Could Probability of Informed Trading Predict Market Volatility?

John Wu

Wes Bethel, David Leinweber

Oliver Rübel, Ming Gu

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

One of world's premier research institutions

Mission: Solve the most pressing and profound scientific problems facing humankind

- *Basic science for a secure energy future*
- *Understand living systems to improve the environment and energy supply*
- *Understand matter and energy in the universe*

16 Nobel Prizes,
2 Elements (Lawrencium & Berkelium)

Pioneer and Center of Excellence in **Data Intensive Science**

People

- *3,863 FTE*
- *3,040 Employees*
- *267 Joint faculty*
- *491 Postdoctoral researchers*
- *328 Graduate students*
- *194 Undergraduates*
- *8,025 Facility users*
- *1,612 Visiting scientists and engineers*

FY10 Total Operating Costs:
$680.6M

LBNL at-a-glance

Advanced Light Source

National Energy Research Scientific Computing Center (NERSC)

88-Inch Cyclotron

Molecular Foundry

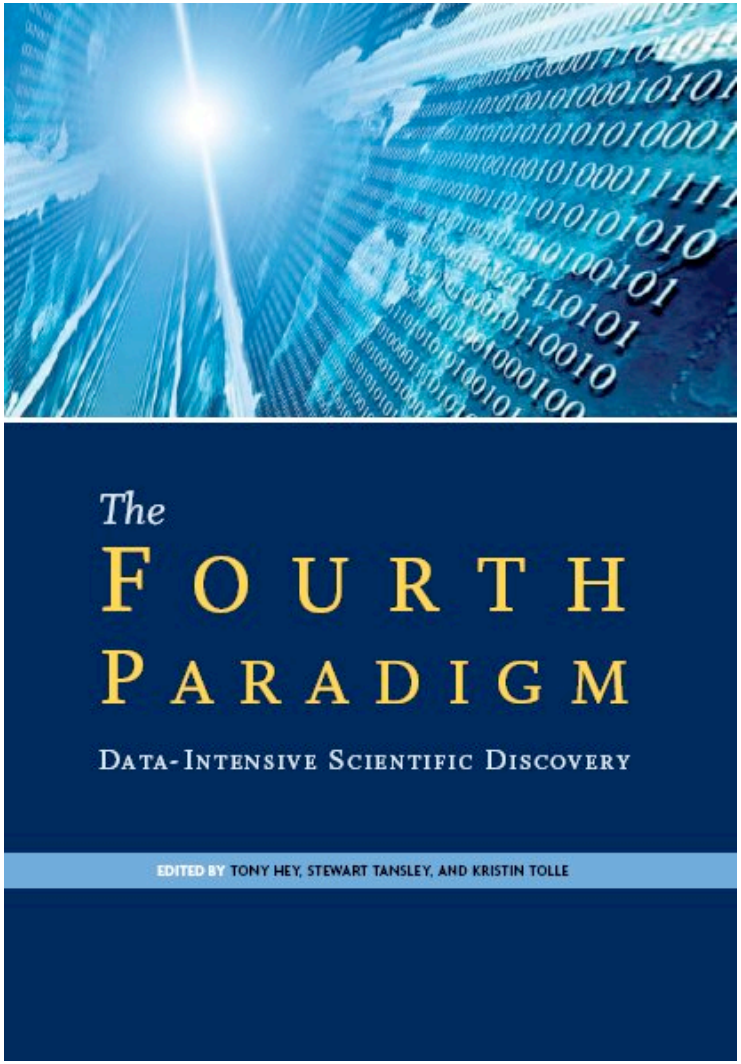
Energy Scences Network (ESnet)

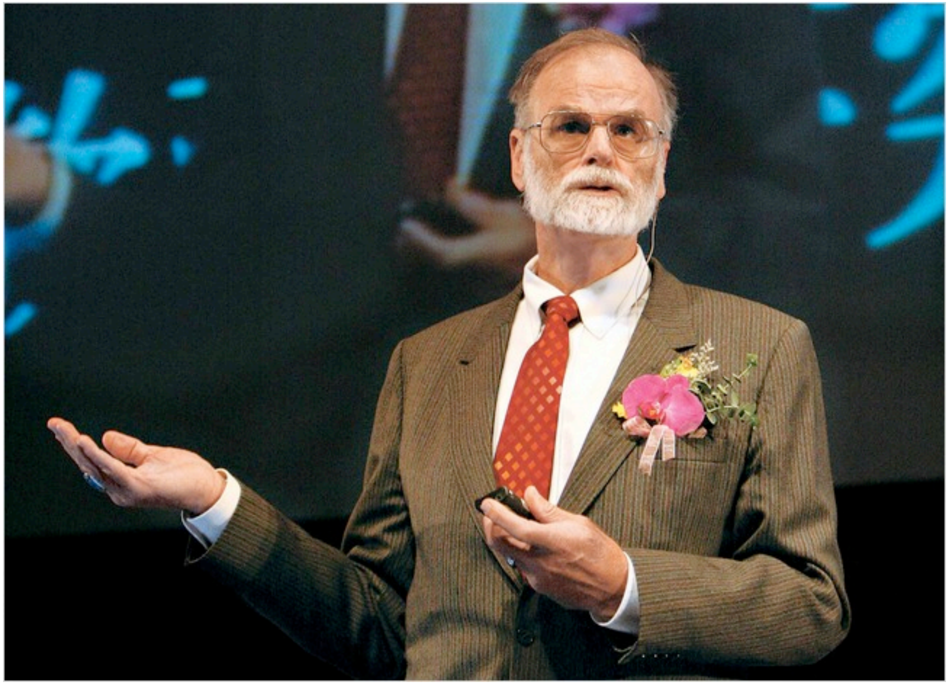
Joint Genome Institute

National Center for Electron Microscopy

Berkeley Lab's largest **research facilities** see more than 25,000 users per year˝

Can we discover the causes by borrowing from data-intensive sciences?


The FOURTH PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



Jim Gray -- Turing Award Winner, 1998

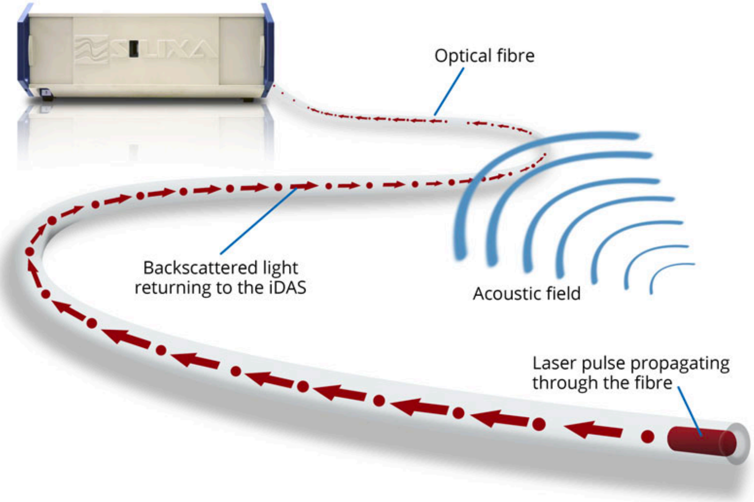
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

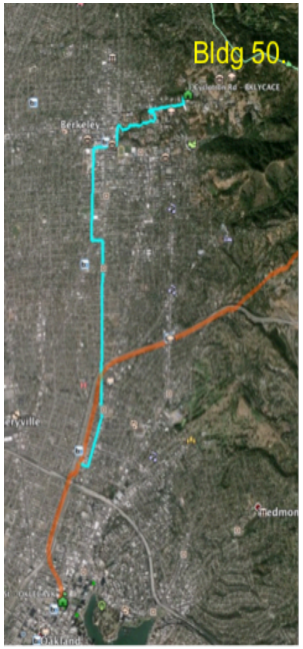
Example: Distributed Acoustic Sensing For Seismic Monitoring

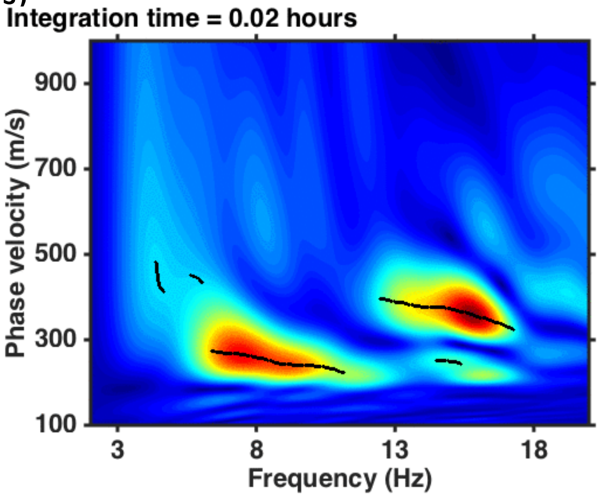
- **Distributed Acoustic Sensing [DAS]** is a rapidly advancing approach for measuring the seismic wavefield using commercial fibers (SM, telecom)

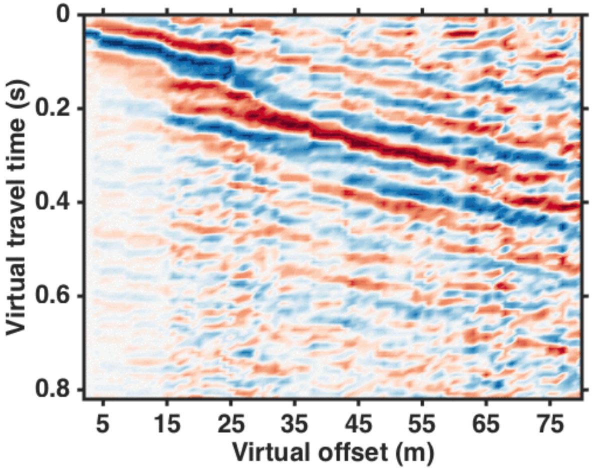
- **Recent** : S/N became sufficient for seismology around 2011. Our work started ~2012/13 out of CO_2 GCS program (borehole applications)

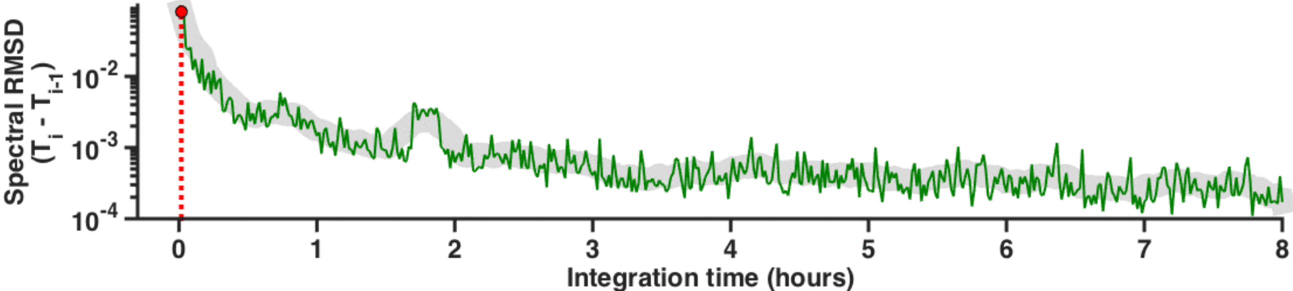
- **Large N** : Easy to deploy in wells, behind casing, 1000s to 100,000s of channels available (TB/day) over 10+ km (collected 0.25PB in 3 months)


SILIXA
Optical fibre
Backscattered light returning to the iDAS
Acoustic field
Laser pulse propagating through the fibre





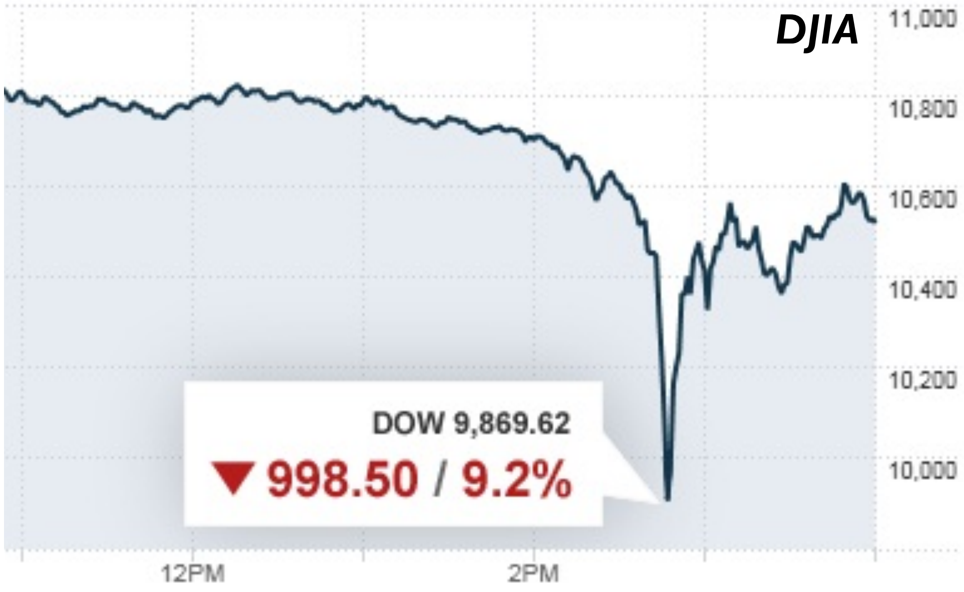




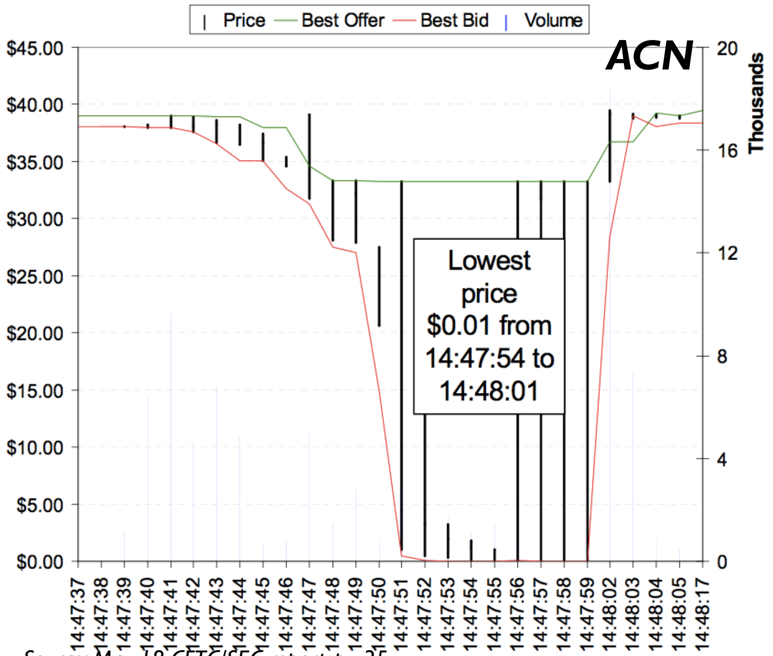
Flash Crash: May 6, 2010

May 6, 2010
Flash Crash

- Dow Jones Industrial Average (DJIA) dropped by nearly 1000 points in minutes, market capitalization decreased by about **1 trillion dollars**
- Many stocks went to pennies. Many didn't.
- Complex unexpected interactions across markets



