Semantics-empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications

Krishnaprasad Thirunarayan and Amit Sheth
Kno.e.sis: Ohio Center of Excellence in Knowledge-enabled Computing
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 V’s 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 new concepts, entities and facts. To handle Veracity, we exploit 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) [CPS-URL] [Sheth et al, 2013] are a new revolution in sensing, computing and communication that brings together a variety of resources ranging from networked embedded computers and mobile devices to multimodal data sources including sensor and social; multiple domains such as medical, geographical, environmental, traffic, and behavioral; and diverse situations and application areas such as system health monitoring, chronic medical condition management, disaster response and threat assessment. The modeling and computing challenges associated with PCSS can be organized in terms of the 5 Vs of Big Data (that is, volume, variety, velocity, veracity and value), which also provides a means to organize our research efforts in addressing Big Data challenges using semantics, network and statistics-empowered Web 3.0. Specifically, in this paper, we discuss the nature of Big Data and address the role of semantics in analyzing and processing Big Data.

Characteristics of the PCSS Big Data
We discuss the primary characteristics of the Big Data problem as it pertains to the 5 V’s.

Volume
The sheer number of sensors and the amount of data reported by sensors is enormous and growing rapidly. For example, 40+ billion sensors have been deployed as of now and about 250TB of sensor data are generated for a NY-LA flight on Boeing 737 [Gigaom-URL]. Parkinson disease dataset that tracked 12 patients with mobile phone sensors over 8 weeks is 150GB in size [MJFF-PC]. However, availability of fine-grained raw data is not sufficient unless we can analyze, summarize or abstract them in meaningful ways that are actionable. For example, from a pilot’s perspective, the sensors data processing should yield insights such as whether the Jet Engine and the flight control surfaces are behaving normally or is there cause for concern? Similarly, can we 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 amount of 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 et al, 2012].

In summary, the volume of data to be processed on available resources creates the following challenges: (1) Ability to abstract the data in a form that summarizes the situation and is actionable, that is, semantic scalability [Sheth, 2011] to transcend from fine-grained machine-accessible data to coarse-grained human comprehensible and actionable abstractions; and (2) Ability to scale computations to take advantage of distributed processing infrastructure and to reason efficiently on mobile devices where appropriate.

**Variety**

PCSS generate and process a variety of multimodal data using heterogeneous background knowledge to interpret the data. For example, traffic dataset (such as from 511.org) contains numeric information about vehicular traffic on roads (e.g., speeds, 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 dataset (such as from Mesowest) provide numeric information about primitive phenomenon (e.g., temperature, precipitation, wind speed) that are required to be combined and abstracted into human comprehensible weather features in textual form. Geosciences datasets also exhibit lot of syntactic diversity, local vocabularies and 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 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) [Accident-URL]. Similarly, our inability to identify 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 suspects inflicted excruciating (albeit needless) pain on the Tripathi family and friends [Misinform-URL].

In summary, the variety in data formats and the nature of available knowledge creates the following challenges: (1) Ability to integrate and interoperate with heterogeneous data (to bridge syntactic diversity, local vocabularies and models, and multimodality); and (2) Semantic scalability to transcend from fine-grained machine-accessible data to coarse-grained human comprehensible, actionable abstractions. (Note that the latter research agenda addresses both volume and variety challenge.)

**Velocity**

Handling of sensor and social data streams in PCSS requires online (as opposed to offline) algorithms to efficiently crawl and filter relevant data sources, detect and track events and anomalies, and collect and update relevant background knowledge. For instance, Wikipedia event pages can be harnessed for relevance ranking of Twitter hashtags, and appropriate physical sensors can be selected and tasked to monitor the health of civil infrastructures during an unfolding disaster situation. Similarly, it is important to determine the entities to be tracked in the context of a natural disaster or a terror attack. For example, during Hurricane Sandy, tweets indicate possible flooding of a subway station, which can lead to finding relevant locations using open source data (e.g., [NYData-URL]), which in turn can help identify sensors that can stream additional data giving us real-time information.

In summary, the rapid change in data and trends creates the following challenges: (1) Ability to focus on and rank the relevant data; (2) Ability to process data in an incremental manner; and (3) Ability to cul, evolve, and hone in on relevant background knowledge.

