VideoLectures Case Study: Supporting Rapid Growth of the Scientific YouTube

Miha Grčar,
Jožef Stefan Institute

Peter Keše,
viidea

Funded by: European Commission – 6th Framework
Project Reference: IST-2004-026460

Outline

- About VideoLectures.net
- Need for semi-automatic categorization

Peter

- Ontology population: TAO approach to categorization
- Evaluation results
- Categorizer in action

Miha

- Value of TAO results for VideoLectures
- VideoLectures in near future

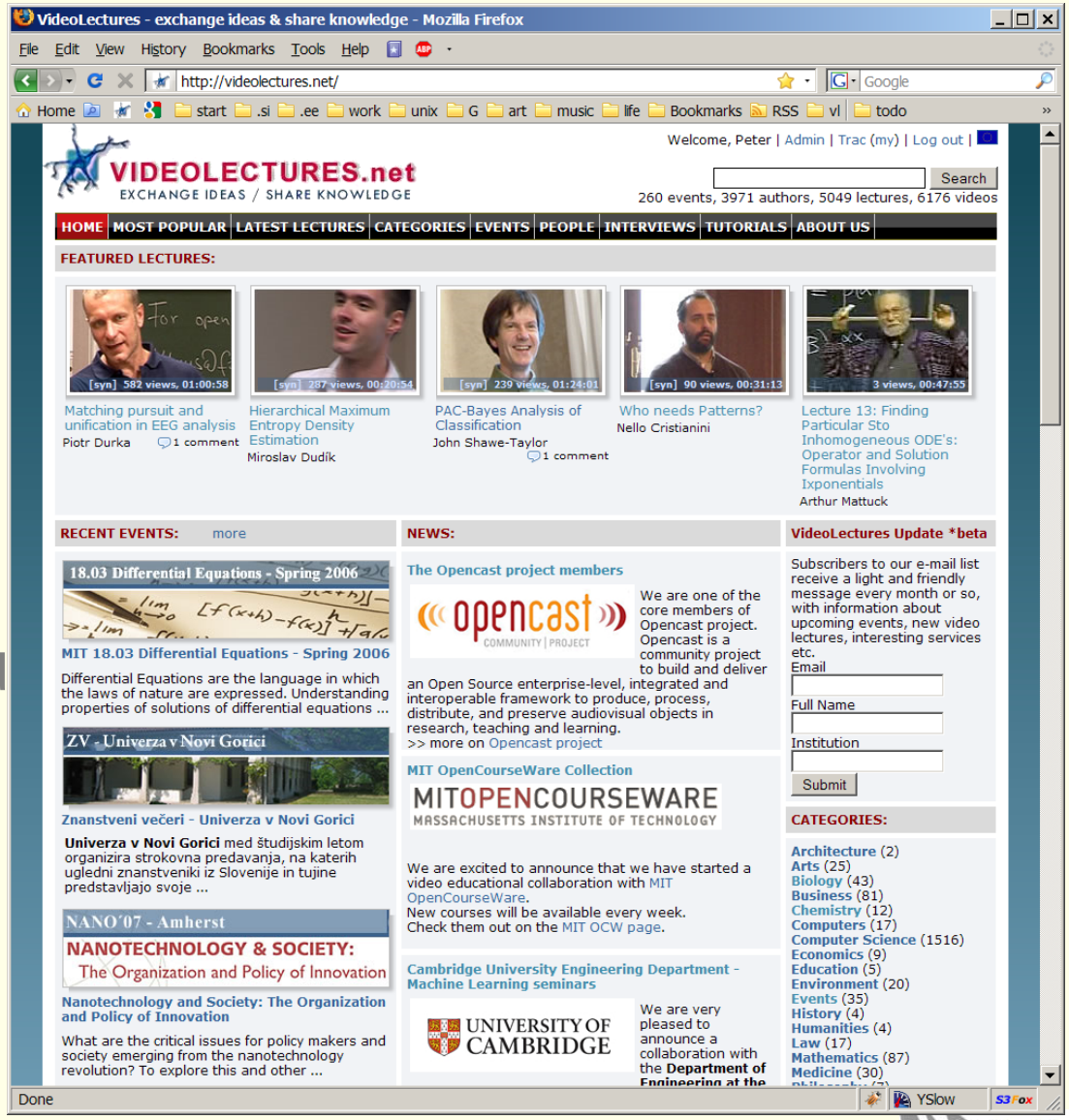
Peter

- Conclusions

Miha

What is VideoLectures.net

- World's largest 'YouTube'-for-science Web site
  - 6,176 videos
  - 3,971 scientists
- Started with coverage of EU research projects
- Now expanding to educational content and non-EU institutions (MIT, CMU, Yale, Korea, CERN) and conferences (ACM, ICWS, ACL, ICML, ECCS, ...)

