# From Web 2.0 to Semantic Web

## A Semi-Automated Approach

Andreas Heß, Christian Maaß and Francis Dierick

Lycos Europe

01/06/2008

# Outline

- Motivation
- Proposals for better tagging
- Tag suggestion / semi-automated tagging
- Tag merging
- Conclusion

# Motivation

- Ontologies: high entrance barriers for ordinary users
- Folksonomies: widely used, low entrance barriers
- Goals
  - Draw benefits from complementary nature
  - Improve quality of folksonomies
    - Annotations
    - Tag Cloud
  - Eventually merge folksonomies and ontologies

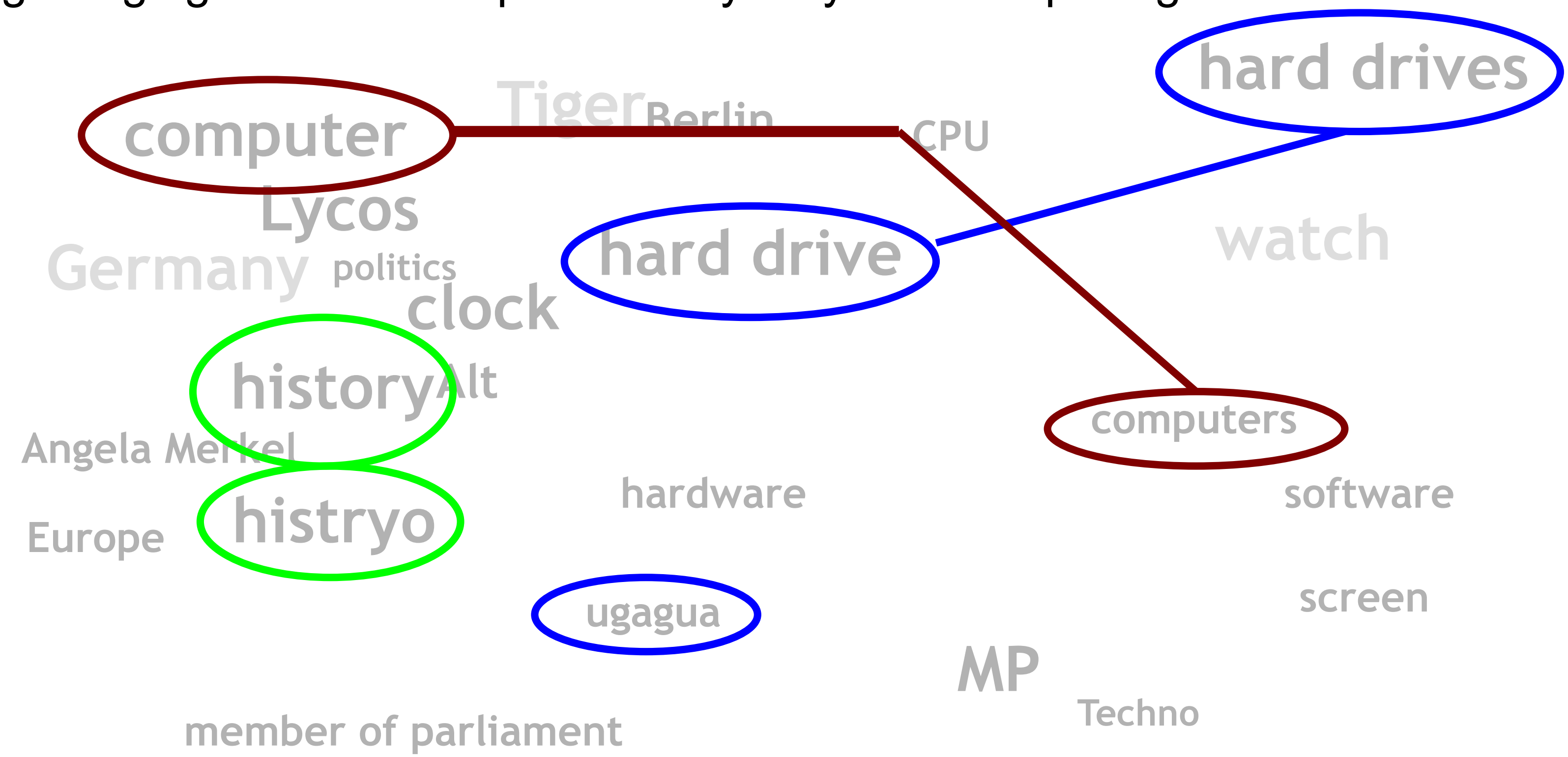
# Moving from Folksonomies to Ontologies: Tag Quality

Tag Merging: Eliminate duplicates / synonyms / misspellings / nonsense

hard drives
Tiger Berlin
computer
CPU
Lycos
Germany politics
hard drive
watch
clock
history Alt
computers
Angela Merkel
hardware
software
Europe
histryo
ugagua
screen
MP
Techno
member of parliament

# Moving from Folksonomies to Ontologies: Tag Quality

Tag Merging: Eliminate duplicates / synonyms / misspellings / nonsense



# Moving from Folksonomies to Ontologies: Tag Quality

Tag Merging: Eliminate duplicates / synonyms / misspellings / nonsense

hard drives
Tiger Berlin
CPU
Lycos
watch
Germany politics
clock
history Alt
computers
Angela Merkel
hardware
software
Europe
screen
MP
Techno
member of parliament

