

Trailblazing, Complex Hypothesis Evaluation, Abductive Reasoning and Semantic Web

- exploring possible synergy

ARO WORKSHOP ON ABDUCTIVE REASONING, REASONING, EVIDENCE AND INTELLIGENT
SYSTEMS, AUGUST 23-24, 2007

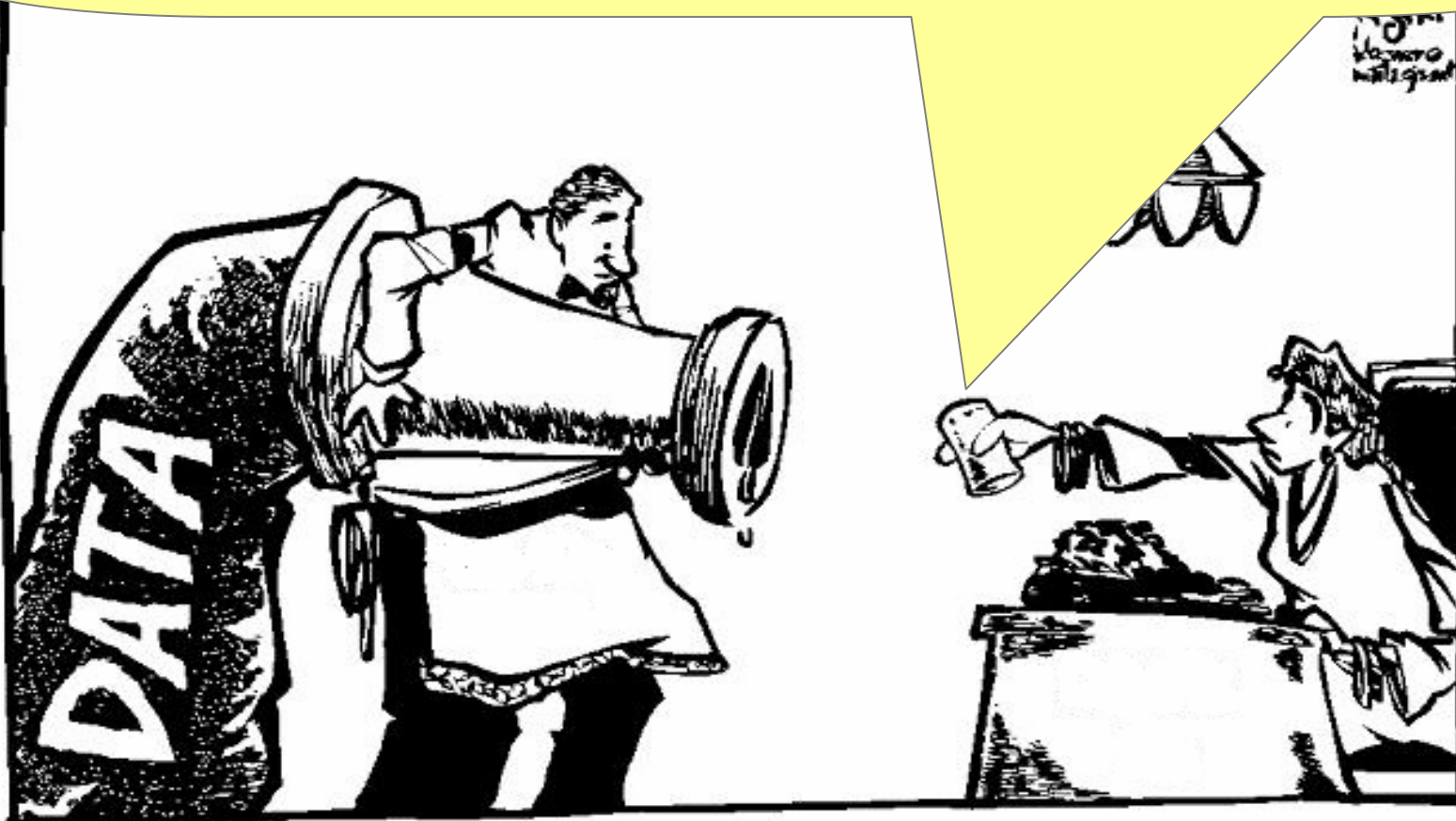
Amit Sheth

Kno.e.sis Center

Wright State University, Dayton, OH

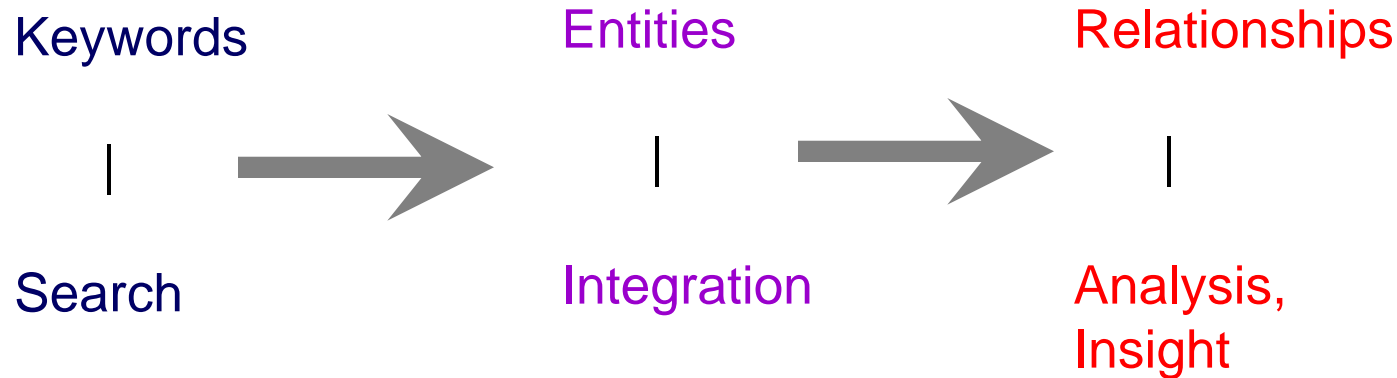
Thanks to the Kno.e.sis team,
esp. Cartic Ramakrishanan and Matt Perry.

