EXPLOITING DARK INFORMATION RESOURCES TO CREATE NEW VALUE ADDED SERVICES TO STUDY EARTH SCIENCE PHENOMENA

Rahul Ramachandran NASA/MSFC, Manil Maskey UAH, Xiang Li UAH, Kaylin Bugbee UAH

Project Team:
MSFC/UAH: Patrick Gatlin, Amanda Weigel, JJ Miller, Ajinkya Kulkarni
GSFC: Steve Kempler, Suhung Shen, Chung-Lin Shie, Maksym Petrenko
RPI: Peter Fox, Stefan Zednik, Anirudh Prabhu

Invited Talk: Earth Observing Data Science IGARSS July 10-15, 2016 Beijing, China
Goals: to make this process efficient, address existing gaps/hurdles, seamlessly integrate new emerging technology, and enable new research capabilities.
1. Project Overview
2. Data Curation Service
3. Rules Engine
4. Application (with Demo)
5. Image Retrieval Service
6. Summary
Part 1: Project Overview
Motivation

• Data preparation steps are **cumbersome** and **time consuming**
  o Covers discovery, access and preprocessing

• Limitations of current Data/Information Systems
  o **Boolean search** on data based on instrument or geophysical or other **keywords**
  o Underlying **assumption** that users have sufficient knowledge of the **domain vocabulary**
  o **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 (CMR) holds >6000 data 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
  o provide novel **data discovery and exploration** capabilities that significantly reduce data preparation time.
  o 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

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 SO₂ mass:
e.g. Chaitén is 10 (kt) = (kilotons) , (1 kt = 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 Data 13.5km
http://disc.gsfc.nasa.gov/datacollection

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

Data Curation Service recommends relevant datasets to support event analysis
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

**Phenomena**

- Hurricane
- Tropical Storm
- Dust Storm
- Volcanic Eruption

**Physical Feature**

- Ash Plume
- Area of High Winds
- Area of Elevated Surface Temperature
- Area of High Particulate Emissions

**Observable Property**

- Temperature
- Radiance
- Wind Speed
- Rain Rate

**Data Variable**

- MOD04_L2:Optical_Depth_Land_and_Ocean_Mean
- MOD02HKM:bands 1, 3, and 4
- OMSO2e:ColumnAmountSO2_PBL
Part 2: Data Curation Algorithm for Phenomena
Data Curation

• Curation is traditionally defined as the process of collecting and organizing information around a common subject matter or a topic of interest and typically occurs in museums, art galleries, and libraries.

• Ramachandran et al. [2015] define geocuration as the act of searching, selecting, and synthesizing Earth science data/metadata and information from across disciplines and repositories into a single, cohesive, and useful collection.
  o Manual
  o Automated
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:
  o Given a phenomenon type (Ex: Hurricane), DCS returns a list of relevant data sets (variables)
    • <list of data sets (variables)> = DCS(Phenomenon Type)
  o 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

- USER TASK
  - Study “Hurricane”
- INFO NEED
  - All data sets useful in studying “Hurricane”
- QUERY
  - By pass this step for the end user

How to define a phenomenon?

How to automatically formulate query? (Misformulation Issue)

Best relevancy ranking algorithm?
Our Approach

USER TASK

INFO NEED

Study “Hurricane”
All data sets useful in studying “Hurricane”

How to define a phenomenon?

How to automatically formulate query?

Best relevancy ranking algorithm?

REFINE

QUERY

SEARCH ENGINE

DOCUMENT COLLECTION

RESULTS

Expert select “bag of words” to define a phenomena

Use Controlled vocabulary (for the “words”)

Develop a custom Relevancy ranking algorithm
Methodology

• Selected three metadata fields
  o Science Keywords
  o Data set long name (title)
  o Data set description

• Developed customized vector space model for each field

• Compared different similarity measures
  o Cosine vs Jaccard

• Used Weighted Zone Ranking (Ensemble)
  o $S_c(e) = w_s \cdot S_c(s) + w_l \cdot S_c(l) + w_d \cdot S_c(d)$
Experiment Setup

- **Datasets**: 200 Randomly selected from ECHO

  - **Binary Labeling**: relevancy to phenomena (Label = Majority: 3 Scientists)