# Moving from Folksonomies to Ontologies: Tag Quality

## Topic Detection

Tiger Berlin

CPU

hard drives

Lycos

watch

Germany politics

clock

history Alt

computers

Angela Merkel

hardware

software

Europe

screen

MP

Techno

member of parliament

# Moving from Folksonomies to Ontologies: Tag Quality

## Topic Detection

hard drives

CPU

computers

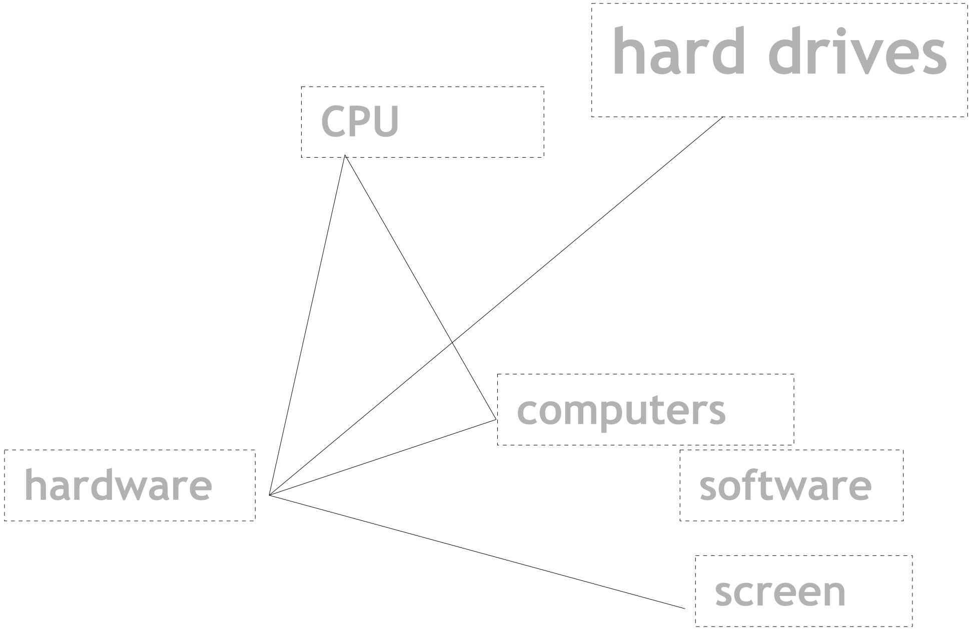
hardware

software

screen

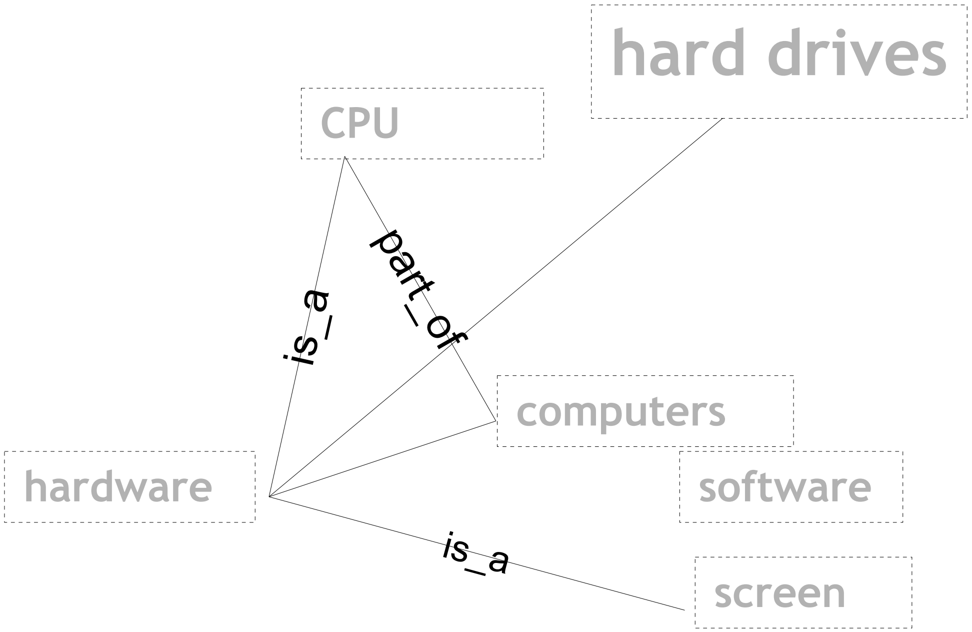
# Moving from Folksonomies to Ontologies: Tag Quality