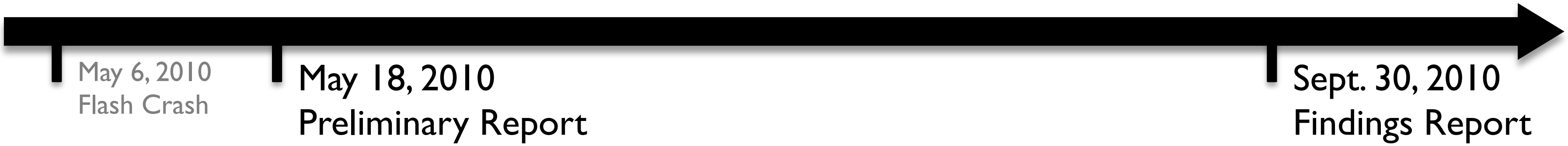
Source: *financial-planning.com*



Lowest price $0.01 from 14:47:54 to 14:48:01

Source: *May 18 CFTC/SEC report, pg 35*

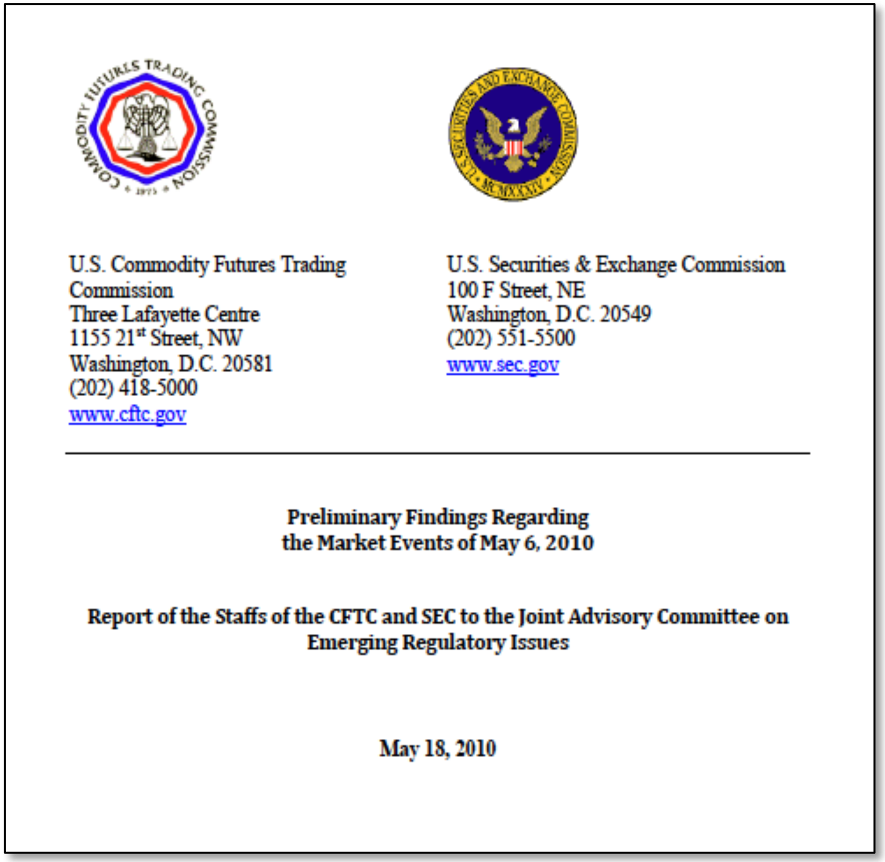
Flash Crash: Official Report Took 5 Months



May 6, 2010
Flash Crash

May 18, 2010
Preliminary Report

Sept. 30, 2010
Findings Report



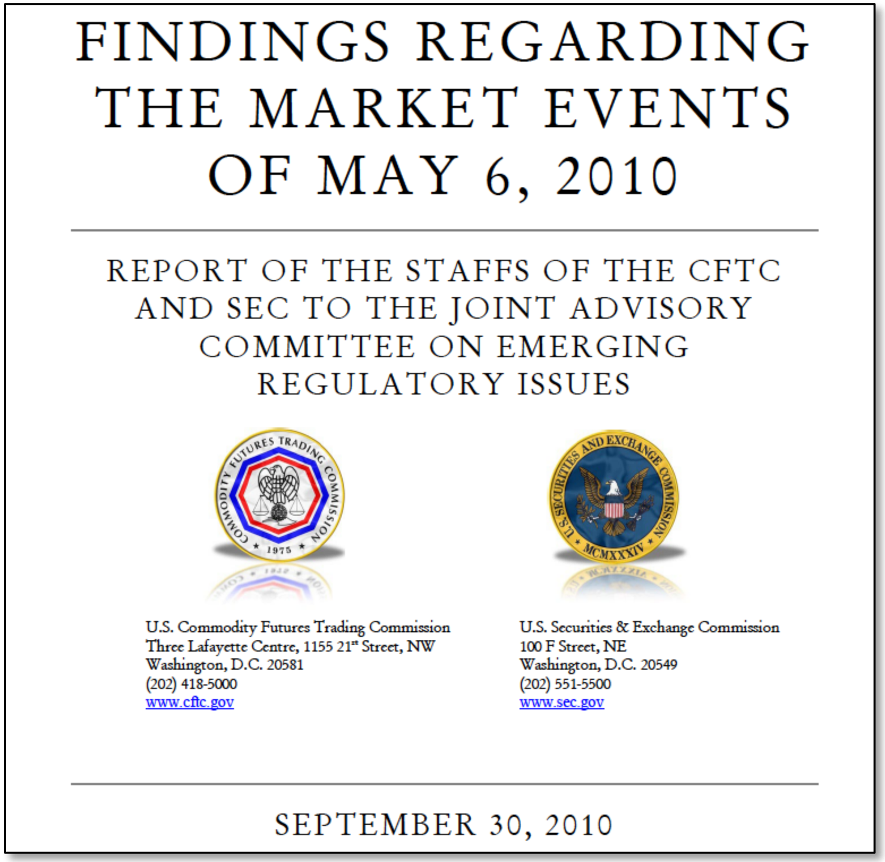
U.S. Commodity Futures Trading Commission
Three Lafayette Centre
1155 21st Street, NW
Washington, D.C. 20581
(202) 418-5000
www.cftc.gov

U.S. Securities & Exchange Commission
100 F Street, NE
Washington, D.C. 20549
(202) 551-5500
www.sec.gov

Preliminary Findings Regarding the Market Events of May 6, 2010

Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues

May 18, 2010



FINDINGS REGARDING THE MARKET EVENTS OF MAY 6, 2010

REPORT OF THE STAFFS OF THE CFTC AND SEC TO THE JOINT ADVISORY COMMITTEE ON EMERGING REGULATORY ISSUES

U.S. Commodity Futures Trading Commission
Three Lafayette Centre, 1155 21st Street, NW
Washington, D.C. 20581
(202) 418-5000
www.cftc.gov

U.S. Securities & Exchange Commission
100 F Street, NE
Washington, D.C. 20549
(202) 551-5500
www.sec.gov

SEPTEMBER 30, 2010

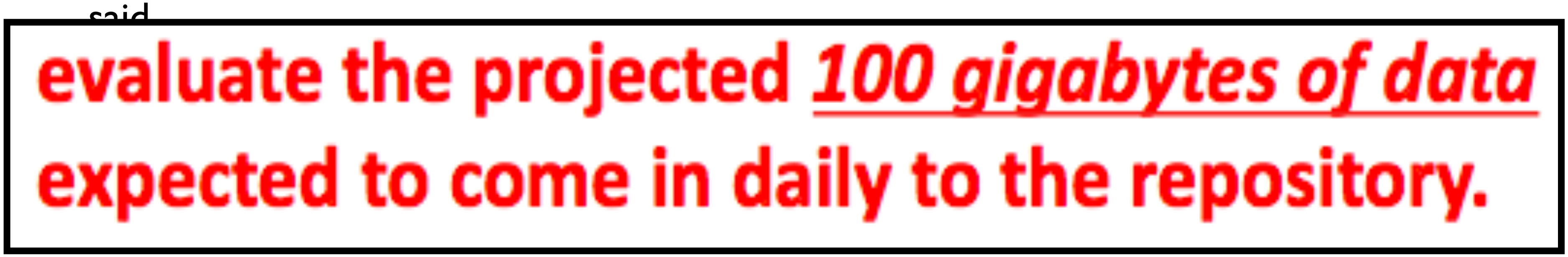
[1] SEC/CFTC. Preliminary report: http://www.sec.gov/sec-cftc-prelimreport.pdf
[2] SEC/CFTC, Findings report: http://www.sec.gov/news/studies/2010/marketevents-report.pdf

"The SEC's efforts to reconstruct the trading on that day are substantially more challenging and time consuming than we would have liked **because no standardized, automated system exists to collect data** across the various trading venues, products and market participants," Schapiro said.

Commissioner Luis Aguilar **questioned, however, whether the SEC would have the human and technological resources to evaluate the projected *100 gigabytes of data* expected to come in daily to the repository.**

"The SEC's staff must be equipped with the best resources to do the job," Aguilar said. **"Most Americans assumed the SEC has these tools. It is shocking that the SEC does not have its own access to this data.**

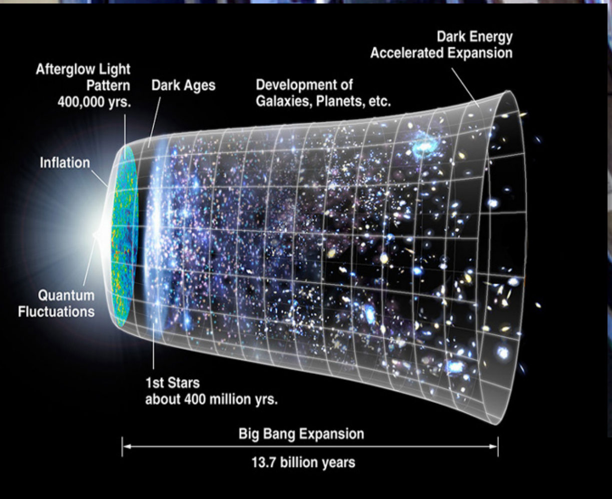
"The SEC must have this data and the tools to identify egregious conduct, identify trends and reconstruct market movements."



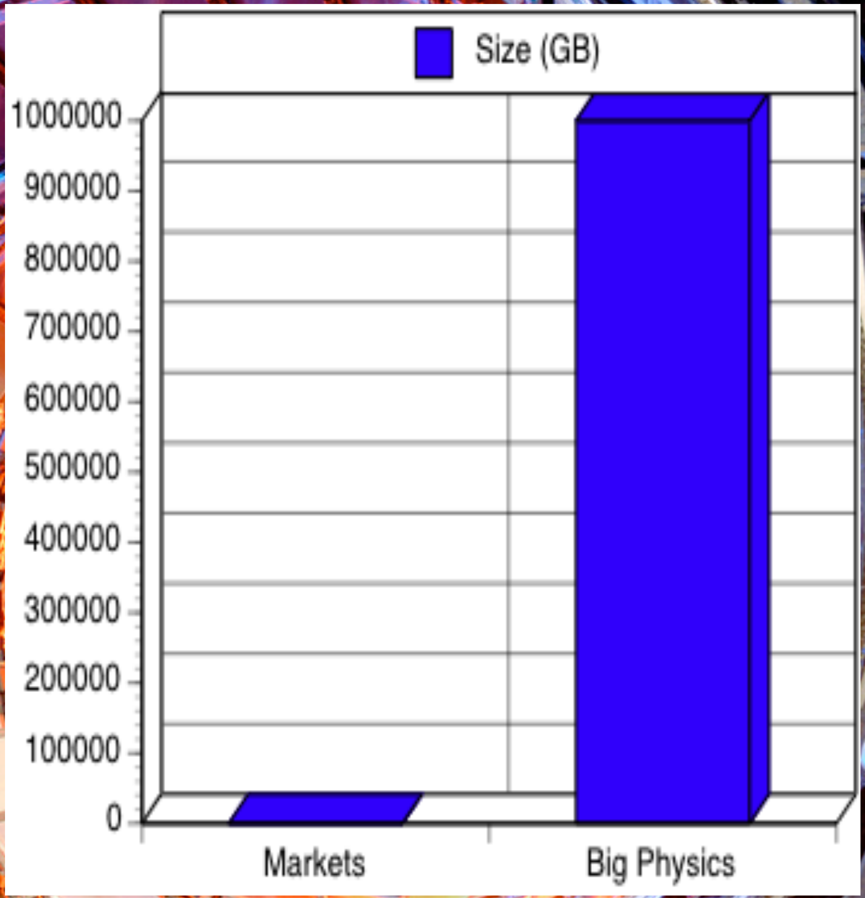
evaluate the projected *100 gigabytes of data* expected to come in daily to the repository.

