

HOW DATA SCIENCE APPROACHES CAN COMPLEMENT SEARCH ENGINE RESEARCH IN INFORMATION SCIENCE

Dirk Lewandowski

Hamburg University of Applied Sciences, Germany

dirk.lewandowski@haw-hamburg.de

**SEARCH
STUDIES**

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation



**HAW
HAMBURG**

AGENDA

1. Data Science, Search Engine Research, and Information Science
2. Data in search engine research
3. Some examples of data science studies from the Search Studies research group
4. Research software
5. Summary and conclusion

1 Data Science, Search Engine Research, and Information Science

“Many of the trumpeted concepts of data science can be seen simply as a rediscovery of existing concepts from traditional fields such as library science, hybridized with computer science and statistics.”

Dov Greenbaum, Mark Gerstein (*Science* 365, 2019, issue 6455, p. 764)



WIKIPEDIA
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Current events](#)

[Random article](#)

[About Wikipedia](#)

[Contact us](#)

[Donate](#)

[Contribute](#)

[Help](#)

[Learn to edit](#)

[Community portal](#)

[Recent changes](#)

[Upload file](#)

[Tools](#)

[What links here](#)

[Related changes](#)

[Special pages](#)

Article [Talk](#)

Read [Edit](#) [View histor](#)

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is an [interdisciplinary](#) field that uses [scientific methods](#), processes, [algorithms](#) and systems to extract or extrapolate [knowledge](#) and insights from noisy, structured and [unstructured data](#),^{[1][2]} and apply knowledge from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).^[3]

Data science is a "concept to unify [statistics](#), [data analysis](#), [informatics](#), and their related [methods](#)" in order to "understand and analyse actual [phenomena](#)" with [data](#).^[4] It uses techniques and theories drawn from many fields within the context of [mathematics](#), statistics, [computer science](#), [information science](#), and [domain knowledge](#).^[3] However, data science is different from [computer science](#) and information science. [Turing Award](#) winner [Jim Gray](#) imagined data science as a "fourth paradigm" of science ([empirical](#), [theoretical](#), [computational](#), and now data-driven) and asserted that "everything about science is changing because of the impact of [information technology](#)" and the [data deluge](#).^{[5][6]}

A **data scientist** is someone who creates programming code and combines it with statistical knowledge to create insights from data.^[7]

https://en.wikipedia.org/wiki/Data_science



WIKIPEDIA
The Free Encyclopedia

- [Main page](#)
 - [Contents](#)
 - [Current events](#)
 - [Random article](#)
 - [About Wikipedia](#)
 - [Contact us](#)
 - [Donate](#)
-
- [Contribute](#)
 - [Help](#)
 - [Learn to edit](#)
 - [Community portal](#)
 - [Recent changes](#)
 - [Upload file](#)
-
- [Tools](#)
 - [What links here](#)
 - [Related changes](#)
 - [Special pages](#)

Article [Talk](#)

[Read](#) [Edit](#) [View histor](#)

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is an interdisciplinary field that uses [scientific methods](#), processes, [algorithms](#) and systems to extract or extrapolate [knowledge](#) and [insights](#) from noisy, structured and unstructured data,^{[1][2]} and apply knowledge from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).^[3]

Data science is a "concept to unify [statistics](#), [data analysis](#), [informatics](#), and their related [methods](#)" in order to "understand and analyse actual [phenomena](#)" with [data](#).^[4] It uses techniques and theories drawn from many fields within the context of [mathematics](#), statistics, [computer science](#), [information science](#), and [domain knowledge](#).^[3] However, data science is different from [computer science](#) and information science. [Turing Award](#) winner [Jim Gray](#) imagined data science as a "fourth paradigm" of science ([empirical](#), [theoretical](#), [computational](#), and now data-driven) and asserted that "everything about science is changing because of the impact of [information technology](#)" and the [data deluge](#).^{[5][6]}

A **data scientist** is someone who creates programming code and combines it with statistical knowledge to create insights from data.^[7]

https://en.wikipedia.org/wiki/Data_science

WHAT DATA?