Relation Extraction



# Moving from Folksonomies to Ontologies: Tag Quality

Relation Qualification



# Proposed Measures

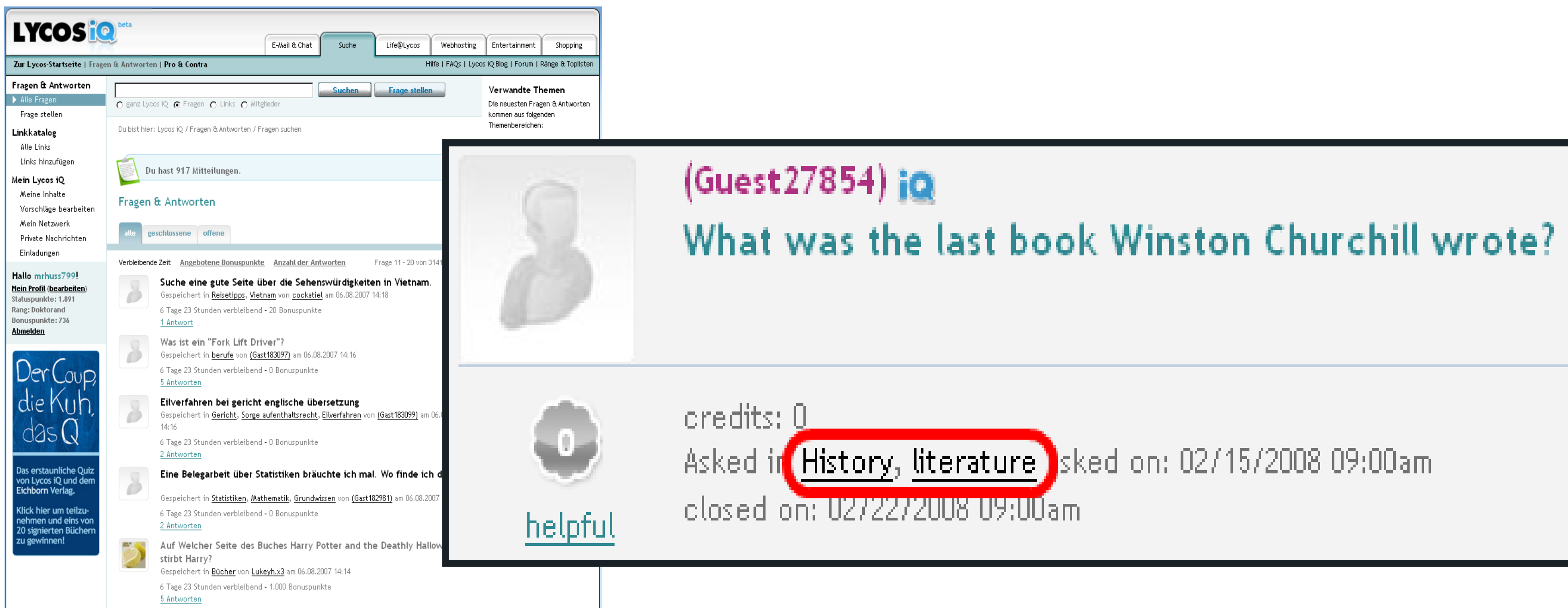
- Semi-Automated Tagging
  - Lower the threshold towards creating meta-data
- Tag Merging
  - Improving tag quality
- Extract Relations
  - First step on the move from folksonomies to more structured form
- User Rating
  - Involve user in refining quality
- Information Extraction
  - Automatically fill blanks

# Proposed Measures

- Semi-Automated Tagging
  - Lower the threshold towards creating meta-data
- Tag Merging
  - Improving tag quality
- Extract Relations
  - First step on the move from folksonomies to more structured form
- User Rating
  - Involve user in refining quality
- Information Extraction
  - Automatically fill blanks

# Semi-Automated Tagging

- Text classification, training data needed
- Semi-automated annotation of very short texts

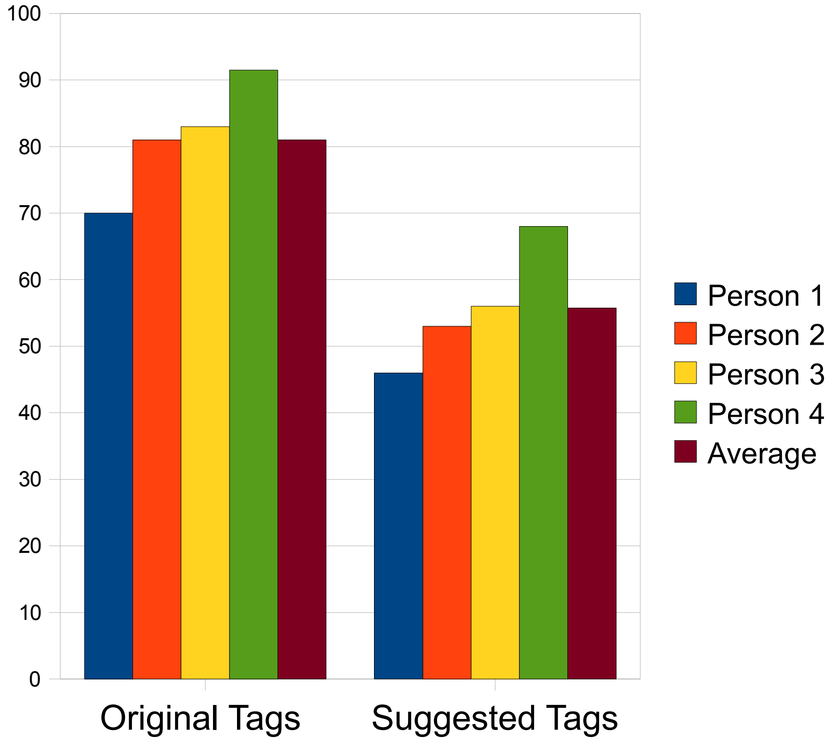


# Choice of Classification Algorithm

- Speed is important
  - Interactive: user does not want to wait
- Use well-known Rocchio text classification algorithm
  - Simple, fast, incremental, suitable for high number of classes
  - Works well only if texts are short and of similar length
  - ... but this is the case here
- Use part-of-speech-tagger for dimensionality reduction
  - Only nouns and proper nouns

