# A Model of the Data Economy

Maryam Farboodi[*] and Laura Veldkamp[†]

July 9, 2024

**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 US 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. 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-date 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 take away 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 calibrates the model using a combination of macroeconomic and financial moments and uses the it to measure 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 by the consumer data acquired by the firms. 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. (2021) or rational attention choices in Maćkowiak and Wiederholt (2009), Matějka and McKay (2015) or Reis (2008). 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., 2019; 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 (2018), Cong et al. (2021) and Cong et al. (2020), 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. (2018) 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., 2019), 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 $\theta_t$ 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 $\theta_0$. 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 $\theta$ 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 $\theta$ 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 chair 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 $\theta_t$. 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.

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 $k_{i,t}^\alpha$:

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

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 $A_{i,t}$.[2] 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 $\theta_t$ 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 $\theta_{t-1}$ 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 $\theta_{t-1}$, but not on any data from before $t$. All past data is just a noisy signal about $\theta_{t-1}$, which the firm now knows. Thus preventing the revelation of $\theta_{t-1}$ 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 $\pi_t$.

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 $\iota$ 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 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

[2] 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.

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[3] $\mathcal{I}_{i,t} = \{\mathcal{I}_{i,t-1}, \{s_{i,t-1,m}\}_{m=1}^{\omega_{i,t-1}}, A_{i,t-1}\}$, where $\omega_{i,t-1}$ is the net number of data points added (or subtracted if $\omega$ 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 $\theta_t$, 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 $\theta_t$. 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 $\theta$, 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 $\theta_t$.

Our data adjustment cost $\Psi$ 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.

[3] 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 $\theta_t$ 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:

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

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.[4] The data price $\pi_t$ 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 $\theta_{t+1}$. 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

[4] Appendix A shows that, because there are infinitely many firms with independent signals and a noisy prior, independent forecast errors imply independence in $A_{i,t}$'s 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, as long as $\iota$ of knowledge is lost, each time a firm sells a data point.

*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 $\theta_t$ is $V[\theta_t|\mathcal{I}_t] := \Omega_t^{-1}$, where $\Omega_t$ 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 $\Omega$ 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 $\Omega_t$. Thus, the depreciation rate, the equivalent of $\delta^k$ 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 $\sigma_\theta^2$), when there is lots of knowledge to begin with (high $\Omega_t$), and when high persistence makes the state a more variable process (high $\rho$).[5]

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.

[5] 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 $\iota$ 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 $\theta_t$ 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 $\theta_t$.

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 $\theta_t$ 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 $\delta$ 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 $\theta_t$.

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 on the 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 asssumption 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 $\Psi$ is sufficiently low, $\bar{P}$ is sufficiently high, and the second derivative of the value function is bounded $V'' \in [\nu, 0)$; and*

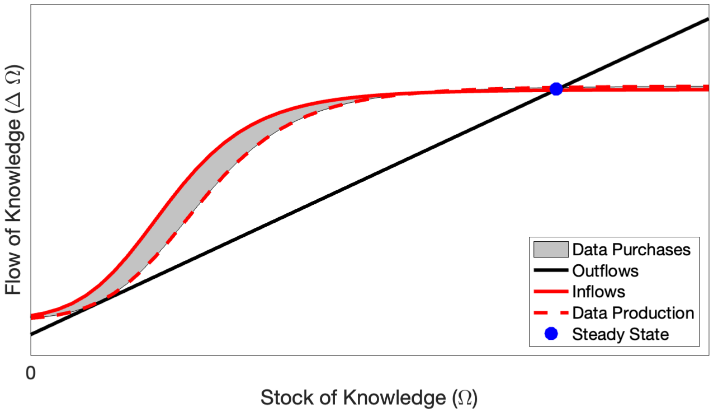

Figure 1: A single new firm grows slowly: Inflows and outflows of one firm's data.
Line labeled inflows plots Equation (7). Line labeled outflows plots (8). Firm $i$ is in an economy where all other firms are in steady state.

**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 $\sigma_\epsilon^{-2}$). 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 $\beta$, 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, $V(0)$, which measures the installed value of any unmeasured assets the firm might have. Therefore, to obtain the book value of a firm, we add the data book value to $V(0)$.

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, Crouzet and Eberly (2021) document that the share of intangibles in firm value rose from 23% to 29% between the late 1980's and 2017. 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 intangible assets not counted for in the firm's accounting/book value rose from 15% to 30% in the model.

The negative profits described in Proposition 2, representing costly experimentation, also show

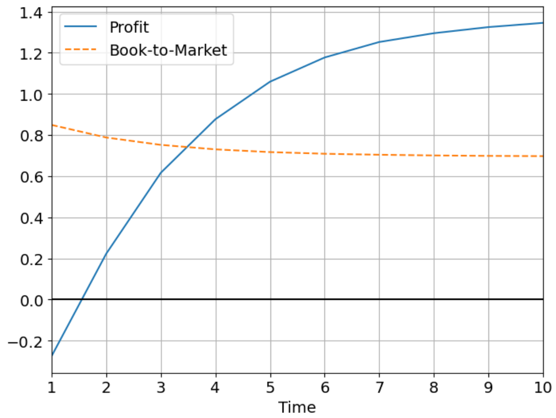

Figure 2: S-shaped growth can create initial profit losses and dampens the book-to-market ratio that follows from the missing value of data in the book value.
Book-to-market value is $(V(0) + \text{Data Book Value}_t)/V(\Omega_t)$. Data book value is defined in (10). Parameters are in Appendix B. Steady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

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 $\pi_{t+1}$. 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., 2018). 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.[6] 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 $\Xi_t$ to be the Borel $\sigma$-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 $v \not\in \Xi_t$.*

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.

**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.*

[6] 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.

2. *Suppose that $\underline{v} = 0$ and the quality function $g$ is finite almost everywhere, except at $\underline{v} = 0$. Productivity-relevant variables ($\theta_t$ 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 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

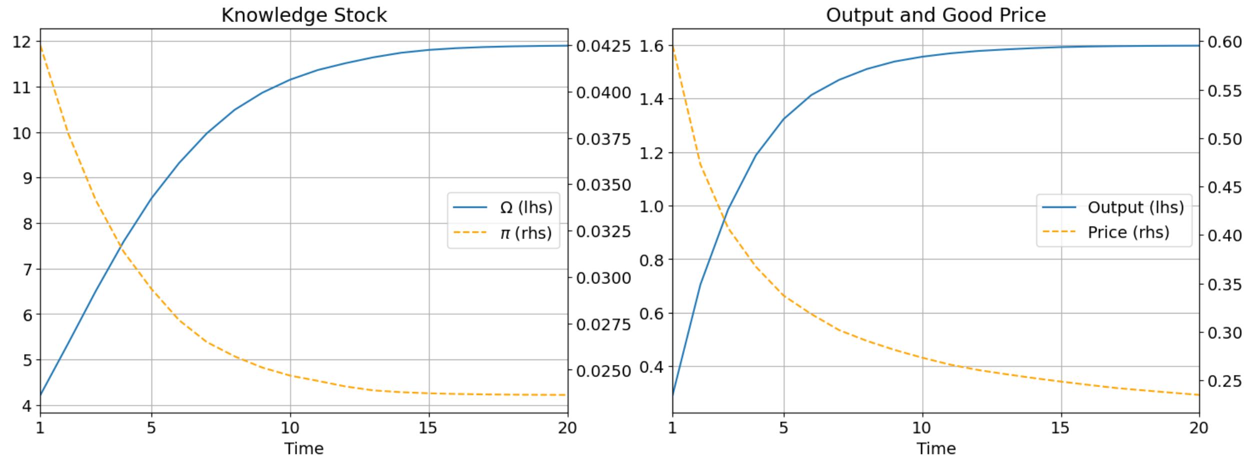

Figure 3: Aggregate Growth Dynamics: Diminishing Returns and Falling Equilibrium Prices. See Appendix B for parameters and numerical solution details.

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. We consider a symmetric firm environment. Since not all firms can buy (or all sell) data, this implies no data trade.

### 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 $\Phi_t$. 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 $\lambda$ 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.[7]

**Equilibrium** We restrict our attention to economies with $\lambda$, $z_H$ 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 $\pi_t$

[7] 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.

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:

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

Notice that the business stealing externality does not change firms' choices because it does not enter in a firm's first order condition.[8] 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.

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: $\int A_{i,t} di = \bar{A}$.

[8] To see why this is the case, note that firm $i$'s actions have a negligible effect on the average productivity term $\int_{j=0}^{1}\big(a_{j,t} - \theta_t - \epsilon_{a,j,t}\big)^2 dj$. 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).

**Proposition 6** ***Welfare with Business Stealing*** *If $b > 0$, there is over-investment in the steady state level of capital and excessive trade in the data market in equilibrium.*

Proposition 6 incorporates two distinct inefficiencies: excessive production and excessive data trade. Higher data production and sales reduces the quality of other firms' goods. Thus, in equilibrium, too much output is produced and too much data is traded.

The idea that firms sell too much data might appear counter-factual, since social networks and search engines do not primarily sell data directly. Instead, they use their data primarily to sell data services to their business customers. For example, Facebook revenue comes primarily from advertising, which is a data service. However, sales of data services is a type of data sales. A formal analysis of the equivalence between data services and data sales is in Admati and Pfleiderer (1990).

## 6 GDP Mis-measurement: Data Barter

A key purpose of building macroeconomic frameworks is to enable measurement. We calibrate the model and use it to estimate the magnitude of GDP mis-measurement that arises from data barter. For this calibration, we consider a linear approximation to the quality function

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

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. For brevity, we summarize the calibration and results in this section. Appendix B provides a detailed description of the measurement procedure and more extensive results.

**Calibration** Table 1 reports the externally calibrated parameter values, the data series we used for matching model implied moments, as well as the parameters calibrated using the model. The first block constitutes the data series that we use.

Since our objective is to value data, not to explain where it comes from, we feed in a measure of firms' data to the model, the US Public firm sales forecast errors from I/B/E/S database. That measure is the forecast errors firms make when reporting their sales revenue. Since firms only

| Model object | Data series | |
|---|---|---|
| $\Omega_t^{-1}$ | Sales forecast errors for US public firms, following Kohlhas and Asriyan (2024) using data from I/B/E/S Guidance | |
| $k_t$ | BEA real net stock of fixed assets | |
| $\dot{p}_t$ | Inflation: Gross Domestic Product Price Deflator | |
| Parameter | Description (Target) | Value / Range |
| $\alpha$ | Capital share of income | 0.4 |
| $\gamma$ | Inverse demand elasticity (Guvenen, 2006) | 0.93 |
| $\rho$, $\sigma_\theta^2$ | AR(1) coefficients from TFP (Fernald, 2014) | 0.98, 0.0026 |
| $\psi_t$ | Data adjustment cost (Brynjolfsson et al., 2021). | 0.5-7.5 |
| $\bar{P}$ | Price level of goods (model moment) | 5.04 |
| $\bar{A}$, $s_\Omega$ | Quality function intercept and slope (model moments) | 1.18, 1.90 |

Table 1: Model calibration targets. See Appendix B for details.

revenue uncertainty in the model comes from uncertainty about the state $\theta_t$, we link sales forecast errors to the conditional variance of firms' $\theta_{t+1}$ forecast at time $t$, which maps directly into an amount of data.[9]

We use the real stock of fixed assets $k_t$, to impute a capital return series $r_t$ that rationalizes the observed capital stock in our model.[10] Matching capital ensures that errors in the estimates of the data value as a fraction of GDP come only from the data part of the economy, not from the capital part. We use GDP price deflator for to computer the inflation series.

The second block in Table 1 reports the calibrated parameters. The first three rows, $\alpha,\ \gamma,\ \rho,\ \sigma_\theta^2$, report the parameters that are are informed by the literature, along with the source and values of each one. The next row corresponds to the data adjustment cost series $\psi_t$. Brynjolfsson et al. (2021) estimates this series using q-theory to impute an adjustment cost for intangible assets related to R&D. We extend their estimate until the end of our time interval. The time series is reported in the appendix.

Finally, $\bar{P}$, $\bar{A}$ and $s_\Omega$ are parameters that we calibrate using moments from the models and the data series reported in the first block of Table 1. Note that these three parameters are key parameters for valuing data: $\bar{P}$ governs the price level of goods, and $\bar{A}$ and $s_\Omega$ govern the $g(.)$ function that maps forecast precision into output quality. We jointly estimate $\bar{P}$, $\bar{A}$ and $s_\Omega$ to

[9] An alternative exercise is to start the model with an initial amount of data and let it predict how much data firms accumulate and value that accumulated data. That exercise is more relevant if one wants to predict the future evolution of the data economy. Figure 3 reports the endogenous data outcome from the calibrated model.

[10] Of course, capital returns are endogenous. There would be a capital friction or wedge that we could calibrate to match the capital stock, in each period, at the equilibrium rental rate. But such an exercise would require much more computation for little gain in a model that is not designed to speak to the physical capital stock.

match three moments of the data: 1) GDP series, 2) BEA reported estimate of 2003 value of firms own data, and 3) sensitive of capital investments to data provision reported by Gorodnichenko et al. (2023).

The GDP series allows our price of goods to have the right scale so that when we report dollar values, they are relative to a GDP level that is plausible. The rationale for matching a 2003 estimate of the one-period value of data value is to get data value on the right scale, but give the model freedom to predict the evolution over time. That value estimate, by the Bureau of Economic Analysis, is based on the cost of accumulating, processing and maintaining a firm's data. Finally, the investment sensitivity to data comes from a randomized control trial by Gorodnichenko et al. (2023). Their experiment treats some firms with data and measures their investment reaction. In our model, the quality function $g$ governs this sensitivity.

