UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Understanding Climate Change from Data Workshop

NSF Expeditions
Data Mining Challenges Panel
August 15-16, 2011
University of Minnesota

Sara J. Graves, PhD

Director, Information Technology and Systems Center

University Professor, Computer Science Department

University of Alabama in Huntsville

Director, Information Technology Research Center

National Space Science and Technology Center

256-824-6064

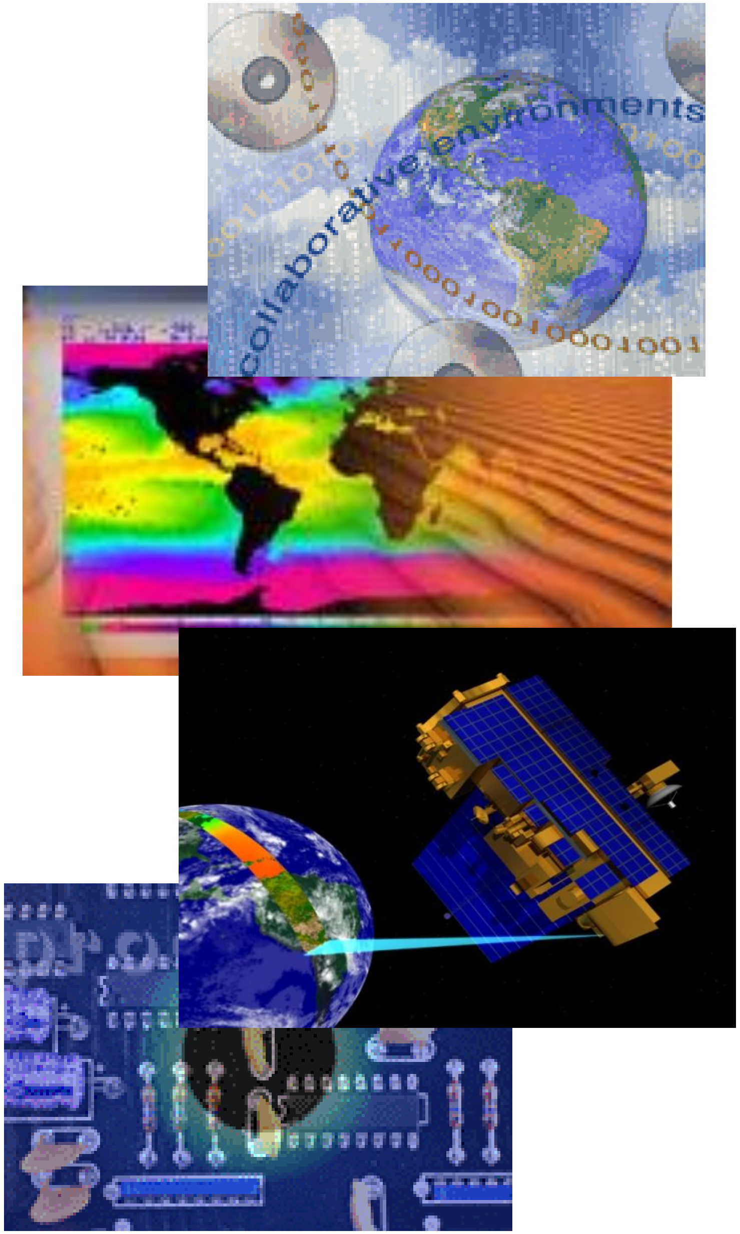
sgraves@itsc.uah.edu

NSSTC Core Facility

http://www.itsc.uah.edu

Primary Research Focus Areas

- Data Mining and Knowledge Discovery
- Visual Analytics
- Modeling and Simulation
- Interoperability
- Knowledge Engineering and Semantic Web
- Information Management and Data Technologies
- Asynchronous Collaboration Technologies / Web 2.0
- Geoinformatics
- Decision Support
- Information Assurance / Cybersecurity
- Near-real-time Processing
- Event-driven / On-demand Processing
- Geospatial Analysis
- Sensor Networking Analysis



Success Built on the Integration of Domain Science and Information Technology

Domain Scientists and Engineers

- Research and Analysis
- Data Set Development

Information Technology Scientists

- Information Science Research
- Knowledge Management
- Data Exploitation

Collaborations

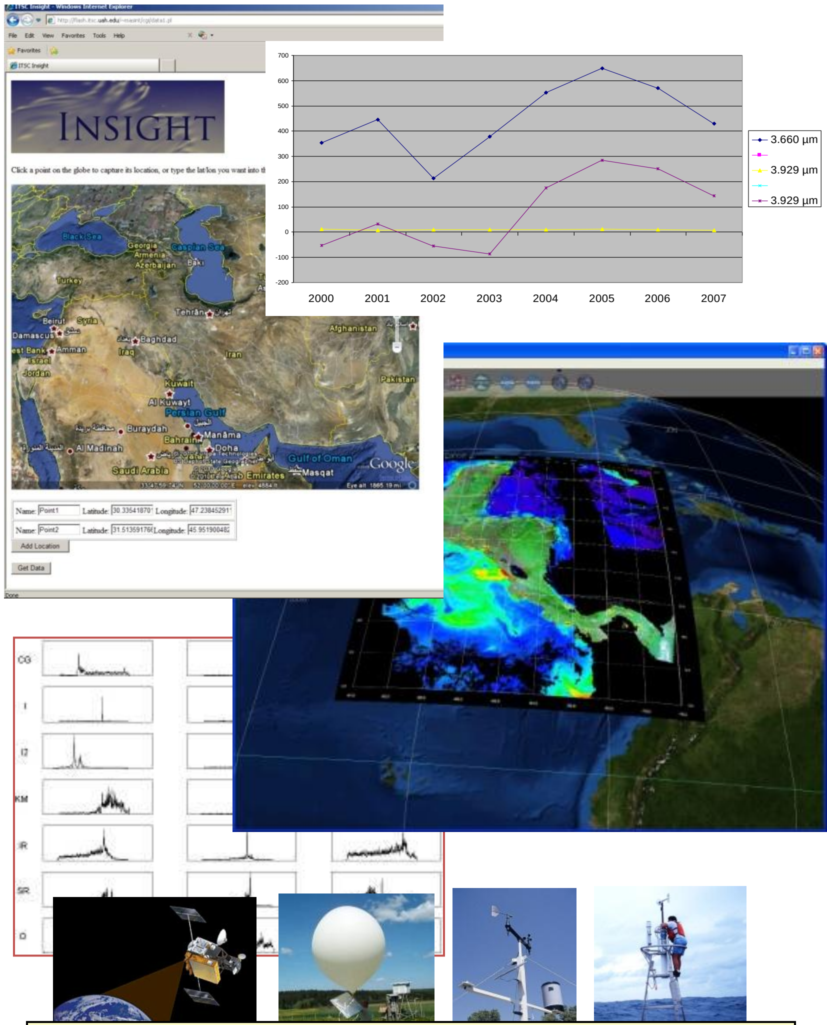
- Accelerate research process
- Maximize knowledge discovery
- Minimize data handling
- Contribute to multiple fields

Data Mining: Situational Awareness and Analysis

How do you get the right information to the right people at the right time?

