Tutorial on **Text Mining** and **Link Analysis** for **Web** and **Semantic Web**

Marko Grobelnik, Dunja Mladenic
Jozef Stefan Institute
Ljubljana, Slovenia

ECML/PKDD 2008, Antwerp, September 15th 2008

Outline

- **Text-Mining**
  - How to deal with text data on various levels?
- **Link-Analysis**
  - How to analyze graphs in the Web context?
- **Semantic-Web**
  - How semantics fits into the picture?
- Wrap-up
  - …what did we learn and where to continue?

Text-Mining

How to deal with text data on various levels?

Why do we analyze text?

- The ultimate goal (or “the mother of all tasks”) is understanding of textual content…
- …but, since this seems to be too hard task, we have number of easier sub-tasks of some importance which we are able to deal with.

What is Text-Mining?

- "…finding **interesting** regularities in large **textual** datasets…" (adapted from Usama Fayad)
  - …where **interesting** means: non-trivial, hidden, previously unknown and potentially useful
- "…finding semantic and abstract information from the surface form of textual data…"

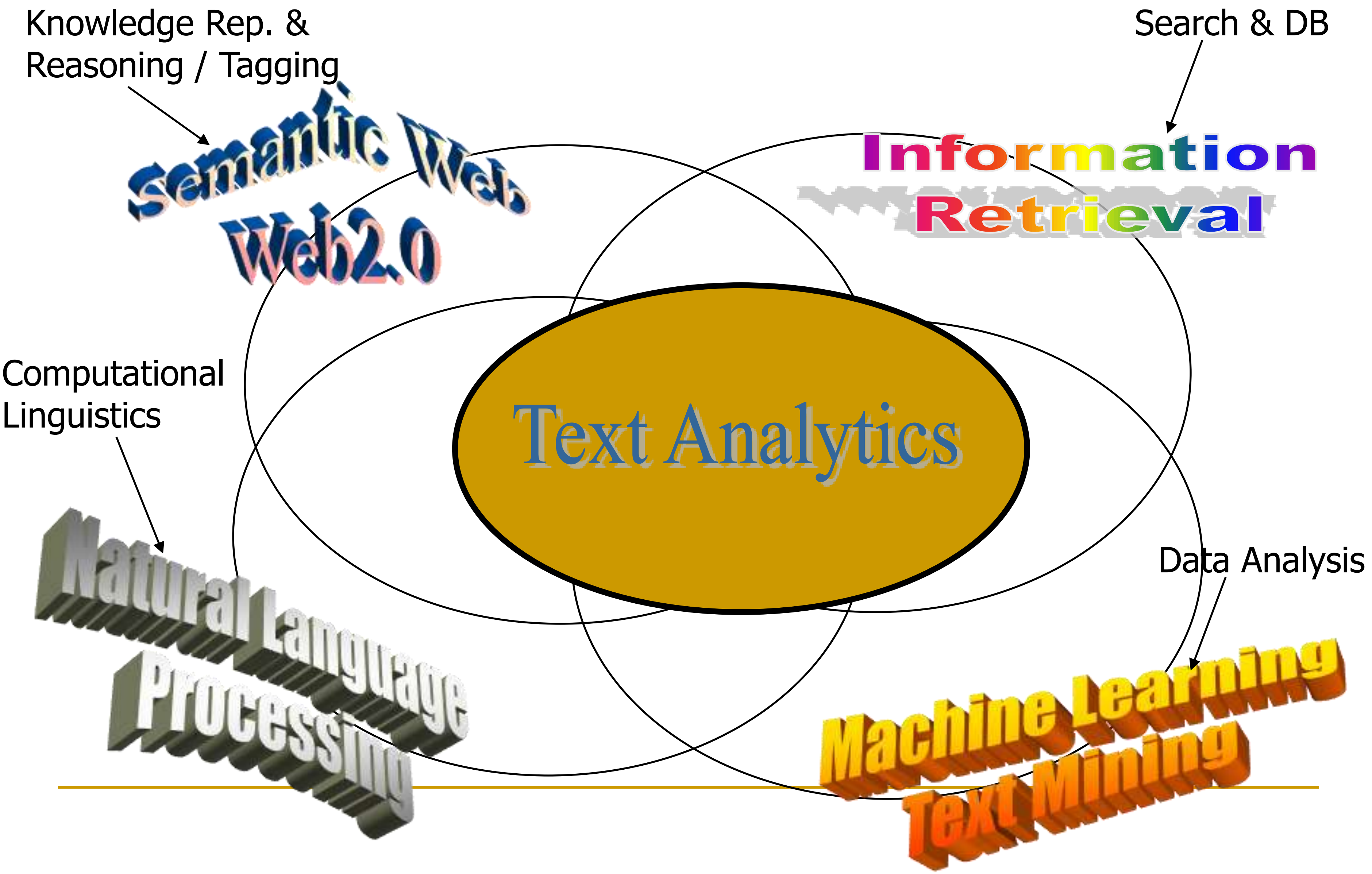
Why dealing with Text is Tough? (M.Hearst 97)

- Abstract concepts are **difficult to represent**
- **“Countless” combinations** of subtle, abstract relationships among concepts
- **Many ways** to represent similar concepts
  - E.g. space ship, flying saucer, UFO
- Concepts are **difficult to visualize**
- **High dimensionality**
- **Tens or hundreds of thousands of features**

Why dealing with Text is Easy? (M.Hearst 97)

- **Highly redundant data**
  - …most of the methods count on this property
- **Just about any simple algorithm can get "good" results for simple tasks:**
  - Pull out "important" phrases
  - Find "meaningfully" related words
  - Create some sort of summary from documents

Who is in the text analysis arena?



Knowledge Rep. & Reasoning / Tagging

Search & DB

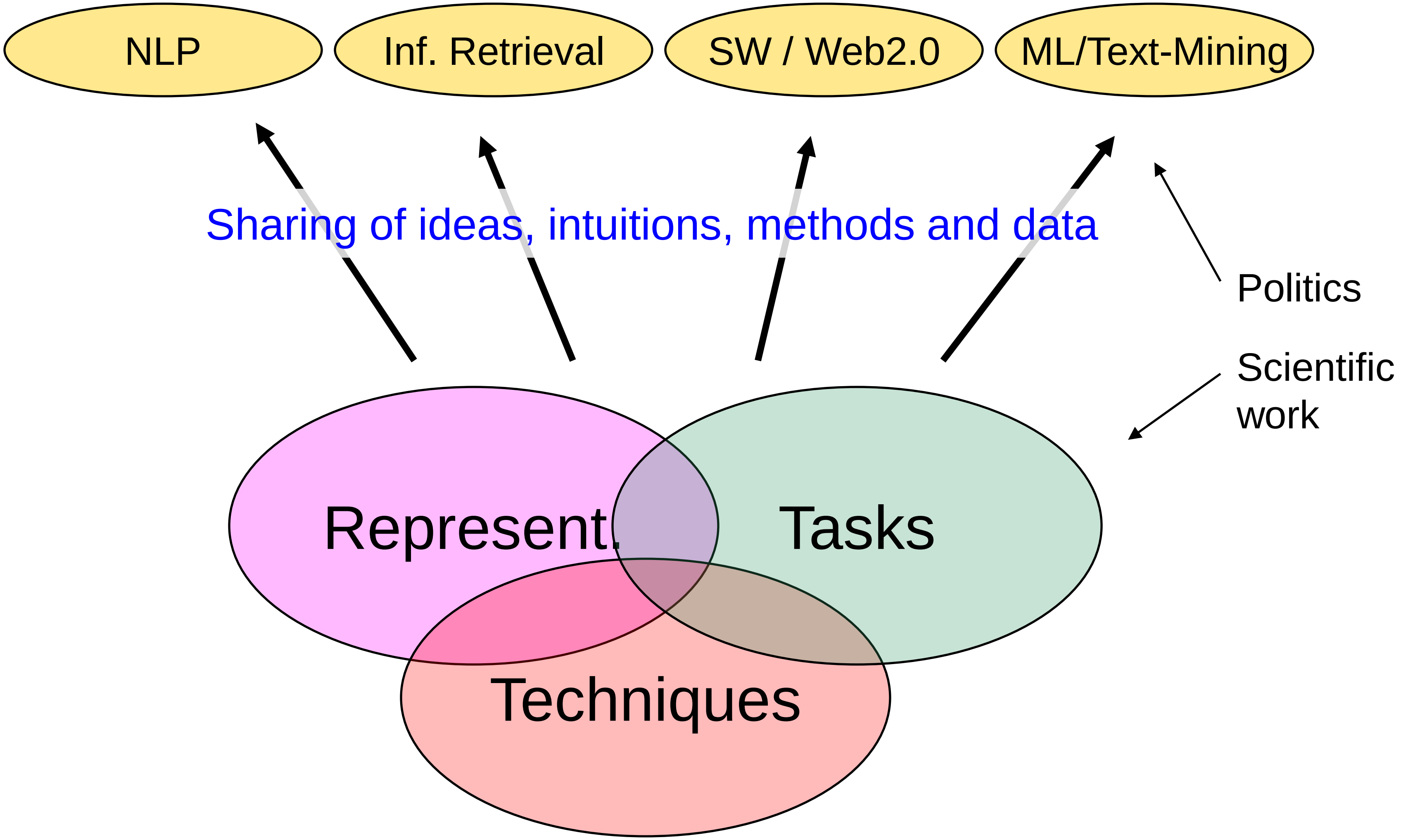
Computational Linguistics

Data Analysis

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 learning to reasoning
  - Tasks
    - …from search, over (un-, semi-) supervised learning, to visualization, summarization, translation …

How dimensions fit to research areas?



NLP

Inf. Retrieval

SW / Web2.0

ML/Text-Mining

Sharing of ideas, intuitions, methods and data

Politics

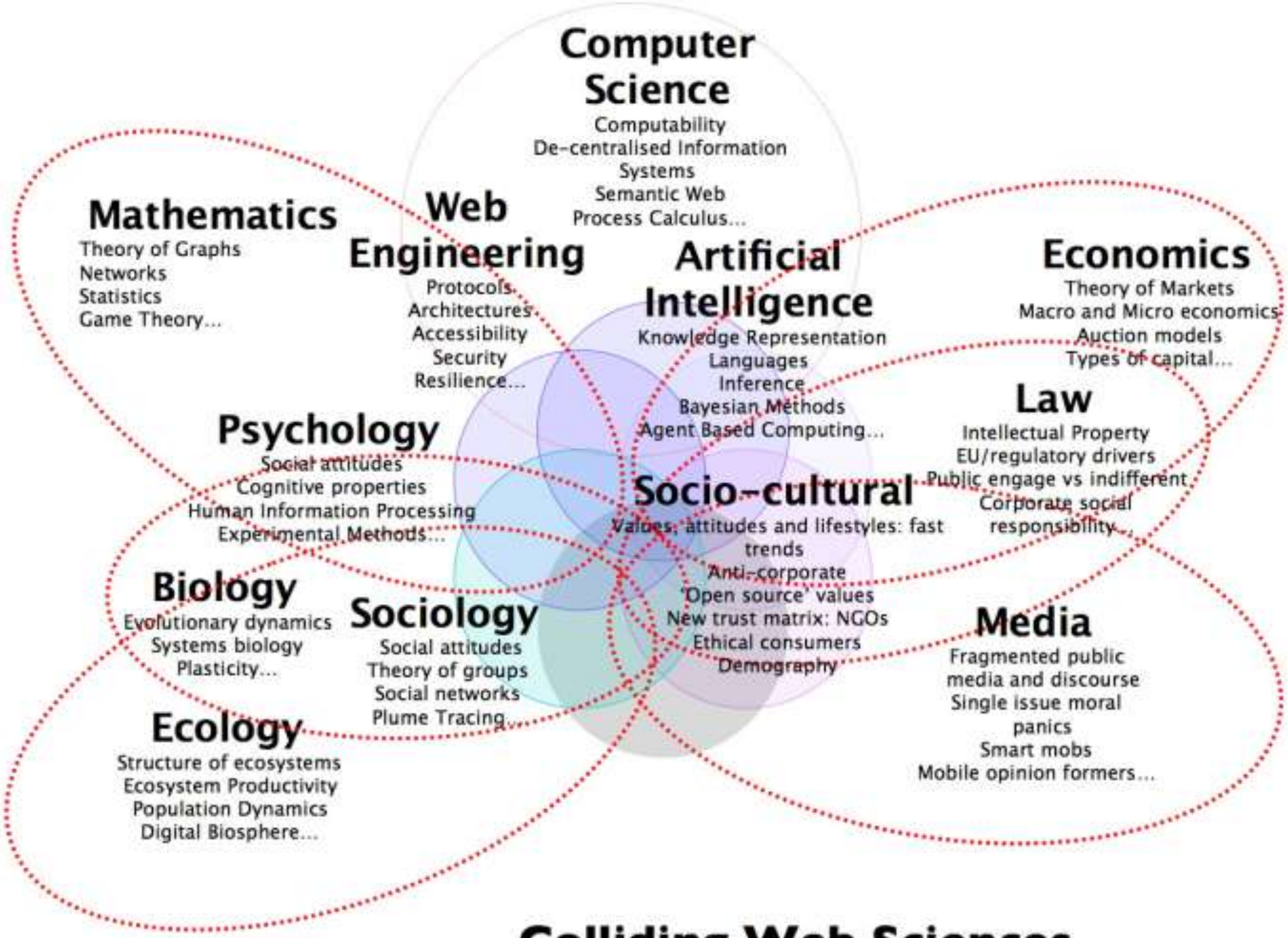
Scientific work

Represent.

Tasks

Techniques

Broader context: Web Science



Colliding Web Sciences

http://webscience.org/

Text-Mining

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

Lexical

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

Syntactic

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

Semantic

Levels of text representations

- **Character**
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

Semantic

Character level

- Character level representation of a text consists from sequences of characters…
  - …a document is represented by a frequency distribution of sequences
  - Usually we deal with contiguous strings…
  - …each character sequence of length 1, 2, 3, … represent a feature with its frequency

Good and bad sides

- Representation has several important strengths:
    - …it is very robust since avoids language morphology
        - (useful for e.g. language identification)
    - …it captures simple patterns on character level
        - (useful for e.g. spam detection, copy detection)
    - …because of redundancy in text data it could be used for many analytic tasks
        - (learning, clustering, search)
        - It is used as a basis for “string kernels” in combination with SVM for capturing complex character sequence patterns
- …for deeper semantic tasks, the representation is too weak

Levels of text representations

- Character
- **Words**
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

Semantic

Word level

- The most common representation of text used for many techniques
  - …there are many tokenization software packages which split text into the words
- Important to know:
  - Word is well defined unit in western languages – e.g. Chinese has different notion of semantic unit

Words Properties

- Relations among word surface forms and their senses:
  - **Homonomy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of an another (e.g. breakfast, meal)

- Word frequencies in texts have **power distribution**:
  - …small number of very frequent words
  - …big number of low frequency words

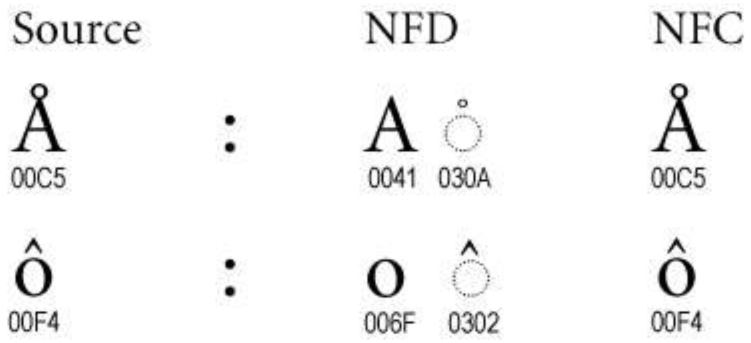
Stop-words

- Stop-words are words that from non-linguistic view do not carry information
  - …they have mainly functional role
  - …usually we remove them to help the methods to perform better

- Stop words are language dependent – examples:
  - **English**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...
  - **Dutch**: de, en, van, ik, te, dat, die, in, een, hij, het, niet, zijn, is, was, op, aan, met, als, voor, had, er, maar, om, hem, dan, zou, of, wat, mijn, men, dit, zo, ...
  - **Slovenian**: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...

Word character level normalization

- Hassle which we usually avoid:
  - Since we have plenty of character encodings in use, it is often nontrivial to identify a word and write it in unique form
  - …e.g. in Unicode the same word could be written in many ways – canonization of words:



Stemming (1/2)

- Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
- Stemming is a process of transforming a word into its stem (normalized form)
  - …stemming provides an inexpensive mechanism to merge

Stemming (2/2)

- For English is mostly used Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- Example cascade rules used in English Porter stemmer

❑ ATIONAL	->	ATE	relational -> relate
❑ TIONAL	->	TION	conditional -> condition
❑ ENCI	->	ENCE	valenci -> valence
❑ ANCI	->	ANCE	hesitanci -> hesitance
❑ IZER	->	IZE	digitizer -> digitize
❑ ABLI	->	ABLE	conformabli -> conformable
❑ ALLI	->	AL	radicalli -> radical
❑ ENTLI	->	ENT	differentli -> different
❑ ELI	->	E	vileli -> vile
❑ OUSLI	->	OUS	analogousli -> analogous

Levels of text representations

- Character
- Words
- **Phrases**
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

Semantic

Phrase level

- Instead of having just single words we can deal with phrases
- We use two types of phrases:
  - Phrases as frequent contiguous word sequences
  - Phrases as frequent non-contiguous word sequences
  - …both types of phrases could be identified by simple dynamic programming algorithm
- The main effect of using phrases is to more precisely identify sense

Google n-gram corpus

- In September 2006 Google announced availability of n-gram corpus:
  - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html#links
  - Some statistics of the corpus:
    - File sizes: approx. 24 GB compressed (gzip'ed) text files
    - Number of tokens: 1,024,908,267,229
    - Number of sentences: 95,119,665,584
    - Number of unigrams: 13,588,391
    - Number of bigrams: 314,843,401
    - Number of trigrams: 977,069,902
    - Number of fourgrams: 1,313,818,354
    - Number of fivegrams: 1,176,470,663

Example: Google n-grams

- ceramics collectables collectibles 55
  ceramics collectables fine 130
  ceramics collected by 52
  ceramics collectible pottery 50
  ceramics collectibles cooking 45
  ceramics collection , 144
  ceramics collection . 247
  ceramics collection </S> 120
  ceramics collection and 43
  ceramics collection at 52
  ceramics collection is 68
  ceramics collection of 76
  ceramics collection | 59
  ceramics collections , 66
  ceramics collections . 60
  ceramics combined with 46
  ceramics come from 69
  ceramics comes from 660
  ceramics community , 109
  ceramics community . 212
  ceramics community for 61
  ceramics companies . 53
  ceramics companies consultants 173
  ceramics company ! 4432
  ceramics company , 133
  ceramics company . 92
  ceramics company </S> 41
  ceramics company facing 145
  ceramics company in 181
  ceramics company started 137
  ceramics company that 87
  ceramics component (76
  ceramics composed of 85

- serve as the incoming 92
  serve as the incubator 99
  serve as the independent 794
  serve as the index 223
  serve as the indication 72
  serve as the indicator 120
  serve as the indicators 45
  serve as the indispensable 111
  serve as the indispensible 40
  serve as the individual 234
  serve as the industrial 52
  serve as the industry 607
  serve as the info 42
  serve as the informal 102
  serve as the information 838
  serve as the informational 41
  serve as the infrastructure 500
  serve as the initial 5331
  serve as the initiating 125
  serve as the initiation 63
  serve as the initiator 81
  serve as the injector 56
  serve as the inlet 41
  serve as the inner 87
  serve as the input 1323
  serve as the inputs 189
  serve as the insertion 49
  serve as the insourced 67
  serve as the inspection 43
  serve as the inspector 66
  serve as the inspiration 1390
  serve as the installation 136
  serve as the institute 187

Levels of text representations

- Character
- Words
- Phrases
- **Part-of-speech tags**
- Taxonomies / thesauri

Lexical

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

Syntactic

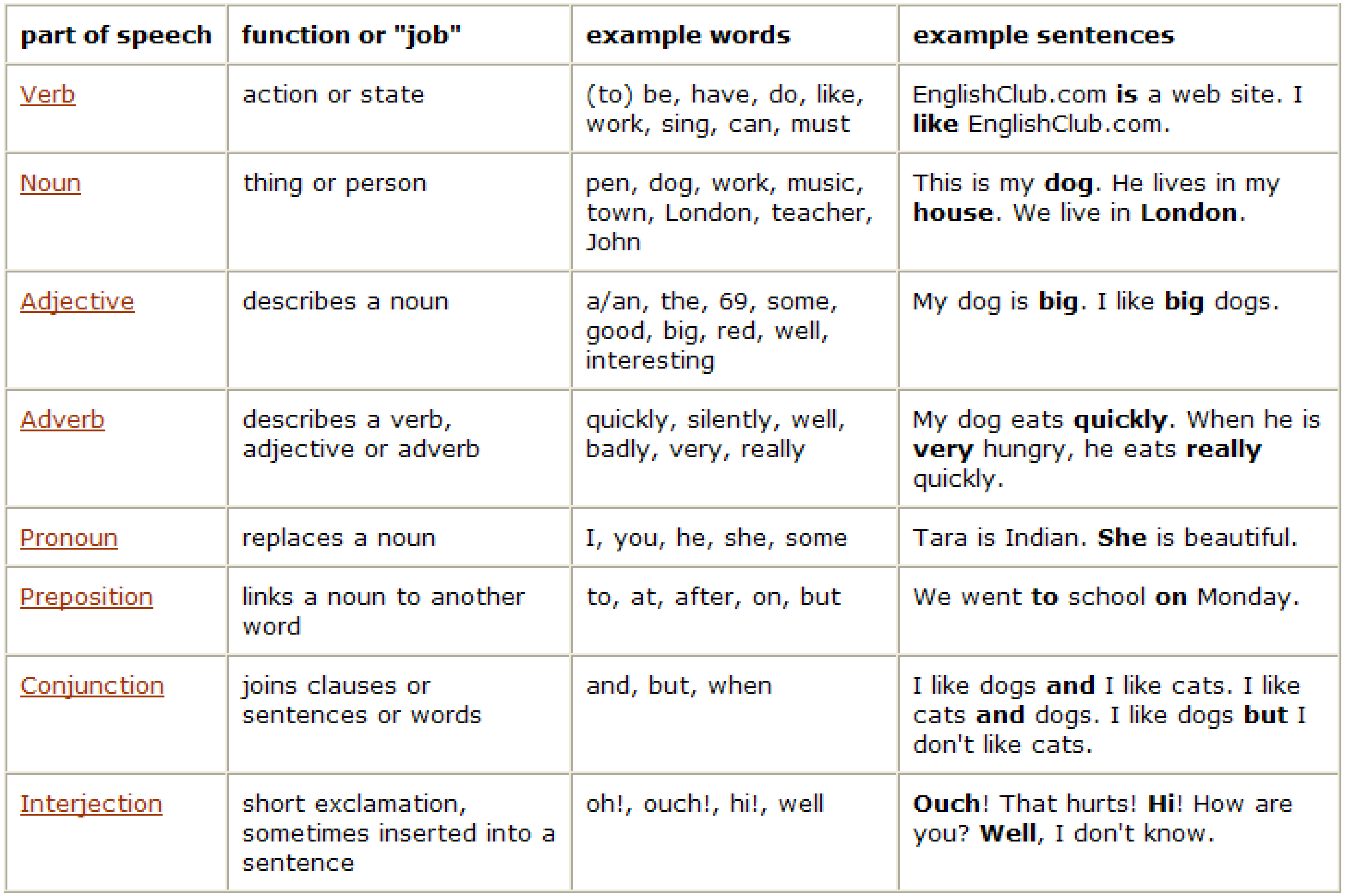
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Semantic

Part-of-Speech level

- By introducing part-of-speech tags we introduce word-types enabling to differentiate words functions
  - For text-analysis part-of-speech information is used mainly for “information extraction” where we are interested in e.g. named entities which are “noun phrases”
  - Another possible use is reduction of the vocabulary (features)
    - …it is known that nouns carry most of the information in text documents
- Part-of-Speech taggers are usually learned by HMM algorithm on manually tagged data

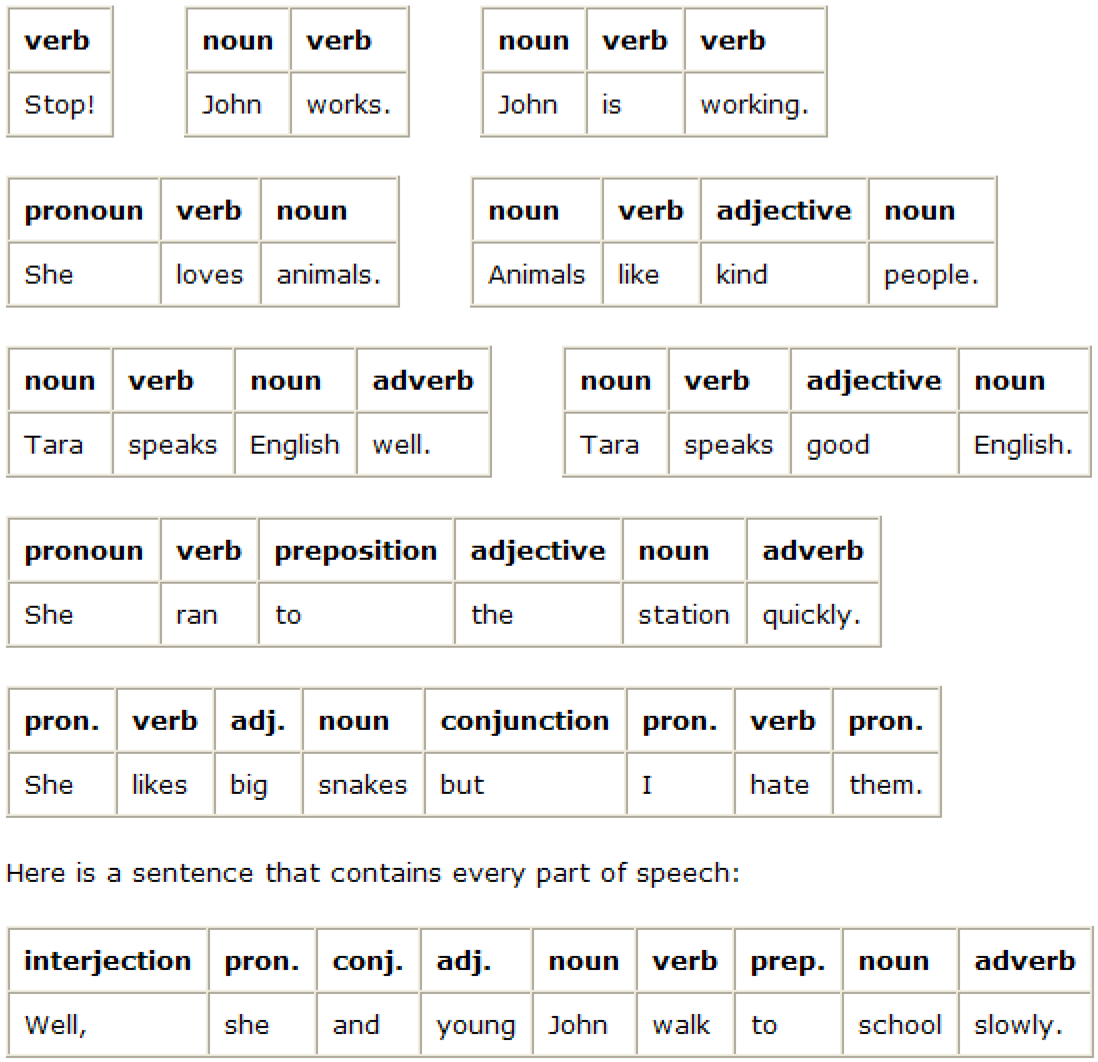
Part-of-Speech Table



part of speech	function or "job"	example words	example sentences
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats **quickly**. When he is **very** hungry, he eats **really** quickly.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	joins clauses or sentences or words	and, but, when	I like dogs **and** I like cats. I like cats **and** dogs. I like dogs **but** I don't like cats.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

http://www.englishclub.com/grammar/parts-of-speech_1.htm

Part-of-Speech examples



verb
Stop!

noun	verb
John	works.

noun	verb	verb
John	is	working.

pronoun	verb	noun
She	loves	animals.

noun	verb	adjective	noun
Animals	like	kind	people.

noun	verb	noun	adverb
Tara	speaks	English	well.

noun	verb	adjective	noun
Tara	speaks	good	English.

pronoun	verb	preposition	adjective	noun	adverb
She	ran	to	the	station	quickly.

pron.	verb	adj.	noun	conjunction	pron.	verb	pron.
She	likes	big	snakes	but	I	hate	them.

Here is a sentence that contains every part of speech:

interjection	pron.	conj.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	young	John	walk	to	school	slowly.

http://www.englishclub.com/grammar/parts-of-speech_2.htm

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- **Taxonomies / thesauri**

Lexical

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

Syntactic

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

Semantic

Taxonomies/thesaurus level

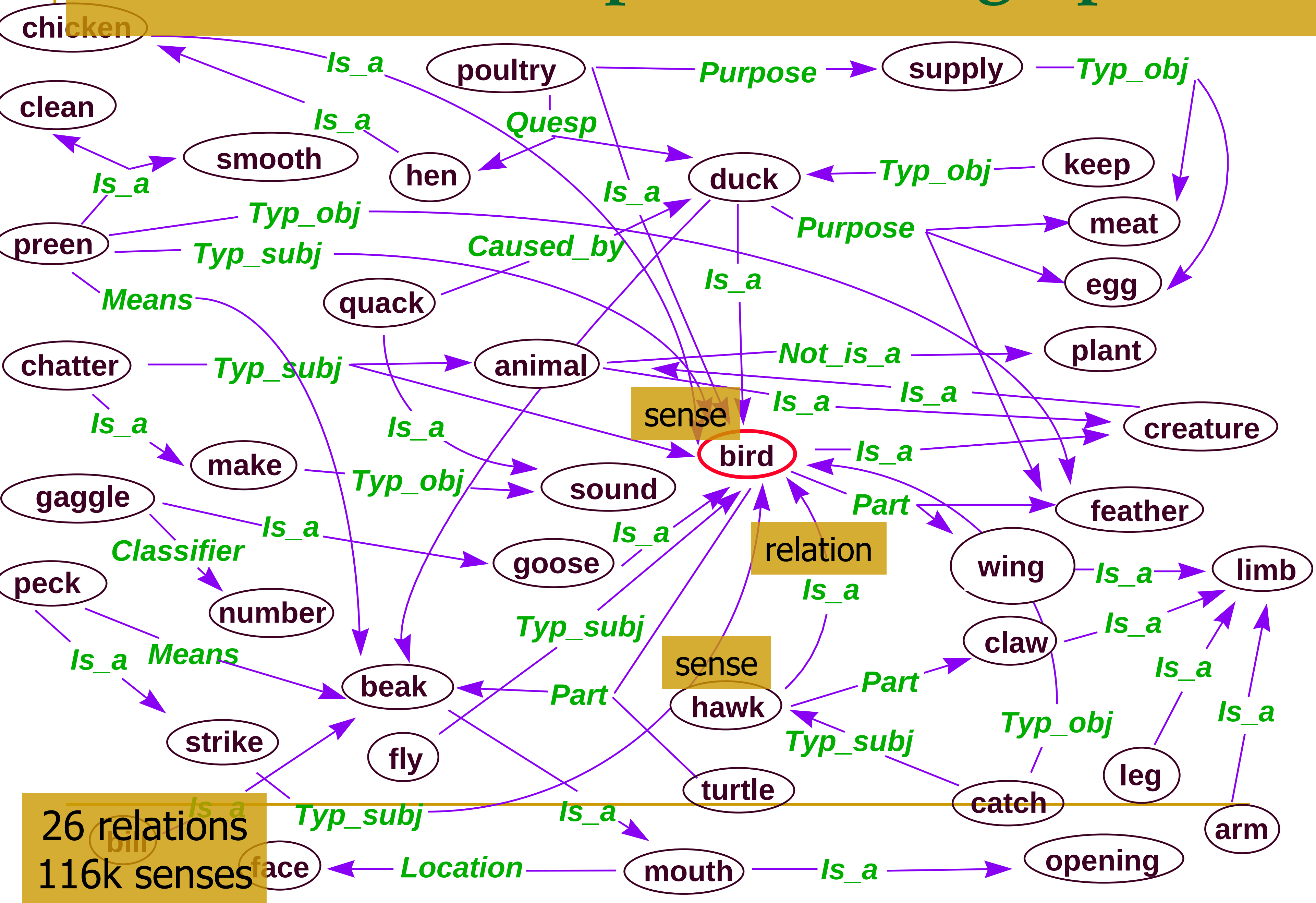
- Thesaurus has a main function to connect different surface word forms with the same meaning into one sense (synonyms)
  - …additionally we often use hypernym relation to relate general-to-specific word senses
  - …by using synonyms and hypernym relation we compact the feature vectors
- The most commonly used general thesaurus is WordNet which exists in many other languages (e.g. EuroWordNet)
  - http://www.illc.uva.nl/EuroWordNet/

WordNet – database of lexical relations

- WordNet is the most well developed and widely used lexical database for English
  - …it consist from 4 databases (nouns, verbs, adjectives, and adverbs)
- Each database consists from sense entries – each sense consists from a set of synonyms, e.g.:
  - musician, instrumentalist, player
  - person, individual, someone
  - life form, organism, being

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

WordNet – excerpt from the graph



sense

relation

sense

26 relations
116k senses

WordNet relations

- Each WordNet entry is connected with other entries in the graph through relations
- Relations in the database of nouns:

Relation	Definition	Example
Hypernym	From lower to higher concepts	breakfast -> meal
Hyponym	From concepts to subordinates	meal -> lunch
Has-Member	From groups to their members	faculty -> professor
Member-Of	From members to their groups	copilot -> crew
Has-Part	From wholes to parts	table -> leg
Part-Of	From parts to wholes	course -> meal
Antonym	Opposites	leader -> follower

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

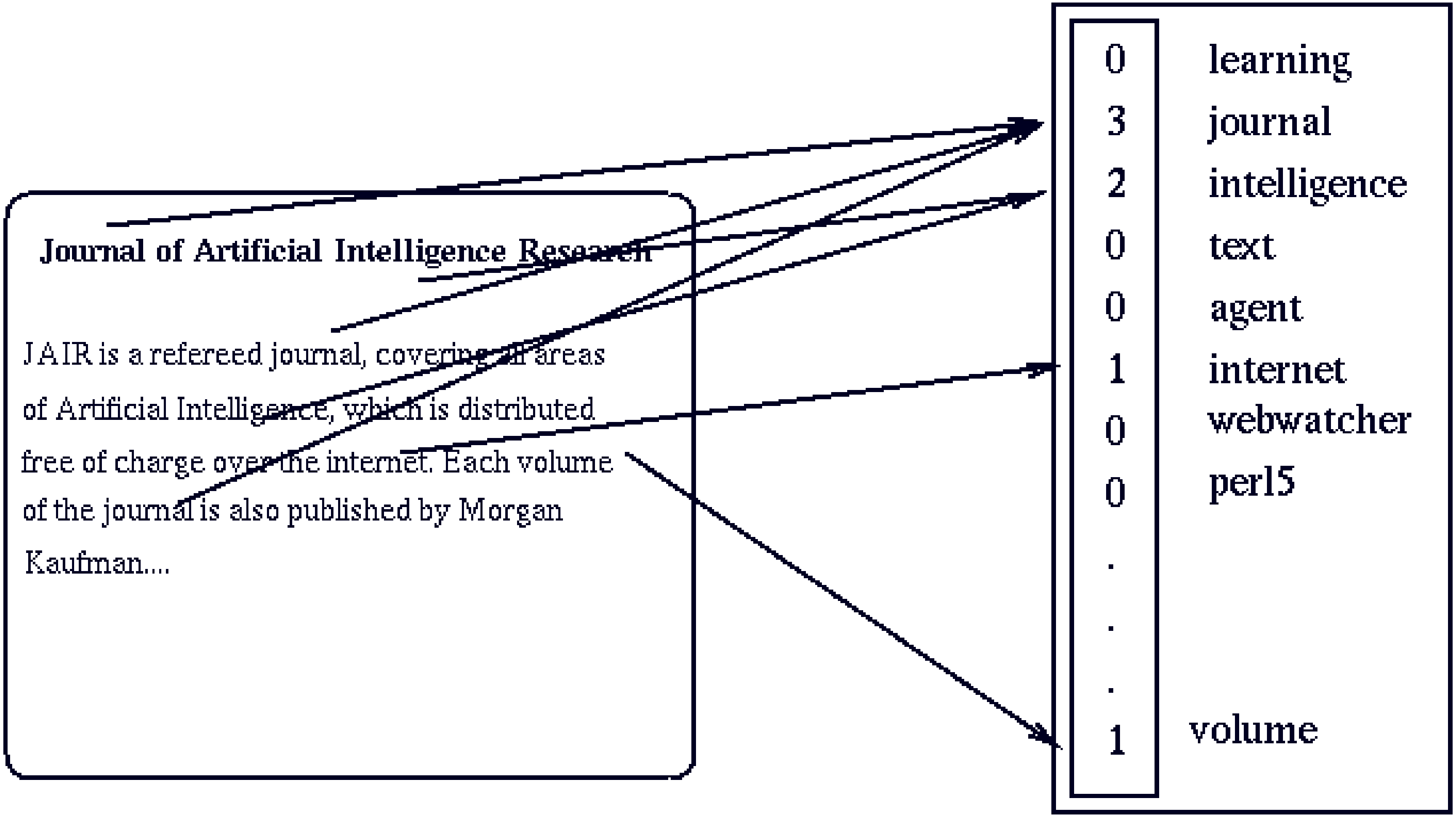
- Collaborative tagging / Web2.0
- Templates / Frames
- Ontologies / First order theories

Semantic

Vector-space model level

- The most common way to deal with documents is first to transform them into **sparse numeric vectors** and then deal with them with **linear algebra operations**
  - …by this, we forget everything about the linguistic structure within the text
  - …this is sometimes called “structural curse” because this way of forgetting about the structure doesn’t harm efficiency of solving many relevant problems
  - This representation is referred to also as “Bag-Of-Words” or “Vector-Space-Model”
  - Typical tasks on vector-space-model are classification, clustering, visualization etc.

Bag-of-words document representation


Journal of Artificial Intelligence Research
JAIR is a refereed journal, covering all areas of Artificial Intelligence, which is distributed free of charge over the internet. Each volume of the journal is also published by Morgan Kaufman....
0 learning
3 journal
2 intelligence
0 text
0 agent
1 internet
0 webwatcher
0 perl5
.
.
.
1 volume

Word weighting

- In the bag-of-words representation each word is represented as a separate variable having numeric weight (importance)
- The most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf \,.\, \log\left(\frac{N}{df(w)}\right)$$

- Tf(w) – term frequency (number of word occurrences in a document)
- Df(w) – document frequency (number of documents containing the word)
- N – number of all documents
- TfIdf(w) – relative importance of the word in the document

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Example document and its vector representation

- TRUMP MAKES BID FOR CONTROL OF RESORTS Casino owner and real estate Donald Trump has offered to acquire all Class B common shares of Resorts International Inc, a spokesman for Trump said. The estate of late Resorts chairman James M. Crosby owns 340,783 of the 752,297 Class B shares. Resorts also has about 6,432,000 Class A common shares outstanding. Each Class B share has 100 times the voting power of a Class A share, giving the Class B stock about 93 pct of Resorts' voting power.

Original text

- [RESORTS:0.624] [CLASS:0.487] [TRUMP:0.367] [VOTING:0.171] [ESTATE:0.166] [POWER:0.134] [CROSBY:0.134] [CASINO:0.119] [DEVELOPER:0.118] [SHARES:0.117] [OWNER:0.102] [DONALD:0.097] [COMMON:0.093] [GIVING:0.081] [OWNS:0.080] [MAKES:0.078] [TIMES:0.075] [SHARE:0.072] [JAMES:0.070] [REAL:0.068] [CONTROL:0.065] [ACQUIRE:0.064] [OFFERED:0.063] [BID:0.063] [LATE:0.062] [OUTSTANDING:0.056] [SPOKESMAN:0.049] [CHAIRMAN:0.049] [INTERNATIONAL:0.041] [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Bag-of-Words representation (high dimensional sparse vector)

Similarity between document vectors

- Each document is represented as a vector of weights D = <x>
- Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
  - …calculates cosine of the angle between document vectors
  - …efficient to calculate (sum of products of intersecting words)
  - …similarity value between 0 (different) and 1 (the same)

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2}\sqrt{\sum_k x_k^2}}$$

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

Semantic

Language model level

- Language modeling is about determining probability of a sequence of words
  - The task typically gets reduced to the estimating probabilities of a next word given two previous words (trigram model):

$$P(w_i|w_{i-2}w_{i-1}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$$

Frequencies of word sequences

  - It has many applications including speech recognition, OCR, handwriting recognition, machine translation and spelling correction

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

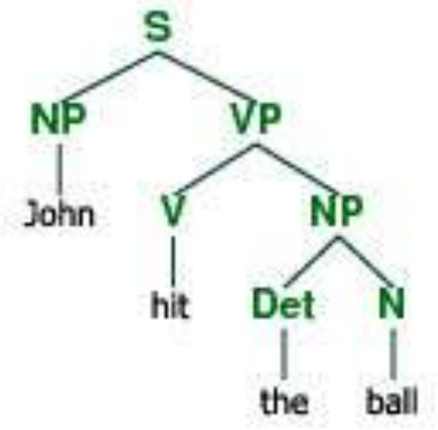
Syntactic

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

Semantic

Full-parsing level

- Parsing provides maximum structural information per sentence
- On the input we get a sentence, on the output we generate a parse tree
- For most of the methods dealing with the text data the information in parse trees is too complex



Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

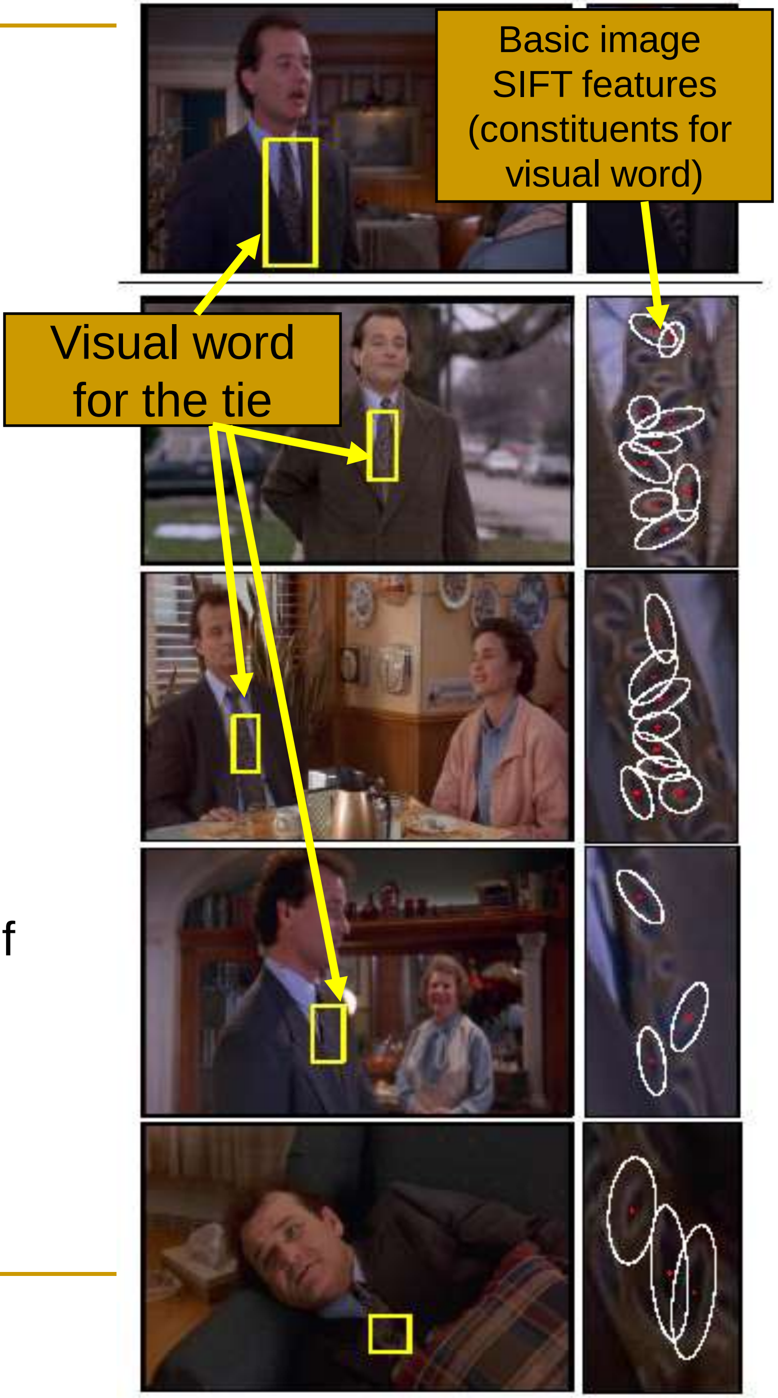
Semantic

Cross-modality level

- It is very often the case that objects are represented with different data types:
  - Text documents
  - Multilingual texts documents
  - Images
  - Video
  - Social networks
  - Sensor networks
- …the question is how to create mappings between different representation so that we can benefit using more information about the same objects

Example: Aligning text with audio, images and video

- The word “**tie**” has several representations (http://www.answers.com/tie&r=67)
  - Textual
  - Multilingual text
    - (tie, kravata, krawatte, …)
  - Audio
  - Image:
    - http://images.google.com/images?hl=en&q=necktie
  - Video (movie on the right)
- Out of each representation we can get set of features and the idea is to correlate them
  - KCCA (Kernel Correlation Analysis) method generates mappings between different representations into “**modality neutral**” data representation



Basic image SIFT features (constituents for visual word)

Visual word for the tie

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

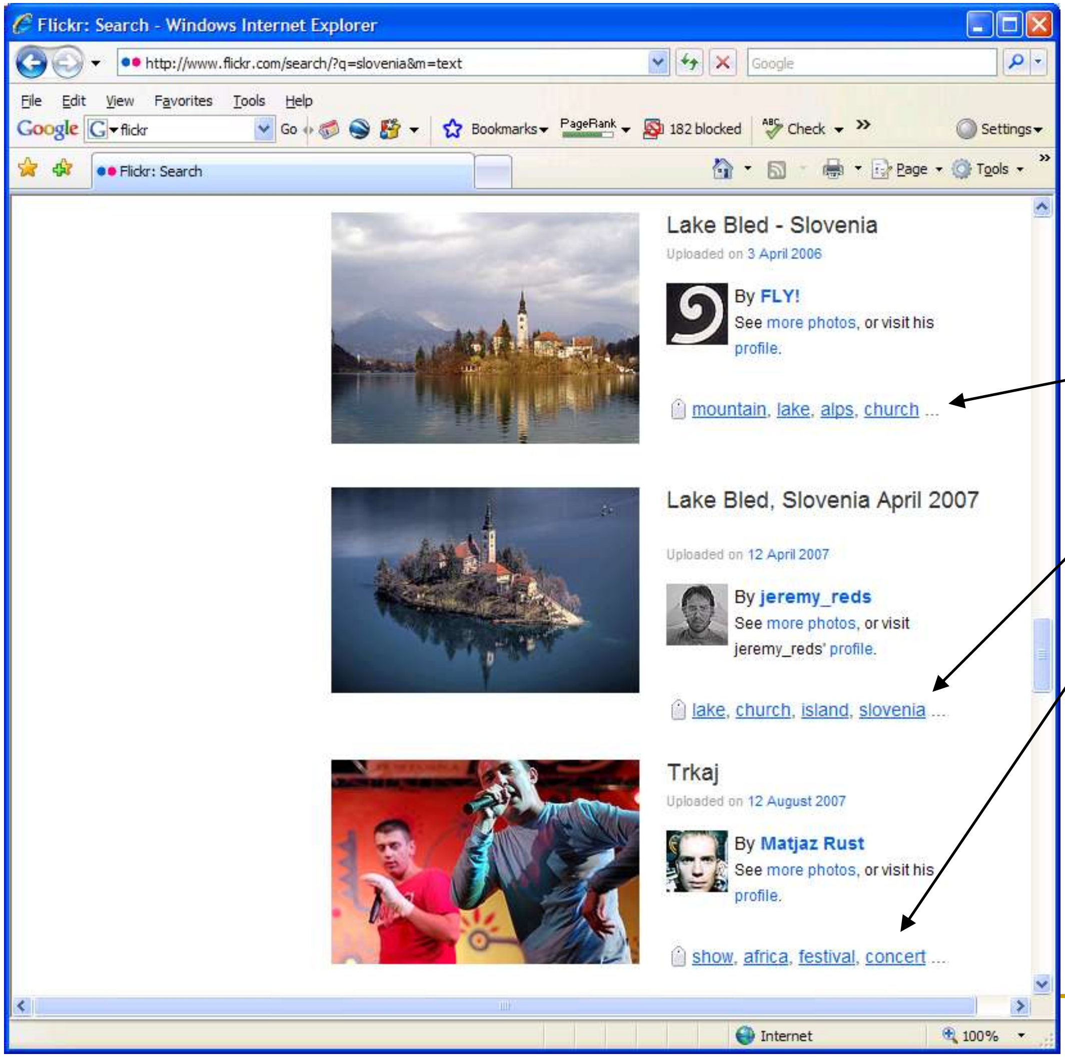
- **Collaborative tagging / Web2.0**
- Templates / Frames
- Ontologies / First order theories

Semantic

Collaborative tagging

- Collaborative tagging is a process of adding metadata to annotate content (e.g. documents, web sites, photos)
  - …metadata is typically in the form of keywords
  - …this is done in a collaborative way by many users from larger community collectively having good coverage of many topics
  - …as a result we get annotated data where tags enable comparability of annotated data entries

Example: flickr.com tagging


Flickr: Search - Windows Internet Explorer
http://www.flickr.com/search/?q=slovenia&m=text
Google
File Edit View Favorites Tools Help
Google flickr Go Bookmarks PageRank 182 blocked Check Settings
Flickr: Search
Page Tools
Internet
100%

Lake Bled - Slovenia

Uploaded on 3 April 2006

By FLY!
See more photos, or visit his profile.

mountain, lake, alps, church ...

Lake Bled, Slovenia April 2007

Uploaded on 12 April 2007

By jeremy_reds
See more photos, or visit jeremy_reds' profile.

lake, church, island, slovenia ...

Trkaj

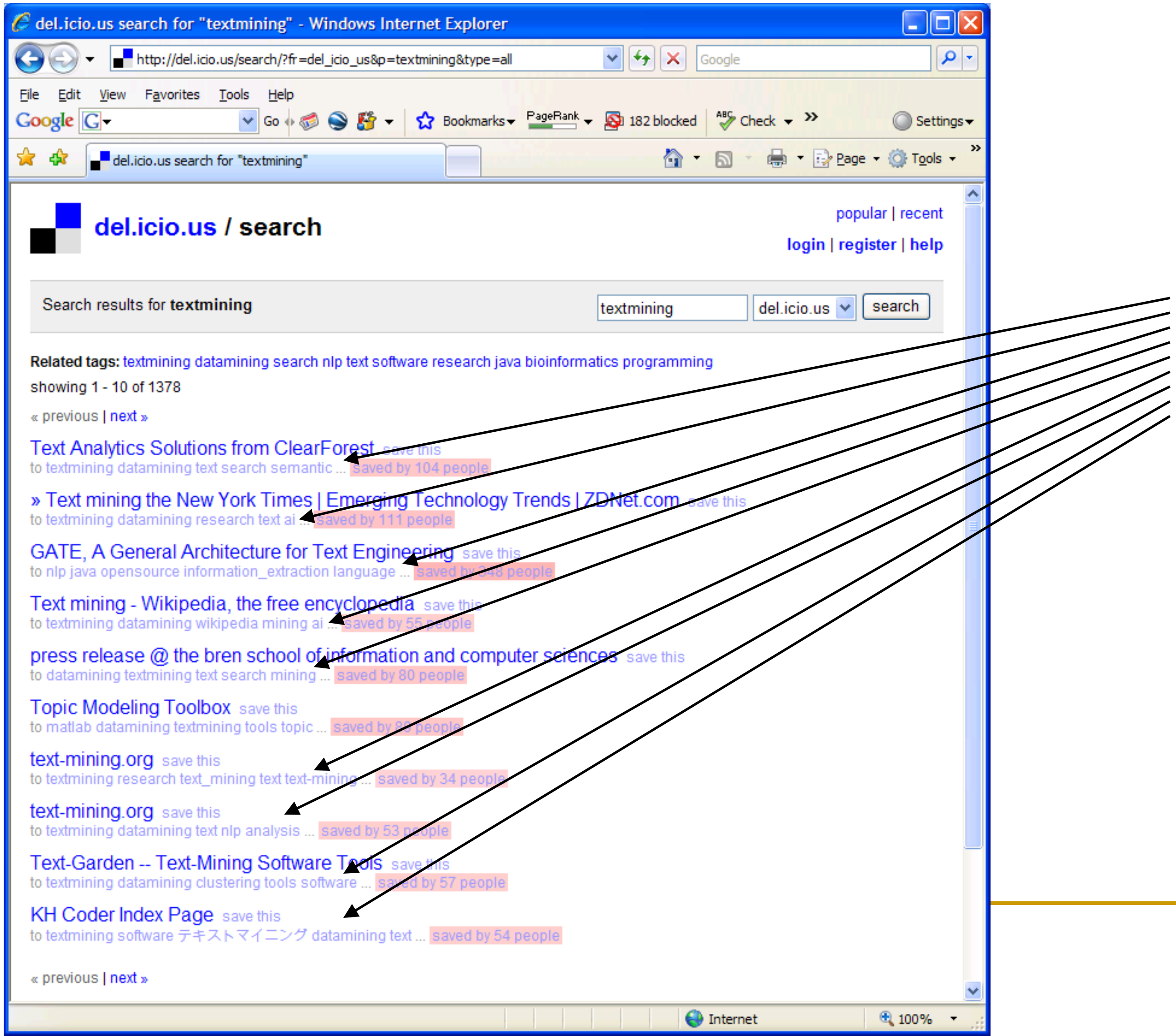
Uploaded on 12 August 2007

By Matjaz Rust
See more photos, or visit his profile.

show, africa, festival, concert ...