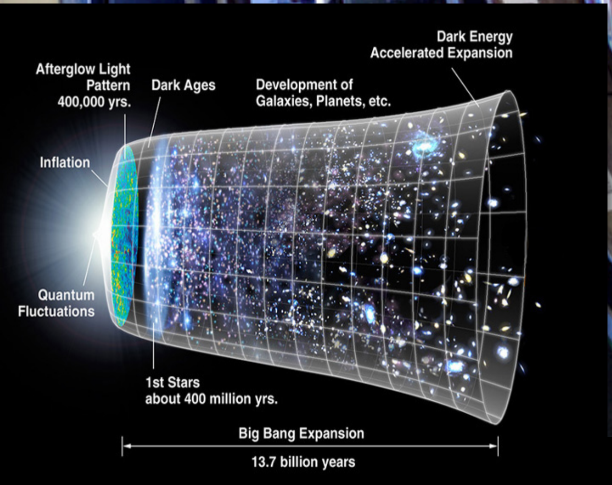
http://www.marketwatch.com/story/sec-to-seek-more-data-about-high-frequency-trades-2010-05-26-105000

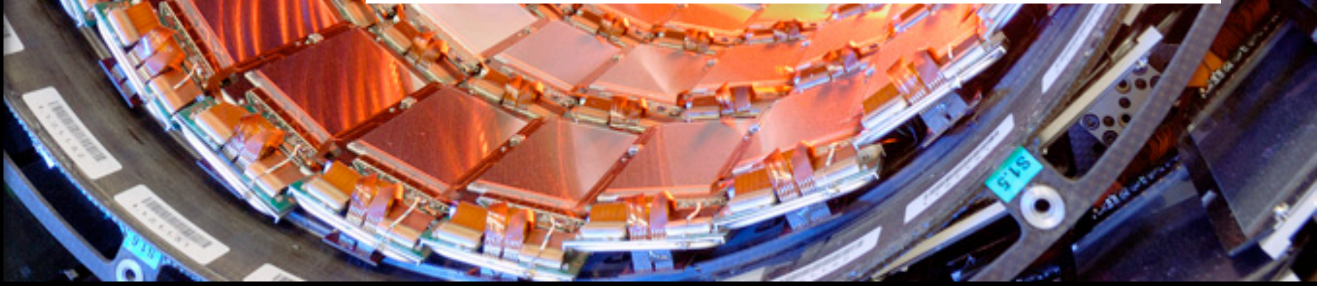
You call *that* big data?



CERN


Size (GB)
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
Markets
Big Physics


Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
400,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
Quantum
Fluctuations
1st Stars
about 400 million yrs.
Big Bang Expansion
13.7 billion years

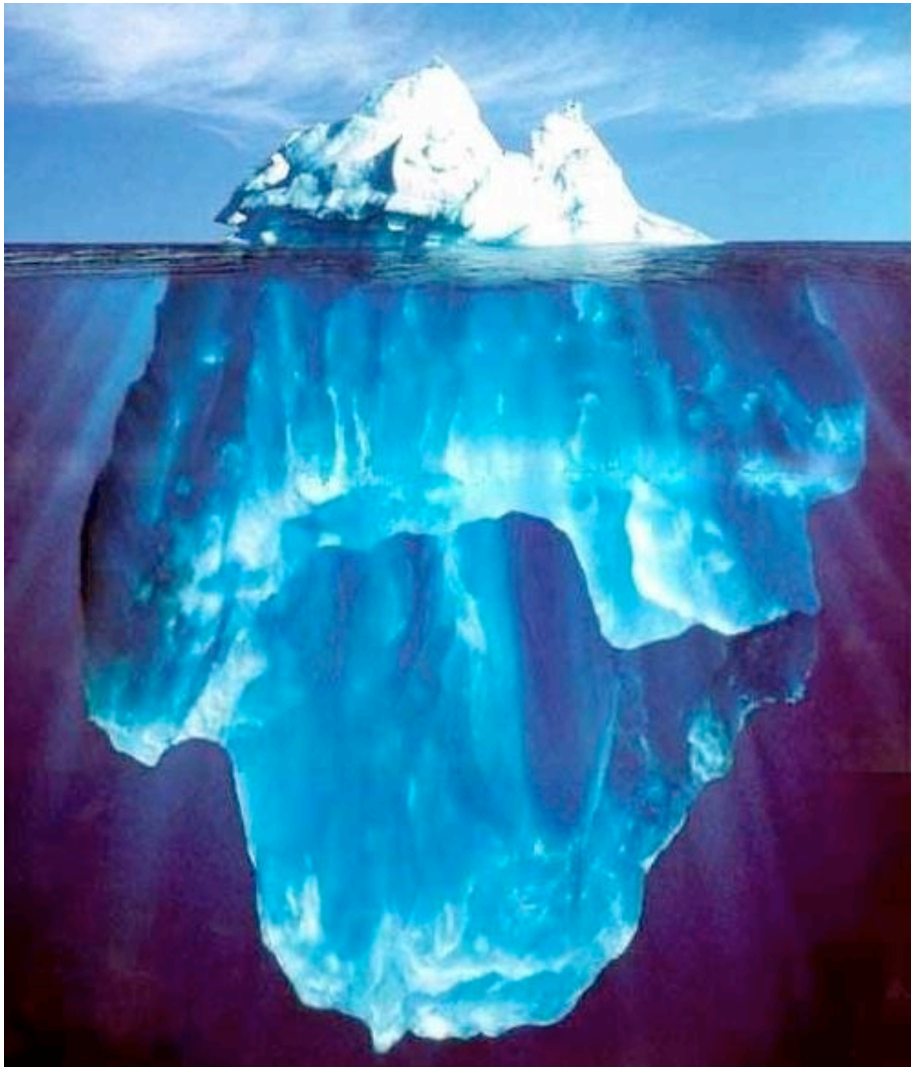


What's In This Work

- Basic understanding the trading data
- Preliminary examination of storage strategy
- Early-warning indicators of flash events
- Interactive exploration with bitmap indexes

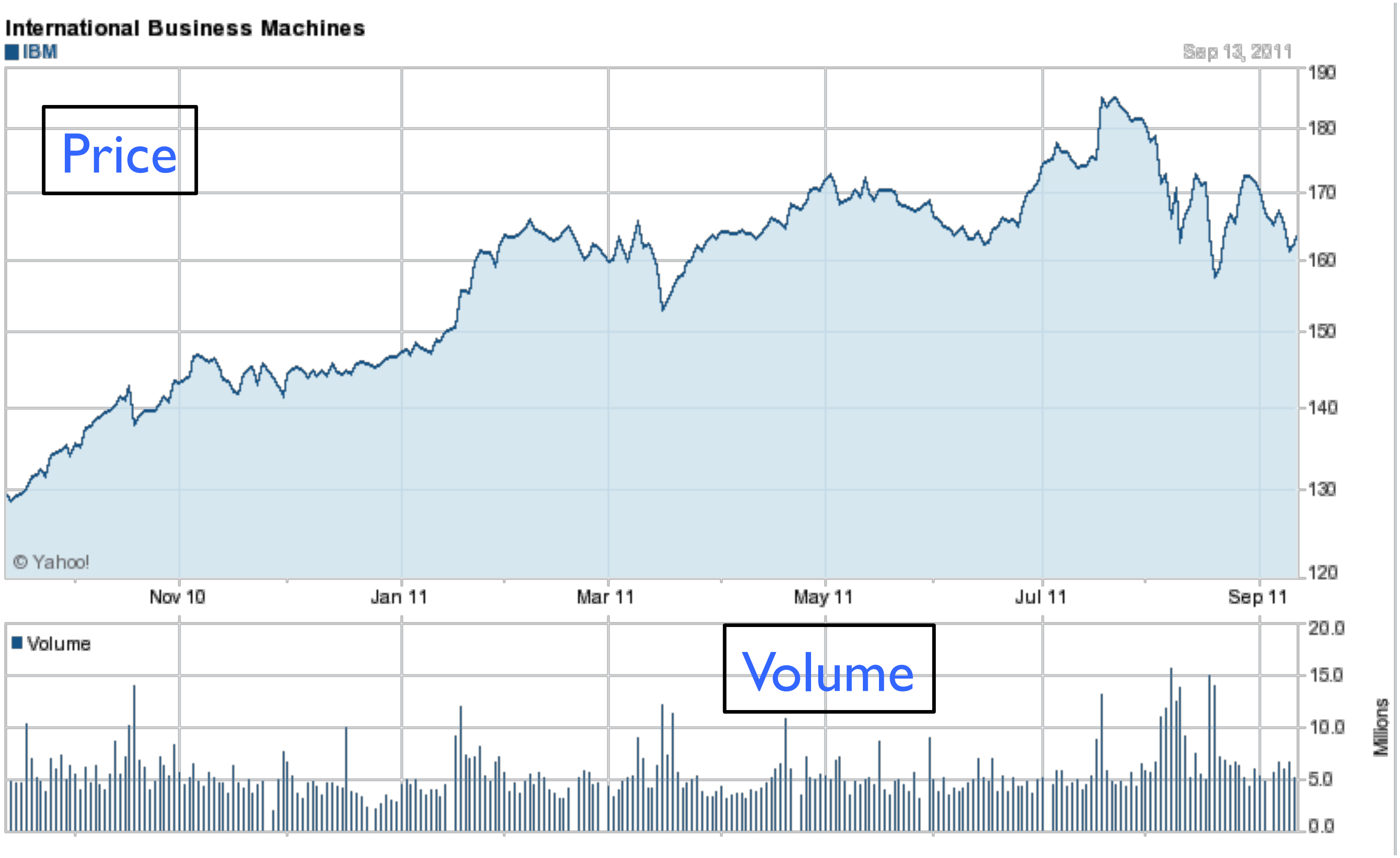
Levels of Financial Market Data

- Level 1: Trades (Transactions)
- Level 2: Best Bid / Offer
- Level 3: Limit Order Book (LOB) Snapshots
- Level 4: Order Flow
- Level 5: Identifying information
- Level 6: System health



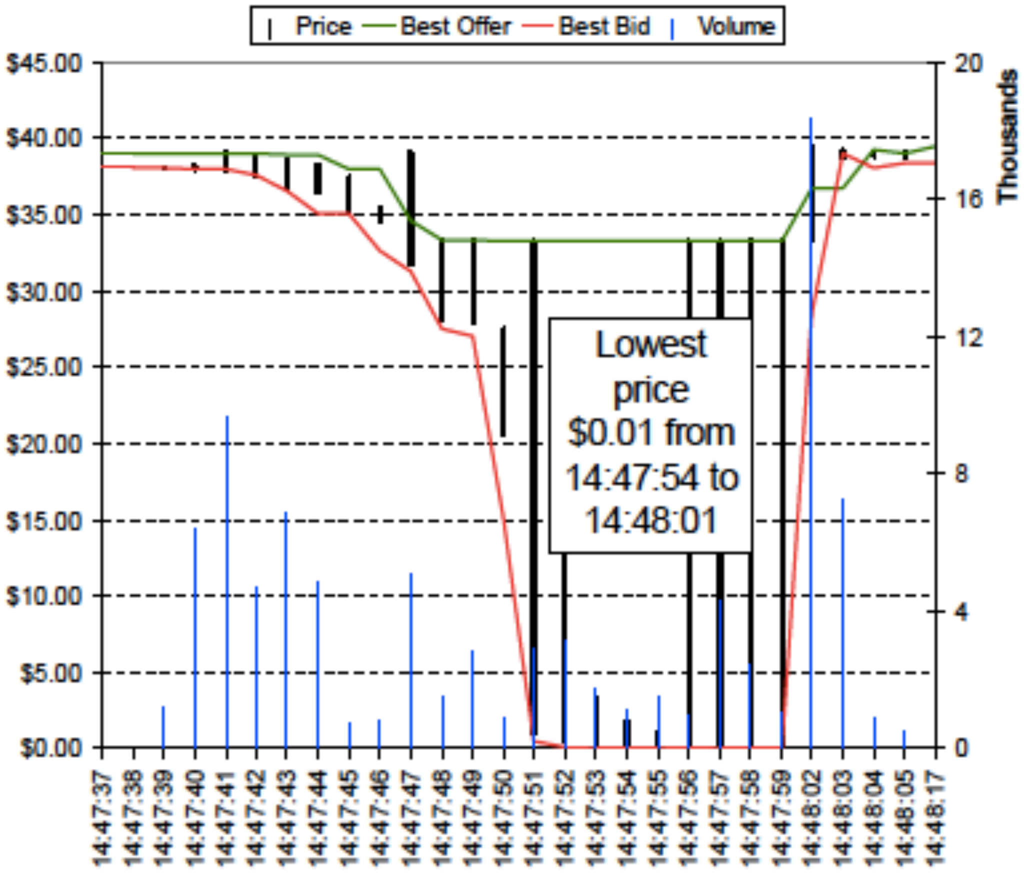
Level I of Market Data - Trades

- Trades (Transactions): trade prices and volumes



Level 2 of Market Data - Best Bid & Offer

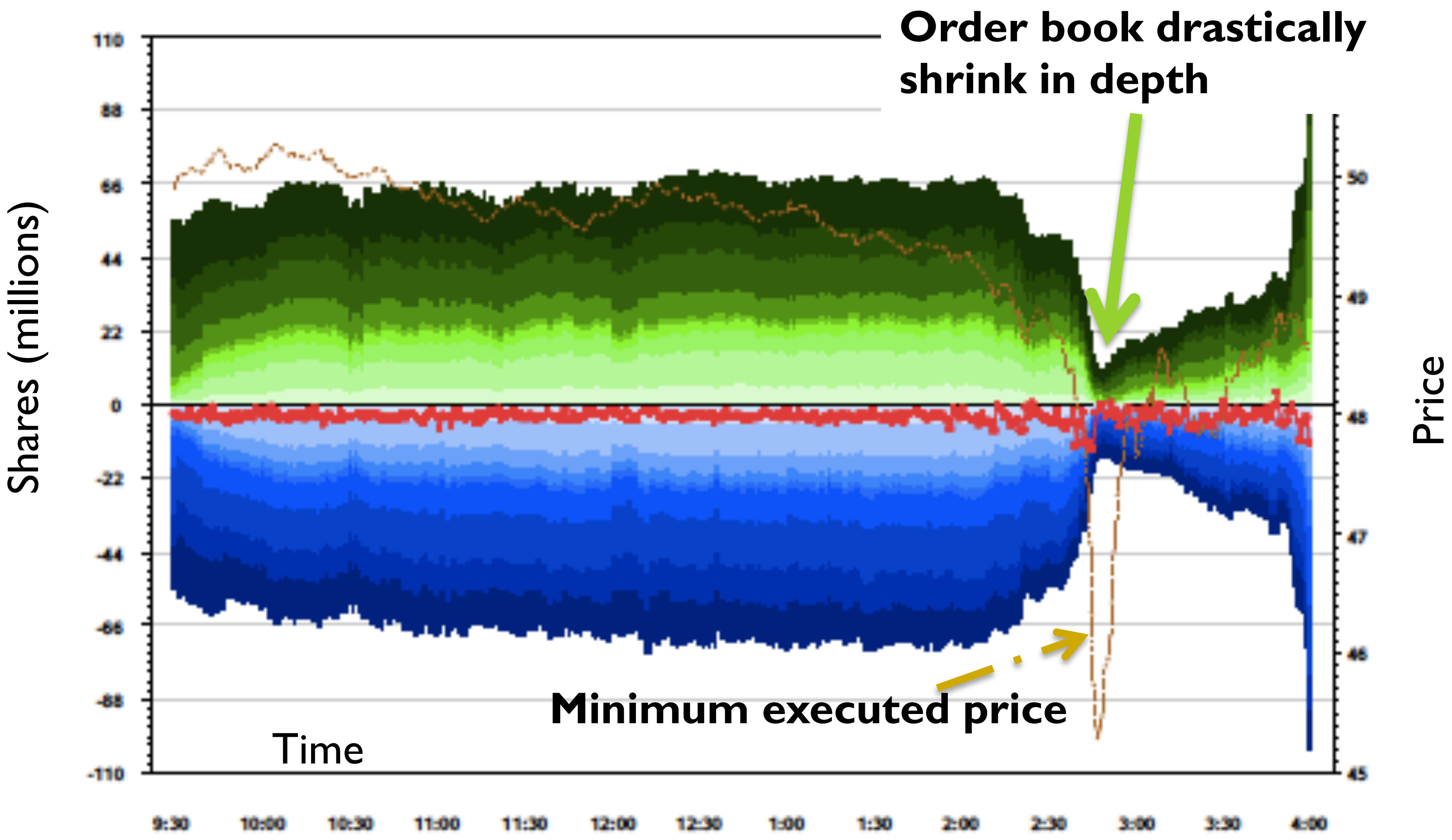
- Level 2 – Adds Best Bid / Offer (BBO) Quotes to Level 1 data
- Figure on the right appeared in the SEC / CFTC preliminary report on May 6 2010 Flash Crash is based on Level 2 data about Accenture (ACN)



http://www.sec.gov/sec-cftc-prelimreport.pdf

Level 3 of Market Data
Limit Order Book Snapshots

- Below is a visualization from the September Report on May 6 2010 Flash Crash using Level 3



