THE PROS AND CONS OF USING BIG DATA IN AUDITING: A SYNTHESIS OF THE LITERATURE AND A RESEARCH AGENDA

Michael Alles

Rutgers Business School

Department of Accounting & Information Systems

One Washington Park, Room 928

Newark, NJ 07102-3122

alles@business.rutgers.edu

Glen L. Gray

Dept. of Accounting & Information Systems

College of Business & Economics

California State University, Northridge

18111 Nordhoff Street

Northridge, CA 91330-8372

glen.gray@csun.edu

September 14, 2015[*]

THE PROS AND CONS OF USING BIG DATA IN AUDITING: A SYNTHESIS OF THE LITERATURE

Abstract With corporate investment in Big Data of $34 billion in 2013 growing to $232 billion through 2016 (Gartner 2012), the Big 4 professional service firms are aiming to be at the forefront of Big Data implementations. Notably, they see Big Data as not only an opportunity for their consulting arms, but also as an increasingly essential part of their assurance practice. We argue that while there is a place for Big Data in auditing its application to auditing is less clear cut than it is in the other fields to which it already has been applied, such as marketing and cancer research. The objectives of this paper are to: (1) provide a balanced discussion of both the pros and the cons regarding incorporating Big Data into financial statement audits; and (2) present a research agenda to identify specific aspects of Big Data that could benefit auditors.

1. Introduction

1.1   Increasing interest in Big Data by Auditors

A survey by Gartner (2014) found that 73% of respondents had invested or planned to invest in Big Data in the next 24 months, up from 64% in 2013. With corporate investment in Big Data of $34 billion in 2013 growing to $232 billion through 2016 (Gartner 2012), the Big 4 professional service firms are eagerly competing with such firms as SAP and IBM for the growing market in Big Data. Notably, they are declaring that Big Data is not only an opportunity for their consulting arms, but also an increasingly essential part of their assurance practice.[1] For example, EY states: *"The audit of the future will bear little resemblance to the traditional audit CFOs are accustomed to receiving today. In fact, the way organizations conduct audits will change more in the next 5-10 years given the evolution of technology and analytics. Data analytics, new technology and access to detailed industry information will all combine to help auditors better understand the business, identify risks and issues and deliver additional insights. Moreover, the ability to review and analyze entire sets of data, rather than applying sampling techniques, will help bring more confidence to the audit."*[2]

Similarly, Deloitte Chairman and CEO Joe Ucuzoglu writes: *"At Deloitte we're investing several hundred million dollars in data analytics and artificial intelligence with some cutting-edge applications that we really believe differentiate us and our audit approach. When we use these tools, we're able to get greater coverage. We're able to more quickly identify risks. We're able to complete the audit with a higher level of quality and ultimately deliver a greater level of insight to our clients."*[3]

In the light of these dramatic and costly developments in practice, we examine in this paper the pros and cons of using Big Data as an audit tool. With so much of the attention emphasizing the former, we unapologetically point out the latter in order to get a more balanced perspective on what promises to be a revolutionary change in auditing. Our objective is not to be naysayer to this change,

---

[1] Although accounting firms could conduct a wide variety of engagements under a variety of professional standards; this paper focuses on financial statement audits conducted in compliance with the Statements of Auditing Standards (SASs) published by the AICPA, Auditing Standards (ASs) published by the PCAOB, and International Standards on Auditing (ISAs) published by the IAASB.

[2]   http://www.ey.com/GL/en/Issues/Managing-finance/EY-cfo-need-to-know-future-of-audit?preview&HL=CON-USDD-9XAN4E&utm_source=eycom&utm_medium=homepage_PF&utm_campaign=Future%20of%20audit#introduction. Accessed 9/1/2015 1:26:52 PM.

[3]   http://www2.deloitte.com/us/en/pages/audit/articles/deloitte-prepares-firm-for-audit-of-the-future-accounting-today-article.html. Accessed 9/1/2015 1:32:53 PM.

but to play the role of a devil's advocate, arguing for the need for auditors to exercise due care when embracing this new methodology.

We will show that while there is a place for Big Data in auditing its application to auditing is less clear cut than it is in the other fields to which it already has been applied, such as marketing and cancer research.[4] To put our approach into proper perspective, we refer to AU Section 230 paragraph .07 which states: *"Professional skepticism is an attitude that includes a questioning mind and a critical assessment of audit evidence. The auditor uses the knowledge, skill, and ability called for by the profession of public accounting to diligently perform, in good faith and with integrity, the gathering and objective evaluation of evidence."*[5] Skepticism is not a negative in auditing, but the positive force from which value is created. Similarly, we bring professional skepticism to bear on the potential role of Big Data in auditing practice in order to better understand when it will add value and when it will not.

1.2   The Inevitably of Big Data in Auditing Versus the Historical Record

At a theoretical or normative level it seems logical that auditors will incorporate Big Data into future financial audits (Littley, 2012; Moffitt and Vasarhelyi, 2013; Setty and Bakhshi, 2013; IIA 2013; Whitehouse 2014; AICPA 2014). Big data potentially paints a very robust picture of the client's activities, far beyond that possible to be captured using only the far smaller population of GAAP-compliant data. A key requirement of auditing standards is setting expectations at the beginning of the audit—an issue that PCAOB inspectors are particularly critical of in their reports. A core component of audit planning is establishing *"expectations developed from data derived from other sources"* (AU 329.16) before collecting and analyzing financial data.[6] Setting expectation based on "disaggregate data" is a reoccurring theme in Staff Audit Practice Alert No. 12 (PCAOB 2014) to improve auditing of revenue, which was motivated by the PCAOB's frequent criticism of revenue auditing in their accounting firm inspection reports. The PCAOB is also pushing accounting firms to boost their fraud detection procedures in compliance with AU 316.

Zhang, Yang and Appelbaum (2015), Cao, Chychyla and Stewart (2015) and Yoon, Hoogduin, and Zhang (2015) argue that Big Data will enable auditors to improve their effectiveness

---

[4] Interestingly, PWC itself advertises their use of Big Data methods to help cancer researchers, presumably as part of their consulting practice: http://extraordinary.pwc.com/?utm_campaign=Extraordinary&utm_medium=Feature&utm_source=Omnibus#medical. Accessed 9/1/2015 1:41:17 PM.
[5] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-00230.pdf. Accessed 9/2/2015 9:24:36 AM.
[6] http://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-00329.pdf. Accessed 9/2/2015 9:26:51 AM.

in all these areas and more. These articles, all in the Big Data Forum published in the June 2015 issue of *Accounting Horizons*, assume the inevitability of incorporating Big Data into financial statement auditing. In September 2015 the American Accounting Association (AAA) held its first "Accounting is Big Data" conference which one of the authors attended.[7] In this conference too it was simply taken for granted that Big Data will be the future of auditing, though few examples of its use specific to accounting or auditing were presented.

In this paper we temper this optimism about the use of Big Data in auditing in light of the empirical evidence that accounting firms have had a mixed history of implementing assurance-related technologies. The 1996 *Special Committee on Assurance Services Report* (the *Elliott Report*) predicted that the annual fees for the assurance services recommended in that report would reach $14 to $21 billion (two to three times the $7 billion annual fees collected for financial statement audits). E-commerce assurance services (later called WebTrust) alone was predicted to generate $1.25 to $6.25 billion annually, far in excess of the reality today.

Regarding technology for financial statement audits specifically, articles equally optimistic to the *Accounting Horizons* Big Data articles (some by the same authors) were published in the past about artificial intelligence (AI), expert systems, data mining, and continuous auditing [CA]. Over 30 years ago, the then Big 6 were spending millions of dollars developing expert system applications. Brown (1991) listed 43 expert systems in use or under development by the then Big 6; yet according to Gray, Chiu, Liu, and Li (2014), none of these applications became part of regular audit activities. Researchers have been promoting continuous auditing and continuous monitoring [CM] for nearly a quarter century (Vasarhelyi and Halper 1991), but as the AICPA states, *"The general view is that not much is currently being done with CA/CM."*[8] More recently, articles (e.g., Debreceny & Gray 2010 and 2011 and Gray and Debreceny 2014) have illustrated how various forms of data mining could be valuable resources for auditors, but generally this technology is isolated to special forensic engagements.

Notwithstanding this discouraging record, we are not saying the accounting firms made mistakes not implementing these and other technologies as a regular part of traditional financial audits. Audit standards AS No. 15, SAS No. 122 (AU-C sec. 500) and ISA 500, all essentially say [Quoting from SAS No 122.04]: *"The objective of the auditor is to design and perform audit procedures that*

enable the auditor to obtain <u>sufficient appropriate</u> audit evidence to be able to draw <u>reasonable conclusions</u> on which to base the auditor's opinion." [Emphasis added]. Since tens of thousands of audits are completed each year, auditors currently believe they are collecting *sufficient appropriate* audit evidence to draw *reasonable conclusions* without these technologies. As such, the overarching research question regarding new technologies and financial statement audits is: how or where can these technologies cost-effectively improve the efficiency and/or effectiveness of audits? More specifically to Big Data analytics: what parts of the current data gathering and analysis procedures can Big Data analytics be cost-effectively *substituted*? We say substituted because we assume Big Data analytics will not simply be added to current audit procedures—and that accounting firms will just increase their audit fees to reflect this addition.

