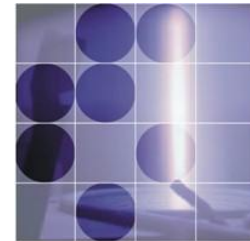


# Text Mining & Light Weight Semantics

Marko Grobelnik ([marko.grobelnik@ijs.si](mailto:marko.grobelnik@ijs.si))

Jozef Stefan Institute (<http://www.ijs.si/>)

Slovenia, Europe



QUINTELLIGENCE



cYcorp



STI · INTERNATIONAL

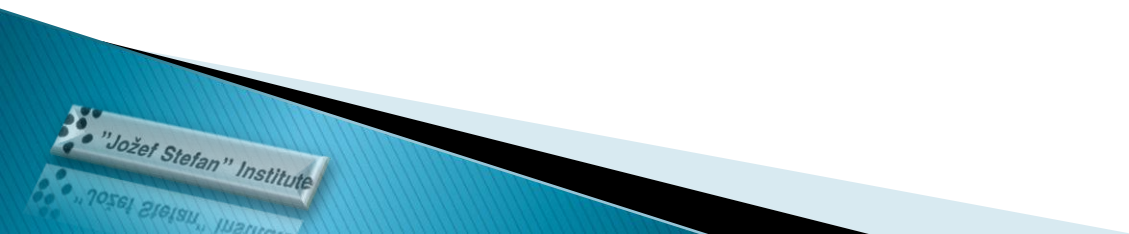


WORLD WIDE WEB  
c o n s o r t i u m

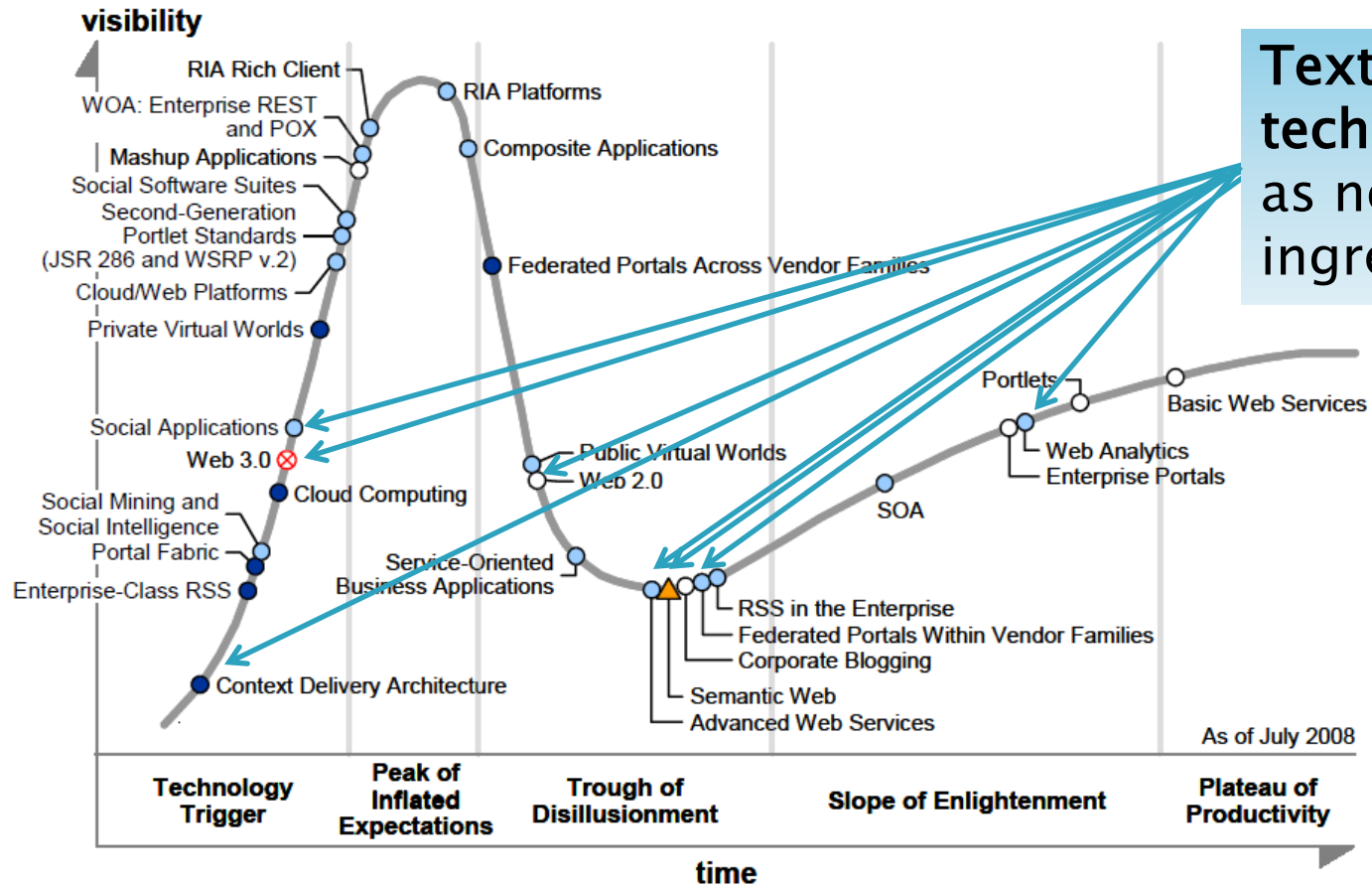
Bled, Sep 4<sup>th</sup> 2009

# Outline

- ▶ Semantic technologies from above
  - ...Gartner hype cycle, research, issues, ...
  - ...Text Analytics
  - ...Semantic Technology Stack
  - ...Beautiful world of Web X.X (Web 1.0, 2.0, 3.0, ...)
- ▶ Demonstrations
  - Data modeling with Support Vector Machines
  - Classification into large taxonomies
  - Contextual Search
  - Multilingual Search
  - News Bias
  - News Visualization
  - Summarization
  - Question Answering
  - Deep Reasoning
- ▶ Further references
  - ...events, books, videos



# Where Text related technologies fit in Gartner Hype Cycle 2008?



Hype Cycle for  
Web and User  
Interaction  
Technologies  
2008

Years to mainstream adoption:

○ less than 2 years

● 2 to 5 years

● 5 to 10 years

▲ more than 10 years

⊗ obsolete  
before plateau

Source: Gartner (July 2008)

"Jožef Stefan" In.

# Priority Matrix for Web and User Interaction Technologies 2008

benefit

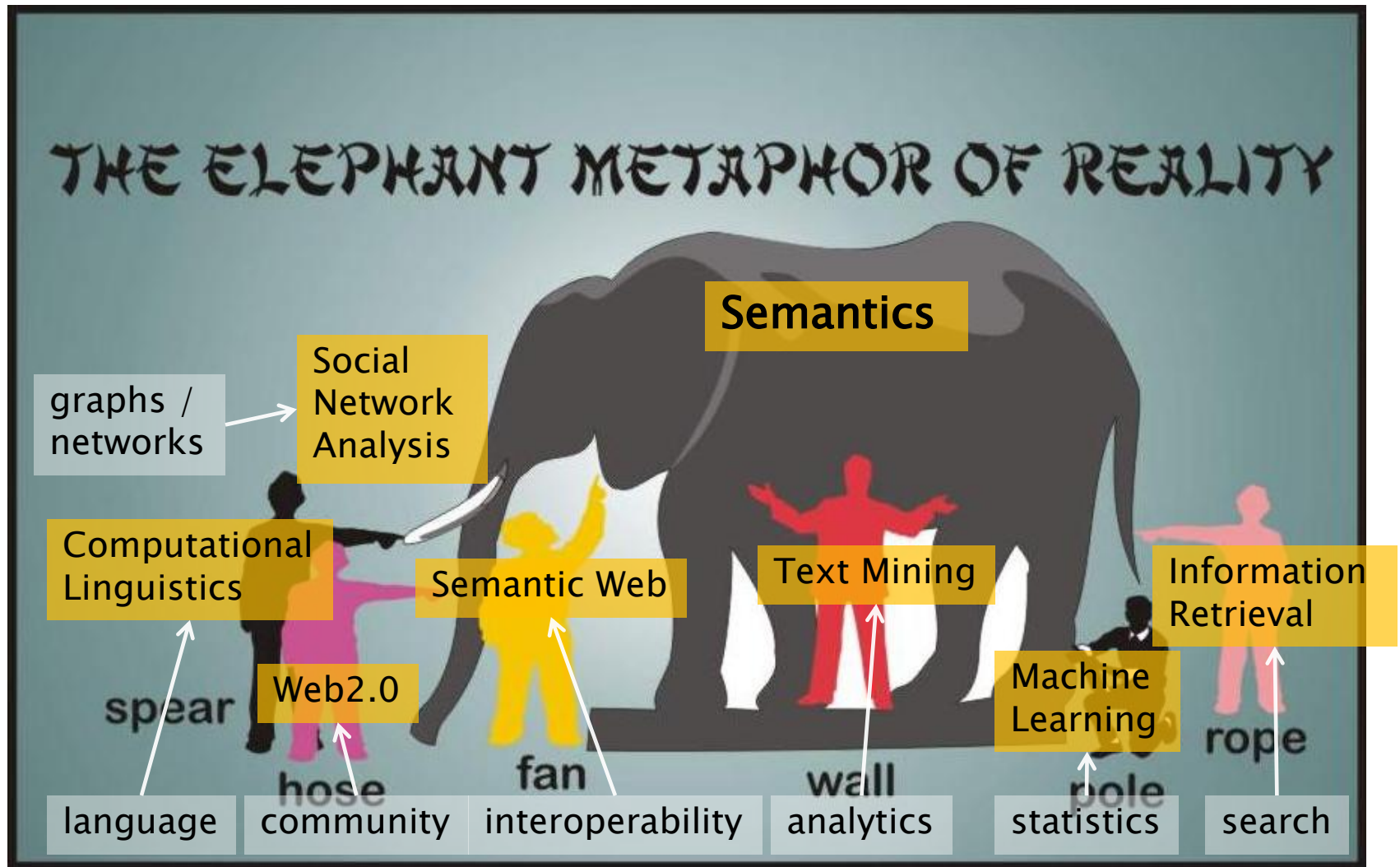
years to mainstream adoption

	less than 2 years	2 to 5 years	5 to 10 years	more than 10 years
transformational	<b>Enterprise Portals</b> <b>Web 2.0</b>	Cloud/Web Platforms Public Virtual Worlds SOA <b>Social Applications</b>	Cloud Computing <b>Context Delivery Architecture</b>	
high	<b>Mashup Applications</b> <b>Portlets</b>	Composite Applications Service-Oriented Business Applications <b>Social Mining and Social Intelligence</b> Social Software Suites WOA: Enterprise REST and POX	Federated Portals Across Vendor Families Portal Fabric Private Virtual Worlds	<b>Semantic Web</b>
moderate	Basic Web Services <b>Corporate Blogging</b>	Advanced Web Services Federated Portals Within Vendor Families RIA Platforms RIA Rich Client RSS in the Enterprise Second-Generation Portlet Standards (JSR 286 and WSRP v.2) <b>Web Analytics</b>	Enterprise-Class RSS	
low				

As of July 2008

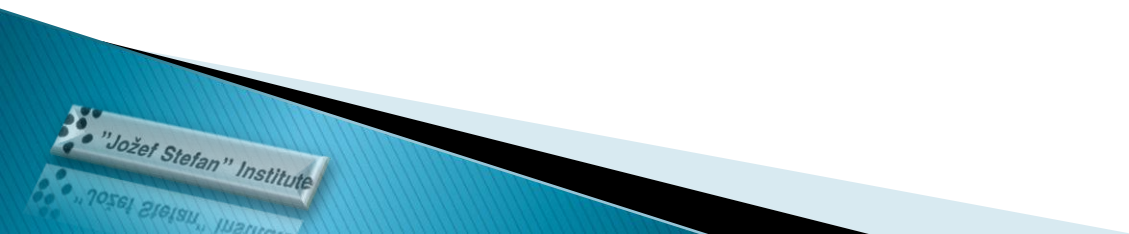


# How technical research areas are approaching text and semantics?



# ...some of the key issues from methodological side

- ▶ Integration of three key scientific paradigms
  - **Top-down approaches**
    - (Semantic Web, KRR)
  - **Bottom-up approaches**
    - (Machine Learning, Data Mining)
  - **Collaborative approaches**
    - (Web2.0, Social Computing)
- ▶ ...integration of ideas opens possibilities to solve problems which were not easy solvable before

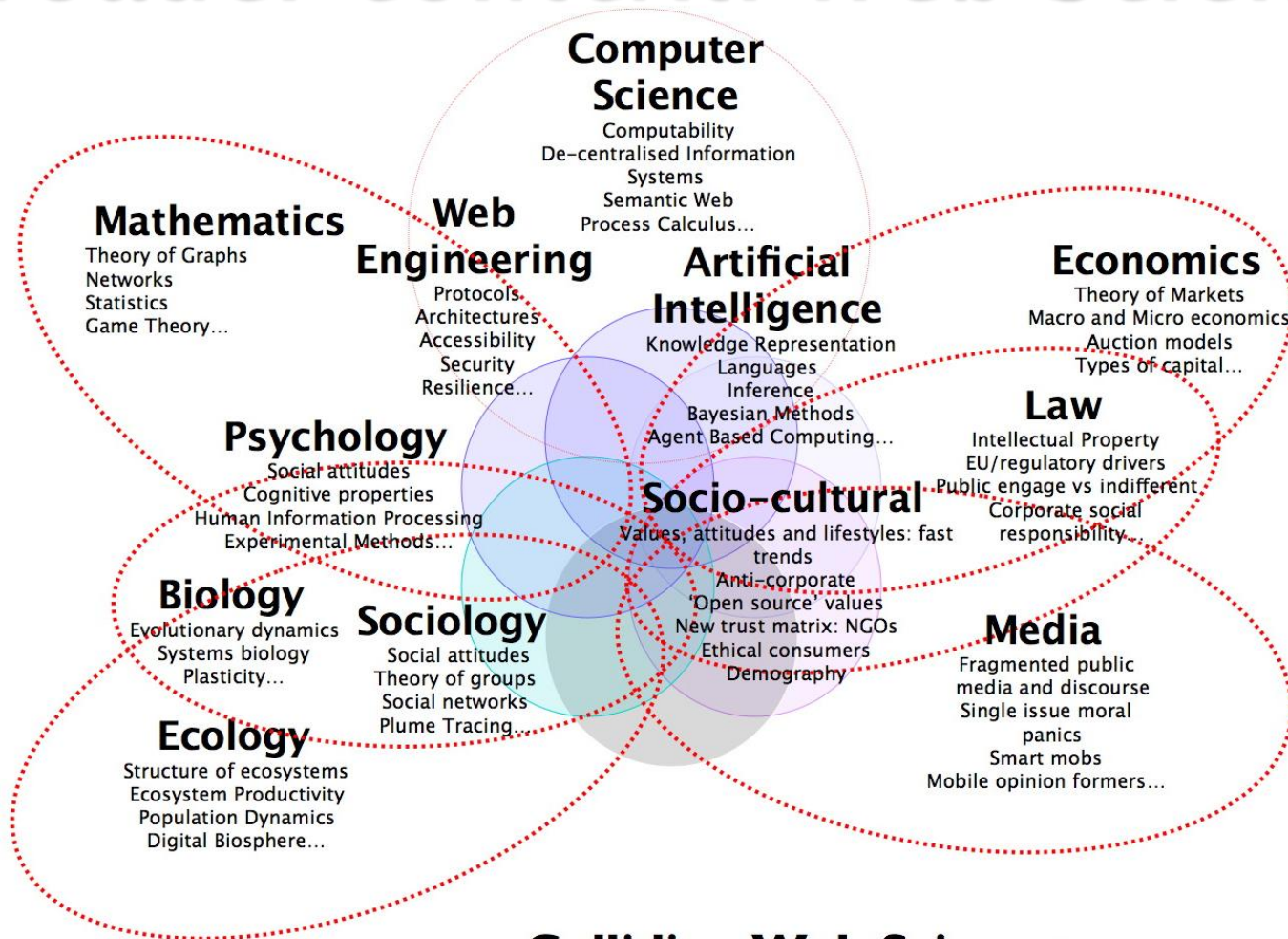


# What are the key issues in today's semantic technologies?

- ▶ **Scale**
  - How to deal with giga & tera bytes of text?
- ▶ **Expressivity**
  - Which representation language to choose? ...not too strong, not too weak!
- ▶ **Interoperability**
  - How to make technologies compatible? ...W3C stack!
- ▶ **Semantic Web Services**
  - How to offer technology and functionality over the Web?
- ▶ **Collaborative approaches**
  - How to organize people to work in the same direction? ...to get new “Wikipedias”
- ▶ **Relation to Computational Linguistics**
  - How to extract knowledge from text? ...still no big progress after decades
- ▶ **Knowledge Acquisition**
  - How to extract knowledge from data? ...by mining? ...still manually? ...cost?
- ▶ **Streams**
  - What if data comes in streams? Can we deal with semantics in real time setting?
- ▶ **Probabilistic approaches**
  - What if the data is not clean, consistent and correct? Still not used enough...

# Text & Semantic Analytics

# Broader context: Web Science



**Colliding Web Sciences**

<http://webscience.org/>



# How do we represent text?

# Levels of text representations

- ▶ Character (character n-grams and sequences)
- ▶ Words (stop-words, stemming, lemmatization)
- ▶ Phrases (word n-grams, proximity features)
- ▶ Part-of-speech tags
- ▶ Taxonomies / thesauri
- ▶ Vector-space model
- ▶ Language models
- ▶ Full-parsing
- ▶ Cross-modality
- ▶ Collaborative tagging / Web2.0
- ▶ Templates / Frames
- ▶ Ontologies / First order theories

Language  
identification,  
Copy detection

Named entity extraction  
(names of people, places,  
organizations)

Text categorization,  
Clustering, Search,  
Summarization, ...

Spam filtering,  
Machine  
translation

Multilingual search,  
Associating text with  
images, ...

