seqout

Search engine for the world's genomic data

Aniruddha Mukherjee
Grad student, IIT Bombay

+

Mukesh Reddy
Undergrad student, NIT Warangal

A bit about us



Aniruddha Mukherjee

MTech student, IIT Bombay

I work on developing software for computational biologists at Saket Lab (saketlab.org). I am `@amkhrjee` everywhere.

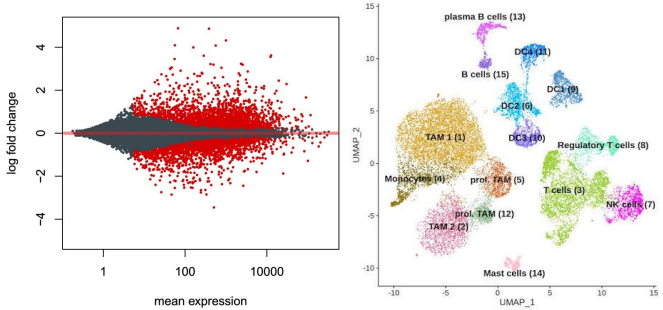


Mukesh Reddy

Undergrad student, NIT Warangal

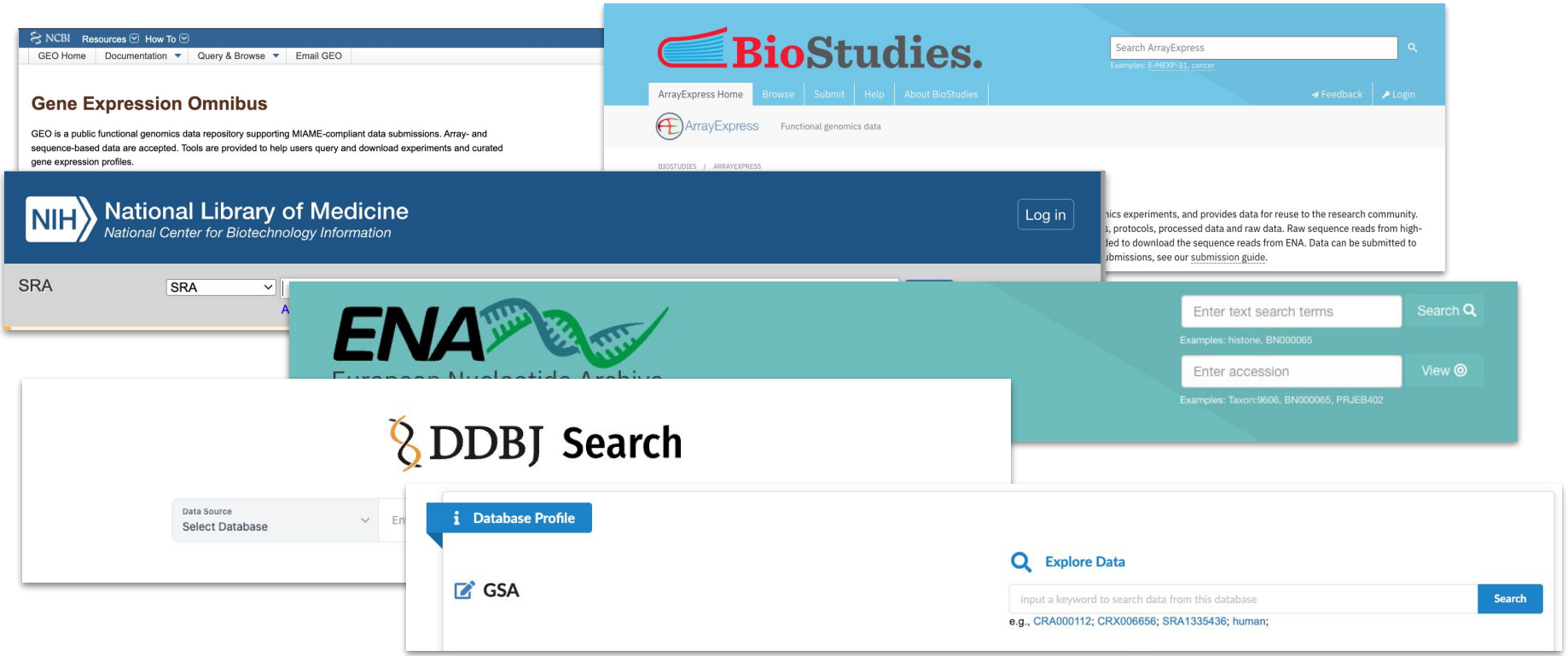
I mainly tinker around with recreational programming and I like cats. More about myself at `0xmukesh.github.io`

So you're looking for genomic datasets...



