Semantics-Empowered Big Data Processing with Applications

Krishnaprasad Thirunarayan, Amit Sheth

■ *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. To handle volume, we advocate semantic perception that can convert low-level observational data to higher-level abstractions more suitable for decision making. To handle variety, we resort to semantic models and annotations of data so that intelligent processing can be done independent of heterogeneity of data formats and media. To handle 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 trust models and approaches to glean trustworthiness. These four of the five v's of big data are harnessed by the semantics-empowered analytics to derive value to support applications transcending the physical-cyber-social continuum.*

Physical-cyber-social systems (PCSS) (Sheth, Anantharam, and Henson 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 PCSS can be organized around the five *v*'s 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.

 ISSN 0738-4602

Characteristics of the Big Data Problem

We discuss the primary characteristics of the big data problem as it pertains to the five *v*'s. (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, more than 2 billion sensors have been deployed and about 250 terabytes of sensor data are generated for a New York to Los Angeles flight on a Boeing 737.[1] The Parkinson's disease data set[2] that tracked 16 people (9 patients + 7 controls) with mobile phones containing 7 sensors over 8 weeks is 12 gigabytes 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 whether there is cause for concern. Similarly, we should be able to measure the symptoms of Parkinson's disease using sensors on a smartphone, monitor the disease's 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 resource constraints negate the use of conventional data formats and algorithms, and instead necessitate the development of novel encoding, indexing, and reasoning techniques (Henson, Thirunarayan, and Sheth 2012).

The volume of data challenges our ability to process them. First, it is difficult to abstract fine-grained machine-accessible data into a 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 (for example, speed, volume, and travel times), as well as textual information about active events (for example, accidents, vehicle breakdowns) and scheduled events (for example, sporting events, music events) (Anantharam, Thirunarayan, and Sheth 2013). Weather data sets (such as from Mesowest) provide numeric information about primitive phenomena (for example, temperature, precipitation, wind speed) that are required to be combined and abstracted into human-comprehensible weather features in textual form. In medical domains (for example, 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-appraisals (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 data sets, and materials and process specifications used for realizing integrated computational materials engineering[3] (ICME) and materials genome initiative[4] (MGI), exhibit a lot of syntactic and semantic variety[5] (Thirunarayan, Berkovich, and Sokol 2005).

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

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 midair 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 traffic collision avoidance system (TCAS) (which was providing a 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 the Boston Marathon bomber.[8] These examples illustrate the difficulties we face while making decisions based on conflicting data from different sources.

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 health care, 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 five *v*'s from Kno.e.sis's 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 2013, Sheth 2011a), 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 the semantic sensor web (SSW) and IntellegO (Henson, Sheth, and Thirunarayan 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 through 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 or visit to a nurse or doctor, or hospitalization. The first example, which follows, can be formalized using our approach with demonstrable benefits, while the subsequent examples require research into high-fidelity models and human mediation for fruition.

Example One: Weather Use Case

This application involves determining and tracking weather features from weather phenomena, with the potential for tasking sensors if additional information is necessary. We have developed the semantically enabled sensor observation service (SemSOS), which 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 the weather use case, we have used rules provided by NOAA to map primitive machine-sensed weather data (for example, wind speed, temperature, precipitation) to human-comprehensible weather features (for example, 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.

Example Two: 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 through sensors and mobile devices. They can also be augmented with patient-reported observations (for example, 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 (for example, using average, peak, rate of change, duration), and extracting human-comprehensible features by integrating them. This approach requires construction of suitable domain models and a 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 eventually generate the minimum explanation that can be used to determine action. For example, abduction can be applied to weather phenomena data (for example, precipitation and temperature) to determine weather features (for example, flurry and blizzard) that can be further disambiguated by making additional observations (for example, 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, microphone, and others are converted into human-perceived features such as tremors, walking style, balance, and slurred speech to diagnose and monitor disease progression and to recommend control options. For heart failure patients, weight change, heart rate, blood pressure, oxygen level, and others 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, sleeplessness, and others can be used to recommend prevention strategies, treatment levels, and control options. The continuous monitoring of a patient, his or her 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 comorbidity, 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 spatiotemporal-thematic contextual knowledge. In medical scenarios, effective treatment also requires personalization on the patient's historical data and the clinician-prescribed current protocol (for example, 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 the data using background knowledge, and return the essence. For example, this is required to address privacy concerns and the 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, Thirunarayan, and Sheth 2012).

