Mega databases, AI and semantic searches: magic buzz words or something real there?

Information Retrieval Meeting (IRM 2024), Cologne, Germany
26 April 2024

Aaron Tay
Head, Data Services
Singapore Management University Libraries

Aarontay@gmail.com
@aarontay
https://musingsaboutlibrarianship.blogspot.com

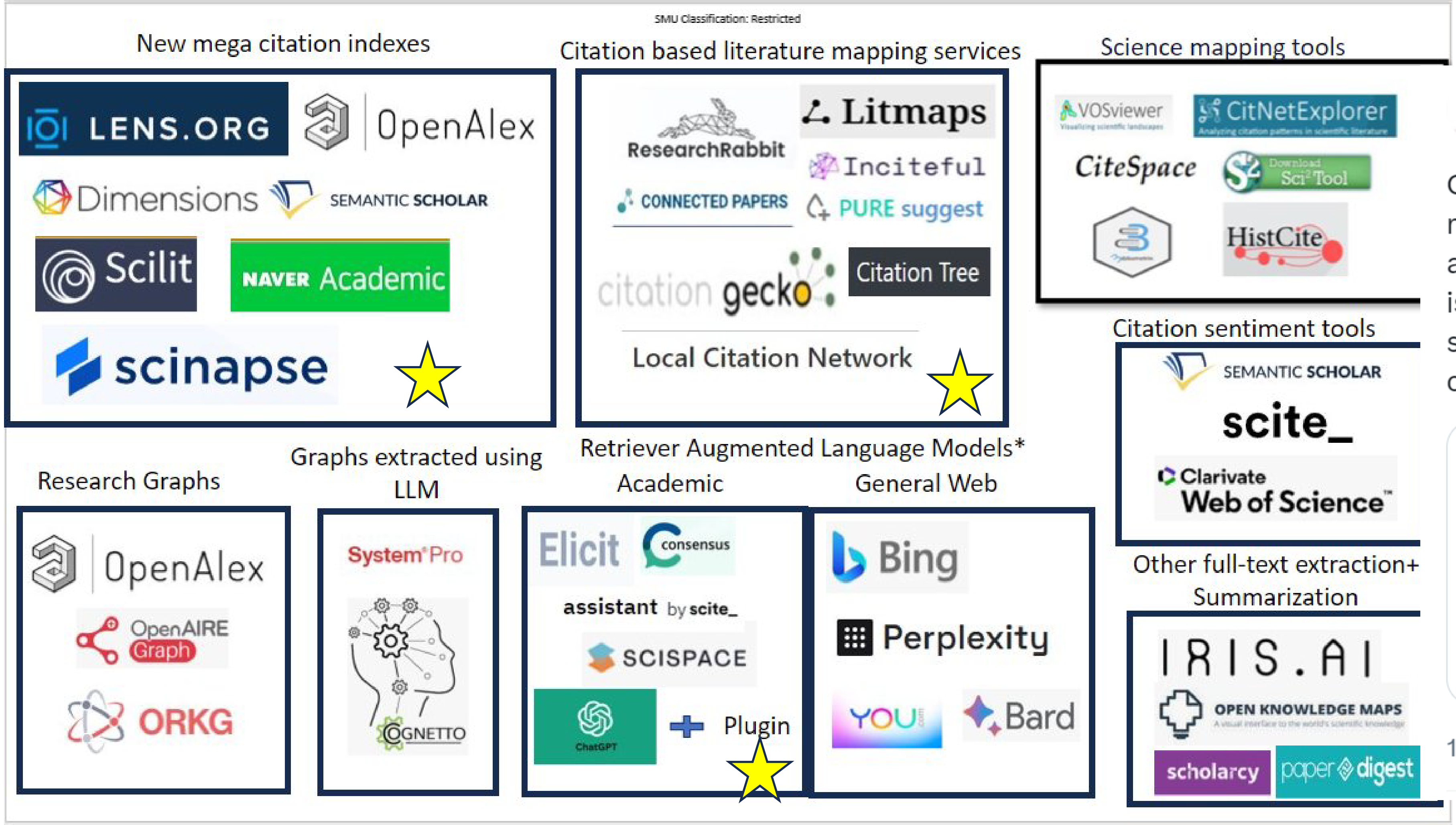
3 Predictions for 2027

1. Semantic Search will become so effective, evidence synthesis will adapt to use it despite it being less transparent.

2. Data Extraction with LLMs will be popular. Either used as a second option check or used alone with human in loop checks

3. Agent LLM based retrieval will start being used but will still be minority

An attempt to classify new discovery tools-June 2023



SMU Classification: Restricted

New mega citation indexes
- LENS.ORG
- OpenAlex
- Dimensions
- SEMANTIC SCHOLAR
- Scilit
- NAVER Academic
- scinapse

Citation based literature mapping services
- ResearchRabbit
- Litmaps
- Inciteful
- CONNECTED PAPERS
- PURE suggest
- citation gecko
- Citation Tree
- Local Citation Network

Science mapping tools
- VOSviewer
- CitNetExplorer
- CiteSpace
- Download Sci2 Tool
- HistCite

Citation sentiment tools
- SEMANTIC SCHOLAR
- scite_
- Clarivate Web of Science™

Research Graphs
- OpenAlex
- OpenAIRE Graph
- ORKG

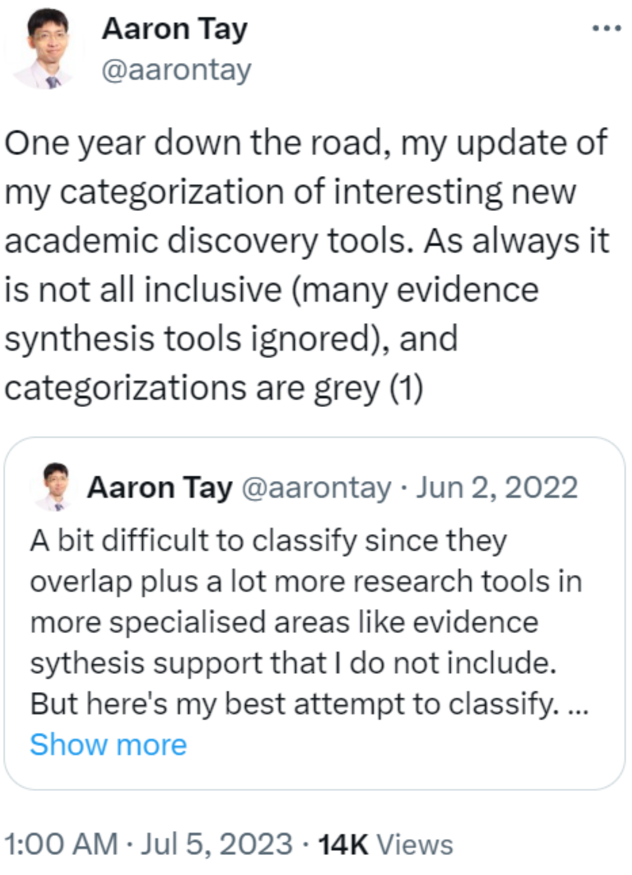
Graphs extracted using LLM
- System Pro
- COGNETTO

Retriever Augmented Language Models*

Academic
- Elicit
- consensus
- assistant by scite_
- SCISPACE
- ChatGPT + Plugin

General Web
- Bing
- Perplexity
- YOU
- Bard

Other full-text extraction+ Summarization
- IRIS.AI
- OPEN KNOWLEDGE MAPS
- scholarcy
- paper digest



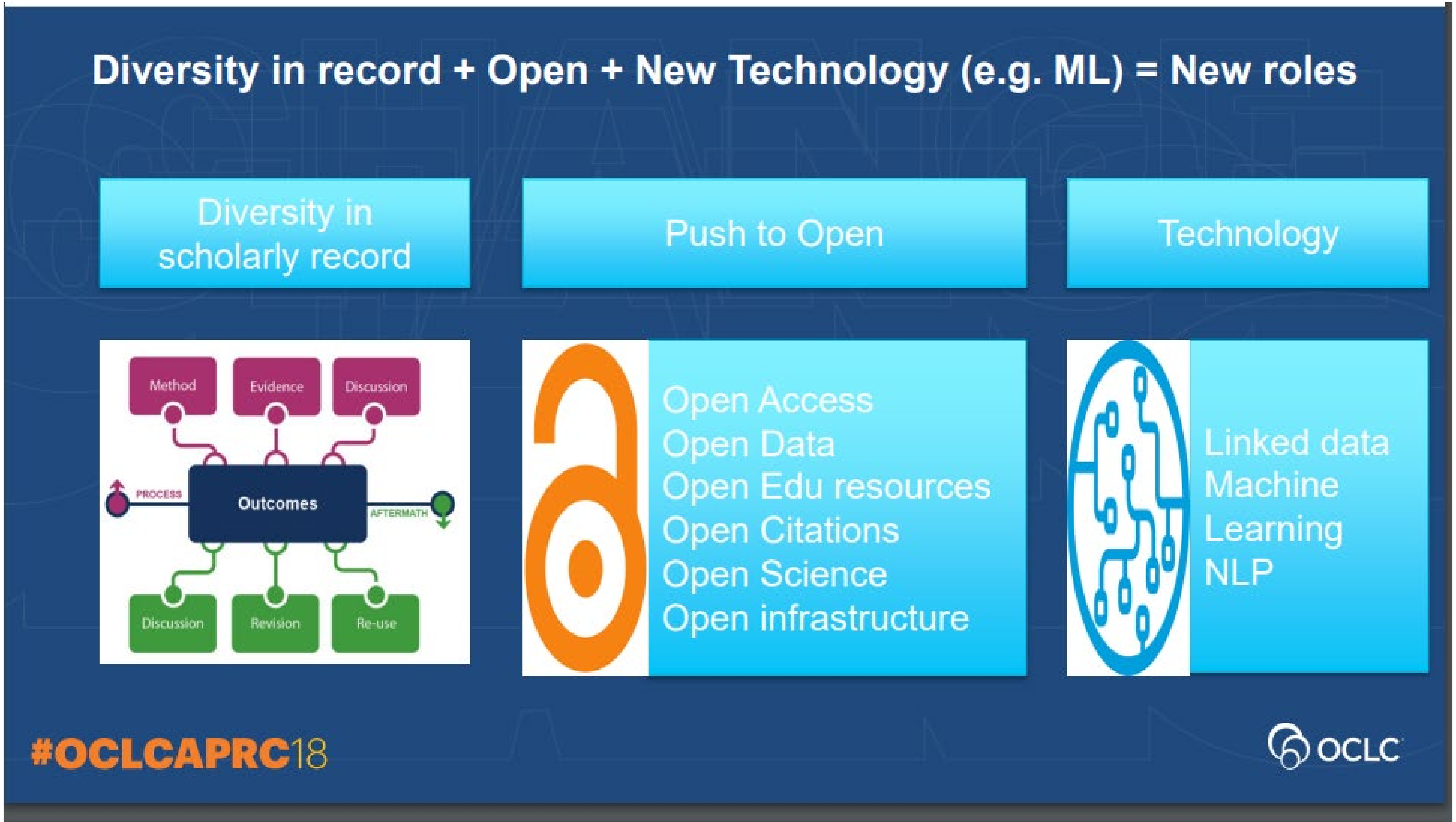
Aaron Tay
@aarontay

One year down the road, my update of my categorization of interesting new academic discovery tools. As always it is not all inclusive (many evidence synthesis tools ignored), and categorizations are grey (1)

Aaron Tay @aarontay · Jun 2, 2022
A bit difficult to classify since they overlap plus a lot more research tools in more specialised areas like evidence sythesis support that I do not include. But here's my best attempt to classify. ...
Show more

1:00 AM · Jul 5, 2023 · **14K** Views

My 2018 prediction about game changers



Diversity in record + Open + New Technology (e.g. ML) = New roles

Diversity in scholarly record

Method
Evidence
Discussion
PROCESS
Outcomes
AFTERMATH
Discussion
Revision
Re-use

Push to Open

- Open Access
- Open Data
- Open Edu resources
- Open Citations
- Open Science
- Open infrastructure

Technology

- Linked data
- Machine Learning
- NLP

#OCLCAPRC18

OCLC

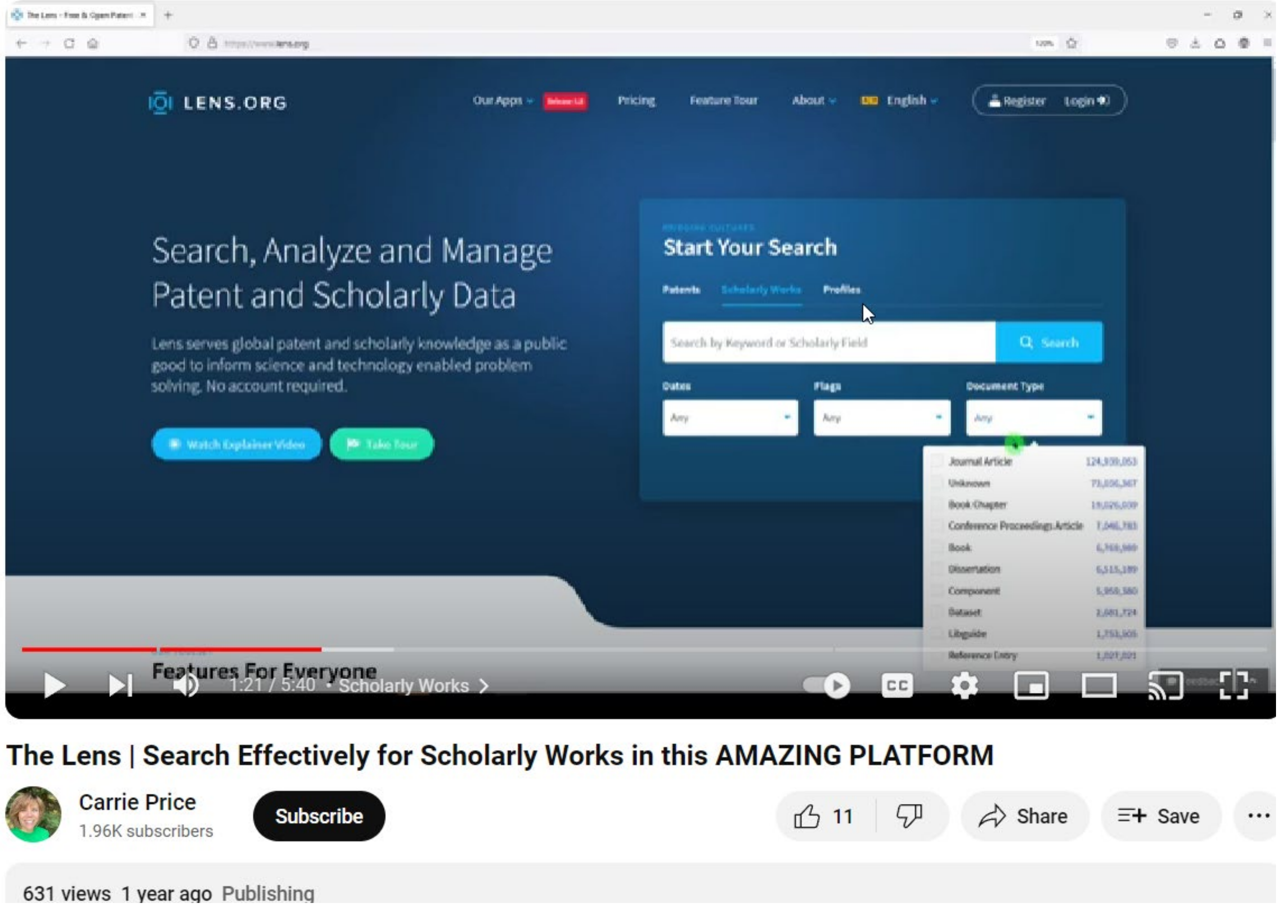
Major milestones in academic search

- **1996** - PubMed
- **2004** – Google Scholar, Scopus launches
- **2009** – Summon discovery service/layer launches
- **2015** – **Semantic Scholar launches – later also Semantic Scholar Open Corpus (S2ORC) and Semantic Scholar Academic Graph (S2AG)**
- **2016** – **Microsoft Academic and Microsoft academic graph (MAG) relaunches**
- **2017** – **Transformer architecture introduced in "Attention is all you need" paper**
- **2017** – **Initiative for Open Citations (I4OC)**
- **2018** – **Lens.org Scholarly work search launches combining MAG+Crossref+Pubmed**
- **2018** – **BERT model introduced**
- **2019** – **Elsevier , ACS make citations open**
- **2020 – Connected papers, ResearchRabbit etc launches**
- **2020 – GPT3**
- **2022 – Crossref makes deposited references mandatory open**
- **2022 – OpenAlex launches, MAG closes**
- **2022 – ChatGPT/GPT3.5 launches, Perplexity.ai, Elicit, Bing Chat launches with RAG**

Push to Open
New technology
New resulting discovery tools

Rise of the "mega-indexes"

Evidence synthesis librarians talk about lens.org



Search, Analyze and Manage Patent and Scholarly Data

Lens serves global patent and scholarly knowledge as a public good to inform science and technology enabled problem solving. No account required.

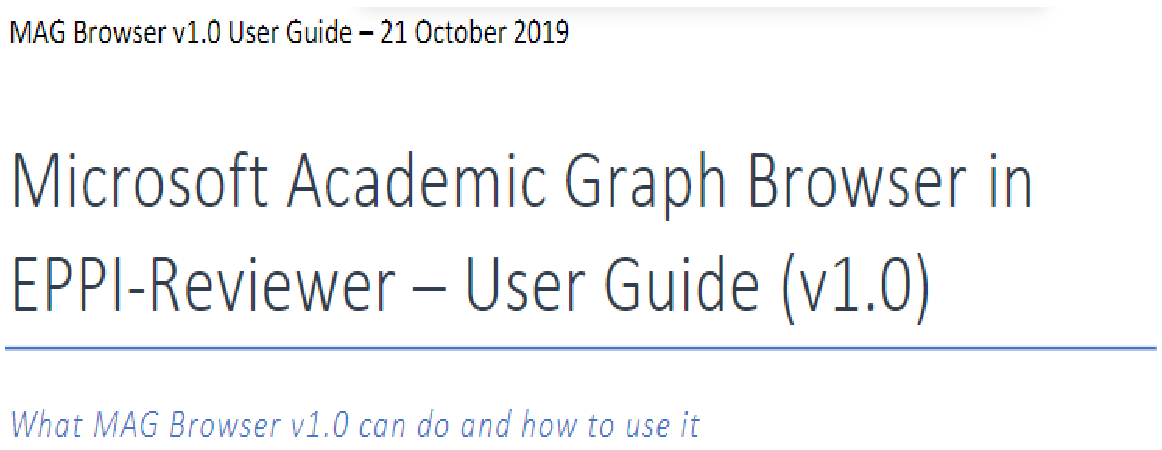
The Lens | Search Effectively for Scholarly Works in this AMAZING PLATFORM

Carrie Price
1.96K subscribers

631 views 1 year ago Publishing

https://www.youtube.com/watch?v=2DDgqPp-hVQ

EPPI reviewer – early pioneer on use of Microsoft Academic graph as far back as 2019

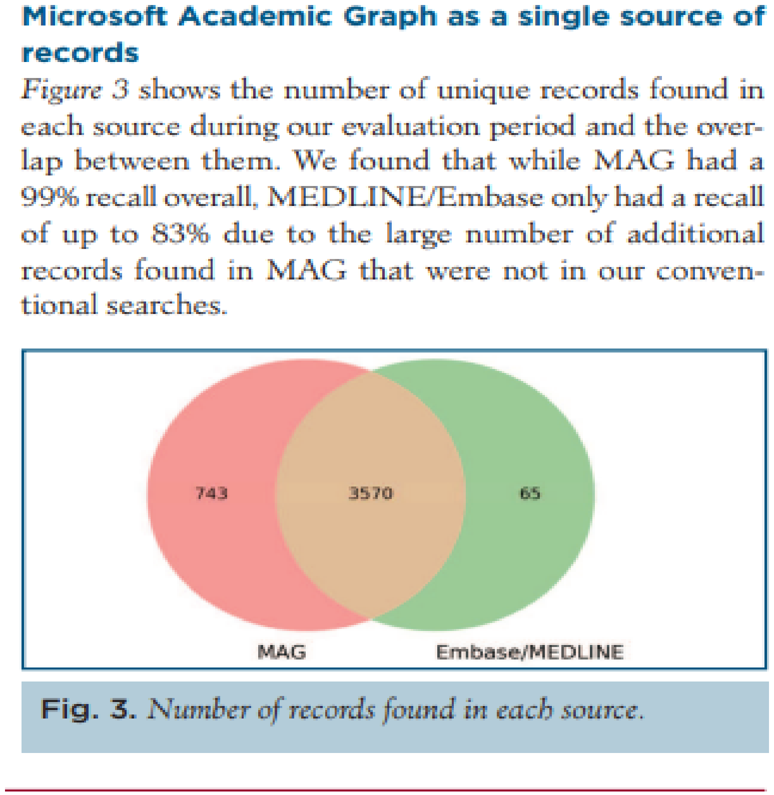


MAG Browser v1.0 User Guide – 21 October 2019

Microsoft Academic Graph Browser in EPPI-Reviewer – User Guide (v1.0)

What MAG Browser v1.0 can do and how to use it

[Finding new reports]



Microsoft Academic Graph as a single source of records

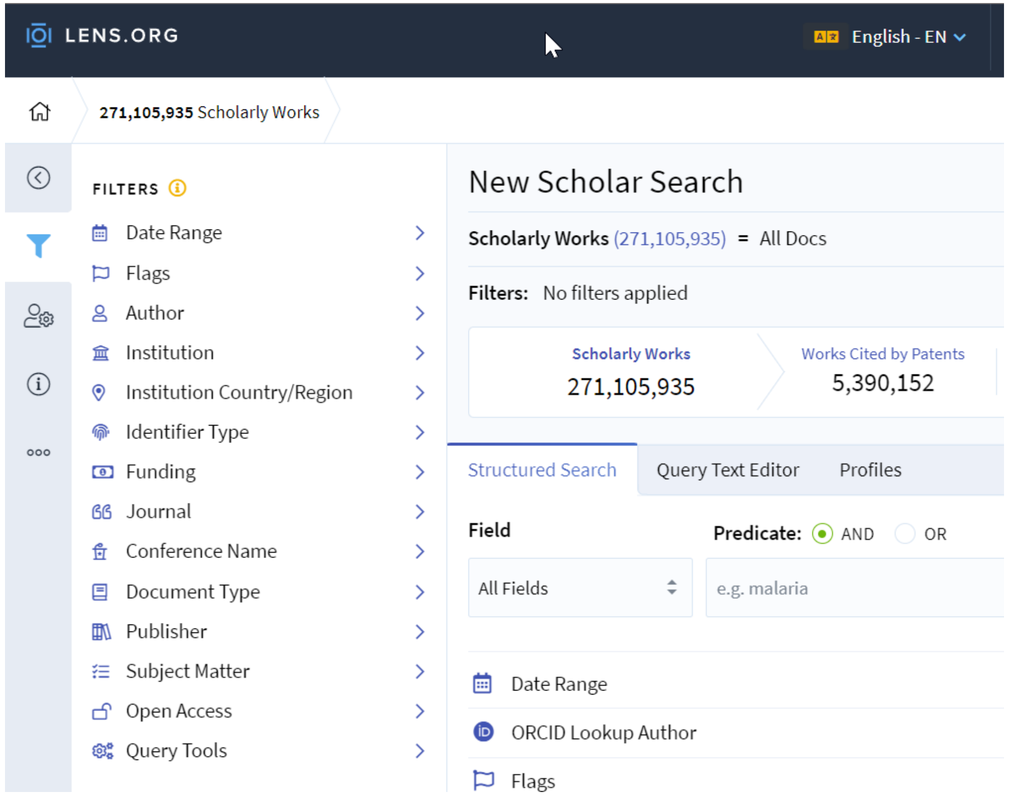
Figure 3 shows the number of unique records found in each source during our evaluation period and the overlap between them. We found that while MAG had a 99% recall overall, MEDLINE/Embase only had a recall of up to 83% due to the large number of additional records found in MAG that were not in our conventional searches.

Fig. 3. *Number of records found in each source.*

Using automation to produce a "living map" of the COVID-19 research literature , Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research

Why use Lens.org/ MAG (now Openalex)

- Large - >150 million, includes some grey literature, includes non-english literature
- Nice web interfaces that support boolean and field searches and or great analytical features (e.g. Lens.org)
- License is relatively open, public domain, CCO etc
- Allows multiple ways to access e.g. OpenAlex, Semantic Scholar
    - Via web interface
    - API (openalex, Semantic Scholar)
    - Data Dumps (OpenAlex)



271,105,935 Scholarly Works

New Scholar Search

Scholarly Works (271,105,935) = All Docs

Filters: No filters applied

Scholarly Works	Works Cited by Patents
271,105,935	5,390,152

Field

Predicate: ● AND ○ OR

All Fields

e.g. malaria

Why NOT use these mega-indexes

- Not curated – WILL contain predatory journals
- Quality of meta-data – might be lower quality than Scopus, Web of Science

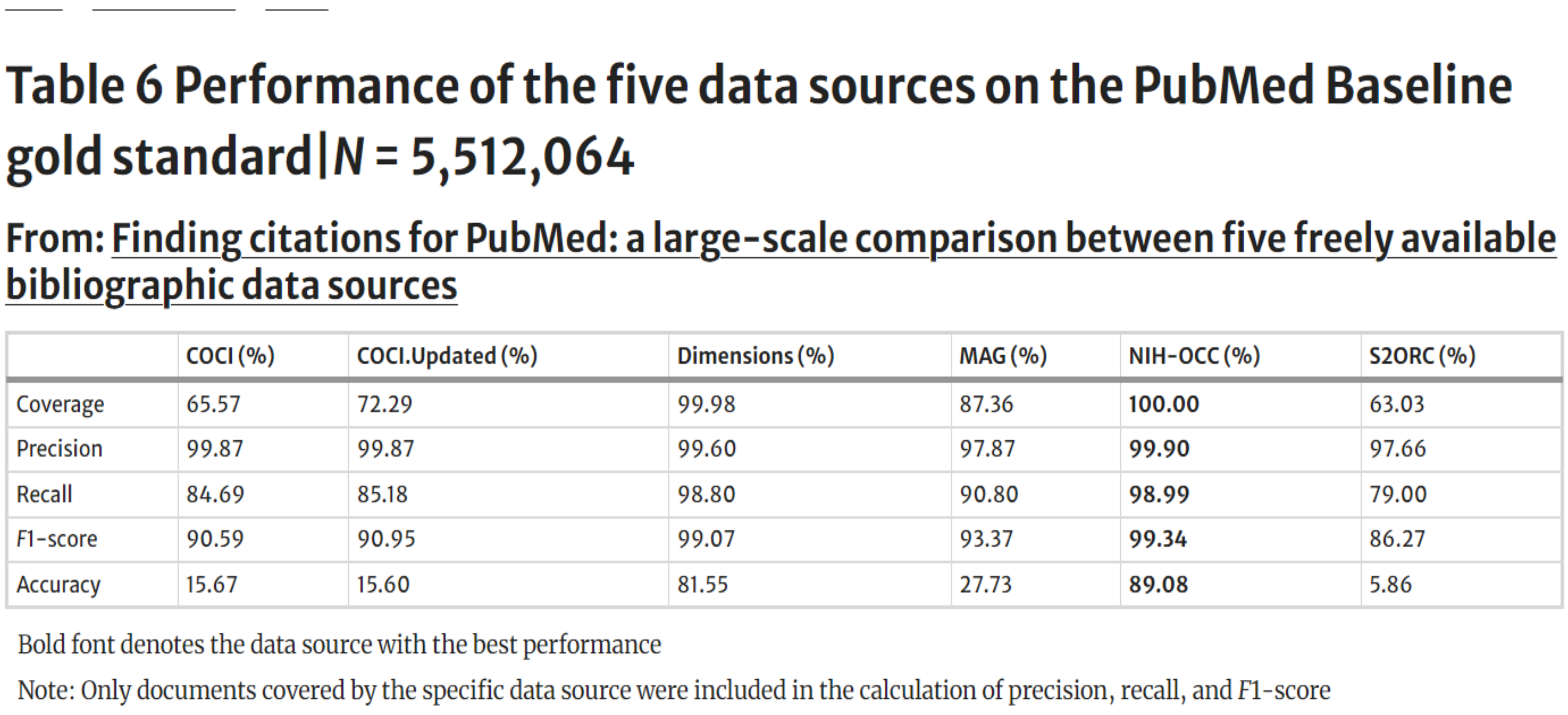


Table 6 Performance of the five data sources on the PubMed Baseline gold standard|*N* = 5,512,064

From: Finding citations for PubMed: a large-scale comparison between five freely available bibliographic data sources

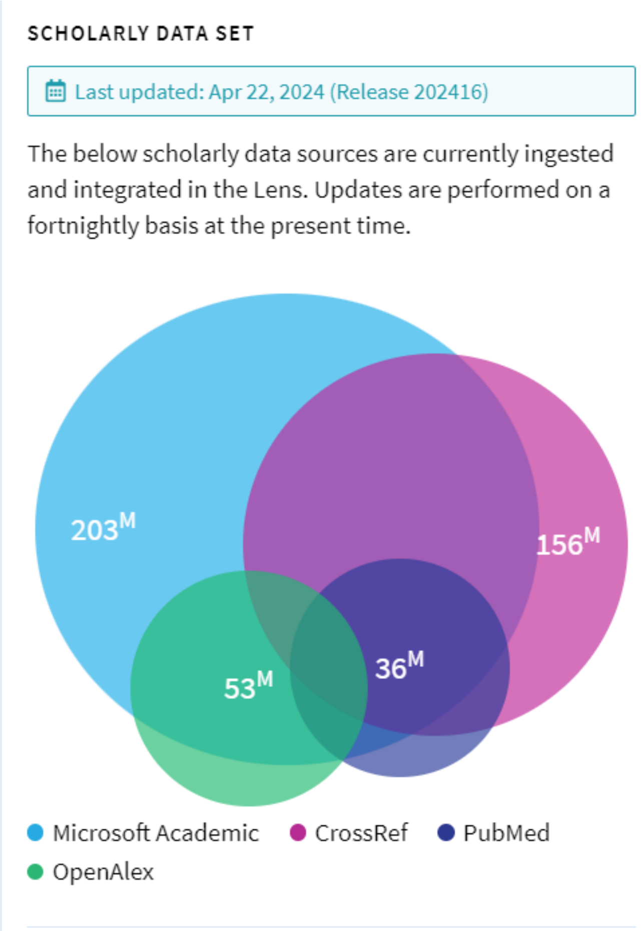
	COCI (%)	COCI.Updated (%)	Dimensions (%)	MAG (%)	NIH-OCC (%)	S2ORC (%)
Coverage	65.57	72.29	99.98	87.36	**100.00**	63.03
Precision	99.87	99.87	99.60	97.87	**99.90**	97.66
Recall	84.69	85.18	98.80	90.80	**98.99**	79.00
*F*1-score	90.59	90.95	99.07	93.37	**99.34**	86.27
Accuracy	15.67	15.60	81.55	27.73	**89.08**	5.86

Bold font denotes the data source with the best performance

Note: Only documents covered by the specific data source were included in the calculation of precision, recall, and *F*1-score

https://link.springer.com/article/10.1007/s11192-021-04191-8/tables/6

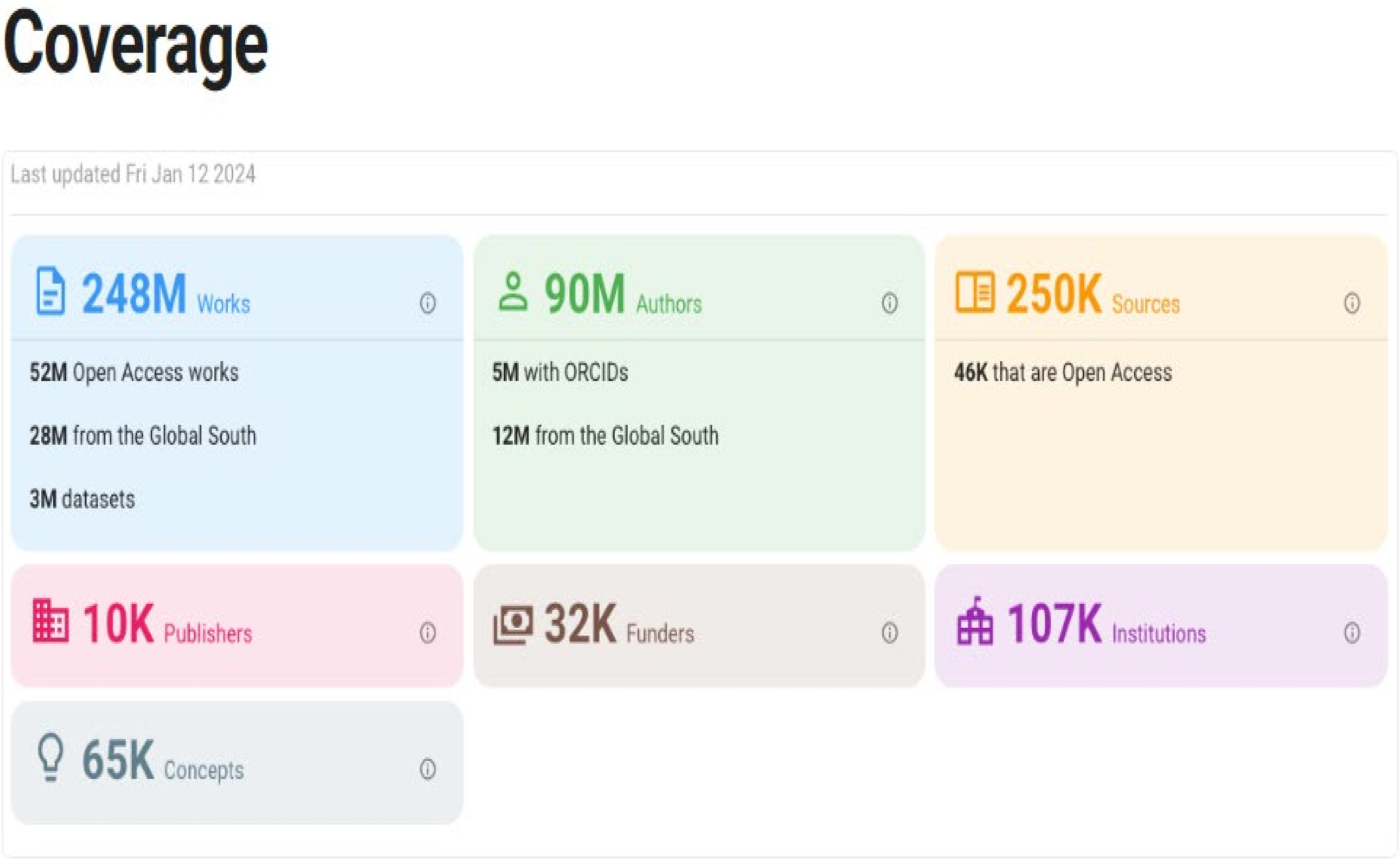
Why "mega"



SCHOLARLY DATA SET

Last updated: Apr 22, 2024 (Release 202416)

The below scholarly data sources are currently ingested and integrated in the Lens. Updates are performed on a fortnightly basis at the present time.



Coverage

Last updated Fri Jan 12 2024

248M Works

52M Open Access works

28M from the Global South

3M datasets

90M Authors

5M with ORCIDs

12M from the Global South

250K Sources

46K that are Open Access

10K Publishers

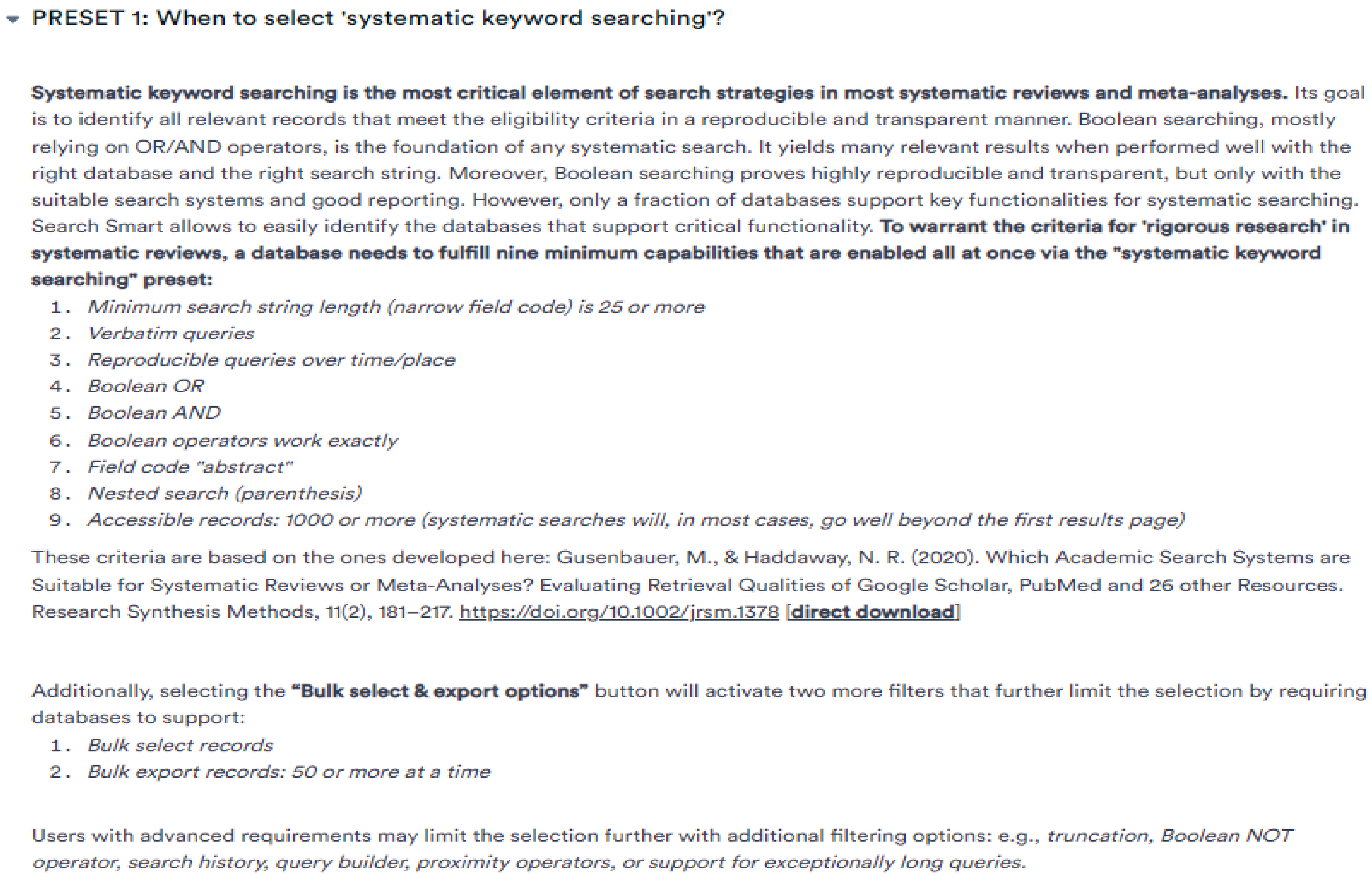
32K Funders

107K Institutions

65K Concepts

[OpenAlex coverage]

[What's in Lens.org]



▾ PRESET 1: When to select 'systematic keyword searching'?

Systematic keyword searching is the most critical element of search strategies in most systematic reviews and meta-analyses. Its goal is to identify all relevant records that meet the eligibility criteria in a reproducible and transparent manner. Boolean searching, mostly relying on OR/AND operators, is the foundation of any systematic search. It yields many relevant results when performed well with the right database and the right search string. Moreover, Boolean searching proves highly reproducible and transparent, but only with the suitable search systems and good reporting. However, only a fraction of databases support key functionalities for systematic searching. Search Smart allows to easily identify the databases that support critical functionality. **To warrant the criteria for 'rigorous research' in systematic reviews, a database needs to fulfill nine minimum capabilities that are enabled all at once via the "systematic keyword searching" preset:**

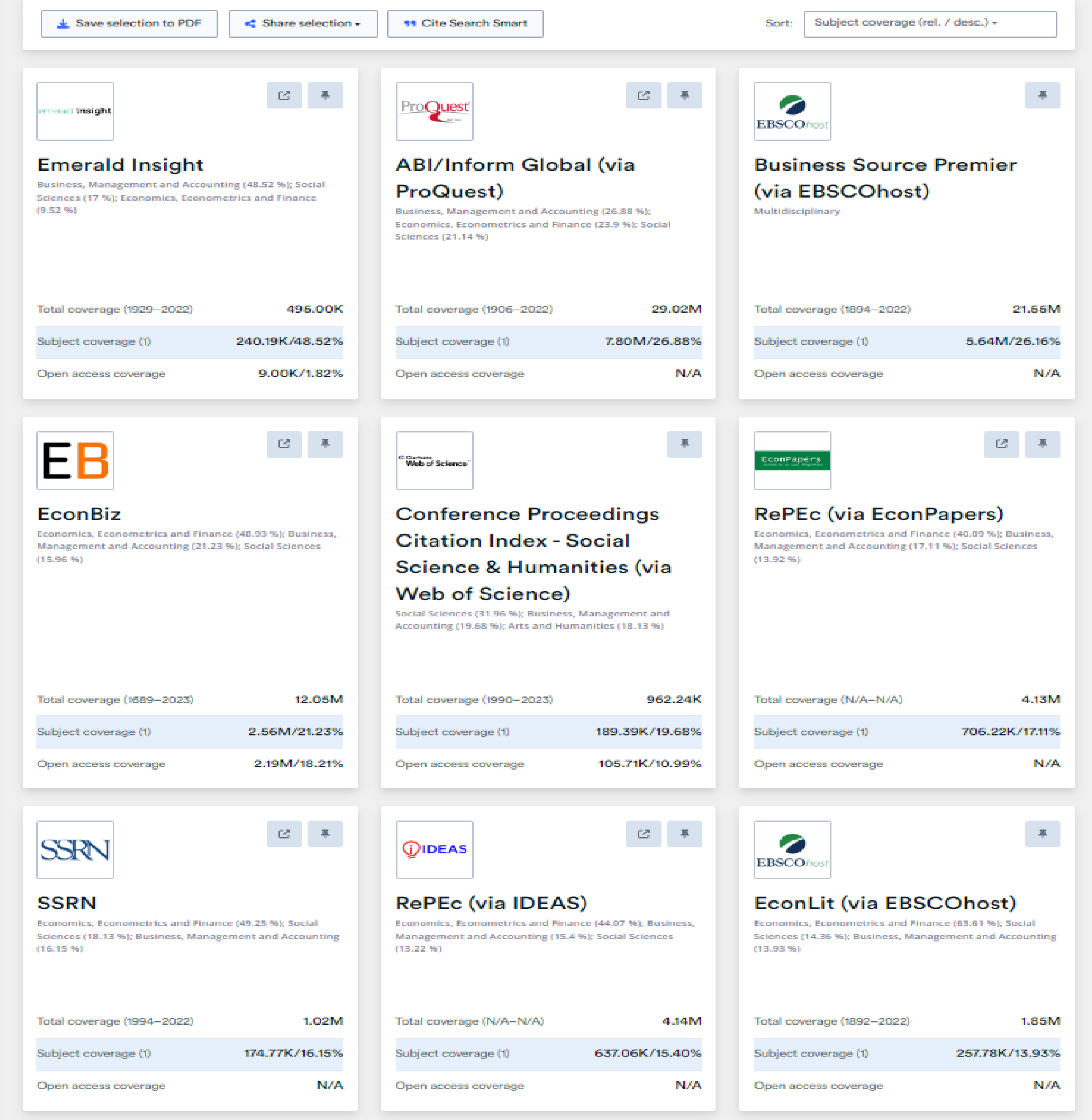
1. *Minimum search string length (narrow field code) is 25 or more*
2. *Verbatim queries*
3. *Reproducible queries over time/place*
4. *Boolean OR*
5. *Boolean AND*
6. *Boolean operators work exactly*
7. *Field code "abstract"*
8. *Nested search (parenthesis)*
9. *Accessible records: 1000 or more (systematic searches will, in most cases, go well beyond the first results page)*

These criteria are based on the ones developed here: Gusenbauer, M., & Haddaway, N. R. (2020). Which Academic Search Systems are Suitable for Systematic Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed and 26 other Resources. Research Synthesis Methods, 11(2), 181–217. https://doi.org/10.1002/jrsm.1378 [**direct download**]

Additionally, selecting the **"Bulk select & export options"** button will activate two more filters that further limit the selection by requiring databases to support:

1. *Bulk select records*
2. *Bulk export records: 50 or more at a time*

Users with advanced requirements may limit the selection further with additional filtering options: e.g., *truncation, Boolean NOT operator, search history, query builder, proximity operators, or support for exceptionally long queries.*



Emerald Insight

Business, Management and Accounting (48.52 %); Social Sciences (17 %); Economics, Econometrics and Finance (9.52 %)

Total coverage (1929–2022)	495.00K
Subject coverage (1)	240.19K/48.52%
Open access coverage	9.00K/1.82%

ABI/Inform Global (via ProQuest)

Business, Management and Accounting (26.88 %); Economics, Econometrics and Finance (23.9 %); Social Sciences (21.14 %)

Total coverage (1906–2022)	29.02M
Subject coverage (1)	7.80M/26.88%
Open access coverage	N/A

Business Source Premier (via EBSCOhost)

Multidisciplinary

Total coverage (1894–2022)	21.55M
Subject coverage (1)	5.64M/26.16%
Open access coverage	N/A

EconBiz

Economics, Econometrics and Finance (48.93 %); Business, Management and Accounting (21.23 %); Social Sciences (15.96 %)

Total coverage (1689–2023)	12.05M
Subject coverage (1)	2.56M/21.23%
Open access coverage	2.19M/18.21%

Conference Proceedings Citation Index - Social Science & Humanities (via Web of Science)

Social Sciences (31.96 %); Business, Management and Accounting (19.68 %); Arts and Humanities (18.13 %)

Total coverage (1990–2023)	962.24K
Subject coverage (1)	189.39K/19.68%
Open access coverage	105.71K/10.99%

RePEc (via EconPapers)

Economics, Econometrics and Finance (40.09 %); Business, Management and Accounting (17.11 %); Social Sciences (13.92 %)

Total coverage (N/A–N/A)	4.13M
Subject coverage (1)	706.22K/17.11%
Open access coverage	N/A

SSRN

Economics, Econometrics and Finance (49.25 %); Social Sciences (18.13 %); Business, Management and Accounting (16.15 %)

Total coverage (1994–2022)	1.02M
Subject coverage (1)	174.77K/16.15%
Open access coverage	N/A

RePEc (via IDEAS)

Economics, Econometrics and Finance (44.07 %); Business, Management and Accounting (15.4 %); Social Sciences (13.22 %)

Total coverage (N/A–N/A)	4.14M
Subject coverage (1)	637.06K/15.40%
Open access coverage	N/A

EconLit (via EBSCOhost)

Economics, Econometrics and Finance (63.61 %); Social Sciences (14.36 %); Business, Management and Accounting (13.93 %)

Total coverage (1892–2022)	1.85M
Subject coverage (1)	257.78K/13.93%
Open access coverage	N/A

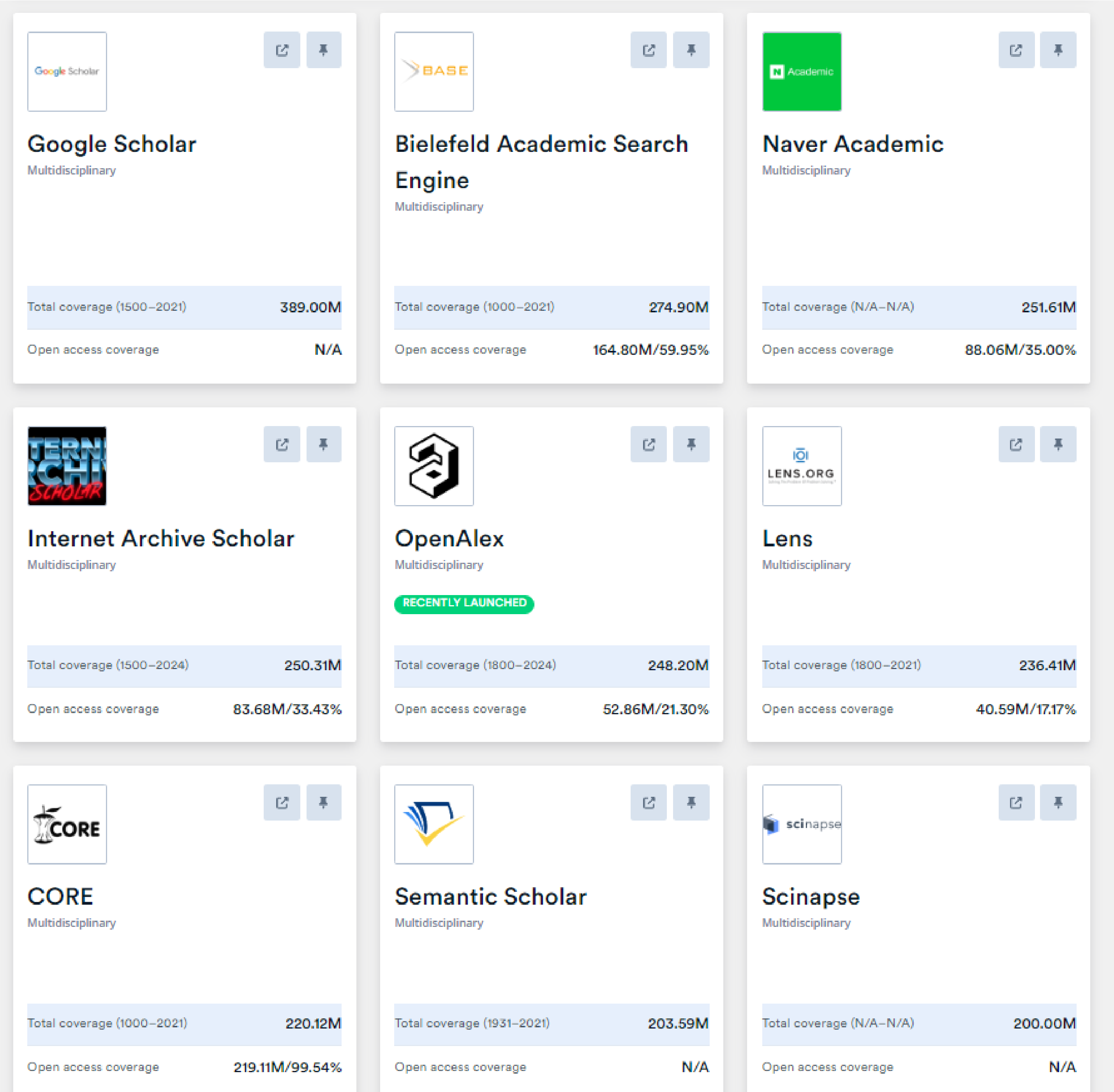
Rigorous methodology to estimate

- Absolute and relative coverage of databases by subject (26 categories)
- Internal and external validity checks, e.g. comparing databases on different platforms
- Allowing sorting by relative subject coverage (example shows Business)