**Data value estimates** 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. This is the payment customers made to firms, above and beyond the monetary price. To value only the data generated in one period, we construct a counter-factual value function $\tilde{V}(\Omega_t)$ that has all the same firm data, except for the $n_t$ data points generated in a single period $t$ (Equation (43)). The present discounted value of data generated in a period $pdv(\Delta\Omega_t)$ is used to pay for goods and services, but is uncounted by GDP. This missing economic value is the difference between firm value with time-$t$ data and without: $pdv(\Delta\Omega_t) = V(\Omega_t) - \tilde{V}(\Omega_t)$, accounting for changes in the aggregate price level.

Figure 4 reports this present discounted value as a percentage of GDP 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

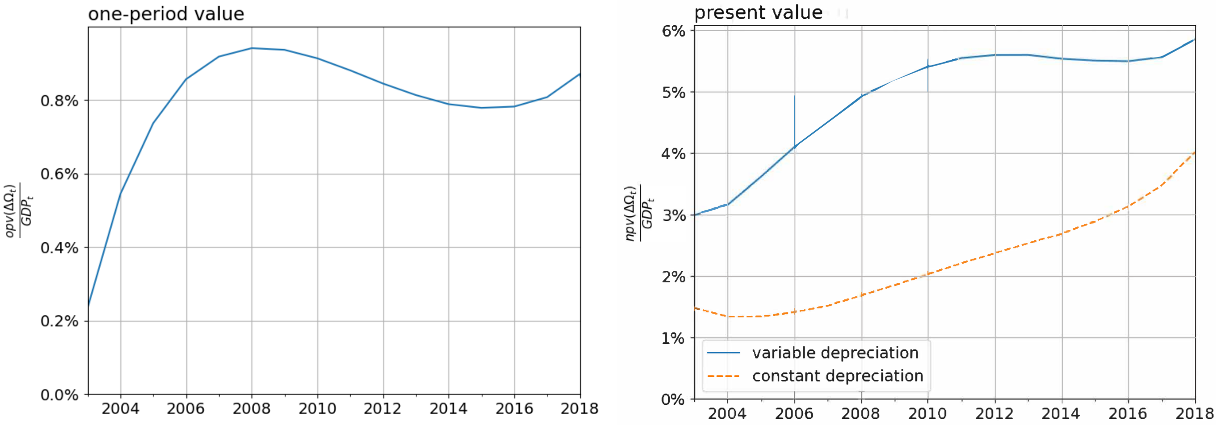

Figure 4: Estimated GDP mis-measurement that comes from bartered data. One-period value (left) and present value (right) of data generated each year. Solid line reports the model-implied value. Dashed line reports the present value of data that depreciates at the constant 6.6% rate, dictated by GAAP accounting rules.

evident in the right side of Figure 4.

**Sensitivity** As the previous discussion shows, the value of data is highly sensitive to the discount rate. The estimates are not very sensitive to $\gamma$ and Y. Appendix B shows that they are also remarkably insensitive to the interest rate $r$. Even doubling the interest rate in the middle of the simulation produces only a modest response in data value. The adjustment cost is important to have, but its precise level and its fluctuations over time make little difference for the calculation.

The one-period estimate of data value is also likely noisy. Fortunately, while it determines the initial present of data, that effect fades within the first 5 years. For the long-run present value of data, the key moments are the investment sensitivity to data and the amount of data.

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

## 7 Industry or Firm-Specific Data and Product Innovation

For simplicity, we started with a tracking problem with only one random variable $\theta_t$ 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 $k$th entry of the $N \times 1$ vector $\theta_t + \epsilon_{it}$. The $N \times 1$ state $\theta_t$ 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, $\Sigma_\theta$ is a diagonal matrix.[11] Firms have a noisy prior about the realization of $\theta_0$. 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 $j$th 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 $j$th 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

[11] 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.

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 $\sigma_\epsilon^{-2}$.

**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:

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

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 $\bar{\Omega}_t$:

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

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, 2022). 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 in AI and data processing technology.* 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.

*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., 2020). 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.

# References

**Admati, Anat and Paul Pfleiderer**, "Direct and Indirect Sale of Information," *Econometrica*, 1990, *58* (4), 901–928.

**Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li**, "A theory of falling growth and rising rents," Technical Report, National Bureau of Economic Research 2019.

**Agrawal, Ajay, John McHale, and Alexander Oettl**, "Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth," in "The Economics of Artificial Intelligence: An Agenda," National Bureau of Economic Research, Inc, 2018.

**_ , 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.

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

**Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki**, "The Impact of Big Data on Firm Performance: An Empirical Investigation," Working Paper 24334, National Bureau of Economic Research February 2018.

**Beraja, Martin, David Y. Yang, and Noam Yuchtman**, "Data-intensive Innovation and the State: Evidence from AI Firms in China," 2020. MIT Working Paper.

**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**, "Information and Wealth Heterogeneity in the Macroeconomy," *Working paper*, 2021.

**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.

**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," 2019. Working Paper, Boston College.

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

**Cong, Lin William, Wenshi Wei, Danxia Xie, and Longtian Zhang**, "Endogenous Growth Under Multiple Uses of Data," 2020.

**Crouzet, Nicolas and Janice C Eberly**, "Rents and Intangible Capital: A Q+ Framework," Working Paper 28988, National Bureau of Economic Research July 2021.

**Eeckhout, Jan and Laura Veldkamp**, "Data and Market Power," 2022. 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.

**Fernald, John**, "A quarterly, utilization-adjusted series on total factor productivity," in "in" Federal Reserve Bank of San Francisco 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, Chad and Chris Tonetti**, "Nonrivalry and the Economics of Data," 2018. Stanford GSB Working Paper.

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

**Kohlhas, Alexandre N. and Vladimir Asriyan**, "Firm Expectations and Misallocation," 2024. Working Paper, University of Oxford.

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

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

**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.

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

**Reis, Ricardo**, "A Sticky-Information General Equilibrium Model for Policy Analysis," Working Papers Central Bank of Chile 495, Central Bank of Chile October 2008.

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

**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.

# A Appendix

## A.1 Proof of Lemma 1: Belief Updating

The information problem of firm $i$ about its optimal technique $\theta_{i,t}$ can be expressed as a Kalman filtering system, with a 2-by-1 observation equation. We start by describing the Kalman system, and show that the sequence of conditional variances is deterministic. Note that all the variables are firm specific, but since the information problem is solved firm-by-firm, for brevity we suppress the dependence on firm index $i$.

*Belief updating.* At time, $t$, the firm takes as given its last-period beliefs, $\widehat{\mu}_{t-1} = \mathbb{E}\big[\theta_{t-1} \mid \mathcal{I}_{i,t-1}\big]$ and $\Omega_{t-1} = Var\big[\theta_{t-1} \mid \mathcal{I}_{i,t-1}\big]^{-1}$.

Next, use the law of motion $\theta_{t+1} = \bar{\theta} + \rho(\theta_t - \bar{\theta}) + \eta_{t+1}$ and take the expectation on both sides of the equation to get: $\mathbb{E}\big[\theta_t \mid \mathcal{I}_{t-1}\big] = \bar{\theta} + \rho \cdot \big(\mathbb{E}\big[\theta_{t-1} \mid \mathcal{I}_{t-1}\big] - \bar{\theta}\big)$. If we take the variance of both sides of the equation, we get $V\big[\theta_t \mid \mathcal{I}_{t-1}\big] = \rho^2\Omega_{t-1}^{-1} + \sigma_\theta^2$.

*Data points* are not depreciated like $\Omega$ because they contain information directly about next period's state $\theta_{t+1}$. Here, we introduce a new piece of notation: the number of new data points added to the firm's data set. $\omega_{i,t}$. For firms that do not trade data, this is $\omega_{i,t} = n_{i,t} = zk_{i,t}^\alpha$. More generally, the number of new data points depends on the amount of data traded:

$$\omega_{i,t} := n_{i,t} + \delta_{i,t}(\mathbf{1}_{\delta_{i,t}>0} + \iota\mathbf{1}_{\delta_{i,t}<0}).$$

The set of signals $\{s_{t,m}\}_{m\in[1:\omega_{i,t}]}$ are informationally equivalent to a single average signal $\bar{s}_t$ such that $\bar{s}_t = \theta_{t+1} + \epsilon_{s,t}$, where $\epsilon_{s,t} \sim \mathcal{N}(0, \sigma_\epsilon^2/\omega_{it})$.

Of course, at the end of date $t$, each firm observes a signal derived from *observing their own output quality* $A$. However, we have assumed that this information is included in the $n_{it}$ signals generated by output. We do recognize that this signal from own quality is not normal. But we assume that the distribution of the other output information complements this information to make the composite information normally distributed. In earlier versions of the paper, we modeled the inference from $A$ separately. It complicates the problem, without providing any additional insight. Results available upon request.

Then, the final step is to use the mean and variance above as prior beliefs and use Bayes law to update them with the average signal $\bar{s}_t$:

$$\widehat{\mu}_t = \mathbb{E}\big[\theta_t \mid \mathcal{I}_t\big] = \frac{\Big[\rho^2\Omega_{t-1}^{-1} + \sigma_\theta^2\Big]^{-1} \cdot \mathbb{E}\big[\theta_t \mid \mathcal{I}_{t-1}\big] + \omega_t\sigma_\epsilon^{-2}\bar{s}_t}{\Big[\rho^2\Omega_{t-1}^{-1} + \sigma_\theta^2\Big]^{-1} + \omega_t\sigma_\epsilon^{-2}} \tag{15}$$

$$\Omega_t^{-1} = Var\big[\theta_t \mid \mathcal{I}_t\big] = \Big\{\Big[\rho^2\Omega_{t-1}^{-1} + \sigma_\theta^2\Big]^{-1} + \omega_t\sigma_\epsilon^{-2}\Big\}^{-1}. \tag{16}$$

Equations (15) and (16) constitute the Kalman filter describing the firm dynamic information problem. Importantly, note that $\Omega_t$ is deterministic.

## A.2 Proof of Lemma 2: Making the Problem Recursive

*Lemma.* The sequence problem of the firm can be solved as a non-stochastic recursive problem with one state variable. Consider the firm sequential problem:

$$\max_{k_t,a_t,delta_t} \sum_{t=0}^{\infty} \left(\frac{1}{1+r}\right)^t \mathbb{E}\left[P_tA_tk_t^\alpha - \Psi(\Delta\Omega_{i,t+1}) - \pi_t\delta_{i,t} - rk_t|\mathcal{I}_{i,t}\right]$$

We can take a first order condition with respect to $a_t$ and get that at any date $t$ and for any level of $k_t$, the optimal choice of technique is

$$a_t^* = \mathbb{E}[\theta_t|\mathcal{I}_t].$$

Given the choice of $a_t$'s, using the law of iterated expectations, we have:

$$\mathbb{E}[(a_t - \theta_t - \epsilon_{a,t})^2|\mathcal{I}_s] = \mathbb{E}[Var[\theta_t + \epsilon_{a,t}|\mathcal{I}_t]|\mathcal{I}_s] = \mathbb{E}[Var[\theta_t|\mathcal{I}_t]|\mathcal{I}_s] + \sigma_u^2,$$

for any date $s \leq t$. We will show that this object is not stochastic and therefore is the same for any information set that does not contain the realization of $\theta_t$.

