Reutlingen University

**DBKDA 2016, Lisbon, 28.06.2016**

# *Current Trends in Information Searching/Query Answering*

*Moderator*

*Fritz Laux, Reutlingen University, Germany*

*Panelists:*

*Jerzy W. Grzymala-Busse, University of Kansas, USA*

*Dimitar Hristovski, University of Ljubljana, Slovenia*

*Andreas Schmidt, Karlsruhe University of Applied Sciences & Karlsruhe Institute of Technology, Germany*



Reutlingen University

## Topics of the Panelists

- *Jerzy W. Grzymala-Busse: „Incompleteness versus inconsistency in Data Mining"*
- *Dimitar Hristovski: „Answering biomedical questions using semantic relations"*
- *Andreas Schmidt: „Named entity recognition and Disambiguation"*
- *Fritz Laux: „Answering Queries with Crowdsourcing"*

2 /10

Reutlingen University

## Crowdsourcing

- *Portmanteau: Crowdsourcing = Crowd + Outsourcing*
- *Definition: Outsourcing tasks to many Web users (the Crowd) Many complicated definitions:*
  - Jeff Howe: "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."[1)]
  - Daren Brabham: ""online, distributed problem-solving and production model."[2)]
  - Christian Papsdorf: Crowdsourcing is the strategy of outsourcing working power by an organization or individual of usually internally performed payed services to a number of unknown persons to gain freely usable and direct economical benefit. (translated from [3)])
- *Using crowd intelligence to execute small, well defined human intelligence tasks (HIT)*

1) *Jeff Howe, "Crowdsourcing: A Definition" , URL: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html, (2006).*
2) *Daren C. Brabham," Crowdsourcing as a Model for Problem Solving", URL: http://www.webcitation.org/67BLxbafe*
3) *Ch. Papsdorf, „Wie Surfen zu Arbeit wird", Crowdsourcing im Web 2.0, Campus Verlag2009, S. 69. ISBN: 359339040X*

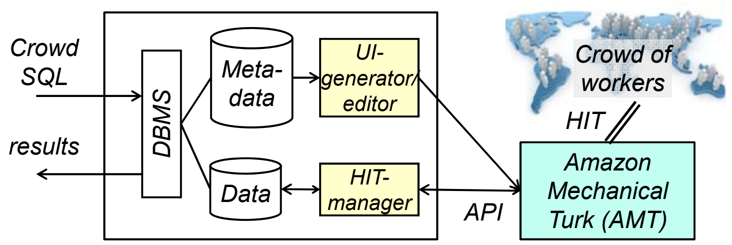
3 /10

© F. Laux

Reutlingen University

## Answering Queries with Crowdsourcing

- *Idea*
  - Combine human work and machine processing to get better semantics and more comprehensive query results
  - Using the Crowd to do work requiring human intelligence and split it into small tasks, called HITs
- *An Example: CrowdDB[1)]*
  - CrowdDB is a database with SQL extension for crowdsourcing. It uses Amazon Mechanical Turk (AMT)




[1)] *M. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin; „CrowdDB: Answering Queries with Crowdsourcing", ACM SIGMOD 2011, Athens, Greece*

4 /10

© F. Laux

Reutlingen University

## CrowdDB

- *New functionality: Answer queries the cannot be answered by computer only, e.g.*
  - Processing requires human input that is missing
  - Performing functions like matching, ranking or summary based on fuzzy criteria
  - Closed world assumption is given up
- *Small extension of SQL required*
  - Crowd colums/tables, CNULL
- *Problems*
  - How to identify and divide the workload for crowdsourcing
  - Formulate precise HITs → cultural background
  - Quality of results

5 /10



Reutlingen University

## SQL Extension[1] of CrowdDB

- *SQL DDL Extensions*
  - *Crowdsourced column (→ missing values suppl. by crowd)*
  - *Crowdsourced table (→ missing records supplied by crowd)*
- *Examples*
  - *SELECT market_capitalization FROM company WHERE name = "I.B.M."*
    - ⇨ *Empty answer if no record for "I.B.M." is found or name was entered differently*
    - ⇨ *Easy to find answer for a person with internet access*
  - *SELECT image FROM picture WHERE subject like "business" ORDER BY relevance LIMIT 1*
    - ⇨ *Easy to answer for humans*
    - ⇨ *If no relevance to a specific topic or if the name instead of the subject has been previously stored this query cannot be answered by the computer*

*[1] M. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin; „CrowdDB: Answering Queries with Crowdsourcing“, ACM SIGMOD 2011, Athens, Greece*

6 /10

Reutlingen University

## Generating query forms for crowdsourcing

- *Query Interfaces for Crowdsourcing generated from the DB-Schema*
  - Similar to other Tools like MS Access, Oracle Forms
- *Helps to produce forms for micro tasks*
- *Only well defined simple tasks like*
  - Searching for simple facts (e.g. phone number, number of inhabitants, prices)
  - Fuzzy tasks (e.g. finding similarities of pictures, providing a short description of a picture, entity disambiguation)