•Sensor Data Integration/Fusion
•Signature Analysis
•Pattern Recognition
•Real-Time Data Analysis

- **ADaM Algorithm Development and Mining toolkit**
- **MASINT Signature Analysis**
  - **Thermal analysis of human activity**
  - **Measuring nuclear, chemical and oil facility usage and production**
  - **Evaluating environmental impacts on national security**
- **OPIR Analysis**
  - **Algorithm Development**
  - **Multi-source integration and fusion**
  - **Signatures of missile systems**
- **NASA/USAID sponsored SERVIR Environmental Data Products for Central America , Africa and Nepal**
  - **Decision Support System for environmental analysis**
- **JCTD EUCOM Efforts**
  - **Providing data products for the Arctic region**
- **NSF Linked Environments for Atmospheric Discovery**
  - **Real-time mining and analysis**
  - **Adaptive processing**

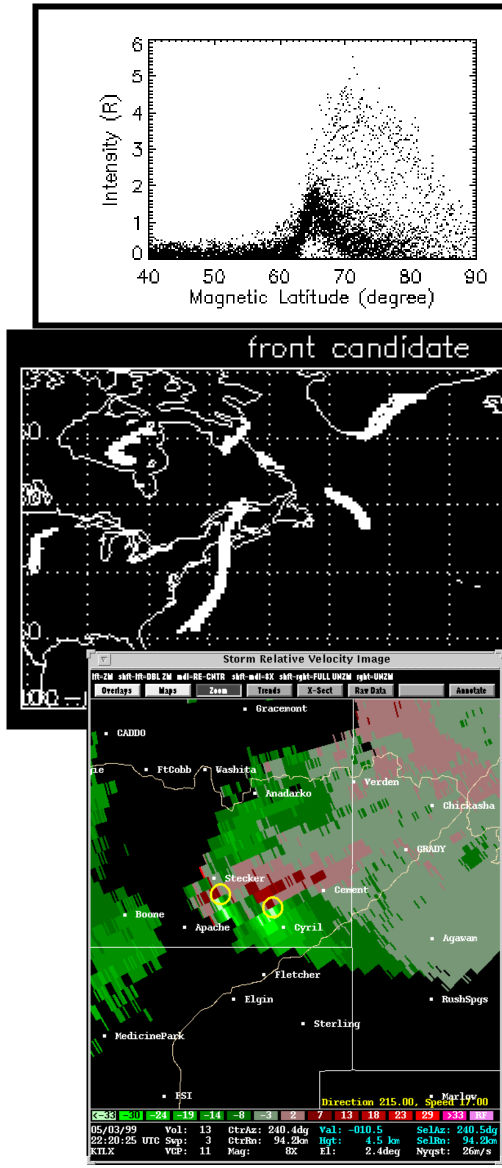


Sensor Data Integration is Critical for Meaningful Situational Awareness

http://www.itsc.uah.edu

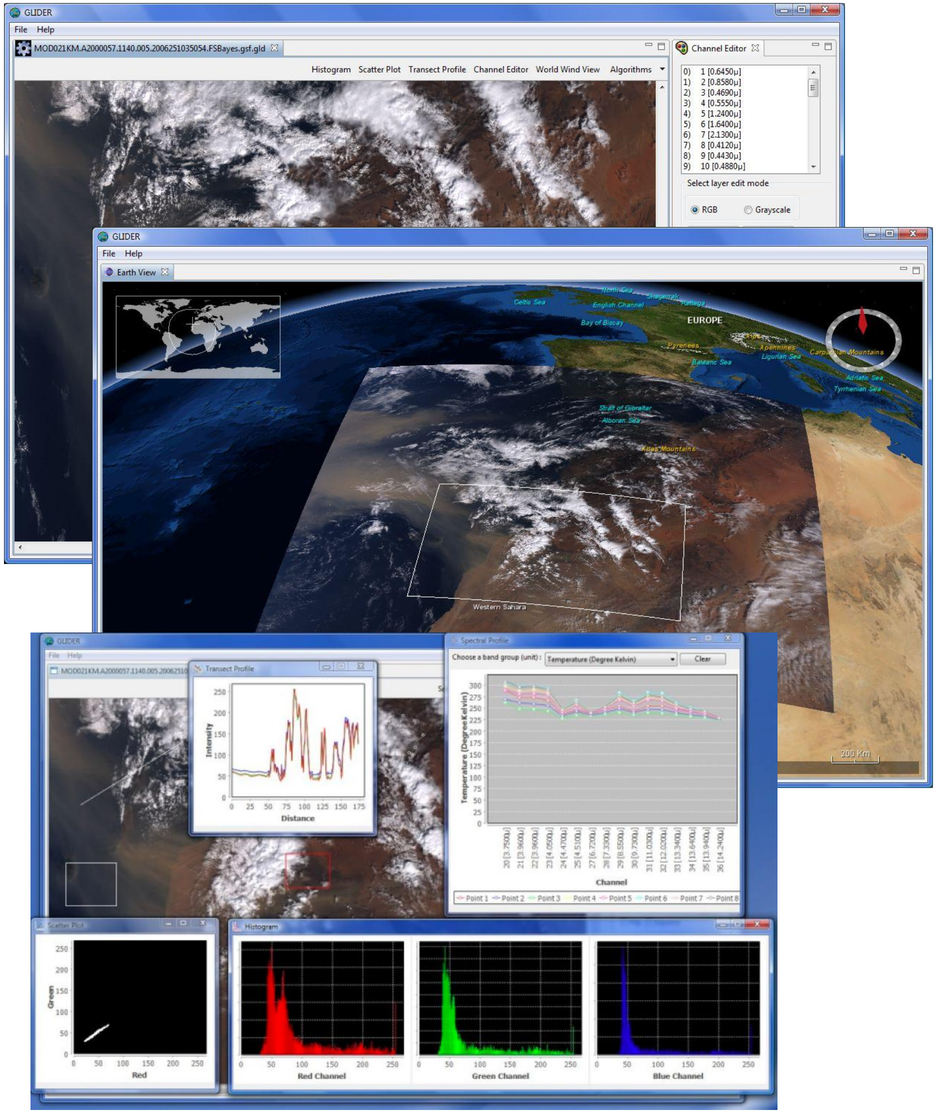
Data Mining: Algorithm Development and Mining (ADaM) Toolkit

- UAHuntsville has been at the forefront of mining sensor data for over 15 years
- ADaM – UAHuntsville developed toolkit with 100+ algorithms, used worldwide
- Automated discovery of patterns, signatures, anomalies
- Derived knowledge for decision making and response
- Allows learning and training for adaptation
- Most cited article in Elsevier *Computers and Geosciences*, 2005-2010



http://www.itsc.uah.edu

GLIDER: Globally Leveraged Integrated Data Explorer for Research



Capabilities:

- ***Visualize and analyze*** satellite data in a native sensor view
- Apply ***image processing algorithms*** on the data
- Apply ***pattern recognition/data mining algorithms*** on the data
- ***3D Globe Visualization*** *of* satellite data, analysis/mining results, and additional layers
- Provides ***multiple views*** to manage, visualize, and analyze data

