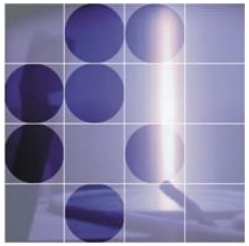
How to Represent Text?
...from Characters to Logic

Marko Grobelnik (marko.grobelnik@ijs.si)
Jozef Stefan Institute (http://www.ijs.si/)
Slovenia, Europe



STI · INTERNATIONAL

W3C WORLD WIDE WEB consortium

cYcorp

QUINTELLIGENCE

Outline

- Some Initial thoughts
  - …quick example why representation matters
- How we represent Text?
  - Big picture
  - Levels of representation:
    - Lexical
    - Syntactic
    - Semantic
- Further references
  - …events, books, videos

Some Initial thoughts

- Two major steps in Machine learning:
  - Choosing representation (feature engineering)
  - Modeling (statistics+optimization)

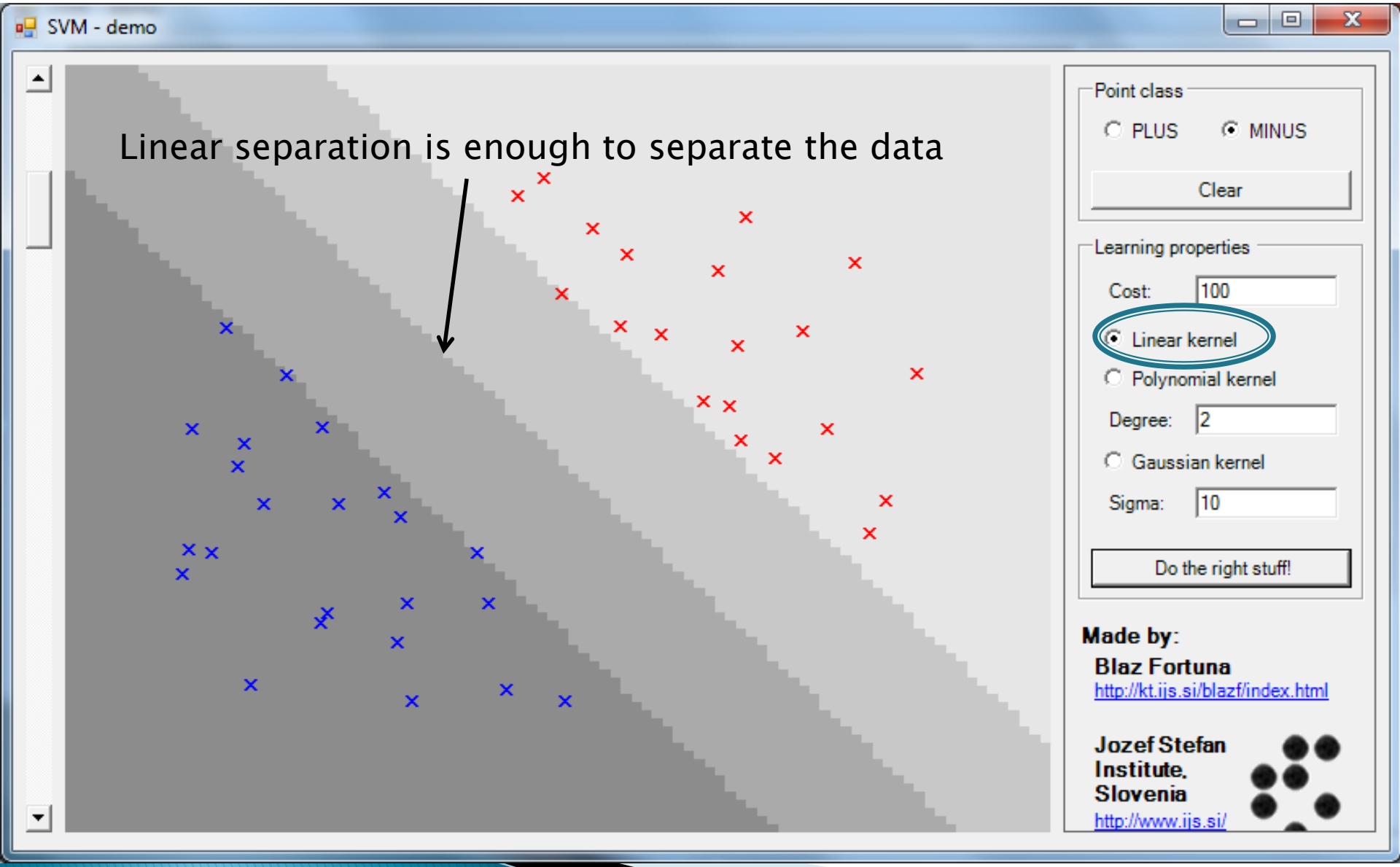
- …typically people do modeling well and often ignore data representation

- But…
  - …**good representation with bad algorithm** gives typically better results then **good algorithm with bad representation**

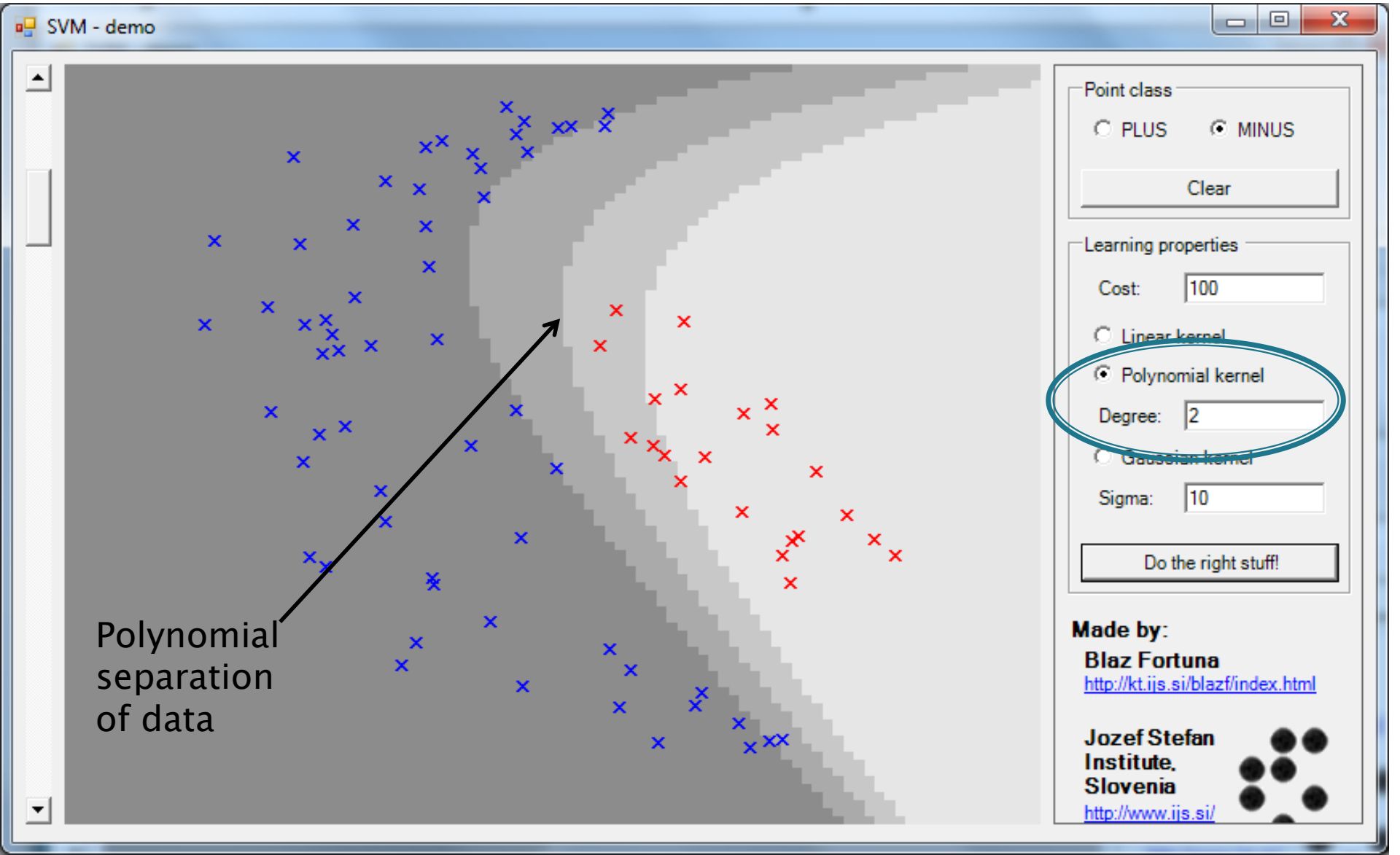
Quick example:
Why representation matters?

- Selection of kernels when using SVM (Support Vector Machine) equals selecting the right representation for data

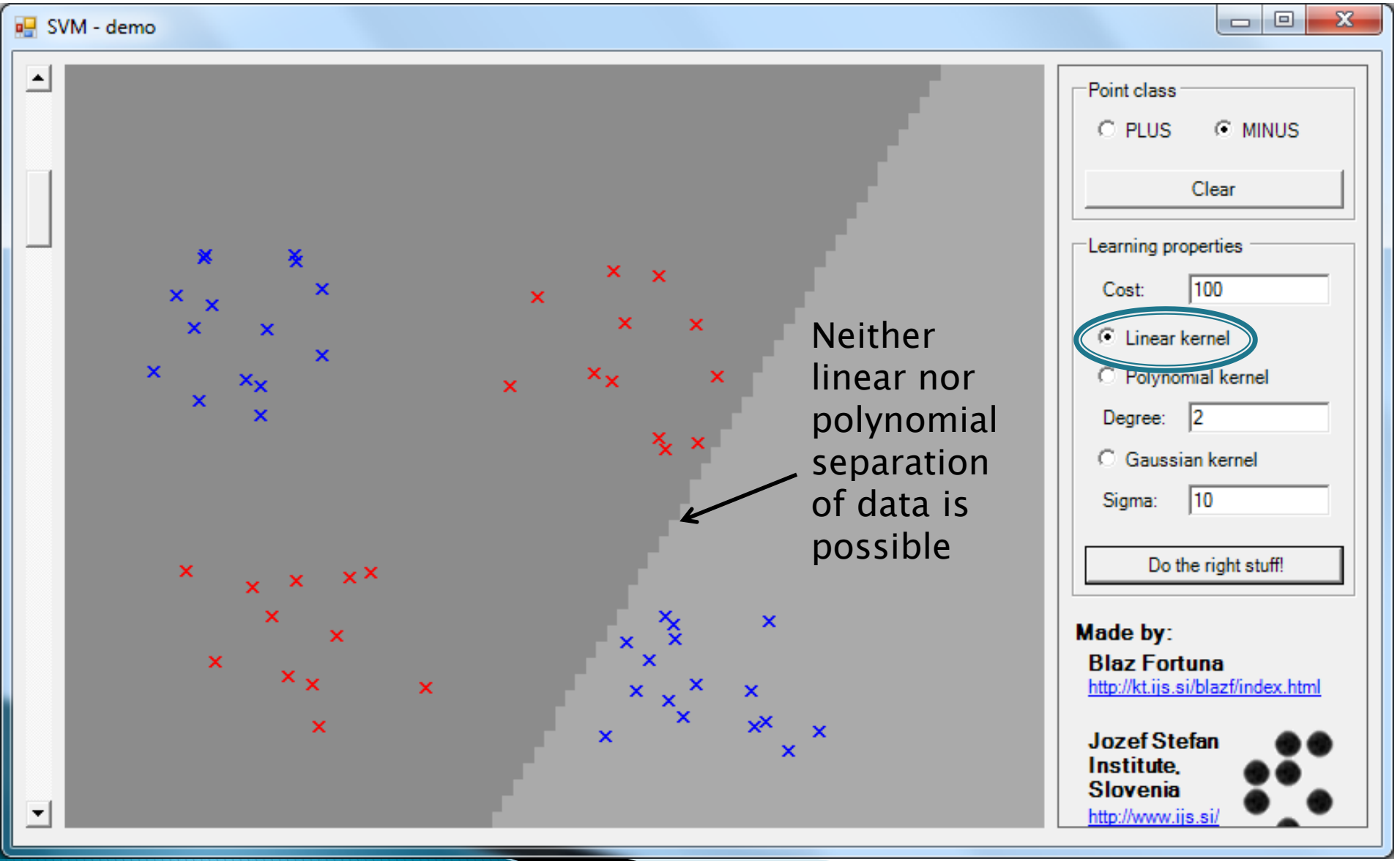
Easy decision problems require simple data representation



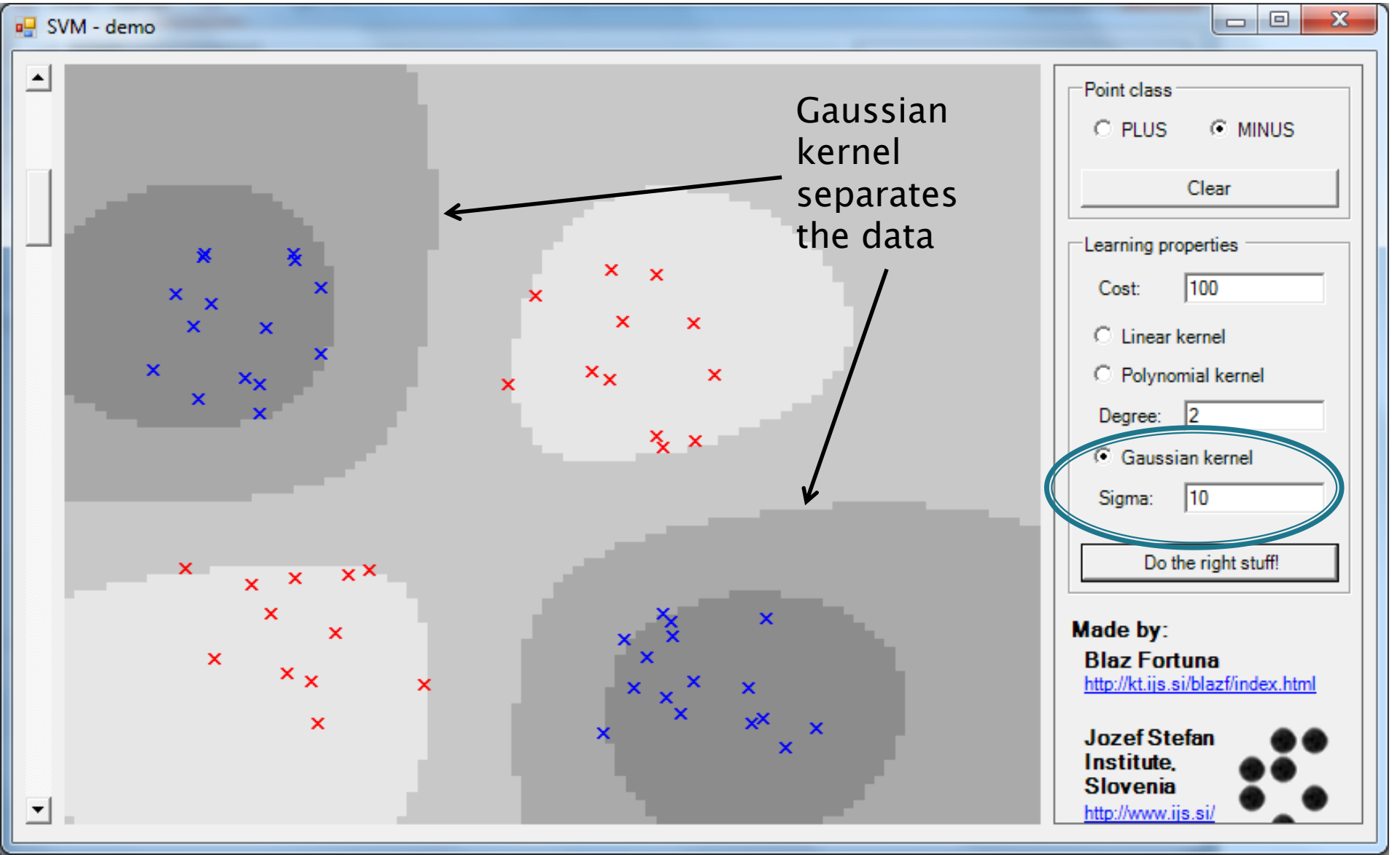
Harder decision problems require better data representation



Hard decision problems require sophisticated data representation



Hard decision problems require sophisticated data representation



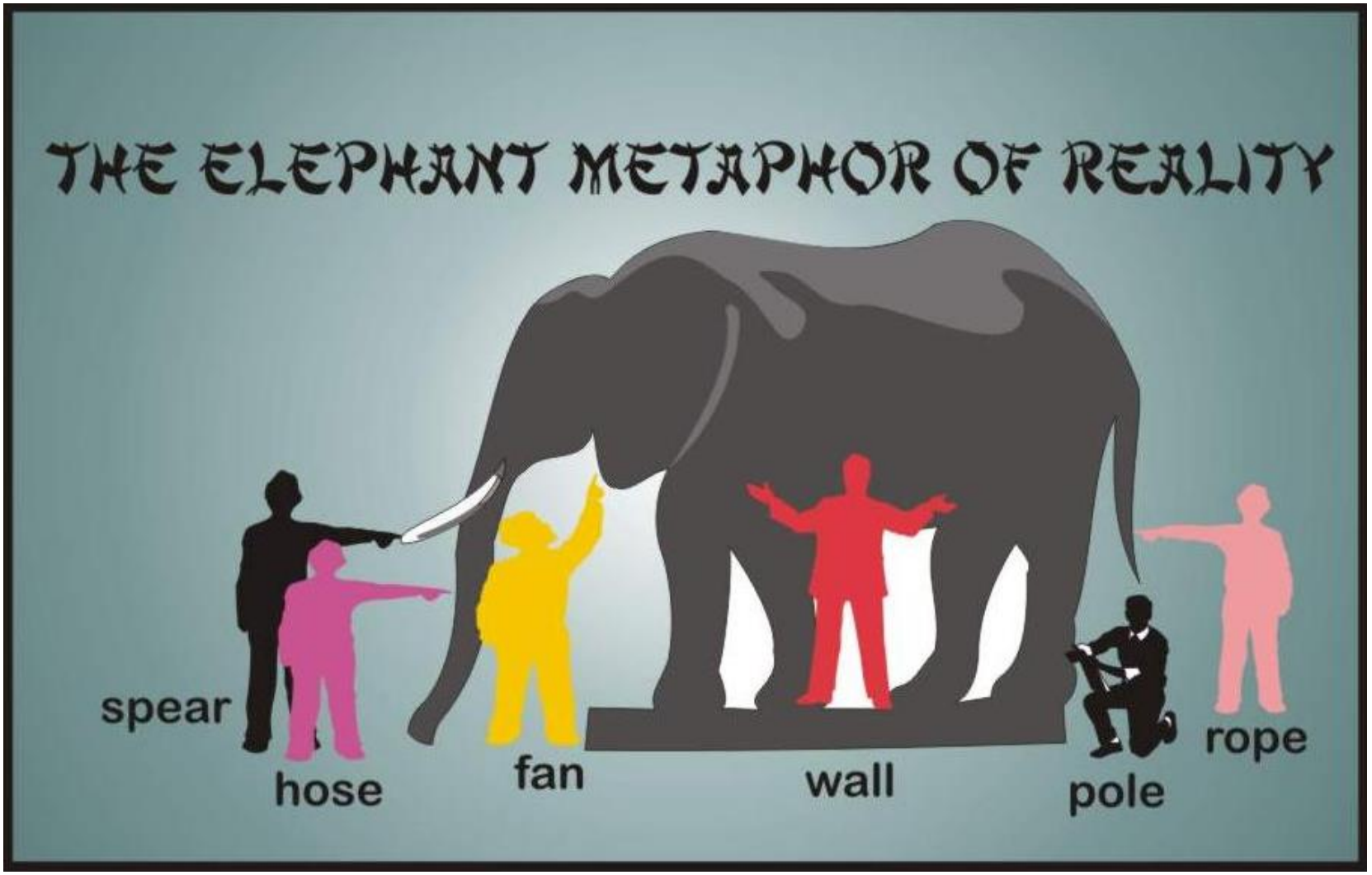
How we represent Text?