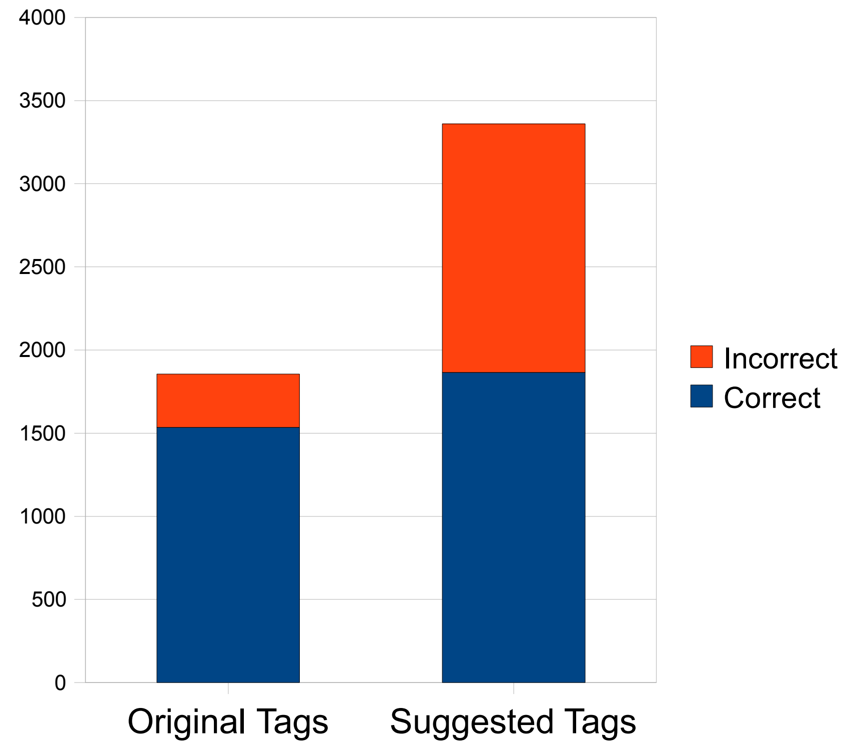
# Evaluation (I): Precision

- Tested precision with 4 test users
- Original tagging far from perfect
- Suggestion quality not great
- But good enough for interactive use
- In 87% at least one correct prediction within top 5



# Evaluation (II): absolute numbers

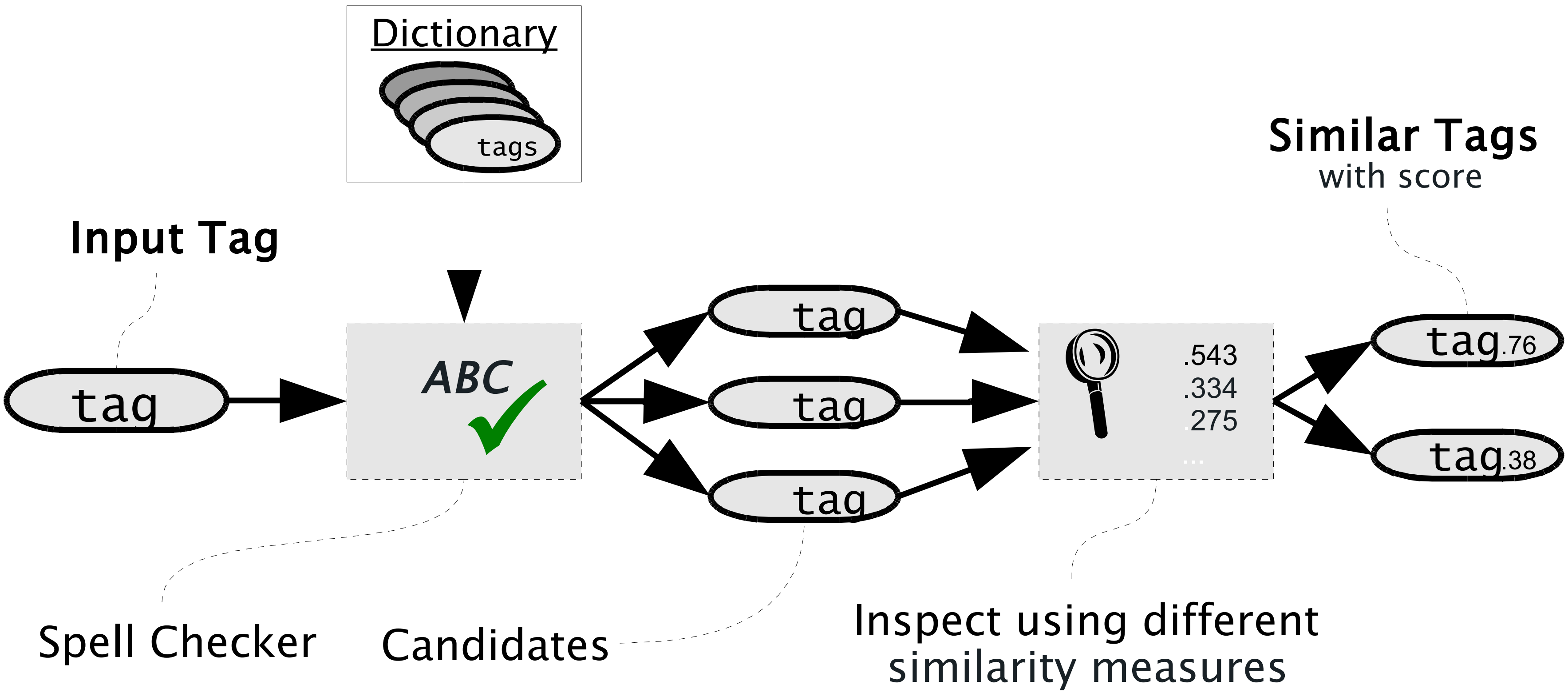
- More *correct* suggestions than original tags *in total*
- Assumption: People will tag more



