https://ntrs.nasa.gov/search.jsp?R=20160005299 2019-08-31T03:22:52+00:00Z

MINING DARK INFORMATION RESOURCES TO DEVELOP NEW INFORMATICS CAPABILITIES TO SUPPORT SCIENCE

Rahul Ramachandran NASA/MSFC, **Manil Maskey** UAH, Kaylin Bugbee UAH

Project Team:

MSFC/UAH: Patrick Gatlin, Xiang Li, Amanda Weigel, JJ Miller, Kaylin Bugbee, Ajinkya Kulkarni

GSFC: Steve Kempler, Suhung Shen, Chung-Lin Shie, Maksym Petrenko

RPI: Peter Fox, Stefan Zednik, Anirudh Prabhu

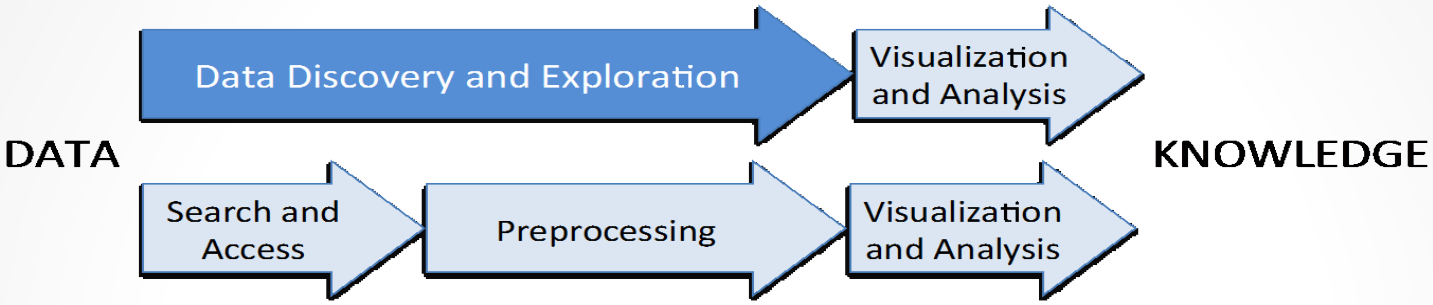
European Geosciences Union General Assembly 2016
Vienna, Austria 17–22 April 2016

Outline

1. Project Overview
2. Data Curation Service
3. Rules Engine
4. Image Retrieval Service
5. Summary

Part 1: Project Overview

Motivation



- Data preparation steps are **cumbersome** and **time consuming**
  - Covers discovery, access and preprocessing
- Limitations of current Data/Information Systems
  - **Boolean search** on data based on instrument or geophysical or other **keywords**
  - Underlying **assumption** that users have sufficient knowledge of the **domain vocabulary**
  - **Lack support** for those **unfamiliar** with the domain vocabulary or the **breadth of relevant data** available

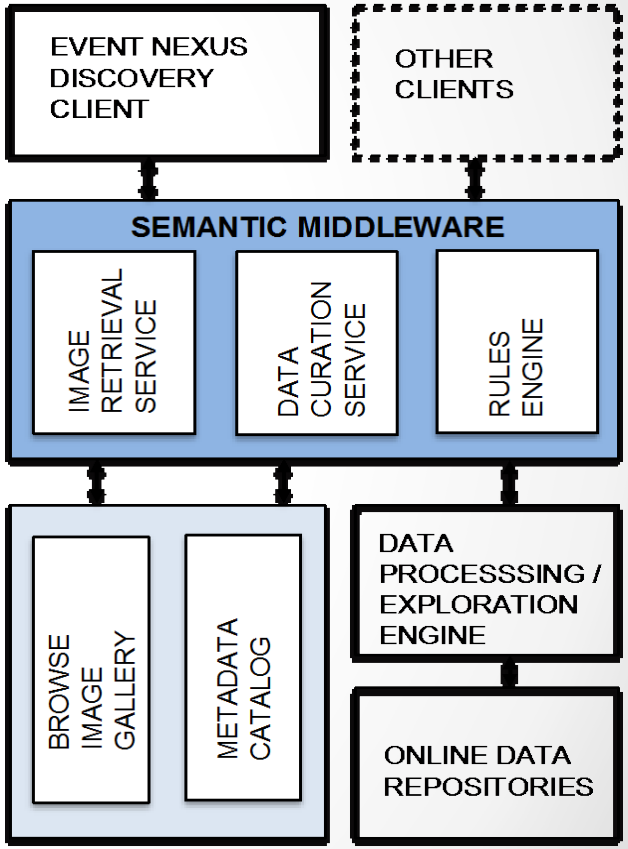
Earth Science Metadata: Dark Resources

- *Dark resources* - information resources that organizations collect, process, and store for regular business or operational activities but fail to utilize for ***other*** purposes
  - Challenge is to recognize, identify and effectively utilize these dark data stores
- Metadata catalogs contain dark resources consisting of structured information, free form descriptions of data and browse images.
  - NASA's Common Metadata Repository (ECHO) holds >6000data collections, 270 million records for individual files and 67 million browse images.

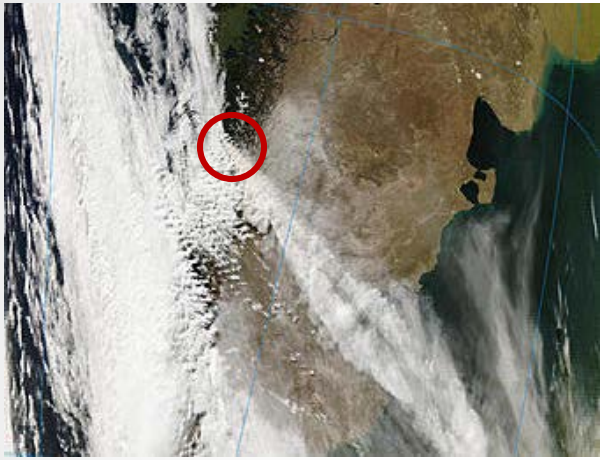
Premise: Metadata catalogs can be utilized *beyond their original design intent* to provide *new data discovery and exploration pathways* to support science and education communities.

Project Goals

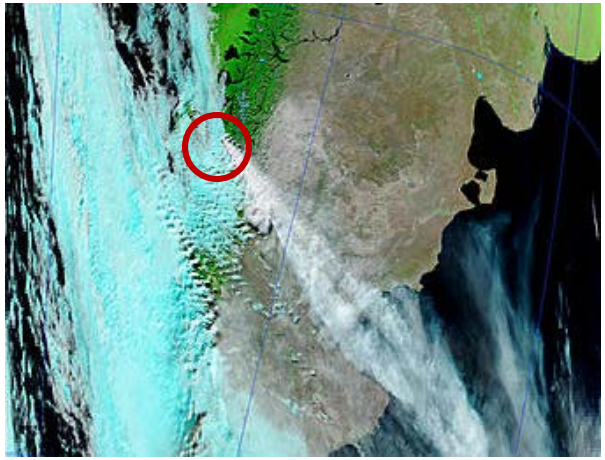
- Design a Semantic Middleware Layer (SML) to exploit these metadata resources
  - provide novel ***data discovery and exploration*** capabilities that significantly reduce data preparation time.
  - utilize a varied set of semantic web, information retrieval and image mining technologies.
- Design SML as a Service Oriented Architecture (SOA) to allow individual components to be used by existing systems



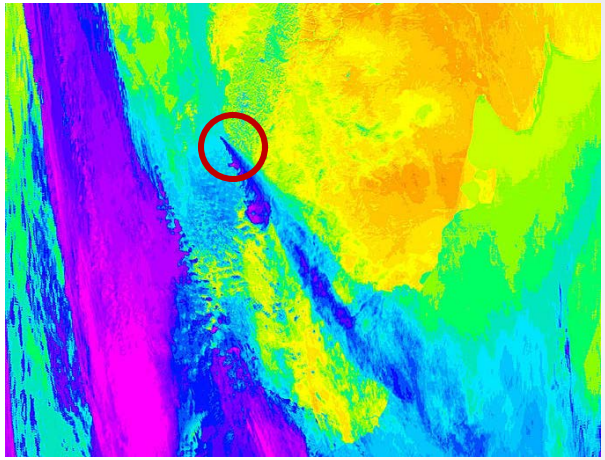
Use Case:
Find Interesting Events from Browse Images



Band 1-4-3 (true color)



Band 7-2-1



LST

Example: MODIS-Aqua 2008-05-03 18:45 UTC

Chaitén Volcano Eruption
Eruption Time period: May 2 – Nov 2008
Location: Andes region, Chile (-42.832778, -72.645833)

Image Retrieval Service can be used to find volcanic ash events in browse imagery

Suggest Relevant Data

Total SO_2 mass:
e.g. **Chaitén** is 10 (kt) =(kilotons) , (1kt= 1000 metric tons)
ftp://measures.gsfc.nasa.gov/data/s4pa/SO2/MSVOLSO2L4.1/MSVOLSO2L4_v01-00-2014m1002.txt

Daily SO2:
OMI/Aura Sulphur Dioxide (SO2) Total Column Daily L2 Global 0.125 deg
http://disc.sci.gsfc.nasa.gov/datacollection/OMSO2G_V003.html

Calibrated Radiances:
MODIS/Aqua Calibrated Radiances 5-Min L1B Swath 1km
http://dx.doi.org/10.5067/modis/myd021km.006

Aerosol Optical Thickness:
MODIS/Aqua Aerosol 5-Min L2 Swath 10km
http://modis-atmos.gsfc.nasa.gov/MOD0
SeaWiFS Deep Blue Aerosol Optical Dept
Data 13.5km
http://disc.gsfc.nasa.gov/datacollection

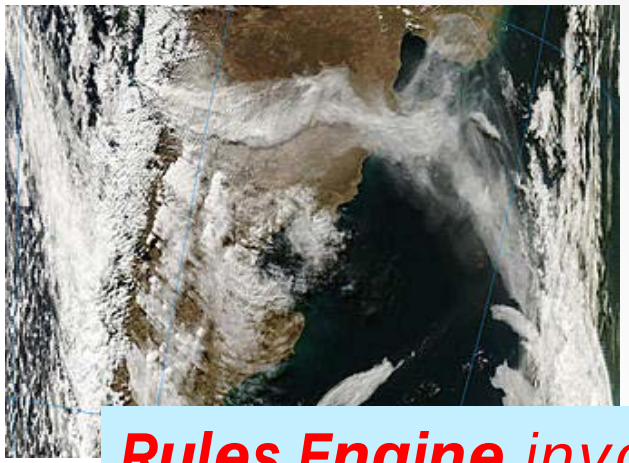
Data Curation Service *recommends* ***relevant datasets*** *to support event analysis*