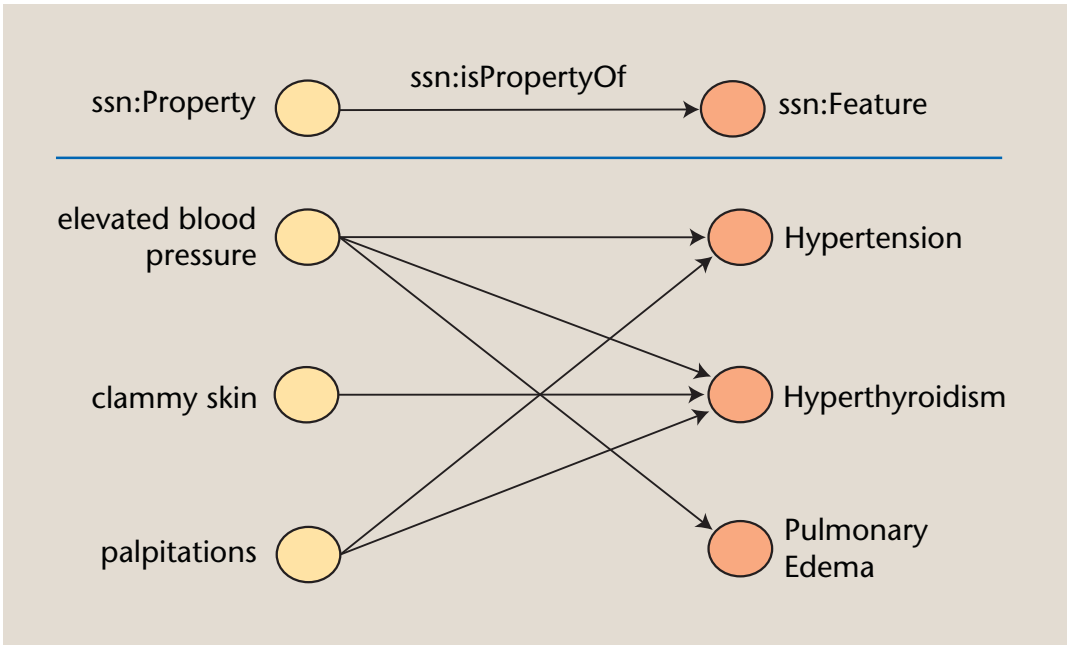

Figure 1. Bipartite Graph Representation of a Simple Cardiology Knowledge Base.

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, Thirunarayan, and Sheth 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 later (Henson, Thirunarayan, and Sheth 2012).

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: The 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 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.

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, Thirunarayan, and Sheth (2012).

Algorithm 1: Explanation

```
[1] input OBSV_BV
[2] define BitVector EXPL_BV
[3] for each j = 0 … |sso:Feature|-1
[4]     EXPL_BV[j] = 1
[5] for each i = 0 … |sso:Property|-1
[6]     if OBSV_BV[i] == 1 then
[7]         EXPL_BV = EXPL_BV AND (row i in KB_BM)
[8] output EXPL_BV
```

Algorithm 1. Explanation.

Algorithm 2: Discrimination

```
[1]  input EXPL_BV, OBSV_BV
[2]  define BitVector DISC_BV
[3]  for each i = 0 … |sso:Property|-1
[4]     DISC_BV[i] = 0
[5]  define BitVector ZERO_BV
[6]  for each j = 0 … |sso:Feature|-1
[7]     ZERO_BV[j] = 0
[8]  for each i = 0 … |OBSV_BV|-1
[9]     if OBSV_BV[i] == 0 then
[10]        BitVector PEXPL_BV =
[11]           EXPL_BV AND (row i in KB_BM)
[12]        if PEXPL_BV != ZERO_BV and
[13]           PEXPL_BV != EXPL_BV then
[14]           DISC_BV[i] = 1
[15] output DISC_BV
```

Algorithm 2. Discrimination.

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 (that is, 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 (that is, there exists an ssn:isPropertyOf(pi,fj) relation). Observed properties are represented as a bit vector $OBSV_{BV}$, where $OBSV_{BV}[i]$ is set to 1 *iff* ObservedProperty(p_i) holds (that is, 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 (that is, the feature f_j explains the set of observed properties represented in $OBSV_{BV}$). Discriminating properties are represented as a bit vector $DISC_{BV}$ where $DISC_{BV}[i]$ is set to 1 *iff* DiscriminatingProperty(p_i) (that is, 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. See algorithm 1.

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 expected for all features nor not applicable for any feature. Note that for a not-yet-observed property at index *i*, and the bit vector $PEXPL_{BV}$: (1) $PEXPL_{BV} = EXPL_{BV}$ holds and the *i*th property is expected; (2) $PEXPL_{BV} = ZERO_{BV}$ holds and the *i*th property is not applicable; or (3) the *i*th property discriminates between the explanatory features. Eventually, properties in $DISC_{BV}$ are each capable of partitioning the set of explanatory features in $EXPL_{BV}$. See algorithm 2.

Illustrative Example

Figure 1 captures the knowledge base (causal relationship) associating observed properties (symptoms) and explanatory features (disorders). For example, 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

