A Model of the Data Economy

Maryam Farboodi* and Laura Veldkamp†

June 23, 2026

Abstract

In a data economy, transactions of goods and services generate data, which is stored, traded and depreciates. How are the economics of this economy different from traditional production economies? How do these differences matter for measurement of GDP, firm values, depreciation rates, welfare and externalities? We incorporate active experimentation and data as an intangible asset to devise a tractable recursive representation of the data economy. The model rationalizes why apps are often "free" and why even non-digital economic activity might be greater than GDP suggests. Calibrating the model using a combination of macroeconomic and financial moments suggests that the mis-measurement in U.S. GDP due to missing value of data has been as high as 6% in 2018.

*MIT Sloan School of Business, NBER, and CEPR; farboodi@mit.edu.

†Columbia Graduate School of Business, NBER, and CEPR, 3022 Broadway, New York, NY 10027; lv2405@columbia.edu.
Thanks to Adrian Casillas, Rebekah Dix and Ran Liu for invaluable research assistance and to participants and discussants at numerous research seminars and conferences for helpful comments and suggestions. Finally, thanks to our editor, Kurt Mitman and our anonymous referees for all their constructive feedback. Keywords: Data, growth, digital economy, data barter.

Does the data economy have new economics? In the age of big data, production increasingly revolves around information. Many firms, particularly the most valuable U.S. firms, are valued primarily for the data they have accumulated. We have known since Wilson (1975) that ideas, data and other non-rival inputs have returns to scale. Because large firms benefit more from data, produce more data and grow bigger, data typically has increasing returns. At the same time, any data scientist will tell you that data has decreasing returns: Most of the predictive value comes from the first few observations. Understanding these opposing forces and what they mean for an economy requires constructing a new, dynamic equilibrium framework, with data as a state variable. Our model of the data economy teaches us that the long-run dynamics and welfare resemble an economy with capital accumulation and decreasing returns. However, the short-run features new dynamics, like increasing returns, negative profits, and the barter of data for goods.

The primary contribution of this paper is a tool to value data, measure its effects and to think clearly about the aggregate economic consequences of data accumulation. Measuring and valuing data are complicated by the fact that customers often provide their data, in exchange for a free digital service. Our value function assigns a positive value to goods and to data, even if they have a zero transaction price. In so doing, it moves aggregate models beyond price-weighted valuation and toward a modern way of thinking about economic value in a data economy.

As such, the contribution is not the particular predictions we explore. Some of our predictions are unsurprising, given the model assumptions. But the realism of the predictions supports the notion that the framework is a relevant and useful one. This degree of realism enables us to calibrate the model to macroeconomic and financial moments, which in turn informs us about the mis-measurement in aggregate GDP due to missing data.

Modeling the data economy is a challenge. A key feature is that firms/customer actions produce data, which is a form of information. When actions are chosen, taking into account the data those actions will generate, this is active experimentation. Micro models of active experimentation are typically challenging to solve (Bergemann and Välimäki, 2000), even without the complicating equilibrium forces. As an additional challenge, a useful model of the data economy should feature data as a long-lived, depreciating and tradeable asset. That calls for a recursive Bellman approach, with a data state variable. Tractably valuing data that a) comes from active experimentation, b) generates value for many periods, c) is traded in markets with equilibrium prices and d) eventually

depreciates, calls for a new set of tools. While the resulting model looks like a standard framework, achieving this degree of simplicity requires care.

The model in Section 1 describes "data" as a particular type of digitized information: Data is the transaction-generated information, used by firms to optimize their business processes, by accurately predicting future outcomes. The data economy blossomed with breakthroughs in machine learning and artificial intelligence. These are prediction algorithms. They require troves of data, which are naturally generated by transactions: buyer characteristics, traffic images, textual analysis of user reviews, click-through-rate data, and other evidence of economic activity. Predictions help firms optimize by forecasting demand, costs, earnings, labor needs, targeting advertising or selecting investments or product lines (Agrawal et al., 2022).

Because of its simple structure, the model can be applied and extended in many ways. We explore some in the paper; others, such as imperfect competition or firm size dispersion, are discussed in the conclusion. While adding features to the main model could allow it to better address one question or another, keeping the model streamlined allows it to be used flexibly.

Section 2 shows how to value and depreciate data, both tough to observe directly. However, our model offers a way to estimate how quickly a particular type of data loses its value. Bayes' Law and its cousin, the Kalman filter dictate the rate at which information precision depreciates depending on the current economic conditions and point us to a simple estimation procedure. Knowing how data depreciates allows us to build up a recursive value function structure that looks similar to ones used to value capital, but embodies the value of production as active experimentation and the unique way in which data depreciates.

Section 3 explores the path a given firm takes when growing to its steady state–the short run. When data is scarce, it may have increasing returns, because of a "data feedback loop." More data makes a firm more productive, which results in more production and transactions, which generate more data, further increasing productivity and data generation. This is the dominant force when data is scarce. Increasing returns also generates poverty traps. Firms with low levels of data earn low profits, which makes little production optimal. But little production generates little data, which keeps the firm data-poor. Firms may even choose to produce with negative profits, as a form of costly investment in data and may still have high equity market valuations, despite having minimal book value. This rationalizes observed "data barter." Many digital services, like apps, which are

costly to develop, are given away to customers at zero price. The exchange of customer data for a service, at a zero monetary price, is a barter trade.

Section 4 examines the data economy in the long run. We find that, in the long run, diminishing returns dominate. The long-run data economy looks like a long-run capital economy, but for different reasons: First, prediction errors can only be reduced to zero which places a natural bound on how much prediction error data can possibly resolve. Second, unforecastable randomness limits how accurate firms' forecasts can possibly be. Either one of these forces ensures that data alone cannot sustain growth. Of course, if we change the model to make data an input into research and development (R&D) it can sustain growth (Section 4.3). The main takeaway is the importance of measuring data used for R&D separately, similar to how we typically distinguish between regular capital investment and R&D investments.

Some of the most heated policy debates today revolve around firms' use of data. Thinking about regulation and welfare requires building out the household side of the model that micro-founds the demand curve. Section 5 does this and finds that, despite the non-rivalry, the increasing returns, and the production of data as a by-product of economic activity, equilibrium choices are efficient. That doesn't mean that data cannot cause harm. It just means that the simple forces our model describes do not compromise welfare, by themselves. When we add externalities, it prompts excessive data trade, which suggests a new direction to look to gauge welfare harms.

Section 6 shows how to use the model for measurement. Using a combination of macroeconomic and financial moments to calibrate the model, we estimate the extent of GDP mis-measurement due to data barter. Our calibration suggests that GDP should be 3-6% higher annually in 2003-2018 due to the missing value of transactions implicitly paid for with consumer data. It also illustrates the quantitative importance of properly depreciating data.

Section 7 extends the model to information that is industry, input or firm-specific and shows how the same model can describe firms that use data for product innovation. Finally, Section 8 provides directions for future research and concludes.

Related literature This work builds on insights from multiple literatures, each of which has some, but not all, of the features of this model. Work on information frictions in business cycles (Caplin and Leahy, 1994; Veldkamp, 2005; Lorenzoni, 2009; Ordonez, 2013; Ilut and Schneider,

2014; Fajgelbaum et al., 2017) have versions of a data feedback loop that operate at the level of the aggregate economy: More data enables more aggregate production, which in turn, produces more data. The key difference is that in those papers information is a public good, not a private asset. The private asset assumption in the current paper changes firms' incentives to produce data, allows data markets to exist and is what raises welfare concerns.

Choosing to acquire data is technically similar to the information choice in Broer et al. (2025) or rational attention choices in Maćkowiak and Wiederholt (2009), Matějka and McKay (2015) or Mankiw and Reis (2007). Our work borrows modeling strategies directly from Morris and Shin (2002) and Angeletos et al. (2006) and shares a focus on the social value of information. Work on media in the macroeconomy (Chahrour et al., 2021; Nimark and Pitschner, 2019) shares our focus on information markets. A novelty of a data economy is that transactions create data.

What differentiates our model from data and growth models is that our data is digitized information. Something is information if it predicts something. In Jones and Tonetti (2020), Cong et al. (2021) and Cong et al. (2022), data contributes directly to productivity. This is okay for their objective – exploring growth versus privacy. But without modeling data as an input into a prediction, they miss the tension between diminishing and increasing returns that is central to data valuation. The insight that the stock of knowledge can serve as a state variable appears in the five-equation toy model sketched in Farboodi et al. (2019).

Work exploring the interactions of data and innovation sounds similar, but has essential differences. For example, in Garicano and Rossi-Hansberg (2012), IT allows agents to accumulate more knowledge, which facilitates innovation. Agrawal et al. (2019) explore how breakthroughs in AI could enhance discovery rates and economic growth. In models of learning-by-doing (Jovanovic and Nyarko, 1996; Oberfield and Venkateswaran, 2018) and organizational capital (Atkeson and Kehoe, 2005; Aghion et al., n.d.), firms also accumulate a form of knowledge. But unlike prediction data, this knowledge is not a tradeable asset. Our work analyzes data accumulation in the absence of technological change. Once we understand this foundation, one can layer these insights about innovation and automation on top.

1 A Data Economy

Because machine learning and AI are prediction technologies, we build a framework in which data is used for prediction. To isolate the effect of data accumulation, the model holds fixed productivity, aside from that which results from data accumulation. There are inflows of data from new economic activity and outflows, as data depreciates. The depreciation comes from the fact that firms are forecasting a moving target. Economic activity many periods ago was quite informative about the state at the time. However, since the state has random drift, such old data is less informative about what the state is today.

The key differences between our data accumulation model and a capital accumulation model are three-fold: 1) Data is used for prediction; 2) data is a by-product of economic activity, and 3) data is, at least partially, non-rival. Multiple firms can use the same data, at the same time. These subtle changes in model assumptions are consequential. They alter the source of diminishing returns, create increasing returns and data barter, and produce returns to specialization.

1.1 Model

Real goods production Time is discrete and infinite. There is a continuum of competitive firms indexed by i. Each firm can produce $k_{i,t}^{\alpha}$ units of goods with $k_{i,t}$ units of capital. These goods have quality $A_{i,t}$. Thus firm i's quality-adjusted output is

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

The quality of a good depends on a firm's choice of a production technique $a_{i,t}$. Each period firm i has one optimal technique, with a persistent and a transitory component: $\theta_t + \epsilon_{a,i,t}$. Neither component is separately observed. The persistent component θ_t follows an AR(1) process: $\theta_t = \bar{\theta} + \rho(\theta_{t-1} - \bar{\theta}) + \eta_t$. The AR(1) innovation $\eta_t \sim N(0, \sigma_\theta^2)$ is *i.i.d.* across time.[1] Firms have a noisy prior about the realization of θ_0. The transitory shock $\epsilon_{a,i,t} \sim N(0, \sigma_u^2)$ is *i.i.d.* across time and firms and is unlearnable.

[1] One might consider different possible correlations of $\eta_{i,t}$ across firms i. An independent θ processes ($corr(\eta_{i,t}, \eta_{j,t}) = 0, \ \forall i \neq j$) would effectively shut down any trade in data. Since buying and selling data happens and is worth exploring, we consider an aggregate θ process ($corr(\eta_{i,t}, \eta_{j,t}) = 1, \ \forall i, j$). It is also possible to achieve an imperfect, but non-zero correlation.

The optimal technique is important for a firm because the quality of a firm's good, $A_{i,t}$, depends on the squared distance between the firm's production technique choice $a_{i,t}$ and the optimal technique $\theta_t + \epsilon_{a,i,t}$:

$$A_{i,t} = g\left((a_{i,t} - \theta_t - \epsilon_{a,i,t})^2\right). \tag{1}$$

The function g is strictly decreasing and known to all agents. A decreasing function means that techniques far away from the optimum result in worse quality goods.

Data The role of data is that better predictions allow firms to choose better production techniques. We are agnostic about whether firms predict demand, transportation logistics, supply chain risks, labor needs, competition or one of the many other uncertainties firms face. Instead, we build a structure where more accurate predictions help firms optimize business processes to be more profitable.

Transactions help to reveal uncertain outcomes, but the economic environment is constantly changing. Firms must continually learn to catch up. Observing production and sales processes at work provides useful information for optimizing business practices. For now, we model data as welfare-enhancing. Section 5 relaxes that assumption.