Not data (search), but **integration**, analysis and **insight**, leading to **decisions** and **discovery**



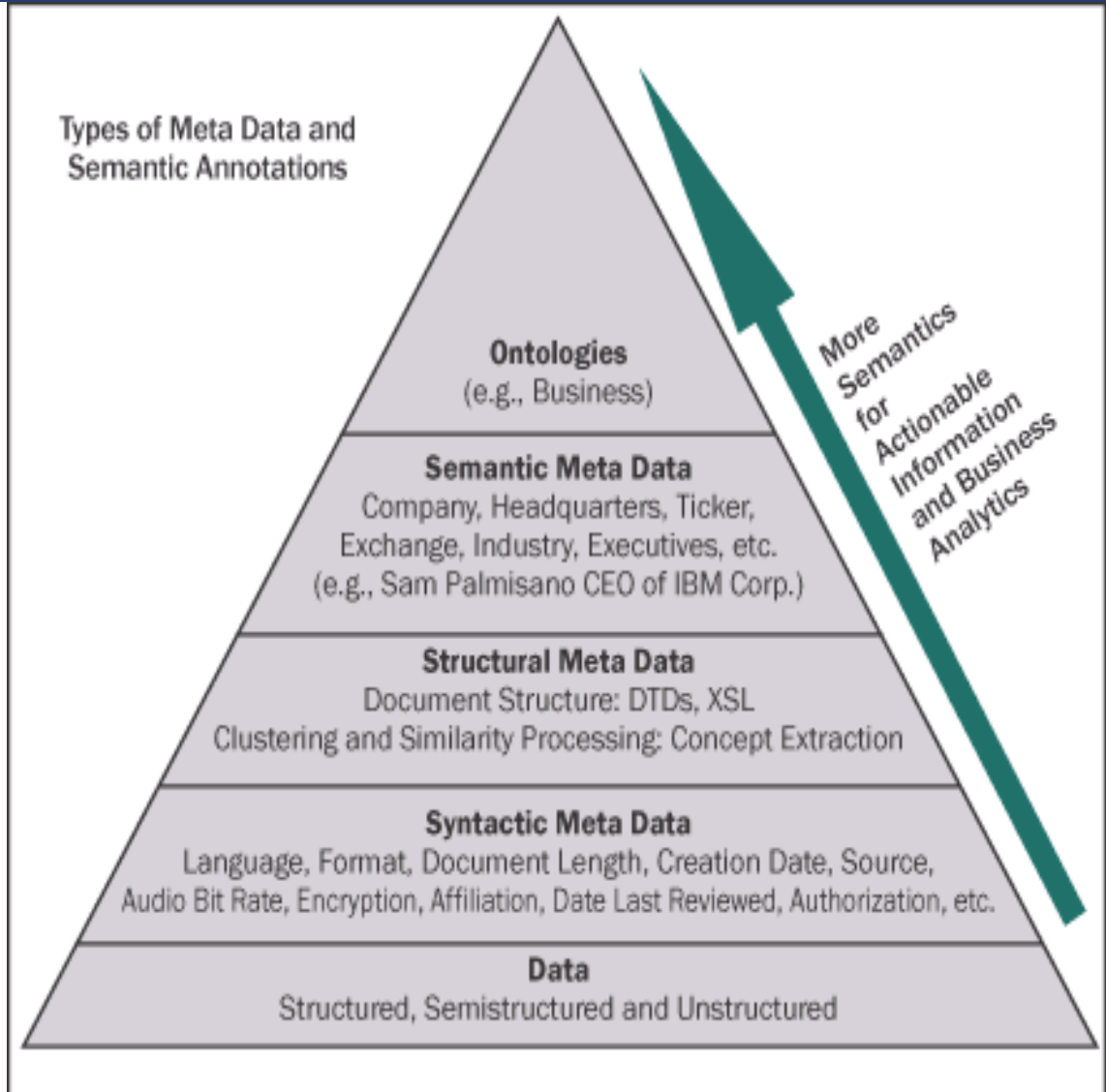
“An object by itself is intensely uninteresting”.

Grady Booch, Object Oriented Design with Applications, 1991



Changing the paradigm from document centric to relationship centric view of information.

Moving from Syntax/Structure to Semantics



Semantics: Meaning & Use of Data

Semantic Web: Labeling data on the Web so both humans and machines can use them more effectively

i.e., Formal, machine processable description \Rightarrow more automation;
emerging standards/technologies

(RDF, OWL, Rules, ...)

How?

Ontology: Agreement with Common Vocabulary & Domain Knowledge

Semantic Annotation: metadata (manual & automatic metadata extraction)

Reasoning: semantics enabled search, integration, analysis, mining, discovery

Time, Space

Gene Ontology, Glycomics, Proteomics

Pharma Drug, Treatment-Diagnosis

Repertoire Management

Equity Markets

Anti-money Laundering, Financial Risk, Terrorism

Biomedicine is one of the most popular domains in which lots of ontologies have been developed and are in use. See:

<http://obo.sourceforge.net/browse.html>

Clinical/medical domain is also a popular domain for ontology development and applications: <http://www.openclinical.org/ontologies.html>

Creation of Metadata/Annotations

Automatic Semantic Metadata Extraction/Annotation – Entity Extraction

Blue-chip bonanza continues

company company company
 Dow above 9,000 as [HP](#), [Home Depot](#) lead advance; [Microsoft](#) upgrade helps techs.
date time
 August 22, 2002: 11:44 AM EDT
phrase phrase
 By Alexandra Twin, CNN/Money Staff Writer
city company
[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 weekday Thursday, with the financial index [Dow Jones industrial average](#) spending time above the 9,000 level.
time financial index
 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 date July 23. The stock exchange [Nasdaq](#) composite gained 9.12 to 1,418.37.
financial index
[The Standard & Poor's 500 index](#) rose 9.61 to 958.97.
company stockSym \$ \$
[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 continent [Europe](#), the region [Middle East](#) and continent [Africa](#).
company stockSym \$ \$
[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 category company
 Tech stocks managed a turnaround. [Software](#) continued to rise after [Salomon Smith Barney](#) upgraded No. 1 software maker company stockSym \$ \$ [Microsoft](#) ([MSFT](#): up \$0.55 to \$52.83, Research, Estimates) to "outperform" from "neutral" and raised its price target to \$56. Business software makers company stockSym \$ \$ [Oracle](#) ([ORCL](#): up \$0.18 to \$10.94, Research, Estimates), company stockSym \$ \$ [PeopleSoft](#) ([PSFT](#): up \$1.17 to \$20.67, Research, Estimates) and company stockSym \$ \$ [BEA Systems](#) ([BEAS](#): up \$0.28 to \$7.12, Research, Estimates) all rose in tandem.
 competes with



