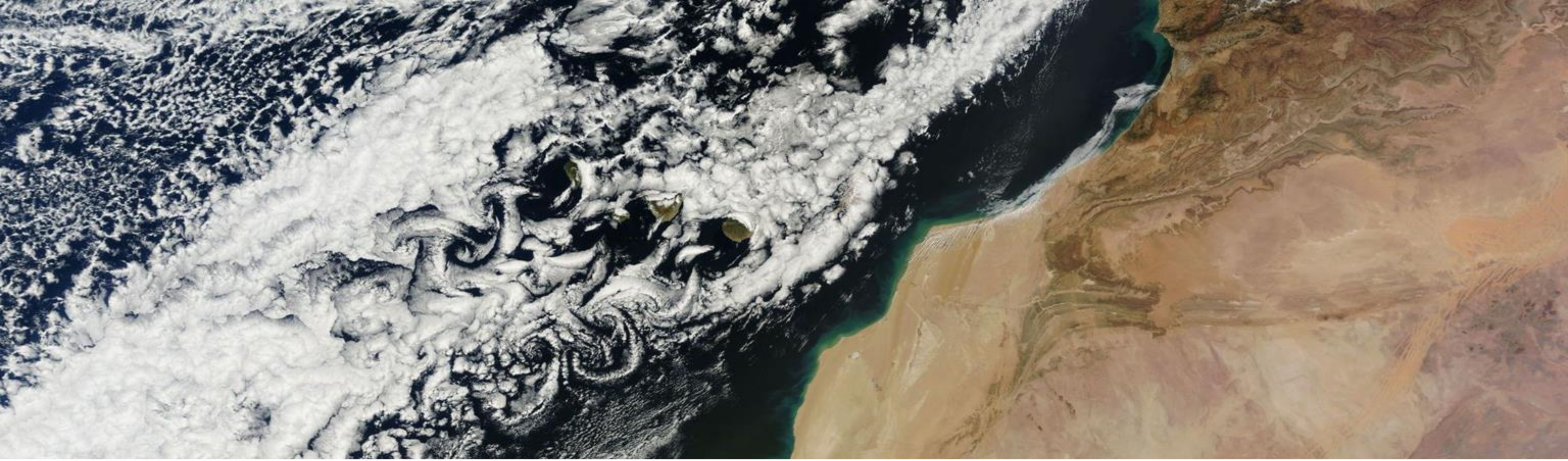


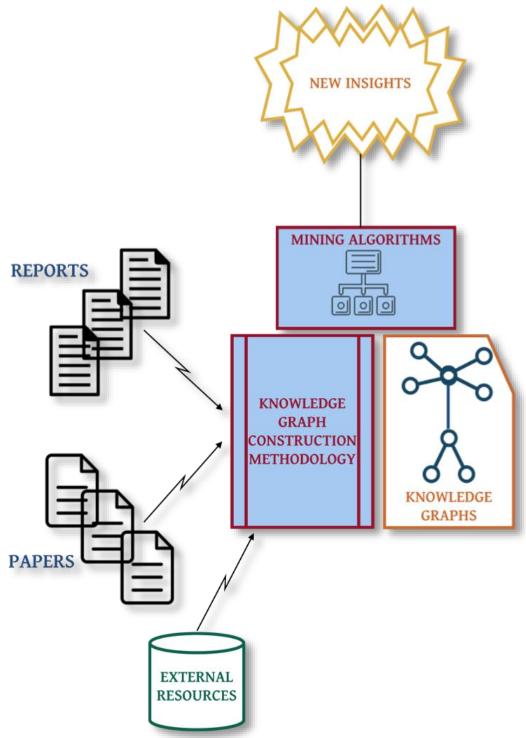
# Building Scalable Knowledge Graphs for Earth Science

Rahul Ramachandran, Manil Maskey Patrick Gatlin (NASA MSFC)
Jia Zhang, Xiaoyi Duan (CMU)
J.J. Miller, Kaylin Bugbee, Sundar Christopher (UAH)

