GES-DISC Graph-Enabled Vector Search

AGU FALL Meeting 2022

Armin Mehrabian[1,2], Irina Gerasimov[1,2], Mohammad Khayat[1,2], Brianna Pagán[1,2]
Binita KC[1,2], Mahabal Hegde[1], and David J Meyer[1]

[1]Code 619, NASA Goddard Space Flight Center
[2]ADNET Systems Inc

Who We Are?

- We are the NASA Goddard Earth Sciences Data and Information Services Center (GES-DISC)
- Located at the Goddard Space Flight Center (GSFC) in Greenbelt, Maryland, USA
- One of 12 NASA Science Mission Directorate Data Centers that provide Earth science data, information, and services
- Our number one goal is to serve your Earth science data and information needs

https://disc.gsfc.nasa.gov/

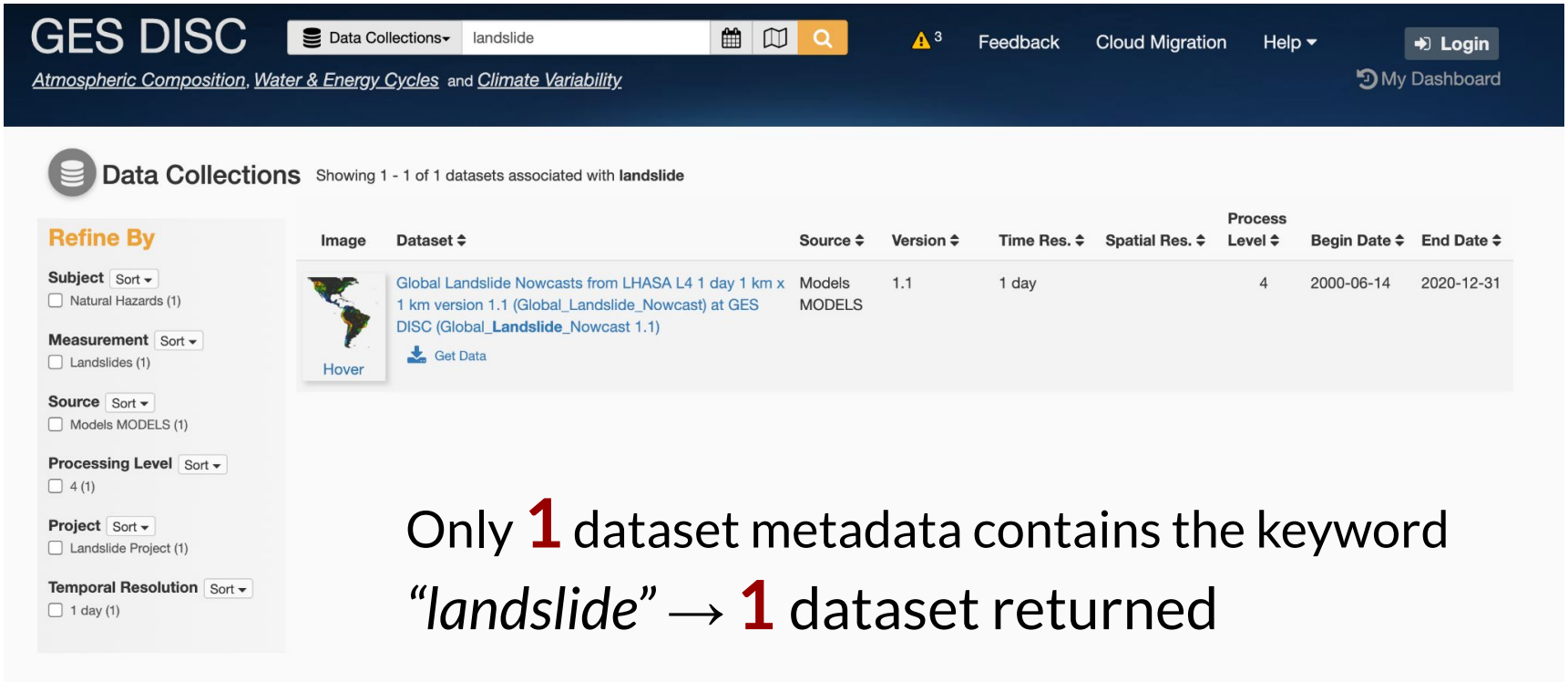
Graph-Enabled

- Connect various data and metadata silos through a ***knowledge graph***
- Infer knowledge from connected data
- Here we focus on connecting ***publications*** that use our data to our dataset metadata

