NASA GES-DISC Knowledge Base: Connecting Science Variables, Measurement, Datasets, and Publications

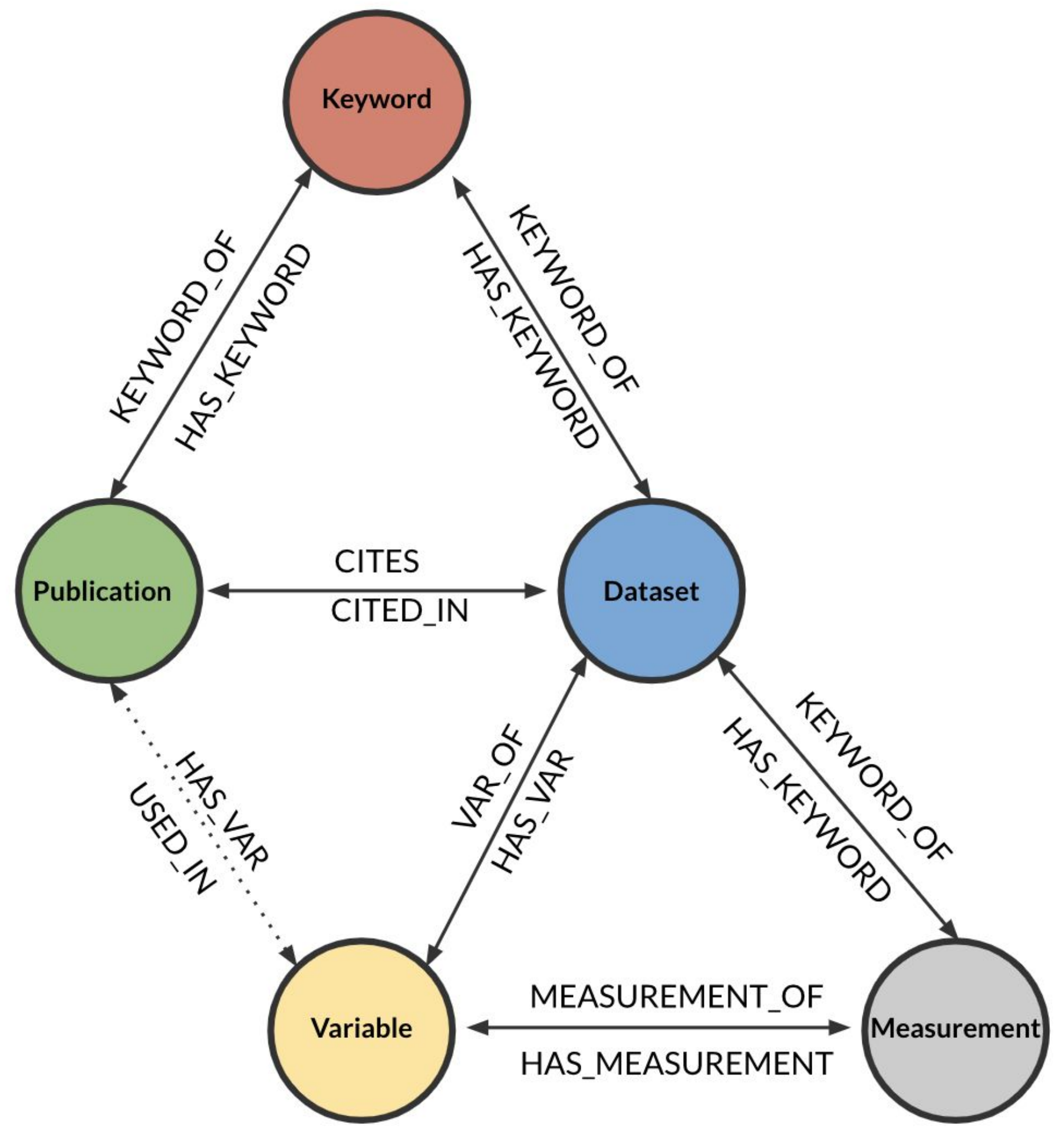
Armin Mehrabian[1,2], Irina Gerasimov[1,2], Mohammad G Khayat[1,2], Brianna Rita Pagán[1,2], Binita KC[1,2], Mahabal Hegde[1], and David J Meyer[1] (1) NASA GES-DISC Code (619) (2) Adnet system Inc.