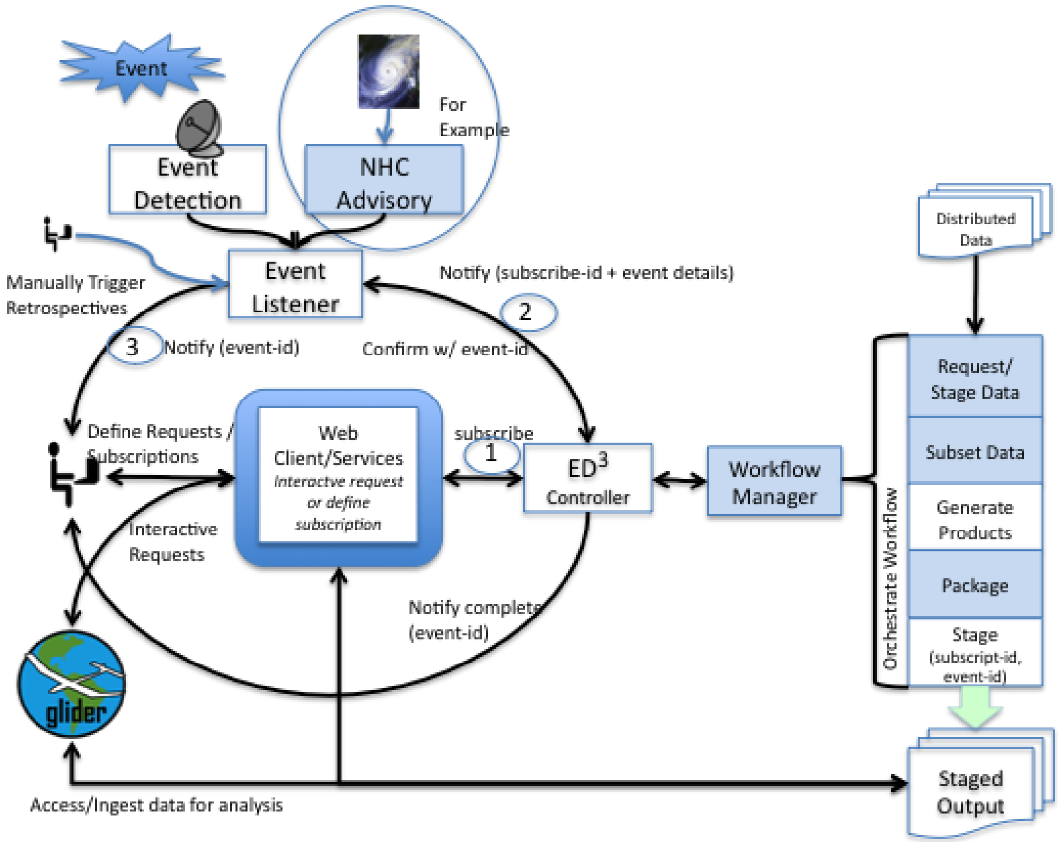
Integrates existing tools:

- ADaM: UAHuntsville's ***Algorithm Development and Mining*** Toolkit
- IVICS: UAHuntsville's ***Interactive Visualizer and Image Classifier for Satellites***
- WorldWind: NASA's ***3-D globe visualization*** system

2010 winner NASA ESDSWG Software Reuse Award and installed at MSIC

Data Management and Dissemination for Analysis and Visualization

Event-Driven Data Delivery (ED^3)



- ***Automated and discrete access*** to remote sensing data (NASA, NOAA, DOD, etc.)
- ***Event-Driven*** Data Delivery based on user inputs or subscriptions
- Enables ***adaptive processing***
- Can be integrated with GLIDER and other tools for ***mining, analysis, and visualization***
- Can be integrated with ***analysis workflow management*** tools

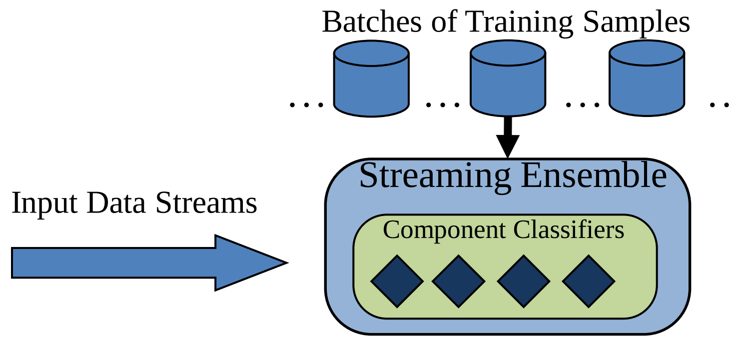
Unclassified

Signature Identification/Characterization using UAH's Coverage Based Ensemble Algorithm (CBEA)

A new ensemble classification method for streaming data developed by UAH

Motivation: Constraints on Streaming Data

- Cannot make multiple passes through all training data
- May only save a small subset of the available samples
- Must make best use of available samples
- Must not forget information provided by old samples
- Can only keep a small number of classifiers
- Must adapt to changing conditions or concepts



CBEA – Training Approach

Characteristics of CBEA

- General purpose ensemble classification method capable of ***incremental learning*** from ***streaming data*** and performing classifications in ***real time*** to provide adaptability
- Handles multiple types of data at ***different resolutions*** of spatial, temporal and other types of information
- Handles ***uneven sampling*** of the classes of interest and the pattern space
  - e.g., if there are not enough truth samples for a particular class or if we are trying to detect a rare event such as nuclear detonations
- Adapts to ***features that change over time***
  - e.g., if the enemy tries to mask or change the weapon signature such as modifying missile propulsion system

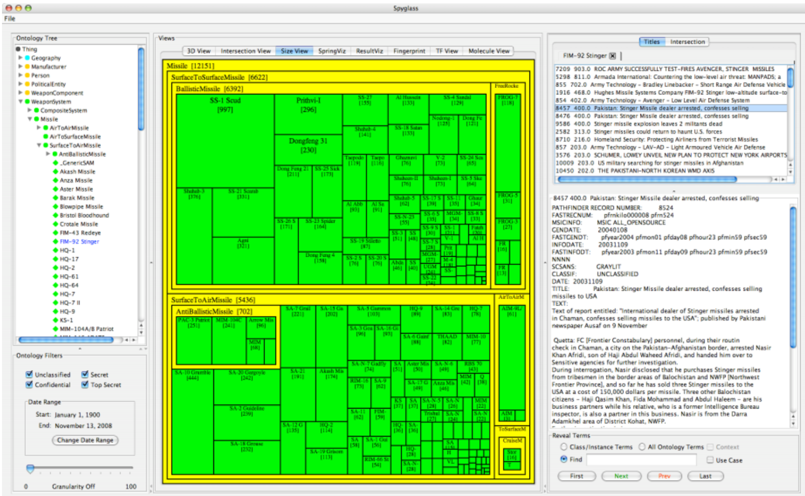
CBEA outperforms Streaming Ensemble Algorithms (SEA) on classification problems with uneven sampling of the pattern space.

SpyGlass
Ontological Text Mining

Multiple Visualization Approaches

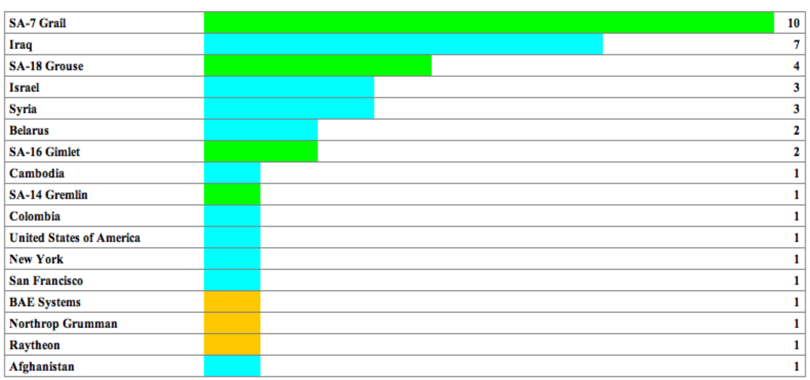
– Show distribution of documents across categories
– Show relationships between documents / categories

Ontology



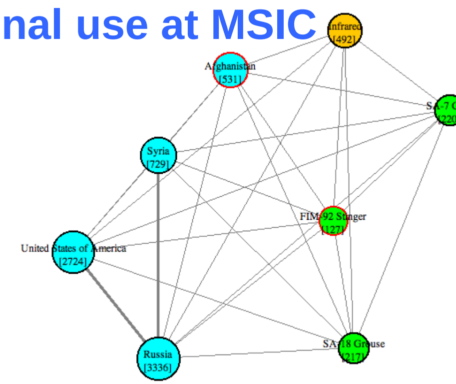
Document Browse View

Size View



Document Fingerprint View

- Captures ***semantic information*** and ***contextual knowledge*** of analysts
- Ontology describes ***entities, concepts and relationships*** in a domain
- ***Constructs document index*** for each term, listing all documents where term occurs
- ***Fast indexing and retrieval***, with high precision and recall
- ***Scores documents*** by number of relevant terms
- ***More powerful*** than simple keyword queries
- ***Possible to reason*** over ontological structures
- **In operational use at MSIC**

