2015

Value Oriented Big Data Processing with Applications

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

FROM INFORMATION TO MEANING

Value-Oriented Big Data Processing with Applications

Krishnaprasad Thirunarayan (T. K. Prasad)

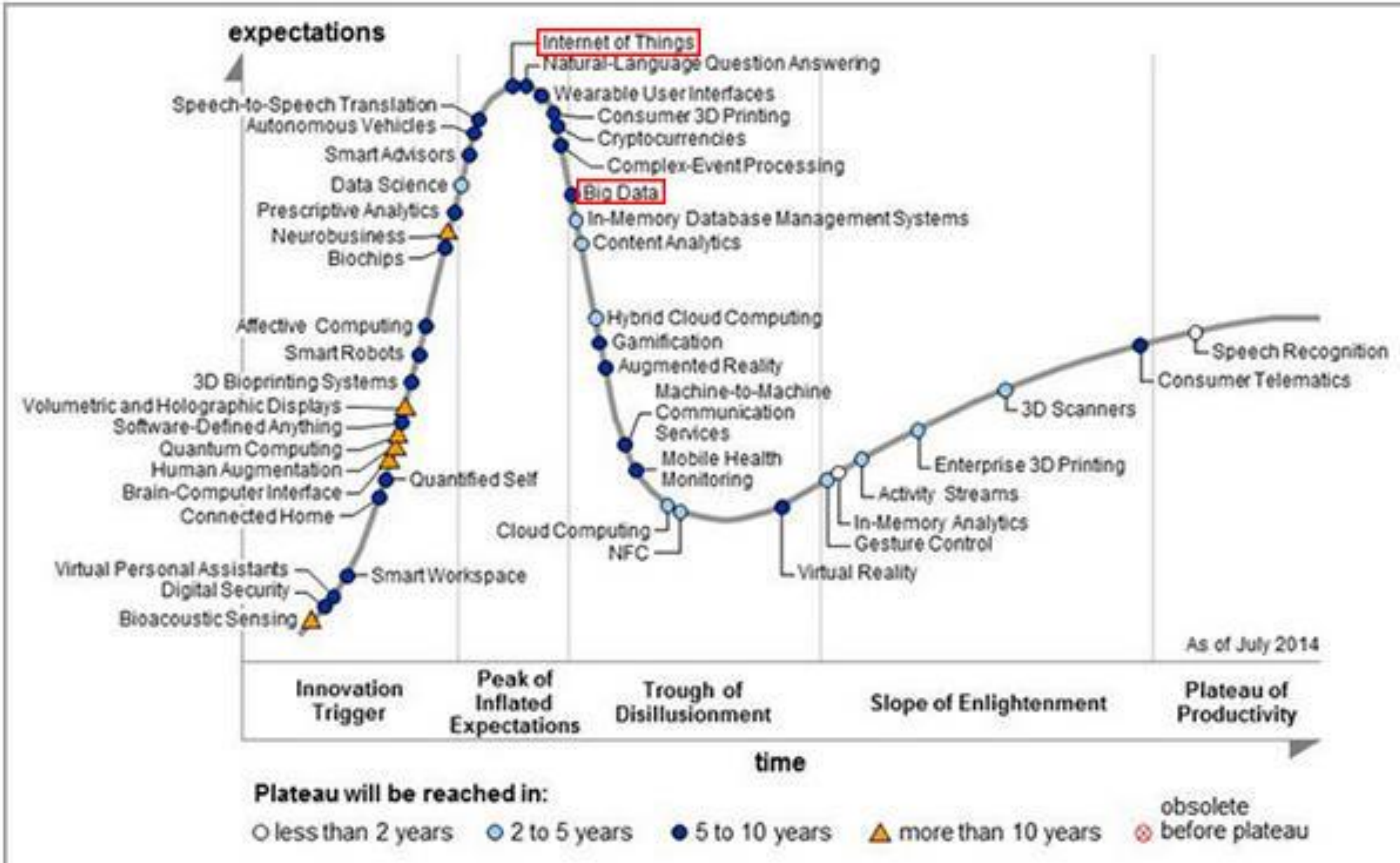
Kno.e.sis – Ohio Center of Excellence in Knowledge-enabled Computing

Wright State University, Dayton, OH-45435

Outline

- 5 V's of Big Data Research
- Semantic Perception for Scalability and Decision Making
- Lightweight semantics to manage heterogeneity
  - Cost-benefit trade-off continuum
- Hybrid Knowledge Representation and Reasoning
  - Anomaly, Correlation, Causation

Gartner's 2014 Hype Cycle for Emerging Technologies



5V's of Big Data Research

Volume

Velocity

Variety

Veracity

Value

Big Data => Smart Data

Volume : Assorted Examples

- 25+ billion sensors deployed by 2015, and 50+ billion by 2020.
- Data Universe would double every two years to reach 40 zettabytes (ZB = 10^{21} bytes) by 2020.
- About 250TB of sensor data are generated for a NY-LA flight on Boeing 737.
- Parkinson disease dataset that tracked 16 patients with mobile phone sensors over 8 weeks is 12GB.