**Veracity**

PCSS systems receive data from sensors subject to 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 in the context of homogeneous sensor networks, while semantic models are necessary for heterogeneous sensor networks [Thirunarayan et al, 2013]. For instance, in the context of applications that involve both humans and sensors systems, the relevant data; (2) Ability to pr...
Value
A key challenge in promoting a PCSS system as an infrastructure, data acquisition and remote monitoring system to a system that provides actionable information and aids humans in decision making is the acquisition, identification (e.g., relevant knowledge on Linked Open Data (LOD)), construction and application of relevant background knowledge needed for data analytics and prediction. The “top down” declarative knowledge creation and curation require experts and the crowds, while “bottom up” statistical knowledge elicitation from raw data requires both domain models and machine learning techniques. Significant benefit of using domain-specific knowledge in addition to machine learning techniques is now well appreciated (e.g., the early work on semantic annotation and search in [Hammond et al, 2002]). Similarly, for leveraging sensor data streams for personalized healthcare, in an effort to reduce readmission rates among cardiac patients, to improve quality of life among asthmatic patients, and to monitor the progression of Parkinson’s disease, requires a hybrid of statistical techniques and declarative knowledge. Similar approaches can be used to leverage Electronic Medical Record (EMR) data to fill gaps in existing declarative knowledge (e.g., [Perera et al, 2012]), and social media data to predict entity-specific sentiments and emotions. Ultimately, the raison d’etre of all analytics on environmental, medical, system, social, and lifestyle data is to derive situational awareness and from it valuable nuggets of wisdom for decision making.

In summary, extracting value using data analytics on sensor and social data creates the following challenges: (1) Ability to acquire and apply knowledge from data and integrate it with declarative domain knowledge; and (2) Ability to learn and apply domain models from novel sensor streams for classification, prediction, decision making, and personalization.

Role of semantics in PCSS Big Data processing
We discuss examples of our early research in developing semantics-empowered techniques to overcome the Big Data problem organized around the 5V’s, while fully realizing that it will require a more extensive survey paper to recognize and organize extensive amount of research many in our community are concurrently pursuing. Most of the examples are from Kno.e.sis’ active multidisciplinary projects summarized at [Knoesis-Multi-URL].

Addressing volume: Semantic scalability
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 [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. The existing approaches use rules that are deductive in nature and are unable to systematically guide towards minimizing incompleteness. 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 that not only enables combining and abstracting multimodal data but also enables seeking relevant information that can reduce ambiguity and minimize incompleteness, a necessary precursor to decision making and taking 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. Specifically, the last step requires mapping causes into categories that tie to the nature of action to be taken and in a form that is easily accessible to the decision maker. 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 naïve heart failure patient might need from a health monitoring application is a suggestion such as whether the condition is as normally expected, or requires a call to a nurse, or a visit to a doctor, or hospitalization. In fact, such “risk assessment” issues arise naturally and similar frameworks can be applied in a wide variety of areas, with different degrees of maturity. As exemplified below, the first two examples can be formalized using our approach with demonstrable benefits, while the subsequent examples require research in high-fidelity models and human mediation for fruition.

(1) Weather use case: Determining and tracking weather features from weather phenomenon, potentially tasking sensors if additional information is necessary for disambiguation.

(2) Health care use case (Diagnosis, Prevention and Cure): Determining disorders afflicting a patient-- their degree of severity and progression -- by monitoring symptoms, normally requires additional physiological observations, personal feedback (e.g., about feeling giddy or tired or depressed that cannot always be ascertained through physical/chemical means), and/or laboratory test results, beyond initial patient reported observations, for disambiguation. For example, consider the sensor data streams resulting from continuous monitoring of patients suffering from Parkinson’s disease, chronic heart failure, diabetes, asthma, etc. These can be further enhanced by
monitoring adherence/compliance to prescribed treatment, and by generating suggestions for avoidance of aggravating factors (to improve the quality of life when a cure is not an option).

(3) Sensor/Social Data summarization use case: Determining patterns in data for generating summaries, potentially requiring conflict resolution techniques and additional probing for disambiguation based on background knowledge.

