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 explore the formalization of trust models and approaches to glean trustworthiness. The above four Vs of Big Data are harnessed by the semantics-empowered analytics to derive *Value* for supporting practical applications transcending physical-cyber-social continuum.

Introduction

Physical-Cyber-Social Systems (PCSS[1]) (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.

Characteristics of the PCSS Big Data

We discuss the primary characteristics of the Big Data problem as it pertains to the 5 V's. (The first 3 V's 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, 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[2]. Parkinson disease dataset[3] that tracked 12 patients with mobile phone sensors over 8 weeks is 150GB in size. 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? Consider a vehicle check engine light analogy. It alerts a driver when a sensor detects one of many faults. After ensuring that the gas tank cap is screwed tightly, a driver can seek a qualified mechanic to run diagnostics to suggest corrective actions such as fixing the malfunctioning light, rebuilding transmission, or replacing fuel injector. Similarly, can we measure the symptoms of Parkinson's disease using

[1] http://en.wikipedia.org/wiki/Cyber-physical_system

[2] http://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/

[3] https://www.michaeljfox.org/page.html?parkinsons-data-challenge

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, 2012a).

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)(Sheth, 2013) 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., speed, volume, and travel times), as well as textual information about active events (e.g., accidents, vehicle breakdowns) and scheduled events (e.g., sporting events, music events) (Anantharam et al, 2013). Weather datasets (such as from Mesowest) provide numeric information about primitive phenomenon (e.g., temperature, precipitation, wind speed) that are required to be combined and abstracted into human comprehensible weather features in textual form. Geoscience datasets also exhibit lot of syntactic and semantic variety[4]. In medical domain (such as related to 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 about how they are feeling (obtained by interviewing them) can be combined and abstracted to obtain their health and well-being. The available knowledge has a mix of declarative and statistical flavor, capturing both qualitative and quantitative aspects that when integrated can provide complementary and corroborative information (Sheth and Thirunarayan, 2012).

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. (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 (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, 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 indicated possible flooding of a subway station, which can lead to finding relevant locations using open data[5], which in turn can help identify sensors for real-time updates. On the other hand, raw speed of interaction is critical for financial market transactions. Another key challenge is the creation of relevant domain model on demand quickly to be useful for semantic searching, browsing, and analysis of real-time content.

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 quickly (such as incrementally) and respond; and (3) Ability to cull, evolve, and hone in on relevant background knowledge.

Veracity

PCSS 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, applications that involve both humans and sensors systems, it is crucial to have trustworthy aggregation of all data and

[4] http://earthcube.ning.com/

[5] E.g. https://nycopendata.socrata.com

control actions. The 2002 Überlingen mid-air collision[6] occurred because the pilot of one of the planes trusted the human air traffic controller (who was ill-informed about the unfolding situation), instead of the electronic TCAS system (which was providing conflicting but correct course of action to avoid collision). Similarly, 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[7].

In summary, determination of veracity of data creates the following challenges: (1) Ability to detect anomalies and inconsistencies in social and sensor data that can be due to defective sensors or uninformed posters or anomalous situations; and (2) Ability to reason about and with trustworthiness that exploits temporal history, collective evidence, context, and conflict resolution strategies for decision making.

Value

A key challenge in transforming PCSS from 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. For example, a hybrid of statistical techniques and declarative knowledge is beneficial for leveraging sensor data streams for personalized healthcare, to reduce readmission rates among cardiac patients, to improve quality of life among asthmatic patients, and to monitor the progression of Parkinson's disease. Similarly, hybrid approaches can also be used to leverage Electronic Medical Record (EMR) data to fill gaps in existing declarative knowledge (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 health, social, and lifestyle data is to derive situational awareness and from it 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 an 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[8].

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. 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 reflect action to be taken and 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 heart failure patients will benefit from are suggestions such as whether the condition is as normally expected, or requires a call/visit to a nurse/doctor, or hospitalization. In fact, such "risk assessment" scenarios arise naturally in a variety of areas. As exemplified below, the first two examples can be formalized using our approach with demonstrable benefits, while the subsequent examples require research into high-fidelity models and human mediation for fruition.

[6] http://en.wikipedia.org/wiki/uberlingen_mid-air_collision

[7] http://www.theatlantic.com/technology/archive/2013/04/it-wasnt-sunil-tripathi-the-anatomy-of-a-misinformation-disaster/275155/

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

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

(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.

(3) *Sensor/Social Data summarization use case*: Determining patterns in data for generating summaries, potentially requiring conflict resolution techniques and additional probing, 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 expressive semantic representation*. For example, in the realm of health care, symptoms and disorders are complex entities with complicated interactions. The acceptable and desirable thresholds for various monitored parameters depend on co-morbidity, especially due to chronic conditions. Any representation must provide the necessary expressivity to accurately formalize the reality of the situation. (2) *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*). 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's historical data and clinician prescribed current protocol (e.g., maintain BP at higher than what is normal for NIH specific guidelines) such as what is in Electronic Medical Records (EMR). (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 information overload. 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[9]. 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. (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 reason with them using background knowledge, and return the essence. For example, this is required to address privacy concerns, need for timely and ubiquitous access to data, using wireless mobile devices. Its realization will also spur use of innovative and specialized inference techniques on resource-constrained devices (Henson et al, 2012a).

Besides using manually curated ontologies and reasoners as discussed above, Linked Open Data (LOD) and Wikipedia can be harnessed to overcome syntactic and semantic heterogeneity with applications from social media to Internet of Things.

