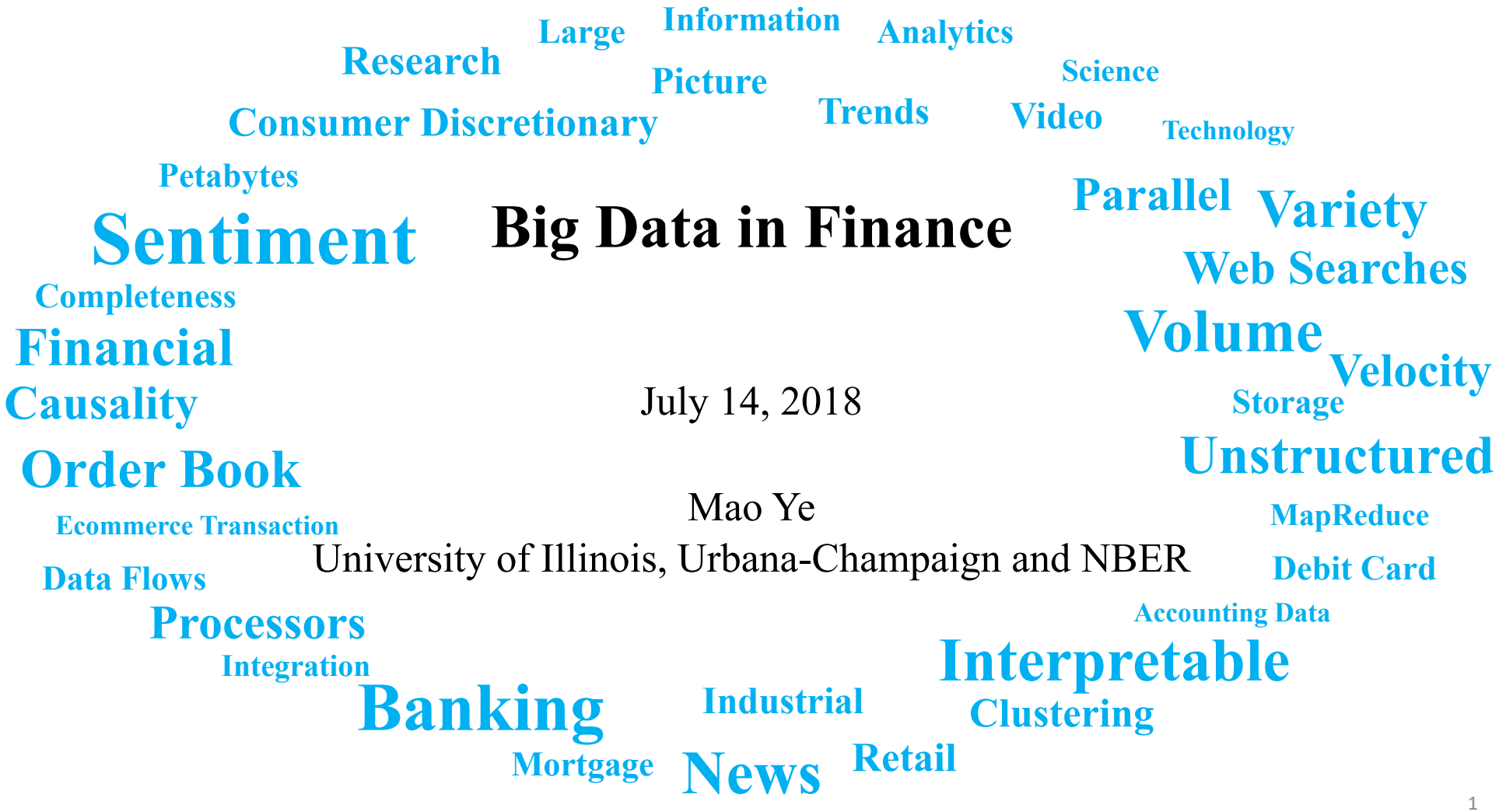

Large
Information
Analytics
Research
Picture
Science
Consumer Discretionary
Trends
Video
Technology
Petabytes
Parallel
Variety
Sentiment
Big Data in Finance
Web Searches
Completeness
Financial
Volume
Velocity
Causality
July 14, 2018
Storage
Order Book
Unstructured
Mao Ye
Ecommerce Transaction
MapReduce
University of Illinois, Urbana-Champaign and NBER
Data Flows
Debit Card
Processors
Accounting Data
Interpretable
Integration
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
  - A large number of variables relative to the sample size
- Complex structure
  - Not in traditional row-column format
- Big data motivate new economic theories

Roadmap

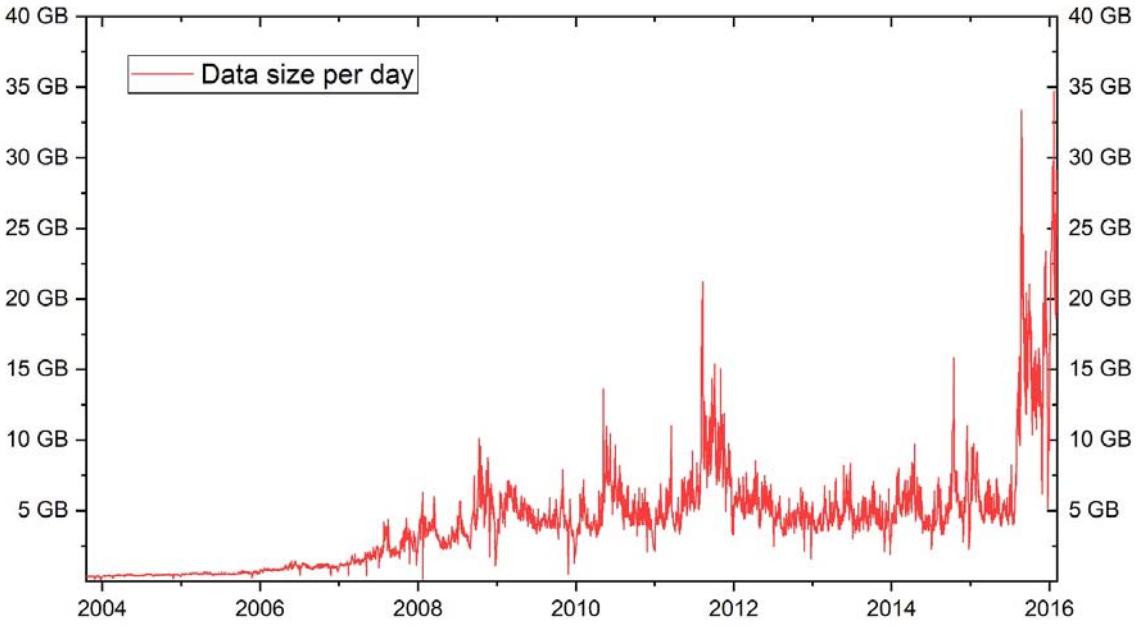
- Large size
- High dimension
  - Large number of variables relative to the sample size
- Complex structure
  - Not in traditional row-column format
- Big data motivate new economic theories

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

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

- Extreme Science and Engineering Discovery Environment (XSEDE)
- First parallel: Day by day
  - Reduce data size to less than 100 gigabytes per day
- Second parallel: Among stocks
  - Each daily file contains 7,000 stocks
    - Some stocks, such as AAPL, are more actively traded than others
    - Divide stock files into equally-sized bundles

