WHAT WE (DON'T) KNOW ABOUT SEARCH ENGINES

Dirk Lewandowski

Hamburg University of Applied Sciences, Germany

dirk.lewandowski@haw-hamburg.de
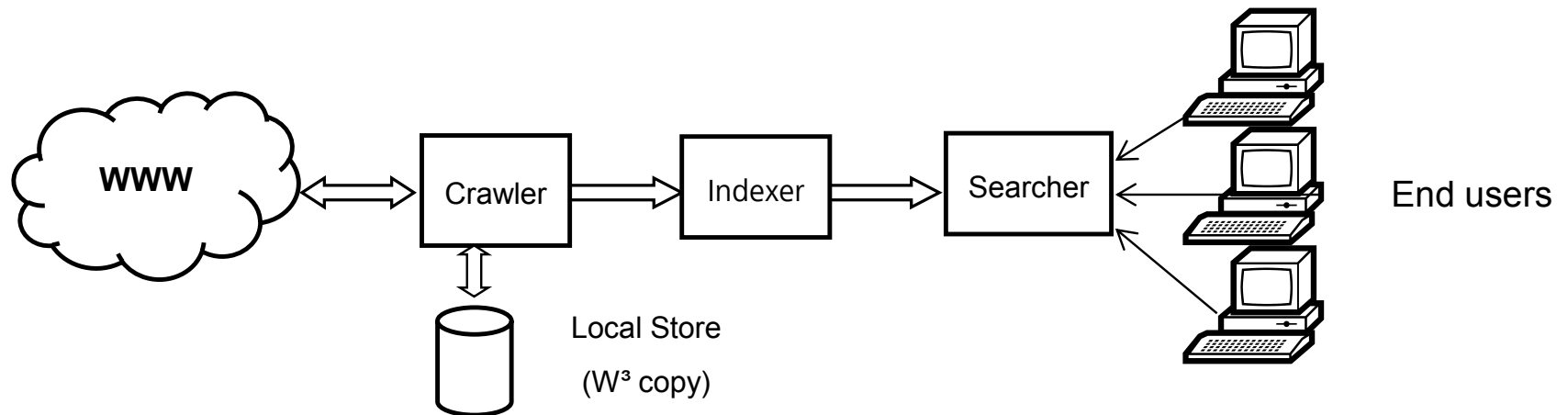
HAW
HAMBURG

AGENDA

1. How search engines work
2. The field of search engine research
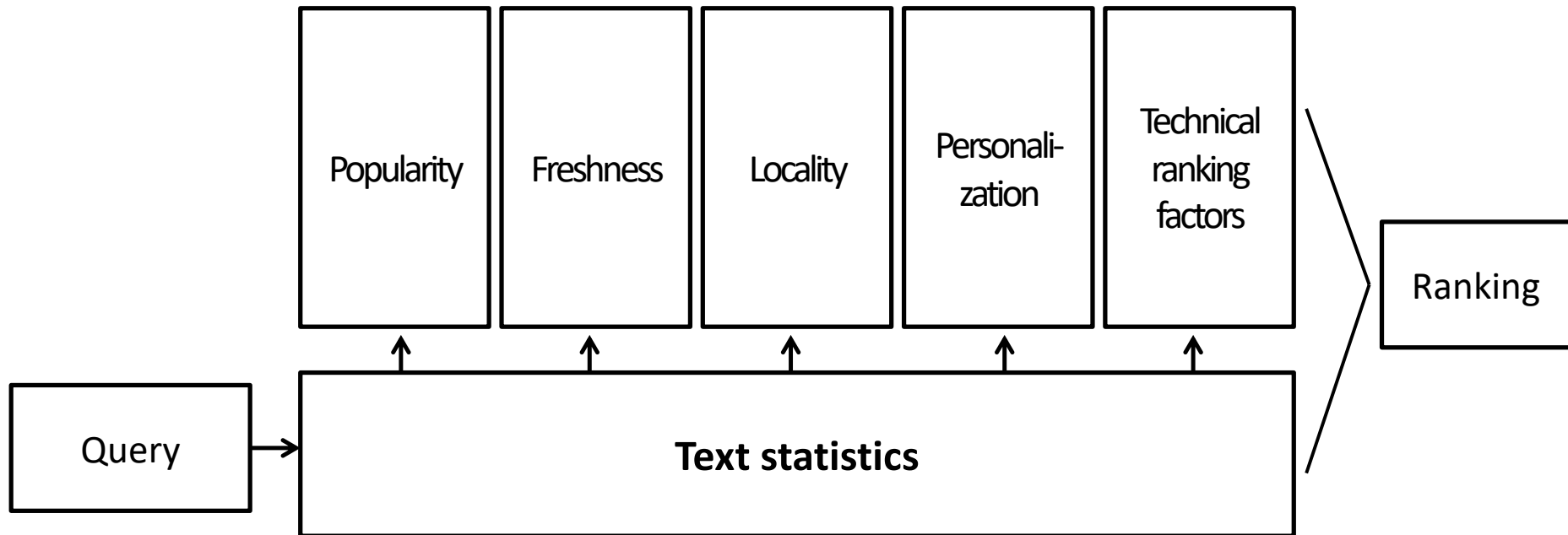3. Research at the search studies research group