Order book drastically shrink in depth

Minimum executed price

http://www.sec.gov/news/studies/2010/marketevents-report.pdf

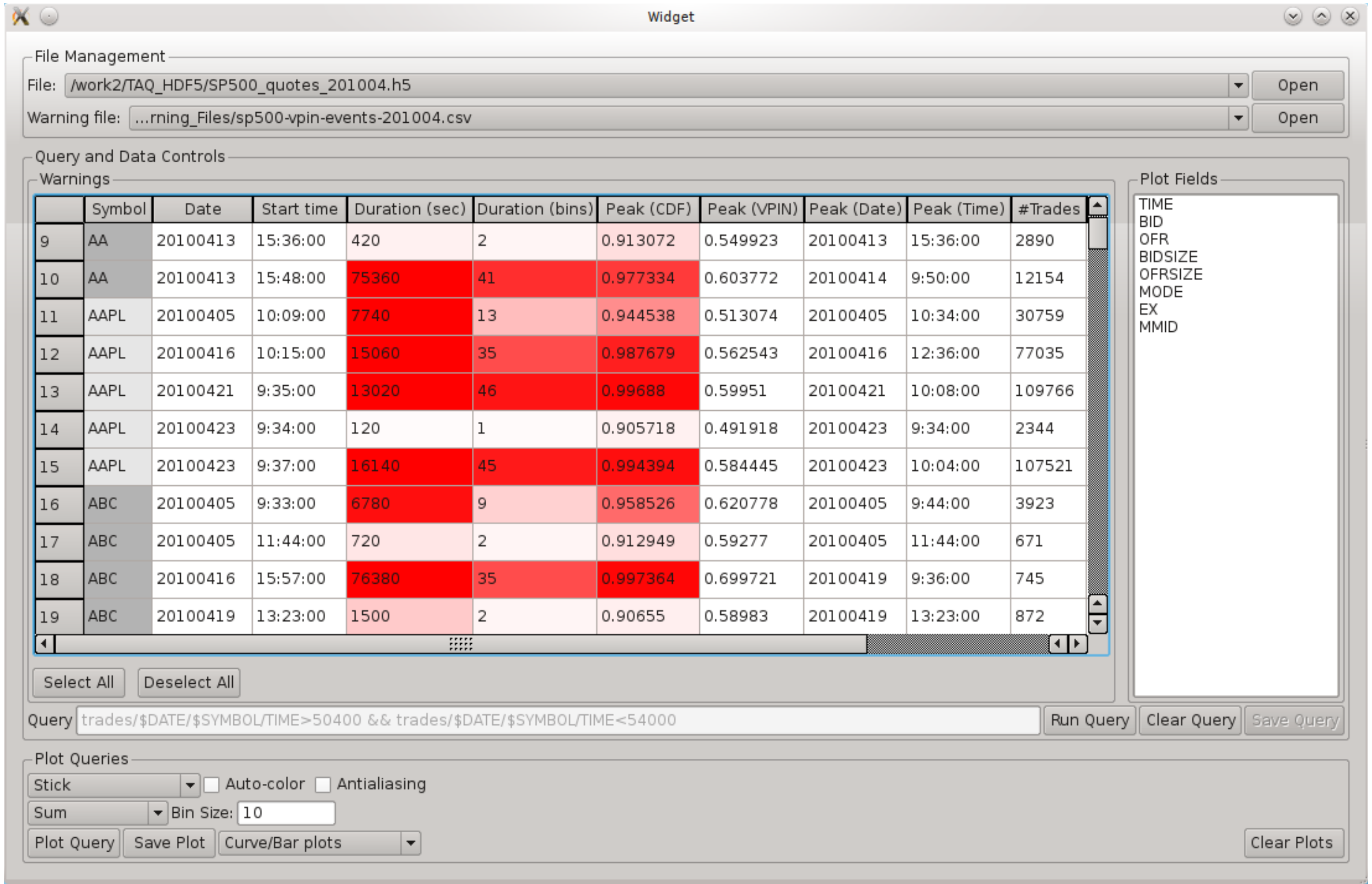
Market Data in a Scientific Data Format

- Most academic research use ASCII data such as Coma Separated Values (CSV), while commercial endeavors usually employ proprietary formats
- We propose to store market data in a widely used scientific data format, HDF5, for reducing disk storage, increasing query performance, making it usable by more tools
  - Compute VPIN on two-month trades of ACN took 142 seconds using CSV, only 0.4 seconds with HDF5

	CSV	CSV(zip)	HDF5	HDF5(SZIP)	Index
Trades	2,769	215	1,326	472	1,803
Quotes	38,566	3,058	28,844	5,377	24,784

Interactive Exploration with Bitmap Indexes

- Develop a simple visualization of the earlier warnings
- In the figure below, warnings with high intensity and long duration are marked with brighter red background



File: /work2/TAQ_HDF5/SP500_quotes_201004.h5

Warning file: ...rning_Files/sp500-vpin-events-201004.csv

	Symbol	Date	Start time	Duration (sec)	Duration (bins)	Peak (CDF)	Peak (VPIN)	Peak (Date)	Peak (Time)	#Trades
9	AA	20100413	15:36:00	420	2	0.913072	0.549923	20100413	15:36:00	2890
10	AA	20100413	15:48:00	75360	41	0.977334	0.603772	20100414	9:50:00	12154
11	AAPL	20100405	10:09:00	7740	13	0.944538	0.513074	20100405	10:34:00	30759
12	AAPL	20100416	10:15:00	15060	35	0.987679	0.562543	20100416	12:36:00	77035
13	AAPL	20100421	9:35:00	13020	46	0.99688	0.59951	20100421	10:08:00	109766
14	AAPL	20100423	9:34:00	120	1	0.905718	0.491918	20100423	9:34:00	2344
15	AAPL	20100423	9:37:00	16140	45	0.994394	0.584445	20100423	10:04:00	107521
16	ABC	20100405	9:33:00	6780	9	0.958526	0.620778	20100405	9:44:00	3923
17	ABC	20100405	11:44:00	720	2	0.912949	0.59277	20100405	11:44:00	671
18	ABC	20100416	15:57:00	76380	35	0.997364	0.699721	20100419	9:36:00	745
19	ABC	20100419	13:23:00	1500	2	0.90655	0.58983	20100419	13:23:00	872

Plot Fields:
TIME
BID
OFR
BIDSIZE
OFRSIZE
MODE
EX
MMID

Query: trades/$DATE/$SYMBOL/TIME>50400 && trades/$DATE/$SYMBOL/TIME<54000

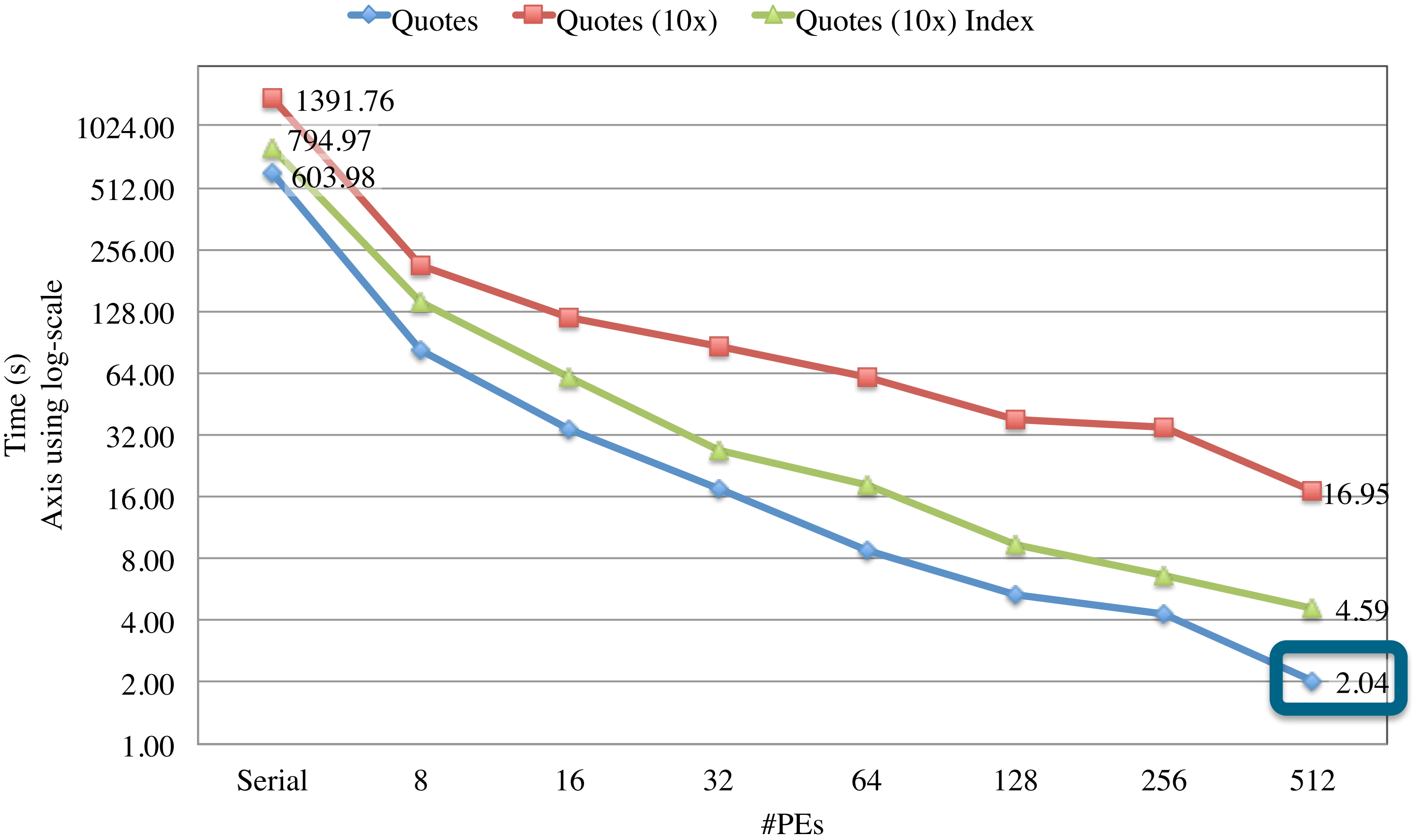
Stick ☐ Auto-color ☐ Antialiasing

Sum Bin Size: 10

Curve/Bar plots

Interactive Exploration with Bitmap Indexes

- Warnings can be used to compose queries on different levels of market data to seek confirming signals
- Significantly speed up query processing with FastBit bitmap indexes



Answer 8000 queries in 2 seconds

Predicting Market Events

- **Question:**
  - Can HPC resources effectively compute market indicators?

