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, \ldots

- more data improves the precision of agents’ posterior belief about the random variable

- **Bayes Law**: posterior precision is additive

$$s^j = z + e^j \quad j = 1, \ldots, n \quad e^j \sim N(0, \Sigma^j)$$

$$\Omega^j = (\Sigma^j)^{-1}$$

$$\Omega_{\text{posterior}} = \sum_{j=1}^{n} \Omega^j$$

- second order approximation is your best friend!
Outline

1 Data Markets
2 Investor Behavior
3 Measurement
4 Concluding Remarks
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

\[ Y_t = f(\{A_{i,t} k_{i,t}\}_i) = \int_i A_{i,t} k_{i,t}^{\alpha} di \]

\[ P_t = \bar{P} Y_t^{-\gamma} \]
**Value Function**

\[
V(\Omega_{i,t}) = \max_{k_{i,t},\delta_{i,t}} P_t E_i \left[ A_{i,t}(\Omega_{i,t}) \right] k_{i,t} \alpha \\
- \Psi(\Delta \Omega_{i,t+1}) - \pi_t \delta_{i,t} - rk_{i,t} + \frac{V(\Omega_{i,t+1})}{1 + r} \\
\Omega_{i,t+1} = \left[ \rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma \theta^2 \right]^{-1} \\
+ \left( z_i k_{i,t}^\alpha + \delta_{it} (1_{\text{data bought}} + \nu 1_{\text{data sold}}) \right) \sigma_\epsilon^{-2}
\]
Inter-firm Data Trade

- $\pi_t$: price of data
- what do the firms use the data for?

$$\Omega_{i,t+1} = \text{discounted current data} + \left( z_i k_{i,t}^{\alpha} + \delta_{i,t} \left( \mathbf{1}_{\text{data bought}} + i \mathbf{1}_{\text{data sold}} \right) \sigma^{-2} \right)$$

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

enhance own future quality

data sales
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 = $i \times$ 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 $\gamma$, $\alpha$ and $\iota$, $\Omega_H < \Omega_L$.

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

![Graph showing Data Profit/Total Profit vs Market Concentration $\lambda$](image-url)
Outline

1. Data Markets
2. Investor Behavior
3. Measurement
4. Concluding Remarks
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
  - $I_{it}$: information set of agent $i$ at time $t$
  - $\Omega_{it}$: 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
  - $\tilde{x}_j + \tilde{x}_{jt}$
  - $\tilde{x}_{jt} \sim N(0, \Sigma_{xj})$
- 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 $z$
  - $\eta_z = \tilde{z} + \tilde{\epsilon}_z$
  - $\tilde{\epsilon}_z \sim iid \mathcal{N}(0, \Omega_z^{-1})$
  - $\Omega_z$: signal precision

- optimal portfolio
  
  $$q_{ijt} = \frac{\mathbb{E}[d_{jt}|I_{it}] - r_{jt}}{\rho \text{Var}[d_{jt}|I_{it}]}$$

  more precise signal $\Rightarrow$ more profitable portfolio

- individual optimization with information acquisition
  
  $$\max_{\Omega_z \geq 0} \ E[U(\tilde{c}_{it})|I_{it}]$$
  s.t. Information Constraint/Cost
Aggregate Trends in Financial Markets

- 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: \( \eta_{\text{fit}} = \tilde{y}_t + \tilde{\epsilon}_{\text{fit}} \)
  - demand: \( \eta_{\text{xit}} = \tilde{x}_t + \tilde{\epsilon}_{\text{xit}} \)

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

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

\[
\frac{p_t - A_t - B(d_{t-1} - \mu) - D_tE[\tilde{x}_t|I_{it}]}{C_t} = \tilde{y}_t + \frac{D_t}{C_t}(\tilde{x}_t - E[\tilde{x}_t|I_{it}])
\]

Demand data↓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
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 ⇒ systematically changes how large and small firm capital is priced

**Key mechanism:** Data resolves risk ⇒ lower risk reduces risk premium ⇒ cost of capital falls ⇒ firm grows more
Outline

1. Data Markets
2. Investor Behavior
3. Measurement
4. Concluding Remarks
Measuring Investor Data: Structural Approach

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

\[
\text{price informativeness}^{j}_{t} = \frac{\sum_{d}^{j} g^{j}_{t}}{\text{StdDev}(p^{j})} \left[ 1 - \frac{\sum_{d}^{j} - 1}{\Omega^{j}} \right]
\]

- 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

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

⇒ 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

\[ U(\mathcal{I}_{it}) = \mathbb{E} [R_t]' \hat{\mathbb{V}}^{-1}_i \mathbb{E} [R_t] + \text{Tr} \left[ (\mathbb{V}[R_t] - \mathbb{V}[R_t | \mathcal{I}_{it}]) \hat{\mathbb{V}}^{-1}_i \right] + r \rho_i \tilde{w}_{it} \]

- Dollar value of data: investor indifferent between having the data \( \equiv \) no data + additional riskless wealth

\[ \text{value of data}_i = \frac{1}{r \rho_i} (\tilde{U}(\mathcal{I}_{it} + \text{data}) - \tilde{U}(\mathcal{I}_{it})) \]
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 $\mathbb{V}[R_{t+1} | I_{it}] \Rightarrow$ it won’t increase utility

- note: price impact (Kyle $\lambda$) 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

  $V[R_{t+1} \mid \mathcal{I}_{it}]$ is the expected squared residual from OLS regression

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

  \[
  V[R_{t+1} \mid \mathcal{I}_{it}] \approx \widehat{\text{Cov}}[\varepsilon_{t}^{Z}] = \frac{1}{T - |Z|} \sum_{t=1}^{T} \varepsilon_{t}^{Z} \varepsilon_{t}^{Z'}
  \]

- conditional variance with data

  \[
  V[R_{t+1} \mid \mathcal{I}_{it} + \text{data}] \approx \widehat{\text{Cov}}[\varepsilon_{t}^{XZ}] = \frac{1}{T - |Z| - |X|} \sum_{t=1}^{T} \varepsilon_{t}^{XZ} \varepsilon_{t}^{XZ'}
  \]
Different Values for the Same Data

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

<table>
<thead>
<tr>
<th>Investment Style</th>
<th>Small</th>
<th>Large</th>
<th>Growth</th>
<th>Value</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perfect Competition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investor with $500,000 Wealth</td>
<td>0.00</td>
<td>$1.7k</td>
<td>$2.5k</td>
<td>$490</td>
<td>$3.5k</td>
</tr>
<tr>
<td>Investor with $250m Wealth</td>
<td>0.00</td>
<td>$566k</td>
<td>$844k</td>
<td>$164k</td>
<td>$1.2m</td>
</tr>
<tr>
<td><strong>With Price Impact</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investor with $500,000 Wealth</td>
<td>0.00</td>
<td>$1.6k</td>
<td>$2.5k</td>
<td>$410</td>
<td>$1.4k</td>
</tr>
<tr>
<td>Investor with $250m Wealth</td>
<td>0.00</td>
<td>$24k</td>
<td>$57k</td>
<td>$1.5k</td>
<td>$253k</td>
</tr>
</tbody>
</table>

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

⇒ **demand elasticity**: inelastic asset demand ⇔ more elastic data demand
Outline

1. Data Markets
2. Investor Behavior
3. Measurement
4. Concluding Remarks
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