Methodology

- Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. Scientometrics, 1-63. https://doi.org/10.1007/s11192-022-04289-7
- My simplified coverage

Databases with estimated >200 million records according to [Searchsmart.org](https://www.searchsmart.org)



Google Scholar
Multidisciplinary
Total coverage (1500–2021) 389.00M
Open access coverage N/A

Bielefeld Academic Search Engine
Multidisciplinary
Total coverage (1000–2021) 274.90M
Open access coverage 164.80M/59.95%

Naver Academic
Multidisciplinary
Total coverage (N/A–N/A) 251.61M
Open access coverage 88.06M/35.00%

Internet Archive Scholar
Multidisciplinary
Total coverage (1500–2024) 250.31M
Open access coverage 83.68M/33.43%

OpenAlex
Multidisciplinary
RECENTLY LAUNCHED
Total coverage (1800–2024) 248.20M
Open access coverage 52.86M/21.30%

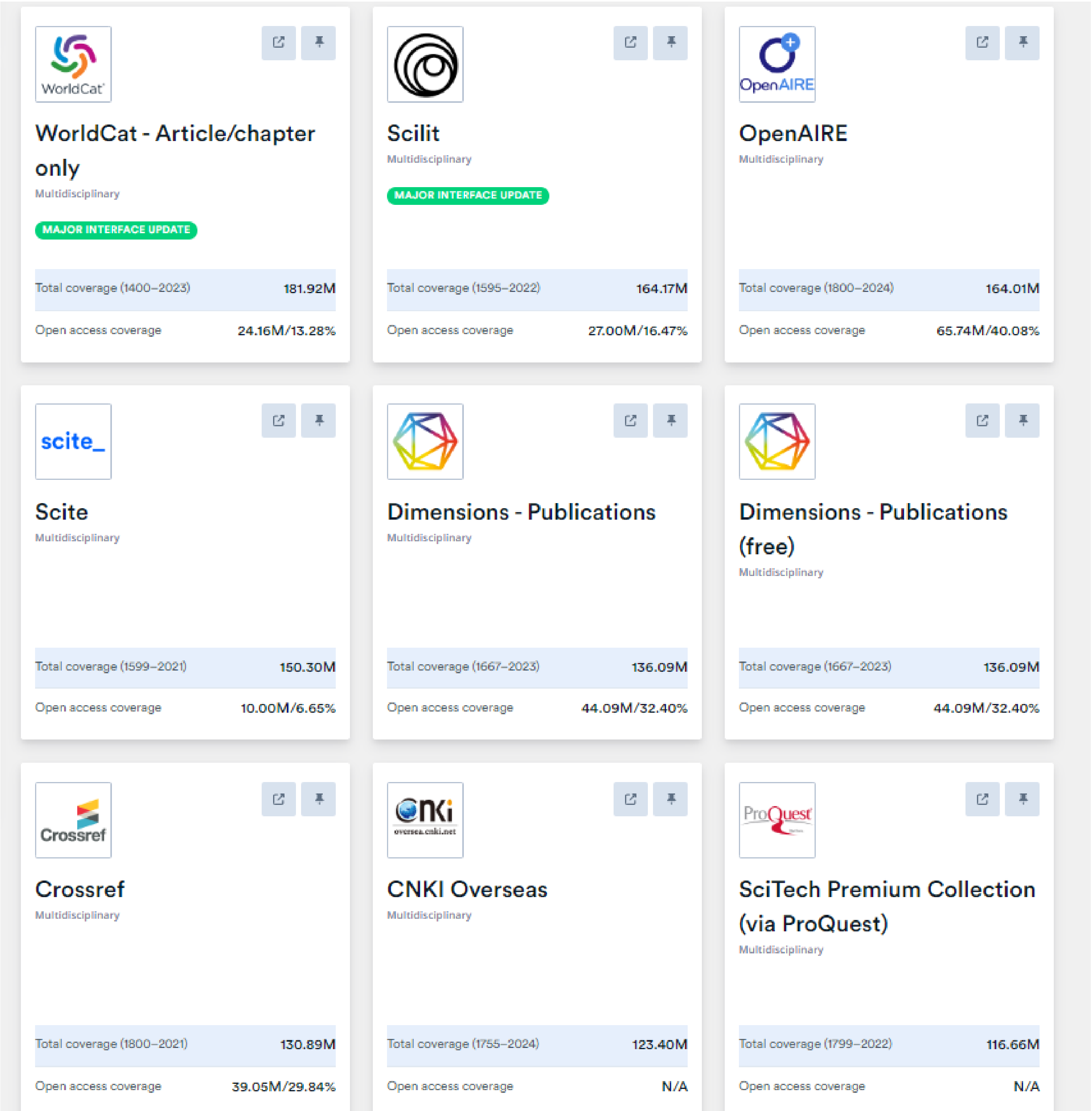
Lens
Multidisciplinary
Total coverage (1800–2021) 236.41M
Open access coverage 40.59M/17.17%

CORE
Multidisciplinary
Total coverage (1000–2021) 220.12M
Open access coverage 219.11M/99.54%

Semantic Scholar
Multidisciplinary
Total coverage (1931–2021) 203.59M
Open access coverage N/A

Scinapse
Multidisciplinary
Total coverage (N/A–N/A) 200.00M
Open access coverage N/A

Databases with estimated 100-200 million records according to Searchsmart.org



WorldCat - Article/chapter only
Multidisciplinary
MAJOR INTERFACE UPDATE
Total coverage (1400–2023) 181.92M
Open access coverage 24.16M/13.28%

Scilit
Multidisciplinary
MAJOR INTERFACE UPDATE
Total coverage (1595–2022) 164.17M
Open access coverage 27.00M/16.47%

OpenAIRE
Multidisciplinary
Total coverage (1800–2024) 164.01M
Open access coverage 65.74M/40.08%

Scite
Multidisciplinary
Total coverage (1599–2021) 150.30M
Open access coverage 10.00M/6.65%

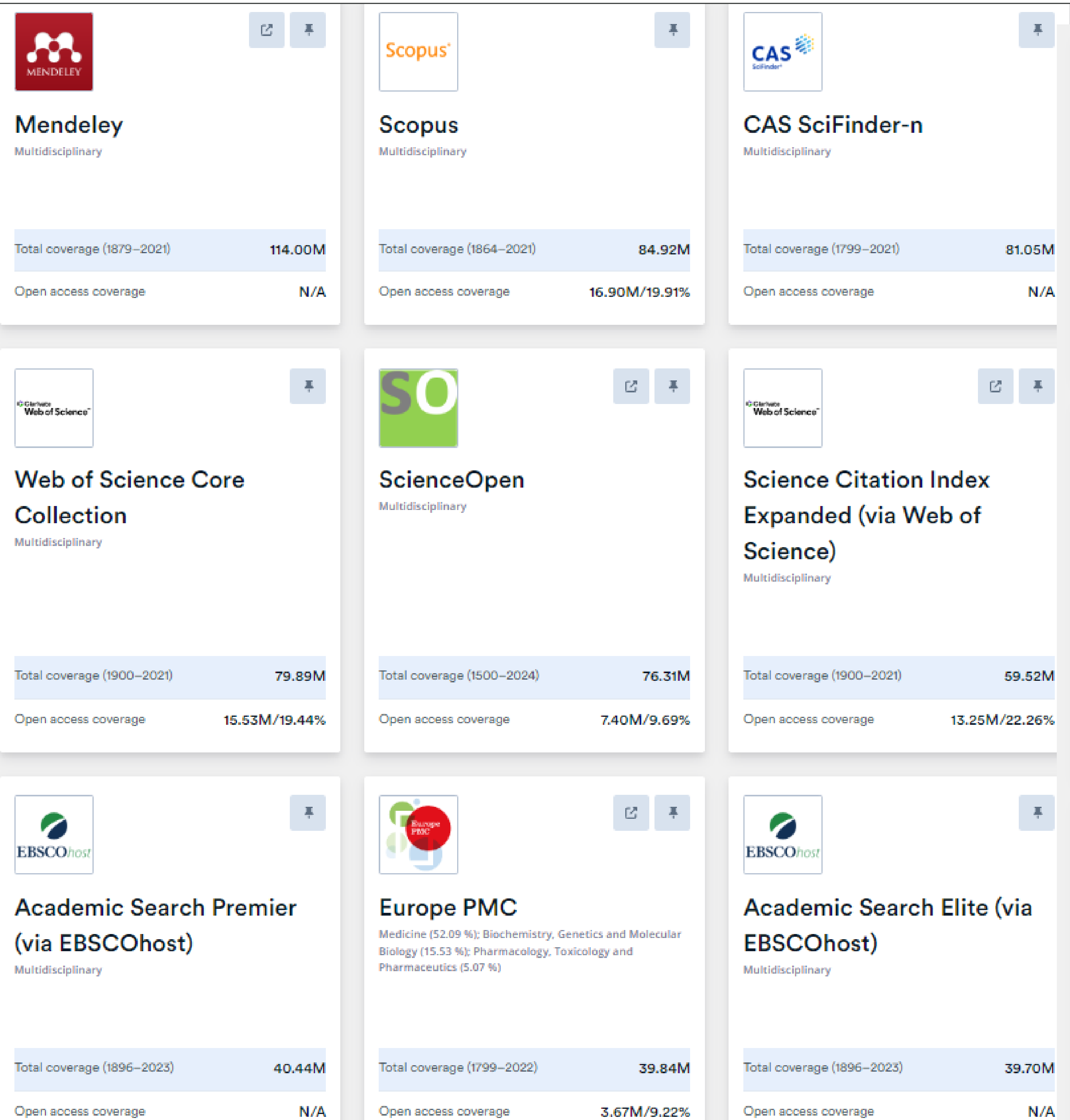
Dimensions - Publications
Multidisciplinary
Total coverage (1667–2023) 136.09M
Open access coverage 44.09M/32.40%

Dimensions - Publications (free)
Multidisciplinary
Total coverage (1667–2023) 136.09M
Open access coverage 44.09M/32.40%

Crossref
Multidisciplinary
Total coverage (1800–2021) 130.89M
Open access coverage 39.05M/29.84%

CNKI Overseas
Multidisciplinary
Total coverage (1755–2024) 123.40M
Open access coverage N/A

SciTech Premium Collection (via ProQuest)
Multidisciplinary
Total coverage (1799–2022) 116.66M
Open access coverage N/A



Mendeley
Multidisciplinary
Total coverage (1879–2021) 114.00M
Open access coverage N/A

Scopus
Multidisciplinary
Total coverage (1864–2021) 84.92M
Open access coverage 16.90M/19.91%

CAS SciFinder-n
Multidisciplinary
Total coverage (1799–2021) 81.05M
Open access coverage N/A

Web of Science Core Collection
Multidisciplinary
Total coverage (1900–2021) 79.89M
Open access coverage 15.53M/19.44%

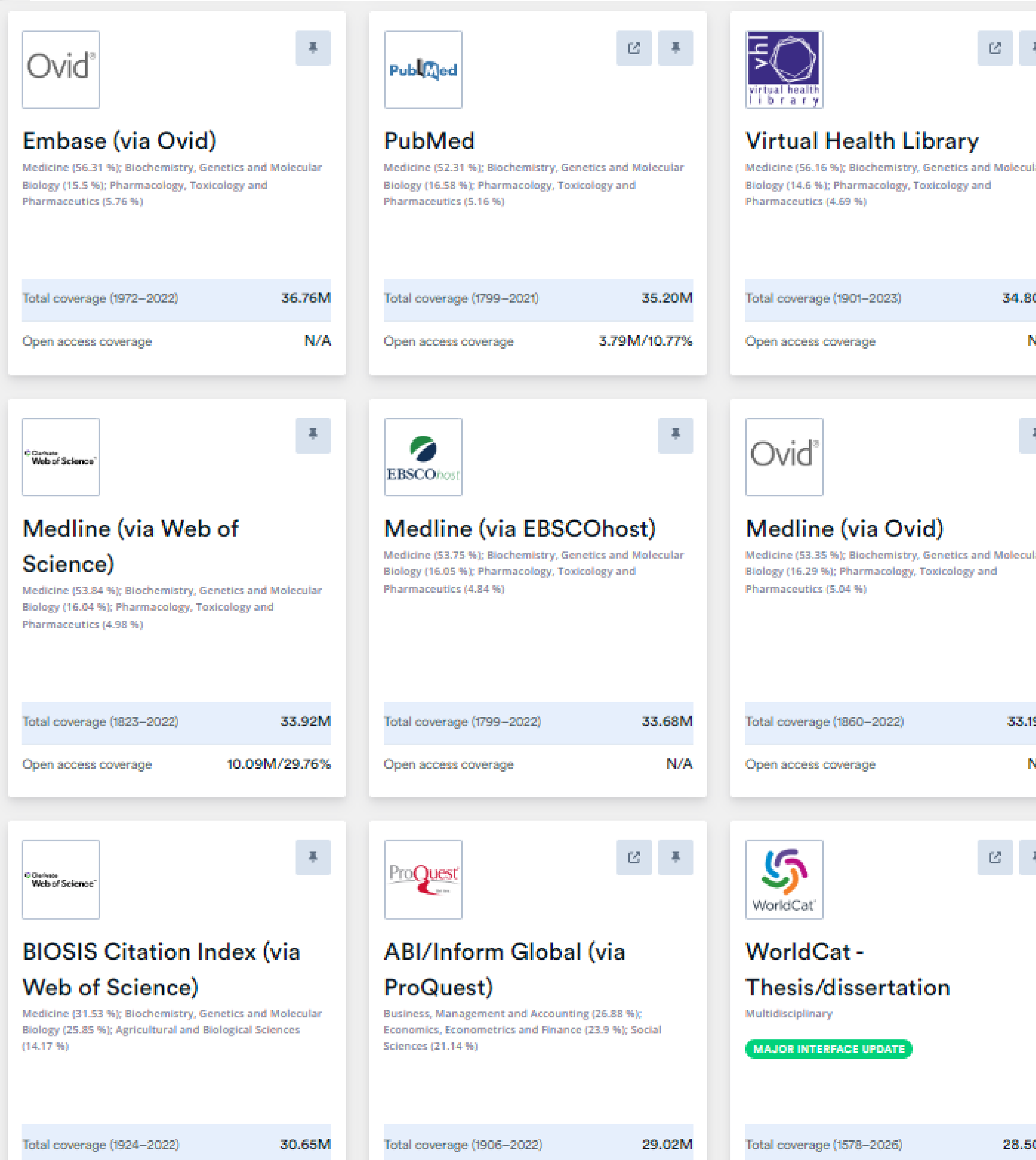
ScienceOpen
Multidisciplinary
Total coverage (1500–2024) 76.31M
Open access coverage 7.40M/9.69%

Science Citation Index Expanded (via Web of Science)
Multidisciplinary
Total coverage (1900–2021) 59.52M
Open access coverage 13.25M/22.26%

Academic Search Premier (via EBSCOhost)
Multidisciplinary
Total coverage (1896–2023) 40.44M
Open access coverage N/A

Europe PMC
Medicine (52.09 %); Biochemistry, Genetics and Molecular Biology (15.53 %); Pharmacology, Toxicology and Pharmaceutics (5.07 %)
Total coverage (1799–2022) 39.84M
Open access coverage 3.67M/9.22%

Academic Search Elite (via EBSCOhost)
Multidisciplinary
Total coverage (1896–2023) 39.70M
Open access coverage N/A



Embase (via Ovid)
Medicine (56.31 %); Biochemistry, Genetics and Molecular Biology (15.5 %); Pharmacology, Toxicology and Pharmaceutics (5.76 %)
Total coverage (1972–2022) 36.76M
Open access coverage N/A

PubMed
Medicine (52.31 %); Biochemistry, Genetics and Molecular Biology (16.58 %); Pharmacology, Toxicology and Pharmaceutics (5.16 %)
Total coverage (1799–2021) 35.20M
Open access coverage 3.79M/10.77%

Virtual Health Library
Medicine (56.16 %); Biochemistry, Genetics and Molecul
Biology (14.6 %); Pharmacology, Toxicology and Pharmaceutics (4.69 %)
Total coverage (1901–2023) 34.8
Open access coverage N

Medline (via Web of Science)
Medicine (53.84 %); Biochemistry, Genetics and Molecular Biology (16.04 %); Pharmacology, Toxicology and Pharmaceutics (4.98 %)
Total coverage (1823–2022) 33.92M
Open access coverage 10.09M/29.76%

Medline (via EBSCOhost)
Medicine (53.75 %); Biochemistry, Genetics and Molecular Biology (16.05 %); Pharmacology, Toxicology and Pharmaceutics (4.84 %)
Total coverage (1799–2022) 33.68M
Open access coverage N/A

Medline (via Ovid)
Medicine (53.35 %); Biochemistry, Genetics and Molecul
Biology (16.29 %); Pharmacology, Toxicology and Pharmaceutics (5.04 %)
Total coverage (1860–2022) 33.1
Open access coverage N

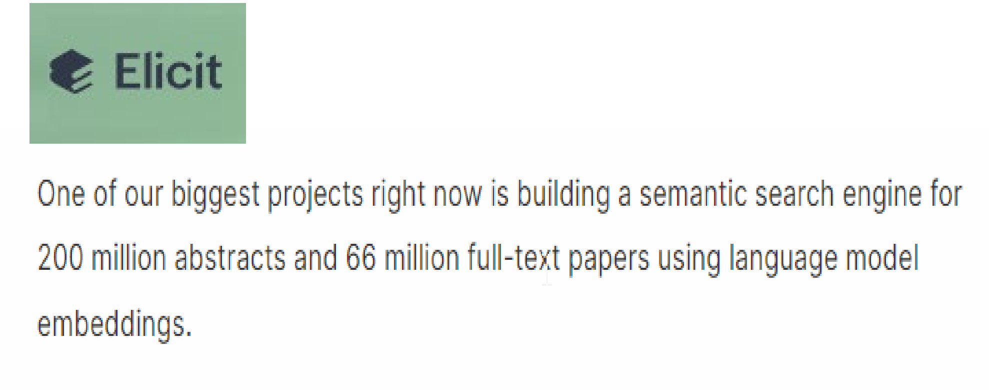
BIOSIS Citation Index (via Web of Science)
Medicine (31.53 %); Biochemistry, Genetics and Molecular Biology (25.85 %); Agricultural and Biological Sciences (14.17 %)
Total coverage (1924–2022) 30.65M

ABI/Inform Global (via ProQuest)
Business, Management and Accounting (26.88 %); Economics, Econometrics and Finance (23.9 %); Social Sciences (21.14 %)
Total coverage (1906–2022) 29.02M

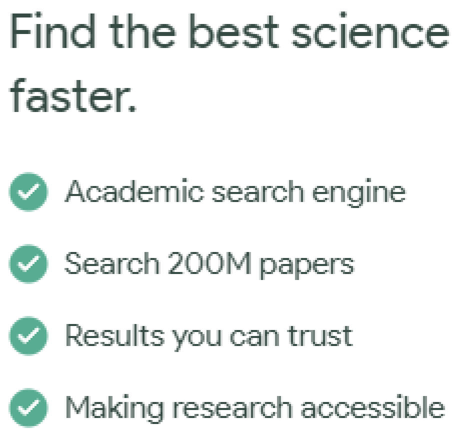
WorldCat - Thesis/dissertation
Multidisciplinary
MAJOR INTERFACE UPDATE
Total coverage (1578–2026) 28.5

This does not even include the new "AI powered search"!

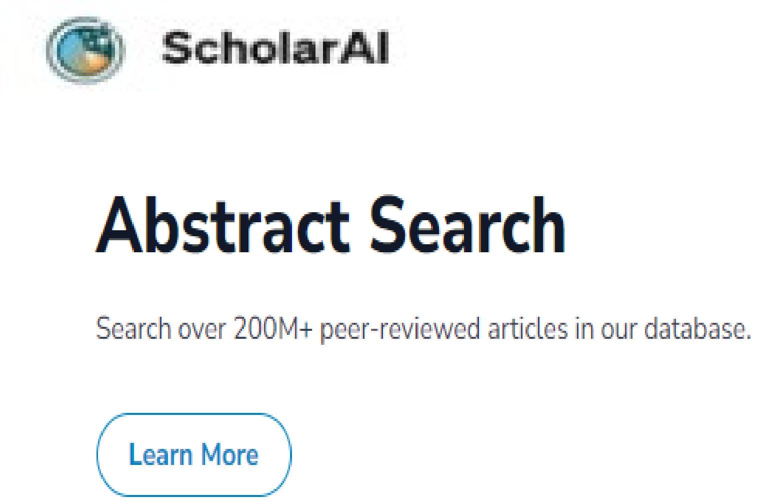
consensus


Elicit

One of our biggest projects right now is building a semantic search engine for 200 million abstracts and 66 million full-text papers using language model embeddings.

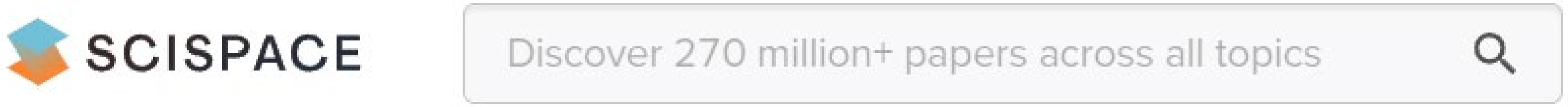

Find the best science faster.

- ✓ Academic search engine
- ✓ Search 200M papers
- ✓ Results you can trust
- ✓ Making research accessible


ScholarAI

Abstract Search

Search over 200M+ peer-reviewed articles in our database.


SCISPACE

Some sources of Open Scholarly Metadata + Applications

Crossref

DataCite
FIND, ACCESS, AND REUSE DATA

ORCID

ROR .community

Microsoft
Academic Graph

OpenAlex

Semantic Scholar

NIH

CORE

unpaywall

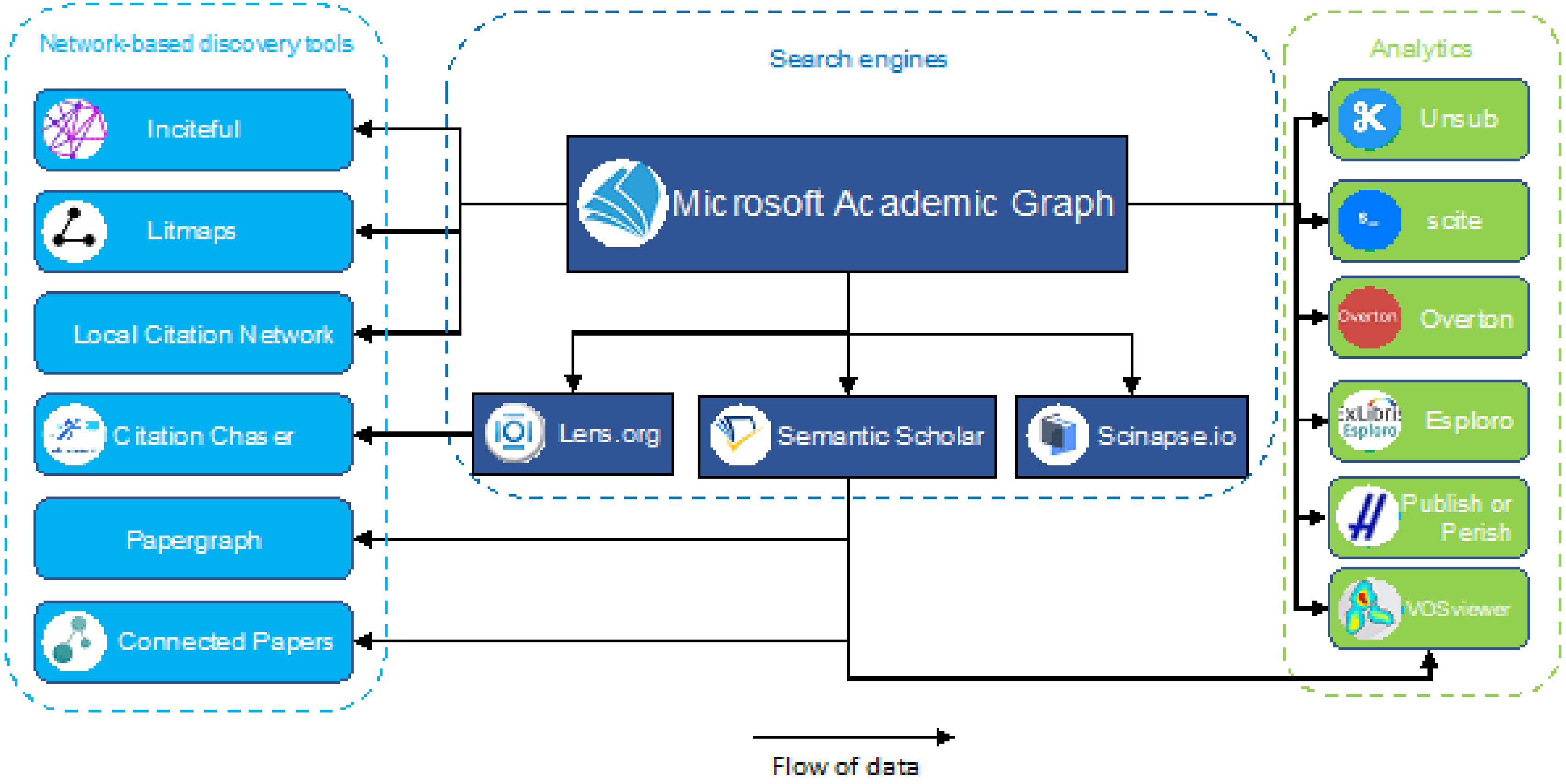
INTERNET
ARCHIVE

WIKIDATA

Open metadata & infrastructure (2018-)

- Initative for Open Citations - successfully hits 90% open citations in Crossref records (Jan 2021)
- Initative for Open Abstract launches (Sept 2020)
- Formal adoption of The **Principles of Open Scholarly Infrastructure** by Crossref, Datacite, ROR, OurResearch etc in 2020

Goodbye, Microsoft Academic – Hello, open research infrastructure? (Dec 2021)



Citation based literature mapping tools

Not just indexes

List of citation-based literature mapping tools

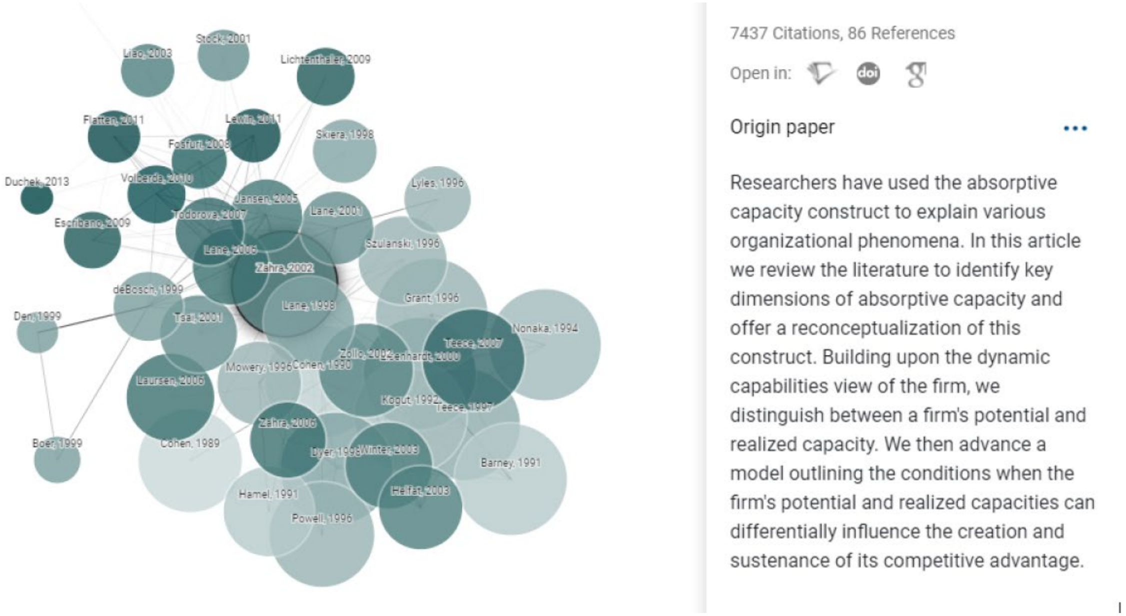
Citation based literature mapping tools

ResearchRabbit
Litmaps
Inciteful
CONNECTED PAPERS
OPEN KNOWLEDGE MAPS
A visual interface to the world's scientific knowledge
PURE suggest
citation gecko
Local Citation Network beta

- Bottom-up approach
- Tends to require you to start by adding a few relevant seed papers (around 10-20)
- Used by any type of researcher who wants to find papers by citation relationships (and some are starting to use text based similarity approaches)
- One step process, comes with a index that you can pull papers from
- User-friendly interface, does not use technical jargon

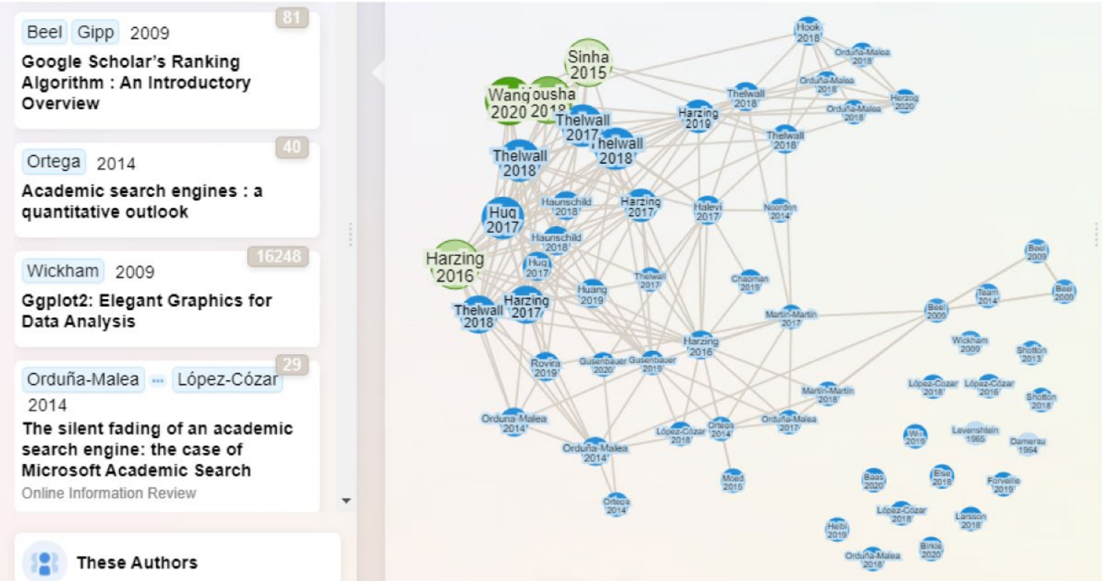
VOSviewer, Citespace, Bibliometrix , CitNetExplorer, Sci2, HistCite, Hazing
Publish or Perish

New innovative tools by startups, hobbyist etc



7437 Citations, 86 References

Researchers have used the absorptive capacity construct to explain various organizational phenomena. In this article we review the literature to identify key dimensions of absorptive capacity and offer a reconceptualization of this construct. Building upon the dynamic capabilities view of the firm, we distinguish between a firm's potential and realized capacity. We then advance a model outlining the conditions when the firm's potential and realized capacities can differentially influence the creation and sustenance of its competitive advantage.



Beel Gipp 2009
Google Scholar's Ranking Algorithm : An Introductory Overview

Ortega 2014
Academic search engines : a quantitative outlook

Wickham 2009
Ggplot2: Elegant Graphics for Data Analysis

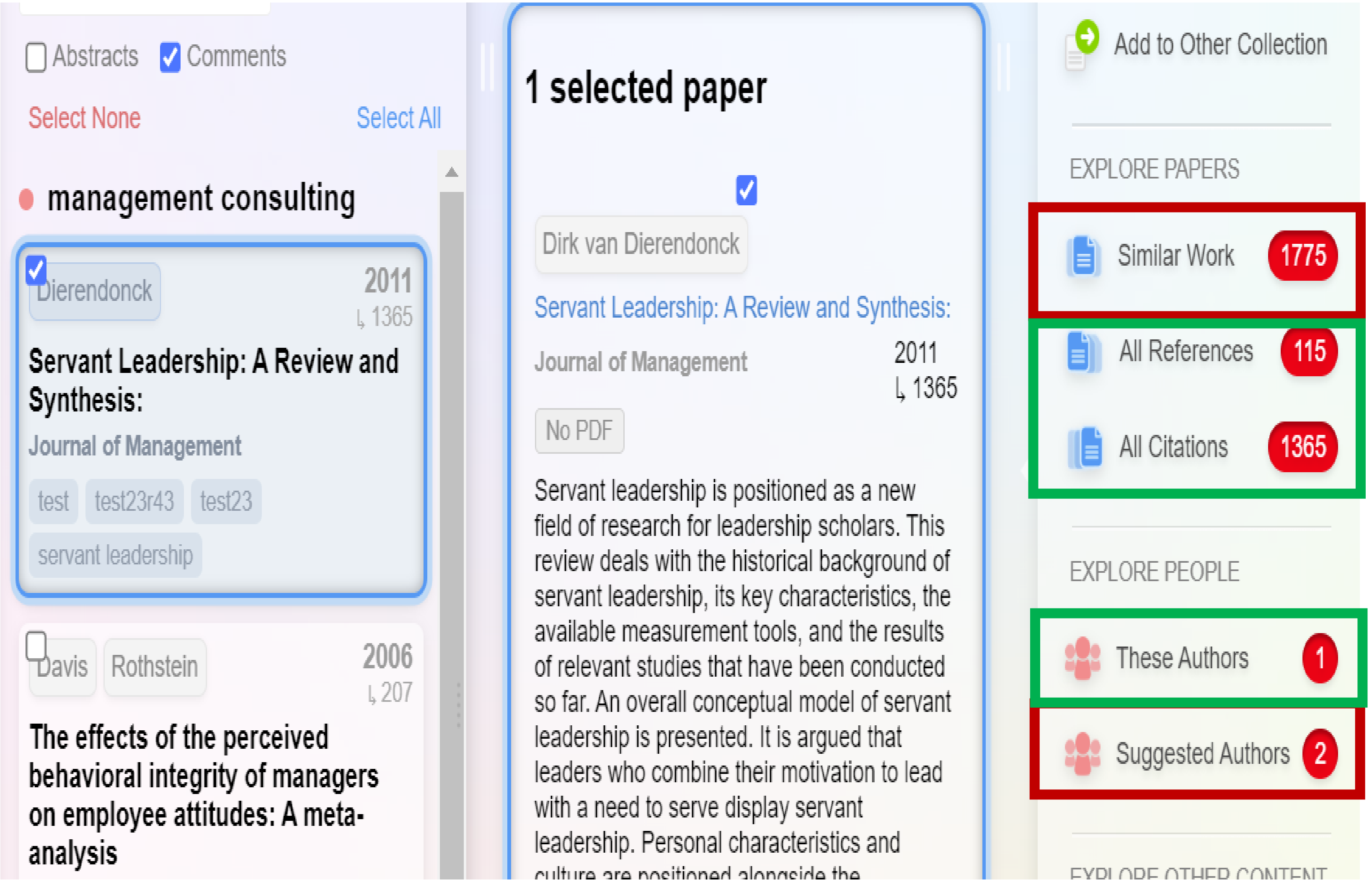
Orduña-Malea López-Cózar 2014
The silent fading of an academic search engine: the case of Microsoft Academic Search
Online Information Review

[ConnectedPapers]

[Research Rabbit]

You can think of this as *citation searching/chasing etc at scale.* However, these tools may use citation/bibliometric based techniques that are not transparent.

Transparency of Research Rabbit – article level



☐ Abstracts ☑ Comments

Select None Select All

● management consulting

☑ Dierendonck 2011 ↳ 1365

Servant Leadership: A Review and Synthesis:

Journal of Management

test test23r43 test23

servant leadership

☐ Davis Rothstein 2006 ↳ 207

The effects of the perceived behavioral integrity of managers on employee attitudes: A meta-analysis

1 selected paper

☑

Dirk van Dierendonck

Servant Leadership: A Review and Synthesis:

Journal of Management 2011 ↳ 1365

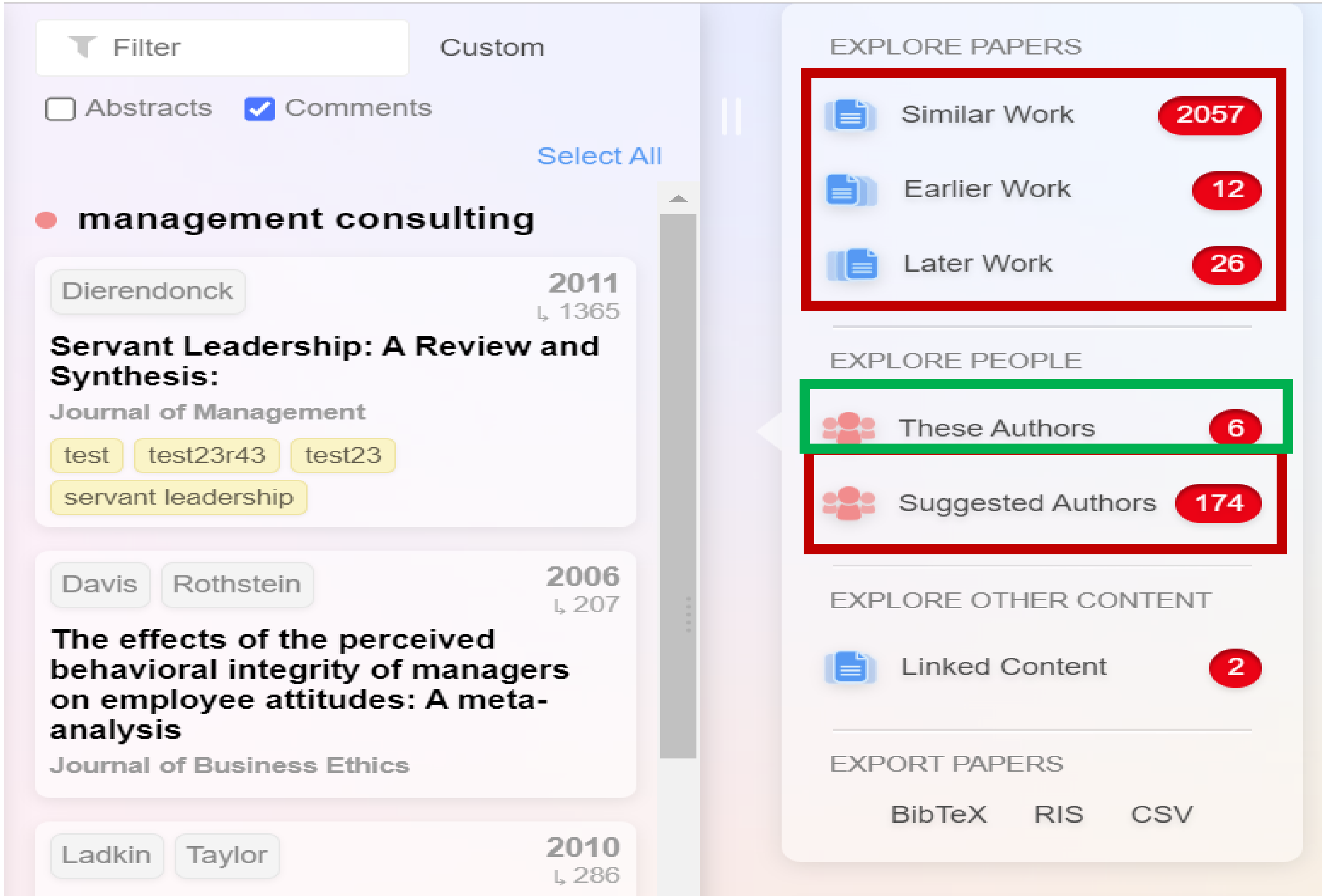
No PDF

Servant leadership is positioned as a new field of research for leadership scholars. This review deals with the historical background of servant leadership, its key characteristics, the available measurement tools, and the results of relevant studies that have been conducted so far. An overall conceptual model of servant leadership is presented. It is argued that leaders who combine their motivation to lead with a need to serve display servant leadership. Personal characteristics and culture are positioned alongside the

Not transparent

Transparent

Transparency of Research Rabbit – collection level



Filter
Custom

☐ Abstracts ☑ Comments

Select All

management consulting

Dierendonck
2011
↳ 1365
Servant Leadership: A Review and Synthesis:
Journal of Management
test test23r43 test23
servant leadership

Davis Rothstein
2006
↳ 207
The effects of the perceived behavioral integrity of managers on employee attitudes: A meta-analysis
Journal of Business Ethics

Ladkin Taylor
2010
↳ 286

EXPLORE PAPERS

Similar Work	2057
Earlier Work	12
Later Work	26

EXPLORE PEOPLE

These Authors	6
Suggested Authors	174

EXPLORE OTHER CONTENT

Linked Content	2

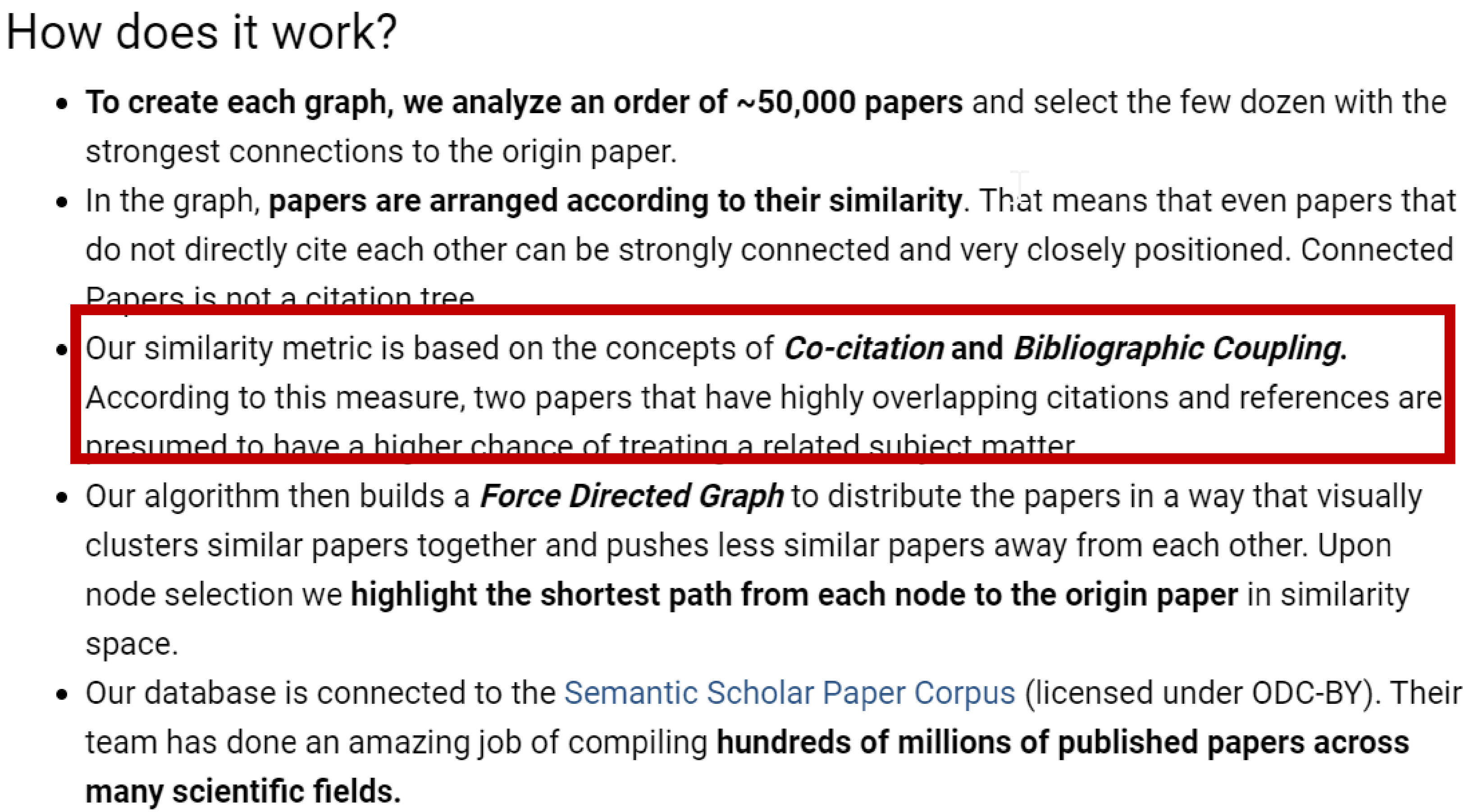
EXPORT PAPERS

BibTeX RIS CSV

🟥 Not transparent

🟩 Transparent

Transparency of Connected papers



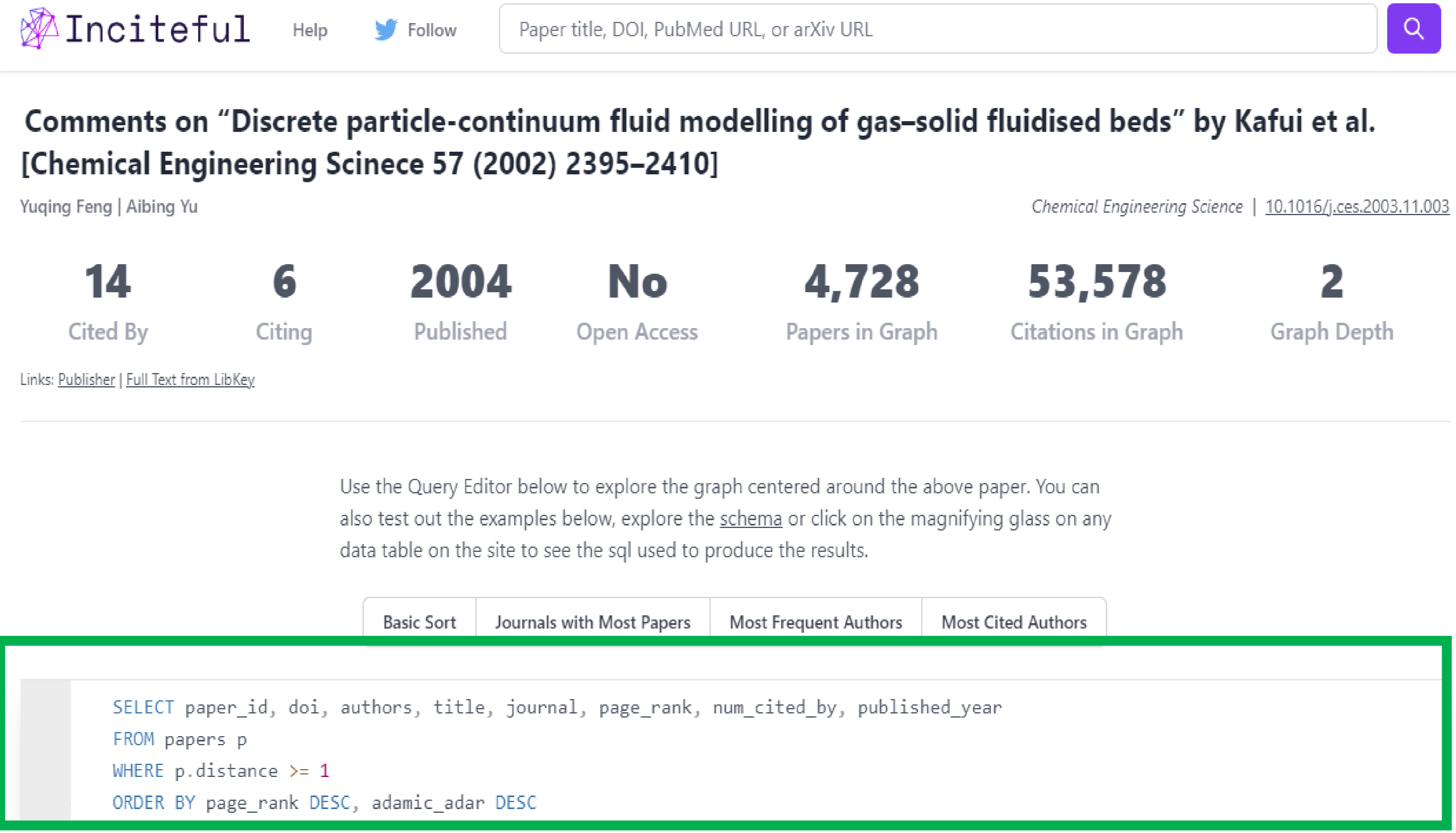
How does it work?

- **To create each graph, we analyze an order of ~50,000 papers** and select the few dozen with the strongest connections to the origin paper.
- In the graph, **papers are arranged according to their similarity**. That means that even papers that do not directly cite each other can be strongly connected and very closely positioned. Connected Papers is not a citation tree
- Our similarity metric is based on the concepts of ***Co-citation*** **and** ***Bibliographic Coupling.*** According to this measure, two papers that have highly overlapping citations and references are presumed to have a higher chance of treating a related subject matter
- Our algorithm then builds a ***Force Directed Graph*** to distribute the papers in a way that visually clusters similar papers together and pushes less similar papers away from each other. Upon node selection we **highlight the shortest path from each node to the origin paper** in similarity space.
- Our database is connected to the Semantic Scholar Paper Corpus (licensed under ODC-BY). Their team has done an amazing job of compiling **hundreds of millions of published papers across many scientific fields.**

Not transparent

Transparent

Transparency of inciteful



Comments on "Discrete particle-continuum fluid modelling of gas–solid fluidised beds" by Kafui et al. [Chemical Engineering Scinece 57 (2002) 2395–2410]

Yuqing Feng | Aibing Yu

Chemical Engineering Science | 10.1016/j.ces.2003.11.003

14	6	2004	No	4,728	53,578	2
Cited By	Citing	Published	Open Access	Papers in Graph	Citations in Graph	Graph Depth

Links: Publisher | Full Text from LibKey

Use the Query Editor below to explore the graph centered around the above paper. You can also test out the examples below, explore the schema or click on the magnifying glass on any data table on the site to see the sql used to produce the results.

