Big Data and Firm Dynamics

By Maryam Farboodi, Roxana Mihet, Thomas Philippon, and Laura Veldkamp*

How does data affect firm dynamics? As markets become more concentrated and dominated by data-savvy firms, it is important to understand the macroeconomic role of data. While Brynjolfsson and McElheran (2016) have already found a connection between information technology and market concentration, a macro framework allows economists to value firms' data, perform counter-factuals, and understand why data has the effects it does. Therefore, we build a framework for measurement and policy analysis.

We argue that data has four important features: (i) data is a by-product of economic activity; (ii) firms use data to increase their efficiency; (iii) data is information, which is distinct from technology; and (iv) accumulated data is a valuable asset. Our objective is to write the simplest framework that includes these features.

We build a model where heterogeneous price-taking firms invest, produce, and accumulate data. Data causes long-lived firms to grow bigger for two reasons. First, data helps firms become more productive. Productive firms invest more, grow larger, and produce more data. This is a "data feedback loop." Second, firms invest more than they otherwise would because additional production generates more data. This is "active experimentation." We also learn that initial size is not the most important factor in the success of a firm. A small firm that uses data efficiently, meaning that it harvests more data per unit of production, may lose money initially while it builds up its data stock. But if the firm can finance this phase, it can quickly outcompete a larger, less data-efficient firm.

* Corresponding author: Laura Veldkamp, Columbia Business School, 3022 Broadway, NY, NY 10027, lv2405@columbia.edu; Farboodi: MIT Sloan, 30 Memorial Drive, Cambridge, MA 02142, farboodi@mit.edu; Mihet: NYU Stern, 44 W4th St, NY, NY 10012; Philippon: NYU Stern, 44 W4th Street, NY, NY 10012.

We build on ideas of others who studied information in the macroeconomy. The growth and learning-by-doing literatures model data as technology-augmenting.[1] Modeling data as information about the firm's optimal choice, allows us to incorporate countervailing forces like the diminishing returns to data.[2] In Veldkamp (2005) and Fajgelbaum, Schaal and Taschereau-Dumouchel (2017) information is a by-product of economic activity and a signal, but the focus is on asymmetric cyclical fluctuations. Our framework differs because of its growth in data and its incorporation of heterogeneous firms. Both are essential to study changing firm dynamics.

I. Setup

We consider a competitive industry. Time is discrete and infinite. There is a continuum of firms indexed by i. Firm i uses $k_{i,t}$ units of capital to produce $k_{i,t}^{\alpha}$ units of goods of quality $A_{i,t}$. Let P_t denote the equilibrium price of quality-adjusted goods. The inverse demand function and the industry quality-adjusted supply are:

$$P_t = \bar{P} Y_t^{-\gamma}, \tag{1}$$

$$Y_t = \int_i A_{i,t} k_{i,t}^{\alpha} di. \tag{2}$$

Firms take the industry price P_t as given and their quality-adjusted outputs are perfect substitutes.

Quality depends on a firm's choice of a production technique $a_{i,t}$. In each period, and for each firm, there is one optimal technique with a persistent and a transitory components: $\theta_{i,t} + \epsilon_{a,i,t}$. The persistent component $\theta_{i,t}$ is unknown and follows an AR(1) process: $\theta_{i,t} = \bar{\theta} + \rho(\theta_{i,t-1} - \bar{\theta}) + \eta_{i,t}$

[1] Jones and Tonetti (2018), Jovanovic and Nyarko (1996), among many others.

[2] Chiou and Tucker (2017) and Bajari et al. (2018).

where $\eta_{i,t}$ is *i.i.d.* across time and firms. The transitory shock $\epsilon_{a,i,t}$ is *i.i.d.* across time and firms and is unlearnable. Deviating from that optimum incurs a quadratic loss in quality:

$$A_{i,t} = \bar{A}_i \left[\hat{A} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2\right]. \tag{3}$$

Data helps firms infer $\theta_{i,t}$ The role of ϵ_a is to prevent firms from inferring $\theta_{i,t}$ at the end of each period. It makes the accumulation of past data a valuable asset. If a firm knew the current value of $\theta_{i,t}$, it would maximize quality by setting $a_{i,t} = \theta_{i,t}$.

