Big Data and the Audit Challenge[1]

Hart Will, Dipl.-Kfm., Ph.D., CPA, CMA.

Professor Emeritus of Information Systems and Computer Auditing
University of Victoria, Canada

Abstract

"Big Data" is a technological term with a seemingly cognitive connotation that masks an ideological orientation of those attempting to be benevolently, criminally of even "innocently" in control of our knowledge and subsequent actions. Without an epistemological foundation "small" and especially "big" data are a myth. When "the truth" becomes "what's on a digital screen" under the control of those in charge of "the cloud" we may be clouding our cultural heritage to an extent that makes us exposed to and manipulated by those screening and displaying our data. As a consequence, subsequent information and knowledge cannot be critically assessed and audited for lack of evidence. All lessons learned during the centuries of enlightening efforts seem to be forgotten and ignored by or unknown to those in control of modern information technology. They act primarily for their own economic, political, and social benefits and feel "justified" by the *big-data-ideology*. Knowledge must remain relevant to, testable and rationally believable by the legitimate recipients of any and all data and information. The audit challenge has become phenomenal during the "digital big data age!" A framework for data governance is overdue in those fields heavily dependent on information technology.

Keywords: Audit theory, big data, data governance, information systems theory

Introduction

"Big Data" is a technological term with a seemingly cognitive connotation that masks an ideological orientation of those attempting to be benevolently, criminally or even "innocently" in control of our knowledge and subsequent actions. Without an epistemological foundation "small" and especially "big" data are a myth. When "the truth" becomes "what's on a digital screen" under the control of those in charge of "the cloud" we may be clouding our cultural heritage to an extent that makes us vulnerable to those screening our data. As a consequence, subsequent information and knowledge cannot be critically assessed and audited for lack of evidence. All lessons learned during the centuries of enlightening efforts seem to be forgotten and ignored by or unknown to technocrats and those in control of modern information technology - primarily for their own economic, political and social benefits and 'justified´ by the *Big-Data-Ideology*:

[1] Keynote Address, **International Workshop on Computer Auditing Education in Vancouver, B.C. at the SFU Downtown Campus on July 9, 2015.**

> Ideology is defined as "the body of doctrine, myth, belief, etc., that guides an individual, social movement, institution, class, or large group… and such a body of doctrine, myth, etc., with reference to some political and social plan, as that of fascism, along with the devices for putting it into operation." (Google)

To unmask the big-data-ideology requires a review of information systems concepts as we have known them for at least 25 years: Observations are not yet data; data are not yet information; information is not yet knowledge; and wisdom is still something quite different. Figure 1 illustrates the relationships between these concepts and constructs as part of an information systems paradigm within a frame of reference or knowledge context. (Will 2000, Fetzer 2000).

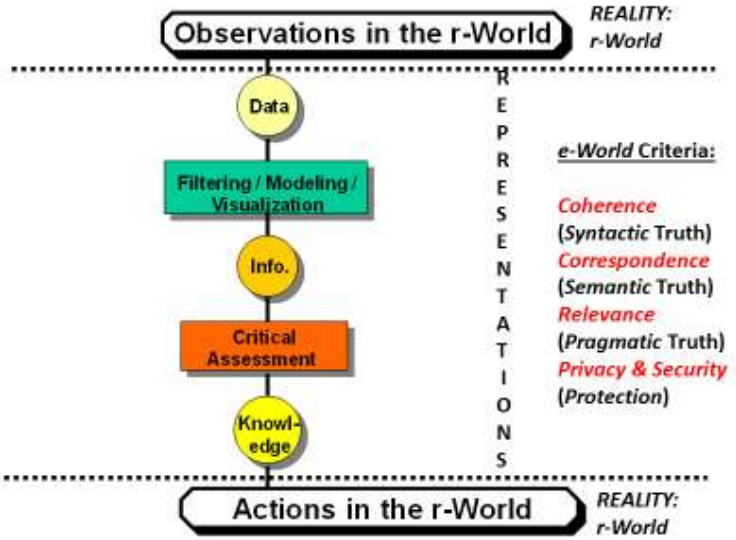

Figure 1: Information Systems Paradigm

If we limit our understanding of the world to digital monitoring, coding, processing and display of data and to their algorithmic manipulation by means of modern computer technology then we may be trapped in our own digitized myths without concern for and reference to ethics, privacy, security, transparency and truth. Therefore, this framework needs to be extended into its meta-dimensions, i.e., *beyond* observations, data, information, and knowledge in order to provide the respective knowledge context for these intellectual efforts as illustrated in Figure 2. (Will 2000, 2006).

Logs represent evidence about the kinds of data processing, information processing and knowledge processing performed within a particular frame of reference. They facilitate assessments and critical re-assessments of any procedures applied. While we make observations with any and all of our senses, controlled by our brains and minds, electronic sensors of various kinds may measure and monitor numerous aspects of reality while missing others which attentive and critical observers would notice. Not only the collected data need to be understood (or ignored) in their respective context, but any further processing when applying algorithms and models must also make sense and be explainable to the recipients of derived and displayed information.

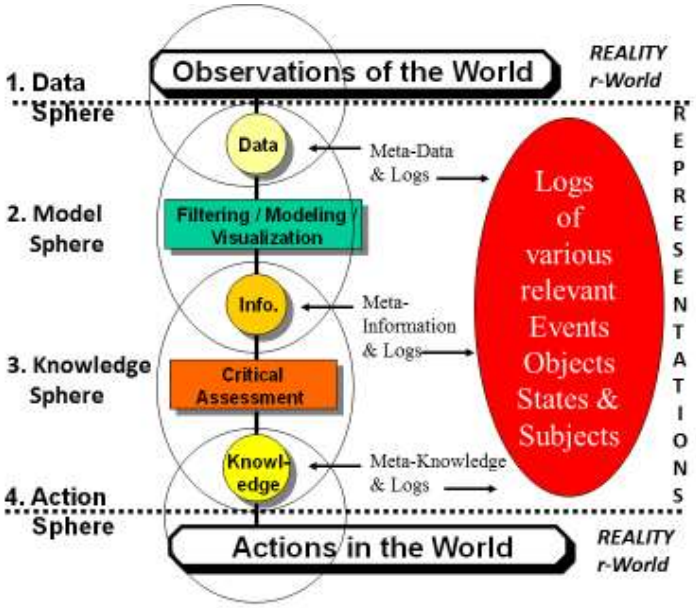

Figure 2: Information Systems Concepts and Constructs

If electronic sensors and monitors are automatically coupled with pre-programmed control systems then their programmers are in control rather than any of the affected people. Before the information can be believed it must therefore be possible to assess it critically. How else can it be accepted, ignored or rejected by the recipients on rational grounds? We need meta-data, meta-information and even meta-knowledge (understood as wisdom) in each specific intellectual context.

