Linking Many Unusual Co-Incidences

Kevin B. Pratt, Chief Scientist ZZAlpha Ltd.

Kevin.pratt@zzalpha.com

Presented to IEEE Big Data Conference

Boston, MA, USA Dec. 2017

ZZAlpha LTD.

Consistent returns from machine learning

The inspirational puppy

Agenda

1. Hypothesis and analogy to “puppy learning”
2. Learning context and goal of stock predictions
3. Innovations

   For algorithm please see paper.
4. Process diagrams
5. Visualizations – bi-partite and swim-lane
6. Clustering of unusual events
7. Results – Monte Carlo and Benchmark

Hypothesis:

a) In the complex US economy and markets, there exist discoverable, transient *sets* of prior *unusual events* that can link to subsequent events in that uncertain environment.

b) Those *links* can be used effectively to predict subsequent events (profitable stock price changes) in a time-lagged re-enforcement learning system. Those links can be distinguished from *spurious "mere-co-incidences."*

Agnostic: we begin with *no model* of how or whether prior and subsequent events may be related.

We assert this analogy:

The hypothesis is similar to the way a new puppy learns on its own in a new and uncertain environment
(but a much less complex environment).

Notions of “puppy learning”:

1. A puppy cannot remember everything, so in the stream of life events it stores in medium term memory what is “**unusual**.” It also discards stale information.

2. A very rare event is not frequent enough to be very useful and is ignored unless it is reinforced enough (to become “unusual” instead of rare).

3. An event is not a trustworthy predictor (a “**link**”) unless the subsequent occurs much of the time when the predicate unusual event occurs .

4. When there is some **set** of trustworthy links, a subsequent becomes more likely to occur in some future time window.

5. It is especially worthwhile for a puppy to be alert for and remember events when the subsequent is an unusual reward or penalty.

Implicit learning hypothesis: Intelligent animals have evolved to have this "puppy learning" as a baseline learning mechanism to solve problems from birth in a changing environment (*without human directed training*).

Explicit hypothesis: This "puppy learning" can be emulated in a computer to offer a simplified, fast, and effective machine learning step-forward method than can be applied to uncertain environments with many things going on concurrently where the data does not satisfy common statistical machine learning requirements.

Time Series Prediction Context:

Task: Each day, identify 5 large cap stocks that will go up in price significantly over the next month.

(Def: Large capitalization stocks are the 100 largest US stocks.)

Data: 5000 time series of events derived from daily stock price and economic indicator motifs for 11 yrs

Measure of success: Significantly beat Monte Carlo simulation and objective stock market benchmark.

Evaluation by rolling and compounding results described later

Why is stock prediction difficult?

Nothing is stationary, Gaussian, or transparent and:

Price movements implemented by bots, committees, and individuals

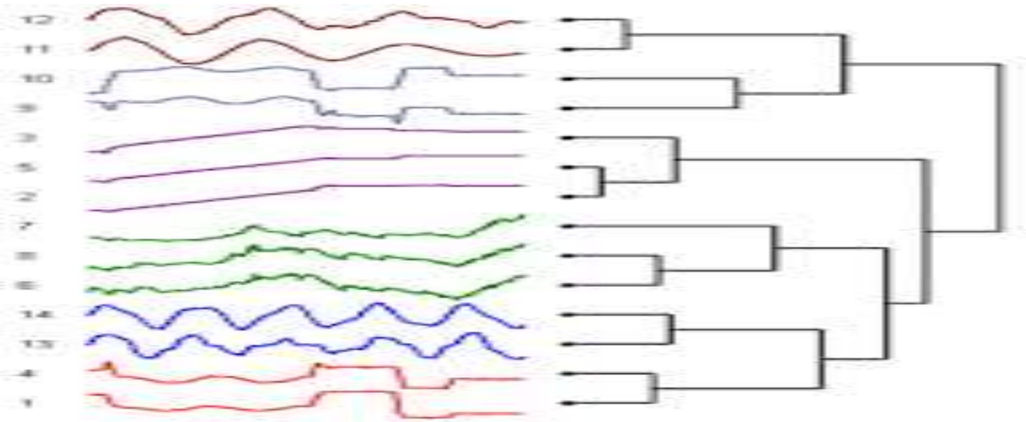
Data is often dirty and delayed in real-time

News impacts stock price **after** prediction (during the hold period)

75% of trades result from algorithmic trading systems
50% of trades result from high frequency trading (HFT)
54% of assets held by institutions
3% of assets held by hedge funds
34% of assets held by individuals/families
10% of assets held by discretionary investors
Sources: Bloomberg,JPMorgan 2013

Predictions for illiquid stocks have illusory value because actual trading opportunity is limited

Typical time series analysis research



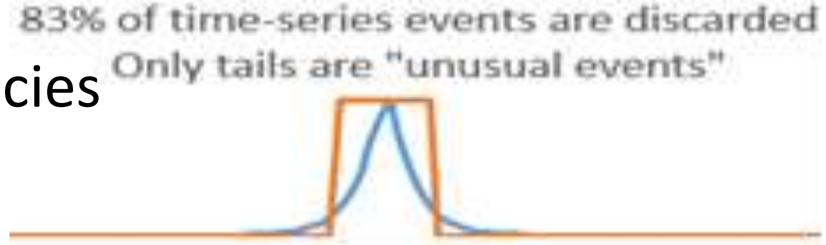
Locate anomalies within a series

Associate similar series (cluster)

Predict next value(s) in a series using that series (or multi-variable series for the same phenomenon)

Big data cross-series analysis

Innovations:

