11-2013

Semantics-Empowered Big Data Processing with Applications

Krishnaprasad Thirunarayan
Wright State University - Main Campus, t.k.prasad@wright.edu

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Semantics-empowered Big Data Processing with Applications

Krishnaprasad Thirunarayan and Amit Sheth
Kno.e.sis : Ohio Center of Excellence in Knowledge-enabled Computing
Department of Computer Science and Engineering, Wright State University, Dayton, OH-45435.
{tkprasad, amit}@knoesis.org

Abstract

We discuss the nature of Big Data and address the role of semantics in analyzing and processing Big Data that arises in the context of Physical-Cyber-Social Systems. We organize our research around the Five Vs of Big Data, where four of the Vs are harnessed to produce the fifth V - value. To handle the challenge of *Volume*, we advocate semantic perception that can convert low-level observational data to higher-level abstractions more suitable for decision-making. To handle the challenge of *Variety*, we resort to the use of semantic models and annotations of data so that much of the intelligent processing can be done at a level independent of heterogeneity of data formats and media. To handle the challenge of *Velocity*, we seek to use continuous semantics capability to dynamically create event or situation specific models and recognize relevant new concepts, entities and facts. To handle *Veracity*, we explore the formalization of trust models and approaches to glean trustworthiness. The above four Vs of Big Data are harnessed by the semantics-empowered analytics to derive *Value* for supporting practical applications transcending physical-cyber-social continuum.

Introduction

Physical-Cyber-Social Systems (PCSS) (Sheth et al, 2013) are a revolution in sensing, computing and communication that brings together a variety of resources. The resources can range from networked embedded computers and mobile devices to multimodal data sources such as sensors and social media. The applications can span multiple domains such as medical, geographical, environmental, traffic, behavioral, disaster response, and system health monitoring. The modeling and computing challenges arising in the context of PCSS can be organized around the Five Vs of Big Data (volume, variety, velocity, veracity and value), which align well with our research efforts that exploit semantics, network and statistics-empowered Web 3.0.

Characteristics of the Big Data Problem

We discuss the primary characteristics of the Big Data problem as it pertains to the Five Vs. (The first three were originally introduced by Doug Laney of Gartner.)

Volume

The sheer number of sensors and the amount of data reported by sensors is enormous and growing rapidly. For example, 25+ billion sensors have been deployed and about 250TB of sensor data are generated for a NY-LA flight on Boeing 737[1]. Parkinson's disease dataset[2] that tracked 16 people (9 patients + 7 control) with mobile phone containing 7 sensors over 8 weeks is 12 GB in size. However, availability of fine-grained raw data is not sufficient unless we can analyze, summarize or abstract them in actionable ways. For example, from a pilot's perspective, the sensors data processing should yield insights about whether the jet engine and the flight control surfaces are behaving normally or is there cause for concern? Similarly, we should be able to measure the symptoms of Parkinson's disease using sensors on a smartphone, monitor its progression, and synthesize actionable suggestions to improve the quality of life of the patient? Cloud computing infrastructure can be deployed for raw processing of massive social and sensor data. However, we still need to investigate how to effectively translate large amounts of machine-sensed data into a few human comprehensible nuggets of information necessary for decision-making. Furthermore, privacy and locality considerations require moving computations closer to the data source, leading to powerful applications on resource-constrained devices. In the latter situation, even though the amount of data is not large by normal standards, the

[1] http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/
[2] https://www.michaeljfox.org/page.html?parkinsons-data-challenge

resource constraints negate the use of conventional data formats and algorithms, and instead necessitate the development of novel encoding, indexing, and reasoning techniques (Henson et al, 2012a).

The volume of data challenges our ability to process them. First, it is difficult to abstract fine-grained machine-accessible data into coarse-grained human comprehensible form that summarizes the situation and is actionable. Second, it is difficult to scale computations to take advantage of distributed processing infrastructure and, where appropriate, exploit reasoning on mobile devices.

Variety

PCSS generate and process a variety of multimodal data using heterogeneous background knowledge to interpret the data. For example, traffic data (such as from 511.org) contains numeric information about vehicular traffic on roads (e.g., speed, volume, and travel times), as well as textual information about active events (e.g., accidents, vehicle breakdowns) and scheduled events (e.g., sporting events, music events) (Anantharam et al, 2013). Weather datasets (such as from Mesowest) provide numeric information about primitive phenomena (e.g., temperature, precipitation, wind speed) that are required to be combined and abstracted into human comprehensible weather features in textual form. In medical domains (e.g., cardiology, asthma, and Parkinson's disease), various physiological, physical and chemical measurements (obtained through on-body sensors, blood tests, and environmental sensors) and patients' feedback and self-appraisal (obtained by interviewing them) can be combined and abstracted to determine their health and well-being. The available knowledge captures both qualitative and quantitative aspects. Such diverse knowledge when integrated can provide complementary and corroborative information (Sheth and Thirunarayan, 2012). Geoscience datasets, and materials and process specifications used for realizing Integrated Computational Materials Engineering[3] (ICME) and Materials Genome Initiative[4] (MGI), exhibit lot of syntactic and semantic variety[5] (Thirunarayan et al, 2005).

The variety in data formats and the nature of available knowledge challenges our ability to integrate and interoperate with heterogeneous data.

[3] http://www.nap.edu/catalog.php?record_id=12199
[4] http://www.whitehouse.gov/mgi
[5] http://earthcube.ning.com/

Velocity

Handling of sensor and social data streams in PCSS requires online (as opposed to offline) algorithms to (1) efficiently crawl and filter relevant data sources, (2) detect and track events and anomalies, and (3) collect and update relevant background knowledge. For instance, Wikipedia event pages can be harnessed for relevance ranking of Twitter hashtags. The semantic similarity of a hashtag to an event can be determined by leveraging the background knowledge in the corresponding event page on Wikipedia. Specifically, we have used the entities that co-occur with the tweets containing the hashtag and the entities present in the Wikipedia event page to determine the relevance ranking (Kapanipathi et al, 2013). Similarly, entities can be tracked in the context of a natural disaster or a terror attack. For example, during Hurricane Sandy, tweets indicated possible flooding of a subway station, whose location obtained using open data[6] helped identify sensors for real-time updates. On the other hand, raw speed of interaction is critical for financial market transactions.

The rapid change in data and trends challenges our ability to process them. First, it is difficult to filter and rank the relevant data incrementally and in real-time. Second, it is difficult to cull and evolve the relevant background knowledge.

