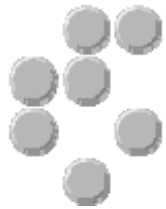


Knowledge Management

Part of
"New Media and eScience"
MSc Programme 2006/07

Nada Lavrač
(with selected slides from
Dunja Mladenić, Marko Grobelnik and Blaž Fortuna)

Jožef Stefan Institute

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

KM Course Outline

- **Nada Lavrač**: Introduction to KM and selected KM tools
- **Peter Ljubič**: Social Network Analysis with Pajek
- **Matjaz Juršič**: Demo of OntoGen ontology construction tool and application in media analysis (press)
- **Ingrid Petrič**: OntoGen application in medical text analzsis (autism)

Knowledge Technologies for KM

JSI Department of Knowledge Technologies

- **Knowledge management - Knowledge technologies relationship:**
  - **Knowledge management**
    - Main topics: knowledge acquisition/ generation, storage/development, transfer, customization/use
    - Three aspects of KM: organizational, technological and sociological
  - **Knowledge technologies**
    - technological aspect of KM – methods, techniques and tools

Knowledge Technologies for KM

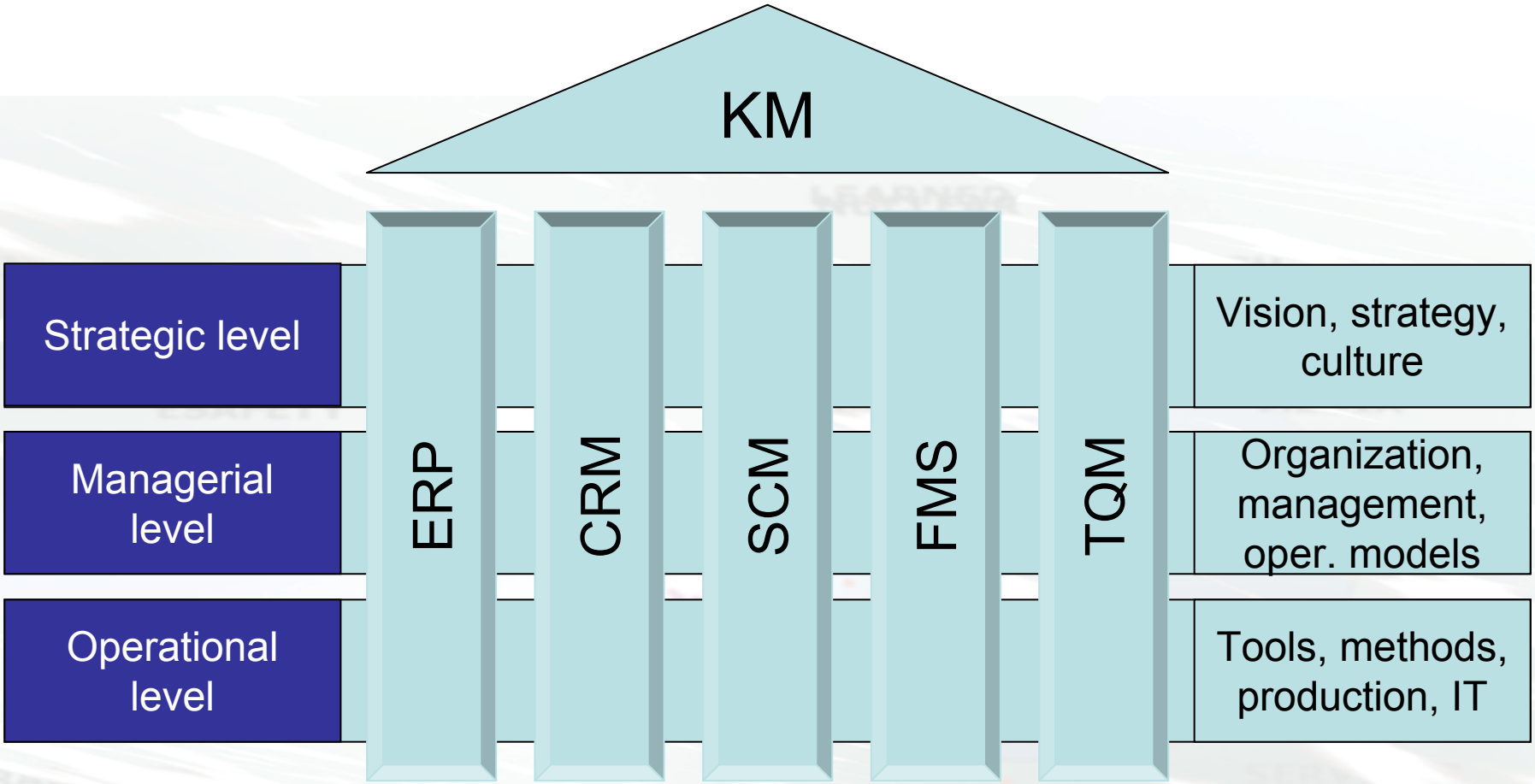
Department of Knowledge Technologies - main research areas

- data (text, web) mining and knowledge discovery
- decision support
- human language technologies
- semantic web
- knowledge representation, logical and probabilistic reasoning, expert systems, artificial intelligence
- Applications and eScience
- eLearning (Center of knowledge transfer in IT)

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis: A case study in semi-automated trust modeling
  - Text mining: A case study in semi-automated ontology construction
  - OntoGen: semi-automated ontology construction tool

Traditional KM

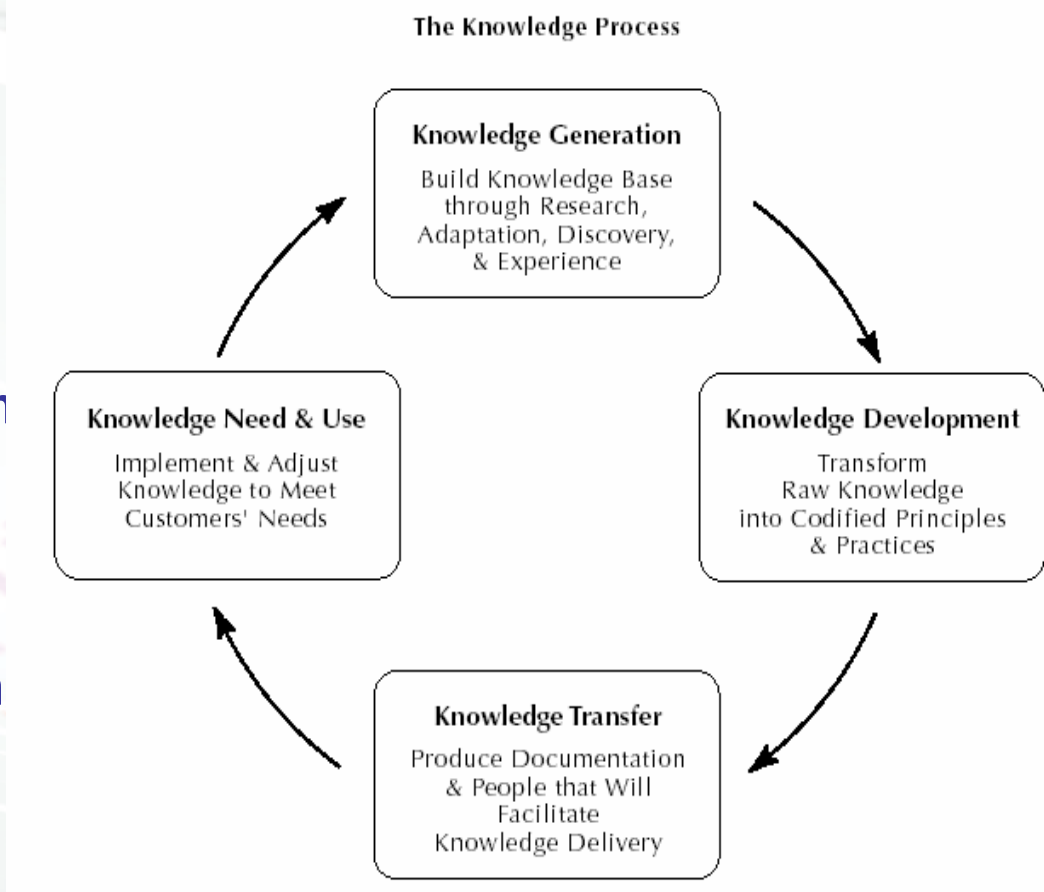


ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), SCM (Supply Chain Management) FMS (Flexible Manufacturing Systems), TQM (Total Quality Management), ...

Traditional KM

Managing knowledge

- generation (acquisition
- storage and development
- transfer
- use and customization



The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation by Ikujiro Nonaka, Hirotaka Takeuchi, 1995

What is KM

Knowledge Management is a systematic approach to improve the way organizations, groups and individuals handle knowledge in all forms, in order to improve effectiveness, innovation and quality.

Knowledge Management aims to transform the intellectual capital of an organization –stored organizational knowledge and tacit knowledge of individuals - into a new corporate value resulting in increased productivity and improved competitiveness. KM teaches all members of an organization how to optimize existing knowledge and how to generate new knowledge as a **collective entity**.

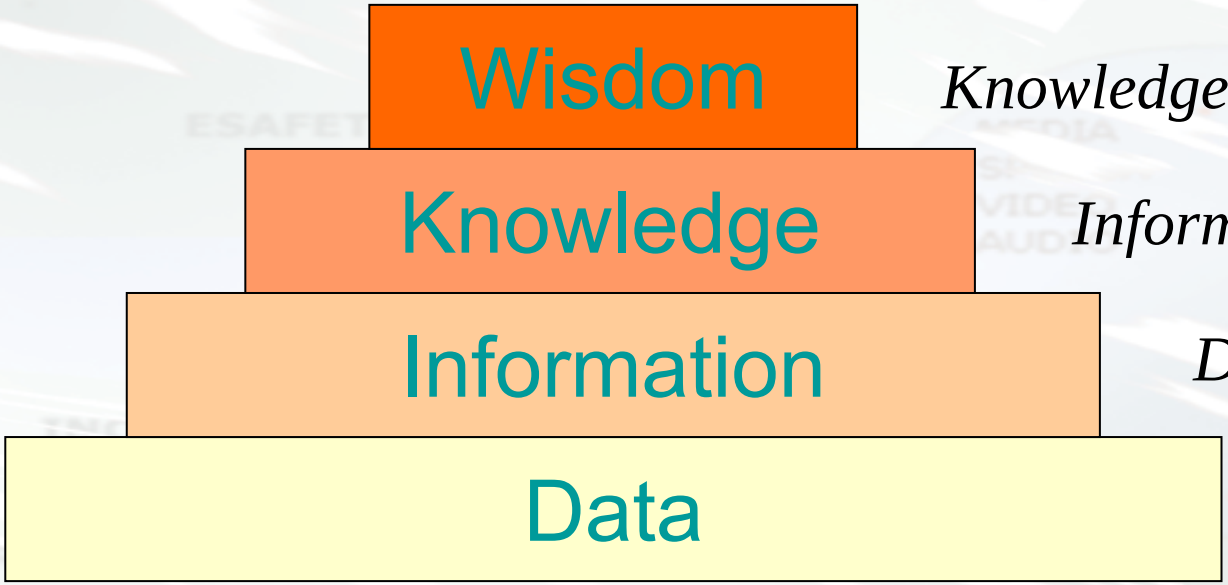
What is knowledge

- Knowledge is a model of (a part of) the reality as perceived by an agent
- Pragmatic definition: Knowledge is the information that confirms itself in use
  - Knowledge can not be uniquely defined, as the definition depends on the characteristics and goals of the organization
  - Knowledge is embedded in organizational processes, products and services

What is knowledge

- Principles
  - Knowledge is expensive to acquire, cheap to exploit
  - Property rights for knowledge are hard to define: IPR
  - Using knowledge does not mean wearing it out, knowledge grows and becomes richer through its use
  - Sharing knowledge with others does not imply losing it, knowledge evolves and multiplies through sharing

Data-Wisdom Pyramid



Knowledge plus experience

Information plus rules

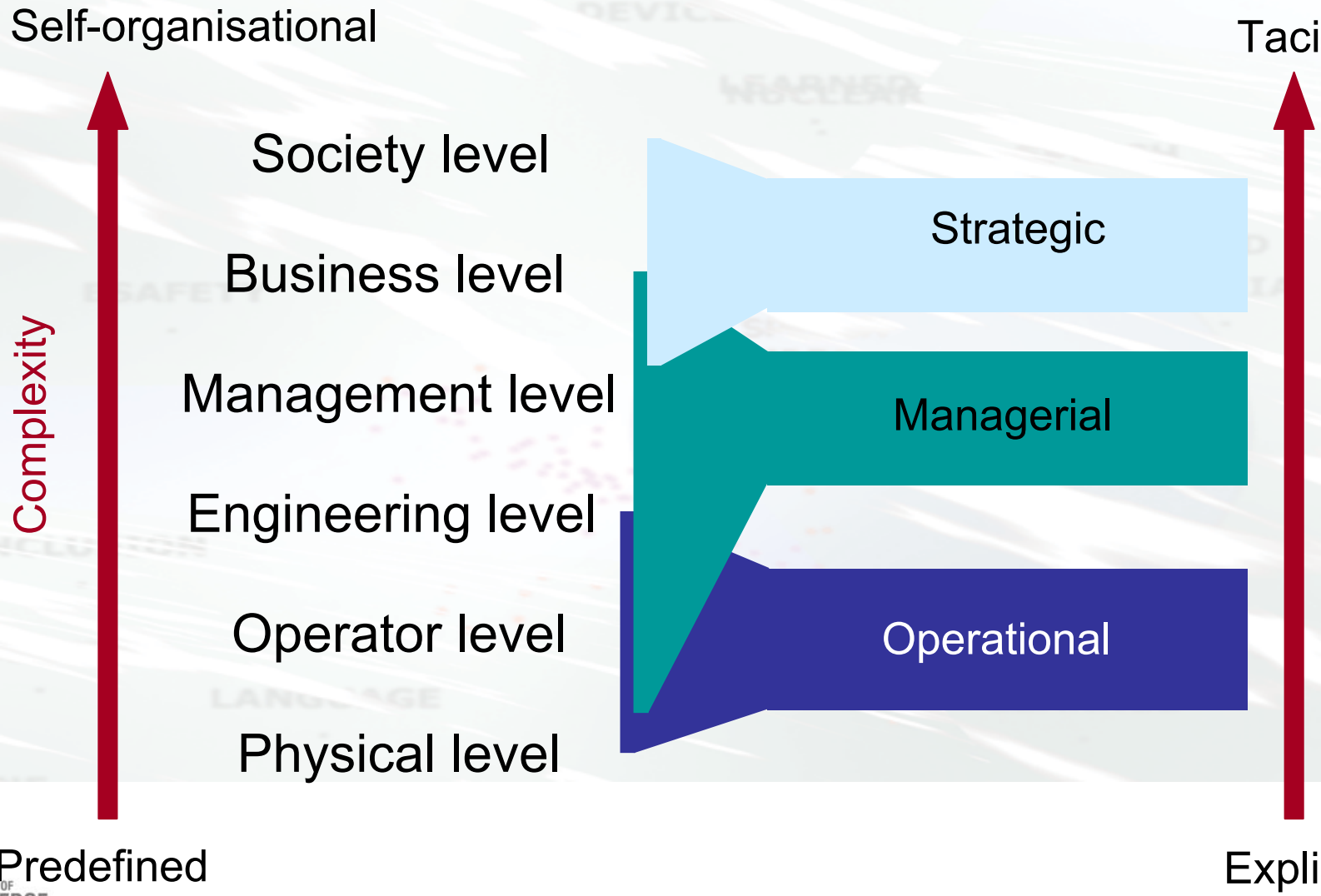
Data plus context

Observations, measurements

Data, information, knowledge

- Data is an individual observation or measurement, that yet needs to be interpreted
- Information is interpreted data – it is ”the difference, which makes the difference”
- Knowledge is the structure from which the meaning of information can be derived (“why” and “what for”) - nothing can become information without pre-knowledge (background knowledge)