It should also be noted that the use of the term big data by auditors—both on the academic and practitioner side—is not always consistent with its definition in the wider data analytics community. In particular, the long reliance of auditors on sampling means that the extension to the analysis of the whole population of accounting transactions may well be considered "big data" in an auditing environment. But, as we show below—despite the AAA's assertion that "Accounting is Big Data—even the entirety of a business's accounting transactions is several orders of magnitude smaller than the data sets that proponents of big data outside auditing have in mind. When discussing big data in an audit context, there is also the need to differentiate between more of the <u>same kind of data</u> that auditors are already using, or more data <u>of a different kind</u> than what auditors have traditionally relied on to give an audit opinion. The former approach would lead, for example, to continuous auditing where the scope of data is not necessarily expanded but measurements are taken more frequently in time (Kogan, Alles, Vasarhelyi and Wu, 2014). By contrast, big data pushes the domain of data far outwards from financial data to non-financial data—from structured to unstructured data and from inside the organization to outside it—to an extent that may well be outside the comfort zone and technical capability of the audit profession as it is currently constituted.[9]

1.3 Costs and Benefits of Using Big Data in Auditing

Just because the benefits of a technology outweigh the costs for a client's organization does not automatically mean that the same technology is cost-beneficial for the auditors, particularly in

---

[9] For example, a senior technical partner at a Big-4 firm known for its statements extolling its commitment to the use of Big Data in auditing, told the authors that in reality the firm focusses almost entirely on structured data and not what he and other consider to be true, unstructured, Big Data,

the heavily regulated and litigious environment in which accounting firms must function. Cost/benefit decisions, by their very nature, are very different for partnerships vs. corporations. For example, consider a situation where the client corporation and their Big 4 auditors are essentially the same size. If someone in the corporation proposes a $25 million technology modernization project, the CEO, CFO, or other executives do not calculate what that's going to personally cost them (What part of that $25 million will come out of my pocket?). At a partnership, partners automatically determine in their head what that project is going to personally cost them based on the prior year's profit distribution formula. In other words, conceptually, each partner does his or her own cost/benefit calculation. Although it's nearly 25-years old, to see an example of this thinking, see the Harvard Business School case study called "KPMG Peat Marwick: The Shadow Partner," (Eccles and Gladstone 1991).[10]

As we discuss in this paper, using a client's Big Data as regular part of that client's financial audits would be a paradigm shift, requiring unprecedented access to proprietary and sensitive client data that are outside tradition data requested during an audit; a significantly higher reliance on non-financial data (NFD), which auditors have been reluctant to use in the past because it's not clear how validate this non-GAAP data; increased technical skills on the part of the audit team; and increased business acumen to determine what data to analyze and how to interpret the results.

To reiterate, however, we are not saying Big Data will not be incorporated into financial statement audits. We have already mentioned the investments in Big Data by Deloitte and E&Y, and PwC is undertaking a pilot study of ten audit clients to which they are applying Big Data techniques.[11] PwC audit partner Mary Grace Davenport states that it inconceivable that auditors would not use Big Data when the marketing departments of their clients rely so heavily upon it.[12] However, we have heard equally optimistic presentations regarding the other technologies mentioned earlier. As Moore (2002) illustrates, there are adoption chasms that must be overcome in any technology innovation diffusion in organizations. These chasms represent barriers or inhibitors to the adoption of a technology that can help explain why a technology can suddenly stop growing at accounting firms. There are adoption chasms between these innovators who speak at conferences and the other technology adopters in the accounting firms.

---

[10] Eccles, Robert G., Jr. "KPMG Peat Marwick: The Shadow Partner." Harvard Business School Case 492-002, July 1991. (Revised October 1995.) [Citations for this case vary. Sometimes Gladstone is first author. The latest seems to drop Gladstone.]

[11] Discussion by PwC partners at American Accounting Association annual meeting in Atlanta, August 2014.

[12] Discussion by PwC partners at American Accounting Association annual meeting in Atlanta, August 2014.

1.4   Outline of the Paper

The objectives of this paper are to: (1) provide a balanced discussion of both the pros and the cons regarding incorporating Big Data into financial statement audits; and (2) present a research agenda to identify specific aspects of Big Data that could benefit auditors. In other words, the incorporation of Big Data into audits is not an all or nothing decision—which can be overwhelming and discouraging—but an incremental incorporation of Big Data will provide "cherry-picking" opportunities that will in turn provide more confidence to move deeper into Big Data.

We begin our analysis by defining the term Big Data, which turns out to be a lot harder than one might imagine, thus illustrating the emerging nature of this field. With our definition as a guide, we survey the Big Data literature in order to understand both what the pros of Big Data as a methodology applied to auditing practice and its cons that arise when comparing its unique characteristics with those of the needs of auditors. While we do examine the accounting research on Big Data much of our analysis is based on the literature written by Big Data practitioners in the new media of blogs, white papers and other non-academic outlets. Our choice in this regard is driven by the fact that right now it is in practice and not [accounting] academia where the cutting edge discussion is taking place in this still new area of Big Data.

The remainder of this paper is organized as follows. Section 2 defines Big Data and places it into the context of the auditing environment. Section 3 identifies how and where Big Data *could* enhance financial audits in a normative sense—in short, the pros of Big Data in Auditing. Then in Section 4 we identify the unique characteristics of what Big Data analysis requires as inputs and what it produces as outputs and compares that with the unique needs and constraints of auditing. As we show, this is where significant cons arise that will require auditors to proceed very carefully and systematically if they are not to make some of the mistakes that have already arisen in other Big Data applications. That naturally leads in Section 5 to a proposed research agenda to identify how to close the gap between traditional audit procedures and the appropriate use of Big Data. Section 6 has concluding comments.

2. An Overview of Big Data

2.1   Defining Big Data

The first issue researchers' face in exploring Big Data in financial audits is that "Big Data" lacks a consistent definition. One website lists 32 definitions and another website had seven more.[13] McKinsey (2011) states that: *"Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."* This is similar to the Wikipedia definition, which presumably represents a consensus perspective: *"Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."[14]*

Although not mutually exclusive, some definitions focus on the dimensions or characteristics of Big Data and other definitions focus more on examples of the contents of Big Data. Frequently, discussions of Big Data define Big Data in terms of volume, velocity, and variety (commonly referred as the "3 Vs") that are generally traced back to a META Group (now Gartner) 2001 blog (Laney 2001), which actually pre-dates the common use of the term Big Data. Volume refers the overall amount of data included in a Big Data database. Velocity is how frequently the data is changing. Many Big Data installations are collecting real-time sensor data that is being continuously updated. Variety is the broad scope of data that organizations are collecting. Over time, other writers and speakers have added additional Vs such as veracity, which relates to the integrity of the data. As we will discuss later, this last V, veracity, may be particularly problematic to auditors—i.e., how does the auditor develop an appropriate level of confidence in a client's Big Data with massive amounts of NFD?

More importantly to auditors, the *data* in Big Data, refers to collections of multiple types of data, which could include some mix of traditional structured financial and non-financial data (NFD), logistics data, sensor data, emails, telephone calls, social media data, blogs, as well as other internal

---

[13] http://www.opentracker.net/article/definitions-big-data; http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html respectively. Accessed 9/3/2015 9:53:42 AM.

[14] http://en.wikipedia.org/wiki/Big_data. Academic researchers will appreciate the amorphous nature of defining a methodology by reference to "typical database software tools" considering that according to such a definition, the commonplace use by empirical capital market researchers today of SAS to analyze COMPUSTAT data would be considered an example of Big Data by the standards of Ball and Brown (1968). This particular definition has the defect of encompassing almost any kind of data analysis more complex than basic regression and any data beyond the general ledger.

and external data. Auditors' traditional focus on transactional data, hence, a particularly relevant *content* definition of Big Data in the auditing context is that by Connolly (2012), which takes transactions as its starting point:[15]

Big Data = Transactions + Interactions + Observations

He goes on to explain and illustrate this equation:

"ERP, SCM, CRM, and transactional Web applications are classic examples of systems processing Transactions. Highly structured data in these systems is typically stored in SQL databases.

Interactions are about how people and things interact with each other or with your business. Web Logs, User Click Streams, Social Interactions & Feeds, and User-Generated Content are classic places to find Interaction data.

Observational data tends to come from the "Internet of Things". Sensors for heat, motion, pressure and RFID and GPS chips within such things as mobile devices, ATM machines, and even aircraft engines provide just some examples of "things" that output Observation data."

---

[15] http://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/ An alternative characterization of these three types of data are Process-Mediated Data, Human-Sourced Information, and Machine-Generated Data.