(4) Threat use case: Determine threats from various evidences and vulnerabilities, subject to historical and surveillance data, cultural and behavioral models.

Some specific research goals to be pursued (that also overlaps with approaches to meet the variety challenge) include: (1) Development and codification of high-fidelity background knowledge. 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 and faithfully formalize the reality of the situation, as opposed to an overly simplified version of the situation (expressive semantic representation). (2) Development of relevant background knowledge that connects active and passive sensor data from readily available sensors to the situation they reflect (that is, acquisition of sensor data patterns and their implications). In other words, how can we determine high-level activities from low-level sensor data by gleaning and characterizing data patterns? (3) Using contextual information and personalization. The interpretation of data is based on contextual information. For example, the notion of anomalous traffic depends on the location and the time of the day. This type of spatio-temporal-thematic contextual knowledge is integral to an accurate interpretation of observations for decision-making. In medical scenarios, effective treatment also requires personalization on patient specific 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 EMR (contextual and personalized interpretation). (4) Effective summarization and justification of recommended action. One of the problems resulting from indiscriminate sensing and logging of observed data is that we have “too much data to assimilate but not enough knowledge to act”. Furthermore, as sensing, mobile computing, wireless networking and communication technologies are becoming cheap and ubiquitous (as embodied by the Internet of Things), we also run the risk of being drowned in the noise [IoT-URL]. 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 to overcome (summary/risk score). (5) Efficient perceptual reasoning on resource-constrained devices: In order to provide “intelligent computing at the edge”, we investigate techniques to collect the data at the edge, intelligently process them using background knowledge and reasoning, and then send back the essential information. For example, this is required to address privacy concerns, need for timely and ubiquitous access to data, given wide availability of wireless mobile devices. Its realization will also spur use of innovative bit-vector-based indexing techniques and graph-based algorithms for performing inferences on resource-constrained devices (efficient perceptual inference) [Henson et al., 2012].

To address knowledge representation and reasoning challenges, semantic technologies can be used to represent facts and assertions, and domain ontologies can be used to encode background knowledge. Besides manually curated ontologies, Linked Open Data (LOD) and Wikipedia can be harnessed to overcome both syntactic and semantic heterogeneity with applications ranging from social media to Internet of Things.

Addressing velocity: Continuous semantics
Formal modeling of evolving, dynamic, domains and events is hard for at least two reasons. First, we do not have many existing ontologies to use as starting point. Second, a diverse set of users will have difficulty committing to the shared worldview, further exacerbated by contentious topics. Building domain models for consensus in such circumstances requires us to pull background knowledge from trusted, uncontroversial sources. Wikipedia, for instance, has shown that it is possible to collaboratively create factual descriptions of entities and events even for contentious topics. Such wide agreement, combined with a category structure and link graph, makes Wikipedia an attractive candidate for knowledge extraction and subsequent enrichment. That is, we can harvest the wisdom of the crowds, or collective intelligence, to build light-weight ontology—an informal domain model—for use in tracking unfolding events, by classifying, annotating and analyzing streaming data. The key challenge to overcome is the creation of relevant domain model on demand very quickly to be useful for semantic searching, browsing, and analysis of real-time content. As part of continuous semantics agenda [Sheth et al., 2010], our research seeks dynamic creation and updating of semantic models from social-knowledge sources such as Wikipedia [Knoesis-Extract-URL] and Linked Open Data (LOD) that 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 cycle to forecasting, tracking and monitoring the aftermath of disasters (such as hurricanes and earthquakes).

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 resulting in fine-grained sensor data streams. This is unprecedented, and hence, the appropriate background knowledge to analyze such multimodal data has not yet been codified. That is, domain models capturing cause-effect relationships and associations among features that are relevant to 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 because they enable 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.

There have been a number of attempts at learning domain models from data as well as specify them declaratively. 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. 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. The data-driven approach uncovers correlations among parameters. In contrast, the declarative approach provides more dependable, human curated facts and rules that can be expected to capture valid relationships. Furthermore, data-driven and declarative approaches can provide complementary information. 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, semantic integration of sensor and social data, using multiple domain ontologies, Semantic Sensor Network ontology, and our IntellegO perceptual reasoning infrastructure, can improve situational awareness.

