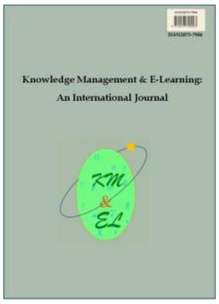
Knowledge Management & E-Learning, Vol.8, No.1. Mar 2016

Knowledge Management & E-Learning



ISSN 2073-7904

Navigating the Benford Labyrinth: A big-data analytic protocol illustrated using the academic library context

Michael Halperin
University of Pennsylvania, Philadelphia, PA, USA
Edward J. Lusk
State University of New York, Plattsburgh, NY, USA
University of Pennsylvania, Philadelphia, PA, USA

Navigating the Benford Labyrinth: A big-data analytic protocol illustrated using the academic library context

Michael Halperin

Lippincott Library
Wharton School of Business
University of Pennsylvania, Philadelphia, PA, USA
E-mail: halperin@upenn.wharton.edu

Edward J. Lusk*

Faculty of Business & Economics
State University of New York, Plattsburgh, NY, USA
Wharton School of Business
University of Pennsylvania, Philadelphia, PA, USA
E-mail: luskej@plattsburgh.edu or lusk@wharton.upenn.edu

*Corresponding author

Abstract: ***Objective*:** Big Data Analytics is a panoply of techniques the principal intention of which is to ferret out dimensions or factors from certain data streamed or available over the WWW. We offer a subset or “second” stage protocol of Big Data Analytics (BDA) that uses these dimensional datasets as benchmarks for profiling related data. We call this Specific Context Benchmarking (SCB). ***Method*:** In effecting this benchmarking objective, we have elected to use a Digital Frequency Profiling (DFP) technique based upon the work of Newcomb and Benford, who have developed a profiling benchmark based upon the Log10 function. We illustrate the various stages of the SCB protocol using the data produced by the Academic Research Libraries to enhance insights regarding the details of the operational benchmarking context and so offer generalizations needed to encourage adoption of SCB across other functional domains. ***Results*:** An illustration of the SCB protocol is offered using the recently developed Benford Practical Profile as the *Conformity Benchmarking Measure*. ***ShareWare*:** We have developed a Decision Support System called: SpecificContextAnalytics (SCA:DSS) to create the various information sets presented in this paper. The SCA:DSS, programmed in Excel™ VBA®, is available from the corresponding author as a free download without restriction to its use. ***Conclusions*:** We note that SCB effected using the DFPs is an enhancement not a replacement for the usual statistical and analytic techniques and fits very well in the BDA milieu.

Keywords: Big-data dataset preparation; Benford expectation intervals; Specific context benchmarking

Biographical notes: Dr. Michael Halperin is the former Director of the Lippincott Library, the Library of the Wharton School. He has published extensively and is co-author of two books: International Business Information and Research Guide to Corporate Acquisitions. He is the creator, with Penn Libraries' Delphine Khanna of the 'Business FAQ', a business knowledge database.

Dr. Edward J. Lusk is Professor of Accounting, the State University of New York (SUNY), College of Business and Economics, and Emeritus: the Department of Statistics, The Wharton School, The University of Pennsylvania. From 2001 to 2006 he held the Chair in Business Administration at the Otto-von-Guericke University, Magdeburg Germany.

1. Introduction: How did we arrive at the big data era?

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) to form such data into decision relevant information. However, according to Lovell (1983) and Porter and 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. In short order, there were thousands of articles on *Data Mining*. For example, we conducted a search on the *Web of Science*™ using the single term Data Mining. From 1992 to 1994 there were ten articles identified; whereas from: 1997 to 1999 there were more than 500 articles in evidence! Many of these Data Mining articles detailed examples of General User Interface (GUI) protocols for creating relevant and reliable dimensional or factor information. This GUI developmental stage was needed as according to Slagter, Hsu, and Chung (2015, p. 489):

> "*Big Data refers to the massive amounts of structured and unstructured data being produced every day from a wide range of sources. Big Data is difficult to work with and needs a large number of machines to process it, as well as software capable of running in a distributed environment.*"

Diebold (2014, p.5), who is principally responsible for coining/popularizing the term Big Data, offers the following regarding the next evolutionary moment leading from Data Mining to 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.*"

1.1. Point of departure

Clear is that the evolutionary trajectory that has led us to Big Data Analytics comes from solid Data Mining roots. The principal thrust of research spawned by Data Mining and now fixed in the discipline area of Big Data Analytics (BDA) has been to address extracting dimensional foci; in this regard, the recent work of Gandomi and Haider (2015); and Yang and Fong (2015) offer insights into the *raison d'être* of BDA and also summarize the technical aspects of the abstracting functionalities employed to glean dimensions of potential interest from the Big Data stream by treating the statistical refinements of the dimensional abstraction so as to avoid the *bane* of Big Data analytics: *spurious association*. Our perspective is slightly different; we are interested in using these BDA-abstracted dimensions as benchmarks for profiling specific related datasets. We call this Specific Context Benchmarking (SCB). SCB is, of course, a subset of the "free-range" Big Data environment, where essentially all the WWW-data "streamed" are possible inputs to the BDA-sifting algorithms such as MapReduce™ as detailed by Slagter, Hsu, and Chung (2015). For SCB, which is our Big Data "carve-out", we elect to focus on creating profiles through benchmarking a specific *a priori* created comparison group using dimensionally derived "peer" datasets. In this regard, we are guided by the work of Akkaya and Uzar (2011, p.49) who offer three essential elements of best-practices BDA which are germane to developing our SCB protocol: (i) identifying and focusing on the *Target* data, (ii) selecting the relevant measurable *Variable Set* and (iii) moving the study forward from *A-Priori Expectations* to a focused set of conclusions. Using this guidance we will:

A. Offer *Benchmarking* as a modeling or profiling focus; benchmarking has started to appear in the literature but to date has not found currency in either Data Mining or Big Data Analytics. As indicated above our benchmarking protocol is called *Specific Context Benchmarking* (SCB). True, benchmarking is certainly not a new analytic concept; however, benchmarking, common though it is, is not employed *per se* as a staple in BDA. Case in point, we searched on ProQuest™ through ABI/INFORM™ as found on WRDS™ on 16 March 2015 using only the terms: Big-Data (*AND*) Benchmark* in the Abstract section and retrieved only 16 articles, the first appearing in 2013. This suggests that the concept of benchmarking is starting to find application.

B. Offer an extension of *Digital Frequency (DF) Testing* often found in Data Mining protocols where we will use a DF screening interval for profiling datasets that has relevance as part of BDA. The screening of Big-Data information sets using Digital Frequency methods has been used extensively and most successfully in forensic studies. See Nigrini (1996); Tam Cho and Gaines (2007); and Rauch, Göttsche, Brähler, and Engel (2011). We are using these DF profile techniques that have been validated in forensic analyses to provide comparative profiles that will offer perspective to the analyst in the Big Data context.