[Hammond et al 2002]

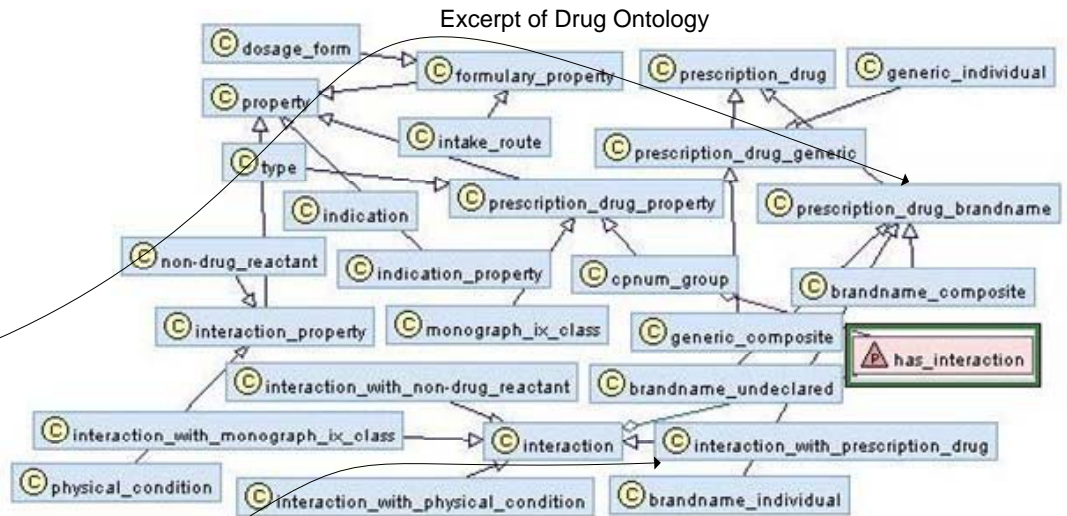
Semantic Annotation – Elsevier’s health care content

<Entity id="494805" class="DrugOntology#Organization">
 Public Health Advisory FDA Announces Important Change
 <Entity id="492805" class="DrugOntology#prescription_drug">
 Selective and Non-Selective Non-Steroidal Anti-Inflammatory
 <Entity id="122805" class="DrugOntology#prescription_drug_brandname">NSAID

Today, the Food and Drug Administration (FDA) is announcing
 <Entity id="474305" class="DrugOntology#Organization">
 to voluntarily withdraw <Entity id="122805" class="DrugOntology#prescription_drug_brandname">Bextra
 </Entity> (<Entity id="10288" class="DrugOntology#prescription_drug">valdecoxib</Entity>) from the market. Pfizer has agreed to
 of Bextra in the <Entity id="7852" class="Sweto#country">
 further discussions with the agency.

.....
 This request is based on:
 -Reports of serious and potentially life-threatening skin reactions, including deaths,
 in patients using Bextra. The risk of these reactions in individual patients is unpredictable,
 occurring in patients with and without a prior history of <Entity id="14280" class="DrugOntology#interaction_with_physical_condition">sulfa allergy
 </Entity>

.....
 Date created: <Regex type="date">April 7, 2005</Regex>



```

Sample Created Metadata

<Entity id="122805"
class="DrugOntology#prescription_drug_brandname">
Bextra
<Relationship id="442134"
class="DrugOntology#has_interaction">
<Entity id="14280" class="DrugOntology
#interaction_with_physical_condition">sulfa allergy
</Entity>
</Relationship>
</Entity>
    
```

NCI

NCI|nci|128|1|v|1|128|1|n|0|3|

NCI|nCi's|128|8|v|1|128|1|b+i|2|3|

NCI|nCis|128|8|v|1|128|1|b+i|2|3|

NCI|National Cancer Institute|128|1|v|1|128|1|b+a|3|1|

NCI|nanocurie|128|1|v|1|128|1|b+a|3|1|

NCI|nanocuries|128|8|v|1|128|1|b+a+i|4|1|

The ambiguity could be resolved through various techniques such as co-reference resolution or evidence based matching, or modeled using probability that the term represents any of the distinct (known) entities.

Insider Threat: an example Semantic Web application that consists of (a) an ontology populated from multiple knowledge sources with heterogeneous representation formats, (b) ontology-supported entity extraction/annotation, (c) computation of semantic associations/relationships to terms in metadata with a (semantic) query represented in terms of ontology and the entities identified in the documents, (d) ranking of documents based on the strength of these semantic associations/relationships

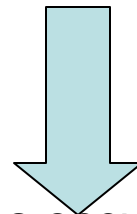
[Demo of Ontological Approach to Assessing Intelligence Analyst Need-to-Know](#)

Extracting relationships (between MeSH terms from PubMed)

- **UMLS** – A high level schema of the biomedical domain
 - 136 classes and 49 relationships
 - Synonyms of all relationship – using variant lookup (tools from NLM)
 - 49 relationship + their synonyms = ~**350** mostly verbs
- **MeSH**
 - 22,000+ topics organized as a forest of 16 trees
 - Used to query PubMed
- **PubMed**
 - Over 16 million abstract
 - Abstracts annotated with one or more MeSH terms

T147—effect
T147—induce
T147—etiology
T147—cause
T147—effecting
T147—induced

[1254239-1] An excessive endogenous or exogenous stimulation by estrogen induces adenomatous hyperplasia of the endometrium.



