Big Data: Immediate Opportunities and Longer Term Challenges

Jens Pohl
Vice President (Senior Technical Advisor)

Kym Jason Pohl
Principal Software Engineer

Tapestry Solutions (a Boeing Company), San Luis Obispo, California 93401, USA

Abstract

The transformation of words, locations, and human interactions into digital data forms the basis of trend detection and information extraction opportunities that can be automated with the increasing availability of relatively inexpensive computer storage and processing technology. Trend detection, which focuses on *what*, is facilitated by the ability to apply analytics to an entire corpus of data instead of a random sample. Since the corpus essentially includes all data within a population there is no need to apply any of the precautions that are in order to ensure the representativeness of a sample in traditional statistical analysis. Several examples are presented to validate the principle that with increasing scale data quality becomes less important.

Information extraction, which focuses on causality or *why*, is concerned with the automated extraction of meaning out of unstructured and structured data. This requires examination of the entities in the context of an entire document. While some of the relationships among the recognized entities may be preserved during extraction, the overall context of a document may not be preserved. The role of information representation in the form of an ontology, as a mechanism for facilitating the collection, extraction, organization, analysis, and retrieval of the semantic content of a sizeable data corpus is described with reference to past research findings.

Keywords

agents, analog, big data, correlation, data, digital, information, information extraction, mathematical models, ontology, prediction, statistics, trend

The Big Data Phenomenon

By far the most noticeable aspect of the Information Age is our increased connectivity. At the beginning of the 21st Century we have convenient access to an unprecedented amount of information over the global Internet. Utilizing one of several search engines we can readily find information about virtually any topic that we might be interested in. This has been made possible by digitized data in combination with electronic networks connecting widely distributed computer servers that can be conveniently accessed with wireless mobile devices. With the aid of satellites we are beginning to track not only ourselves but also the goods that we consume and the conveyances such as aircraft, ships, railcars, and vehicles that transport these goods to us. Often referred to as the 'Internet of Things', these tracking capabilities are starting to attract considerable commercial interest. In fact, there is now increasing concern that there is no standardized protocol for these devices to interact with each other (Reinhardt 2013),

Examples of this data deluge abound (Davenport et al. 2012; Page 2012; Hilbert et al. 2011). Military drones routinely collect several terabytes of data in a single day. The Walmart retail chain processes one million customer transactions every hour, while some real estate firms are collecting anonymous Global Positioning System (GPS) signals from millions of cars to help new home buyers determine their typical drive times to and from work at different times of day.

Twitter receives over 400 million tweets each day. Facebook handles 50 billion photographs from its user base, with 10 million new photographs being uploaded per hour. In addition, Facebook processes 3 billion comments per day. Google processes 24 petabytes of data per day, which is more than the total images in the US Library of Congress. At the same time its 800 million monthly YouTube users upload over an hour of video every second. On the stock exchanges seven billion shares are traded each day. Two thirds of these trades are conducted automatically by computer algorithms based on mathematical models that process huge amounts of data to predict gains.

A comprehensive study undertaken by Martin Hilbert at the University of Southern California to determine the total world-wide volume of stored data produced an estimate of 300 exabytes in 2007. This estimate included books, e-mail, photographs, paintings, music, video, video games, phone calls, mailed letters, car navigation systems, television, radio, and any other stored artifacts (Hilbert et al. 2011). It is interesting to note that in 2007 only 7% of the 300 exabytes of data were analog and 93% were digital. Just seven years earlier in 2000 the analog to digital ratio was very different, with only 25% stored in digital form. Today, in 2013, the amount of stored data is estimated to be 1,200 exabytes and only 2% of that is still in analog form (Mayer-Schönberger and Cukier 2013, 9).

Two Data Utilization Approaches

How can we best utilize this rapidly increasing availability and accessibility of data? We certainly do not want to be encumbered by an overwhelming amount of data. When we use search engines to find some information on the Internet, we typically receive more links to potential information sources (i.e., hits) than we care to look at. We have learned from experience that many of the hits will be disappointing because they do not lead to the information that we are seeking. Soon after the terrorist attacks on the World Trade Center towers in New York City in September 2001, much evidence was found that several warnings of a planned attack were contained in the routinely collected intelligence data, but had been overlooked. Naturally our expectations are that the data will be employed usefully to alert us in time to avoid adverse conditions, for planning purposes, to make better decisions in a timely manner, and to help us to predict trends that will give us a competitive advantage in our business and other endeavors.