Check engine light analogy

Volume : Challenge

- Sensors (due to IoT) offer unprecedented access to granular data that can be transformed into powerful knowledge. *Without an integrated business analytics platform, though, sensor data will just add to information overload and escalating noise.*

http://www.sas.com/en_us/insights/big-data/internet-of-things.html

Volume : (1) Semantic Perception

- Abstracting machine-sensed data
  - E.g., fine-grained to coarse-grained
  - E.g., average, peak, rate of change
- Extracting human-comprehensible features/entities
- Machine perception
  - Derive conclusions using domain models and hybrid abductive/deductive reasoning

Goal: Human accessible situational awareness and actionable intelligence for decision making

Weather Use Case

- Machine-sensed phenomenon
  - temperature, precipitation, humidity, wind speed, etc.
- Human perceived features
  - blizzard, flurry, rain storm, clear, etc.
  - categories of hurricanes (SSHWS)
- Machine perception
  - Using domain models from NOAA
- *Ultimately, generate weather alerts …*

Parkinson's Disease Use Case

- Data from mobile phone sensors
  - accelerometer, GPS, compass, microphone, etc.
- Human perceived features
  - tremor, poor balance, disturbed sleep, slurred speech, fall, etc.
- Machine perception
  - Using domain models *to be created* to diagnose and monitor disease progression
- *Ultimately, recommend options to control chronic conditions …*

Heart Failure Use Case

- Machine-sensed data
  - Weight change, heart rate, blood pressure, oxygen level, etc.
- Human perceived features
  - Risk-level for hospital readmission of CHF/ADHF patient
- Machine perception
  - Using domain models *to be created* to monitor heart condition of a cardiac patient post hospital discharge
- *Ultimately, recommend treatments to reduce preventable hospital readmissions …*

Asthma Use Case

- Data from machine-sensors
  - Environmental sensors, physiological sensors, etc.
- Human perceived features
  - Asthma severity / control level gleaned from frequency of asthma attacks, wheezing, coughing, sleeplessness, etc.
- Machine perception
  - Using domain models *to be created* to monitor asthma patients and their surroundings
- *Ultimately, recommend prevention, treatment, and control options …*[EVIDENCE-BASED APPROACH]

Traffic Use Case

- Data from machine-sensors, social media stream, and planned event schedules
  - Traffic flow sensors : link speed, link volume, Event-specific tweets, etc.
- Human perceived features
  - traffic delays and congestion, etc.
- Machine perception
  - Using domain models *to be created* to understand traffic patterns in response to events
- *Ultimately, recommend traffic management options …*

Heterogeneity in a Physical-Cyber-Social System

linkid	linkspeed	linkvolume	linkoccupancy	linkdelay	linktraveltime	timestamp
107060	18	-1	-1	-1	74	9/30/12 2:20 PM
107070	18	-1	-1	-1	341	9/30/12 2:20 PM
108150	27	6540	29	-1	244	9/30/12 2:20 PM
108420	36	2548	23	-1	216	9/30/12 2:20 PM
119[illegible]	[illegible]	-1	-1	-1	[illegible]	[illegible]

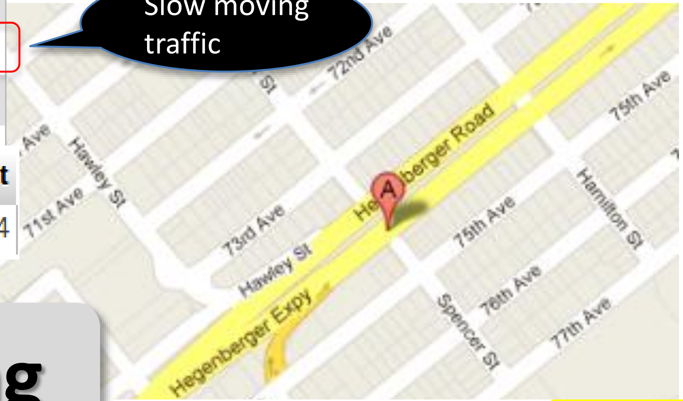
511.org

Slow moving traffic

Link Description

linkid	onstreet	fromstreet	tostreet		speedlimit
108150	I-880 S	66TH AVE	HEGENBERGER RD		104

