JOURNAL OF INFORMATION SYSTEMS
Vol. 27, No. 2
Fall 2013
pp. 1–19

American Accounting Association
DOI: 10.2308/isys-10372

EDITORIAL

AIS in an Age of Big Data

I. INTRODUCTION

Vasarhelyi (2012a) discussed the need for accounting Information systems (AIS) to accommodate business needs generated by rapid changes in technology. It was argued that the real-time economy had generated a different measurement, assurance, and business decision environment. Three core assertions relative to the measurement environment in accounting, the nature of data standards for software-based accounting, and the nature of information provisioning, formatted and semantic, were discussed.

> a. Measurement and representation methods were developed for a different data processing environment. For example, FIFO and LIFO add little value in an era when actual identification, real-time measurement, and real-time market prices are available. (Vasarhelyi 2012a)

This note discusses the effect of Big Data on potential accounting measurements, in particular the potential real-time provisioning of transaction data, the potential provisioning of information data cubes breaking reports into division, product, etc., and the need for different types of accounting standards.

> b. Substantive formalization is necessary for putting in place automation and the dramatic process changes that are necessary. In addition to data processing mechanics, classification structures like taxonomies and hierarchies must be expanded to harden "soft" knowledge into computable structures (Krahel 2011; Vasarhelyi and Krahel 2011; Geerts and McCarthy 2002; Vasarhelyi 2012a; Vasarhelyi 2013).

Argument for formalization of standards is made to bring these into the Big Data digital information provisioning age.

> c. Automatic semantic understanding and natural language processing is necessary to disambiguate representational words in financial statements, evaluative utterances in media reports and financial analyses, potentially damning verbiage in Foreign Corrupt Practices Act (FCPA) violations, etc. (Vasarhelyi 2012a)

Big Data, by its nature, incorporates semantic data from many sources. These utterances are argued to enhance the information content of financial reporting.

The authors are appreciative of the comments from Professor M. Alles, A. Kogan, Kyunghee Yoon, and Roman Chynchyla and the assistance of and Qiao Li.

Published Online: December 2013

In the area of assurance additional concerns were cited:

> d. Traditional procedures in assurance have begun to hinder the performance of their objectives. As an example, confirmations aim to show that reality (bank balances, receivables) is properly represented by the values on the corporations' databases[1] (Romero et al. 2013). This representational check is anachronistically performed through manual (or email aided) confirmations in an era where database-to-database verification with independent trading partners can be implemented (confirmatory extranets; Vasarhelyi et al. [2010]). (Vasarhelyi 2012a)

This note discusses the opportunities and challenges of Big Data in the audit process.

> e. Current auditing cost/benefit trade-offs were likewise calibrated for a different data processing era. The trade-off between cost of verification and the benefits of meta-controls have dimensionally changed. For example, statistical sampling, as a rule, makes little sense in a time when many assertions can be easily checked at the population level. (Vasarhelyi 2012a)

The economics of business will determine the adoption of new accounting and assurance processes integrated/facilitated by Big Data. Socio-technical systems are usually substantively affected by resistance to change.

The pervasive phenomenon of "Big Data" is emerging and coloring these assertions. Typically technology is developed, incorporated into business, and later integrated in accounting and auditing. Organizations have found that in many areas, nontraditional data can be major drivers of multiple business processes. The traditional EDP/ERP environment is typically structured and bound (limited in size with clearly delimited boundaries). Less traditional forms of information (e.g., emails, social media postings, blogs, news pieces, RFID tags) have found their way into business processes to fulfill legal requirements, improve marketing tools, implement environmental scanning methods, and perform many other functions. Eventually, the measurement of business (accounting), the setting of standards for measurement and assurance (FASB and PCAOB), and the assurance function itself (auditing) will become aware of these facts and evolve. Business measurement and assurance are essential for economic production activities and will continue to be performed, but current accounting and auditing methods are in danger of becoming anachronistic, insomuch that they are progressively ignored by economic actions and entities. This could result in tremendous societal costs in terms of societal duplication of measurements and of assurance processes, and cannot be avoided without some degree of standardization, supervision, and comparability.

Within this evolving environment, a large set of interim and longer-term issues emerge. Section II describes the content of this issue of *The Journal of Information Systems*. Section III addresses the basics of Big Data. Some societal-effect illustrations are presented in Section IV, and Big Data in relation to accounting, auditing, standard setting, and its research is discussed in Section V. Section VI concludes.

II. CURRENT ISSUE OUTLINE

The current issue includes a special section on virtual-worlds research edited by Professor William Dilla that encompasses one article plus one "Innovation" section article. Other articles on this topic, also edited by Professor Dilla, may appear in later issues as they are still going

[1] Or, in other words, whether the numbers in the accounting system properly represent real life.

through the editorial process. The ensuing issues will also have special sections on enterprise ontologies (edited by Professor Guido Geerts), and social media (edited by Professor Roger Debreceny). These and others are planned, to be supervised by incoming editors Roger Debreceny and Mary Curtis, most likely covering IT auditing, continuous audit/continuous monitoring, and semantic analysis.

Van der Heijden (2013) examines information dashboards relative to anchoring and presentation format. He examines aspects of dual-performance measures in the context of organizations disclosing operational performance to the general public through information dashboards. Dual-performance measures are measures where performance is a function of two values: one value denoting the percentage of a group to which the measure refers to, and one value denoting the performance level achieved by that particular percentage. A 2×2 experiment, involving performance assessment of a fictional emergency room, varies anchor and presentation format, and measures the effects on subjective performance of the emergency room, as well as perceived informativeness and attractiveness of the dashboard. The results indicate, first, that choice of anchor matters, in the sense that anchor choice can mask or accentuate relevant information, thereby influencing subjective performance. Second, a pictorial unit chart combined with a performance-level anchor is perceived to be the most informative and most attractive dashboard display.

Janvrin, Pinsker, and Mascha (2013) examine XBRL *vis-à-vis* Excel and PDF as the medium of financial statement disclosure. Prior experimental evidence suggests that even when eXtensible Business Reporting Language (XBRL)-enabled technology is available, it is not used by almost 50 percent of participants. The state of XBRL-enabled technology is examined by using an exclusive choice experimental design to examine (1) which reporting technology nonprofessional investors will choose to complete a financial analysis task, and (2) why they choose the specific technology. It was found that 66 percent of nonprofessional investor proxies chose to use XBRL-enabled technology, while 34 percent chose spreadsheets. Participants who chose XBRL-enabled technology perceived that it reduces the time to complete the task (i.e., increases task efficiency), while participants who chose spreadsheets indicated their choice was driven by prior technology experience.

