

ILLUMINATING THE DARKNESS

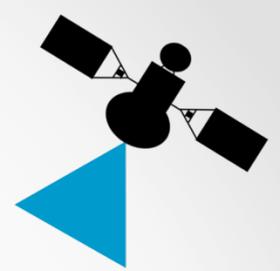
*Exploiting Untapped Data and Information
Resources in Earth Science*

Rahul Ramachandran

NASA/MSFC

rahul.ramachandran@nasa.gov

256-961-7620



Project Team

- Univ. of Alabama in Huntsville
 - Manil Maskey*
 - Xiang Li
- RPI
 - Peter Fox*
 - Stephan Zednik
- NASA/GSFC
 - Steve Kempler (Chris Lynnes*)
 - Suhung Shen
 - Chung-Lin Shie

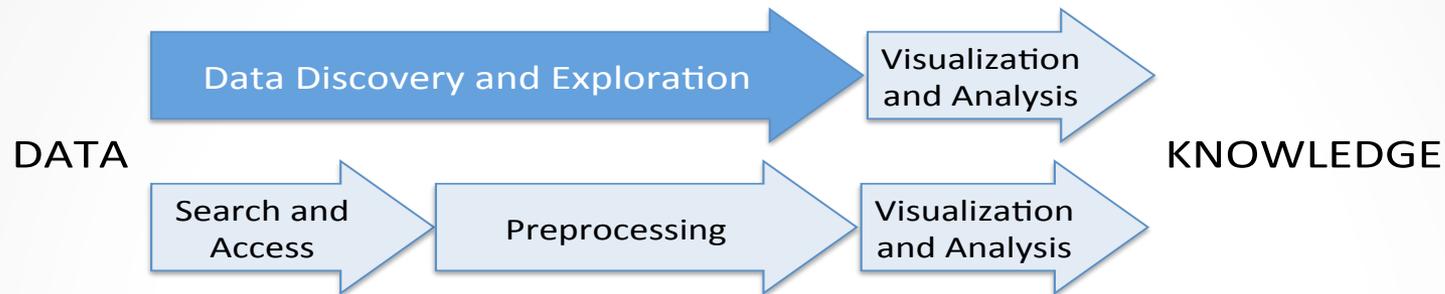
Outline

- Project Overview
- Use Case Deconstruction
- Initial Results from Data Curation Service

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 resources
- Metadata catalogs contain dark resources consisting of structured information, free form descriptions of data and browse images.
 - EOS Clearing House (ECHO) holds 3666 data collections, 127 million records for individual files and 67 million browse images.

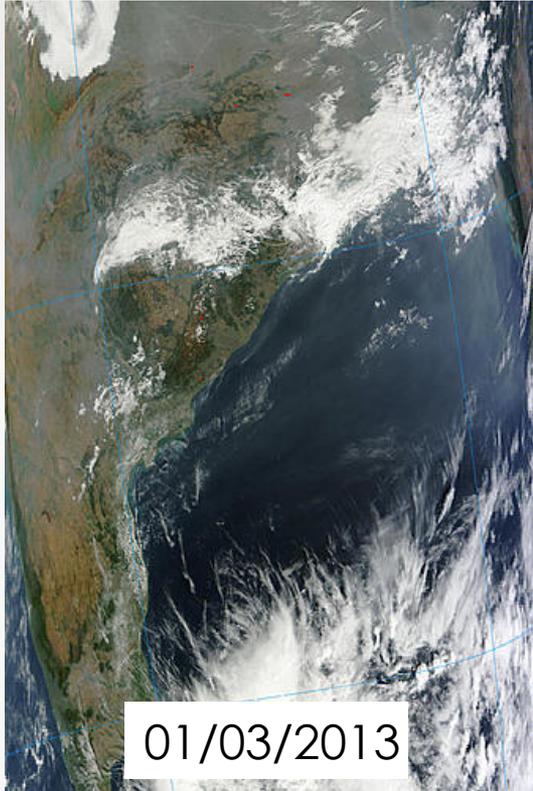
Premise: Metadata catalogs can be utilized *beyond their original design intent* to provide *new data discovery and exploration pathways* to support science and education communities.