SS-Tagger (University of Tokyo)

SS-Parser (University of Tokyo)

- Entities (MeSH terms) in sentences occur in modified forms

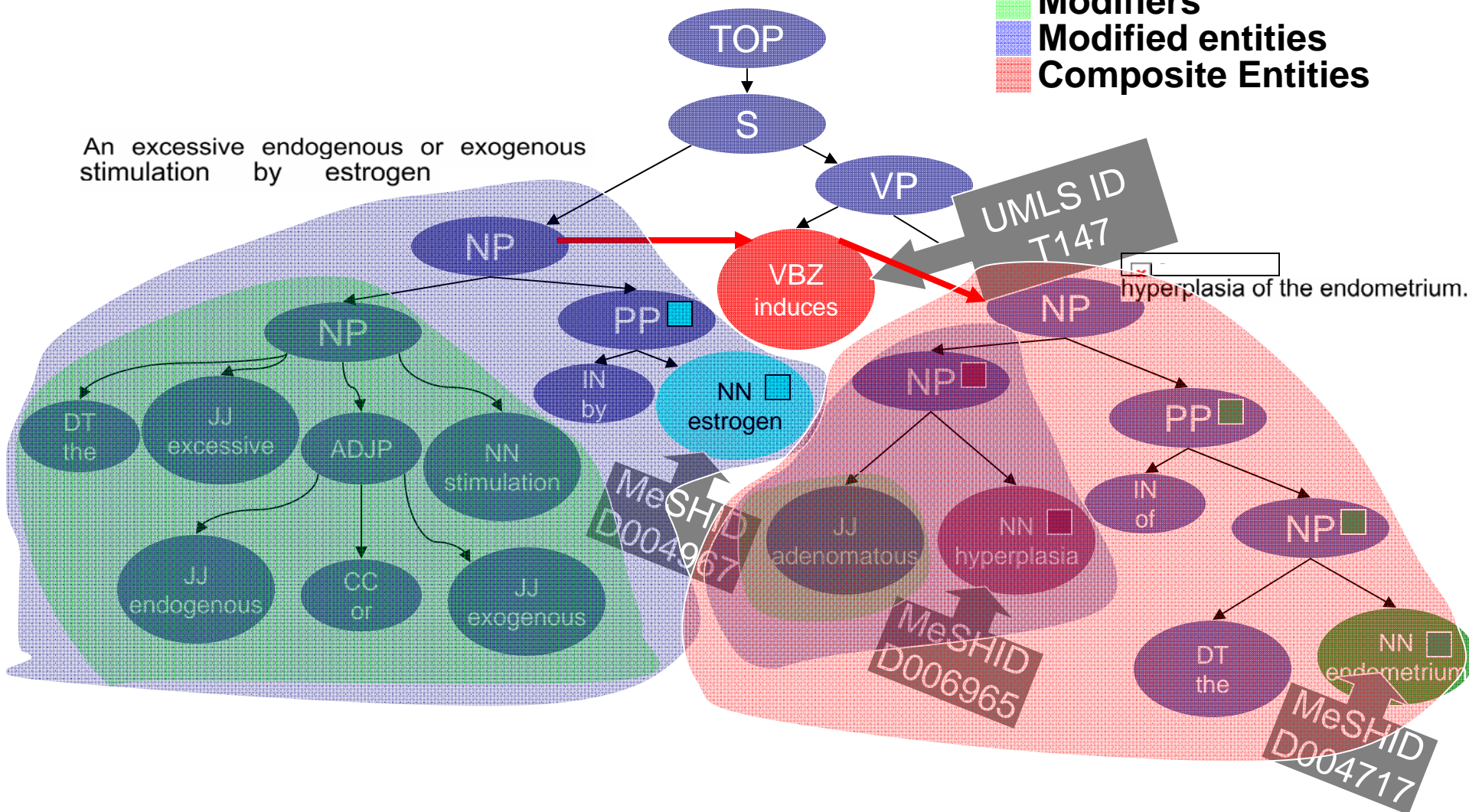
(TOP (S (NP (NP (DT "An") (JJ "excessive") (JJ "endogenous") (CC "or") (JJ "exogenous") (NN "stimulation")) (PP (IN "by") (NP (NN "estrogen"))) (VP (VBZ "induces") (NP (NP (JJ "adenomatous") (NN "hyperplasia")) (PP (IN "of") (NP (DT "the") (NN "endometrium"))))))) (PUNCT "."))

- “adenomatous hyperplasia” and “endometrium” occur as “adenomatous hyperplasia of the endometrium”

