Information
Large Analytics
Research
Science
Picture
Trends Video
Consumer Discretionary
Technology
Petabytes

Big Data in Finance

Parallel Variety
Sentiment
Web Searches
Completeness
Volume
Financial
Velocity
Causality

March 22, 2019

Storage
Unstructured
Order Book

Mao Ye
University of Illinois, Urbana-Champaign and NBER

MapReduce
Ecommerce Transaction
Data Flows
Debit Card
Accounting Data
Processors
Integration
Interpretable
Banking
Industrial
Clustering
Mortgage
News
Retail

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

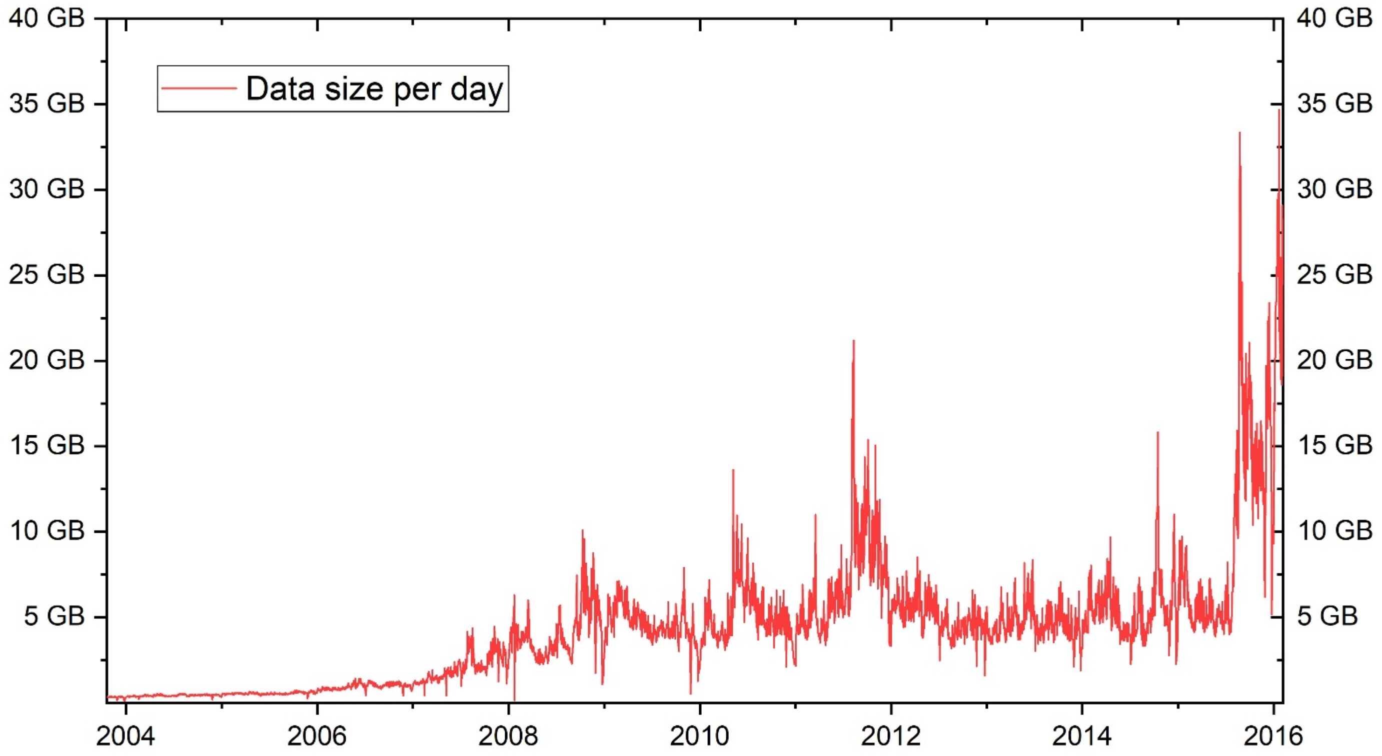
- Large size
- High dimension
  - Large number of variables relative to the sample size
- Complex structure
  - Not in traditional row-column format

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

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

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?

