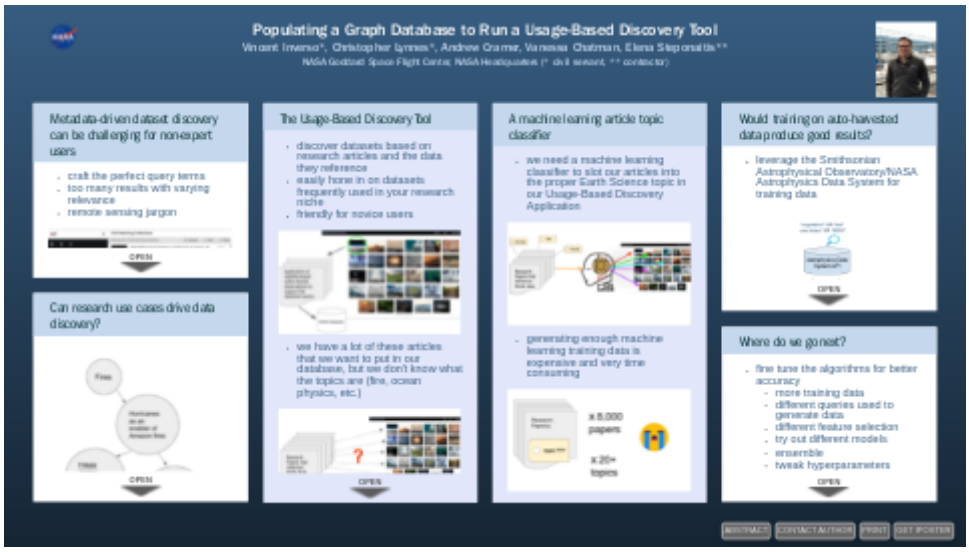
Populating a Graph Database to Run a Usage-Based Discovery Tool



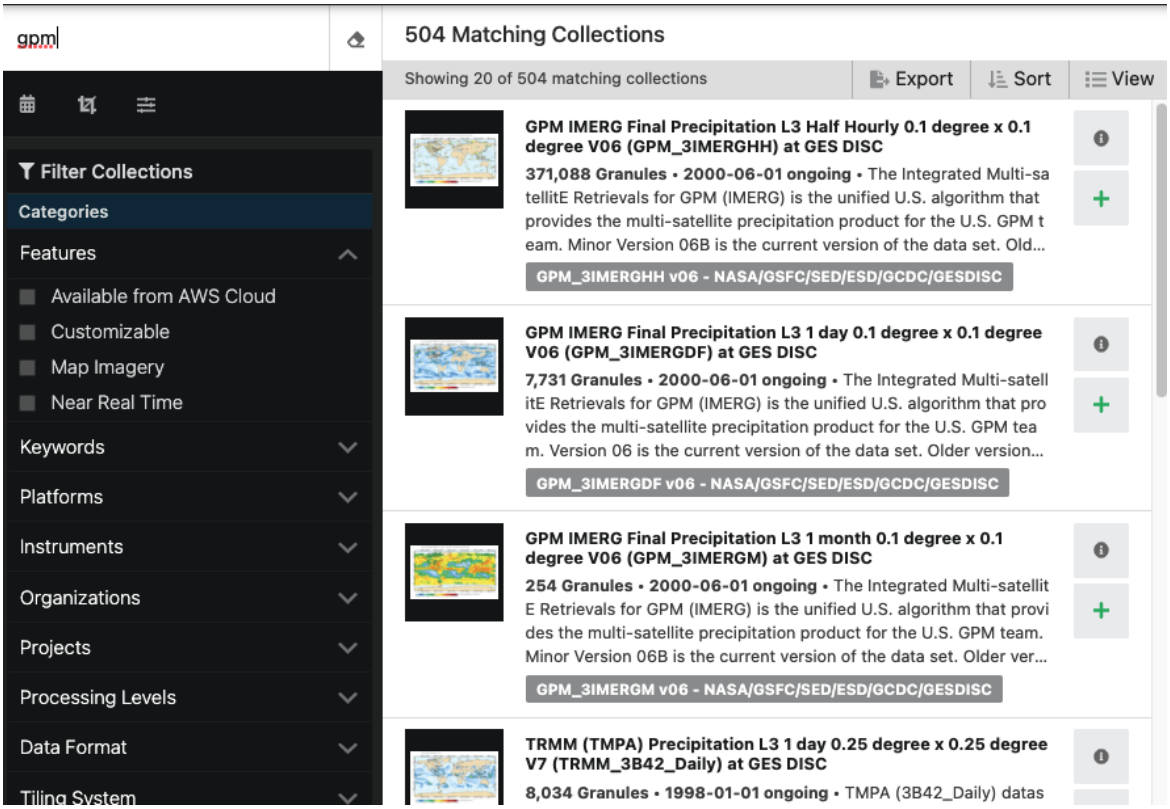
Vincent Inverso, Christopher Lynnes, Andrew Cramer, Vanessa Chatman, Elena Steponaitis

