

# Research in Discrete Sciences

Eduard Hovy

USC/ISI

[hovy@isi.edu](mailto:hovy@isi.edu)

Representing the 4 IDS-UAC Center Directors

# Institute for Discrete Sciences

**IDS focuses on selected topics in computer science, data sciences, machine learning, and discrete mathematics, collectively labeled “discrete sciences”**

RUTGERS

USC  
INFORMATION  
SCIENCES  
INSTITUTE



ILLINOIS

**DyDAn:** The Center for Dynamic Data Analysis



**CKID:** The Center for Knowledge Integration and Discovery



**CERATOPS:** Center for Extraction and Summarization of Events and Opinions in Text



**MIAS:** Multimodal Information Access and Synthesis



# IDS partner network

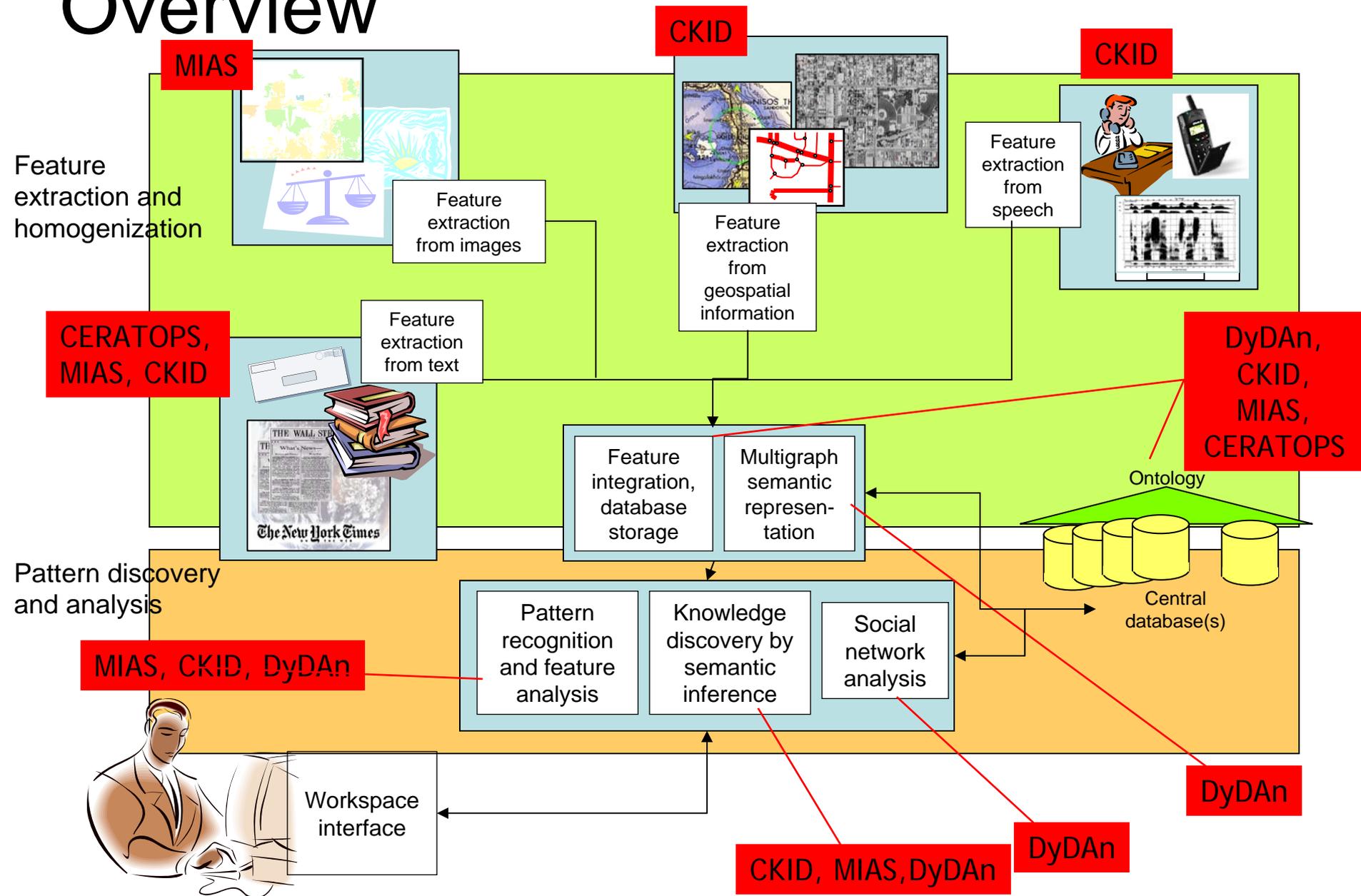