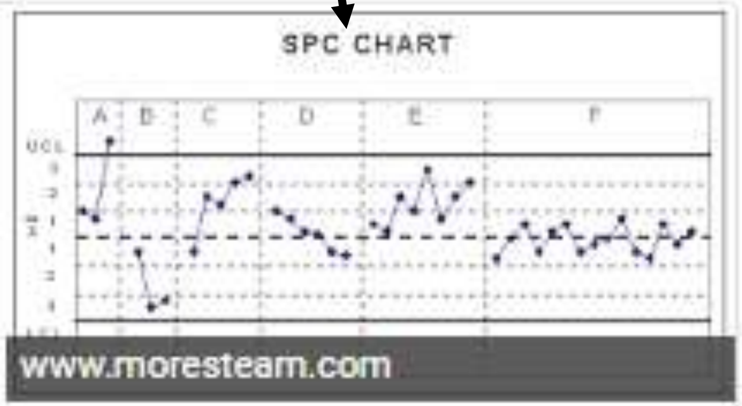


1. 5000 series of **different phenomena** with co-dependencies varying, unknown and with non-stationarities.
2. Utilization of **tails** and discard of 83% of data as weak information value.
3. Sample collected from **across** all the series at each time-step (values quantized)
4. Prediction of **delayed response** (20 time-steps=month) is robust in presence of unknown and varying noise
5. Prediction of **multiple target variables** (100) from varying subsets of a single cross-series sample.
6. Visualizations of prediction **evolution**.

“Unusual event” - examples of motifs extracted from raw time series:

1. A measurement exceeds a threshold control limit in a time series e.g. SPC chart
2. A time series has peaks (local maxima)
3. Management changes
4. Analysts upgrade/downgrade
5. Infrequent SAX word, wavelet, or other motif
6. Z-score thresholds
7. Top/bottom decile
8. Unusualness after transform e.g. to frequency domain
9. High pagerank or other cluster sink

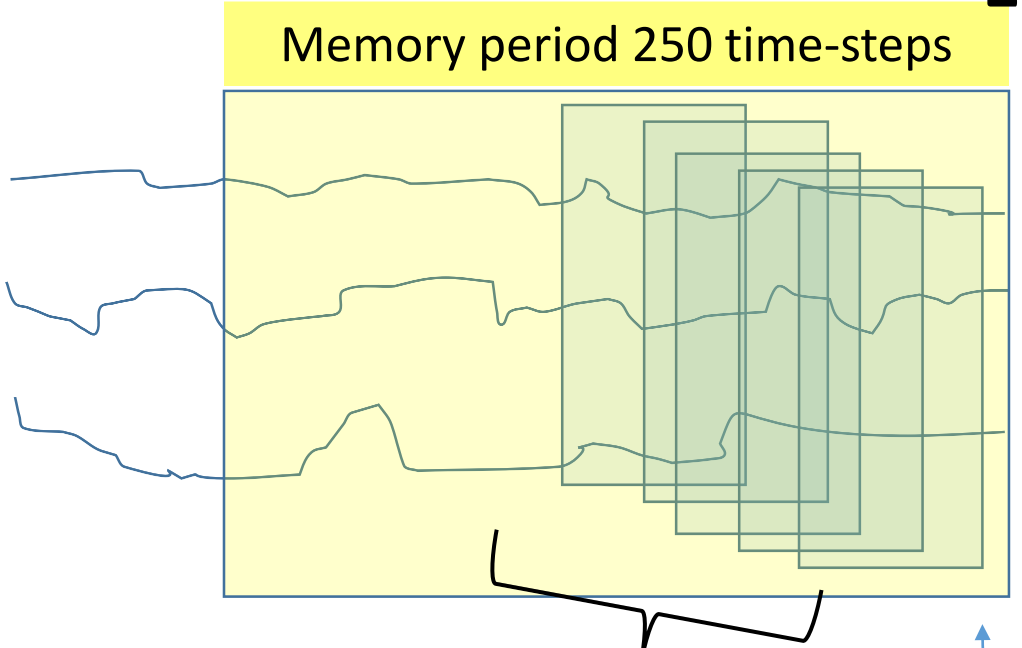
The control chart is a graph used to study how a process changes over time. Data are plotted in time order. A control chart always has a central **line** for the **average**, an upper **line** for the upper control limit and a lower **line** for the lower control limit. These lines are determined from historical data.



Control Chart - Statistical Process Control Charts | ASQ
asq.org/learn-about-quality/data-collection-analysis-tools/.../control-chart.html

Learning and prediction process

Unusual events extracted from Time Series
1,
2,
…
5000



Prior recent time windows for training

Recent time window at time-step t

20 day result delay with noise

Target variables
1,
2,
…
100

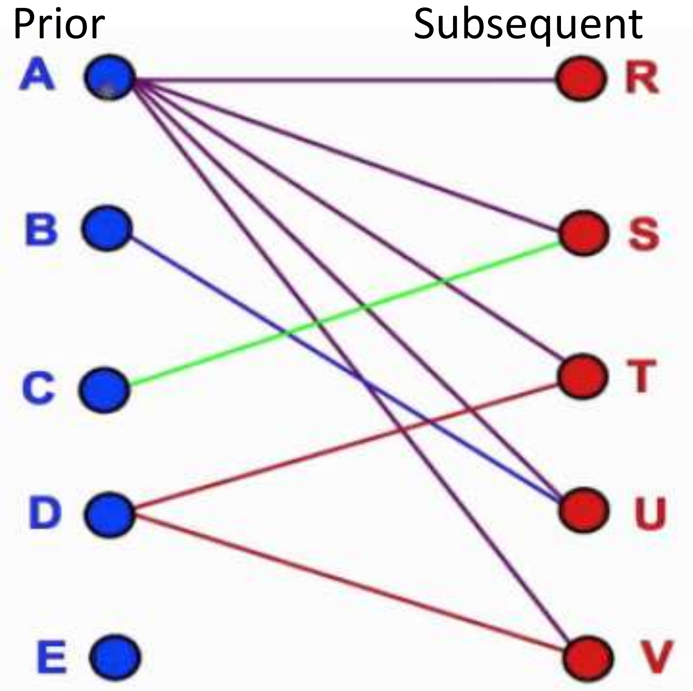
Re-enforcement results become available for time-step t-20

Prediction checklist:
Event unusual ?
Event is durable (within its own series) ?
Link to target is durable?
Aggregate link sets values for targets

Visualizations

Needed a **bi-partite** graph that

- Shows 5,000 nodes on the left
- Shows links to 100+ nodes on the right
- Displays at least three types of meta-data about the node
- Allows the user to “drill into” the nodes and links individually
- Can be animated for view of evolution of the durable nodes and links



Standard bi-partite graph tools did not scale and were not interactive.