When the providers of the knowledge elements and objects are different from any of the sign users at each level of the cognitive hierarchy depicted in Figure 1 and 2, then we are faced with potentially conflicting objectives and purposes as illustrated in Figure 3. (Will 2000, 2006) They can only be rationally reconciled with respect to a common understanding of the truth (see below).

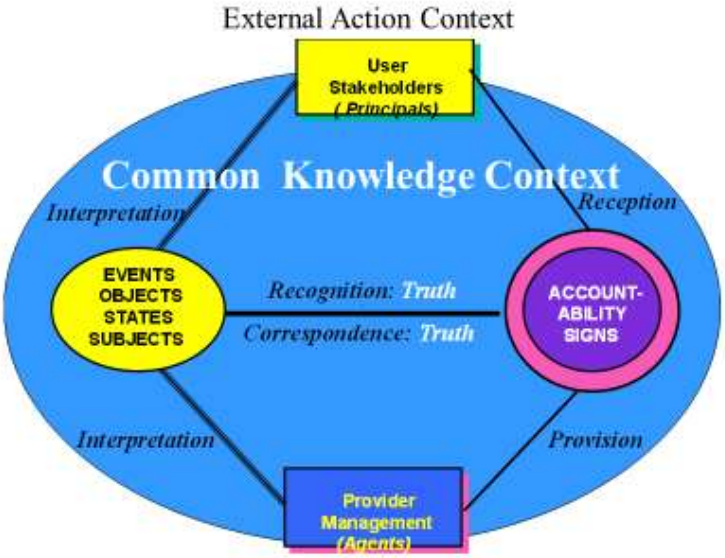

Figure 3: Double-Semiotic Perspective

What is processed and displayed as "computer screen reality" at each level in the

cognitive hierarchy may not meet the recipients' expectations; may cloud their perceptions and inferences; or may even deliberately mislead them. It is therefore important to distinguish various levels of knowledge as justified belief, to recognize the different degrees of certainty associated with each, and to assess their effect on our actions:

> Knowledge of laws of nature which cannot be violated and cannot be changed, takes predictive primacy over knowledge of relative frequencies that have obtained in the past. When we possess knowledge of single-case propensities, therefore, they ought to determine the values of corresponding degrees of belief for inference and decision. When knowledge of single-case propensities is unavailable, however, then degrees of belief should be determined by beliefs about corresponding relative frequencies. In cases where neither knowledge of single-case propensities nor knowledge of relative frequencies happens to be available, however, then decision making depends upon hypothetical reasoning or educated guesswork, where rationality of action tends to be decoupled from rationality of belief. Actions under conditions of this kind are not only extremely risky but are subject to the influence of psychology and ideology. (Eells and Fetzer (eds.), 2010: xxxiii)

Relying naively or uncritically on anonymous computer screen contents implies unintelligent action. If there is only one world-view on display without reference to logs then not even an auditor can provide assurances for lack of evidence. Worse still, providing one's own private data without knowing what is done with it – and even giving up all ownership rights as part of Terms of Service (ToS) agreements - is an abdication of responsibility for the truth about oneself, i.e., for one's own identity! How else would "identity theft" be possible?

When human behavior is automatically monitored and subsequently used to identify or typify persons for such purposes as exposing them "automatically" to tailor-made advertising; to fraudulent collection, payment or phony reward schemes; and to other illegitimate and criminal activities then those in control of the digital screens and their contents can direct the world to their own advantage rather than according to a fair, legal and negotiated (social) contract. In fact, the lengthy ToS for using various internet services seem to be designed and written to be confusing and incomprehensible for a normal person. They are therefore not a normal contract between equal parties. This way, not only numbers and text, but images and voice recordings and even the users' various locations according to GPS coordinates can be monitored. Even their rhythms such as gestures, heart beats, key-strokes, voice inflections, and other behavioral attributes can be measured and recorded for (yet) unknown purposes.

Once these data are openly or surreptitiously collected and stored in special "big-data-formats" in modern computer "memories" they become un-erasable and no longer (easily) traceable for the data subjects themselves and for authorized outsiders like auditors. Their data structures and storage structures are hidden below the information structures derived from them and then displayed selectively on computer screens. To question these information contents becomes difficult and is often impossible without accessible meta-data and meta-information.

Auditing is a rational effort by independent third minds (or parties) to assess the truth of data and any information derived from data that are collected and maintained by providers. They may display exclusively their own "virtual world view" instead of independently observable reality or merely slices of it. Knowledge must remain

relevant, testable and rationally believable to the legitimate recipients of any and all data and information. How else can we try to truly understand our world and act accordingly? The audit challenge has become phenomenal in the "digital age" and needs to be assessed in the context of data (small and big), derived information, and credible knowledge.

Observations, Measurements, and Signs

Living observers are always curious with respect to "anything of survival interest" that is noticeable, measurable, pleasurable, satisfying or threatening - even without a formal purpose and specific hypothesis. This may involve recording of the observations or measurements as signs which can be stored for subsequent uses and listened to or read as digitized messages. While our senses allow us to make all kinds of observations that may be relevant to our success or survival, our minds evaluate these findings in their respective context. We may also communicate our insights and exchange ideas about them with others in order to assess and share our findings critically or uncritically. The "internet of people" assumes communication between intelligent people, possibly supported by artificial intelligence, but certainly not by uncritical intelligence. (Fetzer 1990).

Analogously, computer-monitored observations or measurements may be evaluated automatically by means of control units that trigger the execution of programs as single actions or complicated processes, delayed or immediately. This situation is often referred to as "the internet of things" which implies "artificially intelligent" objects such as fridges or illumination systems.