Knowledge and society



Tacit / implicit vs. Explicit / codified knowledge

- Tacit (silent, mute), Implicit (can not be explicitly articulated)
  - formed of experiences, values, judgments and skills, enabling autonomous triggering and performance of actions. Hard to verify and accept. Two strategies:
    - try making tacit knowledge explicit
    - enable free flow of tacit knowledge
- Explicit (can be explicitly articulated), Codified (explicit, articulated in a specific language)
  - Encoding enables knowledge transfer, provided that the recipient knows the tacit ingredients of encoding used by the encoder
- Knowledge continuum, with barriers to knowledge encoding
  - costs of acquisition of implicit knowledge, codification, learning, problems of misunderstanding and misinterpreting

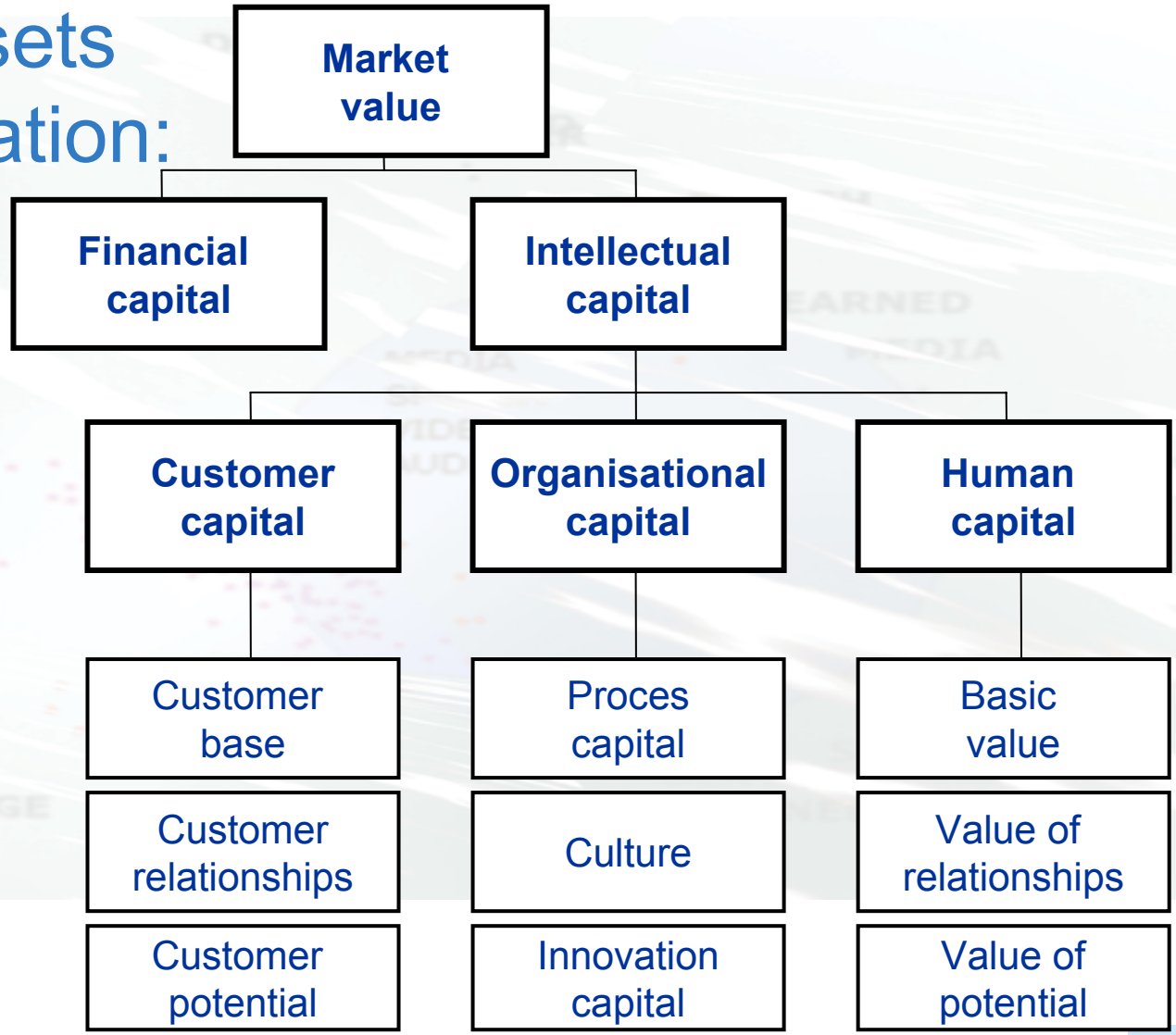
KM in New Economy

- KM: Traditional view
- KM: View shift
  - Information Society - Knowledge society
    - 10% of workforce produces all needed food and material goods, decreased dependence from natural resources (synthetic materials, decoding of human genome, ...), globalization and ease of accessing knowledge through new media, increased amount of people dealing with symbolic descriptions of things rather than things themselves (knowledge workers)
  - New economy - Knowledge economy
    - services rather than production, human networking, large corporations, virtual organizations, rapid changes, lifelong learning, knowledge as a source of intellectual capital

KM in New economy: Intellectual capital

Intangible assets of an organization:

- Internal (organizational structure)
- External (customer structure)
- Human (personnel competencies)



Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis: A case study in semi-automated trust modeling
  - Text mining: A case study in semi-automated ontology construction
  - OntoGen: semi-automated ontology construction tool

KM in New economy: A Networked Organizations Perspective

- **eBusiness, eScience, eMedicine, …**
  - doing business, science, medicine, … in a collaborative setting, supported by new media and computer networks
- **Networked organizations (NOs)**
  - non-static e-collaborative networks of individuals/organizations, enabled by information and communication technologies

NO infrastructures: New media

Infrastructures for KM: New media for eBusiness, eScience, ...

- **New media:** A generic term for many different forms of electronic communications and services that are made possible through the use of Internet technologies
- The term is in relation to “old” media forms, such as newspapers, magazines, radio diffusion and TV

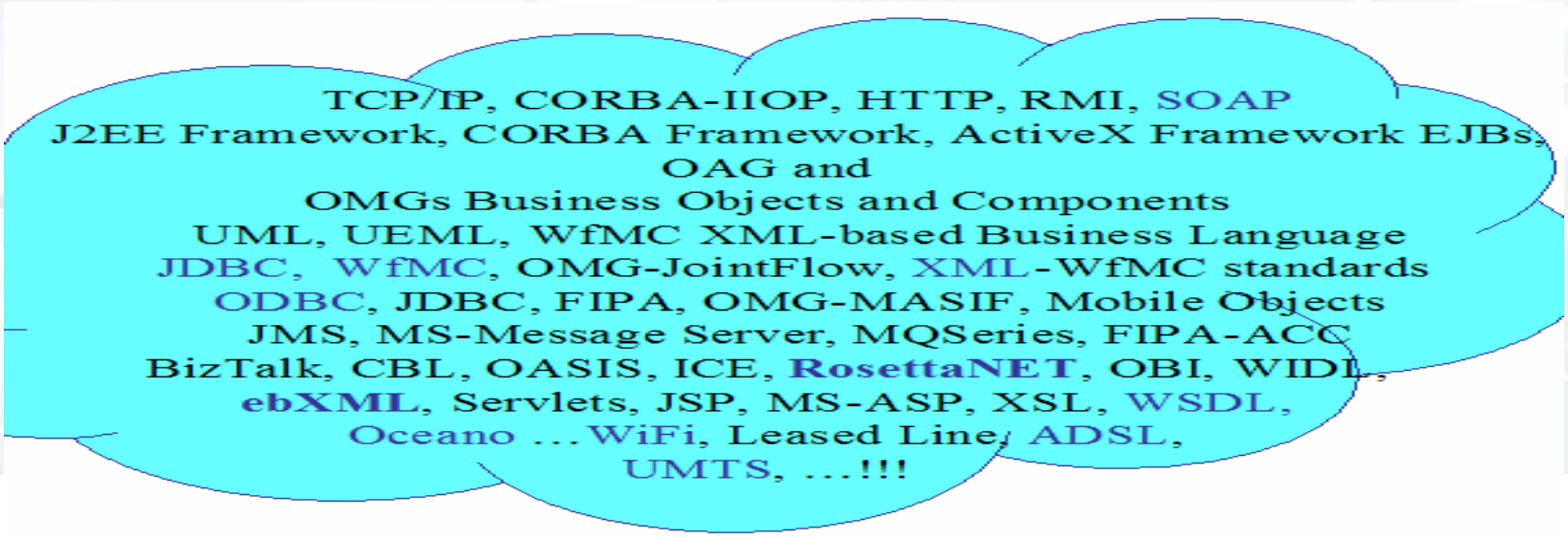
NO infrastructures: New media

- **Infrastructures:**
  - Networks (computer, satellites and telephone networks, cables, …)
  - Digital devices (DVD, CD-ROM, mobile telephones, wearable computers, …)
- **Services:**
  - WWW, internet, intranet, grid computing
  - streaming audio and video
  - chat rooms
  - e-mail
  - online communities
  - Web advertising
  - virtual reality environments
  - integration of digital data with the telephone, such as Internet telephony,
  - digital contents, digital libraries
  - mobile computing, wearable computing, ambient intelligence
  - …

NO infrastructures: Computer networks

Infrastructures for KM: Computer networks for eBusiness, eScience, ...

– ICT technologies, protocols and standards

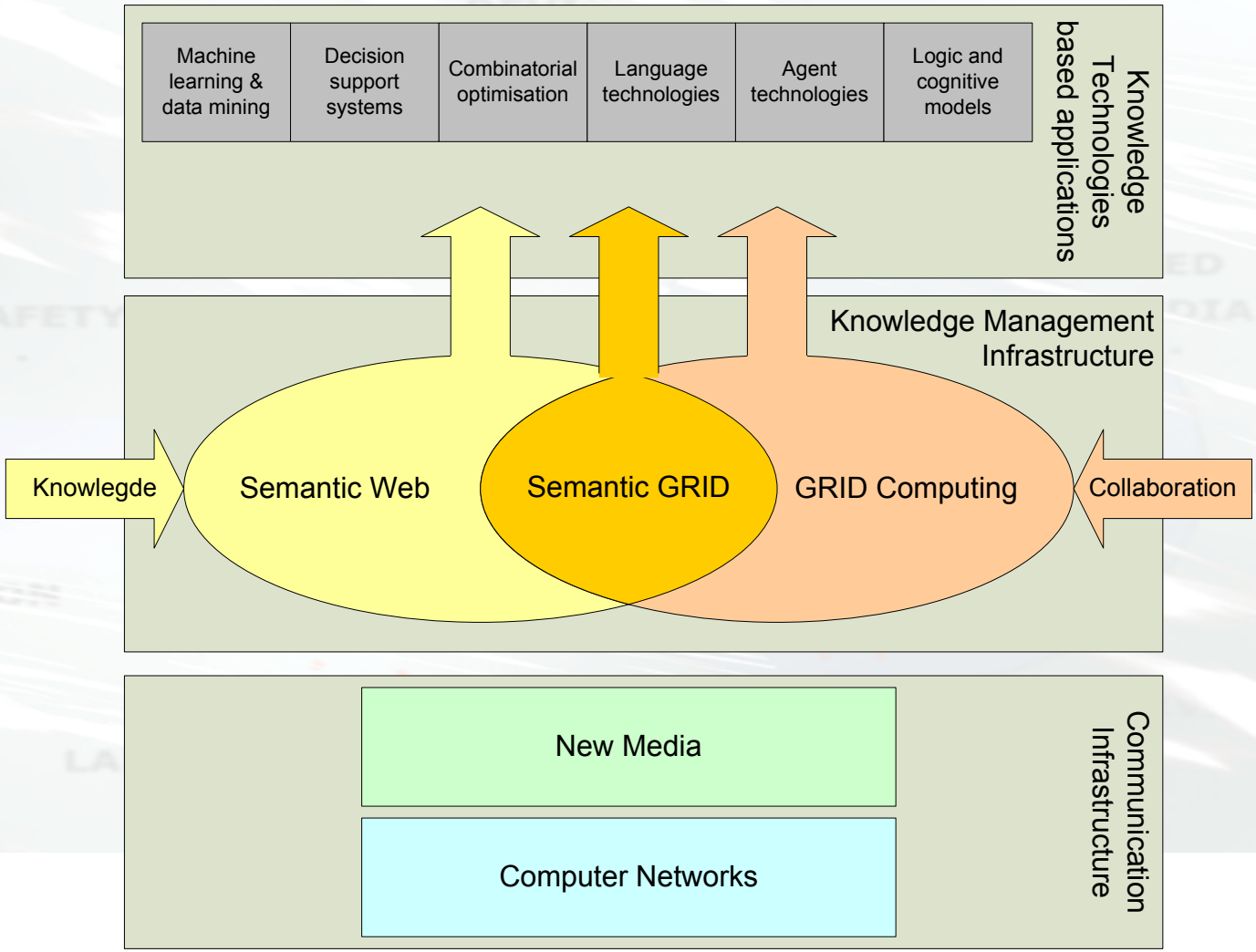


NO infrastructures: Towards the semantic grid

Infrastructures for KM: Semantic grid for eBusiness, eScience, ...

- **Grid computing:** coordinated resource sharing in dynamic, multi-institutional virtual organizations
- **Semantic Web:** extension of the current Web in which information is given a well-defined meaning, enabling data sharing and reasoning
- **Semantic grid:** extension of the current Grid in which information and services are given a well-defined meaning, enabling computers and people to work in collaboration

NO infrastructures and Knowledge technologies



Network economy

- **Network activities are facilitated by the use of shared infrastructure and standards, decreasing risk and costs**
- **Benefits of network membership increase by the number of other individuals and organizations in the network - the larger the network the better:**
  - a larger network is more competitive
  - has greater benefit of applications development
  - stimulates the speed and amount of learning and adapting of new technologies.
  - generates positive feedback where success generates success

Network economy

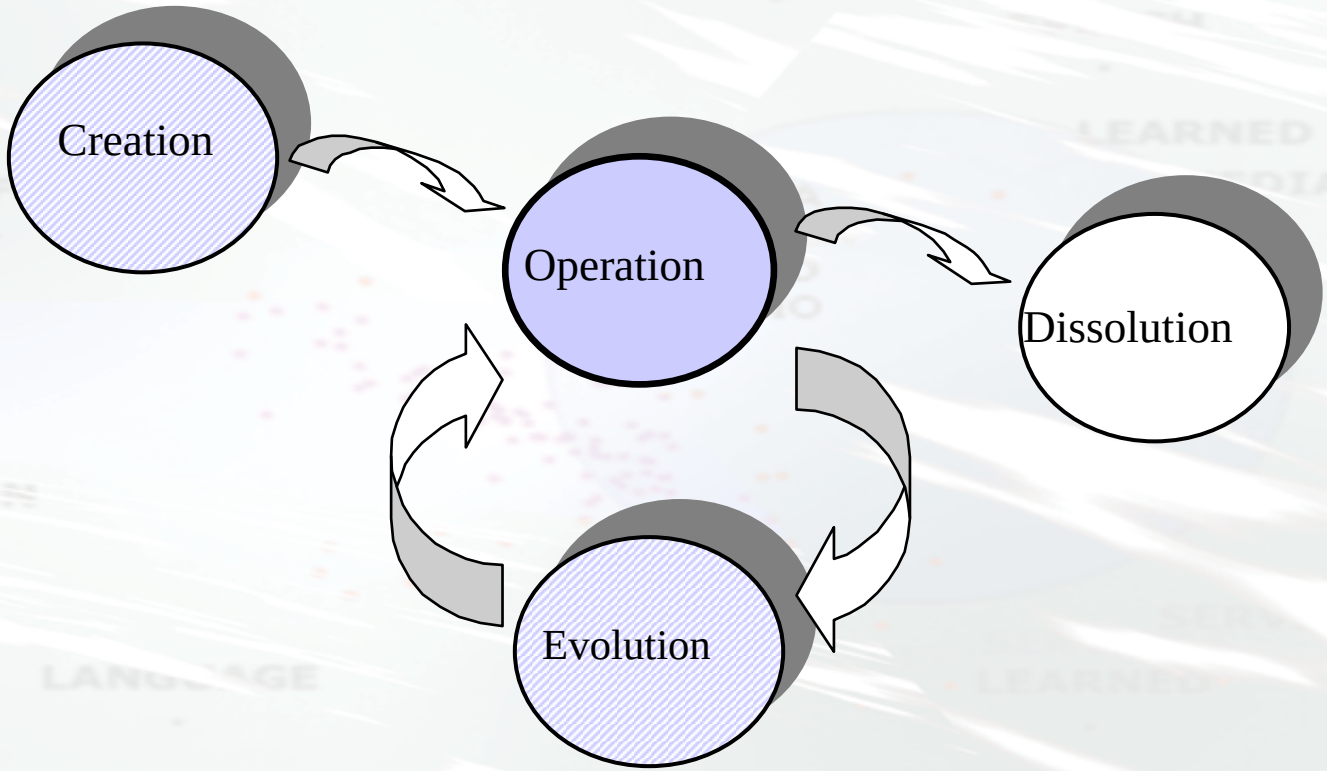
- **But: large networks are more complex to manage:**
  - increased complexity of the business environment and knowledge
  - managing processes instead of resources
  - agents as a source of knowledge
- **A partner in a NO can be viewed as an agent, capable of performing particular tasks**
- **The directing role is performed by an agent (net broker) acting as project leader in the process of:**
  - creating a virtual organization (VO) for a new project
  - planning, leading and controlling processes in a VO

