# Building a Knowledge Graph for Earth Science


Rahul Ramachandran[1], Jia Zhang[2] and Tsengdar Lee[3]

[1] NASA Marshall Space Flight Center

[2] Carnegie Mellon University Silicon Valley

[3] NASA Headquarters



Presented at the NITRD Open Knowledge Network Workshop

Oct 4 – 5, 2017


Electrical & Computer ENGINEERING

DSIG

# Question to Watson and HAL:

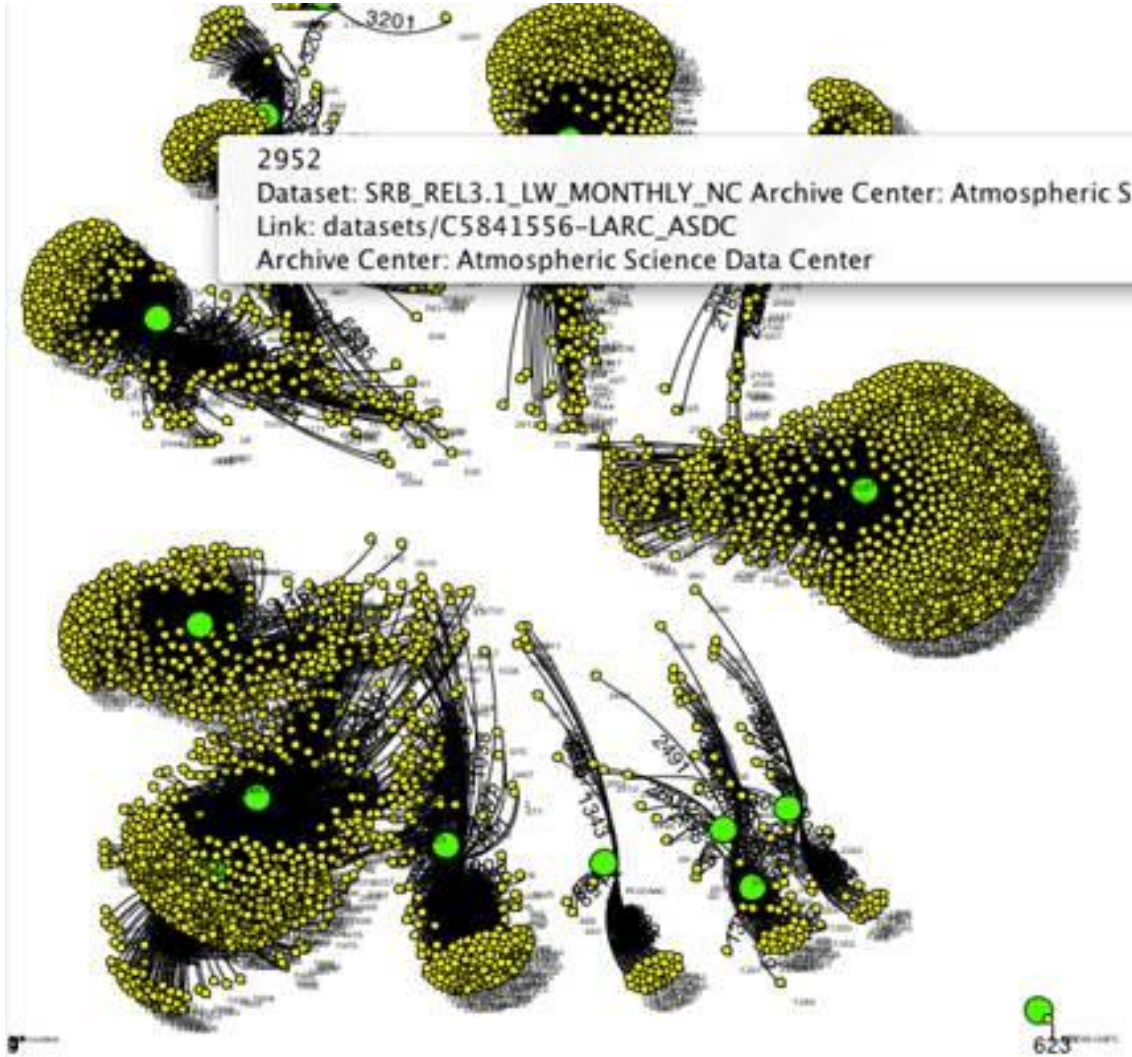
Q: This data set is used for predicting crop yields in CropYieldModelA under the future climate scenario RCP8.5.