The Big picture

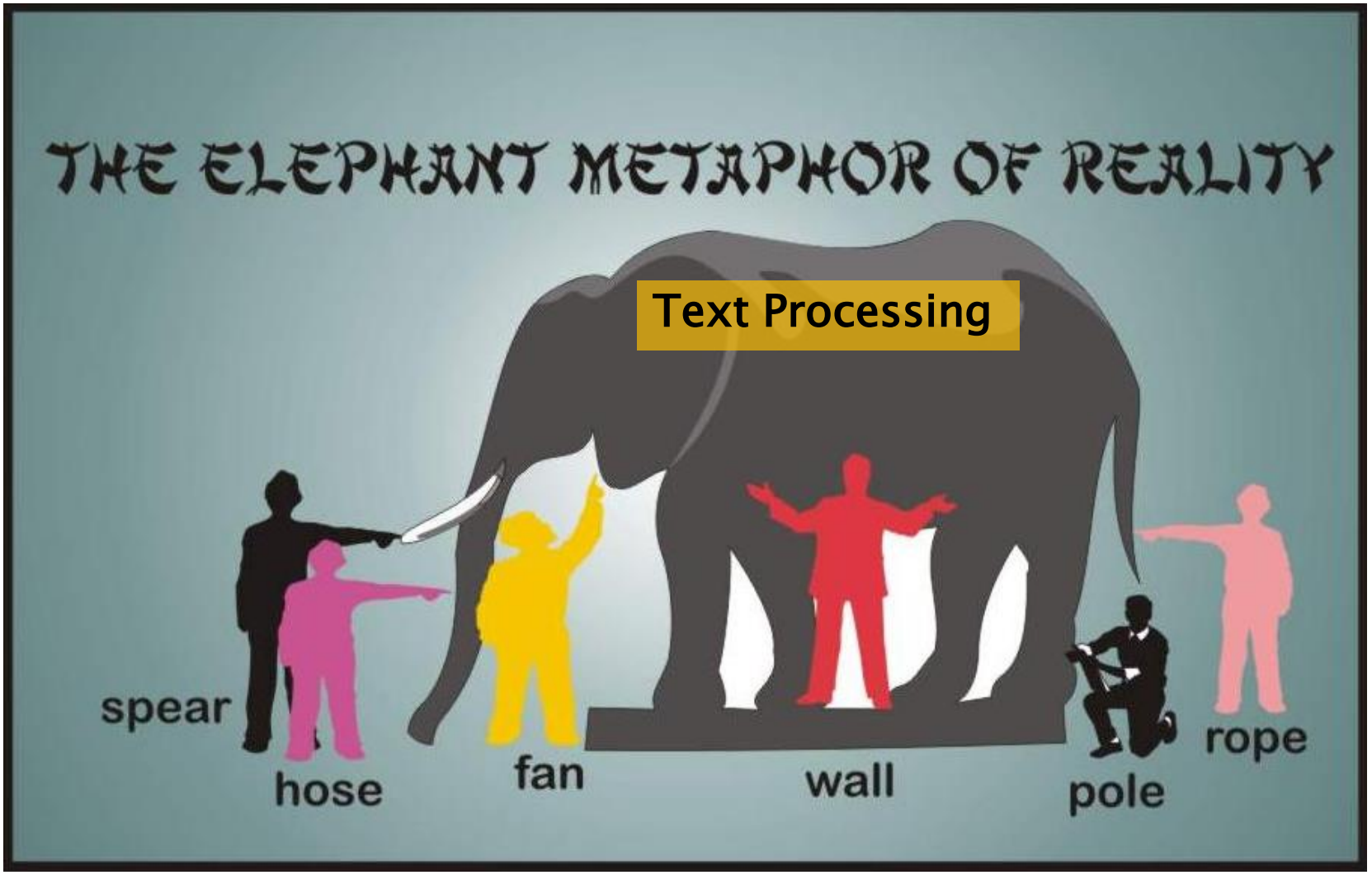
Key paradigms when dealing with data

- Three key scientific paradigms
  - **Top-down approaches (model driven)**
    - (Traditional NLP, KRR, Semantic Web)
  - **Bottom-up approaches (data driven)**
    - (Machine Learning, Data Mining)
  - **Collaborative approaches (socially driven)**
    - (Web2.0, Social Computing)

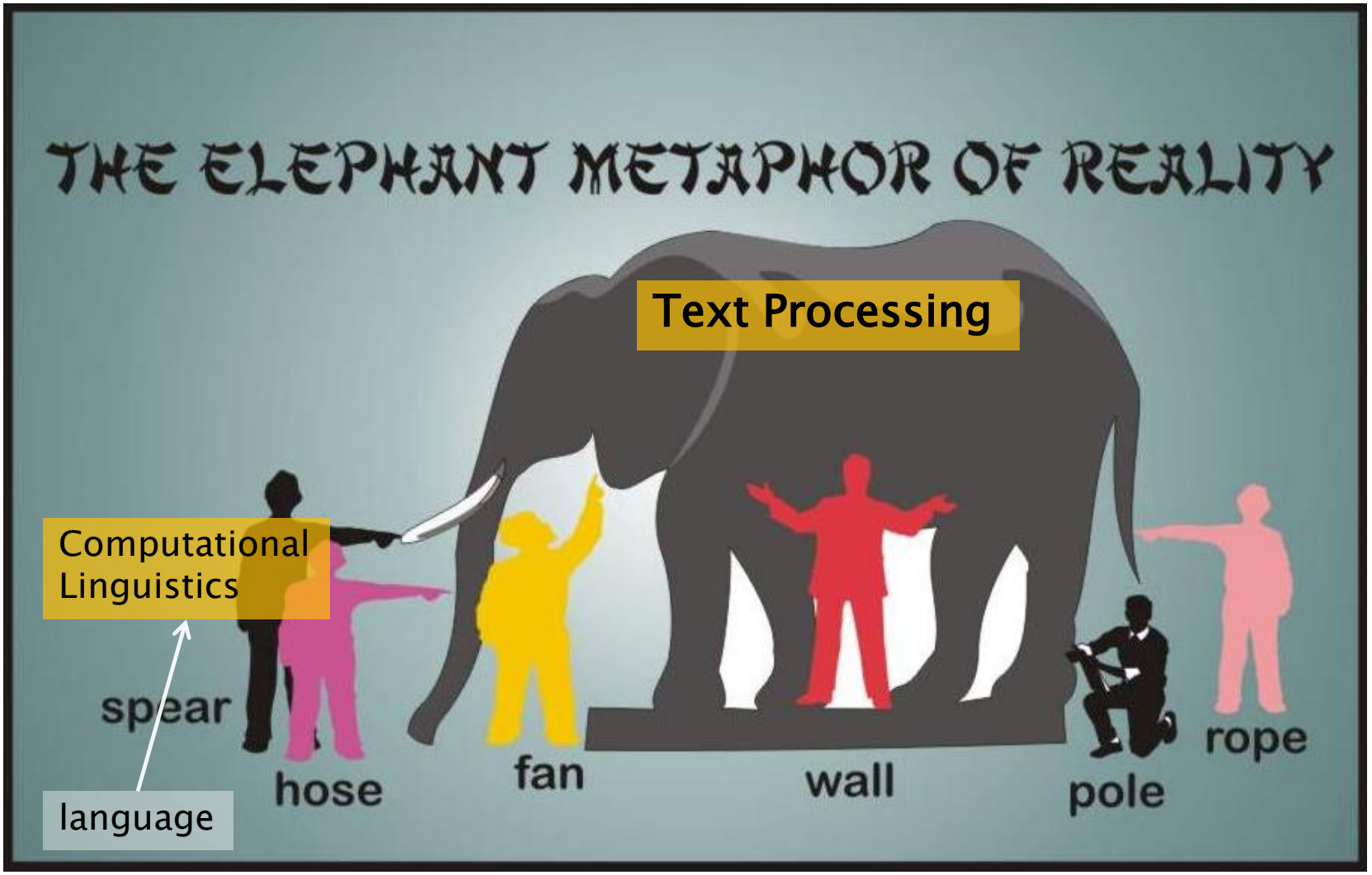
How different research areas approach text?



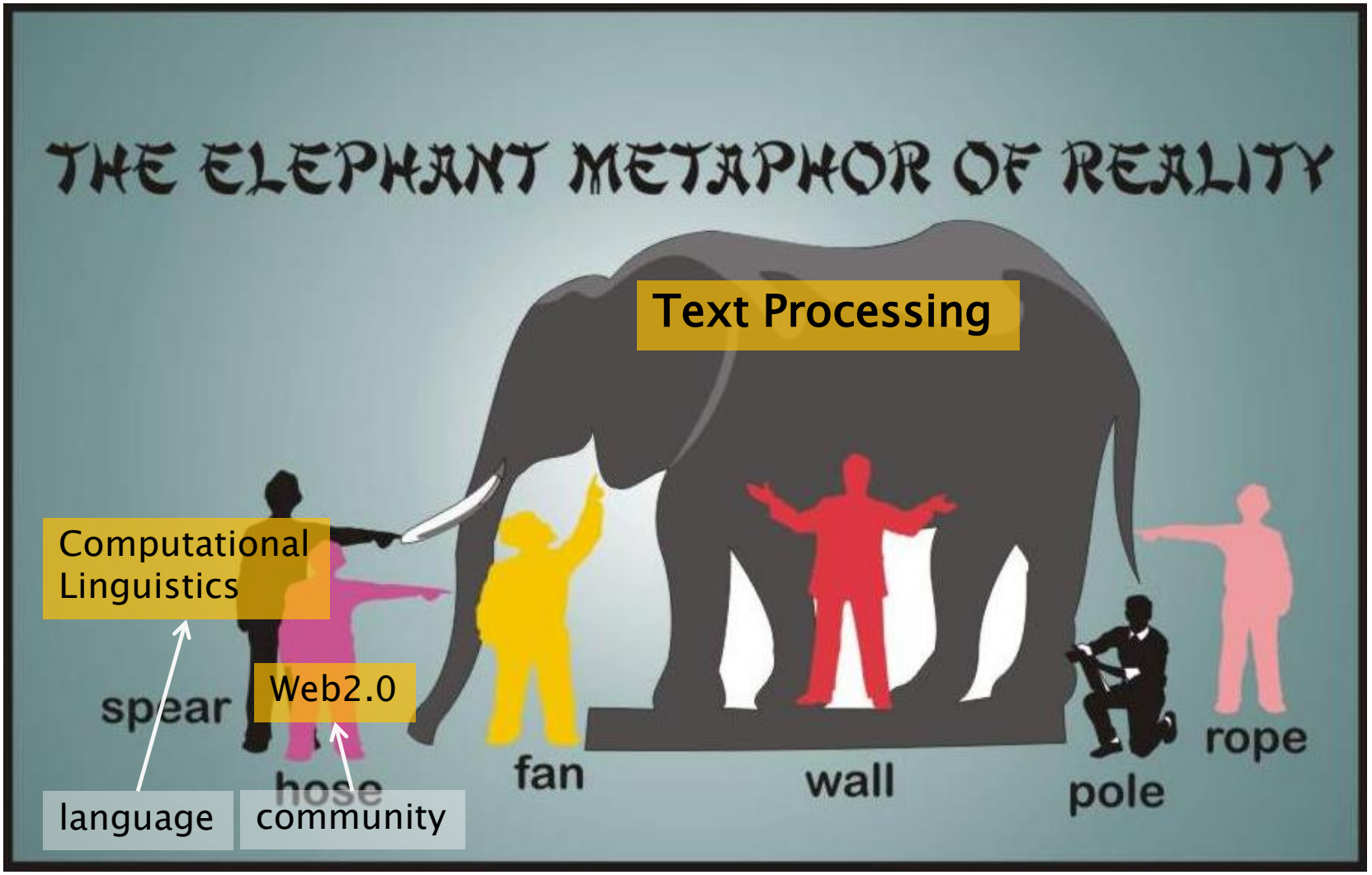
How different research areas approach text?



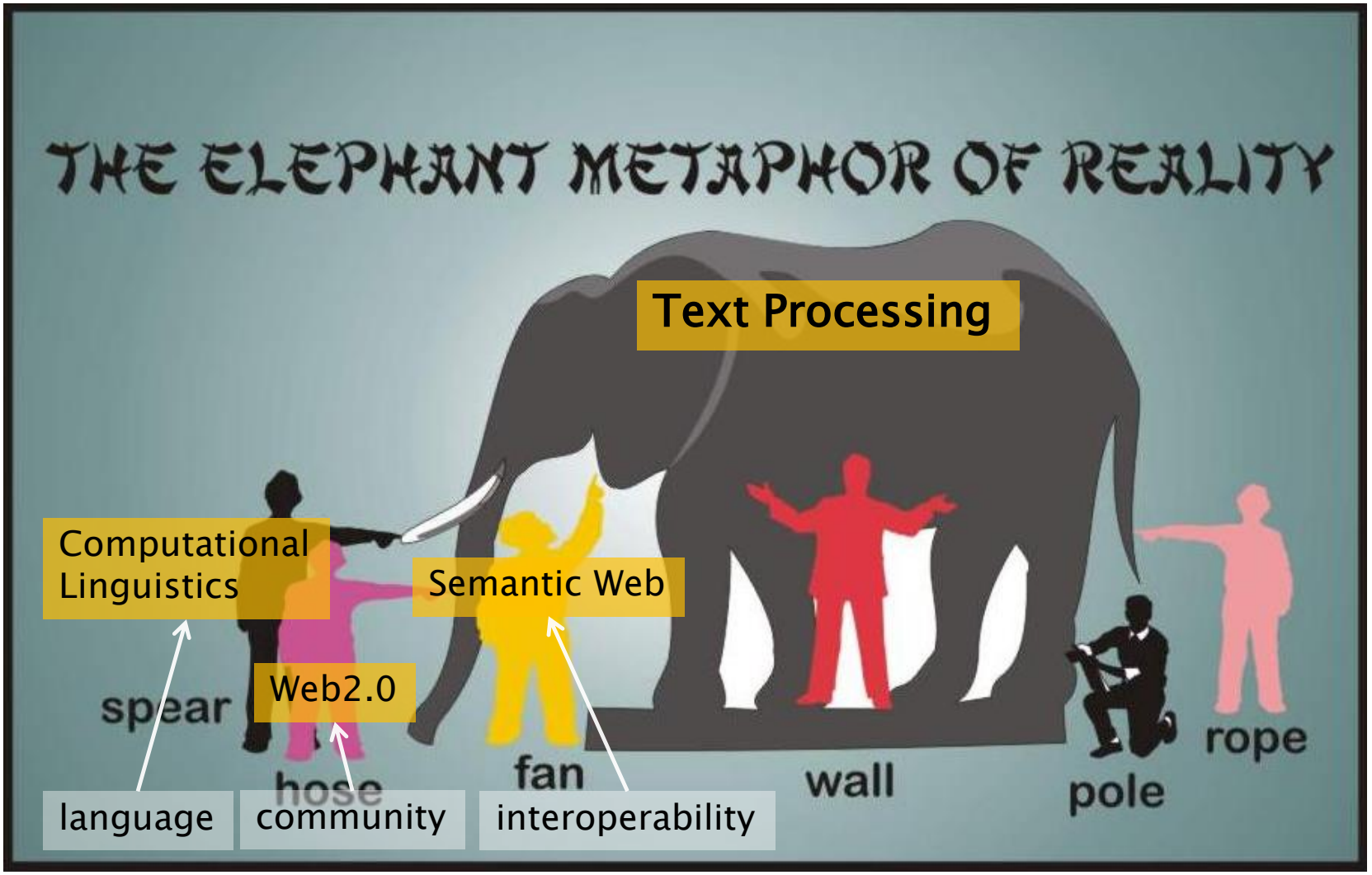
How different research areas approach text?



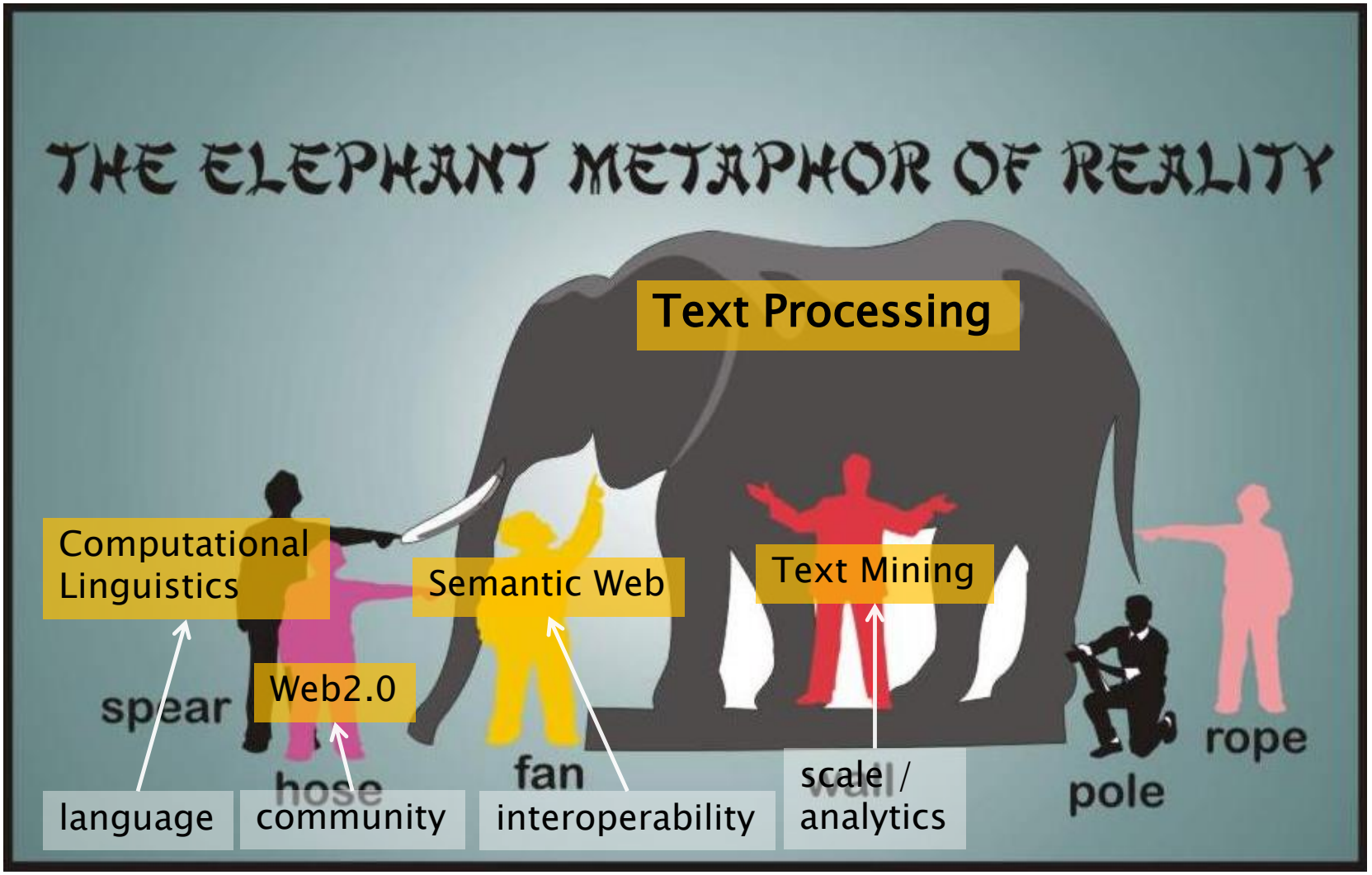
How different research areas approach text?



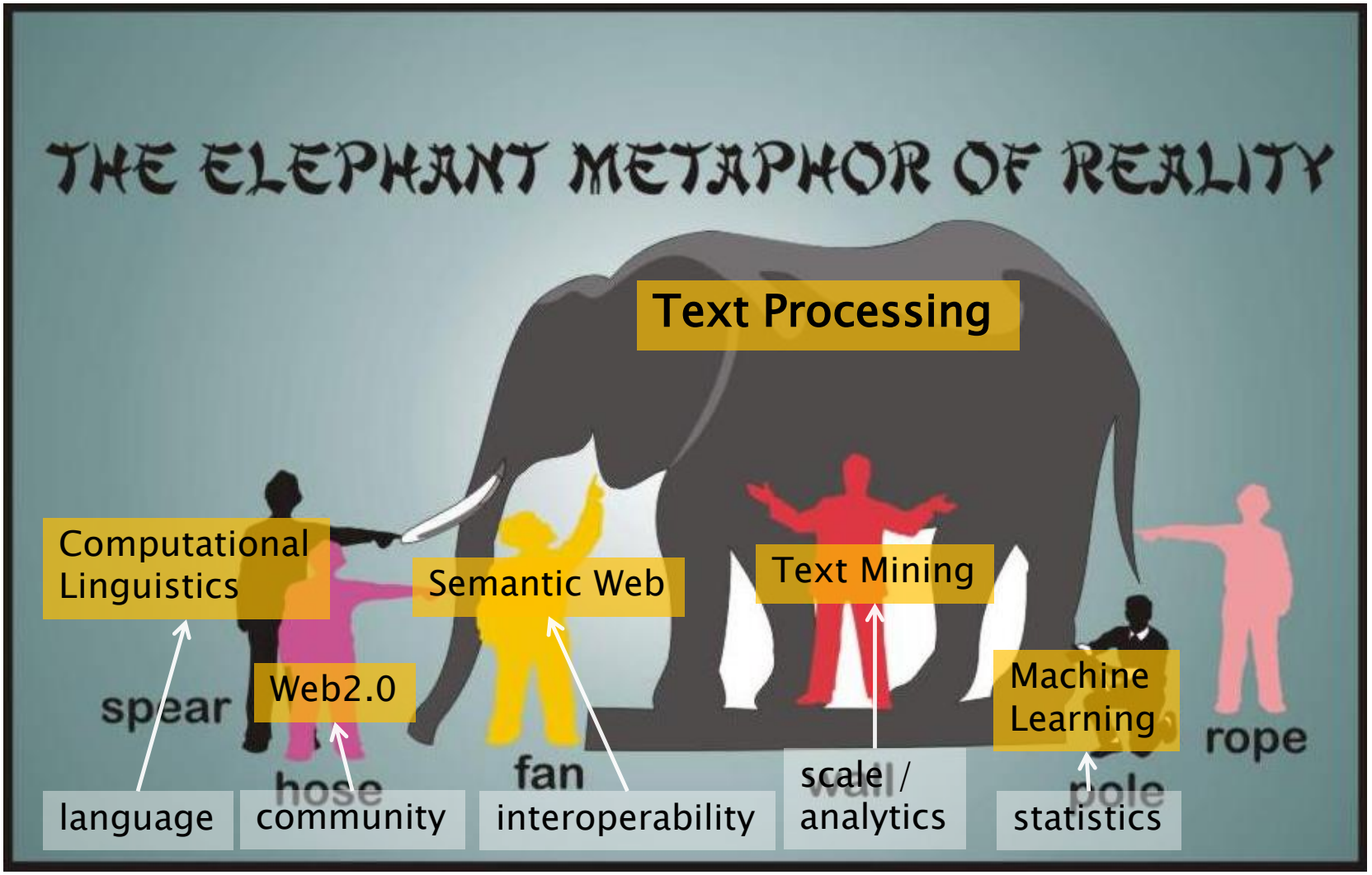
How different research areas approach text?