Networked Organizations

- **Networked organizations (NO)** are non-static e-collaborative networks, enabled by information and communication technologies
- **Types of NO**
  - **Virtual organization (VO)**
  - **Virtual organization breeding environment (VBE)**
    - a cluster/association of organizations willing to collaborate
    - **VOs** are formed from **VBE** when a new business opportunity arises
  - **Professional virtual community (PVC)**

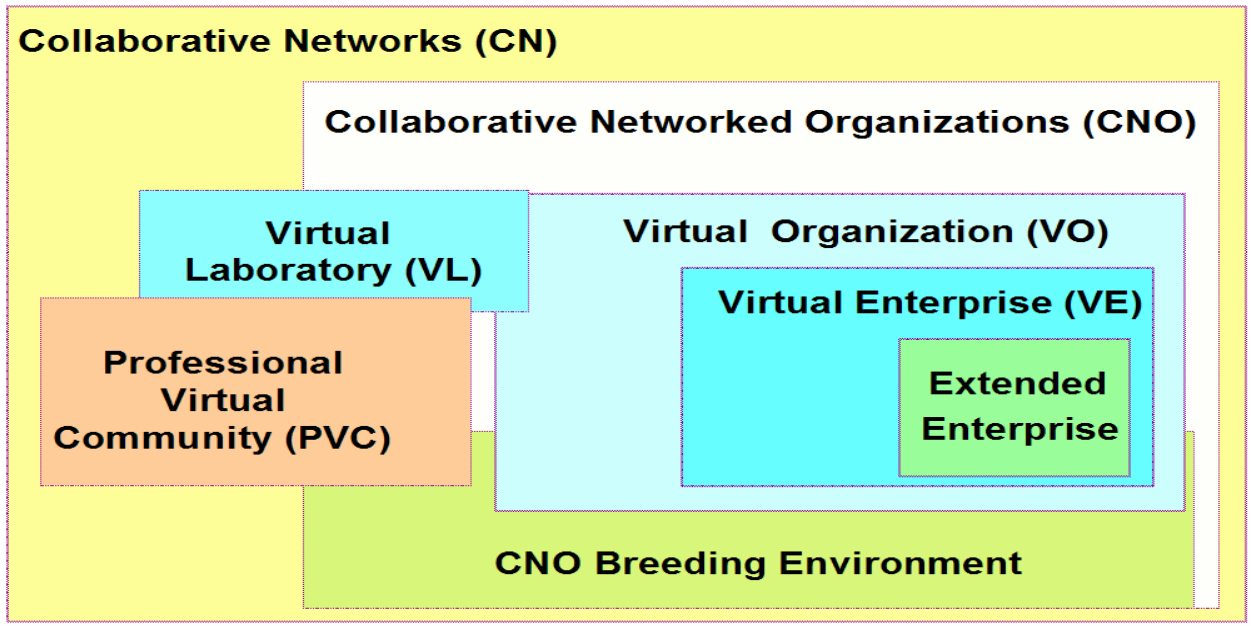
Networked organizations

A typical networked organization lifecycle



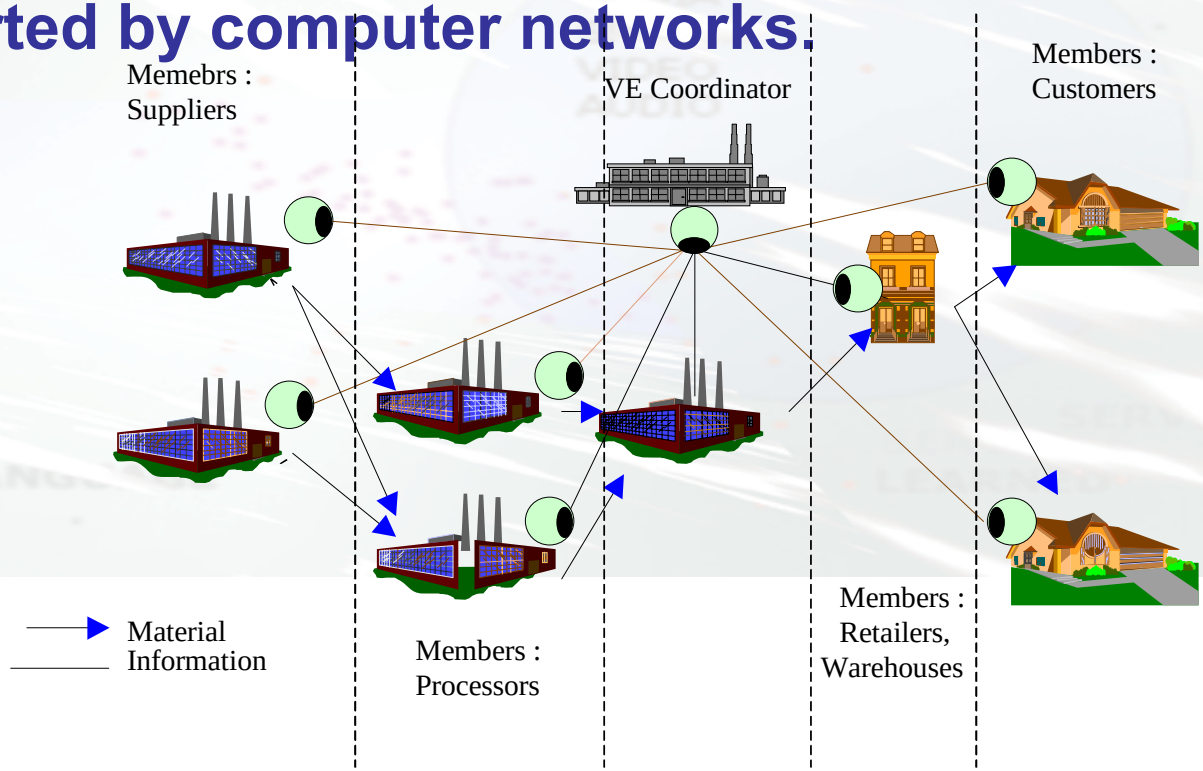
Networked Organizations

UNDERLYING CONCEPTS



Networked Organizations

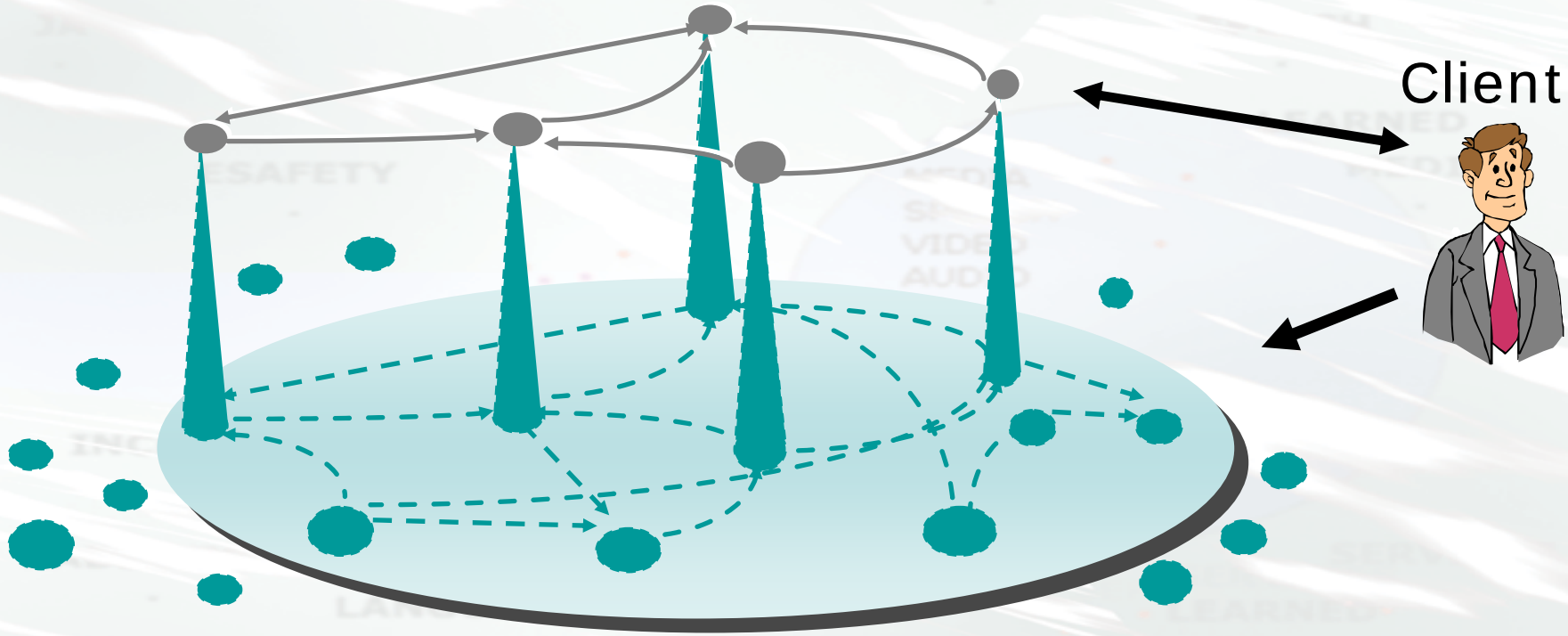
Virtual organization (VO) is a temporary alliance of enterprises/organizations that come together to share skills or core competencies and resources in order to better respond to business opportunities, and whose cooperation is supported by computer networks.



Networked Organizations

- **Virtual Organization Breeding Environment (VBE) represents an association or pool of agents -** organizations, supporting institutions, and individuals **- that have the potential and interest to cooperate.**
- **VBE is an establishment of a base long-term cooperation agreement**
- **When a business opportunity is identified by one member (acting as a broker), a subset of these organizations can be selected to form a VO**

Networked Organizations



Virtual organization Breeding Environment (VBE)

(Loss, 2005 – adapted from Bollhalter, 2004)

KM in NOs

- **Several problems occur:**
  - efficient storage of partners competencies
  - updating, sharing, promoting and transferring of these competencies
- Solved by adequate **knowledge management** using **knowledge technologies**
- **Knowledge map** - a knowledge resource repository is a necessity
  - each partner must have access
  - storing knowledge resources, process costs, resource availability

Introduction to KM: Outline

- What is KM: A traditional view
- KM in New economy: A Networked Organizations (NOs) perspective
- Selected knowledge technologies for KM in NOs:
  - Social Network Analysis: A case study in semi-automated trust modeling
  - Text mining: A case study in semi-automated ontology construction
  - OntoGen: semi-automated ontology construction tool

Knowledge technologies for Knowledge Mapping

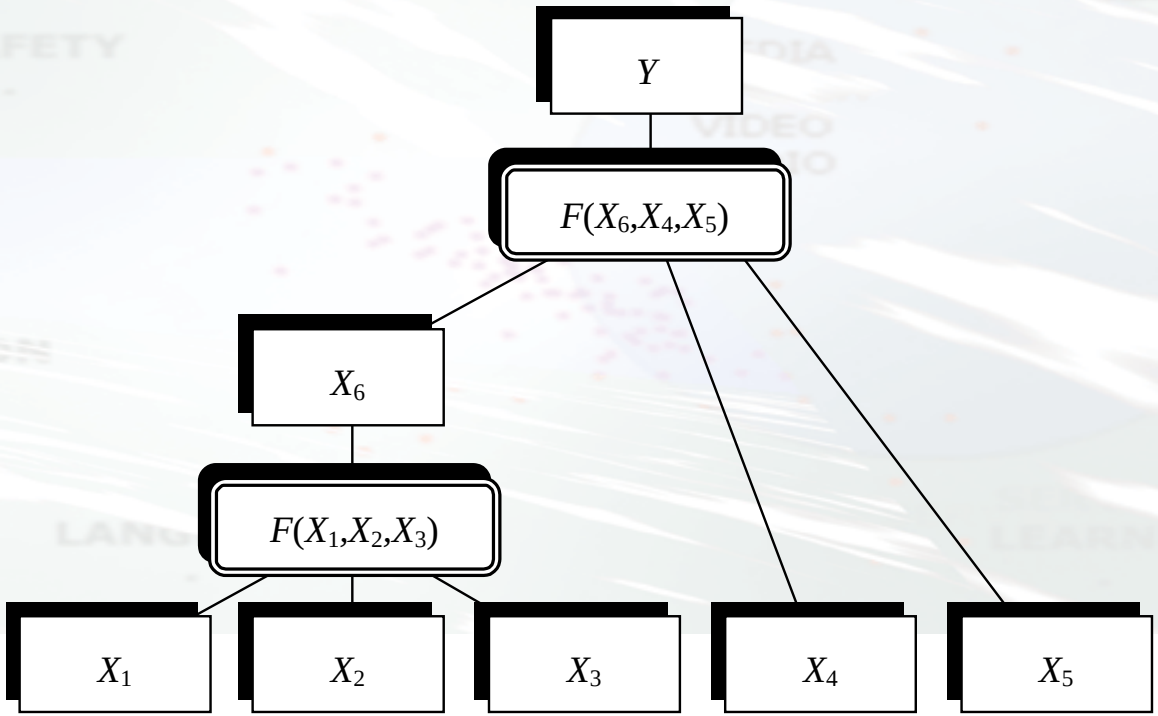
- **Automatic gathering tools:**
  - Web crawling
  - Information and keyword extraction
  - Language technologies (lemmatisation, grammar, dictionary)
- **Data storage** - relational database technology
- **Data analysis and decision support**
  - Social Network Analysis
  - Data,Text and Web mining, clustering
  - Machine learning tools (classification trees,...)
  - Decision support systems and tools
- **Presentation**
  - Visualisation tools (text and data visualisation)
  - Social network visualisation and analysis tools

Knowledge technologies for knowledge mapping in trust modeling

- **Automatic gathering tools:**
  - ☑ Web crawling
    - Information and keyword extraction
    - Language technologies (lemmatisation, grammar, dictionary)
- **Data storage** - relational database technology
- **Data analysis and decision support**
  - ☑ Social Network Analysis
    - Data,Text and Web mining, clustering
    - Machine learning tools (classification trees,...)
  - ☑ Decision support systems and tools
- **Presentation**
    - Visualisation tools (text and data visualisation)
  - ☑ Social network visualisation and analysis tools

A questionnaire-based trust acquisition method

- **Modeling trust between partners (individuals, institutions) using multi-attribute decision support**