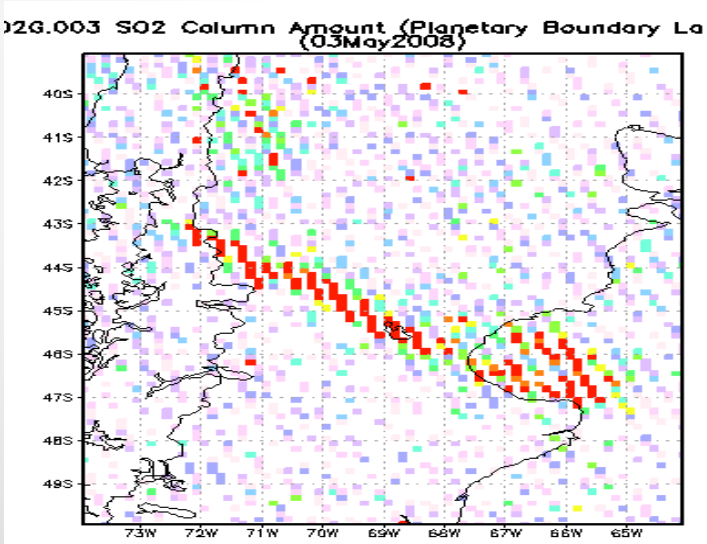
IR Brightness Temperature:
NCEP/CPC 4-km Global (60 deg N - 60 deg S) Merged IR Brightness Temperature Dataset

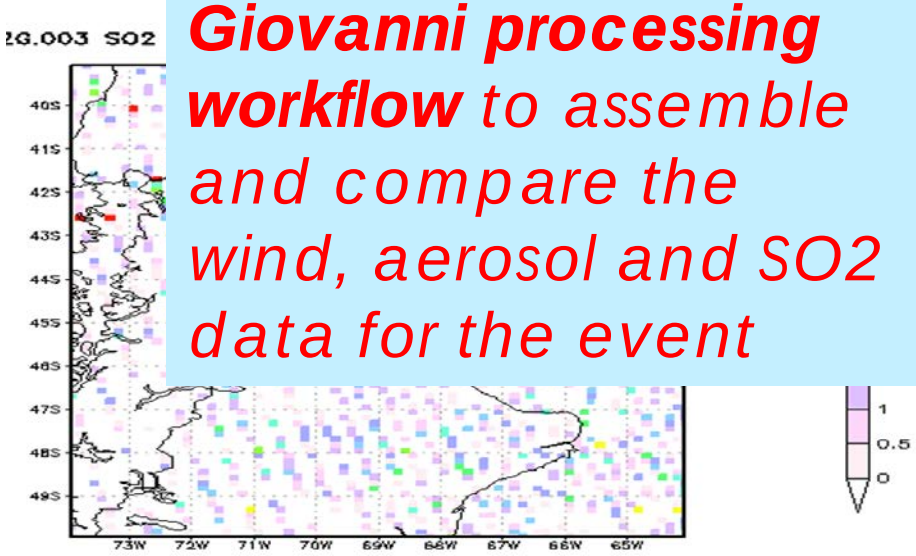
Generate Giovanni SO2 Plots

MODIS-Aqua 2008-05-03 18:45 UTC

MODIS-Aqua 2008-05-05 18:30 UTC





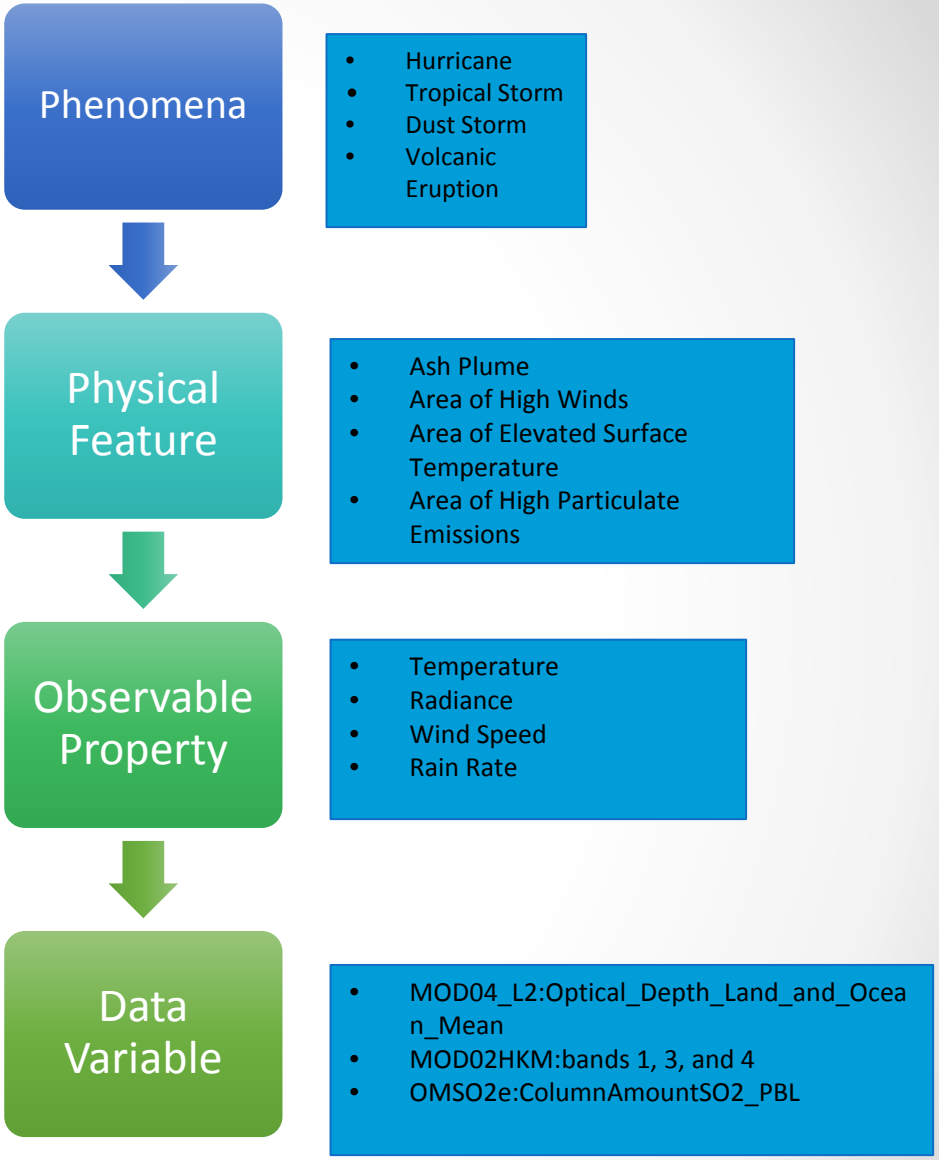


Rules Engine *invokes a* ***Giovanni processing workflow*** *to assemble and compare the wind, aerosol and SO2 data for the event*

http://gdata2.sci.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=omil2g

Conceptual Model

- Phenomena
  - Event type
- Physical Feature
  - Manifestation / Driver of phenomena
  - Has space/time extent
  - Can precede or linger after what is generally thought of as the phenomena event
- Observable Property
  - Characteristic/property of physical feature
- Data Variable
  - Measurement/estimation of observable feature



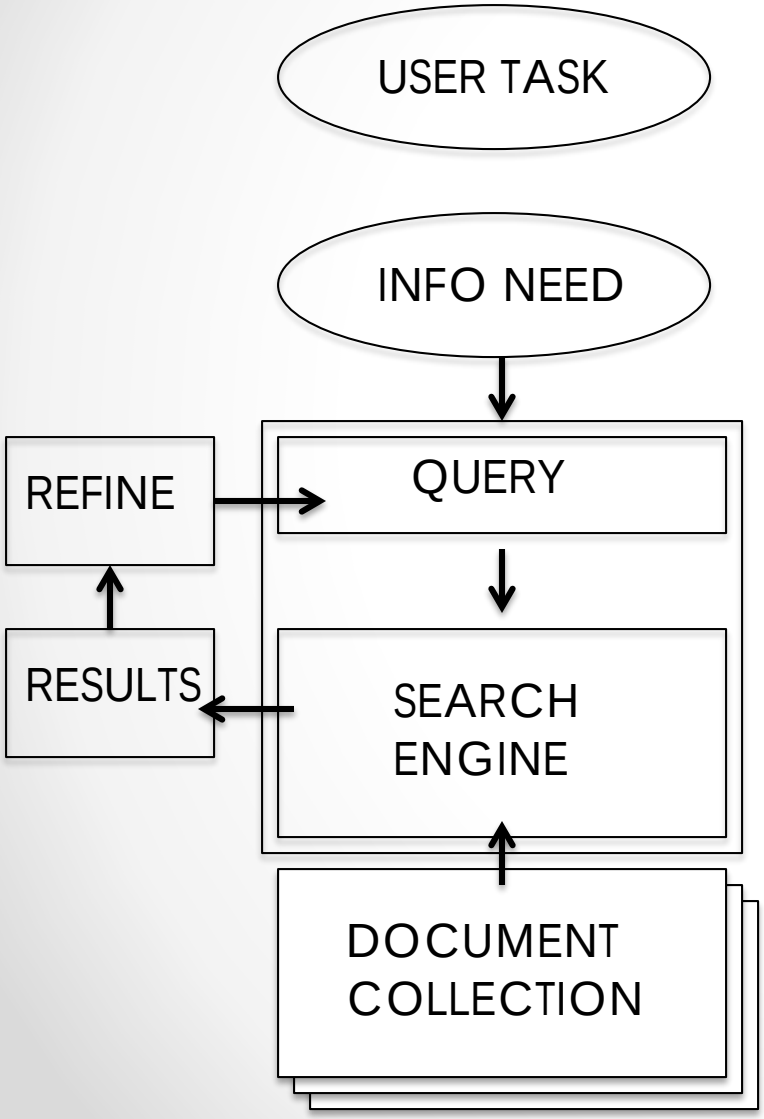
Part 2: Data Curation Algorithm for Phenomena

Objectives

- Design a data curation (relevancy ranking) algorithm for a set of **phenomena**
- Provide the data curation algorithm as a stand alone service

- Envisioned Use:
  - Given a phenomenon type (Ex: Hurricane), DCS returns a list of relevant data sets (variables)
    - <list of data sets (variables)> = DCS(Phenomenon Type)
  - For a specific phenomenon instance (event: Hurricane Katrina), these curated datasets can be filtered based on space/time to get actual granules