We can restate the sequence problem recursively. Let us define the value function as:

$$V(\{s_{t,m}\}_{m\in[1:\omega_t]}, y_{t-1}, \hat{\mu}_{t-1}, \Omega_{t-1}) =$$
$$\max_{k_t,a_t,delta_t} \mathbb{E}\left[P_t A_t k_t^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t\delta_{i,t} - rk_t + \left(\frac{1}{1+r}\right) V(\{s_{t+1,m}\}_{m\in[1:\omega_{t+1}]}, y_t, \hat{\mu}_t, \Omega_t)|\mathcal{I}_{i,t}\right]$$

with $\omega_{i,t}$ being the net amount of data being added to the data stock. Taking a first order condition with respect to the technique choice conditional on $\mathcal{I}_t$ reveals that the optimal technique is $a_t^* = \mathbb{E}[\theta_t|I_t]$. We can substitute the optimal choice of $a_t$ into $A_t$ and rewrite the value function as

$$V(\{s_{t,m}\}_{m\in[1:\omega_t]}, y_{t-1}, \hat{\mu}_{t-1}, \Omega_{t-1}) = \max_{k_t,\delta_t} \mathbb{E}\Big[P_t g\big((\mathbb{E}[\theta_t|I_{i,t}] - \theta_t - \epsilon_{a,t})^2\big)k_t^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t\delta_{i,t} - rk_t$$
$$+ \left(\frac{1}{1+r}\right) V(\{s_{t+1,m}\}_{m\in[1:\omega_{t+1}]}, y_t, \hat{\mu}_t, \Omega_t)|\mathcal{I}_{i,t}\Big].$$

Note that $\epsilon_{a,t}$ is orthogonal to all other signals and shocks and has a zero mean. Thus,

$$\mathbb{E}\Big[(\mathbb{E}[\theta_t|I_t] - \theta_t - \epsilon_{a,t})^2\Big] = \mathbb{E}\Big[(\mathbb{E}[\theta_t|I_{i,t}] - \theta_t)^2\Big] + \sigma_u^2 = \Omega_{i,t}^{-1} + \sigma_u^2$$

$\mathbb{E}[(\mathbb{E}[\theta_t|I_t] - \theta_t)^2|I_{i,t}]$ is the time-$t$ conditional (posterior) variance of $\theta_t$, and the posterior variance of beliefs is $\mathbb{E}[(\mathbb{E}[\theta_t|I_t] - \theta_t)^2] := \Omega_t^{-1}$. Expected productivity determines the within period expected payoff, which using Equation (4) depends on posterior variance. The posterior variance $\Omega_t^{-1}$ is given by the Kalman system Equation (16), which depends only on $\Omega_{t-1}$, $n_t$, and other known parameters. It does not depend on the realization of the data. Thus, $\{s_{t,m}\}_{m\in[1:\omega_t]}, y_{t-1}, \hat{\mu}_t$ do not appear on the right side of the value function equation; they are only relevant for determining the optimal action $a_t$. Therefore, we can rewrite the value function as:

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

Next, we do a change of variables and optimize not over the amount of data purchased or sold $\delta_{i,t}$, but rather the closely related, net change in the data stock $\omega_{i,t}$. We also substitute in $n_{i,t} = z_i k_{i,t}^{\alpha}$ and substitute in the optimal choice of technique $a_{i,t}$. The problem becomes

$$V(\Omega_{i,t}) = \max_{k_{i,t},\omega_{i,t}} \; P_t g\left(\bar{A} - \Omega_{i,t}^{-1} - \sigma_u^2\right) k_{i,t}^{\alpha} - \pi\left(\frac{\omega_{i,t} - z_i k_{i,t}^{\alpha}}{\mathbf{1}_{\omega_{i,t}>n_{i,t}} + \iota\mathbf{1}_{\omega_{i,t}<n_{i,t}}}\right) - rk_{i,t}$$
$$- \Psi\left(\Delta\Omega_{i,t+1}\right) + \left(\frac{1}{1+r}\right) V(\Omega_{i,t+1})$$

$$\text{where} \quad \Omega_{i,t+1} = \left[\rho^2\Omega_{i,t}^{-1} + \sigma_\theta^2\right]^{-1} + \omega_{i,t}\sigma_\epsilon^{-2}$$

Since $\frac{\partial\Omega_{i,t,t+1}}{\partial\omega_{i,t}} = \sigma_\epsilon^{-2}$, the first order condition for the optimal $\omega_{i,t}$ is

$$FOC[\omega_{i,t}]: \quad -\Psi'(\cdot)\sigma_\epsilon^{-2} - \tilde{\pi} + \left(\frac{1}{1+r}\right) V'(\Omega_{i,t+1})\sigma_\epsilon^{-2} = 0$$

where $\tilde{\pi} \equiv \pi/(\mathbf{1}_{\omega_{i,t}>n_{i,t}} + \iota\mathbf{1}_{\omega_{i,t}<n_{i,t}})$ is the price of data, adjusted for non-rivalry. It is lower for data sales since less data is lost per unit of data sold.

## A.3 Lemma 3, 4, 5: Linearity of Data Depreciation

One property of the model that comes up in a few different places is that the depreciation of knowledge (outflows) is approximately a linear function of the stock of knowledge $\Omega_{i,t}$. There are a few different ways to establish this approximation formally. The three results that follow show that the approximation error from a linear function is small i) when the stock of knowledge is small; ii) when the state is not very volatile; and iii) when the stock of knowledge is large.

**Lemma 3** ***Linear Data Outflow with Low Knowledge*** *$\exists\epsilon > 0$ such that $\forall\Omega_{i,t} \in B_\epsilon(0)$, data outflow is approximately linear and the approximation error is bounded from above by $\frac{\rho^4\sigma_\theta^2\epsilon^2}{(1+\rho^2\sigma_\theta^2\epsilon)^3}$. The approximation error is small when $\rho$ or $\sigma_\theta$ is small, or when $\Omega_{i,t}$ is very close to 0.*

Proof:
Recall that data outflows (eq 8) for an individual firm are $d\Omega^-_{i,t} = \Omega_{it} - \left[(\rho^2\Omega_{it})^{-1} + \sigma_\theta^2\right]^{-1} = \Omega_{it} - \frac{\Omega_{it}\rho^2}{\Omega_{it}\rho^2\sigma_\theta^2+1}$. Let $m(\Omega_{it}) \equiv \frac{\Omega_{it}\rho^2}{\Omega_{it}\rho^2\sigma_\theta^2+1}$ be the nonlinear part of data outflows. Its first order Taylor expansion around 0 is $m(\Omega_{it}) = m(0) + m'(0)\Omega_{it} + o(\Omega_{it})$, with $m'(0) = \rho^2$. Thus $\frac{\partial d\Omega^-_{it}}{\partial\Omega_{it}} = 1 - m'(\Omega_{it}) \approx 1 - m'(0)$ for $\Omega_{i,t}$ in a small open ball $B_\epsilon(0)$, $\epsilon > 0$, around 0. The maximum error in approximating $d\Omega^-_{i,t}$ through the first order approximation of $m(\Omega_{i,t})$ is given by $|o(\Omega_{i,t})| = \frac{\rho^4\sigma_\theta^2\Omega^2_{i,t}}{(1+\rho^2\sigma_\theta^2\Omega_{i,t})^3}$. Now, $|o(\Omega_{i,t})| \geq 0$ and equals 0 if and only if $\Omega_{i,t} = 0$. Thus, $\exists\epsilon > 0$ such that $|o(\Omega_{i,t})|$ increases with $\Omega_{i,t}$ for all $\Omega_{i,t} \in B_\epsilon(0)$. Therefore, this error term is bounded above by $|o(\epsilon)| = \frac{\rho^4\sigma_\theta^2\epsilon^2}{(1+\rho^2\sigma_\theta^2\epsilon)^3}$ for all $\Omega_{i,t} \in B_\epsilon(0)$.

**Lemma 4** ***Linear Data Outflow with Small State Innovations*** *$\exists\epsilon_\sigma > 0$ such that $\forall\sigma_\theta \in B_{\epsilon_\sigma}(0)$, data outflows are approximately linear and the approximation error is bounded from above by $\frac{\Omega^2_{i,t}\rho^4\epsilon^2_\sigma}{1+\Omega_{i,t}\rho^2\epsilon^2_\sigma}$. The approximation error is small when $\sigma_\theta$ is close to 0.*

Proof:
Recall that data outflows are $d\Omega^-_{i,t} = \Omega_{it} - \frac{\Omega_{it}\rho^2}{\Omega_{it}\rho^2\sigma_\theta^2+1}$. The non-linear term $m(\Omega_{i,t}) \equiv \frac{\Omega_{it}\rho^2}{\Omega_{it}\rho^2\sigma_\theta^2+1}$ is linear when $\sigma_\theta = 0$. Therefore, $\exists\epsilon_\sigma > 0$ such that $\forall\sigma_\theta \in B_{\epsilon_\sigma}(0)$, $m(\Omega_{i,t})$ is approximately linear. The approximation error $|m(\Omega_{i,t}) - \rho^2\Omega_{i,t}| = \frac{\Omega^2_{i,t}\rho^4\sigma^2_\theta}{1+\Omega_{i,t}\rho^2\sigma^2_\theta}$ is increasing with $\epsilon_\sigma$ and reaches its maximum value at $\sigma_\theta = \epsilon_\sigma$, with value $\frac{\Omega^2_{i,t}\rho^4\epsilon^2_\sigma}{1+\Omega_{i,t}\rho^2\epsilon^2_\sigma}$.

**Lemma 5** ***Linear Data Outflow with Abundant Knowledge*** *When $\Omega_{i,t}\rho^2 \gg \sigma_\theta^{-2}$, discounted data stock is very small relative to $\Omega_{i,t}$, so that data outflows are approximately linear. The approximation error is small when $\rho$ is small or when $\sigma_\theta$ is sufficiently large.*

Proof:
Rearrange data outflows above as $d\Omega^-_{i,t} = \Omega_{it} - \frac{\Omega_{it}\rho^2\sigma_\theta^{-2}}{\Omega_{it}\rho^2+\sigma_\theta^{-2}}$. Let $m(\Omega_{it}) \equiv \frac{\Omega_{it}\rho^2\sigma_\theta^{-2}}{\Omega_{it}\rho^2+\sigma_\theta^{-2}}$ be the nonlinear part of data outflows. Since $(\rho^2\Omega_{it})^{-1} \geq 0$, we have $m(\Omega_{it}) \leq \sigma_\theta^{-2}$.

When $\Omega_{it}\rho^2 \gg \sigma_\theta^{-2}$, $m(\Omega_{it}) \approx \sigma_\theta^{-2}$ which is small when $\sigma_\theta$ is sufficiently large. The constant $m$ implies that outflow is approximately linear.

For low levels of $\rho$, $(\Omega_{it}\rho^2)^{-1}$ is large, $\Omega^-_{it} \approx \Omega_{it} - \Omega_{it}\rho^2$ and the approximation error is $|m(\Omega_{it}) - \rho^2\Omega_{it}| = \frac{\Omega^2_{it}\rho^4}{\sigma_\theta^{-2}+\Omega_{it}\rho^2}$ which is small when $\rho$ is small.

## A.4 Deterministic Aggregate Output.

Why is there no expectation operator around aggregate output, profits or prices? $\Phi_t$ is not random at date $t$ because aggregate quality $\int A_{i,t}di$ converges to a non-random value, even though each $A_{i,t}$ for each firm $i$ is a random variable. The reason is that the random shocks to $A_{i,t}$'s are independent and converge, by the central limit theorem.

Recall that quality is $A_{i,t} = g((a_{it} - \theta_t - \epsilon_{a,i,t})^2)$. The $\epsilon_a$ shocks are obviously idiosyncratic and independent. That is not a cause for concern so we set those aside. However, one might think that shocks to $\theta_t$ would cause $A_{i,t}$ to covary across firms and create aggregate shocks to quality and output. The reason this does not happen is that the action choice $a_{it}$ is firm $i$'s conditional expectation of $\theta_t$. So, $a_{it} - \theta_t$ is a forecast error. The forecast errors are what are independent. What ensures this is the noisy prior assumption made in the model setup. When the prior is noisy, beliefs about $\theta_t$ are the true $\theta_t$, plus idiosyncratic signal noise. Thus, forecast errors are idiosyncratic, or independent. Since any function of an independent random variable or variables is independent, $A_{i,t} = g((a_{it} - \theta_t - \epsilon_{a,i,t})^2)$ is independent across firms. Since the random component of $A_{i,t}$ is independent, its integral over an infinite number of firms, its mean, converges to a constant, by the central limit theorem.

Since we have a continuum of firms, then for any finite types of firms, like the $H$ and $L$ firms later, the quality of each type of firm also has independent noise. Therefore, the type-specific quality averages $A_{L,t}$ and $A_{H,t}$, that we make use of later, will also be non-random variables.

## A.5 Proof of Proposition 1: S-shaped Accumulation of Knowledge

We proceed in two parts: convexity and then concavity.

*Part a. Convexity at low levels of $\Omega_t$. In this part, we first calculate the derivatives of data infow and outflow with respect to $\Omega_{i,t}$, combine them to form the derivative of data net flow, and then show that it is positive in given parameter regions for $\Omega_{i,t} < \hat{\Omega}$.*

Since all other firms, besides firm $i$ are in steady state, we take the prices $\pi_t$ and $P_t$ as given. Since data is sufficiently expensive, data purchases are small. We prove this for zero data trade. By continuity, the result holds for small amounts of traded data.

Recall that data inflow is $d\Omega^+_{i,t} = z_{i,t} k^{\alpha}_{i,t} \sigma_{\epsilon}^{-2}$ and its first derivative is $\frac{\partial d\Omega^+_{i,t}}{\partial \Omega_{i,t}} = \alpha z_{i,t} k^{\alpha-1}_{i,t} \sigma_{\epsilon}^{-2} \frac{\partial k_{i,t}}{\partial \Omega_{i,t}}$. We then need to find $\frac{\partial k_{i,t}}{\partial \Omega_{i,t}}$.