ABSTRACT

The Goddard Earth Sciences Data and Information Services Center (GES-DISC) is a NASA-run data archive center that focuses on topics such as atmospheric composition, water and energy cycles, and climate variability. The center adheres to the FAIR data model, which stresses the importance of finding data. Users can access the center's datasets through a search engine that allows them to locate relevant information. Conventional data discovery methods, which often involve manual curation of metadata, can be limited in their capacity to scale as datasets grow and may not account for linguistic variations. Recent advancements in natural language processing and knowledge graphs have led to the development of NASA GES-DISC vector search, which incorporates these technologies to improve search results by integrating the knowledge of the research community. The citation network is used as supplementary metadata, and knowledge graphs are used to enhance the explanations provided by the search engine. Here we implement our graph on a hybrid of Neo4j and Weaviate vector search. Graph analytics can be performed using Neo4j as a graph-native database. The use of Weaviate as a vector search allows us to perform high-level sparse queries as well as abstract features such as natural language question answering.

GRAPH-ENABLED VECTOR DB

- 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

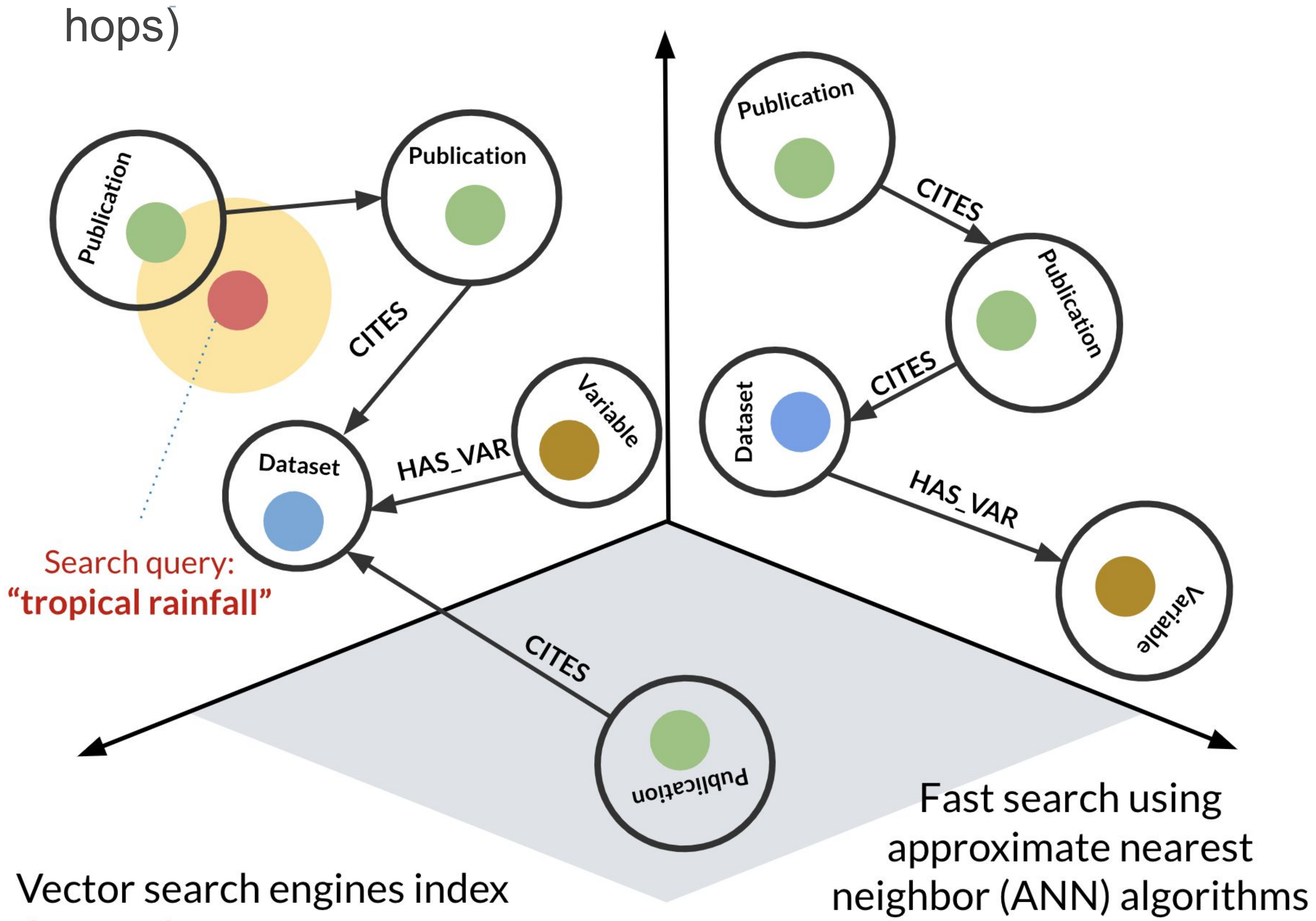


METHODOLOGY

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)