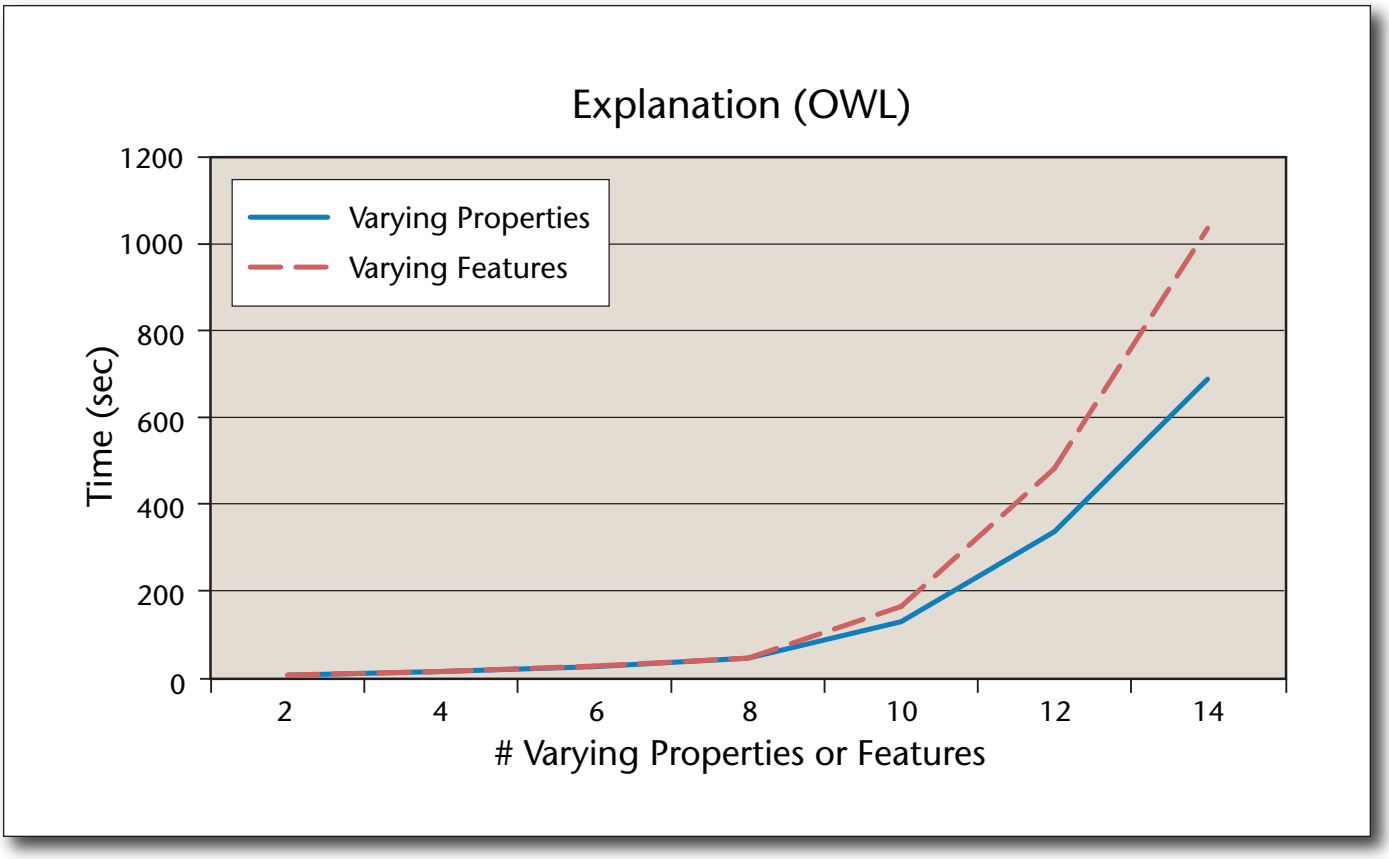

Figure 2. Evaluation.

Explanation (OWL) with O(n^3) growth.

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, Sheth, and Thirunarayan 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, 16 gigabyte storage capacity, and 512 megabytes 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 knowledge base form a complete bipartite 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 knowledge base 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 grow more complex. In the bit-vector approach, as the number of properties increase, the number of rows in KBBM 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 knowledge base of 14 properties and 5 features, and 14 observed properties to be explained, explanation took 688.58 seconds to complete (11.48 minutes); discrimination took 2758.07 seconds (45.97 minutes). With 5 properties and 14 features, and 5 observed properties, explanation took 1036.23 seconds to complete (17.27 minutes); discrimination took 2643.53 seconds (44.06 minutes).