Specifically, data is informative about θ_t. At the start of date t, nature chooses a countably infinite set of potential data points for each firm i: $\zeta_{it} := \{s_{i,t,m}\}_{m=1}^{\infty}$. Each data point m reveals

$$s_{i,t,m} = \theta_{t+1} + \epsilon_{i,t,m},$$

where data noise, $\epsilon_{i,t,m} \sim N(0, \sigma_\epsilon^2)$, is *i.i.d.* across firms, time, and signals.[2]

The next assumption captures the idea that data is a by-product of economic activity. The number of data points n observed by firm i at the end of period t depends on their production $k_{i,t}^{\alpha}$:

$$n_{i,t} = z_i k_{i,t}^{\alpha},$$

[2]Hereafter, we treat the number of data points as a continuous quantity. To account for discreteness, suppose each data point can be subdivided into z finer points. If each original data point has unit precision, we assign each finer point a precision of $1/z$, so that the total information remains constant. Increasing z yields more points, each contributing proportionally less information. In the limit $z \to \infty$, the model becomes quasi-continuous, since the change from z to $z+1$ corresponds to an infinitesimal adjustment in total information.

where z_i is the parameter that governs how much data a firm can mine from its customers. A data mining firm is one that harvests lots of data per unit of output. Thus, firm i's production uncovers signals $\{s_m\}_{m=1}^{n_{i,t}}$.

We assume that the $n_{i,t}$ data points that firm i observes at time t includes the information inferred from the firm's own productivity $A_{i,t}$.[3] The transitory shock $\epsilon_{a,i,t}$ is important in preserving the value of past data and ensuring the $n_{i,t}$ data points the firm gets are not perfectly revealing. It prevents firms, whose payoffs reveal their productivity $A_{i,t}$, from inferring θ_t at the end of each period. Without it, the accumulation of past data would not be a valuable asset. If a firm knew the value of θ_{t-1} at the start of time t, it would maximize quality by conditioning its action $a_{i,t}$ on period-t data $n_{i,t}$ and θ_{t-1}, but not on any data from before t. All past data is just a noisy signal about θ_{t-1}, which the firm now knows. Thus preventing the revelation of θ_{t-1} keeps old data relevant and valuable.

Data trading and non-rivalry Let $\delta_{i,t}$ be the amount of data traded by firm i, after producing in date t. If $\delta_{i,t} < 0$, the firm is selling data. If $\delta_{it} > 0$, the firm purchased data.We restrict $\delta_{i,t} \geq \underline{\delta}$, where $\underline{\delta} \leq 0$. This does not prohibit selling or even selling multiple copies of data. But it does bound sales so that a firm cannot sell so much data that it is left with a negative stock of knowledge. If the firm buys $\delta_{i,t} > 0$ units of data, it adds the data it produced and the data it purchased, $n_{i,t} + \delta_{i,t}$ units of data, to its stock of data.

Let the price of one piece of data be denoted π_t.

Of course, data is non-rival. Some firms use data and also sell that same data to others. If there were no cost to selling one's data, then every firm in this competitive, price-taking environment would sell all its data to all other firms. In reality, that does not happen. Instead, we assume that when a firm sells its data, it loses a fraction ι of the amount of data that it sells to each other firm. Thus if a firm sells an amount of data $\delta_{i,t} < 0$ to other firms, then the firm has $n_{i,t} + \iota\delta_{i,t}$ data points left to add to its own stock of knowledge. Recall that for a data seller, $\iota\delta < 0$ so that the firm has less data than the $n_{i,t}$ points it produced. This loss of data could be a stand-in for the loss of market power that comes from sharing one's own data. It can also represent the extent of privacy

[3]Previous versions of this paper treated information inferred from productivity separately from data generated through transactions. That complicated the exposition and did not change any results. Results available upon request.

regulations that prevent multiple organizations from using some types of personal data. Another interpretation of this assumption is that there is a transaction cost of trading data, proportional to the data value.

Data adjustment and the stock of knowledge The information set of firm i when it chooses its technique $a_{i,t}$ is[4] $\mathcal{I}_{i,t} = \{\mathcal{I}_{i,t-1}, \{s_{i,t-1,m}\}_{m=1}^{\omega_{i,t-1}}\}$, where $\omega_{i,t-1}$ is the net number of data points added (or subtracted if ω is negative), after accounting for data purchases or sales. To make the problem recursive and to define data adjustment costs, we construct a helpful summary statistic for this information, called the "stock of knowledge."

Each firm's flow of $n_{i,t}$ new data points allows it to build up a stock of knowledge $\Omega_{i,t}$ that it uses to forecast future economic outcomes. We define the stock of knowledge of firm i at time t to be $\Omega_{i,t}$. We use the term "stock of knowledge" to mean the precision of firm i's forecast of θ_t, which is formally:

$$\Omega_{i,t} := \mathbb{E}[(\mathbb{E}[\theta_t|\mathcal{I}_{i,t}] - \theta_t)^2]^{-1}. \tag{2}$$

Note that the conditional expectation on the inside of the expression is a forecast. It is the firm's best estimate of θ_t. The difference between the forecast and the realized value, $\mathbb{E}[\theta_t|\mathcal{I}_{i,t}] - \theta_t$, is therefore a forecast error. An expected squared forecast error is the variance of the forecast. It's also called the variance of θ, conditional on the information set $\mathcal{I}_{i,t}$, or the posterior variance. The inverse of a variance is a precision. Thus, this is the precision of firm i's forecast of θ_t.

Our data adjustment cost Ψ captures the idea that if a firm that does not store or analyze any data wants to transform itself to a machine learning powerhouse, it would require new computer systems, workers with different skills, and learning by the management team. As a practical matter, if there is no data adjustment cost, a firm would immediately purchase the optimal amount of data, just as in models of capital investment without capital adjustment costs. Data adjustment costs are important because they make dynamics gradual.

[4]We could include aggregate output and price in this information set as well. We explain in the model solution why observing aggregate variables makes no difference in the agents' beliefs. Therefore, for brevity, we do not include these extraneous variables in the information set.

Equilibrium definition A firm chooses a sequence of production, quality and data-use decisions $k_{i,t}, a_{i,t}, \delta_{i,t}$ to maximize

$$\sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t \mathbb{E}\left[P_t A_{i,t} k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t \delta_{i,t} - r k_{i,t} | \mathcal{I}_{i,t}\right]$$

Firms update beliefs about θ_t using Bayes' law. Each period, firms observe last period's revenues and data, and then choose capital level k and production technique a. The information set of firm i when it chooses its technique $a_{i,t}$ and its investment $k_{i,t}$ is $\mathcal{I}_{i,t}$.

P_t denotes the equilibrium price per quality unit of goods. In other words, the price of a good with quality A is AP_t. By assumption, the inverse demand function and the industry quality-adjusted supply are:

$$\begin{aligned} P_t &= \bar{P} Y_t^{-\gamma}, \\ Y_t &= \int_i A_{i,t} k_{i,t}^{\alpha} di. \end{aligned} \tag{3}$$

Firms take the industry price P_t and the parameter $\bar{P}$ as given. Price is not random because, by the central limit theorem, the aggregate or average A converges to a known value.[5] The data price π_t equates data demand and supply. As in Solow (1956), we take the rental rate of capital as given. This reveals the data-relevant mechanisms as clearly as possible. This could be an industry or a small open economy, facing a world rate of interest r.

1.2 Interpreting Model Assumptions

Alternatives to data as a forecasting tool. In this model, the defining feature of data is that it is a tool to forecast a future state θ_{t+1}. This is not the only way to represent data. As mentioned before, some papers model more data as a direct contribution to TFP, which may well be a useful shorthand for data that is an input into R&D. Another approach to modeling data is as an improved matching technology. It could improve the match between customers and goods or between workers and tasks. Matching and noisy information are not separate phenomena. They are two ways of

[5] Appendix A shows that, because there are infinitely many firms with independent signals and a noisy prior, independent forecast errors imply independence of $\{A_{i,t}\}_i$ and that this implies a deterministic price and aggregate output.

representing an information friction. So, this could be a matching model. In this case, the noisy signal model was a more tractable formulation.

Can data be sold multiple times? Our setting allows this. Whether a firm sells d data points or sells 1 data points d times makes no difference. Each time a firm sells a data point, ι of knowledge is lost, whether it is the same point or a new one.

Investing in data quality. If a firm can pay for a higher z data processing ability, then this will further accentuate the data feedback loop and increasing returns. Larger firms with more transactions to process will get a higher marginal benefit from better data technology and will acquire even more knowledge than small firms. While that additional channel is interesting and may be quantitatively important, it doesn't change any of the ideas we develop in this paper. Therefore, we hold z fixed for simplicity.

Why this formulation of quality? It makes sense to assume g is decreasing because otherwise, worse forecasts improve quality. But the argument of the g function is quadratic in the difference between actions and optimal actions. This quadratic form is an approximation to many relationships. It has a long history in tracking problems like Morris and Shin (2002). Most importantly, this formulation simplifies the solution because it ensures that conditional variance is an approximate sufficient statistic for mapping what a firm knows to their value function.

1.3 Model Solution: Optimal Technique and Expected Quality

A key to simplifying the problem to a one-state variable problem lies in understanding the expected quality that results from the optimal choice of technique.

Taking a first order condition with respect to the technique choice, we find that the optimal technique is $a^*_{i,t} = \mathbb{E}_i[\theta_t|\mathcal{I}_{i,t}]$. Thus, expected quality of firm i's good at time t in (1) can be rewritten as $\mathbb{E}[A_{i,t}] = E\left[g\left((\mathbb{E}_i[\theta_t|\mathcal{I}_{i,t}] - \theta_t - \epsilon_{a,i,t})^2\right)\right]$. The squared term is a squared forecast error. It's expected value is a conditional variance, of $\theta_t + \epsilon_{a,i,t}$. That conditional variance is denoted $\Omega^{-1}_{i,t} + \sigma^2_u$.

To compute expected quality, we first take a second-order Taylor approximation of the quality function, expanding around the expected value of its argument: $g(v) \approx g(\mathbb{E}[v]) + g'(\mathbb{E}[v]) \cdot (v - \mathbb{E}[v]) + (1/2)g''(\mathbb{E}[v]) \cdot (v - \mathbb{E}[v])^2$. Next, we take an expectation of this approximate function: $\mathbb{E}[g(v)] \approx g(\mathbb{E}[v]) + g'(\mathbb{E}[v]) \cdot 0 + (1/2)g''(\mathbb{E}[v]) \cdot var(v)$. Recognizing that the argument v is a

chi-square variable with mean $\Omega_{i,t}^{-1} + \sigma_u^2$ and variance $2(\Omega_{i,t}^{-1} + \sigma_u^2)$, the expected quality of firm i's good at time t in (1) can be approximated as

$$\mathbb{E}[A_{i,t}|\mathcal{I}_{i,t}] \approx g\left(\Omega_{i,t}^{-1} + \sigma_u^2\right) + g''\left(\Omega_{i,t}^{-1} + \sigma_u^2\right) \cdot \left(\Omega_{i,t}^{-1} + \sigma_u^2\right). \tag{4}$$

If the g function is not too convex, then quality is a deceasing function of expected forecast errors. Or put simply, more data precision increases the quality of a firm's good. We will return to the question of highly convex, unbounded g functions in the next section.

2 Valuing and Depreciating Data

Before exploring predictions of the model, we work out what this model structure teaches us about how data should be depreciated and valued.

2.1 Data Depreciation

Solving our dynamic model requires taking a stand on the depreciation rate of data. This depreciation rate estimation is of independent interest. For the most valuable firms in the world, data is arguably their most valuable asset. Yet, data valuation and data accounting are in their infancy. A key question for valuing data is assessing how quickly data depreciates.

Luckily, our model also points us to a method for quantifying depreciation. It teaches us that the depreciation rate of data is a particular function of the persistence and volatility of the environment that data is used to forecast. We derive and explain this depreciation formula, which can be used in this model, or in any environment where data is used for forecasting and where a linear and normal stochastic environment is a reasonable approximation.

To derive this depreciation formula, we start from the state evolution equation. Recall that it is an AR(1): $\theta_{t+1} = \bar{\theta} + \rho(\theta_t - \bar{\theta}) + \eta_{t+1}$. Consider the beliefs about the time-t state and how they change when the same information is used to forecast the $t+1$ state. At the start of date t, the conditional variance of beliefs about the state θ_t is $V[\theta_t|\mathcal{I}_t] := \Omega_t^{-1}$, where Ω_t is what we've called the "stock of knowledge" and is the object we want to depreciate.