A key idea of our model is that data is a by-product of economic activity. Therefore, we assume that the number of data points observed by firm i at time t depends on their $t-1$ production $k^{\alpha}_{i,t-1}$:

$$n_{i,t} = z_i k^{\alpha}_{i,t-1}, \tag{4}$$

where z_i is the parameter that governs how "data-savvy" a firm is. A data-savvy firm is one that harvests lots of data per unit of output.

Each data point $m \in [1 : n_{i,t}]$ reveals

$$s_{i,t,m} = \theta_{i,t} + \epsilon_{i,t,m}, \tag{5}$$

where $\epsilon_{i,t,m}$ is *i.i.d.* across firms, time, and signals. For tractability, we assume that all the shocks in the model are normally distributed: fundamental uncertainty is $\eta_{i,t} \sim N(\mu, \sigma^2_\theta)$, signal noise is $\epsilon_{i,t,m} \sim N(0, \sigma^2_\epsilon)$, and the unlearnable quality shock is $\epsilon_{a,i,t} \sim N(0, \sigma^2_a)$.

Firm Problem. A firm chooses a sequence of production and quality decisions $k_{i,t}, a_{i,t}$ to maximize

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(P_t A_{i,t} k^{\alpha}_{i,t} - r k_{i,t}\right) \tag{6}$$

Firms update beliefs about $\theta_{i,t}$ 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 $a_{i,t}$ is $\mathcal{I}_{i,t} = [\{A_{i,\tau}\}^{t-1}_{\tau=0}; \{\{s_{i,\tau,m}\}^{n_{i,\tau}}_{m=1}\}^{t}_{\tau=0}]$.

Solution

The state variables of the recursive problem are the prior mean and variance of beliefs about $\theta_{i,t-1}$, last period's revenues, and the new data points. Taking a first order condition with respect to the technique choice, we find that the optimal technique is $a^*_{i,t} = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. Let the posterior variance of beliefs be $\Sigma_{i,t} := \mathbb{E}_i[(\mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}] - \theta_{i,t})^2]$. Thus, expected quality is $\mathbb{E}_i[A_{i,t}] = \bar{A} - \Sigma_{i,t} - \sigma^2_a$. We can thus express expected firm value recursively.

LEMMA 1: *The optimal sequence of capital investment choices $\{k_{i,t}\}$ solves the following recursive problem:*

$$\begin{aligned} V_t(\Sigma_{i,t}) = \max_{k_{i,t}} \; & P_t \left(\bar{A} - \Sigma_{i,t} - \sigma^2_a\right) k^{\alpha}_{i,t} \\ & - r k_{i,t} + \beta V_{t+1}(\Sigma_{i,t+1}) \end{aligned} \tag{7}$$

where $n_{i,t+1} = z_i k^{\alpha}_{i,t}$ and

$$\Sigma_{i,t} = \frac{1}{\left[\rho^2(\Sigma^{-1}_{i,t-1} + \sigma_a^{-2})^{-1} + \sigma^2_\theta\right]^{-1} + n_{i,t}\sigma_\epsilon^{-2}} \tag{8}$$

See online Appendix for the proof. This result greatly simplifies the problem by collapsing it to a deterministic problem with only one state variable, $\Sigma_{i,t}$. The reason we can do this is that quality $A_{i,t}$ depends on the conditional variance of $\theta_{i,t}$ and because the information structure is similar to that of a Kalman filter, where the sequence of conditional variances is generally deterministic.[3] This Kalman system has a 2-by-1 observation equation, with $n_{i,t}$ signals about $\theta_{i,t}$ and one signal about $\theta_{i,t-1}$. The signal about $\theta_{i,t-1}$ comes from observing last period's output, which reveals quality $A_{i,t-1}$,