Watson: What is temperature and precipitation data set in IPCC climate projection archive at URL... By the way, the technical document is available at URL…

Q: How can I grow potato to survive, now that I am alone on Mars?

HAL 9000: According to Journal of Aerospace Engineering, User Manual of Water Sequestration System, and Journal of Horticultural Research, Your best chance is to…

# Many Data Sets – Which one should I use?



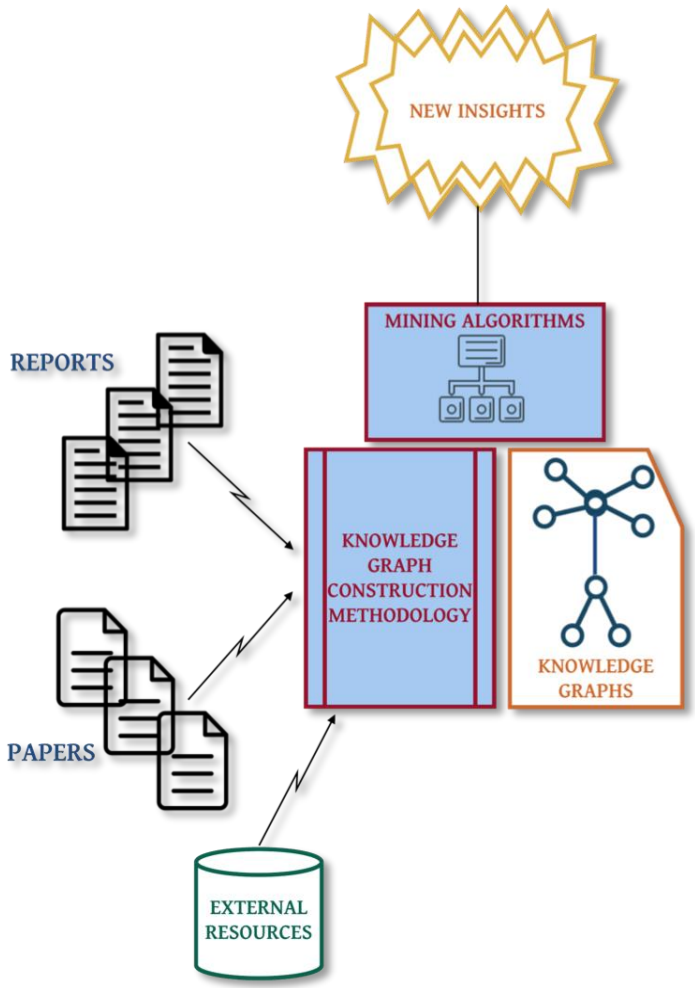

- Graph on the left, made from metadata in GCMD, represents data sets organized by their locations (data centers).
- Additional information may be used to associate the data sets to their use cases.

# Knowledge Graph Project

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

---

- The goal of this project is to develop an end-to-end automated methodology for incrementally constructing Knowledge Graphs for Earth Science.



# Why NASA Needs a Knowledge Graph

- Nearly all of the information contained within research papers is unstructured.
  - Difficult to search
- This project aims to build a traversable graph from unstructured text, tables, and figures within research papers.
- Addresses the challenge in extracting useful knowledge from the increasing volume of data and information.

