Data Mining, Remote Sensing and Biodiversity

Presented at
Biodiversity and Ecosystem Informatics Working Group
February 15, 2005

Sara J. Graves, Ph.D.

Director, Information Technology and Systems Center

University Professor, Computer Science Department

University of Alabama in Huntsville

Director, Information Technology Research Center

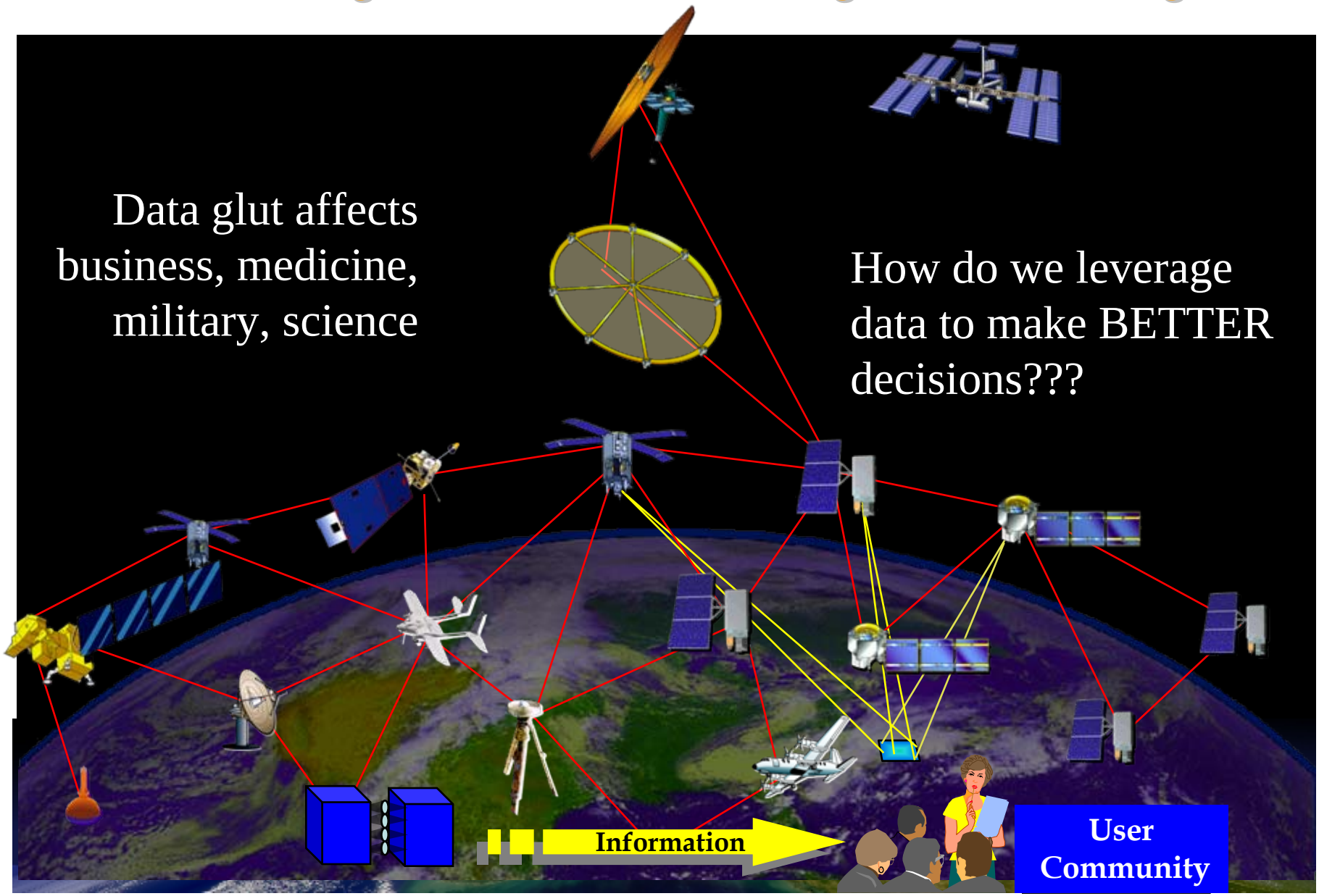
National Space Science and Technology Center

256-824-6064

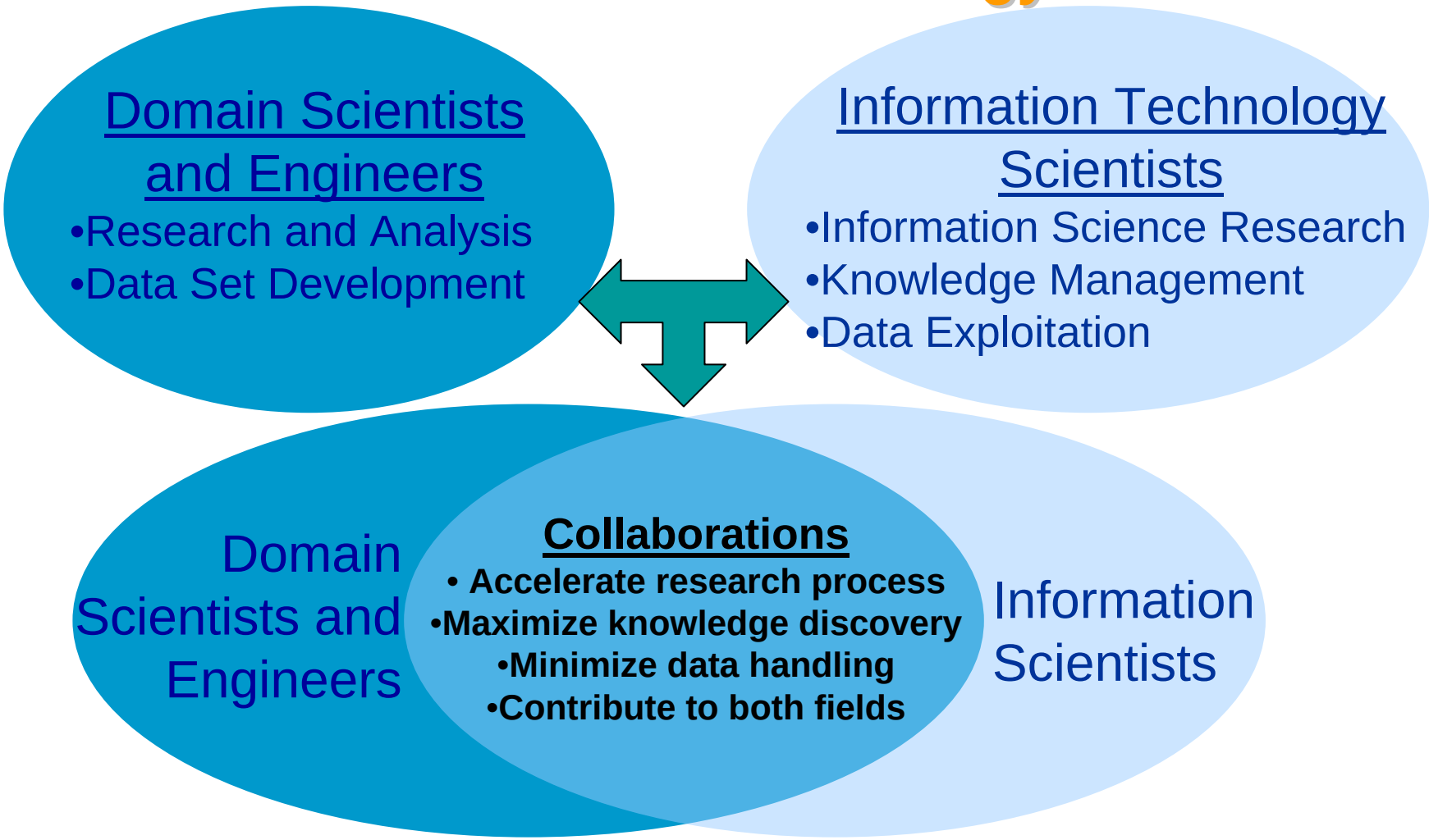
sgraves@itsc.uah.edu

http://www.itsc.uah.edu

"…drowning in data but starving for knowledge"



Data Usability Success Builds on the Integration of Various User Domains and Information Technology



Harnessing Information Technologies

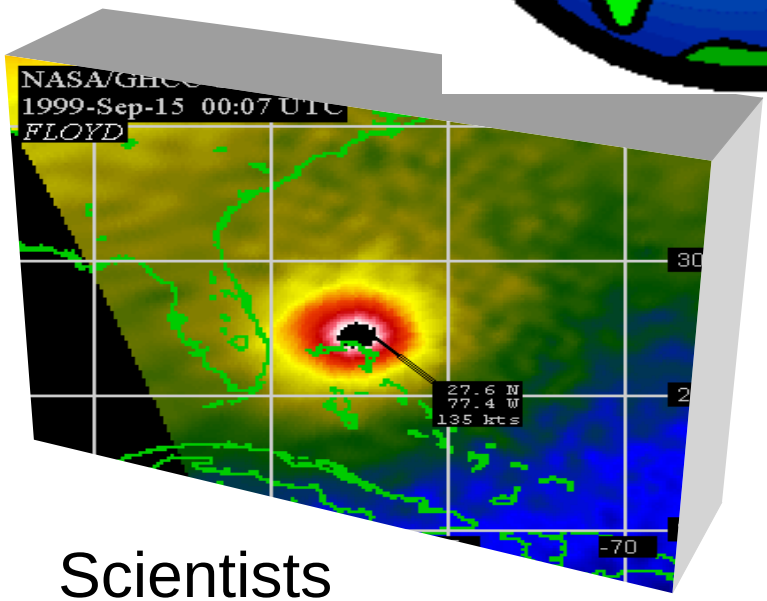
- *Data organization and information management* for archival and analysis
- *Data preparation (e.g., subsetting)* for improved data usability
- *Data Mining* in real-time and for post run analysis
- *Interchange Technologies* for improved data exploitation
- *Semantics* to transform data exploitation via intelligent automated processing
- *On-board processing* to support timeliness of data analysis and delivery and adaptive mission operations

