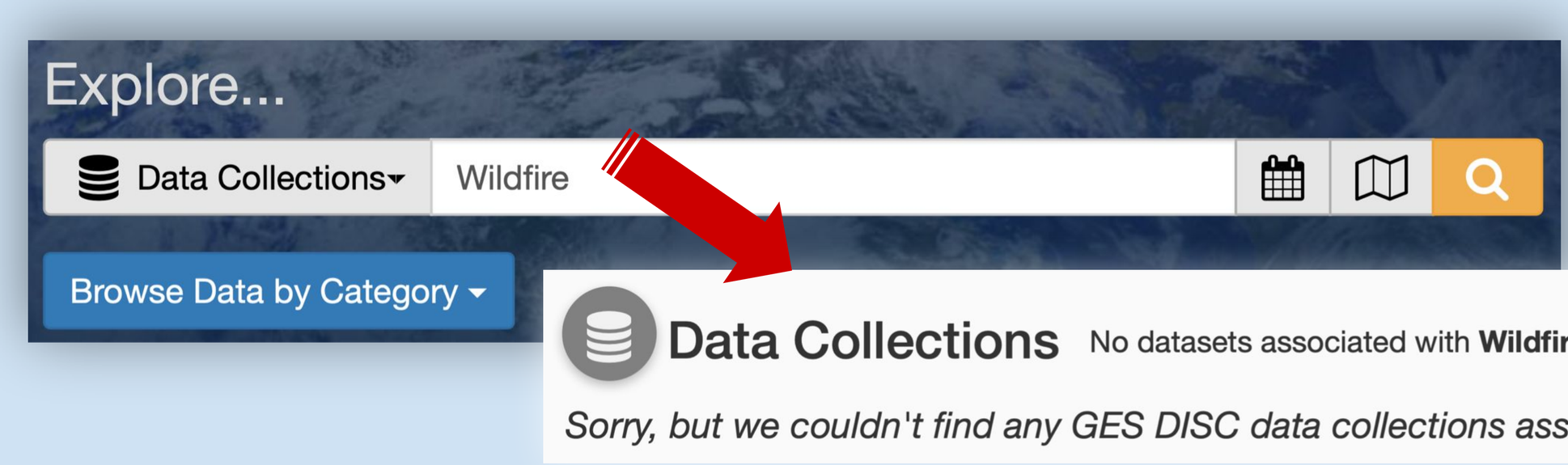
Search Enhancements using NLP Techniques

Armin Mehrabian[1,2], Irina Gerasimov[1,2], Mohammad Khayyat[1,2], Jennifer Wei[1], Binita KC[1,2], and Hegde Mahabaleshwara[1]
1 NASA Goddard Space Flight Center
2 ADNET Systems INC.
Corresponding Author: Armin Mehrabian (armin.mehrabian@nasa.gov)

GES DISC
SUMMER 2021

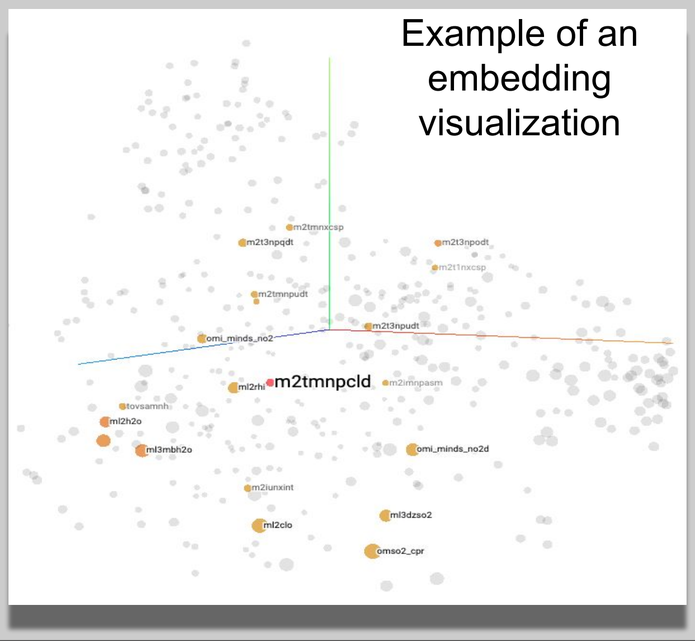
PAIN POINT

- Traditional search engines rely on hard-matching of query string and some form of document metadata
- The above approach fails to find semantic similarities between the search query and metadata that are not exact matches



EMBEDDINGS

- In order to perform mathematical operations in NLP one needs to convert language components such as words, sentences, … into numerical values, namely **vectors** or **embeddings**
- These vectors are high-dimensional (100~300 dimensions)
- Embeddings capture semantics and context. I.E. words "**precipitation**" and "**rainfall**" should have a shorter distance to one another relative to a word such as "**wildfire**" in an embedding space
- **Here** we train a joint embedding from a corpus of relevant scientific text along with a curated text corpus from **GES DISC** collections metadata