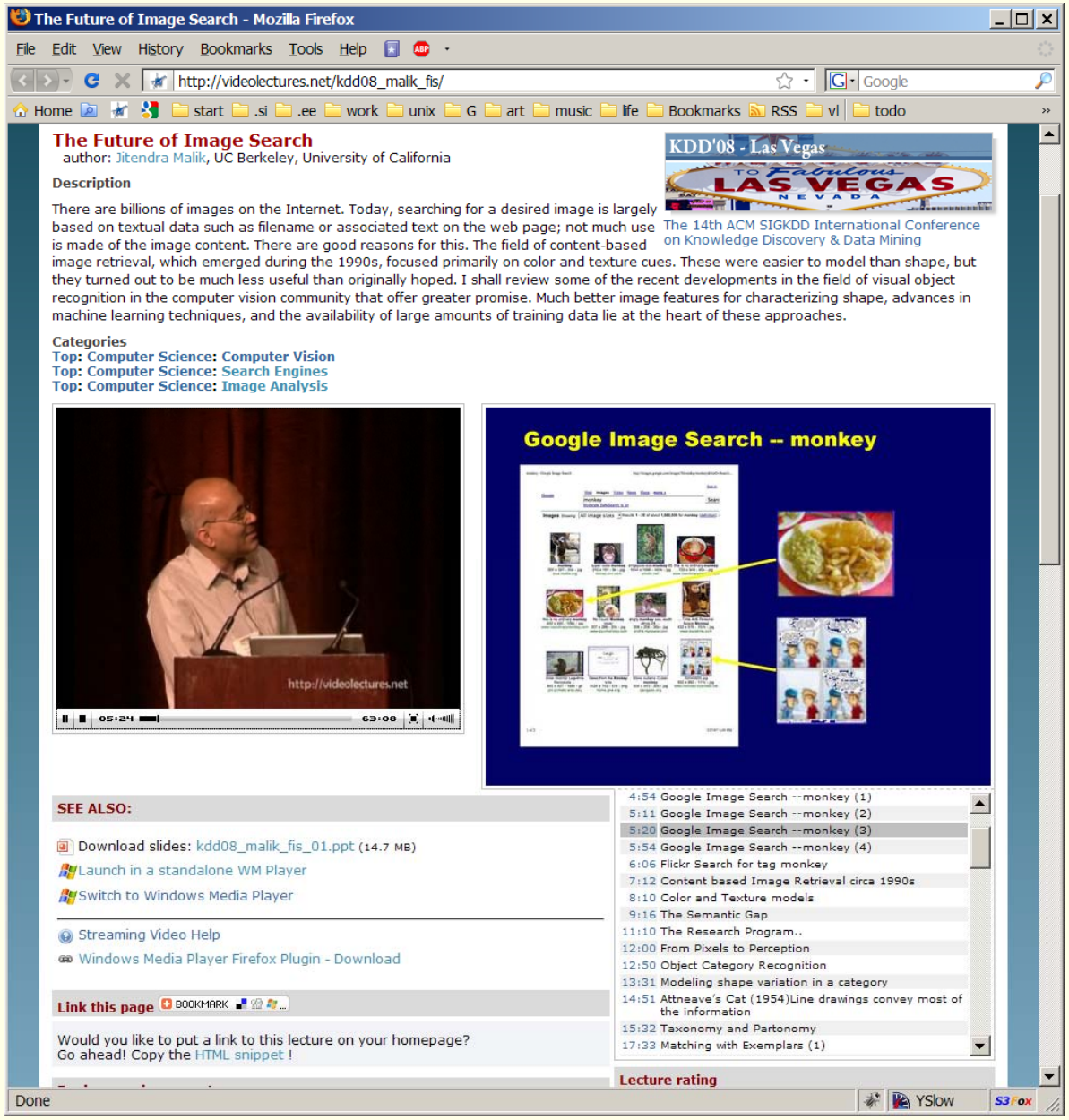


VideoLectures status

- 2 years of existence
- 1.2 million visitors from all over the world
- 3,300 registered users

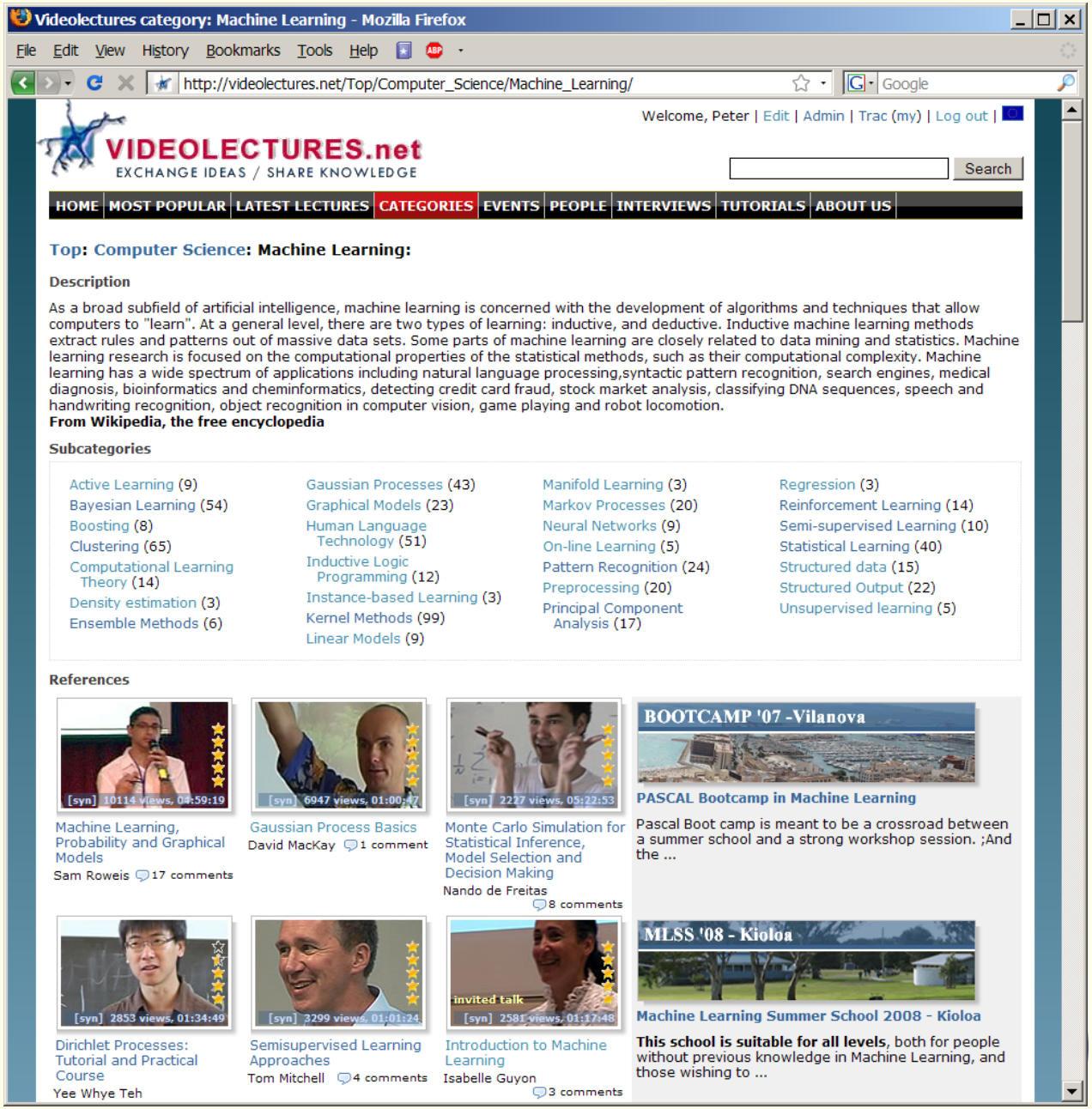
More than just your typical video web site

- Presentations of scientific articles and lectures
- Video + Slides (synchronized)
- Lots of valuable meta-data…
  - Author
  - Institution
  - Abstract
  - Event
  - Slides
  - Downloads
  - Categories



Categories and browsing

- Large and Dynamic Taxonomy
- 200 categories
- 2,200 lectures are categorized



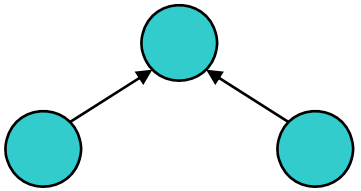
Problem statement

- Manual categorization is difficult and time consuming
- Categorization is important – it allows the users to efficiently browse the content
- Need semi-automatic support for categorization

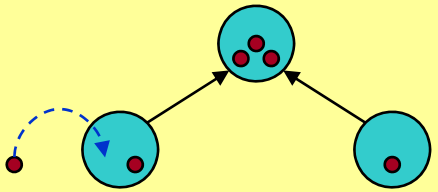
Ontology learning (TAO WP2)

