J Syst Sci Syst Eng (Jun 2013) 22(2): 127-151
DOI: 10.1007/s11518-013-5219-4
ISSN: 1004-3756 (Paper) 1861-9576 (Online)
CN11-2983/N

BIG DATA: UNLEASHING INFORMATION

James M. TIEN

College of Engineering, University of Miami, Coral Gables, Florida 33146, USA

jmtien@miami.edu (✉)

Abstract

At present, it is projected that about 4 zettabytes (or 10**21 bytes) of digital data are being generated per year by everything from underground physics experiments to retail transactions to security cameras to global positioning systems. In the U. S., major research programs are being funded to deal with big data in all five sectors (i.e., services, manufacturing, construction, agriculture and mining) of the economy. Big Data is a term applied to data sets whose size is beyond the ability of available tools to undertake their acquisition, access, analytics and/or application in a reasonable amount of time. Whereas Tien (2003) forewarned about the data rich, information poor (DRIP) problems that have been pervasive since the advent of large-scale data collections or warehouses, the DRIP conundrum has been somewhat mitigated by the Big Data approach which has unleashed information in a manner that can support informed – yet, not necessarily defensible or valid – decisions or choices. Thus, by somewhat overcoming data quality issues with data quantity, data access restrictions with on-demand cloud computing, causative analysis with correlative data analytics, and model-driven with evidence-driven applications, appropriate actions can be undertaken with the obtained information. New acquisition, access, analytics and application technologies are being developed to further Big Data as it is being employed to help resolve the 14 grand challenges (identified by the National Academy of Engineering in 2008), underpin the 10 breakthrough technologies (compiled by the Massachusetts Institute of Technology in 2013) and support the Third Industrial Revolution of mass customization.

Keywords: Big data, data acquisition, data access, data analytics, data application, decision informatics, products, processes, adaptive services, digital manufacturing, mass customization, third industrial revolution

1. Introduction

While the focus of this paper is on Big Data, it should be emphasized that data – especially big data – is worthless from a decision making perspective unless it is analyzed or processed to yield critical information which can then be employed to make informed decisions in a range of areas, including business, science, engineering, defense, education, healthcare and society-at-large. Big Data is poised to add greater value to businesses (which can fathom their transactional data to detect patterns

suggesting the effectiveness of their pricing, marketing and supply chain strategies), to understand planet earth (which is being extensively monitored on the ground, in the air and in the water), to solve science and engineering problems (which have become more data- or empirically-driven), to support modern medicine (which is collecting and mining large amounts of image scans and genetic data), to enhance the World Wide Web (which is amassing terabytes of textual and visual material that is becoming widely available through search engines like Google and Baidu), and to aid intelligence agencies (which are collecting and mining satellite and thermal imagery, audio intercepts and other readily available information, including that found on the World Wide Web).

It is helpful to first define data, which, according to the current version of Wikipedia, "are values of qualitative or quantitative variables, belonging to a set of items". Data are typically the results of measurements and are considered to be raw before they are processed; in fact, the processed data from one stage may be considered to be the raw data to the next stage. Metadata, sometimes referred to as "data about data", describes the content and context of a set or file of data; for example, a photo file's metadata would identify the photographer, the camera settings, the date taken, etc. For the purpose of this paper, our definition of data would include measurements, raw values, processed values, and metavalues. More specifically, the focus of this paper is on digital data, which basic unit of measurement is a bit, an abbreviation for a binary digit that can be stored in a device having two possible distinct values or levels (say, 0 and 1). A byte is a basic unit of information containing 8 bits, which can include 2 raised to the power of 8 or 256 values (say, 0 to 255). Table 1 details the range of digital data sizes, from kilobytes, to megabytes, to gigabytes, to terabytes, to petabytes, to exabytes, to zettabytes, and to yottabytes. Clearly, all this assortment of data can add up and grow very quickly in size; as indicated in Figure 1, the International Data Corporation estimates that, on a world-wide basis, the total amount of digital data created and replicated each year will grow exponentially from 1 zettabyte (or 10**21 bytes) in 2010 to 35 zettabytes in 2020! Thus, in the current year of 2013, it is projected that about 4 zettabytes of digital data are being generated by everything from underground physics experiments to retail

Table 1 Digital data sizes

VALUE	ABBREVIATION	APPELLATION
1000**1	KB	Kilobytes
1000**2	MB	Megabytes
1000**3	GB	Gigabytes
1000**4	TB	Terabytes
1000**5	PB	Petabytes
1000**6	EB	Exabytes
1000**7	ZB	Zettabytes
1000**8	YB	Yottabytes

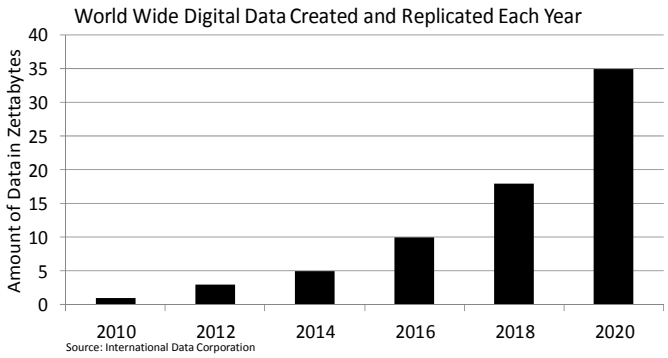

Figure 1 Growth in digital data

transactions to security cameras to global positioning systems.

Mayer-Schonberger & Cukier (2013) credit computer scientist Oren Etzioni with having the foresight to see the world as a series of big data problems, before the concept of Big Data was introduced. In 1994, he helped build one of the Web's first search engines, MetaCrawler, which was sold to InfoSpace; he then co-founded Netbot, the first major Web shopping site, which was sold to Excite; another startup for abstracting meaning from text documents, named ClearForest, was eventually bought by Reuters; and Farecast, a predictive model for airline and other fares, was sold to Microsoft's Bing in 2008 for $110 million. Indeed, Mayer-Schonberger & Cukier (2013) provide not only an historical perspective on the birth and evolution of Big Data but also a persuasive argument about the importance of Big Data, which can offer us insights about "what" is in the data, not necessarily about the "whys" behind the insights; in short, they consider Big Data to be a revolution that will transform how we live, work and think.

Big Data, according to the current version of Wikipedia, "is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time". Obviously, the definition of what constitutes Big Data is shifting as software tools become more powerful; today, depending on the nature and mix of the data, a data set is considered big if it contains a few terabytes to many petabytes of data. Whereas Tien (2003) forewarned about the data rich, information poor (DRIP) problems that have been pervasive since the advent of large-scale data collections or warehouses, the DRIP conundrum has been somewhat mitigated by the Big Data approach which has unleashed information in a manner that can support informed – yet, not necessarily defensible or valid – decisions or choices. The focus, emphasis and scope of Big Data – and its growing impact – are discussed herein. Section 2 considers Big Data from the decision making and data processing perspectives; Section 3 details the Big Data components of acquisition, access, analytics and application; and Section 4 concludes with several critical observations in regard to Big Data, both its benefits and concerns. Finally, it should be noted that inasmuch as Big Data impacts the critical and

related areas of decisions, risk, informatics, services, goods, and customization or personalization, the contents herein reference earlier papers by Tien (2003, 2008, 2011, 2012), Tien & McClure (1986), Tien & Berg (1995, 2003) and Tien et al. (2004) on the same overlapping areas of concern.

2. Considerations

At the beginning of the 21st Century, the growing volumes of data presented a seemingly insoluble problem; storage and CPU (central processing unit) technologies were overwhelmed by the terabytes of data being generated. Fortunately, Moore's law came to the rescue and helped to make storage and CPUs larger, faster, smarter and cheaper. Today, Big Data is no longer a technical problem: it is has become a competitive advantage. As indicated earlier, enterprises are developing and employing Big Data tools for exploring their data trough, to discover new insights that could help them to develop better relationships with their customers, to identify new areas of business opportunities, and to better manage their supply chains, all in an increasingly competitive business environment. In short, Big Data can serve to improve services, products and processes, through its impact on decision making and data processing.

