

# Knowledge Graphs and Their Central Role in Big Data Processing: Past, Present, and Future

7th ACM India Joint International  
Conference on Data Science & Management of Data (CODS-COMAD)  
Indian School of Business, Hyderabad Campus  
5-7 January 2020

Prof. Amit Sheth

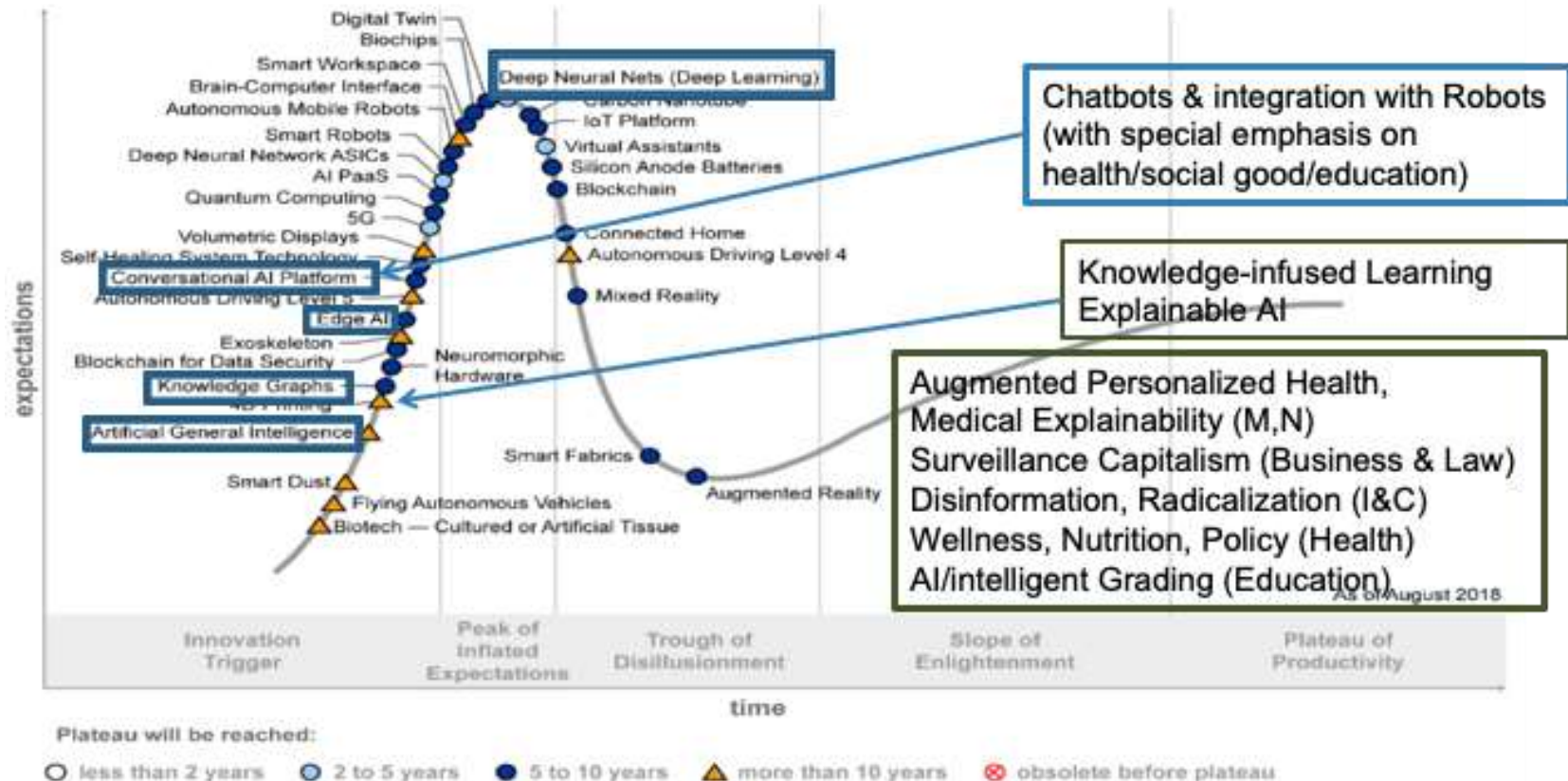
<http://ai.sc.edu/>



**Artificial Intelligence Institute**  
University of South Carolina

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

# AI Institute @ UofSC & AI Landscape



# Interdisciplinary AI: Sample Collaborations

**College of Arts and Sciences:** on misinformation and radicalization (with the Institute of Mind and Brain, Comp Sc & Engg): a MURI proposal submitted (budget \$6.25million) o recovery and resilience from natural disasters (with Hazards and Vulnerability Research Institute, ext partners)

**College of Pharmacy** (includes the South Carolina Colon Cancer Prevention Network): early onset of colorectal cancer on Digestive Inflammation Index

**College of Education:** develop personalized education plan and Analogy based learning in undergraduate courses (eg, biochemistry): using chatbots (virtual agents) for the health of school children through nutrition monitoring

**Mech Engineering** - using chatbot/AI in 2+2 educational in connected manufacturing



# Collaborations...continued

**School of Medicine and College of Nursing:** personalized digital health with virtual assistants (chatbots) for (a) pediatric patients with neutropenia, (b) asthma in children, (c) obesity and hypertension in adults (d) mental health in adults

**Arnold School of Public Health:** o brain imaging (with Aging Brain Cohort and Aphasia Laboratory), o brain aneurysm/ strokes o identification and management of mental health including depression and suicide ideation

**School of Journalism and Mass Communications:** social media insights and harassment of journalists (with Social Media Insights Lab)



# Knowledge - critical role in intelligent computing

**Knowledge - identifying, representing, capturing/creating, representing:** ontology, knowledge graph

**Associating meaning to data, making data interpretable:** e.g., semantic annotations

**Supporting planning, reasoning, analysis, insight, decision making:** semantic applications (search, browsing, ....), improving/enhancing NLP and ML, to activities that support humans

With massive growth in Big Data, this talk is about the central role of Knowledge in computing - in the past, present and future





*The role of knowledge in computing has long been recognized  
- at least since Vannever Bush's 1945 seminal piece: **As We May Think**.*

---

## Knowledge Graphs (KG)

is a structured knowledge in a graphical representation.

## Knowledge Networks (KN)

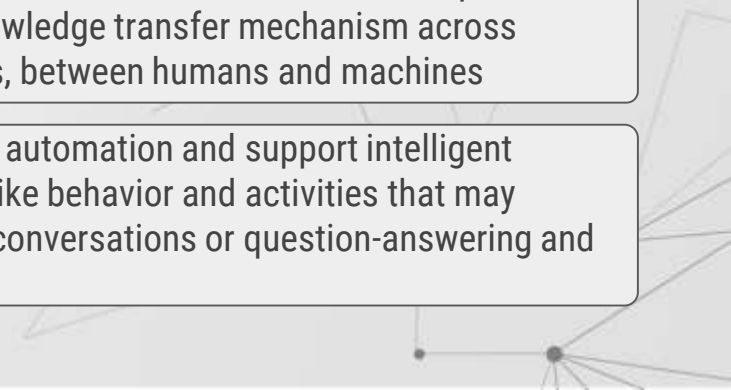
integrate and combine knowledge (usually captured as KGs) from various domains.

Enhanced (semantic) applications such as search, browsing, personalization, recommendation, advertisement, and summarization.

Improve integration of data, including data of diverse modalities and from diverse sources.

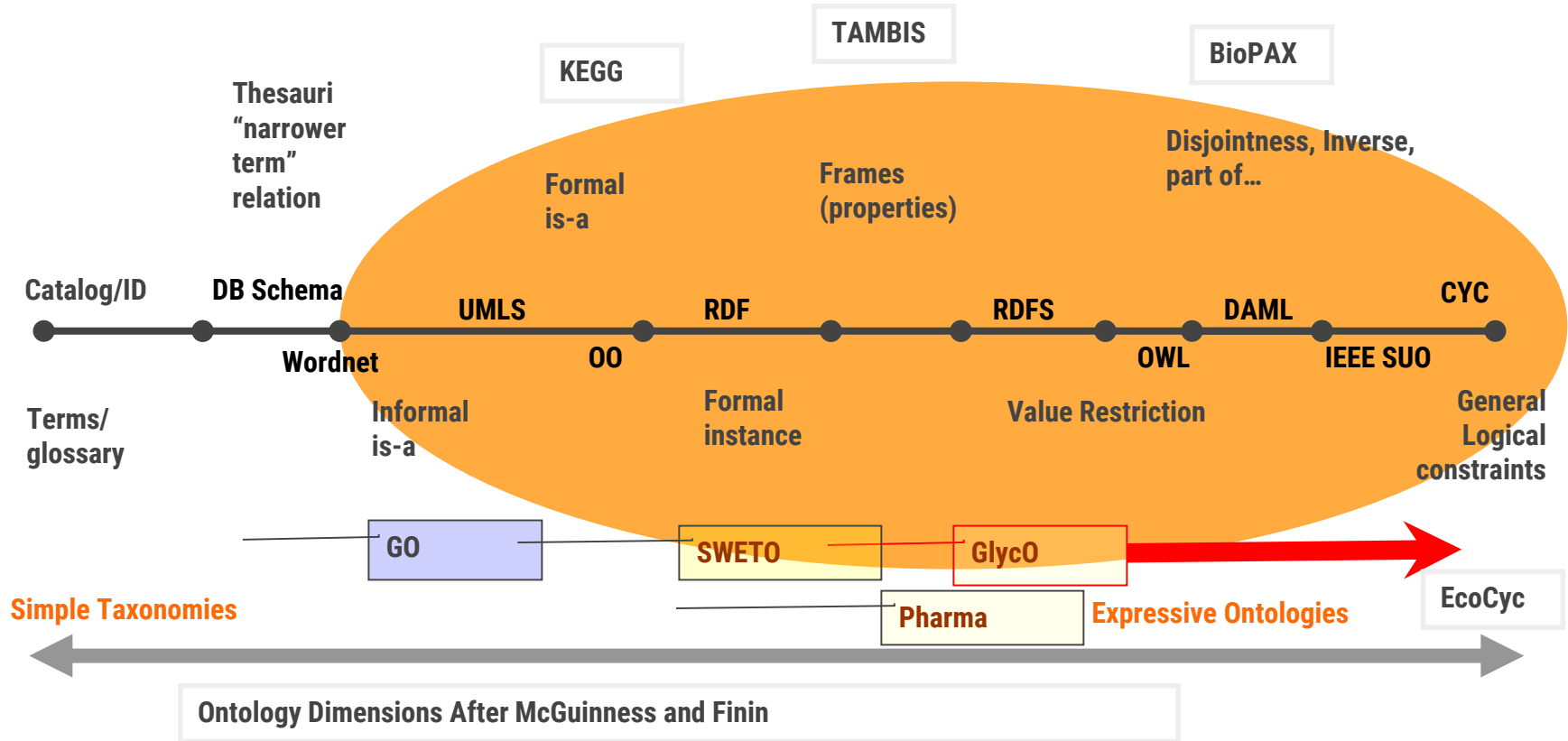
Empower/enhance ML and NLP techniques. Use as a knowledge transfer mechanism across domains, between humans and machines

Improve automation and support intelligent human-like behavior and activities that may involve conversations or question-answering and robots.



Expressiveness Range:

# Knowledge Representation and Ontologies



# Forms of KR pursued

- Simple Facts/Assertions
- Taxonomic
- Schema (description)/instances (description base)
- Logical/multiclass - DL, FOL, higher order
- Graphs - Relationships (non-directional/directional, unlabeled/labeled, and constraints)
- Probabilistic Graphs

What about statistical (i.e., implicit knowledge in statistical AI)?





# Semantic Web – Intelligent Content

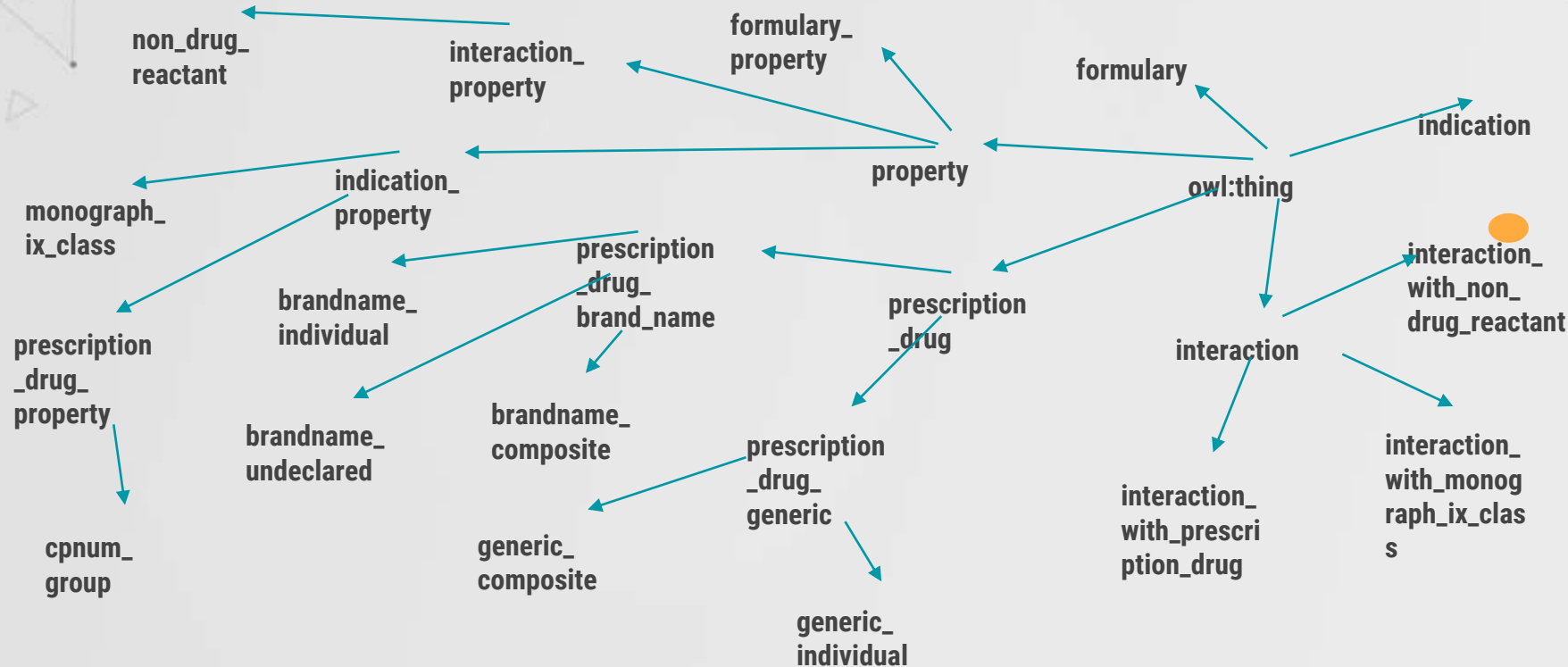
(supported by Taalee Semantic Engine)



Figure: Semantic Web - Semagix

# Drug Ontology Hierarchy

(showing is-a relationships)



# Multimodal KG

SEMAGIX  
POWER • THROUGH • RELEVANCE

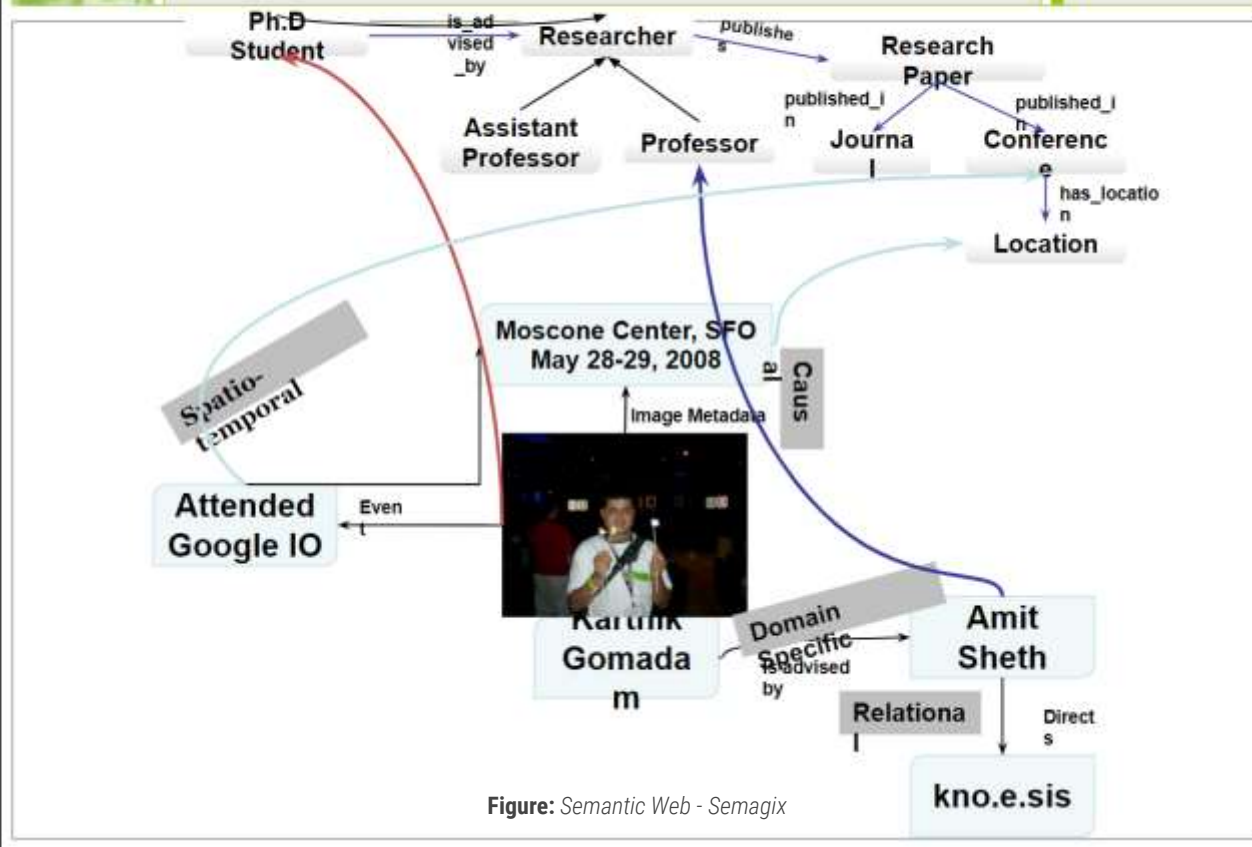


Figure: Semantic Web - Semagix

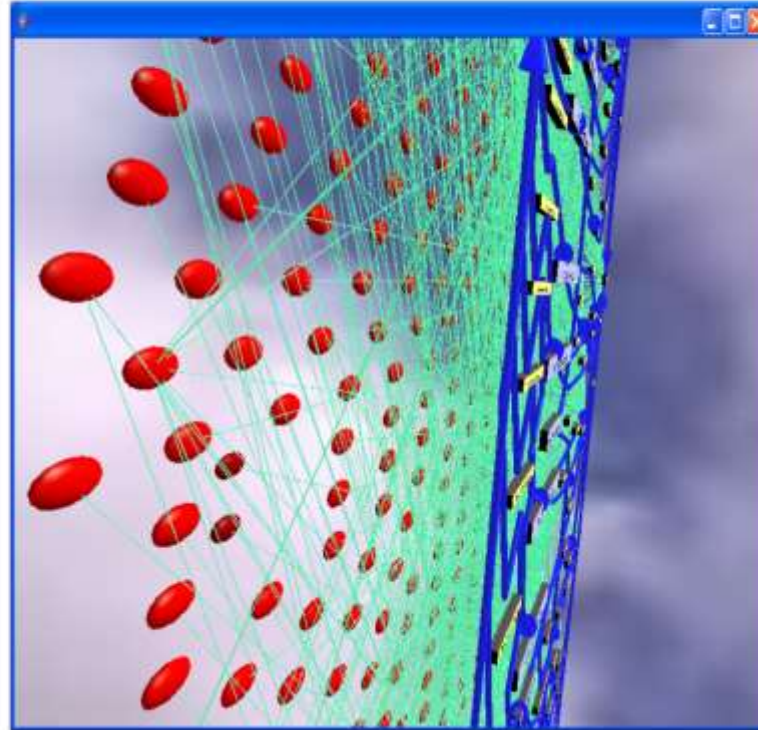
## Space Partitioning

### Foreground

- ◆ visualization of entities and their properties in the foreground.

### Background

- ◆ documents are visualized in the background.