“Data Science revolution”

- Changes in **data collection** (more data + more data sources) and **processing** (more computing power)

What data do we use?

- Data specifically collected for research studies
- Data collected in processes elsewhere (e.g., transaction-log files from web servers, customer data from e-commerce vendors, location data from driverless cars)

HOW MUCH DATA?

	Movies	Music	Search Queries	Clicked Search Results	Web Browsing
Items	17,770	702,896	512,323,034	20,301,327	2,012,617
Users	429,541	2,156,792	57,524,526	57,758,157	109,315
Observations	99,548,085	755,480,158	2,613,137,669	2,491,026,154	287,189,911

Table 1: Descriptive statistics for the five datasets analyzed. Observations correspond to ratings, queries, click events, and page views, as appropriate to the domain. Movie data obtained from Netflix; music, search queries and clicked search results data obtained from Yahoo!; and Web browsing data obtained from the Nielsen Company.

insurance comparison), “Haftpflichtversicherung im Vergleich” (liability insurance in comparison). This procedure identified a total of 121 different search queries. All queries were in German and are here translated for illustrative purposes only.

SEARCH ENGINE RESEARCH

- Traditionally in fields such as Information Retrieval, Information (Seeking) Behaviour
- More and more interest from the social sciences (political science, media and communication, etc.)
- **Common interest in *commercial* search systems**, i.e., systems where researchers do not have access to the system.

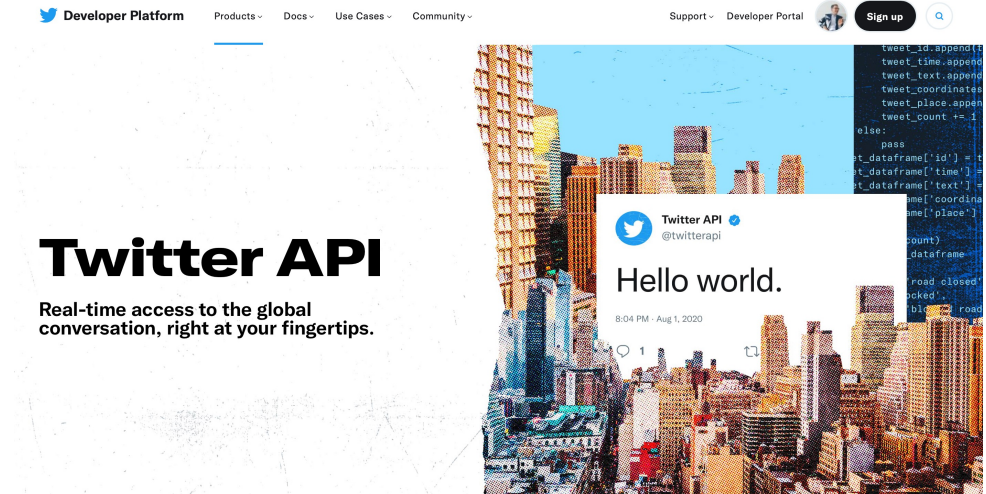
HOW DOES THIS ALL RELATE TO INFORMATION SCIENCE?

- **Data science approaches are worthwhile additions to more traditional methods** used in information science. Main advantage is to scale up studies where manual data collection has been used, where data collection was not possible or required lots of manual labour.
- **The search landscape has changed** over the last 20 years.
- **Number of Google searches** – When people use Google for finding information to such a large degree, this should naturally of interest to information science.
- Search engine research has to a large part moved to more computer-science focused communities like CHIIR but that **research is narrowly focused on improving IR systems**.

2 Data in search engine research

WHY IS THERE SO MUCH SOCIAL MEDIA RESEARCH AND SO LITTLE SEARCH ENGINE RESEARCH?

- Lots of social media research; blind spot when it comes to search engines (e.g., Norocel, 2021)
- Making valid statements about search results is hard as researchers often lack the appropriate tools.