2.1 About Decision Making

As alluded to earlier, the purpose of Big Data is about decision making. More specifically, the overarching reason for undertaking data analysis is to obtain or derive information from data, knowledge from information, and wisdom from knowledge. This sequence of derivations is illustrated in Figure 2. Although the literature does not distinguish between data and information, we attempt to do so in this paper, to the extent practical. In fact, if we were to strictly adhere to such a distinction, we would conclude that given the current state of information technology, it should be referred to as more data – than information – "technology"; however, this would lead to an unnecessary level of confusion, especially in regard to Big Data which is generally considered to be information (i.e., from which informed decisions can be made).

There are actually two approaches to the analysis of data: the scientific and the engineering approach. The scientific focus is on investigating natural phenomena, acquiring new knowledge, correcting and integrating previous knowledge, or understanding the laws of nature; it includes observations, measurements, and experiments, and the formulation, testing, and modification of hypotheses. On the other hand, the engineering approach – which Tien (2003) calls decision informatics – is very purposeful; that is, as per Figure 3, the nature of the required real-time decision (regarding the production and/or delivery of, say, a service or a manufactured good) determines, where appropriate and from a systems engineering perspective, the data to be collected (possibly, from multiple, non-homogeneous sources) and the real-time fusion/analysis to be undertaken to obtain the needed information for input to the modeling effort which, in turn, provides the knowledge to support the required decision in a timely manner.

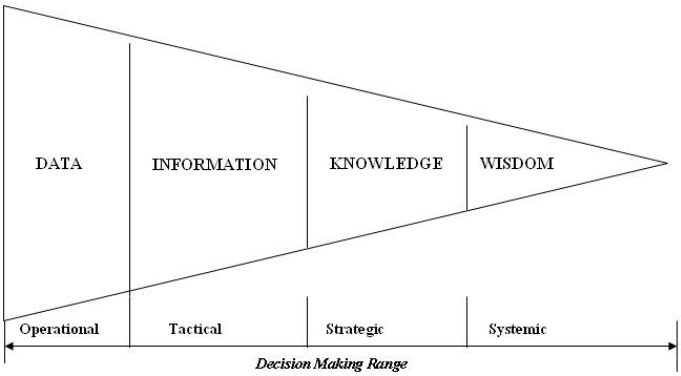

Figure 2 Decision making framework

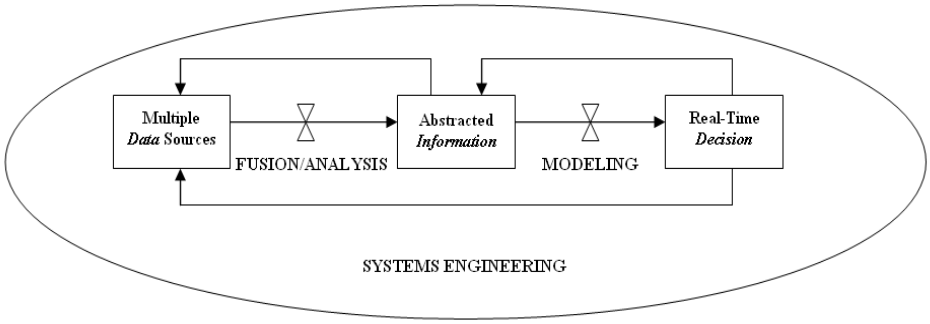

Figure 3 Decision informatics: an engineering approach

The feedback loops in Figure 3 are within the context of systems engineering; they serve to refine the analysis and modeling steps. Thus, decision informatics is supported by two sets of technologies (i.e., information and decision technologies) and underpinned by three disciplines: data fusion/analysis, decision modeling and systems engineering. Data fusion and analysis include the fusing and mining of data, information and knowledge, by employing such methods as probability, statistics, quality, reliability, fuzzy logic, multivariable testing, pattern analysis, etc. On the other hand, real-time data fusion and analysis is more complex and require additional research, especially in the era of Big Data. Decision modeling methods include discrete simulation, finite element analysis, stochastic techniques, neural networks, genetic algorithms, optimization, etc. Further, real-time decision modeling is not just concerned with speeding up the models and solution algorithms; it, like real-time data fusion and analysis, also requires additional research, especially since most steady state models become irrelevant in a real-time Big Data environment. Systems engineering is about integrating products, processes and operations from a holistic perspective, especially human-centered systems that are computationally -intensive and intelligence-oriented. A critical aspect of systems engineering is system performance; it provides an essential framework

for assessing the decisions made – in terms of such issues as satisfaction, convenience, privacy, security, equity, quality, productivity, safety, reliability, etc. Similarly, undertaking systems engineering within a real-time environment will require additional thought and research.

2.2 About Data Processing

Digital data are being generated by many different sources (media, entertainment, healthcare, life sciences, surveillance, transportation, logistics, telecommunication, education, etc.) and for different purposes (tracking, communication, location, investment, learning, etc.). Almost three decades ago, Tien & McClure (1986) identified several data issues from an operational, tactical, strategic and policy perspective. It is instructive to see how Big Data is addressing these issues; as summarized in Table 2 and while the lack of data quality remains a problem, most of the other issues are being mitigated – through the breakthroughs in storage and CPU technologies. In regard to data quality (i.e., accuracy, completeness, consistency, currency, ambiguity, etc.), it is obvious that imperfect data can negatively impact the decision making process, leading to inferences, interpolations or conclusions that are not trustworthy and, of course, unacceptable in critical, life-and-death situations (e.g., defense, air travel, automobiles, healthcare, etc.). Depending on the application and the quality of the data, a number of smart, data-driven analytical methods are being employed to develop and/or combine the data over both space and time, including such methods as Bayes inference of subjective probabilities, Dempster-Shafer theory of plausible belief functions, Black-Scholes formula for option pricing, etc. – from which extraction, estimation and fusion approaches can be further developed.

Before addressing the Big Data components in Section 3, it is helpful to compare the Big Data approach with the traditional data processing approach; as summarized in Table 3, there are major differences between the two approaches. In particular, in contrast to the traditional, mostly statistical, approach, Big Data seeks to unleash information in a manner that can support informed decisions – by somewhat overcoming data quality issues with data quantity, data access restrictions with on-demand cloud computing, causative analysis with correlative data analytics, and model-driven with evidence-driven applications. This somewhat expedient – but not necessarily valid – approach can result in a corresponding set of potential pathologies or concerns, which are further discussed in Section 4. On the other hand, the feasibility – or "good enough" – focus of Big Data is usually more realistic than the optimality focus of traditional, operations research methods; in fact, the steady-state assumption that underpins optimality is, for the most part, unrealistic, especially in real-time environments where values are changing and agent-negotiated solutions are indeed messy and at best only feasible. Nevertheless, no matter what the purpose is for the processing or analysis of the data, it is critical that the data contain the insights or answers being sought; otherwise, we have a case of garbage-in, garbage-out. As a consequence, the aforementioned metadata can play an important role in ascertaining the scope, validity and viability of the data.

3. Components

In general, and as identified in Figure 4, there are four steps or components to Big Data processing: 1) acquisition (including data capture); 2) access (including data indexing, storage, sharing and archiving); 3) analytics (including data analysis and manipulation); and 4) application (including data publication). Although the four steps can be considered to be somewhat sequential in any Big Data application, there is in essence 4-combinatorial-2 or six possible interactions or feedback loops among and between the four components, as illustrated in Figure 4.