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

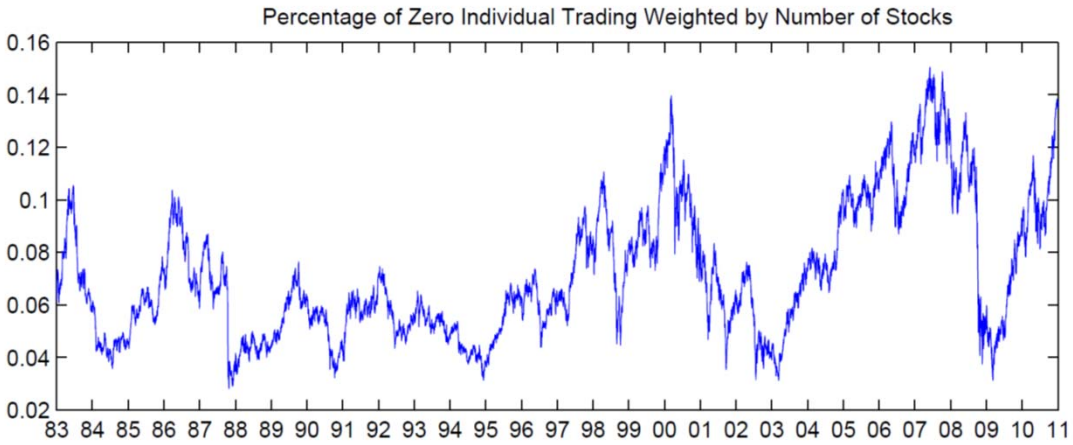
- Computers can reduce large orders to small odd lots
  - Benefit: Hide information
  - Odd lots are more informed than trades greater than or equal to 100 shares
- Policy impact: Regulators reduce report threshold from 100 shares to 1 share

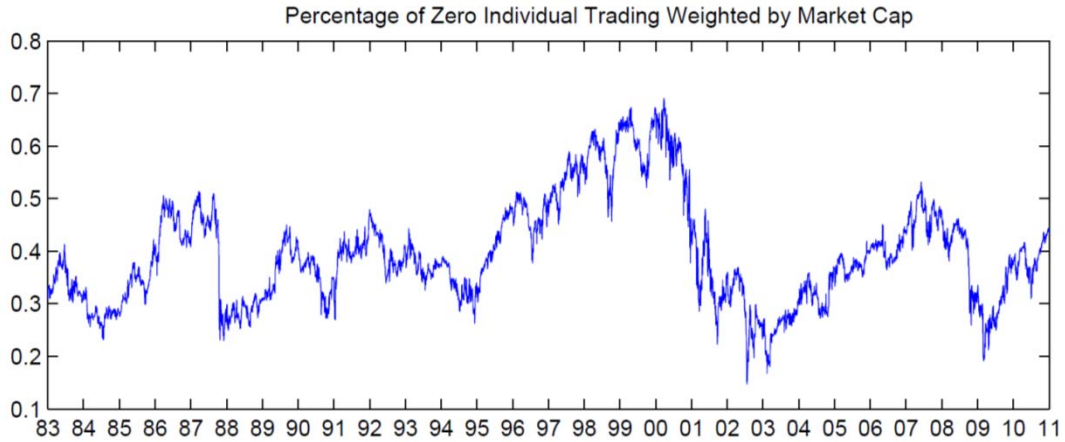
100 Share Cutoff + $5,000 Cutoff

- Lee and Radhakrishna (2000) method
  - Use trades less than $5,000 in TAQ data as a proxy for individual trades

- We find $5,000 cutoff leads to data truncation based on stock price
  - Zero individual trading for any stock with a price higher than $50
  - Truncation does not depend directly on the market share of odd lots
  - O'Hara, Yao, and Ye (2014) estimate the magnitude of truncation based on CRSP

Individual Trading (False Zeroes)





- Up to 15% percent of stocks are truncated
- Those stocks represent up to 70% market cap
- Generate mechanical patterns of retail trading
  - Retail traders trade less in dot-com bubble period

Size Challenges

Techniques

- XSEDE 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
- Big data motivate new economic theories

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 overcome this dimensional challenge

Differences in Approaches

Traditional Approaches

Step 1: Use economic reasoning to **select** x.

Step 2: Use statistical approach to **estimate** x's quality

- Sorting
- Linear regression

Machine Learning Techniques

- Use statistical approach to **select** and **estimate** x
- Handle a large number of x variables
- More flexible functional form

Machine Learning Techniques

- Two common features
  - Cross-validation: Methods are evaluated using out-of-sample predictions
    - Less emphasis on causal inference
    - Belloni, Chernozhukov, and Hansen (2014); Athey and Imbens (2017); Mullainathan and Spies (2017)
  - Regularization: Penalty for complex models to avoid overfitting

- Two variations
  - Functional form: Linear, regression trees, or neural networks
  - The type of regularization

- Example: Chinco, Clark-Joseph, and Ye (2017) use LASSO

Functional Form

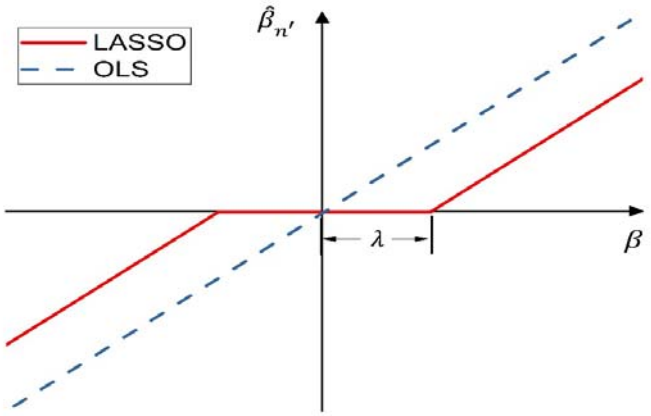
OLS

$$\min_{\beta}\left\{\frac{1}{2\cdot T}\cdot\sum_{t=1}^{T}\left(r_{n,t}-\beta_0-\sum_{n'=1}^{N}\beta_{n'}\cdot r_{n',t-1}\right)^2\right\}$$

Regularization

LASSO is OLS with a penalty function:

$$\min_{\beta}\left\{\frac{1}{2\cdot T}\cdot\sum_{t=1}^{T}\left(r_{n,t}-\beta_0-\sum_{n'=1}^{N}\beta_{n'}\cdot r_{n',t-1}\right)^2+\lambda\sum_{n'=1}^{N}|\beta_{n'}|\right\}$$

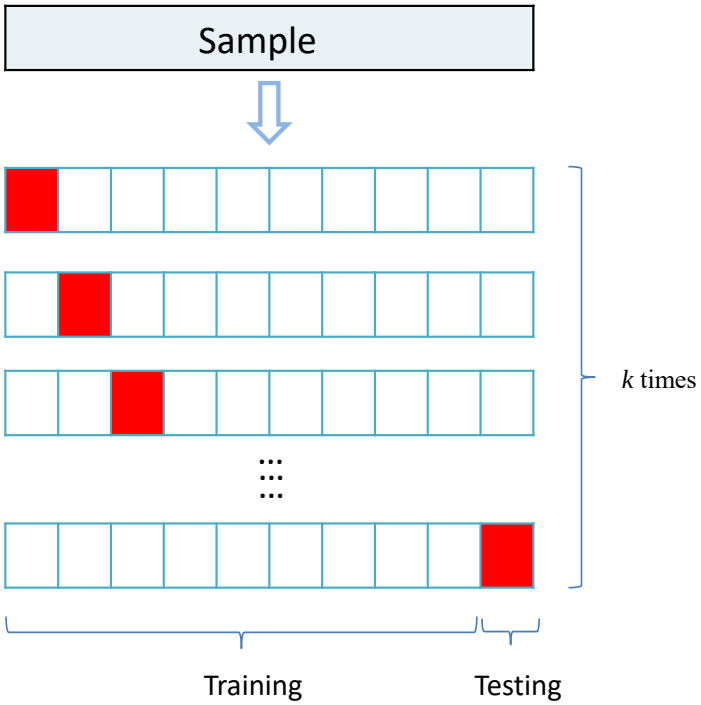


- Variable selection
  - Variables are normalized before regression
  - Small OLS coefficients are set to 0