Scientific Analysis	Data Mining

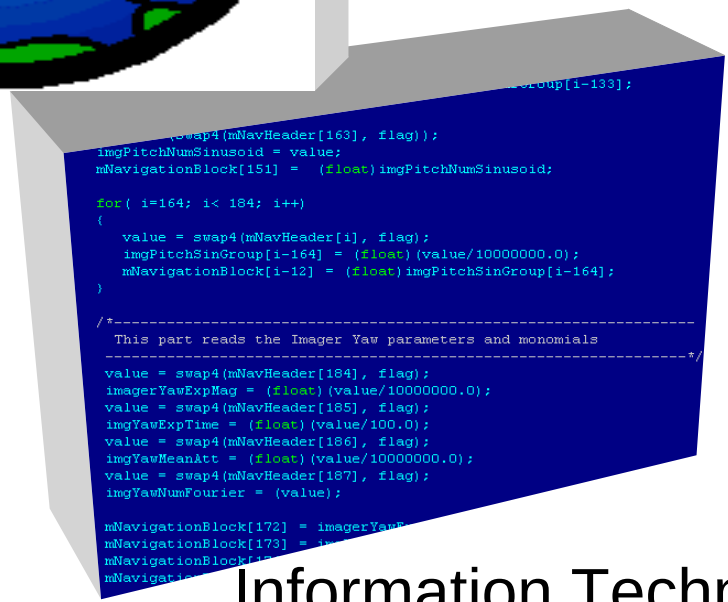
• Harnesses human analysis capabilities
  – Highly creative
• Based on theory and hypothesis formulation
  – Physical basis is normally used for algorithms
• Drawing insights about the underlying phenomena
• Rapidly widening gap between data collection capabilities and the ability to analyze data
• Potential of vast amounts of data to be unused
	• Provides automation of the analysis process
	• Can be used for dimensionality reduction when manual examination of data is impossible
	• Can have limitations
	  – May not utilize domain knowledge
	  – May be difficult to prove validity of the results
	• There may not be a physical basis
	• Should be viewed as complimentary tool and not a replacement for scientific analysis

Key Collaborators

End Users

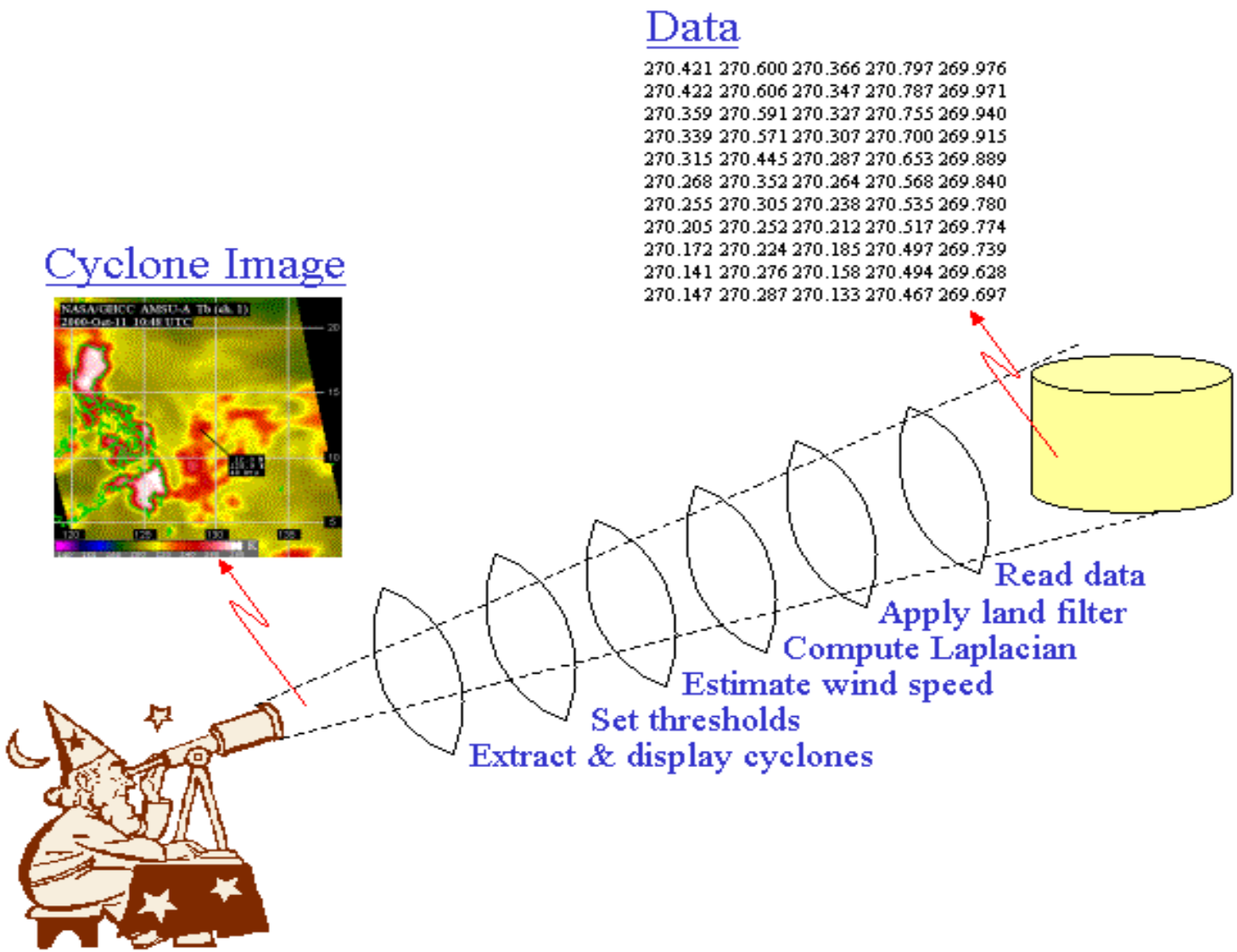


Scientists

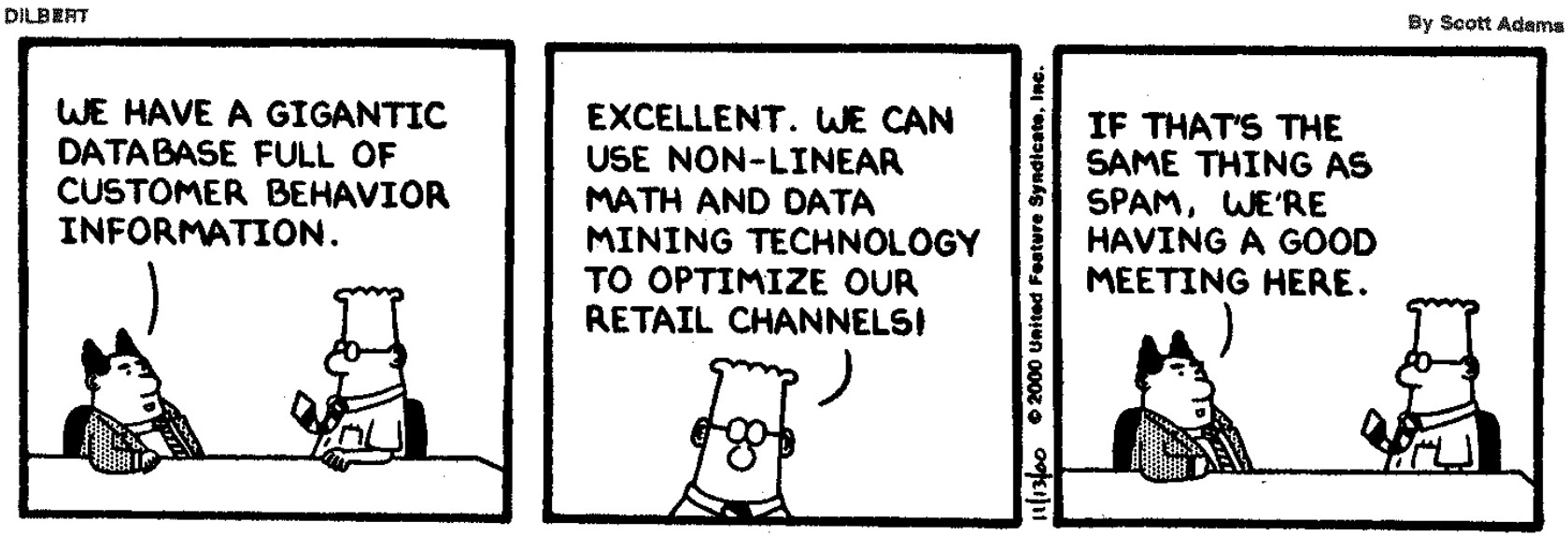


Information Technology Specialists

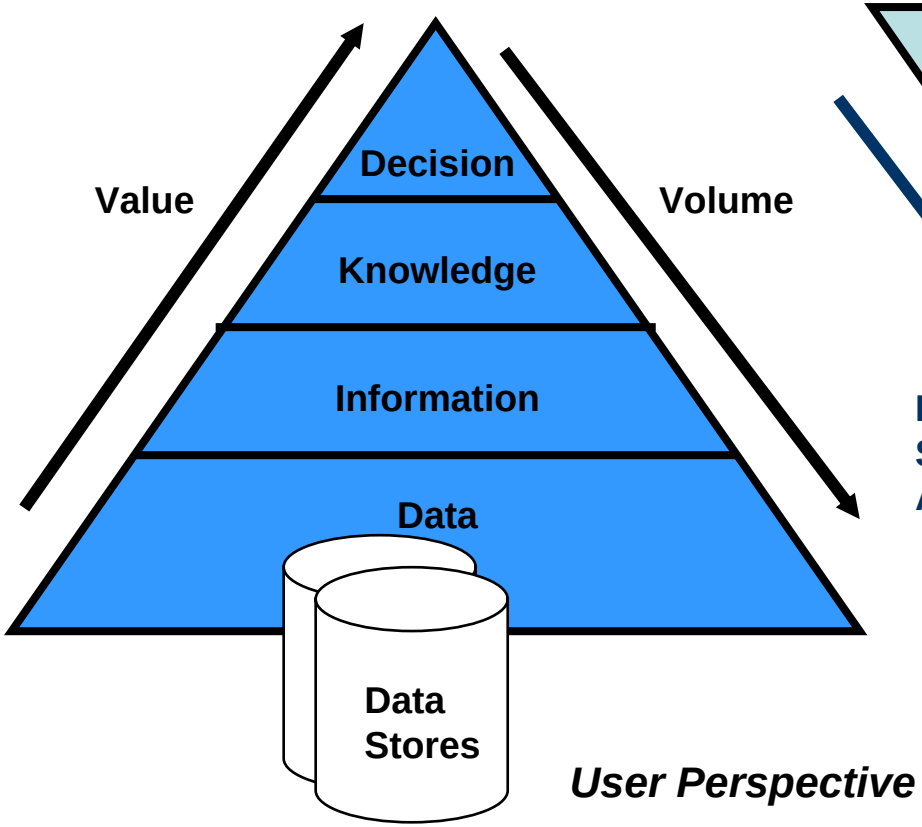
Enhancing Data Usability: Focusing on finding information in data