A bi-partite graph

5000 left x 100 right nodes

(no links shown)

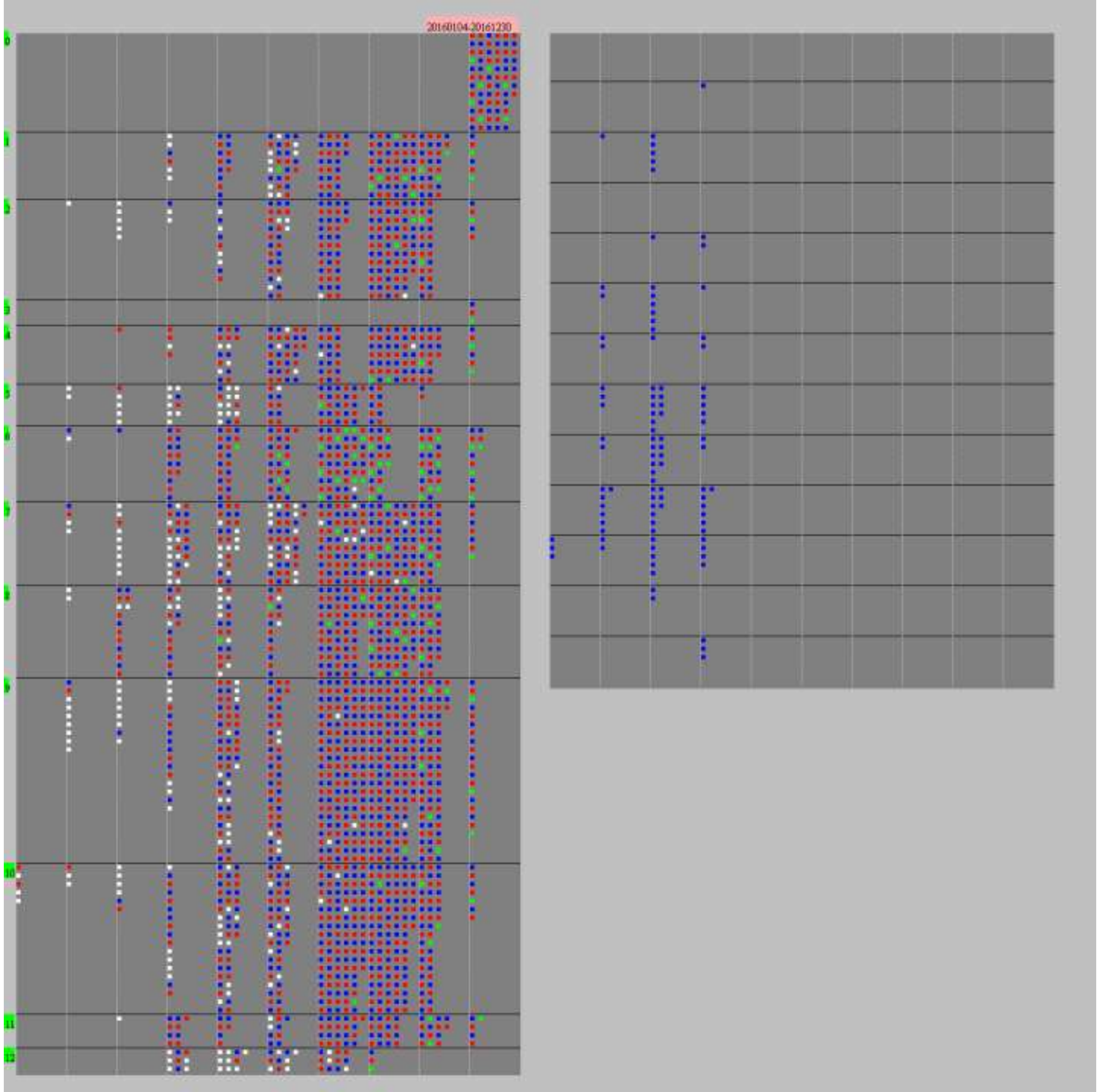
Left= all unusual event types occurring in 2016

Right= unusual price rises (in the largest 100 stocks) in 2016

Node color = an event type group (e.g. “peak”)

