Information Extraction

PD Dr. Günter Neumann
DFKI and Saarland University

Outline

- Overview
- Named Entity Extraction
- Relation Entity Extraction
- Mining Meaning from Wikipedia
- Web Information Extraction & Machine Reading
- Open topics

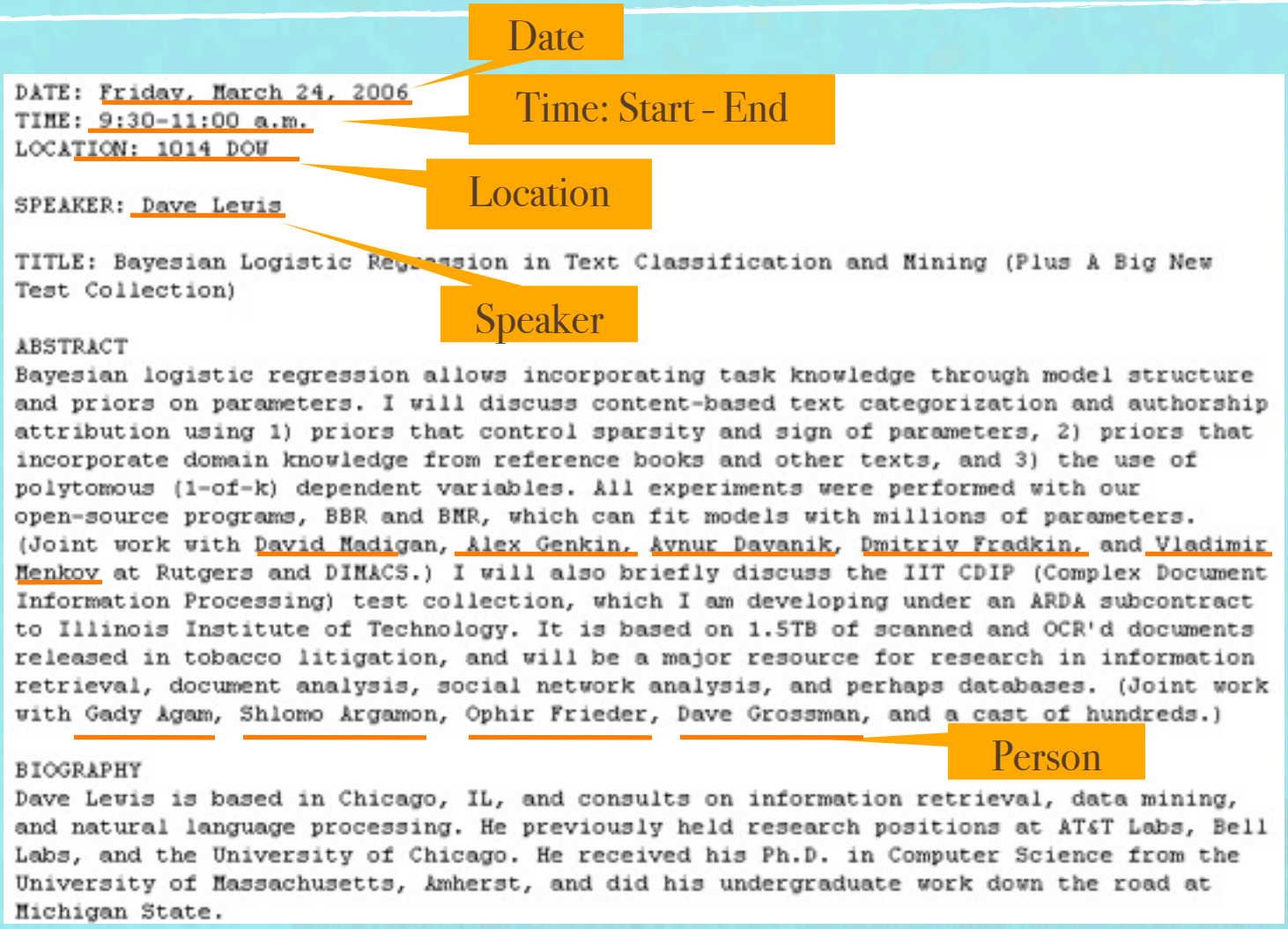
Course Exam ?

- It should be doable -> less effort than a 6/9 CP lecture
- It should be honest -> nothing is for free (but death, and this takes your life)
- Idea:
  - Written exam
    - Topic centric, e.g., only about 1-2 topics (e.g., only relation extraction or Machine Reading)
  - Oral exam
    - at the end of the semester, each student selects a topic, and prepares an oral exam
- When ? March ?