C. Specifically, combining Benchmarking and Digital Frequency Screening we will develop a five-stage protocol for *Specific Context Benchmarking* and illustrate its various functionalities using the voluminous member data produced by the Academic Research Library (ARL) Association. This illustration is central to our study as it offers operational details that are readily transferable across domains.

1.2. Caveat

It is important to bear in mind, as proffered above, that we are not focusing on tools for *sifting the massive volume of e-data* the intention of which is to Zip-Load & Dimensionally Organize thousands of Terabytes of digitized data points such as Near-Real time Stock trading Algorithms popularized by Das, Hanson, Kephart, & Tesauro (2001). Our Big Data focus is formed not on massive streamed datasets but on many large, possibly massive, "population" sized datasets that are "peer" datasets used to benchmark a related dataset for a specific analytic purpose. This contrast is: Big-Data that was birthed by Data-Mining often is used in a discovery mode—to wit: to ferret out data variable relationships ensconced in the Big Data stream. SCB is born of curiosity about possible *a priori* posited relationships of peer selected groups. Therefore, we are focusing on techniques that promote developing a context for further consideration of tested data profiling relationships. This is consistent and effectively motivated by the work of Porter and Gogan (2013, p.59) who note as an important counter-point to the "hype" born of unbridled Big-Data enthusiasm:

> *"Despite media proclamations that big data leaders are already miles ahead, it could be perilous to a company's financial health to try too much too soon. Before scaling the heights of big data, know where the company stands."*

Consider now the Academic Research Library milieu that we will use to illustrate our five-stage SCB-Data protocol.

2. The illustrative context: The big data of the Association of Academic Research Libraries

We have selected to start with a particular case example using the voluminous longitudinal datasets of the Association of Academic Research Libraries (ARL) and to simultaneously build the Specific Context Benchmarking protocol around these ARL datasets. This will enrich, we hope, the exposition and provide, to the readers, more direct access to the concepts of the SCB protocol. Further, after we examine the SCB profiles, we will suggest that these SCB results should be viewed in the light of related statistical analyses so as to enhance the decision relevance of the SCB results. This will be presented in section 5 following.

To be sure, the ARL is a target of opportunity as we have access to these datasets and are most familiar with the Academic Library as an organization; however, the SCB generalizes directly to many other decision-making domains. We will return to this generalizability in the summary section.

2.1. The Association of Academic Research Libraries

The ARL is currently an organization of 126 libraries in the U.S. and Canada. The membership consists of 115 university libraries and 11 public, governmental or nonprofit research libraries. The ARL began collecting and publishing annual data for members in Academic Year: 1961-62. The ARL also makes available annual statistics for university libraries from 1908 to 1962 that were collected by James Gerould, first at the University of Minnesota and later at Princeton University. The ARL statistics are the oldest and most comprehensive library statistics in North America. Currently, they consist of approximately 50 data series. The data is usually grouped as follows:

- Measures of Library Stock: e.g., Collection size and Components
- Measures of Services: e.g., Circulation, Interlibrary Loan, Reference Services
- Library Budget Components: e.g., Expenditures for Salaries, Materials, etc.
- University Statistics: e.g., Numbers of Faculty and Students

We offer that benchmarking is a pivotal decision-making function for the production of such aggregate ARL information. For example, the ARL statistics are frequently used by member and non-member libraries for comparative analyses. Directors of Academic Libraries use the ARL data to: compare their performance with peer institutions, look for trends in expenditure for materials over time, and in particular, justify budget requests. The ARL publishes, as do most industry groups, an annual "Investment Index" using factor principal component scores derived from membership data. The Investment Index is published annually in the *Chronicle of Higher Education* http://chronicle.com/article/Spending-by-University/140753/. For a comprehensive discussion of the Investment Index and details on its use see: Brinley, Cook, Kyrillidou, and Thompson (2010).

There is a perception among library administrators that the ARL statistics, with their emphasis on collection size and output measures, do not provide adequate assessment of process oriented metrics. See the illuminating discussions of this topic offered by: Brinley, Cook, Kyrillidou, and Thompson (2010), Oakleaf (2010) Report, and Koltay and Li (2010). Additionally, according to the excellent benchmarking study of Lewin and Passonneau (2012) and, consistent with our presumption introduced above regarding SCB, there seems a dearth of modeling protocols to view the activity of an ARL in "comparative" relief for purposes of creating decision-making information leading to systemic change-initiatives. Recall in a *Specific Context Benchmarking* protocol one group of institutions in the Big Data aggregate dataset is used to benchmark or create a "profile-in-relief" relative to the data of another sub-group.

The lack of SCB profiling in the ARL Big Data context is surprising because the ARL, as an association, produces a copious amount of summary ARL statistical information over a wide spectrum of activities on a yearly basis. One may perhaps find it anomalous as suggested by Lewin and Passonneau (2012) that academic research librarians, usually a data-driven group of curiosity seekers, have not taken advantage of the plethora of the ARL-summary Big Data population for benchmarking particular activity sets.

We submit that the reasons for the lack of SCB activities in many Big-Data analyses are that: (i) almost by definition industry- or domain-wide Big-Data sets often go beyond the "Apples & Oranges" metaphor; they are by extrapolation "Fruit-Salad", comprised, in the aggregate, of statistics contributed by: Public, Private, and Society enterprises of varying sizes with regional and global dispersion, and (ii) for an individual group desirous of SCB rarely is there a sufficiently long longitudinal time stream of non-event perturbed data to give credence to a benchmarking profile differential.

2.2. Pre-analysis data conditioning

In this regard, recalling the Data Mining discussion of Akkaya and Uzar (2011) and Porter and Gogan (2013), we offer, as a focused extension, that benchmarking SCB protocols in the Big Data milieu require two facilitating data forming actions to create relevant and reliable profile differentials:

A. ***Reasonable Homogeneity*** Certain aggregations from disparate contribution sources, such as those typically found in the Big Data context, may need to be *screened out* so that the benchmarking data-stream used as the SCB profiling contrast is from a generating process that is *en genre* similar to the expected or desired set for the individual group creating the benchmarking profile. *This will then require disaggregation of the peer Big Data "dataset" to achieve the expected profiling homogeneity.* This is not that dissimilar from the sifting algorithms employed in the search for relevant dimensionality. It is, however, not algorithmically driven through a factor model but rather effected by the judgmental intention of the analyst relative to the information to be profiled from the comparative analysis.