HAW HAMBURG

HOW SEARCH ENGINES WORK

Introduction to the technical design of search engines
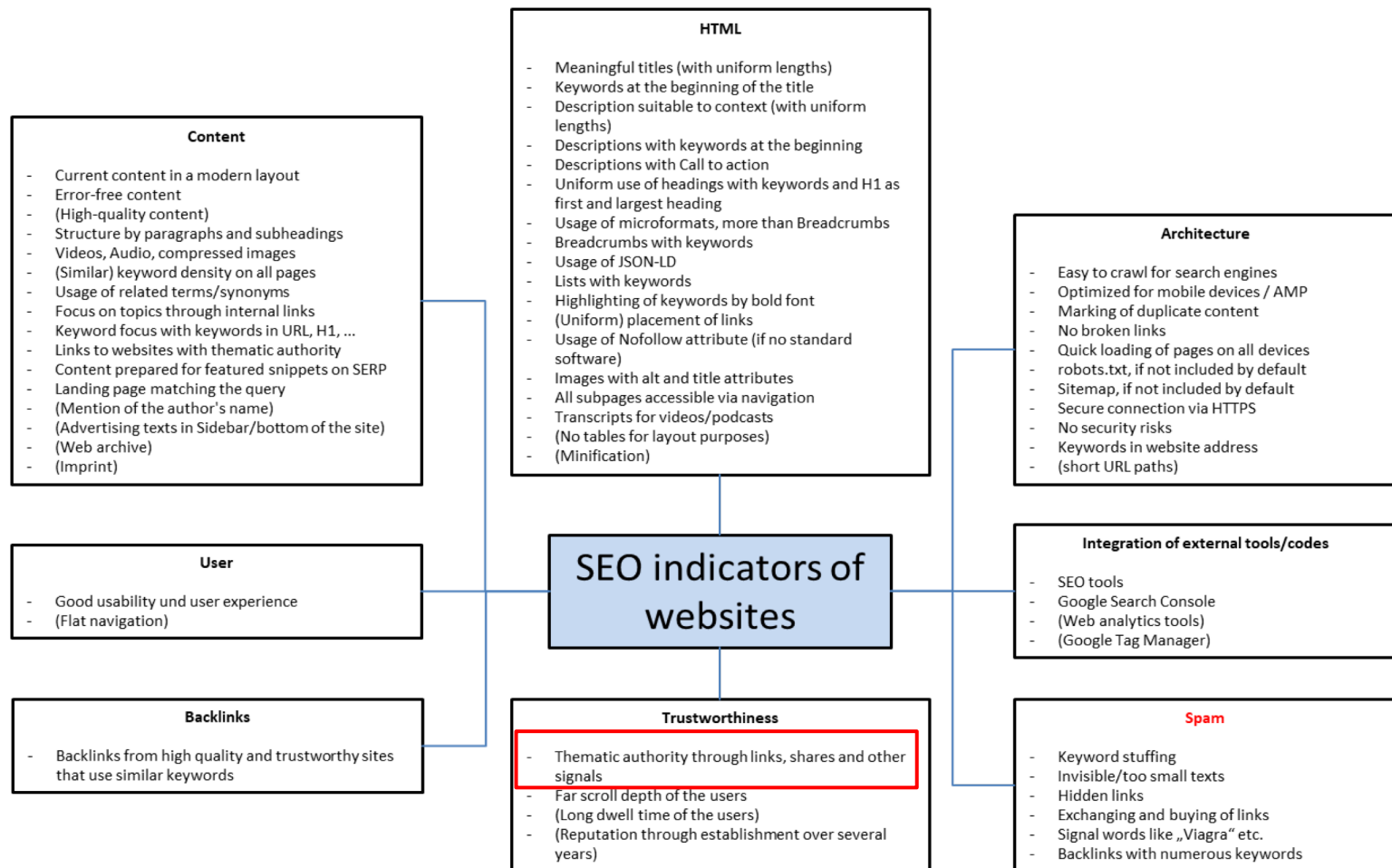
HAW
HAMBURG

GENERAL SEARCH ENGINE ARCHITECTURE

(Risvik & Michelsen, 2002)

SEARCH ENGINE RANKING (SIMPLIFIED)

RANKING FROM THE PERSPECTIVE OF SEARCH ENGINE OPTIMIZATION (SEO)

HTML

- Meaningful titles (with uniform lengths)
- Keywords at the beginning of the title
- Description suitable to context (with uniform lengths)
- Descriptions with keywords at the beginning
- Descriptions with Call to action
- Uniform use of headings with keywords and H1 as first and largest heading
- Usage of microformats, more than Breadcrumbs
- Breadcrumbs with keywords
- Usage of JSON-LD
- Lists with keywords
- Highlighting of keywords by bold font
- (Uniform) placement of links
- Usage of Nofollow attribute (if no standard software)
- Images with alt and title attributes
- All subpages accessible via navigation
- Transcripts for videos/podcasts
- (No tables for layout purposes)
- (Minification)

Content

- Current content in a modern layout
- Error-free content
- (High-quality content)
- Structure by paragraphs and subheadings
- Videos, Audio, compressed images
- (Similar) keyword density on all pages
- Usage of related terms/synonyms
- Focus on topics through internal links
- Keyword focus with keywords in URL, H1, ...
- Links to websites with thematic authority
- Content prepared for featured snippets on SERP
- Landing page matching the query
- (Mention of the author's name)
- (Advertising texts in Sidebar/bottom of the site)
- (Web archive)
- (Imprint)

Architecture

- Easy to crawl for search engines
- Optimized for mobile devices / AMP
- Marking of duplicate content
- No broken links
- Quick loading of pages on all devices
- robots.txt, if not included by default
- Sitemap, if not included by default
- Secure connection via HTTPS
- No security risks
- Keywords in website address
- (short URL paths)

User

- Good usability und user experience
- (Flat navigation)

SEO indicators of websites

Integration of external tools/codes

- SEO tools
- Google Search Console
- (Web analytics tools)
- (Google Tag Manager)

Backlinks

- Backlinks from high quality and trustworthy sites that use similar keywords

Trustworthiness

- Thematic authority through links, shares and other signals
- Far scroll depth of the users
- (Long dwell time of the users)
- (Reputation through establishment over several years)

Spam

- Keyword stuffing
- Invisible/too small texts
- Hidden links
- Exchanging and buying of links
- Signal words like „Viagra" etc.
- Backlinks with numerous keywords

Explanation:
Indicators (or parts thereof) that are in brackets were controversial (agreement by one interviewee, rejection by another)