Next, we simply apply the same conditional variance operator, with the same information set,

to the AR(1) equation above: $V[\theta_{t+1}|\mathcal{I}_t] = \rho^2 V[\theta_t|\mathcal{I}_t] + \sigma_\theta^2 = \rho^2 \Omega_t^{-1} + \sigma_\theta^2$. This holds in the absence of learning any additional information about the state during all of period t. In this no date-t learning case, we invert the variance and rearrange $V[\theta_{t+1}|\mathcal{I}_t]^{-1}$ to get:

$$\Omega_{t+1}^{no\ learning} = \frac{\Omega_t}{\rho^2 + \sigma_\theta^2 \Omega_t}.$$

To be clear, this is not the correct law of motion for the state Ω in this model because firms learn new information every period. But examining the no-learning case is instructive because the only thing changing the stock of knowledge from one period to the next is depreciation. While typically, one would depreciate a capital stock by multiplying capital k_t times a term like $(1 - \delta^k)$. The equivalent multiplicative term here is $(\rho^2 + \sigma_\theta^2 \Omega_t)^{-1}$, which multiplies Ω_t. Thus, the depreciation rate, the equivalent of δ^k in a capital accumulation model, is

$$\text{data depreciation rate} = 1 - \frac{1}{\rho^2 + \sigma_\theta^2 \Omega_t}$$

A larger fraction of the stock of knowledge is lost to depreciation when the state changes lots from one period to the next (high σ_θ^2), when there is lots of knowledge to begin with (high Ω_t), and when high persistence makes the state a more variable process (high ρ).[6]

Depreciation rates are typically linear operators on the stock being depreciated. Appendix A.3 describes three types of economies where the data depreciation rate will be well-approximated by a standard-looking multiplicative constant term.

Accounting rules depreciate all data like software, by amortizing it over three years. That is a depreciation rate of 30% per year. Our results suggest that the depreciation rate of data may vary widely, depending on whether the data is used to forecast something more static, like consumer location or tastes, or something more ephemeral like equity order flow.

[6]One might wonder why this depreciation rate can be negative for small values of $\rho^2 + \sigma_\theta^2 \Omega_t$. These are cases where the firm is so uncertain that its conditional variance is higher than the unconditional variance of next period's outcomes. This is not a scenario that ever arises in our model. If an agent were so uncertain, then simple mean-reversion should reduce their uncertainty. This natural reduction in uncertainty, without any additional data, is what would show up as a negative rate of depreciation.

2.2 A Law of Motion for Data

To get from this depreciation rate to the law of motion for the stock of knowledge requires adding new data from three sources: 1) data that was a by-product of production, 2) data that was bought or sold and 3) data that was inferred from a firm seeing its own quality at the end of the period. These pieces of information are incorporated into beliefs using Bayes' law.

The number of new data points generated by firm i's production, $n_{i,t}$ is assumed to be data mining ability times end of period physical output: $z_i k_{i,t}^{\alpha}$. Bayes law tells us that the posterior precision of a normal variable is the sum of the prior precisions and signal precisions. This means that the sum of the precisions of all the data points, $n_{i,t}\sigma_{\epsilon}^{-2}$, should be added to the stock of knowledge.

At the firm level, data inflows need to be adjusted for data trade. If a firm buys data ($\delta_{i,t} > 0$), we add all the newly-acquired data precision $\delta_{i,t}\sigma_{\epsilon}^{-2}$ to the stock of knowledge. If a firm sells data ($\delta_{i,t} < 0$), we subtract a fraction ι of that signal precision from their stock of knowledge. Since $\delta_{i,t}$ is negative, we add the negative number $\delta_{i,t}\sigma_{\epsilon}^{-2}$ to subtract off the lost knowledge.

Lemma 1 puts the data depreciation and data inflows together. It tells us how the stock of knowledge evolves from one period to the next.

Lemma 1 Evolution of the Stock of Knowledge *In each period t,*

$$\Omega_{i,t+1} = \left[\rho^2 \Omega_{i,t}^{-1} + \sigma_{\theta}^2\right]^{-1} + \left(n_{i,t} + \delta_{i,t}(\mathbb{1}_{\delta_{i,t}>0} + \iota \mathbb{1}_{\delta_{i,t}<0})\right)\sigma_{\epsilon}^{-2} \tag{5}$$

The proof of this lemma and of all the lemmas and propositions that follow are in Appendix A. The proof is an application of Bayes' law, or equivalently, the Ricatti equation of a modified Kalman filter. Because the information structure is similar to that of a Kalman filter, the sequence of conditional variances, or their inverse, the sequence of precisions, is deterministic.

Information from aggregate prices One might wonder why firms do not also learn from seeing aggregate price and the aggregate output. They reflect aggregate quality, which depends on the squared difference between θ_t and other firms' technique a_{jt}. That squared difference reflects how much others know, but not the content of what others know. Because the mean and variance of normal variables are independent, knowing others' forecast precision reveals nothing about θ_t.

Seeing one's own outcome $A_{i,t}$ is informative only because a firm also knows its own production technique choice $a_{i,t}$. Since firms' actions are not observable, aggregate prices or quantities reveal what other firms predicted well. But they convey no useful information about whether θ_t is high or low.

2.3 Valuing Data: A Recursive Representation

One of the most important valuation questions for modern economists, investors and accountants is how to value data. While some data is transacted and might be valued at its price, lots of data is retained by a firm, for its own use. A value function approach assigns a value to a firm with a given amount of data. While that is not a cookbook recipe for assigning a dollar value to data, it offers a first step, a clear way to think about data value and what its components are. Our value function can guide data valuation, in the same way that capital value functions have guided economists' measurement of capital values, for decades.

Lemma 2 *The optimal sequence of capital investment choices* $\{k_{i,t}\}$ *and data sales* $\{\delta_{i,t} \geq -n_{i,t}\}$ *solve the following recursive problem:*

$$V(\Omega_{i,t}) = \max_{k_{i,t},\delta_{i,t}} P_t \mathbb{E}[A_{i,t}|\mathcal{I}_{i,t}]k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t\delta_{i,t} - rk_{i,t} + \left(\frac{1}{1+r}\right)V(\Omega_{i,t+1}) \tag{6}$$

where $\mathbb{E}[A_{i,t}|\mathcal{I}_{i,t}]$ *is an increasing function of* $\Omega_{i,t}$*, given by (4),* $n_{i,t} = z_i k_{i,t}^{\alpha}$*, and the law of motion for* $\Omega_{i,t}$ *is given by (5).*

This result greatly simplifies the problem by collapsing it to a deterministic problem with choice variables k and δ and one state variable, $\Omega_{i,t}$, the stock of knowledge. In expressing the problem this way, we have already substituted in the optimal choice of production technique. The quality $A_{i,t}$ that results from the optimal technique depends on the conditional variance of θ_t.

Since $\Omega_{i,t}$ can be interpreted as a discounted stock of data, $V(\Omega_{i,t})$ captures the value of this data stock. $V(\Omega_{i,t}) - V(0)$ is the present discounted value of the net revenue the firm receives because of its data. Therefore, the marginal value of one additional piece of data, of precision 1, is simply $\partial V_t/\partial\Omega_{i,t}$. When we consider markets for buying and selling data, $\partial V_t/\partial\Omega_{i,t}$ represents the firm's demand, its marginal willingness to pay for data.

3 Transition Path in the Data Economy

A key source of difference between a capital-based and a data economy is the short-run convexity of data accumulation, at the firm level. The convexity is a form of increasing returns that arises from the data feedback loop: Firms with more data produce higher quality goods. The higher profit per unit from higher quality goods induces more production, which results in more transactions and more data. Thus more data begets more data. While that sounds positive, it also creates the possibility of a firm growth trap, with very slow growth and financial losses, early in the lifecycle of a new firm. As a result, the life-cycle path of book-to-market or Tobin's Q of data firms looks very different from capital-intensive firms. Finally, the fact that transactions generate data as a by-product explains why every exchange includes an element of barter, where goods are exchanged for data, frequently at a positive monetary price. But sometimes, the exchange of goods for data happens at a zero monetary price, in which case pure barter arises.

While these results may not be a surprising distance from our assumptions, they all demonstrate the ability of the framework to make sense of and re-interpret new data economy phenomena. Tools to model data phenomena can, in turn, be used to inform ongoing policy debates. Establishing that this is an economically-relevant collection of assumptions is important before using it for measurement or welfare analysis.

3.1 Increasing Returns in the Short Run

Focusing on the dynamics of one firm growing makes forces clearer. The simulated model will show all firms growing. But these results explain the logic behind the transitions. While all others are in steady state, we drop in one, atomless, low-data (low $\Omega_{i,t}$) firm and observe its growth and transition to a high-data firm. For this section, we adopt a linear quality function, for simplicity: $g(x) = \bar{A} - x$. We relax this assumption later on, when we discuss the long run.

Proposition 1 ***S-Shaped Accumulation of Knowledge*** *When all firms are in steady state, except for one firm i, then the firm's net data flow $\Omega_{i,t+1} - \Omega_{i,t}$*

a. *increases with the stock of knowledge $\Omega_{i,t}$ when that stock is low, $\Omega_{i,t} < \hat{\Omega}$, when goods production has sufficient diminishing marginal return, $\alpha < \frac{1}{2}$, adjustment cost Ψ is sufficiently low, $\bar{P}$ is sufficiently high, and the second derivative of the value function is bounded $V'' \in [\nu, 0)$; and*

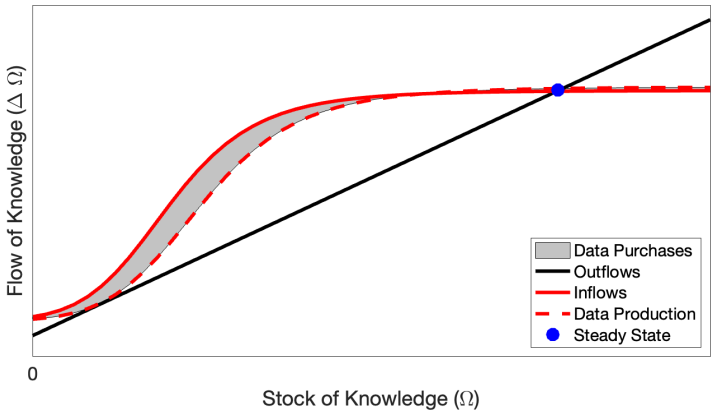

Figure 1: A single new firm grows slowly: Inflows and outflows of one firm's data.
Line labeled inflows plots Equation (7). Line labeled outflows plots Equation (8). Firm i is in an economy where all other firms are in steady state.

b. *decreases with* $\Omega_{i,t}$ *when* $\Omega_{i,t}$ *is larger than* $\hat{\hat{\Omega}}$.

To understand this result, it is helpful to split the stock of knowledge into inflows and outflows. We define the additions to the data stock that are generated by time-t economic activity to be inflows (n_{it} data points, each with precision σ_ϵ^{-2}). We define the total losses due to depreciation (derived in Lemma 1) as outflows.

$$\text{Inflows:} \qquad \Omega_{it}^{+} = \sigma_\epsilon^{-2}\ z_i k_{i,t}^{\alpha} + \delta_{it} \mathbb{1}_{\delta_{i,t}>0} \sigma_\epsilon^{-2} \tag{7}$$

$$\text{Outflows:} \qquad \Omega_{it}^{-} = \Omega_{it} - \left[\rho^2 \Omega_{i,t}^{-1} + \sigma_\theta^2\right]^{-1} + \iota \delta_{it} \mathbb{1}_{\delta_{i,t}<0} \sigma_\epsilon^{-2}. \tag{8}$$

Figure 1 illustrates the inflows, outflows and dynamics of a single firm. This figure illustrates one possible economy. Data production may lie above or below the data outflow line. The difference between data inflows (solid line) and data production (dashed line) is data purchases. These purchases push the inflows line up and help speed up convergence.

The quality-adjusted production path of a single, growing firm mimics the path of its stock of knowledge. The difference between the S-shaped inflows and nearly linear outflows in Figure 1 traces out the S-shaped output path of a new entrant firm in this environment.

Firm size distribution One reason the S-shaped accumulation of data is interesting is that it implies an important role for firm size. Small firms grow slowly because they generate little data.