1. Concept identification

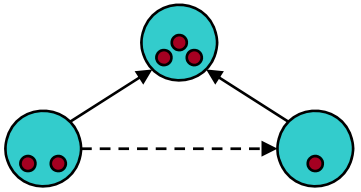
2. Subsumption hierarchy induction



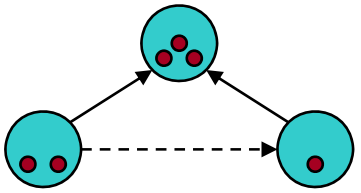
3. Ontology population



4. Relation discovery



5. Additional axiomatization



Task

Kalina Bontcheva, University of Sheffield

Title: **Working with ontologies**

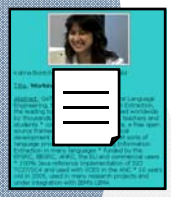
Abstract: GATE is: the Eclipse of Natural Language Engineering, the Lucene of Information Extraction, the leading toolkit for Text Mining * used worldwide by thousands of scientists, companies, teachers and students * comprised of an architecture, a free open source framework (or SDK) and graphical development environment * used for all sorts of language processing tasks, including Information Extraction in many languages * funded by the EPSRC, BBSRC, AHRC, the EU and commercial users * 100% Java reference implementation of ISO TC37/SC4 and used with XCES in the ANC * 10 years old in 2005, used in many research projects and under integration with IBM's UIMA

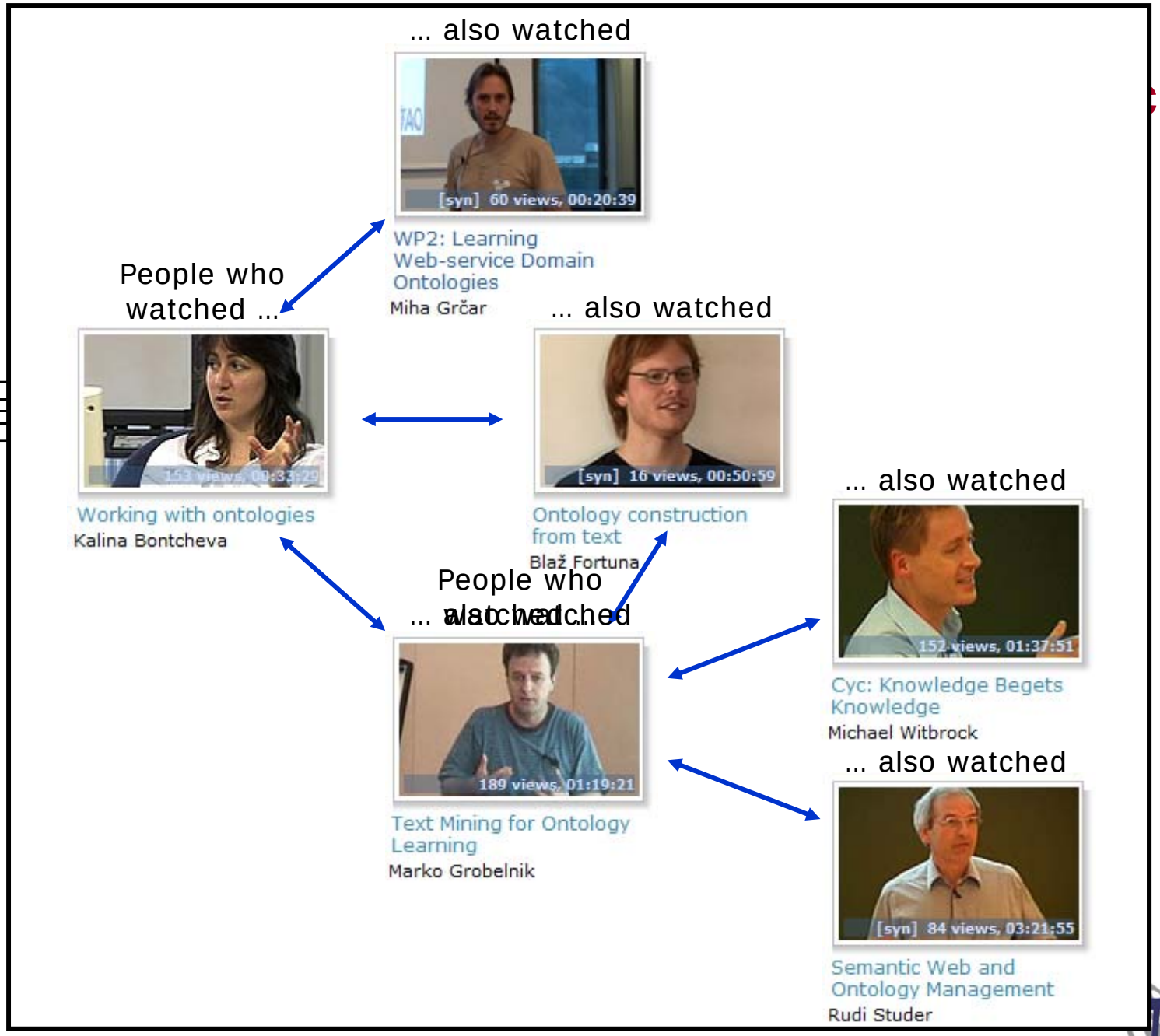
Where?

CATEGORIES:

Arhitecture (2)
Arts (21)
Biology (10)
Business (37)
Chemistry (11)
Computers (10)
Computer Science (1068)
Economics (5)
Education (2)
Environment (8)
Events (27)
History (2)
Law (5)
Mathematics (63)
Medicine (17)
Physics (3)
Psychology (1)
Science (7)
Society (27)
Technology (4)

Approach




... also watched
[syn] 60 views, 00:20:39
WP2: Learning Web-service Domain Ontologies
Miha Grčar
People who watched ...
... also watched
153 views, 00:33:29
Working with ontologies
Kalina Bontcheva
[syn] 16 views, 00:50:59
Ontology construction from text
Blaž Fortuna
People who watched ...
... also watched
... also watched
152 views, 01:37:51
Cyc: Knowledge Begets Knowledge
Michael Witbrock
189 views, 01:19:21
Text Mining for Ontology Learning
Marko Grobelnik
... also watched
[syn] 84 views, 03:21:55
Semantic Web and Ontology Management
Rudi Studer