- **Labeled Datasets**

- **Temporal and Spatial Filtering**

- **Bag-of-keywords**
  - 92 keywords: hurricane
  - 103 keywords: volcano

  - Manually Selected set of GCMD Science Keywords relevant to phenomena

- **Evaluation Metrics**:
  - Jaccard Coeff
  - Cosine Similarity
  - Zone Ranking
  - Ensemble

18
## Comparison of Similarity Measures

<table>
<thead>
<tr>
<th></th>
<th>Hurricane</th>
<th></th>
<th>Volcanic Eruption</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jaccard</td>
<td>Cosine Similarity</td>
<td>Jaccard Coefficient</td>
<td>Cosine Similarity</td>
</tr>
<tr>
<td></td>
<td>Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10 retrieval</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Top 20 retrieval</td>
<td>17</td>
<td>16</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Top 30 retrieval</td>
<td>23</td>
<td>24</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>

- Both of the measures performed similarly
- Selected Cosine Similarity measure because it is commonly used in space vector model information retrieval
# Ranking Results (Top 20) using Ensemble Method

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Optimal Weight</th>
<th>Equal Weight</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Hurricane</td>
<td>90.0%</td>
<td>47.4%</td>
<td>85.0%</td>
</tr>
<tr>
<td>Volcanic Eruption</td>
<td>85.0%</td>
<td>68.0%</td>
<td>80.0%</td>
</tr>
<tr>
<td>Fire</td>
<td>75.0%</td>
<td>30.0%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Flood</td>
<td>65.0%</td>
<td>48.1%</td>
<td>55.0%</td>
</tr>
</tbody>
</table>

- Different numbers of “relevant” data sets, collection size (recall) exist within each truth set for each phenomenon.
- Better to compare the curation results against the random selection rather than compare the performance against each other.
- On average, precision improves about 25% when using our method and recall improves about 16%.
Optimal ensemble weights for each phenomenon

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Optimal Weight Set ($W_{\text{sciencekeyword}}$, $W_{\text{longname}}$, $W_{\text{description}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hurricane</td>
<td>(0.6, 0.1, 0.3)</td>
</tr>
<tr>
<td>Volcanic Eruption</td>
<td>(0.2, 0.6, 0.2)</td>
</tr>
<tr>
<td>Fire</td>
<td>(0.6, 0.2, 0.2)</td>
</tr>
<tr>
<td>Flood</td>
<td>(0.5, 0.4, 0.1)</td>
</tr>
</tbody>
</table>

- Weight for science keyword is largest while the weight for description is smallest
  - Science keywords metadata fields use a controlled vocabulary and should be accurate and consistent
  - Description field is free-text and has the most variability in quality
Methodology Limitations

• **Modeling the search intent is difficult**
  o one may be interested in only a specific aspect of a phenomenon whereas another user may only be interested in some other characteristic of a phenomenon

• **Quality of metadata records is variable**
  o Key assumptions is that the metadata records stored in the CMR catalog are consistent, correct, and complete
  o Launched a project to fix this

• **Granularity of the Controlled Vocabulary**
  o Rich detailed controlled vocabulary provides a better level of annotation granularity to represent different phenomena and help disambiguate data sets

• **Truth set labels may be biased**
  o domain experts on our team have stronger expertise in certain areas such as hurricanes and weaker expertise in others
Next: Find relevant data fields

• Need actual data variables
  o Example: Giovanni uses these fields for visualization

• What we know
  o Data set (Collection) level science keywords (GCMD) – Experts
  o Granule data fields and metadata – Auto extract*

• How do we map?
  o Start with GCMD to CF Standard name
  o Most don’t follow CF Standard names
**Approach**

**Dataset**
- Extract Science Keywords
- Text processing
- Normalization
- Bag-of-words

**Granules**
- Extract Variables and Descriptions
- Text processing
- Look up Table
- Normalization
- Bag-of-words

- OPeNDAP, netCDF Libs, ...
- Remove special characters, Tokenize, ...
- Acronym/Abbreviation expansion, CF
- Remove stopwords/Stem/Lemmatize

**NLP**
Learn Patterns

**Suggest Keywords**
Assess Metadata

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

John.P.Dimarzio.1@nasa.gov

ICESat Science Investigator-led Processing System (I-SIPS)
757-864-1238 (phone)
David.W.Hancock@nasa.gov

