumption, I is Investment, G and $NX = X - M$ are Government Expenditures and Net Export, respectfully. These figures are measured by their market monetary prices.

Our economy produces final goods and data (a type of capital). Define real output and

[2]This discount assumes that the costumers can fully capture the value of the data its transaction provides. The challenge next is to determine what is the value of new data that is deducted from the value of goods. In the other extreme, if the firm does not compensate costumers from the value of data they provide, the firm is able to charge fully for its good, i.e., $P_{i,t}^{\$} = 1$.

the value of data as

$$Y = \sum_i \pi_i k_i^\alpha \tag{12}$$

$$D = \sum_i P_i^D \pi_i k_i^\alpha \tag{13}$$

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 $\sum_i P_i^\$ \pi_i k_i^\alpha = Y - D$. Investment on data, on the other hand, is not recorded as a transaction for its value D. Hence the mismeasured value of data in GDP is $2D$. If customers were not compensated for the value of the data they provide, the value of the good would be correctly measured by its monetary value but still data would be not measured since the transaction is still missing, and then GDP would be mismeasured only by the missing value of the data investment, D.

The correct formulation of GDP inclusive of data, is:

$$\text{GDP}_{\text{exp}}^{\text{corrected}} = (C + D) + (I + D) + G + NX,$$

Notice that under this approach, GDP is undermeasured for two D (value of generated data), and for two motives. One is *consumption bartering*, under which the monetary value of consumption is D less than the true value. The other is *intangible investment*, as the investment in data with value D 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 would be mismeasured, as standard with intangibles.

2.2.2 Income Approach

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

$$\text{GDP}_{\text{inc}} = wL + rK + \Pi + \delta_K K,$$

where wL denotes labor compensation, rK rental return to reproducible capital, Π pre-depreciation corporate profits, and $\delta_K K$ consumption of fixed capital. Because firms report profits net of depreciation, adding back $\delta_K K$ simply yields gross profits.

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

$$\text{GDP}_{\text{inc}}^{\text{corrected}} = wL + rK + (\Pi + D - \delta_D SD_{-1}) + D + \delta_K K + \delta_D SD_{-1},$$

where $\delta_D SD_{-1}$ is the depreciation of the accumulated stock of data capital. Notice that since data depreciation is both added and subtracted, it does not contribute to mismeasurement.

The first time the value of data D 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. Second, firms should deduct the depreciation of their data. The second time D 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 asset

is not subtracted from firm income. 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 two D (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 would be mismeasured, as standard with intangibles.

2.2.3 Firms' Valuations

The previous discussion shows clearly that also firms values are mismeasured. However, the extent of bartering is critical in determining whether this mismeasurement whether firms are over- or under- valued. When there is bartering, firms' values are undermeasured by $D - \delta_D SD_{-1}$. This is an issue only outside the steady state, since in steady state $D = \delta_D SD_{-1}$. 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 should if they would deduct the depreciation of data $\delta_D SD_{-1}$.

2.3 Data Depreciation

Rewriting equation (11) 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 time-$t+1$ stock of data. That law of motion says that we take the stock Ω_t, depreciate it by transforming it into $(\rho^2\Omega_t^{-1}+\sigma_\epsilon^2)^{-1}$ and then add on the flow of new data. New data is like the new investment in the data stock. This implies that for data, the depreciation rate is

$$\zeta_t = 1-(\rho^2+\sigma_\epsilon^2\Omega_t)^{-1}. \tag{14}$$

This depreciation rate teaches us that if the AR(1) process for the unobserved state is highly volatile (high σ_ϵ), or if the stock of data is large, then data will depreciate more quickly.

2.4 Capitalizing the flow value of data

We can now write the *expected per-period profit* from the objective (6) as an explicit function of data stock as

$$\mathbb{E}\left[(A-\gamma\left(a_{i,t}-\theta_t\right)^2)k_{i,t}^\alpha - rk_{i,t}|\mathcal{I}_t\right] = \pi(\Omega_{i,t})k_{i,t}^\alpha - rk_{i,t} \tag{15}$$

where we define $\pi(\Omega_{i,t}^{-1}) = \mathbb{E}_{\Omega_{i,t}}(\pi_{i,t}) = A-\gamma(Var(\hat{\theta}|\mathcal{I}_t)+\sigma_\eta^2) = A-\gamma(\Omega_{i,t}^{-1}+\sigma_\eta^2)$.

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

$$\mathbb{V}(\Omega_{i,t}^{-1}) = \max_{k_{i,t}} \pi(\Omega_{i,t}^{-1})k_{i,t}^\alpha - rk_{i,t} + (1+r)^{-1}\mathbb{V}(\Omega_{i,t+1}^{-1}). \tag{16}$$

This specification collapses into a deterministic problem with choice variable $k_{i,t}$ and state variable $\Omega_{i,t}$. Given a stock of data $\Omega_{i,t}$, optimal capital choice in period t is given by $\alpha\pi(\Omega_{i,t}^{-1})k^{\alpha-1} = r$. Hence,

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

Notice that $\pi(\Omega_{i,t}^{-1})$ is linear in $\Omega_{i,t}^{-1}$ with a negative sign. A reduction of one unit $\Omega_{i,t}^{-1}$ (an

increase in the stock of data or a reduction in forecasts errors) lead to more productive capital and an increase in the firm's scale.