Cong and Romero (2013) focus on information system complexity and vulnerability from the perspective of the enhanced internal controls after Sarbanes-Oxley. It is conjectured that the increased statutory and regulatory requirements on more stringent internal controls increase information systems complexity and, therefore, increase information systems vulnerability. The study hypothesizes that the increased information systems complexity also increases information systems vulnerability, even though the overall internal control is improved. The results of the empirical tests support the hypotheses.

Steinbart, Raschke, Gal, and Dilla (2013) consider information security professionals' perceptions about the relationship between the information security and internal audit functions. The paper presents the results of a survey of information security professionals' perceptions about the nature of the relationship between the information security and internal audit functions in their organization. It finds that information security professionals' perceptions about the level of technical expertise possessed by internal auditors and the extent of internal audit review of information security are positively related to their assessment about the quality of the relationship between the two functions. It also finds that the quality of the relationship between the internal audit and information security functions is positively associated with perceptions about the value provided by internal audit and, most important, with measures of overall effectiveness of the organization's information security endeavors.

Elbashir, Collier, Sutton, Davern, and Leech (2013) discuss business intelligence (BI) in light of shared knowledge and assimilation. Business intelligence systems have attracted significant

interest from senior executives and consultants for their ability to exploit organizational data and provide operational and strategic benefits through improved management control systems. BI has too often failed to support organizations' managerial decision making at both the strategic and operational levels and, thus, failed to enhance business value. *Whether* and *how* organizations achieve business benefits from their BI investments remains unclear. The study draws on the strategic alignment and IT assimilation literature to develop a research model that theorizes the importance of BI systems assimilation and the need for shared knowledge among the strategic and operational level as the drivers of BI business value. Results from the study confirm the crucial role of BI assimilation in translating organizational resources into capabilities that enhance the business value of BI.

Srivastava, Rao, and Mock (2013) use the evidential reasoning approach to evaluate evidence obtained to assess and control the risks of providing assurance on sustainability reports. Sustainability reporting, or corporate sustainability reporting (CSR), provides stakeholders with important information on both financial and nonfinancial factors related to environmental, social, and economic performance. The presented framework is developed from both a Bayesian (probability-based theory) and belief function (Dempster-Shafer theory) perspective. This facilitates application of the framework to cases where the assurance provider prefers to assess risk in terms of probability versus in terms of beliefs. To demonstrate the application of this framework it evaluates assertions, sub-assertions, and audit evidence relevant to CSR based on the G3 Reporting framework developed by the Global Reporting Initiative (GRI).

William Dilla, editor of the special topic on virtual worlds, in his editorial, introduces the tenets of this area and discusses opportunities in accounting research. Professor Dilla's editorial details the content and contribution of the two virtual-worlds papers.

III. BASICS OF BIG DATA

Big Data has recently been the topic of extensive coverage from the press and academia, although the focus on the topic by part of academic accountants has been limited. This note aims to deal with the effect of Big Data on the issues raised in the above introduction. First, it discusses Big Data in general, second it brings out some illustrations of related social issues, and then focuses on their effect on accounting research. The conclusions highlight the effect of Big Data on issues raised in the introduction.

What Is Big Data?

Big Data is defined, in part, by its immense size. Gartner (2011) explains it as data that "exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population." Similarly, the McKinsey Global Institute in May 2011 described it as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Franks 2012).

Furthermore there are many factors that have created, and are intrinsic to, the Big Data phenomenon. Typically Big Data:

1. is automatically machine obtained/generated;
2. may be a traditional form of data now expanded by frequent and expanded collection;
3. may be an entire new source of data;
4. is not formatted for easy usage;
5. can be mostly useless, although Big Data is collected and its economics are positive; and

6. is more useful when connected to structured data in corporate enterprise systems (ERPs) (Franks 2012, adapted).

Big Data has many advantages over traditional structured databases. The properties of Big Data enable analysis for the purpose of assembling a picture of an event, person, or other object of interest from pieces of information that were previously scattered across disparate databases. Big Data is a repository for multi-structure data and presents the ability to draw inferences from correlations not possible with smaller datasets. With Big Data, noise becomes meaningful, and outliers may be included as part of the complete model rather than being discarded. Processing power and storage capacity have been commoditized making Big Data possible for organizations of all sizes (McAfee and Brynjolfsson 2012). Big Data can create/increase profitability for business. This study shows that businesses that use Big Data to inform their decisions have 5–6 percent higher profitability. Large web and IT companies such and IBM, Google, Yahoo, and Amazon have pioneered the efforts of storing and extracting useful information from Big Data, but other industries are now taking advantage of the technology. Big Data is being harnessed by many business sectors including finance and insurance, for risk analysis and fraud detection; utilities and telecom for usage analysis and anomaly detection; and retail and marketing for behavior analysis and product placement.

The Structure of Big Data

Big Data can exist as large structured data (e.g., data that fit into a defined schema, such as relational data), semi-structured data (e.g., data that are tagged with XML), unstructured data (e.g., text and video), and multi-structured data (e.g., integrated data of different types and structural levels). Unstructured data represent the largest proportion of existing data and the greatest opportunity for exploiting Big Data. For example, plain text found in the MD&A section of quarterly and annual reports, press releases, and interviews is completely unstandardized. The context in which this text is presented must be inferred by the type of document on which it is found, titles, subheadings, and words within the text itself. Major themes can be extracted using mathematical and machine learning techniques such as tf-idf (Aizawa 2003), latent semantic analysis (Landauer et al. 1998), and cluster analysis (Thiprungsri and Vasarhelyi 2011). Data with free text can be "tagged" based on the context, but not with the granularity and accuracy of XBRL.[2] Working with textual data provides many opportunities to discover patterns, writing styles, and hidden themes. To further improve the analysis of unstructured data, some attributes can be attached such as the source, date, medium, and location of the data to improve understandability. For example, the structured data records of a client may be linked/hyperlinked to his/her emails to the company, posting of comments in social media, or mentions in the press.

