Linking Many Unusual Co-Incidences

Kevin B. Pratt
ZZAlpha LTD
Tucson, Arizona, USA
kevin.pratt@zzalpha.com

Abstract--We describe how sets of many unusual prior events can be linked and used to predict subsequent unusual events. We apply the technique to extract useful meaning from the massive, seemingly random co-incidences in the dynamics of the US stock market. We show how to qualify the links, to visualize their evolution, and to determine their similarity. We use the similarity for unsupervised clustering of sets of links. We make daily predictions over ten years that, when implemented, show excess returns in large capitalization equities in the US stock market. We begin from an insight about how our new puppy learns that she is about to get her evening walk well before we take out her leash. Our learning algorithm extends her efforts.

Keywords--unusual event, co-incidence, high dimensional visualization, similarity metric, unsupervised clustering, big data, reinforcement learning

I. Introduction and Motivation

My new puppy has already learned without assistance that my switching on the front porch light and closing my laptop and opening the coat closet (regardless of sequence) often means she is going to have a walk. That set of "unusual," ambiguous and disconnected events that she observes (among the many, diverse actions during my busy days) are not spurious co-incidences in her world, but are, together, sufficiently predictive of an interesting opportunity. What is her algorithm, what information does she store and how does she process?

There are thousands of unusual, though not rare, events in US economic and stock market data every week. Can we connect and use thousands of those unusual events to better understand and predict the connected dynamics of that economic/market system and to identify profitable opportunities? Can we distinguish spurious links from useful ones?

The null hypothesis is that apparent links between unusual market events and subsequent investment opportunities are "mere co-incidences" that do not, in the long term, aid in obtaining significant excess returns relative to objective benchmarks from the same historical period.

This paper describes three innovations. First is an algorithm to learn, without a model or *apriori* insights, which sets of linkages are important in a poorly understood, massively connected, externally impacted, constantly adjusting system. In an example described here, the algorithm learns from a small number of time-step samples of many thousands of sparse events and then at each further time-step scores the 100 largest US stocks to obtain returns that significantly exceed benchmarks as well as large, randomized (Monte Carlo) simulations.[10] Second, we propose and use a new similarity measure to cluster large, un-equal sized sets of unusual events through time to help characterize those sets. Third, we describe a visualization of sets linkages through time in large, interactive bi-partite graphs. We also unexpectedly discovered temporal pairings of sets of events.

Much traditional statistics and machine learning seeks to describe and predict from "what is usual." There is also work on exceedingly rare events. [12] . We focus instead on large sets of many ***UN***usual (though not necessarily rare) events that seem to participate in sequential behaviors often called "co-incidences." We intentionally omit the ordinary as overloaded with unnecessary information.

Analysis of many unusual coincidences is needed in fields as diverse as fraud detection[16], machine failures[17], multi-genes links to phenotypes[5], and behavioral economics[7,8,15]. Analysis can be specially challenging in non-stationary time-series contexts such as equities markets or airplane maneuvers, and where types of events at every time-step number in many thousands.

II. Data and Algorithm

A. Event channels

The example described here uses over 5000 "channels" of event types in the US markets and economy. Market information consists of many time-series or streams of information. Within a time-series, one or more pre-defined "events" can occur. For example a local maximum or local minimum can each occur in a stock price time-series, but each lives in its own "channel." An event can be as simple as a numeric threshold exceedance or as complex as a text extraction of a significant fact. We encode each channel as a boolean sequence of time-steps (bitstring) where an "unusual event" is indicated by 'true'. Multiple bitstrings are aligned by the semantics (e.g. time-steps) of the bits. In other domains, channels might be called sensors (electro-mechanical), actions (marketing, terrorism), lab results (medical), delays (logistics), etc. The economy and stock market are more challenging than domains where there are geo-spatial or physical maps from which to obtain intuitions of event links.