NASA Goddard Space Flight Center (civil servant), NASA Goddard Space Flight Center (civil servant), Texas A&M University, Northeastern University, NASA Headquarters (contractor)

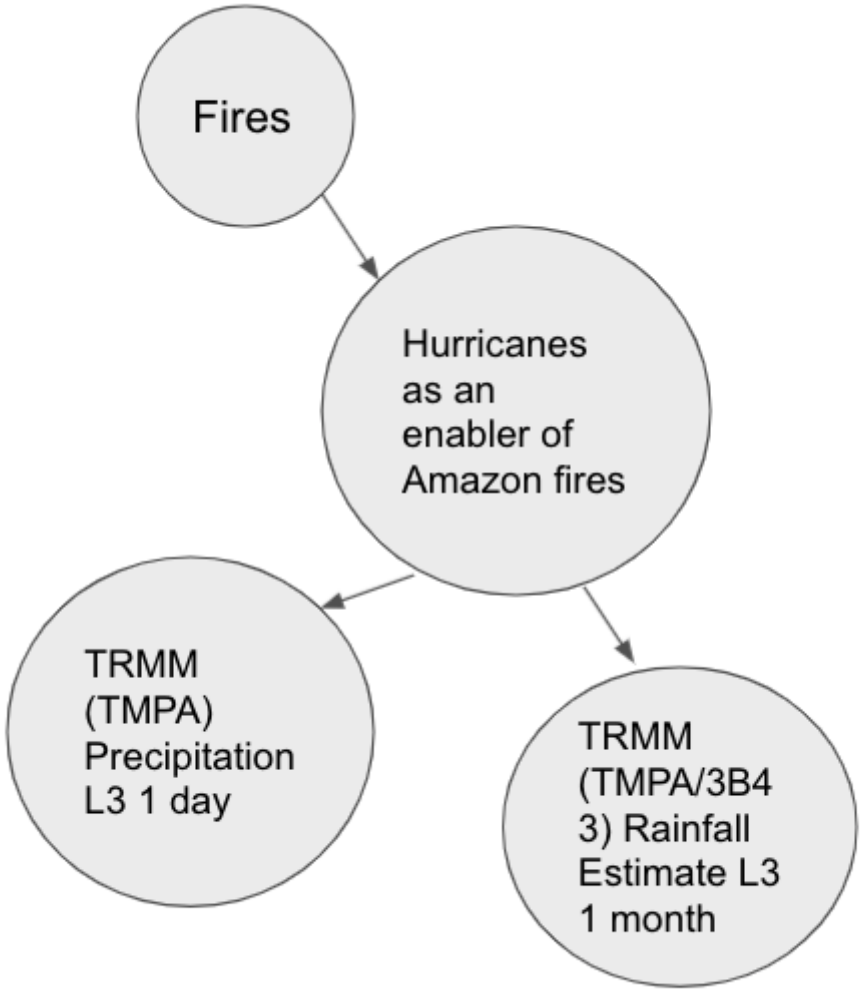
PRESENTED AT:

METADATA-DRIVEN DATASET DISCOVERY CAN BE CHALLENGING FOR NON-EXPERT USERS

- craft the perfect query terms
- too many results with varying relevance
- remote sensing jargon



CAN RESEARCH USE CASES DRIVE DATA DISCOVERY?



Dataset acronyms:

TRMM = The Tropical Rainfall Measuring Mission

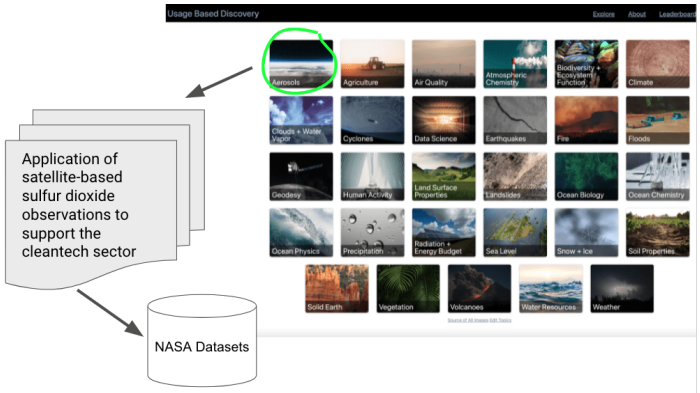
TMPA = TRMM Multi-satellite Precipitation Analysis

