ILLUMINATING THE DARKNESS: EXPLOITING UNTAPPED DATA AND INFORMATION RESOURCES IN EARTH SCIENCE

Manil Maskey* UAH, Rahul Ramachandran NASA/MSFC, Peter Fox RPI, Steven Kempler NASA/GSFC

Project Team:

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

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

RPI: Stefan Zednik, Anirudh Prabhu

Earth Science Technology Forum, 2016

Earth Science Metadata: Dark Resources

- *Dark resources*: not used beyond **intended** purpose
  - Challenge: recognize, identify and effectively utilize for other **purposes**

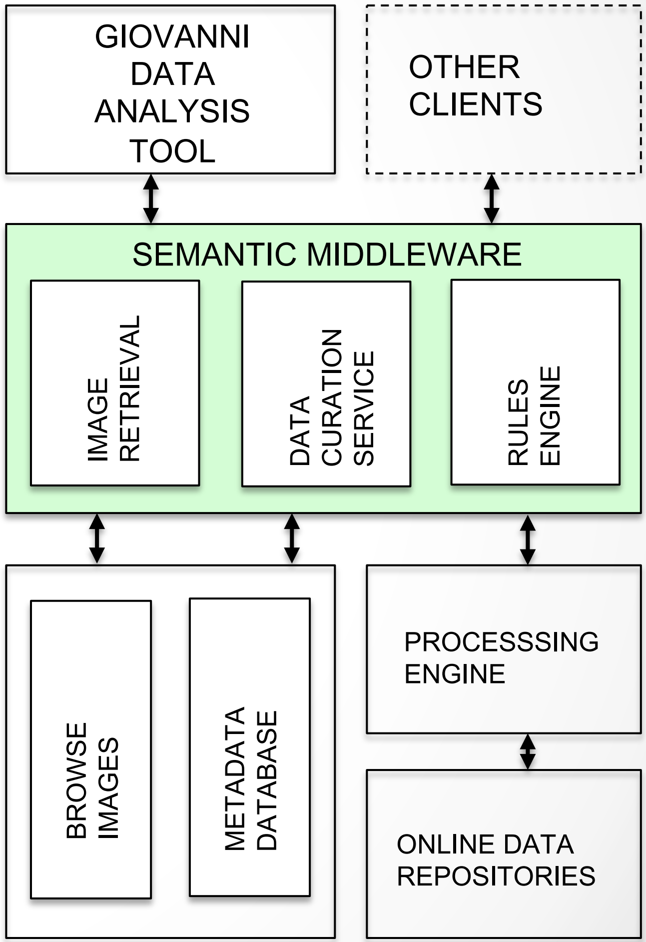
- Metadata catalogs:
  - contain dark resources
  - structured information
  - free form descriptions: data and browse images

- NASA's Common Metadata Repository
  - > 6000 data collections
  - 270 million records for individual files
  - 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 Earth science and education communities.

Project Goals

- Design a Semantic Middleware Layer (SML) to exploit 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.
    - automate
- Design SML as a Service Oriented Architecture