NASA DAAC at the National Snow and Ice Data Center
303-492-6199 (phone)
303-492-2468 (fax)
nsidc@nsidc.org

Science Keywords:
Earth Science  Atmosphere  Clouds

Earth Science  Atmosphere  Aerosols

MONTH

GLAS/ICESat L2 Global Thin Cloud/Aerosol Optical Depths Data (HDF5) V033
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_cld1_od** | Cloud Optical Depth at 532 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
  - more with low scores

- **Serendipitous Discovery** - Data Curation Parameter Mapping
  Algorithm can be used to assess
  - Metadata quality for both dataset and granules
  - Find incorrect/incomplete keyword annotations
  - Automatically suggest science keywords
Parameter Mapping Tool

Data Parameter Mapping Tool

Datasets
- AIRS/Aqua Level 2 Support retrieval (AIRS+AMSU) V005
- GHRST Level 2P USA NASA MODIS Aqua SST:1
- MODIS/Terra Temperature and Water Vapor Profiles 5-Min L2 Swath 5km V005
- LIS/OTD 2.5 DEGREE LOW RESOLUTION DIURNAL CLIMATOLOGY (LDC) V2.3.2013
- MODIS/Terra Aerosol 5-Min L2 Swath 10km V005 NRT

Science Keyword Map

Parameter Mapping

- ATMOSPHERE -> AEROSOLS -> PARTICULATE_MATTER
- Deep_Blue_Aerosol_Optical_Depth_Land_STD : 1
- Aerosol_Cldmask_Byproducts_Ocean : 1
- Aerosol_Cldmask_Byproducts_Land : 1
- Deep_Blue_Aerosol_Optical_Depth_Land : 1
- Aerosol_Type_Land : 1
- Optical_Depth_Small_Average_Ocean : 0
- Asymmetry_Factor_Best_Ocean
- Deep_Blue_Angstrom_Exponent_Land
- Cloud_Fraction_Ocean
- ATMOSPHERE -> AEROSOLS -> CLOUD_CONDENSATION_NUCLEI : 1

Edit/Save Mapping

Mapping Scores Generated by Algorithm
Part 3: Rules Engine
What settings should I use to visualize this event?

Goal: Automate data preprocessing and exploratory analysis and visualization tasks
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.
Survey asked users to rate characteristics of phenomena features