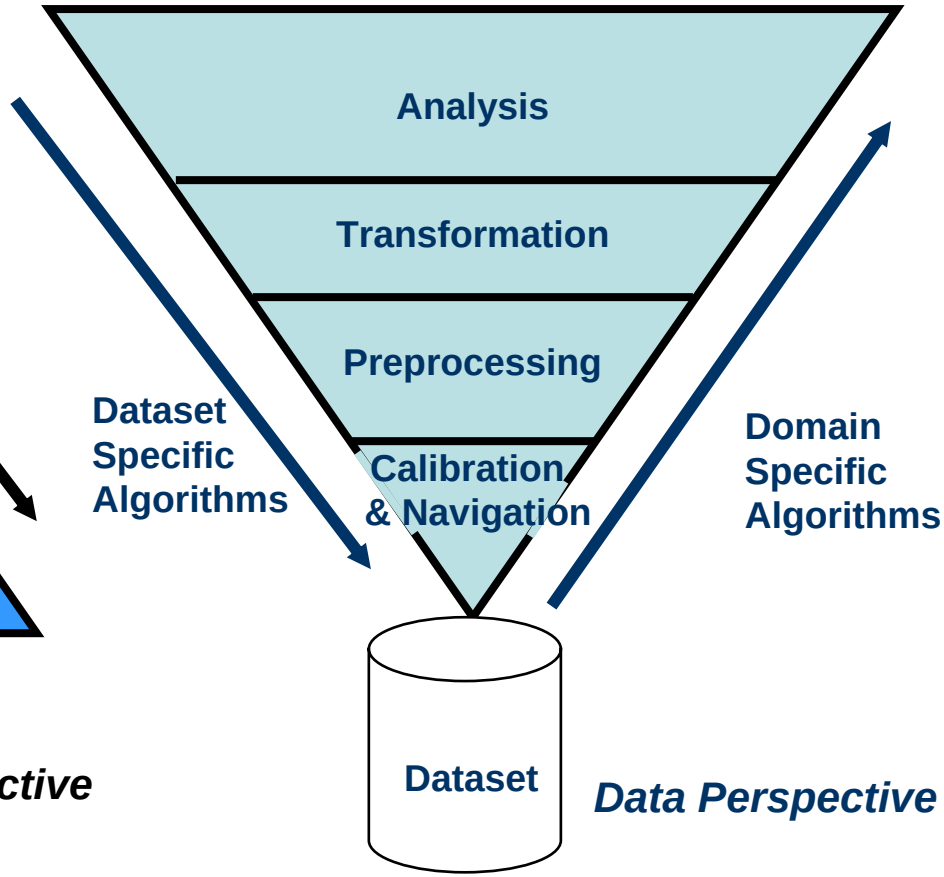
Data Misunderstanding ?



User Perspective and Data Perspective of the Data Mining Process

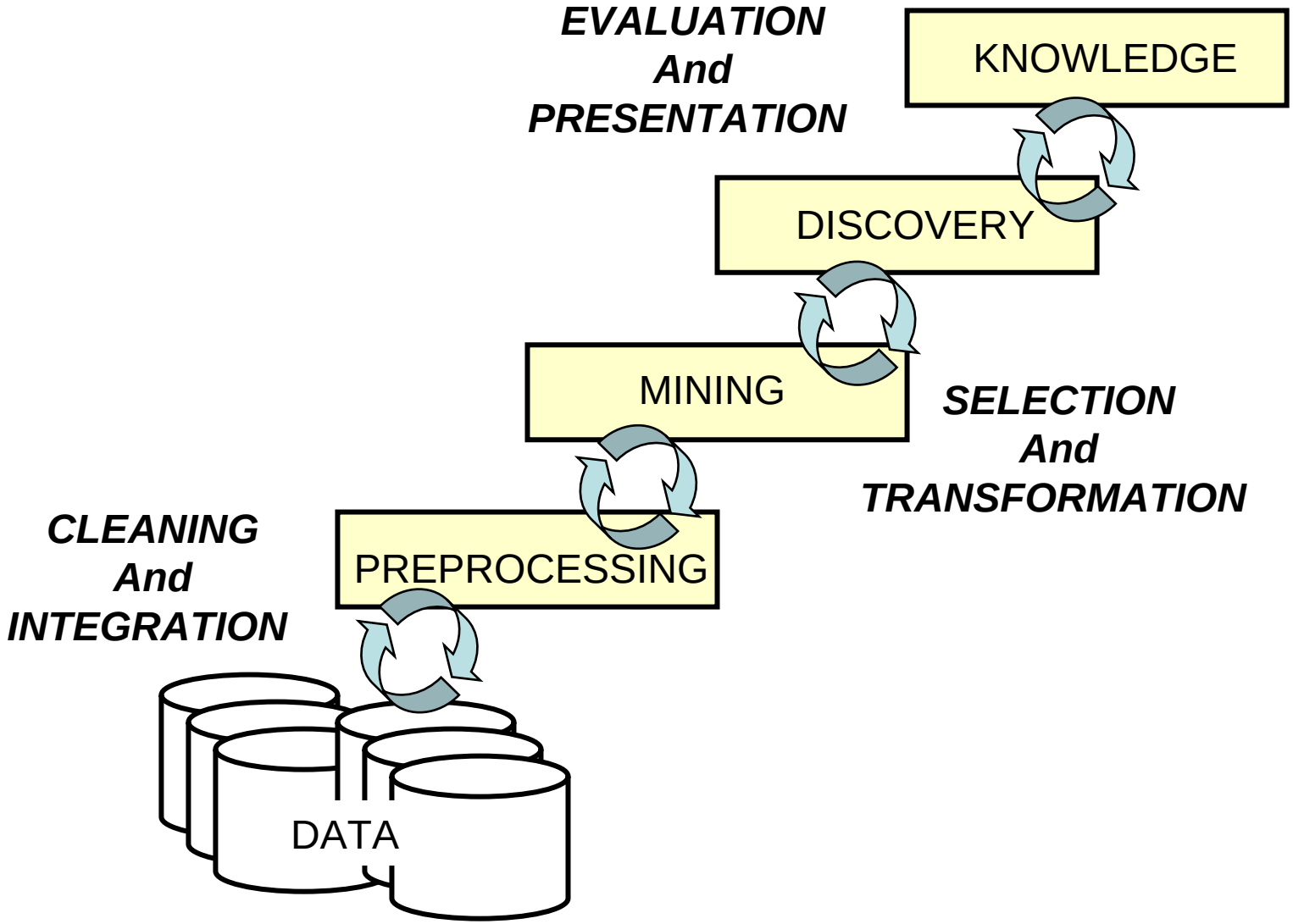


User Perspective



Data Perspective

Iterative Nature of the Data Mining Process



Issues

- Mining
  - Feature extraction
  - Finding anomalies in the data
  - Understandability of the derived model
  - Utilizing domain knowledge effectively
- Scientific Data Mining Environment
  - Ease of use
  - Automation of analysis process
  - Adaptable to new science questions
  - Plug in favorite analysis tools

Mining Environments

Multiple Configurations

- Complete System (Client and Engine)
- Mining Engine (User provides its own client)
- Application Specific Mining Systems
- Operations Tool Kit
- Stand Alone Mining Algorithms

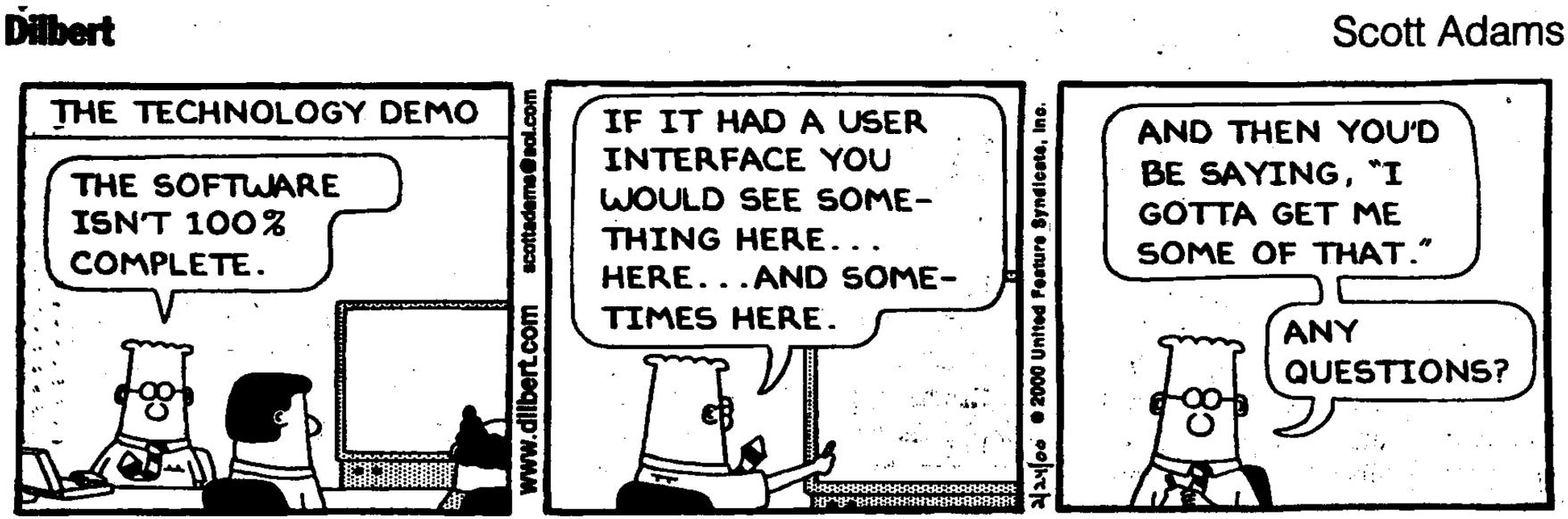
Distributed/Federated/Grid Mining

- Distributed services
- Distributed data
- Chaining using Interchange Technologies