2.5 Approximating the Missing Data Value

As mentioned above, $\pi(\Omega_{i,t}^{-1})$ is linear in $\Omega_{i,t}^{-1}$. A reduction of one unit $\Omega_{i,t}^{-1}$ leads to an increase in γ units in per-unit of production flow profits. We can linearize total profits

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

Linearly approximating the total profits (around no variance, this is around $\Omega_{i,t}^{-1} = 0$) is

$$\pi(\Omega_{i,t}^{-1})k_{i,t}^{\alpha} - rk_{i,t} \approx \tilde{\Pi} + \tilde{\pi}\Omega_{i,t}^{-1}$$

where $\tilde{\Pi} = M\tilde{A}^{\frac{1}{1-\alpha}}$ and $\tilde{\pi} = -\frac{\gamma}{1-\alpha}M\tilde{A}^{\frac{\alpha}{1-\alpha}} < 0$, with $M = \left[\left(\frac{\alpha}{r}\right)^{\frac{\alpha}{1-\alpha}} - r\left(\frac{\alpha}{r}\right)^{\frac{1}{1-\alpha}}\right]$ and $\tilde{A} = A - \gamma\sigma_{\eta}^2$. Notice that this is linear in the forecast errors, hence negative.

Then we can approximate (17), imposing the optimal amount of capital as

$$\mathbb{V}(\Omega_{i,t}^{-1}) \approx \tilde{\Pi} + \tilde{\pi}\Omega_{i,t}^{-1} + (1+r)^{-1}\mathbb{V}(\Omega_{i,t+1}^{-1}). \tag{17}$$

To value a firm's data, consider a firm that does not add any data to its stock in any future period. The value of that firm is the value of its current data. For a firm that adds no new data, the future data stock is the depreciated current data stock: $\Omega_{i,t+1} = (1-\zeta)\Omega_{i,t}$.

Denoting $\beta \equiv (1+r)^{-1}$ and rewriting (17), yields $m + \nu\Omega_{i,t}^{-1} = \tilde{\Pi} + \tilde{\pi}\Omega_{i,t}^{-1} + \beta[m + \nu\Omega_{i,t+1}^{-1}]$. Substituting $\Omega_{i,t+1}^{-1} = \sigma_{\epsilon}^2 + \rho^2\Omega_{i,t}^{-1}$ and rearranging

$$m + \nu\Omega_{i,t}^{-1} = [\tilde{\Pi} + \beta m + \beta\nu\sigma_{\epsilon}^2] + [\tilde{\pi} + \beta\nu\rho^2]\Omega_{i,t}^{-1}$$

Given the approximated linearity in the value function, we can express firm i's value as

$\mathbb{V}(\Omega_{i,t}^{-1}) = m + \nu\Omega_{i,t}^{-1}$, where the coefficients ν and m depend on the rate of time preference β, the persistence ρ and innovation variance σ_ϵ^2 of the state, and the per unit profit $\tilde{\pi}$, as well as the total profit Π of the firm:

$$\nu = \frac{\tilde{\pi}}{1 - \beta\rho^2} < 0 \tag{18}$$

and

$$m = \frac{\tilde{\Pi} + \beta\nu\sigma_\epsilon^2}{1 - \beta} \tag{19}$$

Our approach will be to estimate ν and m to map firm data precision into data value.

2.6 Discussions

2.6.1 Firm Size: Diminishing Returns vs. Scale Economies

Big firms have more data. When a firm already has a large volume of data, its forecast error is already small. Forecast errors cannot fall below zero. Thus, at some point, there must be some diminishing return to data, holding technology fixed.

At the same time, big firms with large revenues extract more value from a given improvement in forecasts. The firm has more production to apply its improved insights to. This is referred to as the scale economy of data.

The diminishing returns make large firms' marginal value of data lower, while economies of scale make the marginal value of data higher. In principle, this could change the value of data barter, as the firm grows. What we find suggests that these two forces roughly cancel. We return to this point in the quantitative results, Section 3.

2.6.2 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. While the simple exercise described below estimates a stable coefficient, one could easily add a time trend, estimate that coefficient separately for different sub-periods or model a productivity process based on productivity-related firm inputs.

3 Quantifying the Sources and Dynamics of Firms' Data

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

3.1 Data Description

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

To estimate the prior variance of revenue, we first create a new variable sales-ratio by dividing the revenue of the firm by it's trend revenue, obtained by using an HP filter. We winsorize this variable at 1% level. We then run an AR(1) regression on this variable by pooling all firm-year observations. The RMSE reported by this regression gives an estimate of the prior standard-deviation. Squaring this is the prior variance. We estimate the prior variance separately for each year, using quarterly data and three-year rolling windows.

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

We retrieved the annual sales guidance from I/B/E/S Guidance via the WRDS platform on April 20, 2023. We use both the "Detail" and "Identifiers" tables. The "Detail" table contains all kinds of financial estimates (e.g., annual or quarterly sales or earnings) made by firms. The "Identifiers" table contains identifying information, including "ticker" and "cusip." We merge the two tables and only keep US firms (observations with "usfirm"=1).

Our sample runs from 2002 to 2021. We chose 2002 as the starting date because the sales guidance data only became available for more than 10 firms after 2001. Each year, the sample size varies from 50 to 550 firms. After 2021, the number of firms with valid sales guidance dropped below 10 again.

Firms revise and update their annual sales guidance over the course of the fiscal year as realized quarterly sales data comes in. Since our analysis focuses on firms' ability to forecast annual sales, we only keep the initial estimates, which are not affected by realized quarterly data. When the sales forecast is expressed as a range, we take the mid-point (the lower and upper bound average). To adjust for inflation, we use the Consumer Price Index for All Urban Consumers (CPIAUCSL) monthly series downloaded from the FRED[3] and take the average in each year. We express all the sales numbers in 2002 dollars. We compare firms' sales guidance with realized sales data (retrieved from Compustat North America) and define the squared relative forecast error as in Equation 20. We winsorize the relative forecast error at the one percent level on the right tail. In each year, we take the average across firms.