Browse Image Example: Understanding regional air pollution from haze



- MODIS 2010 image over India which shows modest level haze pollution
- How often does Haze occur over Indian subcontinent?

Results: Image Retrieval and Metadata



Spatial Distribution Jan-Sept 13

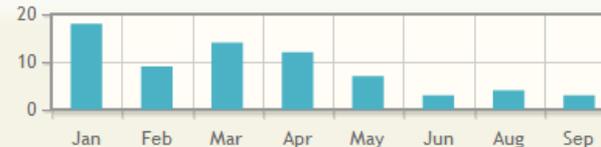


Spatial Distribution Jan 13

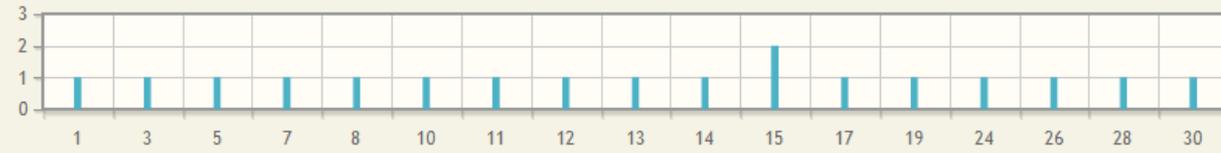
Year Distribution



Month Distribution



Day Distribution



Haze occurs more frequently in Spring than in Summer

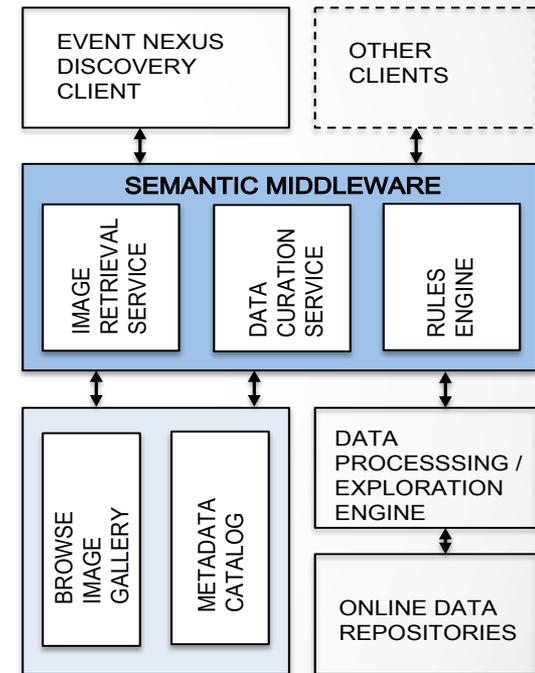
Over half a month in January, haze images were observed in the region

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

Specific Objectives

- Three specific semantic middleware core components
 - *Image retrieval service* - uses browse imagery to enable discovery of possible new case studies and also presents exploratory analytics.
 - *Data curation service* - uses metadata and textual descriptions to find relevant data sets and granules needed to support the analysis of a phenomena or a topic.
 - *Semantic rules engine* - automates data preprocessing and exploratory analysis and visualization tasks.
- Demonstrate value using science use cases



Explore pathways to infuse this technology into existing NASA information and data system