Tags entered by users annotating photos

Example: del.icio.us tagging



del.icio.us search for "textmining" - Windows Internet Explorer
http://del.icio.us/search/?fr=del_icio_us&p=textmining&type=all
Google
File Edit View Favorites Tools Help
Google G Go Bookmarks PageRank 182 blocked Check Settings
del.icio.us search for "textmining"
Page Tools
Internet
100%

del.icio.us / search

popular | recent

login | register | help

Search results for **textmining**

textmining | del.icio.us | search

Related tags: textmining datamining search nlp text software research java bioinformatics programming

showing 1 - 10 of 1378

« previous | next »

Text Analytics Solutions from ClearForest save this
to textmining datamining text search semantic ... saved by 104 people

» Text mining the New York Times | Emerging Technology Trends | ZDNet.com save this
to textmining datamining research text ai ... saved by 111 people

GATE, A General Architecture for Text Engineering save this
to nlp java opensource information_extraction language ... saved by 348 people

Text mining - Wikipedia, the free encyclopedia save this
to textmining datamining wikipedia mining ai ... saved by 55 people

press release @ the bren school of information and computer sciences save this
to datamining textmining text search mining ... saved by 80 people

Topic Modeling Toolbox save this
to matlab datamining textmining tools topic ... saved by 89 people

text-mining.org save this
to textmining research text_mining text text-mining ... saved by 34 people

text-mining.org save this
to textmining datamining text nlp analysis ... saved by 53 people

Text-Garden -- Text-Mining Software Tools save this
to textmining datamining clustering tools software ... saved by 57 people

KH Coder Index Page save this
to textmining software テキストマイニング datamining text ... saved by 54 people

« previous | next »

Tags entered by users annotating Web sites

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

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

Semantic

Template / frames level

- Templates are the mechanism for extracting the information from text
  - …templates always focused on specific domain which includes consistent patterns on where specific information is positioned
  - Templates are one of the basic methods for information extraction

Examples of templates of KnowItAll system

- Generic approach of extracting is described in
  - *Unsupervised named-entity extraction from the Web: An experimental study (Oren Etzioni et al)*
- KnowItAll system uses the following generic templates:
  - NP “and other” <class1>
  - NP “or other” <class1>
  - <class1> “especially” NPList
  - <class1> “including” NPList
  - <class1> “such as” NPList
  - “such” <class1> “as” NPList
  - NP “is a” <class1>
  - NP “is the” <class1>
- …each template represents specific relationship between the words appearing in the variable slots
- From template patterns KnowItAll bootstraps new templates

Levels of text representations

- Character
- Words
- Phrases
- Part-of-speech tags
- Taxonomies / thesauri

Lexical

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

Syntactic

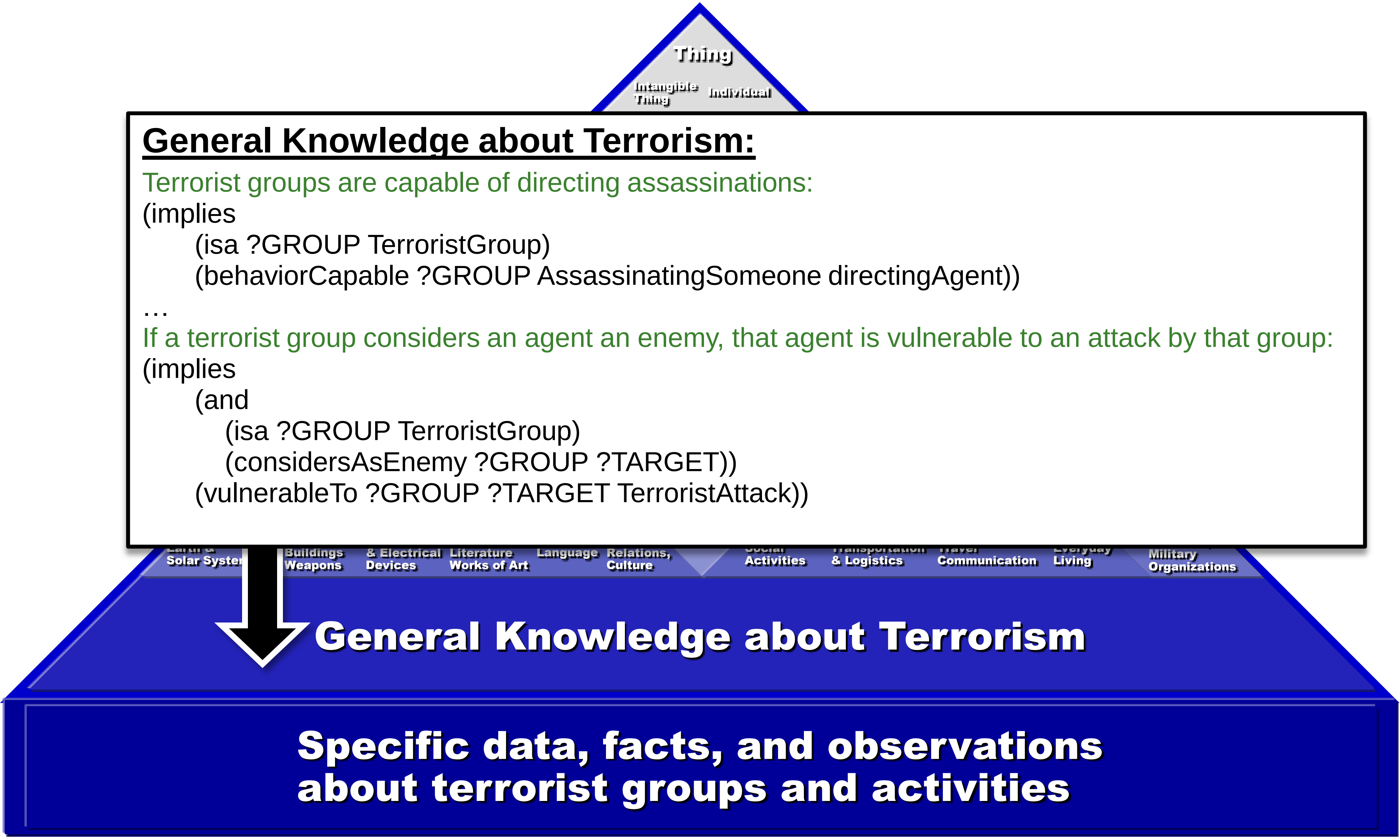
- Collaborative tagging / Web2.0
- Templates / Frames
- **Ontologies / First order theories**

Semantic

Ontologies level

- Ontologies are the most general formalism for describing data objects
  - …in the recent years ontologies got popular through Semantic Web and OWL standard
  - Ontologies can be of various complexity – from relatively simple ones (light weight described with simple) to heavy weight (described with first order theories.
  - Ontologies could be understood also as very generic data-models where we can store extracted information from text