Remote Sensing of Environment 114 (2010) 2575–2583

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

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

Jun Wang [a,b,*], Xiaoguang Xu [a], Robert Spurr [c], Yuxuang Wang [d], Easan Drury [e]

[a] Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, Lincoln, NE, USA
[b] Code 613.2, Climate and Radiation Branch, NASA Goddard Space Flight, Greenbelt, MD, USA
[c] RT Solutions Inc., Cambridge, MA, USA
[d] Department of Environmental Science and Engineering, Tsinghua University, Beijing, China
[e] Strategic Energy Analysis Center, National Renewable Energy Lab, Golden, CO, USA

ARTICLE INFO

*Article history:*
Received 1 February 2010
Received in revised form 27 April 2010
Accepted 26 May 2010

*Keywords:*
Remote sensing of aerosols
Particulate matter air quality
Atmospheric correction in dusty condition
Air quality in China

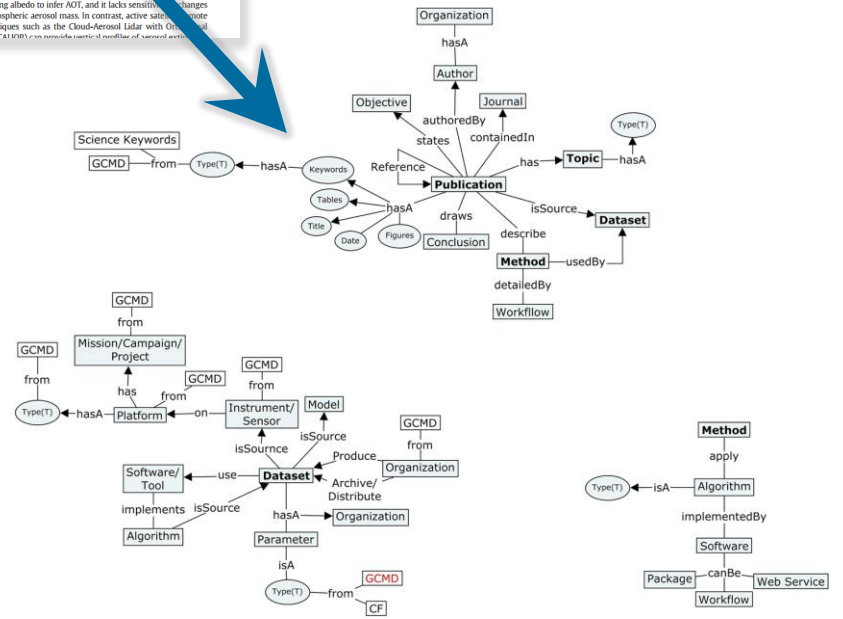
ABSTRACT

A new algorithm, using the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite reflectance and aerosol single scattering properties simulated from a chemistry transport model (GEOS-Chem), is developed to retrieve aerosol optical thickness (AOT) over land in China during the spring dust season. The algorithm first uses a "dynamic lower envelope" approach to sample the MODIS dark-pixel reflectance data in low AOT conditions, to derive the local surface visible (0.65 μm)/near infrared (NIR, 2.1 μm) reflectance ratio. Joint retrievals of AOT at 0.65 μm and surface reflectance at 2.1 μm are then performed, based on the time, location, and spectral-dependent single scattering properties of the dusty atmosphere as simulated by the GEOS-Chem. A linearized vector radiative transfer model (VLIDORT) that simultaneously computes the top-of-atmosphere reflectance and its Jacobian with respect to AOT, is used in the forward component of the inversion of MODIS reflectance to AOT. Comparison of retrieved AOT results in April and [illegible] of 2008 with AERONET observations shows a strong correlation ($R = 0.83$), with small bias (0.01), and [illegible] (0.17); the figures are a substantial improvement over corresponding values obtained with the MODIS C[illegible] 5 AOT algorithm for the same study region and time period. The small bias is partially due to the considerat[illegible]ust effect at 2.1 μm channel, without which the bias is −0.05. The surface $PM_{10}$ (particulate matter with diam[illegible] than 10 μm) concentrations derived using this improved AOT retrieval show better agreement with grou[illegible]ations than those derived from GEOS-Chem simulations alone, or those inferred from the MODIS Collect[illegible]T. This study underscores the value of using satellite reflectance to improve the air quality modeling and m[illegible]

© [illegible]evier Inc. All rights reserved.