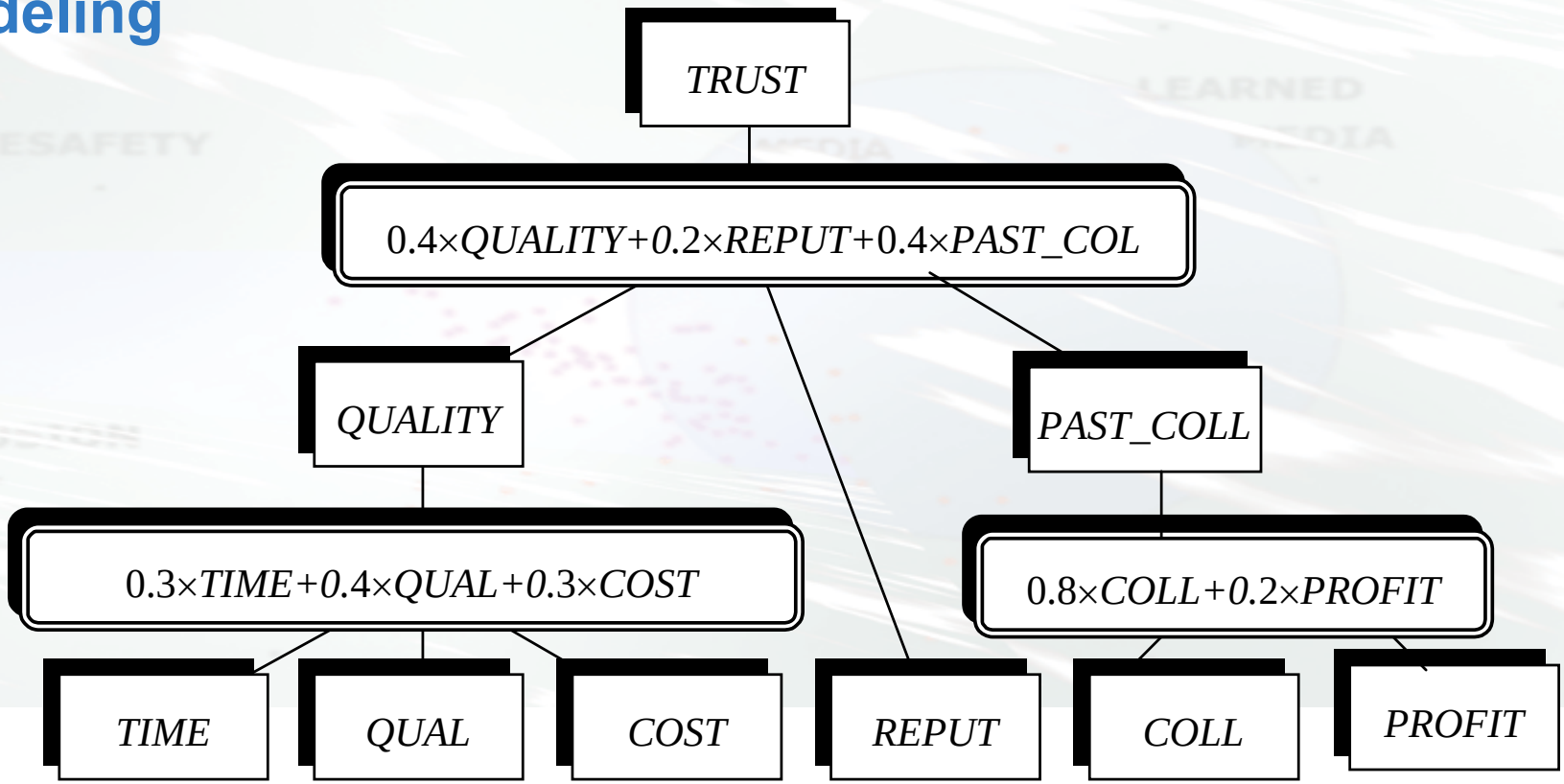
A questionnaire-based trust acquisition method

- E.g., Use user-defined features functions for trust modeling:
  - time
  - quality
  - cost
  - reputation
  - past collaborations
  - profit made in collaborations

$$NormalizedVal = \frac{ActualVal - MinVal}{MaxVal - MinVal}$$

A questionnaire-based trust acquisition method

- **User-defined features and utility functions for trust modeling**



Virtuelle Fabrik

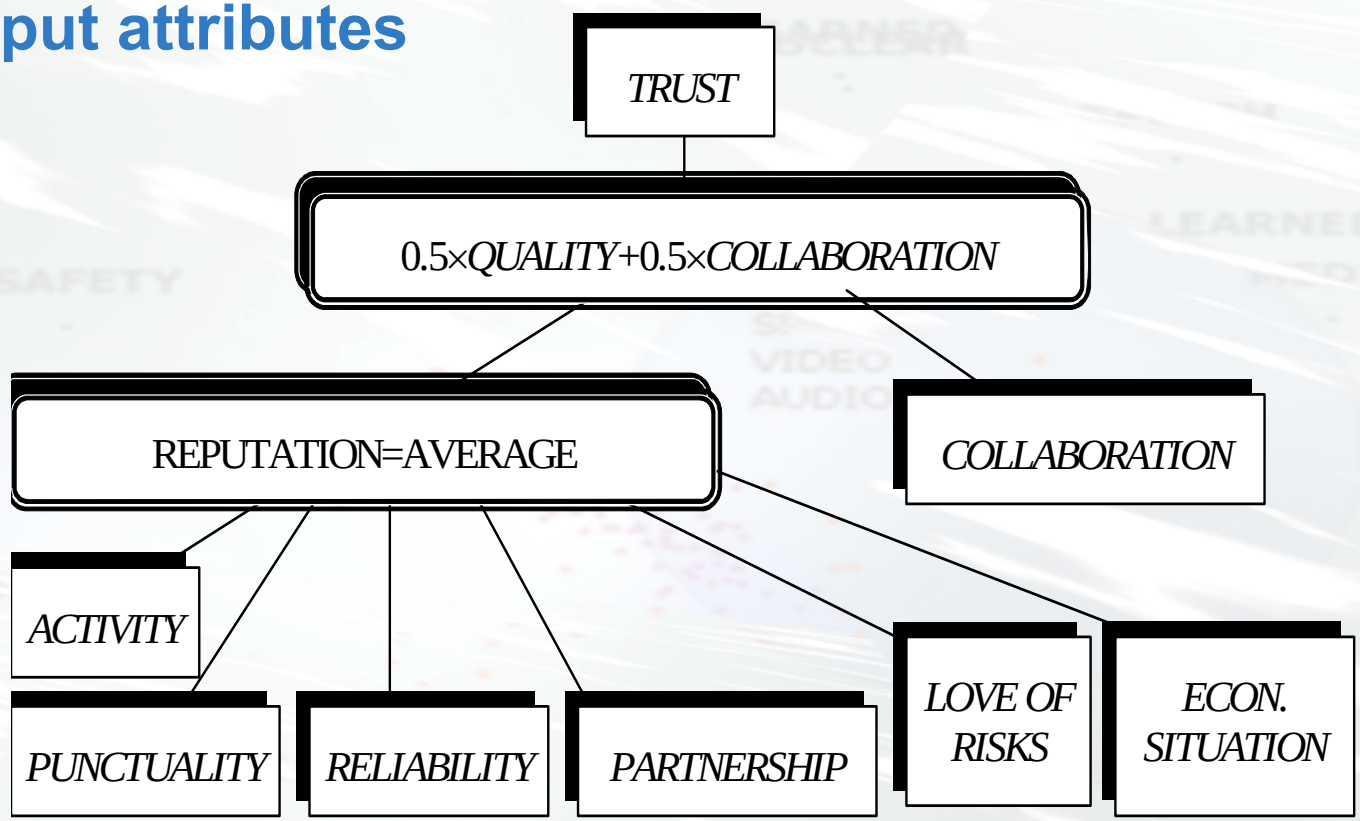
- a Swiss industrial cluster: Virtuelle Fabrik A.G., St. Gallen
- Cluster of partners from mechanical engineering industry
- http://www.virtuelle-fabrik.com
- collaborating expert: Stefan Bolhalter, a VF manager
- The goal of our project: Visualization of partners reputation and collaboration

Virtuelle Fabrik

- Reputation, each of properties has values from 1 to 6 (6 is very good, 1 is very bad)
  - activity
  - punctuality
  - reliability
  - partnership
  - love of risk
  - economical situation
- Collaboration:
  - matrix of collaboration, values from {1, 2, 3}

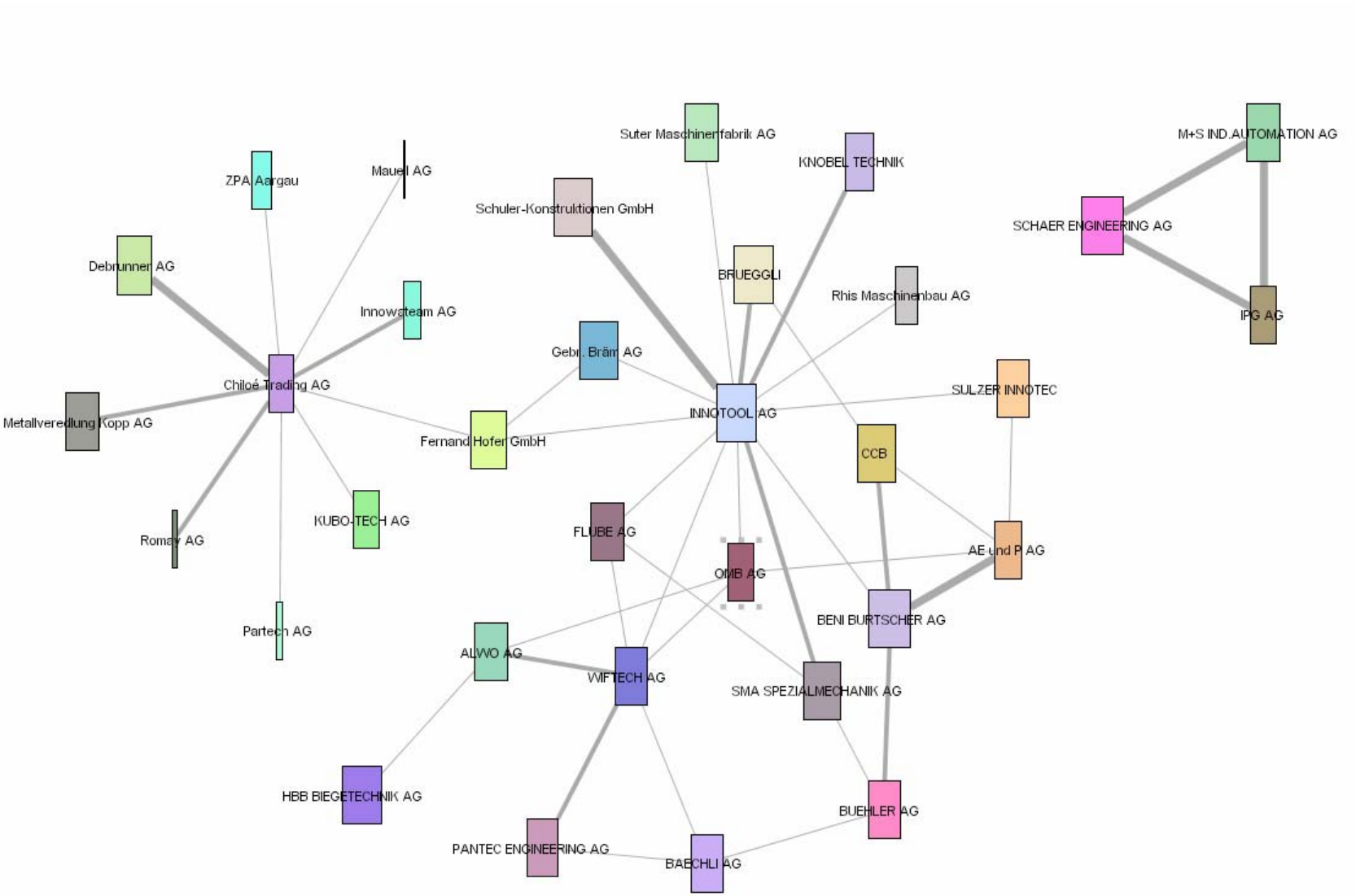
Virtuelle Fabrik

- **Reputation computed as the average of the basic input attributes**



- Other representations possible

Virtuelle Fabrik



Virtuelle Fabrik

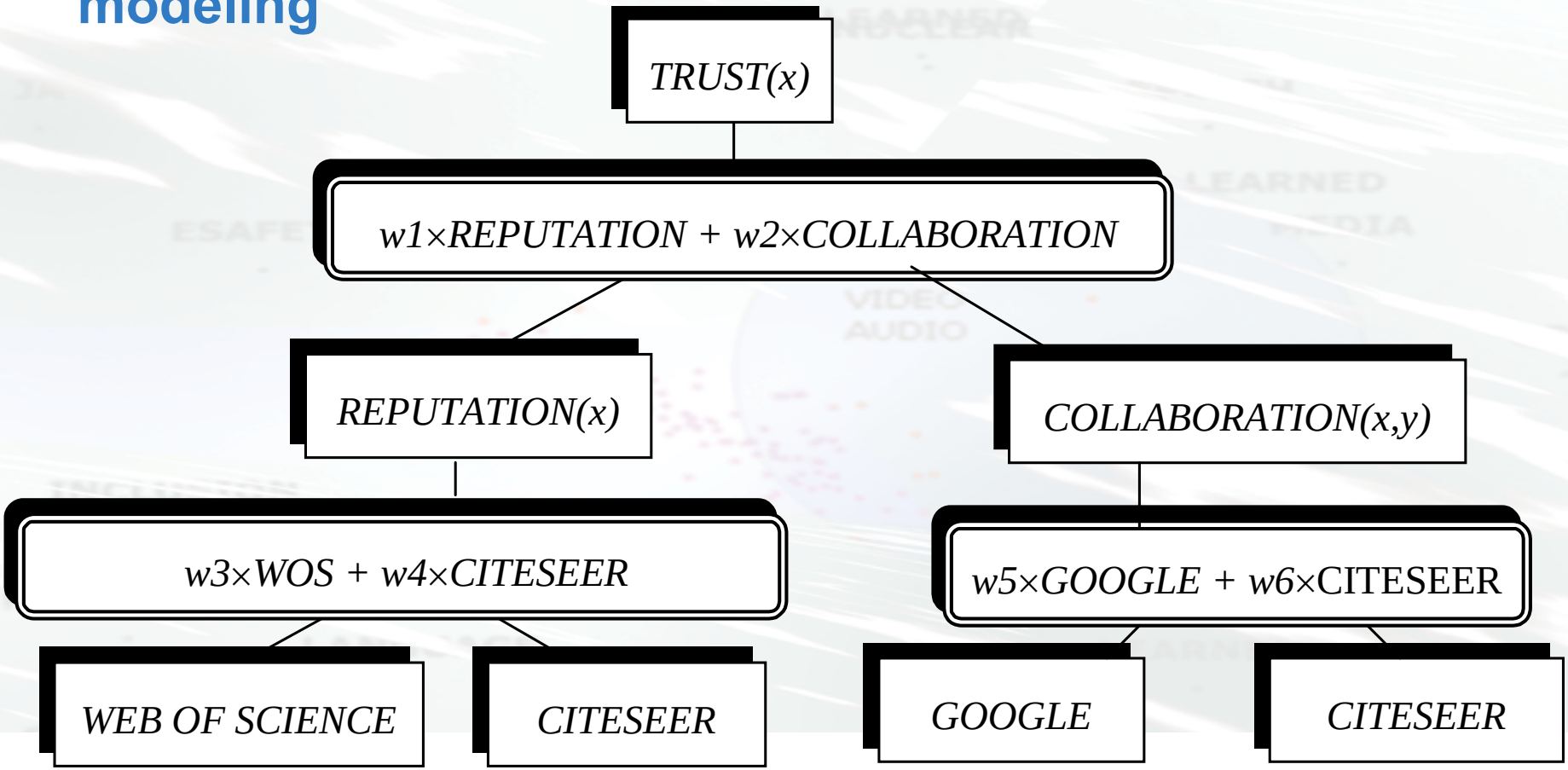
- The proposed decision support approach enables the evaluation and visualization of mutual trust between partners and can be used to find most trusted CNO partners in the process of creating a new VO
- The graph did not show new or surprising relationships to Stefan Bollhalter
- But the graph enabled him to visualize and confirm his knowledge about VF

Trust modeling through Web mining

- Analysis made for 102 individuals from 20 organizations participating in the ECOLEAD project
- Modeling trust between partners (individuals, institutions)
- Trust modeled from two components:
  - Reputation: measured by the # of papers published in SCI journals and # of SCI citations
  - Collaboration: measured by the # of joint papers and # of name co-occurrences on the web