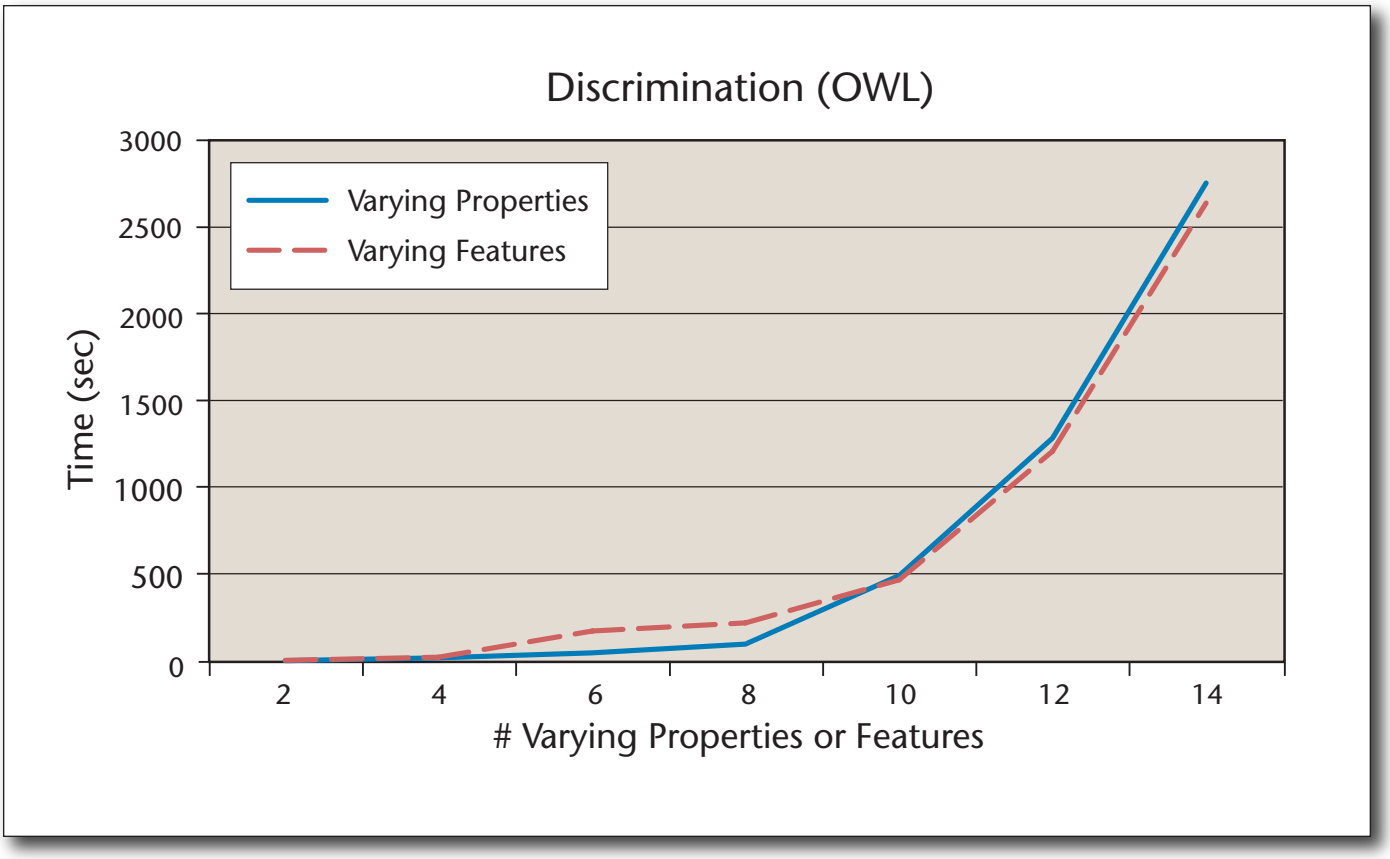

Figure 3. Evaluation.

Discrimination (OWL) with O(n^3) growth.

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 versus features.

Result of Bit-Vector Evaluations

The results from the bit-vector implementations of explanation and discrimination are shown in figures 4 and 5, respectively. With a knowledge base 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 knowledge base (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 knowledge base (up to 11,000 concepts).

Overall Summary

We first developed a declarative specification of the explanation and discrimination steps in first-order logic (Henson, Thirunarayan, and Sheth 2011) and in OWL (Henson, Sheth, and Thirunarayan 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 discrim-

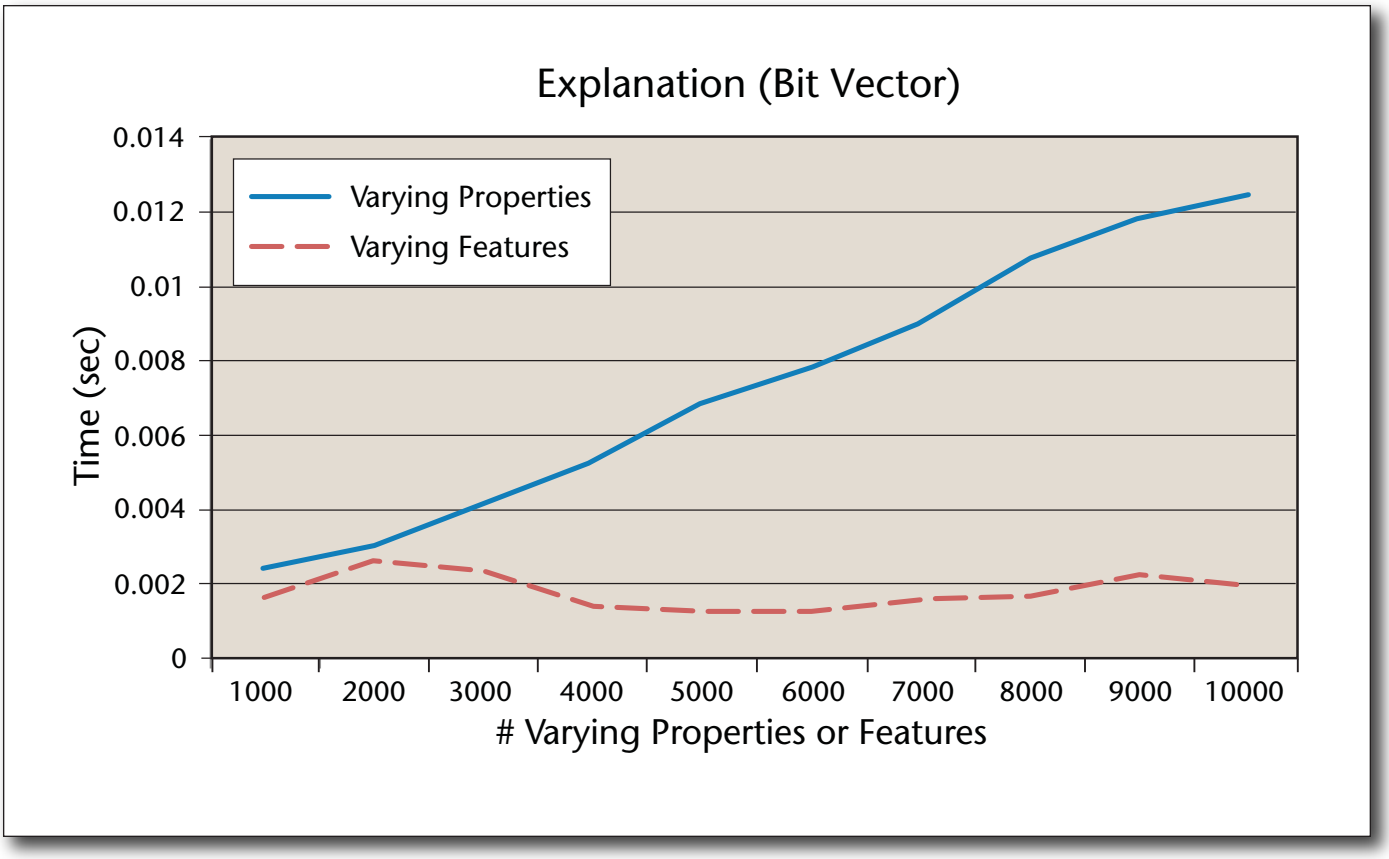

Figure 4. Evaluation.