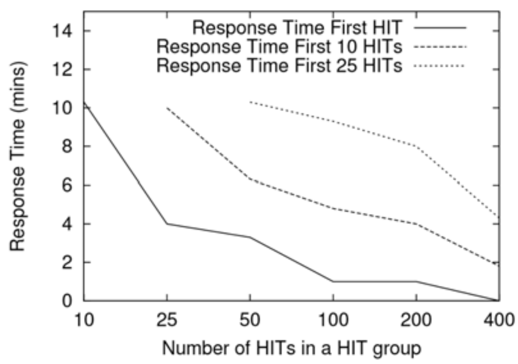
7 /10



Reutlingen University

## Experimental Results I

- *CREATE TABLE company (name VARCHAR PRIMARY KEY, phone_no CROWD VARCHAR(32), address CROWD VARCHAR(256));*
- *SELECT phone_no, address FROM company;*



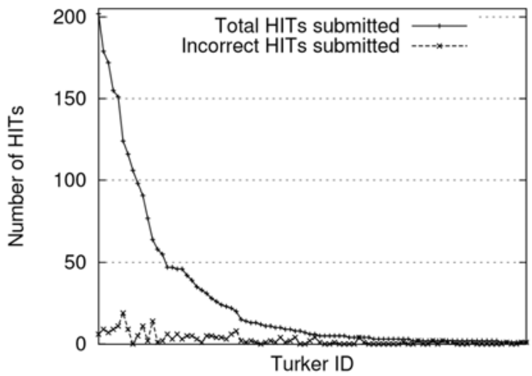

- *Similar tasks (a HIT-group) lead to faster first response*
- *For HIT-groups > 200 HITs a majority result can be expected within < 4 min*

8 /10

Reutlingen University

## Experimental Results II

- *CREATE TABLE company (name VARCHAR PRIMARY KEY, phone_no CROWD VARCHAR(32), address CROWD VARCHAR(256));*
- *SELECT * FROM company;  (5 Assignments/HIT)*




- *Highly skewed performance of workers (community)*
- *Error rate is nearly independent of the performance*

9 /10
© F. Laux

Reutlingen University

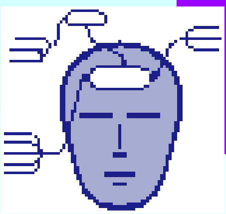
## Experimental Results III

- *SELECT p FROM picture*
  *WHERE subject = "Golden Gate Bridge"*
  *ORDER BY CROWDORDER(p,*
  *"Which picture visualizes better %subject");*


*Pictures of the Golden Gate Bridge from Flickr ordered by workers*
*Legend of the numbers a,b,c: a = votes from workers, b = rank based on votes from workers, c= rank by experts*

- *The Crowd is very good in ordering pictures*

10 /10
© F. Laux



# Biomedical Question Answering using Semantic Relations


Dimitar Hristovski,[1] Dejan Dinevski [2], Andrej Kastrin[1],
Thomas C Rindflesch[3]

*[1]Institute for biostatistics and medical informatics,*
*Medical faculty, University of Ljubljana*
*[2] Medical faculty Maribor, Slovenia*
*[3]National Library of Medicine, National Institutes of Health, Bethesda, MD, U.S.A.*

*e-mail: dimitar.hristovski@mf.uni-lj.si*

# Introduction

- Avalanche of information in biomedicine
- Evidence-based Medicine. Clinical practice should be based on evidence. On average, 2 min available for answering a question, but 30 min needed to find answer
- The importance of bibliographic databases, especially MEDLINE
- Information retrieval (IR) tools (e.g. Entrez for PubMed) most frequently used to search MEDLINE