Data Curation is a Specialized Search Problem



Study "Hurricane"

All data sets useful in studying "Hurricane"

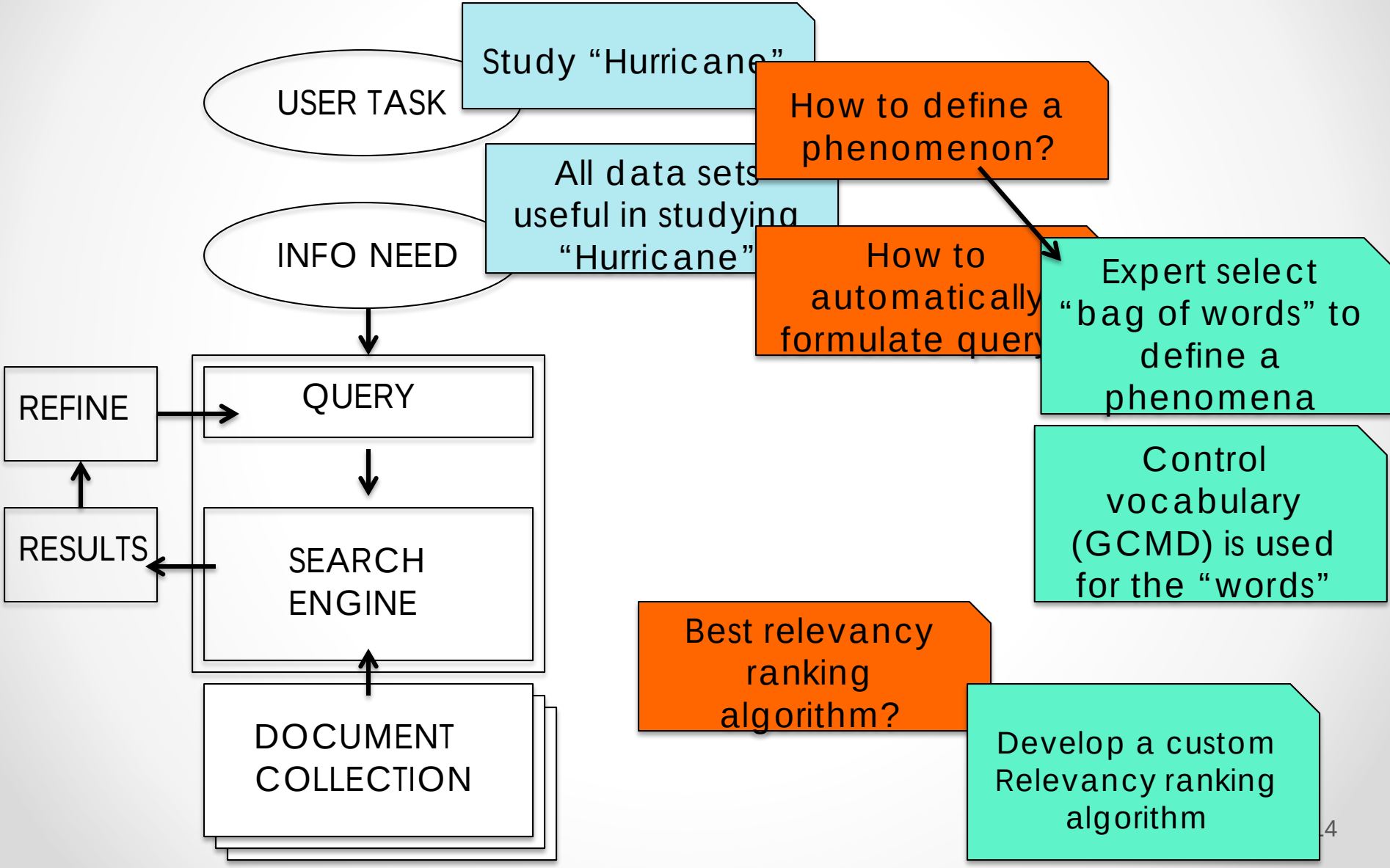
By pass this step for the end user

How to define a phenomenon?

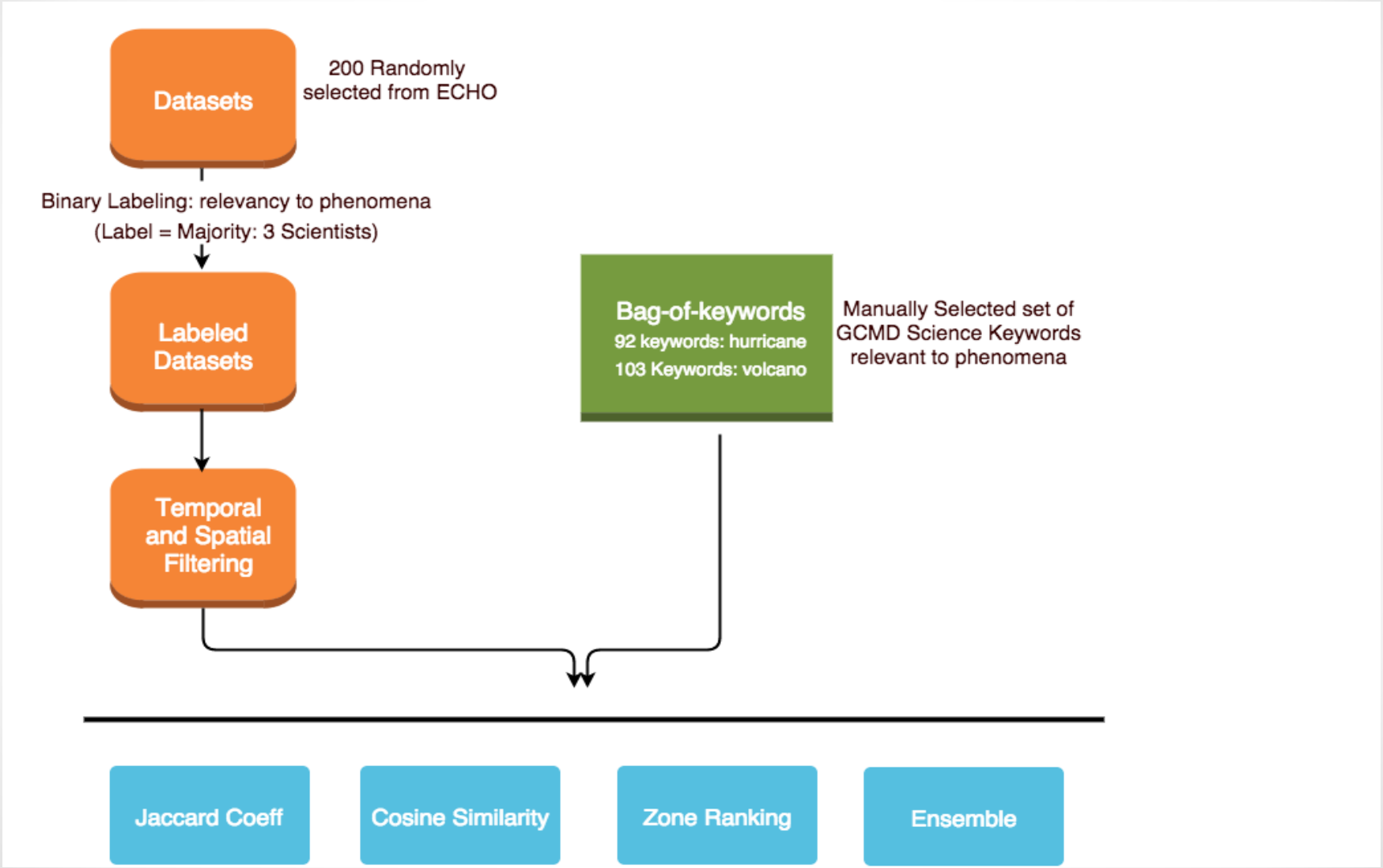
How to automatically formulate query?

Best relevancy ranking algorithm?

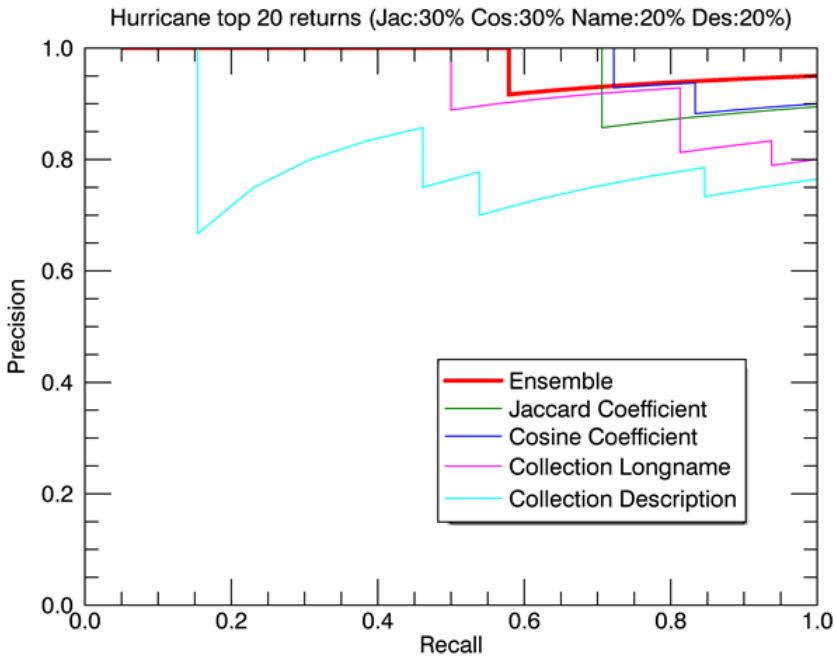
Our Approach

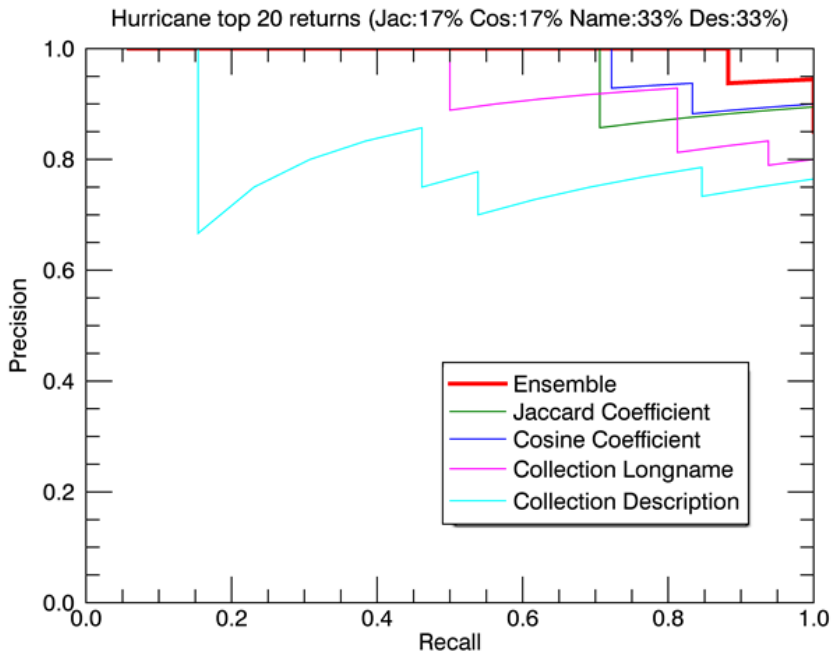


Experiment Setup



Top 20 returns (Hurricane)