Text Exploration →
Important Direction for Our Community

- Many other research communities are looking at how to explore text
  - Most actively, Web, IR (Information Retrieval), AI (Artificial Intelligence), KDD (Knowledge Discovery and Data Mining)
- Important direction for us as well!
  - We have lot to offer, and a lot to gain
- How is text exploited?
  - Text Mining, Information Extraction

The Challenge



DATE: Friday, March 24, 2006
TIME: 9:30-11:00 a.m.
LOCATION: 1014 DOW

SPEAKER: Dave Lewis

TITLE: Bayesian Logistic Regression in Text Classification and Mining (Plus A Big New Test Collection)

ABSTRACT
Bayesian logistic regression allows incorporating task knowledge through model structure and priors on parameters. I will discuss content-based text categorization and authorship attribution using 1) priors that control sparsity and sign of parameters, 2) priors that incorporate domain knowledge from reference books and other texts, and 3) the use of polytomous (1-of-k) dependent variables. All experiments were performed with our open-source programs, BBR and BMR, which can fit models with millions of parameters. (Joint work with David Madigan, Alex Genkin, Aynur Dayanik, Dmitriy Fradkin, and Vladimir Menkov at Rutgers and DIMACS.) I will also briefly discuss the IIT CDIP (Complex Document Information Processing) test collection, which I am developing under an ARDA subcontract to Illinois Institute of Technology. It is based on 1.5TB of scanned and OCR'd documents released in tobacco litigation, and will be a major resource for research in information retrieval, document analysis, social network analysis, and perhaps databases. (Joint work with Gady Agam, Shlomo Argamon, Ophir Frieder, Dave Grossman, and a cast of hundreds.)

BIOGRAPHY
Dave Lewis is based in Chicago, IL, and consults on information retrieval, data mining, and natural language processing. He previously held research positions at AT&T Labs, Bell Labs, and the University of Chicago. He received his Ph.D. in Computer Science from the University of Massachusetts, Amherst, and did his undergraduate work down the road at Michigan State.

What is "Information Extraction"

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

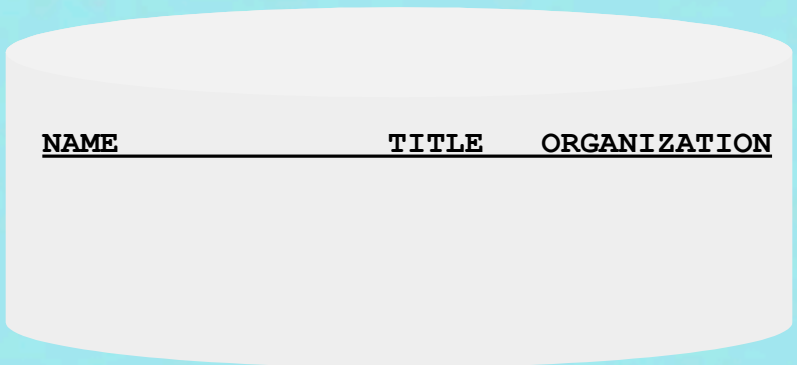
October 14, 2002, 4:00 a.m. PT

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…



From William W. Cohen

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…

IE

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

From William W. Cohen

What is "Information Extraction"

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

aka "named entity recognition"

What is "Information Extraction"

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 F.

What is "Information Extraction"

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 F.

What is "Information Extraction"

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."

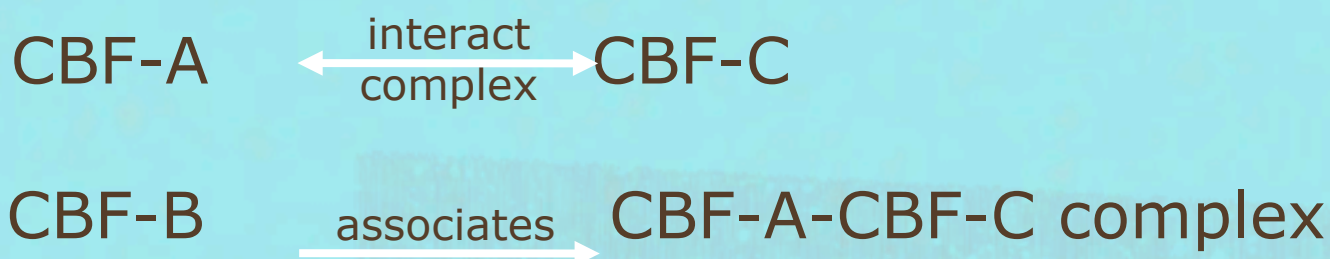
Richard Stallman, founder of the Free Software Foundation, countered saying…

- Microsoft Corporation
  CEO
  Bill Gates
- Microsoft
- Gates
  Microsoft
- Bill Veghte
  Microsoft
  VP
- Richard Stallman
  founder
  Free Software F.