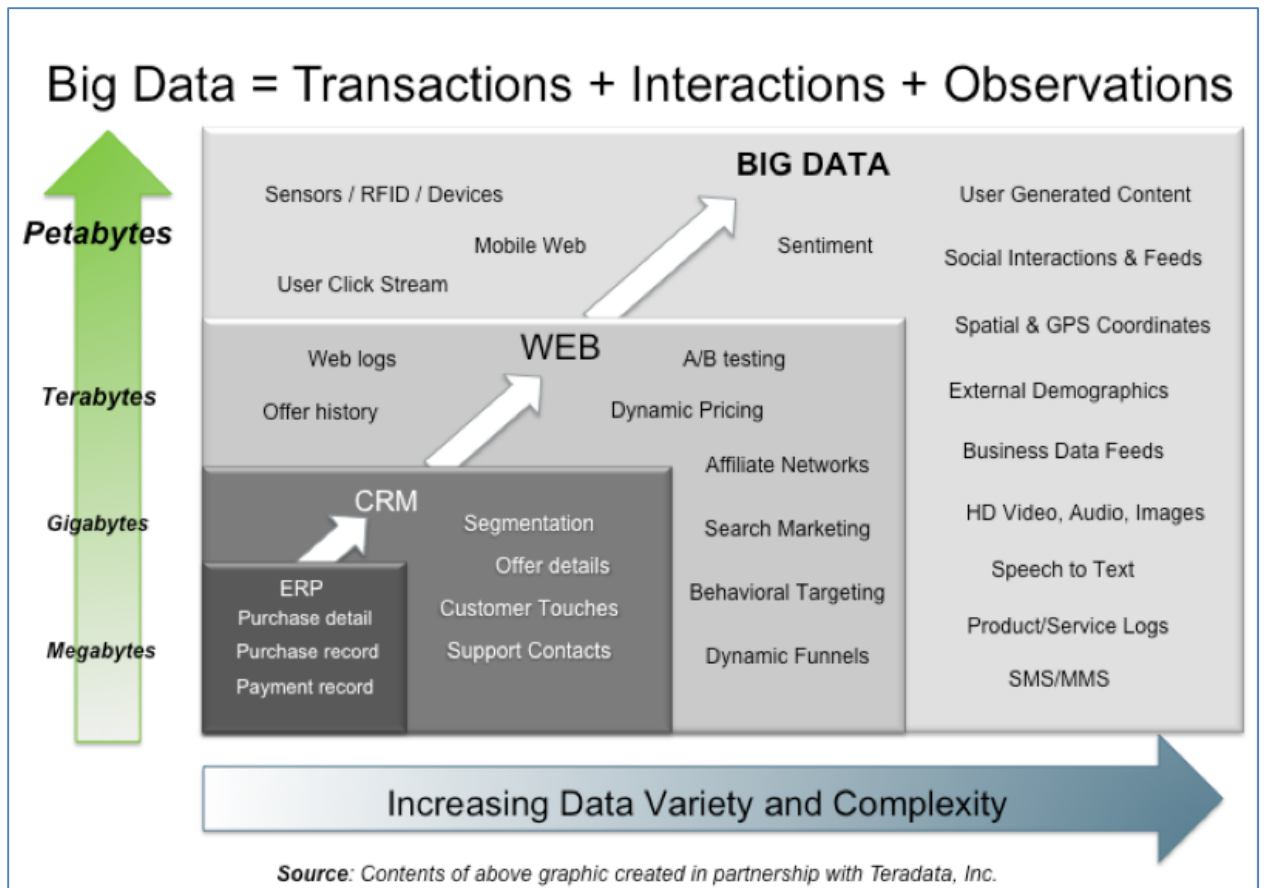**Big Data = Transactions + Interactions + Observations**

Figure 1: Connolly's (2012) definition of Big Data

Connolly's (2012) framework (see Figure 1) is useful because it puts the data currently used by auditors (lower-left corner) into perspective and shows how much additional data that Big Data offers to expand that input into the auditing process. Connolly (2012) goes on to identify what he sees as seven drivers of Big Data in business:

Business

1.   *Opportunity to enable innovative new business models*

2.   *Potential for new insights that drive competitive advantage*

Technical

1.   *Data collected and stored continues to grow exponentially*

2.   *Data is increasingly everywhere and in many formats*

3.   *Traditional solutions are failing under new requirements*

Financial

1.   *Cost of data systems, as a percentage of IT spend, continues to grow*

These drivers are general and are not specific to auditing, however, if Big Data is to impact auditing more directly, then it will be through the first two drivers. How precisely will Big Data do all this? Lucas (2012) has an insightful characterization of Big Data in which he argues that it: *"divides the world by intent and timing rather than the type of data... The 'old world' is about transactions, and by the time these transactions are recorded, it's too late to do anything about them: companies are constantly 'managing out of the rear-view mirror'. In the 'new world,' companies can instead use new 'signal' data to anticipate what's going to happen, and intervene to improve the situation. Examples include tracking brand sentiment on social media (if your 'likes' fall off a cliff, your sales will surely follow) and predictive maintenance (complex algorithms determine when you need to replace an aircraft part, before the plane gets expensively stuck on the runway)."[16]*

The claim of "managing out of the rear-view mirror" is one oft made against accounting information in general, and auditors in particular. As Lucas (2012, emphasis in original) goes on to say: *"**The signals of health in a business we have been trained to look for haven't changed for a century or more, yet the amount of information available today that could indicate the relative health of a business is radically different.** It's sad to say, but Ebenezer Scrooge would be as adept in a board meeting today as the day that Charles Dickens conjured him up in 1843!"*

It is worth quoting at length from Lucas (2012) because he is a Big Data definer who makes his arguments in explicit accounting term. Hence, he makes his case that "every [accounting] firm needs a real time data platform" by stating (emphasis in original):

*"The Income Statement and Balance Sheet are at best rear-view measures of the top line and bottom line. They provide a snapshot in time of all that has **happened**, but very little, if any, indication of what is **happening** in the enterprise. For example, an online retailer with a subscription model experiences a massive drop in stock price because of poor Income Statement results. Further analysis indicates that there was a dramatic drop in subscribers (Churn) in one of their most profitable segments. It would stand to reason that a pre-emptive pulse-check on the churn could have helped stem the bleeding and perhaps prevented the reaction on Wall Street. This churn is an example of a new signal that could have helped this one enterprise run its business more proactively rather than look for explanations with a rear-view perspective. So how does a company gain the ability to know midstream that it has a problem or there is an opportunity going unaddressed? How will a company gain the intelligent actionable insights to*

---

[16]   http://www.saphana.com/community/blogs/blog/2012/08/21/beyond-the-balance-sheet-run-your-business-on-new-signals-in-the-age-of-big-data.

make this happen? What is needed is the ability to ask complex questions going across the volume and variety of data in an interactive manner, and most importantly in real-time."

But as convincing as this argument may seem to be about the long criticized shortcomings of traditional accounting statements, it must be kept in mind that Lucas is talking about accounting and not about auditing. Thus, while marketing managers in a business should certainly pay attention to the churn rate of subscribers, should auditors? Just because Big Data would enable the mangers of an "*enterprise [to] run its business more proactively rather than look for explanations with a rear-view perspective*" does not necessarily imply that auditors also have to adopt such a "proactive" perspective, and they are unlikely to want to do so in the absence of a rigorous analysis of what Big Data brings to auditing.

2.2   The Context for Big Data

Defining Big Data is an incomplete exercise unless that definition is put into context. The emergence of Big Data does not mean that companies are necessarily creating new data elements (although they may be doing that independent of (or loosely coupled to) their Big Data activities). Big data is assembling many different datasets together in one data warehouse to allow analysis of dissimilar data to, in turn, discover new patterns, relationships, and correlations in the data. While it's possible to do such analysis across separate isolated databases using techniques such as open database connectivity (ODBC), this approach would be extremely inefficient across separate massive databases. Besides bringing dissimilar datasets together in one data environment, the incoming data may go through various mappings, transformations, and cleanings to make subsequent data analytics more efficient. Hence, while Big Data certainly has its own unique characteristics it can, nonetheless, be also framed as an evolutionary phenomenon in data storage and data analysis and that evolution can take place at different rates in different companies.

The development of enterprise resource planning (ERP) systems is a major step in that evolutionary process. In the pre-ERP environments, companies were characterized as having numerous islands of technology (silos) with isolated databases specific to isolated applications that do not share data with each other. These pre-ERP databases were a mix of internally developed and proprietary acquired databases, mixed standards for naming and creating data elements (fields) and tables, and a mix of database structures. ERP software integrated the content of many (but not all) of these isolated databases into one large integrated database shared by several applications. As storage costs continued to significantly decreased, massive data warehouses beyond ERP databases

became practical. To deal with these warehouses and new types of data, data mining algorithms and techniques became more sophisticated and efficient. The convergence of these and other technological advancements naturally evolved into what we now call Big Data.

Whatever the characteristics of Big Data are, that is not what defines its value to auditors or any other users. The data→information→knowledge→wisdom hierarchy literature (e.g., see Rowley 2007 and Zins 2007; or in an accounting context, Elliott 1981) can be drawn on to make the point that Big Data is indeed just data. As Zikopoulos, deRoos, Parasuraman, Deutsch, Corrigan, and Giles (2013) say, *"…the term Big Data is a bit of a misnomer"* because the real value of (the real interest in) Big Data is the value of the analytics that can be performed with that data. In other words, Big Data—whether considered as an evolutionary or revolutionary development in technology— remains a means towards an end and not, as the hype sometimes expresses it, as an end in itself. If auditors are to find value in it, Big Data and related analytics must lead to better audit outcomes.

2.3   Extraordinary claims require extraordinary evidence

In this regard it is worth recalling the old adage that "extraordinary claims require extraordinary evidence". The context for Big Data to be of value in auditing is that the converse also holds: **extraordinary use of data requires extraordinary questions that need answering**. What has yet to be articulated is whether auditing contains such extraordinary questions that cannot be answered by fuller use of already available transactional data. Only when the limits of traditional data have been reached are auditors likely to turn to other types of Big Data. Moreover, as Deutsch et al (2013) point out above, what is special about Big Data is the relatively easier ability today to analyze large data sets, not the availability of those data sets. For example, in Figure 1, one component of Connolly's conception of Big Data is RFID chips, but these have been in widespread use for at least a decade. Hence, auditors need not have waited for the emergence of Big Data as a concept to have started to use it in their audit procedures, but there is little evidence that they have chosen to do so.