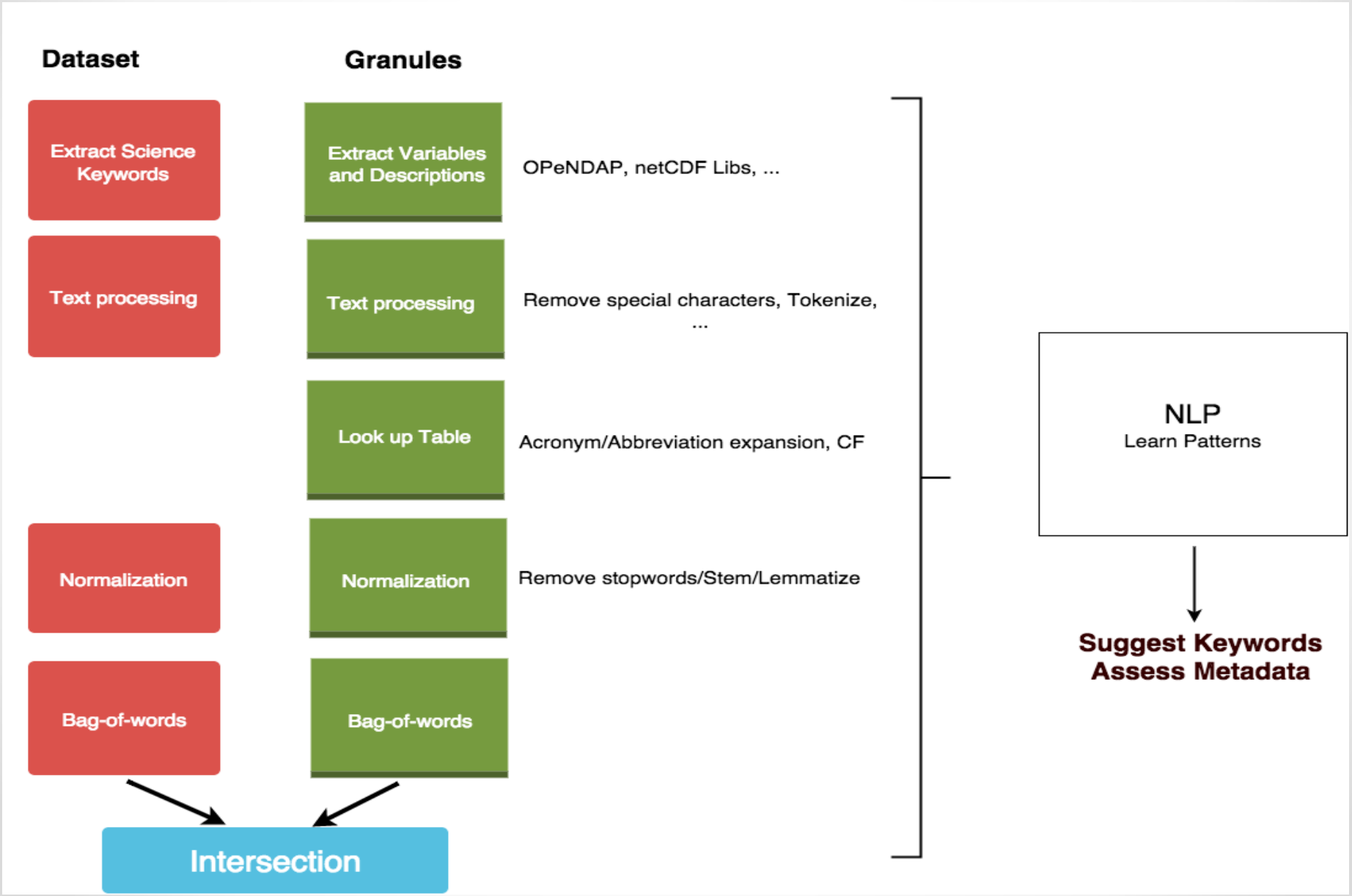




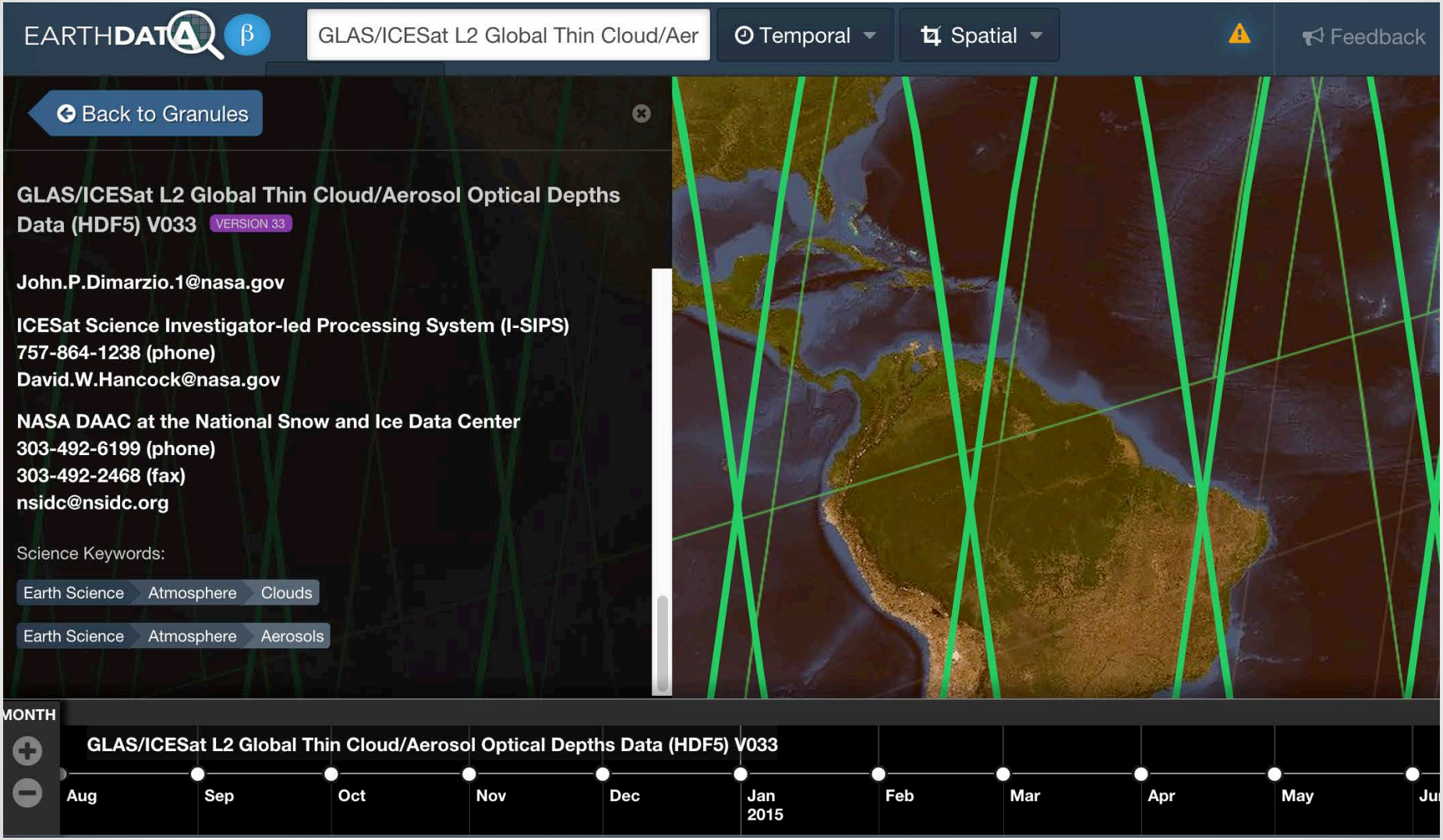
Next: Find relevant data fields

- Need actual data variables
  - Example: Giovanni uses these fields for visualization
- What we know
  - Data set (Collection) level science keywords (GCMD) – Experts
  - Granule data fields and metadata – Auto extract*
- How do we map?
  - Start with GCMD to CF Standard name
  - Most don't follow CF Standard names

