Building Scalable Knowledge Graphs for Earth Science

Rahul Ramachandran[1], Manil Maskey[1], Patrick Gatlin[1], Jia Zhang[2], Xiaoyi Duan[2], J.J. Miller[3], Kaylin Bugbee[3], Sundar Christopher[3], Brian Freitag[3]

1 – NASA Marshall Space Flight Center; 2 – Carnegie Mellon University; 3 – University of Alabama in Huntsville

1. Introduction

Knowledge Graphs link key entities in a specific domain with other entities via relationships. From these relationships, researchers can query knowledge graphs for probabilistic recommendations to infer new knowledge. Scientific papers are an untapped resource which knowledge graphs could leverage to accelerate research discovery.

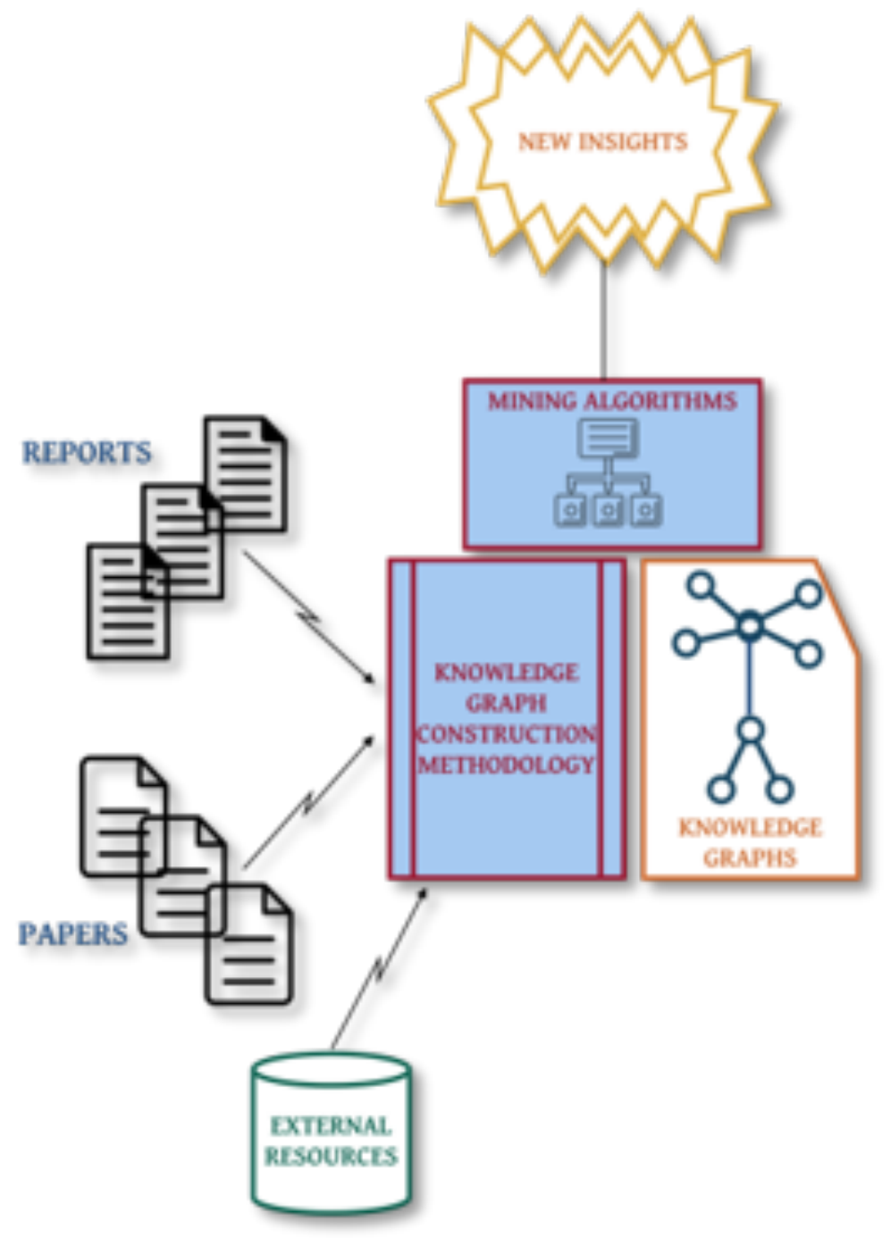
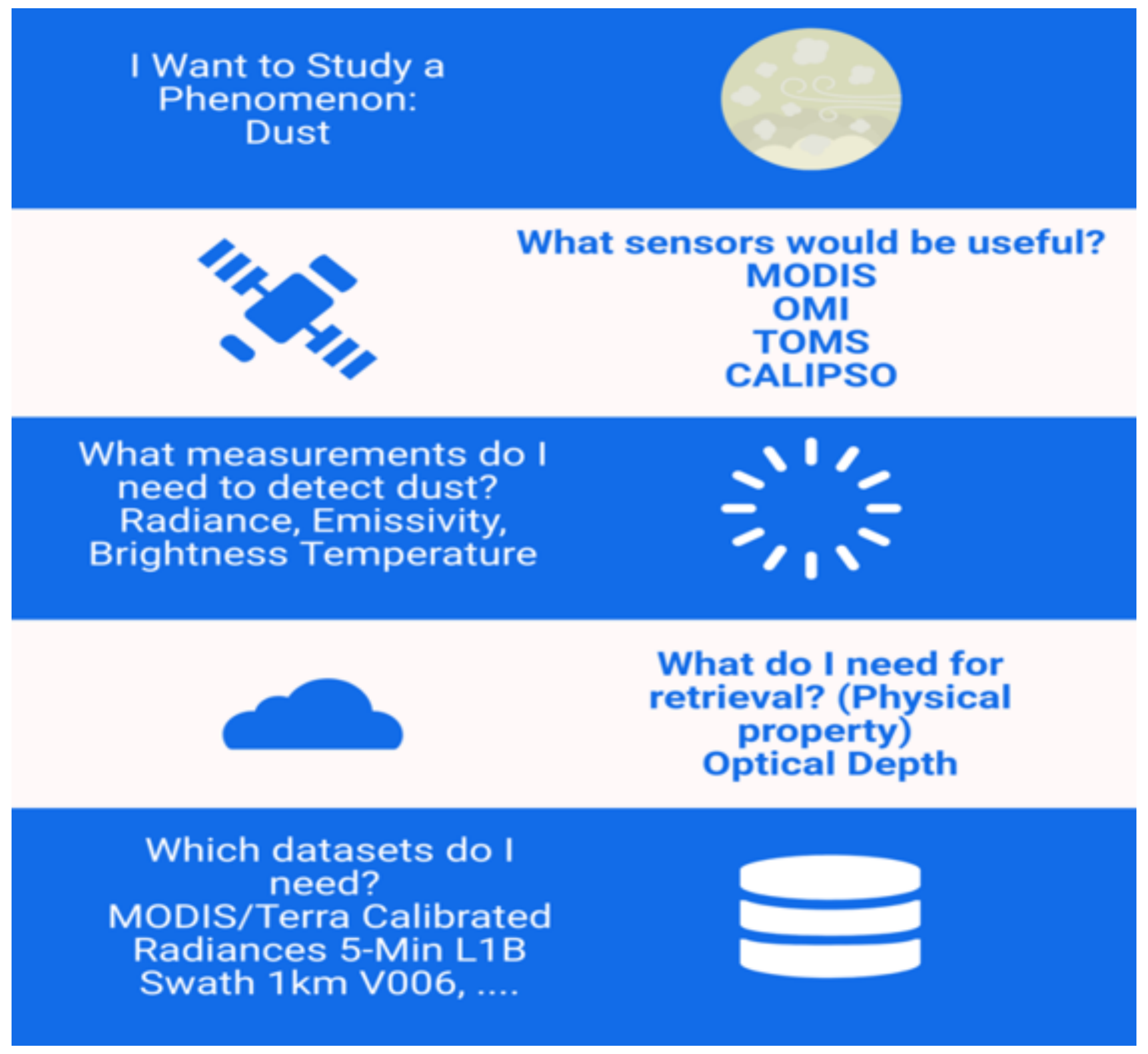


Figure 1. Overall conceptual diagram.

Goal: Develop an end-to-end (semi) automated methodology for constructing Knowledge Graphs for Earth Science.

2. Motivation



I Want to Study a Phenomenon: Dust

What sensors would be useful? MODIS OMI TOMS CALIPSO

What measurements do I need to detect dust? Radiance, Emissivity, Brightness Temperature

What do I need for retrieval? (Physical property) Optical Depth

Which datasets do I need? MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km V006, ...