To gain some perspective on the likelihood of the adoption of the client's Big Data by auditors it is worth noting that auditors were in a position to be pioneers in the analysis of large data sets because they had their own large data sets that includes data collected from hundreds or thousands of clients within and across industries, regions and nations. As Hoitash, Kogan and Vasarhelyi (2006) demonstrated using such peer information can significantly improve audit analytics. Hoitash et al's (2006) work builds on the insights of Foster (1981) more than three decades

ago that information from one firm could be useful for predicting returns in another, and yet, there is very little evidence that audit firms have ever partaken of this internal opportunity.

In other words, it is not the case that in auditing "if you build it they will come". Rather, as Alles, Kogan and Vasarhelyi (2002) stated with regard to continuous auditing: *"Most attention has been paid to the technological aspects of CA, taking for granted the economic architecture that underlies all assurance. We feel that equal weight needs to be placed on the demand for continuous assurance products and the economic infrastructure necessary to produce and pay for CA."* Given the less than whole hearted adoption of continuous auditing in the decade since, a similar caution is surely warranted with regard to Big Data. Alles, Kogan and Vasarhelyi (2008) indicate that ultimately it was the need to implement Section 404 of the Sarbanes Oxley Act that drove demand for continuous auditing techniques and not the needs of the traditional financial statement audit.

3. Pros of Incorporating Big Data into Audit Practice

How Big Data adds value is by finding patterns that are not discernible in smaller data sets with fewer variables to drive unexpected correlations, such as, for example, the kind of sampling that still characterizes much of auditing practice. As Cukier and Mayer-Schoenberger (2013) state: "*It is tempting to understand Big Data solely in terms of size. But that would be misleading. Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before; call it "datafication." For example, location has been datafied, first with the invention of longitude and latitude, and more recently with GPS satellite systems. Words are treated as data when computers mine centuries' worth of books. Even friendships and "likes" are datafied, via Facebook."*

Zikopoulos et al. (2013) provide the following example of structured Facebook data in the JavaScript Object Notation (JSON) format showing the large amount of business related information that is obtainable from even a small example of Big Data:

{"data":[
{"id":"53423432999_23423423_19898799",
"from":{"name":"PaulZikopoulos","id":"Z12"},
"message": "Thinking of surprising my wife with a quality time gift that lets her know she's special, any ideas? I thought about taking her to the driving range, perhaps play a round and caddie my game.", "created_time":
"2012-08-02T21:27:44+0000","likes:5,"

"comments":{"data":[{"id" 2847923942_723423423", "from":{"name":
"MaryAnneInfanti","id":"948574763"}.
"message": "Paul! Purses and gold! Costco's got a great Kate Spade purse
on sale this week that says I love you without having to lift a pen.
If you go with your idea, the only thing driving will be you: alone!"• "created_time 2012-00-
02T11:27:44+0000", "likes:64 }   }

It is this kind of detailed information, not just about preferences for Kate Spate purses, but also contextual information that the husband plays golf while the family has a Costco membership that leads people to say that there is "gold in that data".[17]

Most of the data in Big Data precedes accounting transactions from a few days to one or more years. Companies like Boeing and Airbus have orders for aircraft that will be delivered many years in the future. As such, Big Data contents are powerful predictors to set (and adjust) auditor expectations at the beginning (planning phase) and throughout the audit. Because much of Big Data contents are separated (both physically and conceptually) from accounting data, fraudsters are not likely to be able to manipulate all applicable Big Data contents to disguise their fraud. Because of the substantial contents of Big Data and the concept of "100% sampling" (using the whole population), the discovery of "red flags" and other suspicious outliners are greatly increased. There are many other factors involved, but Big Data analytics—particularly with its emphasis on NFD—clearly had the *potential* to detect historical examples of massive frauds.

Pros	Comments
Strong predictive power, which is a powerful tool for setting expectations for financial statement auditor.	Events/transactions included in Big Data can predate accounting transactions by days, weeks, months, and even years. For example, knowing someone is pregnant (by analyzing sales of pregnancy test kits) enables prediction of a permanent change in purchase patterns.
Rich data sources to identify potential fraudulent activities.	Difficult for fraudster to change all upstream non-financial transactions to cover up financial statement fraud. For example, trade based money laundering can be detected by comparing invoices with the actual weight of shipping containers.
Analyzing all data increases probability of discovering red flags, "smoking guns," and suspicious outliers.	Since fraud represents a very small percentage of transactions and could be easily not included in the small samples auditors traditionally select.

Table 1. Pros of Incorporating Big Data into Audits.

---

[17] http://data-informed.com/understand-your-business-with-behavioral-data/. Accessed 9/2/2015 5:20:44 PM.

However, even with these significant potential benefits of Big Data in financial statement audits, the audit practice community (firms, regulators, and other stakeholders) will have to go through a paradigm shift to exploit those benefits. Table 2 presents the inhibitors that will have to be addressed before these pros of Big Data can be realized in audit practice. In broad terms, auditors will need unprecedented access to client data (which raises concerns for both the audit firms and their clients); increase the audit team's data analytical skills (which can have both short and long-term cost consequences); and increase the understanding of the relationships between NFD and financial data (which means expanding the business acumen of audit team members).

In the third column of Table 2, we present suggestions of how the inhibitors could be addressed to reduce the impact of each inhibitor. The suggested ways to address the inhibitors fall into two broad themes. First, using Big Data is not an all or nothing choice. Auditors can start small in terms of both breadth and depth. Authors can develop their analytical skill by first using accounting-adjacent data (see Figure 3) such as data included in order-entry and purchase order applications and then move out to incorporating data that is further and further removed from accounting data. The value of these suggestions can be explored in future research.

The second theme is the rather pragmatic theme of Big Data will eventually be a normal part of audits (because the changing client IT infrastructure environments) and it is better to be proactive than to be reactive in terms of incorporated Big Data into audits, as Alles (2015) also argues. What is presented in Table 2 are only the usual stumbling blocks that have to be overcome when implementing any new technology. The unique issues that Big Data will pose for its incorporation into auditing practice, its true "cons", are what we discuss next.

Table 2. Inhibitors associated with incorporating Big Data into audits

Inhibitors	Comments	Amelioration
Auditors will need direct access to client's Big Data	• Accounting firms' legal department discourage direct access because of fears of corrupting data and/or negatively impacting performance of the client's system by writing poorly-designed queries. • Client concerned about auditors having unlimited access to propriety data.	Start small. The incorporation of Big Data in financial statement can be an evolutionary (not revolutionary) process. Part of the pre-audit will be to identify NFD that could be valuable based on the pre-audit risk analysis and brainstorming session.
Could "miss" smoking gun in 100% population.	Accounting firms' legal department discourage 100% "samples" because the firm has no defense as to why they missed the smoking gun. Current defense: smoking gun was not in the small sample analyzed during the audit.	Not increasing the population is going to be increasingly harder to defend as Big Data, user-friendly analytical software, and inexpensive computer resources become more available. Firms are going to be driven toward improving their analytical skills.
Requesting specific data send signals to client regarding areas of concern to the auditors.		Pre-requesting board varieties of data at the start of the audit will reduce signaling.
NFD is messy	Could require a significant amount of time to validate the data to determine level of messiness (particularly in terms of accuracy and timing) and usability/applicability to the audit.	The Big 4, in particular, have numerous clients to which they provide internal audit services. Because internal audit performs broader activities and have freer access to internal Big Data. Accounting firms can use these opportunities to hone their skills to deal with messy NFD.
Results can be messy (e.g., numerous false positives		Improving queries (in terms of variables included/excluded in the queries) can greatly reduce false positives. False

4.   Cons of Incorporating Big Data into Audit Practice

4.1   Nature of Audit Decision Making

If Big Data is to establish a place in auditing, it also has to present itself as a superior solution to a problem that they have already addressed. The major constraint that external auditors face in this regard is that their ultimate decision variable is <u>discrete</u> rather than <u>continuous</u>: an unqualified audit opinion or not, with only a small minority of clients receiving either a qualified or adverse opinion. What value does access to more data provide in this context?

For the sake of argument, assume that the auditor's decision whether to issue an unqualified opinion depends on the probability that the client's financial statements are in accordance with GAAP. If the threshold for doing so is, say, 95%, and the auditor's conventional audit practices yield an estimate for a particular client of at least that level, then why should the auditor now turn to Big Data to refine that estimate even further unless the Big Data analysis is likely to actually more effectively and/or efficiently achieve that threshold? As Keltanen (2013) states:[18] *"Diminishing returns still apply. Statements like "data is the new oil" make it sound like data is currency, when it's actually an investment. In all statistical measurements, once enough data points have been collected to establish a result, adding more data points begins to create less accuracy. This should be a pressing concern when you're investing increasing amounts of money, time and resources into capturing and analysing data."*

Of course, the counterargument is that if standard audit techniques yield an estimate of conformance with GAAP below the threshold, then turning to Big Data may prove more valuable. For that matter, Big Data may show that estimates based on standard audit evidence are flawed. Clearly research is needed on these suppositions. On the other hand, auditors obviously already provide clean opinions on the vast majority of their clients so revealed preference does not support the hypothesis that there is an unmet demand for Big Data type analytics. Moreover, the argument for the value added of Big Data in auditing ignores the cost of implementing it, which—as discussed in the prior section—may well be significant.