1. Introduction

Atmospheric aerosol is a major concern for climate prediction and public health, but records of global aerosol distributions have only become available in the last decade from dedicated satellite observations such as MODIS (Remer et al., 2008) and the Multiangle Imaging SpectroRadiometer (MISR) (Kahn et al., 2005). Despite much progress [illegible] the slope of the reflectance is regulated by he[illegible]ependent Rayleigh scattering and aerosol absorption, and this relat[illegible] can be used to estimate the centroid height of absorbing aerosols [illegible]et al., 2007). However, such an algorithm requires *a priori* informa[illegible]on aerosol single scattering albedo to infer AOT, and it lacks sensiti[illegible] changes in lower tropospheric aerosol mass. In contrast, active sate[illegible]mote sensing techniques such as the Cloud-Aerosol Lidar with Or[illegible]al Polarization (CALIOP) can provide vertical profiles of aerosol extin[illegible]



# Example use case for KG Application

**I Want to Study a Phenomenon:**
**Dust**

**What sensors would be useful?**
**MODIS**
**OMI**
**TOMS**
**CALIPSO**

**What do I need to detect dust? (Physical Property)**
**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, ....**

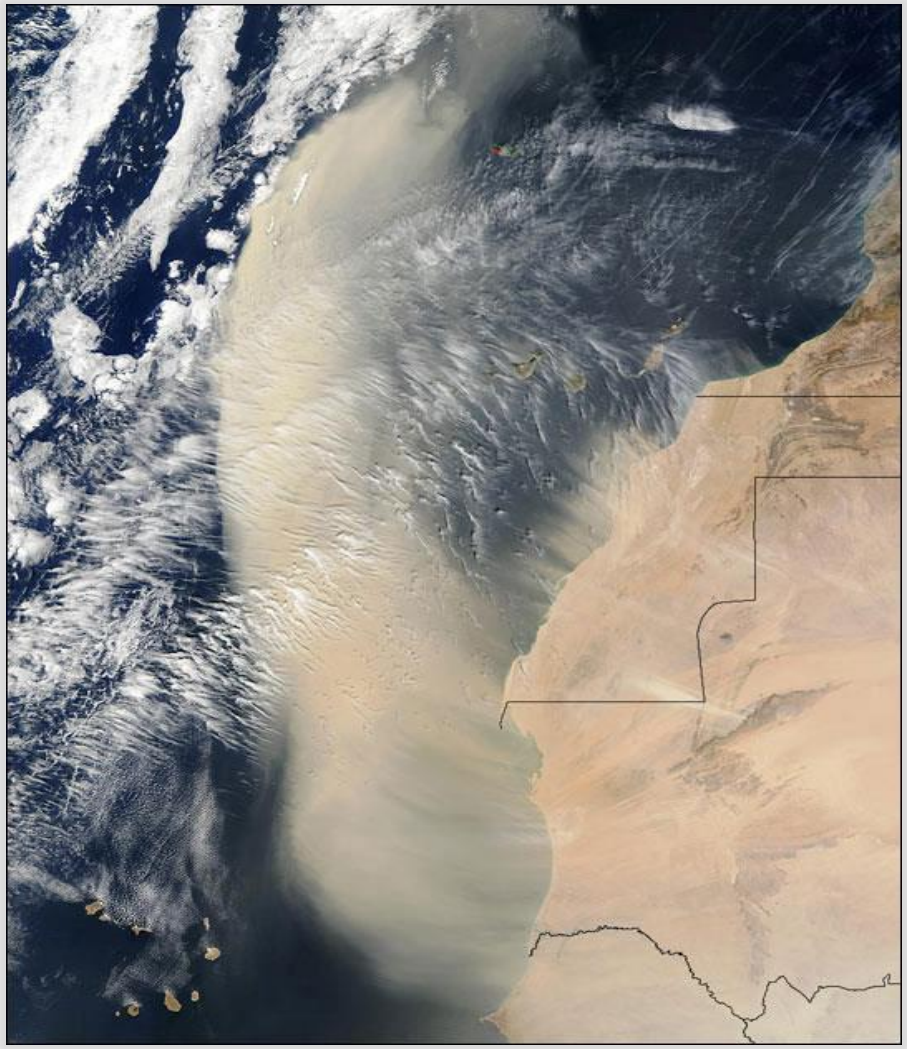
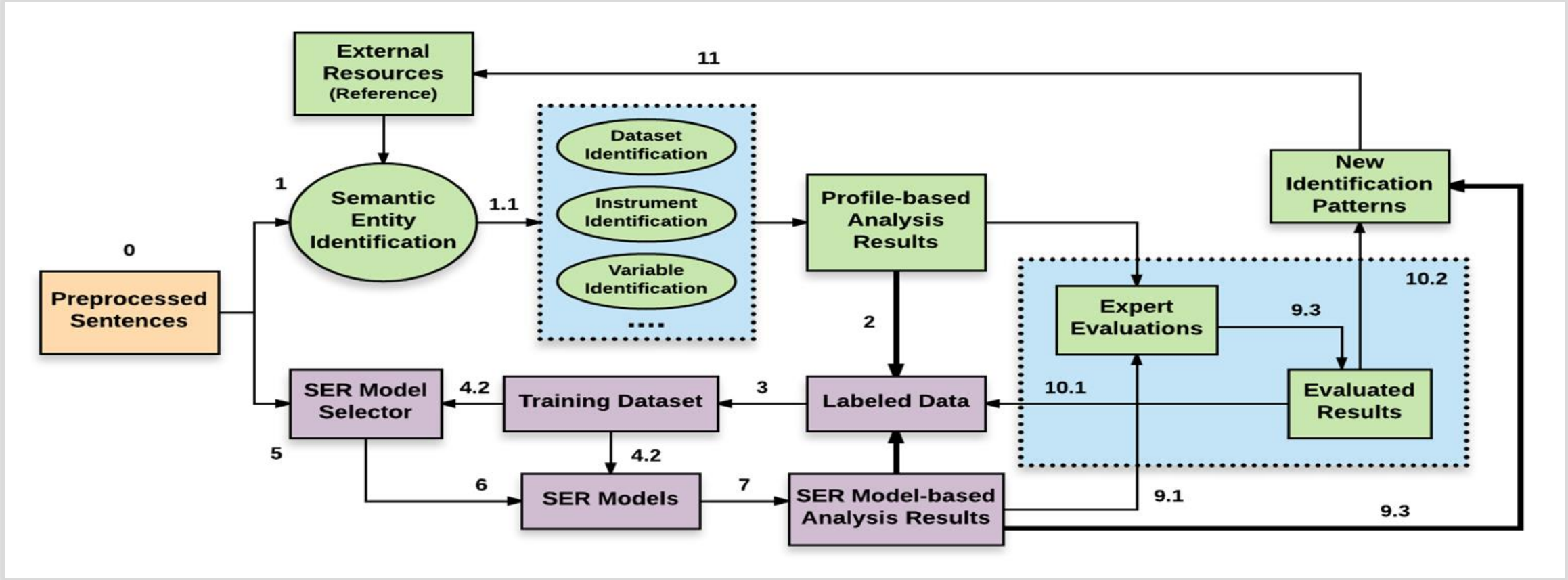


Image source:https://coastal.er.usgs.gov/african_dust/images/Canary.TMOA2004064LG.jpg