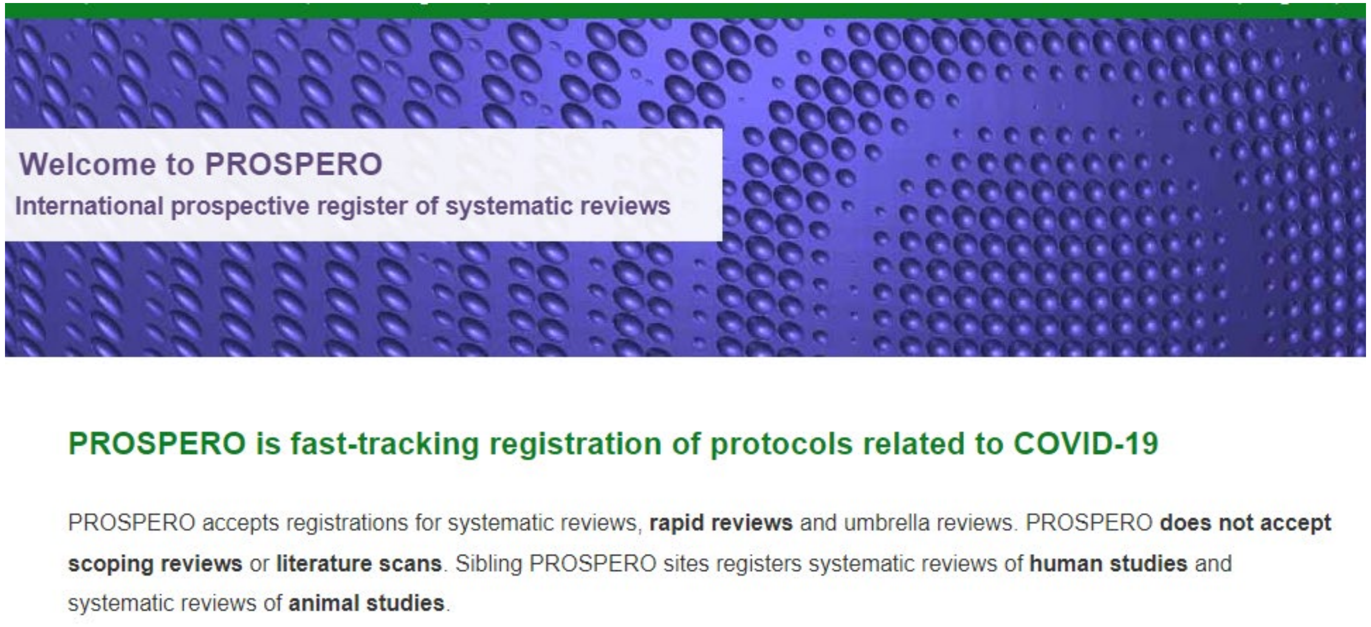
```
SELECT paper_id, doi, authors, title, journal, page_rank, num_cited_by, published_year
FROM papers p
WHERE p.distance >= 1
ORDER BY page_rank DESC, adamic_adar DESC
```

Not transparent

Transparent

How frequent are these tools used in evidence synthesis

How to estimate use?

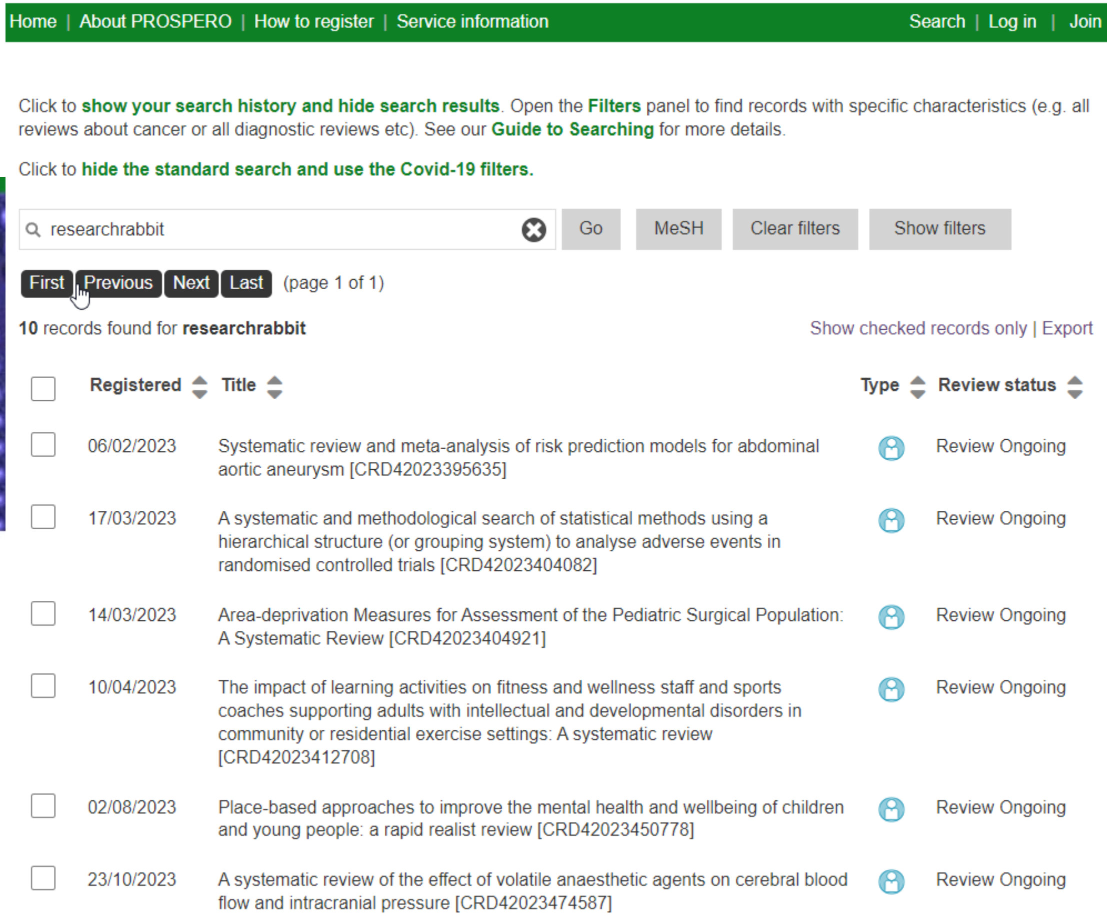


Welcome to PROSPERO

International prospective register of systematic reviews

PROSPERO is fast-tracking registration of protocols related to COVID-19

PROSPERO accepts registrations for systematic reviews, **rapid reviews** and umbrella reviews. PROSPERO **does not accept scoping reviews** or **literature scans**. Sibling PROSPERO sites registers systematic reviews of **human studies** and systematic reviews of **animal studies**.



Home | About PROSPERO | How to register | Service information
Search | Log in | Join

Click to **show your search history and hide search results**. Open the **Filters** panel to find records with specific characteristics (e.g. all reviews about cancer or all diagnostic reviews etc). See our **Guide to Searching** for more details.

Click to **hide the standard search and use the Covid-19 filters.**

researchrabbit

(page 1 of 1)

10 records found for **researchrabbit**

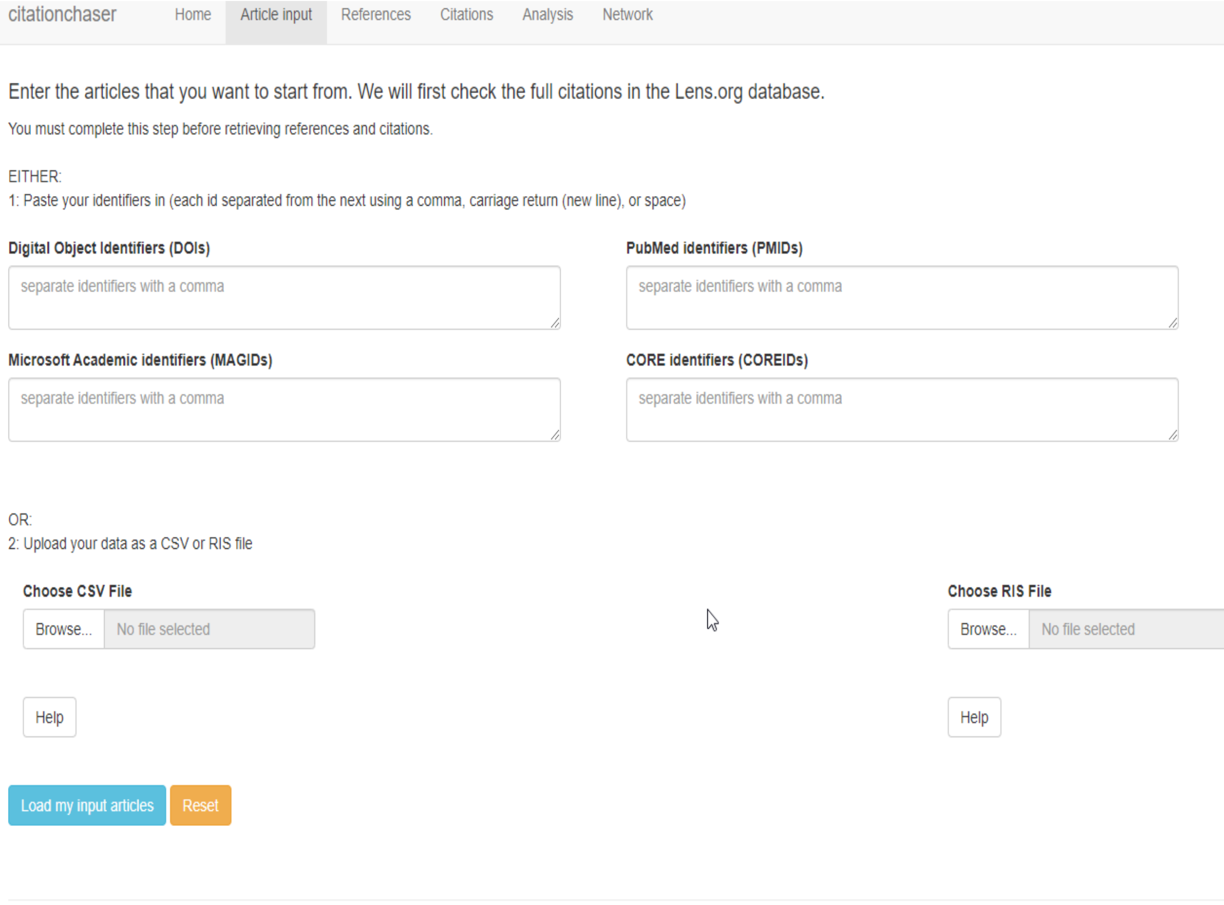
Show checked records only | Export

☐	Registered	Title	Type	Review status
☐	06/02/2023	Systematic review and meta-analysis of risk prediction models for abdominal aortic aneurysm [CRD42023395635]		Review Ongoing
☐	17/03/2023	A systematic and methodological search of statistical methods using a hierarchical structure (or grouping system) to analyse adverse events in randomised controlled trials [CRD42023404082]		Review Ongoing
☐	14/03/2023	Area-deprivation Measures for Assessment of the Pediatric Surgical Population: A Systematic Review [CRD42023404921]		Review Ongoing
☐	10/04/2023	The impact of learning activities on fitness and wellness staff and sports coaches supporting adults with intellectual and developmental disorders in community or residential exercise settings: A systematic review [CRD42023412708]		Review Ongoing
☐	02/08/2023	Place-based approaches to improve the mental health and wellbeing of children and young people: a rapid realist review [CRD42023450778]		Review Ongoing
☐	23/10/2023	A systematic review of the effect of volatile anaesthetic agents on cerebral blood flow and intracranial pressure [CRD42023474587]		Review Ongoing

This method probably undercounts usage of these tools*

Use of citation based literature mapping tools

- **CitationChaser(72)** - 2021-2024
- **ConnectedPapers (24)** - 2020-2024
- **CoCites(11)** - 2020-2023 (discontinued)
- **ResearchRabbit(10)** - 2023-2024
- **Litmaps(6)** - 2023-2024
- **CitationGecko(4)** - 2022-2024



Enter the articles that you want to start from. We will first check the full citations in the Lens.org database.

You must complete this step before retrieving references and citations.

EITHER:
1: Paste your identifiers in (each id separated from the next using a comma, carriage return (new line), or space)

Digital Object Identifiers (DOIs): separate identifiers with a comma

PubMed identifiers (PMIDs): separate identifiers with a comma

Microsoft Academic identifiers (MAGIDs): separate identifiers with a comma

CORE identifiers (COREIDs): separate identifiers with a comma

OR:
2: Upload your data as a CSV or RIS file

Choose CSV File: No file selected

Choose RIS File: No file selected

CitationChaser by Haddaway, N. R., Grainger, M. J., Gray, C. T.

Use of new mega-indexes

- Semantic Scholar (202) - 2018-2024
- Microsoft Academic (90) - 2015-2024
- Lens org(27) - 2018-2024
- Dimensions ai(13) - 2021-2024
- OpenAlex(12) - 2022-2024
- Scinapse(12)- 2020-2023
- Scilit(9) - 2021-2023
- Microsoft Academic Graph(4) - 2020-2022

SEMANTIC SCHOLAR

A free, AI-powered research tool for scientific literature

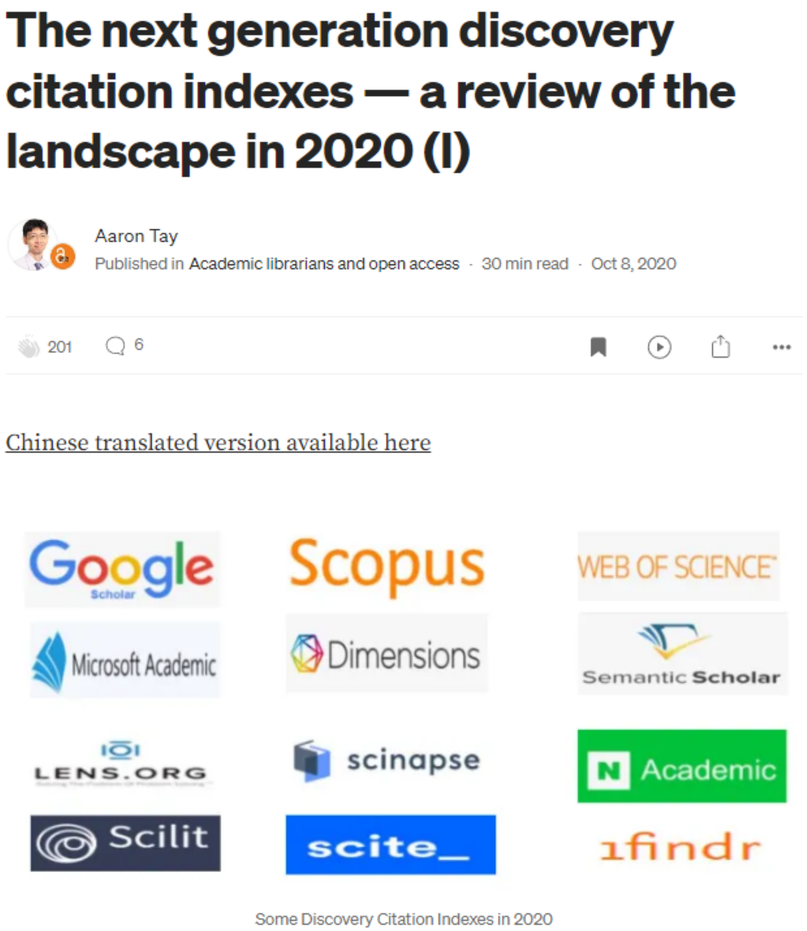
Search 217,973,376 papers from all fields of science

Comparison – Google Scholar mentioned 8k times in 2023 alone!

What are the tools used for

- Normal database search
  - The studies will be searched in PubMed/MEDLINE, AMED, CINAHL, Cochrane Central Register of Controlled Trials (CENTRAL), PsycINFO , ERIC, ScienceDirect, SciencePubCo, Scopus database, JSTOR, ProQuest, Google Scholar, Semantic Scholar, ClinicalTrials.gov.
  - This search will be completed by a search in the world wide web using google and Bing Chat (chatbot based on ChatGPT-4) and in ClinicalTrials.gov.
  - We will screen PubMed, CINAHL, EMBASE, PERPLEXITY without any date restriction.
  - The systematic review will search: AI-assisted tools: Elicit.org, Consensus AI Research Assistant and other relevant GPTs/AI Research Assistants

- Search grey literature
  - Grey literature will be searched on Google Scholar, BASE, Semantic Scholar, OpenGrey and the researchers will also hand-search using reference lists and books.
  - Use of search engines to identify grey literature (not yet completed): Elicit https://elicit.org/ – first 20 results
- Supplementary search
  - In an effort to find the greatest number of related publications, an additional search in Google Scholar, the related search tools in PubMed and connectedpapers.com were used as secondary sources
- Citation chasing
  - Forward /backward citation searches of included studies will be performed in Research Rabbit (https://researchrabbitapp.com/home) and Scite (https://scite.ai/home).
  - In addition, manual searches were carried out using the snowball method, the PubMed “similars” tool and the use of the www.connectedpapers.com tool
- Citation in context
  - Scite and Litsense to capture citation in context.



The next generation discovery citation indexes — a review of the landscape in 2020 (I)

Aaron Tay
Published in Academic librarians and open access · 30 min read · Oct 8, 2020

201 6

Chinese translated version available here

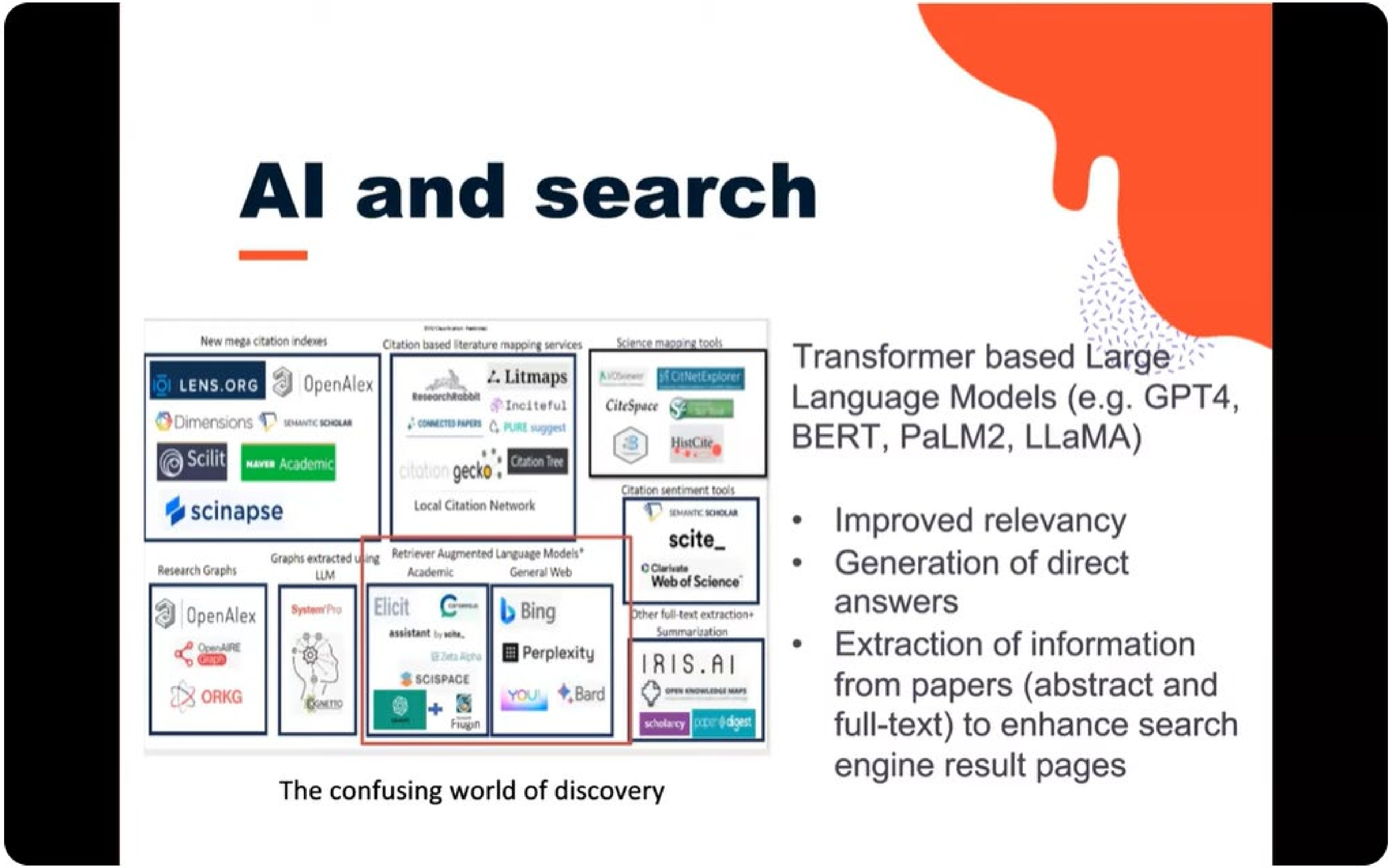
Some Discovery Citation Indexes in 2020

Learn more

Moral of story

1. It is now relatively easy for any startup to create a search engine/index from 200 million academic works
2. This creates the foundation for the rise of AI powered search tools

Three ways "AI" will change search



AI and search

The confusing world of discovery

Transformer based Large Language Models (e.g. GPT4, BERT, PaLM2, LLaMA)

- Improved relevancy
- Generation of direct answers
- Extraction of information from papers (abstract and full-text) to enhance search engine result pages

The possible impact of AI on search and discovery

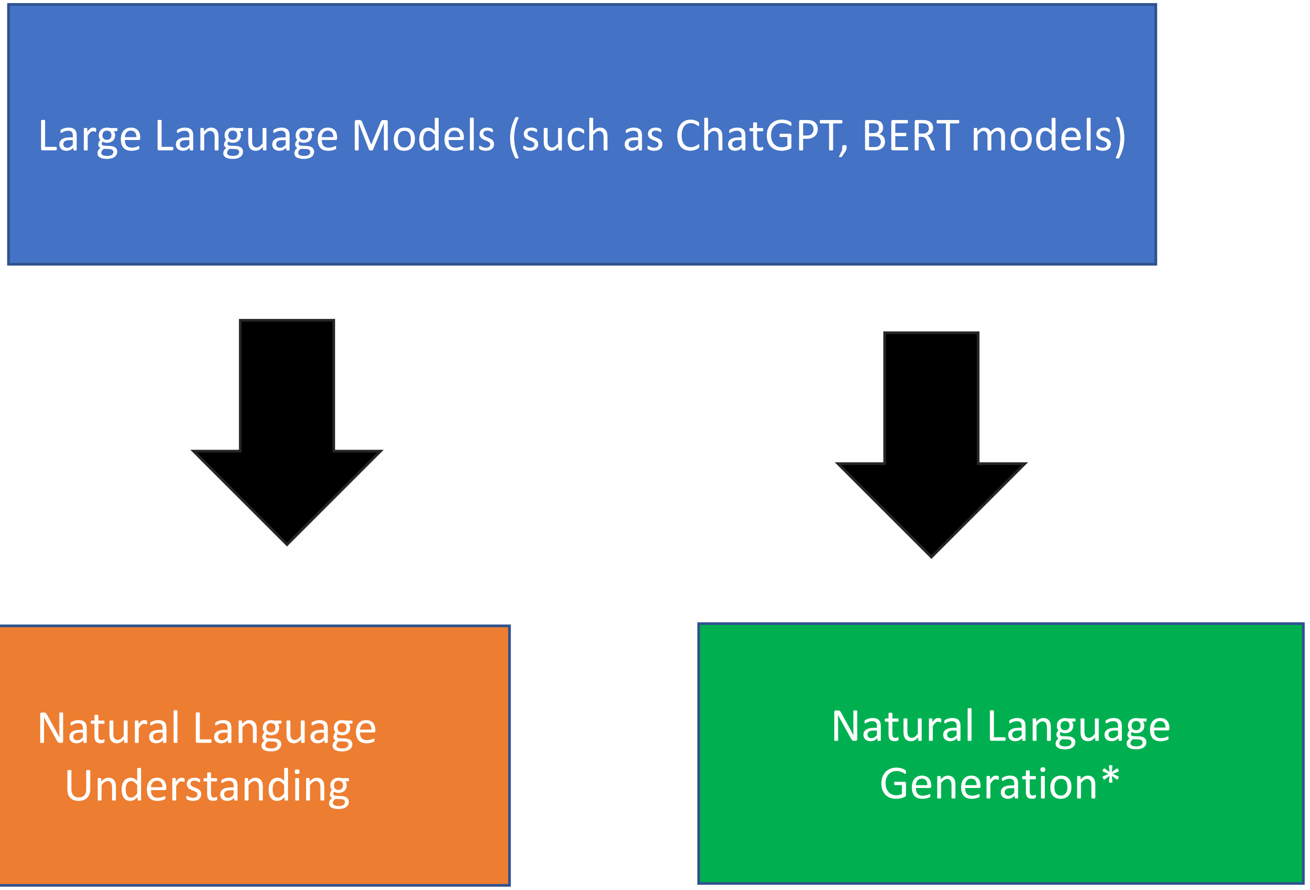
The possible impact of AI on search and discovery (42 minutes)

What do we mean by "generative AI"?

- These days AI = Large Language Models (implemented by Neural nets with Transformer architecture)

- OpenAI's – GPT3.5, GPT4, GPT4-Turno
- Anthropic - Claude 3 Opus, Sonet, Haiku
- Google/Deep Mind – BERT*, T5*, Gemini Ultra, Gemini 1.5 pro
- Meta's – Llama 3 (opensource)
- AI2 - OLMo (Academic Open Language Model)

* Most of the list are decoder only models, BERT is encoder model
and T5 is Encoder decoder model
** Some restrictions



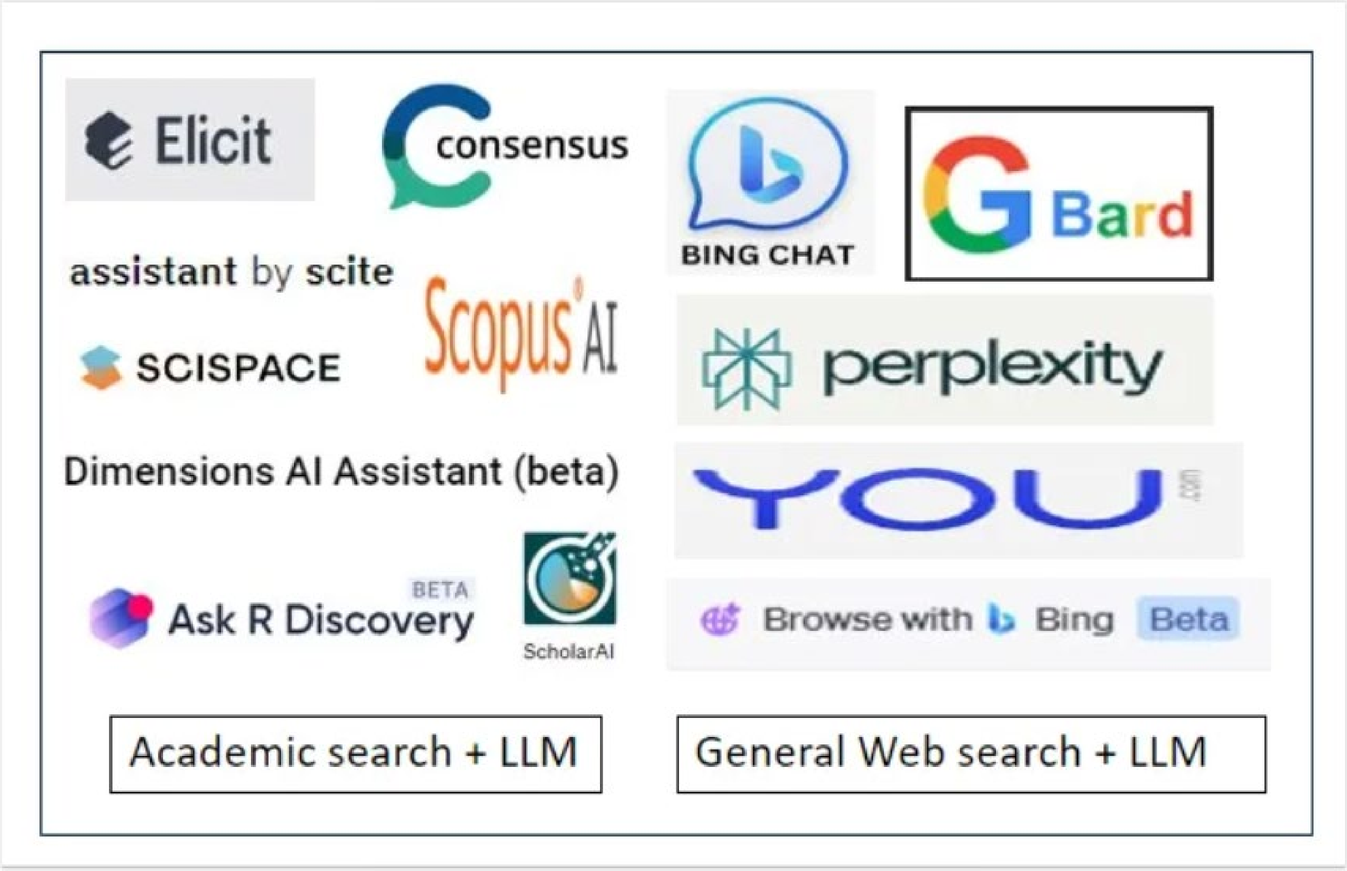
Large Language Models (such as ChatGPT, BERT models)

Natural Language Understanding

Natural Language Generation*

For doubts about "understanding" see - On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? & LLMs can't plan (..but they can help you in planning)

List of academic search engines that use Large Language models*



- Technically I list only academic search engines that use RAG for generation of answers

https://musingsaboutlibrarianship.blogspot.com/p/list-of-academic-search-engines-that.html

How do LLM's improve search (2023)

Current "standard" (2023) use

- **Improved relevancy (through contextual embeddings)**
- **Extraction of information from papers (abstract and full-text) to enhance search engine result pages**
- Generation of direct answers (paragraph or two using RAG)

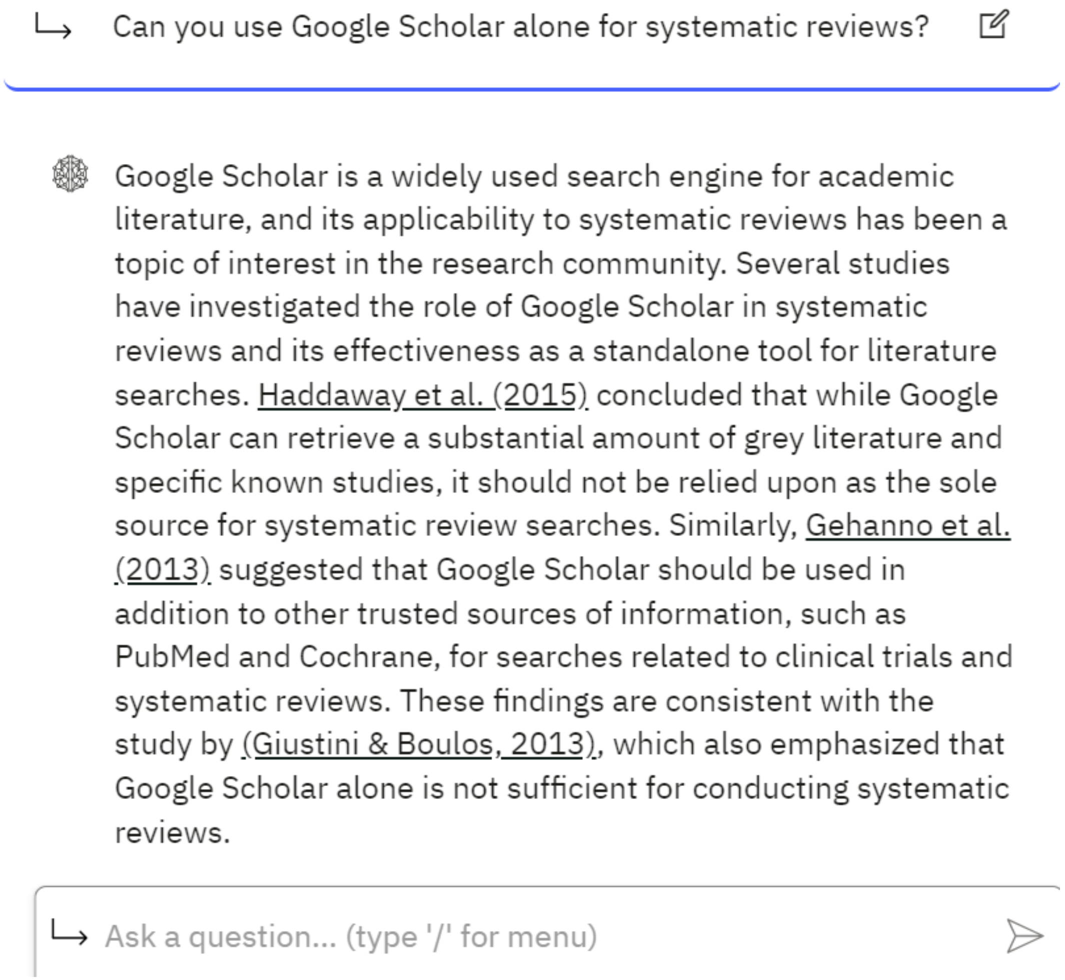
Things people working on

- Semi automation of systematic reviews (agent based)
- Intergration with Knowledge Graph
- **Literature review generation** (holy grail)

Generation of direct answer using Retrieval Augmented Generation

"Google can bring you back 100,000 answers, a librarian can bring you back the right one" – Neil Gaiman

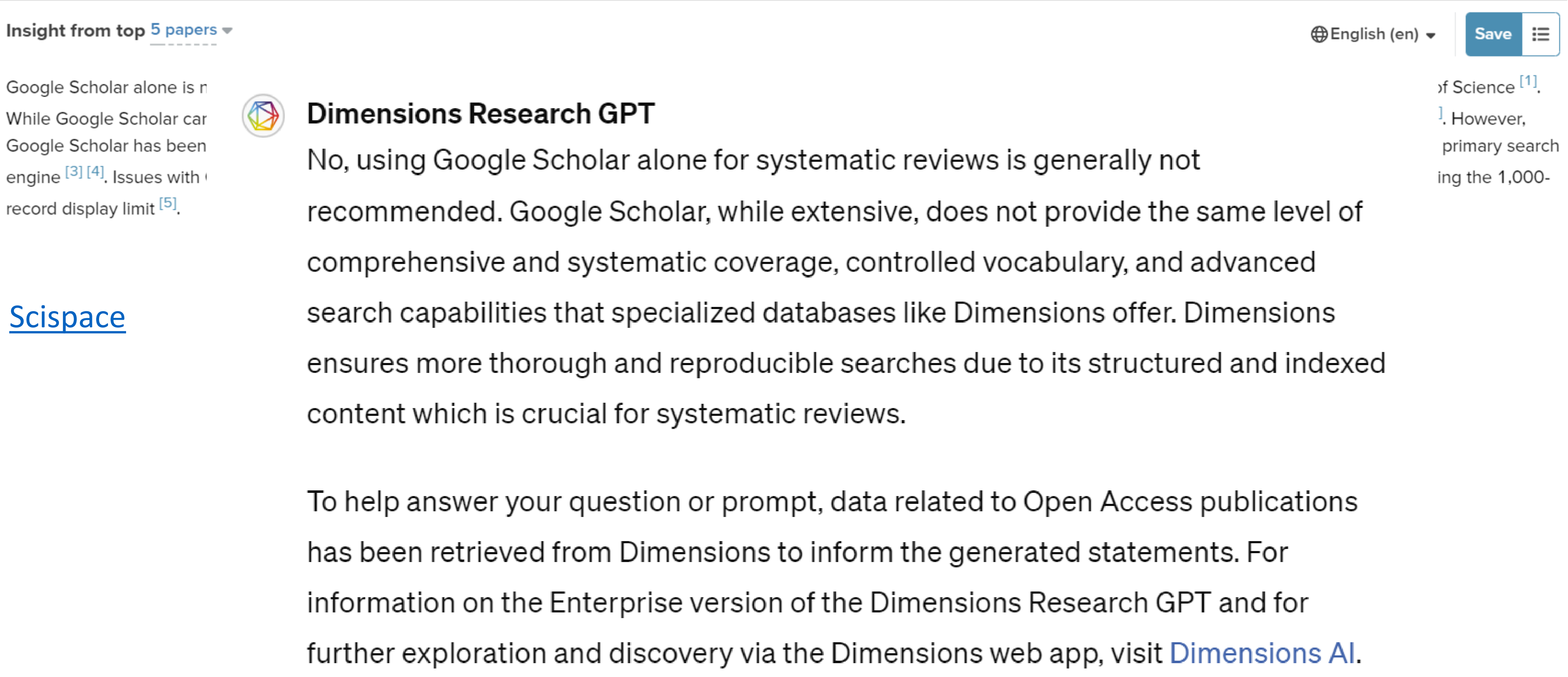
With RAG (retrieval augmented generation) - Scite.ai assistant and other search engines can too!



Can you use Google Scholar alone for systematic reviews?

Google Scholar is a widely used search engine for academic literature, and its applicability to systematic reviews has been a topic of interest in the research community. Several studies have investigated the role of Google Scholar in systematic reviews and its effectiveness as a standalone tool for literature searches. Haddaway et al. (2015) concluded that while Google Scholar can retrieve a substantial amount of grey literature and specific known studies, it should not be relied upon as the sole source for systematic review searches. Similarly, Gehanno et al. (2013) suggested that Google Scholar should be used in addition to other trusted sources of information, such as PubMed and Cochrane, for searches related to clinical trials and systematic reviews. These findings are consistent with the study by (Giustini & Boulos, 2013), which also emphasized that Google Scholar alone is not sufficient for conducting systematic reviews.

Other answers – Dimension Research GPT



Google Scholar alone is r
While Google Scholar car
Google Scholar has been
engine [3] [4]. Issues with
record display limit [5].

›f Science [1].
]. However,
primary search
ing the 1,000-

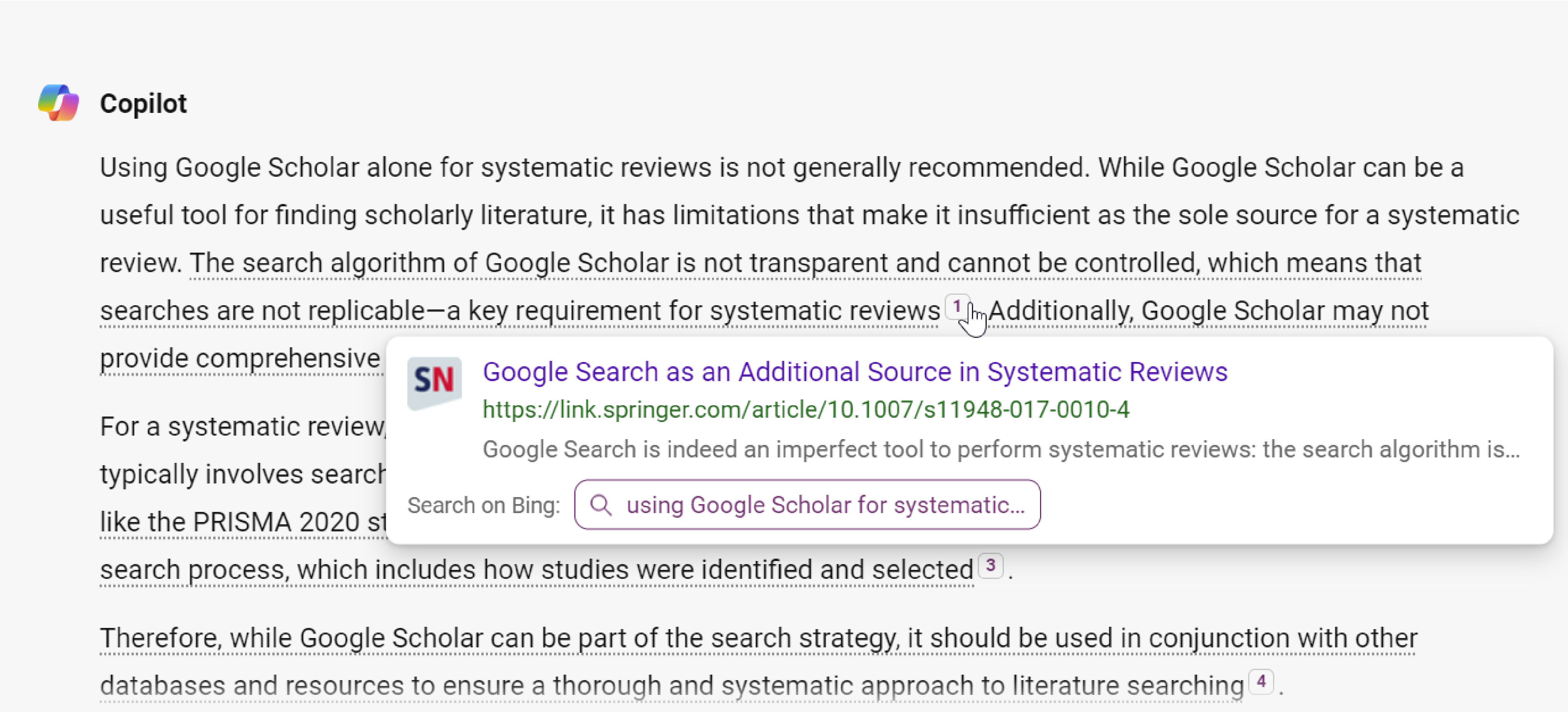
Dimensions Research GPT

No, using Google Scholar alone for systematic reviews is generally not recommended. Google Scholar, while extensive, does not provide the same level of comprehensive and systematic coverage, controlled vocabulary, and advanced search capabilities that specialized databases like Dimensions offer. Dimensions ensures more thorough and reproducible searches due to its structured and indexed content which is crucial for systematic reviews.

To help answer your question or prompt, data related to Open Access publications has been retrieved from Dimensions to inform the generated statements. For information on the Enterprise version of the Dimensions Research GPT and for further exploration and discovery via the Dimensions web app, visit Dimensions AI.

[Scispace]

Bing Chat (Copilot)



Copilot

Using Google Scholar alone for systematic reviews is not generally recommended. While Google Scholar can be a useful tool for finding scholarly literature, it has limitations that make it insufficient as the sole source for a systematic review. The search algorithm of Google Scholar is not transparent and cannot be controlled, which means that searches are not replicable—a key requirement for systematic reviews [1]. Additionally, Google Scholar may not provide comprehensive

For a systematic review,

typically involves search

like the PRISMA 2020 st

search process, which includes how studies were identified and selected [3].

Therefore, while Google Scholar can be part of the search strategy, it should be used in conjunction with other databases and resources to ensure a thorough and systematic approach to literature searching [4].

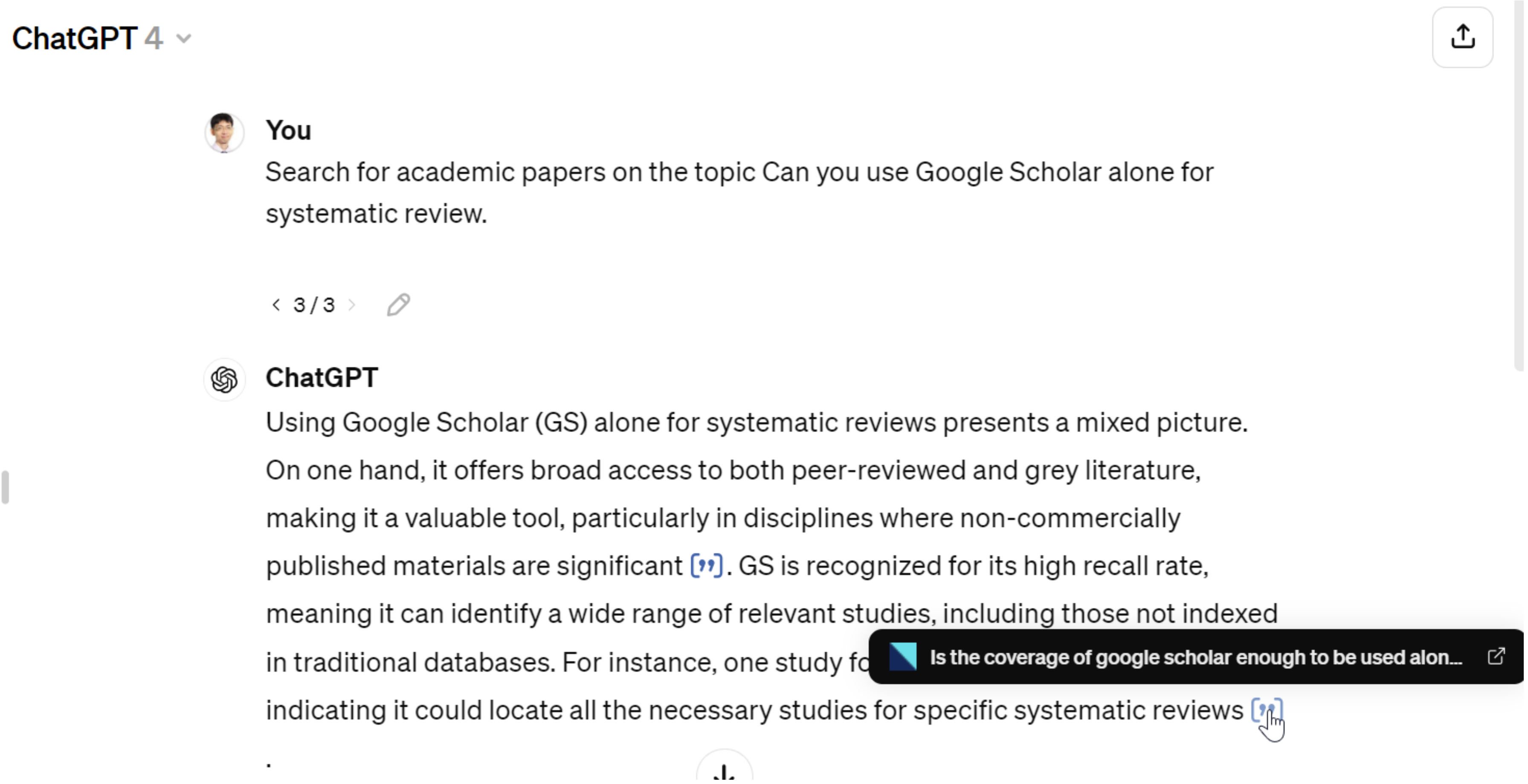
Google Search as an Additional Source in Systematic Reviews

https://link.springer.com/article/10.1007/s11948-017-0010-4

Google Search is indeed an imperfect tool to perform systematic reviews: the search algorithm is...

Search on Bing: using Google Scholar for systematic...

ChatGPT+



You

Search for academic papers on the topic Can you use Google Scholar alone for systematic review.

ChatGPT

Using Google Scholar (GS) alone for systematic reviews presents a mixed picture. On one hand, it offers broad access to both peer-reviewed and grey literature, making it a valuable tool, particularly in disciplines where non-commercially published materials are significant [”]. GS is recognized for its high recall rate, meaning it can identify a wide range of relevant studies, including those not indexed in traditional databases. For instance, one study fo

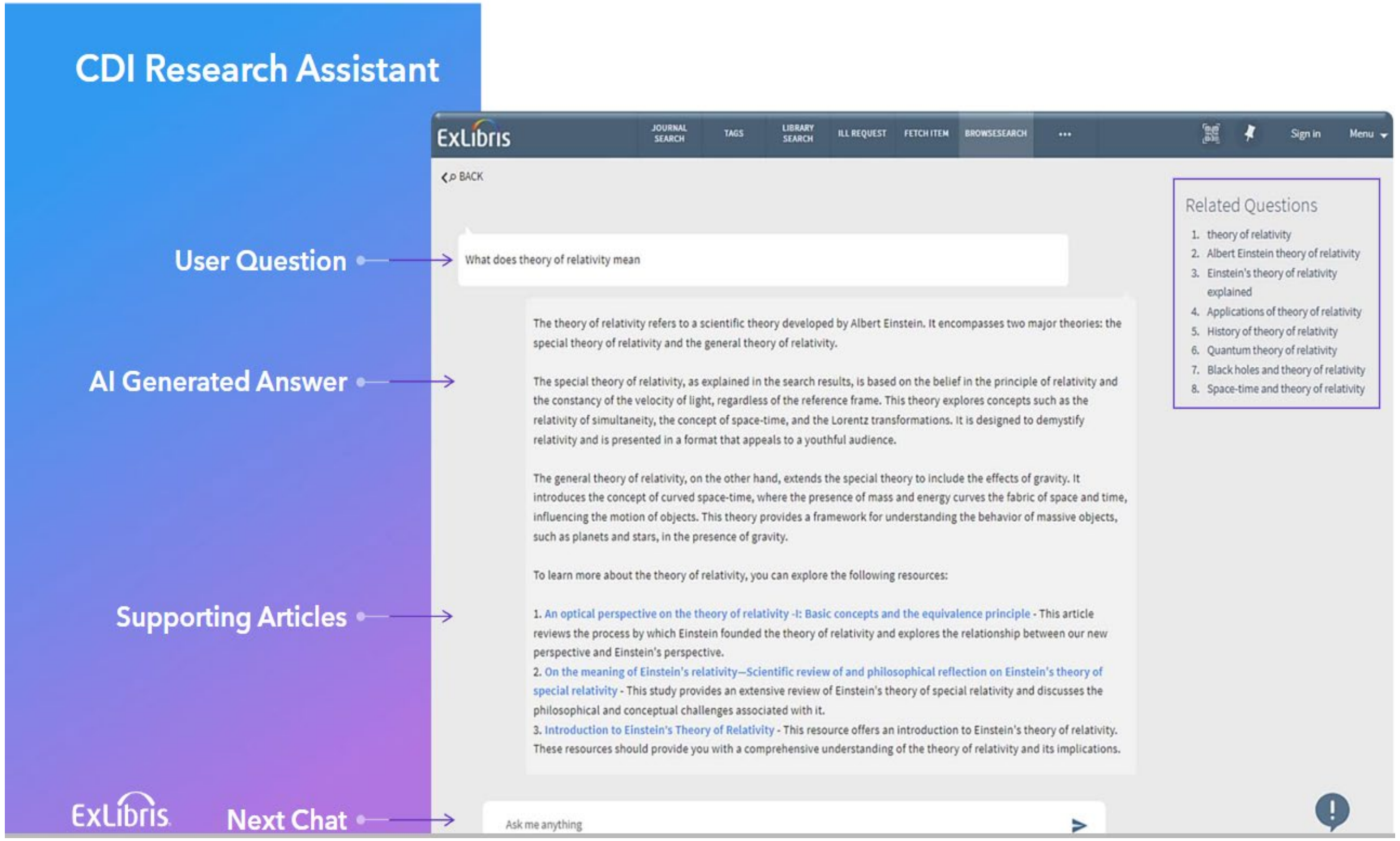
Is the coverage of google scholar enough to be used alon...

indicating it could locate all the necessary studies for specific systematic reviews [”]

.