Example: text represented in the First Order Logic



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))
```

General Knowledge about Terrorism

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

Text-Mining

Typical tasks on text

Document Summarization

Document Summarization

- **Task**: the task is to produce shorter, summary version of an original document
- Two main approaches to the problem:
  - **Selection based** – summary is selection of sentences from an original document
  - **Knowledge rich** – performing semantic analysis, representing the meaning and generating the text satisfying length restriction

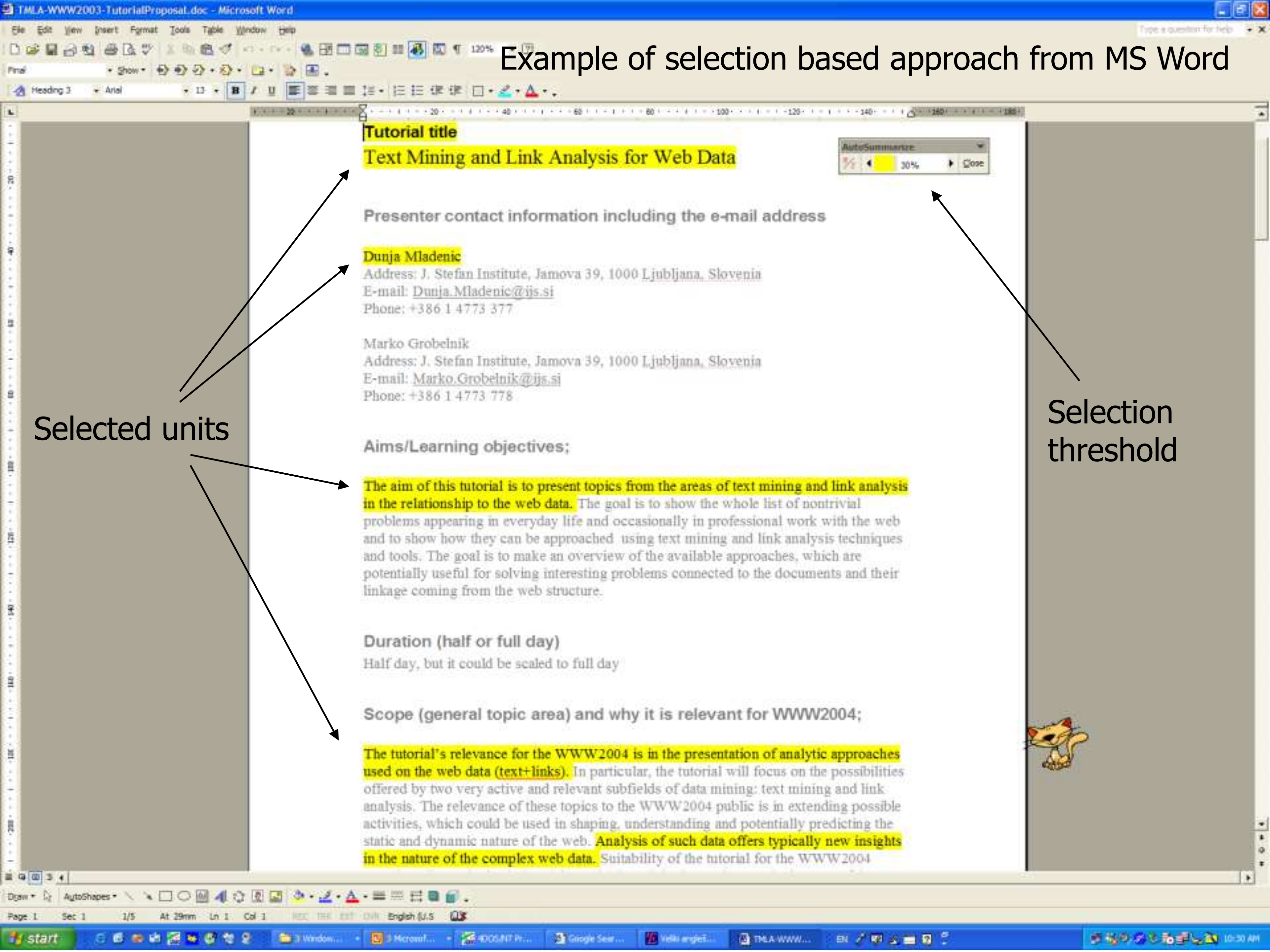
Selection based summarization

- Three main phases:
  - Analyzing the source text
  - Determining its important points (units)
  - Synthesizing an appropriate output
- Most methods adopt linear weighting model – each text unit (sentence) is assessed by the following formula:
  - **Weight(U) = LocationInText(U) + CuePhrase(U) + Statistics(U) + AdditionalPresence(U)**
  - …lot of heuristics and tuning of parameters (also with ML)
- …output consists from topmost text units (sentences)

Selection based summarization

- Three main phases:
  - Analyzing the source text
  - Determining its important points (units)
  - Synthesizing an appropriate output
- Most methods adopt linear weighting model – each text unit (sentence) is assessed by the following formula:
  - **Weight(U) = LocationInText(U) + CuePhrase(U) + Statistics(U) + AdditionalPresence(U)**
  - …lot of heuristics and tuning of parameters (also with Machine learning)
- …output consists from topmost text units (sentences)

Example of selection based approach from MS Word


TMLA-WWW2003-TutorialProposal.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Final
Show
Heading 3
Arial
13
AutoSummarize
30%
Close
Page 1 Sec 1 1/5 At 29mm Ln 1 Col 1 REC TRK EXT OVR English (U.S
start

Tutorial title

Text Mining and Link Analysis for Web Data

Presenter contact information including the e-mail address

Dunja Mladenic
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Dunja.Mladenic@ijs.si
Phone: +386 1 4773 377

Marko Grobelnik
Address: J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Marko.Grobelnik@ijs.si
Phone: +386 1 4773 778

Aims/Learning objectives;

The aim of this tutorial is to present topics from the areas of text mining and link analysis in the relationship to the web data. The goal is to show the whole list of nontrivial problems appearing in everyday life and occasionally in professional work with the web and to show how they can be approached using text mining and link analysis techniques and tools. The goal is to make an overview of the available approaches, which are potentially useful for solving interesting problems connected to the documents and their linkage coming from the web structure.

Duration (half or full day)

Half day, but it could be scaled to full day

Scope (general topic area) and why it is relevant for WWW2004;

The tutorial's relevance for the WWW2004 is in the presentation of analytic approaches used on the web data (text+links). In particular, the tutorial will focus on the possibilities offered by two very active and relevant subfields of data mining: text mining and link analysis. The relevance of these topics to the WWW2004 public is in extending possible activities, which could be used in shaping, understanding and potentially predicting the static and dynamic nature of the web. Analysis of such data offers typically new insights in the nature of the complex web data. Suitability of the tutorial for the WWW2004

Selected units

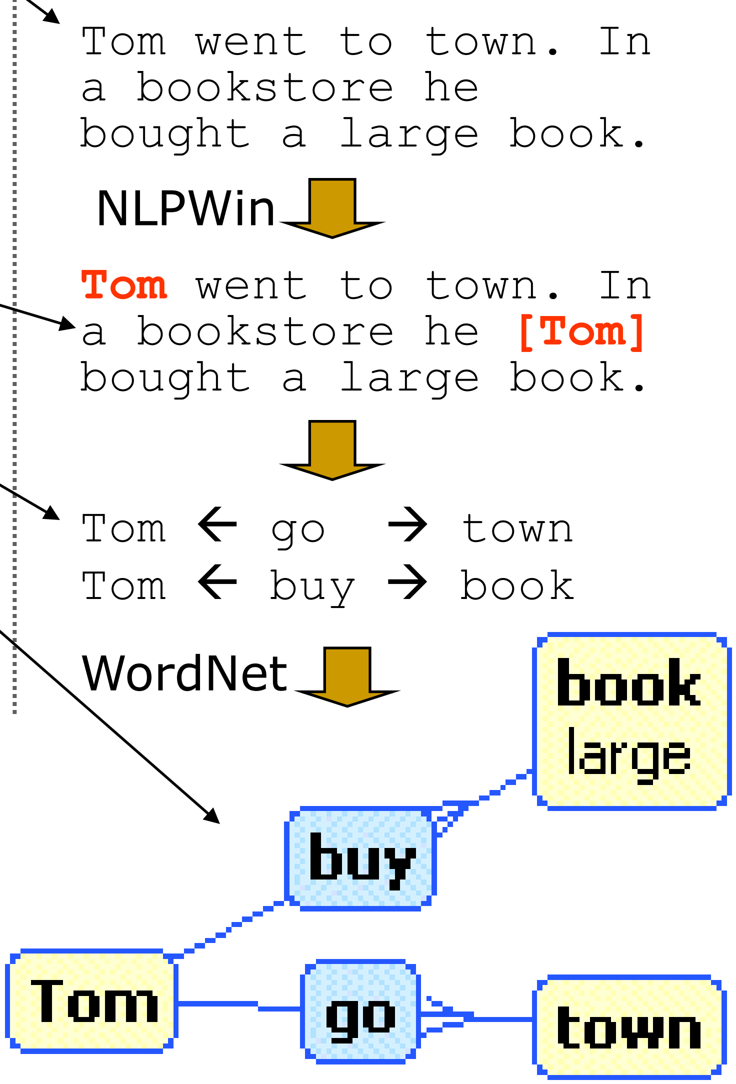
Selection threshold

Knowledge rich summarization

- To generate ‘true’ summary of a document we need to (at least partially) ‘understand’ the document text
  - …the document is to small to count on statistics, we need to identify and use its linguistic and semantic structure
- On the next slides we show an approach from (Leskovec, Grobelnik, Milic-Frayling 2004) using 10 step procedure for extracting semantics from a document:
  - …the approach was evaluated on “Document Understanding Conference” test set of documents and their summaries
  - …the approach extracts semantic network from a document and tries to extract relevant part of the semantic network to represent summary
  - Results achieved 70% recall of and 25% precision on extracted Subject-Predicate-Object triples

Knowledge Rich Summarization Example

1. Input document is split into sentences
2. Each sentence is deep-parsed
3. Name-entities are disambiguated:
   - Determining that 'George Bush' == 'Bush' == 'U.S. president'
4. Performing Anaphora resolution:
   - Pronouns are connected with named-entities
5. Extracting of **Subject-Predicate-Object** triples
6. Constructing a **graph** from triples
7. Each triple in the graph is described with features for learning
8. Using machine learning train a model for classification of triples into the summary
9. Generate a summary graph from selected triples
10. From the summary graph generate textual summary document



Tom went to town. In a bookstore he bought a large book.

NLPWin

Tom went to town. In a bookstore he [Tom] bought a large book.

Tom ← go → town

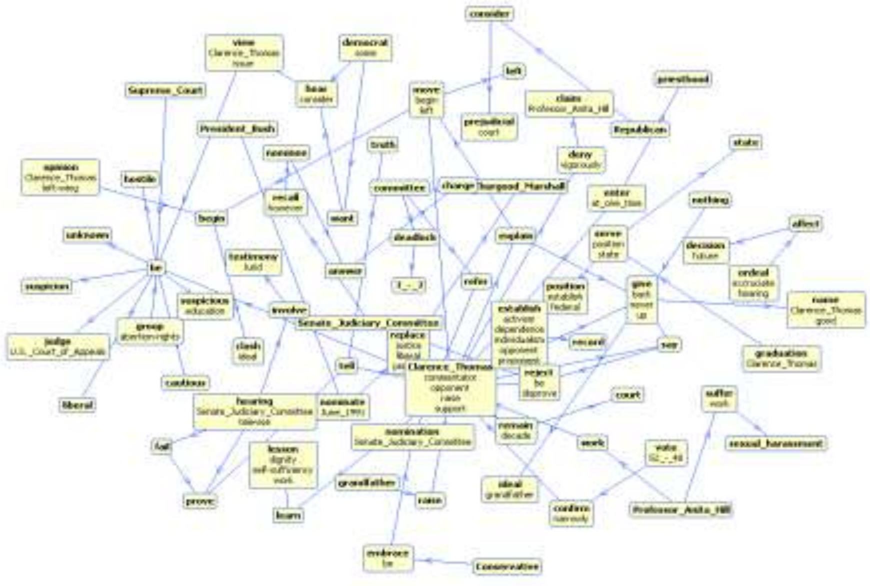
Tom ← buy → book

WordNet

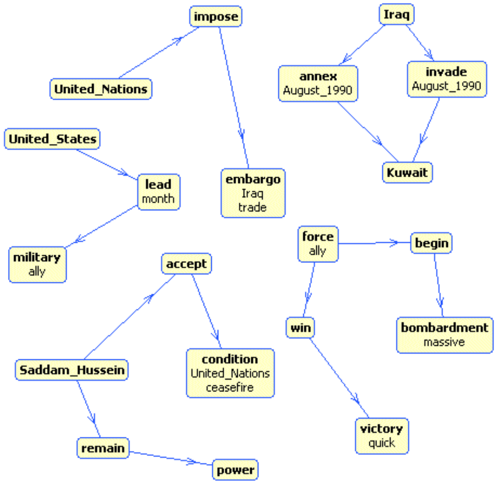
Training of summarization model

- A model was trained deciding which **Subject-Predicate-Object** triple belongs into the target summary
- For training was used Support Vector Machine (SVM) on 400 statistic, linguistic and graph topological features

Document Semantic network



Summary semantic network



Example of summarization

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

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 countries also are trying [illegible] 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 [illegible] reported that [illegible] 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 [illegible] 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. [illegible] 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 [illegible] 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 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.''.

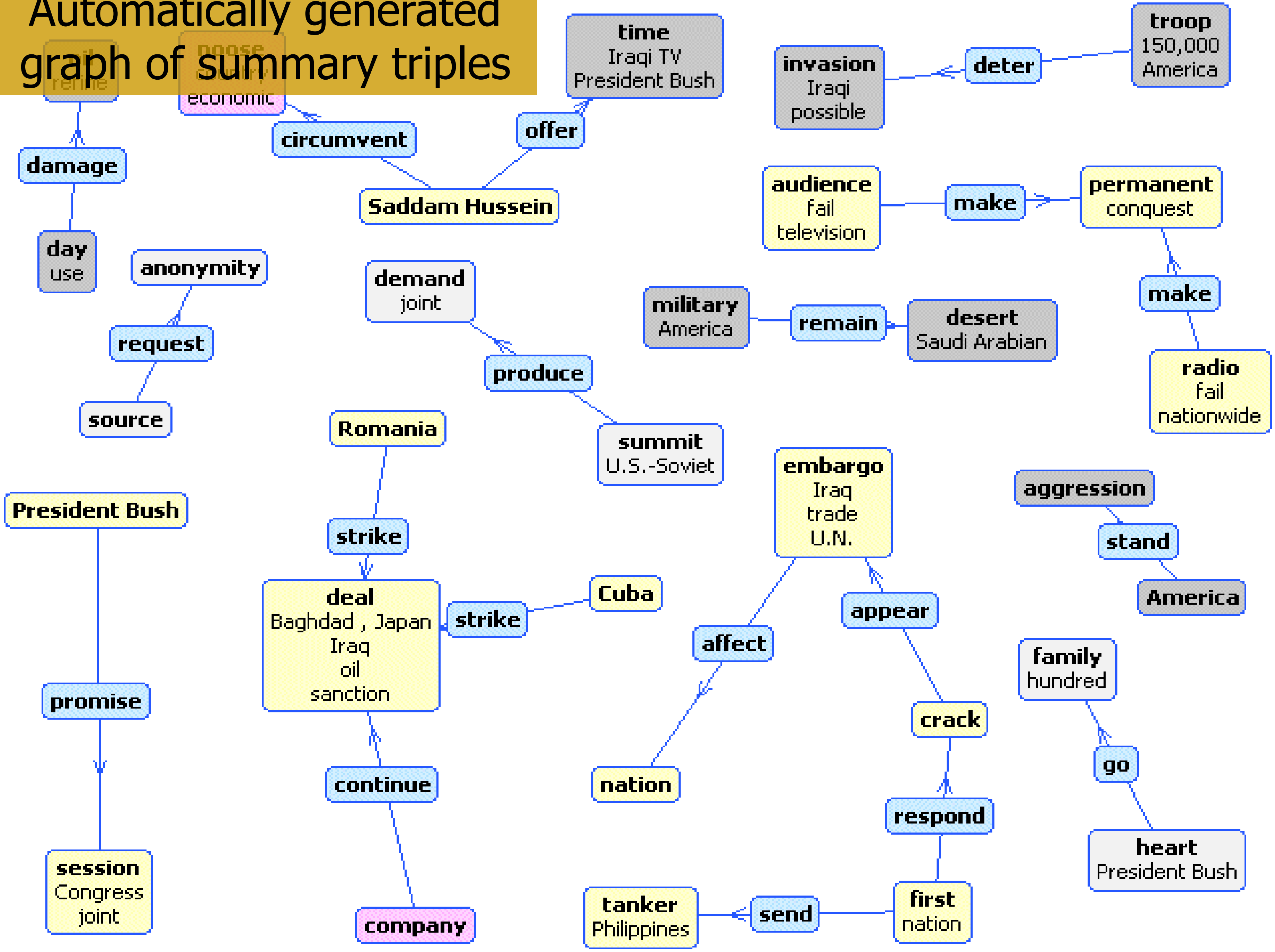
Human written summary

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.**

7800 chars, 1300 words

Automatically generated graph of summary triples



Text Segmentation

Text Segmentation

- **Problem**: divide text that has no given structure into segments with similar content
- Example applications:
  - topic tracking in news (spoken news)
  - identification of topics in large, unstructured text databases

Hearst Algorithm for Text Segmentation

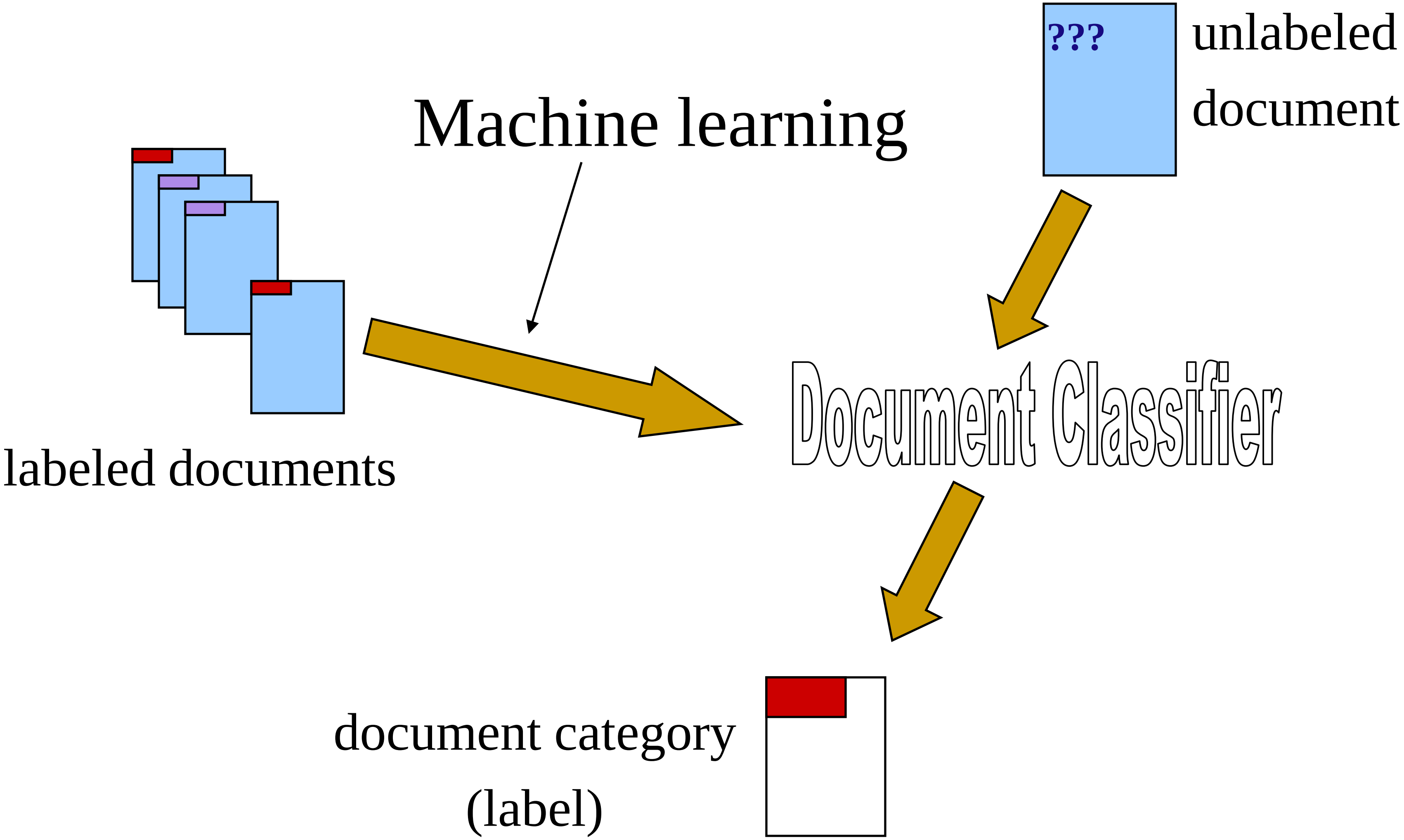
- Algorithm
  - Initial segmentation
    - Divide a text into equal blocks of k words
  - Similarity Computation
    - compute similarity between *m* blocks on the right and the left of the candidate boundary
  - Boundary Detection
    - place a boundary where similarity score reaches local minimum
- …the approach can be defined either as optimization problem or as sliding window

Supervised Learning

Document Categorization Task

- **Given:** set of documents labeled with content categories
- **The goal:** to build a model which would automatically assign right content categories to new unlabeled documents.
- Content categories can be:
  - unstructured (e.g., Reuters) **or**
  - structured (e.g., Yahoo, DMoz, Medline)

Document categorization



Algorithms for learning document classifiers

- Popular algorithms for text categorization:
  - Support Vector Machines
  - Logistic Regression
  - Perceptron algorithm
  - Naive Bayesian classifier
  - Winnow algorithm
  - Nearest Neighbour
  - ....

Example learning algorithm: Perceptron

Input:

- set of documents ***D*** in the form of (e.g. TFIDF) numeric vectors
- each document has label +1 (positive class) or -1 (negative class)

Output:

- linear model w_i (one weight per word from the vocabulary)

Algorithm:

- **Initialize** the model w_i by setting word weights to 0
- **Iterate** through documents N times
  - **For** document ***d*** from ***D***
    - // Using current model w_i classify the document ***d***
    - **if** sum(d_i *w_i) >= 0 **then** classify document as positive
    - **else** classify document as negative
    - **if** document classification is wrong **then**
      - *// adjust weights of all words occurring in the document*
      - $w_{t+1} = w_t$ +sign(true-class) * Beta (input parameter Beta>0)
      - *// where sign(positive) = 1 and sign(negative) =-1*

Measuring success – Model quality estimation

$$Precision(M,targetC) = P(targetC \,|\, \overline{targetC})$$

The truth, and

$$Recall(M,targetC) = P(\overline{targetC} \,|\, targetC)$$

..the whole truth

$$Accuracy(M) = \sum_i P(\overline{C_i}) \times Precision(M,C_i)$$

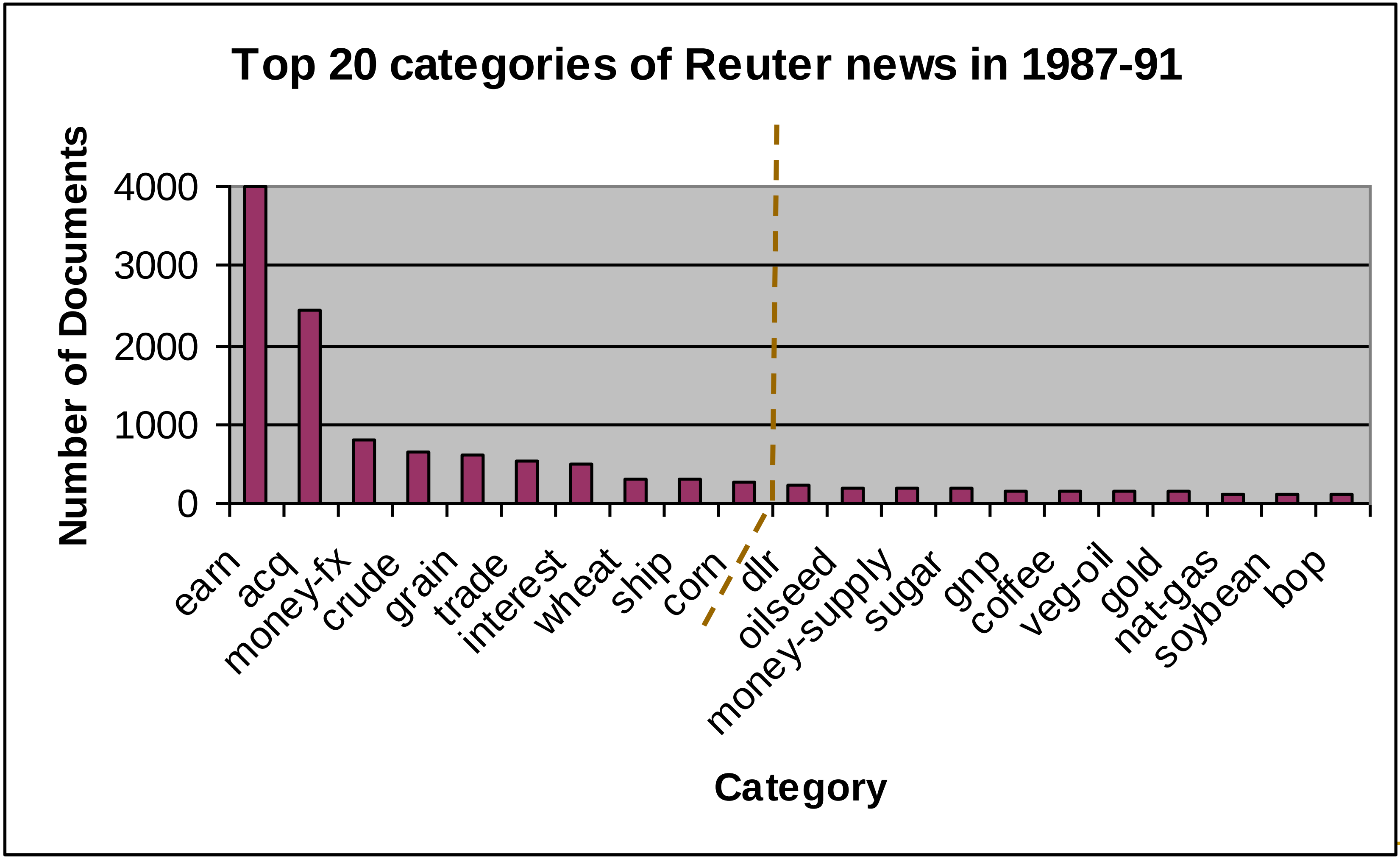
$$F_\beta(M,targetC) = \frac{(1+\beta^2)Precision(M,targetC) \times Recall(M,targetC)}{\beta^2 Precision(M,targetC) + Recall(M,targetC)}$$

- Classification accuracy
- Break-even point (precision=recall)
- F-measure (precision, recall)

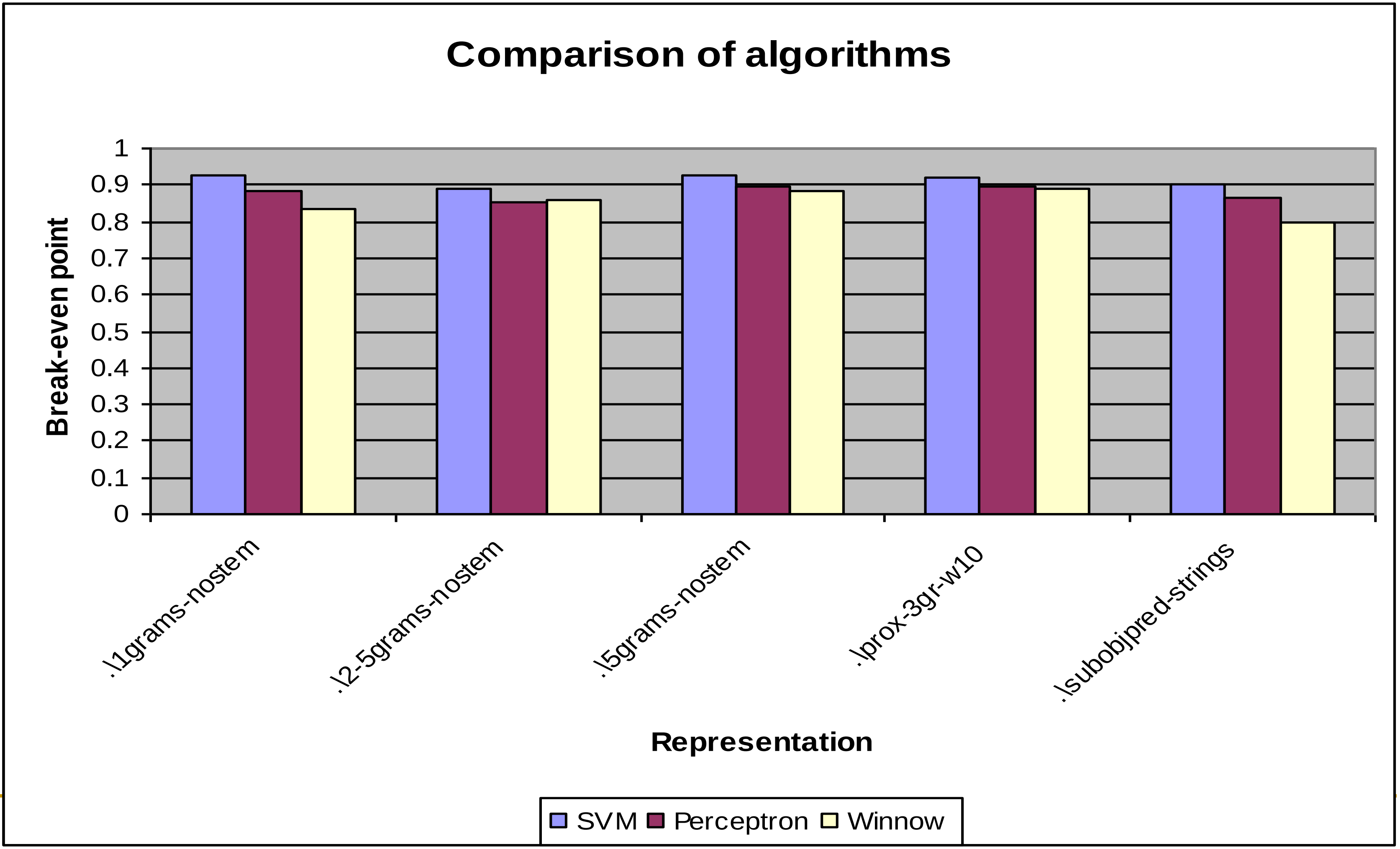
Reuters dataset – Categorization to flat categories

- Documents classified by editors into one or more categories
- Publicly available dataset of Reuters news mainly from 1987:
  - 120 categories giving the document content, such as: *earn, acquire, corn, rice, jobs, oilseeds, gold, coffee, housing, income,...*
- …from 2000 is available new dataset of 830,000 Reuters documents available fo research

Distribution of documents (Reuters-21578)



SVM, Perceptron & Winnow text categorization performance on Reuters-21578 with different representations

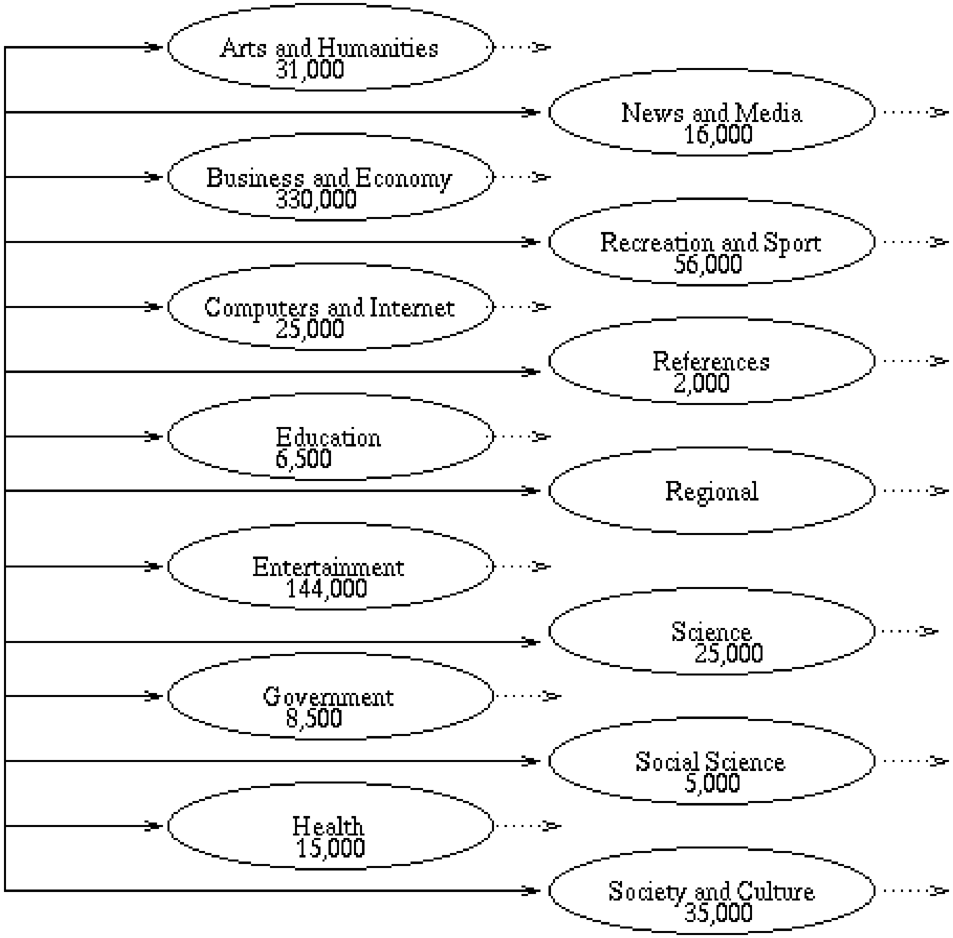


Text Categorization into hierarchy of categories

- There are several hierarchies (taxonomies) of textual documents:
  - Yahoo, DMoz, Medline, …
- Different people use different approaches:
  - …series of hierarchically organized classifiers
  - …set of independent classifiers just for leaves
  - …set of independent classifiers for all nodes

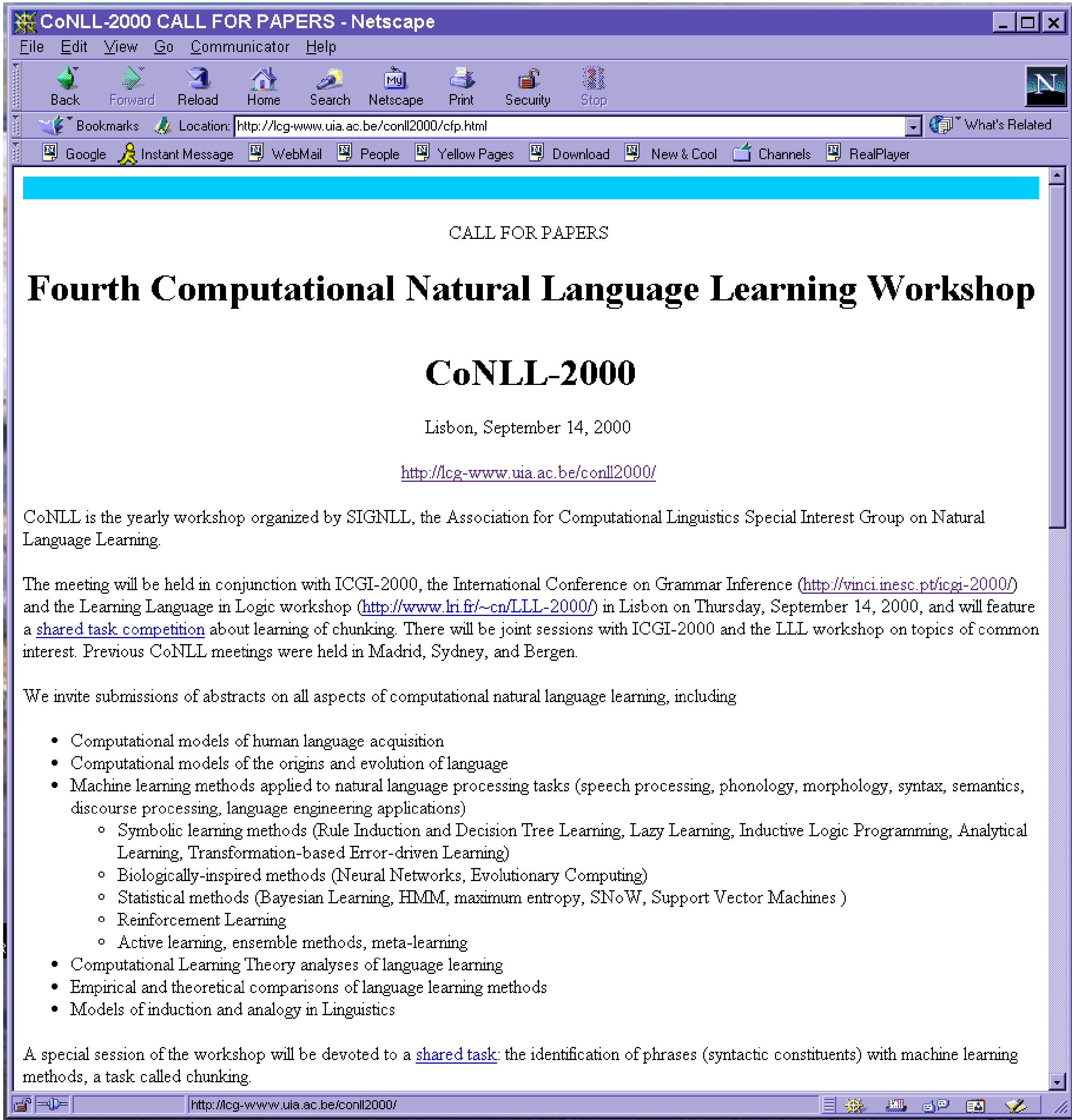
Yahoo! hierarchy (taxonomy)

- human constructed hierarchy of Web-documents
- exists in several languages (we use English)
- easy to access and regularly updated
- captures most of the Web topics
- English version includes over 2M pages categorized into 50,000 categories
- contains about 250Mb of HTML files



Document to categorize:

CFP for CoNLL-2000


CoNLL-2000 CALL FOR PAPERS - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://lcg-www.uia.ac.be/conll2000/cfp.html
What's Related
Google Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
http://lcg-www.uia.ac.be/conll2000/

CALL FOR PAPERS

Fourth Computational Natural Language Learning Workshop

CoNLL-2000

Lisbon, September 14, 2000

http://lcg-www.uia.ac.be/conll2000/

CoNLL is the yearly workshop organized by SIGNLL, the Association for Computational Linguistics Special Interest Group on Natural Language Learning.

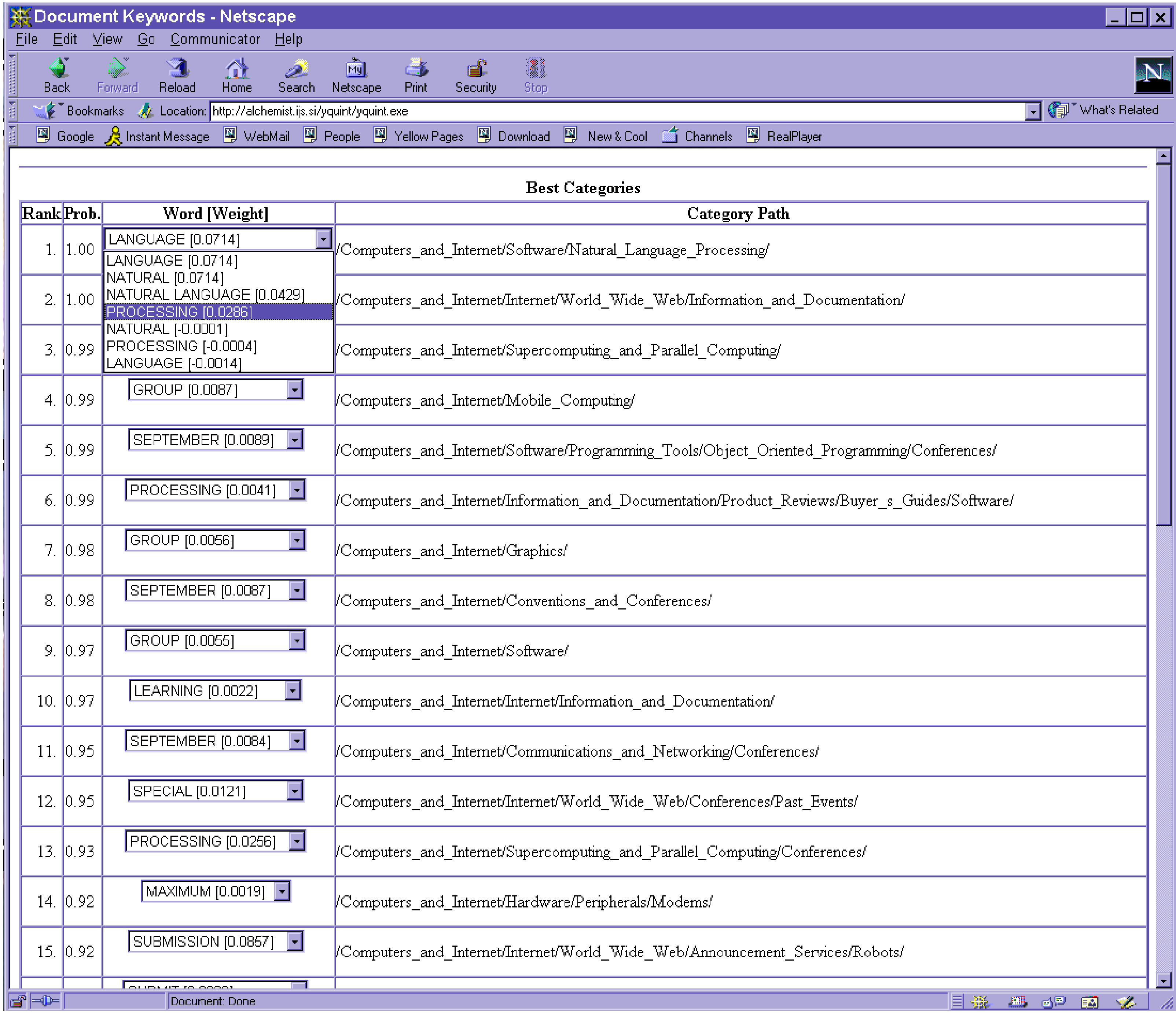
The meeting will be held in conjunction with ICGI-2000, the International Conference on Grammar Inference (http://vinci.inesc.pt/icgi-2000/) and the Learning Language in Logic workshop (http://www.lri.fr/~cn/LLL-2000/) in Lisbon on Thursday, September 14, 2000, and will feature a shared task competition about learning of chunking. There will be joint sessions with ICGI-2000 and the LLL workshop on topics of common interest. Previous CoNLL meetings were held in Madrid, Sydney, and Bergen.

We invite submissions of abstracts on all aspects of computational natural language learning, including

- Computational models of human language acquisition
- Computational models of the origins and evolution of language
- Machine learning methods applied to natural language processing tasks (speech processing, phonology, morphology, syntax, semantics, discourse processing, language engineering applications)
  - Symbolic learning methods (Rule Induction and Decision Tree Learning, Lazy Learning, Inductive Logic Programming, Analytical Learning, Transformation-based Error-driven Learning)
  - Biologically-inspired methods (Neural Networks, Evolutionary Computing)
  - Statistical methods (Bayesian Learning, HMM, maximum entropy, SNoW, Support Vector Machines)
  - Reinforcement Learning
  - Active learning, ensemble methods, meta-learning
- Computational Learning Theory analyses of language learning
- Empirical and theoretical comparisons of language learning methods
- Models of induction and analogy in Linguistics

A special session of the workshop will be devoted to a shared task: the identification of phrases (syntactic constituents) with machine learning methods, a task called chunking.

Some predicted categories

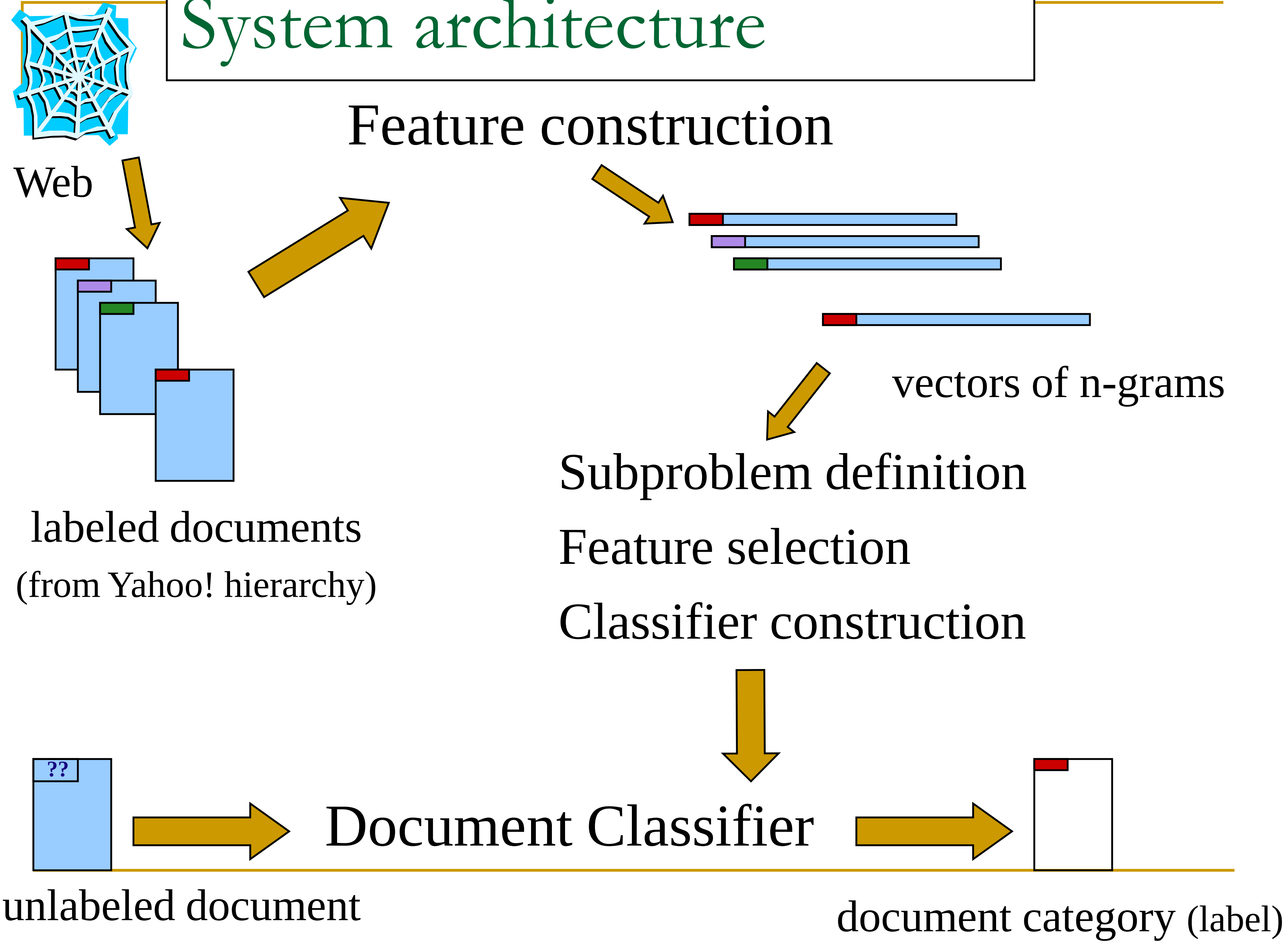

Document Keywords - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks
What's Related
Google Instant Message WebMail People Yellow Pages Download New & Cool Channels RealPlayer
Document: Done

Location: http://alchemist.ijs.si/yquint/yquint.exe

Best Categories

Rank	Prob.	Word [Weight]	Category Path
1.	1.00	LANGUAGE [0.0714] LANGUAGE [0.0714] NATURAL [0.0714] NATURAL LANGUAGE [0.0429] PROCESSING [0.0286] NATURAL [-0.0001] PROCESSING [-0.0004] LANGUAGE [-0.0014]	/Computers_and_Internet/Software/Natural_Language_Processing/
2.	1.00		/Computers_and_Internet/Internet/World_Wide_Web/Information_and_Documentation/
3.	0.99		/Computers_and_Internet/Supercomputing_and_Parallel_Computing/
4.	0.99	GROUP [0.0087]	/Computers_and_Internet/Mobile_Computing/