**Figure:** *Space Partitioning*

## Semantic Analytics Visualization representation of entities and relationships

### Entities

- ◆ blue rectangles

### Relationships

- ◆ arrows between entities - a yellow rectangle above the arrow is the property's label

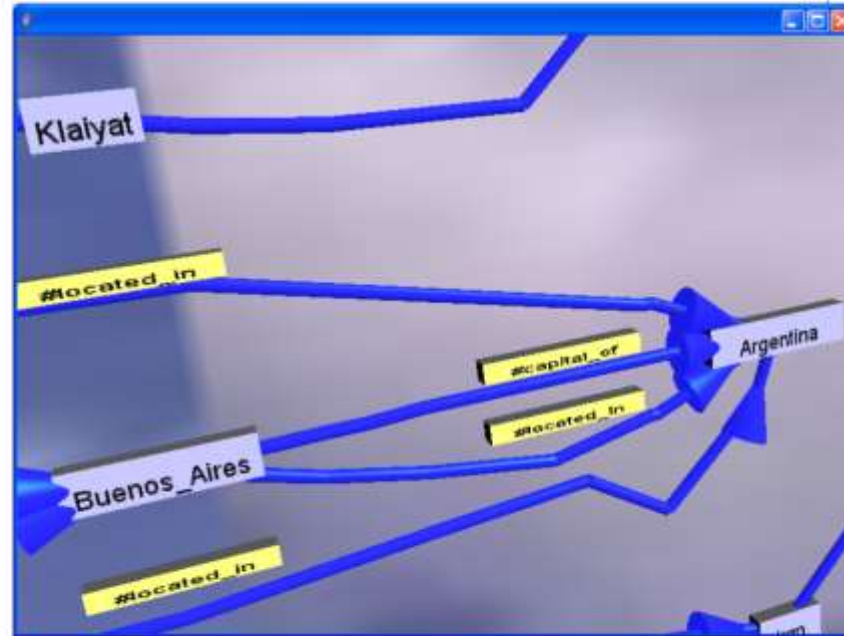
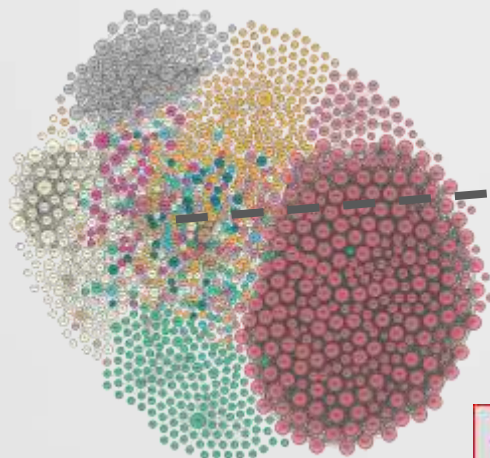


Figure: Semantic Analytics & Visualization





# Knowledge Graphs have become prominent recently



Linked Open Data >  
9960 datasets,  
> 149 B triples



38.3 M entities and 8.8  
B facts



Schema.org annotations

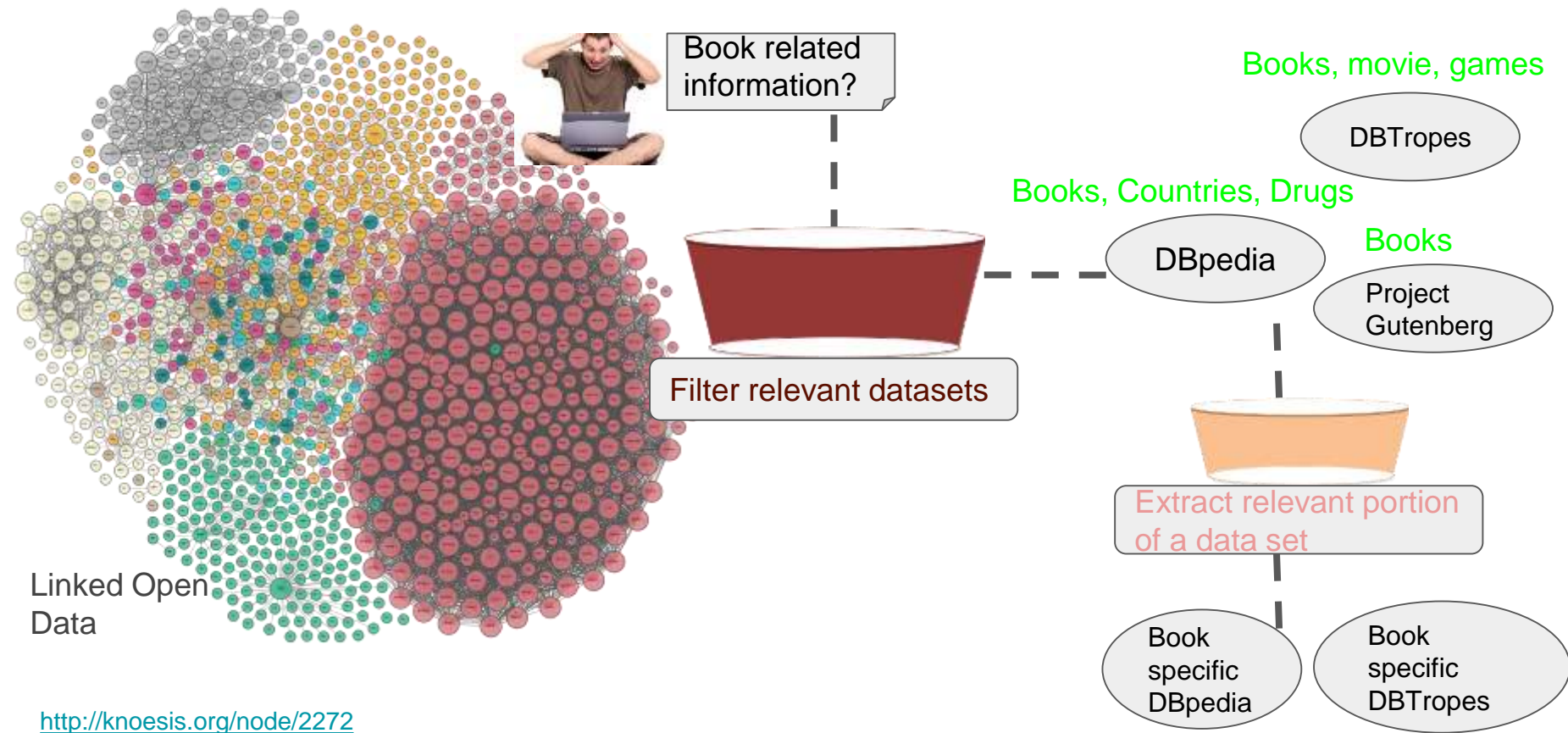


Google Knowledge Graph 570 M  
entities and 18 B facts



LinkedIn Knowledge graph

# Domain-specific knowledge extraction from LOD



<http://knoesis.org/node/2272>

<http://knoesis.org/node/2793>



# It is becoming easier to find or create relevant knowledge for a given application

Existence of large knowledge bases

Ability to search/find a relevant knowledge bases [WI'13]

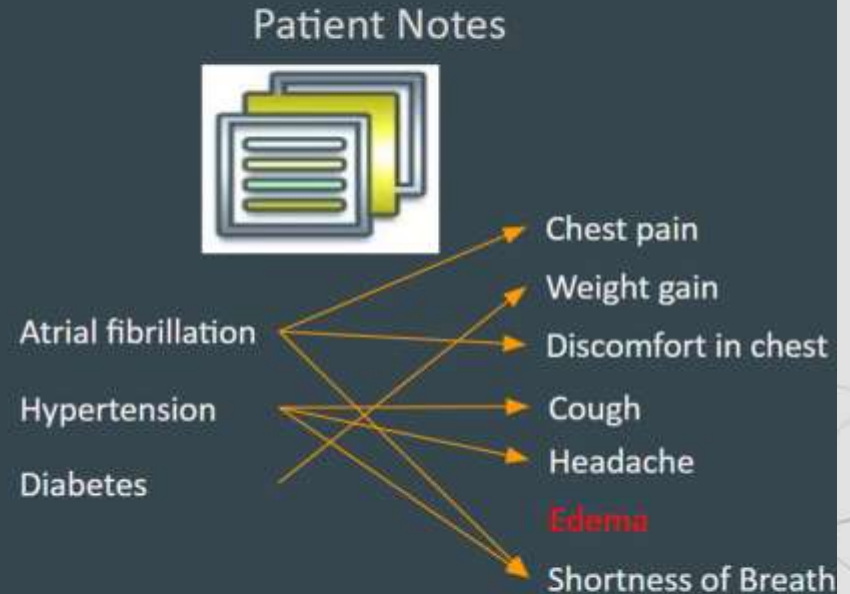
Ability extract a relevant subset [IEEE Big Data'16]

Ability to enrich - by deriving new concepts and new facts [BIBM'12]

**Knowledge graphs** are already playing influential roles in many applications involving big data, starting with search [15 years of search & knowledge graphs].



# Ability to enrich knowledge graphs



Initial knowledge base does not know about edema. Can Edema be a symptom of any of the disorders mentioned according to the patient notes?

## **Semantics for the Semantic Web: The Implicit, the Formal and the Powerful**

Amit Sheth, Cartic Ramakrishnan and Christopher Thomas  
Large Scale Distributed Information Systems lab  
University of Georgia, Athens, GA USA

### ***Abstract:***

Enabling applications that exploit heterogeneous data in the Semantic Web will require us to harness a broad variety of semantics. Considering the role of semantics in a number of research areas in computer science, we organize semantics in three forms: implicit, formal and powerful, and explore their roles in enabling some of the key capabilities related to the Semantic Web. The central message of this paper is that building the Semantic Web purely on description logics will artificially limit its potential, and that we will need to both exploit well known techniques that support implicit semantics, and develop more powerful semantic techniques.

**Keywords:** Semantic Web, Semantic Technology, Formal Semantics, Informal Semantics, Implicit Semantics, Analytical Processing, Document Management and Retrieval, Knowledge Discovery, Soft Computing, Metadata, Semantic Search, Relationship Discovery, Semantic Analytics, Semantic Matching, Semantic Integration

# What propels popularity of KG and its development?

## **KG enabled Web and Enterprise Applications:**

*Microsoft, Siemens, LinkedIn, Airbnb, eBay, and Apple, as well as smaller companies (e.g. ezDI, Fraanz, Metaphactory/Metaphacts GmbH, Semantic Web Company GmbH, Mondeca, Stardog, Diffbot, Siren)*

**Enhanced Learning:** *"Data alone is not enough (Pedro Domingos)"* and the use of domain knowledge improves the results or effectiveness of state of the art ML and NLP techniques.

**Linked Open Data  
(LOD)**

**Schema.org  
([schema.org](https://schema.org))**

**Data Commons  
Knowledge Graph  
(DCKG)  
([datacommons.org](https://datacommons.org))**

**Wikidata  
([wikidata.org](https://wikidata.org))**

# Characteristics of Enterprise KG

|                  | <b>Data model</b>   | <b>Size of the graph</b>  | <b>Development stage</b>                   |
|------------------|---|---|--|
| <b>Microsoft</b> | The types of entities, relations, and attributes in the graph are defined in an ontology.   | ~2 billion primary entities, ~55 billion facts  | Actively used in products                  |
| <b>Google</b>    | Strongly typed entities, relations with domain and range inference  | 1 billion entities, 70 billion assertions   | Actively used in products                  |
| <b>Facebook</b>  | All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal. | ~50 million primary entities, ~500 million assertions   | Actively used in products                  |
| <b>eBay</b>      | Entities and relation, well-structured and strongly typed   | Expect around 100 million products, >1 billion triples  | Early stages of development and deployment |
| <b>IBM</b>       | Entities and relations with evidence information associated with them.  | Various sizes. Proven on scales documents >100 million, relationships >5 billion, entities >100 million | Actively used in products and by clients   |

# PAST

Knowledge Graph for Semantic Applications

---





## Early Use of Knowledge Graphs,

before the start of this century, related to building a knowledge graph **manually** or **semi-automatically** and applying them for *semantic applications*, such as search, browsing, personalization, and advertisement.

*Examples: Taalee, Semagix Semantic Search (2000), InfoBox*

<http://bit.ly/15yrSemS>





## Creation & Use of Knowledge ~2000

### (12) **United States Patent** **Sheth et al.**

---

(54) **SYSTEM AND METHOD FOR CREATING A SEMANTIC WEB AND ITS APPLICATIONS IN BROWSING, SEARCHING, PROFILING, PERSONALIZATION AND ADVERTISING**

(75) Inventors: **Amit Sheth**; **David Avant**, both of Bogart; **Clemens Bertram**, Athens, all of GA (US)







# SEMAGIX

Power Through Relevance

Additional reading: **15 years** of Semantic Search and Ontology-enabled Semantic Applications (<https://www.linkedin.com/pulse/15-years-semantic-search-ontology-enabled-amit-sheth/>)

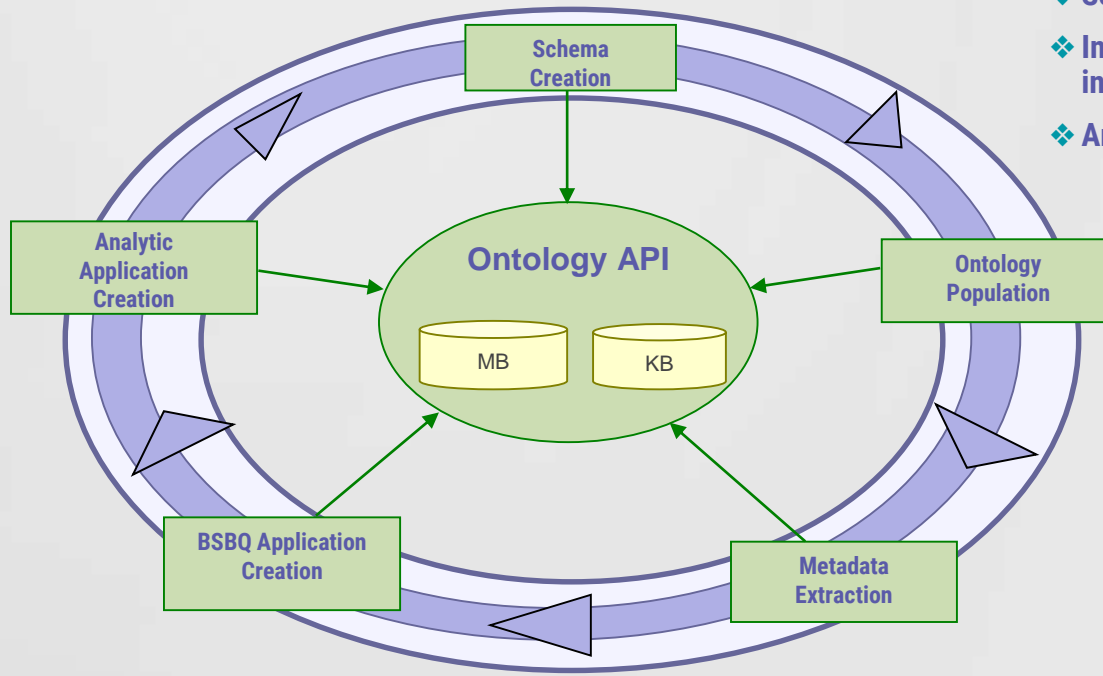
---

# Semantic (Web) Technology

## *State of the Art*

Building a scalable and high performance system with support for:  
Ontology creation and maintenance  
Ontology-driven Semantic Metadata Extraction/Annotation  
Utilizing semantic metadata and ontology

- ❖ Semantic search/querying/browsing
- ❖ Information and application integration - normalization
- ❖ Analysis/Mining/Discovery – relationships

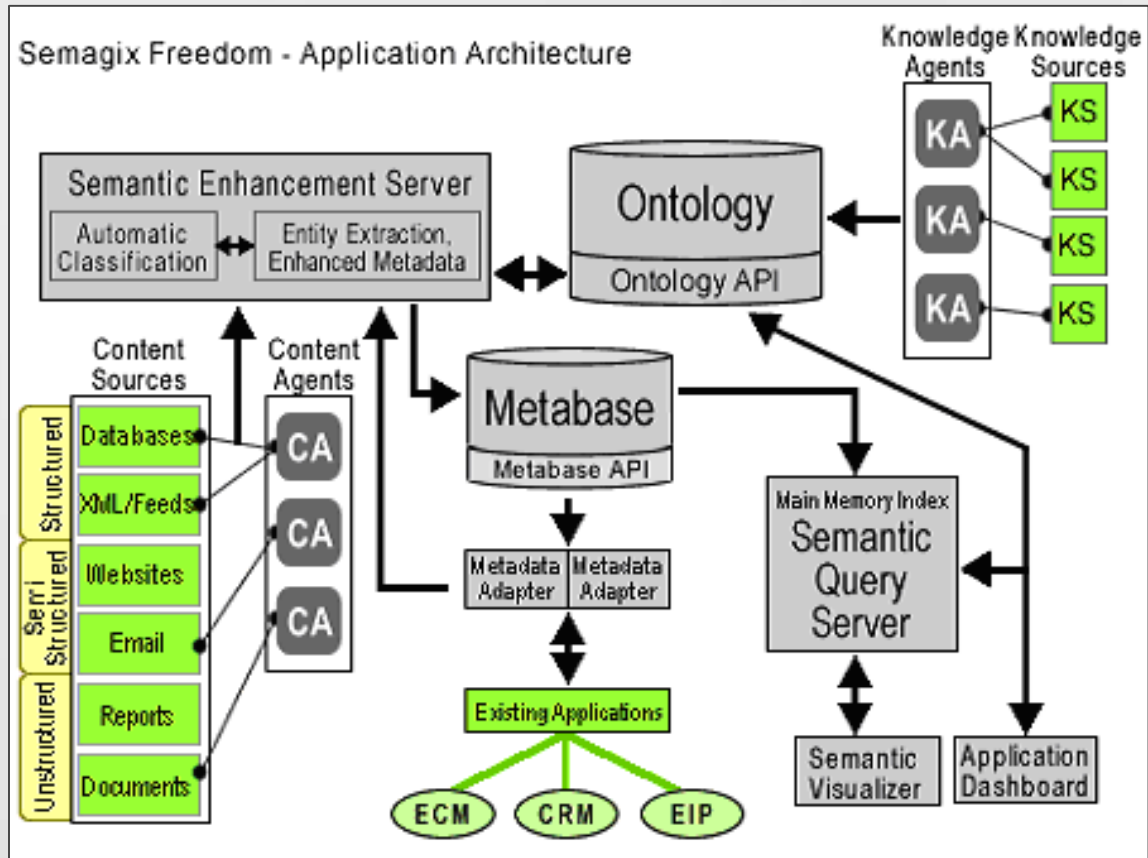


# Semagix Freedom Architecture

for building ontology-driven information system

\*

27



Next Generation:

# Semantic Content Management

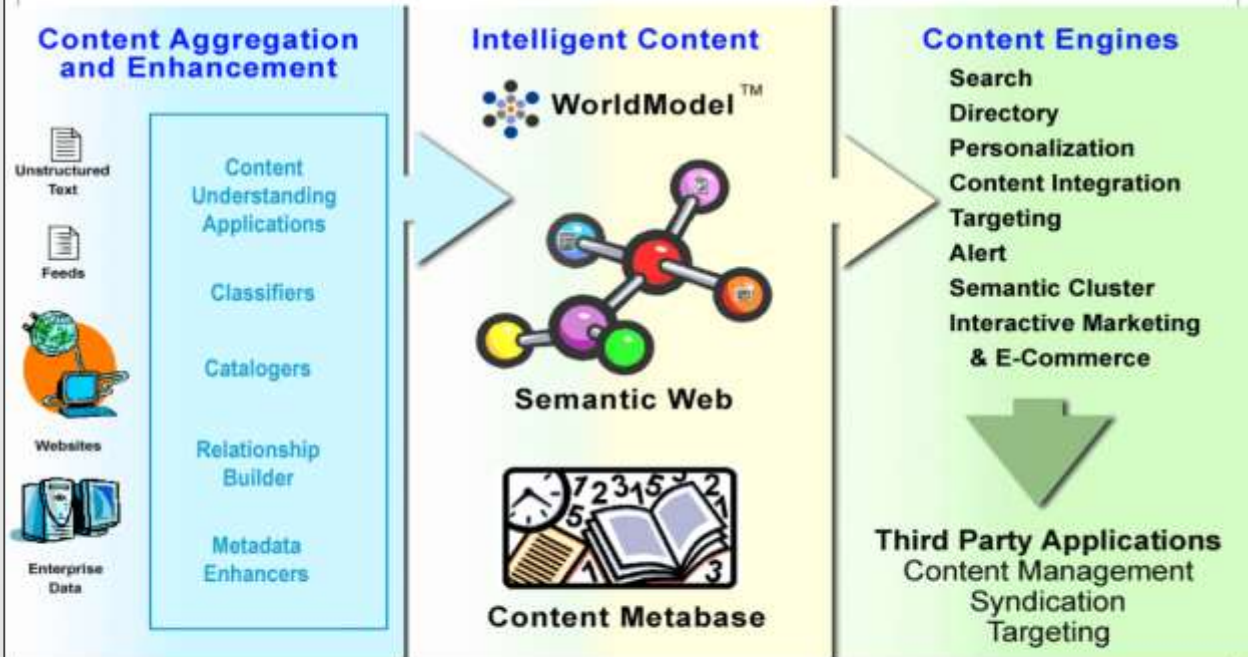


Figure: Semantic Content Management - Semagix

# Semantic Annotation/ Metadata Extraction + Enhancement



Blue-chip bonanza continues

Dow above 9,000 as **HP**, **Home Depot** lead advance; **Microsoft** upgrade helps techs.

August 22, 2002: 11:44 AM EDT

By Alexandra Twin, CNN/Money Staff Writer

**New York** (CNN/Money) - An upgrade of software leader **Microsoft** and strength in blue chips including **Hewlett-Packard** and **Home Depot** were among the factors pushing stocks higher at midday Thursday, with the **Dow Jones industrial average** spending time above the 9,000 level.

Around 11:40 a.m. ET, the **Dow Jones industrial average** gained 65.06 to 9,022.09, continuing a more than 1,300-point resurgence since July 23. The **Nasdaq** composite gained 9.12 to 1,418.37.

**The Standard & Poor's 500 index** rose 9.61 to 958.97.


**Hewlett-Packard** ( **HPQ**: up \$0.33 to \$15.03, Research, Estimates) said a report shows its share of the printer market grew in the second quarter, although another report showed that its share of the computer server market declined in **Europe**, the **Middle East** and **Africa**.

**Home Depot** ( **HD**: up \$1.07 to \$33.75, Research, Estimates) was up for the third straight day after topping fiscal second-quarter earnings estimates on Tuesday.

Tech stocks managed a turnaround. **Software** continued to rise after **Salomon Smith Barney** upgraded No. 1 software maker **Microsoft** ( **MSFT**: up \$0.55 to \$52.83, Research, Estimates) to "outperform" from "neutral" and raised its price target to \$59 from \$56. Business software makers **Oracle** ( **ORCL**: up \$0.18 to \$10.94, Research, Estimates), **PeopleSoft** ( **PSFT**: up \$1.17 to \$20.67, Research, Estimates) and **BEA Systems** ( **BEAS**: up \$0.28 to \$7.12, Research, Estimates) all rose in tandem.

competes with

# Active Semantic Doc with 3 Ontologies

**Athens Heart Center**  
**Jerek Chicken**  
325 PileORice Street, Athens, GA 30606  
SSN: 123-45-6789 MR #: 555555 Sex: M DOB: 01/02/1934 Age: 71

2005 Prince Avenue, Athens, GA, 30606 Phone: 706-208-9700 Fax: 706-208-0806

Visit on 10/28/2005

Other Physicians: Harry Wingate, M.D. E Kevin Adams, M.D. E  
Family Practice  
706-795-9588

Annotate ICD9s

Annotate Doctors

Lexical Annotation

**Problem List:**  
1. Hypertension (365.04) E  
2. Cholecystectomy (576.0) E  
3. Chest Pain E

**Chief Complaint:** Evaluation of abnormal EKG status post abnormal Echo Evaluation of aortic stenosis status post arterial examination Cardiac clearance for aneurysm removal Follow up of recent hospitalization at Barrow Community Hospital for acute myocardial infarction.

**History of Present Illness:** He was evaluated at Athens Regional Medical Center emergency room by Dr. Harry Wingate. He is here today for cardiac clearance for aneurysm removal. The patient reports chronic moderate burning and cramping chest pain located across the chest, which radiates to the arms. He reports that his chest pain is aggravated by movement. The patient is breathing deeply. Patient's history is positive for the following cardiovascular risk factors: diabetes and family history of cor

**Current Medications**  
Actos 30 mg, 1tab E

**Level 3 Drug Interaction**

**Medications Affected**  
Coumadin tablets 10 mg, 1tab 13 F E  
Viagra 50 mg, 1tab 13 F E  
Zyrtec 5 mg, 1tab E  
Zyvox 2 mg/ml, 1inj A E

**Insurance Formulary**

**Allergies:** LINEZOLID  
**Impressions:**  
1. Abdominal aortic aneurysm, advanced secondary to by a positive nuclear scan.  
2. Abnormal cardiac study associated with chest tightness appears to be secondary to a noncardiac cause as evidenced by arterial scan of lower extremities.  
3. Normal cardiac cath.