Many , Many more…

- Scopus AI
- Primo and Summon (coming 2Q 2024)
- OpenRead
- ReadWonder
- Epslion



What does theory of relativity mean

The theory of relativity refers to a scientific theory developed by Albert Einstein. It encompasses two major theories: the special theory of relativity and the general theory of relativity.

The special theory of relativity, as explained in the search results, is based on the belief in the principle of relativity and the constancy of the velocity of light, regardless of the reference frame. This theory explores concepts such as the relativity of simultaneity, the concept of space-time, and the Lorentz transformations. It is designed to demystify relativity and is presented in a format that appeals to a youthful audience.

The general theory of relativity, on the other hand, extends the special theory to include the effects of gravity. It introduces the concept of curved space-time, where the presence of mass and energy curves the fabric of space and time, influencing the motion of objects. This theory provides a framework for understanding the behavior of massive objects, such as planets and stars, in the presence of gravity.

To learn more about the theory of relativity, you can explore the following resources:

1. An optical perspective on the theory of relativity -I: Basic concepts and the equivalence principle - This article reviews the process by which Einstein founded the theory of relativity and explores the relationship between our new perspective and Einstein's perspective.
2. On the meaning of Einstein's relativity—Scientific review of and philosophical reflection on Einstein's theory of special relativity - This study provides an extensive review of Einstein's theory of special relativity and discusses the philosophical and conceptual challenges associated with it.
3. Introduction to Einstein's Theory of Relativity - This resource offers an introduction to Einstein's theory of relativity.

These resources should provide you with a comprehensive understanding of the theory of relativity and its implications.

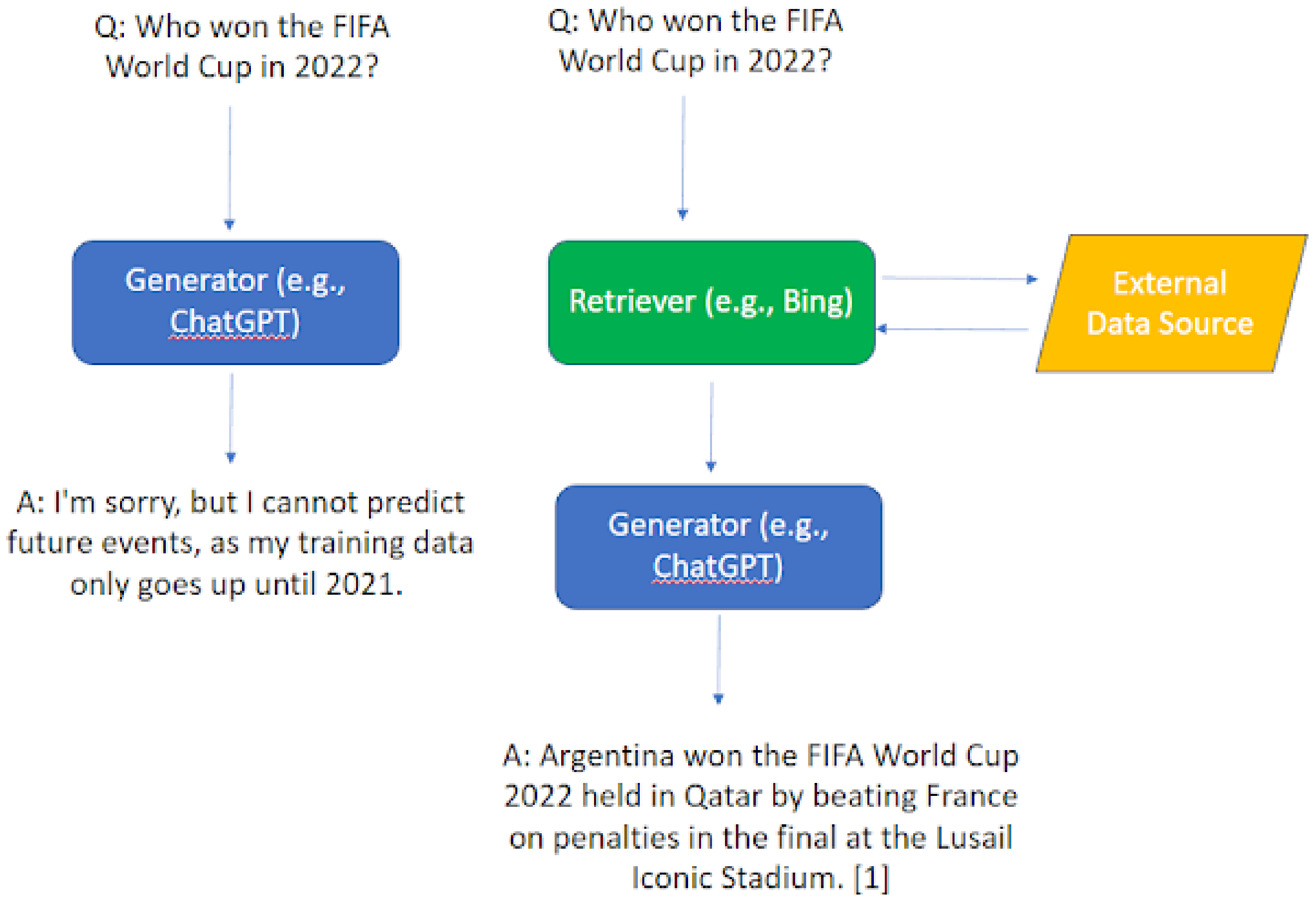
Related Questions

1. theory of relativity
2. Albert Einstein theory of relativity
3. Einstein's theory of relativity explained
4. Applications of theory of relativity
5. History of theory of relativity
6. Quantum theory of relativity
7. Black holes and theory of relativity
8. Space-time and theory of relativity

Primo Research Assistant – 2Q 2024 (on Roadmap)

Retriever augmented generation (RAG)

Idea : Let's ground the generated answer with evidence found via search of external sources (ie context)



Q: Who won the FIFA World Cup in 2022?

Generator (e.g., ChatGPT)

A: I'm sorry, but I cannot predict future events, as my training data only goes up until 2021.

Q: Who won the FIFA World Cup in 2022?

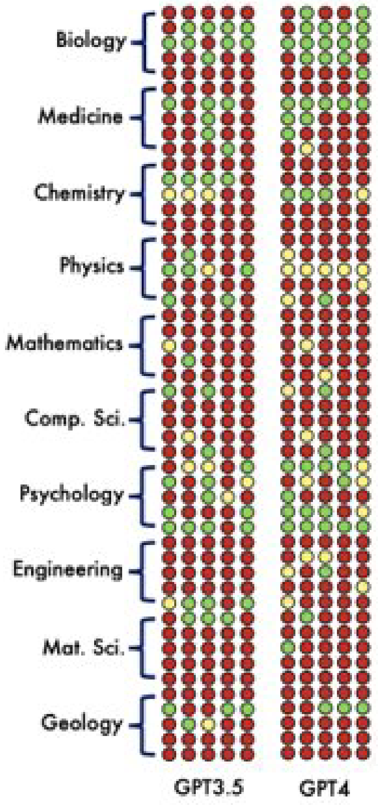
Retriever (e.g., Bing)

External Data Source

Generator (e.g., ChatGPT)

A: Argentina won the FIFA World Cup 2022 held in Qatar by beating France on penalties in the final at the Lusail Iconic Stadium. [1]

ChatGPT (free) without search might make up references

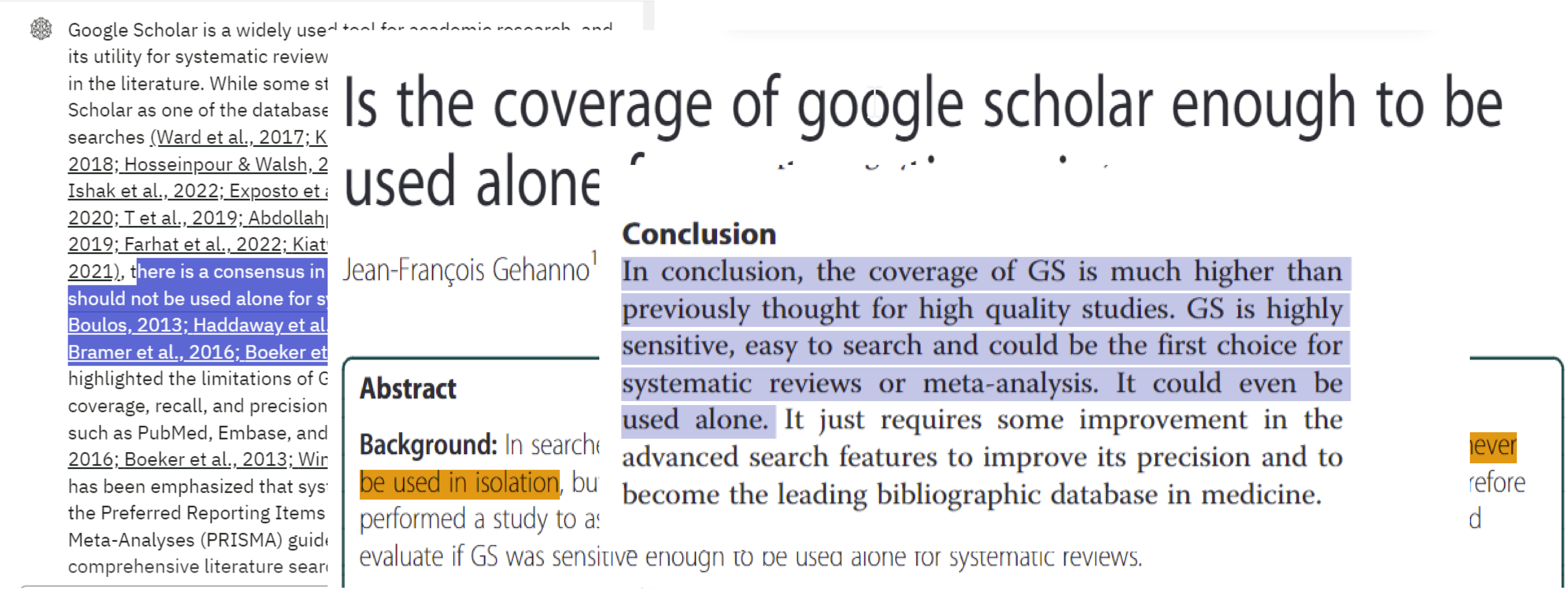


	GPT3.5	GPT4
🔴 Fictional	72.5%	71.2%
🟡 Conflated	5.5%	9.2%
🟢 Factual	22.0%	19.6%

Fig. 1. Citations to answers given by LLMs. Each row represents 5 sources / citations for a single answer. Overall, 72.5% of citations provided by GPT3.5 were fictional. This figure was 71.2% for GPT4

[2307.04683] CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering (arxiv.org)

Retrieval augmented generation does not mean "no hallucination"



Google Scholar is a widely used tool for academic research, and
its utility for systematic review
in the literature. While some st
Scholar as one of the database
searches (Ward et al., 2017; K
2018; Hosseinpour & Walsh, 2
Ishak et al., 2022; Exposto et a
2020; T et al., 2019; Abdollah
2019; Farhat et al., 2022; Kiat
2021), there is a consensus in
should not be used alone for s
Boulos, 2013; Haddaway et al.
Bramer et al., 2016; Boeker et
highlighted the limitations of G
coverage, recall, and precision
such as PubMed, Embase, and
2016; Boeker et al., 2013; Wir
has been emphasized that sys
the Preferred Reporting Items
Meta-Analyses (PRISMA) guid
comprehensive literature sear

Is the coverage of google scholar enough to be used alone

Jean-François Gehanno[1]

Abstract

Background: In search
be used in isolation, bu
performed a study to as
evaluate if GS was sensitive enough to be used alone for systematic reviews.

never
refore
d

Conclusion

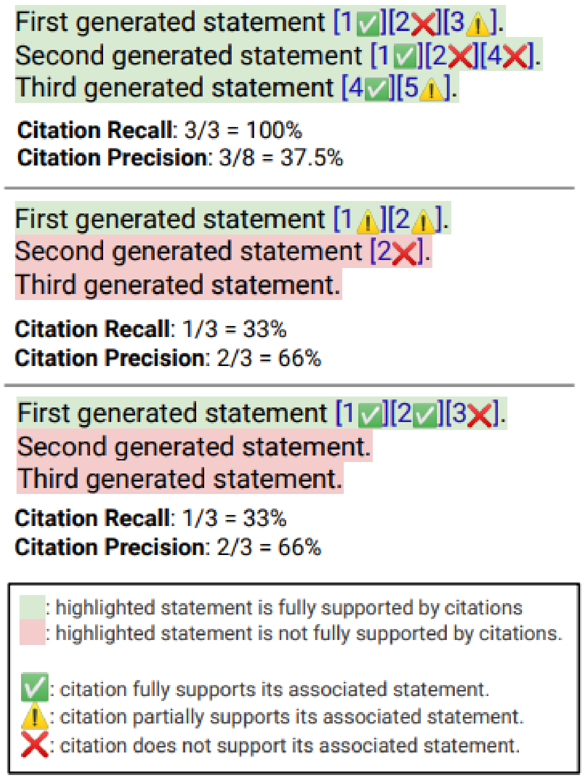
In conclusion, the coverage of GS is much higher than previously thought for high quality studies. GS is highly sensitive, easy to search and could be the first choice for systematic reviews or meta-analysis. It could even be used alone. It just requires some improvement in the advanced search features to improve its precision and to become the leading bibliographic database in medicine.

What does the research say?

Paper comparing Bing Chat, Perplexity, Youchat and NeevaAI

51.5% of generated sentences are fully supported by citations and only **74.5%** of citations support their associated sentence. (Average of 4 search engines)

68.7% of generated sentences are fully supported by citations and only **89.5%** of citations support their associated sentence. (Best of 4 search engines)



First generated statement [1✅][2❌][3⚠].
Second generated statement [1✅][2❌][4❌].
Third generated statement [4✅][5⚠].

Citation Recall: 3/3 = 100%
Citation Precision: 3/8 = 37.5%

First generated statement [1⚠][2⚠].
Second generated statement [2❌].
Third generated statement.

Citation Recall: 1/3 = 33%
Citation Precision: 2/3 = 66%

First generated statement [1✅][2✅][3❌].
Second generated statement.
Third generated statement.

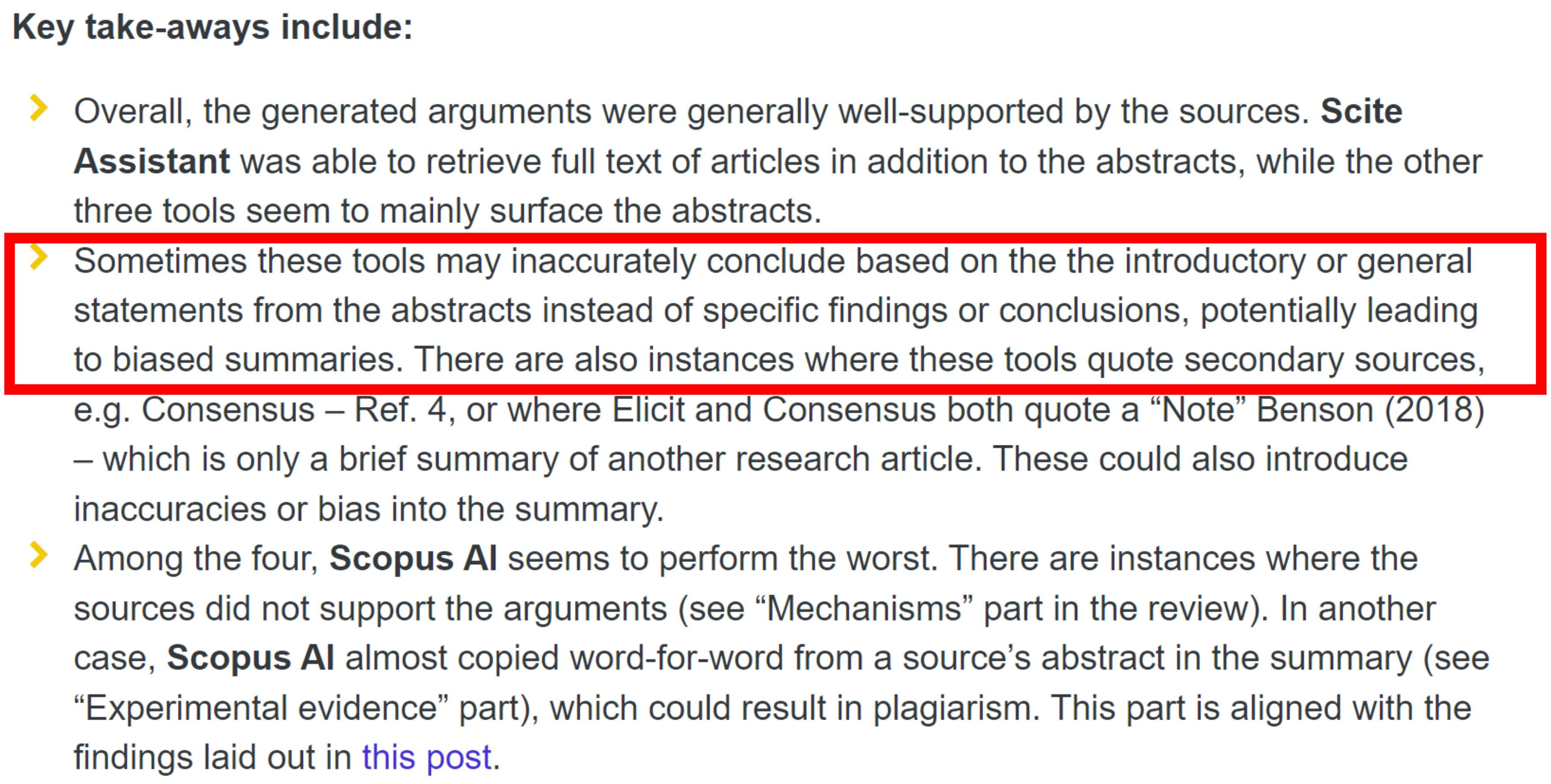
Citation Recall: 1/3 = 33%
Citation Precision: 2/3 = 66%

🟩: highlighted statement is fully supported by citations
🟥: highlighted statement is not fully supported by citations.

✅: citation fully supports its associated statement.
⚠: citation partially supports its associated statement.
❌: citation does not support its associated statement.

Liu, N. F., Zhang, T., & Liang, P. (2023). *Evaluating verifiability in generative search engines*. arXiv. http://arxiv.org/abs/2304.09848

Trust in AI: Evaluating Scite, Elicit, Consensus, and Scopus AI for Generating Literature Reviews



Key take-aways include:

- Overall, the generated arguments were generally well-supported by the sources. **Scite Assistant** was able to retrieve full text of articles in addition to the abstracts, while the other three tools seem to mainly surface the abstracts.
- Sometimes these tools may inaccurately conclude based on the the introductory or general statements from the abstracts instead of specific findings or conclusions, potentially leading to biased summaries. There are also instances where these tools quote secondary sources, e.g. Consensus – Ref. 4, or where Elicit and Consensus both quote a “Note” Benson (2018) – which is only a brief summary of another research article. These could also introduce inaccuracies or bias into the summary.
- Among the four, **Scopus AI** seems to perform the worst. There are instances where the sources did not support the arguments (see “Mechanisms” part in the review). In another case, **Scopus AI** almost copied word-for-word from a source’s abstract in the summary (see “Experimental evidence” part), which could result in plagiarism. This part is aligned with the findings laid out in this post.

https://library.hkust.edu.hk/sc/trust-ai-lit-rev/

Many things can go wrong

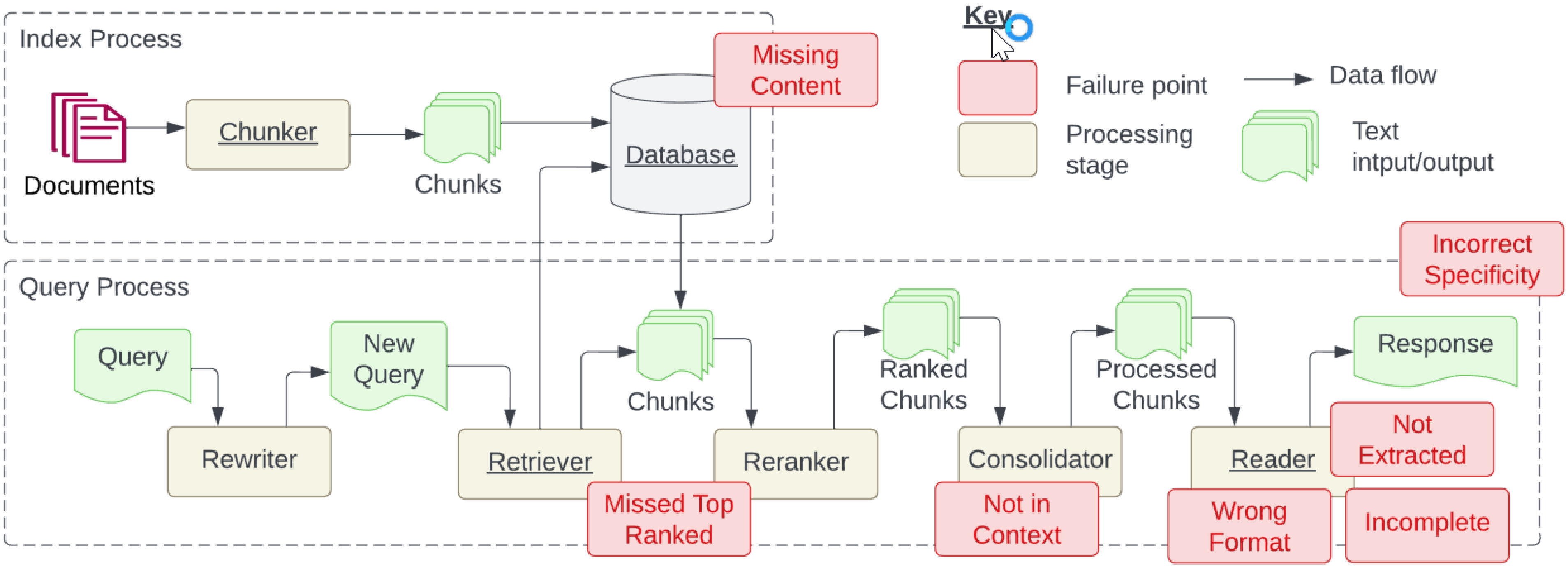
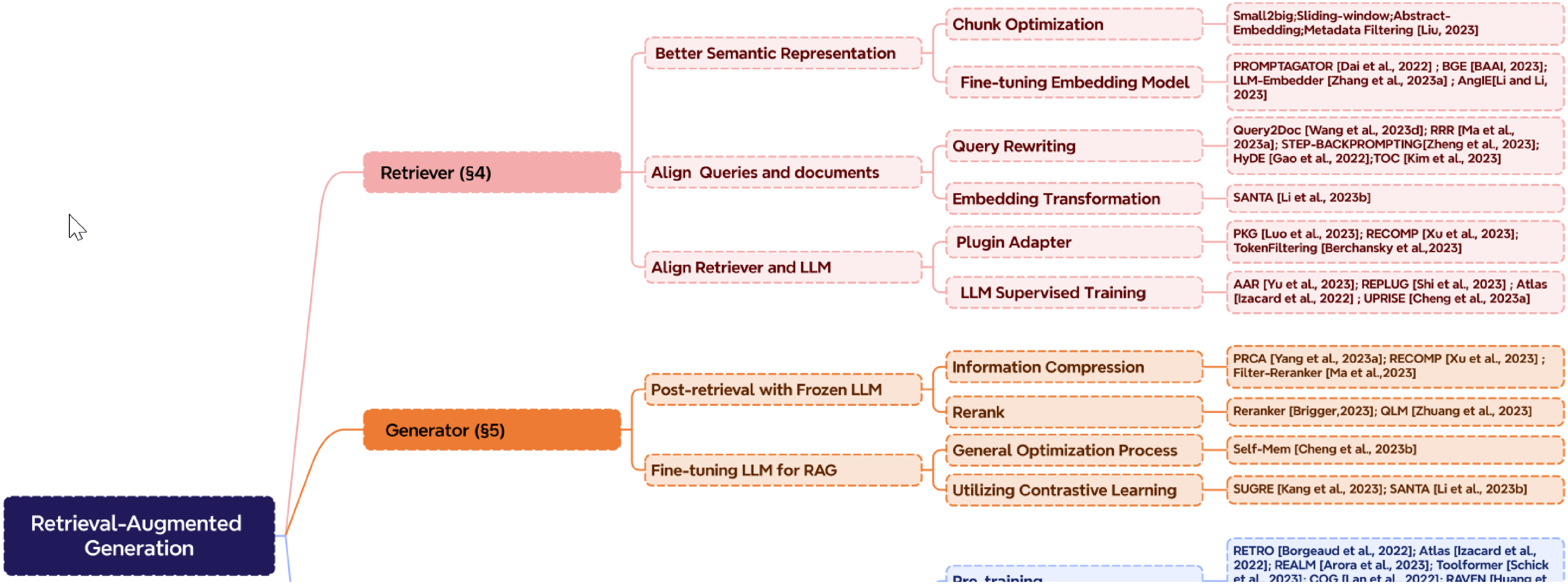


Figure 1: Indexing and Query processes required for creating a Retrieval Augmented Generation (RAG) system. The indexing process is typically done at development time and queries at runtime. Failure points identified in this study are shown in red boxes. All required stages are underlined. Figure expanded from [19].

https://arxiv.org/pdf/2401.05856.pdf

Many workarounds and solutions.. Active research area



- Retrieval-Augmented Generation
  - Retriever (§4)
    - Better Semantic Representation
      - Chunk Optimization
        - Small2big;Sliding-window;Abstract-Embedding;Metadata Filtering [Liu, 2023]
      - Fine-tuning Embedding Model
        - PROMPTAGATOR [Dai et al., 2022] ; BGE [BAAI, 2023]; LLM-Embedder [Zhang et al., 2023a] ; AnglE[Li and Li, 2023]
    - Align Queries and documents
      - Query Rewriting
        - Query2Doc [Wang et al., 2023d]; RRR [Ma et al., 2023a]; STEP-BACKPROMPTING[Zheng et al., 2023]; HyDE [Gao et al., 2022];TOC [Kim et al., 2023]
      - Embedding Transformation
        - SANTA [Li et al., 2023b]
    - Align Retriever and LLM
      - Plugin Adapter
        - PKG [Luo et al., 2023]; RECOMP [Xu et al., 2023]; TokenFiltering [Berchansky et al.,2023]
      - LLM Supervised Training
        - AAR [Yu et al., 2023]; REPLUG [Shi et al., 2023] ; Atlas [Izacard et al., 2022] ; UPRISE [Cheng et al., 2023a]
  - Generator (§5)
    - Post-retrieval with Frozen LLM
      - Information Compression
        - PRCA [Yang et al., 2023a]; RECOMP [Xu et al., 2023] ; Filter-Reranker [Ma et al.,2023]
      - Rerank
        - Reranker [Brigger,2023]; QLM [Zhuang et al., 2023]
    - Fine-tuning LLM for RAG
      - General Optimization Process
        - Self-Mem [Cheng et al., 2023b]
      - Utilizing Contrastive Learning
        - SUGRE [Kang et al., 2023]; SANTA [Li et al., 2023b]
  - Pre-training
    - RETRO [Borgeaud et al., 2022]; Atlas [Izacard et al., 2022]; REALM [Arora et al., 2023]; Toolformer [Schick et al., 2023]; COG [Lan et al., 2022]; RAVEN [Huang et

https://arxiv.org/pdf/2312.10997.pdf

Other thoughts on RAG generated summaries

Often does not find the important, seminal papers to summarise

- Because many tools like Elicit, SciSpace do not by default weight citations and just use topic relevance
- Some tools like scite assistant allow you to upload/select papers, select search strategies ued

Many generated summaries/answers lack coherence...

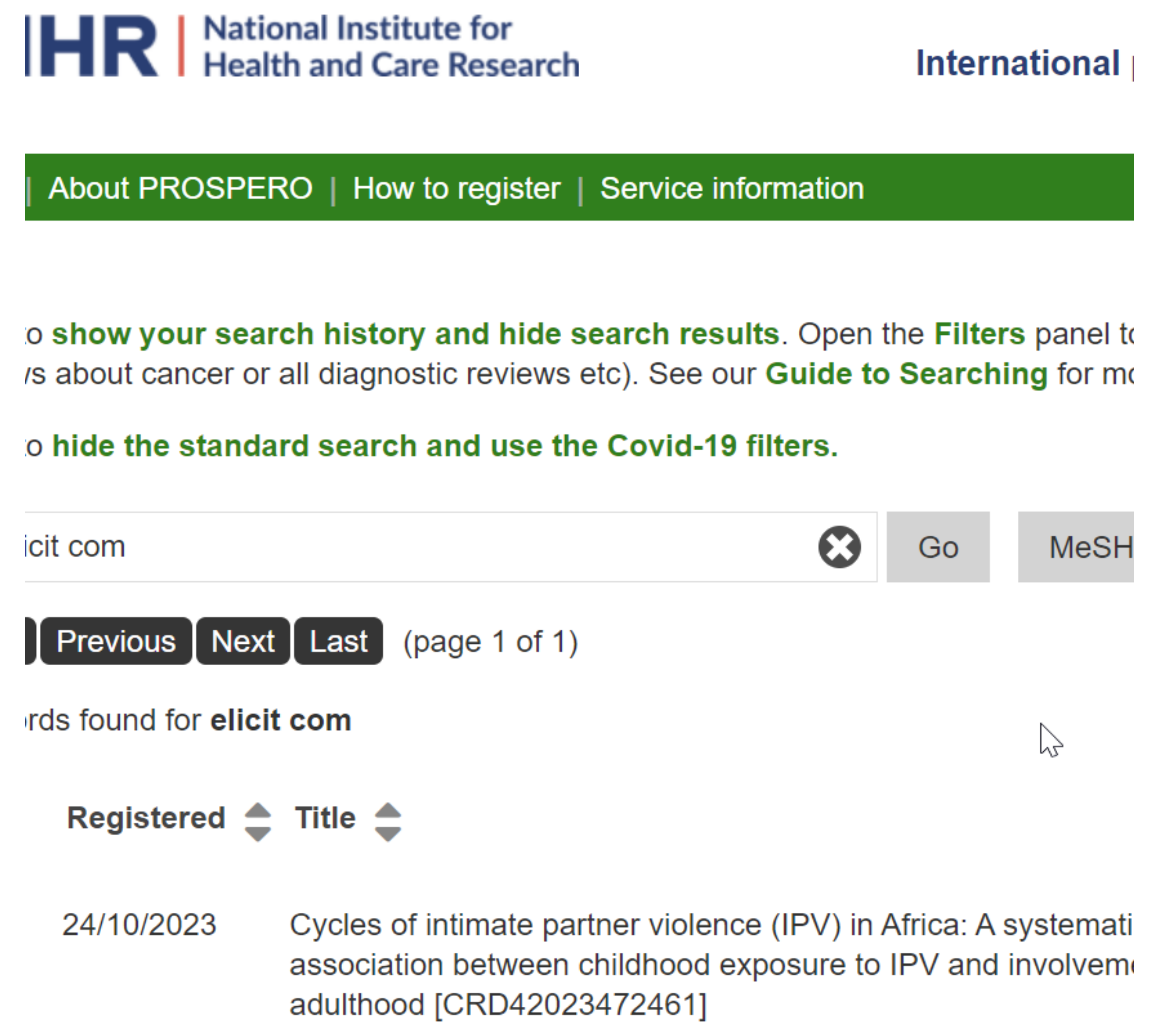
- Example "Paper A said this... Paper B said that"

May not be a big deal in academic search once people get used to it

- Accuracy will improve but academic search typically requires deep exploration not quick answers

Use of AI powered search in Evidence synthesis

- Elicit com or Elicit org (7) - 2022-2024
- SciSpace(4)- 2023-2024
- Scite (2)
- Perplexity.ai (1)
- Concensus.ai(1)
- Bing Chat (1)


IHR | National Institute for Health and Care Research
International |
| About PROSPERO | How to register | Service information
icit com
Go
MeSH
Previous
Next
Last

o **show your search history and hide search results**. Open the **Filters** panel t
's about cancer or all diagnostic reviews etc). See our **Guide to Searching** for m

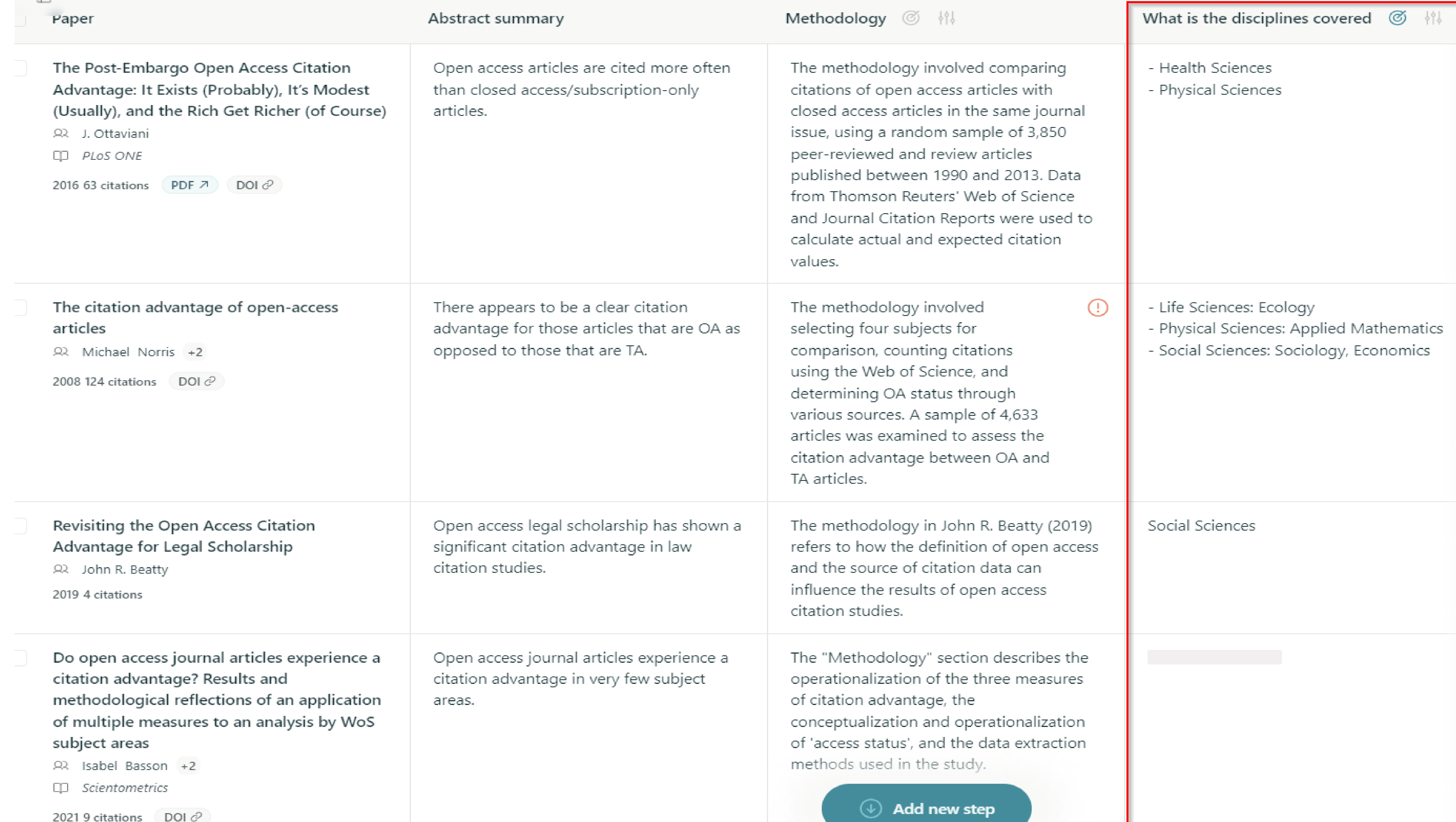
o **hide the standard search and use the Covid-19 filters.**

(page 1 of 1)

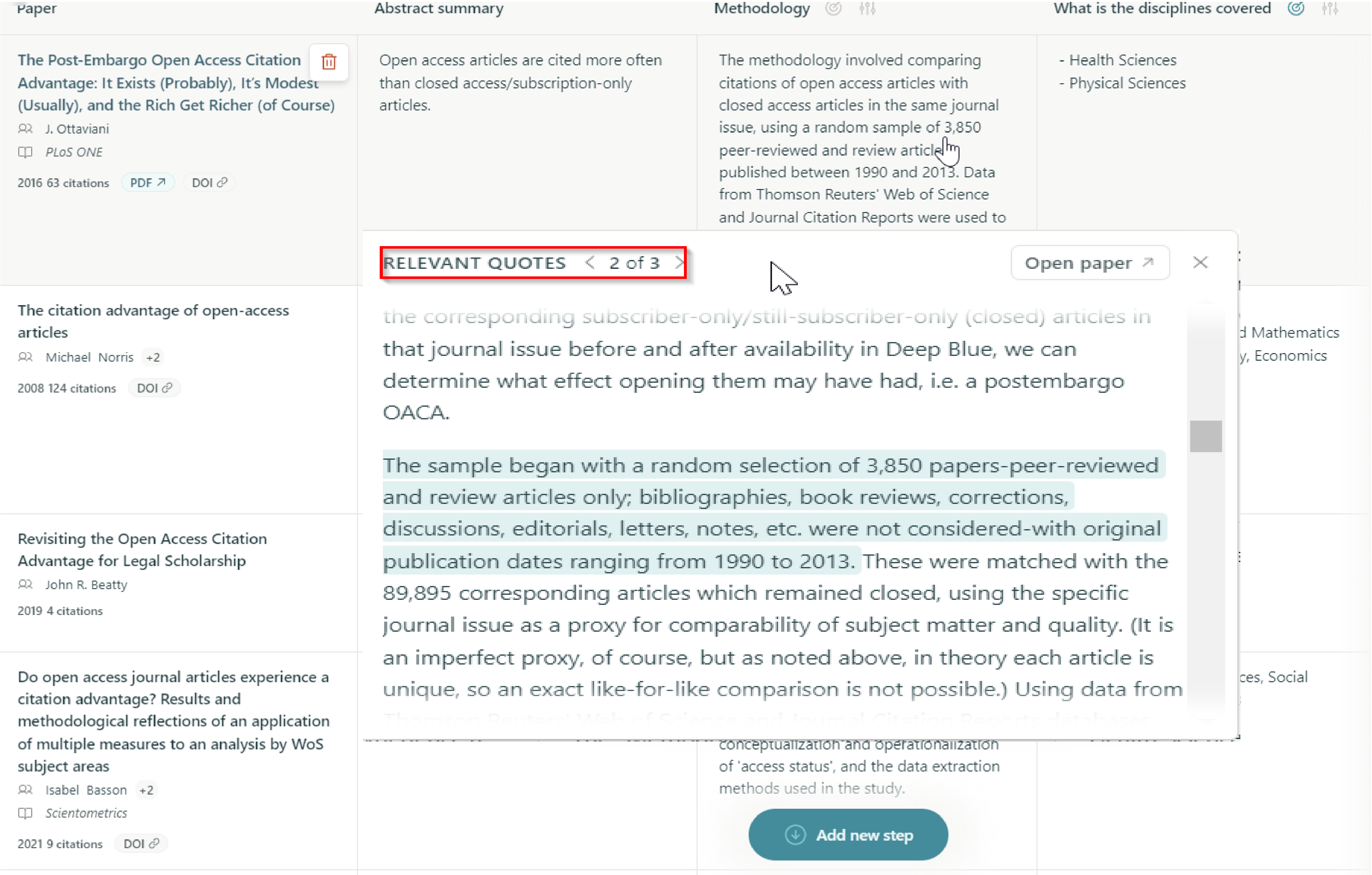
rds found for **elicit com**

Registered	Title
24/10/2023	Cycles of intimate partner violence (IPV) in Africa: A systemati association between childhood exposure to IPV and involvem adulthood [CRD42023472461]

Extraction of information from papers (abstract and full-text) to enhance search engine result pages



Paper	Abstract summary	Methodology	What is the disciplines covered
The Post-Embargo Open Access Citation Advantage: It Exists (Probably), It's Modest (Usually), and the Rich Get Richer (of Course) J. Ottaviani *PLoS ONE* 2016 63 citations PDF ↗ DOI	Open access articles are cited more often than closed access/subscription-only articles.	The methodology involved comparing citations of open access articles with closed access articles in the same journal issue, using a random sample of 3,850 peer-reviewed and review articles published between 1990 and 2013. Data from Thomson Reuters' Web of Science and Journal Citation Reports were used to calculate actual and expected citation values.	- Health Sciences - Physical Sciences
The citation advantage of open-access articles Michael Norris +2 2008 124 citations DOI	There appears to be a clear citation advantage for those articles that are OA as opposed to those that are TA.	The methodology involved selecting four subjects for comparison, counting citations using the Web of Science, and determining OA status through various sources. A sample of 4,633 articles was examined to assess the citation advantage between OA and TA articles.	- Life Sciences: Ecology - Physical Sciences: Applied Mathematics - Social Sciences: Sociology, Economics
Revisiting the Open Access Citation Advantage for Legal Scholarship John R. Beatty 2019 4 citations	Open access legal scholarship has shown a significant citation advantage in law citation studies.	The methodology in John R. Beatty (2019) refers to how the definition of open access and the source of citation data can influence the results of open access citation studies.	Social Sciences
Do open access journal articles experience a citation advantage? Results and methodological reflections of an application of multiple measures to an analysis by WoS subject areas Isabel Basson +2 *Scientometrics* 2021 9 citations DOI	Open access journal articles experience a citation advantage in very few subject areas.	The "Methodology" section describes the operationalization of the three measures of citation advantage, the conceptualization and operationalization of 'access status', and the data extraction methods used in the study.	



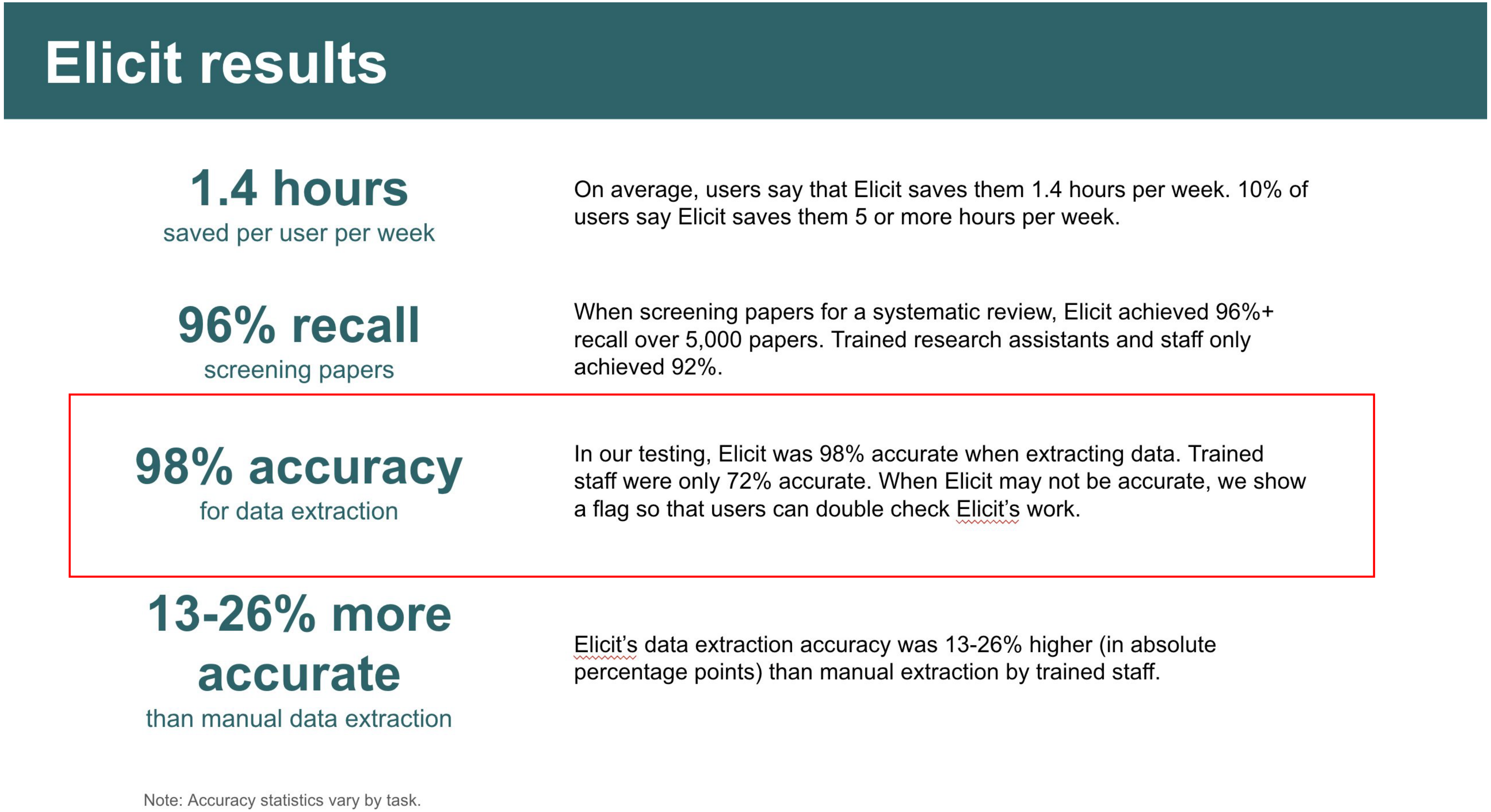
Paper	Abstract summary	Methodology	What is the disciplines covered
The Post-Embargo Open Access Citation Advantage: It Exists (Probably), It's Modest (Usually), and the Rich Get Richer (of Course) J. Ottaviani *PLoS ONE* 2016 63 citations PDF ↗ DOI	Open access articles are cited more often than closed access/subscription-only articles.	The methodology involved comparing citations of open access articles with closed access articles in the same journal issue, using a random sample of 3,850 peer-reviewed and review articles published between 1990 and 2013. Data from Thomson Reuters' Web of Science and Journal Citation Reports were used to	- Health Sciences - Physical Sciences
The citation advantage of open-access articles Michael Norris +2 2008 124 citations DOI			d Mathematics y, Economics
Revisiting the Open Access Citation Advantage for Legal Scholarship John R. Beatty 2019 4 citations			
Do open access journal articles experience a citation advantage? Results and methodological reflections of an application of multiple measures to an analysis by WoS subject areas Isabel Basson +2 *Scientometrics* 2021 9 citations DOI		conceptualization and operationalization of 'access status', and the data extraction methods used in the study.	ces, Social

RELEVANT QUOTES < 2 of 3 >

the corresponding subscriber-only/still-subscriber-only (closed) articles in that journal issue before and after availability in Deep Blue, we can determine what effect opening them may have had, i.e. a postembargo OACA.

The sample began with a random selection of 3,850 papers-peer-reviewed and review articles only; bibliographies, book reviews, corrections, discussions, editorials, letters, notes, etc. were not considered-with original publication dates ranging from 1990 to 2013. These were matched with the 89,895 corresponding articles which remained closed, using the specific journal issue as a proxy for comparability of subject matter and quality. (It is an imperfect proxy, of course, but as noted above, in theory each article is unique, so an exact like-for-like comparison is not possible.) Using data from

Is it really 98% accurate for data extraction?



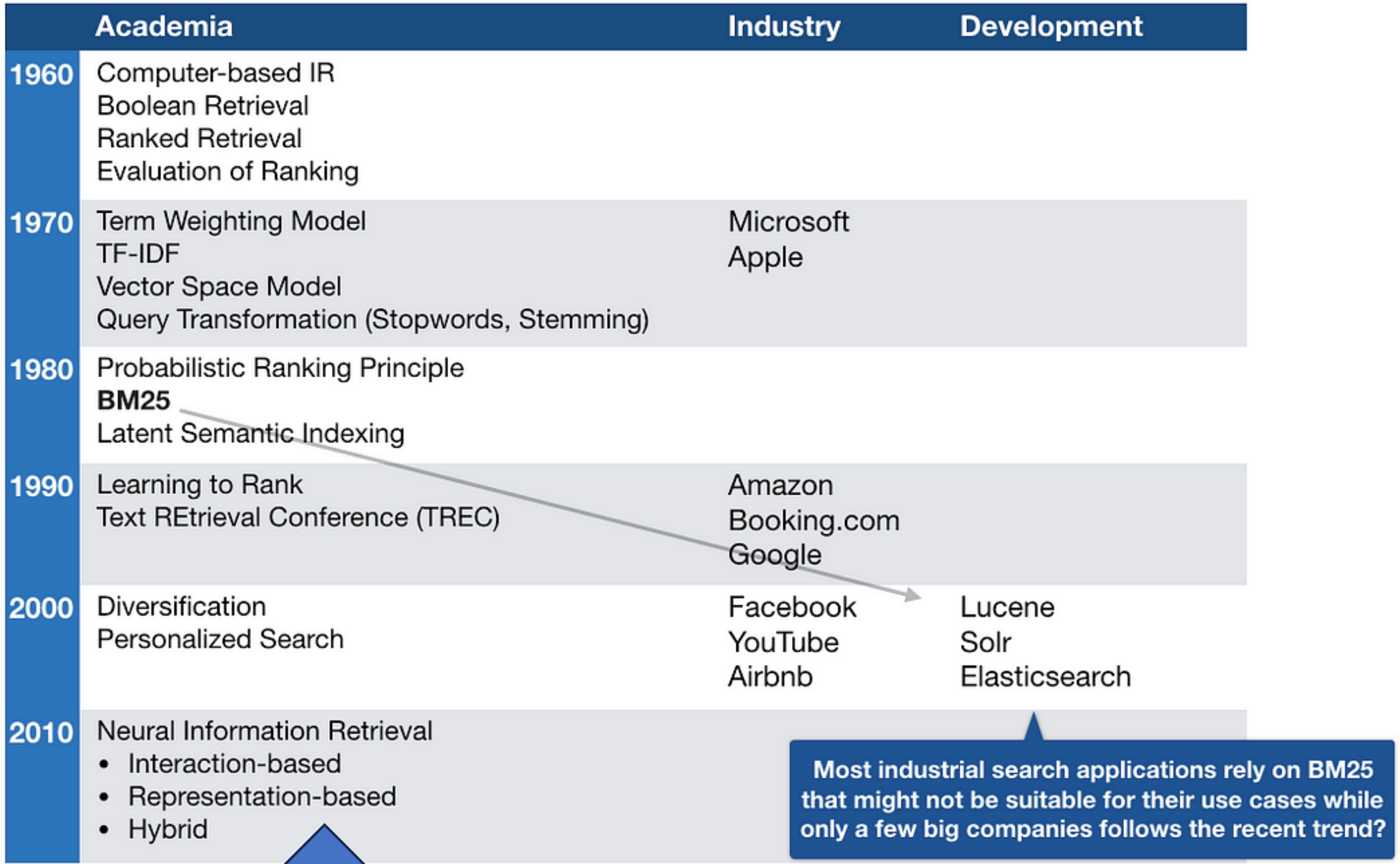
Elicit results

1.4 hours saved per user per week	On average, users say that Elicit saves them 1.4 hours per week. 10% of users say Elicit saves them 5 or more hours per week.
96% recall screening papers	When screening papers for a systematic review, Elicit achieved 96%+ recall over 5,000 papers. Trained research assistants and staff only achieved 92%.
98% accuracy for data extraction	In our testing, Elicit was 98% accurate when extracting data. Trained staff were only 72% accurate. When Elicit may not be accurate, we show a flag so that users can double check Elicit's work.
13-26% more accurate than manual data extraction	Elicit's data extraction accuracy was 13-26% higher (in absolute percentage points) than manual extraction by trained staff.