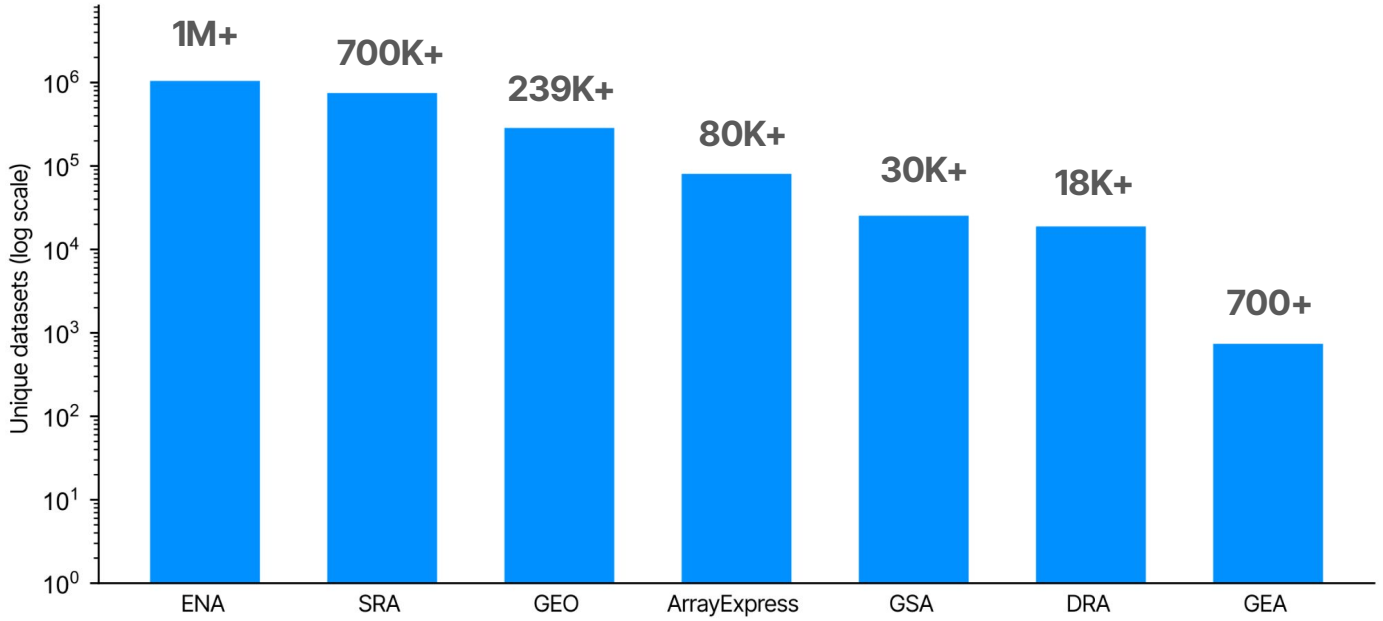
Gene expression data: how different genes are expressed in different conditions in different cells

Where to look for datasets?



All of them have their own unique data!

(that you probably don't want to miss out on)



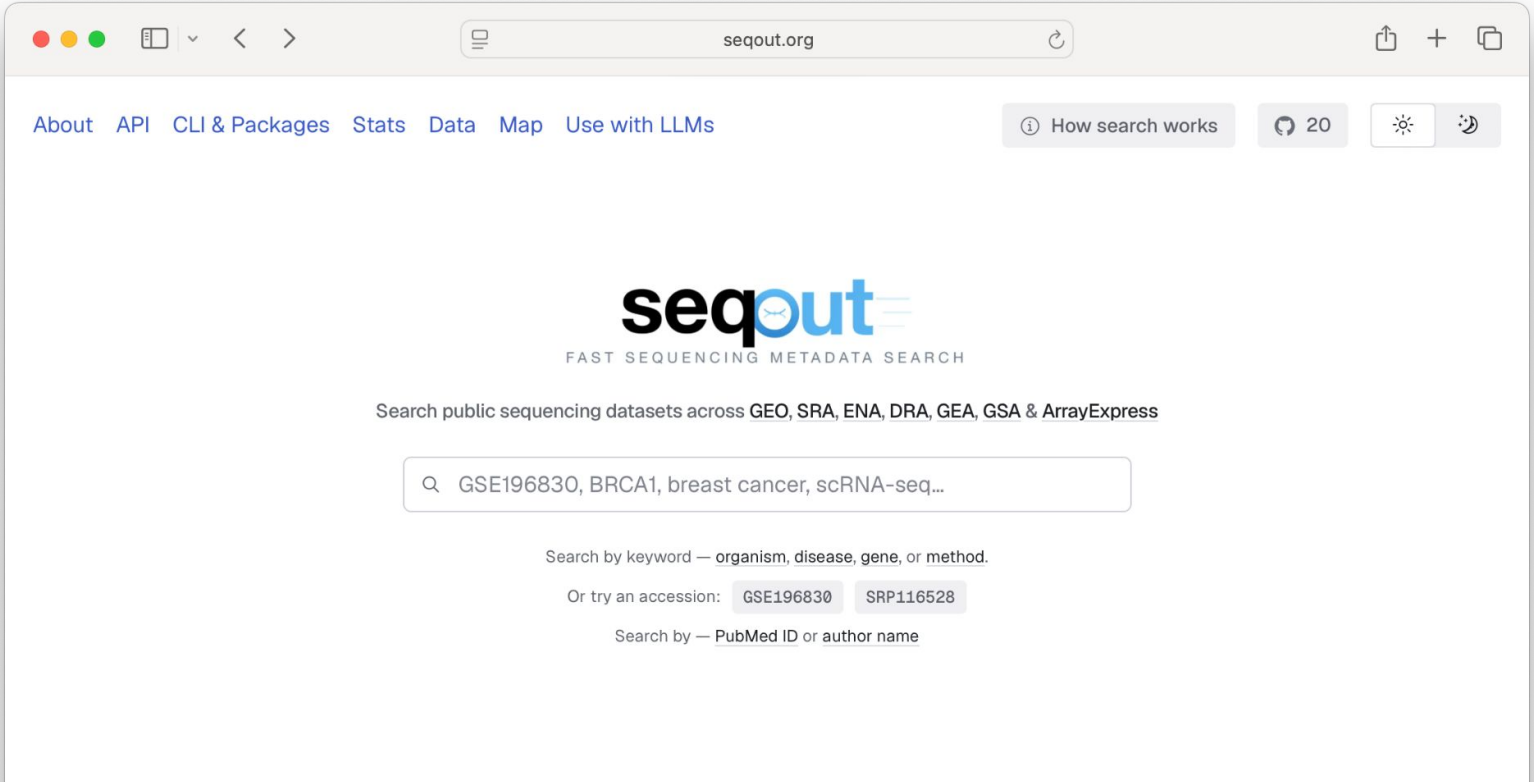
Want to search across all public datasets?

