UNIVERSITY OF TWENTE.

SEARCH SPECIALIZATION & SEARCH DELEGATION

CLEF 2016, 7 September 2016

Djoerd Hiemstra

http://www.cs.utwente.nl/~hiemstra/

HOW DO MACHINES DETERMINE RELEVANCE ... ?

STATISTICS !

IT DEPENDS...

- Ad hoc: Language models, BM25
- Spam filter: Naive Bayes
- Network data: PageRank
- Clicks: Learning to rank
- ...

“It Depends”

-Socrates

THERE IS NO “ONE SIZE FITS ALL”!

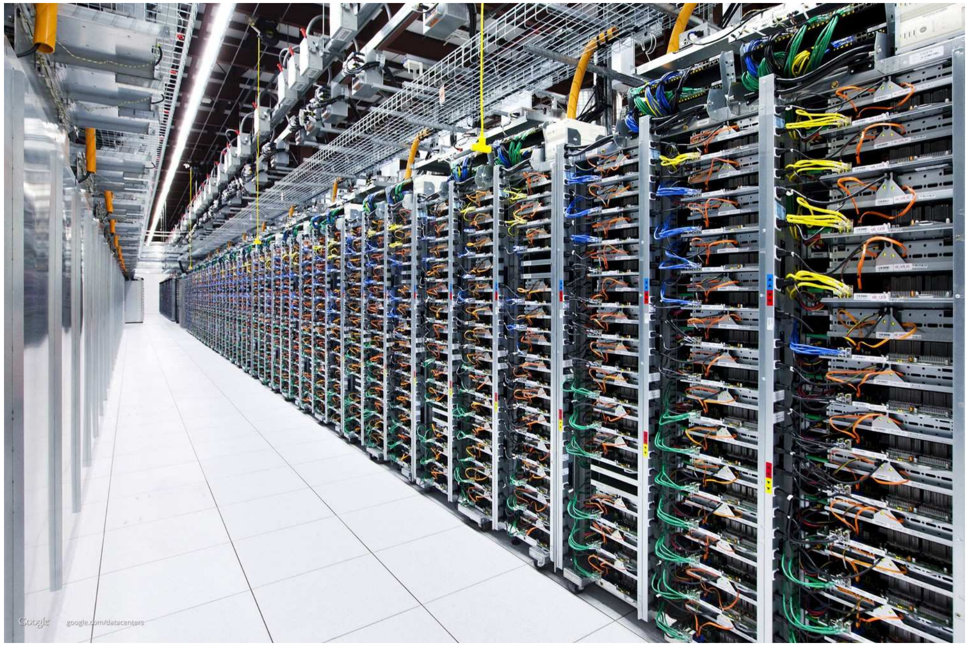
- Web: Web graph, anchor text
- Videos: Views, likes, content-based features
- Advertisements: Bids, click-through-rate
- Tweets: Retweets, likes,
- Scientific papers: Citations
- Restaurants: Geo, reviews
- Products: Price, nr. in stock
- ...

WE NEED SEARCH SPECIALIZATION!

- CLEF eHealth
- ImageCLEF
- LifeCLEF
- Uncovering Plagiarism
- Social Book Search
- News Recommendation
- Living Labs (products, papers)
- …

BIG DATA FALLACY?

- If we have **all data**, we can learn **one** model that...



… participates in every CLEF lab ?

SEARCH SPECIALIZATION & DELEGATION !

UNIVERSITY OF TWENTE.

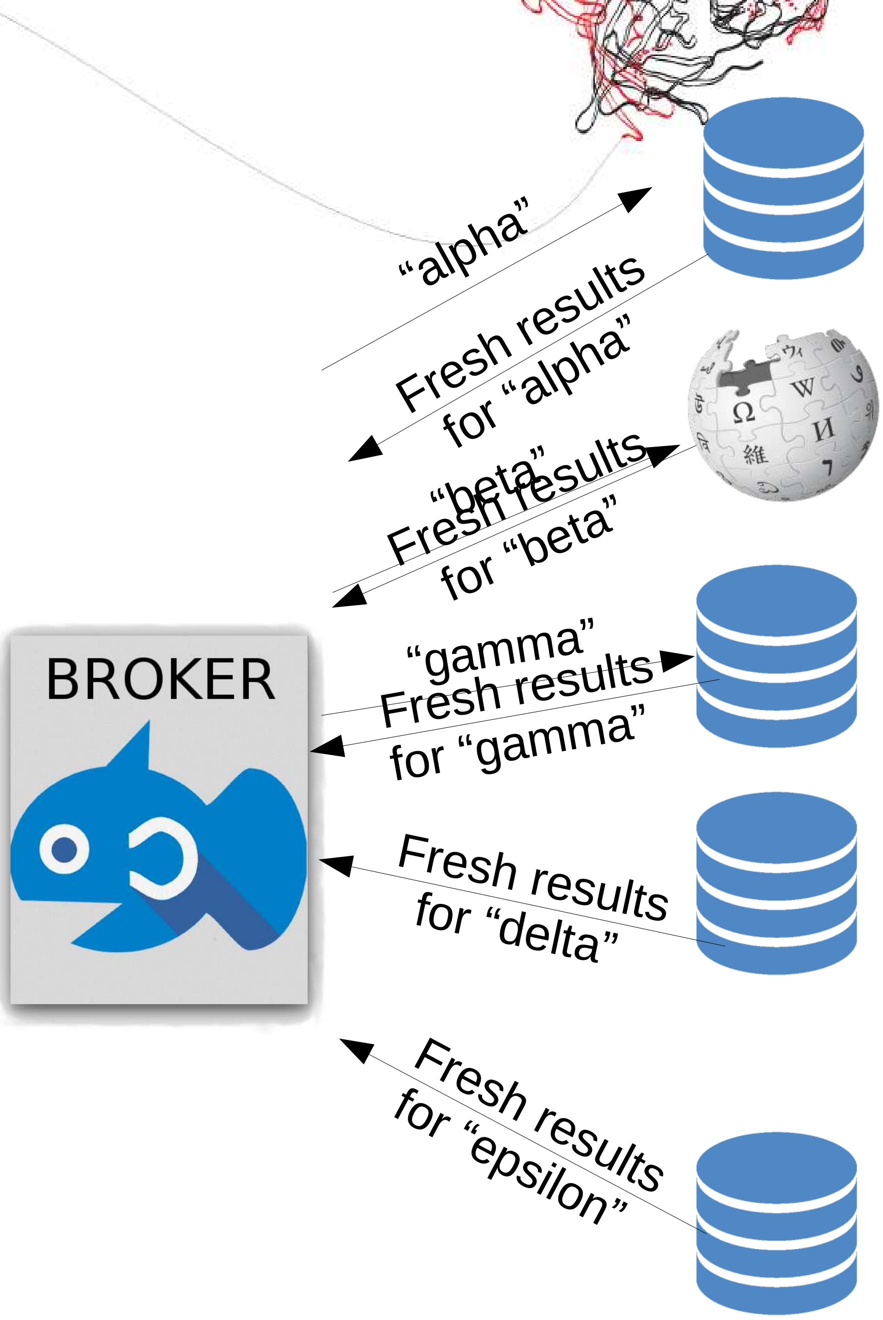
http://search.utwente.nl

QUERY-BASED SAMPLING

- *“Query-based sampling is a (…) method of acquiring resource descriptions that does not require explicit cooperation from resource providers. Instead, resource descriptions are created by running queries and examining the documents that are returned.”*

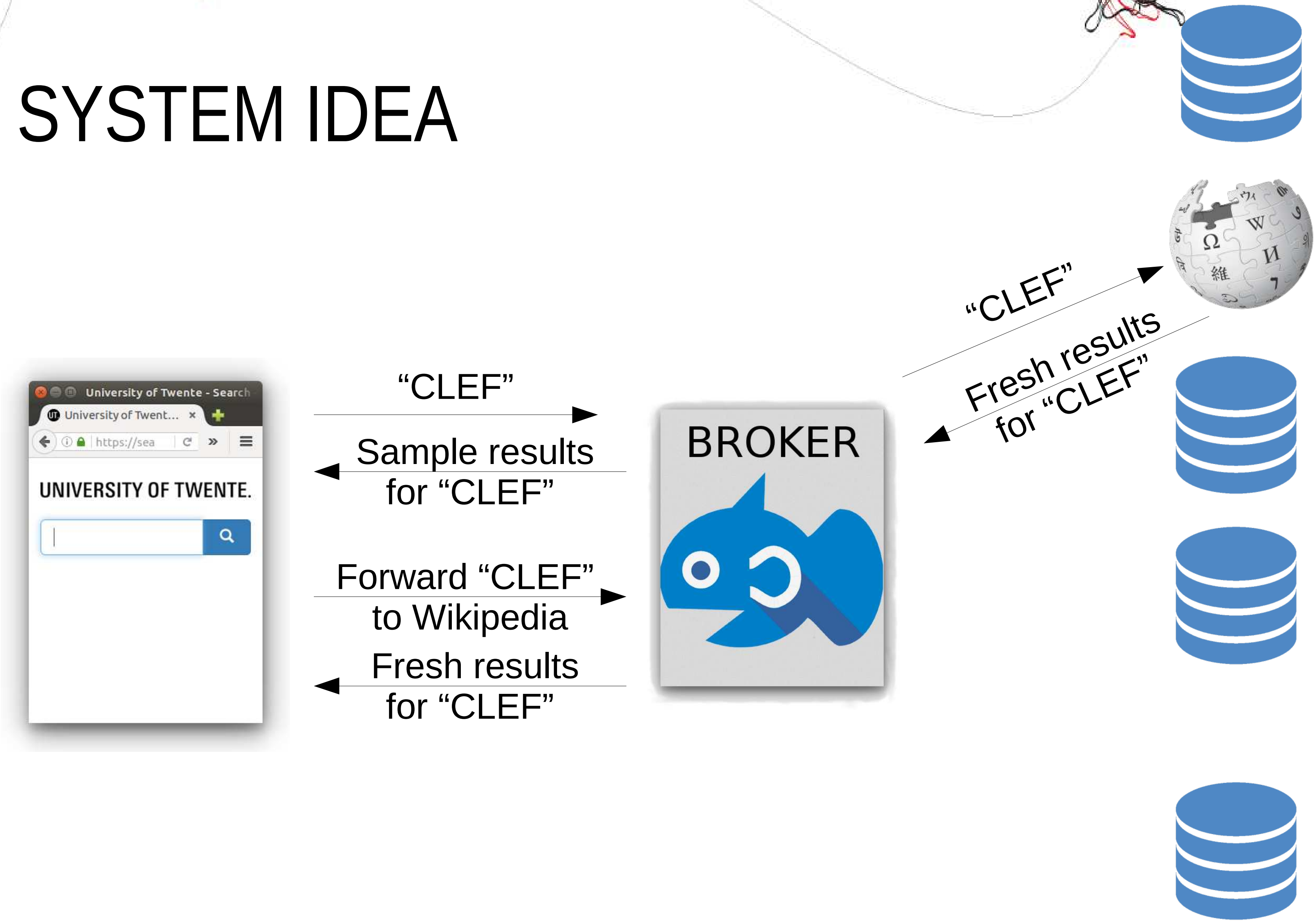
- Jamie Callan and Margaret Connell.
  Query-Based Sampling of Text Databases.
  ACM Transactions on Information Systems 19(2), 2001