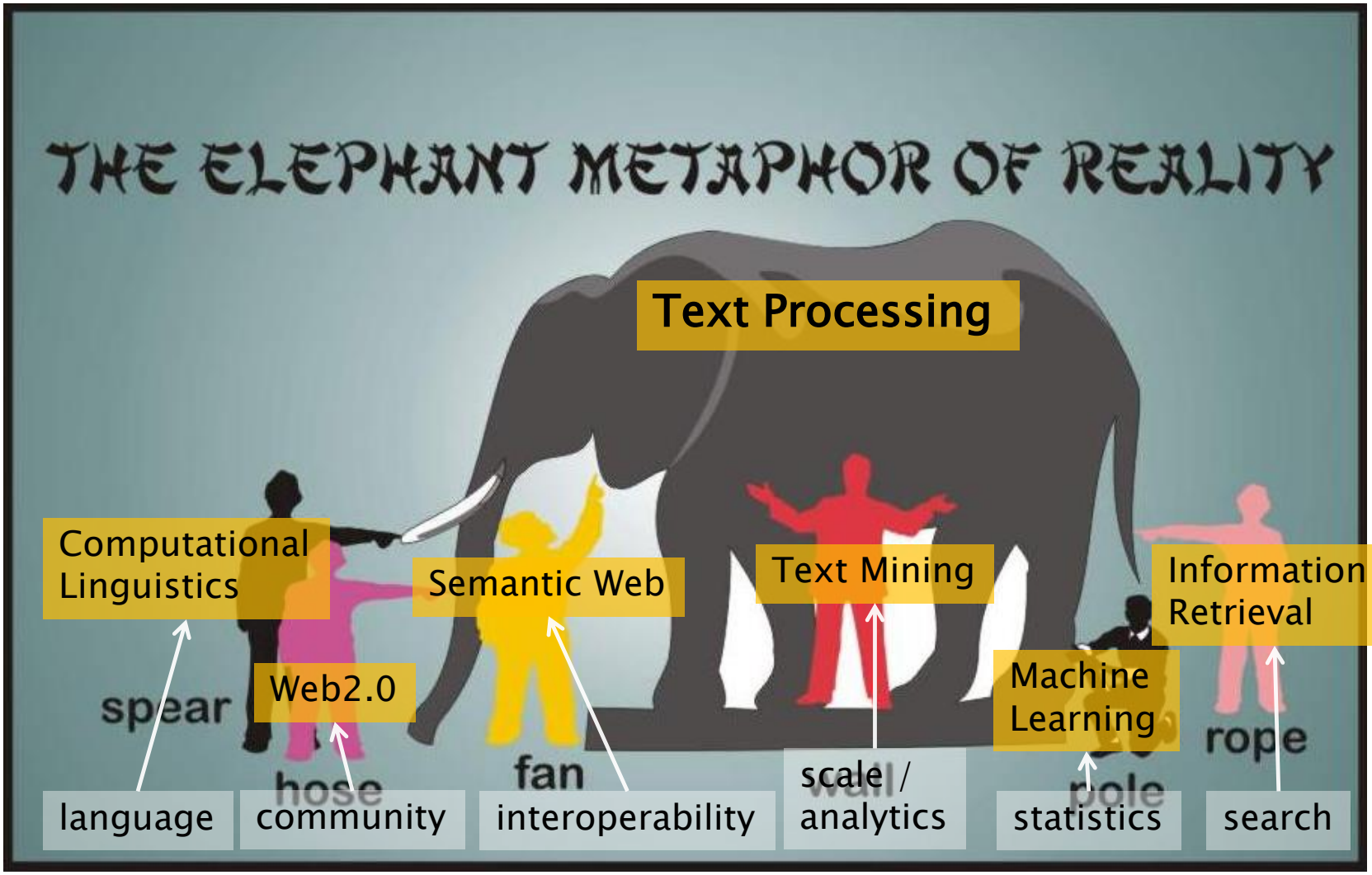
How different research areas approach text?



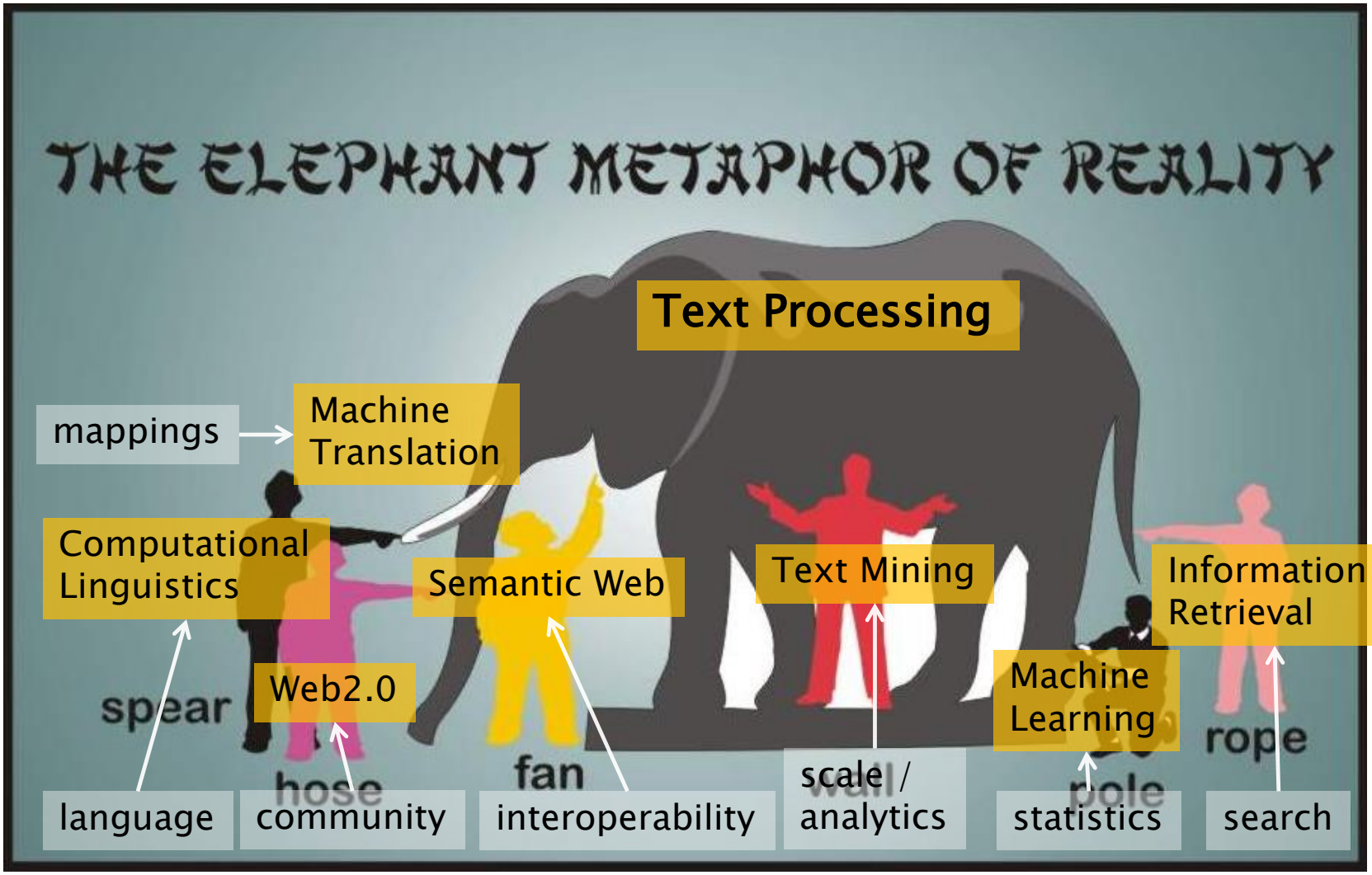
How different research areas approach text?



How different research areas approach text?



How different research areas approach text?



How do we represent text?

Levels of representation

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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Levels of text representations

Language identification, Copy detection

- Character (character n-grams and sequences)
- Words (stop-words, stemming, lemmatization)
- Phrases (word n-grams, proximity features)
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

- Vector-space model
- Language models
- Full-parsing
- Cross-modality

Syntactic

- Collaborative tagging / Web2.0
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Semantic

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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

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

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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

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

Text categorization, Clustering, Search , Summarization, …

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
- Linked Data
- 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

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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

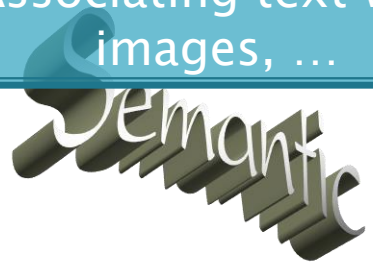
Language identification, Copy detection

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

Text categorization, Clustering, Search , Summarization, …

Spam filtering, Machine translation

Multilingual search, Associating text with images, …



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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

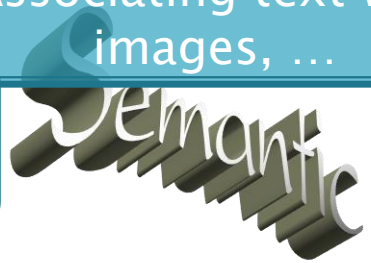
Named entity extraction (names of people, places, organizations)

Text categorization, Clustering, Search , Summarization, …

Spam filtering, Machine translation

Multilingual search, Associating text with images, …

Unifying semantics of data



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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

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

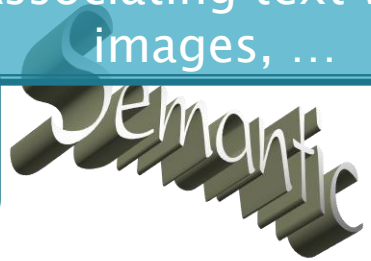
Text categorization, Clustering, Search , Summarization, …

Spam filtering, Machine translation

Multilingual search, Associating text with images, …

Data integration

Unifying semantics of data



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
- Linked Data
- Templates / Frames
- Ontologies / First order theories

Language identification, Copy detection

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

Text categorization, Clustering, Search , Summarization, …

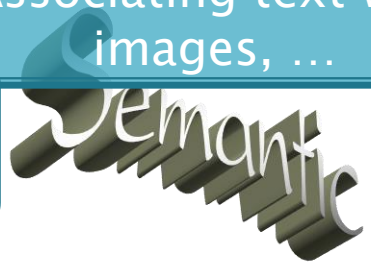
Spam filtering, Machine translation

Multilingual search, Associating text with images, …

Data integration

Unifying semantics of data

Reasoning, Semantic Search



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

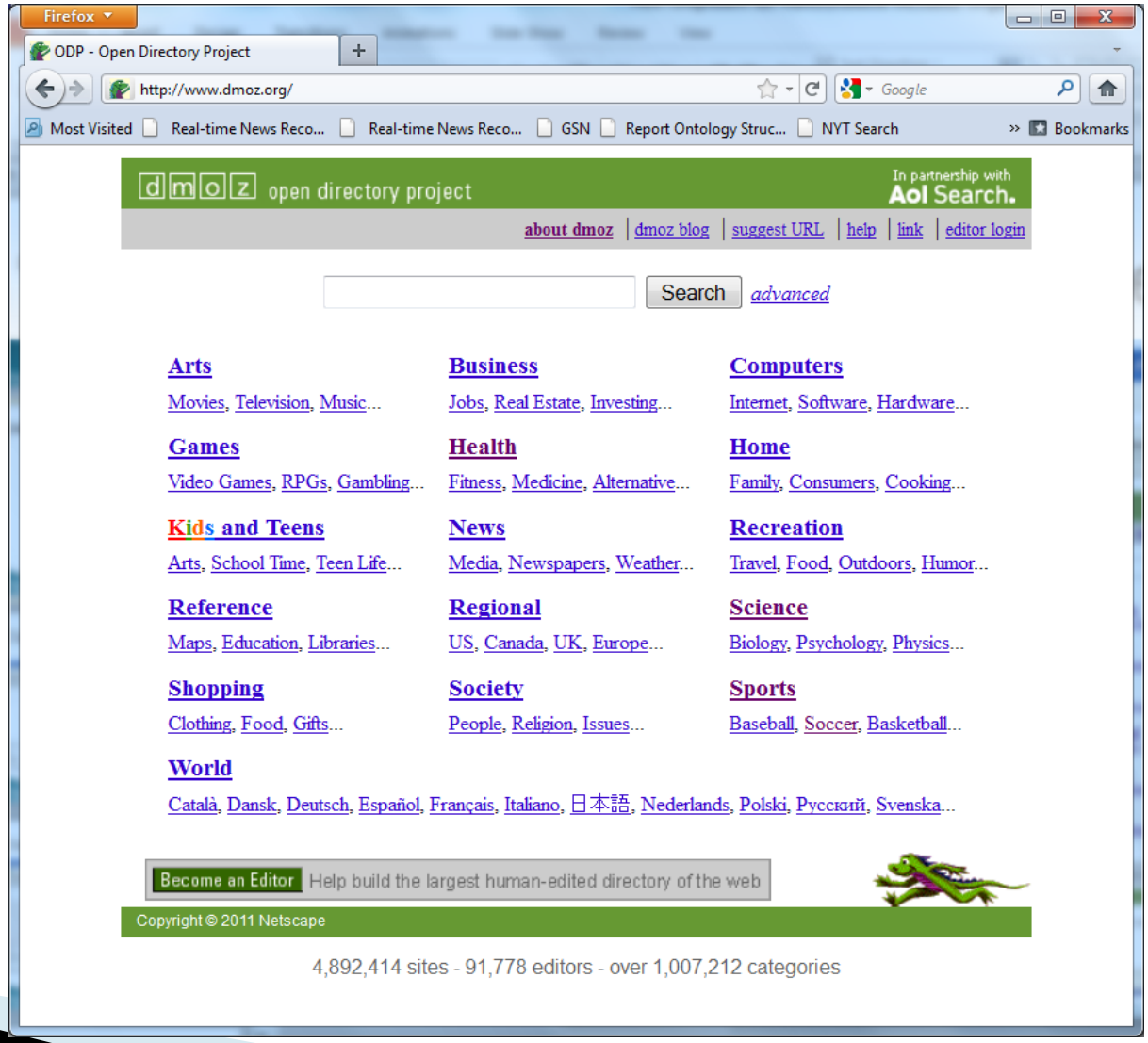
- Vector-space model
- Language models
- Full-parsing
- Cross-modality

- Collaborative tagging / Web2.0
- Linked Data
- Templates / Frames
- Ontologies / First order theories

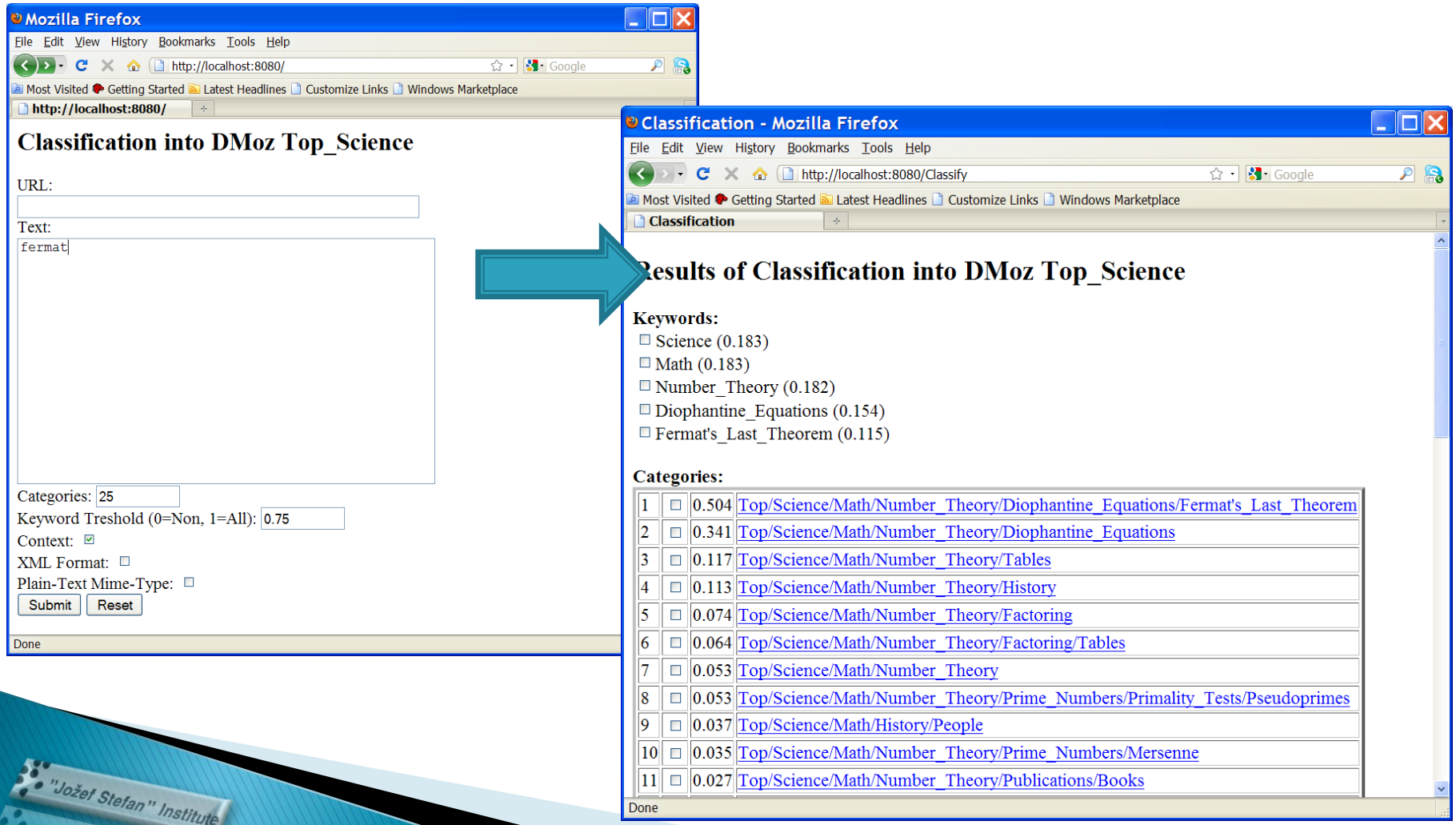
Document Classification into large taxonomies

DMoz (Open Directory Project)
http://dmoz.org

- Largest handcrafted taxonomy on the Web
  - 4,892,414 sites
  - 91,778 editors
  - over 1,007,212 categories
- Data available for download
  - http://www.dmoz.org/rdf.html