B. ***Sufficient Data for Reasonable Inference*** The fundamental assumption underlying benchmarking is to have sufficient data points in both the individual data stream and the selected aggregate benchmark so that the central tendency of their differential is a meaningful reflection of a profiling contrast. However, there is rarely sufficient data for one organization alone to create a rich DF profile. This being the case, there is usually a need to form a group of organizations or a consortium to provide sufficient observations against which to contrast the disaggregated dataset developed in A. above. *This will then require aggregation of individual sources or enterprises drawn from the Big Data milieu to achieve meaningful profiling differentials for the consortium contract with the peer benchmark.*

For our illustrative ARL context, these pre-conditions require that: (i) the ARL specific context benchmark dataset of 126 member libraries will be *dis-aggregated* in the service of *meaningful homogeneity* and (ii) the individual library developing the SCB will associate itself with a meaningful sub-group from the ARL Big Data set of “similar” institutions and use this *aggregated* data as an *analytic “consortium”* in the service of *sufficient data to effect a meaningful benchmark*. Consider now the metric for SCB.

3. Digital frequency profiling: The metric montage for big data reflective profiling

3.1. The measurement metric

An important issue in effecting SCB profiling is: *What metric can be used to facilitate the SCB analytics?* We wish to bring forward from the Data Mining literature an innovative measure called Digital Frequency Profiling that merits inclusion in the panoply of techniques in the Big Data context. See Tam Cho and Gaines (2007) and Kelly (2011) for a discussion of the applicability of Digital Frequency Profiling in Mining examinations.

We wish to give an interesting historical context to introduce Digital Frequency Profiling (DFP). The basis of this profiling technique, a mainstay in the forensic context, was first suggested by Newcomb (1881) and later by Benford (1938). It all begins when Simon Newcomb, mathematician and renowned astronomer, noticed that his book of tables of logarithms, the DSS of the day, with low numbers had pages that were more worn than those pages with higher numbers. Newcomb (1881, p. 39) observes:

> “*That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.*”

Fifty years or so later Benford (1938), an electrical engineer with *General Electric Inc.* with many patents to his credit, who curiously never cites Newcomb, makes and records the same observation. Benford examined thousands of numerical observations as varied as the population of cities, death rates, and physical constants. Newcomb and Benford both arrived at a simple formula to characterize the likely distribution of the nine first digits. To wit the (N-B Profile):

$$\text{Frequency}[d_i] = \text{LOG}_{10}\,(1 + 1/d_i) \text{ for i = 1, 2, - - -, 9} \qquad \text{EQ(1)}$$

This simple formula for forming a DF profile remarkably has been part of the historical record for more than a century! However, only recently has its theoretic underpinning been established as a reasonable surrogate for the generating process the measure of which is the digital frequency profiles produced by EQ(1). The preponderance of this research is due to Hill (1995a; 1995b; 1996; 1998) and Fewster (2009). They show by convincing theoretical argumentation and illustration that the following two conditions seem to result in data profiles, the first digital pattern of which follows, in the main, the Log$_{10}$ formula: *(i) datasets are formed from many different sources [mixing] or, (ii) a kernel data-generating process is subjected to various idiosyncratic constraints that results in base-invariances [scale invariance]*. We shall term this as *Hill-Conformity*.

3.2. A practical extension of the Log$_{10}$ profile

There is an alternative benchmark, due, in fact, to Benford. To give operational validity to the Log$_{10}$ generating function, Benford (1938, Table 1, p. 553) collected 20 samples from an impressive spectrum of generating processes, such as: *River Areas*, *Economic Costs*, and *Atomic Weights* to mention a few. The number of observations, in total, for these 20 datasets is 20 229. The range of the sample sizes for the 20 accruals is [91 to 5000] with a mean of 1012. Therefore, these frequencies as "a realization-profile" could also be used as a benchmark for the *Observed Digital Frequency* profile. However, due to recent research of Lusk and Halperin (2014a), it was reported that the mean frequency profile reported by Benford (1938, Table 1, p. 553) may be refined. Lusk and Halperin (2014a) use this practical dataset developed by Benford to form a screening interval, called the Benford Practical Profile (BPP), which is presented in Table 1.

Table 1
Screening boundary limits for the BPP

First Digit Array	Corrected Means of Benford Datasets, n=20	Lower Benford Screening Window (BSW) Value	Upper Benford Screening Window (BSW) Value
Digit 1	0.289189	0.275377	0.303001
Digit 2	0.194622	0.179919	0.209324
Digit 3	0.126650	0.111340	0.141960
Digit 4	0.090612	0.074990	0.106235
Digit 5	0.075436	0.059684	0.091189
Digit 6	0.064314	0.048467	0.080161
Digit 7	0.054081	0.038147	0.070014
Digit 8	0.054872	0.038945	0.070798
Digit 9	0.050522	0.034558	0.066485

These two benchmarks, the BPP and the Log$_{10}$, are not surprisingly, substantially similar; for example, the sum of the differences over the nine first digits for the BPP (Col2 of Table1) and EQ1 is 0.000298 and the distribution of the signs is as equal as is possible. *For purposes of SCB the decision-maker could use either as the validation* *benchmark*; the main issue is to stay with one benchmark and not to switch between the two for particular analyses or over time. Between the two, our recommendation is to use the BPP as:

1. The BPP was derived using Benford's 20 datasets that were realizations from many different experiential—i.e., real "contexts" –and so embodies the natural variation that may aid the SCB analyst in focusing on practical differences in comparative profiles, and
2. Using the Benford datasets an interval screening test, see Table 1 (BSW: Cols 3 & 4), developed by Lusk and Halperin (2014a; 2014b; 2014c) will greatly facilitate profile differentiation.

As an important point of information, these nine screening digital confidence intervals do not have individual unconditioned statistical properties, See Hill (1995b); for this reason Lusk and Halperin (2014a) have formed a heuristic test using datasets expected to be non-conforming datasets. They find that if overall more than 65.7%, of the individual digits profiled fall outside of the nine Benford confidence intervals of Table 1 then the dataset is likely to be non-conforming. We will note this as the Benford Practical Screening Heuristic (BPSH).

In addition, to this screening procedure there is an inferential measure that can be used in profiling called: the chi-square analysis of the SCB of the DFPs. This is the standard frequency comparison of the two profiles whereas the BPSH is the individual screening for each of the datasets. In the chi-square analysis the frequency profile of the Consortium and the Benchmark are compared; where there are major digital frequency differences the overall chi-square measure can be used to draw an inference if the two datasets are likely to have come from the same population DF profile. We will elaborate on this inferential testing as we present the ARL illustration.

4. Aims: The creation and illustration of a big-data profiling protocol

Following we will introduce the final two components to the Big Data protocol to be used in SCB profiling: The Quadrangle of Profiling Contrasts and the Profiling Screening Recommendations.