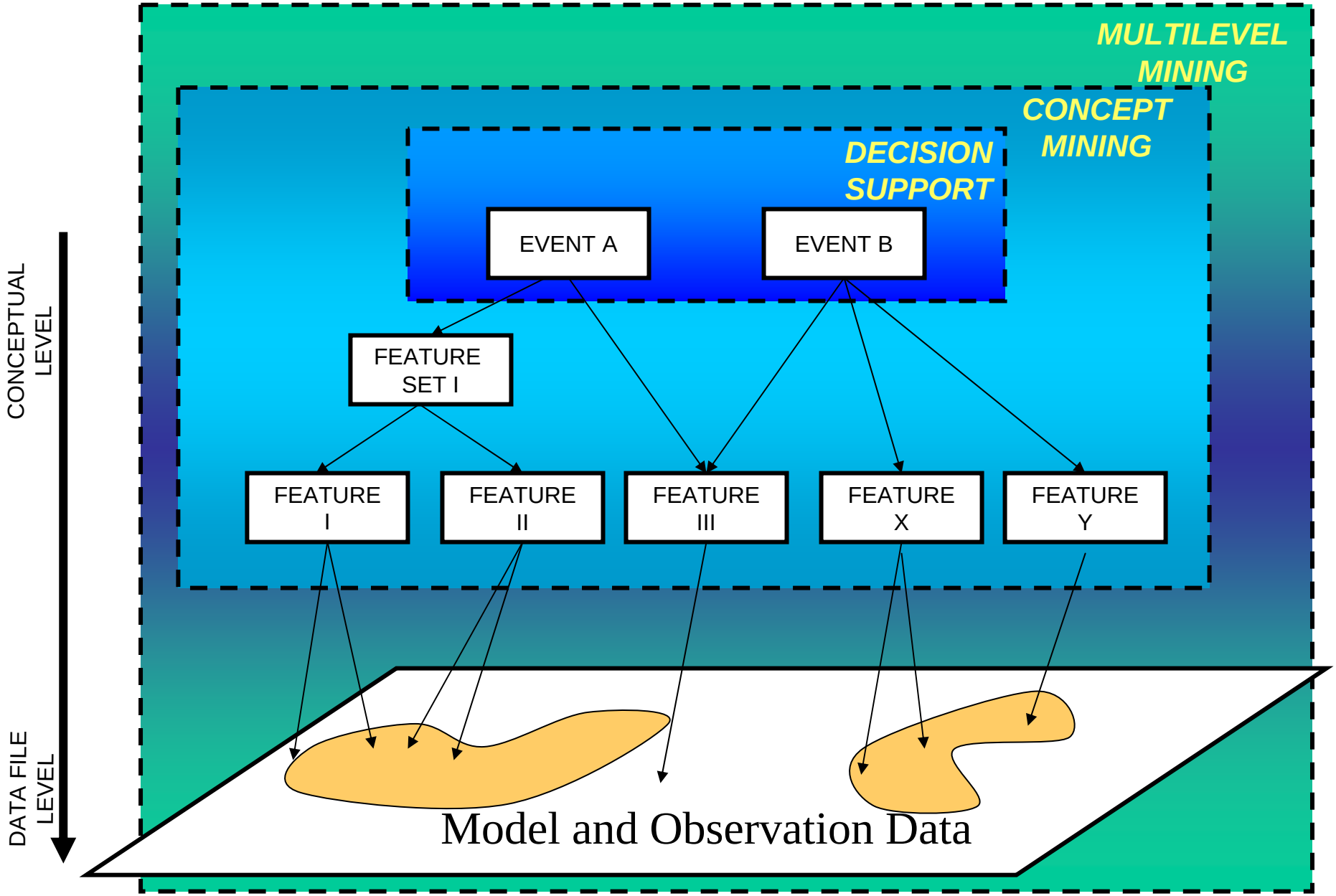
On-board Mining

- Real time and distributed mining
- Processing environment constraints

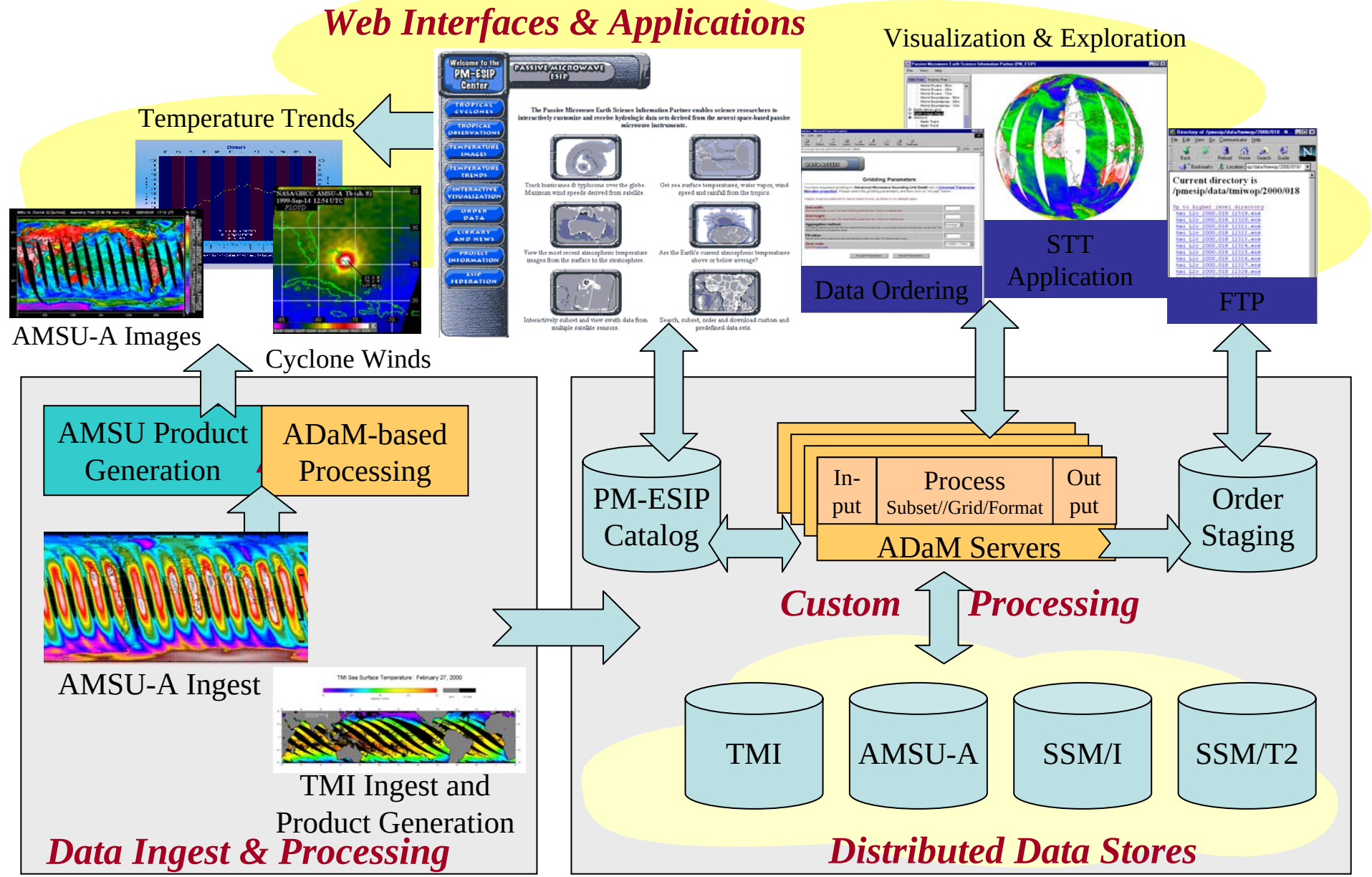
Providing User Interfaces ?



Concept Hierarchy for Data Mining and Fusion

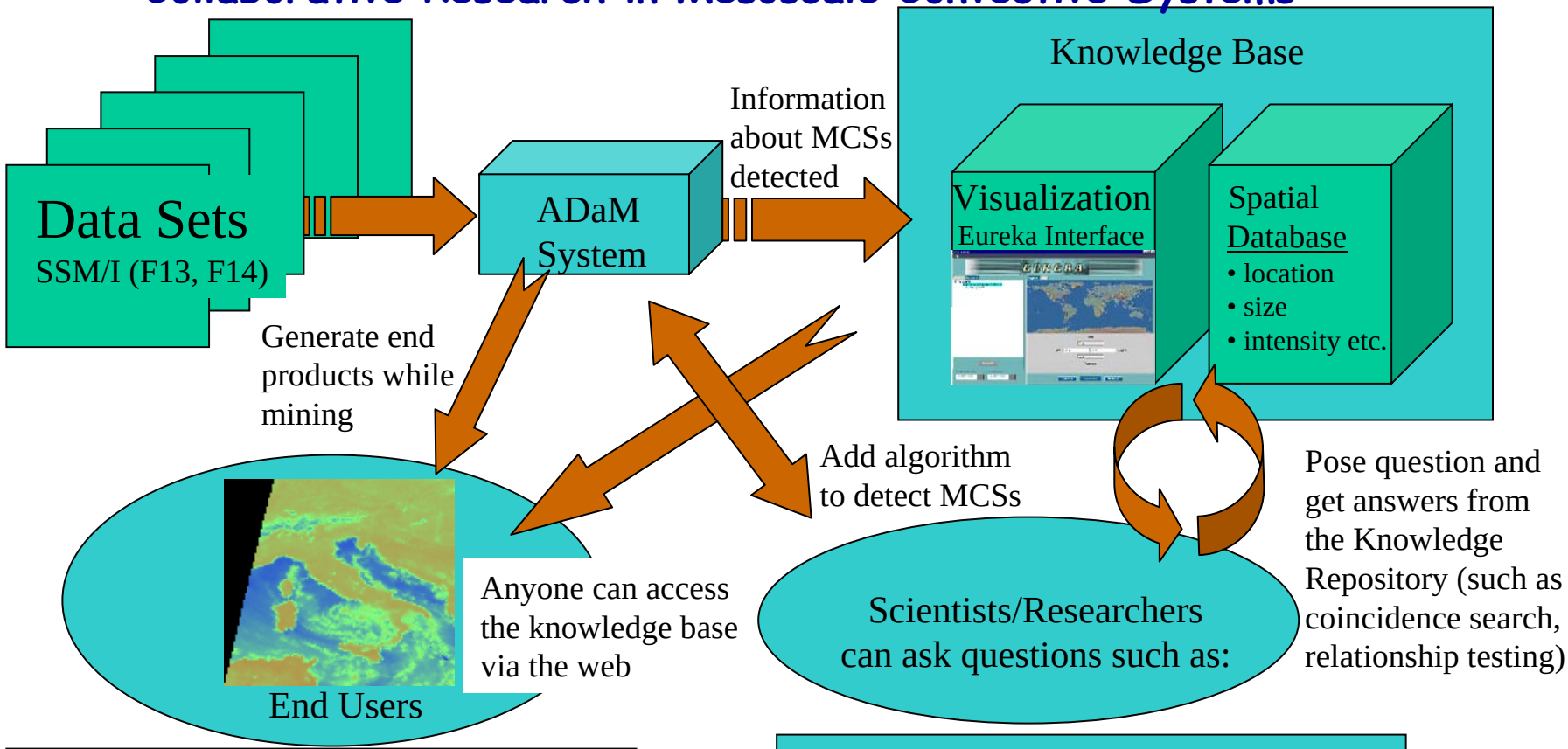


Multiple Mining Environments: Passive Microwave ESIP Information System



Earth Science Example of Developing a Knowledge Network:

Collaborative Research in Mesoscale Convective Systems



- Generate information useful to the general public (students, researchers, policy makers etc)
  - •Images
  - •Forecast aids
  - •General Science information
- Answer the practical side of the problem

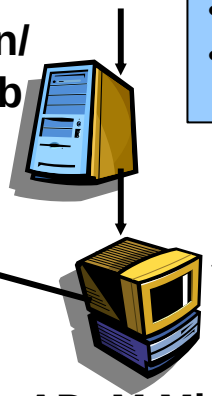
- What is the latitudinal distribution of MCSs?
- Which continent has more MCSs?
- What is the seasonal distribution of MCSs?
- What is the relationship between the number of MCSs and their intensity?

Data Discovery ?