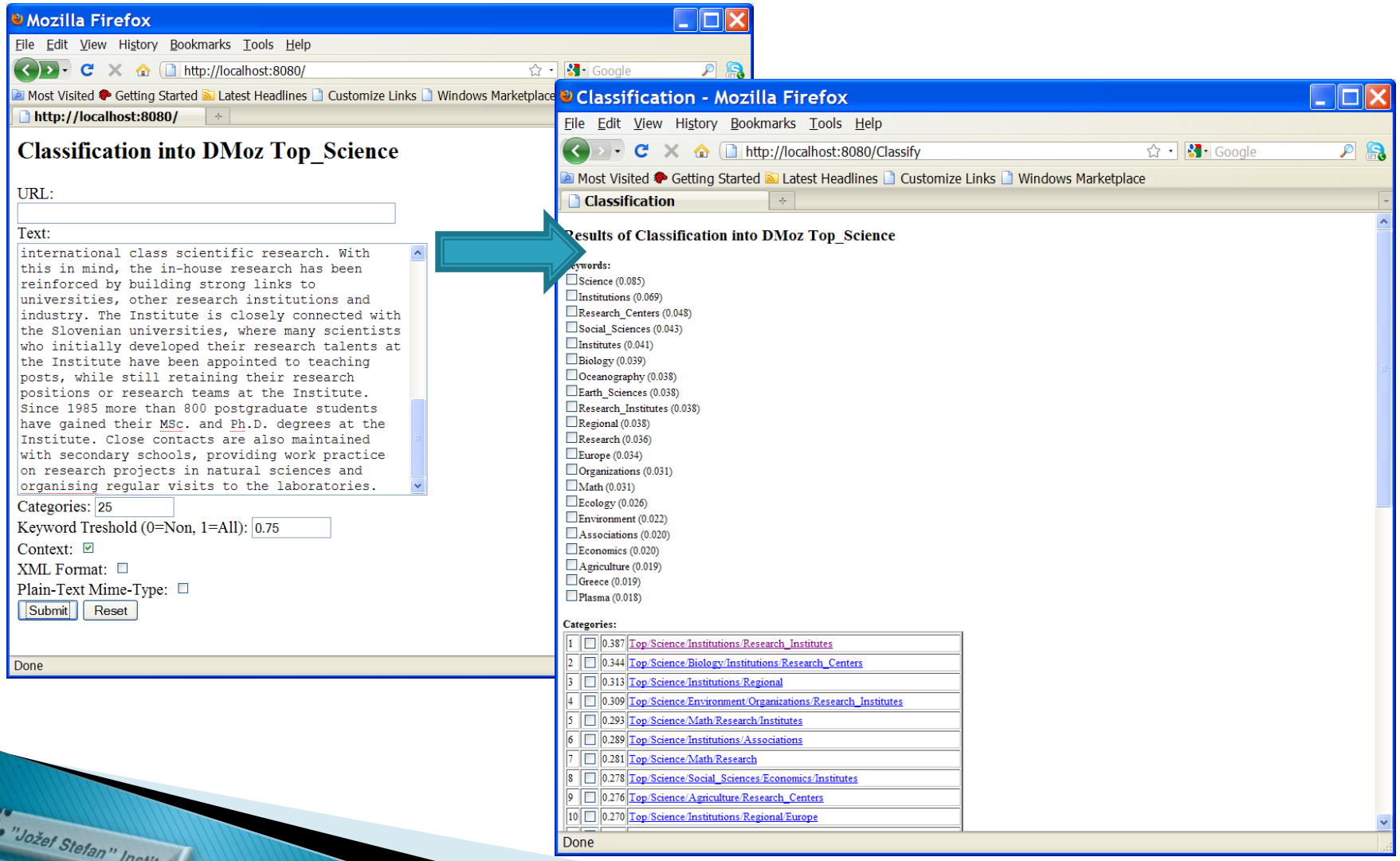


Classification of a query into DMoz



"Jožef Stefan" Institute

Classification of a document into DMoz



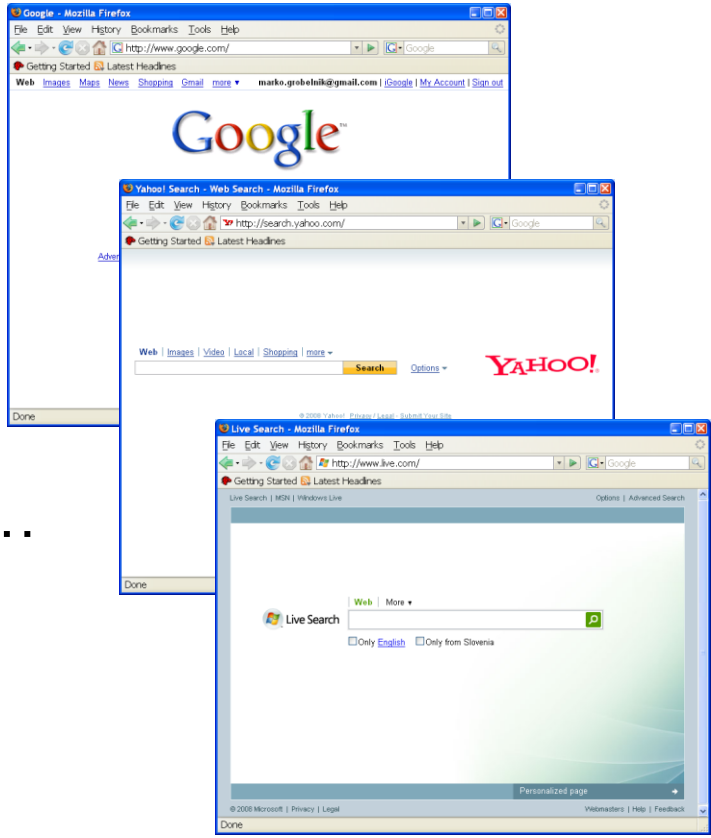
"Jožef Stefan" Institute

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…

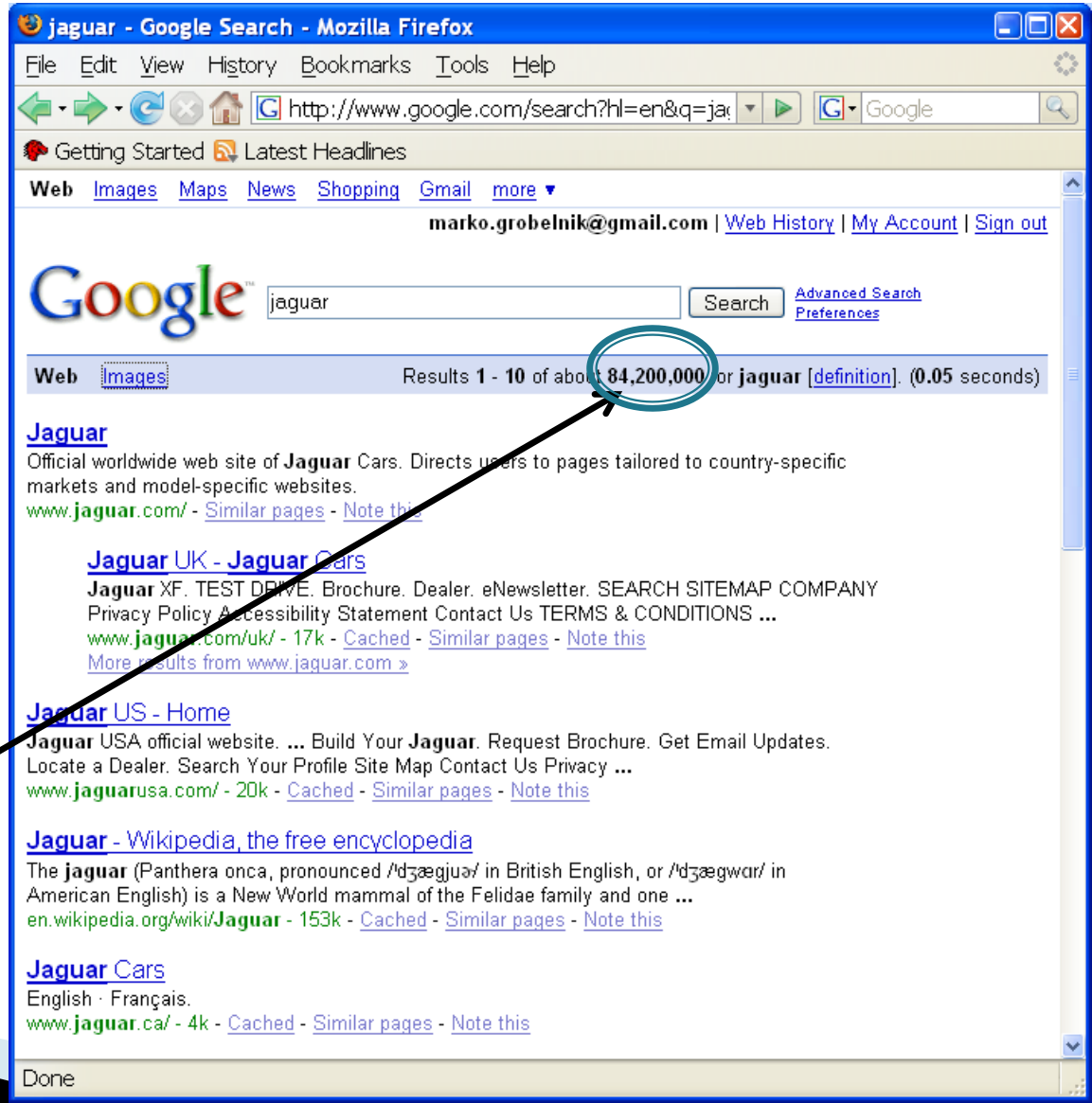


Example: searching for “Jaguar”

- Query “**jaguar**” has many meanings…
- …but the first page of search engines doesn’t provide us with many answers

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



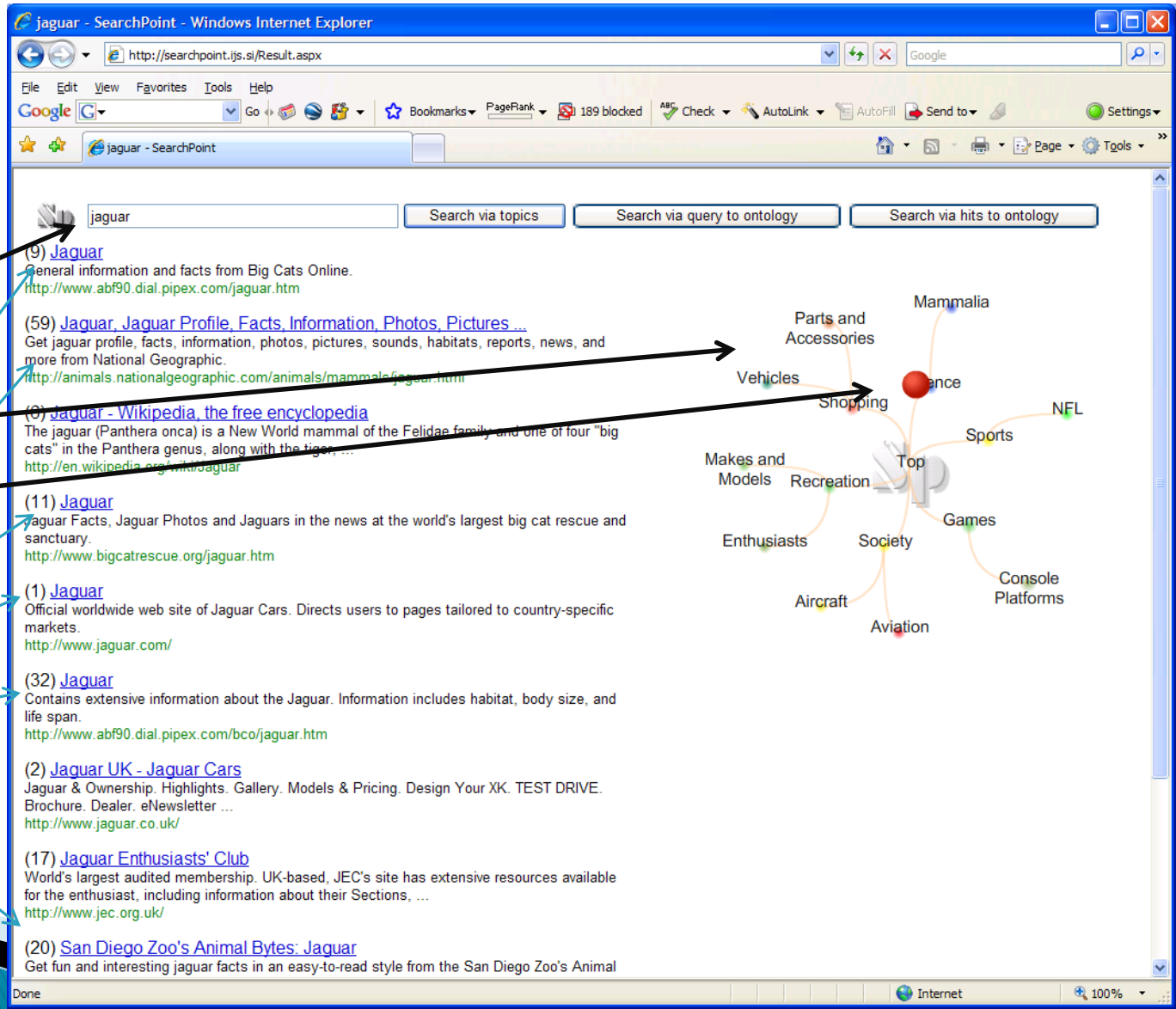
Context sensitive search with http://searchpoint.ijs.si

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point



"Jožef Stefan" Institute

News reporting bias

(Fortuna, Galleguillos, Cristianini 2008)

News Reporting Bias example

UK soldiers cleared in Iraqi death – Seven British soldiers were acquitted on Thursday of charges of beating an innocent Iraqi teenager to death with rifle butts. A judge at a specially convened military court in eastern England ordered the adjudicating panel to return 'not guilty' verdicts against the seven because he did not believe there was sufficient evidence against them, the Ministry of Defence said. . . .

British murderers in Iraq acquitted – The judge at a court-martial on Thursday dismissed murder charges against seven soldiers, from the 3rd Battalion, the Parachute Regiment, who're accused of murdering Iraqi teenager; claiming there's insufficient evidence to secure a conviction, The Associated Press reported Thursday. . . .

Experimental setup

- Time period: March 31st 2005 – April 14th 2006
- Size of collections:

Source	No. of news
Al Jazeera	2142
CNN	6840
Detroit News	2929
International Herald Tribune	9641

- Number of discovered matches:

	AJ	CNN	DN	IHT
AJ	–	816	447	834
CNN	816	–	1103	2437
DN	447	1103	–	895
IHT	834	2437	895	–

Prediction of news source

- **The task**: given a pair of news articles describing the same event, can we predict the news source for each?

- In this experiment we focused on CNN and Al Jazeera.
- SVM linear classifier was used for prediction
  - Evaluation was done using 10-fold cross-validation
  - Significance of results was tested against random matches

Detecting News Reporting Bias

- The task:
  - Given a news story, are we able to say from which news source it came?