# Tag Merging

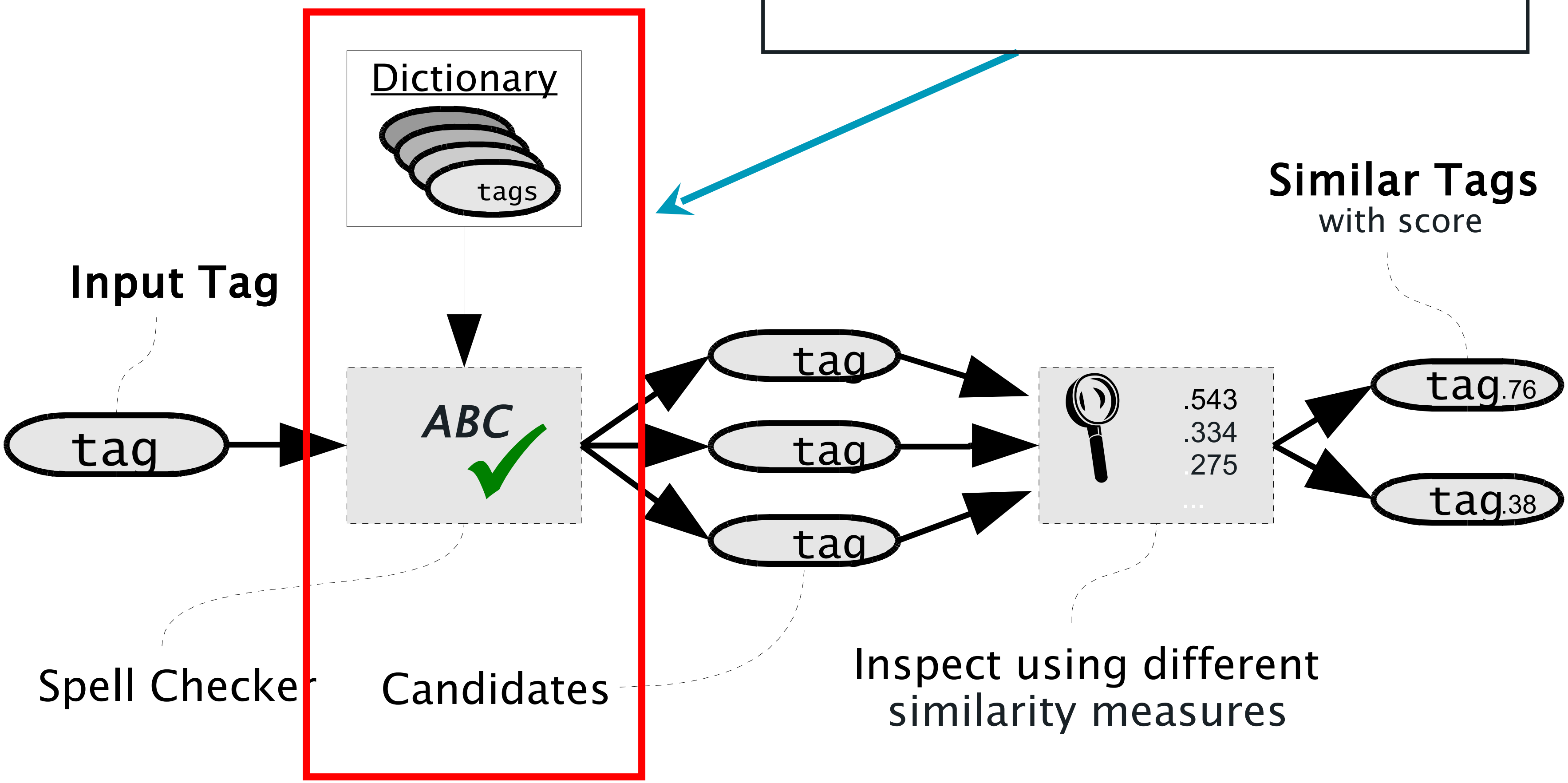
- Goals
  - Elimination and merging of incorrectly spelled tags
  - Merging of different spelling variations
- Example
  - „computer“ vs. „computers“ (singular/plural)