# What is a Knowledge Graph?

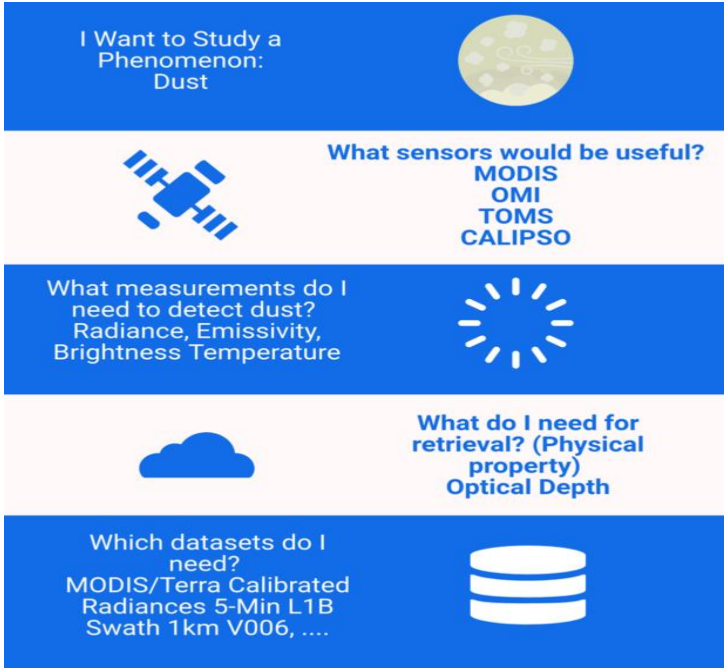
- Knowledge Graphs link key entities in a specific domain with other entities via relationships.
- Researchers can then query these graphs to get probabilistic recommendations and to infer new knowledge.

Can we develop an end-to-end (semi) automated methodology for constructing Knowledge Graphs for Earth Science?



# Why Research Community Needs Knowledge Graphs?

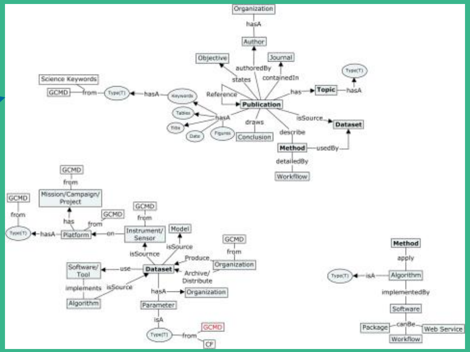
- Untapped resource of knowledge for a given domain is stored in papers and technical reports (unstructured).
    - Difficult to extract and infer knowledge at scale



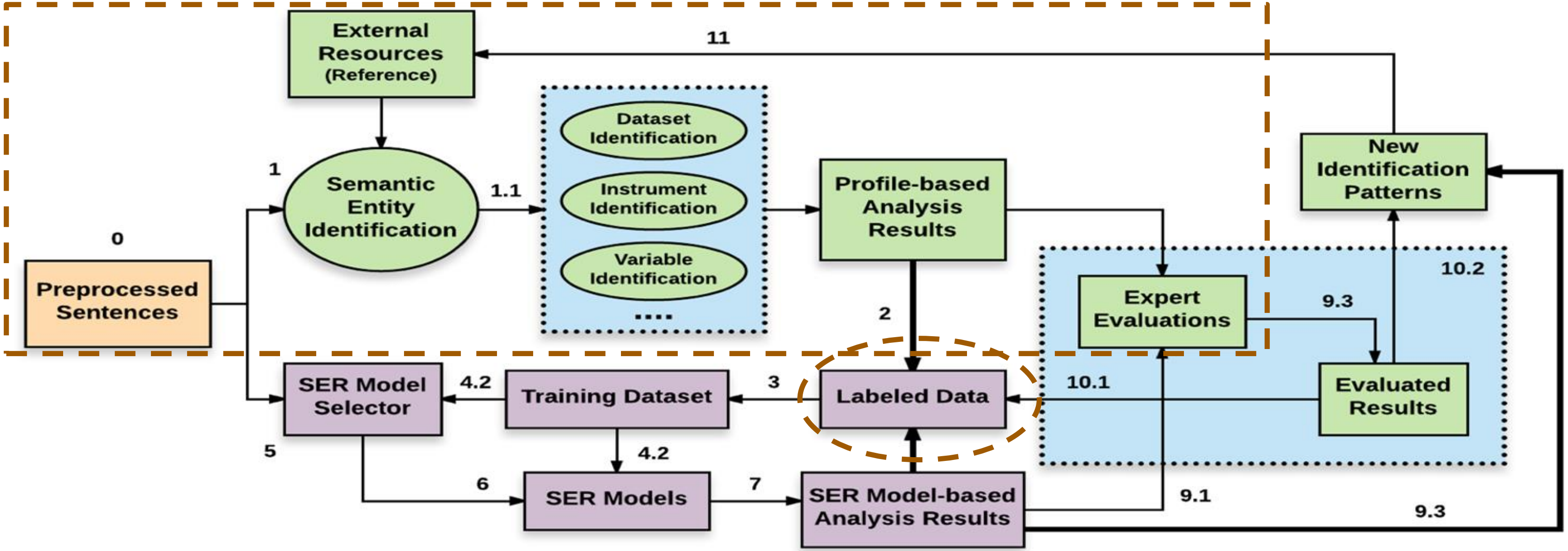


Remote Sensing of Environment

Improved algorithm for MODIS satellite retrievals of aerosol optical thickness over land in dusty atmosphere: Implications for air quality monitoring in China