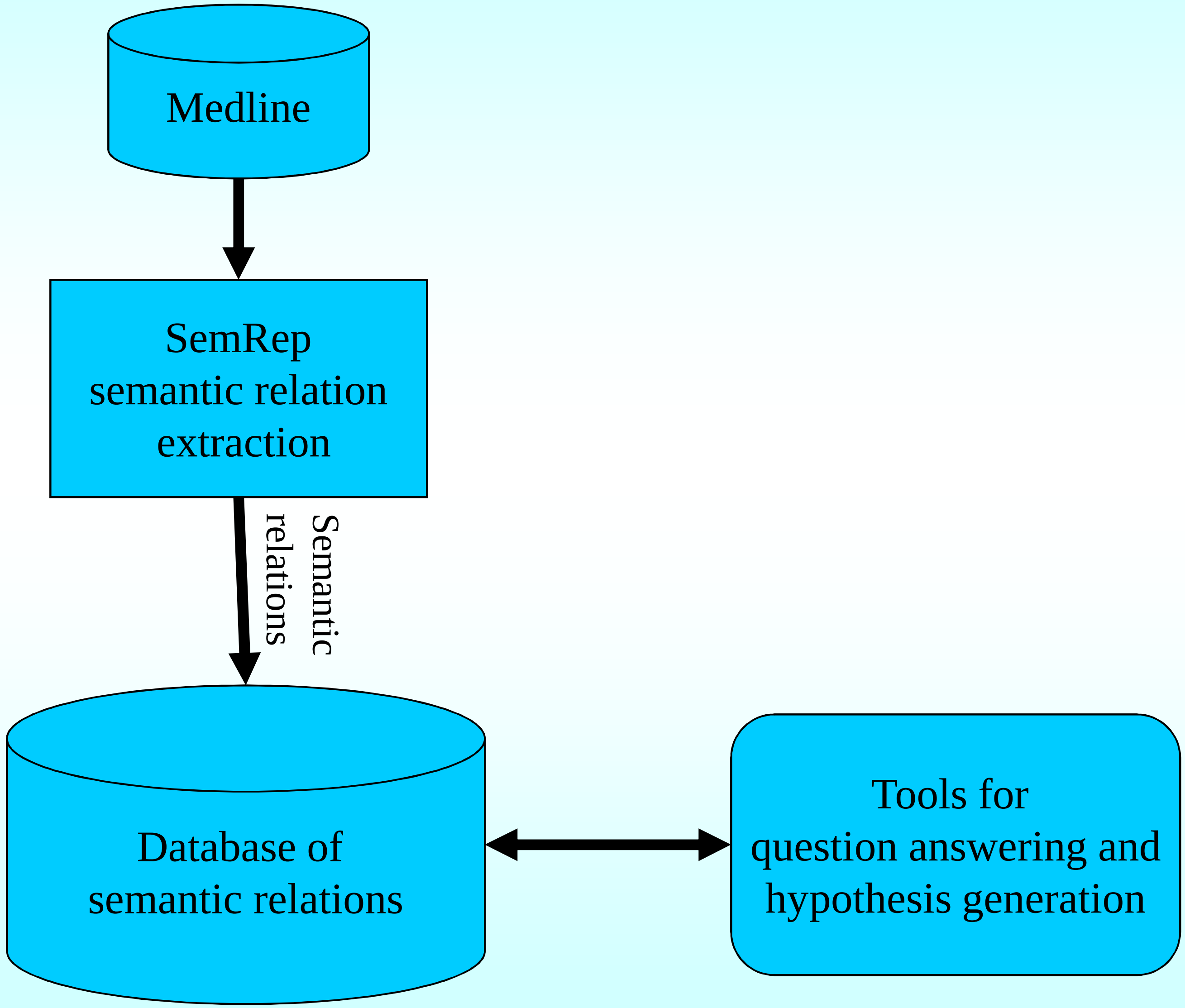
# Information Retrieval Tools

- Fast and robust, but:
- Return hits (bibliographic records) as results
- Users must read the returned hits to extract the facts (<u>answers</u>)
- Users cannot ask: Which drugs are used to treat disease X?
- But only: Find relavant <u>articles</u>, which talk about how to treat disease X!

# Question Answering

- Allows more precise questions with relations between concepts:
  - Which drugs are used to treat disease X?
  - Which diseases are treated with drug Y?
  - What is causing X?
- First returns facts (answers) and then, on demand, the articles
- Benefits for the user:
  - Less to read
  - Faster and easier to more precise answers

# Our Approach - SemBT



# What is used to treat Alzheimer's disease

Relations found: 1145

## SemBT Biomedical Question Answering and Discovery

Semantic relation search

Query:

TREATS Alzheimer's disease

Expand: none Filters:

Microarray Filter

Experiment: none . Limit arguments any to top N 100

all genes at p <= 0.0001 .

Search

*Semantic Relations:*

| Subject | Sem Relation | Object | Frequency |
|---|---|---|---|
| Cholinesterase Inhibitors | TREATS | Alzheimer's Disease | 443 |
| donepezil | TREATS | Alzheimer's Disease | 404 |
| Acetylcholinesterase Inhibitors | TREATS | Alzheimer's Disease | 340 |
| Memantine | TREATS | Alzheimer's Disease | 222 |
| Galantamine | TREATS | Alzheimer's Disease | 216 |
| Tacrine | TREATS | Alzheimer's Disease | 204 |
| rivastigmine | TREATS | Alzheimer's Disease | 196 |
| Intervention regimes | TREATS | Alzheimer's Disease | 165 |
| Immunotherapy | TREATS | Alzheimer's Disease | 132 |
| Assessment procedure | TREATS | Alzheimer's Disease | 119 |
| Application procedure | TREATS | Alzheimer's Disease | 110 |
| Therapeutic agent (substance) | TREATS | Alzheimer's Disease | 92 |
| Diagnosis | TREATS | Alzheimer's Disease | 90 |
| Expression procedure | TREATS | Alzheimer's Disease | 79 |
| Antioxidants | TREATS | Alzheimer's Disease | 75 |
| Pharmacotherapy | TREATS | Alzheimer's Disease | 56 |
| Physostigmine | TREATS | Alzheimer's Disease | 55 |