3.2 Data Measurements

We proceed here to first obtain the precision of firms' data τ in average over the whole sample. First, we estimate expected forecast errors, from forecasters providing firms' forecasts of their future sales revenue. The mean squared relative forecast error is the average over time and over firms,

$$\text{Squared Relative Forecast Error} = \left| \frac{\text{Forecasted Sales} - \text{Actual Sales}}{(\text{Forecasted Sales} + \text{Actual Sales})/2} \right|^2. \tag{20}$$

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

The median squared relative forecast error in the sample 2003 and 2022 is $\sigma^2_{post} = 0.0022$.[4] In our model, all firms use prior information and private data to achieve a forecast error of the state θ_t given by (21).

Second, 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 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. 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

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

The median prior squared relative error estimated across all firms/periods is $\sigma^2_\epsilon = 0.007$. This gives us the prior information precision, $\phi = 1/0.007 = 142.9$.

Finally, te expected squared forecast error is, by definition, a conditional variance.

$$\sigma^2_{post} = \frac{1}{\phi + \tau} \tag{21}$$

Hence

$$0.0022 = \frac{1}{142.9 + \tau} \quad \Longrightarrow \quad \tau = 312.04$$

[4]We provide details on our computation of this error in Appendix A.

This is a first insight: *The precision of private data accrued in the year is twice the precision of prior information.*

What does this improvement in precision given by own data means in terms of dollars? To do this, and following our model, define $\Pi = \log(Revenue)$. We estimated the regression below:

$$\Pi = \hat{\Pi} - \tilde{\pi} FE \tag{22}$$

and obtained $\hat{\Pi} = 6.595$ and $\tilde{\pi} = 2.246$.

This implies that the change in revenue is,

$$\Delta\Pi = \tilde{\pi}\left(\frac{1}{\phi} - \frac{1}{\phi + \tau}\right) = 0.0108 \tag{23}$$

This means that revenues changed 1.08% from having more data compared to just having prior data (when precision was ϕ). Given the median revenue of the firm in our sample for the period 2002-2022 is 692.35 million dollars, this increase of precision from own data represents an increase of $692.35 \times 0.0108 = 7.47$ million dollars for the median firm.

3.2.1 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).

While it is tempting to measure the GDP value of data directly from the GDP of these platforms, an important question is how large this is compared to the value of firms own data, estimated above. To estimate this relative value, we now consider a version of our

model that is still static in nature ($\rho = 0$ and $\sigma_\eta = 0$), but where firms buy data from platforms. We assume a platform that provides data of precision ψ to the average firm. That data generates additional annual revenue for the firm, given by

$$\tilde{\pi}\left(\frac{1}{\phi+\tau}-\frac{1}{\phi+\tau+\psi}\right) \tag{24}$$

where again $\tilde{\pi} = 2.246$ is the parameter that maps precision units into units of additional revenues per year.

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

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

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

This last statement implies that by selling analytics based on unique data from transactions across the board, Visa generates $1.67m additional revenues per firm, which constitutes the value of Visa's data to those firms. We equate this figure to (24), the additional profit

generated by platform revenue, above what the firm would have earned using its own data. As before, we express everything in percentages of firm revenue of the median firm in our sample.

$$1.67/692.35 = 2.246\left(\frac{1}{142.9+312.04} - \frac{1}{142.9+312.04+\psi}\right) \quad \Longrightarrow \quad \psi = 454.15.$$

This is an additional insight: *The precision of data purchased in the year is a bit larger but of the same order of magnitude of own data.*

3.3 Data Dynamics

Now we study the evolution of the different sources of data precision. For lack of infomration about the value of purchased data, we focus on the evolution of ϕ_t and τ_t. To do this we also consider the depreciation of existing data. More formally, we measure each period t's stcok of data, where

$$\Omega_t = \phi_t + \tau_t + (1-\zeta_{t-1})[\Omega_{t-1} - \phi_{t-1}]$$

where ζ_{t-1} is defined by equation 14. Notice that only own data is assumed to depreciate at a rate ζ_t and public data fully depreciates.

The key question in this dynamic value measurement is determining how much data depreciates. Data about yesterday's state is less relevant to today's state because the state is changing. That decline in relevance is the source of data depreciation. While previous studies have guessed a depreciation rate, we will derive the depreciation rate from our model and estimate it with firm revenue data.

For this, we now assume that the state has persistence $\rho > 0$, which makes data from past periods relevant for forecasting today's state. This generalization of the model implies that data obtained in one period has value for multiple future periods. It is this multi-period

valuation that we compute next. We find that the persistence coefficient is estimated to be $\rho = 0.34$.

Figure 1 shows the evolution of the mean squared relative forecast errors $Var(\theta|\mathcal{I}_t))$ and the estimate of the prior precision (these are σ^2_{post} and σ^2_ϵ discussed above) in each period (see Appendix B for details).