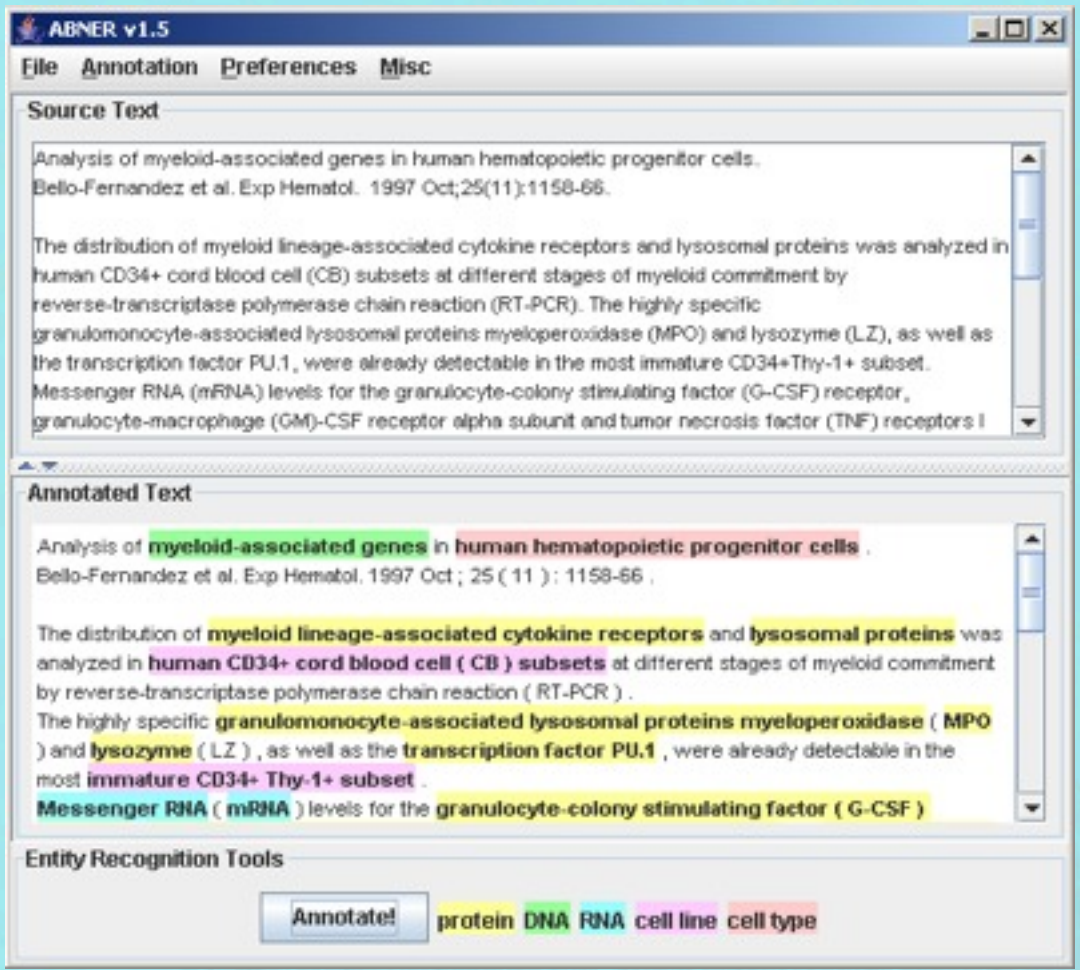
NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Examples of Entity-Relationship

