2013 46th Hawaii International Conference on System Sciences

Big Data: Issues and Challenges Moving Forward

Stephen Kaisler
i_SW Corporation
skaisler1@comcast.net

Frank Armour
American University
fjarmour@gmail.com

J. Alberto Espinosa
American University
alberto@american.edu

William Money,
George Washington University
wmoney@gwu.edu

Abstract

Big data refers to data volumes in the range of exabytes (10^{18}) and beyond. Such volumes exceed the capacity of current on-line storage systems and processing systems. Data, information, and knowledge are being created and collected at a rate that is rapidly approaching the exabyte/year range. But, its creation and aggregation are accelerating and will approach the zettabyte/year range within a few years. Volume is only one aspect of big data; other attributes are variety, velocity, value, and complexity. Storage and data transport are technology issues, which seem to be solvable in the near-term, but represent long-term challenges that require research and new paradigms. We analyze the issues and challenges as we begin a collaborative research program into methodologies for big data analysis and design.

1. Introduction

The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools. Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. The original definition focused on structured data, but most researchers and practitioners have come to realize that most of the world's information resides in massive, unstructured information, largely in the form of text and imagery. The explosion of data has not been accompanied by a corresponding new storage medium.

We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. These imitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and methods) used to analyze it. As with any new problem, the conclusion of how to proceed may lead to a recommendation that new tools need to be forged to perform the new tasks. As little as 5 years ago, we were only thinking of tens to hundreds of gigabytes of storage for our personal computers. Today, we are thinking in tens to hundreds of terabytes. Thus, big data is a moving target. Put another way, it is that amount of data that is just beyond our immediate grasp, e.g., we have to work hard to store it, access it, manage it, and process it.

The current growth rate in the amount of data collected is staggering. A major challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both: (1) design appropriate systems to handle the data effectively and (2) and analyze it to extract relevant meaning for decision making. In this paper we identify critical issues associated with data storage, management, and processing. To the best of our knowledge, the research literature has not effectively addressed these issues,

1.1 Importance of Big Data

In August 2010, the White House, OMB, and OSTP proclaimed that Big Data is a national challenge and priority along with healthcare and national security [1]. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and techniques. Its goal is to advance our "…understanding of the technologies needed to manipulate and mine massive amounts of information; apply that knowledge to other scientific fields "as well as address the national goals in the areas of health energy defense, education and researcher" [14].

The government's emphasis is on how big data creates "value" – both within and across disciplines and domains. Value arises from the ability to analyze the data to develop actionable information. Our survey of the technical literature suggests five generic ways that big data can support value creation for organizations (see Table 1).

1530-1605/12 $26.00 © 2012 IEEE
DOI 10.1109/HICSS.2013.645

Table 1. Value Created from Big Data

Creating transparency by making big data openly available for business and functional analysis (quality, lower costs, reduce time to market, etc.)
Supporting experimental analysis in individual locations that can test decisions or approaches, such as specific market programs
Assisting, based on customer information, in defining market segmentation at more narrow levels
Supporting Real-time analysis and decisions based on sophisticated analytics applied to data sets from customers and embedded sensors
Facilitating computer-assisted innovation in products based on embedded product sensors indicating customer responses

While the government seems to assume that big data users will be more successful, more productive, and have differential impacts across many industries, their underlying concern seems to be a lack of tools and a lack of trained personnel to properly work with big data. Others suggest that the analysis of generic sequences, social media interactions, health records, phone logs, and government records, will not create better tools and services, but may create a new set of privacy incursions and invasive and unwanted marketing.[3] these conflicting concerns drive competing visions of how to deal with big data.

An example from the medical field illustrates how and why big data and new analytics may be truly beneficial. Fox [6] describes how current data in a patient's medical record and current health situation is used to plan and target patient participation in wellness and disease management programs. Fox asserts that doctors (and insurance companies!) must *understand* the patient rather than the disease(s). To do so, they must collect and analyze data - "crucial social and behavioral data that impacts a patient's choice to participate, level of engagement, and appropriateness from public data associating behavior and health data – beyond that solely related to a patient's medical condition". Thus, programs may determine how to better target, retain, and treat people in their programs by leveraging predictive models that could assist doctors and case managers who seek to positively impact the behavior of patients with chronic health disease.