Big Textual Data

Big textual data are available to accounting researchers now. Textual data come from many sources including EDGAR,[3] newspapers, websites, and social media. To increase the utility of the data, the text can be parsed and processed with software. For example, each Item of the 10-K and 10-Q can be tagged (e.g., Item 1, Item 7) and treated separately. Each document can be processed at the document, section, sentence, or word level to extract textual features such as part of speech, readability, cohesion, tone, certainty, tf-idf scores, and other statistical measures. The results can

[2] http://www.xbrl.org
[3] http://www.sec.gov/edgar.shtml

be stored for future querying and analysis. Many of these texts, and text-mining results, would occur/be placed on server clusters for mass availability. Text understanding and vague text understanding can provide the necessary links from textual elements to the more traditional ERP data. Eventually the vocalic and video data would also be progressively linked to the more traditional domains. The AIS, accounting, and finance research communities have already made progress in how to process text (Bovee et al. 2005; Vasarhelyi et al. 1999) and impound it into research.

The Relationship between Big Data and the Cloud

Weinman (2012) calls the cloud both an existential threat and an irresistible opportunity. He points out that most key trend summaries rank cloud computing at or near the top of the list. Most, if not all, of the rest of the top priorities—virtualization, mobility, collaboration, business intelligence—enable, are enabled by, or otherwise relate to the cloud. He also stresses that Rifkin (2001) would consider this to be a natural consequence of "The Age of Access." Rifkin has argued that the market economy—in which people own and trade goods—is being replaced by the network economy—where people pay to access them. The cloud and Big Data are related concepts. While typically the cloud is seen as an ephemeral medium of wide bandwidth and distributed storage, its existence is justified by the need for Big Data.

Weinman (2012) calls the cloud disrupting to every dimension of business, whether it is the research, engineering, or design of new products and services; or their manufacturing, operations, and delivery. The cloud also disrupts a business's interface with the customer and marketing in general including branding, awareness, catalog, trial, customization, order processing, delivery, installation, support, maintenance, and returns. The cloud can be defined with a helpful mnemonic, C.L.O.U.D., reflecting five salient characteristics: (1) Common infrastructure, (2) Location independence, (3) Online accessibility, (4) Utility pricing, and (5) on-Demand resources.

Gilder (2006) calls cloud data centers "information factories" since the cloud can be viewed, in part, as representing the industrialization of IT and the end of the era of artisanal boutiques. Many of the lessons learned in the evolution of manufacturing can be applied to the cloud, as well, including the economies of scale obtained by the cloud and Big Data.

IV. BIG DATA ILLUSTRATIONS

Big Data and the cloud are substantially changing/affecting business, politics, security, and governmental supervision.

Corporations Are People

Big Data in the popular press mainly focuses on knowing all there is to know about individuals (Franks 2012). Emails, phone calls, internet activity, credit card usage, opinions, friends, photographs, videos, passwords, bank account balances, travel history, and more can all be known about an individual with the proper credentials. All of this information can play an important role in painting an accurate picture of who the individual is, what that person has done, and what that person will do in the future. In the eyes of the government it may be advantageous to know if the individual is a friend or foe of the state, and in the eyes of creditors it may be useful to know if the individual will repay a loan.

Accountants must view the possibilities associated with Big Data, of knowing much about a corporation, including knowing a substantive amount about who works in a corporation. While it seems objectionable and invasive that a stranger could know virtually everything about another

person, knowing as much as possible about a corporation is much more palatable. What can be known? One beginning point is to know everything that can be known about the individuals (nodes) within a corporation. Each node should be understood within the context of a corporation's hierarchy to give it proper weight in determining corporate characteristics, yet each node, down to the lowest level, can be impactful. Since information about individuals is now sold like a commodity, the complete analysis of a business must now include such information. Furthermore, the privacy protections of individuals are less within the corporate environment. In general any utterances, documents, and email generated within the corporate structure, using corporate resources, are allowable for scrutiny.

Many questions have yet to be answered regarding increased obtrusive surveillance of companies, and detailing information about employee activities: (1) Should metrics (such as time on the Internet, sites visited, phone calls, geo-location, etc.) be created for employee activities and should they be reported? (2) Should company Big Data be available to investors and auditors? (3) What degree of detail on this data should be made available to the stakeholders/public? (4) Would society as a whole benefit from this information?

Big Data in Surveillance

The U.S. government has confirmed the existence of a system called xKeyscore to glean email and web traffic for surveillance purposes. The National Security Agency (NSA) approach collects phone call data (not the content) of calls through U.S. phone companies and stores them for five years. This is done by the NSA, as the phone companies (due to data size) do not keep them for such a period. This is called the "collect first" (*The Economist* 2013) model where it is available to find relevant data to national security investigations. As there are over 500 million calls a day, five years of this data consists of a very large database and the linkage to other records to make this data relevant. This large quantity of data makes it one of the largest databases existing today. Another database that may be linked to it is the PRISM database, which actually has content from emails and social media (such as Facebook) that are sent or received by foreigners. Although little is known of the details of these systems, their existence and purposes can easily be rethought for the purpose of accounting reporting, assurance, marketing analysis, and sales. Understanding who calls and is called by whom, the volume and timing of calls, can be a treasure trove for understanding the sales of your competitors, confirming volume of sales, predicting macro-economic information, providing leads to your sales force, and detecting product problems, among many other additional applications of the same data.

Edward Snowden (the leaker of Big Data monitoring[4] by the U.S. government; Lee [2013]), and Bradley Manning (the leaker in the Wiki Leaks episode; Dickinson [2011]) are indications that any node within an entity, given the proper motivations, can have a great impact. Their actions have been condemned and lauded, but their impact has not been questioned. From these episodes, and the above facts on government surveillance in the U.S., a few conclusions may be inferred: (1) even low-ranked elements in the chain can have substantive access to information, (2) traditional safeguards of information are not sufficient in the era of Big Data, (3) fraud may not be the main motivation for Big Data security infringement, (4) large businesses' Big Data will have similar degrees of intrusiveness, (5) selective extraction of data from large receptacles may provide troves of highly valuable strategic or casuistic information, and (6) large databases can be used for substantively different purposes of their initial intent.

[4] http://en.wikipedia.org/wiki/PRISM_%28surveillance_program%29

The SEC Audit Quality Model

In early 2013 it was reported that the SEC would roll out a new tool for detecting and preventing fraud (Jones 2013). The name of the tool is the Accounting Quality Model (AQM), dubbed "RoboCop" by some in the press. The tool will draw from XBRL tagged data, data that are currently accessible on the EDGAR database, to identify filers who deviate from reporting patterns compared to industry peers, especially in how discretionary accruals are managed. If significant departures are detected, AQM will flag the disclosure, which may lead to immediate investigation. The SEC is not limited to XBRL filings. As currently defined, AQM does not fit the profile of a Big Data implementation. The Edgar/XBRL data are assumed to be measured in gigabytes, but they will be in terabytes in a few years. A relational database almost certainly serves as the repository for their data, as there would be no need to distribute the data on a cluster of computers.