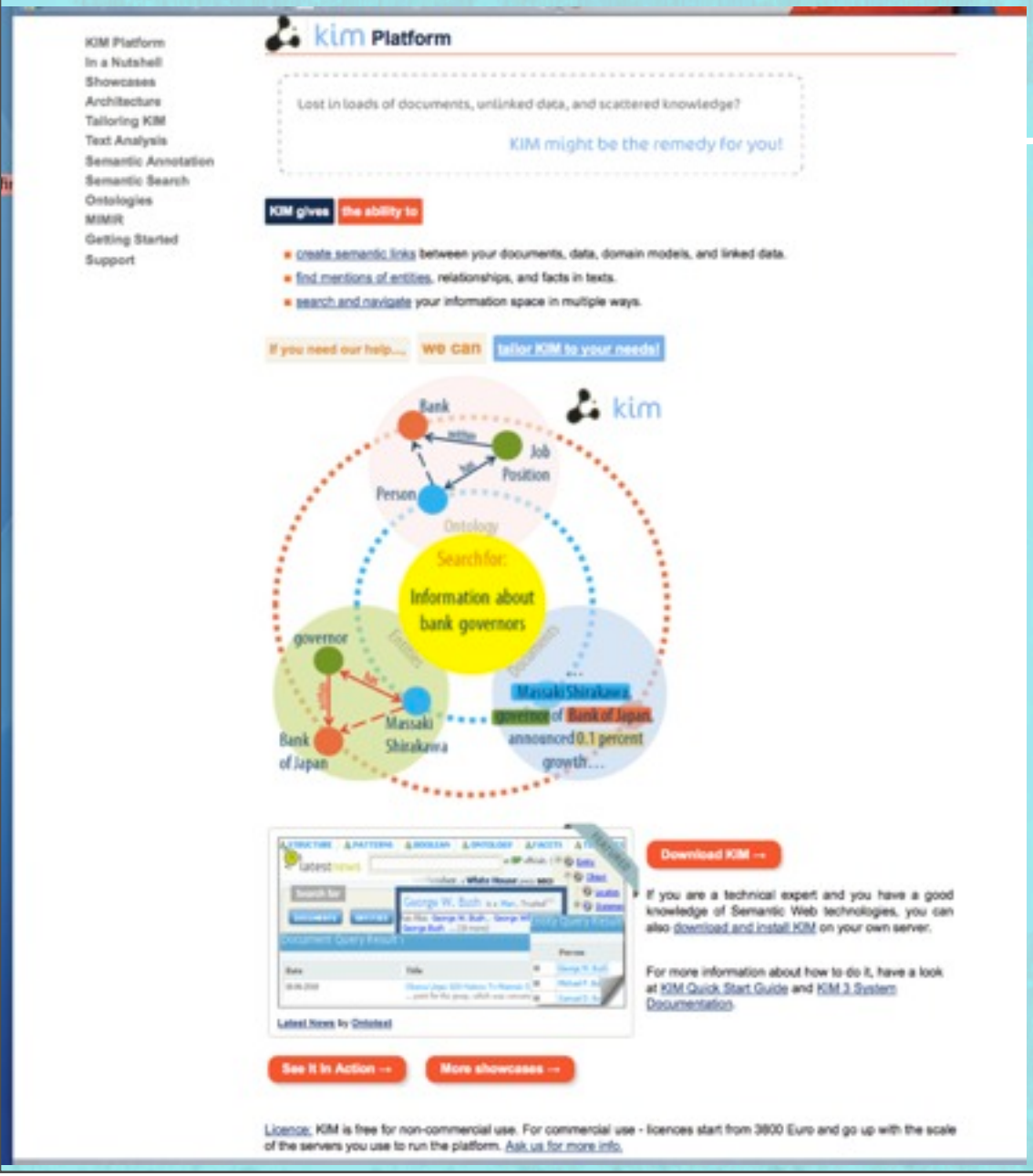
„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."