[3] For any $k_{i,t}$, the optimal choice of technique is always the same: $a^*_{i,t} = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. The way $a_{i,t}$ enters into expected quality $A_{i,t}$ is through $\mathbb{E}[(\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}] - \theta_{i,t})^2]$, which is the conditional variance $\Sigma_{i,t}$. We can replace the entire sequence of $a^*_{i,t}$ with the sequence of variances, which is deterministic here because of normality. The only randomness in this model comes from the signals and their realizations, but they never affect the conditional variance, since normal means and variances are independent. Thus, given $\Sigma_{i,t-1}$, $\Sigma_{i,t}$ is a sufficient statistic for $n_{i,t}$ and $\Sigma_{i,t+1}$. The mean $\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}]$ is not a state variable because it only matters for determining $a_{i,t}$ and does not affect anything else.

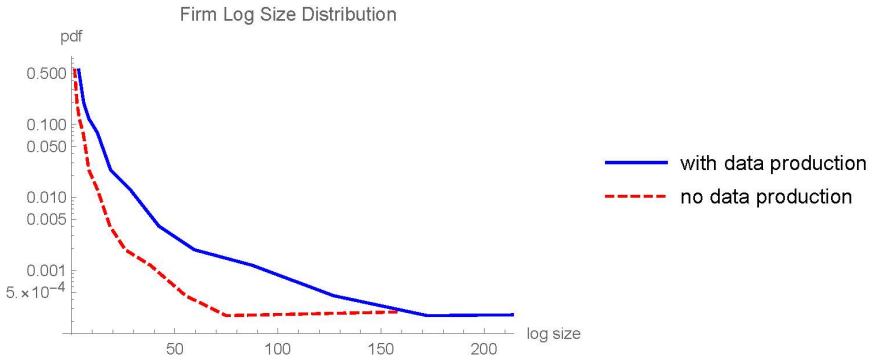

FIGURE 1. STEADY STATE SIZE DISTRIBUTION $k_{i,ss}$ OF FIRMS WITH AND WITHOUT DATA.

Dashed line is without data production ($z = 0$, $\bar{A}_i$'s calibrated to match U.S. firm sizes in 2016). Solid line is with data production ($z = 1$). Parameter values are: $\alpha = 0.5$, $\beta = 0.98$, $\gamma = 0.7$, $\rho = 0.5$, $\hat{A} = 3$, $\bar{P} = 1$, $r = 0.2$, $\sigma_a^2 = 1$, $\sigma_e^2 = 1$, $\sigma_\theta^2 = 1$.

which, in turn, reveals $\theta_{i,t} + \epsilon_{a,i,t}$.[4]

From this recursive expression, we can value data. The marginal value of an additional unit of data, as measured in units of forecast precision is $\partial V_{i,t}/\partial \Sigma_{i,t}^{-1} = \Sigma_{i,t}^2 \left[P_t k_{i,t}^\alpha - \beta V'_{i,t+1}(\Sigma_{i,t+1}^{-1}) \frac{d\Sigma_{i,t+1}}{d\Sigma_{i,t}} \right]$ where $\frac{d\Sigma_{i,t+1}}{d\Sigma_{i,t}} = \Sigma_{i,t+1}^2 \left[\rho^2 (\Sigma_{i,t}^{-1} + \sigma_a^{-2})^{-1} + \sigma_\theta^2 \right]^{-2} \rho^2 (\Sigma_{i,t}^{-1} + \sigma_a^{-2})^{-2} \Sigma_{i,t}^{-2}$.