3.1 Acquisition

Advances in digital sensors, communications, computation and storage have yielded zettabytes of data. Table 4 contains a range of possible digital data sources, including customer order transactions, emails and their attachments, radio frequency identification (RFID) sensors, smart phones, films, video recordings, audio recordings, and genetic sequences. In terms of data acquisition methods, Table 5 includes four example groups of methods, including those

Table 2 Data issues: big data considerations

SYSTEM FOCUS	ISSUES: TIEN & MCCLURE (1986)	2013 STATUS	BIG DATA CONSIDERATIONS
Operational	Lack of data quality (accuracy, completeness, consistency, currency, ambiguity, etc.)	Still Problematic	Mitigated by larger *data acquisition*, including proxy metrics
Tactical	Lack of data processing (timely access, storage capacity, data-user interface, scalability, etc.)	Mostly Overcome	Increasingly more powerful *data access* technologies
Strategic	Lack of decision-support tools (modeling, formulation, monitoring, etc.)	Much Improved	Increasingly more sophisticated *data analytics*
Policy	Lack of policy-support tools (modeling, formulation, monitoring, etc.)	Much Improved	Increasingly more integrated *data application*

Table 3 Data processing approaches: traditional versus Big Data

COMPONENTS	ELEMENTS	TRADITIONAL APPROACH	BIG DATA APPROACH
Acquisition	Focus	Problem-Oriented	Data-Oriented
	Emphasis	Data Quality	Data Quantity
	Scope	Representative Sample	Complete Sample
Access	Focus	On-Supply, Local-Computing	On-Demand, Cloud-Computing
	Emphasis	Over-Time Accessibility	Real-Time Accessibility
	Scope	Personal-Security	Cyber-Security
Analytics	Focus	Analytical Elegance	Empirical Messiness
	Emphasis	Causative Relationship	Correlative Relationship
	Scope	Data-Rich, Information-Poor (DRIP)	Data-Rich, Information-Unleashed (DRIU)
Application	Focus	Steady-State Optimality	Real-Time Feasibility
	Emphasis	Model-Driven	Evidence-Driven
	Scope	Objective Findings	Subjective Findings

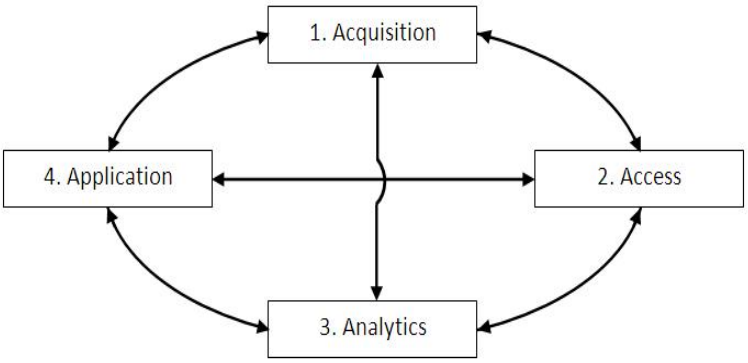

Figure 4 Big Data components

Table 4 Sources of digital data

SOURCES	METRICS	COMPANIES
Transactions	Customer Orders	Walmart
Emails	10-25 MB Attachment Allowed	Google's Gmail
Sensors	Radio Frequency Identification (RFID)	FedEx
Smart Phones	3G, 4G, GPS, Etc.	Apple's iPhone
Films	1-2 GB	Walt Disney Pictures
Video Recordings	Aspect Ratios: 4:3, 16:9	Microsoft's Bing
Audio Recordings	200 Hours = 640MB	LibriVox
Genetic Sequences	3.2B DNA Base Pairs in A Human	Life Technologies

Table 5 Big Data scope: data acquisition

SCOPE	EXAMPLE ACQUISITIONS	EXAMPLE EFFORTS
Data Capture	Keystroke Logger; Clickstream; Smart Sensors; Health Monitors; Drone Sensors; Samples	Monitoring Software; Website Trackers; Smart Phone Apps; RFID; Ornithopters; Memoto; Compressed Samples
Multisensory Data	Visual Detection; Video Cameras; Light-Field Photography; Beyond Video and Audio Telepresence	Thermal Imager; Bug's Eye; Lytro; Internet Transmission of Touch, Smell & Taste Senses
Brain Imaging	Magnetic Response Imaging; Functional MRI (fMRI); Diffusion MRI (dMRI)	U.S.'s Human Connectome Project ($40M); E.U.'s Human Brain Project (Euro 1B); U.S.'s BRAIN Initiative ($100M)
Real-Time Sensing	Real-Time Location Data; Real-Time Image Display; Real-Time Response	Smart Phone-Based, Global Positioning System (GPS); Motion & Image Sensors; OLED TV; Ocean Observatories; Smart Grids; Smart Cities

dealing with data capture, multisensory data, brain imaging, and real-time sensing.

In regard to data capture, the methods include such stealth approaches as keystroke loggers and clickstreams (both of which can provide real-time insights into consumer behavior); smart and RFID sensors (which are becoming ubiquitous in devices, products,

buildings, and even cities); health monitors (for both humans and animals in the monitoring of their body temperature, blood pressure, etc.); drone sensors (including wing flapping ornithopters and stamp-sized Memoto cameras); and samples (from metamaterial that directly compresses a scene's image, thus obviating the need for post processing).

In regard to multisensory data, one can obtain terabytes of data from video surveillance cameras (Chongqing, China, for example, has over half a million cameras in public spaces), thermal imaging (which was responsible for recently capturing the heat signature of Dzhokhar Tsarnaev, one of the April 15, 2013, Boston Marathon bombers), a bug's eye view (which has a dome of 180 microlenses that can capture a 160-degree field of vision), light-field photography or Lytro (which is a multi-lensed camera that can capture all incoming light, including color, intensity, angle, etc., in a 3D pattern which can be endlessly refocused using sophisticated software), and the eventual development of Internet-transmittable senses of touch (haptic), smell (olfactory) and taste.

In regard to brain imaging, U.S. and E.U. have recently provided several large government grants, including the U.S.'s Human Connectome Project at $40M, the E.U.'s Human Brain Project at 1B Euros, and the U.S.'s BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative at $100M for the first year; such efforts may in time make artificial intelligence – whereby an electronic machine is endowed with feelings and consciousness – a reality.

In regard to real-time sensing, the relevant technologies range from global positioning systems which are becoming ever more pervasive (through smart phones) and precise (through powerful satellites), to image display which will become increasingly dominated by organic light-emitting diode (OLED) displays that allow for faster-moving images and are bigger, brighter and thinner than the current liquid crystal displays (LCDs), to ocean observatories which are gathering temperature, pressure and salinity data at depths of 2,000 meters, to smart grids which can capture the behaviors of suppliers and consumers in order to improve the efficiency, reliability, economics, and sustainability of the electrical production and distribution system, and to smart cities like Santander, Spain, which have thousands of electronic sensors that can monitor traffic, parking, electrical consumption, water consumption, waste management, etc., all to increase the efficiency and reduce the stress of urban life.

Four remarks should be made in connection with the astounding growth in Big Data acquisition. First, in order to become truly smart in, say, a smart city sense, all the sensors must be connected or electronically fused on to a common platform that can streamline both the data gathering and the resultant analyses. Second, inasmuch as all sensors must communicate – and they do, mostly in a wireless manner – the question remains: what is the potential health effect from long-term exposure to radio frequency (RF) energy emitted by these sensors? There is, at this time, no long-term health study that can provide a definitive answer. Third, the speed of data acquisition is accelerating; for example, when the human genome was first being decoded in 2003, it

required almost a decade to sequence one person's 3.2 billion base pairs – today, a single facility can sequence an individual's complete genome in a day! Fourth, both the acquisition and subsequent use of personal data raise a number of privacy issues, from misuse to abuse; yet, these same data can save lives or at least help to make lives better, if not safer – clearly, there are trade-offs to be considered.

3.2 Access

Table 6 summarizes Big Data's access performance in terms of data service, data management, platform management and cloud computing. In regard to data service, platform as a service (PaaS) consists of a computing platform and a solution stack as a service; along with software as a service (SaaS) and infrastructure as a service (IaaS), it is a service that is now typically associated with cloud computing. In this service schema, the consumer creates the software using tools and/or libraries from the provider and is also able to control software deployment and configuration settings. The provider – including such powerhouses as Google, VMware, Amazon, Microsoft, HP and Oracle – provides the networks, servers, storage and related services. Thus, as an example, Netflix employs Amazon to stream its on-demand videos and movies to their over 35 million subscribers in North and South America, the Caribbean, and several countries in Europe; indeed, on a typical weeknight, Netfix movies account for about one-third of all downstream internet traffic in North America.