**Drug Allergy**



# Voquette/Taalee's Categorization & Automatic Metadata Creation

**SEMAGIX**  
POWER • THOROUGH • RELEVANCE

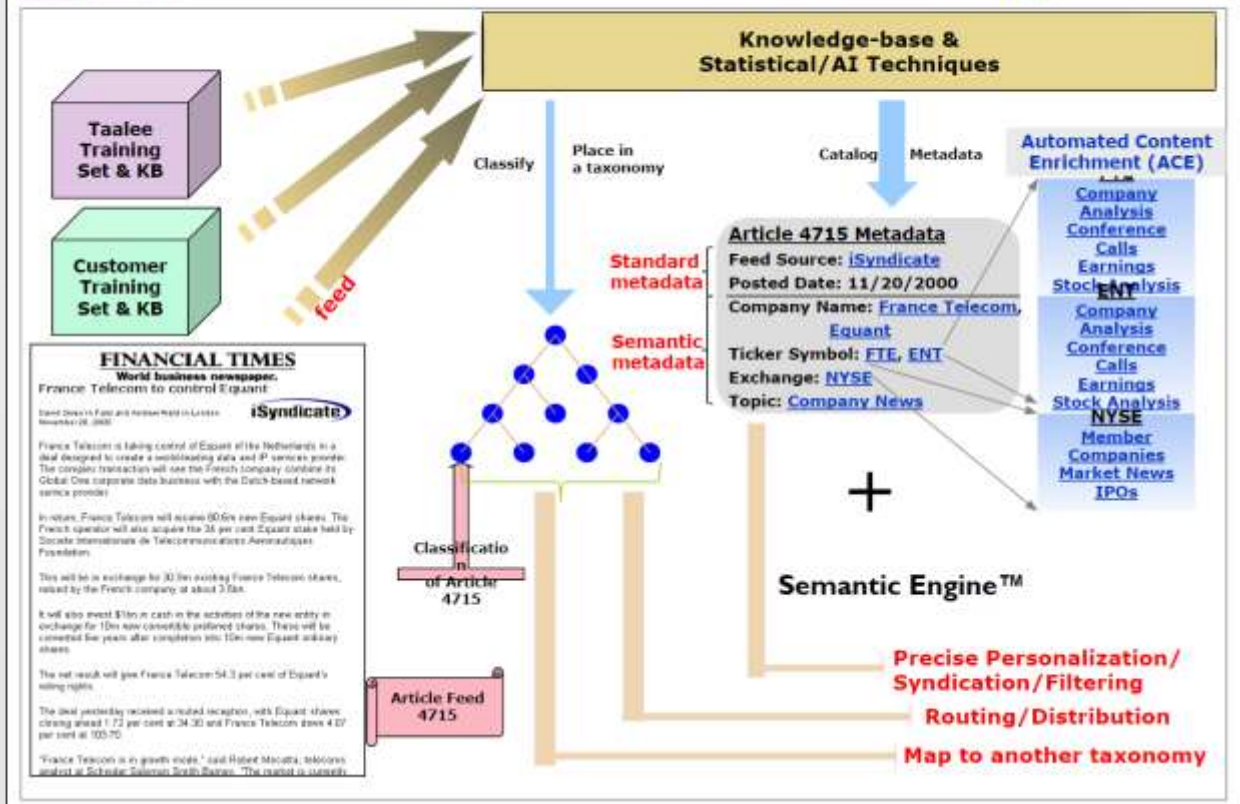


Figure: Semantic Enhancement Engine - Taalee

# Semantic Search

Simply the most precise and freshest A/V search

Choose a category to search in: Music

Enter search term: Song

One field is usually enough.  
Entering more yields more precise,  
but fewer, results.

Artist: Madonna

Album/CD:

Other Search Terms:

Find me: ☒ Video ☒ Audio

**Find It**

Context and Domain Specific Attributes

Search results in **Music** for *madonna*  
16 results found

**Exact match for your search...**

- 1. Material Girl (Madonna)**  
ENGLISH SONGS FOR YOU ABCD'S  
Source: Needt4u  
Category: Music  
Posted: 4/14/2009  
Artist: Madonna
- 2. Holiday**  
Madonna Louise Veronica Ciccone, 16 August 1958, Ft. Michigan, USA. Madonna excelled at dance and drama at high school and during brief periods at colleges in Michigan and North Carolina.  
Source: World Entertainment Network  
Category: Music  
Posted: 5/18/2009  
Artist: Madonna
- 3. "Shower"**  
Heard in background when Dawson is looking for Eve at R...  
Source: <http://www.dawsonsworld.com>  
Category: Music  
Posted: 4/14/2009  
Artist: Madonna
- 4. Madonna - Fata**  
Source: MTV  
Category: Music  
Posted: 4/14/2009  
Artist: Madonna
- 5. The Power Of Good Bye**  
Music video by madonna  
Source: MTV  
Category: Music  
Posted: 4/14/2009  
Artist: Madonna
- 6. Nothing Really Matters**  
Music video by madonna  
Source: MTV  
Category: Music  
Posted: 1/20/2009  
Artist: madonna

Uniform Metadata for Content from Multiple Sources, Can be sorted by any field

**Rich Media Reference**

**Music** RESULTS SEARCH

**Holiday**

[Click to play](#)

[REAL](#) [YouTube](#)

<http://shop.mad.com>

Madonna Louise Veronica Ciccone, 16 August 1958, Rochester, Michigan, USA. Madonna excelled at dance and drama at high school and during brief periods at colleges in Michigan and North Carolina.

Produced by: World Entertainment Network  
Posted Date: 5/18/2009  
Artist: Madonna  
Album: Introspecta Collection  
Song Name: Holiday  
Genre: Music

| Madonna Specials!        | PRICE   | Discography  |
|--------------------------|---------|--------------|
| American Pie DVD         | \$32.99 | Controversy  |
| Immaculate Conception CD | \$13.99 | Videography  |
| Madonna Video Collection | \$17.99 | Bibliography |

[Email this to a friend](#)

Delightful, relevant information,  
exceptional targeting opportunity

Figure: Semantic Search - Semagix



## Metadata

What else can a context do?  
(a commercial perspective)

The screenshot shows an ESPN article titled "Clemens injured" with a "REAL" audio player and a photo of Roger Clemens. A red arrow points from the "ONE-CLICK MEDIA PLAY" annotation to the audio player. Below the article text, a metadata box lists details like "Produced by: ESPN" and "Posted Date: 6/14/2000". A red arrow points from the "City and Team not mentioned in story" annotation to the "Teams: New York Yankees" entry. At the bottom, a yellow box contains links for "Clemens Stats", "Yankees Stats", "Yankees Schedule", and "Buy Yankee Tickets", with a red arrow pointing from the "Category specific sponsorship" annotation to this box.

**Annotations:**

- ONE-CLICK MEDIA PLAY
- Click to play audio
- City and Team not mentioned in story. Taalee Knowledge Experts added these. Other Searchers for 'Yankees' would not find this story.
- Category specific sponsorship dynamically added by Taalee.

Semantic  
Enrichment

Figure: Semantic Metadata - Semagix

Google iPhone 6


Web News Images Videos Shopping More Search tools

About 1,460,000/200 results (0.24 seconds)

**iPhone 6 at Verizon - verizonwireless.com**  
www.verizonwireless.com  
3.5 ★★★★★ - rating for verizonwireless.com  
America's largest & most reliable 4G LTE network for your iPhone  
9 901 S Coast Dr, Ste 120, Costa Mesa, CA - (714) 427-0733  
Verizon Plans iPhone 6  
1GB Bonus Data Promotion iPhone 6 Plus


**iPhone 6: First Impressions, Pre-Order on Sept 12th**  
www.macrumors.com/roundup/iphone-6/ - MacRumors  
Apple has unveiled two new iPhones, the 4.7-inch iPhone 6 and the 5.5-inch iPhone 6 Plus. Along with larger screens and a completely new iPad-style design ...  
Apple Reportedly Launching ... - Apple Likely to Slim Down ... - Transition to '3x'

**News for iPhone 6**

 **T-Mobile offers iPhone 6, 6 Plus on monthly payment plan**  
CNET - 3 hours ago  
T-Mobile announced a pricing program for Apple's new iPhone 6 and iPhone 6 Plus on Thursday that lets users pay off their phones with a ...

**Apple iPhone 6**

The iPhone 6 is a new smartphone by Apple. It was announced on September 9 along with the larger-sized iPhone 6 Plus. It will be available for pre-order on September 12 and released for sale on September 18 - Apple



Price: From \$199  
Screen Size: 4.7 inch  
Screen Resolution: 1334-by-750-pixel resolution at 326 ppi  
Colors: Silver, Gold, Space Gray  
Memory: 16 GB, 64 GB, 128 GB  
Camera Resolution: 8 megapixel  
Weight: 4.55 ounces  
Carriers: AT&T, Sprint, T-Mobile, Verizon

Figure: Google InfoBox

## Example (test on <http://directory.mediaanywhere.com>)

**SEMAGIX**  
POWER • THOROUGH • RELEVANCE

The screenshot displays the Semantic Directory interface within a Microsoft Internet Explorer browser. The main search results page for 'Commerce One, Inc.' is shown, with several annotations in pink callout boxes:

- Links to news on companies that compete against Commerce One**: Points to the 'Related to Commerce One, Inc.' section, which lists companies like Arriba, Inc., and Oracle Corp.
- Crucial news on Commerce One's competitors (Arriba) can be accessed easily and automatically**: Points to the 'Arriba, Inc. vs Company' section, which includes a news item about Arriba's stock being downgraded.

The interface includes a sidebar with navigation links (Sports, Politics, Science, etc.), a search bar, and a list of 'Other Frequently mentioned Companies'.

**Related to Commerce One, Inc.**

the following company/ies:  
[Hoffman Media](#)

the following company/ies competes:  
[Arriba, Inc.](#)  
[Cable One](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)  
[Cable One, Inc.](#)

the Commerce One, Inc., members:  
Corporate, Professor

the Commerce One, Inc., identified:  
[Cable One](#)

the Commerce One, Inc., compete:  
[Arriba, Inc.](#)  
[Oracle Corp.](#)

**Other Frequently mentioned Companies**

1. [Arriba, Inc.](#)
2. [Cable One, Inc.](#)
3. [Cable One, Inc.](#)
4. [Cable One, Inc.](#)
5. [Cable One, Inc.](#)
6. [Cable One, Inc.](#)
7. [Cable One, Inc.](#)
8. [Cable One, Inc.](#)
9. [Cable One, Inc.](#)
10. [Cable One, Inc.](#)

**Arriba, Inc. vs Company**

See [Related to Arriba, Inc.](#)

**Arriba, Inc. vs Company**

Pages: 1 2 3 4 5 6 7 8 9 10 Next [Arriba, Inc. vs Company](#) of 14

1. [Arriba, Inc.](#)  
Off The Move: B2B Marketplace Stocks: Slump on Downgrade  
Source: USA  
Category: Business  
Posted: 10/10/01  
Company Name: Arriba, Inc.
2. [Arriba, Inc. vs Company](#)  
Thomas Weisel Partners Analyst David Gremmels downgrades Arriba, (2 and Rebel from Strong Buy to Buy and keeps his Market Perform.
3. [Arriba, Inc. vs Company](#)  
Thomas Weisel Partners Analyst David Gremmels downgrades Arriba, (2 and Rebel from Strong Buy to Buy and keeps his Market Perform.

**Figure:** Semantic Directory - Semagix

# Knowledge-based and Manual Associations

The image shows two overlapping screenshots of the BBC News website from August 7, 1998. The left window displays the 'WORLD MEDIAWATCH' section with the headline 'Islamic Jihad vows 'revenge''. The right window shows a detailed article titled 'Egypt sentences militants to death'. Blue arrows and red boxes highlight specific text elements, which are then linked to green callout boxes containing manual annotations.

**Manual Annotations:**

- Same entity:** Points to the text 'It is not known who was responsible for the bombings in Tanzania' in the left window.
- led by:** Points to the text 'led by Dr Ayman al-Zawahiri' in the left window.
- Human-assisted inference:** A green box containing this text, with arrows pointing to multiple locations: the article title 'Egypt sentences militants to death', the photo of Osama Bin Laden, the text 'They include the Jihad leader, Ayman al-Zawahiri', and the text 'They included several militants extradited from other countries'.
- Other associations:** Arrows link the text 'Ahmed Ibrahim al Naggar, one of 12 handed over to Egypt from Albania last summer' to the text 'Some of the men have known links with Osama Bin Laden' and 'They included several militants extradited from other countries'. Another arrow links 'death in an Egyptian court' to 'Ahmed Ibrahim al Naggar'.

**Text from the left window (BBC NEWS):**

**WORLD MEDIAWATCH**  
In association with BBC Monitoring  
Friday, August 7, 1998 Published at 22:21 GMT 23:21 UK

**Islamic Jihad vows 'revenge'**

It is not known who was responsible for the bombings in Tanzania.

However, following the bombings, the Islamic Jihad movement warned of "revenge" against the US for its involvement in the extradition to Egypt of several militants living in eastern Europe.

Here are excerpts from a report by based Arabic newspaper Al-Hayat on August 6:

Cairo: The "Islamic Jihad Group" led by Dr Ayman al-Zawahiri - who is currently residing in Taliban-controlled areas in Afghanistan - has vowed to take revenge against the United States, which it is accusing of involvement in the arrest of a number of Egyptian extremists - among them the leading figure in the organization, Ahmad Ibrahim al-Naggar, who was sentenced in absentia to death by the Egyptian Military Court in Cairo in connection with the Khan al-Khalili case - while they were in the Albanian capital Tirana, and in the extradition to Egypt.

**Text from the right window (BBC News):**

Middle East | Egypt sentences militants to death - Micro...

Most of the defendants were said to belong to the armed Islamic movement, Al Jihad, and some of them are known to be closely associated with Osama Bin Laden, the Saudi dissident accused by the United States of masterminding attacks on American embassies in east Africa last year.

**Human-assisted inference**

Al-Jihad, based outside Egypt. They include the Jihad leader, Ayman al-Zawahiri, who is thought to be in Afghanistan and his brother, Mohammed.

Eleven others, convicted of conspiring to overthrow the government, received life sentences with hard labour.

Some of the men have known links with Osama Bin Laden.

They included several militants extradited from other countries.

Ahmed Ibrahim al Naggar, one of 12 handed over to Egypt from Albania last summer, wore the red clothing of a condemned man because he had been sentenced to death in an Egyptian court for plotting an attack on Cairo's Khan al-Khalili market - a major tourist attraction.

Figure: Knowledge-based and Manual Associations



# Blended Semantic Browsing and Querying SEMAGIX

(Intelligence Analyst Workbench)

POWER - THROUGH - RESISTANCE

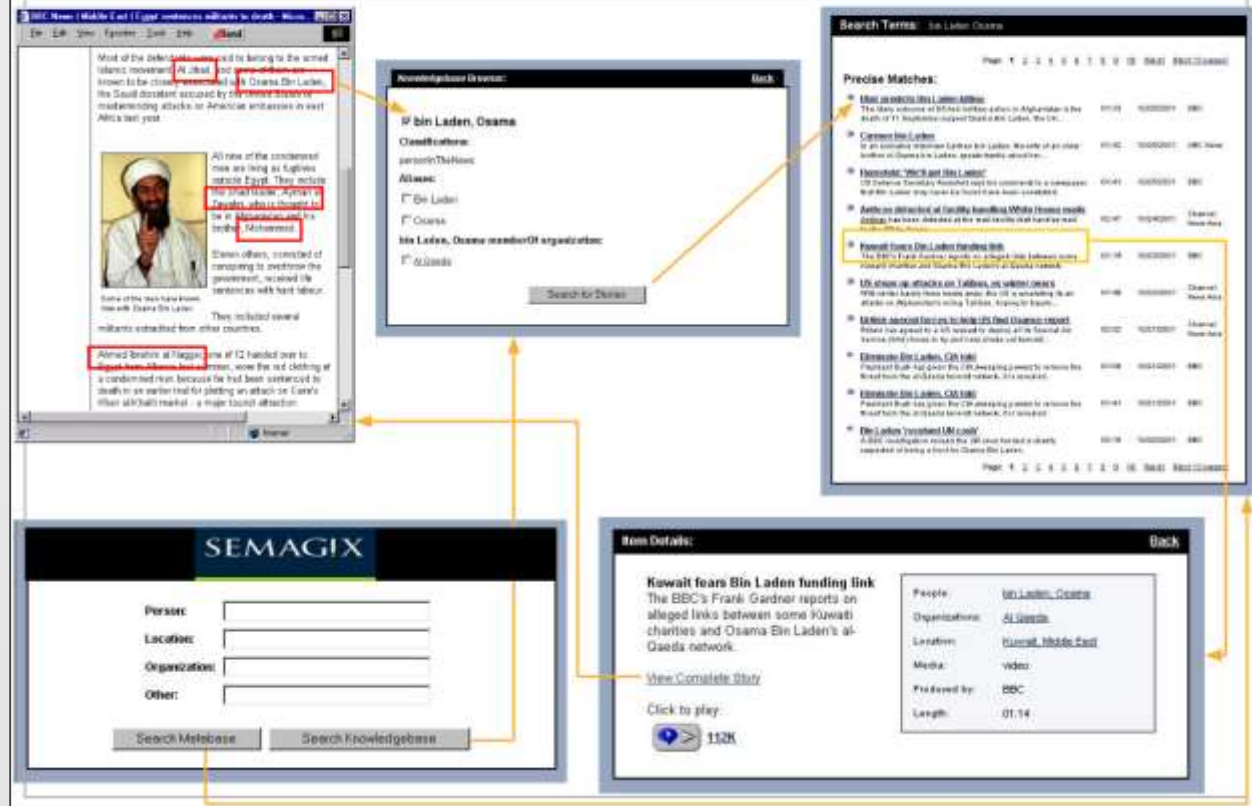


Figure: Blended Semantic Browsing and Querying

## Semantic EventTracker representation of geospatial and temporal dimensions for semantic associations



- Visualization of association unfolding over time
- Integration of associated multimedia content
- Separate **Temporal**, **Geospatial**, and **Thematic** ontologies describe data
- [DEMO](#)

**Figure:** *Semantic EventTracker*

An abstract geometric pattern consisting of a network of interconnected nodes and lines, forming a complex, web-like structure. The nodes are represented by small black dots, and the lines are thin, light gray. The pattern is dense and irregular, with many overlapping triangles and polygons. It occupies the right side of the slide, extending from the top to the bottom.

# Semantics for the Semantic Web

The Implicit, the Formal and the Powerful

---

# Knowledge will Propel Machine Understanding of Content: Extrapolating from Current Examples

Amit Sheth

Kno.e.sis Center, Wright State University  
Dayton, Ohio, USA  
amit@knoesis.org

Sanjaya Wijeratne

Kno.e.sis Center, Wright State University  
Dayton, Ohio, USA  
sanjaya@knoesis.org

Sujan Perera

Kno.e.sis Center, Wright State University  
Dayton, Ohio, USA  
sujan@knoesis.org

Krishnaprasad Thirunarayan

Kno.e.sis Center, Wright State University  
Dayton, Ohio, USA  
tkprasad@knoesis.org

## ABSTRACT

Machine Learning has been a big success story during the AI resurgence. One particular stand out success relates to learning from a massive amount of data. In spite of early assertions of the unreasonable effectiveness of data, there is increasing recognition for utilizing knowledge whenever it is available or can be created purposefully. In this paper, we discuss the indispensable role of knowledge for deeper understanding of content where (i) large amounts of training data are unavailable, (ii) the objects to be recognized are complex, (e.g., implicit entities and highly subjective content), and (iii) applications need to use complementary or related data in

DOI: 10.1145/3106426.3109448

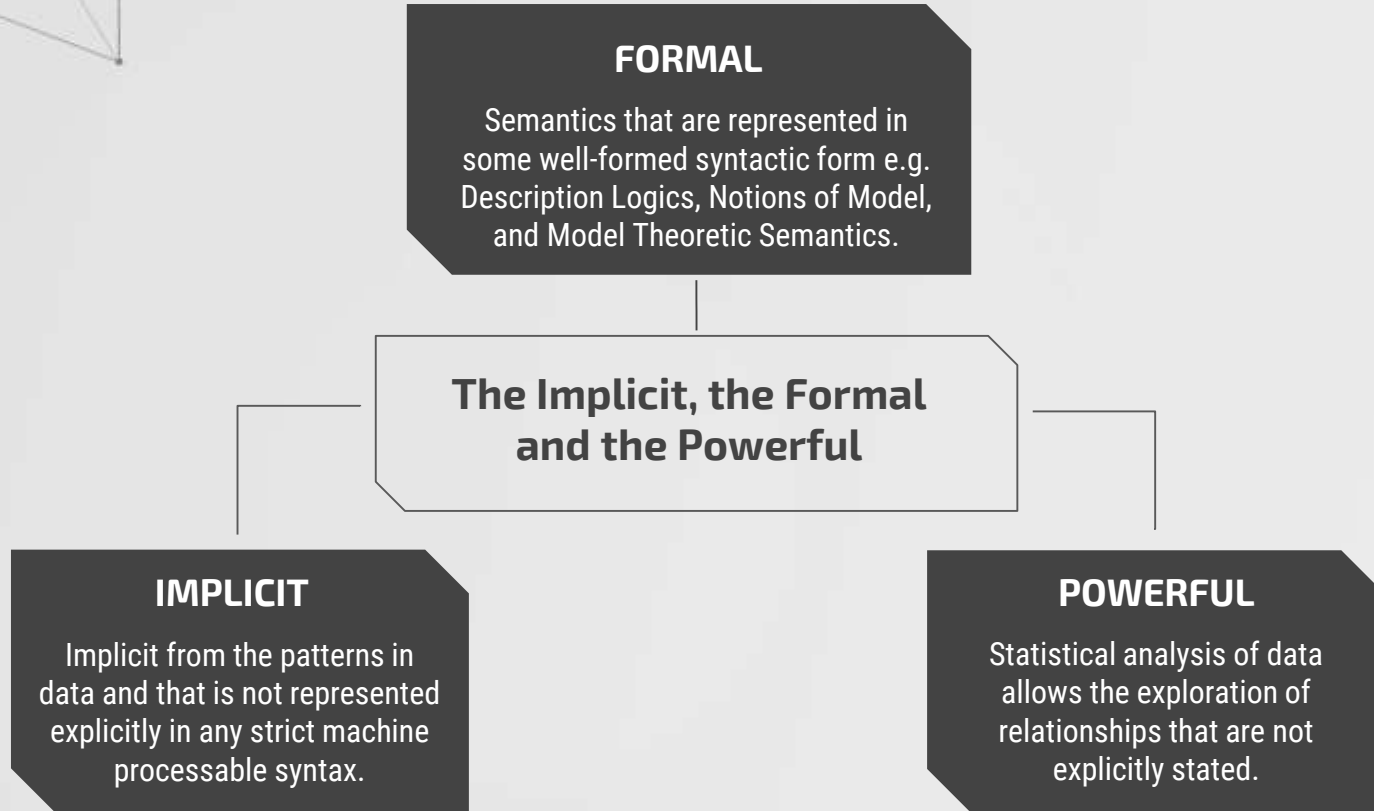
## 1 INTRODUCTION

Recent success in the area of Machine Learning (ML) for Natural Language Processing (NLP) has been largely credited to the availability of enormous training datasets and computing power to train complex computational models [12]. Complex NLP tasks such as statistical machine translation and speech recognition have

<http://knoesis.org/node/2835>



# Semantics for the Semantic Web



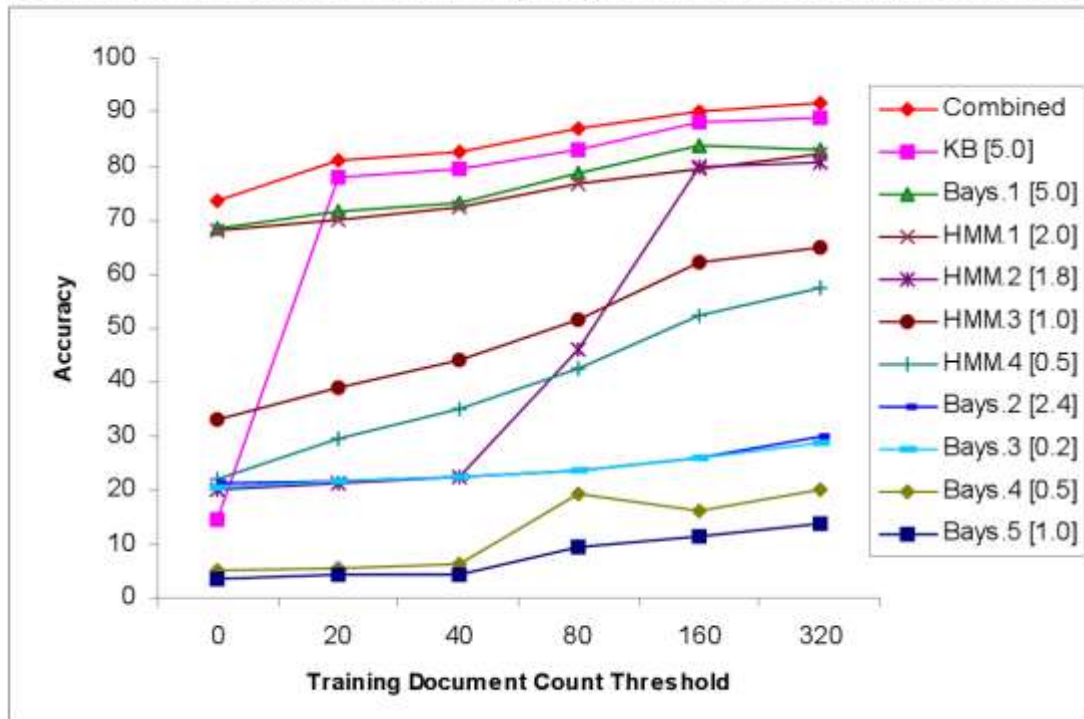
Sheth, A., Ramakrishnan, C., & Thomas, C. (2005). Semantics for the semantic web: The implicit, the formal and the powerful. International Journal on Semantic Web and Information Systems (IJSWIS), 1(1), 1-18.