Since we assumed that $\Psi$ is small, consider the case where $\psi = 0$. In this case, the data adjustment term drops out and the capital first-order condition reduces to

$$k^{1-\alpha}_{i,t} = \frac{\alpha}{r}\left(P_t A_{i,t} + z_i \sigma_{\epsilon}^{-2} \frac{1}{1+r} V'(\Omega_{i,t+1})\right). \tag{17}$$

Differentiating with respect to $\Omega_{i,t}$ on both sides yields

$$\frac{\partial k^{1-\alpha}_{i,t}}{\partial \Omega_{i,t}} = \frac{\partial k^{1-\alpha}_{i,t}}{\partial k_{i,t}} \cdot \frac{\partial k_{i,t}}{\partial \Omega_{i,t}} = (1-\alpha) k^{-\alpha}_{i,t} \cdot \frac{\partial k_{i,t}}{\partial \Omega_{i,t}}$$

Differentiating (17) with respect to $\Omega_{i,t}$ and using the relationships $\frac{\partial A_{i,t}}{\partial \Omega_{i,t}} = {\Omega_{i,t}}^{-2}$ and $\frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} = \rho^2[\rho^2 + \sigma_{\theta}^2 \Omega_{i,t}]^{-2}$, yields

$$\frac{\partial k_{i,t}}{\partial \Omega_{i,t}} = k^{\alpha}_{i,t} \frac{\alpha}{(1-\alpha) r}\left(P_t {\Omega_{i,t}}^{-2} + z_i \sigma_{\epsilon}^{-2} \frac{1}{1+r} V''(\Omega_{i,t+1}) \rho^2 [\rho^2 + \sigma_{\theta}^2 \Omega_{i,t}]^{-2}\right).$$

Therefore,

$$\begin{aligned}\frac{\partial d\Omega^+_{i,t}}{\partial \Omega_{i,t}} &= z_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-2} \frac{\alpha^2}{(1-\alpha) r}\left(P_t {\Omega_{i,t}}^{-2} + z_i \sigma_{\epsilon}^{-2} \frac{1}{1+r} V''(\Omega_{i,t+1}) \rho^2 [\rho^2 + \sigma_{\theta}^2 \Omega_{i,t}]^{-2}\right) \\ &= z_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-2} \frac{\alpha^2}{(1-\alpha) r} P_t {\Omega_{i,t}}^{-2} + z^2_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-4} \frac{\alpha^2}{1-\alpha} \frac{1}{r(1+r)} V''(\Omega_{i,t+1}) \rho^2 [\rho^2 + \sigma_{\theta}^2 \Omega_{i,t}]^{-2}.\end{aligned} \tag{18}$$

Next, take the derivative of data outflow $d\Omega^-_{i,t} = \Omega_{i,t} - \left[(\rho^2 \Omega_{i,t})^{-1} + \sigma_{\theta}^2\right]^{-1}$ with respect to $\Omega_{i,t}$:

$$\frac{\partial d\Omega^-_{i,t}}{\partial \Omega_{i,t}} = 1 - \frac{1}{\rho^2 \Omega^2_{i,t} (\sigma_{\theta}^2 + \rho^{-2} \Omega^{-1}_{i,t})^2}.$$

The derivatives of net data flow is then

$$\begin{aligned}\frac{\partial d\Omega^+_{i,t}}{\partial \Omega_{i,t}} - \frac{\partial d\Omega^-_{i,t}}{\partial \Omega_{i,t}} = \; & z_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-2} \frac{\alpha^2}{(1-\alpha) r} P_t {\Omega_{i,t}}^{-2} + z^2_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-4} \frac{\alpha^2}{1-\alpha} \frac{1}{r(1+r)} V''(\Omega_{i,t+1}) \rho^2 [\rho^2 + \sigma_{\theta}^2 \Omega_{i,t}]^{-2} \\ & + \frac{1}{\rho^2 \Omega^2_{i,t} (\sigma_{\theta}^2 + \rho^{-2} \Omega^{-1}_{i,t})^2} - 1.\end{aligned} \tag{19}$$

For notational convenience, denote the first term in (19) as $M_1 = z_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-2} \frac{\alpha^2}{(1-\alpha) r} P_t {\Omega_{i,t}}^{-2} > 0$, the second term as $M_2 = z^2_{i,t} k^{2\alpha-1}_{i,t} \sigma_{\epsilon}^{-4} \frac{\alpha^2}{1-\alpha} \frac{1}{r(1+r)} V''(\Omega_{i,t+1}) \rho^2 [\rho^2 + \sigma_{\theta}^2 \Omega^{-2}_{i,t})]^{-2} \leq 0$ and the third term as $M_3 = \frac{1}{\rho^2 \Omega^2_{i,t} (\sigma_{\theta}^2 + \rho^{-2} \Omega^{-1}_{i,t})^2} > 0$. Notice that $M_3 - 1 < 0$ always holds, and thus $M_2 + M_3 - 1 < 0$. $\frac{\partial d\Omega^+_{i,t}}{\partial \Omega_{i,t}} - \frac{\partial d\Omega^-_{i,t}}{\partial \Omega_{i,t}} > 0$ only holds when $P_t$ is sufficiently large so that $M_1$ dominates. $P_t$ is sufficiently large when $\bar{P}$ is sufficiently large.

Assume that $V'' \in [\nu, 0)$. Let $h(\Omega_{i,t}) \equiv M_1(\bar{P}) + M_2(\nu)$. Then

$$h'(\Omega_{i,t}) = \qquad (2\alpha-1)z_{i,t}k_{i,t}^{3\alpha-2}\alpha\left(\frac{\alpha}{r(1-\alpha)}\right)^2 \sigma_\epsilon^{-2}\left[\bar{P}\Omega_{i,t}{}^{-2} + z_{i,t}\sigma_\epsilon{}^{-2}\frac{1}{1+r}\nu\rho^2[\rho^2+\sigma_\theta^2\Omega_{i,t}]^{-2}\right]^2$$
$$+z_{i,t}k_{i,t}^{2\alpha-1}\frac{\alpha^2}{(1-\alpha)r}\sigma_\epsilon{}^{-2}\left[-2\bar{P}\Omega_{i,t}^{-3} - z_{i,t}\sigma_\epsilon{}^{-2}\frac{1}{1+r}\nu\rho^2\frac{2\sigma_\theta^2}{(\rho^2+\sigma_\theta^2\Omega_{i,t})^3}\right].$$

The first term is positive when $\alpha > \frac{1}{2}$, and negative when $\alpha < \frac{1}{2}$. And the second term is positive when $\bar{P} < f(\Omega_{i,t})$, and negative when $\bar{P} > f(\Omega_{i,t})$. To see this, note that

$$z_{it}k_{it}^{2\alpha-1}\frac{\alpha^2}{(1-\alpha)r}\sigma_\epsilon{}^{-2}\left[-2\bar{P}\Omega_{it}^{-3} - z_{it}\sigma_\epsilon{}^{-2}\frac{1}{1+r}\nu\rho^2\frac{2\sigma_\theta^2}{\left(\rho^2+\sigma_\theta^2\Omega_{it}\right)^3}\right] > 0$$

if and only if $\bar{P} < f(\Omega_{i,t})$, where

$$f(\Omega_{i,t}) := -z_{it}\sigma_\epsilon{}^{-2}\frac{1}{1+r}\nu\rho^2\Omega_{it}^3\frac{\sigma_\theta^2}{\left(\rho^2+\sigma_\theta^2\Omega_{it}\right)^3}$$

Notice by inspection that $f'(\Omega_{i,t}) < 0$.

Let $\hat{\Omega}$ be the first root of

$$h(\Omega_{i,t}) = 1 - M_3,$$

then if $\alpha < \frac{1}{2}$, when $\Omega_{i,t} < \hat{\Omega}$ and $\bar{P} > f(\hat{\Omega})$, we have that $h(\Omega_{i,t})$ is decreasing in $\Omega_{i,t}$ and $h(\Omega) \geq 1 - M_3$. Since $\nu \leq V''$, we then have $M_1 + M_2 \geq 1 - M_3$, that is $\frac{\partial d\Omega_{i,t}^+}{\partial\Omega_{i,t}} - \frac{\partial d\Omega_{i,t}^-}{\partial\Omega_{i,t}} > 0$. By the same token, if $\alpha > \frac{1}{2}$ and $\bar{P} < f(\Omega_{i,t})$, then $\frac{\partial d\Omega_{i,t}^+}{\partial\Omega_{i,t}} - \frac{\partial d\Omega_{i,t}^-}{\partial\Omega_{i,t}} < 0$.

*Part b. Concavity at high levels of $\Omega_t$. In this part, we first calculate limit of the derivatives of net data flow with respect to $\Omega_{i,t}$ is negative when $\Omega_{i,t}$ goes to infinity and then prove that when $\Omega_{i,t}$ is large enough, $\frac{\partial d\Omega_{i,t}}{\partial\Omega_{i,t}}$ is negative.*

For $\rho \leq 1$ and $\sigma_\theta^2 \geq 0$, data outflows are bounded below by zero. But note that outflows are not bounded above. As the stock of knowledge $\Omega_{i,t} \to \infty$, outflows are of $O(\Omega_{i,t})$ and approach infinity. We have that $\frac{\partial d\Omega_{i,t}^-}{\partial\Omega_{i,t}} = 1 - (\rho\Omega_{i,t})^{-2}(\sigma_\theta^2 + \rho^{-2}\Omega_{i,t}^{-1})^{-2}$. It is easy to see that $\lim_{\Omega_{i,t}\to\infty} \frac{\partial d\Omega_{i,t}^-}{\partial\Omega_{i,t}} = 1$.

For the derivative of data inflow (18), note that $\frac{\partial d\Omega_{i,t}^+}{\partial\Omega_{i,t}} \leq z_{i,t}k_{i,t}^{2\alpha-1}\sigma_\epsilon^{-2}\frac{\alpha^2}{(1-\alpha)r}P_t\Omega_{i,t}{}^{-2}$ because $0 < \alpha < 1$ and $V'' < 0$. Thus $\lim_{\Omega_{i,t}\to\infty} \frac{\partial d\Omega_{i,t}^+}{\partial\Omega_{i,t}} \leq 0$.

Therefore, $\lim_{\Omega_{i,t}\to\infty} \frac{\partial d\Omega_{i,t}+}{\partial\Omega_{i,t}} - \frac{\partial d\Omega_{i,t}-}{\partial\Omega_{i,t}} \leq -1$. Since data outflows and inflows are continuously differentiable, $\exists\hat{\hat{\Omega}} > 0$ such that $\forall\Omega_{i,t} > \hat{\hat{\Omega}}$, we have $\frac{\partial d\Omega_{i,t}+}{\partial\Omega_{i,t}} - \frac{\partial d\Omega_{i,t}-}{\partial\Omega_{i,t}} < 0$, which is the decreasing returns to data when data is abundant.

## A.6 Proof of Proposition 2: New Firms Earn Negative Profits

Without any production or any data purchased, $\Omega_0 = \sigma_\theta^{-2}$, because this is the prior variance of the state $\theta$. This is the case when the firm is entering.

Consider the approximation in Equation (5): $\mathbb{E}_i[A_{i,t}] \approx g\left(\Omega_{i,t}^{-1} + \sigma_u^2\right) + g''\left(\Omega_{i,t}^{-1} + \sigma_u^2\right) \cdot \left(\Omega_{i,t}^{-1} + \sigma_u^2\right)$. $g(v)$ is decreasing. When $g''(.) = 0$ (the standing assumption of this part of the paper), then the second term is zero. Thus $\mathbb{E}[A_{i,0}] = g(\sigma_u^2 + \sigma_\theta^2) < 0$. The inequality is the assumption stated in the proposition.

If expected quality $\mathbb{E}[A_{i,0}]$ is less than zero, then expected profit is negative, for any positive level of production, because the steady state price level for goods is positive $P^{ss} > 0$. This can be seen in (9), noting that adjustment cost $\Psi$, capital rental $r$ and data prices $\pi$ are all non-negative, by assumption or by free disposal.

Of course, a firm can always choose zero production $k_{i,t} = 0$ and zero data to achieve zero profit. A firm that chose this every period, would have no profit ever and thus zero firm value.

Thus, the only way to get to positive firm value is to produce. Either the firm first buys data and then produces, first produces, or does both together. If the firm first buys data, then profit is negative in the period when the firm buys the data and is not yet producing. If the firm produces first, profit is negative because expected quality is negative, as per the argument above. If the firm produces and buys data at the same time, then profit is more negative because of negative expected quality and the cost of the data purchase. In every scenario, the firm must incur some negative profit to achieve positive production and positive firm value.

## A.7 Proof of Proposition 3: Firms Sell Goods at Zero Price (Data Barter)

*Proof:* Suppose the price goods is $P_t = 0$. We want to show that an optimal production/ investment level $K_t$ can be optimal in this environment. Consider a price of data $\pi_t$ is such that firm $i$ finds it optimal to sell a fraction $\chi > 0$ of its data produced in period $t$: $\delta_{i,t} = -\chi n_{i,t}$. In this case, differentiating the value function (6) with respect to $k$ yields $(\pi_t/\iota)\chi z_i \alpha k^{\alpha-1} = r + \frac{\partial\Psi(\Delta\Omega_{i,t+1})}{\partial k_{i,t}}$. Can this optimality condition hold for positive investment level $k$? If $k^{1-\alpha} = \frac{\pi_t \chi z_i \alpha}{\left(r + \frac{\partial\Psi(\Delta\Omega_{i,t+1})}{\partial k_{i,t}}\right)\iota} > 0$, then the firm optimally chooses $k_{i,t} > 0$, at price $P_t = 0$. □

## A.8 Data Accumulation Can be Purely Concave

Data accumulation is not always S-shaped, only for some parameter values. For others, it can be that data accumulation is purely concave. Instead, the net data flow (the slope) decreases with $\Omega_{i,t}$, right from the start.

**Proposition 7** ***Concavity of Data Inflow*** *$\exists\epsilon > 0$ such that $\forall\Omega_{i,t} \in B_\epsilon(0)$, the net data flow decreases with $\Omega_{i,t}$.*

We proceed in two steps. In Step 1, we prove that data outflows are approximately linear when $\Omega_{i,t}$ is small. And then in Step 2, we first calculate the derivative of net data flow with respect to $\Omega_{i,t}$ and then characterize the parameter region where it is negative.

*Step 1: Data outflows are approximately linear when $\Omega_{i,t}$ is small.*
This is proven separately in Lemma 3.