The example described here also uses a universe of 100 subsequent event types ("opportunities") in the US equities market. Those are investments that would gain from significant price increases by the end of 4 weeks. The universe is the 100 largest capitalization US stocks. An example of an opportunity is the action to immediately buy General Motors stock before its price rises at least 3% at the end of 4 weeks. There is no 'magic' in those numbers - the intent is to exemplify a useful benefit that the algorithm learned to predict. In other experiments we have successfully used similar opportunities in the 1500 Standard and Poors' listed large, medium and small stocks.

The opportunity parameters must be precise so that results are deterministically calculable in post-processing by separate trading evaluator software.

The data is eleven years updated daily from January 2006 through December 2016 and is available from public data providers, such as Yahoo Finance, Google Finance, and FederalReserve.gov and from subscription providers such as EODData.com and S&P CapitalIQ [13].

B. Unusual events within channels

"Unusual events" are defined in advance using earlier data. We tune threshold parameters to create "unusual" frequency that is neither rare nor common. Typical "unusual" events are high and low values, high and low variations, and discrete events such as textual facts in government reports and may be applied to state, change, and/or accelerations. Definitions of "unusual" can reflect notions of context and historic distribution of values or discrete occurrences. Implementations may include z-score, decile, absolute, moving and dynamic thresholds[4]. In the channels used in the example here, "unusual" events occur on average in 17% of the time-steps for a channel Fig. 1. The specific selection of which channels, what constitutes an unusual event within the channel, and what defines a recent time window are currently proprietary. (However hundreds of the useful time-series appear in daily financial publications [2]).

We are entirely agnostic as to whether, how and when those prior events relate among themselves and to subsequent opportunities. There is no initial model. The algorithm seeks to learn from scratch using the "big data" in the economy. This "no available model" situation can also arise when sensor architectures are highly complex, incompletely documented, have broken links and data, or are probabilistic[6].

C. Linkages

The algorithm learns from *durable* linkages from sets of prior *durable unusual events* to subsequent *opportunities*. As described below, the algorithm uses notions of *candidate* links, *durable* links to *positive opportunities*, and links to *duds*. Terms in italics are defined in Fig. 2. The tests for *durable* act to qualify the links and discard the spurious. To predict, the algorithm looks at the *durable links* exposed by *durable unusual events* in the most *recent time window*.

The number of potential single linkages of unusual events to opportunities can be hundreds of thousands at each time-step, and the possible combinations of those into sets as large as 1000 are much larger. As Fig. 9 shows, those combinations also cluster and change over the 10 year study period.

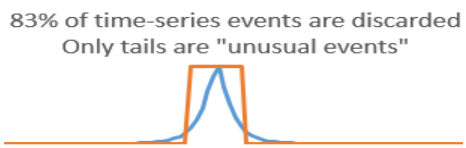

Fig. 1 Typical distribution of events in blue. "Unusual" events are outside red box.

D Reinforcement Learning Algorithm

There can be debate whether this is a reinforcement[14], supervised or unsupervised learning algorithm. The peculiar attributes are that it learns from its environment without a model, but it ignores most of its environment (the ordinary events). At initialization a parameter sets a definition of "good" opportunities. It ingests results as the outcomes of opportunities become available (e.g. after 4 weeks) to use in future training. It adapts, without manual re-tuning, to a morphing probability environment that undermines validity of prior optimizations that would be typical of reinforcement in stationary environments.

E Learning process

We apply some insights from the puppy learning mentioned earlier: the "unusual" matters and we can ignore or forget the rest; simple links are unreliable, ambiguous sets are acceptable; significance changes over time; and simple counts suffice.

The algorithm Fig. 2 uses the notion that as learning occurs, *candidates* (both unusual events and links) can progress to *durable* status, but can revert to *candidate* status if they do not continue to meet *durable* status requirements. They will also eventually be forgotten beyond the *memory period*.