# Evidence for the Answers

Evaluate   You are signed in as: mitko

| donepezil | TREATS | Alzheimer's Disease | Correc |
|---|---|---|---|
| The acetylcholinesterase inhibitors, **donepezil** , galantamine and rivastigmine have been approved for the **treatment** of mild-to-moderate **Alzheimer's dementia** , the NMDA inhibitor memantine is approved for moderate-to-severe Alzheimer's disease. (PMID: 15376702) | | | - |
| RESULTS: Saffron at this dose was found to be effective similar to **donepezil** in the **treatment** of mild-to-moderate **AD** after 22 weeks. (PMID: 19838862) | | | - |
| **Donepezil for Alzheimer's disease** in clinical practice--The DONALD Study. (PMID: 15084792) | | | - |
| AD2000: **donepezil in Alzheimer's disease** . (PMID: 15220027) | | | - |
| **Donepezil for** severe **Alzheimer's disease** . (PMID: 16581383) | | | - |
| Metaanalysis of randomized trials of the efficacy and safety of **donepezil** , galantamine, and rivastigmine for the **treatment** of **Alzheimer disease** . (PMID: 15249273) | | | - |
| OBJECTIVES: To compare directly, in the same patient cohort, the ease of use and tolerability of **donepezil** and galantamine in the **treatment** of **Alzheimer's disease** (AD), and investigate the effects of both treatments on cognition and activities of daily living (ADL). (PMID: 14716700) | | | - |
| Data were pooled from three randomized, placebo-controlled trials of **donepezil for** severe **AD** to further evaluate treatment effects and overall tolerability/safety. (PMID: 19735164) | | | - |
| BACKGROUND: Cholinesterase inhibitors, such as galantamine, **donepezil** and rivastigmine are approved for symptomatic **treatment** of **Alzheimer's Disease** (AD) in Canada. (PMID: 14675494) | | | - |
| **Donepezil** for the **treatment** of mild to moderate **Alzheimer's disease** in France: the economic implications. (PMID: 14560059) | | | - |
| A long-term comparison of galantamine and **donepezil** in the **treatment** of **Alzheimer's disease** . (PMID: 12875613) | | | - |
| [Effect of memantine treatment on patients with moderate-to-severe Alzheimer's disease treated with donepezil]. (PMID: 20201245) | | | - |
| Are there long-term benefits of **donepezil in Alzheimer's disease** ? (PMID: 15559925) | | | - |
| To create a reproducible observation, the sentences occurring at five specific text sites in all 18 RCTs of **donepezil for AD** were tabulated, as were study design, sources of financial support, and outcomes that could | | | - |

# Argument expansion and faceting

Relations found: 15849

## SemBT Biomedical Question Answering and Discovery

**Filters:**

***Subject:***

| Value | Count |
|---|---|
| Lesion | 447 |
| Genes | 59 |
| TP53 | 54 |
| TP53 gene | 53 |
| Smoking | 52 |
| Single Nucleotide Polymorphism | 46 |
| response | 43 |
| Alcohol consumption | 40 |
| Antimicrobial susceptibility | 40 |
| Cigarette Smoking | 40 |
| Observation parameter | 40 |
| MicroRNAs | 38 |
| Proteins | 38 |
| CDKN2A | 37 |
| Consumption-archaic term for TB | 37 |
| Obesity | 37 |
| DNA | 35 |
| Body mass index | 34 |
| EGFR | 34 |
| Behavior | 32 |
| Estrogens | 32 |
| GSTM1 | 32 |
| Carrier of disorder | 31 |

Semantic relation search

Query: arg_name:neoplasms AND relation:PREDISPOSES

Expand: both Filters: ☑

Microarray Filter

Experiment: none . Limit arguments any to top N 100

all genes at p <= 0.0001 .

Search

*Semantic Relations:*

| Subject | Sem Relation | Object | Frequency |
|---|---|---|---|
| Prostate-Specific Antigen | PREDISPOSES | Malignant neoplasm of prostate | 262 |
| Estrogens | PREDISPOSES | Malignant neoplasm of breast | 209 |
| Smoking | PREDISPOSES | Malignant neoplasm of lung | 199 |
| Obesity | PREDISPOSES | Malignant neoplasm of breast | 155 |
| Fibrosis | PREDISPOSES | Primary carcinoma of the liver cells | 148 |
| Alcohol consumption | PREDISPOSES | Malignant neoplasm of breast | 122 |
| Helicobacter Infections | PREDISPOSES | Malignant neoplasm of stomach | 111 |
| BRCA1\|BRCA1 gene | PREDISPOSES | Malignant neoplasm of breast | 110 |
| Carrier of disorder | PREDISPOSES | Malignant neoplasm of breast | 91 |
| Sun Exposure | PREDISPOSES | melanoma | 83 |
| Smoker | PREDISPOSES | Malignant neoplasm of lung | 81 |
| Single Nucleotide Polymorphism | PREDISPOSES | Malignant neoplasm of breast | 79 |
| Cigarette Smoking | PREDISPOSES | Malignant neoplasm of lung | 78 |
| Physical activity | PREDISPOSES | Malignant neoplasm of breast | 77 |
| Genes | PREDISPOSES | Malignant neoplasm of breast | 76 |
| Simian B disease | PREDISPOSES | Primary carcinoma of the liver cells | 74 |

# Conclusion

- Tool for biomedical question answering presented
- Based on semantic relations extracted from the biomedical literature
- Able to answer various biomedical questions
- Complementary to existing information retrieval systems
- Available at: **http://sembt.mf.uni-lj.si**
- Paper: Hristovski2015, BMC Bioinformatics 2015 16:6