## Semantic Enhancement Engine:

### *A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content*

(Brian Hammond, Amit P. Sheth, Krzysztof J. Kochut)

The classification committee consistently outperformed the individual classifiers.



Combining statistical and knowledge-based techniques for NER

Accuracy rates for the individual classifiers on the Reuters set with various threshold values on the number of training documents required.

# PAC Learning

*Probably Approximately Correct (PAC) ---- Need for Knowledge Infusion*

$$samples \geq \frac{\log |C| - \log \delta}{\epsilon}$$

Number of data “samples” required to effectively train a classifier(s) depends on

1. Checking all the possible classifiers ( $|C|$ )
2. True error of the data (human annotation/human bias ( $\delta$ ))
3. Generalization error of the model ( $\epsilon$ )

This equation make two critical inference:

1. **Confidence:** More Certainty (lower  $\delta$ ) means more number of samples.
2. **Complexity:** More complicated hypothesis ( $|C|$ ) means more number of samples

**How do you know that a training set has a good domain coverage?**

A KG (or Ontology) schema is designed by domain experts. It is populated from a representative DB (sets of instances). A KG has very large number of instances (mapping to # of training examples).

---

**How do ensure consistency of labeling, esp when label is not binary?**

**Do labels represent adequate semantics (e.g., number of alternatives)?**

**Do they have adequate domain knowledge?**

**How do you ensure consistency of labeling (interpretation)?**

A good KG has addresses these issues:

- a schema is rich in representation (and captures much more than labeling)
- KG design incorporate substantiate domain knowledge
- instance level knowledge is created through (usually) Collective intelligence and —curation



# Knowledge Will Propel Machine Understanding of Big Data

The indispensable role of knowledge for deeper understanding of content

---

# Knowledge Will Propel Machine Understanding of Big Data

The indispensable role of knowledge for deeper understanding of content where

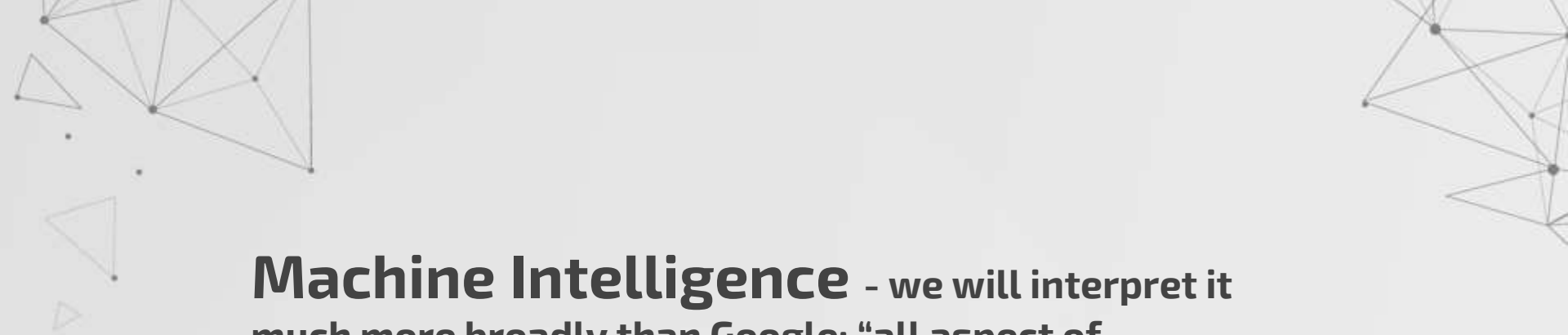
Large amounts of training data are unavailable.

The objects to be recognized are complex  
(e.g., implicit entities and highly subjective content).

Applications need to use complementary or related data in multiple modalities/media.





The top-left and top-right corners of the slide feature decorative geometric patterns. These consist of thin, light-gray lines connecting small black dots to form various triangular and polygonal shapes, creating a network-like or molecular structure.

**Machine Intelligence** - we will interpret it much more broadly than Google: “all aspect of machine learning”... We will define it as machines (any system) performing similar to (nearly emulating) human intelligence.

How will machines “understand” the data/signals/observations, so that it can (help) take timely and good (evidence based) decision and actions.

# Knowledge plays an indispensable role in deeper understanding of content

Knowledge often complements or enhances ML and NLP techniques, supports in contextual “understanding” of data.

Especially interesting situations:

- I. Large amounts of training data are unavailable,
  - II. The objects to be recognized are complex, such as implicit entities and highly subjective content, and
  - III. Applications need to use complementary or related data in multiple modalities/media.
-

# Knowledge-based Approaches and the Resulting Improvements

| Problem Domain  | Use of Knowledge/Knowledge bases   | Problems solved that could not be solved (well) w/o knowledge            |
|---|--|--|
| Implicit Entity recognition and linking                                   | Adapted UMLS definitions for identifying medical entities, and Wikipedia and Twitter data for identifying Twitter entities | Was not solved before  |
| Understanding Drug Abuse-related Discussions on Web forums                | Application of Drug Abuse Ontology along with slang term dictionaries and grammar  | Not solved well at all   |
| Understanding city traffic dynamics using sensor and textual observations | Statistical knowledge extraction and using ontologies for Twitter event extraction   | Multi-modal data stream correlation and explanation virtually impossible |
| Emoji Similarity and Sense Disambiguation                                 | Generation and application of EmojiNet   | Emoji interpretation solved much better                                  |

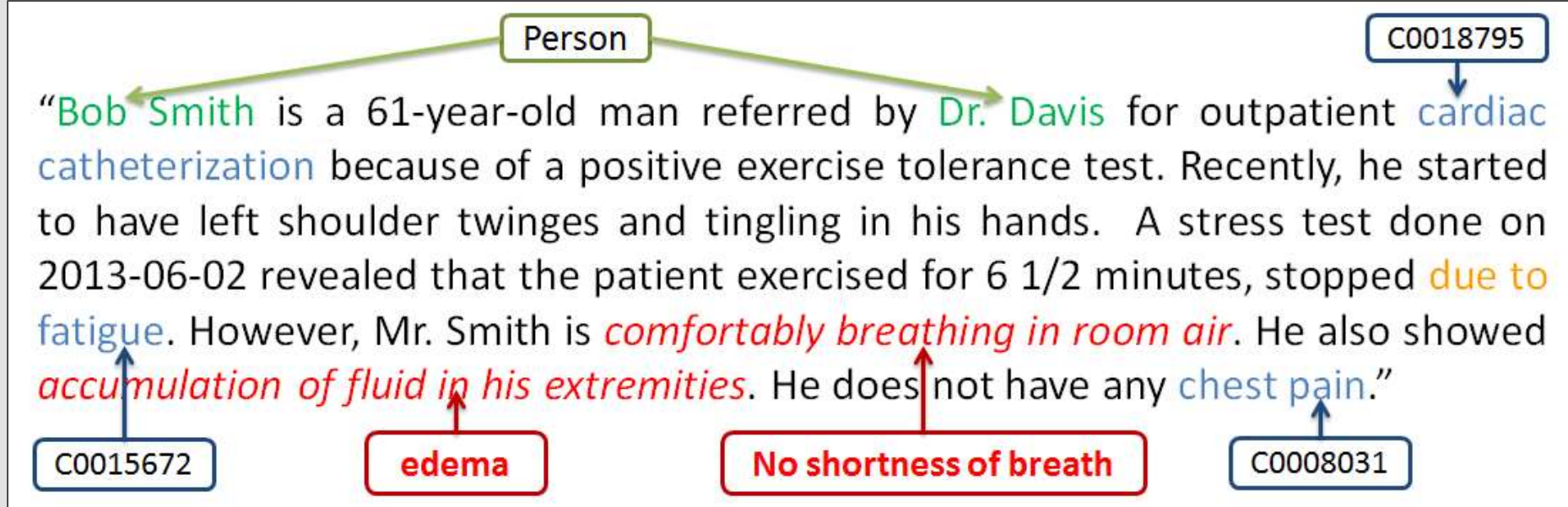


# Implicit Entity Recognition and Linking

Sujan Perera, Pablo N. Mendes, Adarsh Alex, Amit Sheth, Krishnaprasad Thirunarayan.  
Implicit Entity Linking in Tweets. Extended Semantic Web Conference. Heraklion, Crete,  
Greece : Springer; 2016. p. 118-132. <http://knoesis.org/node/2644>

Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex,  
Christopher Heid, Greg Mott. Implicit Entity Recognition in Clinical Documents. 4th Joint  
Conference on Lexical and Computational Semantics (\*SEM) 2015. Denver, CO: Association  
for Computational Linguistics; 2015. p. 228-238. <http://knoesis.org/node/2171>

# Implicit Entity Recognition and Linking



Named Entity Recognition

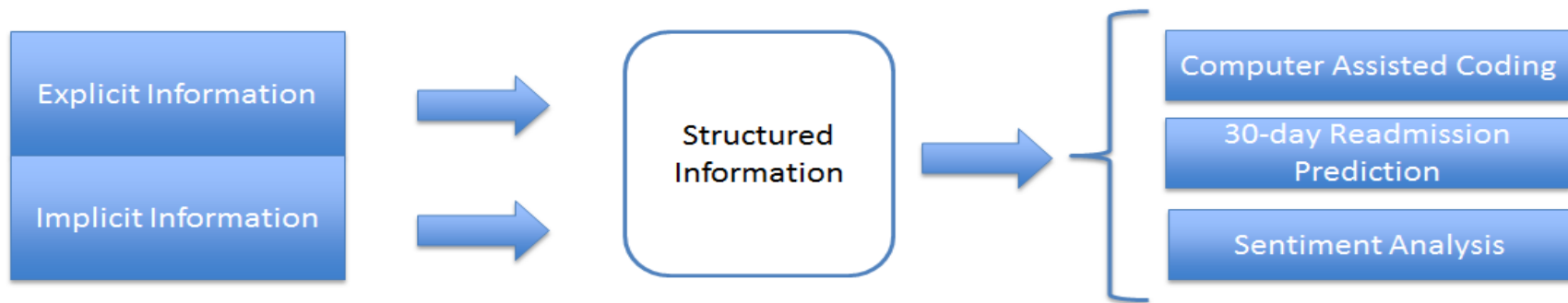
Relationship Extraction

Entity Linking

Implicit information extraction

## Significance

- Volume
  - 20% movie references and 40% book references in tweets
  - 35% edema and 40% shortness of breath references in clinical narratives
- Value



Ignoring implicit information in text would adversely affect downstream applications

## Role of Knowledge



*The patient showed accumulation of fluid in his extremities, but respirations were unlabored and there were no use of accessory muscles.*



## Knowledge

|                     |   |
|---------------------|---|
| Edema               | Accumulation of an excessive amount of watery fluid in cells or intercellular tissues |
| Shortness of breath | Labored or difficult breathing associated with a variety of disorders                 |



*New Sandra Bullock astronaut lost in space movie looks absolutely terrifying*



Sandra Bullock

Gravity

|                         | Entity Type | Without Contextual Knowledge | With Contextual Knowledge |
|-------------------------|-------------|------------------------------|---------------------------|
| Disambiguation Accuracy | Movie       | 51.7%                        | <b>60.97%</b>             |
|                         | Book        | 50.0%                        | <b>61.05%</b>             |





# Understanding and Analyzing Drug Abuse Related Discussions on Web Forums

Cameron, Delroy, Gary A. Smith, Raminta Daniulaityte, Amit P. Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z. Watkins, and Russel Falck. "PREDOSE: a semantic web platform for drug abuse epidemiology using social media." *Journal of biomedical informatics* 46, no. 6 (2013): 985-997. <http://knoesis.org/node/2469>

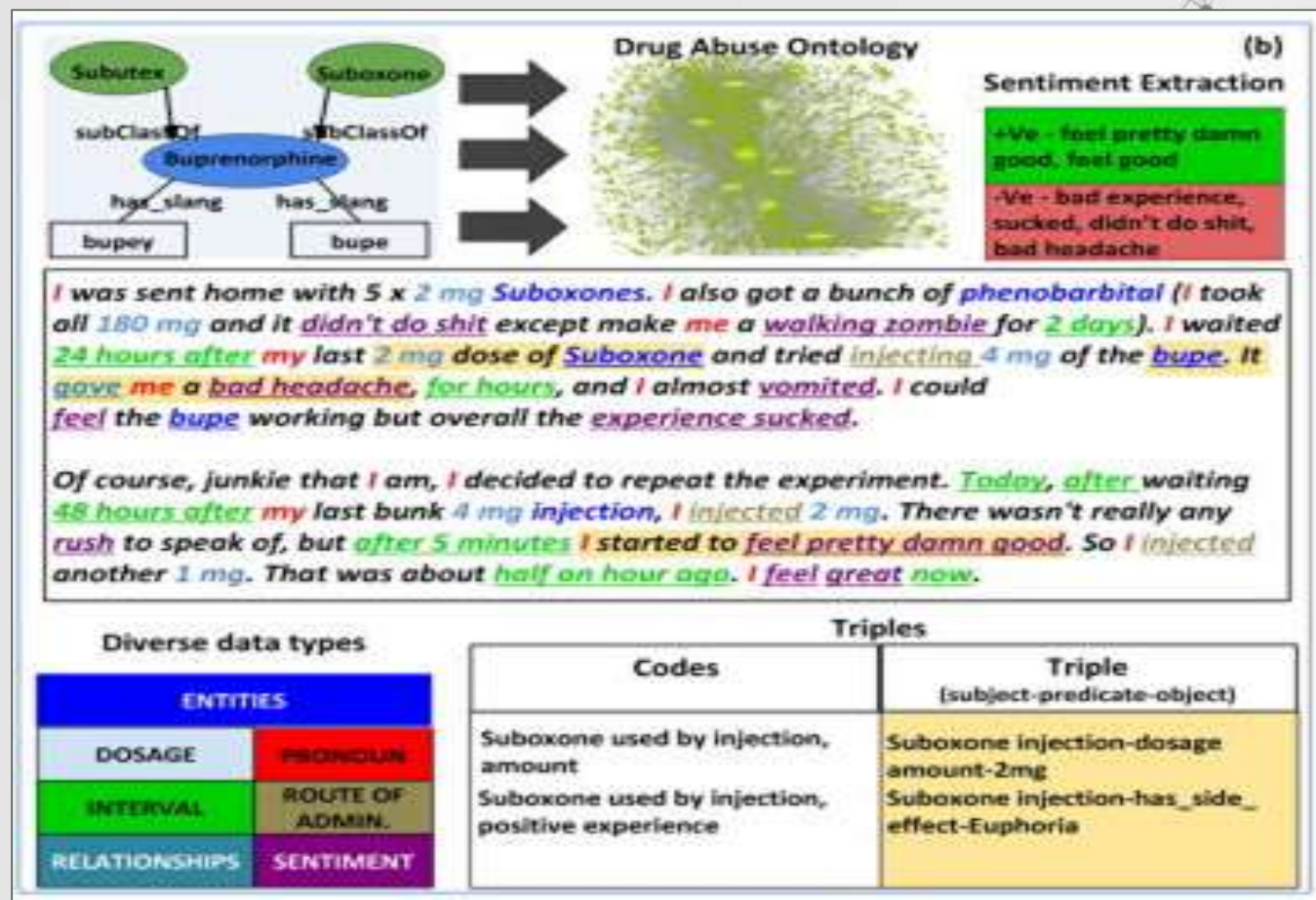
---

## Knowledge Bases:

Application of Drug Abuse  
Ontology along with slang term  
dictionaries and grammar

## Nature of Improvement:

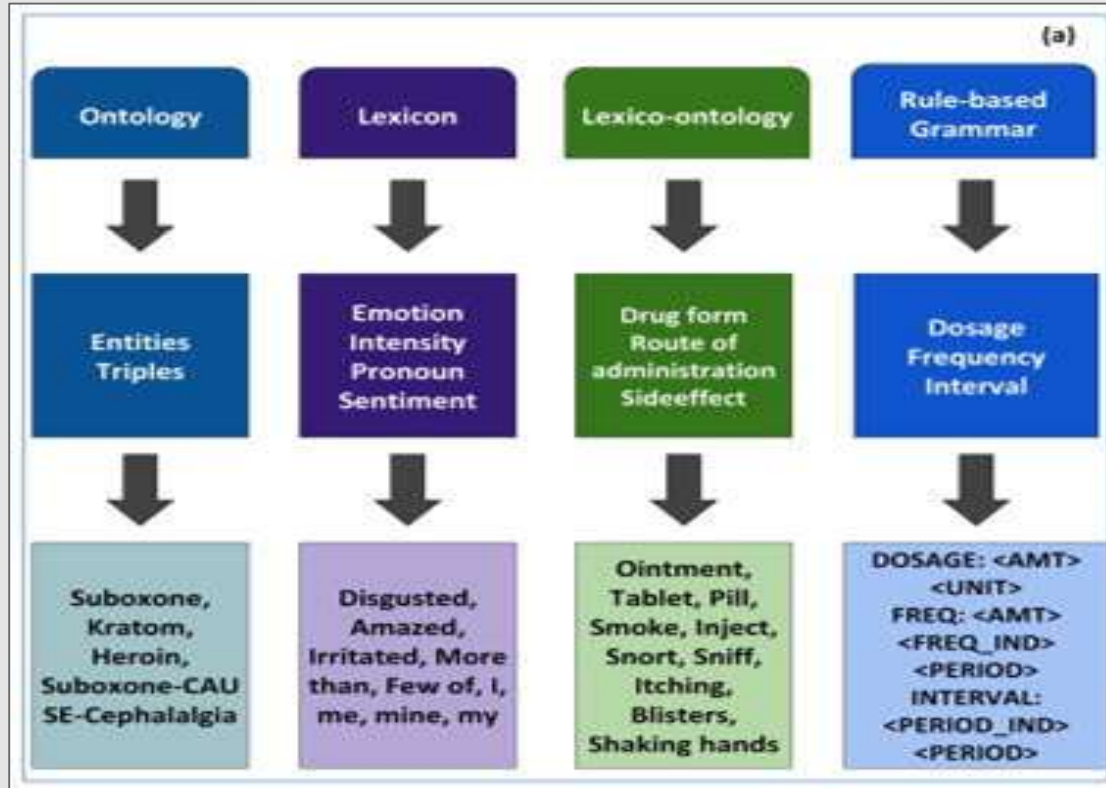
Recall and coverage



**Figure:** Understanding and analyzing drug abuse related discussions on web forums

# PREDOSE:

Smarter Data through Shared Context and Data Integration





# Understanding city traffic using sensor and textual observations

Pramod Anantharam, Krishnaprasad Thirunarayan, Surendra Marupudi, Amit Sheth, Tanvi Banerjee. Understanding City Traffic Dynamics Utilizing Sensor and Textual Observations. In 30th AAAI Conference on Artificial Intelligence (AAAI-16). Phoenix, Arizona; 2016. <http://knoesis.org/node/2145>

Pramod Anantharam, Krishnaprasad Thirunarayan, Amit Sheth. Traffic Analytics using Probabilistic Graphical Models Enhanced with Knowledge Bases. In 2nd International Workshop on Analytics for Cyber-Physical Systems (ACS-2013) at SIAM International Conference on Data Mining (SDM 2013). Austin, Texas; 2013. <http://knoesis.org/node/2476>

---

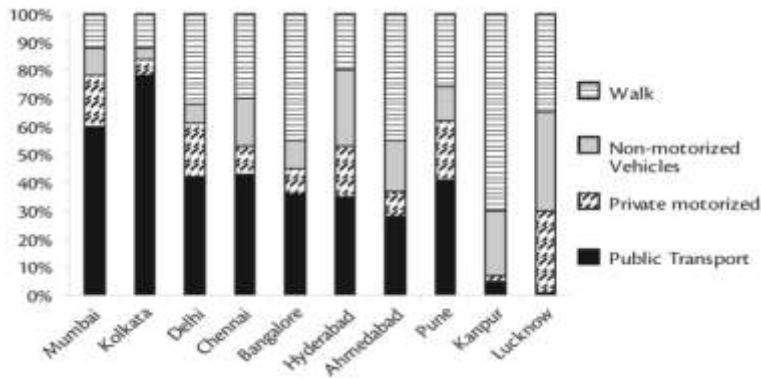
# Severity of the Traffic Problem

By 2001 over 285 million Indians lived in cities, more than in all North American cities combined (Office of the Registrar General of India 2001) <sup>1</sup>.