Vector Search

- Use of language models to scan the constructed knowledge graph
- Query the knowledge graph using natural language
- Question answering capability

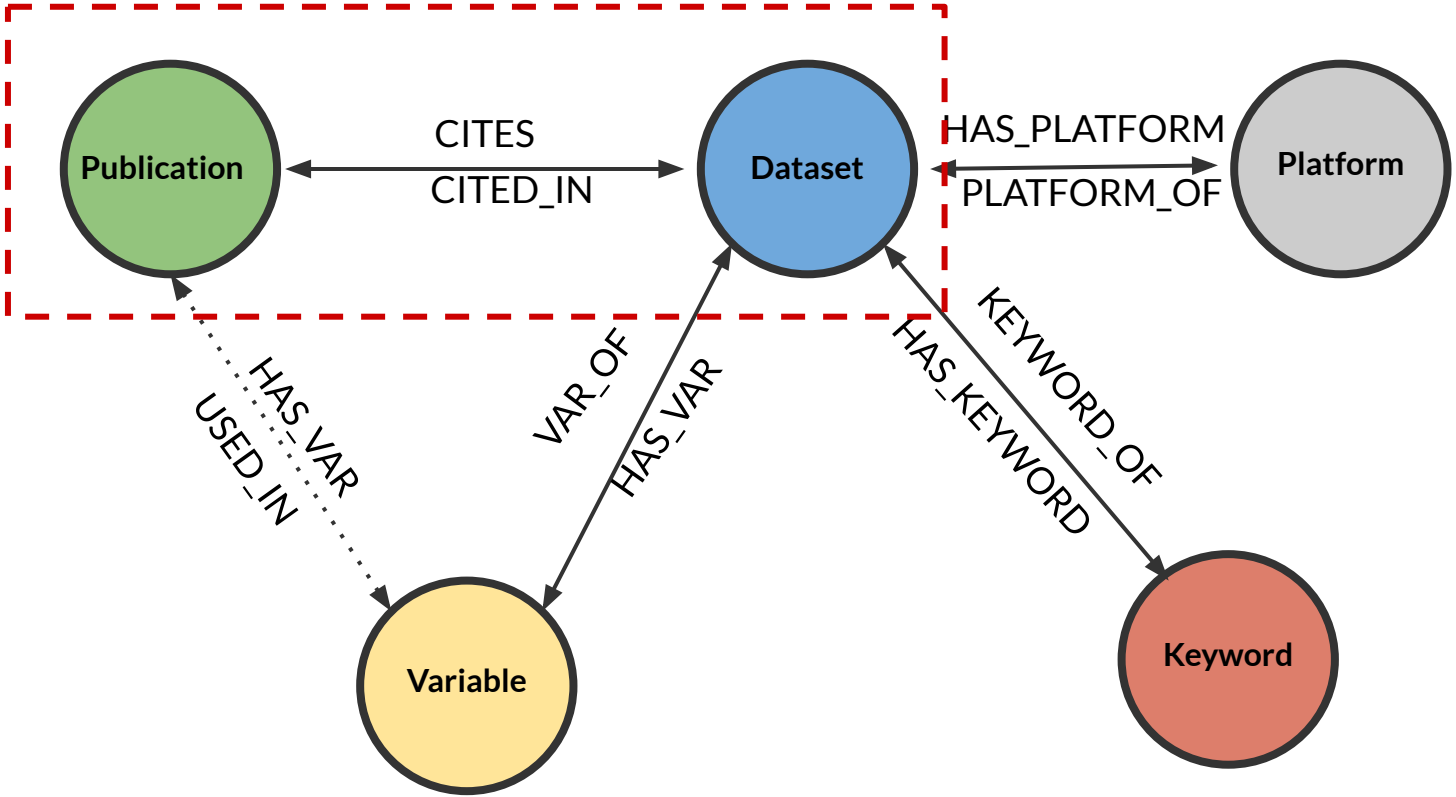
Conventional Full-text Search Engine



Only **1** dataset metadata contains the keyword *"landslide"* → **1** dataset returned