Veracity

PCSS receive data from sensors subject to the vagaries of nature (some sensors may even be compromised), or from crowds with incomplete information (some sources may even be deceitful). Statistical methods can be brought to bear to improve trustworthiness of data in the context of homogeneous sensor networks, while semantic models can be used for heterogeneous sensor networks (Thirunarayan et al, 2013). For instance, for applications that involve both humans and sensors systems, it is crucial to have trustworthy aggregation of all data and control actions. The 2002 Überlingen mid-air collision[7] occurred because the pilot of one of the planes trusted the human air traffic controller (who was ill-informed about the unfolding situation), instead of the electronic TCAS system (which was providing conflicting but correct course of action to avoid collision). Similarly, we were unable to identify and resolve inconsistencies, disagreements and changes in assertions in the aftermath of the rumor about Sunil Tripathi being a potential match for the grainy surveillance photographs of Boston Marathon bomber[8]. These examples illustrate the difficulties we face while making decisions based on conflicting data from different sources.

[6] https://nycopendata.socrata.com
[7] http://en.wikipedia.org/wiki/uberlingen_mid-air_collision
[8] http://bit.ly/1dFi5b3

The determination of veracity of data challenges our ability to detect anomalies and inconsistencies in social and sensor data. Reasoning about trustworthiness of data is also difficult. Fortunately, the latter can exploit temporal history, collective evidence, and context for conflict resolution.

Value

Semantics-empowered analytics of big data can be harnessed to deal with the challenges posed by volume, velocity, variety and veracity to derive value. A key aspect in transforming PCSS to provide actionable information is the construction and application of relevant background knowledge needed for data analytics and prediction. For example, a hybrid of statistical techniques and declarative knowledge can benefit leveraging sensor data streams in a variety of applications ranging from personalized healthcare, to reducing readmission rates among cardiac patients, to improving quality of life among asthmatic patients. Ultimately, the analysis of environmental, medical, system health, and social data enables situational awareness, and derivation of nuggets of wisdom for action.

Extracting value using data analytics on sensor and social data streams challenges our ability to acquire and apply knowledge from data and integrate it with declarative domain knowledge for classification, prediction, decision making, and personalization.

Role of Semantics in Big Data Processing

We discuss examples of our early research in developing semantics-empowered techniques to address the Big Data problem organized around the 5Vs from Kno.e.sis' active multidisciplinary projects[9] (Thirunarayan and Sheth, 2013), while realizing that it will require a longer survey paper to review research being pursued by our community at large.

Addressing Volume: Semantic Scalability

Semantics-based models address the volume challenge by relating how high-level human-sensible abstractions can manifest in terms of low-level sensor observations. This enables filtering of data by determining what to focus on and what to ignore, promoting scalability. Thus, the key to handling volume is to change the level of abstraction for data processing to information that is meaningful to human activity, actions, and decision making. We have called this *semantic perception* (Henson et al, 2013) (Sheth, 2011), which involves semantic integration of large amounts of heterogeneous data and application of perceptual inference using background knowledge to abstract data and derive actionable information. Our work involving Semantic Sensor Web (SSW) and IntellegO (Henson et al, 2012), which is a model of machine perception, integrates both deductive and abductive reasoning into a unified semantic framework. This approach not only combines and abstracts multimodal data but also seeks relevant information that can reduce ambiguity and minimize incompleteness, a necessary precursor to decision and action. Specifically, our approach uses background knowledge, expressed via cause-effect relationships, to convert low-level data into high-level actionable abstractions, using cyclical perceptual reasoning involving predictions, discrimination, and explanation. For instance, in the medical context, symptoms can be monitored using sensors, and plausible disorders that can account for them can be abduced. However, what heart failure patients will benefit from are suggestions such as whether the condition is as normally expected, or requires a call/visit to a nurse/doctor, or hospitalization. The first example below can be formalized using our approach with demonstrable benefits, while the subsequent examples require research into high-fidelity models and human mediation for fruition.

(1) *Weather use case*: This application involves determining and tracking weather features from weather phenomenon, with potential for tasking sensors if additional information is necessary. We have developed Semantically-enabled Sensor Observation Service (SemSOS) that leverages semantic technologies to model the domain of sensors and sensor observations in a suite of ontologies, adding semantic annotations to the sensor data, and reasoning over them (Henson et al, 2009). Specifically, we have extended an open source SOS implementation, 52North, with our semantic knowledge base. For weather use case, we have used rules provided by NOAA to map primitive machine-sensed weather data (e.g., wind speed, temperature, precipitation) to human comprehensible weather features (e.g., blizzard, flurry). SemSOS, provides the ability to query high-level knowledge of the environment as well as low-level raw sensor data using SPARQL. The task of abstracting low-level sensor data to high-level features as explanation is abductive in nature, while disambiguation among multiple explanations requires deduction and selectively seeking additional data.

(2) *Health care use case (Diagnosis, Prevention and Cure):* These applications involve determining disorders afflicting a patient -- their degree of severity and progression -- by monitoring symptoms via sensors and mobile devices. They can also be augmented with patient reported observations (e.g., about feeling giddy or tired or depressed that cannot always be ascertained through physical/chemical means), and/or laboratory test results.

Semantic perception involves abstracting machine-sensed data into coarse-grained form (e.g., using average,

---

[9] http://knoesis.org/projects/multidisciplinary

peak, rate of change, duration), and extracting human comprehensible features by integrating them. This approach requires construction of suitable domain models and hybrid abductive/deductive reasoning framework, which is our current research focus. Abduction generates abstractions of sensor data as explanations. Deduction can be used to discriminate among multiple explanations by predicting and seeking confirmation by tasking appropriate sensors. In general, this iterative and interleaved use of abduction and deduction can be used to eventually generate the minimum explanation that can be used to determine action. For example, abduction can be applied to weather phenomena data (e.g., precipitation and temperature) to determine weather features (e.g., flurry and blizzard) that can be further disambiguated by making additional observations (e.g., wind speed), before taking action. Similarly, abduction can be applied to observed symptoms to determine candidate diseases that can then be disambiguated using the results of additional tests, before one can determine medications and regimen. For Parkinson's disease, data from accelerometer, GPS, compass, and microphone, etc. are converted into human perceived features such as tremors, walking style, balance, and slurred speech, to diagnose and monitor disease progression, and recommend control options. For heart failure patients, weight change, heart rate, blood pressure, oxygen level, etc. are combined and translated into risk-level for hospital readmission (to minimize preventable readmissions). For asthma patients, data from environmental and physiological sensors, and personal feedback about wheezing, coughing, and sleeplessness, etc. can be used to recommend prevention strategies, treatment levels, and control options. The continuous monitoring of a patient and their surroundings, and the associated domain models can be used to determine actionable causes for the symptoms rather than just educated guesses. In general, patients suffering from chronic diseases can benefit from suggestions for avoiding aggravating factors to improve the quality of life, and for enhancing adherence/compliance to prescribed treatment or control options.