↓

↓

-
-

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

↓

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

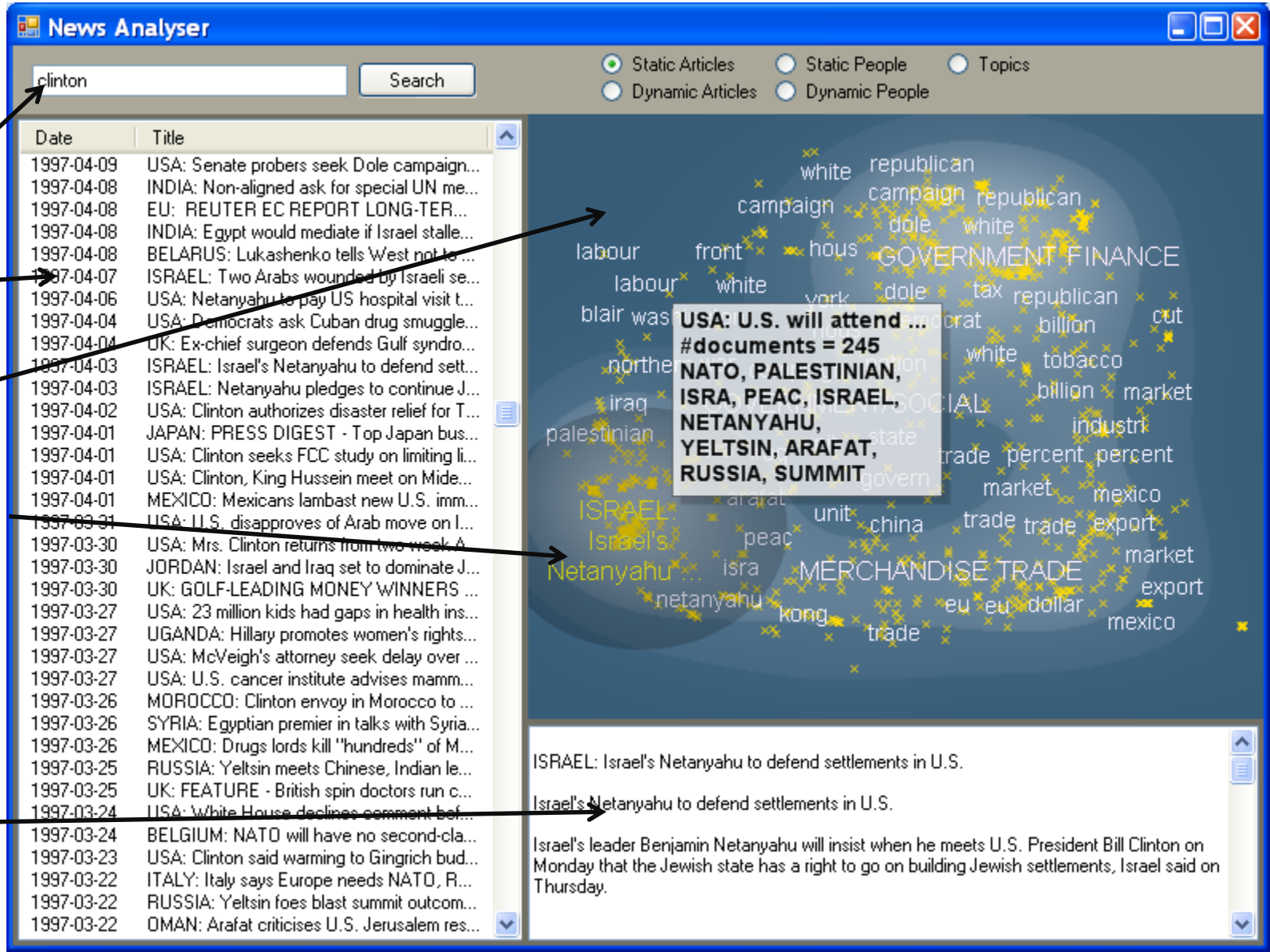
Query

Search Results

Topic Map

Selected group of news

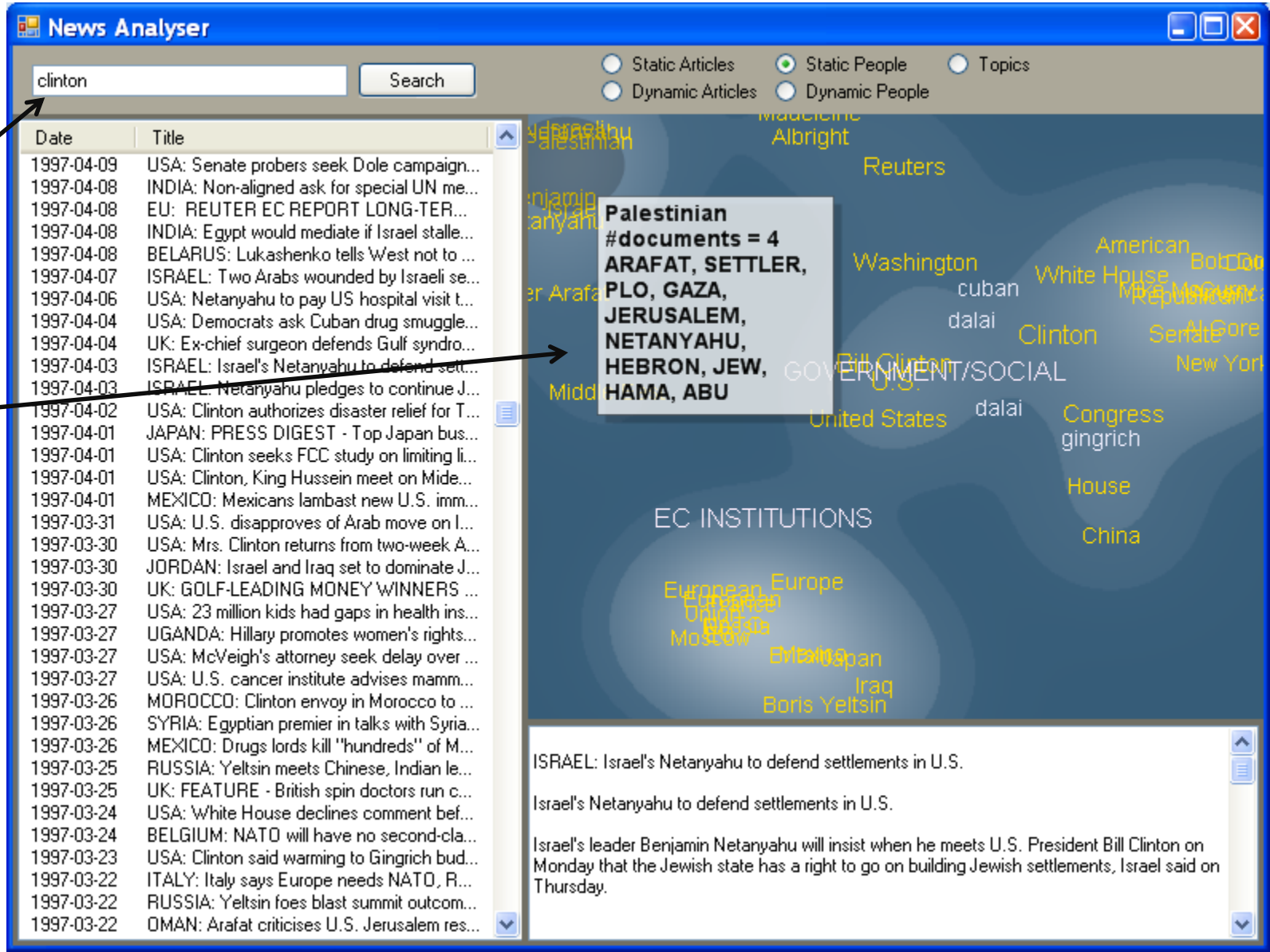
Selected story



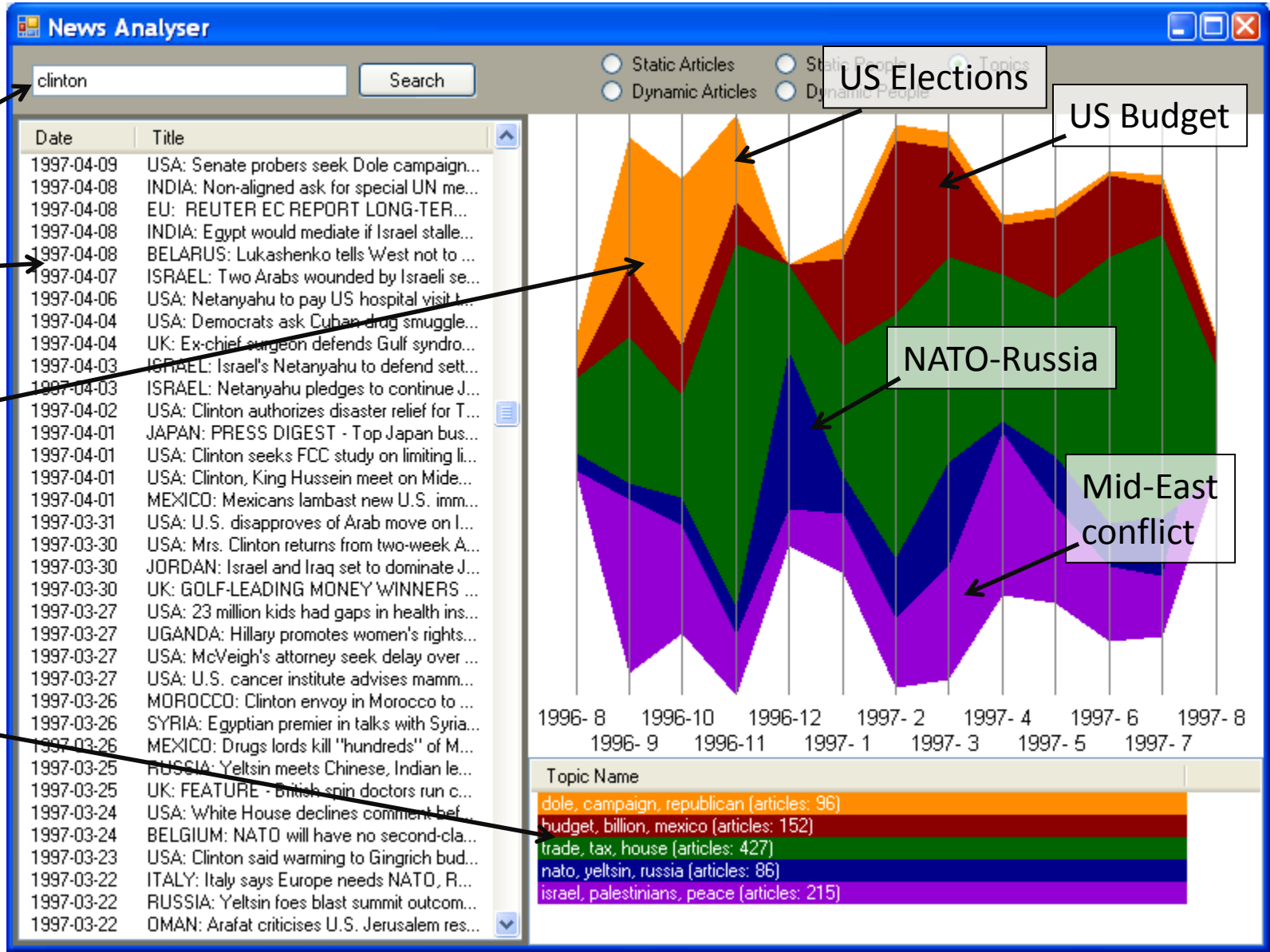
Visualization of social relationships between “Clinton” and other entities

Query

Named entities in relation



Topic Trends Tracking of the documents including “Clinton”



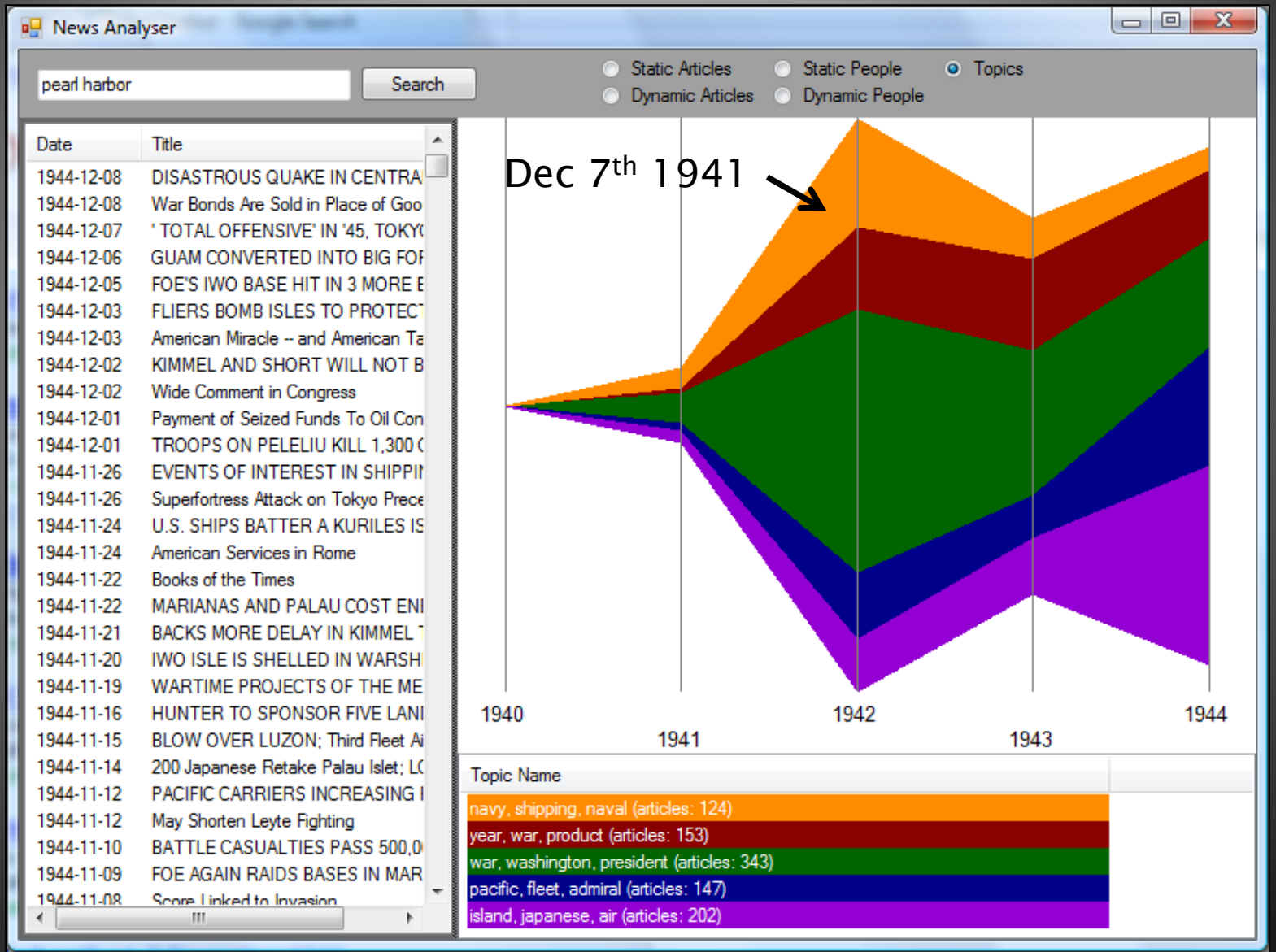
Query

Result set

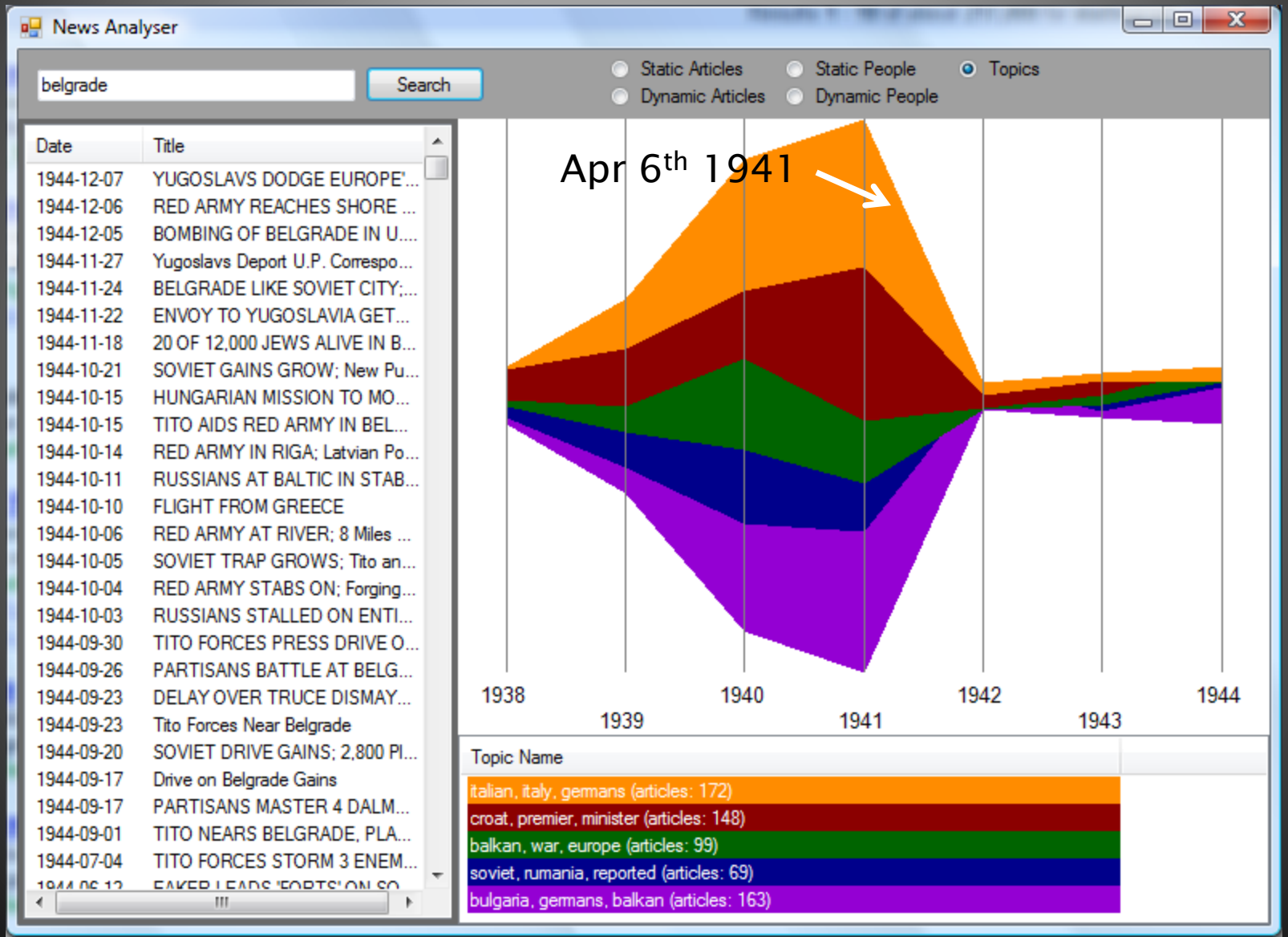
Topic Trends Visualization

Topics description

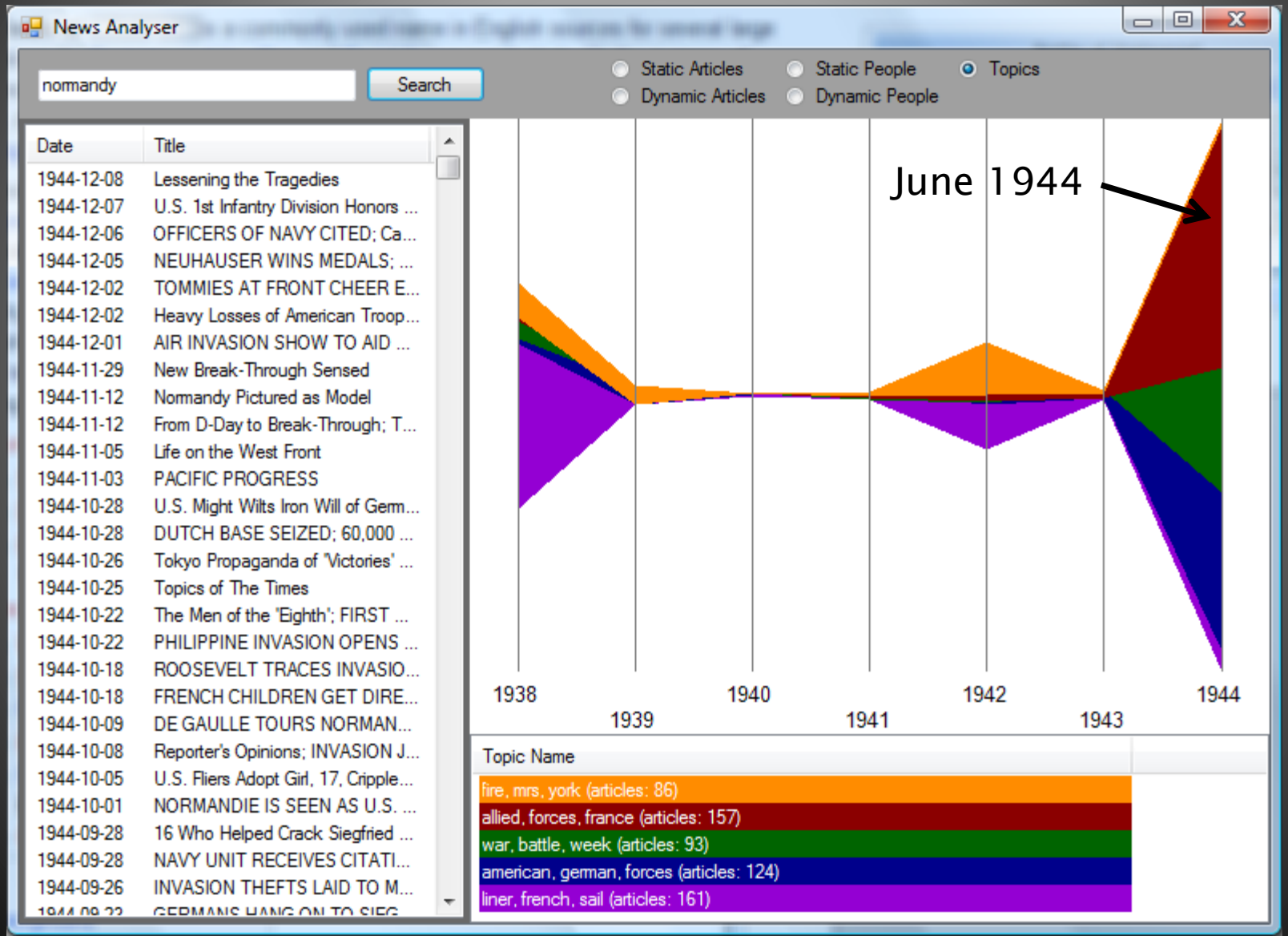
WW2 query “Pearl Harbor” into NYTimes archive



WW2 query “Belgrade” into NYTimes archive



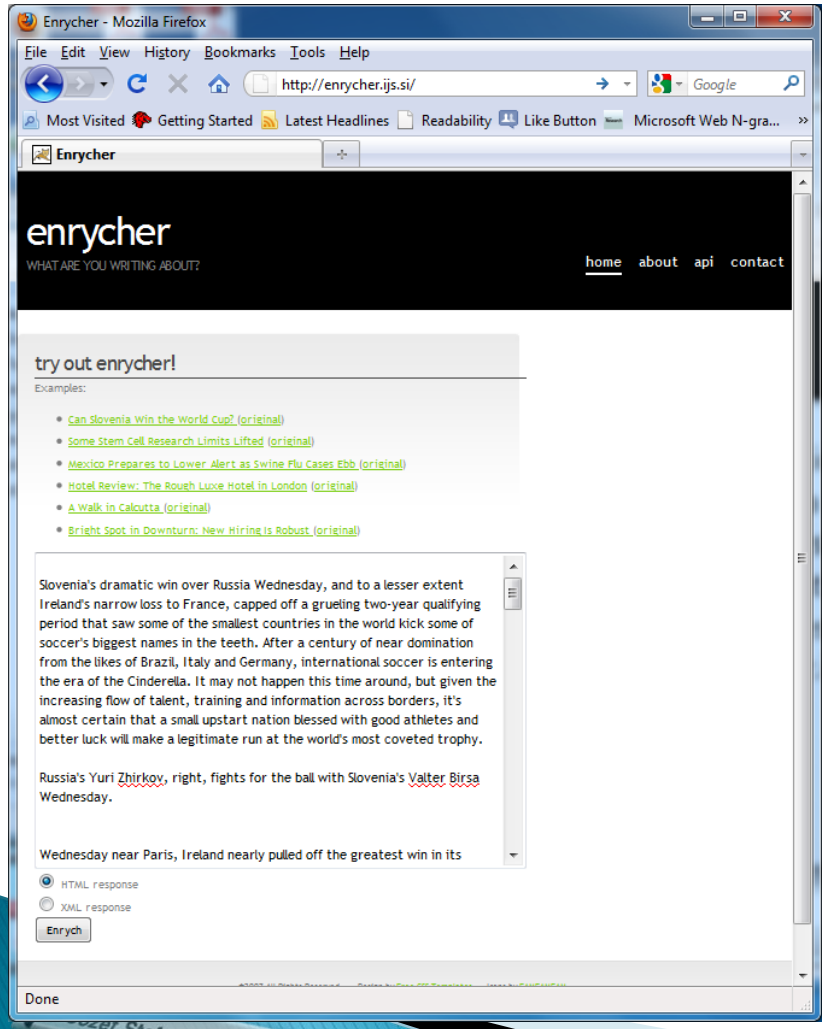
WW2 query “Normandy” into NYTimes archive



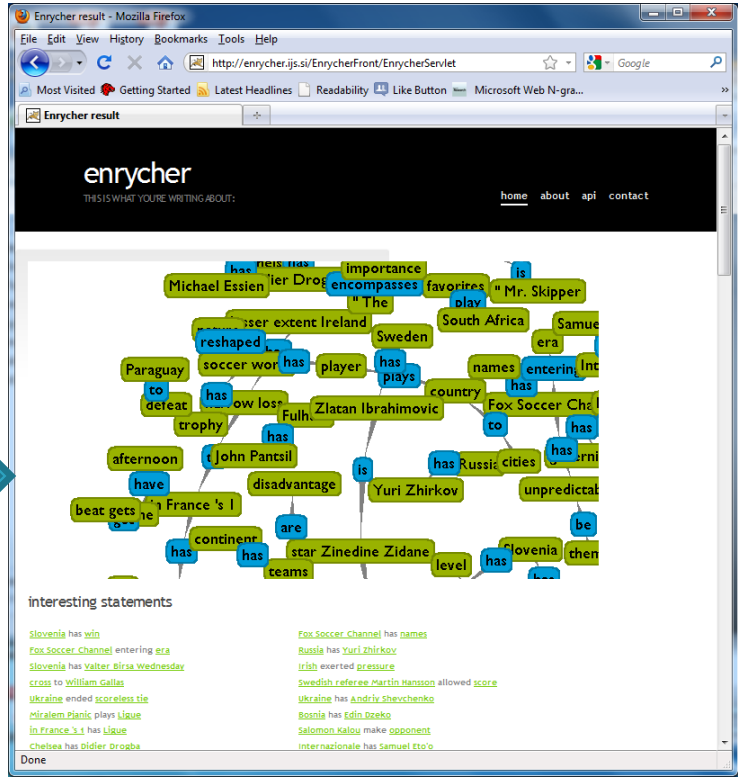
Text Enrichment

WWW2009-SemSearch

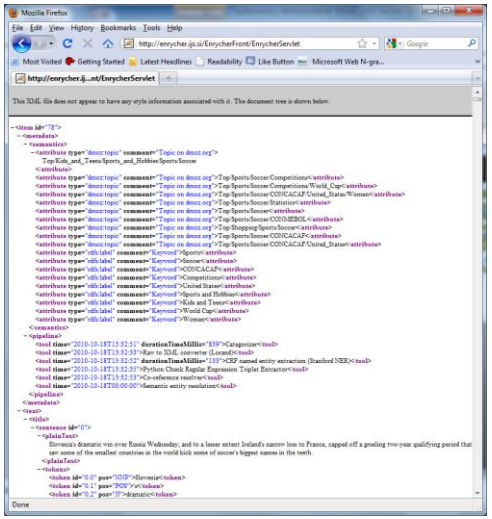
Text enrichment with http://Enrycher.ijs.si