❌ All databases have different search mechanisms

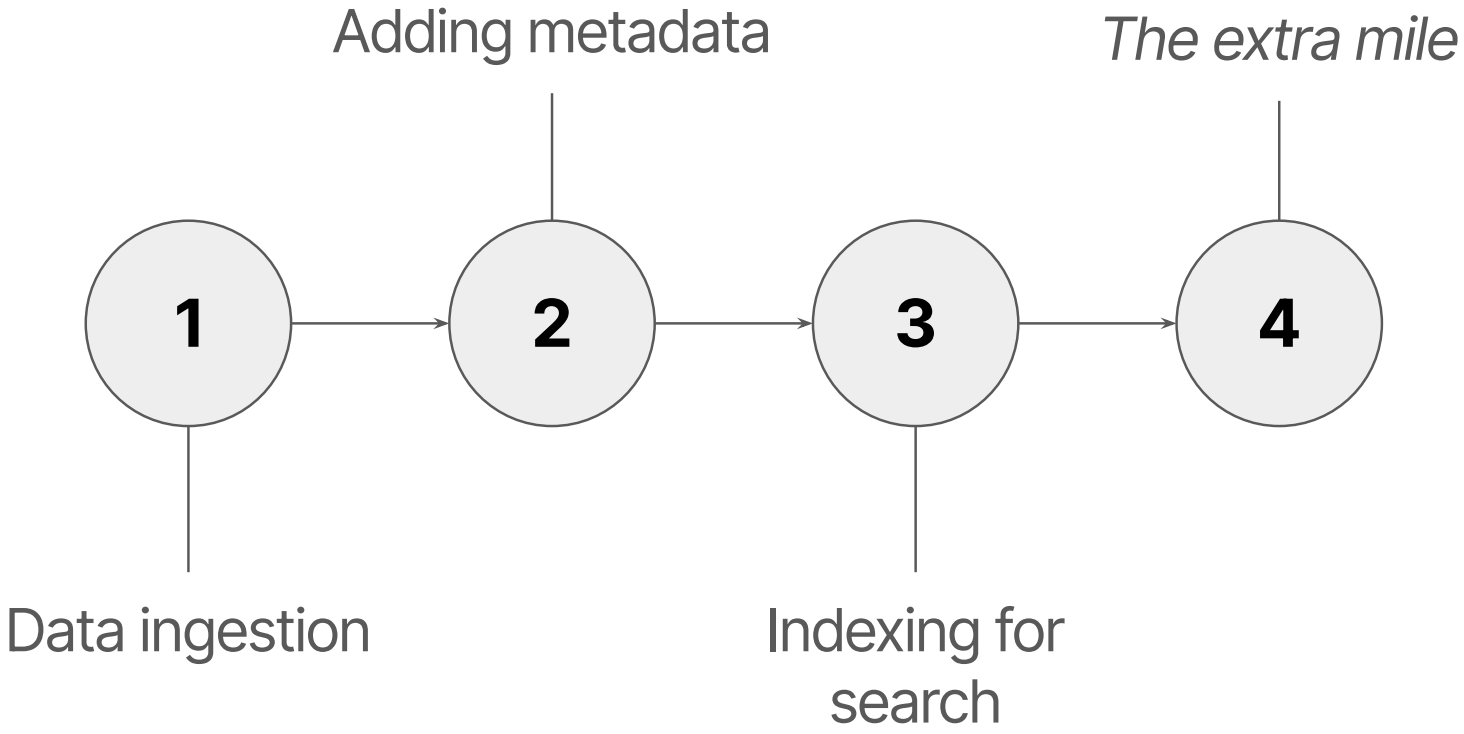
❌ No unified interface that pools data from all of them

❌ Strict API rate limits!

