A Model of the Data Economy

Maryam Farboodi Laura Veldkamp

Is the Data Economy New?

- the economy is changing and we need new tools!
  - the largest firms are valued primarily for their data
  - do the economics change? or is data just new capital?
- challenges
  - economic activity generates informative data

    production is a form of *active experimentation*
  - data is a non-rival good whose value declines when it is sold

    $\rightarrow$ semi-rival
  - value of data: a piece of data is used for multiple periods, how much is it valued?

    $\Rightarrow$ dynamic programming with information as a state variable
  - data depreciation rate depends on economic conditions

This Paper

- theoretical framework to think about the key economic forces
  - useful to think about data markets, policy and measurement
  - have realistic predictions
- model: recursive framework, as tractable as standard DSGE
  - values data and data-intensive firms
  - values zero-price data and digital services
  - informs GDP measurement

A Macro Model of Data

- continuum of competitive firms i
- each uses capital $k_{i,t}$ to produce $k_{i,t}^{\alpha}$ units of goods
- these goods have quality $A_{i,t}$
- Output / demand

$$Y_t = \int_i A_{i,t} k_{i,t}^{\alpha} di$$
$$P_t = \overline{P} Y_t^{-\gamma}$$

Model: Quality Depends on Forecasts

- firm has one optimal technique: $\theta_t + \varepsilon_{a,i,t}$
  - θ_t: AR(1), innovation $\eta_t \sim N(\mu, \sigma_\theta^2)$

$$\theta_t = \bar{\theta} + \rho(\theta_{t-1} - \bar{\theta}) + \eta_t$$

  - $\varepsilon_{a,i,t} \sim N(0, \sigma_a^2)$ is unlearnable and i.i.d.
  - quality depends on chosen production technique $a_{i,t}$ and distance to optimum $(\theta_t + \varepsilon_{a,i,t})$:

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

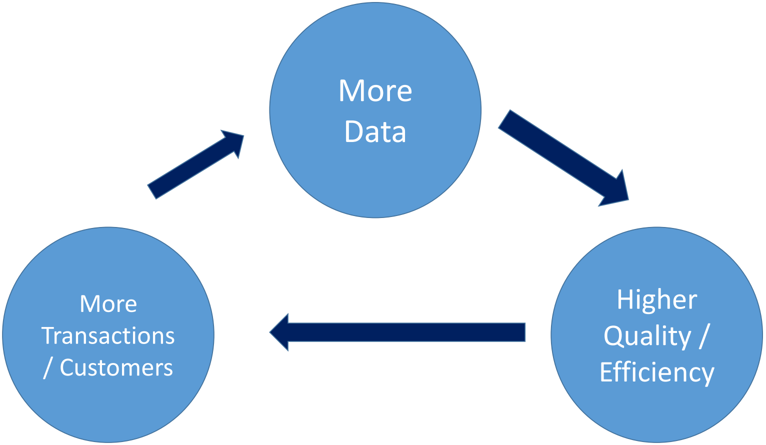
- $g(.)$: monotonically decreasing (accuracy is good)

Model: Data is Information for Forecasting

- at time t, firm obtains $n_{i,t}$ data points about θ_{t+1}
  - $n_{i,t} = z_i k_{i,t}^{\alpha}$
  - data is a bi-product of production with **data-mining ability** z_i
- each data point $m \in [1 : n_{i,t}]$ reveals

$$s_{i,t,m} = \theta_{t+1} + \xi_{i,t,m} \qquad \text{where} \qquad \xi_{i,t,m} \sim N(0, \sigma_{\varepsilon}^2)$$

Data Feedback Loop



MODEL: MARKET FOR DATA

- $\delta_{i,t}$: amount of data traded by firm *i* at time *t*
  - $\delta_{i,t} > 0$: data purchases (< 0: data sales)
  - firm can buy or sell, not both
- data price π_t clears the data market
- multi-use data: firm can sell it and still use it
  - **ι: fraction of sold data that is lost** ($\iota > 0$)
  - many data contracts include prohibitions on seller use, or this captures imperfect competition
- data adjustment cost: $\Psi(\cdot)$: avoid 1-period convergence

Results Overview

- data is an asset: depreciate and value it
- what happens in the long run?
  - diminishing returns: no long-run growth without innovation
  - endogenous growth: data ladder
- what happens in the short run?
  - increasing returns, negative initial losses
  - data barter and book-to-market dynamics
- welfare and business stealing

DATA DEPRECIATION: BAYES LAW

- goal is to forecast

$$\theta_{t+1} = \bar{\theta} + \rho(\theta_t - \bar{\theta}) + \eta_t \qquad\qquad \eta_t \sim N(\mu, \sigma_\theta^2)$$