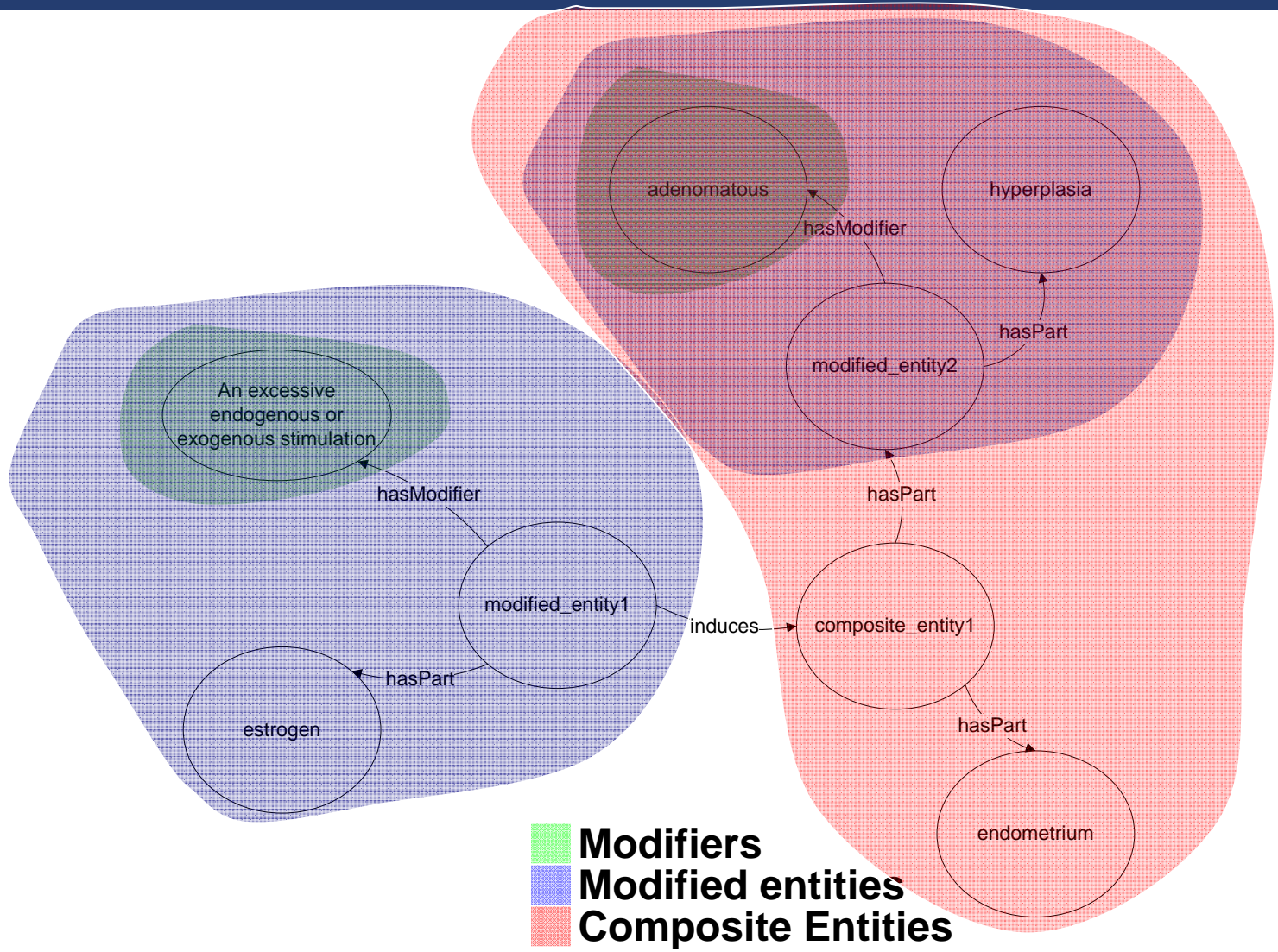
Method – Identify entities and Relationships in Parse Tree

- Modifiers**
- Modified entities**
- Composite Entities**

An excessive endogenous or exogenous stimulation by estrogen



Resulting RDF



Semantic Metadata can be extracted from unstructured (eg, biomedical literature), semi-structured (eg, some of the Web content), structured (eg, databases) data and data of various modalities (eg, sensor data, biomedical experimental data). Focusing on the relationships and the web of their interconnections over entities and facts (knowledge) implicit in data leads to a Relationship Web.

Relationship Web takes you away from “which document” could have information I need, to “what’s in the resources” that gives me the insight and knowledge I need for decision making.

[Amit P. Sheth, Cartic Ramakrishnan: Relationship Web: Blazing Semantic Trails between Web Resources. IEEE Internet Computing, July 2007.](#)

Demonstration of Semantic Trailblazing using a Semantic Browser

This application demonstrating use of ontology-supported relationship extraction (represented in RDF) and their traversal in context (as deemed relevant by the scientists), linking parts of knowledge represented in one biomedical document (currently a sentence in an abstract in Pubmed) to parts of knowledge represented in another document.

This is a prototype and lot more work remains to be done to build a robust system that can support Semantic Trailblazing. For more information:

[Cartic Ramakrishnan](#), [Krys Kochut](#), [Amit P. Sheth](#): A Framework for Schema-Driven Relationship Discovery from Unstructured Text. [International Semantic Web Conference 2006](#): 583-596 [[.pdf](#)]

[Cartic Ramakrishnan](#), [Amit P. Sheth](#): Blazing Semantic Trails in Text: Extracting Complex Relationships from Biomedical Literature. Tech. Report #TR-RS2007 [[.pdf](#)]

QUESTION 1: Given an RDF graph without weights can we use domain knowledge to compute the strength of connection between any two entities?

QUESTION 2: Can we then compute the most “relevant” connections for a given pair of entities?

QUESTION 3: How many such connections can there be? Will this lead to a combinatorial explosion? Can the notion of relevance help?

- **Problem: Discovering relevant connections between entities**
 - All Paths problem is NP-Complete
 - Most informative paths are not necessarily the shortest paths
- **Possible Solution: Heuristics-based Approach***
 - Find a smart, systematic way to weight the edges of the RDF graph so that the most important paths will have highest weight
 - Adopt algorithms for weighted graphs
 - Model graph as an electrical circuit[†] with weight representing conductance and find paths with highest current flow – i.e. top-k

* Cartic Ramakrishnan, William Milnor, Matthew Perry, Amit Sheth. "Discovering Informative Connection Subgraphs in Multi-relational Graphs", SIGKDD Explorations Special Issue on Link Mining, Volume 7, Issue 2, December 2005

† Christos Faloutsos, Kevin S. McCurley, Andrew Tomkins: Fast discovery of connection subgraphs. KDD 2004: 118-127

- What is a good path with respect to *knowledge discovery*?
 - Uses more specific classes and relationships
 - e.g. Employee vs. Assistant Professor
 - Uses rarer facts
 - Analogous to information gain
 - Involves unexpected connections
 - e.g. connects entities from different domains

- More specific classes and properties convey more information

- Specificity of property p_i :