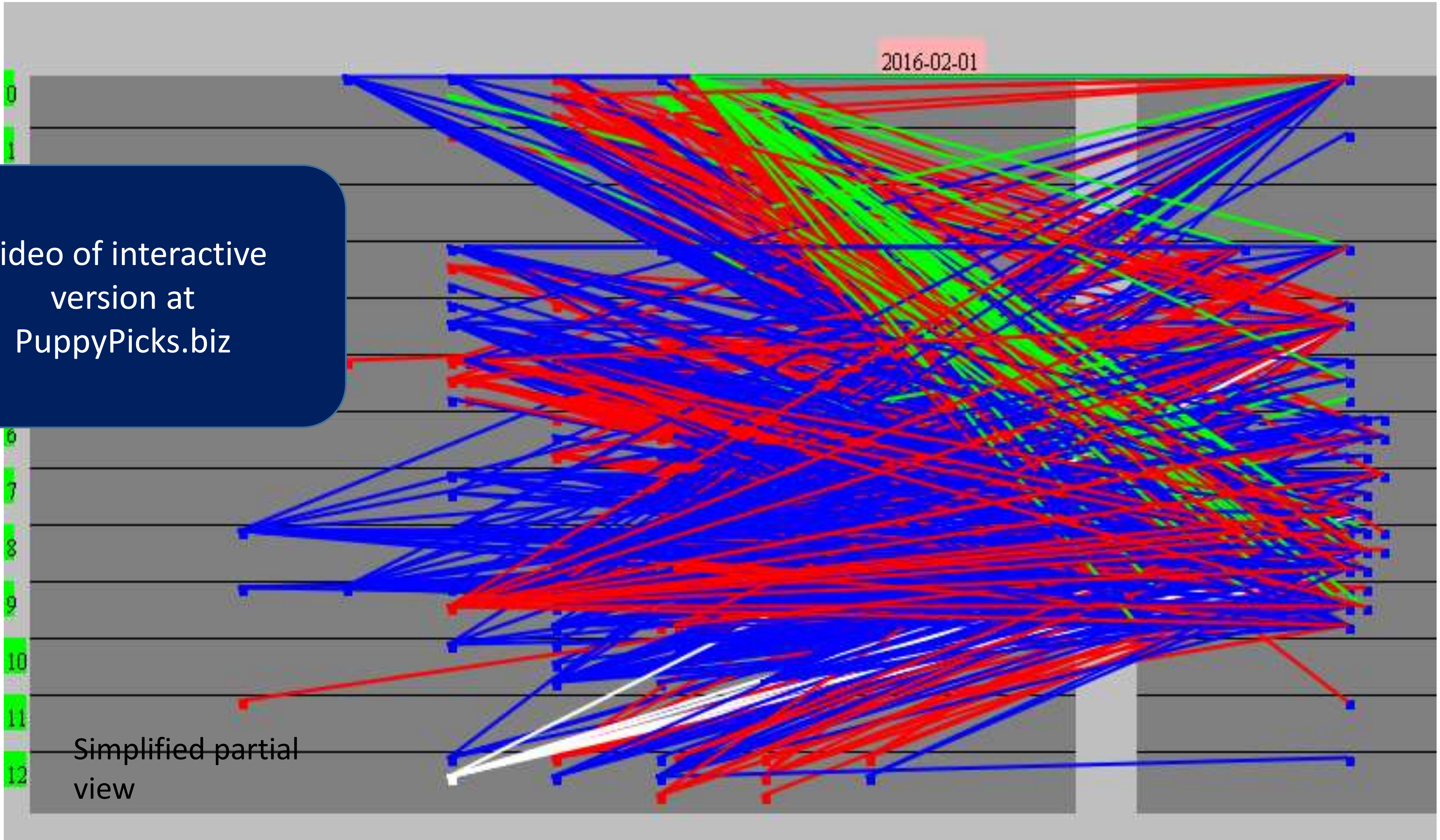
Horizontal groups= economic sectors

Vertical groups = other metadata groups



Single day: linkages of unusual events in the US economy to opportunities

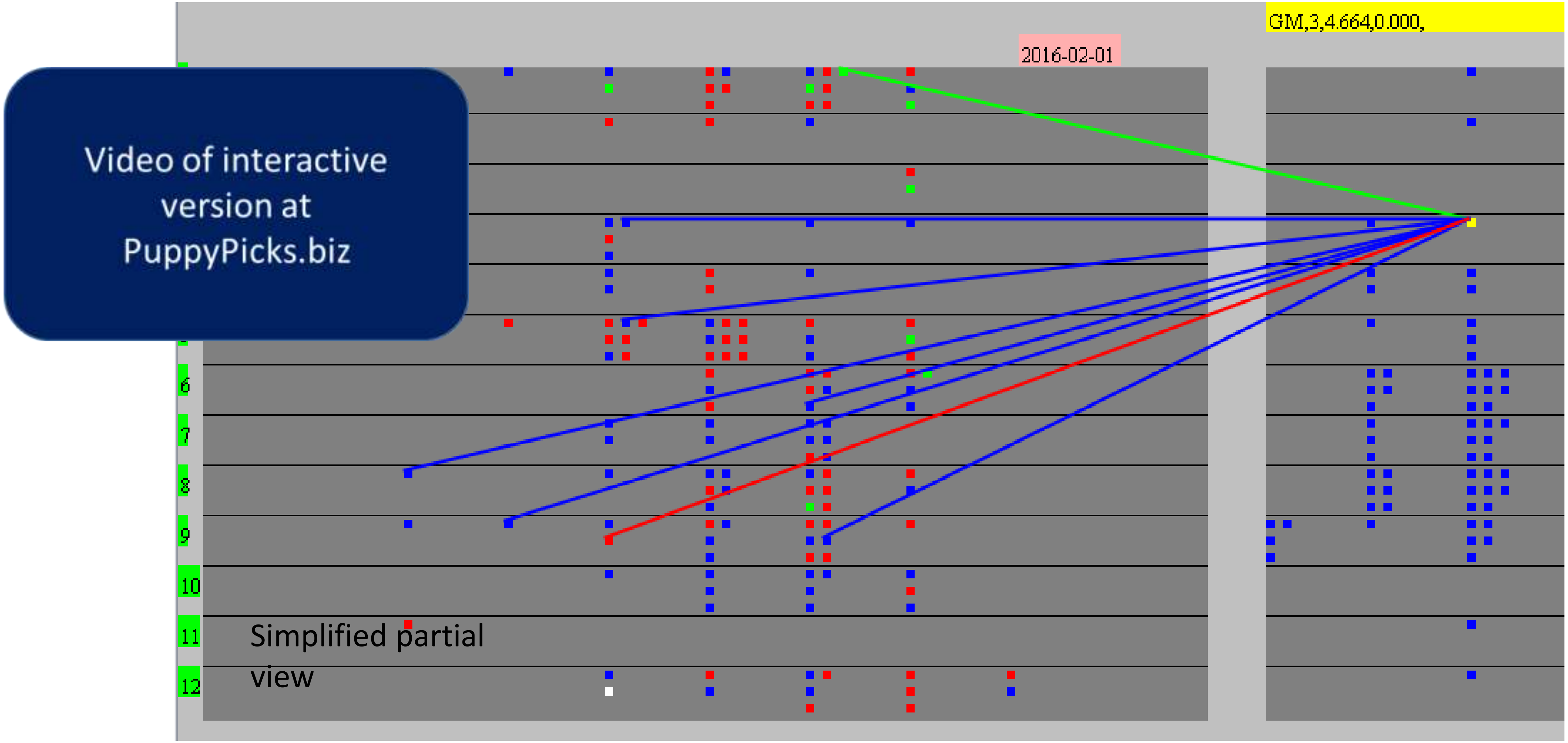
Many to many transient relationships even after many of weak links have been removed.