FAKULTÄT DMI, DEPARTMENT INFORMATION
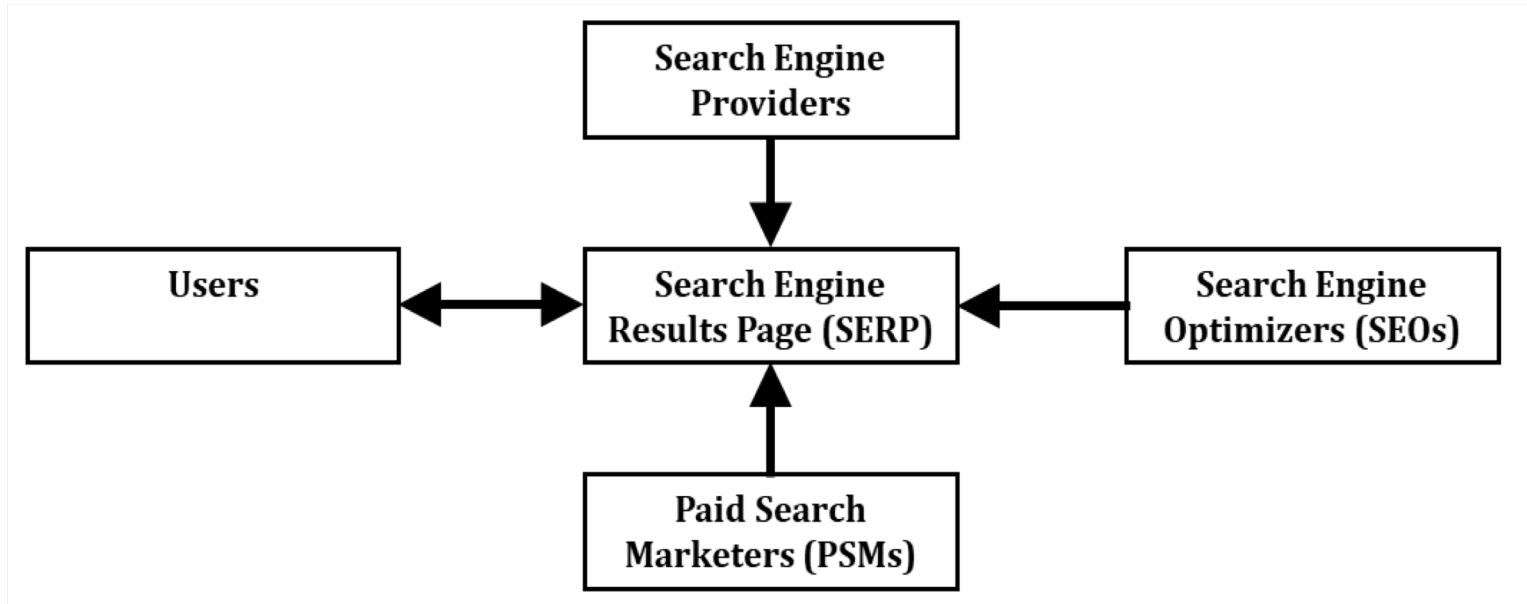Prof. Dr. Dirk Lewandowski

HAW HAMBURG

HOW DO SEARCH ENGINES MEASURE QUALITY?

Different types of "quality" signals

- Text-based signals (e.g., keyword density, text structure)
- Link-based signals (e.g., weighted number of in-links)
- Click-based signals (e.g., clicks on results, dwell time)
- User-based signals (personalization)
- Collection-based signals (e.g., belongs to the collection of news websites)

Measuring quality through these signals generally works well in practice.

HAW
HAMBURG

INFLUENCES ON SEARCH RESULTS

SOME PROBLEMS RESULTING FROM THE "POPULARITY APPROACH" (1)

(2016)

https://searchengineland.com/googles-results-no-longer-in-denial-over-holocaust-265832

HAW HAMBURG

"Search is a reflection of the content that exists on the web. The fact that hate sites appear in search results in no way means that Google endorses these views."

(Google 2016)

SOME PROBLEMS RESULTING FROM THE "POPULARITY APPROACH" (2)

Query: Sportlerinnen [female athletes]

Web results only

HAW HAMBURG

BIAS IN ORGANIC RESULTS IS JUST THE TIP OF THE ICEBERG

Facets of search engine bias

1. **Search engine (engineering)** – Ranking
2. **Users** – Cognitive biases
3. **Search engine providers' self-interests** – Design of search engine result pages
4. **External influences** – Search Engine Optimization (SEO)

HAW
HAMBURG

BIAS ON THE WEB

Figure 1. The vicious cycle of bias on the Web.

(Baeza-Yates, 2018)