Note: Accuracy statistics vary by task.

Improved relevancy (through contextual embeddings)

History of information retrieval



	Academia	Industry	Development
1960	Computer-based IR Boolean Retrieval Ranked Retrieval Evaluation of Ranking		
1970	Term Weighting Model TF-IDF Vector Space Model Query Transformation (Stopwords, Stemming)	Microsoft Apple	
1980	Probabilistic Ranking Principle **BM25** Latent Semantic Indexing		
1990	Learning to Rank Text REtrieval Conference (TREC)	Amazon Booking.com Google	
2000	Diversification Personalized Search	Facebook YouTube Airbnb	Lucene Solr Elasticsearch
2010	Neural Information Retrieval • Interaction-based • Representation-based • Hybrid		

Most industrial search applications rely on BM25 that might not be suitable for their use cases while only a few big companies follows the recent trend?

Early use of NN – RNNs etc

2017 – "Attention is all you need" - Transformers invented

2018 - Bidirectional Encoder Representations from Transformers (BERT)

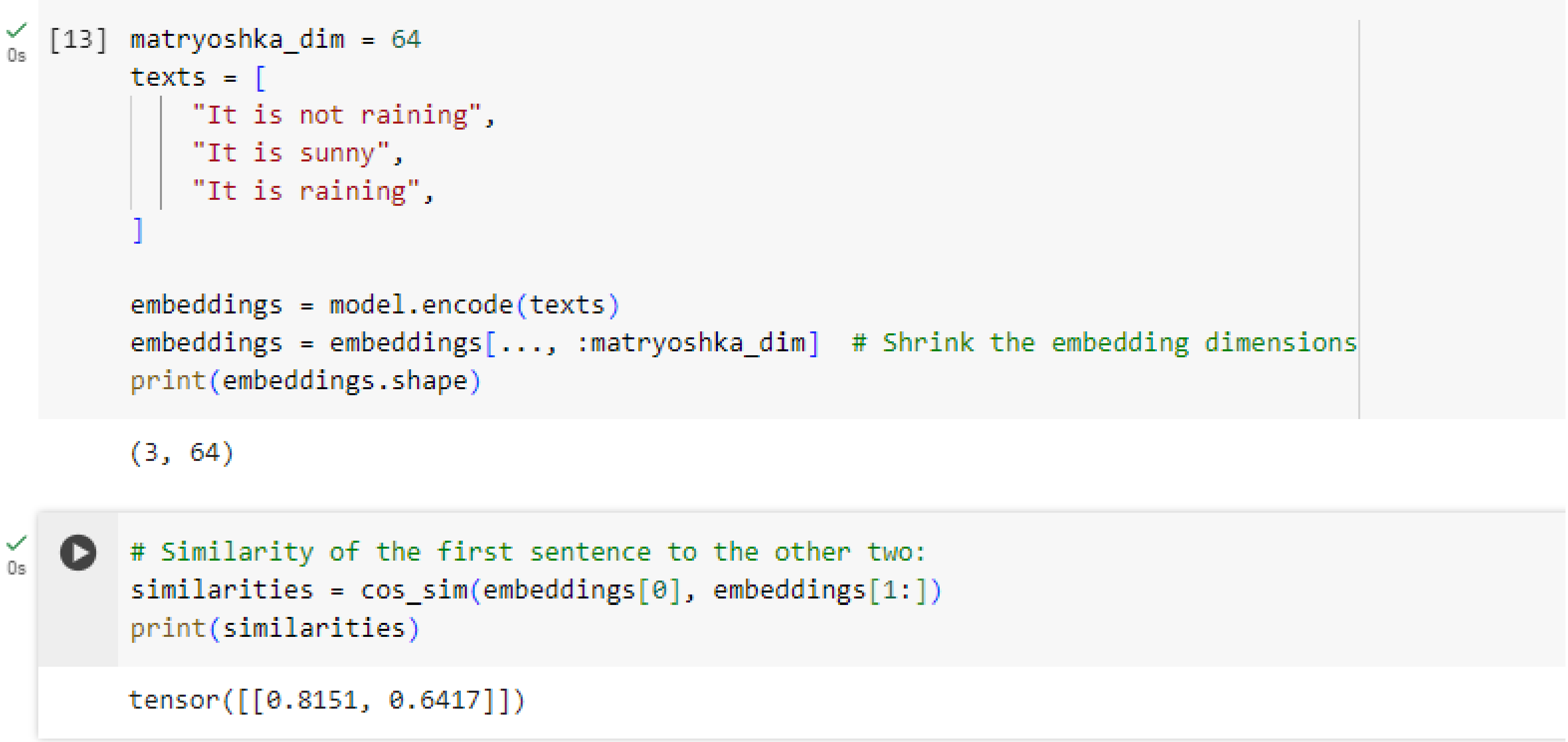
2019 – GPT invented

2020-2023 – First effective "Semantic search" appear using BERT/GPT

Semantic Search – The holy grail

- The system understands “meaning” and doesn’t just took at word by word (bag of words).
- E.g. It must be able to “understand”, “It is **not** raining” is closer to “it is sunny” and not “it is raining”
- Implies – can solve vocabulary mismatch problem (even typos)
- Not the name of a technology but describes what we want to do
- Current state of art involves using *contextual embeddings from transformer models (BERT, GPT) to do semantic search*

Some “Evidence” it works – simple embedding example



```
matryoshka_dim = 64
texts = [
    "It is not raining",
    "It is sunny",
    "It is raining",
]

embeddings = model.encode(texts)
embeddings = embeddings[..., :matryoshka_dim]  # Shrink the embedding dimensions
print(embeddings.shape)
```

```
(3, 64)
```

```
# Similarity of the first sentence to the other two:
similarities = cos_sim(embeddings[0], embeddings[1:])
print(similarities)
```

```
tensor([[0.8151, 0.6417]])
```

Similarity of “It is not raining” and “It is sunny” is 0.8151
Similarity of “It is not raining” and “It is raining” is 0.6417, which is lower despite matching more keywords
This embedding search “understands” the meaning of “not raining”!

Two main types of search algos (not counting hybrids)

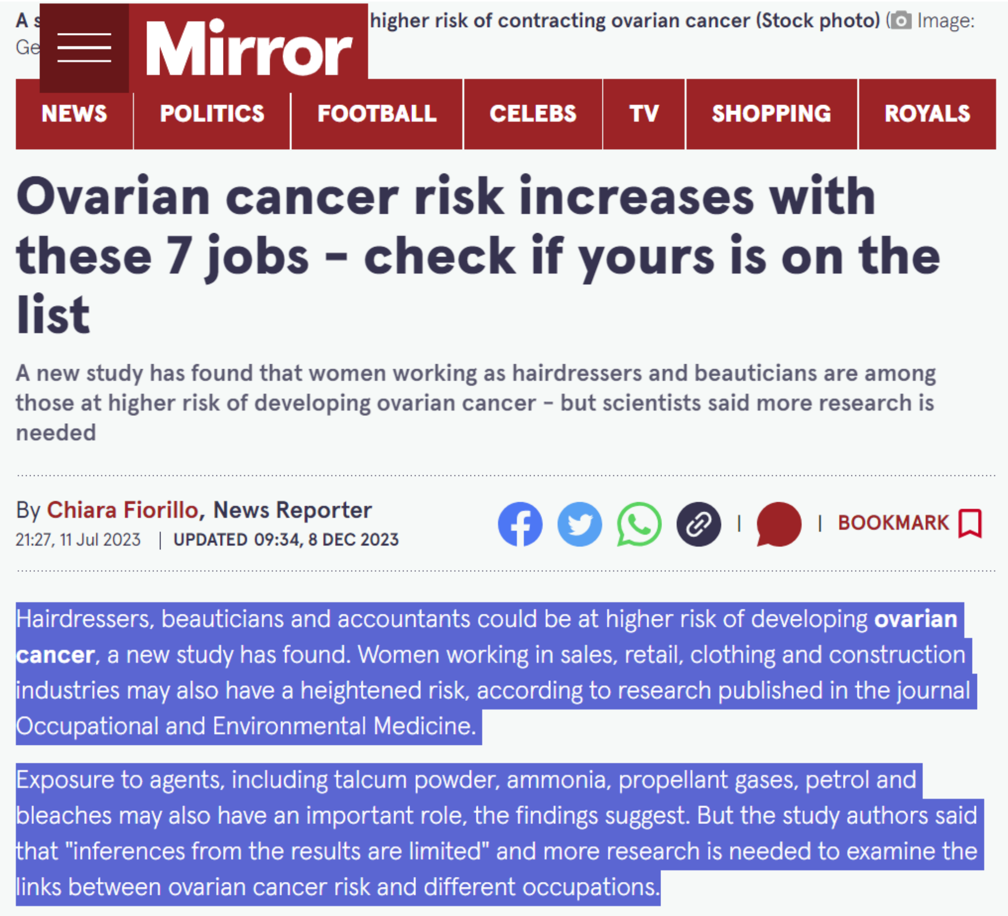
- Lexical/Keyword Search
- Bag of words
- Sparse embeddings/Representation/Retrieval

- Semantic Search/ AI search
- Neural Information Retrieval (Neural IR/Search)
- Dense embeddings/Representation/Retrieval

- Boolean
- TF-IDF/BM25 (probabilistic models)

- "Embedding based"/ "Vector search" eg ada002
- Bi-encoder/Cross-encoder/COLBERT

Finding an article by throwing in full text of news story



A s higher risk of contracting ovarian cancer (Stock photo) (Image:
Ge

Ovarian cancer risk increases with these 7 jobs - check if yours is on the list

A new study has found that women working as hairdressers and beauticians are among those at higher risk of developing ovarian cancer - but scientists said more research is needed

By **Chiara Fiorillo**, News Reporter
21:27, 11 Jul 2023 | UPDATED 09:34, 8 DEC 2023

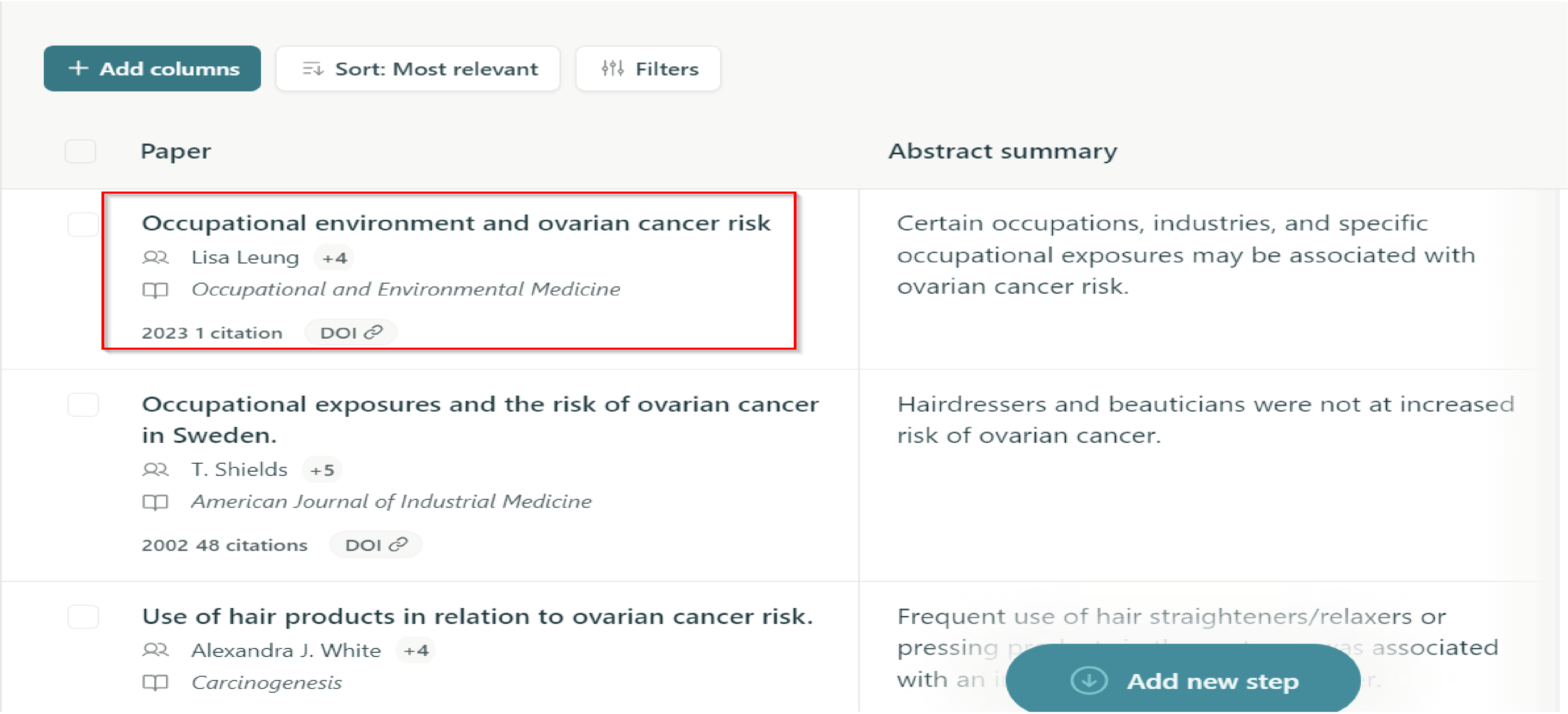
Hairdressers, beauticians and accountants could be at higher risk of developing **ovarian cancer**, a new study has found. Women working in sales, retail, clothing and construction industries may also have a heightened risk, according to research published in the journal Occupational and Environmental Medicine.

Exposure to agents, including talcum powder, ammonia, propellant gases, petrol and bleaches may also have an important role, the findings suggest. But the study authors said that "inferences from the results are limited" and more research is needed to examine the links between ovarian cancer risk and different occupations.

Hairdressers, beauticians and accountants could be at higher risk of developing **ovarian cancer**, a new study has found. Women working in sales, retail, clothing and construction industries may also have a heightened risk, according to research published in the journal Occupational and Environmental Medicine.
Exposure to agents, including talcum powder, ammonia, propellant gases, petrol and bleaches may also have an important role, the findings suggest. But the study authors said that "inferences from the results are limited" and more research is needed to examine the links between ovarian cancer risk and different occupations.

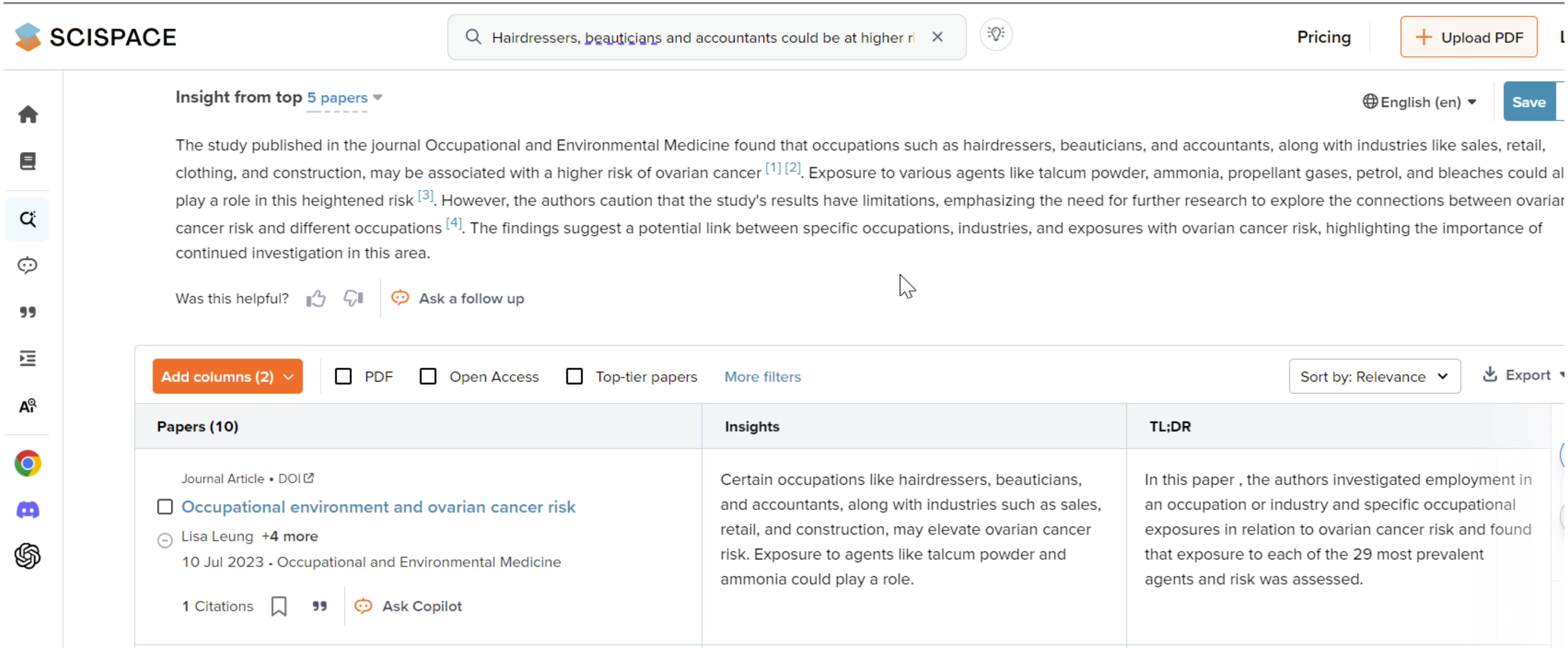
https://www.mirror.co.uk/news/health/ovarian-cancer-risk-increases-7-30444348

Elicit/Scitespace - throw in a long series of text



☐	Paper	Abstract summary
☐	**Occupational environment and ovarian cancer risk** Lisa Leung +4 *Occupational and Environmental Medicine* 2023 1 citation DOI	Certain occupations, industries, and specific occupational exposures may be associated with ovarian cancer risk.
☐	**Occupational exposures and the risk of ovarian cancer in Sweden.** T. Shields +5 *American Journal of Industrial Medicine* 2002 48 citations DOI	Hairdressers and beauticians were not at increased risk of ovarian cancer.
☐	**Use of hair products in relation to ovarian cancer risk.** Alexandra J. White +4 *Carcinogenesis*	Frequent use of hair straighteners/relaxers or pressing pr [illegible] as associated with an i [illegible] r.

Elicit/Scitespace - throw in a long series of text



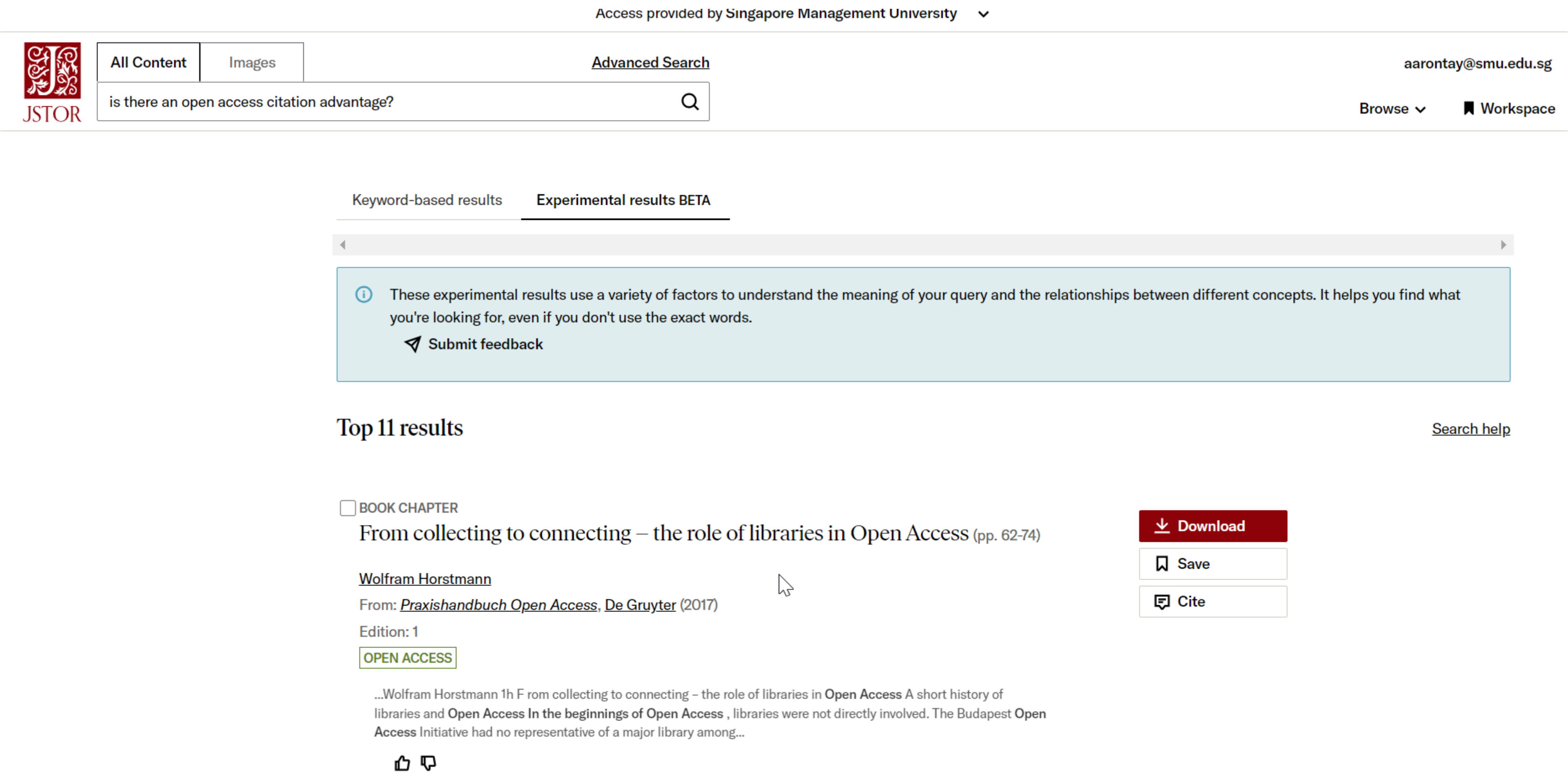
Hairdressers, beauticians and accountants could be at higher ri

Insight from top 5 papers

The study published in the journal Occupational and Environmental Medicine found that occupations such as hairdressers, beauticians, and accountants, along with industries like sales, retail, clothing, and construction, may be associated with a higher risk of ovarian cancer [1] [2]. Exposure to various agents like talcum powder, ammonia, propellant gases, petrol, and bleaches could al play a role in this heightened risk [3]. However, the authors caution that the study's results have limitations, emphasizing the need for further research to explore the connections between ovarian cancer risk and different occupations [4]. The findings suggest a potential link between specific occupations, industries, and exposures with ovarian cancer risk, highlighting the importance of continued investigation in this area.

Papers (10)	Insights	TL;DR
Journal Article • DOI ☐ Occupational environment and ovarian cancer risk Lisa Leung +4 more 10 Jul 2023 - Occupational and Environmental Medicine 1 Citations Ask Copilot	Certain occupations like hairdressers, beauticians, and accountants, along with industries such as sales, retail, and construction, may elevate ovarian cancer risk. Exposure to agents like talcum powder and ammonia could play a role.	In this paper , the authors investigated employment in an occupation or industry and specific occupational exposures in relation to ovarian cancer risk and found that exposure to each of the 29 most prevalent agents and risk was assessed.

JSTOR experimental search



is there an open access citation advantage?

These experimental results use a variety of factors to understand the meaning of your query and the relationships between different concepts. It helps you find what you're looking for, even if you don't use the exact words.

Top 11 results

☐ BOOK CHAPTER

From collecting to connecting – the role of libraries in Open Access (pp. 62-74)

Wolfram Horstmann

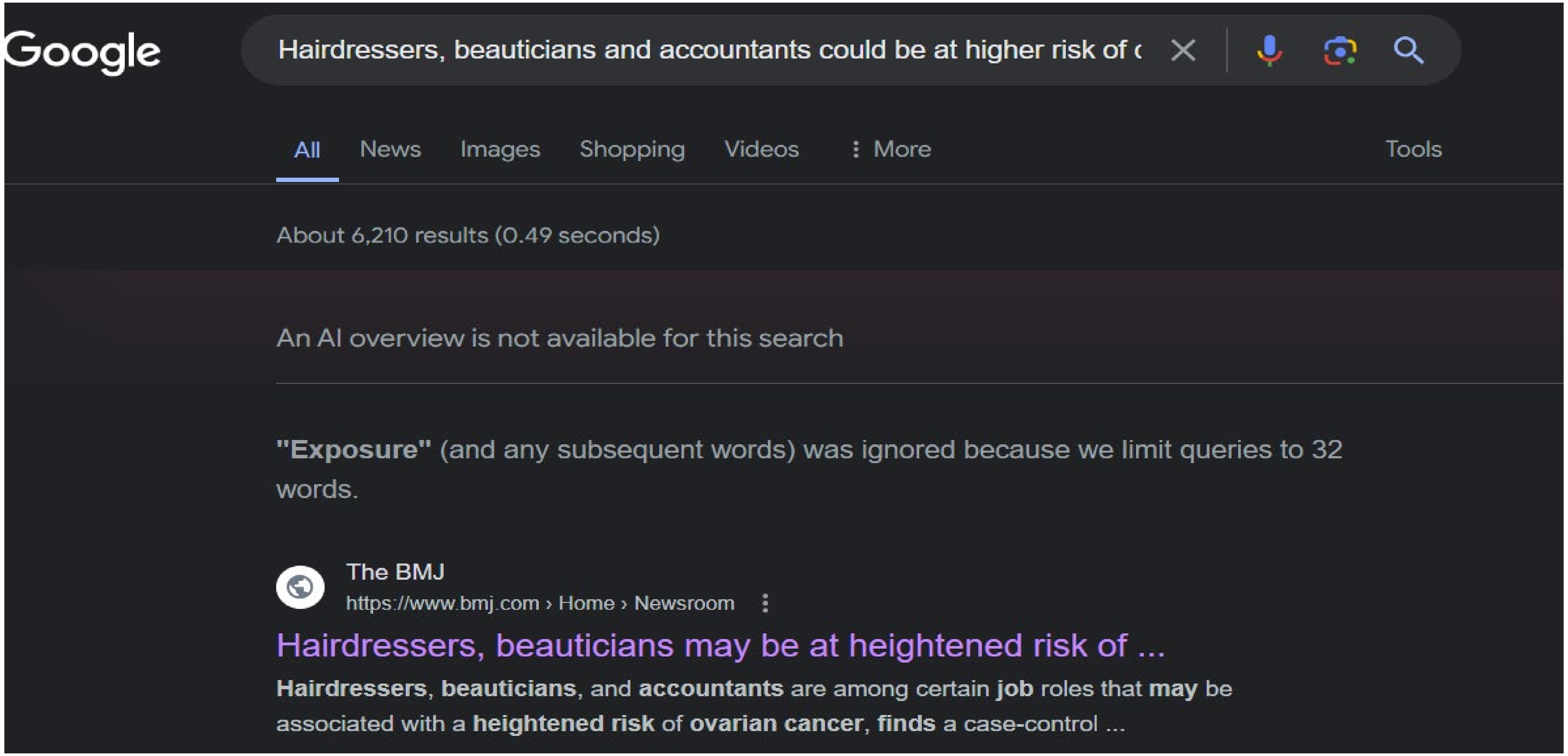
From: *Praxishandbuch Open Access*, De Gruyter (2017)

Edition: 1

OPEN ACCESS

...Wolfram Horstmann 1h F rom collecting to connecting – the role of libraries in **Open Access** A short history of libraries and **Open Access In the beginnings of Open Access** , libraries were not directly involved. The Budapest **Open Access** Initiative had no representative of a major library among...

Google works too


Google
Hairdressers, beauticians and accountants could be at higher risk of c
All
News
Images
Shopping
Videos
More
Tools

About 6,210 results (0.49 seconds)

An AI overview is not available for this search

"Exposure" (and any subsequent words) was ignored because we limit queries to 32 words.

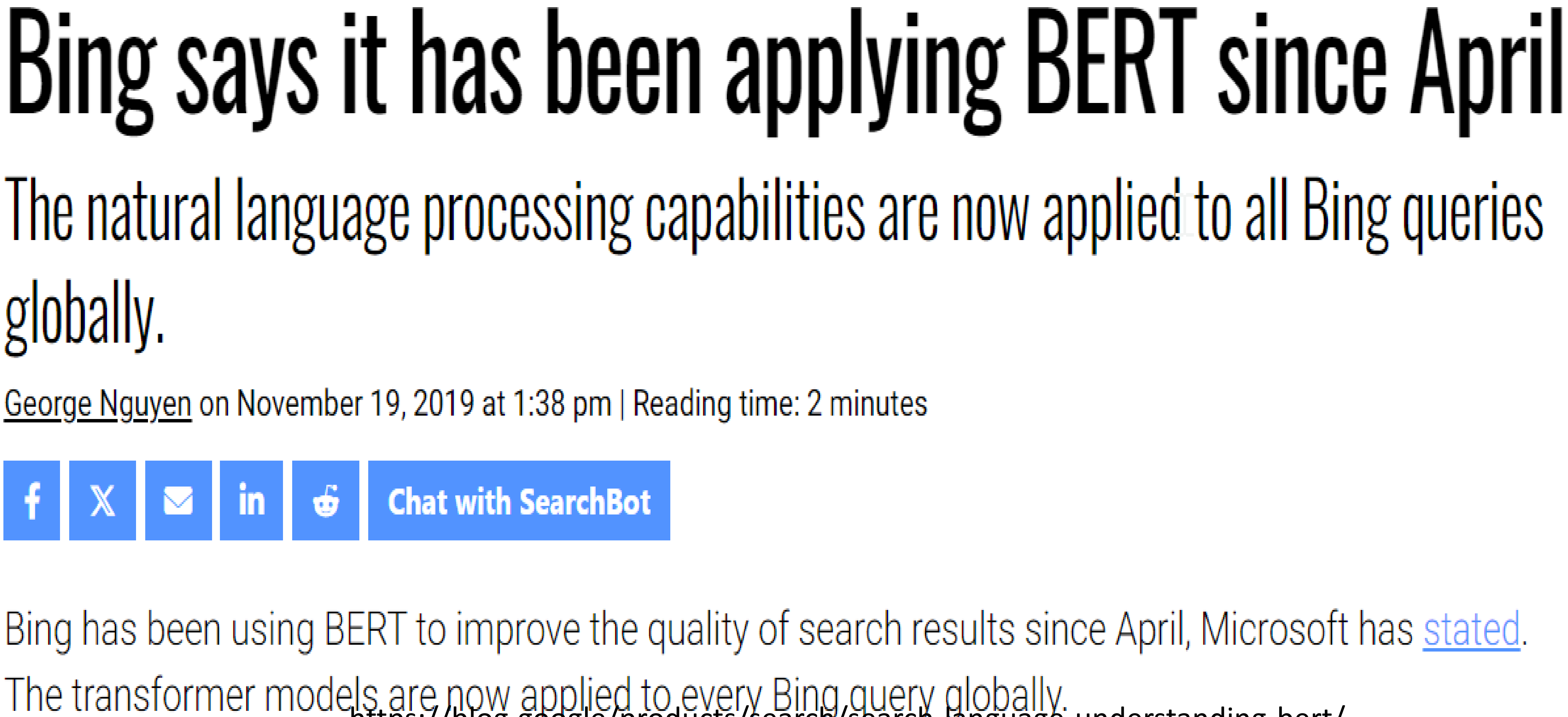
The BMJ
https://www.bmj.com › Home › Newsroom

Hairdressers, beauticians may be at heightened risk of ...

Hairdressers, **beauticians**, and **accountants** are among certain **job** roles that **may** be associated with a **heightened risk** of **ovarian cancer**, **finds** a case-control ...

[Google search]

BERT models applied to Bing/Google in 2019



Bing says it has been applying BERT since April

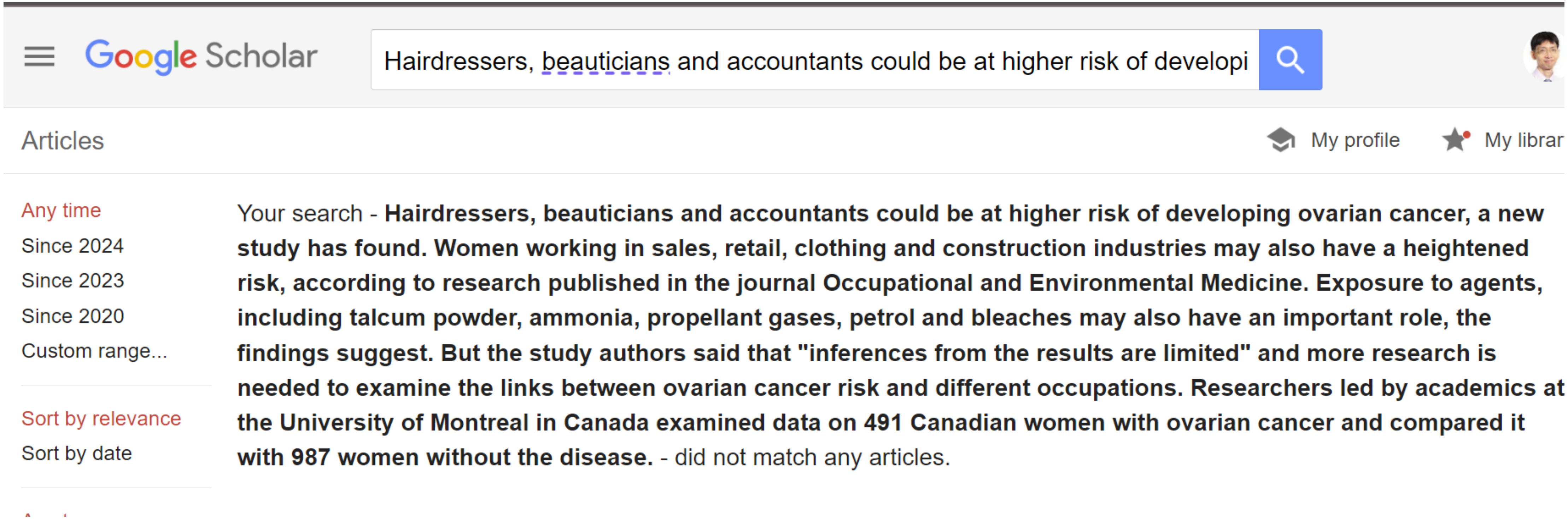
The natural language processing capabilities are now applied to all Bing queries globally.

George Nguyen on November 19, 2019 at 1:38 pm | Reading time: 2 minutes

Bing has been using BERT to improve the quality of search results since April, Microsoft has stated. The transformer models are now applied to every Bing query globally.

https://blog.google/products/search/search-language-understanding-bert/

But Google scholar fails



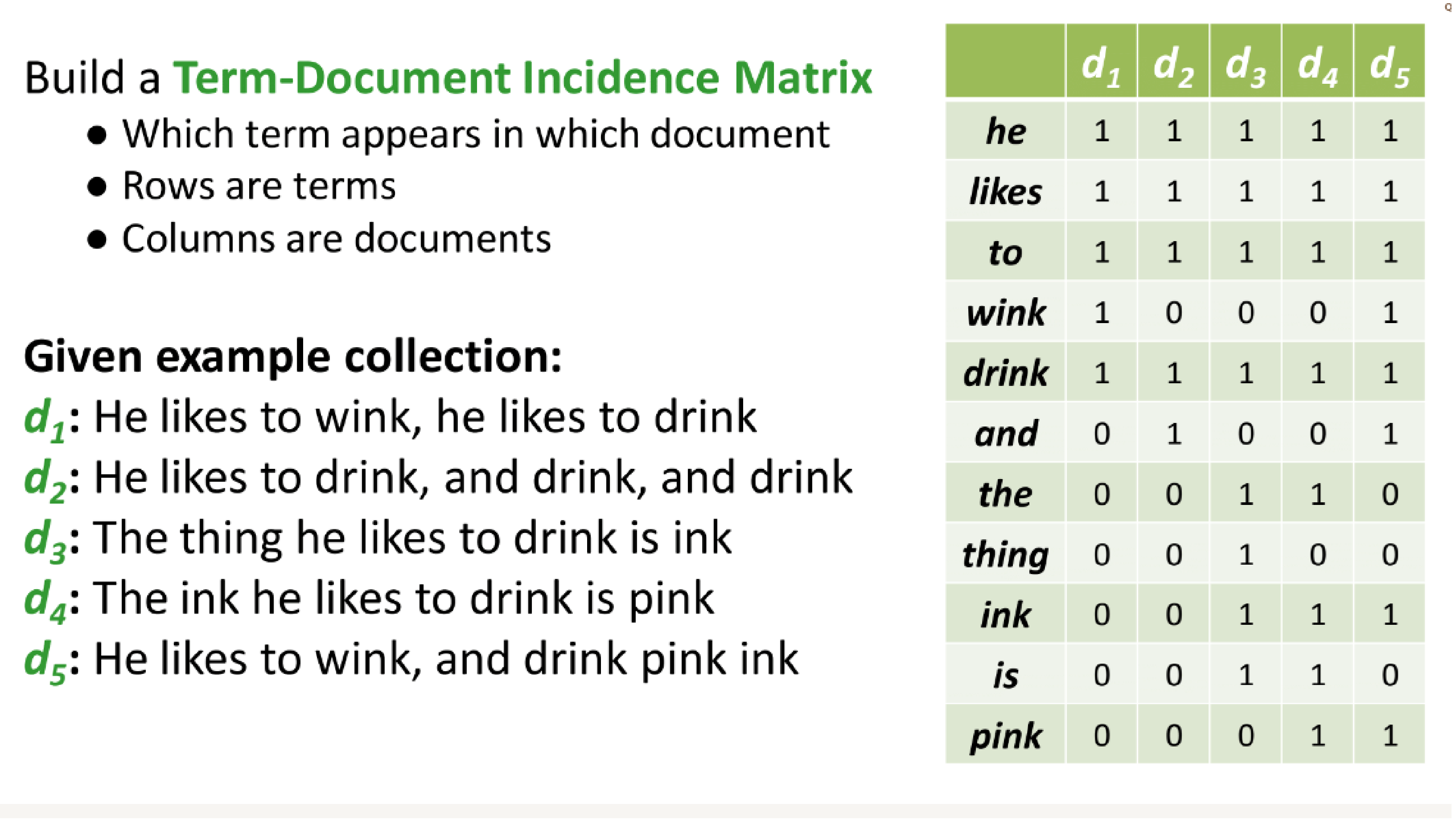
Your search - **Hairdressers, beauticians and accountants could be at higher risk of developing ovarian cancer, a new study has found. Women working in sales, retail, clothing and construction industries may also have a heightened risk, according to research published in the journal Occupational and Environmental Medicine. Exposure to agents, including talcum powder, ammonia, propellant gases, petrol and bleaches may also have an important role, the findings suggest. But the study authors said that "inferences from the results are limited" and more research is needed to examine the links between ovarian cancer risk and different occupations. Researchers led by academics at the University of Montreal in Canada examined data on 491 Canadian women with ovarian cancer and compared it with 987 women without the disease.** - did not match any articles.

So how does "Semantic Search" vs Lexical search work?

Boolean Retrieval Model

- **Queries**: Users express queries as a ***Boolean expression***
  - AND, OR, NOT
  - Can be arbitrarily nested
- Ex. query: ***Qatar*** *AND* ***University*** *AND NOT* ***Street***

Boolean represented as Term-Document matrix



Build a **Term-Document Incidence Matrix**

- Which term appears in which document
- Rows are terms
- Columns are documents

Given example collection:

d_1: He likes to wink, he likes to drink

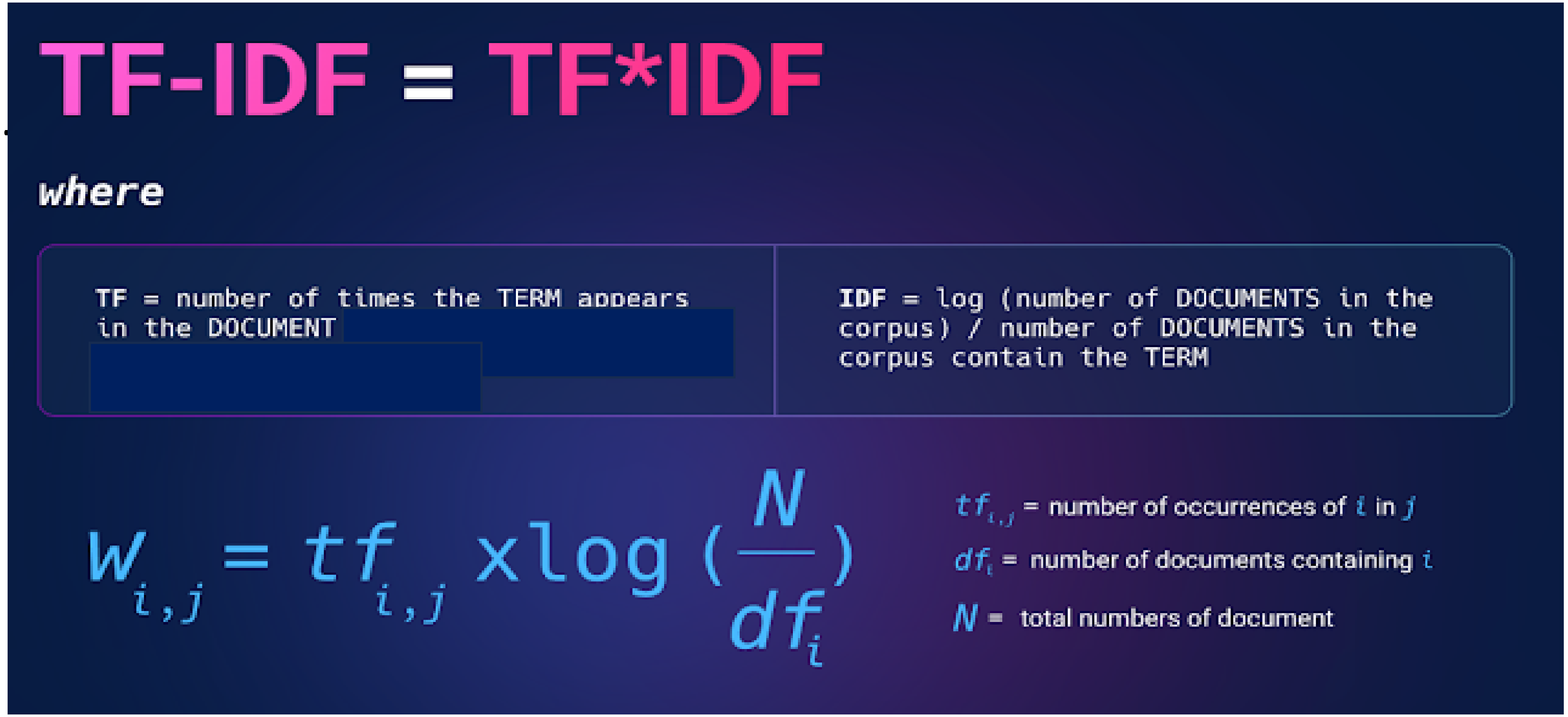
d_2: He likes to drink, and drink, and drink

d_3: The thing he likes to drink is ink

d_4: The ink he likes to drink is pink

d_5: He likes to wink, and drink pink ink

	d_1	d_2	d_3	d_4	d_5
he	1	1	1	1	1
likes	1	1	1	1	1
to	1	1	1	1	1
wink	1	0	0	0	1
drink	1	1	1	1	1
and	0	1	0	0	1
the	0	0	1	1	0
thing	0	0	1	0	0
ink	0	0	1	1	1
is	0	0	1	1	0
pink	0	0	0	1	1



TF-IDF = TF*IDF

where

TF = number of times the TERM appears in the DOCUMENT [redacted]

IDF = log (number of DOCUMENTS in the corpus) / number of DOCUMENTS in the corpus contain the TERM

$$w_{i,j} = tf_{i,j}\ \mathrm{x}\log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total numbers of document

TF= Term frequency (raw) => If Query Term appears in document, N times, TF = N

DF= Document frequency (raw) => If query term appears in N documents, DF = N

The more common a term is in documents, the less important it is, therefore we want to multiply Term frequency **by INVERSE Document frequency**

Example of TF-IDF

- Query term = Singapore University
- N = 1000 document,
  - 30 Documents have the word Singapore so DF(Singapore) = 30
  - 500 documents have the word University so DF(University) = 500
- Matching Singapore is more important because it is rarer than University

Document	**D1**	**D2**
TF (Singapore)	2	1
TF (University)	10	15
TF-IDF score	= 2 * Log (1000/30) + 10 * Log (1000/500) = 10.12+ 10 = **20.12**	= 1 * Log(1000/30) + 15 * Log (1000/500) = 5.06+15 = 20.06

Note: I am using log base 2, but the base will not affect ordering of docs (natural log and base 10 are also used commonly

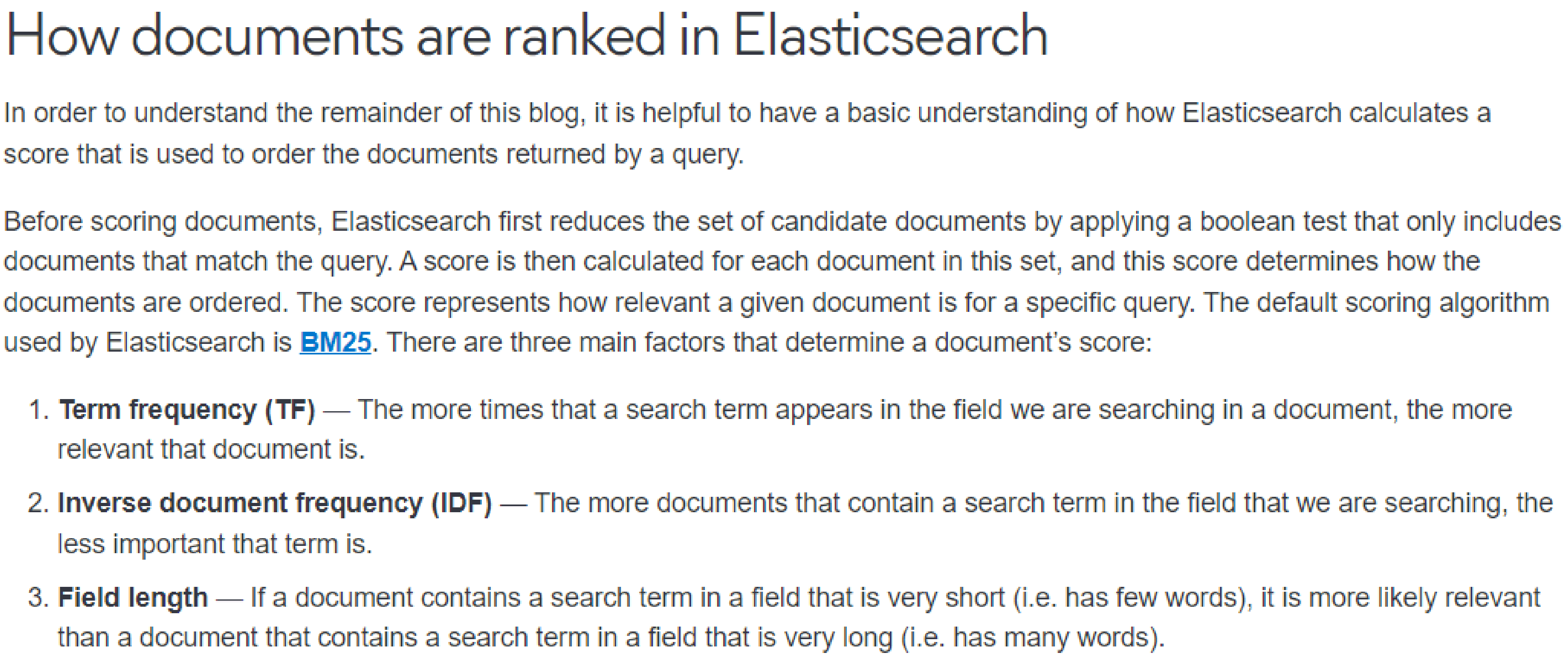
TF-IDF alone does not imply Boolean!

Singapore University implies Singapore AND University?

Document	D1	D3
TF (Singapore)	2	50
TF (University)	10	0
TF-IDF score	= 2 * Log (1000/30) + 10 * Log (1000/500) = 10.12+ 10 = 20.12	= 50 * Log(1000/30) + 0 * Log (1000/500) = 76.1+0 = **50.59**

D3 has higher TF-IDF score than D1 , even thought TF (University) = 0

Boolean + BM25 – in Elastic search



How documents are ranked in Elasticsearch

In order to understand the remainder of this blog, it is helpful to have a basic understanding of how Elasticsearch calculates a score that is used to order the documents returned by a query.