WP2: Learning Web-service Domain Ontologies
Miha Grčar

Working with ontologies
Kalina Bontcheva

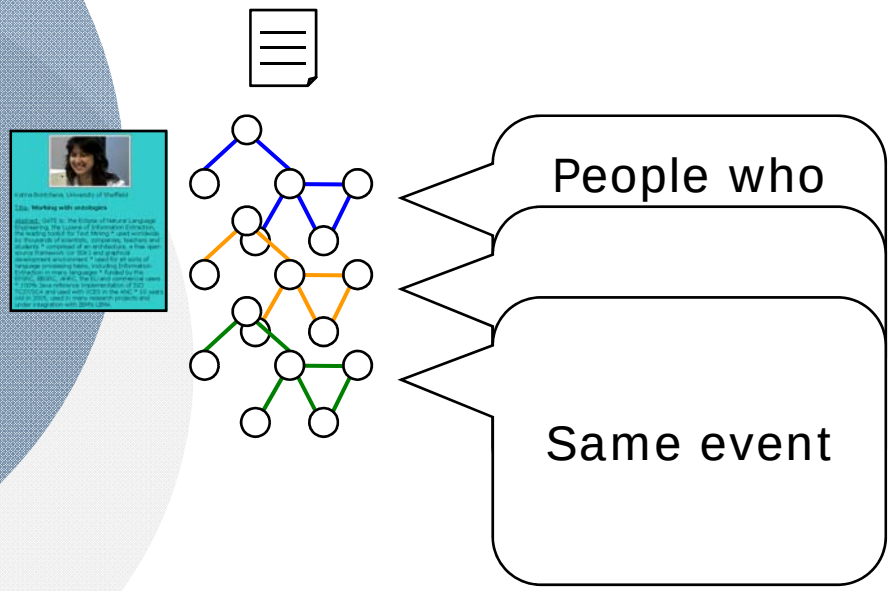
Ontology construction from text
Blaž Fortuna

Text Mining for Ontology Learning
Marko Grobelnik

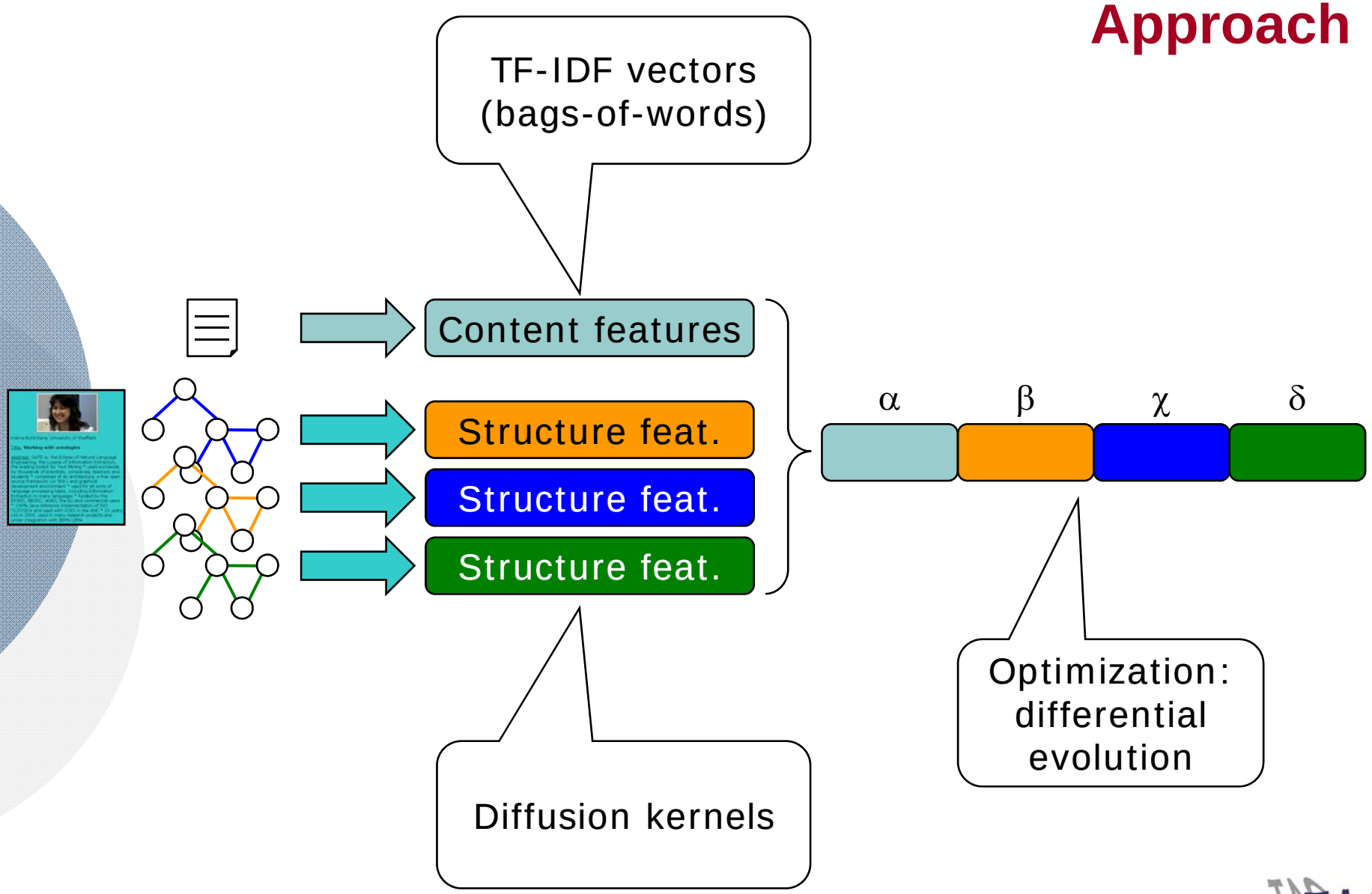
Cyc: Knowledge Begets Knowledge
Michael Witbrock