Unifying  
semantics  
of data

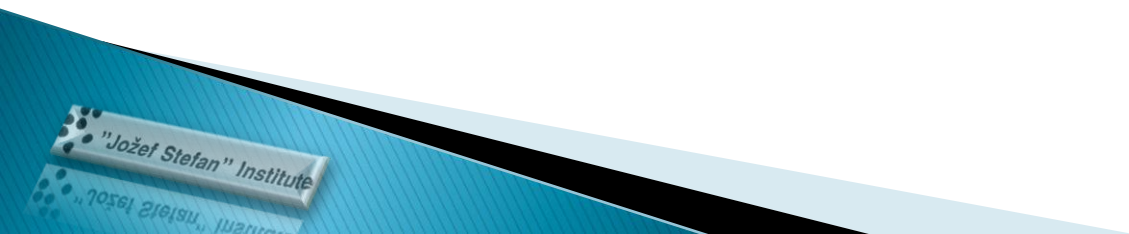
Reasoning,  
Semantic search

Simple

Semantic

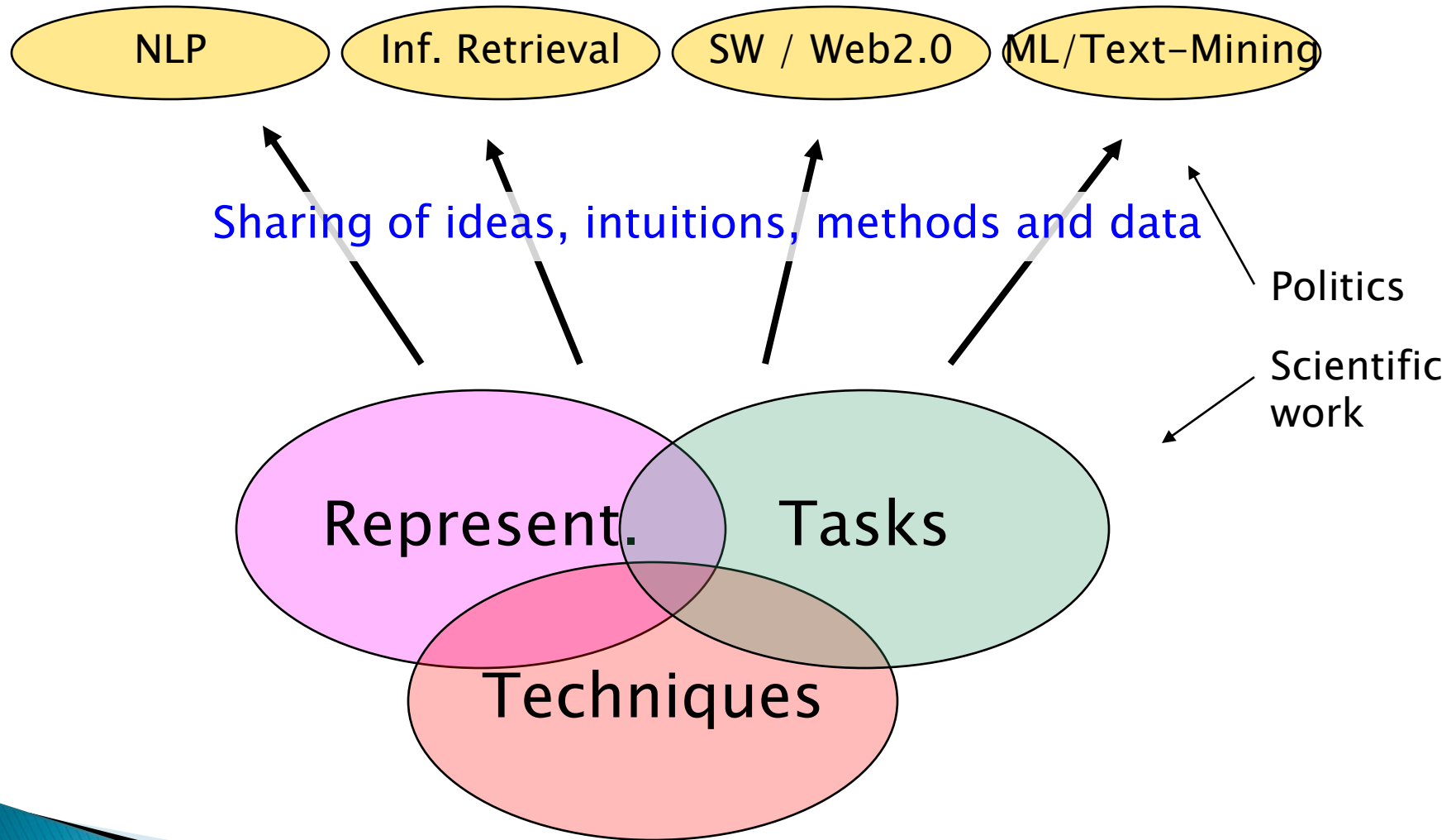
# What dimensions are in text analytics?

- ▶ Three major dimensions of text analytics:
  - Representations
    - ...from character-level to first-order theories
  - Techniques
    - ...from manual work, over analytics to reasoning
  - Tasks
    - ...from search, over (un-, semi-) supervised learning, to visualization, summarization, translation ...





# How dimensions fit to research areas?



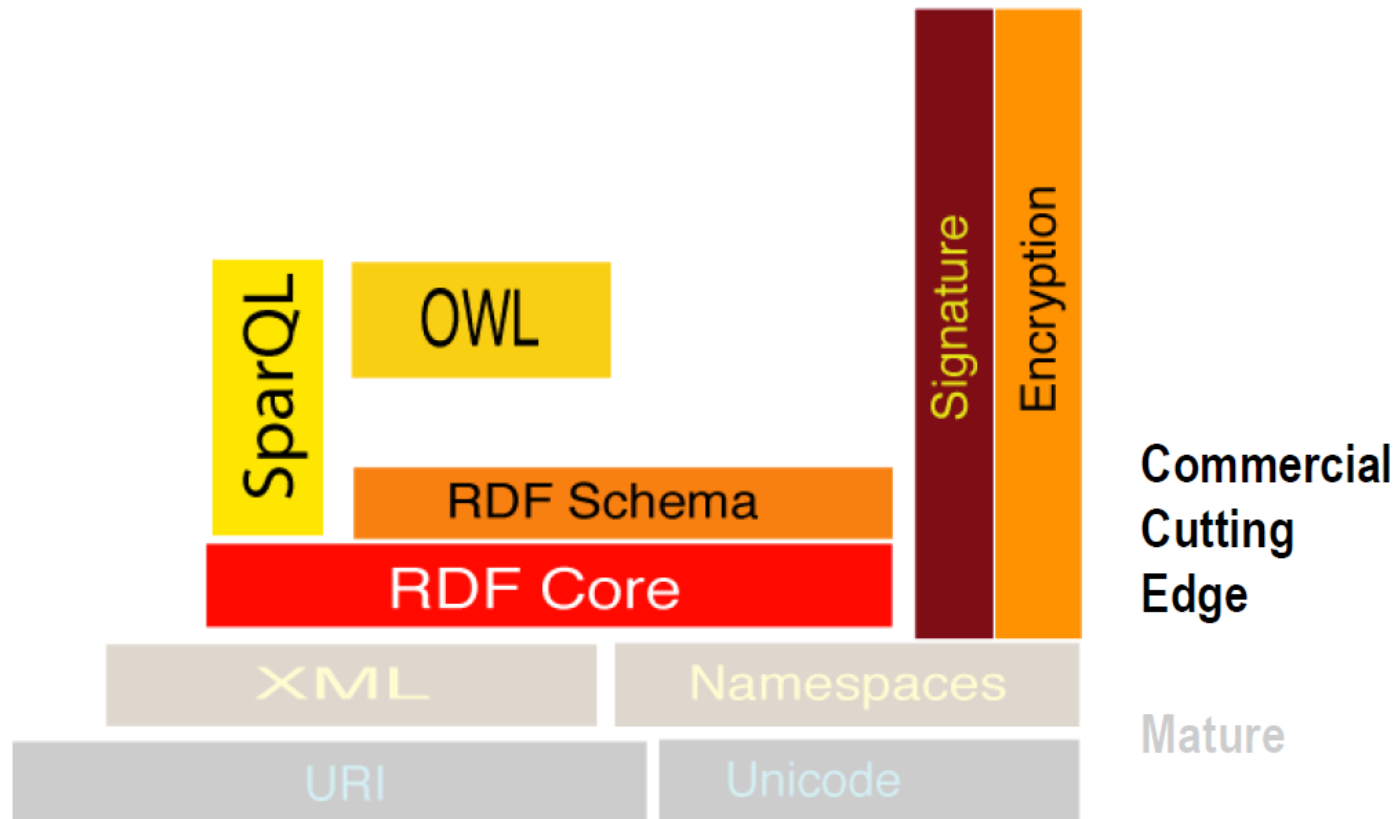
# Semantic Web Technology stack

# The Semantic Web in 2008



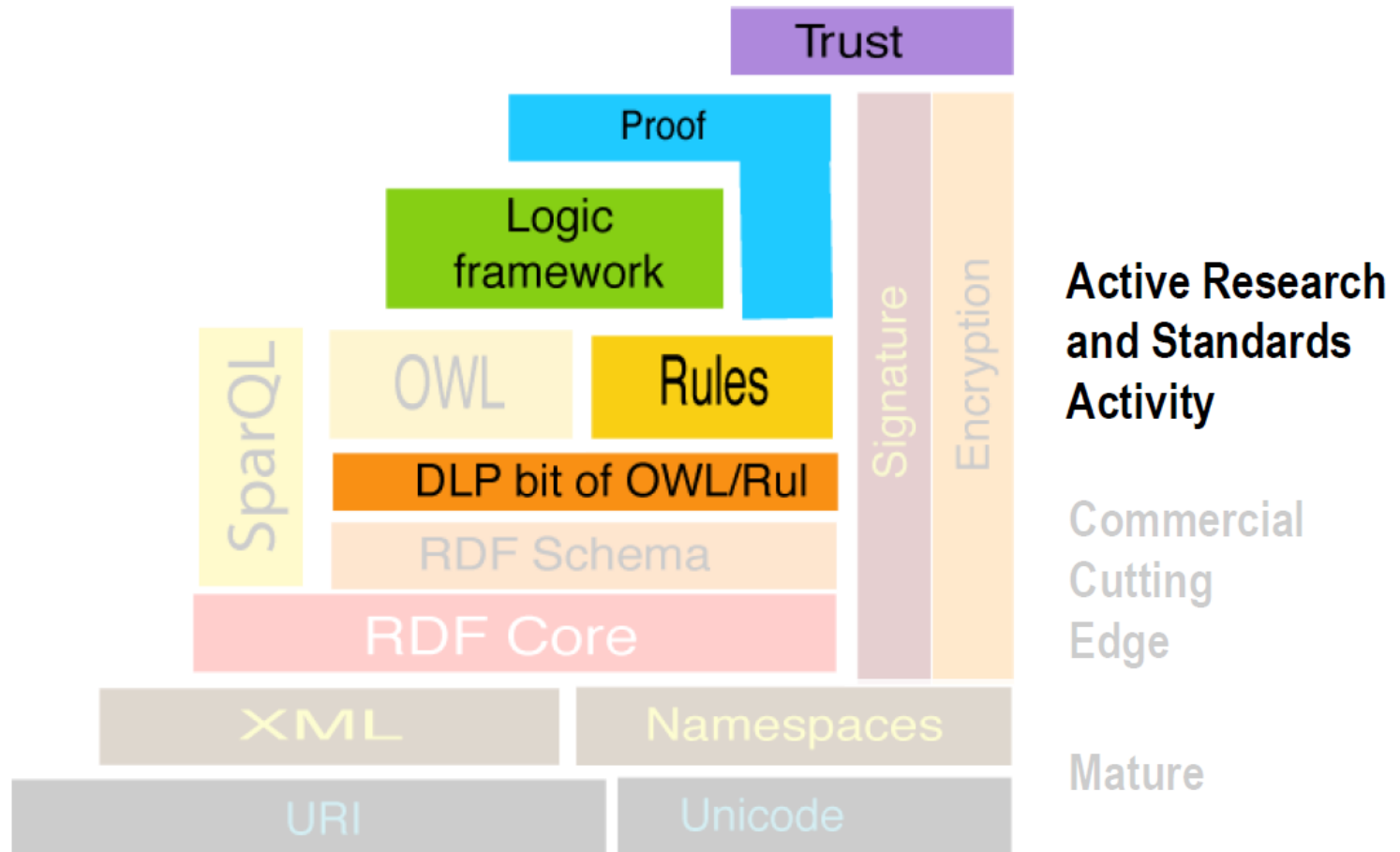
***“The Famous Semantic Web Technology Stack”***

# The Semantic Web in 2008



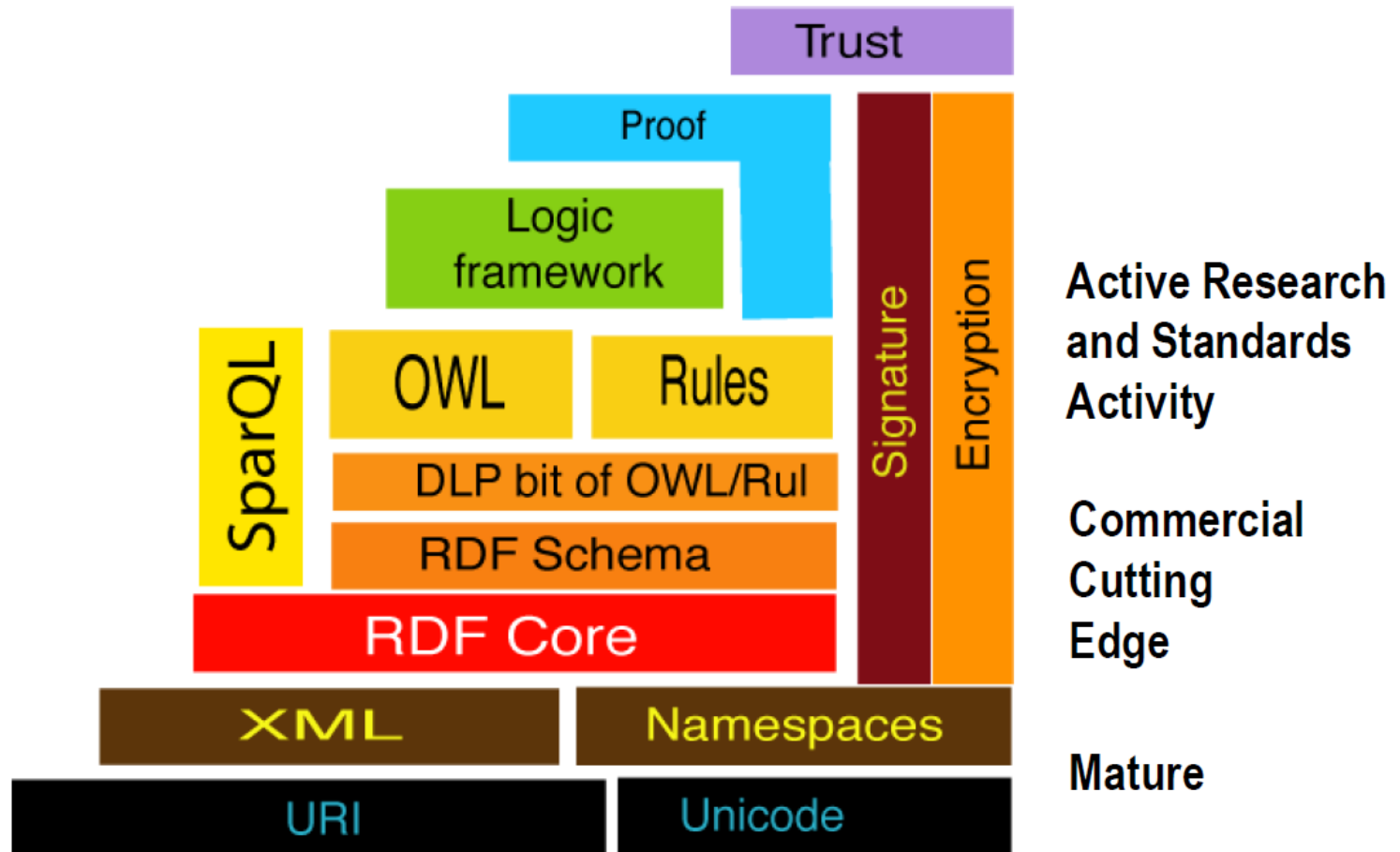
***“The Famous Semantic Web Technology Stack”***

# The Semantic Web in 2008



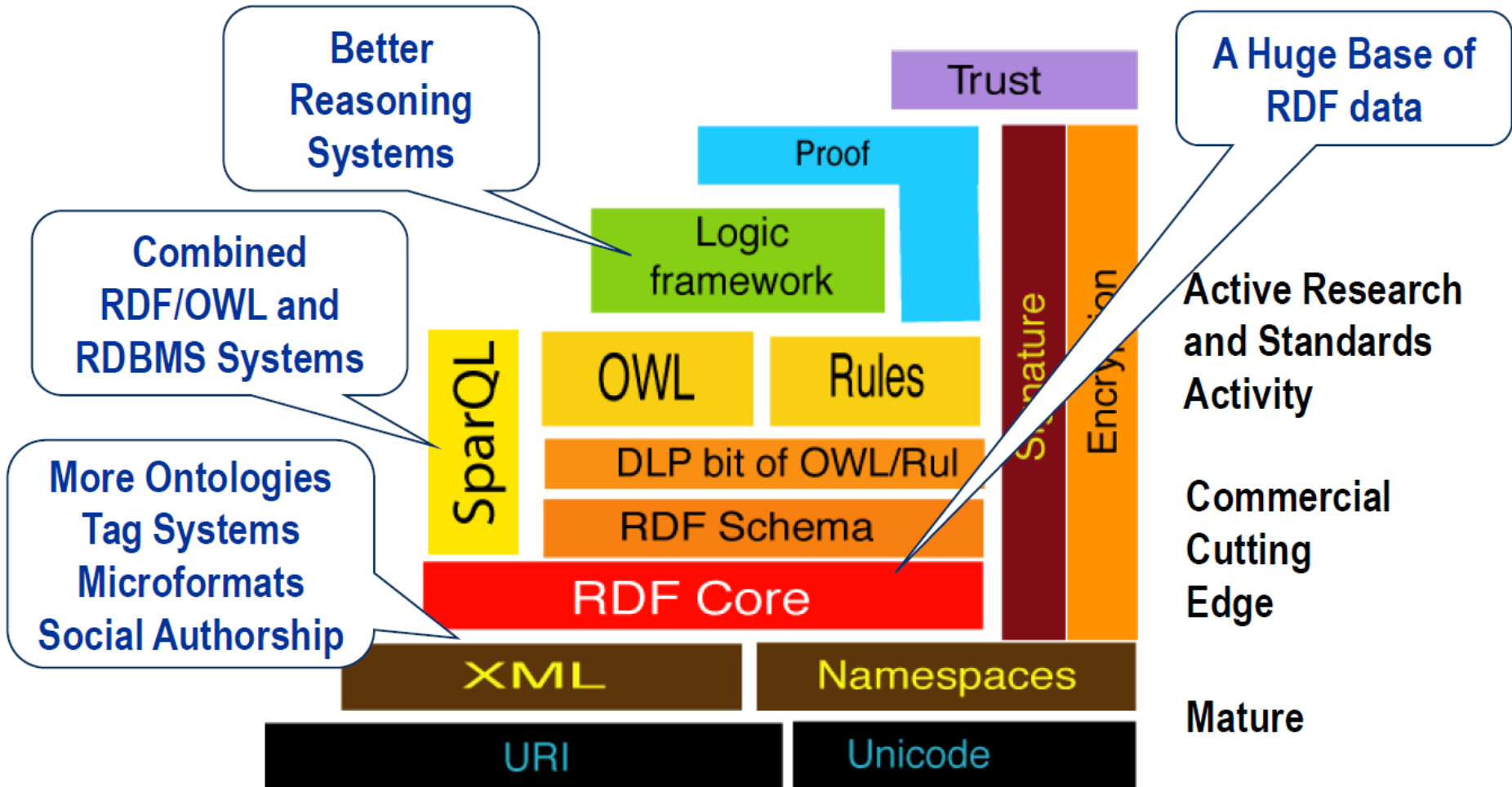
***“The Famous Semantic Web Technology Stack”***

# The Semantic Web in 2008



***“The Famous Semantic Web Technology Stack”***

# Completing the Semantic Web Picture



*Other Technologies Impact the Semantic Web*



# The beautiful world of Web X.X



# The beautiful world of Web X.X versions (...a trial to put all of them on one slide)

	Description	Technologies
<b>Web 1.0</b>	<b>Static</b> HTML pages (web as we first learned it)	HTML, HTTP
<b>Web 1.5</b>	<b>Dynamic</b> HTML content (web as we know it)	Client side (JavaScript, DHTML, Flash, ...), server side (CGI, PHP, Perl, ASP/.NET, JSP, ...)
<b>Web 2.0</b>	<b>Integration</b> on all levels, collaboration, sharing vocabularies (web as it is being sold)	weblogs, social bookmarking, social tagging, wikis, podcasts, RSS feeds, many-to-many publishing, web services, ... URI, XML, RDF, OWL, SparQL, ...
<b>Web 3.0</b>	...adding <b>meaning</b> to semantics - AI dream revival (web as we would need it)	Closest area of a research would be “common sense reasoning” and the “Cyc system” ( <a href="http://www.nytimes.com/2006/11/12/business/12web.html?ref=business">http://www.nytimes.com/2006/11/12/business/12web.html?ref=business</a> )

# Web 2.0 – is there any new quality?

- ▶ With “Web 2.0” the Web community became **really aware** of the importance of global collaborative work
  - ...Web2.0 is next step in globalization of the Web
  - **Bottom-up** “social networking” seems to nicely complement the traditional **top-down** schema design approaches



Visualization of Web 2.0 typical vocabulary  
([http://en.wikipedia.org/wiki/Image:Web20\\_en.png](http://en.wikipedia.org/wiki/Image:Web20_en.png))

# Web 2.0 – the current hype



"semantic web", "web 2.0", "data mining"

Search Trends

Tip: Use commas to compare multiple search terms.