FAKULTÄT DMI, DEPARTMENT INFORMATION
Prof. Dr. Dirk Lewandowski

HAW
HAMBURG
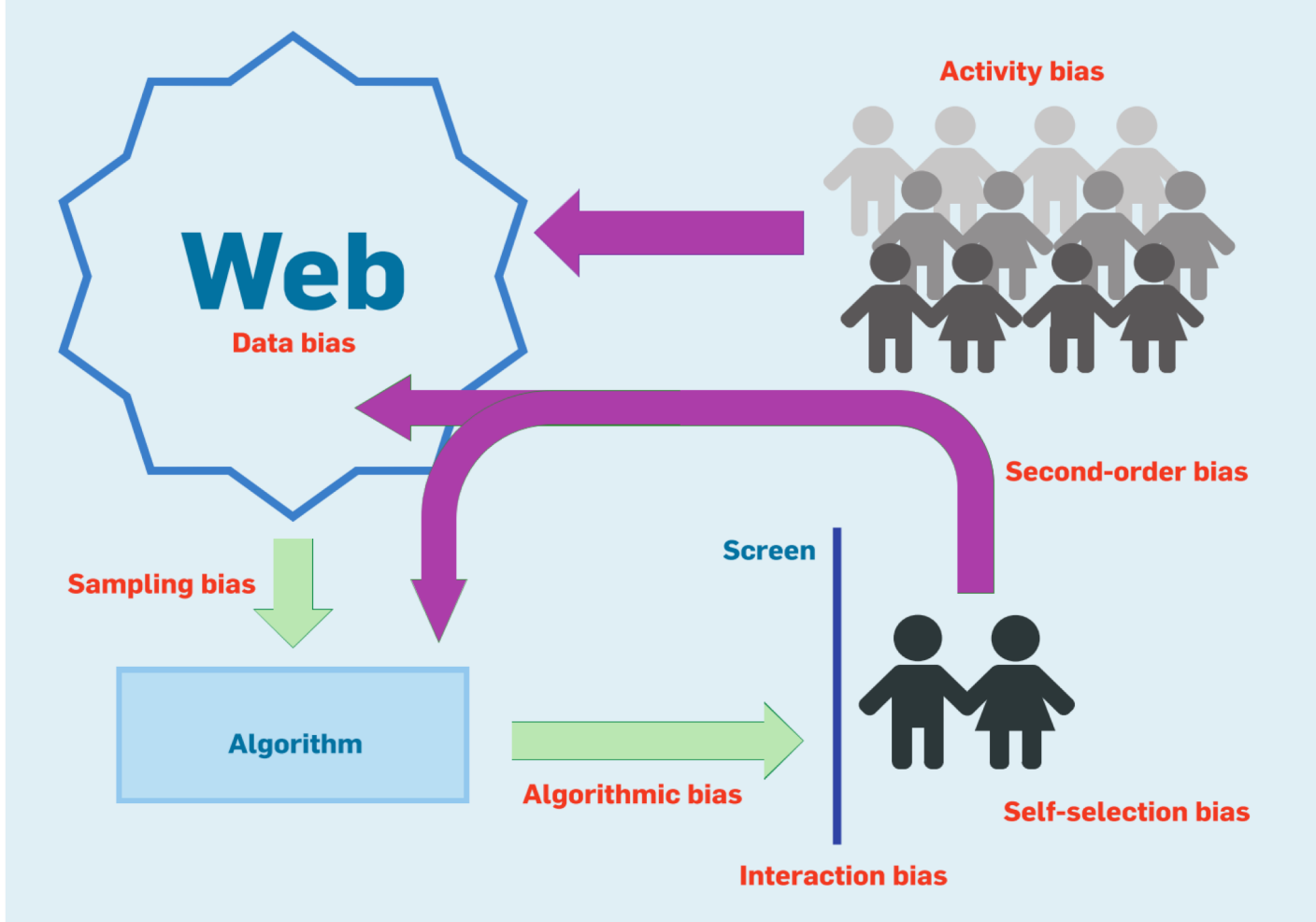
FURTHER READING: SEARCH ENGINE BASICS

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search Engines: Information retrieval in practice*. **Boston, MA: Pearson.**
Good introduction to information retrieval.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. **Cambridge: Cambridge University Press.**
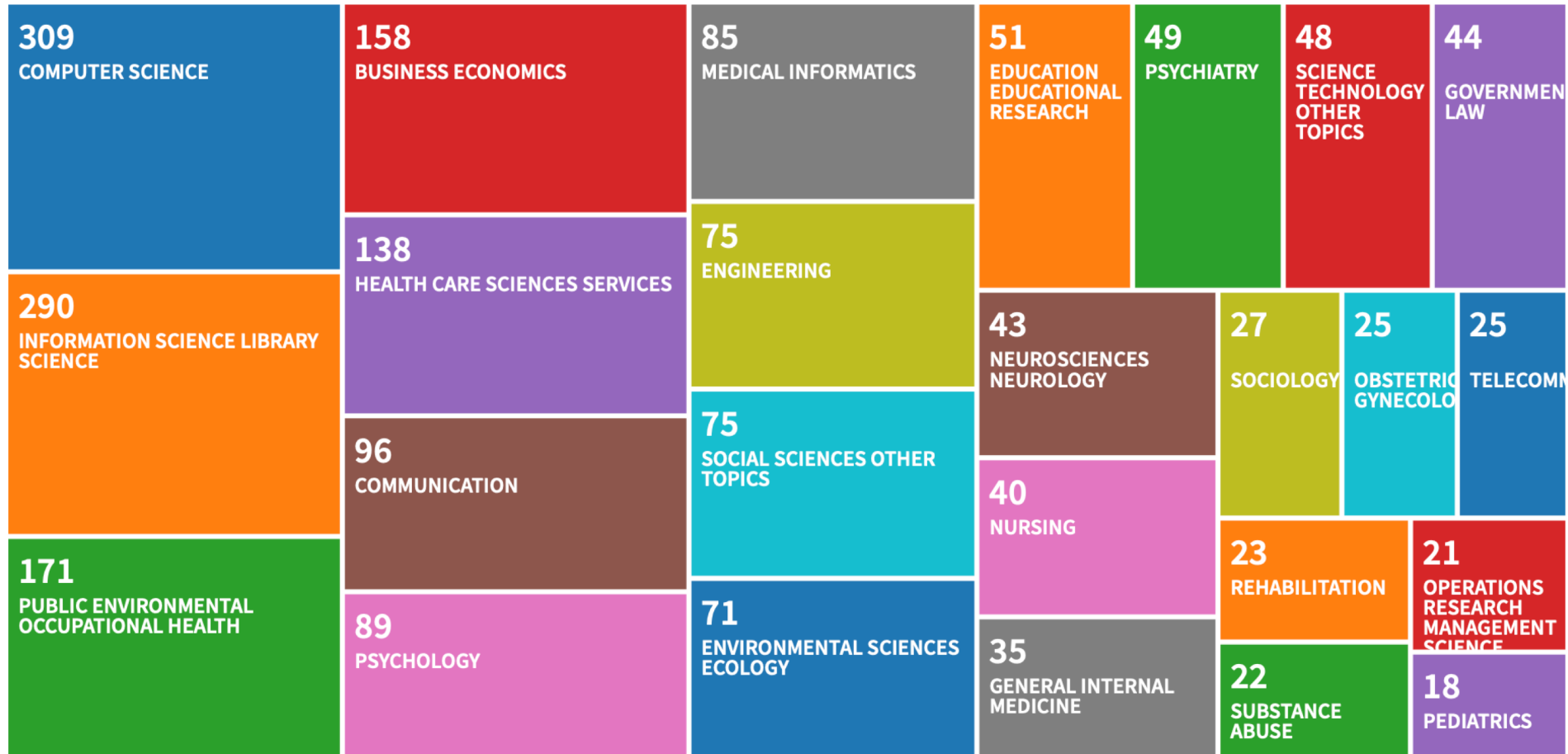Good introduction to information retrieval. Goes a bit deeper than Croft et al.

White, R. W. (2016). *Interactions with Search Systems*. **New York: Cambridge University Press.**
Good overview of interactive information retrieval research. Complete overview of the research literature in the field.