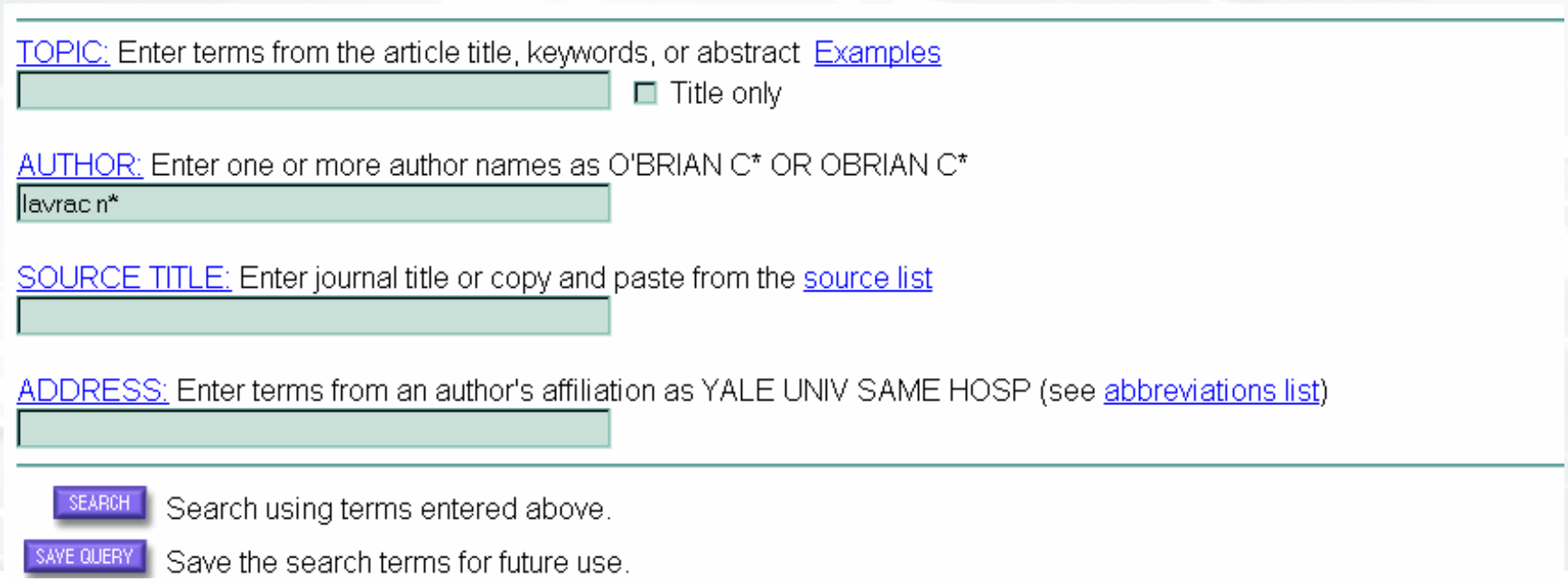
“Trust” computation

- **User-defined features and utility functions for trust modeling**



Reputation

- Citation index
- Taken from:
  - Web of Science, http://wos.izum.si
  - Citeseer, http://citeseer.ist.psu.edu



Reputation

- Citation index
- Taken from:
  - Web of Science, http://wos.izum.si
  - Citeseer, http://citeseer.ist.psu.edu

Collaboration

- Number of co-occurrences in:
  - Citeseer, http://citeseer.ist.psu.edu
  - Google, http://www.google.com

CiteSeer Scientific Literature Digital Library Find: (n lavrac OR nada lavrac) AND (d mlad Documents Citations

Searching for **(n lavrac or nada lavrac) and (d mladenic or dunja mladenic)**.
Restrict to: Header Title Order by: Expected citations Hubs Usage Date Try: Google (CiteSeer) Google (Web) CSB DBLP
2 documents found. **Order: number of citations.**

Feature Subset Selection in Association Rules Learning Systems - Viktor Jovanoski Nada (Correct)
the research. References 1] V. Jovanoski and **N. Lavrac**, Using Association rules for Inductive Concept
Rules Learning Systems Viktor Jovanoski, **Nada Lavrac** Jozef Stefan Institute Jamova 39, 1000
www-ai.ijs.si/MarkoGrobelnik/awamida99/jovanoski.ps

Strojno Ucenje Na Nehomogenih, Distribuiranih Tekstovnih Podatkih - Mladenic (1998) (Correct)
Conference on Machine Learning ICML95. 63] **Lavrac, N**. Dzeroski, S. 1994)Inductive Logic
would like to express my thanks to Assist. Prof. **Nada Lavrac**, my advisor at J. Stefan Institute for her
Conference on Machine Learning ECML98. 79] **Mladenic, D**. 1998)Turning Yahoo into an Automatic
www.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/pww/papers/PhD/PhDFinal.ps.gz

Try your query at: Google (CiteSeer) Google (Web) CSB DBLP

CiteSeer - Copyright NEC and IST

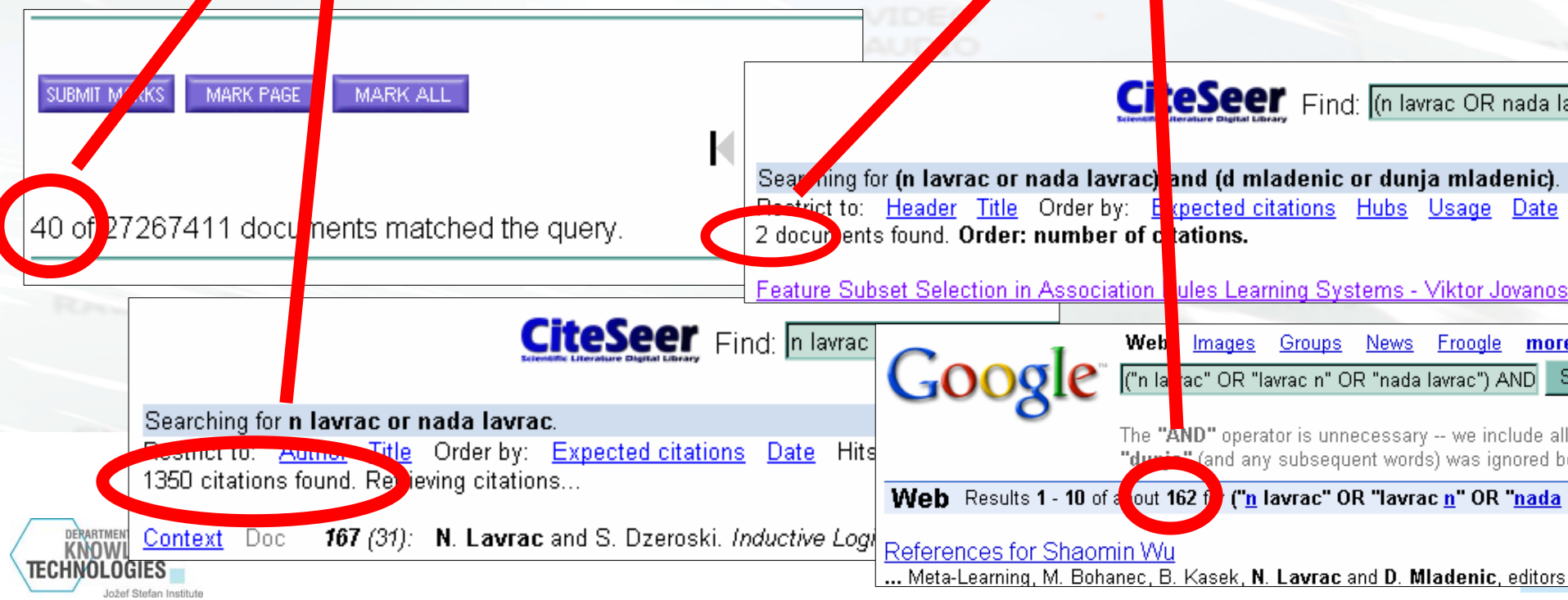
Collaboration

- Number of co-occurrences in:
  - Citeseer, http://citeseer.ist.psu.edu
  - Google, http://www.google.com

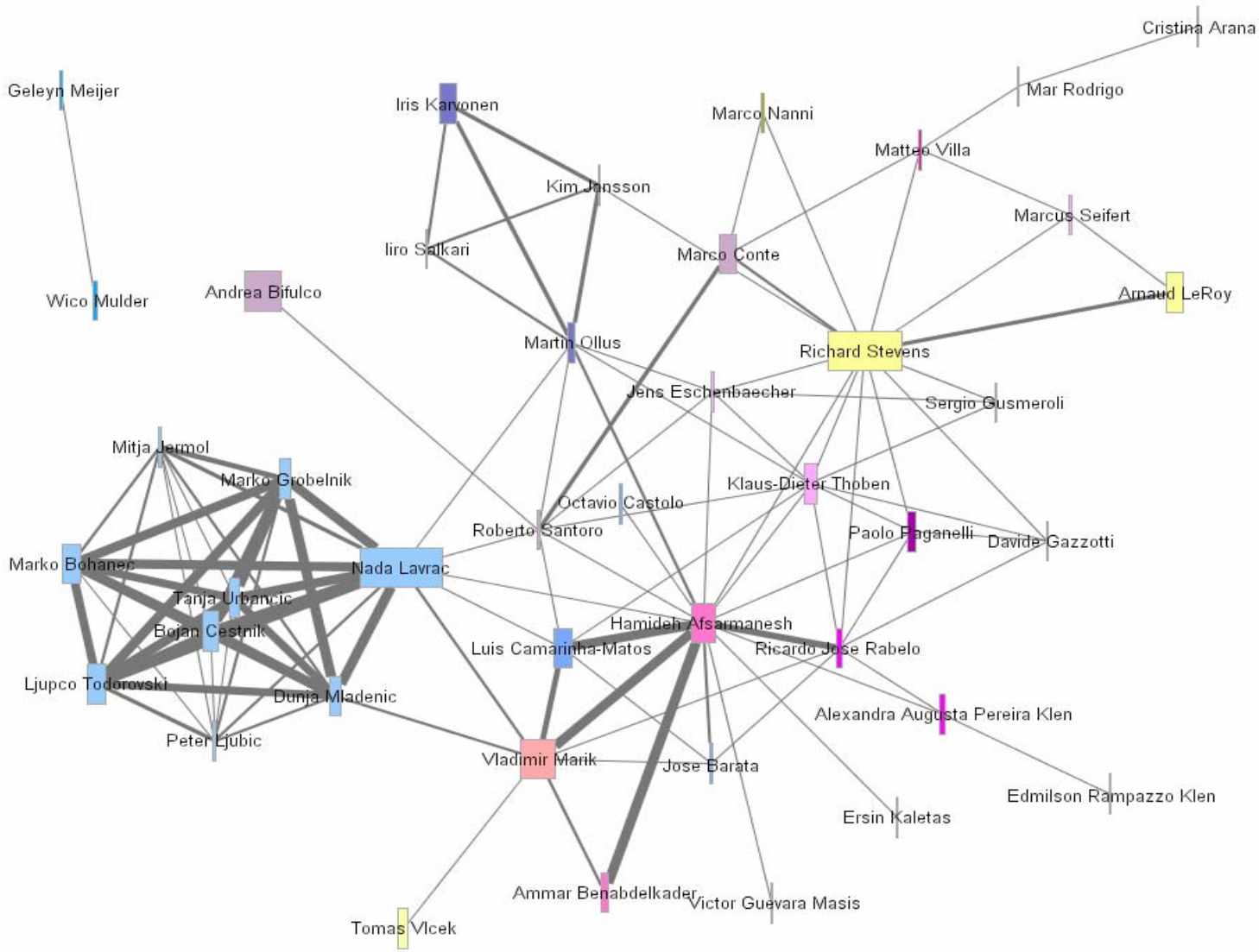
Trust

Reputation

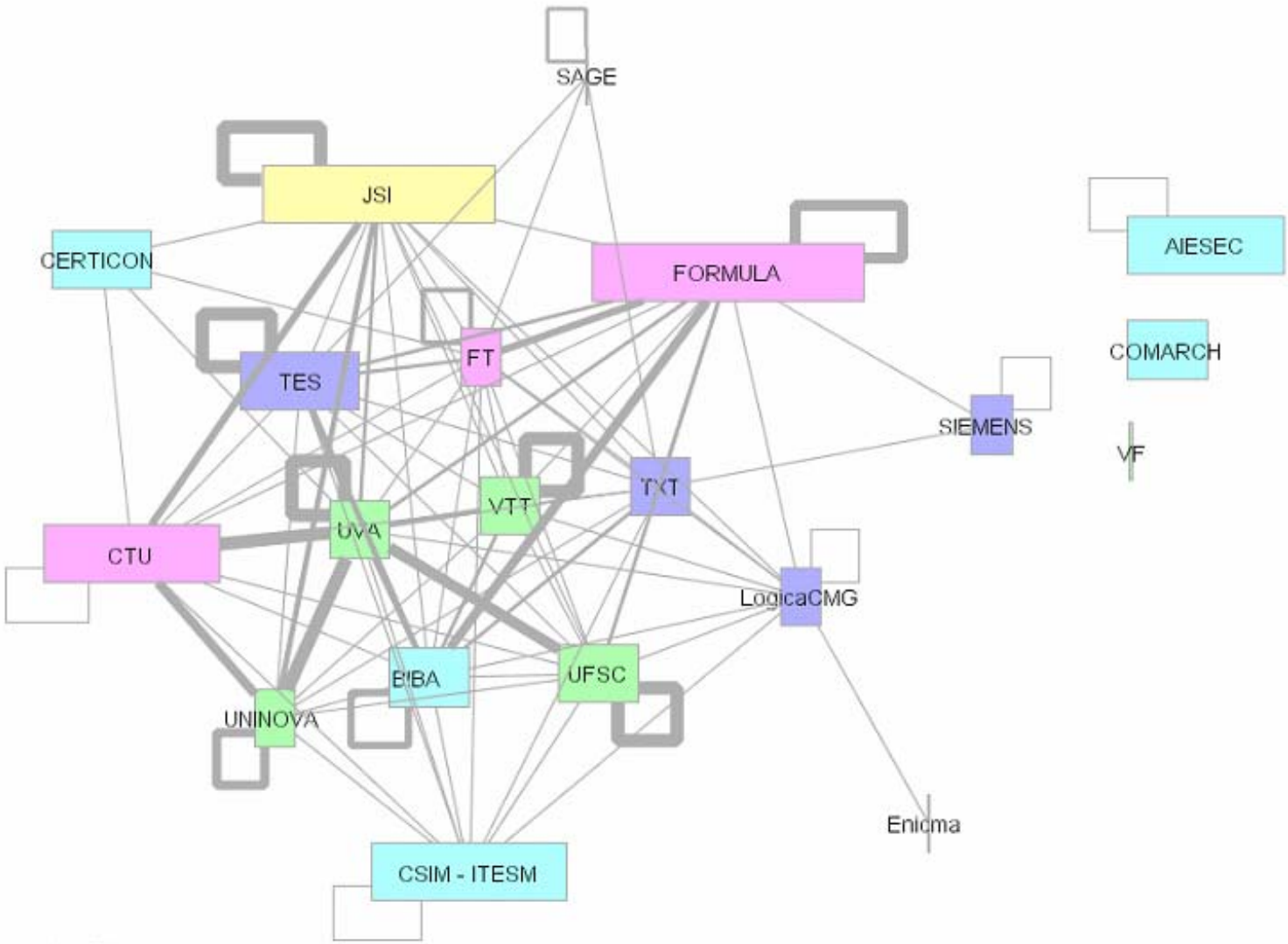
Collaboration



“Trust” between individuals



“Trust” between institutions



Knowledge technologies for knowledge mapping in ontology construction

- **Automatic gathering tools:**
  - ☑ Web crawling
  - ☑ Information and keyword extraction
  - ☑ Language technologies (lemmatisation, grammar, dictionary)
- **Data storage** - relational database technology
- **Data analysis and decision support**
  - Social Network Analysis
  - ☑ Data,Text and Web mining, clustering
  - Machine learning tools (classification trees,...)
  - Decision support systems and tools
- **Presentation**
  - ☑ Visualisation tools (text and data visualisation)
  - Social network visualisation and analysis tools

Ontology construction experiment: Structuring and visualization of NO competencies

- Approach: Applying knowledge mapping tools for competency visualization and structuring of competencies of partners of the Virtuelle Fabrik Swiss industrial cluster