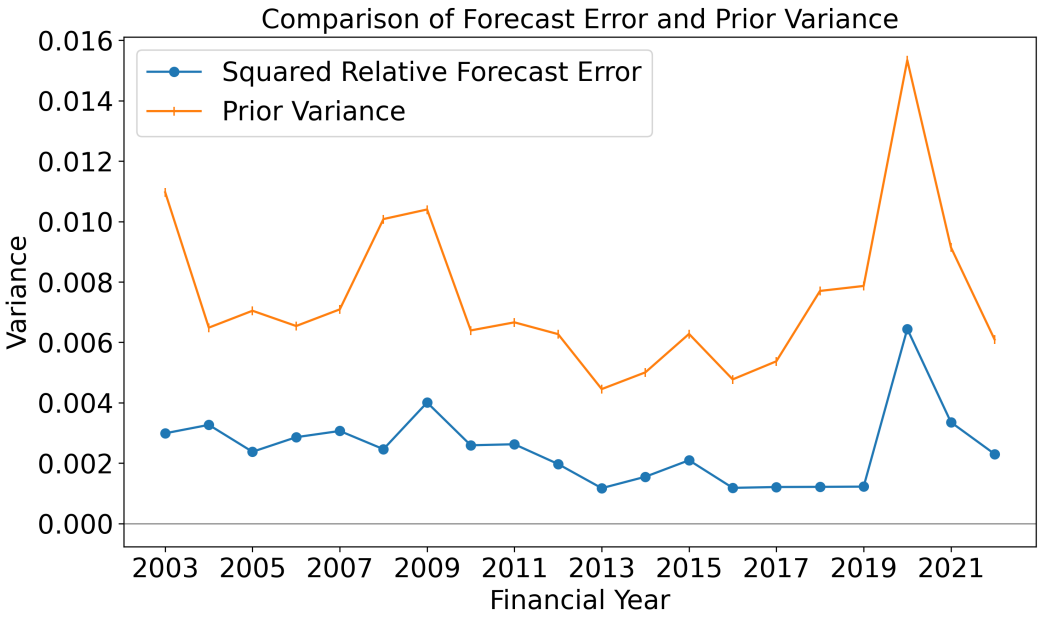


Figure 1: Forecast Errors and Precision of Own Data.

Figure 2 decomposes the data stock into components that come from prior knowledge, which we call public data ϕ_t, newly acquired data flow τ_t and the depreciated data stock carried over from the previous period. Consistent with the previous average, we see that the majority of the precision comes from the new data inflows. Notice that 2020 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.

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

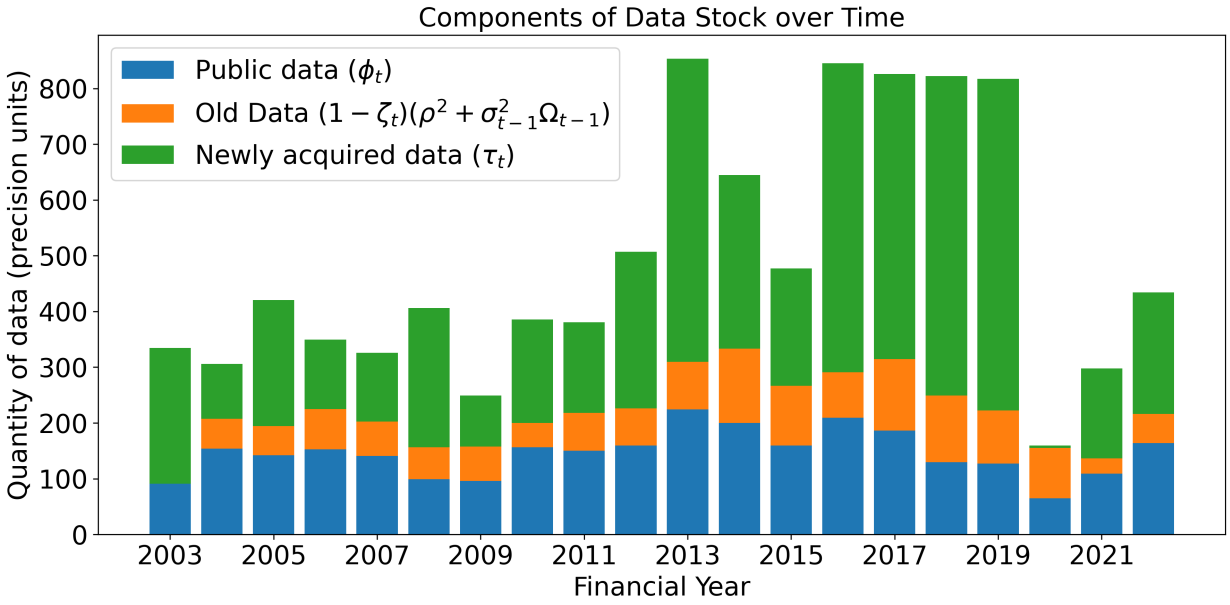


Figure 2: Data Precision: Public, Old and New Data (median firm).

4 Capitalizing and Aggregating the Value of Data

We now compute the capitalized (present discount value) of the flow of new data, the stock of data and the depreciated data for the median firm. We then aggregate by extrapolation these estimations to calculate the missing value of data in GDP measures.

4.1 Data Asset Values for a Median Firm

To quantify mismeasurement, we need to know the value of the data asset being transferred from customers to firms. To value data as an asset, we add up the one-period revenue flows

the data would generate, discounting future revenues and depreciating old data. For the discount factor, we use a constant rate rather than a stochastic discount factor because our framework has no covariance of data value with any aggregate risk. Therefore, we simply discount by the riskless rate: β is the rate of time preference. Future work could introduce stochastic valuation that would realistically discount the value of data for risk.

We compute the capitalized (present discount value, or PDV) of new data added to the firm in period t as

$$D_t = \left(\frac{1}{\Omega_t} - \frac{1}{\Omega_t - \tau_t}\right) \times \tilde{\pi} \times Rev_t \times \frac{1}{1 - \beta\rho^2}$$

where Rev_t is the revenue of the median firm in period t.

Similarly, we also compute the capitalized (present discount value, or PDV) of the stock of data as

$$SV_t = \left(\frac{1}{\Omega_t} - \frac{1}{\phi_t}\right) \times \tilde{\pi} \times Rev_t \times \frac{1}{1 - \beta\rho^2}$$

In the same way we also compute the capitalized (present discount value, or PDV) of the stock of data that depreciates in period t, which we call DEP_t.

Figure 3 displays the dynamics of these values. We see a steady rise in the value of firms' data assets that mirrors the growth in data flow value. The flow value D_t is what is added to the firm's depreciated stock of data from the previous period and the prior knowledge to obtain the new data stock in period t. This flow value is important for national income measurement because it represents the data assets that are bartered for goods in services in period t. However, the value plotted is not the entire missing value of data. It is just the value of data barter for the median firm in our sample.