4.1. The quadrangle of profiling contrasts

To be sure, and as a clarification of the intent of such SCB reflective pondering, we are NOT only looking for *Non-Conformity* between the data reported in the benchmarking sources and a particularity consortium activity set. In reviewing the SCB literature on digital frequency profiling, there seems to be a predilection to focus on *Non-Conformity* as rationalizing reflective brainstorming that may lead to investigative activities and finally to systemic interventions. There are a number of studies the focus of which is exclusively *Non-Conformity* of the observed profile relative to the DF benchmark. This thread of inquiry was essentially started by Newcomb (1881) and enabled by Benford (1938) where the focus was on *Non-Conformity* and continues relatively unabated. See: Nigrini (1996; 1999); Ley (1996); Hill (1998); Geyer and Williamson-Pepple (2004); Tam Cho and Gaines (2007); Hickman and Rice (2010); and Reddy and Sebastin (2012).

This focus on *Non-Conformity*, as an investigative aberration, we feel misses the point of reflective benchmarking which is:

> *To generate an information profile from the Big Data set of information as a comparison relative to an* ***a priori*** *expectation either for a benchmarking profile where there is expected Non-Conformity* ***or*** *alternatively Conformity.*

Consider the following Table 2 where the exhaustive sets of foci that may be productively treated are summarized:

Table 2
The exhaustive investigative quadrangle of action plan profiles

	Expected *Conformity*	Expected *Non-Conformity*
Actual *Conformity*	No Investigative Actions	Investigative Actions
Actual *Non-Conformity*	Investigative Actions	No Investigative Actions

With such flexibility, the analyst can ask: *What do I learn from the comparative profiling?* For example, consider the action scripted in quadrant (Actual *Conformity*, Expected *Non-Conformity*). Referencing our ARL illustrative context, assume that we have as the consortium the Ivy League ARLs and for the benchmark the ARL aggregate dataset: *ARL Reported Professional Salaries from 2000 to 2013* where we have screened out AR-libraries not, in nature, similar to our Ivy League group—e.g., public libraries. If our ARL consortium group aligns well in SCB terms with the benchmark but it was our *a priori* expectation/desire that we should not conform to the ARL aggregate benchmarking dataset, then that could be a signal that the processes at the consortium level are NOT working as desired as we do not expect that we should profile as conforming to the ARL benchmarking activity set. This unexpected *Conformity* would have us consider actions of organizational re-deployment of key resources or other logistical considerations.

4.2. Profiling screening recommendations

The last component of the Big Data montage is the Testing Taxonomy to classify these Aggregate Datasets as Conforming or Non-Conforming. Above we have indicated that there are two profiling contexts: (i) The N-B Log$_{10}$ which is essentially a context-free theoretical functionality profile or (ii) the BPP suggested by Lusk and Halperin (2014a). Additionally, there are two screening modalities: (i) the screening intervals formed by the Benford aggregation of 20 disparate sampled datasets that present inherent variation that can be used to form a practical screening interval, to wit: the BPSH or (ii) the chi-square inference measure. Finally, there are two ways that the dataset comparisons can be effected using the chi-square inference measure: (i) tested as random samples one against the other, or (ii) the consortium data benchmarked directly against an ARL dataset. As there are a number of benchmarking profiles that can be put into play, we wish to narrow the focus and select the profiling set that we suggest as effective and also efficient for creation of profiling information. The following schema is our suggested taxonomy (see Table3):

Table 3
Schema for big data DF-profiling

	Log_{10} or BPP Benchmark	Benford CIs as the Inference Screen:BPSH	Chi-square Inference Measure	
			Two Random Samples	ARL as the Benchmark
Aggregated Consortium Dataset	BPP Preferred	Preferred	N/A	N/A
Disaggregated ALR Dataset	BPP Preferred	Preferred	N/A	N/A
Aggregated Consortium Dataset *relative profile* Disaggregated ALR Dataset	N/A	N/A	Preferred Inference Modality*	Risks the False Positive Error Anomaly

*Using sample size control by selecting from large dataset samples in the range [315 to 440]. See Lusk and Halperin (2014b).

The rationalization of this screening schema information is best discussed by referencing the studies that were used to create this taxonomic profile. The Log_{10} screen is an absolute point process screen and therefore, lacks sufficient practical variation to effectively screen using confidence intervals in the BPSH mode. See Lusk and Halperin (2014a). If one elects to use the BPP of the 20 sample accruals offered by Benford that has inherent variation and so is more likely to follow the Hill-mixing paradigm, then the confidence intervals offered by Lusk and Halperin (2014a) in Table 1 are the logical choice. Also, as another form of the inference calibration one could elect to use the chi-square inference measure. In this case, one must be cognizant of the fact that the chi-square inference measure is very sensitive to the sample size used in the inferential comparison. See Tam Cho and Gaines (2007). In this regard, Lusk and Halperin (2014b) offer a sampling range of [315 to 440] which is argued as a range that effectively controls the False Positive (FP) and the False Negative (FN) Errors for inferential comparisons. In this context of using the overall chi-square as the inferential basis of comparative analysis, Tamhane and Dunlop (TD) (2000, p.324) suggest an individual chi-square cell value sensitivity heuristic. They suggest that individual chi-square cell values are important signals of specific inherent variation from expectation. This can aid the ARL analyst in focusing the investigation. The TD heuristic is: *Any chi-square cell contribution greater than 1.0 is of interest as an indicator or signal of an important variance of* ***expectation*** *from* ***actual***. We will be using this heuristic on a-cell-by-cell basis consistent with the recommendation of Tamhane and Dunlop as it logically focuses on the particular digits that are likely candidates for investigation over the two datasets. To be clear, there is NO statistical inference attached to this TD-signaling protocol—it is their heuristic. What still governs is the overall chi-square; *this is the only statistically-based inference signal that can be used*. Finally, it is also the case that direct benchmarking creates a risk for the FP error anomaly as illustrated by Lusk and Halperin (2014b; 2014c) where they argue for two random samples with sample size control in the range [315 to 440]. This then rationalizes the various cell profiles that we will now use in our ARL profiling. This is an excellent juncture to summarize the components of the Big Data Protocol. There are five **stages** as an elaboration and extension of the recommendations of Akkaya and Uzar (2011) in the Big Data Profiling Montage which address these issues.

4.3. Specific context benchmarking

These five stages of the suggested protocol are:

A. Develop an *A-Priori* Expectation of benchmarking *Conformity* or *Non-Conformity* from the suggested Quadrangle in Table 2.

B. Select the *Variable Set* of Interest Relative to the *A-Priori* Expectation.

C. Develop the *Benchmark*: Disaggregation of the Big Data population & Develop the *Consortium*: Aggregation by selection of specific institutions/organizational entities from the Big Data population.