# Methodology to Build Knowledge Graphs

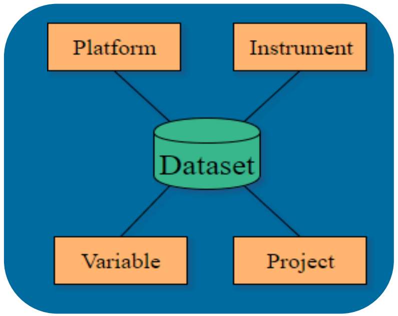



- Consists of two stages
  - Development of Heuristic algorithms to perform Semantic Entity Identification (Phenomena, Dataset, Instrument, Variable (Physical Property)...) to assist human experts in building training data [Steps 0-2] [**Focus of this Poster**]
  - Use Deep Learning Algorithms to improve results [Steps 3-7]

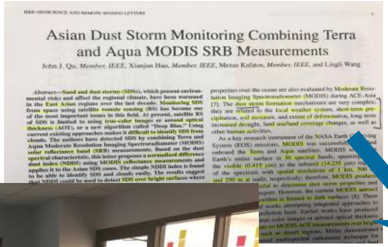
# Heuristic Algorithm Development Strategy

- Goal:
  1. Develop a set of algorithms to extract different semantic entities to build a training dataset
     - Phenomena, Property (Variable), Process, Projects, Instruments, Places
  2. Develop "profiles" to march relevant datasets to papers
- Explore the use of existing taxonomies (GCMD, CF, SWEET)

Dataset Profile

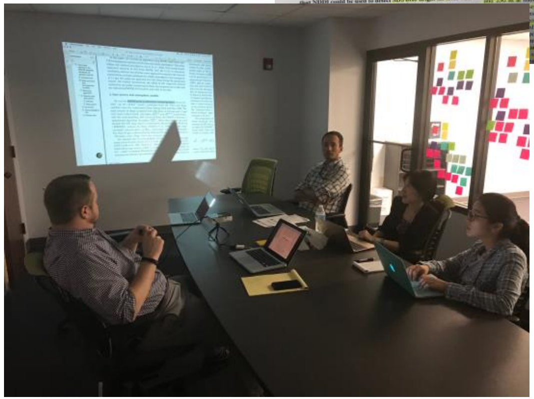


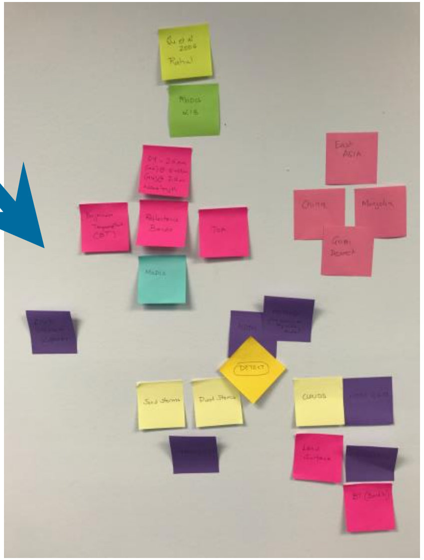

- Use curated set papers as a benchmark for a specific topic – "Airborne Dust Retrieval from Satellites"
- Experts manually extract key entities from these papers


IEEE GEOSCIENCE AND REMOTE SENSING LETTERS

Asian Dust Storm Monitoring Combining Terra and Aqua MODIS SRB Measurements

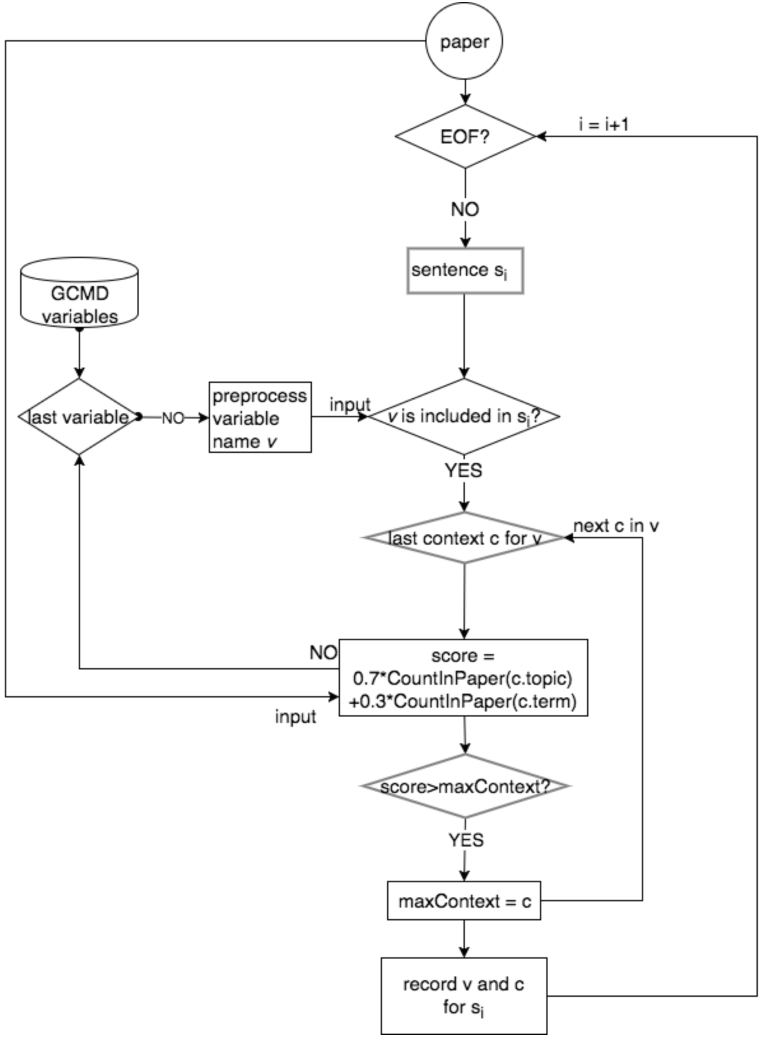
John J. Qu, Member, IEEE, Xianjun Hao, Member, IEEE, Menas Kafatos, Member, IEEE, and Lingli Wang