4.2 Scaling Up: From the Median firm to GDP

Ideally, we would like to estimate the data value associated with each firm size, use the whole firm size distribution to impute a distribution of data values and add up these values.

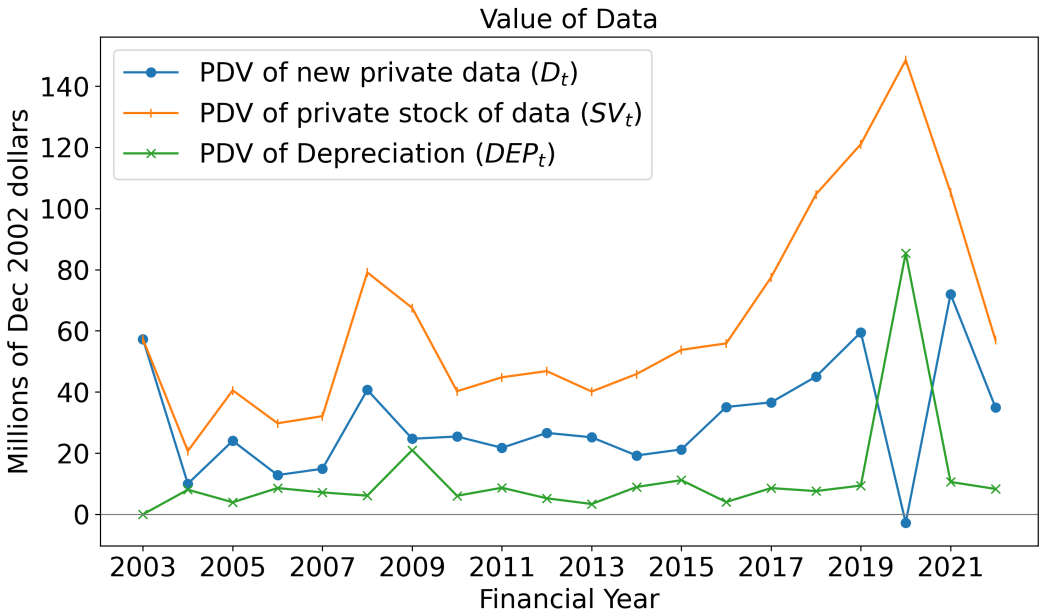


Figure 3: Asset Value of New Data, Data Stock and Depreciation (median firm).

Instead, we approximate data value heterogeneity using the five size quintiles of firms in our sample.[5] We project the estimated data value for each quintile to obtain an aggregate data value in three ways.

The first projection is an underestimate. It counts data value, assuming that no firms outside of our sample have any valuable data. It simply takes the median firm's data value in each quintile and multiplies this value by the number of firms in our sample.

The second projection considers all public firms. Our data is the subsample of public firms who issue sales forecasts. To extrapolate this value to public firms, we consider which of our size bins each public firm would fall in and assign that firm the data value of that size bin. For our sample, these bins are quintiles, each with an equal number of firms. However, in the sample of all public firms, the number of small firms is modestly higher. Thus, we add up the number of public firms in each bin and multiply this by the data value associated with that size bin. This estimate assumes that all private firms have zero data value. The

[5]In future work, we will explore the top size quintile of firms with more granularity. There is a highly-skewed right tail of firms with high valuations and large data sets that matter for aggregate data value. Our current approach misses these large data stocks by grouping these firms with the top quintile. Thus the median data value for the top quintile is surely an underestimate.

details can be found in Appendix C.

Public firms are less than half the economy. Their value added is only about 25% of GDP and their revenues are 40-50% of GDP. The third projection uses the employment share of the firms in our sample to scale up the data value of our median firm to the entire economy. This is likely an overestimate because small, private firms may have less data then their same-sized public counterparts. We consider the first and third methods lower and upper bounds and the second method closer to a baseline.

The estimates of missing GDP from all three methods are plotted in Figure 4. We find that the missing share of GDP is between 3-8% in most years.