- priors: $E[\theta_t|\mathscr{I}_t]$ and $V[\theta_t|\mathscr{I}_t] := \Omega_t^{-1}$
- Ω_t: "stock of knowledge"
- Bayes law for normal variables

  posterior precision $=$ prior precision $+$ signal precision
- law of motion for stock of knowledge (Kalman filter, Ricatti eqn):

$$\Omega_{t+1} = (\rho^2 \Omega_t^{-1} + \sigma_\theta^2)^{-1} + \text{signal precision}$$

discount more when

1) persistence is low, $\rho \downarrow$;   2) innovation is volatile, $\sigma_\theta^2 \uparrow$

Valuing Data: A Recursive Solution

- $a^*_{i,t} = \mathbb{E}[\theta_t + \varepsilon_{i,t} | \mathscr{I}_{i,t}] \to$ Quality $A_{i,t} \approx$ a fn of squared forecast error

- **state variable: stock of knowledge**

$$\Omega_{i,t} \equiv \mathbb{E}\left[\left(\mathbb{E}[\theta_t | \mathscr{I}_{i,t}] - \theta_t\right)^2 | \mathscr{I}_{i,t}\right]^{-1} \qquad \text{(posterior precision)}$$

Lemma

optimal sequence of capital / data choices $\{k_{i,t}, \delta_{i,t}\}$ solves:

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

where (Kalman filter)

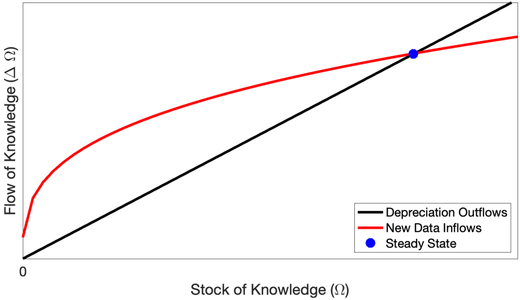
$$\Omega_{i,t+1} = \left[\rho^2(\Omega_{i,t} + \tilde{\sigma}_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left(z_i k_{i,t}^{\alpha} + \delta_{it}(\mathbf{1}_{\delta_{it}>0} + \iota \mathbf{1}_{\delta_{it}<0})\right) \sigma_\varepsilon^{-2}$$

Semi-Rivalry and Data Market

- benefit to buying one unit of data: $V'(\Omega_t) - \pi_t$
- cost of selling one unit of data: $-\iota V'(\Omega_t) + \pi_t$
- *negative* bid-ask spread
- data market active even in steady state with identical firms

Understanding Growth. Data Inflows and Outflows

- **inflow**: $z_i k_{it}^{\alpha} \sigma_{\varepsilon}^{-2}$ (# of data points $\times$ precision)
- **outflow**: data depreciation



- steady state: inflows $=$ outflows $\rightarrow$ **growth stops**

How General is Diminishing Returns?

for sustained growth $g_t > \underline{g} > 0$:

Proposition

1. **infinite output from one-period-ahead forecasts:** *the quality function has to approach infinity*
2. **no fundamental randomness:** *even if $g(0) \to \infty$, quality function $g\Big((a_{t+1} - \theta_{t+1} - \varepsilon_{a,t+1})^2\Big)$ has no time-t fundamental randomness*

ENDOGENOUS GROWTH

- alternative quality formulation: data for idea creation

$$A_{i,t} = A_{i,t-1} + \max\{0, \hat{\Delta} A_{i,t}\}$$
$$\hat{\Delta} A_{i,t} = \bar{A} - (a_{i,t} - \theta_t - \varepsilon_{a,i,t})^2$$

- data increases step size in a quality ladder $\rightarrow$ growth
- data reduces the variance: R&D that focuses on risk-reduction

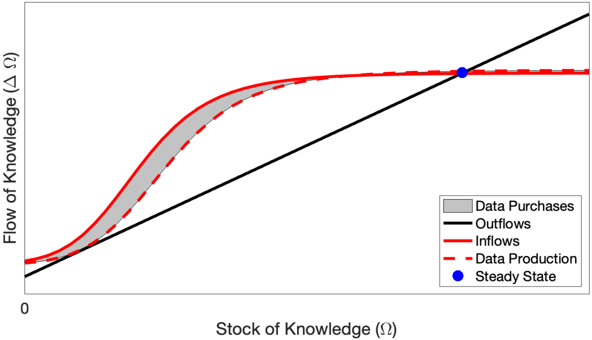
long run: data looks similar capital (except the data market)