Only later, when they are larger and generate more data can they grow quickly. This lends itself to a bifurcated firm size distribution. There are many new firms that are stuck small and data-poor. Then, there are firms that have reached the explosive growth phase in the middle of the S-curve and grew large. In a world with increasing and then decreasing returns, firms do not remain mid-sized for long.

Single firms can have decreasing returns For some parameter values, the diminishing returns to data is always stronger than the data feedback loop. Proposition 7 in the appendix shows that, when learnable risk is abundant, knowledge accumulation is concave. In such cases, each firm's trajectory looks like the concave aggregate path in Figure 3. But the appendix describes the set of parameters that make the data feedback loop sufficiently strong, to make data inflows convex at low levels of knowledge.

3.2 New Entrant Profits, Book Value and Market Value

In a data economy, the trajectory of a single firm's profits, book value and market value are quite different from those in an economy driven by capital accumulation. Since empirical evidence on profits, book value and market value are easily available, it is useful to explore the model's predictions along these dimensions. In doing so, we relate to the literature on using Tobin's Q to measure intangible capital.

In a standard model, a young, capital-poor firm has a high marginal productivity of capital. The firm offers high returns to its owners and has a book and market value that differ only by the capital adjustment cost. In a data economy, data scarcity makes a young firm's quality and profits low. In fact, there is a range of parameters for which young firms cannot possibly make positive initial profits. Start by defining a firm's profit:

$$\text{Profit}_t = P_t A_{i,t} k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t \delta_{i,t} - r k_{i,t}. \tag{9}$$

Proposition 2 ***Negative Profits for New Entrants.*** *Assume that $g(\sigma_u^2 + \sigma_\theta^2) < 0$. Then for a firm entering with zero data, $\Omega_{i,0} = \sigma_\theta^{-2}$, the firm cannot make positive expected profit at any period t unless it has made strictly negative expected profit at some $t' < t$.*

The reason such a firm produces even though producing loses money, is that production generates data, which has future value to the firm. This firm is doing costly experimentation. This is like a bandit problem. There is value in taking risky, negative expected value actions because they generate data–active experimentation. Costly production at time t is effective payment to generate data, which will allow the firm to be profitable in the future. The reason that the firm's production loses money is that if $g(\sigma_u^2 + \sigma_\theta^2) < 0$, the initial expected quality of the firm's good is too low to earn a profit. But production in one period generates information for the next, which raises the average quality of the firm's goods, and enables future profits.

The idea that data unlocks future firm value implies that in order to increase its stock of knowledge, a new firm both produces low quality goods to self-produce data, and buys some data on the data market, as depicted in Figure 1. The two mechanisms of building stock of knowledge lead to a discrepancy between a firm's book value and market value. It is so because accounting rules do not allow a firm's book value to include data, unless that data was purchased. Therefore, we define the firm book value to be the discounted value of all purchased data. The indicator function $\mathbf{1}_{\delta_{i,t}>0}$ captures only data purchases, not self-produced data. If we equate the book value depreciation rate to the household's rate of time preference β, then

$$\text{Data Book Value}_t = \sum_{\tau=0}^{t} \beta^{t-\tau} \pi_\tau \delta_\tau \mathbf{1}_{\delta_{i,\tau}>0}. \tag{10}$$

The market value of the firm is the Bellman equation value function $V(\Omega)$ in (6). In the context of our simple model, the firm rents but does not own any capital. However, a firm without data does have value, $V(\Omega_0)$, which measures the installed value of any unmeasured assets the firm might have.[7] Therefore, to obtain the book value of a firm, we add the data book value to $V(\Omega_0)$.

Figure 2 plots the book-to-market value and profits of a young firm, over time. The ratio of the market value to the book value of a firm is used to measure intangible assets. Using a Q-theory approach and R&D expenditures, Crouzet and Eberly (2023) document that the share of intangibles in firm value rose from 17% to 29% between the late 1980's and 2015. When they include a portion of firms' overhead expenses as well, this figure nearly doubles. Our book-to-market ratio starts at 0.849 and falls to 0.697. This implies that the fraction of market value accounted for by the model's

[7]For a firm with no data, Ω_0 is not zero because zero precision would imply that the firm is infinitely uncertain. Firms always know their prior beliefs. Thus, zero data means that Ω_0 is the inverse of the prior variance of θ_0.

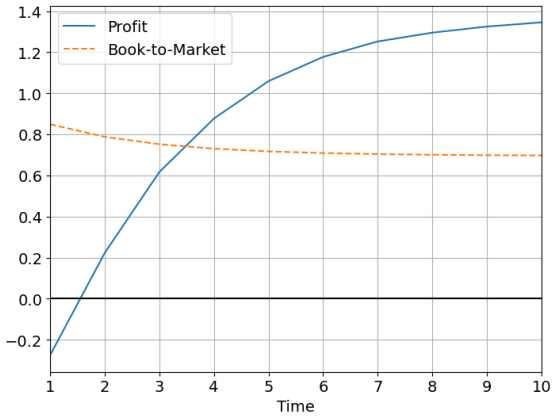

Figure 2: S-shaped growth can create initial profit losses and dampens the book-to-market ratio that follows from the missing value of data in the book value.
Book-to-market value is $(V(0) + \text{Data Book Value}_t)/V(\Omega_t)$. Data book value is defined in (10). Parameters are in Appendix B. Steady state prices of goods and data are reported as the end points of the dashed lines in Figure 3.

intangible data assets rose from 15% to 30%. While the nature of the two measures and what they capture is quite different, and other intangible assets surely account for much of firm value, data is an input for research and frequently used for product development (Babina et al., 2024). Thus, the similarity in magnitudes and rates of increase is reassuring.

The negative profits described in Proposition 2, representing costly experimentation, also show up in Figure 2, in the first period. Producing goods at a loss eventually pays off for this firm. It generates data that allows the firm to become profitable. This situation looks like Amazon at its inception. In its early days, Amazon lost $2.8 billion before turning an enormous profit.

3.3 Data Barter and Missing GDP

Data barter arises when goods are exchanged for customer data, at a zero price. While this is a knife-edge possibility in this model, it is an interesting outcome because it illustrates a phenomenon we see in reality. In many cases, digital products, like apps, are being developed at great cost to a company and then given away "for free." Free here means zero monetary price. But obtaining the app does involve giving one's data in return. That sort of exchange, with no monetary price attached, is a classic barter trade. The possibility of barter is not shocking, given the assumptions. But the result demonstrates the plausibility of the framework, by showing how it speaks to data-specific phenomena we see.

The analysis also reveals that not only are zero-price transactions, like free apps, being missed,

every transaction, in principle, has a data barter element to it. Every firm should charge slightly less for every product, because of the value of the data that accompanies its sale. In practice, a whole segment of the economy is not being captured by traditional GDP measures because the transactions price misses the value of data being paid.

Proposition 3 ***Bartering Goods for Data*** *It is possible that a firm will optimally choose positive production $k_{i,t}^{\alpha} > 0$, even if its price per unit is zero: $P_t = 0$.*

At $P_t = 0$, the marginal benefit of investment is additional data that can be sold tomorrow, at price π_{t+1}. If the price of data is sufficiently high, and/or the firm is a sufficiently productive data producer (high z_i), then the firm should engage in costly production, even at a zero goods price, to generate the accompanying data. Our framework allows us to assign a value to such barter trades and partial-barter trades, despite their zero monetary price.

These results could enable better measurement of GDP. Investment in a stock of valuable knowledge is missing from aggregate measures of economic activity. Even if we cannot observe the data-adjusted true price of a transaction, if we can measure the value of the asset being generated, we can fill in this missing value. The value of the knowledge asset generated by all this barter trade is $V(\Omega_{i,t}) - V(\Omega_{i,t-1})$, for each firm i. Typical numerical approaches to approximating a value function could be applied to $V(\Omega_{i,t})$. Alternatively, one might use revenue data, use hiring and wages of workers who maintain data stocks and work with data, or look for the covariance of a firms' choices with the random variables it needs to forecast. A detailed discussion of the myriad of approaches to measure this value function is beyond the scope of this paper. However, frameworks like this are important inputs into digital economy measurement because they guide our thinking about what is missing and how to infer this missing aggregate economic activity.

4 Long-Run Features of a Data Economy

While the previous section emphasized the contrasts, this section highlights similarities between the data economy and a capital-based production economy. Within the model, there is no long run growth because data has diminishing returns, a property documented empirically by (Bajari et al., 2019). To explore this, we describe a general class of models in which the accumulation of data

does and does not enable long-run growth. The non-rivalry of data does not sustain growth because non-rivalry simply allows something to be used by many and therefore abundant. The following results show that no matter how abundant data is, its potential is limited, unless it facilitates technological innovation.

4.1 Diminishing Returns and Zero Long Run Growth

Conceptually, data has diminishing returns because its ability to reduce variance gets smaller and smaller as beliefs become more precise. Is there some other model, without innovation, where data accumulation can sustain growth? For sustained growth to be possible, two things must both be true: 1) Perfect one-period-ahead foresight implies infinite real output; and 2) the future is a deterministic function of today's observable data.[8] Both conditions are at odds with most theories.

In order to formalize this idea, we start with two definitions.

Definition 1 (Sustainable Growth) *Let* $Y_t = \int_i \mathbb{E}[A_{i,t}]k_{i,t}^{\alpha} di$, *such that* $ln(Y_{t+1}) - ln(Y_t)$ *is the aggregate growth rate of expected output. A data economy can sustain a minimum growth rate* $\underline{g} > 0$ *if* $\exists\, T$ *such that in each period* $t > T$, $ln(Y_{t+1}) - ln(Y_t) > \underline{g}$.

The next definition, "fundamental randomness," formalizes the notion of *learnability* in the data economy. Recall that $\zeta_{i,t}$ is the set of all signals that nature draws for firm i. These are all potentially observable signals. Not all will be observed. Define Ξ_t to be the Borel σ-algebra generated by $\{\zeta_{i,t} \cup \mathcal{I}_{i,t}\}_{i=1}^{\infty}$. This is the set of all variables that can be perfectly predicted with $\mathcal{I}_{i,t}$ and time-t observable data.

Definition 2 (Fundamental Randomness) v *has time-t fundamental randomness if* $v \notin \Xi_t$.

Fundamentally random variables are simply those that are not perfectly learnable. In our model, fundamental randomness or unlearnable risk is present when $\sigma_u^2 > 0$.

We now use the the above two definitions to provide general conditions under which positive growth can be permanently sustained in the data economy.

[8] It is also true that inflow concavity comes from capital having diminishing returns. The exponent in the production function is $\alpha < 1$. But that is a separate force. Even if capital did not have diminishing marginal returns, inflows would still exhibit concavity.

Proposition 4 ***Sustainable Growth*** *In our data economy, sustainable growth requires the following two conditions to hold simultaneously*

1. *There exists a* $\underline{v}$ *such that as* $v \to \underline{v}$ *the quality function approaches infinity* $g(v) \to \infty$*; i.e., forecasts must enable infinite output.*

2. *Suppose that* $\underline{v} = 0$ *and the quality function* g *is finite almost everywhere, except at* $\underline{v} = 0$*. Productivity-relevant variables (*θ_t *and* $\varepsilon_{a,i,t}$*) have no time-*$(t-1)$ *fundamental randomness.*

The first conditions says that growth can only be sustained if $\mathbb{E}[A_{i,t}]$ can become infinite in the high-data limit. The reason is that expected aggregate output is $\int_i \mathbb{E}[A_{i,t}]k_{i,t}^{\alpha} di$. From the capital first order condition, we know that capital choice $k_{i,t}$ will be finite, as long as expected quality $\mathbb{E}[A_{i,t}]$ is finite. If output is finite, sustained growth is not possible.

If society as a whole knows tomorrow's state, they can simply produce today what they would otherwise be able to produce tomorrow. Thus, imposing finite real output at zero forecast error is a sensible assumption. But this common-sense assumption then leads to the conclusion that data has diminishing returns.

The second condition relates to the observation that realistically, not everything can be perfectly learned in the economy. Note that the assumption that g is finite-valued, except at zero, simply rules out the possibility that firms that have imperfect forecasts and still make mistakes can still achieve perfect, infinite quality. Under this assumption, the second condition asserts that even if you believe perfect one-period-ahead forecasts can produce infinite output, you still get diminishing returns because of the existence of fundamental, unlearnable randomness.

To sum up, if one believes that some events tomorrow are fundamentally random, data must have diminishing returns. Conversely, even if one believes that nothing is truly random, but they believe that with one-period ahead knowledge, an economy can only produce the finite amount today that they would otherwise produce tomorrow, then data must also have diminishing returns.