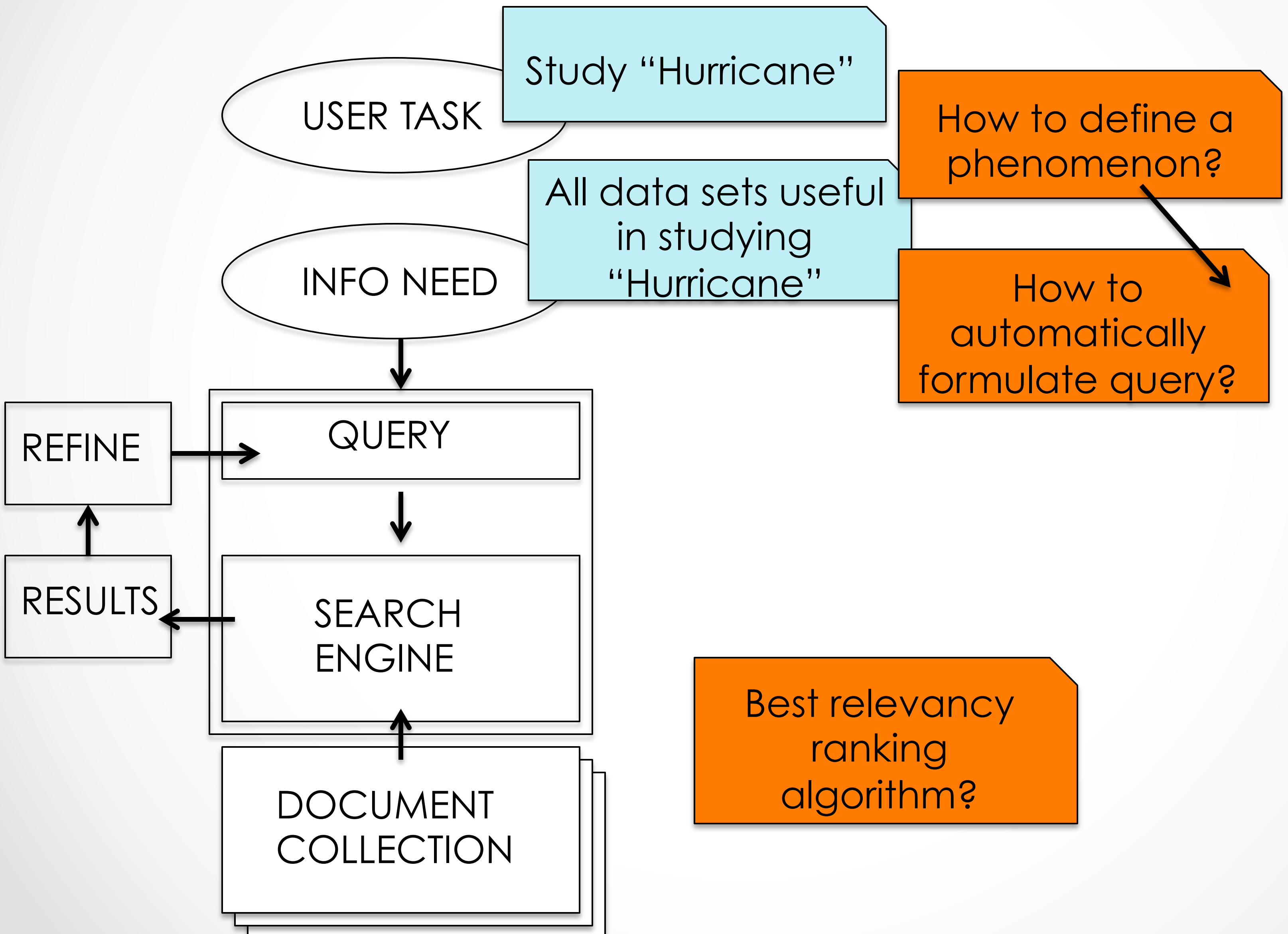


Data Curation Service

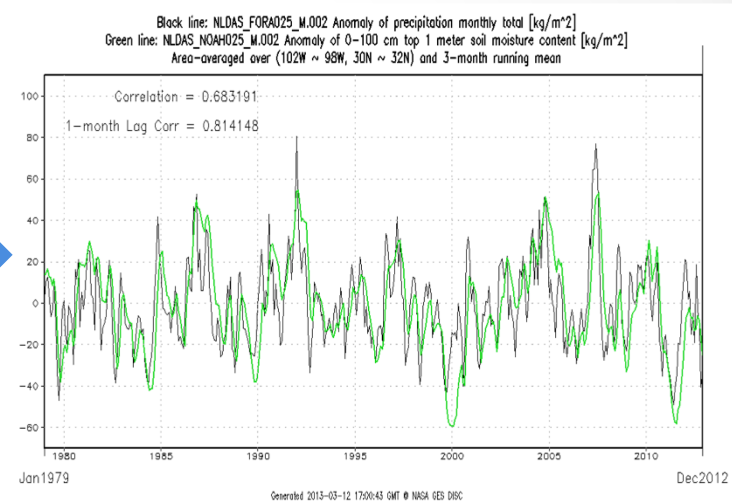
- Relevancy ranking algorithm for a set of **phenomena**
- Stand alone service

- 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 Approach



Rules Engine: What settings should I use to visualize this event?



Data Variable ?

Dataset ?

Visualization Type?

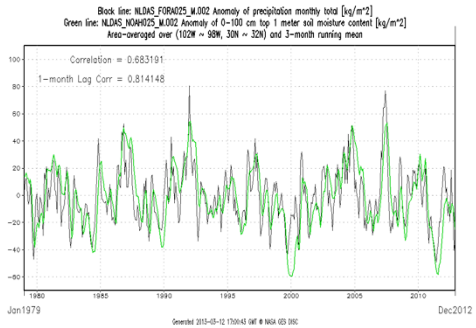
Goal: Automate data preprocessing and exploratory analysis and visualization tasks

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

Compute Compatibility

Use rules to make **assertions** about **compatibility based on multiple factors**