- **Candidate Market Indicators**:
  - **Volume-Synchronized Probability of Informed Trading (VPIN)** *[1]*: Measures imbalance between buy and sell activities in volume time.

$$VPIN = \frac{\left|V_{buy} - V_{sell}\right|}{V_{total}}$$

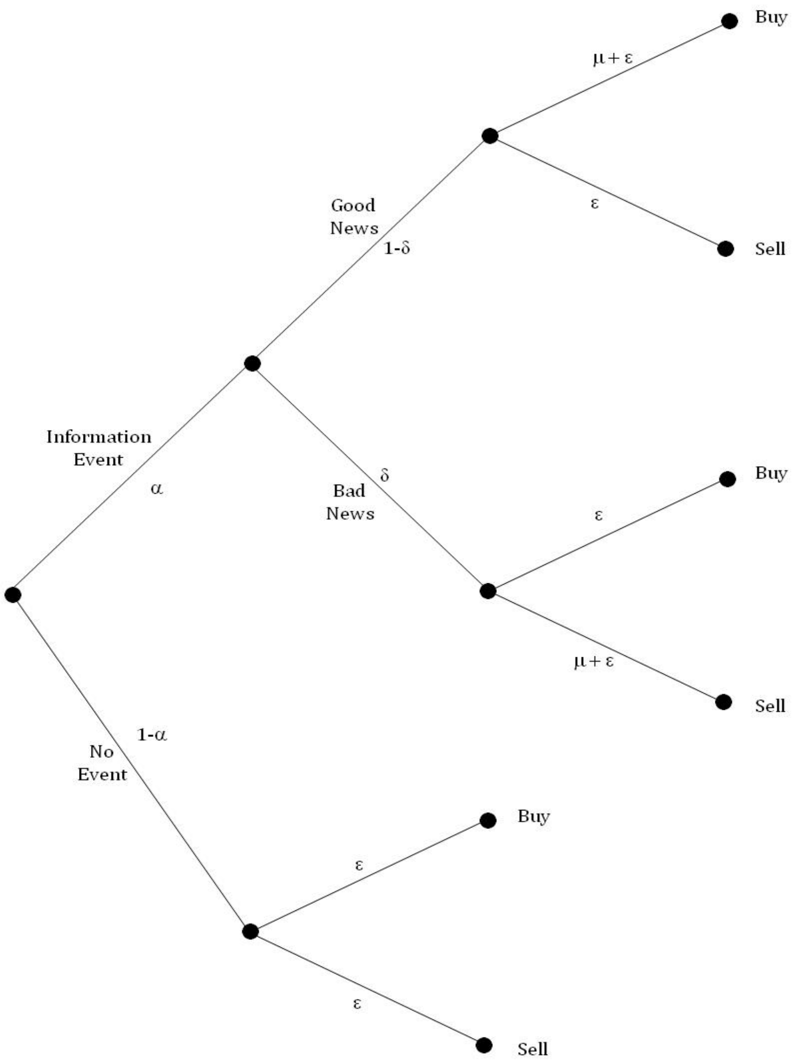
  - **Volume Herfindahl-Hirschman Index (HHI)** *[2]*: is a measure for the fragmentation of the market.

$$HHI = \left(\frac{V_{NYSE}}{V_{NYSE} + \ldots + V_{NASDAQ}}\right)^2 + \ldots + \left(\frac{V_{NASDAQ}}{V_{NYSE} + \ldots + V_{NASDAQ}}\right)^2$$

[1] D. Easley, M. M. Lopez de Prado, and M. O'Hara. Flow Toxicity and Liquidity in a High Frequency World. *Review of Financial Studies, Vol. 25, No. 5, pp. 1457-1493, 2012.* SSRN 1695596

[2] A. Madhavan. Exchange-traded funds, market structure and the flash crash. BlackRock, 2011. SSRN 1932925,

Theory: Probability of Informed Trading



$$E[S_i \mid t] = P_n(t)S_i^* + P_b(t)\underline{S}_i + P_g(t)\overline{S}_i$$

$$B(t) = E[S_i \mid t] - \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)}\left[E[S_i \mid t] - \underline{S}_i\right]$$

$$A(t) = E[S_i \mid t] + \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)}\left[\overline{S}_i - E[S_i \mid t]\right]$$

$$\Sigma(t) = \frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)}\left[\overline{S}_i - E[S_i \mid t]\right] + \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)}\left[E[S_i \mid t] - \underline{S}_i\right]$$

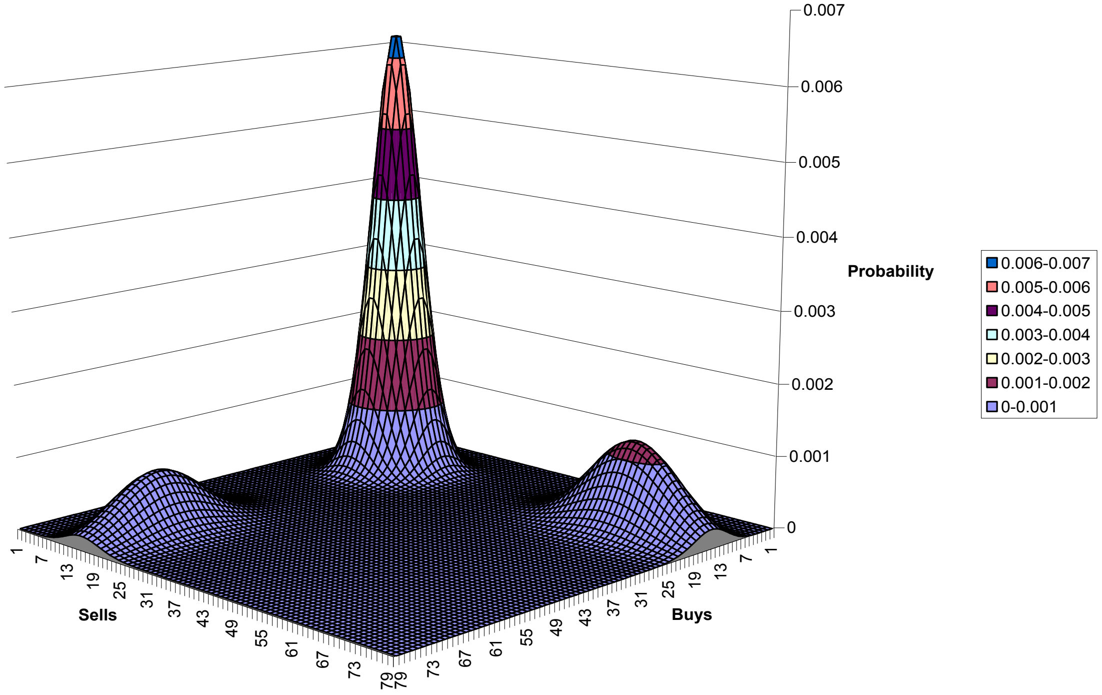
$$If\ \delta = \frac{1}{2} \Rightarrow \Sigma = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon}\left[\overline{S_i} - \underline{S_i}\right]$$

$$PIN = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon}$$

Source: M. Lopez-DePrado, Quant Congress 2011

How can PIN be estimated – Low Frequency

$$P(V^B, V^S) \\ = (1-\alpha)P(V^B,\varepsilon)P(V^S,\varepsilon) \\ + \alpha[\delta P(V^B,\varepsilon)P(V^S,\mu+\varepsilon) + (1-\delta)P(V^B,\mu+\varepsilon)P(V^S,\varepsilon)]$$



Source: M. Lopez-DePrado, Quant Congress 2011

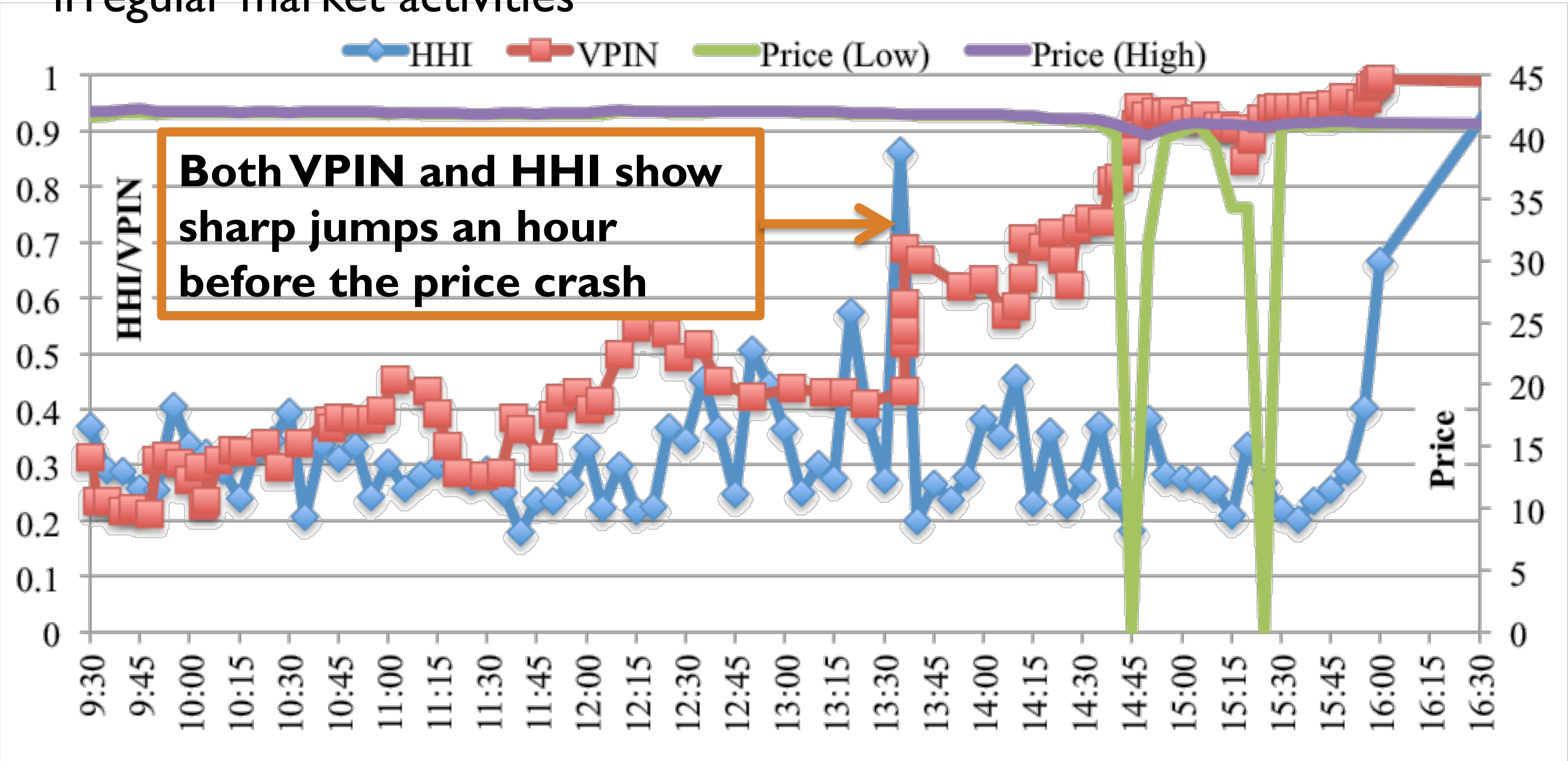
Which can be fitted for (α , δ , μ , ε) on low frequency data through ML (Easley, Kiefer, O'Hara, Paperman, 1996), EM (Kokot, 2004) or dynamically (Easley, Engle, O'Hara, Wu, 2008).

However, these procedures tend to be unstable when applied on high frequency data.

→ Solution: volume-time

VPIN Got Noticed!

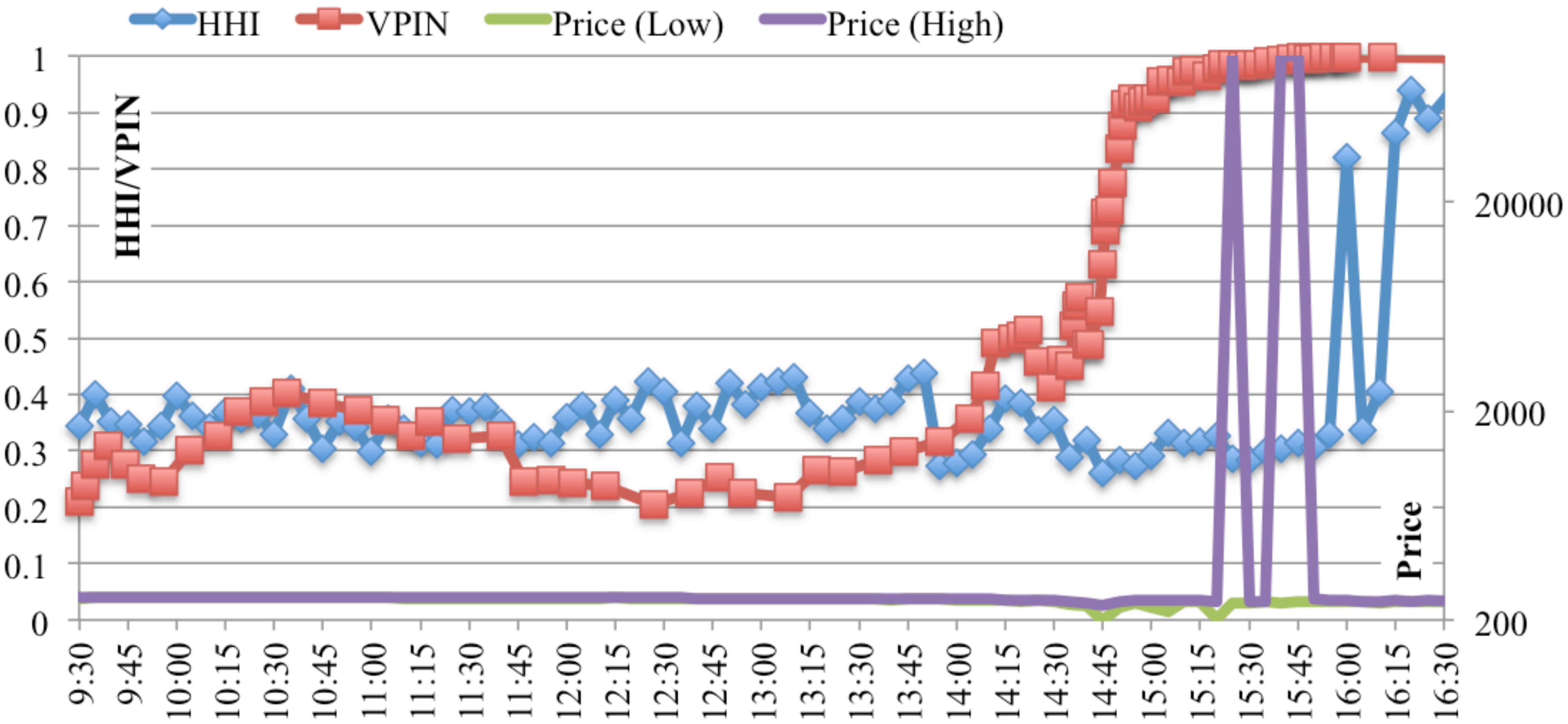
- VPIN (Volume Synchronized Probability of Informed Trading, Easley, de Prado and O'Hara 2011)
- HHI (Herfindahl-Hirschman Index for volume fragmentation, Madhavan 2011)
- Computed on Level 1 data, could raise warnings about upcoming irregular market activities