Norocel, O. C. (2021, June 7). Imbalanced Agendas—Search engines still in the shadows in Sweden. *In Search of Search (& Its Engines)*. <https://medium.com/in-search-of-search/imbalanced-agendas-search-engines-still-in-the-shadows-in-sweden-9df21086dc55>

SOFTWARE FOR SEARCH ENGINE RESEARCH

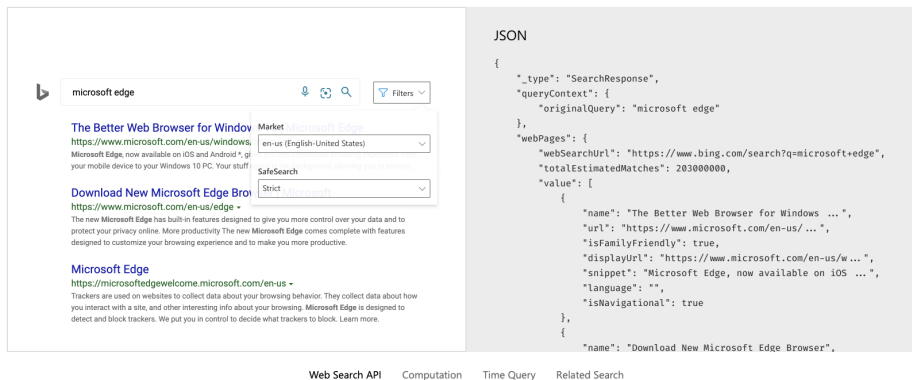
Bing Web Search API

Bring intelligent search to your apps and harness the ability to comb billions of webpages, images, videos, and news with a single API call.

TRY NOW >

Get enhanced search details from billions of web documents

Enable safe, ad-free, location-aware search results, surfacing relevant information from billions of web documents.



The screenshot shows a Bing search for "microsoft edge". The search results include links to "The Better Web Browser for Windows" and "Download New Microsoft Edge Browser". A JSON response is overlaid on the right side of the page, showing search details for the first result.

```
JSON
{
  "_type": "SearchResponse",
  "queryContext": {
    "originalQuery": "microsoft edge"
  },
  "webPages": {
    "webSearchUrl": "https://www.bing.com/search?q=microsoft+edge",
    "totalEstimatedMatches": 203000000,
    "value": [
      {
        "name": "The Better Web Browser for Windows ...",
        "url": "https://www.microsoft.com/en-us/...",
        "isFamilyFriendly": true,
        "displayUrl": "https://www.microsoft.com/en-us/w...",
        "snippet": "Microsoft Edge, now available on iOS ...",
        "language": "",
        "isNavigational": true
      },
      {
        "name": "Download New Microsoft Edge Browser",

```



DMI Tools

Media Analysis: Media Monitoring | Mapping | Clouding |
Data Treatment: Data Collection | Data Analysis | Informa
Natively Digital: The Link | The URL | The Tag | The Dorr
Device Centric: Google | Google Images | Google News |