Cross Validation

We search the best penalty parameter λ through k-fold cross validation

- We choose $k = 10$



1. For every $k = 1,\ldots, 10$, use the k^{th} portion as the testing sample, the rest as training sample

2. Use training sample to calculate the LASSO estimator

3. Use testing sample to calculate the mean squared error:

$$Q(k,\lambda) = \sum_{t=1}^{T}\left(r_{n,t} - \beta_0(k,\lambda) - \sum_{n'=1}^{N}\widehat{\beta_{n'}}(k,\lambda)\cdot r_{n',t-1}\right)^2$$

4. Repeat steps 2–3 ten times to get the average:

$$\bar{Q}(\lambda) = \frac{1}{10}\sum_{k=1}^{10} Q(k,\lambda)$$

5. Pick the λ with the best overall performance:

$$\hat{\lambda}_{min} = \arg\min_{\lambda} \bar{Q}(\lambda)$$

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)

Choice of Predictors

- Unexpected
  - LASSO ignores well-known weekly or monthly predictors
  - Reason: LASSO typically ignores predictors weaker than 2.5% per month
- Short-lived: 95% of LASSO predictors last less than 14.2 minutes
- Sparse: LASSO uses only 12.7 predictors on average
- 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

Machine Learning vs. News

- Big data incorporate information faster than news announcements
- 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
  - LASSO is more likely to pick a stock as a predictor in the same minute as scheduled news

Three Open Questions

- Other horizons
  - Feng, Giglio, and Xiu (2018); Freyberger, Neuhier, and Weber (2018); Han, He, Rapach, and Zhou (2018)
- Other regularizations
  - Kozak, Nagel, and Santosh (2018): Ridge regression
  - $\min_{\beta}\left\{\frac{1}{2\cdot T}\cdot\sum_{t=1}^{T}\left(r_{n,t}-\beta_0-\sum_{n'=1}^{N}\beta_{n'}\cdot r_{n',t-1}\right)^2+\lambda\sum_{n'=1}^{N}\beta_{n'}^2\right\}$
- Other functional forms
  - Capture important nonlinearities and interactions
    - Gu, Kelly, and Xiu (2018)
  - Titanic example (Varian, 2014)
    - Logistic regression predicts "age barely matters for survival rate"
    - Regression tree predicts survival rate is very high for passengers less than 8.5 years old

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

Types of Unstructured Data (Kolanovic and Rajesh, 2017)

- Generated by individuals
  - Social media posts, news, product reviews, web search records, mobile apps, personal pictures/videos/audios
- Generated by business transactions and government filings
  - Supermarket scanner data, SEC filings
- Generated by sensors
  - Satellite images, weather and pollution sensors

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
    - Many vendors transfer unstructured data to structured data (e.g., RavenPack)
    - A comprehensive list of 500 alternative data vendors
    - J.P. Morgan's Big Data and AI Strategies (2017)
  - Another solution: Interdisciplinary collaboration
- Economics: Do unstructured data generate unique measures of economic activity?
  - More challenging
- Example: Da, Nitesh, Xu, and Ye (2017)


AccountingData
EcommerceTransaction
Orderbook
SocialMedia
Geolocation
WebSearches
Maritime
Blogs
Video
RealEstate
Sentiment
Flow
HealthCare
DebitCard
Trends
Metal&Mining
Financials
TextParsing
Reviews
Satellites
CreditCard
Picture
Agriculture
Industrial
Energy
Utilities
News
China/Japan
Telecommunication
Material
CompanyParking
Macroeconomic
Mortgage

Two Challenges

- Techniques: How to extract information from unstructured data?
  - One solution: Find a data vendor
    - Many vendors transfer unstructured data to structured data. E.g. Ravenpack
    - A comprehensive list of 500 alternative data vendors
    - J.P. Morgan's Big Data and AI Strategies (2017)
  - 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 …

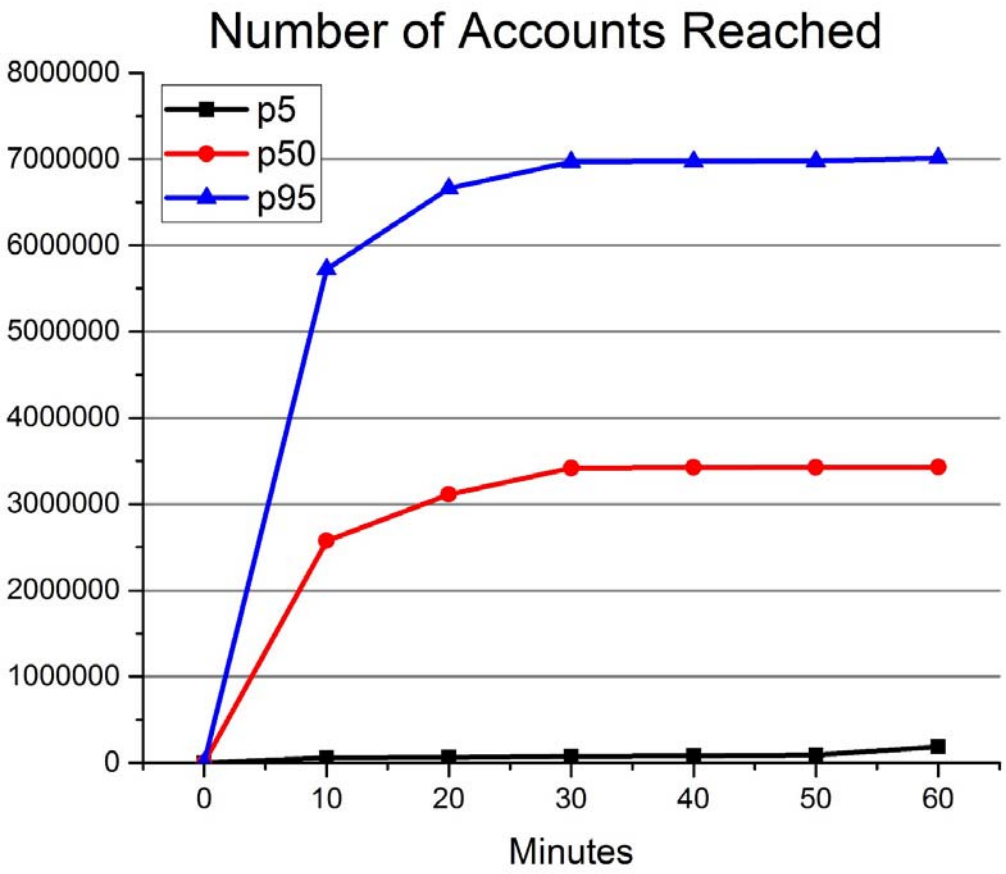
Useful Fields in the Original Tweet


```
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"}
```

One Retweet


```
twitter_public_stream.20140128-220104.json:{"created_at":"Wed Jan 29 21:14:33 +0000
2014","id":428637311690366976,"id_str":"428637311690366976","text":"RT @BreakingNews: Facebook earnings: Q4
EPS $0.31 ex-items v. $0.27 estimate; revenues $2.59 billion v. $2.33 billion estimate - @CNBC
http\u2026","source":"\u003ca href=\"http:\/\/blackberry.com\/twitter\" rel=\"nofollow\"\u003eTwitter for
BlackBerry\u00ae\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":nu
ll,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":99439
5594,"id_str":"994395594","name":"Xee' Sovereign","screen_name":"04Frequency","location":"Nigeria","url":null,
"description":"Student","protected":false,"followers_count":26,"friends_count":124,"listed_count":0,"created_a
t":"Fri Dec 07 05:18:25 +0000 2012","favourites_count":4,"utc_offset":null,"time_zone":null,"geo_enabled":fals
e,"verified":false,"statuses_count":1201,"lang":"en","contributors_enabled":false,"is_translator":false,"is_tr
anslation_enabled":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.
twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/
images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/
profile_images\/427887935523155968\/HC1pZ4sC_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/
profile_images\/427887935523155968\/HC1pZ4sC_normal.jpeg","profile_banner_url":"https:\/\/pbs.twimg.com\/profi
le_banners\/994395594\/1390851403","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","prof
ile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"default_pr
ofile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"g
eo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"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
```

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?