*Step 2: Characterize the parameter region where the derivative of net data flow with respect to $\Omega_{i,t}$ is negative. A negative least upper bound is sufficient for it be negative.*

Recall that the derivative of data inflows with respect to the current stock of knowledge $\Omega_t$ is $\frac{\partial d\Omega^+_{i,t}}{\partial\Omega_{i,t}} = \rho^2\left[\rho^2 + \sigma^2_\theta\Omega_{i,t}\right]^{-2} > 0$ (see the Proof of Proposition 1 for details). Thus

$$\frac{\partial d\Omega^+_{i,t}}{\partial\Omega_{i,t}} - \frac{\partial d\Omega^-_{i,t}}{\partial\Omega_{i,t}} \approx \rho^2\left[\rho^2 + \sigma^2_\theta\Omega_{i,t}\right]^{-2} + \rho^2 - 1.$$

Since this derivative increases in $\rho^2$ and decreases in $\Omega_{i,t} = 0$, so its least upper bound 1 is achieved when $\rho^2 = 1$ and $\Omega_{i,t} = 0$. A non-negative least upper bound requires $0 \geq \sigma^2_\theta$. Since $\sigma^2_\theta > 0$, the supreme of $\frac{\partial d\Omega^+_{i,t}}{\partial\Omega_{i,t}} - \frac{\partial d\Omega^-_{i,t}}{\partial\Omega_{i,t}}$ is negative, so it will always be negative $\forall\Omega_{i,t} \in B_\epsilon(0)$.

## A.9 Proof of Proposition 4

**Part 1** Suppose not. Then, for every firm $i \in I$, with $\int_{i\notin I} di = 0$, producing infinite data $n_{i,t} \to \infty$ implies finite firm output $y_{i,t} < \infty$. Thus $M_y \equiv \sup_i\{y_{i,t}\} + 1$ exists and is finite. By definition, $y_{i,t} < M_y$, $\forall i$. If the measure of all firms is also finite, that is $\exists 0 < N < \infty$ such that $\int_i di < N$. As a result, the aggregate output is also finite in any period $t+s$, $\forall s > 0$:

$$Y_{t+s} = \int_i y_{i,t} di < M_y \int_i di < M_y N < \infty. \tag{20}$$

On the other hand, given that the aggregate growth rate of output $\ln(Y_{t+1}) - \ln(Y_t) > \underline{g} > 0$, we have that in period $t+s$, $\forall s > 0$, output growth is $\ln(Y_{t+s}) - \ln(Y_t) = [\ln(Y_{t+s}) - \ln(Y_{t+s-1})] + \cdots + [\overline{\ln}(Y_{t+1}) - \ln(Y_t)] > \underline{g}s$. This implies that $Y_{t+s} > Y_t e^{\underline{g}s}$.. Thus for $\forall s > \underline{s} \equiv \lceil\frac{\ln(MN)-\ln(Y_t)}{\underline{g}}\rceil$,

$$Y_{t+s} > Y_t e^{\underline{g}s} > Y_t e^{\underline{g}\underline{s}} > Y_t e^{\underline{g}\frac{\ln(M_y N)-\ln(Y_t)}{\underline{g}}} = M_y N,$$

which contradicts (20).

**Part 2** We break this part into two sub-parts. Part (a) of the result is that in order to have infinite output in the limit, an economy will need $(a_{i,t} - \theta_t - \epsilon_{a,i,t})^2$ to approach zero.

Part (b) says: For $(a_{i,t} - \theta_t - \epsilon_{a,i,t})^2$ to approach zero, marginal utility relevant variables $\theta_t$ and $\epsilon_{a,i,t}$ must be in the set $\Xi_{t-1}$.

*Proof part a*: From Proposition Part 1, we know that sustaining aggregate growth above any lower bound $\underline{g} > 0$ arises only if a data economy achieves infinite output $Y_t \to \infty$ when some firm has infinite data $n_{i,t} \to \infty$. Since $Y_t$ is a finite-valued function, except at 0, infinite output requires that the argument of $g$, which is $(a_{i,t} - \theta_t - \epsilon_{a,i,t})^2$ becomes arbitrarilty close to zero.

*Proof of part b.* Suppose not. The optimal action that can achieve infinite output when $g$ is not finite-valued is $a_t^* = \theta_t + \epsilon_{a,i,t}$. If the optimal action is not in $\Xi_{t-1}$, then it is not a $t$-measurable action. There is some unforecastable error such that $\mathbb{E}[(a_{i,t} - \theta_t - \epsilon_{a,i,t})^2] > \underline{z} > 0$.

If it is not a measurable action, it cannot be chosen with strictly positive probability in a continuous action space. Since the optimal action must be in $\Xi_{t-1}$, then $\theta_t + \epsilon_{a,i,t}$ must be in $\Xi_{t-1}$ as well. Since $\theta_t$ and $\epsilon_{a,i,t}$ are unconditionally and conditionally independent, for the sum to be perfectly predictable, each element must also be perfectly predictable. Thus, $\theta_t$ and $\epsilon_{a,i,t}$ must be in $\Xi_{t-1}$.

## A.10 Competitive Equilibrium

In order to prove our welfare result, we begin by characterizing competitive equilibrium. Then we characterize the solution to the social planner problem. Finally, we compare the two solutions to determine the efficiency of the equilibrium outcome.

**Household problem** Let $\Gamma_t$ denote the Lagrangian multiplier of the individual problem on his budget constraint. Individual problem can be written as:

$$\max_{c_t, m_t} \sum_{t=0}^{+\infty} \frac{1}{(1+r)^t} \left(u(c_t) + m_t\right) \qquad \text{with } u(c_t) = \bar{P} \frac{c_t^{1-\gamma}}{1-\gamma}$$
$$\text{s.t.} \quad P_t c_t + m_t = \Phi_t \qquad \forall t$$

where $\Phi_t$ is the aggregate profit of all firms:

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

The first order conditions for optimal household choices of consumption of $c_t$ and the numeraire good $m_t$ are

$$c_t: \quad \frac{1}{(1+r)^t} u'(c_t) = P_t \Gamma_t,$$
$$m_t: \quad \Gamma_t = \frac{1}{(1+r)^t},$$

The first order conditions imply that agents equate their marginal utility of $c$ to its price: $P_t = u'(c_t)$.

**Firm problem** Firms' sequential optimization problem is

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

Equivalently, in recursive form

$$V(\Omega_{i,t}) = \max_{k_{i,t}, \delta_{i,t}} P_t \mathbb{E}[A_{i,t}|\mathcal{I}_{i,t}] k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi_t \delta_{i,t} - r k_{i,t} + \frac{V(\Omega_{i,t+1})}{1+r} \tag{21}$$

$$\text{s.t.} \quad \Omega_{i,t+1} = \left(\rho^2 \Omega_{i,t}^{-1} + \sigma_\theta^2\right)^{-1} + \left(z_i k_{i,t}^{\alpha} + \left(\mathbf{1}_{\delta_{i,t}>0} + \iota \mathbf{1}_{\delta_{i,t}<0}\right) \delta_{i,t}\right) \sigma_\epsilon^{-2} \tag{22}$$

The profits of the firm at time $t$ are $\Phi_{i,t} = P_t A_{i,t} k_{i,t}^{\alpha} - \Psi(\Delta\Omega_{i,t+1}) - \pi \delta_{i,t} - r k_{i,t}$.

**Market clearing (resource constraint)** is given by

$$\begin{aligned}
&\text{retail good :} && c_t = \int_i A_{i,t} k_{i,t}^{\alpha} di,\\
&\text{numeraire good :} && m_t + \int_i \big(r k_{i,t} + \Psi(\Delta\Omega_{i,t+1})\big) di = 0\\
&\text{data :} && \int_i \delta_{i,t} di = 0.
\end{aligned}$$

The adjustment cost $\Psi$ is incorporated in the market clearing/resource constraint for the numeraire good so that it shows up in the planner's objective function.

**Steady state** In equilibrium, households (HHs, hereafter) maximize utility by choosing $c_t$ and $m_t$, firms maximize profits by choosing $\{k_{i,t}, \delta_{i,t}\}_{i=L,H}$, and markets clear.

In this section we focus on steady state equilibrium outcomes with two types of firms, $i = L, H$. HH budget constraint simplifies to

$$\begin{aligned}
&P^{eq} c^{eq} + m^{eq} = \Phi^{eq}\\
&\Phi^{eq} = P^{eq}\Big(\lambda \mathbb{E}[A_L^{eq}](k_L^{eq})^{\alpha} + (1-\lambda)\mathbb{E}[A_H^{eq}](k_H^{eq})^{\alpha}\Big) - r\Big(\lambda k_L^{eq} + (1-\lambda) k_H^{eq}\Big)
\end{aligned}$$

where HH optimization implies $P^{eq} = u'(c^{eq})$. In steady state, the market clearing conditions simplify to

$$\begin{aligned}
&\text{retail good :} && c^{eq} = \lambda \mathbb{E}[A_L^{eq}](k_L^{eq})^{\alpha} + (1-\lambda)\mathbb{E}[A_H^{eq}](k_H^{eq})^{\alpha},\\
&\text{numeraire good :} && m^{eq} + r\Big(\lambda k_L^{eq} + (1-\lambda) k_H^{eq}\Big) = 0\\
&\text{data :} && \lambda \delta_L^{eq} + (1-\lambda)\delta_H^{eq} = 0.
\end{aligned}$$

**Firms' optimal capital choices** There are two equations for first order condition (FOC) with respect to $k_i$, $i = L, H$. We will use the sequential problem to get this first order condition. Consider FOC of firm $i$ with respect to $k_{i,t}$:

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

Substitute

$$\frac{\partial \mathbb{E}[A_{i,t+1}|\mathcal{I}_{i,t}]}{\partial k_{i,t}} = \alpha z_i \sigma_{\epsilon}^{-2} k_{i,t}^{\alpha-1} \Omega_{i,t+1}^{-2} g\prime.$$

Multiply both sides by $\frac{1}{(1+r)^t}$. Steady state implies a stable level of knowledge ($\Delta\Omega = 0$). With a quadratic adjustment cost function that is 0 at 0, $\Psi'(0) = 0$. Thus, in the steady state $\frac{\partial \Psi(\Delta\Omega_{i,t+2})}{\partial k_{i,t}} = \frac{\partial \Psi(\Delta\Omega_{i,t+1})}{\partial k_{i,t}} = 0$. Imposing this condition simplifies the firm's FOC:

$$\alpha P k_i^{\alpha-1}\Big(\mathbb{E}[A_i] + \frac{z_i \sigma_{\epsilon}^{-2}}{1+r}\Omega_i^{-2} g\prime k_i^{\alpha}\Big) = r. \tag{23}$$

**Firm's optimal data choices.** In the steady state, where the adjustment cost is zero, the firm's FOC with respect to data purchases/sales is $\pi_t = \frac{1}{1+r} V'(\Omega_{i,t+1})\sigma_{\epsilon}^{-2}(\mathbb{1}_{\delta_{i,t}>0} + \iota \mathbb{1}_{\delta_{i,t}<0})$, which can be rearranged as

$$V'(\Omega_{i,t+1}) = \frac{(1+r)\pi_t}{\sigma_{\epsilon}^{-2}(\mathbb{1}_{\delta_{i,t}>0} + \iota \mathbb{1}_{\delta_{i,t}<0})} \tag{24}$$

Next, differentiate the value function of the firm with respect to $\Omega_{i,t}$ and use the envelope condition to hold the choice variables constant:

$$V'(\Omega_{i,t}) = P_t k_{i,t}^{\alpha} \Omega_{i,t}^{-2} + \frac{1}{1+r} V'(\Omega_{i,t+1}) \frac{\partial \Omega_{i,t+1}}{\Omega_{i,t}}, \tag{25}$$

Differentiating Equation (22) with respect to $\Omega_{i,t}$,

$$\frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}} = \frac{\rho^2}{\left(\rho^2 + \sigma_\theta^2 \Omega_{i,t}\right)^2}. \tag{26}$$

Substitute Equation (24) for $V'(\Omega_{i,t}) = V'(\Omega_{i,t+1})$ (in steady state) in (25):

$$\left(1 - \frac{1}{1+r} \frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}}\right) V'(\Omega_{i,t}) = P_t k_{i,t}^\alpha \Omega_{i,t}^{-2}$$

Next substitute for $V'(\Omega_{i,t})$ in Equation (24), using the expression for $\frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}}$ from Equation (26). Then, multiply through by $1+r$, and re-arrange. This yields one condition for the optimal capital-knowledge ratio for $L$ firms and one for $H$ firms:

$$\left(1 + r - \frac{\rho^2}{(\rho^2 + \sigma_\theta^2 \Omega_i)^2}\right) \frac{\pi}{P \sigma_\epsilon^{-2} (\mathbb{1}_{\delta_i > 0} + \iota \mathbb{1}_{\delta_i < 0})} = k_i^\alpha \Omega_i^{-2} \qquad i = L, H$$

If we guess and verify that $H$ firms will sell data and $L$ firms will buy it, then we can simplify $(\mathbb{1}_{\delta_i > 0} + \iota \mathbb{1}_{\delta_i < 0})$, by equating it to 1 for $L$ firms and $\iota$ for $H$ firms. Taking the ratio of the $L$ and $H$ optimality conditions allows us to cancel out $P_t$, which delivers Equation (29).