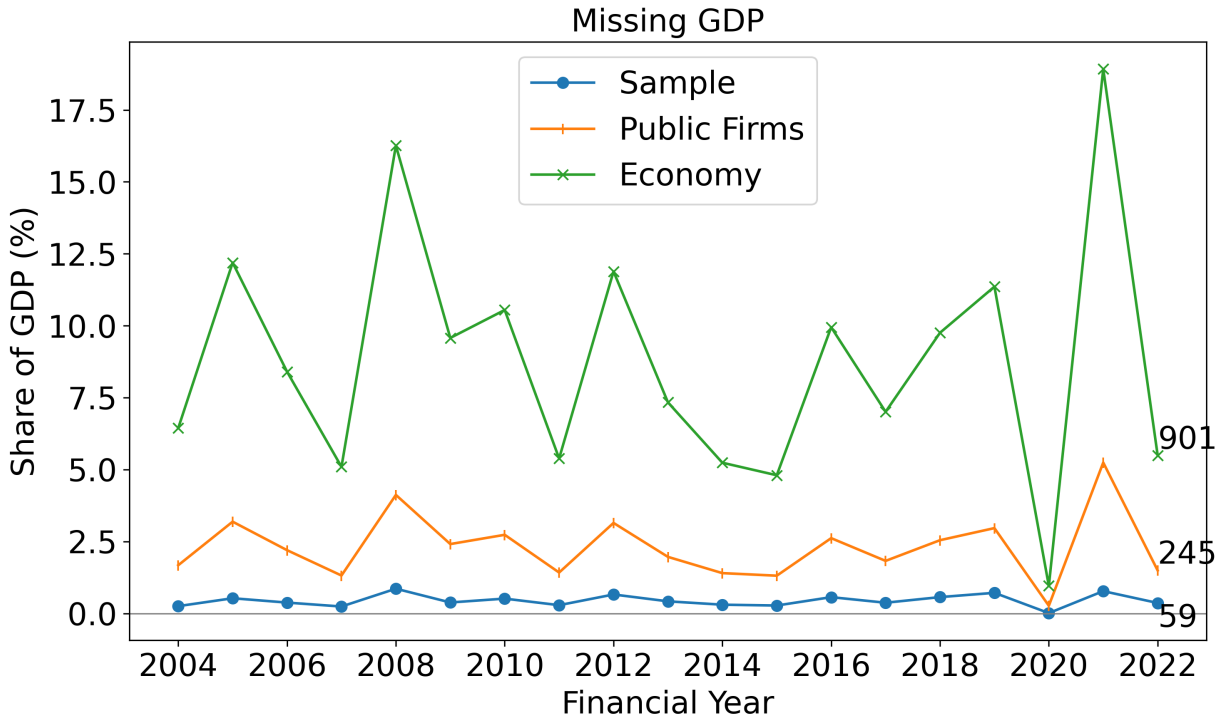


Figure 4: Estimates of missing data GDP measure with three methods.
The first method adds the value of data for firms in our sample. The second method multiplies the value of data in each quintile of our sample times the number of public firms of that size. The last estimate multiplies the public firm value by the inverse employment share of all public firms relative to total employment.

4.3 Comparing Data Growth to the Productivity Slowdown

The productivity slowdown refers to a drop of 1.3-1.5% of total factor productivity growth, starting around 2004 (Byrne et al., 2016; Fernald, 2014; Gordon, 2015). This phenomenon is observed across many high-income countries (Cette et al., 2016). Might data accumulation depress GDP measurement enough to account for this finding?

There are two clear outliers in our data valuations: The initial estimate and the 2020 COVID-19 year. The initial estimate of data flow value is too high because we initialize the model with zero data stock in 2002. Thus, all forecasting power of firms in 2003 is attributed to new data. That's a burn-in period which we ignore. By 2004, the data value settles down to be in line with later estimates. The other outlier period is 2020 when firm uncertainty was extremely high, due to structural changes brought on by the pandemic. That year should also be ignored or considered separately.

Let's start in 2004. In that year, about $12 million worth of data was transferred from customers to firms. Suppose this is a competitive market. If so, firms much forfeit approximately $12 million of revenue in discounts to obtain that data. The average public firm in 2004 had $1.2 billion in revenue. The lost revenue due to strategic discounts, given for the purpose of data accumulation, would be 1% of the firm's revenue.

On average from 2005-2019, the value of this missing revenue rises from $12-$50 million. That is an annual growth rate of about 11% per year. Thus, each period firms are missing 11% more revenue that should have been counted as sales of goods and services, as well as a separate payment to consumers for their data. The implied lost rate of growth of output is just under 1%, which is less than the magnitude of the productivity puzzle, but of the right order of magnitude to explain much of it.

Of course, assuming perfect competition between firms gives us an upper bound on the price discount firms offer consumers for data. On the other hand, recall that the missing GDP from a static barter trade is twice the value of the barter trade. Future work on this

project will estimate the average degree of market power and what it implies for data barter discounts.

5 Valuing the Data of Heterogeneous Firms

The previous analysis is based on taking averages across firms that, in principle, are quite heterogeneous regarding economic activity, technology, size, and dimensions of uncertainty. To better support the idea that perhaps slowing productivity is related to data accumulation, we dig into which types of firms accumulate the most data and exhibit the greatest productivity slowdowns.

5.1 Data Values by Size Quintile

To grasp the extent of heterogeneity in forecast errors and their evolution over time, we split firms by employment quintiles in 2005; we computed their forecast errors relative to other firm sizes. Then, we computed the flow value of the implied data separately for each quintile. In Figure 5, we plot the values of data for size quintiles 1, 3, and 5 only.

The striking feature is the overwhelming data advantage of large firms. The value of their data sets are an order of magnitude larger than those of the median firm. This points to the importance of considering firm heterogeneity when aggregating up the value of data in national income accounts.

Since employment in 2022 is larger than in 2005, we included all the firms in 2022 with more employment than the 2005 maximum in the last bin. Hence, a somewhat higher number of firms in the last bin.

These results suggest that the differences in data available to forecast were determined by size over the whole distribution in 2005 but not in 2022, except for the very small firms. Indeed, the differences in forecast errors for firms other than the smallest became very similar, perhaps pointing towards the dissipation of the largest firms' advantage in gathering data.