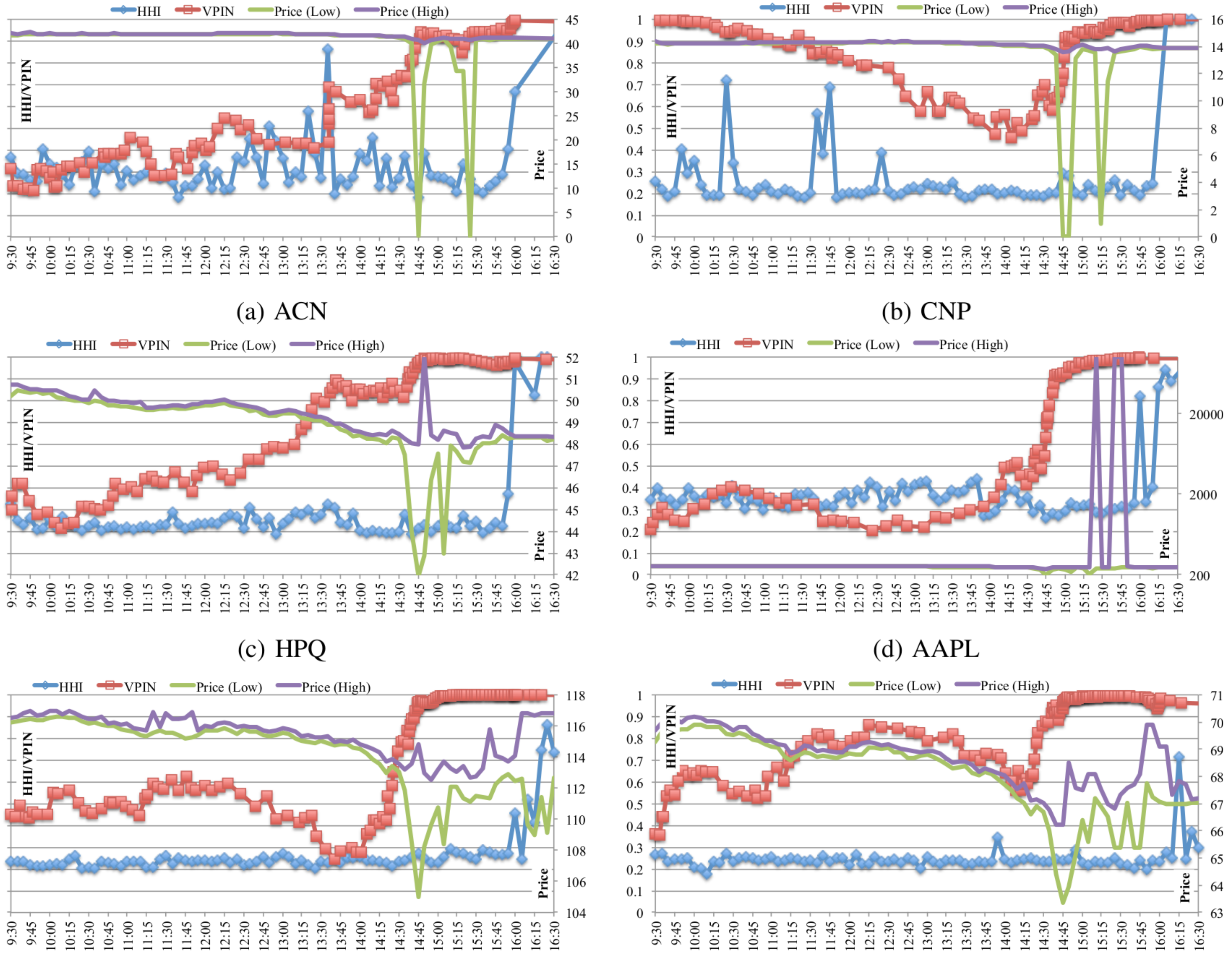


Both VPIN and HHI show sharp jumps an hour before the price crash

Another VPIN Example

- VPIN (in red) rises to a high level about 45 min before Apple share rise to $100,000





(a) ACN

(b) CNP

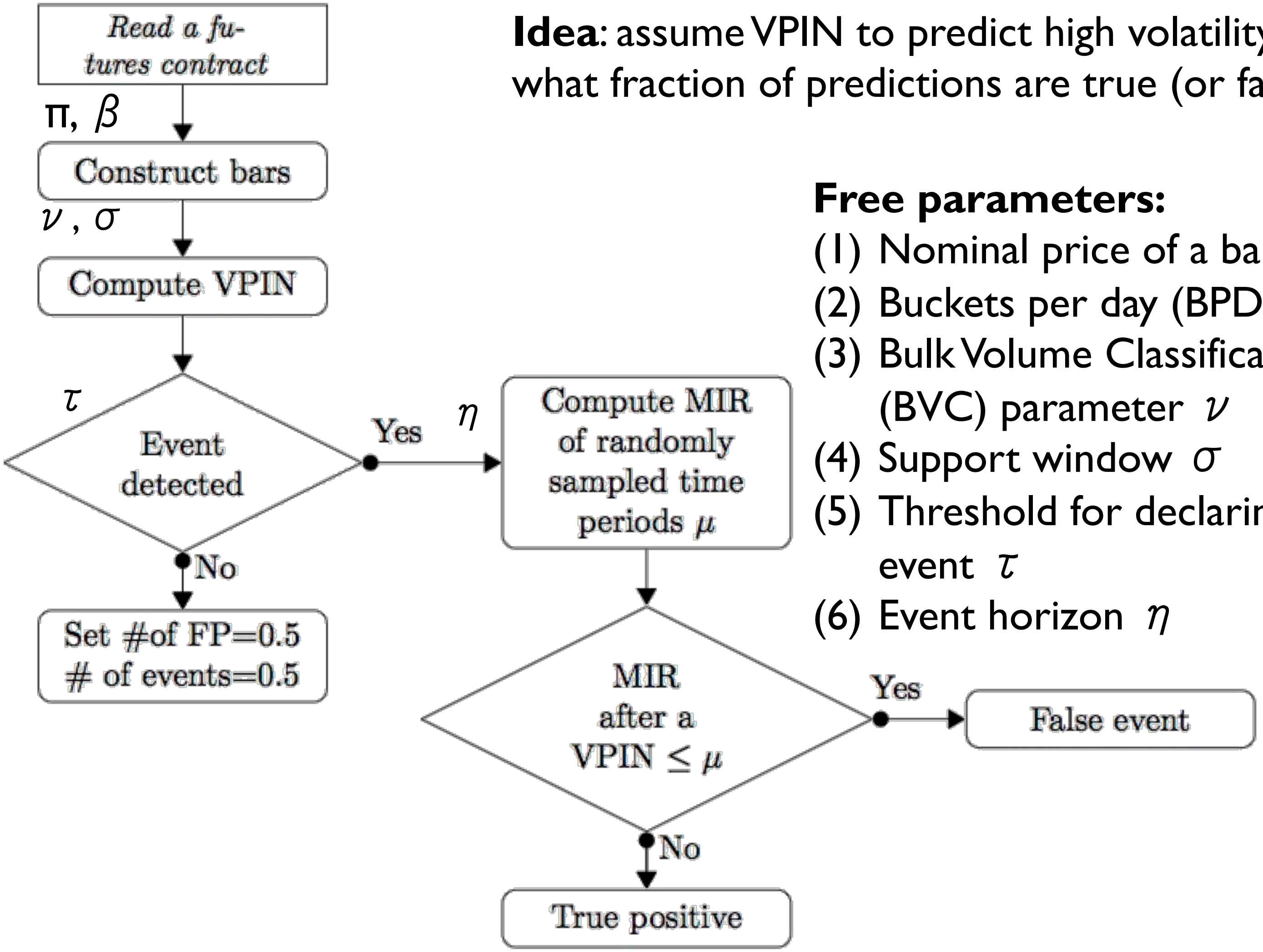
(c) HPQ

(d) AAPL

(e) SPY

(f) IWM

Quantifying Effectiveness of VPIN



Read a futures contract
π, β
Construct bars
ν, σ
Compute VPIN
τ
Event detected
Yes
η
Compute MIR of randomly sampled time periods μ
No
Set #of FP=0.5 # of events=0.5
MIR after a VPIN $\leq \mu$
Yes
False event
No
True positive

Idea: assume VPIN to predict high volatility events, what fraction of predictions are true (or false)?

Free parameters:

(1) Nominal price of a bar π

(2) Buckets per day (BPD) β

(3) Bulk Volume Classification (BVC) parameter ν

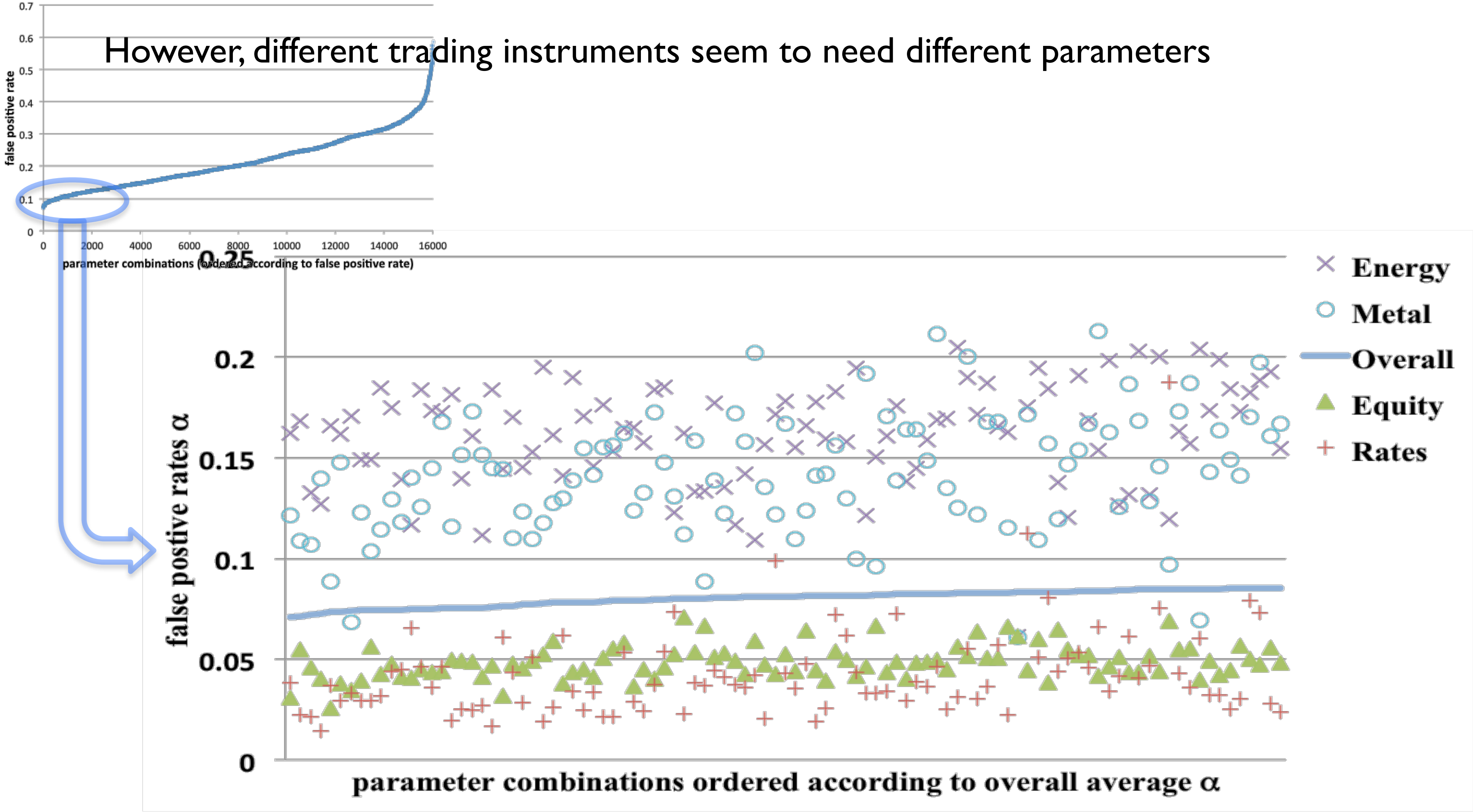
(4) Support window σ

(5) Threshold for declaring VPIN event τ

(6) Event horizon η

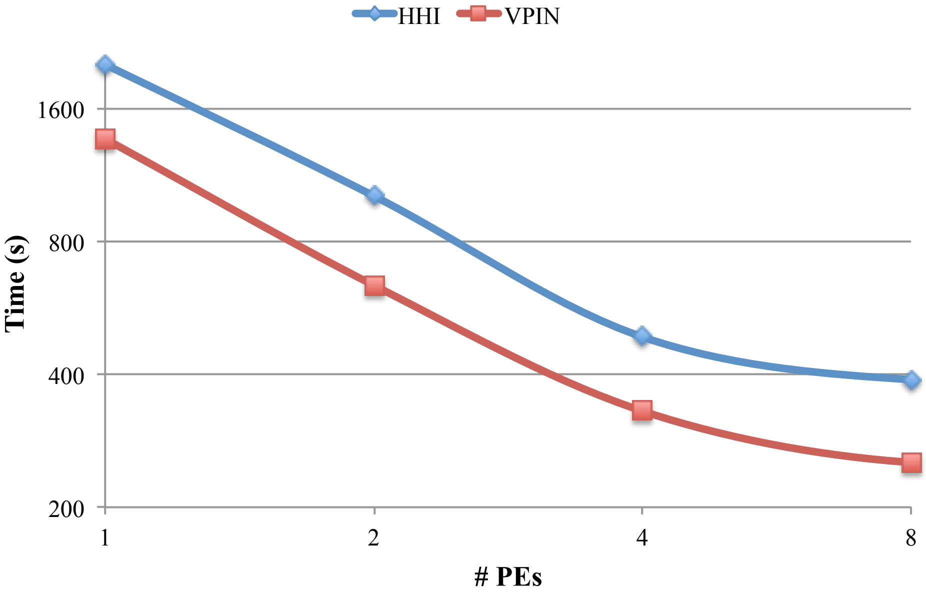
Lots of Parameter Values to Choose from

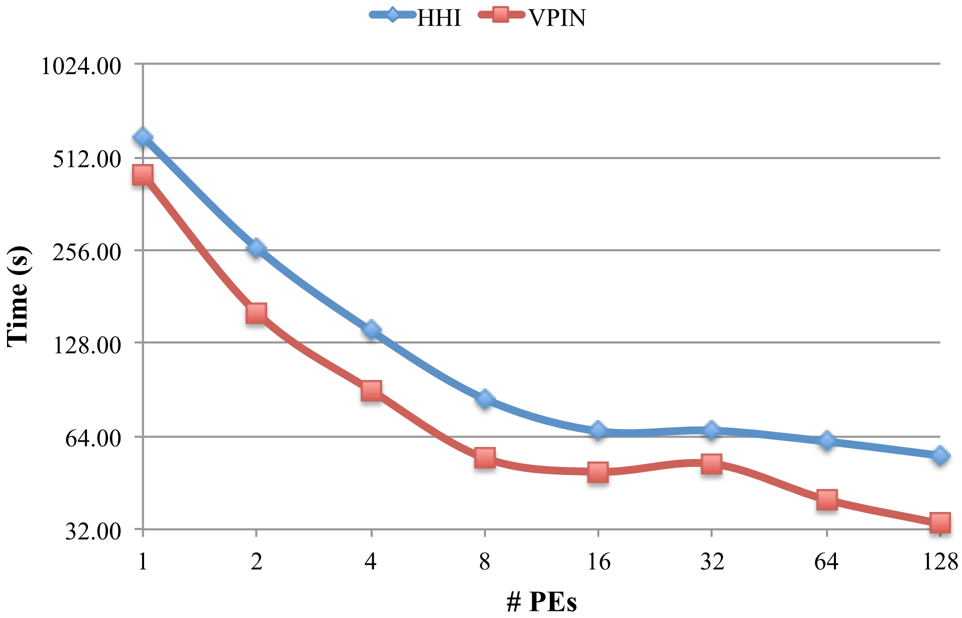
However, different trading instruments seem to need different parameters