Science Use Cases

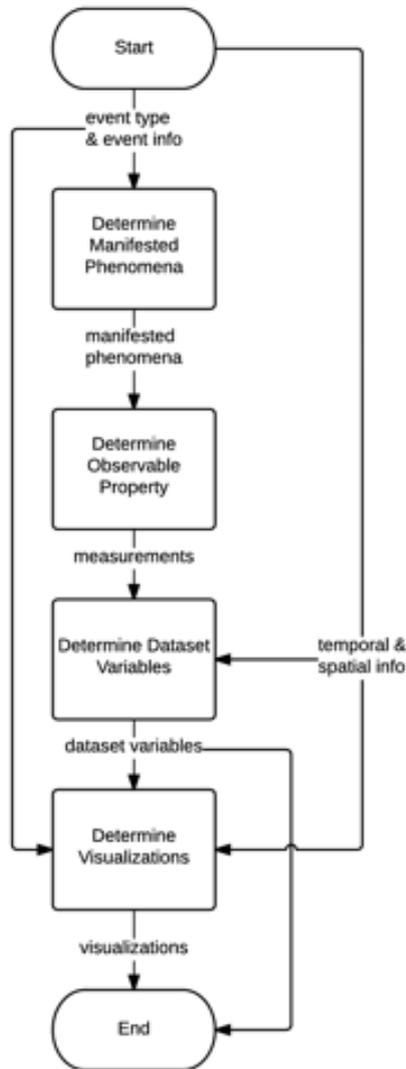
- Phenomena: Dust storms, Volcanic Eruptions, Tropical Storms/Hurricanes
- *Volcanic Eruptions:*
 - Emit a variety of gases as well as volcanic ash, which are in turn affected by atmospheric conditions such as winds.
 - Role of Components
 - Image Retrieval Service is used to find volcanic ash events in browse imagery
 - Data Curation Service provides the relevant datasets to support event analysis
 - Rules Engine invokes a Giovanni processing workflow to assemble and compare the wind, aerosol and SO₂ data for the vent

Part 2: Use Case Deconstruction

...