Open for Innovation

KNIME



RELEVANCE ASSESSMENT TOOL

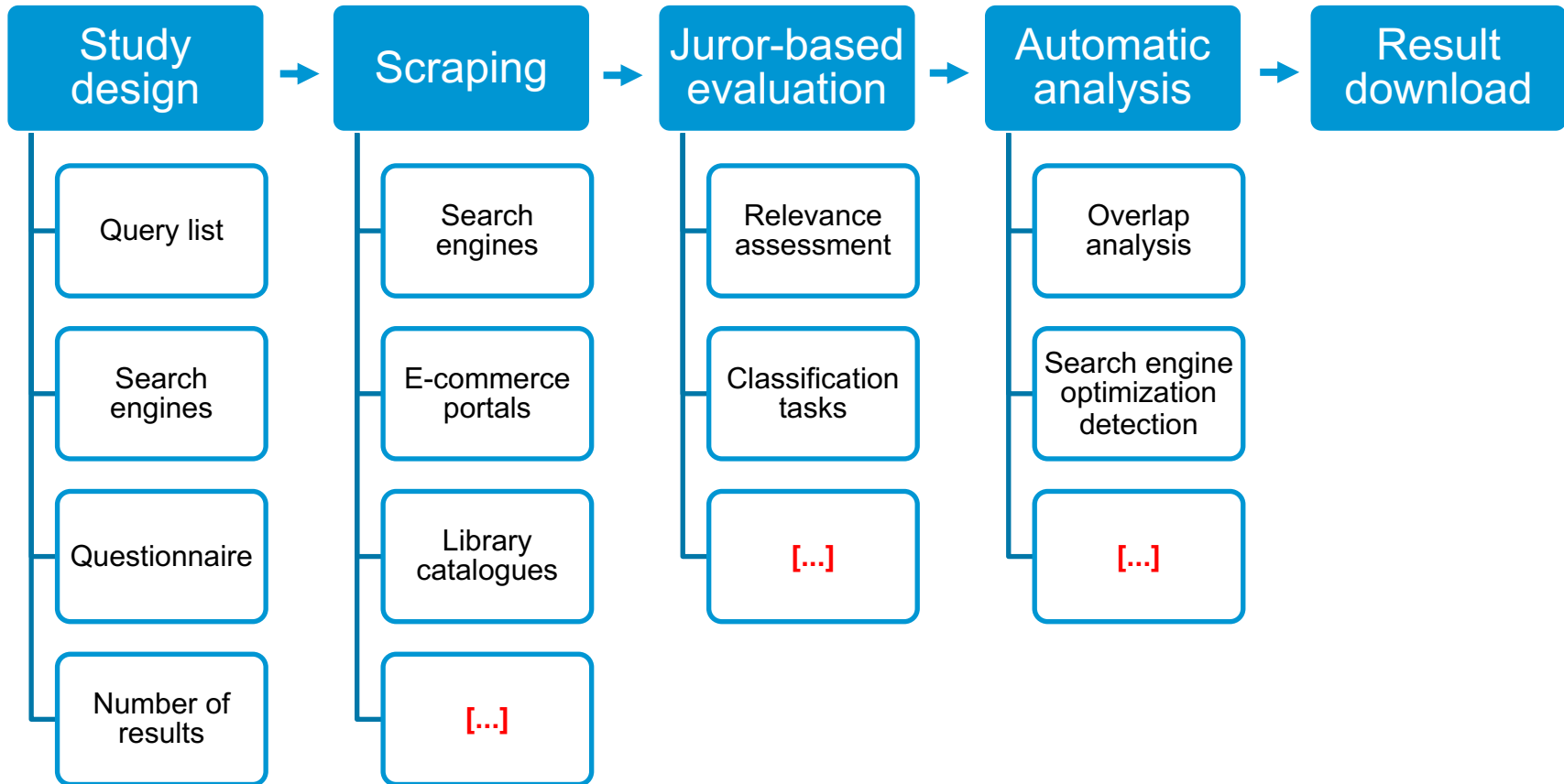
SCREEN SCRAPING

The screenshot shows a Google search for "swedish school of library and information science". The search results page displays the title "About SSLIS - University of Borås" and a snippet of text: "The Swedish School of Library and Information Science (SSLIS) is one of a kind in Sweden. Within Library and Information Science, our research, programmes and courses are the most extensive in the country, and are being developed in close interaction with the global society around us." Below the snippet is the URL "https://www.hb.se/about-the-swedish-school-of-library-a...".

The browser's developer tool is open, showing the HTML source code for the search results. The code includes a Facebook link: ``. The developer tool interface includes a left sidebar with "Breakpoints" and "search" sections, and a main area displaying the HTML code with line numbers from 4674 to 4690.