In regard to data management (including indexing, warehousing, searching and navigating), images are especially hard to index; a chicken-and-egg problem develops when there are insufficient images to train an algorithmic classifier – as a result, a learning approach must be employed to effect such training. Adobe provides an integrated data warehousing approach with its digital management schema. Microsoft Office 365, as a cloud server, allows access to the Office suite from anywhere, anytime. As Furtado (2009) details, there are a number of reasons for selecting a parallel or a distributed warehouse, depending on data size, query processing speed and organizational structure. An increasing number of the data warehouses are being located in the cloud

Table 6 Big Data scope: data access

SCOPE	EXAMPLE ACCESSES	EXAMPLE EFFORTS
Data Service	Platform As A Service (PaaS); Software As A Service (SaaS); Infrastructure As A Service (IaaS)	Google, VMware; Amazon; Microsoft; Google; Globus Online; Amazon; HP; Oracle
Data Management	Data & Image Indexing; Enterprise Data Warehouses; Database Search & Navigation	Microsoft's Bing; Adobe; SAS, Microsoft Office 365; VMware Inc; Visualization; SAP; Splunk
Platform Management	Accessibility; Scalability; Quantum Computing; Security	Google Fiber (Kansas City, Austin); Petascale to Exascale Computer; D-Wave; State-Backed Hackers
Cloud Computing	Private Clouds; Public Clouds; Hybrid Clouds	Cloudcor; NEC; Google; OpenStack; Amazon; Rackspace

(Schadt et al. 2010), where there is the potential of accessing an unlimited amount of processing and storage power that can be reconfigured on an on-demand basis and which cost can be charged on a pay-as-you-use basis, much like the use of electricity. Splunk, for example, is a 10-year old software company focused on capturing, indexing and correlating real-time data in a searchable repository from which it can generate graphs, reports, alerts, dashboards and visualizations.

In regard to platform management (including accessibility, scalability and security), it is obvious that bandwidth is a critical issue; Google has just selected Austin, Texas, as its second city (after Kansas City) to be wired to receive both downloads and uploads at a gigabyte per second, fast enough to transmit a high-definition, feature-length film in a second or two – AT&T declared that it too is planning to bring the same broadband service to Austin. The benefit-cost impact of such an investment is yet to be determined. Even the successful social networks – including Facebook, MySpace, Twitter, Ning, LinkeIn and Zanga – are still developing an effective and viable business model based on attracting advertising dollars through targeted ads; nevertheless, as their subscribers and Big Data problems multiply, they are increasingly migrating to the cloud. The Oak Ridge National Laboratory's Titan is today's fastest computer at 18 petaflops (10**15 floating point operations per second); however, scientists and engineers are clamoring for an exaflop machine that could better model climate changes, cryptography, quantum mechanics, nanodevices and advanced biofuel engines. The exascale machine will probably employ quantum information processing (Stern & Lindner 2013); it must have a scalable architecture and error correction that can be performed in parallel with computation – D-Wave's quantum computer is especially adept at solving discrete optimization problems. On the other hand, SAP is developing faster database software to deal with the growing storage and processing needs. Big Data security is obviously a matter of grave concern; Shostack (2012, p.8) appropriately states that "information security is the assurance and reality that information systems can operate as intended in a hostile environment". Thus, research is ongoing to build smarter storage systems that can adaptively reconfigure as they learn from how they are being used and make run-time changes to better meet goals for dependability, availability, performance, privacy, and security. A related issue is about benchmarking Big Data systems; Baru et al. (2013) are beginning to lay the groundwork.

In regard to cloud computing in private, public or hybrid clouds, it should be noted that large or big data sets are becoming more the norm as scientists try to gain insights into global warming, meteorology, neurology, genomics and nanotechnology, and as engineers and other decision makers seek answers to the development of new drugs, novel materials, more efficient supply chains and more effective sensors. As noted earlier, the big data files are increasingly being archived in the cloud. For example, the 1996 established Internet Archive, a 501(c)(3) non-profit, resides in the cloud; it is building a digital library of Internet sites and other cultural artifacts in digital form, and it is providing free access to researchers, historians,

scholars, and the general public. Cloud computing – provided by such companies as Cloudcor, NEC, Google, OpenStack, Amazon and Rackspace – is growing in size as technical and security issues are being resolved and enterprises become dependent on the cloud for their growth, system efficiencies and new products and processes; it is projected that by 2015, more than 2.5 billion users and 15 billion devices will be accessing cloud services.

Three remarks should be made in connection with the astounding growth in Big Data access. First, big media firms are worried that broadband access may cause greater video piracy, as was the case in South Korea where the home entertainment industry was decimated by digital piracy, supposedly enabled by the widely available high-speed Internet – obviously, piracy must be prevented, most likely by a technological solution that is yet to be developed. Second, there remains a policy question regarding cyber-security and whether the U. S. government is responsible for protecting commerce (especially financial businesses) from cyber-attacks, just as the U. S. military is responsible for defending the air and sea lanes from an invasion. Third, as with Big Data acquisition, Big Data access is also subject to the same privacy and confidentiality concerns.

3.3 Analytics

When analyzing Big Data, we must necessarily employ the computer, resulting in the term data analytics, which, according to the current version of Wikipedia, "is the application of computer technology, operational research, and statistics to solve problems in business and industry". Analytics is carried out within a computer-based information system, while in the past statistics and mathematics could be studied without computers and software. Of course, mathematics underpins the methods and algorithms employed in analytics and the science of analytics is concerned with extracting useful insights or properties of the data using computable functions. It should be explicitly noted that although the underlying mathematics may still be valid for Big Data, the software application may come undone if the dataset is too large (Jacobs 2009). Indeed, there are other scalability issues in connection with Big Data (Allen et al. 2012).

It should be noted that Big Data analytics in the business environment has been referred to by other names, including Business Intelligence (Luhn 1958), Decision Informatics (Tien 2003) and Business Analytics (Davenport & Harris 2007). In addition to a decision-driven focus that permeates business and engineering, Big Data analytics has also been employed to gain scientific insights concerning, say, the laws of nature, genomics and human behavior. Actually, the first time that computers or analytics were employed to prove a scientific theorem was when Appel & Haken (1977) proved the four color theorem using a special-purpose computer program to empirically confirm that no more than four colors are required to color the regions of a map so that no two adjacent regions have the same color. Initially, their proof was not accepted by the theoretical mathematicians because the computer-assisted proof was not able to be checked by hand (Swart 1980). Since then the proof has gained wider acceptance, although some doubt still remains (Wilson 2002).

With the advent of Big Data analytics, a number of niche analytics have been developed, including in retail sales, financial services, risk and credit, marketing, buying behavior, loan collections, fraud, pricing, telecommunications, supply chain, demand chain, transportation, and visualization. Early efforts at risk management were focused on risk avoidance; however, lost opportunities should also be factored into such efforts. To process Big Data within tolerable elapsed times, Manyika et al. (2011) suggest several appropriate technologies, including association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization. Additional technologies being applied to Big Data include massively parallel-processing databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems.

Table 7 identifies some categories of Big Data analytics, including correlational analysis, pattern recognition, evidence-driven, and analytics accreditation. In regard to correlational analysis, a number of companies offer their statistical, operations research and management science services in this domain, ranging from statistical algorithms that fill in for incomplete, inconsistent or inaccurate data sets; to data fusion for differentiated direct marketing strategies (van Hattum & Hoijtink 2008); and to simulation which Coca-Cola employs to combine a myriad of flavors and consumer preferences to arrive at a Black Book of new offerings. Lavalle et al. (2011) provide other examples of how to transition from possible correlations to plausible ones to probable ones to credible insights using Big Data analytics.