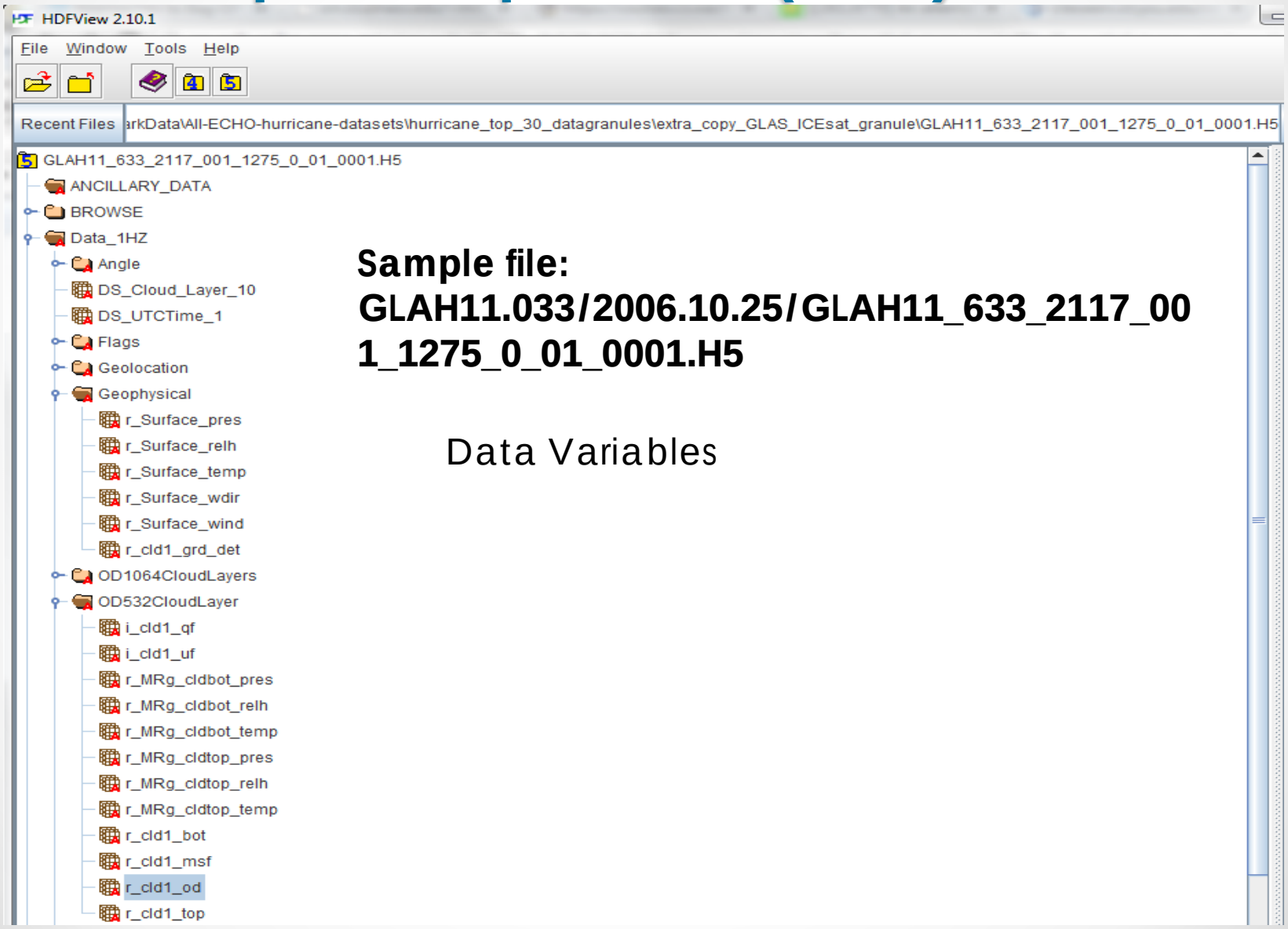
Approach


Dataset
Granules
Extract Science Keywords
Extract Variables and Descriptions
OPeNDAP, netCDF Libs, ...
Text processing
Text processing
Remove special characters, Tokenize, ...
Look up Table
Acronym/Abbreviation expansion, CF
Normalization
Normalization
Remove stopwords/Stem/Lemmatize
Bag-of-words
Bag-of-words
Intersection
NLP
Learn Patterns
Suggest Keywords
Assess Metadata

Example: GLAS/ICESat L2 Global Thin Cloud/Aerosol Optical Depths Data (HDF5) V033 – Dataset Metadata



Example: GLAS/ICESat L2 Global Thin Cloud/Aerosol Optical Depths Data (HDF5) V033



Sample file: GLAH11.033/2006.10.25/GLAH11_633_2117_001_1275_0_01_0001.H5

Data Variables

Example: GLASICESat L2 Global Thin Cloud Aerosol Optical Depths Data (HDF5) V033

Science keyword to variable mapping

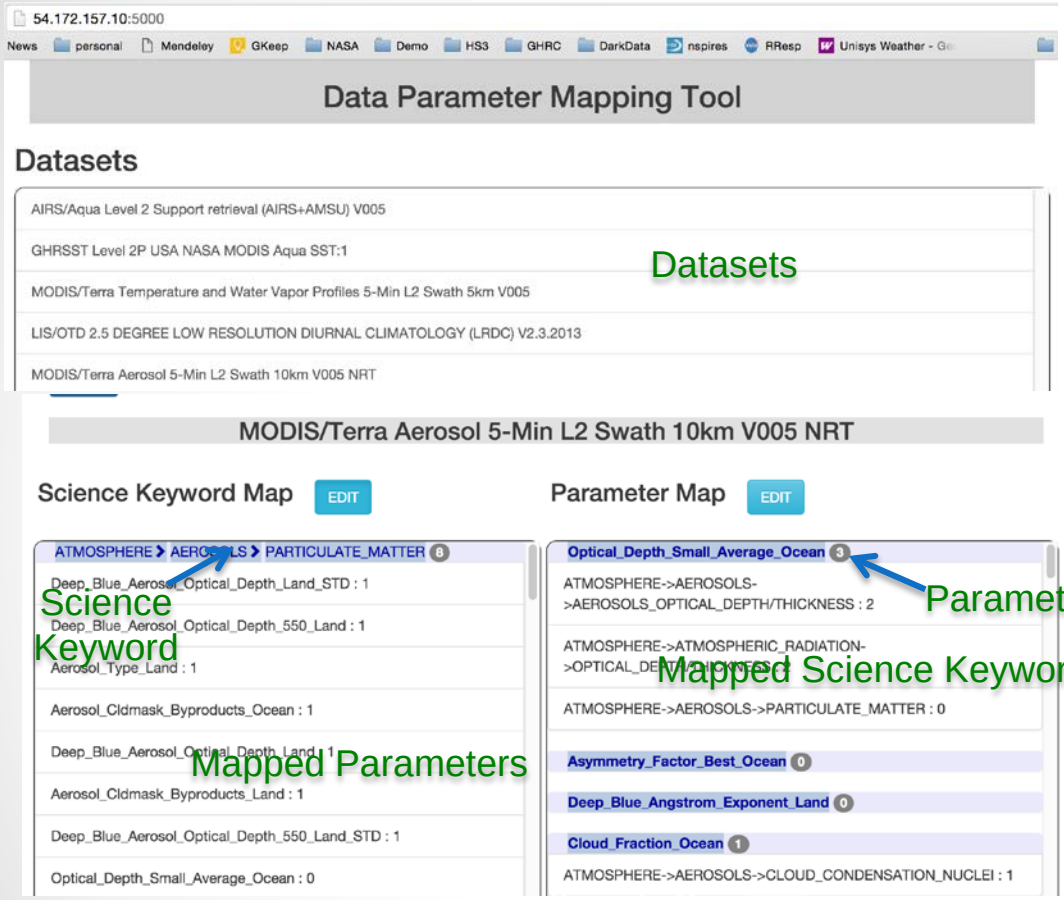
- r_Surface_relh | Surface Relative Humidity
  - No match
- r_Surface_temp | Surface Temperature
  - No match
- r_Surface_wind | Surface Wind Speed
  - No match
- r_cld1_od | Cloud Optical Depth at 532 nm
  - Score=3 keyword: ATMOSPHERE->CLOUDS->CLOUD OPTICAL DEPTH/THICKNESS
  - Score=2 keyword: ATMOSPHERE->AEROSOLS->AEROSOL OPTICAL DEPTH/THICKNESS

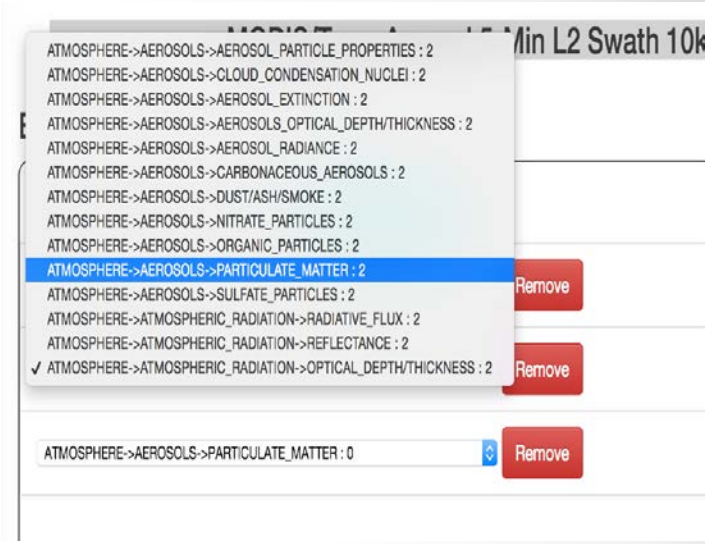
Variable to keyword mapping

- ATMOSPHERE->CLOUDS->CLOUD OPTICAL DEPTH/THICKNESS
  - Score=3 name: r_cld_ir_OD | Cloud Optical Depth at 1064 nm
  - score=3 name:i_cld1_qf | Cloud optical depth flag for 532 nm
  - Score=3 name:i_cld1_uf | Cloud optical depth flag for 532 nm
  - Score=3 name:r_cld1_od | Cloud Optical Depth at 532 nm

- *Found incorrect/incomplete keyword annotation*
- *Can be used **assess metadata quality** and **suggest keyword annotation!!***

Parameter Mapping Tool



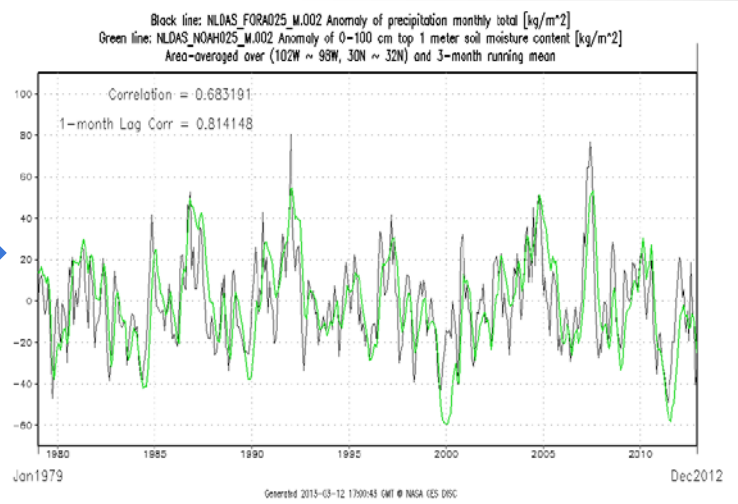


Edit/Save Mapping

Mapping Scores Generated by Algorithm

Part 3: Rules Engine

What settings should I use to visualize this event?



Data Variable ?

Dataset ?

Visualization Type?

