



# Language model for Earth Science for semantic search

EGU, 2022

**Rahul Ramachandran\*, Muthukumaran Ramasubramanian, Prasanna Koirala, Iksha Gurung, and Manil Maskey\***

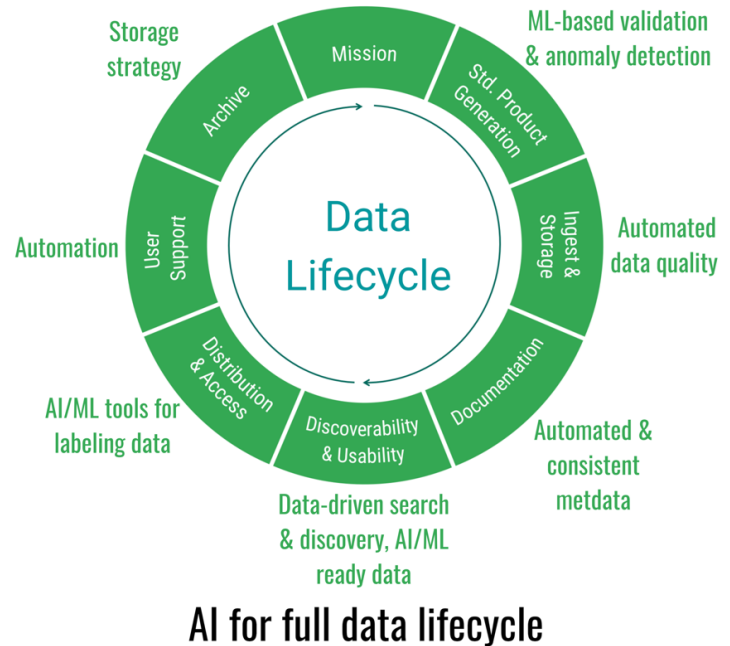
\*NASA/MSFC

The University of Alabama in Huntsville



# Drivers for Semantics/Knowledge Models

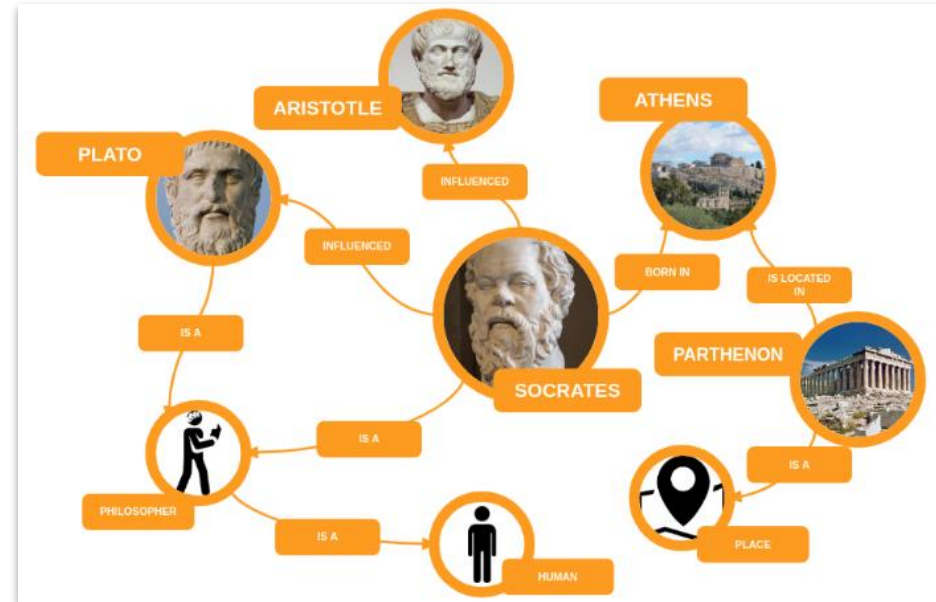
- Enable latent knowledge discovery
  - Ex: knowledge present in the material science scientific literature is used to recommend materials for functional applications several years before their discovery
- Improve data/information discovery and use
  - Ex: bridge the gap between a user's science application need and the appropriate instruments or datasets required to study that need.
- Augmenting data stewardship activities
  - Ex: keyword recommendations, summarizations for metadata



# Knowledge Graph

Definition: Aggregates structured and detailed information about a defined topic, enabling users to resolve their query without having to navigate and assemble information manually

- Top Down
  - Manually constructed
  - Curated by subject matter experts
- **Bottom Up**
  - Data-driven
  - Minimal human interaction
- **Hybrid**



# What Is a Word Embedding

*You shall know a word by the company it keeps (Firth, J. R. 1957:11)*

A way to encode **words** and entire documents as **numerical values** that capture the surrounding context.