- $d(p_i)$ is the depth of p_i
- $d(p_{iH})$ is the depth of the property hierarchy

$$\mu(p_i) = \frac{d(p_i)}{d(p_{iH})}$$

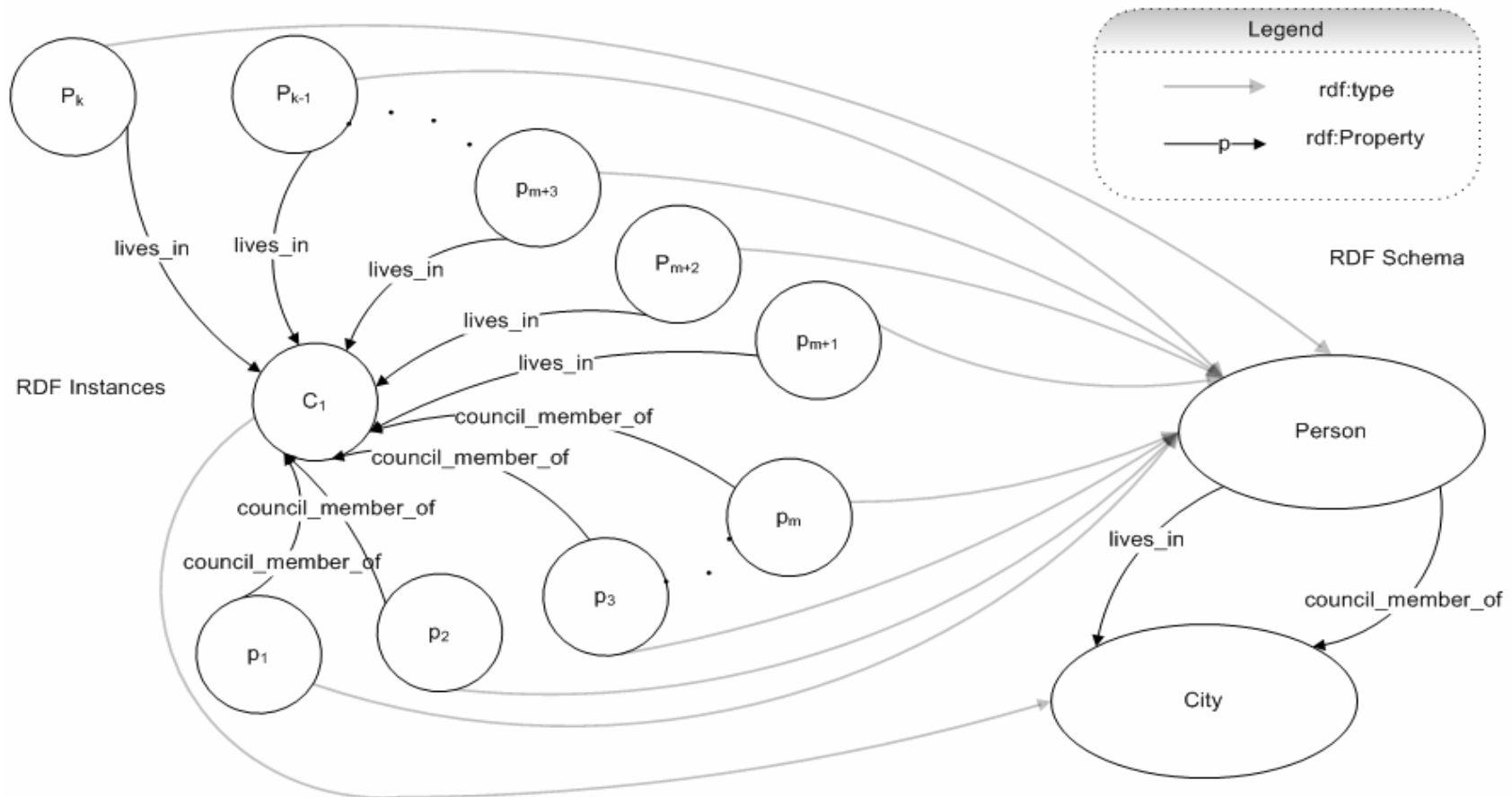
- Specificity of class c_j :

- $d(c_j)$ is the depth of c_j
- $d(c_{jH'})$ is the depth of the class hierarchy

$$\mu(c_j) = \frac{d(c_j)}{d(c_{jH'})}$$

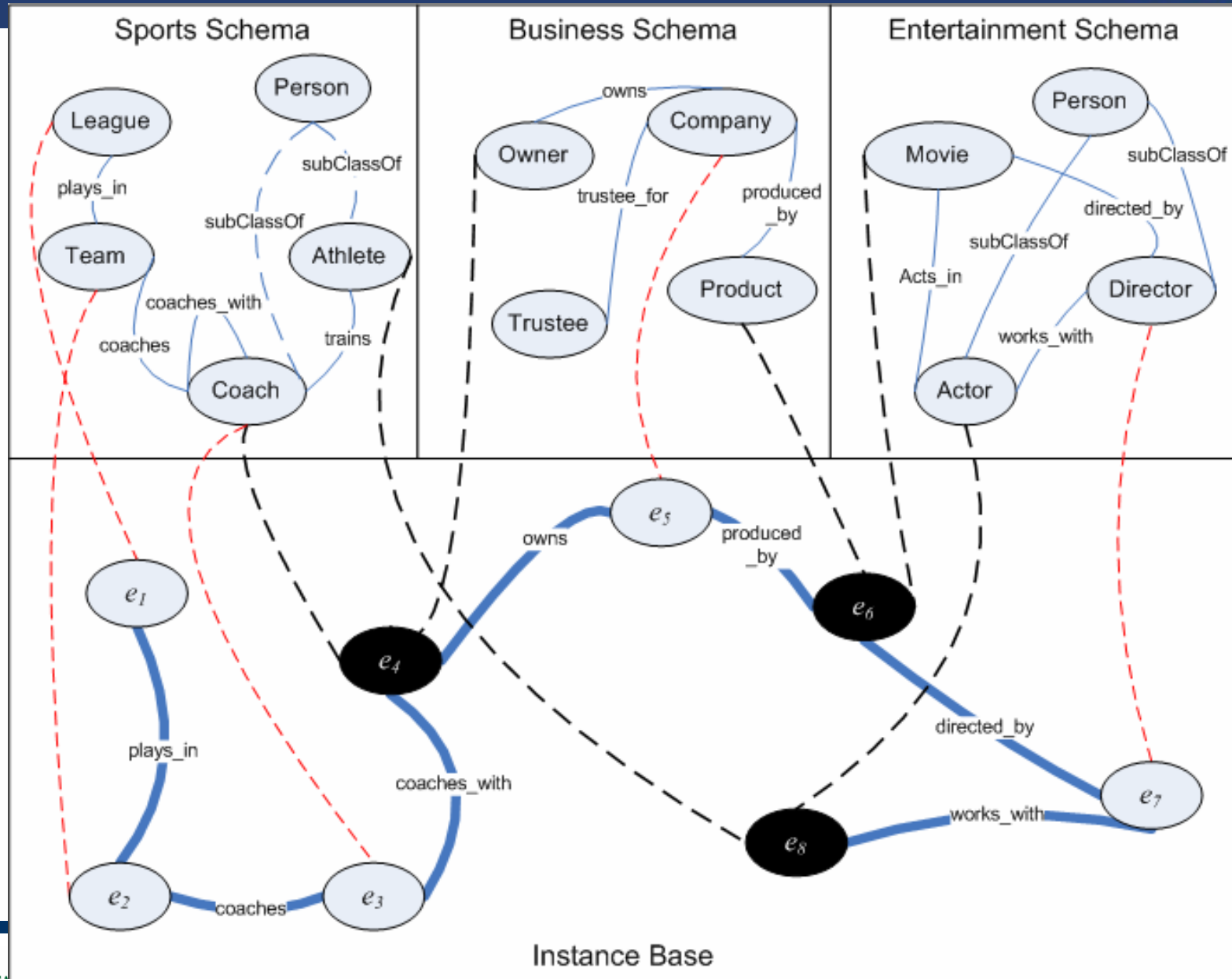
- *Node is weighted and this weight is propagated to edges incident to the node*