Using Publications for Enhanced Data Discovery

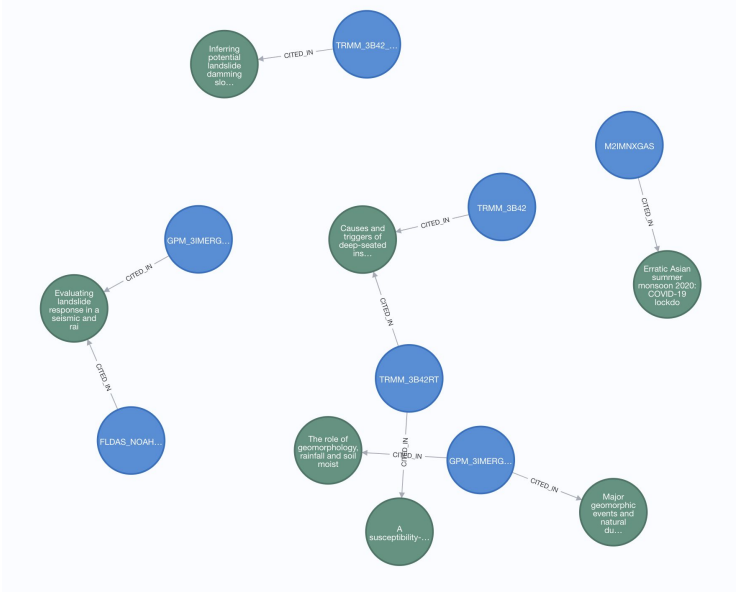
Data Model


Publication
CITES
CITED_IN
Dataset
HAS_PLATFORM
PLATFORM_OF
Platform
HAS_VAR
USED_IN
VAR_OF
HAS_VAR
HAS_KEYWORD
KEYWORD_OF
Variable
Keyword

Search Query “LANDSLIDE”

```
MATCH (d:Dataset)-[i:CITED_IN]->(p:Publication) WHERE p.abstract CONTAINS 'landslide' RETURN d.shortName, p.title
```

	d.shortName	p.title
1	"GPM_3IMERGHH"	"The role of geomorphology, rainfall and soil moisture in the occurrence of landslides triggered by 2018 Typhoon Mangkhut in the Philippines"
2	"GPM_3IMERGHH"	"Major geomorphic events and natural hazards during monsoonal precipitation 2018 in the Kali Gandaki Valley, Nepal Himalaya"
3	"M2IMNXGAS"	"Erratic Asian summer monsoon 2020: COVID-19 lockdown initiatives possible cause for these episodes?"
4	"TRMM_3B42"	"Causes and triggers of deep-seated hillslope instability in the tropics Insights from a 60-year record of Ikoma landslide (DR Congo)"
5	"TRMM_3B42_Daily"	"Inferring potential landslide damming using slope stability, geomorphic constraints, and run-out analysis: a case study from the NW Himalaya"
6	"FLDAS_NOAH01_C_GL_M"	"Evaluating landslide response in a seismic and rainfall regime: a case study from the SE Carpathians, Romania"
7	"GPM_3IMERGDF"	"Evaluating landslide response in a seismic and rainfall regime: a case study from the SE Carpathians, Romania"
8	"TRMM_3B42RT"	"Causes and triggers of deep-seated hillslope instability in the tropics Insights from a 60-year record of Ikoma landslide (DR Congo)"
9	"TRMM_3B42RT"	"A susceptibility-based rainfall threshold approach for landslide occurrence"



Search Query “ALGAL BLOOM”