While the primary source of data is currently XBRL, it is conceivable that the SEC will begin to gather much more data on companies in the same way data are gathered about individuals. Phone records, Internet activity, purchasing behaviors, and social media activities could all be gathered at the company level. Individual employee behavior and customer behavior may also be of interest to the SEC, as employees have increased ability to impact a company's bottom line with the growth of social media. The Justice Department has often requested some of these data in the case of litigation. Google and internet service providers often have requests from the government and courts of very large data troves. As the diversity and amount of data grow, the SEC will be required to implement and maintain Big Data platforms, whether it is under the auspices of the AQM program, or a related one. As continuous auditing and continuous control monitoring are increasingly implemented, the SEC may request access to these data to protect the investors. There are many effects to this expanded scope of data activity. For example, foreign companies may balk at making information available, German laws are very restrictive on database information dissemination, etc.

Many questions arise in this domain: (1) Should the government's tools (like AQM) have unfettered access to corporate information systems? (2) Should government's supervision be automated and real time? (3) Can algorithms/analytics be developed that provide quality exception reporting minimizing false positives from this ocean of data? (4) Should the laws/regulations be different for the government, organizations, and individuals?

V. BIG DATA AND ACCOUNTING RESEARCH

Inevitably, accounting researchers are going to use forms of Big Data in their research, although enormous difficulties exist. For example, empirical research data could be updated in real time as companies release their financial results, and other forms of relevant data (e.g., trading volumes, prices, news pieces) emerge. This would increase the reliability of accounting models and show that they withstand the test of time. When these become outdated as behaviors change, new models can be introduced or inferred from the data itself.

Much in the same way the SEC harnesses data for their Audit Quality Model, the EDGAR and WRDS databases could become a local Big Data repository for accounting researchers. Data from other sources would also enrich future analyses including any company documentation (including archives of their web content), data produced by third parties such as Yahoo-Finance, transcripts and audio from quarterly calls to analysts, news headlines, Twitter feeds, and social media. To date there is limited accounting research that uses Big Data to derive results, probably due to a lack of access to Big Data in the way corporations, governments, and not-for-profits have it. With open-source software and commoditized hardware, Big Data should be available for accounting research. Other data-intensive disciplines already use Big Data, such as medicine, physics, meteorology, and biology. The challenge for accounting research is to become data intensive when

(apart from public filings) organizational data is not always easy to obtain. Furthermore, another serious challenge is that most data are not standardized and there may be substantive cost in their pre-processing. Public good would be clearly served if large research-oriented public financial related databases could be made available to the accounting research community.

Accounting and Audit Practice with Big Data

Given that accounting is the art of recording business operations and reporting these results, ERP systems are already collecting a wide new range of data. In fact, the alternate data surrounding operations can be much more informative in terms of measurement of business than exclusively financial transactions. As *accountants* are responsible for gathering and reporting information that is useful to management, then there is a role for them in Big Data and Data Analytics.

Ideally, all organizational data should be available for data mining to improve recording of events, reporting to regulators, and enforcing internal controls. As an illustration, new analytic technologies facilitated by emerging data processing methods can be used. Process mining (Jans et al. 2010) of Big Data can ensure efficiency, reduce waste, provide assurance for separation of duties, and discover asset misappropriations.

Auditors can also take advantage of Big Data. Auditors should seek to verify transactions, not with just an invoice and receipt, but multi-modal evidence that a transaction took place. Photo, video, GPS location, and other meta data could accompany transaction data.[5] If implemented, continuous monitoring and auditing (Vasarhelyi et al. 2010) will create Big Data that will help improve audit efficiency and efficacy. Data from each audit should be digitized and stored for retrieval and analysis. From the sales pitch to the working papers, everything could be recorded and later analyzed to improve the entire auditing process. These issues are further expanded in the ensuing section.

Big Data in Business Measurement, Assurance, and Standard Setting

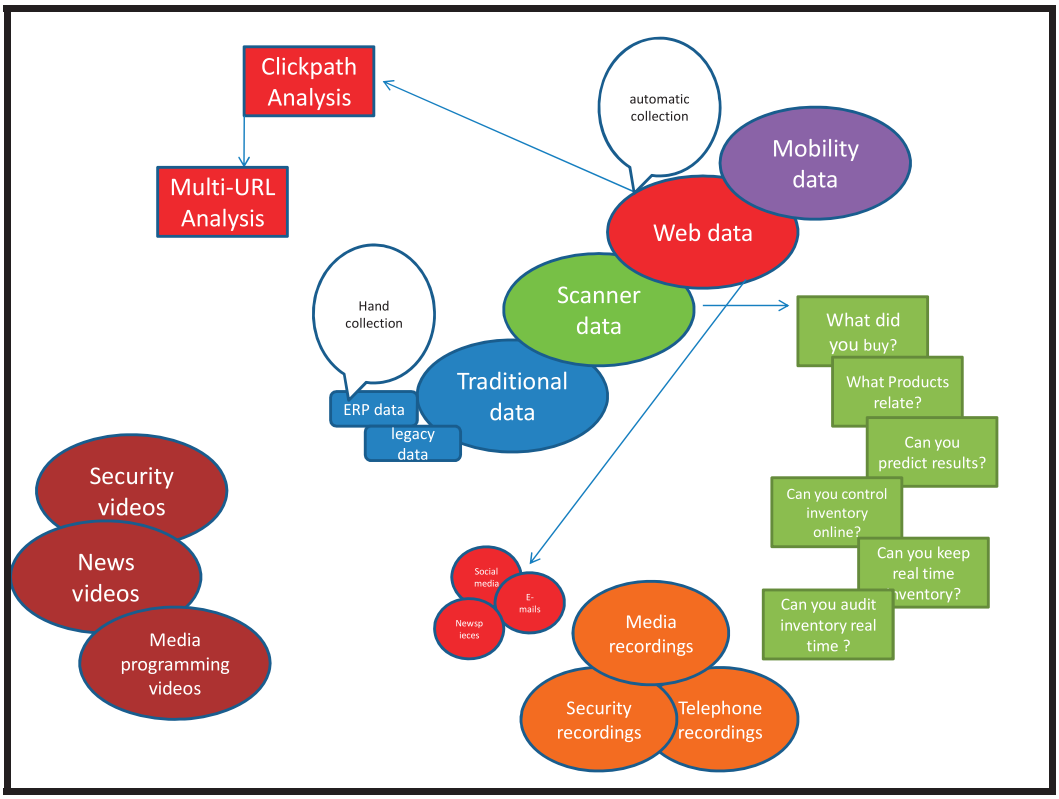
Corporate usage of data has evolved dramatically toward a much larger set of uses. The Enterprise Data Ecosystem (EDE) is exponentially expanding, and this environment presents a dynamically changing set of characteristics that most likely will require the development of enhanced theory of information (Shannon and Weaver 1949). This theory will require recognition of the nature of the data capture (manual versus automatic), the volume of the data, the efficiency of the integration with the existing data corpus, the efficiency of the transformation of information into data, the granularity of data, types of operation/decision supported, and other variables.