1.2 Big Data Characteristics

One view, espoused by Gartner's Doug Laney describes Big Data as having three dimensions: volume, variety, and velocity. Thus, IDC defined it: "*Big data technologies describe a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis."* [8] Two other characteristics seem relevant: value and complexity. We summarize these characteristics in table 2.

1.3 Big Data – Where is it?

Big data surrounds us, although we may not immediately realize it (see Table 3). Part of the problem is that, except in unusual circumstances, most of us don't deal with large amounts of data in our everyday lives. Lacking this immediate experience, we often fail to understand both opportunities as well challenges presented by big data. Because of these characteristics,there are currently a number of issues and challenges in addressing these characteristics going forward.

1.4 Issues

We suggest there are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues. Each of these represents a large set of technical research problems in its own right.

Table 2. Big Data Characteristics

Data Volume**:** Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.
Data Velocity**:** Data velocity measures the speed of data creation, streaming, and aggregation. eCommerce has rapidly increased the speed and richness of data used for different business transactions (for example, web-site clicks). Data velocity management is much more than a bandwidth issue; it is also an ingest issue (extract-transform-load.
Data Variety**:** Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

***Data Value*:** Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data.
***Complexity*:** Complexity measures the degree of interconnectedness (possibly very large) and interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all.

1.4.1 Storage and Transport Issues

The quantity of data has exploded each time we have invented a new storage medium. What is different about the most recent explosion – due largely to social media – is that there has been no new storage medium. Moreover, data is being created by everyone and everything (e.g., devices, etc) – not just, as heretofore, by professionals such as scientist, journalists, writers, etc.

Table 3. Some Examples of Big Data

Data Set/Domain	Description
Large Hadron Collider/Particle Physics (CERN)	13-15 petabytes in 2010
Internet Communications (Cisco)	667 exabytes in 2013
Social Media	12+ Tbytes of tweets every day and growing. Average retweets are 144 per tweet.
Human Digital Universe	1.7 Zbytes (2011) -> 7.9 Zbytes in 2015 (Gantz and Reinsel 2011)
British Library UK Website Crawl	~ 110 TBytes per domain crawl to be archived
Other	RFIDS, smart electric meters, 4.6 billion camera phones w/ GPS

Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take longer to transmit the data from a collection or storage point to a processing point than it would to actually process it!

Two solutions manifest themselves. First, process the data "in place" and transmit only the resulting information. In other words, "bring the code to the data", vs. the traditional method of "bring the data to the code." Second, perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data.

1.4.2 Management Issues

Management will, perhaps, be the most difficult problem to address with big data. This problem first surfaced a decade ago in the UK eScience initiatives where data was distributed geographically and "owned" and "managed" by multiple entities. Resolving issues of access, metadata, utilization, updating, governance, and reference (in publications) have proven to be major stumbling blocks.

Unlike the collection of data by manual methods, where rigorous protocols are often followed in order to ensure accuracy and validity, digital data collection is much more relaxed. The richness of digital data representation prohibits a bespoke methodology for data collection. Data qualification often focuses more on missing data or outliers than trying to validate every item. Data is often very fine-grained such as clickstream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed.

The sources of this data are varied - both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models, software behaviors, user interface designs, etc – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance. Yet, all this data is readily available for inspection and analysis.

Going forward, data and information provenance will become a critical issue. JASON has noted [10] that "there is no universally accepted way to store raw data, … reduced data, and … the code and parameter choices that produced the data." Further, they note:

"We are unaware of any robust, open source, platform-independent solution to this problem." As far as we know, this remains true today. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled.

1.4.3 Processing Issues

Assume that an exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information.

2. Dynamic Design Challenges

There are numerous challenges requiring long-term research to working with big data. Stonebreaker and Hong [18] argue that the design for the systems and components that work with big data will require an understanding of both the needs of the users and the technologies that can be used to solve the problem being investigated – i.e., not all big data and its requirements are the same. In this instance, since the data that is newly created (envisioned and collected), is neither truly known or well understood, designers will need to consider interfaces, graphics, and icons; application organization; and conceptual models, metaphors, and functionality. Because the end users will not often be the system designers, this presents an additional design challenge.

There are unknown challenges that will arise with each increase in scale and development of new analytics. Some of these challenges will be intractable with the tools and techniques at hand. We believe these challenges to be just "over the horizon" with the next jump to zettabyte-size data sets.