Some specific research goals to be pursued to realize semantics-based analytics (that also overlap with approaches to meet the variety challenge) include: (1) *Development and codification of high-fidelity background knowledge for processing sensor data streams using expressive semantic representations.* For example, in the realm of health care, symptoms and disorders are complex entities with complicated interactions. The acceptable and desirable thresholds for various monitored parameters depend on co-morbidity, especially due to chronic conditions. Any representation must provide the necessary expressivity to accurately formalize the reality of the situation. (2) *Using contextual information and personalization*. An accurate interpretation of data is based on spatio-temporal-thematic contextual knowledge. In medical scenarios, effective treatment also requires personalization on patient's historical data and clinician prescribed current protocol (e.g., maintain BP at higher than what is normal for NIH specific guidelines) such as what is in Electronic Medical Records (EMR). (3) *Effective summarization and justification of recommended action*. One of the problems resulting from indiscriminate sensing and logging of observed data due to ubiquity of mobile computing, wireless networking and communication technologies is that we are drowned in the noise[10]. The ability to determine the nature and severity of a situation from a glut of data, and to issue an informative alert or summary that is accessible to and actionable by the end users is a critical challenge we are addressing in the kHealth project. (4) *Efficient perceptual reasoning on resource-constrained devices*. In order to provide "intelligent computing at the edge", we need techniques to collect the data at the edge, intelligently reason with them using background knowledge, and return the essence. For example, this is required to address privacy concerns, need for timely and ubiquitous access to data, using wireless mobile devices. Its realization will also spur use of innovative and specialized inference techniques on resource-constrained devices as described in the next section (Henson et al, 2012a).

An Efficient Approach to Semantics-based Machine Perception in Resource-Constrained Devices

We employed OWL to formally define the two inference tasks needed for machine perception – *explanation* and *discrimination* (Henson, et al, 2011). Unfortunately, this declarative specification does not run as is on extant mobile devices using a standard reasoner as its memory and time requirements far exceed the capacity provided by the popular configurations of the mobile devices. This hurdle has been overcome using bit-vector encoding based algorithms for explanation and discrimination tasks as summarized below (Henson, et al, 2012a).

Semantic Sensor Ontology: The SSN ontology serves as a foundation to formalize the semantics of perception. An observation (`ssn:Observation`) is defined as a situation that describes an observed feature, an observed property, the sensor used, and a value resulting from the observation (note: prefix `ssn` is used to denote concepts from the SSN ontology). A feature (`ssn:Feature`) is an object or event in an environment, and a property (`ssn:Property`) is an observable attribute of a feature. For example, in cardiology, elevated blood pressure is a

[10]http://www.cio.co.uk/insight/r-and-d/internet-of-everything-tweeting-tweets/

property of the feature Hyperthyroidism. In SSN, knowledge of the environment is represented as a relation (`ssn:isPropertyOf`) between a property and a feature. To enable integration with other ontological knowledge on the Web, this knowledge is aligned with concepts in the DOLCE Ultra Lite ontology[11]. Figure 1 provides a simple example from the cardiology domain.

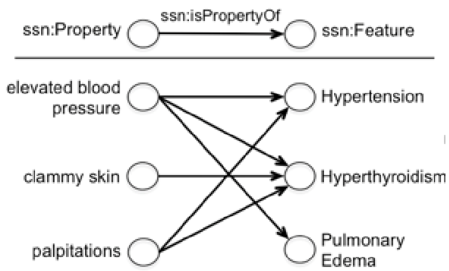

Figure 1. Bipartite graph representation of a simple cardiology knowledge base

Semantics of Machine Perception: A feature is said to *explain* an observed property if the property is related to the feature through an `ssn:isPropertyOf` relation. In Figure 1, Hyperthyroidism explains the observed properties elevated blood pressure, clammy skin, and palpitations. Since several features may be capable of explaining a given set of observed properties, explanation is most accurately defined as an abductive process. For example, the observed properties, elevated blood pressure and palpitations, are explained by the features Hypertension and Hyperthyroidism. A property is said to *discriminate* between a set of features if its presence can reduce the set of explanatory features. In Figure 1, the property clammy skin discriminates between the features, Hypertension and Hyperthyroidism. For a detailed formal description of explanation and discrimination tasks in OWL, see (Henson, et al, 2012a).

Efficient Algorithms for Machine Perception: To implement machine perception on resource-constrained devices, we developed bit-vector based algorithms for explanation and discrimination, satisfying a *single-feature assumption* (i.e., one feature is sufficient to account for all the observed properties).

To preserve the ability to share and integrate with knowledge on the Web, lifting and lowering mappings between the semantic representations (in RDF) and bit vector representations were developed. An environmental knowledge base is represented as a bit matrix KB_{BM}, with rows representing properties and columns representing features. $KB_{BM}[i][j]$ is set to 1 (true) *iff* the property p_i is a property of feature f_j (i.e., there exists a `ssn:isPropertyOf(p_i,f_j)` relation). Observed properties are represented as a bit vector $OBSV_{BV}$, where $OBSV_{BV}[i]$ is set to 1 *iff* `ObservedProperty(p_i)` holds (i.e., property p_i has been observed). Explanatory features are represented as a bit vector $EXPL_{BV}$. $EXPL_{BV}[j]$ is set to 1 *iff* `ExplanatoryFeature(f_j)` holds (i.e., the feature f_j explains the set of observed properties represented in $OBSV_{BV}$). Discriminating properties are

Algorithm 1: Explanation

```
input OBSV_BV
define BitVector EXPL_BV
for each j := 0 … |ssn:Feature|-1
   EXPL_BV[j] := 1
for each i := 0 … |ssn:Property|-1
   if OBSV_BV[i] = 1 then
      EXPL_BV := EXPL_BV AND (row i in KB_BM)
output EXPL_BV
```

represented as a bit vector $DISC_{BV}$ where $DISC_{BV}[i]$ is set to 1 *iff* `DiscriminatingProperty(p_i)` (i.e., the property p_i discriminates between the set of explanatory features represented in $EXPL_{BV}$).

Algorithm for Explanation: The strategy employed for efficient implementation of the explanation task relies on the use of the bit vector AND operation to discover and *dismiss* those features that cannot explain the set of observed properties. It begins with all the features as potentially explanatory, and iteratively dismisses those features that cannot explain an observed property. Eventually, for each index position in $EXPL_{BV}$ that is set to 1, the corresponding feature explains all the observed properties.

Algorithm for Discrimination: The strategy employed for efficient implementation of the discrimination task relies on the use of the bit vector AND operation to discover and indirectly *assemble* those properties that discriminate between a set of explanatory features. The discriminating properties are those that are determined to be neither

Algorithm 2: Discrimination

```
input EXPL_BV, OBSV_BV
define BitVector DISC_BV
for each i := 0 … |ssn:Property|-1
   DISC_BV[i] := 0
define BitVector ZERO_BV
for each j := 0 … |ssn:Feature|-1
   ZERO_BV[j] := 0
for each i := 0 … |OBSV_BV|-1
   if OBSV_BV[i] = 0 then
      BitVector PEXPL_BV :=
         EXPL_BV AND (row i in KB_BM)
      if PEXPL_BV != EXPL_BV and
         PEXPL_BV != ZERO_BV then
         DISC_BV[i] := 1
output DISC_BV
```

expected for all feature nor not-applicable for any feature. Note that for a not-yet-observed property at index k_i, and

[11] http://www.loa-cnr.it/ontologies/DUL.owl

the bit vector $PEXPL_{BV}$: (i) $PEXPL_{BV} = EXPL_{BV}$ holds and the k_i^{th} property is expected; (ii) $PEXPL_{BV} = ZERO_{BV}$ holds and the k_i^{th} property is not-applicable; or (iii) the k_i^{th} property discriminates between the explanatory features. Eventually, properties in $DISC_{BV}$ are each capable of partitioning the set of explanatory features in $EXPL_{BV}$.