Table 7 Big Data scope: data analytics

SCOPE	EXAMPLE ANALYTICS	EXAMPLE EFFORTS
Correlational Analysis	Statistics; Visualization; Operations Research; Simulation; Management Science; Algorithms	Data Fusion; Data Cave; SAS; IBM; GE; VMware; Terradata; Amazon; Coca-Cola; Splunk; Twitter; Zynga; Risk Management
Pattern Recognition	Tracking; Disease Spread Topology; Simulation; Modeling Real-Time Search	ShopperTrak; Facebook's Timeline Google; Ayasdi's Software; Ansys' Simulator; Sparse Fourier Transform; IBM's Watson
Evidence-Driven	Marketing (Behavior, Attitude); Predicting (Savvy, Statistics); Technical Assistant; Answering Questions	Facebook's Graph Search; E-Marketing; Crowdsourcing; Pricing; Obama Elections; Apple's Siri; Google's MapReduce; Hadoop
Analytics Accreditation	PStat (Accredited Professional Statistician); CAP (Certified Analytics Professional); Niche Analytics	By ASA (American Statistical Association); By INFORMS (Institute for OR & MS); Practiced by IBM, SAS, Etc.

In regard to pattern recognition, ShopperTrak is not only about traffic counting but also tracks security breaches and inventory shrinkages; Facebook's Timeline is collecting and analyzing consumer data on an unprecedented scale, presumably to better target advertisements; by employing a range of predictive models which correlated 45 search terms with actual flu cases documented by the Centers for Disease Control (CDC) in 2007 and 2008, Google was able to predict and locate the 2009 H1N1flu spread on a near real-time basis; employing the mathematics of topology, Ayasdi seeks to draw novel patterns and relationships among data points; Ansys, like SolidWorks, provides software that can simulate products and processes in order to build better prototypes; in geophysics, sophisticated models of large data sets are simulating seismic activities in order to better understand earthquakes (Barbot et al. 2012, Segall 2012); in transportation, multiagent-based simulations are being employed to better understand driver behavior (Hattori et al. 2011); sparse Fourier transforms (SFTs) are allowing for more real-time analysis and decision support; and in 2011 IBM's Watson computer was able to beat two former Jeopardy! champions by first understanding the questions being asked (in natural language) and then sifting through terabytes of data to find the best answers. As a footnote, Watson is now being used to suggest cancer treatments.

In regard to evidence-driven analytics, Ahlquist & Saagar (2013) are integrating behavior and attitude data to gain insights into customers for, say, marketing segmentation and messaging purposes; Facebook is employing graph search and the earlier cited Timeline techniques to create a virtual feedback loop between Facebook users and advertisers; by dissecting behavioral data, e-marketers are arming websites with effective micro-marketing abilities; Silver (2012), a renowned statistician, examines past predictions and suggests that most predictions fail because we have a poor understanding of probability and uncertainty and we lack the humility to learn from our failures; Netflix crowdsourced a web event that awarded one million U.S. dollars for an algorithm that could best Netflix's own algorithm for predicting film ratings by at least 10% – the winner was able to best it by10.06%; Black & Scholes (1973) established a cornerstone of modern finance by developing a future option pricing formula (which, surprisingly and because of a hedging assumption, does not depend on the expected return of the underlying share) that won them a Nobel Prize in 1997; although largely unseen, Siegel (2013) is convinced that Big Data analytics is driving millions of decisions, determining who to call, mail, investigate, incarcerate, set up on a date, or medicate; SAS advertises its analytics prowess in visualization, marketing and business; microtargeting played a critical role in both of President Barack Obama's presidential campaigns (Samuelson 2013); Apple's intelligent personal assistant, Siri, is able to listen and communicate in several languages, while undertaking semantic-based management, search, and retrieval of available data; and open-source Hadoop (Zikopoulos et al. 2012) – which is at the core of the Watson computer – and Google's MapReduce attempt to find answers to queries by quickly and cheaply sifting through massive data sets that are

residing in different data centers. Parenthetically, Hadoop was the name of the toy elephant belonging to the son of Douglas Cutting, an advocate and creator of the open-source search technology.

In regard to analytics accreditation, the American Statistical Association (ASA) developed in 2010 a PStat (Accredited Professional Statistician) designation; it is a voluntary portfolio-based rather than examination-based designation, covering the candidate's education, experience, and demonstrated statistical competence as well as their agreement to abide by ethical standards of practice. The Institute of Operations Research and Management Science (INFORMS) developed in 2013 a CAP (Certified Analytics Professional) designation that is also based on a portfolio of experience and skills (i.e., attainment of at least a BS or BA degree, 3-7 years of professional analytics work experience depending on the degree field and level, and confirmation of acceptable "soft skills" by the candidate's current or previous employers), followed by a 100-question, task-and-knowledge examination. Furthermore, in order to maintain the CAP designation, a minimum of 30 professional development units (PDUs) are required in a 3-year renewal period; the PDUs may be achieved through formal professional education courses, self-directed learning, new analytics knowledge development, volunteer service, and/or relevant work experience, with every hour of activity being equal to one PDU. Additionally, other niche analytics are being practiced by IBM, SAS, etc. without the benefit of certified experts; hopefully, as with any new technology, may the best practices be documented and promulgated.

Four remarks should be made in connection with the astounding growth in Big Data analytics. First, there is a concern that modeling or design support software like Ansys and SolidWorks may undermine the need for young engineers to engage in hands-on activities; on the other hand, such support may provide more time for the aspiring engineers to become more involved in understanding the physical complexities of their software-supported designs. Second, a similar concern is that powerful machines like Watson might displace human workers; again, machines which can satisfy the Turing (1950) test of artificial intelligence are yet to be built by humans – meanwhile, existing machines are, for the most part, only doing the drudgery work that humans dislike, including searching, mining and matching. In fact, Watson's new offering, known as IBM Watson Engagement Advisor, will give customer service transactions a layer of cognitive computing help, leveraging Watson's unique skills to semantically answer questions. Third, the question may well be asked about the usefulness or impact of Big Data analytics; Tables 8 and 9 respectively assess the potential impact in regard to the 14 grand challenges (promulgated by the National Academy of Engineering) and the 10 breakthrough technologies (reported by Technology Review, a publication by the Massachusetts Institute of Technology) –resulting in an impact valuation, ranging from medium to high. Fourth, as with Big Data acquisition and access, Big Data analytics is also subject to the same privacy and confidentiality concerns.

Table 8 Big Data impact: 14 NAE grand challenges

CATEGORY	GRAND CHALLENGES	FOCUS	IMPACT
Healthcare & Technobiology	1. Advance Health Informatics	Detect, Track and Mitigate Hazards	High (3)
	2. Engineer Better Medicines	Develop Personalized Treatment	Medium (2)
	3. Reverse-Engineer The Brain	Allow Machines to Learn & Think	High (3)
Informatics & Risk	4. Secure Cyberspace	Enhance Privacy & Security	High (3)
	5. Enhance Virtual Reality	Test Design & Ergonomics Schemes	High (3)
	6. Advance Personal Learning	Allow Anytime, Anywhere Learning	High (3)
	7. Engineer Discovery Tools	Experiment, Create, Design and Build	Medium (2)
	8. Prevent Nuclear Terror	Identify& Secure Nuclear Material	Low (1)
Sustainable Systems	9. Make Solar Energy Economical	Improve Solar Cell Efficiency	Low (1)
	10. Provide Energy From Fusion	Improve Fusion Control & Safety	Low (1)
	11. Develop Sequestration Methods	Improve Carbon Dioxide Storage	Low (1)
	12. Manage The Nitrogen Cycle	Create Nitrogen, Not Nitrogen Oxide	Low (1)
	13. Provide Access To Clean Water	Improve Decontamination/Desalination	Low (1)
	14. Improve Urban Infrastructure	Restore Road, Sewer, Energy, Etc. Grids	Medium (2)
Average Impact			Medium (1.9)