# Incompleteness versus Inconsistency in Data Mining

**Jerzy W. Grzymala-Busse'''**

' University of Kansas, Lawrence, KS 66045, USA

'' Department of Expert Systems and Artificial Intelligence,

University of Information Technology and Management, 35-225 Rzeszow, Poland

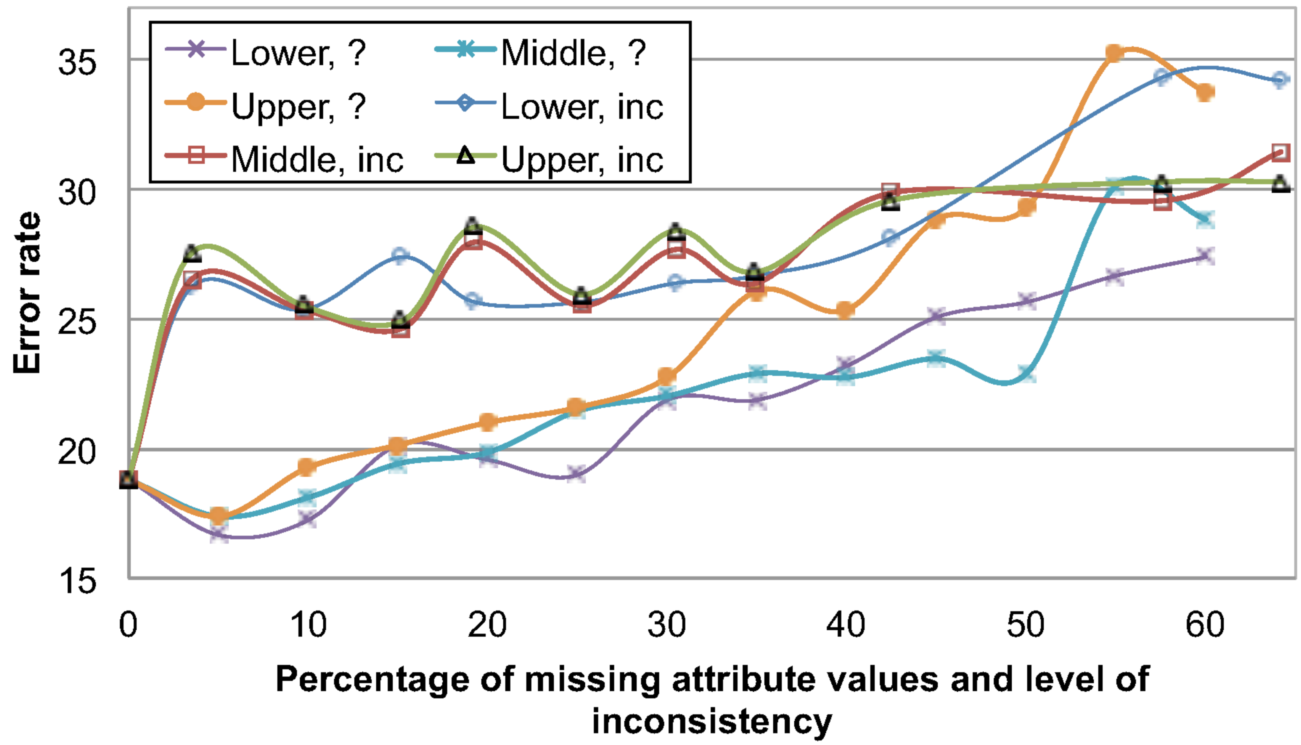
# An Incomplete Data Set

| | Attributes | | | Decision |
|---|---|---|---|---|
| Case | Age | Cholesterol | Weight | Risk |
| 1 | ? | 180..210 | 140..170 | low |
| 2 | 45..60 | ? | 170..210 | low |
| 3 | 20..45 | ? | ? | low |
| 4 | 45..60 | 180..210 | 170..210 | low |
| 5 | 45..60 | ? | 170..210 | high |
| 6 | ? | 210..220 | ? | high |
| 7 | 45..60 | 180..210 | ? | high |

# An Inconsistent Data Set

| | Attributes | | | Decision |
|---|---|---|---|---|
| Case | Age | Cholesterol | Weight | Risk |
| 1 | 20..45 | 180..210 | 140..210 | low |
| 2 | 45..60 | 180..210 | 140..210 | low |
| 3 | 45..60 | 180..210 | 140..210 | low |
| 4 | 45..60 | 210..220 | 140..210 | high |
| 5 | 45..60 | 180..210 | 210..220 | high |
| 6 | 20..45 | 210..220 | 140..210 | high |
| 7 | 20..45 | 210..220 | 140..210 | low |

# *Australian* data set



Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

**The Eight International Conference on Advances in Databases, Knowledge, and Data Applications**

**June 26 - 30, 2016 - Lisbon, Portugal**

# PANEL on DBKDA/GraphSM/WEB

# Topic: Current Trends on Information Searching/Query Answering

**Andreas Schmidt**

**Department of Informatics and**
**Business Information Systems**
**University of Applied Sciences Karlsruhe**
**Germany**