5.	0.99	SEPTEMBER [0.0089]	/Computers_and_Internet/Software/Programming_Tools/Object_Oriented_Programming/Conferences/
6.	0.99	PROCESSING [0.0041]	/Computers_and_Internet/Information_and_Documentation/Product_Reviews/Buyer_s_Guides/Software/
7.	0.98	GROUP [0.0056]	/Computers_and_Internet/Graphics/
8.	0.98	SEPTEMBER [0.0087]	/Computers_and_Internet/Conventions_and_Conferences/
9.	0.97	GROUP [0.0055]	/Computers_and_Internet/Software/
10.	0.97	LEARNING [0.0022]	/Computers_and_Internet/Internet/Information_and_Documentation/
11.	0.95	SEPTEMBER [0.0084]	/Computers_and_Internet/Communications_and_Networking/Conferences/
12.	0.95	SPECIAL [0.0121]	/Computers_and_Internet/Internet/World_Wide_Web/Conferences/Past_Events/
13.	0.93	PROCESSING [0.0256]	/Computers_and_Internet/Supercomputing_and_Parallel_Computing/Conferences/
14.	0.92	MAXIMUM [0.0019]	/Computers_and_Internet/Hardware/Peripherals/Modems/
15.	0.92	SUBMISSION [0.0857]	/Computers_and_Internet/Internet/World_Wide_Web/Announcement_Services/Robots/

System architecture



Web

labeled documents

(from Yahoo! hierarchy)

Feature construction

vectors of n-grams

Subproblem definition

Feature selection

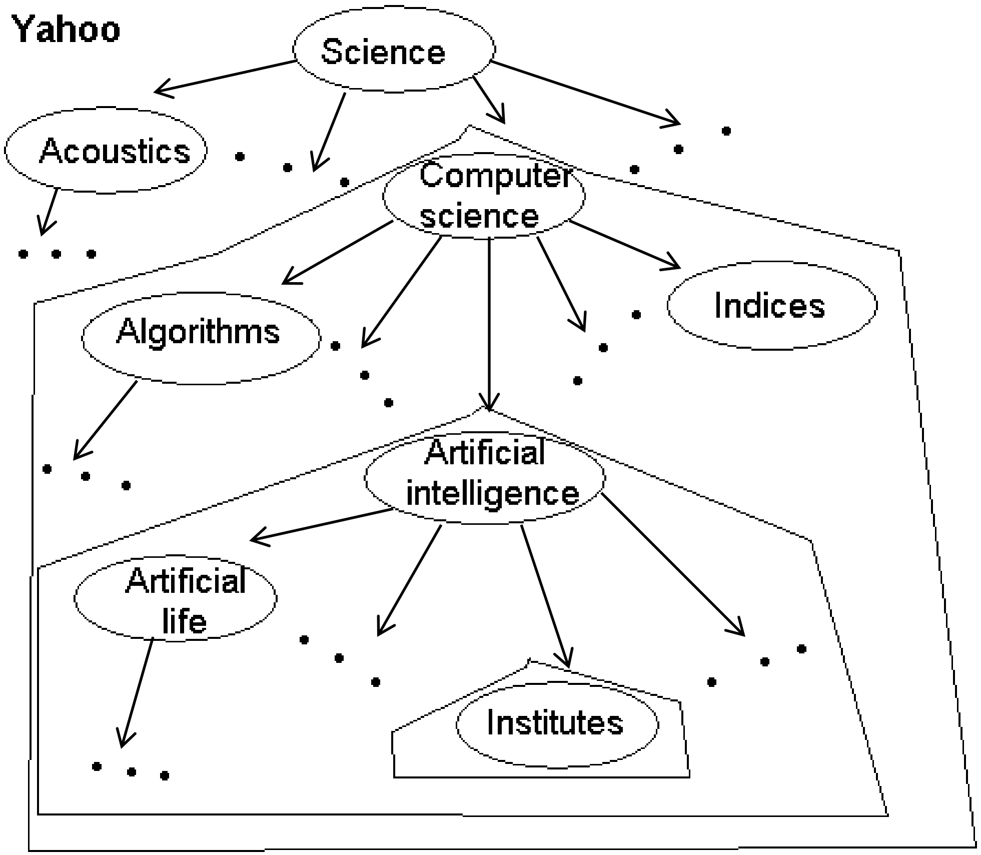
Classifier construction

unlabeled document

Document Classifier

document category (label)

Content categories



- For each content category generate a separate classifier that predicts probability for a new document to belong to its category

Considering promising categories only (classification by Naive Bayes)

$$P(C|D) = \frac{P(C)\prod_{W \in Doc} P(W|C)}{\sum_{i} P(C_i)\prod_{W \in Doc} P(W|C_i)}$$

- Document is represented as a set of word sequences W
- Each classifier has two distributions: P(W|pos), P(W|neg)
- Promising category:
  - calculated P(pos|Doc) is high meaning that the classifier has P(W|pos)>0 for at least some W from the document (otherwise, the prior probability is returned, P(neg) is about 0.90)

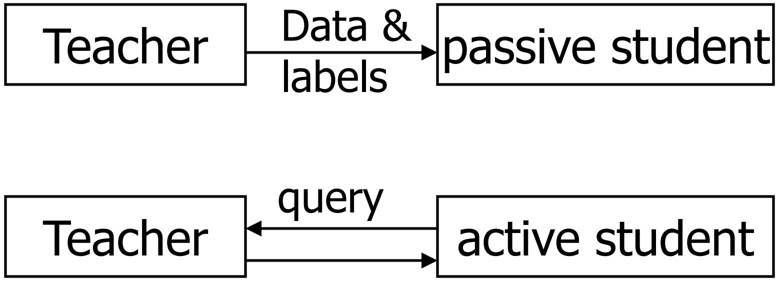
Summary of experimental results

Domain	probability	rank	precision	recall
Entertain.	0.96	16	0.44	0.80
Arts	0.99	10	0.40	0.83
Computers	0.98	12	0.40	0.84
Education	0.99	9	0.57	0.65
Reference	0.99	3	0.51	0.81

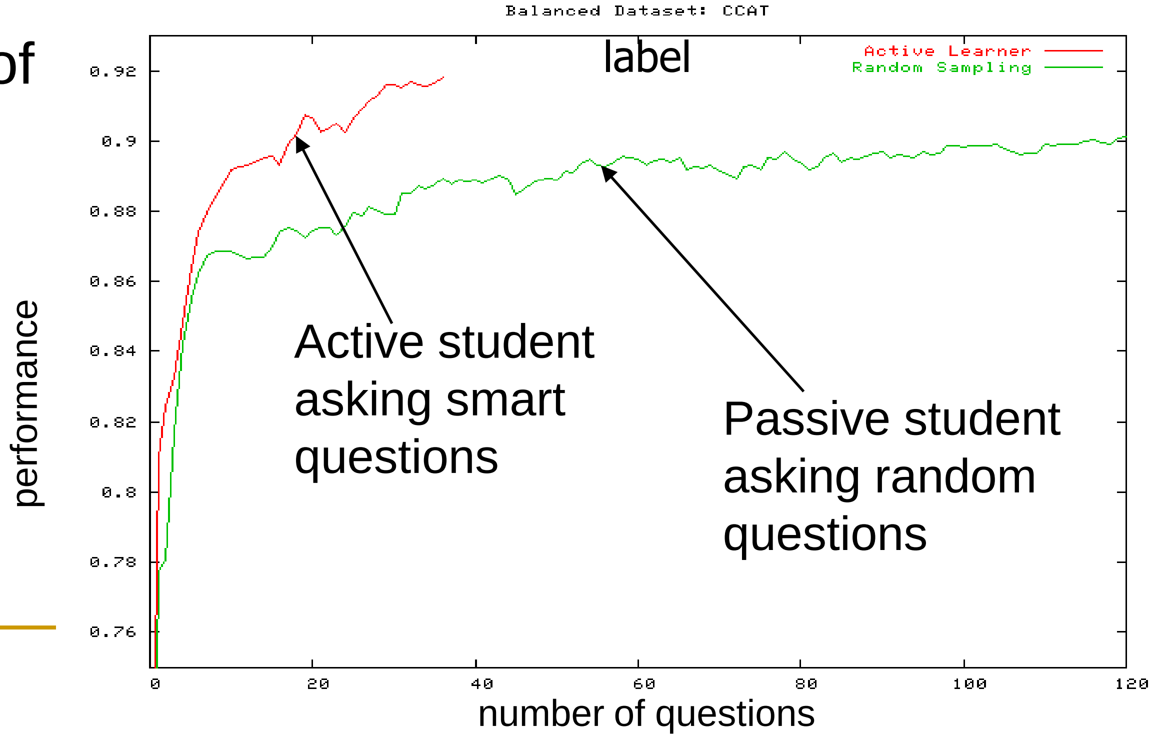
Active Learning

Active Learning

- We use this methods whenever hand-labeled data are rare or expensive to obtain
- Interactive method



- Requests only labeling of “interesting” objects
- Much less human work needed for the same result compared to arbitrary labeling examples



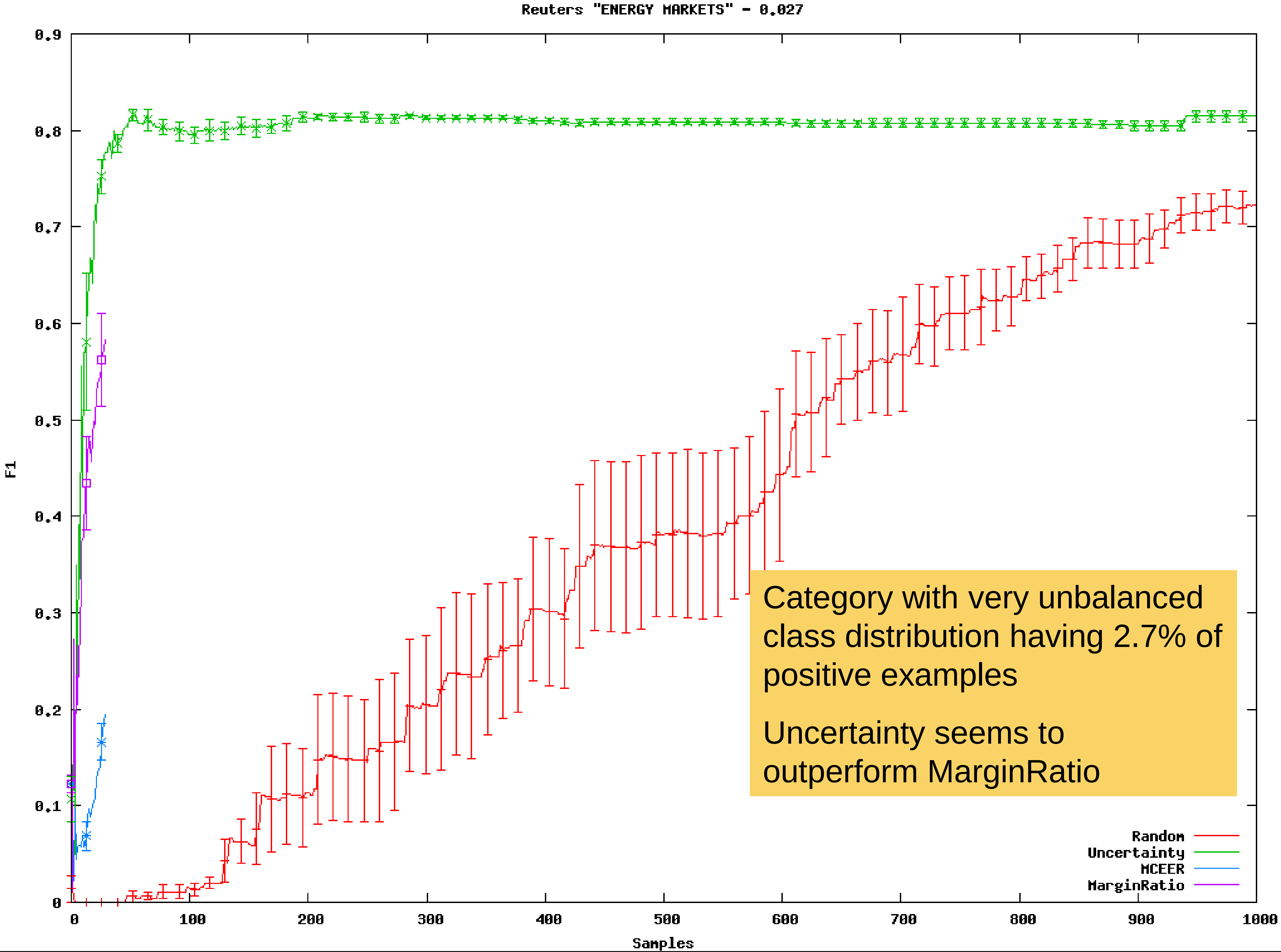
Active student asking smart questions

Passive student asking random questions

Some approaches to Active Learning

- **Uncertainty sampling** (efficient)
  - select example closest to the decision hyperplane (or the one with classification probability closest to P=0.5) (Tong & Koller 2000 Stanford)
- **Maximum margin ratio change**
  - select example with the largest predicted impact on the margin size if selected (Tong & Koller 2000 Stanford)
- **Monte Carlo Estimation of Error Reduction**
  - select example that reinforces our current beliefs (Roy & McCallum 2001, CMU)
- **Random sampling** as baseline

- Experimental evaluation (using F1-measure) of the four listed approaches shown on three categories from Reuters-2000 dataset
  - average over 10 random samples of 5000 training (out of 500k) and 10k testing (out of 300k) examples
  - the last two methods are rather time consuming, thus we run them for including the first 50 unlabeled examples
  - experiments show that active learning is especially useful for unbalanced data



Category with very unbalanced class distribution having 2.7% of positive examples

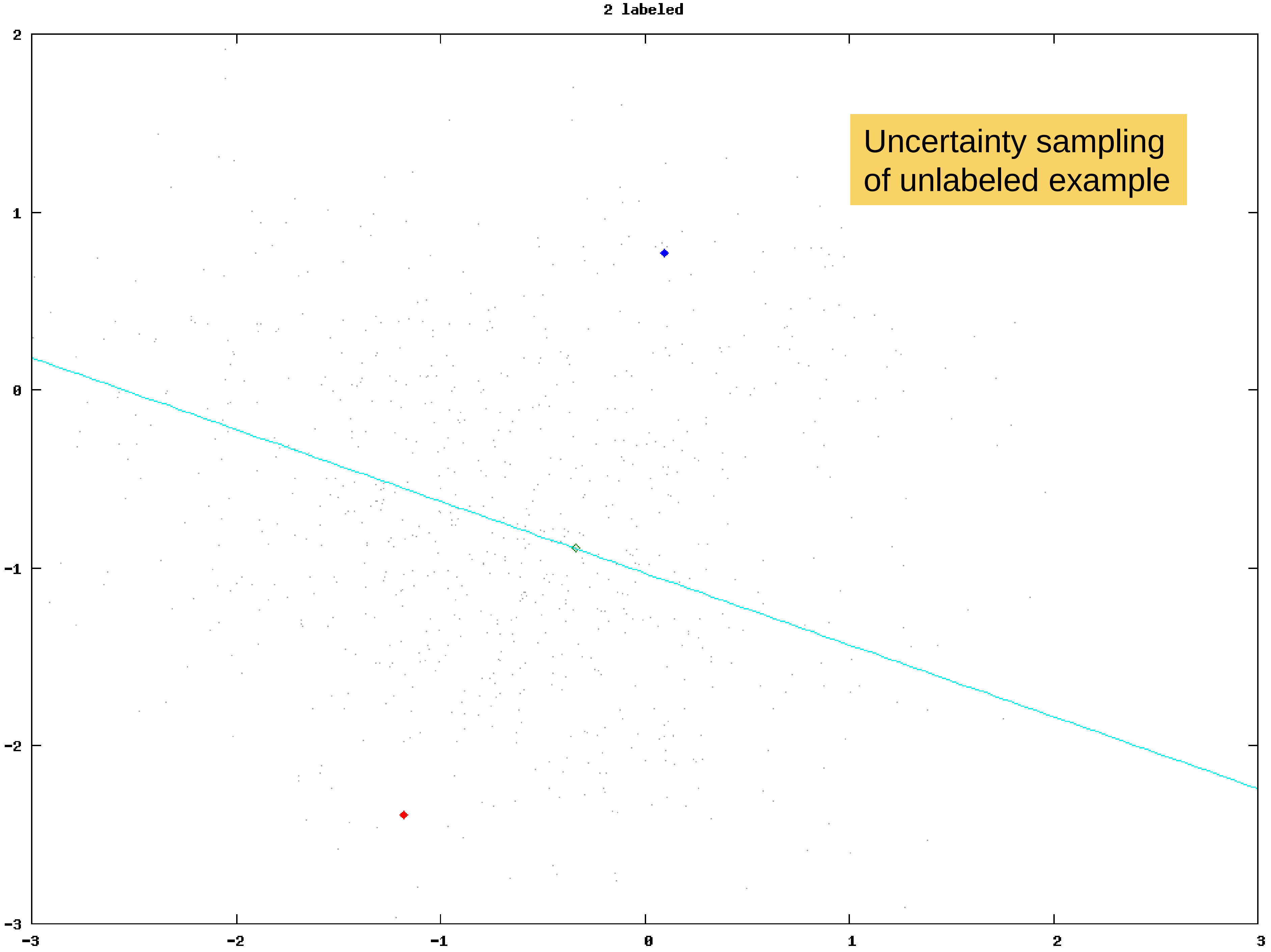
Uncertainty seems to outperform MarginRatio

Illustration of Active learning

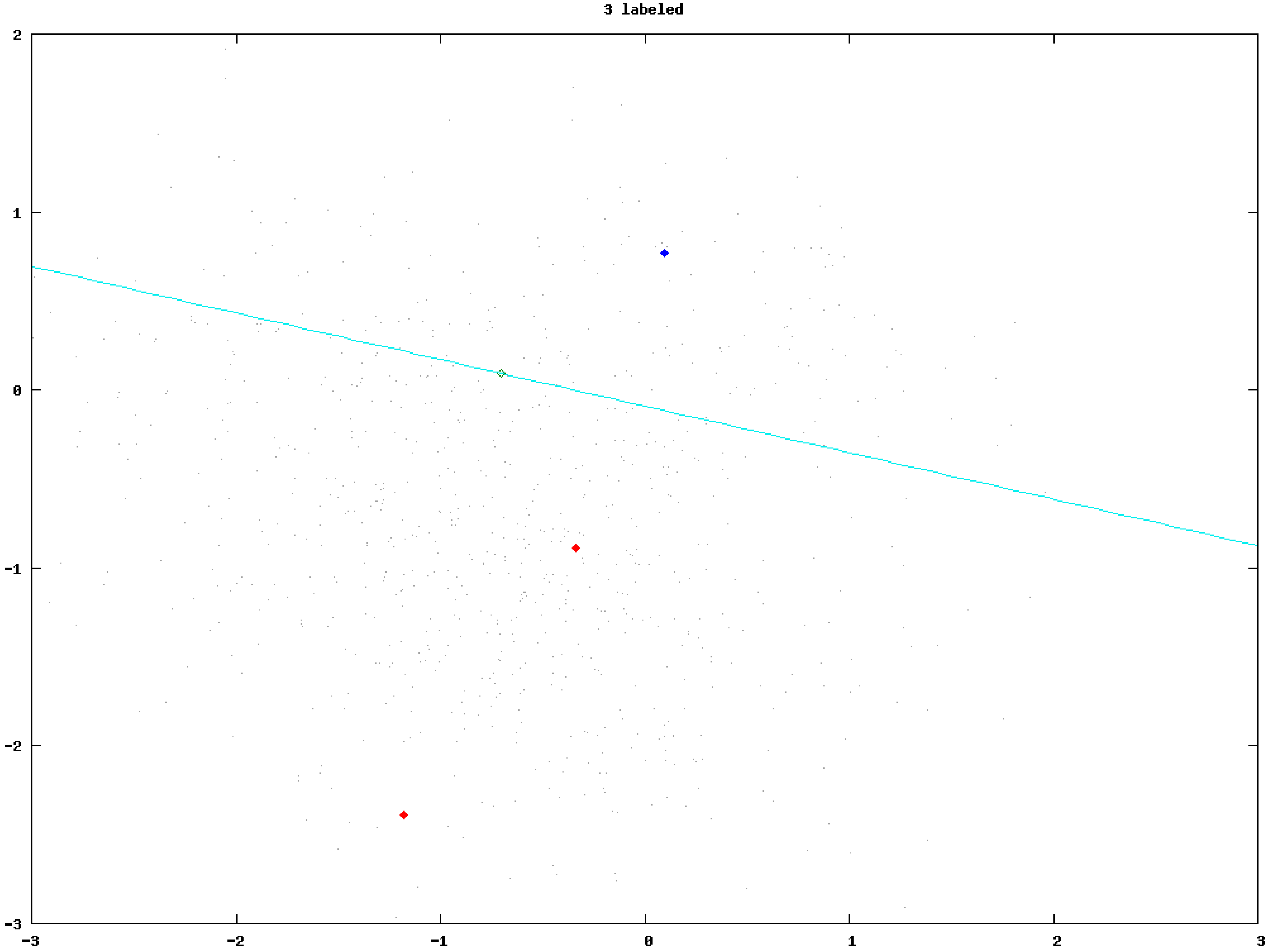
- starting with one labeled example from each class (red and blue)
- select one example for labeling (green circle)
- request label and add re-generate the model using the extended labeled data

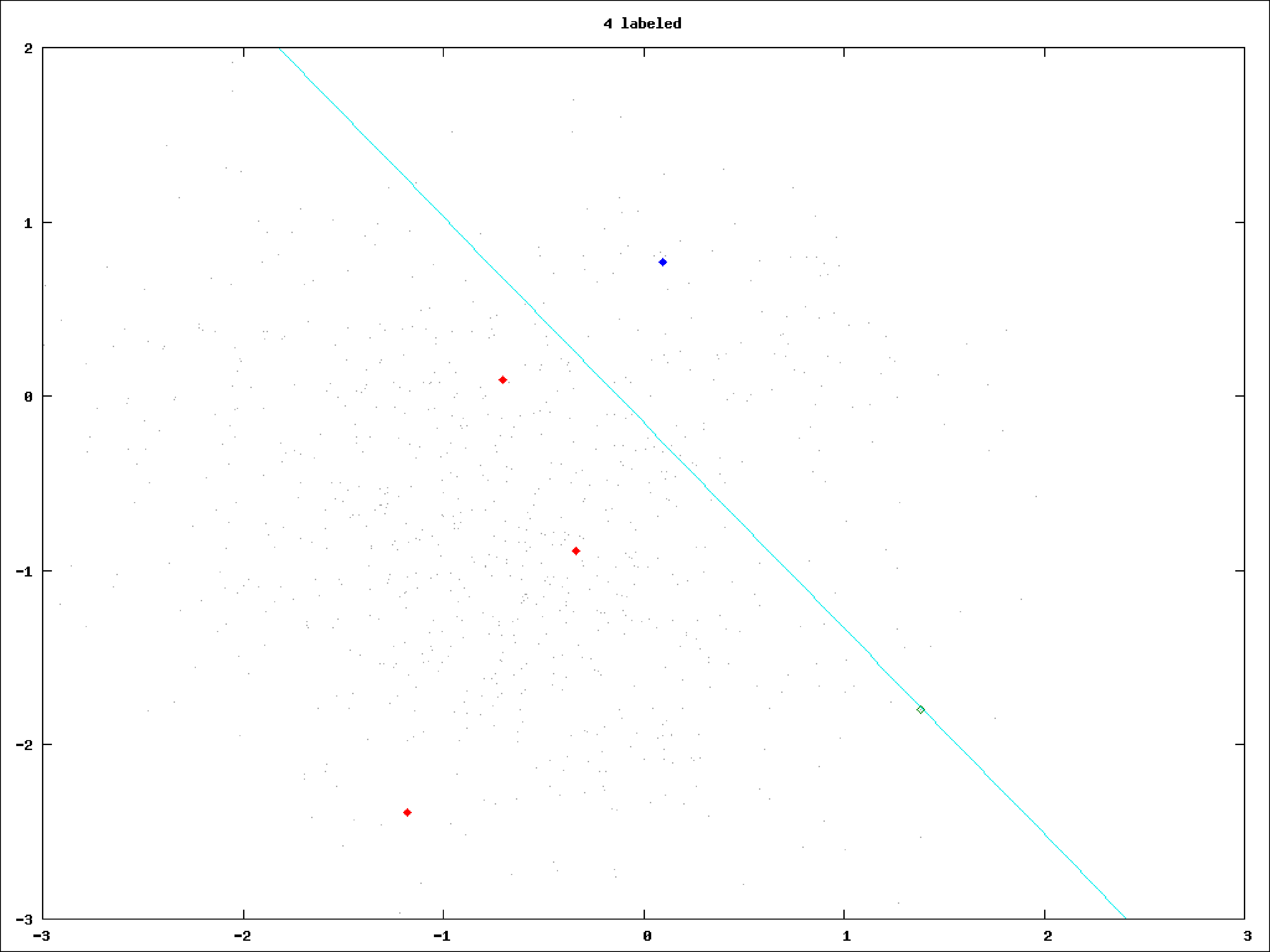
Illustration of linear SVM model using

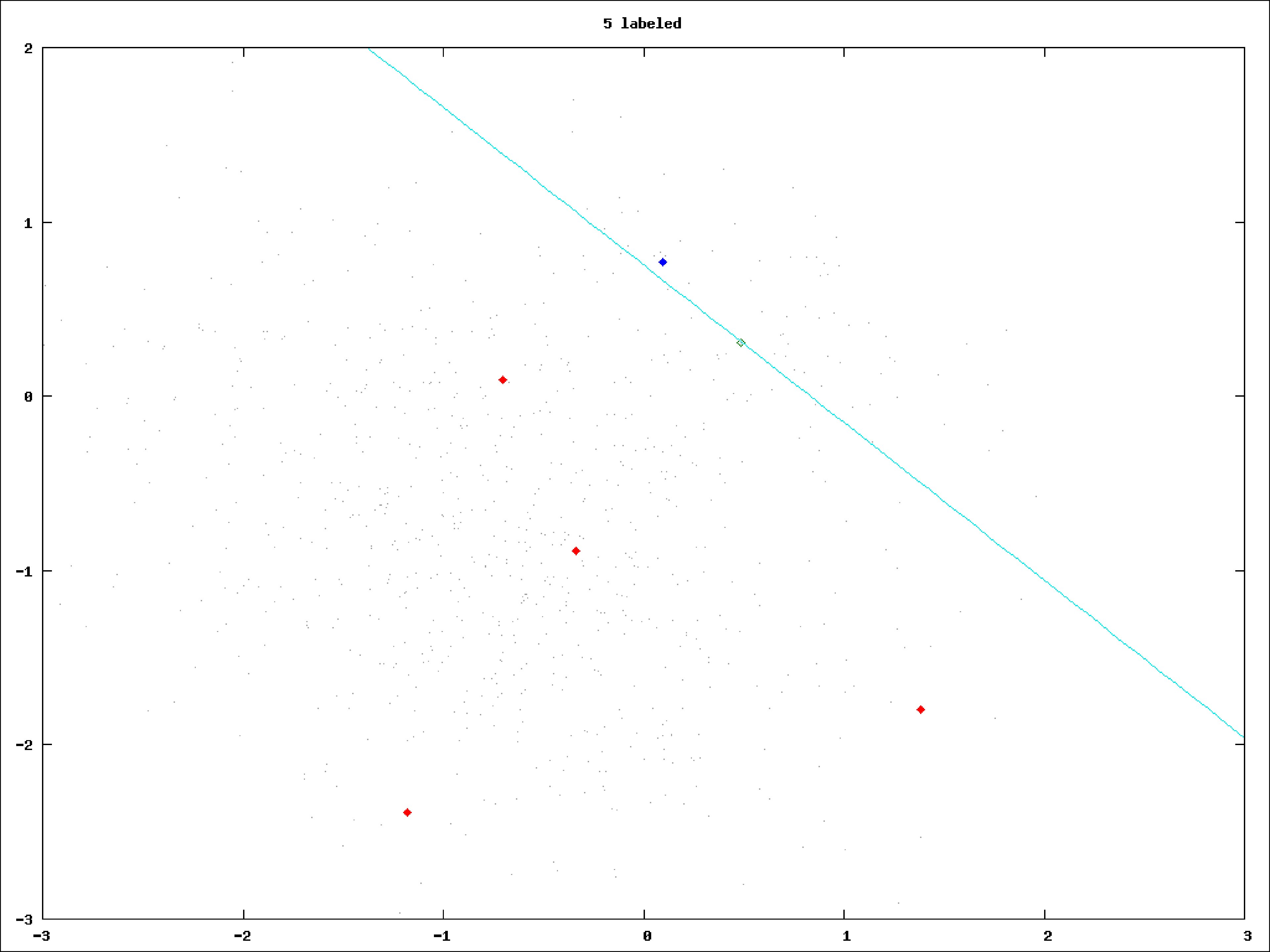
- arbitrary selection of unlabeled examples (random)
- active learning selecting the most uncertain examples (closest to the decision hyperplane)

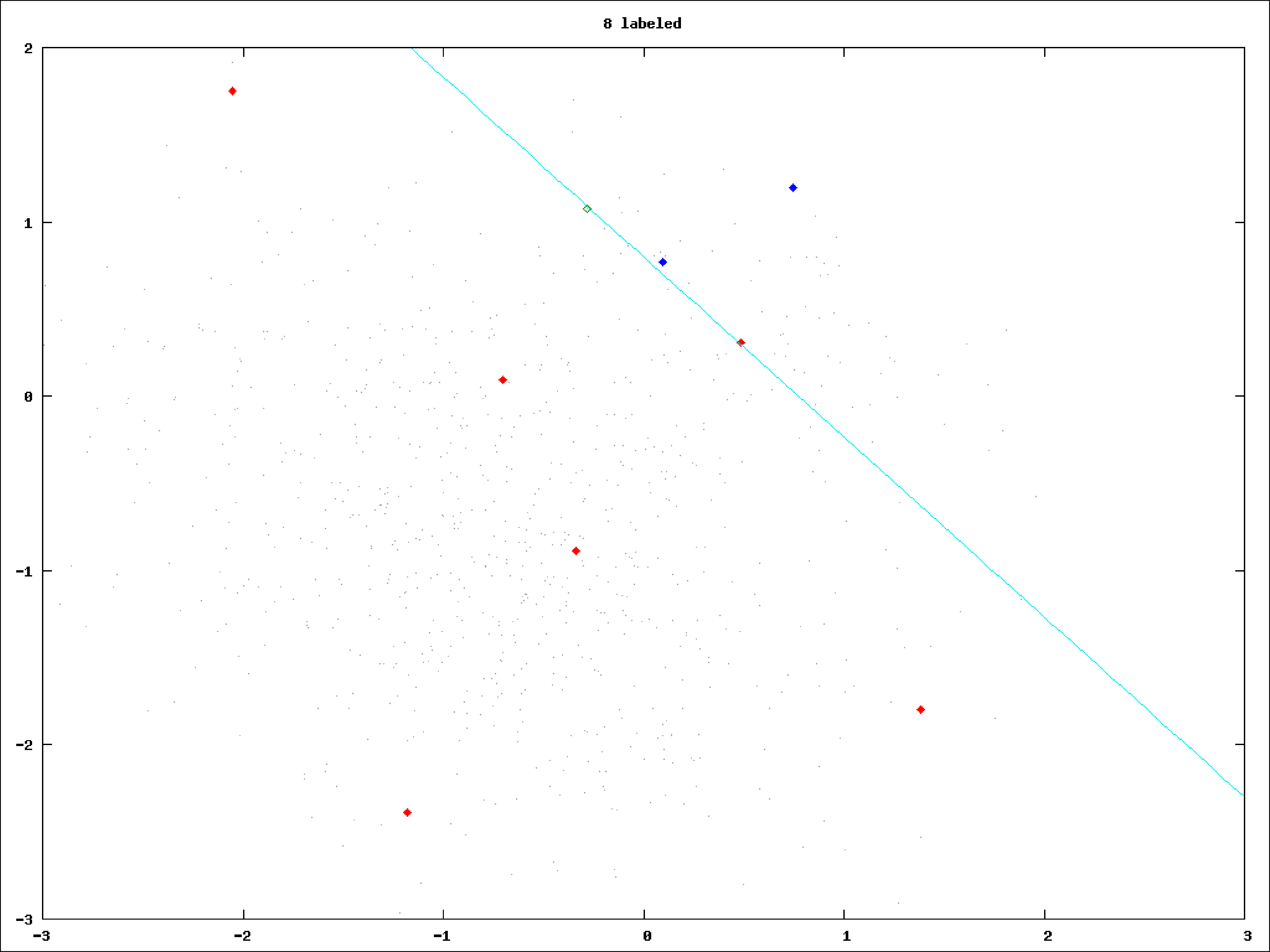


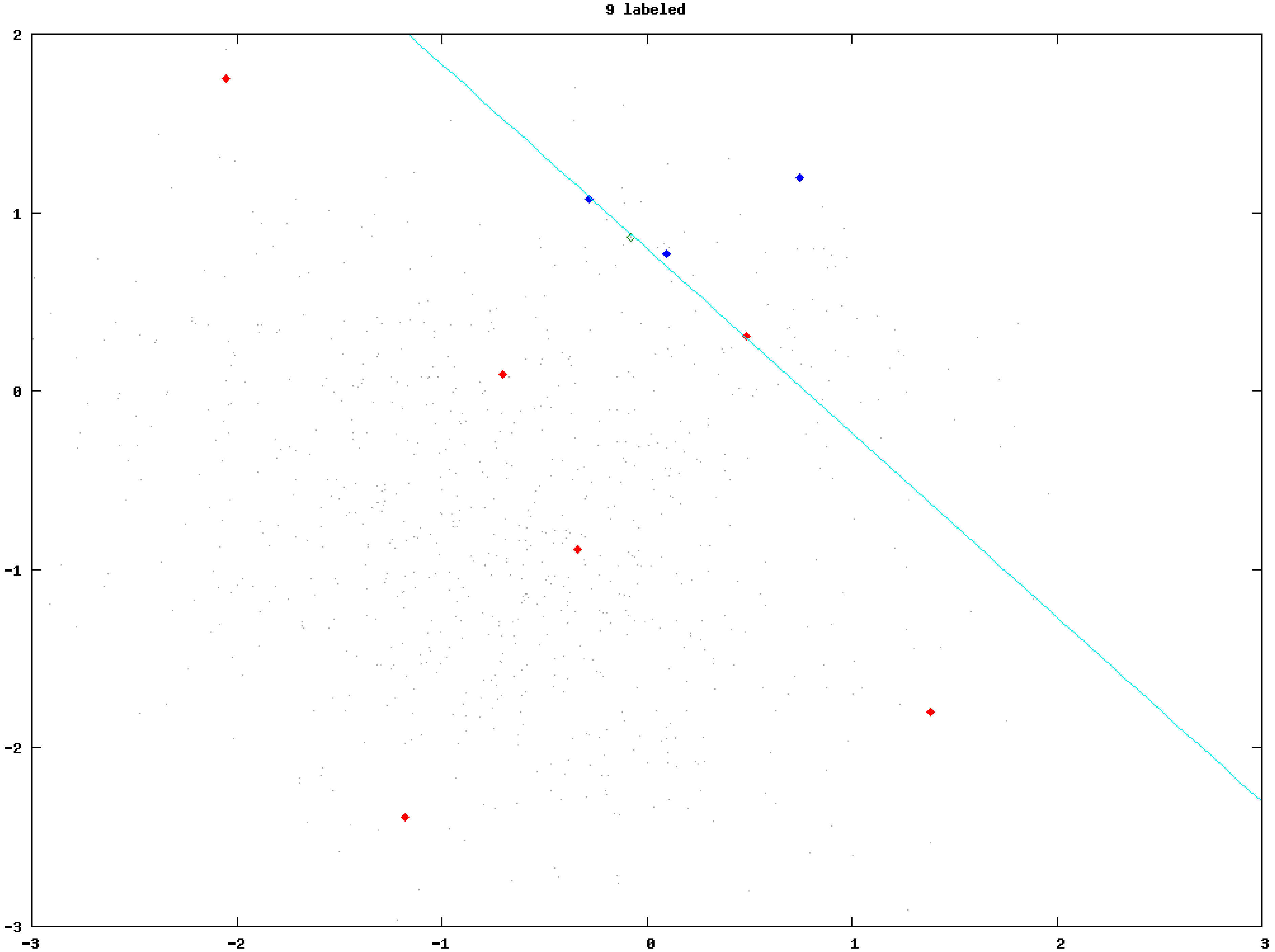
Uncertainty sampling of unlabeled example

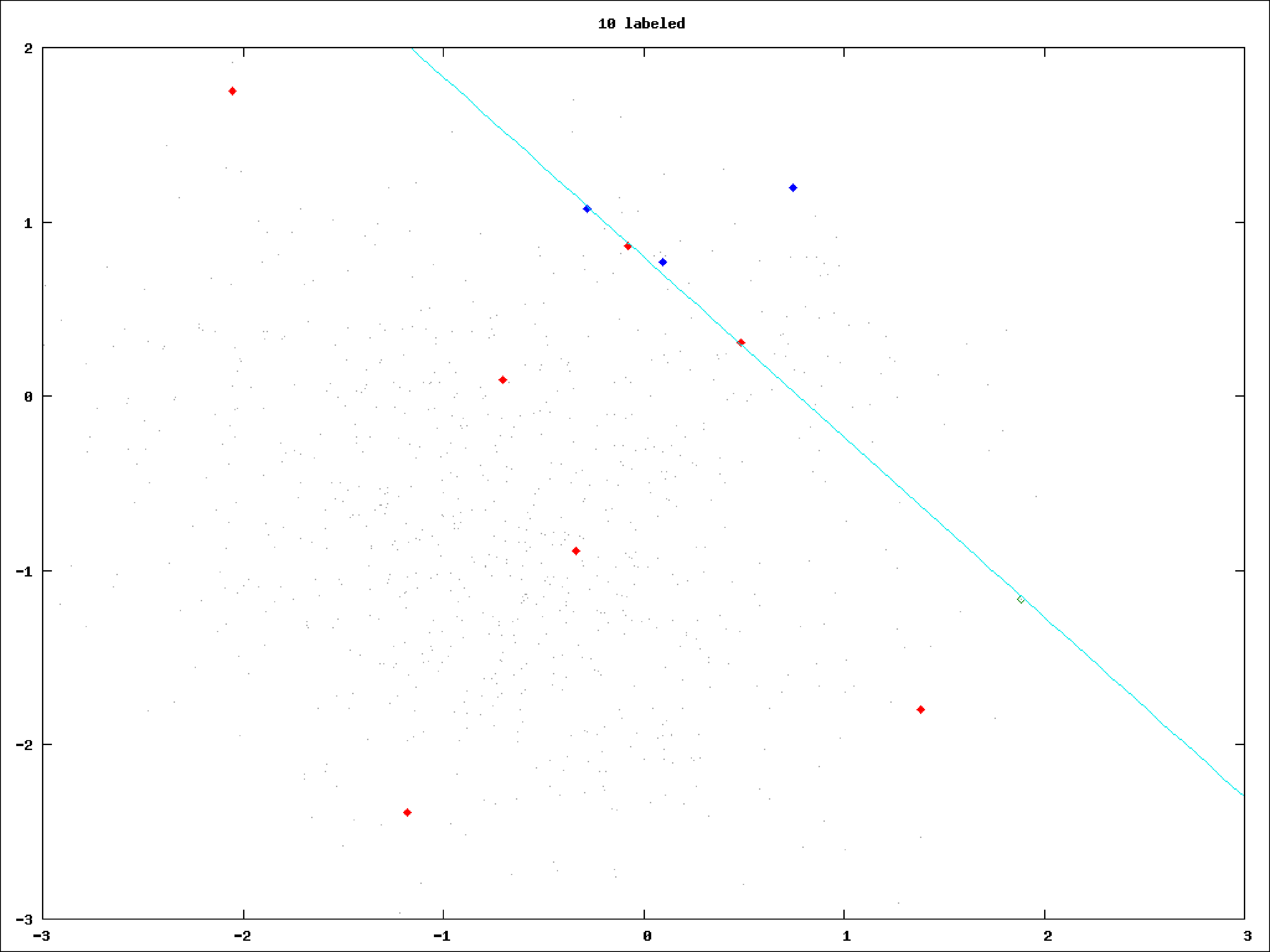

3 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

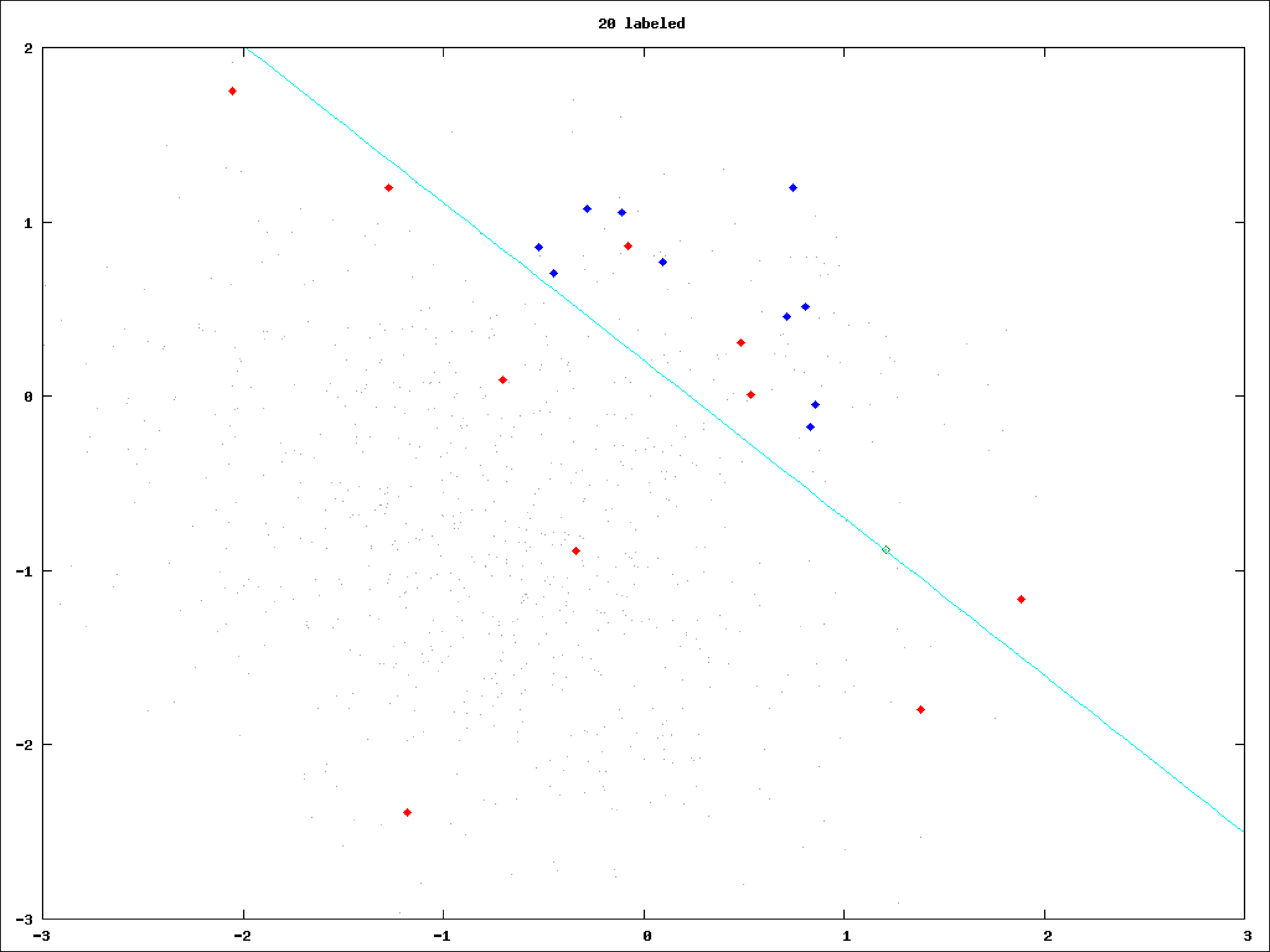

4 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

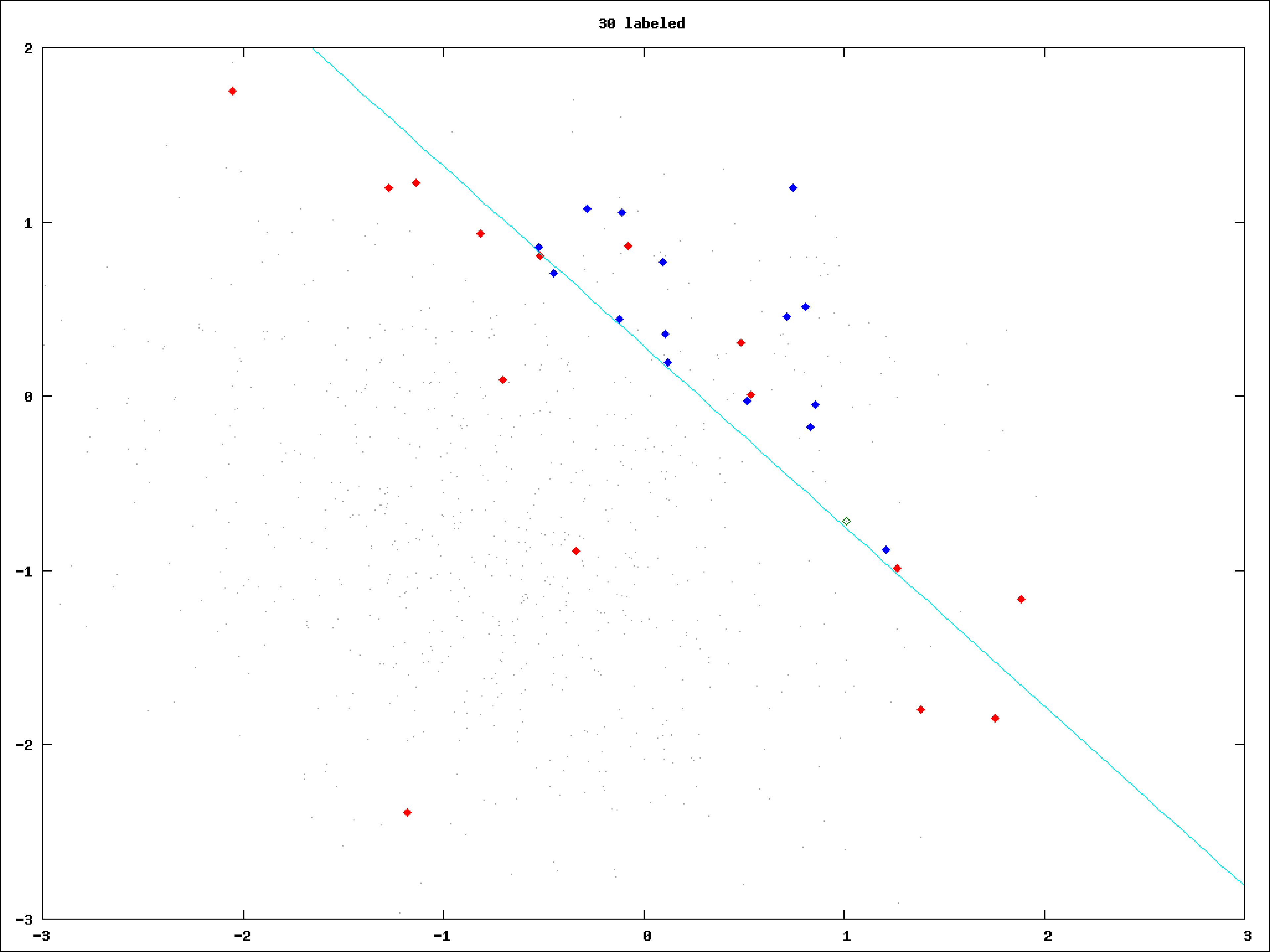

5 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

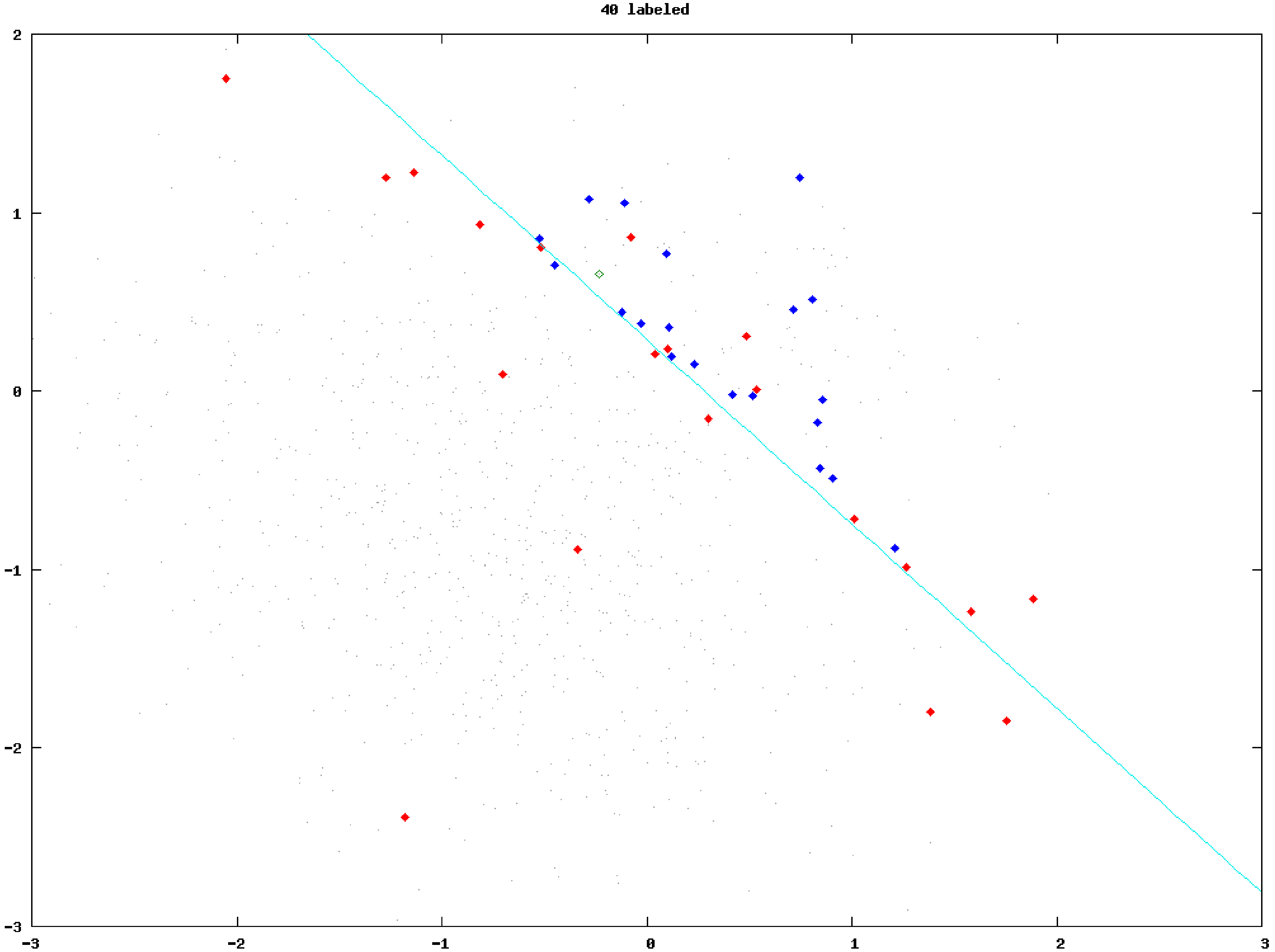

8 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

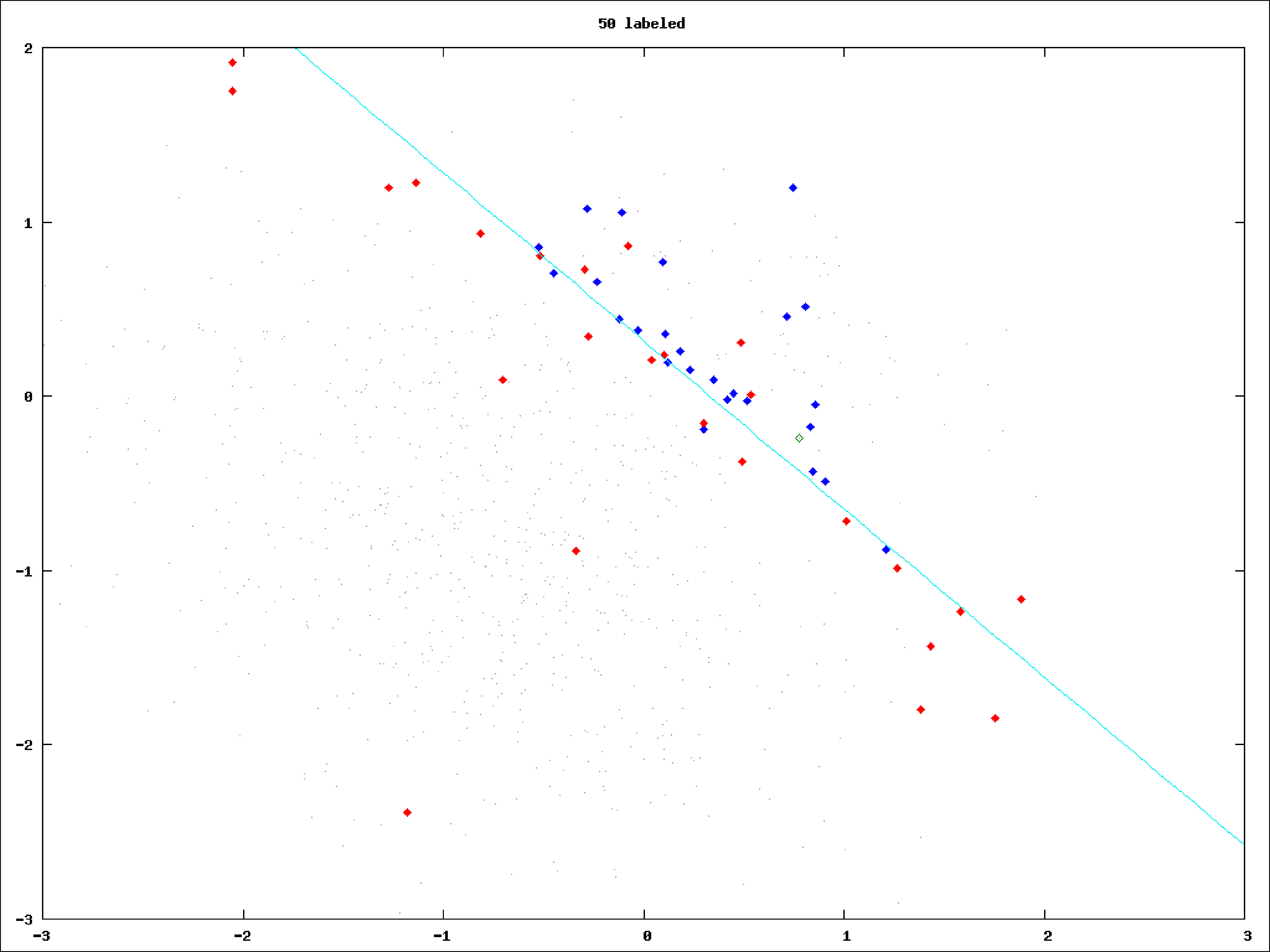

9 labeled
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

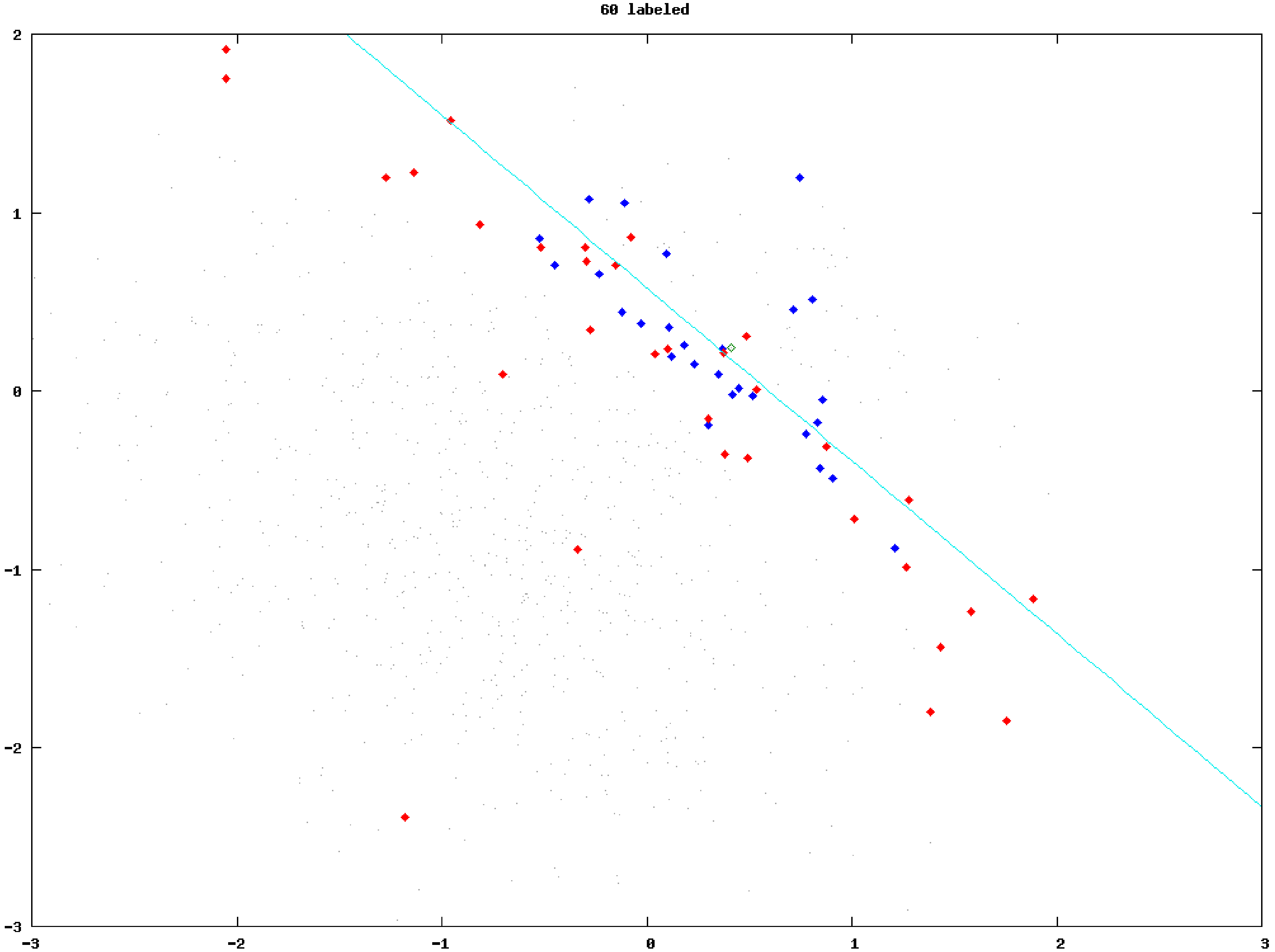

10 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

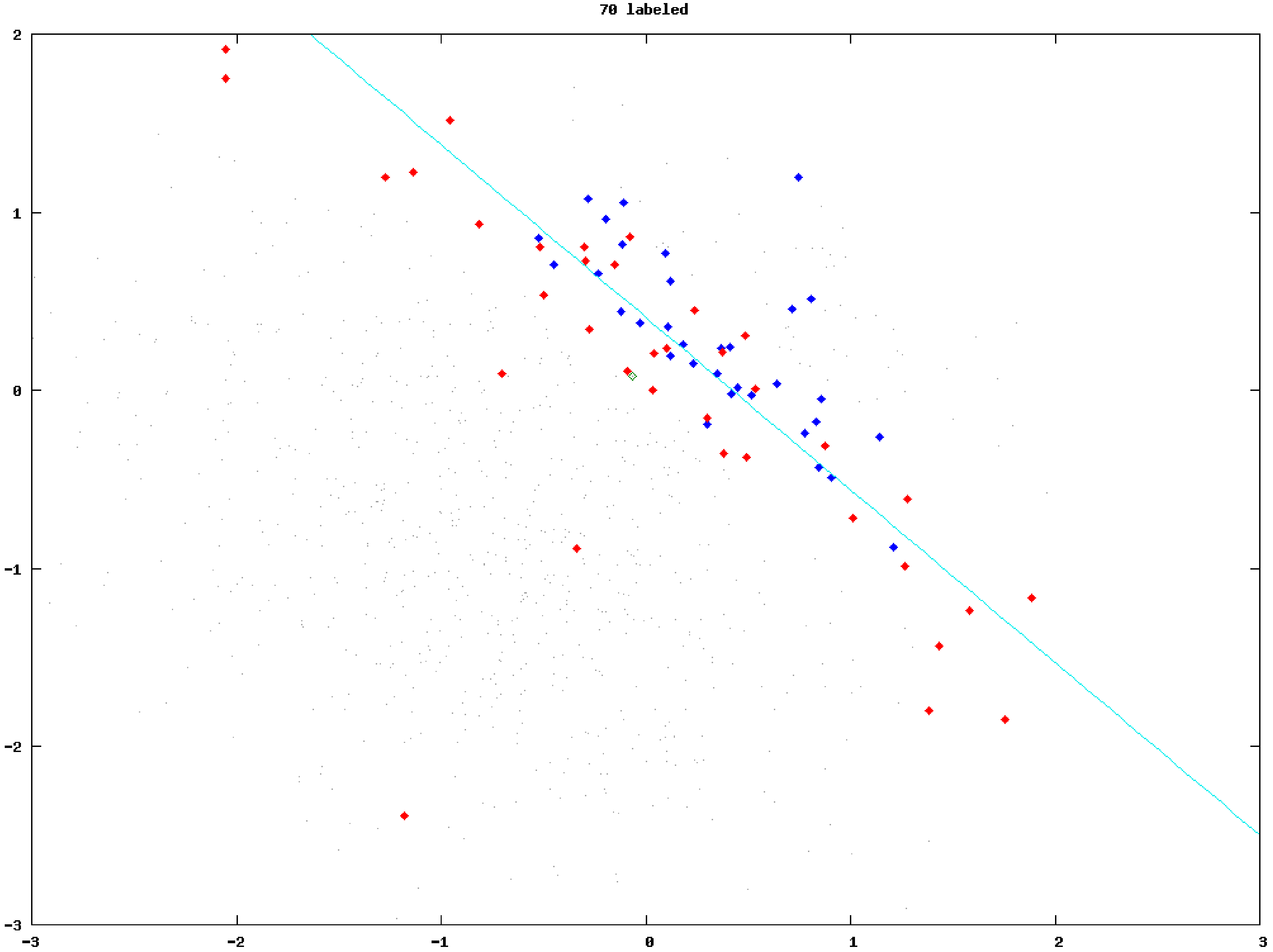

20 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

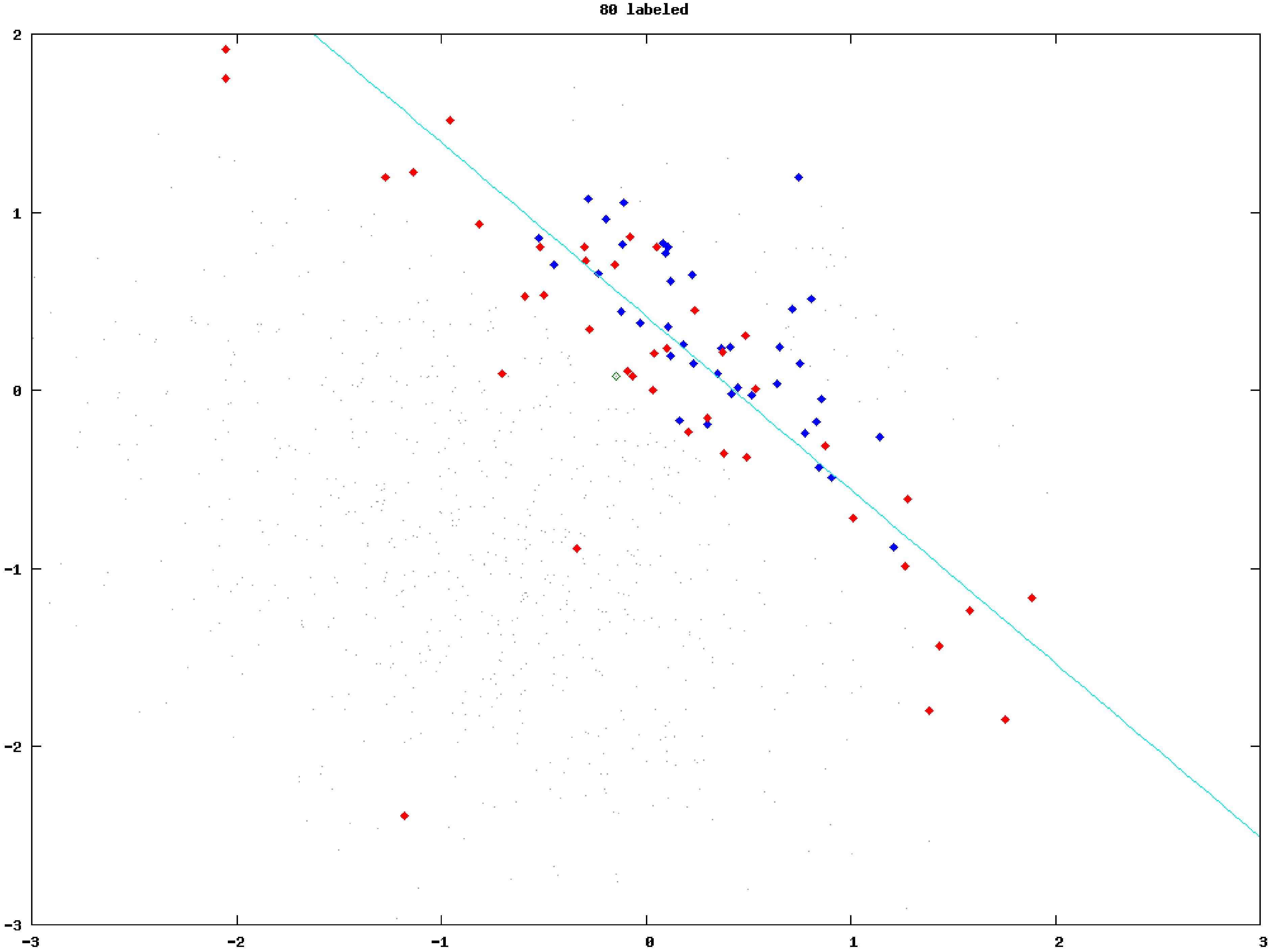

30 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


40 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

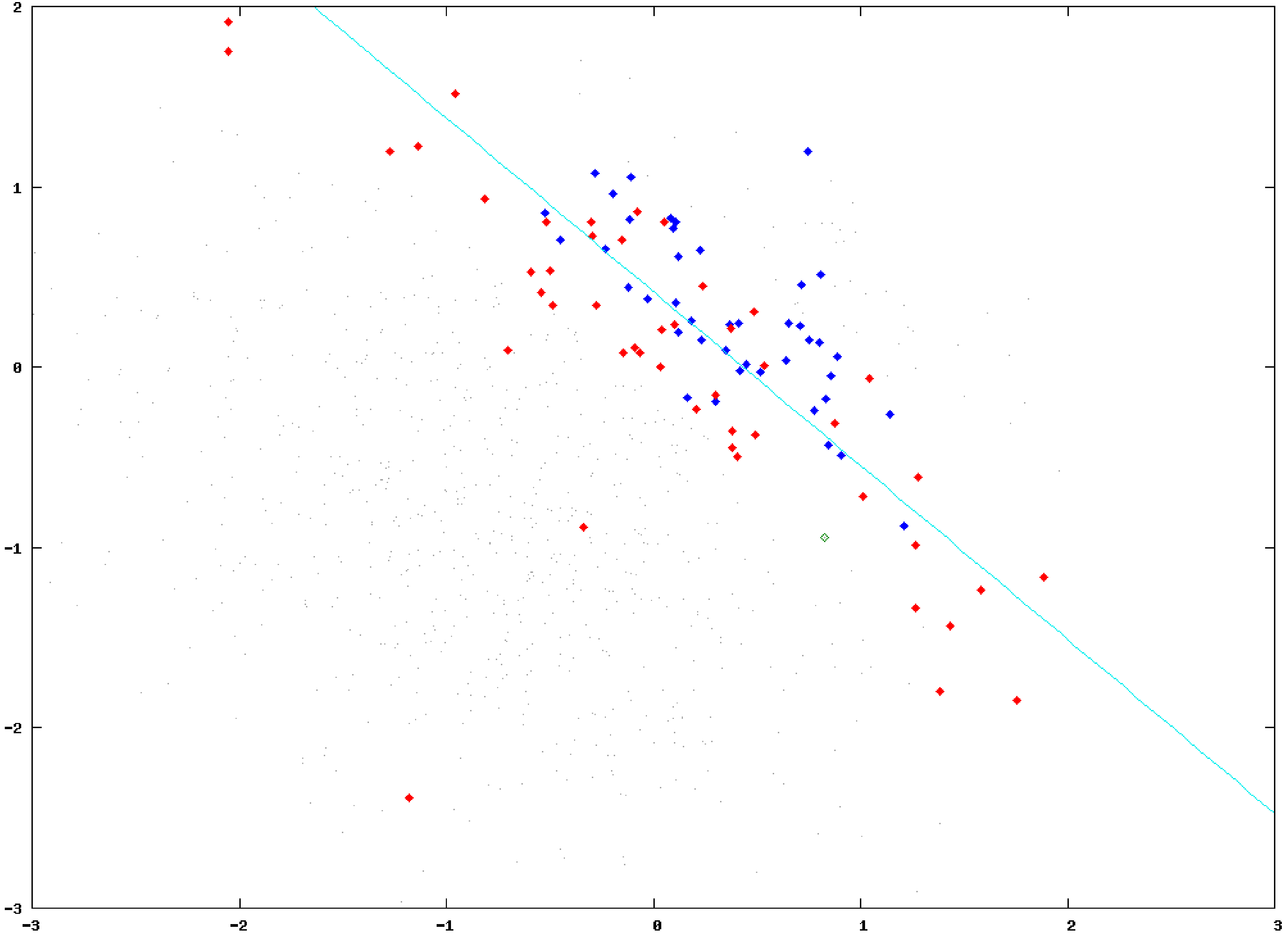

50 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

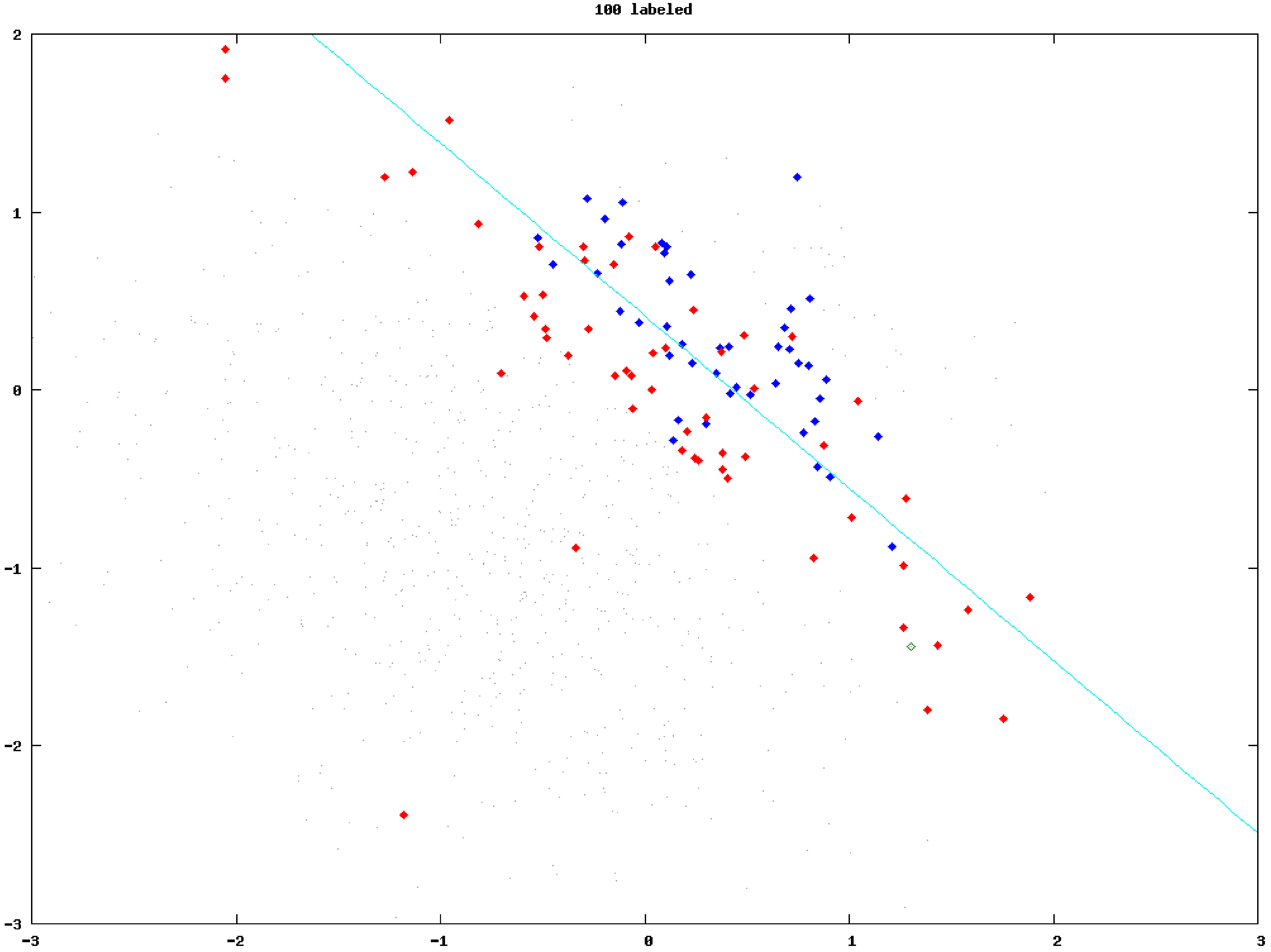

60 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

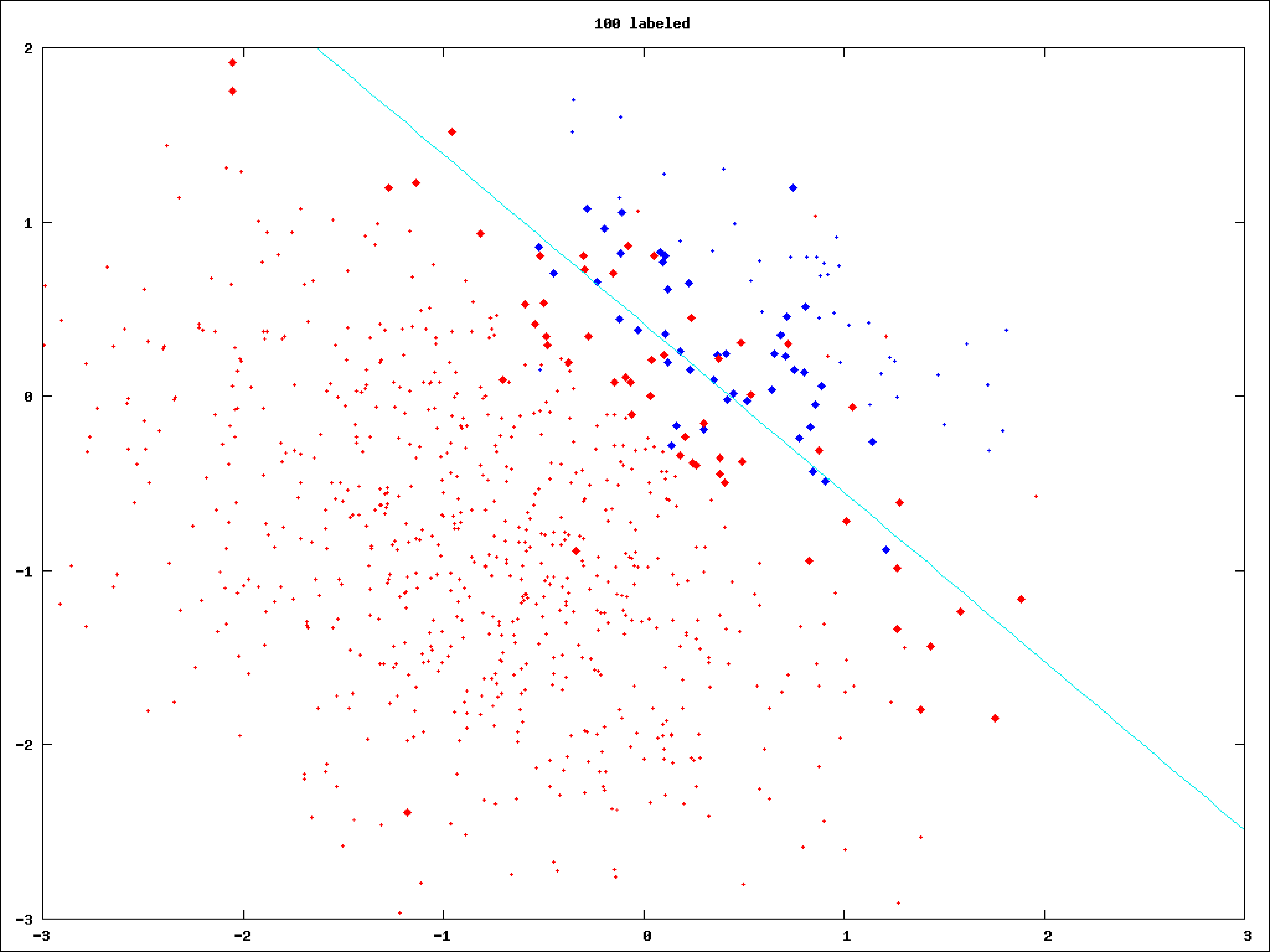

70 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


80 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

90 labeled




100 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


100 labeled
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Unsupervised Learning

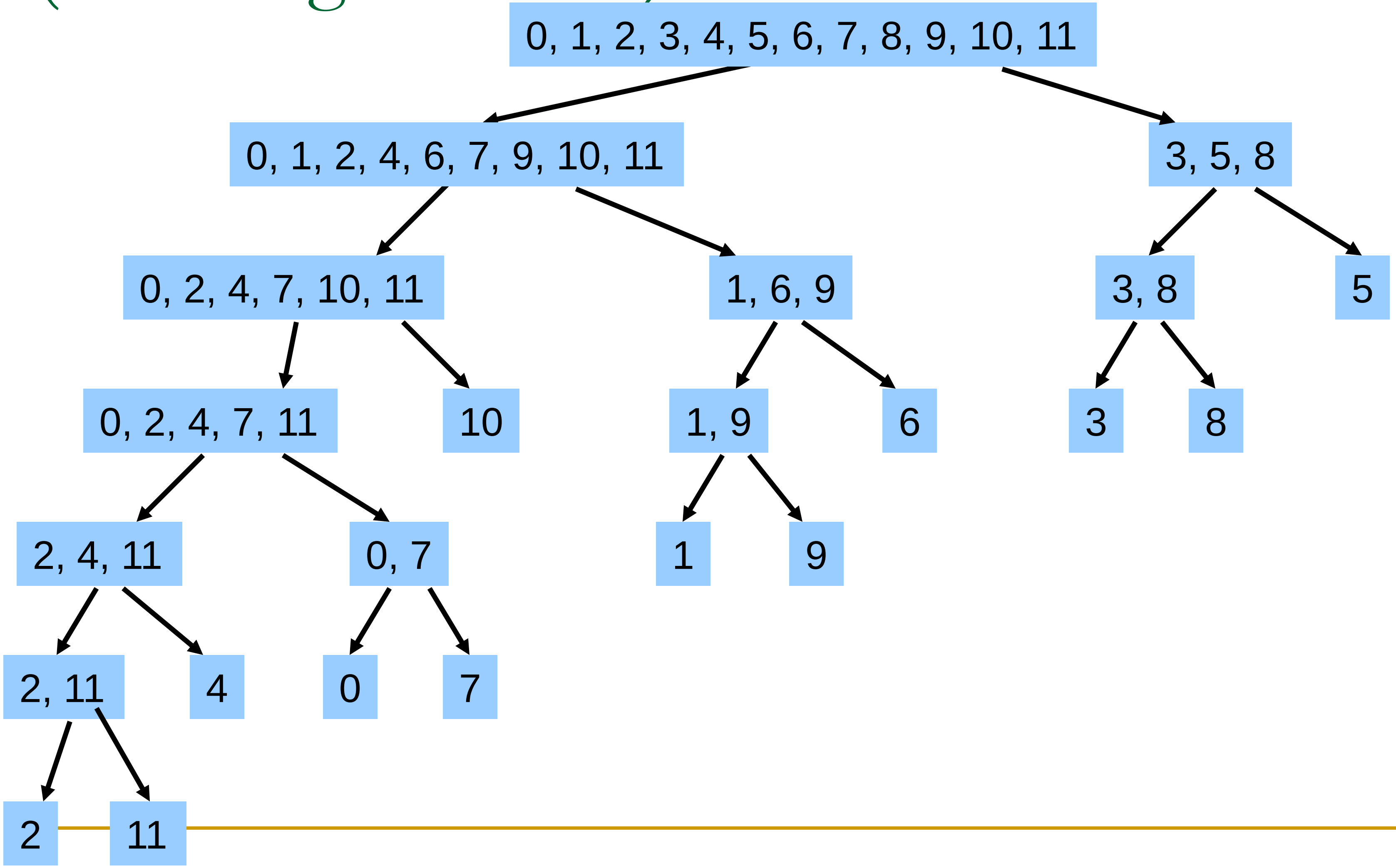
Document Clustering

- Clustering is a process of finding natural groups in the data in a unsupervised way (no class labels are pre-assigned to documents)
- Key element is similarity measure
  - In document clustering cosine similarity is most widely used
- Most popular clustering methods are:
  - K-Means clustering (flat, hierarchical)
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - …

K-Means clustering algorithm

- **Given**:
  - set of documents (e.g. TFIDF vectors),
  - distance measure (e.g. cosine)
  - ***K*** (number of groups)
- **For each** of ***K*** groups initialize its centroid with a random document
- **While** not converging
  - Each document is assigned to the nearest group (represented by its centroid)
  - For each group calculate new centroid (group mass point, average document in the group)

Example of hierarchical clustering (bisecting k-means)



0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

0, 1, 2, 4, 6, 7, 9, 10, 11

3, 5, 8

0, 2, 4, 7, 10, 11

1, 6, 9

3, 8

5

0, 2, 4, 7, 11

10

1, 9

6

3

8

2, 4, 11

0, 7

1

9

2, 11

4

0

7

2

11

Latent Semantic Indexing

- LSI is a statistical technique that attempts to estimate the hidden content structure within documents:
  - …it uses linear algebra technique Singular-Value-Decomposition (SVD)
  - …it discovers statistically most significant co-occurrences of terms

LSI Example

Original document-term mantrix

	d1	d2	d3	d4	d5	d6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

Rescaled document matrix,
Reduced into two dimensions

	d1	d2	d3	d4	d5	d6
Dim 1	-1.62	-0.60	-0.04	-0.97	-0.71	-0.26
Dim 2	-0.46	-0.84	-0.30	1.00	0.35	0.65

High correlation although d2 and d3 don't share any word

Correlation matrix

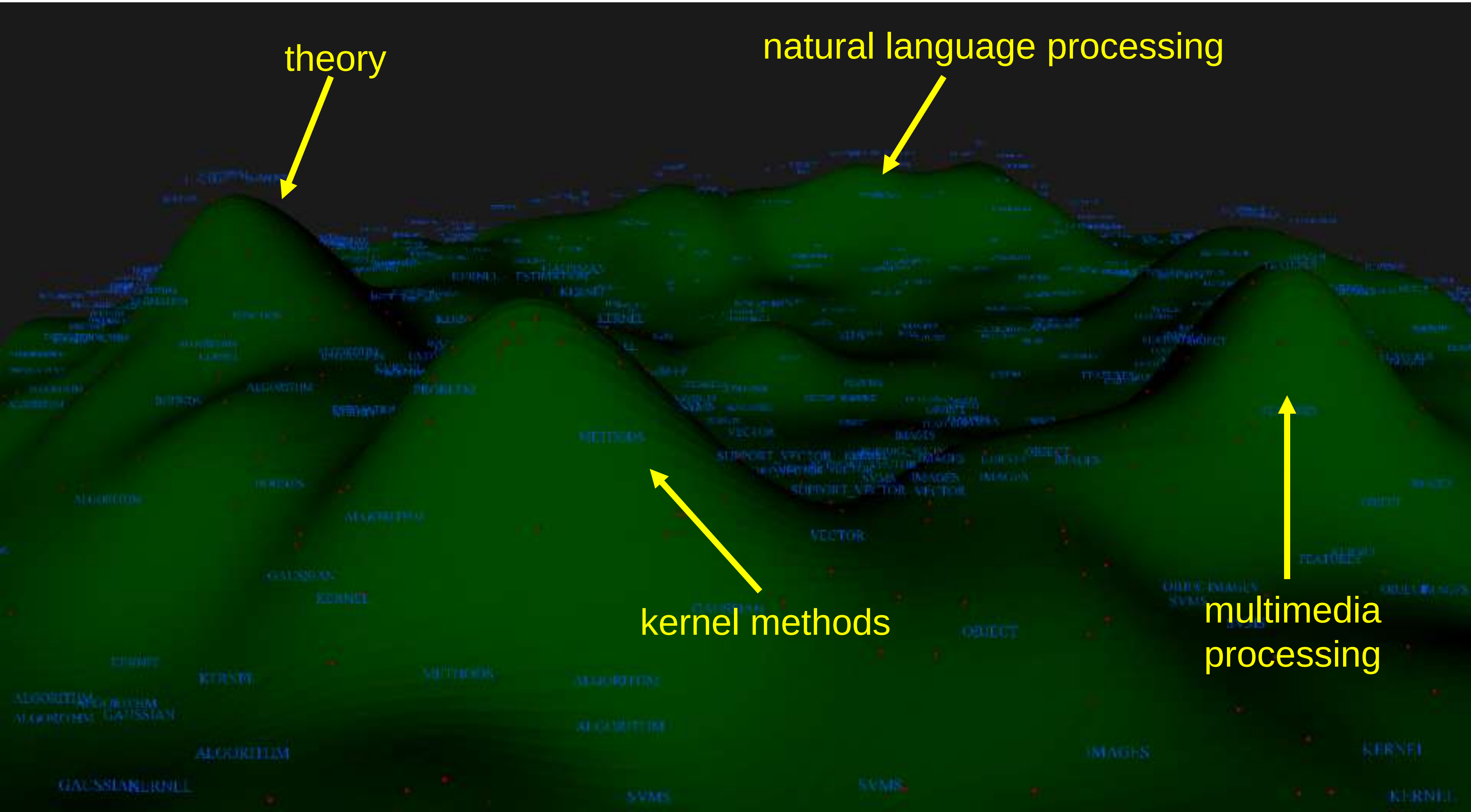
	d1	d2	d3	d4	d5	d6
d1	1.00					
d2	0.8	1.00				
d3	0.4	0.9	1.00			
d4	0.5	-0.2	-0.6	1.00		
d5	0.7	0.2	-0.3	0.9	1.00	
d6	0.1	-0.5	-0.9	0.9	0.7	1.00

Visualization

Why visualizing text?

- ...to have a top level view of the topics in the corpora
- ...to see relationships between the topics and objects in the corpora
- ...to understand better what's going on in the corpora
- ...to show highly structured nature of textual contents in a simplified way
- ...to show main dimensions of highly dimensional space of textual documents
- ...because it's fun!

Example: Visualization of PASCAL project research topics (based on published papers abstracts)



theory

natural language processing

kernel methods

multimedia processing

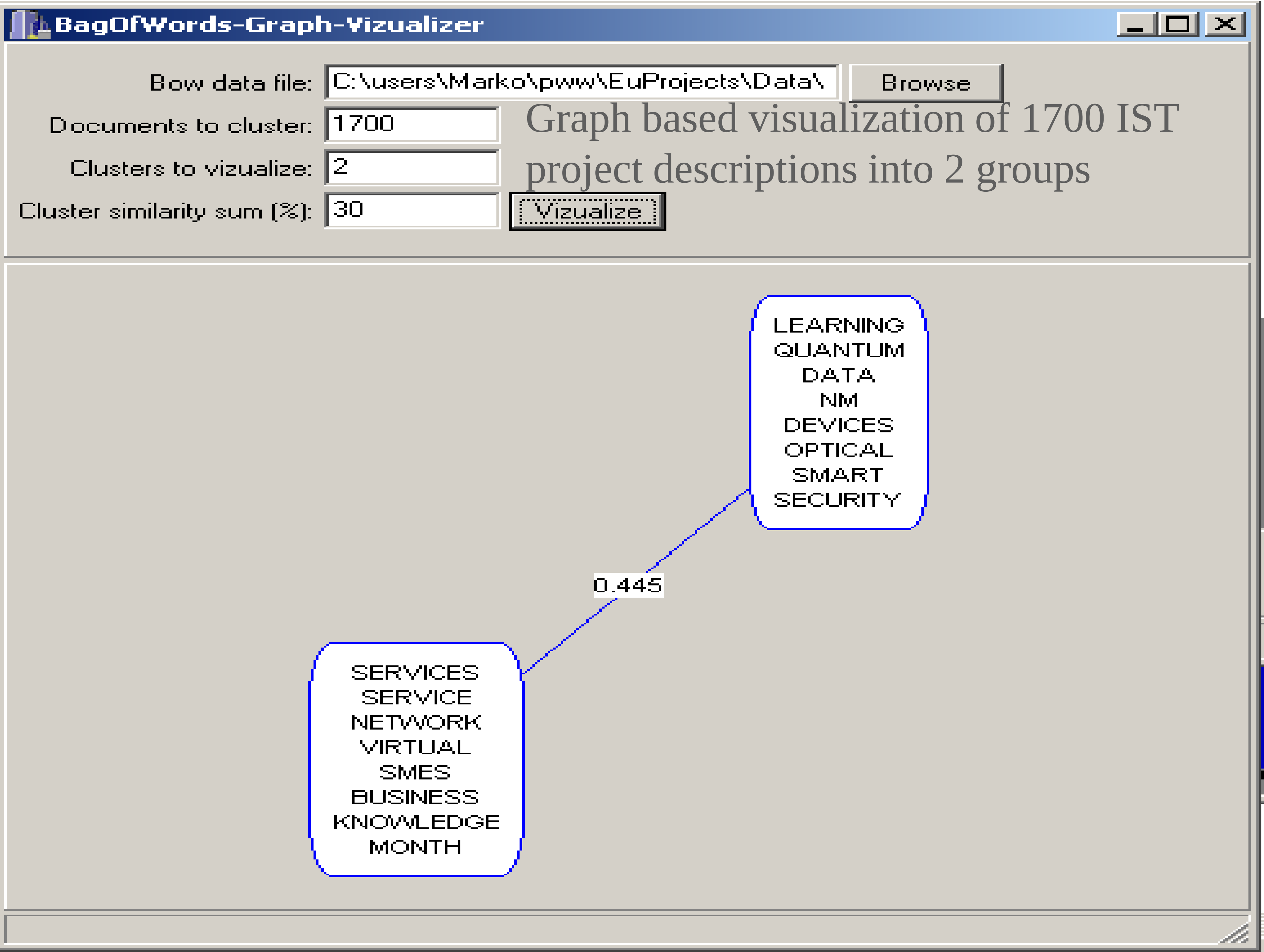
…typical way of doing text visualization

- By having text in the sparse vector Bag-of-Words representation we usually perform so kind of **clustering algorithm** identify structure which is then mapped into 2D or 3D space (e.g. using MDS)
- …other typical way of visualization of text is to find frequent co-occurrences of words and phrases which are visualized e.g. as graphs
- Typical visualization scenarios:
  - Visualization of document collections
  - Visualization of search results
  - Visualization of document timeline

Graph based visualization

- The sketch of the algorithm:
  1. Documents are transformed into the bag-of-words sparse-vectors representation
     - Words in the vectors are weighted using TFIDF
  2. K-Means clustering algorithm splits the documents into K groups
     - Each group consists from similar documents
     - Documents are compared using cosine similarity
  3. K groups form a graph:
     - Groups are nodes in graph; similar groups are linked
     - Each group is represented by characteristic keywords
  4. Using simulated annealing draw a graph

Graph based visualization of 1700 IST project descriptions into 2 groups



BagOfWords-Graph-Vizualizer

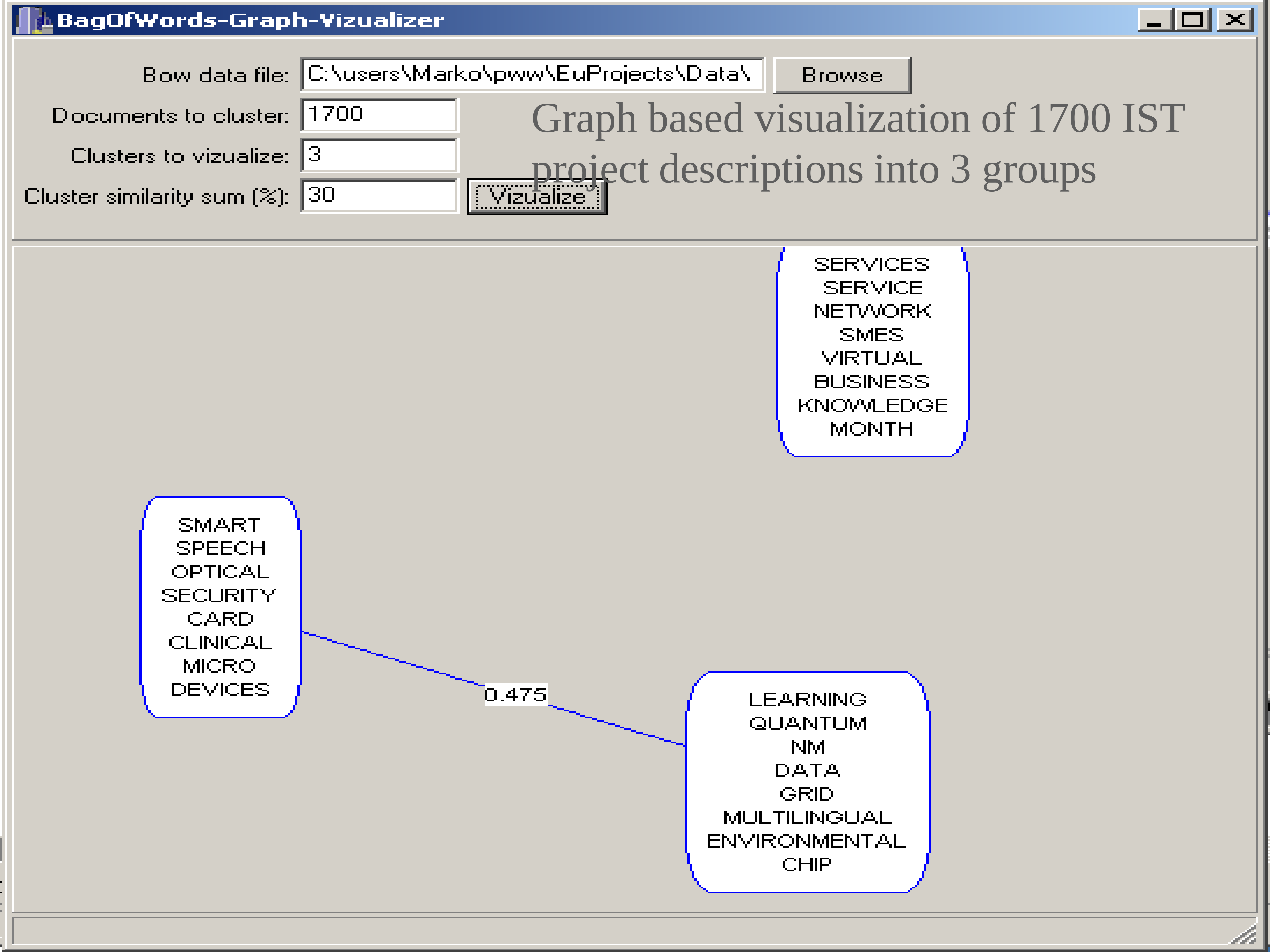
Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Clusters to vizualize: 2

Cluster similarity sum (%): 30

Graph based visualization of 1700 IST project descriptions into 3 groups



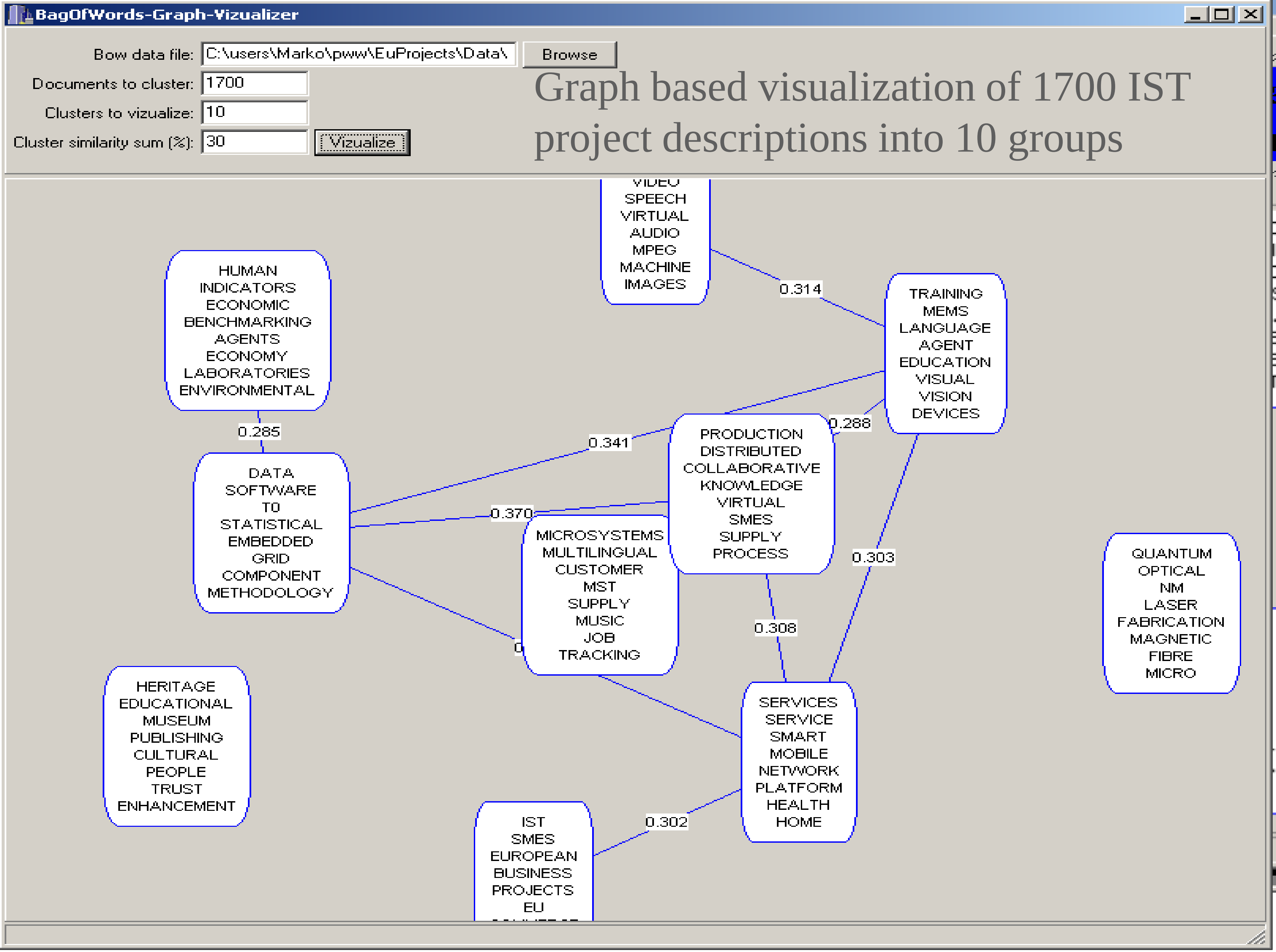
Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Clusters to vizualize: 3

Cluster similarity sum (%): 30

Graph based visualization of 1700 IST project descriptions into 10 groups



BagOfWords-Graph-Vizualizer

Bow data file: C:\users\Marko\pww\EuProjects\Data\

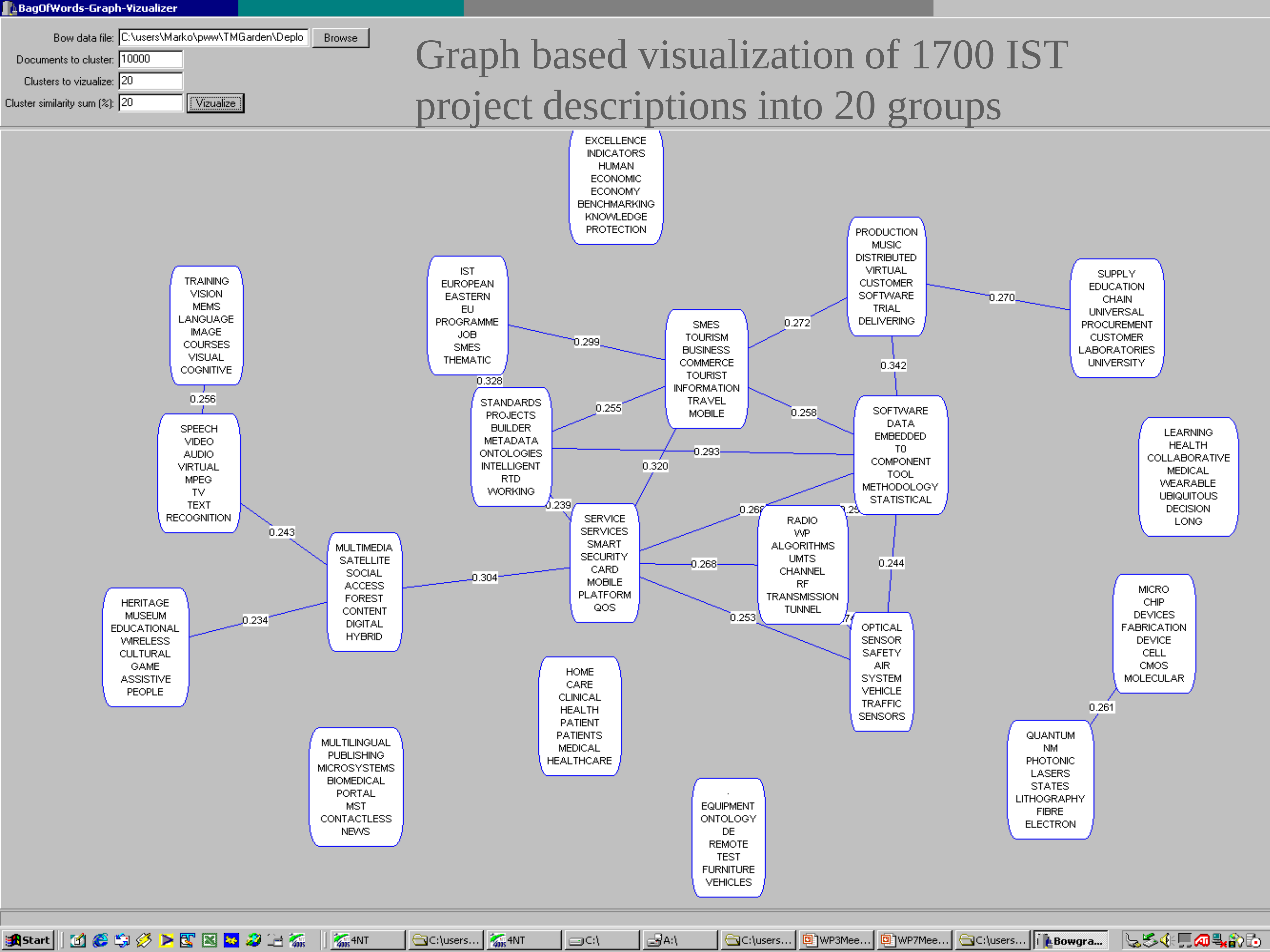
Documents to cluster: 1700

Clusters to vizualize: 10

Cluster similarity sum (%): 30

Browse
Vizualize
VIDEO SPEECH VIRTUAL AUDIO MPEG MACHINE IMAGES
0.314
TRAINING MEMS LANGUAGE AGENT EDUCATION VISUAL VISION DEVICES
HUMAN INDICATORS ECONOMIC BENCHMARKING AGENTS ECONOMY LABORATORIES ENVIRONMENTAL
0.285
0.288
0.341
PRODUCTION DISTRIBUTED COLLABORATIVE KNOWLEDGE VIRTUAL SMES SUPPLY PROCESS
DATA SOFTWARE TO STATISTICAL EMBEDDED GRID COMPONENT METHODOLOGY
0.370
MICROSYSTEMS MULTILINGUAL CUSTOMER MST SUPPLY MUSIC JOB TRACKING
0
0.303
QUANTUM OPTICAL NM LASER FABRICATION MAGNETIC FIBRE MICRO
0.308
HERITAGE EDUCATIONAL MUSEUM PUBLISHING CULTURAL PEOPLE TRUST ENHANCEMENT
SERVICES SERVICE SMART MOBILE NETWORK PLATFORM HEALTH HOME
0.302
IST SMES EUROPEAN BUSINESS PROJECTS EU

Graph based visualization of 1700 IST project descriptions into 20 groups



BagOfWords-Graph-Vizualizer

Bow data file: C:\users\Marko\pww\TMGarden\Deplo

Documents to cluster: 10000

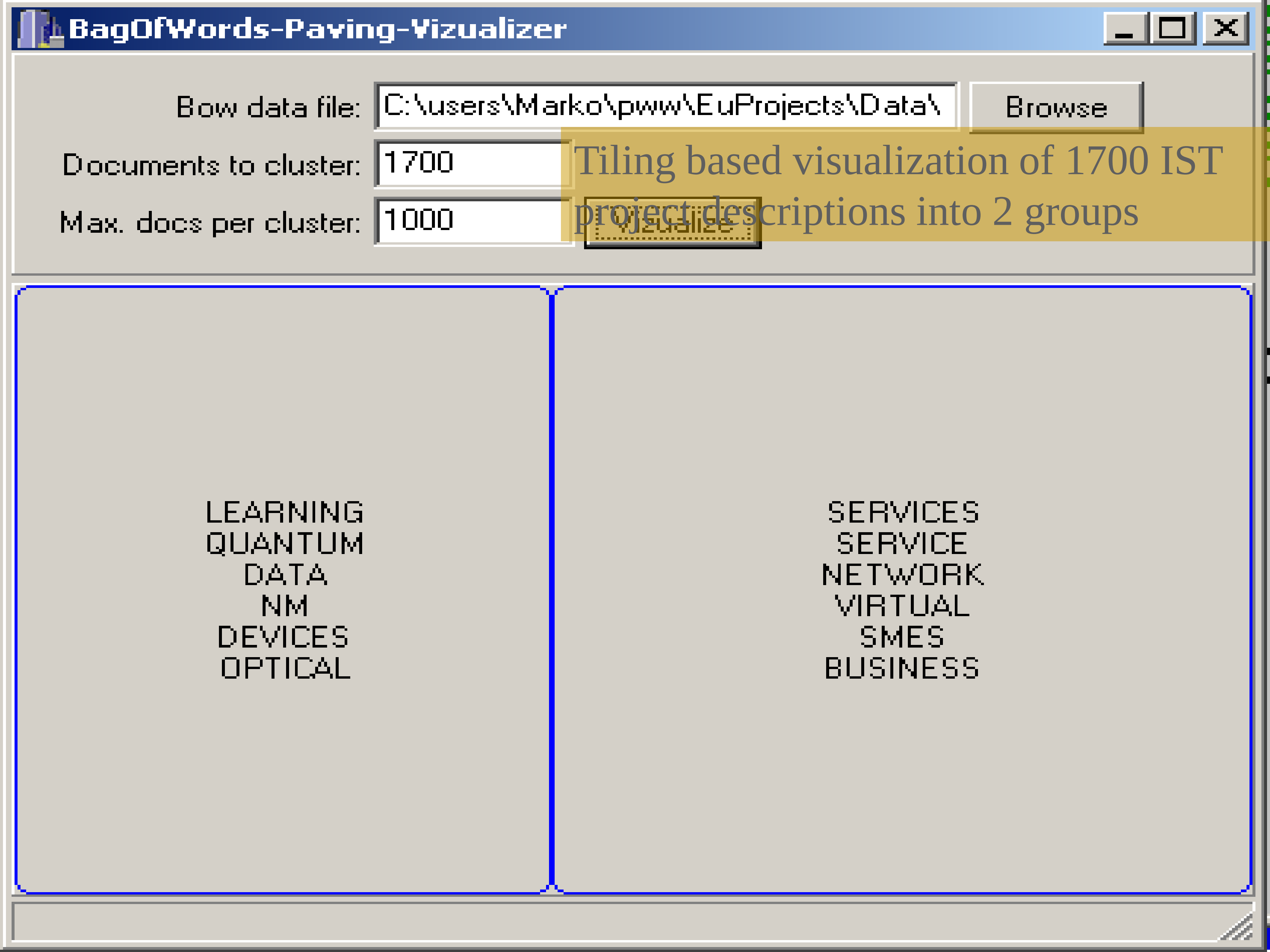
Clusters to vizualize: 20

Cluster similarity sum (%): 20

Browse
Vizualize
EXCELLENCE INDICATORS HUMAN ECONOMIC ECONOMY BENCHMARKING KNOWLEDGE PROTECTION
PRODUCTION MUSIC DISTRIBUTED VIRTUAL CUSTOMER SOFTWARE TRIAL DELIVERING
SUPPLY EDUCATION CHAIN UNIVERSAL PROCUREMENT CUSTOMER LABORATORIES UNIVERSITY
0.270
IST EUROPEAN EASTERN EU PROGRAMME JOB SMES THEMATIC
TRAINING VISION MEMS LANGUAGE IMAGE COURSES VISUAL COGNITIVE
0.299
SMES TOURISM BUSINESS COMMERCE TOURIST INFORMATION TRAVEL MOBILE
0.272
0.342
0.328
0.256
0.255
0.258
STANDARDS PROJECTS BUILDER METADATA ONTOLOGIES INTELLIGENT RTD WORKING
SPEECH VIDEO AUDIO VIRTUAL MPEG TV TEXT RECOGNITION
SOFTWARE DATA EMBEDDED TO COMPONENT TOOL METHODOLOGY STATISTICAL
LEARNING HEALTH COLLABORATIVE MEDICAL WEARABLE UBIQUITOUS DECISION LONG
0.293
0.320
0.239
0.26
0.25
SERVICE SERVICES SMART SECURITY CARD MOBILE PLATFORM QOS
RADIO WP ALGORITHMS UMTS CHANNEL RF TRANSMISSION TUNNEL
0.243
MULTIMEDIA SATELLITE SOCIAL ACCESS FOREST CONTENT DIGITAL HYBRID
0.268
0.244
0.304
MICRO CHIP DEVICES FABRICATION DEVICE CELL CMOS MOLECULAR
HERITAGE MUSEUM EDUCATIONAL WIRELESS CULTURAL GAME ASSISTIVE PEOPLE
0.234
0.253
7
OPTICAL SENSOR SAFETY AIR SYSTEM VEHICLE TRAFFIC SENSORS
HOME CARE CLINICAL HEALTH PATIENT PATIENTS MEDICAL HEALTHCARE
0.261
MULTILINGUAL PUBLISHING MICROSYSTEMS BIOMEDICAL PORTAL MST CONTACTLESS NEWS
QUANTUM NM PHOTONIC LASERS STATES LITHOGRAPHY FIBRE ELECTRON
EQUIPMENT ONTOLOGY DE REMOTE TEST FURNITURE VEHICLES
Start
4NT
C:\users...
4NT
C:\
A:\
C:\users...
WP3Mee...
WP7Mee...
C:\users...
Bowgra...

Tiling based visualization

- The sketch of the algorithm:
  1. Documents are transformed into the bag-of-words sparse-vectors representation
     - Words in the vectors are weighted using TFIDF
  2. Hierarchical top-down two-wise K-Means clustering algorithm builds a hierarchy of clusters
     - The hierarchy is an artificial equivalent of hierarchical subject index (Yahoo like)
  3. The leaf nodes of the hierarchy (bottom level) are used to visualize the documents
     - Each leaf is represented by characteristic keywords
     - Each hierarchical binary split splits recursively the rectangular area into two sub-areas



Tiling based visualization of 1700 IST project descriptions into 2 groups

BagOfWords-Paving-Vizualizer

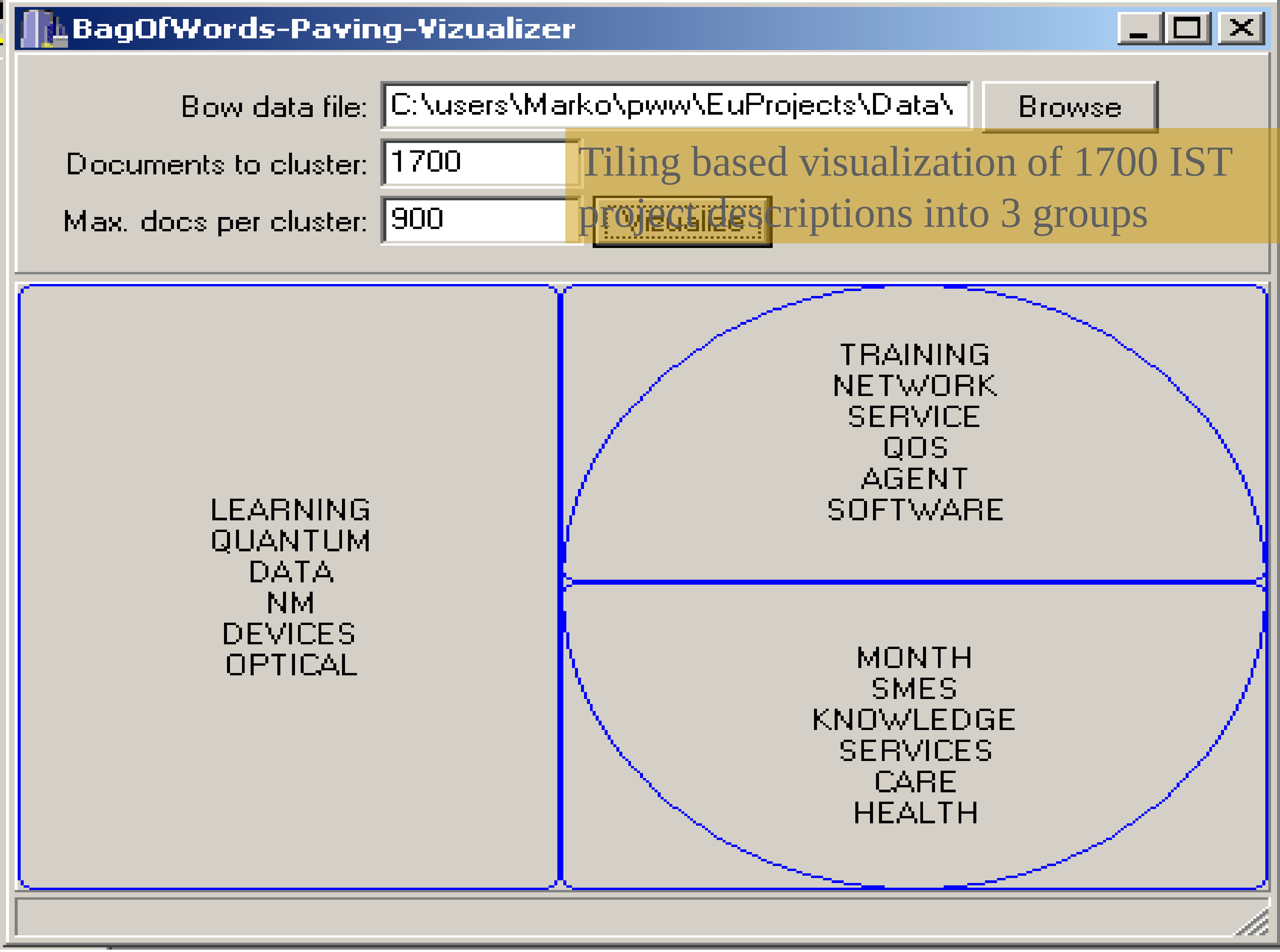
Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Max. docs per cluster: 1000

LEARNING
QUANTUM
DATA
NM
DEVICES
OPTICAL

SERVICES
SERVICE
NETWORK
VIRTUAL
SMES
BUSINESS