In essence our expectations can be categorized into two kinds of data utilization approaches. The detection of trends that may have adverse or beneficial implications are focused on *what* is happening now or is likely to occur in the future, while the extraction of useful information and the interpretation of the meaning of data are focused on *why* something is occurring. While the pursuit of *why* is steeped deep in the scientific tradition, we have only recently become aware that there may be a great deal of value in knowing *what* without being able to explain its cause (i.e., *why*). The scientific method typically begins with the formulation of a hypothesis based on one or more theories. In this respect the hypothesis is really an abstract idea of *why*. This is followed by the collection of data in the form of observations that normally include measurements of some kind, taking great care that the measurements are as accurate as possible. We then proceed to perform a correlation analysis with the objective of verifying that the data do confirm the hypothesis and underlying theory. If the hypothesis appears to fail we will first test our data to confirm their accuracy and if the data survive this reexamination we will attempt to amend the initially established hypothesis and its underlying theories (Figure 1).

Historical Antecedents

The newfound global connectivity in combination with relatively low cost electronic storage and computing power is bringing with it an awareness that data have value beyond their normal role as the observations and measurements that form the basis of the traditional scientific method. On reflection the scientific method is largely a consequence of our human situatedness. We are situated in our environment and react to it with our senses. In our intrinsic desire to interpret and understand our surroundings we observe and measure its behavior. This not only satisfies our curiosity but also promotes our ability to adapt and survive. In fact it is often argued that our intellectual abilities are very much a product of the complexity and challenges posed by our environment.

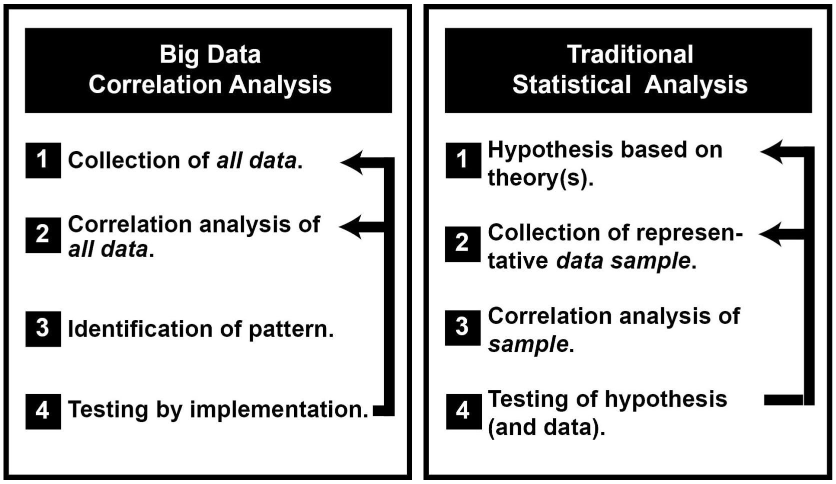
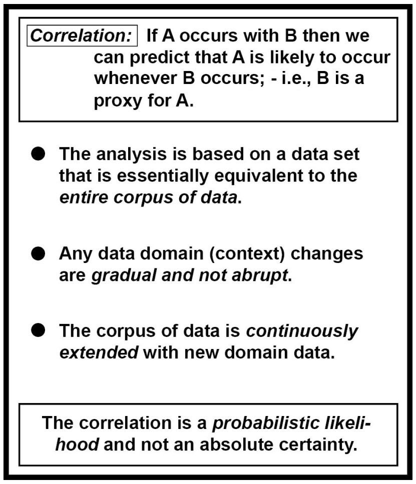

Big Data
Correlation Analysis
1 Collection of *all data*.
2 Correlation analysis of *all data*.
3 Identification of pattern.
4 Testing by implementation.
Traditional
Statistical Analysis
1 Hypothesis based on theory(s).
2 Collection of representative *data sample*.
3 Correlation analysis of *sample*.
4 Testing of hypothesis (and data).

Figure 1: *What* and *Why* methodologies



Correlation: If A occurs with B then we can predict that A is likely to occur whenever B occurs; - i.e., B is a proxy for A.

- The analysis is based on a data set that is essentially equivalent to the *entire corpus of data.*
- Any data domain (context) changes are *gradual and not abrupt.*
- The corpus of data is *continuously extended* with new domain data.

The correlation is a *probabilistic likelihood* and not an absolute certainty.