Sequence	Symbol	Hour	Minute	Second	Millisecond	Shares	Buy/Sell	Price	Type
1	AAPL	13	59	1	107	20	S	125.00	HN
2	AAPL	13	59	1	107	10	S	125.00	HN
………									
108	AAPL	13	59	1	107	50	S	125.00	HN
109	AAPL	13	59	1	107	50	S	125.00	HN
110	AAPL	13	59	1	107	30	S	125.00	HN
111	AAPL	13	59	1	107	3	S	125.00	HN
112	AAPL	13	59	1	110	47	S	125.00	HN
113	AAPL	13	59	1	110	80	S	125.00	HN
114	AAPL	13	59	1	110	80	S	125.00	HN
……									
210	AAPL	13	59	1	110	5	S	125.00	HN
211	AAPL	13	59	1	110	25	S	125.00	HN
212	AAPL	13	59	1	110	50	S	125.00	HN
213	AAPL	13	59	1	110	12	S	125.00	HN

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

Does Machine Learning Capture Any Economic Signal?

- 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.”
  - *The Wall Street Journal* (May 21, 2017)

- Chinco, Clark-Joseph, and Ye (2017)
  - Examine this question at minute-by-minute horizon

High Dimensional Challenges

- Basic idea**:** Use lagged stock returns to forecast $r_{n,t+1}$
- Data**:** One-minute returns of other (≈ 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

LASSO-Implied Trading Strategy: 2005-2012

Forecast-Implied Performance Net of Trading Costs

Annualized Sharpe Ratios

S&P 500	LASSO
0.123	1.791

LASSO-Implied Strategy Abnormal Returns [%/yr]	α	Mkt	HmL	SmB	Mom
Market	2.709 (0.034)	0.004 (0.002)			
3-Factor Model	2.713 (0.034)	0.004 (0.002)	−0.004 (0.004)	0.000 (0.003)	
4-Factor Model	2.707 (0.034)	0.005 (0.002)	−0.004 (0.004)	0.003 (0.004)	0.003 (0.004)

Economic Foundation

- 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


```
twitter_public_stream.20140128-220104.json:{"created_at":"Wed Jan 29 21:14:11 +0000
2014","id":428637220338425856,"id_str":"428637220338425856","text":"Facebook earnings: Q4 EPS $0.31 ex-items
v. $0.27 estimate; revenues $2.59 billion v. $2.33 billion estimate - @CNBC http:\/\/t.co\/
sNqDbtfyzv","source":"\u003ca href=\"http:\/\/www.breakingnews.com\" rel=\"nofollow\"\u003ebreakingnews.
com\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_
to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":6017542,"id_str":"
6017542","name":"Breaking News","screen_name":"BreakingNews","location":"Global","url":"http:\/\/www.
breakingnews.com\/about\/mobile","description":"Introducing our new iOS app and http:\/\/BreakingNews.com
that lets you control the breaking news you want to
see.","protected":false,"followers_count":6483805,"friends_count":475,"listed_count":85853,"created_at":"Sun
May 13 23:06:45 +0000 2007","favourites_count":51,"utc_offset":-18000,"time_zone":"Eastern Time (US & Canada)"
,"geo_enabled":false,"verified":true,"statuses_count":82721,"lang":"en","contributors_enabled":false,"is_trans
lator":false,"is_translation_enabled":true,"profile_background_color":"EEEEEE","profile_background_image_url":
"http:\/\/a0.twimg.com\/profile_background_images\/661943965\/2eu2ntwqt6ereyyumm38.
png","profile_background_image_url_https":"https:\/\/si0.twimg.com\/
profile_background_images\/661943965\/2eu2ntwqt6ereyyumm38.
png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/
profile_images\/378800000700003994\/53d967d27656bd5941e7e1fcddf47e0b_normal.
png","profile_image_url_https":"https:\/\/pbs.twimg.com\/
profile_images\/378800000700003994\/53d967d27656bd5941e7e1fcddf47e0b_normal.
png","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/6017542\/1383589267","profile_link_color"
:"CC0000","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"F3F3F3","profile_text_color":"
333333","profile_use_background_image":true,"default_profile":false,"default_profile_image":false,"following":
null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors
":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"symbols":[],"urls":[{"url":"http:\/\/t.
co\/sNqDbtfyzv","expanded_url":"http:\/\/bit.ly\/1nlzmNA","display_url":"bit.ly\/1nlzmNA","indices":[117,139]}
],"user_mentions":[{"screen_name":"CNBC","name":"CNBC","id":20402945,"id_str":"20402945","indices":[111,116]}]
},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"medium","lang":"en"}
```

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 ...