SYSTEM IDEA



“alpha”
Fresh results for “alpha”
“beta”
Fresh results for “beta”
BROKER
“gamma”
Fresh results for “gamma”
Fresh results for “delta”
Fresh results for “epsilon”

SYSTEM IDEA



“CLEF”

Sample results for “CLEF”

Forward “CLEF” to Wikipedia

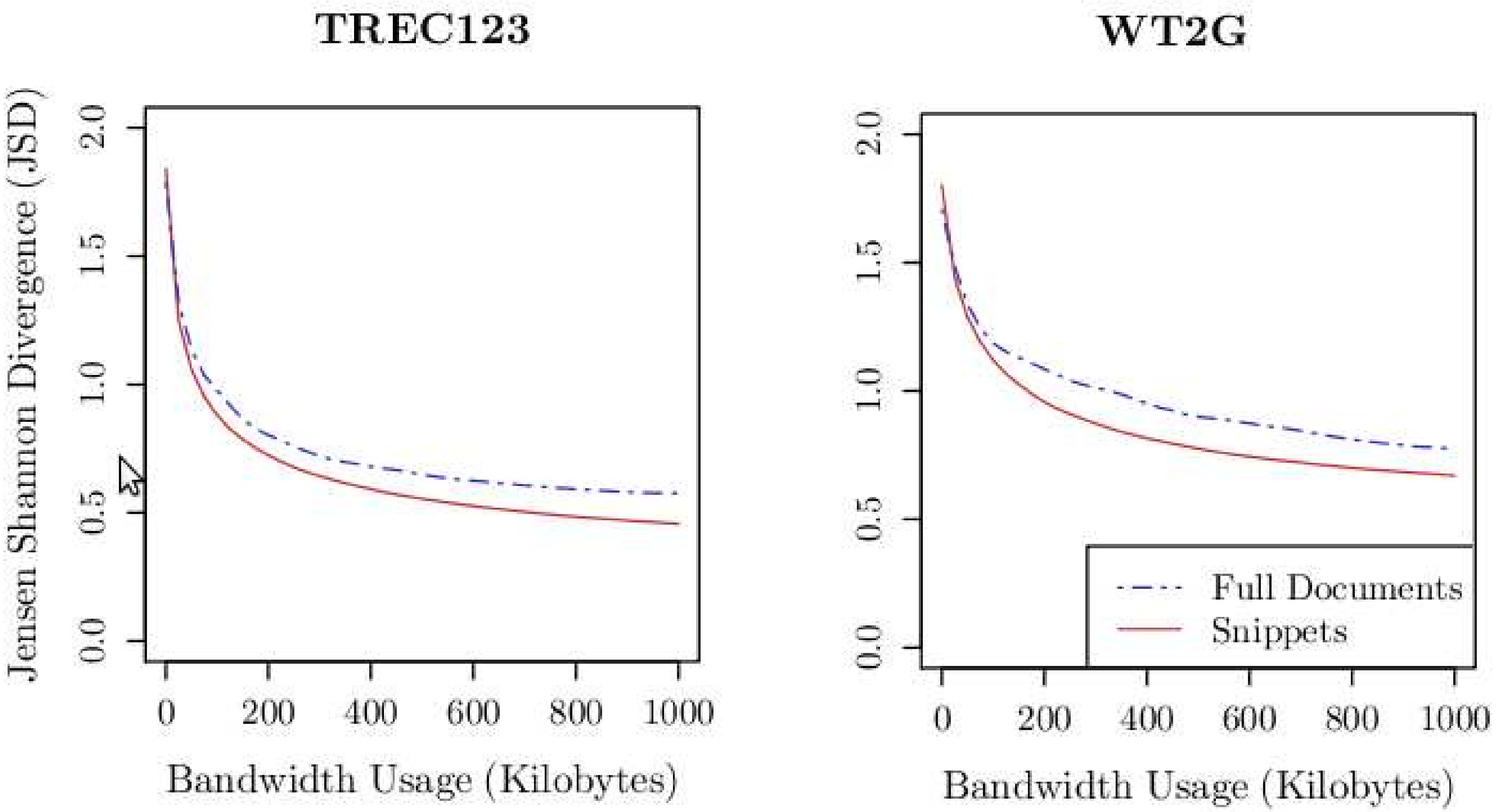
Fresh results for “CLEF”

BROKER

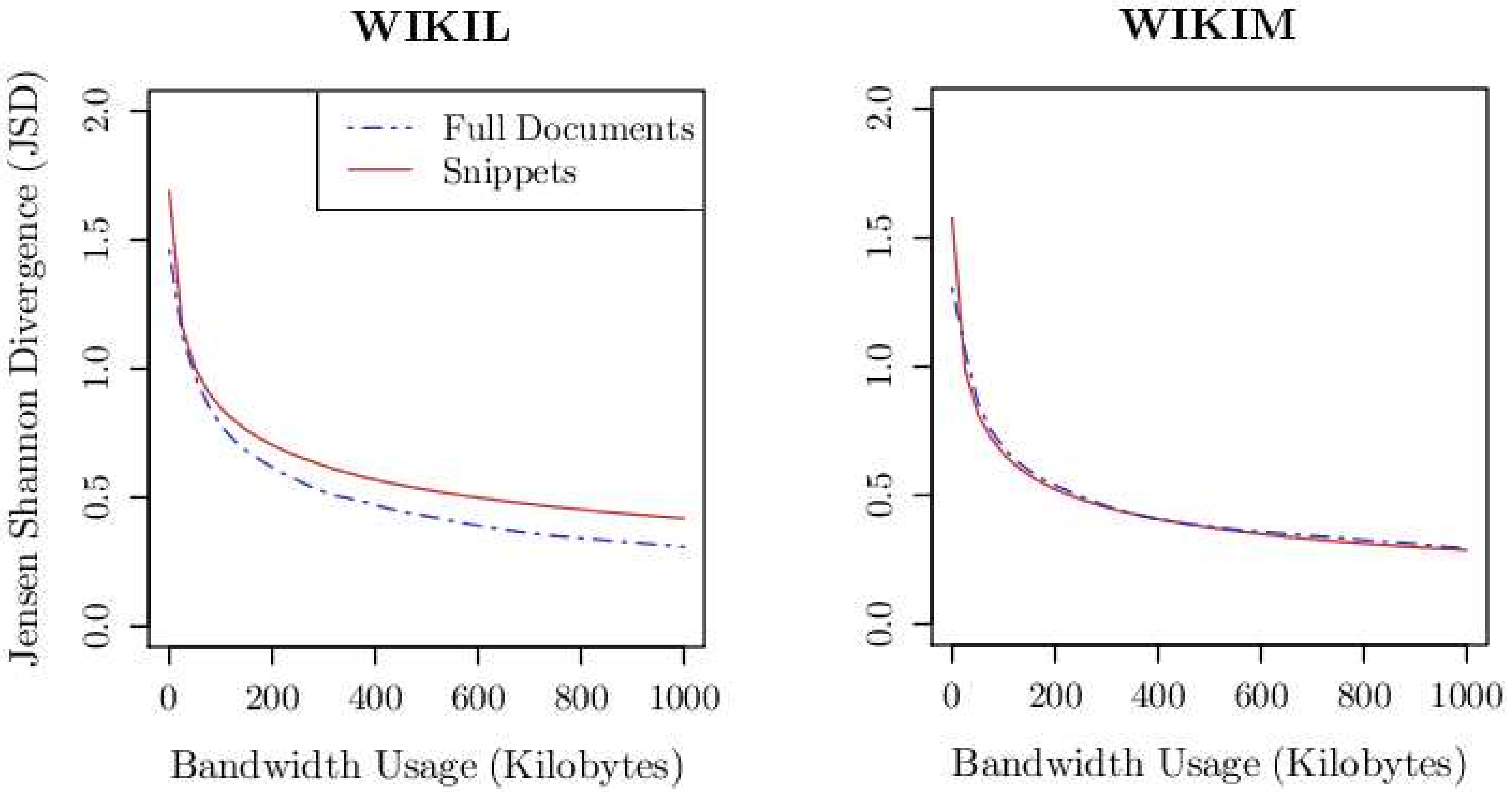
“CLEF”

Fresh results for “CLEF”

DO SAMPLES RESEMBLE THE FULL INDEX?