2.1 Data Input and Output Processes

A major issue raised in big data design is the output process. Jacobs [9] summarized the issue very succinctly – "…its easier to get the data in than out." His work shows that data entry and storage can be handled with processes currently used for relational databases. But, the tools designed for transaction processing that add, update, search for, and retrieve small to large amounts of data are not capable of extracting the huge volumes and cannot be executed in seconds or a few minutes.

How to access very large quantities of semi- or unstructured data, and how to utilize as yet unknown tool designs is not known. It is clear the problem may neither be solved by dimensional modeling and online analytical processing (OLAP), which may be slow or have limited functionality, nor by simply reading all the data into memory. Technical considerations that must be factored into the design include the ratio of the speed of sequential disk reads to the speed of random memory access. The current technology shows that random access to memory is 150,000 times slower than sequential access. Joined tables, an assumed requirement of associating large volumes of disparate but somehow related data, perhaps by observations over time alone, will come at further huge performance costs. (Jacobs, 2009)

2.2 Quality versus Quantity

An emerging challenge for big data users is "quantity vs. quality". As users acquire and have access to more data (quantity), they often want even more. For some users, the acquisition of data has become an addiction. Perhaps, because they believe that with enough data, they will be able to perfectly explain whatever phenomenon they are interested in.

Conversely, a big data user may focus on quality which means not having all the data available, but having a (very) large quantity of high quality data that can be used to draw precise and high-valued conclusions. (see Table 3) .

Another way of looking at this problem is, what is the level of precision that the user requires? For example, trend analysis may not require the precision that traditional DB systems provide, but which requires massive processing in a Big Data environment. This problem also manifests itself in the "speed versus scale" challenge discussed below.

Table 4. Some Quantity and Quality Challenges

How do we decide which data is irrelevant versus selecting the most relevant data?
How do we ensure that all data of a given type is reliable and accurate? Or, maybe just approximately accurate?
How much data is enough to make an estimate or prediction of the specific probability and accuracy of a given event?
How do we assess the "value" of data in decision making? Is more necessarily better?

2.3 Data Growth versus Data Expansion

Most organizations expect their data to grow over their lifetime as the organization increases its services, its business and business partners and clients, its projects and facilities, and its employees. Few businesses adequately consider data expansion, which occurs when the data records grow in richness, when they evolve over time with additional information as new techniques, processes and information demands evolve. Most data is time-varying – the same data items can be collected over and over with different values based on a timestamp. Much of this data is required for retrospective analysis – particularly that which is used in estimative and predictive analytics.

2.4 Speed versus Scale

As the volume of data grows, the "big" may morph from the scale of the data warehouse to the amount of data that can be processed in a given interval, say 24 hours. Gaining insight into the problem being analyzed is often more important than processing all of the data. Time-to-information is critical when one considers (near) real-time processes that generate near-continuous data, such as radio frequency identifiers (RFIDs – used to read electronic data wirelessly, such as with EZPass tags) and other types of sensors. An organization must determine how much data is enough in setting its processing interval because this will drive the processing system architecture, the characteristics of the computational engines, and the algorithm structure and implementation.

That said, another major challenge is *data dissemination*. The bottleneck is the communications middleware. While communication hardware speeds are increasing with new technologies, message handling speeds are decreasing only slowly. The computation versus communication dichotomy has not been fully resolved by large data store systems such as HDFS or Accumulo for exabyte-sized data sets.

2.5 Structured versus Unstructured Data

Translation between structured data with well-defined data definitions (often in tables) as stored in relational databases, and unstructured data (e.g., free text, graphics, multi-media, etc.) suitable for analytics can impede end-to-end processing performance. The emergence of non-relational, distributed, analytics-oriented databases such as NoSQL, MongoDB, SciDB and linked data DBs provides dynamic flexibility in representing and organizing information.

Unlike a data set, a *data source* has no beginning and no end. One begins collecting and continues to do so until one has enough data or runs out of patience or money or both. The data streams in with varied speed, frequency, volume, and complexity. The data stream may dynamically change in two ways: (1) the data formats change, necessitating changes in the way the analytics process the data, or (2) the data itself changes necessitating different analytics to process it. A complicating factor is the implicit assumption that the data streams are well-behaved and that the data arrive more or less in order. In reality, data streams are not so well-behaved and often experience disruptions and mixed-in data, possibly unrelated, to the primary data of interest. There is a need to rethink data stream processing to, perhaps, emphasize continuous analytics over discontinuous and distributed data streams.