Explanation (bit vector) with O(*n*) growth.

ination 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 10s to 1000s. See figures 2, 3, 4, and 5 for details. The prototyped approach holds promise for applications of contemporary relevance (for example, health care/cardiology).

Addressing Velocity: Continuous Semantics

Velocity can be perceived as either (1) handling large amount of streaming information for real-time analysis (for example, Super Bowl generated 17,000 tweets/second) or (2) analyzing and delivering timely information (for example, detect people in trouble and respond through 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.

As part of our continuous semantics agenda (Sheth, Thomas, and Mehra 2010; Sheth 2011b), 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 deter-

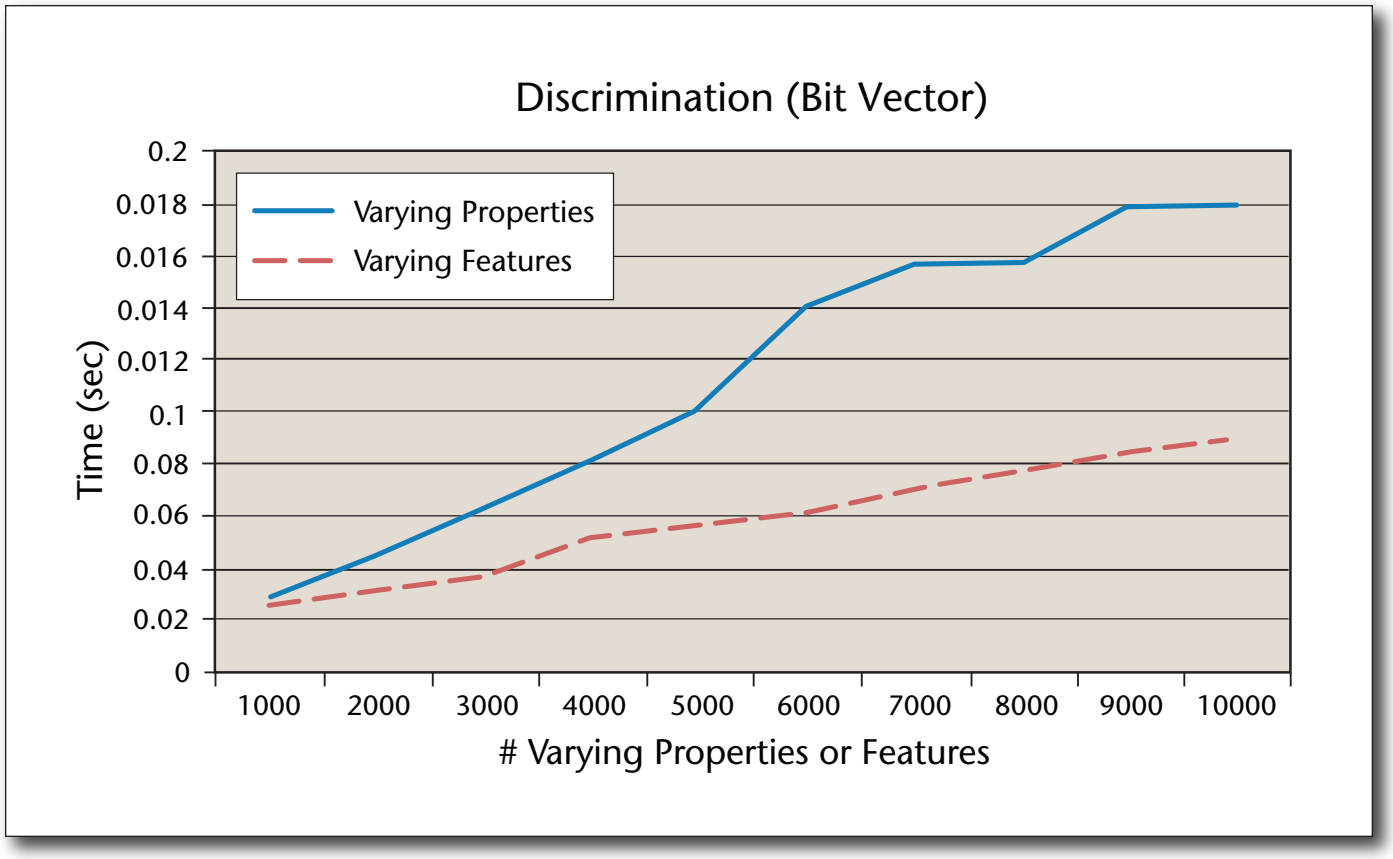

Figure 5. Evaluation.