D. Determine the profiling testing montage as presented in Schema in Table 3.

E. Effect a Succinct Summary Analysis relative to the *A-Priori* expectation for the *Conformity*/*Non-Conformity*: BPSH and paired contrasts using the chi-square inference measure.

Consider now our illustrative context: The ARL Big Data population as analyzed using these five stages.

5. Results and discussion: Navigating the ARL-DFP-Labyrinth: An ARL illustration of the preferred screening profiles as presented in the screening taxonomy

To illustrate the functionality of the SCB analysis, we will conduct an investigation and provide the rationalization for the selection of the datasets and the development of the inferences produced by the SCB analysis. We wish to note that all of the information generated as part of the following illustrative analysis is generated using a Decision Support System called **SCA:DSS** that is available without cost or restriction from the authors. For each of the analyses that were produced by the **SCA:DSS,** we will note the specific worksheet that was used, such as *Tab:SampeSize* indicating that information being presented was generated by the **SCA:DSS**, Worksheet:SampleSize. Consider now the recommended stages for conducting the SCB analysis.

5.1. Specific context benchmarking: The stages of the ARL montage

Stage I: *Develop an A-Priori Expectation of benchmarking Conformity or Non-Conformity from the suggested Quadrangle in Table 2.* It is essential to form an expectation before conducting the SCB analysis so as to benefit from the relationships between what is the inferential result realized from the DFP of the SCB and one's initial expectation of the expected relationships. For our illustrative context we are interested in the *Professional Salary* dimension between a *Consortium of Ivy League ARLs* and the Specific Context Benchmark: *Selected other ARL reported by the ARLA respecting Professional Salaries*. The information for the various illustrations is found in the Appendix I. In this context:

> *We expect that the Ivy Consortium would differ in Salary Profile from the BPP and also from the ARL-Benchmark essentially due to: (i) expected uniformity—i.e., a lack of mixing—of the library generating processes for the Consortium thus creating, one would expect, BPP Non-Conformity, and (ii) the differences in the nature of the Service Profile of the Ivy Consortium compared to the ARL-benchmarking institutions.*

These dual-conditioned expectations are what we would consider as a *desirable state of nature*; therefore, if these expectations are realized in the SCB analysis this would not signal the need for considering possible continued investigative actions. In this case, then, we are in Cell (Expected *Non-Conformity*, Actual *Non-Conformity*).

***Stage II**: Select the Variable Set of Interest Relative to the A-Priori Expectation* The principal dataset to be used as the catalyst of reflective thinking for the SCB analysis is: *ARL Salaries of the Professional Staff*. As related contextual information, we have selected *ARL Services Reference Transactions*. We have selected the *Services* dataset as we suggest that this is the "kernel" generating process; after all, the *reason d'être* of the library system is the reference activity in the service of the client/stakeholder base. Therefore, we are interested in how the SCB of ARL Professional *Salaries* profiles in relief to this kernel generating process of *Services*.

Stage III: Develop the Benchmark: Disaggregation of the Big Data population & Develop the Consortium: Aggregation by selection of specific institutions/organizational entities from the Big Data population As we are interested in an SCB comparative analysis between the Ivy League institutions and other selected ARL institutions for *Professional Salaries* and *Reference Services* for the time-inclusive periods: 2000 to 2013, we made the following decisions: There are usually four ARLs that are added as part of the Ivy-8: *Duke*, *Chicago*, *MIT* and *Northwestern*. This is the aggregation stage where the ARL Consortium is formed; this is labeled *SalIvy*$^{+}$. As for the disaggregation stage, there are a number of ARL datasets that were deemed to not provide an interesting benchmark regarding professional salaries; specifically: all the Canadian members of the ARL were screened out, as were Public Libraries and Library Societies. This created the Disaggregated *Professional Salary* ARL dataset of: ninety-eight institutions, referred to as: *Sal98*, accounted for as follows: The original ARL: Professional Salaries download had 126 ARLs 16 of which were screened out as were the Ivy-8 plus *Chicago, Duke, MIT and Northwestern* or in total, 28 [16 + 12] yielding the dataset: Sal98: [126 –28].

Principal Analysis: ARL Professional Salaries: Download [SalDL], ARL Professional Salaries: Disaggregated [Sal98], and ARL Professional Salaries: Consortium [SalIvy^{+}].

Contextual Analysis: ARL Services Reference Transactions: Download [STrDL], ARL Services: Disaggregated [STr98], and ARL Services: Consortium [STrIvy^{+}].

The initial analysis is to examine if these six datasets are conforming to the BPP. In this regard the BPP confidence intervals suggested by Lusk and Halperin (2014a) as presented in Table 1 will be used in the BPSH mode. Recall, given the research of Hill (1995a; 1995b; 1996; 1998) and the demonstration of Fewster (2009) datasets that *do not conform* are likely to result from a generating process that is constrained in some way whereas those datasets that are not constrained are *likely to conform*.

Stage IV: Determine the profiling testing montage as presented in Schema in Table 3.

***Stage IV.a**: Conformity Analysis* Following is the analysis of the six ARL datasets under examination and their BPSH-*Conformity* profiles. Recall that we are using the 65.7% specific digit BPP containment as the cut-point for *Conformity*. Therefore, for this SCB, if six (6) or more digits are not in the BPP intervals of Table 1 then the dataset is labeled as: *Not-Conforming*; if 5 or less are not in the BPP then the dataset is: *Conforming*. The *Conformity* profile is coded in Table 4 using Tab: *ComputationsBSW* & Tab: *BenfordCalibrationTests*. In the Header row are the column variable designations, the number of institutions in the dataset, and the number of values contributed in total. For example, SalDL is the professional salary variable from the DownLoad, where there were 126 institutions and, in total, 1,686 reported professional salaries. In the Results row *Non-*

Conformity and *Conformity* are noted as: NonC[x] and C[x] respectively, and x represents the number of digits not in the BPSH screening intervals.