511.org



Traffic Monitoring

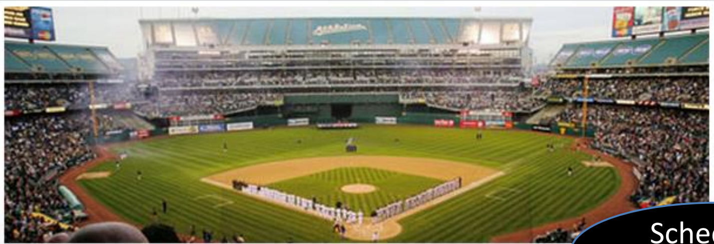
Scheduled Event

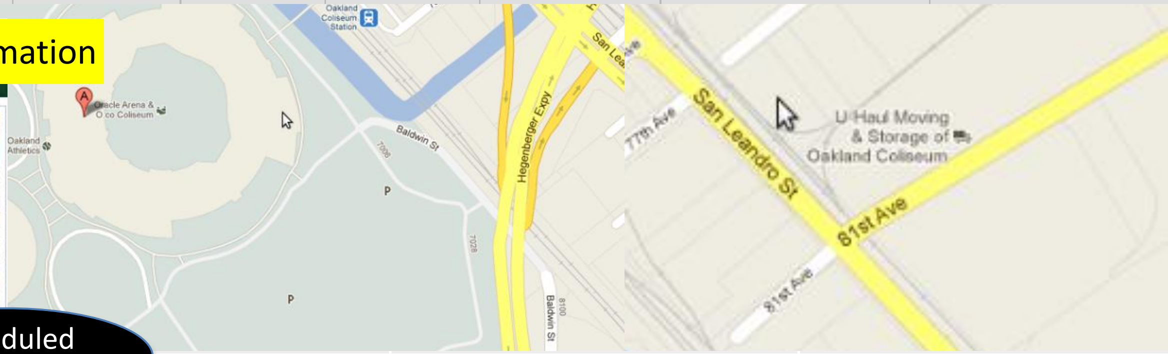
scheduleid	eventtype					ong	starttime	endtime
2012040510161401002076	baseball-game	NULL	NULL	NULL	37750956	-122202232	2012-09-30T11:59:00.0000	2012-09-30T17:00:00.0000

511.org

THE COLISEUM

Schedule Information

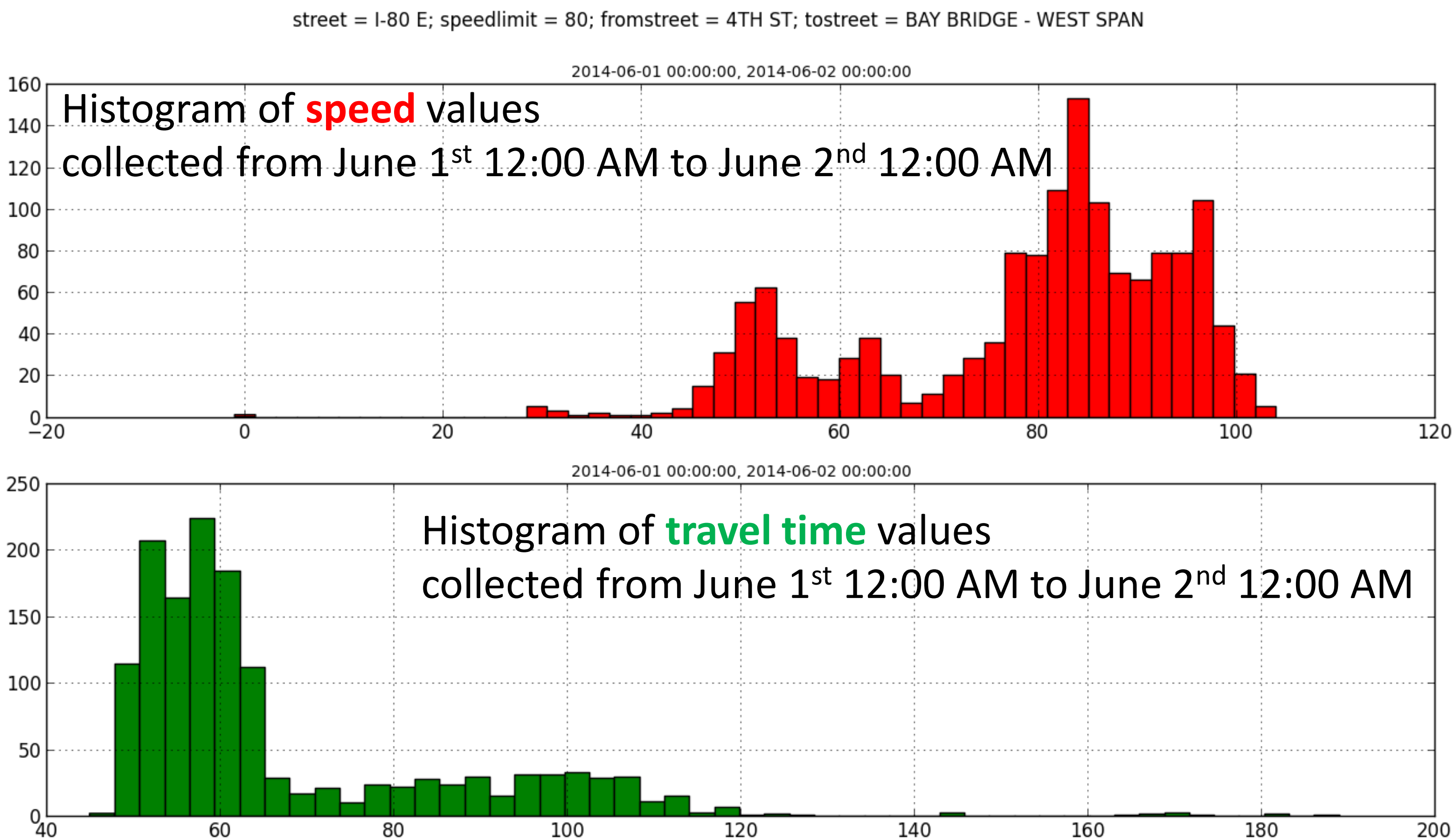




Scheduled Event

Fri, 9/28	Mariners		89-68	Griffin (7-1)	Beavan (10-11)
Sat, 9/29	Mariners	W 7-4	90-68	Balfour (3-2)	Perez (1-3)
Sun, 9/30	Mariners	W 5-2	91-68	Doolittle (2-1)	Kelley (2-4)
Mon, 10/1	Rangers	W 4-3	92-68	Parker (13-8)	Perez (1-4)
Tue, 10/2	Rangers	W 3-1	93-68	Blackley (6-4)	Harrison (18-11)