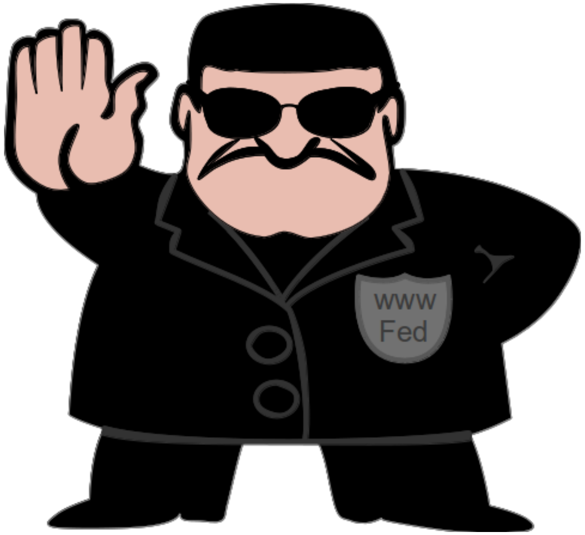
DO SAMPLES RESEMBLE THE FULL INDEX?



Almer Tigelaar and Djoerd Hiemstra, “Query-Based Sampling using Snippets'', In Proceedings of the SIGIR Workshop on Large-Scale Distributed Systems for Information Retrieval, 2010.

TREC “FEDWEB” TRACK

- Large-scale federated search evaluation:
  - □ Resource selection
  - □ Result merging
  - □ Vertical selection



RESOURCES

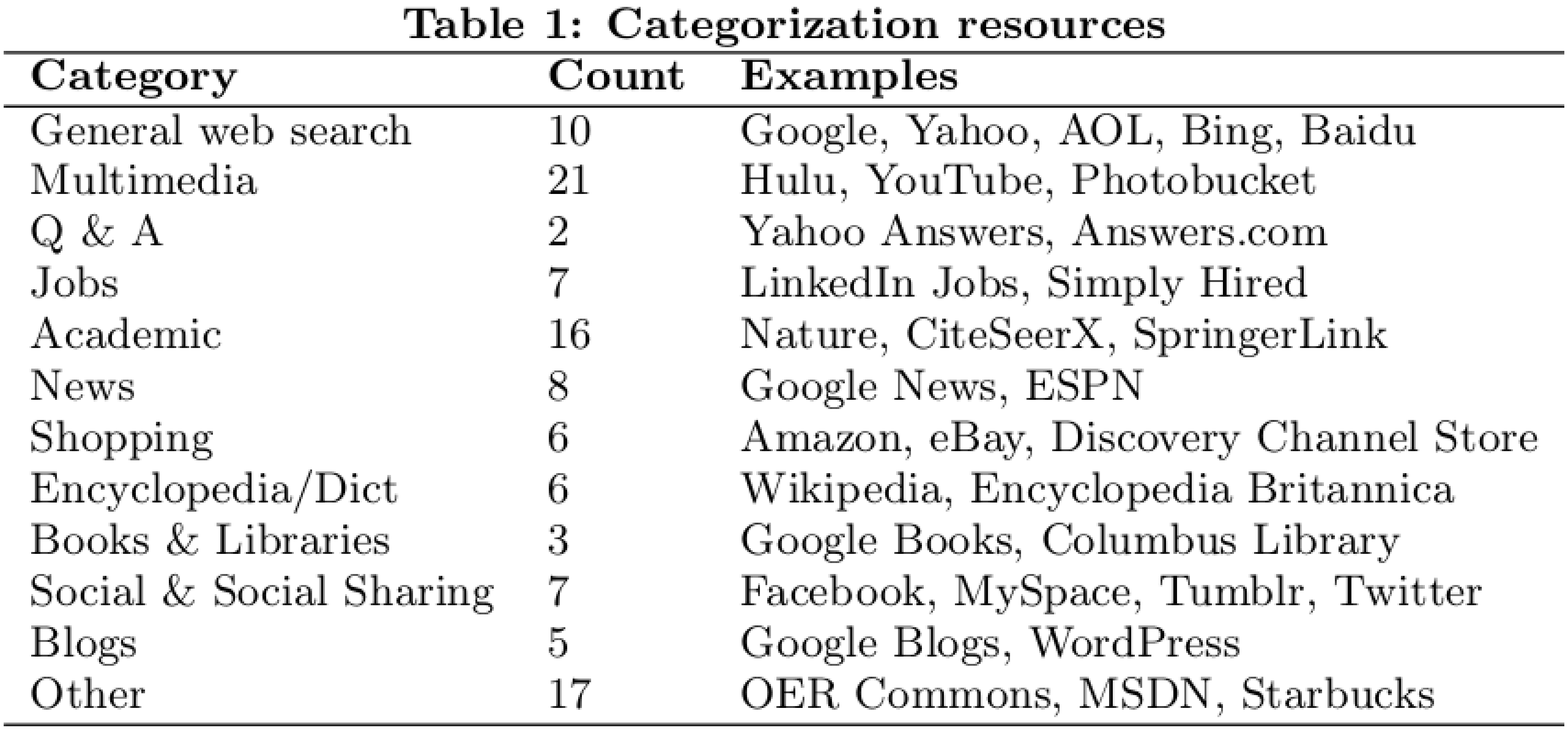
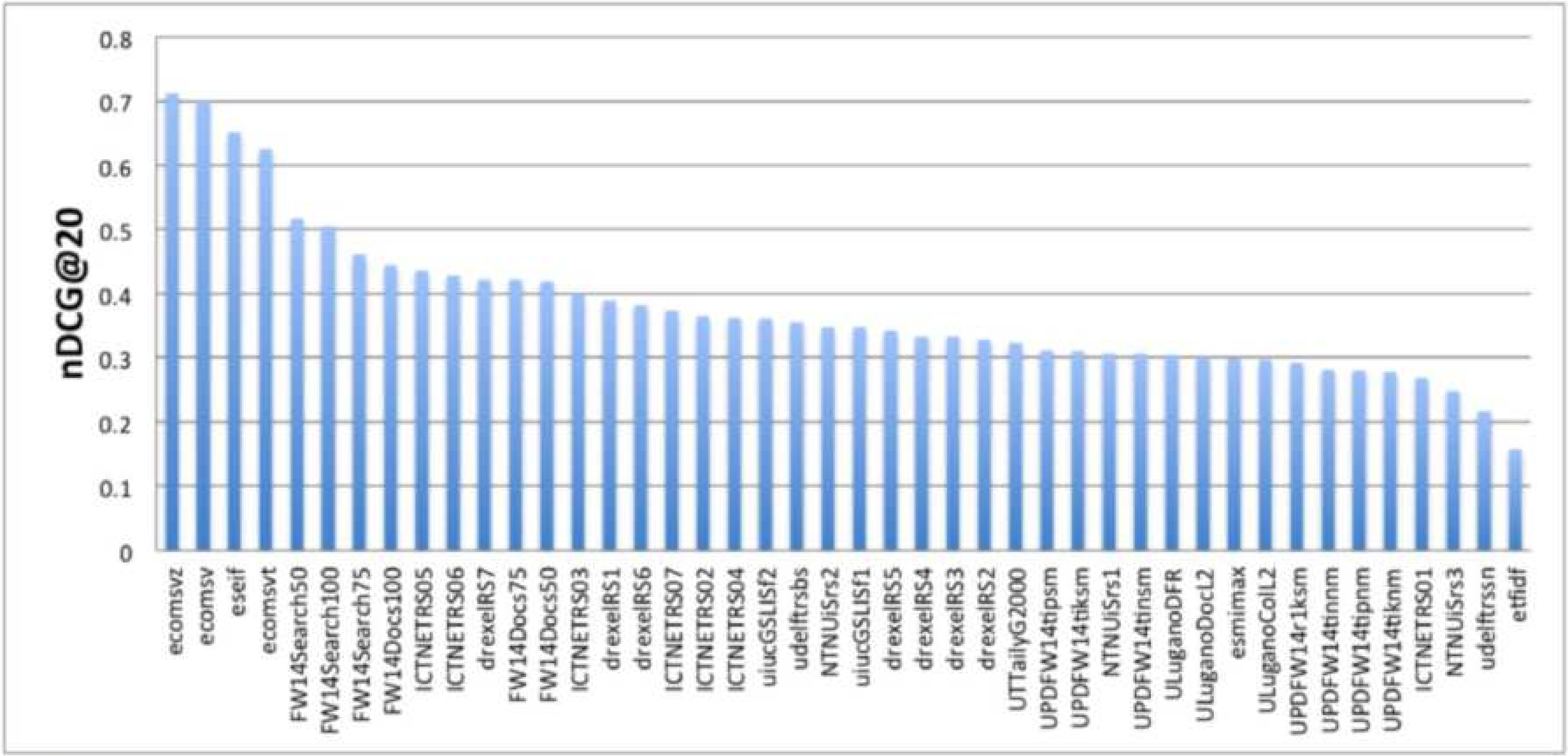


Table 1: Categorization resources

Category	Count	Examples
General web search	10	Google, Yahoo, AOL, Bing, Baidu
Multimedia	21	Hulu, YouTube, Photobucket
Q & A	2	Yahoo Answers, Answers.com
Jobs	7	LinkedIn Jobs, Simply Hired
Academic	16	Nature, CiteSeerX, SpringerLink
News	8	Google News, ESPN
Shopping	6	Amazon, eBay, Discovery Channel Store
Encyclopedia/Dict	6	Wikipedia, Encyclopedia Britannica
Books & Libraries	3	Google Books, Columbus Library
Social & Social Sharing	7	Facebook, MySpace, Tumblr, Twitter
Blogs	5	Google Blogs, WordPress
Other	17	OER Commons, MSDN, Starbucks

Resource Selection Results (44 runs)