Video of interactive version at PuppyPicks.biz

Simplified partial view

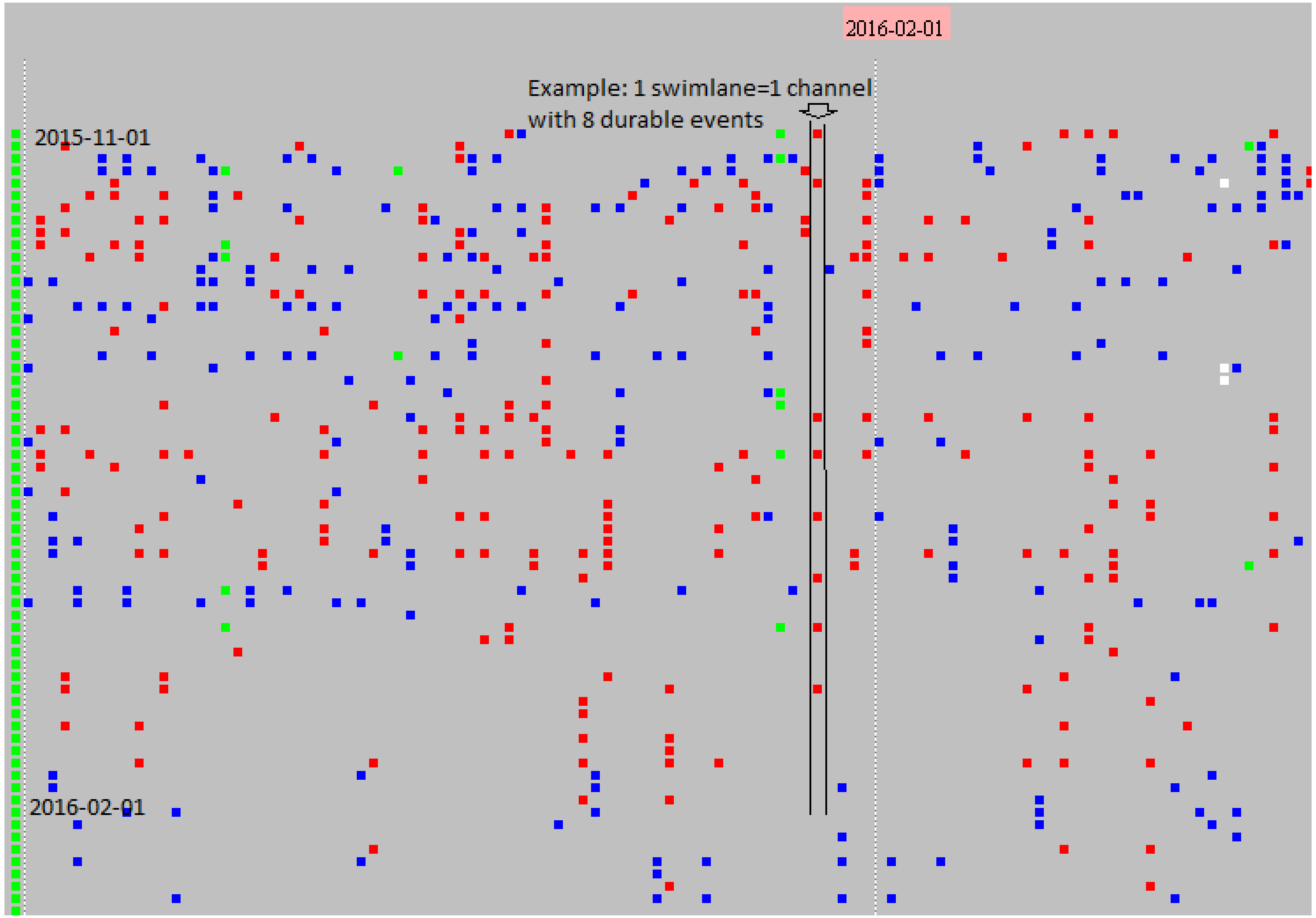
Interactive and animated: User clicks on historic stock price rise event on right (GM) to see the predicate unusual events (that existed that specific day)



Interactive "biopsy" of pathological durable events:

Swim lane view of unusual events leading up to current day

Partial view of interactive, scrollable large graph



Columns are event types ("channels")

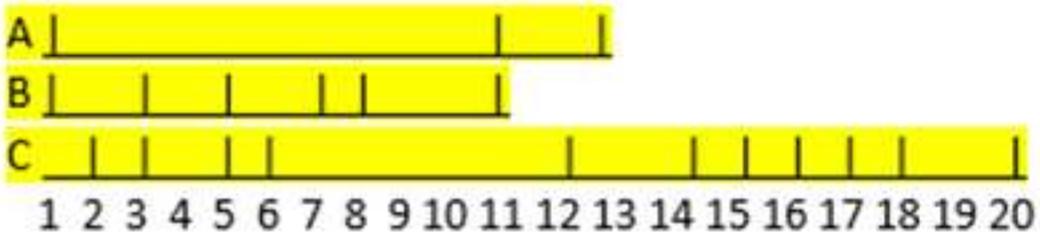
Most current day at bottom row

Similarity and clustering

We define:

Similarity= likeness - unlikeness/(number of possible occurrences - likeness)

where *likeness* = count of intersections of the two bitstrings and *unlikeness* = count of XOR of the two bitstrings.