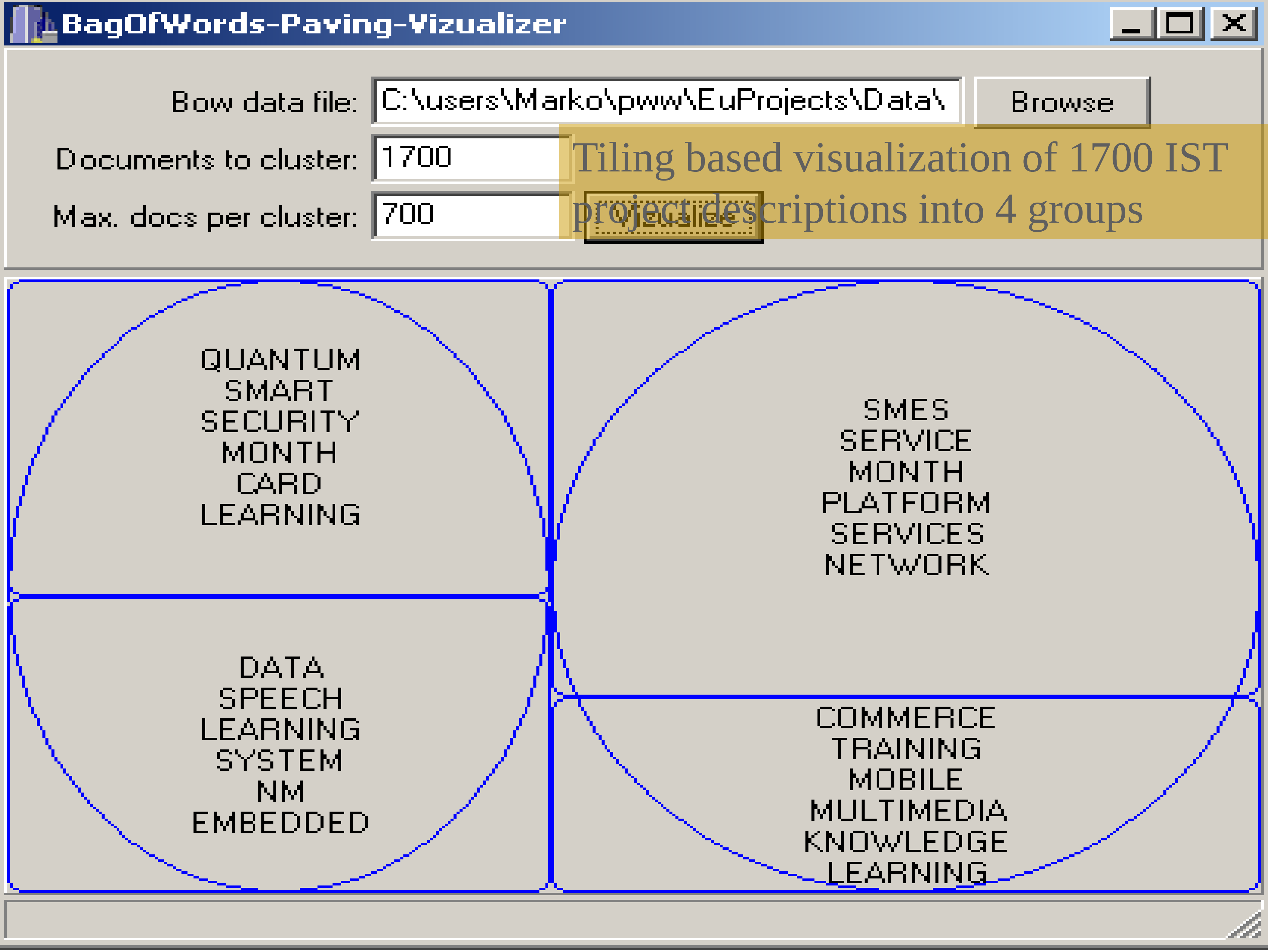
Tiling based visualization of 1700 IST project descriptions into 3 groups

BagOfWords-Paving-Vizualizer

Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Max. docs per cluster: 900

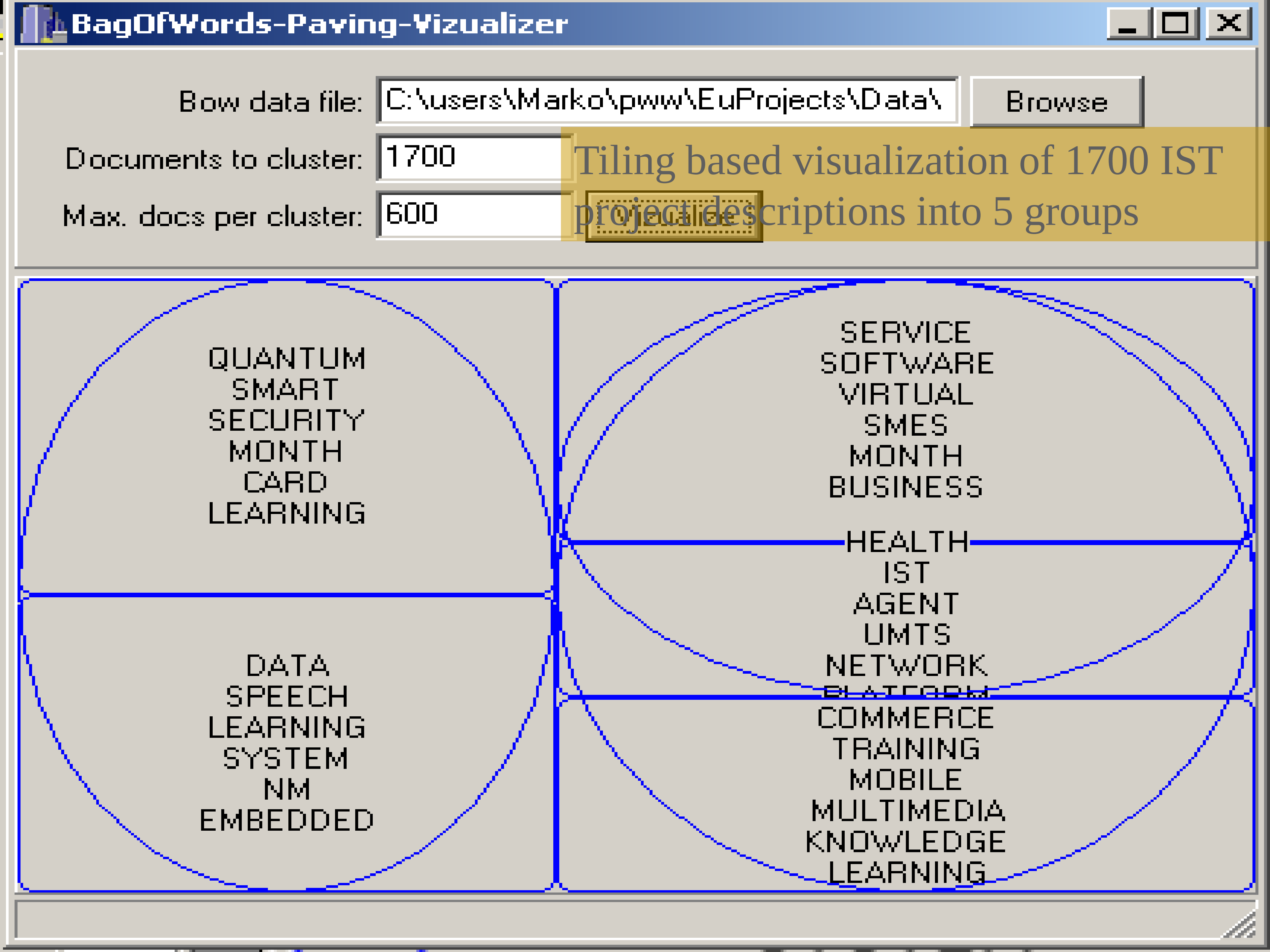


Tiling based visualization of 1700 IST project descriptions into 4 groups

Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Max. docs per cluster: 700



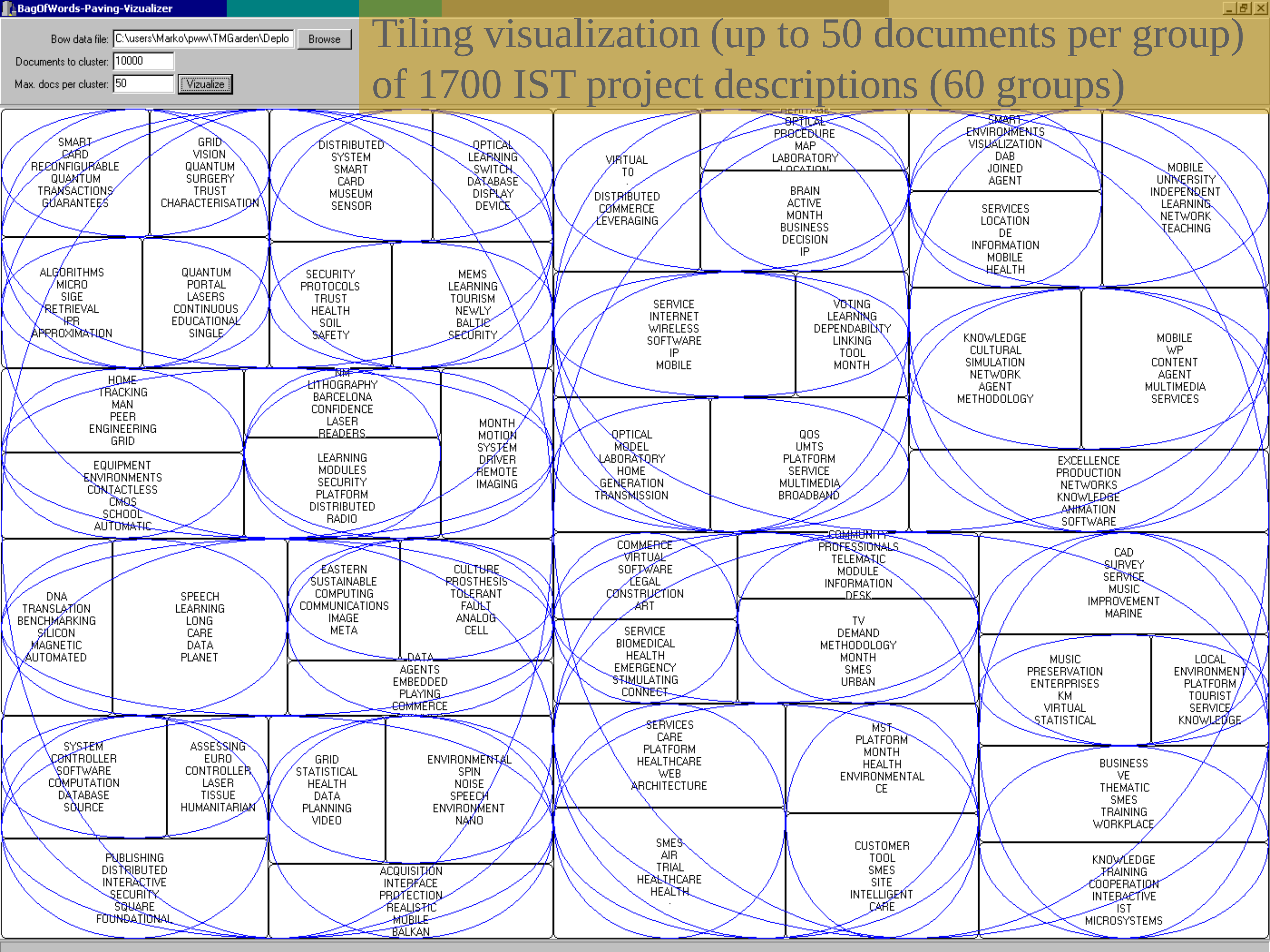
Bow data file: C:\users\Marko\pww\EuProjects\Data\

Documents to cluster: 1700

Max. docs per cluster: 600

Tiling based visualization of 1700 IST project descriptions into 5 groups

Tiling visualization (up to 50 documents per group) of 1700 IST project descriptions (60 groups)



BagOfWords-Paving-Vizualizer

Bow data file: C:\users\Marko\pww\TMGarden\Deplo [Browse]

Documents to cluster: 10000

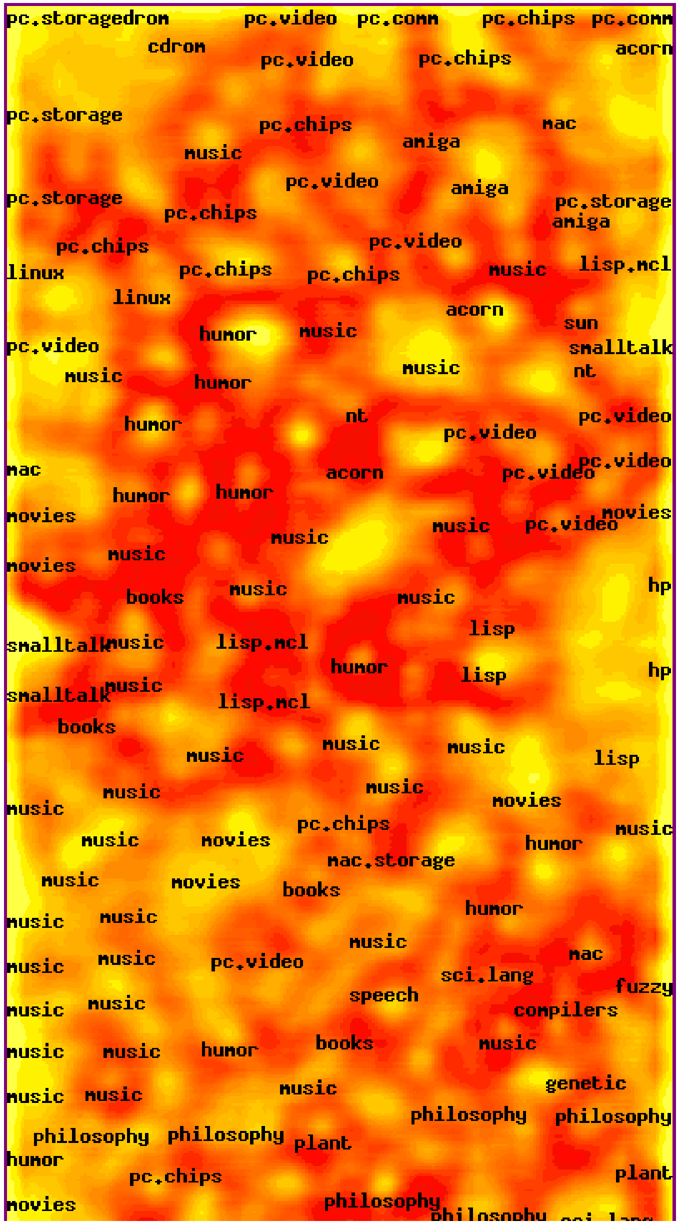
Max. docs per cluster: 50 [Vizualize]

SMART CARD RECONFIGURABLE QUANTUM TRANSACTIONS GUARANTEES
GRID VISION QUANTUM SURGERY TRUST CHARACTERISATION
DISTRIBUTED SYSTEM SMART CARD MUSEUM SENSOR
OPTICAL LEARNING SWITCH DATABASE DISPLAY DEVICE
ALGORITHMS MICRO SIGE RETRIEVAL IPR APPROXIMATION
QUANTUM PORTAL LASERS CONTINUOUS EDUCATIONAL SINGLE
SECURITY PROTOCOLS TRUST HEALTH SOIL SAFETY
MEMS LEARNING TOURISM NEWLY BALTIC SECURITY
HOME TRACKING MAN PEER ENGINEERING GRID
NM LITHOGRAPHY BARCELONA CONFIDENCE LASER READERS
MONTH MOTION SYSTEM DRIVER REMOTE IMAGING
EQUIPMENT ENVIRONMENTS CONTACTLESS CMOS SCHOOL AUTOMATIC
LEARNING MODULES SECURITY PLATFORM DISTRIBUTED RADIO
DNA TRANSLATION BENCHMARKING SILICON MAGNETIC AUTOMATED
SPEECH LEARNING LONG CARE DATA PLANET
EASTERN SUSTAINABLE COMPUTING COMMUNICATIONS IMAGE META
CULTURE PROSTHESIS TOLERANT FAULT ANALOG CELL
DATA AGENTS EMBEDDED PLAYING COMMERCE
SYSTEM CONTROLLER SOFTWARE COMPUTATION DATABASE SOURCE
ASSESSING EURO CONTROLLER LASER TISSUE HUMANITARIAN
GRID STATISTICAL HEALTH DATA PLANNING VIDEO
ENVIRONMENTAL SPIN NOISE SPEECH ENVIRONMENT NANO
PUBLISHING DISTRIBUTED INTERACTIVE SECURITY SQUARE FOUNDATIONAL
ACQUISITION INTERFACE PROTECTION REALISTIC MOBILE BALKAN
VIRTUAL TO DISTRIBUTED COMMERCE LEVERAGING
HERITAGE OPTICAL PROCEDURE MAP LABORATORY LOCATION
BRAIN ACTIVE MONTH BUSINESS DECISION IP
SMART ENVIRONMENTS VISUALIZATION DAB JOINED AGENT
SERVICES LOCATION DE INFORMATION MOBILE HEALTH
MOBILE UNIVERSITY INDEPENDENT LEARNING NETWORK TEACHING
SERVICE INTERNET WIRELESS SOFTWARE IP MOBILE
VOTING LEARNING DEPENDABILITY LINKING TOOL MONTH
KNOWLEDGE CULTURAL SIMULATION NETWORK AGENT METHODOLOGY
MOBILE WP CONTENT AGENT MULTIMEDIA SERVICES
OPTICAL MODEL LABORATORY HOME GENERATION TRANSMISSION
QOS UMTS PLATFORM SERVICE MULTIMEDIA BROADBAND
EXCELLENCE PRODUCTION NETWORKS KNOWLEDGE ANIMATION SOFTWARE
COMMERCE VIRTUAL SOFTWARE LEGAL CONSTRUCTION ART
COMMUNITY PROFESSIONALS TELEMATIC MODULE INFORMATION DESK
CAD SURVEY SERVICE MUSIC IMPROVEMENT MARINE
SERVICE BIOMEDICAL HEALTH EMERGENCY STIMULATING CONNECT
TV DEMAND METHODOLOGY MONTH SMES URBAN
MUSIC PRESERVATION ENTERPRISES KM VIRTUAL STATISTICAL
LOCAL ENVIRONMENT PLATFORM TOURIST SERVICE KNOWLEDGE
SERVICES CARE PLATFORM HEALTHCARE WEB ARCHITECTURE
MST PLATFORM MONTH HEALTH ENVIRONMENTAL CE
BUSINESS VE THEMATIC SMES TRAINING WORKPLACE
SMES AIR TRIAL HEALTHCARE HEALTH
CUSTOMER TOOL SMES SITE INTELLIGENT CARE
KNOWLEDGE TRAINING COOPERATION INTERACTIVE IST MICROSYSTEMS

WebSOM

- Self-Organizing Maps for Internet Exploration
  - …algorithm that automatically organizes the documents onto a two-dimensional grid so that related documents appear close to each other
  - … based on Kohonen’s Self-Organizing Maps
  - Demo at http://websom.hut.fi/websom/

WebSOM visualization



Explanation of the symbols on the map

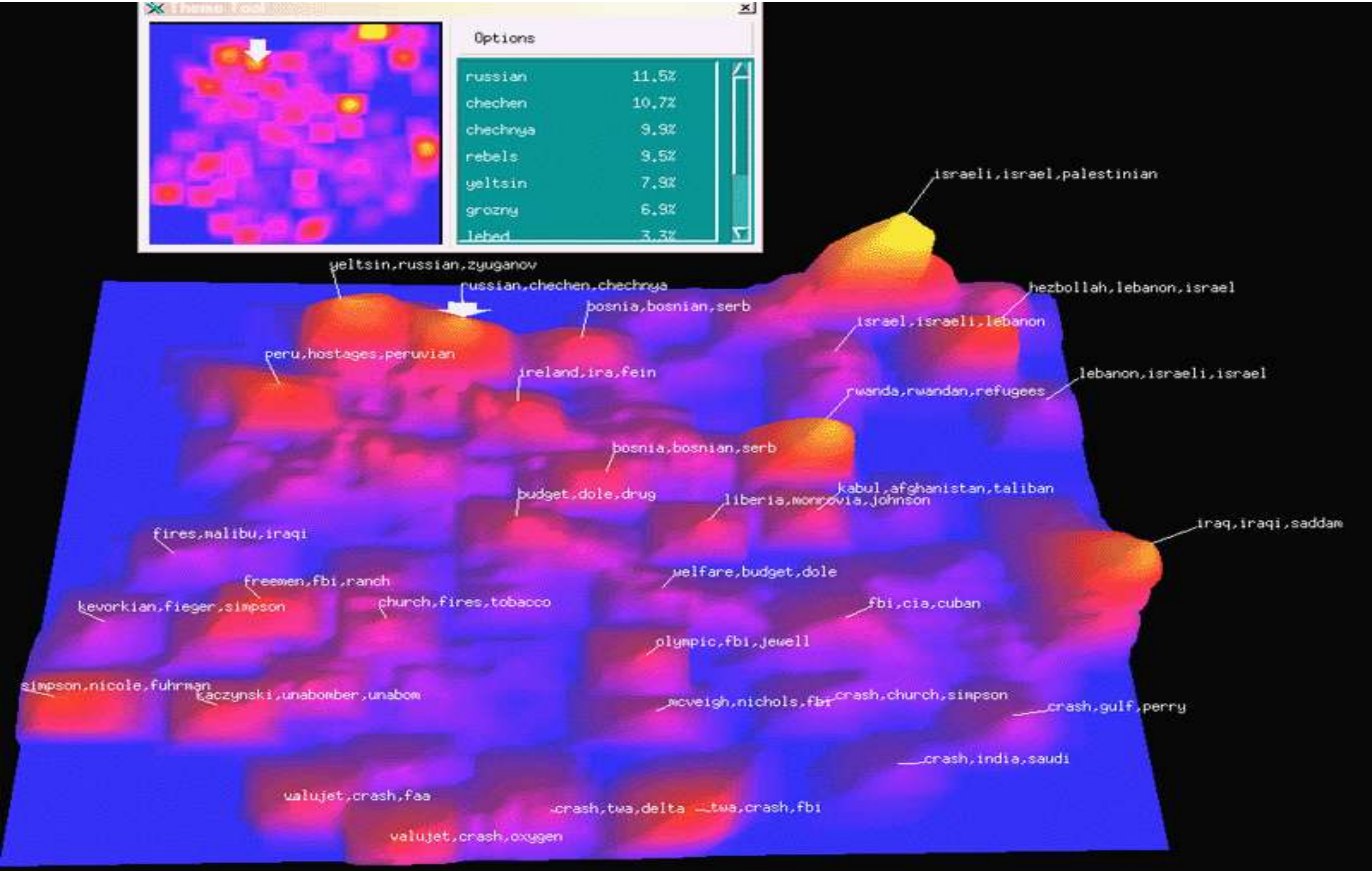
acorn - comp.sys.acorn.hardware
amiga - comp.sys.amiga.hardware
booksEach white20 - rec.arts.books
cdrom - comp.publish.cdrom.hardware
compilers - comp.compilers
fuzzy - comp.ai.fuzzy
genetic - comp.ai.genetic
hp - comp.sys.hp.hardware
humor - rec.humor
lang.eiffel - comp.lang.eiffel
lang.ml - comp.lang.ml
linux - comp.os.linux.hardware
lisp - comp.lang.lisp
lisp.mcl - comp.lang.lisp.mcl
mac - comp.sys.mac.hardware.misc
mac.storage - comp.sys.mac.hardware.storage
movies - movies
music - music
nt - comp.os.ms-windows.nt.setup.hardware
pc.cdrom - comp.sys.ibm.pc.hardware.cd-rom
pc.chips - comp.sys.ibm.pc.hardware.chips
pc.comm - comp.sys.ibm.pc.hardware.comm
pc.storage - comp.sys.ibm.pc.hardware.storage
pc.video - comp.sys.ibm.pc.hardware.video
philosophy - philosophy
plant - bionet.biology.plant
prolog - comp.lang.prolog
sci.lang - sci.lang
smalltalk - comp.lang.smalltalk

ThemeScape

- Graphically displays images based on word similarities and themes in text
- Themes within the document spaces appear on the computer screen as a relief map of natural terrain
  - The mountains in indicate where themes are dominant - valleys indicate weak themes
  - Themes close in content will be close visually based on the many relationships within the text spaces
  - Algorithm is based on K-means clustering

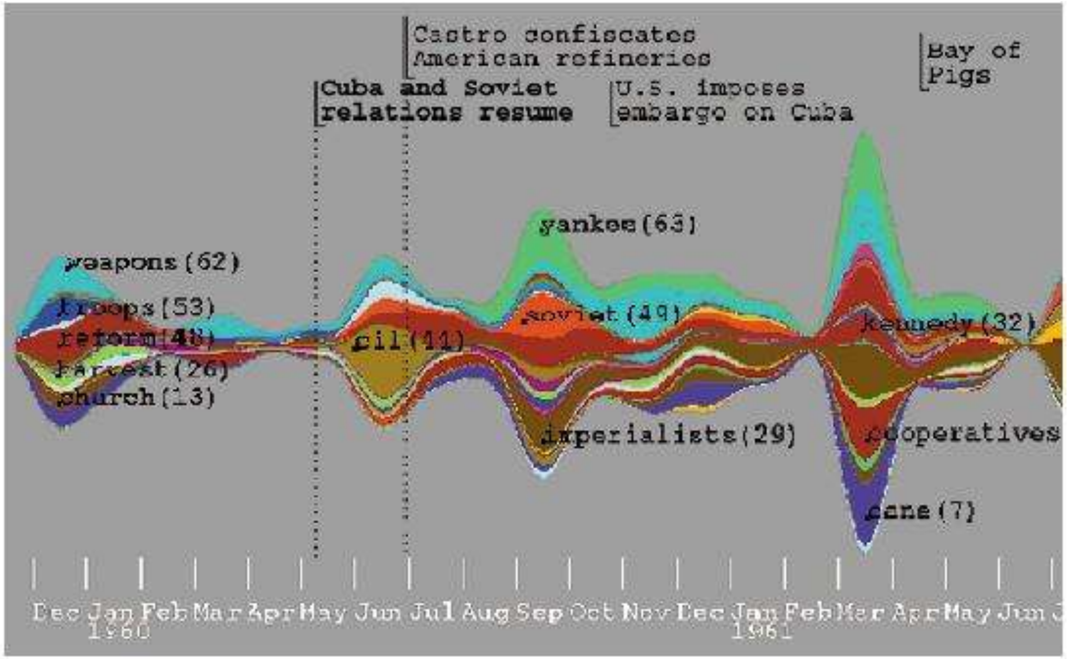
http://www.pnl.gov/infoviz/technologies.html

ThemeScape Document visualization



russian	11.5%
chechen	10.7%
chechnya	9.9%
rebels	9.5%
yeltsin	7.9%
grozny	6.9%
lebed	3.3%

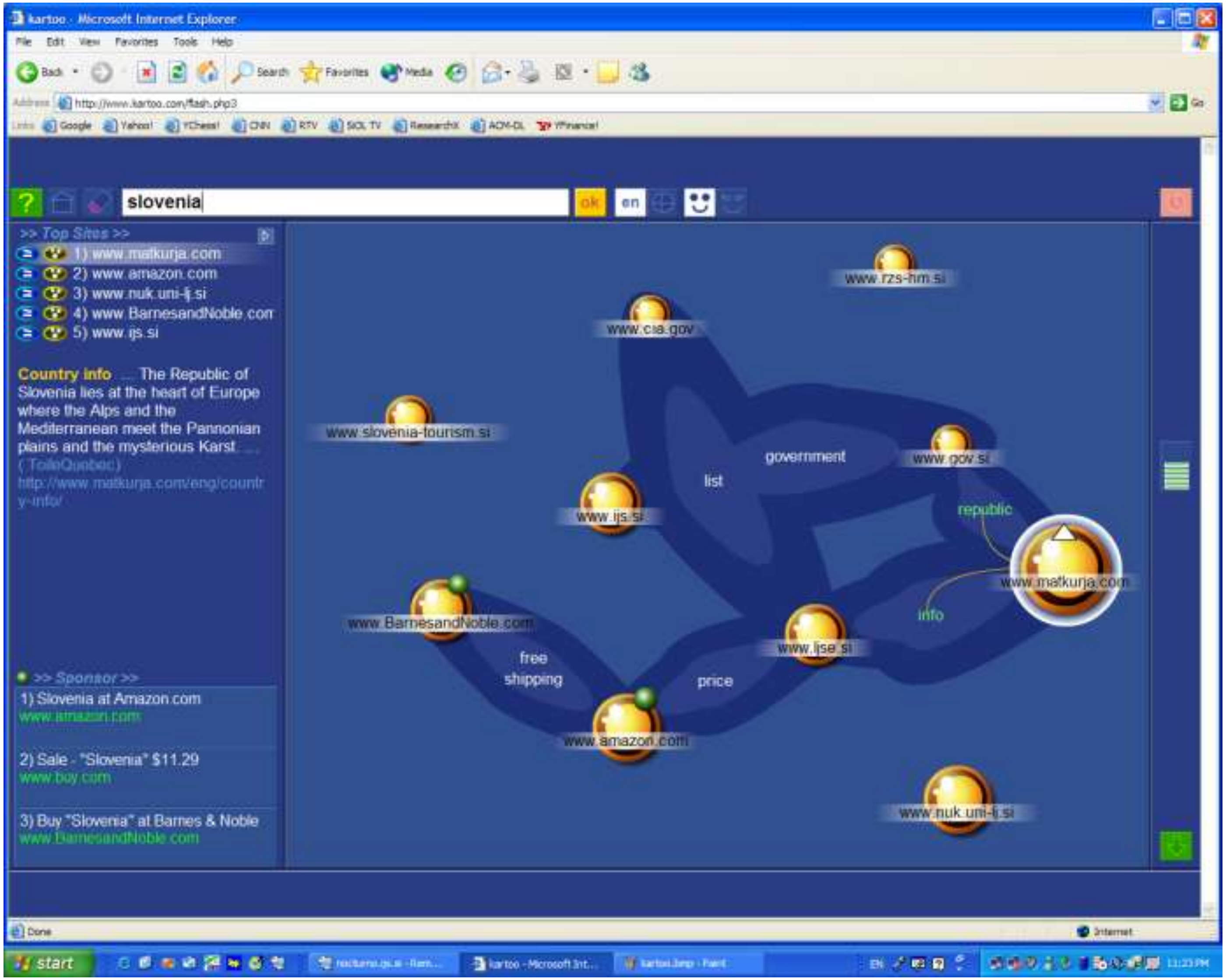
ThemeRiver
topic stream visualization



- The ThemeRiver visualization helps users identify time-related patterns, trends, and relationships across a large collection of documents.
- The themes in the collection are represented by a "river" that flows left to right through time.
- The theme currents narrow or widen to indicate changes in individual theme strength at any point in time.

http://www.pnl.gov/infoviz/technologies.html

Kartoo.com – visualization of search results



slovenia

>> Top Sites >>

1) www.matkurja.com
2) www.amazon.com
3) www.nuk.uni-lj.si
4) www.BarnesandNoble.com
5) www.ijs.si

Country info ... The Republic of Slovenia lies at the heart of Europe where the Alps and the Mediterranean meet the Pannonian plains and the mysterious Karst ... (ToileQuebec) http://www.matkurja.com/eng/country-info/

>> Sponsor >>

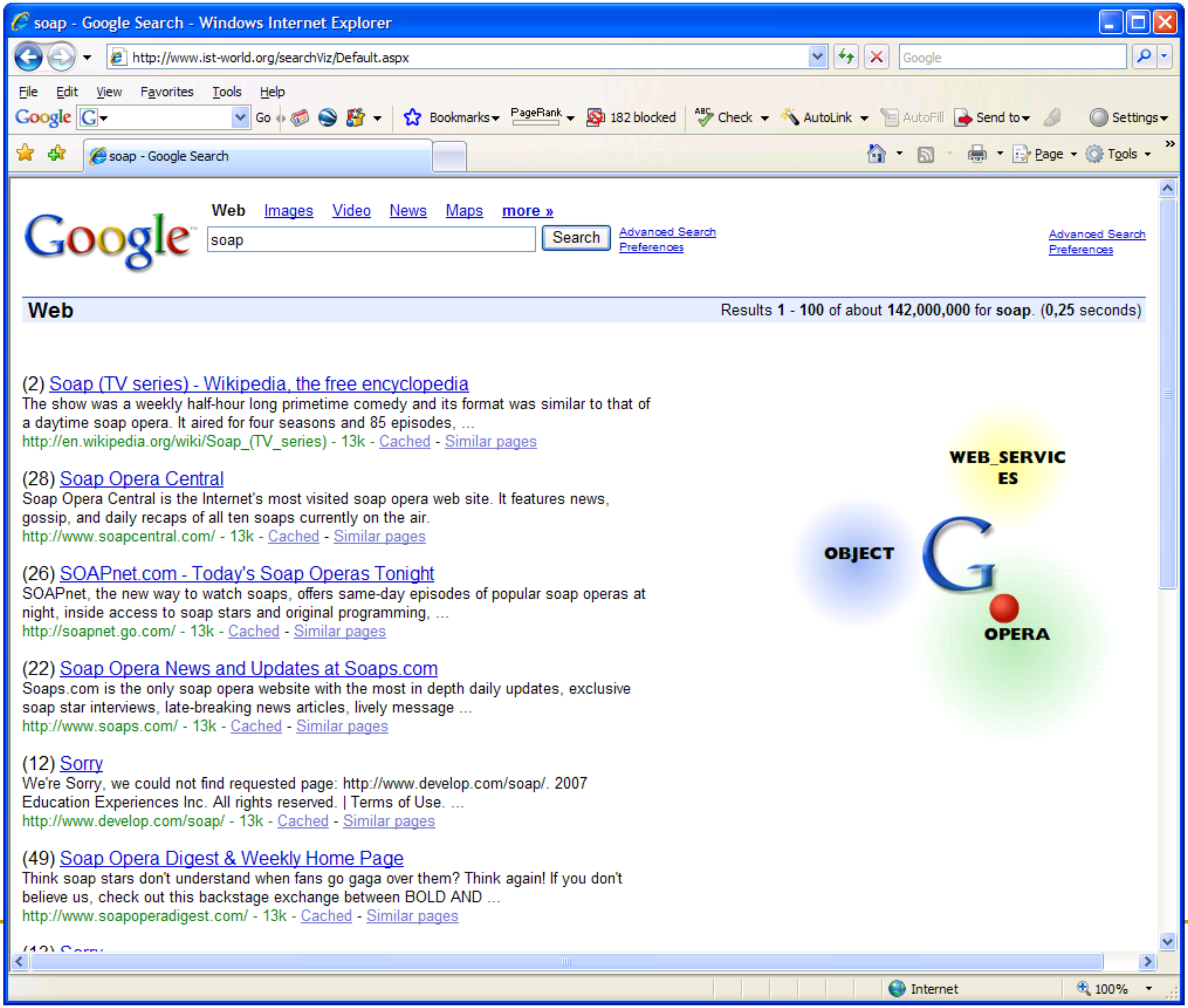
1) Slovenia at Amazon.com
www.amazon.com

2) Sale - "Slovenia" $11.29
www.buy.com

3) Buy "Slovenia" at Barnes & Noble
www.BarnesandNoble.com

http://kartoo.com/

SearchPoint – re-ranking of search results


soap - Google Search - Windows Internet Explorer
http://www.ist-world.org/searchViz/Default.aspx
File Edit View Favorites Tools Help
Bookmarks PageRank 182 blocked Check AutoLink AutoFill Send to Settings
soap - Google Search
Page Tools
Web Images Video News Maps more »
soap
Search
Advanced Search
Preferences
Internet
100%

Web Results **1** - **100** of about **142,000,000** for **soap**. (**0,25** seconds)

(2) Soap (TV series) - Wikipedia, the free encyclopedia
The show was a weekly half-hour long primetime comedy and its format was similar to that of a daytime soap opera. It aired for four seasons and 85 episodes, ...
http://en.wikipedia.org/wiki/Soap_(TV_series) - 13k - Cached - Similar pages

(28) Soap Opera Central
Soap Opera Central is the Internet's most visited soap opera web site. It features news, gossip, and daily recaps of all ten soaps currently on the air.
http://www.soapcentral.com/ - 13k - Cached - Similar pages

(26) SOAPnet.com - Today's Soap Operas Tonight
SOAPnet, the new way to watch soaps, offers same-day episodes of popular soap operas at night, inside access to soap stars and original programming, ...
http://soapnet.go.com/ - 13k - Cached - Similar pages

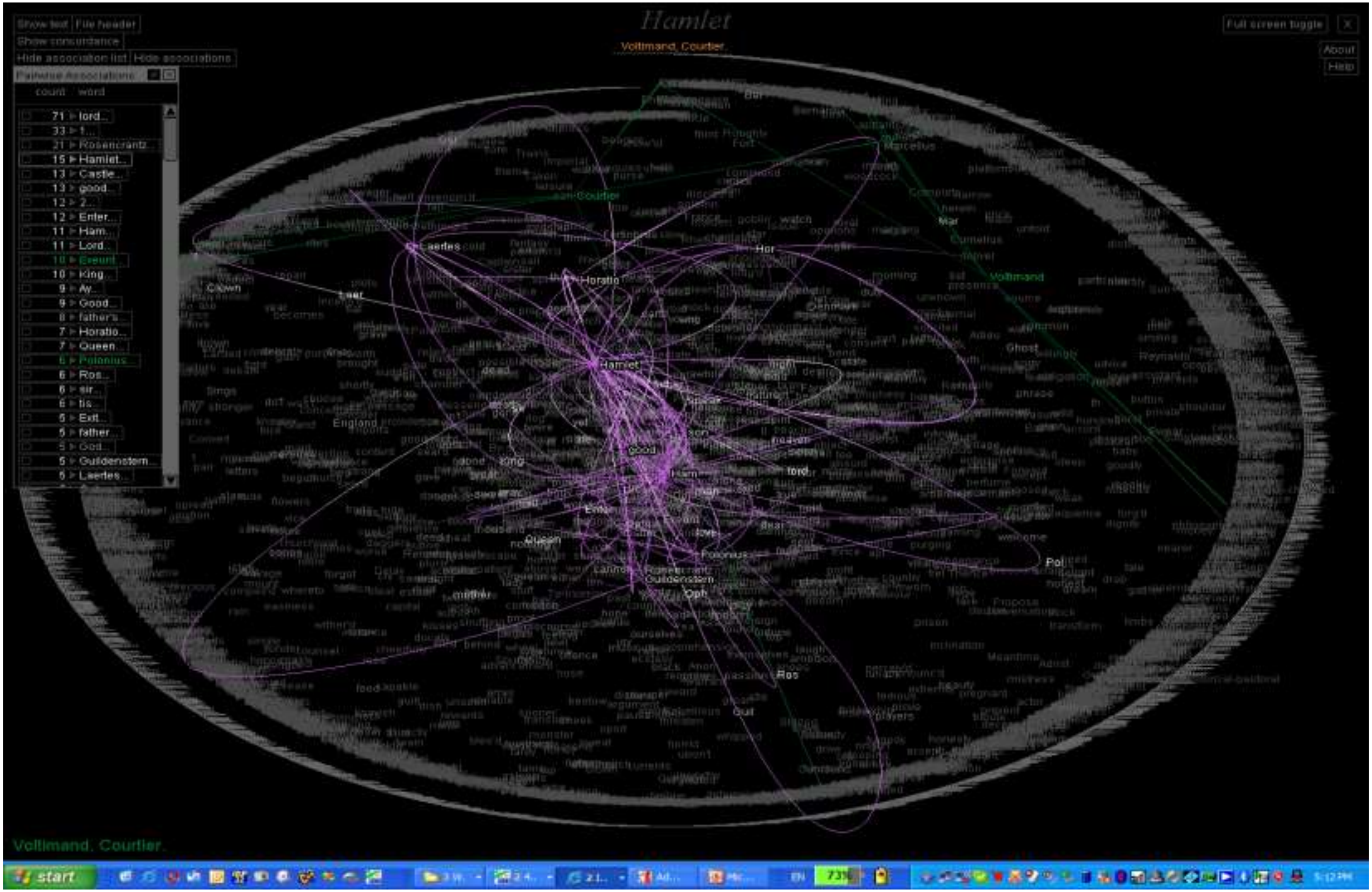
(22) Soap Opera News and Updates at Soaps.com
Soaps.com is the only soap opera website with the most in depth daily updates, exclusive soap star interviews, late-breaking news articles, lively message ...
http://www.soaps.com/ - 13k - Cached - Similar pages

(12) Sorry
We're Sorry, we could not find requested page: http://www.develop.com/soap/. 2007 Education Experiences Inc. All rights reserved. | Terms of Use. ...
http://www.develop.com/soap/ - 13k - Cached - Similar pages

(49) Soap Opera Digest & Weekly Home Page
Think soap stars don't understand when fans go gaga over them? Think again! If you don't believe us, check out this backstage exchange between BOLD AND ...
http://www.soapoperadigest.com/ - 13k - Cached - Similar pages

(13) Sorry

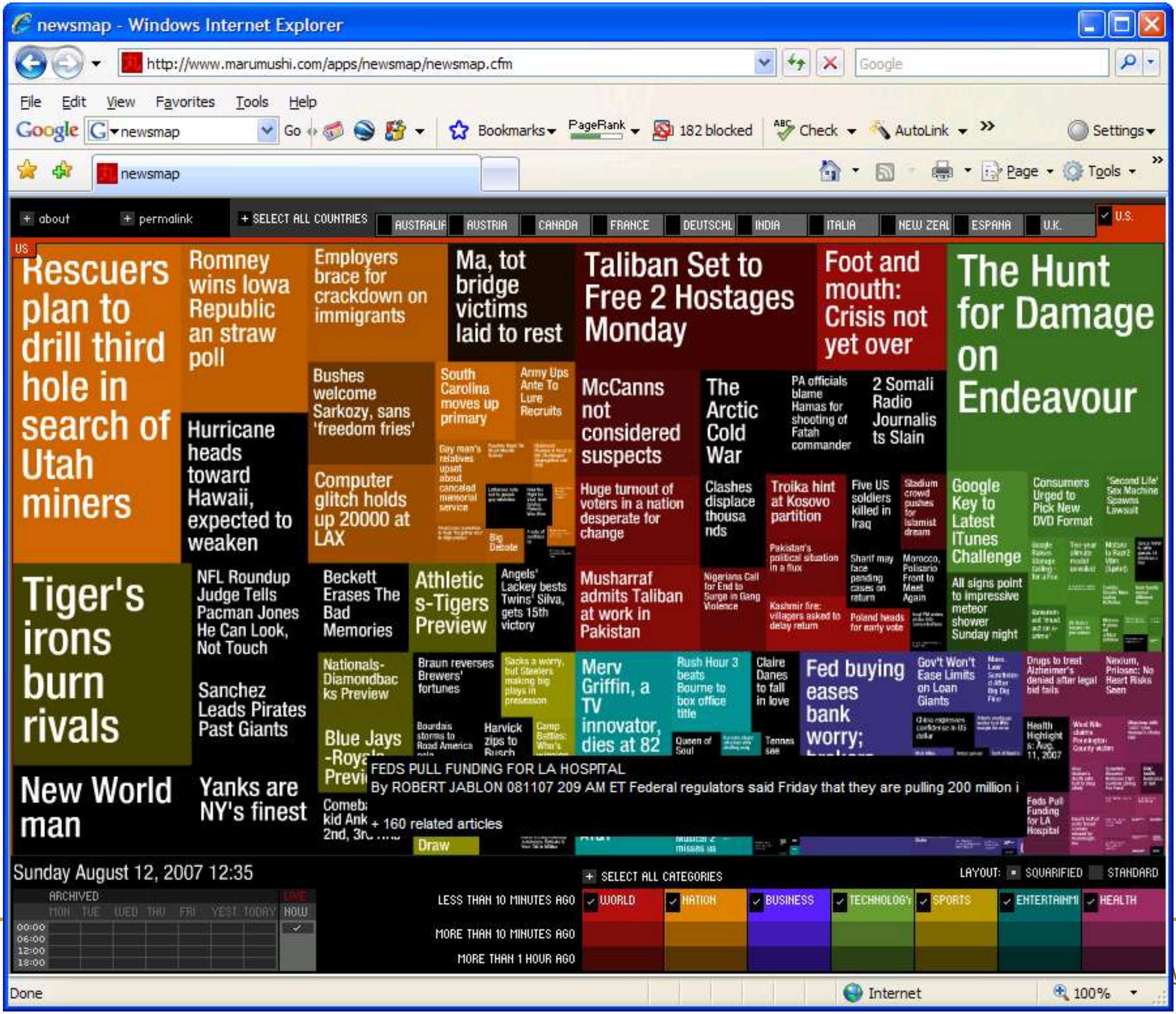
TextArc – visualization of word occurrences



count	word
71	▸ lord...
33	▸ 1...
21	▸ Rosencrantz...
15	▸ Hamlet...
13	▸ Castle...
13	▸ good...
12	▸ 2...
12	▸ Enter...
11	▸ Ham...
11	▸ Lord...
10	▸ Exeunt...
10	▸ King...
9	▸ Ay...
9	▸ Good...
8	▸ father's...
7	▸ Horatio...
7	▸ Queen...
6	▸ Polonius...
6	▸ Ros...
6	▸ sir...
6	▸ tis...
5	▸ Exit...
5	▸ father...
5	▸ God...
5	▸ Guildenstern...
5	▸ Laertes...

http://www.textarc.org/

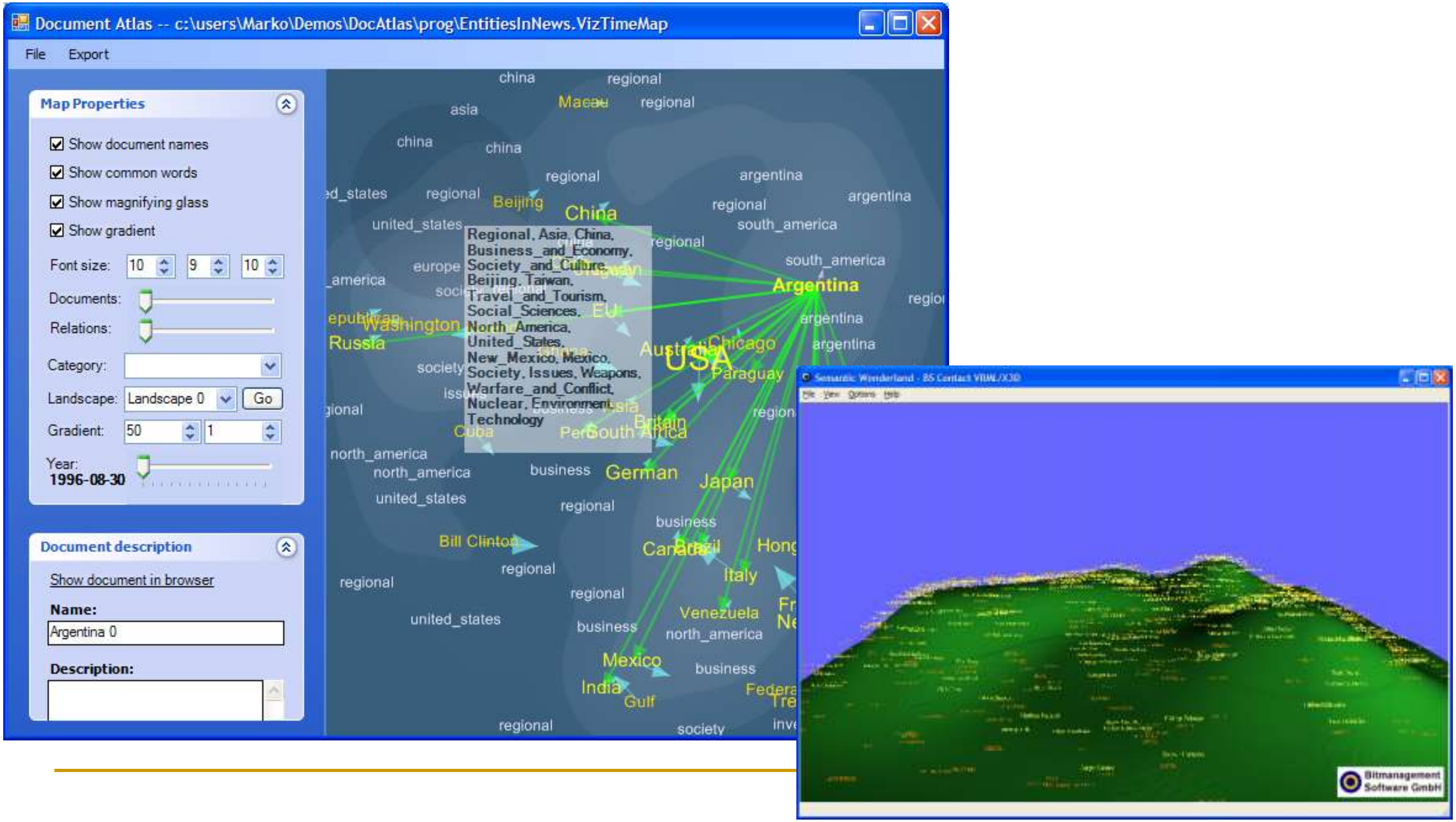
NewsMap – visualization of news articles


newsmap - Windows Internet Explorer
http://www.marumushi.com/apps/newsmap/newsmap.cfm
File Edit View Favorites Tools Help
newsmap
+ about + permalink + SELECT ALL COUNTRIES
AUSTRALIA AUSTRIA CANADA FRANCE DEUTSCHL INDIA ITALIA NEW ZEAL ESPAÑA U.K. ✓ U.S.
US
Rescuers plan to drill third hole in search of Utah miners
Romney wins Iowa Republican straw poll
Employers brace for crackdown on immigrants
Ma, tot bridge victims laid to rest
Taliban Set to Free 2 Hostages Monday
Foot and mouth: Crisis not yet over
The Hunt for Damage on Endeavour
Bushes welcome Sarkozy, sans 'freedom fries'
South Carolina moves up primary
Army Ups Ante To Lure Recruits
McCanns not considered suspects
The Arctic Cold War
PA officials blame Hamas for shooting of Fatah commander
2 Somali Radio Journalists Slain
Hurricane heads toward Hawaii, expected to weaken
Computer glitch holds up 20000 at LAX
Huge turnout of voters in a nation desperate for change
Clashes displace thousands
Troika hint at Kosovo partition
Five US soldiers killed in Iraq
Google Key to Latest ITunes Challenge
Consumers Urged to Pick New DVD Format
'Second Life' Sex Machine Spawns Lawsuit
Tiger's irons burn rivals
NFL Roundup Judge Tells Pacman Jones He Can Look, Not Touch
Beckett Erases The Bad Memories
Athletic s-Tigers Preview
Angels' Lackey bests Twins' Silva, gets 15th victory
Musharraf admits Taliban at work in Pakistan
All signs point to impressive meteor shower Sunday night
Sanchez Leads Pirates Past Giants
Nationals-Diamondbacks Preview
Braun reverses Brewers' fortunes
Merv Griffin, a TV innovator, dies at 82
Rush Hour 3 beats Bourne to box office title
Claire Danes to fall in love
Fed buying eases bank worry;
Gov't Won't Ease Limits on Loan Giants
Drugs to treat Alzheimer's denied after legal bid fails
Nexium, Prilosec: No Heart Risks Seen
Blue Jays -Royals Preview
New World man
Yanks are NY's finest
Feds Pull Funding for LA Hospital
Sunday August 12, 2007 12:35
ARCHIVED LIVE
MON TUE WED THU FRI YEST TODAY NOW
00:00 06:00 12:00 18:00
+ SELECT ALL CATEGORIES
LAYOUT: ■ SQUARIFIED STANDARD
LESS THAN 10 MINUTES AGO
MORE THAN 10 MINUTES AGO
MORE THAN 1 HOUR AGO
✓ WORLD ✓ NATION ✓ BUSINESS ✓ TECHNOLOGY ✓ SPORTS ✓ ENTERTAINMENT ✓ HEALTH
Done Internet 100%

FEDS PULL FUNDING FOR LA HOSPITAL
By ROBERT JABLON 081107 209 AM ET Federal regulators said Friday that they are pulling 200 million i
+ 160 related articles

wsmap.cfm

Document Atlas – visualization of document collections and their structure



Document Atlas -- c:\users\Marko\Demos\DocAtlas\prog\EntitiesInNews.VizTimeMap

Map Properties

- ☑ Show document names
- ☑ Show common words
- ☑ Show magnifying glass
- ☑ Show gradient

Font size: 10 9 10

Documents:

Relations:

Category:

Landscape: Landscape 0 Go

Gradient: 50 1

Year: 1996-08-30

Document description

Show document in browser

Name: Argentina 0

Description:

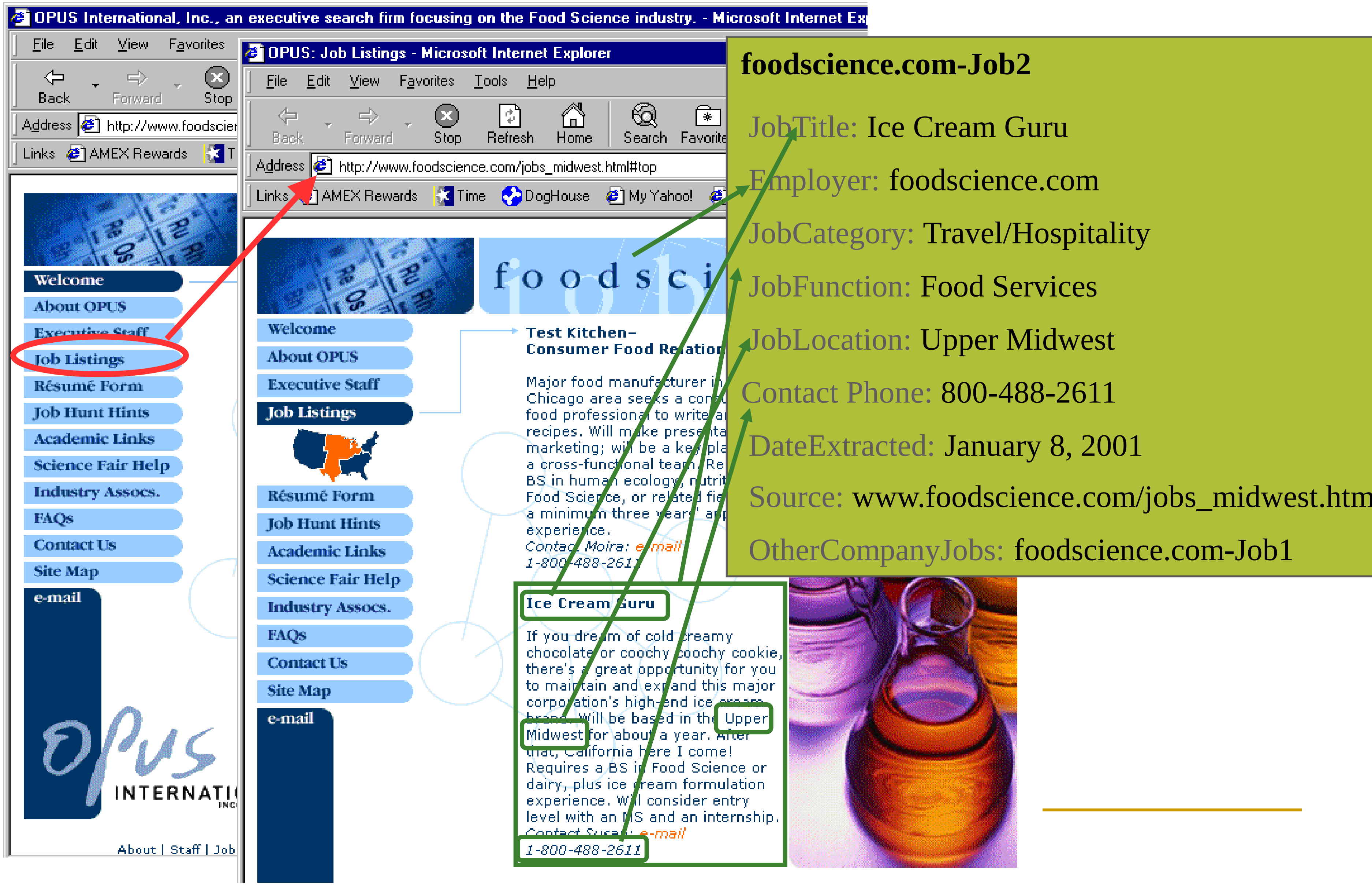
Regional, Asia, China, Business_and_Economy, Society_and_Culture, Beijing, Taiwan, Travel_and_Tourism, Social_Sciences, North_America, United_States, New_Mexico, Mexico, Society, Issues, Weapons, Warfare_and_Conflict, Nuclear, Environment, Technology

http://docatlas.ijs.si

Information Extraction

(slides borrowed from William Cohen’s Tutorial on IE)

Example: Extracting Job Openings from the Web



OPUS International, Inc., an executive search firm focusing on the Food Science industry. - Microsoft Internet Ex

Address http://www.foodscier

Welcome
About OPUS
Executive Staff
Job Listings
Résumé Form
Job Hunt Hints
Academic Links
Science Fair Help
Industry Assocs.
FAQs
Contact Us
Site Map
e-mail

About | Staff | Job

OPUS: Job Listings - Microsoft Internet Explorer

Address http://www.foodscience.com/jobs_midwest.html#top

Links AMEX Rewards Time DogHouse My Yahoo!

Test Kitchen–
Consumer Food Relation

Major food manufacturer i
Chicago area seeks a con
food professional to write a
recipes. Will make presenta
marketing; will be a key pla
a cross-functional team. Re
BS in human ecology, nutrit
Food Science, or related fie
a minimum three years' ap
experience.
Contact Moira: e-mail
1-800-488-2611

Ice Cream Guru

If you dream of cold creamy chocolate or coochy coochy cookie, there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream brand. Will be based in the Upper Midwest for about a year. After that, California here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an MS and an internship.
Contact Susan: e-mail
1-800-488-2611

foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

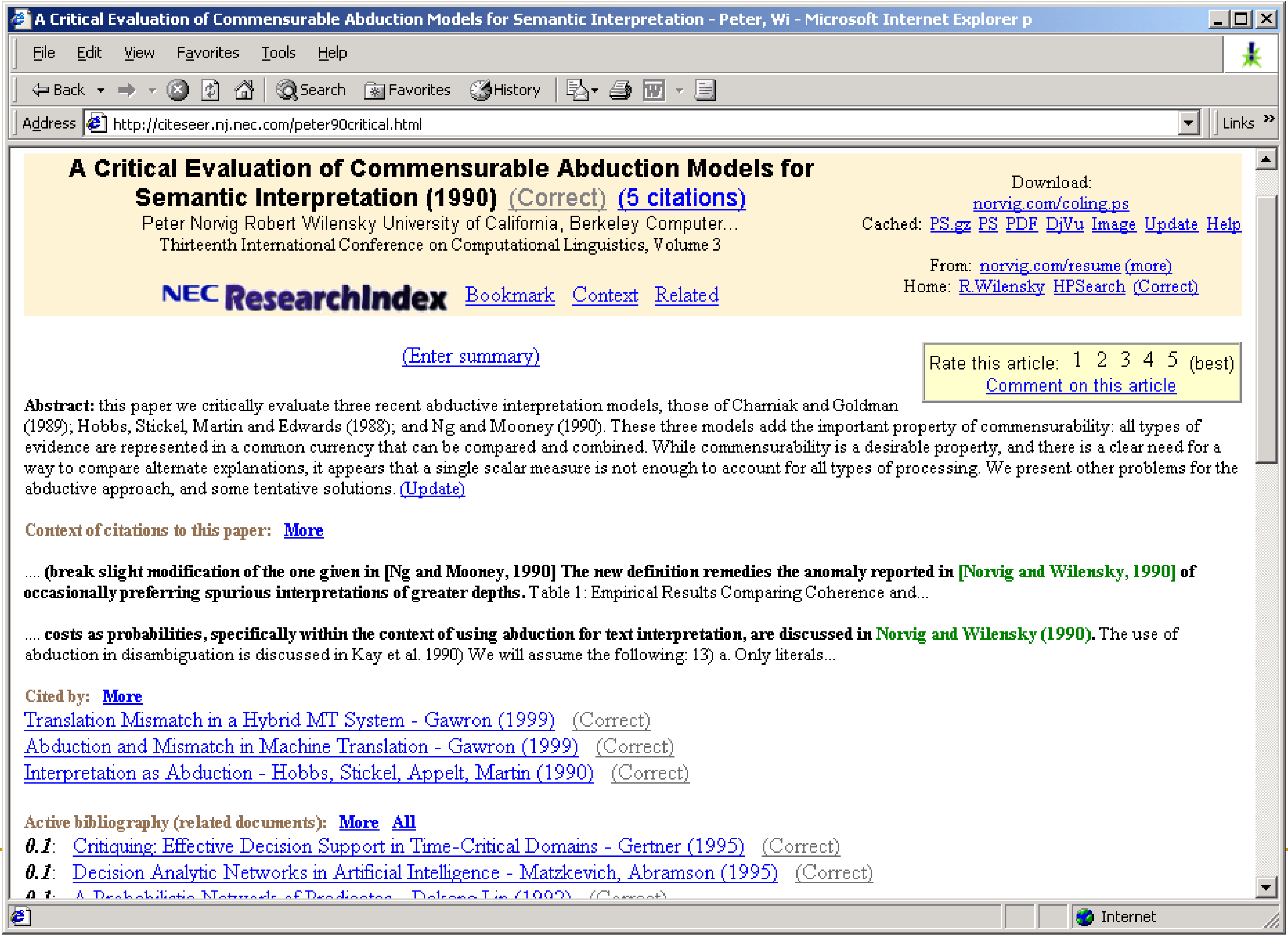
Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1

Example: IE from Research Papers


A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address