3 Research software

RAT SOFTWARE MODULES

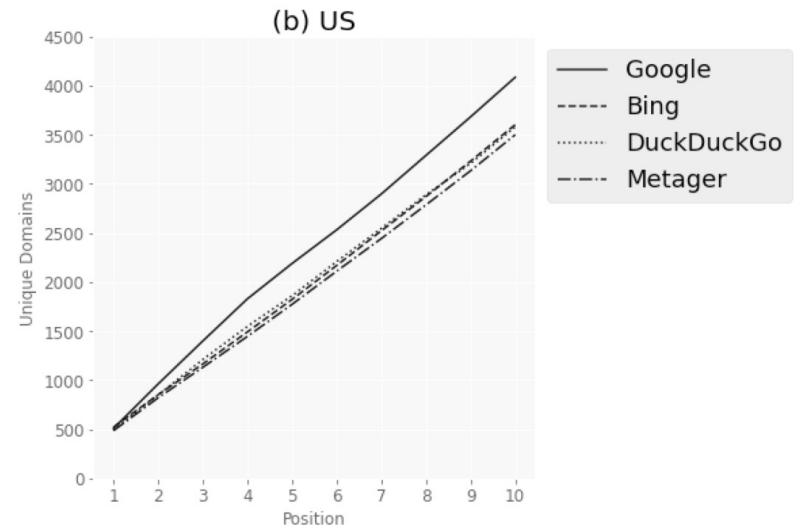


4 Some examples of data science studies from the Search Studies research group

SEARCH ENGINE OVERLAP

Do search engines produce different (top) results? Is it worthwhile using another search engine than Google?

- Comparison between four search engines
- 1,672 queries and 66,880 results for Germany and 1,865 queries and 74,600 results for the US
- Automatic analyses: Number of unique domains, most popular domains per search engine, top domains per search engine that are not a top domain in the other engines investigated, source distribution (Gini)

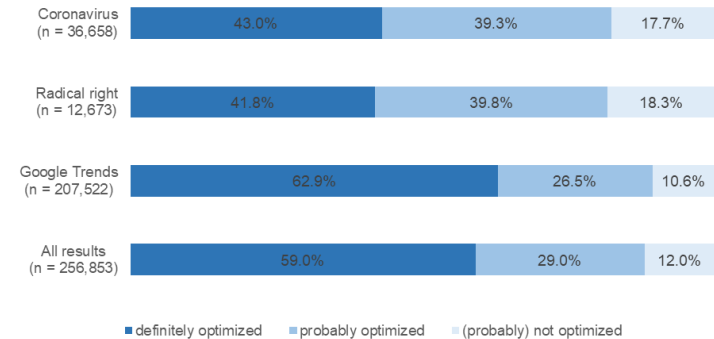


Yagci, N., Sünkler, S., Häußler, H., & Lewandowski, D. (2022). A Comparison of Source Distribution and Result Overlap in Web Search Engines. *Proceedings of the 85th Annual Meeting of the Association for Information Science & Technology*, 343–353.

SEARCH ENGINE OPTIMIZATION

How do external actors influence Google's results?

- Three datasets: Google Trends, radical right queries, Corona queries
- 1,914 queries and 256,853 results
- Detection of search engine optimization based on 21 indicators
- Shows that the majority of results in top positions is optimized; also holds true for lower-ranked results
- In health information seeking, optimized sites outrank government websites, which are predominantly not optimized (Schultheiß, Häußler & Lewandowski, 2022)



Lewandowski, D., Sünkler, S., & Yagci, N. (2021). The influence of search engine optimization on Google's results. *13th ACM Web Science Conference 2021*, 12–20. <https://doi.org/10.1145/3447535.3462479>

Schultheiß, S., Häußler, H., & Lewandowski, D. (2022). Does Search Engine Optimization come along with high-quality content? *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 123–134. <https://doi.org/10.1145/3498366.3505811>