Figure 1 represents the expanding ecosystem of corporate data usage. At the core are *traditional data*, which typically incorporate legacy and ERP data. These data were preponderantly, manually acquired in the initiation of transactions. With the advent of *scanners*, data expanded in scope by collecting details, for example, of items being acquired at the cash register. This allowed substantive increase in data analysis applications including inventory control, and detecting related products and individual product preferences. This data collection tends to be automatic and, consequently, presents substantively fewer economic reasons for limiting its capture. Linked with traditional data, opportunities are very rich for additional analytics and for increasing operational value. Data automatically mined from the World Wide Web can include URL, click-path, and content data, adding information from a substantively different source. Click-path data are particularly interesting, as one may follow step by step the customer information acquisition and decision process. The advent of ubiquitous cell phone use (mobility data), presents an additional

[5] Data used for accounting objectives may also serve for security, to drive marketing, or to manage operations.

American Accounting Association

FIGURE 1
Emerging Sources of Data



major type of data collection that includes, among other fields, the location of the user. Although locational information may be inferred from other sources, mobility data provide close-time intervals location data. Linking this to the previous types of data could provide powerful explanations and predictive information on customer behavior.

Furthermore, two very large data domains, which in volume dwarf the above-mentioned data, are still separate, but eventually will become part of the Big Data ecosystem: video and audio data. Audio streams (such as utterance or telephone recordings), media audio streams, and audio surveillance streams, can be recorded, eventually, automatically transcribed into text, and associated to client data. Vocalic data from the audio streams, such as pitch and intonation, can relate stress and vocal dissonance to each utterance. Video surveillance, news-piece videos, and cell phone video recordings represent an enormous, continuous flow of data that is still mostly self-contained and separate from other data sources. However, with tools such as automatic face recognition, video threat assessment software, and other evolving tools, they can be progressively linked to the EDE.

Although many early Big Data applications come from the physical sciences and from other self-contained domains, business is finding greater value when it links automatically captured and outside data sources. For example, a bank would *traditionally* capture loan applicant data in forms that would then be entered into their systems manually. Loan decisions would be made after applying analytics and standardized business rules to these data. Now, with automatically captured

data integrated into the EDE including tracked visits and clicks in websites, online questionnaire responses, location information captured through cell phones, client telephone calls, and video captured from surveillance cameras, a much more complete picture of loan applicants can be developed.

Table 1 examines some of the expanded data sources, the multiplicity of parameters that can be captured at limited cost, the meta data explaining these files, the sources of the data, and the linkages facilitated by the EDE. Several insights may be of value to mention:

1. There is a progressive extension of the feasible dataset. Inclusion of sources is mainly an economic and legal issue and not one of feasibility.
2. Newly included data structures contain a wide set of not previously determined/used parameters that by themselves may be informational.
3. Prior existing data may expand in informational value by their collection in more granular form, by their additional parameterization to allow the creation of more explanatory data cubes, by their analysis in time sequence or in parallel with nontraditional data sources, etc.
4. Mathematical relationships can be derived among different elements or processes linking their effects or validating their measurement (Kogan et al. 2011).

In Table 1, the enhanced EDE, with particular emphasis in accounting and auditing, is exemplified. The "Sources of Data" expand with more content of existing data (additional parameters are kept as their collection is costless). For example, the time of data entry, approval, and execution are kept. Furthermore, now emails and social media comments can be obtained and connected to the transactions. The metadata is informing about the nature of the data fields and other tags. This information is not only informative, but also if formalized and standardized (such as XBRL) can be processed and used by automation tools. For example, names and formats of fields are kept and may be automatically used to access, interchange, and store data from external sources. The "Meta-Meta Data" uses the expanded dataset on a collective basis for more aggregate, strategic evaluation. For example, data can be gathered of purchase paths and qualified by product. "New Sources of Business Data" are a narrower form of the first column. The advent of interconnectivity allowed elements such as RFID chips and real-time trading prices to be included in transaction databases and data streams. A particular case of these is the "New Sources of Financial Data" that now include many of these sources and data such as XBRL and EDGAR filings. Finally, "New Linkages Facilitated by IT and Analytic Technologies," include a very rich potential set of new information that will emerge and eventually will be fully automated into the EDE.

Business Measurement (Accounting) and Big Data

Big Data offers tremendous opportunities in the area of business measurement, although some of these changes are so fundamental that they will take time to achieve. For example:

1. The current basic posting element is the journal entry but some entries in large corporations have hundreds or thousands of lines in one posting. Automatic data collection allows for a richer set of attributes. A finer measurement may be desirable. Journal entries can tie directly to supplemental measures (elements of an invoice, physical parts received, and their RFIDs).
2. Account names and content are inherited from the era of manual accounting. Today, substantive new detail can be provided at small cost and great explanatory value. For example, values such as the type of product in inventory, the location of this inventory, the supplier of the parts, and the physical age of inventory can enrich the information in existing data structures.

TABLE 1

Illustration of Data Expansion and Measurement Consequences

Sources, Content, and Enhanced Content (Content)	Parameters (Meta Data)	Meta-Meta Data	New Sources of Business Data	New Sources of Financial Data	New Linkages Facilitated by IT and Analytic Technologies
News pieces			RFID data		
	Source		Detailed transaction data		Use news mentions, or emails, or click- path analysis to predict sales
	Time		B2B transaction data	B2B transaction data	
	Publication		Blog postings		
	Topic	Frequency over time	Emails		
		Change in nature over time	Social media postings		
Emails			Emails linked to transactions with attachments	XBRL/FR data	Automatic classification and addressing of emails
	To whom			XBRL/GL data	
	From whom			EDGAR data	
	Topic			FDIC call reports	
		Frequency over time			
		Change in nature over time			
Social Media		Groups of people that behave similarly		Comments on blogs about companies	Automatic response and tone detection
	To whom				
	From whom				
	Topic				
Click-path		Purchase paths by product		Paths of fraudulent behavior	Identify fraudulent user groups
	Website visited				
	Pages visited				

3. Analytically derived estimates of sales, costs, product mix, human resource turnarounds, etc. can be provided at very little cost.
4. Drill-downs of an explanatory nature, tempered to avoid competitive impairment, can be provided.