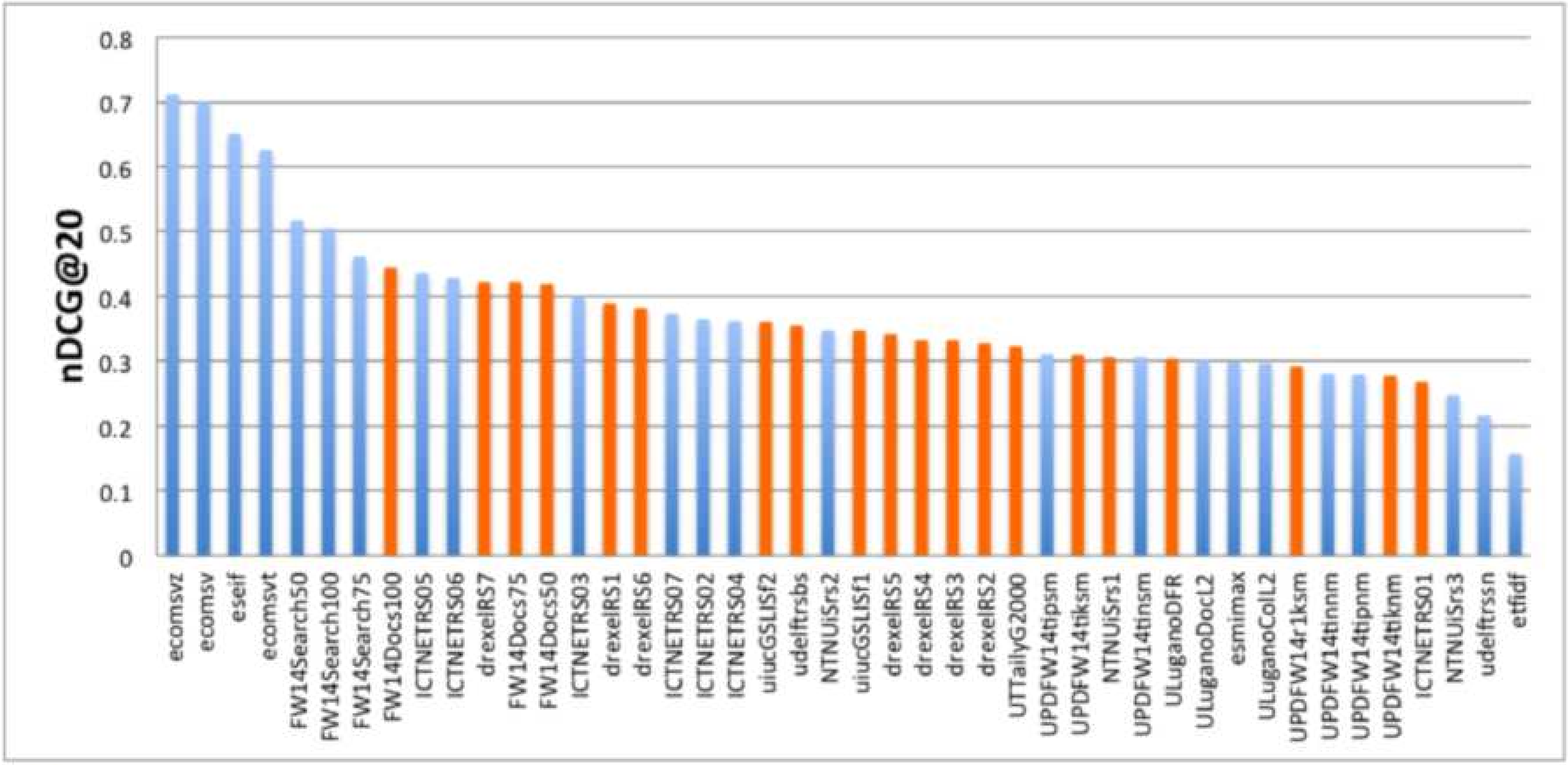


UNIVERSITEIT GENT

UNIVERSITY OF TWENTE.

YAHOO! LABS

Resource Selection Results: use of sampled pages only

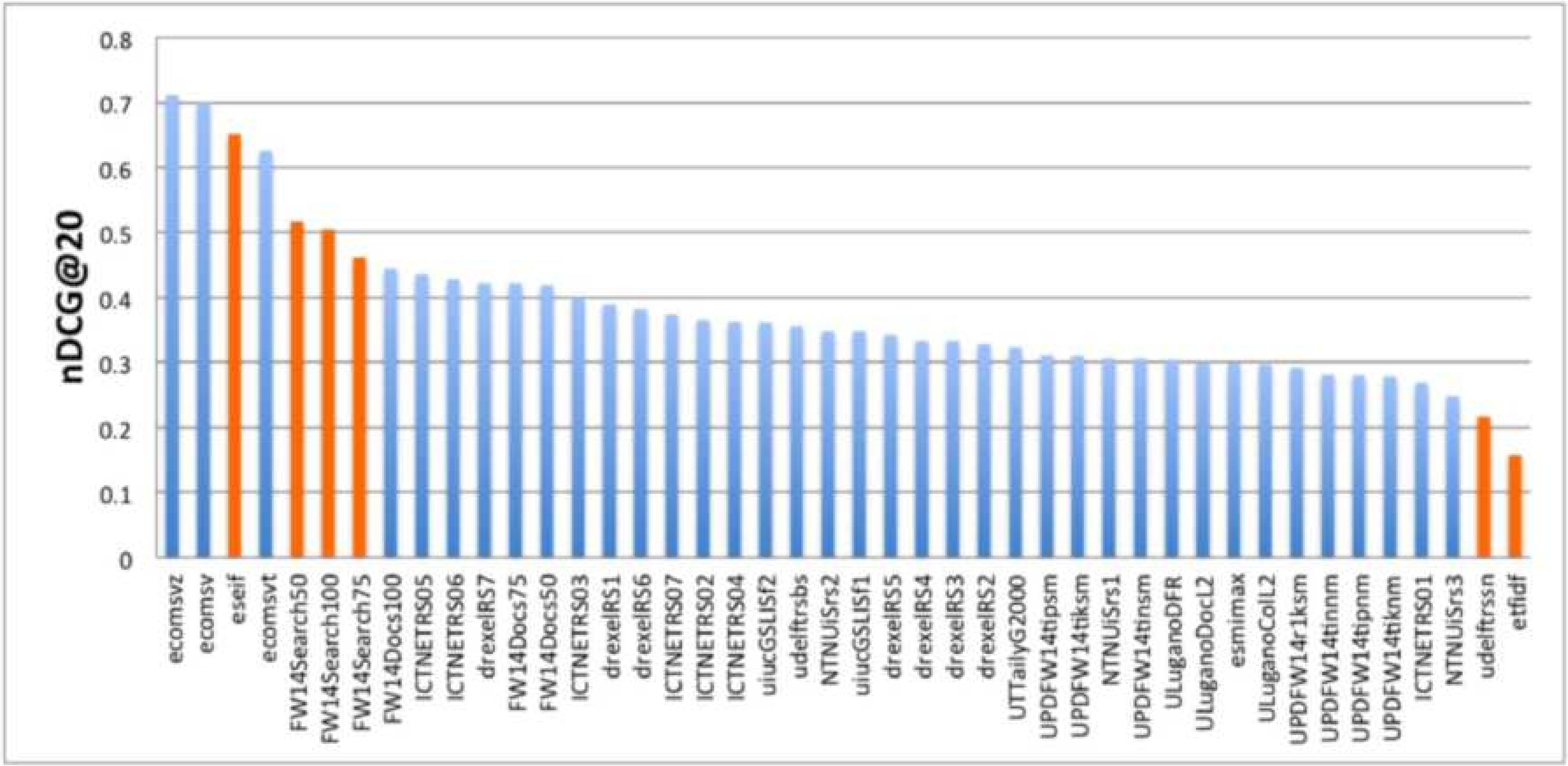


UNIVERSITEIT GENT

UNIVERSITY OF TWENTE.