<sup>1</sup> The Crisis of Public Transport in India.

<sup>2</sup> IBM Smarter Traffic.

## Modes of Transportation in Indian Cities



Today there are more than [2011]  
**one billion cars** on the road.  
That number **will double by 2030**.

Traffic jumped **236%** as population grew nearly 20% between 1982 and 2001 in the U.S.

## The Texas Transportation Institute (TTI) Congestion report for the United States

The TTI report analyzed the impact of public transportation in 439 metropolitan areas, categorized as very large, large, medium and small.

- Very large areas (3+ million): Public transportation saved 557 million hours of delay and \$11.9 billion in congestion costs.
- Large areas (1-3 million): 59 million hours of delay and \$1.2 billion saved.
- Medium urban areas (500,000-1 million): 13 million hours and \$259 million saved.
- Small areas (less than 500,000): Public transportation saved 2 million hours of delay and \$31 million.



# Questions Asked Daily

- What time to start?
- What route to take?
- What is the reason for traffic?
- Wait for some time or re-route?





# Complementary Data Sources

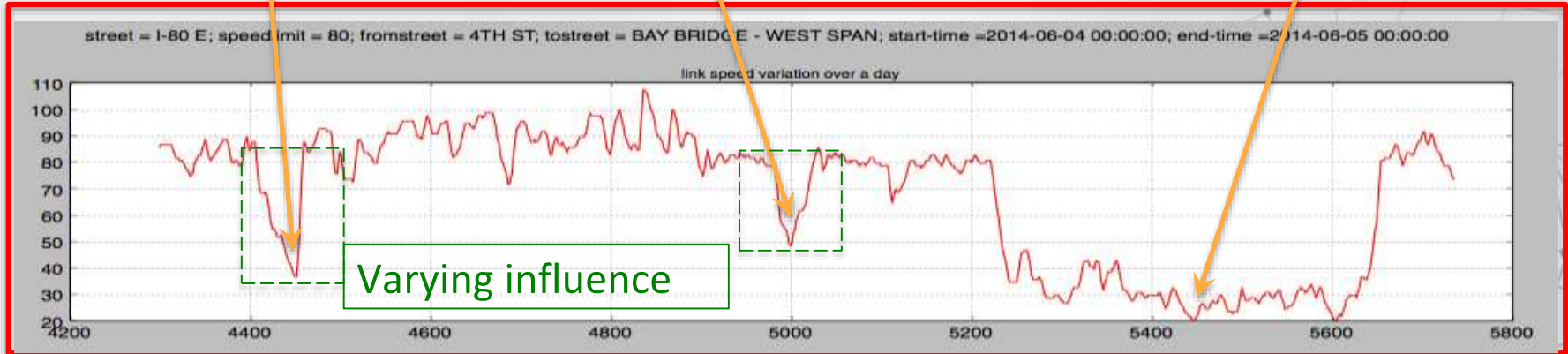


# Challenge: Non-linearity in Traffic Dynamics

Multiple Events



Interact with Each Other

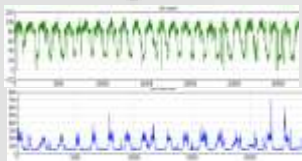


# Learning Context-specific LDS Models

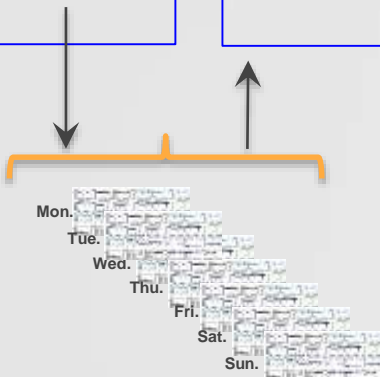
**Step 1:** Index data for **each link** for **day of week** and **hour of day** utilizing the traffic domain knowledge for piece-wise linear approximation

**Step 2:** Find the “typical” dynamics by computing the *mean* and choosing the *medoid* for each hour of day and day of week

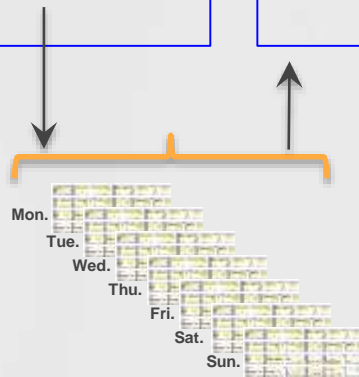
**Step 3:** Learn LDS parameters for the *medoid* for each hour of day (24 hours) and each day of week (7 days) resulting in  $24 \times 7 = 168$  models for each link



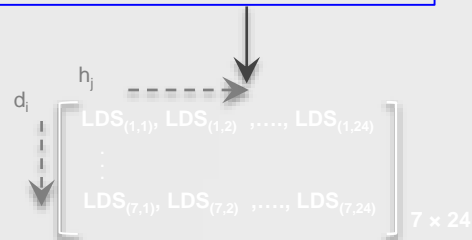
Speed/travel-time time series data from a link.



Time series data for each hour of day (1-24) for each day of week (Monday – Sunday).

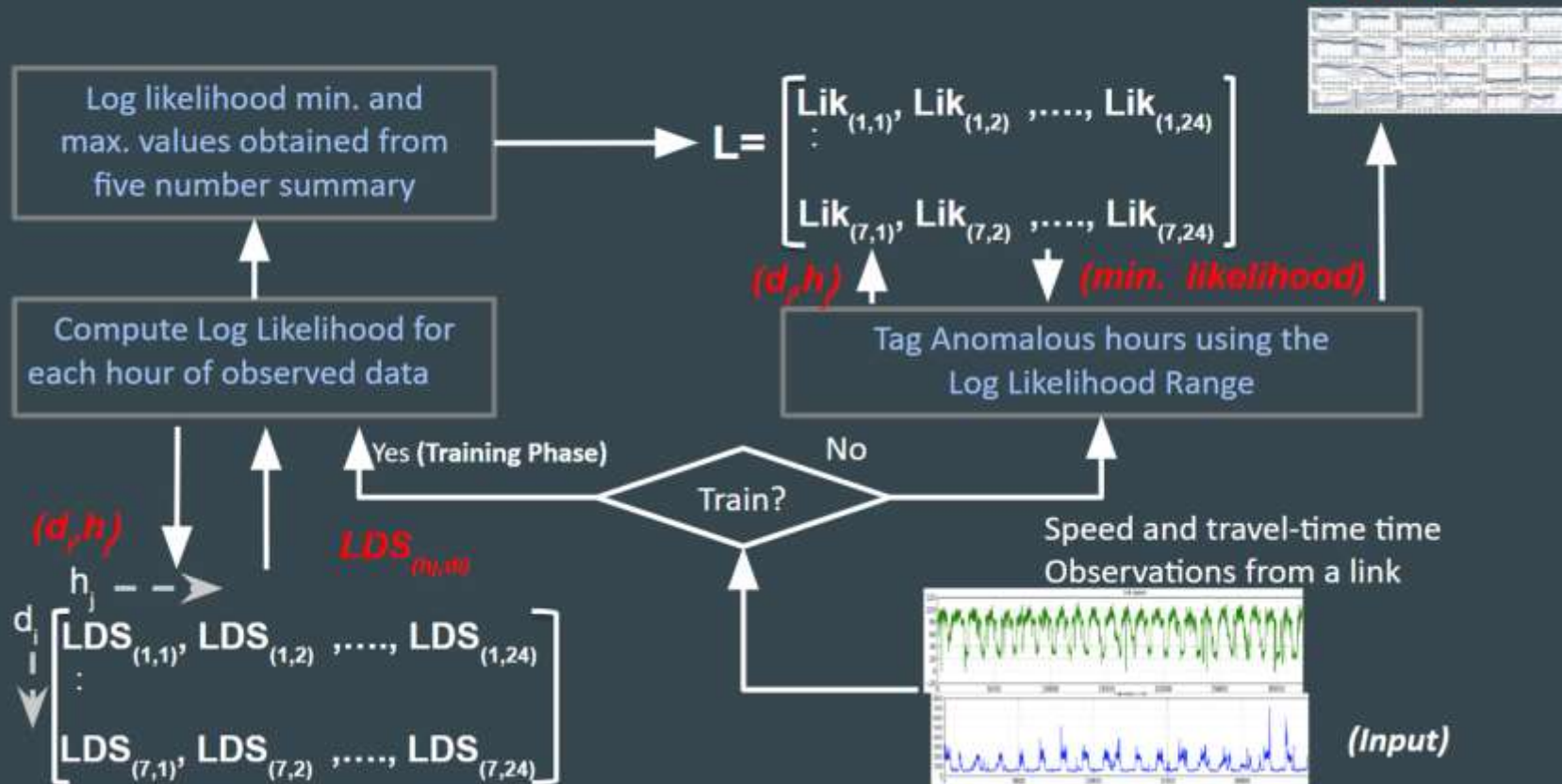


Mean time series computed for each day of week and hour of day along with the medoid.



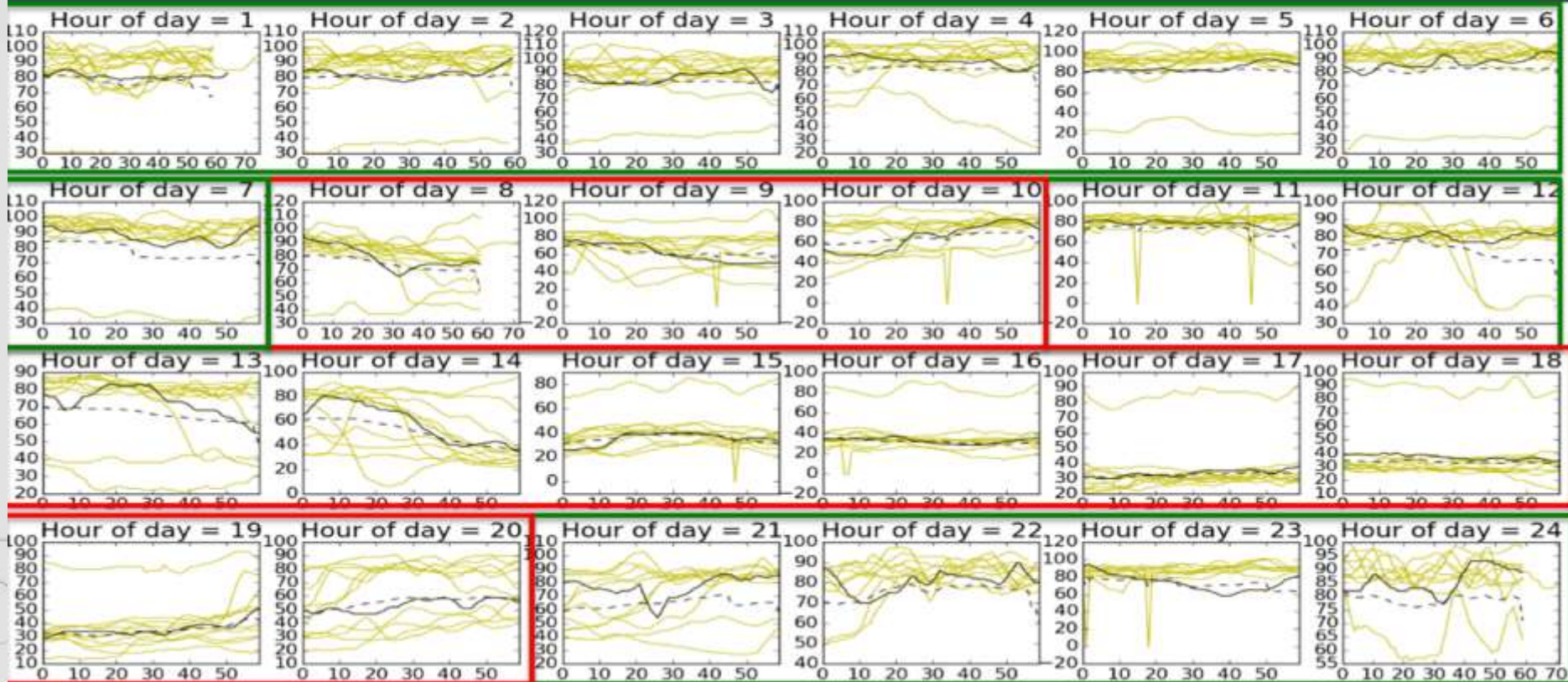
168 LDS models for each link; Total models learned = **425,712** i.e.,  $(2,534 \text{ links} \times 168 \text{ models per link})$ .

# Tagging Anomalies with LDS Models



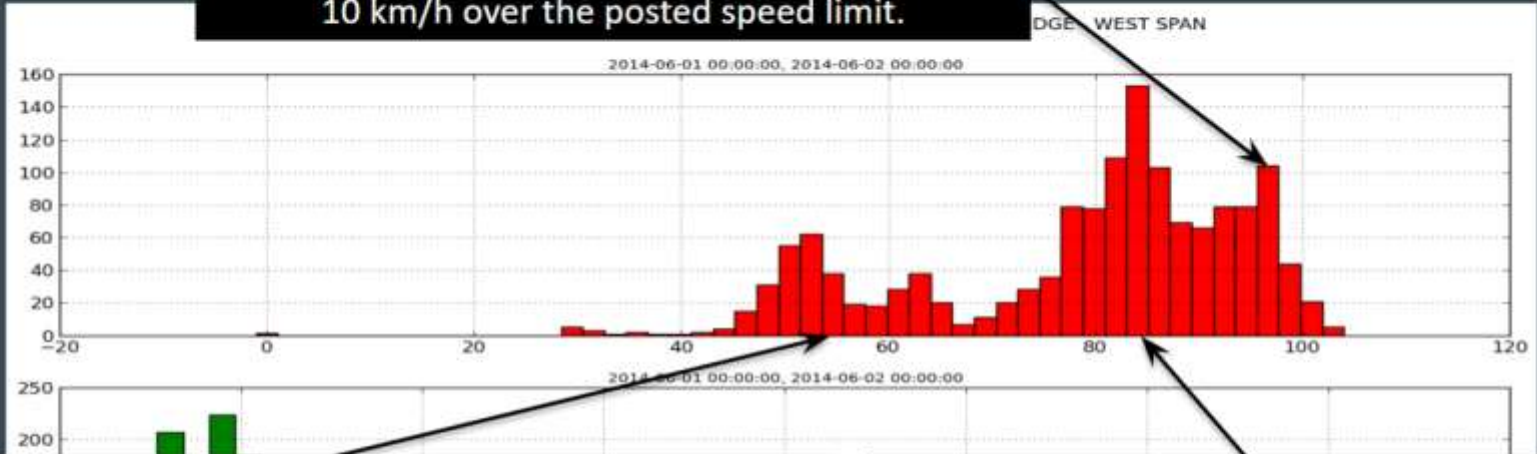


# Hourly Traffic Dynamics Over a Day



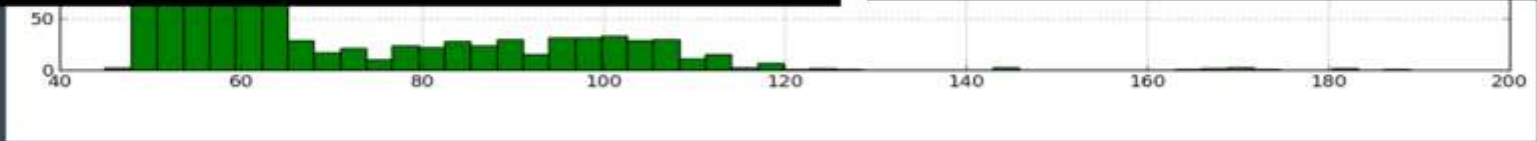
# Traffic Data: Possible Explanation

There are relatively few drivers who go more than 10 km/h over the posted speed limit.



There are situations in a day where the drivers are going (forced) below the speed limit e.g., rush hour traffic.

Most of the drivers tend to go 5 km/h over the posted speed limit.



Do these histograms resemble any probability distribution?



# Twitter as a Source of City Events

## Public Safety

**Post Local** @postlocal 10 Sep  
Small explosion outside Montgomery County's **public safety** building, possibly a firecracker wapo.st/19ExveU  
View summary Reply Retweet Favorite More

## Urban Planning

**Kansas City Scout** @KansasCityScout 5m  
My KC Scout Traffic Alert Cleared: **Roadwork** on I-435 NB (East of KC) PAST 24 HWY  
Message #65072-142  
Expand Reply Retweet Favorite More

## Gov. & Agency Admin.

**Oliver Galak** @OliverGalak 4 Jun  
"AFA, whose president is a firm ally of the **Kirchner administration**", dice el aliento fétido de ABC news abcnews.go.com/ABC\_Univision/...  
@LarustaMoore  
View summary Reply Retweet Favorite More

## Social Programs

**Elle A from Elle A** @elle\_amm\_alch 19m  
hungry w/ no hope & no **social programs** to access if you suffer from hardship. They want you terminally ill w/o healthcare.  
Expand Reply Retweet Favorite More

## Healthcare

**Jane Frawley** @JaneFrawley 1m  
Australian Government urged to invest more in complementary **healthcare** foodnavigator-asia.com/Policy/New-Aus... via @FoodNavAsia  
Expand Reply Retweet Favorite More

## Education

**Greg Gorman** @ggorman 1h  
Being a Courageous Educator and School Leader in an Era of Foolish **Education Policy** the21stcenturyprincipal.blogspot.com/2013/09/being-... via @Digg  
Expand Reply Retweet Favorite More

## Energy & water

**Alan Hawkins** @ECJourn 5 Sep  
**City Power** workers turn off the lights in parts of @Johannesburg because of a new shift system #strikeseason news24.com/SouthAfrica/Ne...  
View summary Reply Retweet Favorite More



**Kim Hernandez** @Kimm31 16 Sep  
@unitedutilities please please will you fix the air in the **water supply problem**, CH46, my kids refuse to drink the water  
Expand Reply Retweet Favorite More

## Environmental

**Helaman Copp** @HelamanCopp 13 Sep  
e're at Muthurwana Secondary School in Sundari Village, **Waste management** is a problem, for one, uQshdgb7k  
Expand Reply Retweet Favorite More



**Autogas Limited** @AutogasLPG 31 Aug  
@Warwick resident? Poor air quality is a **problem** in the town. Switching to LPG could help tackle **air pollution** ow.ly/oIRrs  
Expand Reply Retweet Favorite More

## Transportation

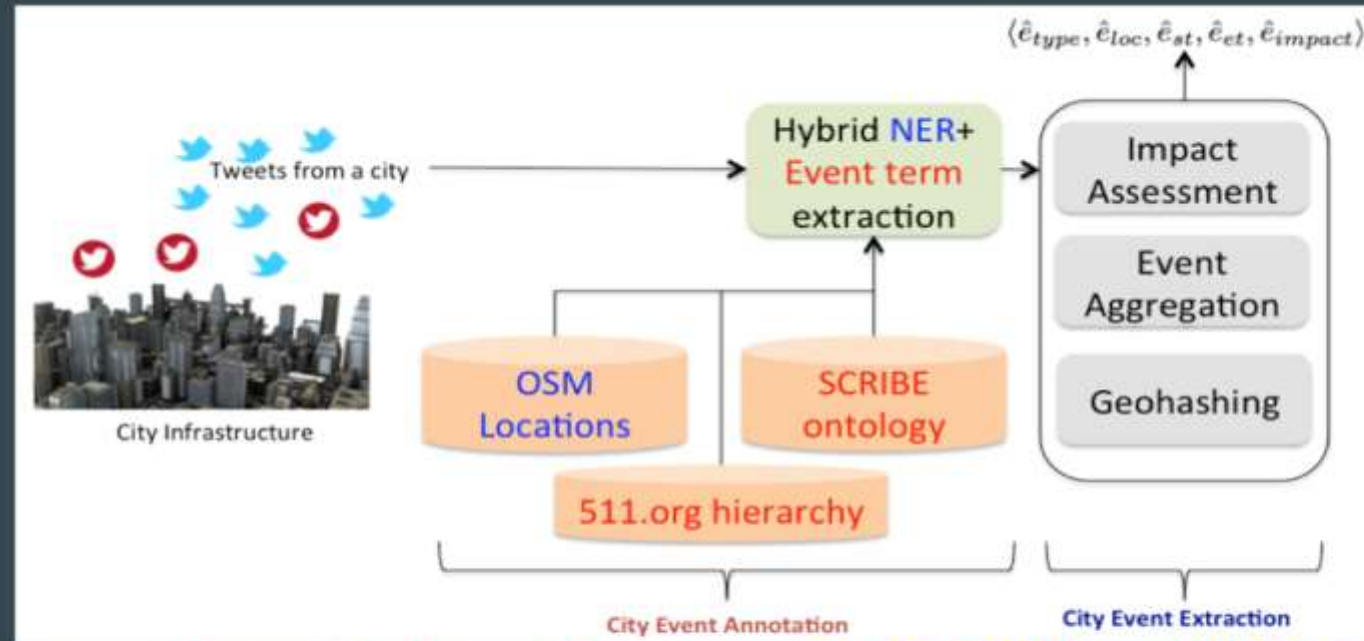


**Ram N Kumar** @ramnkumar 51s  
Crossing Yamuna now, it is mirred with **traffic Jam** and hooliganism by people going for Ganpati idol immersion at... fb.me/2iZjrGDO  
Expand Reply Retweet Favorite More



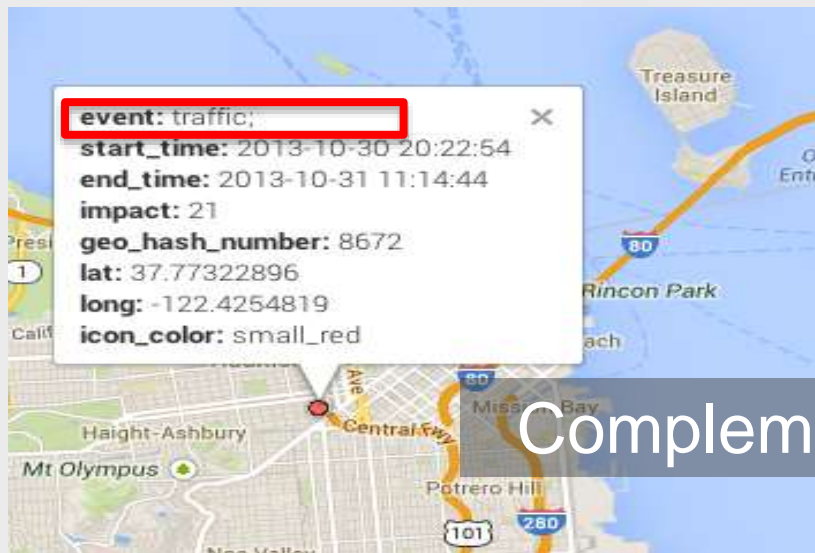
**Erikaaaa** @erikasbtr97 2h  
@louisacc im at the station now! My **bus** came ahha beware there are loads of **delays** at the station  
View conversation Reply Retweet Favorite More