In the context of big data generated by PCSS, 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, as suggested by David Brooks of New York Times [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, it is also important to guard against unintentional data skew or improper sampling (e.g., to understand the life cycle of Sun (resp. human), observations over a human’s (resp. fruit fly’s) life span is woefully inadequate), by whetting gleaned correlations, before being put into practice for prediction. In general, big data can be noisy, skewed, inaccurate, and incomplete.

Correlations between two concepts can arise for various different reasons such as: (i) Causal (or due to common origin or shared cause) that is consistent with cause-effect declarative knowledge (e.g., tides and ebbs are caused by the alignment of earth, sun and moon, around full moon and new moon; “anomalous” orbits of Solar system planets w.r.t. the “circular” motion of stars in geocentric theory (‘planet’ is ‘wanderer’ in Greek) was significantly simplified and satisfactorily explained by heliocentrism and theory of gravitation, and subsequently, the “anomalous” precision of Mercury’s orbit clarified by General Theory of Relativity; C-peptide protein can be used to estimate insulin produced by a patient’s pancreas); (ii) Coincidental due to data skew or misrepresentation such as the “data-empowered” conflicting claims about the impact of an economic policy in politically charged debates [Klass, 2008] [Cayo, 2013], or improper use of historical precedents with temporal nearness and chance over shadowing actual similarity [Stauffer, 2002][Christensen, 1997]; (iii) Coincidental new discovery (e.g., the market basket problem that associated beer and diapers, Wal-Mart executives who associated approaching hurricanes with people buying large quantities of Strawberry Pop-Tarts [Brooks-2, 2013]); or (iv) Anomalous and accidental (e.g., Since the 1950s, both the atmospheric Carbon Dioxide level and obesity levels have increased sharply. Hence, atmospheric Carbon Dioxide causes obesity [Correlation-URL].) Pavlovian learning induced conditional reflex, and some of the financial market moves, seem to be classic cases of correlation turning into causation! Even though correlations can provide some valuable insights, they can at best serve as valuable hypothesis or deserve explaining from a background theory before we can have full faith in them. For instance, consider the controversies surrounding
assertions such as ‘smoking causes cancer’, ‘high debt causes low growth’, ‘low growth causes high debt’, and ‘religious fanaticism breeds terrorists’.

Combining data-driven 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 include: (i) Gleaning of data-driven qualitative dependencies and integration with qualitative declarative knowledge, which are at the same level of granularity and abstraction, and (ii) Use of these seed models to learn parameters for reliable fit with the data. For instance, 511.org traffic data can be analyzed to obtain progressively expressive models starting from gleaning undirected correlations among various parameters, to updating it further using declarative knowledge from ConceptNet to orient the dependencies among parameters, to quantifying dependencies. The hybridization of qualitative and quantitative analysis should eventually provide an ability to rank various options in a manner that is human comprehensible for decision making and acting. In more technical terms, we investigate principled ways to integrating declarative approach with progressively expressive probabilistic models for analyzing heterogeneous data such as: (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 further take into account the evolution of PCSS. We have applied this approach to fine-grained analysis of Kinect data streams obtained from 10 sensors for homogeneous populations to predict whether the population consists of humans or aliens [Koller, 2012]. In the long run, such techniques can be scaled and applied for activity recognition that can serve as the foundation for applications ranging from monitoring Parkinson Disease/Alzheimer patients to monitoring traffic and system health.

Addressing veracity: Gleaning trustworthiness
Actionable information from multiple sources requires abstracting, arbitrating, and integrating heterogeneous and sometimes conflicting and unreliable data. A semantics-empowered integration of physical and citizen sensor data can improve assessing data trustworthiness. 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 (e.g., the cautionary tale out of recent Boston Marathon bombing [Misinformed-URL]), but 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 by virtue of the fact that it may also represent abnormal situation. (As an aside, trending topic abuses are common during disasters.) In general, 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 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 aggregate information from multiple sources (including human-in-the-loop) and over time, to compute the trustworthiness of aggregated data and their sources. 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] [Thirunarayan et al, 2013]. Ultimately, the Holy Grail of trust research is to develop expressive trust frameworks that have both declarative/axiomatic and computational specification, and to devise methodologies for instantiating them for practical use, by justifying automatic trust/trustworthiness inference in terms of application-oriented semantics of trust (i.e., in terms of vulnerabilities and risk tolerance) requiring human in the loop.