+



=

Phenomena: Volcano - Ash Plume

Service - Area Averaged Time Series

Temporal Evolution	Detection of Events
Strong	**Strong**

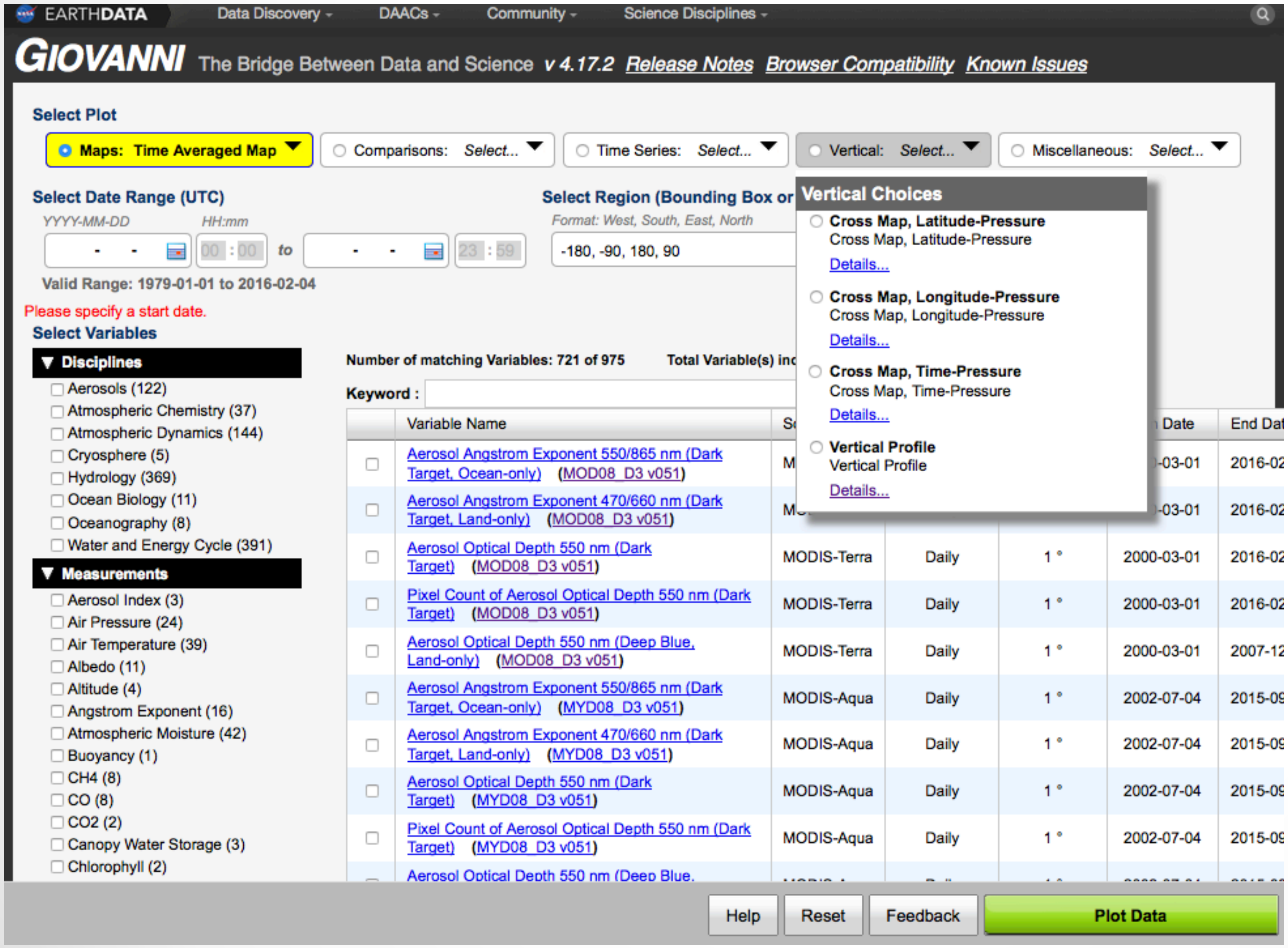
Area Averaged Time Series : bestFor →	Temporal evolution; Detection of events

STRONG COMPATIBILITY x2

Service to generate and rank candidate workflow configurations

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

Giovanni – Standard Edition



User needs to decide:

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

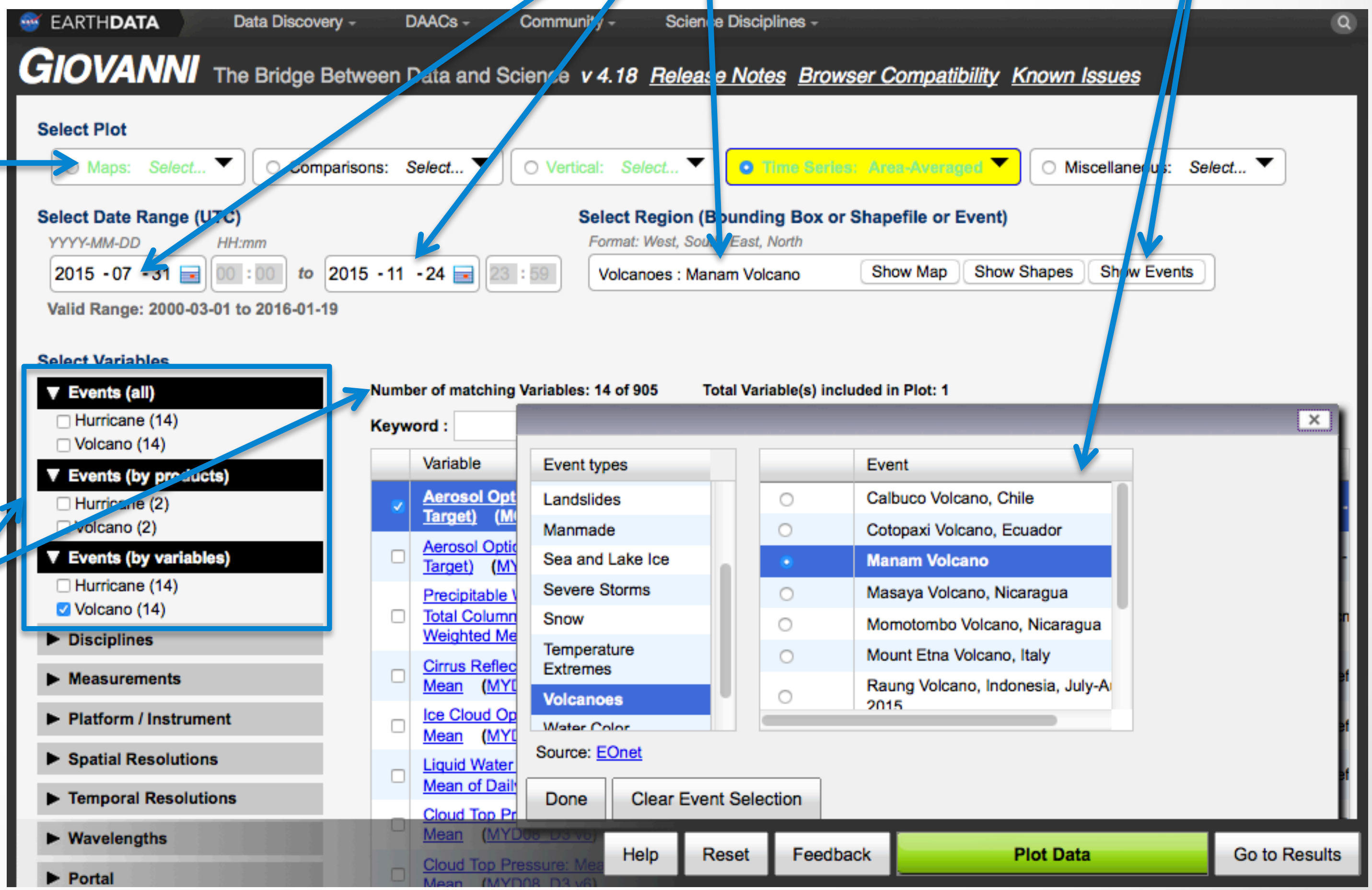
http://giovanni.sci.gsfc.nasa.gov/giovanni/