Structuring and visualization of VF competencies

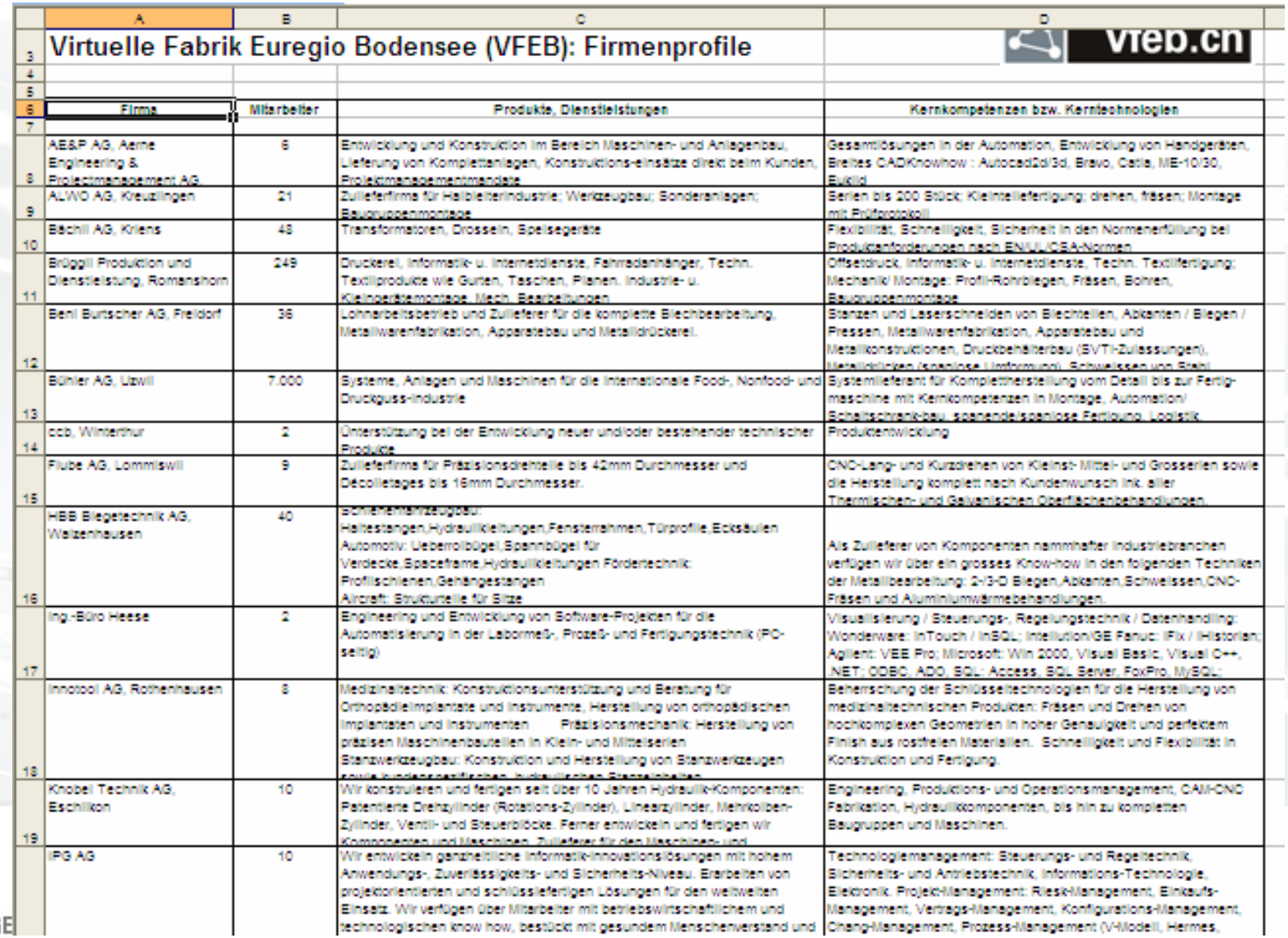
- Structuring the expertise of companies: Analysis of VF partners business data (a subset of VF industrial cluster - 20 companies from the Bodensee sub-cluster)

Table 2 - Company identifiers assigned to company names.

1 AE&P	2 ALWO	3 Bachli	4 Bruggli	5 Beni
6 Buchler	7 Ccb	8 Flube	9 KBB	10 Heese
11 Innotool	12 Knobel	13 IPG	14 M+S	15 OMB
16 Pantec	17 Schar	18 SMA	19 Sulzer	20 Wiftech

- Our approach: Apply hierarchical k-means document clustering and visualization

Descriptions of 20 VF partners



	A	B	C	D
3	Virtuelle Fabrik Euregio Bodensee (VFEB): Firmenprofile			vfeb.cn
4				
5				
6	Firma	Mitarbeiter	Produkte, Dienstleistungen	Kernkompetenzen bzw. Kerntechnologien
7				
8	AE&P AG, Aerne Engineering & Projectmanagement AG	6	Entwicklung und Konstruktion im Bereich Maschinen- und Anlagenbau, Lieferung von Komplettanlagen, Konstruktions-einsätze direkt beim Kunden, Projektmanagementmandate	Gesamtlösungen in der Automation, Entwicklung von Handgeräten, Breites CADKnowhow : Autocad2d/3d, Bravo, Catia, ME-10/30, Euklid
9	ALWO AG, Kreuzlingen	21	Zulieferfirma für Halbleiterindustrie; Werkzeugbau; Sonderanlagen; Baugruppenmontage	Serien bis 200 Stück; Kleinteilefertigung; drehen, fräsen; Montage mit Prüfprotokoll
10	Bächli AG, Kriens	48	Transformatoren, Drosseln, Speisegeräte	Flexibilität, Schnelligkeit, Sicherheit in den Normenerfüllung bei Produktentwicklungen nach EN/IEC/CSA-Normen
11	Brüggli Produktion und Dienstleistung, Romanshorn	249	Druckerei, Informatik- u. Internetdienste, Fahrradanhänger, Techn. Textilprodukte wie Gurten, Taschen, Planen, Industrie- u. Kleingerätemontage, Mech. Bearbeitungen	Offsetdruck, Informatik- u. Internetdienste, Techn. Textilfertigung; Mechanik/ Montage: Profil-Rohrbiegen, Fräsen, Bohren, Baugruppenmontage
12	Beni Burtscher AG, Freidorf	36	Lohnarbeitsbetrieb und Zulieferer für die komplette Blechbearbeitung, Metallwarenfabrikation, Apparatebau und Metalldrückerei.	Stanzen und Laserschneiden von Blechteilen, Abkanten / Biegen / Pressen, Metallwarenfabrikation, Apparatebau und Metallkonstruktionen, Druckbehälterbau (SVTI-Zulassungen), Metalldrücken (spanlose Umformung), Schweissen von Stahl
13	Bühler AG, Uzwil	7.000	Systeme, Anlagen und Maschinen für die internationale Food-, Nonfood- und Druckguss-Industrie	Systemlieferant für Komplettherstellung vom Detail bis zur Fertigmaschine mit Kernkompetenzen in Montage, Automation/ Schaltschrank-bau, spanende/spanlose Fertigung, Logistik
14	ccb, Winterthur	2	Unterstützung bei der Entwicklung neuer und/oder bestehender technischer Produkte	Produktentwicklung
15	Flube AG, Lommiswil	9	Zulieferfirma für Präzisionsdrehteile bis 42mm Durchmesser und Décolletages bis 16mm Durchmesser.	CNC-Lang- und Kurzdrehen von Kleinst- Mittel- und Grosserien sowie die Herstellung komplett nach Kundenwunsch inkl. aller Thermischen- und Galvanischen Oberflächenbehandlungen.
16	HBB Biegetechnik AG, Walzenhausen	40	Schienenfahrzeugbau: Haltestangen,Hydraulikleitungen,Fensterrahmen,Türprofile,Eckstäulen Automotiv: Ueberrollbügel,Spannbügel für Verdecke,Spaceframe,Hydraulikleitungen Fördertechnik: Profilschienen,Gehängestangen Aircraft: Strukturteile für Sitze	Als Zulieferer von Komponenten nammhafter Industriebranchen verfügen wir über ein grosses Know-how in den folgenden Techniken der Metallbearbeitung: 2-/3-D Biegen,Abkanten,Schweissen,CNC-Fräsen und Aluminiumwärmebehandlungen.
17	Ing.-Büro Heese	2	Engineering und Entwicklung von Software-Projekten für die Automatisierung in der Labormeß-, Prozeß- und Fertigungstechnik (PC-seitig)	Visualisierung / Steuerungs-, Regelungstechnik / Datenhandling: Wonderware: InTouch / InSQL; Intellution/GE Fanuc: iFix / iHistorian; Agilent: VEE Pro; Microsoft: Win 2000, Visual Basic, Visual C++, .NET; ODBC, ADO, SQL, Access, SQL Server, FoxPro, MySQL;
18	Innotool AG, Rothenhausen	8	Medizinaltechnik: Konstruktionsunterstützung und Beratung für Orthopädieimplantate und Instrumente, Herstellung von orthopädischen Implantaten und Instrumenten Präzisionsmechanik: Herstellung von präzisen Maschinenbauteilen in Klein- und Mittelserien Stanzwerkzeugbau: Konstruktion und Herstellung von Stanzwerkzeugen sowie kundenspezifischen hydraulischen Stanzpressen	Beherrschung der Schlüsseltechnologien für die Herstellung von medizinaltechnischen Produkten: Fräsen und Drehen von hochkomplexen Geometrien in hoher Genauigkeit und perfektem Finish aus rostfreien Materialien. Schnelligkeit und Flexibilität in Konstruktion und Fertigung.
19	Knobel Technik AG, Eschlikon	10	Wir konstruieren und fertigen seit über 10 Jahren Hydraulik-Komponenten: Patentierte Drehzylinder (Rotations-Zylinder), Linearzylinder, Mehrkolben-Zylinder, Ventil- und Steuerblöcke. Ferner entwickeln und fertigen wir Komponenten und Maschinen. Zulieferer für den Maschinen- und	Engineering, Produktions- und Operationsmanagement, CAM-CNC Fabrikation, Hydraulikkomponenten, bis hin zu kompletten Baugruppen und Maschinen.
	IPG AG	10	Wir entwickeln ganzheitliche Informatik-Innovationslösungen mit hohem Anwendungs-, Zuverlässigkeits- und Sicherheits-Niveau. Erarbeiten von projektorientierten und schlüsselfertigen Lösungen für den weltweiten Einsatz. Wir verfügen über Mitarbeiter mit betriebswirtschaftlichem und technologischen know how, bestückt mit gesundem Menschenverstand und	Technologiemanagement: Steuerungs- und Regeltechnik, Sicherheits- und Antriebstechnik, Informations-Technologie, Elektronik. Projekt-Management: Riesk-Management, Einkaufs-Management, Vertrags-Management, Konfigurations-Management, Change-Management, Prozess-Management (V-Modell, Hermes,

Text Mining: Levels of Text Processing

- **Word Level**
  - **Words Properties**
  - **Stop-Words**
  - **Stemming**
  - **Frequent N-Grams**
  - **Thesaurus (WordNet)**
- Sentence Level
- Document Level
- Document-Collection Level

Stemming and Lemmatization

- Different forms of the same word usually problematic for text data analysis
  - because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
  - usually treated as completely unrelated words
- Stemming is a process of transforming a word into its stem
  - cutting off a suffix (eg., smejala -> smej)
- Lemmatization is a process of transforming a word into its normalized form
  - replacing the word, most often replacing a suffix (eg., smejala -> smejati)

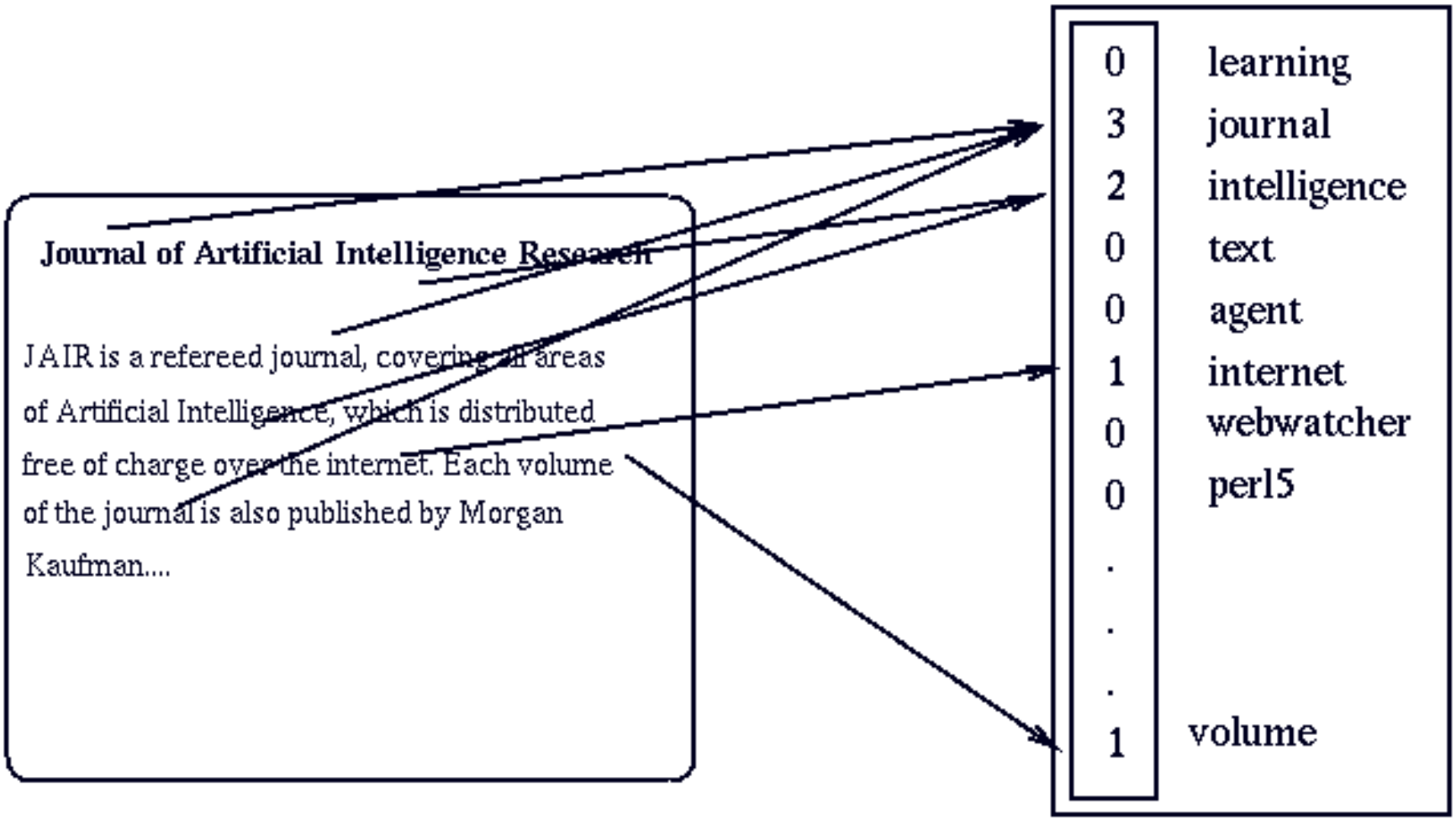
Stemming

- For English it is not a big problem - publicly available algorithms give good results
  - Most widely used is Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- In Slovenian language 10-20 different forms correspond to the same word:
  - ("to laugh" in Slovenian): smej, smejal, smejala, smejale, smejali, smejalo, smejati, smejejo, smejeta, smejete, smejeva, smeješ, smejemo, smejiš, smeje, smejoč, smejta, smejte, smejva

Text Mining: Levels of Text Processing

- Word Level
- Sentence Level
- Document Level
- Document-Collection Level
  - Representation
  - Feature Selection
  - Document Similarity
  - Categorization
  - Clustering
  - Summarization

Bag-of-words document representation



Word weighting

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

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

- 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 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.