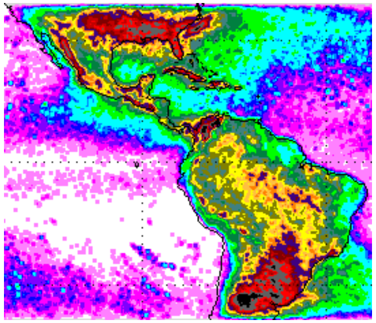


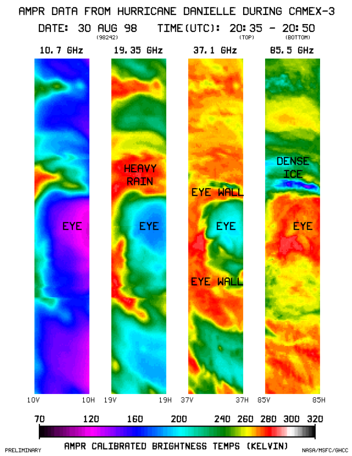
Molecule View of Concept Associations

UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Global Hydrology Resource Center

- **Partnership** between NASA and UAHuntsville to apply **advanced information technologies** to a variety of **science data projects**, thereby enabling research and scientific discovery
- **One of twelve full service NASA data centers** providing data ingest, routine and custom processing, archive, distribution, user support, and science data services
  - *Passive Microwave Data*
    - Fifteen-year inventory of satellite and aircraft based data
  - *Lightning Imaging Sensor Science Computing Facility*
    - National lightning data center for the TRMM Lightning Imaging Sensor and validation networks, satellite lightning observations back to 1973
  - *AMSR-E Science Investigator-led Processing System*
    - Generates swath, daily, and monthly products of precipitation, sea ice, water vapor, cloud water, sea surface temps, etc.
    - Near-real-time processing and distribution capability
  - *Field Campaigns*:
    - Web-based collaboration for science before, during, and after experiments. Data acquisition, integration, archive and distribution
    - CAMEX (1998, 2001), ACES (2002), TCSP (2005), NAMMA (2006), TC4 (2007), ARCTAS (2008), GRIP (2010), MC3E (2011)





Science Data Provenance

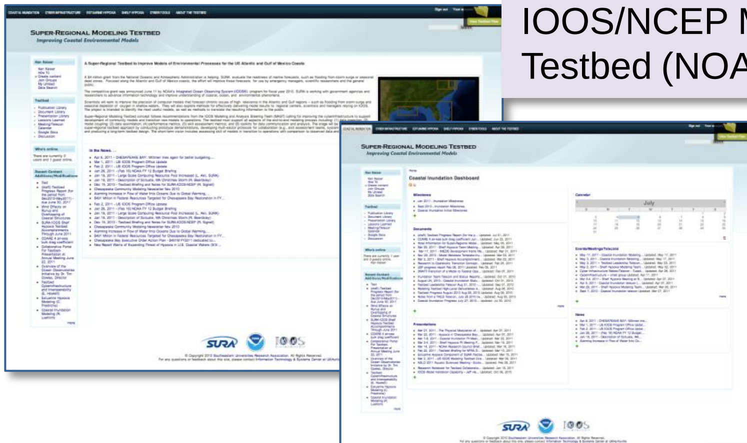
- ***Data lineage***: product generation "recipe" (data inputs, software and hardware). Try to capture at moment of creation where most knowledge about product is present.
- Additional ***knowledge*** about science algorithms, instrument variations, etc.
- ***Lots of information already available, but scattered across multiple locations***
  - Processing system configuration
  - Dataset and file level metadata
  - Processing history information
  - Quality assurance information
  - Software documentation (e.g., algorithm theoretical basis documents, release notes)
  - Data documentation (e.g., guide documents, README files)
- ***Instant Karma tool aims to collate and organize information from multiple sources at earliest stage of data's life.***

Noesis: Ontology Based Search and Resource Aggregation Tool

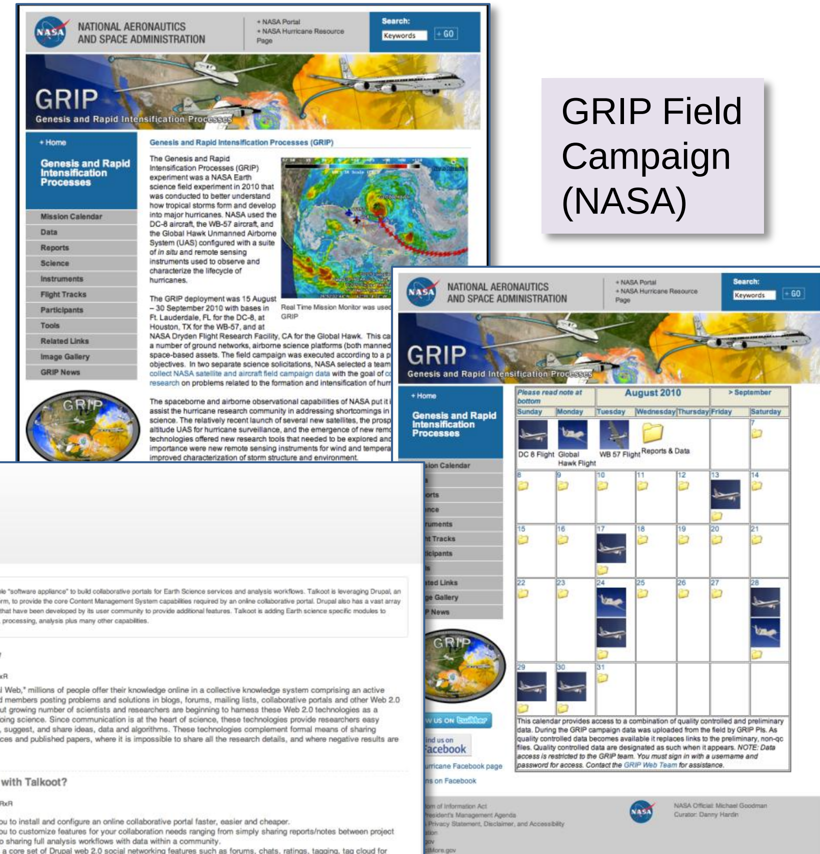
Noesis, developed for LEAD, is also being used by NASA ESIP Federation, Gulf of Mexico Regional Collaboration Project and others

- Information searching can be enhanced considerably through the ***integration of ontologies into search systems*** – Noesis is one example
- Ontologies allow searches to be conducted in terms of concepts (unambiguous denotations of entities of interest) rather than words, to reduce the ambiguity problem
- The use of ontologies while searching has only minimal effect on precision but has significant effect on recall
- Domain information can be used to search heterogeneous databases to collect and aggregate results
- Our ongoing research is focusing on using ontologies in both ***query expansion*** and ***ranking results*** using different metrics and techniques (relevance aura, text mining approaches, etc.)