Figure 2. Flow chart detailing scientific discovery

3. Methodology

Semantic Entity Identification

- **Step 0**: Parse paper for potential entities (Phenomena, Datasets, Instruments, Variables)
- **Step 1**: Weighted heuristic algorithm applied to scientific papers to identify entities to build a training dataset using existing taxonomies (GCMD, CF, SWEET)
- **Step 2**: Algorithm entity extraction evaluated by scientific experts and classified

Deep Learning Integration

- **Steps 3-11**: Use results from the heuristic algorithm to build training data for deep learning algorithms

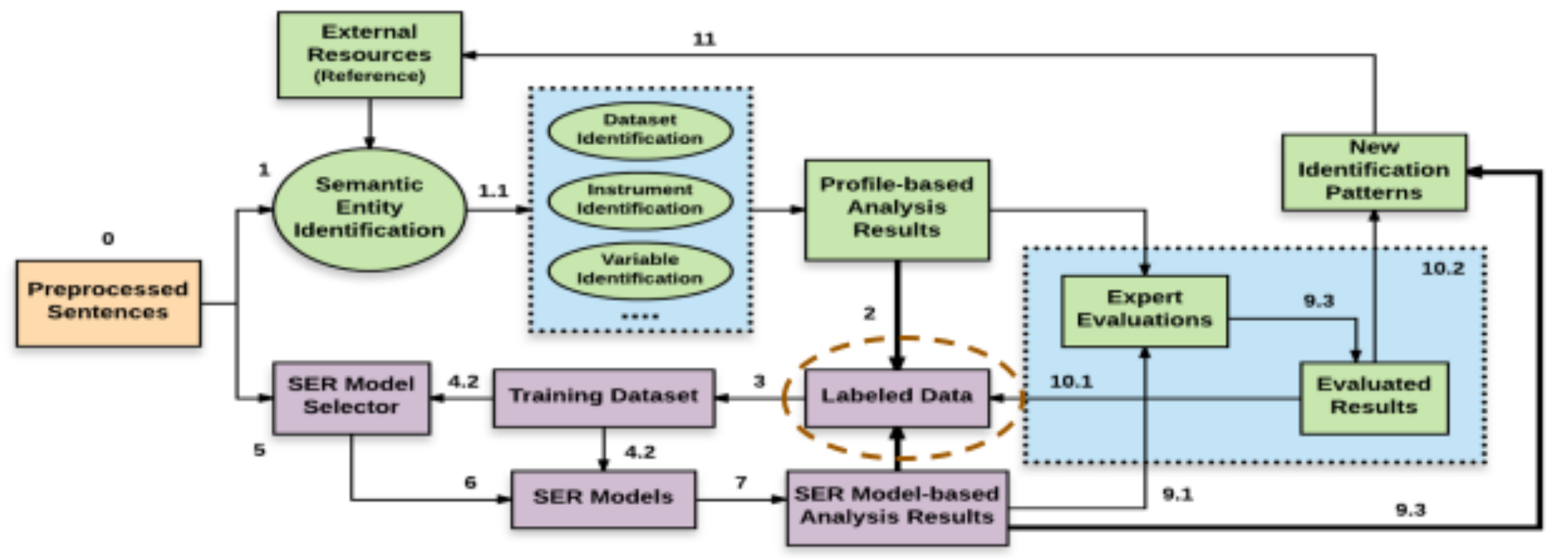


Figure 3. Methodology to build knowledge graphs.

4. Initial Analysis

Semantic entity identification performed on 13 papers focusing on airborne dust retrieval from satellites. Papers were manually parsed by science experts to qualitatively validate identifications from heuristic algorithms to optimize identification algorithm output. Statistical methods were applied to identification algorithm output to organize retrieved entities based on importance (Raw Count, Normalized TF, TF-IDF).