http://citeseer.nj.nec.com/peter90critical.html
Links
Internet

A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990) (Correct) (5 citations)

Peter Norvig Robert Wilensky University of California, Berkeley Computer...
Thirteenth International Conference on Computational Linguistics, Volume 3

NEC ResearchIndex Bookmark Context Related

Download:
norvig.com/coling.ps
Cached: PS.gz PS PDF DjVu Image Update Help

From: norvig.com/resume (more)
Home: R.Wilensky HPSearch (Correct)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)
Comment on this article

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: **More**

.... **(break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths.** Table 1: Empirical Results Comparing Coherence and...

.... **costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990).** The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

Cited by: **More**

Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)
Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)
Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (1990) (Correct)

Active bibliography (related documents): **More** **All**

0.1: Critiquing: Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)
0.1: Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson (1995) (Correct)
0.1: A Probabilistic Network of Predicates - Dekang Lin (1992) (Correct)

What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

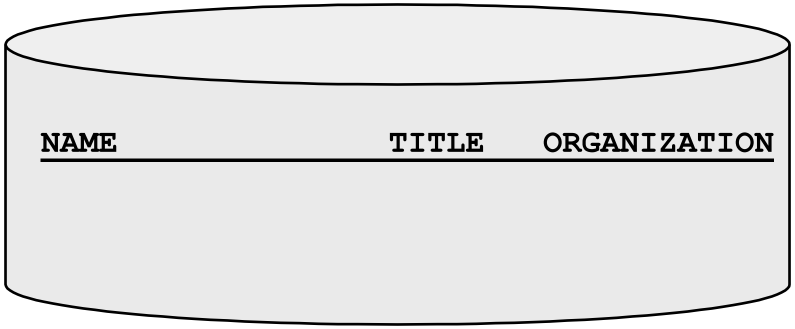
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION

What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

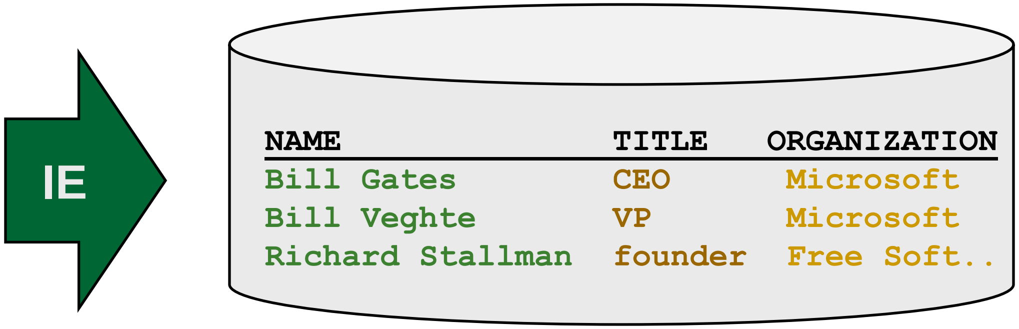
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka “named entity extraction”

What is "Information Extraction"

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is "Information Extraction"

As a family of techniques:

Information Extraction = segmentation + classification + association + clustering

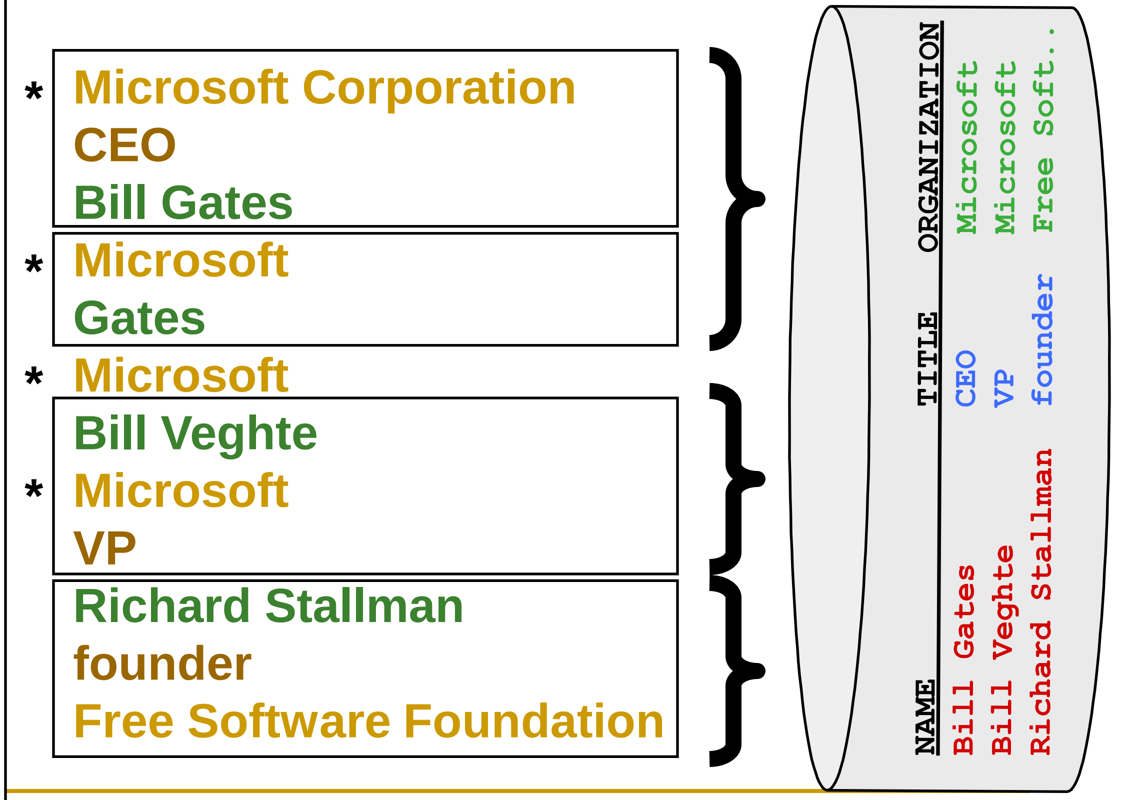
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

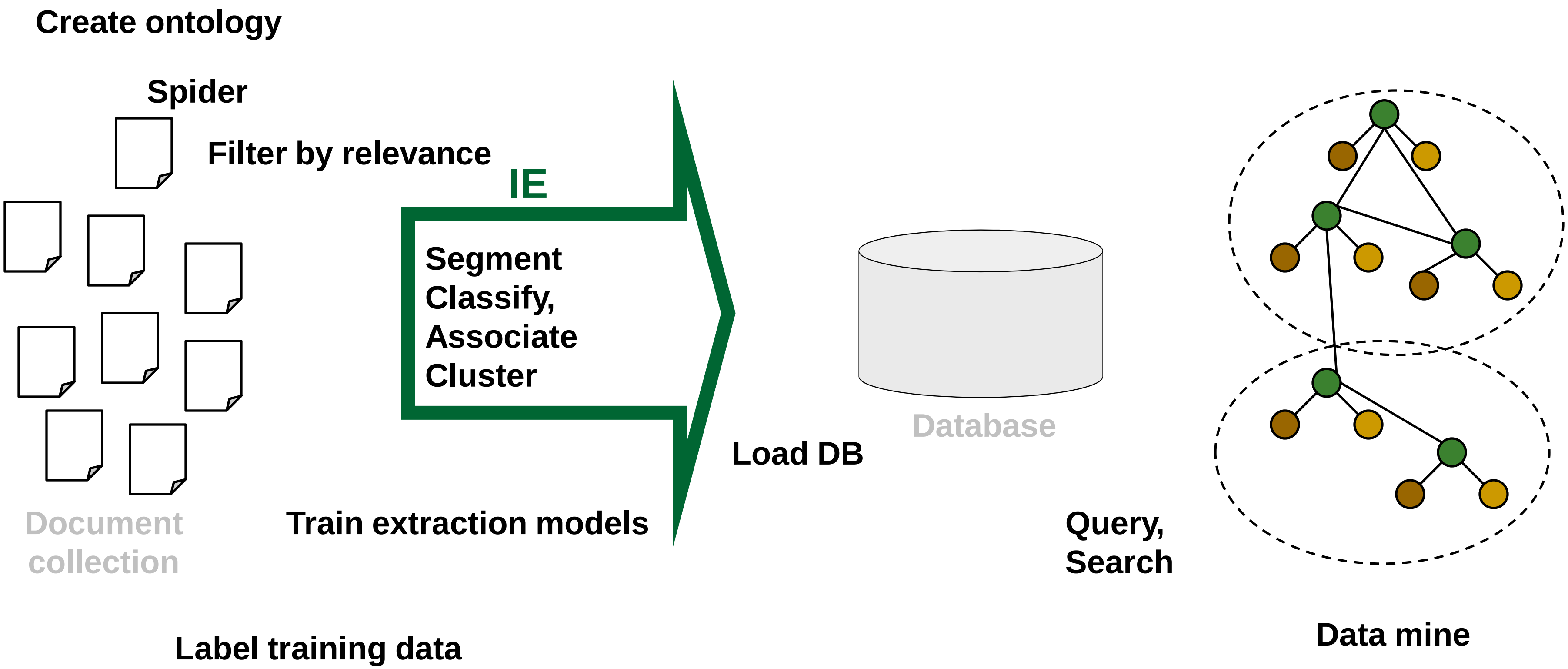
"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

IE in Context



Create ontology

Spider

Filter by relevance

IE

Segment
Classify,
Associate
Cluster

Document
collection

Train extraction models

Load DB

Database

Query,
Search

Label training data

Data mine

Typical approaches to IE

- Hand-built rules/models for extraction
  - …usually extended regexp rules
  - …GATE system from U. Sheffield (http://gate.ac.uk/)
- Machine learning used on manually labelled data:
  - Classification problem on sliding window
    - …examples are taken from sliding window
    - …models classify short segments of text such as title, name, institution, …
    - …limitation of sliding window because it does not take into account sequential nature of text
  - Training stochastic finite state machines (e.g. HMM)
    - …probabilistic reconstruction of parsing sequence

Link-Analysis

How to analyze graphs in the Web context?

What is Link Analysis?

- Link Analysis is exploring associations between the objects
  - …most characteristic for the area is **graph** representation of the data
  - Category of graphs which attract recently the most interest are the ones which are generated by some social process (**social networks**) – this would include web

- Synonyms for **Link Analysis** or at least very related areas are **Graph Mining**, **Network Analysis**, **Social Network Analysis**

- In the next slides we'll present some of the typical definitions, ideas and algorithms

What is Power Law?

- Power law describes relations between the objects in the network
  - …it is very characteristic for the networks generated within some kind of social process
  - …it describes **scale invariance** found in many natural phenomena (including physics, biology, sociology, economy and linguistics)

- In Link Analysis we usually deal with power law distributed graphs

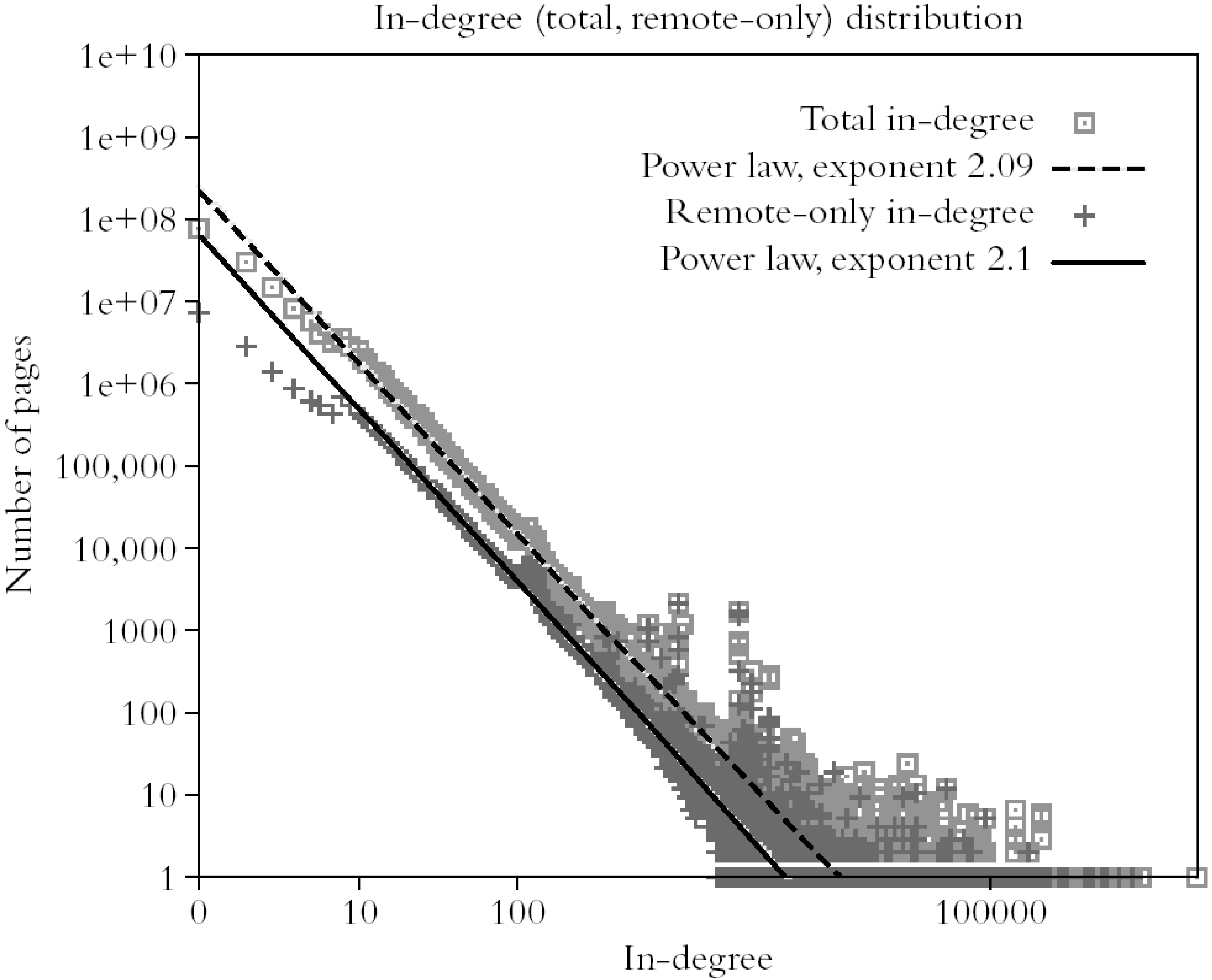
Power-Law on the Web

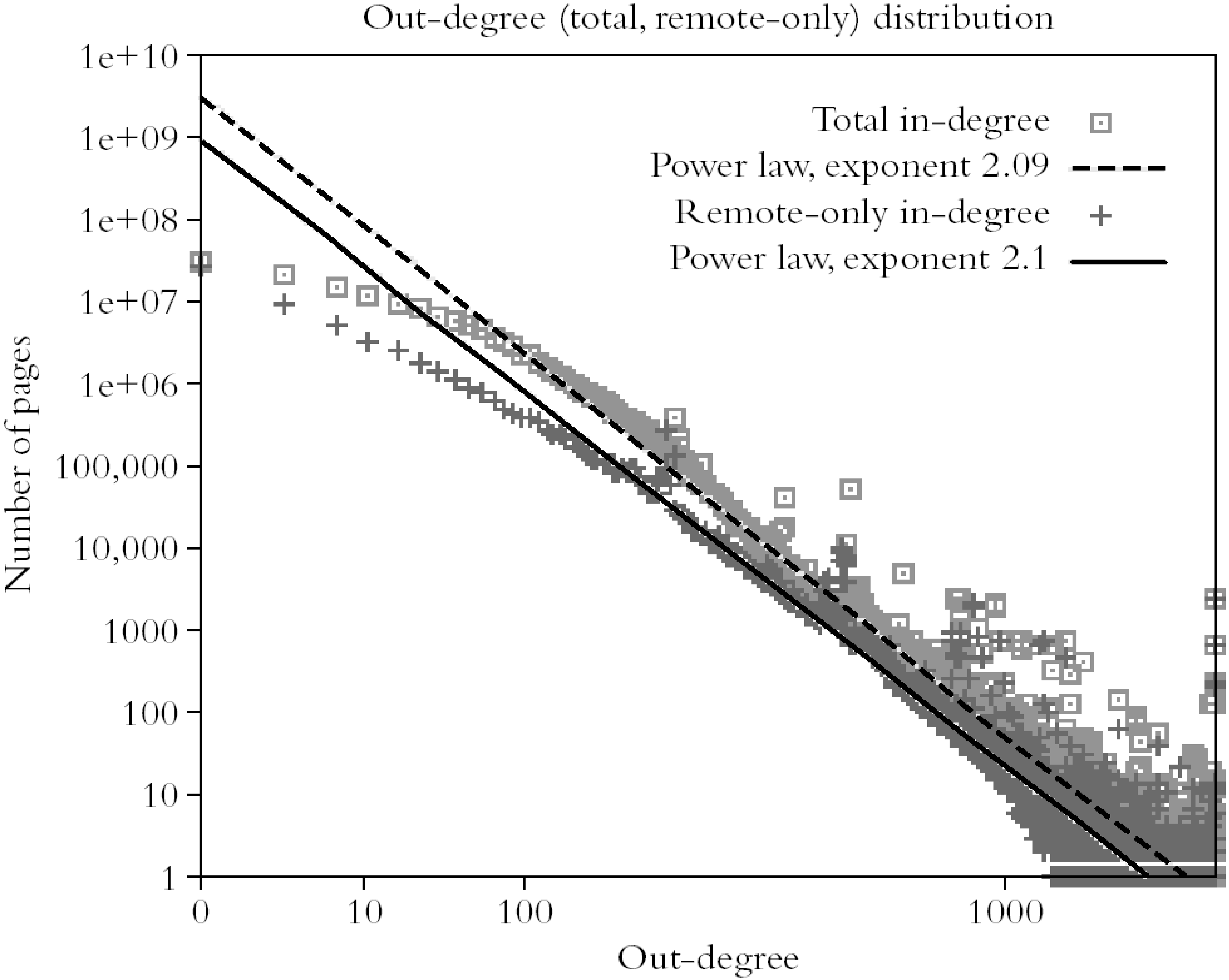
- In the context of Web the power-law appears in many cases:
  - Web pages sizes
  - Web page connectivity
  - Web connected components' size
  - Web page access statistics
  - Web Browsing behavior
- Formally, power law describing web page degrees are:

$$\Pr(\text{out-degree is } k) \propto 1/k^{a_{\text{out}}}$$

$$\Pr(\text{in-degree is } k) \propto 1/k^{a_{\text{in}}}$$

(This property has been preserved as the Web has grown)


In-degree (total, remote-only) distribution
Total in-degree ◻
Power law, exponent 2.09
Remote-only in-degree +
Power law, exponent 2.1
1e+10
1e+09
1e+08
1e+07
1e+06
100,000
10,000
1000
100
10
1
Number of pages
0
10
100
100000
In-degree


Out-degree (total, remote-only) distribution
Total in-degree
Power law, exponent 2.09
Remote-only in-degree
Power law, exponent 2.1
Number of pages
1e+10
1e+09
1e+08
1e+07
1e+06
100,000
10,000
1000
100
10
1
0
10
100
1000
Out-degree

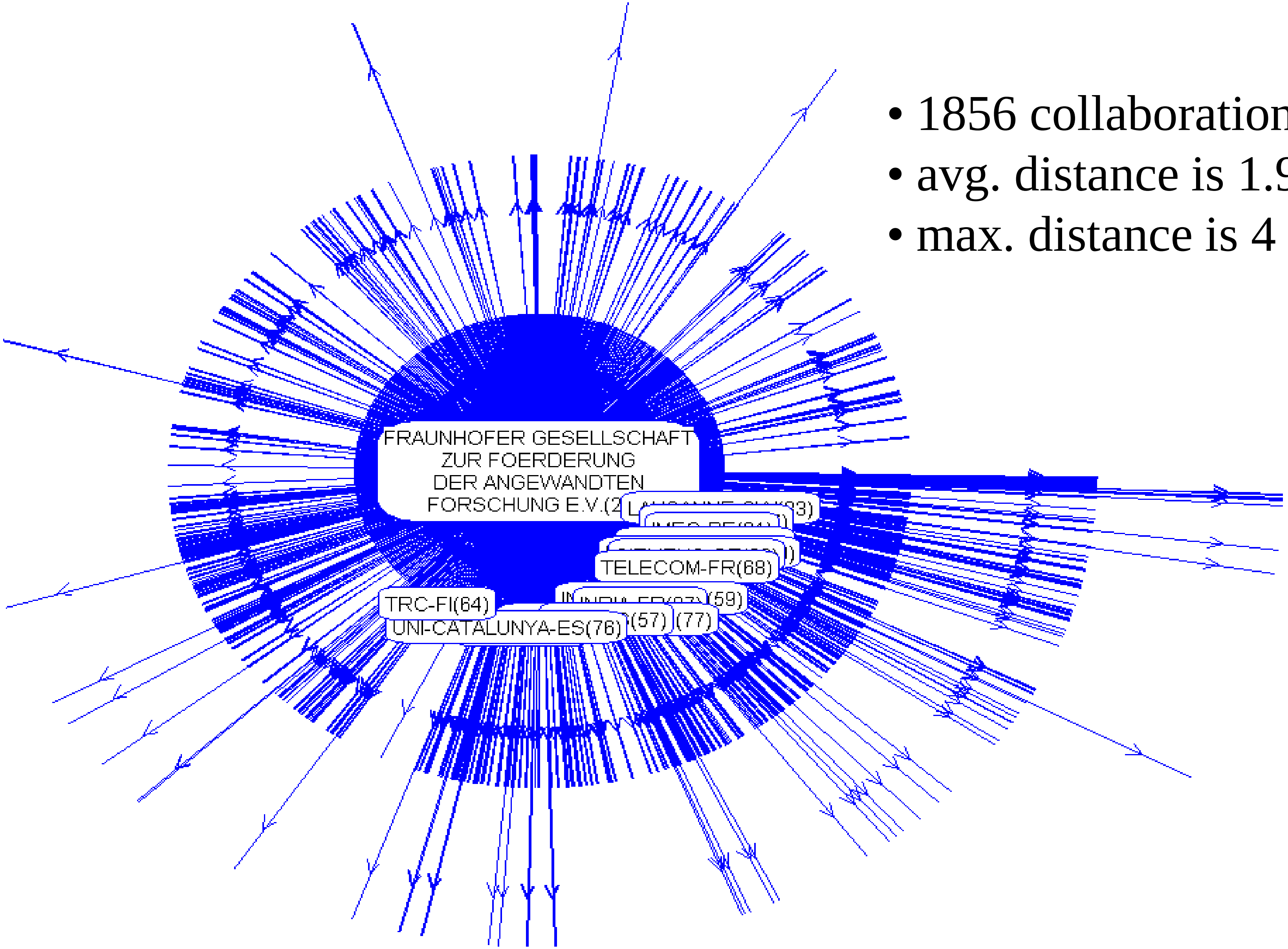
Small World Networks

- Empirical observation for the Web-Graph is that the diameter of the Web-Graph is small relative to the size of the network
  - …this property is called “Small World”
  - …formally, small-world networks have diameter exponentially smaller then the size
- By simulation it was shown that for the Web-size of 1B pages the diameter is approx. 19 steps
  - …empirical studies confirmed the findings

Example of Small World: project collaboration network

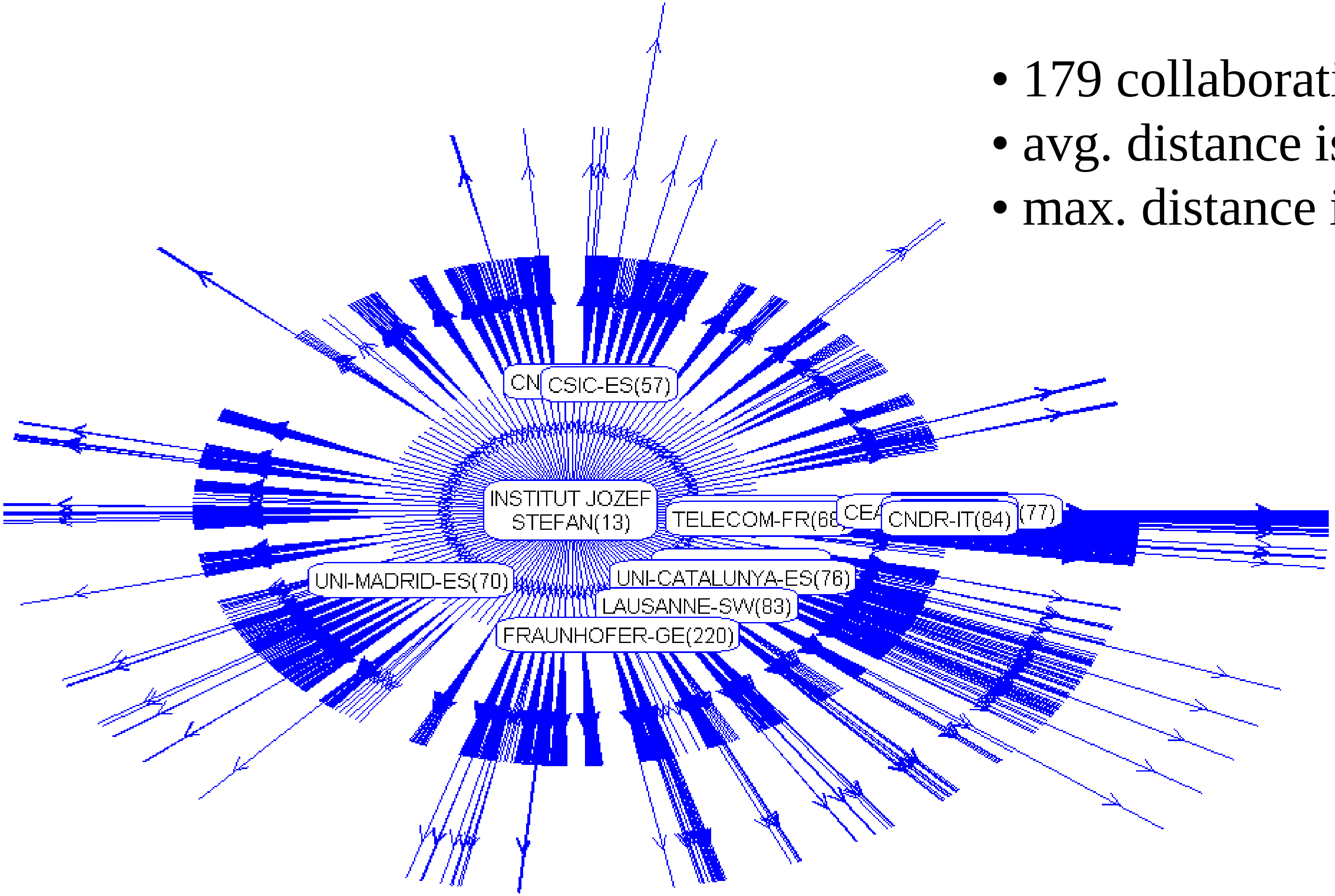
- The network represents collaboration between institutions on projects funded by European Union
  - …there are 7886 organizations collaborating on 2786 projects
  - …in the network, each node is an organization, two organizations are connected if they collaborate on at least one project
- Small world properties of the collaboration network:
  - **Main connected part** of the network contains 94% of the nodes
  - **Max distance** between any two organizations is 7 steps … meaning that any organization can be reached in up to 7 steps from any other organization
  - **Average distance** between any two organizations is 3.15 steps (with standard deviation 0.38)
  - 38% (2770) of organizations have avg. distance 3 or less

Connectedness of the most connected institution


FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.(2
TELECOM-FR(68)
TRC-FI(64)
UNI-CATALUNYA-ES(76)
(57)
(77)
(59)
3)

- 1856 collaborations
- avg. distance is 1.95
- max. distance is 4

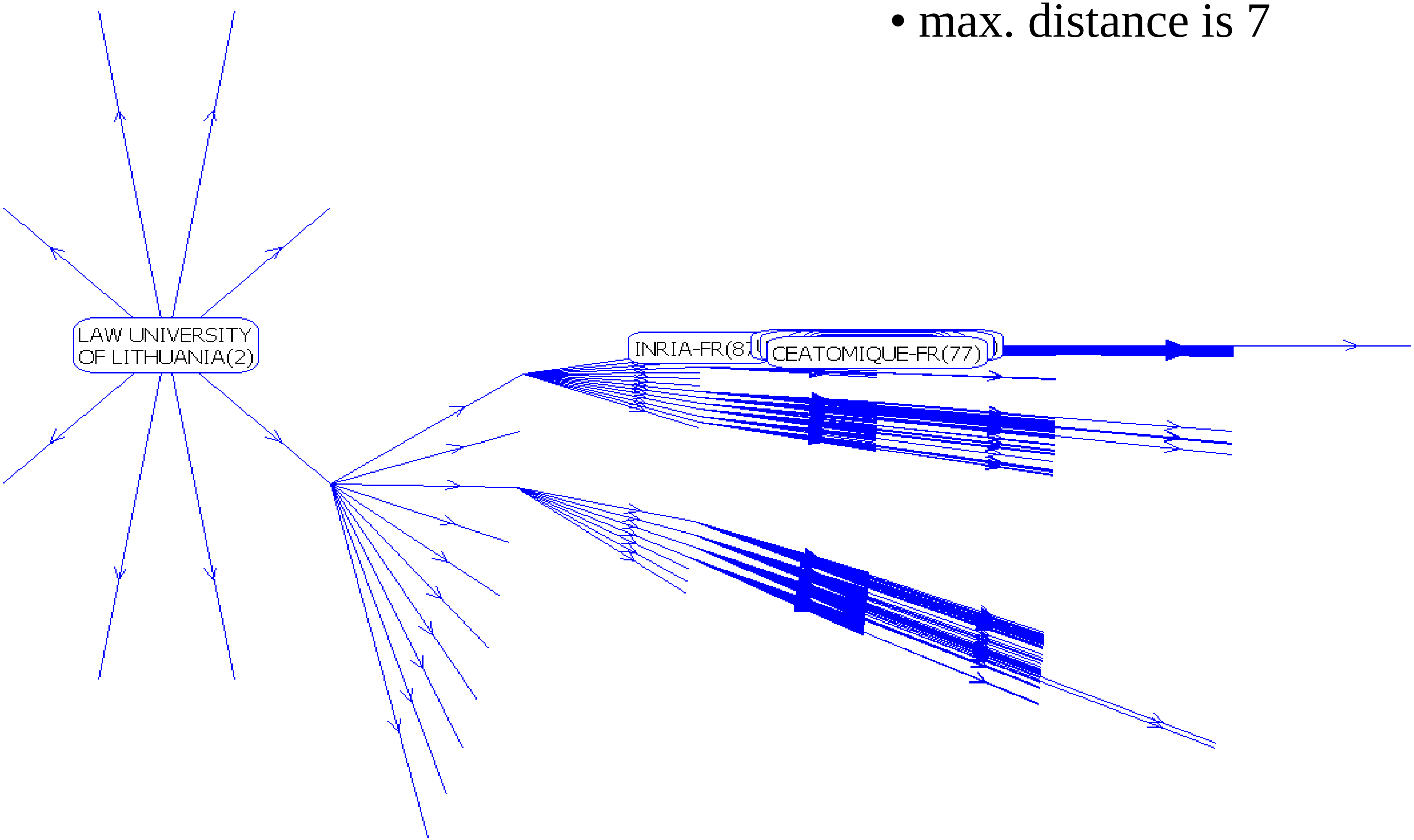
Connectedness of semi connected institution



- 179 collaborations
- avg. distance is 2.42
- max. distance is 4

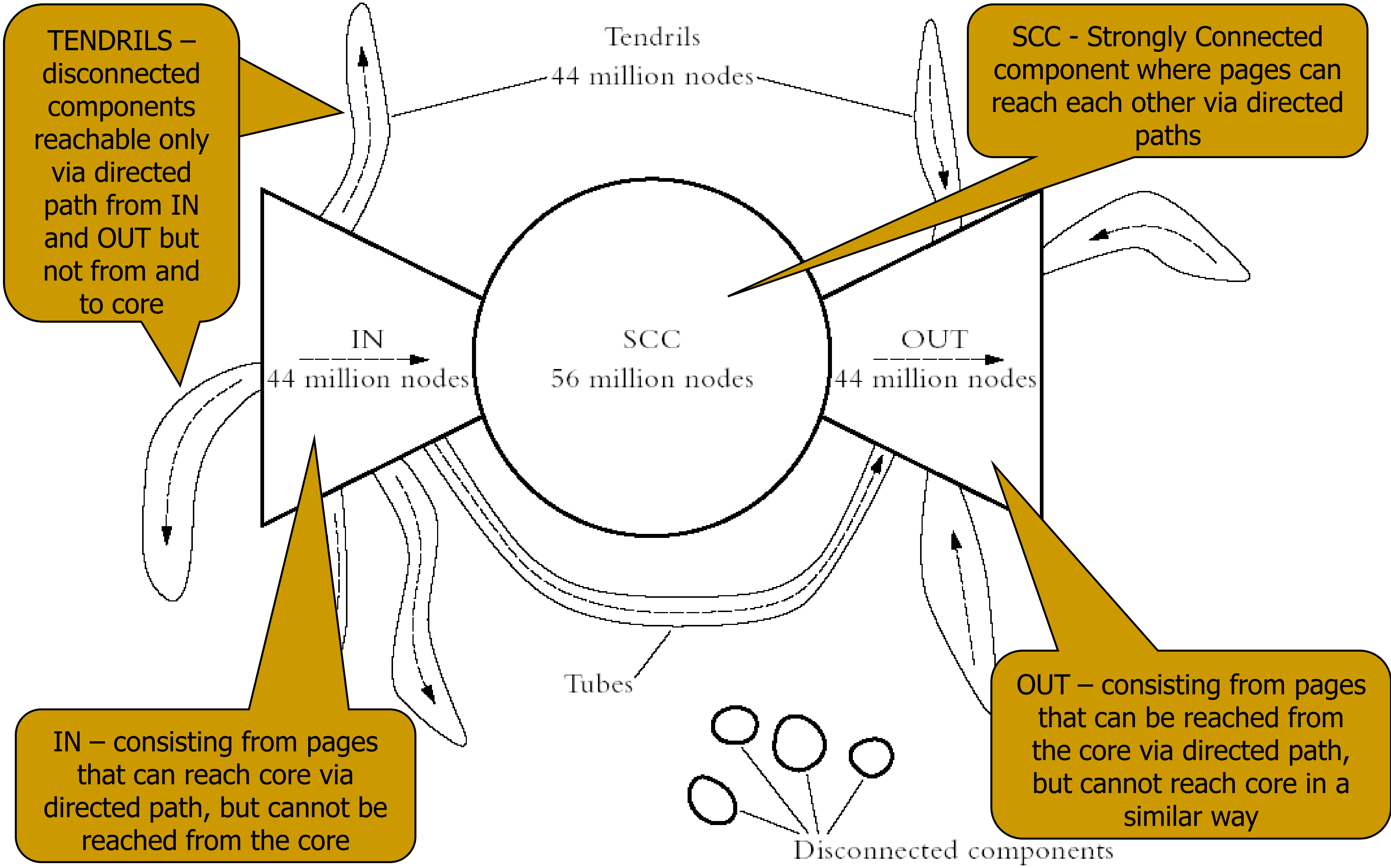
Connectedness of min. connected institution

- 8 collaborations
- max. distance is 7



Structure of the Web – “Bow Tie” model

- In November 1999 large scale study using AltaVista crawls in the size of over 200M nodes and 1.5B links reported “bow tie” structure of web links
  - …we suspect, because of the scale free nature of the Web, this structure is still preserved



TENDRILS – disconnected components reachable only via directed path from IN and OUT but not from and to core

SCC - Strongly Connected component where pages can reach each other via directed paths

IN – consisting from pages that can reach core via directed path, but cannot be reached from the core

OUT – consisting from pages that can be reached from the core via directed path, but cannot reach core in a similar way

Region:	SCC	IN	OUT	Tendrils	Disconnected	Total
Size:	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

Modeling the Web Growth

- Links/Edges in the Web-Graph are not created at random
  - …probability that a new page gets attached to one of the more popular pages is higher then to a one of the less popular pages
  - Intuition: “rich gets richer” or “winners takes all”
  - Simple algorithm “Preferential Attachment Model” (Barabasi, Albert) efficiently simulates Web-Growth

“Preferential Attachment Model” Algorithm

- $\mathbf{M_0}$ vertices (pages) at time 0
- At each time step new vertex (page) is generated with $\mathbf{m \leq M_0}$ edges to **m** random vertices
  - …probability for selection a vertex for the edge is proportional to its degree
- …after **t** time steps, the network has $\mathbf{M_0+t}$ vertices (pages) and **mt** edges
  - …probability that a vertex has connectivity **k** follows the power-law

Estimating importance of the web pages

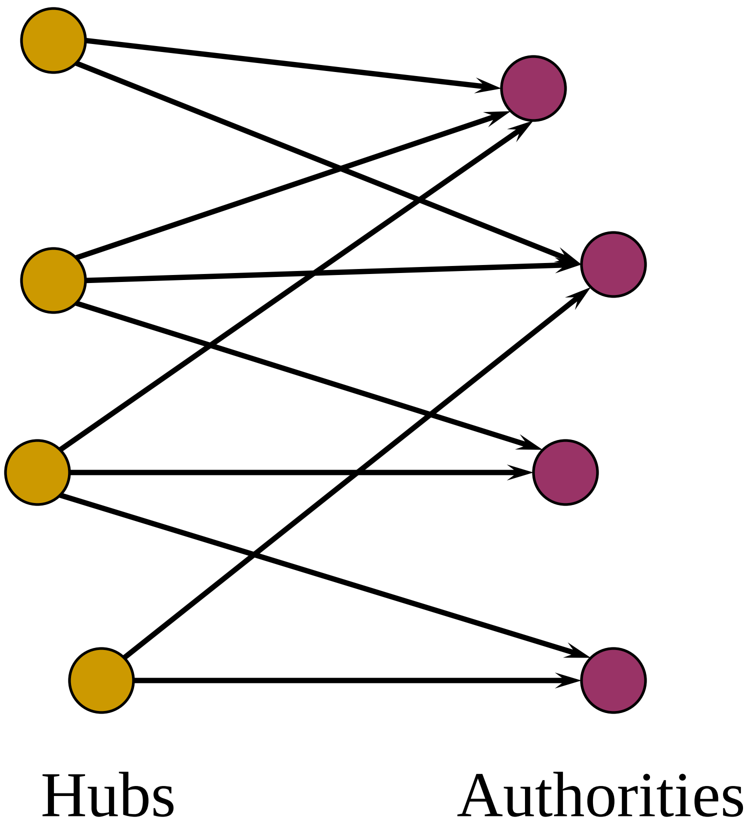
- Two main approaches, both based on eigenvector decomposition of the graph adjacency matrix
    - Hubs and Authorities (HITS)
    - PageRank – used by Google

Hubs and Authorities

- Intuition behind HITS is that each web page has two natures:
    - …being good content page (authority weight)
    - …being good hub (hub weight)
- …and the idea behind the algorithm:
    - …good authority page is pointed to by good hub pages
    - …good hub page is pointing to good authority pages

Hubs and Authorities

(Kleinberg 1998)



"Hubs and authorities exhibit what could be called a *mutually reinforcing* relationship"

Iterative relaxation:

$$\mathrm{Hub}(p) = \sum_{q:p\to q} \mathrm{Authority}(q)$$

$$\mathrm{Authority}(p) = \sum_{q:q\to p} \mathrm{Hub}(q)$$

Semantic-Web

How semantics fits into the picture?

What is Semantic Web? (informal)

- Informal statements:
  - “…if the ordinary web is mainly for **computer-to-human** communication, then the semantic web aims primarily at **computer-to-computer** communication
  - The idea is to establish infrastructure for dealing with common vocabularies
  - The goal is to overcome surface syntax representation of the data and deal with the “semantics” of the data
    - …as an example, one should be able to make a “semantic link” from a database column with the name “ZIP-Code” and a GUI form with a “ZIP” field since they actually mean the same – they both describe the same abstract concept
  - **Semantic Web is mainly about integration and standards!**

What is Semantic Web? (formal)

- Formal statement (from http://www.w3.org/2001/sw/):
    - “The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries.”
    - “It is a collaborative effort led by **W3C** with participation from a large number of researchers and industrial partners.”

What is the link between Text-Mining, Link Analysis and Semantic Web?

- **Text-Mining, Link-Analysis** and other analytic techniques deal mainly with extracting and aggregating the information from raw data
  - …they maximize the quality of extracted information
- **Semantic Web**, on the other hand, deals mainly with the integration and representation of the given data
  - …it maximizes reusability of the given information

- **Both areas** are very much complementary and necessary for operational information engineering

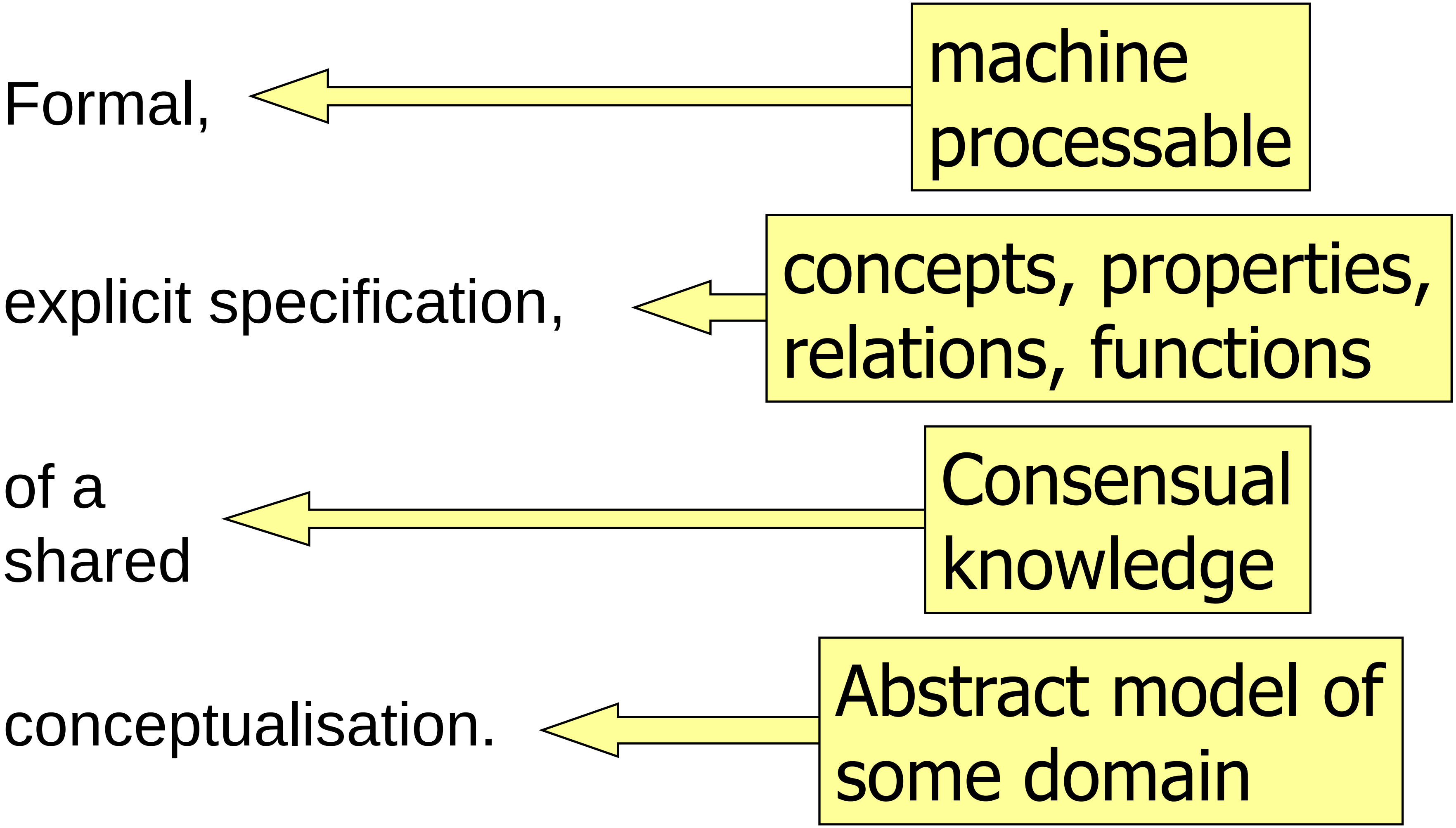
Semantic Web

Ontologies
(formalization of semantics)

Ontologies – central objects in SW

- Ontologies are central formal objects within Semantic Web
  - Ontologies have origin in philosophy, but within computer science they represent a **data model** that represents a domain and is used to **reason** about the objects in that domain and the **relations** between them
  - …their main aim is to describe and represent an area of **knowledge in a formal way**
  - Most of the Semantic Web standards/languages (XML, RDF, OWL) are concerned with some level of ontological representation of the knowledge

What is an ontology?



Formal, ← machine processable

explicit specification, ← concepts, properties, relations, functions

of a shared ← Consensual knowledge

conceptualisation. ← Abstract model of some domain

Frank.van Harmelen 2003: http://seminars.ijs.si/sekt

Which elements represent an ontology?

- An ontology typically consists of the following elements:
  - **Instances** – the basic or “ground level” objects
  - **Classes** – sets, collections, or types of objects
  - **Attributes** – properties, features, characteristics, or parameters that objects can have and share
  - **Relations** – ways that objects can be related to one another

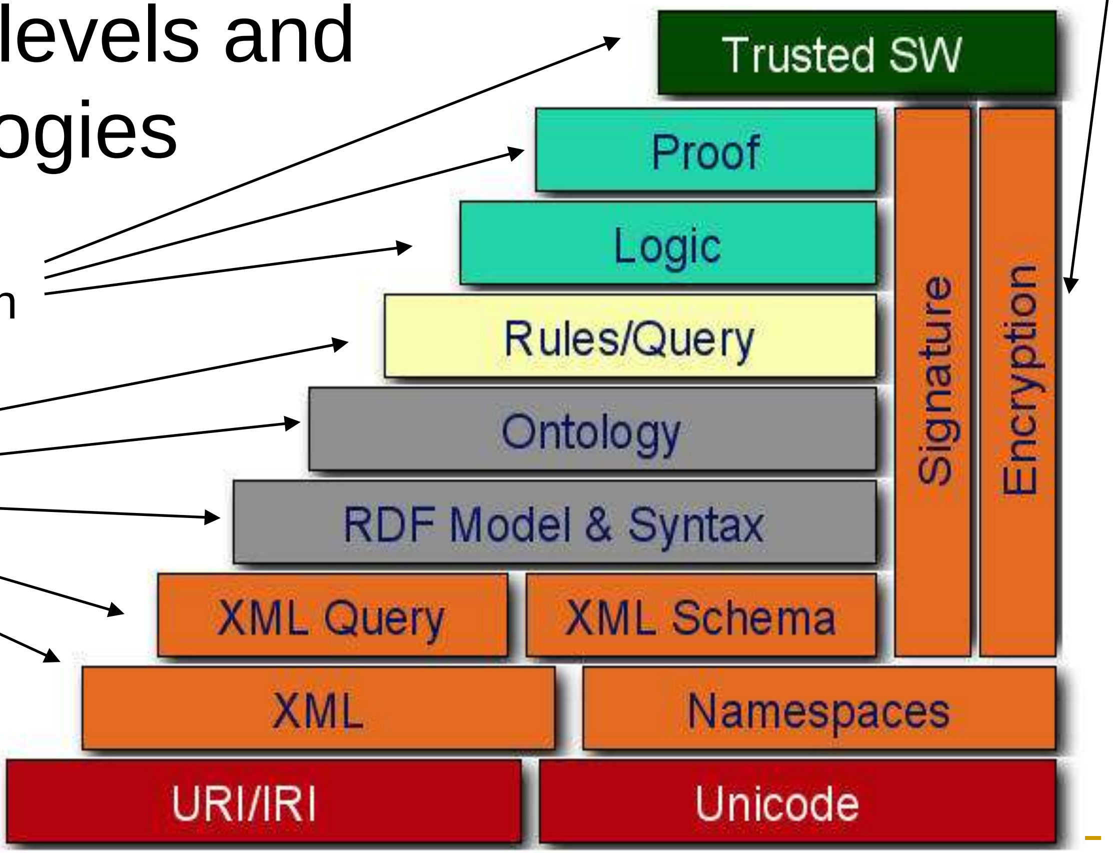
- Analogies between *ontologies* and *relational databases*:
  - **Instances** correspond to **records**
  - **Classes** correspond to **tables**
  - **Attributes** correspond to **record fields**
  - **Relations** correspond to **relations between the tables**

Semantic Web Languages (XML, RDF, OWL)

Which levels Semantic Web is dealing with?

- The famous "**Semantic Web Layer Cake**" shows representation levels and related technologies

Infrastructure



Higher level of representation and reasoning

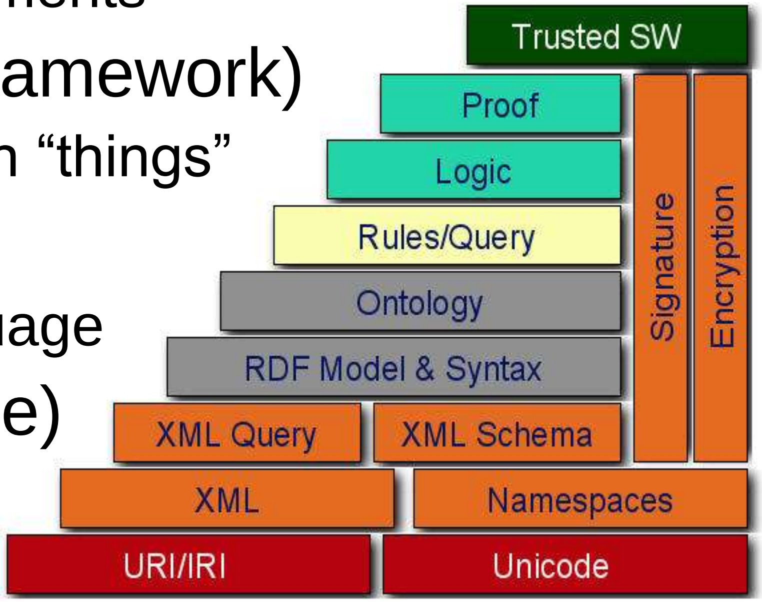
Different Levels of Semantic Abstraction

Addressing the information

Character Level Encoding

Stack of Semantic Web Languages

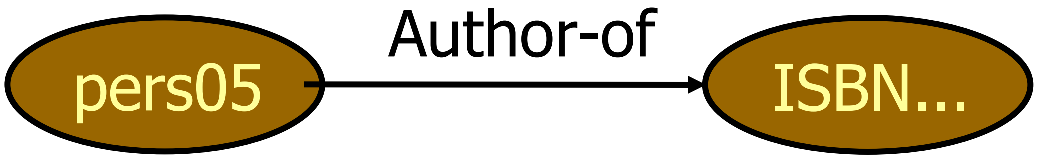
- **XML** (eXtended Markup Language)
  - Surface syntax, no semantics
- **XML Schema**
  - Describes structure of XML documents
- **RDF** (Resource Description Framework)
  - Datamodel for “relations” between “things”
- **RDF Schema**
  - RDF Vocabulary Definition Language
- **OWL** (Web Ontology Language)
  - A more expressive
    Vocabulary Definition Language



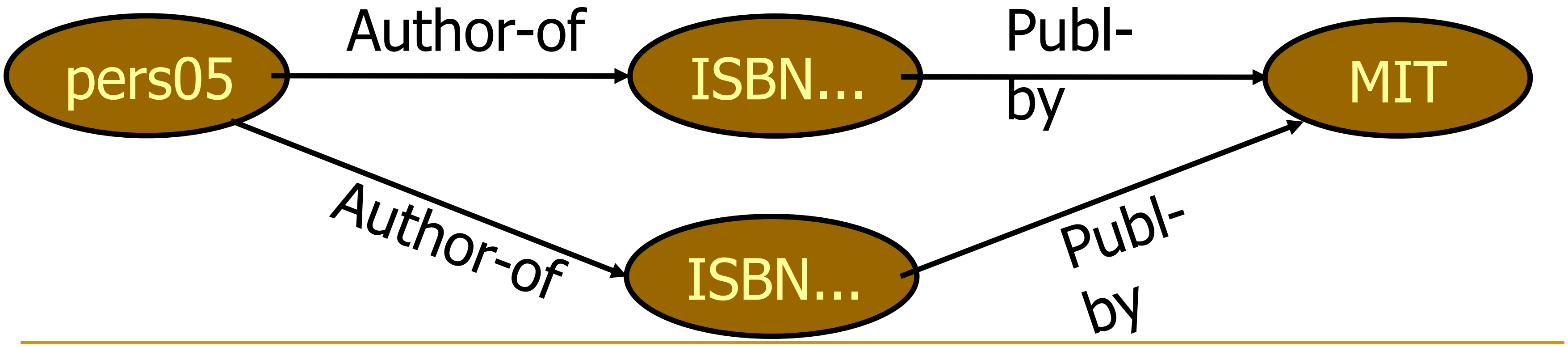
Frank.van Harmelen 2003: http://seminars.ijs.si/sekt

Bluffer's guide to RDF (1/2)

- **Object ->Attribute-> Value** triples



- objects are **web-resources**
- Value is again an Object:
  - triples can be **linked**
  - data-model = graph



Bluffer's guide to RDF (2/2)

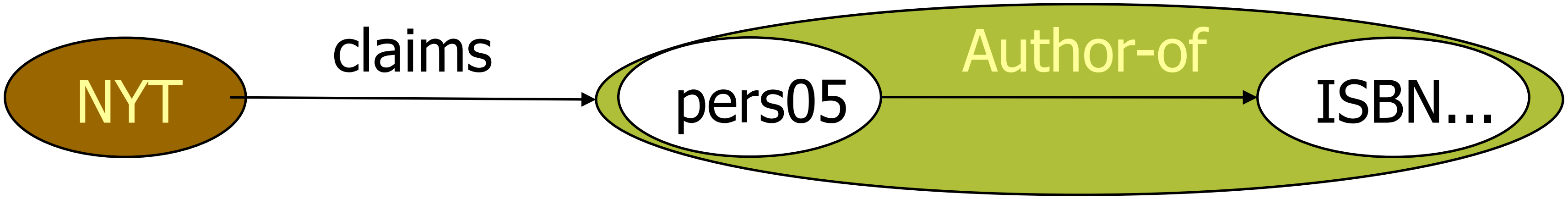
- Every identifier is a URL

  = world-wide unique naming!

- Has XML syntax