# Overall Methodology

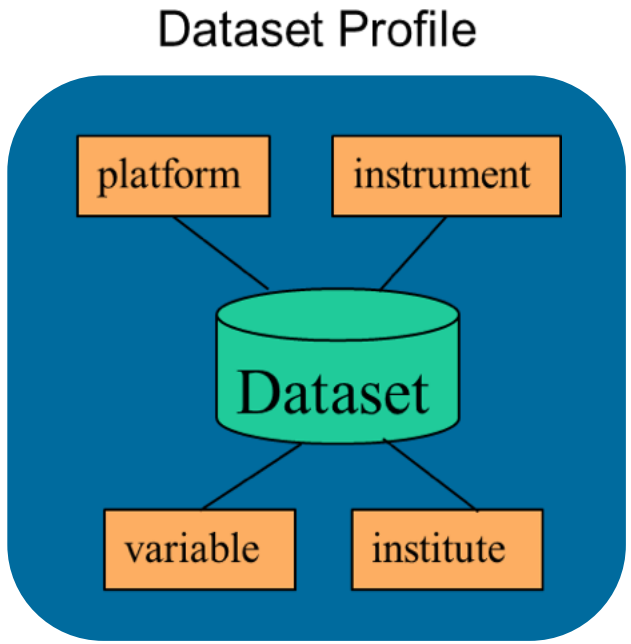



- Consists of two stages
  - Development of Heuristic algorithms to perform Semantic Entity Identification (Phenomena, Dataset, Instrument, Variable (Physical Property), Workflow...) to assist a human expert in building training data [Steps 0-2]
  - Use Deep Learning Algorithms to improve results [Steps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

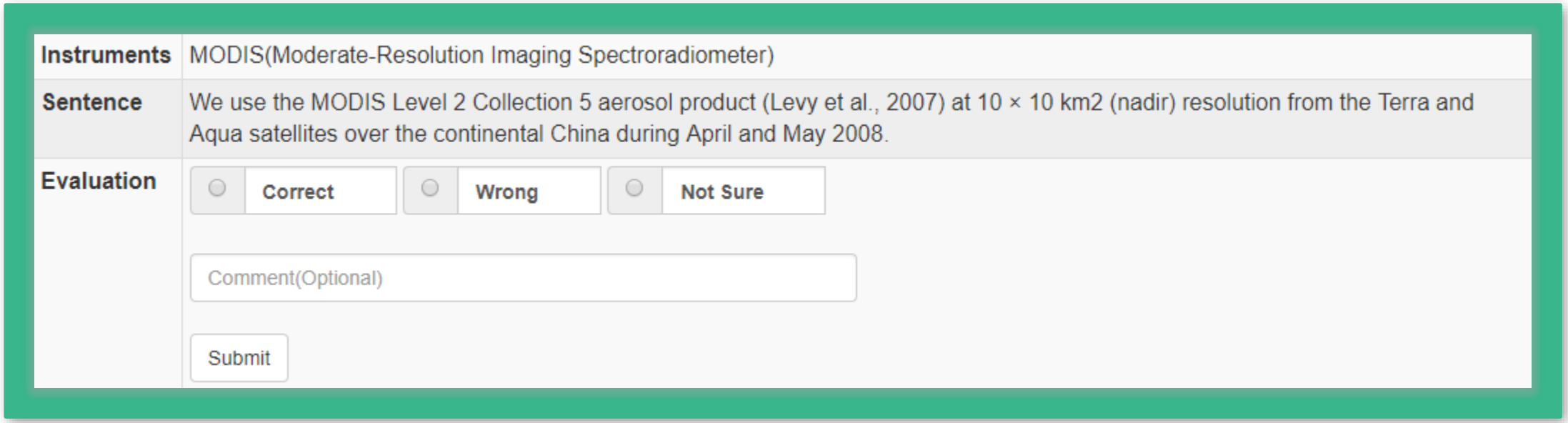
# Semantic Entity Identification (SEI) Algorithm

- Goal: extract entities to build a training dataset
- Use existing taxonomies where available
    - GCMD (variables, instruments, etc…)
    - SWEET (phenomena)
- Data set identification
    - Build profile based on extracted entities
    - Use profile to search NASA catalog for the most relevant dataset