Algorithm parameters to set:

i. *Event channels.* Define the universe of channels. Within a time-series, define what is an event and how events will qualify to be tagged as "unusual" within a channel.

ii. Duration of *recent time window* for collecting unusual events. Each *time-step* has a *recent time window*.

iii. Minimum number of occurrences of unusual events needed to move an event from *candidate* to *durable* status.

iv. Number of *time-steps* in the longer *memory period* to use in counting event occurrences and opportunity occurrences.

v. *Opportunity universe.* Define the universe of opportunities. Define an *opportunity window* and required change. A *positive opportunity* is one which exceeds a change requirement within the opportunity window. A *dud* is one which fails the change requirement.

vi. Define the minimum ratio of specific *positive opportunity* occurrences to specific *durable unusual events* that will adjust link status from or to *candidate* or *durable*.

vii. Method for combining actual results of links (within the *memory period*) for a *score value* to be applied to an opportunity.

Algorithm learning process:

1. Start recording
2. Record unusual events that occurred in *recent time window*. These are *candidate* unusual events.
3. For the *opportunity window* that has closed in this *time-step*, record actual result for every opportunity in the *opportunity universe*. Note the *positive opportunities*.
4. Record *candidate* links, which are those links from the earlier *candidate unusual events* to *positive opportunities* (Note: must align end of the earlier *recent time window* with start of *opportunity period*).
5. Within each channel, count unusual event occurrences in *memory period*. Update status (*candidate* or *durable*) of unusual events according to settings.

6. Count link occurrences in *memory period*. Update status of links (*candidate* or *durable*) according to the minimum ratio setting.
7. Increment time-step.
8. If a *memory period* is fully populated with *time-steps*, a training set is complete and the *durable links* knowledge is available for scoring. ASSERT: Most or all *opportunities* in the opportunity universe are *durably linked* to a set of *recent durable events*. Note: There can be over 1000 *durable links* to a *positive opportunity*, but sometimes there is only one. Now **Invoke scoring process.**
9. "Forget" information in oldest *time-step* in *memory period.*
10. Iterate from 2.

Algorithm scoring process:
1. Identify *durable* unusual events in *most recent time window.*
2. Locate all *durable links* from those *durable unusual events.* These point to *opportunities*.
3. Evaluate historic actual results of linkages of those *durable unusual events* to each of the *opportunities.* Note: include in actual results both the *positive opportunity* results (successful) and the *dud* results (losing).
4. Combine the historic (i.e. in the *memory period*) actual results of sets of *durable links* to *opportunity* into a *score*.
5. Rank the *opportunities* by *score*.
6. Act upon the topN scored *opportunities*.

Fig. 2 Learning and scoring algorithm

F. Parameter settings

In the experiment shown here, we require 12 occurrences of the unusual event in the memory period to advance to durable status. Memory period is one year. Time-step is daily. 100 big cap stocks are opportunities, with 3% gain in 4 weeks as positive opportunities. At least half the durable unusual events of an event channel must link to the positive opportunity in order to deem a link durable. Arithmetic mean is the combination method for actual results. We use five as topN. Learning and scoring a time-step takes about 1 minute of wall clock time.

Traditional statistics may assert that the channel samples (*durable events*) supporting a link are too small (as few as 6), that the over 5000 sparse event channels just contribute noise, and that limiting to 250 daily samples for training make prediction unreliable. Results here challenge those assertions.

III. LARGE BIPARTITE GRAPH VISUALIZATION

Constantly changing, many-to-many relationships in big data can be difficult to grasp. We needed to understand the evolution of the sets of prior events that foreshadow an opportunity in order to propose more hypotheses for testing. The visualization tool is *not* used for prediction.

With 5000 possible prior events channels, 100 possible subsequent opportunities and potentially 100,000+ links, an ordinary bi-partite graph representation was unworkable. We also wanted to interactively investigate which recent event "linked" to possibly multiple subsequent opportunities, and vice versa. The graph needed to display various metadata about the channels, e.g. which were typically categorized as "energy sector related" or "financial indicators", etc.

Fig. 3 shows our interactive visualization tool applied to one time-step with all of the links to positive opportunities. This tool was immediately useful for discarding an early hypothesis that an opportunity would mostly be preceded by activity in closely related economic categories. Clicking on several opportunities, Fig. 3, quickly showed that the "fans" vary tremendously among the opportunities, even within a single time-step. ("Fan" refers to the visual effect of a set of durable links spreading from a box (i.e. an event) to other events.)