Before scoring documents, Elasticsearch first reduces the set of candidate documents by applying a boolean test that only includes documents that match the query. A score is then calculated for each document in this set, and this score determines how the documents are ordered. The score represents how relevant a given document is for a specific query. The default scoring algorithm used by Elasticsearch is BM25. There are three main factors that determine a document's score:

1. **Term frequency (TF)** — The more times that a search term appears in the field we are searching in a document, the more relevant that document is.
2. **Inverse document frequency (IDF)** — The more documents that contain a search term in the field that we are searching, the less important that term is.
3. **Field length** — If a document contains a search term in a field that is very short (i.e. has few words), it is more likely relevant than a document that contains a search term in a field that is very long (i.e. has many words).

https://www.elastic.co/blog/how-to-improve-elasticsearch-search-relevance-with-boolean-queries

BM25 - Evolution of TF-IDF – corrects for

- **Term saturation** – Compare Term frequency 1 vs 10 vs 100 vs 1000, does it make sense the relevancy score should scale linearly? – Control this with K factor in formula

- **Length of document** – Longer documents have more words = more chance of matching query terms, we should adjust for it! Control this with b factor in formula

Lexical search – Boolean/TF-IDF/BM25 main issue

- **Documents**: Views each document as a ***"bag" of words***
- **Main idea**: Re-ordering doesn't destroy the topic
  - individual words are "building blocks"
  - "bag" of words: a "composition" of "meanings"

Almost all Lexical or Keyword based methods are "bag of word" methods – ie does not take into account order of words or context of words

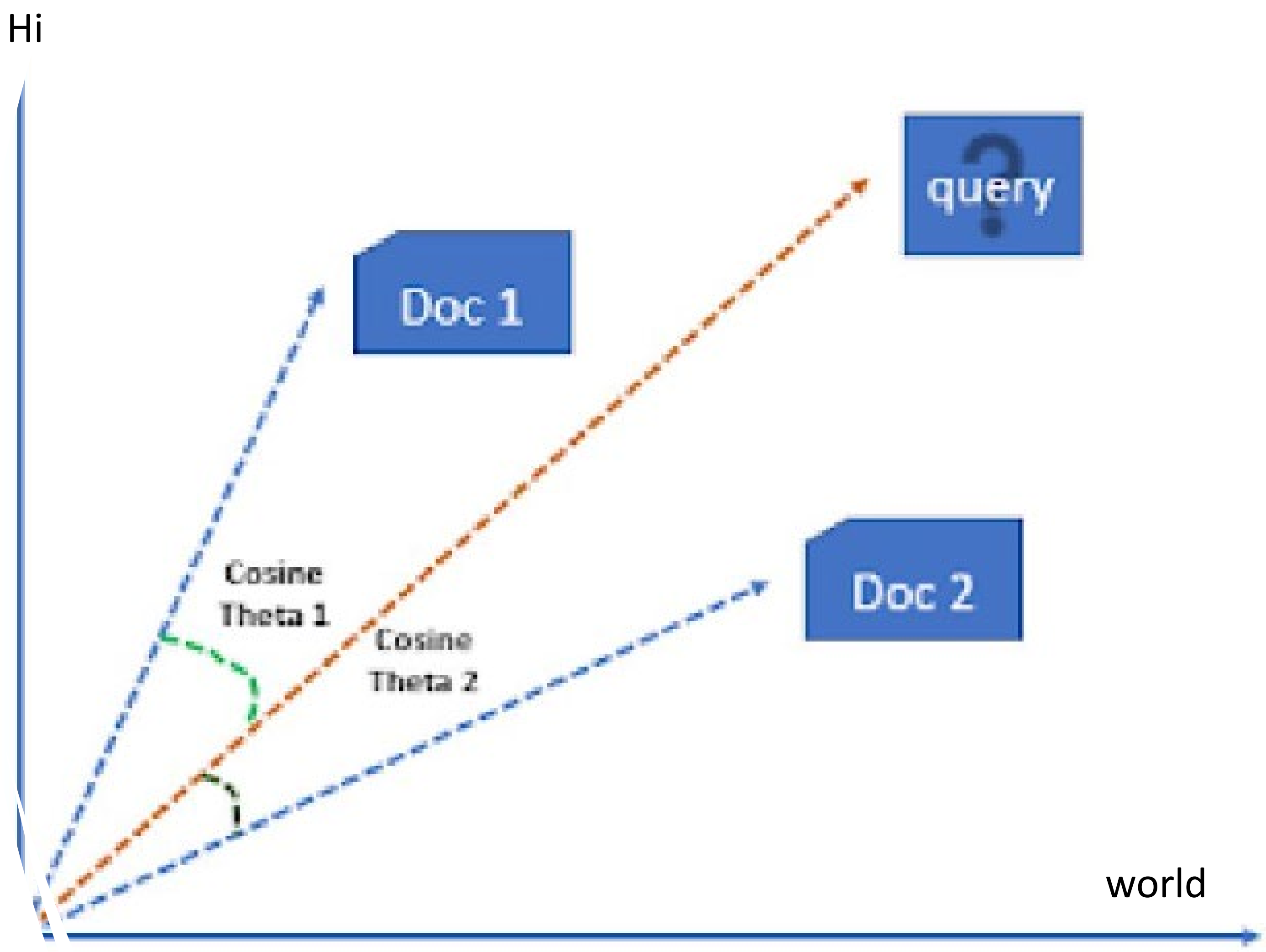
Words are tricky – Technical term = **Vocabulary mismatch problem**

- Polysemy (same word can have many diff word senses, e.g river “bank” vs financial “bank”)
- Synoymy (diff word can have same meaning, e.g. “car” vs “automobile”
- Jargon
- Spelling variations (UK vs US) - Morphological

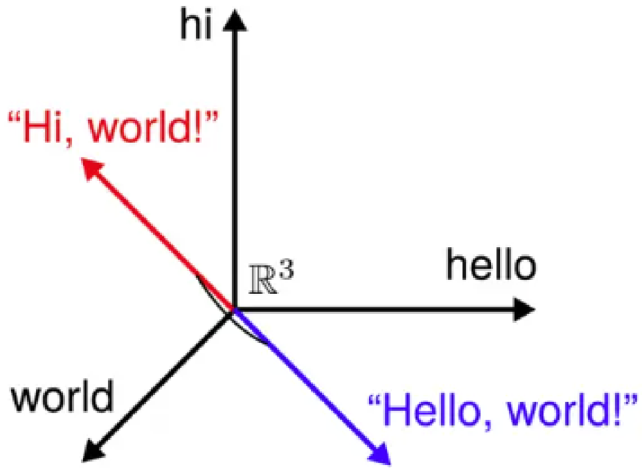
Bag of words approach - doesn’t take into account order of words, that’s why we remove stopwords

Idea of vector space model

- Represent document as a vector (series of numbers)
- Represent query as a vector (series of numbers)
- Axis represent – Term frequency
- Measure similarity of vector by angle size between vector



What if there are 3 terms?

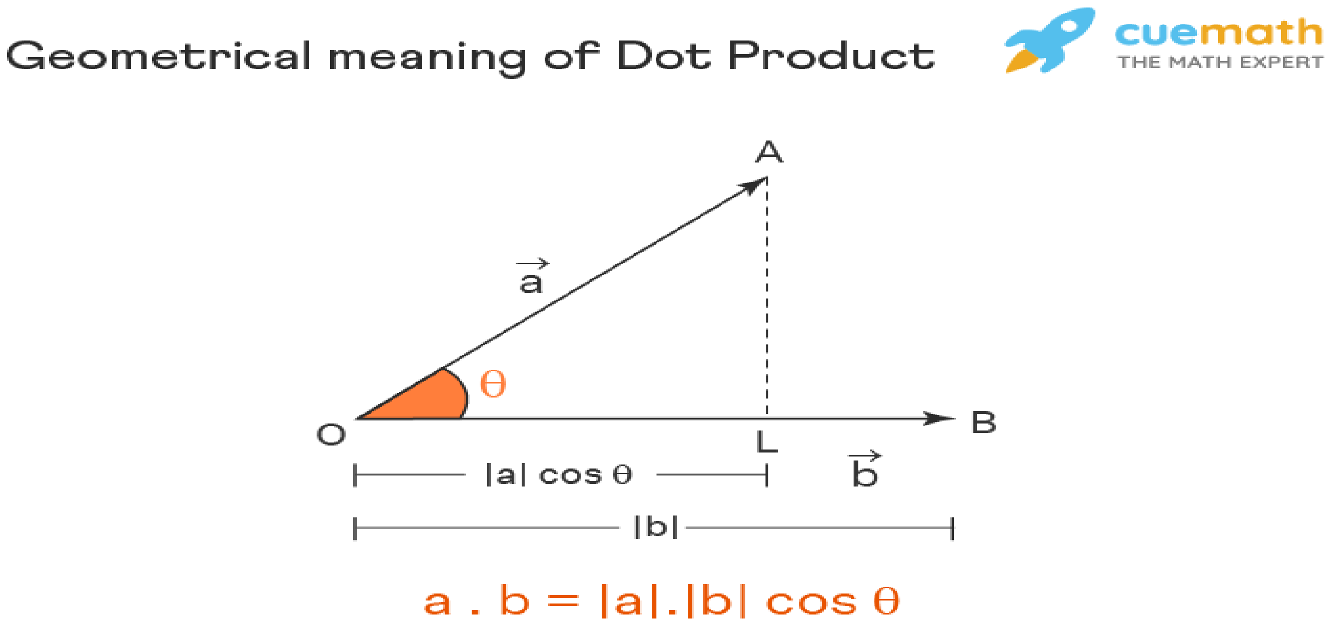


Cosine Similarity

	D1	D2
Hi	2	2
Hello	6	3
World	2	0

$$\begin{Bmatrix}2\\6\\2\end{Bmatrix} \begin{Bmatrix}2\\3\\0\end{Bmatrix}$$

What if there are >3 terms?



Dot product

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

What if there are >3 terms?

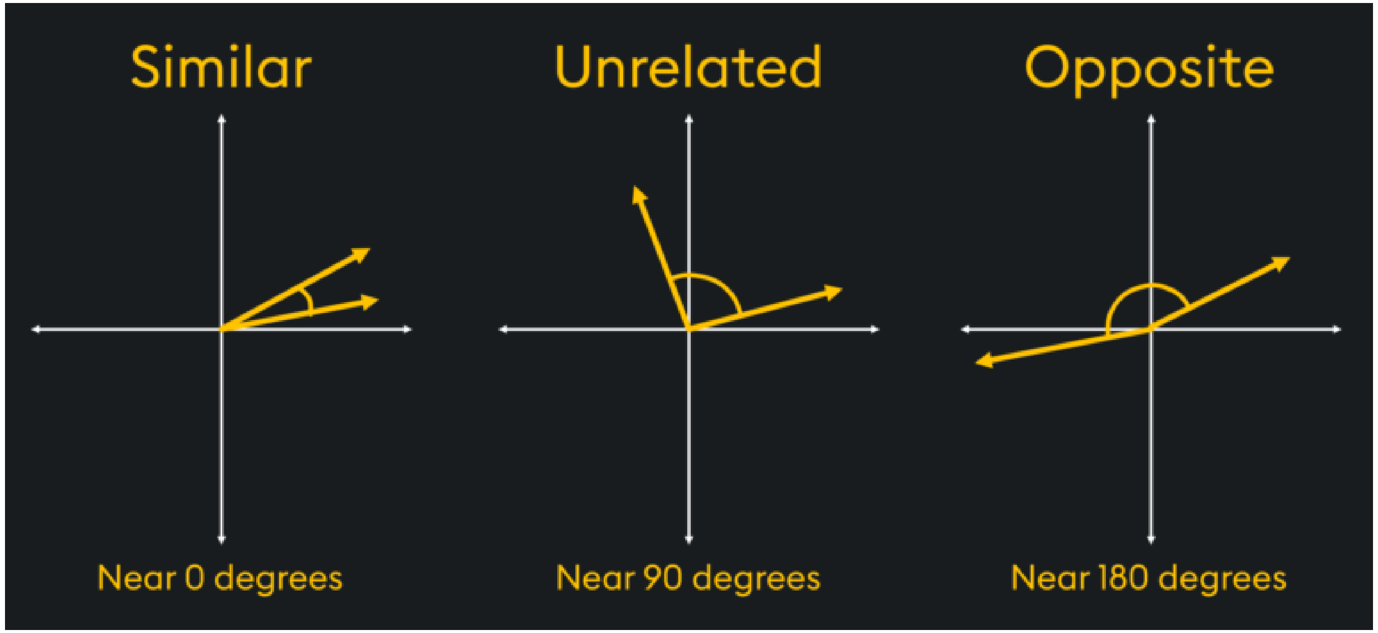
Cosine similarity

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\,\|\mathbf{B}\|}$$

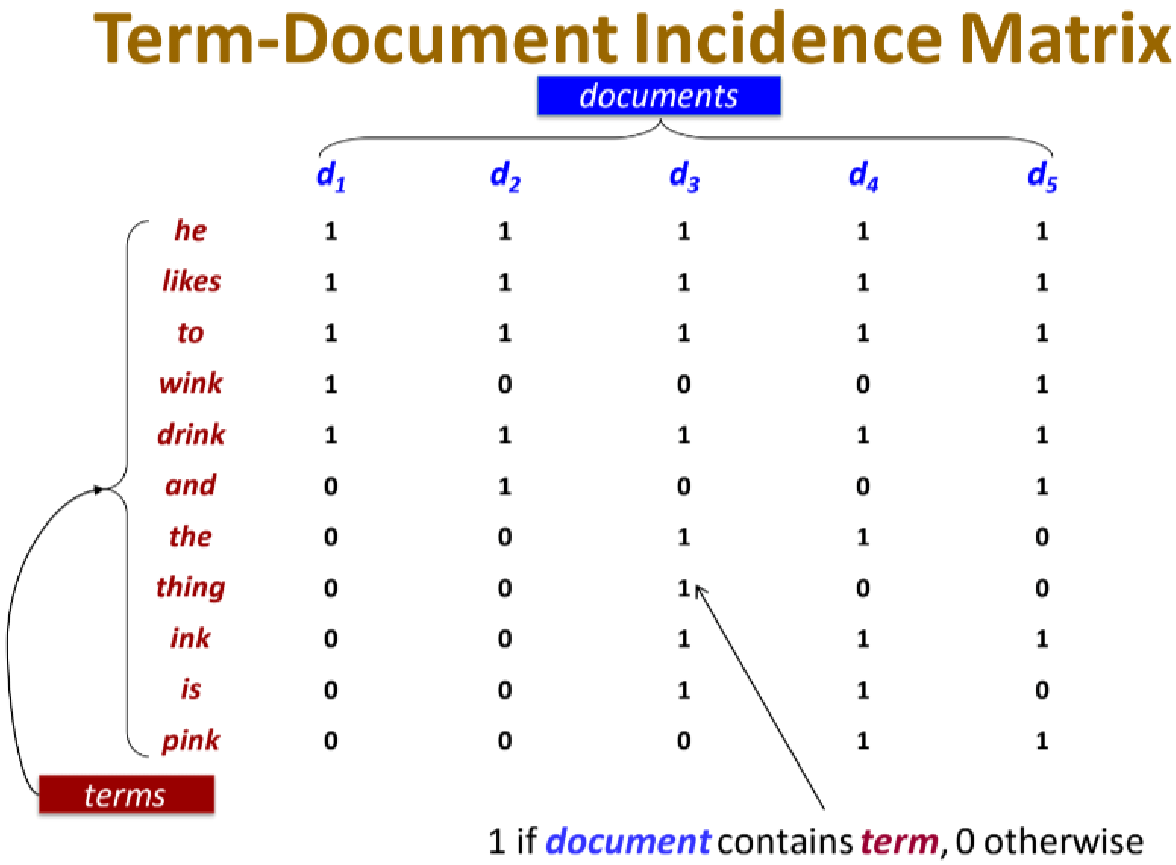
Cosine similarity range

Cosine similarity calculates the cosine of the angle between these two vectors to determine how similar they are.

- When two vectors overlap, cosine(0) = 1 => Maximum Similarity
- When two vectors are at 90 degree, Cosine (90) = 0 => Minimum Similarity
- Term frequencies and document-specific weights, cannot be negative. Thus, in practice, for document similarity calculations using TF-IDF, the cosine similarity values you'll encounter will range from 0 to 1.



Boolean Term-document matrix = "Sparse"



Term-Document Incidence Matrix

documents

	d_1	d_2	d_3	d_4	d_5
he	1	1	1	1	1
likes	1	1	1	1	1
to	1	1	1	1	1
wink	1	0	0	0	1
drink	1	1	1	1	1
and	0	1	0	0	1
the	0	0	1	1	0
thing	0	0	1	0	0
ink	0	0	1	1	1
is	0	0	1	1	0
pink	0	0	0	1	1

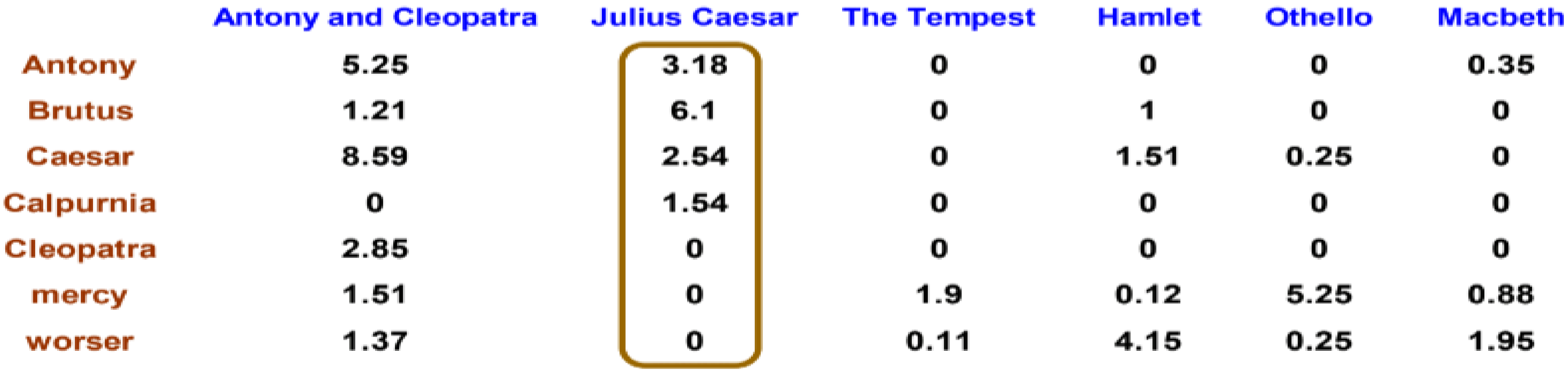
terms

1 if ***document*** contains ***term***, 0 otherwise

- Lexical/Keyword techniques are also known as "Sparse" representations or embedding (explain later).

- "Sparse" because most cells in the table are zero.

Values in cells can be binary (1 or 0), word count (e.g Term frequency) or any type of weights (e.g. TF-IDF)

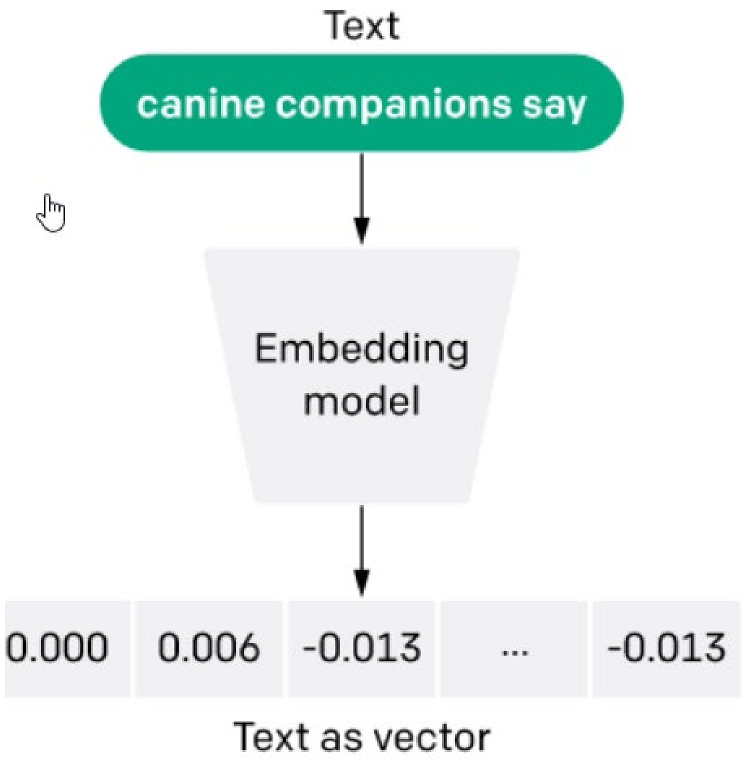


	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

Each document is now represented by a real-valued vector of tf-idf weights ∈ $R^{|V|}$

Do you understand why these are called "Sparse" embeddings/representations?

Neural search/Vector search/Semantic Search uses embeddings....

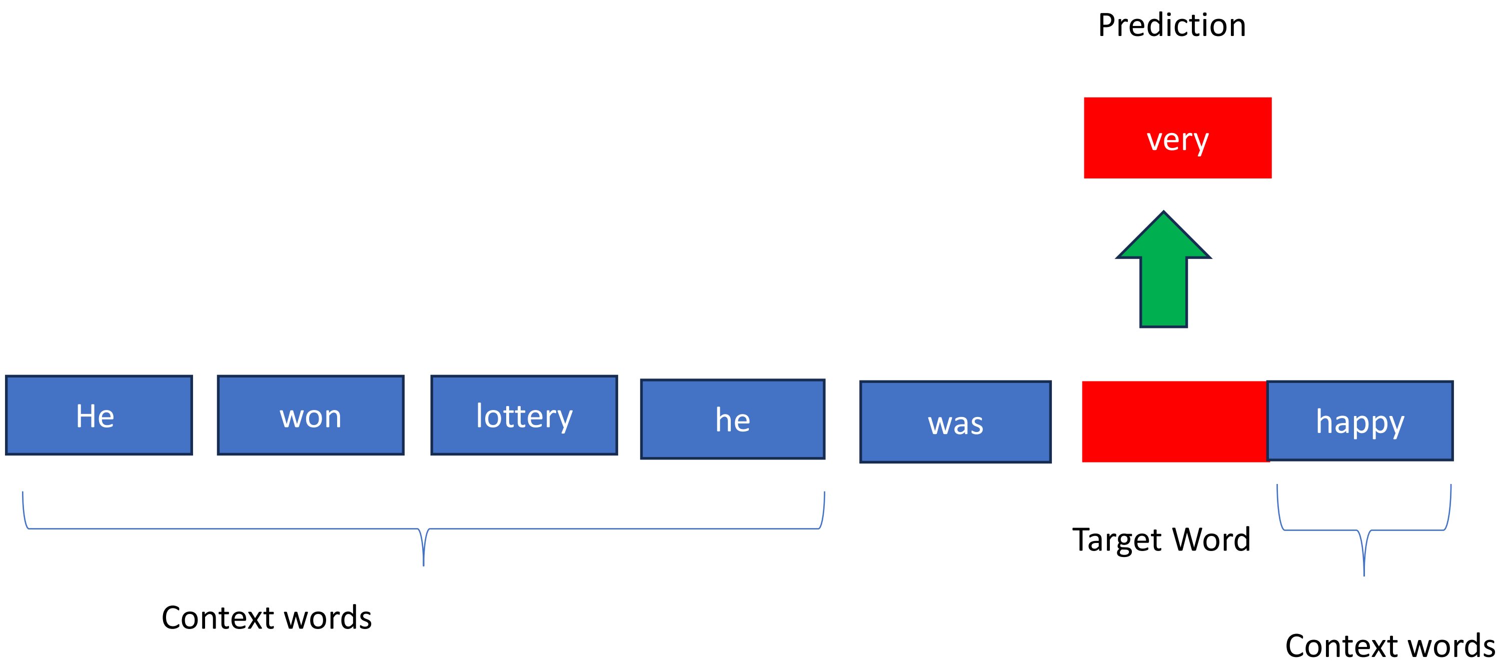


IDEA: Use neural networks to learn to represent words with string of numbers

You can specify the embedding to be fixed length of n numbers.

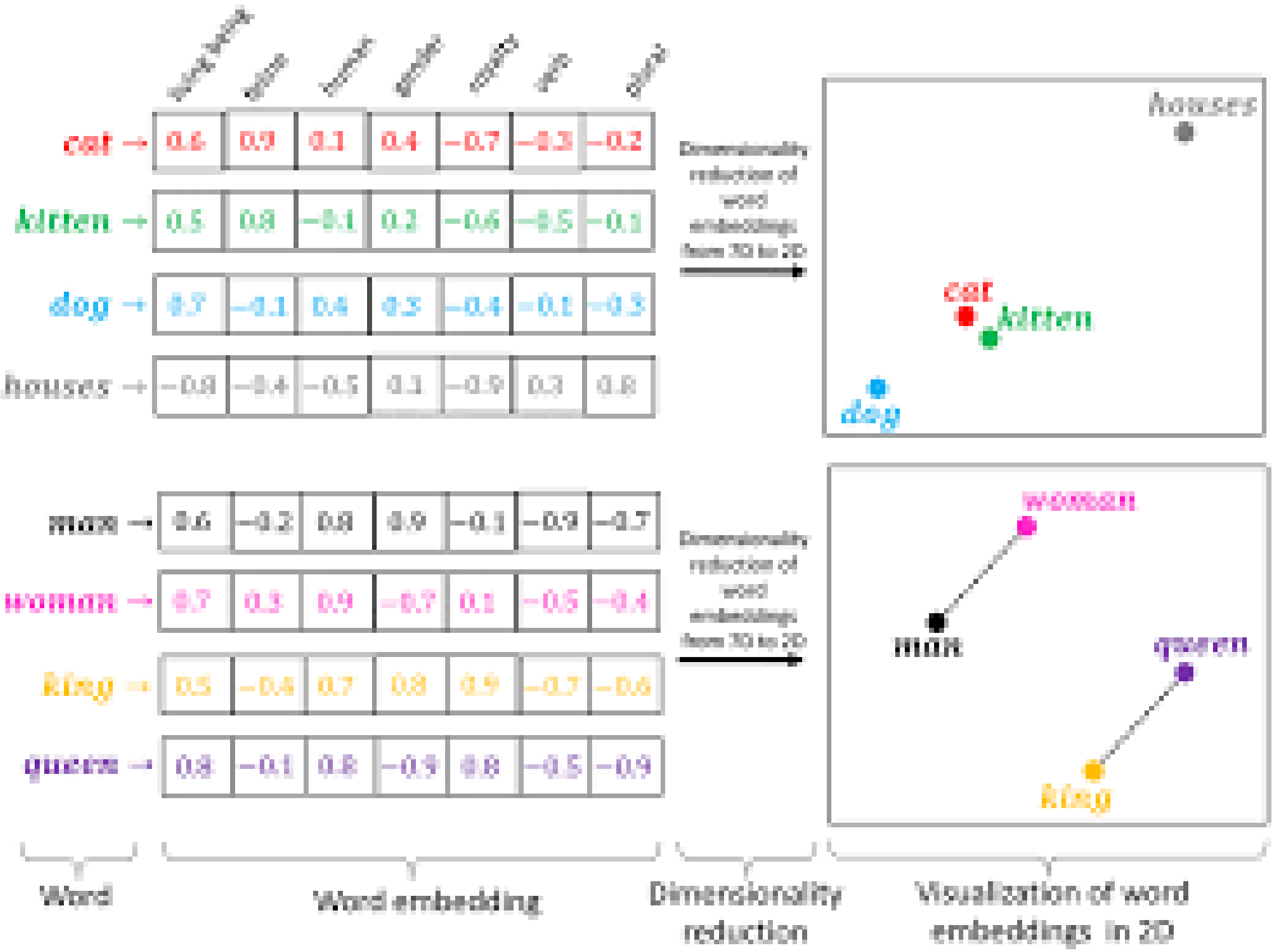
https://www.ai-contentlab.com/2023/05/introduction-to-vector-embedding.html

Rough idea of training



"a word is characterized by the company it keeps" (Firth)

Word2Vec (2013) - earlier embedding

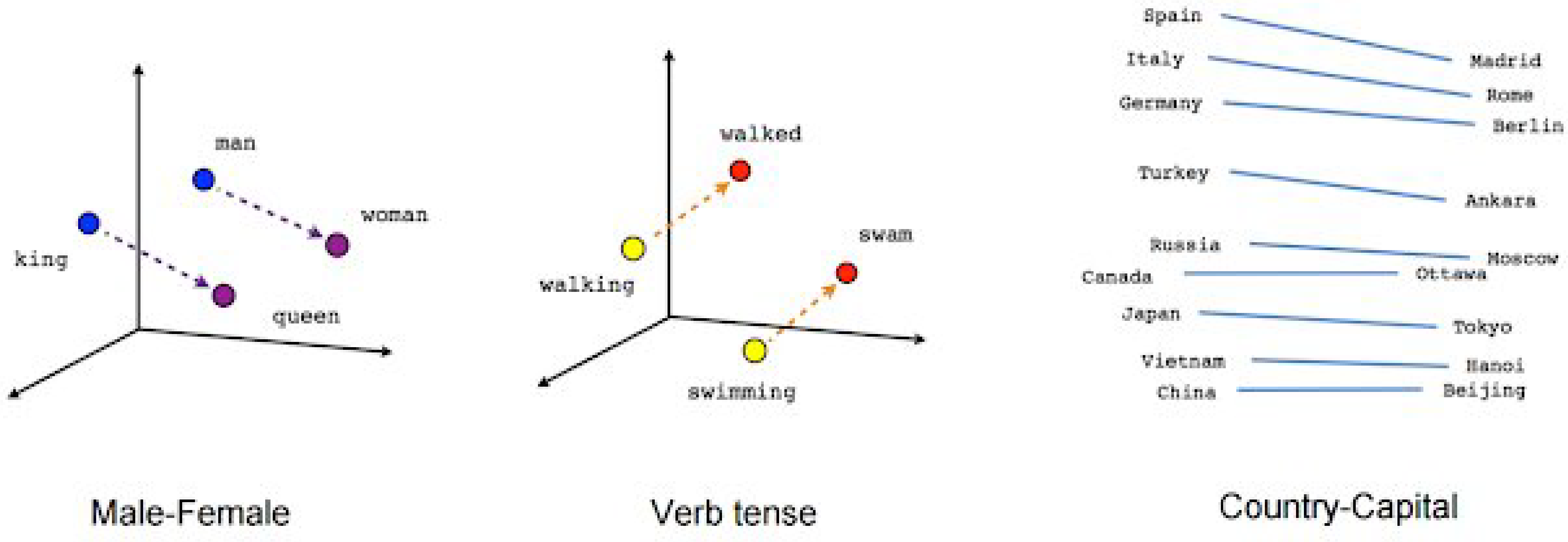


- Represent docs and queries as embedding vectors (series of numbers)

- The values in the embeddings are learnt from neural nets instead of using statistical models like TF-IDF, BM25

- Otherwise similar concept, find similarity using cosine similarity

Word embedding in Word2vec make "sense"



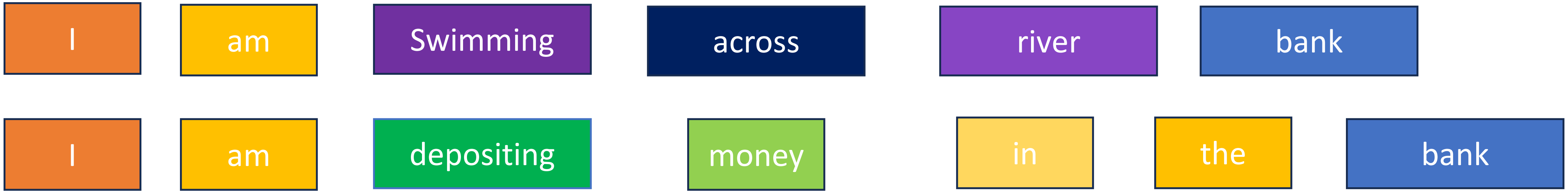
Male-Female

Verb tense

Country-Capital

MAN-KING+WOMAN = QUEEN

Word2vec and other static Embedding (2015s).e.g GLOVE, Fastext

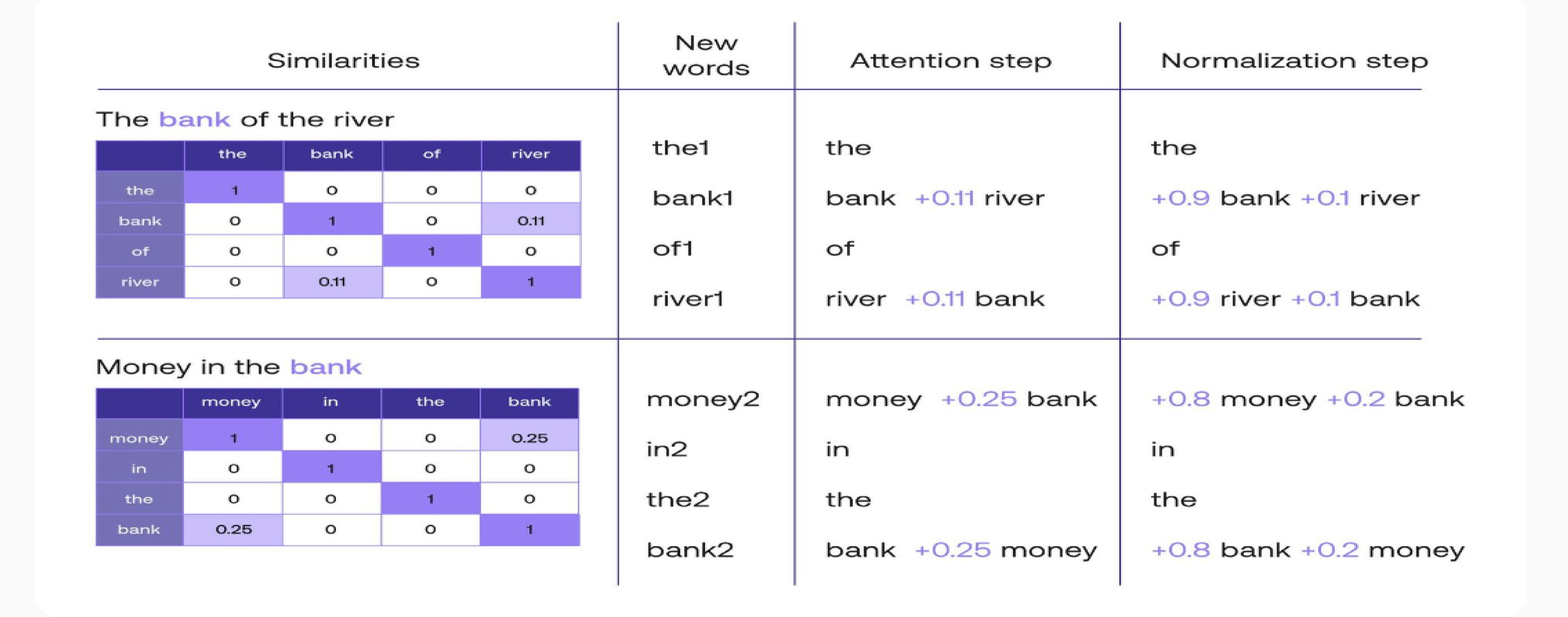


I am Swimming across river bank

I am depositing money in the bank

Main weakness – static embedding – the embedding for “bank” is the same when used in two diff sentences

Attention mechanism from transformers (2017)…



Similarities	New words	Attention step	Normalization step
The bank of the river	the1	the	the
	bank1	bank +0.11 river	+0.9 bank +0.1 river
	of1	of	of
	river1	river +0.11 bank	+0.9 river +0.1 bank
Money in the bank	money2	money +0.25 bank	+0.8 money +0.2 bank
	in2	in	in
	the2	the	the
	bank2	bank +0.25 money	+0.8 bank +0.2 money

The bank of the river

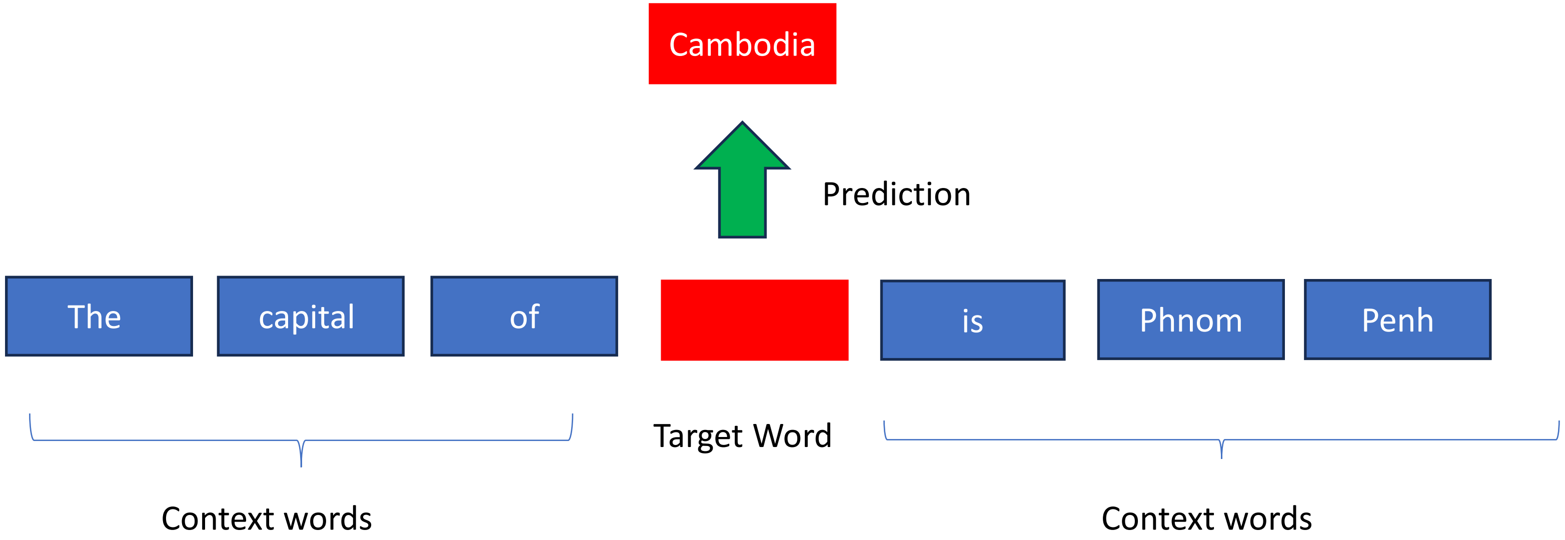
	the	bank	of	river
the	1	0	0	0
bank	0	1	0	0.11
of	0	0	1	0
river	0	0.11	0	1

Money in the bank

	money	in	the	bank
money	1	0	0	0.25
in	0	1	0	0
the	0	0	1	0
bank	0.25	0	0	1

https://cohere.com/blog/what-is-attention-in-language-models

Masked language model (simplified)



Cambodia

Prediction

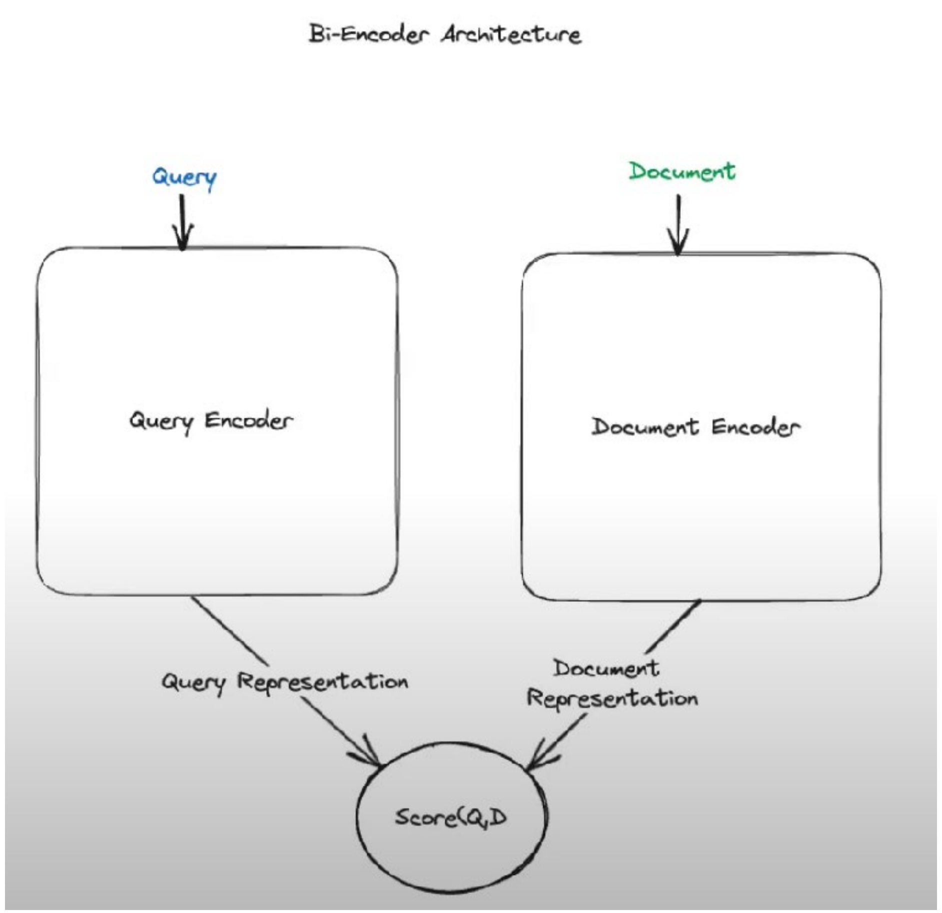
The capital of 🟥 is Phnom Penh

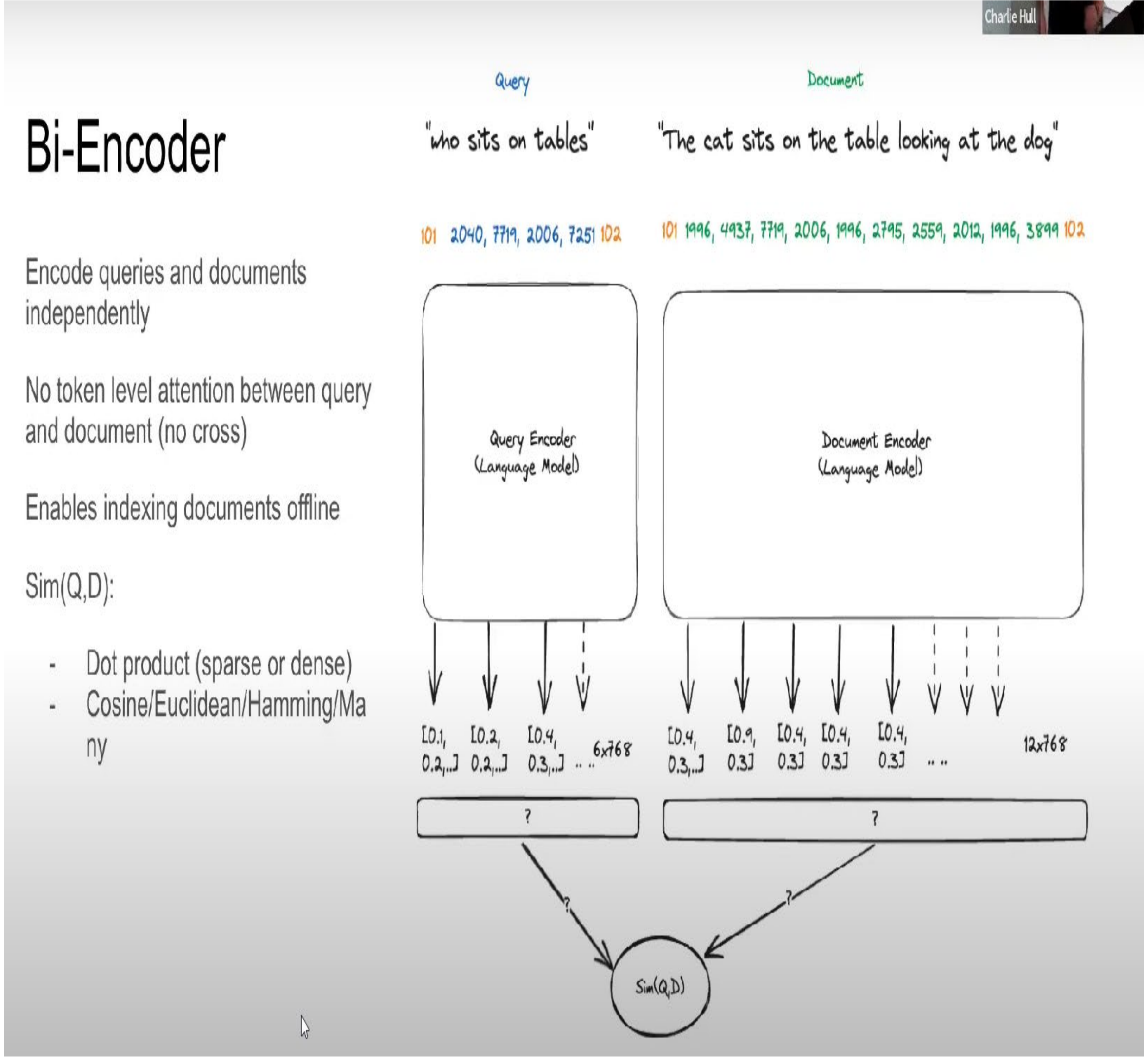
Context words

Target Word

Context words

Bi-encoder





Bi-Encoder

Encode queries and documents independently

No token level attention between query and document (no cross)

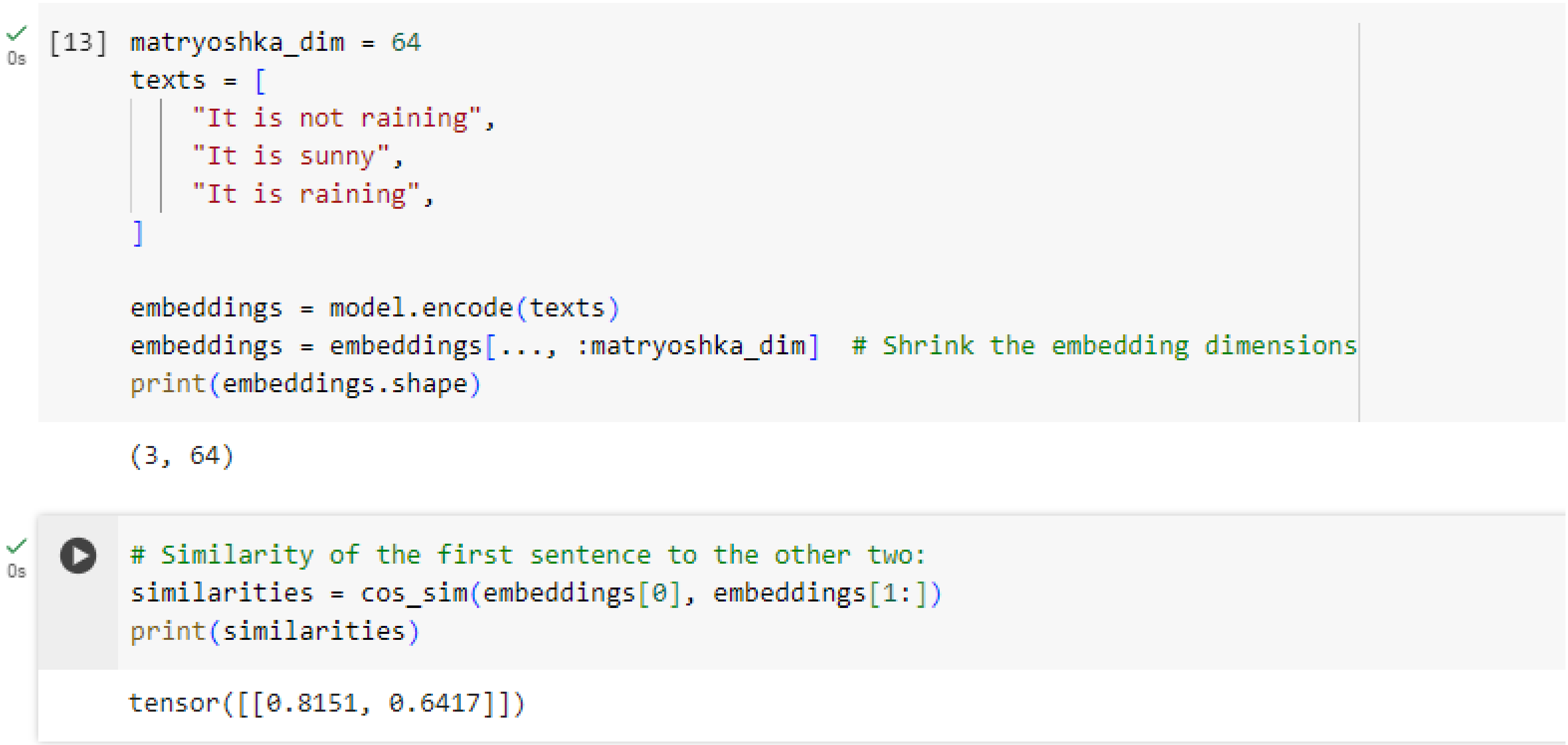
Enables indexing documents offline

Sim(Q,D):

- Dot product (sparse or dense)
- Cosine/Euclidean/Hamming/Many

Find cosine similarity of query embedding vector and doc embedding vector

Some “Evidence” it works – simple embedding example



```
matryoshka_dim = 64
texts = [
    "It is not raining",
    "It is sunny",
    "It is raining",
]

embeddings = model.encode(texts)
embeddings = embeddings[..., :matryoshka_dim]  # Shrink the embedding dimensions
print(embeddings.shape)
```

```
(3, 64)
```

```
# Similarity of the first sentence to the other two:
similarities = cos_sim(embeddings[0], embeddings[1:])
print(similarities)
```

```
tensor([[0.8151, 0.6417]])
```

Similarity of “It is not raining” and “It is sunny” is 0.8151
Similarity of “It is not raining” and “It is raining” is 0.6417, which is lower despite matching more keywords
This embedding search “understands” the meaning of “not raining”!