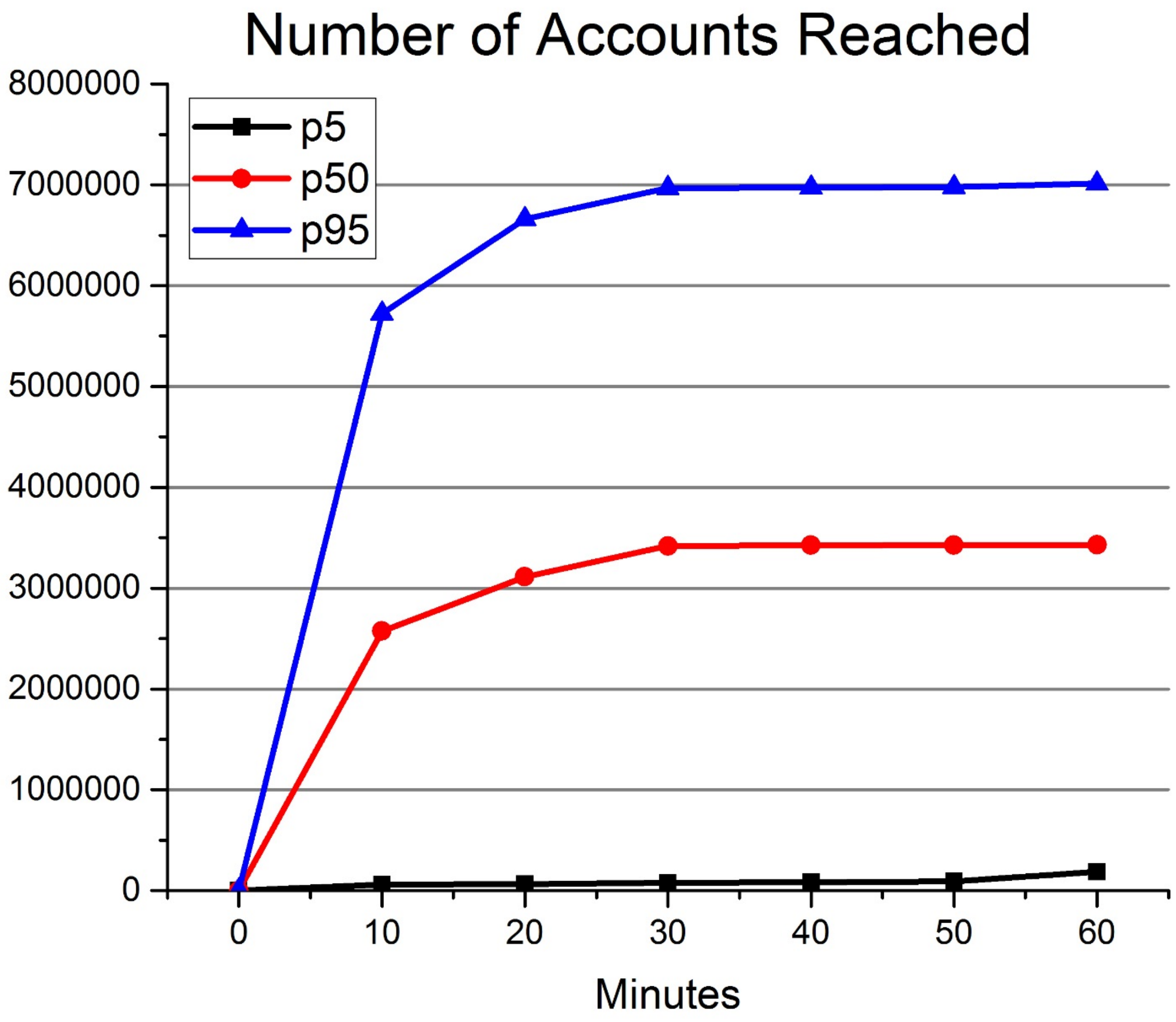
Nitesh has 100,000 followers

@Nitesh @Zhi: Twitter …

Jian has 5,000 followers

@Jian @Nitesh @Zhi: Twitter …

Speed of Information Diffusion



Da, Nitesh, Xu, and Ye (2017)

- 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)
  - Tetlock (2011) sample period: 1996–2008

- 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

High-frequency Traders

3 Months

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
- Alterative 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