Bitstrings aligned by semantics of bits

A,B,C are channels. Marks indicate events.

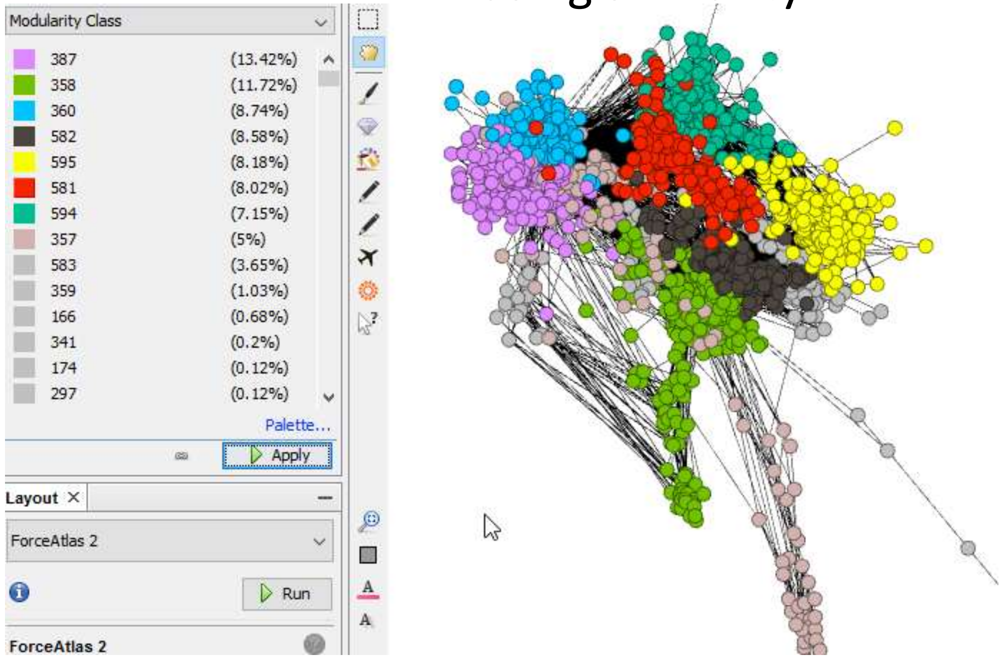
A:B Ham: 5 Jac: 2/(2+ 5)=0.29 Sim: 2- 5/(13-2)= 1.55

B:C Ham: 13 Jac: 2/(2+13)=0.13 Sim: 2- 13/(20-2)= 1.28

A:C Ham: 14 Jac: 0 =0.00 Sim: 0-14/(20-0)= -0.30

Fig. 7 Example of similarity measure compared to Hamming and Jaccard. In our similarity the integer portion encodes likeness and the fractional portion encodes unlikeness. It preserves ordering on likeness.

Gephi – application of Blondel-Louvain unsupervised modularity clustering using similarity



Predicate sets of unusual events cluster over time.

More recent research shows clusters align to results (levels of stock price change).