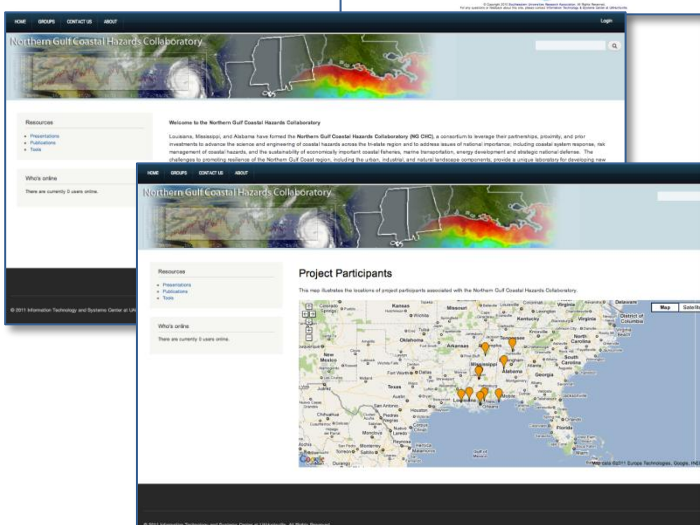
Community Portal Technology



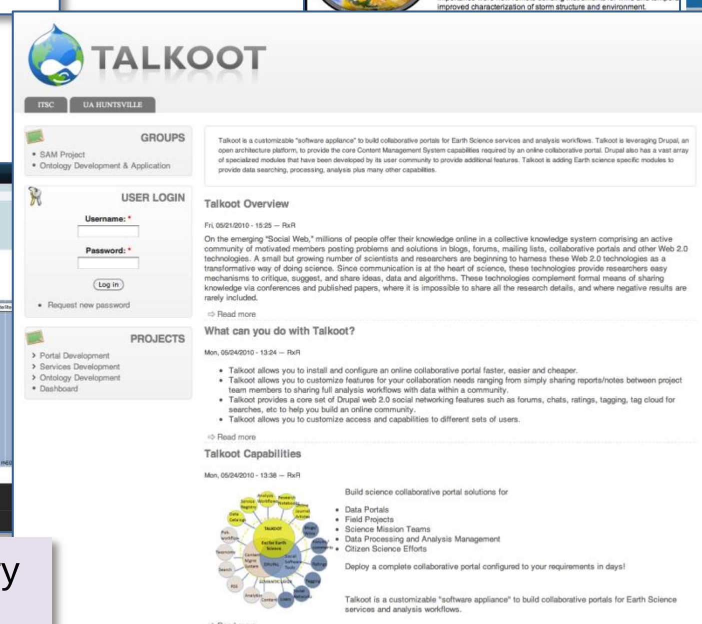
IOOS/NCEP Modeling Testbed (NOAA)



GRIP Field Campaign (NASA)



Northern Gulf Coastal Hazards Collaboratory (NSF)



Talkoot Collaborative Framework

Northern Gulf Coastal Hazards Collaboratory

The Northern Gulf Coastal Hazards Collaboratory (NG-CHC), funded by NSF EPSCoR

- A cyberinfrastructure is being created to catalyze collaborative research and education and reduce risks to coastal vulnerabilities.
- Researchers in Louisiana, Mississippi, and Alabama are working to advance the science and engineering of coastal hazards across the region and address problems of national importance, including engineering design, coastal system response, and risk management of coastal hazards.
- A collaborative environment provides needed capabilities for enhancing linkages between modeling and observations in a multidisciplinary environment and allows researchers to organize, discover and share information about data, models, tools and other resources; discuss project activities and results; view publications, presentations and other documents; and track the history of project activities.

From the Multi-angle Imaging SpectroRadiometer highlights coastal areas of four states along the Gulf of Mexico: Louisiana, Mississippi, Alabama, and part of the Florida panhandle. The images were acquired on October 15, 2001 and represent an area of 345 kilometers x 315 kilometers.
Credits: NASA/GSFC/LaRC/JPL, MISR Team

Sediment plume of the Mississippi River
Credits: SeaWiFS Project, NASA/Goddard Space Flight Center, and ORBIMAGE

SERVIR

An Environmental Monitoring and Decision Support System for Central America

UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Earth Observatories

Electronic Transfer

SERVIR Node @ NSSTC

(NASA/MSFC and U. Alabama in Huntsville)

Product Generation System
- Ingest Data
- Subset Data Over C. Amer.
- Mine Data for Events
- Generate Products

Web Server
servir.nsstc.nasa.gov
- Distribute Products
- Archive Products

Visualization System

Source Data Archive

Product Archive

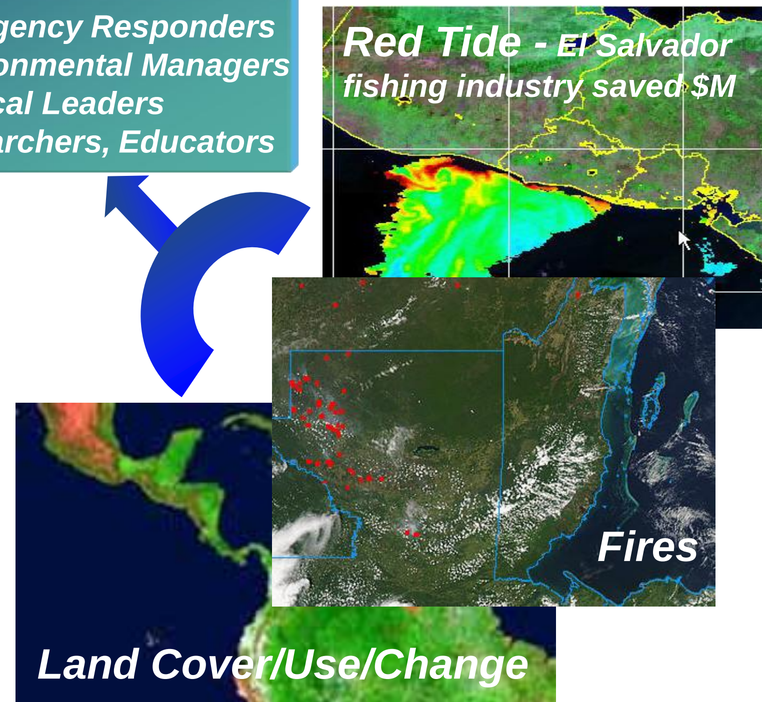
Data & Algorithms

SERVIR Partners

Central American Commission for Environment and Development

- Emergency Responders
- Environmental Managers
- Political Leaders
- Researchers, Educators

Environmental Monitoring & Decision Support Products



Rapid Response
ftp, e-mail, etc.

SERVIR Node in Panama

University of Arkansas
(World Bank Funding)

- Geographic Info Systems
- Decision Support Systems
- Environmental Data from Central American countries