2.6 Data Ownership

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the servers of Facebook, MySpace, and Twitter, it is not really owned by them (although they may contend so because of residency). Certainly, the "owners" of the pages or accounts believe they own the data. This dichotomy will have to be resolved in court. Kaisler, Money and Cohen [12] addressed this issue with respect to cloud computing as well as other legal aspects that we will not delve into here.

With ownership comes a modicum of responsibility for ensuring its accuracy. This may not be required of individuals, but almost certainly is so of businesses and public organizations. However, enforcement of such an assumption (much less a policy) is extremely difficult. Simple user agreements will not suffice since no social media purveyor has the resources to check every data item on its servers.

With the advent of numerous social media sites, there is a trend in big data analytics towards mixing of first-party, reasonably verified data, with public and third-part external data, which has largely not been validated and verified by any formal methodology. The addition of unverified data: compromises the fidelity of the dataset; may introduce non-relevant entities; and may lead to erroneous linkages among entities. As a result, the accuracy of conclusions drawn from processing this mixed data varies widely.

Table 5. Some Big Data Ownership Challenges

When does the validity of (publicly available) data expire?
If data validity is expired, should the data be removed from public-facing websites or data sets?
Where and how do we archive expired data? Should we archive it?
Who has responsibility for the fidelity and accuracy of the data? Or, it a case of user beware?

2.7 Compliance and Security

In certain domains, such as social media and health information, as more data is accumulated about individuals , there is a fear that certain organizations will know too much about individuals. For example, data collected in electronic health record systems in accordance with HIPAA/HITECH provisions is already raising concerns about violations of one's privacy. Developing algorithms that randomize personal data among a large data set enough to ensure privacy is a key research problem.

Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. For example, the State of Maryland became the first state to prohibit by law employers asking for Facebook and other social media passwords during employment interviews and afterwards.

International Data Corporation (IDC) coined the term "digital shadow" to reflect the amount of data concerning an individual which has been collected, organized, and perhaps analyzed, to form an aggregate "picture" of the individual. It is the information about you that is much greater than the information you create and/or release about yourself. A key problem is how much of this information – either original or derived – do we want to remain private?

Clearly, some big data must be secured with respect to privacy and security laws and regulations. IDC suggested five levels of increasing security [8]: privacy, compliance-driven, custodial, confidential, and lockdown. Further research is required to clearly define these security levels and map them against both current law and current analytics. For example, in Facebook, one can restrict pages to 'friends'. But, if Facebook runs an analytic over its databases to extract all the friend's linkages in an expanding graph, at what security level should that analytic operate? e.g., how many of an individual's friends should be revealed by such an analytic at a given level if the individual (has the ability to and) has marked those friends at certain security levels?

2.8 The Value of "Some Data" versus "All Data"

Not all data is created equal; some data is more valuable than other data – temporally, spatially, contextually, etc. Previously, storage limitations required data filtering and deciding what data to keep. Historically, we converted what we could and threw the rest away (figuratively, and often, literally).

Table 6. Some Big Data Compliance Challenges

What rules and regulations should exist regarding combining data from multiple sources about individuals into a single repository?
Do compliance laws (such as HIPAA) apply to the entire data warehouse or just to those parts containing relevant data?
What rules and regulations should exist for prohibiting the collection and storage of data about individuals – either centralized or distributed?
Should an aggregation of data be secured at a higher level than its constituent elements?
Given IDC's security categorization, what percentage of data should reside in each category? What mechanisms will allow data to move between categories?

The concept of "quantitative qualitative computation" suggests that we need new mechanisms for converting latent, unstructured text, image or audio information into numerical indicators to make them computationally tractable. With big data and our enhanced analytical capabilities, the trend is towards keeping everything with the assumption that analytical significance will emerge over time. However, at any point in time the amount of data we need to analyze for specific decisions represents only a very small fraction of all the data available in a data source and most data will go un-analyzed.