# GCMD Variable Extraction Algorithm

- Match variable name
- Some variables appear in the collection multiple times
    - Find the most related context:
        - 0.7*topic_count + 0.3*term_count



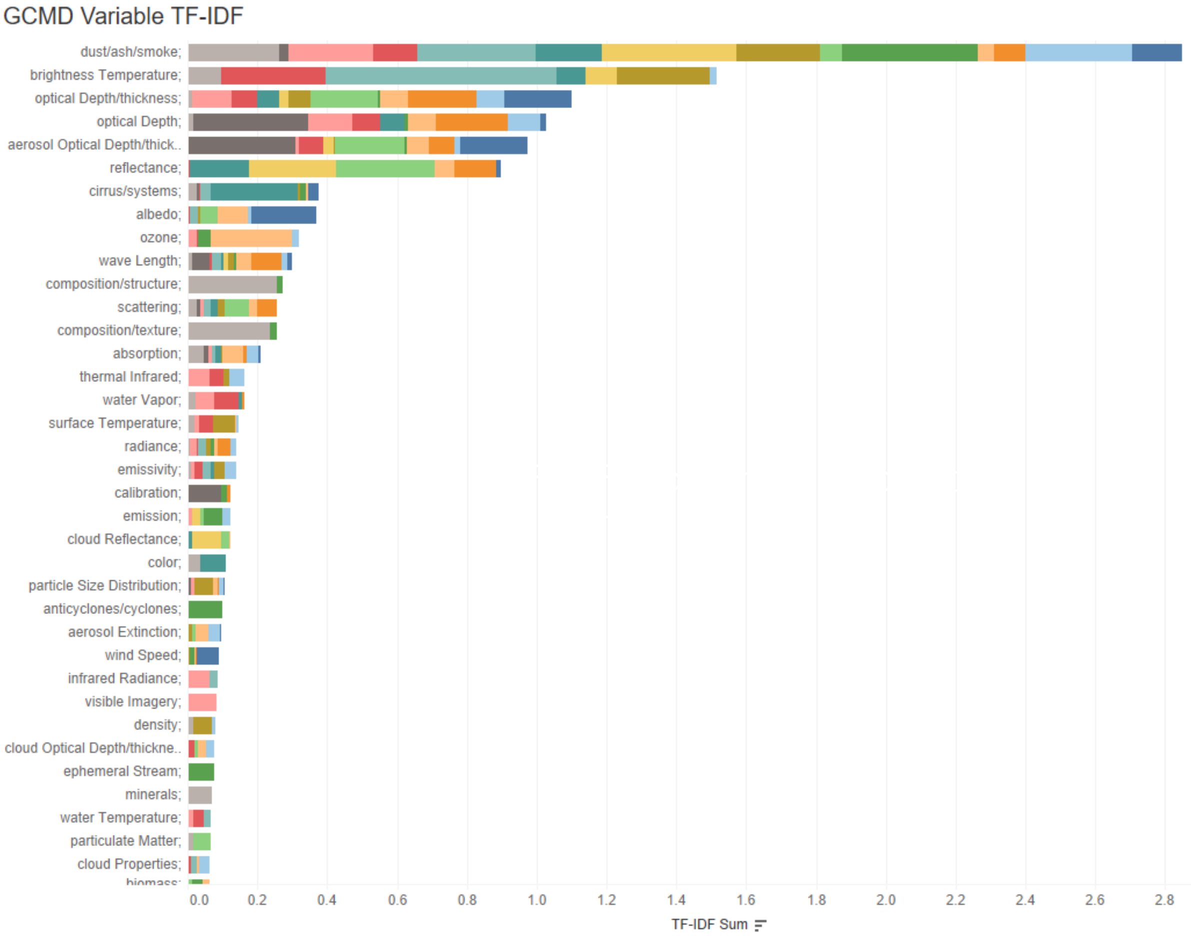
# Extraction Results

**Good:**

- TF/IDF better than total counts
- Brightness temp is ranked higher than in the total counts result
- Uncovered errors in paper: "Dust has a higher albedo at 12 microns instead of 11"
    - Should be temperature, not albedo