ABNER - A Biomedical Named Entity

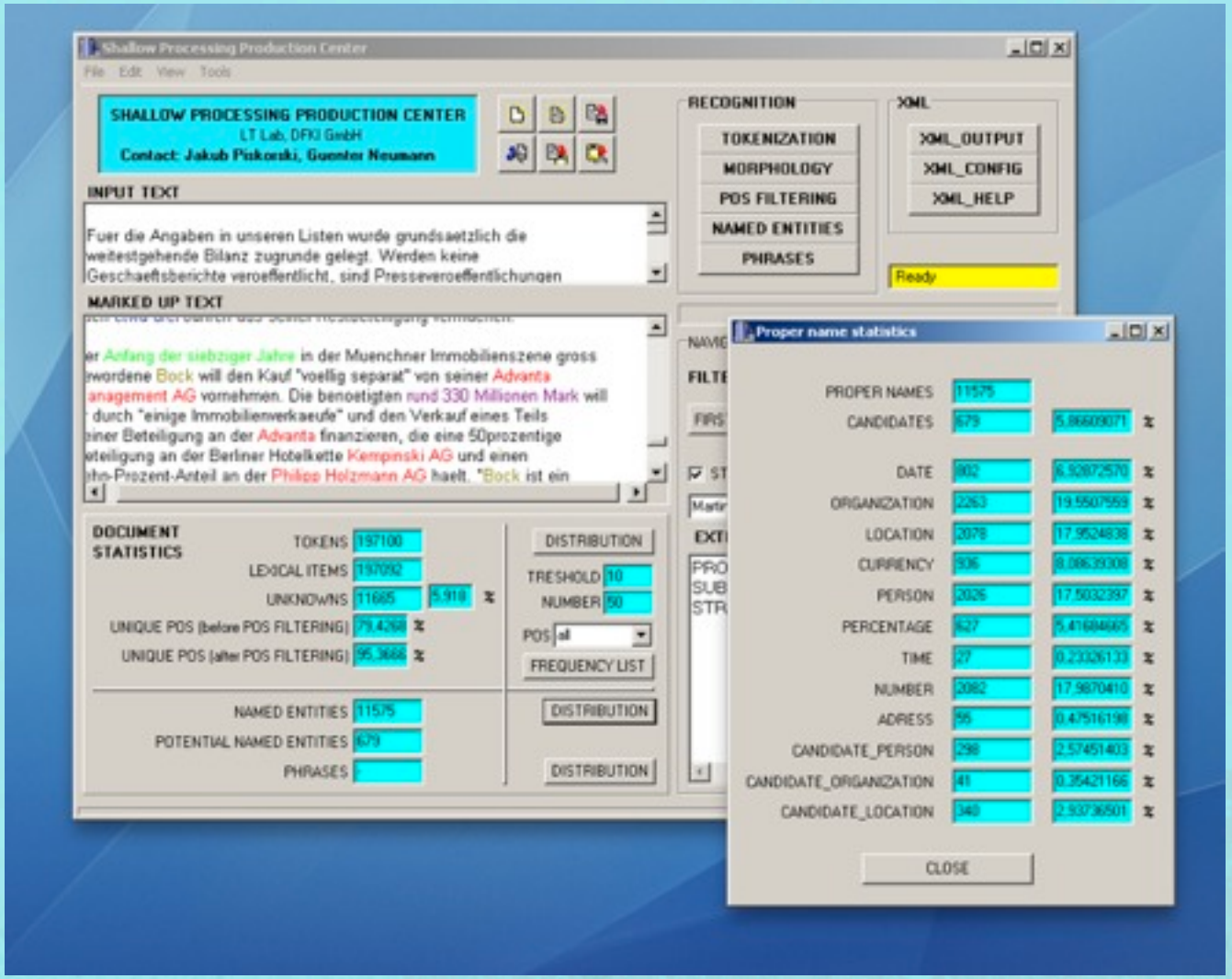


Application Example - KIM



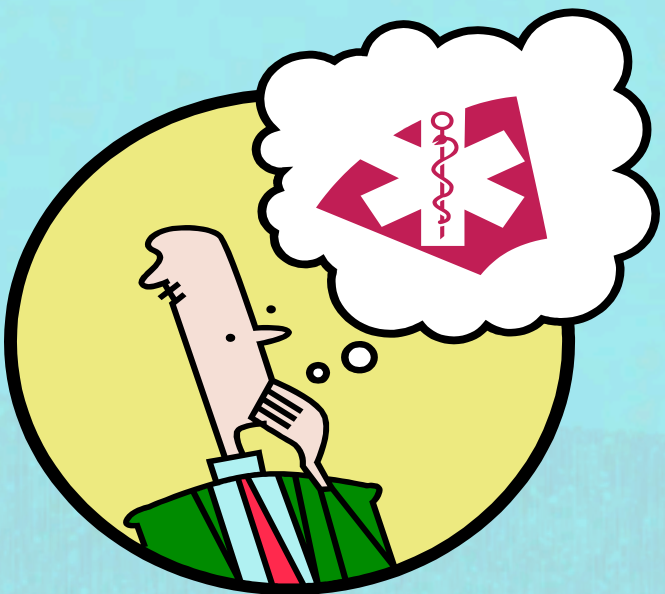
http://www.ontotext.com/kim

SPPC - German NE recognizer



Mining Medical Literature

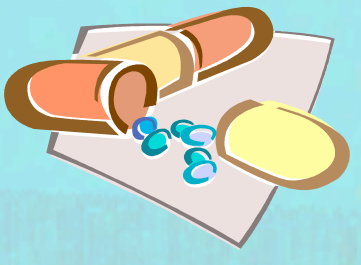
- **Medical research**
- **Find causal links between symptoms or diseases and drugs or chemicals.**



A Classical Example

- **Research objective:**
  - Follow chains of causal implication to discover a relationship between migraines and biochemical levels.

- **Data:**
  - medical research papers, medical news **(unstructured text information)**



- **Key concept types:**
  - symptoms, drugs, diseases, chemicals...
  - These have to be identified and analysed

Relationship of IE to other NL-related application areas

(1) Information Retrieval (IR)

Identify and extract documents as answers of an information request.

(2) Passage Retrieval

Identify and extract document snippets as answers of an information request.

(3) Information Extraction (IE)

Identify and extract relevant textual passages used for filling up a pre-defined data record/template.

(4) Textual Question-Answering

Answer an arbitrary question by using textual documents as knowledge base: Fact retrieval, combination of IR & IE.

(5) Text understanding

Interpret texts like humans do: Artificial Intelligence

Interpretation of NL-documents

(1) Information Retrieval (IR)

User

(2) Passage Retrieval

User

(3) Information Extraction (IE)

System (static, pre-defined)

(4) Textual Question/Answering

System (dynamic, facts/relations)

(5) Text understanding

System (complete)

NL analysis as step-wise normalization

- Tokenization

  9.11.2000, 11/9/2000 →
  {day: 9, month: 11, year: 2000}

- Morphological analysis:
  - Determination of lexical stems
    - Inflection:
      supporting → *to support*
      Häuser → *haus*
    - German compounds:
      Informationstechnologiezentrum →
      {*Information, Technologie, Zentrum*}

NL analysis as step-wise normalization

- Special phrases (word groups):
  - date and time expressions:

    18.12.98 und *Friday, December the 18th 1998*

    ⇨ <type=date, year=1998, month=12, day=18, weekday=5>
  - proper names: persons, institutions, companies, locations, products, ...
  - number expressions, addresses, mathematical expressions, ...