Semantic Web and Ontology Management
Rudi Studer

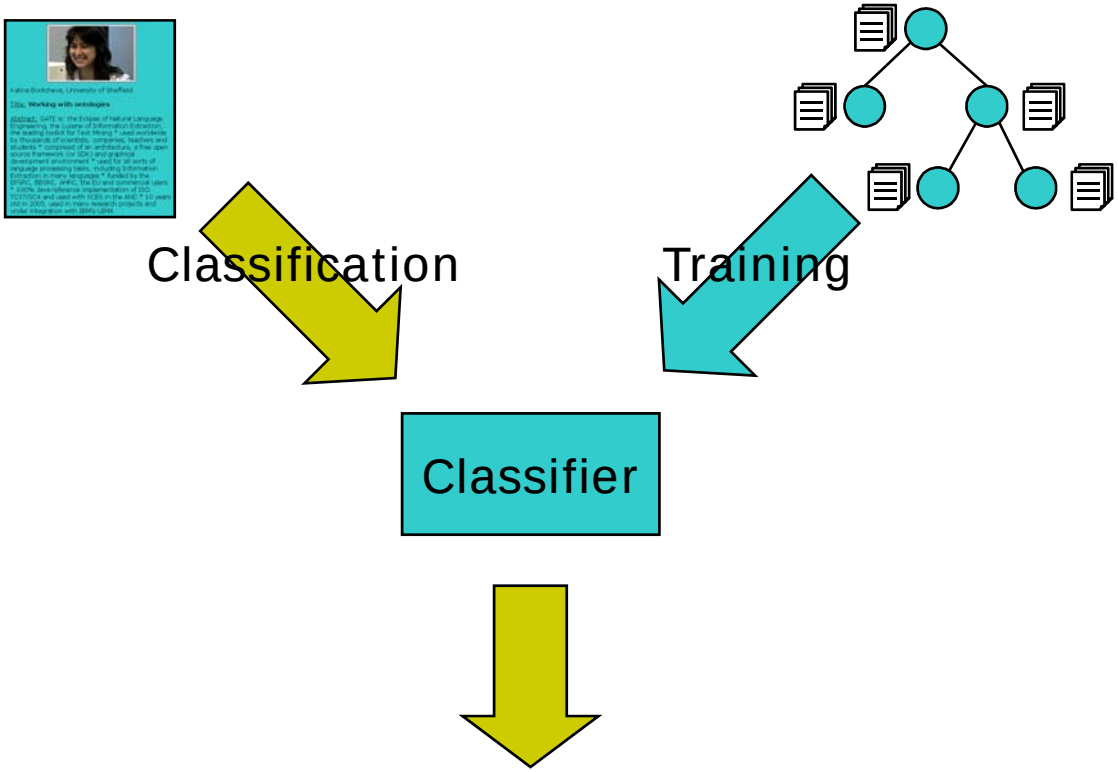
Approach



Approach

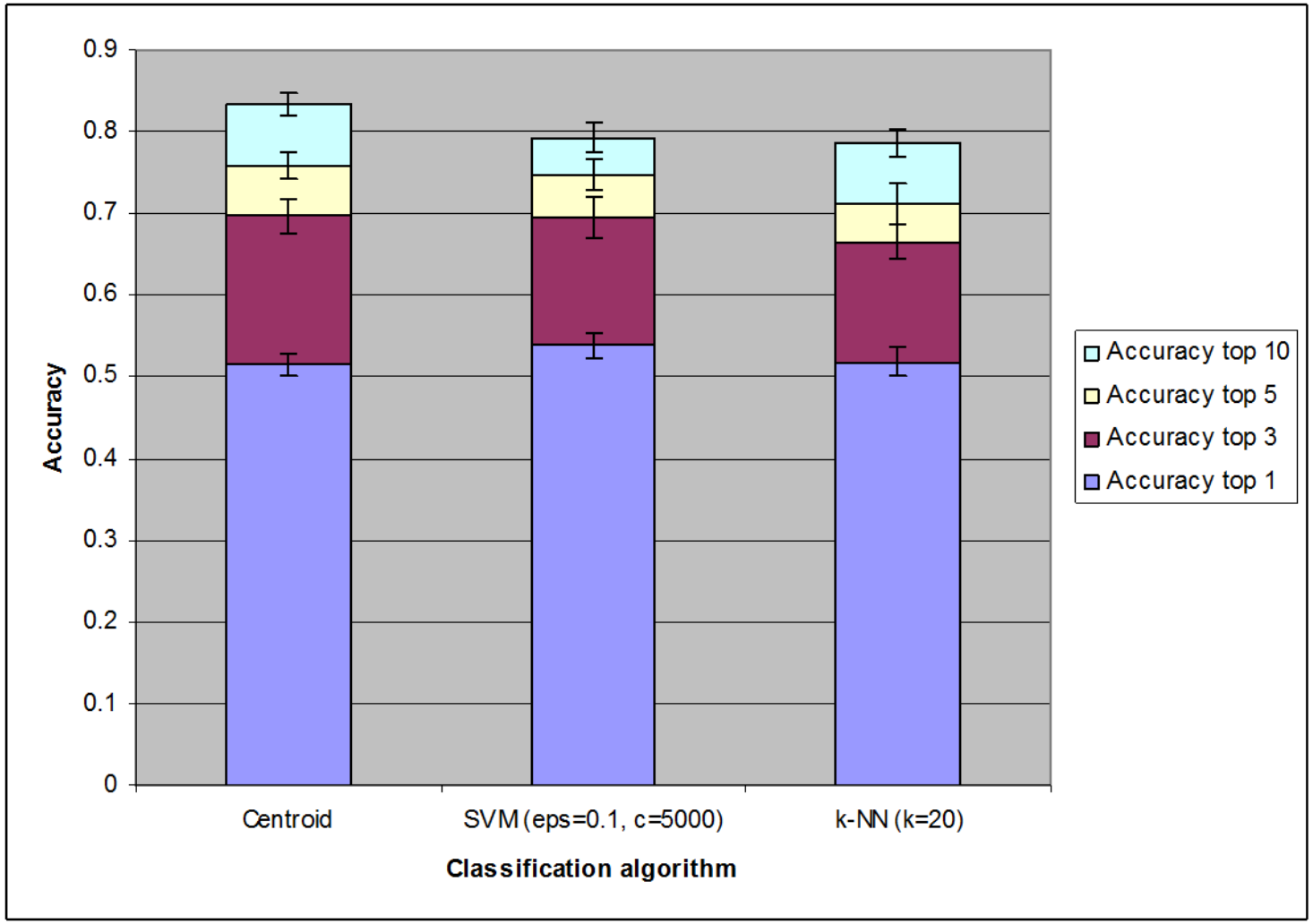

TF-IDF vectors
(bags-of-words)
Content features
Structure feat.
Structure feat.
Structure feat.
α
β
χ
δ
Optimization:
differential
evolution
Diffusion kernels