To describe the context and purpose of original observations generally as *data signs*, i.e., syntactically, semantically, and pragmatically, we can identify their cognitive effects (awareness, correspondence, purpose); their uses (perception, recognition, interpretation); and their semiotic properties (causation, ground, *interpretant*) as illustrated in Figure 4. (Fetzer 1990, 2000, Will 2006).

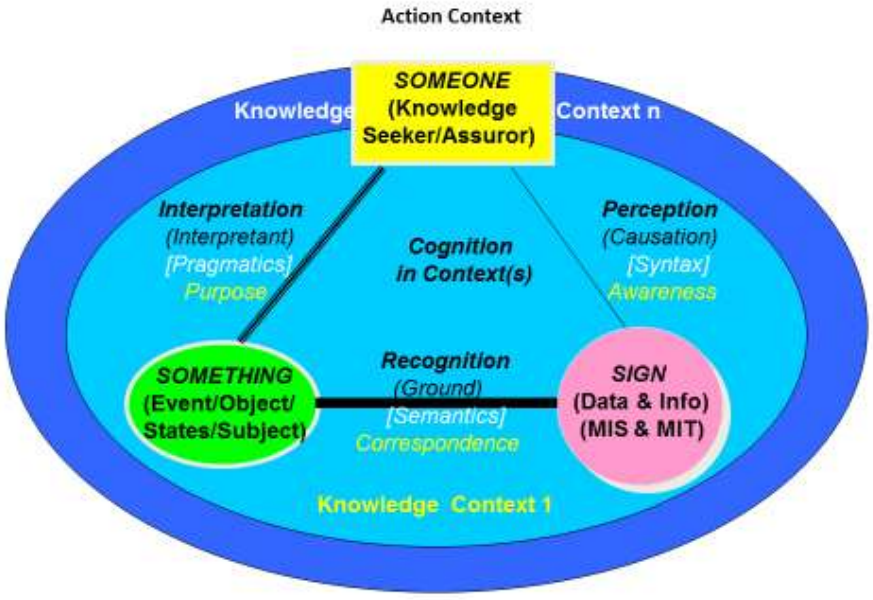

Figure 4: The Semiotic Perspective (Detail)

As signs in a semiotic sense data are made up syntactically of symbols of and in a language. They have a meaning in that language which is understood (semantically) by all competent users of that language; and they have a pragmatic purpose beyond their meaning both for the originators and the recipients, although these purposes may be contradictory. It is therefore important to understand the true meaning of data as signs from the perspectives of both their originators and their users; whether they are private and protected, or publicly available and technically accessible.

A syntactically correct expression may have different meanings and can therefore be semantically ambiguous at best and false at worst. Similarly, a syntactically and semantically correct expression may be used for legitimate and legal or for illegitimate and illegal purposes. One-sided semiotic depictions of data by their originators need to be avoided since identical data can have different meanings and different purposes for the recipients of the data as illustrated in Figure 5 in an accountability context. (Will 2006, 2015). After all, the recipients of the accountability signs need confidence and ideally trust in the provider.

Figure 5: Accountability

For example, accounting transactions describe the observed or automatic (preprogrammed) exchange of goods or services for monetary resources between economic agents at specific dates (and times) in specific markets; however, the recorded economic history may be erroneous, incomplete or even deliberately misleading. Syntactically correct “debit” or “credit” entries in the accounts can have different meanings either as true economic transactions or as actual or hypothetical adjustments to the accounts in recognition of depreciation, receivables, shrinkage or theft and various effects on asset values and economic performance. Each adjustment can have different purposes depending on its effect on calculated profit or loss and the expected reactions of the intended recipient of the information. Thus, accounting transactions need to be (and are commonly) auditable by a third party in order to determine their linguistic structure, their true meaning and their evident (or hidden) purpose(s) both for the collectors and the recipients of derived information.

Data and Meta-Data

Evidently, to understand data requires knowledge about their structure (syntax), their meaning (semantics), and their purpose (pragmatics) as illustrated above in Figure 4. Meta-Data are therefore required if anyone but the originator needs or wants to understand the data. For example, accounting data refer to the types of objects traded, produced and inventoried and the corresponding kinds of payments made or received at specific times in various degrees of detail for processing by human accountants or machine accounting programs.

Thus, accountants describe their data by means of formal symbols and references to accounts and ledgers within their double-entry language such as "debits" and corresponding "credits" in order to be able to trace them and to control their accuracy. The familiar charts of accounts, journals, ledgers, sub-ledgers and trial balance provide structural meta-data in financial accounting. We can therefore understand accounting data as records of *real* transactions; however, understanding the *hypothetical* adjustments made by accountants requires meta-data of a different kind, namely about the age and (ab)use of machinery, the shrinking of inventories, the cash collection experience in general and with respect to individual customers, exchange rates of different currencies at various times, market conditions in different markets, etc. Not to recognize these differences in terms of data and meta-data will cloud the derived information that may appear on accountants' screens as their financial statements.

Critical data reviews are therefore the cognitive foundation of modern auditing of private and public sector financial and tax accounting. Auditors must be able to trace both the transaction data to their origins in reality and the adjustment data to the assumptions made. The data structures employed must be accessible and understandable, regardless whether operated in manual or electronic modes. (Not too long ago auditors even advocated "auditing around the computer"!) In order to structure data electronically, various data models have been used to organize small data: linear lists, hierarchic and network structures and relations. All of them are transparent with modern audit software which allows them to be read "in the original," selected and connected or networked according to numerous criteria and even renamed for more understandable audit logs and reviews (see also www.acl.com).

New methods and technology have provided opportunities to centralize and arrange small data into even more complex data structures, now known as "big data;" however, little consideration seems to be given to their transparency and even less to their auditability.

Big Data and Meta-Big-Data

The modern term "big data" implies, of course, that there exist also "small data" as we have known and stored them since Babylonian clay tablets, papyrus and paper records, on punch cards and more recently on magnetic tapes, discs, computer chips and now even in "the cloud." So what is so mythical or even mystical about big data?