```
<rdf:Description rdf:about="#pers05">
   <authorOf>ISBN...</authorOf>
</rdf:Description>
```

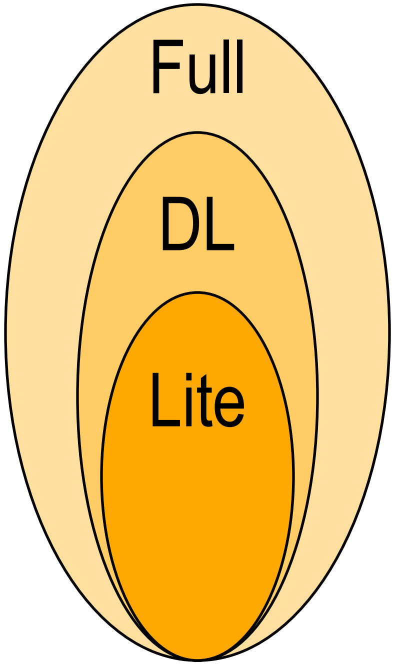
- Any statement can be an object
  - …graphs can be **nested**



Frank.van Harmelen 2003: http://seminars.ijs.si/sekt

OWL Layers

- **OWL Lite:**
  - Classification hierarchy
  - Simple constraints
- **OWL DL:**
  - Maximal expressiveness
  - While maintaining tractability
  - Standard formalisation
- **OWL Full:**
  - Very high expressiveness
  - Loosing tractability
  - Non-standard formalisation
  - All syntactic freedom of RDF (self-modifying)



Frank.van Harmelen 2003: http://seminars.ijs.si/sekt

Semantic Web

OntoGen system (example of ontology learning)

Ontology learning

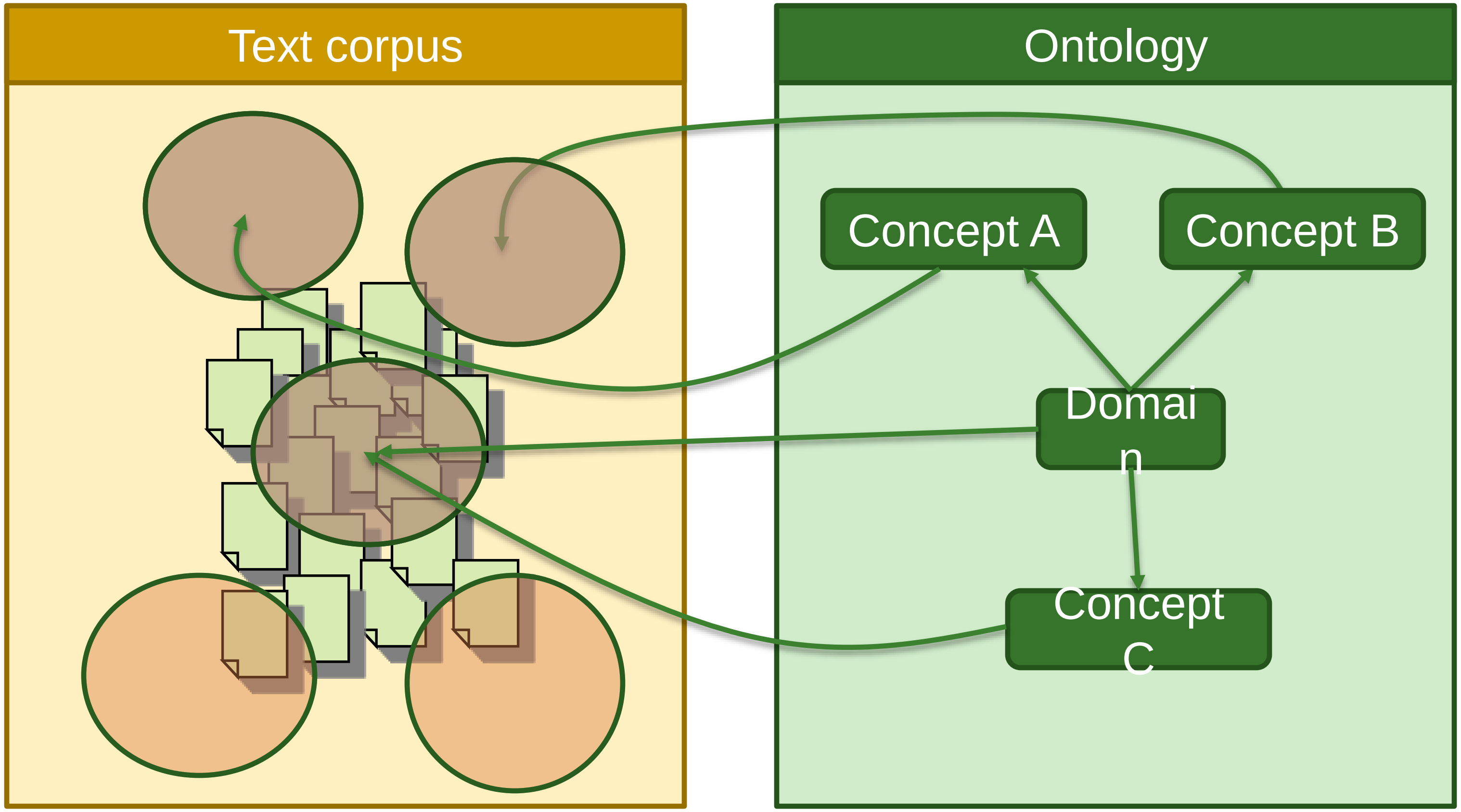
- Ontology learning task aims at extracting structure in the given data and save the structure in the form of an ontology
- Two systems for ontology learning from documents:
  - OntoGen (http://ontogen.ijs.si)
    - …extracts the structure by using machine learning techniques (clustering, active learning, visualization, …)
  - Text2Onto (http://ontoware.org/projects/text2onto/)
    - …extracts the structure from text by using linguistic patterns

OntoGen – main scenarios using

- Given a corpus of documents a user can interactively…
  - …construct new classes by
    - …clustering of documents into topics and subtopics
    - …active learning when user wants to extract structure
    - …selecting data on visualized map of documents
    - …mapping proposed concepts to existing ontologies
  - …populate new documents into an ontology by
    - …by categorization of documents into hierarchy
  - …summarize ontology by
    - …keyword extraction techniques
    - …visualization of the structure
  - …save constructed ontology as
    - Semantic Web formalism (RDF, OWL, Prolog)
    - statistical model

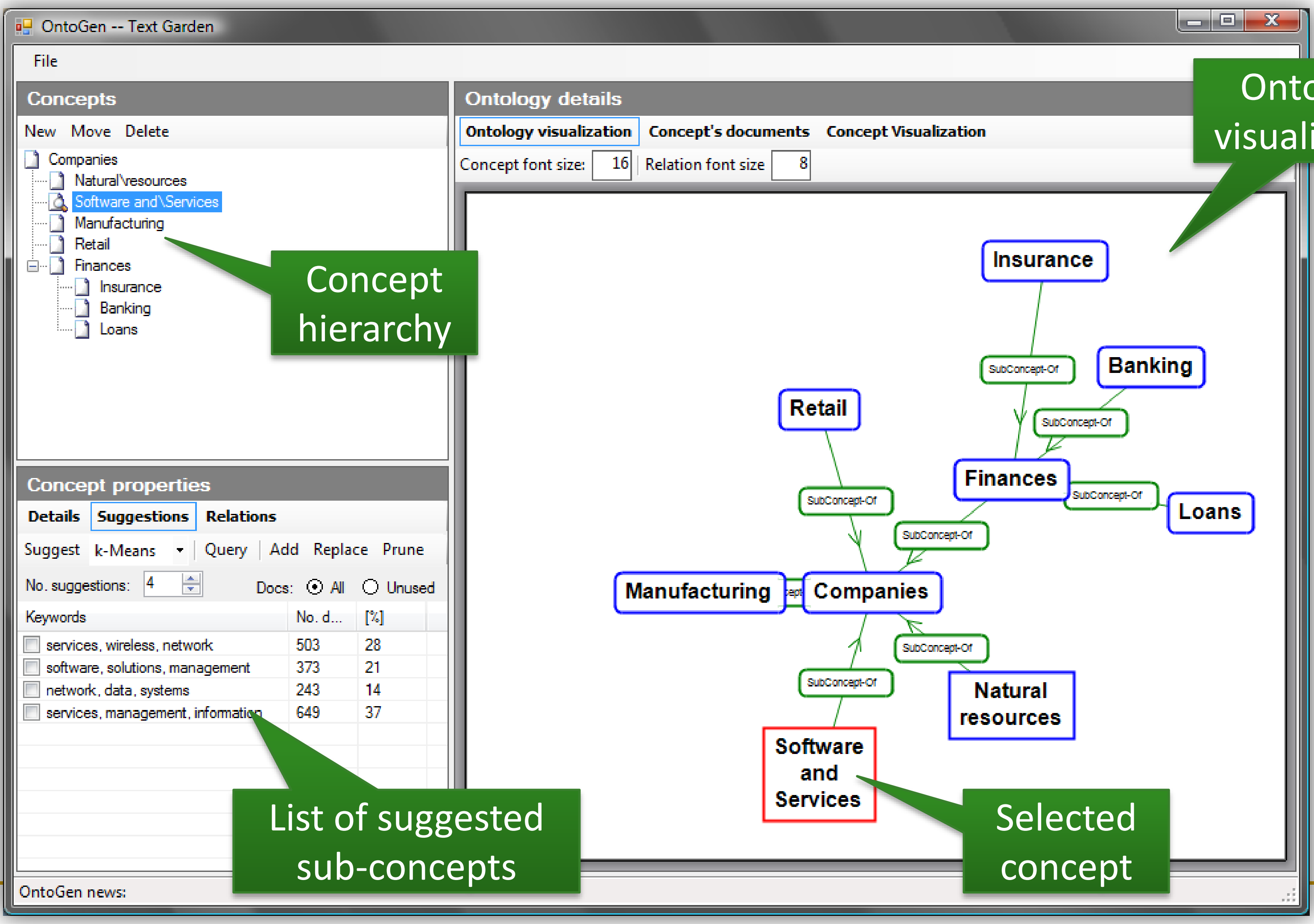
OntoGen – main scenario

- Given a text corpus, construct semi-automatically a taxonomic ontology where each of the documents belongs to a certain class



Blaz Fortuna et al, HCII2007

OntoGen – main screen



OntoGen -- Text Garden

File

Concepts

New Move Delete

- Companies
  - Natural\resources
  - Software and\Services
  - Manufacturing
  - Retail
  - Finances
    - Insurance
    - Banking
    - Loans

Concept properties

Details | Suggestions | Relations

Suggest: k-Means | Query | Add | Replace | Prune

No. suggestions: 4 | Docs: ● All ○ Unused

Keywords	No. d...	[%]
□ services, wireless, network	503	28
□ software, solutions, management	373	21
□ network, data, systems	243	14
□ services, management, information	649	37

Ontology details

Ontology visualization | Concept's documents | Concept Visualization

Concept font size: 16 | Relation font size 8

OntoGen news:

Ontology visualization

Concept hierarchy

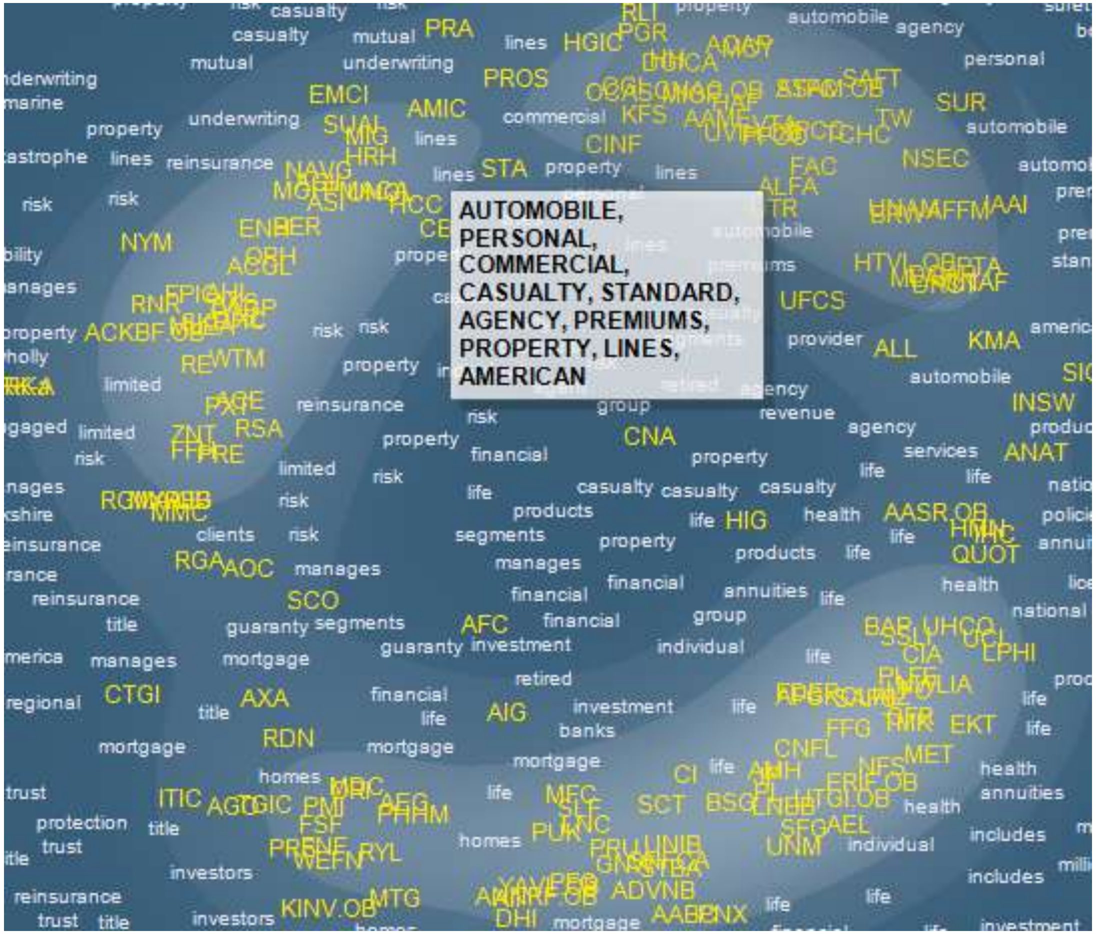
List of suggested sub-concepts

Selected concept

Blaz Fortuna et al, HCII2007

Ontology construction from content visualization

- Documents are visualized as points on 2D map
  - The distance between two instances on the map correspond to their content similarity
  - Characteristic keywords are shown for all parts of the map
- User can select groups of instances on the map to create sub-concepts



Blaz Fortuna et al, HCII2007

Semantic Web

Cyc system (example of deep reasoning)

Cyc …a little bit of historical context

- Older AI-ers know about Cyc:
  - …one of the boldest attempts in AI history to encode common sense knowledge in one KB
  - The project started in 1984 at Stanford as US response to Japan's project on "5th Generation Computer Systems"
  - In 1994 the company Cycorp was established (in Austin, TX)
  - In 2005 Cyc KB gets opened and available for research
    - OpenCyc (http://www.opencyc.org/)
    - ResearchCyc (http://research.cyc.com/)
  - In 2006 Cyc-Europe was established (in Ljubljana, Slovenia)
  - Till 2006 ~$80M was spent into the KB

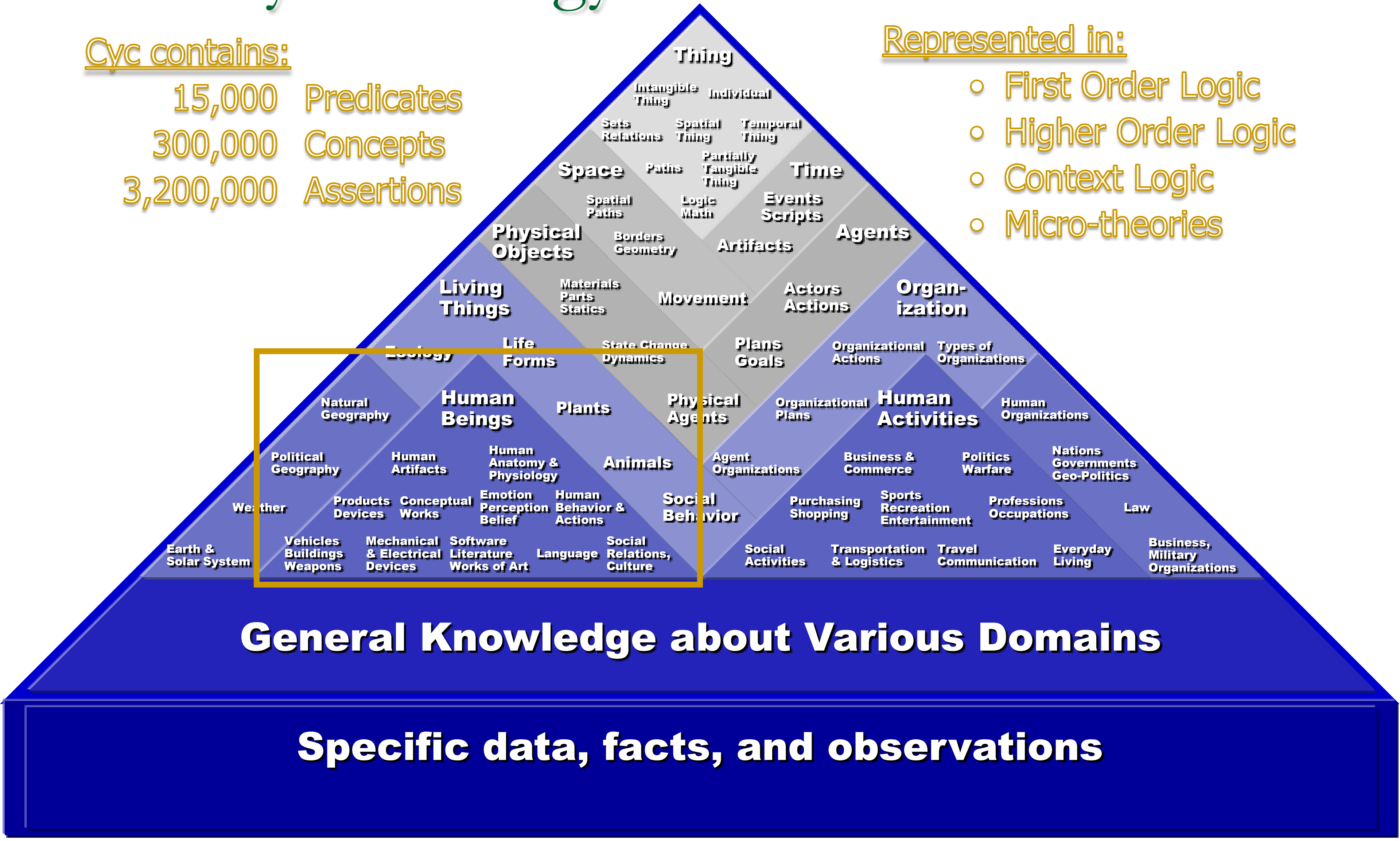
The Cyc Ontology

Cyc contains:

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

Represented in:

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



Thing

Intangible Thing

Individual

Sets Relations

Spatial Thing

Temporal Thing

Space

Paths

Partially Tangible Thing

Time

Spatial Paths

Logic Math

Events Scripts

Physical Objects

Borders Geometry

Artifacts

Agents

Living Things

Materials Parts Statics

Movement

Actors Actions

Organ-ization

Ecology

Life Forms

State Change Dynamics

Plans Goals

Organizational Actions

Types of Organizations

Natural Geography

Human Beings

Plants

Physical Agents

Organizational Plans

Human Activities

Human Organizations

Political Geography

Human Artifacts

Human Anatomy & Physiology

Animals

Agent Organizations

Business & Commerce

Politics Warfare

Nations Governments Geo-Politics

Weather

Products Devices

Conceptual Works

Emotion Perception Belief

Human Behavior & Actions

Social Behavior

Purchasing Shopping

Sports Recreation Entertainment

Professions Occupations

Law

Earth & Solar System

Vehicles Buildings Weapons

Mechanical & Electrical Devices

Software Literature Works of Art

Language

Social Relations, Culture

Social Activities

Transportation & Logistics

Travel Communication

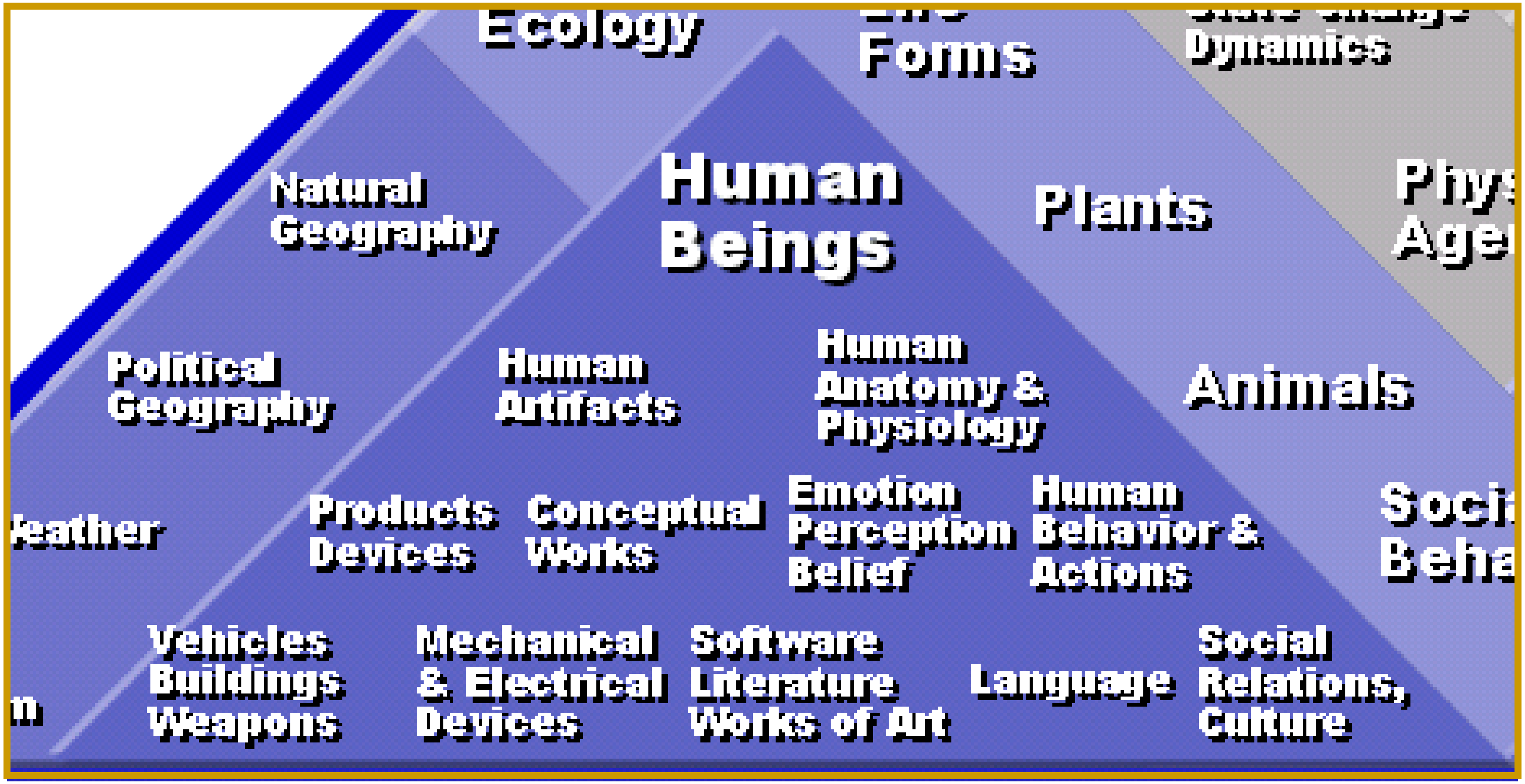
Everyday Living

Business, Military Organizations

General Knowledge about Various Domains

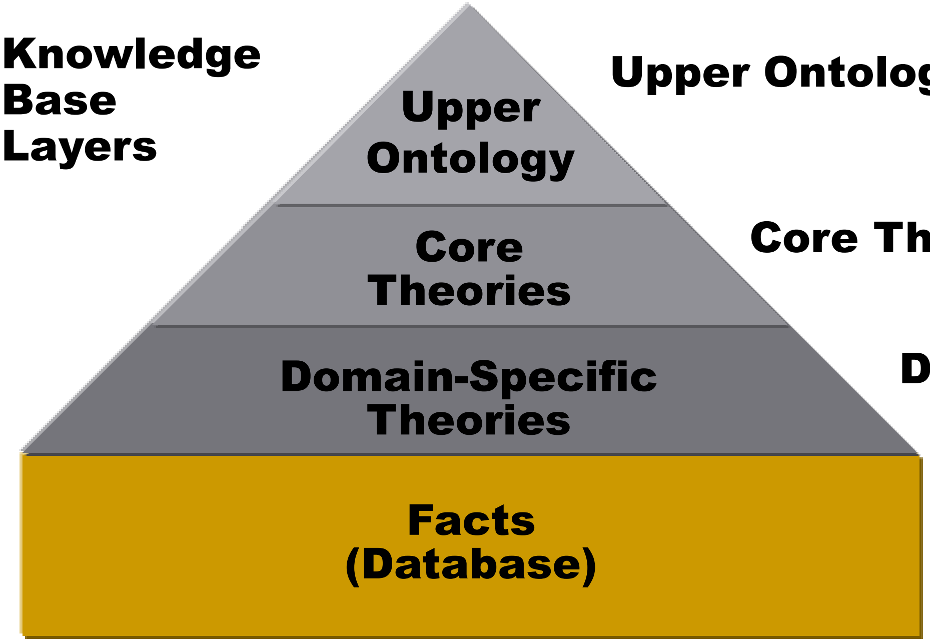
Specific data, facts, and observations

…part of Cyc Ontology on Human Beings



Structure of Cyc Ontology

Knowledge
Base
Layers



Upper Ontology

Core Theories

Domain-Specific Theories

Facts (Database)

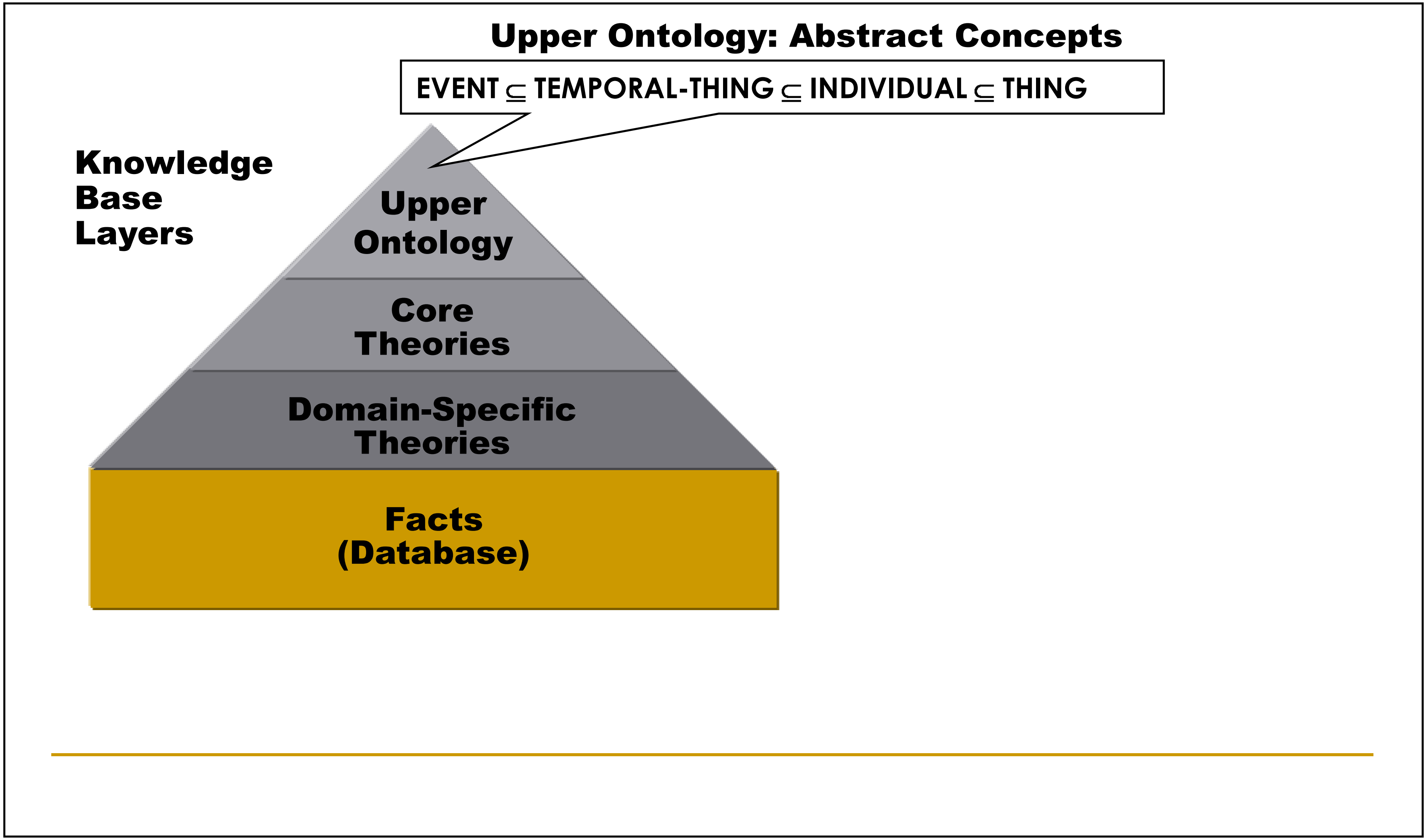
Upper Ontology: Abstract Concepts

Core Theories: Space, Time, Causality, …

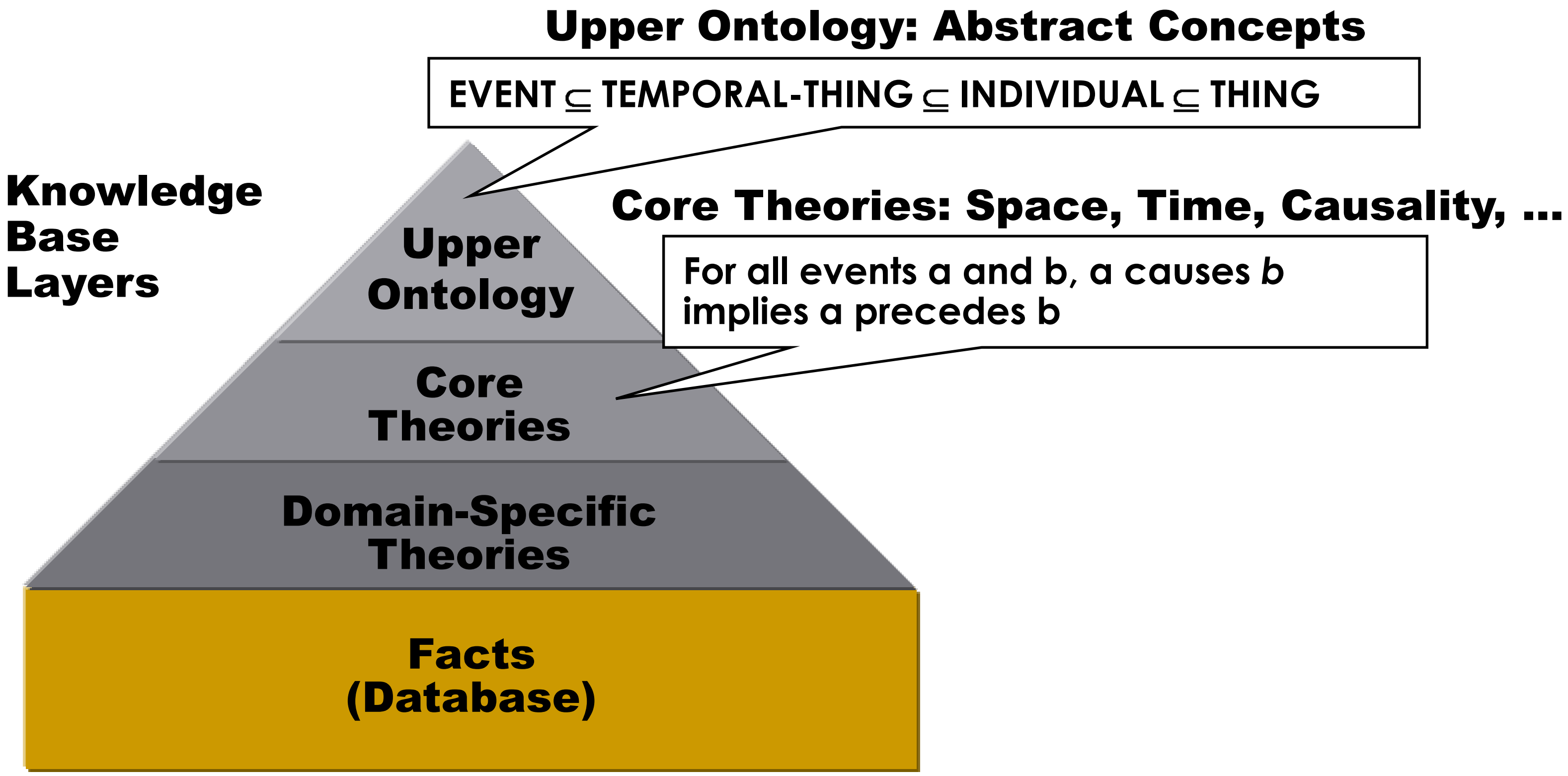
Domain-Specific Theories

Facts: Instances

Structure of Cyc Ontology



Structure of Cyc Ontology



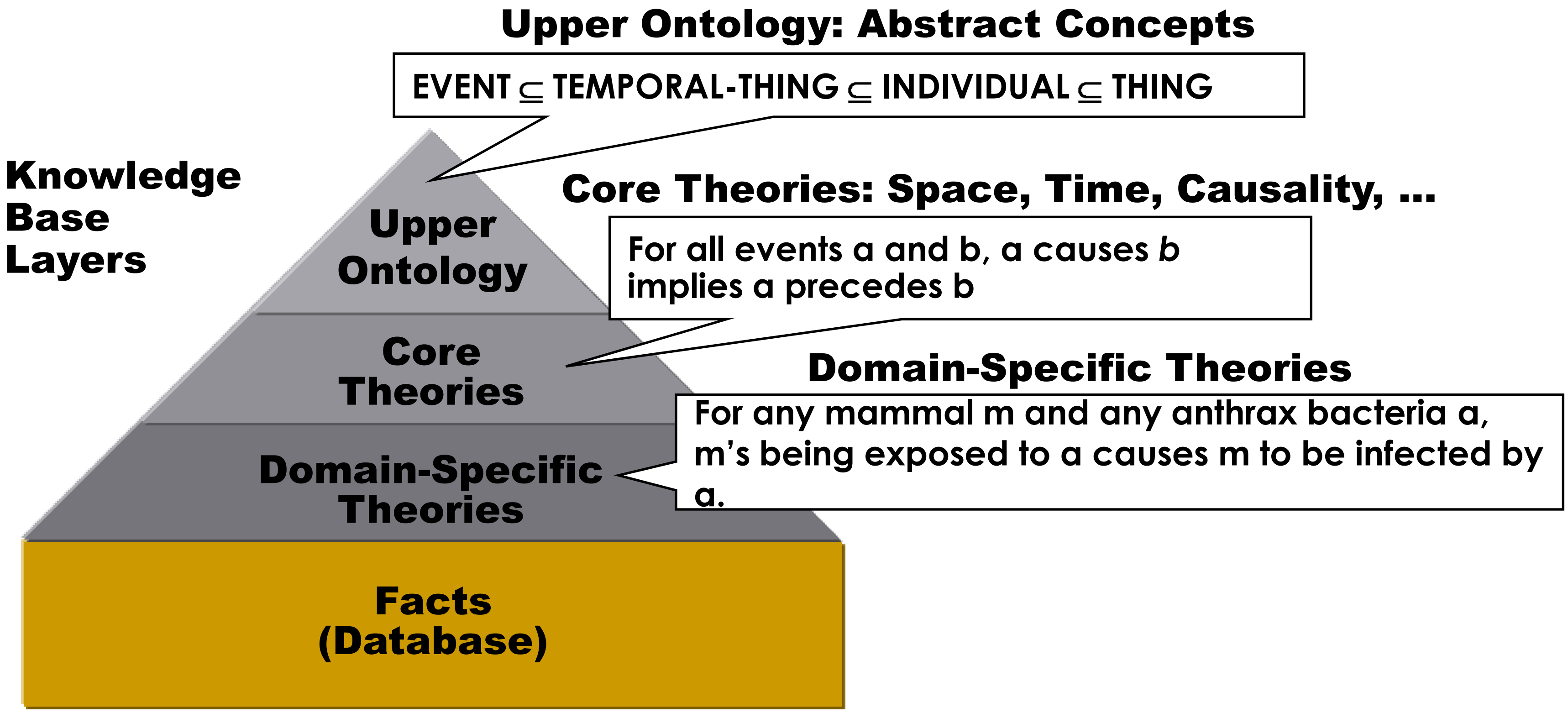
Upper Ontology: Abstract Concepts

EVENT ⊆ TEMPORAL-THING ⊆ INDIVIDUAL ⊆ THING

Core Theories: Space, Time, Causality, …

For all events a and b, a causes *b* implies a precedes b

Structure of Cyc Ontology



Upper Ontology: Abstract Concepts

EVENT ⊆ TEMPORAL-THING ⊆ INDIVIDUAL ⊆ THING

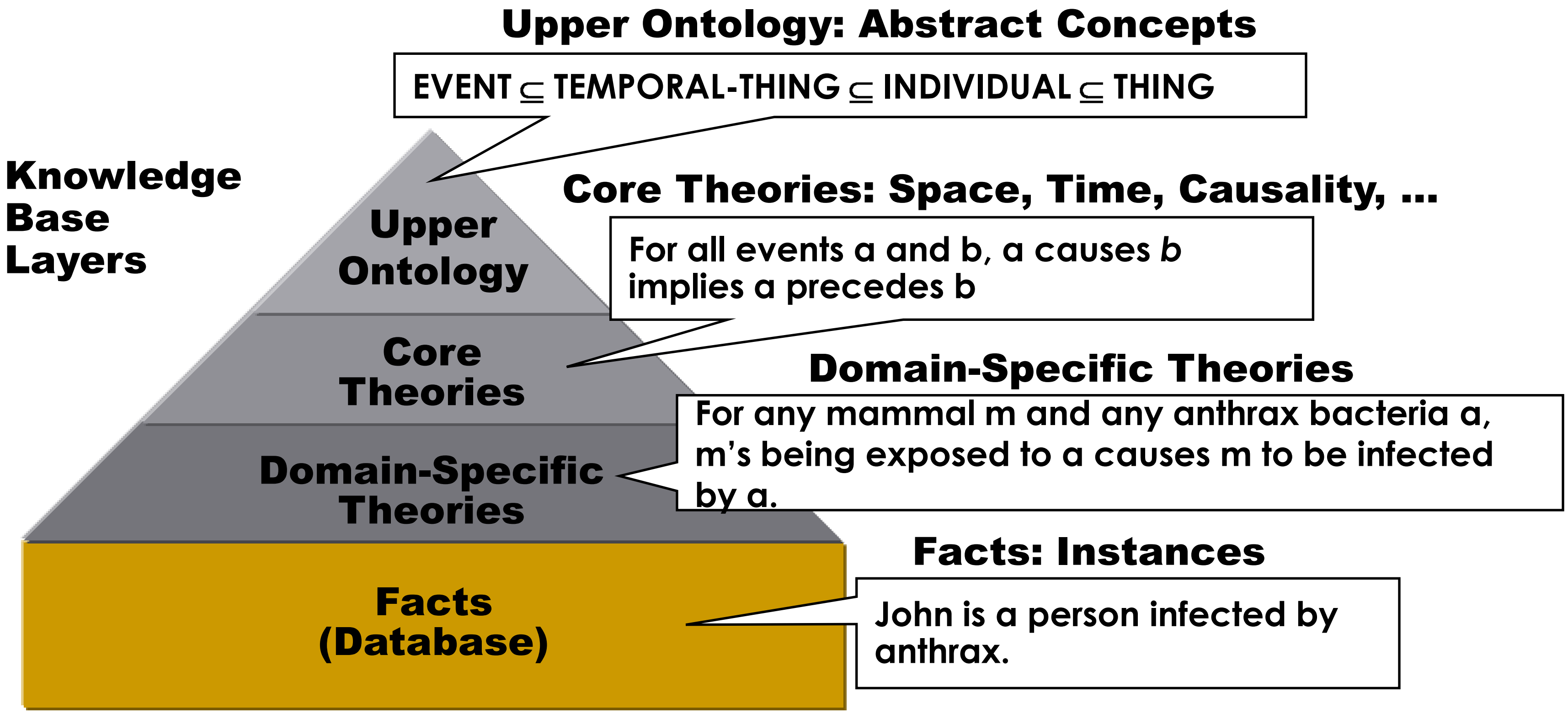
Core Theories: Space, Time, Causality, …

For all events a and b, a causes *b* implies a precedes b

Domain-Specific Theories

For any mammal m and any anthrax bacteria a, m’s being exposed to a causes m to be infected by a.

Structure of Cyc Ontology



Upper Ontology: Abstract Concepts

EVENT ⊆ TEMPORAL-THING ⊆ INDIVIDUAL ⊆ THING

Core Theories: Space, Time, Causality, …

For all events a and b, a causes *b* implies a precedes b

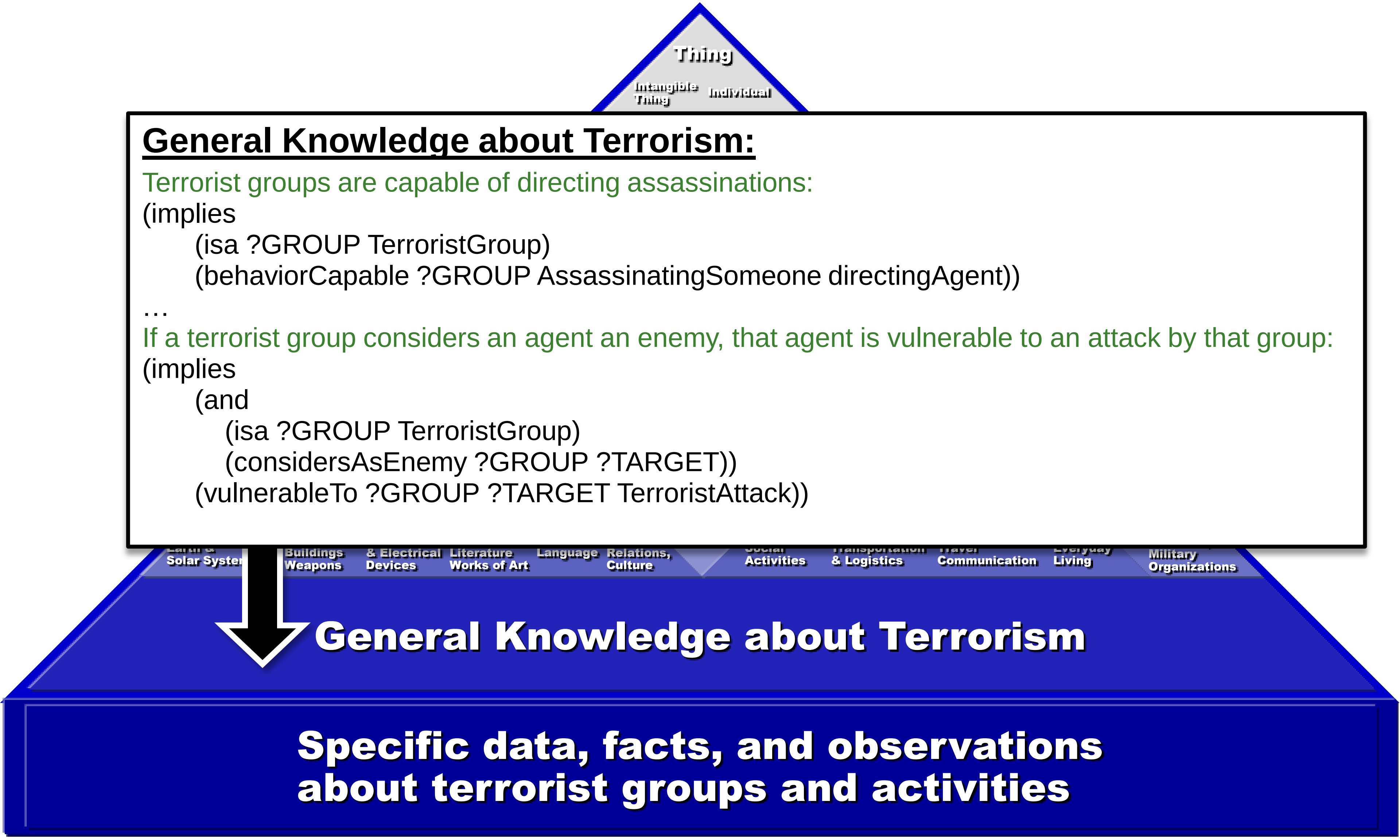
Domain-Specific Theories

For any mammal m and any anthrax bacteria a, m's being exposed to a causes m to be infected by a.

Facts: Instances

John is a person infected by anthrax.

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))
…
```

