An Empirical Contextual Validation of the CapitalCube™ Market Trading Variables as Reflected in a 10-year Panel of the S&P500: Vetting for Inference Testing

Edward J. Lusk[1] & Michael Halperin[2]

[1] The State University of New York (SUNY) at Plattsburgh, NY, USA & Emeritus, Department of Statistics, The Wharton School: University of Pennsylvania, Philadelphia, PA, USA

[2] Director Lippincott Library of the Wharton School [Retired], University of Pennsylvania, Philadelphia, PA, USA

Correspondence: E. Lusk, SBE SUNY Plattsburgh 101 Broad St. Plattsburgh, NY, USA 12901. Tel: 1-518-564-4190 or 1-215-898-6803

Received: November 6, 2015 Accepted: December 2, 2015 Online Published: December 8, 2015

doi:10.5430/afr.v5n1p15 URL: http://dx.doi.org/10.5430/afr.v5n1p15

Abstract

Introduction: We have accepted an Assurance engagement to examine the CapitalCube™ market analytics platform as the lead navigation tool offered by AnalytixInsight™. This first vetting is to examine the variables which constitute the measures used in creating decision-making inferential information. **Study Precise**: Our reasoning in examining the reasonability of the CapitalCube variable set is that in the Big Data world spurious associations, the bane of relevance, are expected. We examined independently the four ***Context Variables*** and the four ***Decision-making Variables*** offered by AnalytixInsight. For the former, we used Spearman ρ screens to eliminate firms that were dynamically not in-sync with expectations of the CapitalCube Panel as expressed through the S&P500 Panel. Further, we examined inferential power issues for extended analyses. For the Decision-making variables, we used Harman Factor results to test various *a priori* hypothesized profiles. **Results**: For the Spearman screens, we eliminated only 1% of the firms in the Panel; for the Power screens, we eliminated eight firms resulting in a Panel of 487 firms. For the Decision-making variables, the Factor profiles were strongly supportive of expectations. **Impact**: These eight CapitalCube variables are arguably in-sync with the empirical trajectory of the S&P500 Panel over the 10-year accrual period starting in 2002. Therefore, these CapitalCube-variables seem capable of market discrimination. This variable vetting is the critical first step in evaluating any analytic platform in the trading market milieu where Big Data rules of engagement must be serviced.

Keywords: Trading Market Platform Analytics, Big Data

1. Introduction: Setting the Analytic Context

1.1 Précis: The Operational Prologue

1.1.1 Introductory Section Following on the cautionary work of Fan, Han, & Liu (2014), relative to spurious inference in Big Data sets, we offer an important perspective on market analytics in the Big Data milieu. This overview will rationalize and so provide the context for the technical aspects of this paper which are addressed to providing reasonable assurance that the CapitalCube [http://www.capitalcube.com/] market analytics platform is formed of structural coherent and sensitive variables—clearly a pre-condition for engaging an analytics platform in the market trading world.

1.1.2 Conditional Screening of Firms We use the set of CapitalCube context variables as a filter to screen firms that are not consistently tracking with the S&P500 that is the data-panel for expression of the CapitalCube variable-set.

1.1.3 Vetting the Decision-making Variable Set We examine the four CapitalCube decision-making variables as they relate to *a priori* expected tracking over the S&P500 panel. In this section, we test the structural integrity of this inferential variable set; these are the details of the conditional vetting to examine the reasonability of the CapitalCube analytics platform as an inference platform in the Bid Data context.

1.1.4 Summary The concluding section is formed around the conditional imperative of sensible assurance as argued by Kwon, Lee, & Shin (2014). We summarize the various statistical tests employed in the service of the necessary vetting of market analytics Decision Support System [DSS] platforms such as those offered by CapitalCube.

1.2 The Big Data World: Vast Opportunities & Treacherous Pitfalls In the Big Data global market trading milieu, the critical issue is not finding data but rather processing the terabytes of streaming data to ***cull*** and ***glean*** in the service of creating longitudinal profiles of organizations so as to form reasonable action plans. Big Data is the popular linguistic currency of the long standing "Data Mining" protocols.

According to Lusk & Halperin (2015):

> *The lineage of Big Data Analytics [BDA] traces back to the single-portal linkage of the uncountable number of e-networks, such as Intra-Nets, LANs and W-Area Networks that effectively became the WWW circa 1993. At the dawn of this new information age there were a dearth of agile analytic tools to enable managers to (i) access this web-based new world of effectively unlimited data or (ii) form such data into decision relevant information. However, according to Lovell (1983) and Porter, & Gogan (2013, p.59) the Excel™ platforms of the 1980s would soon be the progenitors of the first generation Data-Manipulation packages that would be the platforms for Data Mining.*

Data Mining, relatively early in its pre-adolescence, was cast in the pejorative shadow of "mindless searching for theory in the data-dumpster" where one can always find a few spurious association sets upon which to form the next "theoretical" breakthrough. Data Mining was in need of a make-over to regain respectability. See: Kimble & Milolidakis (2015).

Diebold (2014, p.5), in a relatively unabashed fashion, details the essence of the Data Mining chrysalis which spawns Big Data analytics:

> *Now consider the emerging Big Data discipline. It leaves me with mixed, but ultimately positive, feelings. At first pass it sounds like frivolous fluff, as do other information technology sub-disciplines with catchy names like artificial intelligence, data mining and machine learning. Indeed it's hard to resist smirking when told that Big Data has now arrived as a new discipline and business, and that major firms are rushing to create new executive titles like "Vice President for Big Data." But as I have argued, the phenomenon behind the term is very real, so it may be natural and desirable for a corresponding new discipline to emerge, whatever its executive titles.'*

This ocean of streaming data created a need for eddies of relative calm where reflective contemplation could spawn data profiles in-sync with decision imperatives. Also see: Akkaya & Uzar (2011) and Yang & Fong (2015).

1.3 Market Analytics: Historical Perspective and Current Players This mother of necessity vacuum was the progenitor of many organizations offering services formed around General User Interfaces [GUI] that were essentially the DSS that served as the "life-rafts " of managers awash in data. See: Niessing & Walker (2015) and Slagter, Hsu, & Chung (2015). Interestingly, two of the very earliest data-analytic platforms pre-date the Internet by three decades: specifically COMPUSTAT™[http://www.spcapitaliq.com/our-capabilities/our-capabilities.html?product=compustat-research-in sight], and CRSP™ [http://www.crsp.com] both of which offered "download" capabilities that soon became the life-sustaining umbilical cords to market-trading studies. This was the embryo that would mature into Big Data Analytics. The next evolutionary step was a GUI-friendly Platform DSS that would facilitate the seamless profiling of the data stream which now resides in the Cloud or Virtual Platform. Some of the ground-breakers in the platform-oriented Data Analytics' milieu were:

Bloomberg™[http://www.bloomberg.com/markets/world],

Morningstar™[http://www.morningstar.com/],

Cable News Network, Inc.[CNN™ http://money.cnn.com/data/markets/sandp/]

WRDS™ [http://wrds-web.wharton.upenn.edu/wrds/], and

AuditAnaltyics™ [http://www.auditanalytics.com/].