**Bad:**

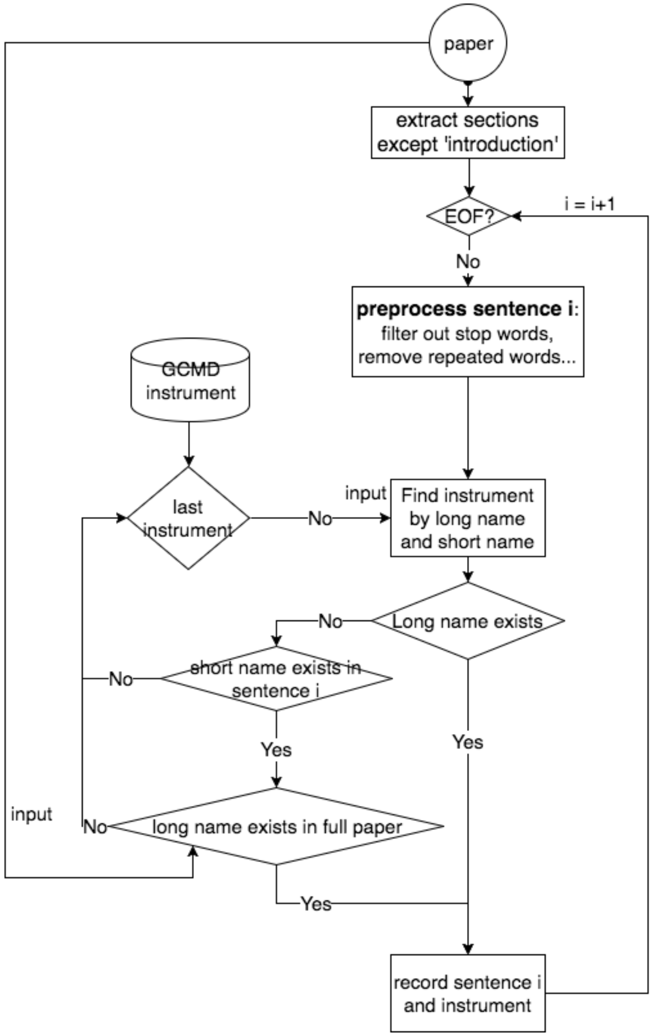
- GCMD does not differentiate between entity types: physical property, phenomena etc
- Emissivity and radiance are important properties but are ranked low
- Dust/ash/smoke gives big picture but not really useful for analysis



# Instrument and Project Extraction Algorithm

Entity name: short name (S), long name (L)

- Instrument and project extraction from each sentence:
  - if find L, record;
  - if find S, check L in full text