Short Run: Increasing Returns

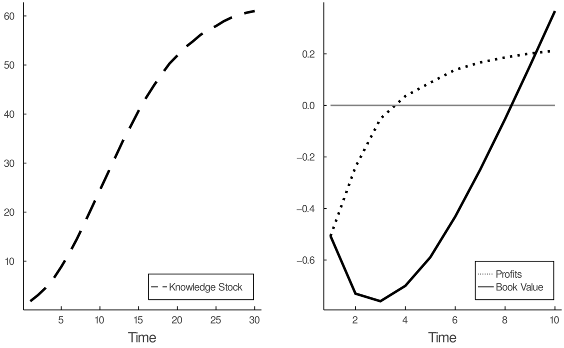
single firm enters a steady state

Proposition (Convex Data Flow)

there exist parameters such that when knowledge is scarce $\Omega_{it} < \hat{\Omega}$, *net data flow* $d\Omega_{it}$ *increases over time.*



Initial Losses and Low Book-to-Market



- early profit losses are an investment in data: Amazon!
- book value: only includes purchased data
- $\tilde{v}_{it} =$ pdv cost of purchased data, up until date $t =$ Book Value$_t$

Data Barter. Why Produce At a Loss?

- *barter*: data is "exchanged" for the good
  - at good price $P_t = 0$
- **result:** data barter arises early in a firm's life
  - firms produce goods at a loss to generate data

$$\partial V_t / \partial \Omega_{i,t} > 0$$

- reality: lots of data is bartered for services (phone apps)
- GDP is missing lots of digital economic activity because price does not reflect value

DECENTRALIZED PROBLEM: 2 TYPES OF FIRMS

- **household problem**

$$\max_{c_t, m_t} \sum_{t=0}^{+\infty} \frac{u(c_t) + m_t}{(1+r)^t}$$

$$\text{s.t.} \quad P_t c_t + m_t = \Phi_t = \text{aggregate profits of all firms} \qquad \forall t$$

- **(retail) firm problem: efficient and inefficient data-miners**

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

$$\Omega_{i,t+1} = \left[\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2 \right]^{-1} + \left(z_i k_{it}^{\alpha} + \delta_{it} (\mathbf{1}_{\delta_{it} > 0} + \iota \mathbf{1}_{\delta_{it} < 0}) \right) \sigma_\varepsilon^{-2}$$

- **market clearing**

$$c_t = \lambda A_{L,t} k_{L,t}^{\alpha} + (1-\lambda) A_{H,t} k_{H,t}^{\alpha} \qquad \text{(retail good)}$$

$$m_t + r \left(\lambda k_{L,t} + (1-\lambda) k_{H,t} \right) + \Sigma_i \lambda_i \Psi(\Delta\Omega_{i,t+1}) = 0 \qquad \text{(numeraire good)}$$

$$\lambda \delta_{L,t} + (1-\lambda) \delta_{H,t} = 0 \qquad \text{(data)}$$

Social Planner Problem

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

$$\text{s.t.} \quad c_t = \lambda A_{L,t} k_{L,t}^{\alpha} + (1-\lambda) A_{H,t} k_{H,t}^{\alpha} \qquad \text{(retail good)}$$

$$\lambda \delta_{L,t} + (1-\lambda)\delta_{H,t} = 0 \qquad \text{(data)}$$

$$\Omega_{i,t+1} = \left[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left(z_i k_{it}^{\alpha} + \delta_{it}(\mathbf{1}_{\delta_{it}>0} + \iota \mathbf{1}_{\delta_{it}<0})\right)\sigma_\varepsilon^{-2}$$

- equilibrium is efficient

Data as a Business Stealing Technology

- lots of data used for advertising, maybe not quality enhancing?
- data processing helps the firm that uses it, but has no social value
  Morris-Shin (2002)

$$A_{i,t} = \bar{A} - \left(a_{i,t} - \theta_t - \varepsilon_{a,i,t}\right)^2 + \int_{j=0}^{1} \left(a_{j,t} - \theta_{j,t} - \varepsilon_{a,j,t}\right)^2 dj$$

- *keeping of with Joneses*
- unchanged: firm choices, firm dynamics, aggregate quality
- changed: welfare

CONCLUSIONS

- macroeconomics of big data
- knowledge economies have tricky features:

  economic transactions generate data, semi-rivalry, data accumulation and depreciation, increasing and decreasing returns

- flexible tool that captures many features of the data economy:

  endog growth, data platforms, data barter, business stealing, welfare/opt policy

- lots of new directions to explore:

  measurement, data pricing and valuation theory, firms dynamics with entry/exit, imperfect competition