Moreover, while costless estimates based on more data are always likely to lead to better decision outcomes, the benefits are likely to be less with discrete decision problems than with continuous ones—and the examples often used to promote Big Data are definitely continuous. For example, using social media data to refine marketing campaigns or visualization techniques to

---

[18] http://www.theguardian.com/media-network/media-network-blog/2013/apr/16/big-data-lean-strategy-business.

identify which set of patients are likely to most benefit from a drug regime will yield superior outcomes if the Big Data results in a 1% change in probabilities, something which is not always the case in auditing.

Of course not all decisions in auditing are dichotomous, as the examples from the standards above indicate. Hence there clearly is scope for auditing to play a role in refining fraud detection, for instance, or developing better analytical procedures. But these are means towards the end of an audit opinion and not ends in themselves, as auditors point out when they say that if fraud detection is the main objective then they should be hired to do a forensic audit rather than a regular one. As Noreen (2011) put it: *"To the auditing industry, the fact that investors tend to blame auditors when frauds go undetected reflects unrealistic expectations, not bad work by the auditors. The rules say auditors are supposed to have a "healthy degree of skepticism," but not to detect all frauds. "There is a significant expectations gap between what various stakeholders believe auditors do or should do in detecting fraud, and what audit networks are actually capable of doing, at the prices that companies or investors are willing to pay for audits," stated a position paper issued in 2006 by the chief executives of the six largest audit networks.*

Note that last part. They suggested that if investors were really worried about fraud, they should consider paying more for a "forensic audit" that would have a better — but not guaranteed — chance of spotting fraud. Don't like our work? Pay us more."[19]

There is scope for accounting firms to sell Big Data services as a consulting activity, but that is not auditing related. Looking at Figure 1, we can see that data is hierarchically in order of magnitude from ERP data to CRM information, data drawn from the web and what Connolly (2012) labels true Big Data, which are observations of behavior and processes. Without exception, all the work in the current focus of applying technology to auditing, including continuous auditing, has been based on the lowest level of this hierarchy, ERP data (Brown, Wong and Baldwin, 2007). Given the limited success in the adoption of CA, what are the prospects that auditors will turn to analyzing non-ERP based Big Data?

4.2   Discrete vs. Open-Ended Decisions

One of the factors in this regard is that the type of decisions that auditors are concerned with are constrained by accounting rule and regulations and are not *open-ended* as the optimization of marketing strategies or the causes of heart disease might be. Such problems are deemed "open-

---

[19] http://www.nytimes.com/2011/06/10/business/10norris.html.

ended" in as much it is hard to know what variables will prove to be correlated with the dependent variable. By contrast, most auditing problems arise in a much more ordered environment, with less independent variables likely to impact upon the problem. That raises the question of how useful non-transactional data or data from outside the company will be in establishing whether, for instance, segregation of duty controls are being violated or financial statements are in accordance with GAAP, though this is ultimately an issue that requires empirical investigation., The most obvious use of Big Data may be in fraud detection (EY 2014). For example, in Brazil it is commonplace for employers to require that employees use a given bank in order that the employer can then check that account for evidence of corrupt practices, such as expenditures in excess of salary. Apart from the fact that this practice would never be allowed in most Western nations due to privacy concerns, this is still a well-defined problem where it is reasonable to imagine modeling all the variables whose examination will determine the outcome. But this is virtually antithetical to the mindset of Big Data as we shall discuss below.

If there is an open-ended type question in auditing, it is the going concern opinion. Especially given the criticism auditors faced for not issuing going concern opinions for businesses that failed in the recent financial crisis such as Bear Stearns, improving auditor assessment of firm risk may be a promising application of Big Data. On the other hand, it has to be kept in mind that this represents only a small part of the audit engagement and that the number of going concern opinions issued has declined lately to a small minority of all audit clients examined.[20] Nonetheless, this may be the most immediately promising area for research into incorporating Big Data techniques into audit practice.

4.3   Implications of "Datafication": Correlation over Causation

As summarized by Gray and Debreceny (2014), traditional data mining approaches can be classified into two broad categories: directed (also top-down or hypotheses testing) and undirected (also bottom-up or hypotheses generating). With directed data mining, specific relationships with preselected variables (e.g., how has the relationship between revenue and cost of goods changed from year to year?). With undirected data mining, specific relations are not being tested; instead the auditor wants the software to find relationships between populations of variables—and, particularly, to identify outliers to those newly discovered relationships. It was up to the data miner (i.e., the auditor) to develop a hypothesis as to why there is a relationship between to variables and what does

---

[20] http://www.complianceweek.com/going-concern-warnings-drop-to-12-year-low/article/273186/.

the outliers mean in that context. The phenomenon of "datafication" subsequently drives the way in which Big Data is analyzed, with pattern recognition or (undirected) hypothesis generating taking precedence over (directed) hypothesis testing. As Cukier and Mayer-Schoenberger (2013) go on to say:

"Using great volumes of information in this way requires three profound changes in how we approach data. The first is to collect and use a lot of data rather than settle for small amounts or samples, as statisticians have done for well over a century. The second is to shed our preference for highly curated and pristine data and instead accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. Third, in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations. With Big Data, instead of trying to understand precisely why an engine breaks down or why a drug's side effect disappears, researchers can instead collect and analyze massive quantities of information about such events and everything that is associated with them, looking for patterns that might help predict future occurrences. Big data helps answer what, not why, and often that's good enough."

These three profound changes listed by Cukier and Mayer-Schoenberger (2013)—collecting more data, messy data, and correlation over causation—pose increasing levels of difficulty for acceptance by auditors and we examine each in turn.

I.   More Data

While storing large amounts of data is no longer very costly, that only takes into account the direct cost of data storage. Another factor that would particularly concern auditors would be maintaining security of ever larger amounts of data and ensuring that there are no breaches of privacy, most especially concerning information drawn from social networks and other interactions that identify individuals outside the auditee.

II.   Messy Data

While the more-data change may be addressed by rigorous data protection controls, the next change that Cukier and Mayer-Schoenberger (2013) postulate is necessary to succeed at Big Data poses a more challenging hurdle for auditors: *"In the past, when people collected only a little data, they often had to decide at the outset what to collect and how it would be used. Today, when we gather all the data, we do not need to know beforehand what we plan to use it for."* While this approach may make sense in the case of

scientists or internet companies, it is hard to imagine that auditors would feel inclined to request/collect data without a clear reason to do so. Doing so also goes counter to the desire of auditors to be conservative and to be able to clearly justify any decisions that they make. More to the point, if the data is to come from the client, then obtaining information at all, let alone with no explanation provided is difficult to attain. Indeed, the development of the Audit Data Standard (AICPA 2013) indicates that auditors have difficulty obtaining even transactional data: "*One reason data standards are needed is to address the ongoing challenge that management and internal and external auditors face in the efficient exchange of a company's data. This process is complicated by the fact that accounting and IT personnel approach requests for such information from different perspectives. For example, in some cases, audit-related data requests are forwarded directly to a company's IT department, with limited further involvement from the accounting or finance department. In many cases, the burden is on the auditors to acquire the data.*"

The problems that auditors would have with this approach towards non-curated data gathering can be gauged by what Cukier and Mayer-Schoenberger (2013, emphasis added) say next: "*When we increase the scale by orders of magnitude,* **we might have to give up on clean, carefully curated data and tolerate some messiness***. This idea runs counter to how people have tried to work with data for centuries. Yet the obsession with accuracy and precision is in some ways an artifact of an information constrained environment. When there was not that much data around, researchers had to make sure that the figures they bothered to collect were as exact as possible. Tapping vastly more data means that* **we can now allow some inaccuracies to slip in** *(provided the data set is not completely incorrect), in return for benefiting from the insights that a massive body of data provides.*"

While researchers or managers in search of new marketing strategies may tolerate such ambiguity it is probable that auditors fall into that category of people for whom it runs counter to how they have worked with data for many years. Allowing "some inaccuracies to slip in" is difficult to reconcile with the focus in auditing on data integrity.[21]

AICPA (2014) recognizes these inherent difficulties that Big Data (or "predictive analytics" as they call it) will pose for CPAs:

"*This all sounds easy enough. But, in practice, it is difficult to build and validate a model. Some reasons for this difficulty include:*

---

[21] For example, the Audit Data Standard includes a questionnaire designed to "help auditors better understand the data and assess its completeness and integrity."
http://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/pages/auditdatastandardworkinggroup.aspx

1. It often is difficult to understand causality. *Does a fast car cause a driver to exceed the speed limit?*

2. Data quality issues affect model results. *One source rounded dollars up to millions but the other used standard rounding.*

3. It often is difficult to obtain the data that you want. *You did not collect historical data from one of the locations in your model.*

4. The real world is the lab — you can't always be flexible with experiment design. *Would there have been fewer audit findings if you had used a different sampling system?*

5. Independent variables affect each other. *The type of customer who is complaining behaves differently during price increases.*

6. The relationships between variables often are not linear. *How much would a client care about a price change of 1%, 5% or 500%?*