Searches Websites

All regions

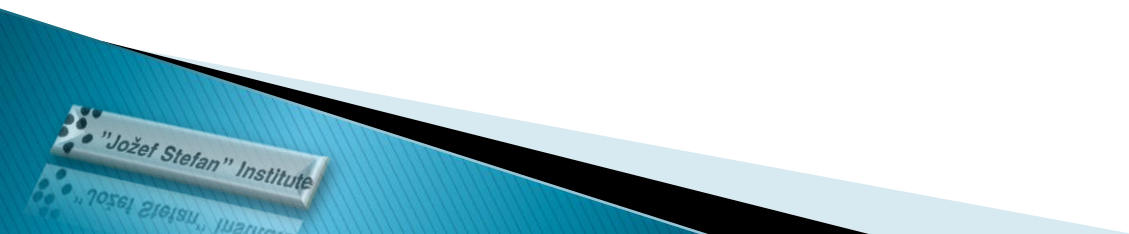
Scale is based on the average worldwide traffic of **"semantic web"** in all years. [Learn more](#)

**"semantic web"** — 1.00    **"web 2.0"** — 5.30    **"data mining"** — 3.90



# ...scale and dynamics of Web 2.0

- ▶ Per minute, there are:
  - 100 edits in Wikipedia (1 44K/day)
  - 200 tags in del.icio.us (288K/day)
  - 270 image uploads to flickr (388K/day)
  - 1 100 blog entries (1.6M/day)
- ▶ Are we prepared to deal with this amounts?
  - ...yes, but not in a simplistic way!
  - The key is to use technologies and ideas from different research areas!



# Some demonstrations...

...from light-weight to heavy-weight

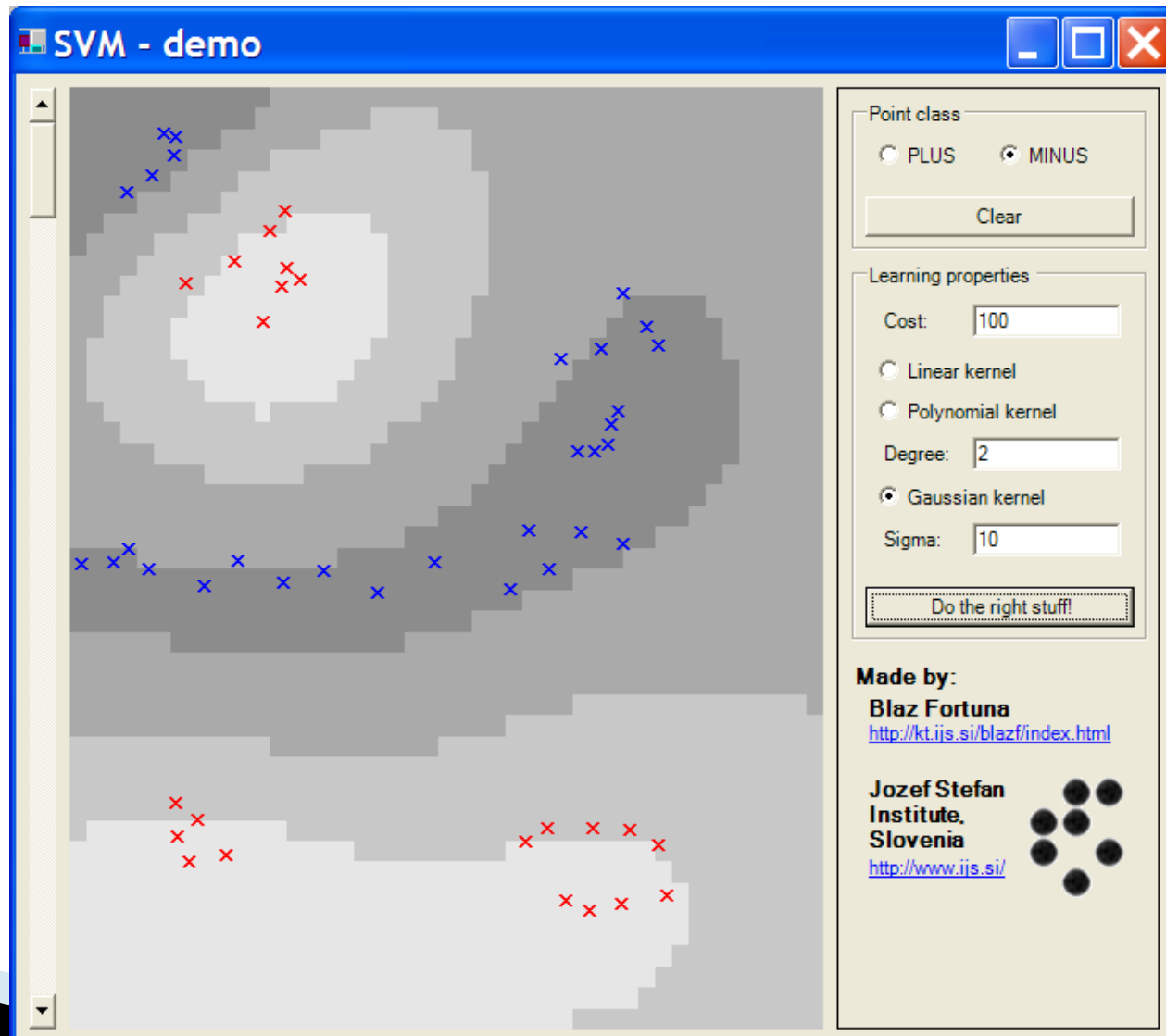
# Modeling the data

(learning discrimination models)



# Modeling the data with Support Vector Machine (SVM)

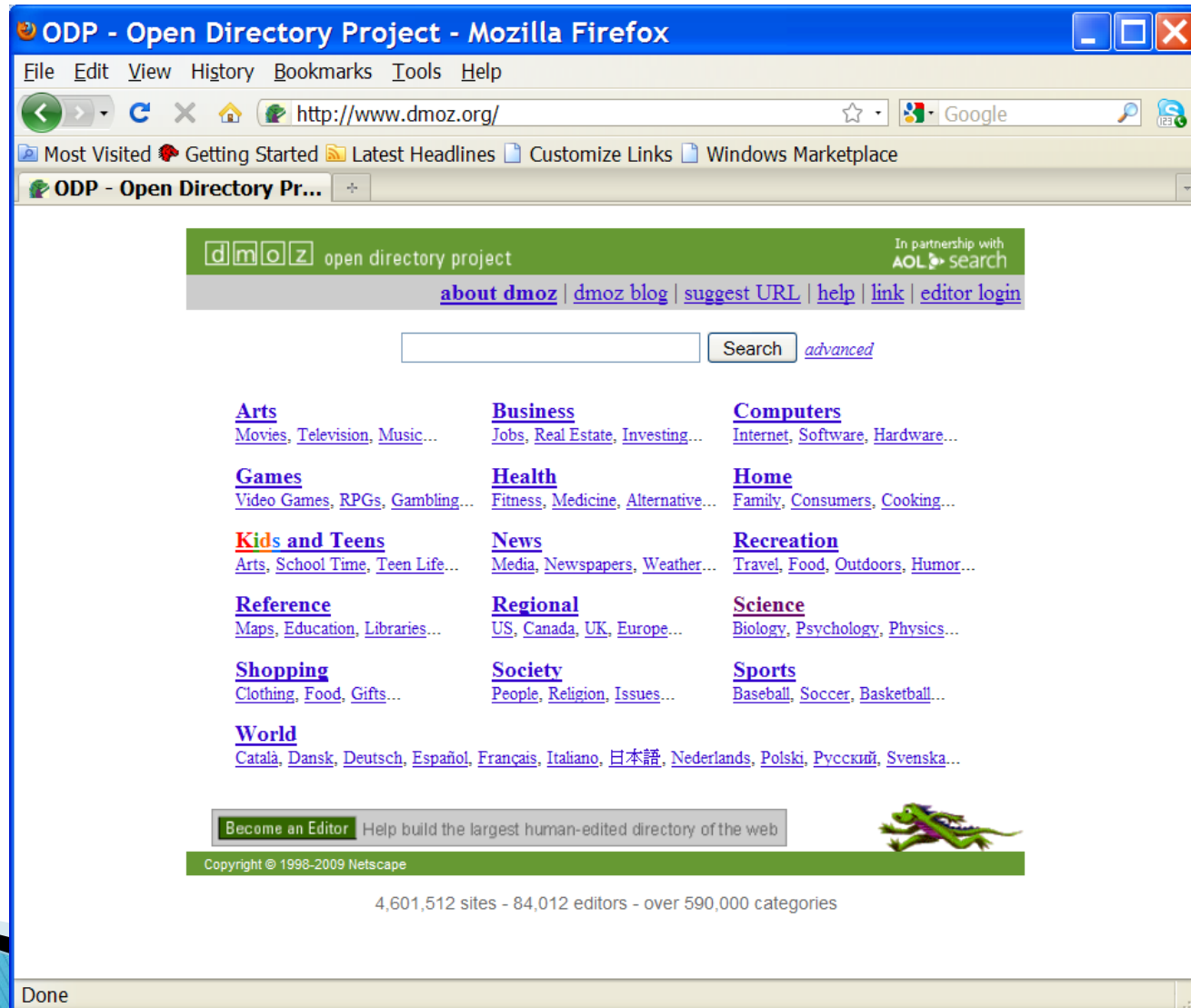
- ▶ One of the basic building blocks in most techniques is learning a function for splitting positive and negative data instances
- ▶ Support Vector Machine is most popular method for learning such functions



# Classification into large taxonomies



# DMoz (Open Directory Project)



# Classification of a query into DMoz

**Mozilla Firefox**

File Edit View History Bookmarks Tools Help

http://localhost:8080/

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

## Classification into DMoz Top\_Science

URL:

Text:

Categories:

Keyword Treshold (0=Non, 1=All):

Context: ☒

XML Format: ☐

Plain-Text Mime-Type: ☐

Done



**Classification - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

http://localhost:8080/Classify

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

## Results of Classification into DMoz Top\_Science

**Keywords:**

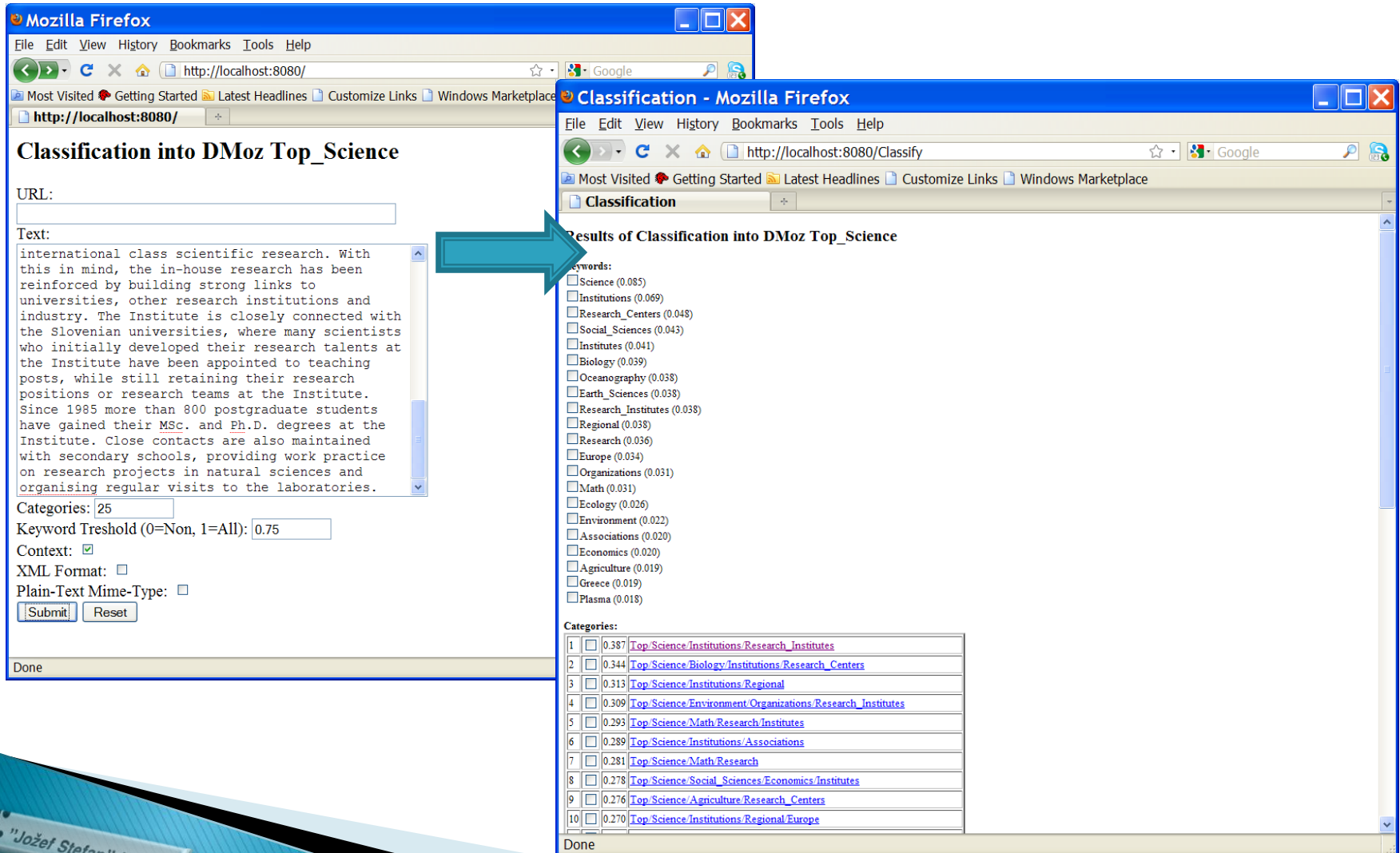
- ☐ Science (0.183)
- ☐ Math (0.183)
- ☐ Number\_Theory (0.182)
- ☐ Diophantine\_Equations (0.154)
- ☐ Fermat's\_Last\_Theorem (0.115)

**Categories:**

1	<input type="checkbox"/>	0.504	<a href="#">Top/Science/Math/Number_Theory/Diophantine_Equations/Fermat's_Last_Theorem</a>
2	<input type="checkbox"/>	0.341	<a href="#">Top/Science/Math/Number_Theory/Diophantine_Equations</a>
3	<input type="checkbox"/>	0.117	<a href="#">Top/Science/Math/Number_Theory/Tables</a>
4	<input type="checkbox"/>	0.113	<a href="#">Top/Science/Math/Number_Theory/History</a>
5	<input type="checkbox"/>	0.074	<a href="#">Top/Science/Math/Number_Theory/Factoring</a>
6	<input type="checkbox"/>	0.064	<a href="#">Top/Science/Math/Number_Theory/Factoring/Tables</a>
7	<input type="checkbox"/>	0.053	<a href="#">Top/Science/Math/Number_Theory</a>
8	<input type="checkbox"/>	0.053	<a href="#">Top/Science/Math/Number_Theory/Prime_Numbers/Primality_Tests/Pseudoprimes</a>
9	<input type="checkbox"/>	0.037	<a href="#">Top/Science/Math/History/People</a>
10	<input type="checkbox"/>	0.035	<a href="#">Top/Science/Math/Number_Theory/Prime_Numbers/Mersenne</a>
11	<input type="checkbox"/>	0.027	<a href="#">Top/Science/Math/Number_Theory/Publications/Books</a>

Done

# Classification of a document into DMoz



**Classification into DMoz Top\_Science**

URL:

Text:

Categories:

Keyword Threshold (0=Non, 1=All):

Context: ☒

XML Format: ☐

Plain-Text Mime-Type: ☐

Done

**Classification - Mozilla Firefox**

http://localhost:8080/Classify

**Classification**

**Results of Classification into DMoz Top\_Science**

Keywords:

- ☐ Science (0.085)
- ☐ Institutions (0.069)
- ☐ Research\_Centers (0.048)
- ☐ Social\_Sciences (0.043)
- ☐ Institutes (0.041)
- ☐ Biology (0.039)
- ☐ Oceanography (0.038)
- ☐ Earth\_Sciences (0.038)
- ☐ Research\_Institutes (0.038)
- ☐ Regional (0.038)
- ☐ Research (0.036)
- ☐ Europe (0.034)
- ☐ Organizations (0.031)
- ☐ Math (0.031)
- ☐ Ecology (0.026)
- ☐ Environment (0.022)
- ☐ Associations (0.020)
- ☐ Economics (0.020)
- ☐ Agriculture (0.019)
- ☐ Greece (0.019)
- ☐ Plasma (0.018)

Categories:

1	<input type="checkbox"/> 0.387 <a href="#">Top_Science/Institutions/Research_Institutes</a>
2	<input type="checkbox"/> 0.344 <a href="#">Top_Science/Biology/Institutions/Research_Centers</a>
3	<input type="checkbox"/> 0.313 <a href="#">Top_Science/Institutions/Regional</a>
4	<input type="checkbox"/> 0.309 <a href="#">Top_Science/Environment/Organizations/Research_Institutes</a>
5	<input type="checkbox"/> 0.293 <a href="#">Top_Science/Math/Research_Institutes</a>
6	<input type="checkbox"/> 0.289 <a href="#">Top_Science/Institutions/Associations</a>
7	<input type="checkbox"/> 0.281 <a href="#">Top_Science/Math/Research</a>
8	<input type="checkbox"/> 0.278 <a href="#">Top_Science/Social_Sciences/Economics/Institutes</a>
9	<input type="checkbox"/> 0.276 <a href="#">Top_Science/Agriculture/Research_Centers</a>
10	<input type="checkbox"/> 0.270 <a href="#">Top_Science/Institutions/Regional/Europe</a>

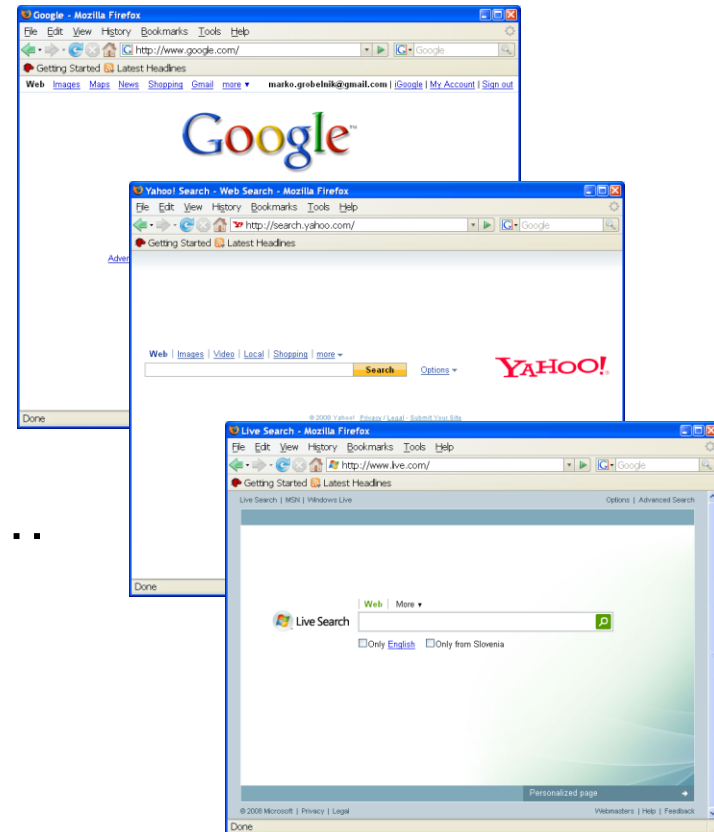
Done

# Visual & Contextual Search

(WWW2008)

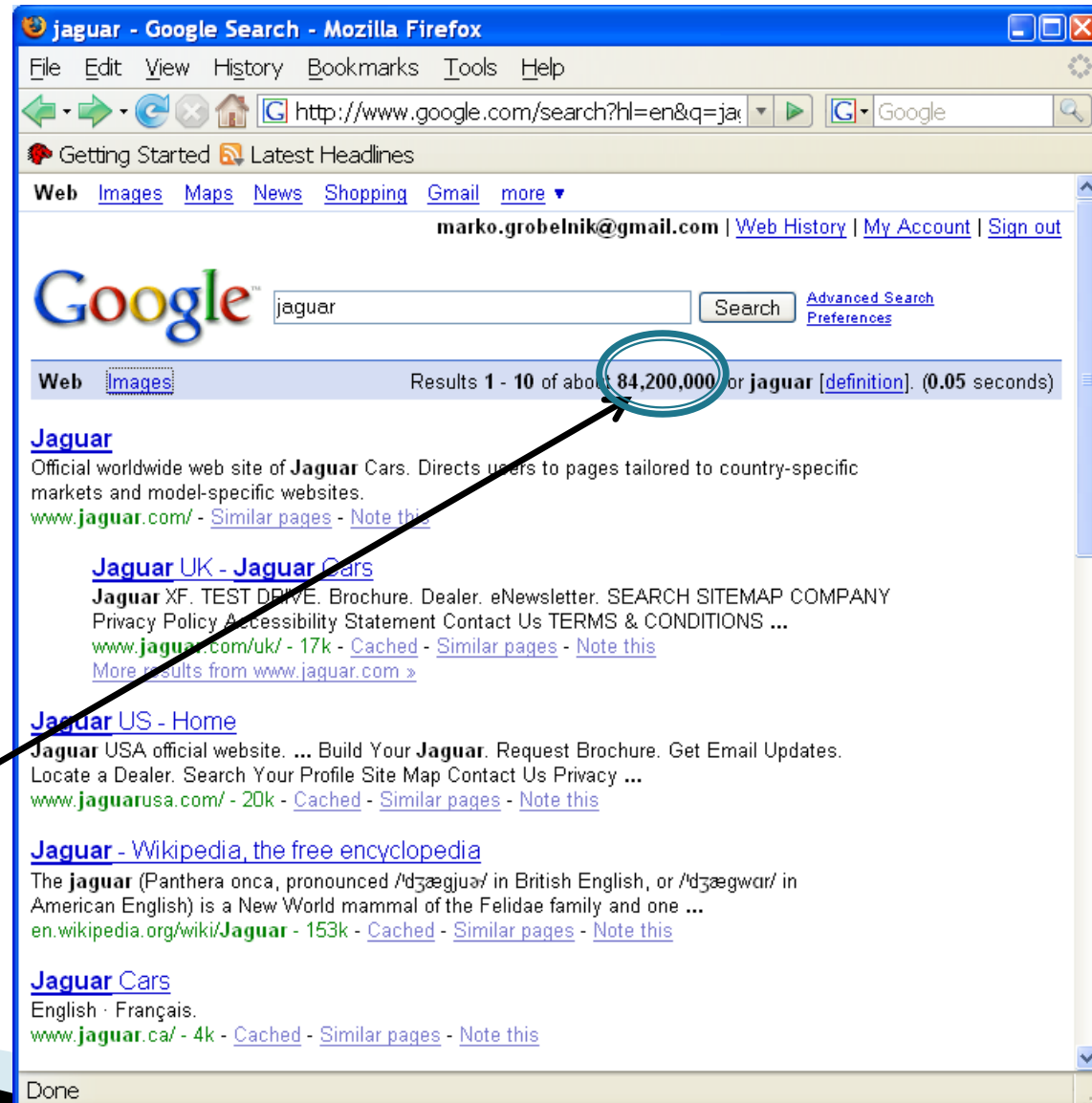
# Contextualized search

- ▶ What is the most common tasks where we manipulate text in everyday life?
  - “Internet search”!
- ▶ ...but – how smart is search technology today?
  - ...not too smart!
  - It is sophisticated, but not smart...



# Example: searching for “Jaguar”

- ▶ Query “jaguar” has many meanings...
- ▶ ...but the first page of search engines doesn't provide us with many answers
- ▶ ...there are 84M more results





# Context sensitive search with

## <http://searchpoint.ijs.si>

The screenshot shows the SearchPoint search engine interface in a Windows Internet Explorer browser window. The address bar displays <http://searchpoint.ijs.si/Result.aspx>. The search bar contains the query "jaguar". Below the search bar, there are three buttons: "Search via topics", "Search via query to ontology", and "Search via hits to ontology". The search results are listed on the left, with the top result being "(9) Jaguar" from Big Cats Online. To the right of the search results is a conceptual map. The map is a network of nodes and edges. The central node is "Top". Other nodes include "Mammalia", "Vehicles", "Shopping", "Sports", "Games", "Console Platforms", "Aviation", "Society", "Recreation", "Models", "Enthusiasts", "Aircraft", "Parts and Accessories", "NFL", and "Dance". The map is a conceptual map of the search results, showing the relationships between different topics related to "jaguar".

Query

Conceptual map

Search Point

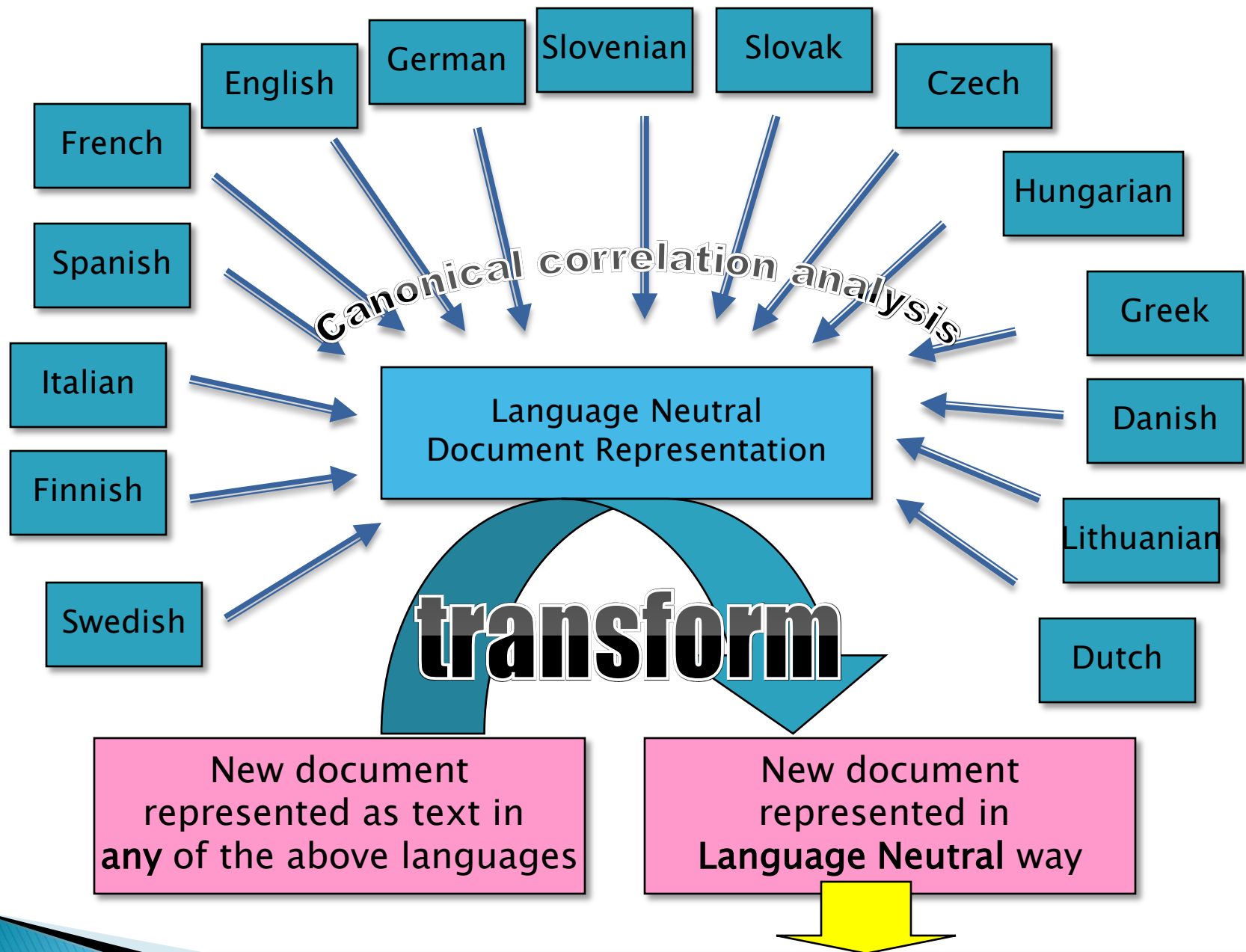
Dynamic contextual ranking based on the search point

# Multilingual search

<http://www.smart-project.eu/>







...enables cross-lingual retrieval, categorization, clustering, ...

# Example of cross-lingual information retrieval on Reuters news corpus (using KCCA)

The image shows two overlapping Mozilla Firefox browser windows. The background window displays search results for 'horse' in German. The foreground window displays the search interface for 'stock exchange' in English, with language options 'en' (selected) and 'de', and a 'Bili Search' button. Below the button, a table of search results is shown.

Weight	Document Name
0.311615	143935.en.txt
REPUBLIC OF IRELAND: Countyglen Irish Stock Exchange listing cancelled.	
Countyglen Irish Stock Exchange listing cancelled.	
The Irish Stock Exchange said in a statement on Friday that the listing for Countyglen Plc had been cancelled with effect from 1500 GMT (1600 local time).	
-- Dublin Newsroom +353 1 676 9775	
0.240322	328995.de.txt
BORSE ANDERT PLANE FUR OPTIONSSCHEINHANDEL.	
Frankfurt (Reuter) - Die Deutsche Borse AG hat ihre Plane fur den elektronischen Handel mit Optionsscheinen geandert. Das ursprunglich angedachte Computersystem OHS werde nicht verwirklicht, weil die Anforderungen der Marktteilnehmer hier nicht zu sinnvollen Investitions-und Betriebskosten hatten verwirklicht werden konnen, erklarte die Borse am Freitag in Frankfurt. Stattdessen soll der elektronische Handel mit Optionsscheinen im Rahmen des geplanten allgemeinen Elektronischen Handelssystems (EHS) der Borse realisiert und im Laufe des	

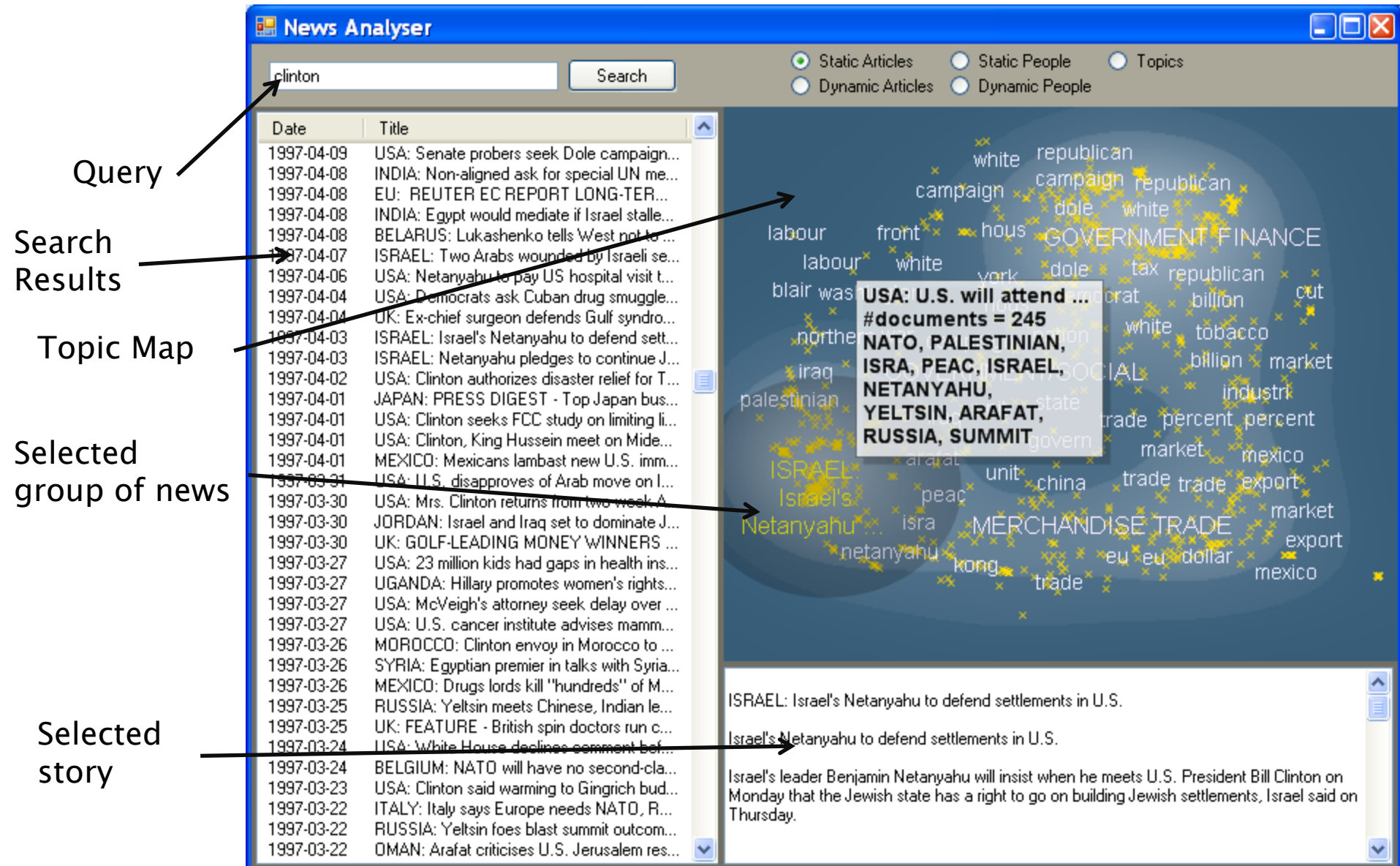
# News reporting bias

# Detecting News Reporting Bias

- ▶ The task:
  - Given a news story, are we able to say from which news source it came?
- ▶ We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
  - ...300 aligned articles describing the same story from both sources
- ▶ The same topics are expressed in both sources with the following keywords:
  - CNN with:
    - **Insurgents**, Troops, Baghdad, Iran, **Militant**, Police, **Suicide**, **Terrorist**, United, National, Hussein, **Alleged**, Israeli, Syria, Terrorism...
  - Aljazeera with:
    - Attacks, Claims, **Rebels**, Withdrawing, Report, **Fighters**, President, **Resistance**, Occupation, Injured, Army, Demanded, Hit, Muslim, ...