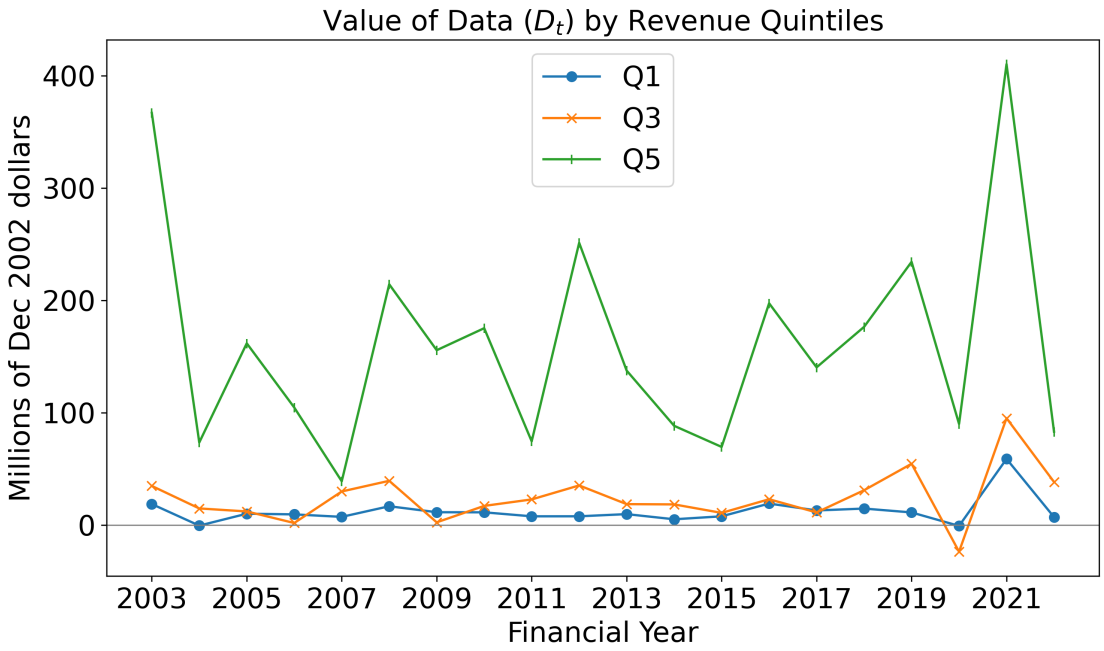


Figure 5: Data Values by Firm Size 2003-2022

6 Conclusion and Next Steps

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

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

Addressing these measurement gaps is essential for policymakers, economists, and corporate decision-makers. More accurate valuation of data assets can improve financial market

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

There are three key measurement issues that we plan to address in future versions of this paper.

1. Adjust 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. This assumption about how surplus is split also captures monopoly power that customers have when selling their data – if they are a natural monopolist over the facts that describe them as a unique individual they have some monopoly power over the provision of their data that they can use to extract surplus. While perfect competition is a natural starting point, it surely over-estimates the size of the data barter. To adjust our estimates for market power, we plan to use data on profits, as a fraction of revenue, to pin down a parameter that governs the fraction of surplus that firms retain.

2. Incorporate data processing costs. Another reason that customers to not receive the entire value of data as a payment, or discount on their purchase, is that firms must incur costs to turn raw data into forecasts. The primary cost is labor. To address this, we will use estimates of the number of data scientists hired across industries, and estimates of their earnings, to form a data processing cost measure that varies by firm size and industry. This should be subtracted from the data value to estimate the data

barter payment to customers.

3. Reverse causality. We use the covariance of forecast errors and firm profits to estimate the value of improved forecasts. However, more profitable firms may also acquire and process more or better data. To tease apart these mechanisms, we plan to build both effects into our model and estimate lead-lag relationships that distinguish the two directions of effects. Once our model teaches us what moments differ, depending on the nature of the effect, we will estimate those moments in our data and use them to further refine the model calibration.

References

Abis, Simona and Laura Veldkamp, "The changing economics of knowledge production," *Available at SSRN 3570130*, 2021.

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 in the digital economy," Technical Report, National Bureau of Economic Research 2019.

Byrne, David M., Stephen D. Oliner, and Daniel E. Sichel, "Does the United States Have a Productivity Slowdown or a Measurement Problem?," *Brookings Papers on Economic Activity*, 2016, *2016* (1), 109–182.

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.

Cette, Gilbert, John Fernald, and Benoît Mojon, "The Pre-Great Recession Slowdown in Productivity," *European Economic Review*, 2016, *88*, 3–20.

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

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

_ , Janice C Eberly, Andrea L Eisfeldt, and Dimitris Papanikolaou, "A model of intangible capital," Technical Report, National Bureau of Economic Research 2022.

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.

Fernald, John G., "Productivity and Potential Output Before, During, and After the Great Recession," *NBER Macroeconomics Annual*, 2014, *29* (1), 1–51.

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

Gordon, Robert J., "Secular Stagnation: A Supply-side View," *American Economic Review*, 2015, *105* (5), 54–59.

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 Calibration Details

Once we have the data, we calculate the Squared Relative Forecast Error for every firm-year observation as follows

$$FE = \Big| \frac{\text{Forecasted Sales} - \text{Actual Sales}}{(\text{Forecasted Sales} + \text{Actual Sales})/2} \Big|^2$$

We take this error as our estimate of the posterior variance. Next, we need estimates for two other parameters in our model: the persistence of the state (ρ) and prior variance (σ_ϵ^2). To obtain this, 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.

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

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

- Median Squared Relative Forecast Error in the data σ^2_{post}: 0.0022
- Median Squared Relative Error (Prior) - estimated through AR1 regression σ^2_ϵ: 0.0070
- New estimate of $\phi = 1/\sigma^2_\epsilon = 1/0.0070 = 142.9$
- New estimate of $\tau = 1/\sigma^2_{post} - \phi = 312.04$
- Define $\Pi = \log(Revenue)$. We estimated the regression below:

$$\Pi = \hat{\Pi} - \tilde{\pi} FE \tag{25}$$

  and obtained $\hat{\Pi} = 6.595$ and $\tilde{\pi} = 2.246$

- Change in revenue from data:

$$\Delta\Pi = \tilde{\pi}\left(\frac{1}{\phi} - \frac{1}{\phi + \tau}\right) = 0.0108 \tag{26}$$

  This means that revenues changed 1.08% from having more data compared to without additional data (when precision was ϕ).

- We can also directly compute the change in revenues from:

$$Rev_{prior} = \exp(6.595 - 2.246 \times \frac{1}{142.9}) = 720.02$$
$$Rev_{post} = \exp(6.595 - 2.246 \times \frac{1}{142.9 + 312.04}) = 727.83$$

  Then, the value of such data is around 7.8 million dollars.