HAW
HAMBURG

THE FIELD OF SEARCH ENGINE RESEARCH

HAW HAMBURG

A QUICK OVERVIEW OF SEARCH ENGINE RESEARCH (2016-2020)

309 COMPUTER SCIENCE

290 INFORMATION SCIENCE LIBRARY SCIENCE

171 PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH

158 BUSINESS ECONOMICS

138 HEALTH CARE SCIENCES SERVICES

96 COMMUNICATION

89 PSYCHOLOGY

85 MEDICAL INFORMATICS

75 ENGINEERING

75 SOCIAL SCIENCES OTHER TOPICS

71 ENVIRONMENTAL SCIENCES ECOLOGY

51 EDUCATION EDUCATIONAL RESEARCH

49 PSYCHIATRY

48 SCIENCE TECHNOLOGY OTHER TOPICS

44 GOVERNMEN LAW

43 NEUROSCIENCES NEUROLOGY

40 NURSING

35 GENERAL INTERNAL MEDICINE

27 SOCIOLOGY

25 OBSTETRIC GYNECOLO

25 TELECOMM

23 REHABILITATION

22 SUBSTANCE ABUSE

21 OPERATIONS RESEARCH MANAGEMENT SCIENCE

18 PEDIATRICS

(Data from Web of Science; topic "search engine*")

HAW HAMBURG

THE FIELD OF SEARCH ENGINE RESEARCH

General characteristics

- Multidisciplinary field
- Output is growing but not very fast
- Research focussing on commercial search engines vs. research and development in the field of information retrieval more generally

"Data problems"

- Lack of query data (real queries from real users)
- Very limited access to search results data (Bing API)
- Data access through cooperation with search engine companies

HAW
HAMBURG

TOPICAL CLUSTERS (NOT SYSTEMATIC)

- Using Google Trends to predict real-world events (stock prices, illnesses, etc.)
- Lots of computer science research focusing on developing search engine components
- Improving the effectiveness of search engine advertising

- Search engine user behaviour
- Search engine bias
- Evaluating the quality of Web search engines
- Search engine optimization (SEO)
- Search engines and information literacy

- Search engine market and regulation

- …

HAW
HAMBURG

RESEARCH TOPICS AND THE EVIDENTIAL VALUE OF USER STUDIES
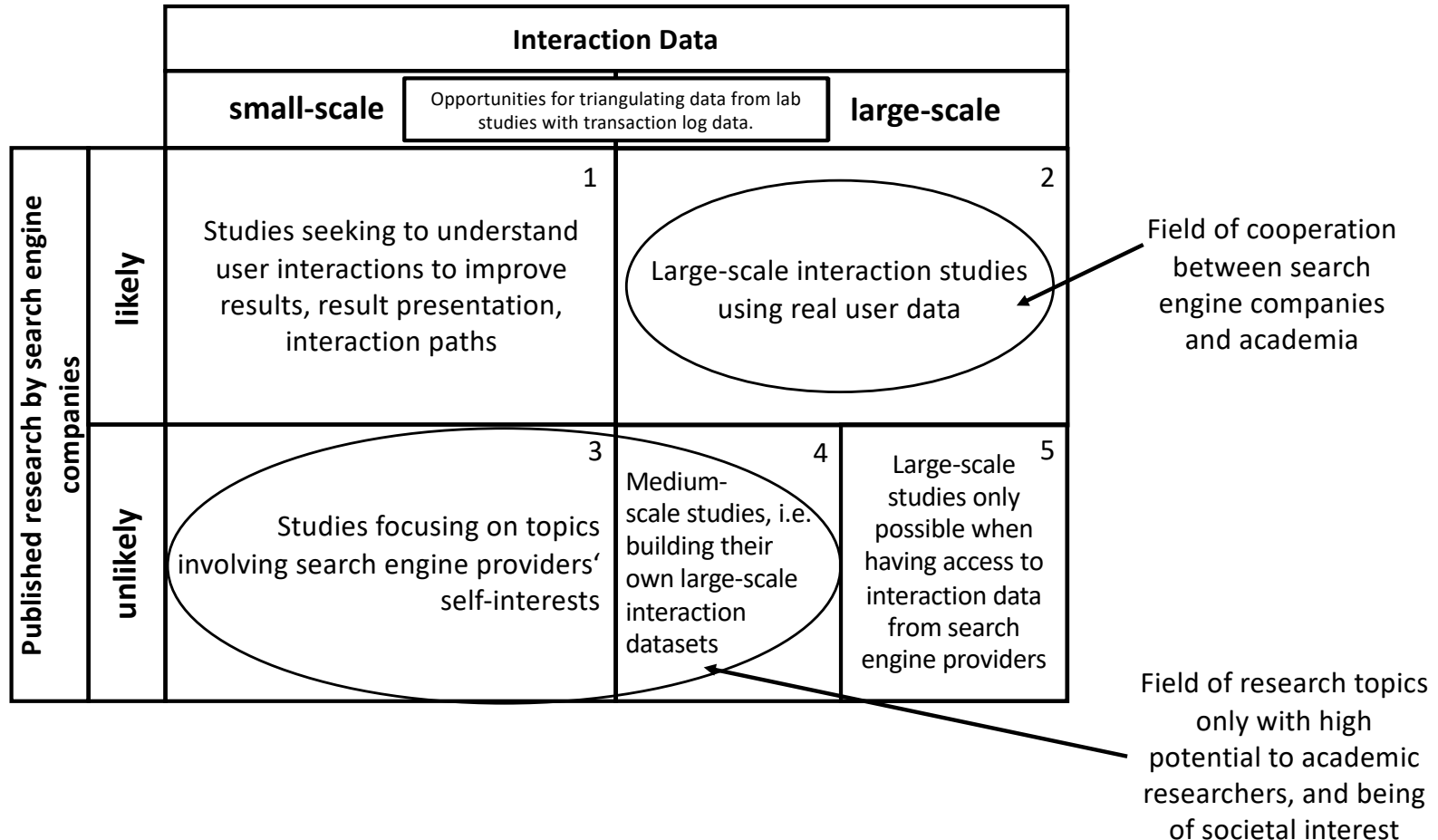
(Lewandowski, Sünkler & Schultheiß, 2020)

FAKULTÄT DMI, DEPARTMENT INFORMATION
Prof. Dr. Dirk Lewandowski

HAW HAMBURG

BOOKS ON SEARCH ENGINES 2016-2020

White, R. W. (2016). *Interactions with Search Systems*. **New York: Cambridge University Press.**
Good overview of interactive information retrieval research. Complete overview of the research literature in the field.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. **New York, USA: New York University Press.**
Illustrates the problem of search engine bias but many of the examples are hopelessly outdated.