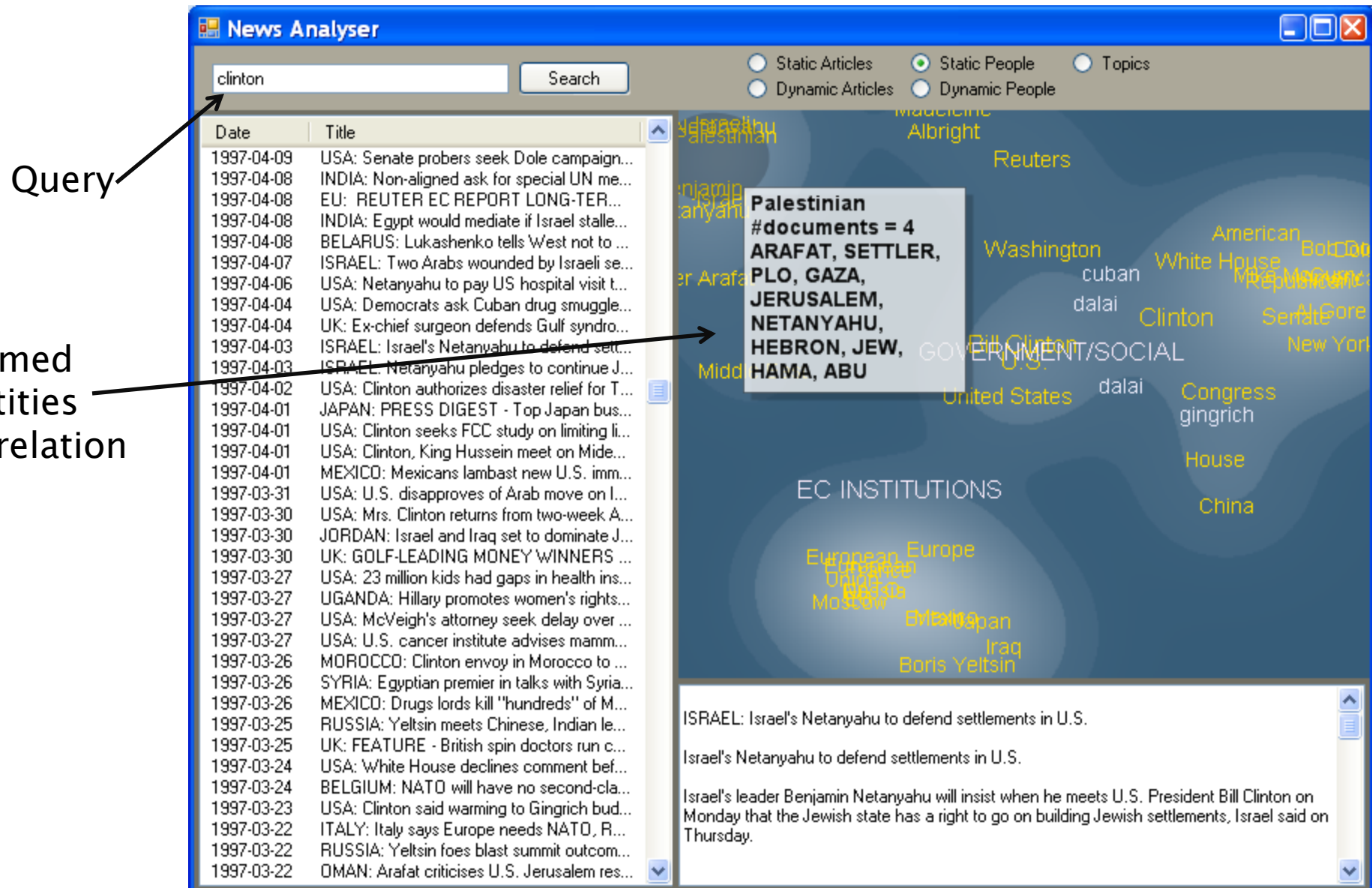
# News Visualization

# Topic landscape of the query “Clinton” from Reuters news 1996–1997

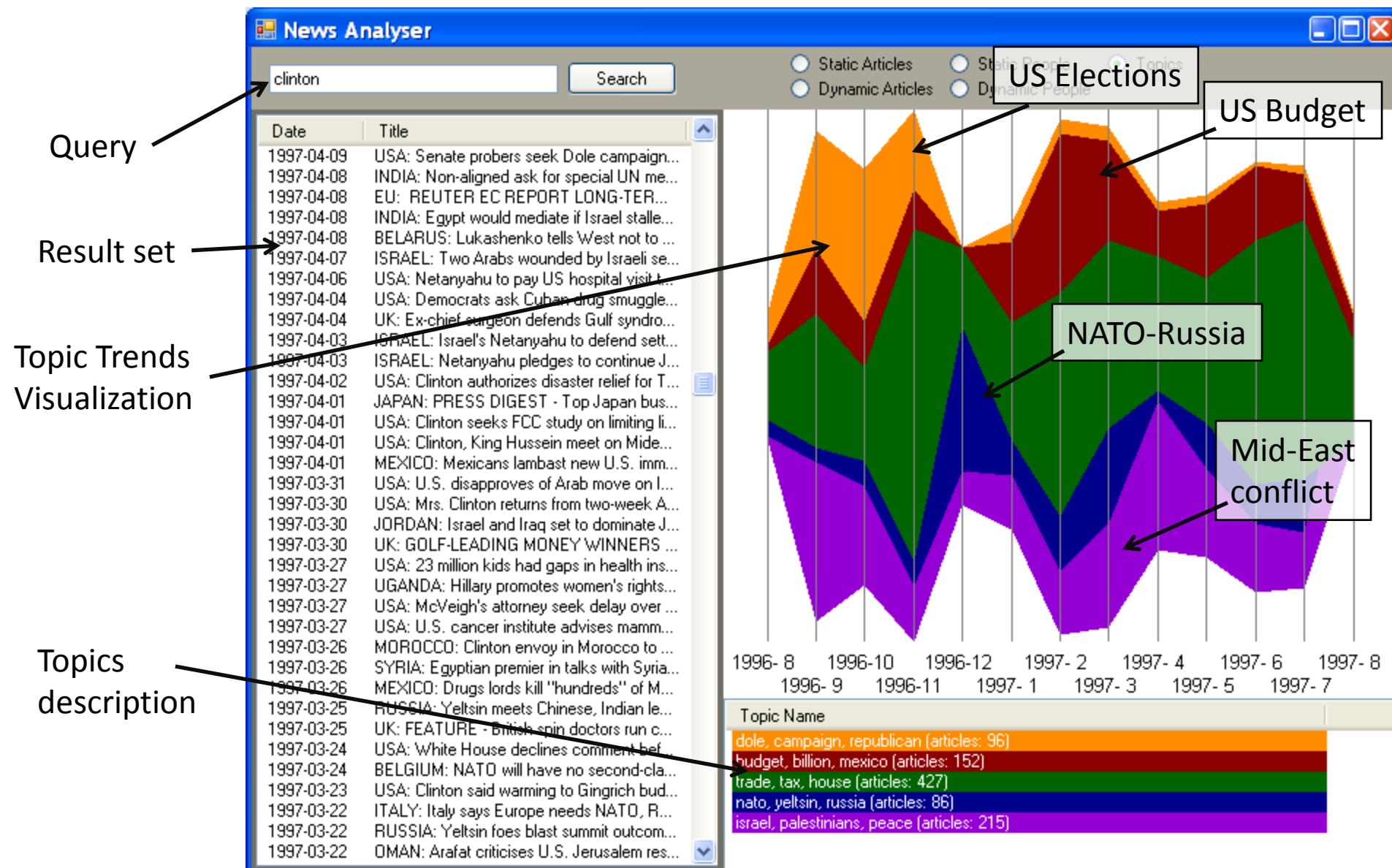




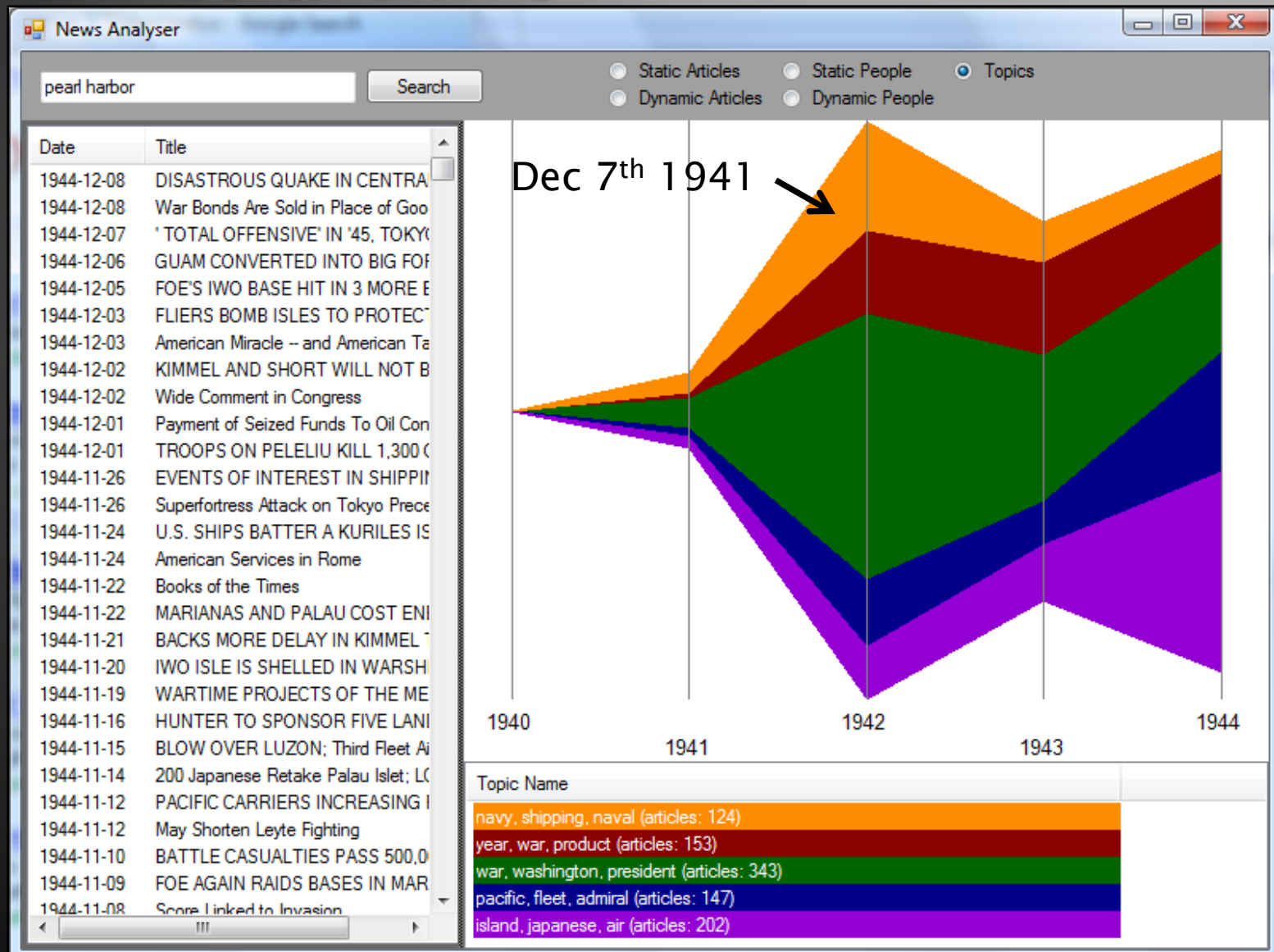
# Visualization of social relationships between “Clinton” and other entities



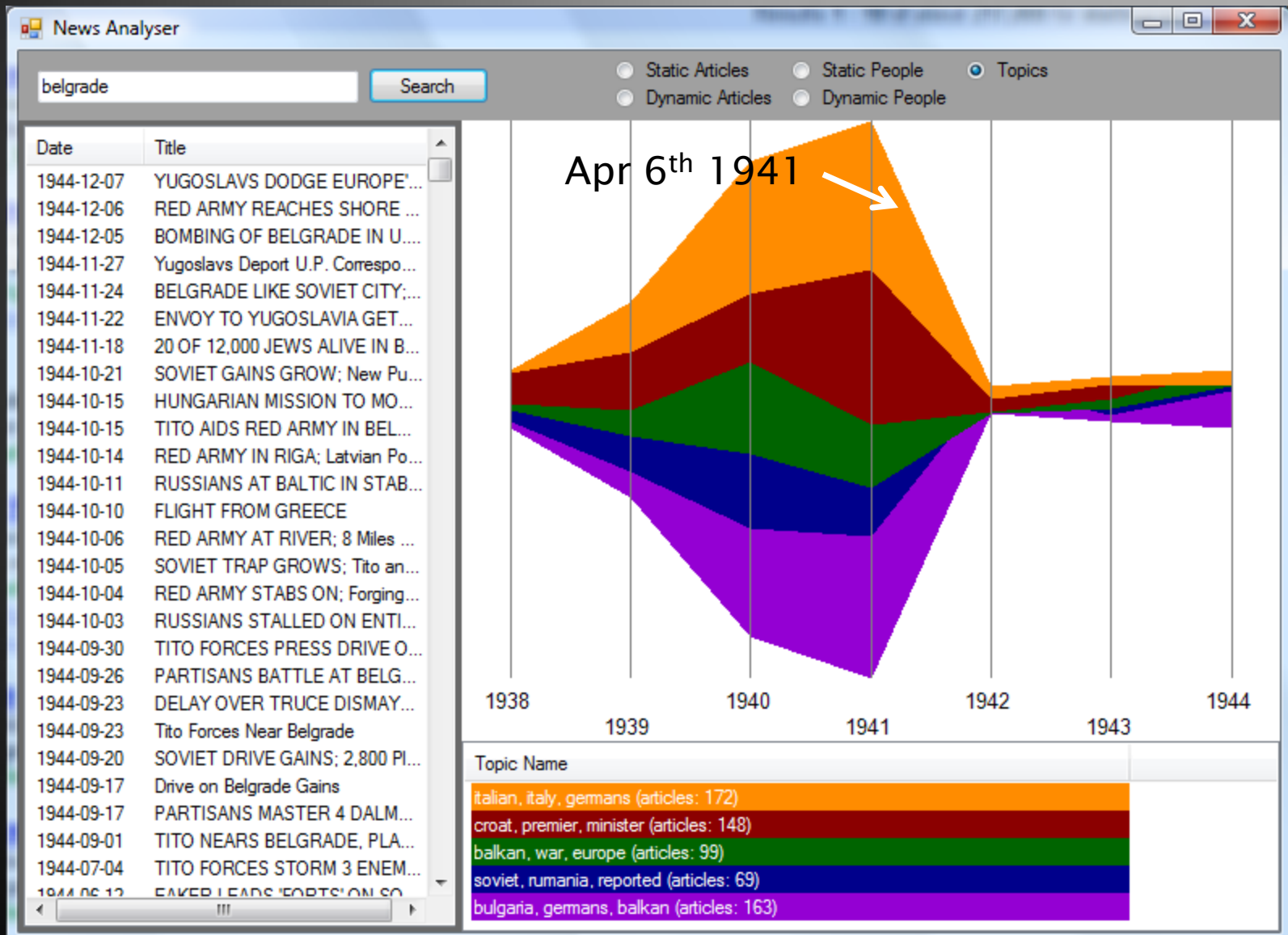
# Topic Trends Tracking of the documents including "Clinton"



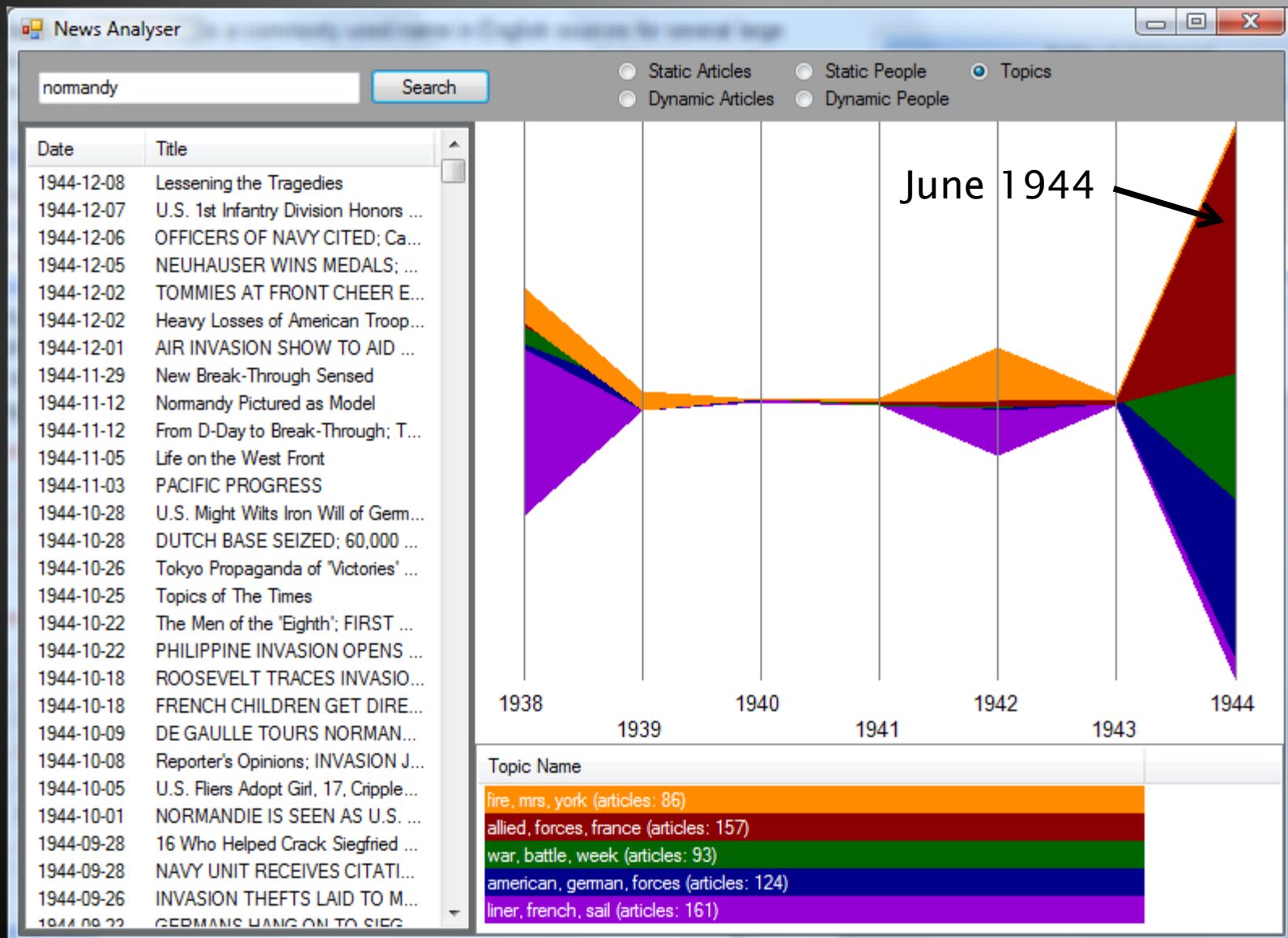
# WW2 query “Pearl Harbor” into NYTimes archive



# WW2 query “Belgrade” into NYTimes archive



# WW2 query “normandy” into NYTimes archive



# LiveNetLife



# LiveNetLife connects people on similar information

## Awareness



searching



blogging



selling

Posted by admin on August 4, 2008

### Are You Looking For A Cheap Car Rental Service?

In today's world of competitive marketing and cutthroat competition, a customer can't he/she is closing will benefit her or not, most of the time most, consumers, just go on w hassle of inquiring for more economic alternatives, or discounts.

This also applies to car rentals, nowadays, it's hard to call anything a cheap car rental, w continuing gas price increase it's hard to keep product prices down, so here are a few tips even if every other factor is making it expensive.

## Semantic Similarity

enterprise

Rent a Car Business Rentals Rent a Truck Buy a Car

Buy a Car

[Used Car Deals](#)

Join Our **Business Rental Program**  
[Learn more](#)

Long Term Rentals  
[We can help](#)

Make a car rental reservation

Rent a car in: **US** | [Canada](#) | [UK](#) |

US includes Puerto Rico

Location Enter either a City, State or an [Airport](#) or [Port of Call](#).

