The Big Data Revolution: Data Markets and Finance

Maryam Farboodi

MFR Summer School

August 2022

What is Big data?

I. **two** things
   - large volume of digitized datasets
   - accompanied technological innovation that is necessary to process, analyze and manage them

II. what is data used for in this "big data revolution"?
   - **prediction**

III. data is a new asset class

IV. what should we expect as data availability and technology of data processing improves?
   - markets for data trade
   - investor behavior
   - measurement

Bayes Law
Bread and Butter in Economics of Prediction

- agents want to predict random variable z
- **data:** signals $s^1, s^2, ..., s^n$ about z

  where do signals come from? prior knowledge, information acquisition, production, God sent them, $\cdots$
- more data improves the precision of agents' pesterior belief about the random variable
- **Bayes Law: posterior precision is additive**

$$s^j = z + \epsilon^j \qquad j = 1, \cdots, n \quad \epsilon^j \sim N(0, \Sigma^j)$$
$$\Omega^j = (\Sigma^j)^{-1}$$
$$\Omega^{\text{posterior}} = \sum_{j=1}^{n} \Omega^j$$

- **second order approximation is your best friend!**

Outline

Framework Reminder!

- continuum of firms i
  - use capital to produce goods and data
  - data is a byproduct of transactions
    - firms can be different in their big data technology: z_i
  - data is non-exclusive/non-rival: seller keeps $1 - \iota$ fraction
  - data used for prediction: improve product quality tomorrow $A_{i,t+1}$
  - firms can use the data they produce and/or sell it to other firms

- aggregate output

$$Y_t = f\left(\{A_{i,t} k_{i,t}\}_i\right) = \int_i A_{i,t} k_{i,t}^{\alpha} di$$

$$P_t = \overline{P} Y_t^{-\gamma}$$

Value Function

$$
\begin{aligned}
V(\Omega_{i,t}) = \max_{k_{i,t},\delta_{i,t}} \; & P_t \mathbb{E}_i\left[A_{i,t}(\Omega_{i,t})\right] k_{i,t}\alpha \\
& - \Psi(\Delta\Omega_{i,t+1}) - \pi_t \delta_{i,t} - r k_{i,t} + \frac{V(\Omega_{i,t+1})}{1+r} \\
\Omega_{i,t+1} = & \left[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma\theta^2\right]^{-1} \\
& + \left(z_i k_{i,t}^{\alpha} + \delta_{it}(\mathbf{1}_{\text{data bought}} + \iota \mathbf{1}_{\text{data sold}})\right) \sigma_{\epsilon}^{-2}
\end{aligned}
$$

Inter-firm Data Trade

- π_t: price of data
- what do the firms use the data for?

$$\Omega_{i,t+1} = \text{discounted current data} + \underbrace{\left(z_i k_{i,t}^{\alpha} + \delta_{it}(\mathbf{1}_{\text{data bought}} + \iota \overbrace{\mathbf{1}_{\text{data sold}}}^{\text{data sales}})\sigma_{\epsilon}^{-2} \right)}_{\text{enhance own future quality}}$$

  - green: improve own product quality tomorrow & more profits on good market tomorrow: $P_{t+1} \times \Delta(\mathbb{E}\left[A_{t+1}|\Omega_{i,t+1}\right] k_{i,t+1}^{\alpha})$)
  - red: data sales & profits today: $\pi_t \delta_{i,t}$

Data Market: Homogeneous Big Data Technology

- Farboodi (2022), "Data Markets and Intermediation"
- firm i has produced a unit of data
- price of data = benefit of selling one unit = cost of buying one unit
- **non-exclusivity of data**
  - cost of selling the unit of data $= \iota \times$ benefit of buying the unit of data
  - there is always a price in between that equates the supply and demand of data on the data market

$\Rightarrow$ **no equilibrium where data market is not active**

Open Banking

- bank data sharing regulation
- when data is non-exclusive, data sharing enhances welfare
- designing an efficient interbank data market should lead to voluntary data sharing $\Rightarrow$ **data market design**
- why is it that banks do not want to share their data?
- **entry:** who are the new fintech entrants targeting? who is the policy targeting?