- Rare facts are more informative than frequent facts
- Define a *type* of an statement RDF $\langle s,p,o \rangle$
 - Triple $\pi = \langle C_i, p_j, C_k \rangle$
 - $typeOf(s) = C_i$
 - $typeOf(o) = C_k$
- $|\pi|$ = number of statements of type π in an RDF instance base
- *ISP* for a statement: $\sigma_\pi = 1/|\pi|$



- $\pi = \langle \text{Person}, \text{lives_in}, \text{City} \rangle$
- $\pi' = \langle \text{Person}, \text{council_member_of}, \text{City} \rangle$
- $\sigma_{\pi} = 1/(k-m)$ and $\sigma_{\pi'} = 1/m$, and if $k-m > m$ then $\sigma_{\pi'} > \sigma_{\pi}$

- **RDF allows Multiple classification of entities**
 - Possibly classified in different schemas
 - Tie different schemas together
- ***Refraction* is Indicative of anomalous paths**
- **SPAN favors *refracting* paths**
 - Give extra weight to multi-classified nodes and propagate it to the incident edges



- What if we are not just interested in knowledge discovery style searches?
- Can we provide a mechanism to adjust relevance measures with respect to users' needs?
 - Conventional Search vs. Discovery Search

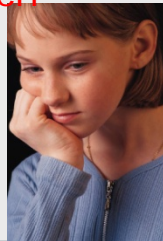
Yes! ... SemRank*

* Kemafor Anyanwu, Angela Maduko, Amit Sheth. “**SemRank: Ranking Complex Relationship Search Results on the Semantic Web**”, The 14th International World Wide Web Conference, (WWW2005), Chiba, Japan, May 10-14, 2005



SSARK SSARK *Semantic Searching of a Different Kind*

Low Information Gain
Low Refraction Count
High S-Match



Query Specification

Select Search Mode: %

Conventional Discovery

High Information Gain
High Refraction Count
High S-Match



adjustable search mode

Enter the two resources

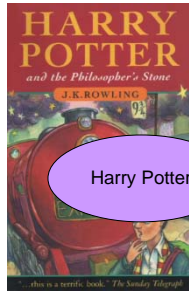
Relevance Specification (seperated with spaces)

-
-
-
-
-

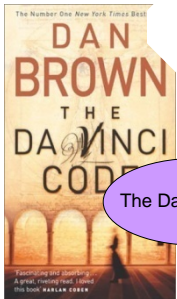
Example of Relevant Subgraph Discovery based on evidence

Anecdotal Example

UNDISCOVERED PUBLIC KNOWLEDGE
Discovering connections hidden in text



Harry Potter



The Da Vinci code



Et in Arcadia Ego



Santa Maria delle Grazie

Priority of Sion - Wikipedia, the free encyclopedia - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

W http://en.wikipedia.org/wiki/Priority_of_Sion

Customize Links Free Hotmail My Yahoo! Windows Marketplace Windows Media Windows Yahoo! Bookmarks Yahoo! Mail Distributed data stru... Yahoo!

Search MSD Text Search for OR Get PDB:

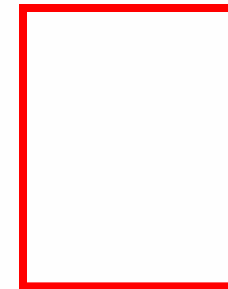
The Priory of Sion was supposedly led by a [Grand Master](#) or [Nautonnier](#).