Halavais, A. (2018). *Search Engine Society* **(Second edition). Cambridge, UK: Polity.**
Good starting point as the author discusses a lot of aspects relevant to search engine research. A bit weak on the technical side.

Haider, J., and O. Sundin (2019), *Invisible Search and Online Search Engines*. **Oxford, New York: Routledge.**

FAKULTÄT DMI, DEPARTMENT INFORMATION
Prof. Dr. Dirk Lewandowski

HAW
HAMBURG

RESEARCH AT THE SEARCH STUDIES RESEARCH GROUP

TWO PILLARS

User studies	Search engine data analysis
- Task-based surveys, lab studies (incl. eye tracking) - Focus on representative samples (surveys), diverse samples (lab studies)	- Automatic data collection (screen scraping SERPs) - (Semi-)automatic analysis of URLs and result documents - Software development

HAW
HAMBURG

CURRENT RESEARCH PROJECTS

1. Contextual advertising

How well do users understand advertising on search engine result pages?

What effect do ads have on users' selection behaviour?

2. Search result analysis

What is actually shown on SERPs?

3. SEO-Effekt: Influence of search engine optimization on the search results

SEO from the perspective of optimizers and users

Software for identifying optimized documents in search results

HAW
HAMBURG

1. CONTEXTUAL ADVERTISING

APPROACH

Two large-scale studies representative of the German online population (data from 2014, 2020)

1. Users' knowledge on Google's business model
2. Four tasks where users had to mark all ads on SERPs
3. Experiment with manipulated SERPs (2014 only)

→ Results from the 2020 study (as yet unpublished) confirm the older results.

Eye-tracking study (data from 2019)

HAW
HAMBURG

SAMPLE TASK: MARK ALL THE ADVERTISEMENTS ON THIS SEARCH ENGINE RESULTS PAGE

N=1,000
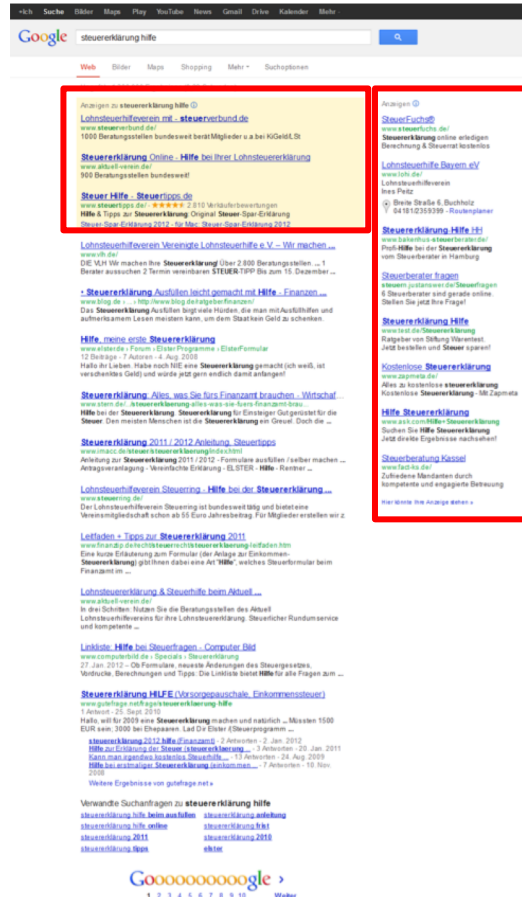
35% marked all ads correctly.

18% marked at least some organic results as ads.

Still, more than 90% of users regard themselves as competent when it comes to using search engines.

Lewandowski, D., Kerkmann, F., Rümmele, S., & Sünkler, S. (2018). An empirical investigation on search engine ad disclosure. *Journal of the Association for Information Science and Technology*, *69*(3), 420–437. https://doi.org/10.1002/asi.23963

25

HAW HAMBURG

RESULTS FROM THE EXPERIMENT

N=1,000
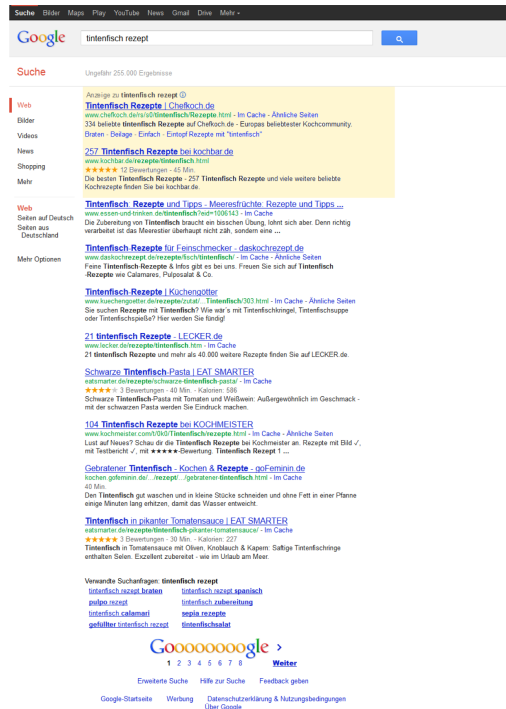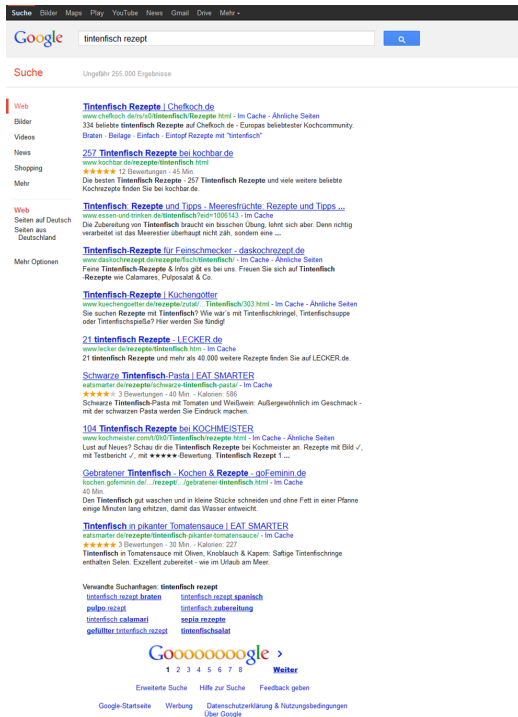
Users not able to distinguish ads from organic results clicked on the first ad about twice as often (40.3% vs. 21.6%).