A recent "New-Kid-on-the-Block" is *AnalytixInsight*™ [AI] [http://www.analytixinsight.com/]. AI offers an analytics-platform called: *CapitalCube*™ [http://www.capitalcube.com/] which is the lead-link in AI front-door's *Platform & Products* pull-down. The CapitalCube link creates a plethora of information as a "carve-out" of the Big Data market-trading stream. We are mentioning this as "kitchen-sink" web-links where the decision-maker has access to "everything" in the terabyte world guarantees the Claude Shannon and Warren Weaver nightmare of Information Overload. See the work of van Bussel, Smit & van de Pas (2015). The CapitalCube navigation dashboard seems to be carefully tuned to pre-selected and well-researched expert systems' guidelines and so *should*

be capable of steering the analytical ship between the *Charybdis*: [being swamped by a *whirlpool of irreverent data* and so essentially losing any chance to form reasonable action-oriented data protocols] and the *Scylla*: [of *"push-button" analytics* where the decision-maker [DM] gives discretion over to the "DSS" which produces only "here and there" relevant data protocols but *loses the possibility to use directed DM-intel to form a rich set of action dynamics*.] One, of course, searches in the market-analytics world for the *Goldilocks* platform-set: Not too much data & Not too prescriptive but rather the *just-right* interactive set of "carve-out" navigation tools that are DM-driven. This is the point of departure of our paper.

2. Analysis of the CapitalCube platform

We, the authors, together with Ms. Marjorie Churgin, *Director of Development*, AnalytixInsight, Inc. and Mr. Gautam Pasupuleti *COO*, AnalytixInsight. Inc. discussed over a number of months the possibility of analyzing and evaluating the CapitalCube analytics-platform. This resulted in a letter of agreement addressed to Mr. Chaith Kondragunta, *CEO*: of: AnalytixInsight, Inc & CapitalCube Corporation [AI&CC] where the CapitalCube data capture of the S&P500 would be given to us for our analysis. This is an Assurance engagement of an academic nature to wit: there are no monetary, *quid-pro-quo* compensations between us and AI&CC nor are there oversight approvals or veto provisions granted to AI&CC. Therefore, we are independent evaluators of this CapitalCube S&P capture; however, for clarification and accuracy checking, we sent reader/comment-drafts to the above-mentioned individuals. We have archived their comments and they are available upon request.

2.1 Specifics of the CapitalCube Dataset [CCD] We received the CCD as a download from AI&CC that comprised all the firms listed on the S&P 500 as of 9 April 2015; *in round* R×C-dimensions, the CCD is: 180,000 × 80. The longitudinal dimension starts on 2 Jan 2005 and terminates at 20 March 2015; the data from inception to and including 2013 are monthly time series; starting in 2014 to the last data point in 2015 the time series is more or less daily with the usual non-contiguous gaps for non-trading days. For most of the firms there are 371 data points.

2.2 Measured Continuous Point-Process Variables in the CCD

There are four ***decision-making variables*** created by CapitalCube (Note 1):

1. *Current Price Level Annual* [CPLA]; This is a ratio formed as the bell-price on a particular day as benchmarked by the Range of previous trading-day values going back one year in time. As such, basically the range of CPLA is [0 to 1].
2. *Scaled Earnings Score Average Latest* [SESAL]: This starts with the reported earnings of the firm and uses 50 or so calibration variables such as *Working Capital*; *Earnings Growth* & *Revenue Growth* to create an aggregate rolling benchmark that scales the reported earnings most always in the Range [1 to 100].
3. *Previous Day Closing Price Latest* [PDCPL]: Is the bell-price as adjusted for Stock splits and any sort of Stock spin-offs going back a number of years. The distribution of PDCPL resembling a Poisson pdf starting in the positive quadrant to the right of zero and tailing off into the low thousands.
4. *CapitalCube Price Latest* [CCPL]: This is a projective rolling variable—i.e., longitudinal—adjusted for Split/Spins, and benchmarked by a large number of market performance measures. The CCPL is projective in nature and used, for example, to index the *Under-and Over-Priced* labeling. The CCPL index-labeling employs a sensitivity analysis using a range around the mid-point of measured values of CCPL extending out to Min and Max boundaries. The CCPL indexing seems to be in nature the same performance indexing as one finds for the Tukey Box-Plot in SAS™ for outlier identification. The CCPL resembles a Poisson pdf starting in the positive quadrant to the right of zero and tailing off into the low thousands.

There are also four ***context variables*** that may be used in calibrating the four decision-making variables so as to create domain context information (Note 2).

1. [Fifty-Two Week Low] & [Fifty-Two Week High] both of which pertain to the CPLA;
2. [Capital Cube Price Range: Min] & [Capital Cube Price Range: Max]

3. Preliminary Contextual Analysis of the CapitalCube Data Set: Screening Protocol

To be clear: The purpose of what follows is the condition-testing of the CapitalCube Dataset. We are interested in the "inference-condition", in a Shannon-Weaver sense, of the CapitalCube Dataset. This means that before we investigate the information content of the AI{CC} platforms for the eight-variables noted above, we need to have reasonable assurance that these variables are in-sync with the market generation processes that underlie the dynamics of the S&P500 over the Panel. The condition-testing will be produced by a set of *a-priori* screens that we have

created to remove firms that: (i) are outliers in their context variable space, and (ii) compromise power in extended analyses. This screening will provide validation of the nature of the CapitalCube context variable set as one can proffer dynamic interactions that certainly should obtain for the S&P500 Panel. Failing to observe these expected market dynamics would not bode well for the inferential and projective validity of the CapitalCube variable sets. This is then the first aspect of the vetting of the CapitalCube Dataset[CCD].