Capturing **semantics**

Three ways to generate encodings

One-hot  
encoding



[1,0,0 ... 1]

Frequency  
based



Count

Concurrence

TF-IDF

Prediction  
based

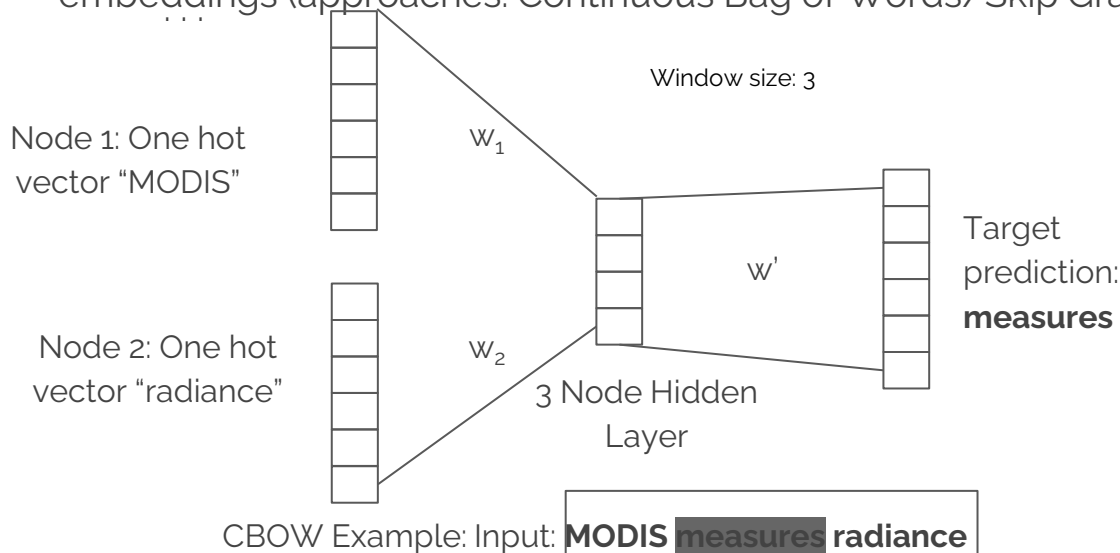


**Deep learning  
Machine learning**

# Word2Vec and Transformer Based Language Models

**Word2Vec** - Two layer neural network to generate word embeddings for a corpus of text

- NN tries to ensure words with similar context have similar embeddings (approaches: Continuous Bag of Words/Skip Gram)



[Final wts  $W \times$  [MODIS One Hot Vector]  $\Rightarrow$  MODIS (embedding)]

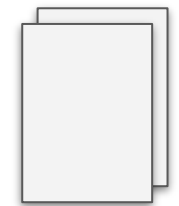
## Transformers (BERT, GPT-3)

- A new neural network which is based on attention based encoder/decoder type architecture
- "Attention" - ability to reference (attend) previous words (context) for the generated word
  - Learned during training
  - Unlike RNN, GRU and LSTMs it has infinite window (not limited to sentence) but can use entire context

"Attention is all you need" - Vaswani et al (2017)  
<https://arxiv.org/abs/1706.03762>