Figure 6 is a modification of Figure 2 to illustrate that the data, model, and knowledge spheres are now conflated without reference to any of the respective logs and their meta-features. By providing only provider-designed information screen images

without references, even the knowledge and action contexts are indistinguishable such that preferred action is suggested to the information user. Moreover, without knowledge about the size of the available or searched data base, such "suggestions" by monopolistic providers may be not only biased in favor of the respective provider, but deliberately designed to eliminate any competition. (EU vs. Google, 2015)

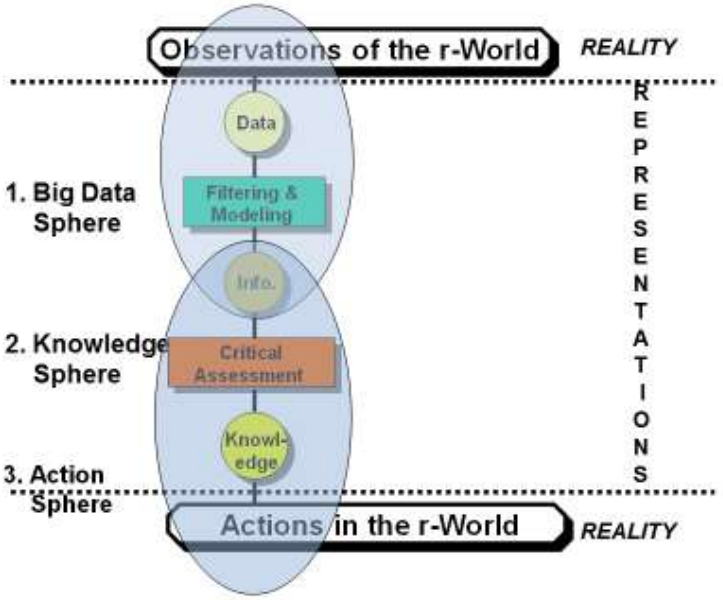

Figure 6: Big Data Sphere

The following definitions from Wikipedia allude to the connection between small and big data; however, they are unfortunately neither enlightening nor illuminating:

> Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS". (Wikipedia)

The primary definitions of big data conflate the data structural problems of information administration (organizationally) and information management (technologically) and the related data processing issues encountered:

> Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. (Wikipedia)

Since when have data processing applications been "difficult to process using traditional data processing applications"? - The IT industry has been constantly evolving and innovative during the last half-century (as some of us can still vividly remember). Are these challenges now more easily overcome by mythical big data analyses?

> The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. 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, prevent diseases, combat crime and so on."(Wikipedia – footnotes omitted).

Another definition of big data from Wikipedia addresses even some of the epistemological issues faced in the modern digitized world; however, as we will see, to no intellectual avail:

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." (Gartner). Additionally, a new V "Veracity" is added by some organizations to describe it. (Wikipedia – Footnotes omitted)

If big data is explained information-technologically as „high volume, high velocity and/or high variety information assets" although one wants to answer epistemological questions such as „enhanced decision-making," „insight discovery," „process optimization" und most recently also „veracity" then we are required to ask and consider more profound questions:

1. Who or what is depicted by the data: animate subjects or inanimate objects – and what rights are associated with each?
2. Which features of data subjects or data objects are observed, measured, monitored, collected and stored by whom, how and when?
3. How are the data structured, and can these structures be standardized and publicized for audit, security, and transparency purposes?
4. Who owns the data?
5. Who uses the data for what purposes, where and when?
6. How are the data protected against abuse, misuse and illegal access?
7. Do legal rights and obligations exist to guide and protect owners, users and auditors?

Answers to these questions would not only demystify and demythologize the big-data-ideology as a grandiose IT scam, but also provide the basis for an overdue comprehensive data governance framework. (Hua ???)

Another attempt to distinguish the concepts of "big data" and "business intelligence" with respect to statistical methodology matches and mixes different cognitive categories – an epistemologically inadmissible method: As we all know, even descriptive statistics are inductive; statistical „laws" are not laws of nature that cannot be violated (although statisticians seem to believe in their equivalent); correlations are not causations; and clusters may improve „information density", but they do not represent single-case propensities to improve on relative frequencies that where observable in the past.

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors. (Wikipedia – Footnotes omitted).

As we all know, even descriptive statistics are inductive; statistical laws are not laws of nature that cannot be violated (although many statisticians seem to believe in their equivalent); correlations are not causations; and clusters may improve „information density," but they do not represent single-case propensities to improve on relative frequencies that where observable in the past.

„Revelation of relationships and dependencies" is based on specific purposes such as advertising and marketing to specific types of persons who belong to various clusters

such as „people who are most likely to be African American or Hispanic, working parents of teenage kids, and lower middle class and shop at discount stores“ or „Caucasian, high-school educated, rural, family oriented, and interested in hunting, fishing and watching NASCA.“ (Goodman, 2015: 67) – Of course, whether the members of these clusters behave really as expected must be ascertained or audited independently.

„Predictions of outcomes and behaviour“ become possible if people in such clusters can be monitored and compared with each other in similar situations such that their behavior becomes predictable according to specific laws of nature (e.g., caused by birth defects, diseases, handicaps, etc.); according to specific propensities (e.g., various kinds of addictions or typical consumer behavior); and with relative frequencies. This means that the more we know about a person's attributes, behavior and possessions in various situations and locations the easier it becomes to predict and control her or his behavior. The data privacy questions have become the central issue.

A whole industry has evolved to condense „free“ small data into profitable „proprietary“ big data for sale to and subsequently by data brokers for profit:

> The goal of Acxiom and other data brokers is to provide what is alternatively called „behavioral targeting,“ „predictive targeting,“ or „premium proprietary insights“ on you and your life. In plain English this means understanding you with extreme precision so that data brokers can sell the information they aggregate at the highest price to advertisers, marketers, and other companies for their decision-making purposes. (Goodman, 2015: 67)

Big data have also been circumscribed only recently (2014) as both a major information-technological challenge and a privacy concern by Eric Schmidt, Executive Chairman of Google, and one of the most successful entrepreneurial information technocrats from Silicon Valley:

> This is the „big data“ challenge that government bodies and other institutions around the world are facing: How can intelligence agencies, military divisions and law enforcement integrate all of their digital databases into a centralized structure so that the right dots can be connected without violating citizens‘ privacy? (Schmidt and Cohen, 2014:174)

Notice that creating one or more centralized and integrated super-data-structures for „intelligence agencies, military divisions and law enforcement“ does not necessarily protect the citizen's privacy without proper and accessible meta-big-data that can be used for legally required audits and compliance tests! (See below)

Interestingly, even someone very much aware of the risks associated with modern information technology and its uses and abuses in the digital age does not define big-data either, but circumscribes it as an almost pious belief and naïve expectation and refers to their economic uses as follows:

> The promise of big data is that long-standing complex problems become quantifiable and thus empirically solvable... Across all industries, whether retail, transportation, or pharmaceuticals, there will be tremendous economic value realized as a result of big data, so much so that the World Economic Forum recently dubbed it „the new oil.“ (Goodman, 2015:85).