Fig. 4 shows a view of a day in history and shows both durable unusual events in the recent time frame (on the left)

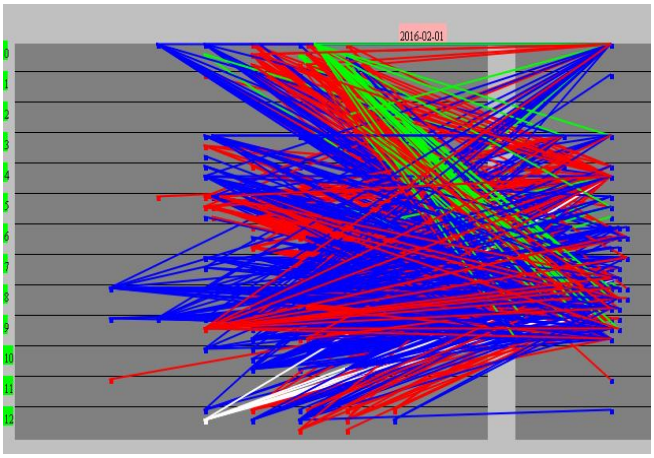

Fig. 3 This visualization shows all links among about 200 durable event occurrences in a recent time frame (on left) to about 60 positive opportunities in large cap stocks (on right) as of 1 Feb 2016. Each box on the left represents a recent durable event in its own channel. The other about 4800 channels did not have durable events in the recent time frame and so are not displayed here. The colors indicate channel types of high-blue, low-red, active-green and inactive-white. Numbers on left reference various economic categories. The graph is interactive: one can click on a box or link to view identity, specific link sets and metadata.

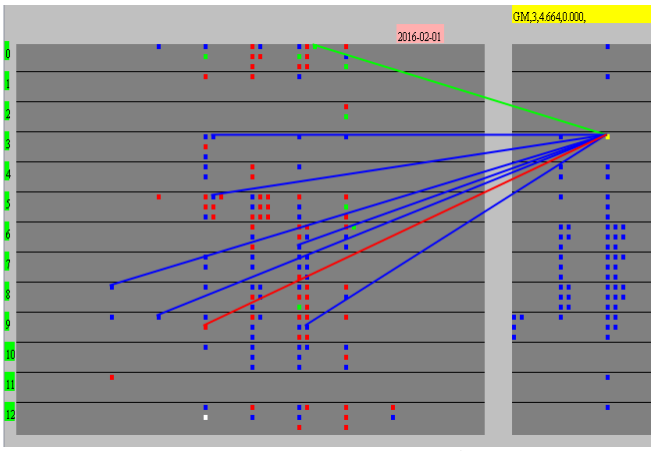

Fig. 4 Here, a user has selected the box on the right for GM stock purchase opportunity for 1 Feb 2016. The selection causes display of the "fan" (i.e. a set) of links from recent unusual events on the left. Metadata is displayed for the box selected. A user can also select a channel on the left and view the fan to a set of opportunities on the right, or can inspect a specific link.

relative to 1 Feb 2016 and the positive opportunities (on the right) that had in fact completed 4 weeks later.

A. Visualizing historic event occurrences supporting durable status

When inspecting the durable events on the left side of Fig. 3 and 4, one wants to know how these durable events came to be and whether pathological patterns appear in their histories. For that purpose we developed a "swim lane" graph of the history of those events Fig. 5. Fig. 5 in effect drills into the history behind the "fan" shown in Fig. 4, and earlier time-step fans for the specific opportunity.

Fig. 5 confirms that with few exceptions, the durable events preceding the current GM stock opportunity acquire and maintain their durability over a scattering of time-steps, rather than in vertical streaks that may be associated with a channel's "trend." Auto-correlation trends in market channels can be undesirably brittle [15].