Giovanni – Dark Data Edition

Selected event & its time

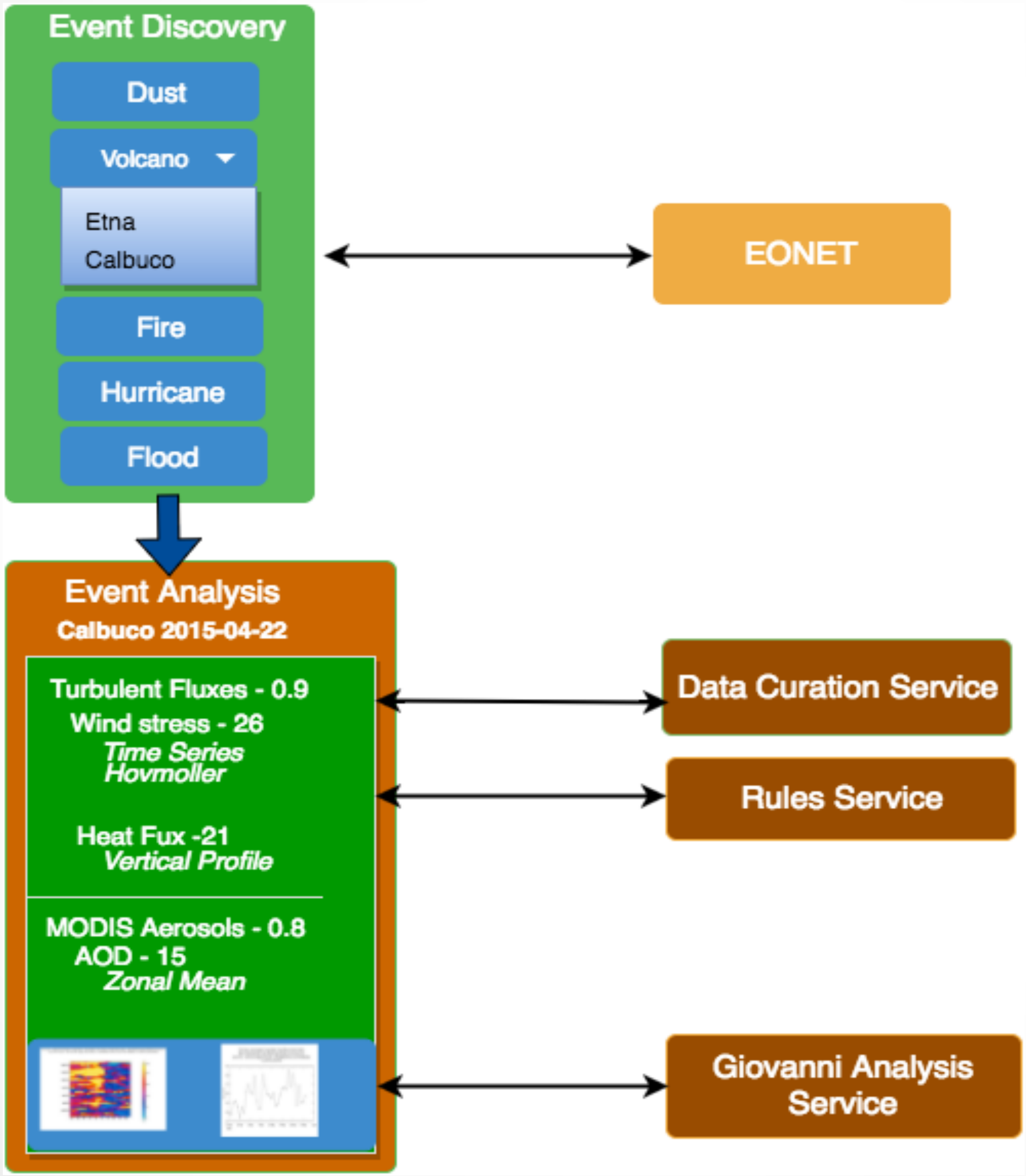
Event Client

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



Curation Service: event type filters relevant variables

Giovanni - Dark Data Edition



Event Analysis Workflow

Serendipitous Discovery

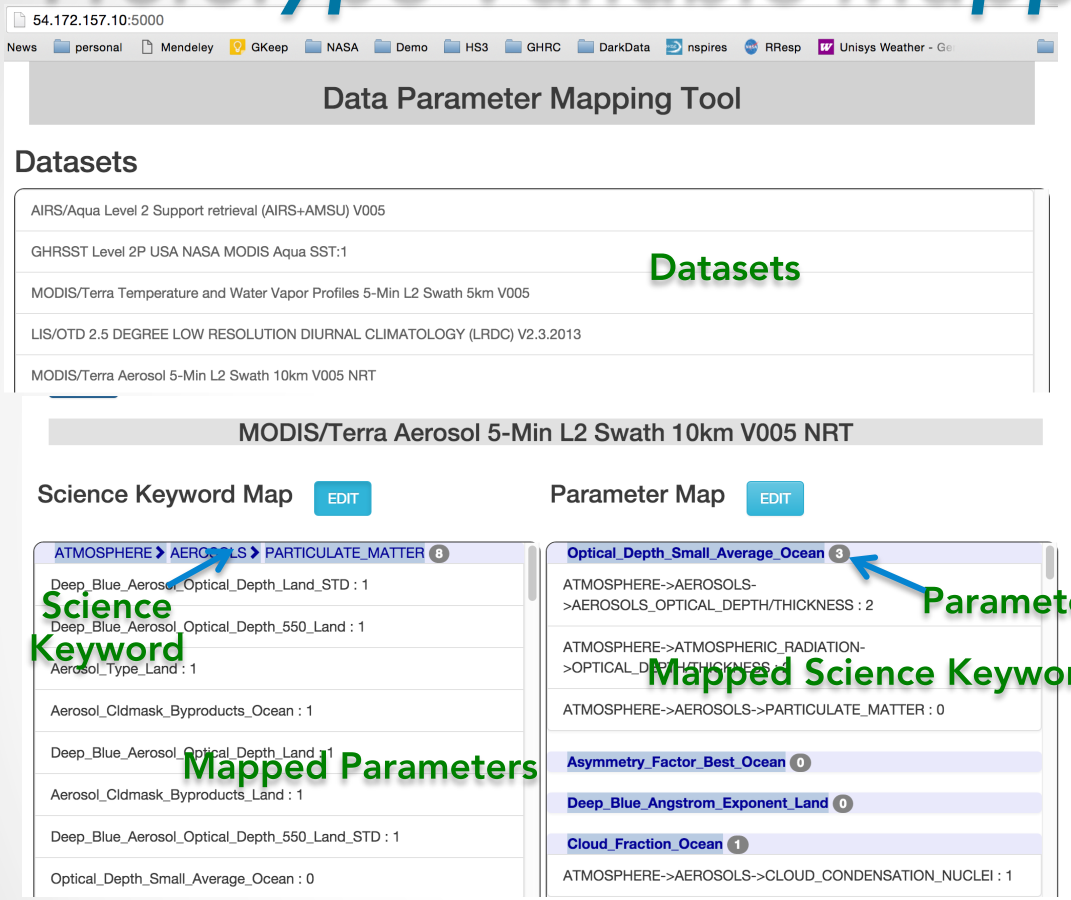
Data Curation Goal: map dataset keywords to granule variables

Application of Data Curation for Operational Use:

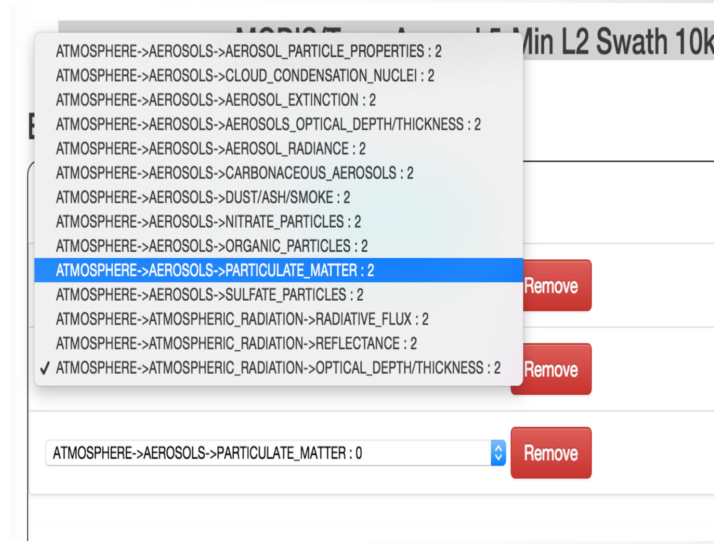
Data Curation Algorithm can be used to assess

- Metadata quality for both dataset and granules
    - Find incorrect/incomplete keyword annotations
- Automatically suggest science keywords

Operational Use:
Prototype Variable Mapping



Mapping Scores Generated by Algorithm



Edit/Save Mapping

Opportunity to develop this prototype and infuse into operational use at DAACs to improve metadata quality