Figure 2: The correlation concept

As we organized ourselves into larger communities for reasons of safety, to assure a more dependable food supply, and to foster specialization for creating more functionally useful tools and products, effective governing and planning required the analysis of greater amounts of data. In fact, as the scope of the environment increased the amount of data that was required to be collected interpreted and understood increased disproportionally. We did not have the tools necessary to collect and analyze very large (i.e., population size) data sets. For example, in 1880 the US Government census took eight years to process, making the information that was collected partly obsolete before it became available[1]. At the time it was estimated that the 1890 census would take 13 years to process[2] (US Census Bureau 2012). Such difficulties encountered in dealing with large data sets led to the concept of sampling and the mathematics of predicting the characteristics of the entire corpus of data from a very much smaller sample.

The results provided by inferential statistics, the name given to this field of mathematics, can be surprisingly accurate as long as the sample is representative of the population (i.e., entire data set). To ensure this key requirement there was a need for a method of selecting a sample without any bias. This problem was solved by the adoption of a procedure by which samples could be collected randomly. Although random sampling has in the past and continues today to serve as

[1] The United States Constitution requires a population census to be taken every 10 years.

[2] Herman Hollerith introduced punched cards for the 1890 census to reduce processing to one year. This was a tremendous achievement and marked the beginning of data automation. However, the punched card process was still very labor intensive and expensive.

well, it is only a second best solution to the problem of not being able to collect and analyze the entire corpus of data. Today, due to connectivity and relatively inexpensive electronic storage and computing capabilities we have the ability to process very large data sets with billions of data elements and correlate these through the application of mathematical modeling techniques with historical data patterns to identify and predict trends in near real-time.

The *What* Approach to Trend Prediction

The concept underlying the *what* approach to trend prediction[3] can be expressed succinctly as follows: if A often occurs with B then we can predict that A is likely to occur whenever B occurs. This is also referred to as a correlation, with B becoming a proxy for A. It must be noted that such correlations do not predict the future with absolute certainty, but only with a probabilistic likelihood (Figure 2). What adds validity to the correlation is that it was based on the analysis of a data set that is essentially equivalent to the entire corpus of data. In support of this assumption it has been found in practice that accuracy increases with randomness rather than larger sample size (Kruskal et al. 1980). For example, a random sample of 1,100 observations on a yes/no question has typically only a 3% margin of error, while the margin of error of a random sample of 11,000 observations will be virtually the same. In this respect Big Data analysis overcomes some of the weaknesses of random sampling. With random sampling the accuracy or margin of error depends on achieving randomness when collecting the sample. Systemic biases such as election polling using landline telephones when an appreciable proportion of younger voters are likely to use cell phones are difficult to avoid.

Analyzing a large corpus of monetary transactions, Xoom, a company that specializes in international money transfers raised the alarm in 2011 when it noticed a slightly higher than average number of Discover credit card transactions originating from New Jersey (Economist 2012a). It found a pattern that should not have existed and that would likely have been missed by sampling. While each individual transaction looked legitimate further investigation by law enforcement discovered the involvement of a criminal group.

While Big Data analysis normally involves very large data sets, this is not always the case. For some time Sumo Wrestling in Japan had been under suspicion of match fixing, but no convincing evidence could be found. An analysis of 11 years of Sumo Wrestling involved only about 64,000 wrestling matches. Analysis of this corpus of data found a prevalence of match fixing in lower level bouts and not championship bouts. In this case the explanation was also immediately apparent. In Sumo Wrestling the ranking of wrestlers is based on the number of wins. Therefore, a win is much more important to a 5-5 (win-loss) wrestler than an 8-5 wrestler. The analysis showed that a 5-5 wrestler was 25% more likely to win then should have been expected and that when the same two wrestled the next time in a championship bout the previous winner invariably lost (Duggan et al. 2002).

Another example where the analysis of all data led to the identification of a phenomenon that would not have been recognized with sampling involved the examination of all mobile phone calls in a large region in Europe over a four-month period. The results indicated that the removal of callers who had many links within the community degraded the network, but the network did not collapse. However, the removal of callers who had many links outside of their community caused the network to disintegrate (Onnela et al. 2007).

[3] The *what* approach is used in this paper as being synonymous to the term *Big Data analysis*.