3.1 Screening Firms on the Context Variables Logically, we selected the pairs:

1. [Fifty-Two Week Low] & [Fifty-Two Week High];
2. [Capital Cube Price Range Min] & [Capital Cube Price Range Max]

as these pairings have a natural expected associational relationship over the various firms in the CCD. To examine these dynamic relationships, we selected associational tests. Specifically, we used the conservative [relative to power] Spearman ρ [Rank Order] test as the Pearson Test assumes: (i) linear associational spacing, which seems too restrictive for a screening protocol for market studies as non-linear ordered differences are the likely norm, and (ii) in that likely context assumed bi-variate normality will slightly understate associational relationships for screening purposes. Therefore, the Spearman ρ is the preferred ranking model for screening purposes as it only assumes ordinal placement and not "equal" spaced Cartesian-Coordinate-orientation and so seems ideal for the CCD as relational order should underlie the relationship between High and Low point longitudinal movement. For the screening of firms, we have selected the following ρ cut-point: Any firm for which ρ is less than +0.15 will result in that firm being removed from the CCD. *Rationale*: The positive sign on ρ means that we expect that for market trading firms at time *t* the distance between the Low & High Points and the ordinal orientation of these points will be such that at time *t+1* the inter-point range may change but that there will be only a few instances where the points at time *t+1* are not uniformly greater than the points at time *t*. This is consistent with the market generating process driving most firms which has trended in a positive direction during the accrual period. Regarding the magnitude of ρ, we selected 0.15 as sample size of 371 points per firm gives power of 90% for a FPE of 5% [http://www.statstodo.com/SSizCorr_Pgm.php] relative to the Null of no association which is certainly adequate as a screening filter.

3.2 Screening Results We produced the screenings by running all the 500 firms in the CCD through the Spearman ρ functionality of JMP™/SASv.12. These results were then passed to a programmed-VBA™ module coded to identify all the firms with ρ < +0.15. For the first variable set this resulted in the following as presented in Table 1

Table 1. CCD Screening results for [Fifty-Two Week Low] & [Fifty-Two Week High]

Firms Screened [Tix*]	Spearman ρ	Sample Size	p-value
ADT	0.0919	283	0.1229
NWSA	–0.1616	274	0.0074
QEP	–0.3819	310	<0.0001
WU	0.1264	352	0.0196
ZTS	0.0569	279	0.3436

*See the Appendix for the SIC and URL information for these firms

The statistical profile of the 495 firms not screened was:

Table 2. Firms Not Screened: In Spearman Profile Respecting the [Fifty-Two Week Low] & [Fifty-Two Week High] Screen

Remaining Firms In Profile	Results
Number of Firms	495
Range of ρ	[0.1728 to 0.9999]
Median/Mode of p-values on ρ	<0.0001 / <0.0001
Range of p-values	[.0008 to <0.0001]

Next we executed the same screen protocol on the modified CCD [n = 495] using the variable set: [Capital Cube Price Range Min] & [Capital Cube Price Range Max]. The results were that no firms had a ρ < +0.15. The profile of these 495 firms was:

Table 3. The Spearman Profile Respecting the [Capital Cube Price Range Min] & [Capital Cube Price Range Max] Screen

Firms In Profile	Results
Number of Firms	495
Range of ρ	[0.1771 to 0.9999]
Median/Mode of p-values on ρ	<0.0001 / <0.0001
Range of p-values	[.0006 to <0.0001]

3.3 COMPUSTAT Profiling: Screening for Power Comparability The next screening protocol is to examine the number of firms that could support extended analysis. Specifically, at some point we will examine CapitalCube's extensive category index which has very creative category indices such as:

M&A Action Linguistic Coding: [Acquirer & Target}

Capital Investing Strategy Linguistic Coding: [Betting on the Future; Maintenance Mode; Milking the Business & Supporting Growth]

Borrowing Capacity Linguistic Coding: [Constrained; Limited Flexibility; Some Capacity & Quick and Able]

The related variable set for the comparative analysis of the category indices of the CCD will be a panel download from COMPUSTAT™ [WRDS™] [CuStP] where we have selected, *in round*, 30 market-performance variables in the rubric set: *Asset*, *Liability*, *Deb*t, *Return*, *Profitability* and *Nature of the Audit Opinion*; in this screening, we want to examine the number of firms in the COMPUSTAT and CCD panels that could provide a reasonable number of points over the various event spaces in the accrual period from 2005 to 2015. In this case, we examined the joint of the proportion of data points in CuSTP ∩ CCD. The screening criteria was that if either the CCD or the [CuStP] had *less than* 50% of the data populated where most all of the CCD had 371 data points and most of the [CuStP] had 17 [Yearly summary information], then this firm would be eliminated from the analysis. Of the 495 remaining after the first screening there were eight firms that lacked sufficient observations to form a rich comparative analysis set. These firms are: ABBV; ALLE; FB; KORS; KRFT; MNK; PSX & TRIP. [See the Appendix for the SIC and URL information.] This then removed these eight additional firms and the CCD now had 487 firms for analysis.

3.4 Screening Summary ***The Spearman screening*** served two purposes: The first is to remove firms that may be outliers respecting the general profile of the firms that one would expect to constitute the S&P500. Recall that we essentially removed firms that had longitudinal ρ-association that was clearly at variance with the empirical profile of the market over the panel. This screening moves in the direction of increasing the efficiency of the inferential dimension of the analysis in that it is almost always the case that screening "outliers" or, in our case, "noisy analytic objects" removes more variation than is lost in inferential of power from a reduction in sample size. The second purpose was a testing of the reasonably of performance benchmarking these CapitalCube context variables using the COMPUSTAT variables. Overall there were very few firms removed for the two Spearman screens. In fact, for the Spearman screens we lost only 1.0% of the firms in the CCD. This then suggests that variables offered by CapitalCube as context variables:

1. [Fifty-Two Week Low] & [Fifty-Two Week High];
2. [Capital Cube Price Range Min] & [Capital Cube Price Range Max]

are, in the main, four variables that fit the ex-post empirical evidence relative to the trend of the market for S&P500 firms. This suggests strongly that the four CapitalCube context measures as profiled through 99% of the firms in the S&P500 panel are in-sync with the independent empirical evidence of the panel. Here we are using concurrent validation in that the CapitalCube variables are being profiled with an independent filter, the S&P500, that has rigorous listing criteria. Indeed, the S&P500 is a Darwinian-index as the inclusion/listing bar is exclusive to only the best 500 firms. Recall that the CAPM modelers of the late 1960s selected the S&P500 as the benchmark due to the fact that firms on the S&P500 *a priori* were expected to be well-managed and "main-stays" in their SIC or NAICS grouping. According to cfTechnology: BrainBank: [www.cftech.com/BrainBank/FINANCE/SandP500Hist.html]

The Standard & Poor's 500 is a market-value-weighted index (shares outstanding multiplied by stock price) of 500 stocks that are traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and the NASDAQ National Market System. The weightings make each company's influence on Index performance directly proportional to that company's market value. - - - Companies selected for the Standard & Poor's 500 Index are not chosen because they are the largest companies in terms of market value, or sales, or profits. Rather, the companies

included in the Index tend to be representative of important industries within the U.S. economy and many also are the leaders of their industries. When the U.S. Department of Commerce developed its Index of Leading Economic Indicators in 1968 to signal potential turning points in the national economy, it chose the S&P 500 Index as one of the components.