Fig. 6 shows the temporal distribution of the cardinality of durable events relating to a specific opportunity (in this example Dow Chemical Company stock) compared to the temporal distribution of all the durable events. This graph (and similar) dispelled a hypothesis that event cardinality would show periodicity due to quarterly financial and economic reporting.

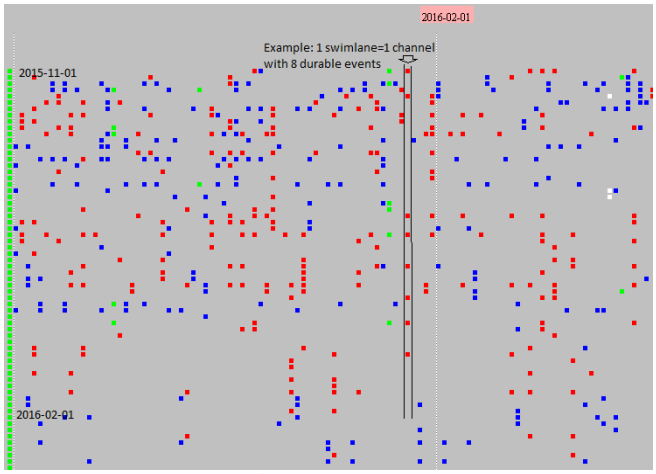

Fig. 5 A "swim lane" view of the GM stock related durable unusual events in the 3 months (66 time-steps) leading down to 1 Feb 2016 (bottom-most row). Each column is an unusual event channel and the green boxes on the left side indicate a sequence of days. As with Fig 3 and 4, the colors indicate channel types. This figure is a clip out of a large 8000 x 6000 pixel scrollable view of a full year of all channels pertaining to GM stock opportunities over that time. A user can pre-select the opportunity of interest and the number of time-steps desired, and can interactively click on boxes to see dates and other metadata.

IV. Clustering

There are many flavors of clusters to explore in the data: co-occurrence of opportunities, co-occurrence of unusual events, clusters of days based on events, and clusters of days with respect to a specific opportunity. We explore the latter here because we want to understand whether the durable unusual events cluster in time in some fashion suggestive of better "opportunity periods" for a specific opportunity.

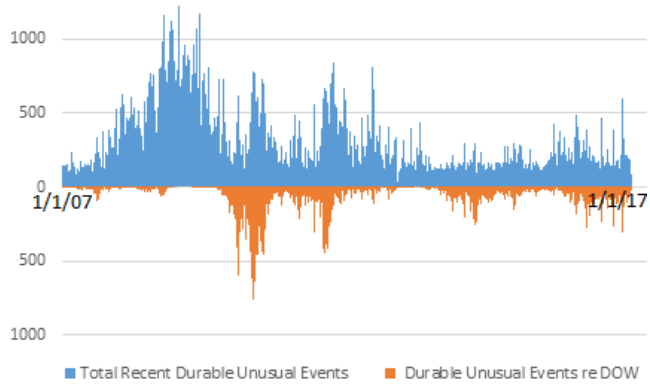

Fig. 6 This butterfly chart shows ten years (2007-2016) of relative occurrence of total recent durable unusual events vs those preceding DOW stock opportunities. The relationship is variable and neither shows reliable periodicity. However, this visualization helps confirm that sets of unusual events can be specific to opportunities and not correlated to the total events numbers.

A. Similarity measure

We developed a similarity measure that would emphasize likeness while also permitting fine granular ordering based on unlikeness. More common Hamming or Jaccard distance did not fit our needs Fig. 7. Hamming expects equal length strings without missing data and measures *dis*-similarity using XOR. Jaccard measures similarity using intersection but normalizes over union. When strings are of unequal length, and there is any density of events in the longer string, the Jaccard measure of similarity reduces linearly in the excess length of the longer string and produces skewed measures: the large dissimilarity denominator can overwhelm the similarity numerator even when significant portions of the two strings match.

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.

