Big Data in Finance

March 22, 2019

Mao Ye
University of Illinois, Urbana-Champaign and NBER

Sentiment
Completeness
Financial
Causality
Order Book
Ecommerce Transaction
Data Flows
Processors
Integration
Banking
Mortgage
News
Retail
Interpretable
Industrial
Clustering

Petabytes
Large
Information
Analytics
Science
Technology

Consumer Discretionary
Large
Picture
Trends
Video

Parallel
Variety
Web Searches
Volume
Storage
Velocity
Unstructured

MapReduce
Debit Card
Accounting Data

Accounting Data

Integration

Debit Card

Volume

Ecommerce Transaction

Processors

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration

Debit Card

Accounting Data

Integration
Three Aspects of Big Data

• Large size

• High dimension
  – A large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format
  – Satellite images, social media, and credit card transactions
Roadmap

• Large size

• High dimension
  – Large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format
Small vs. Large Data

- Smaller datasets often involve selection processes from larger datasets
  - Smaller sample size
  - Fewer variables
  - Aggregations of economic activity
  - Snapshot of economic activity

- Are there sample selection biases in smaller datasets?
Size of Trade and Quote Data (TAQ)

- NYSE, NASDAQ, and regional exchange listed securities
- All trades and quotes reported to the consolidated tape
## Larger Data: Order Level Data

<table>
<thead>
<tr>
<th>Type</th>
<th>Timestamp (nanoseconds)</th>
<th>Order Reference Number</th>
<th>Buy/Sell</th>
<th>Shares</th>
<th>Stock</th>
<th>Price</th>
<th>Original Order Reference Number</th>
<th>Market Participant ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>53435.759668667</td>
<td>335531633</td>
<td>S</td>
<td>300</td>
<td>EWA</td>
<td>19.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>40607.031257842</td>
<td>168914198</td>
<td>B</td>
<td>100</td>
<td>NOK</td>
<td>9.38</td>
<td></td>
<td>UBSS</td>
</tr>
<tr>
<td>U</td>
<td>53520.367102587</td>
<td>336529765</td>
<td></td>
<td>300</td>
<td></td>
<td>19.45</td>
<td>335531633</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>53676.740300677</td>
<td>336529765</td>
<td></td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>57603.003717685</td>
<td>625843333</td>
<td></td>
<td>100</td>
<td></td>
<td>32.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>53676.638521222</td>
<td>336529765</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>53676.740851701</td>
<td>336529765</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Add order anonymously</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Add order with market participant ID</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Update: replace old order with a new order</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Order execution</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Order executed with price message</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Partial cancellation</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Order deletion</td>
<td></td>
</tr>
</tbody>
</table>
Research Question

• Are there selection biases in TAQ data?

• Method: Compare TAQ data with order level data
  – A large dataset and a larger dataset

• Solution: high performance computing
Selection Bias Led by Regulations

• Previous regulations: No need to report trades less than 100 shares (odd lots)
  – Rationale: Odd lots are from small retail traders

• Consequence: Odd lots are missing from TAQ data

• O’Hara, Yao, and Ye (2014) find:
  – 25% of trades are unreported in 2011
  – More trades are missing for high-priced stocks
    • Google: 53% of trades, 23% of volume
    • Apple: 38% of trades, 14% of volume
Are Odd Lots from Retail Traders?

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Symbol</th>
<th>Hour</th>
<th>Minute</th>
<th>Second</th>
<th>Millisecond</th>
<th>Shares</th>
<th>Buy/Sell</th>
<th>Price</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>20</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>2</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>10</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>109</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>110</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>30</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>111</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>3</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>112</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>47</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>113</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>80</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>114</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>80</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>210</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>5</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>211</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>25</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>212</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>213</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>12</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
</tbody>
</table>
Machines Challenge Regulations

• Computers can reduce large orders to small odd lots
  – Benefit: Hide information
  – Odd lots are more informed than trades greater than or equal to 100 shares

• Policy impact: Regulators reduce report threshold from 100 shares to 1 share
Size Challenges

Techniques
• High performance computing helps to overcome size challenges

Economic insights
• Open question for policy
  – Many regulations were designed for humans
  – Should regulations be revised for machines?

• Are there selection biases in other “small” datasets?
  – Can larger datasets lead to different results?
Roadmap

• Large size

• High dimension
  – Large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format
Firms that use machine-learning techniques to make investment decisions, such as Renaissance Technologies and Two Sigma Investments, operate at timescales ranging “anywhere from a few minutes to a few months.”