Illustrative Example: Figure 1 captures the knowledge base (causal relationship) associating observed properties (symptoms) and explanatory features (disorders). E.g., the observation palpitations is explained by both Hypertension and Hyperthyroidism. Similarly, the observations {elevated blood pressure, and palpitations can be explained by the three disorders Hypertension, Hyperthyroidism, and Pulmonary Edema. Viewing it another way, the observed properties elevated blood pressure and palpitations are both expected properties of the features Hypertension and Hyperthyroidism, and hence the former properties cannot be used to discriminate the latter features. The observed property clammy skin is not applicable to the features Hypertension and Hyperthyroidism because the latter does not cause the former. Hence the former property cannot be used to discriminate the latter features. Discriminating properties are those that are neither expected nor not applicable. Thus, the observation clammy skin can be used to discriminate between Hypertension and Hyperthyroidism because clammy skin is caused by Hyperthyroidism but not by Hypertension.

Evaluation: We compared the use of OWL reasoner for running our OWL specifications with the bit vector-based algorithms. (Recall that these algorithms have been shown to be formally correct with respect to the declarative specification in OWL (Henson et al, 2012).) Both implementations are coded in Java, compiled and run on a Dalvik VM for Android phone. The OWL implementation uses Androjena[12], a port of the Jena Semantic Web Framework for Android OS. The Samsung Infuse[13] phone had a 1.2 GHz processor, 16GB storage capacity, and 512MB of internal memory.

To test the efficiency of the two approaches, we timed and averaged 10 executions of each inference task. To test the scalability and evaluate worst-case complexity, the set of relations between properties and features in the KB form a complete bi-partite graph. In addition, for the explanation evaluations, every property is initialized as an observed property; for the discrimination evaluations, every feature is initialized as an explanatory feature. We varied the size of the KB along two dimensions – properties and features. In the OWL approach, as the number of observed properties increase, the `ExplanatoryFeature` class grows more complex (with more conjoined clauses in the complex class definition). As the number of features increase, the `ExpectedProperty` class and `NotApplicableProperty` class grows more complex. In the bit vector approach, as the number of properties increase, the number of rows in KB_{BM} grows. As the number of features increase, the number of columns grows.

Result of OWL evaluations: The results from the OWL implementations of explanation and discrimination are shown in Figures 2 and 3, respectively. With a KB of 14 properties and 5 features, and 14 observed properties to be explained, explanation took 688.58 seconds to complete (11.48 min); discrimination took 2758.07 seconds (45.97 min). With 5 properties and 14 features, and 5 observed properties, explanation took 1036.23 seconds to complete (17.27 min); discrimination took 2643.53 seconds (44.06 min). In each of these experiments, the mobile device runs out of memory if the number of properties or features exceeds 14. The results of varying both properties and features show greater than cubic growth-rate ($O(n^3)$ or worse). For explanation, the effect of features dominates; for discrimination, we are unable to discern any significant difference in computation time between an increase in the number of properties vs. features.

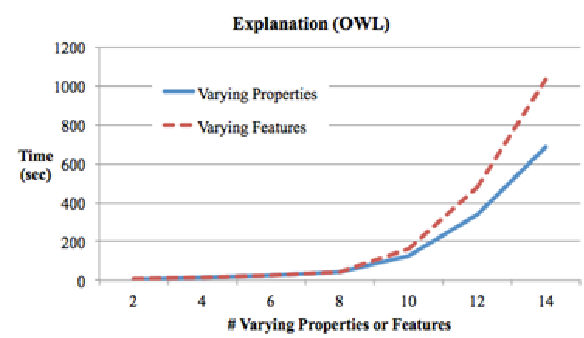

Figure 2. Evaluation: Explanation (OWL) with $O(n^3)$ growth.

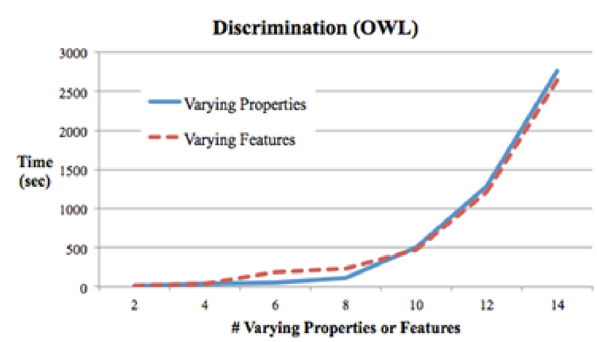

Figure 3. Evaluation: Discrimination (OWL) with $O(n^3)$ growth.

Result of bit vector evaluations: The results from the bit vector implementations of explanation and discrimination

[12] http://code.google.com/p/androjena/

[13] http://www.samsung.com/us/mobile/cell-phones/SGH-I997ZKAATT

are shown in Figures 4 and 5, respectively. With a KB of 10,000 properties and 1,000 features, and 10,000 observed properties to be explained, explanation took 0.0125 seconds to complete; discrimination took 0.1796 seconds. With 1,000 properties and 10,000 features, and 1,000 observed properties, explanation took 0.002 seconds to complete; discrimination took 0.0898 seconds. The results of varying both properties and features show linear growth-rate (O(n)); and the effect of properties dominates.

Discussion of results: The evaluation demonstrates orders of magnitude improvement in both efficiency and scalability. The inference tasks implemented using an OWL reasoner both show greater than cubic growth-rate ($O(n^3)$ or worse), and take many minutes to complete with a small number of observed properties (up to 14) and small KB (up to 19 concepts; #properties + #features). On the other hand, the bit vector implementations show linear growth-rate (O(n)), and take milliseconds to complete with a large number of observed properties (up to 10,000) and large KB (up to 11,000 concepts).

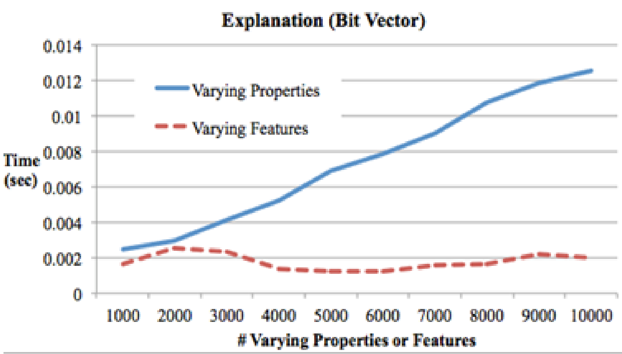

Figure 4. Evaluation: Explanation (bit vector) with O(n) growth.