The world's "wise" people who met in Davos seem to have overlooked that data – big or small – need to be transformed into information which can serve numerous purposes, some beneficial and some detrimental. Since we can observe more differences between undefined big data and refined oils, the analogy does not seem to

hold, unless either is viewed as an exploitable resource for those who possess and treat it in a capitalistic sense. Although oil can also be mined (cracked) and refined in various ways, it makes some people enormously rich and others dependent, disadvantaged and poor. The ethical and social dimensions of the use of such precious resources remain conveniently ignored by those in control.

One of the most serious issues is the lack of control over data collectors such as Facebook, Google, Twitter, and Yahoo and over subsequent data brokers:

> Today's modern data brokers, unlike credit reporting agencies, are almost entirely unregulated by the government. There are no laws, such as the Fair Credit Reporting Act, that require them to safeguard a consumer's privacy, correct any factual errors, or even reveal what information is contained within their systems on you and your family. (Goodman 2015: 68).

Of course, if the users of search engines, social media and various apps give up all rights to their behavioral data (heart beats, measurements, pictures, temperatures, text, voice, and words in various ways) "voluntarily" by signing complicated, lengthy and incomprehensible terms of service agreements (ToS) such as the ones demanded by Google then we may become "transparent" and "objectively" known, but lose practically all control over our own identity:

> When you upload or otherwise submit content to our services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify and create derivative works, such as those resulting from translations, adaptations or other changes and license to communicate, publish, publicly perform, publicly display and distribute such content. (Quoted in Goodman, 2015:59)

Data seem to have lost their value for those who provide them ignorantly, naively, and voluntarily - according to the Hippie mentality of "couch surfing" or "sharing everything without protection?" – solely for the convenience of surfing the internet and searching the world wide web at various levels. By connecting "free" small data into centralized and integrated big data as targeted information, the collectors of data about our behavior, the data brokers, and their customers have not only become immensely wealthy but also dangerously powerful and uncontrollable. They are the modern oracles – not of the Delphi kind, but of the Silicon Valley variety – and that in the age of enlightenment!

Information & Meta-Information

A proper understanding of information systems as both a cognitive challenge and a technological construct way beyond traditional accounting has led to modern forms of (organizational) data administration, (technological) data management, and sophisticated computer programming of filters and models in terms of algorithms. It is useful to remember the admonition of one of the leading scholars of computer programming and his definition of an algorithm before we fall for a mystical explanation related to big data:

> An algorithm must be seen to be believed, and the best way to learn what an algorithm is all about is to try it… The modern meaning for algorithm is quite similar to that of *recipe, process, method, technique, procedure, routine,* except that the word "algorithm" connotes something just a little different. Besides merely being a finite set of rules which gives a sequence of operations for solving a specific type of problem, an algorithm has five important features: … Finiteness…Definiteness… Input…Output…Effectiveness. (Knuth, Vol. 1, 1973: 4-9).

When data are filtered, condensed or modeled into information structures of less detailed forms, then the semiotic dimensions illustrated in Figures 4 and 5 still apply. Informative signs such as data need to be perceived by someone who is causally and syntactically aware of them in a knowledge context; they are grounded in real or well-defined hypothetical events, objects states or subjects and recognized as such semantically in languages that represent this correspondence; and they are pragmatically interpreted by a knowledge seeker for specific purposes.

Big data complicate the issues since we have hardly any meta-big-data available to enlighten us. Not knowing the algorithms used to generate such information profiles as the mentioned clusters for targeting advertising or marketing nor the actual data and meta-data applied ought to make us hesitant to believe the information. Also, not knowing the providers of the information (such as search engine and social network suppliers or data brokers) and their reasons for processing the variously connected and centralized data is no basis for belief in any of it, much less reason for providing the often private small data voluntarily. As Goodman suggests, it would be much less risky to pay for the use of such products and services and to protect our privacy and security legally, accountably, and auditably.

Already the annoyance felt when we use the internet and are bombarded with targeted advertising should make us hesitant to share our behavioral data voluntarily on the one hand. On the other hand, "hits" recorded on cluster-based-selected advertisements around an internet user's screen image may not be causally related to purchases. As long as the advertisers believe the big-data-mystique, they will continue to pay for the "refined" data, but at our "cost" or "donation."

Although the traditional audit trail details consisting of financial data and relevant meta-data may be still commonly available and relatively easily retraceable, we cannot assume that the information automatically derived from big data is believable as such (knowledge) and provides a rational basis for subsequent action.

> The problem is that we are leading lives fully intermediated by screens and other technologies that, although they give the appearance of transparency, are in fact programmed, controlled, and operated by others. Worse, none of us have a freaking clue as to how any of it works. (Goodman, 2015: 165)

To forget or ignore the interest of the users of information systems while concentrating on the providers and those maintaining them administratively and technically means to abdicate responsibility for critical and ethical thinking about the truth associated with information derived from any data.