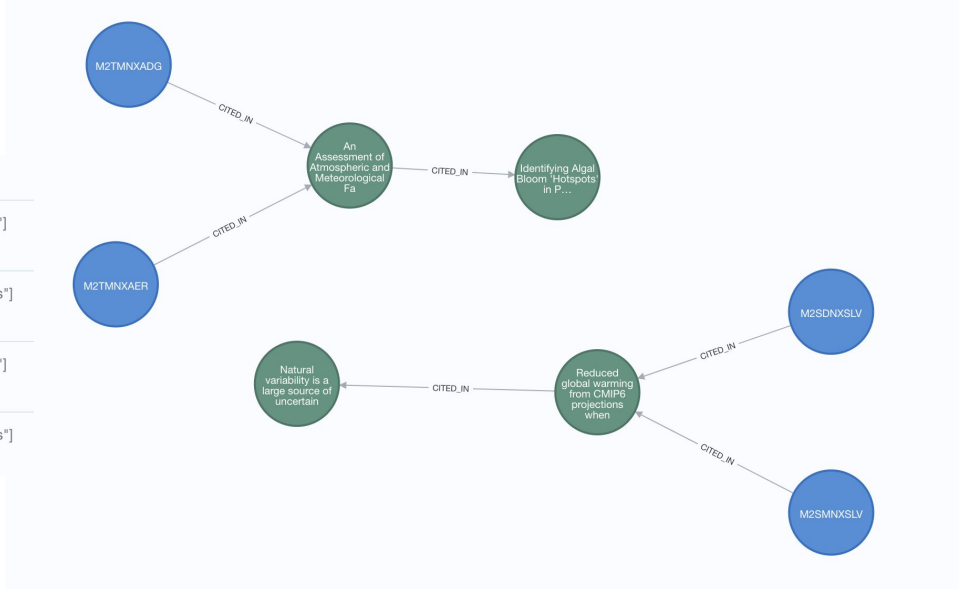
```
MATCH (d:Dataset)-[i:CITED_IN]->(p:Publication) WHERE p.abstract CONTAINS 'algal bloom' RETURN
d.shortName, p.title
```

0 Publications → 0 Datasets

Search Query “ALGAL BLOOM”

```
MATCH (d:Dataset)-[i:CITED_IN*1..2]->(p:Publication) WHERE p.abstract
CONTAINS 'algal bloom' RETURN d.shortName, p.title
```

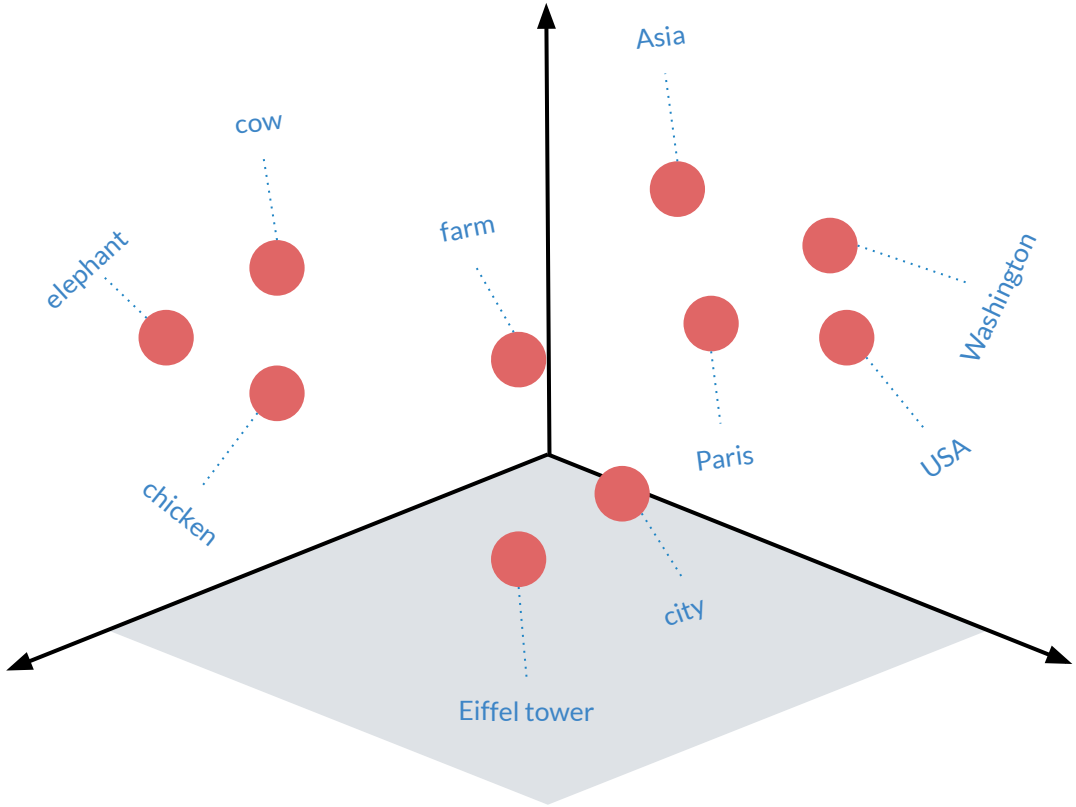
	d.shortName	p.title
1	"M2SDNXSLV"	["Natural variability is a large source of uncertainty in future projections of hypoxia in the Baltic Sea"]
2	"M2TMNXAER"	["Identifying Algal Bloom 'Hotspots' in Marginal Productive Seas: A Review and Geospatial Analysis"]
3	"M2SMNXSLV"	["Natural variability is a large source of uncertainty in future projections of hypoxia in the Baltic Sea"]
4	"M2TMNXADG"	["Identifying Algal Bloom 'Hotspots' in Marginal Productive Seas: A Review and Geospatial Analysis"]