DEMO

SML Components

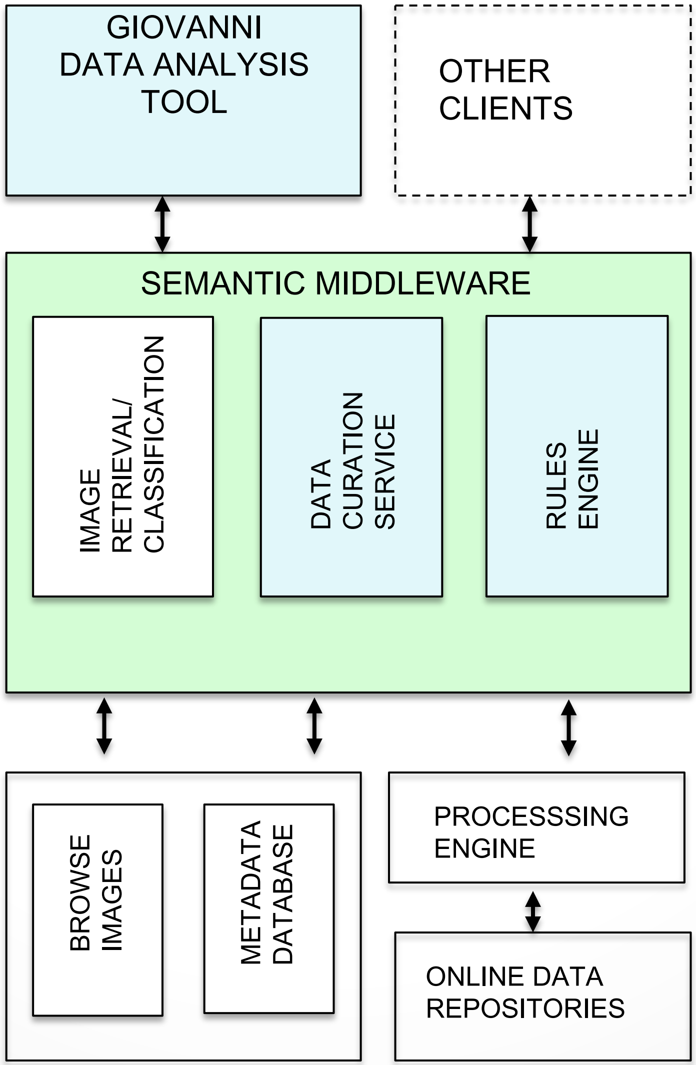


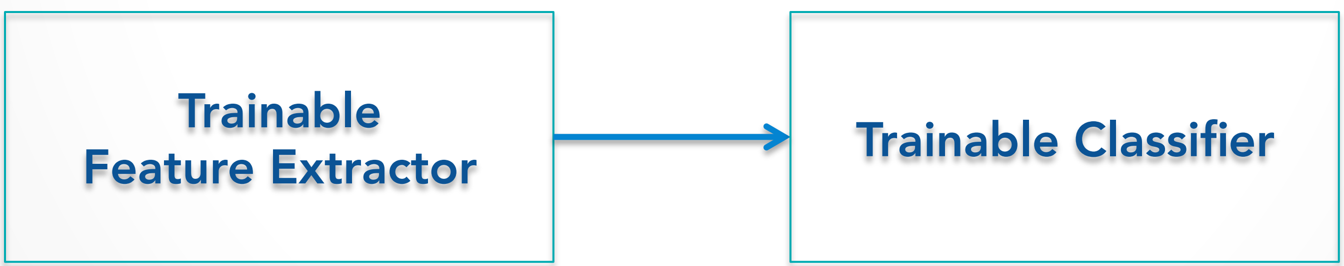
Image Classification

- Goal: label images in archives with a known Earth science phenomenon

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

"Deep" Architecture

- Features are key to recognition
- What about learning the features?
- Deep Learning
  - Hierarchical Learning
  - Mimics the human brain that is organized in a deep architecture
  - Uultiple stages of representation



Convolutional Neural Network (CNN)

- *Applicable to Images*
- *Supervised*

Transfer Learning

- CNN requires large number of parameters
- Transfer learning
  - Use internal representation learned from one classification task to another
- Faster learning
- Better accuracy

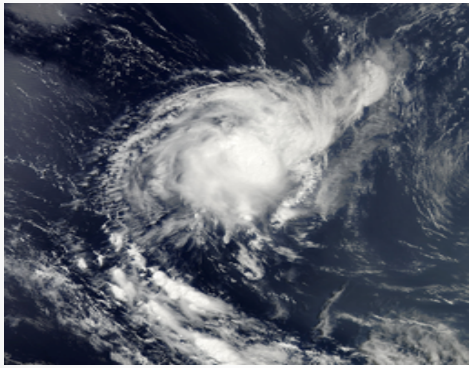
Applications: Searching for Events

- Detection of phenomena in Browse Imagery

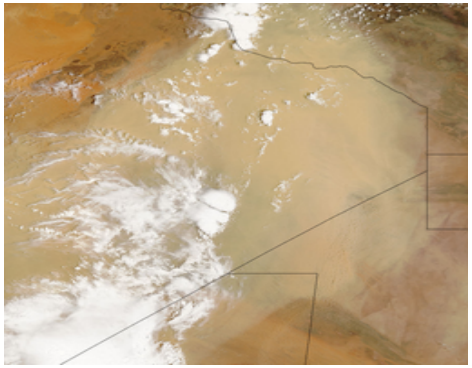
True/Pred	Dust	Hurricane	Smoke	Other
Dust	**287**	8	32	33
Hurricane	0	**379**	1	10
Smoke	12	12	**443**	9
Other	33	9	23	**211**