**Institute for Applied Computer Sciences**
**Karlsruhe Institute of Technologie**
**Germany**

# Semantic Search

- Understanding the semantic of text (content analysis) is an essential key asset for advanced searching
- To understand the semantic of text, we have to determine
  - the identity of the (main) entities (i.e. Paul -> 'Pope John Paul')
  - the relations between the identified entities ( <a> lives_in <b>)
  - the hierarchies of the identified relations

Named Entity Disambiguation

# Named Entity Recognition (NER)

- Discovering single-word or multi-word phrase entities (mentions) in text, like
  - Persons
  - Organizations
  - Locations
  - Temporal expressions
  - Works of Art
  - Product names

accuracy > 90%
Popular system:
Stanford NER [FGM05]

- Approaches based on
  - handcrafted, rule based algorithms
  - linguistic grammar-based techniques
  - statistical models (machine learning)

## Example (from [AIDA])

Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.

Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

# Step 1 - NER

**mentions**

Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.

# Step 2- NED

[Michael Jackson]

[Michael Moore]

[Michael Stonebraker]

[Michael Pollan]

[ ... ]

Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.

[ ... ]

[William Berkeley (governor)]

[Berkeley, California]

[University of California, Berkeley]

[Berkeley Square]

[Lennox Berkeley]

Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.

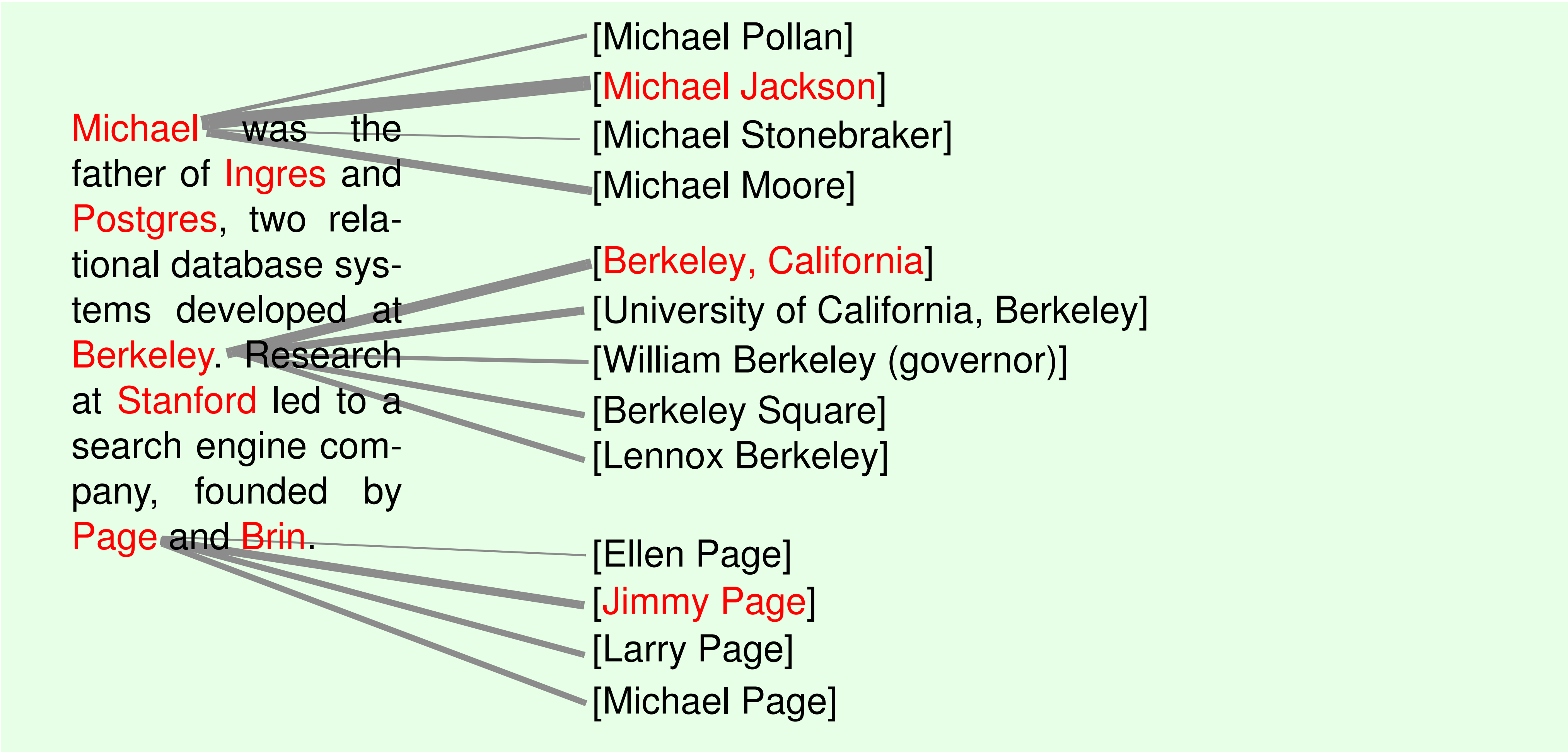
[Michael Pollan]
[Michael Jackson]
[Michael Stonebraker]
[Michael Moore]

[Berkeley, California]
[University of California, Berkeley]
[William Berkeley (governor)]
[Berkeley Square]
[Lennox Berkeley]