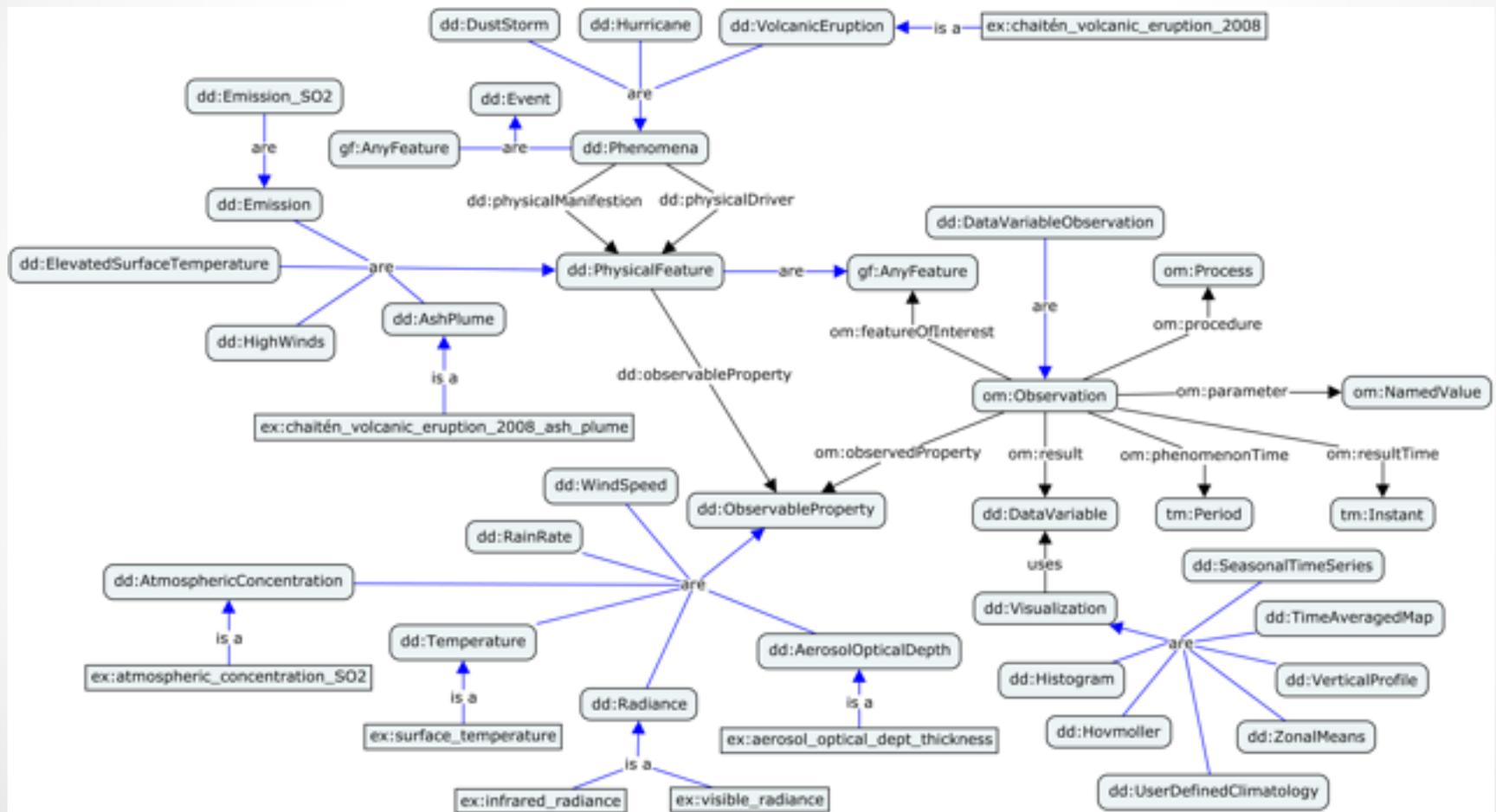
Volcanic Eruptions

Conceptual Flow and Data Dictionary

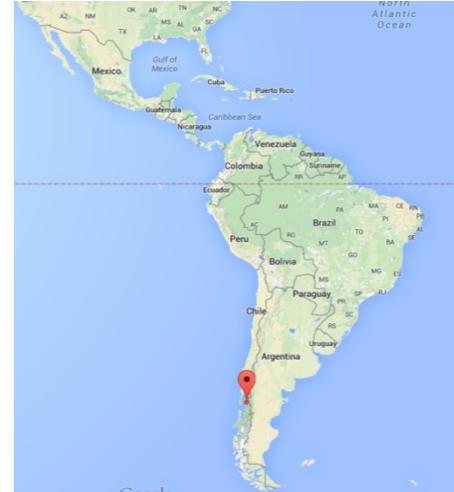


Phenomena : As commonly used in weather observing practice, an observable occurrence of particular physical significance	<ol style="list-style-type: none"> Volcanic eruption Hurricane
Event : Instance of a natural phenomena	<ol style="list-style-type: none"> 2008 Chaitén Volcanic eruption, Hurricane Katrina
Physical Manifestation : feature characteristic – a distinct attribute	Volcano: Ash plume Hurricane: Wind Fields Eye (Atmospheric Pressure)
Instance (time and space) of physical manifestation	<ol style="list-style-type: none"> 2008 Chaitén ash plume Wind speeds in and around Hurricane Katrina
Measurements (Observable Property) : How an instrument observes the Physical Manifestation	<ol style="list-style-type: none"> Volcanic Eruption: SO2 Column, Aerosol Optical Depth Hurricane Rainrate Wind speed/direction
Data Set Variable : Representation of the measurement in a data file, variables within an actual data file	OMSO2e:ColumnAmountSO2_PBL MOD08:Optical_Depth_Land_and_Ocean_Mean Precipitation/Visible Frequencies, Pressure

Initial Model



Volcanic Eruption: Chaitén 2008



The Chaitén Volcano seen from a commercial flight, October 2008. It was into eruptive phase for the first time in about 9,500 years on the morning of May 2, 2008.