7. Relationships are not necessarily stable over time. *Who is watching the same advertisement at 2pm and 6pm?"*

III.   From Causation to Correlation

The emphasis on quantity rather than quality in Big Data leads to the third major change identified by Cukier and Mayer-Schoenberger (2013, emphasis added): *"These two shifts in how we think about data—from some to all and from clean to messy—give rise to a third change:* **from causation to correlation***. This represents a move away from always trying to understand the deeper reasons behind how the world works to* **simply learning about an association among phenomena and using that to get things done***."*

This idea that data is sufficient in itself is a persistent theme in the Big Data literature. For example, Anderson (2008) in an article with the telling title "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" argues that *"Correlation is enough. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."* The approach of using correlation from Big Data to draw inferences has had considerable success. For example, Ginsberg, Mohebbil, Patel, Brammer, Smolinski and Brilliant (2009) reported on the use of Google searches for flu related terms to plot incidences of influenza. As the authors state: *"Because*

the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day." The methods of this study have now been made publicly available through the Google Flu Trends site.[22]

Whatever the successes of this approach, however, it is certainly startling to researchers who had the adage "correlation is not causation" drummed into them as doctoral students. And auditors too would be skeptical about basing conclusions on correlations that cannot be proven to be based on any kind of causality—for example, if the decisions based on those Big Data results were ever to be subject of subsequent litigation.[23]

The difference between managers using Big Data to create a new marketing strategies or doctors tracking disease is that while they are using correlation to drive future action, auditors use information to create controls. Thus, if Amazon finds a correlation between college-educated women under 30 who do yoga, purchase Lululemon clothing and eat organic food, then they will tailor a marketing strategy accordingly. Of course, it is only a hypothesis and the strategy really an experiment, but if the correlation proves valid then more sales are achieved, while if it turns out to be spurious, then Amazon would move on to the next experiment.

By contrast, if auditors using Big Data find a correlation between newly divorced employees and overbilling of expenses what precisely are they supposed to do with that information? Recommend that divorcing employees be subject to great scrutiny or be fired? It is one thing for auditors to use correlation information to suggest new areas to investigate further, but quite another to institute formal controls based on that information. Verifiability and credibility are vital in auditing if action against particular employees is to be initiated, or even when broad controls are instituted because, unlike with a marketing strategy, actions by auditors may have to be defended in litigation. Of course, this is not true for all audit findings and it is not to say that Big Data has no role to play in auditing. But the "correlation not causation" mantra of Big Data poses a bigger hurdle in auditing than it would in the management side of the business.

4.4   The Role of Theory in Interpreting Big Data Results

---

[22] http://www.google.org/flutrends/us/#US.

[23] As one practitioner referee of an earlier version of this paper stated in this regard: *"Definitely true, how can you justify a conclusion based on a correlation that may or may not exist in the data. The judges and lawyers could have a difficult time with this. Also if all correlations had to be proven 100%, the efficiencies could easily be lost."*

It is worth noting that there is a backlash against the correlation over causation approach in the Big Data literature itself. Keltanen (2013) writes: *"…the lightest, simplest way to achieve your data analysis goals is the best one…The dirty secret of Big Data is that no algorithm can tell you what's significant, or what it means. Data then becomes another problem for you to solve. A lean data approach suggests starting with questions relevant to your business and finding ways to answer them through data, rather than sifting through countless data sets. Furthermore, purely algorithmic extraction of rules from data is prone to creating spurious connections, such as false correlations… today's Big Data hype seems more concerned with indiscriminate hoarding than helping businesses make the right decisions."*

David Brooks (2013) writing in the New York Times first describes the correlation focused case for Big Data and then discusses what Big Data can and cannot do:[24] *"The theory of Big Data is to have no theory, at least about human nature. You just gather huge amounts of information, observe the patterns and estimate probabilities about how people will act in the future… As Viktor Mayer-Schönberger and Kenneth Cukier write in their book, "Big Data,"[25] this movement asks us to move from causation to correlation. People using Big Data are not like novelists, ministers, psychologists, memoirists or gossips, coming up with intuitive narratives to explain the causal chains of why things are happening. "Contrary to conventional wisdom, such human intuiting of causality does not deepen our understanding of the world," they write. Instead, they aim to stand back nonjudgmentally and observe linkages: "Correlations are powerful not only because they offer insights, but also because the insights they offer are relatively clear. These insights often get obscured when we bring causality back into the picture… One limit [of Big Data] is that correlations are actually not all that clear. A zillion things can correlate with each other, depending on how you structure the data and what you compare. To discern meaningful correlations from meaningless ones, you often have to rely on some causal hypothesis about what is leading to what. You wind up back in the land of human theorizing."*

A debate that auditors need to pay attention to in this regard is that taking place in the Big Data community about the role of theory in interpreting Big Data results. Crawford (2013) writes:[26] *"Can numbers actually speak for themselves? Sadly, they can't. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves. For example, consider the Twitter data generated by Hurricane Sandy, more than 20 million tweets between October 27 and November 1. A fascinating study combining*

---

[24] http://www.nytimes.com/2013/04/16/opinion/brooks-what-youll-do-next.html?hp.
[25] Mayer-Schoenberger and Cukier (2013).
[26] http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/.

Sandy-related Twitter and Foursquare data produced some expected findings (grocery shopping peaks the night before the storm) and some surprising ones (nightlife picked up the day after — presumably when cabin fever strikes). But these data don't represent the whole picture. The greatest number of tweets about Sandy came from Manhattan. This makes sense given the city's high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster."

The economist Susan Athey in an interview with The Region journal said the following (Clement, 2013):[27] "*Region: "What are your thoughts about Big Data? Does it portend, as some have suggested, an "end to theory"?*

Athey: Absolutely not an end to theory. In fact, the need for theory is in some ways magnified by having large amounts of data. When you have a small amount of data, you can just look at the data and build your intuition from it. When you have very large amounts of data, just taking an average can cost thousands of dollars of computer time. So you'd better have an idea of what you're doing and why before you go out to take those averages. The importance of theory to create conceptual frameworks to know what to look for has never been larger, I think.

Region: And yet some have argued that because data exist in increasingly large quantities, all you really need is to "see what the data say."

Athey: I think what is true is that when you have large amounts of data, if you ask it the right questions, you have a greater ability to let the data speak, and so you can be much less reliant on assumptions. But you still need a strong conceptual framework to understand what's coming out."

While one may question whether taking an average can cost thousands of dollars of computer time even with very large data sets, the key takeaway from both these citations is that the correlation not causation school of Big Data is not the only perspective and that making use of Big Data results may require a proper understanding of the context from which those results are obtained. As Crawford (2013) argues: *"We get a much richer sense of the world when we ask people the why and the how not just the "how many". This goes beyond merely conducting focus groups to confirm what you already want to see in a Big Data set. It means complementing data sources with rigorous qualitative research. Social science methodologies may make the challenge of understanding Big Data more complex, but they also bring context-awareness to our research to address serious signal problems. Then we can move from the focus on merely "big" data towards something more three-dimensional: data with depth."*

---

[27] http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=5112.

A theory-driven approach may prove to be a better fit in the application of Big Data in auditing. For one thing, developing a theory to explain Big Data results found in one client engagement would facilitate the transfer of that learning to other clients. Moreover, auditors unlike other Big Data users may feel more comfortable specifying to at least some extent what they are looking for (a directed, hypothesis testing approach) rather than leaving it all to the data to speak (an undirected, hypothesis generating approach). Keltanen (2013) argued that it is more efficient to start with relevant questions and find ways to answer them through data, rather than sifting through countless data sets in the hope of finding something useful. But it may also fit better with the conservative culture of auditing, with "conservative" not meant as a pejorative but in due recognition of the greater societal and economic consequences of audit decisions.

Indeed, Crawford (2013) offers a corrective to the previously cited Big Data success story of Google Flu Trends: *"But as we increasingly rely on Big Data's numbers to speak for themselves, we risk misunderstanding the results and in turn misallocating important public resources. This could well have been the case had public health officials relied exclusively on Google Flu Trends, which mistakenly estimated that peak flu levels reached 11% of the US public this flu season, almost double the CDC's estimate of about 6%. While Google will not comment on the reason for the overestimation, it seems likely that it was caused by the extensive media coverage of the flu season, creating a spike in search queries."*

A recent New York Times (2014) article provides auditors with insights both into the potential and the pitfalls of using Big Data.[28] The article points out that Big Data analytics enables the search of millions of patient records for diagnoses and treatments, something which is particularly helpful in the case of rare conditions for which it is difficult to make the economic case for a targeted research program: *"…large, costly and time-consuming clinical trials are rarely carried out for uncommon complications of this sort. In the absence of such focused research, doctors and scientists are increasingly dipping into enormous troves of data that already exist — namely the aggregated medical records of thousands or even millions of patients to uncover patterns that might help steer care."* An important questions that auditors will have to ask is whether the situations they face is more akin to treatment of a rare condition or whether audit engagements share broadly similar issues that can better investigated more systematically than through Big Data techniques. Certainly the article makes the point that Big Data is not the ultimate analytic method that supersedes all others (such as formal clinical trials) but one that adds value in some circumstances and not in others.

---

[28] http://www.nytimes.com/2014/10/05/magazine/can-big-data-tell-us-what-clinical-trials-dont.html?ref=magazine.