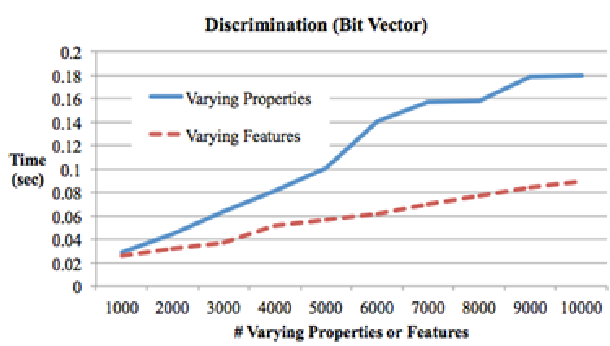

Figure 5. Evaluation: Discrimination (bit vector) with O(n) growth.

Overall Summary:

We first developed a declarative specification of the explanation and discrimination steps in first-order logic (Henson et al, 2011) and in OWL (Henson et al, 2012). We demonstrated that, under single-feature (single-disorder) assumption, the explanation generation (an abductive task) can be carried out by a (deductive) OWL reasoner. We then developed bit-vector encoding as (significantly more) efficient approach to computing the explanation. Specifically, the OWL language and reasoner is more expressive than our limited framework as far as deductive inferences are concerned. However, this reasoner is inadequate for efficiently carrying out the explanation and discrimination steps we need for our use cases on resource-constrained devices as discussed. In fact, the (perception cycle) computation that yields minimum explanation (consisting of single entity/feature) is iterative and requires interleaved use of explanation (abduction) and discrimination (deduction) steps.

For the explanation and discrimination inference tasks executed on a resource-constrained mobile device, the evaluation highlights both the limitations of OWL reasoning and the efficacy of specialized algorithms utilizing bit vector operations. The bit vector encodings and algorithms yield *significant* and *necessary* computational enhancements – including asymptotic order of magnitude improvement, with running times reduced from minutes to milliseconds, and problem size increased from 10's to 1000's. See Figures 2, 3, 4 and 5 for details. The prototyped approach holds promise for applications of contemporary relevance (e.g., healthcare/cardiology).

Addressing Velocity: Continuous Semantics

Velocity can be perceived as either (1) handling large amount of streaming information for real-time analysis (e.g., Superbowl generated 17000 tweets/second) or (2) analyzing and delivering "timely" information (e.g., detect people in trouble and respond via social media to help them out during disasters). In our work, we have focused more on dealing with the latter challenge. For real-time analysis of social-data (Twitter) during events, it is necessary to keep the data filter (crawler) abreast of the happenings of the event. For example, during "Hurricane Sandy", the focus on changing locations (path of the hurricane) and happenings (power cut, flooding, fire) has to be adapted to keep the analysis up to date with the event.

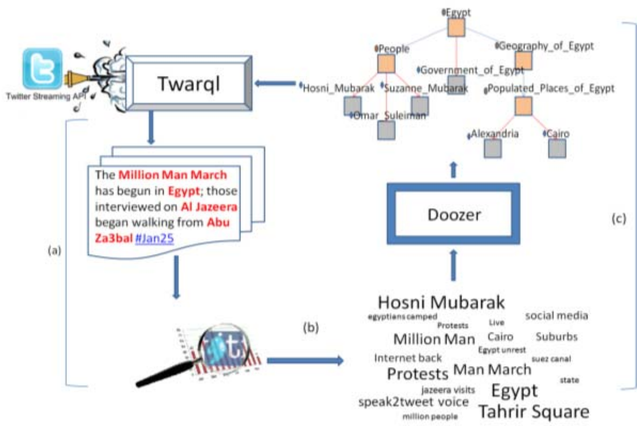

Figure 6: Pipeline for event descriptions using Continuous Semantics.

As part of our Continuous Semantics agenda (Sheth et al, 2010) (Sheth, 2011a), we support dynamic creation and updating of semantic models from social-knowledge sources such as Wikipedia and LOD. These offer exciting new capabilities in making real-time social and sensor data more meaningful and useful for advanced situational-awareness, analysis and decision making. Example applications can be as diverse as following election cycles to forecasting, tracking and monitoring the aftermath of disasters. In Figure 6, Twarql (Mendes et al, 2010) is a social data stream filtering application that utilizes domain models to determine the appropriate key terms to filtering topically relevant tweets. However, given that many events (e.g., disasters, unrests and social movements) change in unanticipated ways, having a static pre-defined model would reduce the recall and consequently miss temporally relevant information (tweets) of the event. In Continuous Semantics, the tweets themselves are used in conjunction with Wikipedia for dynamic model creation by Doozer (Sheth et al, 2010). Such a dynamic domain model is then leveraged for crawling temporally relevant tweets by Twarql. For example, during the Egypt revolt, when the term "million man march" appeared on January 29, 2011, the day before this suddenly planned event, we used the tweets to find frequently occurring terms to generate a temporally relevant domain model. The domain model consisted of "Heliopolis" as a concept relevant to the Egypt revolt. "Heliopolis" is a suburb in Egypt and was the destination of "million man march". This helped to crawl more tweets that mentioned the term relevant to the event. A preliminary study of determining evolving key terms (hashtags) for events was done on US Presidential Elections and Hurricane Sandy. Our approach is able to improve recall and crawl for (on an average) 90% precise tweets using the top-5 relevant hashtags[14].

[14] http://j.mp/C-crawling

Addressing Variety: Hybrid Representation and Reasoning

Use of semantic metadata to describe, integrate, and interoperate between heterogeneous data and services can be very powerful in the big data context, especially if annotations can be generated automatically or with some manual guidance and disambiguation (Sheth and Thirunarayan, 2012). Continuous monitoring of PCSS is producing fine-grained sensor data streams, which is unprecedented. Hence, domain models capturing cause-effect relationships and associations between features and data patterns gleaned from the recently available sensors and sensor modalities have not been uncovered and formalized hitherto. Such properly vetted domain models are, however, critical for prediction, explanation, and ultimately, decision making in real-time from the sensed data. Further, objective physical sensors (e.g., weather sensors, structural integrity sensors) provide quantitative observations. In contrast, subjective citizen sensors (e.g., Tweets) provide qualitative "high-level" interpretation of a situation. For example, a sensed slow moving traffic can result from rush hour, fallen trees, or icy conditions that can be determined from postings on social media. Thus physical and citizen sensors can provide complementary and corroborative information enabling disambiguation. *Specifically, we have sought semantic integration of sensor and social data, using multiple domain ontologies and our IntellegO perceptual reasoning infrastructure, to improve situational awareness.*

Learning domain models from data as well as specifying them declaratively has been widely studied (Domingo and Kersting, 2013). The former approach is "bottom-up", machine driven, correlation-based and statistical in nature, while the latter approach is "top-down", manual, causal and logical in nature. Significant benefit of using domain-specific knowledge in addition to machine learning techniques is now well appreciated (e.g., (Hammond et al, 2002)). The data-driven approach (e.g., exemplified by probabilistic graphical models (Koller and Friedman, 2009)) can be further divided into two levels: (i) *structure learning* that derives qualitative dependencies and (ii) *parameter learning* that quantifies dependencies. We have investigated how to combine these approaches to obtain more complete and reliable situational awareness exploiting mutually corroborative as well as disambiguation information. *Specifically, correlational structure gleaned from data provides the right-level of abstraction for refinement and enhancement using declarative knowledge, prior to parameter estimation in order to learn reliable probabilistic graphical models* (Anantharam et al, 2013).