Graph-Enabled Vector Search Engines

Combining NLP and Graph Capabilities

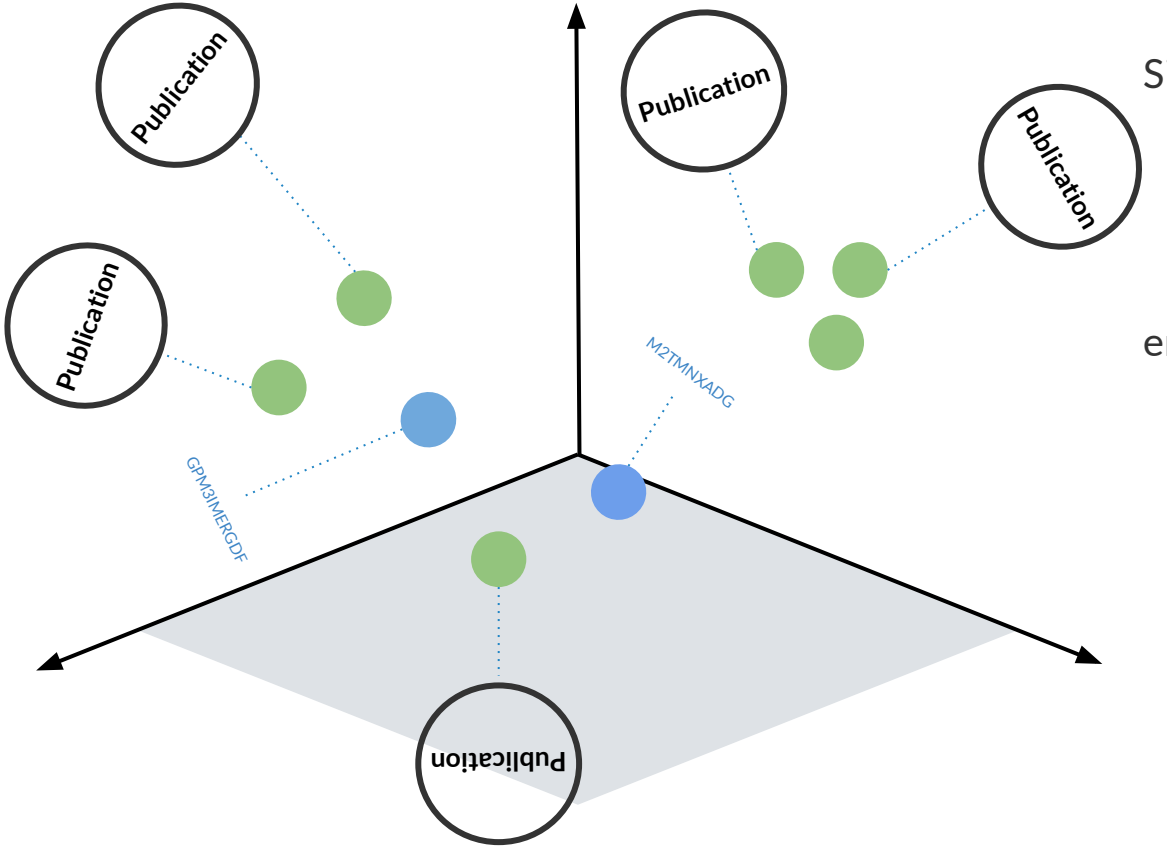
Language embeddings



We can represent every word, sentence, phrase, … document with a meaningful vector

elephant = [0.31, 0.62, … , 0.87]
cheetah = [0.94, 0.54, …, 0.88]
Eiffel tower = [0.45, 0.67, …, 0.87]
France = [0.56, 0.83, …, 0.22]
farm = [0.32, 0.68, … , 0.76]