Lewandowski, D. (2017). Users' Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice*, *5*(4), 6–25. https://doi.org/10.1633/JISTaP.2017.5.4.1

HAW HAMBURG

UNDERSTANDING OF ADS INFLUENCES USERS' VISUAL BEHAVIOUR

N=100

Participants with a low level of knowledge on search advertising are more likely to click on ads than subjects with a high level of knowledge.

Subjects with little knowledge show less willingness to scroll down to organic results.

Larger effect on the mobile device

Q14: "Imagine you want to find information about Work-&-Travel programs in Australia. A Google search returned the following results. Please click on a result."
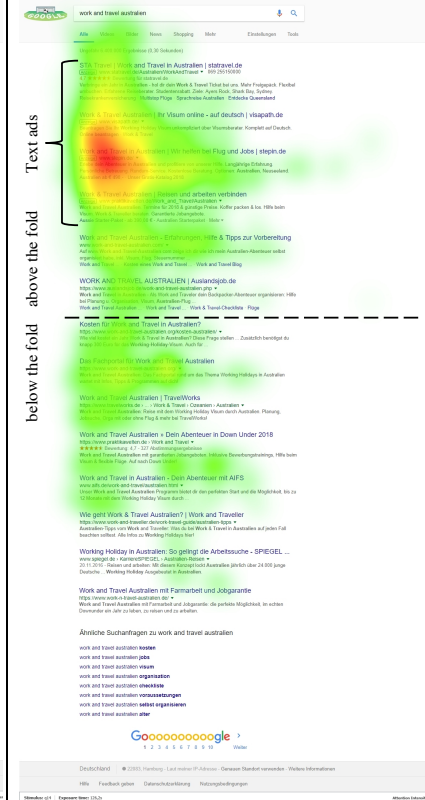
(Schultheiß & Lewandowski, 2020)

HAW HAMBURG

2. SEARCH RESULT ANALYSIS

TOOLS FOR COLLECTING AND ANALYSING RESULTS FROM COMMERCIAL SEARCH ENGINES

Relevance Assessment Tool (RAT; Lewandowski & Sünkler, 2013)

- Screen-scraping SERPs: Collecting URLs, downloading result documents
- Test-design interface
- Juror interface for collecting (relevance) assessments

Some tools for analysing search results developed in different projects

- Online shop identification, grocery shop identification
- Imprint data
- Indicators for search engine optimization

HAW
HAMBURG

STUDIES USING SCREEN SCRAPING AND RESULT ANALYSIS (1/2)

German Federal Elections 2017

Use of info boxes (principles and priorities) by political candidates

Those who are more established make more considerable efforts to achieve visibility on the result pages.

The content analysis shows different degrees of internal pluralism of parties

(Hinz, Sünkler & Lewandowski, 2020)

# STUDIES USING SCREEN SCRAPING AND RESULT ANALYSIS (2/2)

Insurance offering

- Setof 121 queries containing "insurance" and "comparison", e.g., "car insurance comparison sites", "buildings insurance comparison".

- Shows a high concentration of some commercial domains within the top search results (not necessarily providing fair and unbiased insurance comparisons)
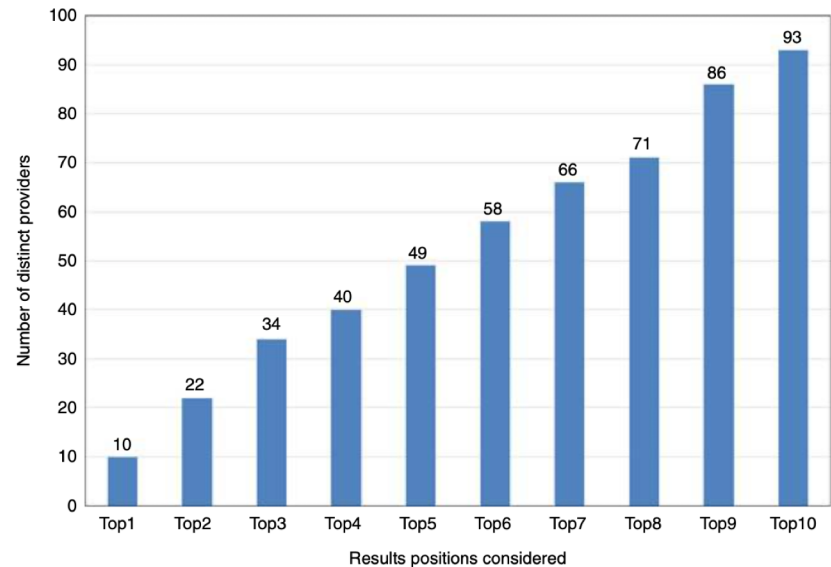
(Lewandowski & Sünkler, 2019)