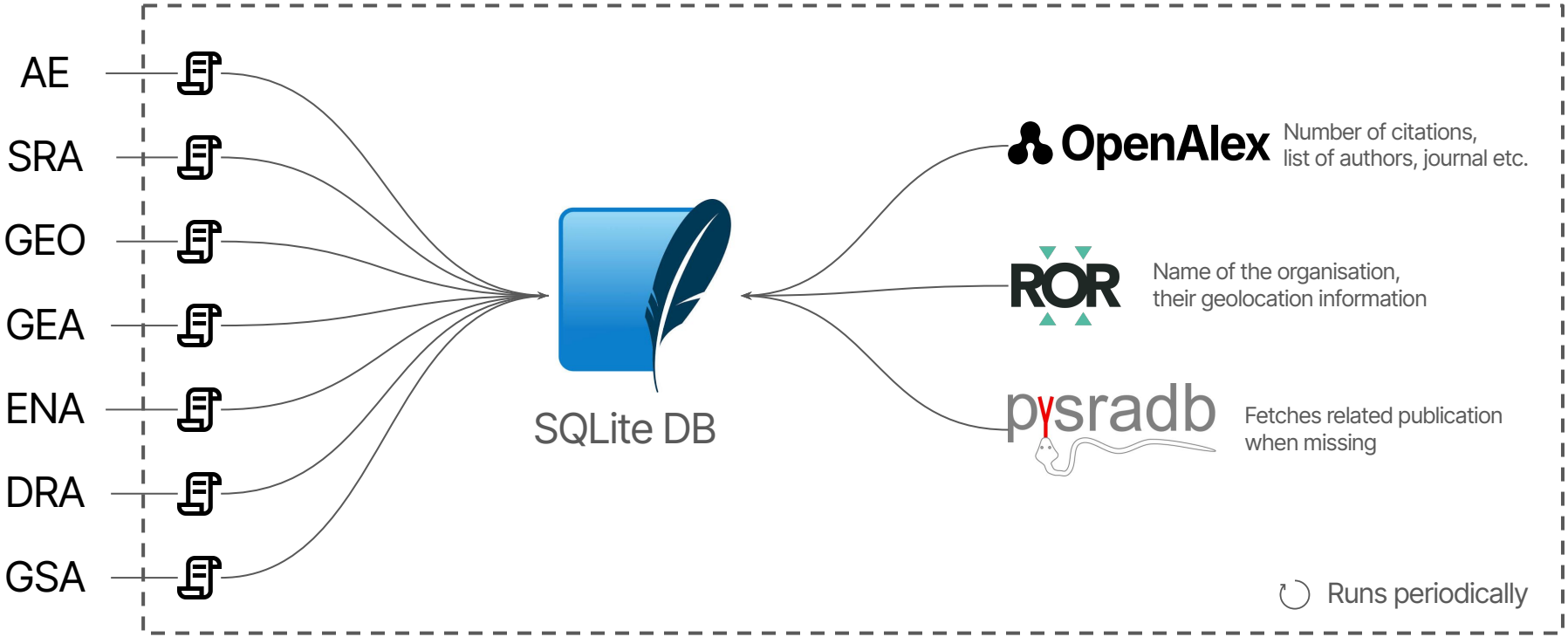
Now you can!


seqout.org
About
API
CLI & Packages
Stats
Data
Map
Use with LLMs
How search works
20
seqout
FAST SEQUENCING METADATA SEARCH
Search public sequencing datasets across GEO, SRA, ENA, DRA, GEA, GSA & ArrayExpress
GSE196830, BRCA1, breast cancer, scRNA-seq...
Search by keyword — organism, disease, gene, or method.
Or try an accession:
GSE196830
SRP116528
Search by — PubMed ID or author name