Thus the 6 equilibrium steady state real variables, $(\Omega_L^{eq}, \Omega_H^{eq}, k_L^{eq}, k_H^{eq}, \delta_L^{eq}, \delta_H^{eq})$ are determined by the following system of 6 equations. Note that (27) and (28) represent two equations each.

$$\Omega_i^{eq} = \left[\rho^2 (\Omega_i^{eq})^{-1} + \sigma_\theta^2\right]^{-1} + \left(z_i (k_i^{eq})^\alpha + \delta_i^{eq} (\mathbf{1}_{\delta_i^{eq} > 0} + \iota \mathbf{1}_{\delta_i^{eq} < 0})\right) \sigma_\epsilon^{-2} \qquad i = L, H \tag{27}$$

$$r = \alpha \bar{P} (c^{eq})^{-\gamma} (k_i^{eq})^{\alpha - 1} \left[\mathbb{E}[A_i^{eq}] + \frac{z_i \sigma_\epsilon^{-2}}{1+r} (k_i^{eq})^\alpha (\Omega_i^{eq})^{-2}\right] \qquad i = L, H \tag{28}$$

$$\frac{(k_L^{eq})^\alpha / (\Omega_L^{eq})^2}{\iota (k_H^{eq})^\alpha / (\Omega_H^{eq})^2} = \frac{1 + r - \rho^2 (\rho^2 + \sigma_\theta^2 \Omega_L^{eq})^{-2}}{1 + r - \rho^2 (\rho^2 + \sigma_\theta^2 \Omega_H^{eq})^{-2}} \tag{29}$$

$$\lambda \delta_L^{eq} + (1 - \lambda) \delta_H^{eq} = 0. \tag{30}$$

Equation (27) represents the two law of motions for stock of knowledge, one for each type of firm $i = L, H$. Equation (28) comes from (23) with $P' = u(c)$ and $u'(c) = \bar{P} c^{-\gamma}$ substituted in. It represents the two first order conditions for capital choice, one for each type of firm $i = L, H$. (29) is a single equation, the ratio of first order conditions for the data choice for the two types of firm. Taking the ratio enables us to eliminate the steady state data price $\pi^{eq}$ from the system of equations. Finally, Equation (30) is the resource constraint for the total traded data, which should be zero.

## A.11 Social Planner Problem

The planner maximizes HH total discounted utility, taking the resource constraints into account. Thus planner's problem can be written as

$$\max_{\{k_{i,t}, \delta_{i,t}\}_{i=L,H}} \sum_{t=0}^{\infty} \frac{1}{(1+r)^t} \left(u(c_t) - r \int_i k_{i,t} di - \int_i \Psi(\Delta \Omega_{i,t+1}) di\right)$$

or in recursive form

$$\begin{aligned}
V^P(\{\Omega_{i,t}\}_i) = \max_{\{k_{i,t}, \delta_{i,t}\}_i} & u(c_t) - r \int_i k_{i,t} di - \int_i \Psi(\Delta \Omega_{i,t+1}) di + \frac{1}{1+r} V^P(\{\Omega_{i,t+1}\}_i) \\
\text{s.t.} \qquad & c_t = \int_i A_{i,t} k_{i,t}^\alpha di \qquad (\text{ with multiplier } \Xi_t) \qquad \forall t \\
& \int_i \delta_{i,t} di = 0 \qquad (\text{ with multiplier } \eta_t) \qquad \forall t \\
& \Omega_{i,t+1} = \left[\rho^2 \Omega_{i,t}^{-1} + \sigma_\theta^2\right]^{-1} + \left(z_i (k_{i,t})^\alpha + \delta_{i,t} (\mathbf{1}_{\delta_{i,t} > 0} + \iota \mathbf{1}_{\delta_{i,t} < 0})\right) \sigma_\epsilon^{-2} \quad \forall i, t \\
& \mathbb{E}[A_{i,t}] \approx g\left(\Omega_{i,t}^{-1} + \sigma_u^2\right) \qquad \forall i, t.
\end{aligned}$$

Similar to equilibrium, as the household consumption equal the aggregate production of a continuum of firms, it is deterministic at each time $t$.

**Social Planner's optimal capital choice.** The planner's first order condition with respect to $k_{i,t}$ is

$$r\lambda_i = \frac{\partial u(c_t)}{\partial k_{i,t}} + \frac{1}{1+r}\frac{\partial u(c_{t+1})}{\partial k_{i,t}} \qquad \text{for } \ i = L, H \tag{31}$$

Again, focus on two types of firm $i = L, H$ where the firms in each group are identical. Then $\lambda_i = \lambda$ when $i = L$ and $\lambda_i = 1 - \lambda$ when $i = H$. The planner objective simplifies to

$$\begin{aligned} V^P(\Omega_{L,t}, \Omega_{H,t}) = \max_{\{k_{i,t},\delta_{i,t}\}_{i=L,H}} & u(c_t) - r\Big(\lambda k_{L,t} + (1-\lambda)k_{H,t}\Big) - \Big(\lambda\Psi(\Delta\Omega_{L,t+1}) + (1-\lambda)\Psi(\Delta\Omega_{H,t+1})\Big) \\ & + \frac{1}{1+r}V^P(\Omega_{L,t+1}, \Omega_{H,t+1}) \end{aligned}$$

Furthermore, $c_t = \lambda\mathbb{E}[A_{L,t}]k_{L,t}^{\alpha} + (1-\lambda)\mathbb{E}[A_{H,t}]k_{H,t}^{\alpha}$. Thus

$$\begin{aligned} \frac{\partial c_t}{\partial k_{i,t}} &= \alpha\lambda_i\mathbb{E}[A_{i,t}]k_{i,t}^{\alpha-1} \\ \frac{\partial c_t}{\partial \Omega_{i,t}} &= \lambda_i\frac{\partial \mathbb{E}[A_{i,t}]}{\partial \Omega_{i,t}}k_{i,t}^{\alpha} = \lambda_i\Omega_{i,t}^{-2}g\prime k_{i,t}^{\alpha} \end{aligned} \tag{32}$$

In steady state, substitute in the expressions above into (31),

$$r = \alpha\bar{P}(c^{opt})^{-\gamma}(k_i^{opt})^{\alpha-1}\left[\mathbb{E}[A_i^{opt}] + \frac{z_i\sigma_\epsilon^{-2}}{1+r}(k_i^{opt})^{\alpha}(\Omega_i^{opt})^{-2}g\prime\right] \qquad i = L, H \tag{33}$$

This is the same as Equation (28). Thus the capital FOCs are the same between optimum and equilibrium.

**Social Planner's optimal data choice.** Let $V_i^P$ denote the derivative of the social planner value function with respect to $\Omega_{i,t}$, $i = L, H$. To solve for $V_i^P$ in steady state, differentiate the value function and apply the envelope condition to get:

$$V_i^P(\Omega_{i,t}, \Omega_{-i,t}) = \frac{\partial u(c_t)}{\partial \Omega_{i,t}} + \frac{1}{1+r}V_i^{P\prime}(\Omega_{i,t+1}, \Omega_{-i,t+1})\frac{\partial \Omega_{i,t+1}}{\partial \Omega_{i,t}}$$

The data first order condition reveals that the Lagrange multiplier $\eta_t$ on the data constraint is

$$\lambda_i\eta_t = \frac{1}{1+r}V_i^P(\Omega_{L,t+1}, \Omega_{H,t+1})\sigma_\epsilon^{-2}(\mathbb{1}_{\delta_{i,t}>0} + \iota\mathbb{1}_{\delta_{i,t}<0}). \tag{34}$$

In steady state, $V_i^p(\Omega_{i,t}, \Omega_{-i,t}) = V_i^P(\Omega_{i,t+1}, \Omega_{-i,t+1})$. Use this equality and Equations (26) and (34) to replace for $\frac{\partial\Omega_{i,t+1}}{\partial\Omega_{i,t}}$ and $V_i^P(\Omega_{i,t+1}, \Omega_{-i,t+1})$ to get

$$\left(1 + r - \rho^2(\rho^2 + \sigma_\theta^2\Omega_{i,t})^{-2}\right)\frac{\eta_t\lambda_i}{\sigma_\epsilon^{-2}(\mathbb{1}_{\delta_{i,t}>0} + \iota\mathbb{1}_{\delta_{i,t}<0})} = \frac{\partial u(c_t)}{\partial \Omega_{i,t}} \qquad i = L, H \tag{35}$$

which in steady state can be written as

$$\left(1 + r - \rho^2(\rho^2 + \sigma_\theta^2\Omega_{i,t}^{opt})^{-2}\right)\frac{\eta\lambda_i}{\bar{P}(c^{opt})^{-\gamma}\sigma_\epsilon^{-2}(\mathbb{1}_{\delta_i^{opt}>0} + \iota\mathbb{1}_{\delta_i^{opt}<0})} = \lambda_i(k_i^{opt})^{\alpha}(\Omega_i^{opt})^{-2}g\prime \qquad i = L, H \tag{36}$$

In steady state, $H$ firms sell data. For them, $(\mathbb{1}_{\delta_i^{opt}>0} + \iota\mathbb{1}_{\delta_i^{opt}<0}) = \iota$ while $L$ firms buy data. For them, $(\mathbb{1}_{\delta_i^{opt}>0} + \iota\mathbb{1}_{\delta_i^{opt}<0}) = 1$. Next take the ratio of the $H$ and $L$ conditions from (36). $(c^{opt})^{-\gamma}$ and the Lagrange multiplier $\eta$ both drop out of the resulting equation, thus we have

$$\frac{(k_L^{opt})^{\alpha}/(\Omega_L^{opt})^2}{\iota(k_H^{opt})^{\alpha}/(\Omega_H^{opt})^2} = \frac{1 + r - \rho^2(\rho^2 + \sigma_\theta^2\Omega_L^{opt})^{-2}}{1 + r - \rho^2(\rho^2 + \sigma_\theta^2\Omega_H^{opt})^{-2}}, \tag{37}$$

which is the same as Equation (29).

Finally, the planner's first order conditions with respect to consumption choice tells us that the Lagrange multiplier on the consumption resource constraint is $\Xi_t = u'(c_t)$.

## A.12 Proof of Proposition 5: Efficient Equilibrium

The decentralized equilibrium is characterized by Equations (27) for $i = L, H$, (29), (30), and (28) for $i = L, H$.

The social planner's optimum is characterized by Equations (27) for $i = L, H$ and (30) (all for optimum variables), Equation (33) for $i = L, H$, and Equation (37).

The resulting capital first order conditions for each form $i = L, H$, as well as the ratio of the data first order conditions across two types of firms, for both problems are the same. Thus, the equilibrium is efficient because the decentralized economy and the social planner end up making the same choices.

## A.13 Proof of Proposition 6: Inefficiency with Business Stealing

With business stealing externality, i.e. when $b = 1$, the only difference is that $A_i$ is determined by Equation (14). Thus in a symmetric allocation, with 2 types, where all firms of type $i$ are the same, in equilibrium we have

$$\mathbb{E}[A_{i,t}] = \left(\bar{A} - (\Omega_{i,t})^{-1} - \sigma_u^2\right) + \Big(\lambda_i \left(\Omega_{i,t}^{-1} + \sigma_u^2\right) + (1-\lambda_i)\left(\Omega_{-i,t}^{-1} + \sigma_u^2\right)\Big) = \bar{A} - (1-\lambda_i)(\Omega_{i,t}^{-1} - \Omega_{-i,t}^{-1}).$$

By construction, aside from the change in the equilibrium steady state value of $\mathbb{E}[A_i^{eq}]$, the business stealing externality does not change the firm optimization problem. In particular, it does not affect any of the first order condition, such as $\frac{\partial \mathbb{E}[A_{i,t+1}]}{\partial k_{i,t}}$. Thus the equilibrium is still characterized by Equations (27) for $i = L, H$, (29), (30), and (28) for $i = L, H$.

For the optimum, Equations (27) for $i = L, H$ and (30) clearly remains the same. The other optimum equations change as the quality of every firm is affected by the capital and data choices of each individual firm $i$.

**Planner's Optimal Data with Business Stealing** Observe that the amount of data traded by firm $i$ at time $t$, $\delta_{i,t}$ does not affect the stock of knowledge of firm $j$ at $t+1$, $\Omega_{j,t+1}$ conditional on $\delta_{j,t}$. Furthermore, $\Omega_{i,t}$ does not affect $\Omega_{j,t+1}$, $j \neq i$. However, $\frac{\partial c_t}{\partial \Omega_{i,t}}$ is adjusted to reflect data used for business stealing:

$$\begin{aligned}\frac{\partial c_t}{\partial \Omega_{i,t}} &= \lambda_i k_{i,t}^{\alpha} \frac{\partial \mathbb{E}[A_{i,t}]}{\partial \Omega_{i,t}} + (1-\lambda_i) k_{-i,t}^{\alpha} \frac{\partial \mathbb{E}[A_{-i,t}]}{\partial \Omega_{i,t}} = \lambda_i (1-\lambda_i) k_{i,t}^{\alpha} \Omega_{i,t}^{-2} - (1-\lambda_i)^2 k_{-i,t}^{\alpha} \Omega_{i,t}^{-2} \\ &= (1-\lambda_i)\Omega_{i,t}^{-2}\left(\lambda_i k_{i,t}^{\alpha} - (1-\lambda_i) k_{-i,t}^{\alpha}\right). \end{aligned} \tag{38}$$

Comparing Equations (32) and (38) clarifies that data with business stealing, data is less useful to increase the consumption level. The firms do not internalize that selling data ot others decreases their quality. Thus, there is an over-supply of data on the data market, and too much data trade. With business stealing, Equations (36) and (37) change to

$$\begin{aligned}&\left(1 + r - \frac{\rho^2}{(\rho^2 + \sigma_\theta^2 \Omega_i^{opt})^2}\right) \frac{\eta \lambda_i}{\bar{P}(c^{opt})^{-\gamma} \sigma_\epsilon^{-2} (\mathbb{1}_{\delta_i^{opt} > 0} + \iota \mathbb{1}_{\delta_i^{opt} < 0})} \\ &\qquad\qquad\qquad\qquad = (1-\lambda_i)(\Omega_i^{opt})^{-2}\left(\lambda_i (k_i^{opt})^{\alpha} - (1-\lambda_i)(k_{-i}^{opt})^{\alpha}\right) \quad \forall i\end{aligned}$$

$$\left(\frac{1-\lambda}{\lambda}\right)^2 \frac{\left(\lambda (k_L^{opt})^{\alpha} + (1-\lambda)(k_H^{opt})^{\alpha}\right)(\Omega_L^{opt})^{-2}}{\iota\left((1-\lambda)(k_H^{opt})^{\alpha} + \lambda (k_L^{opt})^{\alpha}\right)(\Omega_H^{opt})^{-2}} = \frac{1 + r - \rho^2(\rho^2 + \sigma_\theta^2 \Omega_L^{opt})^{-2}}{1 + r - \rho^2(\rho^2 + \sigma_\theta^2 \Omega_H^{opt})^{-2}}, \tag{39}$$

Equation (39) is different from equilibrium Equation (29) on the left hand side.