# Extracting City Events from Textual Data

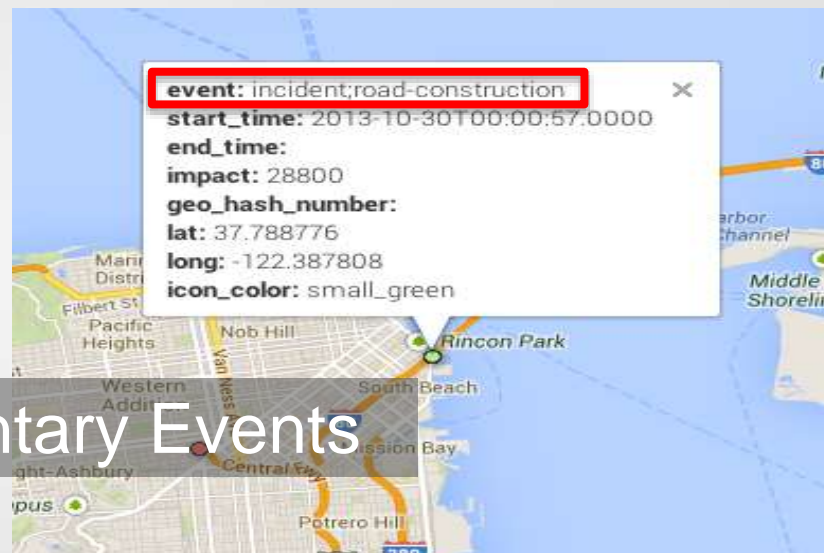


Last night in CA... (@ Half B-LOCATION Moon I-LOCATION Bay  
B-LOCATION Brewing I-LOCATION Company w/ 8 others)  
<http://t.co/w0eGEJjApY>

# Textual Events from Tweets vs. 511.org: Complementary



Traffic



Incident;  
road-construction

Complementary Events



# Understanding: Semantic Annotation of Sensor + Textual Data Utilizing Background Knowledge



- Thing
  - area-wide-information
  - closures
  - collision
  - delay-status-car
  - device-status
  - disasters
  - disturbances
  - incident
    - abandoned-vehicle
    - accident
    - accident-cleared
    - accident-investigation-work
    - accident-involving-a-bicycle
    - accident-involving-a-bus
    - accident-involving-a-motorcycle
    - accident-involving-a-pedestrian
    - accident-involving-a-train
    - accident-involving-a-truck
    - accident-involving-hazardous-materials
    - acid-spill
    - bus-fire
    - bus-stuck-under-bridge
    - chemical-spill

Domain knowledge in the form of traffic vocabulary

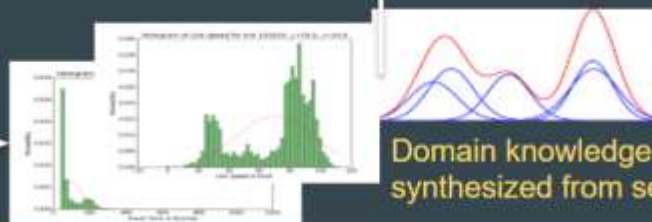
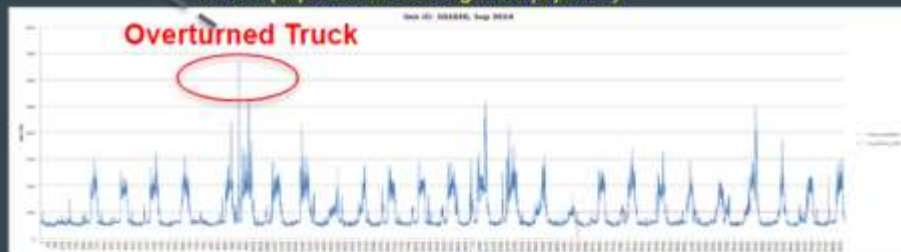
Accident **B-EVENT** on **O** the **O** Golden **B-LOCATION** Highway **I-LOCATION** at **O** the **O** Viking **O** robots **O** in **O** Devland **O** JHB **O** , **O** ambo **O** truck **O** , **O** injured **B-EVENT** treating **O** themselves **O**

Explained-by

Horizontal operator: relating/mapping data from different modality to a concept (theme) within a spatio-temporal context;  
Spatial context even include what it means to have a slow traffic for the type of road (<http://wiki.knoesis.org/index.php/PCS>)



Image Credit:  
<http://traffic.511.org/index>



Domain knowledge of traffic flow synthesized from sensor data

# How traffic analysis captures complexity of the real-world?

This example demonstrates use of:

- Multimodal data streams (types of events from text - signature from sensor data).
- Multiple sources of declarative knowledge/ontologies.
- Semantic annotations and enrichments.
- Use of rich representation (PGM)
  - learned probabilistic models improved using declarative knowledge
- Statistical approach to create normalcy models and understand anomalies using historical data. Explain anomalies using extracted events.
  - use declarative knowledge to approximate nonlinear models using a collection of linear dynamical systems
- Provide actionable information.





# **Semantic, Cognitive, and Perceptual Computing**

Paradigms That Shape Human Experience

---

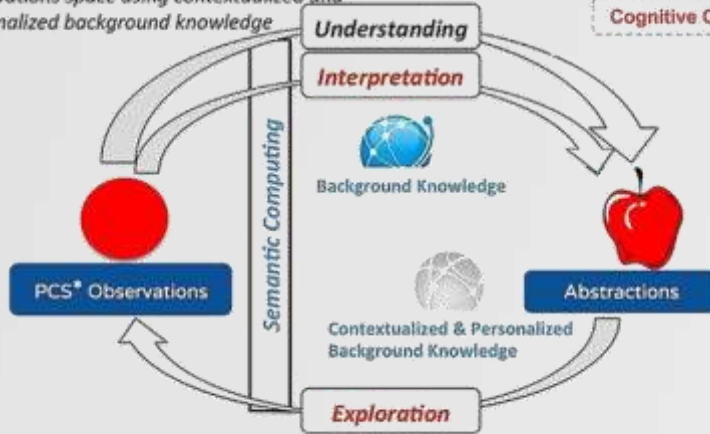
# Semantic, Cognitive, and Perceptual Computing: Paradigms That Shape Human Experience

## Perceptual Computing

Perceptual computing is characterized by a cyclical process of **interpretation** and **exploration** of observations space using contextualized and personalized background knowledge

Cognitive computing is a technology for **understanding** observations utilizing techniques that loosely mimic human cognition

## Cognitive Computing



\*Real-world events manifest in observations spanning Physical-Cyber-Social (PCS) modalities

<http://bit.ly/SCPComputing>

Humans are interested in **high-level concepts** (phenotypic characteristics).

**Semantic** Computing: Assign labels and associate meanings (representation & contextualization).

**Cognitive** Computing: Interpretation of data with respect to perspectives, constraints, domain knowledge, and personal context.

**Perceptual** Computing: A cyclical process of semantic-cognitive computing for higher level of perception and reasoning (abstraction & action).



# Interplay between Semantic, Cognitive, and Perceptual Computing

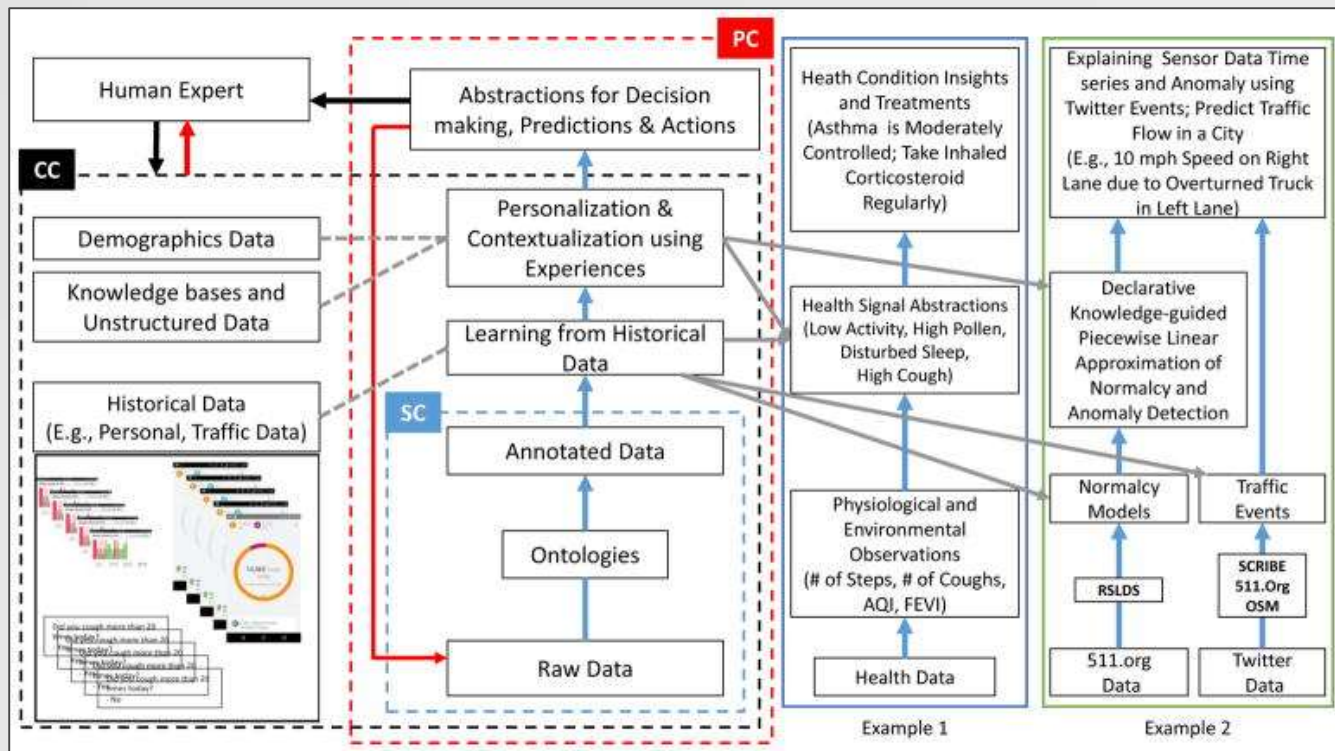


Figure: SCP with Examples on Asthma and Traffic Use-cases



# **PRESENT**

Knowledge Graph (Infusion) for Deep Learning

---



# BIG DATA

**Top-down** symbolic approach (concepts, rules) in data reasoning, inferencing, and deduction.

**SEMANTIC &  
KNOWLEDGE GRAPH**

**Bottom-up** statistical approach in searching, analyzing and deriving insights from Big Data.

**MACHINE/ DEEP  
LEARNING**



# Deep Learning Critical Review

- **Deep Learning thus far is data hungry** : Human beings can learn abstract relationships in a few trials (explicit and implicit definitions).
- **Deep Learning thus far is shallow and has limited capacity for transfer** : Each model is very specific for the task and generalization is not straightforward.
- **Deep Learning thus far has no natural way to deal with hierarchical structure** : Hierarchical structure in English sentences aren't captured in sequential techniques (which motivate move from Sequential LSTMs to Tree LSTMs).

<https://arxiv.org/pdf/1503.00075.pdf>



# Deep Learning Critical Review (Continued...)

- **Deep Learning thus far has struggled with open-ended inference:** Current machine reading systems have achieved some degree of success in tasks like **SQuAD**, in which the answer to a given question is explicitly contained within a text, but far less success in tasks in which inference goes beyond what is explicit in a text, either by combining multiple sentences (so called multi-hop inference) or by combining explicit sentences with background knowledge that is not stated in a specific text selection.
- **Deep Learning thus far has not been well integrated with prior knowledge:** Consider a set of easily-drawn inferences that people can readily answer without anything like direct training, such as Who is taller, Prince William or his baby son Prince George? Can you make a salad out of a polyester shirt? If you stick a pin into a carrot, does it make a hole in the carrot or in the pin? As far as I know, nobody has even tried to tackle this sort of thing with deep learning. Such apparently simple problems require humans to integrate knowledge across vastly disparate sources, and as such are a long way from the sweet spot of deep learning-style perceptual classification.



The top corners of the slide feature decorative geometric patterns. On the left, there are several interconnected triangles and lines forming a network-like structure. On the right, a similar but more complex network of lines and dots is visible. These elements are rendered in a light gray color, blending into the background.

# WHY KNOWLEDGE-INFUSED LEARNING?

- **Ambiguous** online healthcare communications and **difficult** to engineer discriminative features.
- Domain-specific embedding models provide a **shallow infusion** of knowledge.
- Decrease the **dependence** on large datasets
- **Reduce bias** in the dataset (ie: potentially avoid social discrimination and unfair treatment)
- Provide **information provenance**: Allowing explainability of a model
- **Improve information coverage** specific to a domain that would be missed otherwise
- **Reduce time and space complexity** of the models architecture
- Improve models **sensitivity** and **specificity**

# KNOWLEDGE-INFUSED LEARNING



## **SHALLOW** Infusion

of knowledge graphs to improve the semantic and conceptual processing of data.

Deeper and congruent incorporation or integration of the knowledge graphs in the learning techniques.

## **SEMI-DEEP** Infusion



## **DEEP** Infusion

*(Part of Future KG Strategy)*

combines statistical AI (bottom-up) and symbolic AI learning techniques (top-down) for hybrid and integrated intelligent systems.



# Shallow Infusion

*In shallow infusion,  
both the external  
information and  
method of knowledge  
infusion is shallow.*

---

# Shallow External Knowledge

- Shallow external knowledge is described as those form of information which are extracted from text based on some heuristics, often designed for task-specific problems:
  - Bag of Words/Phrases from Corpus
  - Bag of Words/Phrases from Semantic Lexicons
  - Count of Nouns, Pronouns, Verbs
  - Sentiment and Emotions of the sentence
  - Latent topics describing the documents
  - Label assignment to words or phrases in sentence:
    - Mary sold the book to John
    - Mary (agent), the book (theme), John (recipient), sold (predicate)



# Shallow Method of Knowledge Infusion

- **Statistical NLP Methods:**

- Term Frequency and Inverse Document Frequency
- Latent Dirichlet Allocation/ Hierarchical Dirichlet Process
- Semantic Role Labeling
- Zipf Law

- **Neural NLP Methods:**

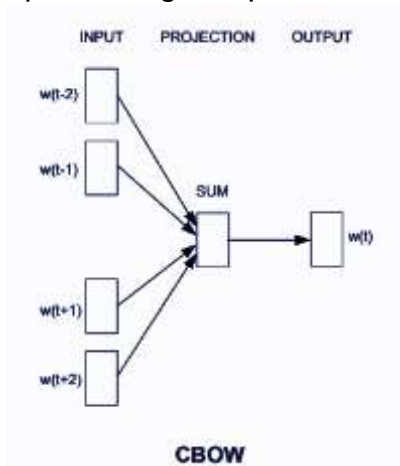
- Word2Vec, GLoVE
- Sentence2Vec, Doc2Vec
- BERT, ELMo



# Examples of Shallow Infusion from NLP domain

## Word2Vec

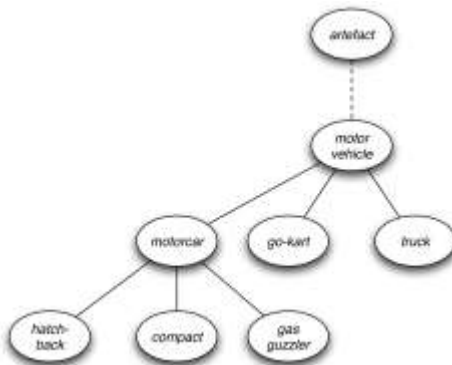
**Knowledge:** Domain specific large corpora



*“Context is represented by a set of words for a given target word”*

## Retrofitting

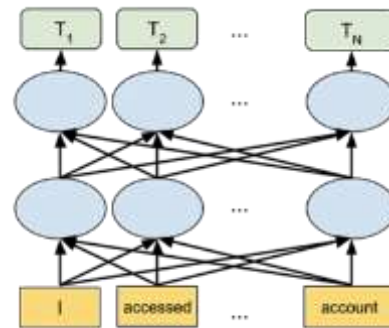
**Knowledge:** pre-trained embeddings + semantic lexicons



*“Learned embeddings are further **enriched** by using semantic lexicons”*

## ★ BERT

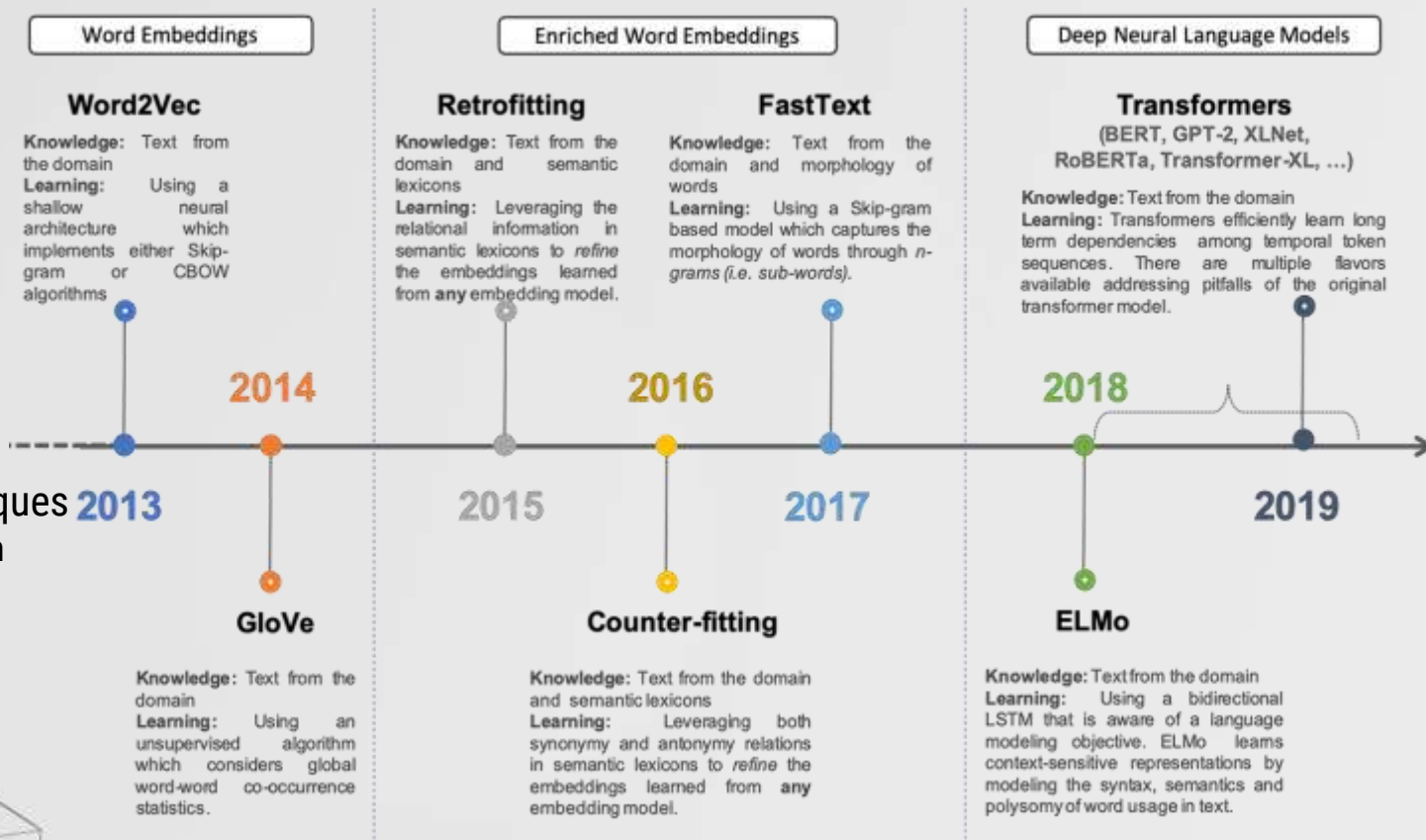
**Knowledge:** Domain specific large corpora



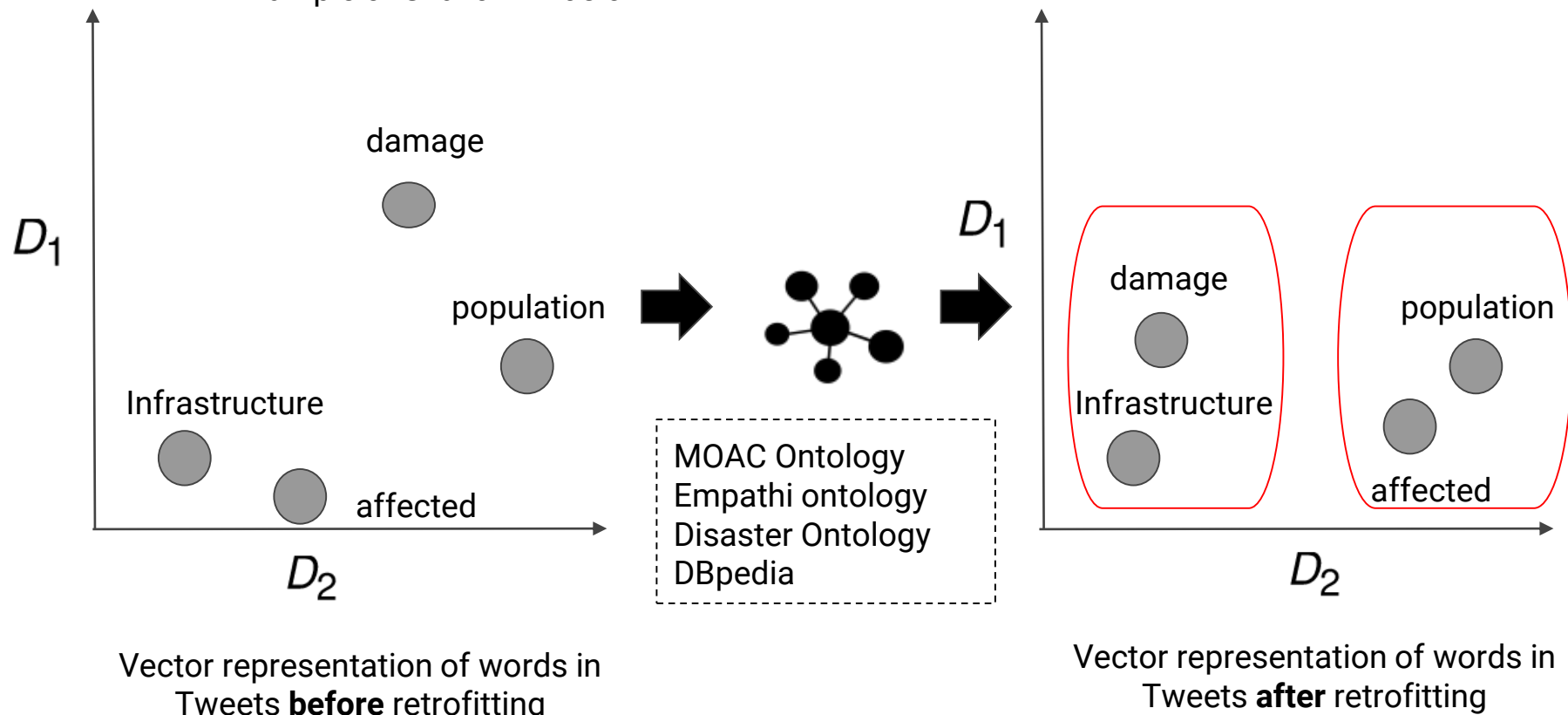
*“Uses language modeling objective to learn the contextual representations”*