Knowledge and Wisdom

At least since the classical Greek period in the fourth century B.C. do we know the distinction between *Mythos* and *Logos.* Rather than asking the Oracle of Delphi, we now ask the oracles of Silicon Valley for answers to our questions; however, while viewing electronic screens filled by algorithmic and digital wonders we may forget to ask critical questions. *Where is the Socrates of the digital age?*

When confronted with insights based on big data and displayed on electronic screens

we seem to suppress doubts and ignore critical thinking, as if they were superfluous when viewed against the mythical dogma of "overwhelming big-data-evidence:"

> It is as if we have transformed into an "in screen we trust" culture. If something is on a screen, whether it be a computer, iPad, industrial control system, street sign, GPS device, radar installation, or mobile phone, our first inclination is to trust what we see before us. However, we have shown time and time again that everything from our friends on Facebook to the numbers we dial on our mobile phones can be rigged to deceive us. (Goodman, 2015: 165)

We seem to have forgotten to think about knowledge as the result of critical thinking and logical reasoning applied to data and information in order to assess their relevance, validity and truth.

> If we identify "thinking" with the context of discovery and "reasoning" with the context of justification, then it is indeed correct that logical reasoning is not sufficiently flexible to serve as a foundation for *thinking*. But it certainly does not follow that logical consistency and deductive closure are therefore properties that are neither available nor desirable within contexts of *reasoning* in general. (Fetzer, 1990: 231-232)

While we can formulate hypotheses and theories without concern for completeness and consistence, their rational appraisal is dependent on it:

> An expert whose knowledge could not possibly be true because it was inconsistent would not be worth the bother. An expert whose knowledge was consistent but unsupported by the available evidence would hardly be worth utilizing. Perhaps the underlying moral that emerges from this discussion, therefore, that the most important decisions confronting those working in this field [of artificial intelligence] involve determining exactly what "knowledge" is worth presenting. (Fetzer 1990: 232).

To believe information derived from data without sufficient meta-data as evidence to assess their relevance, completeness, consistency and truth would be irrational although modern audit software applied by critical and logical minds can help quite a bit in analyzing computer-based data and filling gaps of belief. To believe information without sufficient meta-information about the algorithmic structures and procedures employed without being able to reproduce them with available data would not only be irrational but also unwise as an assessment of the results.

Wisdom is the final stage in our depiction of the knowledge hierarchy.

> It has been defined as the quality or state of being wise; knowledge of what is true or right coupled with just judgment as to action; sagacity, discernment, or insight. (Google reference).

It may be wise not to apply logical conclusions derived from big data to action since logical consistency and deductive closure may not be certain. It takes wisdom to see through the big data ideology, to de-mystify and to de-mythologize it sufficiently to create a legal framework analogous to financial auditing which could be called *data governance*. It may also be wise to give up providing "free" data and information to the various operators of internet services and of the internet of things in order to protect our personal authenticity, privacy, property, and security.

The ultimate irony with respect to big data and modern information technology is the following definition and equation:

Internet + Internet of Things = Wisdom of the Earth.
(Wen Jiabao, Chinese Premier, quoted in Goodman, 2014: 259).

Audit Challenges

Auditing by third parties, understood as critical analysis of information provided by someone (or some computer system) to someone else as legitimate user, is only possible if the underlying observations, data and information are properly documented within a frame of reference such as manual or electronic financial accounting and data governance in various fields such as medical and pharmaceutical research treatment, to name just a few. There should not be a secret or tacit disconnect with respect of the truth between legitimate suppliers and users of the data and derived and displayed information. Although most accountants can be trusted by the users of their information, historical experience has shown that deception and fraud cannot be excluded in accounting either, although the accounting algorithms and models are fairly transparent and accounting data can be traced through such systems. While accounting has a tradition of transparency with respect to real transactions, adjustments made to the data prior to sharing them with others have often been secretly motivated.

The audit situation is illustrated in Figure 7 in the context of discovery and justification where the audit opinion, understood as true belief and assurance, is derived from accountability signs (see Will 2000, 2006, 2015).

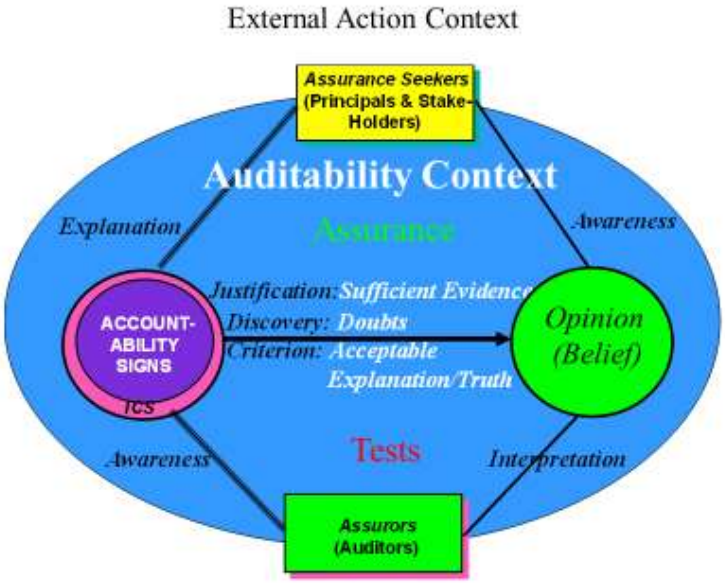

Figure 7: Auditability

For example, in the context of discovery within the financial accounting frame of reference auditors try to identify and to find misrepresented assets, liabilities and net worth or misrepresented revenues, expenses and profit or loss within the "small" data bases. What is contained in the "big data base" is still not officially known and can therefore not be used to formulate hypotheses to be tested independently.

The context of justification allows us to examine the semiotic (syntactic, semantic and pragmatic) foundations of the small data and to re-compute the accounting information trying to falsify it or to corroborate it when we cannot falsify it. (see Karl Popper 1965, 1968). When there is no frame of reference for the big data; no standard filter, model or algorithm to compute any expected results; and no legal framework to authorize audits by critical thinkers, then no independent opinion can be sought and justified.

In contrast to this depiction, the "bible of auditing" defines the nature of auditing by conflating the context of discovery and the context of justification as follows:

> Auditing is analytical, not constructive; it is critical, investigative, concerned with the basis for accounting measurements and assertions. Auditing emphasizes proof, the support of statements and data. Thus auditing has its principal roots, not in accounting which it reviews, but in logic on which it leans heavily for ideas and methods. (Mautz and Sharaf, 1961: 14)

Whereas the financial accounting context is part of a relatively well-defined frame of reference within which auditors can discover violations of the truth and justify their rational assessment of the truth, the big data frame of reference remains a myth because neither the context of discovery nor the context of justification can be (or has been) defined. Moreover, the kind of truth to be pursued – if at all - remains unclear.