Approach

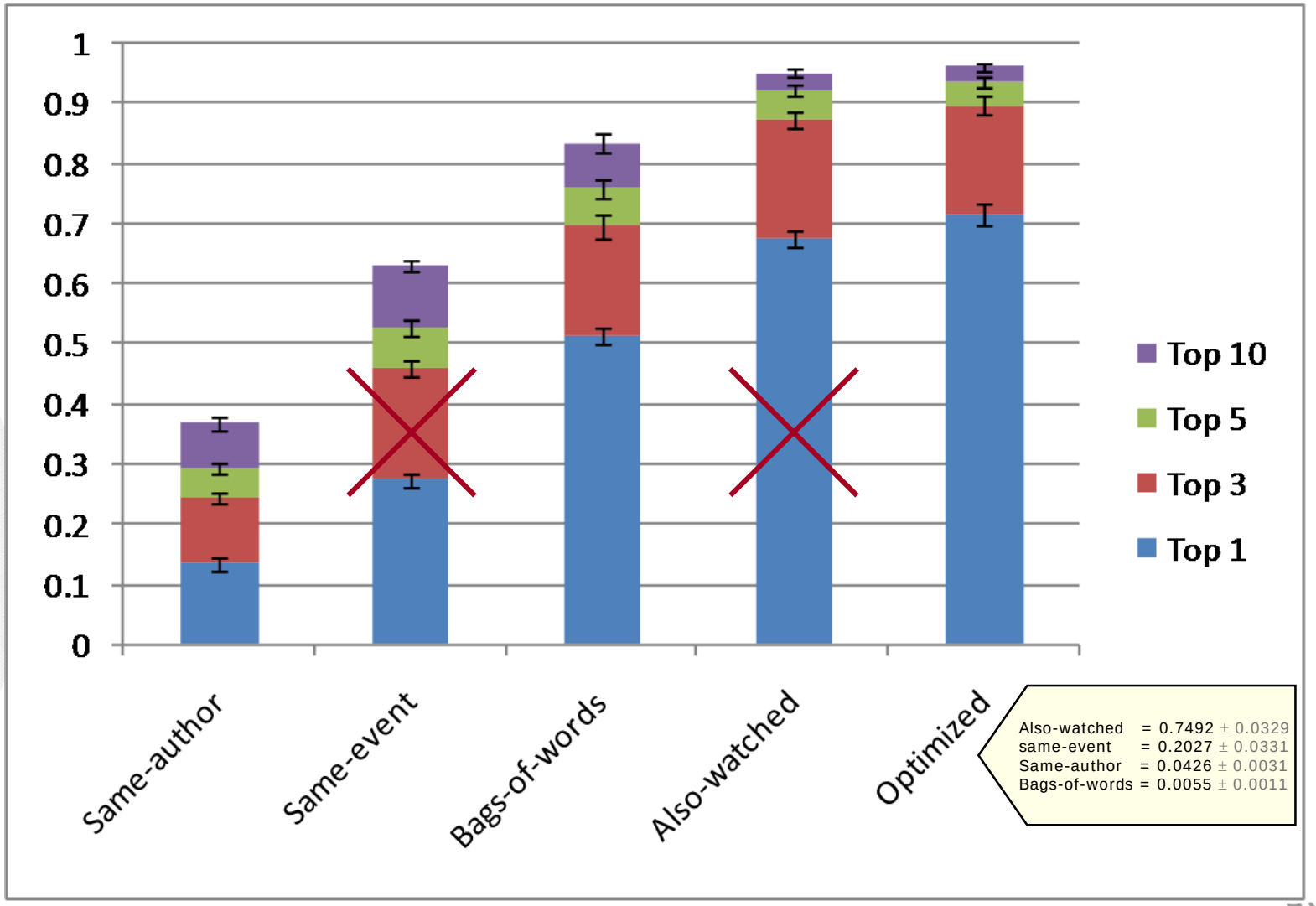


1. Computer science / Semantic Web / Ontologies
2. Computer Science / Semantic Web
3. Computer Science / Software Tools
4. Computer Science / Information Extraction
5. Computer Science / Natural Language Processing
6. …

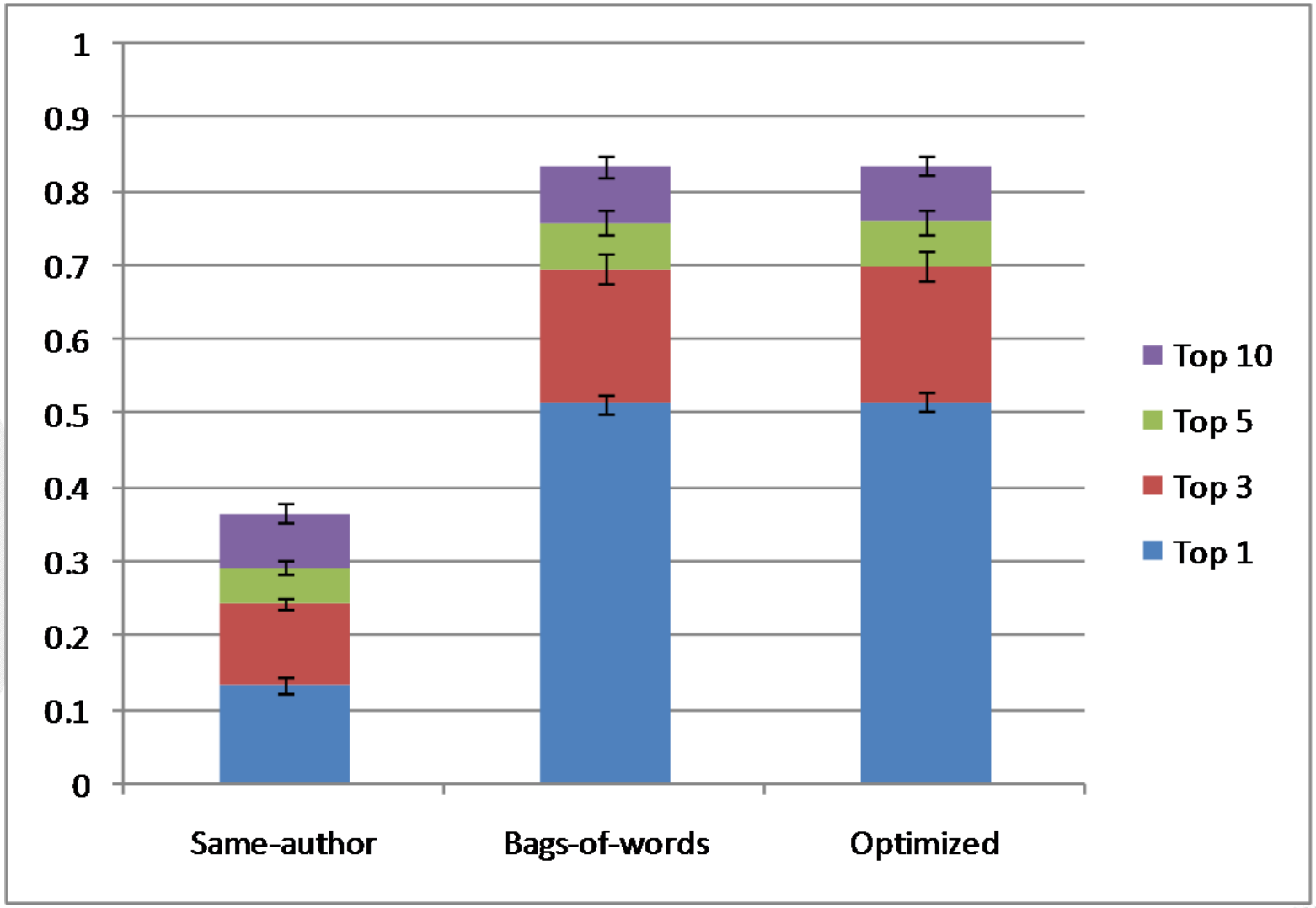
Evaluation Baseline



Evaluation
Final Results



Evaluation
Final Results



DBpedia: A Nucleus for a Web of Open Data - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://dev.videolectures.net/iswc07_soeren_nwo/

Google

Getting Started Latest Headlines

DBpedia: A Nucleus for a Web of Open Data

author: Sören Auer, University of Pennsylvania

Description

*You have **authorship rights** on this lecture, therefore you can edit its **description**...*

Categories [edit...]