The solution to the firm's investment problem comes from the first-order condition, $\partial V_{i,t}/\partial k_{i,t} = 0$ and the Euler equation, $\alpha p_t \left(\bar{A} - \Sigma_{i,t} - \sigma_a^2 \right) + \beta V'_{i,t+1} \frac{\partial \Sigma_{i,t+1}}{\partial k_{i,t}} = r k_{i,t}^{1-\alpha}$. Substituting expressions above yields

$$rk_{i,t}^{1-\alpha} = \alpha P_t \left(\bar{A} - \Sigma_{i,t} - \sigma_a^2 \right) + \alpha\beta z_i \Sigma_{i,t+1}^2 \sigma_\epsilon^{-2} P_{t+1} k_{i,t+1}^\alpha \quad (9)$$

The first term on the right is the added contemporaneous value from additional investment. The second term represents gains from experimentation. Firms invest more to improve their future data set.

DATA CHANGES STEADY-STATE FIRM SIZE

Our first numerical experiment studies how improvements in firm data processing are changing the size distribution of firms. We start by calibrating firm sizes in a model with no data processing, to match the size distribution of U.S. firms. Then, we turn on data processing to observe how sizes change. For most parameters, we choose round numbers that deliver sensible outcomes.

What governs the steady-state size of a firm is its product quality parameter $\bar{A}_i$. We choose twelve levels of $\bar{A}_i$ as follows: We match the steady-state $k_{i,t}$ of a firm, with that $\bar{A}_i$ and with no data production ($z_i = 0$), to the average size of the firm in each of twelve firm-size categories, as defined by the 2016 Longitudinal Business Database. Each of these sizes has a market share associated with it, which is the number of firms in that size category, divided by the total number of firms in 2016. In Figure 1, the dashed line labeled "no data production" plots size and market share, as reported by our data.

To compute the change in size of firms when all firms process data, we change one parameter and re-compute the steady state. Instead of $z = 0$, we set $z = 1$ for all firms, so that production generates usable data. Figure 1 shows that the new firm size distribution (solid line, labeled "with data production") has more large firms. The very largest firms get substantially larger.

While one expects larger firms with more

[4]Firms observe $(\theta_{i,t} + \epsilon_{a,i,t})^2$. For tractability, we assume that firms know whether the root is positive or negative. For more on this and for the derivation of the belief updating equations, see online Appendix.

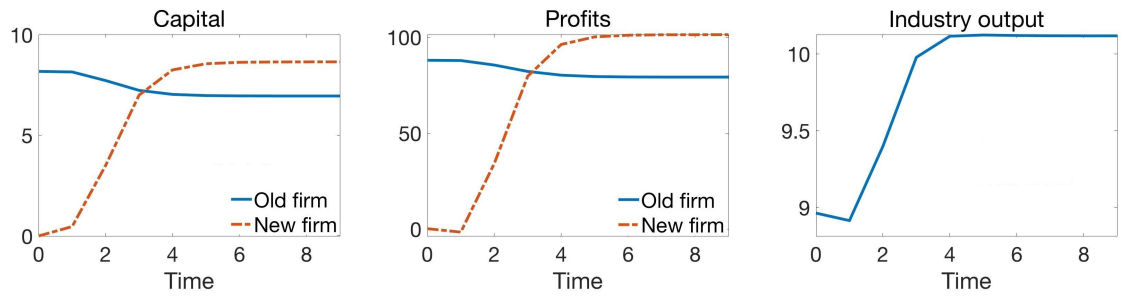

FIGURE 2. DYNAMICS WHEN NEW, DATA-SAVVY FIRMS ENTER AT TIME $t = 0$.

Solid line is old incumbents with z_{old}. Dashed line is new, data-savvy entrants with $z_{new} > z_{old}$. Parameters are as in Figure 1, except $\bar{A} = 3$, $\hat{A} = 2.5$, $\gamma = 0.5$, $\bar{P} = 10$ and $\rho = 0.99$, for all firms.

data to benefit most, there is a counteracting force. Bayes' law tells us that each unit of data increases the precision of a forecast by less and less. Diminishing returns to data works against increasing returns to scale. Our results suggest that scale wins. Of course, to make precise quantitative statements, future work should calibrate the model more carefully. But the exercise illustrates how one might use the framework for measurement and makes the point that the effect of data on firm size may be sizable.