[Ellen Page]
[Jimmy Page]
[Larry Page]
[Michael Page]

# **NED -** How to disambiguate a mention?

- Baseline:
  - Choose the most prominent entity (longest wiki article, arti-cle with most inlinks, biggest city, ...)
  - Choose the entity that uses the mention most frequently as link anchor

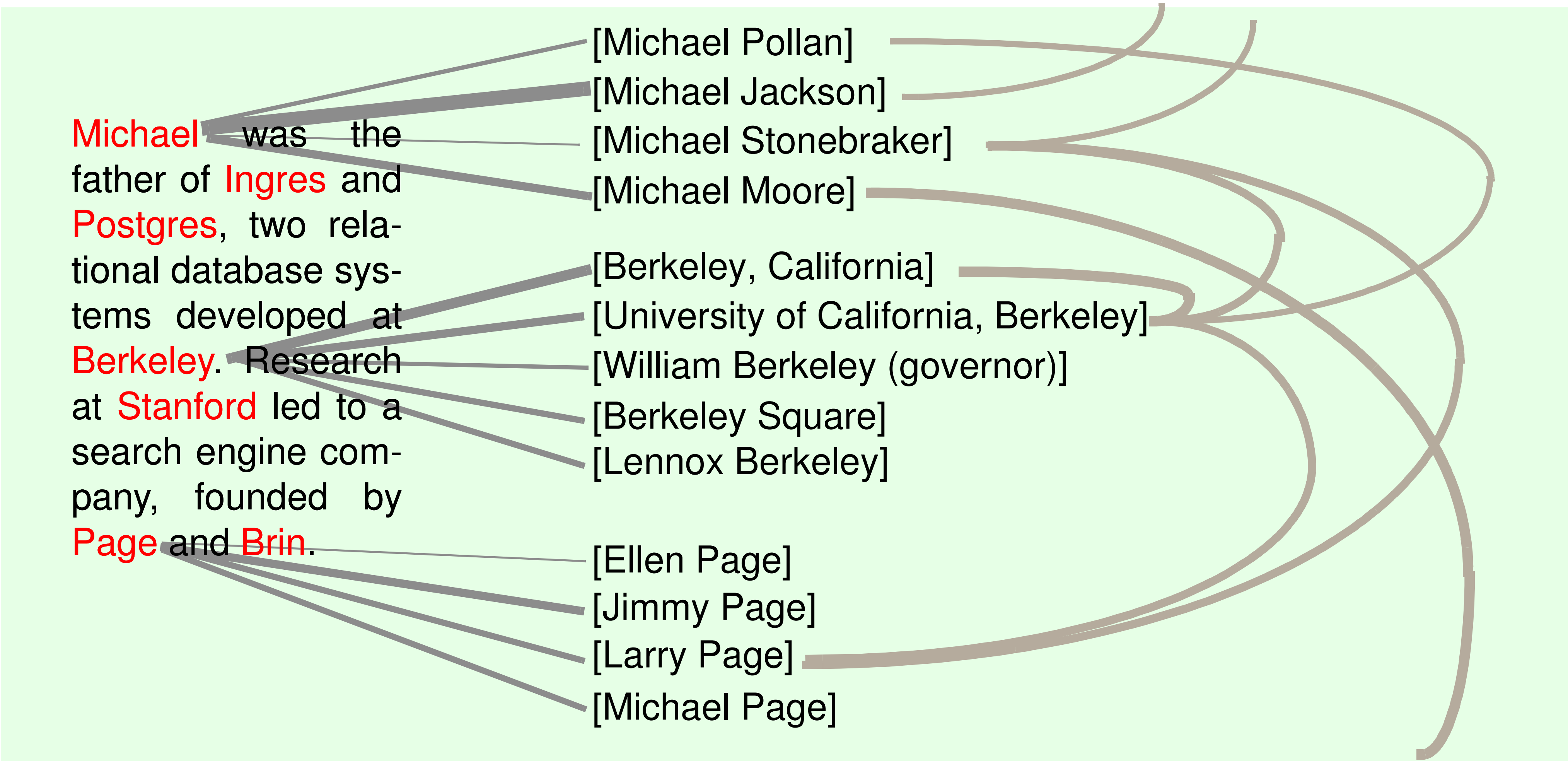

Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.
[Michael Pollan]
[Michael Jackson]
[Michael Stonebraker]
[Michael Moore]
[Berkeley, California]
[University of California, Berkeley]
[William Berkeley (governor)]
[Berkeley Square]
[Lennox Berkeley]
[Ellen Page]
[Jimmy Page]
[Larry Page]
[Michael Page]