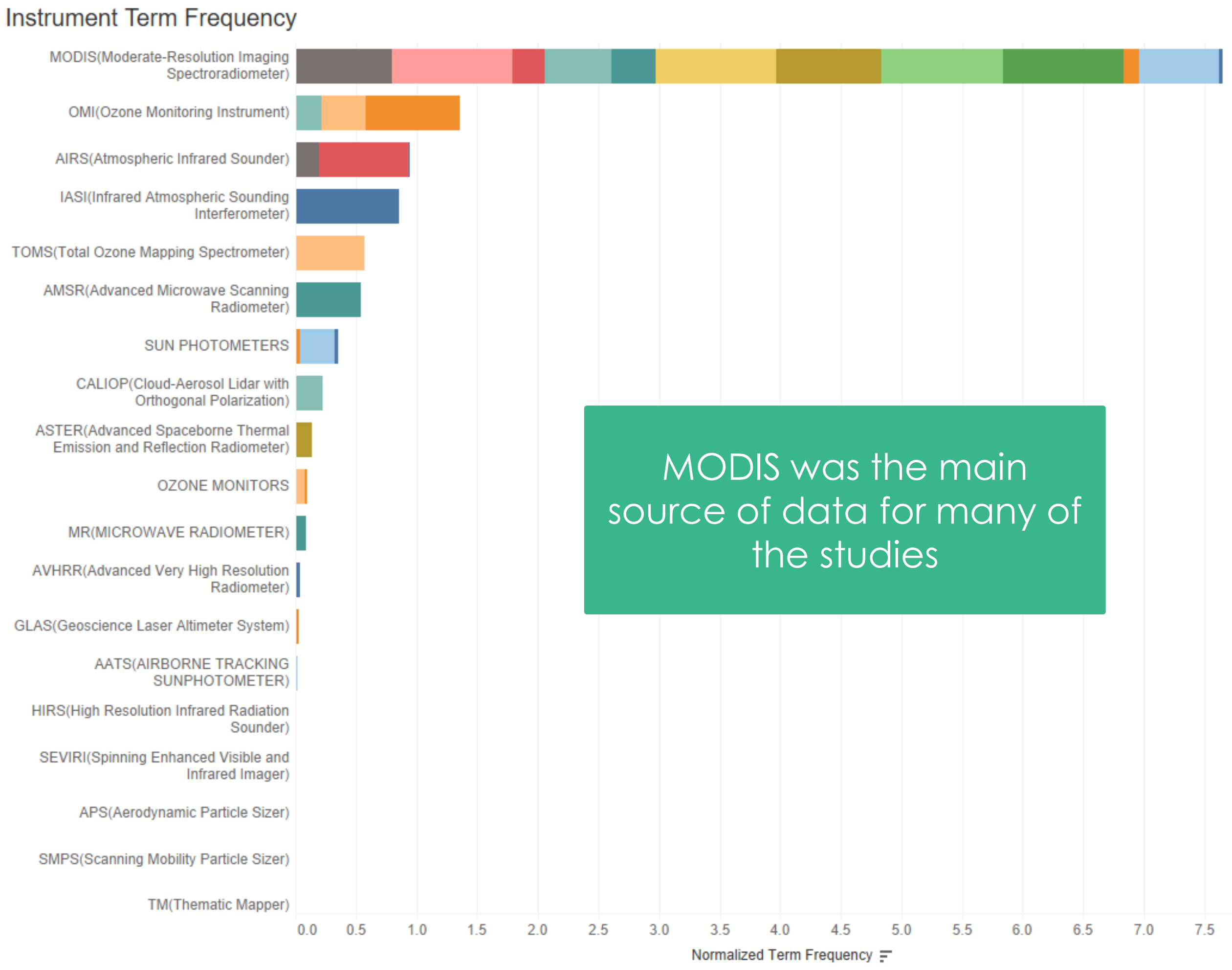
# Extraction Results

**<u>Good:</u>**

- MODIS/OMI are top instruments for dust

**<u>Bad:</u>**

- MR (Microwave radiometer from GCMD) incorrectly matched with AMSR related sentence
  - "The AMSR-E is a conical scanning total power passive **<u>microwave radiometer</u>** sensing (brightness temperatures) at 6 frequencies ranging from 6.9 to 89.0 GHz."




MODIS was the main source of data for many of the studies

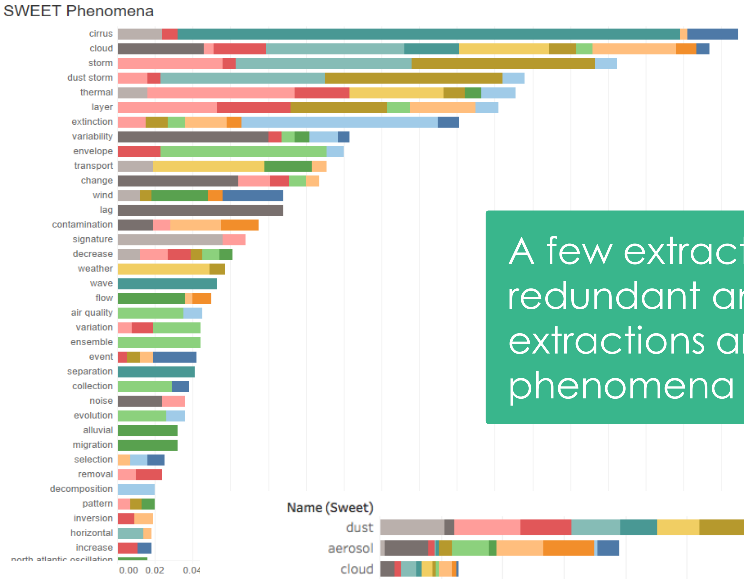
# SWEET Phenomena

**Good:**

- A few of the papers talk about differentiating cirrus clouds from dust
- Dust storm in top 4

**Bad:**

- Quite a few don't even appear to be phenomena
  - Thermal, decrease, layer, etc…
- Redundant extraction