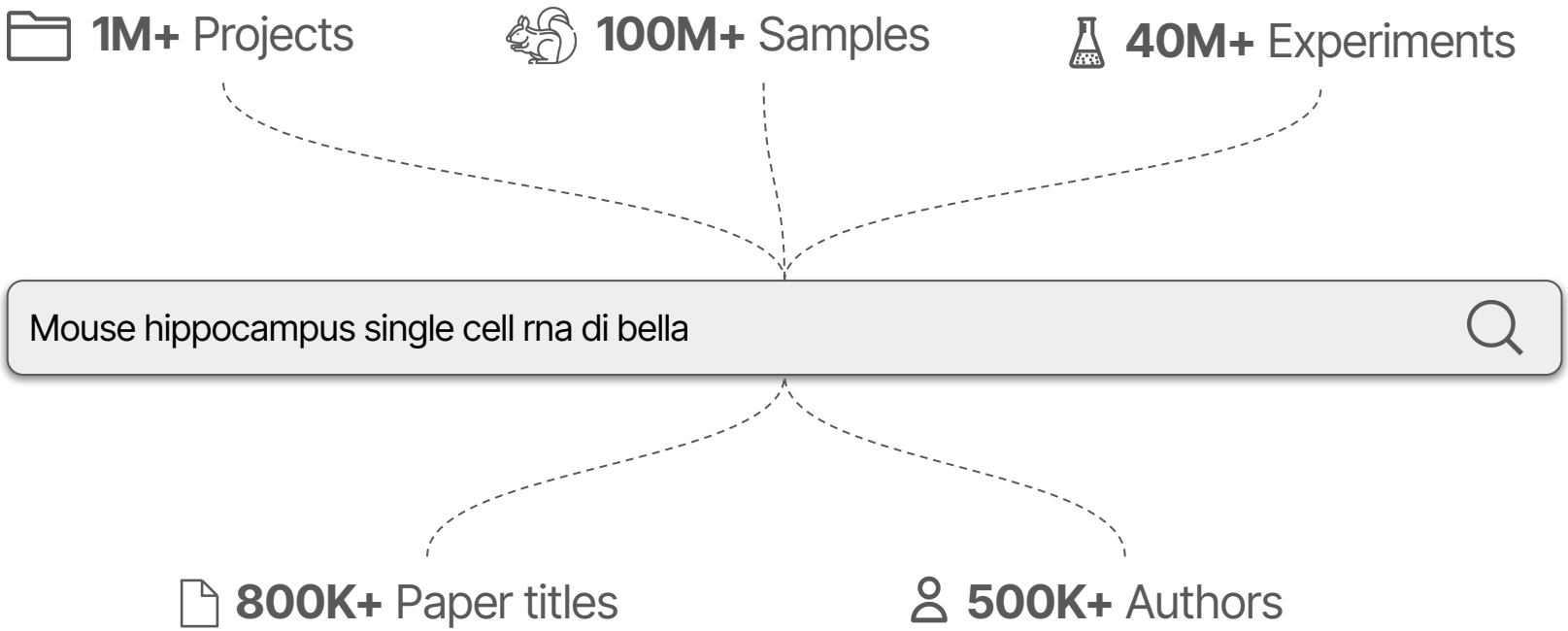
Search Engine 101



Data ingestion & metadata enrichment

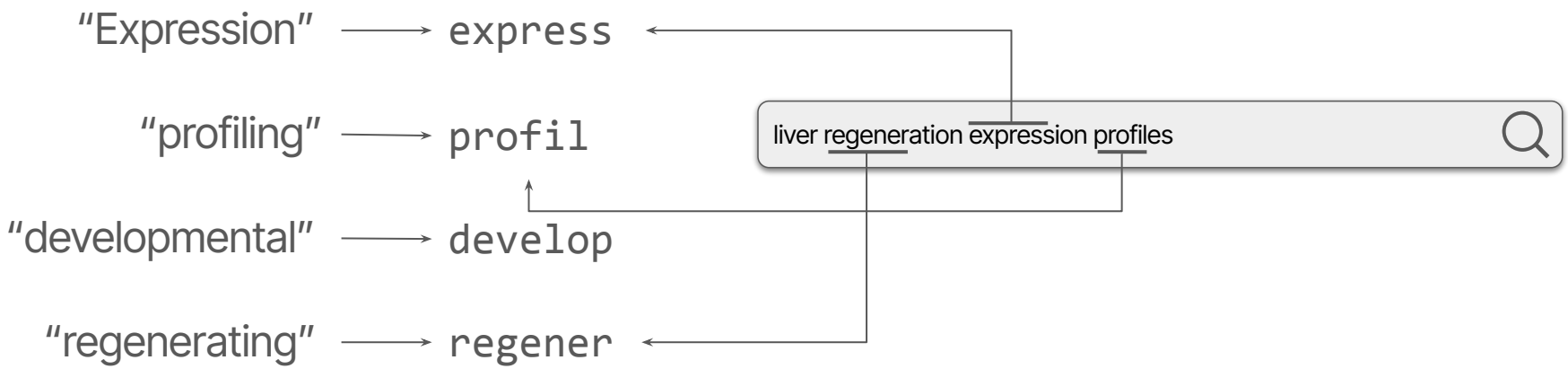


Searching through different metadata



Word stemming

GSE6998	*Expression profiling of developmental and regenerating liver in mice*



Generalized Inverted Index (aka GIN Index)