Statistical and machine learning techniques can be brought to bear to discover correlations among various sensor modalities. Use of data to validate domain models has been the hallmark of modern physics and it is imperative for Data Science as well (Brooks, 2013): "*Data can help compensate for our overconfidence in our own intuitions and can help reduce the extent to which our desires distort our perceptions*." However, big data can be noisy, skewed, inaccurate, and incomplete. Technically speaking, this can confound probability estimates by implicitly conditioning it.

Correlations between two concepts can arise for different reasons. (i) Correlations may be *causal* in nature that is consistent with cause-effect declarative knowledge. For example, "anomalous" motion of Solar system planets w.r.t. earth can be satisfactorily explained by heliocentrism and theory of gravitation, and the "anomalous" precision of Mercury's orbit can be clarified by General Theory of Relativity. C-peptide protein can be used to estimate insulin produced by a patient's pancreas. (ii) Correlations may be *coincidental due to data skew or misrepresentation.* For example, "data-empowered" conflicting claims have been made with improper use of historical precedents (Klass, 2008) (Cayo, 2013) (Stauffer, 2002) (Christensen, 1997). (iii) Correlations may be *coincidental new discoveries.* For example, Wal-Mart executives associated approaching hurricanes with people buying large quantities of Strawberry Pop-Tarts (Brooks, 2013a). (iv) Correlations may be *anomalous and accidental.* For example, since the 1950s, both the atmospheric Carbon Dioxide level and obesity levels have increased sharply. (v) Pavlovian learning induced conditional reflex, and some of the financial market moves, are classic cases of correlation turning into causation!

Even though correlations can provide valuable insights, they can at best serve as valuable hypothesis or deserve explaining from a background semantic theory before we can have full faith in them. For example, consider controversies surrounding assertions such as 'high debt causes low growth', and 'low growth causes high debt'. On the other hand, stress/spicy foods are correlated with peptic ulcers, but the latter are caused by Helicobacter Pyroli[15].

In essence, all these anecdotal examples show possible pitfalls that can also befall big data analytics and predictions, and potential benefits that can accrue.

Combining a statistical approach with declarative logical approach has been a Holy Grail of Knowledge Representation and Reasoning (Domingo and Lowd, 2009). Some specific research goals to be pursued here to improve the quality, generality, and dependability of background knowledge can include: (i) *Gleaning of data driven qualitative dependencies, and integration with qualitative declarative knowledge* that are at the same level of granularity and abstraction. (ii) *Use of these seed models to learn parameters for reliable fit with the data*. For instance, 511.org data (for Bay Area road traffic network) can be analyzed to obtain progressively expressive models starting from gleaning undirected correlations among concepts, to updating (enhancing and correcting) it further using declarative knowledge from ConceptNet[16] to orient the dependencies among concepts, to quantifying dependencies (Anantharam et al, 2013). Specifically, 511.org data can enable us to determine correlation between a number of random variables such as Travel Time, Volume, Speed, Delay, Active Event, Scheduled Event, Day of the Week, and Time of day, associated with every road link. A Bayesian network can be gleaned from 511.org data and enhanced with explicitly provided declarative knowledge by humans or available in ConceptNet (Liu and Singh, 2004). These enhancements can be in the form of correcting edges, orienting undirected edges, and adding new edges. For instance, the enhanced Bayesian network includes edges such as 'baseball-game → traffic jam', 'traffic jam → slow traffic', and 'bad weather → slow traffic' (from ConceptNet), and 'Time of Event → Active Event', 'Volume' → 'Speed', and 'Speed → Travel Time', and 'Scheduled Event → Event' (from 511.org).

We encourage principled ways *to integrate declarative approach with progressively expressive probabilistic models* for analyzing heterogeneous data (Domingo and Lowd, 2009): (1) Naive Bayes that treats all the features as independent; (2) Conditional Linear Gaussian that accommodates boolean random variables; (3) Linear Gaussian that learns both structure and parameters; and (4) Temporal enrichments to these models that can account for the evolution in PCSS. We have applied this approach to fine-grained analysis of Kinect data streams by building models to predict whether a pose belongs to a human or an alien (Koller, 2012). Such techniques can also be applied for activity recognition – ranging from monitoring Parkinson's/Alzheimer's patients to monitoring traffic and system health.

Orthogonal to these efforts are our research initiatives to deal with variety issue cropping up in formalizing materials and process specifications (specs). This can arise in the context of Integrated Computational Materials Engineering (ICME) and Materials Genome Initiative (MGI). We are developing a continuum of light-weight ontologies to annotate documents and embed data semantics to deal with heterogeneity. For example, a spec can be annotated to different levels of detail. The simplest

[15] http://www.nobelprize.org/nobel_prizes/medicine/laureates/2005/press.html

[16] http://csc.media.mit.edu/conceptnet

approach is to make explicit the source and nature of a spec (e.g., AMS 4967 Ti Alloy in the form of bar, wire, etc.). The next refinement can determine the names of the processing steps the spec describes (e.g., composition, heat treatment). A really detailed approach can aggregate all the required parameters for carrying out a process/test (e.g., annealing, tensile test). Our approaches present cost-benefit trade-offs accommodating various application scenarios from indexing and semantic search, to content extraction, to data integration (Thirunarayan, et al, 2005). Further, tabular data are compact and highly irregular (Thirunarayan, 2005a) because they are meant for human consumption. Developing *regular* data structures with well-defined semantics as targets for table translation is an active area of research (Thirunarayan and Sheth, 2013).

Addressing Veracity: Gleaning Trustworthiness

A semantics-empowered integration of physical and citizen sensor data can improve assessing data trustworthiness by correlating data from different modalities. For example, during disaster scenarios, physical sensing may be prone to vagaries of the environment, whereas citizen sensing can be prone to rumors and inaccuracies. So combining their complementary strengths can enable robust situational awareness.