Goal: Automate data preprocessing and exploratory analysis and visualization tasks

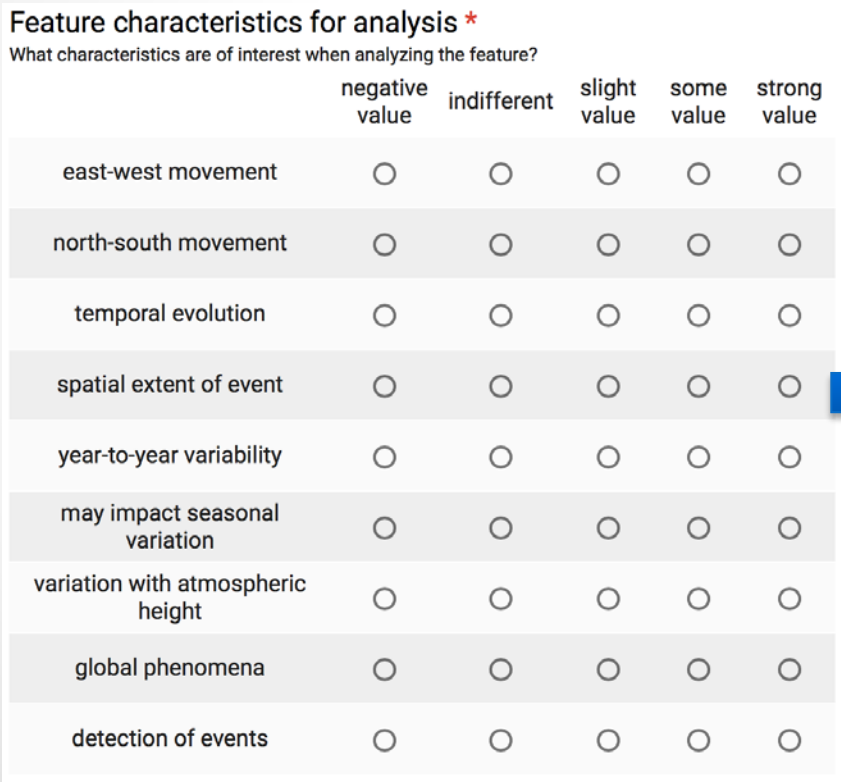
Images from : http://globe-views.com/dcim/dreams/volcano/volcano-03.jpg , http://grecaira.users37.interdns.co.uk/essay/images/confused.png , http://disc.sci.gsfc.nasa.gov/datareleases/images/nldas_monthly_climatology_figure_9.gif

Strategy

- Service to generate and rank candidate workflow configurations
- Use rules to make **assertions** about **compatibility based on multiple factors**
  - does this data variable make sense for this feature?
  - does this visualization type make sense for this feature?
  - does the temporal / spatial resolution of this dataset make sense for this feature?
- Each compatibility assertion type is assigned weights.
  - ex: Strong = 5, Some = 3, Slight = 1, Indifferent = 0, Negative = -1.
- Based on the aggregated compatibility assertions, we calculate the score for each visualization candidate.

Ruleset Development

Survey asked users to rate characteristics of phenomena features



Feature characteristics for analysis *

What characteristics are of interest when analyzing the feature?

	negative value	indifferent	slight value	some value	strong value
east-west movement	○	○	○	○	○
north-south movement	○	○	○	○	○
temporal evolution	○	○	○	○	○
spatial extent of event	○	○	○	○	○
year-to-year variability	○	○	○	○	○
may impact seasonal variation	○	○	○	○	○
variation with atmospheric height	○	○	○	○	○
global phenomena	○	○	○	○	○
detection of events	○	○	○	○	○

Survey results used to formulate rules

```
[rule1:
  (?feature rdf:type
dd:AshPlume)
  ->
  (?feature
dd:strongCompatibilityFor
dd:temporal_evolution),
  (?feature
dd:indifferentCompatibilityFor
dd:east-west-movement),
  …
]
```

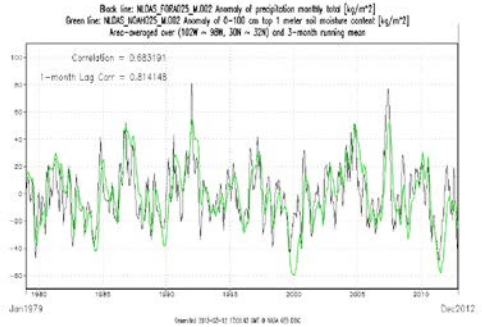
Phenomena Feature Characteristic Mappings

Phenomena	East-West Movement	North-South Movement	Temporal Evolution	Spatial Extent of Event	Year-to-Year Variability	May Impact Seasonal Variation	Variation with Atmospheric Height	Global Phenomena	Detection of Events
Volcano - Ash Plume	Indifferent	Indifferent	Strong	Slight	Strong	Strong	Strong	Strong	Strong
Flood	Some	Some	Strong	Some	Some	Strong	Some	Slight	Some
Dust Storm	Strong	Strong	Strong	Strong	Indifferent	Indifferent	Strong	Indifferent	Some

Service to Characteristic Mappings

Service	Visualization	East-West Movement	North-South Movement	Temporal Evolution	Spatial Extent of Event	Year-to-Year Variability	Seasonal Variation	Variation with Atmospheric Height	Global Phenomena	Detection of Events
Time-averaged Map	Color-Slice Map				✓					
Area-averaged Time Series	Time Series			✓						✓
User-defined Climatology	Color-Slice Map						✓			
Vertical Profile	Line Plot							✓		
Seasonal Time Series	Time Series					✓				
Zonal Means	Line Plot								✓	
Hovmoller (Longitude)	Color-Slice Grid	✓								
Hovmoller (Latitude)	Color-Slice Grid		✓							

Compute Compatibility



Phenomena: Volcano - Ash Plume

Service - Area Averaged Time Series

Temporal Evolution	Detection of Events
Strong	**Strong**

Area Averaged Time Series : bestFor →	Temporal evolution; Detection of events

STRONG COMPATIBILITY x2

Images from , http://disc.sci.gsfc.nasa.gov/datareleases/images/nldas_monthly_climatology_figure_9.gif, http://www.clipartbest.com/cliparts/biy/bAX/biybAXGiL.png

volcanic ash image - By Boaworm (Own work) [CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)], via Wikimedia Commons

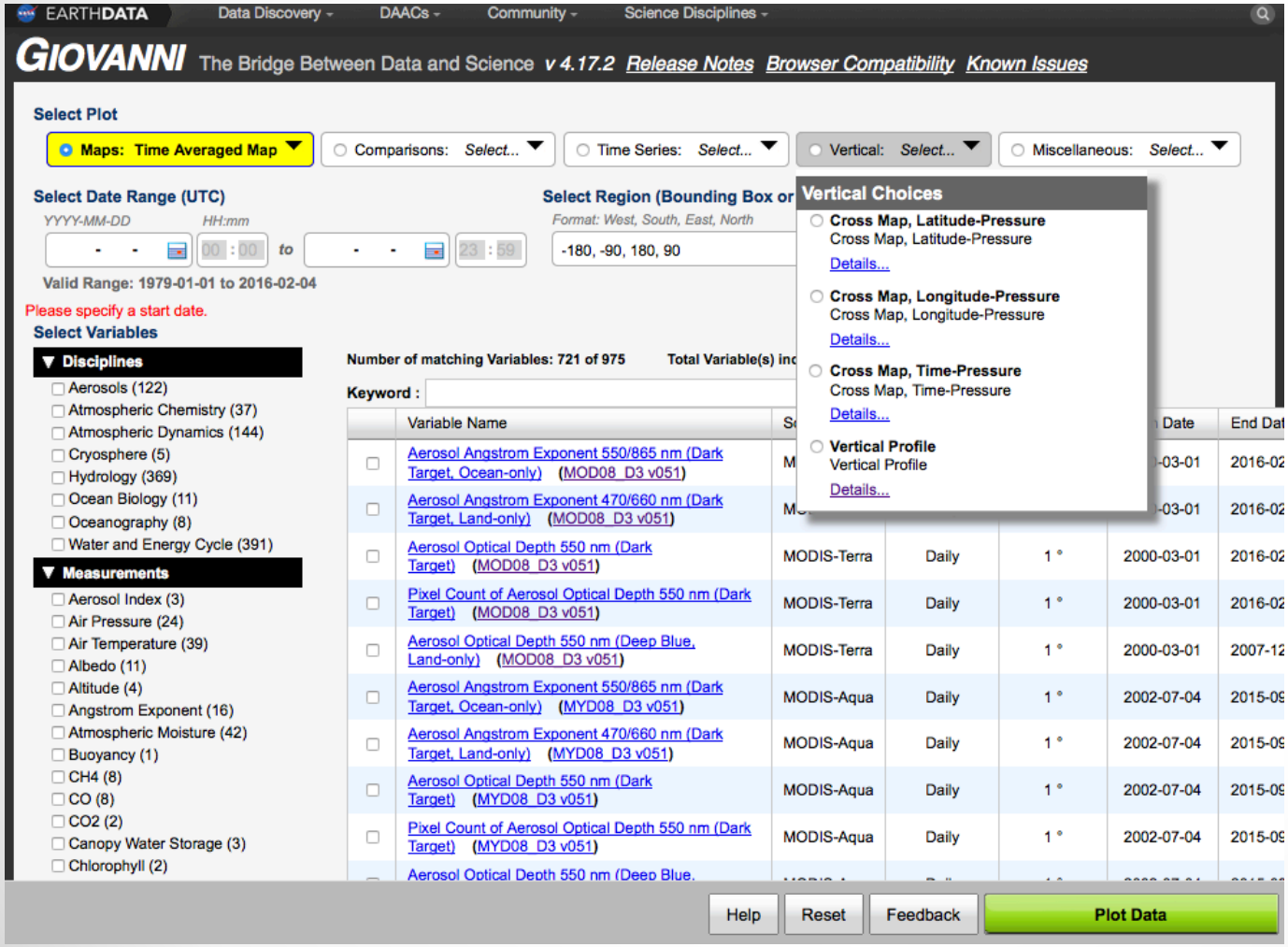
Integrating Services in Giovanni

- **Tool**: Giovanni is a popular on-line environment that lets users discover, plot, and download a number of geophysical parameters (data variables)

- **Goal**: Leverage Dark Data services and technologies to assist Giovanni users in discovering and exploring data

'Success will be realized when Giovanni requests can be automatically invoked with the appropriate spatial and temporal extents, variables and workflow / visualization type for a particular event'

Giovanni – Standard Edition



User needs to decide:

- Variable(s)
- Time
- Space
- Plot type

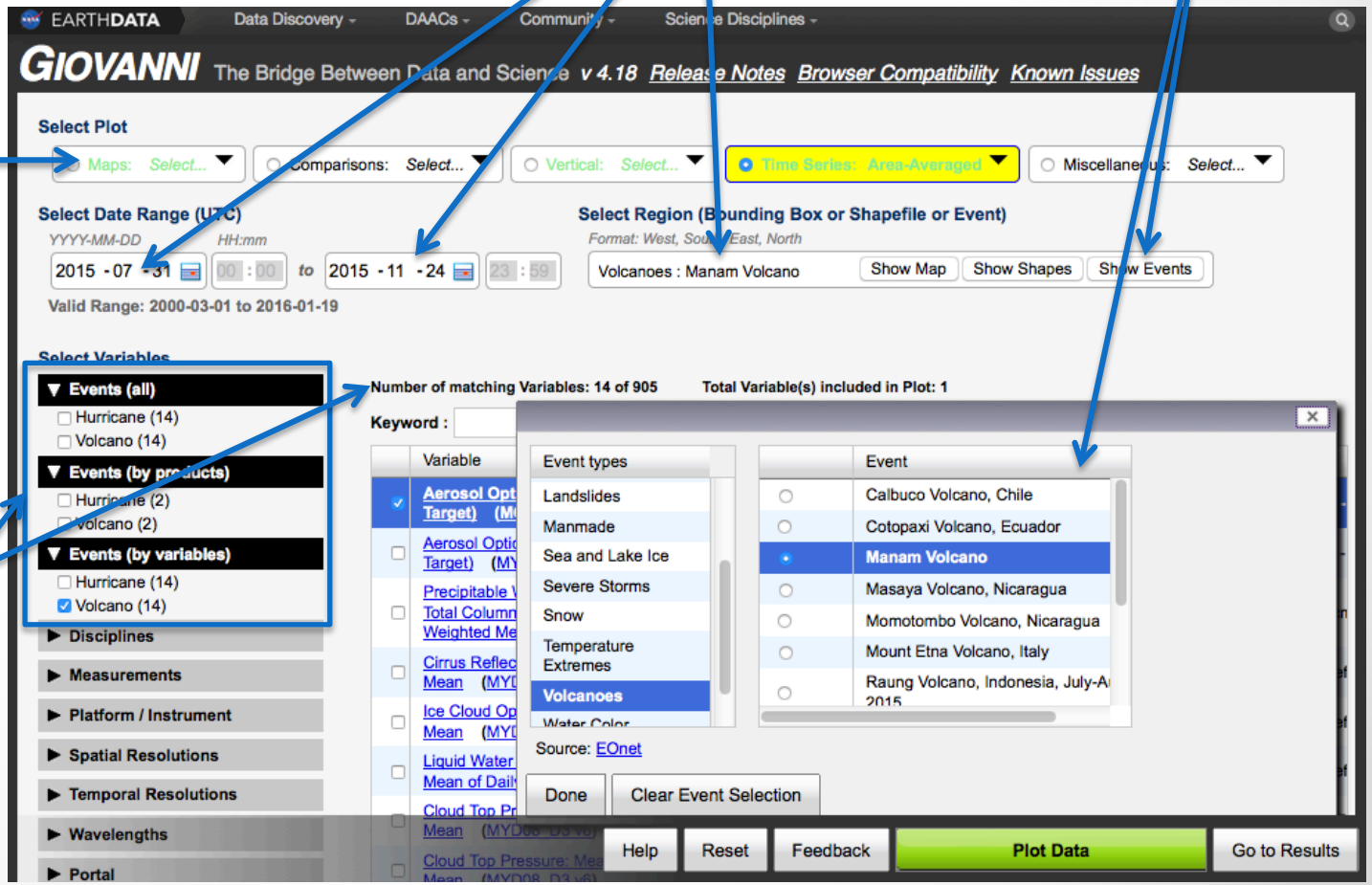
http://giovanni.sci.gsfc.nasa.gov/giovanni/

Giovanni – Dark Data Edition

Selected event & its time

Event Client

Rules Service: highlights suitable plots based on selected event & variables



Curation Service: event type filters relevant variables

Part 4: Image Retrieval

...

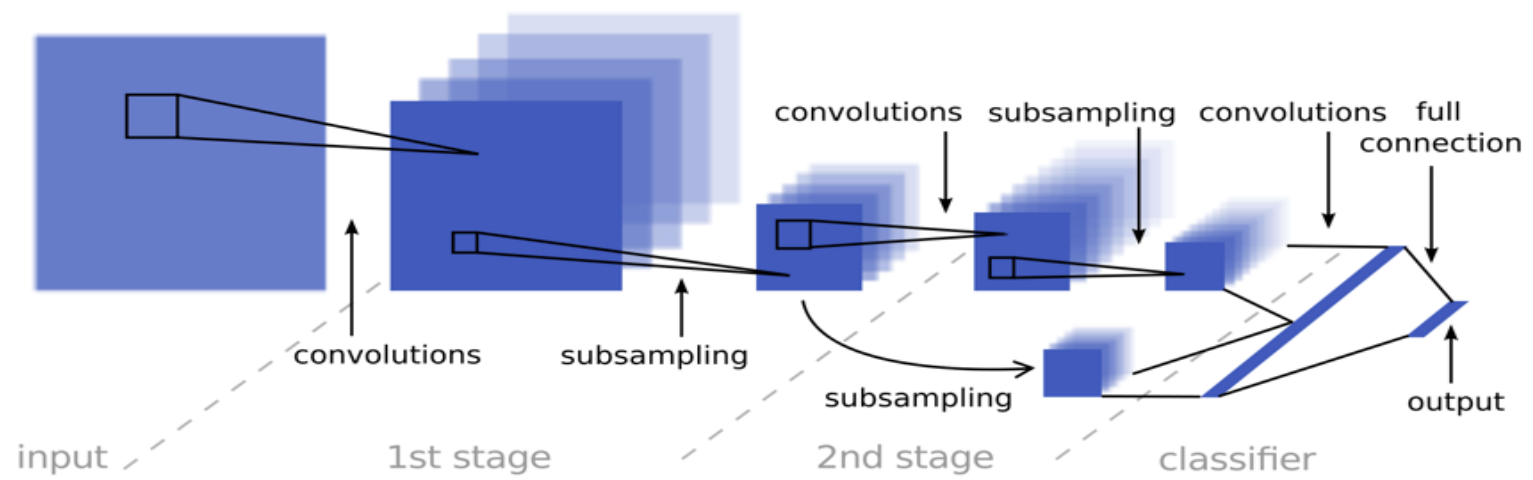
Image Retrieval

- Goal: given an image of Earth science phenomenon retrieve similar images

- Challenge: “semantic gap”
  - low-level image pixels and high-level semantic concepts perceived by humans

Deep Learning

- Mimics the human brain that is organized in a deep architecture
  - Processes information through multiple stages of transformation and representation
- Learns complex functions that directly map pixels to the output, without relying on *human-crafted features*



Convolution neural network

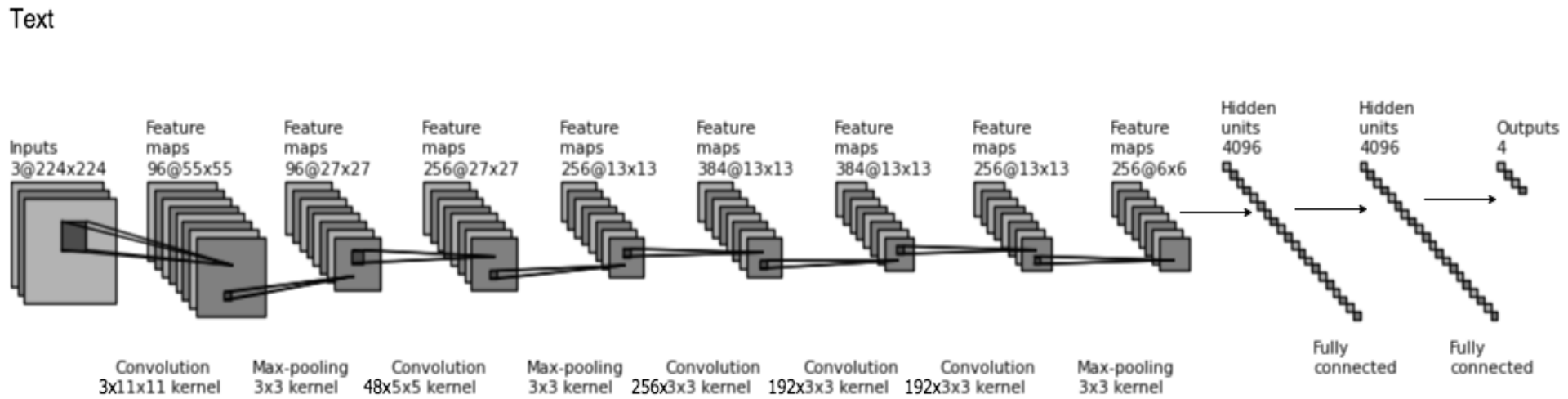
Source: Google Research, CVPR 2014

Transfer Learning

- CNN requires large number of parameters
- Learning parameters from *a few thousand training samples* is unrealistic
- Transfer learning
  - Use internal representation learned from one classification task to another
  - AlexNet architecture - Krizhevsky et. al.
  - Weights learned from ImageNet 1.3 million high-resolution images
  - State-of-the-art classification accuracy

A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp: 1097-1105, 2012.