YAHOO! LABS

Resource Selection Results: use of sampled snippets only

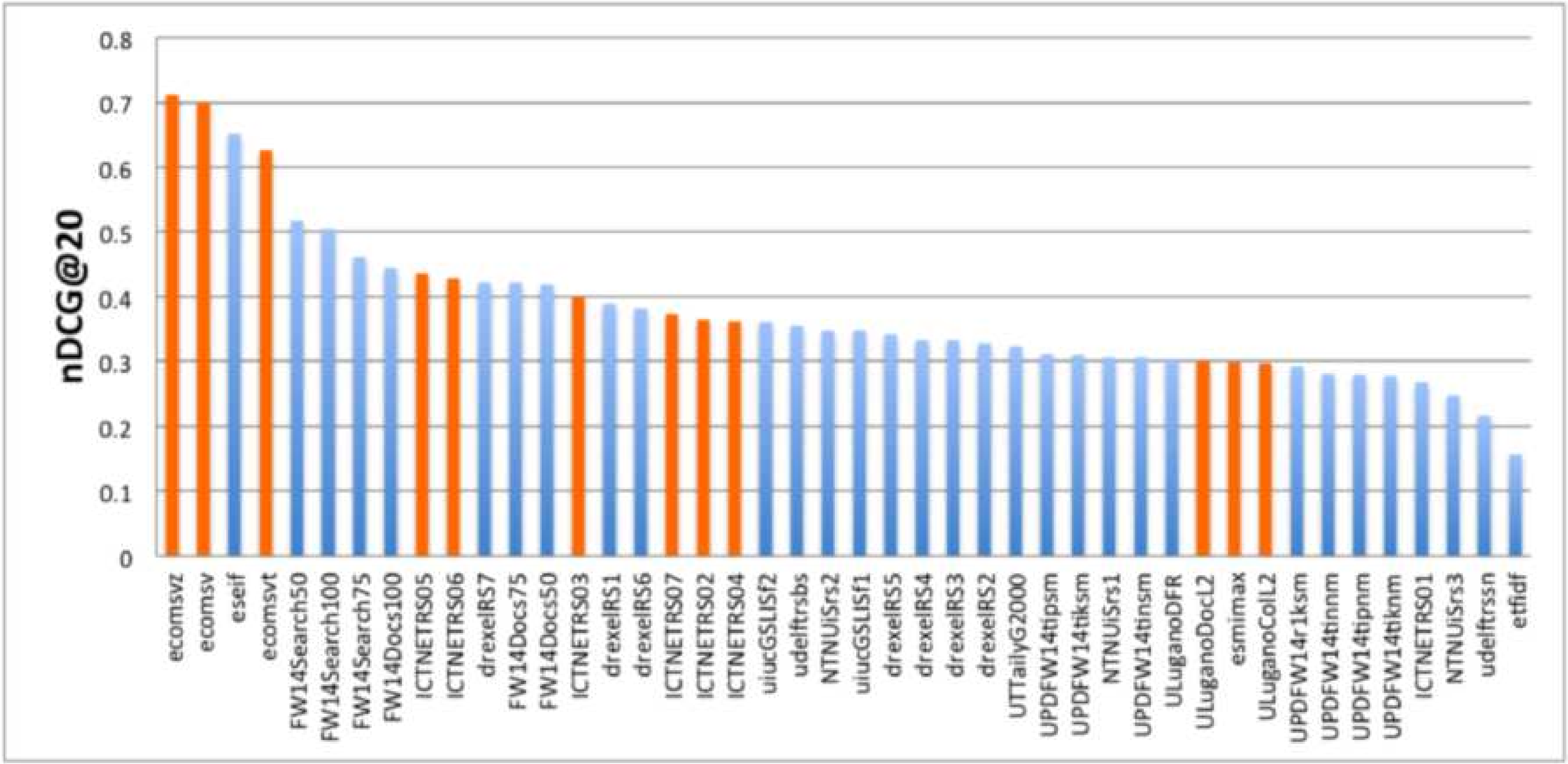


UNIVERSITEIT GENT

UNIVERSITY OF TWENTE.

YAHOO! LABS

Resource Selection Results: use of external resources



UNIVERSITEIT GENT

UNIVERSITY OF TWENTE.

YAHOO! LABS

QUERY-BASED SAMPLING: DISCUSSION

1. Sampling snippets is as effective as sampling full documents
2. Can be done at no extra costs(!)

FEDWEB GREATEST HITS

https://fedwebgh.intec.ugent.be

- A citable (static) dataset!
  - Thomas Demeester, Dolf Trieschnigg, Ke Zhou, Dong Nguyen, Djoerd. Hiemstra. FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search. In *WWW* 2015.

FEDWEB GREATEST HITS

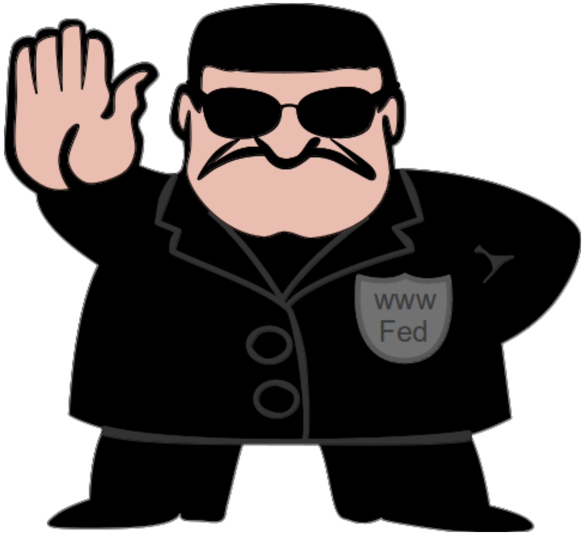
https://fedwebgh.intec.ugent.be

- >50 topics (queries) per year (2013, 2014)
- Result pages for 150 resources
- Click-through for snippets
- Relevance judgments for pages
- Duplicates detected
- Results and pages for sample queries
- Additional (duplicate) judgments

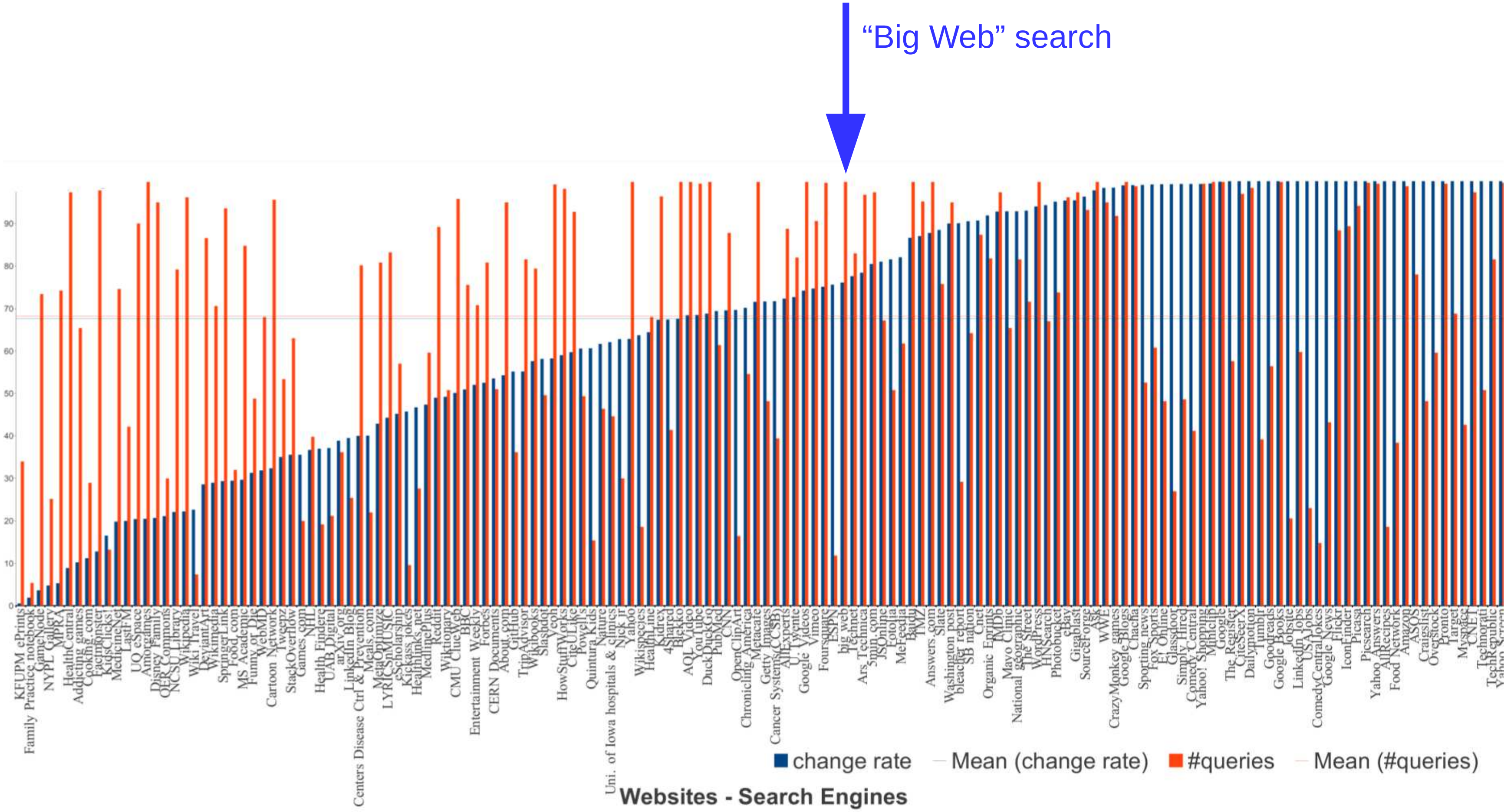
INVENT YOUR OWN RESEARCH

https://fedwebgh.intec.ugent.be

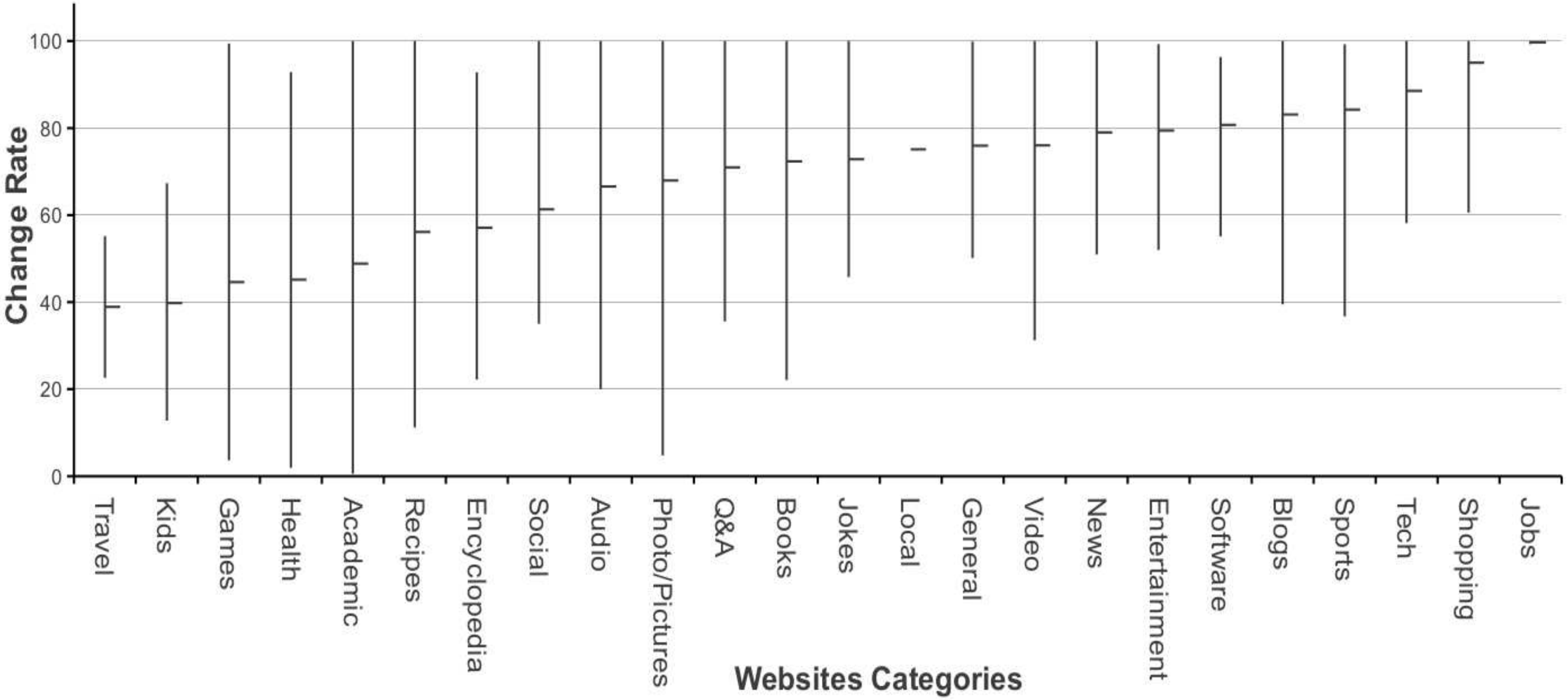
- Monitoring: What changed in 1 year?
- Clicks vs. Page relevance
- Web search without web search engines
- Size estimation
- …