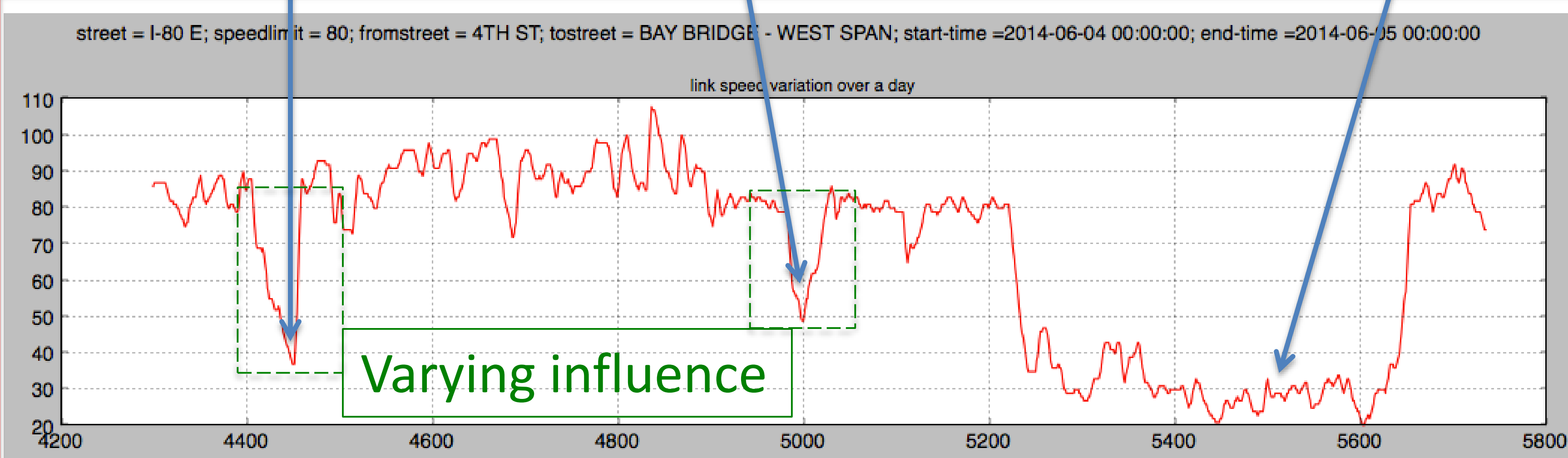
Traffic Data Analysis



Relating Sensor Time Series Data to Scheduled/Unscheduled Events

Multiple events

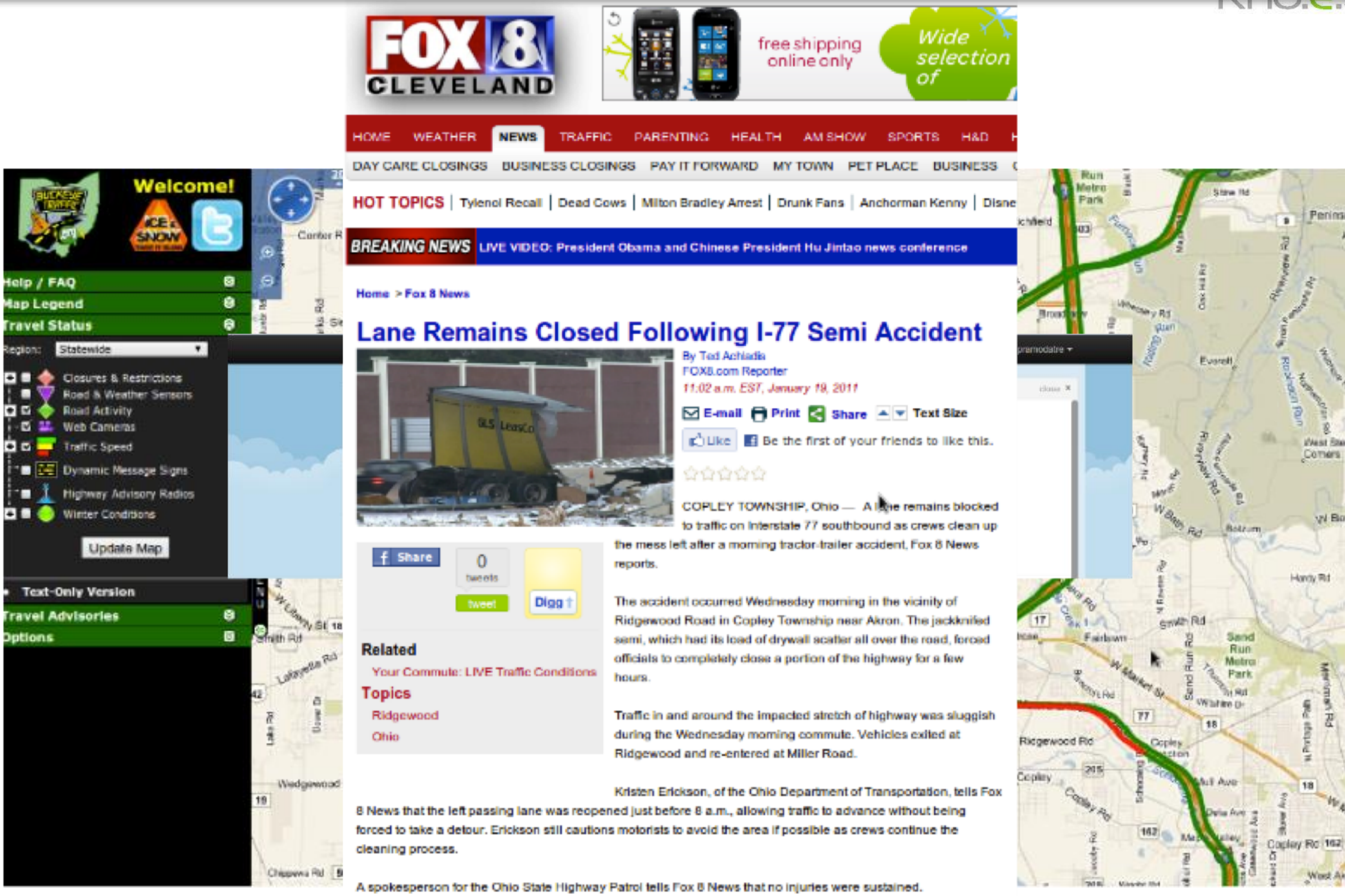
interact with each other



Varying influence

Image credit: http://traffic.511.org/index

Heterogeneity in a Physical-Cyber-Social System



Volume : (2) Exploiting Embarrassing Parallelism

- Cloud Computing