# Data-Driven Knowledge Graph Using NLP

## Idea



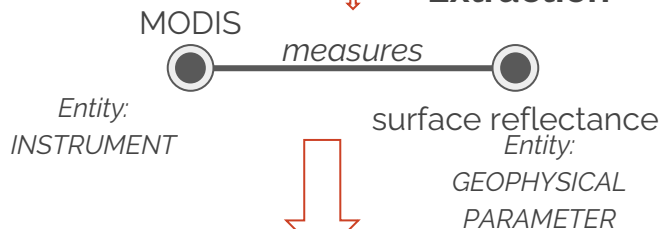
Knowledge  
Stored in  
Documents



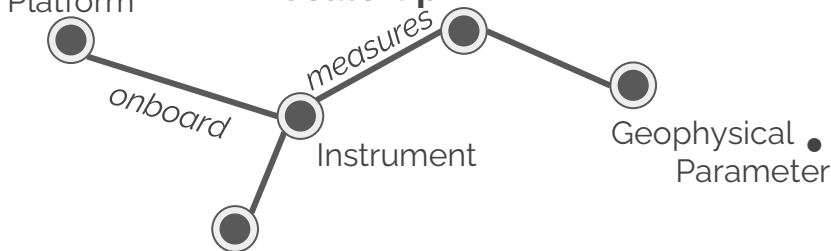
"MODIS measures surface reflectance"



## Entity/Relationship Extraction



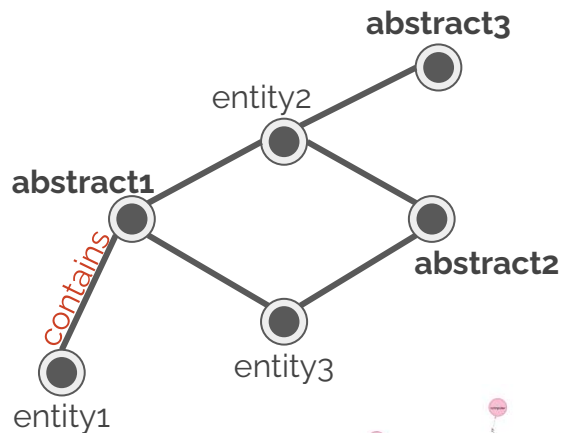
## Scale Up



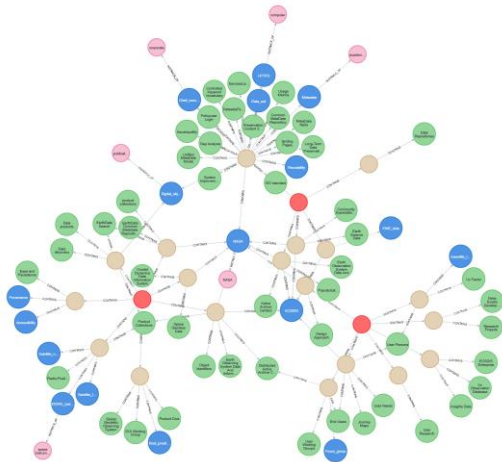
- Scientists write using **complex sentences**.
  - Subject-verb-object is not a common sentence construction in real-world articles.
- Entity extraction for scientific concepts is **non-trivial**.
  - Use existing vocabularies to help
  - Off-the-shelf libraries can be limiting
- While in theory entities in a domain are limited, verbs are **unbounded** leads to large, unusable graphs (Low signal to noise)
- **Rethink** the granularity of the graph and its conceptual model for it to be **practical** (useful)

Simplifying the graph

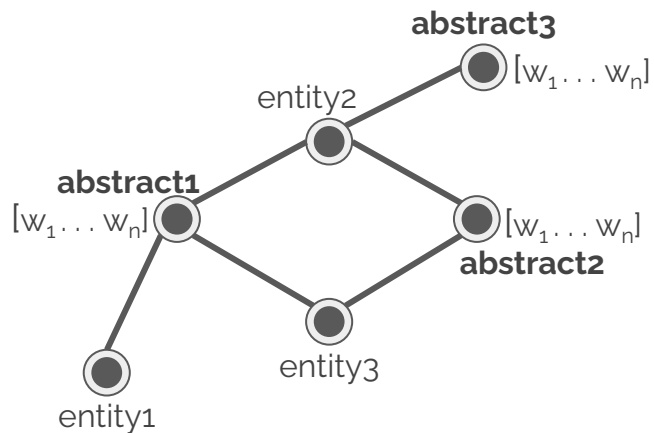
## Solutions



- Abstracts
- Sentences
- Entities
- Instances



Convolving embedding with the graph to improve semantics



# Putting It All Together

Given a query: "What instrument measures surface temperature?"