Besides the *redundant theory of truth* which acknowledges merely different perspectives on reality, there exist at least five different theories of truth (as convincing beliefs according to logical rules). They ought to be known to auditors in order to make them cognizant of the dimensions and kinds of their intellectual contributions both in the financial accounting context and beyond: The *coherence theory*, the *correspondence theory*, the *semantic conception*; the *pragmatic conception,* and the *collective theory of truth*. Each of these are epistemologically explained and their insights can be clearly expressed and demonstrated:

> [The coherence theory of truth] defines „true" as a property of sets of beliefs that are mutually reinforcing (or „hang together") while satisfying conditions of logical consistency (where it is not the case that, for any belief *b* , both *b* and its negation, not-*b*, are accepted at the same time) and of deductive closure (where, if the truth of belief *b1* logically requires the truth of belief *b2*, then *b2* must also be accepted whenever *b1* is accepted). Since one person at two different times and two persons at the same time are entitled to completely different beliefs as long as their beliefs are coherent, the coherence theory does not entail the correspondence theory. [Fetzer and Almeder, 1993: 134]

What is coherent from a data collector's point of view (such as that of a bookkeeper or a data administrator) and that of an information provider's perspective (such as an accountant's or a programmer's) does not have to be „automatically so" from a user's point of view. Whereas the coherence theory describes „subjective truths" the correspondence theory is oriented on „objective facts:"

> [The correspondence theory of truth] defines „true" as designating the property of a declarative sentence when what it asserts to be the case is the case. Such a sentence („John is a bachelor") is true when the world (or **reality**) is the way it is thereby described as being or when that sentence „corresponds" to the world (because, in this case, John is a bachelor). The semantic theory of truth is a refinement of the correspondence theory. [Fetzer and Almeder,

1993: 135]

Since we are dealing with digital data (small or big) that are described not only by means of various bit and byte conventions for graphic, linguistic, numeric and voice expressions, but also by means of different programming languages with various conventions for data descriptions, the semantic conception of the truth is also relevant here:

> [The semantic conception of truth] maintains that truth ought to be interpreted as a metalinguistic predicate in order to avoid various semantic paradoxes (such as the sentence that asserts of itself, „This sentence is false," which is true if it is false and false if it is true)). Truth is viewed as a predicate that occurs in a **metalanguage** to describe sentences that occur in an **object-language**. Truth ascriptions are relative to a language and require adequate translations in the language in which they are expressed. The sentence, 'Schnee ist weiss' is true in German if and only if snow is white," thus specifies necessary and sufficient conditions of truth for the sentence „Schnee ist weiss" in German provided that it is properly tranlated within the meta-language of English by the sentence „Snow is white." [Fetzer and Almeder 1993: 136].

If we want to test expressions in the object language of accounting (*accountese*)such as „profit is $1 million", then we must translate it correctly in to a meta-language such as *ACL (Audit Command Language* – www.acl.com*)* and can then express the truth (or falsity) of the *accountese* statement by the *ACL* expression „profit is (not) $1 million." - The meta-language has to be syntactically and semantically at least as powerful as the object language. In other words, an audit language must be able to represent all small and big data understandably to allow an auditor to discover the truth of object language expressions and statements.

If the truth depends on expressions referring to convictions that the available evidence is sufficient to justify the conclusions or to accept them as false or true then we are dealing with the pragmatic truth theory:

> [The pragmatic theory of truth] defines „true" as designating the property a declarative sentence has when its assertion (or acceptance) is fully warranted. This requires that the available evidence is sufficient to justify its assertion (or acceptance). Yet it differs from the correspondence theory insofar as sentences whose assertion is fully warranted might not describe (or „correspond to") the world. (Fetzer and Almeder, 1993: 135-136]

To believe uncritically in the dogma „big data provides sufficient evidence" eliminates any audit requirements; however, professional auditors are not normally „believers" of this kind and ought to speak up when critical thinking is (to be) replaced by dogma, ideology or myth!

If we are dealing with more than one data or information user incl. auditors, then the truth may depend on the convictions of more than one member of a „guild." A variation of the pragmatic theory of truth is the „collective theory of truth" by Charles S. Peirce:

> The Peirsean theory of truth] defines „true" as a property of those beliefs that the community of inquirers is ultimately destined to accept or to agree upon in the long run (that is the opinion that they will share in common as a result of applying scientific methods to answerable questions concerning the world forever). Alternatively, it is the opinon that they *would* share if they *were* to apply scientific methods to answerable questions concerning the world forever. In either formulation, those beliefs are thought to „correspond" to the world.

> Strictly speaking, truth in Peirce's sense does not guarantee correspondence, in the meanwhile, rational beliefs are those whose acceptance is suitably warranted by the available relevant evidence. [Fetzer and Almeder, 1993: 135]

Truth implies evidence of the circumstances and the context of knowledge. When the frame of reference is not known to the providers nor to the recipients of the information; when the data elements and their description in a computer language are hidden; and when the filter, model or algorithm applied to the data are unspecified, then IT personnel arrogates to itself the monopolistic use of computer technology under the guise of the big data mystique and algorithmic omnipotence. Nobody seems to be able to question them critically, intelligently and seriously. - Truth seems to be redundant in the digital big-data-age!?

If data is collected by monitoring the users of IT with reference to their economic value as marketing or advertising "fodder" after their sale to, and further manipulation by, data brokers, then the IT users are "on display" in more than one sense in the "big-data-zoo." The majority does not yet seem, or want, to know why they are "observed" or why they are "bombarded" with seemingly relevant advertising whenever they open their computer… only to be observed further.

The prime criterion for collecting information about users of technology (social media, etc.) and for claiming all rights to the information forever and irrevocably is evidently to sell these data for a profit to data brokers. They can manipulate them further and sell them for another profit to advertisers and marketers who try to influence the behavior of the data subjects. Thus, the social contract between IT suppliers and IT users is completely one-sided and fully in favor of those in control of IT. Instead, their responsibility could and should be the protection of the privacy, property, and security of their users by providing relevant, protected and auditable data and software.