  - Hardware : Networked Stock PCs
  - Middleware: Replicated storage and restarted computations for fault tolerance
    - E.g., Hadoop file system, Google file system
  - Application Programming: Models / languages for distributed computation
    - E.g., Map-Reduce, PIG, HIVE

Volume with a Twist

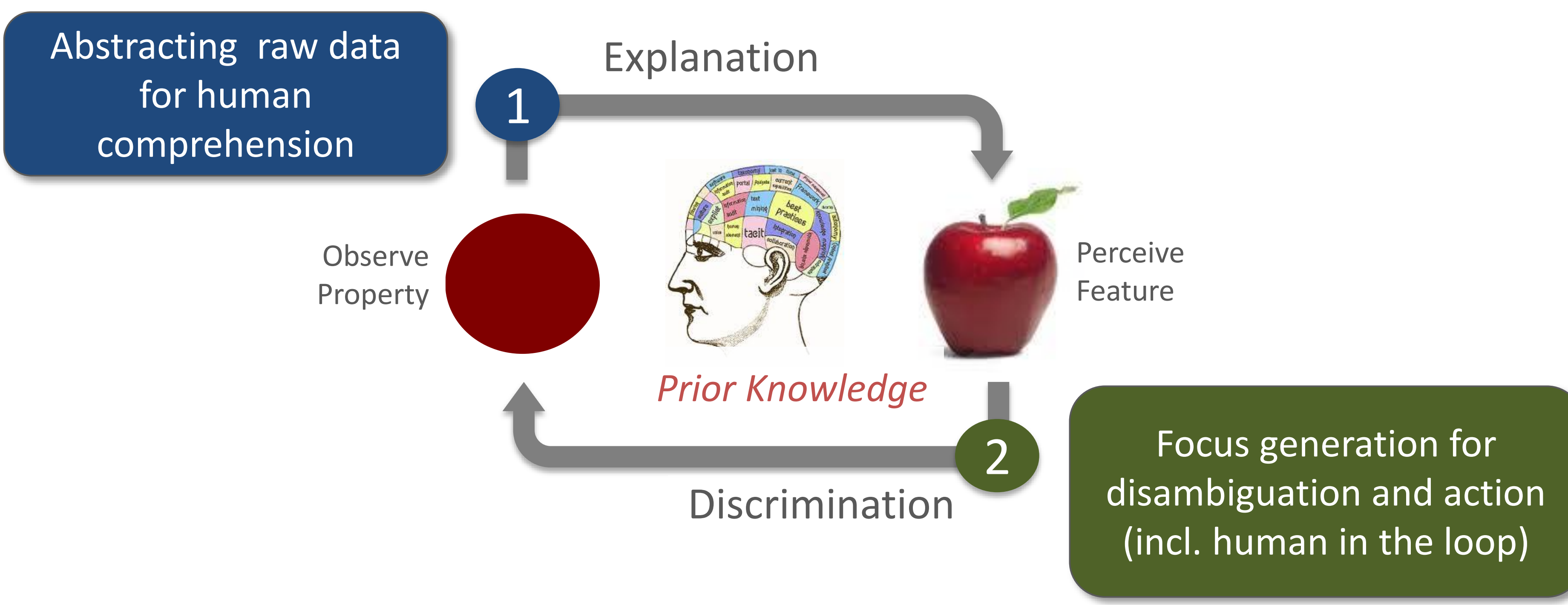
Resource-constrained reasoning on mobile-devices