# SHALLOW Infusion of Knowledge for Machine/ Deep Learning in Brief

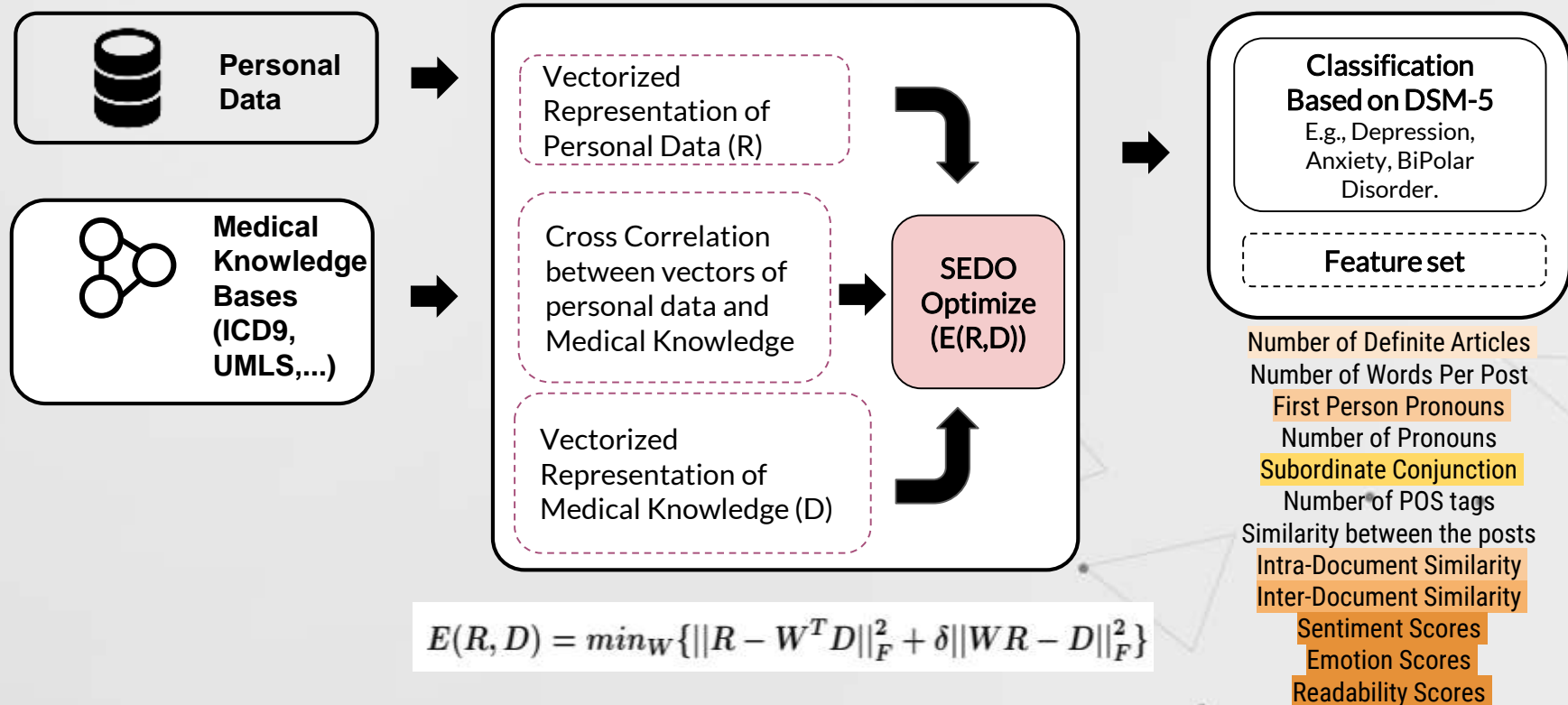
Chronological  
arrangement of  
shallow  
Infusion techniques  
From NLP domain



## Example of Shallow Infusion



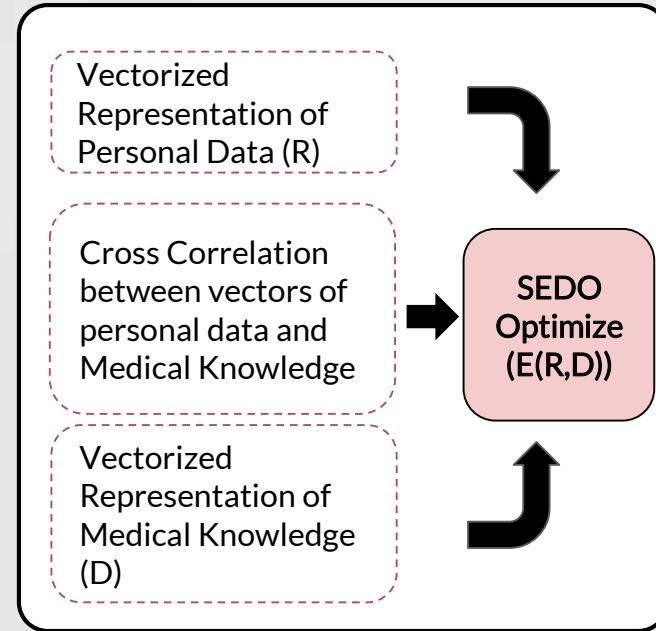
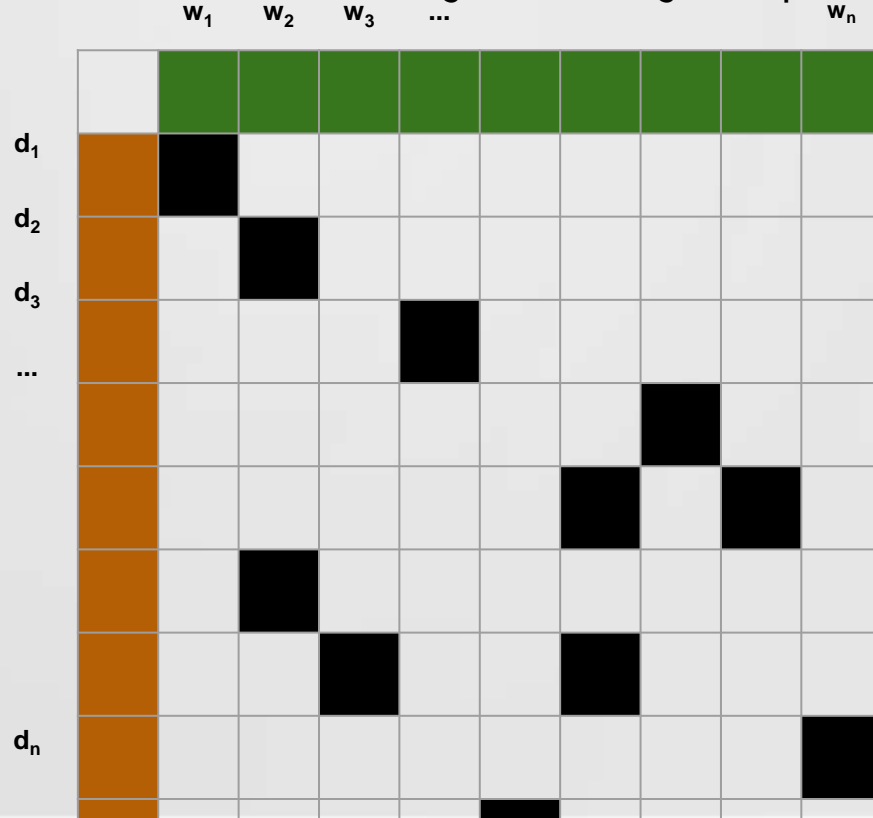
# Explaining the prediction of mental health disorders (CIKM 2018)



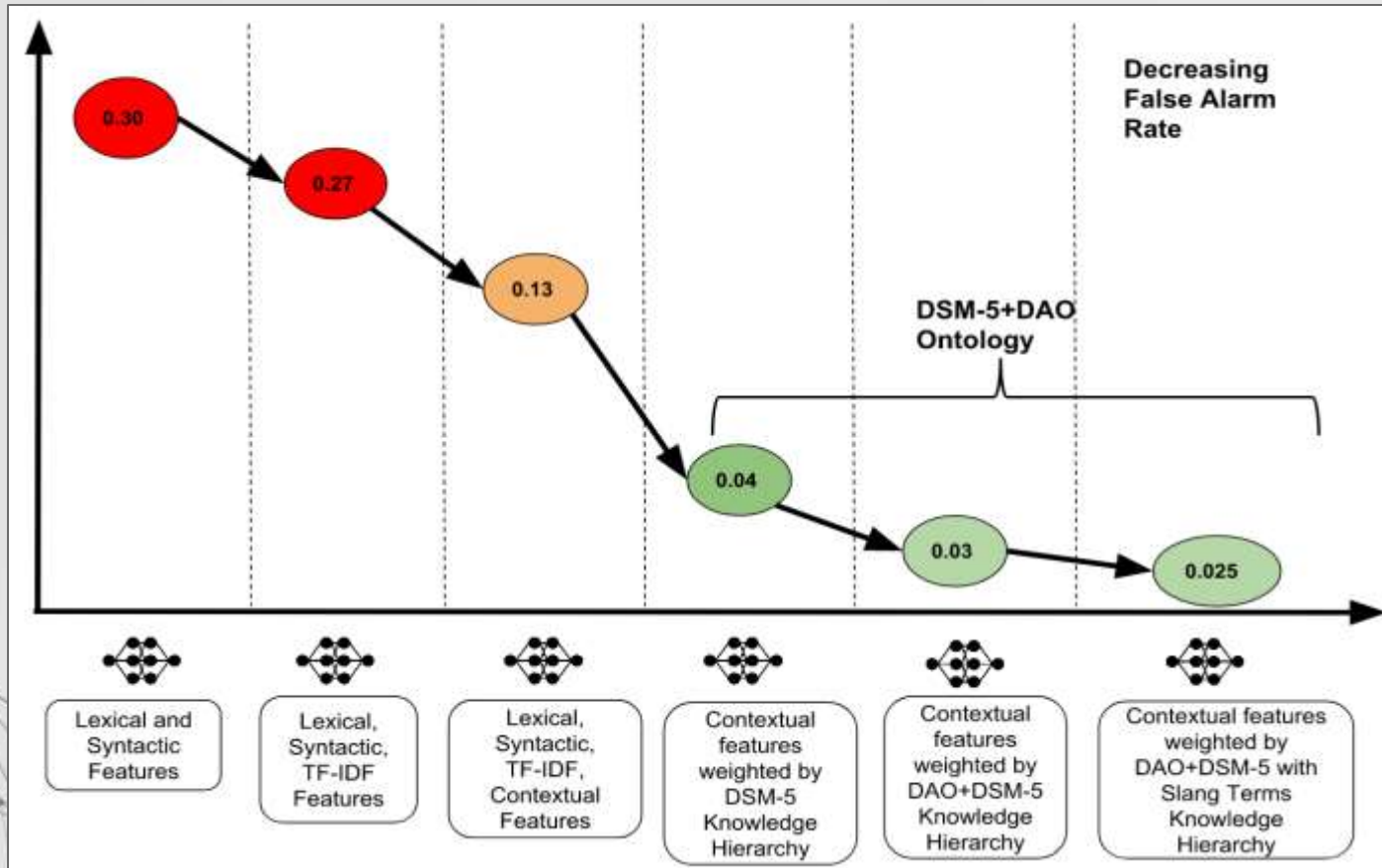


## *Continued: Explaining the prediction of mental health disorders (CIKM 2018)*

- Explanation through word features created through Semantic Encoding & Decoding Optimization.
  - Semantic encoding of personal data into knowledge space
  - Semantic decoding of knowledge into personal data space.



## Continued: Explaining the prediction of mental health disorders (CIKM 2018)



# Semi-Deep Infusion

*In semi-deep infusion,  
external knowledge is  
involved through attention  
mechanism or learnable  
knowledge constraints  
acting as a sentinel to  
guide model learning.*

---

# SEMI-DEEP Infusion of Knowledge for Machine/ Deep Learning in Brief

## Forcing Methods

### Teacher Forcing

**Knowledge:** the input word embedding to the model.

**Learning:** Encoding and Decoding, where the knowledge is given as input at the decoding stage.

1988

2017

### Professor Forcing

**Knowledge:** word vectors of the input

**Learning:** Generative Adversarial Network based learning using encoder and decoder.

## Neural Attention Models

### Neural Self-Attention

**Knowledge:** Weighted average of different sequences of a sentence

**Learning:** correlation between current word and previous part of sentence in an LSTM

2016

2016

### Knowledge-guided Neural Attention

**Knowledge:** The dependency parse tree of the sentence

**Learning:** joint learning in RNN with structural linguistic property and sentence sequence.

## Knowledge-based Models

### Knowledge-based LSTMs

**Knowledge:** External knowledge base

**Learning:** Each LSTM cell is augmented with a Knowledge module which is triggered based on attention probability acting as sentinel to attend or ignore background knowledge.

2017

2018

### Knowledge-based GANs

**Knowledge:** ground truth sentences

**Learning:** Though generative models are effective in capturing knowledge through min-max loss, leveraging specialized losses using ground truth knowledge propel generality in model.



# External Knowledge through Attention

- A neural attention mechanism equips a neural network with the ability to focus on a subset of its inputs (or features):
  - **Hard Attention** or Position specific attention : location of important entities and relationship in the text are hard-coded in the model. Thus allowing efficiency in feature engineering, however, the model suffer from exposure bias.
  - **Soft Attention**: The model learns to attend to specific parts of the text while generating the word describing that part (following distributional semantics).
  - **Attention with Knowledge base**: background knowledge is integrated using an attention mechanism, which decide whether to attend to background knowledge and which information from KBs is useful.



# External Knowledge through Learnable Constraints

- Learnable constraints are empirical thresholds (probabilistic value) learnt by the model which allows it to adaptively learn.
- It can be done in following ways:
  - Learning based on pre-structured axiomatic rules - axiomatic knowledge
  - Learning based on difference in content similarity - KL Divergence, Cross-entropy loss
  - Learning based on commonsense knowledge - ConceptNet
  - Learning over different permutations of text generated through synonyms, antonyms, and homonyms.



Template: fill in the blanks

\_\_\_\_\_ meant to \_\_\_\_\_ not to \_\_\_\_\_

Generative  
Model

It was meant to dazzle not to make it

Target:

It was meant to dazzle not to make  
sense

Infilling Content  
Matching through  
averaged KL  
Divergence

Learnable knowledge  
constraint module

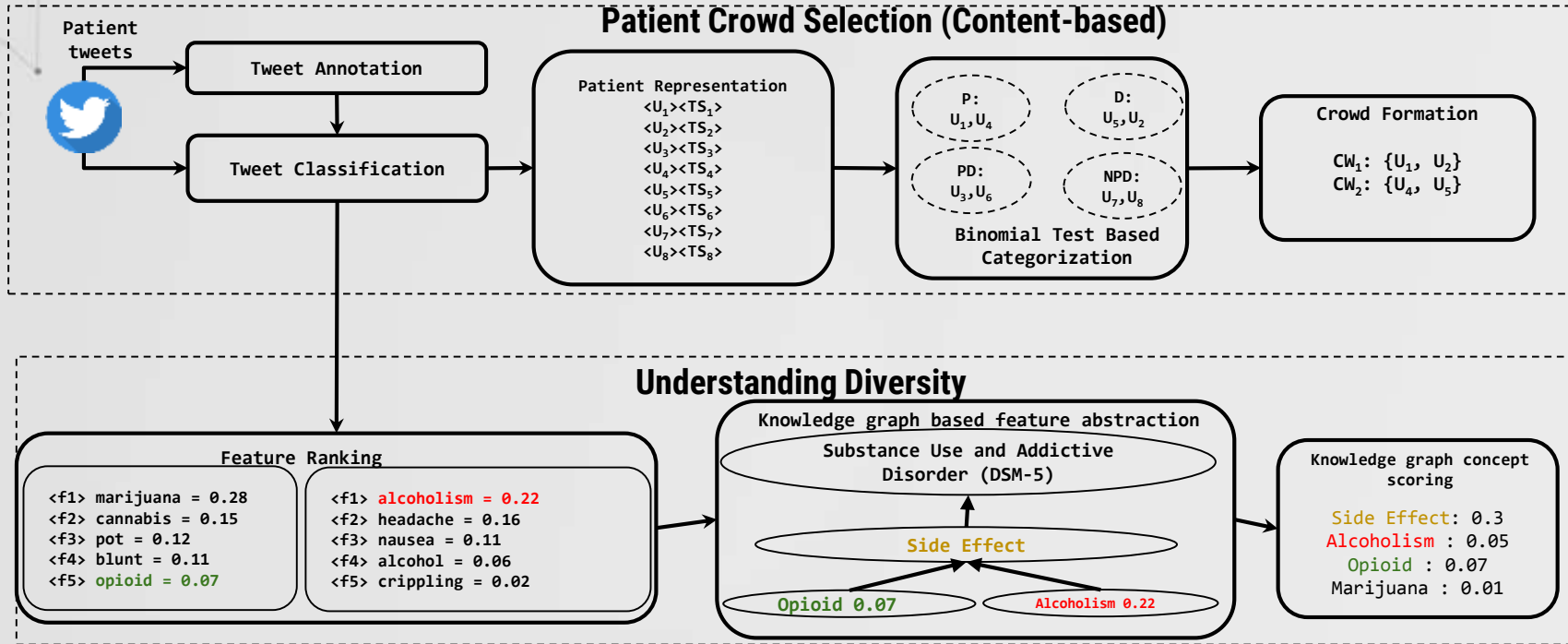




# Relevant research:

## Explaining Mental Health Classification

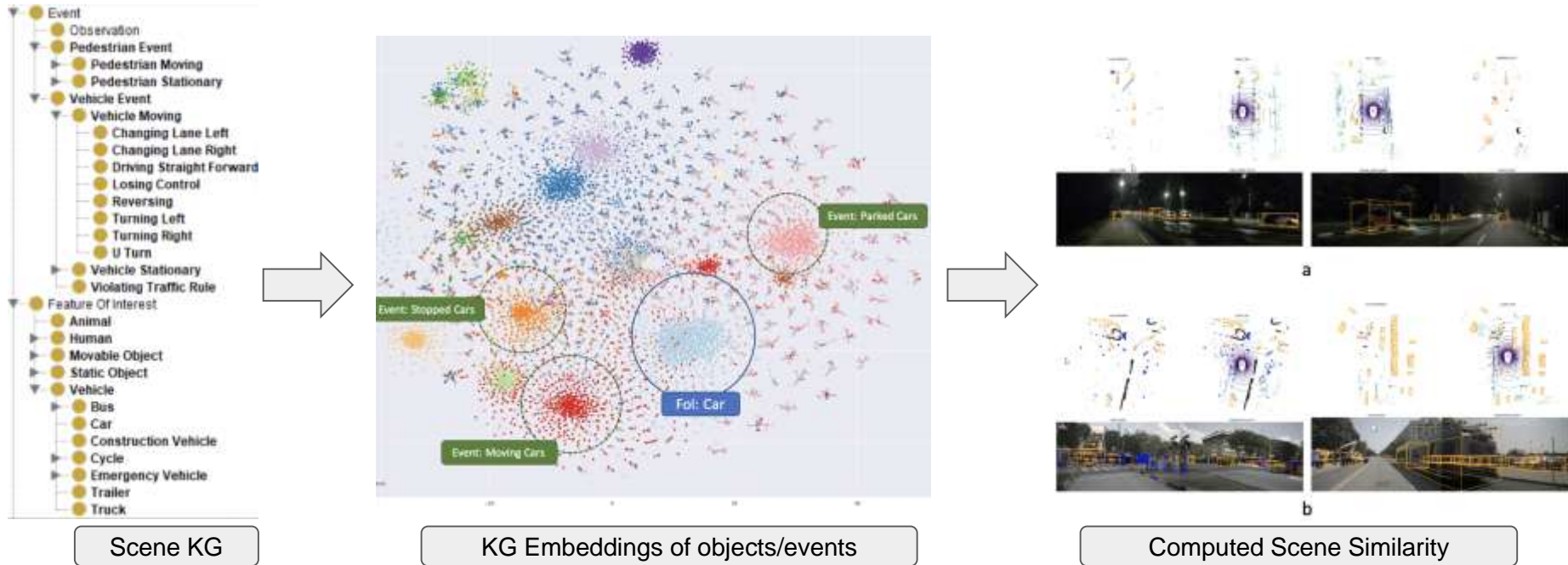
by adapting the prediction for wisdom of crowd



Bhatt, S., Gaur, M., Bullemer, B., Shalin, V., Sheth, A., & Minnery, B. (2018, December). Enhancing crowd wisdom using explainable diversity inferred from social media. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 293-300). IEEE.

Relevant Research:

# Knowledge Graph Embeddings for Autonomous Driving



Wickramarachchi, Ruwan., Henson, Cory., and Sheth, Amit. An evaluation of knowledge graph embeddings for autonomous driving data: Experience and practice. In AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020).