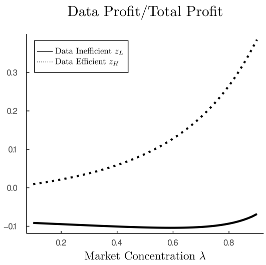
Heterogeneous Big Data Technology Comparative Advantage & Specialization, Concentration

- big data technology: $z_L < z_H$, $\lambda =$ measure of z_L firms

Proposition (Data Efficient Firms Accumulate Less Knowledge)

For sufficiently low γ, α and ι, $\Omega_H < \Omega_L$.

- few efficient data producers
  $\equiv$ high concentration
  $\Rightarrow$ more specialization



Outline

Some Notation!

- random variables to learn about
  - y : firm fundamental, x : market demand
- financial variables
  - R_t : asset return, d_t : firm earnings, g_t : earning growth
- data
  - $\mathcal{I}_{it}$: information set of agent i at time t
  - Ω_{it} : stock of knowledge of agent i at time t (posterior precision)

Standard REE Model

- continuum of investors i
- preferences: $U(\tilde{c}_{it}) = -e^{-\rho\tilde{c}_{it}}$
  - $\rho > 0$: risk aversion
- endowments: e_{it}
- n asset, indexed by j
  - dividend $\tilde{d}_{jt} = H(\mu, d_{jt-1}) + \tilde{y}_{jt}$
  - $\tilde{y}_{jt} \sim N(0, \Sigma_{dj})$. $\sim$ means unknown start of period.
- budget: $\tilde{c}_{it} = r\left(e_{it} - \sum_{j=1}^{n} q_{ijt} p_{jt}\right) + \sum_{j=1}^{n} q_{ijt} \times \text{asset } j \text{ payoff}$
  - $q_{it} = [q_{ijt}]_{j=1}^{n}$: portfolio choice vector
  - $r > 1$: rate of time preference
- stochastic supply
  - $\bar{x}_j + \tilde{x}_{jt}$ $\tilde{x}_{jt} \sim N(0, \Sigma_{xj})$
- market clears: $\int_i q_{it} di - \tilde{x}_t = 1$
- **price:**

$$p_t = A_t + B(d_t - \bar{d}) + C_t y_{t+1} + D_t x_{t+1}$$

Information

- signal about random variable z
  - $\eta_z = \tilde{z} + \tilde{\epsilon}_z \qquad \tilde{\epsilon}_z \sim iid\ N(0, \Omega_z^{-1})$
  - Ω_z: signal precision

- optimal portfolio

$$q_{ijt} = \frac{\mathbb{E}[d_{jt}|\mathcal{I}_{it}] - rp_{jt}}{\rho Var[d_{jt}|\mathcal{I}_{it}]}$$

  more precise signal $\Rightarrow$ more profitable portfolio

- individual optimization with information acquisition

$$\begin{array}{ll} \max_{\Omega_z \geq 0} & E[U(\tilde{c}_{it})|\mathcal{I}_{it}] \\ \text{s.t.} & \text{Information Constraint/Cost} \end{array}$$

Aggregate Trends in Financial Markets

- Farboodi Veldkamp (AER 2020), "Long Run Growth of Financial Data Technology"
- aggregate consequences of technological progress in data analysis in financial markets
- growth theory describes how technology boosts efficiency.
  but in finance, technology (IT) is blamed for volatility, illiquidity and inefficiency SEC ('15), Ben-David et al ('12), Zhang ('06)
- concern: big data is changing not only how much data we see, but also what kinds of data we choose to use
  - big data can predict asset payoffs, or market demand/sentiments

Fundamental versus Demand Data

- dynamic economy
- data processing technology grows exogenously
- investors choose how much to learn, and about what
  - fundamental: $\eta_{fit} = \tilde{y}_t + \tilde{\epsilon}_{fit}$
  - demand: $\eta_{xit} = \tilde{x}_t + \tilde{\epsilon}_{xit}$
- what does one do with demand data?
  - "dumb" order flow: trade against it (market-making?)
  - extract what others know (remove noise from price)
- key insight: demand data expressed as fundamental data: MRT $(C/D)^2$ in precision

$$p_t = A_t + B(d_{t-1} - \mu) + C_t \tilde{y}_t + D_t \tilde{x}_t$$

$$\frac{(p_t - A_t - B(d_{t-1} - \mu) - D_t E[\tilde{x}_t | \mathcal{I}_{it}])}{C_t} = \tilde{y}_t + \underbrace{\frac{D_t}{C_t}(\tilde{x}_t - E[\tilde{x}_t | \mathcal{I}_{it}])}_{\text{demand data } \downarrow\text{es signal noise}}$$