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),
  ...
]
```
<table>
<thead>
<tr>
<th>Phenomena</th>
<th>East-West Movement</th>
<th>North-South Movement</th>
<th>Temporal Evolution</th>
<th>Spatial Extent of Event</th>
<th>Year-to-Year Variability</th>
<th>May Impact Seasonal Variation</th>
<th>Variation with Atmospheric Height</th>
<th>Global Phenomena</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volcano - Ash Plume</td>
<td>Indifferent</td>
<td>Indifferent</td>
<td>Strong</td>
<td>Slight</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
</tr>
<tr>
<td>Flood</td>
<td>Some</td>
<td>Some</td>
<td>Strong</td>
<td>Some</td>
<td>Some</td>
<td>Strong</td>
<td>Some</td>
<td>Slight</td>
<td>Some</td>
</tr>
<tr>
<td>Dust Storm</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Strong</td>
<td>Indifferent</td>
<td>Indifferent</td>
<td>Strong</td>
<td>Indifferent</td>
<td>Some</td>
</tr>
</tbody>
</table>
## Service to Characteristic Mappings

<table>
<thead>
<tr>
<th>Service</th>
<th>Visualizatio</th>
<th>East-West Movement</th>
<th>North-South Movement</th>
<th>Temporal Evolution</th>
<th>Spatial Extent of Event</th>
<th>Year-to-Year Variability</th>
<th>Seasonal Variation</th>
<th>Variation with Atmospheric Height</th>
<th>Global Phenomena</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-averaged Map</td>
<td>Color-Slice Map</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area-averaged Time Series</td>
<td>Time Series</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>User-defined Climatology</td>
<td>Color-Slice Map</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Vertical Profile</td>
<td>Line Plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Seasonal Time Series</td>
<td>Time Series</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zonal Means</td>
<td>Line Plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hovmoller (Longitude)</td>
<td>Color-Slice Grid</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hovmoller (Latitude)</td>
<td>Color-Slice Grid</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Compute Compatibility

Phenomena: Volcano - Ash Plume

Service - Area Averaged Time Series

<table>
<thead>
<tr>
<th>Temporal Evolution</th>
<th>Detection of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Strong</td>
</tr>
</tbody>
</table>

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

STRONG COMPATIBILITY x2

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

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

Curation Service: event type filters relevant variables

Selected event & its time

Event Client
Giovanni - Dark Data Edition

Event Discovery
- Dust
- Volcano
- ENS
- Calbuco
- Fire
- Hurricane
- Flood

Event Analysis
Calbuco 2015-04-22
- Turbulent Fluxes - 0.9
- Wind stress - 26
  Time Series
- How2Mlter
- Heat Flux - 21
  Vertical Profile
- MODIS Aerosols - 0.8
  AOD - 15
  Zonal Mean

Data Curation Service
Rules Service
Giovanni Analysis Service

Event Analysis Workflow
Part 5: Image Retrieval
Image Retrieval

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

• Challenge: “semantic gap”
  o 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
  o Hierarchical Learning
  o Mimics the human brain that is organized in a deep architecture
  o Processes information through multiple stages of transformation and representation

Convolutional Neural Network (CNN) - Applicable to Images
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
    o AlexNet architecture - Krizhevsky et. al.
    o Weights learned from ImageNet 1.3 million high-resolution images
    o State-of-the-art classification accuracy
Experiment: CNN Configuration

- AlexNet architecture
  - Initialized weights with ImageNet trained model
  - Adaptive learning rate
  - 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)

<table>
<thead>
<tr>
<th>True/Pred</th>
<th>Dust</th>
<th>Hurricane</th>
<th>Smoke</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust</td>
<td>287</td>
<td>8</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>Hurricane</td>
<td>0</td>
<td>379</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Smoke</td>
<td>12</td>
<td>12</td>
<td>443</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>33</td>
<td>9</td>
<td>23</td>
<td>211</td>
</tr>
</tbody>
</table>

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
Applications: Enabling new science

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

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>Dust</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust</td>
<td>1379</td>
<td>379</td>
<td>1758</td>
</tr>
<tr>
<td>Other</td>
<td>260</td>
<td>4932</td>
<td>5192</td>
</tr>
<tr>
<td></td>
<td>1639</td>
<td>5311</td>
<td>6950</td>
</tr>
</tbody>
</table>

Validation Accuracy = 91%

Confusion Matrix

Based on GIBS
Applications: Improving forecast operations

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

<table>
<thead>
<tr>
<th>True\Predicted</th>
<th>td</th>
<th>ts</th>
<th>h1</th>
<th>h2</th>
<th>h3</th>
<th>h4</th>
<th>h5</th>
<th>no_cat</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>td</td>
<td>3168</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>3510</td>
</tr>
<tr>
<td>ts</td>
<td>489</td>
<td>4823</td>
<td>159</td>
<td>5</td>
<td>11</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>5496</td>
</tr>
<tr>
<td>h1</td>
<td>9</td>
<td>484</td>
<td>1158</td>
<td>92</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1770</td>
</tr>
<tr>
<td>h2</td>
<td>3</td>
<td>76</td>
<td>214</td>
<td>513</td>
<td>145</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>960</td>
</tr>
<tr>
<td>h3</td>
<td>6</td>
<td>40</td>
<td>33</td>
<td>155</td>
<td>689</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>978</td>
</tr>
<tr>
<td>h4</td>
<td>1</td>
<td>18</td>
<td>17</td>
<td>12</td>
<td>142</td>
<td>810</td>
<td>32</td>
<td>0</td>
<td>1032</td>
</tr>
<tr>
<td>h5</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>59</td>
<td>216</td>
<td>0</td>
<td>306</td>
</tr>
<tr>
<td>no_cat</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>54</td>
</tr>
</tbody>
</table>

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

Data: NRL Images, HURDAT
Ongoing Work

Browse Image-based Event Explorer

Arabian Dust Event, 07/24/2013

GIBS Tile

SML Data Curation
Rule Engine
Giovanni

July

33
Summary

- Science data and information systems need to evolve to enable better data search, access and usability!
- Need operational services like – Data Curation Service, Rules Engine and Image Retrieval
Questions

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/