MENY



THE SWEDISH SCHOOL OF LIBRARY
AND INFORMATION SCIENCE
UNIVERSITY OF BORÅS

Elemente Netzwerk Quellen Speicher Konsole

```

3 <!DOCTYPE html>
4 <html lang="en">
5 <head>
6   <meta charset="utf-8">
7   <meta http-equiv="X-UA-Compatible" content="IE=edge">
8   <title>The Swedish School of Library and Information Science - University of Borås</title>
9   <link rel="stylesheet" href="/assets/build/css/main.css?v=1.6.6.0">
10
11
12   <link rel="apple-touch-icon" sizes="180x180" href="/assets/img/favicons/apple-touch-icon.png">
13   <link rel="icon" type="image/png" sizes="32x32" href="/assets/img/favicons/favicon-32x32.png">
14   <link rel="icon" type="image/png" sizes="16x16" href="/assets/img/favicons/favicon-16x16.png">
15   <link rel="manifest" href="/assets/img/favicons/site.webmanifest">
16   <link rel="mask-icon" href="/assets/img/favicons/safari-pinned-tab.svg">
17
18   <link rel="stylesheet" href="https://use.typekit.net/ck12ezo.css">
19
20   <meta name="msapplication-TileColor" content="#2f2f2f">
21   <meta name="theme-color" content="#ffffff">
22
23   <meta name="viewport" content="width=device-width, initial-scale=1">
24   <meta name="description" content="The Swedish School of Library an Information Science (SSLIS) is one of a kind in Sweden. Within Library and Information Science, our research, prog...">
25
26   <meta property="og:title" content="The Swedish School of Library and Information Science" />
27   <meta property="og:type" content="website" />
28   <meta property="og:url" content="https://www.hb.se/en/the-swedish-school-of-library-and-information-science-sslis/" />
29   <meta property="og:image:height" content="2000" />
30   <meta property="og:image:width" content="2000" />
31
32
33   <meta property="og:site_name" content="Högskolan i Borås">
34   <meta property="og:description" content="The Swedish School of Library an Information Science (SSLIS) is one of a kind in Sweden. Within Library and Information Science, our research, prog..." />
35
36   <meta name="twitter:card" content="summary">
37   <meta name="twitter:site" content="@hogskolaniboras">
38   <meta name="twitter:creator" content="@hogskolaniboras">
39   <meta name="twitter:description" content="The Swedish School of Library an Information Science (SSLIS) is one of a kind in Sweden. Within Library and Information Science, our research, prog...">
40   <meta name="twitter:title" content="The Swedish School of Library and Information Science">
41
42

```

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: ✓

Home | Biographie | Veröffentlichungen |
WS 2014/15 | Veranstaltungsarchiv | Lehrmaterialien nach Themen | Bachelor-/Masterarbeiten |
Projekte | Auslandspraktikum | Studienberatung / BAföG-Dozentin |



Prof. Dr. Ulrike Spree

Lehrgebiete:

Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Design Medien Information
- Department Information -
Mediencampus Finkenau
Finkenau 35
22081 Hamburg

Informationsdienstleistung und Informationsvermittlung
Wissensorganisation und Records Management
Information Research

Telefon: 040/42875/3607
Mail: ulrike dot spree at haw-hamburg dot de

Sprechstunde:
mittwochs 15:00 bis 16:30
Am 03.12.2014 muss die Sprechstunde leider ausfallen, da ich dienstlich unterwegs bin.

Aktuelle Informationen:
Kurse in Moodle
Themenvorschläge Bachelor- und Masterarbeiten
Infos zum BAföG-Antrag

Aktuelle Veröffentlichungen und Vorträge:

- Alexander, Fran; Spree, Ulrike (Hgg.): *Aslib Journal of Information Management. Special Issue: Semantic Search*. Vol. 66, 2014, Nr. 5 (Beginn ab S.473)
<http://www.emeraldinsight.com/toc/ajim/66/5>
- Spree, Ulrike: *How Readers Shape the Content of an Encyclopedia: A Case Study Comparing the German Meyers Konversationslexikon (1885-1890) with Wikipedia (2002-2013)*. In: *Culture Unbound : Journal of Current Cultural Research*; Vol. 6, Article 29, 2014
<http://www.cultureunbound.ep.liu.se/v6/a29/cu14v6a29.pdf>
- Spree, Ulrike: *Kulturelle Teilhabe ermöglichen : Möglichkeiten und Grenzen zielgruppenorientierter Bibliotheksarbeit für Menschen mit Demenz*. In: *Medien & Altern* 2013; 3. S. 77-93
[preprint]
- Spree, Ulrike: *Wörterbücher und Enzyklopädien*. In: *Grundlagen der praktischen Information und Dokumentation* / hrsg. von Kühlen, Rainer. 6., völlig neu gefasste Ausgabe. - Berlin, Boston, Mass. : De Gruyter Saur, 2012, S. 550-559. [preprint]
- Ulrike Spree: *Mutually dependent : Encyclopedias and their audiences*. Vortrag gehalten am 09.11.2012 auf der Konferenz "Encyclopedias and beyond" in der National Library Oslo
- Ulrike Spree; Jutta Lindenthal; Andje Knaack: *Wortnetz Kultur - ein Thesaurusprojekt zur kollaborativen Erschließung von Fachinformationen des kulturellen Erbes*. In: *Information, Wissenschaft & Praxis* 2012; 63(1). S.23-36. [preprint]

... weitere Publikationen

Mein Motto:

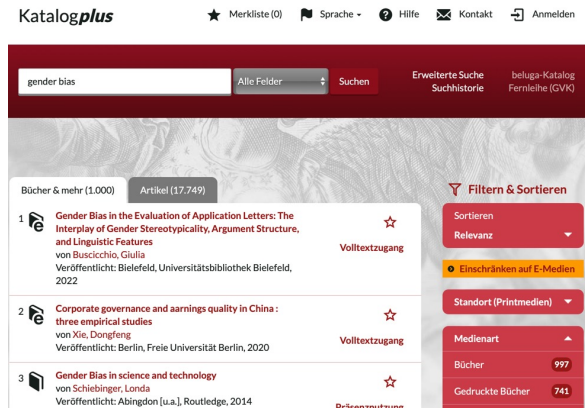
When action grows unprofitable, gather information; when information grows unprofitable, sleep.

(Demo tool available at <http://5.189.155.20:5000>)