Table 9 Big Data impact: 10 Technology Review breakthrough technologies

CATEGORY	BREAKTHROUGH TECHNOLOGIES	FOCUS	IMPACT
Healthcare & Technobiology	1. Deep Learning	Mimic The Brain Through Digital Patterns	High (3)
	2. Prenatal DNA Sequencing	Determine Genetic Destiny of Unborn	Medium (2)
	3. Memory Implants	Form Memories Despite Brain Damage	Low (1)
Informatics & Risk	4. Baxter: The Blue Collar Robot	Reprogram Robotic Functions As Needed	High (3)
	5. Big Data From Cheap Phones	Detect Disease Spread By Mobility Data	High (3)
	6. Temporary Social Media	Maintain Privacy By Self-Destruct Tweets	Medium (2)
	7. Smart Watches	Allow Easy-to-Use Interface to Phone Data	High (3)
Sustainable Systems	8. Ultra-Efficient Solar Power	Improve Solar Cell Efficiency	Medium (2)
	9. Supergrids	Integrate Wind & Solar By DC Grid	Medium (2)
	10. Additive Manufacturing	Make Complex Parts By 3D Printing	High (3)
Average Impact			Medium (2.4)

3.4 Application

To begin, it is, of course, difficult to separate Big Data analytics from Big Data applications; however, for the purpose of this paper, Section 3 is focused on the prominent types of analytics, while the current section considers the broader application areas that may be based on one or more of the individual analytics – nevertheless, there may be areas of overlap.

Table 10 identifies some topics of Big Data application, including smart innovations, data-driven insights, data-driven decisions and mass customization. In regard to smart innovations, IBM is credited with coining and trademarking the term "Smarter Planet", which is their overarching goal for collaborating with companies, cities and communities in establishing social businesses (by developing a social network within the enterprise and empowering customers and partners to crowdsource the enterprise); a range of driver-assist technologies (e.g., reverse-park, crash-avoidance, skid-avoidance, etc.), including those being developed by Google, will essentially constitute an autopilot for cars, which will in turn lower congestion, pollution, injuries,

deaths and, ultimately, insurance costs; Intel's algorithmically-designed 3D transistors are faster, more energy-efficient and more powerful than the traditional 2D devices (and, moreover, will extend the life of Moore's Law which predicts that the number of transistors per chip will double every two years); Rethink Robotics' Baxter is not only an innovative industrial robot, it is also a platform that can be employed to develop other robotic applications; Google Glass will not only record audio and video but also provide total recall and support augmented reality and cognition (i.e., a brain prosthesis); and telemedicine can improve access to medical services and save lives in critical care and emergency situations, especially when it is augmented by an Internet that can transmit the sense of touch.

In regard to data-driven insights, PECOTA (Player Empirical Comparison and Optimization Test Algorithm) is a system developed by Silver (2003) for forecasting Major League Baseball player performance that exemplifies the importance of understanding probability, uncertainty and Bayesian reasoning with well-reasoned priors; machine learning, a branch of artificial intelligence, will benefit from quantum computing which allows for the simultaneous analyses of data; autonomous systems – like driverless cars and drones – are, of course, data-driven in its performance; and a number of consulting companies – like McKinsey, Boston Consulting Group and Bain – are always ready to provide services that help organizations to deal with new government regulations, including the Dodd-Frank financial reform and the Obama Care health reform.

In regard to data-driven decisions, McAfee & Brynjolfsson (2012) caution that a change in the enterprise's decision making culture must occur before the benefits of Big Data can revolutionize company management and performance; as more local, state and federal agencies make their data public or digitally available, more new Big Data related businesses will flourish (e.g., car-navigation, precision farming, property valuation, matching suppliers and consumers, etc.); human resource data can be mined for insights on promotion and retention; the healthcare industry can be helped

Table 10 Big Data scope: data application

SCOPE	EXAMPLE APPLICATIONS	EXAMPLE EFFORTS
Smart Innovations	Smart Buildings & Power Grids; Smarter Planet Smart Devices & Cell Phones; Mobile Internet; Robots; Reality Augmentation; Telemedicine	Global Positioning System; Driverless Cars; IBM Apple's iPhone 5; Intel's 3D Transistors; Rethink Robotics' Baxter; Google Glass
Data-Driven Insights	Probability; Uncertainty; Bayes; Machine Learning; Autonomous Systems Consulting on Dodd-Frank Reform, Obama Care	PECOTA, Baseball Prospectus; Algorithmic Trading; Drones McKinsey; Boston Consulting Group; Bain
Data-Driven Decisions	Economic Development in All 5 Sectors; Improved Health Throughout The Globe; Enhanced Quality of Life	Resource Management; Open Public Data Anticipating Disease; Consumer Choice; Reverse Engineering The Brain
Mass Customization	Big Data Analytics; Adaptive Services; Digital Manufacturing	3D Imaging & Multimedia Information; Nanopore DNA Sequencing; Social Business; Additive Manufacturing; 3D/4D Printing

in its transformation by tracking encounters over time and establishing personalized recommendations and aggregated benchmarks; and by carefully analyzing millions of neuroimages, we can better understand how we think, how we learn, how we remember, and how we can control our muscles and interface with bionic devices.

In regard to mass customization and as depicted in Figure 5, Tien (2012) augurs that Big Data (especially Big Data analytics) is the foundation for personalization or mass customization, which in turn is the basis for the Third Industrial Revolution (TIR). The first two industrial revolutions focused on goods (especially manufactured goods), while the third is focused on services and goods (especially the integration of services and/or goods). The First Industrial Revolution (FIR) focused on textiles and iron making; it sought to enhance productivity in production, employed mechanical tools that had life cycles on the order of decades, depended mostly on muscle power, embraced a living standard concerned with subsistence, and had its initial impact in Britain in 1750. The Second Industrial Revolution (SIR) focused on assembly lines (especially as conceived by Henry Ford) and steel making; it sought to enhance productivity in mass production, employed electromechanical tools that had life cycles on the order of years, depended mostly on both muscle and brain power, embraced a living standard concerned with quality of goods, and had its initial impact in the U.S. and Germany in 1860. The Third Industrial Revolution (TIR) is focused on adaptive services and digital manufacturing; it seeks to enhance productivity in mass customization (especially in regard to the integration of services and manufactured goods), employs computer and communication technologies that have life cycles on the order of months, depends mostly on brain power, embraces a living standard concerned with the quality of life, and had its initial impact in the U.S. in 2010. In contrast to the first two industrial revolutions, which emphasized the production of goods (from agriculture, mining, manufacturing and construction), the third industrial revolution builds on the extensive development of customized services and integrates it with the production of customized goods. Thus, every nation has gone or will go through these three industrial revolutions, consistent with their transition through the three stages of economic evolution: today, the underdeveloped nations are still at the mechanical stage or experiencing FIR; most of

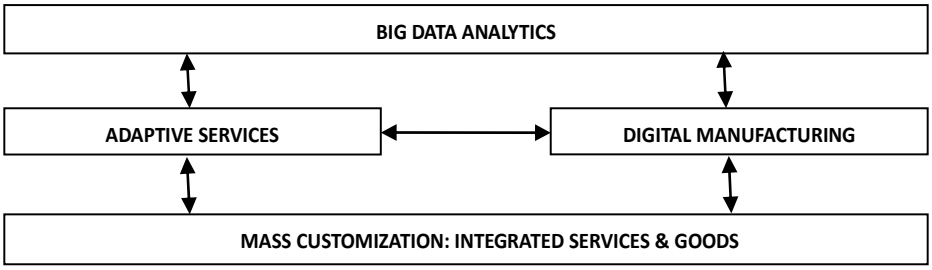

Figure 5 Third industrial revolution enablers

the developed nations are at the electrical stage or experiencing SIR; while the economically advanced nations are, as noted above, well into the services stage and, with the advent of Big Data, beginning to experience TIR.