The fact that a very large corpus of data is likely to be imprecise due to irregularities (i.e., messiness) has little impact on the results of an analysis. The reason is that we are dealing with all of the data and not a sample that might no longer be representative of the total data. For example, in the late 1980s researchers at IBM applied statistical probability with a large corpus of data to machine translation from French to English with quite promising results. The data consisted of three million sentence pairs taken from Canadian parliamentary transcripts published in English and French. The data set was unusually precise due to the fact that the sentences originated from legal documents. In comparison, in 2006 Google researchers used a much larger and much messier data set with even better results. The data set comprised 95 billion English sentences taken directly from the Internet (Halery et al. 2009). Google's much larger corpus of multi-language translations, although of dubious quality, still provided better results than IBM's smaller corpus of much cleaner data. In other words, with increasing scale data accuracy becomes less important (Helland 2011).

This is quite contrary to the traditional mindset that has originated with sampling, where the implementation of error reducing strategies is a necessary and often costly undertaking. As an example, every month the US Bureau of Labor Statistics publishes the Consumer Price Index (CPI), which is used to calculate the inflation rate. This work involves hundreds of staff who collect around 80,000 prices from vegetables to clothing in 90 US cities; - an effort that takes a few weeks and costs around $250 million per year. Pursuing an alternative path, two researchers at the Massachusetts Institute of Technology (MIT) used software to collect half a million prices of goods sold every day[4]. By combining this messy data with mathematical analysis they were able to detect inflationary and deflationary trends within days. Their work has resulted in a commercial venture that now processes millions of product sales in 19 countries each day, with results that are available in near real-time and more accurate than the governmental statistics (Economist 2012b; Lowrey 2010).

As industry recognizes the business opportunities that can derive from the analysis of large data sets the role that data plays is changing from their primary most often transaction-based use to their potential future use. This change in mindset has several impacts. Businesses are increasingly looking upon data as a valuable asset that can be exploited for financial gain. This encourages the archiving of data for future uses that may not have been considered at the collection stage of the data. Some businesses are beginning to either license their data to other businesses that then use the data for entirely different purposes, or exploit the data themselves. For example: the analysis of GPS readings that indicate the location of delivery vehicles together with time of day, weather conditions, and traffic conditions will allow the optimization of routes and the optimization of delivery sequences; the analysis of past search queries on Google allowed the near real-time prediction of a flu epidemic in 2009 (Ginsburg et al. 2009; Dugar et al. 2012); analysis of past airline reservations data allowed the near real-time prediction of fare prices (Cukier 2010); and, the analysis of sensor data from machines and structures can facilitate preventative maintenance to avoid operational failure.

Data do not lose their value after use in the way physical commodities deteriorate after repeated use. Even though most data lose at least some of their value over time, they can often be reused for other purposes that are not impacted by this obsolescence factor. For example, while the houses may change in the images collected by Google's Street View cars, the GPS data do not change. Nevertheless, it is important to continuously cull the data that have changed unless they

[4] The researchers were Alberto Cavallo and Robert Rigobon and their work became the genesis of Price-Stats.

are used for the purpose of identifying change patterns. The obsolescence factor is mitigated at least to some degree by the natural increase in the volume of historical data over time. This provides the corpus of data with an automatic adjustment mechanism that is capable of accounting for at least gradual changes in context. However, the vulnerability to abrupt changes in context remains a legitimate criticism and potential weakness of Big Data analysis. Fortunately, most business and societal changes tend to occur over a period of years rather than weeks.

In the realm of Big Data analysis all kinds of data can become useful. Even data that are a byproduct of the actions and movements of persons have value. Previously considered to be throw-away data, the term assigned to such data today is *data exhaust* (Siegel 2013, 74). E-book readers (kindles) can capture how much time persons take to read one page, in what sequence they read a book, whether they complete a book or stop after the first few chapters, and so on. All of these secondary data have value that can be exploited in analyses that may be used for very different purposes and possibly in combination with other data sets. Similarly, in on-line educational courses the errors that students make in their exercises can be used to predict the probability that if students read a particular posting they will gain an understanding of the subject matter. Another highly innovative example is the creation of Google's spellchecker facilities. Rather than invest in the creation of a spellchecker dictionary, Google generated its spellchecker capabilities automatically and in multiple languages by analyzing the billions of misspellings of search words by millions of its users (Cukier 2010; Mayer-Schönberger and Cukier 2013, 112).