TRAINING

270K AGU Abstracts

GES DISC collections metadata

NLP Preprocessing

Flattened and curated **collection metadata corpus**

Train Joint **Embedding** (FastText, BERT, …)

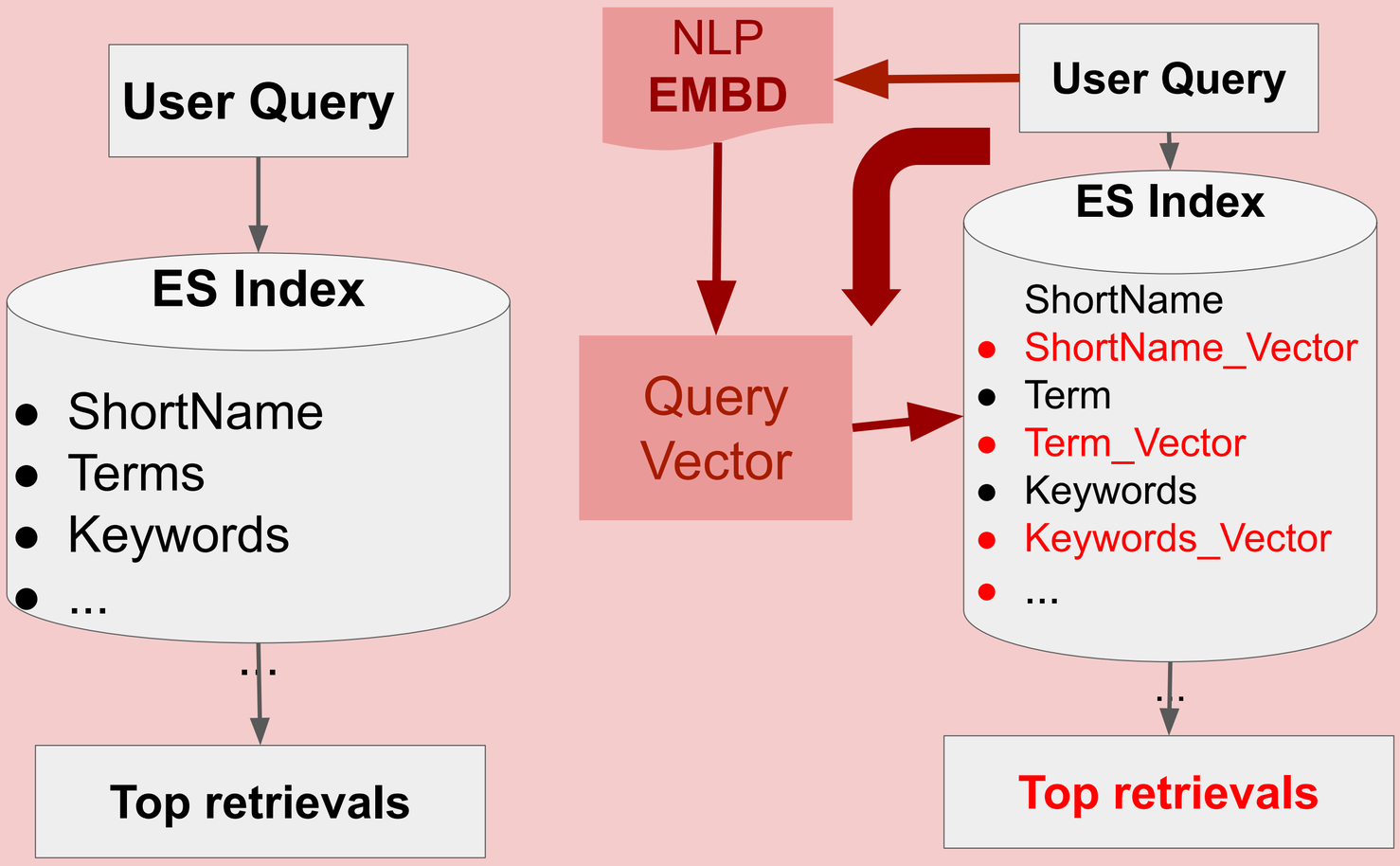
NLP **EMBD**

Wildfire = [0.26, 0.11, 0.54, … 0.32]

S5P_L2__SO2 = [0.22, 0.87, 0.43, … 0.31]

Every word receives an embedding vector including collections metadata components such as dataset short names.

ELASTICSEARCH



- Elasticsearch offers "**dense_vector**" field types to store embedding vectors
- For a curated set of fields in the collections metadata we ingest the fields accompanied by their embedding vectors
- During "**query time**", the user query string is converted into its vector representation
- A similarly score is then calculated based on the query vector and the vectors of ingested fields
- This allow us to have a search return based on similarity score even in the absence of exact matches