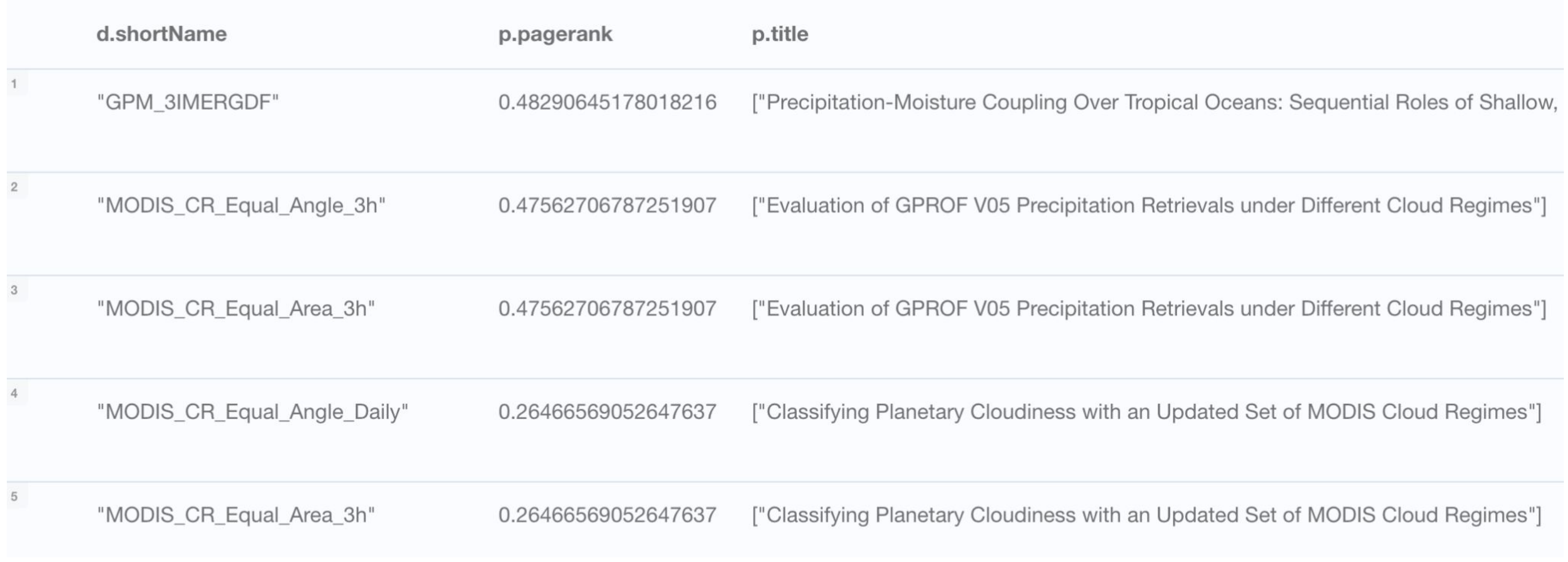


Vector search engines index data as *dense vectors*

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

SEARCH USING GRAPH ANALYTICS
(i.e. Pagerank centrality)

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



	d.shortName	p.pagerank	p.title
1	"GPM_3IMERGDF"	0.48290645178018216	["Precipitation-Moisture Coupling Over Tropical Oceans: Sequential Roles of Shallow,
2	"MODIS_CR_Equal_Angle_3h"	0.47562706787251907	["Evaluation of GPROF V05 Precipitation Retrievals under Different Cloud Regimes"]
3	"MODIS_CR_Equal_Area_3h"	0.47562706787251907	["Evaluation of GPROF V05 Precipitation Retrievals under Different Cloud Regimes"]
4	"MODIS_CR_Equal_Angle_Daily"	0.26466569052647637	["Classifying Planetary Cloudiness with an Updated Set of MODIS Cloud Regimes"]
5	"MODIS_CR_Equal_Area_3h"	0.26466569052647637	["Classifying Planetary Cloudiness with an Updated Set of MODIS Cloud Regimes"]

HIGH-LEVEL NATURAL LANGUAGE FEATURES

1. Complex queries

Example Query: "rainfall and cloud type relationship"

Top result from a publication abstract:

> 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

2. Question answering

Example Query: "How do TROPOMI and OMI products differ?"

Top result from publication abstracts:

- Answer: *"spatial and temporal scales"*
  Certainty: 0.65
- Answer: *"tropomi shows a superior performance compared with omi -qa4ecv and operates as anticipated from instrument specifications"*
  Certainty: 0.30

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?	✗	✗	✓
Graph analytic capabilities	✗	✓	✗