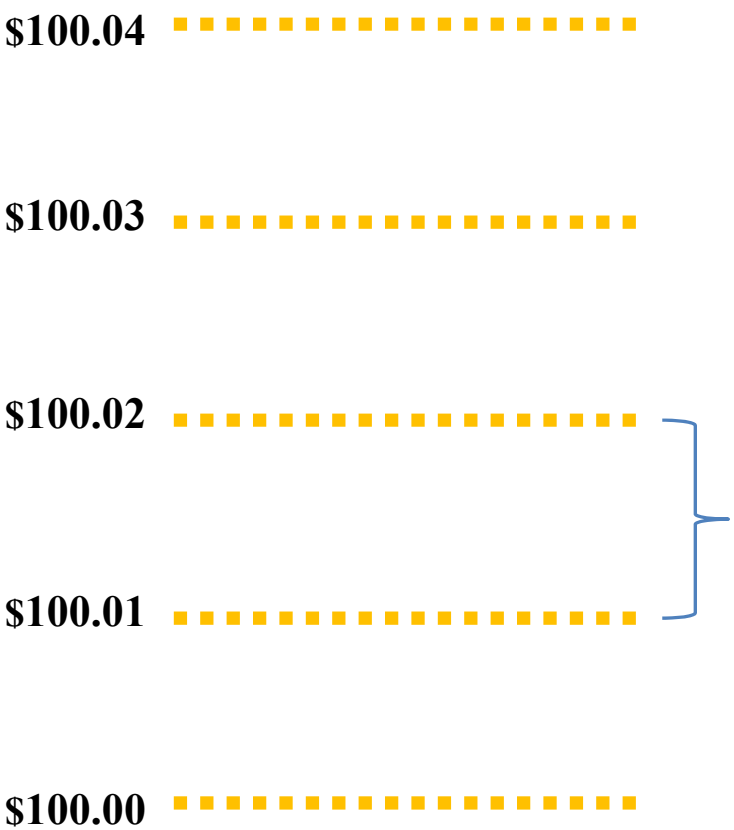
Structure Challenges

- Techniques
  - Find an alternative data vendor
  - Work with experts in other fields
- Economic insights
  - Unstructured data create unique measures of economic activity
  - Unstructured data help financial economists to test economic theory

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

What Drives the Arms Race in Speed?



Standard Walrasian equilibrium

- Continuous price

Reality: Price is discrete

- Tick size
- Minimum price increment imposed by SEC Rule 612

Liquidity Supplier: Limit Order Submitter

Limit order
An offer to buy or sell

$100.02

$100.01

$100.00

Toni:
An offer to buy 100 shares at $100.00

$99.99

Liquidity Demander: Market Order Submitter

Limit order
An offer to buy or sell

$100.02

$100.01

$100.00

Market order
Accepts the limit order

$99.99

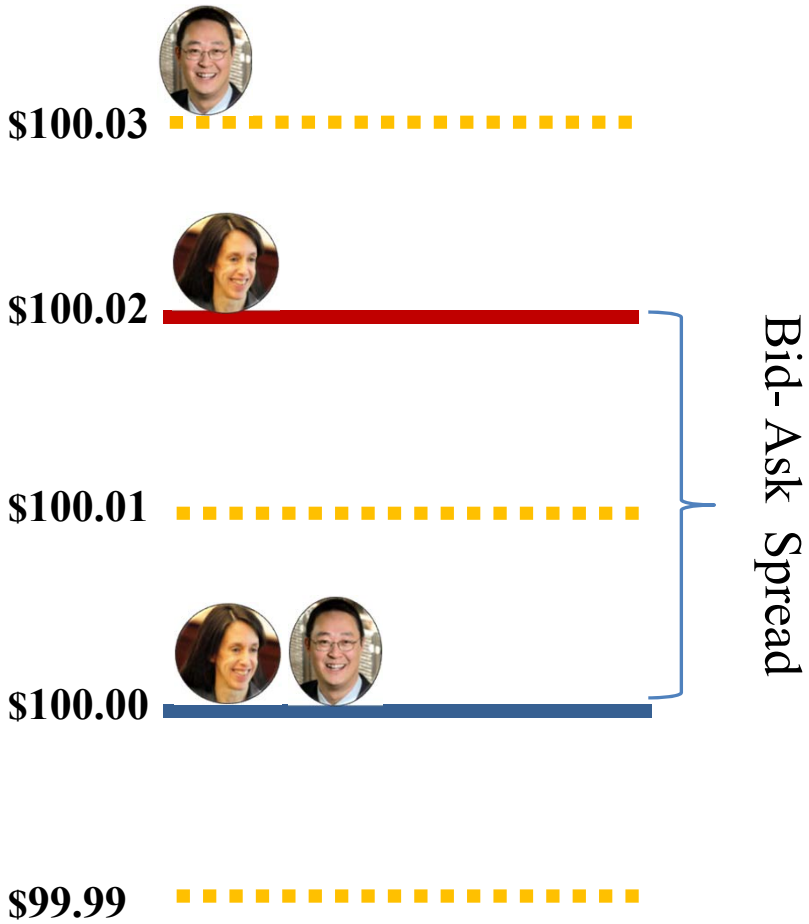
Toni:
An offer to buy 100 shares at $100.00

Mao:
Sells 100 shares at $100.00

Execution Priority for Liquidity Suppliers

1. Price: First priority
   - Better-priced limit orders execute first
     - Limit sells at lower prices
     - Limit buys at higher prices