HAW HAMBURG

SEO-EFFEKT: INFLUENCE OF SEARCH ENGINE OPTIMIZATION ON THE SEARCH RESULTS

HAW
HAMBURG

SEO-EFFEKT

HAW
HAMBURG
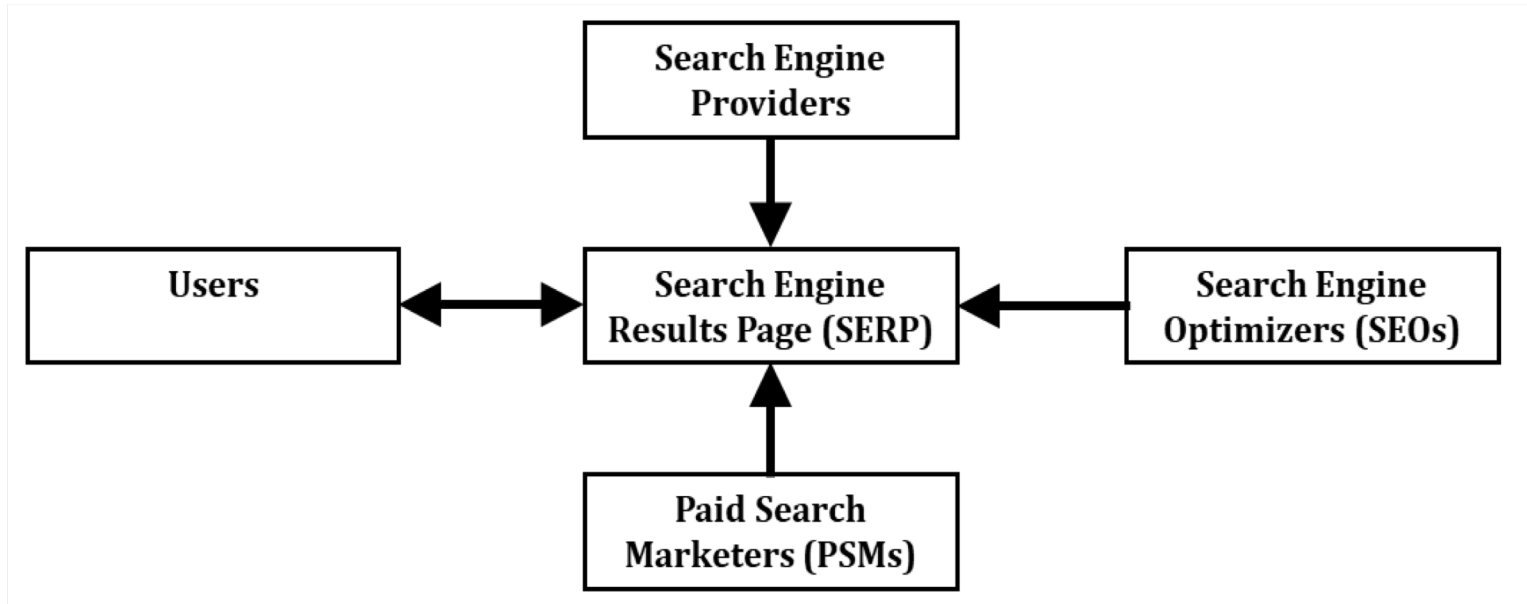
SEO-EFFEKT: SOME FINDINGS

Expert interviews (Schultheiß & Lewandowski, submitted)

- Search engine optimizers regard SEO as a means to providing users with better results.
- SEO is considered necessary for content providers to ensure visibility.

Representative user study (publication in preparation)

- Users have a very limited understanding of SEO.

Trust in Google (results from representative user study; submited)

- Users with little search engine knowledge are more likely to trust and use Google than users with more knowledge.

Data analysis (publication in preparation)

- Tool used to extract information from HTML code, e.g., use of specialized SEO tools, analytics tools + URL classification (pre-defined sets) [+machine learning].
- Demo tool: http://5.189.155.20:5000

34

HAW
HAMBURG

DEMO TOOL RESULTS

Results for https://www2.bui.haw-hamburg.de/pers/ulrike.spree/

Tools & Plugins

SEO Tools: ✗
Analytics Tools: ✗

URL Classification

Online Marketing Customer: ✗
News Service: ✗
Ads on Website: ✗
Company Website: ✗
Online Shop: ✗
Not optimized: ✗

Indicators for SEO

https: ✔
Description: ✗
Title: ✔
robots.txt:: ✗
Sitemap: ✗
nofollow: ✗
speed: ✔
canonical: ✗
viewport: ✗
Micros: ✗

SEO Assessment

Most probably: ✗
Most probably not: ✗
Probably: ✗
Likely not optimized: ✔
Uncertain: ✔

FAKULTÄT DMI, DEPARTMENT INFORMATION
Prof. Dr. Dirk Lewandowski

HAW HAMBURG

SUMMARY OF RESEARCH PROJECTS

- Combining social science and computer science methods provides a fruitful basis for conducting research on commercial search engines.
- Collecting data from commercial search engines is cumbersome but is the only way forward given that search engine companies will only provide data for research projects they endorse.

HAW
HAMBURG

THANK YOU

Dirk Lewandowski
Hamburg University of Applied Sciences, Hamburg, Germany
dirk.lewandowski@haw-hamburg.de
www.searchstudies.org/dirk

HAW
HAMBURG

REFERENCES

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search Engines: Information retrieval in practice*. Boston, MA: Pearson.

Haider, J., & Sundin, O. (2019). *Invisible Search and Online Search Engines*. Oxford, New York: Routledge.

Halavais, A. (2018). *Search Engine Society* (Second edition). Cambridge, UK: Polity.

Hinz, K., Sünkler, S., & Lewandowski, D. (2020). Selbstdarstellung und Positionierung von Kandidatinnen und Kandidaten zur Bundestagswahl 2017 in Google-Infoboxen. *Medien & Kommunikationswissenschaft*, *68*(1–2), 94–112. https://doi.org/10.5771/1615-634X-2020-1-2-94

Lewandowski, D. (2017). Users' Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice*, *5*(4), 6–25. https://doi.org/10.1633/JISTaP.2017.5.4.1

Lewandowski, D., Kerkmann, F., Rümmele, S., & Sünkler, S. (2018). An empirical investigation on search engine ad disclosure. *Journal of the Association for Information Science and Technology*, *69*(3), 420–437. https://doi.org/10.1002/asi.23963

Lewandowski, D., & Sünkler, S. (2019). What does Google recommend when you want to compare insurance offerings? *Aslib Journal of Information Management*, *71*(3), 310–324. https://doi.org/10.1108/AJIM-07-2018-0172

Lewandowski, D., Sünkler, S., & Schultheiß, S. (2020). Studies on Search: Designing Meaningful IIR Studies on Commercial Search Engines. *Datenbank-Spektrum*, *20*(1), 5–15. https://doi.org/10.1007/s13222-020-00331-1

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, USA: New York University Press.

Risvik, K. M., & Michelsen, R. (2002). Search engines and web dynamics. *Computer Networks*, *39*(3), 289–302.

Schultheiß, S., & Lewandowski, D. (2020). How users' knowledge of advertisements influences their viewing and selection behavior in search engines. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24410

White, R. W. (2016). *Interactions with Search Systems*. New York: Cambridge University Press.

FAKULTÄT DMI, DEPARTMENT INFORMATION
Prof. Dr. Dirk Lewandowski

HAW
HAMBURG