Language embeddings



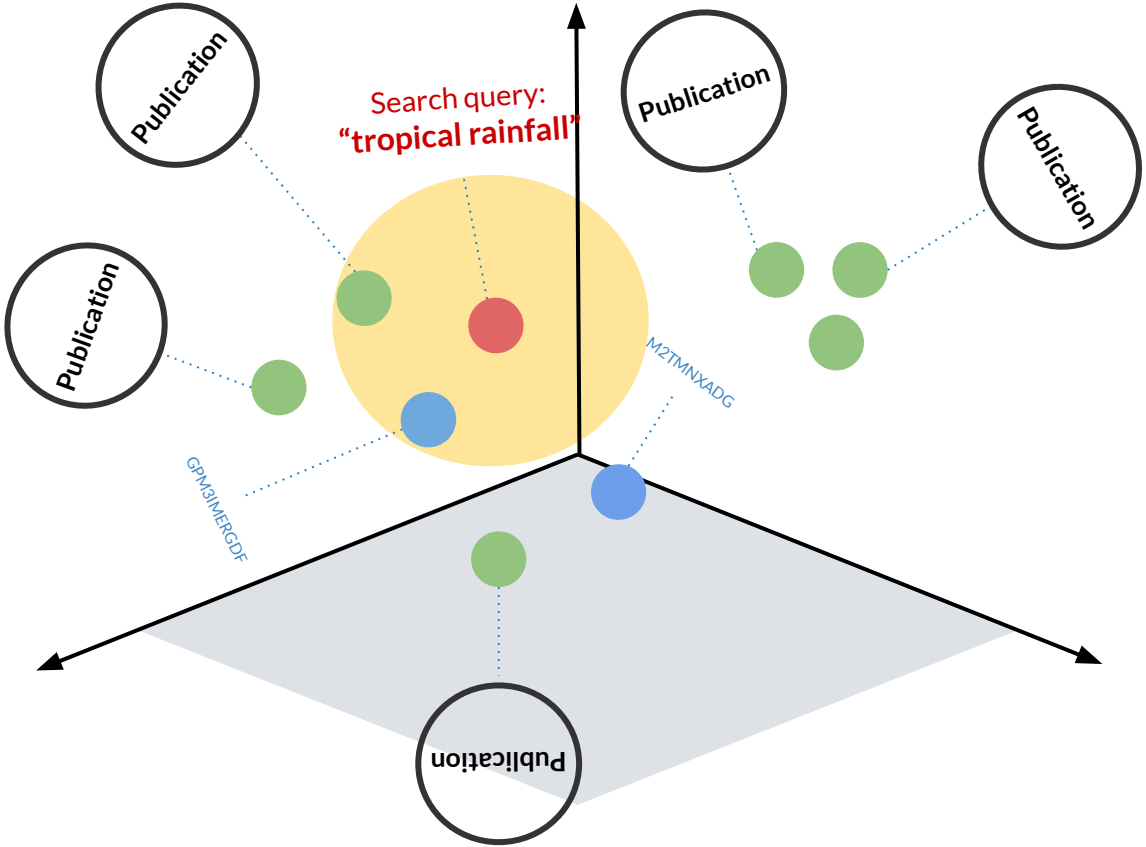
Similarly we can map

- Dataset metadata
- Publications
- User search queries

onto this space using language embeddings

- Contextionary-based (Fasttext, w2vec, …)
- Transformer-based (BERT, xlm, GPTs, …)