This is particularly the case when one considers the drawbacks of Big Data, for the New York Times the article goes on to make some of the caveats that we have discussed above, such as the problems of "less than pristine data": *"In the lab, ensuring that the data-mining conclusions hold water can also be tricky. By definition, a medical-records database contains information only on sick people who sought help, so it is inherently incomplete. Also, they lack the controls of a clinical study and are full of other confounding factors that might trip up unwary researchers. Daniel Rubin, a professor of bioinformatics at Stanford, also warns that there have been no studies of data-driven medicine to determine whether it leads to positive outcomes more often than not. Because historical evidence is of "inferior quality," he says, it has the potential to lead care astray."*

As with the prior section on pros of Big Data, this discussion is not an exhaustive examination of its cons relative to auditing practice, and nor is it meant to suggest that there is no possibility of reconciling the two. Rather, this discussion is meant as a corrective to the presumption that because Big Data is being embraced by businesses implies that it will be facile to apply it to auditing too. As before, more research is needed on this topic, as we discuss next.

5. Research Opportunities in the Application of Big Data to Auditing

The cons of the implementation of Big Data in auditing arise from the contrast between the unique characteristics of each field. That obviously suggests the need for further research on better understanding that conflict and how it can be best overcome. To begin that process, we pose four broad and interrelated questions about the degree and speed at which Big Data will be incorporated into financial statement audits:

1. What are the unique characteristics of Big Data that have the greatest potential to improve the efficiency and effectiveness of financial statement audits?
2. What are the tangible and intangible benefits of incorporating Big Data into financial statement audits in terms of audit quality, efficiency, and deterrent effect?
3. What are the costs (software, training, and staffing) of incorporating Big Data into financial statement audit practice, and what are the technical and cognitive skill sets auditors will need in a Big Data environment?
4. What would be the optimum approach to begin incorporating Big Data into financial statement audits, particularly, placing specific components of Big Data into specific phases of the audit process?

Based on our review of the literature and the comments made by practitioners at the 2014

and 2015 AAA Annual Meetinsg, accounting firms are in the very early and limited stages of incorporating Big Data into their audits. Because of the newness of these activities at the accounting firms there is likely to be a lack of quantitative data in the immediate future to conduct empirical research. Hence, most of the Big Data research that can be conducted in the near term will be some mix of design science, interviews, case studies, surveys, and other qualitative research, such as focus groups. A variety of research subjects will have to be examined in order to obtain insights and perspectives to answer the questions posed above. Specifically, companies, in general, as well as internal auditors are ahead of external auditors in their use of Big Data and, as such, can potentially provide rich research environments to help predict the answers to the above questions in the external financial audit domain.

I.   Research outside of the external audit domain

There is a vast and growing body of literature, both academic and nonacademic, that discusses how various industries are using Big Data as well as benefits, costs, issues, and concerns of using Big Data. Most of the Big Data activities discussed in the existing business literature are in the marketing domain, but those activities should not be discounted in the audit domain. Instead, researchers should look for parallels to auditing practice for insights on how that knowledge can be transferred from one domain to the other. For example, Big Data marketing research is searching for patterns in human behavior that are correlated with purchase practices and to subsequently predict future decisions. Thus, LinkedIn Corp. (LinkedIn.com) uses Big Data to determine the "People You May Know" list on their Web site (Dwoskin 2014 (WSJ 8/9/2014)). Auditing might take a related, but opposite perspective in terms of using Big Data to identify those who do not seem to follow predicted behavior in order to isolate outliers. This is already being done in the data mining of emails where various cliques are recognized in order to subsequently identify those who seem to be in a clique that they don't logically belong to (Debreceny & Gray 2011). For example, why is the CFO sending emails to a warehouse worker? Audit researchers could study corporate Big Data activities and specifically identify parallels to financial statement audits and then analyze how these activities could be modified for use in auditing practice. Interviews and case studies would be particularly valuable here.

Although general corporate Big Data activities may appear to differ from financial statement auditing, on closer examination there are parallels, particularly to the auditor's assessment of the effectiveness of the client's internal controls over its financial reporting (ICFR) process. Besides

marketing activities, corporate Big Data activities include analyzing compliance with corporate policies and procedures (identifying employees who are not following corporate procedures and policies) and improving the efficiency of logistics by systematically analyzing routing and warehousing activities—for example, by using GPS coordinates to determine which employees are not following prescribed routes (or visiting prohibited locations, such as bars). These and other related activities are essentially control monitoring/evaluation activities that could inspire ICFR-related audit procedures.

II.   Research internal auditor Big Data activities

Compared to the general Big Data uses described above, the characteristics of the internal auditor domain are much closer to the external auditor domain. However, there are two things different about the internal auditor domain that will be very valuable to researchers. First, internal auditors can potentially access their organization's Big Data more easily and more fully than external auditors. Both the Institute of Internal Auditors (IIA) and ISACA have published articles and other publications regarding Big Data and internal auditing, as mentioned above. Many parallels can be drawn from these publications and be applied to external auditors. These articles and publications might also help identify specific internal auditors that could be interviewed and organizations that could provide a basis for case studies. Both the IIA and ISACA have been very helpful in working with academics and could be approached about helping organize focus groups and invite members to complete online surveys.

Second, the internal auditors' use of Big Data would generally be broader than that of external auditors. For example, internal auditors might use a broad selection of nonfinancial data as part of an operational audit (as opposed to a financial audit). The experiences of internal auditors using nonfinancial data could provide valuable guidance for the external auditor's use of nonfinancial data.

III.   Accounting firms experiences with Big Data outside of financial statement auditing

Large accounting firms provide Big Data consulting services and their forensic groups are expanding their use of Big Data for their analytical activities. Through interviews, case studies, focus groups, and other qualitative research methods, valuable information could be gleaned from these non-audit activities that could answer the four questions posed above. Of particular interest might be the cost benefit parameters that could be derived from those non-audit activities. It's clear that as

more Big Data analytics are incorporated into financial statement auditing there will have to be a change in the mix of technical skills on the audit team. As we mentioned earlier in this paper, there is a growing shortage of Big Data analysts and, as such, these analysts are paid premium salaries. So, even if Big Data analytics could improve the efficiency and effectiveness of the audit, a critical research question is going to be whether this improvement in efficiency and effectiveness can justify the cost of the people needed to conduct the Big Data analytics. The kinds of skills these new Big Data auditors will need to have, both technical and cognitive, and how the accounting curriculum will have to change to teach them are all matters that warrant further research.

IV.   Gaining insights directly from external auditors

The incorporation of Big Data into financial statement audits is probably more likely to succeed if it is seen as an evolution in audit practice driven by easier access to a wider set of data rather than as a technology-driven revolution in analytics. Nearly every auditor has experience using client data and, maybe, external data to perform analytical procedures required for every audit. A growing body of auditors is expanding the data and the data analysis and data mining tools they use as part of an audit. As such, the experiences and opinions regarding drivers and inhibitors of current auditors can provide insights into the Big Data evolutionary process. These auditors can particularly address the paradigm-shifting aspects of Big Data, such as, moving away from sampling, the changing mix of skills needed on audit teams, and the increasing requirement of direct access to client data (particular NFD). Although the other groups of the above subjects could have direct or indirect experience or opinions, the following questions could be particularly addressed by experienced external auditors:

    a) Given the hierarchy of data in the Connolly (2012) framework (Figures 1 and 3), what is the marginal value to auditing of going beyond transactional data? Are higher levels of data substitutes or complements for traditional data inputs into the audit process?

    b) What unique aspects of Big Data are likely to have the biggest impact (positive or negative) on audit practice?

    c) To what extent is data integrity central to auditing and how can data integrity be addressed when dealing with "messier" non-financial and non-transactional data?

    d) Is unguided information collection appropriate or feasible in the auditing context?

e) Is the shift from causation to correlation inescapable in auditing and can audit practice evolve to deal that shift? Can auditors justify basing decisions on patterns observed in data for which they cannot offer an underlying theory?

f) Is it the case that Big Data requires stability in the environment and does that accurately describe a typical audit setting? When there is dynamic change in the environment, how should both Big Data and auditing respond?

6. Conclusion

Big Data is becoming an indispensable resource to many companies and has the potential to be an extremely valuable resource to financial statement auditors. But that presumption must not be taken as a given without question, skepticism or further research. Otherwise there is the danger that Big Data will succumb to the same forces that have stalled the adoption of prior technologies equally lauded in their time as having the potential to transform auditing practice.

We conclude with an example from our own experience that illustrates the importance of gaining a full understanding of what Big Data is and what its pros and cons are before applying it to auditing. Recall that earlier in the paper we cited Crawford (2013) who discusses a flawed Big Data analysis of twitter data to assess the impact of super storm Sandy. In that case the prevalence of twitter users in Manhattan gave the false impression that it was there that the most damage took place when in reality it was NJ that took the brunt of the storm. The point he and we were making was that in the absence of theory as to what Big Data analytics should be run and on what data, it is all too easy to fall into simplistic and erroneous interpretations of spurious or misleading correlations.

This particular illustration drew a pointed rejoinder from a referee of an earlier version of this paper, who wrote (emphasis added): *"some of the examples given are straw man arguments. For example, the discussion… concerning Twitter and Foursquare data creating the illusion that Manhattan was the hub of the hurricane Sandy disaster is suspect—wouldn't big data include an analysis of weather maps?"* We could have responded to that referee that the Big Data analysis Crawford (2013) cited was trying to assess the damage caused by the storm rather than its intensity, which is why the analysts turned to social media data. We could also have responded that the original study was conducted by a team of computer science researchers and practitioners with likely far better Big Data skills than most accounting researchers or auditors in an engagement.