# **NED -** How to disambiguate a mention?

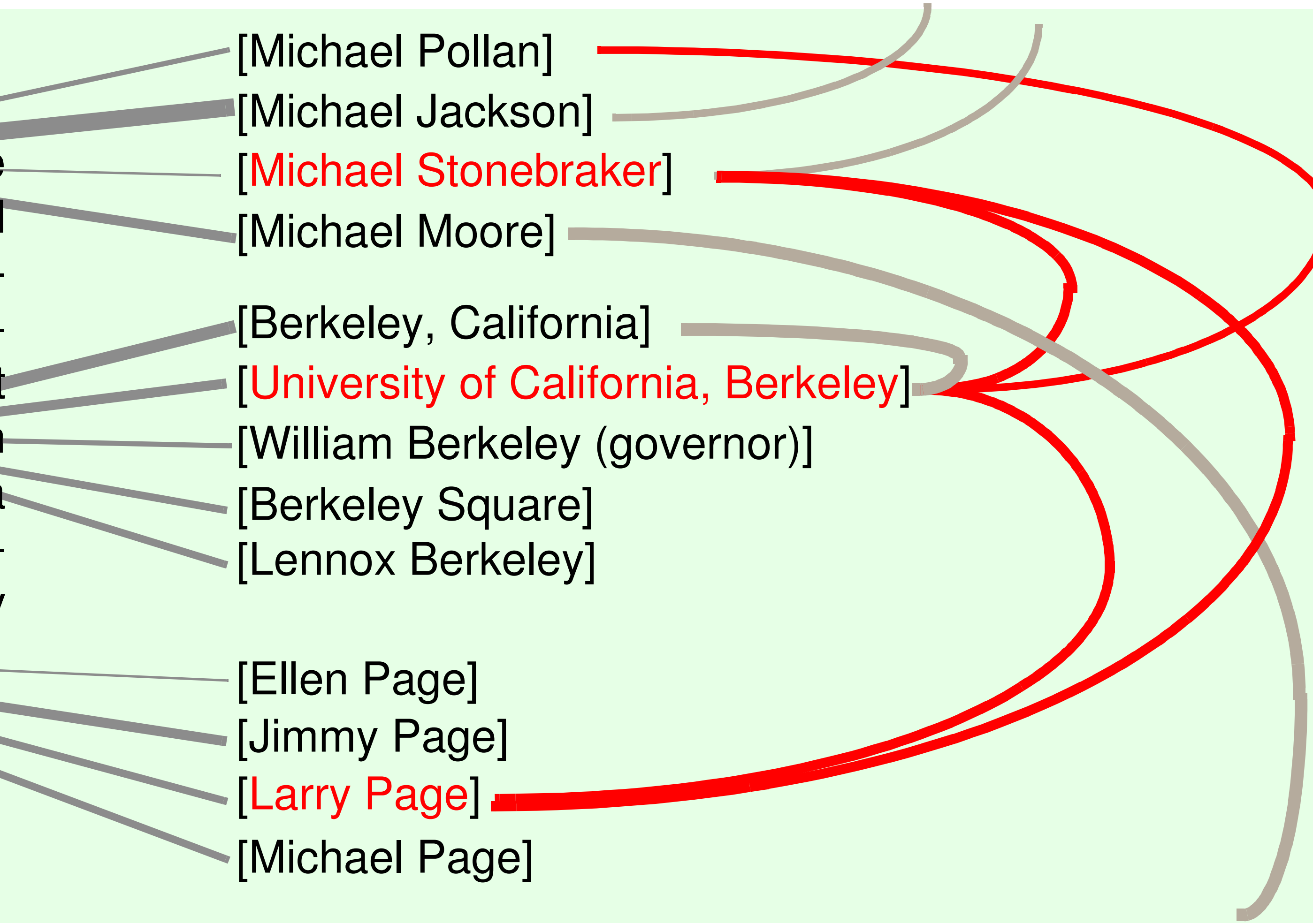
- Baseline:
  - Choose the most prominent entity (longest wiki article, article with most inlinks, biggest city, ...)
  - Choose the entity that uses the mention most frequently as link an-chor
- Improvement:
  - Consider multiple mentions at once and consider the semantic re-latedness between the possible entities

Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES



Michael was the father of Ingres and Postgres, two relational database systems developed at Berkeley. Research at Stanford led to a search engine company, founded by Page and Brin.



Karlsruhe Institute of Technology

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

# Further (difficult) Examples of Disambiguation (from [Hof15])

- Italy recalled Marcello Cuttitta on Friday for their friendly against Scotland at Murrayfield [...].

Murrayfield_Stadium

Scotish_Rugby_Team

Marcello_Cuttitta_(Rugby Player)

Italian_Rugby_Team

# Can you disambiguate this?

When Page played Kashmir at Knebwort, his Les Paul was uniquely tuned

Karlsruhe Institute of Technology

Hochschule Karlsruhe Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

# Can you disambiguate this?

When Page played Kashmir at Knebwort, his Les Paul was uniquely tuned

Jimmy Page

Kashmir_(Song)

E-Guitar

Rock_Festival

# References


[Hof15] J. Hoffart, Discovering and Disambiguating Named Entities in Text, Dissertation, University of Saarland, 2015

[FGM05] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, ACL 2005, University of Michigan, USA, 2005.

[Hof11] Robust Disambiguation of Named Entities in Text. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 2011

[MW08] David Milne and Ian H Witten. An Eective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In Proceedings of the AAAI 2008 Workshop on Wikipedia and Articial Intelligence,Chicago, USA, 2008.

[AIDA] https://gate.d5.mpi-inf.mpg.de/webaida/