NL analysis as step-wise normalization

- General phrases:
  - nominal phrases, prepositional phrases, verb groups
    - For the new economy
    - ➡<head=for, comp=<head=economy, quant=def, mod=new>>
- complex flat sentence structure
- domain specific templates (integration of ontology)

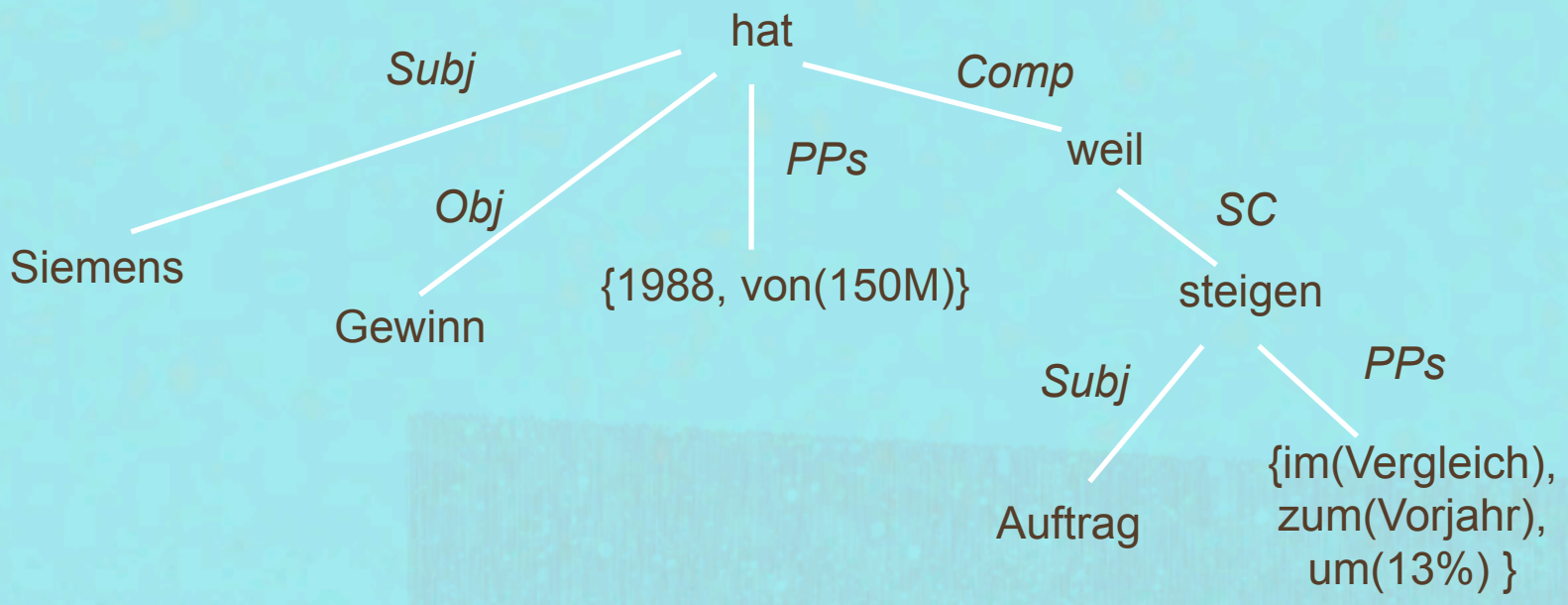
type	=	turnover	c-name	=	Possehl1
year	=	1995/1	amount	=	1.3e+9DM
tendency	=	+	diff	=	+23%