A few extractions seem redundant and some extractions aren't even phenomena

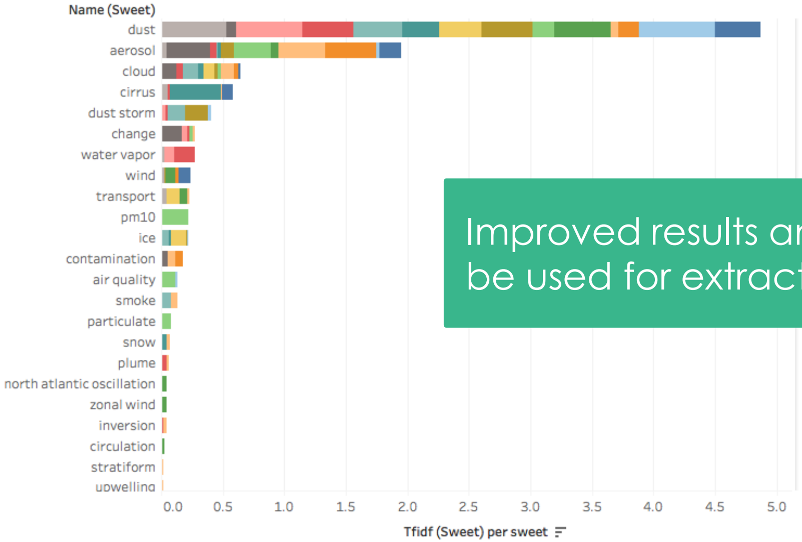
# Curated SWEET Phenomena

**Good:**

- Top ~7 results make sense scientifically

**Bad:**

- Still too generic to be helpful
- Contamination, transport may be helpful but not without context




Improved results and can be used for extractions

# Location extraction using named entity recognition (NER)

**Good:**

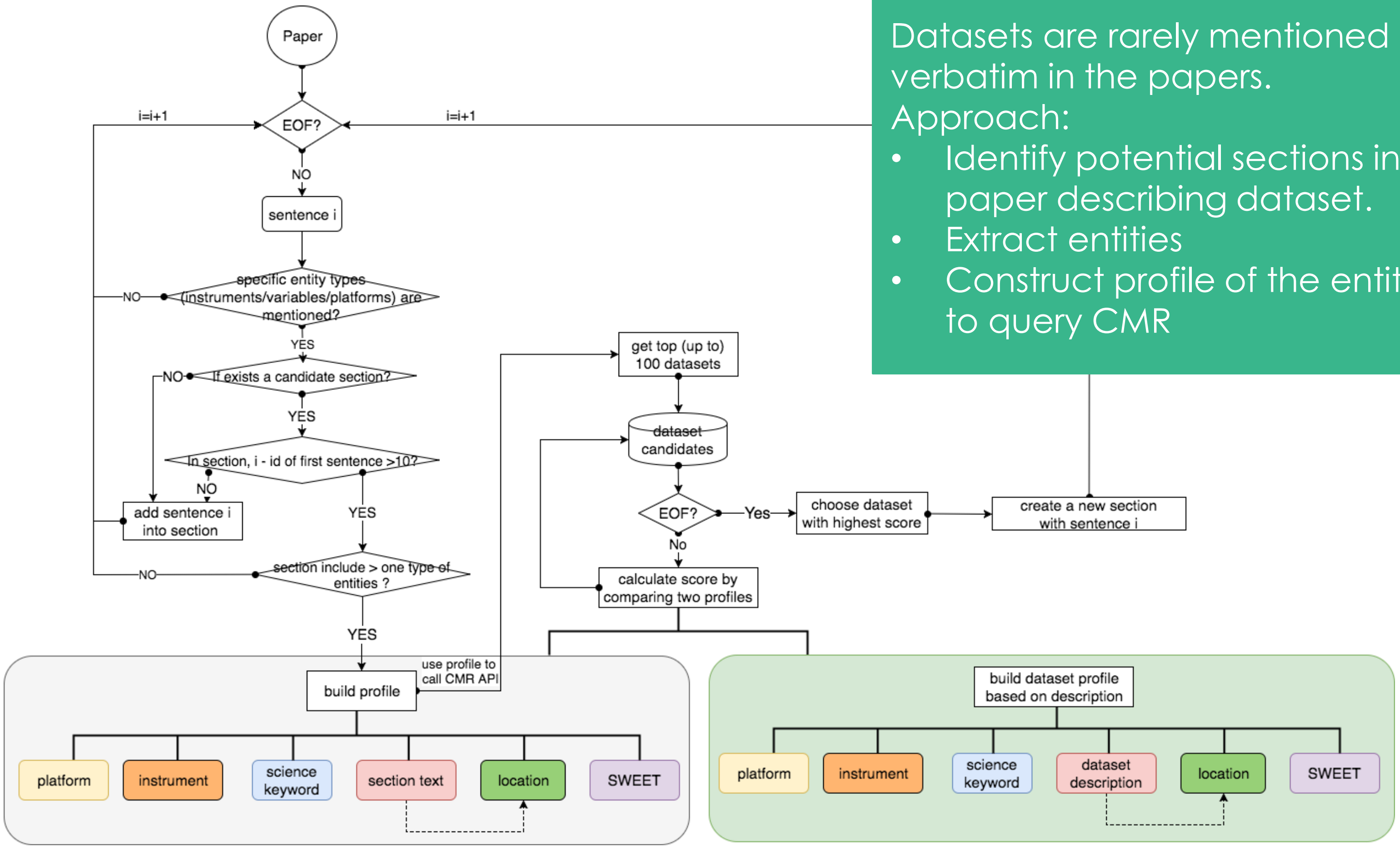
- Many of the locations are deserts or regions where deserts are located.
- Majority of the studies took place in China.