Discrimination (bit vector) with O(*n*) growth.

mine the appropriate key terms to filtering topically relevant tweets. However, given that many events (for example, disasters, unrests, and social movements) change in unanticipated ways, having a static predefined 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, Thomas, and Mehra 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 U.S. presidential elections and Hurricane Sandy. Our approach is able to improve recall and crawl for (on an average) 90 percent precise tweets using the top five relevant hashtags.[14]

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 (for example, weather sensors, structural integrity sensors) provide quantitative observations. In contrast, subjective citizen sensors (for example, tweets) pro-

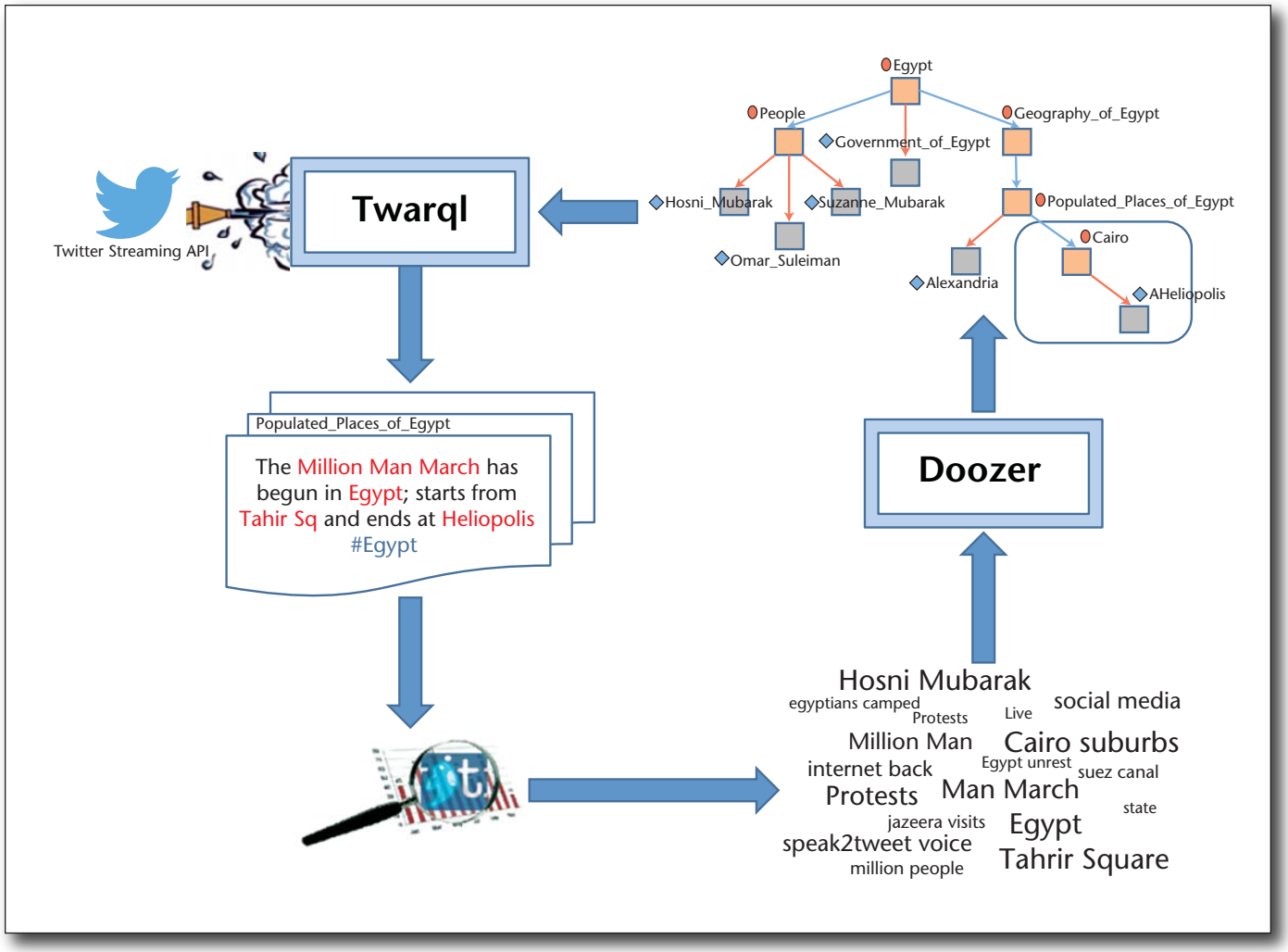

Figure 6. Pipeline for Event Descriptions Using Continuous Semantics.

vide 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 (for example, Hammond, Sheth, and Kochut [2002]). The data-driven approach (for example, exemplified by probabilistic graphical models [Koller and Friedman 2009]) can be further divided into two levels: (1) structure learning that derives qualitative dependencies and (2) 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, Thirunarayan, and Sheth 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 2013a): 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. First, correlations may be causal in nature that is consistent with cause-effect declarative knowledge. For example, anomalous motion of solar system planets with regard to Earth can be satisfactorily explained by heliocentrism and theory of gravitation, and the anomalous precision of Mercury's orbit can be clarified by the general theory of relativity. C-peptide protein can be used to estimate insulin produced by a patient's pancreas. Second, 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 (Cayo 2013, Stauffer 2002, Christensen 1997).[15] Third, correlations may be coincidental new discoveries. For example, Wal-Mart executives associated approaching hurricanes with people buying large quantities of strawberry Pop-Tarts (Brooks 2013b). Fourth, correlations may be anomalous and accidental. For example, since the 1950s, both the atmospheric carbon dioxide level and obesity levels have increased sharply. Finally, 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 a 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*.[16]

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: (1) Gleaning of data-driven qualitative dependencies, and integration with qualitative declarative knowledge, that are at the same level of granularity and abstraction. (2) 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[17] to orient the dependencies among concepts, to quantifying dependencies (Anantharam, Thirunarayan, and Sheth 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 four principled ways to integrate the 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.[18] Such techniques can also be applied for activity recognition — ranging from monitoring Parkinson's or Alzheimer's patients to monitoring traffic and system health.