- [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] … [STOCK:0.035] [YORK:0.035] [PCT:0.022] [MARCH:0.011]

Cosine similarity between document vectors

- Each document is represented as a vector of weights D = <x>
- Similarity between two vectors is estimated by the similarity between their vector representations (cosine of the angle between the two vectors):

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

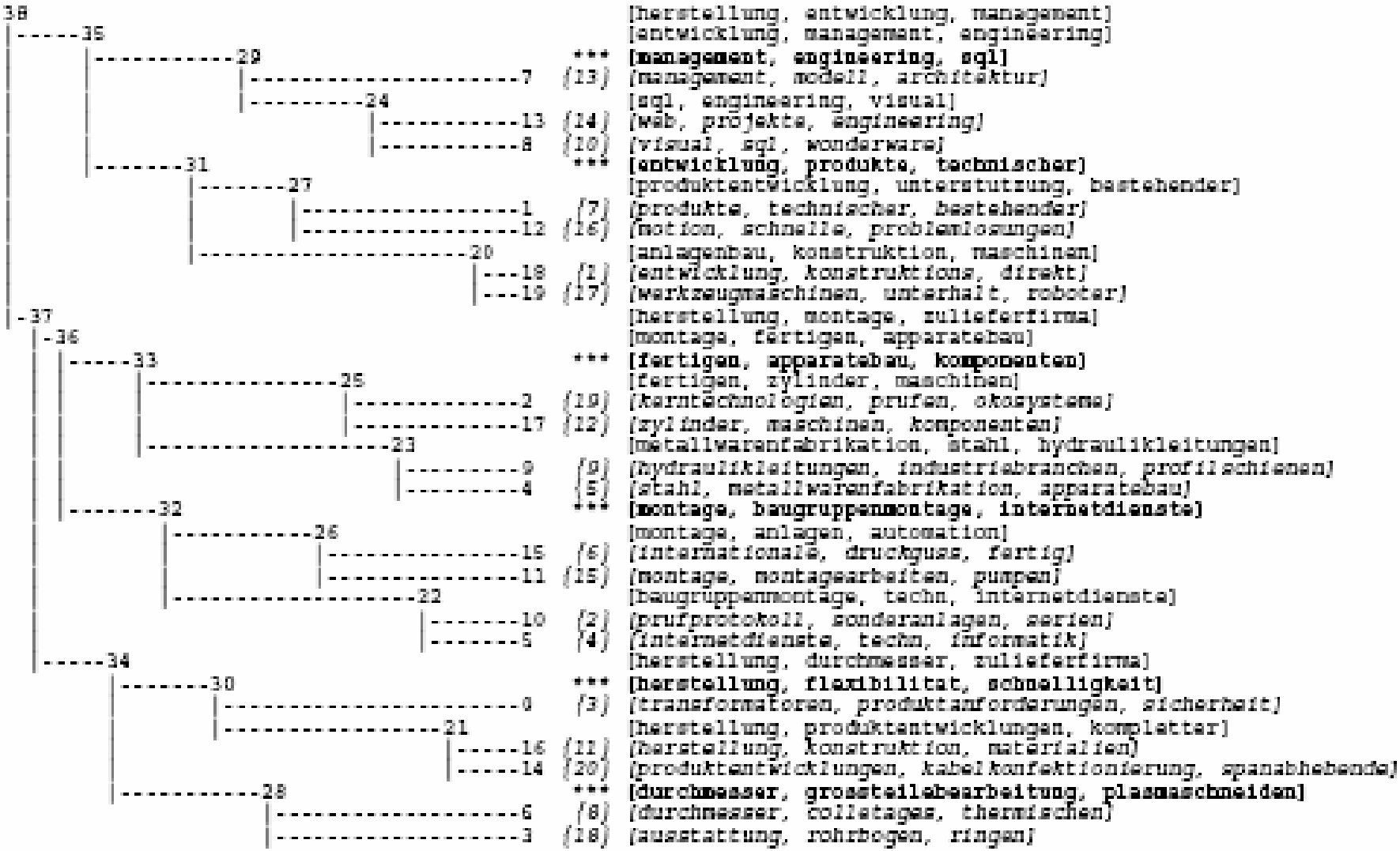
Document Clustering

- Clustering is a process of finding natural groups in data in a unsupervised way (no class labels pre-assigned to documents)
- Document similarity is used
- Most popular clustering methods are:
  - K-Means clustering
  - Agglomerative hierarchical clustering
  - EM (Gaussian Mixture)
  - …

K-Means clustering

- Given:
  - set of documents (eg., word-vectors with TFIDF),
  - distance measure (eg., cosine similarity)
  - K - number of groups
- For each group 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)

VF partners clustering



VF partners competency visualization

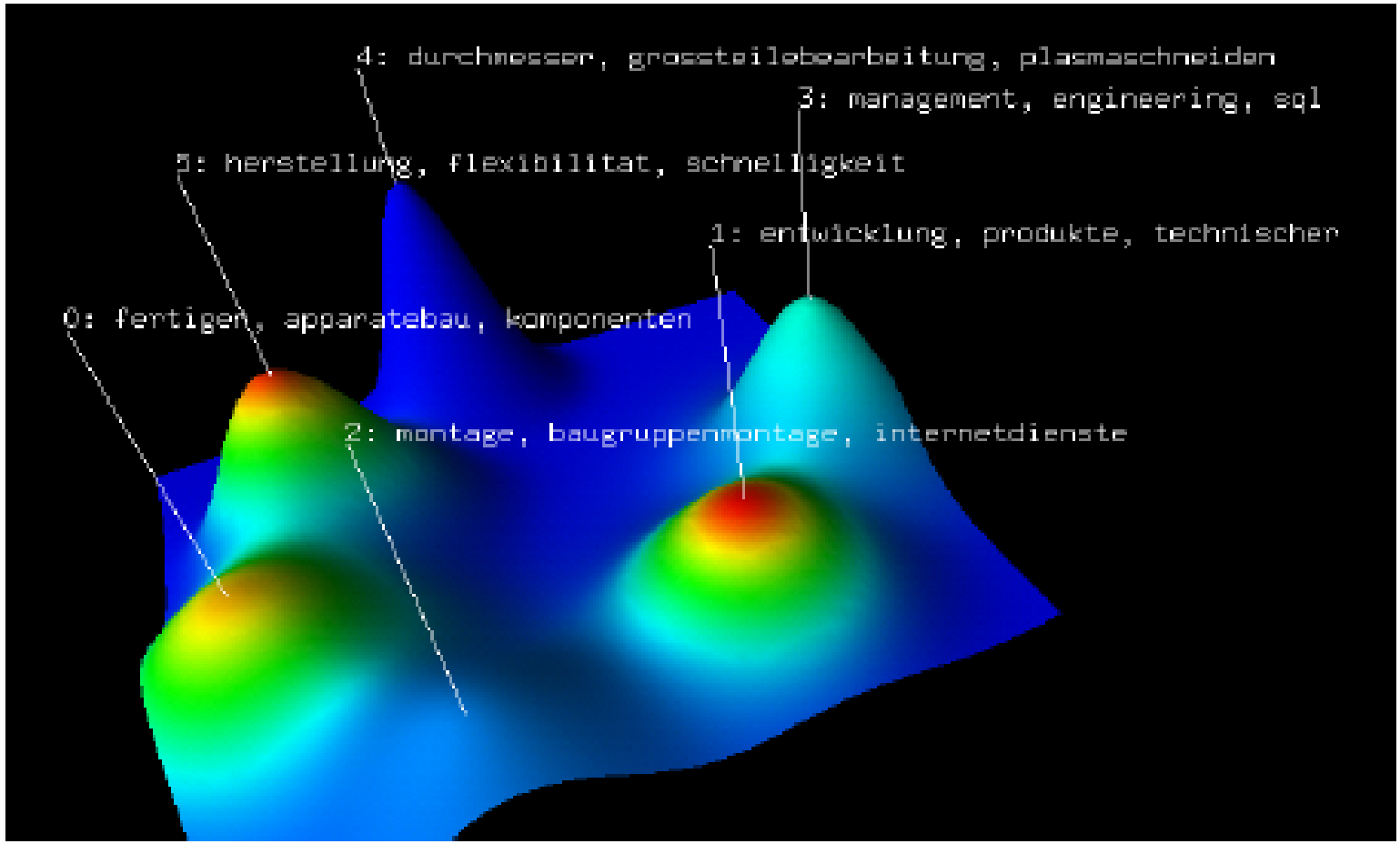
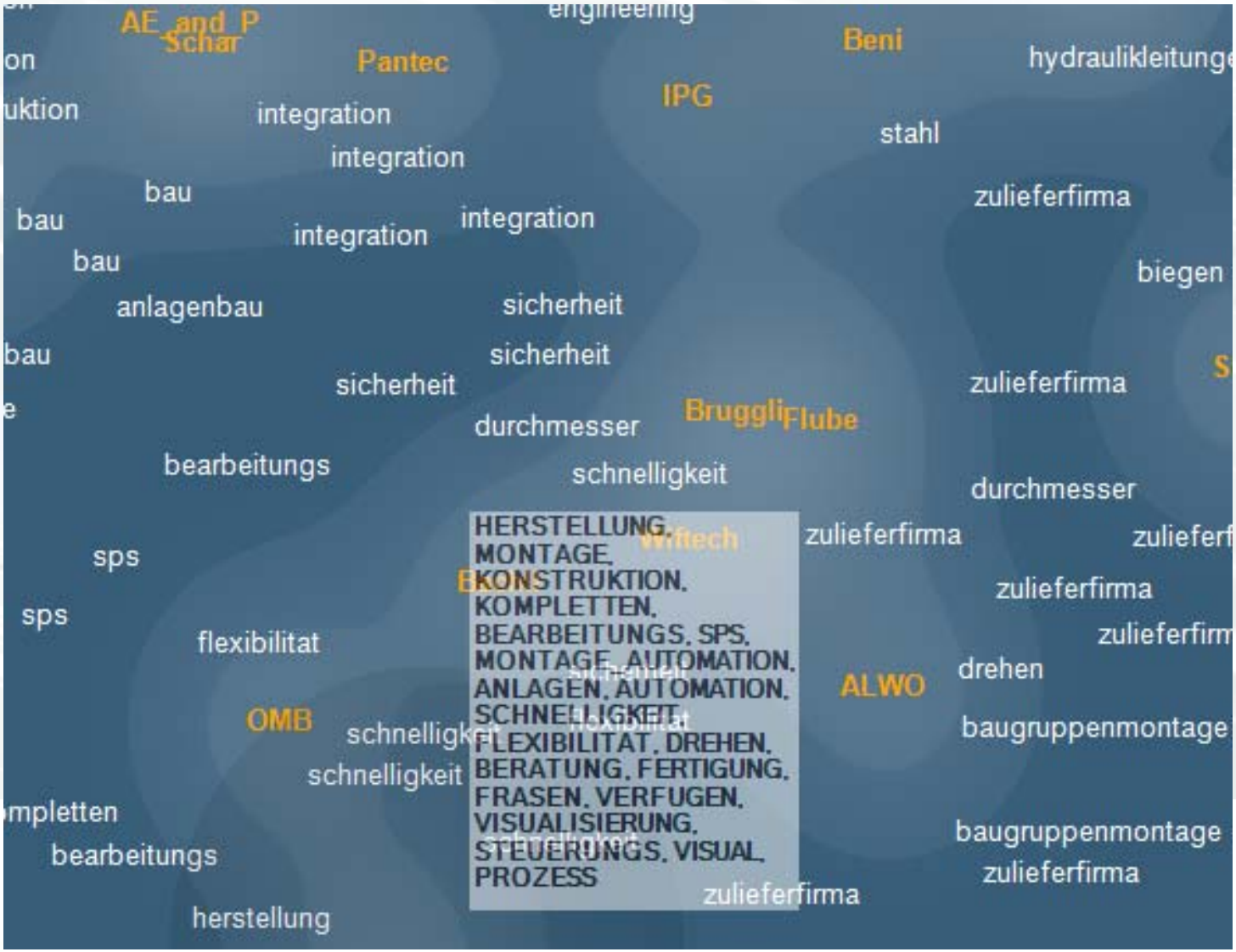


Figure 3 - Mountain visualization of six clusters, described with most descriptive words.

VF partners competency visualization



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Summary

- What is knowledge
- Traditional view of KM
- KM in the new economy: A networked organizations perspective
- Selected knowledge technologies for KM
- Using Web crawling and social network analysis for trust modeling and competency structuring through ontology construction

OntoGen:
System for semi-automatic ontology construction

Blaž Fortuna
Dunja Mladenić
Marko Grobelnik

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Why *semi*-automatic?

There is a big divide between *unsupervised* and *fully supervised* construction tools.

Both approaches have weak points:

- it is difficult to obtain desired results using unsupervised methods, e.g. limited background knowledge
- manual tools (e.g. Protégé, OntoStudio) are time consuming, user needs to know the entire domain.

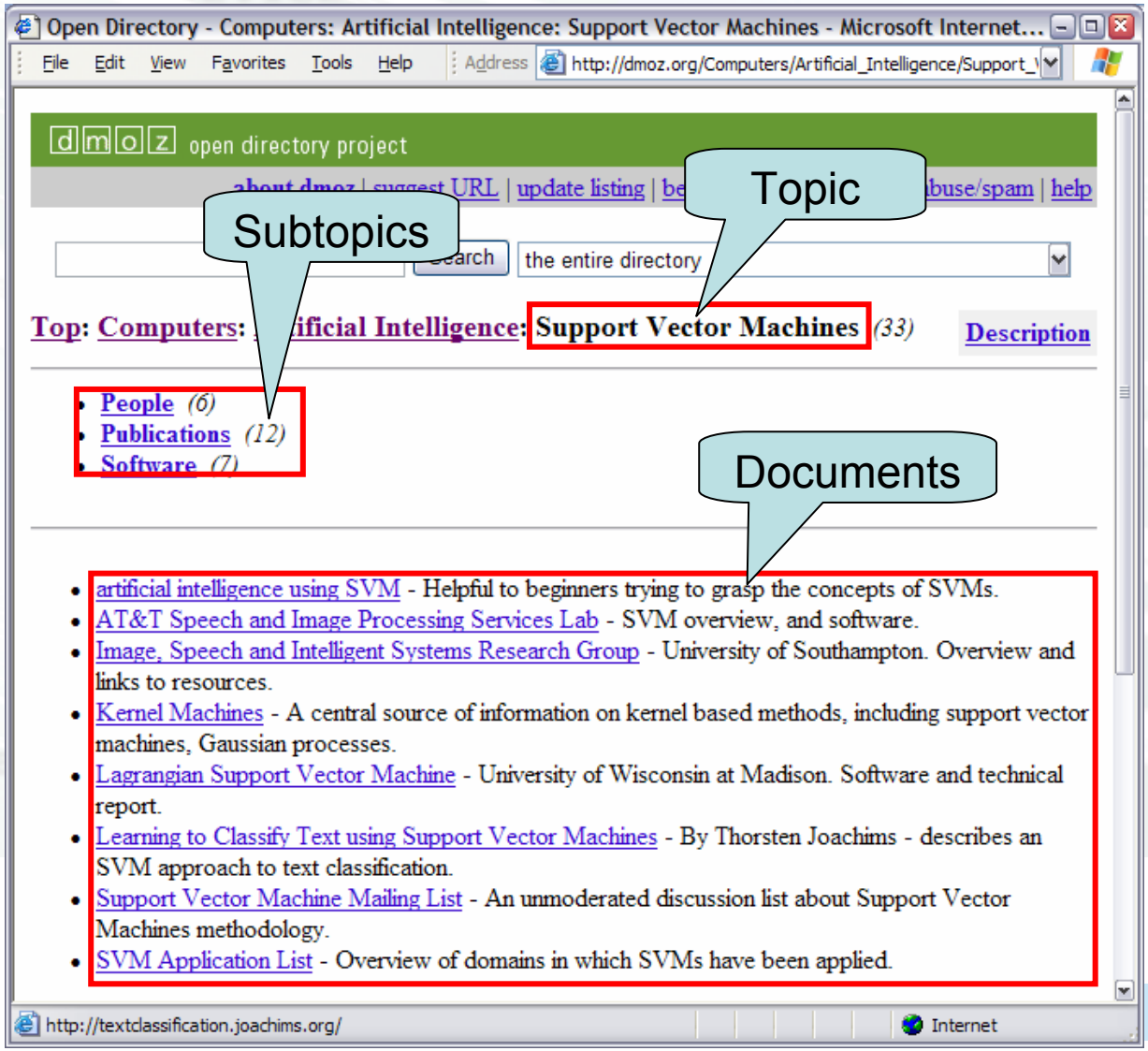
We combined these two approaches in order to eliminate these weaknesses:

- the user guides the construction process,
- the system helps the user with suggestions based on the document collection.

What is topic ontology?

It consists of:

- set of **topics**,
- set of **'Subtopic-of'** relations between topics,
- collection of **documents**,
- set of **'Subject-of'** relations between documents and topics.