4.2 Equilibrium Price Effects

While Figure 1 represented a single firm's transition, Figure 3 illustrates the transition of a whole economy of symmetric firms, growing together. The difference between the two is the effect of equilibrium goods and data prices. When all firms are data-poor, all goods are low quality. While

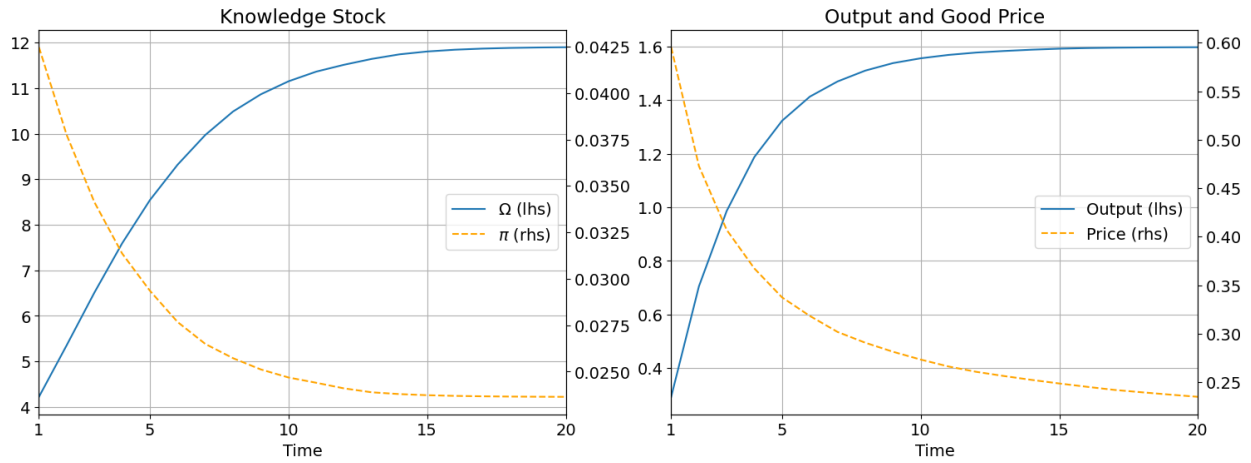

Figure 3: Aggregate Growth Dynamics: Diminishing Returns and Falling Equilibrium Prices. See Appendix B for parameters and numerical solution details.

aggregate knowledge and output exhibit growth with diminishing returns, the prices of data and goods fall, as they become more abundant. These changing prices create two equilibrium effects, both of which speed up growth. Goods prices are high initially because quality units are scarce. The high price of goods induces these firms to produce abundant goods, creating data and speeding growth. In contrast, when the single firm enters, others are already data-rich. Quality goods are abundant, so prices are low. This absence of the equilibrium price effect in the one-firm case makes it costlier and slower for the single firm to grow. The second equilibrium price effect comes from the price of data. The high initial price of scarce data also induces firms to produce more, for the purpose of generating valuable data.

The reason knowledge and output plateau in both settings is that eventually, every firms' inflows and outflows (Equations (7) and (8)) cross at the steady state. The equilibrium effects govern what happens early in the transition, when data is scarce.

4.3 Endogenous Growth

If data is used for research and development, data accumulation can sustain growth. Following a logic similar to Grossman and Helpman (1991), assume that instead of Equation (1), product quality follows a non-decreasing process:

$$A_{i,t} = A_{i,t-1} + \max\{0, \Delta A_{i,t}\} \qquad \text{with} \qquad \Delta A_{i,t} = g\left((a_{i,t} - \theta_t - \epsilon_{a,i,t})^2\right).$$

The solution inherits the same structure as before: the expected change in quality of firm i's good at time t, $\mathbb{E}[\Delta A_{i,t}|\mathcal{I}_{i,t}]$, can be approximated by $\left(\Omega_{i,t}^{-1} + \sigma_u^2\right)$. The interpretation is that more data allows for more precisely targeted innovations, which increase the size of the technology advance. An illustrative example is when $g\left((a_{i,t} - \theta_t - \epsilon_{a,i,t})^2\right) = \bar{A} - (a_{i,t} - \theta_t - \epsilon_{a,i,t})^2$. With this formulation, depending on $\bar{A}$, more data can make the innovation viable: $\mathbb{E}[\Delta A_{i,t}|\mathcal{I}_{i,t}] > 0$. A similar structure with multiplicative $\Delta A_{i,t}$ could sustain exponential growth.

This extension teaches us that data used for research should be measured separately from data used for other purposes, just like economists typically do for capital expenditures. Of course, for this formulation to make sense, one needs to believe that information resulting from transactions can be used to discover growth-sustaining technologies.

5 Welfare and Data Externalities

Before now, our framework lacked two important features needed to assess welfare and consider optimal policy. The first is micro-foundations for demand, which reveal consumer utility. The other feature is a negative externality of data. Incorporating these assumptions justify the previous model by delivering the same inverse demand as in (3). They also reveal that the only source of inefficiency is the data externality.

5.1 A Micro-founded Model for Welfare Analysis

Consider an economy with two goods: a numeraire good, m_t and a retail good c_t, that is produced using capital and data. Let P_t denote the price of the retail good in terms of the numeraire.

Households There is a continuum of homogeneous infinitely lived households, with quasi-linear preferences over consumption of the retail good c_t and the numeraire good m_t. Households have CRRA utility for retail good consumption: $u(c_t) = \bar{P}\frac{c_t^{1-\gamma}}{1-\gamma}$. The representative household's optimization problem is

$$\max_{c_t, m_t} \sum_{t=0}^{+\infty} \frac{u(c_t) + m_t}{(1+r)^t} \qquad \text{s.t.} \quad P_t c_t + m_t = \Phi_t \qquad \forall t \tag{11}$$

The budget constraint equates consumption expenditures on the two goods to household income,

which is firm profits Φ_t. Since aggregate output is non-random, as argued earlier, aggregate profits and the optimization problem are also not random, within each period t.

Retail Good Production The producers of the retail goods live forever. They use capital, rented at rate r, trade data, and produce the retail good using their capital and data. There are two types of retail firms. They are identical, except for their, z_i, the efficiency with which they produce data. We consider a measure λ of low data-productivity firms with $z_i = z_L$, and a measure $(1-\lambda)$ of high data-productivity firms with $z_i = z_H$, where $z_L < z_H$.

Profit is revenue minus adjustment costs, minus data costs (if $\delta > 0$) or plus revenue from data sales (if $\delta < 0$), minus the cost of capital, $\Phi_{it} := P_t A_{i,t} k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t \delta_{i,t} - r k_{i,t}$. The profit the households get is the aggregate firm profit,

$$\Phi_t = \int \Phi_{it} di = P_t \int_i A_{i,t} k_{i,t}^{\alpha} di - \int_i \Psi(\Delta\Omega_{i,t+1}) di - r \int_i k_{i,t} di,$$

Firms maximize the expected present discounted value of their profit:

$$\max_{\{k_{i,t},\delta_{i,t}\}_{t=0}^{\infty}} V(\Omega_{i,0}) = \sum_{t=0}^{+\infty} \frac{1}{(1+r)^t} \left(P_t \mathbb{E}[A_{i,t}|\mathcal{I}_{i,t}] k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi \delta_{i,t} - r k_{i,t} \right). \tag{12}$$

Data governs the expected quality of goods, $\mathbb{E}[A_{i,t}]$. To simplify the exposition, we use the following specification for $g(.)$ in Equation (1):

$$A_{i,t} = \bar{A} - \left(a_{i,t} - \theta_t - \epsilon_{a,i,t} \right)^2. \tag{13}$$

The law of motion for data is expressed in Equation (5).

The retail sector represents an industry where consumption and data are industry-specific, but capital is rented from an inter-industry market, at rate r, paid in units of numeraire.[9]

[9]Equivalently, we can interpret this as a small, open economy where capital and numeraire goods are tradeable and retail goods are non-tradeable. The world rental rate of capital is r. This simplification puts the focus on data. An endogenously determined rental rate of capital would increase when firms are more productive. This would create a wealth effect for capital owners. These equilibrium effects are well-studied in previous frameworks, but are not related to economics of data.

Equilibrium We restrict our attention to economies with λ, z_H and z_L such that there exists a symmetric, pure-strategy equilibrium, where all firms of the same type make the same choices; if $z_i = z_j$, then $\delta_{i,t} = \delta_{j,t}$ and $k_{i,t} = k_{j,t}$ $\forall t$. An equilibrium is household choices of c_t and m_t that maximize (11), firm choices of capital $k_{i,t}$ and data $\delta_{i,t}$ that maximize (12) and prices P_t and π_t that clear markets (they satisfy aggregate resource constaints):

$$\begin{aligned}
&\text{Retail good :} && c_t = \lambda A_{L,t} k_{L,t}^{\alpha} + (1-\lambda) A_{H,t} k_{H,t}^{\alpha}, \\
&\text{Numeraire good :} && m_t + r\left(\lambda k_{L,t} + (1-\lambda) k_{H,t}\right) + \Big(\lambda \Psi(\Delta\Omega_{L,t+1}) + (1-\lambda)\Psi(\Delta\Omega_{H,t+1})\Big) = 0 \\
&\text{Data :} && \lambda \delta_{L,t} + (1-\lambda)\delta_{H,t} = 0.
\end{aligned}$$

Proposition 5 ***Welfare*** *The steady state allocation is socially efficient.*

Equilibrium capital investment and data production are efficient because there are no externalities. The constraint, that data may only be produced through the production of goods, is a constraint that is faced both by the planner and the firm. Prices of goods and data reflect their marginal social value. This aligns the private and social incentives for production.

5.2 Data for Business Stealing

When data can be used for marketing or other forms of business stealing, firms' use of data harms others. Using data for business stealing can be represented through a quality externality:

$$A_{i,t} = \bar{A} - \left(a_{i,t} - \theta_t - \epsilon_{a,i,t}\right)^2 - b \int_{j=0}^{1} \left(a_{j,t} - \theta_t - \epsilon_{a,j,t}\right)^2 dj \qquad \text{for } b \in [0,1] \tag{14}$$

Notice that the business stealing externality does not change firms' choices because it does not enter in a firm's first order condition.[10] Therefore, it does not change data inflows, outflows, data sales or capital choices, at a given set of prices. However, it does influence aggregate good quality. The baseline model is represented by $b = 0$. In this case, Equations (14) and (13) are identical and there is no externality.

[10] To see why this is the case, note that firm i's actions have a negligible effect on the average productivity term $\int_{j=0}^{1} \left(a_{j,t} - \theta_t - \epsilon_{a,j,t}\right)^2 dj$. So the derivative of that new externality term with respect to i's choice variables is zero. If the term is zero in the first order condition, it means it has no effect on choices of the firm. This formulation of the externality is inspired by Morris and Shin (2002).

If $b > 0$, this captures the idea that when one firm uses data to market effectively, it reduces the ability of all other firms to generate value by reaching their preferred customers. The extreme case where data does not have any social value is $b = 1$. The aggregate losses from business stealing entirely cancel out the productivity gains from data: $\int A_{i,t} di = \bar{A}$.

Proposition 6 ***Welfare with Business Stealing*** *If $b > 0$, the steady state features over-investment in capital and excessive data production.*

Proposition 6 incorporates two distinct inefficiencies: excessive output production and excessive data production. Higher output has an externality, which is greater data production. This is a negative externality because more data reduces the quality of other firms' goods.

6 GDP Mis-measurement

Building macroeconomic frameworks enables measurement. We calibrate the model and use it to estimate the magnitude of GDP mis-measurement that arises from data barter. The total amount by which real goods and services are under-priced is equal to the value of the data transferred from customers to firms in that year. The data itself is a business input. It is the final goods that are under-valued by GDP because the price paid is the price of the good, net of the value of the data firms rebate to consumers, in the form of a discount. Our calibration suggests that GDP should be 3-6% higher annually in 2003-2018 due to the missing value of transactions implicitly paid by the data exchanged.[11]

6.1 Externally Calibrated Parameters

The model has 8 parameters: α, γ, ρ, σ_θ^2, ψ_t (a time-series), $\bar{P}$, $\bar{A}$, and s_Ω. We calibrate the first five parameters externally, either directly from the literature or using procedures suggested in previous work. After describing these external parameters, the next section calibrates the last three parameters, $\bar{P}$, $\bar{A}$ and s_Ω, to jointly match three moments to endogenous model outputs.