# Tag Merging - Algorithm



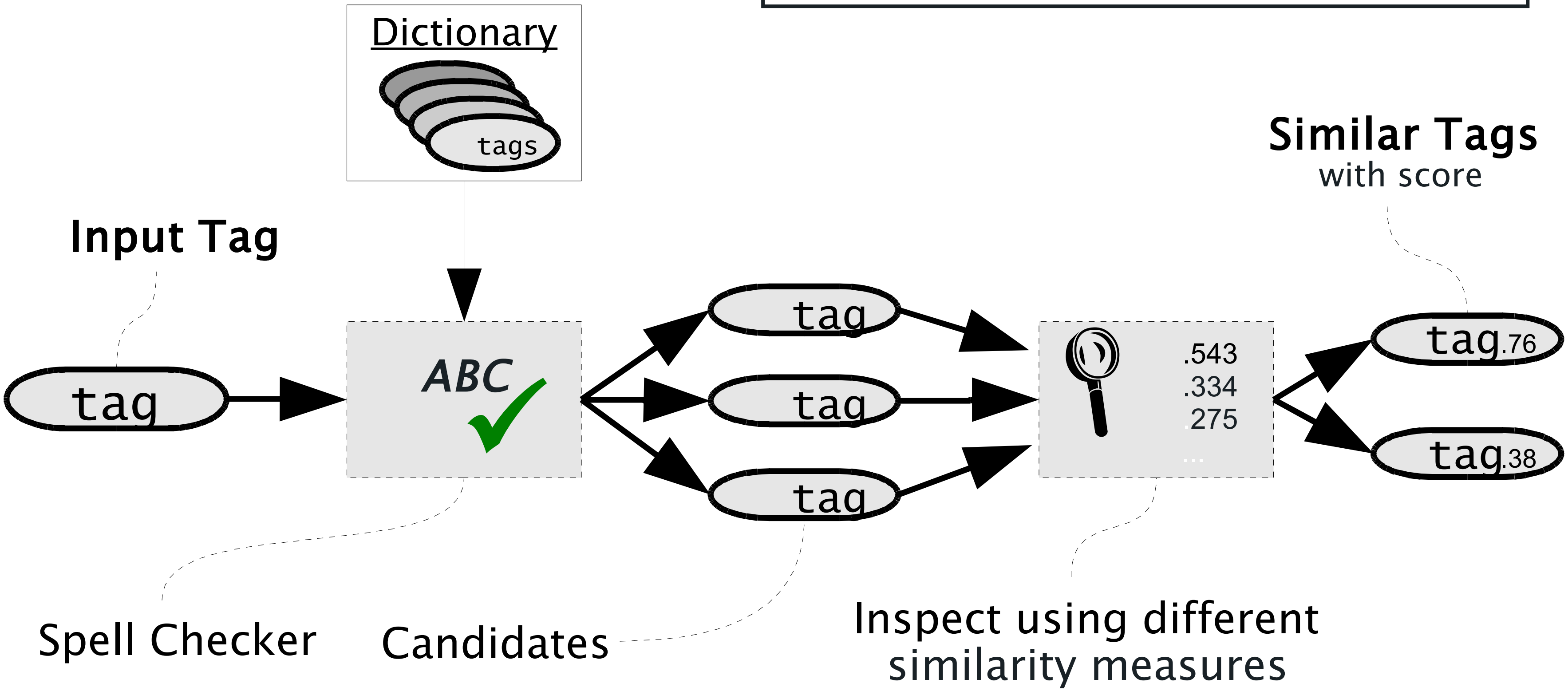
# Tag Merging - Algorithm

» **Why this extra step?**

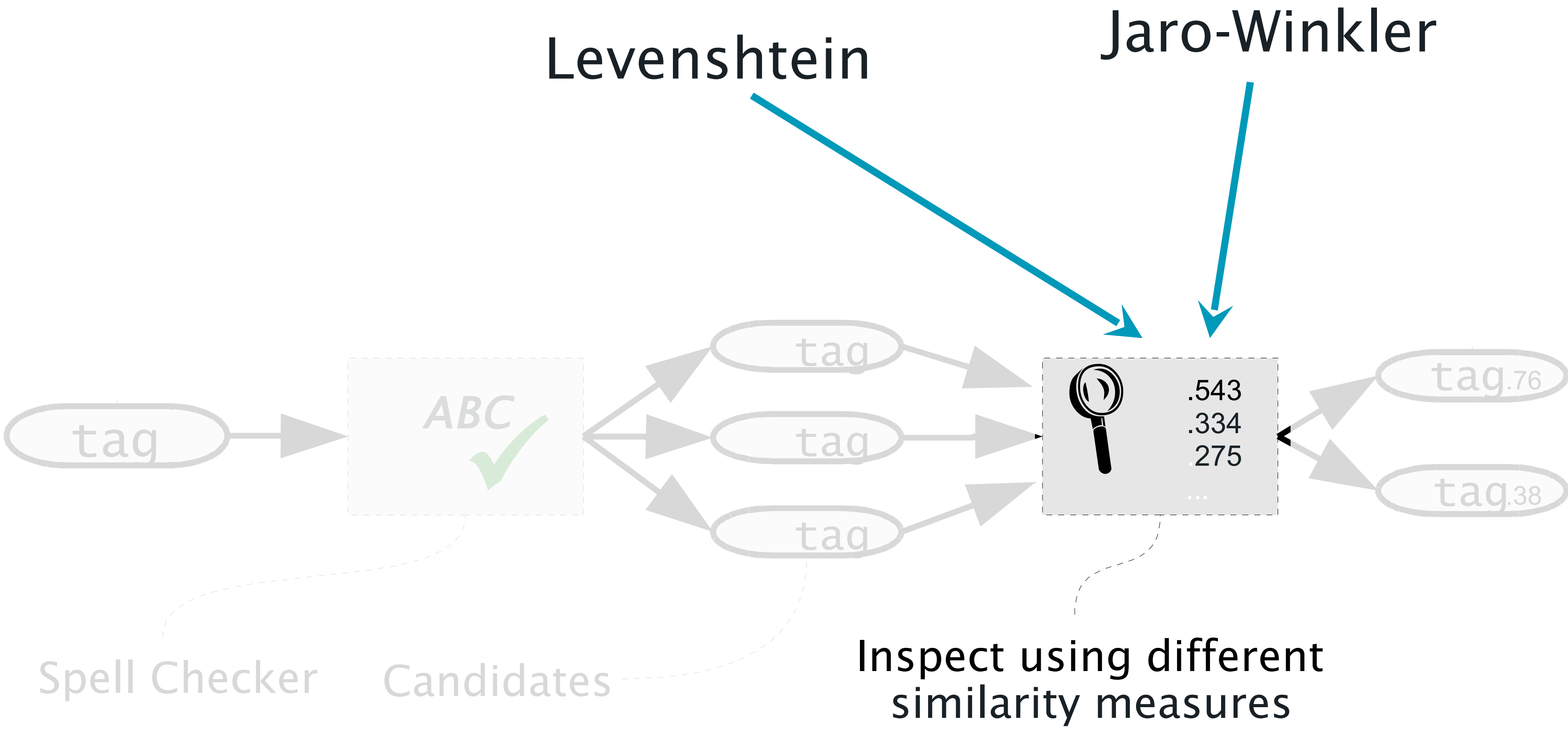


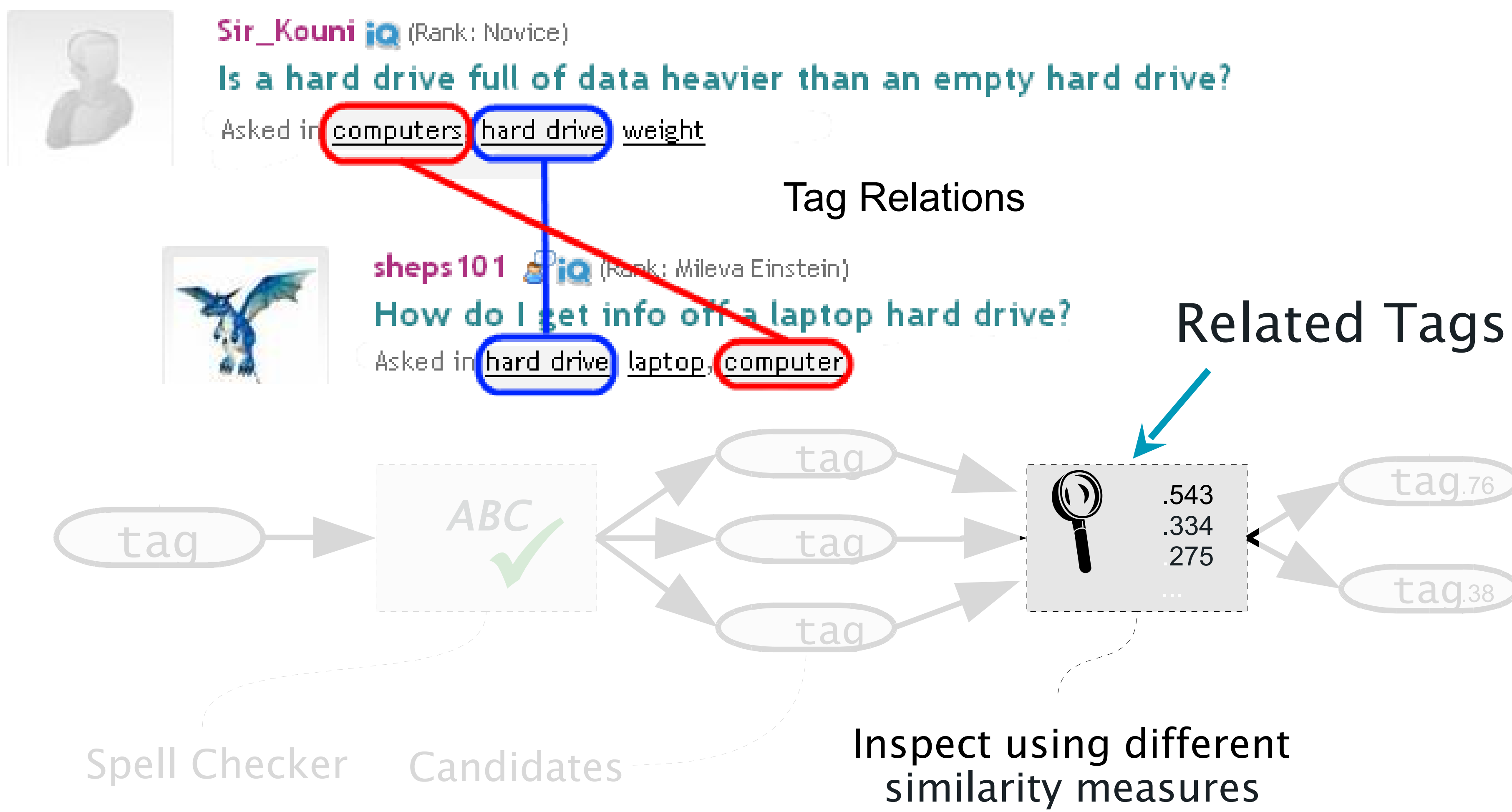
# Tag Merging - Algorithm

- Computing similarities is slow
- Pairwise checking is $\Theta(n^2)$

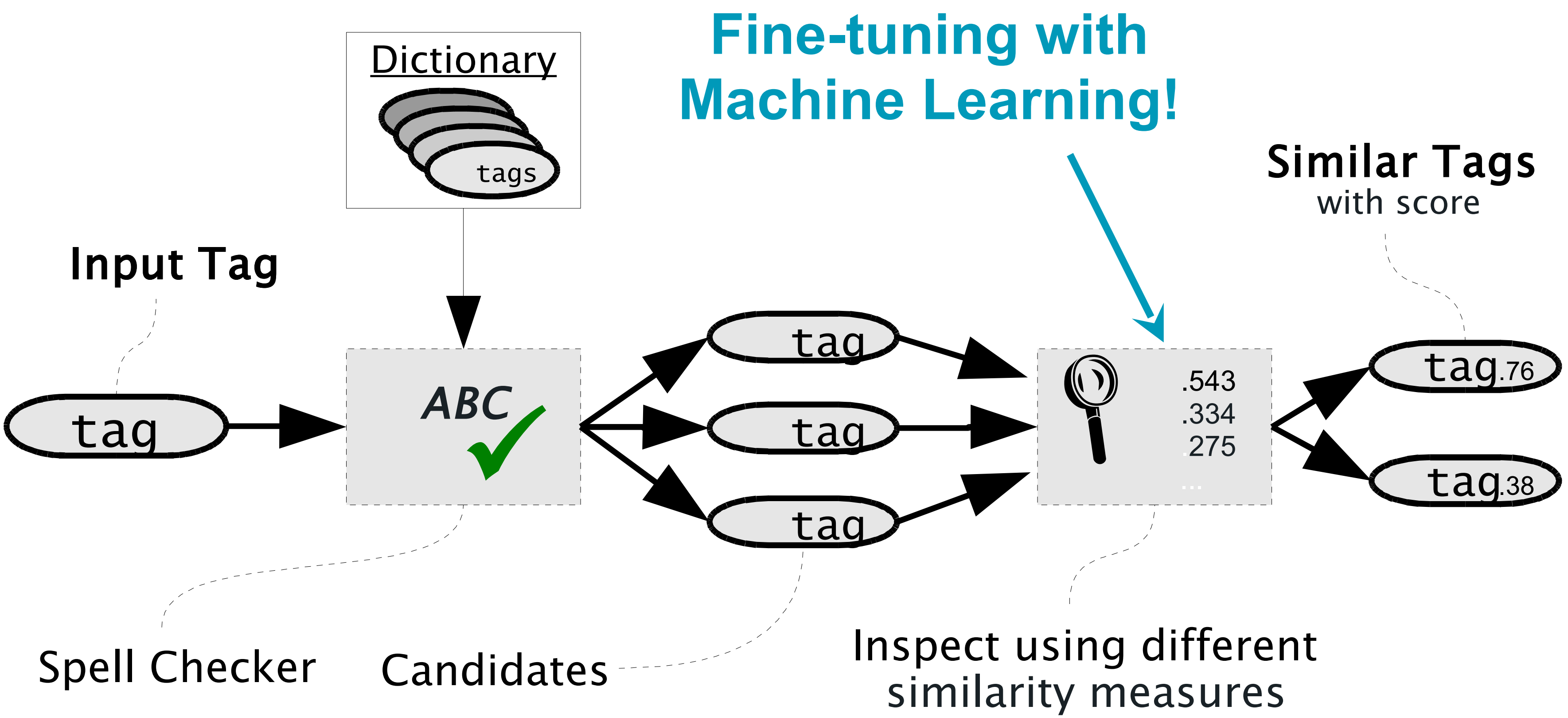


# Tag Merging - Algorithm




Sir_Kouni (Rank: Novice)
Is a hard drive full of data heavier than an empty hard drive?
Asked in computers hard drive weight
Tag Relations
sheps101 (Rank: Mileva Einstein)
How do I get info off a laptop hard drive?
Asked in hard drive laptop, computer
Related Tags
tag
ABC
tag
tag
tag
.543
.334
.275
...
tag.76
tag.38
Spell Checker
Candidates
Inspect using different similarity measures

# Tag Merging - Algorithm



# Tag Merging - Evaluation

- Can reach high precision by fine tuning with machine learning
- Trade-off between precision and recall tunable
- Precision in sample (100 tags): **95%**
- Fully automated batch processing possible
- With this setting **12%** smaller tag cloud

# Conclusion

» Proposed ways to combine strengths of folksonomies and ontologies

» Semi-automated Tagging and ...

» Tag Merging to increase folksonomy quality

» Outlined plan for future work

# Thank You for Your Attention!

» Questions?

# Tag Suggestions - Algorithm

» Rocchio with dimensionality reduction