Detection of anomalous (machine/human) sensor data is fundamental to determining the trustworthiness of a sensor. For densely populated sensor networks, one can expect spatio-temporal coherence among sensor data generated by sensors in spatio-temporal proximity. Similarly, domain models can be used to correlate sensor data from heterogeneous sensors. However, anomaly detection in both social and sensor data is complicated as it may also represent an abnormal situation. (As an aside, trending topic abuses are common during disasters and political events/upheavals as illustrated by the infamous Kenneth Cole tweet (Anantharam et al, 2012).) It may not be possible to distinguish an abnormal situation from a sensor fault or plausible rumor purely on the basis of observational data (e.g., freezing temperature in April vs. stuck-at-zero fault). This may require exploring robust domain models for PCSS that can distinguish data reported by compromised sensors (resp. malicious agents) from legitimate data signaling abnormal situation (resp. unlikely event) or erroneous data from faulty sensors (resp. uninformed public).

Reputation-based approaches can be adapted to deal with data from multiple sources (including human-in-the-loop) and over time, to compute the trustworthiness of aggregated data and their sources. Provenance tracking and representation can be the basis for gleaning trustworthiness (Perez, 2010) (Gil, 2012). We have developed upper-level trust ontology and a comparative analysis of several approaches to binary and multi-valued trust, and analyzed their robustness to various attacks (Thirunarayan et al, 2013). Specifically, we have used Bayesian foundation in the form of Beta-distribution to formalize binary trust and Dirichlet-distribution to formalize multi-valued trust. For example, for the binary case, the dynamic trustworthiness of an agent (e.g., sensor, vendor) can be characterized using Beta-PDF Beta(a,b), whose parameters can be gleaned from total number of correct observations $r = (a - 1)$ and total number of erroneous observations $s = (b - 1)$ so far. The overall trustworthiness (reputation) can then be equated to its mean: $a/(a+b)$. We have also analyzed the pros and the cons of several approaches to computing direct trust and (inferred) indirect trust. The indirect trust is computed using trust propagation rules for sequential chaining of edges and parallel aggregation of paths. We have also developed algorithms for computing K-level trust metric based on Dirichlet-distribution incorporating temporal decay, to make it robust with respect to various well-known attacks in trust networks (Thirunarayan et al, 2013). Unfortunately, there is neither a universal notion of trust that is applicable to all domains nor a clear explication of its semantics or computation in many situations (Josang, 2009) (Thirunarayan, 2012).

Trust issues are crucial to big data analytics where we aggregate and integrate data from multiple sources, and in different contexts. The Holy Grail of trust research is to develop *expressive trust frameworks* that have both declarative/axiomatic and computational specification. Furthermore, we need to devise methodologies for instantiating them for practical use by justifying automatic trust inference in terms of application-oriented *semantics of trust* (i.e., vulnerabilities and risk tolerance).

Deriving Value: Evolving Background Knowledge, Actionable Intelligence and Decision Making

The aforementioned research should yield new background knowledge applicable to big data instances and that can benefit end users decision-making (Sheth, 2013). For specificity, here are some concrete examples of applications impacted by our line of research.

Our first example is the Health and wellbeing of patients afflicted with chronic conditions that can be improved by empowering patients to be more proactive and participatory in their own health-care. Development of such mobile applications requires:

(i) Building background knowledge/ontology involving disorders, causative triggers, symptoms and medications.

(ii) Using environmental and on-body sensors, background knowledge, and patient health history to prescribe a regimen to avoid triggers, improve resistance, and treat symptoms.

As a second example, consider the acquisition of new background knowledge to improve coverage by exploiting EMR data (e.g., in the cardiology context). Specifically, our research elicits missing knowledge by leveraging EMR data to hypothesize plausible relationships, gleaned through statistical correlations. These can be validated by domain experts with reduced manual effort (Perera et al, 2012).

As a third example, our research leveraged massive amounts of user generated content to build high-quality prediction models. For example, Twitter and author-provided emotion hashtags can be harnessed for sentiment/emotion identification in tweets (Wang et al, 2012).

The observations and interactions in PCSS are characterized by three attributes. They are *incomplete* due to partial observation from the real world. There is *uncertainty* due to inherent randomness involved in the sensing process (noise in machine sensors and bias in citizen sensors). It is *dynamic* because of the ever changing and non-deterministic conditions of the physical world. Graphical models can be used to deal with incompleteness, uncertainty, and dynamism in many diverse domains. Unfortunately, extracting structure is very challenging due to data sparseness and difficulty in detecting causal links (Anantharam et al, 2013). Declarative domain knowledge can obviate the need to learn everything from data. In addition, correlations derivable from data can be further consolidated if the declarative knowledge base provides evidence for it. Similarly to the traffic use case discussed before, we believe that leveraging domain ontologies and data sets published on the LOD cloud and integrating it with data-driven correlations will increase the fidelity of graphical models, improving their predictive and analytical power.

Conclusions

We have outlined how semantic models and technologies can be, and in many cases are being, used to address various problems associated with big data. We overcome *volume* by enabling abstraction to achieve semantic scalability for decision making. We defined two operations – explanation and discrimination – that underlie the semantics of machine perception, and showed how they can be implemented efficiently on resourced-constrained devices. *Variety* challenges can be overcome using a continuum of light-weight semantics to achieve semantic integration and interoperability. We benefitted from combining statistical as well as declarative knowledge, to improve coverage, reliability, and semantic scalability. We employed dynamically constructed domain models for semantic filtering to deal with *velocity*. To improve *veracity*, we have used Bayesian foundation to deal with homogeneous sensor networks, and semantics for cross checking multimodal data against constraints. We achieved *value* by enriching background knowledge to make them comprehensive for better decision making. Given Kno.e.sis' empirically driven multidisciplinary research, we seek to harness semantics for big data that can impact a wide variety of application areas including medicine, health and wellbeing, disaster and crisis management, environment and weather, Internet of Things, sustainability and smart city infrastructure.

Acknowledgements

We acknowledge Cory Henson for significant contributions on Semantic Perception, Pramod Anantharam on hybridization of statistical and logic-based techniques, and in dealing with real-world sensor and social data, and Pavan Kapanipathi and Sanjaya Wijeratne on Continuous Semantics. We acknowledge partial support from the National Science Foundation (NSF) awards IIS-1111182: SoCS: Social Media Enhanced Organizational Sensemaking in Emergency Response and IIS-1143717: EAGER - Expressive Scalable Querying over Integrated Linked Open Data. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Anantharam, P., Thirunarayan, K., and Sheth, A. 2012. Topical Anomaly Detection for Twitter Stream, *Proceedings of ACM Web Science 2012*, 11-14.

Anantharam, P., Thirunarayan, K., and Sheth, A. 2013. Traffic Analytics using Probabilistic Graphical Models Enhanced with Knowledge Bases, *Proceedings of the 2nd International Workshop on Analytics for Cyber-Physical Systems* (ACS-2013) at SIAM International Conference on Data Mining, 13-20.

Brooks, D. 2013. What Data Can't Do? http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html?_r=0

Brooks, D. 2013a. What You'll Do Next http://www.nytimes.com/2013/04/16/opinion/brooks-what-youll-do-next.html

Cayo, D. 2013. Bad Data Leads to Bad Decisions on Foreign Aid, SFU Economist. http://www.vancouversun.com/business/story.html?id=8271632#ixzz2SRB49mOM.