Underspecified functional description for sentences

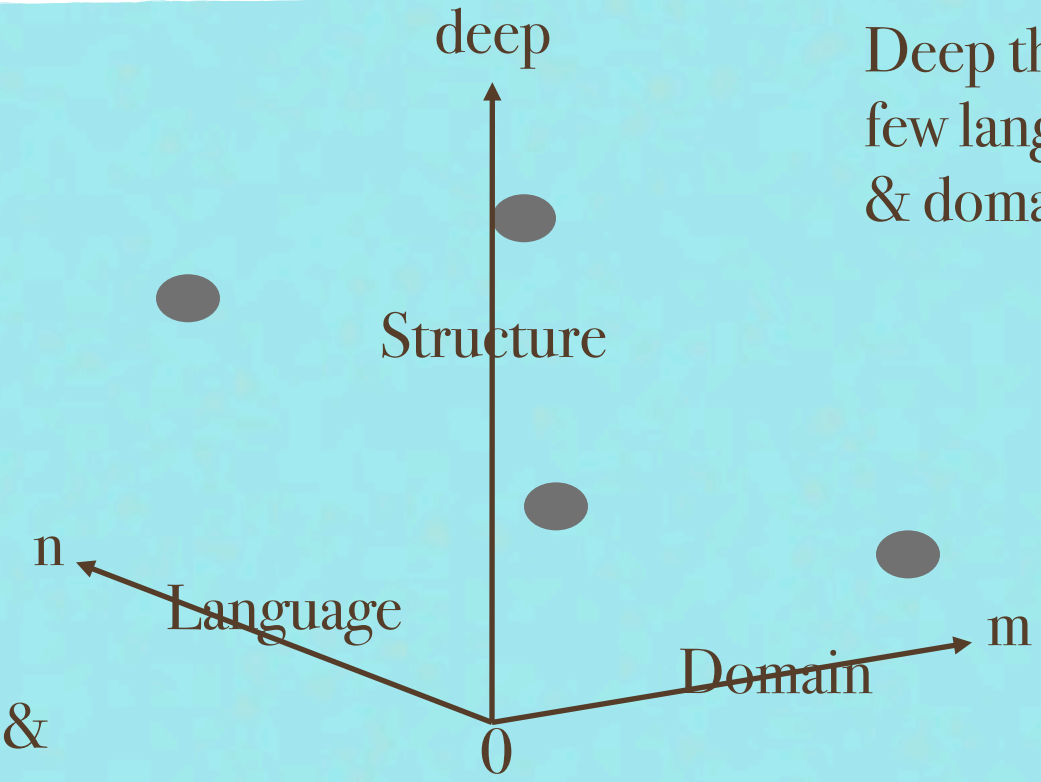
Flat dependency-based structure, only upper bounds for attachment and scoping:

[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM], [$_{Comp}$weil] [$_{NP}$die Aufträge] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].

"The siemens company has made a revenue of 150 million marks in 1988, since the orders increased by 13% compared to last year."



Complexity of IE



Deep then
few languages
& domains

Many languages &
domains
then very shallow

Data - Knowledge - Information

- Main task of an information system
  - Maintain knowledge in digitalized form as data
  - Provide knowledge as useful information to a user

Data - Knowledge - Information

Information = Data + Knowledge.

- Data:
  - recorded facts or figures
- Knowledge:
  - the understanding required to convert data into information and to apply it to real-world situations
- Information:
  - the value derived from data through the application of knowledge

Data vs. Knowledge

New Dehli‘s latitude

Character sequence

28081749

Birthday of Goethe

Knowledge is data with meaning, e.g., a property (or feature) of an object (size of a human, name of a company). Note that the same data element might have several possible interpretations.

11:15

Time expression

game result

Knowledge vs. Information

- Knowledge:
  - A model of the world (structural and functional properties of the real world)
- Information:
  - Is that part of knowledge which is used to solve a certain problem (Information System view).
  - Information only exists in concrete problem situations.
- Information systems extract that knowledge „just in time“, a user needs in context of a given situation.
  - If the information search is done, then the information is unnecessary.
  - Seen so, information need not necessarily be stored; only if it is new knowledge. In this case information turned to knowledge.

SDI: Standard Definition of Information, Floridi, 2005

- Intuitively: „information“ means
  - Non-mental, user-independent, declarative, semantic content
  - Embedded in some physical implementation
  - Information as cognitive units which can be generated and carried by texts/news
- DOS
  - Declarative, objective, semantic information

SDI means:

- Let „infons“ be discrete elements of information (independenly of a specific semantic encoding or physical implementation).
- „infon“ is an instance of DOS, iff
  - SDI.1: „infon“ consists of N data
  - SDI.2: the data are wellformed
  - SDI.3: the wellformedness is significant, i.e., not arbritary

SDI.1 means that

- Information is not dataless, but the concrete data type is not important.
  - This means: information exists, because data exists.
- Distinguish:
  - Primary data: the implemented data types, e.g., numbers, texts, DB entries
  - Meta data: secondary indicators about the nature of the primary data, e.g., location, formats, updates, copyrights
  - Operational data: data about the use of data, e.g., wrt whole system, its performance
  - Derived data: data which are derived/computed from the above data