Until recent years, the DJIA was the only stock market indicator widely quoted by the general news media. The Standard & Poor's 500 was rarely mentioned in news summaries, although the Index was the most common benchmark used by investment professionals to measure the performance of their portfolios. That use of the S&P 500 as the proxy for the overall stock market predates the widespread adoption of the Capital Asset Pricing Model (CAPM) in the 1970s. By convention, the S&P 500 Index was used as the market portfolio in tests of the CAPM. Betas of individual stocks were then calculated against the S&P 500 Index, which by definition had a portfolio beta of 1.00. As the amount of money invested in the equity markets grew, the need for a broad market indicator that reflected how people actually invest in equities was needed.

The S&P-listing protocol is found at:[http://www.standardandpoors.com/ratings/articles/en/us/?assetID=1245218657627**]**

Simply, the CapitalCube context variables are empirically in-sync with the *a priori* expectation of firm behavior in a market now controlled by AS 2 and AS 5 that are the audit standard rules promulgated by the Public Company Accounting Oversite Board [PCAOB] under the Federal legislation of Sarbanes-Oxley:2002 [HR: Source: U.S. Congress, *Sarbanes-Oxley Act of 2002*, Pub. L. 107-204, 116 Stat/ 745 (2002)]. This is another way of affirming that if we were to have found that these CapitalCube context variables eliminated 30 or 35% of the firms that would be strong evidence that these four context variables are not likely in tune with the usual market realties over the panel period. This then argues that indeed *these four **context variables** seem capable of market discrimination*.

As for the COMPUSTAT elimination this was a practical reality screen recognizing that for extended analyses using the COMPUSTAT variables, to be effected subsequently, we wanted to use variable sets of comparable power. Thus we eliminated eight firms that were not sufficiently populated with all the variable information. This then resulted in a CCD with 487 firms.

The next set of analyses will be addressed to the four ***decision-making variables***. We will test the reasonability of these decision-making variables by offering a set of associational relationships in much that same way that we did for the Spearman ρ-screens. Then we will test these associational expectations over the 487 firms in the CCD. Here we are not interested to screen a particular firm as we were in the previous screening analysis. In this case, we will look at these variables as profiled by the actual activity over the aggregation of firms in the CCD.

4. Preliminary Contextual Analysis: **Specific Profile Hypotheses for the Four CapitalCube Decision-Making Variables**

4.1 Hypotheses for Decision-Making Variables In the same condition-testing mode as was developed for the Context variable set we will now examine the following four CapitalCube Decision-making variables:

V1: *Current Price Level Annual* [CPLA]

V2: *Scaled Earnings Score Average Latest* [SESAL]

V3: *Previous Day Closing Price Latest* [PDCPL]

V4: *CapitalCube Price Latest* [CCPL]

In this case we will form the following *a-priori* hypotheses, present their rationalizations, and the inference protocol for the aggregate testing of the profiles of: {V1, V2, V3 & V4}.

H1*[Dual Conditioned]*: We expect that: (i) more than 50% of the S&P500 Firms in the CCD will have First Factor positive Loadings for V3 & V4, and (ii) both variable loadings will be greater than the Harman recommended cut-off of (5.)^.5 or +0.7072 which is the exclusive cut-point.

Justification We expect V3 & V4 [on a Firm by Firm basis in the S&P500 Panel] to be highly Pearson correlated as they both are: (i) current/latest and so should be "directionally in-sync" over the Panel as we saw from the previous context screening analysis, (ii) re-calibrated relative to splits & spins, (iii) most like a Poisson class of probability density functions, and (iv) likely Box-Jenkins [BJ] ARIMA[1or2, 0, 0]—i.e., linked by a very strong AR process as the expected market generating process over the accrual period has trended in a positive direction. If these four pre-conditions obtain, then V3 & V4 will be highly associated and so should consistently define the First Factor with positive loadings. ***The inference test*** will be the test of proportions where we have formed the Null as chance—i.e.,

50%—equal proportional loading over the Factors. In this case, any proportional loading of [V3 & V4] outside the (1-α)CI will rationalize the rejection of the Null of 50% in favor of support of H1. This will be a two-tailed test as high or low percentages of factor loadings as a proportion will suggest a relative exclusive loading either on Factor 1 (or perhaps on Factor 2). We have, however, formed H1 as V3 & V4 jointly loading on Factor 1; should the Harman projection fix V3 & V4 on Factor 2 we will note that this is consistent with the H1 as a related event occurrence.

H2 For V1: Current Price Level Annual [CPLA] we do not expect that V1 will systematically follow or favor a particular factor: **Justification** V1 is ratio-benchmarked by a slowly rolling annual range and is re-calibrated to form a (0 to1) range and so V1 is inherently smoothed in a Moving Average sense. Therefore, for relatively long panel sections in a relatively stable market there is likely to be Factor association on a firm basis of V1 with {V3 & V4} which are AR driven. In this case, some percentage of the time, V1 may align with the V3 & V4 as all three are market pricing variables and so over relatively long stable sections of the Panel the three variables {V1 & {V3 &V4}} are likely to exhibit AR-affinity resulting in all three grouping on the same Factor. In the case, where V1 is not in-sync with V3 or V4, then it most likely will align with V2:SESAL which is another smoothed variable and has as a driver, *Earnings*, which is likely to produce a lagged relationship to the latest price calibration of V3 and V4. This lag and smoothing for V2:SESAL, in shorter longitudinal segments of the Panel, may both disconnect from V3 & V4 and connect with V1 that is also smoothed. This may, depending on the relative segment-volatility of the market, produce an equal distribution over the two Factors for V1 as proffered. ***The inference test*** will be the one-tailed test of proportions where we have formed the Null as a Factor Loading on the First Factor of 80% or a relative high proportional loading profile on a particular Factor—in this case the first Factor. Any proportional loading of V1 outside on the LHS of the (1-α)CI[centered at 80%] will rationalize the rejection of the Null of unequal Factor profiles in favor of support of H2.