- Use heuristics to prevent noise
    - Ignore "Background" sections of paper
- *Heuristic algorithms are brittle and need to be replaced with Deep Learning algorithms*



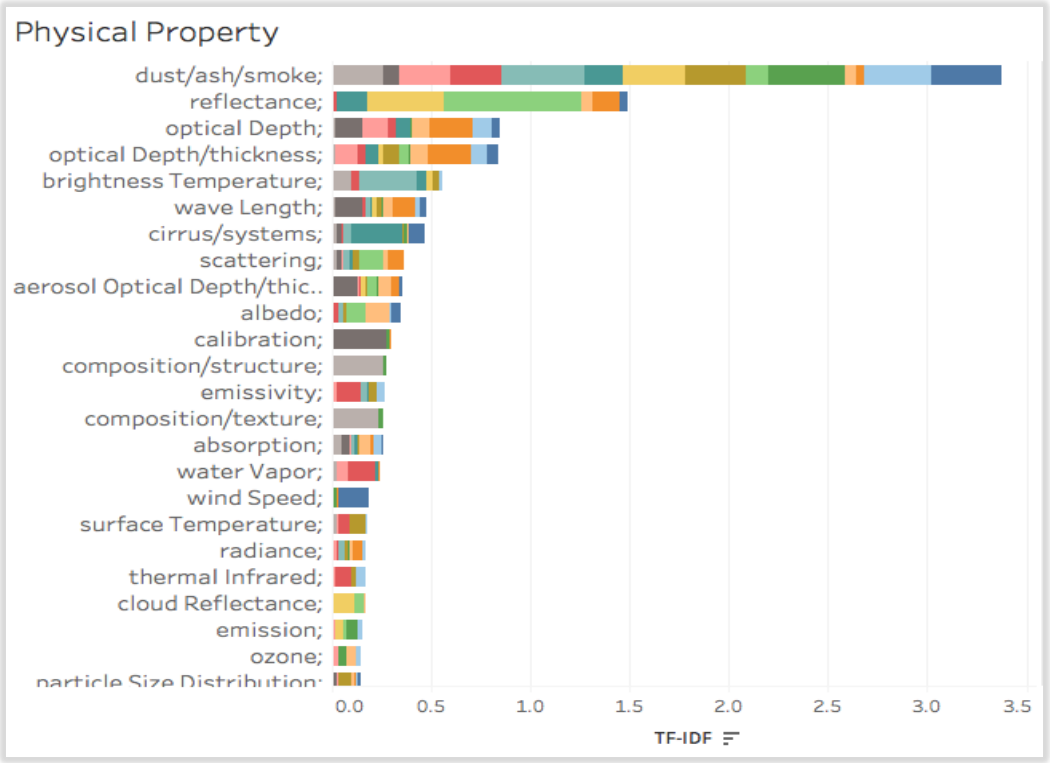
# Scaling up: Building Training Dataset for Deep Learning