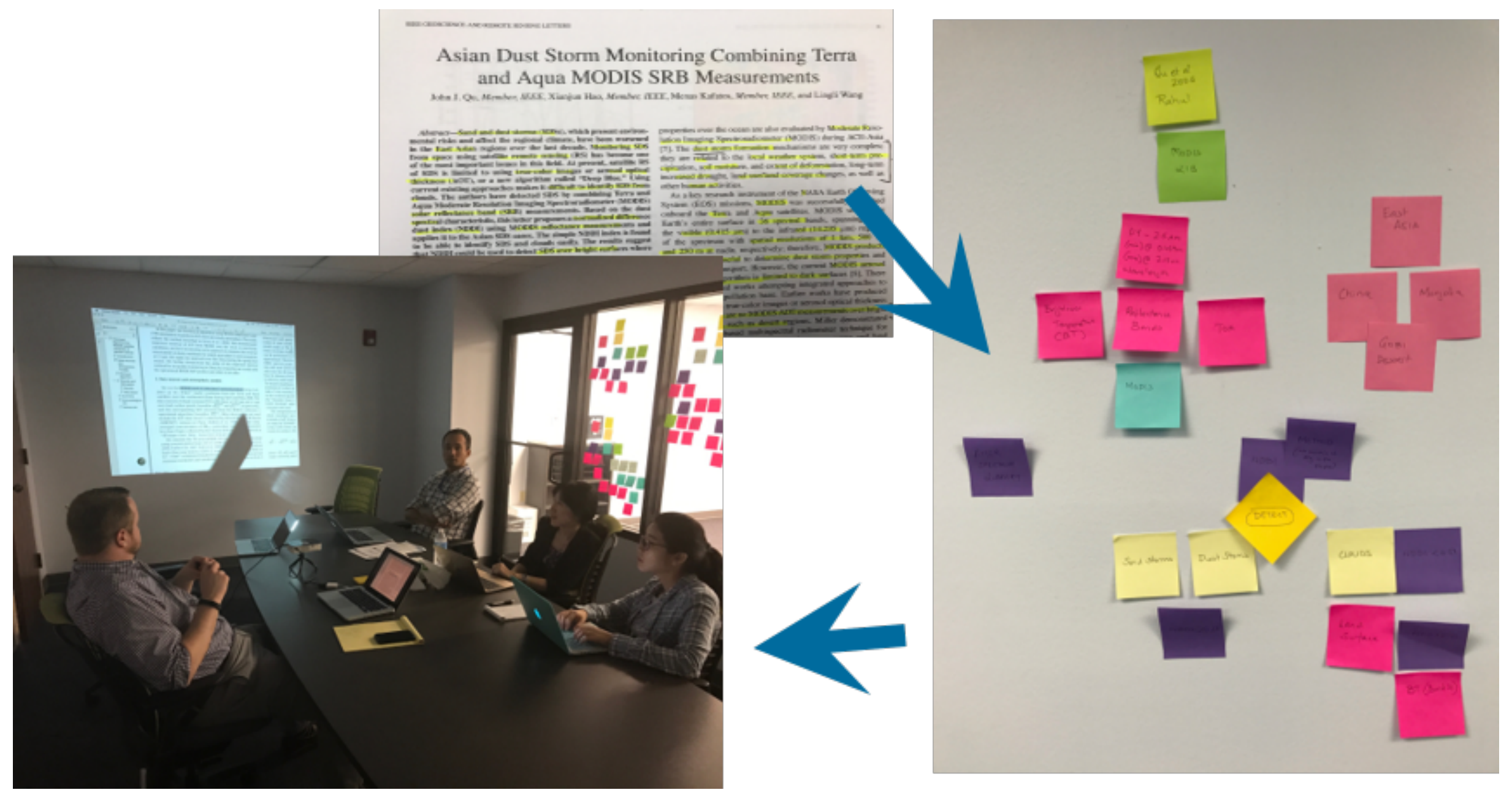

Asian Dust Storm Monitoring Combining Terra and Aqua MODIS SRB Measurements

Figure 4. Expert evaluation of semantic entity identification algorithm.

5. Results

Heuristics based entity identification algorithms provides mixed results. While key entities identified by expert reviewers are identified, where the algorithm ranks these entities in importance is not sufficient. Additionally, more generic entities are typically ranked higher than more relevant entities and multiple entities are extracted representing the same variable (e.g. optical depth, optical depth/thickness, aerosol optical depth, etc.)