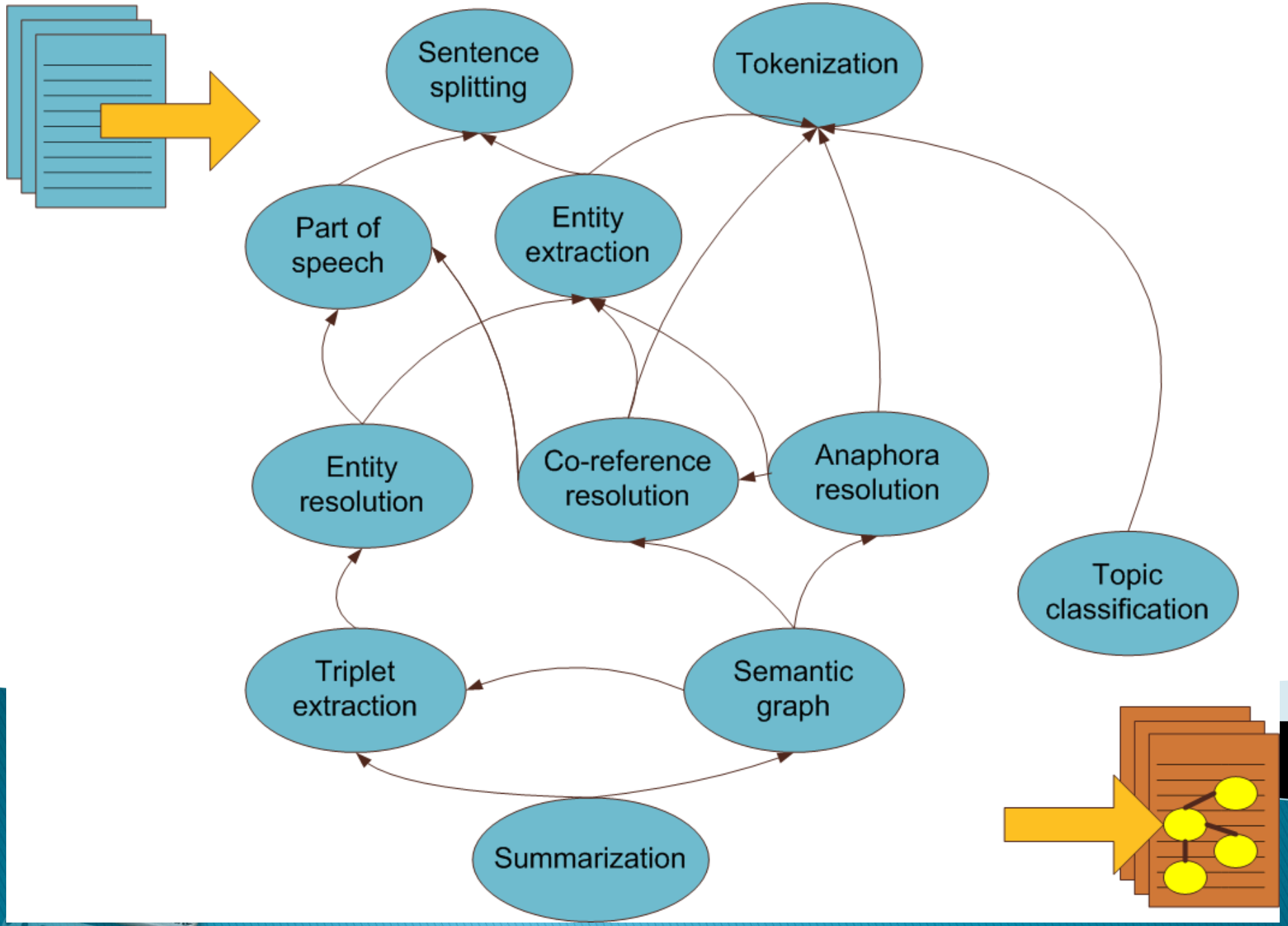


HTML



Web service




Sentence splitting
Tokenization
Part of speech
Entity extraction
Entity resolution
Co-reference resolution
Anaphora resolution
Topic classification
Triplet extraction
Semantic graph
Summarization

Knowledge based summarization

AAAI 2005

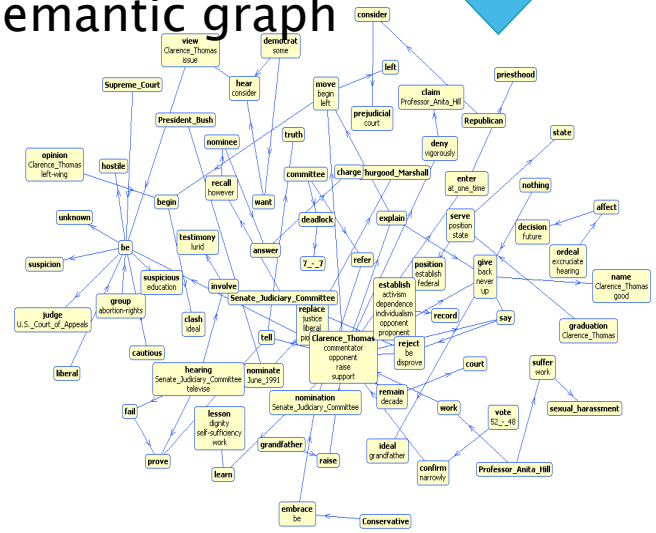
Summarization via semantic graphs

Original document

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

Hussein sought to circumvent the economic noose around his country. Japan, enforcing the sanctions. Hoping to defuse criticism that it is not doing its share s most affected by the U.N. embargo on Iraq. President Bush on Tuesday night on audience that ``Saddam Hussein will fail'' to make his conquest of Kuwait ush, who added that the U.S. military may remain in the Saudi Arabian desert o withdraw from Kuwait,'' Bush said. More than 150,000 U.S. troops have been bia. Bush's aides said the president would follow his address to Congress with a their government's invasion of Kuwait. Saddam had offered Bush time on Iraqi pond to an offer Monday by Saddam of free oil _ in exchange for sending their as a none-too-subtle attempt to bypass the U.N. embargo, in effect since four ck 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 European countries also are trying to maintain their military sales to Iraq. A well-informed source in Tehran told The Associated Press that Iran has agreed to an Iraqi request to exchange food and medicine for up to 200,000 barrels of refined oil a day and cash payments. There was no official comment from Tehran or Baghdad on the reported food-for-oil deal. But the source, who requested anonymity, said the deal was struck during Iraqi Foreign Minister Tariq Aziz's visit Sunday to Tehran, the first by a senior Iraqi official since the 1980-88 gulf war. After the visit, the two countries announced they would resume diplomatic relations. Well-informed oil industry sources in the region, contacted by The AP, said that although Iran is a major oil exporter itself, it currently has to import about 150,000 barrels of refined oil a day for domestic use because of damages to refineries in the gulf war. Along similar lines, ABC News reported that following Aziz's visit, Iraq is apparently prepared to give Iran all the oil it wants to make up for the damage Iraq inflicted on Iran during their conflict. Secretary of State James A. Baker III, meanwhile, met in Moscow with Soviet Foreign Minister Eduard Shevardnadze, two days after the U.S.-Soviet summit that produced a joint demand that Iraq withdraw from Kuwait. During the summit, Bush encouraged Mikhail Gorbachev to withdraw 190 Soviet military specialists from Iraq, where they remain to fulfill contracts. Shevardnadze told the Soviet parliament Tuesday the specialists had not reneged on those contracts for fear it would jeopardize the 5,800 Soviet citizens in Iraq. In his speech, Bush said his heart went out to the families of the hundreds of Americans held hostage by Iraq, but he declared, ``Our policy cannot change, and it will not change. America and the world will not be blackmailed.'' The president added: ``Vital issues of principle are at stake. Saddam Hussein is literally trying to wipe a country off the face of the Earth.'' In other developments: _A U.S. diplomat in Baghdad said Tuesday up to 800 Americans and Britons will fly out of Iraqi-occupied Kuwait this week, most of them women and children leaving their husbands behind. Saddam has said he is keeping foreign men as human shields against attack. On Monday, a planeload of 164 Westerners arrived in Baltimore from Iraq. Evacuees spoke of food shortages in Kuwait, nighttime gunfire and Iraqi roundups of young people suspected of involvement in the resistance. ``There is no law and order,'' said Thuraya, 19, who would not give her last name. ``A soldier can rape a father's daughter in front of him and he can't do anything about it.''_The State Department said Iraq had told U.S. officials that American males residing in Iraq and Kuwait who were born in Arab countries will be allowed to leave. Iraq generally has not let American males leave. It was not known how many men the Iraqi move could affect. _A Pentagon spokesman said ``some increase in military activity'' had been detected inside Iraq near its borders with Turkey and Syria. He said there was little indication hostilities are imminent. Defense Secretary Dick Cheney said the cost of the U.S. military buildup in the Middle East was rising above the $1 billion-a-month estimate generally used by government officials. He said the total cost _ if no shooting war breaks out _ could total $15 billion in the next fiscal year beginning Oct. 1. Cheney promised disgruntled lawmakers ``a significant increase'' in help from Arab nations and other U.S. allies for Operation Desert Shield. Japan, which has been accused of responding too slowly to the crisis in the gulf, said Tuesday it may give $2 billion to Egypt, Jordan and Turkey, hit hardest by the U.N. prohibition on trade with Iraq. ``The pressure from abroad is getting so strong,'' said Hiroyasu Horio, an official with the Ministry of International Trade and Industry. Local news reports said the aid would be extended through the World Bank and International Monetary Fund, and $600 million would be sent as early as mid-September. On Friday, Treasury Secretary Nicholas Brady visited Tokyo on a world tour seeking $10.5 billion to help Egypt, Jordan and Turkey. Japan has already promised a $1 billion aid package for multinational peacekeeping forces in Saudi Arabia, including food, water, vehicles and prefabricated housing for non-military uses. But critics in the United States have said Japan should do more because its economy depends heavily on oil from the Middle East. Japan imports 99 percent of its oil. Japan's constitution bans the use of force in settling international disputes and Japanese law restricts the military to Japanese territory, except for ceremonial occasions. On Monday, Saddam offered developing nations free oil if they would send their tankers to pick it up. The first two countries to respond Tuesday _ the Philippines and Namibia _ said no. Manila said it had already fulfilled its oil requirements, and Namibia said it would not ``sell its sovereignty'' for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a ``propaganda ploy.'' Venezuela, an OPEC 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 re[illegible]ment of Iraqi petroleum since U.N. sanctions were imposed five weeks ago. And Romania, it said, expects to receive oil indirectly from Iraq. R[illegible]bassador to the [illegible] and without four[illegible]

Linguistic processing and Creation of semantic graph



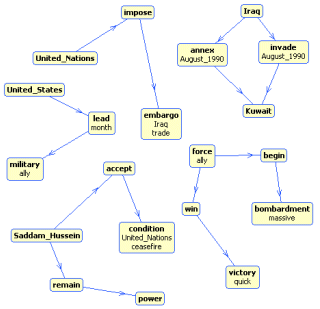
Automatically generated document summary

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

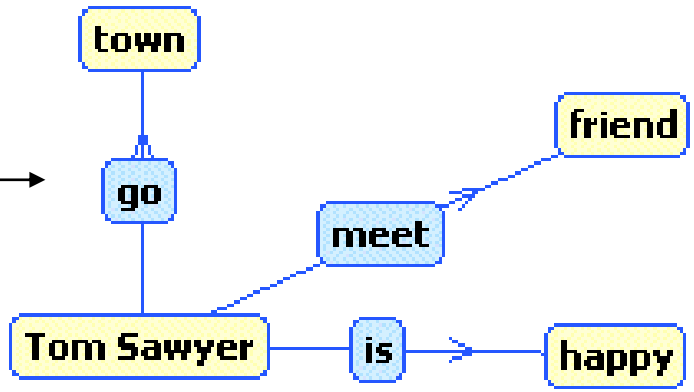
Generate a **summary graph**

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 ← go → town
Tom ← meet → friend
Tom ← is → happy



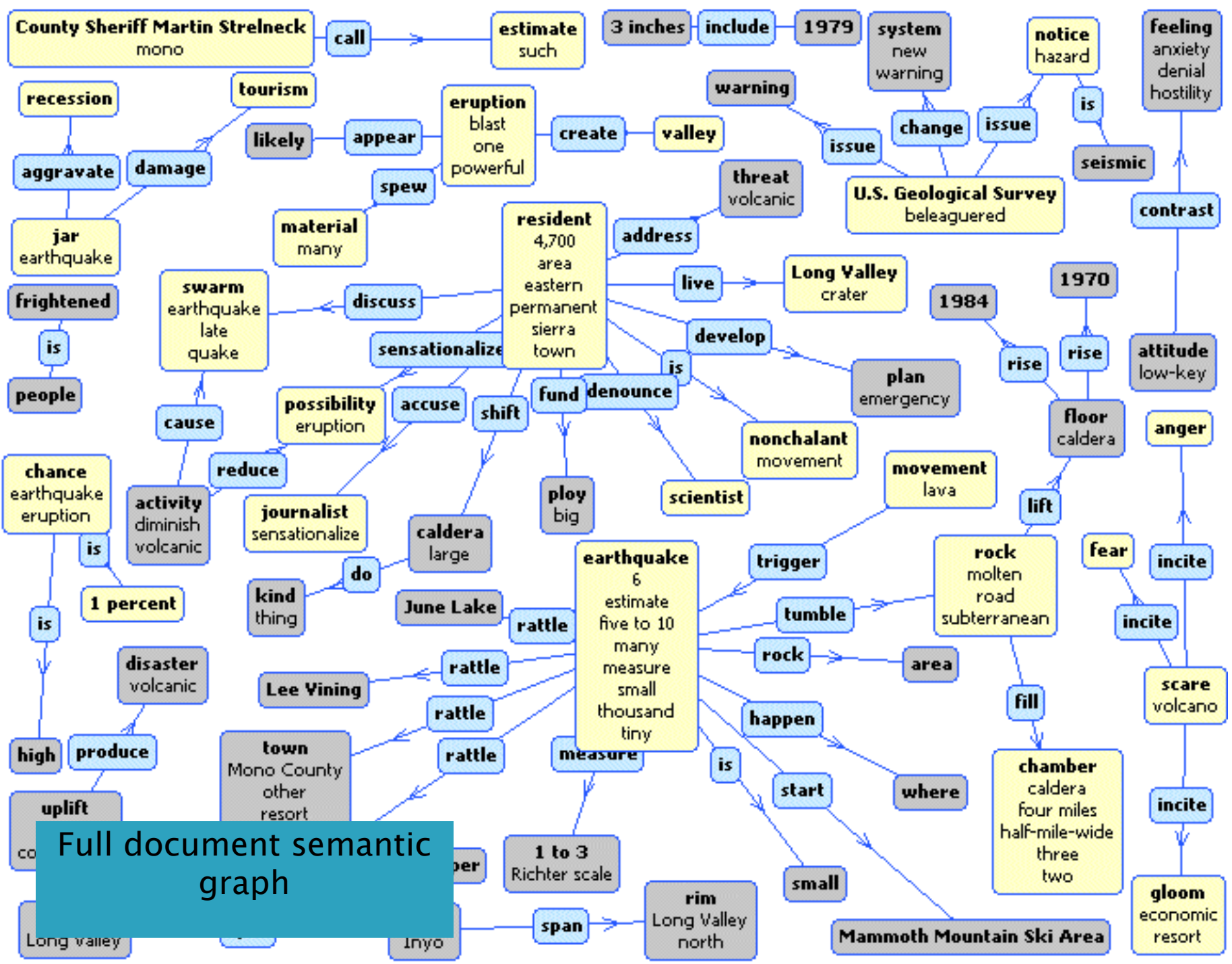
Example of automatic summary

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

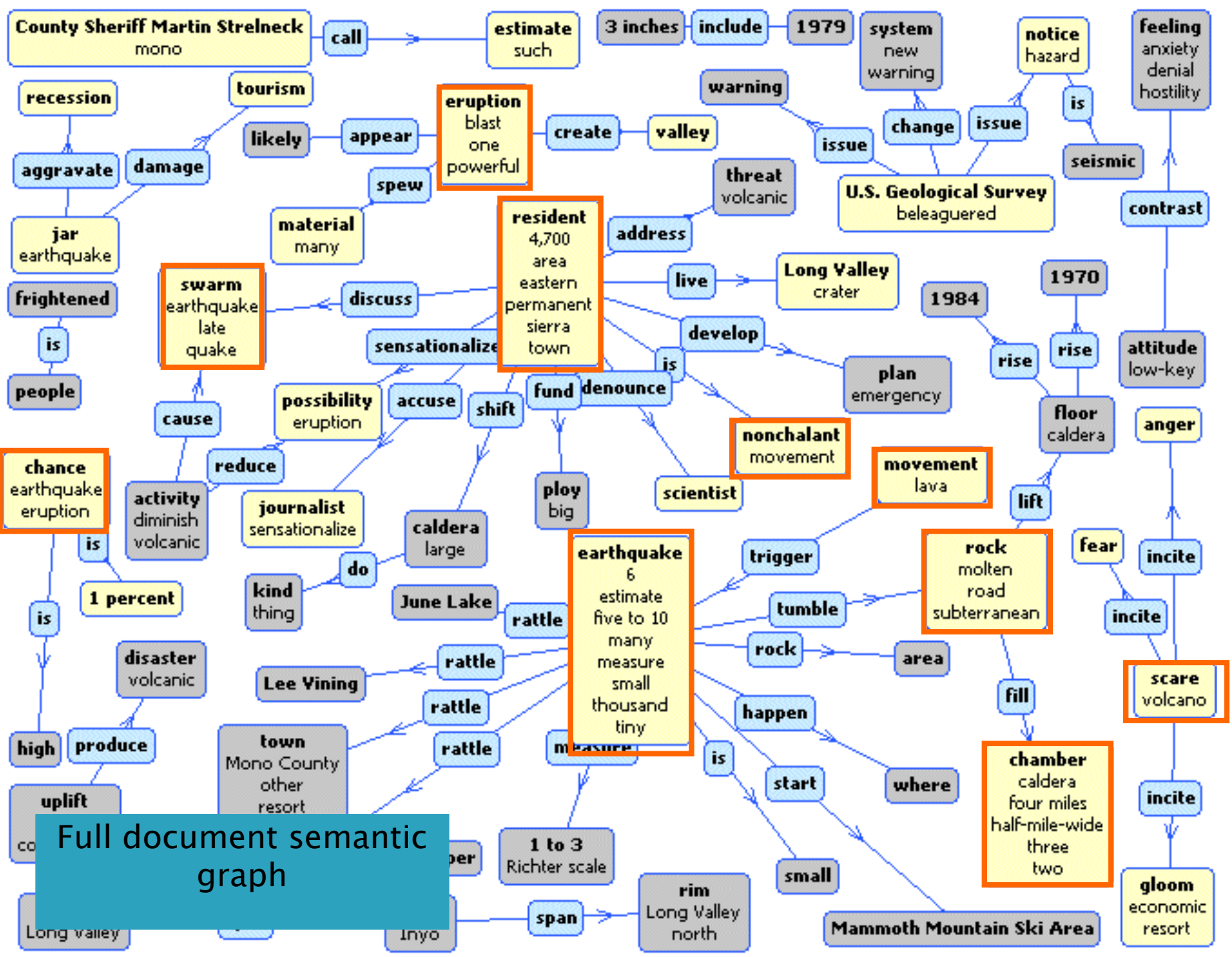
Human extracted summary

Cracks appeared Tuesday in the U.N. trade embargo against Iraq as Saddam Hussein sought to circumvent the economic noose 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 its share 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 poor 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 European ... Iran has agreed to an Iraqi requ... ment from Tehran or Baghdad ... ter Tariq Aziz's visit Sunday to Tehran ... iplomatic relations. Well-informed ... mport about 150,000 barrels of ... following Aziz's visit, Iraq is apparen... State James A. Baker III, meanwhil... nt demand that Iraq withdraw ... re they remain to fulfill contracts. ... e the 5,800 Soviet citizens in Iraq. In... ``Our policy cannot change, and it will ... am Hussein is literally trying to ... ns and Britons will fly out of Iraqi-occu... n men as human shields against at... ighttime gunfire and Iraqi roundups ... ve her last name. ``A soldier ca... icials that American males residing i... was not known how many men the I... s borders with Turkey and Syria. He ... in the Middle East was rising abo... t _ could total $15 billion in the next f... er U.S. allies for Operation Desert Sh... pt, Jordan and Turkey, hit hardest b... the Ministry of Internatio... Fund, and $600 million would be ... on to help Egypt, Jordan and Turke... ater, vehicles and prefabrica... s heavily on oil from the Middle Ea... law restricts the military to Japanese ... ers to pick it up. The first two count... mibia said it would not ``sell its sovereignty'' for Iraqi oil. Venezuelan President Carlos Andres Perez dismissed Saddam's offer of free oil as a ``propaganda ploy.'' Venezuela, an OPEC 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.''.

- 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.