2.9 Distributed Data and Distributed Processing

The allure of hardware replication and system expandability as represented by cloud computing along with the MapReduce and Message Passing Interface (MPI) parallel programming systems offers one solution to these challenges by utilizing a distributed approach. Even with this approach, significant performance degradation can still occur

because of the need for communication between the nodes.

Table 7. Some Data Value Challenges

For a given problem domain, what is the minimum data volume required for descriptive, estimative, predictive and prescriptive analytics and decision modeling with a specified accuracy?
For a given data velocity, how do we update our data volume to ensure continued accuracy and support (near) real-time processing?
For a given problem domain, what constitutes an analytic science for non-numerical data?
"What if we know everything?" – What do we do next?

An open research question is, which big data problems are "MapReducible"? Specialized distributed algorithms, not necessarily based on the MapReduce or MPI paradigms, may be required to complete tasks to minimize the communication needed between the nodes. Finally, if distributed processing is viewed as an alternative, the overall reliability of the system will need to be increased to assure that no simple node or required communication fails. Both Google's and Hadoop's MapReduce systems have taken initial steps to ensure fault-tolerance for hardware, system software, and to some extent, for the algorithms in user application software.

3. Processing Big Data: Analytics Challenges

Processing big data is a major challenge, perhaps more so than the storage or management problem. There are many types of analytics: descriptive, estimative, predictive, and prescriptive, leading to various types of decision and optimization models. Some common business analytics are depicted in Figure 1. Kaisler [11] presents another decomposition of analytics into 16 categories based on the types of problems to be addressed, including econometric models, game theory, control theory, evolutionary computation, and simulation models. The new normal is agile, advanced, predictive analytics that adapt readily to changing data sets and streams and yield information and knowledge to improve services and operations across academia, industry, and government.

3.1 Scaling

A critical issue is whether or not an analytic process scales as the data set increases by orders of magnitude. Every algorithm has a "knee" – the point at which the algorithm's performance ceases to increase linearly with increasing computational resources and starts to plateau or, worse yet, peak, turn over, and start decreasing. Solving this problem requires a new algorithm for the problem, or rewriting the current algorithm to "translate" the knee farther up the scale. An open research question is whether for any given algorithm, there is a fundamental limit to its scalability. These limits are known for specific algorithms with specific implementations on specific machines at specific scales. General computational solutions, particularly using unstructured data, are not yet known. Table 8 gives some examples of analytic approaches that may not scale linearly. Simplistically, the processing of big data can be characterized in one of three ways as shown in table 9.

Figure 1. Examples of Types of Analytics

Table 8. Examples of Analytics That May Not Scale to Zettabytes

Machine Learning Techniques
Unstructured Text/Image/Video Analytics
Visualization
Cloud Computing
Data Mining
Graph and Mesh Algorithms
Joining Algorithms Across Structured Data Sets

Figure 2. The Needle in the Haystack

Table 9. Big Data Processing

1. Finding the needle in the haystack	The objective is to discover and extract the critical piece of information that provides the user with leverage in some situation
2. Turning straw into gold	There is no point solution, but a myriad of solutions depending on how the problem is presented. The objective is to select the best, but not necessarily the optimal, solution.
3. A hybrid of techniques	The objective is to converge on the answers at the same time we converge on the question(s) – an *outer-in* strategy, but this is likely to yield many answers.

3.2 Finding the Needle in the Haystack

This challenge focuses on finding the key data that provides leverage for decision-making within a problem space. A needle-in-a-haystack problem is one in which the right answer is very difficult to determine in advance, but very easy to verify once you know where the needle is [5]. Suppose we characterize it as finding the one right answer within a pool of 1,000,000 wrong answers. If the decision process is wrong 0.1% of the time, then of the 100 answers proposed, there are 101 answers, only one of which is 'correct". As Felten notes, any research area depending on this approach will suffer this problem, especially if it relies on statistical analysis. There are only two ways out of this problem: reduce the size of the haystack, or improve our search, analysis and decision-making procedures.

3.3 Turning Straw into Gold

This challenge focuses on processing a large set of discrete data points into high-valued data. Consider figure 3 below – a data visualization of Kenneth Freeman's Facebook Friends in December 2011 [7]. It represents a small subset of the hundreds of millions or so people using Facebook. As the number of edges emanating from "central" nodes increases, the overall mesh complexity increases nonlinearly. Finding subgraphs within graphs with particular sets of features may not be a linearly computationally tractable problem using standard graph-traversal and analysis algorithms.