TRMM_3B43 = NASA short name / identifier for dataset

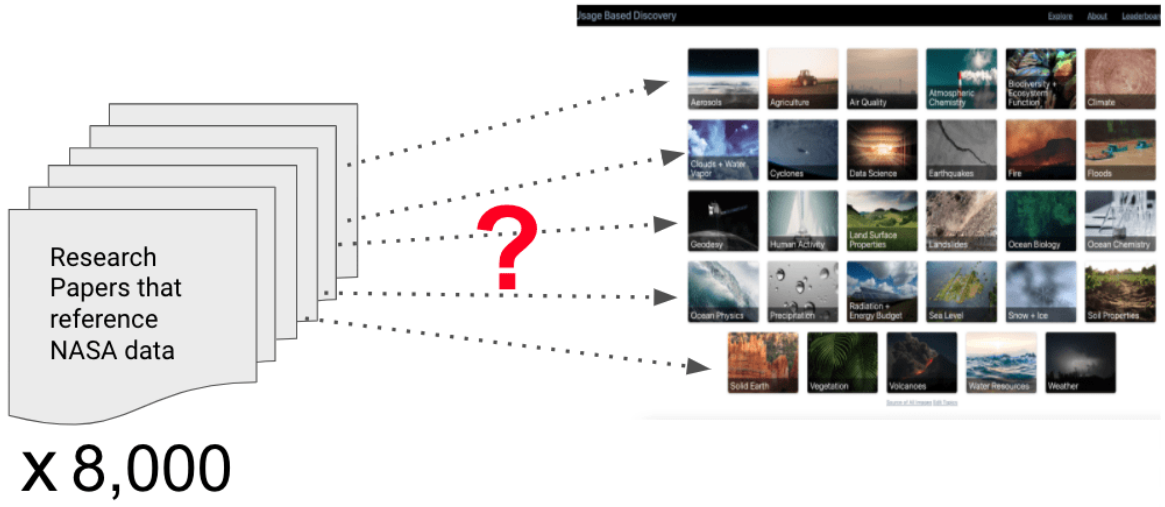
L3 = level three

THE USAGE-BASED DISCOVERY TOOL

- discover datasets based on research articles and the data they reference
- easily hone in on datasets frequently used in your research niche
- friendly for novice users

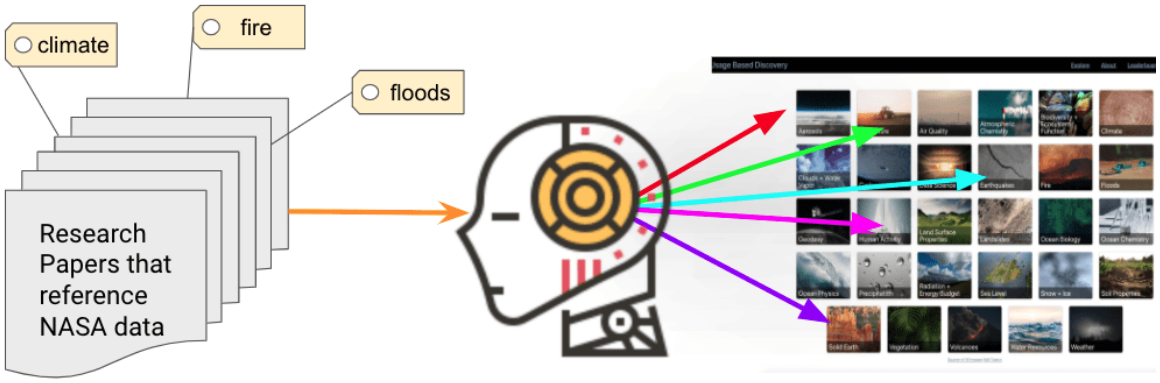


- we have a lot of these articles that we want to put in our database, but we don't know what the topics are (fire, ocean physics, etc.)