How do we help the user?

Our Topic ontology construction tool helps the user by:

- … identifying the topics and relations:
  →*hierarchical clustering, latent semantic analysis*
- … naming the topics:
  →*keywords extraction*.
- … visualizing the topic ontology,
- … outlier detection.

Topic and relation identification

- k-means clustering used to identify topics:
  - cluster of documents => *topic*
  - documents are assigned to clusters => *'subject-of' relation*
  - We can repeat clustering on a subset of documents assigned to a specific topic => identifies *subtopics* and *'subtopic-of' relation*

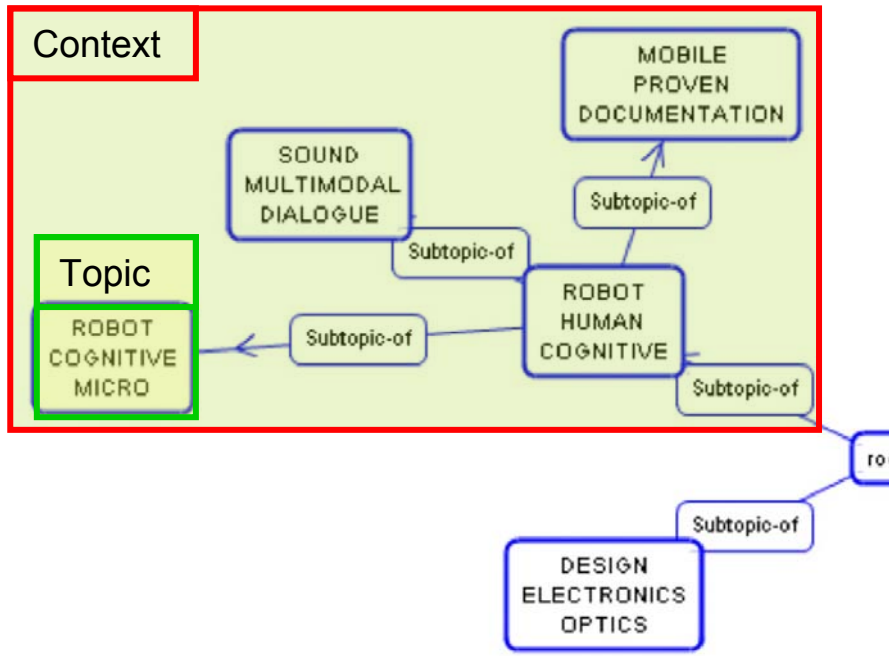
Keywords extraction

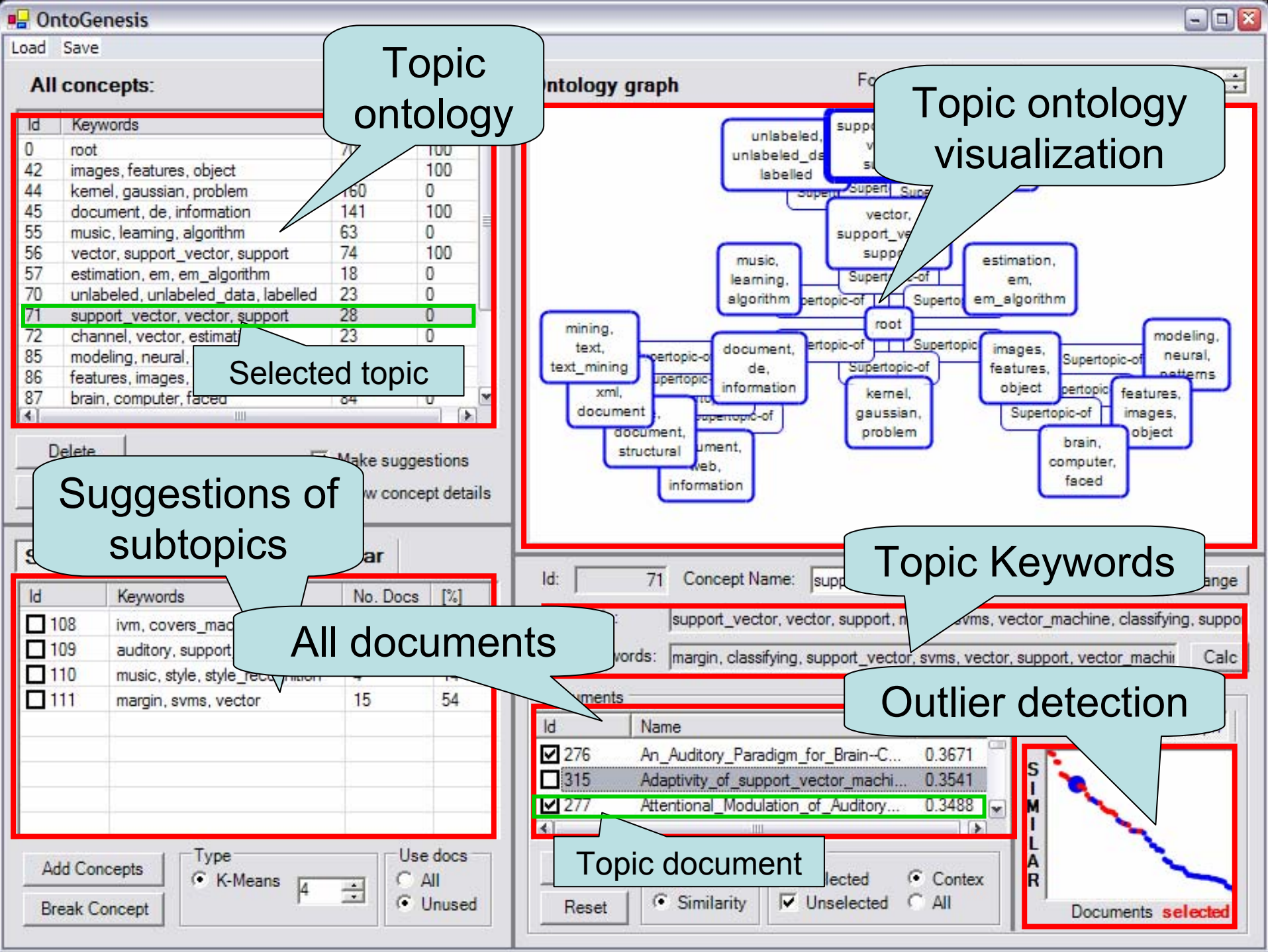
… using centroid vector:

- A *centroid vector* of a given topic is the average document from this topic (normalised sum of topic's documents)
- Most descriptive keywords for a given topic are the words with the highest weights in the centroid vector.

… using linear SVM classifier:

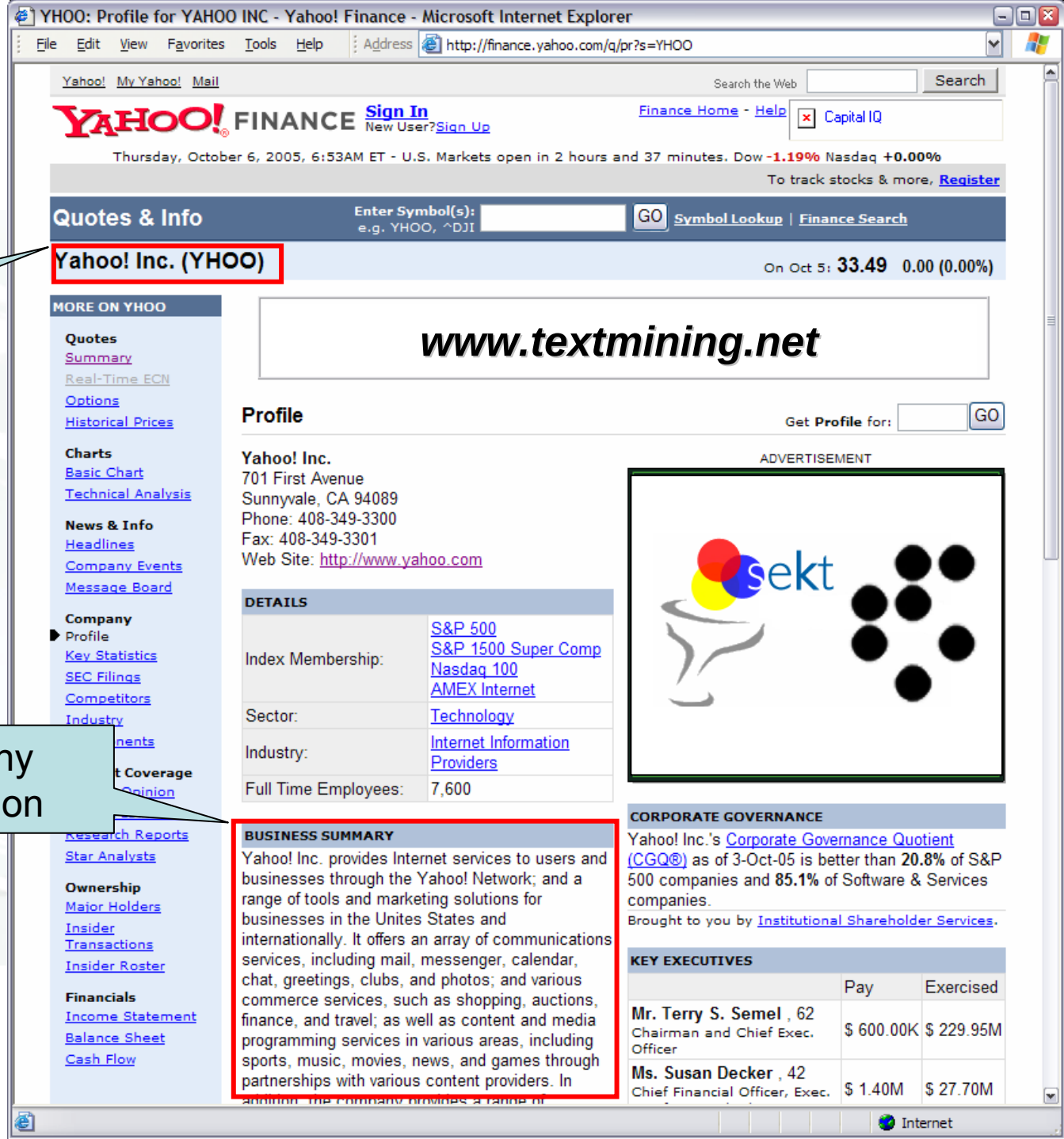
- SVM classifier is trained to seperate documents of the *given topic* from the other document *in the context*
- Words that are found most important for the classification are selected as keywords for the topic



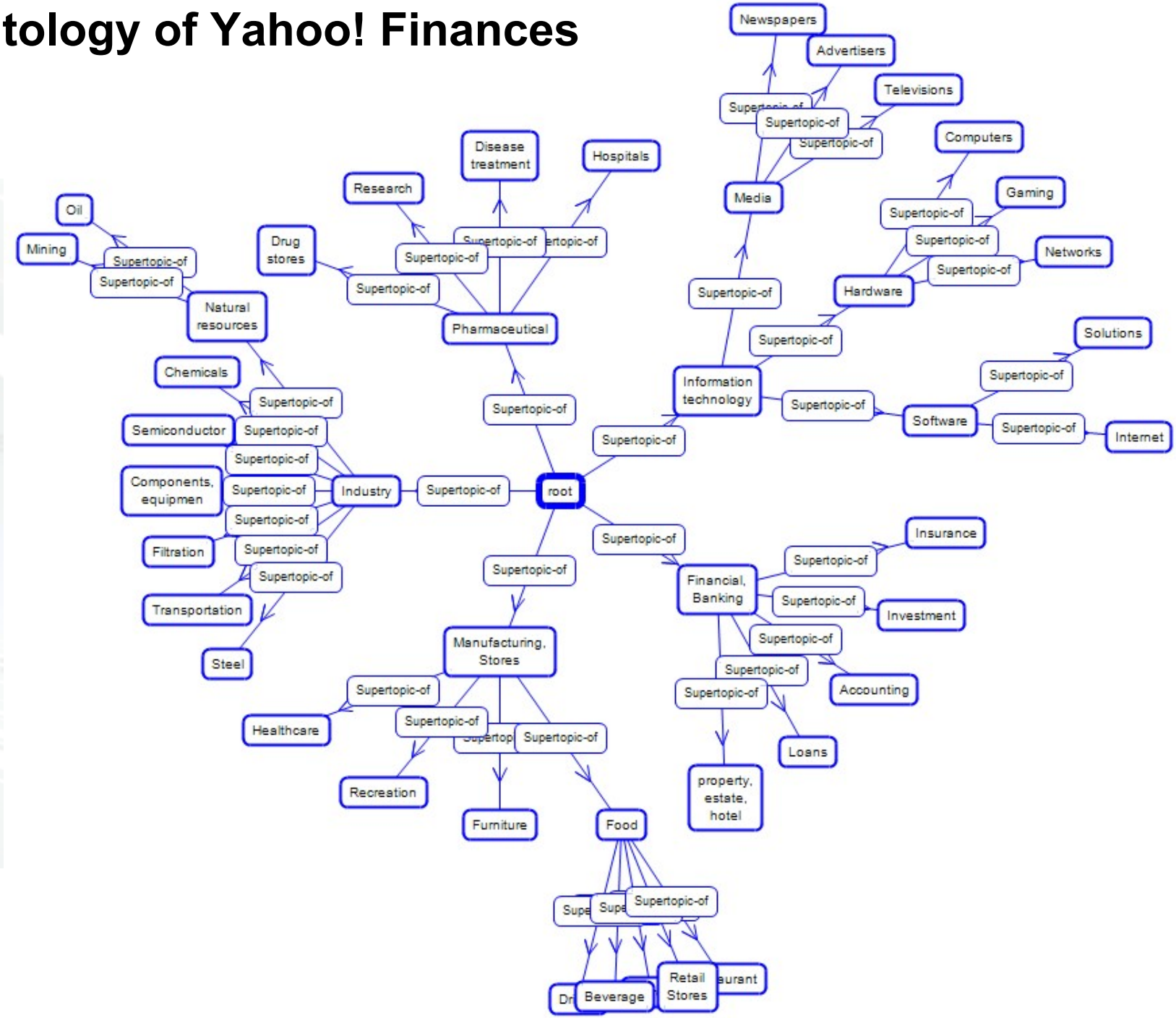

OntoGenesis
Load
Save
All concepts:
Topic ontology
Id
Keywords
0 root
42 images, features, object 100
44 kernel, gaussian, problem 0
45 document, de, information 141 100
55 music, learning, algorithm 63 0
56 vector, support_vector, support 74 100
57 estimation, em, em_algorithm 18 0
70 unlabeled, unlabeled_data, labelled 23 0
71 support_vector, vector, support 28 0
72 channel, vector, estimat 23 0
85 modeling, neural,
86 features, images,
87 brain, computer, faced
Selected topic
Delete
Make suggestions
w concept details
Suggestions of subtopics
Id
Keywords
No. Docs
[%]
108 ivm, covers_mac
109 auditory, support
110 music, style, style_rec
111 margin, svms, vector 15 54
All documents
Add Concepts
Break Concept
Type
K-Means
4
Use docs
All
Unused
ntology graph
Topic ontology visualization
unlabeled, unlabeled_de labelled
vector, support_ve supp
music, learning, algorithm
estimation, em, em_algorithm
root
mining, text, text_mining
xml, document
document, de, information
document, structural
kernel, gaussian, problem
images, features, object
modeling, neural, patterns
features, images, object
brain, computer, faced
Supertopic-of
Id: 71
Concept Name:
Topic Keywords
support_vector, vector, support,
svms, vector_machine, classifying, suppo
margin, classifying, support_vector, svms, vector, support, vector_machi
Calc
Outlier detection
Id
Name
276 An_Auditory_Paradigm_for_Brain-C... 0.3671
315 Adaptivity_of_support_vector_machi... 0.3541
277 Attentional_Modulation_of_Auditory... 0.3488
Topic document
Reset
Similarity
Unselected
Contex
All
SIMILAR
Documents selected

Company name

Company description



Topic ontology of Yahoo! Finances



Other features & Future work

Other features:

- Reads and writes to RDF
- Can work as plug-in for OntoStudio
- Already actively used in more projects and by different people

Future work

- Identification of relations
  - based on context of instances
- Extend to more general types of ontologies