Vectorize

$[w_{11}, w_{12}, w_{13} \dots w_{1n}, w_{21}, w_{22} \dots w_{2n}, w_{31} \dots w_{3n}, w_{41}]$



Cosine similarity against convolved (weighted) embeddings of adjacent nodes



Match

Best result

# Example Query and Result

**Query:** What is the aerodynamic diameter of fine particulate matter?

## **Result (convolution layer 3)**

**Source:** Aerosol optical and microphysical properties over the Atlantic Ocean during the 19th cruise of the Research Vessel Akademik Sergey Vavilov

This Paper presents Aerosol optical Depths in the total atmospheric Column, Aerosol Size Distributions, Number Concentrations and Black Carbon Mass Concentrations at the Deck Level measured in October-December 2004 on board the R/V Akademik Sergey Vavilov. Aerosol optical Depths measured within the Spectral Range 0.34-4.0  $\mu\text{m}$  were close to background oceanic Conditions ( $\sim 0.04$ - $0.08$ ) in the high-latitude Southern Atlantic. Angstrom Parameters derived within 440-870 nm and 870-2150 nm Spectral Ranges did not exceed 0.6, yielding Averages of 0.34 and 0.12, respectively. The Mass Concentration of Black Carbon varied within the Range 0.02-0.08  $\text{mg}/\text{m}^3$  in the 34-55 degS latitudinal Belt. The Average of 0.04  $\text{mg}/\text{m}^3$  (s.d.  $\sim 0.015$ ) is close to the reported Results for the remote Areas of the South Indian Ocean. **Aerosol Volume Size Distributions measured within the Size Range of 0.4-10  $\mu\text{m}$  can be characterized by a geometric Volume mean Radius  $\sim 3 \mu\text{m}$ .** This is consistent with the columnar Retrievals reported by the Aerosol Robotic Network (AERONET).

**Sentence score:** 2, **Abstract score:** 2

Rudimentary prototype: <http://qes.nasa-impact.net/>

# BERT- E

**Objective:** Create a Earth science specific language model using the literature corpus from AGU and AMS

## Approach

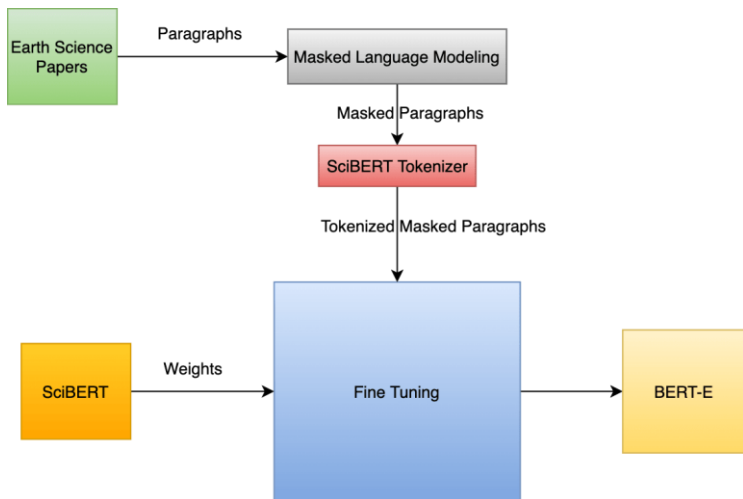
- Tuned existing SciBERT using 250K Earth science journal papers using a masked language model
- Tested BERT-E on the downstream application GCMD Science Keyword Recommender

## Results

- Results better than Word2Vec and SciBERT
- Value in fine tuning BERT using domain specific corpus
- Provides a benchmark dataset for Earth science

## Issues

- Tokenizer - we want control/ learn optimal tokens from our corpus of 200k specific papers
- Generic models are more prone to hallucinations and can be improved by training the model from scratch



<https://huggingface.co/nasa-impact/bert-e-base-mlm>

**Thank you.**

---