To summarize, Big Data analysis is based on the correlation between data sets that may appear to have no direct relationship. The correlation simply represents a statistical estimate of the likelihood of a direct relationship between A and B (Mayer-Schönberger 2013, 53). This leaves Big Data analysis open to the just criticism that two data sets may appear to be related when in fact they are just coincidental. The mitigating factors in favor of the validity of Big Data analysis are four-fold. First, the analysis is being applied to what is essentially the entire corpus of data, rather than a sample. This overcomes any uncertainty over whether the data set is in fact representative. Second, the calculated correlation value is a direct assessment of the strength of the pattern that has been identified in the data and not an estimate derived statistically from a subset of the data. Third, it is the nature of Big Data analysis that its conclusions are normally immediately applied. The resultant feedback serves as a continuous verification mechanism. Fourth, as long as the corpus of data is continuously extended with the new data that are being collected and there are no abrupt major changes in the environment in which the new data are being generated, then there can be confidence that the conclusions of the Big Data analysis continue to be relevant.

Finally the question arises, when is it more important to know *what* than *why* and when is it important to know *why*? Big Data analysis can be the initial step by identifying *what*, thereby allowing us to devise and implement plans for reacting to predicted events before undertaking the usually much more resource and time consuming task of gaining an understanding of *why*. In particular, knowing *what* without an understanding of *why* is valuable in time critical situations. Understanding the reasons *why* a certain correlation exists is more important in the longer term to guide strategic planning, validate existing theories, and build new theories.

The Problem with the *What* Approach of Big Data Analysis

While identifying and subsequently modeling the *What* trends that exist within a given community of data can be a valuable mechanism for predicting the future, the success of such models are traditionally predicated on the assumption that the future will for the most part mimic the past. However, in many real life situations this kind of stability is far from guaranteed. In the world of Big Data analysis the inevitable variability in dependent conditions is endeavored to be addressed by repeatedly regenerating the predictive model in its entirety based upon progressively more current data. While the computing power available today is typically sufficient to support relatively timely model turnarounds, this is nevertheless a brute force approach that is wasteful of computing resources.

Holding true in most areas of engineering, and Big Data analysis should be no exception, the conservation of resources is a philosophy that can separate a superficial expedient fix from a more effective solution that readily adapts to a variety of circumstances. Experience has shown that brute force strategies are progressively replaced with more elegant and efficient solutions. The remainder of this paper will hypothesize an approach that bridges the *What* objective of Big Data analysis to the *Why* objectives of context-based Information Extraction.

Grounding Predictive Models in Context

This section will propose a technological strategy for addressing the dilemma that the future does not necessarily mimic the past and, in fact, in many cases can deviate dramatically from both the past and the present. With this shortcoming firmly in mind, the proposed method attempts to infuse the predictive models powering Big Data analysis with the rich representational qualities of context-based modeling, or more specifically ontologies (Chandrasekaran et al. 1999). An ontology is a powerful form of context modeling and can be effectively used to express the concepts and entities inherent within a given domain (e.g., bank lending, building design, facilities management, logistical planning, etc.). Fundamental to such models is not only the expression of these discrete elements, but perhaps more importantly the capturing of the relationships that bind these elements together. The result is a contextual fabric that richly expresses the elaborate qualities and characteristics inherent within the particular domain. It is within this fabric that the proposed approach effectively grounds the predictive models produced by Big Data analysis. That is, the factors upon which predictive logic is based are captured and consequently expressed as rich context that can be readily understood by both the human user as well as machine-learning enabled software utilizing agent methodologies (Wooldridge and Jennings 1995). Such access to the contextual backdrop of dependent factors is consequently leveraged by the predictive logic powering Big Data analysis as a means of continually remaining in-tune with the evolution of the particular reality. As changes in the context occur, subsequent execution of connected logic will in turn produce revised predictions that more accurately reflect the current state of the reality. Figure 3 provides a conceptual architecture in support of grounding predictive logic within a contextual backdrop of dependent factors. In particular attention is drawn in Figure 3 to the use of reasoning-based technology as a more agile and adaptive alternative to decision trees that are often used in Big Data analysis (Siegel 2013, 111).

Equipping Contextually-Grounded Predictive Models with Self-Synchronization

To fully exploit the context-based method described above, the predictive model and contextual fabric in which its dependent factors are expressed can be engineered to be self-synchronizing. This requires the engineering of two distinct mechanisms. The first of these mechanisms deals with the contextual fabric in relation to the data with which it is progressively populated. Among the set of data producers that provide this content there are likely to exist opportunities to instrument such feeds with the ability to automatically trigger updates to relevant portions of the associated context model, whenever new content becomes available. In this manner, the modeled context upon which predictive logic operates is automatically aligned with the evolving state of the target reality.