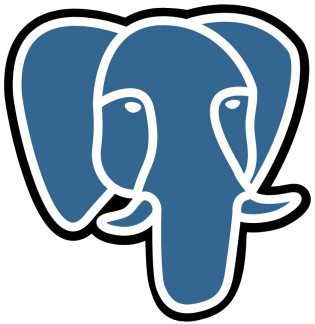


Full-text Search (FTS)

elastic

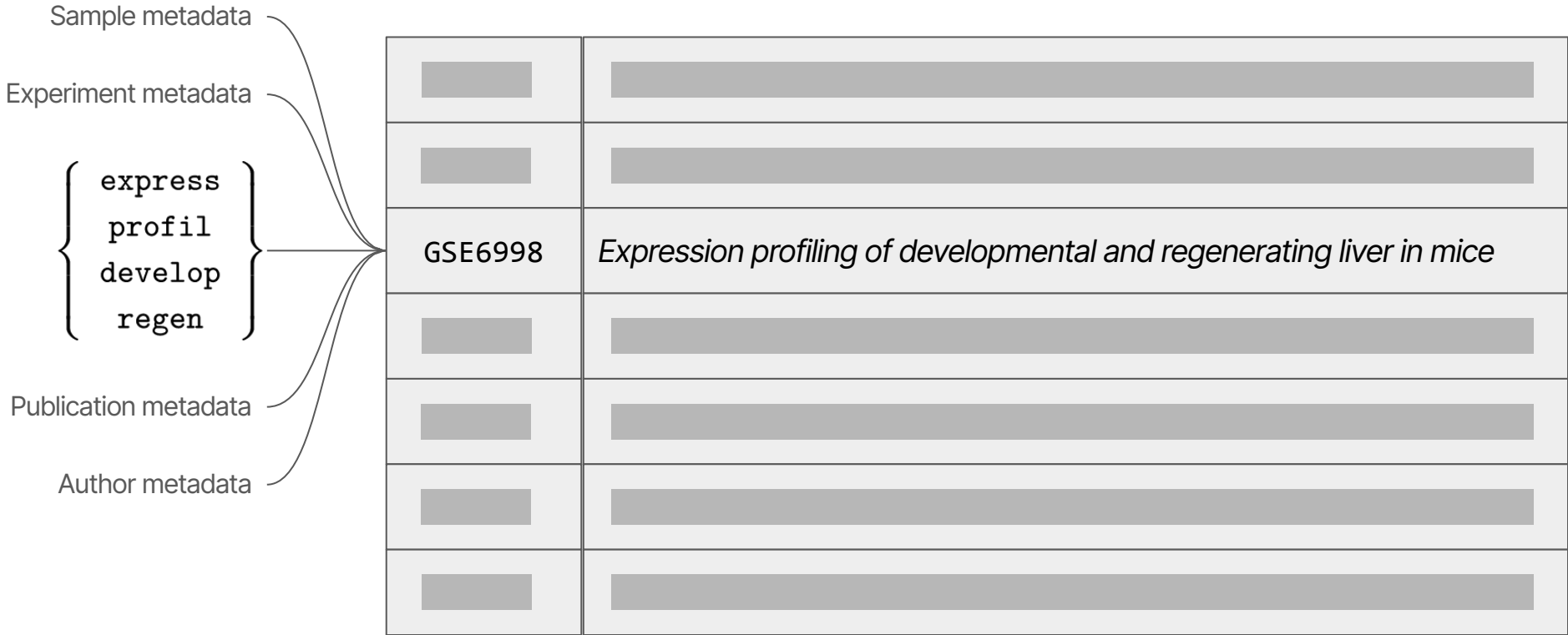
OpenSearch

meilisearch

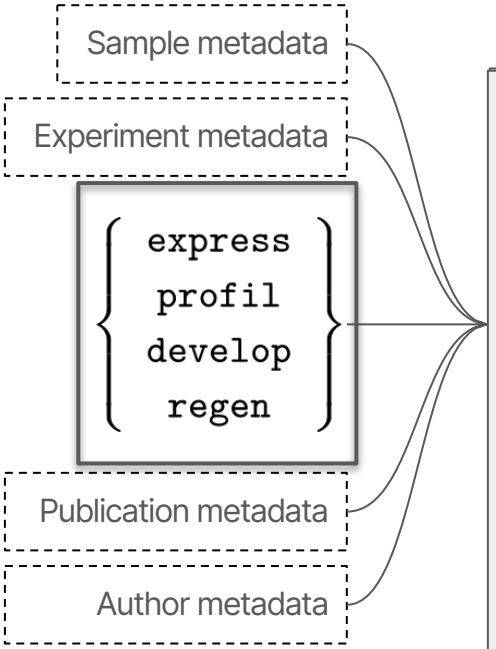


✅ Free & open source
✅ Built directly into Postgres
✅ Supports SQL
✅ No additional setup required

GIN Indexes with all metadata



GIN Indexes with all metadata (with weights!)




```
setweight(to_tsvector('english', title), 'A') ||
setweight(to_tsvector('english', sample_metadata), 'B') ||
setweight(to_tsvector('english', experiment_metadata), 'C') ||
setweight(to_tsvector('english', author_metadata), 'D') ||
setweight(to_tsvector('english', paper_metadata), 'D'),
```

But, is string matching enough?