Confusion Matrix

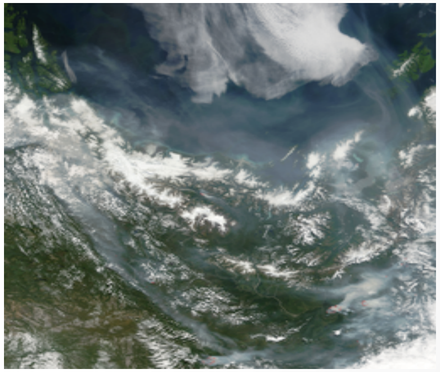
Overall Accuracy = **87.88%**



Hurricane – True Positive



Dust – True Positive



Smoke– True Positive

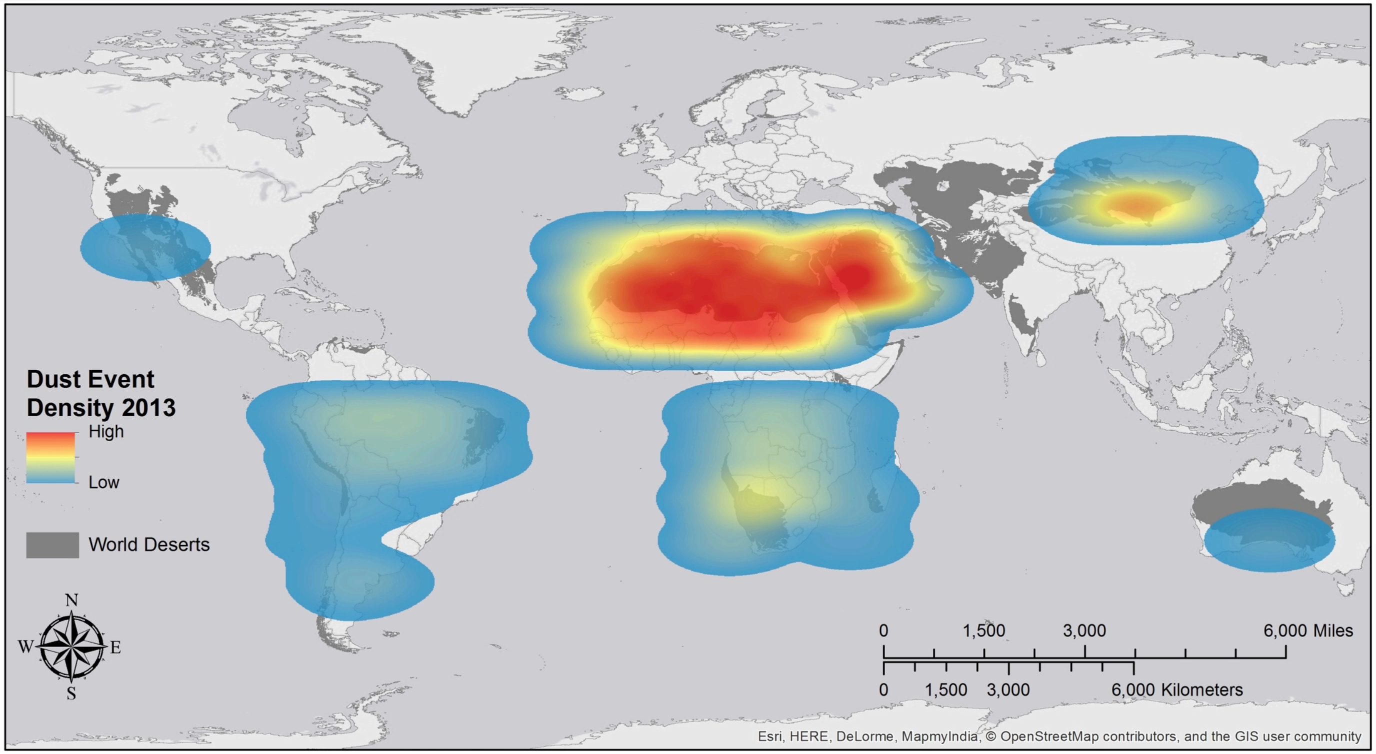
Applications: Enabling new science

- Dust climatology – Collaboration with Sundar Christopher, UAH Atmospheric Science Professor

True\Predicted	**Dust**	**Other**	Total
Dust	**1379**	379	1758
Other	260	**4932**	5192
	1639	5311	6950

Confusion Matrix

Validation Accuracy = **91%**

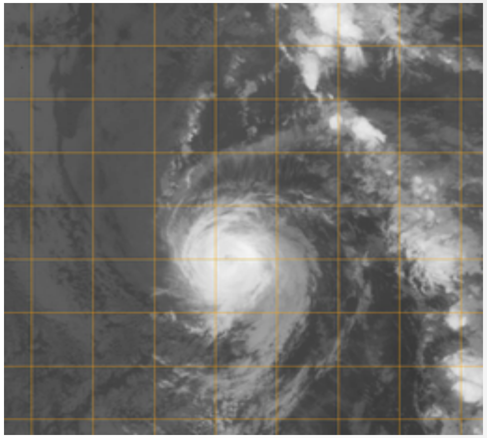


Based on GIBS

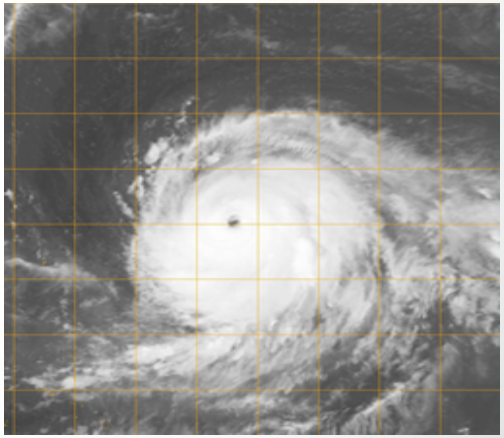
Applications: Improving forecast operations

- Hurricane intensity estimation - Collaboration with Dan Cecil, NASA/MSFC Atmospheric Scientist