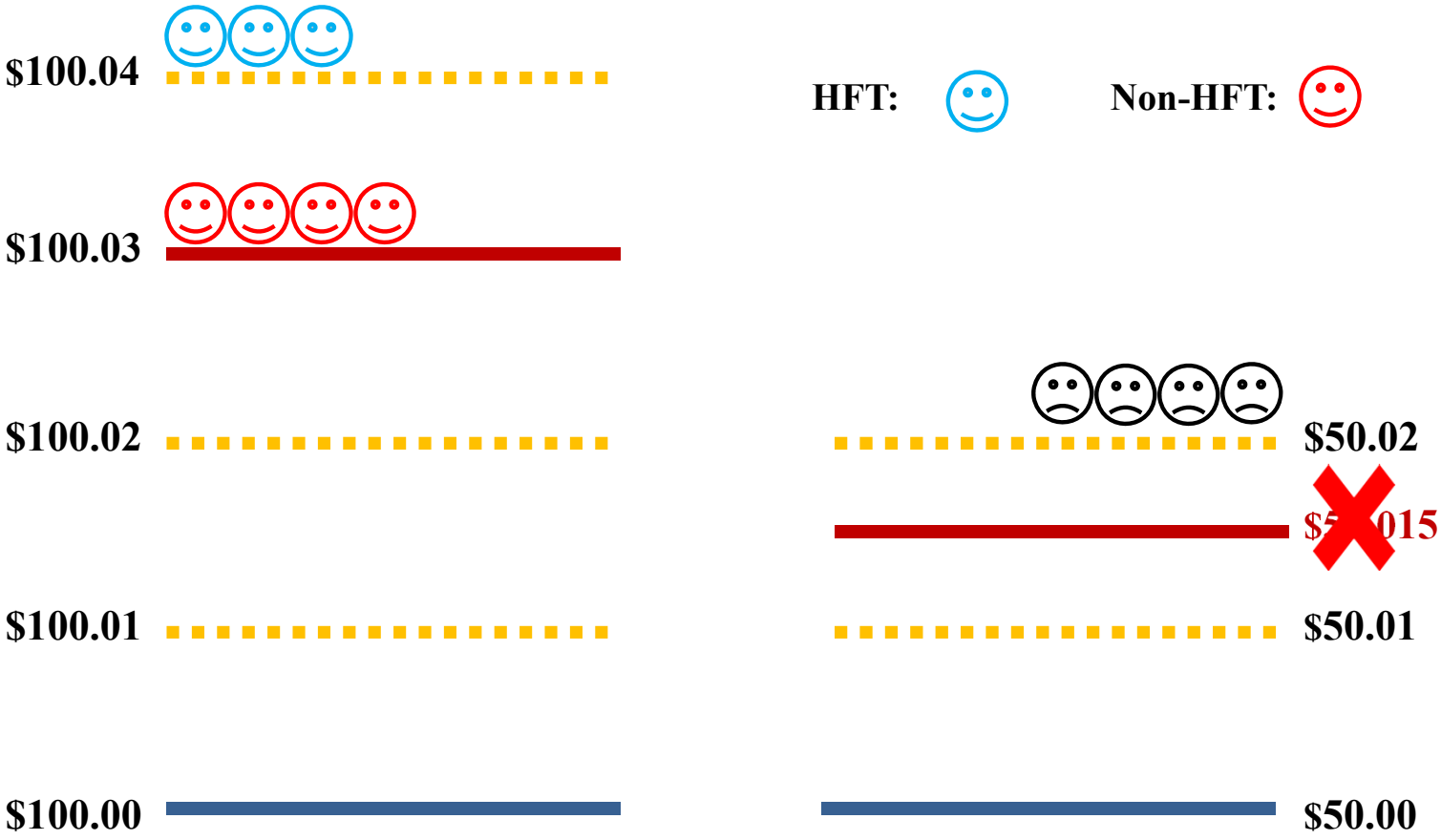
2. Time: Second priority
   - At the same price: first come, first served



Identification: ETF Splits/Reverse Splits

- Hypothesis: High-frequency traders (HFTs) provide more liquidity for low-priced securities
  - One cent tick size is more binding
- Treatment group: ETFs that split/reverse split
  - Splits decrease price
  - Reverse splits increase price
- Control group: ETFs track the same index but do not split/reverse split
- Tens of TBs of trading data (Yao and Ye, 2018)
  - Supercomputer helps to analyze 64 splits/reverse splits in four years

Price vs. Speed Competition



Puzzles

- Who are these non-HFTs?
- Why do they quote better prices than HFTs?
- Analysis of big data → new theory → new analysis of big data

Two Types of Traders

- HFTs: Computers

- Humans

The Third Type

- HFTs: Computers
- Half human, half computer
- Humans

Half Human, Half Computer

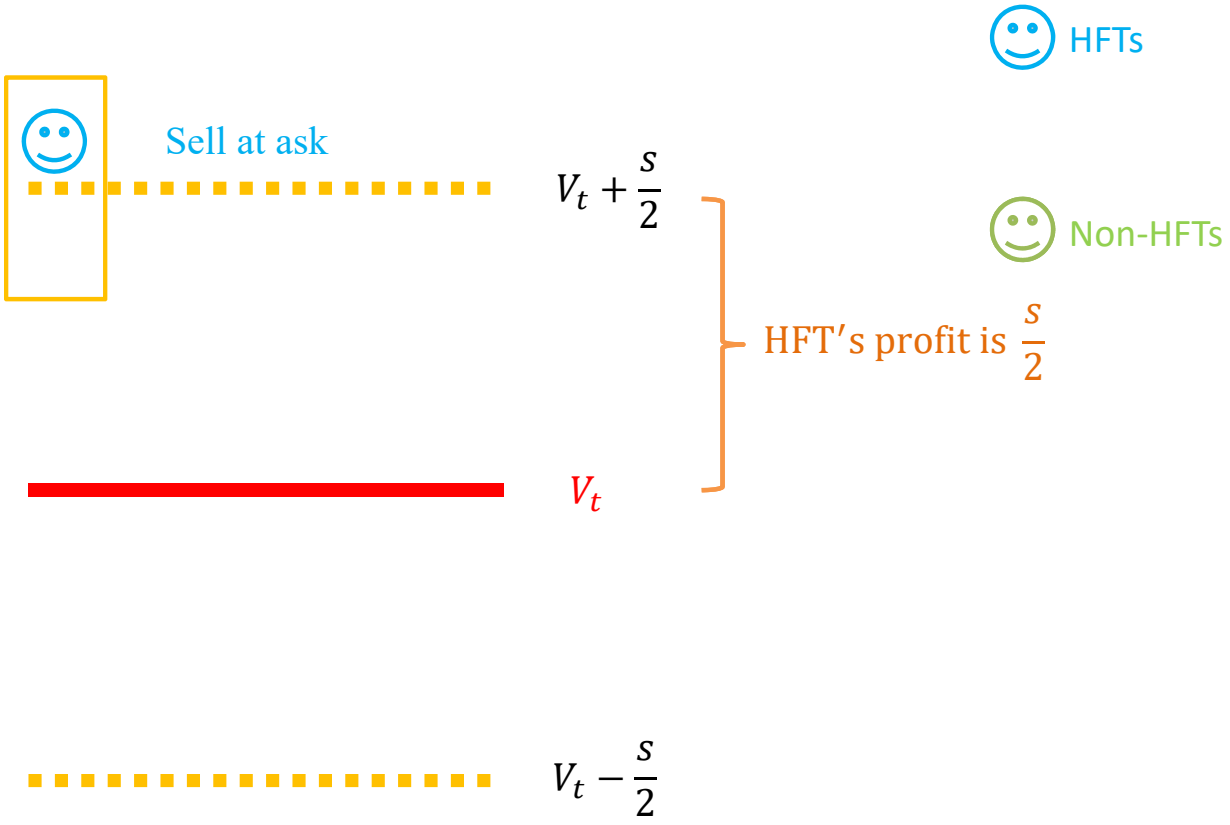
Buy-side Algorithmic Traders (BATs)

- BATs: Half human and half computer
- Humans (e.g., portfolio managers) make investment decisions
- Algorithms execute orders for portfolio managers
  - Demand or provide liquidity to minimize transaction costs
- BATs are faster than humans
- BATs are slower than HFTs
  - No need to be as fast as micro or nanoseconds