Christensen, C. M. 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Harvard Business School Press. [Domingo and Lowd, 2009]

Domingo, P. and Lowd, D. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. San Rafael, CA: Morgan & Claypool.

Domingo, P. and Kersting, K. 2013. *Combining Logic and Probability: Languages, Algorithms and Applications*. AAAI-2013 Tutorial. http://www-kd.iai.uni-bonn.de/index.php?page=news_details&id=30

Gil, Y. 2012. http://www.isi.edu/~gil/research/provenance.html.

Hammond, B., Sheth, A., and Kochut, K. 2002. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content, In: *Real World Semantic Web Applications*, V. Kashyap and L. Shklar (Eds.), Frontiers in Artificial Intelligence and Applications, vol. 92, Amsterdam: IOS Press, 29-49.

Henson, C., Pschorr, J., Sheth, A., and Thirunarayan, K. 2009. SemSOS: Semantic Sensor Observation Service, In *Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems* (CTS 2009), Baltimore, MD.

Henson, C., Thirunarayan, K., and Sheth, A. 2011. An Ontological Approach to Focusing Attention and Enhancing Machine Perception on the Web. *Applied Ontology*, 6(4):345-376.

Henson, C. Sheth, A., and Thirunarayan, K. 2012. Semantic Perception: Converting Sensory Observations to Abstractions, *IEEE Internet Computing*, 16(2):26-34.

Henson, C., Thirunarayan, K., and Sheth, A. 2012a. An Efficient Bit Vector Approach to Semantics-Based Machine Perception in Resource-Constrained Devices. *International Semantic Web Conference* (1) 2012: 149-164.

Henson, C. 2013. A Semantics-based Approach to Machine Perception. Doctoral Dissertation. Department of Computer Science and Engineering, Wright State University, Dayton, OH.

Josang, A. 2009. Trust and Reputation Systems, http://folk.uio.no/josang/tr/IFIPTM2009-TrustRepSys.pdf, Invited Tutorial at IFIPTM-2009.

Kapanipathi, P., Thirunarayan, K., Sheth, A., and Hitzler, P. Leveraging Semantics for Detection of Event-Descriptors on Twitter. Kno.e.sis Center, Wright State University, Technical Report 2013. (See also: http://wiki.knoesis.org/index.php/Continuous_Semantic_Crawling_Events)

Klass, G. 2008. Just Plain Data Analysis: Common Statistical Fallacies in Analyses of Social Indicator Data. http://polmeth.wustl.edu/media/Paper/2008KlassASA2.pdf

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.

Koller, D. 2012 *Programming Graphical Models course*, http://online.stanford.edu/pgm-fa12; https://www.coursera.org/course/pgm

Liu, H., and Singh, P. ConceptNet—A Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 22.4 (2004): 211-226.

Mendes, Pablo N., et al. "Linked open social signals." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010. APA

Perera, S., Henson, C., Thirunarayan, K., Sheth, A, and Nair, S. 2012. Data Driven Knowledge Acquisition Method for Domain Knowledge Enrichment in the Healthcare, *6th International Conference on Bioinformatics and Biomedicine* BIBM12. 197-205. (Extended version to appear as: Semantics Driven Approach for Knowledge Acquisition from EMRs, in *IEEE Journal of Biomedical and Health Informatics*.)

Perez, J. M. G. 2010. Provenance and Trust. http://www.slideshare.net/jmgomez23/provenance-and-trust.

Sheth, A., Thomas, C., and Mehra, P. 2010. Continuous Semantics to Analyze Real-Time Data, *IEEE Internet Computing*, 14 (6), 84-89. http://wiki.knoesis.org/index.php/Continuous_Semantics_to_Analyze_Real_Time_Data

Sheth, A. 2011. Semantics Scales Up: Beyond Search in Web 3.0. http://www.computer.org/csdl/mags/ic/2011/06/mic2011060003-abs.html

Sheth, A. 2011a. Citizen Sensing-Opportunities and Challenges in Mining Social Signals and Perceptions, Invited Talk at Microsoft Research Faculty Summit 2011, Redmond, WA.

Sheth, A. and Thirunarayan, K. 2012. *Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications*. Synthesis Lectures on Data Management, Morgan & Claypool Publishers.

Sheth, A., Anantharam, P., and Henson, C. 2013. Physical-Cyber-Social Computing: An Early 21st Century Approach, *IEEE Intelligent Systems*, 79-82, with extended version at: http://wiki.knoesis.org/index.php/PCS

Sheth, A. 2013. Transforming Big Data into Smart Data: Deriving Value via harnessing Volume, Variety and Velocity using semantics and Semantic Web, Keynote at the 21st Italian Symposium on Advanced Database Systems, Roccella Jonica, Italy. http://j.mp/SmatData

Stauffer, D. 2002. How Good Data Leads to Bad Decisions, *Harvard Business Publishing Newsletters*, 3 pages.

Thirunarayan, K., Berkovich, A., and Sokol, D. Z. 2005. An Information Extraction Approach to Reorganizing and Summarizing Specifications. Information & Software Technology 47(4), pp. 215-232.

Thirunarayan, K. 2005a. On Embedding Machine-Processable Semantics into Documents, *IEEE Transactions on Knowledge and Data Engineering*, 17(7), pp. 1014-1018.

Thirunarayan, K. 2012. Trust Networks Tutorial, http://www.slideshare.net/knoesis/trust-networks, Invited Tutorial at CTS-2012.

Thirunarayan, K., and Sheth, A. 2013. Semantics-empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications, In: *Proceedings of AAAI 2013 Fall Symposium on Semantics for Big Data*, Arlington, Virginia, November 15-17, 2013, 8 pages.

Thirunarayan, K., Anantharam, P., Henson, C., and Sheth, A. 2013. Comparative Trust Management with Applications: Bayesian Approaches Emphasis, *Future Generation Computer Systems*.18 pages. http://dx.doi.org/10.1016/j.future.2013.05.006

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. 2012. Harnessing Twitter 'Big Data' for Automatic Emotion Identification. *International Conference on Social Computing* (SocialCom).

Brief Biographies:

Krishnaprasad Thirunarayan is a Professor of Computer Science and Engineering at Kno.e.sis - the Ohio Center of Excellence in Knowledge-enabled Computing. His research interests are in Big Data (including social and sensor data analytics), Hybrid Knowledge Representation and Efficient Reasoning, IR, Trust Networks, Internet of Things, and Web 3.0. (http://knoesis.org/tkprasad)

Amit Sheth is the LexisNexis Ohio Eminent Scholar, and directs Kno.e.sis Center. His research interests are in Computing for Human Experience, Big and Smart Data Sciences, Physical Cyber Social Computing, Web 3.0, Internet of Things, and Semantic Web. He is one of the highly cited authors in Computer Science. (http://knoesis.org/amit)