RUTGERS



CORNELL

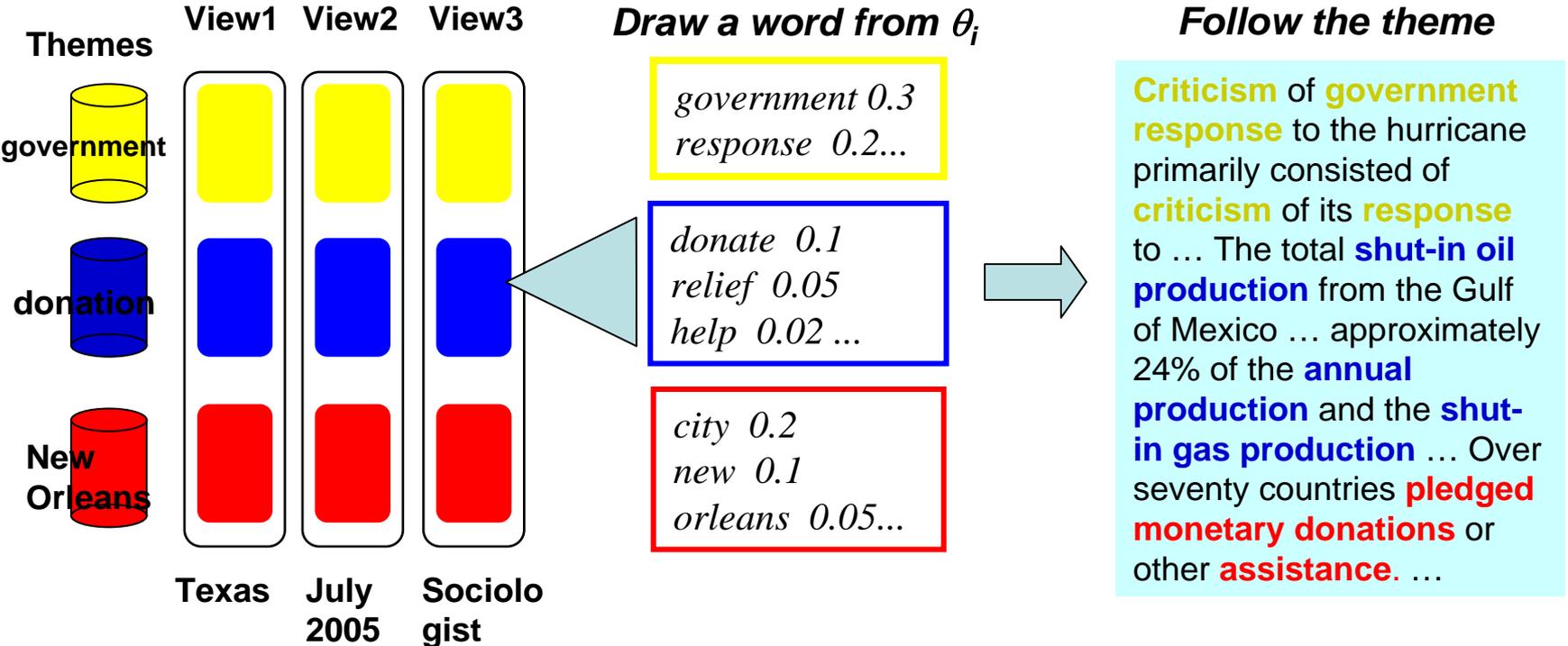
# Overview



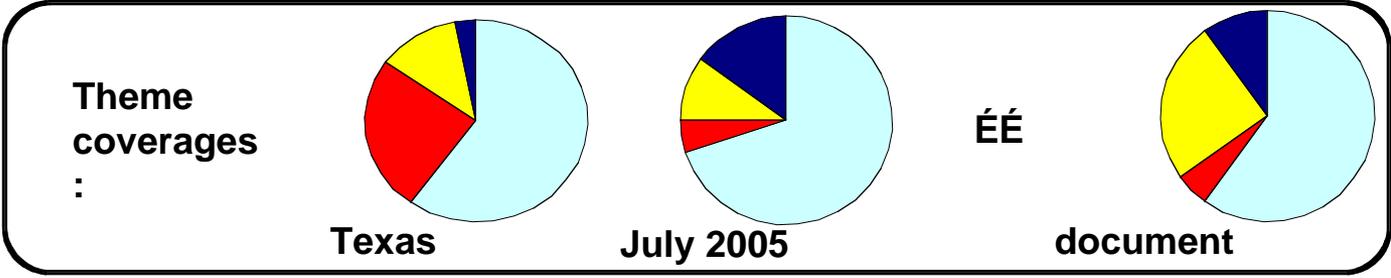
# Theme 1: Contextual text analysis

- Useful info: Docs often in context (meta-data)
  - Direct context: time, location, source, authors...
  - Indirect context: events, policies...
- Many applications require “contextual text analysis”:
  - Discovering topics from text in a context-sensitive way
  - Analyzing variations of topics over different contexts
  - Revealing interesting patterns (e.g., topic evolution, topic variations, topic communities)
  - Learning how they change over time
- **Can we develop a single unified way to handle all this?**
- Researchers:
  - MIAS: ChengXiang Zhai and Dan Roth (U of Illinois, Urbana-Champaign)
  - CERATOPS: Jan Wiebe (U of Pittsburgh), Claire Cardie (Cornell U), and Ellen Riloff (U of Utah)
  - CKID: Eduard Hovy and Patrick Pantel (USC/ISI)

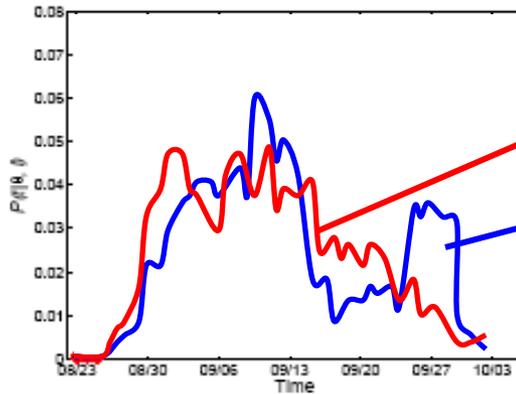
# Example technique: Contextual probabilistic Latent Semantics Analysis



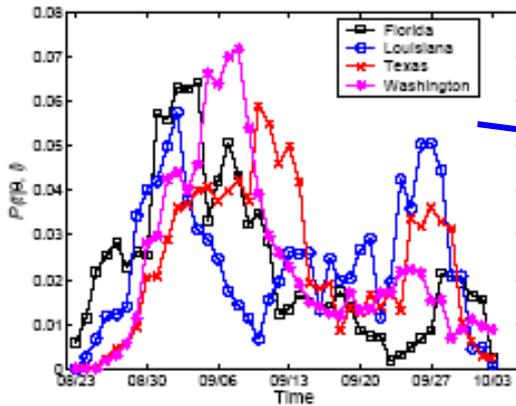
## Choose a view



# Topic lifecycles (“Hurricane Katrina”)



(a) Theme life cycles in Texas (Hurricane Katrina)



(b) Theme “New Orleans” over states (Hurricane Katrina)

*Oil Price*

*New Orleans*

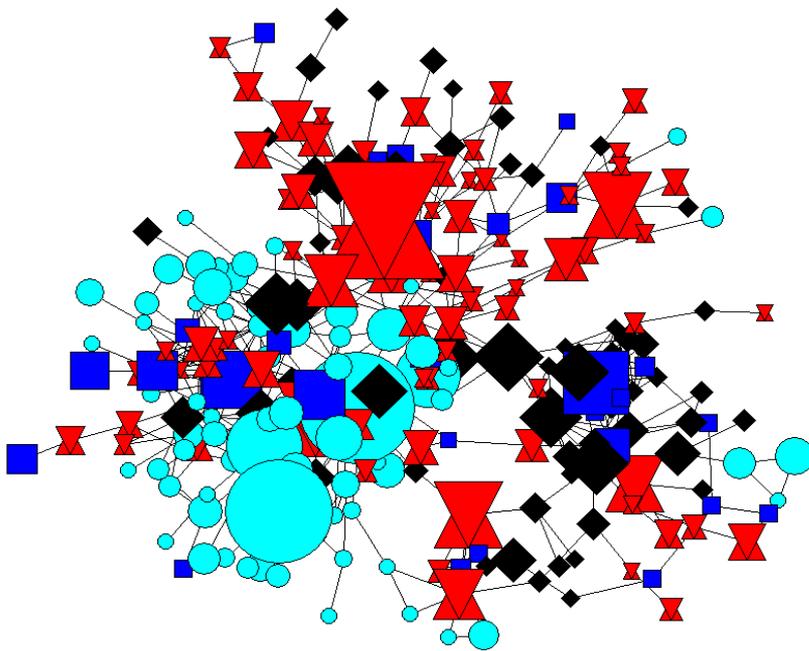
*price 0.0772*  
*oil 0.0643*  
*gas 0.0454*  
*increase 0.0210*  
*product 0.0203*  
*fuel 0.0188*  
*company 0.0182*  
 ...