Chinco, Clark-Joseph, and Ye (2017)

– Examine this question at minute-by-minute horizon
• Basic idea: Use lagged stock returns to forecast $r_{n,t+1}$

• Data: One-minute returns of other ($\approx 2,000$) NYSE-listed stocks

• OLS requires at least 2,000 observations (six trading days)
  • Too many RHS variables for OLS
  • Hard-to-capture signals that are unexpected and short-lived

• We use machine learning techniques to reduce dimensions

Forecast-Implied Performance Net of Trading Costs

Annualized Sharpe Ratios

<table>
<thead>
<tr>
<th></th>
<th>S&amp;P 500</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.123</td>
<td>1.791</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LASSO-Implied Strategy Abnormal Returns [%/yr]</th>
<th>α</th>
<th>Mkt</th>
<th>HmL</th>
<th>SmB</th>
<th>Mom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>2.709 (0.034)</td>
<td>0.004 (0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Factor Model</td>
<td>2.713 (0.034)</td>
<td>0.004 (0.002)</td>
<td>-0.004 (0.004)</td>
<td>0.000 (0.003)</td>
<td></td>
</tr>
<tr>
<td>4-Factor Model</td>
<td>2.707 (0.034)</td>
<td>0.005 (0.002)</td>
<td>-0.004 (0.004)</td>
<td>0.003 (0.004)</td>
<td>0.003 (0.004)</td>
</tr>
</tbody>
</table>
• LASSO is more likely to pick a stock as a predictor before its news announcements
  – Even if we use the millisecond news feeds like RavenPack

• Big data incorporate information faster than news announcements
  – A story

• Writing news articles takes time, especially for unscheduled events
  – The difference between public information and news

• Empirical evidence
  – LASSO is more likely to pick a stock as a predictor before unscheduled news
High Dimensional Challenges

• Techniques
  – Machine learning techniques deal with high dimensional data

• Economic insights
  – Determining economic interpretations is a higher hurdle
Roadmap

• Large size

• High dimension
  – A large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format

• Big data motivate new economic theories
Example: Twitter Data
Two Challenges

- Techniques: How to extract information from unstructured data?
  - One solution: Find a data vendor
    - J.P. Morgan’s *Big Data and AI Strategies* (2017) provides a list of 500 alternative data vendors
    - Many vendors transfer unstructured data to structured data.
  - Another solution: interdisciplinary collaboration

- Economics: Do unstructured data generate unique measures of economic activity?
  - More challenging

- Example: Da, Nitesh, Xu, and Ye (2017)
Unique Measures from Big Data

- Information diffusion
  - Word-of-mouth communication: No direct measure without big data

- Two traditional solutions
  - Proxies: Physical proximity (Hong, Kubik, and Stein, 2005; Ivkovich and Weisbenner, 2007; Brown et al., 2008) and common schooling (Cohen, Frazzini, and Malloy, 2008)
  - Criminal investigations (Rantala, 2015; Ahern, 2016)

- Big data solution
  - Measure information diffusion using tweets and retweets
Information Diffusion through Retweets

Zhi has 10,000 followers

@Zhi: Twitter data are unstructured ...

Nitesh has 100,000 followers

@Nitesh @Zhi: Twitter ...

Jian has 5,000 followers

@Jian @Nitesh @Zhi: Twitter …
Speed of Information Diffusion

Number of Accounts Reached

- p5
- p50
- p95

Minutes
• Social media can spread stale news
  – When someone retweets news, it is already stale
    • Stale: Ten minutes after the initial release from a news outlet
  – Retail traders still respond
    • Create temporal price pressures
    • Prices first overshoot then revert to the next day

• Smart traders should trade against stale news
  – Profit opportunity: Sell after stale good news and quickly buy back
Machines vs. Humans?

• Reversion speed in our sample period (2013–2014) is much faster than reported in Tetlock (2011)

• Open question: Are smart traders machines?

• Broader questions
  – Do machines trade against human behavioral biases?
  – Are markets more efficient due to the rise of machines?
Structure Challenges

• Techniques
  – Find an alternative data vendor
  – Work with experts in other fields

• Economic insights
  – Unstructured data create unique measures of economic activity
The Future: Understanding Financial Market Ecosystem

Underexplored Territories

Microseconds  3 Months
High-frequency Traders  13F Data

- Paucity of studies on traders who are slower than HFTs but faster than a quarter
  - Execution algorithms who operate at timescales of milliseconds or seconds
  - Traders who use machine-learning techniques operate at timescales of “anywhere from a few minutes to a few months.”
- Half machine, half human
Terminators?
Conclusion: Big Data Challenges and Opportunities

Techniques

• High-performance computing mitigates the size challenges
• Machine learning alleviates the high dimensional challenges
• Alternative data vendors or interdisciplinary collaborations mitigate the structure challenges

Big data opportunities

• Reduce sample selection bias
• Machine learning: foundation for “algorithmic behavioral finance”?
  – Psychology: foundation of behavioral finance
• Create unique measures to test theories