Amazon is a TIR company; among other distinctions, it stands out as the publisher with the most ambitious goal of digitizing all published material and then making it available on an on-demand basis. On-demand or customized printing allows text, graphics and images to be assembled per the customer's need; thus, a professor can assemble a customized textbook for her class or a marketer can assemble a set of personalized or customized letters, each with the same basic layout but printed with a different salutation, name and address. It should be noted that such a digital publishing or printing approach is an adaptive service that employs a digital manufacturing technology, leading to mass customization or an on-demand production venue. TIR is flourishing, as Big Data, adaptive services and digital manufacturing are being enhanced; for example, 3D printing is making way for 4D printing (i.e., 3D printing over time).

Five remarks should be made in connection with the astounding growth in Big Data application. First, there is a concern that some smart innovations (e.g., smart power grids) are compromising privacy, raising costs rather than lowering them, and, as noted in Section 3.1, threatening health with electromagnetic fields from the smart sensors used to collect and transmit the data. Second, a similar concern is that smart innovations like robots and driverless cars will be difficult to accept; however, it should be noted that robots (including furniture assemblers and autistic children entertainers), airplanes and bullet trains have all been subject to autonomous control – yes, accidents may still happen but their casualty rate is much less than that for human-controlled systems. On the other hand, it is true that humans have to learn to trust the autonomous systems and overturn, for example, decades of road-safety legislation. Third, as Big Data is encompassing in its approach, it can facilitate interdisciplinary activities; indeed, Nobelist McFadden (2013) is advocating for broadening the traditional theory of consumer choice (which is based on the assumption that consumers are driven by self-interest) to include the evidence being provided by cognitive psychology, anthropology, evolutionary biology, and neurology. Fourth, it should be noted that the term TIR was first coined by the economist, Rifkin (2011); he considers TIR to be the integration of Internet technology (an adaptive service) with renewable energy (a digitally manufactured good) – which, in essence, is consistent with this paper's view of the Third Industrial Revolution. Fifth, as with Big Data acquisition, access and analytics, Big Data application is also subject to the same privacy and confidentiality concerns; perhaps more so because applications have a direct impact on the individual.

4. Conclusion

In conclusion, it is helpful to briefly consider the benefits of and concerns with Big Data. A number of papers discuss the challenges and opportunities posed by Big Data (Chen et al. 2012, Bizer et al. 2013), namely the engineering challenge of efficiently managing large data sets and the semantics challenge of locating and

meaningfully integrating the data of interest. In regard to the benefits, Big Data allows for a) better integration or fusion and subsequent analysis of quantitative and qualitative data; b) better observation of rare but great impact events or "black swans" (Taleb 2010); c) greater system and system-of-systems efficiency and effectiveness; d) better evidence-based – "data rich, information unleashed" (DRIU) – decisions that can overcome the prejudices of the unconscious mind (Mlodinow 2011); and e) more messy (including data quality concerns) findings but yet good enough to support informed decisions.

In regard to concerns, Table 11 provides a summary in terms of the four Big Data components: acquisition (i.e., Big Data does not imply big/complete understanding of underlying problem; Big Data quantity does not imply Big Data quality; and Big Data sample does not imply a representative or even a complete sample); access (i.e., Big Data's on-demand accessibility may create privacy concerns; Big Data's real-time abilities may obscure past and future concerns; and Big Data's cyber-security concerns may overlook personal-security concerns); analytics (i.e., Big Data's inherent messiness may obscure underlying relationships; Big Data's correlational finding may result in an unintended causal consequence; and Big Data's unleashing of information may obscure underlying knowledge); and application (i.e., Big Data's feasible explanations may obscure more probable explanations; Big Data's evidence-driven findings may obscure underlying factual knowledge; and Big Data's subjective, consumer-centric findings may obscure objective findings). Other concerns include surveillance by autocratic governments and processing data in an increasingly unfocused, unproductive and generally "shallow manner" (Carr 2010). Of course, the potential Big Data concerns or problems can be mitigated with thoughtful and effective approaches and practices; for example, legislation could be promulgated and passed to forbid the invasion of privacy and to dispense severe sanctions against those who break the law.

Table 11 Big Data: potential concerns

COMPONENTS	ELEMENTS	POTENTIAL CONCERNS
Acquisition	Focus	Big Data Does Not Imply Big/Complete Understanding of Underlying Problem
	Emphasis	Big Data Quantity Does Not Imply Big Data Quality
	Scope	Big Data Sample Does Not Imply A Representative or Even A Complete Sample
Access	Focus	Big Data's On-Demand Accessibility May Create Privacy Concerns
	Emphasis	Big Data's Real-Time Abilities May Obscure Past and Future Concerns
	Scope	Big Data's Cyber-Security Concerns May Overlook Personal-Security Concerns
Analytics	Focus	Big Data's Inherent Messiness May Obscure Underlying Relationships
	Emphasis	Big Data's Correlational Finding May Result In An Unintended Causal Consequence
	Scope	Big Data's Unleashing of Information May Obscure Underlying Knowledge
Application	Focus	Big Data's Feasible Explanations May Obscure More Probable Explanations
	Emphasis	Big Data's Evidence-Driven Findings May Obscure Underlying Factual Knowledge
	Scope	Big Data's Subjective, Consumer-Centric Findings May Obscure Objective Findings

It is also helpful to qualitatively compare the two data approaches – traditional and Big Data – on four overarching criteria: 1) usefulness (i.e., degree to which traditional and Big Data acquisition, access, analytics and application are useful); 2) timeliness (i.e., degree to which traditional and Big Data acquisition, access, analytics and application are timely in their impact); 3) privacy-sensitivity (i.e., degree to which traditional and Big Data acquisition, access, analytics and application are sensitive to privacy issues); and 4) benefit-cost (i.e., degree to which traditional and Big Data acquisition, access, analytics and application have a positive benefit to cost impact). Table 12 provides the author's subjective impact assessments; it is interesting to note that Big Data's impact is medium-high, as compared to the traditional approach's impact of medium. Clearly, the Big Data approach – with its technology-enhanced acquisition, access, analytics and application abilities – is superior to the traditional approach that has been employed to date.

Within just the past year, Table 13 lists four recent Big Data initiatives, if not investments. The Simons Foundation selected the University of California, Berkeley, as host for an ambitious new $60M Simons Institute for the Theory of Computing, whereby an interdisciplinary team of scientists and engineers will tackle problems in fields as diverse as healthcare, astrophysics, genetics and economics. Boston University received $15M to establish the Rafik B. Hariri Institute for Computing and Computational Science and Engineering, an interdisciplinary research center for discoveries through the use of computational and data-driven approaches and for advances in the science of computing. GE decided to open a 400-person, $1B Global Software Center in San Ramon, California, near

Table 12 Traditional versus Big Data impact

COMPONENTS	ELEMENTS	TRADITIONAL	BIG DATA
Acquisition	Usefulness	Medium (2)	High (3)
	Timeliness	Low (1)	High (3)
	Privacy-Sensitivity	High (3)	Low (1)
	Benefit-Cost	Medium (2)	Medium (4)
Access	Usefulness	Medium (2)	High (3)
	Timeliness	Low (1)	High (3)
	Privacy-Sensitivity	High (3)	Low (1)
	Benefit-Cost	Medium (2)	High (3)
Analytics	Usefulness	Medium (2)	Medium (2)
	Timeliness	Medium (2)	High (3)
	Privacy-Sensitivity	Medium (2)	Medium (2)
	Benefit-Cost	Medium (2)	Medium (2)
Application	Usefulness	Medium (2)	High (3)
	Timeliness	Low (1)	High (3)
	Privacy-Sensitivity	Medium (2)	Medium (2)
	Benefit-Cost	Medium (2)	High (3)
Average Impact		Medium (1.9)	Medium-High (2.5)

Table 13 Recent Big Data efforts in U.S.