Tutorials

Categorize this lecture:
navigate to chosen category and click the 'Add' button.

Top

Architecture	Computer Science	Innovation Journalism	Psychology
Arts	Economics	Law	Recreation
Biology	Education	Mathematics	Regional
Business	Environment	Medicine	Science
Chemistry	Events	Philosophy	Society
Computers	History	Physics	Technology

Quick links:

- Top: Computer Science: Semantic Web
- Top: Computer Science: Semantic Web: Ontologies
- Top: Computer Science: Semantic Web: Semantic Web Services
- Top: Computer Science: Web Mining
- Top: Events
- Top: Computer Science: Machine Learning: Unsupervised learning
- Top: Computer Science: Machine Learning: Human Language Technology
- Top: Computer Science: Data Mining
- Top: Computer Science: Natural Language Processing
- Top: Computer Science: Semantic Web: Annotation
- Top: Computer Science: Data Visualisation
- Top: Computer Science: Semantic Web: Annotation: Tagging

Add selected category: **Top** Add Close

SEE ALSO:

Launch in a standalone WM Player

Switch to Windows Media Player

Download iswc07_soeren_nwo_01.flv (Flash video 86.0 MB)

Download iswc07_soeren_nwo_01.wmv (Windows Media video 272.6 MB)

Streaming Video Help

Windows Media Player Firefox Plugin - Download

Lecture rating

Done

The categorizer from the perspective of VideoLectures.net

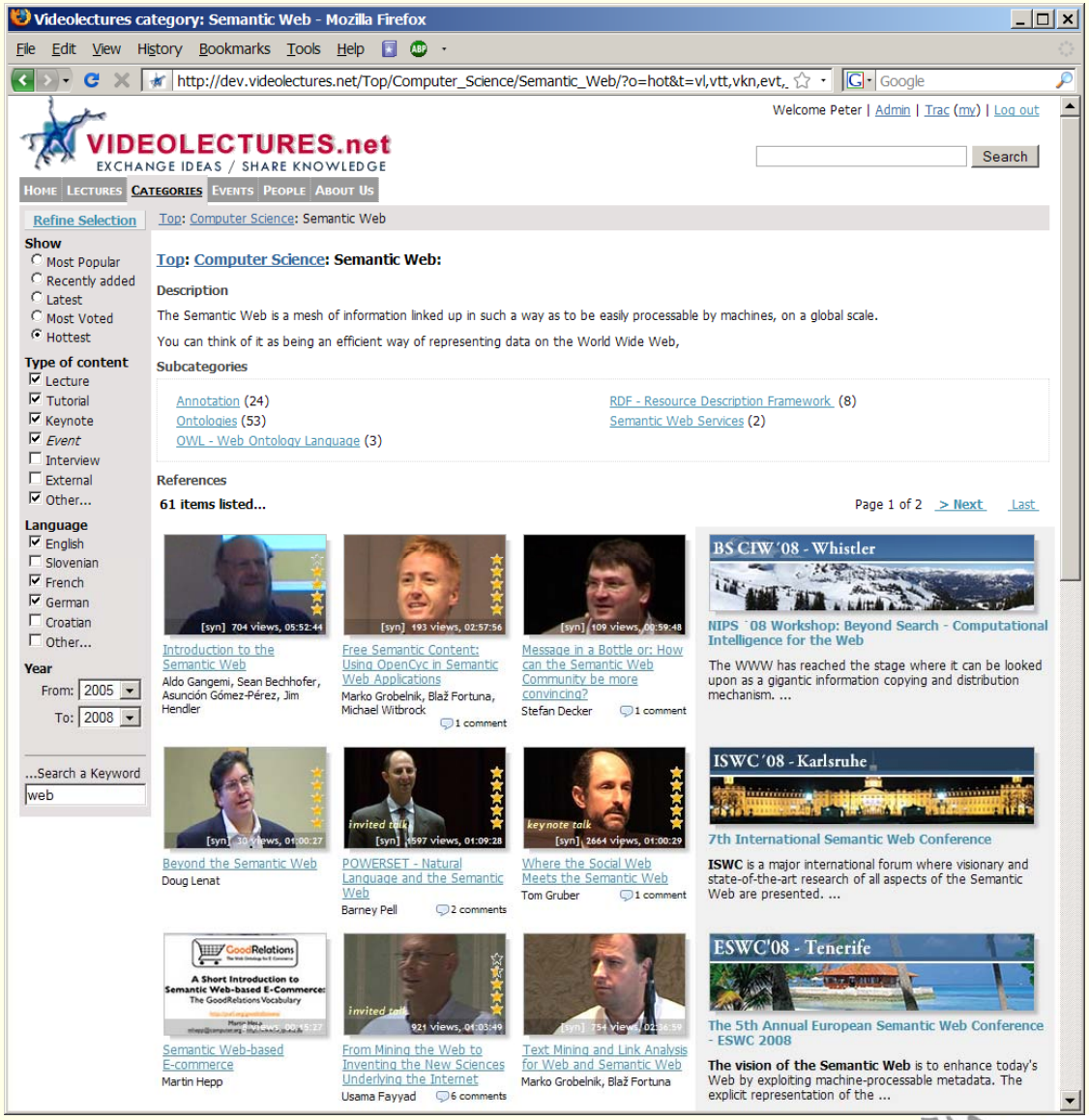
- An extremely valuable contribution to VideoLectures
- Easier for visitors and staff to categorize:
  - More accurate categorization even without deep knowledge of the scientific topic
  - No need to get familiar with the taxonomy
  - In most cases, the categorizer gives the correct suggestion on top-most position

VideoLectures.net already analyzing ways of improvement

- Provide better graph structures to the algorithm
  - collect more user click statistics
- Provide more and better texts
  - Extract more text by parsing slides and articles
  - Use OCR (Optical Character Recognition) on slides and videos
  - Use Speech Recognition

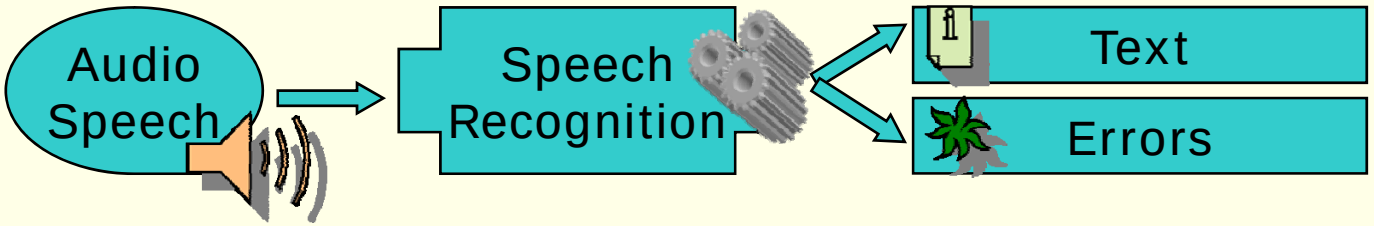
VideoLectures' view of the future

- Determined to provide even better semantic facilities
- Extract more information from the lectures
- Organize content even better
- Provide easier and more functional navigation