These are just a few potential changes, many of which are not revolutionary or are already contained in the ERPs that currently run business, and which databases are already enhanced by online analytical processing (OLAP) cubes (Vasarhelyi and Alles 2006) and business intelligence (Sutton et al. 2013). This extraordinary potential expansion of business reporting clearly will be balanced by organizational disclosure reticence in exchange for corporate and stakeholders needs.

When many of the present accounting and auditing standards were enacted, current information technologies did not exist (Titera 2013). The trade-offs between the costs and benefits of disclosure dramatically changed (Vasarhelyi and Alles 2006) and were not yet reflected in the accounting model. In general, if a database can be made available with requisite filtering (Gal 2008), the rules of disclosure do not matter, as they can be created from the raw data (Vasarhelyi 2012b). Therefore, accounting standards will have to deal with the content of the databases and allowable sets of extractions but not with the particular rules of account disclosure. Technologies such as relational databases, linkages with textual data, textual data analytics, drill-downs, census-like filtering of details (Gal 2008), and XBRL can be extensively used for disclosure.

Assurance and Big Data

Big Data also offers tremendous opportunities for the area of business assurance. The need for evolution and facilitation of assurance has been illustrated by the AICPA's issuance of the Audit Data Standard (Zhang et al. 2012; Titera 2013), which focuses on detailed data specification in different formats (flat file and XBRL/GL), not on aggregate measures. Many other enhancements on the assurance model may be desirable. For example,

1. Auditors may have to be able to acquire extensions of the current corporate data that are not in the financial domain to confirm the existence of events. In particular, sales of electronic goods where there is no storage and physical flow have to be verified by different methodologies including the usage of extended Big Data that is not currently kept in corporate records. This would imply the need for auditors to request and receive data from corporate data centers that are not kept today for operational purposes.
2. Auditors may have to rely on analytic models that link markets, operational data, sales, and post-sales efforts to validate reporting elements (Kogan et al. 2011).
3. Auditors may have to suggest new processes of verification due to the unauditability of large data systems in their current form. Among these processes are automatic confirmation (Vasarhelyi 2003) and process mining (Jans et al. 2010).
4. Many auditors' interests and needs coincide with those of management. New mechanisms to assure objectivity (not independence) need to be developed to facilitate and enrich the internal/external audit function and to avoid the duplication of analytic efforts by managers and assurers.
5. Big Data substantially expands the scope of corporate data collection and retention. This scope must be often validated, in particular endogenous data, to support the numbers derived by corporate models. On a lesser scale this effort is already being performed, mainly in an *ad hoc* manner for the lack of specific guidance, by auditors verifying derivative instruments.
6. New forms of audit evidence such as alarms/alerts (Vasarhelyi and Halper 1991), text mining, E-discovery, continuity equations (Kogan et al. 2011), and the search for

exceptional exceptions (Issa 2013) will arise and be used to complement or to replace certain forms of traditional audit evidence.

These are just a few potential changes that may need to be enacted in the future. These changes will necessarily change the focus of the assurance process. Automatic confirmation will limit the need for verification of population and data integrity by extending the boundaries of the audit process to outside of the entities' environment. Database field controls will further limit the need for verification of the value of transactions. Continuity equations will provide dimensional checks of the value of accounts. These changes will have to be reflected in minimal assurance standards.

Accounting and Auditing Standards and Big Data

It is not Big Data *per se*, but the restructuring and reconceptualizing of accounting and auditing that drives the efforts in standard setting in the future information environment. Krahel (2011) argued that *de facto* ERP developers drive the transformation of fuzzy rules issued by standard setters into the "rule based" (Schipper 2003) software that executes most accounting standards. Consequently, much of the effort by standard setters is focused on clarification of the original rules. In addition to the need of specificity in the formulation of rules, in order for these to be implemented in ERP environments, (now with the enhanced Big Data environment) a set of more drastic changes toward the formulation of standards is needed:

1. Standards in financial reporting:
   a. Notwithstanding the de-emphasis on specific rules of disclosure, as the financial report is "just a layer" (Vasarhelyi 2012a), comparability will continue to be a driving force. Without comparability, assessment for resource allocation and transparency for stakeholder assessment become very difficult. But comparable disclosures are to be just one of the elements of the reporting layer, not the only report.
   b. Disclosure rules, in the measurement domain, will have to focus on basic data to be provided, in particular to its content, timing, and level of aggregation.
   c. Rules of disclosure will have to deal with a much finer level of disclosure of the entity being measured (business, business unit, subdivision, product, etc.).
   d. Alternate guidance will have to be provided for disclosures to different stakeholder groups.
   e. The "one report for all" approach needs to be changed allowing for strategic drill-downs for relevant details. Special reports for comparability are needed, but do not serve for all purposes.
2. Standards in auditing standards:
   a. Assurance will have to abandon the traditional concept of sampling the population and understand the implication of Big Data and the ability of obtaining full population analyses.
   b. Population integrity and data computation integrity will be largely achieved with automatic confirmation and the usage of electronic data processing.
   c. Methodologies will have to be developed to deal with cloud storage integrity (Weinman 2012), necessary data redundancy, distribution of data, and country-related differences.
   d. Methodologies will have to be developed for much more frequent, and eventually real-time (Vasarhelyi et al. 2010), assurance measures.
   e. Preventive and predictive audit approaches (Kuenkaikaew 2013) may be formulated and guidance issued. With this guidance, issues of independence of auditors, materiality of estimates, coincidence of management, and auditing procedures must be considered.

f. In general, with the change in information processing technology, the quantity of data and structures of the business, economics, and objectives of the assurance (verification) process must be rethought.

New Roles, Tools, and Evidence in Business Measurement and Assurance Due to Big Data

The difficulties of auditor data procurement have been extensively discussed in the literature (Vasarhelyi et al. 2012). Big Data adds a new dimension to this problem, as the sources are larger, their extraction more complex, the costs of this access substantive and, most of all, the data more revealing. Furthermore, new methodologies of auditing, such as continuity equations, allow for substantive linkages between variables and processes but require the auditor to request data that the company may not normally retain.