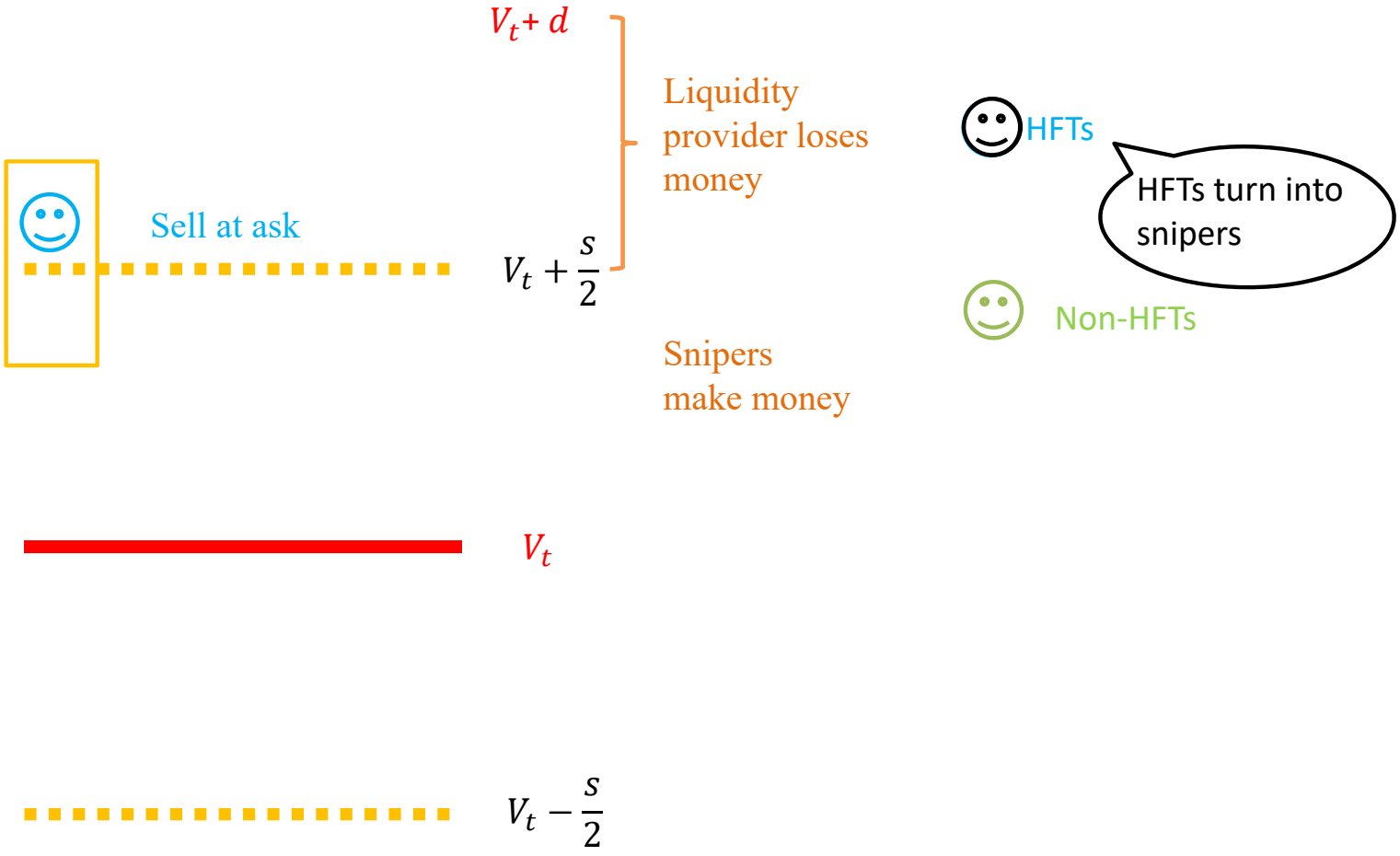
Benchmark: Budish, Cramton, and Shim (2015)

- Continuous time, continuous price, and two types of traders
- HFTs: Continually monitor the market
  - Supply or demand liquidity if the expected profit is above 0
- Non-HFTs arrive with an inelastic need to buy/sell one share
  - Arrival intensity λ_I
  - Only demand liquidity
- Security value v_t evolves as a compound Poisson process
  - v_t is public information
  - Intensity of the jump event: λ_J
  - Size of the jump: d or $-d$

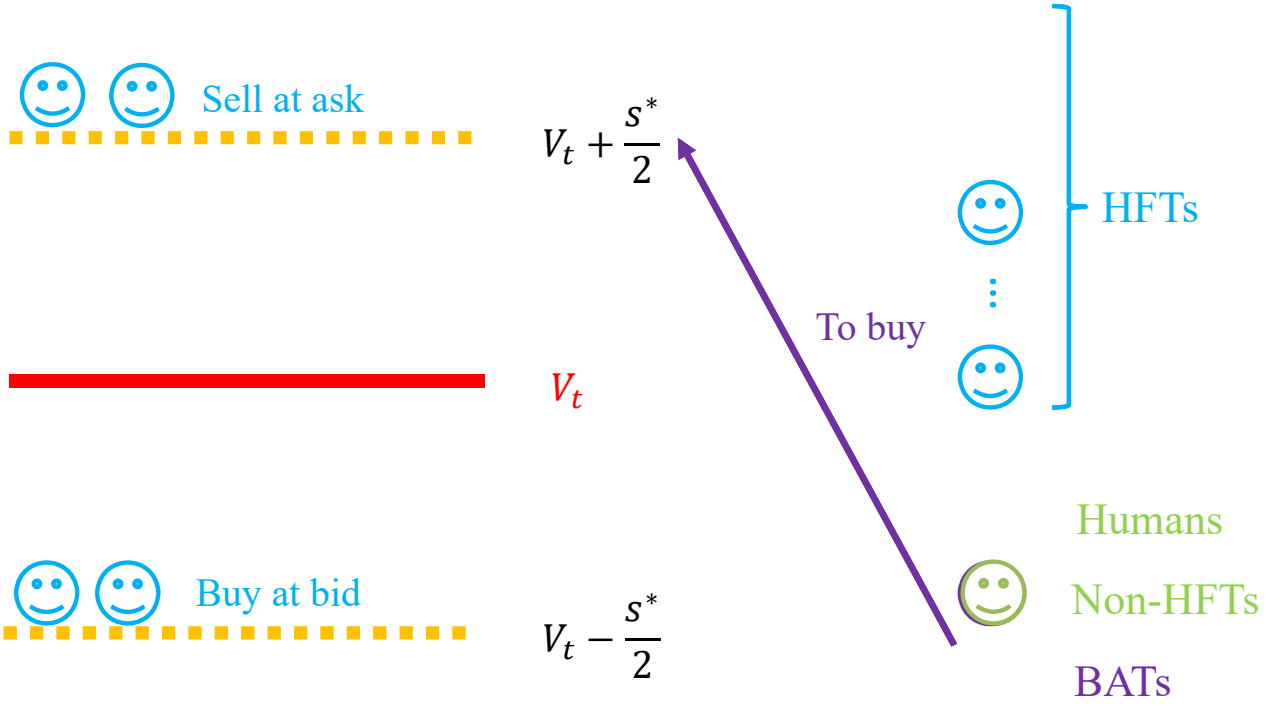
Revenue to Supply Liquidity to a Non-HFT