21

BIAS IN LIBRARY DISCOVERY SYSTEMS

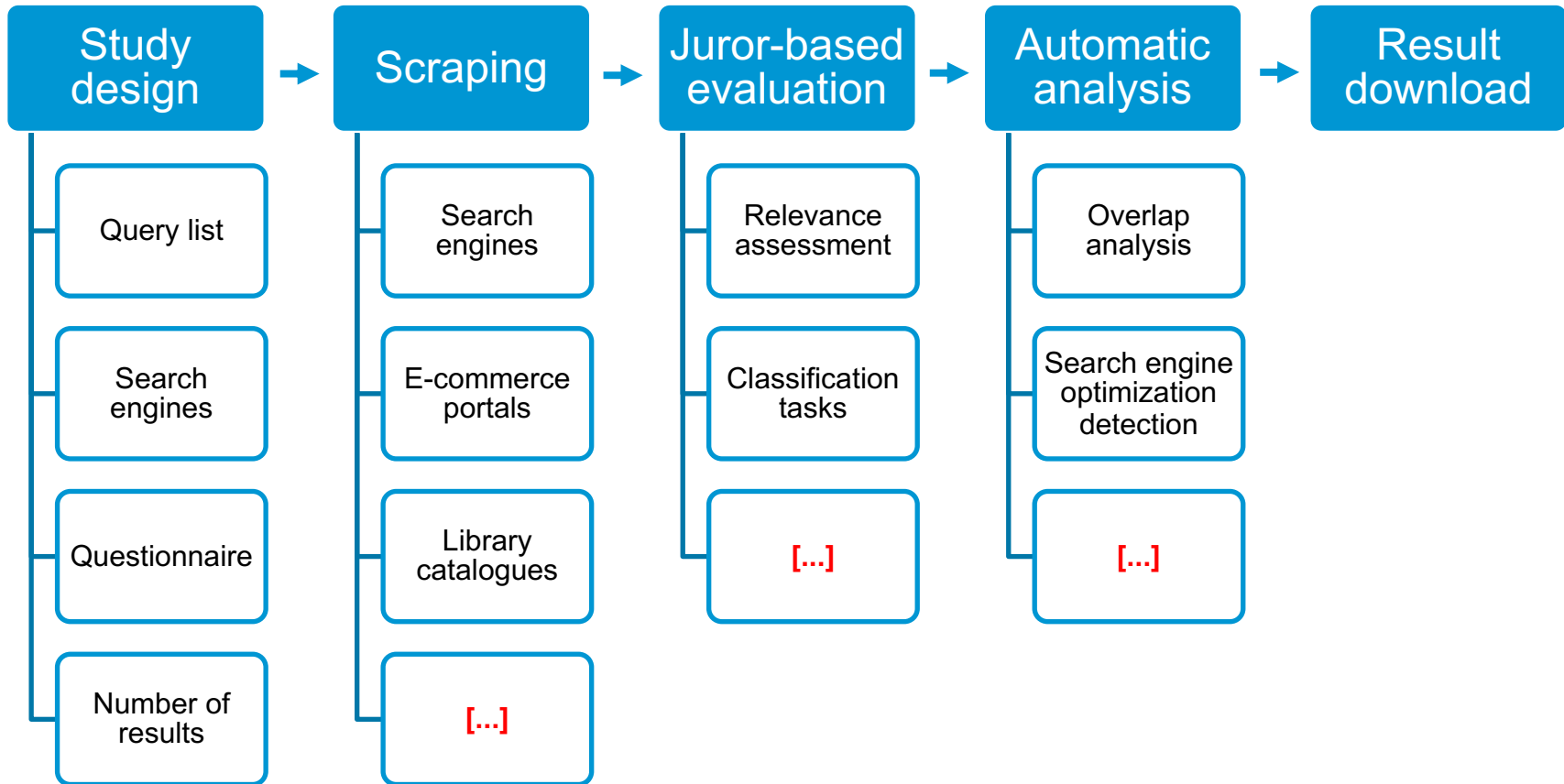
Are results equally represented in the library discovery system of the Hamburg University Library?



- Hypothesis: In an equal representation, there should be no differences over result ranks regarding formal criteria of the results. (There still may be some intended difference, as implemented in the ranking algorithms.)
- Method: Comparing a set of top results with a random set of lower-ranked results
- 5,948 results
- Results: Over-representation of current literature, over-representation of female authors in top results, under-representation of German-language material

Student work by Inga Albrecht, Daniel Klein, Paulina Triesch & Torge Plückhahn (2022)

RAT SOFTWARE MODULES



5 Summary and conclusion

SUMMARY AND CONCLUSION

- With data science approaches, we can **collect and analyse large amounts of data**. Data science approaches can complement information science studies by scaling them.
- A modular software like **RAT allows for many different research designs** and makes it easy for researchers to design and conduct their own studies.
- As part of the RAT project, we offer **support for researchers** doing their own studies. We are actively looking for researchers doing their studies with RAT and will support you doing your studies.
- A search result study might be a good **addition** to your research
- **More to come** – A demo tool will be available from early October, more modules to be added continuously. Tutorials, how-to's, and a knowledge base will be available soon.
- More information on the RAT software: <https://searchstudies.org/research/rat/>

THANK YOU

Dirk Lewandowski

Hamburg University of Applied Sciences, Hamburg, Germany

dirk.lewandowski@haw-hamburg.de

www.searchstudies.org/dirk

SEARCH
STUDIES

Funded by

DFG

Deutsche
Forschungsgemeinschaft
German Research Foundation



HAW
HAMBURG