How to Examine More Options: **Compute Faster**

- The procedure of computing VPIN and HHI can effectively take advantage of parallel machines
- The left figure shows the time needed to compute VPIN and HHI on 25 most frequently traded Electronically Traded Funds (ETF) using 10-year trades: 5 X speedup on 8 cores
- The right figure shows the time needed to compute VPIN and HHI on 500 stocks in SP500 using 3-year trades: 11 X (HHI), 13 X (VPIN) on 128 cores



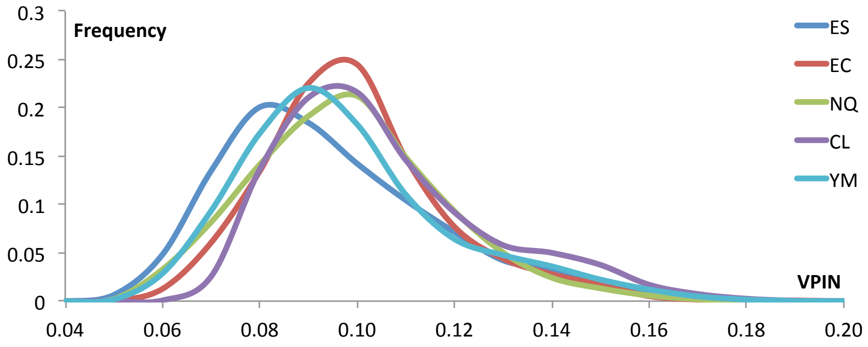


Faster Computation Leads to More Information about VPIN

Distribution of raw VPIN values

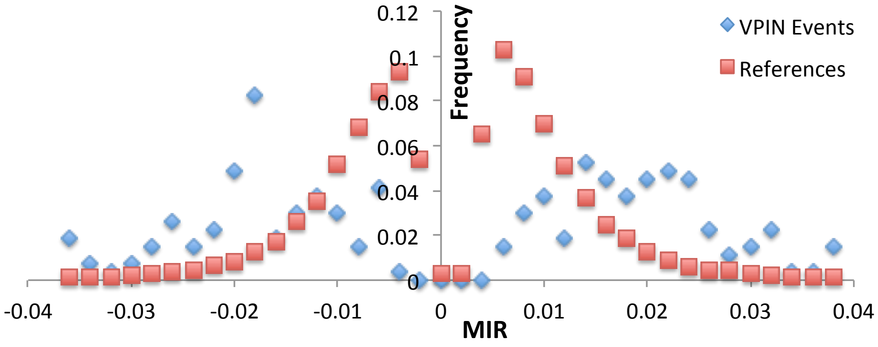
Since each trading instrument creates its own distribution, need a way to normalize the values.

→ VPIN values are normalized as a expected percentile (assuming the values are distributed normally)

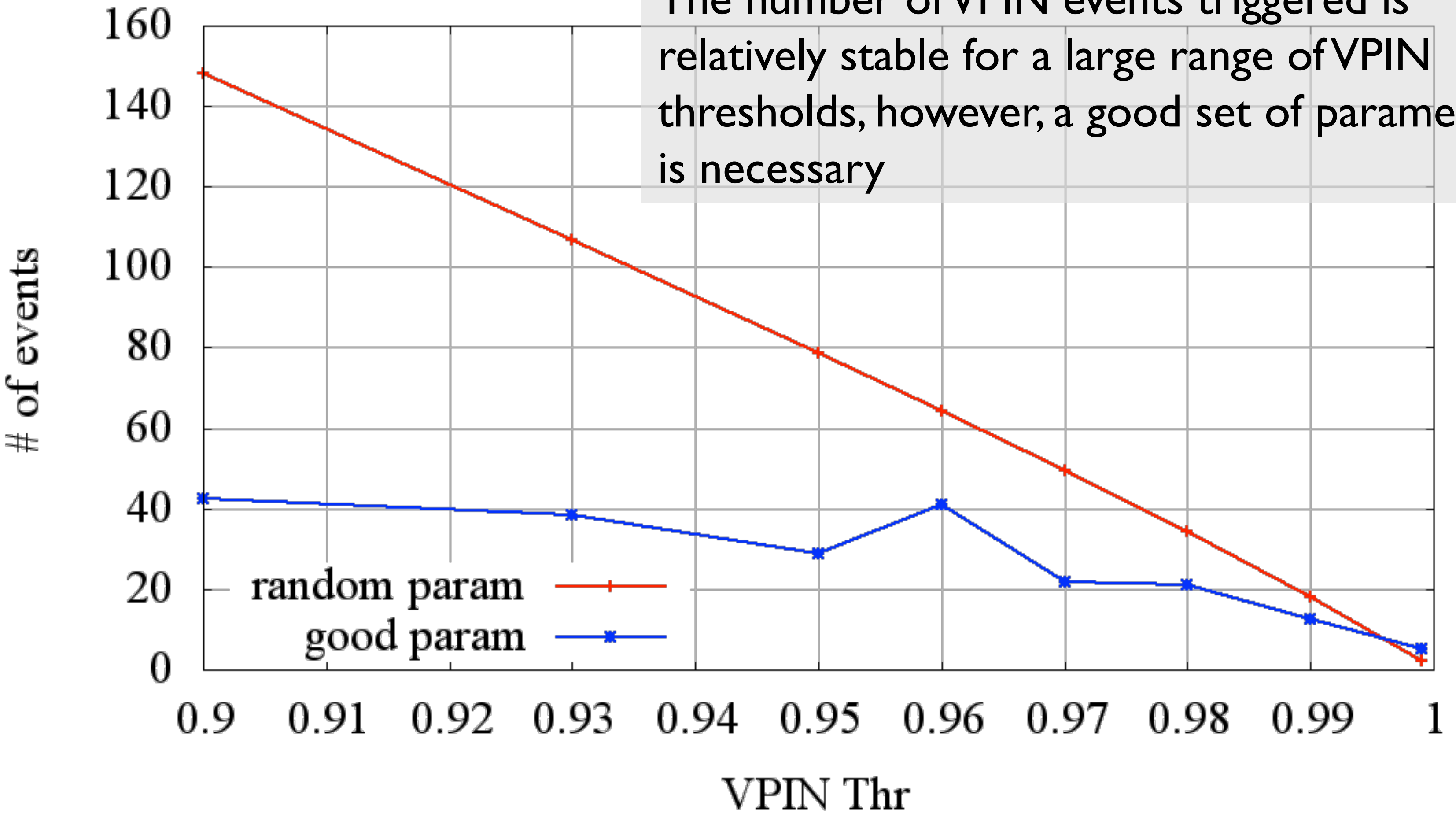


Maximum intermediate return (MIR) values on trading data follow different distribution than randomize sequences of the same values

→ Large (absolute) values in MIR indicate something special



Number of VPIN Events Stable



The number of VPIN events triggered is relatively stable for a large range of VPIN thresholds, however, a good set of parameters is necessary

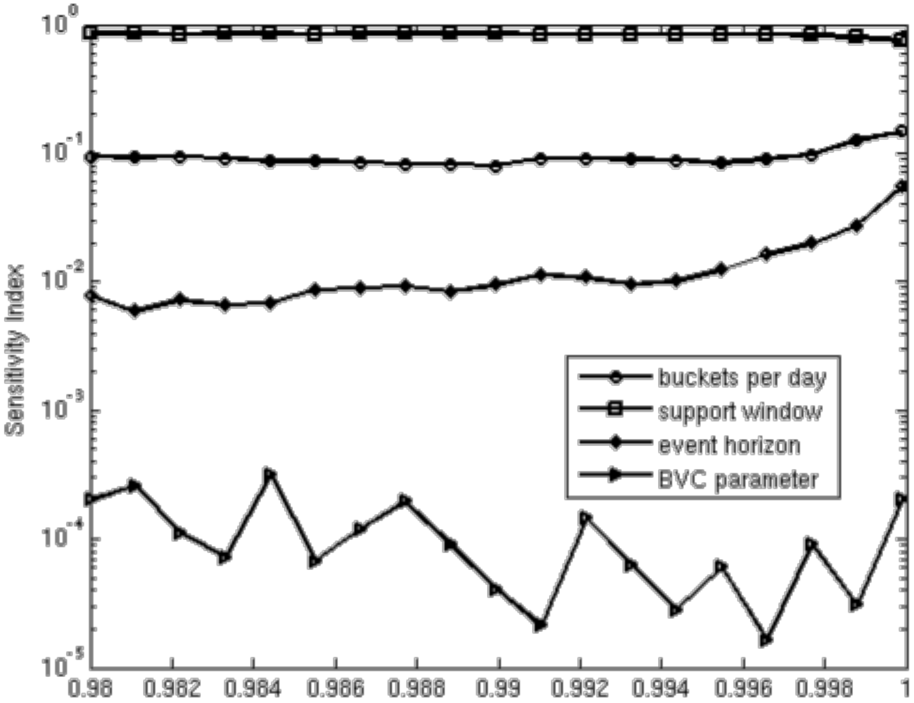
Which Parameters Are Important

Sensitivity analysis performed with UQTK (Uncertainty Quantification Toolkit)

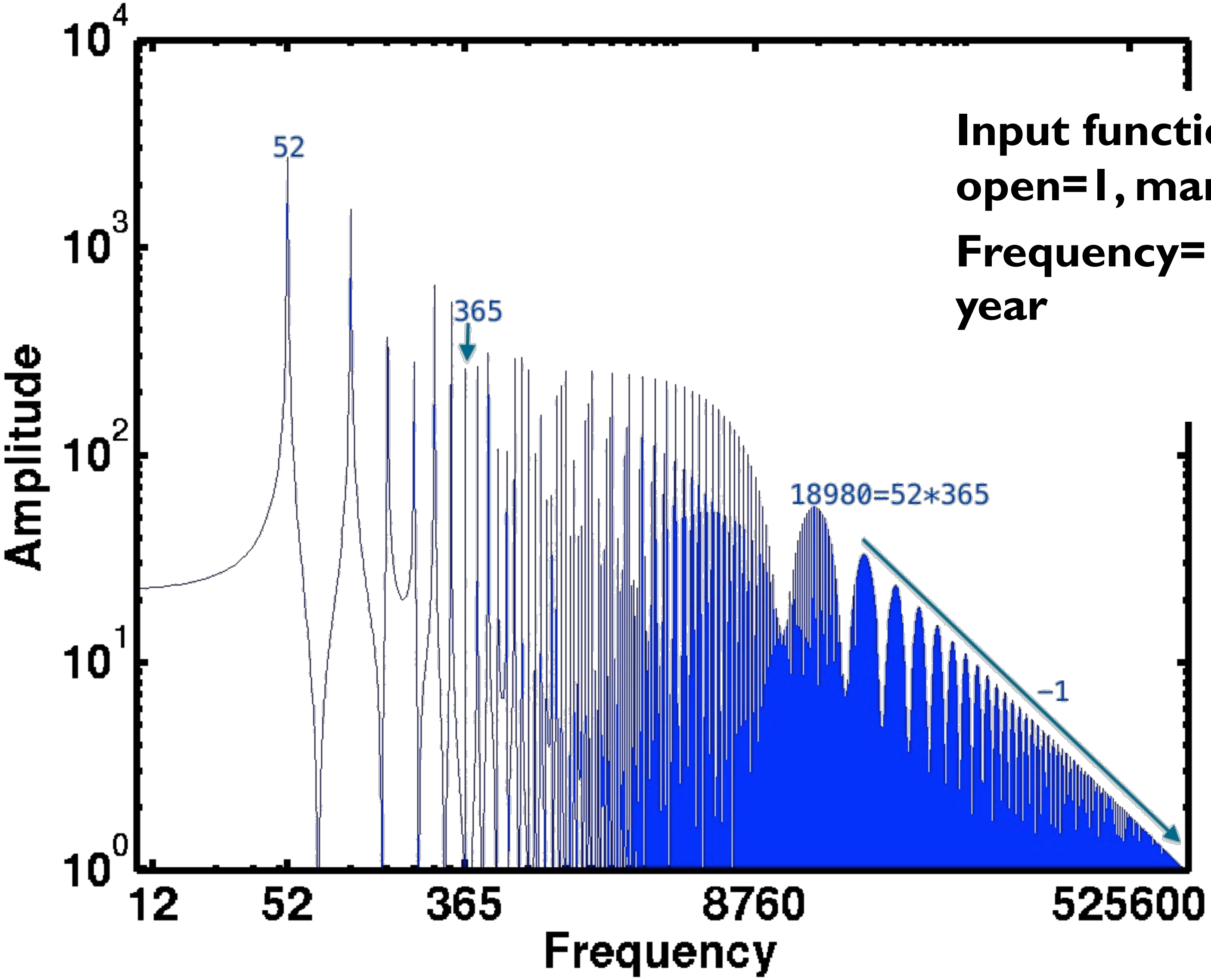
- Compute Sobol indices to measure the sensitivity of parameters using polynomial chaos expansion
- C++ implementation by Debusschere, Najm, Pébay, Knio, Ghanem, and Le Maître [2004]

Given a VPIN threshold,

- ~ 90% of variance explained by buckets per day β
- ~ 10% of variance explained by support window size σ