ENTRY OF NEW TECH FIRMS

We learned that data processing makes firms larger. But what are the dynamics of the transition to the new big-data steady state? To learn about how the economy might behave in the transition, we consider two types of firms. Type *old* firms are old-economy incumbents that either do not generate much data or do not make good use of the data they get. These firms have a low z_{old}, meaning that they get few data points, per unit of output. Type *new* firms are data-savvy and have higher value of z_{new}. They start small, with low capital, but they scrape lots of data from the transactions that they generate.

We consider an industry in steady state with a mass one of identical, old-economy firms. We then drop a mass M of new, data-savvy firms that have not accumulated any data yet. We solve for the dynamic transition path to the new steady state, with both types of firms in the economy.[5]

When new firms enter, they have no data to guide their choice of technique. They do not know what consumers want, thus they *experiment*. They supply goods and services of random quality and do not generate much contemporaneous value, on average. Figure 2 (dashed lines) shows the initial low capital investment and negative profits of the entrants. But these new firms learn quickly over time. As they accumulate data, their productivity improves. They scale up and generate even more data.

As soon as new firms enter, the market value of old firms drops. Old firms anticipate the rise of the new firms and expect their capital to generate less profit in the future. They cut investment and produce less. The output quality of new firms is initially low, so the industry-wide, quality-adjusted output initially falls (Figure 2, right panel) and the industry price initially rises. Output then expands as new firms learn, improve their quality and invest more.

[5]There is a continuum of old and new firms so we can apply the law of large numbers to each group. The industry equilibrium (output and price) is deterministic, although individual firms' output and productivity are random. We solve for eight unknowns, P_t, Y_t, $A_{old,t}$, $A_{new,t}$, $\Sigma_{old,t}$, $\Sigma_{new,t}$, $k_{old,t}$, and $k_{new,t}$. The old firms start with the stock of capital (and data) they had in the old industry steady state. The new firms start with $k_{new,0}$ close to zero, thus they have little data. Each type of firm has its own Bellman equation and anticipates correctly the future path of the price level.

How quickly the entrants overtake the incumbents depends on their data accumulation advantage $z_{new} - z_{old}$, as well as on the persistence of the optimal technique ρ. If the state is persistent, old data remains useful and it takes more time for the entrant to overtake the incumbents. In fast-changing environments, new data is more valuable and the transition is quicker.

II. Future Research

Data is changing how firms operate and compete. We offered a simple framework that can help us to think about these changes systematically. One could use this framework for many purposes. One would be to estimate the value of data. If we rewrite the value function as a function of precision, $V(\Sigma^{-1})$, rather than variance, then $V'(\Sigma^{-1})$ is the marginal value of additional data precision. Since Bayes' law tells us that precision is linear in signals, this is equivalent to measuring the marginal value of data, where the quantity of data has some natural economic interpretation as additional units of forecast precision.

Another feature that would be natural to add is to relax the perfect competition assumption and explore strategic firm behavior. Surely, there would be some interaction between market structure and the effects of data. Similarly, data can be bought and sold. One could add to (6) a term representing revenues from selling or a cost of buying additional data. That term would be a price of data, times a net quantity of data transferred. Small firms' ability to buy data could change the competitive benefits of size.

Policy questions about data regulation abound. Without equilibrium reasoning, it is difficult to say much about potential consequences. A model like this can help us think through the non-obvious consequences of market-wide regulatory changes. Theory models of big data are essential because theory guides thinking in environments where the future may look quite different from the past.

REFERENCES

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

Brynjolfsson, Erik, and Kristina McElheran. 2016. "The Rapid Adoption of Data-Driven Decision-Making." *American Economic Review*, 106(5): 133–39.

Chiou, Lesley, and Catherine Tucker. 2017. "Search Engines and Data Retention: Implications for Privacy and Antitrust." National Bureau of Economic Research Working Paper 23815.

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

Jones, Charles, and Christopher Tonetti. 2018. "Nonrivalry and the Economics of Data." Society for Economic Dynamics 2018 Meeting Papers 477.

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

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