Big Data substantively expands the scope of potential analytical usage and data to be used in the audit process. It is not only restricted to the usage of new tools or data sources, but also expands the potential usefulness of existing tools that can integrate them into new methodologies. The large stores of data create, and greatly expand, the potential of Exploratory Data Analysis (Tukey 1977; Liu 2013b) in an analogous approach to what is now called data mining (Hand et al. 2001). A conceptual confusion arises from the often advocated *a priori* theory construction to be followed by capture of data and then Confirmatory Data Analysis (CDA). The advent of Big Data allows for a pragmatic exploration of data to develop testable assertions without the fear of over fitting. By and large, narrowly focused CDA tests are a product of the paucity of data available and the difficulty of calculations prior to modern computer technology. New audit analytic techniques and the creation of new audit evidence are of major interest.

It is important for accountants and accounting researchers to understand the issues around establishing Big Data repositories, populating the repositories, and querying and analyzing those repositories using modern tools such as NoSQL,[6] cutting edge machine learning algorithms, and traditional statistics. Repositories with flexible schema, distributed across many nodes, are necessary to handle the volume and complexity of the data. Facebook-developed Cassandra,[7] and Apache's HBase[8] are two well-known nonrelational databases that may be of help. MapReduce[9] as an approach for managing the computational problems of Big Data may be desirable. Hadoop,[10] the most popular implementation of MapReduce, could serve as the core technology and Amazon's cloud services[11] could host the Hadoop implementation.

Emerging Audit Analytics

Kogan et al. (2011) have proposed the usage of continuity equations to link processes, to model lagged information processes, and to perform automatic error correction. With Big Data many unorthodox linkages may be attempted, for example (1) calls recorded in a telephone switch and collection, (2) calls recorded and customer support, and (3) client complaints in blogs with customer care.

Thiprungsri and Vasarhelyi (2011) have used cluster analysis for anomaly detection in accounting data. Multidimensional clustering can serve on many aspects of the assurance and be

[6] http://nosql-database.org/
[7] http://Cassandra.apache.org
[8] http://Hbase.apache.org
[9] http://research.google.com/archive/mapreduce.html
[10] http://Hadoop.apache.org
[11] http://Aws.amazon.com

used as an Exploratory Data Analysis (Tukey 1977; Liu 2013a) tool leading to the development of testable assertions and hypotheses (Liu 2013b).

Jans et al. (2010) focused on applying process mining on the audit process. This entails extracting process logs of a company's ERP and examining the path followed by a transaction being processed. This approach is technologically facilitated and allows for the examination of how transactions are executed notwithstanding the content. Several fields are using this approach and many tools are being developed that can be used in assurance processes.

New Forms of Audit Evidence

New forms of audit evidence are progressively emerging to complement and replace old approaches and are covering a more recent set of risks. In addition to the evidence generated by the three forms of analytics described above, we may obtain evidence such as (1) alarms and alerts (Vasarhelyi and Halper 1991), (2) text mining, (3) E-discovery, and (4) massive data and exceptional exceptions, etc. Furthermore, the traditional evidential data used were mainly from internal (endogenous) sources, usually relying on confirmations for external system validation. Big Data will place at the hands of the accountant/auditor, enormous amounts of external (endogenous or exogenous) data that can serve in model building or on the creation of competitive baselines. Businesses can also provide analytics about their performance without providing direct data (transactions, customers, suppliers) using KPIs, KRIs, or other summary information.

VI. CONCLUSIONS

The advent of massive data stores and ubiquitous access is a paradigmatic change in the operations of organizations. This change is already progressing into financial processes but has not yet been substantively impounded into accounting and auditing. This note describes Big Data and its features as it is affecting organizations, and prepares a framework of potential research topics to be examined when Big Data is reflected/integrated into accounting, auditing, and standards. The key issues raised by Vasarhelyi (2012a) and discussed in the introduction of this note were discussed *vis-à-vis* Big Data. Some of the inferences drawn, that have deep implications in accounting research are:

a. Measurement and representation methods
   - The accounting model must evolve/be changed to focus on data content, atomicity, data linkages, etc.
   - Accounting standards will have to deal with the content of large databases and allowable sets of extractions, not with extant rules of account disclosure.
   - Process mining can add a new dimension to process management and assurance by focusing on transaction paths, not the content of information.
   - Disclosure rules, in the measurement domain, will have to focus on basic data to be provided, in particular to its content, timing, and level of aggregation.
   - Rules of disclosure will have to deal with a much finer level of disclosure of the entity being measured (business, business unit, subdivision, product, etc.).

b. Formalization
   - Database field controls will further limit the need for verification of the value of transactions.
   - Continuity equations will provide dimensional checks of the value of accounts.
   - Meta data and "meta-meta" data are allowing for a higher level of formalization (and automation) of data collection, reporting, and decision making.

c. Semantic understanding

- Big textual data are available to accounting researchers now.
- Text understanding and vague text understanding can provide the necessary links from textual elements to the more traditional ERP data.
- Eventually, the vocalic and video data would also be progressively linked to the more traditional domains.
- The AIS, accounting, and finance research communities have already made progress in how to process and impound it into research.

d. Assurance procedures
- Automatic confirmation will limit the need for verification of population and data integrity.
- Auditors should seek to verify transactions not with just an invoice and receipt, but with multi-modal evidence that a transaction took place. Photo, video, GPS location, and other meta data could accompany transaction data.
- Auditors may have to be able to acquire extensions of the current corporate data.
- New forms of audit evidence are to complement and replace old approaches, and are covering a more recent set of risks.

e. Audit economics, social welfare, and other added issues
- Public good would be served if large research-oriented public financial related databases could be made available to the accounting research community.
- Accounting education will have to evolve educating faculty, professionals, and students in the issues of Big Data and data analytics.

—Kevin C. Moffitt, Assistant Professor
—Miklos A. Vasarhelyi, Editor
Rutgers, The State University of New Jersey

REFERENCES

Aizawa, A. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing and Management* (January): 45–65.

Bovee, M, A. Kogan, K. Nelson, R. P. Srivastava, and M. Vasarhelyi. 2005. Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and Extensible Business Reporting Language (XBRL). *Journal of Information Systems* 19 (1): 19–41.

Cong, Y., and J. Romero. 2013. On information system complexity and vulnerability. *Journal of Information Systems* (Fall).