H3 As for V2: SESAL this is: (i) formed around the *Earnings* of the firm, (ii) calibrated by a number of variables, (iii) boxed into a range (1, 100), and (iv) lagged (respecting market price) variable and so may naturally detach from the price-driven variables: {V1, V3, V4]. As such its relative tracking may not follow the market pricing variables and so we expect ***to see V2 as predominately loading on the Second/Other Factor*** and from time to time for V1, also a boxed-price-driven variable to share the factor loading profile of V2 as argued above. ***The inference test*** will be the one-tailed test of proportions where we have formed the Null as a Factor Loading on the Second Factor of 50% or equal loading. Any proportional loading of V2 outside on the RHS of the (1-α)CI[centered at 50%] will rationalize the rejection of the Null as equal Factor profiles in favor of support of H3

In this regard, for inferential purposes, we will use the standard Harman (1967) Factor model as programmed in JMP/SAS v.12. ***Caveat***: *A priori* for the variable set {V1, V2, V3 & V4} we have proffered the relationships for H1, H2 & H3. These condition, then by extension, that the rotation space to form the factors will have two projection axes—i.e., two Factors. For the firms in the CCD, we tested the reasonability of this pre-inference condition. We created the eigenvalue profile for all of the 487 firms; the Mean of the total cumulative percentage explained for the second factor was 81.96%. For the first Factor alone the explanation percent was a little over 50% at 55.65%; this clearly rationalizes the reasonability and also the need for two factors in rotation as 80% will provide a robust inference profile. Further, we will use Harman's standard and recommended screening criteria: (i) Rotate using Varimax Orthogonal projections on Pearson Product moment correlations, and (ii) Index the Factor using the variables that load [in projection] > then .5^.5 which is an exclusive loading in that no other Factor can have a value greater than that loading. Here we will be using the Pearson model and not the Spearman model as this will give a conservative rendering if non-linear order is the case as expected. This means that non-linear relative placements will reduce the Pearson product moment correlations and so diminish slightly the eigenvalues. This seems wise as now the inference will not be on a Firm basis as it was in the Spearman screening context but rather overall. Therefore, support for the Hs, as formed, will be slightly conservative—a higher chance in the False Negative Error direction or a conservative rendering of the inference profiles.

At this juncture we will follow the advice of one of our pre-submission technical readers to include an illustrative graphical context for the Harman Factor Profiles.

4.2 Illustrative Examples For these examples, we will present graphical profiles of the variable sets as we have discussed them above relative to the Hs. Also, only for purposes of presentation and not for any of the analyses, we scale-aligned the CapitalCube decision-making variables to an average of 100 using the usual presentation transformation: *Each data point in its variable series is divided by the Mean of the variable and multiplied by 100*. This then results in all four of the series having Mean 100 which greatly facilitates reading their graphical presentation. We have randomly selected two firms WDC [Western Digital Corp [3572]] and VIAB [Viacom, Inc.

[4841]] for this illustrative presentation. We will examine the graphs of V1 & V2 and V3 & V4 and also give their individual Harman Factor Profiles [HFP]; this will offer a visual context for the variable dynamics.

Consider the firm WDC. The bi-variate graphs of V1 & V2: Figure 1 which are "boxed benchmarked variables" and V3 & V4: Figure 2 which are "free range" in nature compared to V1 & V2 over the longitudinal panel are most instructive and are typical of most of the firms in profile.