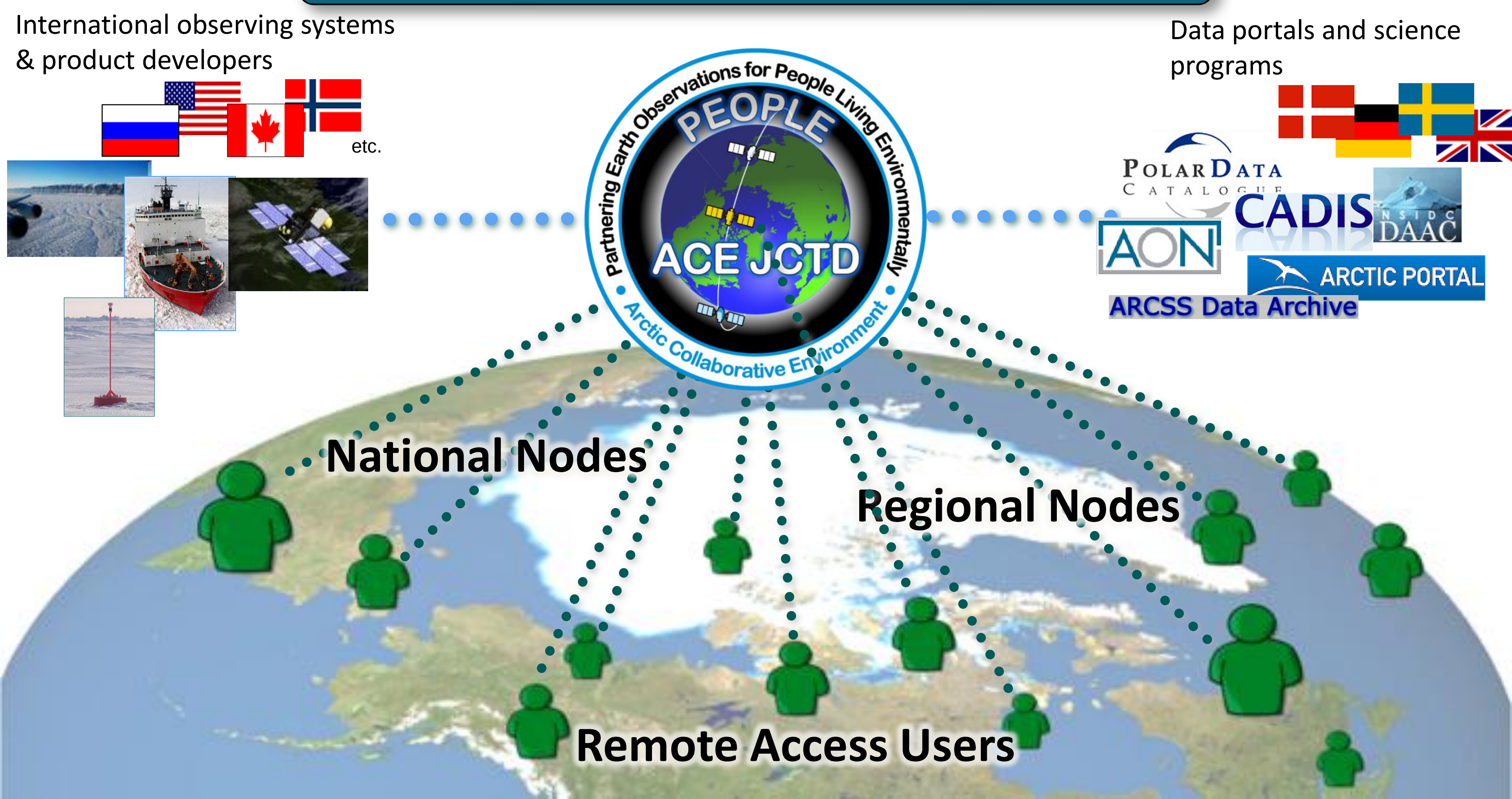
Goals

- Rapid Response
- Corridor Preservation
- Species Preservation
- Sustained Development
- Better Living Conditions
- Policy Changes

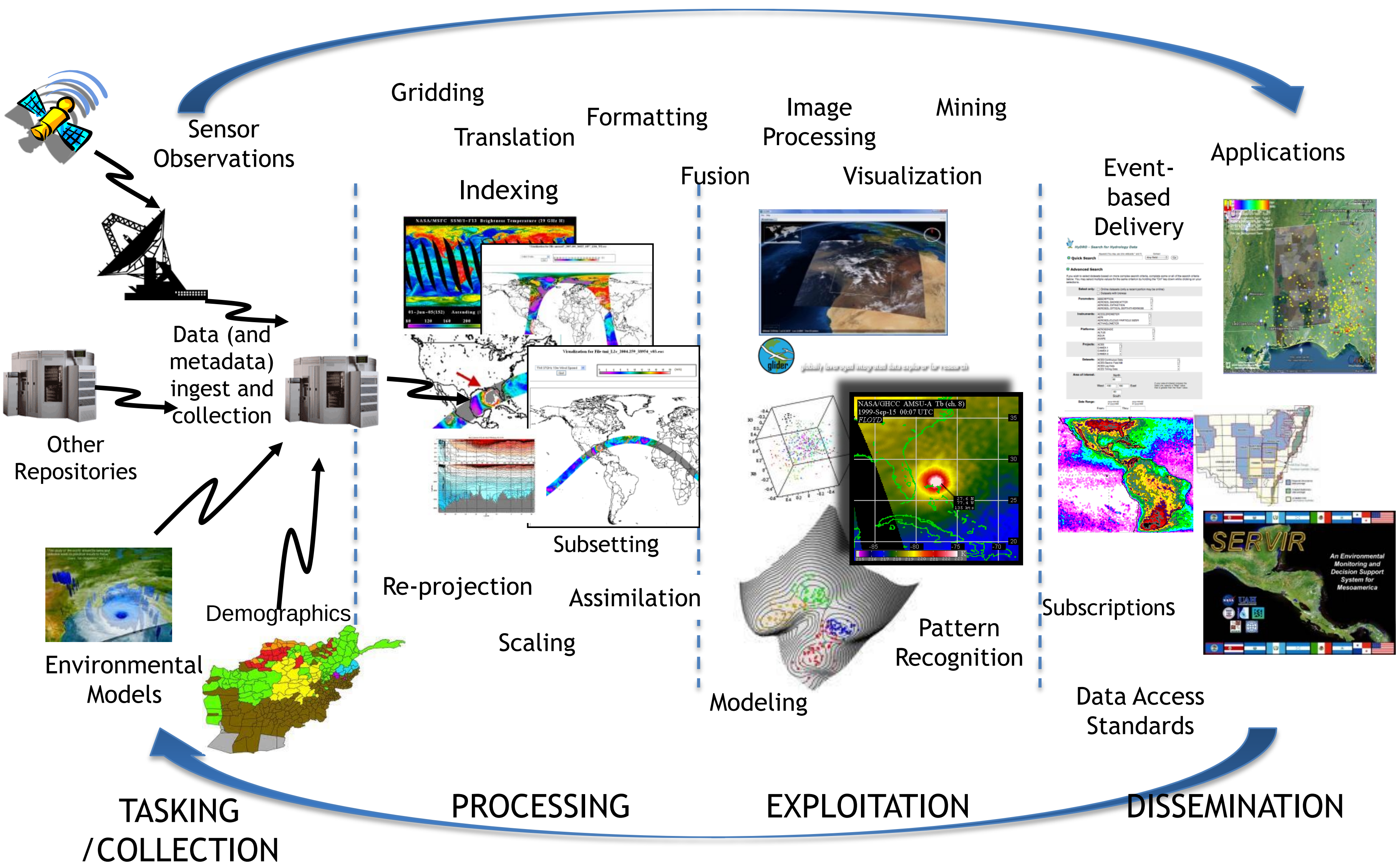
UAHuntsville
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Links with other efforts: PEOPLE-ACE JCTD

An open source, web-based, multi-national environmental monitoring, research, and decisions support system to enable development of advanced value-added products