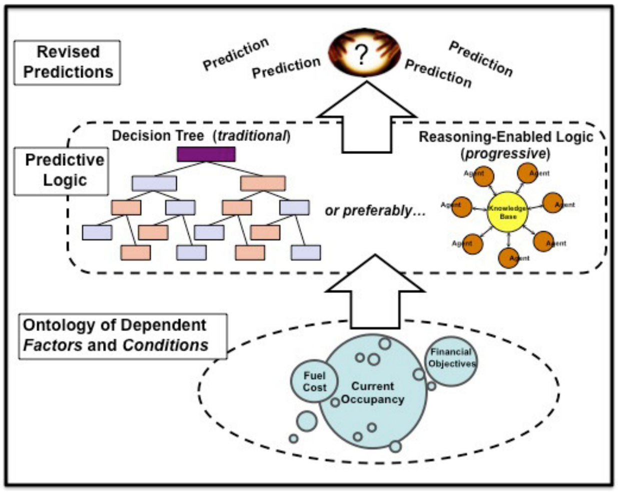


Figure 3: Conceptual architecture

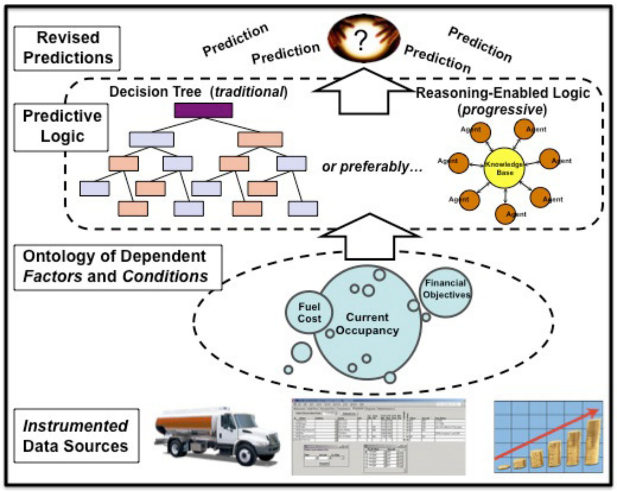


Figure 4: Extended architecture with instrumentation and synchronization

At this point it should be noted that the contextual-synchronization described thus far is focused on the values that populate a primarily static expression (or blueprint) of the various factors and conditions upon which the predictive logic is based. Although by no means a trivial task, this capability could in theory be extended to include the adjustment of the blueprint itself. In other words, when realigning modeled context within highly dynamic domains it may be advantageous, albeit even necessary, to modify not only a concept's values, but the very manner in which the concept is expressed. Achieving this level of synchronization relies on the work being performed in the area of dynamic ontologies (Zhou 2007) and is presently beyond the scope of the solution being presented. Yet, extending the solution in this manner would be an interesting and highly challenging area of research and could result in significant capabilities in the face of a somewhat unpredictable domain that embodies multiple, progressively changing perspectives (e.g., artistic value, etc.).

Regardless of the level of dynamics achieved in updating the context model portion of the system, the next step in the synchronization approach is to equip the predictive logic component with the ability to automatically respond to changes occurring within the underlying context. Achieving this step will require the ability for the predictive logic to be responsive (i.e., listen) to the relevant components of the context model upon which it depends. Once modifications to dependent factors and conditions are detected, such changes would in turn stimulate the execution of related logic resulting in the automatic revision of corresponding predictions. This process could continue in a self-governing fashion or in situations where specific levels of oversight are warranted, to be integrated into a more elaborate control system. Figure 4 depicts in conceptual terms an extended form of the original architecture that is proposed for providing the more automated inter-layer connections discussed within this section.

Combining the grounding of predictive logic in a contextual backdrop with the automatic refreshing capabilities of self-synchronization, changes in market trends, resource utilization, or any other aspects upon which Big Data predictions are made could be automatically detected and responded to in a manner that continually adjusts resulting predictions to better reflect current circumstances.

Concluding Remarks

In his keynote presentation at the Strata 2012 conference entitled *A Big Data Imperative: Driving Big Action* Google's Avinash Kaushik opened with a quotation that he attributed to a Kenyan farmer: "Information is powerful, it's what we use it for that will define us" (Minelli et al. 2013, 169). In this regard the most significant initial impact of Big Data analysis is likely to be in the commercial arena. By analyzing very large corpuses of data it will be possible to identify trends that can be readily exploited for commercial gain. While some of these trends could have been detected through standard statistical analysis based on random sampling, Big Data analysis will provide insights on a scale that has been neither contemplated nor practical in the past. The beginnings of this analysis are already apparent when we receive e-mail messages from retailers with offers of products that are related to our recent Internet searches and on-line purchases. Much of the Big Data analysis will be directed to uncovering patterns and correlations that lead to entirely new services and products or the enhancement of existing services to provide a competitive edge.