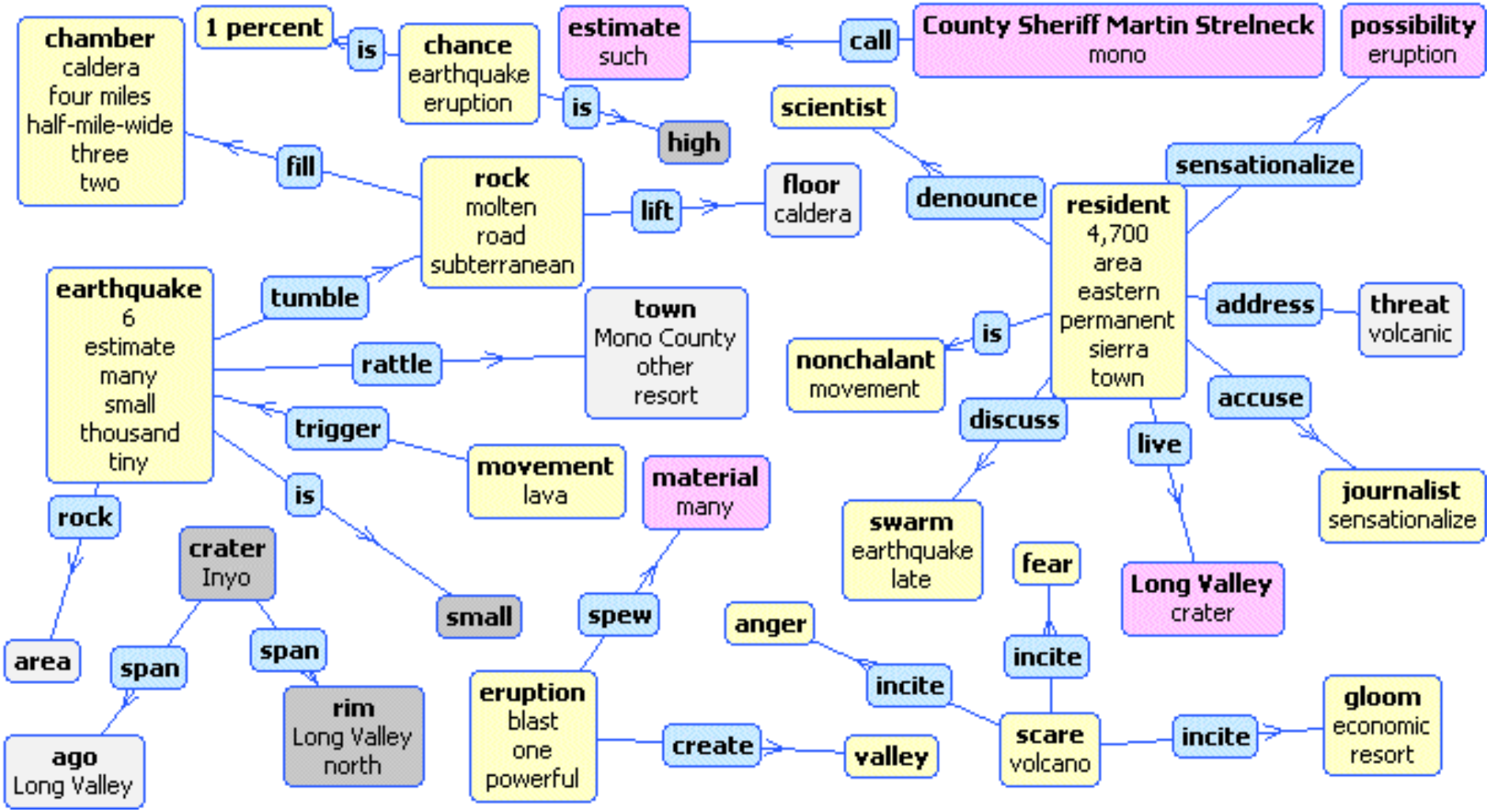


Full document semantic graph

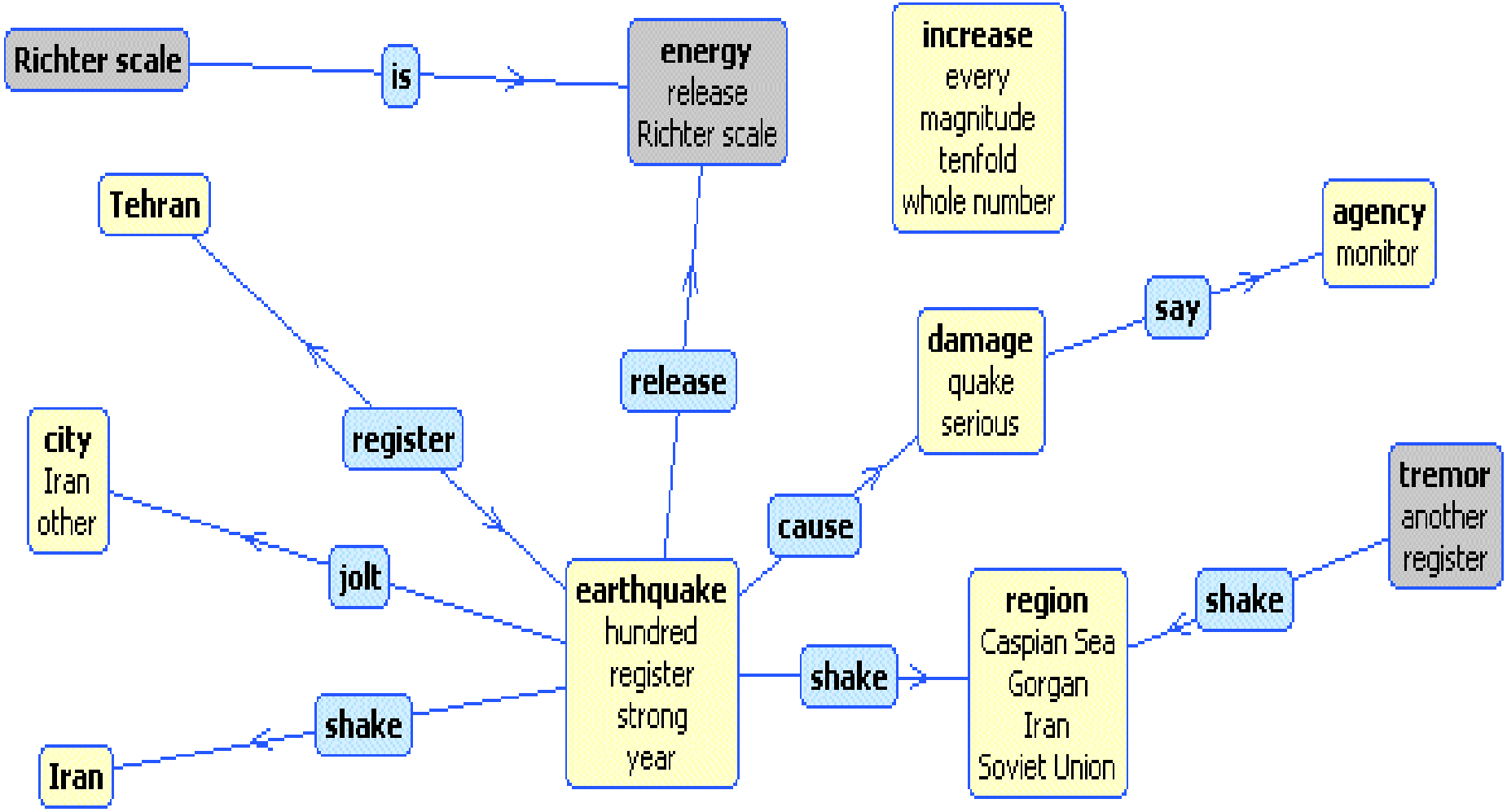


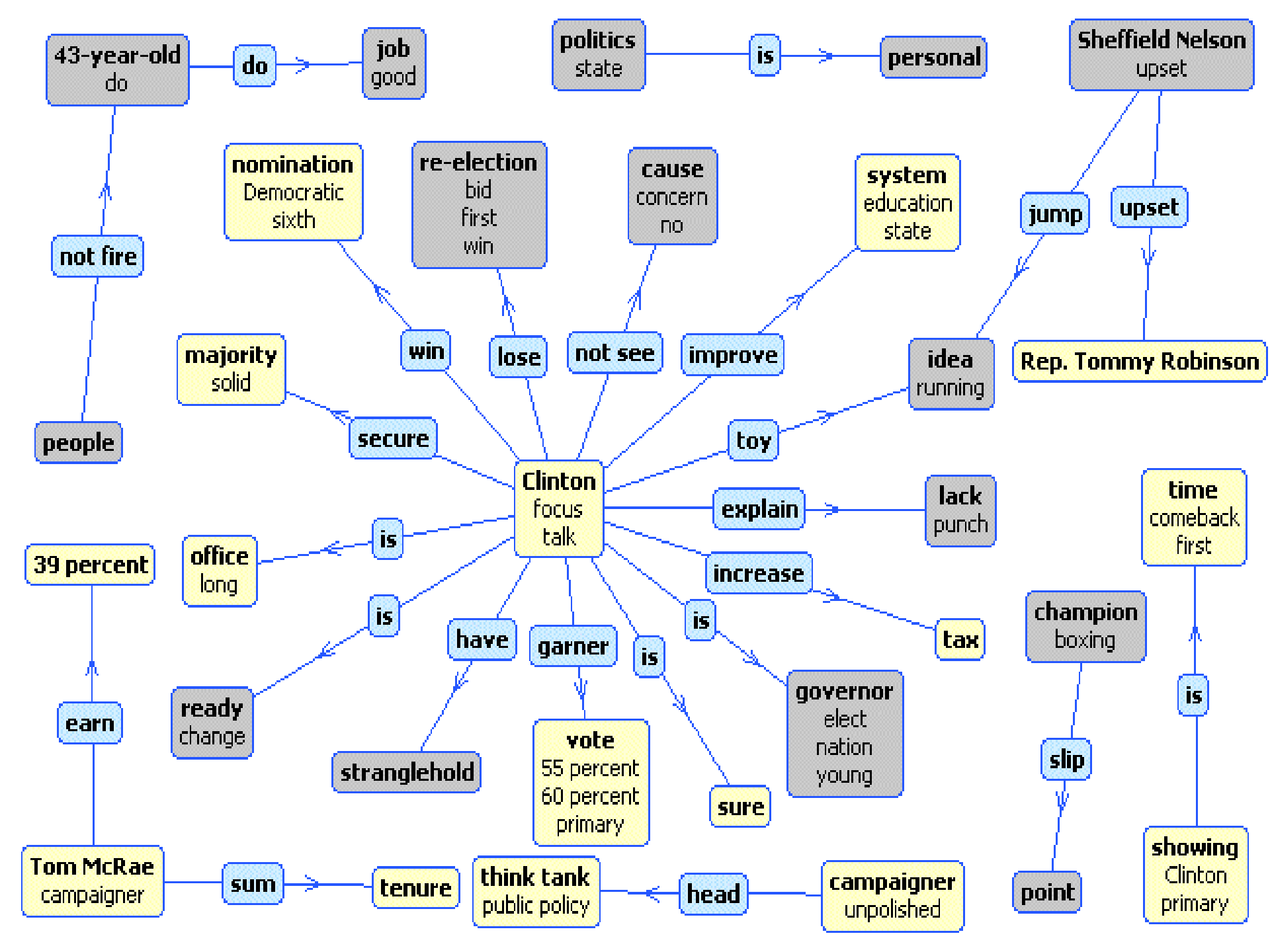
Full document semantic graph

Automatically generated summary graph



More examples




43-year-old
do
do
job
good
politics
state
is
personal
Sheffield Nelson
upset
not fire
people
nomination
Democratic
sixth
re-election
bid
first
win
cause
concern
no
system
education
state
jump
upset
win
lose
not see
improve
idea
running
Rep. Tommy Robinson
majority
solid
secure
toy
Clinton
focus
talk
explain
lack
punch
time
comeback
first
39 percent
office
long
is
increase
champion
boxing
is
tax
is
have
garner
is
is
earn
ready
change
governor
elect
nation
young
slip
is
stranglehold
vote
55 percent
60 percent
primary
sure
Tom McRae
campaigner
sum
tenure
think tank
public policy
head
campaigner
unpolished
point
showing
Clinton
primary

Cyc Knowledge Base and Reasoning

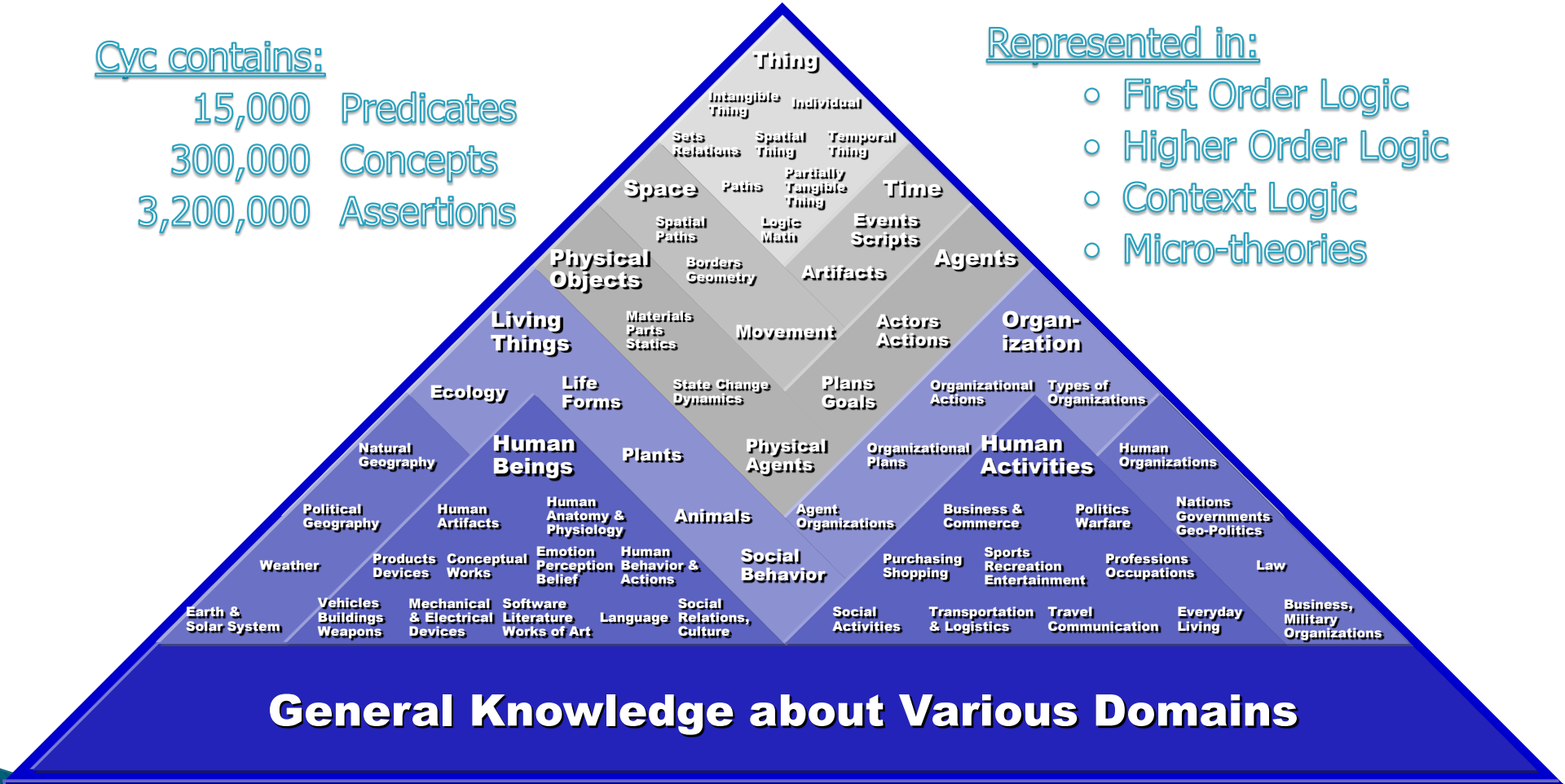
The Cyc Ontology

Cyc contains:

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

Represented in:

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



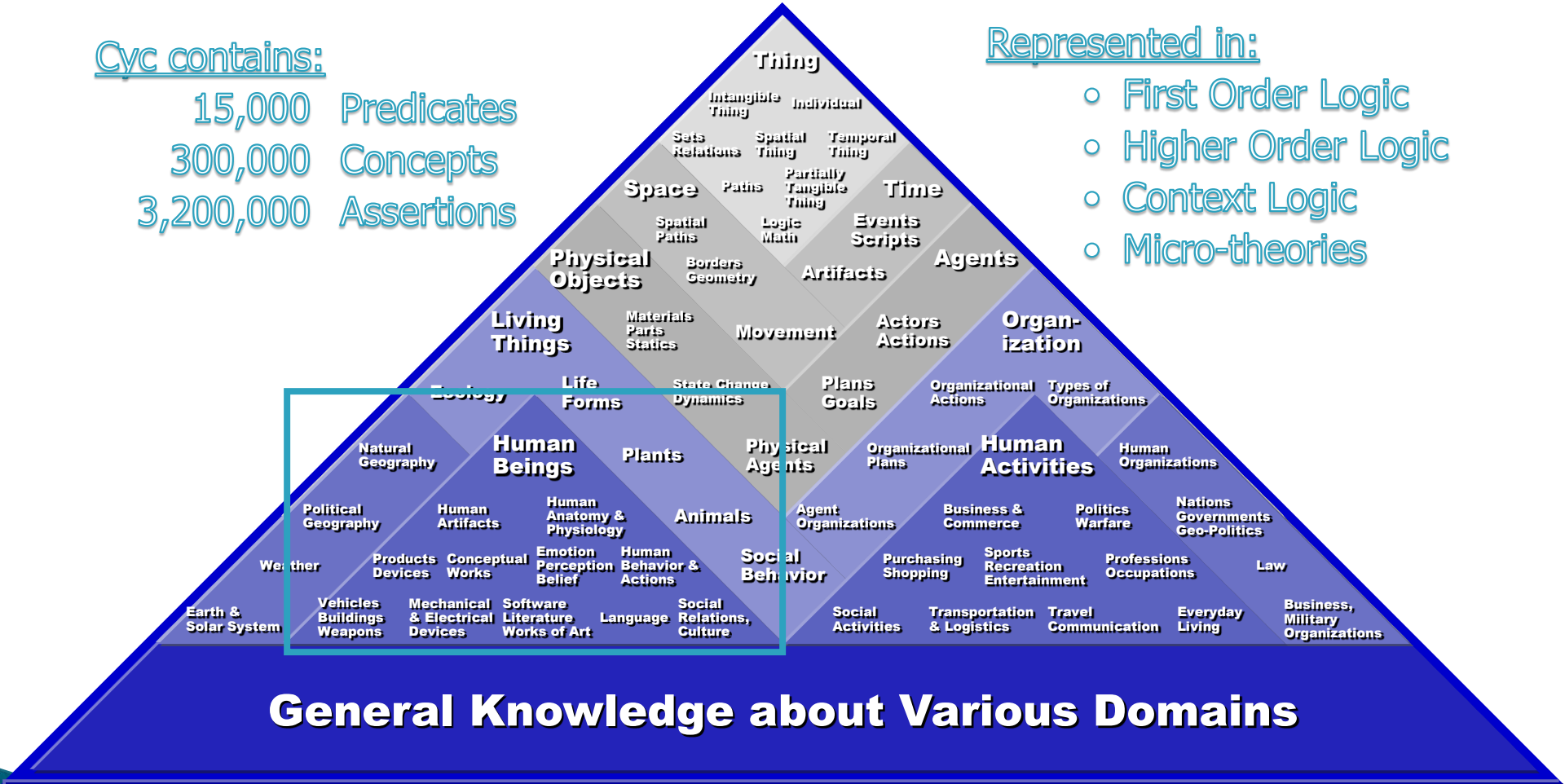
The Cyc Ontology

Cyc contains:

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

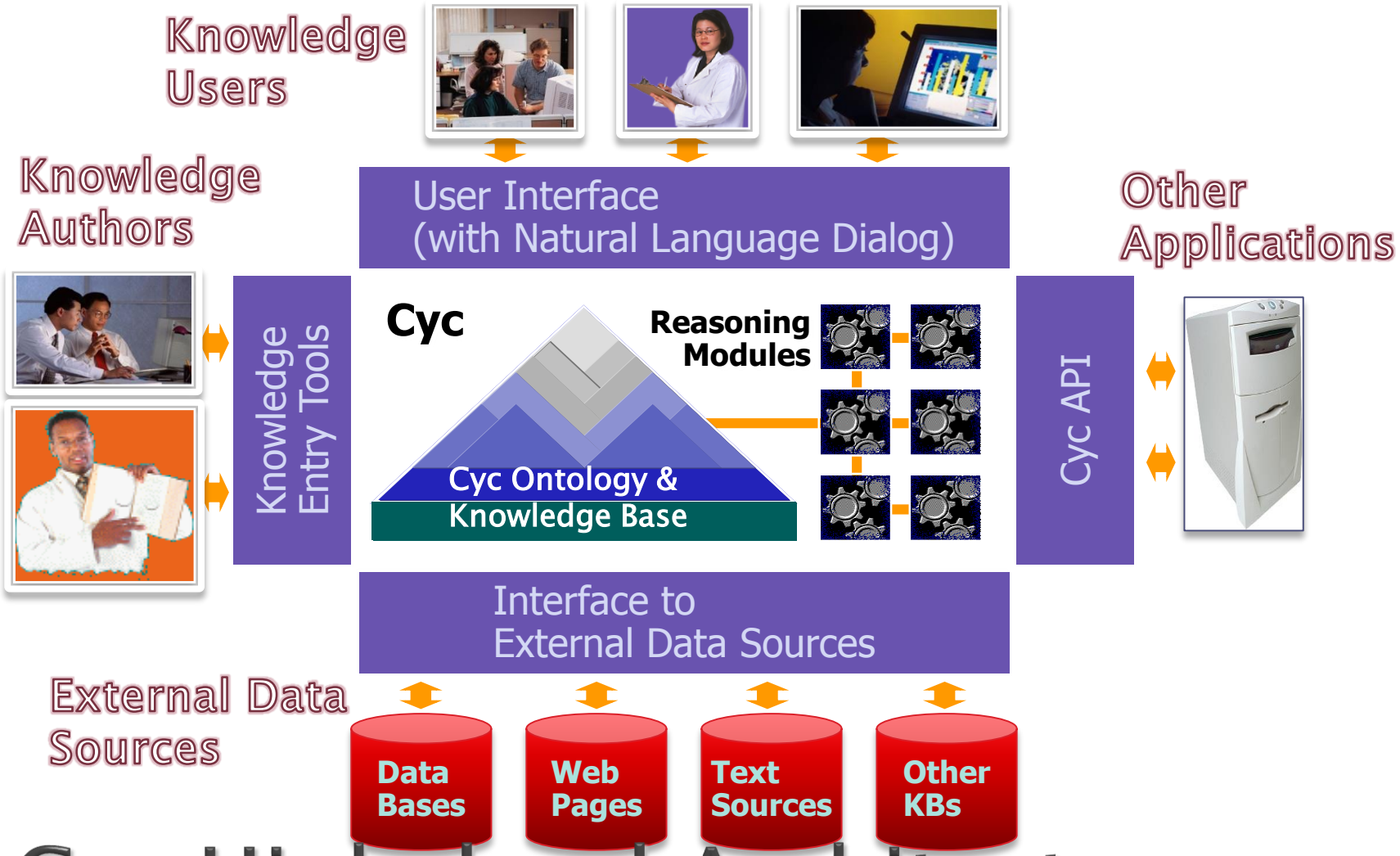
Represented in:

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



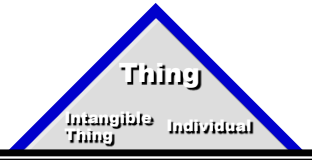
General Knowledge about Various Domains

Specific data, facts, and observations



Cyc High-level Architecture

Cyc KB Extended w/Domain Knowledge



General Knowledge about Terrorism:

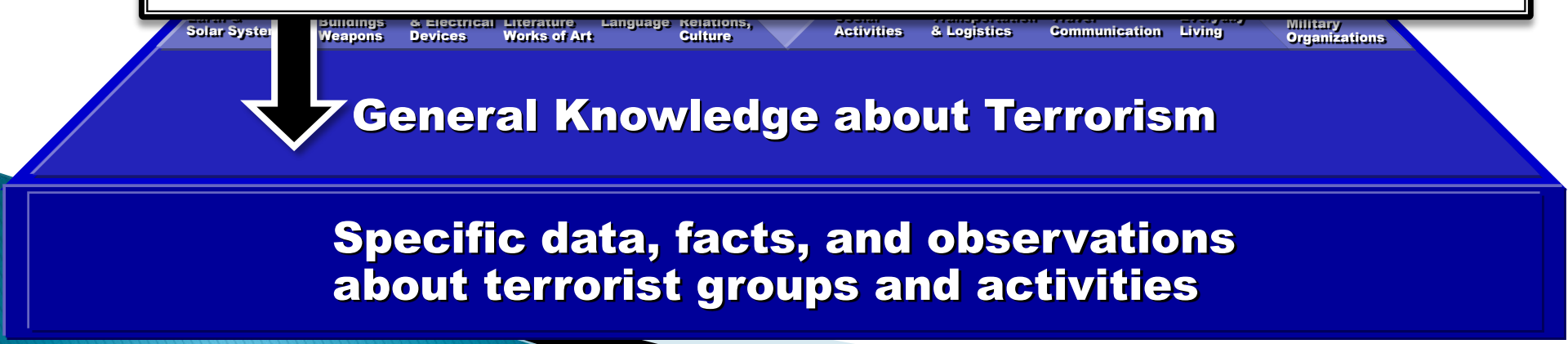
Terrorist groups are capable of directing assassinations:

```
(implies
      (isa ?GROUP TerroristGroup)
      (behaviorCapable ?GROUP AssassinatingSomeone directingAgent))
```

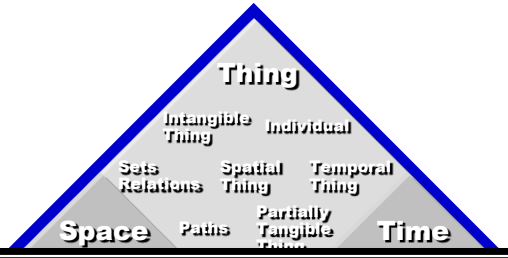
…

If a terrorist group considers an agent an enemy, that agent is vulnerable to an attack by that group:

```
(implies
      (and
         (isa ?GROUP TerroristGroup)
         (considersAsEnemy ?GROUP ?TARGET))
      (vulnerableTo ?GROUP ?TARGET TerroristAttack))
```



Cyc KB Extended w/Domain Knowledge



Specific Facts about Al Qaida:

(basedInRegion AlQaida Afghanistan) Al-Qaida is based in Afghanistan.
(hasBeliefSystems AlQaida IslamicFundamentalistBeliefs) Al-Qaida has Islamic fundamentalist beliefs.
(hasLeaders AlQaida OsamaBinLaden) Al-Qaida is led by Osama bin Laden.
…
(affiliatedWith AlQaida AlQudsMosqueOrganization) Al-Qaida is affiliated with the Al Quds Mosque.
(affiliatedWith AlQaida SudaneseIntelligenceService) Al-Qaida is affiliated with the Sudanese Intell Service
…
(sponsors AlQaida HarakatUlAnsar) Al-Qaida sponsors Harakat ul-Ansar.
(sponsors AlQaida LaskarJihad) Al-Qaida sponsors Laskar Jihad.
…
(performedBy EmbassyBombingInNairobi AlQaida) Al-Qaida bombed the Embassy in Nairobi.
(performedBy EmbassyBombingInTanzania AlQaida) Al-Qaida bombed the Embassy in Tanzania.

General Knowledge about Terrorism

Specific data, facts, and observations about terrorist groups and activities

Example of automatic translating text into Cyc Logic

Source: "Galileo Galilei was an Italian physicist and astronomer."

Learn Logic:
```
(#$and (#$isa #$GalileoGalilei #$ItalianPerson)
       (#$isa #$GalileoGalilei #$Physicist)
       (#$isa #$GalileoGalilei #$Astronomer))
```

Fact: Galileo was an Italian, a physicist, and an astronomer.

Source: "Galileo was born in Pisa on Feburary 15, 1564."

Learn Logic:
```
(#$and (#$birthDate #$GalileoGalilei
                    (#$DayFn 15
                    (#$MonthFn #$February
                    (#$YearFn 1564))))
       (#$birthPlace #$GalileoGalilei #$CityOfPisaItaly))
```

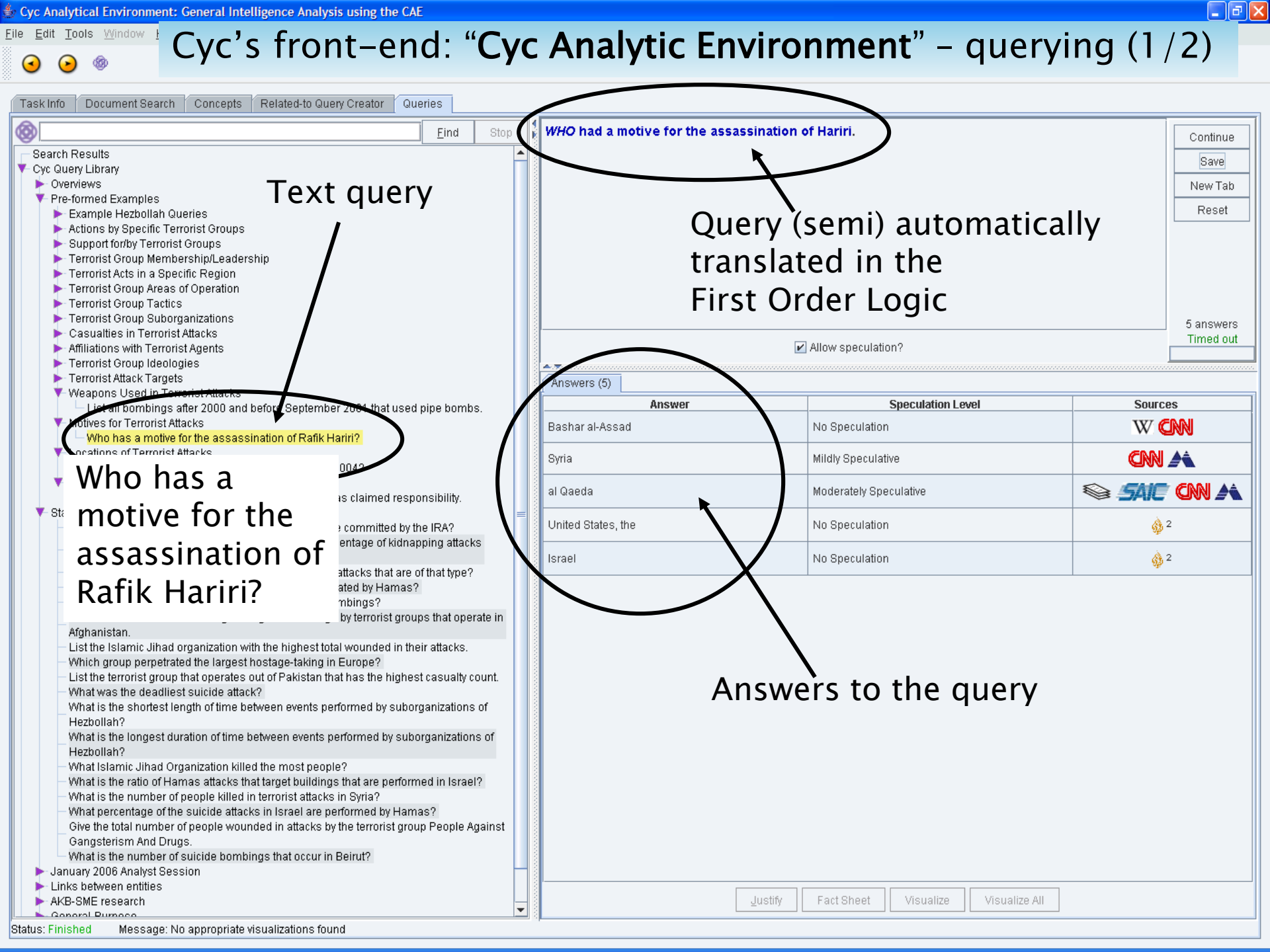
Fact: Galileo was born on February 15, 1564 and he was born in Pisa.

Source: "Albert Einstein was born in 1879 in Ulm, Germany."

Learn Logic: `(#$birthDate #$AlbertEinstein (#$YearFn 1879))`

Fact: Albert Einstein was born in 1879.

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



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

Cyc Analytical Environment: General Intelligence Analysis using the CAE

File Edit Tools Wind

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?
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

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.

▼ Options

▼ Options

Detailed Justification:
▶ Al Qaeda had a motive for the assassination of Hariri.
External Sources:

1 Gary C. Gambill, "Dossier: Rafiq Hariri", *United States Committee for a Free Lebanon,* July 2001, http://www.meib.org/articles/0107_ld1.htm .

2 "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/ .

Query & Answer

Justification

Sources for Reasoning and Justification

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

Cyc Analytical Environment: General Intelligence Analysis using the CAE

File Edit Tools Wind

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?
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

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.

▼ Options

▼ Options

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_ld1.htm .

"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/ .

Query & Answer

Justification

Sources for Reasoning and Justification

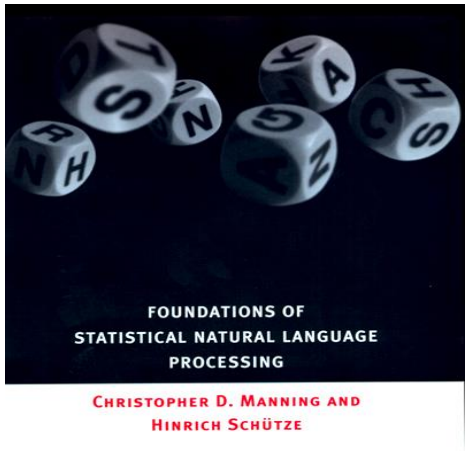
If

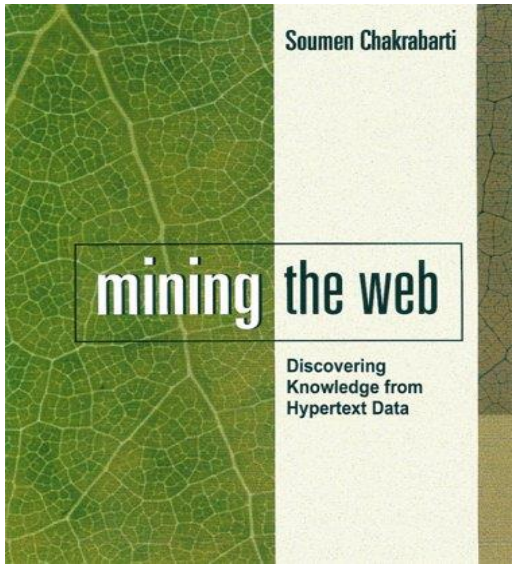
- some intelligent agent opposes some policy,
- and some other intelligent agent *VICTIM* is an advocate of that policy,
- and some other intelligent agent *ADOPTER* is responsible for according with the policy,
- and it is adopted by *ADOPTER* in any some *ADOPT-TYPE*,
- and some *ACT* prevents *VICTIM*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

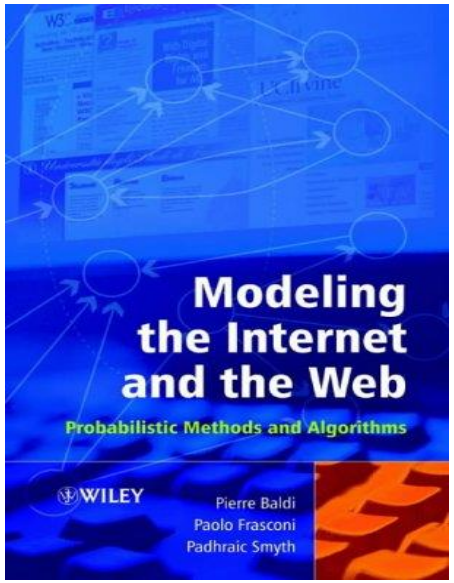
then that intelligent agent had a motive for *ACT*.

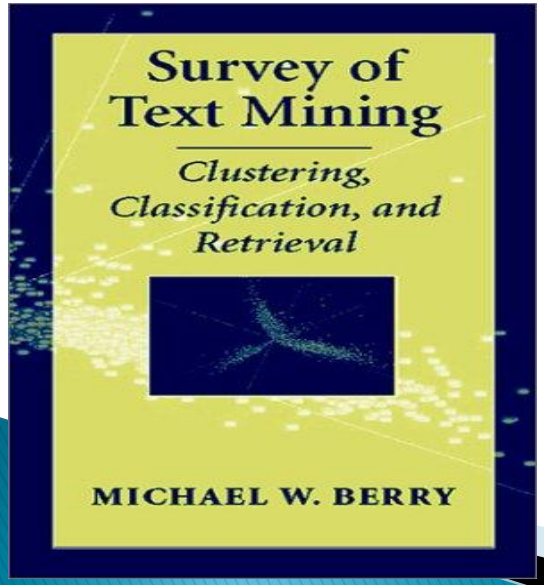
Further references…

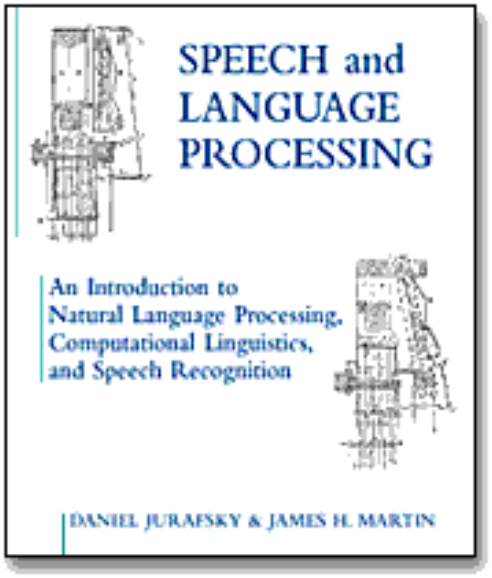
References to some Text-Mining books

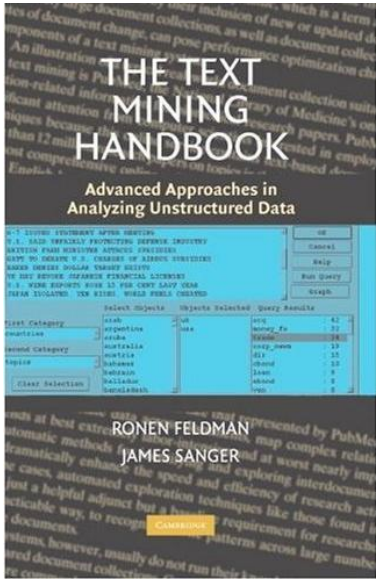




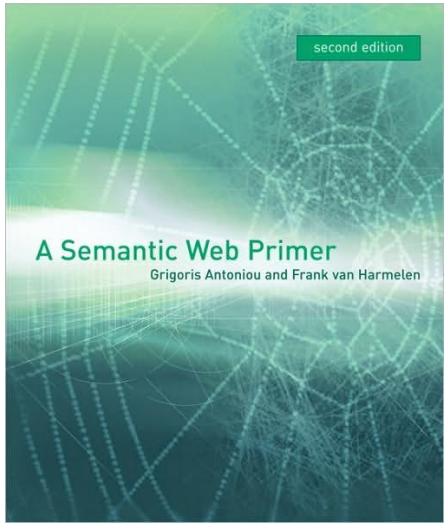


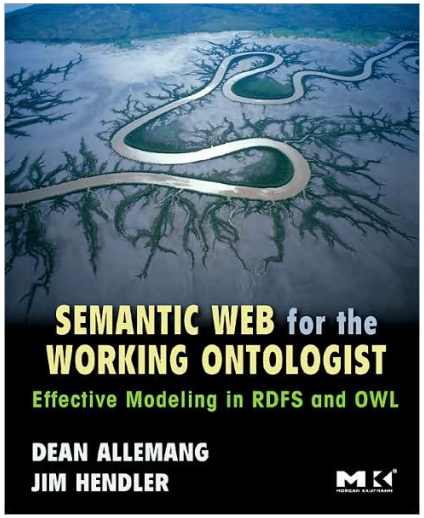


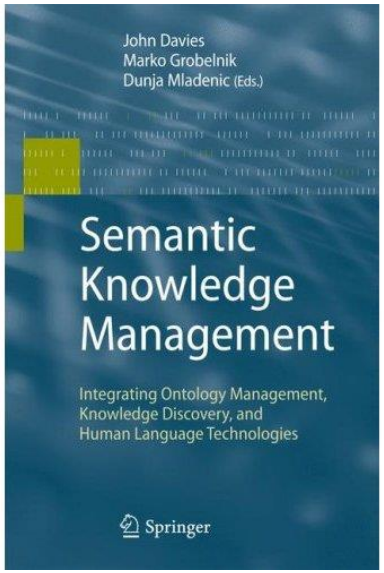


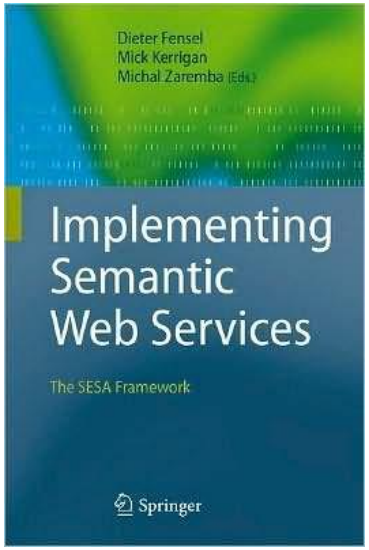


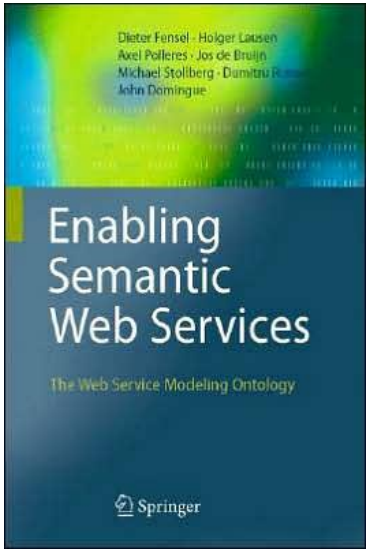
Books on Semantic Technologies

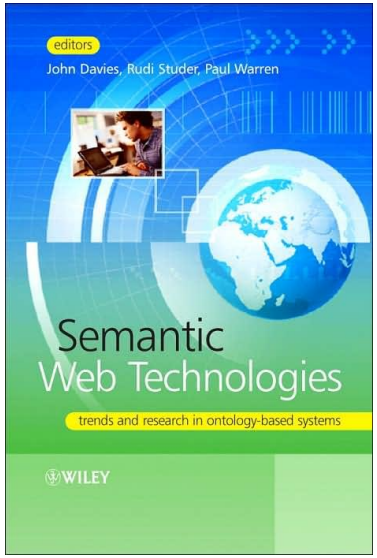












References to the main conferences

- 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://videolectures.net/Top/Computer_Science/Semantic_Web/

- Recorded tutorials, workshops, conferences, summer schools available from http://videolectures.net