# Deep Infusion

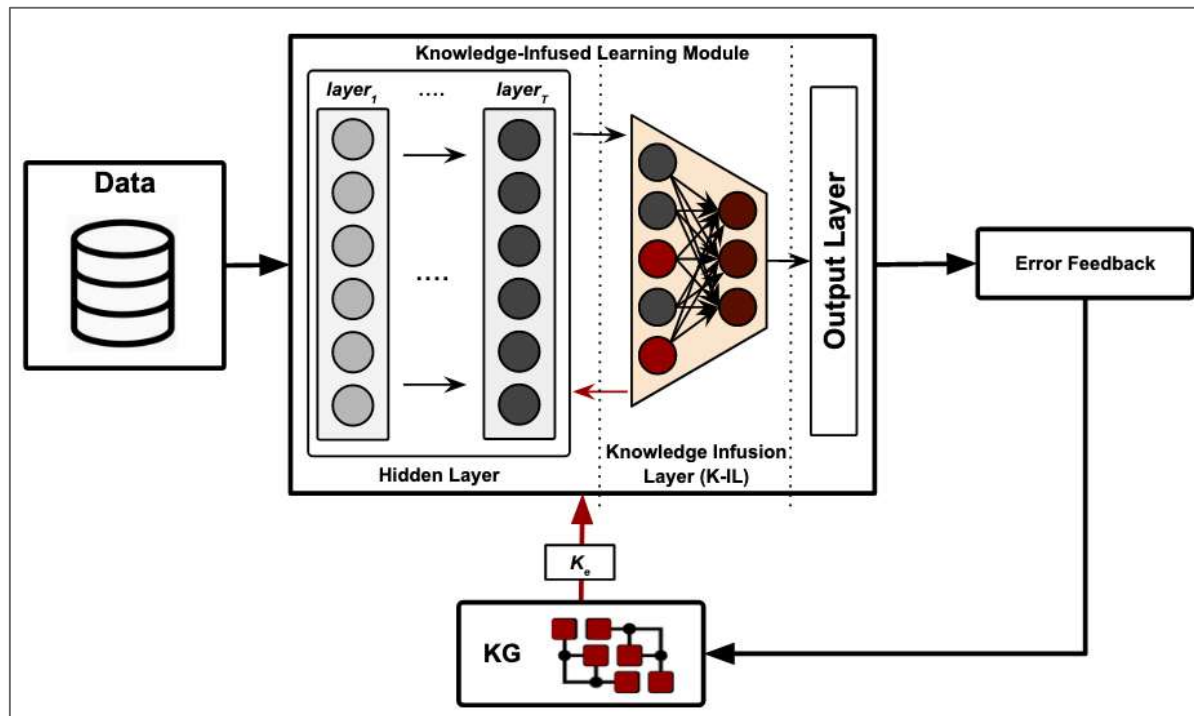
*In deep infusion, external knowledge is incorporated within the neural network. The latent representation is transmitted between layers including hidden layers, while achieving enhanced abstractions.*

---

## DEEP Infusion of Knowledge for Machine/ Deep Learning

122

- Infusion can take place (i) **before the output layer** (e.g., Soft-Max), (ii) between hidden layers (e.g., reinforcing the gates of an NLM layer, modulating the hidden states of NLM layers)
- Classification **error** and **KG** determine the need for infusing knowledge. The Knowledge Infusion Layer incorporates the knowledge in the latent representation before output layer.




Ugur Kursuncu, Manas Gaur, and Amit Sheth. "Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning." AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering. Palo Alto, California, USA. 2020.

# Open Research Questions:

## Knowledge-Infused Learning

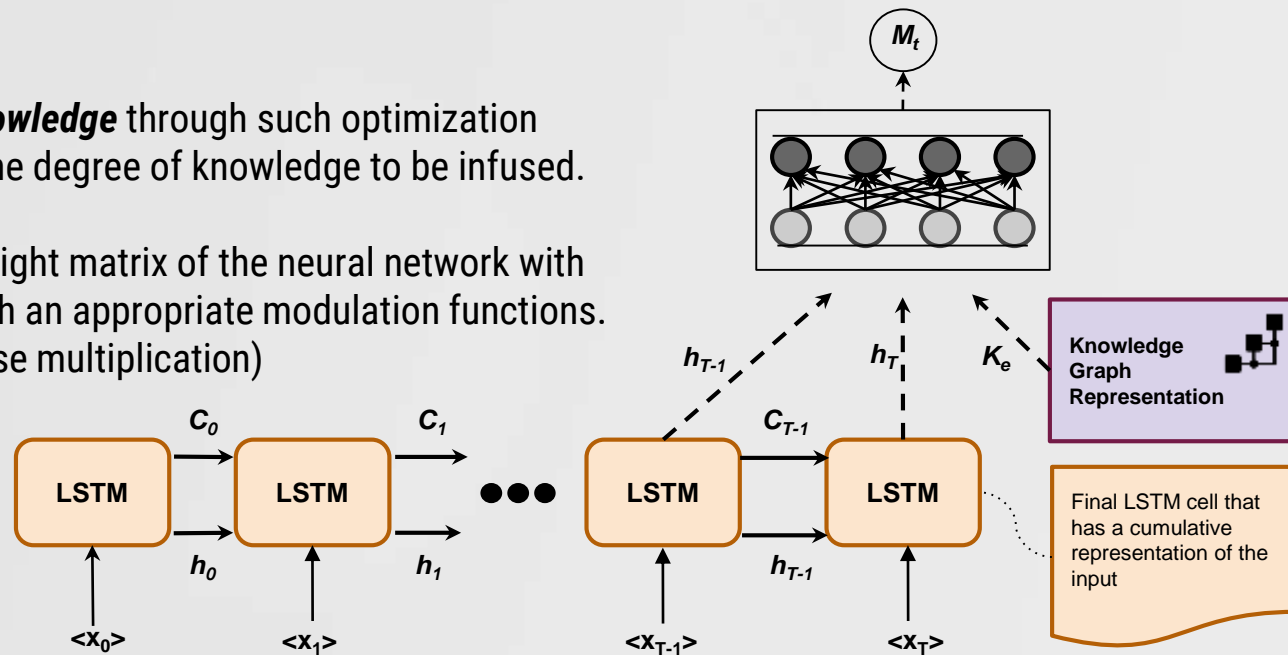
- How to decide whether to infuse knowledge or not at a particular stage in learning between layers, and how to measure the incorporation of knowledge?
- How to merge latent representations with knowledge representations, and how to propagate the knowledge through the learned representation?

The Learning is aligned with stratified knowledge, and knowledge structure and abstractions are significantly retained. Specific functions and how they can be operationalized.

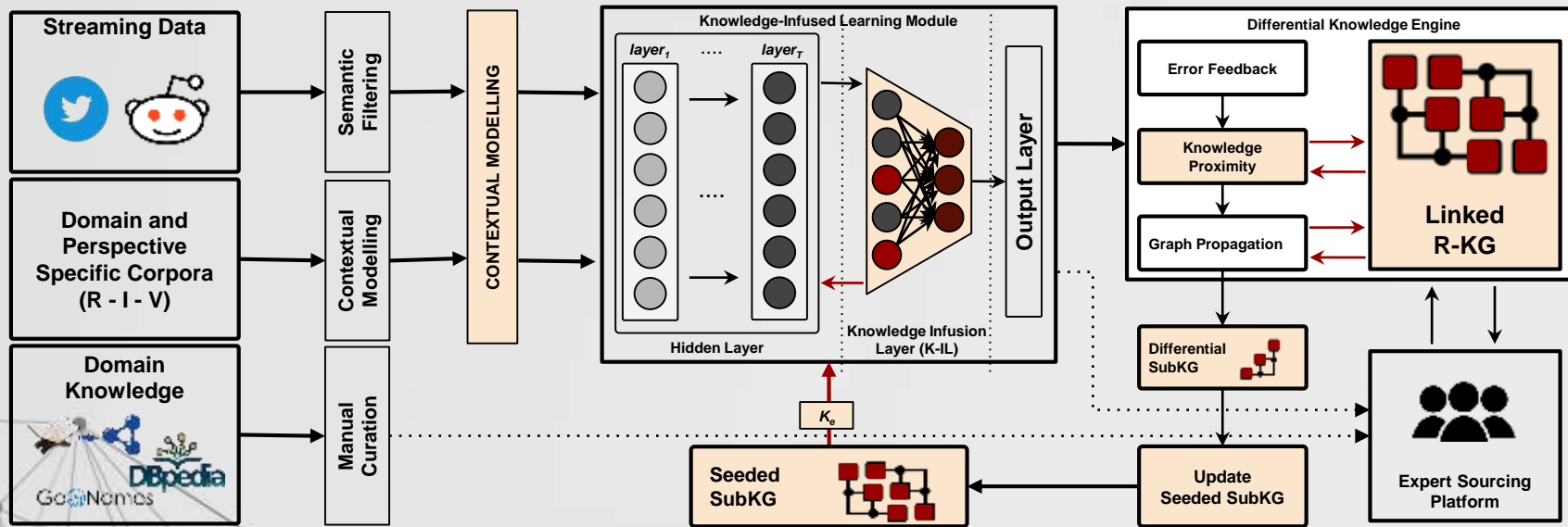
- 
- Knowledge-Aware Loss Function
  - Knowledge Modulation Function

# Knowledge-Infused Learning Layer

- Optimize the loss function as the difference between the latent representations ( $h_{T-1}$   $h_T$ ) of the last cell is reduced with respect to the knowledge representation, e.g., through KL divergence.
- Compute **differential knowledge** through such optimization approach; determining the degree of knowledge to be infused.
- Modulate the learned weight matrix of the neural network with the hidden vector through an appropriate modulation functions. (e.g., Hadamard pointwise multiplication)



# Broader Perspective: Knowledge-Infused Learning







# FUTURE

---

Computing for Human Experience

*My vision for semantic technologies is for it to find a place in future more powerful AI resulting from a synergy between the **top-down processing** (symbolic AI) and **bottom-up processing** (statistical AI) in human brains.*

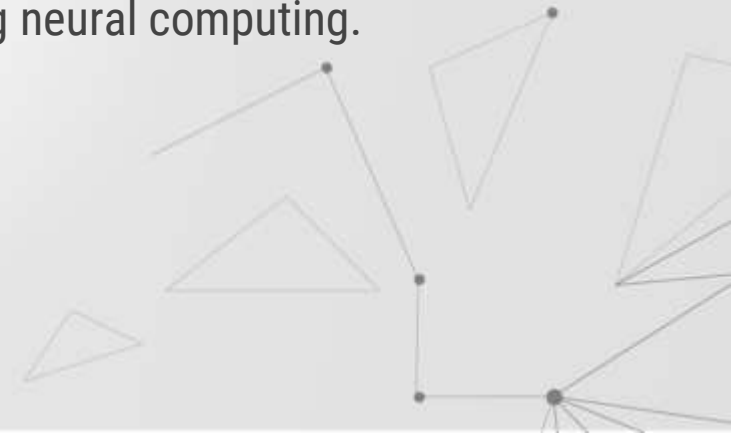



---

# Knowledge Graphs

will play an increasing role in developing hybrid neuro-symbolic systems (that is bottom-up deep learning with top-down symbolic computing) as well as in building explainable AI systems for which KGs will provide scaffolding for punctuating neural computing.

---





# KG: The Glue in Developing Hybrid AI Systems

## SYMBOLIC AI


FORMAL

*"Unreasonable effectiveness of small data"*  
in human decision making - can this be  
emulated to power top down processing?

## STATISTICAL AI

CONNECTIONIST

*"Unreasonable effectiveness of big data"*  
in machine processing &  
powering bottom up processing



# Explainability, Traceability, and Interpretability in AI

A hybrid approach integrating top-down with bottom up

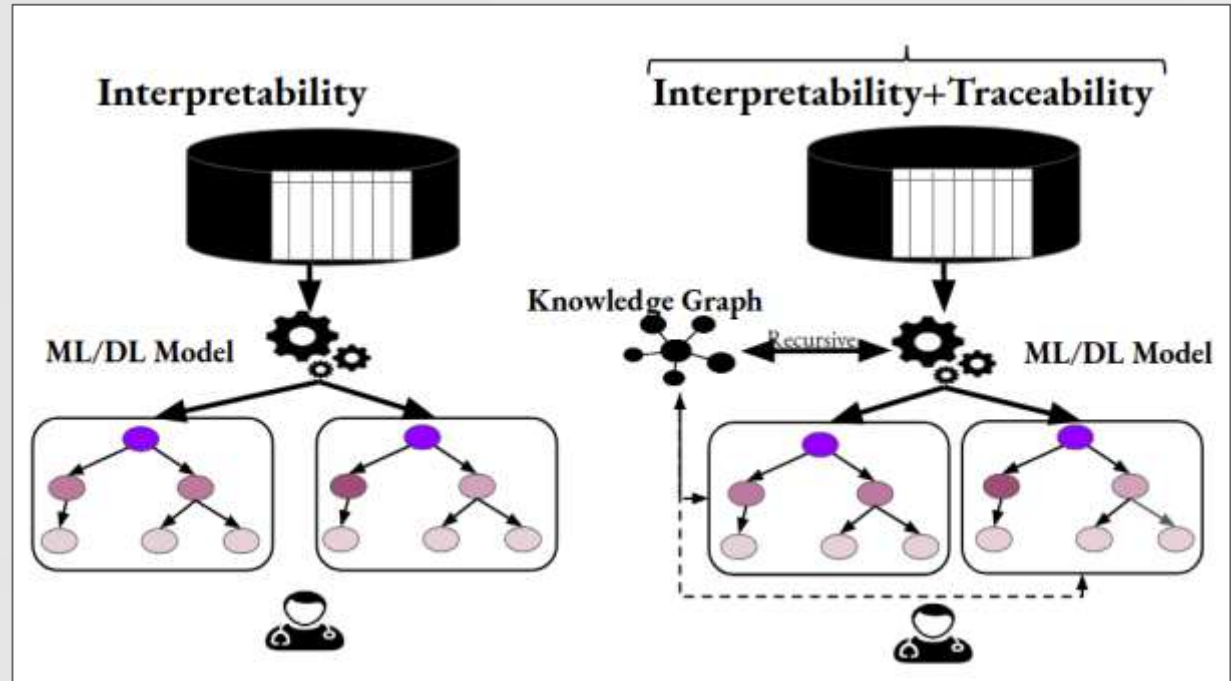


Figure: KG-infused Machine/ Deep Learning

# PROMISING KG IMPACTS

## ROBOTICS

### Cross-domain Knowledge

- 1) Observational (sensory data) and common-sense knowledge to perceive the surrounding environment
- 2) Knowledge representation to model the knowledge concerning the surrounding environment
- 3) Appropriate cross-domain knowledge reasoning mechanisms

## COGNITIVE SCIENCE

### Human Intelligence

“Inject” human intelligence into AI assistants such as Amazon Alexa, utilization of cross-domain knowledge of social interactions, emotions and linguistic variations of natural language.

## SELF-DRIVING CARS

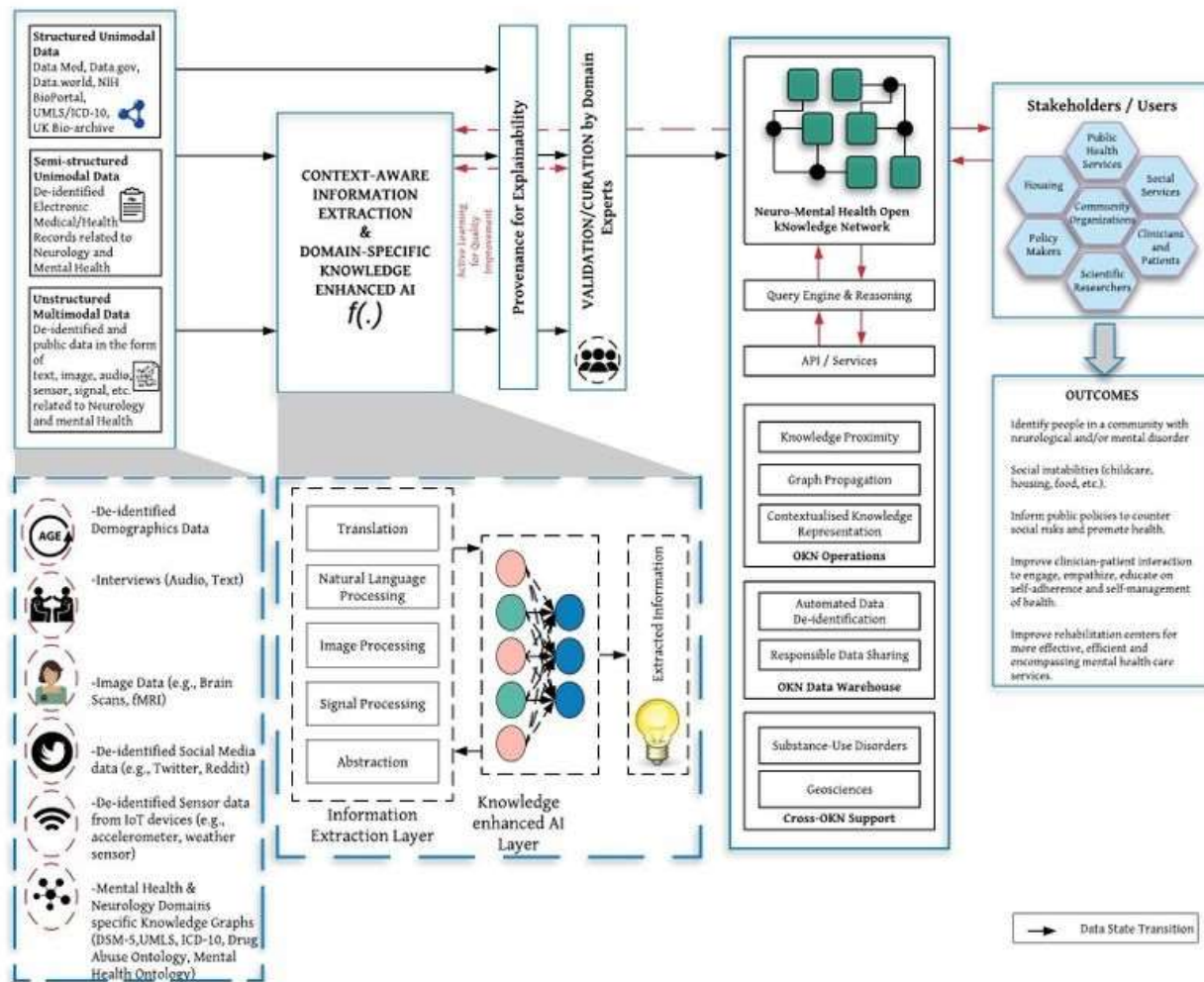
### Empathy and Morality

AI agents to mimic human emotions and decisions, we need to model human emotional knowledge of empathy, moral, and ethics.

## PERSONAL ASSISTANT

### Personalization

Smart **health** agents are adapting to answer real-world personalized complex health queries in simple interactive language. Requires patients' environmental knowledge, health data, and coordination with their healthcare physicians.



**Figure:** Illustrative example of OKN for neuro-mental domain.



**CAPTURING  
CONTEXT**

**01**

**DOMAIN-SPECIFIC  
KNOWLEDGE  
EXTRACTION**

**02**

**KNOWLEDGE  
ALIGNMENT**

**03**

## **CHALLENGES For KG**

**04**


**REAL-TIME KG FOR  
FAST DATA**

**05**

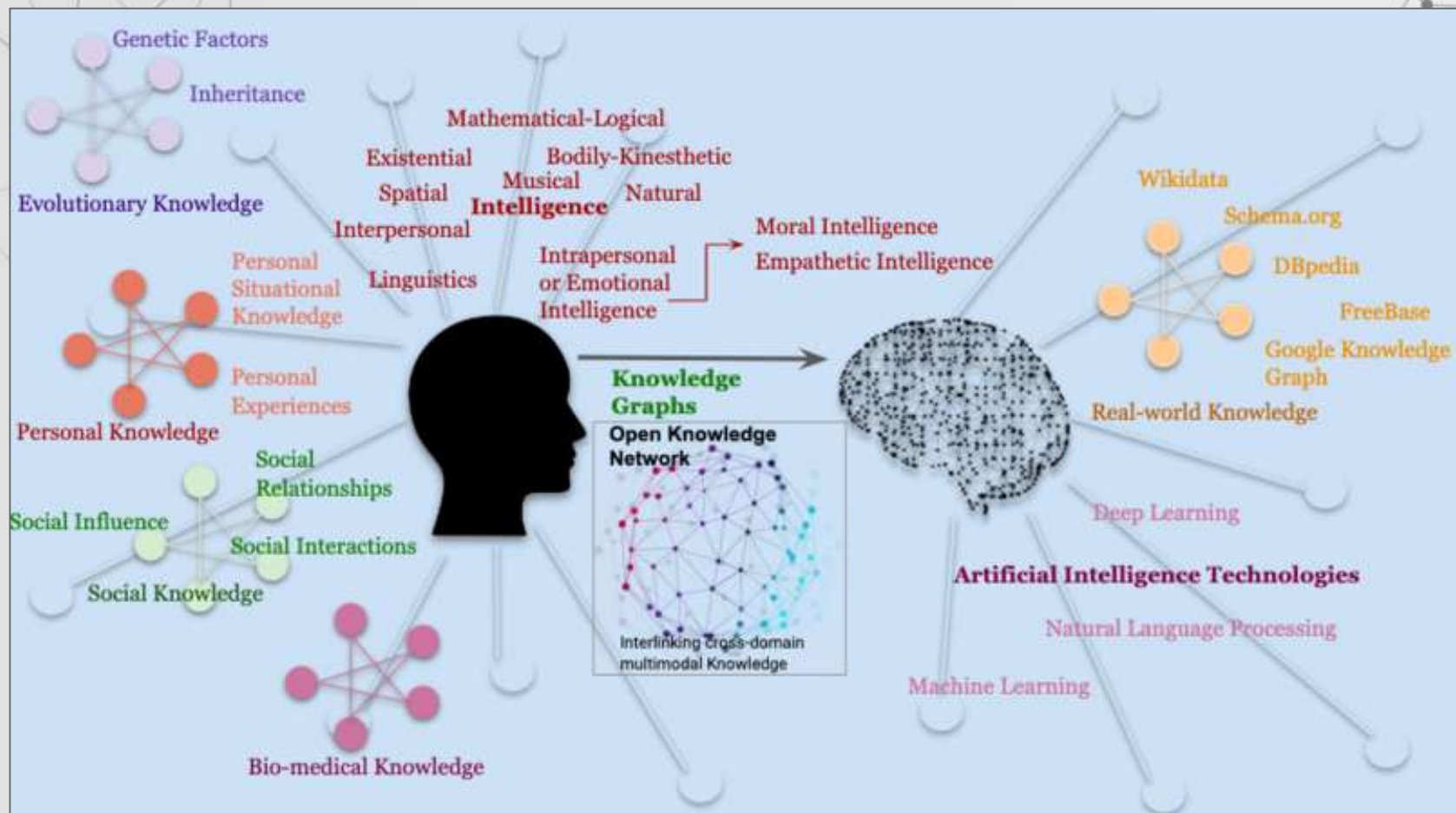
**QUALITY &  
VALIDITY OF KGs**

**06**

**ADAPTIVE  
KNOWLEDGE  
NETWORK**







**Figure:** An expanding role of Knowledge in future AI systems.




# Conclusion & Take away

**“Data alone is not enough”:**

<https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>

Consider **combining data-centric**/bottom up/statistical learning **with knowledge-based**/top down techniques

- To improve understanding of simpler content
  - To understand complex content and concepts
  - To understand heterogeneous/multimodal content
  - and a lot more
- 



South Carolina

*Thank You!*

[ai.sc.edu](http://ai.sc.edu)

# References

- Valiant, Leslie G. "Robust logics." *Artificial Intelligence* 117.2 (2000): 231-253.
- Valiant, Leslie. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.
- Valiant, Leslie G. "A neuroidal architecture for cognitive computation." *Journal of the ACM (JACM)* 47.5 (2000): 854-882.
- Davis, Ernest, and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence." *Commun. ACM* 58.9 (2015): 92-103.
- Xu, Keyulu, et al. "What Can Neural Networks Reason About?." *arXiv preprint arXiv:1905.13211* (2019).
- Marcus, Gary. "Deep learning: A critical appraisal." *arXiv preprint arXiv:1801.00631* (2018).