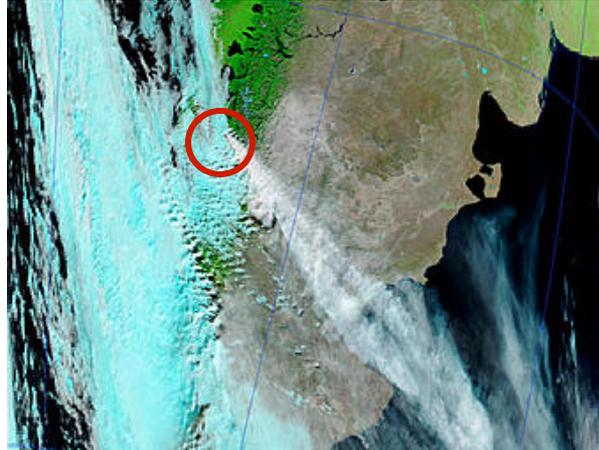
Eruption Time period: May 2 – Nov 2008

Location: Andes region, Chile (-42.832778, -72.645833)

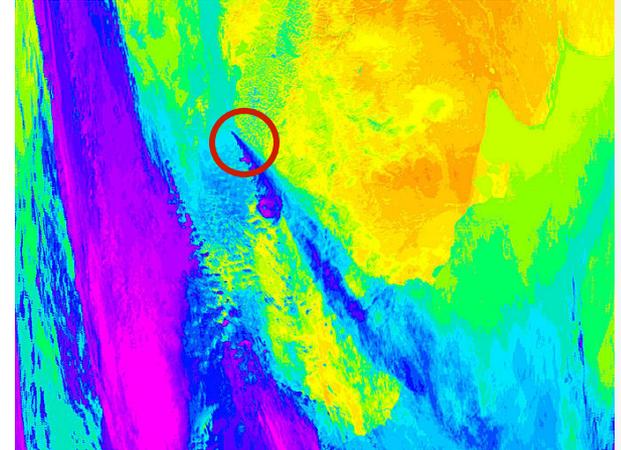
Browse Images



Band 1-4-3 (true color)



Band 7-2-1



LST

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

<http://lance-modis.eosdis.nasa.gov/cgi-bin/imagery/realtime.cgi?date=2008124>

Example Relevant Data

Total SO₂ 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/MOD04_L2/

SeaWiFS Deep Blue Aerosol Optical Depth and Angstrom Exponent Level 2 Data 13.5km

http://disc.gsfc.nasa.gov/datacollection/SWDB_L2_V004.shtml

IR Brightness Temperature:

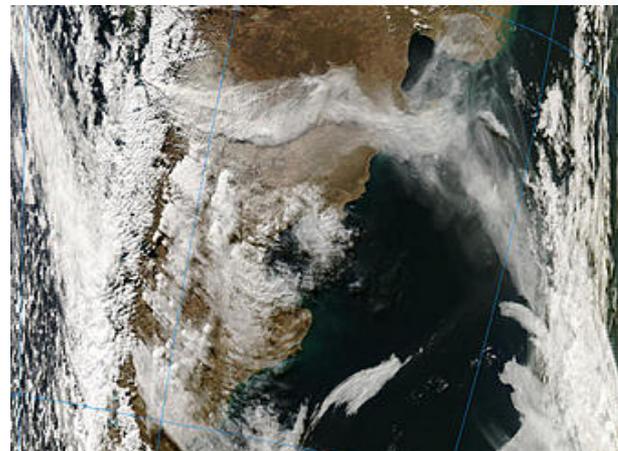
NCEP/CPC 4-km Global (60 deg N - 60 deg S) Merged IR Brightness Temperature Dataset