FINDINGS

- different phases of data analysis
  1. first fundamental analysis
  2. followed by demand/sentiment analysis
  3. finally balanced growth
- aggregate price informativeness grows
- market becomes illiquid before reverting and becoming more liquid
- **future information risk:** data double-edged sword in risk resolution

Cross Sectional Trends

- Farboodi, Matray, Veldkamp, Venkateswaran (RFS 2022), "Where Has All the Data Gone?"
- S&P500 price informativeness has improved over time
  but average price informativeness over all public firms has deteriorated!
- large degree of heterogeneity in cross-section of firm

  have all the firms benefited the same from progress in big data technology?
- **structural approach**
- **finding:** divergence in data and informational efficiency of prices
  - most data processing by investors is about *large growth* firms
  - why? investors process that that is most valuable to them
  - size and growth interact to make data more valuable
  - measuring investor data

Spillover from Financial Markets to Firm Distribution

- Begenau, Farboodi, Veldkamp (JME) "Big Data in Finance and Growth of Large Firms "
- small firms are being displaced by larger ones
- **big data technology** benefits growth of large firms disproportionately
  - data comes from economic transactions
  - big firms, with many transactions, produce a lot of data
  - big data technology allows investors to process all of this data $\Rightarrow$ systematically changes how large and small firm capital is priced

key mechanism: data resolves risk $\Rightarrow$ lower risk reduces risk premium $\Rightarrow$ cost of capital falls $\Rightarrow$ firm grows more

Outline

Measuring Investor Data: Structural Approach

- group stocks into four groups j:
  {Small-Growth, Large-Growth, Small-Value, Large-Value}
- informativeness of stock prices

$$\boxed{\text{price informativeness}_t^j = \underbrace{\frac{\Sigma_d^j}{StdDev(p^j)}}_{\text{volatility}} \underbrace{\frac{{g^j}^t}{r - g^j}}_{\text{growth}} \underbrace{\left[1 - \frac{{\Sigma_d^j}^{-1}}{\overline{\Omega}^j}\right]}_{\textit{data}}}$$

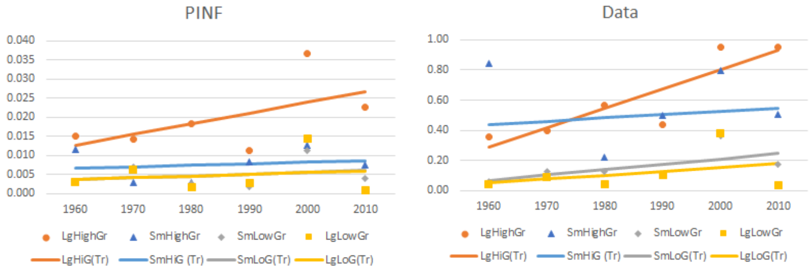
- estimate for each decade (stock f, group j, time t):

$$\frac{EBIT_{fjt+s}}{ASSET_{fjt}} = \alpha_{js} + \beta_{js} log\left(\frac{MKVAL_{fjt}}{ASSET_{fjt}}\right) + \gamma_j X_{fjt} + \epsilon_{fjts}$$

- price informativeness

$$PINF_{js} = \beta_{js} \cdot StdDev\left(\frac{MKVAL_{fjt}}{ASSET_{fjt}}\right)$$

Investor Data



cross-sectional divergence in financial data

Measuring Value of Data to Investors: Statistical Approach

- Farboodi, Singal, Veldkamp, Venkateswaran (2022) "Valuing Financial Data"
- what is an investor's willingness to pay for data?
- demand, not equilibrium transactions price
- why is it hard?
  - individual investor profit from data depends on who else knows that data, who knows similar data, and how aggressively they will trade on it

$\Rightarrow$ statistical approach to bypass the need to know others' information sets and characteristics

front and center: investor heterogeneity: wealth, investment style (mandate), price impact of trades

Statistical Estimation

- general utility function: second order Taylor approximation
- ex-ante expected utility of data for an investor in a GE REE framework

$$U(\mathcal{I}_{it}) = \underbrace{\mathbb{E}\left[R_t\right]' \hat{\mathbb{V}}_i^{-1} \mathbb{E}\left[R_t\right]}_{\text{(Sharp ratio)}^2} + \text{Tr}\left[\underbrace{\left(\mathbb{V}[R_t] - \mathbb{V}[R_t \mid \mathcal{I}_{it}]\right)}_{\text{variance reduction}} \hat{\mathbb{V}}_i^{-1}\right] + r\rho_i \bar{w}_{it}$$

- **Dollar value of data:** investor indifferent between

  having the data $\equiv$ no data$+$ additional riskless wealth

$$\$\text{value of } \text{data}_i = \frac{1}{r\rho_i}\left(\tilde{U}(\mathcal{I}_{it} + \text{data}) - \tilde{U}(\mathcal{I}_{it})\right)$$

Where is the Data?