Goal: Boolean encodings to ensure feasibility, efficiency, and economy

Cory Henson's Thesis Statement

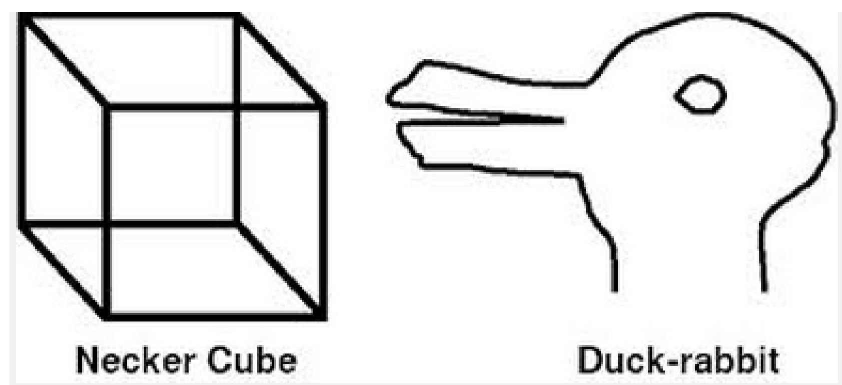
Machine perception can be formalized using semantic web technologies to derive abstractions from sensor data using background knowledge on the Web, and efficiently executed on resource-constrained devices.

Perception Cycle* that exploits background knowledge / domain models



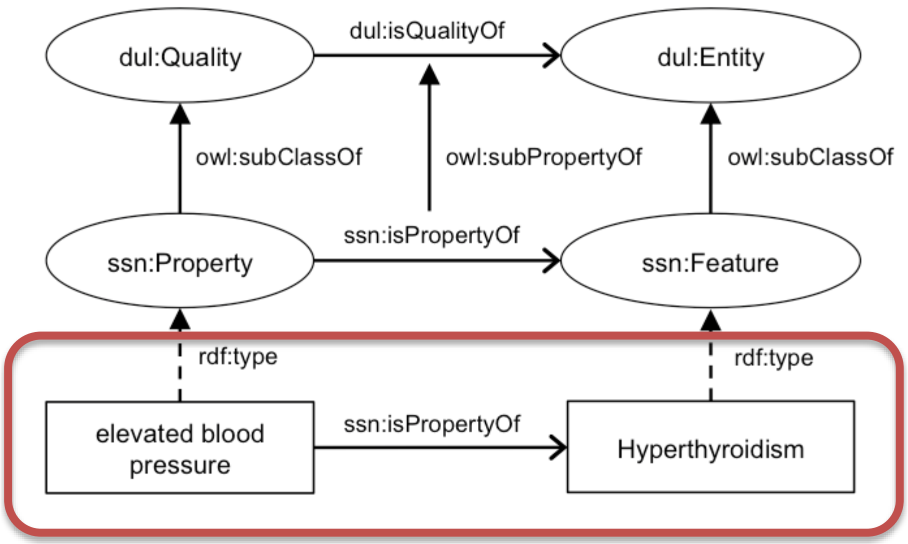
* based on Neisser's cognitive model of perception