True\Predicted	td	ts	h1	h2	h3	h4	h5	no_cat	total
td	3168	335	0	1	0	0	0	6	**3510**
ts	489	4823	159	5	11	3	6	0	**5496**
h1	9	484	1158	92	20	6	1	0	**1770**
h2	3	76	214	513	145	4	0	5	**960**
h3	6	40	33	155	689	55	0	0	**978**
h4	1	18	17	12	142	810	32	0	**1032**
h5	2	2	0	0	27	59	216	0	**306**
no_cat	22	0	0	0	0	0	0	32	**54**
	3700	**5778**	**1581**	**778**	**1034**	**937**	**255**	**43**	**14106**



Cat 2 Hurricane

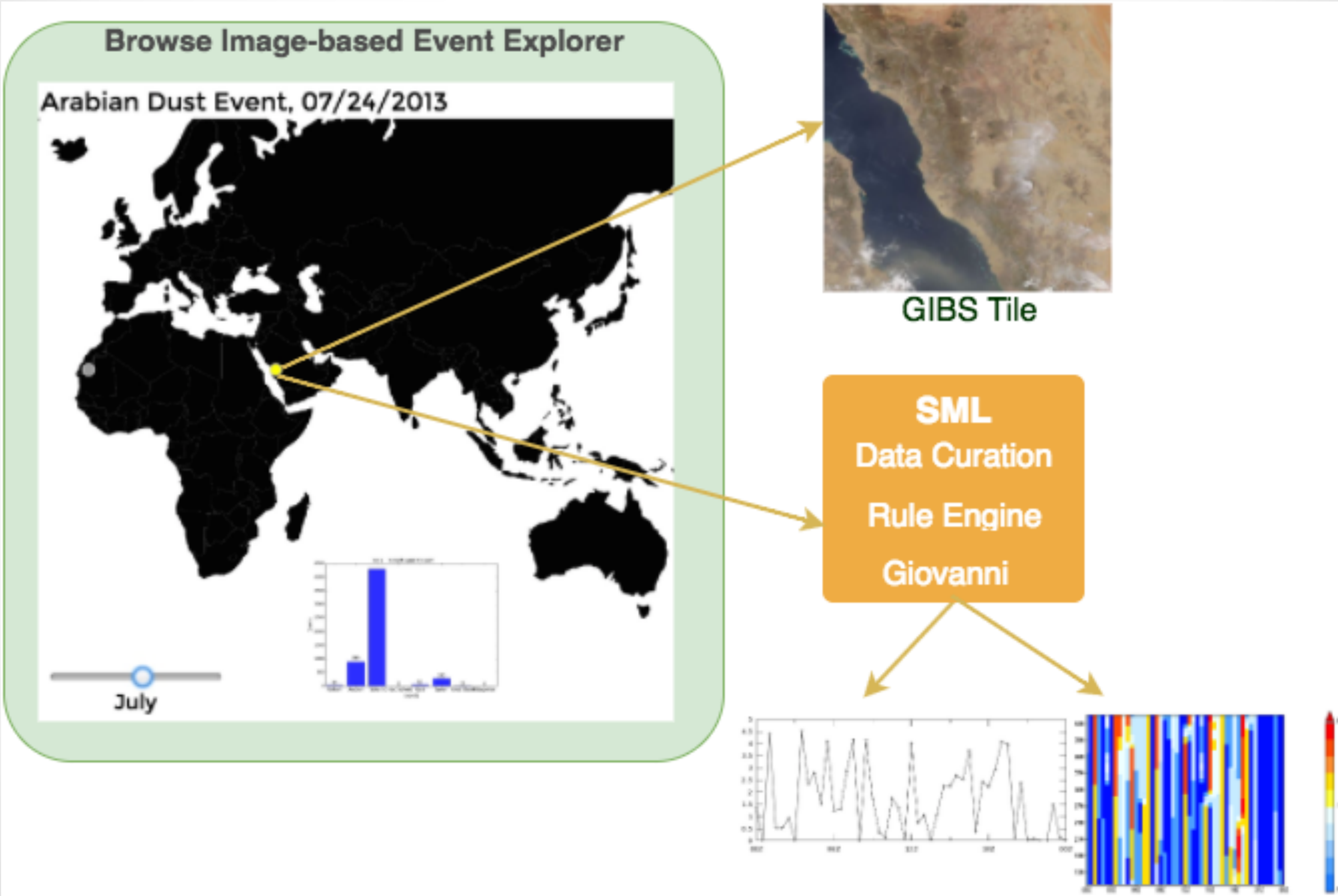


Cat 4 Hurricane

Overall Accuracy : 81 %(Top 2 Probabilities 95.73%)

Data: NRL Images, HURDAT

Ongoing Work



Journal Publications

- Submitted:
  - *Deep Learning for Phenomena-based Classification of Earth Science Images* - IEEE Geoscience and Remote Sensing Letters
- In Progress:
  - *Relevancy Algorithm to Curate Earth Science Data for Different Phenomena* – to be submitted to Computers and Geoscience
  - *Detecting Transverse Cirrus Bands using Deep Learning* - IEEE Geoscience and Remote Sensing Letters (collaboration with U.S. Nair)
- Planned:
  - *Dust climatology* (collaboration with Sundar Christopher)
  - *Hurricane intensity estimation* (collaboration Dan Cecil)

Thanks to NASA ESTO for their support

Contact

Manil Maskey
manil.maskey@nasa.gov

Rahul Ramachandran
rahul.ramachandran@nasa.gov