Mining on Data Ingest: Tropical Cyclone Detection

Advanced Microwave Sounding Unit (AMSU-A) Data

Calibration/ Limb Correction/ Converted to Tb

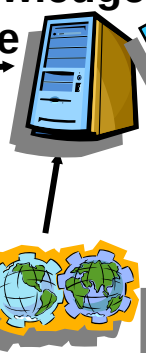


ADaM Mining Environment

Data Archive

Mining Plan:

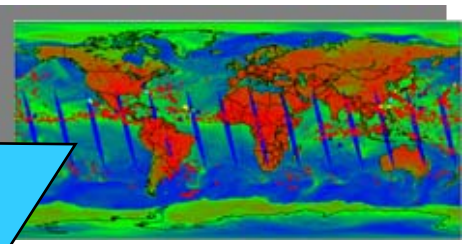
- **Water cover mask to eliminate land**
- **Laplacian filter to compute temperature gradients**
- **Science Algorithm to estimate wind speed**
- **Contiguous regions with wind speeds above a desired threshold identified**
- **Additional test to eliminate false positives**
- **Maximum wind speed and location produced**

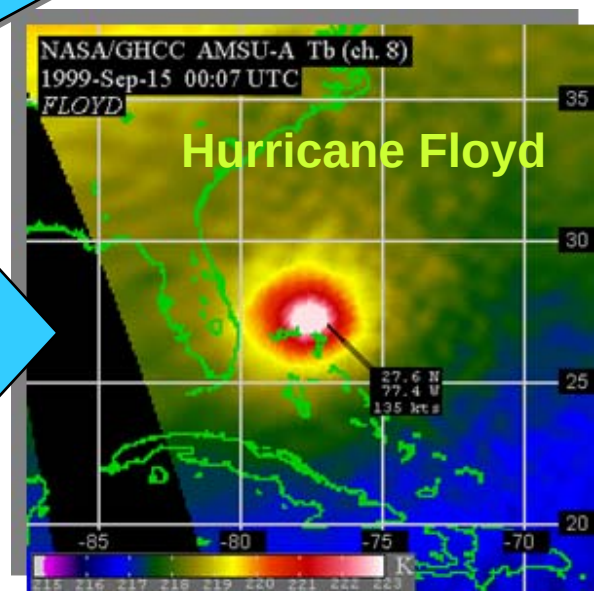


Knowledge Base

Result

Further Analysis



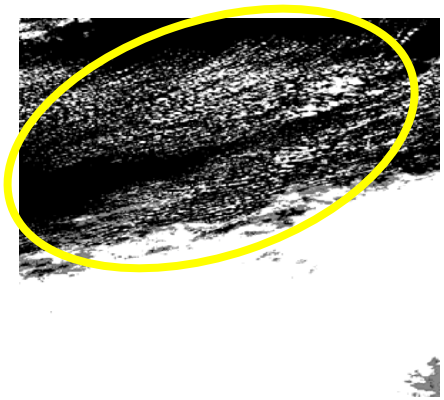


Results are placed on the web, made available to National Hurricane Center & Joint Typhoon Warning Center, and stored for further analysis

http://pm-esip.msfc.nasa.gov/cyclone

Cumulus Cloud Classification

- Science Rationale: Man-made changes to land use cause changes in weather patterns, especially cumulus clouds
- ADaM allows comparison of many different classification techniques based on accuracy of detection and amount of time required to classify
- Best algorithm can be used to create cloud mask product

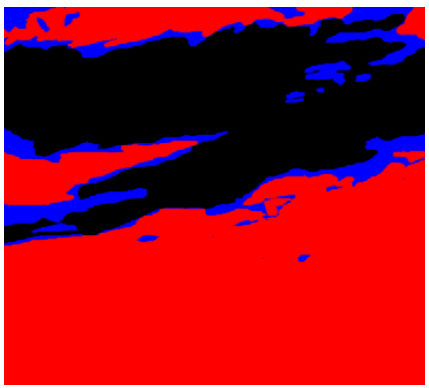


Original

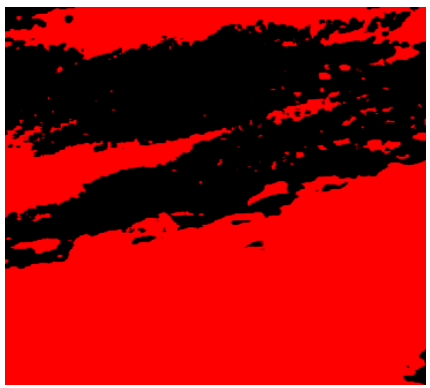
GLRL

Association Rules

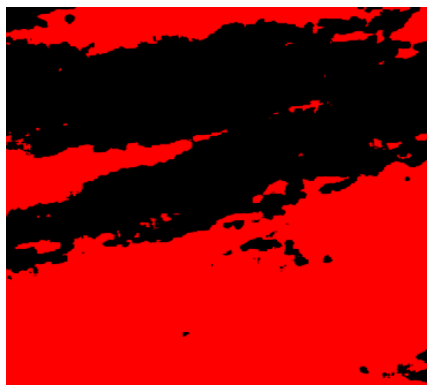
GLCM



Expert Labeled



Sobel

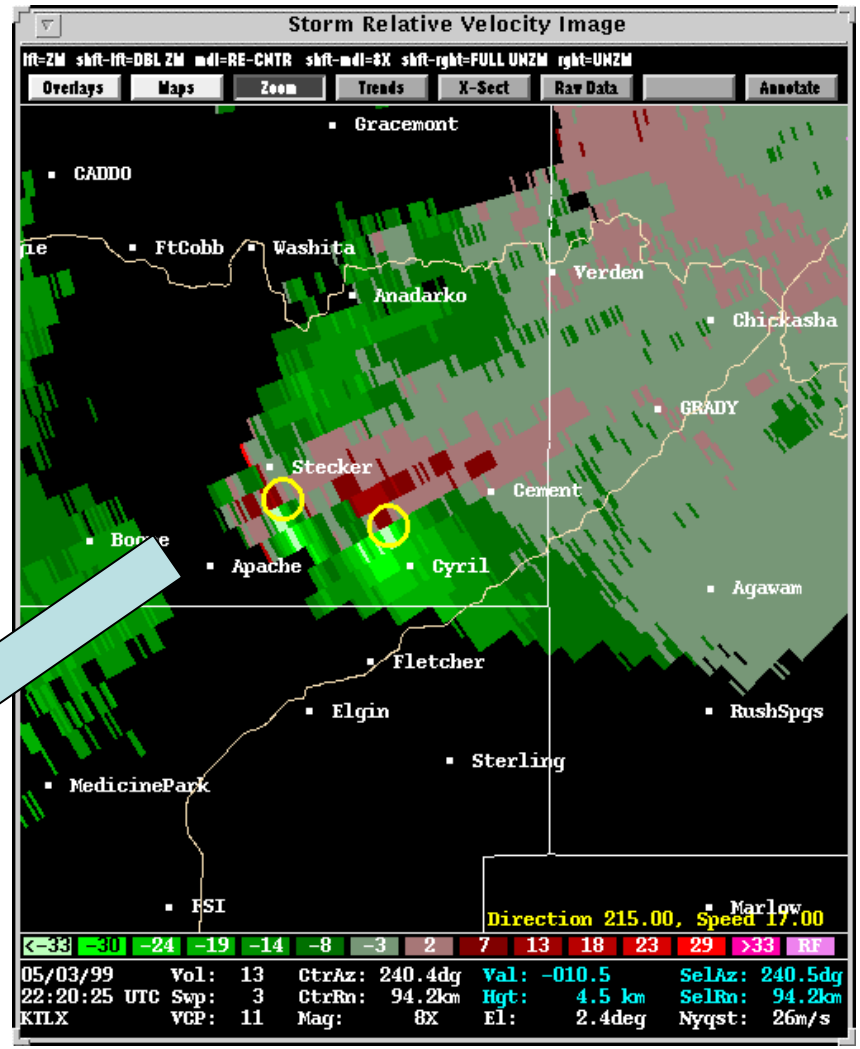


Sobel + Laplacian

Laplacian

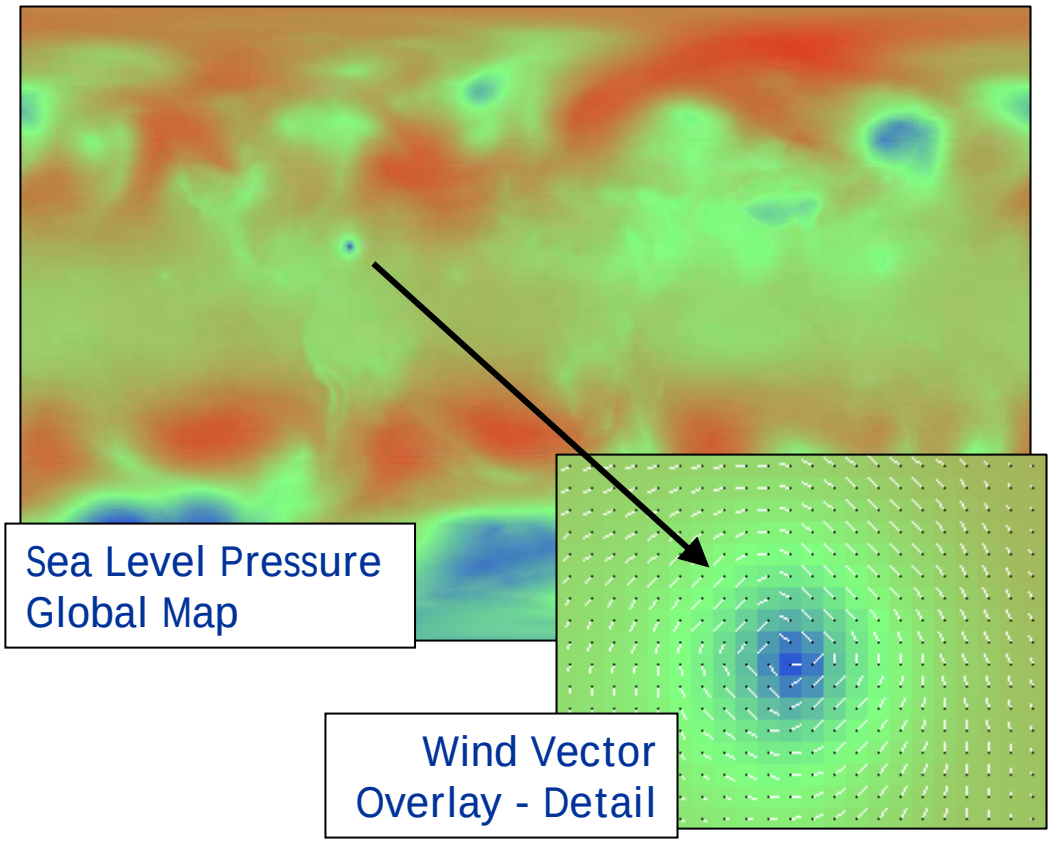
Mesocyclone Signatures

- *Problem*: Detecting mesocyclone signatures in Radar data
- *Science Rationale*: Improved accuracy and reduced false alarm rate for indicators of tornadic activity
- *Technique*: Developing an algorithm based on wind velocity shear signatures



Mining Model Data

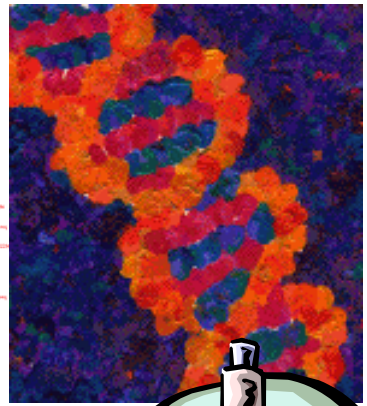
To advance the capacity for information extraction from models, NASA/ARC, the Global Modeling and Assimilation Office at NASA/GSFC, ITSC and Simpson Weather Associates are applying data mining frameworks for the analysis and extraction of information from numerical model output data generated or archived at the GMAO. The team is conducting experiments focusing on the automated detection and mining of atmospheric phenomena relationships within the model data.