Giovanni SO2 Plots

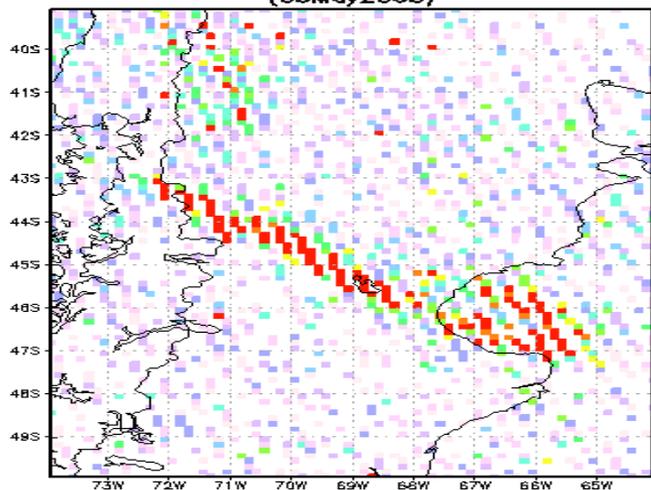
MODIS-Aqua 2008-05-03 18:45 UTC



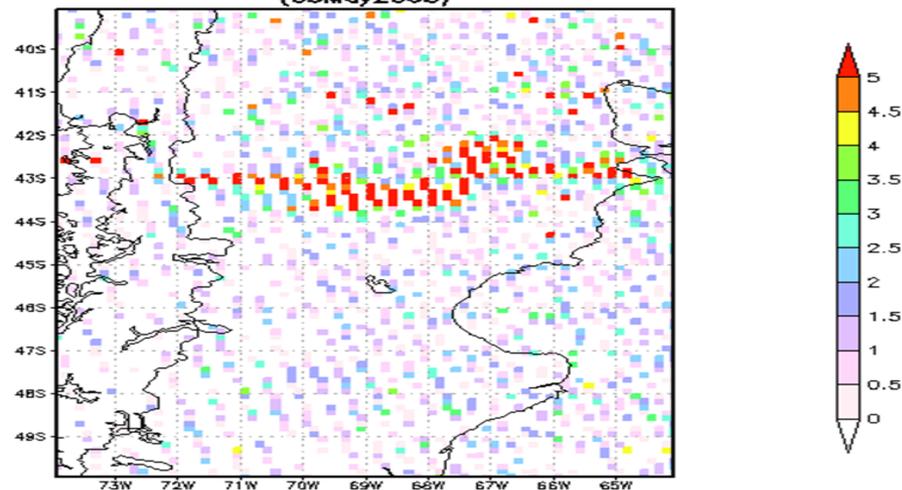
MODIS-Aqua 2008-05-05 18:30 UTC



2G.003 SO2 Column Amount (Planetary Boundary Layer) (03May2008)



2G.003 SO2 Column Amount (Planetary Boundary Layer) [DU] (05May2008)



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

Giovanni Infrared Data Plot

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

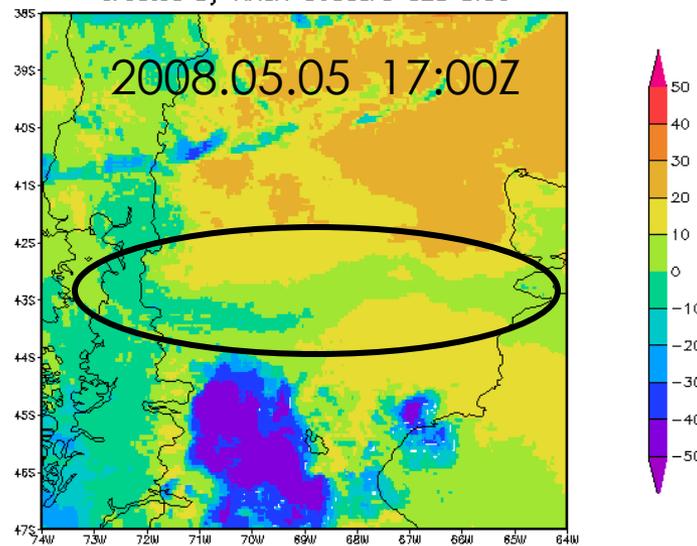
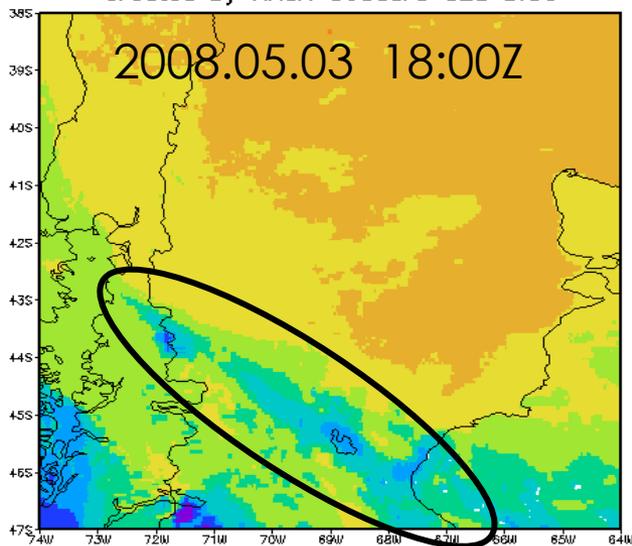