SDI.2 means that

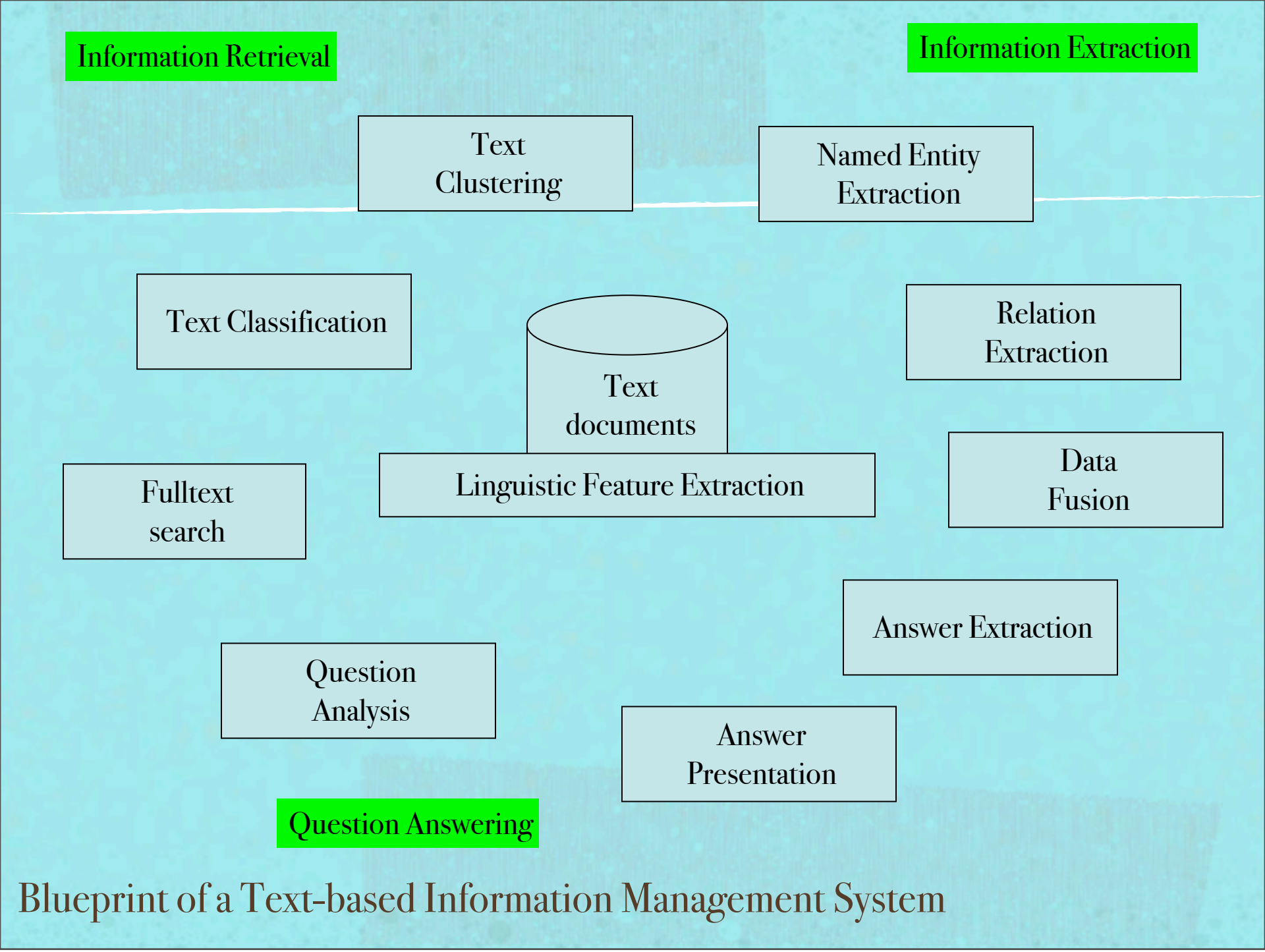
- Information is usually transmitted by means of large groups of patterns of wellformed, coded data, very often alhpanumerically
- Information depends the occurrency of syntactically wellformed groups, strings or patterns of data, and that they are physically implemented, where the concrete implementation might be differently
- No information without data representation
  - Quasi bodyless information is not possible

SDI.3 means that

- Information is the name of the meaning that is exchanged
- Information is „the difference about the difference“
- Difference is a discrete state, i.e., a date and „making the difference“ means, that the date is significant at least potentially.
- Information exists with an informed subject.

Text-based Information Management (TIM)

- Main tasks
  - To maintain the information which is represented in digital form in data
  - To identify and collect the relevant information for a user request
  - To present that information to a user in an understandable form.
- Text-based means
  - The information is encoded mainly in natural language in texts and has to be transformed into data.
- This requires NLP tools of different granularity depending on the depth of the structure that has to be determined in NL texts.



Blueprint of a Text-based Information Management System

Text Mining Definition

Many definitions in the literature

- The non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data

- An exploration and analysis of textual (natural-language) data by automatic and semi automatic means to discover new knowledge.

- What is "previously unknown" information?
  - Strict definition
    - Information that even the writer does not know
  - Lenient definition
    - Rediscover the information that the author encoded in the text
    - Unfold implicit relationships mentioned via textual entities and make them explicit to the reader

Text Mining Process

- Text Preprocessing
  - Syntactic/Semantic Text Analys
- Feature Generation
  - Bag of Words
- Feature Selection
  - Simple Counting
  - Statistics
- Text/Data Mining
  - Classification- Supervised Learning
  - Clustering- Unsupervised Learning
- Analyzing Results