Table 4
Conformity / non-conformity screening using the BPP

First Digit Array	SalDL	Sal98	SalIvy$^+$	STrDL	STr98	STrIvy$^+$
	n:126_1,686	**n=98_1,227**	**n=12_167**	**n=126_1,641**	**n=98_1,211**	**n=12_134**
Digit 1	0.123	0.315	0.503	0.325	0.299	0.381
Digit 2	0.160	0.068	0.138	0.127	0.113	0.239
Digit 3	0.198	0.010	0.072	0.102	0.114	0.067
Digit 4	0.157	0.046	0.012	0.093	0.102	0.052
Digit 5	0.124	0.130	0.036	0.080	0.084	0.052
Digit 6	0.093	0.112	0.060	0.086	0.093	0.075
Digit 7	0.068	0.104	0.048	0.066	0.070	0.060
Digit 8	0.046	0.108	0.090	0.069	0.074	0.030
Digit 9	0.032	0.106	0.042	0.052	0.050	0.045
Result	**NonC[7]**	**NonC[9]**	**NonC[6]**	**C[4]**	**C[4]**	**NonC[6]**

Initially we will consider the comparative analysis of the two ARL-Downloads and then, *with that information,* we will examine the intra-context analysis. As a point of information, we recommend selecting a context for the SCB principal variable analysis as this will help give boundaries of reasonability and, in general, enrich the inferential nuances of the analysis. This is to say that in SBC benchmarking, context aids in making a determination of the nature of the effects that underlie the results on the principal variable under analysis. In this regard, we have selected *Service Transactions* as the context for the *Salary* analysis; we suppose that there is a structural relationship between the provision of library *Services* and the *Professional Salaries* required to deliver such services. It is not likely to be “one to one” but we suppose, *a priori*, that there is at least a meaningful direct associational relationship. We did examine this assumption by computing the Spearman Correlation coefficient as the assumptions underlying the Pearson (1900) version did not seem to hold. The Spearman correlations for [Salary, Service] for the Ivy+ and the Benchmark datasets were 0.53 and 0.27 respectively. Both p-values were <0.0001 as tested against the Null thus supporting our supposition.

Specifically, this SCB profiling for the Downloads suggests the following. The two downloads are different with respect to the BPP. The *Salary* Download, SalDL, follows, for the first six digits: {1, 2 , 3, 4, 5, 6}, the Hill uniform or Lottery-model of equal overall likelihood of 11.1%; whereas the *Services* Download, STrDL, is conforming suggesting that it follows the Hill-*Conformity* profile characterized by mixing. The interesting inference gleaned from this aspect of the SCB analysis is that, given that *Services* follows the Hill-Profile, likely effected by mixing and so resulting in a base-invariant profiling, then because *Salaries* are equally blocked over most of the initial six digit ranges this strongly suggests that the 126 ARL institutions have average professional-salary profiles that are likely different over important sub-groups. The reasoning for this is that if the average professional salary profile was the same for all the ARL institutions then this would be effectively a constant multiplier of the StrDL profile which due to the base invariance character of the Hill-Profile would leave the Salary profile similar to the Hill-*Conformity* profile. This means that under an equal salary profile one expects that the Professional Salary would be a *Conforming* dataset; as this is not the case, this suggests that the likely salary profiles are variable in some systematic way. *Summary Insight for the Downloads: We observe that Salaries appear to be controlled in a uniform manner possibly due to a few uniform Salary levels or groupings which are magnitude shift adjusted in unit-salary-scales thus producing the Hill-Lottery profile for most of the data frequencies; this seems to be the case as the related Services are in fact a Conforming dataset. Therefore, the various ARLs in the download have different salary scales and also different service configurations.*

The next aspect to consider is the Intra-context profiles of Services. Considering Str98 and STrIvy$^+$, we observe that, consistent with the Download results, there seems to be mixing in that the STr98 retains its Hill-*Conformity*. However, most interestingly the STrIvy$^+$ partition is Non-conforming suggesting that the STrIvy$^+$ group, as a partition of the STrDL are not mixed and are rather more of a monolith or homogenous in the delivery of their services. This suggests strongly that the STrIvy$^+$ group represents institutions that have relatively the same kernel generating process and/or that there is not sufficient mixing to create Hill-*Conformity*. In this regard, observe that over 60% of the DFP is in the first two digits and the other 40% or so is relatively uniformly spread over the last seven digits at, on average, 5.5% uniformly. This suggests that the STrIvy$^+$ group has effectively the same service configuration and this uniformity produces a lack of mixing and so creates the *Non-Conformity*. The opposite is the case for the disaggregated group, STr98, where these 98 institutions have sufficient variation so that the Hill-Mixing is in-evidence over the aggregate and so there is *Conformity*. *Summary Insight: For the Intra-Service Context analysis, we observe that services appear to be controlled or uniform in the STrIvy+ context whereas such kernel blocking over the mix of institutes does not seem to have inhibited the variation of services over the aggregated ARL: STr98.*

Stage IV.b*: Intra-Salaries SCB analysis*, we see that the segregation of the datasets dramatically changes the DFP relative to the download profile; however, even though the Sal98 and SalIvy$^+$ are non-conforming their DF profiles are very different. For the aggregated: Sal98 there is a reversal of uniform DF loading relative to the SalDL. In the download, most of the mass was uniformly spread over the first six digits; now for the Sal98 institutions the last five digits have this uniform profile that produces the *Non-Conformity*! For the SalIvy$^+$, the *Non-Conformity* is clearly evident in that 50% of the digital mass is located on the first digit. This fits very well with the Service results where for the STrIvy$^+$ 60% of the service DFP was on the first two digits and so strongly suggests that there is uniformity—non-mixing—not only in the service configuration but also in the salary profile as they both profile in much the same way. This echoes the salary-sub-group profile variability that was in evidence in the Download analysis. *Summary Insight: For the Intra-Salary Context analysis, we observe that salaries appear to be controlled or uniform for both the 98 institutions and for the Ivy Plus group. For the SalIvy$^+$ context we see the striking similarity between the digital profiles of the Services and the Salaries where both place about 60% of the frequencies in the first two digits. For the dis-aggregated institutions, Sal98, the salaries are uniformly spread over the last five digits; and in a parallel manner most of the digital mass in the last set of digits is also relatively uniformly distributed for the services aspect. These are certainly most interesting parallels for the DFP of the two groups with respect to the Download analysis presented above.*

Stage IV.c*: Statistical Profiling using the chi-square inference structure* The next step is to determine if there is evidence that the non-conforming datasets differ from each other using another inference measure. This chi-square analysis of the DFPs is a recommended *robustness check* on the SCB-analysis. Often one could, in fact, stop the SCB analysis at the BPSH stage. Certainly we have learned a great deal from this BPSH SCB analysis. However, to be complete we will continue the SCB analysis focusing here only on the salary component as this was our principal variable in the SCB ARL-analysis. To examine these two datasets: Sal98 and SalIvy^{+}, we will use the chi-square distribution as calibrated by Lusk and Halperin (2014b). This is the final step in the SCB DFP analysis and here we will use the comparative chi-square analysis. In this case, as the Sal98 dataset has more than 1,000 observations, we will take a random sample in the range [315 to 440] as is recommended by Lusk and Halperin (2014b). In this regard, we used the Excel Worksheet function: *RANDBETWEEN[x=315, y=440]*. This produced a recommended random sample of 364. [Table: *SampelSize*] Using a sampling function that is part of the **SCA:DSS**, we produced the following sampled data profiles for the two datasets as presented in Table 5. [Table: *ComputationsTwoDataSetsOnly*] Note, that the profile for Sal98 is slightly different than the profile reported in Table 4 as we took random samples from the Sal98 dataset. There is an alternative to taking random samples to form the two-sample comparison; to effect this alternative, one uses the actual profile for the datasets and fixes the sample size for the computation to some number in the range [315 to 440] and re-creates the number of observations.