The externally calibrated parameters are: Capital share of income α, inverse demand elasticity γ, parameters of the AR(1) process for TFP, ρ and σ_θ^2, and marginal adjustment cost of data ψ_t. A capital share of 40% is used in many papers, including the handbook of macroeconomics.

[11] Appendix B provides a detailed description of the measurement procedure and more extensive results.

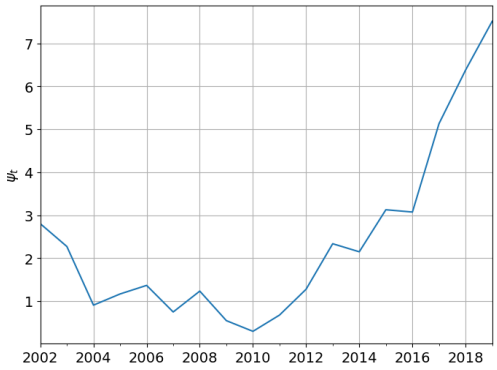

Figure 4: Calibrated adjustment cost series ψ_t using the methodology in Brynjolfsson et al. (2021).

The demand elasticity parameter γ is the exponent on quantity in the price function. In a consumption-based, dynamic model, this curvature parameter γ is also the inverse of the elasticity of intertemporal substitution (IES). Guvenen (2006) survey measures of the IES and report that, if one wants to match the macro evidence with one IES value, then a value that fits the evidence well is 1.07. That implies $\gamma = 1/1.07 = 0.93$. For our purposes, this parameter is not very important. Since it mostly just governs the price level, changing elasticity causes a re-calibration the price scaling parameter $\bar{P}$ to match the same moment, delivering the same results. We have explored a value ten times smaller and, after re-calibrating $\bar{P}$, obtained results that are visually indistinguishable.

The persistence and innovation variance of the optimal technique process, ρ and σ_θ^2, come from fitting an AR(1) to the productivity process estimated by Fernald (2014)[12] The argument is not that technique and productivity are the same, but rather that a major source of changes in technique might be technological and thus the processes would have similar properties.

To estimate the data adjustment costs ψ_t, we follow the procedure in Brynjolfsson et al. (2021). The estimated cost is a coefficient from a cross-sectional regression of the market value of a firm on its R&D expenses (with firm fixed effects and controlling for overhead costs). Brynjolfsson et al. (2021) estimate the adjustment cost annually for their sample. We extend their estimation through 2018.[13] The argument for why this measures AI costs is that, according to q-theory, incurring

[12]The updated dataset is available here: https://www.johnfernald.net/TFP. We use the observations that correspond to our sample period 2003-2018.

[13]The packaged implementation of this exercise uses two public inputs: `Brynjolfsson2021.csv`, a Compustat-based file constructed following Brynjolfsson et al. (2021), and `IPO-age.csv`, drawn from the original replication materials archived through openICPSR (openICPSR, 2023).

an investment cost should only increase firm value at the margin, if there is some unmeasured adjustment cost that prevents the firm from investing more. Figure 4 plots the calibrated data adjustment cost series.

6.2 Indirectly Calibrating Three Parameters

There are three parameters left to estimate using model equations: The maximum product quality, $\bar{A}$, the marginal effect of forecast errors on product quality, s_Ω, and the multiplier that determines the level of goods price, $\bar{P}$. We use the model to match three moments: 1) mean-squared error between realized and model estimated time-series of real US GDP during 2003-2018, 2) BEA estimate of firm investment in own-account data assets in 2003, 3) sensitivity of capital investments with respect to uncertainty, reported by Gorodnichenko et al. (2023).

First, we need to express the model in steady state and in nominal dollars, to be comparable with our data moments. Assume that the data adjustment cost takes a quadratic form, $\Psi(\Delta\Omega_{t+1}) = \psi_t(\frac{\Omega_{t+1}-\Omega_t}{\Omega_t})^2$, while the quality function is a linear approximation to the quality function in Equation (1):

$$g(\Omega_t^{-1}) = \bar{A} - s_\Omega(\Omega_t^{-1})$$

We assume $\iota = 1$ to calibrate the model so that in steady state there is no data trade. Thus, the nominal value function in Equation (6) simplifies to:

$$V(\Omega_t) = \max_{k_t} \; P_t(\bar{A} - s_\Omega(\Omega_t^{-1} + \sigma_u^2))k_t^\alpha - \Psi(\Delta\Omega_{t+1}) - r_t k_t + \left(\frac{1}{1+r_t}\right) V(\Omega_{t+1}),$$

Finally, we convert the nominal value function to a real one, $\bar{V}_t = \frac{V_t}{P_t}$. The Bellman equation in real terms can be written as

$$\bar{V}(\Omega_t) = \max_{k_t} \; (\bar{A} - s_\Omega(\Omega_t^{-1} + \sigma_u^2))k_t^\alpha - \frac{1}{P_t}\Psi(\Delta\Omega_{t+1}) + r_t k_t + \left(\frac{1}{1+r_t}\right) \bar{V}(\Omega_{t+1}), \qquad (15)$$

To determine the price level of goods P_t, we cumulate the inflation rates between dates $\tau = 0, \cdots, t$, using $P_t = P_0 \prod_{\tau=0}^{t-1}(1+\dot{p}_\tau)$, where $\dot{p}_t$ is the time-t inflation rate and $P_0 = \bar{P}((\bar{A} - s_\Omega(\Omega_0^{-1} + \sigma_u^2))k_0^\alpha)^{-\gamma}$ is the date $t = 0$ price index.

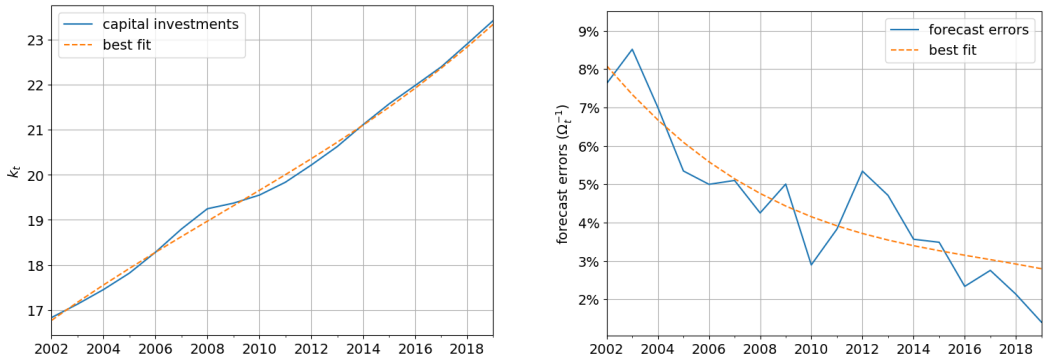

Figure 5: Calibration targets: Capital Stock and Forecast Errors. The left panel is the net stock of nonresidential fixed assets from BEA's Fixed Assets Accounts (U.S. Bureau of Economic Analysis, 2023). The right panel is the time-series of US public firm forecast errors calculated using data from I/B/E/S earnings guidance via WRDS (LSEG Data & Analytics, 2023), following Asriyan and Kohlhas (2024). Forecast errors are winsorized at the one percent level on the right tail.

The first order condition with respect to the capital choice k_t is

$$\alpha(\bar{A} - s_\Omega(\Omega_t^{-1} + \sigma_u^2))k_t^{\alpha-1} - r_t + \frac{d\Omega_{t+1}}{dk_t}\left[\frac{1}{1+r_t}\bar{V}'(\Omega_{t+1}) - \frac{2\psi_t}{P_t}\left(\frac{\Omega_{t+1}-\Omega_t}{\Omega_t^2}\right)\right] = 0. \tag{16}$$

We use three data series to perform the calibration— the time-series for capital, inflation, and firms' sales forecast error. For capital, k_t, we use the net stock of nonresidential fixed assets from BEA during 2003-2018.[14]. The left panel of Figure 5 depicts this time-series and its best affine fit. For inflation, $\dot{p}_t$, we use the percent change in FRED's Gross Domestic Product: Implicit Price Deflator (GDPDEF) (Federal Reserve Bank of St. Louis, 2023b).[15]

The stock of knowledge, Ω_t, is the conditional precision of firms' forecasts about the learnable component of their optimal technique θ_t. Their precision is the inverse of the firms' expected squared forecast error. The technique uncertainty, Ω_t^1 also has an unlearnable component σ_u^2. Taken together, $\Omega_t^1 + \sigma_u^2$, are the only source of uncertainty in firms' revenues. Therefore, technique uncertainty $\Omega_t^1 + \sigma_u^2$ is proportional to revenue uncertainty, as measured by the expected squared error of a firm's revenue forecast. Since the two types of uncertainty enter additively throughout the model, we calibrate their sum, but do not need to decompose the two.

To compute firm forecast errors, we follow Asriyan and Kohlhas (2024) and use the data from I/B/E/S Guidance via WRDS (LSEG Data & Analytics, 2023) to measure sales forecast accuracy of

[14]U.S. Bureau of Economic Analysis (U.S. Bureau of Economic Analysis, 2023), Table 1.2. Chain-Type Quantity Indexes for Net Stock of Fixed Assets and Consumer Durable Goods, Line 4

[15]Inflation is calculated using the data reported in https://fred.stlouisfed.org/series/GDPDEF.

US public firms. I/B/E/S Guidance extracts quantitative company expectations from press releases and transcripts of corporate events. Asriyan and Kohlhas (2024) show that forecast accuracy is correlated with features of firms that ought to predict their information choices. We use data of US firms between 2002-2021 because the sales guidance data only became available for more than 10 firms after 2001. When the sales forecast is expressed as a range, we take the mid-point (the average of the lower bound and upper bound). We express all the sales numbers in 2002 dollars, deflating with the Consumer Price Index for All Urban Consumers (CPIAUCSL) monthly series from FRED (Federal Reserve Bank of St. Louis, 2023a). We compare firms' sales guidance with realized sales data from Compustat North America (S&P Global, 2023) and define the squared relative forecast error as:

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

In each year, we take the average of across firms. The right panel of Figure 5 plots the squared relative forecast error, averaged across all firms, in each year, as well as its exponential trend.

Of course, a decline in forecast error could come from a decline in firms' earnings volatility, rather than more data. However, the variance of earnings innovations is not trending down.[16] Declining forecast errors could also be a result of better data technology. Technology improvements do not distort our measure of data value. Better technology raises the value of data. We measure one unit of data as however much data improves forecast precision by one unit. Whether that precision comes from 10 bits or 1 bit of underlying data does not matter for valuing the one unit improvement in precision.

We further need the prevailing capital rental rate from the data to calibrate the model. In order to get that, we use the net stock of nonresidential fixed assets from BEA Fixed Assets Accounts (U.S. Bureau of Economic Analysis, 2023) during 2003-2018 to determine the capital rental rate r_t that rationalize the observed capital stock. The left panel of Figure 6 compares the model-implied r_t with empirical estimates of weighted average cost of capital, implying that the general equilibrium capital rental rates backed out from the model are plausible.[17]

[16] When we estimate a 5-year variance of each firm's earnings and then average across the sample of firms that are also in the forecast data, we found a positive slope over time. This held with and without fixed effects. This rising earnings volatility means that, if anything, we are understating the rise of data.

[17] Thanks to Kurt Mitman for suggesting this exercise.

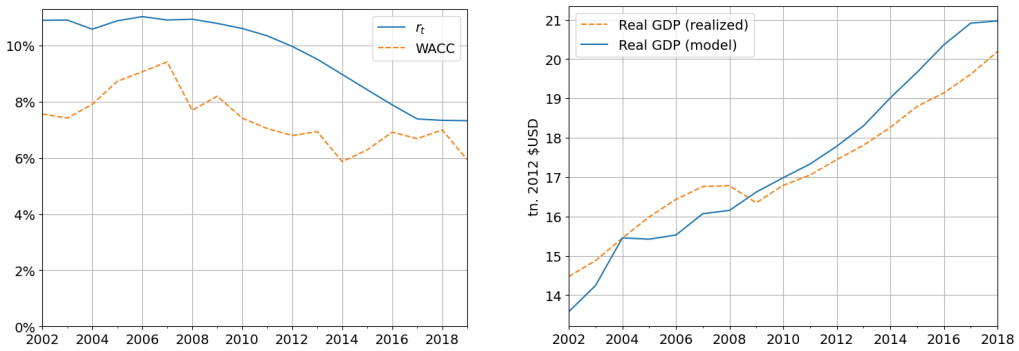

Figure 6: Interest and Output: The left panel depicts the model-implied and data series for cost of capital. The right panel depicts the analogous two series for real GDP.