Orthogonal to these efforts are our research initiatives to deal with the variety of issues 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 lightweight 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 approach is to make explicit the source and nature of a spec (for example, AMS 4967 Ti Alloy in the form of bar, wire, and others). The next refinement can determine the names of the processing steps the spec describes (for example, composition, heat treatment). A really detailed approach can aggregate all the required parameters for carrying out a process/test (for example, 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, Berkovich, and Sokol 2005). Further, tabular data are compact and highly irregular (Thirunarayan 2005) 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 or human) sensor data is fundamental to determining the trustworthiness of a sensor. For densely populated sensor networks, one can expect spatiotemporal coherence among sensor data generated by sensors in spatiotemporal 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, Thirunarayan, and Sheth 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 (for example, freezing temperature in April versus stuck-at-zero fault). This may require exploring robust domain models for PCSS that can distinguish data reported by compromised sensors (respectively, malicious agents) from legitimate data signaling abnormal situation (respectively, unlikely event) or erroneous data from faulty sensors (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.[19, 20] We have developed an upper-level trust ontology and a comparative analysis of several approaches to binary and multivalued trust and analyzed their robustness to various attacks (Thirunarayan et al. 2013). Specifically, we have used a Bayesian foundation in the form of beta distribution to formalize binary trust and Dirichlet distribution to formalize multivalued trust. For example, for the binary case, the dynamic trustworthiness of an agent (for example, sensor, vendor) can be characterized using Beta-PDF Beta(a, b), whose parameters can be gleaned from the total number of correct observations $r = (a - 1)$ and the 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 the 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 specifications. 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 (that is, 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 affected by our line of research.

Our first example is the health and well-being 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 (1) building background knowledge/ontology involving disorders, causative triggers, symptoms, and medications; and (2) 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 (for example, 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. 2014).

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 nondeterministic 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, Thirunarayan, and Sheth 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 lightweight 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 a 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 the knowledge comprehensive for better decision making. Given Kno.e.sis's empirically driven multidisciplinary research, we seek to harness semantics for big data that can affect a wide variety of application areas including medicine, health and well-being, 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 authors and do not necessarily reflect the views of the NSF.

Notes

1. gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2.

2. www.michaeljfox.org/page.html?parkinsons-data-challenge.

3. www.nap.edu/catalog.php?record_id=12199.

4. www.whitehouse.gov/mgi.

5. earthcube.ning.com.

6. nycopendata.socrata.com.

7. en.wikipedia.org/wiki/uberlingen_mid-air_collision.

8. bit.ly/1dFi5b3.

9. knoesis.org/projects/multidisciplinary.

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

11. www.loa-cnr.it/ontologies/DUL.owl.

12. code.google.com/p/androjena.

13. www.samsung.com/us/mobile/cell-phones/SGH-I997ZKAATT.

14. j.mp/C-crawling.

15. See Just Plain Data Analysis: Common Statistical Fallacies in Analyses of Social Indicator Data. Unpublished mss. (polmeth.wustl.edu/media/Paper/2008KlassASA2.pdf).

16. www.nobelprize.org/nobel_prizes/medicine/laureates/2005/press.html.

17. csc.media.mit.edu/conceptnet.

18. See D. Koller's Programming Graphical Models course, (online.stanford.edu/pgm-fa12); (www.coursera.org/course/pgm).

19. See www.isi.edu/~gil/research/provenance.html.

20. See J. M. G. Perez, Provenance and Trust, www.slideshare.net/jmgomez23/provenance-and-trust.

References

Anantharam, P.; Thirunarayan, K.; and Sheth, A. 2012. Topical Anomaly Detection for Twitter Stream. In *Proceedings of ACM Web Science* 2012, 11–14. New York: Association for Computer Machinery.

Anantharam, P.; Thirunarayan, K.; and Sheth, A. 2013. Traffic Analytics Using Probabilistic Graphical Models Enhanced with Knowledge Bases. Paper presented at the 2nd International workshop on Analytics for Cyber-Physical

Systems, 2013 SIAM International Conference on Data Mining, Austin Texas, May 2–4.

Brooks, D. 2013a. What Data Can't Do? *New York Times* (February 18). www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html?_r=0

Brooks, D. 2013b. What You'll Do Next. *New York Times* April 13 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. *Vancouver Sun* (April 19), www.vancouversun.com/business/Cayo+data+leads+decisions+foreign/8268906/story.html?__federated=1

Christensen, C. M. 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Cambridge, MA: Harvard Business School Press.

Domingo, P., and Kersting, K. 2013. Combining Logic and Probability: Languages, Algorithms, and Applications. Tutorial Presentation at the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WA, July 14. (Slides: www-kd.iai.uni-bonn.de/index.php?page=news_details&id=30)

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

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, Frontiers in Artificial Intelligence and Applications,* vol. 92, ed. V. Kashyap and L. Shklar, 29–49. Amsterdam: IOS Press.

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

Henson, C.; Pschorr, J.; Sheth, A.; and Thirunarayan, K. 2009. SemSOS: Semantic Sensor Observation Service. In *Proceedings of the 2009 IEEE International Symposium on Collaborative Technologies and Systems* (CTS 2009). Piscataway, NJ: Institute for Electrical and Electronics Engineers.