HOW MUCH CHANGED IN 1 YEAR?


“Big Web” search

CHANGE RATE PER CATEGORY



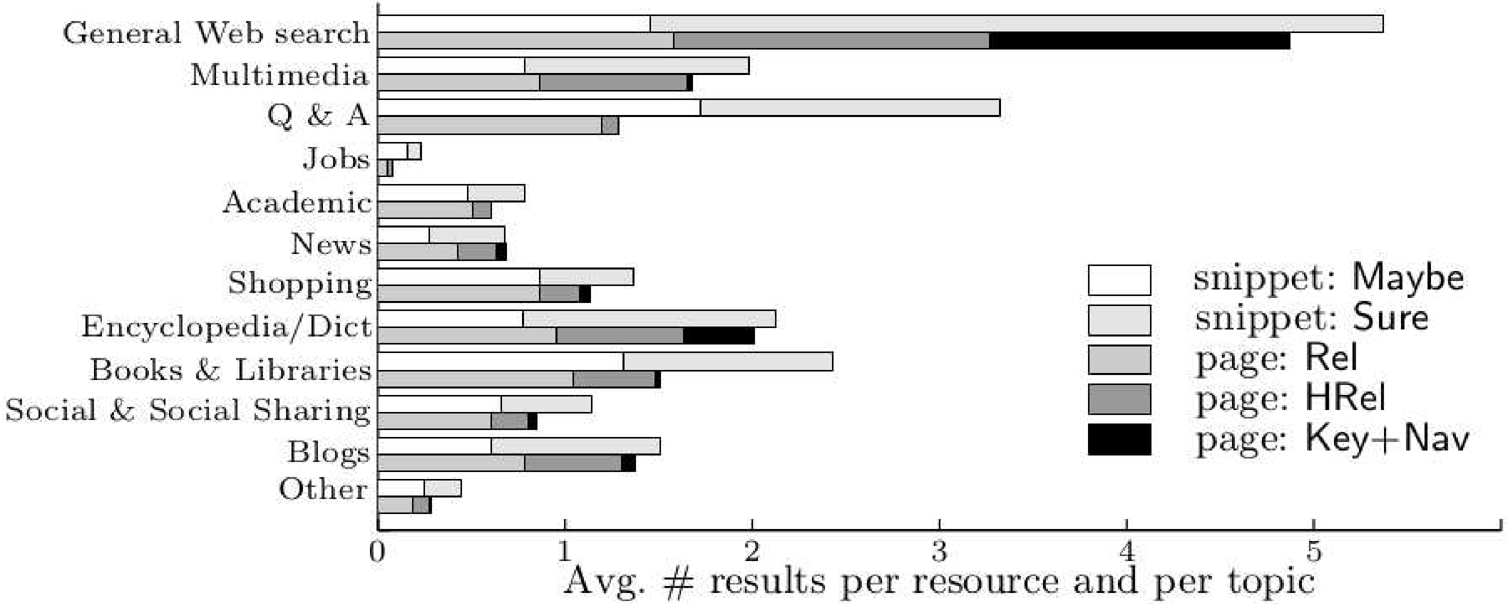
UNIVERSITY OF TWENTE.

DISCUSSION (1)

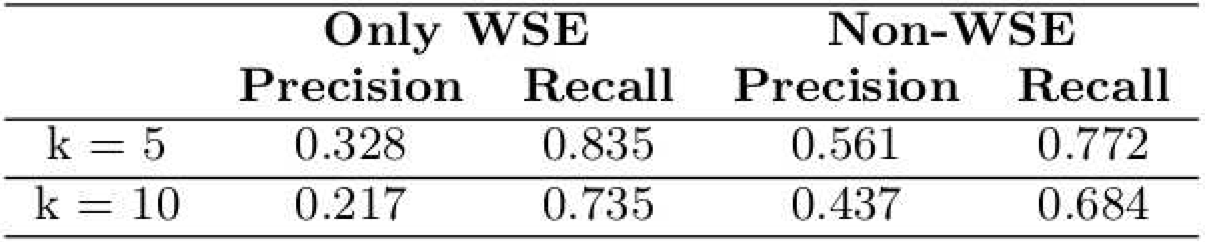
- Many things change in 1 year
  - □ Big differences per resource (Jobs vs. travel)
- Challenge:
  - □ Remove outdated results from sample index
  - □ Learn change rate!

Mohammadreza Khelghati, Djoerd Hiemstra and Maurice van Keulen. Efficient Web Harvesting Strategies for Monitoring Deep Web Content. (Submitted for publication)

RELEVANT RESULTS?



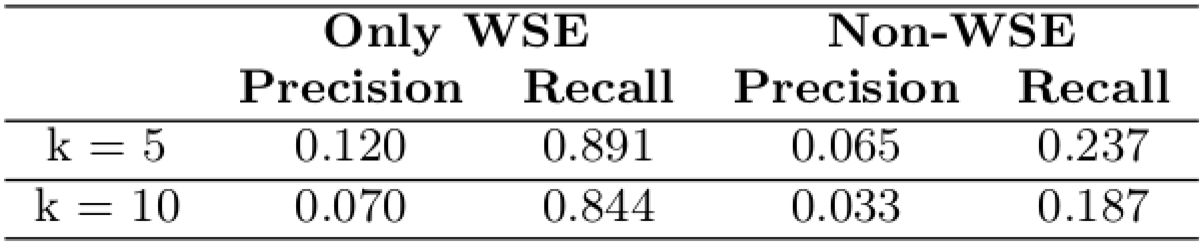
CAN WE DO WITHOUT LARGE SEARCH ENGINES?



	Only WSE		Non-WSE	
	Precision	Recall	Precision	Recall
k = 5	0.328	0.835	0.561	0.772
k = 10	0.217	0.735	0.437	0.684

Table 4: Oracle experiment - Rel or better

CAN WE DO WITHOUT LARGE SEARCH ENGINES?



	Only WSE		Non-WSE	
	Precision	Recall	Precision	Recall
k = 5	0.120	0.891	0.065	0.237
k = 10	0.070	0.844	0.033	0.187

Table 5: Oracle experiment - Better than HRel

DISCUSSION (2)

- Relevant results in all resource categories
- General web search engines needed for top results

DO CLICKS IMPLY RELEVANCE?