Cross-Encoder

Encodes both query and document at the same time (cross)

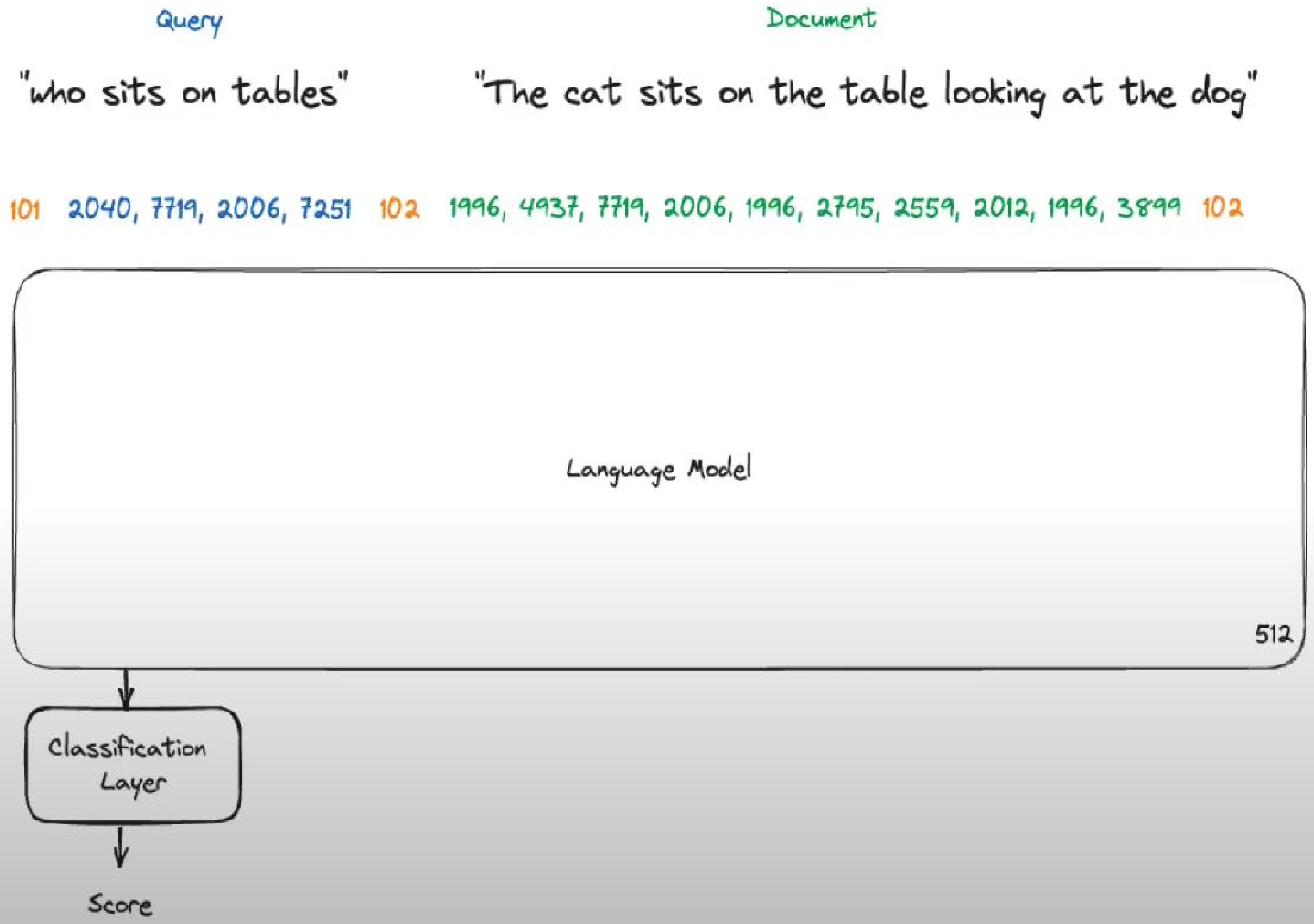
all-to-all attention between all tokens in query and document

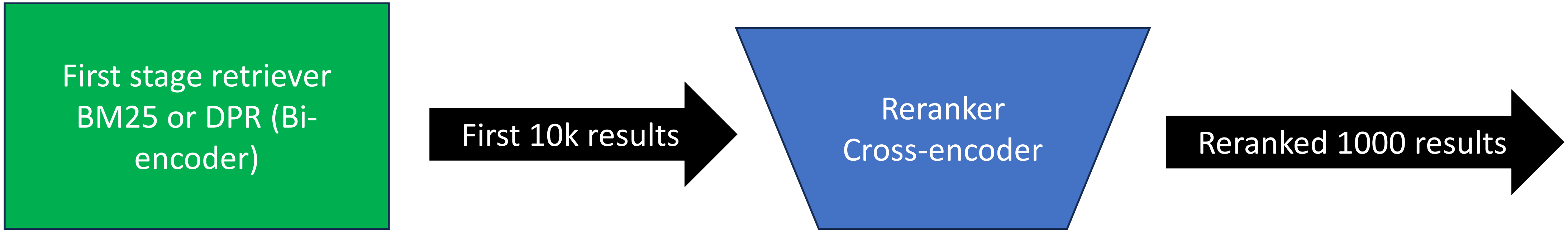
Most effective on IR benchmarks (nDCG)

High compute complexity (n^2)

No efficient way to “index”

Distillation can be used, where cross-encoder “teaches” bi-encoder





First stage retriever BM25 or DPR (Bi-encoder)

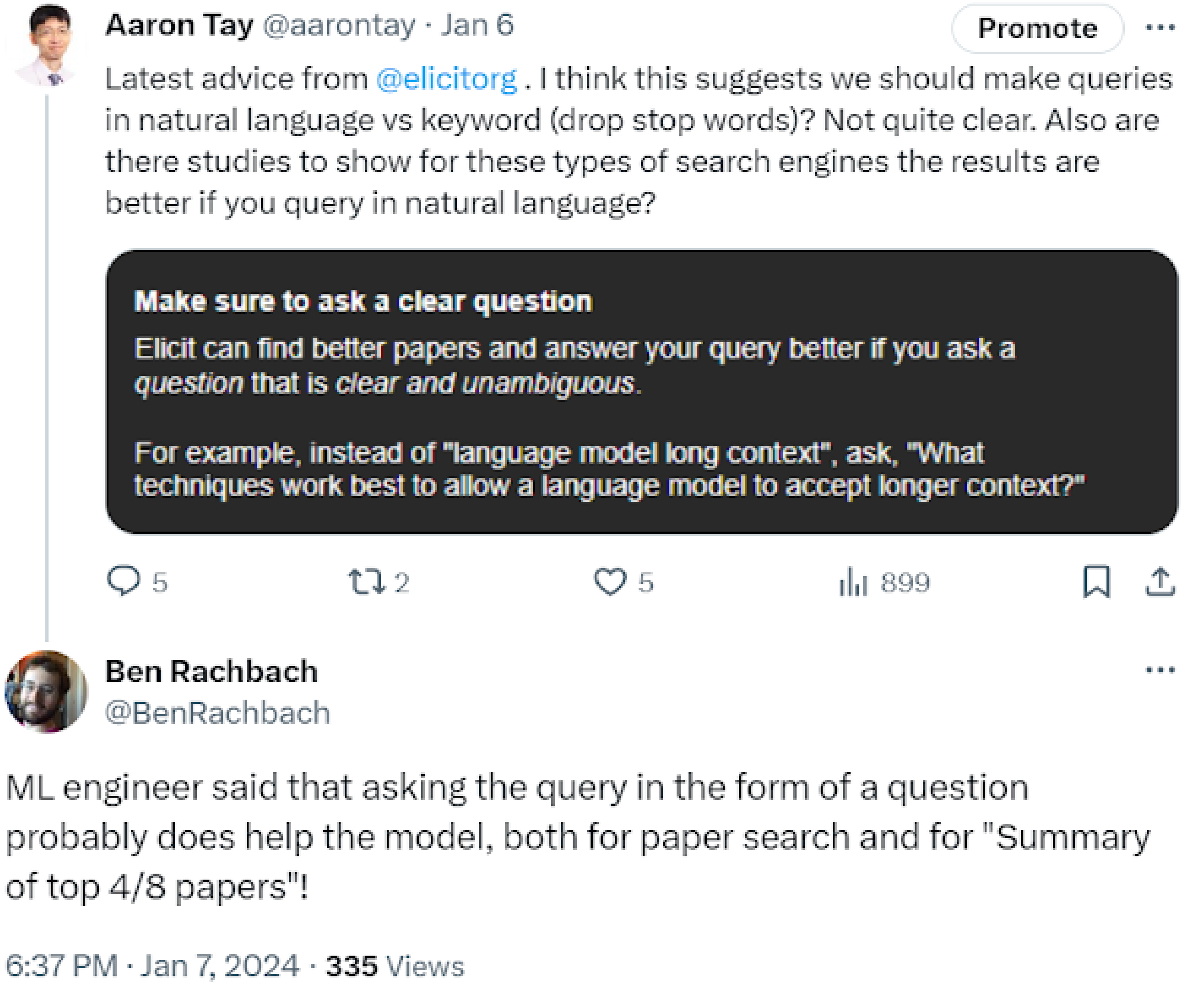
First 10k results

Reranker
Cross-encoder

Reranked 1000 results

Questions about semantic search

- Do results work better in natural language ?
- Are the results interpretable and predictable?
- Do we need to change the way we search??



Aaron Tay @aarontay · Jan 6

Latest advice from @elicitorg . I think this suggests we should make queries in natural language vs keyword (drop stop words)? Not quite clear. Also are there studies to show for these types of search engines the results are better if you query in natural language?

Make sure to ask a clear question

Elicit can find better papers and answer your query better if you ask a *question* that is *clear and unambiguous*.

For example, instead of "language model long context", ask, "What techniques work best to allow a language model to accept longer context?"

Ben Rachbach
@BenRachbach

ML engineer said that asking the query in the form of a question probably does help the model, both for paper search and for "Summary of top 4/8 papers"!

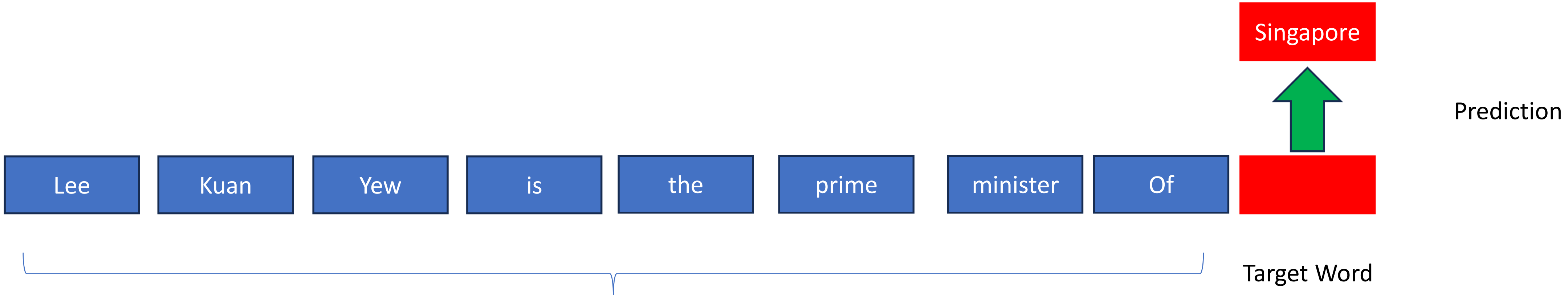
6:37 PM · Jan 7, 2024 · **335** Views

Can we just use semantic search and throw away lexical search?

Can we throw away lexical search and just do biencoders+cross encoders?

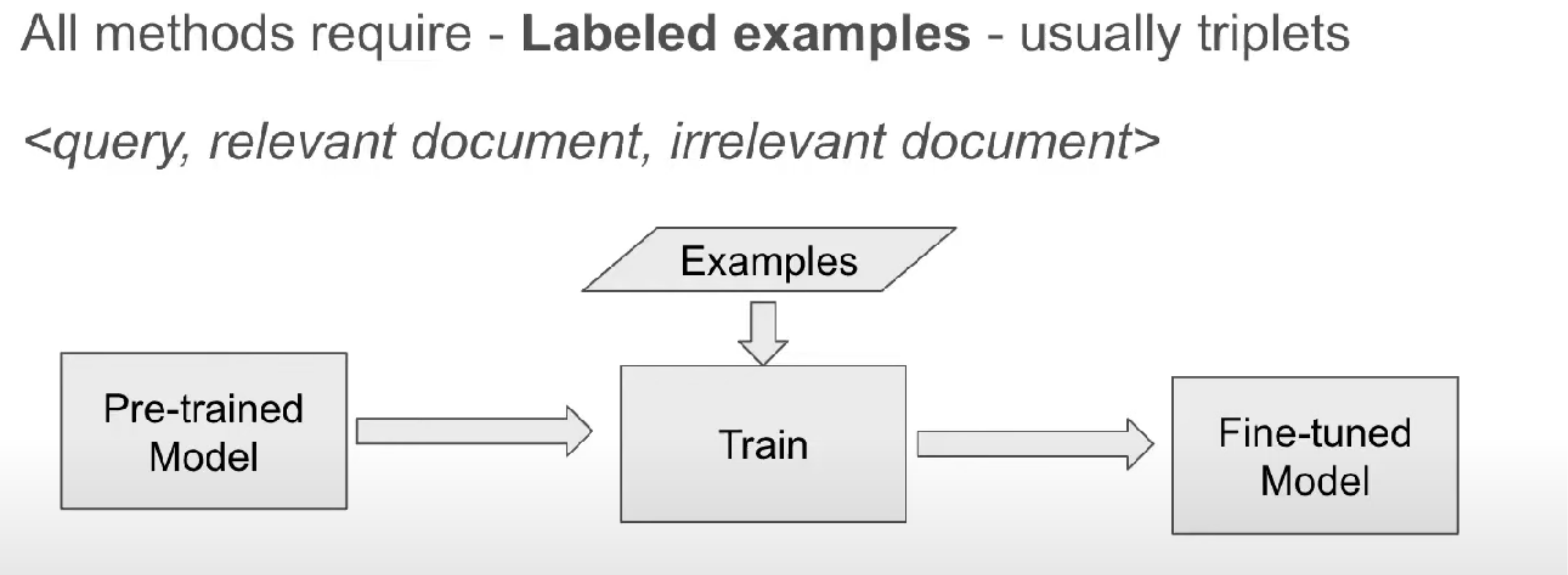
- Probably not
- Bi-encoders are still slower than lexical search
- Lexical search are better at string and exact match search for out of domain (not trained strings)
- Semantic search work better for in-domain tasks

Reminder embedding depend on the text they are trained on.



Embedding	Trained on
Word2Vec	Google News
bert-base-uncased	English Wikipedia
OpenAI's Ada-002	GPT based
Allen institute for AI's SPECTER2	Academic content

Important : Dense embeddings are further trained on specifical labelled examples



All methods require - **Labeled examples** - usually triplets

<query, relevant document, irrelevant document>

Examples

Pre-trained Model

Train

Fine-tuned Model

A lot of different techniques to create such labelled examples, weakly supervised methods and different sampling strategies (e.g. selecting hard, very hard samples) – constrastive training

Up to recently, sparse embedding based on BM25 are not trained on labelled examples, this has changed in 2022-2023, with new “learned sparse representations” e.g. SPLADE, SparseEmbed etc

Dense embedding methods work well on one test but how well do they do across many?

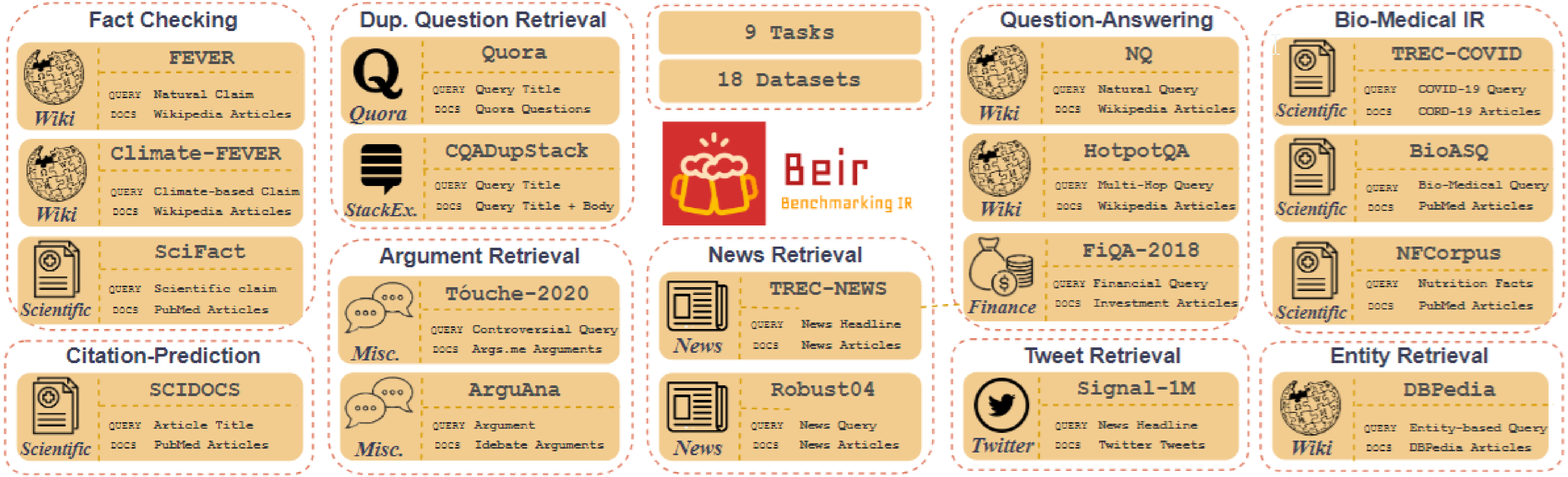
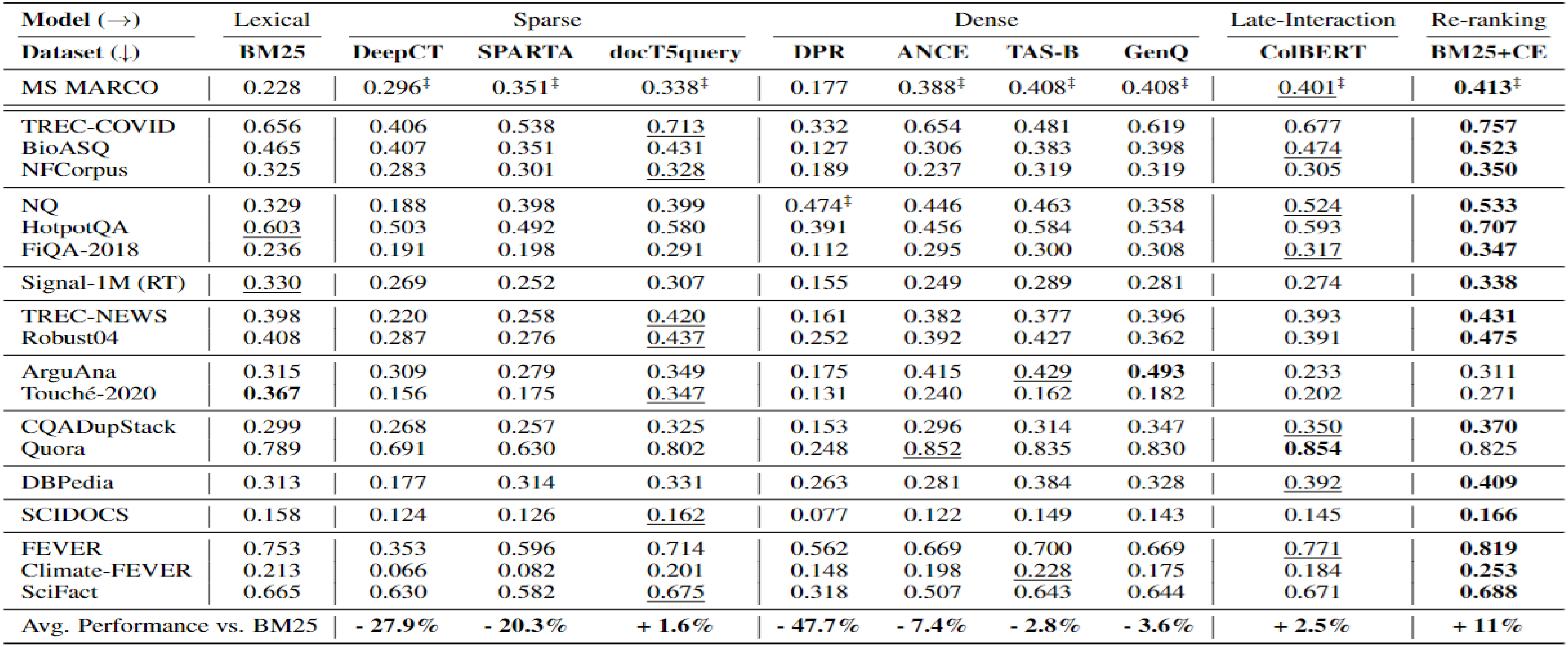


Figure 1: An overview of the diverse tasks and datasets in BEIR benchmark.

Now replaced by the even larger MTEB benchmark (see earlier)

https://arxiv.org/abs/2104.08663

BEIR – large benchmark, BM25 is hard to beat on average – only ColBERT wins on average



Model (→)	Lexical	Sparse			Dense				Late-Interaction	Re-ranking
Dataset (↓)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**GenQ**	**ColBERT**	**BM25+CE**
MS MARCO	0.228	0.296‡	0.351‡	0.338‡	0.177	0.388‡	0.408‡	0.408‡	0.401‡	**0.413‡**
TREC-COVID	0.656	0.406	0.538	0.713	0.332	0.654	0.481	0.619	0.677	**0.757**
BioASQ	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398	0.474	**0.523**
NFCorpus	0.325	0.283	0.301	0.328	0.189	0.237	0.319	0.319	0.305	**0.350**
NQ	0.329	0.188	0.398	0.399	0.474‡	0.446	0.463	0.358	0.524	**0.533**
HotpotQA	0.603	0.503	0.492	0.580	0.391	0.456	0.584	0.534	0.593	**0.707**
FiQA-2018	0.236	0.191	0.198	0.291	0.112	0.295	0.300	0.308	0.317	**0.347**
Signal-1M (RT)	0.330	0.269	0.252	0.307	0.155	0.249	0.289	0.281	0.274	**0.338**
TREC-NEWS	0.398	0.220	0.258	0.420	0.161	0.382	0.377	0.396	0.393	**0.431**
Robust04	0.408	0.287	0.276	0.437	0.252	0.392	0.427	0.362	0.391	**0.475**
ArguAna	0.315	0.309	0.279	0.349	0.175	0.415	0.429	**0.493**	0.233	0.311
Touché-2020	**0.367**	0.156	0.175	0.347	0.131	0.240	0.162	0.182	0.202	0.271
CQADupStack	0.299	0.268	0.257	0.325	0.153	0.296	0.314	0.347	0.350	**0.370**
Quora	0.789	0.691	0.630	0.802	0.248	0.852	0.835	0.830	**0.854**	0.825
DBPedia	0.313	0.177	0.314	0.331	0.263	0.281	0.384	0.328	0.392	**0.409**
SCIDOCS	0.158	0.124	0.126	0.162	0.077	0.122	0.149	0.143	0.145	**0.166**
FEVER	0.753	0.353	0.596	0.714	0.562	0.669	0.700	0.669	0.771	**0.819**
Climate-FEVER	0.213	0.066	0.082	0.201	0.148	0.198	0.228	0.175	0.184	**0.253**
SciFact	0.665	0.630	0.582	0.675	0.318	0.507	0.643	0.644	0.671	**0.688**
Avg. Performance vs. BM25		**- 27.9%**	**- 20.3%**	**+ 1.6%**	**- 47.7%**	**- 7.4%**	**- 2.8%**	**- 3.6%**	**+ 2.5%**	**+ 11%**

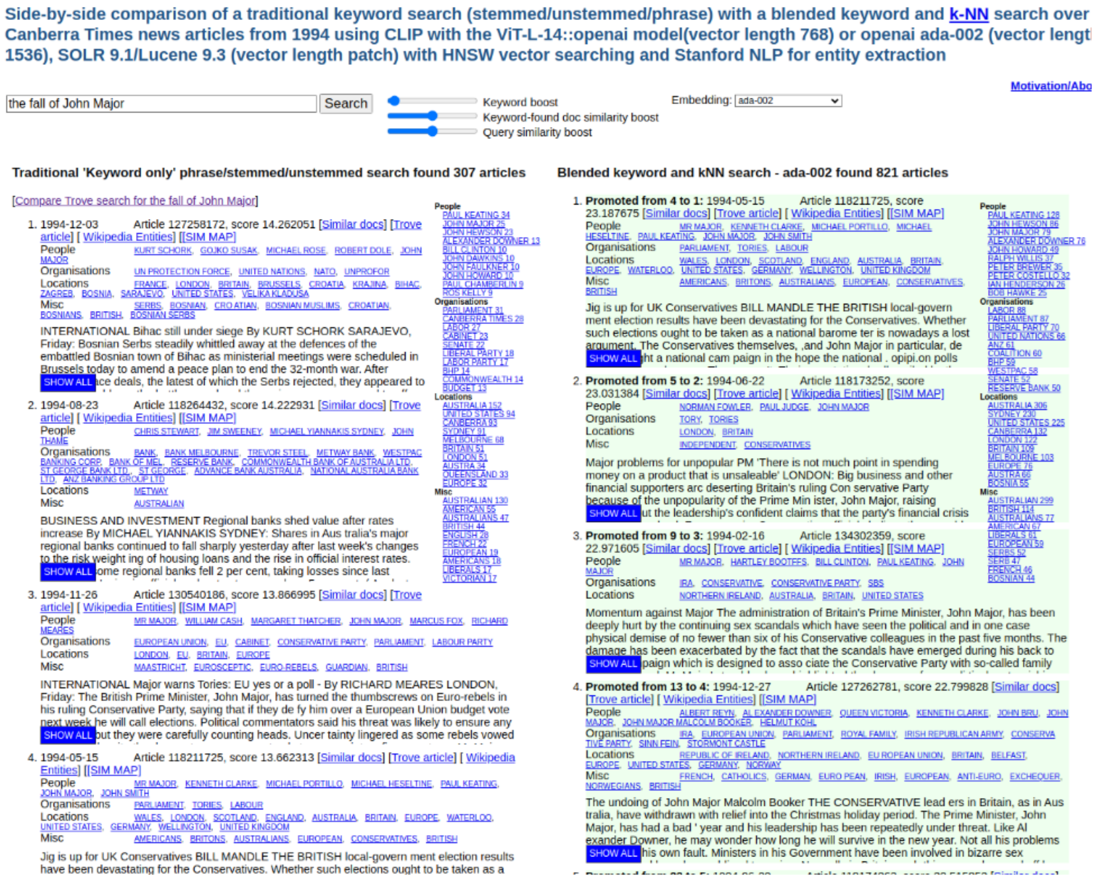
All dense embedding methods are poorer than BM25 on average across diverse tests. Only COLBERT and BM25+Cross encoder are better!

MTEB leaderboard

- Large benchmark of many test suites
- Allows you to know the performance of different embeddings for different NLP tasks.
- Some NLP tasks performances measured are
  - Retrieval – Finding relevant articles
  - Rerank – Reranking relevant articles
  - STS (Semantic Text Similarity) – Scoring similar sentences
  - Other NLP tasks- Summarization, Clustering, Translation, classification

E5 embedding chosen by Govtech for Hasnard search is one of the top5 embedding for retrieval!

Hybrid search is the way to go - example



Side-by-side comparison of a traditional keyword search (stemmed/unstemmed/phrase) with a blended keyword and k-NN search over Canberra Times news articles from 1994 using CLIP with the ViT-L-14::openai model(vector length 768) or openai ada-002 (vector lengt 1536), SOLR 9.1/Lucene 9.3 (vector length patch) with HNSW vector searching and Stanford NLP for entity extraction

Traditional 'Keyword only' phrase/stemmed/unstemmed search found 307 articles

[Compare Trove search for the fall of John Major]

1. 1994-12-03 Article 127258172, score 14.262051 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People KURT SCHORK, GOJKO SUSAK, MICHAEL ROSE, ROBERT DOLE, JOHN MAJOR
   Organisations UN PROTECTION FORCE, UNITED NATIONS, NATO, UNPROFOR
   Locations FRANCE, LONDON, BRITAIN, BRUSSELS, CROATIA, KRAJINA, BIHAC, ZAGREB, BOSNIA, SARAJEVO, UNITED STATES, VELIKA KLADUSA
   Misc SERBS, BOSNIAN, CRO ATIAN, BOSNIAN MUSLIMS, CROATIAN, BOSNIANS, BRITISH, BOSNIAN SERBS
   INTERNATIONAL Bihac still under siege By KURT SCHORK SARAJEVO, Friday: Bosnian Serbs steadily whittled away at the defences of the embattled Bosnian town of Bihac as ministerial meetings were scheduled in Brussels today to amend a peace plan to end the 32-month war. After SHOW ALL ce deals, the latest of which the Serbs rejected, they appeared to
2. 1994-08-23 Article 118264432, score 14.222931 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People CHRIS STEWART, JIM SWEENEY, MICHAEL YIANNAKIS SYDNEY, JOHN THAME
   Organisations BANK, BANK MELBOURNE, TREVOR STEEL, METWAY BANK, WESTPAC BANKING CORP, BANK OF MEL, RESERVE BANK, COMMONWEALTH BANK OF AUSTRALIA LTD, ST GEORGE BANK LTD, ST GEORGE, ADVANCE BANK AUSTRALIA, NATIONAL AUSTRALIA BANK LTD, ANZ BANKING GROUP LTD
   Locations METWAY
   Misc AUSTRALIAN
   BUSINESS AND INVESTMENT Regional banks shed value after rates increase By MICHAEL YIANNAKIS SYDNEY: Shares in Aus tralia's major regional banks continued to fall sharply yesterday after last week's changes to the risk weight ing of housing loans and the rise in official interest rates. SHOW ALL ome regional banks fell 2 per cent, taking losses since last
3. 1994-11-26 Article 130540186, score 13.866995 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People MR MAJOR, WILLIAM CASH, MARGARET THATCHER, JOHN MAJOR, MARCUS FOX, RICHARD MEARES
   Organisations EUROPEAN UNION, EU, CABINET, CONSERVATIVE PARTY, PARLIAMENT, LABOUR PARTY
   Locations LONDON, EU, BRITAIN, EUROPE
   Misc MAASTRICHT, EUROSCEPTIC, EURO-REBELS, GUARDIAN, BRITISH
   INTERNATIONAL Major warns Tories: EU yes or a poll - By RICHARD MEARES LONDON, Friday: The British Prime Minister, John Major, has turned the thumbscrews on Euro-rebels in his ruling Conservative Party, saying that if they de fy him over a European Union budget vote next week he will call elections. Political commentators said his threat was likely to ensure any SHOW ALL but they were carefully counting heads. Uncer tainty lingered as some rebels vowed
4. 1994-05-15 Article 118211725, score 13.662313 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People MR MAJOR, KENNETH CLARKE, MICHAEL PORTILLO, MICHAEL HESELTINE, PAUL KEATING, JOHN MAJOR, JOHN SMITH
   Organisations PARLIAMENT, TORIES, LABOUR
   Locations WALES, LONDON, SCOTLAND, ENGLAND, AUSTRALIA, BRITAIN, EUROPE, WATERLOO, UNITED STATES, GERMANY, WELLINGTON, UNITED KINGDOM
   Misc AMERICANS, BRITONS, AUSTRALIANS, EUROPEAN, CONSERVATIVES, BRITISH
   Jig is up for UK Conservatives BILL MANDLE THE BRITISH local-govern ment election results have been devastating for the Conservatives. Whether such elections ought to be taken as a

People
PAUL KEATING 34
JOHN MAJOR 25
JOHN HEWSON 23
ALEXANDER DOWNER 13
BILL CLINTON 10
JOHN DAWKINS 10
JOHN FAULKNER 10
JOHN HOWARD 10
PAUL CHAMBERLIN 9
ROS KELLY 9

Organisations
PARLIAMENT 31
CANBERRA TIMES 28
LABOR 27
CABINET 23
SENATE 22
LIBERAL PARTY 18
LABOR PARTY 17
BHP 14
COMMONWEALTH 14
BUDGET 13

Locations
AUSTRALIA 152
UNITED STATES 94
CANBERRA 93
SYDNEY 91
MELBOURNE 68
BRITAIN 51
LONDON 51
AUSTRA 34
QUEENSLAND 33
EUROPE 32

Misc
AUSTRALIAN 130
AMERICAN 55
AUSTRALIANS 47
BRITISH 44
ENGLISH 28
FRENCH 22
EUROPEAN 19
AMERICANS 18
LIBERALS 17
VICTORIAN 17

Blended keyword and kNN search - ada-002 found 821 articles

1. **Promoted from 4 to 1:** 1994-05-15 Article 118211725, score 23.187675 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People MR MAJOR, KENNETH CLARKE, MICHAEL PORTILLO, MICHAEL HESELTINE, PAUL KEATING, JOHN MAJOR, JOHN SMITH
   Organisations PARLIAMENT, TORIES, LABOUR
   Locations WALES, LONDON, SCOTLAND, ENGLAND, AUSTRALIA, BRITAIN, EUROPE, WATERLOO, UNITED STATES, GERMANY, WELLINGTON, UNITED KINGDOM
   Misc AMERICANS, BRITONS, AUSTRALIANS, EUROPEAN, CONSERVATIVES, BRITISH
   Jig is up for UK Conservatives BILL MANDLE THE BRITISH local-govern ment election results have been devastating for the Conservatives. Whether such elections ought to be taken as a national barome ter is nowadays a lost argument. The Conservatives themselves, ,and John Major in particular, de SHOW ALL ght a national cam paign in the hope the national . opipi.on polls
2. **Promoted from 5 to 2:** 1994-06-22 Article 118173252, score 23.031384 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People NORMAN FOWLER, PAUL JUDGE, JOHN MAJOR
   Organisations TORY, TORIES
   Locations LONDON, BRITAIN
   Misc INDEPENDENT, CONSERVATIVES
   Major problems for unpopular PM 'There is not much point in spending money on a product that is unsaleable' LONDON: Big business and other financial supporters arc deserting Britain's ruling Con servative Party because of the unpopularity of the Prime Min ister, John Major, raising SHOW ALL ut the leadership's confident claims that the party's financial crisis
3. **Promoted from 9 to 3:** 1994-02-16 Article 134302359, score 22.971605 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People MR MAJOR, HARTLEY BOOTFFS, BILL CLINTON, PAUL KEATING, JOHN MAJOR
   Organisations IRA, CONSERVATIVE, CONSERVATIVE PARTY, SBS
   Locations NORTHERN IRELAND, AUSTRALIA, BRITAIN, UNITED STATES
   Momentum against Major The administration of Britain's Prime Minister, John Major, has been deeply hurt by the continuing sex scandals which have seen the political and in one case physical demise of no fewer than six of his Conservative colleagues in the past five months. The damage has been exacerbated by the fact that the scandals have emerged during his back to SHOW ALL paign which is designed to asso ciate the Conservative Party with so-called family
4. **Promoted from 13 to 4:** 1994-12-27 Article 127262781, score 22.799828 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]
   People ALBERT REYN, AL EXANDER DOWNER, QUEEN VICTORIA, KENNETH CLARKE, JOHN BRU, JOHN MAJOR, JOHN MAJOR MALCOLM BOOKER, HELMUT KOHL
   Organisations IRA, EUROPEAN UNION, PARLIAMENT, ROYAL FAMILY, IRISH REPUBLICAN ARMY, CONSERVA TIVE PARTY, SINN FEIN, STORMONT CASTLE
   Locations REPUBLIC OF IRELAND, NORTHERN IRELAND, EU ROPEAN UNION, BRITAIN, BELFAST, EUROPE, UNITED STATES, GERMANY, NORWAY
   Misc FRENCH, CATHOLICS, GERMAN, EURO PEAN, IRISH, EUROPEAN, ANTI-EURO, EXCHEQUER, NORWEGIANS, BRITISH
   The undoing of John Major Malcolm Booker THE CONSERVATIVE lead ers in Britain, as in Aus tralia, have withdrawn with relief into the Christmas holiday period. The Prime Minister, John Major, has had a bad ' year and his leadership has been repeatedly under threat. Like Al exander Downer, he may wonder how long he will survive in the new year. Not all his problems SHOW ALL his own fault. Ministers in his Government have been involved in bizarre sex

People
PAUL KEATING 128
JOHN HEWSON 86
JOHN MAJOR 79
ALEXANDER DOWNER 76
JOHN HOWARD 49
RALPH WILLIS 37
PETER BREWER 35
PETER COSTELLO 32
IAN HENDERSON 26
BOB HAWKE 25

Organisations
LABOR 88
PARLIAMENT 87
LIBERAL PARTY 70
UNITED NATIONS 66
ANZ 61
COALITION 60
BHP 59
WESTPAC 58
SENATE 52
RESERVE BANK 50

Locations
AUSTRALIA 306
SYDNEY 230
UNITED STATES 225
CANBERRA 132
LONDON 122
BRITAIN 109
MELBOURNE 103
EUROPE 76
AUSTRA 66
BOSNIA 55

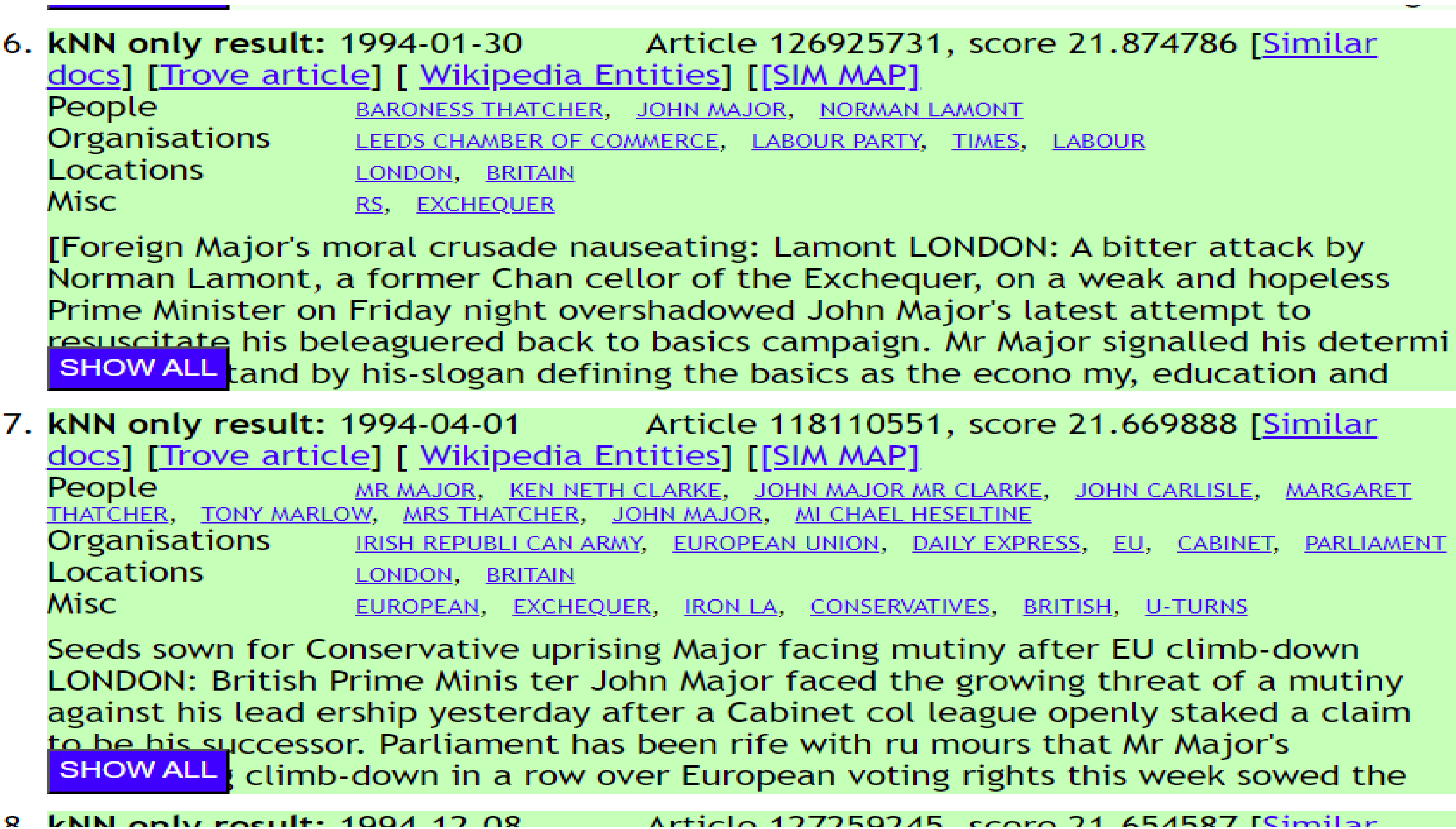
Misc
AUSTRALIAN 299
BRITISH 114
AUSTRALIANS 77
AMERICAN 67
LIBERALS 61
EUROPEAN 59
SERBS 52
SERB 47
FRENCH 46
BOSNIAN 44

https://nla-overproof.projectcomputing.com/knnBlend?set=1994&embedding=ada-002&stxt=the%20fall%20of%20John%20Major

The proof-of-concept explores a blended keyword and semantic search. It does this by:

1. Issuing a "standard" keyword and phrase text search with standard keyword ranking.
2. For each of the top-10 keyword-ranked results, fetch that document's embedding and use this to issue a semantic search to find other documents with embeddings most similar to it. The intent is to "enrich" the result candidates by including documents very similar to those best keyword results but which may not contain all the keywords.
3. Creating an embedding of the original search query and issuing a semantic search to find documents with embedding most similar to it.

Each of these searches produces a ranked list of documents with a search score (in this case, generated by Lucene). The score for the first search (the "standard" keyword and phrase search) is calculated by Lucene based on its default BM25 [10] ranking using keyword repository and document frequencies with a boost applied if all keywords were found nearby (ie, a phrase-like boost). The scores for the second and third searches are also calculated by Lucene based on the distance in vector-space between the embeddings of the search embeddings and the document embeddings. The proof-of-concept then applies a separate weight to the scores produced by the three types of searches and adds the results across all three searches to generate a document result set for ranking.



6. **kNN only result:** 1994-01-30 Article 126925731, score 21.874786 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]

People	BARONESS THATCHER, JOHN MAJOR, NORMAN LAMONT
Organisations	LEEDS CHAMBER OF COMMERCE, LABOUR PARTY, TIMES, LABOUR
Locations	LONDON, BRITAIN
Misc	RS, EXCHEQUER

[Foreign Major's moral crusade nauseating: Lamont LONDON: A bitter attack by Norman Lamont, a former Chan cellor of the Exchequer, on a weak and hopeless Prime Minister on Friday night overshadowed John Major's latest attempt to resuscitate his beleaguered back to basics campaign. Mr Major signalled his determi

SHOW ALL

tand by his-slogan defining the basics as the econo my, education and

7. **kNN only result:** 1994-04-01 Article 118110551, score 21.669888 [Similar docs] [Trove article] [Wikipedia Entities] [[SIM MAP]

People	MR MAJOR, KEN NETH CLARKE, JOHN MAJOR MR CLARKE, JOHN CARLISLE, MARGARET THATCHER, TONY MARLOW, MRS THATCHER, JOHN MAJOR, MI CHAEL HESELTINE
Organisations	IRISH REPUBLI CAN ARMY, EUROPEAN UNION, DAILY EXPRESS, EU, CABINET, PARLIAMENT
Locations	LONDON, BRITAIN
Misc	EUROPEAN, EXCHEQUER, IRON LA, CONSERVATIVES, BRITISH, U-TURNS

Seeds sown for Conservative uprising Major facing mutiny after EU climb-down LONDON: British Prime Minis ter John Major faced the growing threat of a mutiny against his lead ership yesterday after a Cabinet col league openly staked a claim to be his successor. Parliament has been rife with ru mours that Mr Major's

SHOW ALL

climb-down in a row over European voting rights this week sowed the

8. kNN only result: 1994-12-08 Article 127259245, score 21.654587 [Similar

Lexical vs semantic

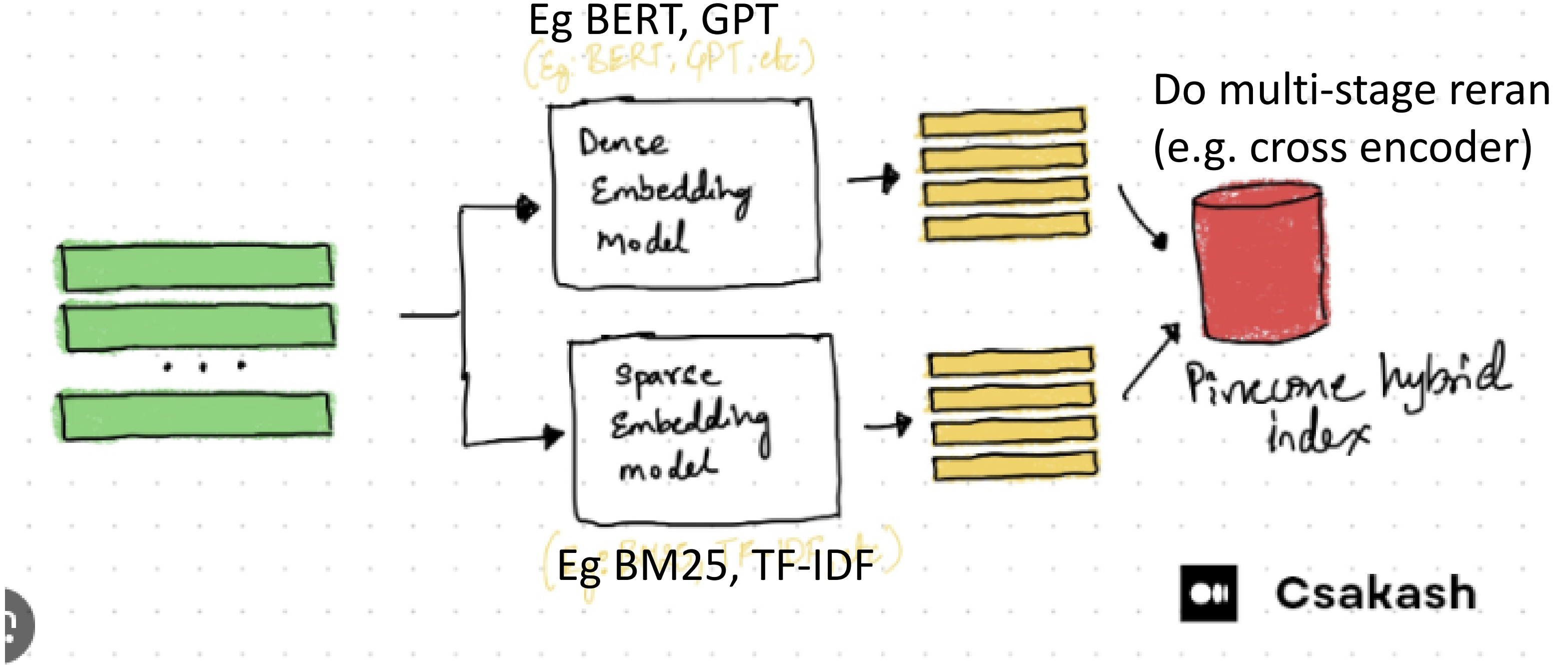
- Searching for a specific name like "John Major" favours lexical search as it involves direct word matches with little semantic complexity.
- A search for a single word like "kamikaze" also favors lexical search as it's likely the searcher wants documents containing that exact word.
- A query like "train crash" may involve broader semantic interpretations such as railway accidents or specific types of crashes, indicating the potential benefit of semantic search

Lexical vs semantic

- For longer queries, there is more “semantic intent” that is really hard to represent fairly with simple keywords, even when attempting to automatically expand the search with keywords.
- For example - "the fall of John Major", may need to match articles that talk about the ousting or downfall or undoing or unravelling or humiliation or collapse of support for John Major

- Allows you to find docs with newer jargon like "Cyber attack"

Hybrid search is the way to go to retrieve pool of documents to rank



Eg BERT, GPT

Do multi-stage reranking (e.g. cross encoder)

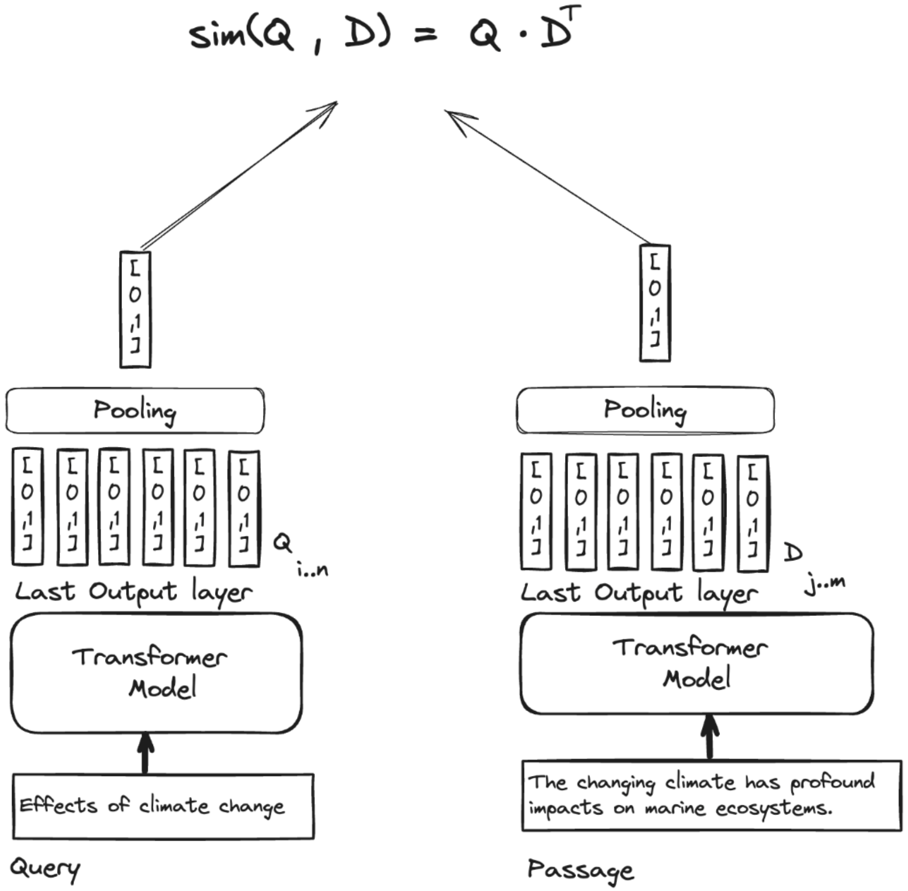
Eg BM25, TF-IDF

Empirically, first stage retrievers using BM25 and reranking top 10k results with cross encoders do well, but it is probably better to hedge bets by using lexical AND Semantic/dense embedding methods to retrieve documents

Can we have more transparent search?