Clearly, data will be increasingly valued as a desirable asset rather than an encumbrance. In particular the combination of multiple data corpuses will lead to the discovery of apparent relationships that are not only unforeseen but questionable. This is why validation of the apparent correlations will become as important as the initial identification of the pattern or trend. The context-based modeling approach proposed in the second part of this paper is presented as a possible validation mechanism. By expressing the notions and entities of the particular domain in which an apparent relationship has been detected in machine-readable form, the validation process can be integrated with the analysis process. In this way the validation becomes an integral part of the Big Data analysis, with the objective of preventing the analysis from straying into areas that are not supported by the context model.

In this regard the self-synchronizing capability alluded to earlier in this paper is intended to ensure that data changes over time are captured in the context model on a near real-time basis. This will require mechanisms for the automatic modification of the context model (i.e., ontology) itself. While the dynamic generation and extension of ontologies has been a research area of intensive interest for the past two decades, progress to date has been rather disappointing. Research findings have shown that the automated extraction of context within a given corpus of data is a complex undertaking. For this reason existing context-based software systems typically incorporate static ontologies that are manually updated whenever extensions or revisions are required. Even in the case of research projects that focus on the automated extraction of context it is common practice to start off with a predefined high level ontology to guide the progressive refinement and extension attempts (Assal et al. 2013).

Apart from these technical challenges there is the social issue of privacy that is starting to raise concerns. Whether or not these concerns are legitimate is yet to be determined. A clear distinction can be drawn between the analysis of data that are directly or indirectly traceable to individual persons such as mobile telephone calls, and analyses that are undertaken solely for the purpose of predicting trends such as the spread of a virus that could cause an epidemic. Even though in the first instance the collection of the telephone data may be restricted to billing information such as telephone number and duration of call, subsequent use of the data for homeland security purposes may lead to much more invasive actions. For example, counterterrorism intelligence may yield the telephone number of a person who is suspected of being involved in terrorist activities. By querying the corpus of telephone data it is possible to immediately identify all persons who have either received or placed calls to that number. This implies a potential criminal association that may be completely unwarranted. On the other hand, it can be argued that the availability of the telephone data for exactly this purpose is essential for the protection of the public from terrorist threats.

The debate on privacy will increase in intensity over the next several years as the exploitation of data pervades virtually all human activities.

References

Assal H., K. Pohl, F. Kurfess, E. Schwarz and J. Seng (2013); 'Enhancing Information Extraction with Context and Inference: The ODIX Platform'; in Ordonez de Pablos P., H. Nigro, R. Tennyson, S. Gonzalez Cisaro and W. Karwowski (eds.), 'Advancing Information Management Through Semantic Web Concepts and Ontologies', Chapter 11, Information Science Reference, IGI Global, Hershey, Pennsylvania (pp. 195-220).

Ayres I. (2007); 'Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart'; Bantam Books, New York, New York.

Berger A., P. Brown, V. Pietra, S. Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz and L. Ures (1994); 'The Candide System for Machine Translation'; Proceedings of the 1994 ARPA Workshop on Human Language Technology, DARPA, Washington, DC.

Breslin J., A. Passant and S Decker (2009); 'The Social Semantic Web'; Springer Verlag, Berlin, Germany.

Chandrasekaran B, J. Josephson and V. Benjamins (1999); 'What are Ontologies, and Why Do We Need Them?'; IEEE Intelligent Systems, January-February.

Collins K. (2012); 'Use High-Performance Analytics to Unlock the Power of Big Data'; SAS High Performance Analytics, sas.com/hpa-insights, SAS Institute Inc.

Cukier K. (2010); 'Data, Data Everywhere'; The Economist Special Report, 27 February (pp. 1-14).

Davenport T., P. Bart and R. Bean (2012); 'How Big Data is Different'; Sloan Review, July 30 (pp. 43-6).

Davenport T., J. Harris and R. Morison (2010); 'Analytics at Work: Smarter Decisions, Better Results'; Harvard Business Press, Boston, Massachusetts.