One approach to solving the representation problem is to parse semistructured text and convert it to linked data using the Resource Description Framework (RDF) triple format. The explosion of resulting text is often on the order of 10:1 due to the use of RDF tags to identify components of an RDF structure. This translates the problem from processing semistructured text to finding relationships over a very large, real-world, partially connected mesh. Extracting mesh structural features is critical to identifying patterns and anomalies. Inference across mesh substructures is akin to 'guilt by association", e.g., if a person is a drug abuser, it is likely his friends are as well. Beliefs are propagated across the mesh resulting in a further explosion of data. [13].

Another challenge is the time-varying nature of very large graphs. Freeman's graph is a snapshot. Determining the change between two snapshots – either statically or continuously for a given interval – is a computationally explosive problem. This type of problem occurs frequently, but requires more intensive computation and new algorithms when applied in near-real-time analytics such as network attack monitoring.

To us, it is clear: *Gold Mining is not equal to Data Mining*! Different algorithms with greater reliance on reasoning (machine learning – symbolic, not statistical), computational social science, and domain-based analysis are essential to seeing the "big picture" in order to interpret and extract actionable patterns of behavior, meaning and nuggets of intelligence for informed decision-making.

3.4 A Hybrid of Techniques

Given a very large heterogeneous data set, a major challenge is to figure out what data one has and how to analyze it. Unlike the previous two sections, hybrid data sets combined from many other data sets hold more surprises than immediate answers. To analyze these data will require adapting and integrating multiple analytic techniques to "see around corners", e.g., to realize that new knowledge is likely to emerge in a non-linear way. It is not clear that statistical analysis methods, as Ayres [2] argues, are or can be the whole answer.

Nassim Taleb [19] addressed the potential for undirected and unpredicted effects arising from events that are outliers that lie outside the realm of regular expectations, because nothing in the past can convincingly point to their possibility. Such events often have an extreme impact – a "shock" to the system that can force new behaviors. Because we do not expect it, we cannot predict it. Thus, we can only try to explain what happened retrospectively. With more data, the likelihood of identifying such events rises and will force us to re-evaluate our estimative and predictive analytical tools.

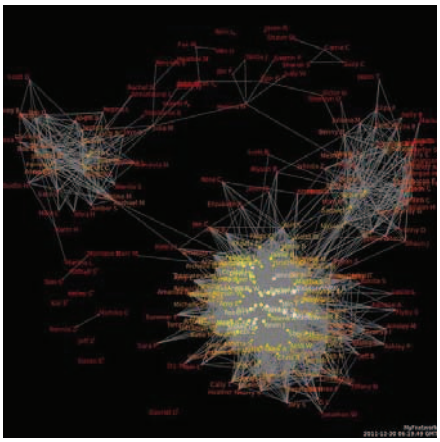

Figure 3. Freeman's Facebook friends

3.5 Know the World

With the (over)abundance of data available to us, a major research question is: can we model world systems, say, on the order of our desire to model and predict the weather? For example, can we forecast global political and/or economic stability at a given temporal interval? Underlying this challenge are questions of modeling natural science, social and cultural interactions at different scales; understanding how societies function and the causes of global unrest; and understanding how human societies produce and consume resources and the resource flows around the world. Such questions are important to transnational and global businesses determining where to allocate resources and invest in infrastructure.

Creating models in a computer is standard science. But, creating world-encompassing models (or, even, domain-encompassing) models is not yet feasible. Nevertheless, the challenge is to begin building such models that will allow us to comprehend systems at both the scope and granularity necessary to answer fundamental questions of cause and effect.

Consider the effects of natural disasters (such as the tsunami affecting Japan's decisions regarding nuclear power), economic system failure (the US housing and banking crisis), the recession/depression and the Arab Spring causing major evolution in governments, or technological innovation (such as the rise of social media). The key question for many decision makers – business, academia, government – is, "what does it all mean?" followed by "what is likely to happen next?"