Table 5
The sampled profiles from the Sal98 and the SalIvy+ datasets

First Digit Array	Digital Profile for Sal98; n=364	Digital profile SalIvy+; n=167	Cell Chi-square Values
Digit 1	0.30769	0.50299	**3.7/8.1**
Digit 2	0.06044	0.13772	**2.5/5.5**
Digit 3	0.01923	0.07186	**2.8/6.1**
Digit 4	0.03297	0.01198	0.6/**1.3**
Digit 5	0.12088	0.03593	**2.8/6.0**
Digit 6	0.10714	0.05988	0.9/**1.9**
Digit 7	0.12363	0.04790	**2.1/4.5**
Digit 8	0.12637	0.08982	0.4/0.9
Digit 9	0.10165	0.04192	**1.6/3.4**

In this case, the comparisons between the two datasets are noted in the fourth column; these are the individual chi-square Cell-Contributions. We have **bolded** the particular cell contributions that are greater than the TD-heuristic of 1.0. The overall comparative result is that the chi-square comparison has an overall chi-square value for inferential purposes of 55.05 which suggests that the two datasets are not likely to have been produced by similar generating processes—i.e., the Null of no difference may be confidently rejected. Therefore, this being the case the TD-Chi-square values are useful information. Specifically, we see that most of the digits have a differential frequency profile; in fact only Digit 8 is not flagged as different between the two datasets. To make the simplest selection, we notice that the largest sum of the digits is for Digit 1 with a sum of the Cell-Contributions of 11.8 which is 32.6% larger than the second largest sum. This then suggests that these two non-conforming datasets also differ from each other in most respects but in particular with respect to the first digit where we see the Ivy Group relative to the ARL-Disaggregated Group is: [50.3% vs. 30.8%]. This is confirmatory of the information gleaned using the BPSH analysis and can be used as a point of reference as to the robustness of the insights that we have offered above relative to the Salary realizations as examined in the BPSH phase.

***Stage V**: Effect a Succinct Summary Analysis relative to the A-Priori expectation for the Conformity/Non-Conformity: BPSH and paired contrasts using the chi-square inference measure* The summary image gleaned from the information generated at the various stages of the SCB analysis is that there are relative uniformities in the systems both for the *Services* and the related *Salaries* and that these intra-group uniformities are different over the two groups. The Ivy consortium is, in Salary and Service configuration, different from the benchmarking group of 98 libraries as well as presenting a non-conforming set of data for both Salary and Services. Most simply stated:

> *The IvyPlus Consortium group seems more uniform compared to the Benchmark in both the **Service** profile and in the related **Salary** committed to sustain the service profile.*

Possibly this is due to the relative endowment profiles between the Ivy League and the other Benchmarking institutions. Considering only universities with endowments of over 1 Billion USDs, the average (mean) endowment of the IvyPlus group is about $12.2 billion compared to $2.8 billion for the other ARL libraries. Therefore, given the likely funding/budget differential it seems that this *Non-Conformity* and the directed differential discussed above seem logical and certainly *consistent with our initial expectation* and so there is not likely to be a reason to consider investigations regarding changing of the IVYPlus systemic profile as far as *Salary* and *Services* are concerned.

5.2. Validity robustness check: Reliability calibration

As a simple validity or credibility check on these SCB relative profiles, we also computed the statistical demographic profiles for *Salaries*. This statistical re-analysis is intended to provide another reflective modality of the SCB results. If the usual statistical analyses are supportive of the SCB profile then this should reinforce the meaningfulness of the SCB results. If not then perhaps the SCB is leading in a False Positive Investigative direction. However, the point that we are making is that the SCB is the first "cut" of the analysis and that these SCB results should be further validated by other statistical methods as a reliability calibration. This is consistent with "best practices" analytics where there is clear jeopardy in not "looking" at the results in a variety of different dimensions—after all dimensional profiling is the nature of the BDA modality.

This re-analysis of the Salary profile which was one of the basic foci of the SBP is presented in Table 6.

Table 6
Relative summary financial profiles: *Salary* (in Millions)

	SalIvy+	Sal98
Mean/Median	19.2/13.3	10.4/9.0
St Error[Mean]	1.0	0.15
Range	59.6	29.7
DFT Containment	50.3%	31.5%

This financial profile is a simple statistical validity check on the DF profiling developed above using the BPSH and the chi-square inference measure. In considering the relative Means and Medians and the differences in the standard error of the mean it is clear that there is a "bunching-up" of the salary expenditures for the SalIvy+ in the 10s of Millions and for the Aggregate [Sal98] less but close to the 10s of millions which are consistent with the profiles that we found and reported using the SCB for the DFPs. In fact, the last measure that we computed, the digital frequency transaction (DFT) containment, is just the percentage of salary budgets in the range: 10.0 <= Salary Budgets <= 20.0; for the SalIvy+ it is 50.3% and for the aggregate it is 31.5% which are almost identical to the Digital Frequencies for the Benford analysis: See the first row for Digit 1 in Table 5. This analysis then is certainly consistent with the SCB that we formed for salary and so will, we suggest, enhance the decision-making relevance of the SCB results.

6. Overall summary for reflective consideration for the SCB analysis

In summary, we have offered the technical underpinnings of SCB in the Big Data context. Additionally we have suggested and illustrated, for the first time, Digital Frequency Profiling as a measure that aids in focusing inferential information derived from SCB profiling. We have undertaken this discussion and illustration of the technical aspects of the SCB as it may play out in the ARL context in the case that those in the ARL community may find Digital Frequency Profiling an analytic technique that may be of interest in creating reflective thinking profiles of the benchmarking for consortium groups. Even though we focused on the ARL datasets, one may generalize this to any Big Data environment where industry summary statistics and sub-group comparisons could create useful decision-making information—i.e., the SCB focus. For example, for purposes of developing Balanced Scorecard profiles one could use the NAICS industrial grouping or the SIC groupings as the peer profiling groups.

Echoing the cautionary advice of Porter and Grogan (2013), Niessing and Walker (2015) note:

> *"At bottom, debunking these myths is about discarding blind faith that the formulae for business success are set down in the data. Truth is, Big Data is a tool in itself, like a computer or smartphone- an awesome, game-changing tool, but only when wielded by people who know the right commands and coordinates."*

As for the future, it would be most useful in this "fledgling" sort of analysis where DFP and SCB are used to create decision-making information in the library Big Data context that as individuals create information relative to their SCB analyses that they make such analyses available on their Commons-Links so that DFP can be refined and developed further in the service of SCB in the ARL context.