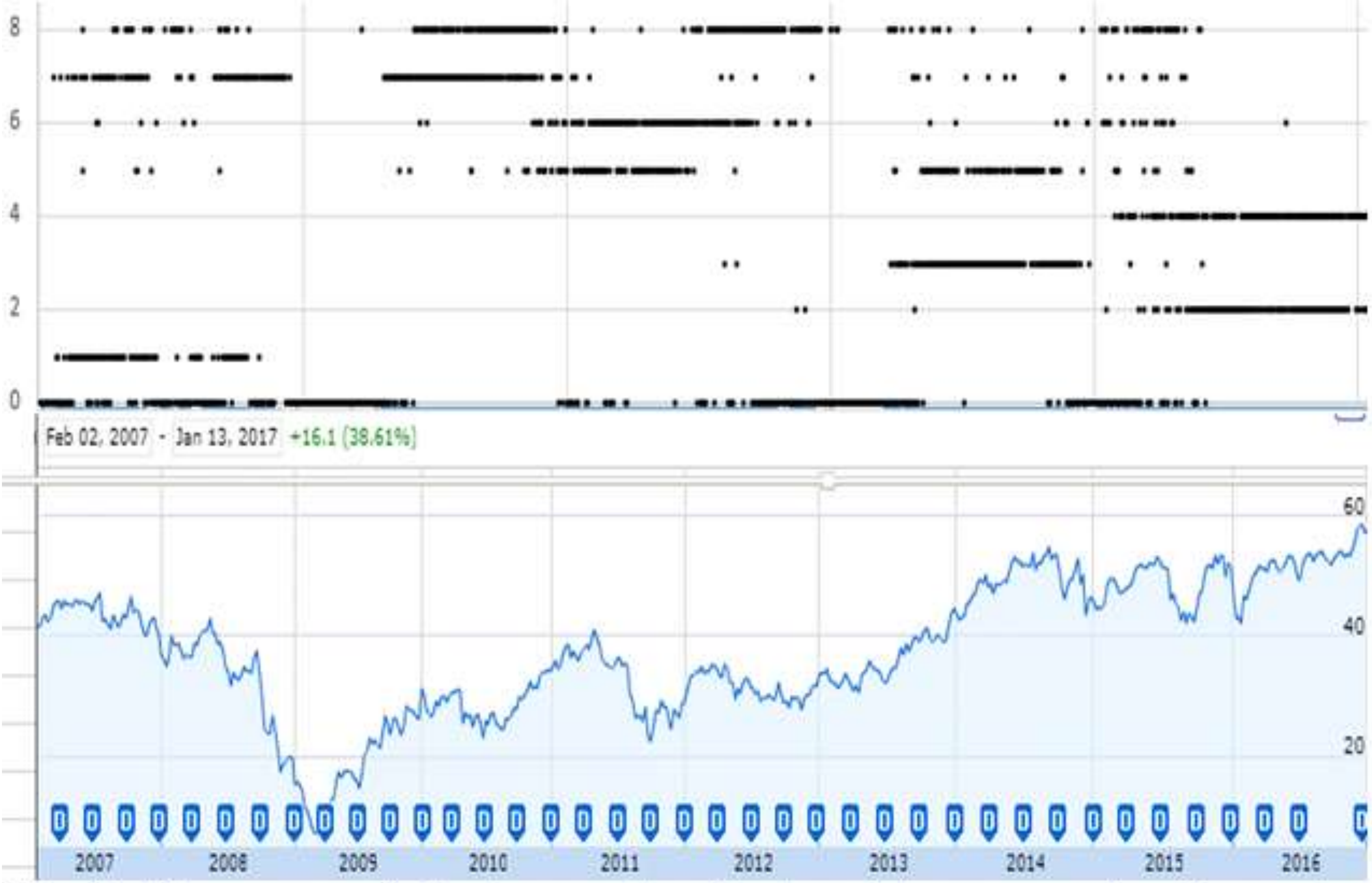


Fig. 9 Ten years of cluster membership by date compared with a price history for the DOW stock. The upper left label indicates the cluster IDs.

Objective function implemented in separate software stack:

Limited resource ("initial pot") adjusted by gains and losses.

Portfolio aggregation: k targets (e.g.5) will be evaluated together at each time step.

Deterministic change evaluation using next day opening price and strict fixed day hold and sale at closing price.

Cumulative aggregation by step-forward rolling results (with rebalancing) and compounding.

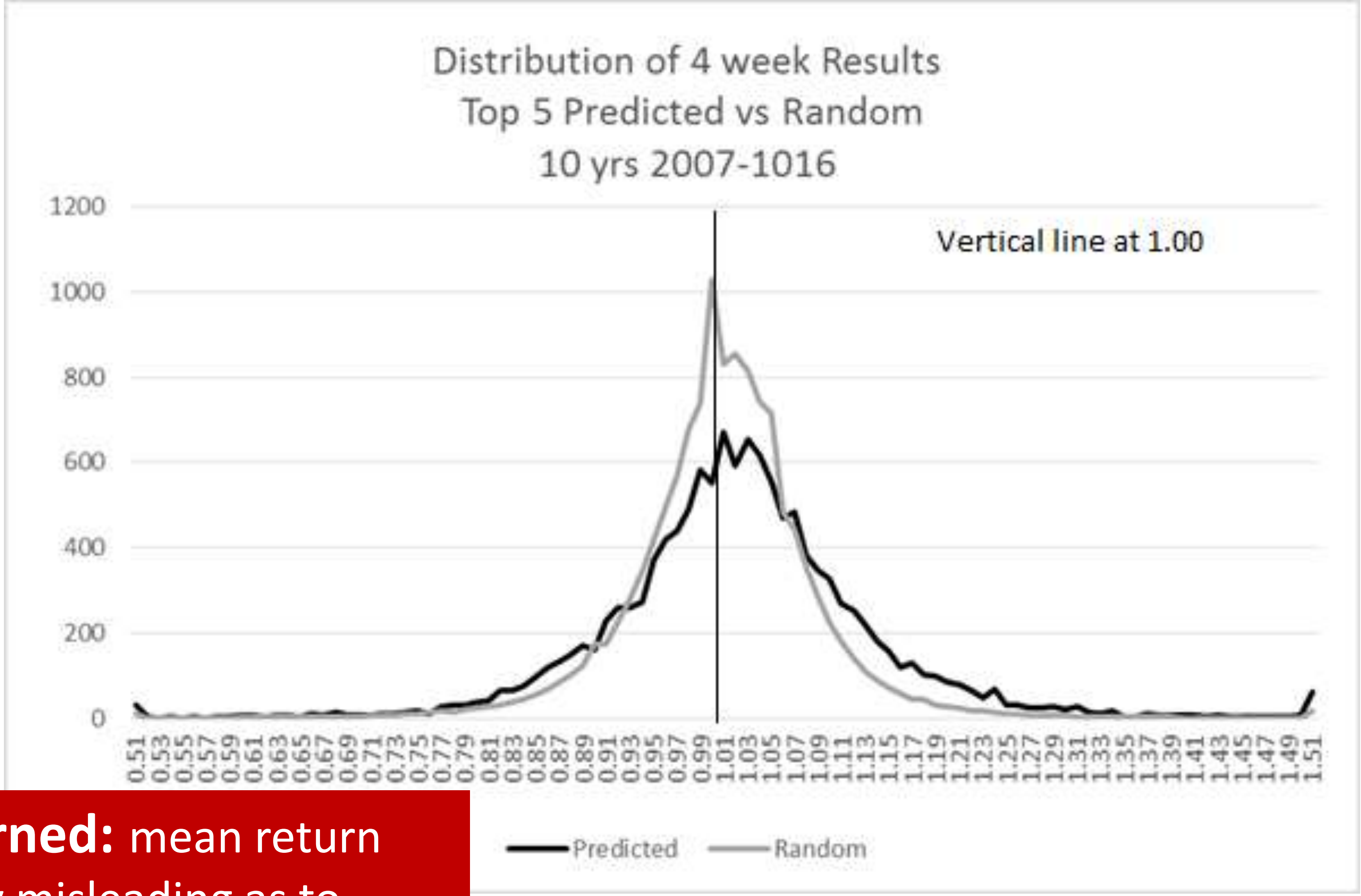
Trading costs applied.

Gains and losses are measured in dollars, not in normalized units (e.g. Z-score).

Measurements are on predicted daily topN of target set.