- The median revenue of the firm in our sample for the period 2002-2022 is 692.35 million dollars. The increase in precision from data represents an average increase of 692.35 $\times$ 0.0108 = 7.47 million dollars for the median firm.

B Calculation procedure for value of data

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.
- PDV of data generated at t=1 is $D_1 = \left(\frac{1}{\Omega_1} - \frac{1}{\Omega_1 - \tau_1}\right) \times \tilde{\pi} \times Rev_1 \times \frac{1}{1-\beta\rho^2}$
- PDV of private data stock at t=1 is $SV_1 = \left(\frac{1}{\Omega_1} - \frac{1}{\phi_1}\right) \times \tilde{\pi} \times Rev_1 \times \frac{1}{1-\beta\rho^2}$

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 $\frac{1}{\Omega_t} - \frac{1}{\Omega_{t-1}} = FE_t - FE_{t-1}$
- PDV of data generated at t is $D_t = \left(\frac{1}{\Omega_t} - \frac{1}{\Omega_t - \tau_t}\right) \times \tilde{\pi} \times Rev_t \times \frac{1}{1-\beta\rho^2}$
- PDV of private data stock at t is $SV_t = \left(\frac{1}{\Omega_t} - \frac{1}{\phi_t}\right) \times \tilde{\pi} \times Rev_t \times \frac{1}{1-\beta\rho^2}$

C GDP Mismeasurement

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 give guidance consistently every year, we cannot calculate the value of data at the firm level D_t. To get around this issue while still accounting for heterogeneity, we divide the data into 5 equal buckets (quintiles) every year based on revenue. We run the above procedure for every quintile separately and obtain the amount of mis-measured GDP in each quintile in our sample as $2 \times D_{q,t} \times N_{q,t}$ where $D_{q,t}$ as represents value of newly acquired data for the median private firm in quintile q and $N_{q,t}$ represents the number of firms in our sample in quintile q. Then, the total mis-measured GDP in our data in year t is given by

$$\text{missing-GDP}_t^{sample} = 2 \sum_q D_{q,t} \times N_{q,t}$$

Extending to public firms: To obtain the corresponding estimate for public firms, we calculate the number of public firms ($N_{q,t}^{pub}$) that would fall in each of the 5 buckets above i.e. we use the same revenue thresholds as used above for creating the quintiles for each year, and then calculate the number of public firms in each bucket. The total missing GDP among public firms is then given as:

$$\text{missing-GDP}_t^{pub} = 2 \sum_q D_{q,t} \times N_{q,t}^{pub}$$

where $D_{q,t}$ is still the median estimate for quintile q from our sample but $N_{q,t}^{pub}$ is the number of public firms in the quintile instead of the number of firms in our sample.

Figure 6 illustrates the difference in firm sizes in our sample of firms that issue revenue forecasts and all public firms.

Extending to the economy: As there are no good estimates of revenue at the aggregate level, we rely on employment data. For every year, we calculate the total number of workers employed by the public firms. Then, we use the total employment yearly series (CE160V) to obtain the total number of workers in the economy. Our estimate for the missing GDP in the economy is then given by:

$$\text{missing-GDP}_t^{econ} = \text{missing-GDP}_t^{pub} \times \frac{\text{Total workers employed by public firms in year t}}{\text{Total workers employed in the economy in year t}}$$

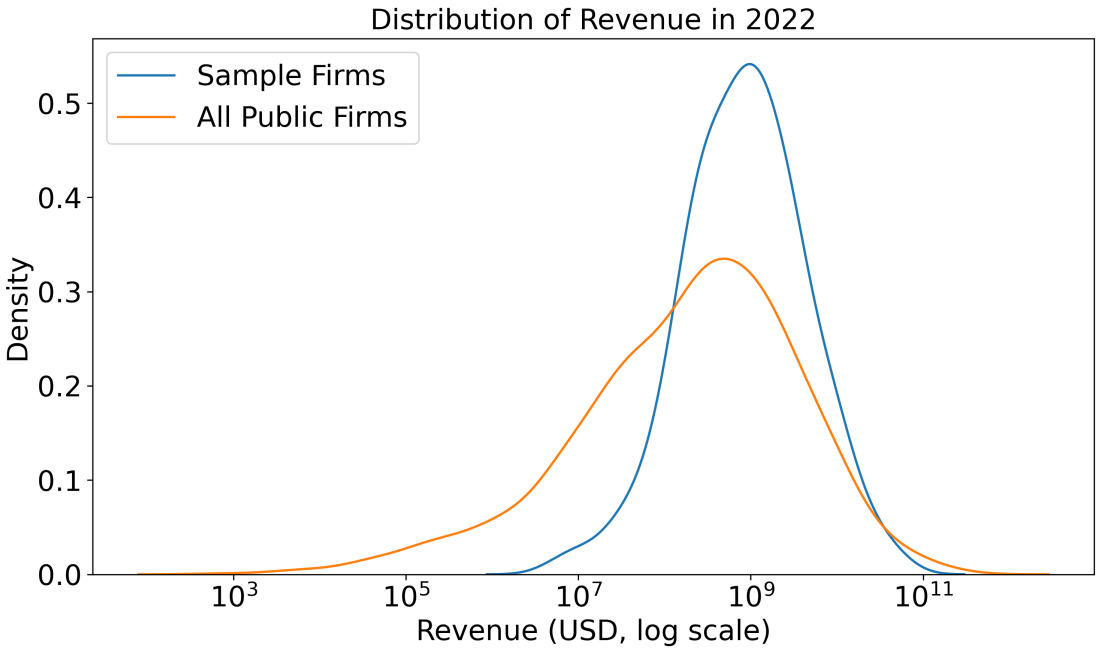


Figure 6: The size distribution of firms in our sample and all public firms. Size is measured by the annual firm revenue in 2022.