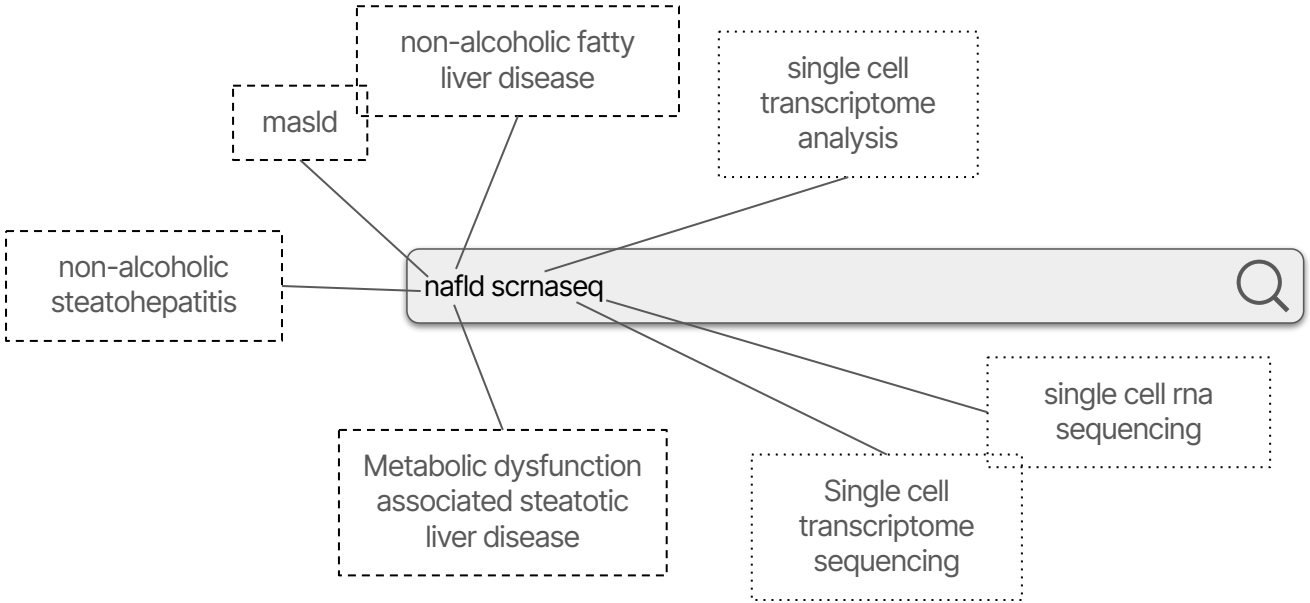


Metabolic dysfunction associated steatotic liver disease ....
..... single cell transcriptome analysis

GSEXXXXX | Journal of ... | 20 citations

With exact string matching, we will always miss out on this dataset!

Biological terms have lots of synonyms



Graph database to the rescue!

mondo

uberon

efo

HGNC
HUGO Gene Nomenclature Committee

MeSH

Disease terms

Anatomy terms

Experimental terms

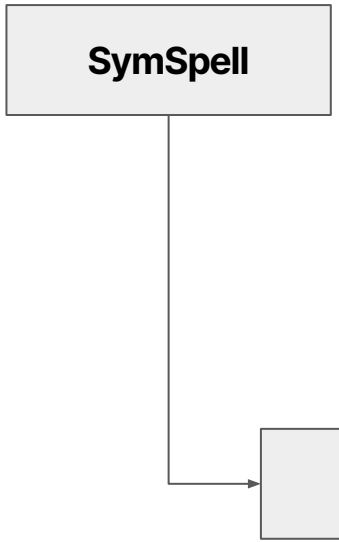
Gene names

Biomedical terms



neo4j graph of ontology terms

But what about typos?



Precomputes variants of dictionary words for fast spelling correction

“Livre” gets corrected to “Liver”

But what if it’s not a dictionary term?

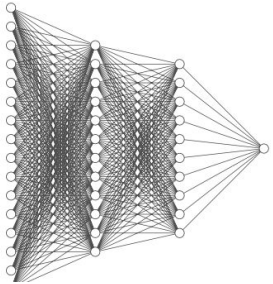
Trigram-based fuzzy search over dataset metadata

“HEP2G” gets corrected to “HEPG2”