Acknowledgements

We wish to thank Dr. H. Wright, *Boston University*: Department of Mathematics and Statistics and the participants at the Research Workshop at the SUNY:Plattsburgh for their most helpful suggestions. Additionally, we wish to thank the two anonymous reviewers selected by *Knowledge Management and E-Learning* for their detailed and very constructive suggestions. Finally, thanks to Dr. Minhong (Maggie) Wang, Editor-in-Chief of *Knowledge Management and E-Learning* for her excellent stewardship of our paper.

References

Akkaya, G. C., & Uzar, C. (2011). Data mining: Concept, techniques and applications. *GSTF Business Review (GBR), 1*, 47–50.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78*(4), 551–572.

Brinley, F., Cook. C., Kyrillidou, M., & Thompson, B. (2010). Library investment index – Why is it important? In *Proceedings of the International Conference on QQML* (pp. 243–249).

Das, R., Hanson, J. E., Kephart, J. O., & Tesauro, G. (2001). Agent-human interactions in the continuous double auction. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)* (pp. 45–49).

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

Fewster, R. M. (2009). A simple explanation of Benford's Law. *The American Statistician, 63*, 26–32.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Geyer, C. L., & Williamson, P. P. (2004). Detecting fraud in data sets using Benford's law. *Communications in Statistics - Simulation and Computation, 33*, 229–246.

Hickman, M. J., & Rice, S. K. (2010). Digital analysis of crime statistics: Does crime conform to Benford's law? *Journal of Quantitative Criminology, 26*(3), 333–349.

Hill, T. P. (1995a). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society, 12 3*, 887–895.

Hill, T. P. (1995b). The significant-digit phenomenon. *The American Mathematical Monthly, 102*(4), 322–327.

Hill, T. P. (1996). A statistical derivation of the significant-digit law. *Statistical Science, 10*(4), 354–363.

Hill, T. P. (1998). The first digit phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist, 86*(4), 358–363.

Kelly, C. (2011). Benford's law to the rescue: Taking the popular analysis method to a deeper level can uncover well-hidden fraud and control failures. *Internal Auditor: IT Audit, 68*, 25–27.

Koltay, K., & Li, X. (2010). *SPEC KIT 318: Impact measures in research libraries*. Washington, D.C.: Association of Research Libraries. Retrieved from http://publications.arl.org/Impact-Measures-in-Research-Libraries-SPEC-Kit-318/

Lewin, H. S., & Passonneau, S. M. (2012). An analysis of academic research libraries assessment data: A look at professional models and benchmarking data. *The Journal of Academic Librarianship, 38*(2), 85–93.

Ley, E. (1996). On the peculiar distribution of the U.S. stock indexes' digits. *The American Statistician, 50*(4), 311–313.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics, 65*, 1–12.

Lusk, E. J., & Halperin, M. (2014a). Using the Benford datasets and the Reddy & Sebastin results to form an audit alert screening heuristic: An appraisal. *IUP Journal of Accounting Research & Audit Practices, 8*(3), 56–69.

Lusk, E. J., & Halperin, M. (2014b). Detecting digital frequencies anomalies as benchmarked against the Newcomb-Benford theoretical frequencies: Calibrating the x^2 test: A note. *International Business Research, 7*(2), 72–86.

Lusk, E. J., & Halperin, M. (2014c). Detecting Newcomb-Benford digital frequency anomalies in the audit context: Suggested x^2 test possibilities. *Accounting and Finance Research, 3*(2), 191–205.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics, 4*, 39–40.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's law. *The Journal of American Taxation Association, 18*, 72–91.

Nigrini, M. J. (1999). I've got your number. *Journal of Accountancy, 187*(5), 79–83.

Niessing, J., & Walker, J. (2015). *The eight most common Big Data myths*. INSEAD Knowledge. Retrieved from http://knowledge.insead.edu/blog/insead-blog/the-eight-most-common-big-data-myths-3878

Oakleaf, M. (2010). *Value of academic libraries: A comprehensive research review and report*. Chicago, IL: Association of College and Research Libraries.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5, 50*(302), 157–175.

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

Rauch, B., Göttsche, M., Brähler, B., & Engel, S. (2011). Fact and fiction in EU-Governmental economic data. *German Economic Review, 12*(3), 243–255.

Reddy, Y. V., & Sebastin, A. (2012). Entropic analysis in financial forensics. *The IUP Journal of Accounting Research & Audit Practices, 6*(3), 42–57.

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*(3), 489–507.

Tam Cho, W. K., & Gaines, B.J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician, 61*(3), 218–223.

Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate* (1st ed.). Upper Saddle River, NJ: Prentice Hall. ISBN-10: 0-13-744426-5

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.

Appendix I. Digital frequencies for 12 ARL datasets from 2000 to 2013*

Digits	BPS	BTE	BTS	SCir	SGP	SRT	SN-P	SProS	SSA	PhDA	USTS	UTE
1	**0.12**	0.41	0.33	0.15	0.32	**0.32**	0.46	0.31	0.22	0.27	0.33	0.31
2	**0.16**	0.30	0.09	0.19	0.09	**0.13**	0.12	0.05	0.10	0.22	0.36	0.14
3	**0.20**	0.13	0.03	0.21	0.05	**0.10**	0.06	0.04	0.13	0.16	0.15	0.08
4	**0.16**	0.07	0.05	0.15	0.09	**0.09**	0.02	0.08	0.15	0.12	0.06	0.08
5	**0.12**	0.03	0.11	0.09	0.09	**0.08**	0.04	0.16	0.13	0.06	0.02	0.09
6	**0.09**	0.02	0.10	0.07	0.11	**0.09**	0.05	0.13	0.10	0.06	0.02	0.09
7	**0.07**	0.01	0.10	0.05	0.10	**0.07**	0.07	0.09	0.07	0.05	0.02	0.09
8	**0.05**	0.01	0.10	0.05	0.09	**0.07**	0.07	0.08	0.06	0.04	0.01	0.08
9	**0.03**	0.01	0.09	0.04	0.07	**0.05**	0.10	0.08	0.05	0.03	0.02	0.05

*Source: ARL Statistics – Institutional Data (http://www.arlstatistics.org/analytics)

Legend: **BPS (Budget Professional Salary)**, BTE (Budget Total Expenditures), BTS (Budget Total Salaries), SCir (Services Circulation), SGP (Services Group Presentations), **SRT (Services Reference Transactions)**, SN-P (Salaries Non-Professional), SProS (Salaries Professional Staff), SSA (Staff Student Assistants), PhDA (PhD Awarded), USTS (University Statistics Total Students), UTE (University Total Expenditures). The two datasets that were used as part of the SCB illustration are here **bolded** in the above Table. The full datasets are included in the SCA:DSS.