12
A B C
14

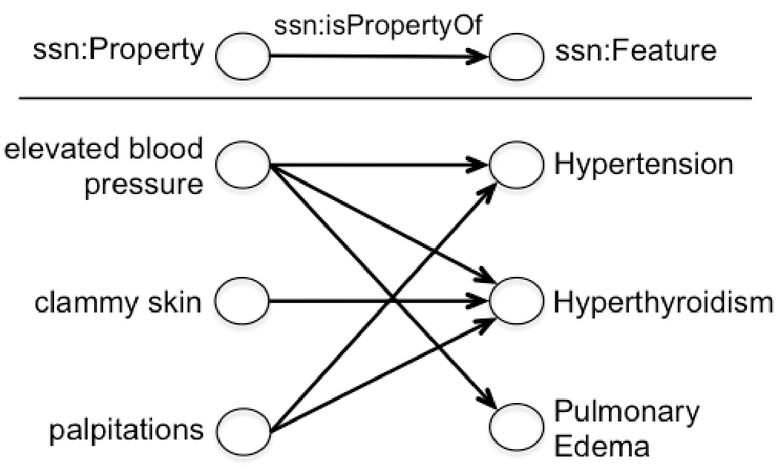

Necker Cube
Duck-rabbit

Prior knowledge on the Web

W3C Semantic Sensor Network (SSN) Ontology

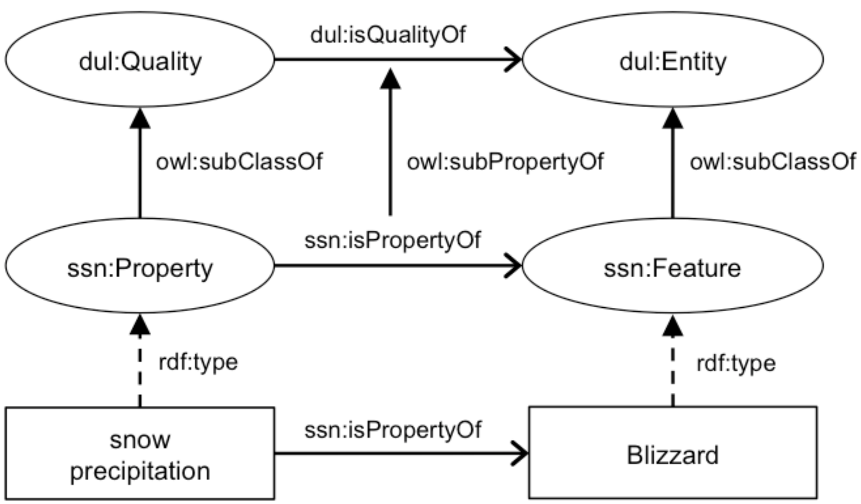


Bi-partite Graph

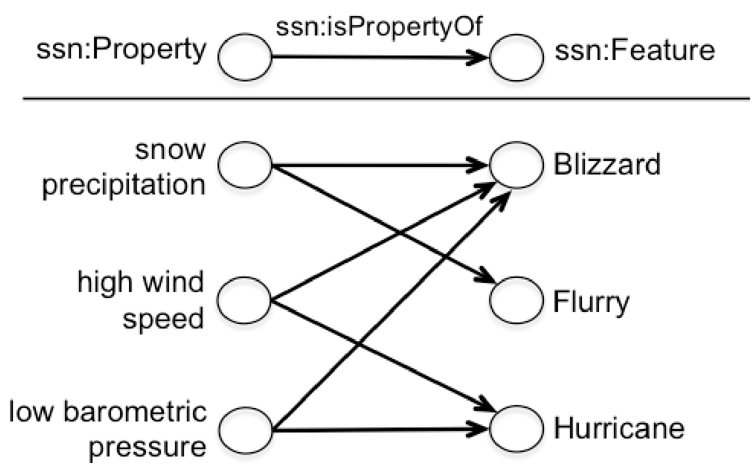


Prior knowledge on the Web

W3C Semantic Sensor Network (SSN) Ontology



Bi-partite Graph

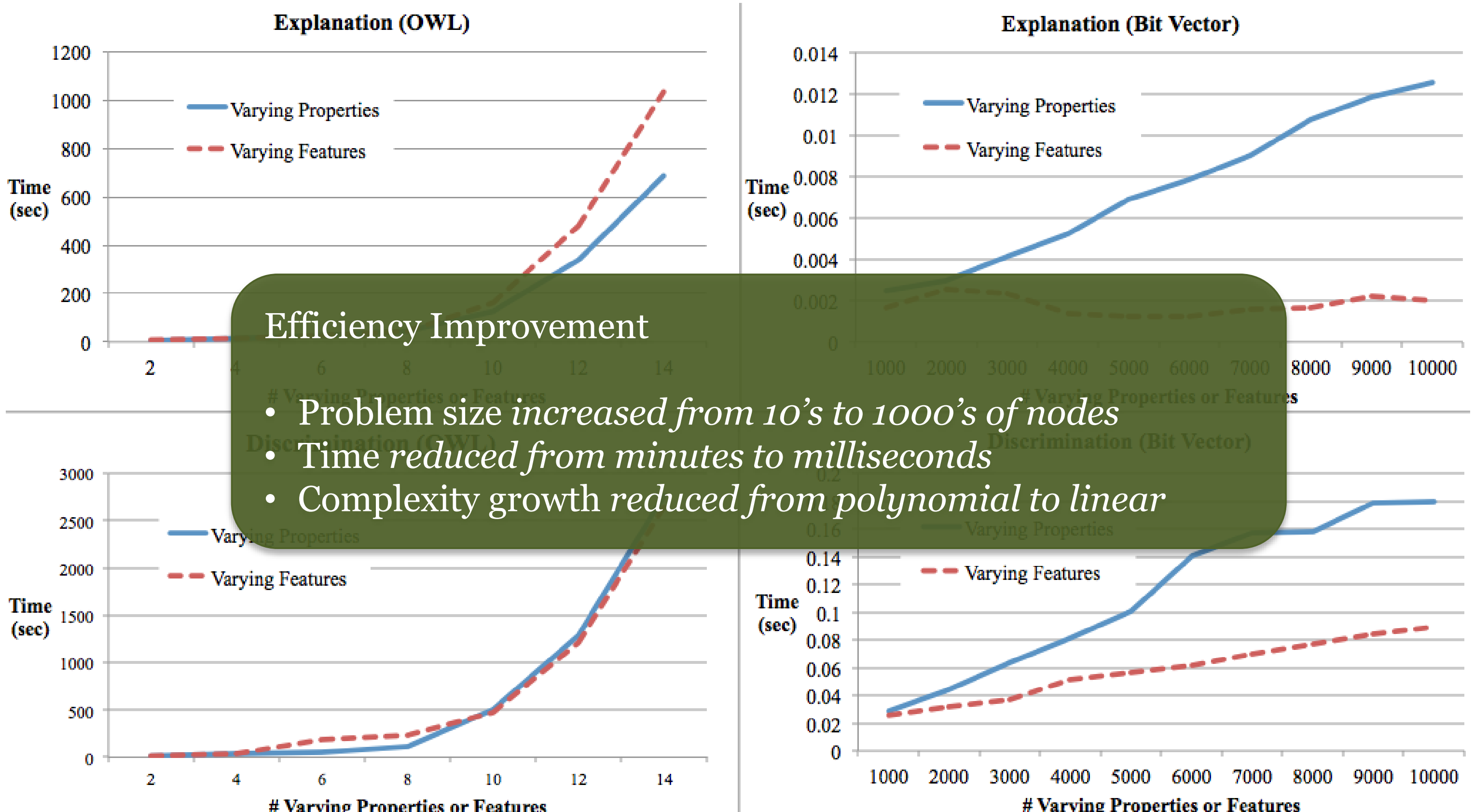


Virtues of Our Approach to Semantic Perception

Blends **simplicity, effectiveness,** and **scalability**.