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. 2011. An Ontological Approach to Focusing Attention and Enhancing Machine Perception on the Web. *Applied Ontology* 6(4): 345–376.

Henson, C.; Thirunarayan, K.; and Sheth, A. 2012. An Efficient Bit Vector Approach to Semantics-Based Machine Perception in Resource-Constrained Devices. In *The Semantic Web,* Lecture Notes in Computer Science Volume 7649. Berlin: Springer.

Josang, A. 2009. Trust and Reputation Systems. Invited Tutorial presented at the Third International Federation for Information Processing Conference on Trust Management. West Lafayette, IN, June 16–19.

Kapanipathi, P.; Thirunarayan, K.; Sheth, A.; and Hitzler, P. 2013. Leveraging Semantics for Detection of Event-Descriptors on Twitter. Kno.e.sis Center, Wright State University, Technical Report.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models — Principles and Techniques*. Cambridge, MA: The MIT Press.

Liu, H., and Singh, P. 2004. ConceptNet—A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal* 22(4): 211–226.

Mendes, P. N.; Passant, A.; Kapanipathi, P.; and Sheth, A. P. 2010. Linked Open Social Signals. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* vol. 1. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Perera, S.; Henson, C.; Thirunarayan, K.; Sheth, A.; and Nair, S. 2014. Semantics Driven Approach for Knowledge Acquisition from EMRs *IEEE Journal of Biomedical and Health Informatics* 18(2): 515–24. dx.doi.org/10.1109/JBHI.2013.2282125.

Sheth, A. 2013. Transforming Big Data into Smart Data: Deriving Value via Harnessing Volume, Variety, and Velocity Using Semantics and Semantic Web. Keynote talk presented at the 21st Italian Symposium on Advanced Database Systems, Roccella Jonica, Italy, 30 June.

Sheth, A. 2011a. Semantics Scales Up: Beyond Search in Web 3.0. *IEEE Internet Computer* 15(6): 3–6. dx.doi.org/10.1109/MIC.2011.157

Sheth, A. 2011b. Citizen Sensing-Opportunities and Challenges in Mining Social Signals and Perceptions. Invited Talk presented at the Microsoft Research Faculty Summit 2011, 18–20 July, 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 and Claypool Publishers.

Sheth, A.; Anantharam, P.; and Henson, C. 2013. Physical-Cyber-Social Computing: An Early 21st Century Approach, *IEEE Intelligent Systems* 28(1): 79–82.

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

Stauffer, D. 2002. How Good Data Leads to Bad Decisions. *Harvard Mangement Update Newsletter* December: 25–27.

Thirunarayan, K. 2012. Trust Networks Tutorial presented at the International Conference on Collaboration Technologies and Systems, Denver, CO, 21–25 May. (www.slideshare.net/knoesis/trust-networks)

Thirunarayan, K. 2005. On Embedding Machine-Processable Semantics into Documents. *IEEE Transactions on Knowledge and Data Engineering* 17(7): 1014–1018. dx.doi.org/10.1109/TKDE.2005.113

Thirunarayan, K., and Sheth, A. 2013. Semantics-Empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications, In Semantics for Big Data: Papers from the AAAI Symposium. AAAI Technical Report FS-13-04, 68–75.

Thirunarayan, K.; Anantharam, P.; Henson, C.; and Sheth, A. 2013. Comparative Trust Management with Applications: Bayesian Approaches Emphasis. *Future Generation Computer Systems* 31(February): 182–199. dx.doi.org/10.1016/j.future.2013.05.006

Thirunarayan, K.; Berkovich, A.; and Sokol, D. Z. 2005. An Information Extraction Approach to Reorganizing and Summarizing Specifications. *Information & Software Technology* 47(4): 215–232. dx.doi.org/10.1016/j.infsof.2004.08.003

Please Join Us for the Ninth International AAAI Conference on Weblogs and Social Media

26–29 May 2015
Oxford, United Kingdom

The Ninth International AAAI Conference on Weblogs and Social Media (ICWSM) will be held at Oxford University in Oxford, United Kingdom from May 26–29. This interdisciplinary conference is a forum for researchers in computer science and social science to come together to share knowledge, discuss ideas, exchange information, and learn about cutting-edge research in diverse fields with the common theme of online social media. This overall theme includes research in new perspectives in social theories, as well as computational algorithms for analyzing social media. ICWSM is a singularly fitting venue for research that blends social science and computational approaches to answer important and challenging questions about human social behavior through social media while advancing computational tools for vast and unstructured data.

ICWSM-15 will include a lively program of technical talks and posters, invited presentations, and keynote talks from prominent social scientists and technologists. The ICWSM Workshop program will return in 2016 and will be held on the first day of the conference, May 26.

Please see individual websites for workshop submission deadlines. Registration information will be available at the ICWSM-15 website in March. The early registration deadline is April 10, and the late registration deadline is May 1. For full details about the conference program, please visit the ICWSM-15 website (icwsm.org) or write to icwsm15@aaai.org.

Wang, W.; Chen, L.; Thirunarayan, K.; and Sheth, A. 2012. Harnessing Twitter Big Data for Automatic Emotion Identification. In *Proceedings of the IEEE 2012 International Conference on Social Computing,* 587–592. Piscataway, NJ: Institute for Electrical and Electronics Engineers.

Krishnaprasad Thirunarayan (knoesis.org/tkprasad) 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 3.0.

Amit Sheth is the LexisNexis Ohio Eminent Scholar and directs the 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 the semantic web.