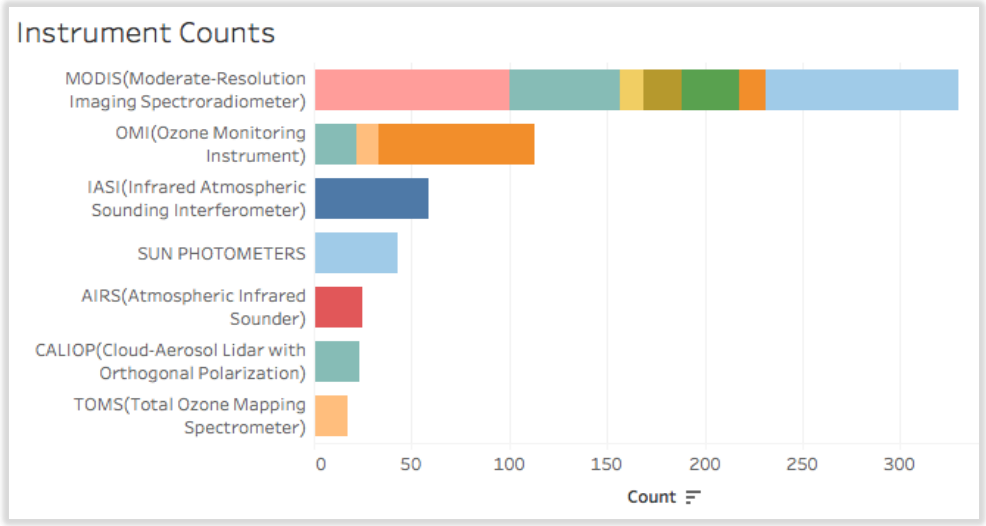



- Provide an easy to use tool for human experts to label results from the heuristic
- Build training data set for Deep Learning algorithm
- Can we use reCAPTCHA to scale up?

# Preliminary Results from SEI and Key Findings

Based on selected set of papers focused on study of dust from satellites



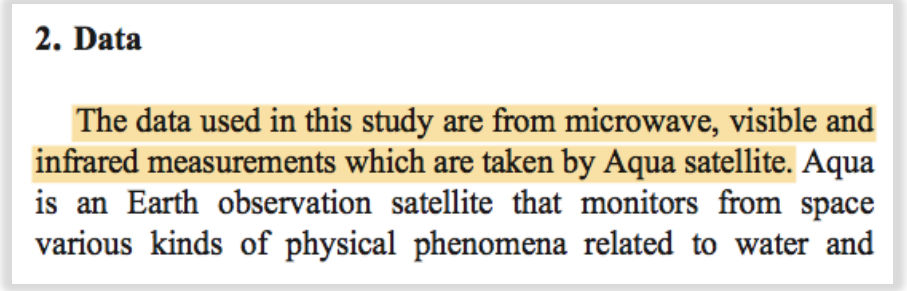




Key takeaways:

- Semantic entity identification is a difficult problem
- Use of existing taxonomies can help for specific instances (instruments/platforms)
- However, quality of the taxonomy can impact results for other instances (Physical Property, Phenomena)
- Dataset profile approach is dependent on the metadata quality in a data catalog

# Other interesting findings

- Dataset names are very rarely mentioned in the text. When they are, key information is usually missing.



**2. Data**

The data used in this study are from microwave, visible and infrared measurements which are taken by Aqua satellite. Aqua is an Earth observation satellite that monitors from space various kinds of physical phenomena related to water and

- The way the data is described in the metadata catalog does not match the way the data is described in the papers.
  - Example: MODIS Level 1B science keywords in the metadata are Spectral/Engineering > Infrared Wavelengths and Spectral/Engineering > Visible Wavelengths versus dust, brightness temperature, etc… extractions
- Terminology used in the papers can be inconsistent.
  - Synonyms for the same phenomena
  - Dust, dust storm, wind borne minerals, mineral dust

# How Big Data May Help

Steve Papa: Data volume is cumulative, analysis possibilities are combinatorial (video)

- Problem with BI is the idea of semantic layer.  Problem with semantic layer is you have to define the relationships for everything up front. This is directly at odds with the idea that with the information you can't know all the questions you want to ask up front. In fact every time you find anything interesting it begs for more questions which means the semantic layer is out of date.
- [Company] creates a crowd source semantic layer. Everyone uses a crowd source semantic layer everyday. It's called Google type-ahead search. Every time someone searches for something that has not been searched before you are going to benefit from what you are looking for from that type-ahead. Every incremental query grows the knowledge base. We can apply the similar idea to enterprise schema where someone comes up with a new query, there is a new relationship there someone can discover.

# Data & Challenges

## Current Access to Data

- NASA technical library
- AGU/Wiley published literatures

## Challenges:

- Data considered proprietary. Difficult to gather and difficult to share. The research is hard to scale.
- NLP is still far from NLU. We need significant advancement in NLU.