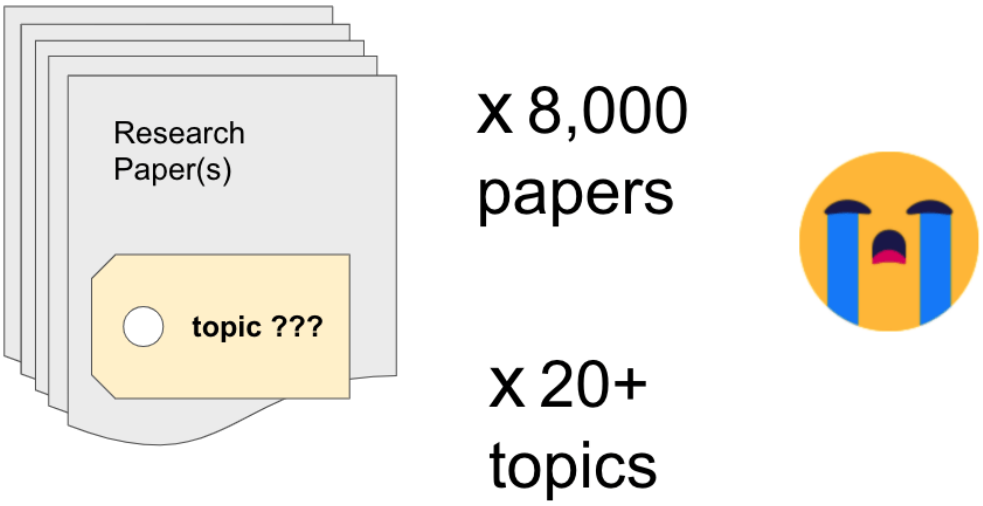


A MACHINE LEARNING ARTICLE TOPIC CLASSIFIER

- we need a machine learning classifier to slot our articles into the proper Earth Science topic in our Usage-Based Discovery Application

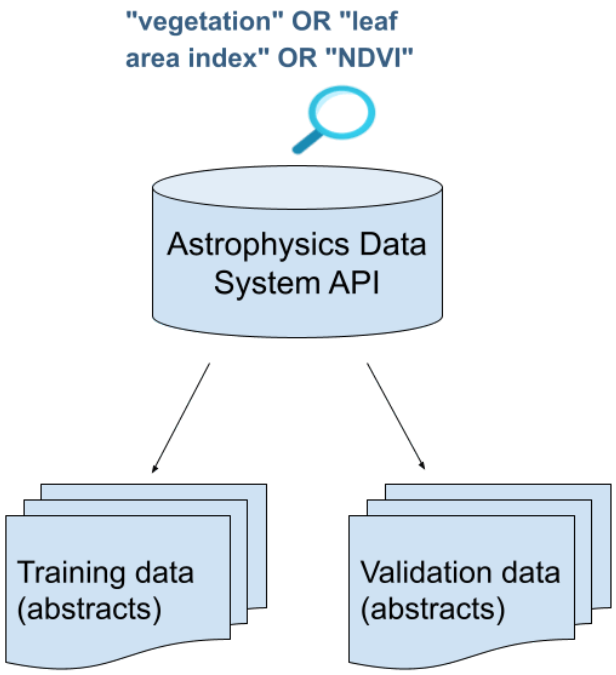


- generating enough machine learning training data is expensive and very time consuming



WOULD TRAINING ON AUTO-HARVESTED DATA PRODUCE GOOD RESULTS?

- leverage the Smithsonian Astrophysical Observatory/NASA Astrophysics Data System for training data



- manually curate the test (holdout) dataset for final evaluation of our models (k nearest neighbors, random forest, decision trees, transformers)
- 80%+ test accuracy for initial binary classification results

Acronyms:

NDVI = Normalized Difference Vegetation Index

API = Application Programming Interface

NASA = National Aeronautics and Space Administration

WHERE DO WE GO NEXT?

- fine tune the algorithms for better accuracy
  - more training data
  - different queries used to generate data
  - different feature selection
  - try out different models
  - ensemble
  - tweak hyperparameters
- run the final model(s) on our corpus of 8,000 research articles
- populate the Usage-Based Discovery graph!

ABSTRACT

Most dataset discovery tools for Earth Observation data rely on descriptions and other metadata of the datasets, using keyword searches or attribute filtering to determine relevance. However, these descriptions often do not include the potential uses of the data. Thus, a user working on floods will rarely see few if any rainfall datasets show up in such a search. The Usage Based Discovery tool, on the other hand, offers usage instances to the user, either research articles or applications, along with the datasets that those usage instances used. This allows a user, particularly one new to the world of Earth Observation data, to investigate which datasets are used in similar cases.

The information that powers Usage-Based Discovery is a graph database of relationships of usage to dataset and usage to topic, allowing the user to narrow their search for similar cases. In order to scale out to a graph database rich enough to provide a satisfactory user experience, we combine manual and automated processes to populate the graph.

The initial content of the graph has been seeded primarily via human-aided data curation methods, using sites like Google Scholar. To scale up this effort, we've employed crowdsourcing. It is easy for anyone to contribute to our graph using their Open Researcher and Contributor Identifier for authorization. We're now experimenting with Machine Learning and Natural Language Processing to help automate population of the graph, starting with the classification of research articles by topic. Finding adequate training data in the absence of a comprehensive and open research article API continues to be a significant challenge.