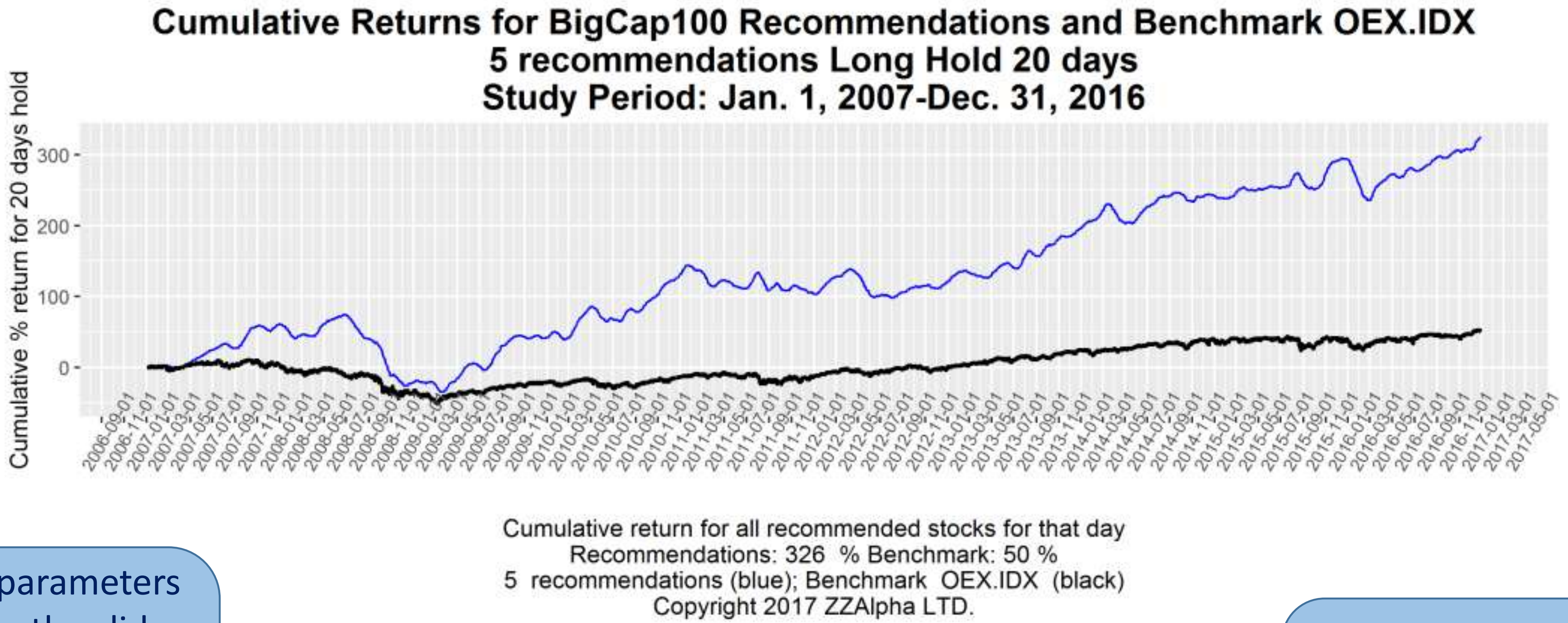
Time-stamp certification in production system

Results compared against Monte Carlo trials



Lesson learned: mean return can be highly misleading as to cumulative return.

Results compared to objective stock market benchmark



Tuning parameters shown in the slides and paer are for example and are not intended as optimal.

Wall clock time for daily ingest, train, predict and publish: 30 seconds

Algorithm is in production:

Sample commercial product:

Thank you.

ZZAlpha LTD.

Consistent returns from machine learning

Conservative Stock Picks

Solid, profitable recommendations in a turbulent world.

10 stocks for trading on Nov 29 2017

LONG Portfolio BigCap100 Lg for 20 days

List of today's FRESH recommendations:

AAPL,ABBV,AVGO,BA,BAC,C,CAT,MCD,NFLX,NVDA, Total stocks recommended LONG today: 10

REMINDER: do not use yesterday's recommendations that have now gone stale and can lead to trading errors!

		Cap	Sector	Industry	SIC	Vol(10d)	Last EA	Liq	Trnd	Rec	Score	Close
AAPL	Apple	$898299 m	Technology	Communications Equip	3663	22261 k	2017-11-02	$3895 m	9.9	C	4.9	173.07
ABBV	AbbVie Inc	$150594 m	Health Care	Biotechnology & Drugs	2834	4055 k	2017-10-27	$384m	6.5	C	5.0	95.42
AVGO	Broadcom Ltd	$112345 m	Technology	Semiconductors	3674	2320 k	2017-08-24	$655m	13.8	C	5.1	277.40
BA	Boeing Co	$157917 m	Capital Goods	Aerospace and Defense	3721	2062 k	2017-10-25	$548m	3.4	C	5.1	267.99
BAC	Bank of America Corp	$278080 m	Financial	Regional Banks	6029	53632 k	2017-10-13	$1426 m	-0.4	C	4.7	27.64
C	Citigroup Inc	$191055 m	Financial	Regional Banks	6029	11477 k	2017-10-12	$826m	-0.1	C	4.8	73.70
CAT	Caterpillar	$82106 m	Capital Goods	Construction & Agri Machinery	3531	2600 k	2017-10-24	$357m	1.5	C	4.8	138.99
MCD	McDonald's	$136929 m	Services	Restaurants	5812	2433 k	2017-10-24	$411m	4.5	C	4.2	171.34
NFLX	Netflix	$84953 m	Services	Broadcasting & Cable TV	4841	4514 k	2017-10-16	$883m	2.0	C	6.0	199.18
NVDA	NVIDIA	$130247 m	Technology	Semiconductors	3674	12892	2017-11-09	$2797 m	7.7	C	15.5	210.71

RECENT PERFORMANCE	5 Yr(annualized)	One Yr	Yr to Date
BigCap100(Top 5 Score)	19.3	27.2	20.9
BigCap100(10)	16.2	22.2	16.9
Benchmark: OEX.IDX	12.5	19.0	16.2