Conclusions

Big data are not auditable as long as there is no ethical and legal basis for providing original small data (incl. meta-data) and more informative big data (incl. meta-big-data) to enlightened and autonomous users:

> We allowed ourselves to be monetized and productized on the cheap, giving away billions of dollars of our personal data to new classes of elite who saw an opportunity and seized it. We accepted all their one-sided ToS (Terms of Service) without ever reading them, and they maximized their profits, unencumbered by regulation or oversight. To be sure, we got some pretty cool products out of the deal… [b]ut now that we've given all these data away, we find ourselves at the mercy of powerful data behemoths with near-government-level power who do as they please with our information and our lives. (Goodman, 2015: 79)

Sharing data naively ("nothing to hide") and freely for mere IT convenience means sacrificing one's individuality, privacy and security in favor of exploitation by technologists and those in absolute and possibly criminal control of data and their use to the possible detriment of "transparent" data subjects.

Are we willing to give up the fruits of centuries of philosophical struggling for and with enlightenment issues to those collecting data about our behavior and reaping huge economic, organizational, personal and social benefits in a modern capitalistic

manner? The issues are more profound than monopolistic tendencies of search engine and social network providers! (EU vs. Google). To hope for improved objectivity, transparency and trust with ill-defined big data is also an illusion (Will 2015). – We need a major societal effort to establish a rational data governance framework as part of a new social contract between data and information providers and their users in the digital age in analogy to the accountability and auditability framework depicted in Figure 8. (Will, 2015). – Auditors need to think critically, speak up fearlessly, and be heard clearly since so much is at stake in the context of the big data ideology!

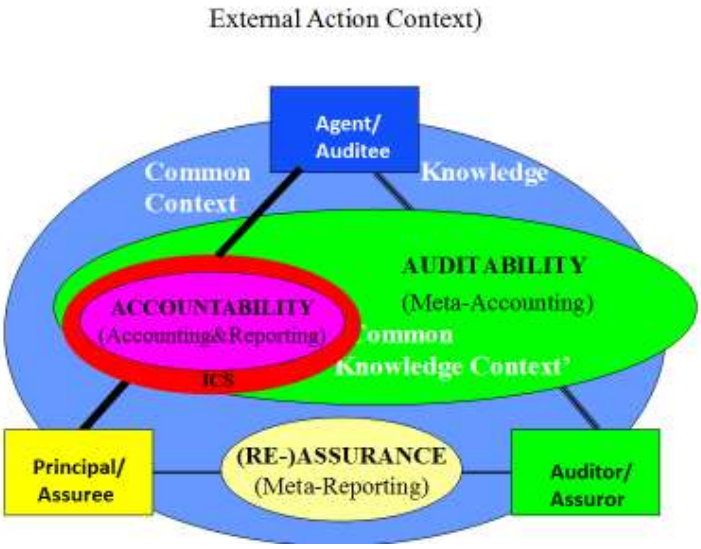

Figure 8: Accountability and Auditability Framework

References

[1] "Big Data and other definitions." *Wikipedia.* Wikimedia Foundation 2015.

[2] Eells, Ellery, and James H. Fetzer. (2010). *The Place of Probability in Science: In Honor of Ellery Eells (1953-2006).* Dordrecht: Springer, 2010.

[3] Fetzer, James H. (1990). *Artificial Intelligence: Its Scope and Limits*. Dordrecht: Kluwer Academic, 1990.

[4] Fetzer, James H. (1993). *Philosophy of Science*. New York: Paragon House, 1993.

[5] Fetzer, James H. (2000). "Information and Representation." *Proceedings World Multiconference on Systems, Cybernetics and Informatics:* Orlando, Florida, USA, July 23-26, 2000: Volume X: pp. 472-477.

[6] Fetzer, James H. (2005). *The Evolution of Intelligence: Are Humans the Only Animals with Minds?* Chicago: Open Court, 2005.

[7] Fetzer, James H. and Robert F. Almeder (1993). *Glossary of Epistemology/Philosophy of Science.* New York: Paragon House, 1993.

[8] Goodman, Marc (2015). *Future Crimes: Everything Is Connected, Everyone Is Vulnerable and What We Can Do about It.* Canada: Doubleday, 2015.

[9] Hua, Jing-Shiuan (2013). "An Innovative Framework of Data Governance." (Ph.D. Thesis, Institute of information Management, National Chung Cheng University, Taiwan).

[10] Knuth, Donald Ervin. (1973). *The Art of Computer Programming.* Reading, MA: Addison-Wesley Pub., 1973.

[11] Mautz, R.K., and Hussein A. Sharaf. (1961). *The Philosophy of Auditing*. Madison, Wis.: American Accounting Association, 1961.

[12] Popper, K. R. (1965). *The Logic of Scientific Discovery*. New York: Harper and Row, 1965.

[13] Popper, K.R. (1968). *Conjectures and Refutations*. New York: Harper and Row, 1968.

[14] Schmidt, Eric, and Jared Cohen (2013). *The New Digital Age: Transforming Nations, Businesses, and Our Lives.* London: John Murray Publishers, 2013.

[15] Will, Hart J. (2000). "Auditability and Controllership: Extracting Knowledge from Accounting Information" in: *Proceedings World Multiconference on Systems, Cybernetics and Informatics.* Orlando, Florida, USA, July 23-26, 2000: Volume X: pp. 483-488.

[16] Will, Hart and Darren Whobrey (2003). "The Assurance Paradigm & Organizational Semiotics: A New Applications Domain" in: *Proceedings 6th International Workshop on Organizational Semiotics (IWOS 2003)*. July 12-13, 2003. Reading, UK: Reading University.

[17] Will, H. (2006). "Knowledge management and administration depend on semiotic information systems." *International Journal of Management and Decision Making* 7(1), pp. 36-57.

[18] Will, Hartmut J. (2015). "Erhoehte Objektivitaet, Transparenz und Vertrauen mit Big Data?! (Increased Objectivity, Transparency and Trust with Big Data?!)" in: Deggendorfer Forum zur digitalen Datenanalyse e.V. [editor]: *GoBD and Big Data*, Berlin: Erich Schmidt Verlag, pp. 11-56.