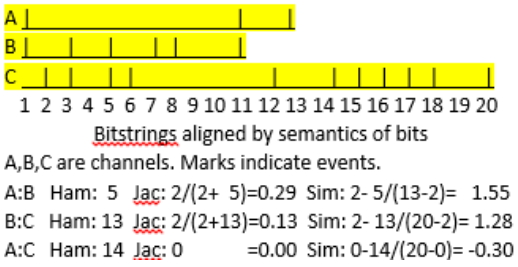

Fig. 7 Example of similarity measure compared to Hamming and Jaccard. In our similarity the integer portion encodes likeness and the fractional portion encodes unlikeness. It preserves ordering on likeness.

For comparison of sets of durable events on two dates with respect to each other, we represented the unusual events on a date as a bitstring of length n where n is the global number of distinct channels. (In effect comparing two *rows* in Fig 5.) The date become the node identifier and the bitstring its attribute.

Because calculations on bitstring representations are very fast, we can quickly determine all pair similarities even on big data with many thousands of channels and dates.

Bitstring representation may remind industrial readers of exceedance/alarm tags on the control charts long used in Statistical Process Control for industrial processes[17].

B. Unsupervised clustering

The nodes and weighted links (dates as nodes, and links having bitstring similarities of events as weights) were imported to the freely available tool Gephi [3] for visualization and clustering. To improve visibility, we pruned the links of each node to the five most similar nodes. Gephi provides unsupervised modularity clustering using the Blondel-Louvain algorithm[1,9]. Fig. 8 shows a Gephi representation of the clusters, with nodes colored by cluster (modularity class number).

Fig. 8 shows both that the clusters are relatively tight internally, and some have common links to other clusters. Unlike k-means, the unsupervised clustering algorithm does not require setting a pre-determined number of clusters. The

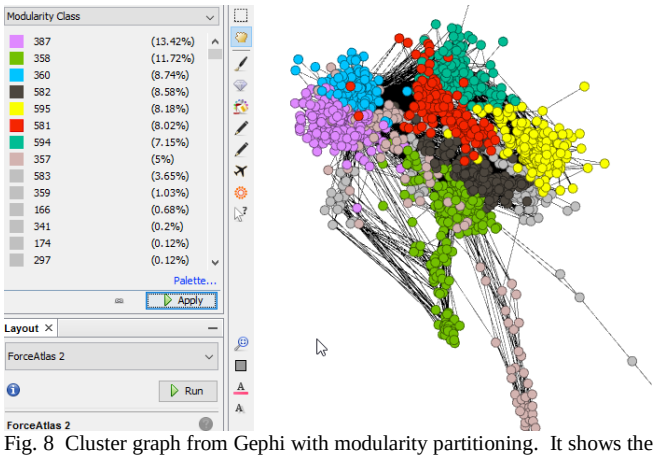

Fig. 8 Cluster graph from Gephi with modularity partitioning. It shows the top 8 largest clusters in distinct colors (color key and clusterID on left). Gephi layout for this view is ForceAtlas2.

only clustering parameter setting needed is resolution=0.75. Gephi supports interactivity and a variety of statistics. However, one can simply export the nodes with the newly assigned modularity to other visualization tools.

Fig. 9 shows the clustering of the days across time (with respect to the DOW stock opportunity). The upper portion shows the presence of the top 8 modularity classes during the ten years. The lower portion is a ten year stock price line graph from finance.google.com for Dow Chemical.

We were surprised by Fig. 9. What is unexpected is a) that pairs of clusters often persist for several months without much interleaving of other cluster classes (cluster classes represent similar durable event sets) and b) when pairings expire, they seldom (with some exceptions) occur in later times. Recall that the clustering relied solely on attributes which are the likeness of occurrences of unusual events and NOT the dates they occurred. This helps confirm a hypothesis that the linkages of sets of unusual events are indeed durable for some time, but also eventually "die off" as the economy evolves. This confirms the need for the algorithm to continually update learning from the moving memory period.

V. Sample Financial Results

The example provides results from evaluating ten years (2519 market days) of predictions. Results compare against common investment benchmarks S&P 100 Index (OEX), a S&P 500 Exchange Traded Fund (SPY, which includes gains from dividends), and random selection (1000 Monte Carlo trials) from the 100 big cap stocks (capitalization determined annually).