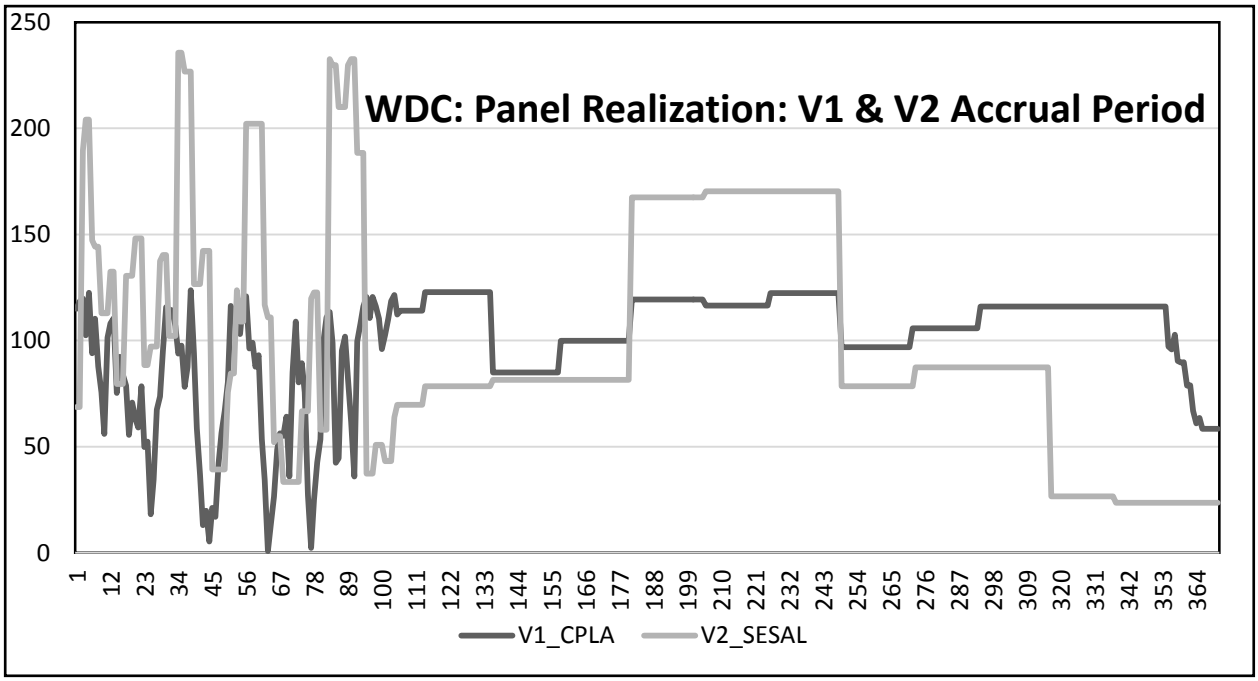

Figure 1. Panel Profile of WDC for V1: CPLA & V2: SESAL

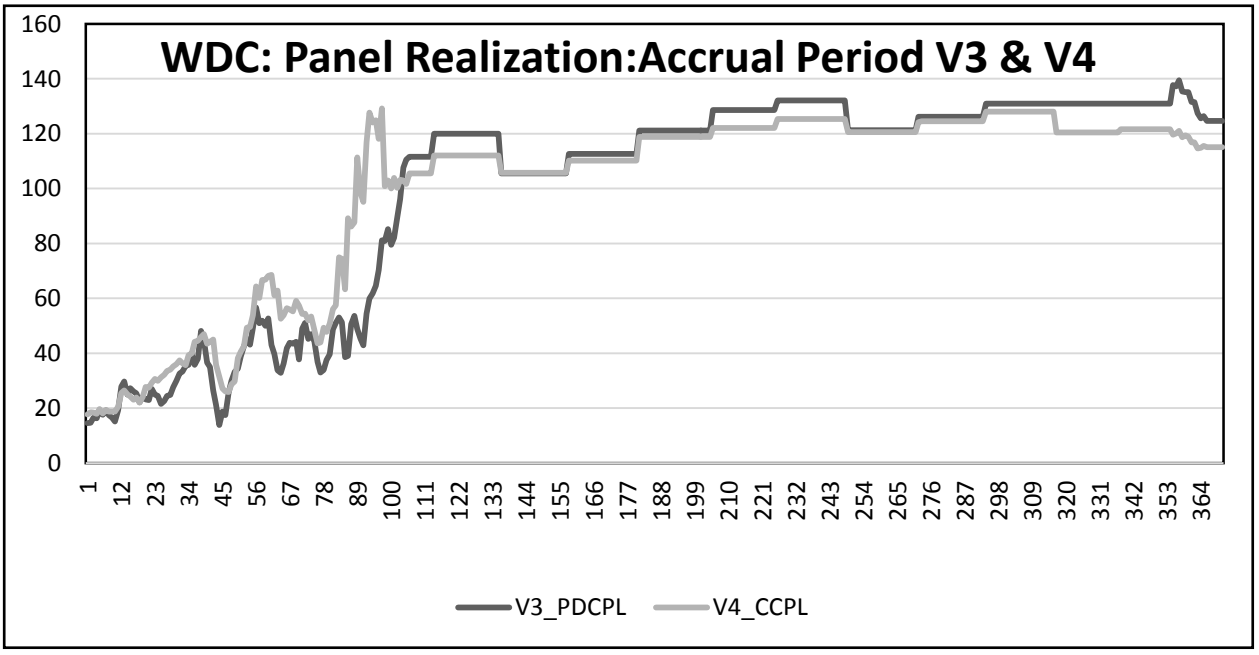

Figure 2. Panel profile of WDC for V3: PDCPL & V4: CCPL

Here we see that V1 and V2 which are boxed variables track very differently one from the other compared to V3 and V4 which track together and are in the BJ:(1or2, 0, 0) mode. Both V1 & V2 are more in the Moving average mode or

BJ: (0, 0, 1or2)—i.e., they both track along the horizon plane. V1 is in the median projection position and V2 has relatively high motion around the median track of V1. The motion of V1 seems more aligned with V3 & V4 than with V2. In the case of WDC we find the following HFP:

Table 4. The HFP Eigenvalue Profile of WDC for the Four Decision-making Variables

Variables	Loadings Factor 1	Loadings Factor 2	Variate Grouping	Eigenvalue
V1:CPLA	**0.8086**	0.2590	**First**	**2.4584(62%)[62%]**
V2:SESAL	-0.0571	**0.9708**	**Second**	**1.0548(26%)[88%]**
V3:PDCPL	**0.9451**	-0.2286	**Third**	0.4480(11%)**[99%]**
V4:CCPL	**0.9349**	-0.1645	**Fourth**	0.0388(1%)**[100%]**

Here we see the expectation that the V3 & V4 group together as they are relatively in lock-sync and exhibit a dramatic rise over run for the first third of the accrual period which, as expected, has a strong AR 1 or 2 effect. In this case V1:CPLA in its relative median projection disconnects with V2 and seems to align with the V3 &V4 grouping which plays out in the above HFP loadings. However, V2:SESA is clearly not aligned with V1:CPLA and by linkage therefore not in-sync group V3&V4 and so defines Factor two.

As an example of another factor profile, consider the firm VIAB. The graphs V3 & V4 for VIAB are relatively similar to those for WDC as presented in Figure 3. However, for V1 and V2 the BJ (0, 0, 1or2) horizontal plane orientation tracking pattern is reversed.

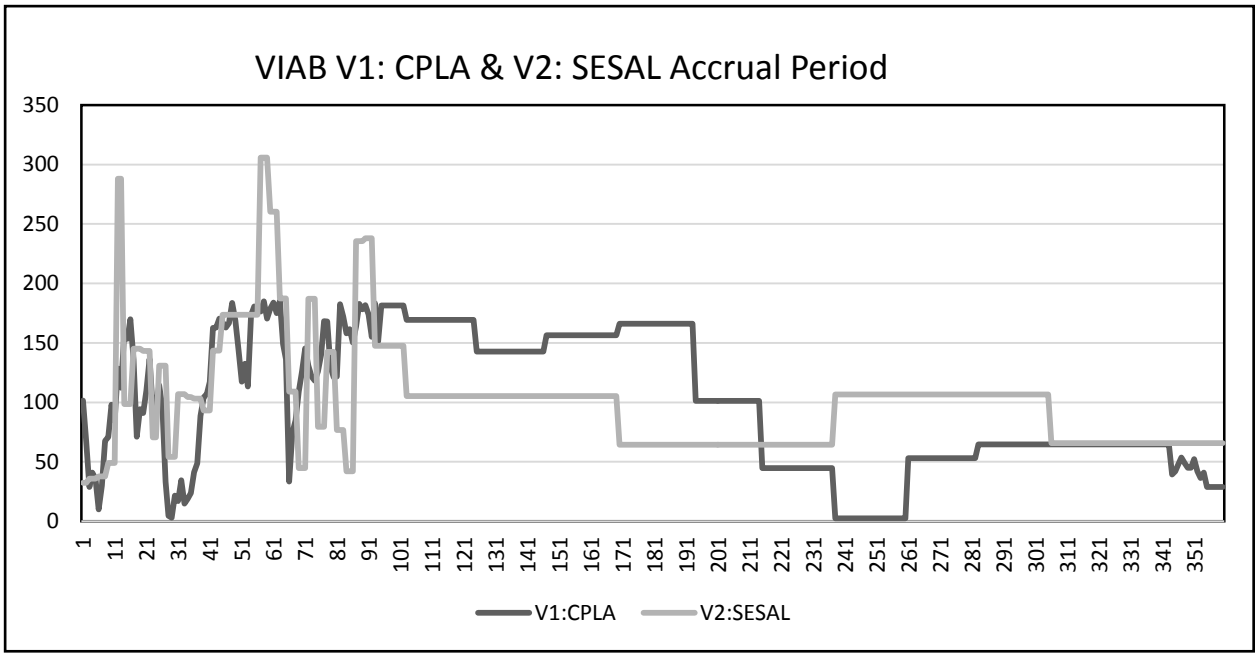

Figure 3. Panel Profile for VIAB of V1: CPLA & V2: SESAL

In this case it is V2: SESAL which is in the Median tracking position and V1: CPLA exhibits motion around V2. However, probably because the relative motion after the 100th point or so of V1 is not in-sync with the BJ: (1or2, 0, 0) modeling form of V3 &V4 but is associated with V2 likely a MA model form, then V1 & V2 are grouped together in association as they do not share the in-sync motion of V3 & V4. We see this in the following HFP:

Table 5. The HFP Eigenvalue Profile of VIAB for the Four Decision-making Variables

Variables	Loadings Factor 1	Loadings Factor 2	Variate Grouping	Eigenvalue
V1:CPLA	0.1345	**0.8086**	**First**	**2.1118(53%)[53%]**
V2:SESAL	0.0409	**0.9924**	**Second**	**1.2984(33%)[86%]**
V3:PDCPL	**0.9924**	-0.3217	**Third**	0.5658(14%)**[99%]**
V4:CCPL	**0.9543**	-0.1854	**Fourth**	0.0240(1%)**[100%]**

Having illustrated the panel profile which are typical for these variables, we now will provide the summary data for the firms in the CCD relative to the Hs. This will be the second test of the data reasonability so as to move to the next stage of the analysis which will be the testing phase on the category variables as profiles over the decision-making variables that we have tested in this paper.

4.3 Summary of the Factor Profile for the Decision-making Variables We ran for each of the 487 firms in the CCD the standard Harman Factor analysis as it is coded in the Multivariate module of JMP/SAS v.12. This data was captured in an Excel file and then processed using a VBA module to record:

1. The number of occurrences for: V3 & V4 on the First Factor where both are > than (5.)^.5 referred to as Harman Loading [HL] *as well as* on Factor 2.
2. The number of occurrences for V1 of HL on Factor 1 *as well as* The number of occurrences for V1 of HL on Factor 2
3. The number of HL occurrences for V2 on Factor 1 *as well as* on Factor 2.

The HL-profile as presented in Table 6:

Table 6. Harman Loading Profiles for the CapitalCube Decision-making Variables

Profile	Factor 1	Factor 2	Aggregate	Hypotheses
V1:CPLA	**40.7%**	**35.1%**	**75.8%**	**H2 Supported**
V2:SESAL	8.4%	**72.3%**	**80.7%**	**H3 Supported**
V3&V4	**82.5%**	1.6%	**84.2%**	**H1 Supported**

will then be used to test H1, H2 & H3.

H1 Results We used as the Null expectation 50%, meaning that the distribution of the HFP is equally distributed over the two factors. In this case, the two tailed fail-to-reject-the-Null i.e.,—fail-to-support-H1—region for a FPE of 5% is: [45.6% to 54.4%]. As the realization is that 82.5% of the time V3&V4 aligned in HFP on Factor 1 and 82.5% $\not\subset$ [45.6% to 54.4%] and is exterior on the RHS this provides strong support for rejecting the Null [p-value <0.001] and so suggesting that H1 is the likely case.

H2 Results In this case, we assumed for testing purposes that Null expectation was 80% for a HL on Factor 1. As presented in **Table 6** the actual distributions of V1:CPLA over Factors 1 & 2 are: 53.7% [40.7% /75.8%] and 46.3% [35.1%/75.8%]. Therefore, the population inference test interval for the Null of 80%—the fail-to-support H2 region—for a one-tailed FPE of 5% is: [74% to 86%]. The realizations for either Factor are outside this directional 95% CI and so we reject the Null in favor of support for H2. For example, 53.7% $\not\subset$ [74% to 86%] and, of course, the same is true for 46.3% which is (100% - 53.7%). H2 is strongly supported as the p-value for the Null is < 0.001.

H3 Results We used the population Null expectation of 50% meaning that the distribution of the HFP of V2:SESAL is equally distributed over the two factors. The fail-to-support the one-tailed 5% FPE H3 region is: [46.3% to 53.7%]. In this case, we find that 72.3% $\not\subset$ [46.3% to 53.7%] and is RHS in orientation and so provides strong support [p-value <0.001] for rejecting the Null suggesting that H3 is the likely case.

5. Context, Overall Summary, and Conclusion

5.1 Context In this paper we have presented a detailed examination of the principal variables that are mainstays of the CapitalCube market analytics platform as expressed through the S&P500 Panel. Following on the work of Kwon, Lee & Shin (2014), our analyses addressed the *Necessary Condition* testing of any analytic-navigation platform: *As a pre-condition to engaging analytic-platforms to ferret out useful decision-making relationships one needs reasonable assurance that the variable data sets are relevant and reliable in the market context of their realization.* Therefore, we focused this, the first in a series of papers, on the empirical profile of the dynamics of the CapitalCube variables to address *the fundamental question* underlying any inferential analytics in the Big Data world; to wit:

For the inference profile that was produced by the analytic platforms through which I passed my data capture do I have confidence that the usual FPE and FNE conditionals are realistic to guide my decision-making?

This question is a PC-repackaging of the age-old adage: ***Garbage-In Garbage-Out***. This obligatory analytic "heads-ups" is, now more than ever, in play in the Big Data world where spurious associations muddy the waters of inferential profiling by creating a paradox of structure—i.e., the *Type 3* error: Believing in the Wrong Model. In the paradox space, the FPE and the FNE are not indications of inferential guidance that leads to useful inference the expected percentage of the time. See also, Gandomi & Haider (2015). An example will here be instructive.

Assume that we are in the *Olympia Stadium*, Berlin Germany for the Final of the 2015 Champions League match between *Barcelona FC* and *Juventus FC*. There are 74,442 in attendance. An Analyst asks everyone to stand-up and to predict if a coin flip will result in a *Head* or a *Tail*. If they are wrong they are to sit down. The first flip is a Tail so those who predicted a Head sit down. The second flip is a Head and all of those who guessed Tails sit down. This continues for 11 flips. There are now 36 people standing. The analyst interviews these individuals so as to glean the underlying fundamentals of their predictive process so as to develop an DSS for Predicting Coin Flips reasoning that these 36 individuals must be experts in predicting coin flips as their chance of guessing right 11 times in a row for a fair coin has a p-value < 0.0005. *Moral: There are many apparent patterns in the Big Data world even when random chance is the data generating process. So the first pre-inferential screen is to rule out if the data-driver is random chance.* See also the excellent related example offered by Pinder (2014).