- individual investor's data
  - adjust the variance in the Sharp ratio
  - variance reduction
- where did everyone else's data go?
  - it did not disappear! it matters through R_{t+1}
  - data others know is in prices $p_t \Rightarrow$ does not forecast returns beyond that
  - conditioning on it will not affect $\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}] \Rightarrow$ it won't increase utility
- note: price impact (Kyle λ) is also in the adjusted variance

Estimation: Procedure

- data to be valued X_t, existing data Z_t

$$R_{t+1} = \beta_1 X_t + \beta_2 Z_t + \varepsilon_t^{XZ}$$
$$R_{t+1} = \gamma_2 Z_t + \varepsilon_t^{Z}$$

- **insight:** for linear Normal variables: Bayes law and OLS coincide
  $\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}]$ is the expected squared residual from OLS regression

- conditional variance without data we're valuing for a sample $1, \ldots, T$

$$\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}] \approx \widehat{\text{Cov}}[\varepsilon_t^Z] = \frac{1}{T - |Z|} \sum_{t=1}^{T} \varepsilon_t^Z \varepsilon_t^{Z\,\prime}$$

- conditional variance with data

$$\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it} + \text{data}] \approx \widehat{\text{Cov}}[\varepsilon_t^{XZ}] = \frac{1}{T - |Z| - |X|} \sum_{t=1}^{T} \varepsilon_t^{XZ} \varepsilon_t^{XZ\,\prime}$$

Different Values for the Same Data

how much are IBES forecasts worth to an investor who only knows some aggregate(concurrent) variables?

	Investment Style				
	Small	Large	Growth	Value	All
Perfect Competition					
Investor with \$500,000 Wealth	0.00	\$1.7k	\$2.5k	\$490	\$3.5k
Investor with \$250m Wealth	0.00	\$566k	\$844k	\$164k	\$1.2m
With Price Impact					
Investor with \$500,000 Wealth	0.00	\$1.6k	\$2.5k	\$410	\$1.4k
Investor with \$250m Wealth	0.00	\$24k	\$57k	\$1.5k	\$253k

- dispersion of valuations for the same data is immense
- data valuations become less heterogeneous with price impact $\rightarrow$ higher price elasticity of data demand

$\Rightarrow$ **demand elasticity:** inelastic asset demand $\Leftrightarrow$ more elastic data demand

Outline

Big Data & Big Data Technology as Traded Products

- market for data is on the rise
- **data intermediaries**
  - data brokers sell consumer data to firms
  - open banking
  - firms buy transaction data statistics from data intermediaries such as Amazon
- market for digital services is also growing
- **digital intermediaries:** large tech firm like Amazon, Google and Microsoft
  - large investment in digital infrastructure
  - rent out cloud storage and computing to other firms
  - build an ecosystem
- fintech industry is changing how the financial market functions

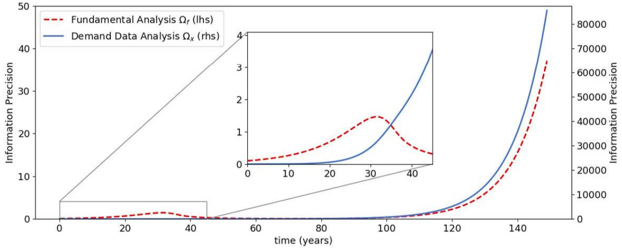
Concluding Remarks. The Big Data Research Agenda

- numerous shifts in the financial and real sector are a logical consequence of emergence of big data
- data is changing how firms operate: "Data Is the New Oil"
- data measurement is far from obvious

Big Data is transforming markets. We need theory and measurement to make sense of a constantly evolving landscape!

Growth of Financial Data Processing

- phases of data analysis



- price informativeness and liquidity