Dugas A., Y. Hsieh, S. Levin, J. Pines, D. Mareiniss, A. Mohareb, C. Gaydos, T. Perl and R. Rothman (2012); 'Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics'; http://cid.oxfordjournals.org/content/early/2012/01/02/cid.cir883.full, CID Advanced Access, January.

Duhigg C. (2012); 'The Power of Habit: Why We Do What We Do in Life and Business'; Random House, New York, New York.

Duggan M. and S. Levitt (2002); 'Winning Isn't Everything: Corruption in Sumo Wrestling'; American Economic Review, 92 (pp. 1594-1605).

Economist (2012a); ‘Special Report: International Banking’; Special Report, John Rosenthal, 19 May (pp. 7-8).

Economist (2012b); ‘Official Statistics: Don’t lie to me, Argentina’; The Economist, 25 February, www.economist.com/node/21548242.

FCW (2012); ‘Big Data’; Federal Computer Week (FCW) Research Report, FCW Custom Report, www.FCW.com/BigDataResearch.

Flanagan G. (2011); ‘Information Extraction: A Survey of the State of the Art’; Collaborative Agent Design Research Center (CADRC) White Paper, Cal Poly, San Luis Obispo, California.

Frei P., A. Poulsen, C. Johansen, J. Olsen, M. Stedig-Jessen and J. Schüz (2011); ‘Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study’; British Medical Journal, vol. 343, October, www.bmj.com/content/343/bmj.d6387.

Ginsberg J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski and L. Brilliant (2009); ‘Detecting Influenza Epidemics Using Search Engine Query Data’; Nature 457 (pp. 1012-14).

Halevy A., P. Norvig and F. Pereira (2009); ‘The Unreasonable Effectiveness of Data’; IEEE Intelligent Systems, March-April (pp. 8-12).

Helland P. (2011); ‘If You Have Too Much Data Then Good Enough is Good Enough’; Communications of the ACM, June (pp. 40-1).

Hilbert M. and P. Lopez (2011); ‘The World’s Technological Capacity to Store, Communicate, and Compute Information’; Science, April (pp. 60-5).

Janert P. (2011); ‘Data Analysis With Open Source Tools: A Hands-On Guide for Programmers and Data Scientists’; O’Reilly Media, Sebastopol, California.

Kruskal W. and F. Mosteller (1980); ‘Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939’; International Statistical Review, 48 (pp. 169-195).

Lane A. (2013); ‘Security Implications of Big Data Strategies’; Dark Reading, Information Week reports, Reports.InformationWeek.com, March (pp. 12).

Lord Kelvin (1883); ‘Electrical Units of Measurement’; Lecture presented on 3 May 1883, quoted in Stellmann J. (1998), Encyclopaedia of Occupational Health and Safety (pp. 1992).

Lowrey A. (2010); ‘Economist’s Programs are Beating U.S. at Tracking Inflation’; Washington Post, www.washingtonpost.com/wp-dyn/content/article/2010/12/25/AR2010122502600.html, 25 December.

Mayer-Schönberger V. and K. Cukier (2013); ‘Big Data: A Revolution that Will Transform How We Live, Work, and Think’; Houghton Mifflin Harcourt, New York, New York.

Mehta A. (2011); ‘Big Data: Powering the Next Industrial Revolution’; Tableau Software White Paper, www.tableausoftware.com/learn/whitepapers/big-data-revolution.

Minelli M, M. Chambers and A Dhiraj (2013); ‘Big Data, Big Analytics’; Wiley, Hoboken, New Jersey (pp. 169-174).

Onnela J, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész and A.Barabási (2007); ‘Structure and Tie Strengths in Mobile Communication Networks’; Proceedings of the National Academy of Sciences, 104(18) (pp.7332-6).

Page L. (2012); 'Update from the CEO'; http://investor.google.com/corporate/2012/ceo.letter.html, Google, April.

Reinhardt H. (2013); 'IoT Tech Talk Follow-Up'; www.layer7tech.com/blogs/index.php/author/holger/.

US Census Bureau (2012); 'The Hollerith Machine'; Online History, www.census.gov/history.

Siegel E. (2013); 'Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die'; Wiley, Hoboken, New Jersey.

Wooldridge M. and N. Jennings (1995); 'Intelligent Agents: Theory and Practice'; The Knowledge Engineering Review, Vol. 10(2) (pp. 115-152).

Zhou L. (2007); 'Ontology Learning: State of the Art and Open Issues'; Springer-Verlag, Berlin, Germany.