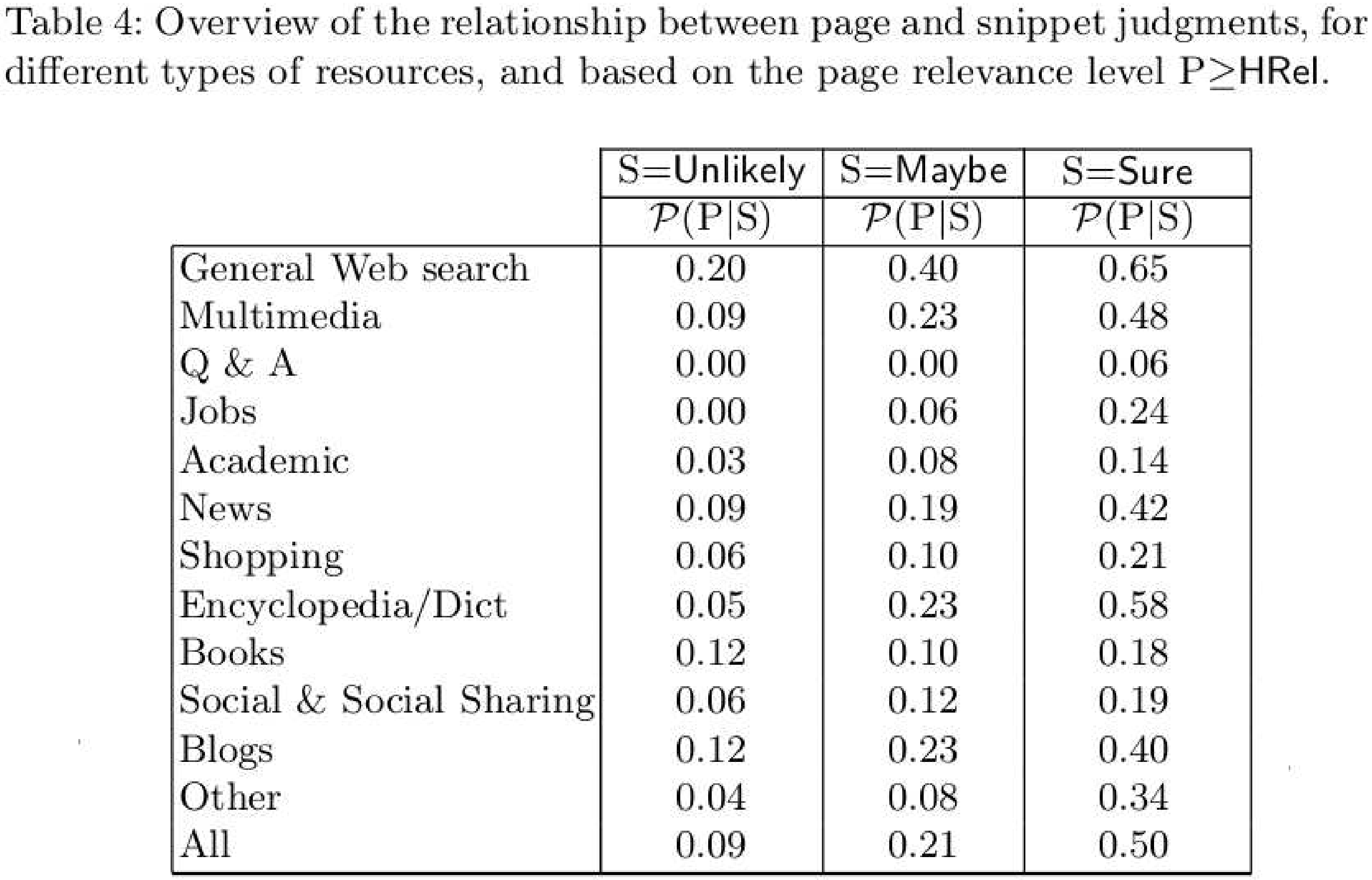


Table 4: Overview of the relationship between page and snippet judgments, for different types of resources, and based on the page relevance level P≥HRel.

	S=Unlikely	S=Maybe	S=Sure
	$\mathcal{P}$(P|S)	$\mathcal{P}$(P|S)	$\mathcal{P}$(P|S)
General Web search	0.20	0.40	0.65
Multimedia	0.09	0.23	0.48
Q & A	0.00	0.00	0.06
Jobs	0.00	0.06	0.24
Academic	0.03	0.08	0.14
News	0.09	0.19	0.42
Shopping	0.06	0.10	0.21
Encyclopedia/Dict	0.05	0.23	0.58
Books	0.12	0.10	0.18
Social & Social Sharing	0.06	0.12	0.19
Blogs	0.12	0.23	0.40
Other	0.04	0.08	0.34
All	0.09	0.21	0.50

CLICKS ON U. TWENTE SEARCH

- 84.4% of clicks are on result 1 (considering resources as results)

CLICKS ON U. TWENTE SEARCH

- Skip/click pairs

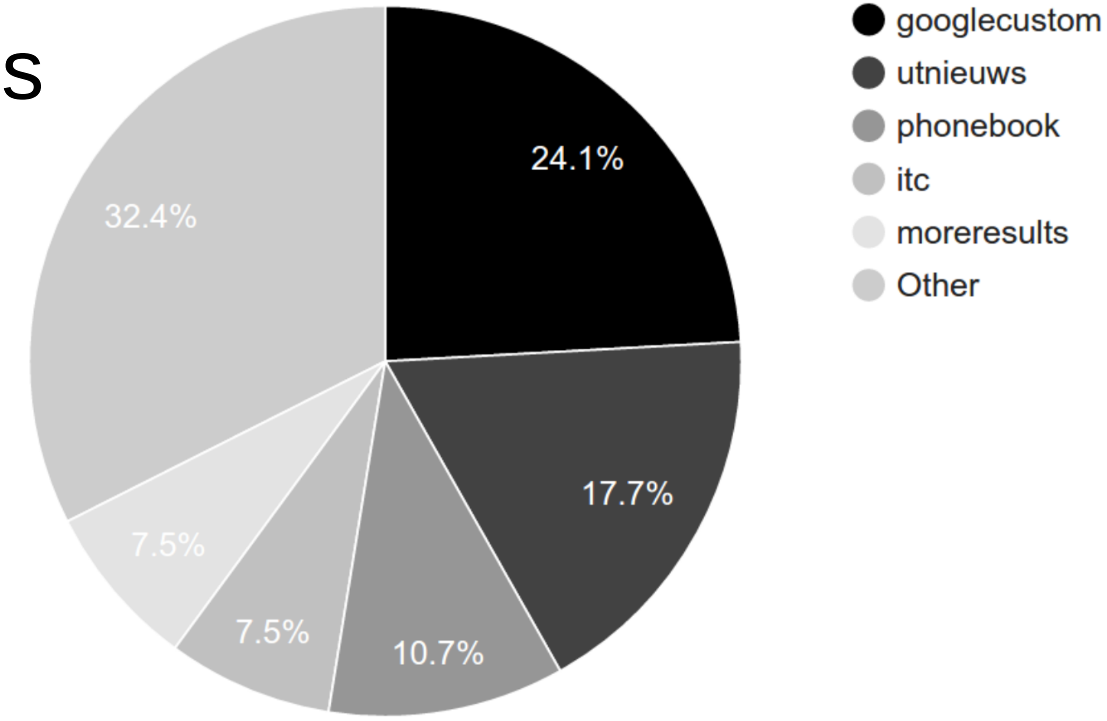


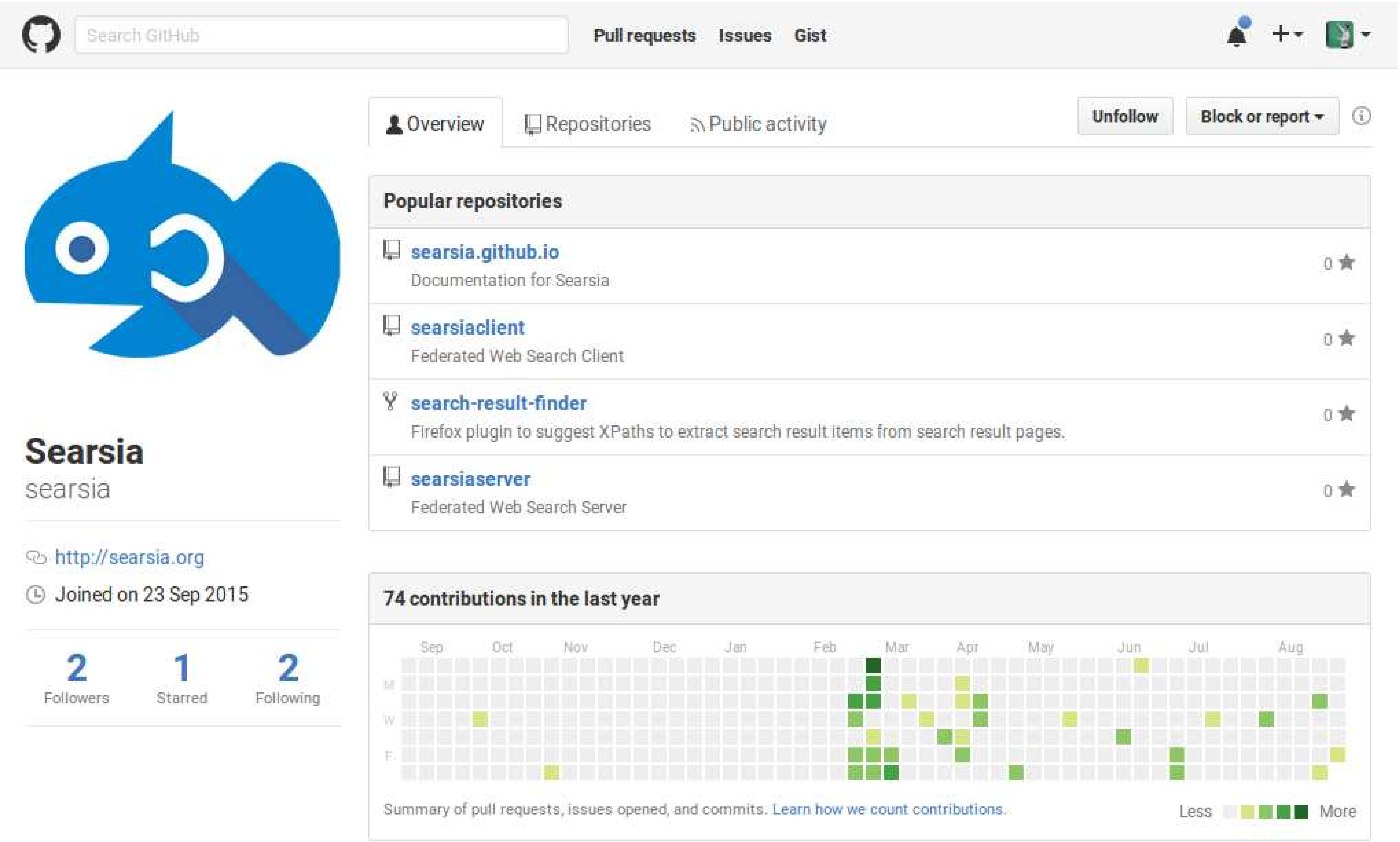
Figure 1. Percentage of clicks on a search engine if the rank is larger than one (everything below 5% is in the other section)

DISCUSSION (3)

- Clicks
  - Bias might be severe in real system
  - Bias not present in test collection (by design!)
  - Clicks are noisy prediction of relevance
- Challenge
  - Study click bias
  - Learn from click data