Dickinson, E. 2011. The first Wiki Leaks revolution. *Foreign Policy* (January).

Economist, The. 2013. *In the Secret State*. Available at: http://www.economist.com/news/united-states/21582536-public-opinion-may-be-shifting-last-against-government-intrusiveness-secret

Franks, Bill. 2012. *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. New York, NY: Wiley.

Gal, G. 2008. Query issues in continuous reporting systems. *Journal of Emerging Technologies in Accounting* 5 (1): 81–97.

Gartner. 2011. *CEO Advisory: "Big Data" Equals Big Opportunity*. Available at: http://www.gartner.com/id=1614215

Geerts, G. L., and W. McCarthy. 2002. An ontological analysis of the economic primitives of the extended-REA enterprise information architecture. *International Journal of Accounting Information Systems* 3: 1–16.

Gilder, G. 2006. The information factories. *Wired* (October).

Hand, D. J., H. Mannila, and P. Smyth. 2001. *Principles of Data Mining (Adaptive Computation and Machine Learning)*. Cambridge, MA: MIT Press.

Issa, H. A. 2013. *Exceptional Exceptions*. Ph.D. dissertation draft, Rutgers Business School.

Jans, M., M. Alles, and M. A. Vasarhelyi. 2010. *Process Mining of Event Logs in Auditing: Opportunities and Challenges*. Working paper, Hasselt University.

Janvrin, D., R. Pinsker, and M. Mascha. 2013. XBRL-enabled, Excel, or PDF? Factors influencing exclusive user choice of reporting technology for financial analysis. *Journal of Information Systems* (Fall).

Jones, A. 2013. SEC to roll out Robocop against fraud. *Financial Times* (February).

Kogan, A., M. G. Alles, M. A. Vasarhelyi, and J. Wu. 2011. *Analytical Procedures for Continuous Data Level Auditing: Continuity Equations*. Working paper, Rutgers Accounting Research Center.

Krahel, J. P. 2011. *Formalization of Accounting Standards*. Dissertation proposal, Rutgers Business School.

Kuenkaikaew, S. 2013. *Predictive Audit*. Ph.D. dissertation draft, Rutgers Business School.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25 (2–3).

Lee, T. B. 2013. Here's everything we know about PRISM to date. *The Washington Post* (June).

Liu, Q. 2013a. *Exploratory Data Analysis in Accounting*. Ph.D. dissertation proposal, Rutgers Business School.

Liu, Q. 2013b. *The Application of Exploratory Data Analysis in Auditing*. Dissertation proposal, Rutgers Business School.

McAfee, A., and E. Brynjolfsson. 2012. Big data: The management revolution. *Harvard Business Review* (October): 60–66.

Rifkin, J. 2001. *The Age of Access: The New Culture of Hyper Capitalism, Where All of Life is a Paid-for Experience*. New York, NY: Tarcher.

Romero, S., G. Gal, T. J. Mock, and M. A. Vasarhelyi. 2013. A measurement theory perspective on business measurement. *Journal of Emerging Technologies in Accounting* (forthcoming).

Schipper, K. 2003. Principles-based accounting standards. *Accounting Horizons* 17 (1): 61–72.

Shannon, C. E., and W. Weaver. 1949. *The Mathematical Theory of Information*. Urbana, IL: University of Illinois Press.

Srivastava, R., S. Rao, and T. J. Mock. 2013. Planning and evaluation of assurance services for sustainability reporting: An evidential reasoning approach. *Journal of Information Systems* (Fall).

Steinbart, P., R. Raschke, G. Gal, and W. Dilla. 2013. Information security professionals' perceptions about the relationship between the information security and internal audit functions. *Journal of Information Systems* (Fall).

Sutton, S., M. Elbashir, P. Collier, M. Davern, and S. Leech. 2013. Enhancing the business value of business intelligence: The role of shared knowledge and assimilation. *Journal of Information Systems* (Fall).

Thiprungsri, S., and M. Vasarhelyi. 2011. Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research* (July).

Titera, W. R. 2013. Updating audit standard—Enabling audit data analysis. *Journal of Information Systems* (Spring).

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Van der Heijden, H. 2013. Evaluating dual performance measures on information dashboards: Effects of anchoring and presentation format. *Journal of Information Systems* (Fall).

Vasarhelyi M. A., M. G. Alles, and K. T Williams. 2010. *Continuous Assurance for the Now Economy*. Sydney, Australia: Institute of Chartered Accountants in Australia.

Vasarhelyi, M. A. 2003. *Confirmatory Extranets: A Methodology of Automatic Confirmations*. Grant proposal, Rutgers Accounting Research Center.

Vasarhelyi, M. A. 2012a. AIS in a more rapidly evolving era. *Journal of Information Systems* (Spring).

Vasarhelyi, M. A. 2012b. Financial accounting standards do not matter: It's just a layer. *Journal of Information Systems* (Fall).

Vasarhelyi, M. A. 2013. Formalization of standards, automation, robots, and IT governance. *Journal of Information Systems* (Spring).

Vasarhelyi, M. A., and F. B. Halper. 1991. The continuous audit of online systems. *Auditing: A Journal of Practice & Theory* 10 (1): 110–125.

Vasarhelyi, M. A., K. Nelson, A. Kogan, and R. Srivastava. 1999. Inquiring information systems in the boundary–less world: The FRAANK example. Proceedings of the 1999 Americas Conference on Information Systems (AMCIS), Milwaukee, WI, August.

Vasarhelyi, M. and M. Alles. 2006. *The Galileo Disclosure Model (GDM): Reengineering Business Reporting Through Using New Technology and a Demand Driven Process Perspective to Radically Transform the Reporting Environment for the 21st Century*. Available at: http://raw.rutgers.edu/gdl/Galileo.

Vasarhelyi, M., and J. P. Krahel. 2011. Digital standard setting: The inevitable paradigm. *International Journal of Economics and Accounting* 2 (3): 242–254.

Vasarhelyi, M., S. Romero, S. Kuenkaikaew, and J. Littley. 2012. Adopting continuous audit/continuous monitoring in internal Audit. *Information Systems Audit and Control Association Journal* 3.

Weinman, J. 2012. *Cloudonomics: The Business Value of Cloud Computing*. New York, NY: John Wiley & Sons.

Zhang, L., A. R. Pawlicki, D. McQuilken, and W. R. Titera. 2012. The AICPA assurance services executive committee emerging assurance technologies task force: The audit data standards (ADS) initiative. *Journal of Information Systems* (Spring): 199–205.