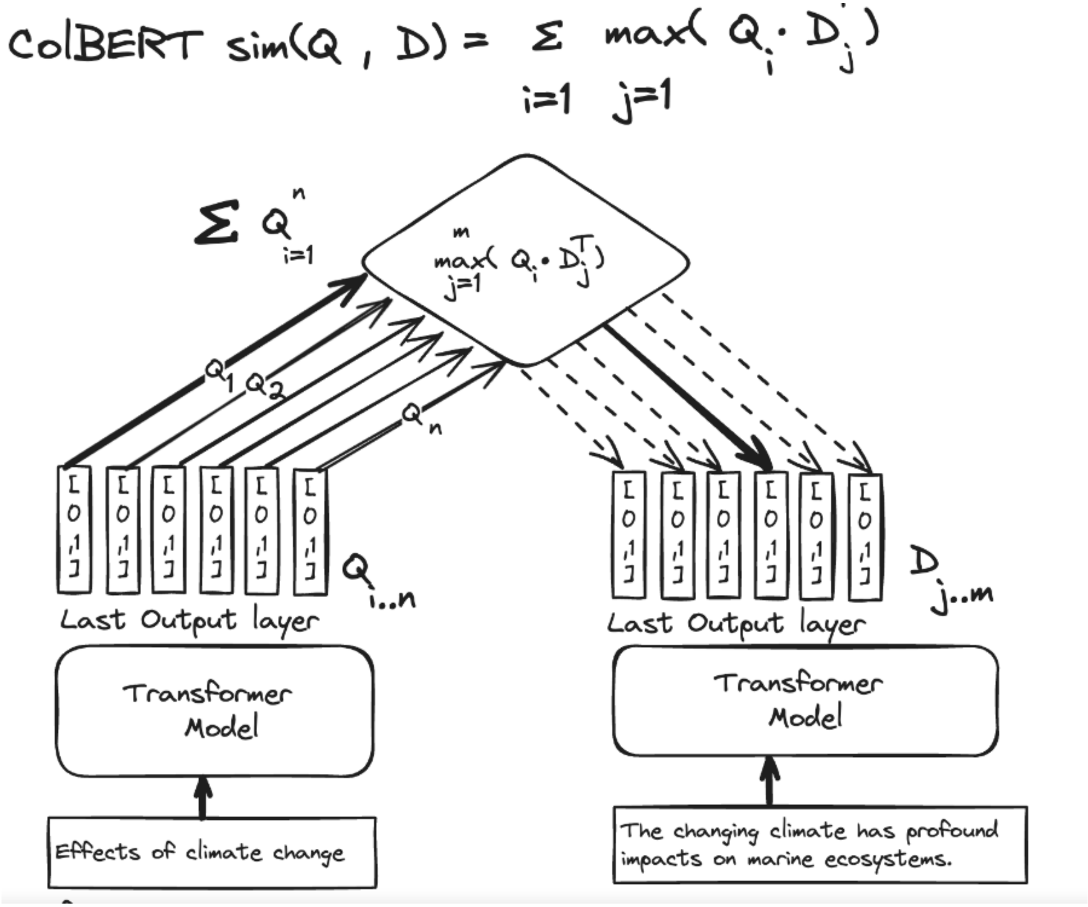
Bi-encoder = One embedding represent one doc is stored in index

Pooling is a way to average out multiple embeddings into one (see notebook for examples of different ways to pool)

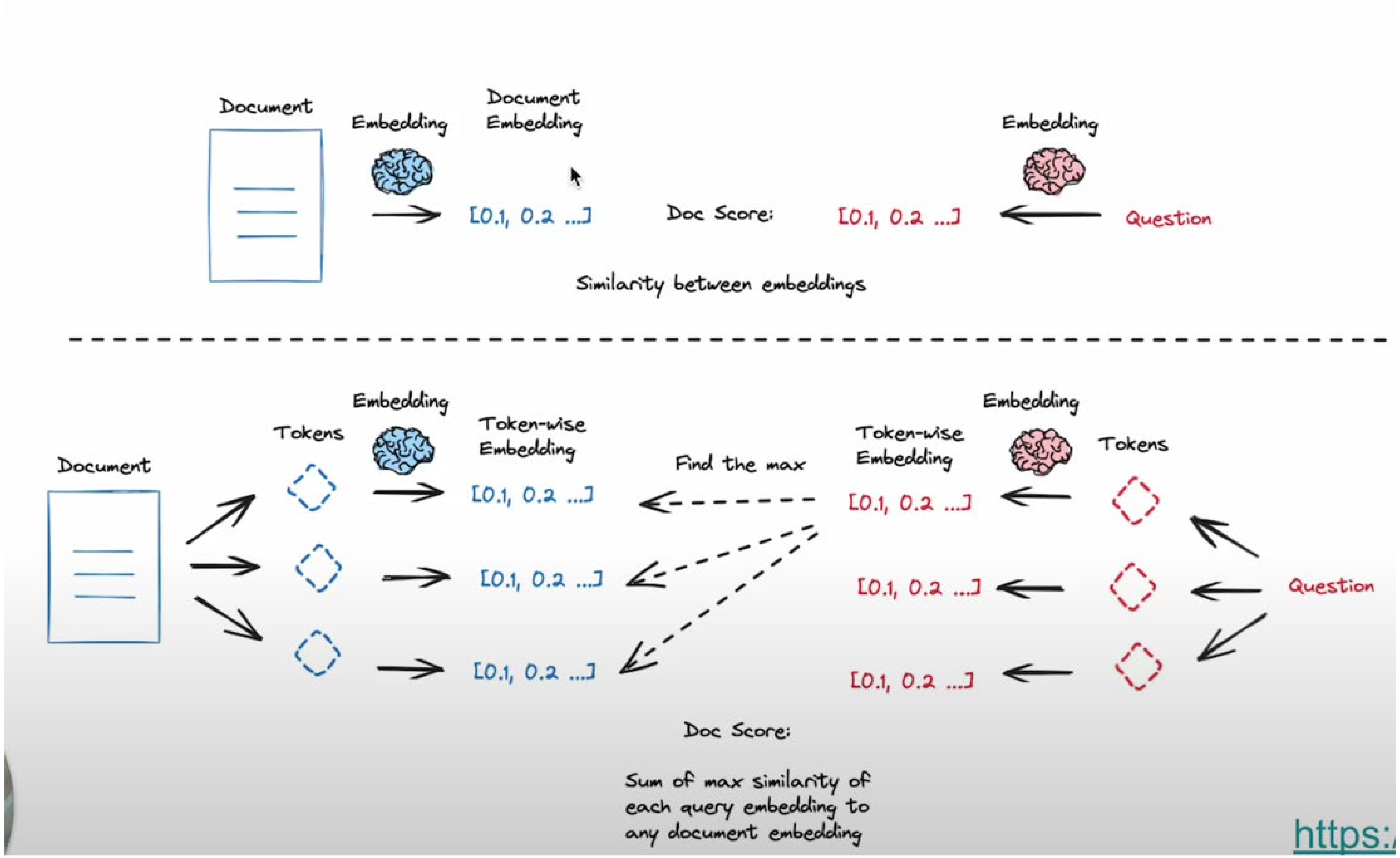


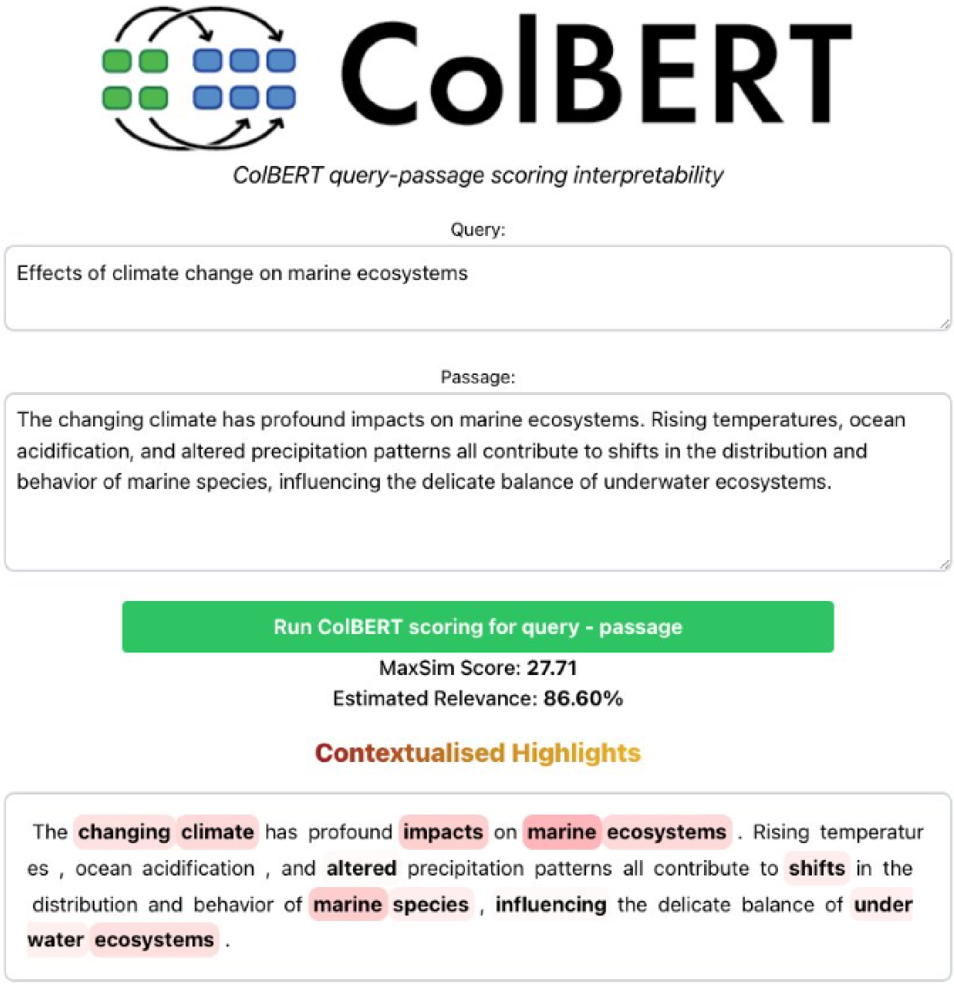
1. Can one embedding represent a doc well enough?
2. By representing the doc and query as individual embeddings and doing a similarity match, we are unable to see WHY the system thinks query and doc are similar because there is no token level information

COLBERT – Each token in doc is stored as embedding in index


ColBERT $sim(Q, D) = \sum_{i=1} \max_{j=1}(Q_i \cdot D_j)$
$\sum_{i=1} Q^n$
$\max_{j=1}^{m}(Q_i \cdot D_j^T)$
Q_1
Q_2
Q_n
[0,1]
$Q_{i..n}$
$D_{j..m}$
Last Output layer
Last Output layer
Transformer Model
Transformer Model
Effects of climate change
The changing climate has profound impacts on marine ecosystems.

COLBERT hybrid





ColBERT query-passage scoring interpretability

Query:

Effects of climate change on marine ecosystems

Passage:

The changing climate has profound impacts on marine ecosystems. Rising temperatures, ocean acidification, and altered precipitation patterns all contribute to shifts in the distribution and behavior of marine species, influencing the delicate balance of underwater ecosystems.

MaxSim Score: **27.71**

Estimated Relevance: **86.60%**

Contextualised Highlights

The **changing** **climate** has profound **impacts** on **marine** **ecosystems** . Rising temperatur es , ocean acidification , and **altered** precipitation patterns all contribute to **shifts** in the distribution and behavior of **marine** **species** , **influencing** the delicate balance of **under** **water** **ecosystems** .

https://colbert.aiserv.cloud/

COLBERT is interpretable because you can trace embedding into token level

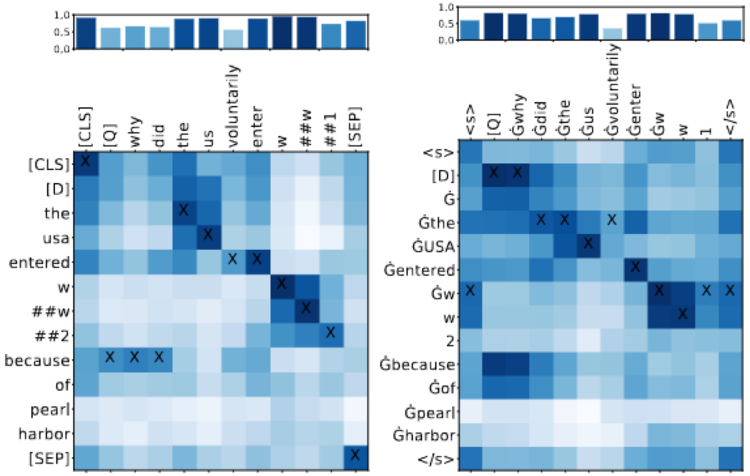
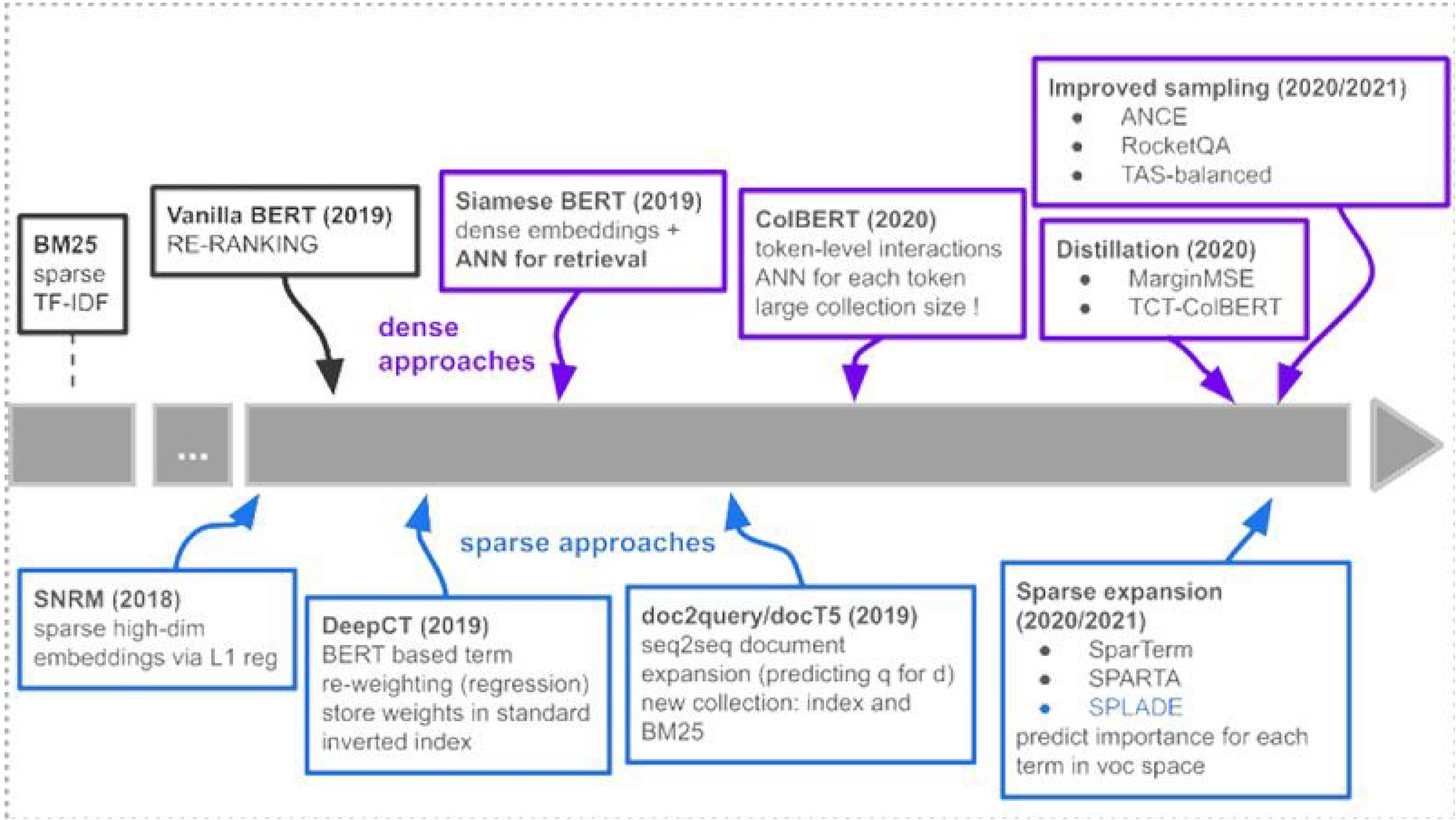


Figure 3: Late interaction diagrams for ColBERT and ColRoBERTa models between the query: *why did the us voluntarily enter ww1* and the document: *the usa entered ww2 because of pearl harbor*. For each column, the heatmap indicates the similarity scores among all the document embeddings for each query embedding, where the highest similarity score is highlighted with the symbol `X`. The top histogram depicts the magnitude of the contribution of the maximum similarity of each query embedding for the final relevance score between the query and document. The `[MASK]` tokens are omitted.

Improvements on sparse embedding using dense retrieval



BM25
sparse
TF-IDF

Vanilla BERT (2019)
RE-RANKING

Siamese BERT (2019)
dense embeddings +
ANN for retrieval

ColBERT (2020)
token-level interactions
ANN for each token
large collection size !

Improved sampling (2020/2021)
- ANCE
- RocketQA
- TAS-balanced

Distillation (2020)
- MarginMSE
- TCT-ColBERT

dense approaches

sparse approaches

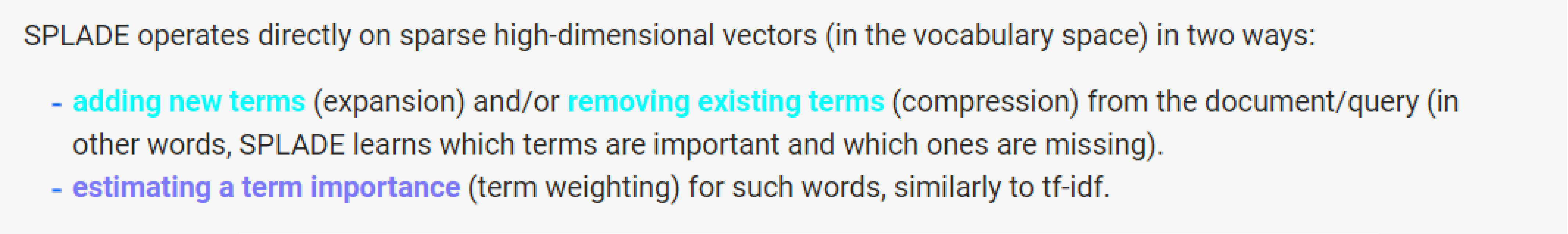
SNRM (2018)
sparse high-dim
embeddings via L1 reg

DeepCT (2019)
BERT based term
re-weighting (regression)
store weights in standard
inverted index

doc2query/docT5 (2019)
seq2seq document
expansion (predicting q for d)
new collection: index and
BM25

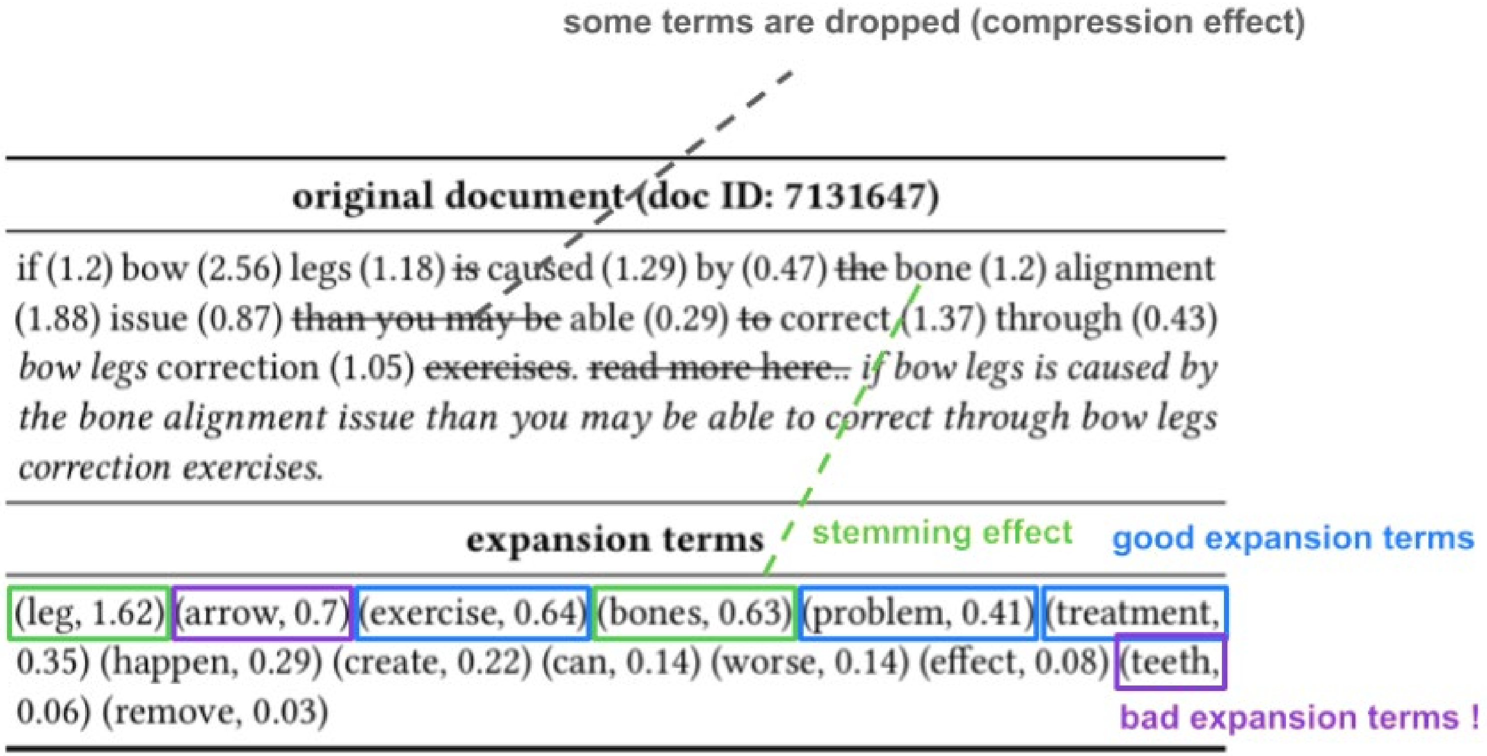
Sparse expansion (2020/2021)
- SparTerm
- SPARTA
- SPLADE

predict importance for each
term in voc space



SPLADE operates directly on sparse high-dimensional vectors (in the vocabulary space) in two ways:

- **adding new terms** (expansion) and/or **removing existing terms** (compression) from the document/query (in other words, SPLADE learns which terms are important and which ones are missing).
- **estimating a term importance** (term weighting) for such words, similarly to tf-idf.



some terms are dropped (compression effect)

original document (doc ID: 7131647)

if (1.2) bow (2.56) legs (1.18) ~~is~~ caused (1.29) by (0.47) ~~the~~ bone (1.2) alignment (1.88) issue (0.87) ~~than you may be~~ able (0.29) ~~to~~ correct (1.37) through (0.43) *bow legs* correction (1.05) ~~exercises~~. ~~read more here.~~. *if bow legs is caused by the bone alignment issue than you may be able to correct through bow legs correction exercises.*

expansion terms

stemming effect

good expansion terms

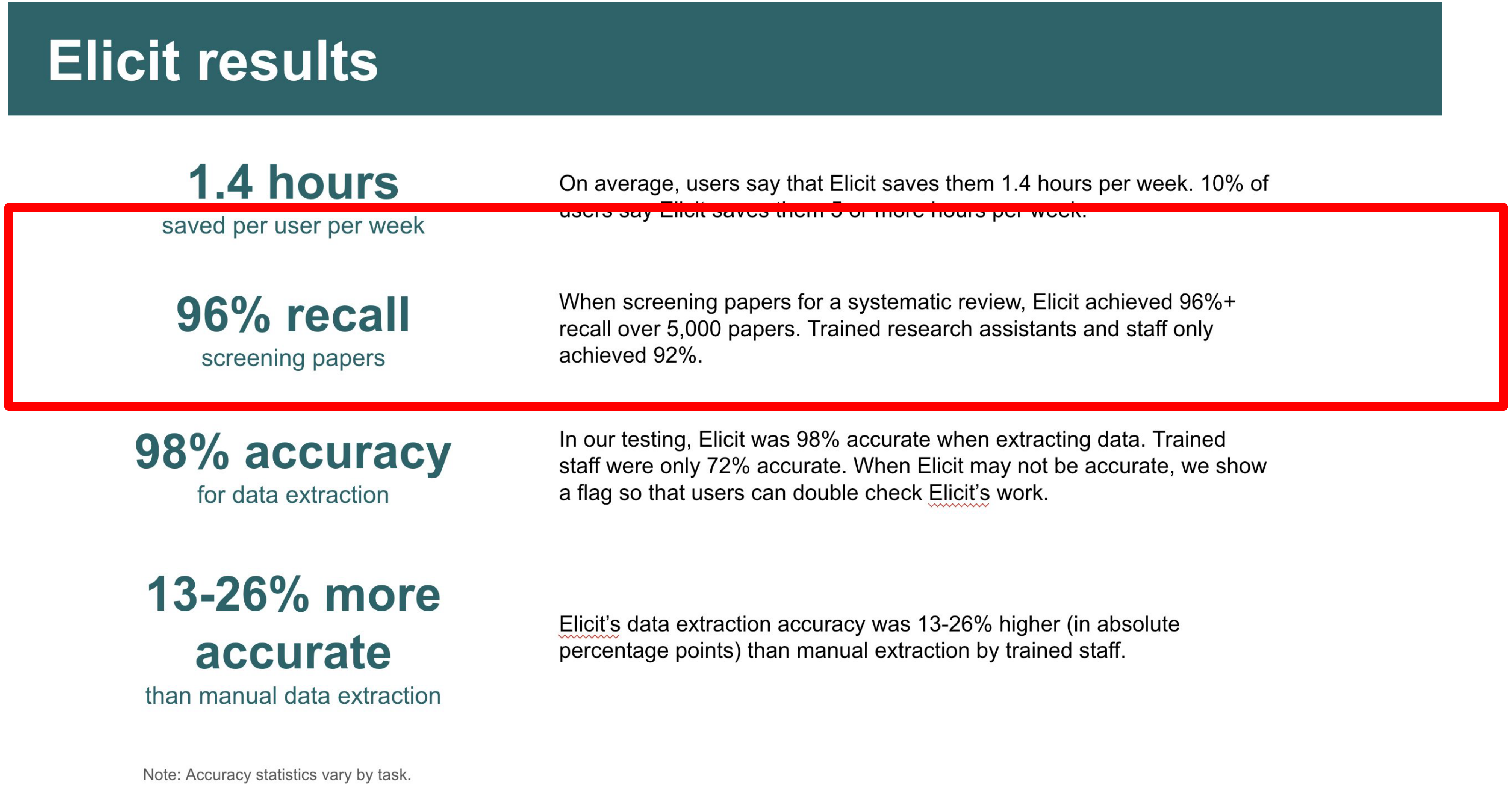
(leg, 1.62) (arrow, 0.7) (exercise, 0.64) (bones, 0.63) (problem, 0.41) (treatment, 0.35) (happen, 0.29) (create, 0.22) (can, 0.14) (worse, 0.14) (effect, 0.08) (teeth, 0.06) (remove, 0.03)

bad expansion terms !

SPLADE is interpretable!

Agent based search – Going beyond single search

Is this truly possible?



Elicit results

1.4 hours
saved per user per week

On average, users say that Elicit saves them 1.4 hours per week. 10% of users say Elicit saves them 5 or more hours per week.

96% recall
screening papers

When screening papers for a systematic review, Elicit achieved 96%+ recall over 5,000 papers. Trained research assistants and staff only achieved 92%.

98% accuracy
for data extraction

In our testing, Elicit was 98% accurate when extracting data. Trained staff were only 72% accurate. When Elicit may not be accurate, we show a flag so that users can double check Elicit's work.

13-26% more accurate
than manual data extraction

Elicit's data extraction accuracy was 13-26% higher (in absolute percentage points) than manual extraction by trained staff.

Note: Accuracy statistics vary by task.

Elicit etc despite it's innovations is still following the Google Scholar paradigm

- Does one single, non-iterative search
- Not optimised for recall, but precision
- Assesses relevancy based on a score e.g. cosine similarity, BM25
- Must return results as fast as possible (less than 2s)

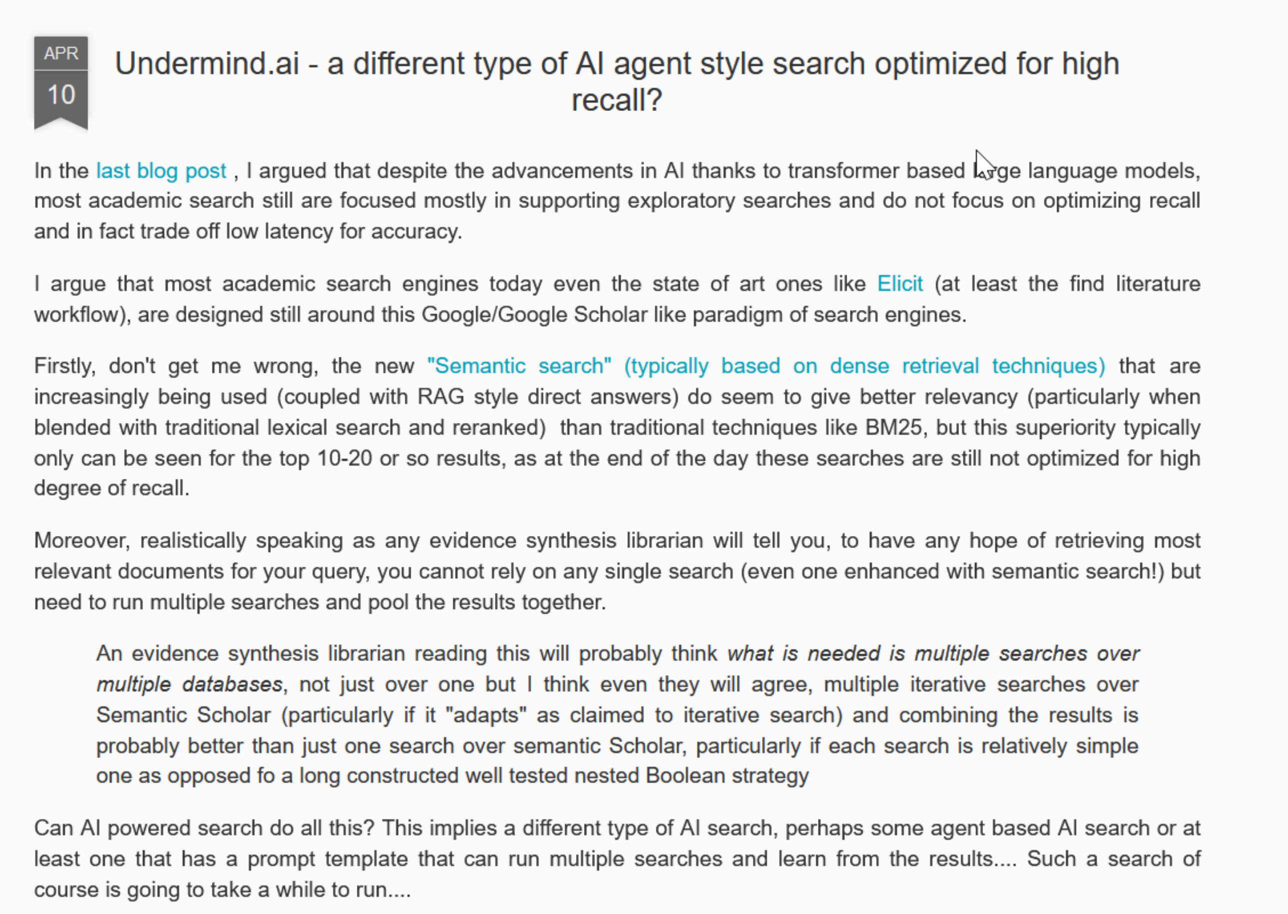
New agent based paradigm

- Does multiple, iterative searches
  - Does citation searching of papers

- Optimised for recall
  - Instead of using cosine similarity etc, using GPT4 level LLM to reason over papers - "Is paper X, relevant to the query?"

- Takes longer time to return results
  - >5 minutes

The tool is called undermind.ai = Semantic Scholar data + AI agent



APR
10

Undermind.ai - a different type of AI agent style search optimized for high recall?

In the last blog post , I argued that despite the advancements in AI thanks to transformer based large language models, most academic search still are focused mostly in supporting exploratory searches and do not focus on optimizing recall and in fact trade off low latency for accuracy.

I argue that most academic search engines today even the state of art ones like Elicit (at least the find literature workflow), are designed still around this Google/Google Scholar like paradigm of search engines.

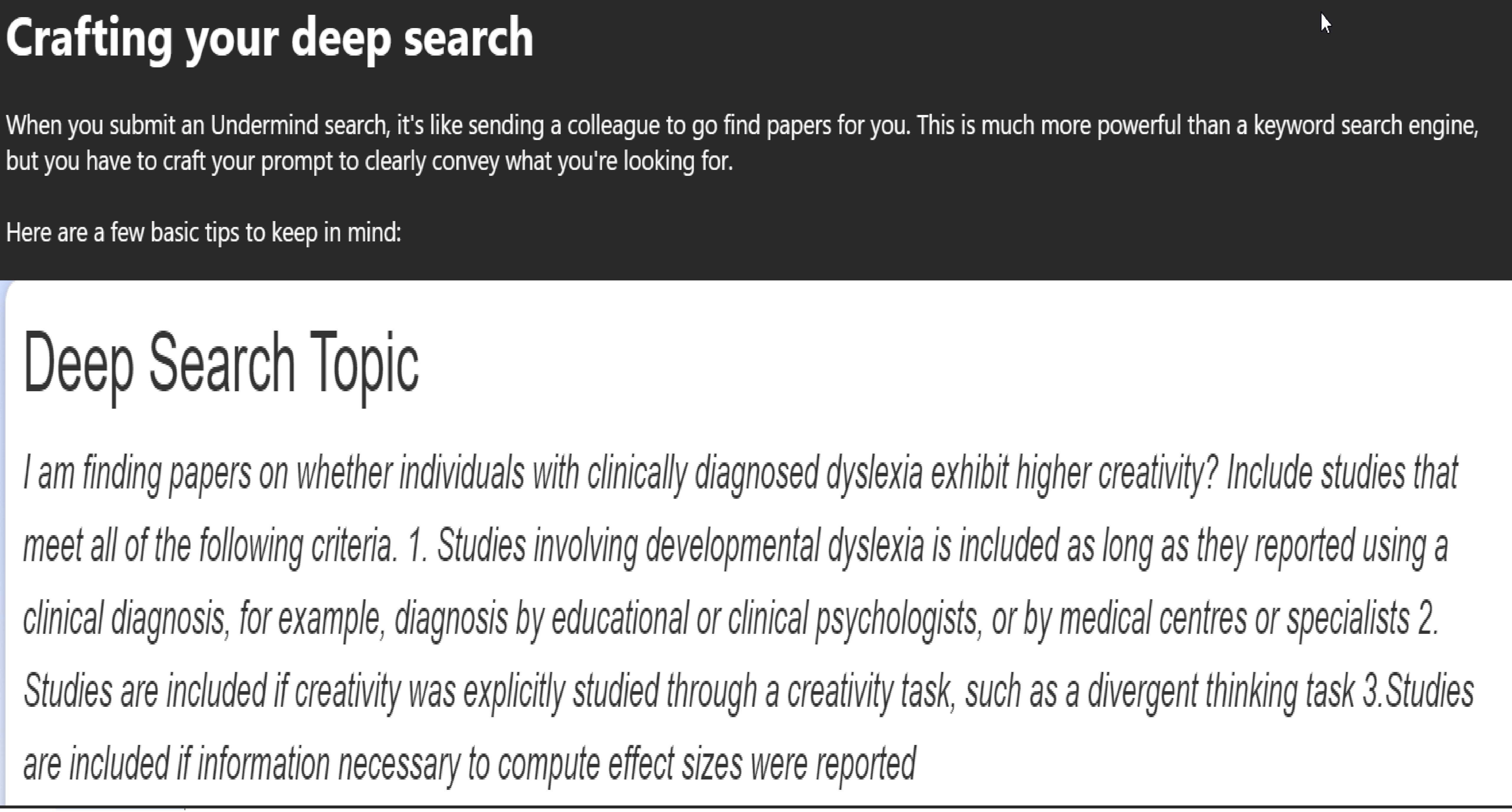
Firstly, don't get me wrong, the new "Semantic search" (typically based on dense retrieval techniques) that are increasingly being used (coupled with RAG style direct answers) do seem to give better relevancy (particularly when blended with traditional lexical search and reranked) than traditional techniques like BM25, but this superiority typically only can be seen for the top 10-20 or so results, as at the end of the day these searches are still not optimized for high degree of recall.

Moreover, realistically speaking as any evidence synthesis librarian will tell you, to have any hope of retrieving most relevant documents for your query, you cannot rely on any single search (even one enhanced with semantic search!) but need to run multiple searches and pool the results together.

> An evidence synthesis librarian reading this will probably think *what is needed is multiple searches over multiple databases*, not just over one but I think even they will agree, multiple iterative searches over Semantic Scholar (particularly if it "adapts" as claimed to iterative search) and combining the results is probably better than just one search over semantic Scholar, particularly if each search is relatively simple one as opposed fo a long constructed well tested nested Boolean strategy

Can AI powered search do all this? This implies a different type of AI search, perhaps some agent based AI search or at least one that has a prompt template that can run multiple searches and learn from the results.... Such a search of course is going to take a while to run....

My blog post explaining how Undermind.ai works



Crafting your deep search

When you submit an Undermind search, it's like sending a colleague to go find papers for you. This is much more powerful than a keyword search engine, but you have to craft your prompt to clearly convey what you're looking for.

Here are a few basic tips to keep in mind:

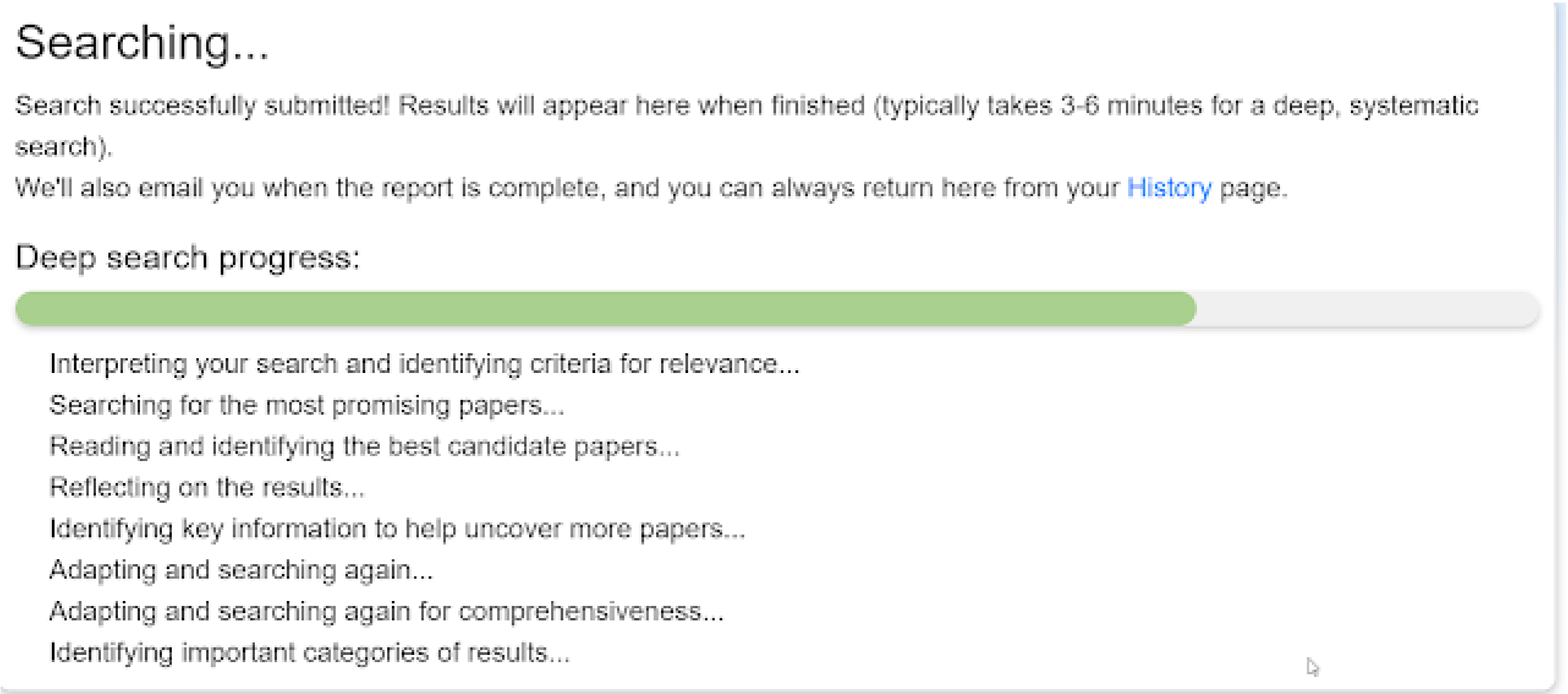
Deep Search Topic

I am finding papers on whether individuals with clinically diagnosed dyslexia exhibit higher creativity? Include studies that meet all of the following criteria. 1. Studies involving developmental dyslexia is included as long as they reported using a clinical diagnosis, for example, diagnosis by educational or clinical psychologists, or by medical centres or specialists 2. Studies are included if creativity was explicitly studied through a creativity task, such as a divergent thinking task 3.Studies are included if information necessary to compute effect sizes were reported

1.1 How it works

There are four steps to Undermind's algorithm:

1. **Basic search:** We identify promising candidate papers using a custom algorithm that combines semantic vector embeddings, citations, and language model reasoning.
2. **Relevance classification:** Given your search query, a high quality language model (GPT-4) accurately classifies each candidate paper based on its full text into 3 categories: highly relevant, closely related (meaning relevant, but slightly off-topic), or ignorable. See Appendix 3.2 for classification accuracy statistics.[2]
3. **Adaptation and exploration:** The algorithm adapts and searches again based on the relevant content it has discovered. This adaptation, which mimics a human's discovery process, makes it possible to uncover every relevant result.
4. **Estimating comprehensiveness:** Undermind tracks how frequently it discovers relevant papers during each search. Undermind initially finds many relevant results, but over time diminishing returns set in, empirically leading to "discovery curves" which are exponential in form (see Fig. 1(b)). Modeling this process allows us to determine when Undermind has found nearly all the relevant works.



Searching...

Search successfully submitted! Results will appear here when finished (typically takes 3-6 minutes for a deep, systematic search).

We'll also email you when the report is complete, and you can always return here from your History page.

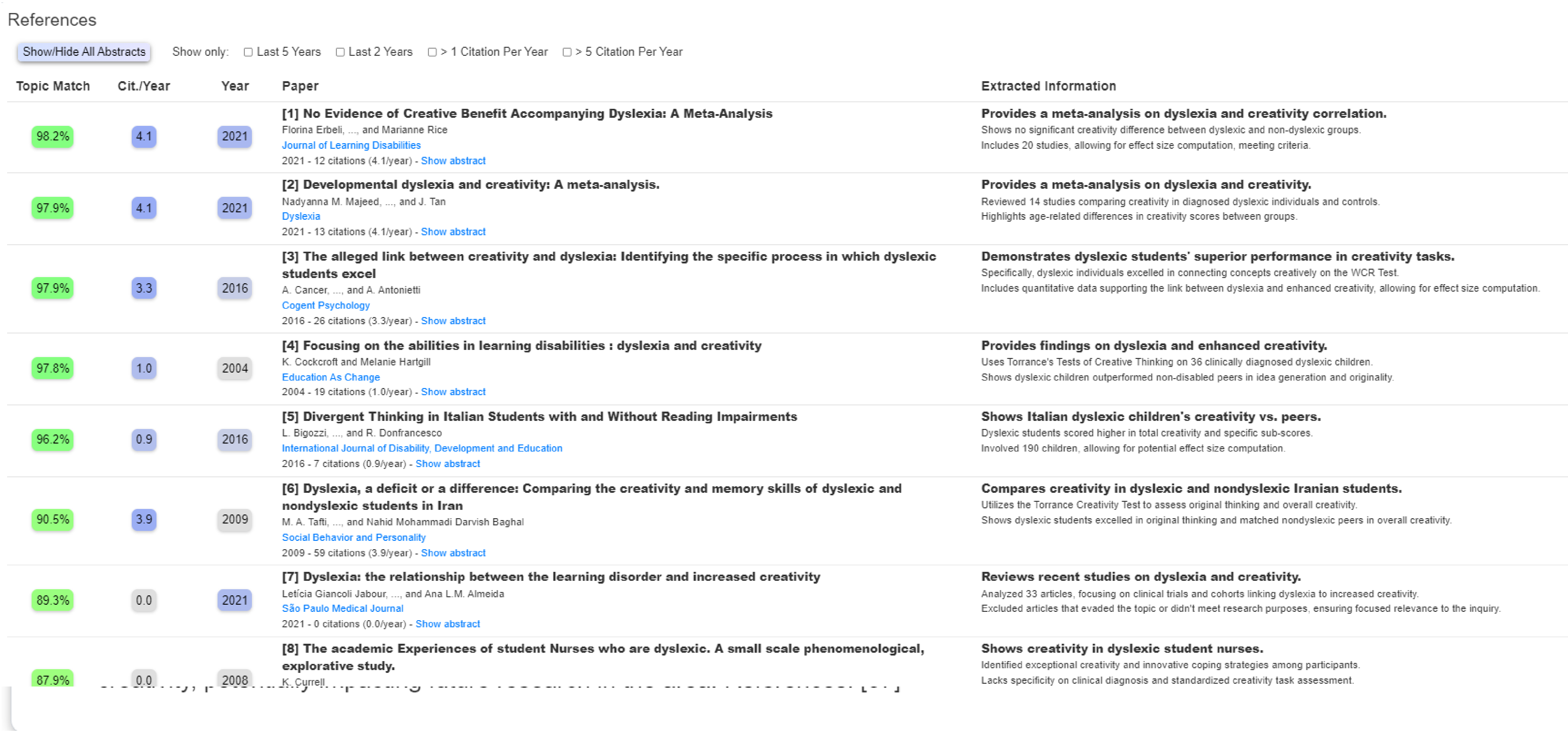
Deep search progress:

Interpreting your search and identifying criteria for relevance...
Searching for the most promising papers...
Reading and identifying the best candidate papers...
Reflecting on the results...
Identifying key information to help uncover more papers...
Adapting and searching again...
Adapting and searching again for comprehensiveness...
Identifying important categories of results...

Human Judgment	**Undermind Classification Probability**		
	Highly relevant	Closely related	Not relevant
Highly relevant	$75.9\%^{+9.8}_{-6.4}$	$22.0\%^{+9.6}_{-6.2}$	$1.8\%^{+4.7}_{-1.0}$
Closely related	$17.3\%^{+9.7}_{-5.8}$	$73.0\%^{+10.9}_{-7.0}$	$9.2\%^{+8.0}_{-4.1}$

Undermind Judgment	**Human Classification Probability**		
	Highly relevant	Closely related	Not relevant
Highly relevant	$83.3\%^{+9.4}_{-5.6}$	$16.7\%^{+9.4}_{-5.6}$	$0.0\%^{+3.8}_{-0.0}$
Closely related	$24.0\%^{+10.2}_{-6.5}$	$69.0\%^{+10.9}_{-7.2}$	$7.6\%^{+7.2}_{-3.6}$

Table 2: **Conditional classification rates.** Top: Undermind classification probabilities conditioned on human judgements. Associated upper and lower 95% confidence intervals are shown. Bottom: Human classification probabilities conditioned on Undermind judgement of a paper as highly relevant or closely related. For each table, note the far right column, which gives the probability that a truly relevant paper is missed (upper table) or the probability that a paper emphasized by Undermind is irrelevant (lower table).



References

Show only: ☐ Last 5 Years ☐ Last 2 Years ☐ > 1 Citation Per Year ☐ > 5 Citation Per Year

Topic Match	Cit./Year	Year	Paper	Extracted Information
98.2%	4.1	2021	**[1] No Evidence of Creative Benefit Accompanying Dyslexia: A Meta-Analysis** Florina Erbeli, ..., and Marianne Rice Journal of Learning Disabilities 2021 - 12 citations (4.1/year) - Show abstract	**Provides a meta-analysis on dyslexia and creativity correlation.** Shows no significant creativity difference between dyslexic and non-dyslexic groups. Includes 20 studies, allowing for effect size computation, meeting criteria.
97.9%	4.1	2021	**[2] Developmental dyslexia and creativity: A meta-analysis.** Nadyanna M. Majeed, ..., and J. Tan Dyslexia 2021 - 13 citations (4.1/year) - Show abstract	**Provides a meta-analysis on dyslexia and creativity.** Reviewed 14 studies comparing creativity in diagnosed dyslexic individuals and controls. Highlights age-related differences in creativity scores between groups.
97.9%	3.3	2016	**[3] The alleged link between creativity and dyslexia: Identifying the specific process in which dyslexic students excel** A. Cancer, ..., and A. Antonietti Cogent Psychology 2016 - 26 citations (3.3/year) - Show abstract	**Demonstrates dyslexic students' superior performance in creativity tasks.** Specifically, dyslexic individuals excelled in connecting concepts creatively on the WCR Test. Includes quantitative data supporting the link between dyslexia and enhanced creativity, allowing for effect size computation.
97.8%	1.0	2004	**[4] Focusing on the abilities in learning disabilities : dyslexia and creativity** K. Cockcroft and Melanie Hartgill Education As Change 2004 - 19 citations (1.0/year) - Show abstract	**Provides findings on dyslexia and enhanced creativity.** Uses Torrance's Tests of Creative Thinking on 36 clinically diagnosed dyslexic children. Shows dyslexic children outperformed non-disabled peers in idea generation and originality.
96.2%	0.9	2016	**[5] Divergent Thinking in Italian Students with and Without Reading Impairments** L. Bigozzi, ..., and R. Donfrancesco International Journal of Disability, Development and Education 2016 - 7 citations (0.9/year) - Show abstract	**Shows Italian dyslexic children's creativity vs. peers.** Dyslexic students scored higher in total creativity and specific sub-scores. Involved 190 children, allowing for potential effect size computation.
90.5%	3.9	2009	**[6] Dyslexia, a deficit or a difference: Comparing the creativity and memory skills of dyslexic and nondyslexic students in Iran** M. A. Tafti, ..., and Nahid Mohammadi Darvish Baghal Social Behavior and Personality 2009 - 59 citations (3.9/year) - Show abstract	**Compares creativity in dyslexic and nondyslexic Iranian students.** Utilizes the Torrance Creativity Test to assess original thinking and overall creativity. Shows dyslexic students excelled in original thinking and matched nondyslexic peers in overall creativity.
89.3%	0.0	2021	**[7] Dyslexia: the relationship between the learning disorder and increased creativity** Letícia Giancoli Jabour, ..., and Ana L.M. Almeida São Paulo Medical Journal 2021 - 0 citations (0.0/year) - Show abstract	**Reviews recent studies on dyslexia and creativity.** Analyzed 33 articles, focusing on clinical trials and cohorts linking dyslexia to increased creativity. Excluded articles that evaded the topic or didn't meet research purposes, ensuring focused relevance to the inquiry.
87.9%	0.0	2008	**[8] The academic Experiences of student Nurses who are dyslexic. A small scale phenomenological, explorative study.** K. Currell	**Shows creativity in dyslexic student nurses.** Identified exceptional creativity and innovative coping strategies among participants. Lacks specificity on clinical diagnosis and standardized creativity task assessment.

How can information specialist adapt

- Study information retrieval formally
- Advocate for features and interface changes for users
- Test formally and with rigor new tools like Elicit and Undermind.ai

Questions??