SOFTWARE

http://github.com/searsia



Search GitHub

Pull requests Issues Gist

Overview Repositories Public activity Unfollow Block or report

Popular repositories

- **searsia.github.io** — Documentation for Searsia — 0 ★
- **searsiaclient** — Federated Web Search Client — 0 ★
- **search-result-finder** — Firefox plugin to suggest XPaths to extract search result items from search result pages. — 0 ★
- **searsiaserver** — Federated Web Search Server — 0 ★

Searsia
searsia

http://searsia.org
Joined on 23 Sep 2015

2 Followers · 1 Starred · 2 Following

74 contributions in the last year

Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug

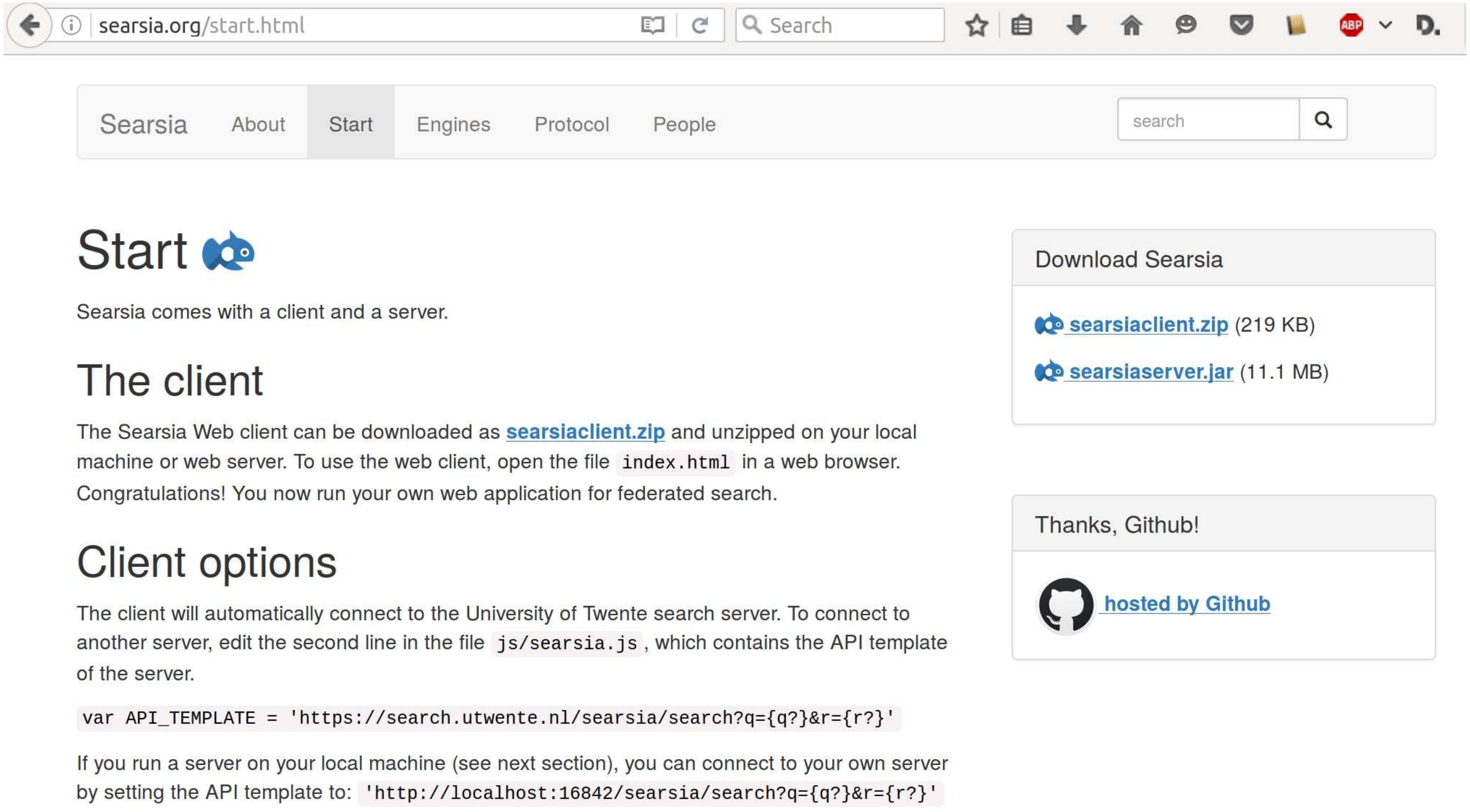
M W F

Summary of pull requests, issues opened, and commits. Learn how we count contributions.

Less ⬜ 🟩 🟩 🟩 🟩 More

DOCUMENTATION

http://searsia.org



Start

Searsia comes with a client and a server.

The client

The Searsia Web client can be downloaded as searsiaclient.zip and unzipped on your local machine or web server. To use the web client, open the file `index.html` in a web browser. Congratulations! You now run your own web application for federated search.

Client options

The client will automatically connect to the University of Twente search server. To connect to another server, edit the second line in the file `js/searsia.js`, which contains the API template of the server.

```
var API_TEMPLATE = 'https://search.utwente.nl/searsia/search?q={q?}&r={r?}'
```

If you run a server on your local machine (see next section), you can connect to your own server by setting the API template to: `'http://localhost:16842/searsia/search?q={q?}&r={r?}'`

Download Searsia

searsiaclient.zip (219 KB)

searsiaserver.jar (11.1 MB)

Thanks, Github!

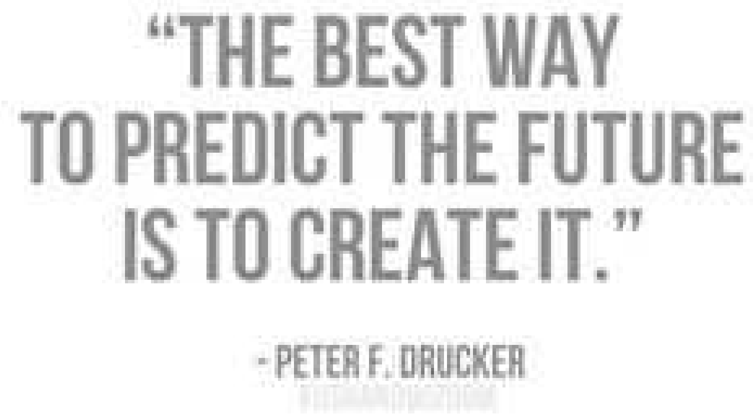
hosted by Github

CONCLUSIONS

- Federated Search as a Living Lab
  - □ Data for Federated Search
  - □ Software for Federated Search
  - □ Coming up: Experiments with our own interaction data

OUTLOOK

- “Green” (no need to crawl & store everything)
- “Democratic” (resources vote for results)
- “Cheap” (for Searsia; costly for NSA ;-))



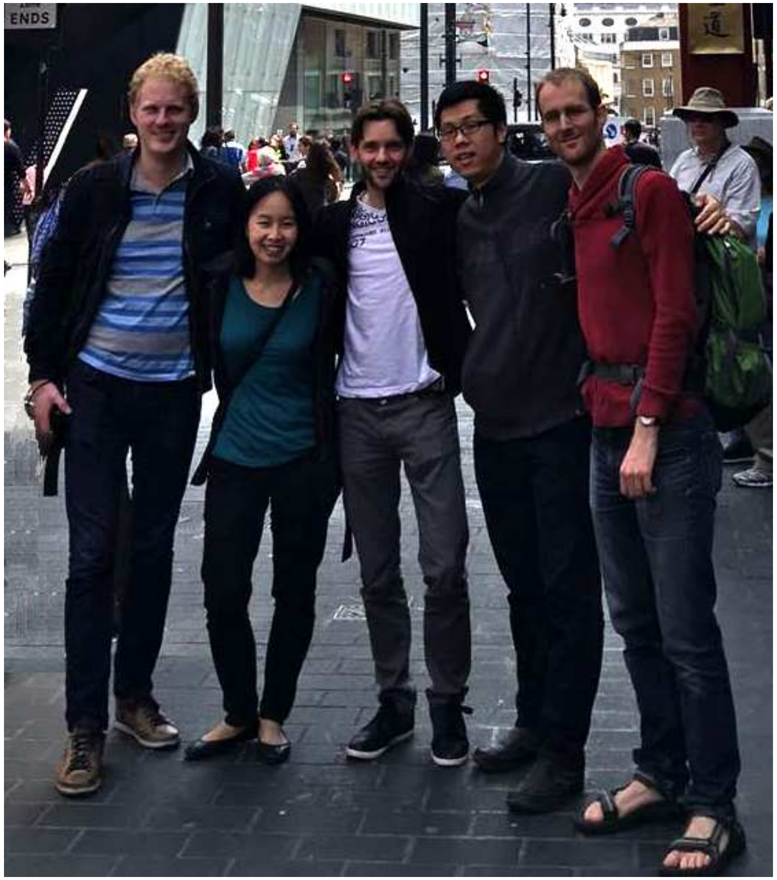
“THE BEST WAY TO PREDICT THE FUTURE IS TO CREATE IT.”

- PETER F. DRUCKER

PUBLICATIONS

- Almer Tigelaar and Djoerd Hiemstra, “Query-Based Sampling using Snippets’’, SIGIR LSDSIR 2010.
- Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. *Federated Search in the Wild: CIKM 2012*.
- Thomas Demeester, et al. “Overview of the TREC Federated Web Search Track’’. *TREC 2015*.
- Thomas Demeester, Dolf Trieschnigg, Ke Zhou, Dong Nguyen, and Djoerd. Hiemstra. FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search. In *WWW* 2015.
- Thomas Demeester, et al., “Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation’’. *Information Retrieval Journal 19*, 2016.

ACKNOWLEDGEMENTS



- Almer Tigelaar
- Mohammad Khelghati
- Rob Stortelder
- Dong Nguyen
- Thomas Demeester
- Ke Zhou
- Dolf Trieschnigg

UNIVERSITY OF TWENTE. NWO COMMIT/ YAHOO! UNIVERSITEIT GENT