As our first moment, we use the time-series of real GDP for the 2003-2018 period from FRED (Federal Reserve Bank of St. Louis, 2023c). We minimize the mean-squared error (MSE) between realized and estimated real GDP. Our model implied real GDP is $\max_{k_t} \; g(\Omega_t^{-1})k_t^{\alpha} - \frac{1}{P_t}\Psi(\Omega_{t+1}) - r_t k_t$. We minimize the MSE between the model-implied and realized real GDP across the sample period. The fit to annual real GDP is shown in Figure 6.

As our second moment, we match the one-period value of data to the BEA estimate. The BEA estimated that firm investments in own-account data assets had a flow value of $72bn in 2003 (0.25% of GDP) (Calderón and Rassier, 2022).[18] We match this to the model's one-period value of data, as a fraction of the model GDP: $opv(\Delta\Omega_t)/\text{GDP}_{2003}$ and derived below in (18). The model's estimated one-period values of data are shown in the left panel of Figure 7.

As our third and last moment, we match the sensitivity of capital investments with respect to uncertainty. Gorodnichenko et al. (2023) estimate that a one percentage point increase in macroeconomic uncertainty results in a 7.5% decrease in firm capital investments. We increase Ω^{-1} by 1pp and use Equation 16 to solve for k_t^* under this increased uncertainty. We want $\frac{k_t^*}{k_t}$ to imply a 7.5% decrease in optimal capital investment given a 1pp increase in uncertainty in 2018, which is close to the date of their survey.

Table 1 reports the data series used for both external and indirect calibration, as well as the values of the calibrated parameters.

[18] See https://www.bea.gov/system/files/2022-05/BEA-ACM-Data-Assets-Presentation-05132022.pdf.

Model object	Data series	
Ω_t^{-1}	Sales forecast errors for US public firms, following Asriyan and Kohlhas (2024) using data from I/B/E/S Guidance via WRDS (LSEG Data & Analytics, 2023)	
k_t	BEA real net stock of fixed assets (U.S. Bureau of Economic Analysis, 2023)	
$\dot{p}_t$	Inflation: Gross Domestic Product Price Deflator (Federal Reserve Bank of St. Louis, 2023b)	
Parameter	Description (Target)	Value / Range
α	Capital share of income	0.4
γ	Inverse demand elasticity (Guvenen, 2006)	0.93
ρ, σ_θ^2	AR(1) coefficients from TFP (Fernald, 2014)	0.98, 0.0026
ψ_t	Data adjustment cost (Brynjolfsson et al., 2021).	0.5-7.5
$\bar{P}$	Price level of goods (model moment)	5.04
$\bar{A}$, s_Ω	Quality function intercept and slope (model moments)	1.18, 1.90

Table 1: Model calibration targets.

6.3 Estimating the Value of Data

To measure the uncounted GDP that arises from data barter, we need to know the value of all the data consumers transferred to firms in a year. In order to measure the present value of the data generated in a year, we construct a counter-factual value function without one year's worth of new data. We introduce an unexpected loss of the new data that the representative firm acquires in a single year. If a firm receives no new data in a period, then their stock of knowledge in the next period is the depreciated current stock: $\tilde{\Omega}_{t+1} = \frac{\Omega_t}{\rho^2 + \sigma_\theta^2 \Omega_t} = (1 - \delta_t^o)\Omega_t$. This loss of knowledge stock changes firm value going forward. Let $\tilde{V}$ denote the firm value function without time-t generated data:

$$\tilde{V}(\Omega_t) = \max_{k_t} g(\Omega_t) k_t^\alpha - \frac{1}{P_t}\Psi(\tilde{\Omega}_{t+1} - \Omega_t) - r_t k_t + \frac{1}{1+r_t}\bar{V}(\tilde{\Omega}_{t+1}) \tag{17}$$

Let $\tilde{V}(\Omega_t)$, $pdv(\Delta\Omega_t) = \bar{V}(\Omega_t) - \tilde{V}(\Omega_t)$ denote the difference between the actual real value of data, $\bar{V}(\Omega_t)$, and this counter-factual value. It represents the net present discounted value of the bartered data acquired in period t.

The share of economic value which comes from bartered data goods at time-t is the present discounted value, $pdv(\Delta\Omega_t)/GDP_t$, because this is the value of the data asset transferred from customers to firms.

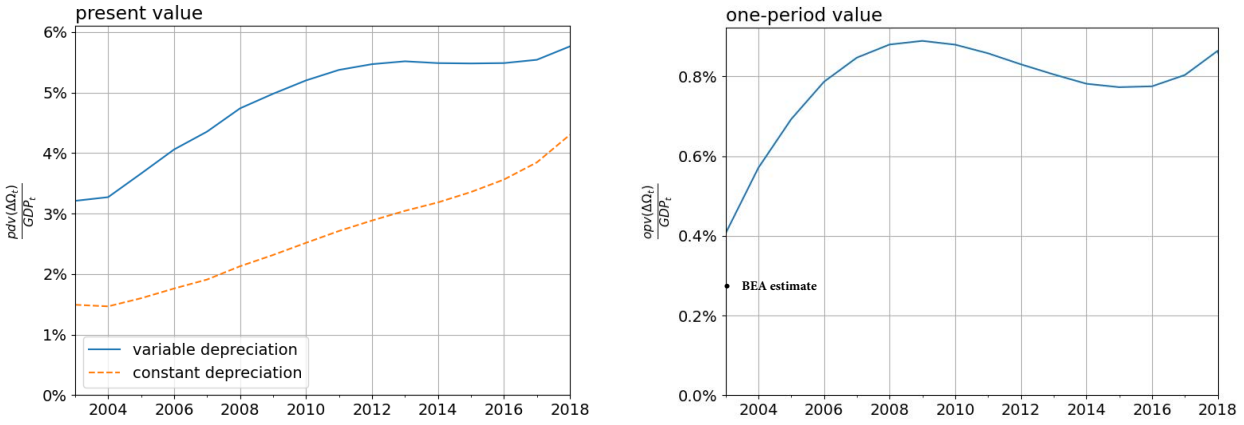

Figure 7: Estimated GDP mis-measurement that comes from bartered data– present value (left) and one-period value (right) of data generated each year. Solid line reports the model-implied value. Dashed line reports the present value of data that depreciates at the constant 6.6% rate, dictated by GAAP accounting rules.

The left panel of Figure 7 reports the share of economic value due to bartered goods between 2003 and 2018. Our calculations suggest that GDP should be 3-6% higher because the value of a transaction is measured by the value of the payment and the data exchanged. While the present value of data is much higher than its one-period value, the rate of growth of that present value is slower. While the one-period value of data has quadrupled in the last two decades, the present value has doubled. This discrepancy is largely because abundant data depreciates faster. The constant depreciation value starts lower, but rises at a similar rate to the one-period value. Firms need to accumulate more and more data to gain a small improvement in their predictions. This illustrates the importance of the model's explanation for how data depreciates. Because it is information, Bayes' law tells us that data depreciates in a way that is fundamentally different from capital. That difference is evident in the left panel of Figure 7.

Alternatively, the value of data within a single period, $opv(\Delta\Omega_t)$, follows the same logic, except that two period from today (at $t+2$) the stock of data has to revert to its original value:

$$opv(\Delta\Omega_t) = pdv(\Delta\Omega_t) - \frac{1}{1+r_t}\bar{V}(\tilde{\Omega}_{t+1}) \tag{18}$$
$$+ \frac{1}{1+r_t}[\max_{k_{t+1}} A(\tilde{\Omega}_{t+1})k_{t+1}^{\alpha} - \frac{\Psi(\tilde{\Omega}_{t+1}-\Omega_{t+2})}{P_{t+1}} - r_{t+1}k_{t+1} + \frac{1}{1+r_{t+1}}\bar{V}(\Omega_{t+2})]$$

The right panel of Figure 7 illustrates this one-period discounted value.

To put this value in context, consider the predicted value of GDP in two extremes of the data economy. If no firms had any data, GDP would be 12% lower in 2003 and 22% lower by 2018. In contrast, if firms had infinite data, GDP would be almost 30% higher in 2003, but only 12% higher by 2018.

6.4 Sensitivity

Our estimates are not very sensitive to the demand elasticity parameter γ and the maximum productivity $\bar{A}$. If one changes one of these values, but then re-calibrates the remaining endogenous moments, the predictions look similar.

Appendix B explores the role of three parameters or moments that play a more important role: the interest (time discount) rate, the data adjustment cost and the one-period data value. The value of data is highly sensitive to the time discount rate when data is scarce. This can be seen from the fact that the 2003 one-period value of data on the right of Figure 7 is about a tenth of the 2003 present value of data assets on the left. If future payoffs were highly discounted, the asset values would be more similar to the one-period values. Most of the value of scarce data is in its ability to generate future value and future data. This future value is interest-rate sensitive. When current-period data revenue is more abundant, the interest rate matters less. The Appendix shows that even doubling the interest rate in the middle of the simulation affects data value by less than 1% of GDP.

The data adjustment cost ψ_t is important to have and its level effects the magnitude of estimate data values. However, data adjustments have surprisingly little effect on the trajectory of data values.

The BEA's one-period estimate of data value is important to examine because it is likely noisy. Fortunately, while it determines the initial value of data, that effect fades almost entirely within the first 5 years. For the long-run present value of data, the key moments are the investment sensitivity to data, s_Ω, and the stock of data Ω_t.

The appendix further explores the interest rate and the predicted rate of data growth as over-identifying moments that reveal the model to have plausible predictions.

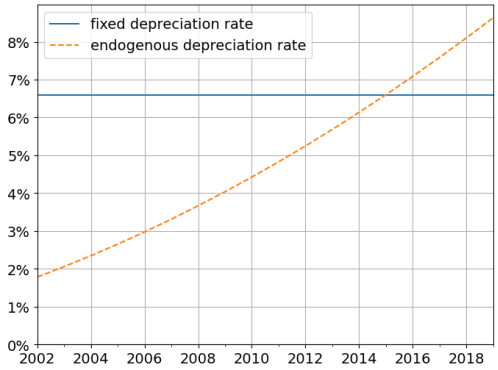

Figure 8: Model-implied versus constant annual data depreciation rate.

The Importance of Data Depreciation One of the contributions of the paper was to derive a depreciation rate that is not constant over time, but varies with the stock of data. Our last exercise uses the qualitative model to show the importance of this depreciation measure. According to GAAP accounting rules, intangible assets like data and software should be amortized over 15 years. That is a 6.6% rate of depreciation per year. Figure 7 shows that using the constant 6.6% depreciation rate, instead of the model-implied rate results in a valuation for data that is about 50% too low.

7 Industry or Firm-Specific Data and Product Innovation

For simplicity, we started with a tracking problem with only one random variable θ_t to forecast. However, firms learn about industry, input-specific or firm-specific conditions as well. When a firm has many attributes to learn about, they not only choose how much to produce, but also choose what to produce. They use data to do product design and innovation. An extension of the model to N dimensions can capture such problems, without losing any of the tractability of the original model.

Model setup Consider N products whose profits depend on N attributes. These attributes could be related to cost and optimal operations. They could be related to fads and fashion, or they could represent dimensions of worker skills and human resources decisions a firm must make. For each attribute, there is a optimal action: the best supplier of a material, a hottest color, the optimal degree of quant versus verbal skill than a manager should have. For attribute k, this

optimal choice is the kth entry of the $N \times 1$ vector $\theta_t + \epsilon_{it}$. The $N \times 1$ state θ_t follows the AR(1) process $\theta_t = \bar{\theta} + \rho(\theta_{t-1} - \bar{\theta}) + \eta_t$. The $N \times 1$ innovation vector $\eta_t \sim N(0, \Sigma_\theta)$ is $i.i.d.$ across time. The innovations are independent across attributes. In other words, Σ_θ is a diagonal matrix.[19] Firms have a noisy prior about the realization of θ_0. The transitory $N \times 1$ shock $\epsilon_{a,i,t} \sim N(0, \sigma_u^2 I)$ is $i.i.d.$ across time and firms and is unlearnable.