Language embeddings

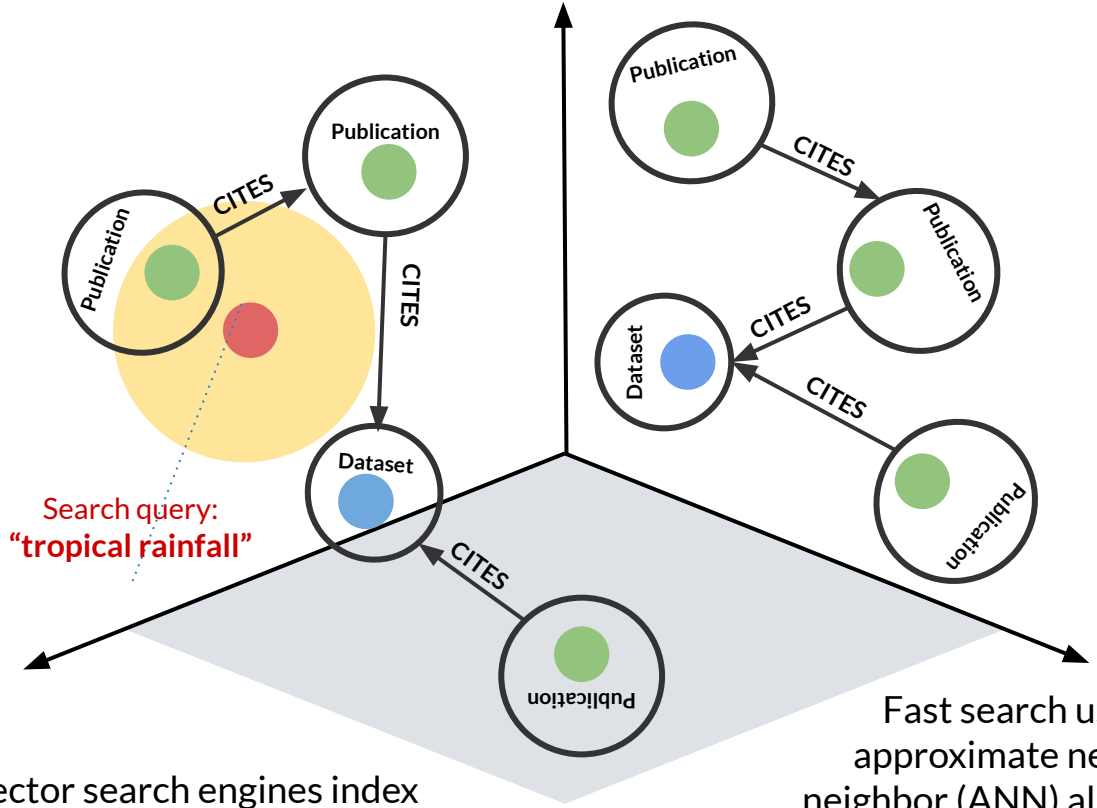


For a given user query i.e.

"tropical rainfall"

1. The query is mapped in the vector space
2. All objects including publications and dataset metadata within a radius of the query are identified.

Graph-enabled Vector Search



Vector search engines index data as *dense vectors*

Fast search using approximate nearest neighbor (ANN) algorithms such as *HSNW (<100ms)*

Graph-enables vector search tools such as [Weaviate] create a graph layer on top of the vector search.

For a given user query i.e.

"tropical rainfall"

1. The query is placed in the vector space
2. All objects including publications and datasets within a radius of the query are identified.
3. If identified object is of type "publication", we can traverse graph to find the closest dataset (minimum hops)

Query: rainfall and cloud type relationship

The exact query string does NOT exist in any publication, dataset metadata, etc. However, using vector search,

```
{
  Explore (
    limit: 2,
    nearText: {
      concepts: ["rainfall and cloud type relationship"],
      certainty: 0.7
    },
  ) {
    beacon
    certainty
    className

  }
}
```


```
{
  "data": {
    "Explore": [
      {
        "beacon": "weaviate://localhost/167910bc-3135-55e5-801f-ab77e8473ee1",
        "certainty": 0.8890923,
        "className": "Publication"
      },
      {
        "beacon": "weaviate://localhost/0a65b2d9-74d7-5fa2-a58f-8ac291d7b9fa",
        "certainty": 0.8522953,
        "className": "Publication"
      }
    ]
  }
}
```

Query: rainfall and cloud type relationship

Example returned publications

Abstract 1

Three years of reanalysis and ground-based observations collected at the Eastern North Atlantic (ENA) observatory are **analyzed to document the properties of rain and boundary layer clouds** and their relationship with the large-scale environment during general subsidence conditions and following cold front passages. Clouds in the wake of cold fronts exhibit on average a 10% higher propensity to precipitate and higher rain-to-cloud fraction than cloud found in general subsidence conditions. Similarities in the seasonal cycle of rain and of large-scale properties suggest that the large-scale conditions created by the cold front passage are responsible for the unique properties of the rain forming in its wake. **The identification of monotonic relationships between rain-to-cloud fraction and rain rate with surface forcing and boundary layer stability parameters as well as between virga base height with stability and humidity measures further supports that large-scale conditions impact precipitation variability.** That being said, these relationships between the large-scale and rain properties are less clear than **those established between cloud and rain properties, suggesting that cloud macrophysics have a more direct impact on the properties of rain than the large-scale environment.** The applicability of previously documented **relationships between cloud thickness and rain properties is tested** and the relationships adjusted to accommodate the complex shallow clouds and melting precipitation observed to occur in the ENA region. Establishing these relationships opens up opportunities for parametrization development and suggests that a realistic representation of precipitation properties in models relies on the accurate representation of both clouds and the large-scale environment.