**Issues:**

- Some of the locations are very general (Earth, Atlantic, etc…)
- IR (Infra Red) acronym confused for Iran
- Redundancy: some locations mean the same thing but are worded differently ( Mongolia, Inner Mongolia)


- Many locations are deserts
- Some locations are too general

# Dataset Extraction Algorithm




Datasets are rarely mentioned verbatim in the papers.
Approach:

- Identify potential sections in the paper describing dataset.
- Extract entities
- Construct profile of the entities in to query CMR

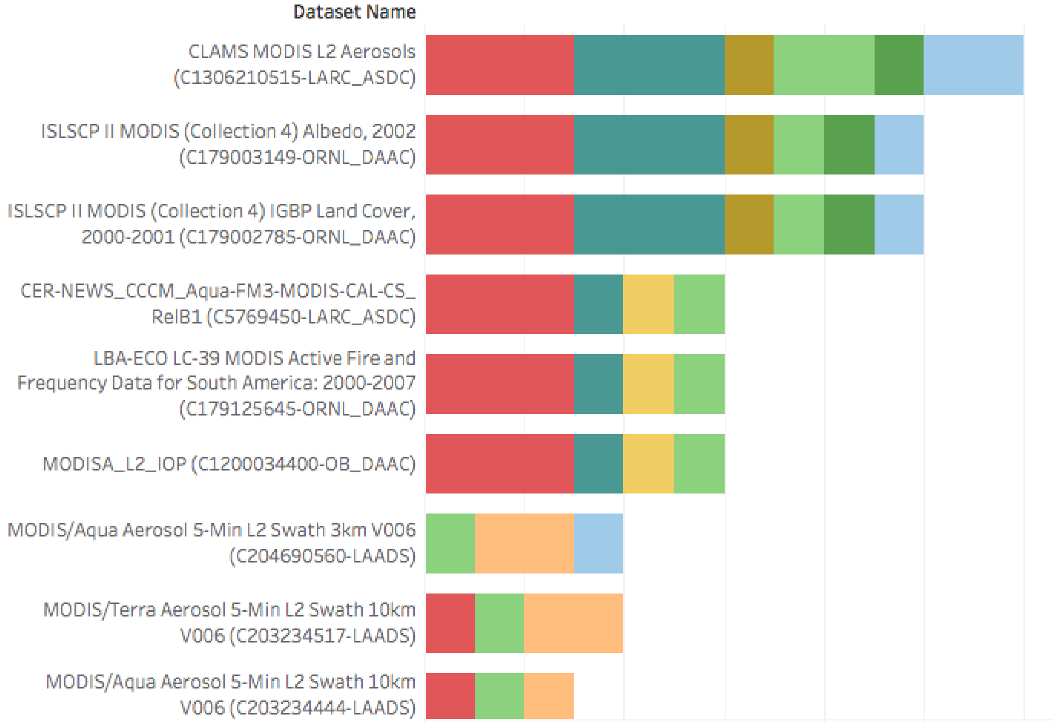
# Extraction Results

**Good:**

- Most of the datasets are dust or aerosol related
- Lists all MODIS datasets

**Issues:**

- Some datasets don’t make sense for dust studies
- Slight differences in the API query can provide very different results




- MODIS L1B data is what is used on mosot of the papers
- Dataset extraction results depends on Instrument/platform context and the precision of other entities extracted

# Lessons Learned

- Semantic entity identification is a difficult problem and heuristics based algorithms are brittle
- Use of existing taxonomies is helpful for specific entities (instruments/platforms) and less helpful for others (physical property/phenomena..)
    - Quality of the taxonomy impacts extraction results
    - CF is the least useful
    - SWEET covers most concepts and has the best potential for use
- Dataset profile approach is dependent on both the metadata and entity extraction quality
    - Metadata creators view dataset keywords differently than dataset users

# Next Steps: Begin Machine Learning Phase

- Use these algorithms to semi-automate training set generation
  - Have Atmospheric Science students provide URLs to 5-10 papers from their research area
  - Provide extractions and have students label results
- Train Deep Neural Networks for entity extraction
  - Evaluate results
- Build verb extraction and categorization to identify relationships between different entity types