1

☐ Show airport locations only

Dates & Times  
Start

Google

Web [Maps](#) [News](#)

**Rent a Car**  
[Rent.A.Car](#) [Hertz](#) [si](#) Ugoden Rent a Car! Veliko število vozil vam je na razpolago 2

Looking for local results for **rent a car**?

☒ Remember this location

**Hertz** [f](#) [x](#)  
Reserve a rental **car** from Hertz car rental and get a great rate online. Find out how eas to book a hybrid, convertible or luxury **car** today.  
[www.hertz.com/](#) - 145k - [Cached](#) - [Similar pages](#) - [f](#) [x](#)

**Enterprise Rent-A-Car** - Rental Cars at Low Rates [f](#) [x](#)  
Book a car rental from Enterprise **Rent-A-Car** at low rates. Rental car classes include economy cars, luxury car rental, cargo vans, pickup trucks and more.  
[www.enterprise.com/](#) - 52k - [Cached](#) - [Similar pages](#) - [f](#) [x](#)

**Avis Rent A Car** [f](#) [x](#)

# LiveNetLife in ACTIVE

## Awareness



search for  
information



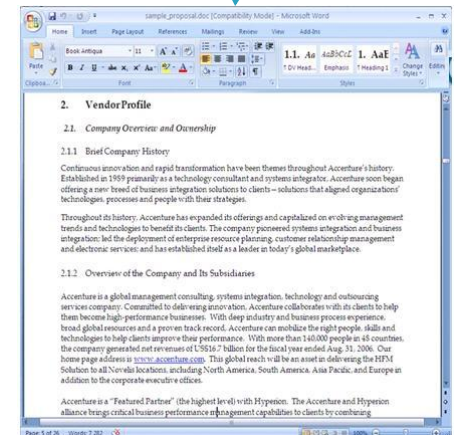
bid  
preparation



knowledge  
formalization



context /  
task  
similarity



# LiveNetLife demonstration

## ► Web page/browser/office plugin



## ► Video Demonstration

- <http://www.facebook.com/video/video.php?v=49721269044&subj=528454044>

## ► Live Demonstration

- <http://videolectures.net/>



# Knowledge based summarizaiton

AAAI 2005

[illegible]

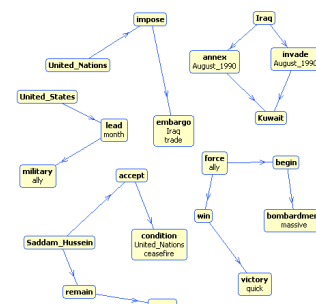
Automatically  
generated  
document summary

Cracks appeared in the U.N. trade embargo against Iraq. The State Department reports that Cuba and Romania have struck oil deals with Iraq as others attempt to trade with Baghdad in defiance of the sanctions. Iran has agreed to exchange food and medicine for Iraqi oil. Saddam has offered developing nations free oil if they send their tankers to pick it up. Thus far, none has accepted. Japan, accused of responding too slowly to the Gulf crisis, has promised \$2 billion in aid to countries hit hardest by the Iraqi trade embargo. President Bush has promised that Saddam's aggression will not succeed.

# Sentence extraction rule

Select a sub-graph that  
would characterize  
extracted summaries

## Learn the sub-graph selection model



# Detailed Summarization Procedure

Linguistic analysis of the text

- Deep parsing of sentences

Refinement of the text parse

- Named-entity consolidation

Determine that 'George Bush' = 'Bush'  
= 'U.S. president'

- Anaphora resolution

Link pronouns with name-entities

Extract **Subject-Predicate-Object** triples

Compose a **graph** from triples

Describe each triple with a set of features for learning

**Learn** a model to classify triples into the summary

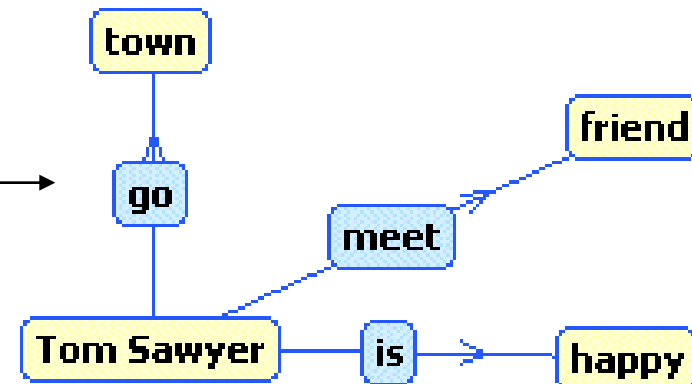
Generate a **summary graph**

Use summary graph to generate textual document summary

Tom Sawyer went to town. He met a friend. Tom was happy. ...

**Tom Sawyer** went to town. He [**Tom Sawyer**] met a friend. Tom [**Tom Sawyer**] was happy. ...

Tom  $\leftarrow$  go  $\rightarrow$  town  
Tom  $\leftarrow$  meet  $\rightarrow$  friend  
Tom  $\leftarrow$  is  $\rightarrow$  happy





# Example of automatic summary

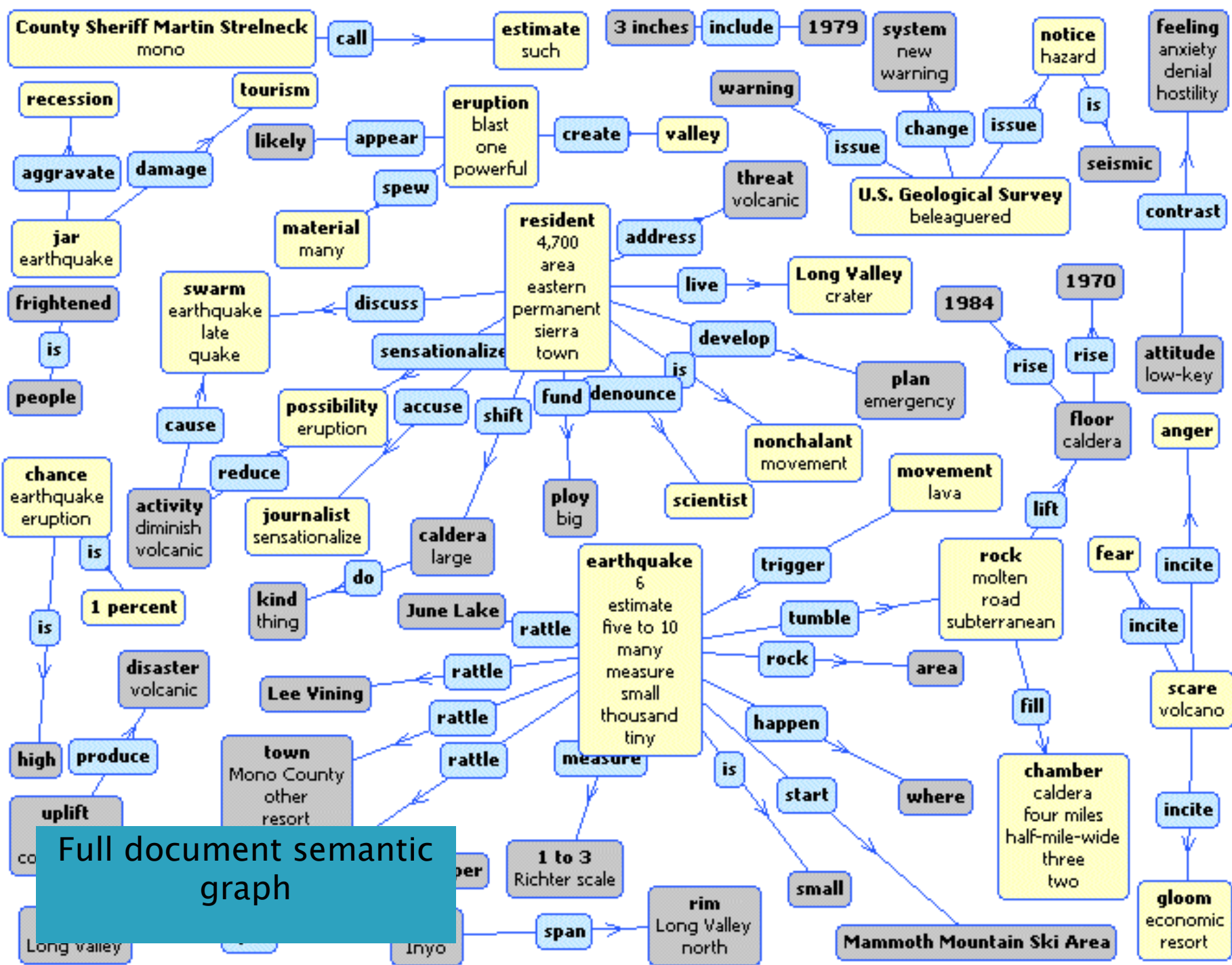
## Cracks Appear in U.N. Trade Embargo Against Iraq.

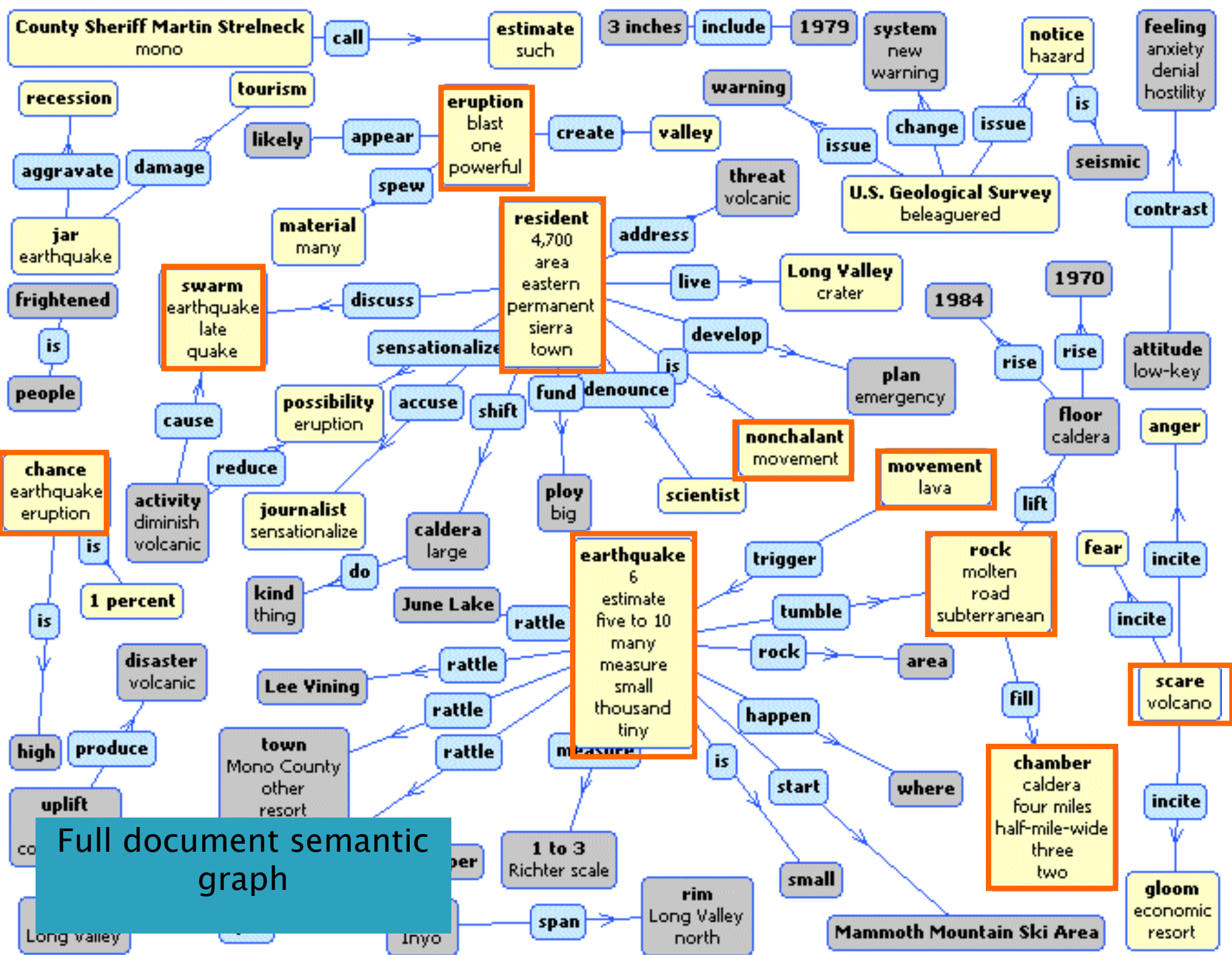
# Human extracted summary

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Saddam Hussein sought to circumvent the economic no-fly zone around his country. Japan, meanwhile, announced it would increase its aid to countries hardest hit by enforcing the sanctions. Hoping to defuse criticism that it is not doing enough to oppose Baghdad, Japan said up to \$2 billion in aid may be sent to nations most affected by the U.N. embargo on Iraq. President Bush on Tuesday night promised a joint session of Congress and a nationwide radio and television audience that "Saddam Hussein will fail" to make his conquest of Kuwait permanent. "America must stand up to aggression, and we will," said Bush, who added that the U.S. military may remain in the Saudi Arabian desert indefinitely. "I cannot predict just how long it will take to convince Iraq to withdraw from Kuwait," Bush said. More than 150,000 U.S. troops have been sent to the Persian Gulf region to deter a possible Iraqi invasion of Saudi Arabia. Bush's aides said the president would follow his address to Congress with a televised message for the Iraqi people, declaring the world is united against their government's invasion of Kuwait. Saddam had offered Bush time on Iraqi TV. The Philippines and Namibia, the first of the developing nations to respond to an offer Monday by Saddam of free oil in exchange for sending their own tankers to get it, said no to the Iraqi leader. Saddam's offer was seen as a none-too-subtle attempt to bypass the U.N. embargo, in effect since four days after Iraq's Aug. 2 invasion of Kuwait, by getting other countries to dock their tankers in Iraq. But according to a State Department survey, Cuba and Romania have struck oil deals with Iraq and companies elsewhere are trying to continue trade with Baghdad, all in defiance of U.N. sanctions. Romania denies the allegation. The report, made available to The Associated Press, said some Eastern

- Eight years after a volcano scare incited fear, anger and economic gloom in Sierra resorts, residents are nonchalant about renewed underground lava movement that is triggering thousands of tiny earthquakes.
- The resort town's 4,700 permanent residents live in Long Valley, a 19-mile-long, 9-mile wide volcanic crater known as a caldera.
- The Earth's crust is being stretched apart in the region, allowing molten rock to fill half-mile-wide chambers under the caldera.
- The valley was created 730,000 years ago by one of Earth's most powerful eruptions, a blast that spewed 600 times more material than the May 1980 eruption of Mount St. Helens in Washington state.
- Despite the current activity, the probability of a major earthquake or a volcanic eruption in the area is "less than 1 percent each year," said David Hill, the U.S. Geological Survey geophysicist in charge of research at Long Valley. Mono County Sheriff Martin Strelneck called such estimates "a scientific guessing game," and said area residents rarely discuss the latest swarm of earthquakes, which started in May 1989.
- As a result, the Geological Survey issued a "notice of potential volcanic hazard" for Long Valley in May 1982.
- That warning, coupled with jarring earthquakes, damaged tourism and aggravated a recession in the once-booming real estate market.

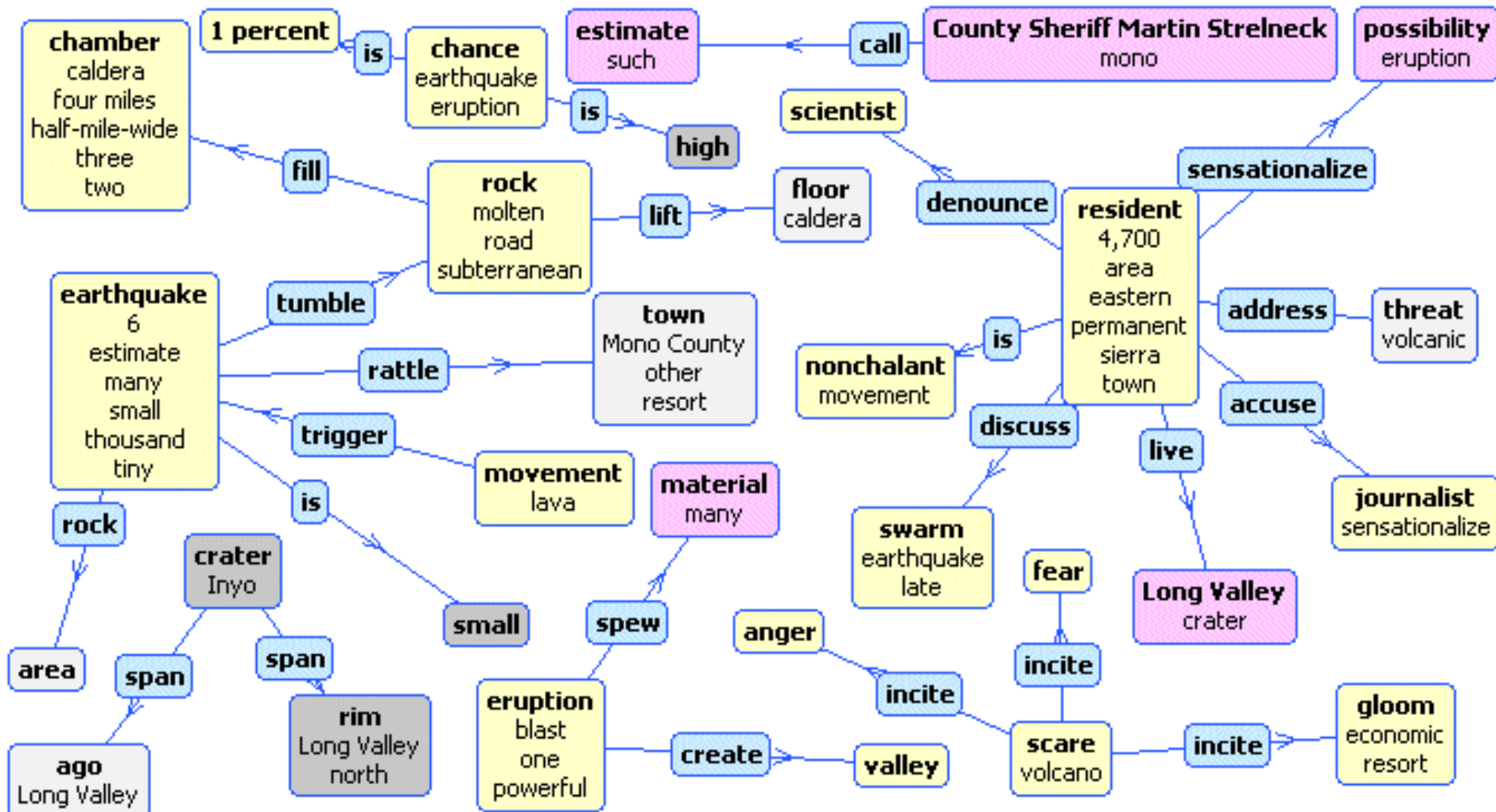
member, has led a drive among oil-producing nations to boost production to make up for the shortfall caused by the loss of Iraqi and Kuwaiti oil from the world market. Their oil makes up 20 percent of the world's oil reserves. Only Saudi Arabia has higher reserves. But according to the State Department, Cuba, which faces an oil deficit because of reduced Soviet deliveries, has received a shipment of Iraqi petroleum since U.N. sanctions were imposed five weeks ago. And Romania, it said, expects to receive oil indirectly from Iraq. Romania's ambassador to the United States, Virgil Constantinescu, denied that claim Tuesday, calling it "absolutely false and without foundation."