1. [Ugo de Blancheford \(1150-1151\)](#)
2. [Bernard de Tremblay \(1151-1153\)](#)
3. [Guillaume de Chanailleilles \(1153-1154\)](#)
4. [Evrard de N...? \(1154-1154\)](#)
5. [Andr  de Montbard \(1155-1156\)](#)
6. [Bertand de Blancheford \(1156-1169\)](#)
7. [Philippe de Milly \(1169-1170\)](#)
8. [Eudes de Saint-Amand \(1170-1180\)](#)
9. [Arnaud de Toroge \(1181-1184\)](#)
10. [G rard de Rideford \(1184-1188\)](#)
11. [Jean de Gisors \(1188-1220\)](#)
12. [Marie de Saint-Clair \(1220-1266\)](#)
13. [Guillaume de Gisors \(1266-1307\)](#)
14. [Edouard de Bar \(1307-1336\)](#)
15. [Jeanne de Bar \(1336-1351\)](#)
16. [Jean de Saint-Clair \(1351-1366\)](#)
17. [Blanche d'Evreux \(1366-1398\)](#)
18. [Nicolas Flamej \(1398-1418\)](#)
19. [Rene d'Anjou \(1418-1480\)](#)
20. [Iolande de Bar \(1480-1483\)](#)
21. [Sandro Filioepi AKA Botticelli \(1483-1510\)](#)
22. [Leonardo da Vinci \(1510-1519\)](#)
23. [Charles III \(Duke of Bourbon-Montpensier\) \(1519-1527\)](#)
24. [Ferdinand de Gonzague \(1527-1556\)](#)
25. [Michel de Notre-Dame AKA Nostradamus\(1556-1566\)](#)
26. [Duc de Longueville & Nicolas Froumenteau \(1566-1575\)](#)
27. [Louis de Nevers \(1575-1595\)](#)
28. [Robert Fludd \(1595-1637\)](#)
29. [Johann Valentin Andrea \(1637-1654\)](#)
30. [Robert Boyle \(1654-1691\)](#)
31. [Isaac Newton \(1691-1727\)](#)
32. [Charles Radclyffe \(1727-1746\)](#)
33. [Charles de Lorraine \(1746-1780\)](#)
34. [Maximillian de Lorraine \(1780-1801\)](#)
35. [Charles Nodier \(1801-1844\)](#)
36. [Victor Hugo \(1844-1885\)](#)
37. [Claude Debussy \(1885-1918\)](#)
38. [Jean Cocteau \(1918-1963\)](#)
39. [Pierre Plantard \(1963-1989\)](#)

A second List of the Grand Masters of the Priory of Sion that included the names of Roger Patrice Pelat and Thomas Plantard appeared in 1989. but it should not be confused with the

Done

start Discovering Informati... KnowledgeDiscovery... Drawing1 - Microsoft ... Yahoo! Messenger S Firefox 1:45 PM

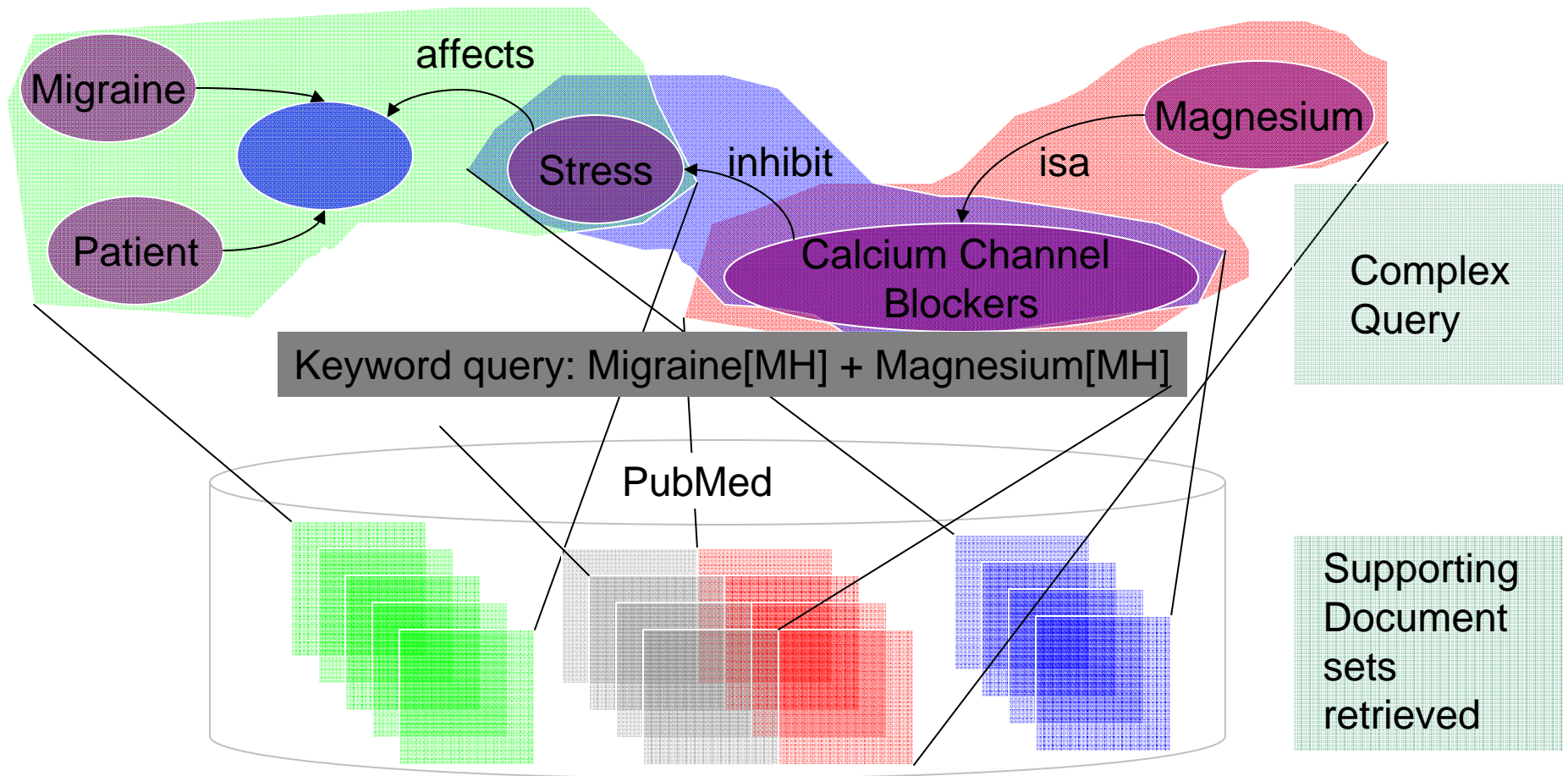


d_at



Ontology supported text retrieval and hypothesis validation

Complex Hypothesis Evaluation over Scientific Literature



We discuss some scenarios tying evidence based reasoning and the need to add representations and reasoning that involve approximate information in the context of current research in Semantic Web

Knowledge enable Information & Services Science Center:
<http://knoesis.wright.edu>