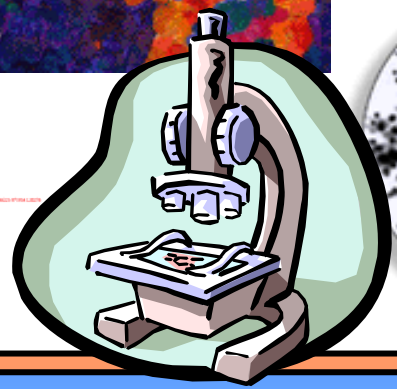
Tropical Cyclone Identification

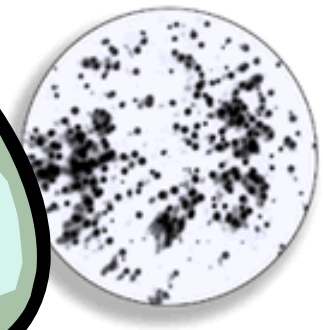
- The heuristic procedure considered all tropical ocean pixels and accepted those that:
  - Had surface pressure below a certain threshold (990)
  - Had vorticity above a certain threshold (15)
- As an alternative to the heuristic procedure, a clustering algorithm was used to derive the signature of the cyclones
  - Using pressure, vorticity
  - Using pressure, vorticity, temperature, cloud total
  - Using pressure, vorticity, cloud low

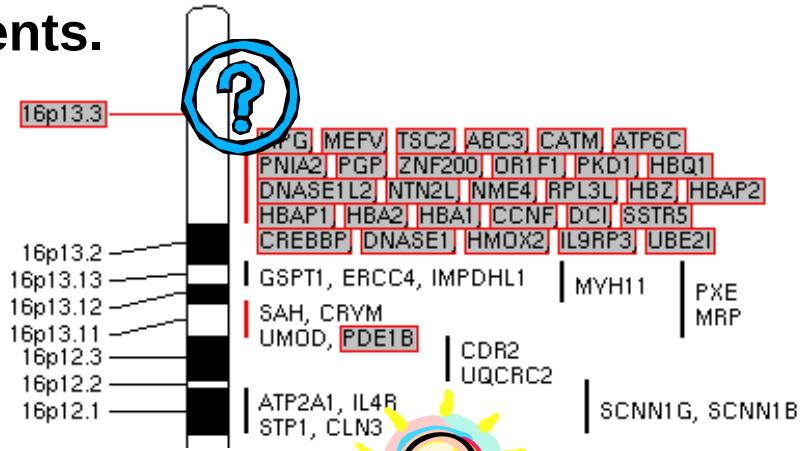
BioInformatics: Genome Patterns

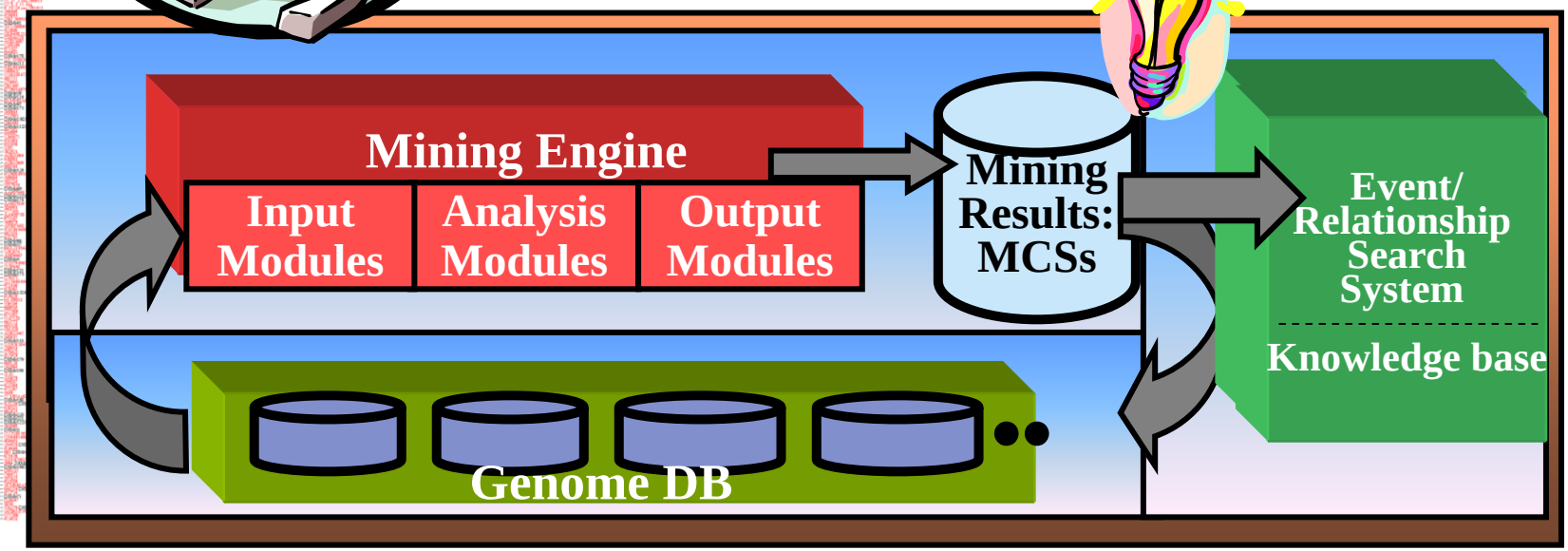


<u>Text Pattern Recognition</u>:
Used to search for text patterns in bioscience data as well as other text documents.