Abstract 2

Four years of CloudSat cloud and precipitation observations are combined with CALIPSO lidar, Moderate Resolution Imaging Spectroradiometer (MODIS) radiance, and Global Precipitation Measurement (GPM) precipitation data to document the cloud properties of precipitation confined to latitudes between 30°N and 30°S. **The relations between two different cloud top heights (CTHs) and precipitation are examined.** The maximum CTH observed in the column is one measure (referred to as the highest CTH, HCTH) and the second is the minimum CTH within the same raining column, interpreted to be the tops of the rain-bearing clouds in the column (referred to as the raining cloud top height, RCTH). Although a broad relation between rain intensity and CTH is shown to exist, especially for shallower warm clouds, the HCTH of the deepest, raining clouds in the tropics is shown to be a poor indicator of precipitation intensity. T**he implication of the difference between HCTH and RCTH is that for all but the deepest convection, the height of raining clouds is significantly overestimated from observing systems that cannot see below upper cloud layers. The vertical profile of CTHs is shown to be distinctly bimodal with RCTH profiles having a large maximum associated with shallow precipitating clouds, whereas the HCTH distribution has its maximum in the upper troposphere. The influence of this vertical profile information on radiative and latent heating profiles results in a nonnegligible shift in latent heating from an upper level maximum to a more bimodal profile reflecting the increased contribution of shallow raining clouds.**

Question Answering

Query: what is the correlation between human population and wildfire?

Returns the starting and ending position of text from the documents that contain the answers

```
{
  Get {
    Publication(
      ask: {
        question: "what is the correlation between human population and wildfire?",
        properties: ["abstract"],
        rerank: true # supported from v1.10.0 on
      },
      limit: 5
    ) {
      title
      _additional {
        answer {
          hasAnswer
          certainty
          property
          result
          startPosition
          endPosition
        }
      }
    }
  }
}
```


```
"data": {
  "Get": {
    "Publication": [
      {
        "_additional": {
          "answer": {
            "certainty": 0.4658797264099121,
            "endPosition": 1973,
            "hasAnswer": true,
            "property": "abstract",
            "result": "positively with burnt area only in densely forested regions",
            "startPosition": 1914
          }
        },
        "title": "Understanding and modelling wildfire regimes: an ecological perspective"
      },
```

Query: In what ways TROPOMI and OMI products differ?

1. Answer: *"spatial and temporal scales"*
   Certainty: **0.65**
   Pub Title: *"Comparative assessment of TROPOMI and OMI formaldehyde observations and validation against MAX-DOAS network column measurements"*

2. Answer: *"intercompared at regional daily and monthly temporal scales, as well as globally at monthly and seasonal scales"*
   Certainty: **0.52**
   Pub Title: *"TROPOMI aerosol products: evaluation and observations of synoptic-scale carbonaceous aerosol plumes during 2018-2020"*

3. Answer: *"tropomi shows a superior performance compared with omi - qa4ecv and operates as anticipated from instrument specifications"*
   Certainty: **0.30**
   Pub Title: *"S5P TROPOMI NO2 slant column retrieval: method, stability, uncertainties and comparisons with OMI"*

Search Queries Matrix

Search Query	Doc Search (Elasticsearch)	Graph Search (Neo4j)	Vector Graph Search (Weaviate)
Precipitation	✓	✓	✓
Wildfire	✓ (1 dataset)	✓	✓
Air pollution	✗	✓	✓
Algal Bloom	✗	✓ (multi-hop)	✓
Rainfall and cloud type relationship	✗	✗	✓
what is the correlation between human population and wildfire?	✗	✗	✓

THANK YOU

Please feel free to contact me if you have any questions.

armin.mehrabian@nasa.gov