Firms use data for product innovation and design. After observing and analyzing their data, they choose a location in the product space, represented by the $N \times 1$ vector x_{it}. The jth entry of x_{it} reports the weight firm i's product places on attribute j. The quality of firm i's product is then $x_{it}' A_{it}$. To have a distinct notion of quantity and product location, we normalize the sum of weights x to one: $x_{it}' \mathbb{1}_N = 1$.

In order to add richness and still see the mechanisms clearly, we simplify. From here on, we assume that the production technology for goods has an Ak structure ($\alpha = 1$). To focus on product choice, we shut down data markets ($\iota = 1$). Since data is no longer traded, we replace the adjustment cost of data with a quadratic investment cost rk_{it}^2 to keep the problem concave ($\psi(\cdot) = 0$). Finally, the quality of attribute j produced by firm i at time t is the jth entry of the vector $A_{it} = \bar{A} - (a_{it} - \theta_t - \epsilon_{it}) \odot (a_{it} - \theta_t - \epsilon_{it})$, where $\odot$ denotes the Hadamard product (element-by-element multiplication). This quality expression represents the same squared loss function as in the univariate case.

Firms get data about the optimal attribute production technique, for every attribute they produce. They get more data about attributes their good loads on more heavily. The effective number of data points a firm sees about each attribute is the vector $n_{it}' = z_i k_{it} x_{it}'$. As before, each data point has precision σ_ϵ^{-2}.

Equilibrium The equilibrium price of each attribute P_t depends on the aggregate supply of that attribute. As before, $P_t = \bar{P} Y_t^{-\gamma}$. Y_t is an $N \times 1$ vector of the equilibrium prices and supply of each of the N attributes:

$$Y_t = \int_i (x_{it} \odot A_{i,t}) k_{i,t} di$$

[19] This is without loss of generality. If attributes have correlated innovations, we could construct a new linear combination of goods that does have independent innovations and call that the attributes. For example, we might think of attributes as the principal components of the variance of shocks.

The price of the good that firm i produces is the linear combination of its attributes and the price of each attribute, $x'_{it}P_t$.

Firms update beliefs with Bayes' law. The evolution of the stock of knowledge is the same as in the uni-variate problem, but with a vector-matrix representation:

$$\Omega_{i,t+1} = \left[\rho^2\Omega_{i,t}^{-1} + \Sigma_\theta\right]^{-1} + n_{i,t}\sigma_\epsilon^{-2}.$$

The sequential problem of a firm can be expressed recursively in terms of firm i's data and an approximate sufficient statistic of other firms' data $\bar{\Omega}_t$:

$$V(\Omega_{it}, \bar{\Omega}_t) = \max_{x_{it}, k_{it}} x'_{it}(P_t \odot E[A_{it}|\mathcal{I}])k_{it} - rk_{it}^2 + \left(\frac{1}{1+r}\right) V(\Omega_{t+1}, \bar{\Omega}_{t+1}).$$

First, solve for the firm's joint choice of quantity and product location: $\tilde{x}_{it} := x_{it}k_{it}$, using the first order condition:

$$\tilde{x}_{it} = \frac{1}{2r}\left[P_t \odot E[A_{it}|\mathcal{I}] + \left(\frac{1}{1+r}\right)\frac{\partial V(\Omega_{t+1}, \bar{\Omega}_{t+1})}{\partial \Omega_{i,t+1}}\sigma_\epsilon^{-2}\right].$$

To recover product design and quantity separately, recognize that if the elements of x_{it} must sum to one, then $k_{it} = \tilde{x}'_{it}\mathbb{1}$ is the sum of the entries of $\tilde{x}_{it}$ and the product choice is $x_{it} = \tilde{x}_{it}/k_{it}$.

This problem is separable in attributes because attributes are defined as dimensions with independent shocks and independent data. Thus, the choice of x is simply N parallel choices of the single-state model we described at the start.

8 Conclusion

The economics of transactions data bears some resemblance to technology and some to capital. It is not identical to either. Data has the diminishing returns of capital, in the long run. But it has the increasing returns of ideas and technologies, early in the transition path to steady state. Data generated from economic activity also changes firms' choices of production over their life-cycle. Thus, while the accumulation and analysis of data may be the hallmark of the "new economy," this new economy has many economic forces at work that are old and familiar.

We conclude with future research possibilities that our framework could enable.

Firm size dispersion. One of the biggest questions in macroeconomics and industrial organization is: What is the source of the bifurcation in firm size? As Section 3.1 explains, one possible source is the accumulation of data. Future work might quantify this effect.

Firm competition. Instead of assuming price taking behavior, one could model a finite number of firms that consider the price impact of their production decisions. Firms' data affect the how they compete (Eeckhout and Veldkamp, 2025). Alternatively, a monopolist may price discriminate (Farboodi et al., 2024). Placing these mechanisms in a recursive setting like this one, could give us insights about how data changes firms' dynamic competitive strategies.

Investment and data. The fixed data productivity parameter z_i represents the idea that certain industries will spin off more data than others. A firm could invest in collecting and analyzing the data by choosing its data processing technology, z_i, at a cost. In such a problem, endogenizing the interest rate with a capital market would be central. That would enlarge the state space, but would also reveal equilibrium interactions between investment and data accumulation.

Optimal data policy. A benevolent government might adopt a data policy to promote the growth of small and mid-size firms. The policy solution to increasing return-growth traps is typically a form of big push investment. In the context of data investment, the government could collect data itself, from taxes or reporting requirements, and share it with firms. For example, China shares data with some firms, in a way that seems to facilitate their growth (Beraja et al., 2023). Alternatively, the government might facilitate or promote data sharing among firms or act to prevent data from being exported to foreign firms.

This simple framework enables research on many data-related phenomena. It can be a foundation for thinking about many more.

The data and code underlying this research is available on Zenodo at https://doi.org/10.5281/zenodo.18378460

References

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li, "A Theory of Falling Growth and Rising Rents," *Review of Economic Studies*, *90* (6), 2675–2702.

Agrawal, Ajay, John McHale, and Alex Oettl, "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 149–174.

_ , Joshua Gans, and Avi Goldfarb, *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*, Harvard Business Press, 2022.

Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan, "Signaling in a Global Game: Coordination and Policy Traps," *Journal of Political Economy*, 2006, *114* (3), 452–484.

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

Atkeson, Andrew and Patrick J Kehoe, "Modeling and Measuring Organization Capital," *Journal of political Economy*, 2005, *113* (5), 1026–1053.

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

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, "The Impact of Big Data on Firm Performance: An Empirical Investigation," *AEA Papers and Proceedings*, 2019, *109*, 33–37.

Beraja, Martin, David Y. Yang, and Noam Yuchtman, "Data-intensive Innovation and the State: Evidence from AI Firms in China," *Review of Economic Studies*, 2023, *90* (4), 1701–1723.

Bergemann, Dirk and Juuso Välimäki, "Experimentation in Markets," *The Review of Economic Studies*, 04 2000, *67* (2), 213–234.

Broer, Tobias, Alexandre Kohlhas, Kurt Mitman, and Kathrin Schlafmann, "Expectation and Wealth Heterogeneity in the Macroeconomy," *Working paper*, 2025.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, January 2021, *13* (1), 333–72.

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.

Caplin, Andrew and John Leahy, "Business as Usual, Market Crashes, and Wisdom After the Fact," *American Economic Review*, 1994, *84* (3), 548–565.

Chahrour, Ryan, Kristoffer Nimark, and Stefan Pitschner, "Sectoral Media Focus and Aggregate Fluctuations," *American Economic Review*, 2021, *111* (12), 3872–3922.

Cong, Lin William, Danxia Xie, and Longtian Zhang, "Knowledge Accumulation, Privacy, and Growth in a Data Economy," *Management Science*, 2021, *67* (10), 5969–6627.

_ , Wenshi Wei, Danxia Xie, and Longtian Zhang, "Endogenous Growth Under Multiple Uses of Data," *Journal of Economic Dynamics & Control*, 2022, *141.*

Crouzet, Nicolas and Janice Eberly, "Rents and Intangible Capital: A Q+ Framework," *The Journal of Finance*, 2023, *78* (4), 1873–1916.

Damodaran, Aswath, "Total U.S. Market Weighted Average Cost of Capital," 2024. Accessed April 2024.

Eeckhout, Jan and Laura Veldkamp, "Data and Markups: A Macro-Finance Perspective," 2025. NBER Working Paper.

Fajgelbaum, Pablo D., Edouard Schaal, and Mathieu Taschereau-Dumouchel, "Uncertainty Traps," *The Quarterly Journal of Economics*, 2017, *132* (4), 1641–1692.

Farboodi, Maryam, Nima Haghpanah, and Ali Shourideh, "Price Discrimination: Who Benefits from the Data?," 2024. Working Paper.

_ , **Roxana Mihet, Thomas Philippon, and Laura Veldkamp**, "Big Data and Firm Dynamics," *American Economic Association Papers and Proceedings*, May 2019.

Federal Reserve Bank of St. Louis, "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average, series CPIAUCSL," 2023. Accessed August 2023.

_ , "Gross Domestic Product: Implicit Price Deflator, series GDPDEF," 2023. Accessed August 2023.

_ , "Gross Domestic Product, series GDPA," 2023. Accessed August 2023.

Fernald, John G., "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity," Technical Report 2012-19, Federal Reserve Bank of San Francisco April 2014.

Garicano, Luis and Esteban Rossi-Hansberg, "Organizing growth," *Journal of Economic Theory*, 2012, *147* (2), 623–656.

Gorodnichenko, Yuriy, Saten Kumar, and Olivier Coibion, "The Effect of Macroeconomic Uncertainty on Firm Decisions," *Econometrica*, 2023, *91* (4), 1297–1332.

Grossman, Gene M and Elhanan Helpman, "Quality ladders in the theory of growth," *The review of economic studies*, 1991, *58* (1), 43–61.

Guvenen, Fatih, "Reconciling conflicting evidence on the elasticity of intertemporal substitution: A macroeconomic perspective," *Journal of Monetary Economics*, 2006, *53* (7), 1451–1472.

Ilut, Cosmin and Martin Schneider, "Ambiguous Business Cycles," *American Economic Review*, August 2014, *104* (8), 2368–99.

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

Jovanovic, Boyan and Yaw Nyarko, "Learning by Doing and the Choice of Technology," *Econometrica*, 1996, *64* (6), 1299–1310.

Lorenzoni, Guido, "A Theory of Demand Shocks," *American Economic Review*, December 2009, *99* (5), 2050–84.

LSEG Data & Analytics, "I/B/E/S Academic Database," 2023. Accessed via WRDS on April 10, 2023.

Maćkowiak, Bartosz and Mirko Wiederholt, "Optimal sticky prices under rational inattention," *American Economic Review*, 2009, *99 (3)*, 769–803.

Mankiw, N. Gregory and Ricardo Reis, "Sticky Information in General Equilibrium," *Journal of the European Economic Association*, 2007, *5* (2-3), 603–613.

Matějka, Filip and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, January 2015, *105* (1), 272–98.

Morris, Stephen and Hyun Song Shin, "Social value of public information," *The American Economic Review*, 2002, *92* (5), 1521–1534.

Nimark, Kristoffer P. and Stefan Pitschner, "News media and delegated information choice," *Journal of Economic Theory*, 2019, *181*, 160–196.

Oberfield, Ezra and Venky Venkateswaran, "Expertise and Firm Dynamics," 2018 Meeting Papers 1132, Society for Economic Dynamics 2018.

openICPSR, "Brynjolfsson et al. replication materials source for IPO-age.csv," 2023. Accessed July 27, 2023.

Ordonez, Guillermo, "The Asymmetric Effects of Financial Frictions," *Journal of Political Economy*, 2013, *121* (5), 844–895.

Solow, Robert M., "A Contribution to the Theory of Economic Growth," *The Quarterly Journal of Economics*, 02 1956, *70* (1), 65–94.

S&P Global, "Compustat North America Fundamentals Annual Database," 2023. Accessed via WRDS on April 10, 2023.

U.S. Bureau of Economic Analysis, "Fixed Assets Accounts, Table 1.2: Chain-Type Quantity Indexes for Net Stock of Fixed Assets and Consumer Durable Goods," 2023. Accessed August 2023.

Veldkamp, Laura, "Slow Boom, Sudden Crash," *Journal of Economic Theory*, 2005, *124(2)*, 230–257.

Wilson, Robert, "Informational Economies of Scale," *Bell Journal of Economics*, 1975, *6*, 184–95.