GCMD Taxonomy

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

Issues

- 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

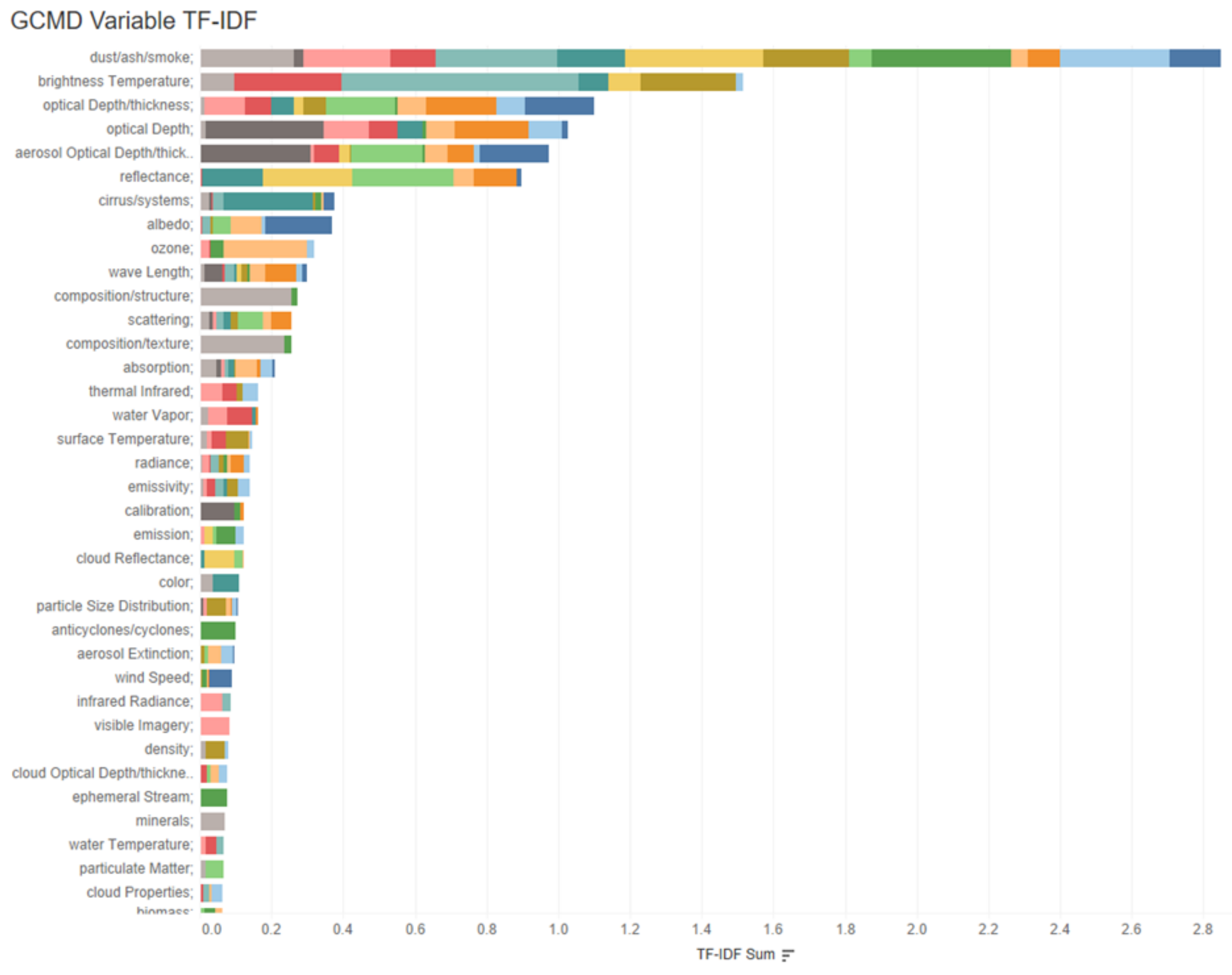


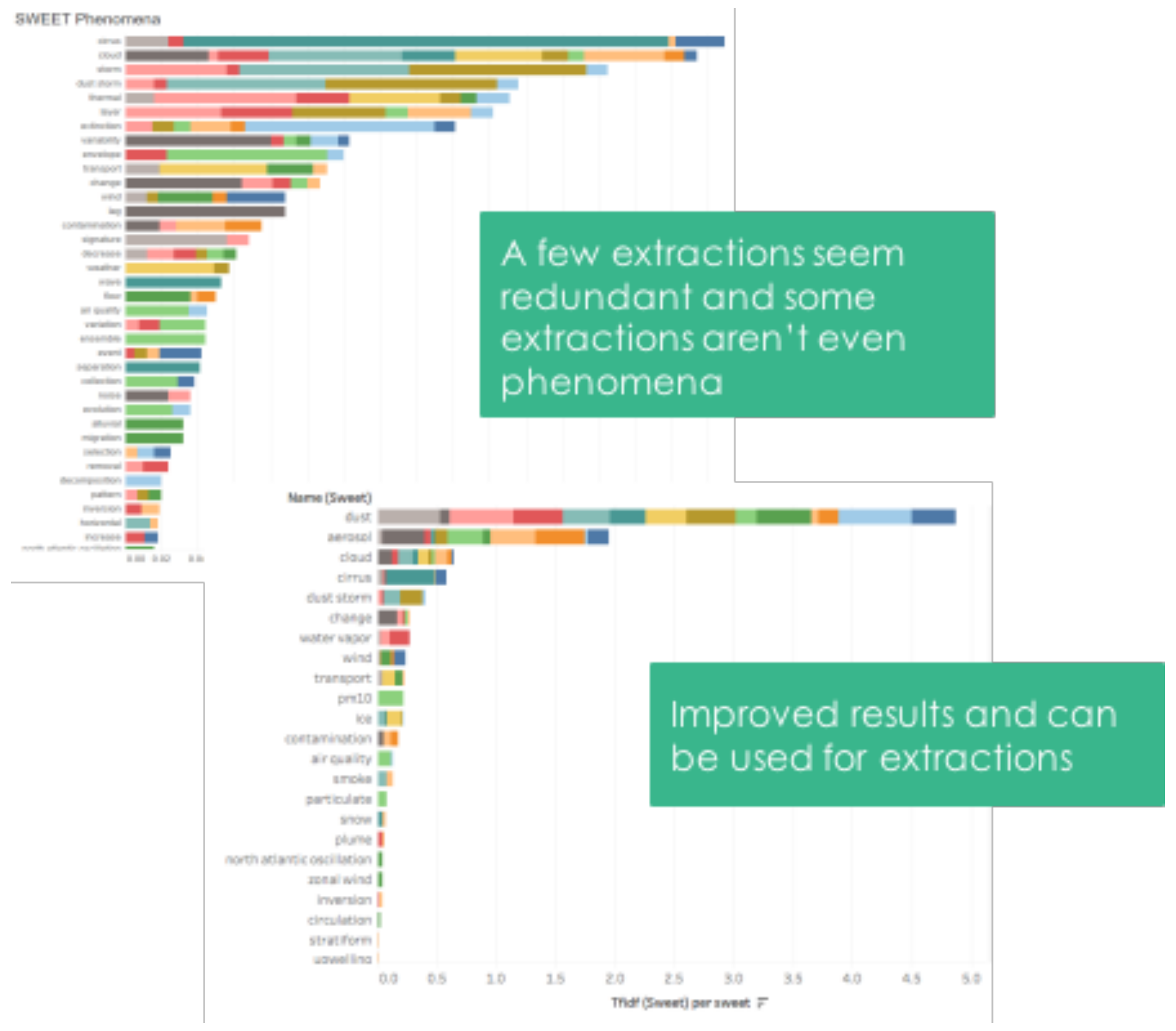
Figure 5. Algorithm performance using GCMD

SWEET Phenomena

- Many entities identified by the algorithm are not phenomena (e.g. thermal, decrease, layer, etc.)
- SWEET as a taxonomy is noisy which limits the performance of the identification algorithm

Curated SWEET Phenomena

- Top ~7 results make sense scientifically
- Cleaned up SWEET holds the best potential



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

Improved results and can be used for extractions

Figure 6 Algorithm performance using SWEET taxonomy.

6. Conclusion and Next Steps

Semantic entity identification was proven to be a difficult problem not easily addressed with heuristic algorithms. Entity identification in selected papers and validated by experts showed existing taxonomies can be useful for identifying specific entities, but not all. Performance of the heuristic identification algorithm was dependent upon the quality of the taxonomy. We plan to use these algorithms to semi-automate the process of building a training set and then train a Deep Neural Network for improved entity extraction

Contact: rahul.ramachandran@nasa.gov