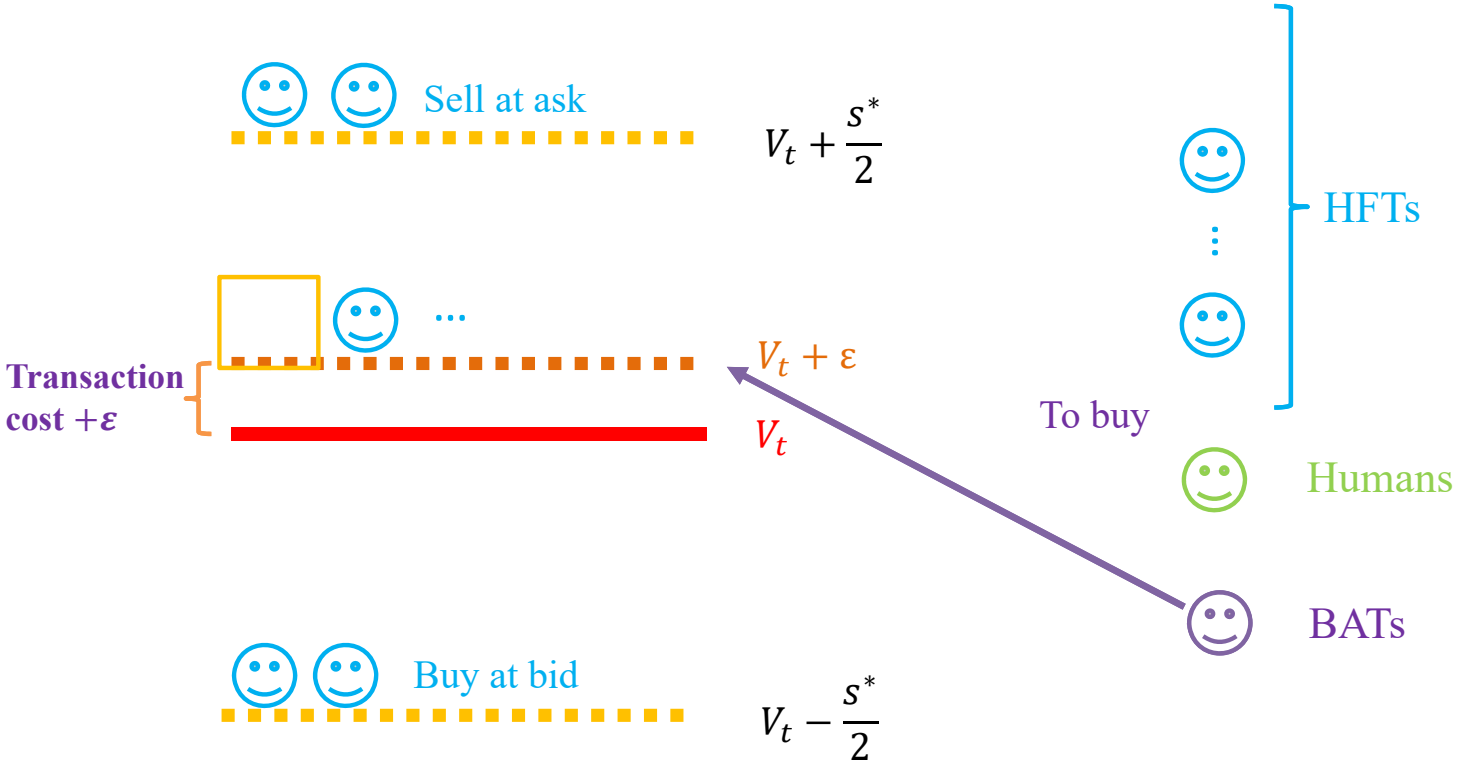
Cost of Being Sniped



Adding BATs (Li, Wang, and Ye, 2018)



Continuous Price: BATs always Supply Liquidity



Machine-Machine Interactions

- BATs always provide liquidity
  - Lower opportunity costs for providing liquidity
  - BATs have to buy or sell

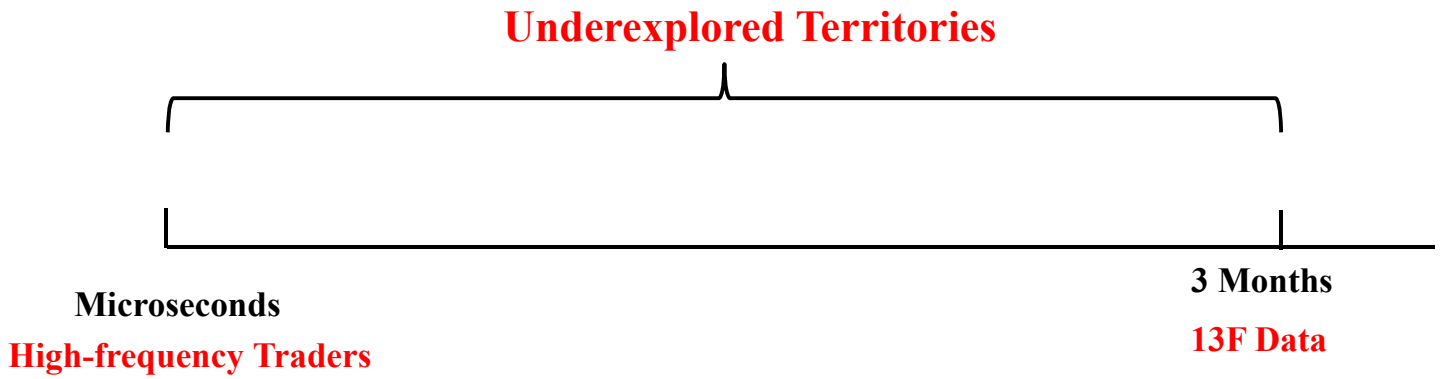
- HFTs' strategy to sell
  - Offer to sell at $V_t + \frac{s^*}{2}$
    - Limit sell orders are more likely to be executed when v_t jumps upward
  - Accept BATs' offer at $V_t + \varepsilon$
    - Immediate execution with no sniping risk

- Machine-machine interactions blur the definition of liquidity provision
  - BATs arrive first
  - HFTs immediately respond

Predictions and Policy Implications

- Li, Wang, and Ye (2018) explain several existing puzzles
  - Non-HFTs quote better prices due to lower opportunity costs
- New predictions after adding discrete size: Four types of equilibria
  - Theory works well to predict who provides liquidity and when
  - Machine-machine interactions provide a clean environment to test theories
- Policy implication: SEC's tick size pilot program
  - Increase the tick size to 5 cents for 1,200 pilot stocks
  - Implication from the model: A large tick size would fuel speed competition

Financial Market Ecosystem

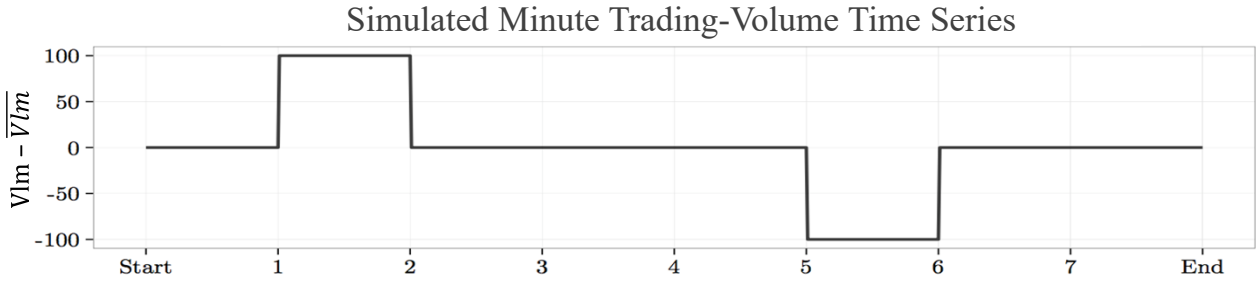


- Paucity of studies on traders who are slower than HFTs but faster than a quarter
  - BATs operate at timescales of milliseconds or seconds
  - Traders who use machine-learning techniques operate at timescales of "anywhere from a few minutes to a few months."