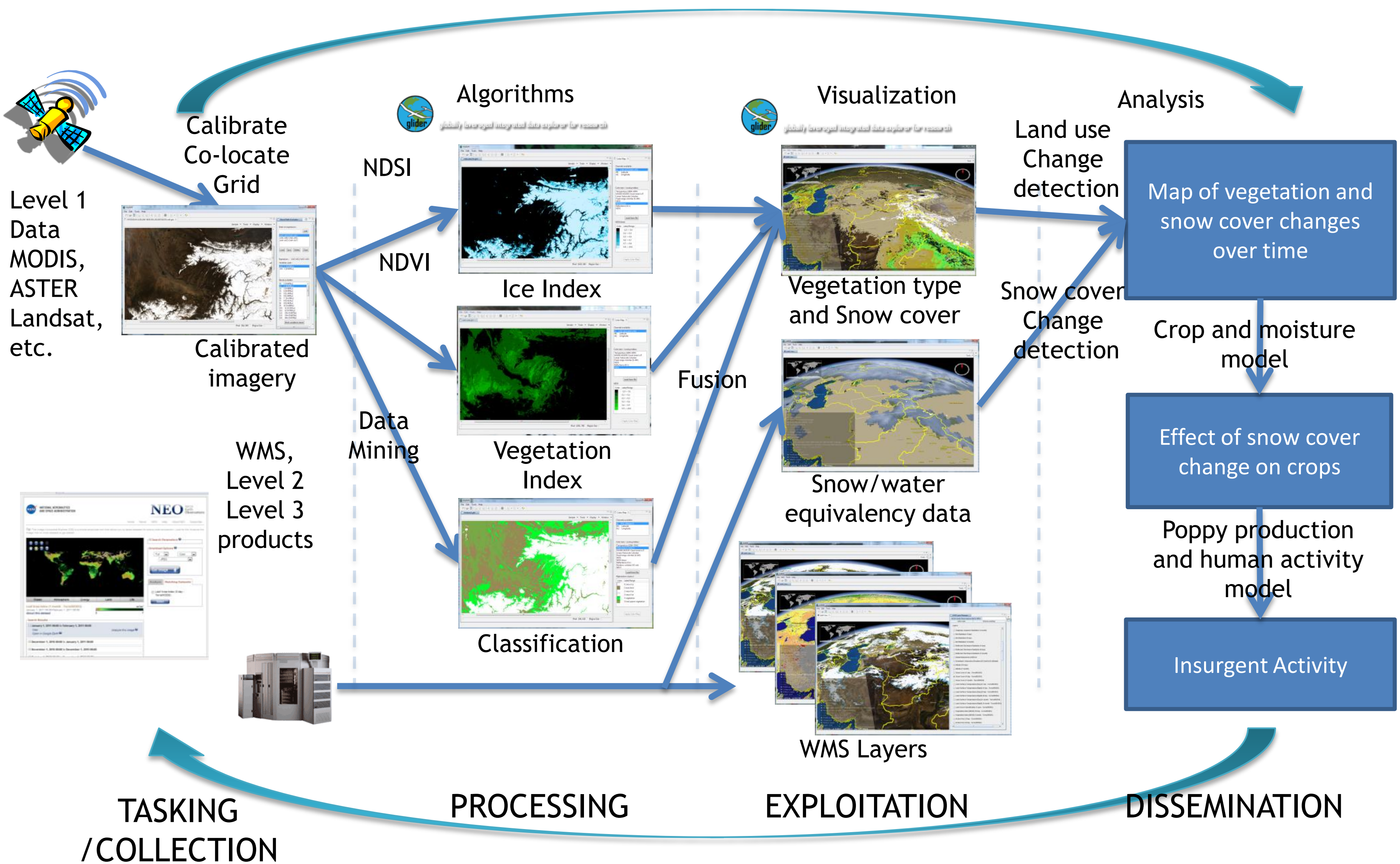
Conceptual Framework for Multi-source, Multi-function Analysis



Mountain Snow Cover in Afghanistan

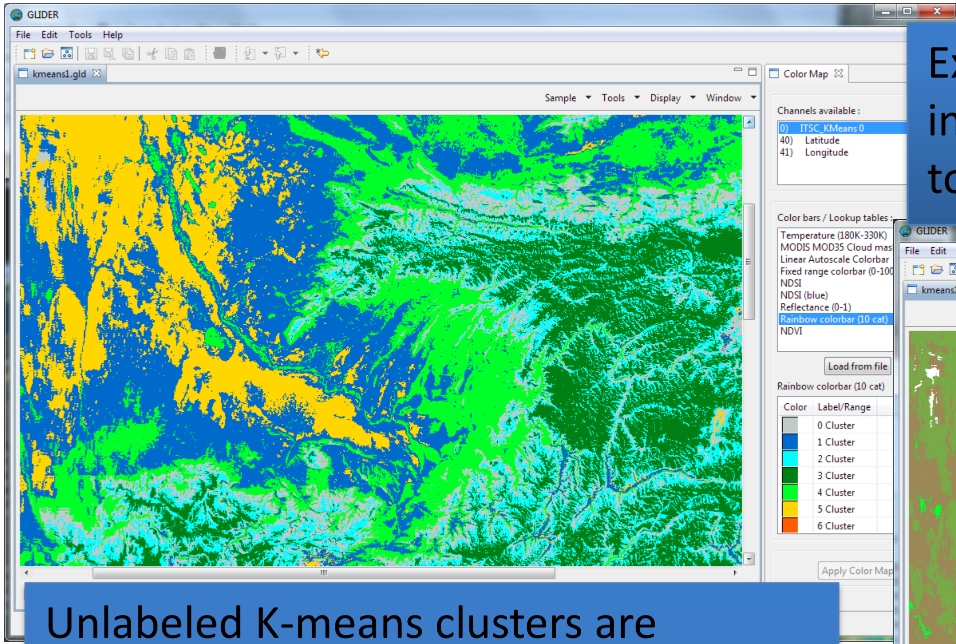
- Coupled System Hypothesis: Winter mountain snow cover and poppy crop production
- Assumed human impact: Increase in snow cover leads to increased poppy and food crop yields
  - Poppy crop production funds insurgent activity
  - Increase in poppy crop leads to increased insurgent activity
  - Decrease in food crops causes instability, aiding insurgents
- Possible mitigation: Analysis of possible crop yields may allow planning for counter insurgent activity

Mountain Snow Cover Process

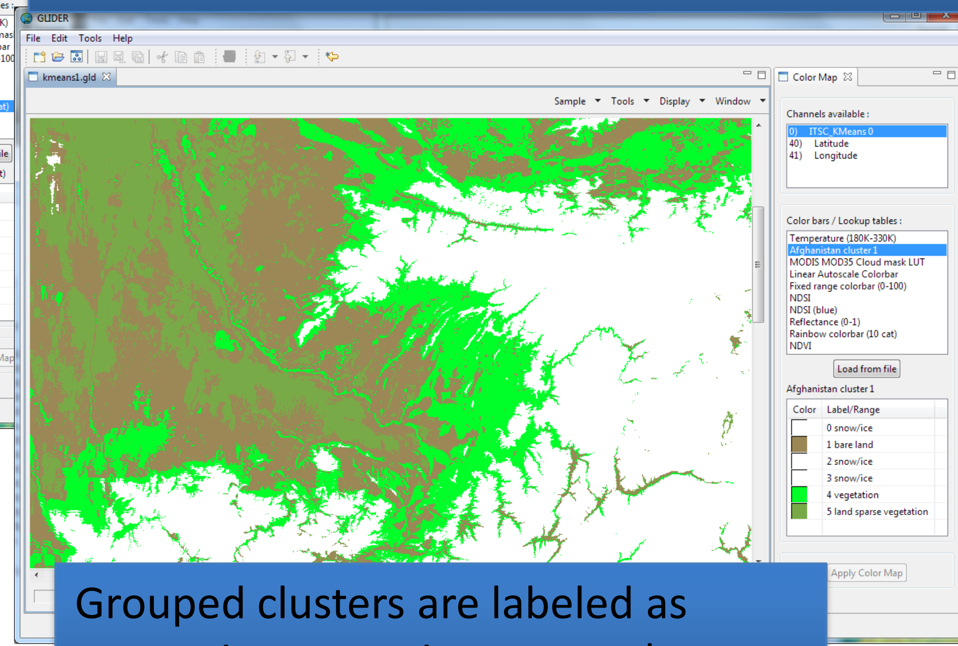


Thematic Map of Surface Types Using Glider

- Clustering algorithms may be used to provide the classification capabilities required to determine crop type in Afghanistan



Example using K-means to cluster regions of interest together. Glider is providing the basic tools for this analysis.