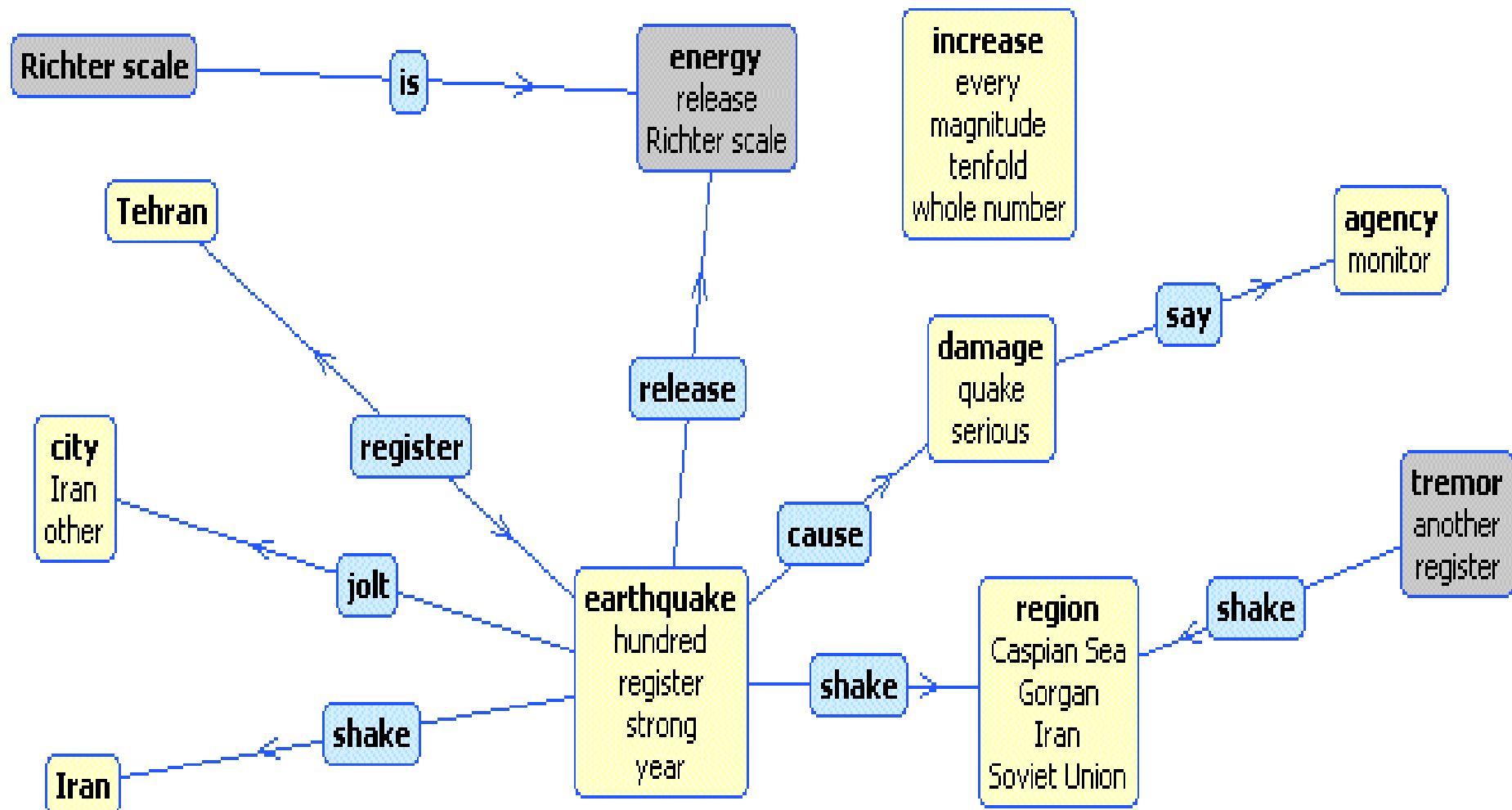


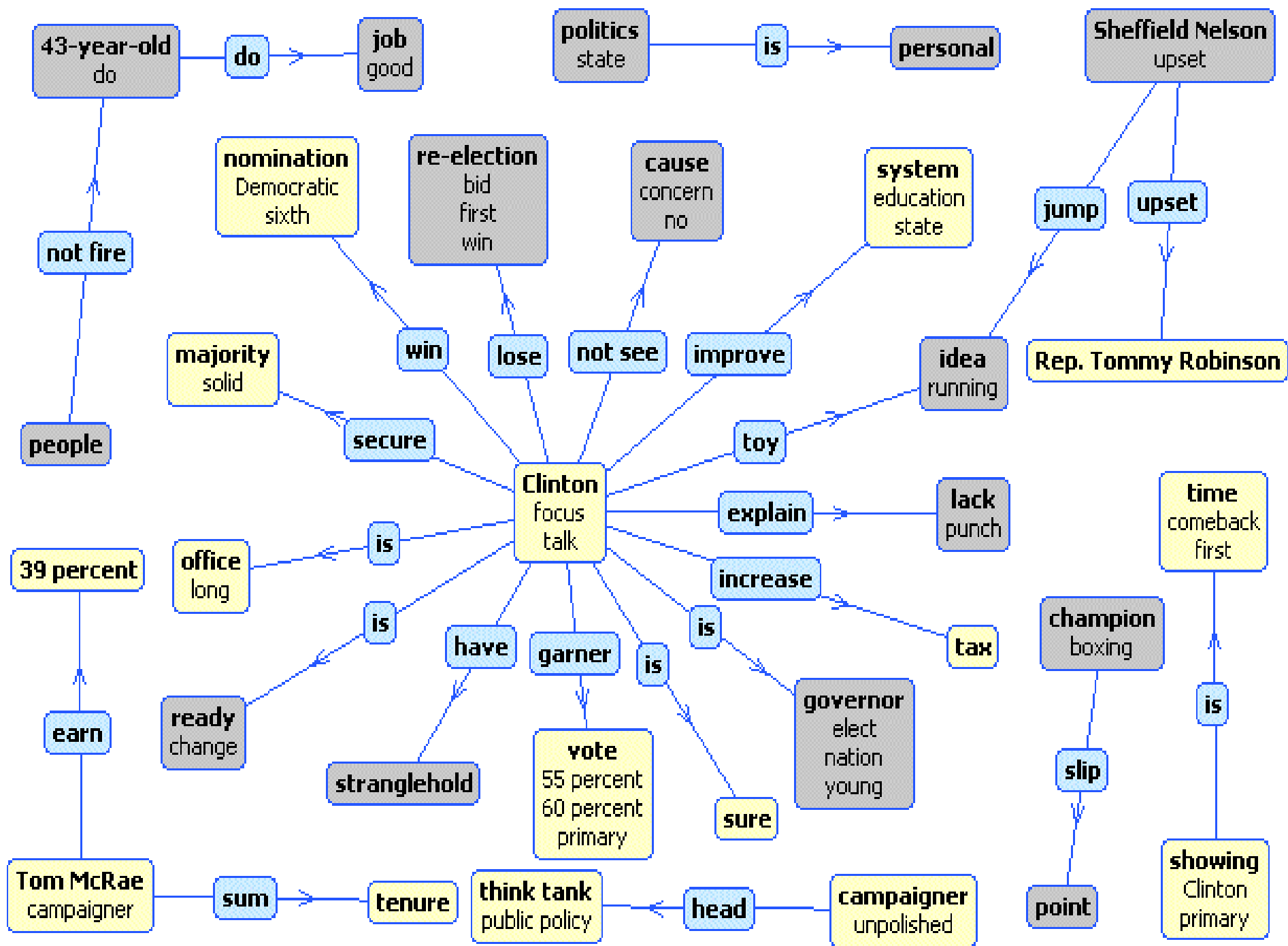


# Automatically generated summary graph



# More examples







# Question Answering

<http://AnswerArt.net>

# Asking a question in natural language

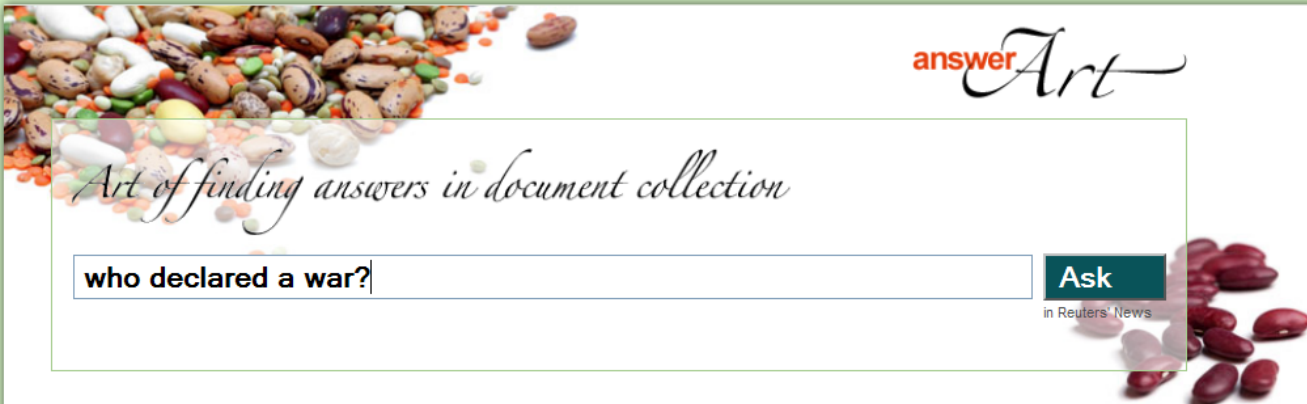
AnswerArt - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://answerart.net/ Google

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

AnswerArt



*Art of finding answers in document collection*

who declared a war?

Ask  
in Reuters' News

Get the answers for your questions.

*We found that*

Eddie Murphy did the following

played characters, roles, League, season, remainder, baseball, man, Chandler Jarrell, drums, negotiator, Ed, prince, dragon, Clarence,

Try this:

Who declared war?  
Did Bavaria declare war?  
What is an embargo?  
Who is Lenin?  
Where do birds fly?

Broaden your knowledge with related information.

*Find related information by asking*

What did Eddie Murphy perform ?  
Who performed stand-up ?  
What did Eddie Murphy do to stand-up ?  
Who is Eddie Murphy ?  
Who performed something ?  
Who did something to Eddie Murphy ?

Try this:

Who needs some sunshine?  
What did Clinton say?  
Where did an earthquake happen?  
What do people drink?  
Was Clinton visiting Venezuela?

Make sense of a document contents in a second.

AnswerArt - Document Overview example

Try this:

Italian film director Ferreri dies in Paris.  
Russian parliament opposes efforts to bury Lenin.  
American Air jet turned back by bird in engine.  
Bavaria declares war on canned beer.  
Argentine weekly official crop progress.

About AnswerArt - Contact - Terms of Use - Interactive Document Analysis

Done

# ...result set is set of mentions of potential answers

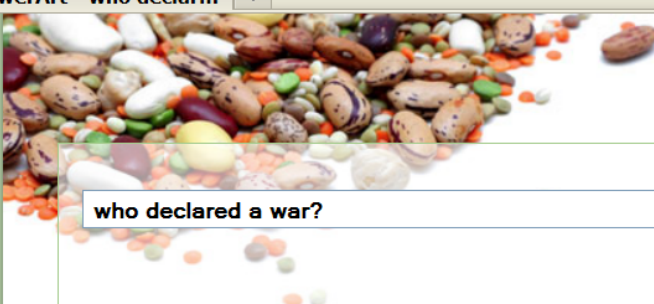
AnswerArt - who declared a war? - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://answerart.net/qa/?q=who+declared+a+war%3F

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace


AnswerArt - who declar...

 answerArt

who declared a war? **Ask**

**We found that**

the following	declared	war
Israel	declaring, declares, declared	war, Middle East war
France	declared	war, Gulf War
plan	announced	Gulf War
Republican Army	announced, declare, declares	war, guerrilla war
president	declared	war
state	declaring	war
government	declares, declaring	war
We	declared	war
President Boris Yeltsin	declared	war

 Related documents

**Israel** EGYPT: Arafat says Israel declares war on Palestinians. Israel is declaring war against the Palestinian people.

**Israel** EGYPT: Arafat says Israel declares war on Palestinians. Arafat says Israel declares war on Palestinians.

**Israel** ISRAEL: Arafat aide says Israel has declared war. Arafat aide says Israel has declared war.

**Israel** EGYPT: Moslem cleric says force last option for Jerusalem. Israel annexed East Jerusalem and declared all of the city its eternal and indivisible capital after the 1967 Middle East war.

**Israel** ISRAEL: Israel upbeat, PLO glum after U.S.-brokered talks. Israel annexed East Jerusalem and declared all of the city its eternal and indivisible capital after the 1967 Middle East war.

Done

# Document browser with summarization

AnswerArt - Document Overview - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://answerart.net/docview/?view=1&id=19970802\810803newsML.xml

Most Visited Getting Started Latest Headlines Customize Links Windows Marketplace

AnswerArt - Document ...

**DOCUMENT OVERVIEW**

**FACTS**

- Israel declares war
- Saturday Israel declared war
- Saturday Israel declared Palestinians
- west Jerusalem took wake
- suicide bombing west Jerusalem
- President Yasser Arafat told reporters
- President Yasser Arafat told talks

**SUMMARY**

Palestinian President Yasser Arafat said on Saturday Israel had declared war on Palestinians rather than "terrorism" by a series of measures it took in the wake of a double suicide bombing in west Jerusalem. The Israeli government used this situation and this incident (the suicide bombing) to declare war on the Palestinian people, their National Authority and their leadership instead of declaring it against terrorism," Arafat told reporters after talks with Egyptian President Hosni Mubarak.

Arafat arrived in Egypt earlier on Saturday and went straight into talks with Mubarak on the crisis in the Middle East peace talks, Egyptian state television said.

"I told President Mubarak about all the dangerous measures announced by Israel after the terrorist operation that happened recently in Jerusalem," Arafat said. He has since returned to Gaza.

Arafat described as "collective punishment" the measures taken by Israel since the suicide bombing and said Mubarak told him he would "begin to decrease the suffering of the Palestinian people," but did not give details.

Egypt has been trying to jump-start Israeli-Palestinian peace talks which collapsed in March after Israel began to build a new Jewish settlement in Arab East Jerusalem.

Arafat, who met Mubarak in the Egyptian leader's private residence in Burg el-Arab, near the Mediterranean city of Alexandria, was accompanied by senior Palestinian officials Saeb

Done

# Cyc Knowledge Base and Reasoning



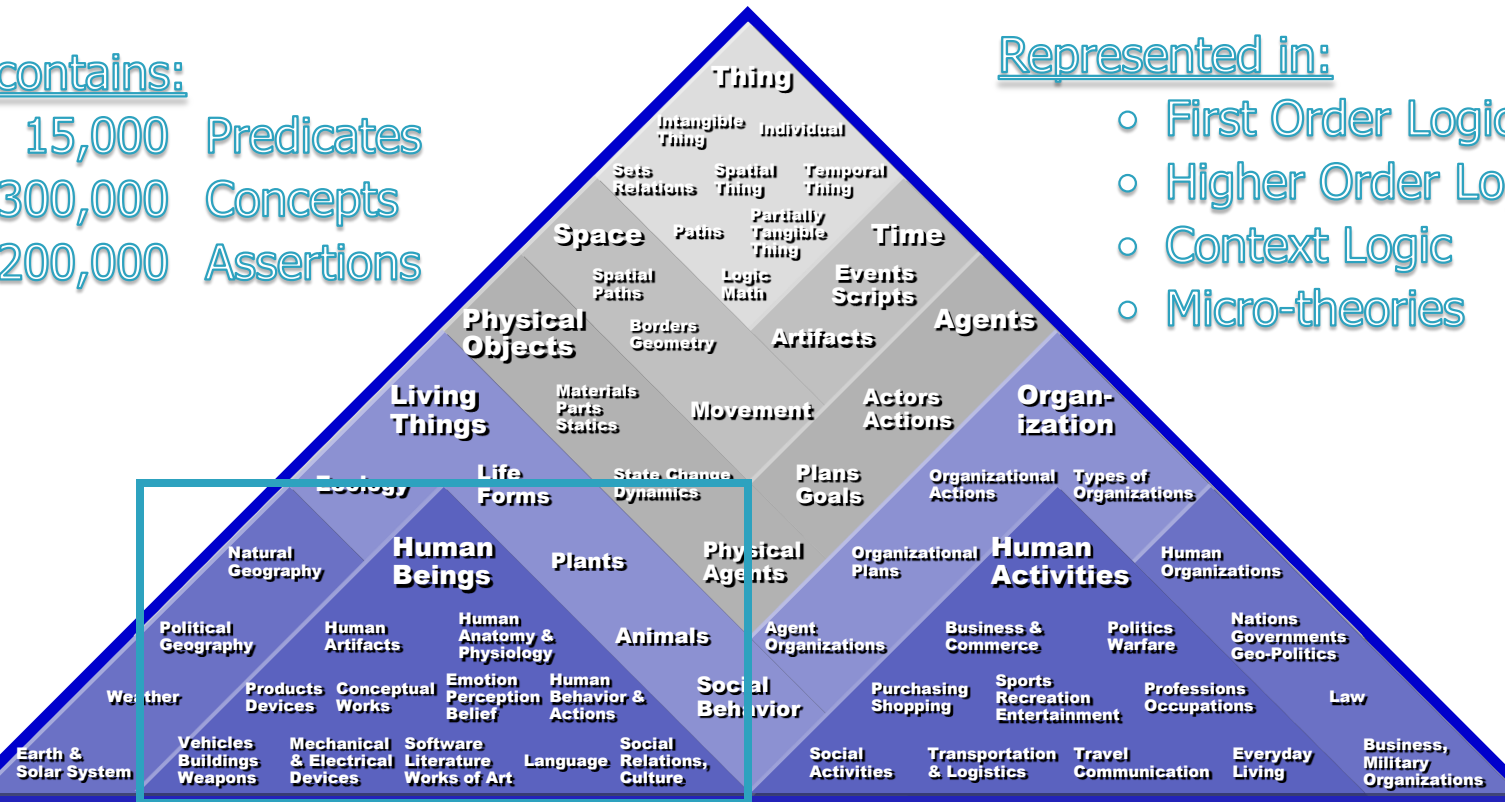
# The Cyc Ontology

## Cyc contains:

15,000 Predicates  
300,000 Concepts  
3,200,000 Assertions

## Represented in:

- First Order Logic
- Higher Order Logic
- Context Logic
- Micro-theories



**General Knowledge about Various Domains**

**Specific data, facts, and observations**



## Cyc's front-end: "Cyc Analytic Environment" – querying (1 / 2)

Task Info Document Search Concepts Related-to Query Creator Queries

Find Stop

WHO had a motive for the assassination of Hariri.