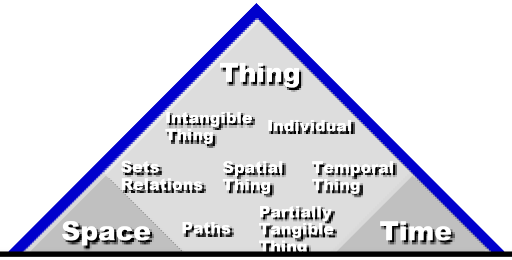
If a terrorist group considers an agent an enemy, that agent is vulnerable to an attack by that group:

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

General Knowledge about Terrorism

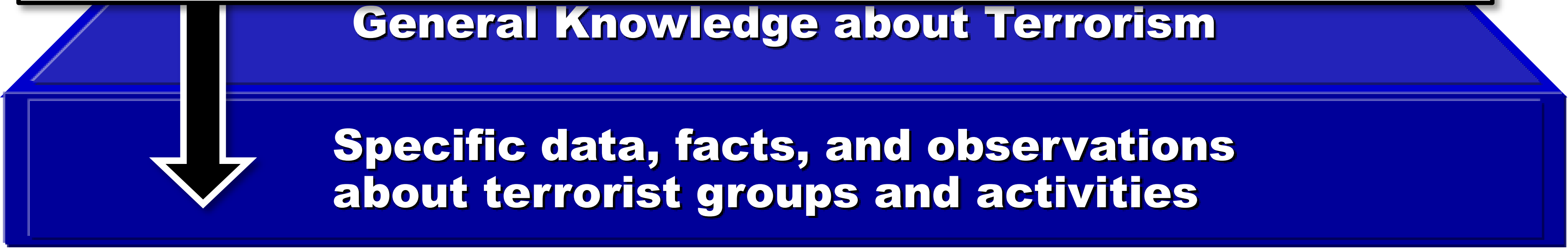
Specific data, facts, and observations about terrorist groups and activities

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

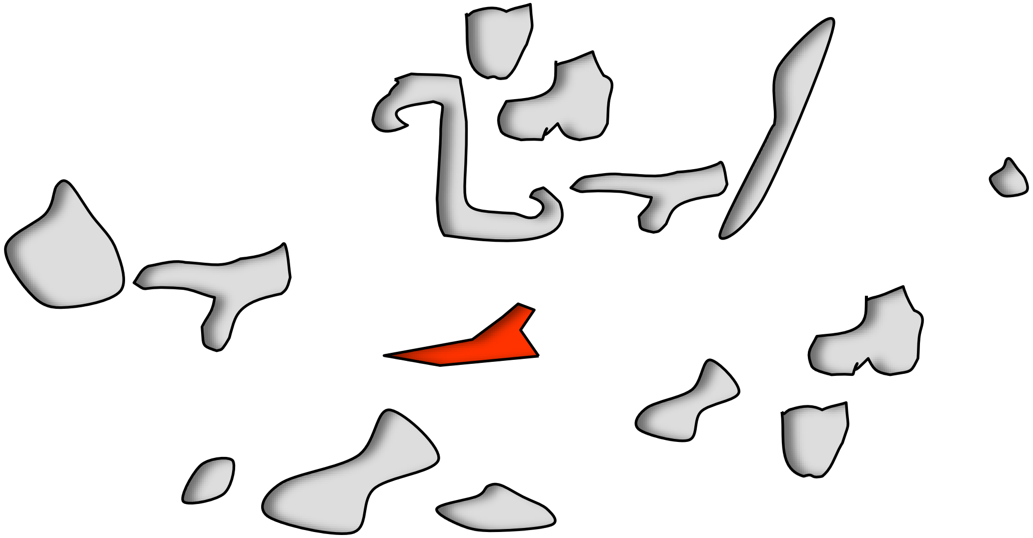
Specific data, facts, and observations
about terrorist groups and activities

An example of Psychoanalyst's Cyc taxonomic context

```
#$Psychoanalyst (lexical representation: “psychoanalyst”, “psychoanalysts”)
    specialization-of #$MedicalCareProfessional
    |    specialization-of #$HealthProfessional
    |        specialization-of #$Professional-Adult
    |            specialization-of #$Professional
    specialization-of #$Psychologist
    |    specialization-of #$Scientist
    |        specialization-of #$Researcher
    |        |    specialization-of #$PersonWithOccupation
    |        |    |    specialization-of #$Person
    |        |    |    |    specialization-of #$HomoSapiens
    |        |    |    |    |    instance-of #$BiologicalSpecies
    |        |    |    |    |    |    specialization-of #$BiologicalTaxon
    |        |    |    |    |    instance-of #$SomeSampleKindsOfMammal-Biology-Topic
    |        |    specialization-of #$AdultAnimal
    |        |    |    specialization-of #$Animal
    |        |    |    |    specialization-of #$SolidTangibleThing
    |        |    |    |        instance-of #$StatesOfMatter-Material-Topic
    |        specialization-of (#$GraduateFn #$University)
    |            specialization-of (#$Graduate #$DegreeGrantingHigherEducationInstitution)
    specialization-of #$Counselor-Psychological
```

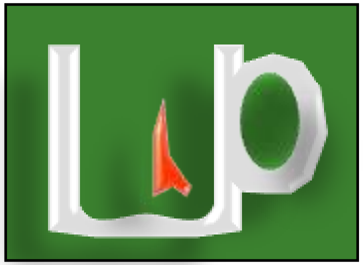
Example Vocabulary: Senses of ‘**In**’ relation (1/3)

- Can the inner object leave by passing between members of the outer group?
  - Yes -- Try **#$in-Among**

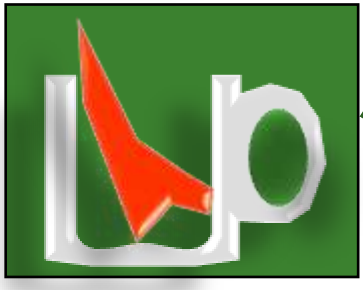


Example Vocabulary: Senses of ‘In’ relation (2/3)

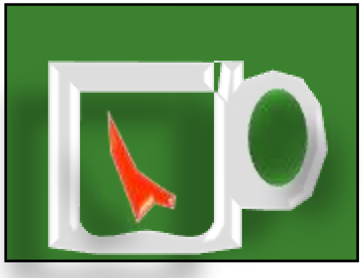
- Does part of the inner object stick out of the container?
  - ❑ None of it. -- Try **#$in-ContCompletely**



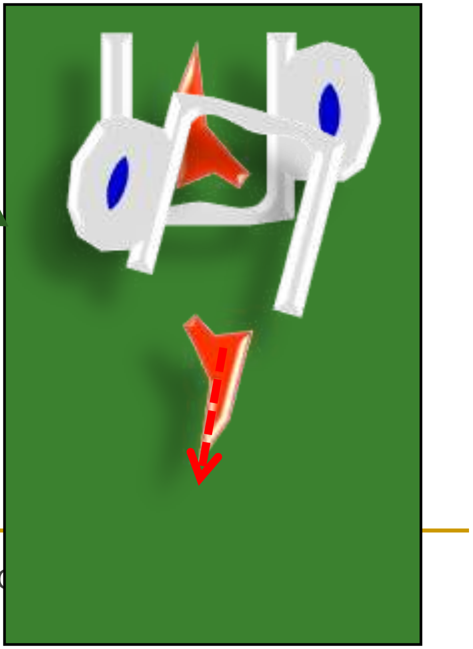
  - ❑ Yes -- Try **#$in-ContPartially**



  - ❑ No -- Try
  - **#$in-ContClosed**



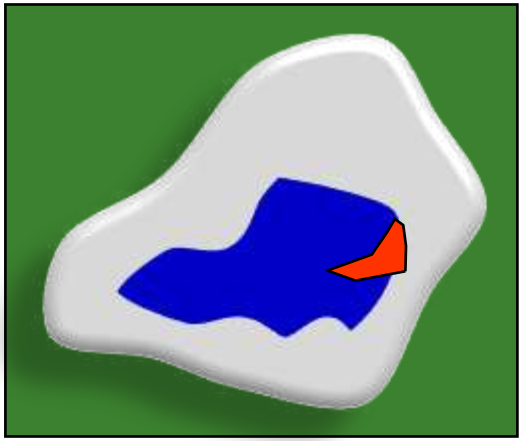
- ❑ If the container were turned around could the contained object fall out?
  - Yes -- Try **#$in-ContOpen**



Example Vocabulary: Senses of ‘In’ relation (3/3)

Is it attached to the inside of the outer object?

- Yes -- Try **#$connectedToInside**

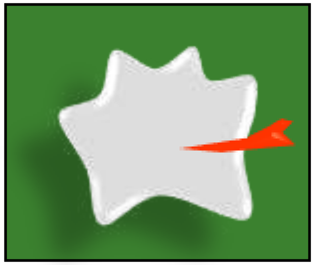


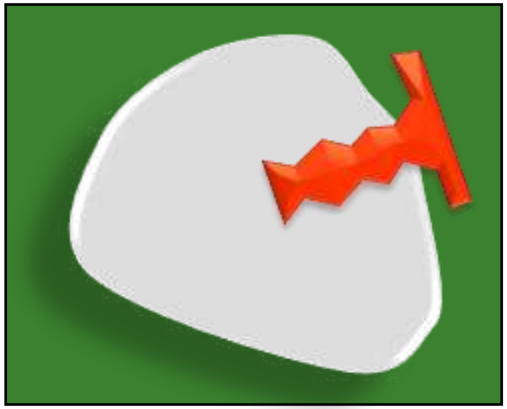
Can it be removed by pulling, if enough force is used, without damaging either object?

- No -- Try **#$in-Snugly** or **#$screwedIn**

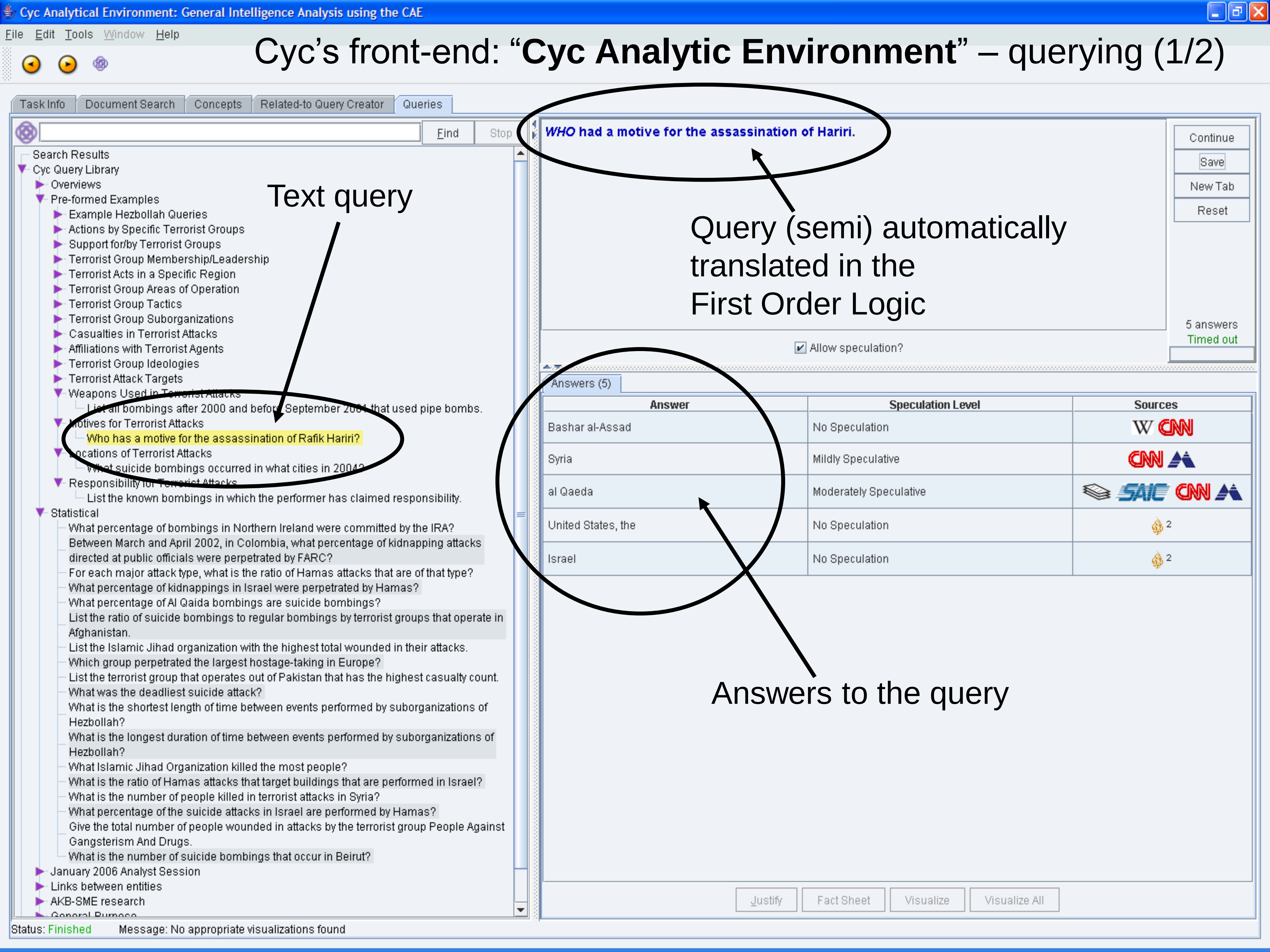
Does the inner object stick into the outer object?

- Yes – Try **#$sticksInto**





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



- 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 2001 that used pipe bombs.
    - Motives for Terrorist Attacks
      - Who has a motive for the assassination of Rafik Hariri?
    - Locations of Terrorist Attacks
      - What suicide bombings occurred in what cities in 2004?
    - Responsibility for Terrorist Attacks
      - List the known bombings in which the performer has claimed responsibility.
  - Statistical
    - What percentage of bombings in Northern Ireland were committed by the IRA?
    - Between March and April 2002, in Colombia, what percentage of kidnapping attacks directed at public officials were perpetrated by FARC?
    - For each major attack type, what is the ratio of Hamas attacks that are of that type?
    - What percentage of kidnappings in Israel were perpetrated by Hamas?
    - What percentage of Al Qaida bombings are suicide bombings?
    - List the ratio of suicide bombings to regular bombings 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

***WHO* had a motive for the assassination of Hariri.**

5 answers
Timed out

☑ Allow speculation?

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

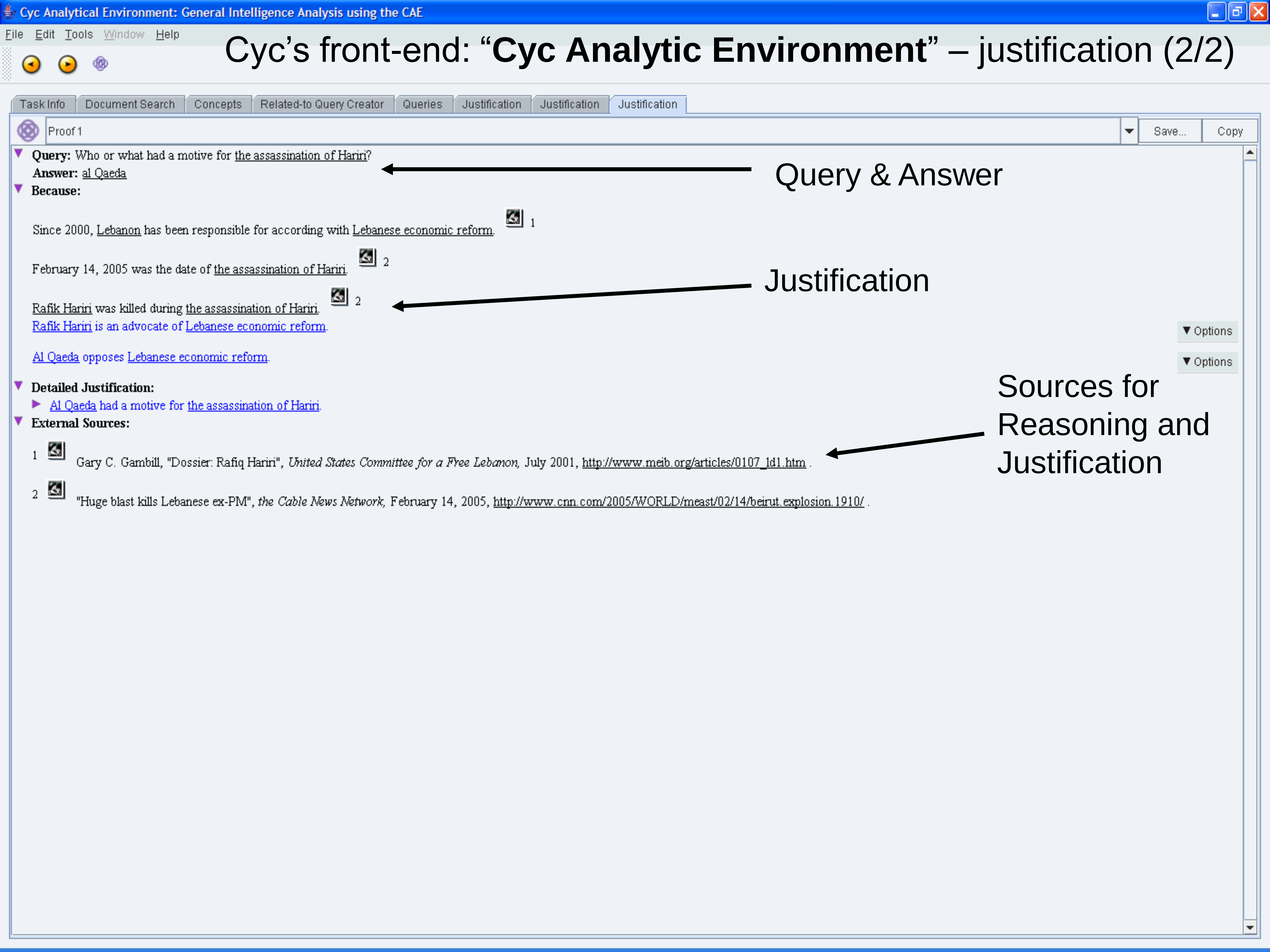
Status: Finished Message: No appropriate visualizations found

Text query

Query (semi) automatically translated in the First Order Logic

Answers to the query

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



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.

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

Semantic Web

Web X.X versions (past and current trends)

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

	Description	Technologies
Web 1.0	**Static** HTML pages (web as we first learned it)	HTML, HTTP
Web 1.5	**Dynamic** HTML content (web as we know it)	Client side (JavaScript, DHTML, Flash, …), server side (CGI, PHP, Perl, ASP/.NET, JSP, …)
Web 2.0	**Integration** 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, …
Web 3.0	…adding **meaning** 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” (http://www.nytimes.com/2006/11/12/business/12web.html?ref=business)

Web 2.0 –is there any new quality?

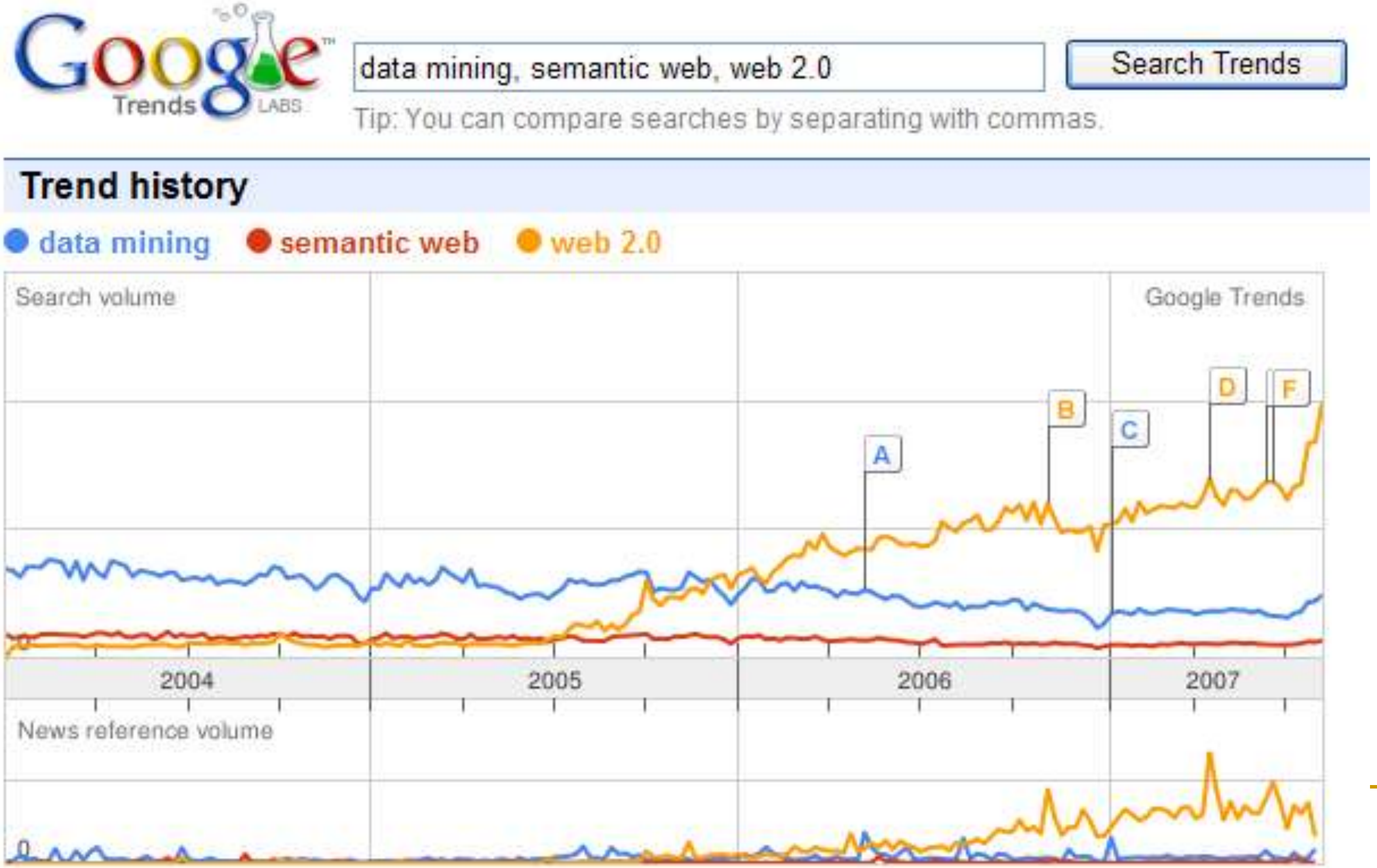
- With “Web 2.0” the Web community became **really aware** of the importance of the global collaborative work
  - …next step in the globalization of the Web
  - **Bottom-up** “social networking” seems to nicely complement the traditional **top-down** schema design approaches

BASECAMP
BACKPACK
CHALK.IT
REMEMBERTHEMILK
WRITEBOARD
TADALIST
GOOGLE MAPS
KIKO
WRITELEY
GTD
GMAIL
GOOGLE RSSREADER
GOOGLE IG
UNWIELDY
TAGTAGGER
ZIMBRA
MEEBO
TAGGLING
CLOUDALICIOUS
WIKIPEDIA
Focus on Simplicity
PROTOPAGE
TAGCLOUD
Wikis
ZOOZIO
FLOCK
DELICIOUS
COLLABORATION
WIDGETS
Joy of Use
NETVIBES
STUMBLEUPON
FLICKR
Folksonomy
SHARING
LIVE.COM
AJAX
43THINGS
DIGG
SQUIDOO
PAGERANK
Usability
FLICKR
NING
EBAY
Recommendation
USER CENTERED
SOFLOW
LINKEDIN
FOAF
BLINKSALE
FOTOLIA
ETSY
OPENBC
ORKUT
Social Software
SIXDEGREES
PERPETUAL BETA
PAYPAL
The Long Tail
EBAY
UPCOMING
FRIENDSTER
XFN
Participation
TRUST
Economy
AMAZON
MTURK
PLANZO
BLOGGER
Blogs
AGGREGATION
VC
Affiliation
ITUNES
TYPEPAD
TECHNORATI
LASTFM
READ.IO
WORDPRESS
MINT
Web2.0
beta
COSTPERCLICK
ADSENSE
MEASUREMAP
IPODDER
PODCASTING
IM
PANDORA
Audio
ODEO
SKYPE GTALK JABBER
LIVELINESS
SIMPLEBITS
Convergence
ITUNES
DATA INSIDE
SEO
Design
STYLEGALA
VIMEO
Video
VIDEOCASTING
BIT TORRENT
GRANULARITY
MODULARITY
SIMPLICITY
9RULES
YOUTUBE
CASTPOST
WIFI
Mobility
UMTS
Remixability
SYNDICATION
SOAP
REST
RUBYONRAILS
CSSBEAUTY
GOOGLE MAPS
XML
CREATIVECOMMONS
A9
PLAZES
RSS
Standardization
ZENGARDEN
YAHOO MAPS
OpenAPIs
TAGCLOUD
SEMANTIC
CSS Design
NEWSGATOR
BLOGLINES
AMAZON
DataDriven
XHTML
ACCESSIBILITY
WASP
FEEDSTER
ROJO
GOOGLE MAPS
Web Standards
ALISTAPART
FEEDBURNER
GMAPS MASHUPS
TECHNORATI
FIREFOX
MYYAHOO
GOOGLE NEWS
GOOGLE
A9
Microformats
BUZZTRACKER
ICEROCKET
INORDER
WINK
ALEXA
MICROFORMATS.ORG
ROLLYO
TRENDMAPPER
PUBSUB
MARKUS.ANGERMEIER @ APERTO.DE
RELEASED UNDER CC 2.0 DE ATTRIBUTION SHARE ALIKE 11.11.05

Visualization of Web 2.0 typical vocabulary (http://en.wikipedia.org/wiki/Image:Web20_en.png)

Web 2.0 – the current hype!

Google search volume of “**data mining**” vs. “**Web 2.0**” vs. “**semantic web**” (http://www.google.com/trends?q=data+mining%2C+semantic+web%2C+web+2.0)



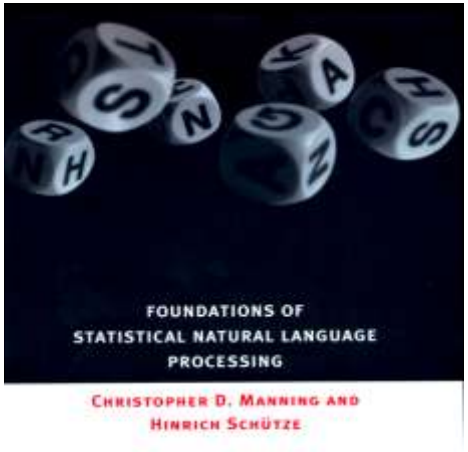
What about Web 4.0? ☺

- Citation from some blog:
  - "…*Web 4.0 is the impending state at which all information converges into a great ball of benevolent self-aware light, and solves every problem from world peace to* …" http://blogs.intel.com/it/2006/11/web_40_a_new_hype.html
- Ultimate stage in web development…
  - …will prevent Web 5.0 to happen since everything will be resolved already by Web 4.0.

Wrap-up

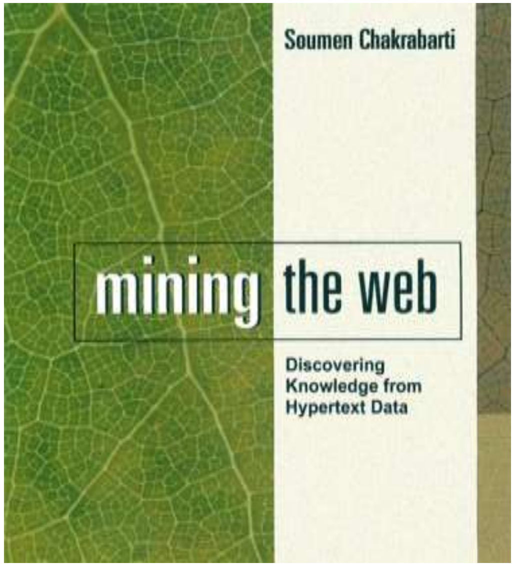
…what did we learn and where to continue?

References to some Text-Mining & Link Analysis Books



FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

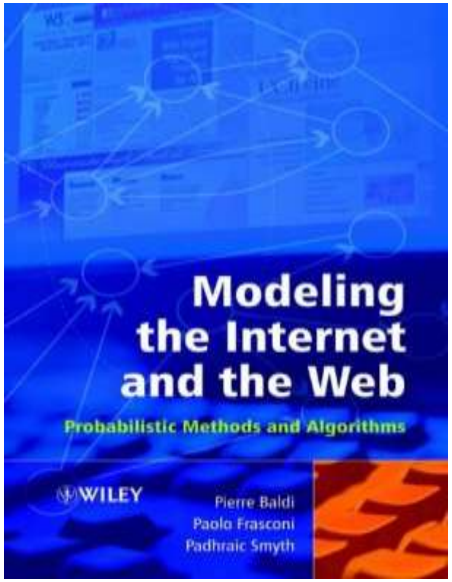
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE



Soumen Chakrabarti

mining the web

Discovering Knowledge from Hypertext Data

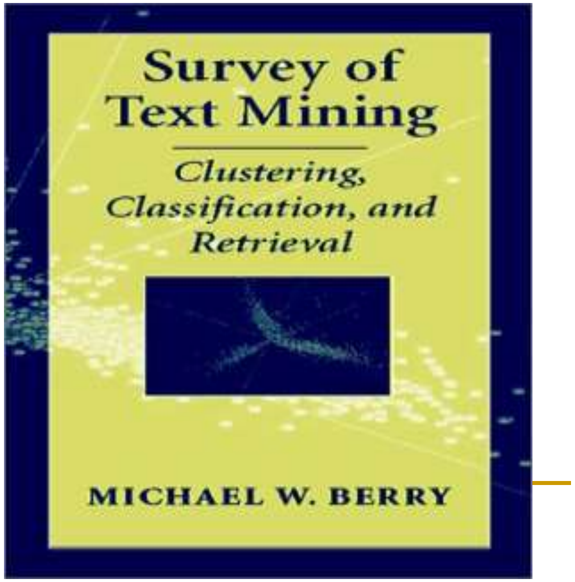


Modeling the Internet and the Web

Probabilistic Methods and Algorithms

WILEY

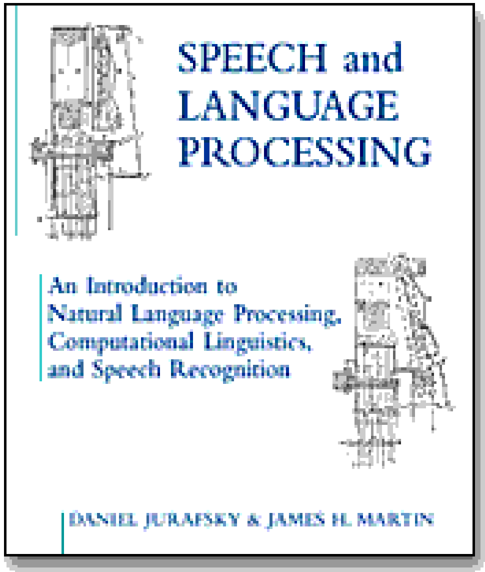
Pierre Baldi
Paolo Frasconi
Padhraic Smyth



Survey of Text Mining

Clustering, Classification, and Retrieval

MICHAEL W. BERRY



SPEECH and LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

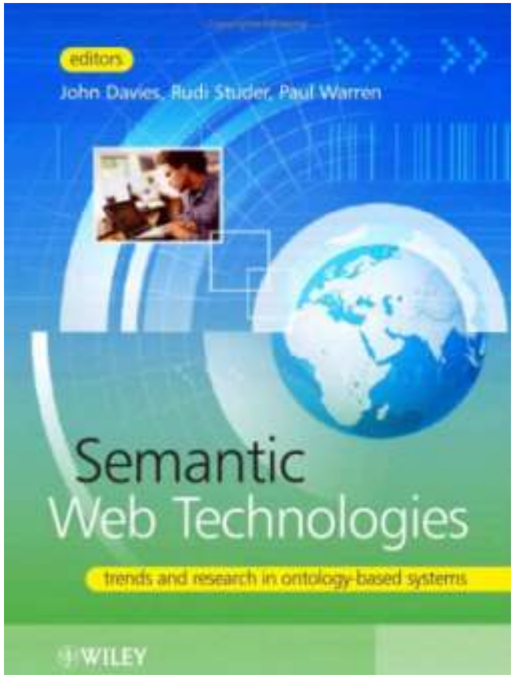
DANIEL JURAFSKY & JAMES H. MARTIN

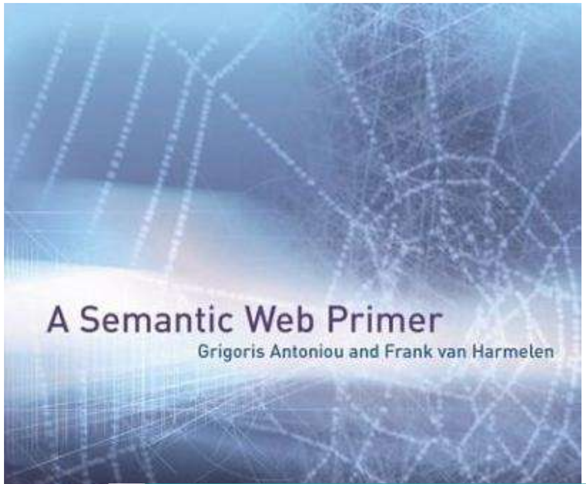
THE TEXT MINING HANDBOOK

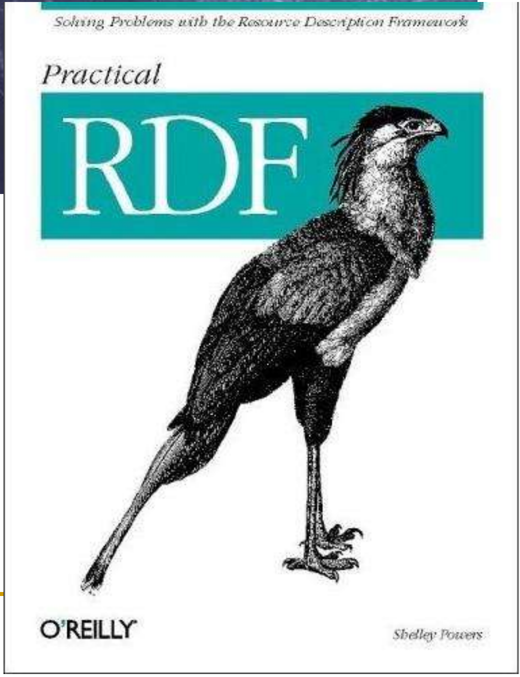
Advanced Approaches in Analyzing Unstructured Data

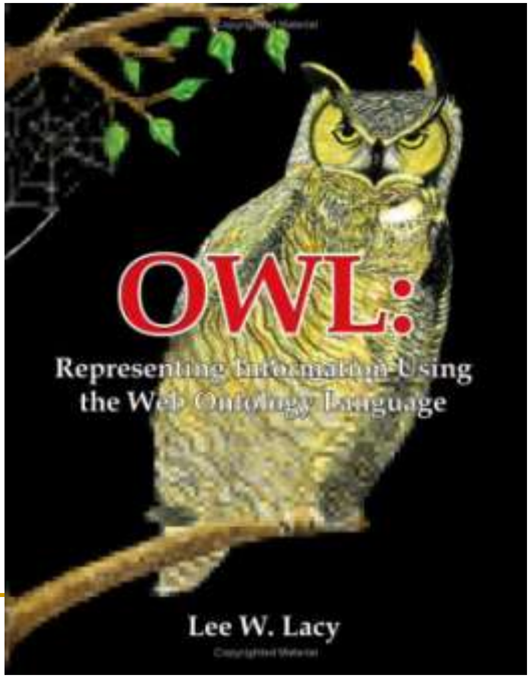
RONEN FELDMAN
JAMES SANGER

References to some Semantic Web Books


editors
John Davies, Rudi Studer, Paul Warren
Semantic
Web Technologies
trends and research in ontology-based systems
WILEY







References to the main conferences

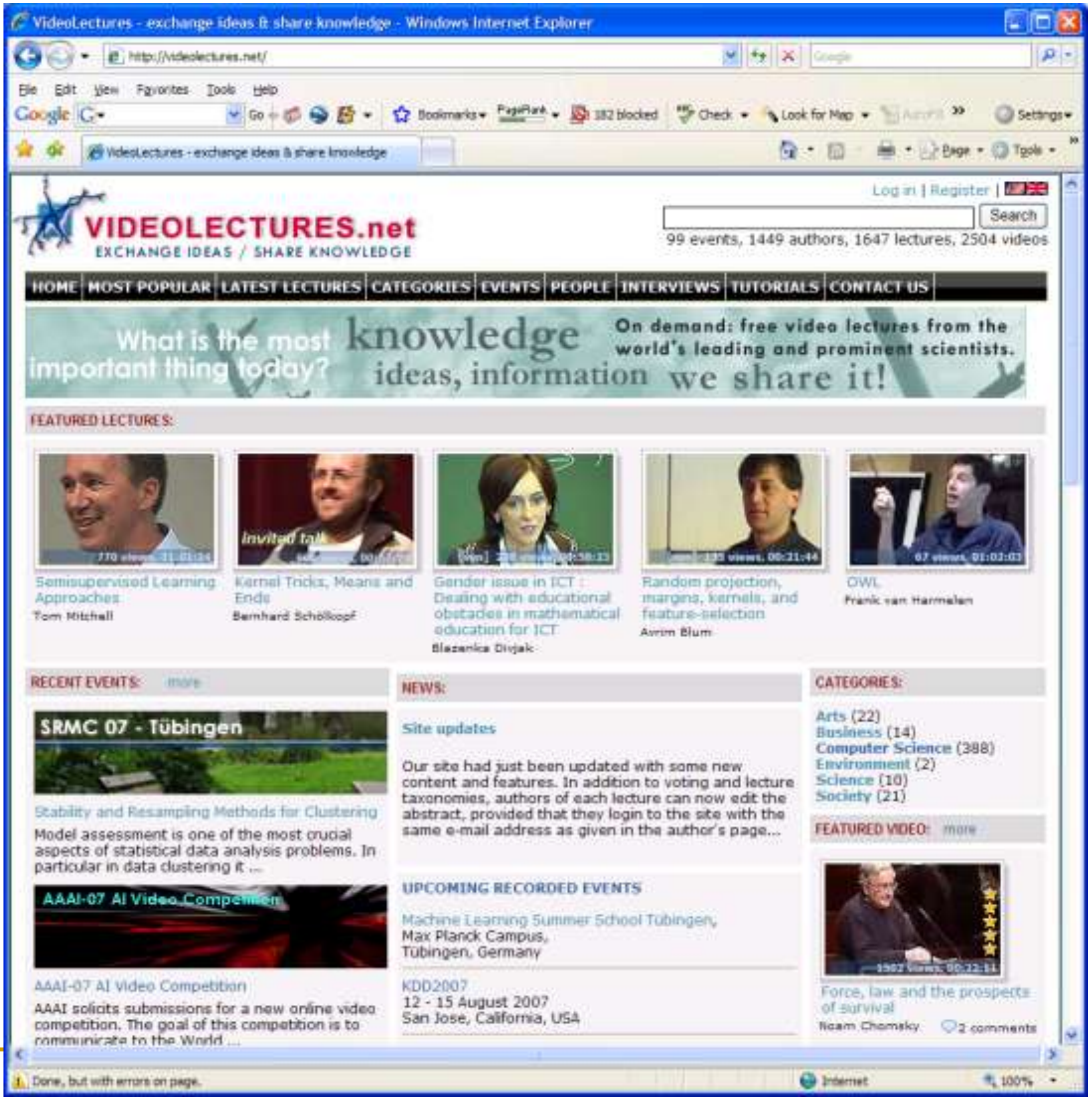
- Information Retrieval:
  - SIGIR, ECIR
- Machine Learning/Data Mining:
  - ICML, ECML/PKDD, KDD, ICDM, SDM
- Computational Linguistics:
  - ACL, EACL, NAACL
- Semantic Web:
  - ISWC, ESWS

References to some of the Text-Mining & Link Analysis workshops at KDD, ICDM, ICML and IJCAI conferences (available online)

- ICML-1999 Workshop on Machine Learning in Text Data Analysis (TextML-1999) (http://www-ai.ijs.si/DunjaMladenic/ICML99/TLWsh99.html), Bled 1999
- KDD-2000 Workshop on Text Mining (TextKDD-2000) (http://www.cs.cmu.edu/~dunja/WshKDD2000.html), Boston 2000
- ICDM-2001 Workshop on Text Mining (TextKDD-2001) (http://www-ai.ijs.si/DunjaMladenic/TextDM01/), San Jose 2001
- ICML-2002 Workshop on Text Learning (TextML-2002) (http://www-ai.ijs.si/DunjaMladenic/TextML02/), Sydney 2002
- IJCAI-2003 Workshop on Text-Mining and Link-Analysis (TextLink-2003) (http://www.cs.cmu.edu/~dunja/TextLink2003/), Acapulco 2003
- KDD-2003 Workshop on Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003) (http://www.cs.cmu.edu/~dunja/LinkKDD2003/), Washington DC 2003
- KDD-2004 Workshop on Workshop on Link Analysis and Group Detection (LinkKDD2004) (http://www.cs.cmu.edu/~dunja/LinkKDD2004/), Seattle 2004
- KDD-2005 Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005) (http://www.isi.edu/LinkKDD-05/), Chicago 2005
- KDD-2006 Workshop on Link Analysis: Dynamics and Statics of Large Networks (LinkKDD 2006) (http://kt.ijs.si/Dunja/LinkKDD2006/), Philadelphia 2006
- IJCAI-2007 Workshop on Text-Mining & Link-Analysis (TextLink 2007) (http://kt.ijs.si/dunja/textlink2007/), Hyderabad 2007

References to video content

- Many scientific events are recorded and freely available from http://videolectures.net/
  - …videos categorized by a subject http://videolectures.net/Top/Computer_Science/


VideoLectures - exchange ideas & share knowledge - Windows Internet Explorer
http://videolectures.net/
File Edit View Favorites Tools Help
Google Go Bookmarks PageRank 182 blocked Check Look for Map Settings
VideoLectures - exchange ideas & share knowledge
Page Tools
Log in | Register
Search
HOME MOST POPULAR LATEST LECTURES CATEGORIES EVENTS PEOPLE INTERVIEWS TUTORIALS CONTACT US
What is the most important thing today?
knowledge ideas, information we share it!
Invited talk
770 views, 01:01:34
135 views, 00:21:44
67 views, 01:02:03
1902 views, 00:22:11
Done, but with errors on page.
Internet
100%

VIDEOLECTURES.net
EXCHANGE IDEAS / SHARE KNOWLEDGE

99 events, 1449 authors, 1647 lectures, 2504 videos

On demand: free video lectures from the world's leading and prominent scientists.

FEATURED LECTURES:

Semisupervised Learning Approaches
Tom Mitchell

Kernel Tricks, Means and Ends
Bernhard Schölkopf

Gender issue in ICT : Dealing with educational obstacles in mathematical education for ICT
Blazenka Divjak

Random projection, margins, kernels, and feature-selection
Avrim Blum

OWL
Frank van Harmelen

RECENT EVENTS: more

SRMC 07 - Tübingen

Stability and Resampling Methods for Clustering

Model assessment is one of the most crucial aspects of statistical data analysis problems. In particular in data clustering it ...

AAAI-07 AI Video Competition

AAAI-07 AI Video Competition

AAAI solicits submissions for a new online video competition. The goal of this competition is to communicate to the World ...

NEWS:

Site updates

Our site had just been updated with some new content and features. In addition to voting and lecture taxonomies, authors of each lecture can now edit the abstract, provided that they login to the site with the same e-mail address as given in the author's page...

UPCOMING RECORDED EVENTS

Machine Learning Summer School Tübingen,
Max Planck Campus,
Tübingen, Germany

KDD2007
12 - 15 August 2007
San Jose, California, USA

CATEGORIES:

Arts (22)
Business (14)
Computer Science (388)
Environment (2)
Science (10)
Society (21)

FEATURED VIDEO: more

Force, law and the prospects of survival
Noam Chomsky 2 comments

Some of the Products

- Authonomy
- ClearForest
- Megaputer
- SAS – Enterprise-Miner
- SPSS – Clementine, LexiQuest
- Oracle – ConText
- IBM - Intelligent Miner for Text, UIMA
- Microsoft – SQL Server

Major Databases & Text-Mining

- **Oracle** – includes some functionality within the database engine (e.g. classification with SVM, clustering, …)
- **IBM DB2** – text mining appears as a database extender accessible through several SQL functions
  - …a lot of functionality is included in WebFountain and UIMA environments
- **Microsoft SQL Server** – text processing is available as a preprocessing stage in Data-Transformation Services module

Final Remarks

- In the future we can expect stronger integration and **bigger overlap** between Text-Mining, Information-Retrieval, Natural-Language-Processing and Semantic-Web…
- …the technology and solutions will try to **capture deeper semantics** within the text
- …**integration of various** data sources (where text and graphs are just two of the modalities) is becoming increasingly important.