More than just keyword search

Dataset metadata

Project Title
Dataset Description
Experimental Design
⋮



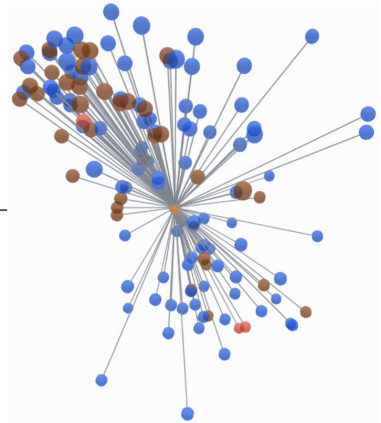
Hugging Face

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \end{bmatrix}$$

4096-dimensional embedding vector

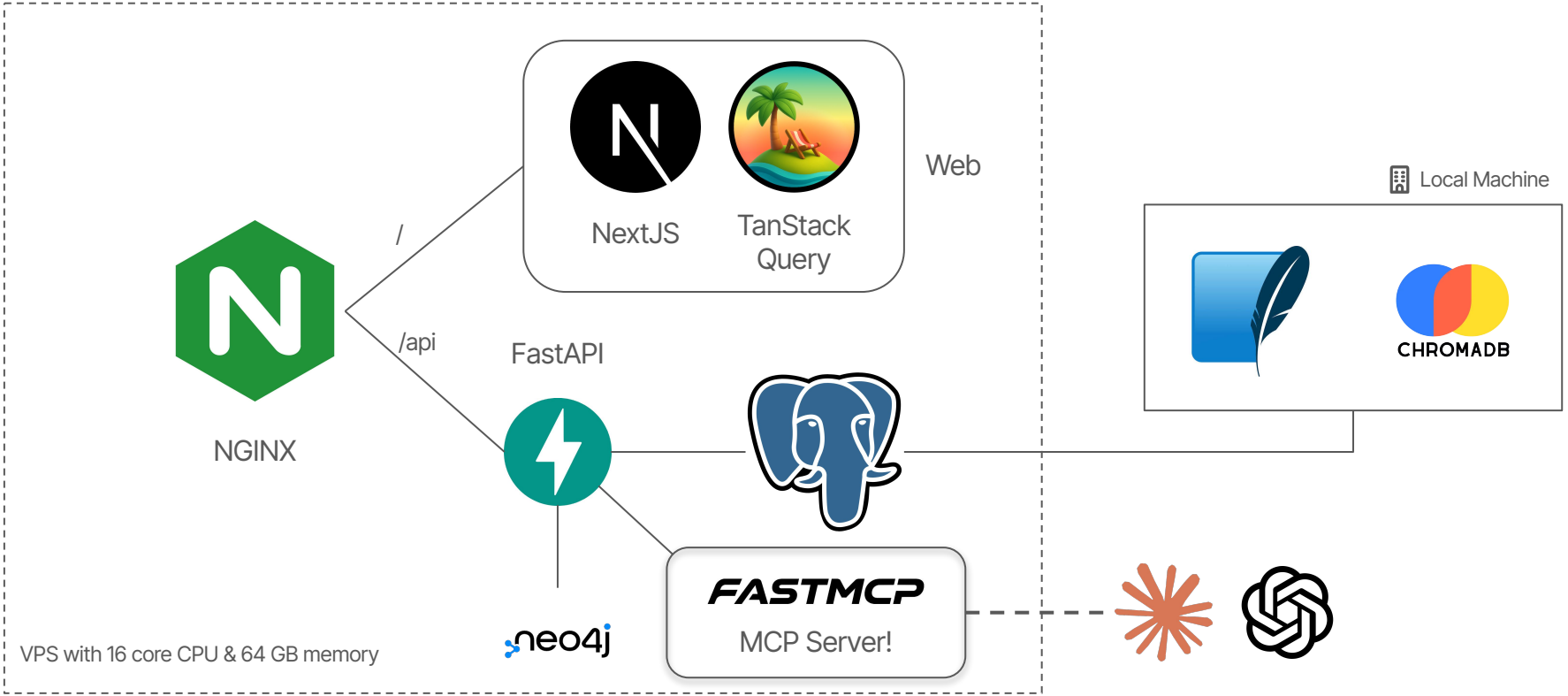
CHROMADB

Project to a 3D space



Store the nearest neighbor projections

Architecture



Nuanced queries in natural language

Find two single cell RNA seq datasets for human glioblastoma

Used seqout.org integration, loaded tools >

Here are two good options, both from primary patient tumors with processed count matrices available on GEO.

1. GSE182109 — Glioma immune cell atlas (whole tumor, 10x Genomics)
"A glioma immune cell atlas identifies prognostic macrophage signatures and a novel immunotherapy target, S100A4" (Abdelfattah et al., *Nature Communications*, 2022). It has 44 samples covering newly diagnosed GBM (ndGBM), recurrent GBM (rGBM), and lower-grade glioma, with several recurrent tumors sampled at multiple regions (for example rGBM-01-A

What we learned

Users throw all kinds of queries at your search bar!

("pancreatic ductal adenocarcinoma"[MeSH Terms] OR PDAC) AND ("normal pancreas" OR "pancreatic benig
("pancreatic ductal adenocarcinoma"[MeSH Terms] OR PDAC) AND ("normal pancreas") AND ("Homo sapiens"
(PDAC) AND ("normal pancreas") AND ("Homo sapiens"[Organism])
137540
42413807
ASM2538384
ASM2538384v1
Arabidopsis thaliana IAA RNA-seq

- Continual process of learning from user behaviour and patterns
- Sometimes adapting to their patterns
- Other times making it easier for users to learn the rules!

Reception so far...



User verdict

End

To put it simply, if you deal with GEO/SRA daily, spending five minutes opening seqout.org and trying a search will likely make you unable to go back to NCBI's tabbed interface. It's open source, and the team has provided standardized referencing methods.

https://mp.weixin.qq.com/s/mzyEwjJEp8agqRt2Gnn9bA

Thank you!