This is the first externality. With business stealing, the planner internalizes that the data that a firms sells on the data market, decreases its own quality. Since firms do to internalize this effect, they sell more data on the data market than what is efficient. There is excessive data trade.

**Planner's Optimal Capital with Business Stealing** The first order condition for the planner's capital choice becomes $r\lambda_i = \partial u(c_t)/\partial k_{i,t} + \frac{1}{1+r} \partial u(c_{t+1})/\partial k_{i,t}$ for $i = L, H$. Substituting in the same expressions for marginal utility as before yields

$$r = \alpha \bar{P} (k_i^{opt})^{\alpha - 1} (c^{opt})^{-\gamma} \left[\mathbb{E}[A_i^{opt}] + \frac{z_i \sigma_\epsilon^{-2}(1-\lambda_i)}{1+r} \Big((k_i^{opt})^{\alpha} - \frac{1-\lambda_i}{\lambda_i}(k_{-i}^{opt})^{\alpha}\Big)(\Omega_i^{opt})^{-2}\right]. \qquad i = L, H \tag{40}$$

Equation (40) is different from Equation (28).

This is the second externality. With business stealing, the planner internalizes that an increase in capital of firm $i$, increases data production, which decreases the quality of every other firm in the sector. Since firms do to internalize this effect, they over-invest in capital to get more data than what is efficient. There is excessive production.

# B Model Calibration

The model has 8 parameters: $\alpha$, $\gamma$, $\rho$, $\sigma_\theta^2$, $\psi_t$ (a series), $\bar{P}$, $\bar{A}$, and $s_\Omega$, whose values are summarized in Table 1.

We calibrate the first five parameters externally, either directly from the literature or using procedures suggested in previous work. To calibrate the last three parameters, $\bar{P}$, $\bar{A}$ and $s_\Omega$, we use model equations 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). We then use the calibrated model to estimate the GDP mis-measurement due to data barter during 2003-2018. To do this, we execute the following steps.

**Conversion of nominal to real value function** Begin with the Bellman equation in which firm $i$'s value at time $t$ is the recursive and deterministic result of a firm's stock of knowledge $\Omega$ and their capital choice $k_t$. We consider an economy where symmetric firms do not trade data and assume that the data adjustment cost takes a quadratic form, $\Psi(\Omega_{i,t+1}) = \psi_t(\frac{\Omega_{i,t+1}-\Omega_{i,t}}{\Omega_{i,t}})^2$, while the quality function is linear in knowledge, a generalized version of Equation (13):

$$g(\Omega_{i,t}^{-1} + \sigma_u^2) = \bar{A} - s_\Omega(\Omega_{i,t}^{-1} + \sigma_u^2)$$

We assume $\iota = 1$ to calibrate the model so that in steady state there is no data trade. Thus, the nominal value function can be written as:

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

This measures aggregate economic value, across all firms in a given period. Let $\bar{V}_t$ denote the real value of data:

$$\bar{V}_t = \frac{V_t}{P_t}$$

The Bellman equation for firm value, normalized by the price of goods, is:

$$\bar{V}(\Omega_t) = \max_{k_t} \; (\bar{A} - s_\Omega(\Omega_{i,t}^{-1} + \sigma_u^2))k_t^\alpha - \frac{1}{P_t}(\psi_t(\frac{\Omega_{i,t+1} - \Omega_{i,t}}{\Omega_{i,t}})^2) + r_t k_t + \left(\frac{1}{1+r_t}\right) \bar{V}(\Omega_{t+1}) \tag{41}$$

The inflation rate determines the price level of goods as: $P_t = P_0 \prod_{\tau=0}^{t-1}(1 + \dot{p}_\tau)$, where the initial price is $P_0 = \bar{P}((\bar{A} - s_\Omega(\Omega_{i,0}^{-1} + \sigma_u^2))k_0^\alpha)^{-\gamma}$ and where $\dot{p}_t$ is the time-t inflation rate.

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

$$\alpha(\bar{A} - s_\Omega(\Omega_{i,t}^{-1} + \sigma_u^2))k_t^{\alpha-1} - \frac{1}{P_t}(2\psi_t(\frac{\Omega_{i,t+1} - \Omega_{i,t}}{\Omega_{i,t}^2})\frac{d\Omega_{t+1}}{dk_t}) + r_t + \frac{1}{1+r_t}\bar{V}'(\Omega_{t+1})\frac{d\Omega_{t+1}}{dk_t} = 0. \tag{42}$$

**External calibration** We calibrate five of the model parameters externality using existing values from the literature: the capital share of income $\alpha$, inverse demand elasticity $\gamma$, the AR(1) coefficients of the state $\rho$ and $\sigma_\theta^2$, and the marginal adjustment cost of data $\psi_t$.

A capital share of 40% is used in many papers, including the handbook of macroeconomics. The demand elasticity parameter $\gamma$ is the exponent on quantity in the price function. While elasticity estimates can be controversial, for our purposes, this parameter is not very important. Because it mostly just governs the price level, if we choose a different elasticity, we end of re-calibrating the price scaling parameter $\bar{P}$ to achieve the same results. That being said, we still choose the elasticity parameter with care. This price function is an inverse demand curve. It takes in the supply $Y$, which is the quantity demanded if the market clears, and spits out a price. When demand elasticity is high, it means that small changes in the price deliver large changes in quantity demanded. But in the pricing function, this implies that large changes in quantity $Y$ deliver small changes in price. Thus, $\gamma$ is the inverse of price elasticity. In the micro-founded model of Section 5, this inverse demand comes from a CRRA household utility function. The curvature in that utility function is also $\gamma$. Optimizing households purchase until price is marginal utility. The marginal utility of the household is $c^{-\gamma}$. Thus, the pricing function takes the form $p = c^{-\gamma}$. Since markets clear, this is $P = Y^{-\gamma}$. In a dynamic model, this curvature parameter $\gamma$ 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 $\gamma = 1/1.07 = 0.93$. 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, $\rho$ and $\sigma_\theta^2$, come from fitting an AR(1) process to the productivity process estimated via Fernald (2014)[12], for our sample period. 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.

The data adjustment cost parameter $\psi_t$ follows the estimation procedure in Brynjolfsson et al. (2021). It 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. The argument for why this measures AI costs is that, according to q-theory, incurring 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 5 plots the calibrated data adjustment cost series.

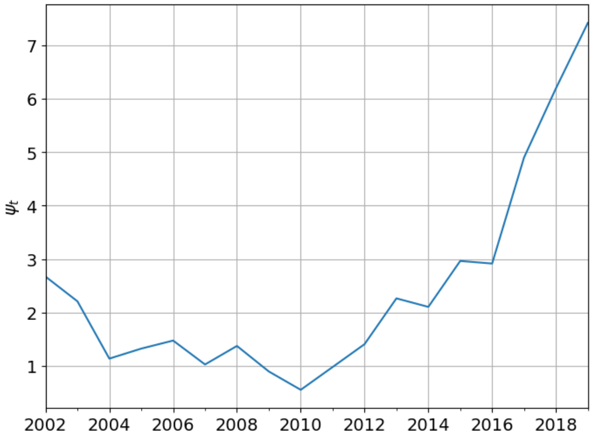

Figure 5: Calibrated adjustment cost series $\psi_t$ using the methodology in Brynjolfsson et al. (2021).

**Time-series input data** In order to calibrate the remaining parameters of the model, we uese the time-series for capital, inflation, and firms' sales forecast error.

The stock of knowledge, $\Omega_t$, is the conditional precision of firms' forecasts about the learnable component of their optimal technique $\theta_t$. Their precision is the inverse of the firms' expected squared forecast error. The technique uncertainty, $\Omega_t^1$ also has an unlearnable component $\sigma_u^2$. 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 Kohlhas and Asriyan (2024) and use the data from I/B/E/S Guidance to measure sales forecast accuracy of US public firms. I/B/E/S Guidance extracts quantitative company expectations from press releases and transcripts of corporate events. Kohlhas and Asriyan (2024) show that forecast accuracy is correlated with features of firms that ought to predict their information choices.

We retrieved the annual sales guidance from I/B/E/S Guidance via the WRDS platform on April 20, 2023. We use both the "Detail" and "Identifiers" tables. The "Detail" table contains financial estimates (e.g., annual or quarterly sales or earnings) made by firms. The "Identifiers" table contains "ticker" and "cusip" identifiers. We merge the two tables and only keep US firms (observations with "usfirm"=1). Our sample runs from 2002 to 2021. We choose 2002 as the starting date because the sales guidance data only became available for more than 10 firms after 2001. Firms revise and update their annual sales guidance over the course of the fiscal year as realized quarterly sales data comes in. Since our analysis focuses on firms' ability to forecast annual sales, we only keep the initial estimates, which are not affected by realized quarterly data. When the sales forecast is expressed as a range, we take the mid-point (the 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 downloaded from the FRED.

Average sales forecast increased from 800 million dollars in 2002 to 1600 million dollars in 2018, then declining to 1000 million dollars in 2021 because of the pandemic and adverse macroeconomic conditions. We compare firms' sales guidance with realized sales data (retrieved from Compustat North America) and define the squared relative

[12] The updated dataset is available here: https://www.johnfernald.net/TFP.

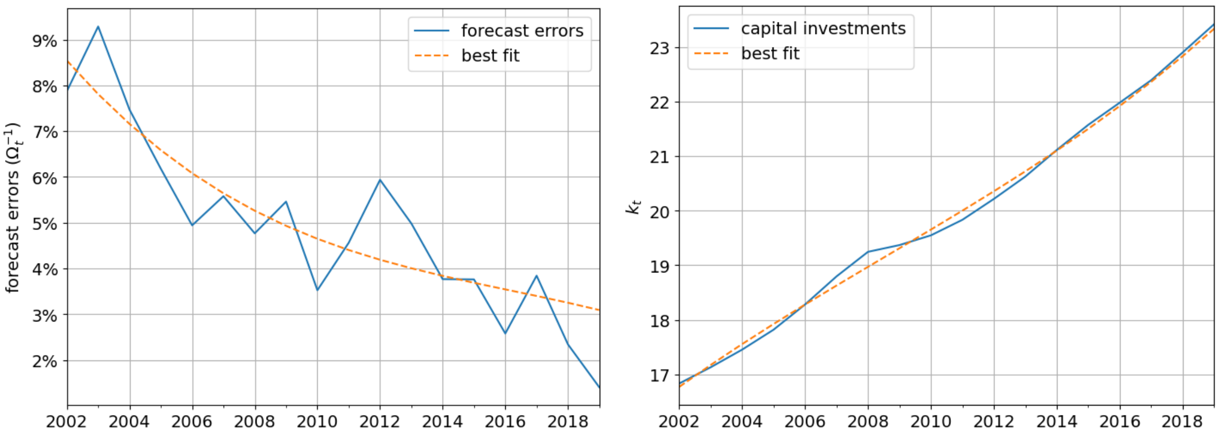

Figure 6: Time series of US public firm forecast errors calculated following Kohlhas and Asriyan (2024), using data from I/B/E/S Guidance (left) and capital stock as the net stock of nonresidential fixed assets from BEA's Fixed Assets Accounts (right).

forecast error as:

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

We winsorize the relative forecast error at the one percent level on the right tail. In each year, we take the average of the squared relative forecast error across firms. The time series is shown in Figure 6. We can see that the error decreases over time. However, there is a sudden upsurge in uncertainty during the pandemic.