This was the motivation of this report on the CapitalCube Data Capture. So before we engage the CapitalCube Dataset to determine its information content or utility, we first examined the reasonability of the variables as expressed through the S&P500 the results of which are detailed following.

5.1.1 Spearman Screening Results Using the ***Context Variable*** pairings:

1. [Fifty-Two Week Low] & [Fifty-Two Week High];
2. [Capital Cube Price Range Min] & [Capital Cube Price Range Max],

we screened only 1% of the firms in the S&P500 Panel because they failed to exhibit rational measured-value patterns for these CapitalCube context parings. The fact that only 1% of the firms in the S&P500 Panel were screened as outliers on the two CapitalCube screening variable sets is strong evidence that the context variables are dynamically in-sync with the empirical profile as we know it for the tracking of the S&P500 over the accrual period.

5.1.2 Factor Profiling of the CapitalCube ***Decision-making Variables*** Using the variables:

V1: *Current Price Level Annual* [CPLA]

V2: *Scaled Earnings Score Average Latest* [SESAL]

V3: *Previous Day Closing Price Latest* [PDCPL]

V4: *CapitalCube Price Latest* [CCPL]

and hypothesis-driven factor screens, we examined the reasonability of these four variables as expressed through the S&P500 Panel. As we rejected the Nulls of the expectations formed for the three empirical validation hypotheses in support of the factor profiling hypotheses, there is strong evidence for the support of the expected structural nature of {V1, V2, V3 & V4} in that they behave in an expected manner given the usual AR & Fixed-Effects character of a Panel of traded Firms.

5.2 Conclusion One may reject, with a high degree of assurance, the random or chance generating process as the driver of these eight variables in that they behave as one would expect for a Panel of traded firms. **Implication: *The CapitalCube variable set, herein examined, is structurally in-sync with the expected market generating process(es) and therefore, in this sense, the CapitalCube variable set represents variables from which longitudinal market performance information can likely be gleaned.***

Acknowledgments

We wish to thank Dr. H. Wright, *Boston University*: Department of Mathematics and Statistics, the participants at the SBE Research Workshop at SUNY: Plattsburgh, Mr. Manuel Bern, *Deloitte Touche* LLP, Audit & Assurance Services, Frankfurt [Main], Germany, and the two anonymous reviewers from *Accounting and Finance Research* for their detailed comments and suggestions.

References

Akkaya, G., & Uzar, C. (2011). Data mining: Concept, techniques and applications. *Global Science and Technology Forum*, *1*, 47-50.

Diebold, F. (2014). On the origin(s) and development of the term "Big Data". *Penn Institute for Economic Research*, *12*, 1-6.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*, 293-314. http://dx.doi.org/10.1093/nsr/nwt032

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management*, *35*, 137-144. http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007

Lusk, E., & Halperin, M. (2015). Navigating the Benford labyrinth: A Big-Data analytic protocol illustrated using the Academic Library Context, *Knowledge Management & E-Learning*, forthcoming.

Harman, H. (1967). *Modern factor analysis*, 2nd Ed. University of Chicago Press, Chicago, USA.

Kimble, C., & Milolidakis G. (2015). Big data and business intelligence: Debunking the myths. *Global Business & Organizational Excellence*, *35*, 23-34. http://dx.doi.org/10.1002/joe.21642

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics, *International Journal of Information Management, 34*, 387-403. http://dx.doi.org/10.1016/j.ijinfomgt.2014.02.002

Lovell, M. (1983). Data Mining. *Review of Economics and Statistics, 65*, 1-12. http://dx.doi.org/10.2307/1924403

Niessing, J., & Walker, J. (2015). The eight most common Big Data myths, INSEAD Publication Series, *INSEAD Articles*; Fontainebleau, FR, Web Access: ABI/INFORM Global; 19March2015

Pinder, J.P. (2014). A demonstration of regression false positive selection in data mining. *Decision Sciences Journal of Innovative Education*, *12*, 199-217. http://dx.doi.org/10.1111/dsji.12037

Porter, L., & Gogan, J. (2013). Before racing up Big Data Mountain, look around. *Financial Executive*, *6*, 59-61.

Slagter, K., Hsu, C-H., & Chung, Y-C. (2015). An adaptive and memory efficient sampling mechanism for partitioning in MapReduce. *International Journal of Parallel Programming*, *43*, 489-507. http://dx.doi.org/10.1007/s10766-013-0288-z

van Bussel, G., Smit, N., & van de Pas, J. (2015). Digital archiving, green IT and environment. Deleting data to manage critical effects of the data deluge. *Electronic Journal of Information Systems Evaluation*, *18*, 187-197.

Yang, H., & Fong, S. (2015). Countering the concept-drift problems in big data by an incrementally optimized stream mining model. *Journal of Systems and Software*, *102*, 158-166. http://dx.doi.org/10.1109/bigdata.congress.2013.25

Notes

Note 1. All of these four decision-making variables are the intellectual property of the AI&CC. Mutually, we agreed to set aside the specifics of the functionalities of these intellectual property variables. The COO however was agreeable to giving informative disclosure as to the nature of the creation of the variables. This is what is reported in this section.

Note 2. There was a ninth point process variable: *Dividend Yield Latest*. For this variable there were a relatively high number of missing and zero values in the CCD. For this reason we did not use this relatively "noisy" variable in the analysis to be reported.

Appendix Tickers, SIC and URL Information for Removed Firms

Ticker[EDGAR]	SIC	HomePage URL
ABBV	2834	http://www.abbvie.com/
ALLE	6381	http://www.allegion.com/corp/en/home.html
FB	7370	https://www.facebook.com/
KORS	3100	http://www.michaelkors.com/
KRFT	2000	http://www.kraftfoodsgroup.com/home/index.aspx
MNK	2834	http://www.mallinckrodt.com/
PSX	2911	http://www.phillips66.com/EN/Pages/index.aspx
TRIP	7370	http://ir.tripadvisor.com/
ADT	7381	http://www.adt.com/
NWSA	2711	http://newscorp.com/
QEP	1311	http://www.qepres.com/
WU	7389	http://corporate.westernunion.com/
ZTS	2834	https://www.zoetis.com/