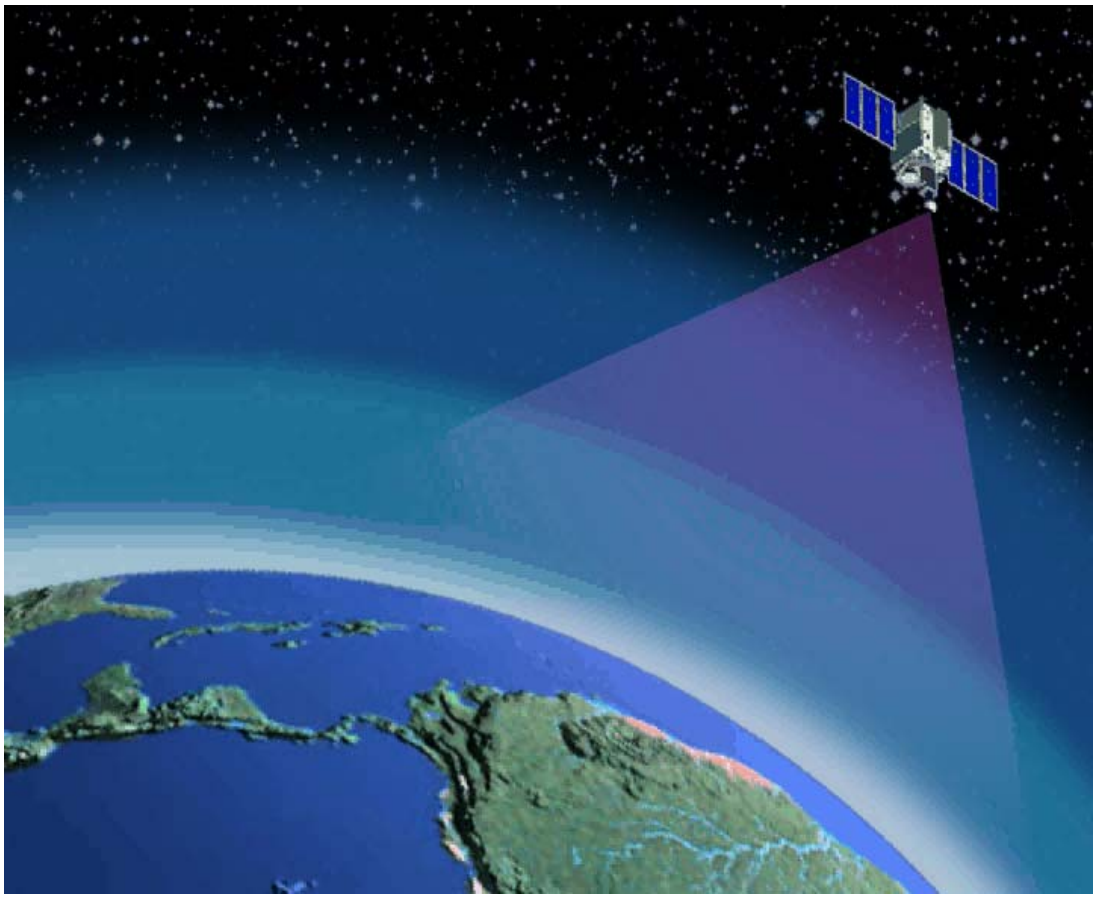






On-Board Real-Time Processing Sensor Control/Targeting

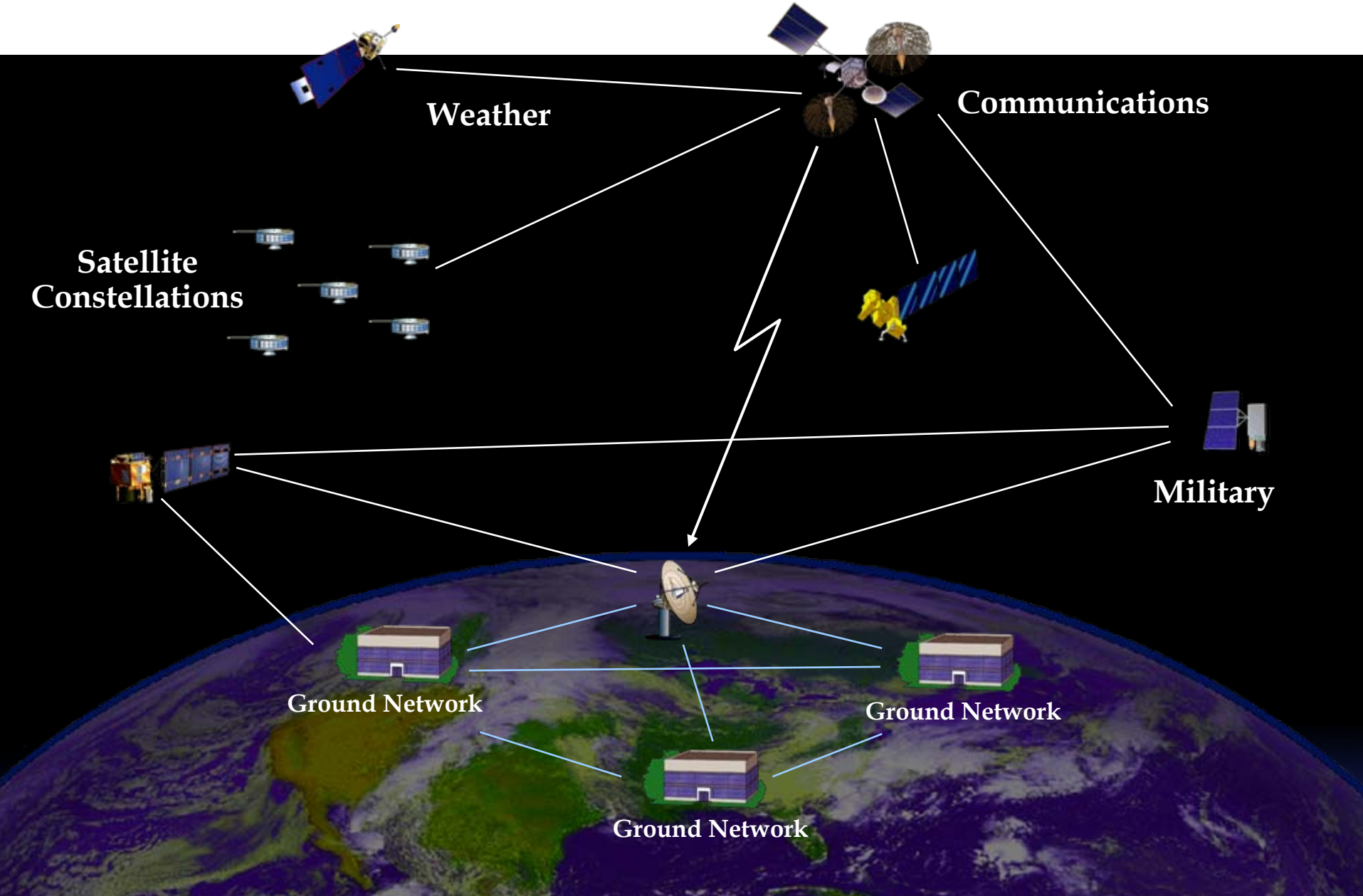
EVE – Environment for On-board Processing



- **Anomaly detection**
- **Data Mining**
- **Autonomous Decision Making**
- **Immediate response**
- **Direct satellite to Earth delivery of results**

www.itsc.uah.edu/eve

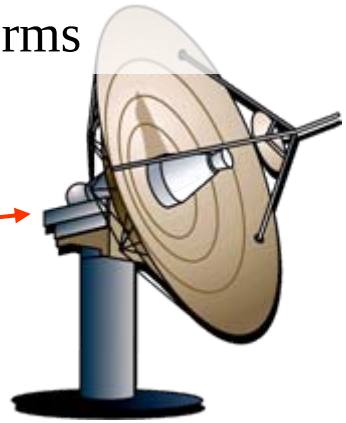
A Reconfigurable Web of Interacting Sensors

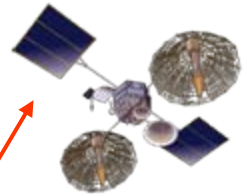


Example Application of EVE Technology: Lightning Detection During Tornadic Activity

1) The user creates a mining plan using the EVE editor

2) The Ground Station uploads the plan to multiple on-board platforms



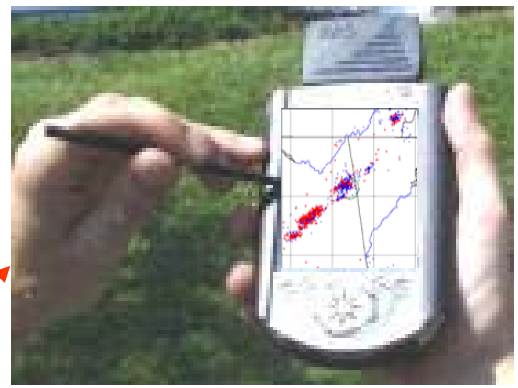


3) On-board Platform 1 uses its sensor to watch for lightning events

4) Platform 1 notifies Platform 2 of the event

5) Platform 2 requests subsetting web services from an NSSTC server

NSSTC Core Facility



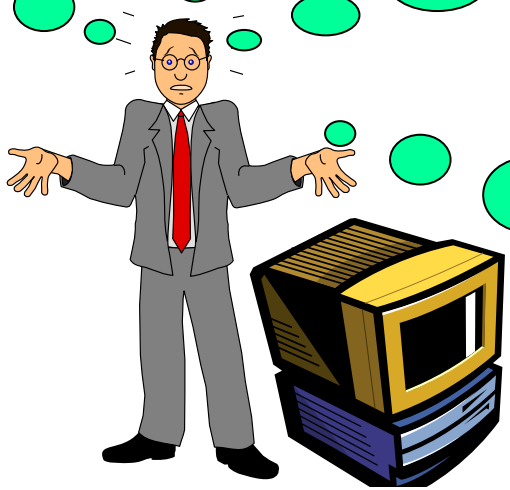
6) The results are sent back to Platform 1 for display and further processing

Mining Environment: When, Where, Who and Why?

WHEN
- Real Time
- On-Ingest
- On-Demand
- Repeatedly

WHERE
- User Workstation
- Data Mining Center
- GRID

WHO
- End Users
- Domain Experts
- Mining Experts



WHY
- Event
- Relationship
- Association
- Corroboration
- Collaboration

Data Mining

Emerging Cyberinfrastructures and Research Communities

NOMADS
The NOAA Operational Model Archive and Distribution System

GriPhyN
Data Intensive Science

Homeland Security

GEON
CYBERINFRASTRUCTURE FOR THE GEOSCIENCES

NVODS
The National Virtual Ocean Data System

MADIS

NSDI

TERAGRID

DLESE
A community-centered resource for anyone interested in learning more about the Earth.
Supported by the National Science Foundation

earthscope

NVO
NATIONAL VIRTUAL OBSERVATORY

Earth System Grid

Building the National Virtual Collaborat
for Earthquake Engineering Research
NEESgrid

ESIP
FEDERATION

nbii
National Biological Information Infrastructure

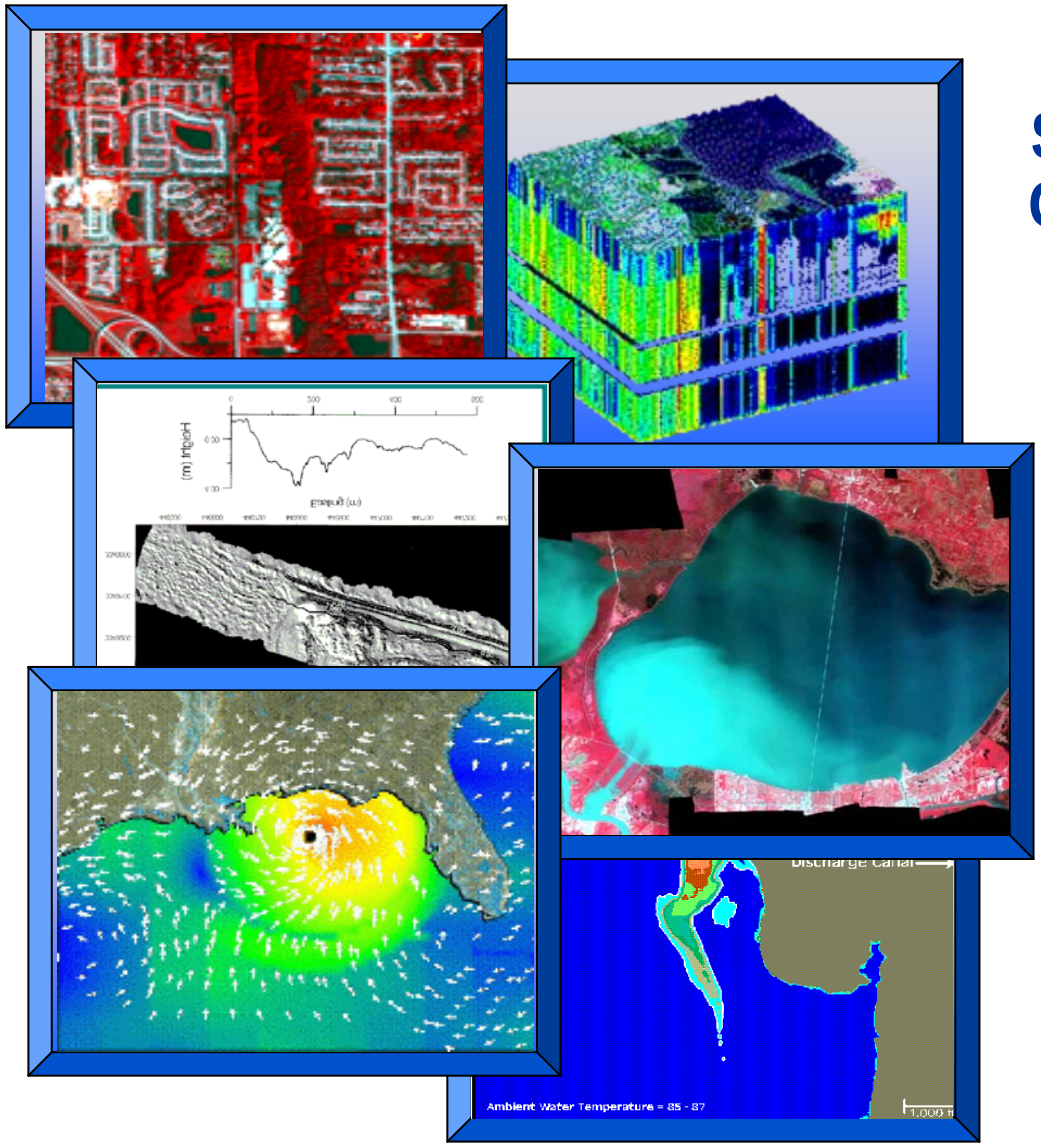
VGEE
Visual Geophysical Exploration Environ.
Home

Achieving Usability ?