- Declarative specification of *explanation* and *discrimination;*
- With contemporary relevant applications (e.g., healthcare);
- Using improved encodings/algorithms that are *significant* (asymptotic order of magnitude gain) and *necessary* (“tractable” resource needs for typical problem sizes); and
- Prototyped using extant PCs and mobile devices.

Evaluation on a mobile device



Efficiency Improvement

- Problem size *increased from 10's to 1000's of nodes*
- Time *reduced from minutes to milliseconds*
- Complexity growth *reduced from polynomial to linear*

$O(n^3) < x < O(n^4)$

$O(n)$

Variety

Syntactic and semantic heterogeneity

- in textual and sensor data,
- in (legacy) materials data
- in (long tail) geosciences data

Idea: Semantics-empowered integration

Variety (What?): Materials/Geosciences Use Case

- *Structured Data* (e.g., relational)
- *Semi-structured, Heterogeneous Documents* (e.g., Publications and technical specs, which usually include text, numerics, maps and images)
- *Tabular data* (e.g., *ad hoc* spreadsheets and complex tables incorporating “irregular” entries)

Variety (How?): (1) Granularity of Semantics & Applications

- *Lightweight semantics*: File and document-level annotation to enable discovery and sharing
- *Richer semantics*: Data-level annotation and extraction for semantic search and summarization
- *Fine-grained semantics*: Data integration, interoperability and reasoning in Linked Open Data

Cost-benefit trade-off continuum

Variety (What?) : Sensor Data Use Case

Develop/learn domain models to exploit complementary and corroborative information to obtain improved situational awareness

- To relate patterns in multimodal data to "situation"
- To integrate machine sensed and human sensed data
- Example Application:

SemSOS :

Semantic Sensor Observation Service

Variety: (2) Hybrid KRR

Blending data-driven models with declarative knowledge

- Data-driven: Bottom-up, correlation-based, statistical
- Declarative: Top-down, causal/taxonomical, logical
- Refine structure to better estimate parameters

E.g., Traffic Analytics using PGMs + KBs

Variety (Why?): Hybrid KRR

Data can help compensate for our overconfidence in our own intuitions and reduce the extent to which our desires distort our perceptions.

-- David Brooks of New York Times

However, inferred correlations require clear justification that they are not coincidental, to inspire confidence.

Variety (How?): Hybrid KRR

Blending data-driven models with declarative knowledge

- Structure learning from data
- Enhance structure
  - By refining direction of dependency
    - Disambiguation
    - Filtering
  - By augmenting with taxonomy
    - nomenclature and relationships
- Improved Parameter learning from data

E.g., Traffic Analytics using PGMs + KBs

WRIGHT STATE UNIVERSITY

KNO.E.SIS

Anomalies, Correlations, Causation

- Due to common cause or origin
  - E.g., Planets: Copernicus > Kepler > Newton > Einstein
- Coincidental due to data skew or misrepres… tation
  - E.g., Tall policy claims made by politicia…
- Coincidental new discovery
  - E.g., Hurricanes and St… …arts Sales
- Strong correlati… …ation
  - E.g., S… …elicobacter Pyroli : Stomach Ulcers
- A… …d accidental
  - … CO_2 levels and Obesity
- Correlation turning into causations
  - E.g., Pavlovian learning: conditional reflex

Paradoxes: The Seeds of Progress

Veracity

Lot of existing work on Trust ontologies, metrics and models, and on Provenance tracking

- Homogeneous data: Statistical techniques
- Heterogeneous data: Semantic models

Open Problem:

- Develop (application-specific but defensible) semantics of trust using expressive frameworks that are both declarative and computational
- To make explicit all aspects that go into trust formation, to inspire confidence in inferences

Veracity: Confession of sorts!

Trust is well-known,

but is not well-understood.

The utility of a notion testifies not to its clarity but rather to the philosophical importance of clarifying it.

-- Nelson Goodman

(Fact, Fiction and Forecast, 1955)

(More on) Value

Learning domain models from “big data” for prediction

E.g., Harnessing Twitter "Big Data" for Automatic Emotion Identification

Idea: Exploit tweets with “emotion-hashtag” as training dataset

(More on) Value

Discovering gaps and enriching domain models using data

E.g., Data driven knowledge acquisition method for domain knowledge enrichment in the healthcare

Idea: Use associations between diseases, symptoms and medications in EMR documents

Conclusions

- Glimpse of our research organized around the 5 V's of Big Data
- Discussed role in harnessing Value
  - Semantic Perception (Volume)
  - Continuum of Semantic models to manage Heterogeneity (Variety)
  - Hybrid KRR: Probabilistic + Logical (Variety)
  - Continuous Semantics (Velocity)
  - Trust Models (Veracity)

Kno.e.sis: Ohio Center of Excellence in Knowledge-enabled Computing

Thank You

http://knoesis.wright.edu/tkprasad

Krishnaprasad Thirunarayan, Amit P. Sheth: Semantics-Empowered Big Data Processing with Applications. AI Magazine 36(1): 39-54 (2015)

Special Thanks to: Pramod Anantharam, Dr. Cory Henson

Department of Computer Science and Engineering
Wright State University, Dayton, Ohio, USA