The mean random result value is 0.7% increase per 4 weeks and the mean algorithm result is 1.5% increase per 4 weeks. Fig. 10 shows the comparison of distributions. (We assume investment in the top 5 selections from the 100 opportunities every day). Fig. 11 compares the ten year cumulative gains of the OEX and SPY benchmarks against the selections made by using the algorithm scores.

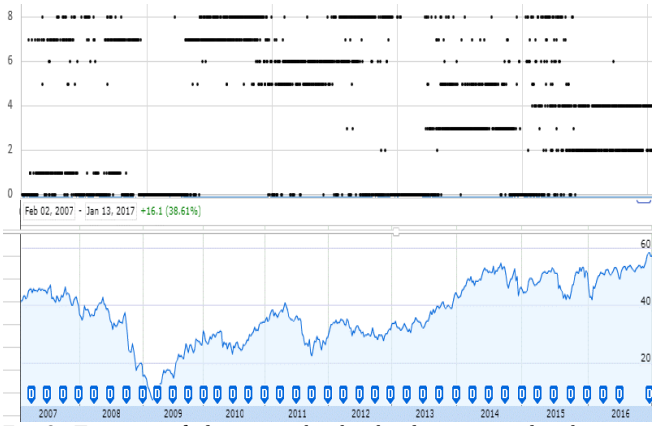

Fig. 9 Ten years of cluster membership by date compared with a price history for the DOW stock. The upper left label indicates the cluster IDs.

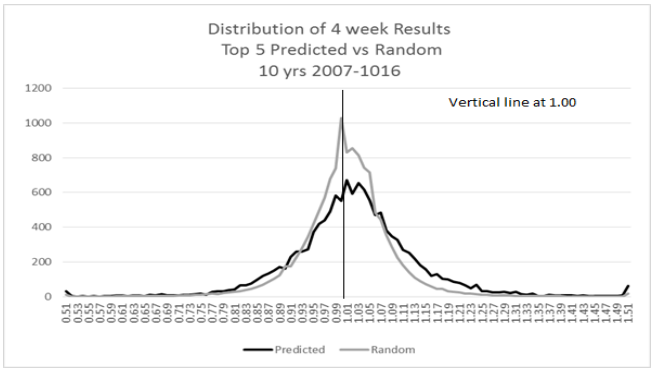

Fig. 10 Use of top-5 predictions cause the distribution of results to shift to the right relative to the random selection distribution. The "spike" at 1.0 (no gain or loss) in the Monte Carlo distribution is caused by missing results in the data caused by mergers, ticker changes, and missing/erroneous contemporary data. The data used for the big 100 results contained the same missing entries. "Dirty data" is a constant irritant when doing real-time scoring in the stock market.

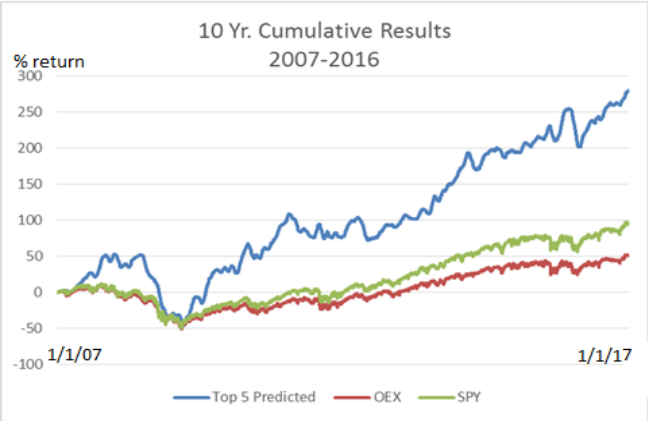

Fig. 11 Ten year cumulative results.