Unlabeled K-means clusters are displayed as different colors. These clusters can be grouped and labeled as different land/surface types.

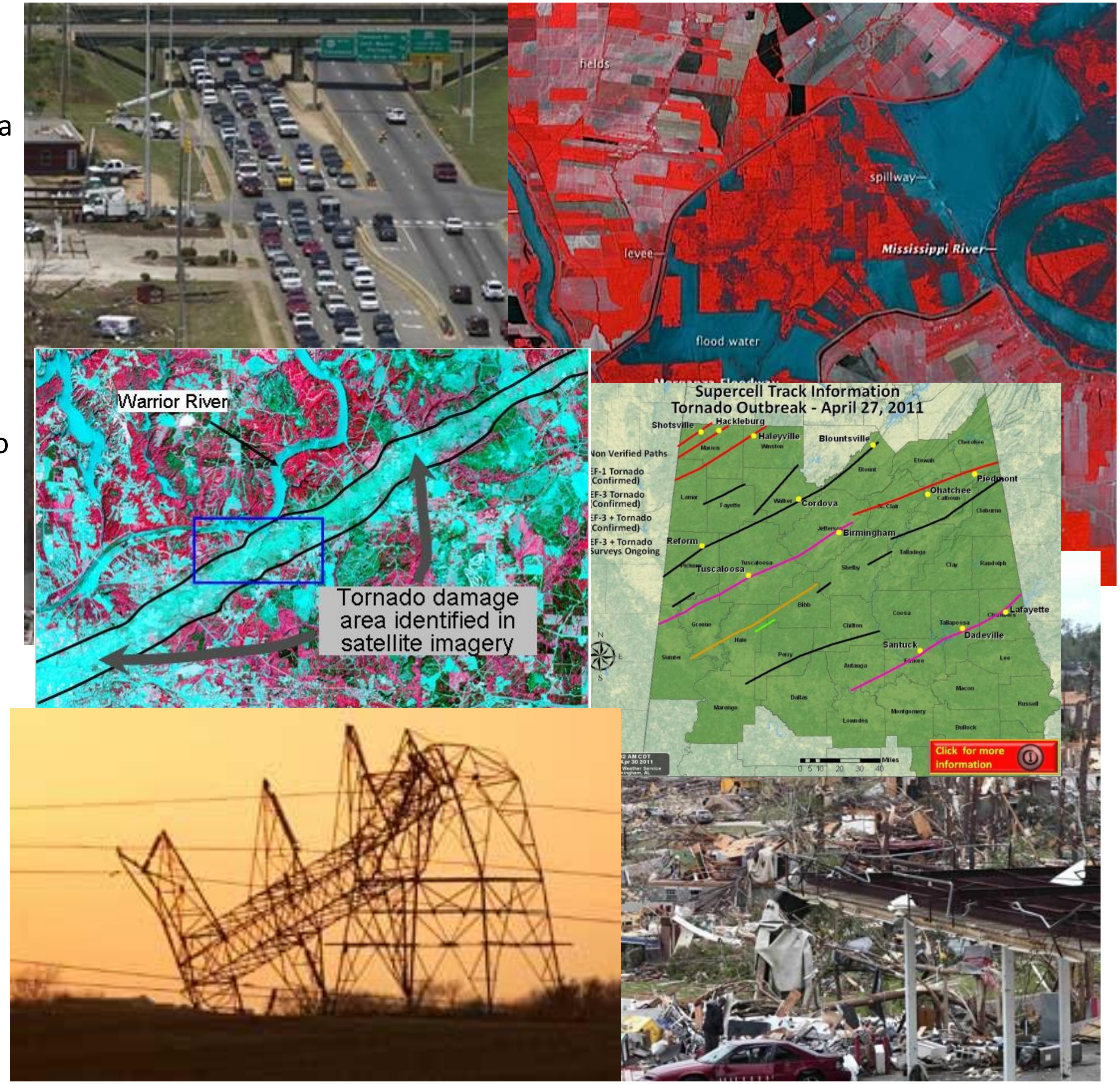
Grouped clusters are labeled as vegetation types, ice, etc. and displayed as specified colors.

Impact of Natural Disasters

- Tornadoes and floods devastated large portions of Alabama, Louisiana and Missouri in April and May 2011
  - Power outages
  - Drinking Water
  - Refugees
  - Disease outbreaks

- Catastrophic events such as tornado outbreaks have profound national security implications
  - Death and destruction
  - Over-stretched law enforcement and emergency responders
  - Food, water, fuel shortages
  - Civil unrest, rioting, price gouging, looting

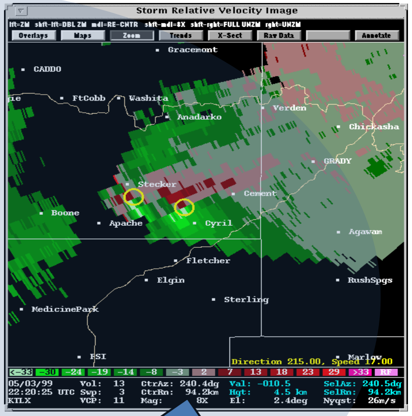
- Consider the impact in less developed areas of the world



Information Technology and Systems Center

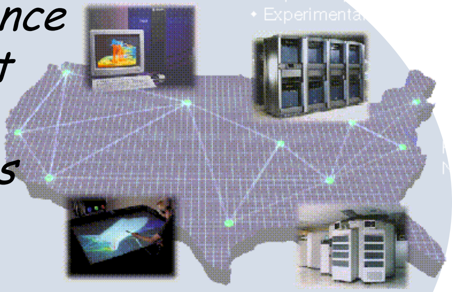
Data Mining

- Clustering
- Classification
- Anomaly Detection
- Association Rules
- Pattern Recognition
- Feature Selection
- Image Processing
- Text
- Texture
- Ontology driven



Cybersecurity

- IA Center of Excellence
- Identity Management
- Metrics
- Trustworthy Systems
- Policy Development
- Risk Management
- Vulnerability Analysis
- Situational Awareness
- Privacy
- FISMA compliance

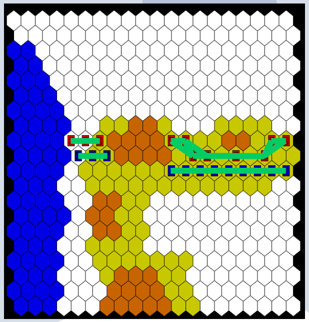


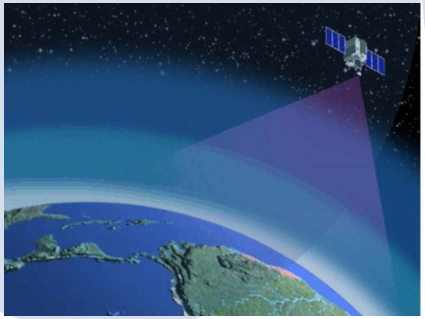
On-Demand Processing

- Real-time operations
- MultiSensor Fusion
- Signature Intelligence
- Unmanned Systems
- Sensor Networks

Advanced Computational Methods

- Artificial Intelligence for Wargaming
- Semantics
- Training Systems
- Path Determination
- Knowledge Networking
- Data Exploitation
- Adaptive Processing





Strategic and Tactical Coordination

- Collaborative Environments
- Remote Mission Management
- Information Acquisition and Integration
- Urban Environments
- Emergency Response