Deriving value: Evolving background knowledge, obtaining actionable intelligence and decision making
The aforementioned research should yield new background knowledge applicable to PCSS that is rooted in sensor data correlations and that can provide actionable intelligence for decision-making, and ultimately, benefit end users. For specificity, here are some concrete examples of applications benefitted/impacted by our line of research:
(1) Health and wellbeing of patients afflicted with chronic conditions such as heart failure and asthma by empowering patients to be more proactive and participatory in their own health-care. For example, mobile applications enabled by our research have the potential to exploit commonly available sensors to monitor patients and their environment continuously, to help minimize the debilitating effects of asthma by minimizing/preventing contact with allergens, proactively suggesting medications to reduce allergic reaction, and determining/carrying out action plans to build
resistance to or avoid asthmatic attacks. Development of such mobile applications to assist asthma patients requires:

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

(ii) Deploying mobile applications that can use environmental and on-body sensors, background knowledge, and patient health history to prescribe immediate and future course of actions to avoid allergens, improve resistance, and treat symptoms.

(2) Acquisition of new background knowledge to fill gaps in available knowledge and improve coverage by exploiting Electronic Medical Record (EMR) data (e.g., in the cardiology context). Specifically, our research will elicit missing knowledge by leveraging EMR data to hypothesize plausible relationships, gleaned through statistical correlations, for validation by domain experts, which can significantly reduce domain knowledge creation effort, without sacrificing quality and reliability. Note that existing knowledge bases in health care domain are rich in taxonomic relationships, but they lack non-taxonomic (domain) causal relationships [Perera et al, 2012].

Similarly, our research can enable leveraging massive amounts of user generated data in building high-quality prediction models. Specifically, social media data such as from Twitter can be harnessed to develop models for sentiment analysis and fine-grained emotion identification in tweets, and that can be further repurposed across different domains to deal with blogs and documents [Wang et al, 2012].

Recently, researchers discovered drug-drug interaction between the antidepressant, paroxetine, and the cholesterol lowering drug, pravastatin, that causes high blood sugar, by analyzing searches for both terms, and for words and phrases like “hyperglycemia”, “high blood sugar” or “blurry vision” [SideEffects-URL].

(3) The observations and interactions in a PCSS are characterized by: (i) incompleteness due to partial observation from the real world, (ii) uncertainty due to inherent randomness involved in the sensing process (noise in case of machine sensors and bias in case of citizen sensors), and (iii) dynamism from 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 but 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; otherwise, it may be coincidental or due to data skew. Furthermore, declarative knowledge (including causal relationships) is increasingly being published using open data standards on the Semantic Web including knowledge bases and many domain ontologies and data sets published on the LOD cloud. We believe that our research that leverages such knowledge and integrates it with data-driven correlations will increase the fidelity of graphical models, which will in turn improve predictive and analytical power.

Conclusions

In this paper, we outlined how semantic models and technologies can be, and in many cases are being, used to address various problems associated with big data -- volume by enabling abstraction to achieve semantic scalability (for decision making), variety by enabling overcoming syntactic and semantic heterogeneity to achieve semantic integration and interoperability, velocity by enabling ranking to achieve semantic filtering and focus, veracity by cross checking multimodal sensor data with semantic constraints, and value by enriching semantic models to make them more expressive and comprehensive. Given Kno.e.sis’ empirically driven multidisciplinary research [Knoesis-multi-URL], 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, traffic and smart city infrastructure.

Acknowledgements

We are grateful to Pramod Anantharam for enlightening discussions on hybridization of statistical and logic-based techniques, and in dealing with real-world sensor and social data, and to Cory Henson for insights on semantic perception and its applications to analysis of sensor data for health and wellbeing. We also 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 National Science Foundation.

References