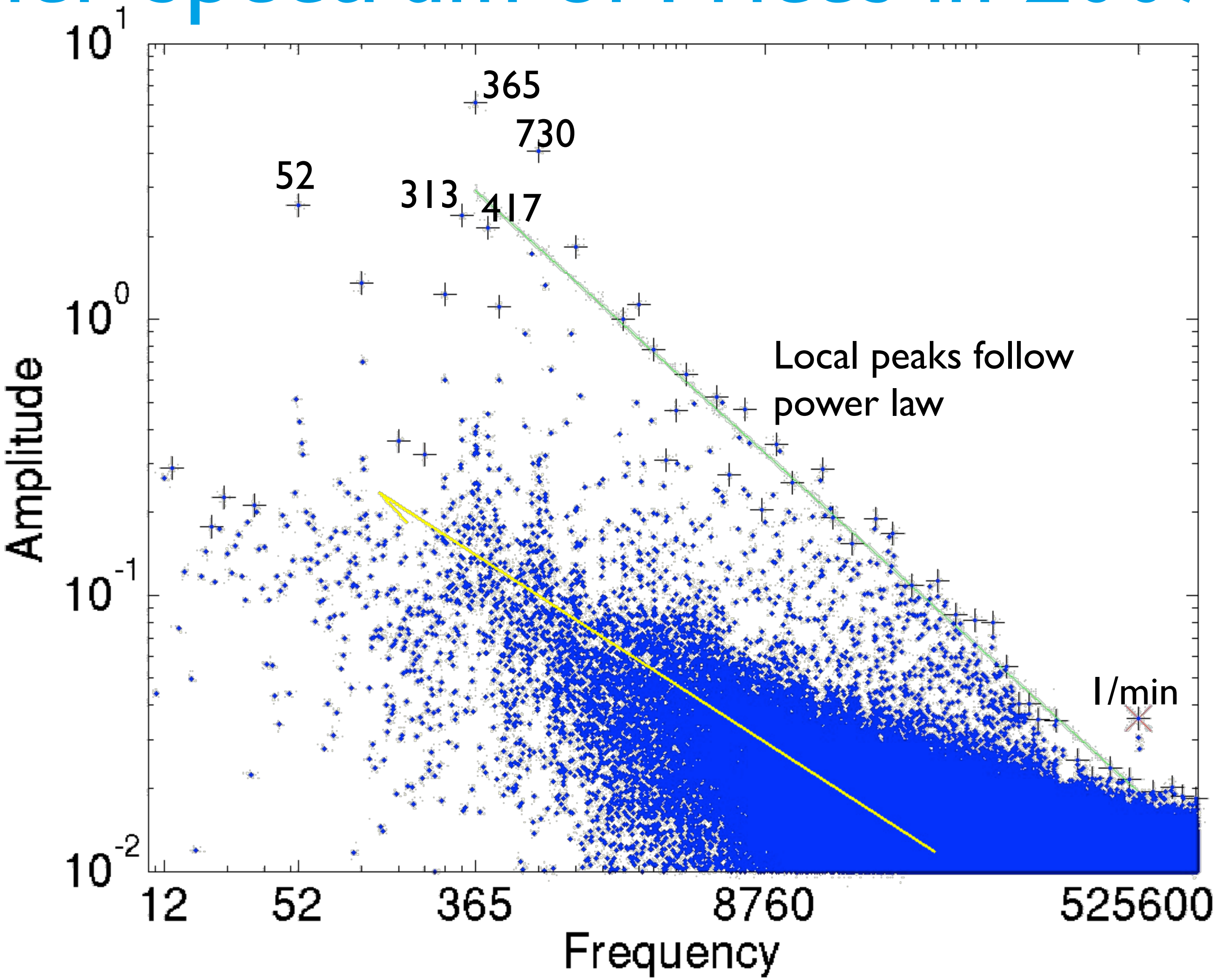
Another Tool for High-Frequency Data -- Fourier Analysis



Input function: market open=1, market closed=0

Frequency=1, once per year

Fourier Spectrum of Prices in 2007

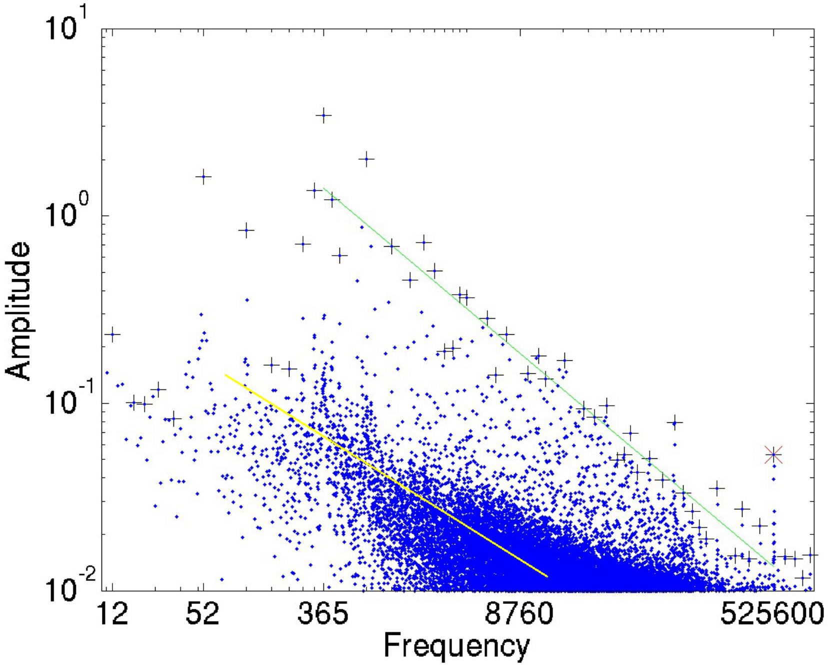


Strongest amplitude at frequency of 365, once per day

Fourier Spectra of Prices

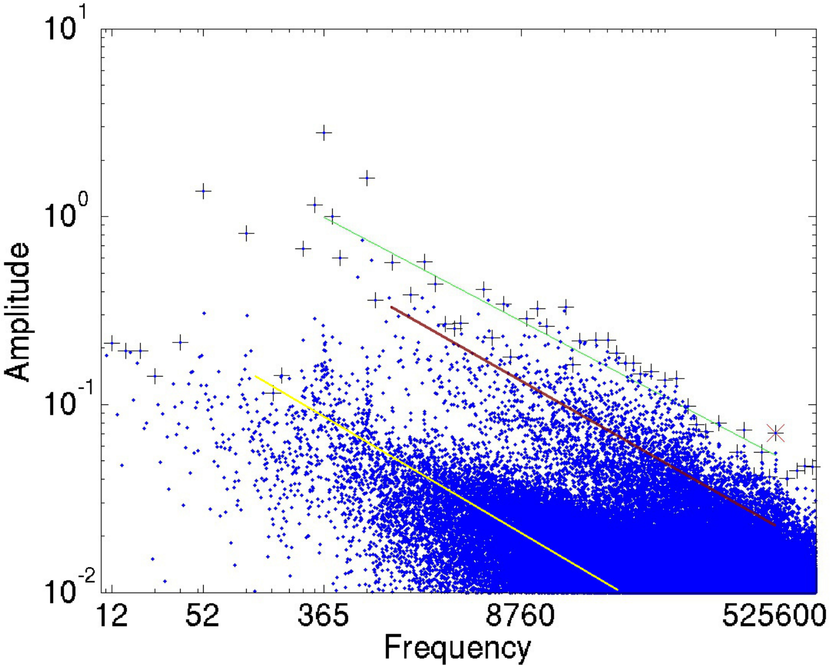
2010

Frequencies with highest amplitudes: 365, 730, 52, 313, 417



2013

Frequencies with highest amplitudes: 365, 730, 52, 313, 417



Fourier Spectra of Trading Volumes

2010

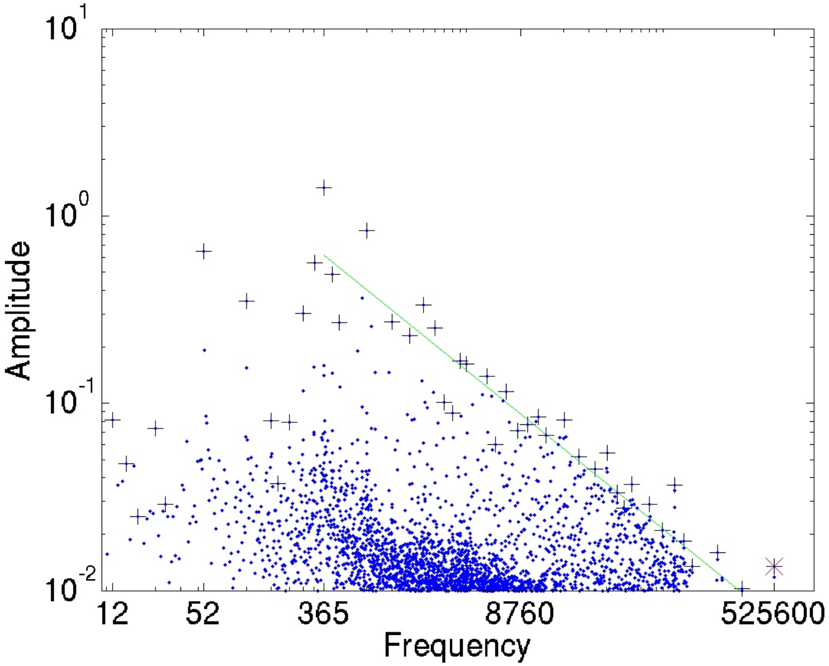
Frequencies with highest amplitudes: 365, 730, 52, 313, 417

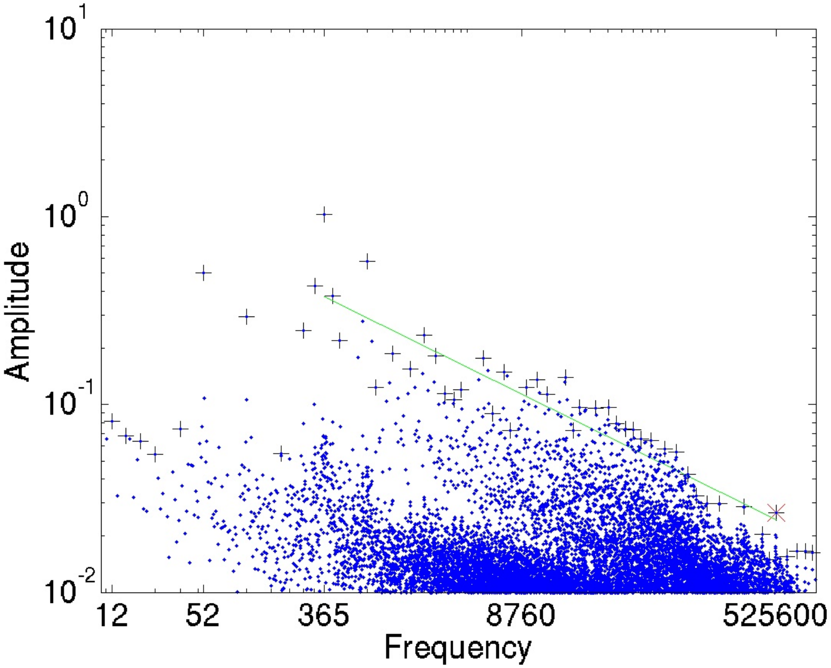
Same as spectrum of prices

2013

Frequencies with highest amplitudes: 365, 730, 52, 313, 417

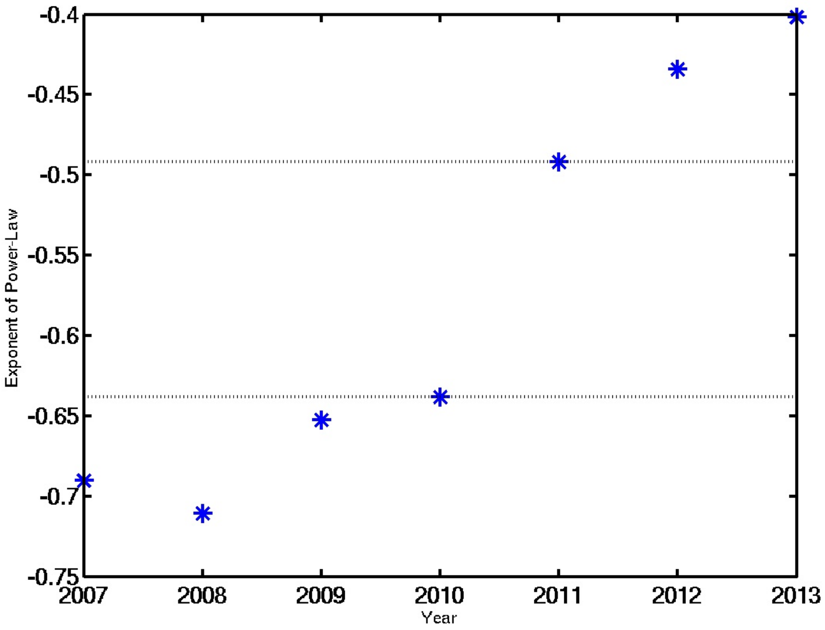
Same as spectrum of prices





Fourier Spectra of Prices

Exponents of the power law distribution of the local peaks separate into two groups: three recent years have large values



The frequency at 1/min has much higher amplitude than nearby frequencies: relative strengths are more pronounced in the five recent years

Year	Frequency	Rel Strength
2007	525600	6.7
2008	527040	5.1
2009	525600	13.7
2010	525600	20.3
2011	525600	15.6
2012	527040	15.7
2013	525600	15.4

Summary and Future Work

- Scientific data format HDF5 is shown to be more effective than CSV
- Early-warning indicators can be found, even on simple "Low Level" data -- investigated VPIN and HHI
- Computations can be parallelized to take advantage of high-performance computers
- Ultimate goal is to develop an early warning system that can be the basis of a "yellow flag" to augment the current circuit breaker for financial market

Additional Information

- Papers: DOI:10.1145/2088256.2088267, DOI: 10.3233/AF-13030
- Author emails
  - Wes Bethel EWBethel@lbl.gov
  - Ming Gu MGu@berkeley.edu
  - David Leinweber DLeinweber@lbl.gov
  - Oliver Ruebel ORuebel@lbl.gov
  - K. "John" Wu KWu@lbl.gov

- Computational Intelligence and Forecasting Technologies http://crd.lbl.gov/cift
- Scientific Data Management research group http://crd.lbl.gov/sdm/