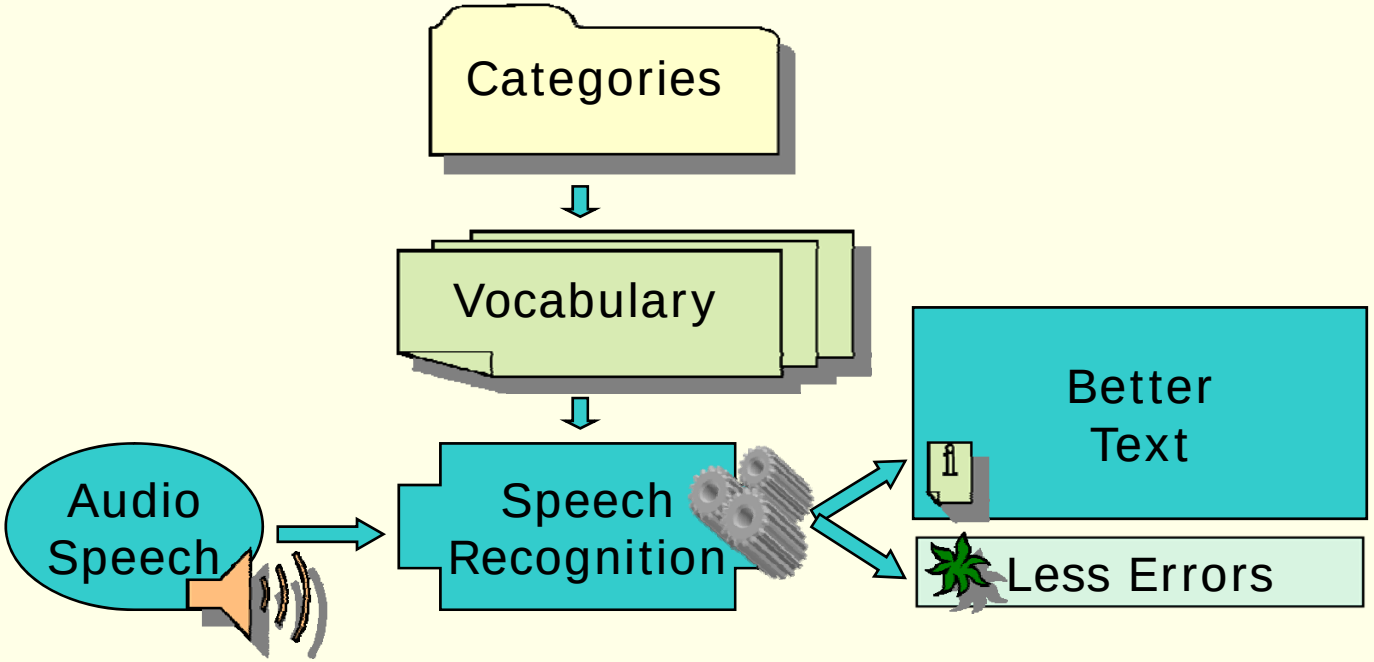
Speech recognition pilot

- Trying to provide keyword search through videos by analyzing speech



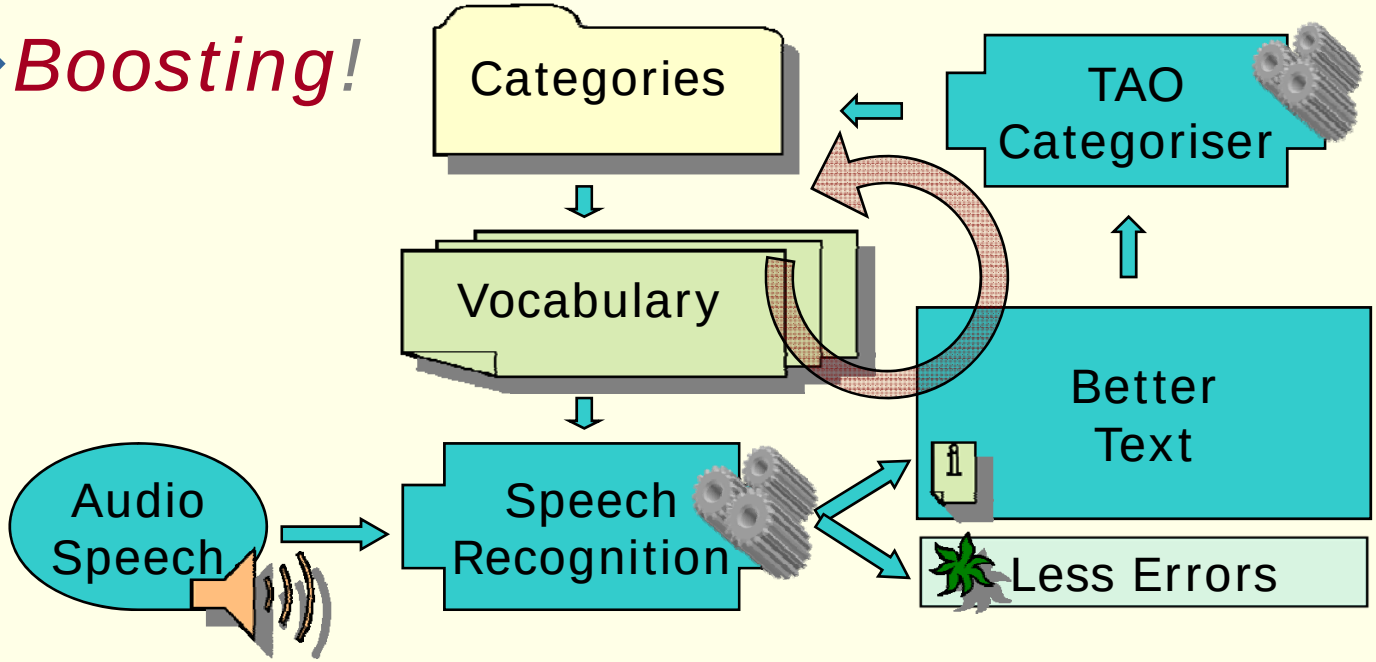
Speech recognition pilot

- Trying to provide keyword search through videos by analyzing speech
- Improving results by using Categories



Speech recognition pilot

- Trying to provide keyword search through videos by analyzing speech
- Improving results by using Categories
- *Boosting!*



Conclusions

- Purpose
  - Support categorization of lectures
- Success!
  - Significant increase in accuracy (12–20% over the baseline based solely on text mining)
  - Robustness in terms of missing data
- Current status and future plans
  - Integrated into the VL authoring module
  - Will be employed to boost the accuracy of the speech-to-text process

That's it

- Thank you for your attention
- Questions?

Check out http://www.VideoLectures.net