Addressing Velocity: Continuous Semantics

Formal modeling of evolving, dynamic, domains and events is hard. First, we do not have many existing ontologies to use as a starting point. Second, diverse users will have difficulty committing to the shared worldview, further exacerbated by contentious topics. Building domain models for consensus requires us to pull background knowledge from trusted, uncontroversial sources. Wikipedia 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. As part of continuous semantics agenda (Sheth et al, 2010)(Sheth, 2011a), our research seeks dynamic creation and updating of semantic models from social-knowledge sources such as Wikipedia[10] and LOD that offer exciting new capabilities in making real-time social and sensor data more meaningful and useful for

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

[10] http://www.knoesis.org/research/semweb/projects/knowledge-extraction

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). We have used our on-demand domain hierarchy creation application, DOOZER, to generate domain models[11] in the realm of oncology, cognitive performance, and ad hoc IR tasks. Orthogonal to the aforementioned issues of quality and precision, we also need to address the speed of response which is beyond the scope of the current paper.

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 between features and data patterns gleaned from the recently available sensors and sensor modalities have not been uncovered and formalized hitherto. Such properly vetted domain models are however critical for prediction, explanation, and ultimately, decision making in real-time from the sensed data. Further, objective physical sensors (e.g., weather sensors, structural integrity sensors) provide quantitative observations. In contrast, subjective citizen sensors (e.g., Tweets) provide qualitative "high-level" interpretation of a situation. For example, a sensed slow moving traffic can result from rush hour, fallen trees, or icy conditions that can be determined from postings on social media. Thus physical and citizen sensors can provide complementary and corroborative information enabling disambiguation. Specifically, we have sought semantic integration of sensor and social data, using multiple domain ontologies and our IntellegO perceptual reasoning infrastructure, to improve situational awareness.

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

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, by vetting gleaned correlations, before being put into practice for prediction. For example, to understand the life cycle of Sun (resp. human), observations over a human's (resp. fruit fly's) life span is woefully inadequate. In general, 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 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) can be significantly simplified and satisfactorily explained by heliocentrism and theory of gravitation, and the "anomalous" precision of Mercury's orbit can be clarified by General Theory of Relativity; C-peptide protein can be used to estimate insulin produced by a patient's pancreas); (ii) *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 overshadowing 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, 2013a)); or (iv) *Anomalous and accidental* (e.g., Since the 1950s, both the

[11] http://www.knoesis.org/research/semweb/projects/knowledge-extraction

atmospheric Carbon Dioxide level and obesity levels have increased sharply. However, atmospheric Carbon Dioxide does not cause obesity, and vice versa[12].) 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 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. In other words, correlations require clear justification that they are not coincidental to inspire sufficient confidence. Hence, discovering "unexpected" correlations, and then seeking a transparent basis for them, seems worthy of pursuit. 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'. Stress/spicy foods are correlated with peptic ulcers, but the latter are caused by Helicobacter Pyroli as demonstrated by Nobel Prize winning works of Marshall and Warren[13].

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 concepts, to updating it further using declarative knowledge from ConceptNet[14] to orient the dependencies among concepts, to quantifying dependencies. The hybridization of qualitative and quantitative analysis should eventually provide an ability to rank various options that is human comprehensible for decision making and acting. We encourage principled ways *to integrate declarative approach with progressively expressive probabilistic models* for analyzing heterogeneous data (Domingo and Lowd, 2009). For example: (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 by building models to predict whether a pose belongs to a human or an alien (Koller, 2012). Such techniques can be applied for activity recognition – 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/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[15]), 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 and political events/upheavals as illustrated by the infamous Kenneth Cole tweet (Anantharam et al, 2012).) It may not be possible to distinguish an abnormal situation from a sensor fault or plausible rumor purely on the basis of observational data (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 deal with data from multiple sources (including human-in-the-loop) and over time, to compute the trustworthiness of aggregated data and their sources. Provenance tracking and representation can be the basis for gleaning trustworthiness (Perez, 2010) (Gil, 2012). 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). The Holy Grail of trust research is to develop *expressive trust frameworks* that have both declarative/axiomatic and

[12] http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

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

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

[15] http://www.theatlantic.com/technology/archive/2013/04/it-wasnt-sunil-tripathi-the-anatomy-of-a-misinformation-disaster/275155/

computational specification, and to devise methodologies for instantiating them for practical use, by justifying automatic trust inference in terms of application-oriented *semantics of trust* (i.e., vulnerabilities and risk tolerance) (Thirunarayan et al, 2013).

Deriving Value: Evolving Background Knowledge, 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 (Sheth, 2013). 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 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 improve coverage by exploiting EMR data (e.g., in the cardiology context). Specifically, our research elicits missing knowledge by leveraging EMR data to hypothesize plausible relationships, gleaned through statistical correlations, for validation by domain experts, which can significantly reduce manual 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 has enabled leveraging massive amounts of user generated content 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 repurposed for use 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"[16].

(3) The observations and interactions in 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 machine sensors and bias in 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 leveraging such knowledge and integrating it with data-driven correlations will increase the fidelity of graphical models, which in turn will improve 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 -- *volume* by enabling abstraction to achieve semantic scalability (for decision making), *variety* by 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 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[17], 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.

[16] http://www.nytimes.com/2013/03/07/science/unreported-side-effects-of-drugs-found-using-internet-data-study-finds.html

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

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

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

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

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

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

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

Christensen, C. M. 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Harvard Business School Press. [Domingo and Lowd, 2009] Domingo, P. and Lowd, D. 2009. *Markov Logic: An Interface Layer for Artificial Intelligence*. San Rafael, CA: Morgan & Claypool.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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