EFFORT	LOCATION	AMOUNT	FUNDER
Simons Institute For The Theory of Computing	U.C., Berkeley	$60M	Simons Foundation
Institute for Computational Science & Engineering	Boston U	$15M	Rafik B. Hariri
Global Software Center	San Ramon, CA	$1B	GE
Various Other Big Data Initiatives	Mostly At Universities	$1B+ Per Year	U.S. Agencies

Silicon Valley; its focus is on Big Data analytics. Led by the U.S. Office of Science and Technology Policy, the Departments of Defense, Energy, Health and Human Services (including the National Institutes of Health) and other funding agencies (including the National Science Foundation) are providing over $1B to fund mostly academic research in the core technologies associated with Big Data analytics and cloud computing

A final point should be made about Big Data; as suggested throughout this paper, it has to be regarded as a permanent disruptive innovation or transformation. That is, data must be constantly acquired, accessed, analyzed and applied, resulting in new – and changing – insights that might be disruptive in nature. To profit from Big Data, one must accept uncertainty and change as a permanent state of affairs; it must be a part of any enterprise's DNA. Indeed, some companies invite such changes by adopting processes that enable variation, not eliminate it, and by valuing disruptions over the relentless pursuit of a single vision (e.g., efficiency). As an example, Google encourages some of the company's workers to spend 20 percent of their time on projects of their own choosing and provides additional resources to those ideas with the most merit. In short, change is the only constant; companies which do not embrace it will face the same demise as Kodak, Digital Equipment Corporation and Atari. On the other hand, those which allow for disruptive innovations, like GE and IBM, have not only survived but thrived.

References

[1] Ahlquist, J. & Saagar, K. (May/June 2013). Comprehending the complete customer. Analytics Magazine, 36-50

[2] Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K. & Tuecke, S. (2012). Software as a service for data scientists. Communications of the ACM, 55 (2): 81-88

[3] Appel, K. & Haken, W. (October 1977). Solution of the four color map problem. Scientific American, 237 (4): 108-121

[4] Barbot, S., Lapusta, N. & Avouac, J.-P. (2012). Under the hood of the earthquake machine: toward predictive modeling of the seismic cycle. Science, 336: 707-710

[5] Baru, C., Bhandarkar, M., Nambiar, R., Poess, M. & Rabl, T. (March 2013).

Benchmarking big data systems and the bigdata top 100 list. Big Data, 60-64

[6] Bizer, C., Boncz, P., Bodie, M.L. & Erling, O. (December 2011). The meaningful use of big data: four perspectives – four challenges. SIGMOD Record, 40 (4): 56-60

[7] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81: 637-654

[8] Carr, N. (2010). The Shallows: What the Internet Is Doing to Our Brains. Norton, New York, NY

[9] Chen, H., Chiang, R.H.L. & Storey, V.C. (2012). Business intelligence and analytics: from big data to big impact. MIS Quarterly, 36 (4): 1165-1188

[10] Davenport, T.H. & Harris, J.G. (2007). Competing on Analytics: The New Science of Winning. Harvard Business School Press, Cambridge, MA

[11] Futardo, P. (2009). A survey of parallel and distributed data warehouses. International Journal of Data Warehousing and Mining, 5 (2): 57-77

[12] Hattori, H., Nakajima, Y. & Ishida, T. (2011). Learning from humans: agent modeling with individual human behaviors. IEEE Transactions on Systems, Man and Cybernetics – Part A, 41 (1): 1-9

[13] Jacobs, A. (2009). The pathologies of big data. Communications of the ACM, 52 (8): 36-44

[14] Lavalle, S., Lesser, E., Shockley, R., Hopkins, M.S. & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. MIT Sloan Management Review, 52 (2): 21-31

[15] Luhn, H.P. (1958). A business intelligence system. IBM Journal, 2 (4): 314-350

[16] Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxbury, C. & Byers, A.H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, New York, NY

[17] Mayer-Schonberger, V. & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. Houghton Mifflin Harcourt Publishing Company, New York, NY

[18] McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. Harvard Business Review, October, 3-9

[19] McFadden, D.L. (2013). The new science of pleasure. National Bureasu of Economic Research. Working paper 18687

[20] Mlodinow, L. (2012). Subliminal: How Your Unconscious Mind Rules Your Behavior. Pantheon Books, New York, NY

[21] Rifkin, J. (2011). The Third Industrial Revolution: How Lateral Power Is Transforming Energy, the Economy, and the World. Palgrave Macmillan, New York, NY

[22] Samuelson, D.A. (2013). Analytics: key to Obama's victory. OR/MS Today, February, 20-24

[23] Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. & Nolan, G.P. (2010). Computational solutions to large-scale data management and analysis. Nature Reviews, 11: 647-657

[24] Segall, P. (2012). Understanding earthquakes. Science, 336: 676-710

[25] Shostack, A. (2012). The evolution of information security. The Next Wave, 19 (2): 6-11

[26] Siegel, E. (2013). Predictive Analytics: The

Power to Predict Who Will Click, Buy, Lie, or Die. John Wiley & Sons, New York, NY

[27]Silver, N. (2003). Introducing PECOTA. In: Huckabay, G., Kahrl, C., Pease, D. (eds.), Baseball Prospectus 2003, pp. 507-514. Dulles, VA: Brassey's Publishers

[28]Silver, N. (2012). The Signal and the Noise: Why So Many Predictions Fail – But Some Don't. The Penguin Press, New York, NY

[29]Stern, A. & Lindner, N.H. (8 March 2013). Topological quantum computation – from basic concepts to first experiments. Science, 339: 1179-1184

[30]Swart, E.R. (1980). The philosophical implications of the four-color problem. American Mathematical Monthly, 87 (9): 697-702

[31]Taleb, N.N. (2010). The Black Swan: Second Edition. Random House, Inc., New York, NY

[32]Tien, J.M. (2003). Toward a decision informatics paradigm: a real-time information based approach to decision making. IEEE Transactions on Systems, Man and Cybernetics, Part C, Special Issue, 33 (1): 102-113

[33]Tien, J.M. (2008). On integration and adaptation in complex service systems. Journal of Systems Science and Systems Engineering, 17 (2): 1-31

[34]Tien, J.M. (2011). Manufacturing and services: from mass production to mass customization. Journal of Systems Science and Systems Engineering, 20 (2): 129-154

[35]Tien, J.M. (2012). The next industrial revolution: integrated services and goods. Journal of Systems Science and Systems Engineering, 21 (3): 257-296

[36]Tien, J.M. & Berg, D. (1995). Systems engineering in the growing service economy. IEEE Transactions on Systems, Man, and Cybernetics, 25 (5): 321-326

[37]Tien, J.M. & Berg, D. (2003). A case for service systems engineering. International Journal of Systems Engineering, 12 (1): 13-39

[38]Tien, J.M., Krishnamurthy, A. & Yasar, A. (2004). Towards real time management of supply and demand chains. Journal of Systems Science and Systems Engineering, 13 (3): 257-278

[39]Tien, J.M. & McClure, J.A. (1986). Towards an operations-oriented approach to information systems design in public organizations. Public Administration Review, Special Issue on Public Managemnt Information Systems, 27 (7): 553-562

[40]Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59: 433-460

[41]van Hattum, P. & Hoijtink, H. (2008). Data fusion: an application in marketing. Database Marketing & Customer Strategy Management, 15 (4): 267-284

[42]Wilson, R. (2002). Four Colors Suffice. Penguin Books, London, England

[43]Zikopoulos, P.C., Eaton, C., DeRoos, D., Deutsch, T. & Lapis, G. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. The McGraw-Hill Companies, New York, NY

James M. Tien received the BEE from Rensselaer Polytechnic Institute (RPI) and the SM, EE and PhD from the Massachusetts Institute of Technology. He has held leadership

positions at Bell Telephone Laboratories, at the Rand Corporation, and at Structured Decisions Corporation. He joined the Department of Electrical, Computer and Systems Engineering at RPI in 1977, became Acting Chair of the department, joined a unique interdisciplinary Department of Decision Sciences and Engineering Systems as its founding Chair, and twice served as the Acting Dean of Engineering. In 2007, he was recruited by the University of Miami to be a Distinguished Professor and Dean of its College of Engineering. He has been awarded the IEEE Joseph G. Wohl Outstanding Career Award, the IEEE Major Educational Innovation Award, the IEEE Norbert Wiener Award, and the IBM Faculty Award. He is also an elected member of the prestigious U. S. National Academy of Engineering.