Continue  
Save  
New Tab  
Reset

5 answers  
Timed out

Text query

Query (semi) automatically translated in the First Order Logic

Answers (5)

Answer	Speculation Level	Sources
Bashar al-Assad	No Speculation	W CNN
Syria	Mildly Speculative	CNN
al Qaeda	Moderately Speculative	SAIC CNN
United States, the	No Speculation	2
Israel	No Speculation	2

Justify Fact Sheet Visualize Visualize All

Status: Finished Message: No appropriate visualizations found

Search Results

Cyc Query Library

Overviews

Pre-formed Examples

Example Hezbollah Queries

Actions by Specific Terrorist Groups

Support for/by Terrorist Groups

Terrorist Group Membership/Leadership

Terrorist Acts in a Specific Region

Terrorist Group Areas of Operation

Terrorist Group Tactics

Terrorist Group Suborganizations

Casualties in Terrorist Attacks

Affiliations with Terrorist Agents

Terrorist Group Ideologies

Terrorist Attack Targets

Weapons Used in Terrorist Attacks

List all bombings after 2000 and before September 2004 that used pipe bombs.

Motives for Terrorist Attacks

Who has a motive for the assassination of Rafik Hariri?

Locations of Terrorist Attacks

Who has a motive for the assassination of Rafik Hariri?

is claimed responsibility.

committed by the IRA?

entage of kidnapping attacks

attacks that are of that type?

ated by Hamas?

mbings?

by terrorist groups that operate in

Afghanistan.

List the Islamic Jihad organization with the highest total wounded in their attacks.

Which group perpetrated the largest hostage-taking in Europe?

List the terrorist group that operates out of Pakistan that has the highest casualty count.

What was the deadliest suicide attack?

What is the shortest length of time between events performed by suborganizations of Hezbollah?

What is the longest duration of time between events performed by suborganizations of Hezbollah?

What Islamic Jihad Organization killed the most people?

What is the ratio of Hamas attacks that target buildings that are performed in Israel?

What is the number of people killed in terrorist attacks in Syria?

What percentage of the suicide attacks in Israel are performed by Hamas?

Give the total number of people wounded in attacks by the terrorist group People Against Gangsterism And Drugs.

What is the number of suicide bombings that occur in Beirut?

January 2006 Analyst Session

Links between entities

AKB-SME research

General Purpose


## Cyc's front-end: "Cyc Analytic Environment" – justification (2/2)


Task Info Document Search Concepts Related-to Query Creator Queries Justification Justification Justification


Proof 1 Save... Copy

▼ Query: Who or what had a motive for the assassination of Hariri? ← Query & Answer  
 Answer: al Qaeda

▼ Because:

Since 2000, Lebanon has been responsible for according with Lebanese economic reform.  1

February 14, 2005 was the date of the assassination of Hariri.  2 ← Justification

Rafik Hariri was killed during the assassination of Hariri.  2  
Rafik Hariri is an advocate of Lebanese economic reform.

Al Qaeda opposes Lebanese economic reform.

▼ Detailed Justification:  
 ► Al Qaeda had a motive for the assassination of Hariri.

▼ External Sources:

► Gary C. Gambill, "Dossier: Rafiq Hariri", *United States Committee for a Free Lebanon*, July 2001, [http://www.meib.org/articles/0107\\_id1.htm](http://www.meib.org/articles/0107_id1.htm). ← Sources for Reasoning and Justification

► "Huge blast kills Lebanese ex-PM", *the Cable News Network*, February 14, 2005, <http://www.cnn.com/2005/WORLD/meast/02/14/beirut.explosion.1910/>.

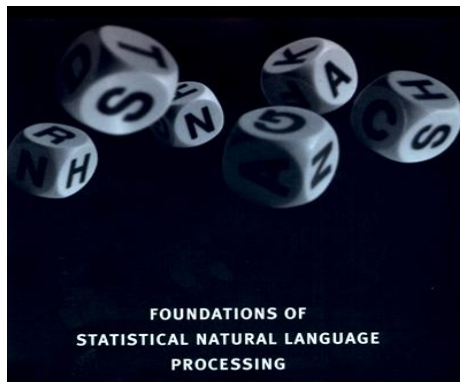
► If

- some intelligent agent opposes some policy,
- and some other intelligent agent *VICTIM* is an advocate of that policy,
- and some other intelligent agent *ADOPTER* is responsible for according with the policy,
- and it is adopted by *ADOPTER* in any some *ADOPT-TYPE*,
- and some *ACT* prevents *VICTIM* from playing the role "key participant" in any *ADOPT-TYPE*,

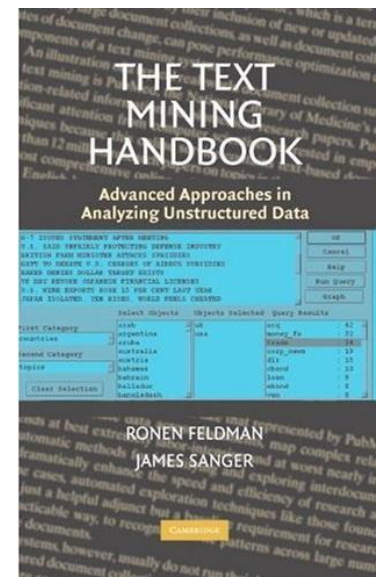
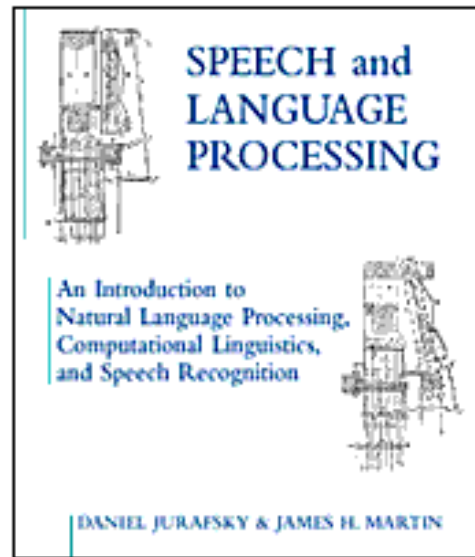
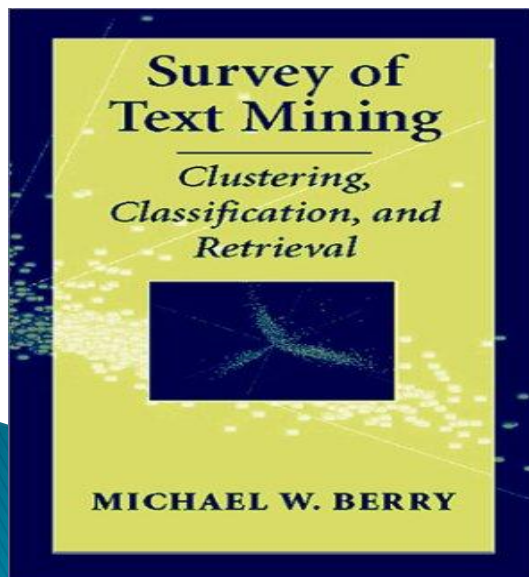
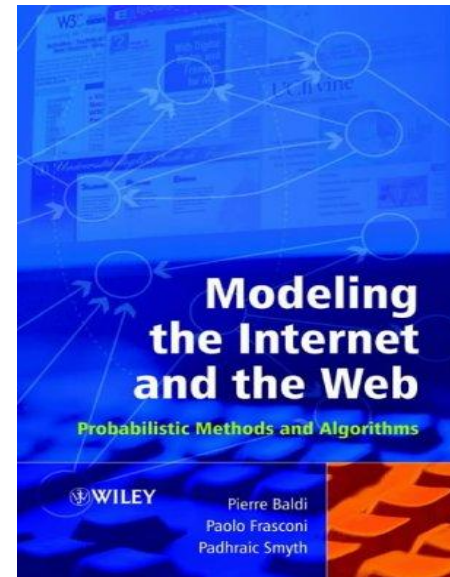
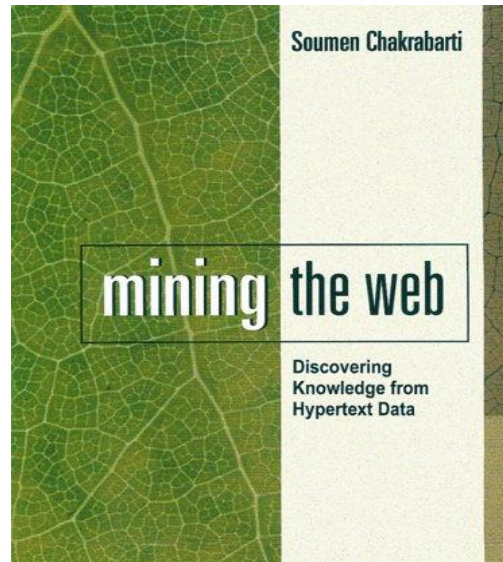
then that intelligent agent had a motive for *ACT*.

# Further references...

# References to some Text-Mining books

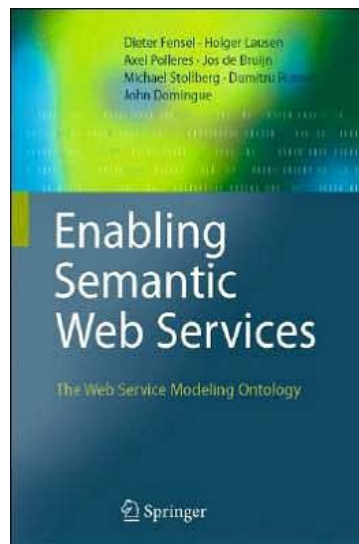
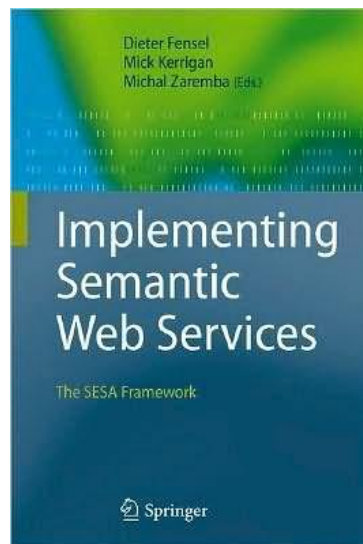
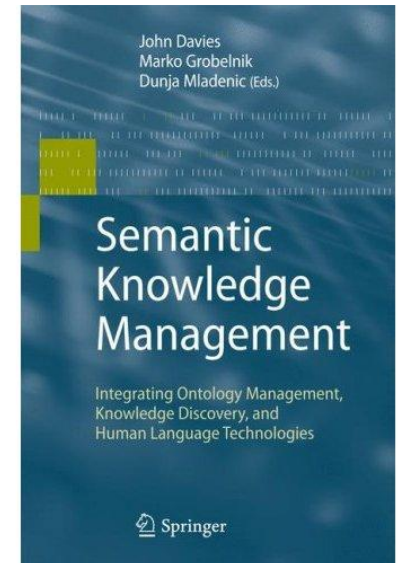
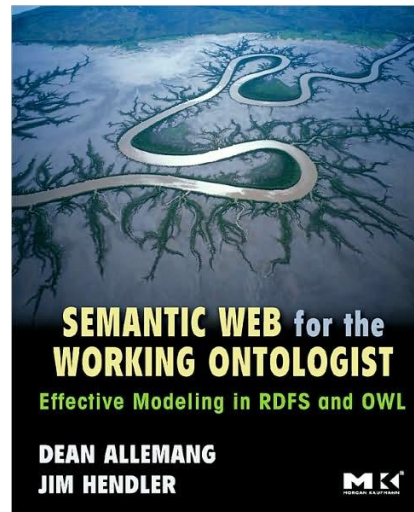
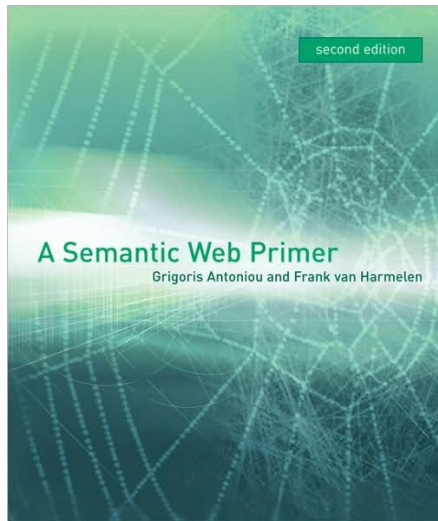


CHRISTOPHER D. MANNING AND  
HINRICH SCHÜTZE



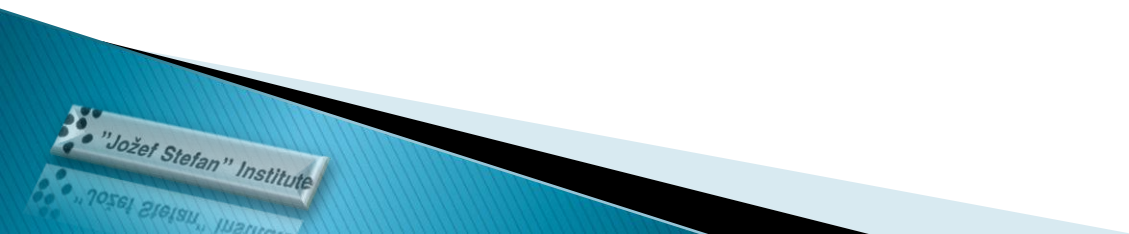


# Books on Semantic Technologies



# References to the main conferences

- ▶ Information Retrieval:
  - SIGIR, ECIR
- ▶ Machine Learning/Data Mining:
  - ICML, ECML/PKDD, KDD, ICDM, SDM
- ▶ Computational Linguistics:
  - ACL, EACL, NAACL
- ▶ Semantic Web:
  - ISWC, ESWC, ASWC





# Videos on Text and Semantic Technologies

[http://videlectures.net/Top/Computer\\_Science/Semantic\\_Web/](http://videlectures.net/Top/Computer_Science/Semantic_Web/)

- ▶ Recorded tutorials, workshops, conferences, summer schools available from <http://videlectures.net>

The screenshot shows the VideoLectures website in a Mozilla Firefox browser window. The browser's address bar displays <http://videlectures.net/>. The website's header includes the title "VideoLectures - exchange ideas & share knowledge - Mozilla Firefox" and a navigation menu with links: File, Edit, View, History, Bookmarks, Tools, Help. The main content area features a search bar, a welcome message for "Davor", and a list of featured lectures. Each lecture entry includes a thumbnail image, a title, a brief description, and view/comment statistics. The website also has sections for "RECENT EVENTS", "NEWS", and "CATEGORIES". The "RECENT EVENTS" section highlights the "ISWC '08 - Karlsruhe" and "18.085 Computational Science and Engineering I". The "NEWS" section mentions the "MIT OpenCourseWare Collection" and "Cambridge University Engineering Department - Machine Learning seminars". The "CATEGORIES" section lists various fields such as Architecture, Arts, Biology, Business, Chemistry, Computers, Computer Science, Economics, Education, Environment, Events, History, Law, Mathematics, Medicine, Philosophy, Physics, Psychology, Science, Society, and Technology. The "FEATURED" section includes a video titled "Universal Access to All Knowledge - Archive.org".

VideoLectures - exchange ideas & share knowledge - Mozilla Firefox

File Edit View History Bookmarks Tools Help

<http://videlectures.net/>

Most Visited Getting Started Latest Headlines

Welcome, Davor | [Admin](#) | [Trac \(my\)](#) | [Log out](#) | [Google](#)

202 events, 3705 authors, 4191 lectures, 5755 videos

**VIDELECTURES.net**  
EXCHANGE IDEAS / SHARE KNOWLEDGE

[HOME](#) [MOST POPULAR](#) [LATEST LECTURES](#) [CATEGORIES](#) [EVENTS](#) [PEOPLE](#) [INTERVIEWS](#) [TUTORIALS](#) [ABOUT US](#)

**FEATURED LECTURES:**

- POWERSET - Natural Language and the Semantic Web**  
Barney Pell | 1461 views, 01:09:28 | 2 comments
- Leonardo: Goal assistance with divergent beliefs**  
Jesse Gray, Matt Berlin, Cynthia Breazeal | 346 views, 00:03:31
- Bayesian Kernel Methods**  
Alexander J. Smola | 164 views, 04:32:50
- Functional Analysis in Data Modelling**  
Tony Dodd | 90 views, 01:06:03
- Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology**  
Jarno Vanhatalo | 42 views, 00:05:51

**RECENT EVENTS:** [more](#)

- ISWC '08 - Karlsruhe**  
7th International Semantic Web Conference  
ISWC is a major international forum where visionary and state-of-the-art research of all aspects of the Semantic Web are presented. ...
- MIT 18.085 Computational Science and Engineering I - Fall 2007**  
This course provides a review of linear algebra, including applications to networks, structures, and estimation, Lagrange multipliers. Also covered are: ...
- ESTC '08 - Vienna**  
2nd European Semantic Technology Conference  
The 2008 European Semantic Technology Conference is the only European conference that brings together a global community of ...

**NEWS:**

- MIT OpenCourseWare Collection**  
**MITOPENCOURSEWARE**  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
We are excited to announce that we have started a video educational collaboration with MIT OpenCourseWare.  
New courses will be available every week. Check them out on the MIT OCW page.
- Cambridge University Engineering Department - Machine Learning seminars**  
 **UNIVERSITY OF CAMBRIDGE**  
We are very pleased to announce a collaboration with the Department of Engineering at the University of Cambridge to make Machine Learning @ CUED talks also available via VideoLectures.net.  
[more on > http://videlectures.net/mlcued08\\_cambridge/](http://videlectures.net/mlcued08_cambridge/)
- Carnegie Mellon Machine Learning Lunch seminar**  
 **Carnegie Mellon**  
We established the collaboration with Machine Learning Department at Carnegie Mellon University to make Machine Learning Lunch seminar talks also available via VideoLectures.net.  
[more on > http://videlectures.net/cmlls](http://videlectures.net/cmlls)

**CATEGORIES:**

- Architecture (2)
- Arts (24)
- Biology (38)
- Business (70)
- Chemistry (12)
- Computers (16)
- Computer Science (1375)
- Economics (9)
- Education (4)
- Environment (12)
- Events (35)
- History (2)
- Law (12)
- Mathematics (73)
- Medicine (29)
- Philosophy (7)
- Physics (15)
- Psychology (24)
- Science (41)
- Society (33)
- Technology (8)

**FEATURED:** [more](#)

- Universal Access to All Knowledge - Archive.org**  
172 views, 00:26:53

Transferring data from maps.amung.us...