*city 0.0634*  
*orleans 0.0541*  
*new 0.0342*  
*louisiana 0.0235*  
*flood 0.0227*  
*evacuate 0.0211*  
*storm 0.0177*  
 ...

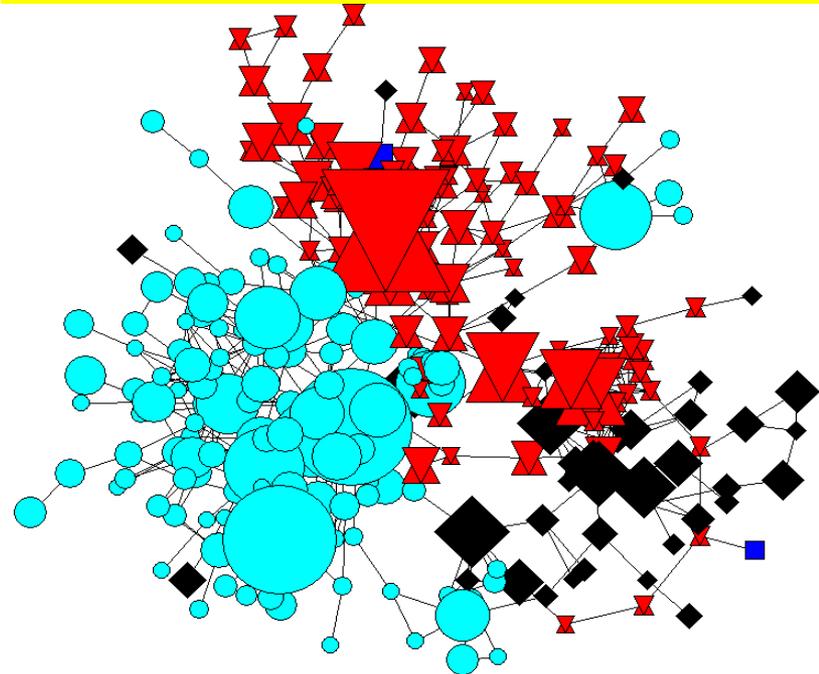
# Topic modeling + Social Networks

- Authors writing about the same topic form a community
- Separation of 3 communities:

Topic Model Only

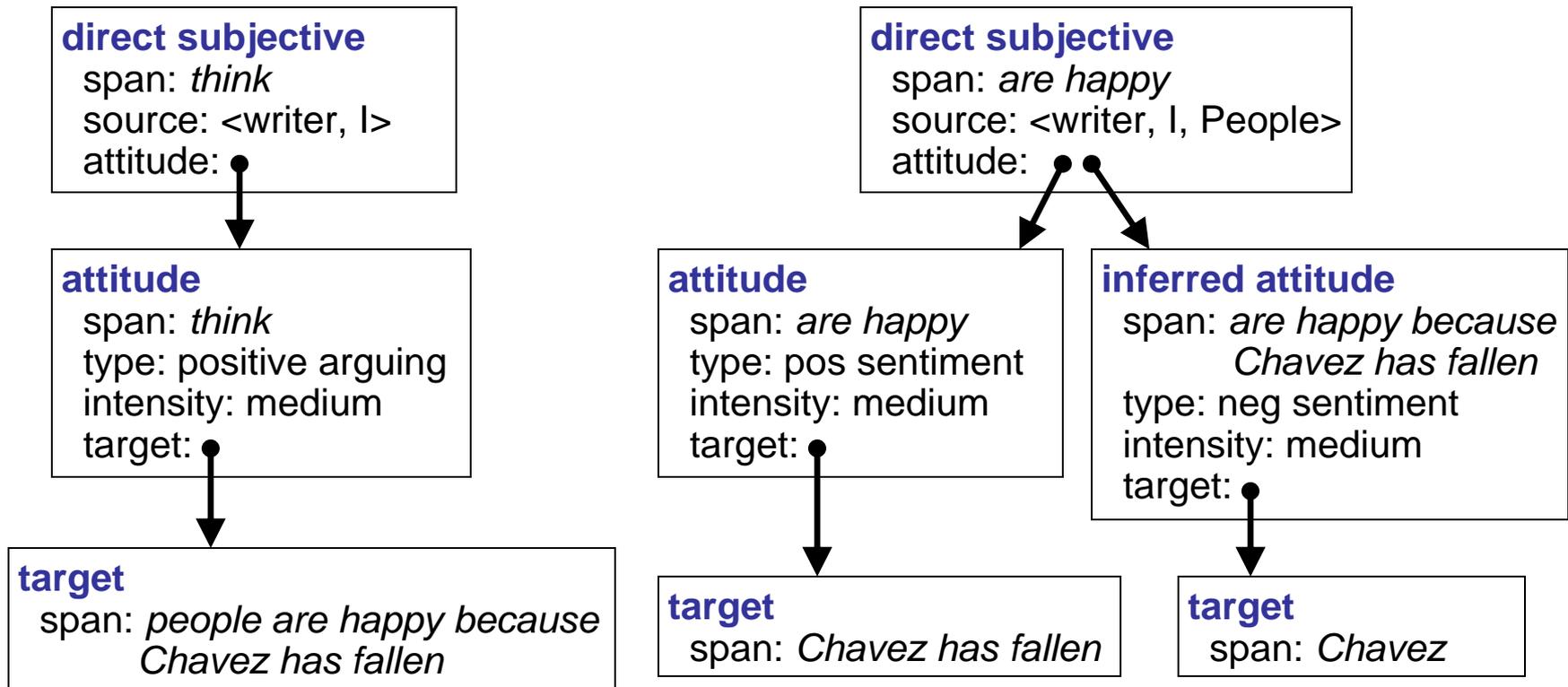


Topic Model + Social Network



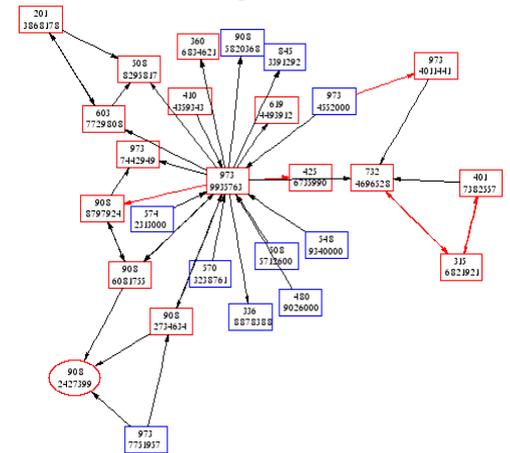
# Example: Judgment opinions

I think people are happy because Chavez has fallen.



# Theme 2: Multigraph analysis

- Need to understand massive amounts of data:
  - Data takes numerous forms; requires data mining methods that span the modalities
  - Data inherently distributed from multiple sources
  - Data arrives rapidly and continuously
- Seek anomalies, patterns, emerging events
  - Continuously monitor incoming data stream
- Multigraphs represent source information, cross-linked multiple times by various relations
- Researchers:
  - DyDAn: Fred Rogers et al. (Rutgers, Princeton, AT&T)
  - MIAS: Dan Roth (U of Illinois, Urbana-Champaign)
  - CKID: Hans Chalupsky (USC/ISI)



# Work on large multigraphs

- Information of interest to DHS is often stored using “shallow” (but still useful) representations
  - Much information is English tags: susceptible to ambiguity, incompleteness, etc.
- Alleviate by augmenting with rich ontologies that describe and prescribe how a domain works
  - Can discover information inherent in shallow information
  - Can expose inconsistencies in shallow information
- Problem: Reasoning with rich information is difficult and computationally expensive. So, research on:
  - Extending OWL Web Ontology Language, a powerful ontology language for use with shallow information
  - Extending and specializing theory of Distributed Description Logics (DDLs), designed to limit interactions to lessen computational load
  - Developing and extending a highly-optimized reasoner to improve its performance with large amounts

# IDS educational programs

- Sampling of events:
  - Summer 2007, 2008: 8-week school (UIUC + teachers from other centers)
  - Summer 2008: 1-week short course on Biosurveillance (DyDAn)
  - November 2007: Tutorial: Balancing Data Confidentiality and Data Quality (DyDAn)
  - 2008: Scholar program and summer interns: (DyDAn, USC, UIUC)
  - February 2008: Workshop: Privacy Preserving Data Analysis (DyDAn)



**Open to All**

# Please contact us!

Discrete science is a “cross-cutting” capability that can provide support to each of the DHS divisions

**DyDAn:** The Center for Dynamic Data Analysis

Fred Roberts (froberts@dimacs.rutgers.edu)



**CKID:** The Center for Knowledge Integration and Discovery

Eduard Hovy (hovy@isi.edu)



**MIAS:** Multimodal Information Access and Synthesis

Dan Roth (danr@cs.uiuc.edu)



**MIAS**  
MULTIMODAL  
INFORMATION ACCESS  
AND SYNTHESIS

**CERATOPS:** Center for Extraction and Summarization of Events and Opinions in Text

Jan Wiebe (wiebe@cs.pitt.edu)