We use the capital stock (net stock of nonresidential fixed assets) $k_t$ from BEA Fixed Assets Accounts (2003-2018) to determine the capital rental rate $r_t$ that rationalize the observed capital stock. (U.S. Bureau of Economic Analysis, Table 1.2. Chain-Type Quantity Indexes for Net Stock of Fixed Assets and Consumer Durable Goods, Line 4.) The reason is that when we report the value of data, as a fraction of GDP, it is a more useful measure if it doesn't have additional noise that arises from a capital series in the model that does not correspond to empirical measures. The model is not a rich enough model to capture capital dynamics well because it is designed to focus on the dynamics of data. Therefore,

We use inflation, $\dot{p}_t$, to turn the nominal value function into the real one. It is the percent change in FRED's Gross Domestic Product: Implicit Price Deflator (GDPDEF).[13]

One could go one step further and design a series of investment frictions that would produce the rental rate series $r_t$ as an endogenous outcome, that would in turn, explain the observed capital stock. But that adds complication for no additional insight. Instead, we feed in a price series that ensures capital is not a source of noise that still allows for meaningful measurement of data value. In Section B.1, we check if the resulting capital rental rates are plausible.[14]

**Calibration using model moments** 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, $s_\Omega$, and the multiplier that determines the level of goods price, $\bar{P}$. Since we want a meaningful comparison with real economic data, we choose a price multiplier that comes close to matching the level of real GDP. Then, we need two moments related to the effect of data that allow us to pin down the quality function parameters $\bar{A}$ and $s_\Omega$. We use estimates from the BEA and Gorodnichenko et al. (2023) of the aggregate and firm-specific effects of data to jointly calibrate the parameters. The specific moments we match are as follows.

As our first moment, we use the time-series of real GDP for the 2003-2018 period. We minimize the mean-squared error between realized and model estimated real GDP. Our model implied real GDP is $\max_{k_t} \; g(\Omega_t^{-1} + \sigma_a^2)k_t^\alpha - \frac{1}{P_t}\Psi(\Omega_{t+1}) - r_t k_t$. We want to minimize the MSE between the modeled and realized real GDP across the sample period. The fit to annual real GDP is shown in Figure 7.

As our second moment, we match the one-period value of data to the BEA estimate. The BEA estimated that

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

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

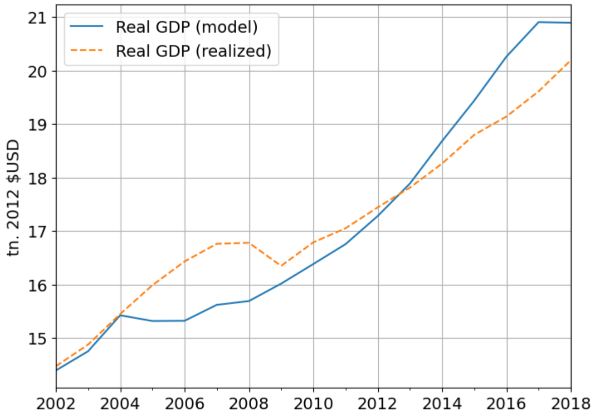

Figure 7: Calibration fit. Real GDP in the model and data.

firm investments in own-account data assets had a flow value of \$72bn in 2003 (0.25% of GDP).[15] 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 (44). The model's estimated one-period values of data are shown in the left panel of Figure 4.

The downside of this approach is that is rests on an external cost and markup-based approach to data valuation. The upside is that it is a value clearly related to data. We calibrate to the 2003 data value and let the model predict the rest of the evolution of data value. After a couple of decades, the data value is not so sensitive to the starting value.

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 $\Omega^{-1}$ by 1pp and use Equation 42 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.

*Estimation details.* For each guess of $\bar{A}$, $s_\Omega$, and $\bar{P}$, we estimate $\bar{V}$ and $r$ iteratively using Equations (41) and (42). We start with an initial guess of $\bar{V}$ which is strictly increasing on $\Omega$ and solve for its sequence of implied rental rates. $\bar{V}$ and $r$ are continuously reestimated until $\bar{V}$ converges. We run this estimation procedure on a grid of $\bar{A}$, $s_\Omega$, and $\bar{P}$ and select the combination of parameters which optimize the criteria listed above.

**Valuing data** The reason for doing this calibration is to use the model to estimate the value of data. Specifically, we compute the value of all data received by all firms, in a year. There are two ways to express the value of data. One is the value the firm derives from the data, in the current period. The other is to compute the present discounted value of all the revenue that will be derived from the data they currently own, in all future periods. We report both in Figure 4.

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:

$$\tilde{V}(\Omega_t) = \max_{k_t} g(\Omega_t) k_t^\alpha - \frac{1}{P_t}\Psi(\tilde{\Omega}_{t+1}) - r_t k_t + \frac{1}{1 + r_t}\bar{V}(\tilde{\Omega}_{t+1}) \tag{43}$$

The difference between the actual real value of data, $\bar{V}(\Omega_t)$, and this counter-factual value, $\tilde{V}(\Omega_t)$, and is the net present discounted value of the bartered data acquired in period $t$:

$$pdv(\Delta\Omega_t) = \bar{V}(\Omega_t) - \tilde{V}(\Omega_t).$$

Alternatively, the value derived from a year of data in one period (opv) is the same as $pdv(\Delta\Omega_t)$ at time $t$ (when the usable data is the same) and $t + 1$ (when the lost data compromises product quality), but reverts to its original value

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

in period $t+2$:

$$\begin{aligned} opv(\Delta\Omega_t) = & \max_{k_t} g(\Omega_t)k_t^\alpha - \frac{1}{P_t}(\Psi(\Omega_{t+1}) + r_t k_t) \\ & + \bar{V}(\Omega_t) - \frac{1}{1+r_t}[\max_{k_{t+1}} A(\tilde{\Omega}_{t+1})k_{t+1}^\alpha - \frac{\Psi(\Omega_{t+2})}{P_{t+1}} + r_{t+1}k_{t+1} + \frac{1}{1+r_{t+1}}\bar{V}(\Omega_{t+2})] \end{aligned} \tag{44}$$

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.

## B.1 Cost of Capital

**Over-identifying moment** In our calibration exercise, we do not match $r_t$ to any sequence of rental rate or cost of capital. Instead we compare the estimated sequence of $r$ to the 2003-2018 U.S. cost of capital, estimated by Aswath Damodaran.[16] The comparison between model-implied rental rates and measured firm cost of capital is shown in Figure 9.

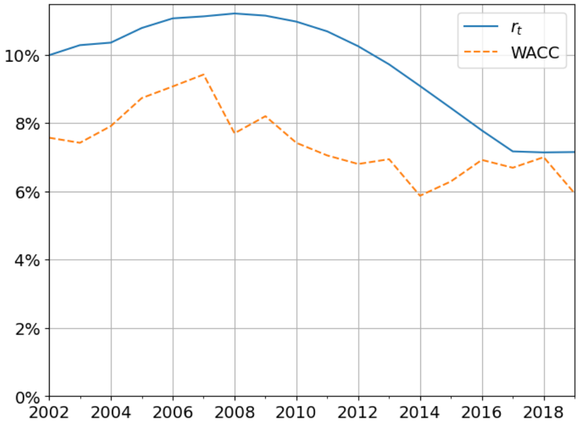

Figure 8: Model-implied compared to empirical estimates of real cost of capital.

The results suggest that our simple approach to calibrating capital is not perfect, but is unlikely to grossly mis-estimate aggregate values.

**Model predicted versus empirical growth of stock of knowledge** Our calibration strategy is designed to measure and value data, not to predict it. As such, we set the stock of knowledge for the representative firm to match the measured forecast errors or an average firm, to value a measured amount of data. However, the model also predicts the amount of data that a firm should have.

Instead of feeding in an exogenous (empirically observed) data series as we have done in the calibration exercise, one can start the model off with the same initial conditions and predict the amount of data. Figure 3 in the main text reports the results of this exercise. In the endogenous data model, the stock of data, which is the precision of the forecast, rises from 4 to 11, representing a 2-3 fold increase. In the data series used for calibration, the squared forecast error in Figure B fall from about 9% to 4% in the first 10 periods (2003-2013), representing a doubling of precision. The similarity between the predicted and measured changes in firm forecast accuracy are one more piece of evidence that the model is a useful measurement tool.

## B.2 The Importance of the Depreciation Rate

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 4 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 high.

[16] See https://pages.stern.nyu.edu/ adamodar/New_Home_Page/datafile/wacc.html for detail of data construction.

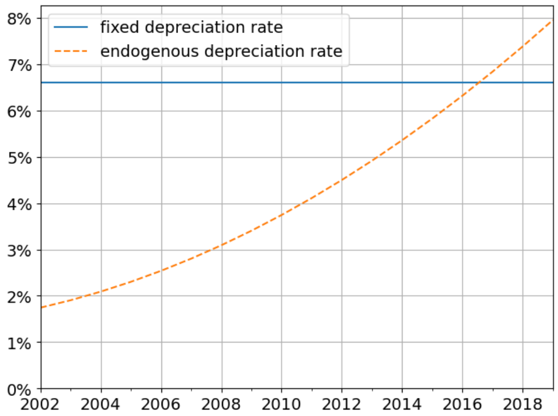

Figure 9: Model-implied versus constant annual data depreciation rate.

## B.3 Robustness

**Constant data adjustment cost** Since the estimated adjustment costs vary greatly over this time period, one might be concerned that these fluctuations drive the results. Therefore, we fix the adjustment cost at the constant level reported by Brynjolfsson et al. (2021) that best fits the data in a pooled regression of firm market value on its R&D expenses. This value is $\psi = 2.73$. Then, we follow the same procedure described above to re-calibrate the model. Three additional parameters change: $\bar{P}$, which governs the price level of goods is 3.77; $\bar{A}$ and $s_\Omega$, which represent the quality function intercept and slope are now 1.23 and $-2.52$.

Both the one-period and present value of the data are between 50-100% higher in this calibration. The trajectories are similar. The estimates in the main text are therefore conservative estimates of GDP mis-measurement.

**Interest rate (rental rate of capital)** Understanding the sensitivity of results to the interest rate is of particular importance because most of the analysis holds the interest rate fixed. The exercises below allow us to understand: If we model capital markets and allow $r$ to be determined endogenously, how much might this matter for the value of data? To answer this, we explore halving the interest rate, doubling the interest rate, and shocking the interest rate halfway through the simulation. We find that our results are surprisingly insensitive to the interest rate. This gives us confidence that endogenizing the interest rate would have little impact on our results.

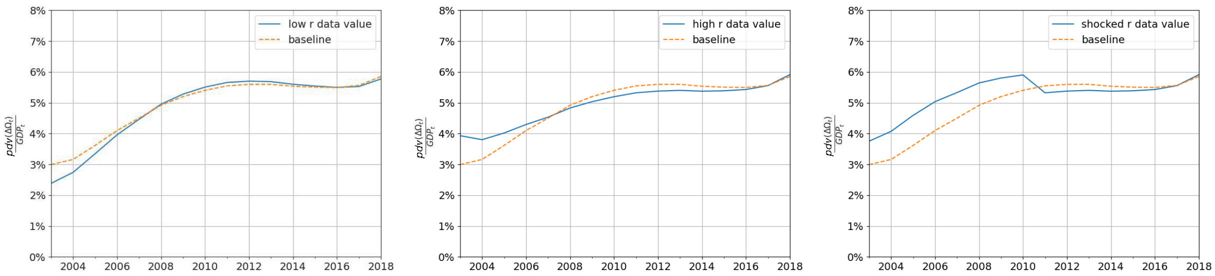

Figure 10: Interest rate insensitivity. Estimated GDP mis-measurement that comes from bartered data, with a lower (left), higher (middle) and shocked (right) interest rate. All panels report the present value of data generated each year, $pdv(\Delta\Omega_t/GDP_t)$. Left panel is half of calibrated $r_t$ series, reported in Figure 9. Middle panel uses $2\,r_t$. The right panel uses $r_t$ until 2010, where we double the interest rate and use $2r_t$ thereafter.

Figure 10 shows that the results are not very sensitive to the interest rate. While the value of data in the low interest rate regime is lower at the start when data is scarce, by 2006, the results are nearly identical. The reason interest rate sensitivity is high at the start is that 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 hardly matters. The right panel of Figure 10 shows that even doubling the interest rate in the middle of the simulation affects data value by less than 1% of GDP.

**One-period data value** One of the trickier moments to match to the model is the estimate of the one period value of data. It is surely mismeasured in some way. That might seem problematic because changing the initial value of data might shift the present discounted value of data up or down for the entire simulation. However, if we change the one-period data value and, at the same time, allow the other parameters to be jointly re-estimated, the effect is modest and fades away after a few years.

To show this, we ran a series of experiments. Initially, a halving of the one-period data value raises the 2003 present value of data by only one-half of one percent. When we calibrate to $opv_{2003} = 0.125\%$, instead of 0.25%, we obtain parameter estimates of $\bar{A} = 1.24$, $s_\Omega = -2.41$ and $\bar{P} = 7.34$. All three parameters become greater in magnitude. However, the largest change is in the price of physical goods parameter $\bar{P}$. That makes physical goods more valuable and brings down the one-period value of data, as a fraction of measured gdp. However, in the long run, the greater value of good raises the value of data, which helps to produce those goods more efficiently, which raises the value of data.

More importantly, from 2008 on, the present value of data is indistinguishable from the baseline results. The 2018 present value of data is 5.82%, which is within a tenth of a percentage point of the original finding. Exploring the space of initial data value calibration targets shows that the invariance of longer-run data present value is generally insensitive to the one-period value calibration target. When we take the initial data value all the way up to 1% of GDP, the re-calibrated parameters become $\bar{A} = 1.31$, $s_\Omega = -4.02$ and $\bar{P} = 1.79$. However, the 2018 pdv of data is still just below 5%, falling within 1% of the initial estimate.

In short, while we need some initial data value to run the exercise, the importance of this initial value quickly fades. The long-run present value of data is governed by the sensitivity of investment to data, not by the initial one-period value.