Result values come from trading evaluator software that reads the top 5 selections produced by the algorithm, rolls and compounds 4 week outcomes (both successful and losing), rebalances dollar results, and subtracts trading commissions of each trade. The trading evaluator is entirely distinct software from the algorithm described here and is applied after the prediction process completes for the entire 10 year period. [11]

The investment domain is less interested in confusion matrix accuracies and more in compounded results and short and mid-term relative financial risk. Deep and long duration "valleys" as seen in Fig. 11 indicate risk. Faster recovery can offset risk fears.

VI. Conclusion and Future Work

We are unaware of other work evaluating and predicting economic opportunities from large sets of many thousands of channels of ***un***usual events. Our work validates this approach using an example that identifies opportunities for excess market gain in large stocks. There are likely many possible optimizations and tunings.

In a dynamic environment, the realized value of an opportunity can depend on other externalities (e.g. world news and fake news) that do not fall in the categories of economic events. Those could be added as channels.

The clustering shown here uses "hindsight clusters". It would be more useful to watch the clusters evolve at each time-step, and to evaluate using the real-time clusters in the scoring algorithm.

This algorithm and clusters of economic features may be a useful alternative for understanding multi-causal behaviors of the US economy. The approaches described here may also be useful in other big data, complex linkage challenges. Some examples are the diverse maintenance sensors on a large airframe and its power units, nervous system components in animals, and drug side-effects across a diverse patient population. Or maybe it is simply good for appreciating a new puppy.

Some related presentation slides, videos of visualization interactivity and animation are available at www.puppypicks.biz.

References

[1] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre, "Fast unfolding of communities in large networks," J Statistical Mechanics: Theory and Experiment, vol.10, 2008, p1000.

[2] Simon Constable and Robert E. Wright, Guide to the 50 Economic Indicators That Really Matter, NY, USA : Harper Collins, 2011.

[3] Gephi 0.9.1, at www.gephi.org, 2017.

[4] Eugene Fink, Kevin B. Pratt, and Harith Suman Gandhi, "Indexing of time series by major minima and maxima", In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2003, pp. 4280-4287.

[5] Leland H. Hartwell, John J. Hopfield, Stanislas Leibler and Andrew W Murray, "From molecular to modular cell biology," Nature vol.402, 2 Dec. 1999.

[6] C. S. Holling, "Understanding the complexity of economic, ecological and social systems," Ecosystems vol. 4(5), 2001, pp.390-405.

[7] Daniel Kahneman, Paul Slovic and Amos Tversky, Eds., Judgment under uncertainty: Heuristics and biases, Cambridge U. Press, 1982.

[8] Daniel Kahneman, and Amos Tversky, Eds., Choices, Values, and Frames, Cambridge U. Press, 2000.

[9] R. Lambiotte, J.-C. Delvenne, and M. Barahona, "Laplacian Dynamics and Multiscale Modular Structure in Networks", IEEE Transactions on Network Science and Engr, vol.1(2), 2015, pp.76-90.

[10] Andrew W. Lo and A. Craig MacKinlay, A Non-Random Walk Down Wall Street, Princeton NJ, USA: Princeton U. Press, 1999.

[11] Kevin B. Pratt, "Proof protocol for a machine learning technique making longitudinal predictions in dynamic contexts," In Proceedings of the 21st ACM SIGKDD Conference, 2015, pp.2049-2058.

[12] Nassim Nicholas Taleb, Black Swan: The Impact of the Highly Improbable, 2d Ed. NY,USA: Random House, 2010.

[13] James P. O'Shaughnessy, What works on Wall Street, NY,USA: McGraw Hill, 2011.

[14] Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction 2d ed, in progress, 2012.

[15] Richard H. Thaler, Winner's Curse: Paradoxes and Anomalies of Economic Life, NY, USA: Princeton U Press, 1994.

[16] Dallas Thornton, Roland M Mueller, Paulus Schoutsen and Jos van Hillegersberg, "Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection", In Procedia Technology vol.9, 2013, pp.1252-1264.

[17] William H. Woodall, "Controversies and contradictions in statistical process control," J. of Quality Technology, American Statistical Association, Oct 2000.