These are all "wicked problems" as defined by Ritchey [16]. A *wicked problem* is one which has incomplete, contradictory and often changing requirements [17]. Because of the complex interdependencies of their elements, it is often difficult to recognize that one has achieved even a partial solution. Moreover, while attempting to solve a wicked problem, the partial solution often reveals or creates even more complex problems. The underlying systems are emergent, adaptive systems meaning that the system dynamically changes its behavior and its ability to adapt to new situations. Modeling these types of systems must continually evolve in order to support the decision-maker's wide area situation awareness.

4. Conclusions and Future Work

Big data is the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different".

"Big data" is just the beginning of the problem. Technology evolution and placement guarantee that in a few years more data will be available in a year than has been collected since the dawn of man. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years (perhaps 3-5) we are indeed facing the challenge of "big data becoming really big data".

We – as a global society – are evolving from a data-centric to a knowledge-centric community. Our knowledge is widely distributed and equally widely accessible. One program that is addressing this problem is The Federal Semantic Interoperability Community of Practice (SICoP) which supports an evolving model: Citizen-Centric Government – Systems That Know; Advanced Analytics – Systems That Learn; and Smart Operations – Systems That Reason. These systems will require big data. The data will not be stored in one or even a few locations; it will not be just one or even a few types and formats; it will not be amenable to analysis by just one or a few analytics; and there will not be just one or a few cross-linkages among different data elements. Thus, it is an exemplar of some of the issues we have addressed in this paper. Solving the issues and challenges addressed in this paper will require a concerted research effort – one which we expect to evolve over the next several years

This paper initiates a collaborative research effort to begin examining big data issues and challenges. We identified some of the major issues in big data storage, management, and processing. We also identified some of the major challenges – going forward – that we believe must be addressed within the next decade and

which will establish a framework for our Big Data minitrack in future HICSS sessions. Our future research will concentrate on developing a more complete understanding of the issues associated with big data, and those factors that may contribute to a need for a big data analysis and design methodology. We will begin to explore solutions to some of the issues that we have raised in this paper through our collaborative research effort.

5. References

[1] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[2] Ayres, I. 2007. *Supercrunchers*, Bantam Books, New York, NY

[3] Boyd, D. and K. Craford. 2011. "Six Provocations for Big Data", Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society"

[4] The Economist. 2010. "Data, Data Everywhere", (online edition, February 28)

http://www.economist.com/node/15557443

[5] Felten, E. 2010. "Needle in a Haystack Problems",

https://freedom-to-tinker.com/blog/felten/needle-haystack-problems/

[6] Fox, B. 2011. "Leveraging Big Data for Big Impact", Health Management Technology, http://www.healthmgttech.com/

[7] Freeman, K. 2011. http://en.wikipedia.org/wiki/File:Kencf0618FacebookNetwork.jpg

[8] Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC

[9] Jacobs, A. 2009. "Pathologies of Big Data", *Communications of the ACM*, 52(8):36-44

[10] JASON. 2008. "Data Analysis Challenges", The Mitre Corporation, McLean, VA, JSR-08-142

[11] Kaisler, S. 2012. "Advanced Analytics", CATALYST Technical Report, i_SW Corporation, Arlington, VA

[12] Kaisler, S., W. Money, and S. J. Cohen. 2012. "A Decision Framework for Cloud Computing", *45th Hawaii International Conference on System Sciences*, Grand Wailea, Maui, HI, Jan 4-7, 2012

[13] Kang, U. 2012. "Mining Tera-scale Graphs with MapReduce: Theory, Engineering, and Discoveries", PhD. Thesis, Computer Science, Carnegie-Mellon University, Pittsburgh, PA

[14] Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012

[15] Popp, R., S. Kaisler, et al. 2006. "Assessing Nation-State Fragility and Instability", *IEEE Aerospace Conference*, 2006, Big Sky, MT

[16] Ritchey, T. 2005. "Wicked Problems: Structuring Social Messes with Morphological Analysis", Swedish Morphological Society, http://www.swemorph.com/wp.html

[17] Rittel, H. and M. Webber. *1973. "*Dilemmas in a General theory of Planning*", in Policy Sciences*, Vol. 4, Elsevier Scientific, Amsterdam, the Netherlands, pp. 155-169

[18] Stonebraker, M. and J. Hong. 2012. "Researchers' Big Data Crisis; Understanding Design and Functionality", Communications of the ACM, 55(2):10-11

[19] Taleb, N. 2010. *The Black Swan: The Impact of the Highly Improbable*, Random House, New York, NY