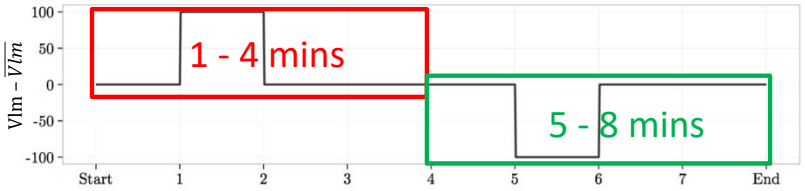
Our Solution: Wavelet Estimator

- Aggregate each stock's trading-volume data to the one-minute timescale from TAQ data
- Decompose each stock's trading-volume variance into timescale-specific components with the wavelet-variance estimator
  - Chinco and Ye (2018)

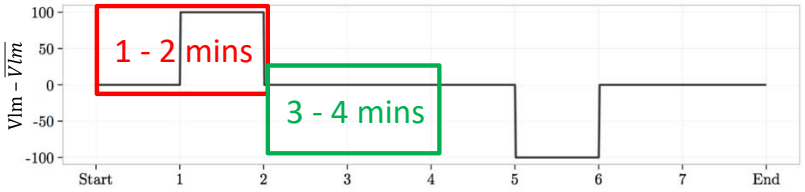
Intuition



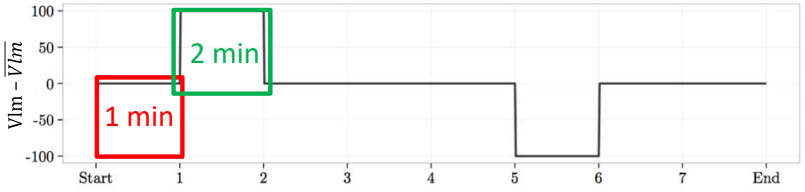
What happens in different horizons?



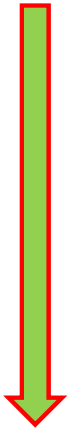
Low frequency



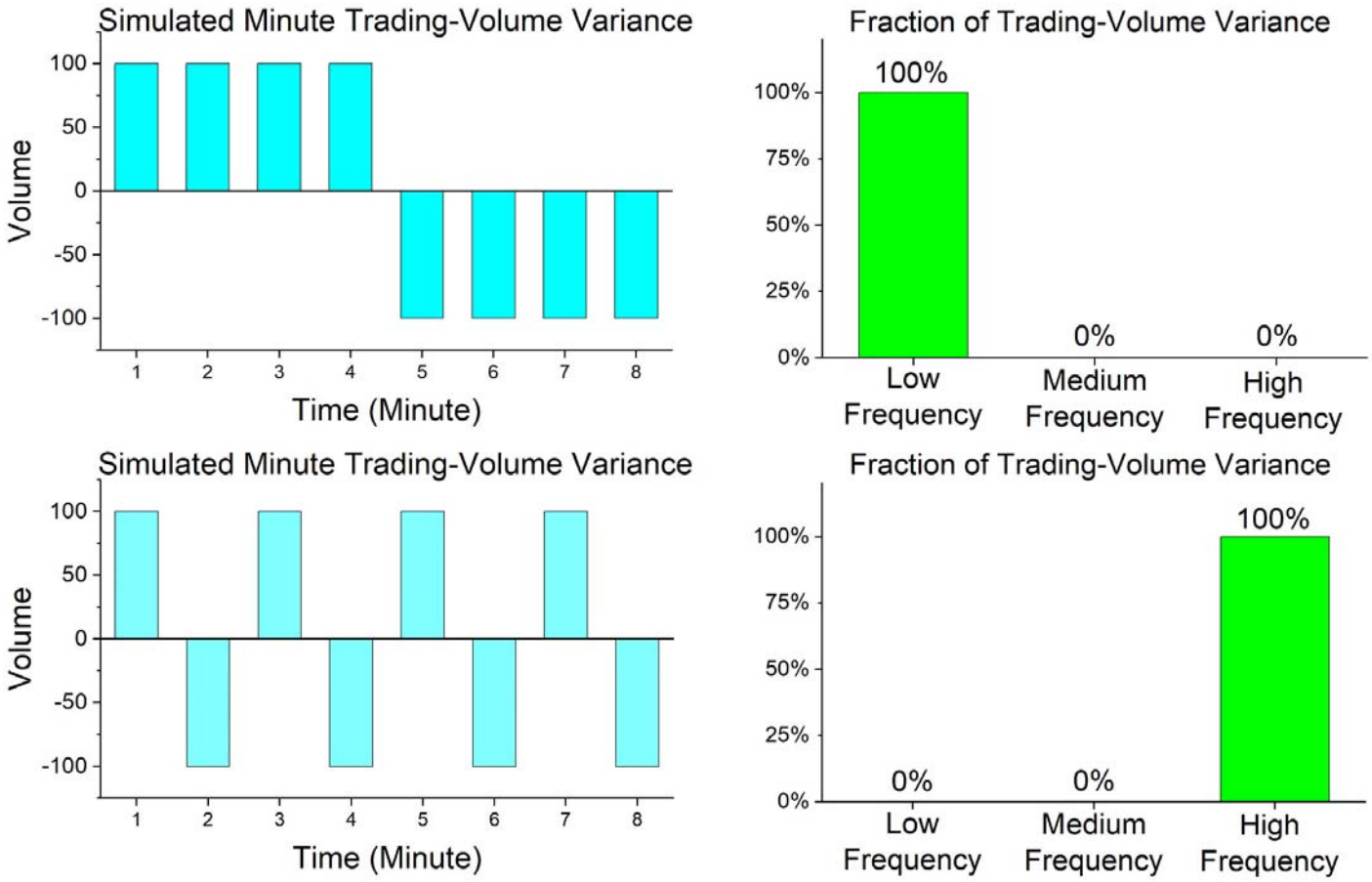
Median frequency



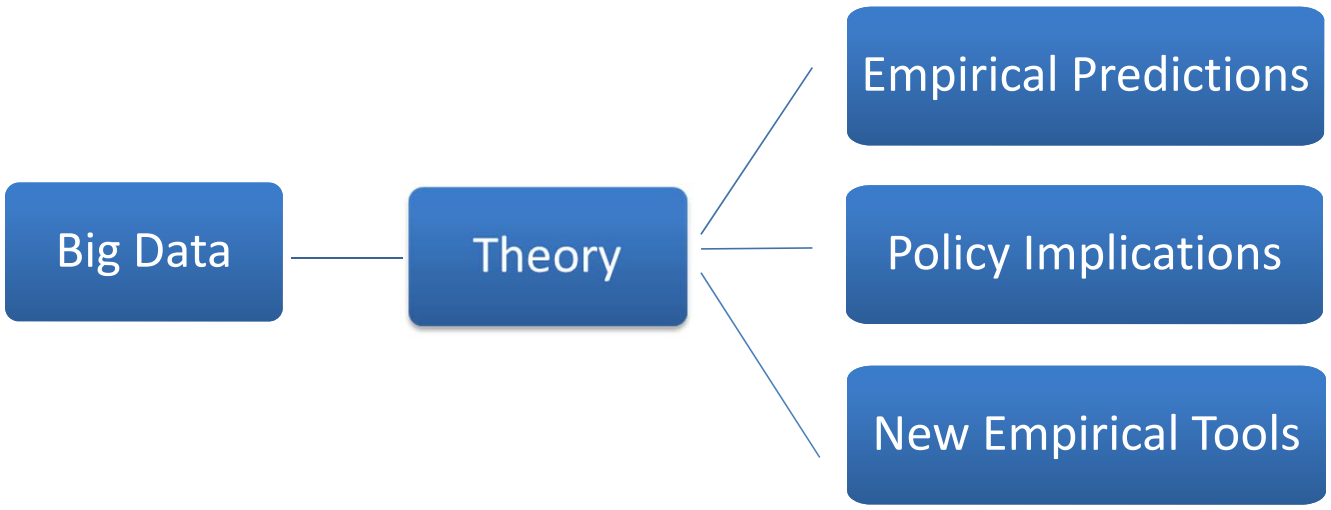
High Frequency



Wavelet Decomposition



Big Data Research Strategy



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 open doors for new research questions

- Document new empirical regularities and inform public policy
- Motivate us to find economic interpretations of new data
- Create unique measures to test theories and motivate us to construct new theories