But really, such rational responses are pointless, because when the referee says *"wouldn't big data include an analysis of weather maps?"* he or she is practicing what is properly called "magical thinking". Like many Big Data proponents who don't know what Big Data really is, the referee has an almost supernatural belief that Big Data will miraculously include all relevant data and provide perfectly accurate results as a consequence. There is no recognition that Big Data is only a means towards an end and how well it works depends on the choices made by the analyst about what data to include and how that data is analyzed. And the failure to grasp that fact that can make even highly capable statisticians make fundamental mistakes, which should raise doubts about what will happen when auditors in the field are urged to develop Big Data queries continuously.[29]

We wrote this paper precisely in order to counter this kind of magical thinking about Big Data that is all too prevalent among some of its advocates today and that referee's statement makes the case for why our paper is needed better than anything we could have written.

REFERENCES

1. AICPA. 2013. Audit Data Standard. Audit Data Standard Working Group. http://www.aicpa.org/InterestAreas/FRC/AssuranceAdvisoryServices/Pages/AuditDataStandardWorkingGroup.aspx.

2. AICPA, 2014. Why Predictive Analytics Should be "A CPA Thing". http://www.aicpa.org/InterestAreas/InformationTechnology/Resources/BusinessIntelligence/DownloadableDocuments/Predictive_Analytics_Final.pdf.

3. Alles, M. 2015. Drivers of the Use and Facilitators of the Evolution of Big Data by the Audit Profession. Accounting Horizons. Vol. 29, No. 2, pp. 439-449.

4. Alles, M, H. Issa and M. Vasarhelyi. 2014. An Analysis of the Design and Impact of the AICPA Audit Data Standard (ADS). Unpublished working paper, Rutgers Business School.

5. Alles M., A. Kogan, and M. Vasarhelyi. 2002. Feasibility and Economics of Continuous Assurance. Auditing: A Journal of Practice and Theory. 21 125-138.

6. Alles M., A. Kogan, and M. Vasarhelyi. 2008. Putting Continuous Auditing Theory Into Practice: Lessons From Two Pilot Implementations. Journal of Information Systems. 22 (2) 195-214.

---

[29] This is not idle speculation. Speaking at the EY reception at the 2015 American Accounting Association annual meeting the head of the firm's Tax practice stated that he expected each of this associates to come up with a question that required Big Data analysis every single day.

7. Anderson, C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine. July 16.
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

8. Ball, R. and P. Brown. 1968. An Empirical Evaluation of Accounting Income Numbers, Journal of Accounting Research. 6 (2) 159-178.

9. Brown, C. E. (1991). Expert systems in public accounting: current practice and future directions. **Expert Systems with Applications**, 3 (1): 3–18.

10. Brown, C., J. Wong and A. Baldwin. 2007. A Review and Analysis of the Existing Research Streams in Continuous Auditing. Journal of Emerging Technologies in Accounting. 4 1-28.

11. Cao, M., R. Chychyla and T. Stewart. 2015. Big Data analytics in financial statement audits, Accounting Horizons, Vol. 29, No. 2, pp. 423-429.

12. Clement, D. 2013. Interview with Susan Athey. The Region: The Federal Reserve Bank of Minneapolis. June 10.
http://www.minneapolisfed.org/publications_papers/pub_display.cfm?id=5112.

13. Connolly, S. 2012. 7 Key Drivers for the Big Data Market. Blog posting, May 14th.
http://hortonworks.com/blog/7-key-drivers-for-the-big-data-market/.

14. Crawford, K. 2013. The Hidden Biases in Big Data. Harvard Business Review Blog Network. April 1. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/

15. Cukier, K. and V. Mayer-Schoenberger. 2013. The Rise of Big Data: How It's Changing the Way We Think About the World. Foreign Affairs. May/June.

16. Debreceny. R.S. and G.L. Gray. 2010. Data mining journal entries for fraud detection: An exploratory study. International Journal of Accounting Information Systems 11 (3), 157-181

17. Debreceny. R.S. and G.L. Gray. 2011. Data mining of electronic mail and auditing: A research agenda, - Journal of Information Systems, 25 (2): 195-226.

18. EY. 2014. Global Forensics Data Analytics Survey.
http://www.ey.com/GL/en/Services/Assurance/Fraud-Investigation---Dispute-Services/EY-Global-Forensic-Data-Analytics-Survey-2014

19. Foster, G. 1981. Intra-industry information transfers associated with earnings releases. Journal of Accounting and Economics 3 201–232.

20. Gartner, 2012. Gartner Says Big Data Will Drive $28 Billion of IT Spending in 2012. Press Release, October 17th. http://www.gartner.com/newsroom/id/2200815.

21. Gartner, 2014. Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years. Press Release, September 17th. http://www.gartner.com/newsroom/id/2848718.

22. Ginsberg, J., M. Mohebbil, R. Patel, L. Brammer, M. Smolinski and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature. 457 1012-1014.

23. Gray, G.L. and R.S. Debreceny. 2014. A taxonomy to guide research on the application of Data Mining to Fraud Detection in Financial Statement Audits. International Journal of Accounting Information Systems. Volume 15, Issue 4, Pages 357–380.

24. Gray, G.L., V. Chiu, Q. Liu, and P. Li. 2014. The expert systems life cycle in AIS research: What does it mean for future AIS research? International Journal of Accounting Information Systems. Volume 15, Issue 4, Pages 423–451.

25. Hoitash, R., A. Kogan, and M. Vasarhelyi. 2006. Peer-Based Approach for Analytical Procedures. Auditing A Journal of Practice & Theory. 25 (2) 53-84.

26. IIA. 2013. Big Data: Collect It, Respect It. Institute of Internal Auditors: Tone at the Top. Issue 63. August.
https://global.theiia.org/knowledge/Public%20Documents/TaT_August_2013.pdf.

27. Keltanen, M. 2013. Why 'lean data' beats Big Data. Guardian Newspaper. http://www.theguardian.com/media-network/media-network-blog/2013/apr/16/big-data-lean-strategy-business.

28. Kogan. A., M. Alles, M. Vasarhelyi and J. Wu 2014. Design and Evaluation of a Continuous Data Level Auditing System. Auditing: A Journal of Theory and Practice. Vol. 33, No. 4, pp. 221-245.

29. Littley, J. 2012. Leveraging data analytics and continuous auditing processes for improved audit planning, effectiveness and efficiency. KPMG white paper. http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/data-analytics-continuous-auditing.pdf.

30. Lucas, S. 2012. Beyond the Balance Sheet: Run Your Business on New Signals in the Age of Big Data. Blog posting, August 21st. http://www.saphana.com/community/blogs/blog/2012/08/21/beyond-the-balance-sheet-run-your-business-on-new-signals-in-the-age-of-big-data.

31. Mayer-Schoenberger, V. and K. Cukier. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt. ISBN: 9780544002692.

32. McKinsey, 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, June. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

33. Moffitt, K. and M. Vasarhelyi. 2013. AIS in an Age of Big Data. Journal of Information Systems. Fall. Vol. 27, No. 2, pp. 1-19.

34. Moore, G. A. 2002. Crossing the Chasm. New York: Harper Business.

35. New York Times. 2014. Can Big Data Tell Us What Clinical Trials Don't? October 5th. http://www.nytimes.com/2014/10/05/magazine/can-big-data-tell-us-what-clinical-trials-dont.html?ref=magazine.

36. pwcNoreen, F. 2011. Troubled Audit Opinions. New York Times. June 9. http://www.nytimes.com/2011/06/10/business/10norris.html.

37. PCAOB. 2014. Matters Related to Auditing Revenue in an Audit of Financial Statements, Staff Audit Practice Alert No. 12. September 9.

38. Rowley, J. 2007. "The wisdom hierarchy: representations of the DIKW hierarchy". Journal of Information Science 33 (2): 163–180.

39. Setty, K. and R. Bakhshi. 2013. What Is Big Data and What Does It Have to Do With IT Audit? ISACA Journal. Vpl. 3. pp. 1-3.

40. Vasarhelyi, M.A and F.B. Halper. 1991. The Continuous Audit of Online Systems. Auditing: A Journal of Practice and Theory. 10 (Spring), 110–125.

41. Whitehouse, T. 2014. Auditing in the Era of Big Data. Compliance Week. April 29. http://www.complianceweek.com/auditing-in-the-era-of-big-data/article/7XljSe2GNuI1/t/.

42. Yoon, K., L. Hoogduin, and L. Zhang. 2015. Integrating different forms of data for audit evidence: markets research becoming relevant to assurance, Accounting Horizons, Vol. 29, No. 2, pp. 431-438.

43. Zikopoulos, P. C., D. deRoos, K. Parasuraman, T. Deutsch, D. Corrigan, and J. Giles. 2013. Harness the Power of Big Data: The IBM Big Data Platform. McGraw-Hill. New York, NY.

44. Zhang, J., Yang, X. and Appelbaum, D. 2015. Towards effective Big Data analysis in Continuous Auditing. Accounting Horizons, Vol. 29, No. 2, pp. 469-476.

45. Zins, C. 2007. "Conceptual Approaches for Defining Data, Information, and Knowledge". Journal of the American Society for Information Science and Technology (Wiley Periodicals, Inc.) 58 (4): 479–493.