Global Merged IR (00min18Z03MAY2008)
Created by NASA Goddard GES DISC

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



Global Merged IR (00min17Z05MAY2008)
Created by NASA Goddard GES DISC



http://disc.sci.gsfc.nasa.gov/daac-bin/hurricane_data_analysis_tool.pl

Part 3: Data Curation Algorithm for Phenomena

...

Initial Results

Data Curation Algorithm Approaches

- Text mining
 - Pros: Don't need to explicitly define the phenomena
 - Cons: Dependent of the truth set; Catalog is dynamic and new data may never get classified
- Ontology Based
 - Pros: Best precision and recall
 - Cons: Labor intensive to build an explicit model
- Information Retrieval
 - Boolean (Faceted) Search
 - Pros: Simple to implement
 - Cons: Phenomena can be complex; User may not know all the right keywords
 - **Relevancy Ranking Algorithm**
 - Pros: List most relevant data first
 - Cons: Requires a custom algorithm

Assumptions/Observations

- Catalog metadata (ECHO) is rich and most metadata records have been tagged with appropriate vocabulary terms (GCMD)
- Phenomena *can be* defined using a bag of keywords using vocabulary terms
 - User's information need can be captured by using this broad query
- Keywords (tags) in the metadata and the unstructured text (description) can be used
- Keyword is only used once per metadata record
 - Term frequency does not matter
- Document frequency for keywords can be used
 - Some keywords may occur in many metadata records

Experiment Setup and Approach

- Randomly select 200 sample dataset metadata from ECHO
- Label 200 datasets
 - binary: relevant to phenomena/not relevant to phenomena (Hurricane)
- Compile set of keywords (GCMD) relevant to Hurricane – “bag of words” model
- Filter
 - Spatial filter
 - Temporal resolution
 - “<= daily”
 - 85 datasets filtered out
- Apply algorithms on remaining 115 datasets
 - Jaccard coefficient-based ranking
 - Vector Space Model using Cosine similarity-based ranking