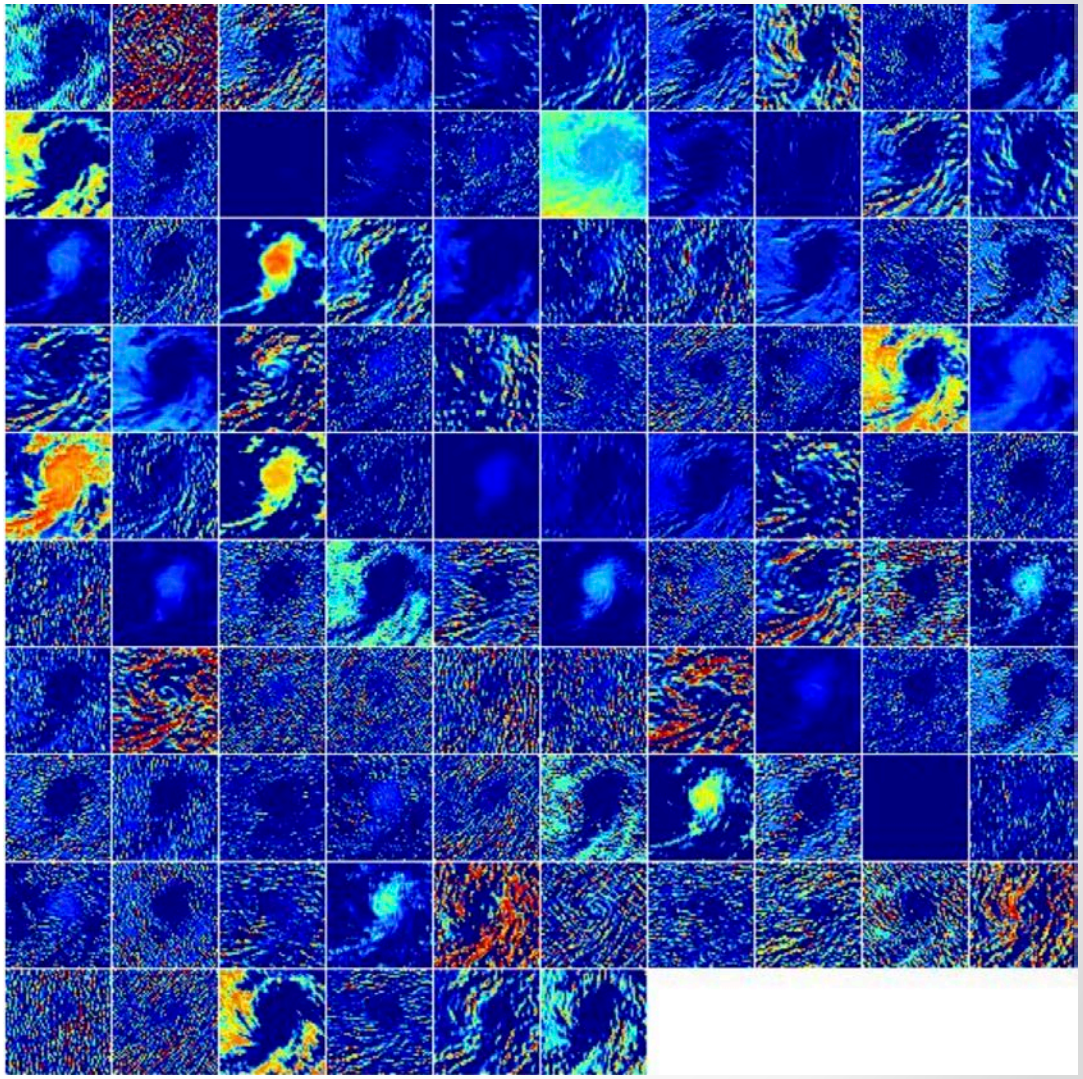
Experiment: CNN Configuration



- AlexNet architecture
  - o Initialized weights with ImageNet trained model
  - o Adaptive learning rate
  - o GPU implementation

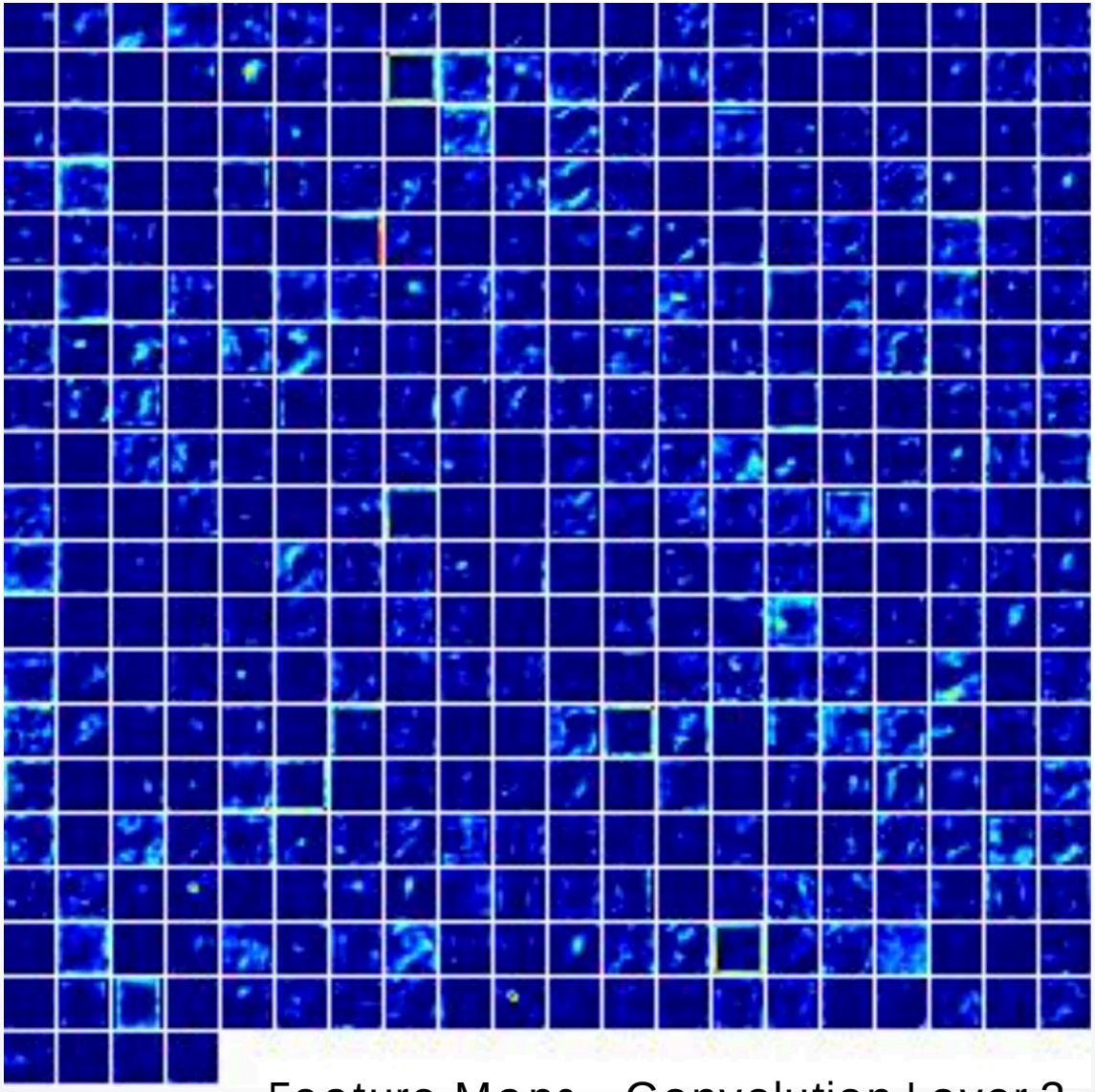
Experiment CNN – Visualization

Input Image



Feature Maps – Convolution Layer 1

Experiment CNN – Visualization



Feature Maps – Convolution Layer 3

Results: Confusion Matrix

MODIS Rapid Response Test Images (Images are New to Trained CNN)

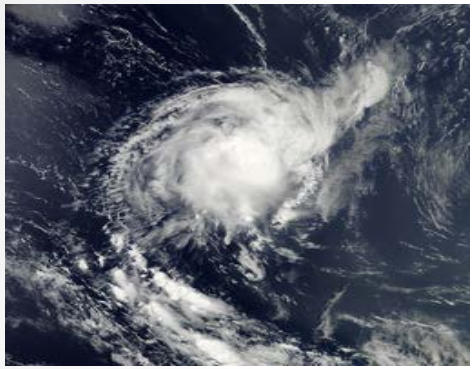
True/Pred	Dust	Hurricane	Smoke	Other
Dust	**287**	8	32	33
Hurricane	0	**379**	1	10
Smoke	12	12	**443**	9
Other	33	9	23	**211**

Overall Accuracy = **87.88**%

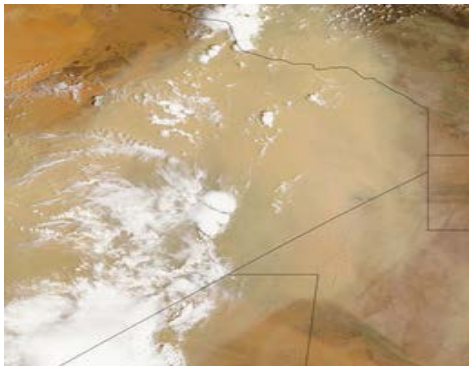
Producer's Accuracy
Dust 86.45%
Hurricane 92.89%
Smoke 88.78%
Other 80.23

User's Accuracy
Dust 79.72%
Hurricane 97.18%
Smoke 93.07%
Other 76.45%

Results (MODIS Rapid Response)

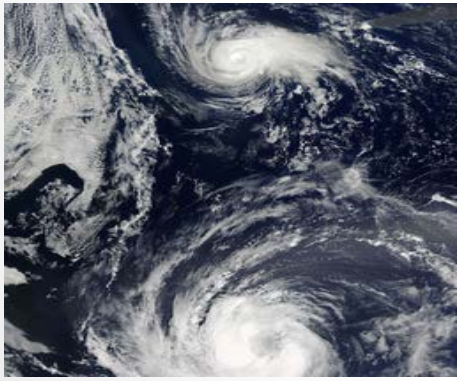


Hurricane – True Positive

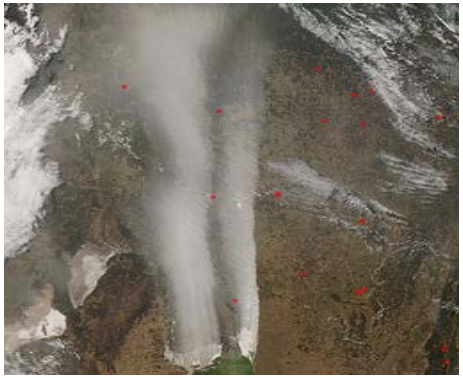


Dust – True Positive

Smoke– True Positive



Hurricane – False Negative



Dust – False Positive

Smoke– False Positive

Summary

- Building three specific semantic middleware components
  - *Image retrieval service*
  - *Data curation service*
  - *Semantic rules engine*

Infuse the entire middleware or the components into existing NASA data and information system

Dr. Rahul Ramachandran
Deputy Editor, Earth Science Informatics Journal
Manager, Global Hydrology Resource Center
NASA's Distributed Active Archive Center
Earth Science Office (ZP11)
NASA / Marshall Space Flight Center
Huntsville, Alabama 35812, USA
(w) 256.961.7620
(c) 256.226.6854
OrcId: orcid.org/0000-0002-0647-1941
www.linkedin.com/in/rramachandran05/