Algorithms

Jaccard Coefficient

$$J(A,B) = |A \cap B| / |A \cup B|$$

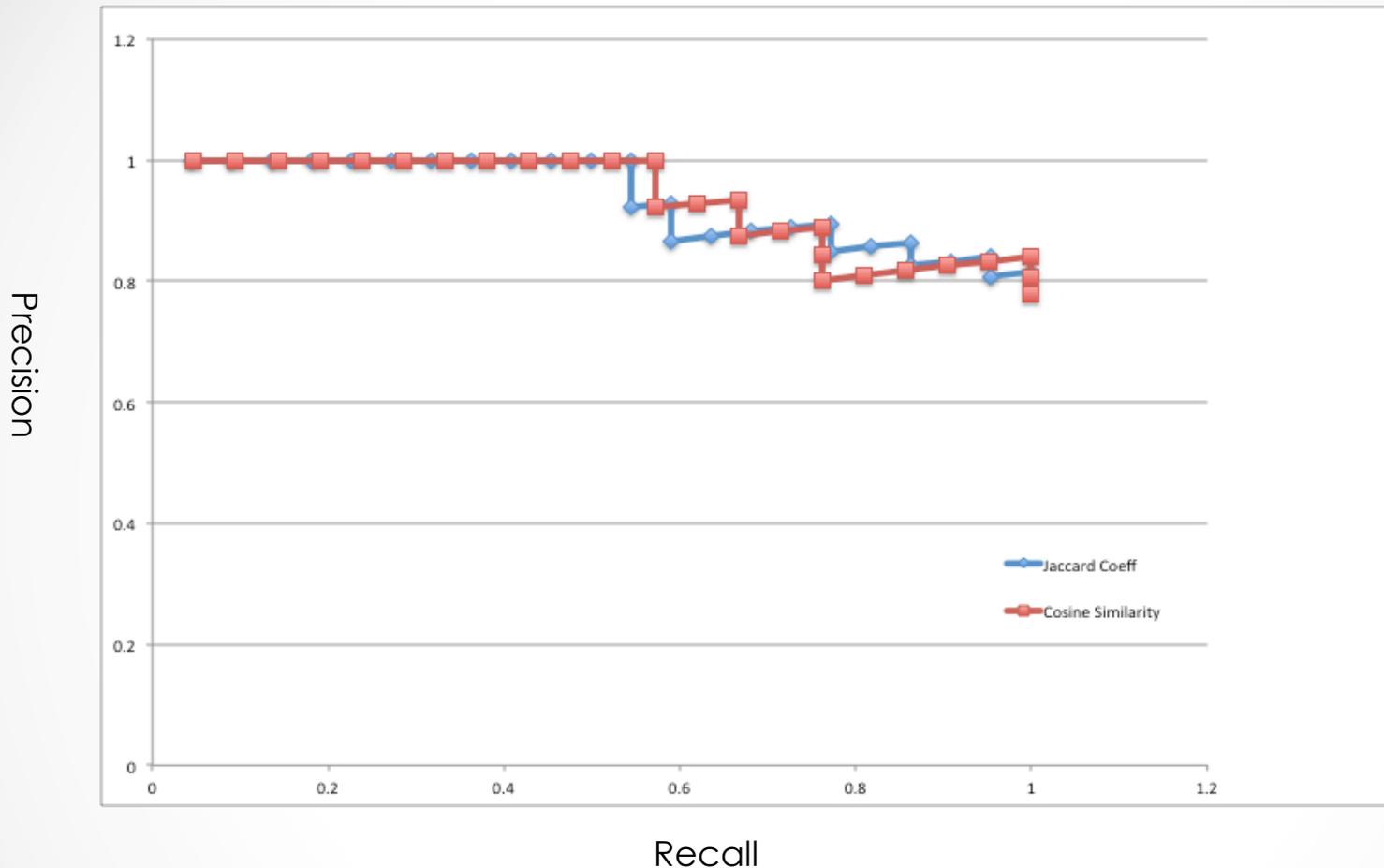
Where:

- A - keywords defining a phenomena
- B - keywords in a given dataset

Vector Space Model

- Determine term frequency (tf): (1 in our case)
- Determine inverse document frequency (idf): number of metadata records that contain the keyword
- Calculate Cosine similarity
 - Sum (tf x idf) for each keyword

Retrieval Results



Top 20 Returns
80 % precision with a 75% recall - 75% of relevant data are retrieved with 80% precision

Milestones

- Kickoff 02/15
- Ontology based data curation service 11/15
- Rule based processing prototype 02/16
- Content-based image retrieval service 06/16
- Improved rule-based processing 12/16
- Data exploration demo: 3 use cases (TRL 3/4) 12/16

Questions



Contact Info:

rahul.ramachandran@nasa.gov

256-961-7620