Improving Data Usability

- **Advanced Applications Development**
  - *Data organization and management* for archival and analysis
  - *Data Mining* in real-time and for post run analysis
  - *Interchange Technologies* for improved data exploitation
  - *Semantics* to transform data exploitation via intelligent automated processing
- **Infrastructure Development**
  - *Grid technologies* for seamless access to multiple computational and data resources into a virtual computing environment
  - *Cluster technologies* for high speed parallel computation, for multiple agent computations, and other applications
  - *High-performance networking* for advanced applications development and high performance connectivity
  - Next generation technologies in *videoconferencing and electronic collaboration*

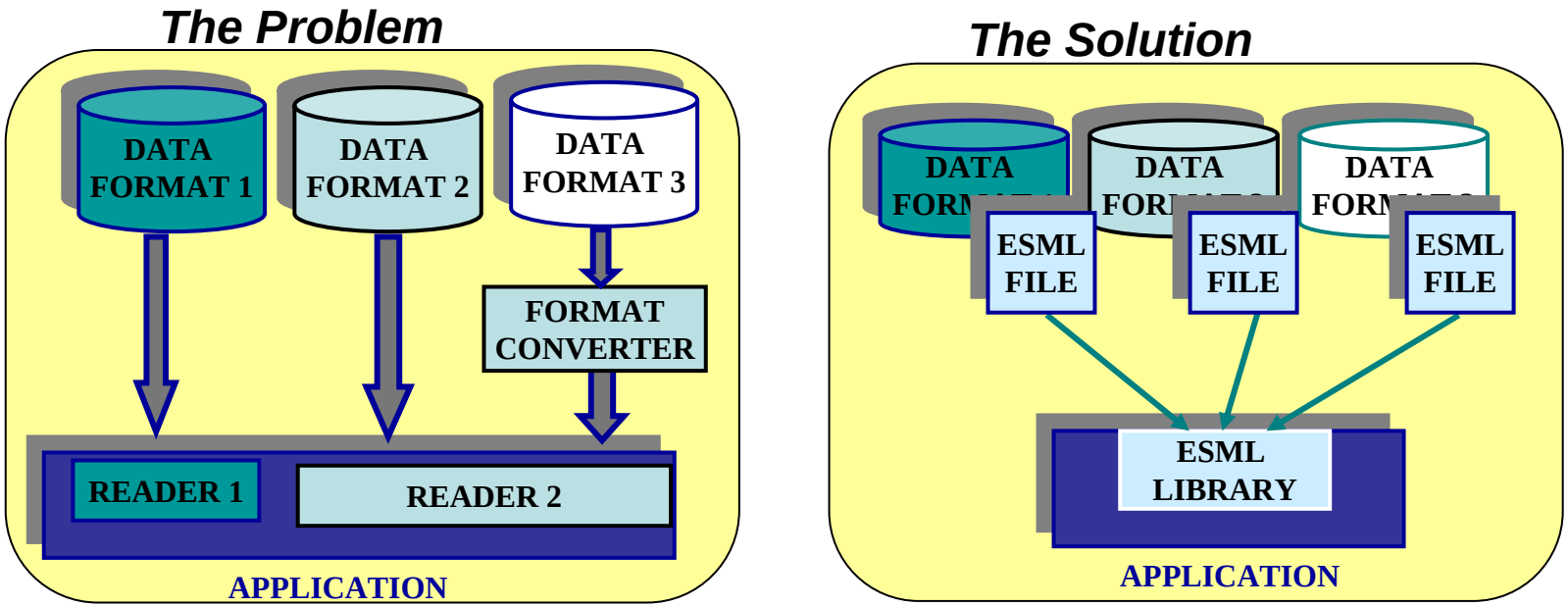
Heterogeneity Leads to Data Usability Problems



Science Data Characteristics

- Many different formats, types and structures (18 and counting for atmospheric science alone!)
- Different states of processing (raw, calibrated, derived, modeled or interpreted)
- Enormous volumes

Interoperability: Accessing Heterogeneous Data



One approach: Enforce a standard data format, *but*…

- Difficult to implement and enforce
- Can't anticipate all needs
- Some data can't be modeled or is lost in translation
- Converting legacy data is costly

A better approach: Interchange Technologies

- Earth Science Markup Language

What is ESML?

- It is a **specialized markup language** for Earth Science metadata based on XML - NOT another data format.
- It is a machine-readable and -interpretable representation of the **structure, semantics and content** of any data file, regardless of data format
- ESML description files contain **external metadata** that can be generated by either data producer or data consumer (at collection, data set, and/or granule level)
- ESML provides the **benefits of a standard, self-describing data format** (like HDF, HDF-EOS, netCDF, geoTIFF, …) without the cost of data conversion
- ESML is the basis for **core Interchange Technology** that allows data/application interoperability
- ESML **complements and extends data catalogs** such as FGDC and GCMD by providing the use/access information those directories lack.

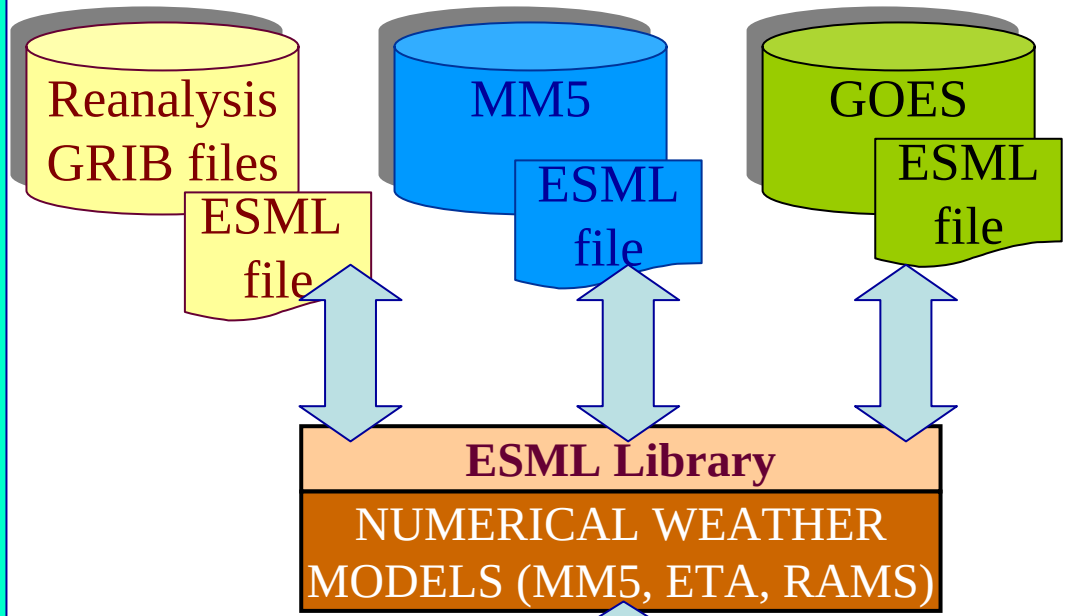
http://esml.itsc.uah.edu

ESML IN ACTION:

Ingest surface skin temperature data in Numerical Models

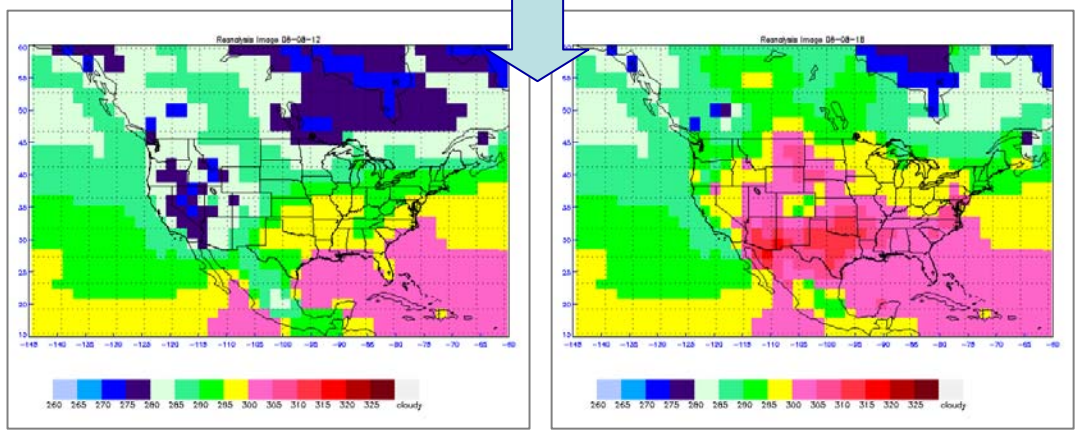
Purpose:

- *Use ESML to incorporate observational data into the numerical models for simulation*
- *Skin temperatures come in a variety of data formats*
  - *GOES – McIDAS*
  - *Reanalysis Data - GRIB*
  - *MM5 Model - Binary*
  - *AVHRR – HDF*
  - *MODIS - EOS-HDF*



Scientists can:

- *Select remote files across the network*
- *Select different observational data to increase the model prediction accuracy*



http://vortex.nsstc.uah.edu/~sud/web/default.htm

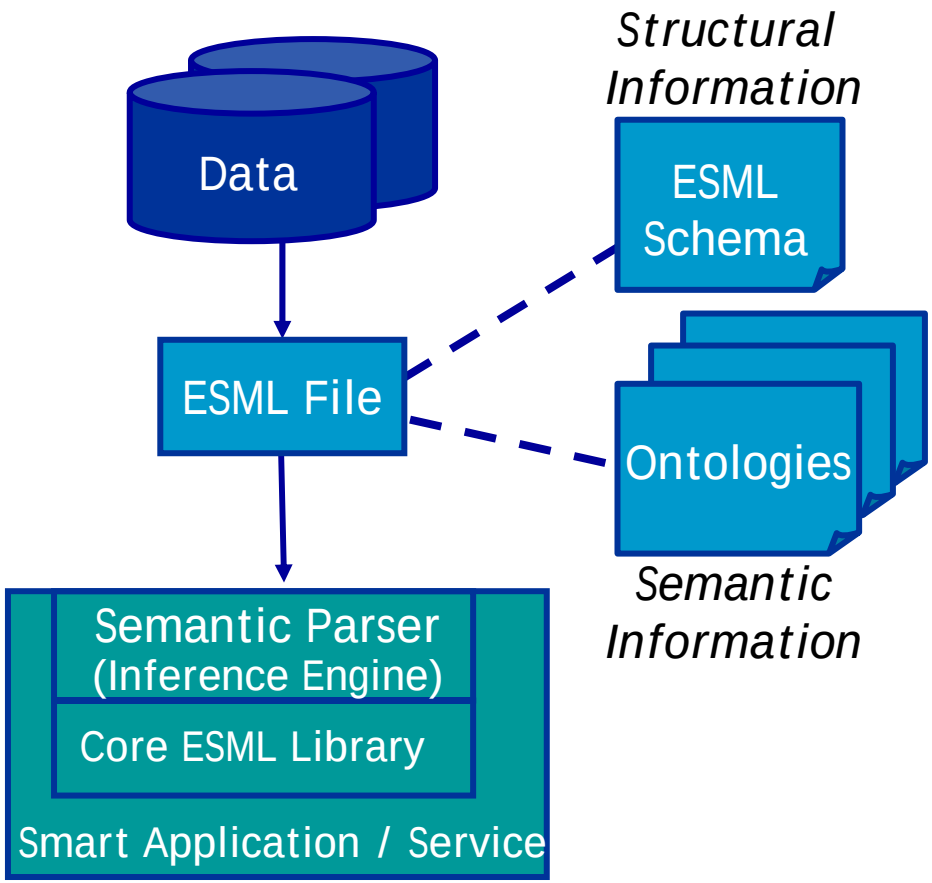
Ontology Definition

- Philosophy
  - Science of Being [Aristotole]
- Machine Learning/AI/Intelligent Systems
  - "An Ontology is a FORMAL, EXPLICIT specification of a SHARED conceptualization" [Gruber, 1993]
    - Explicit – type of concept and constraints of use are explicitly defined
    - Formal – should be machine understandable
    - Shared – captures consensual knowledge
- Ontology consists of concepts and their relationships

Uses of Ontology

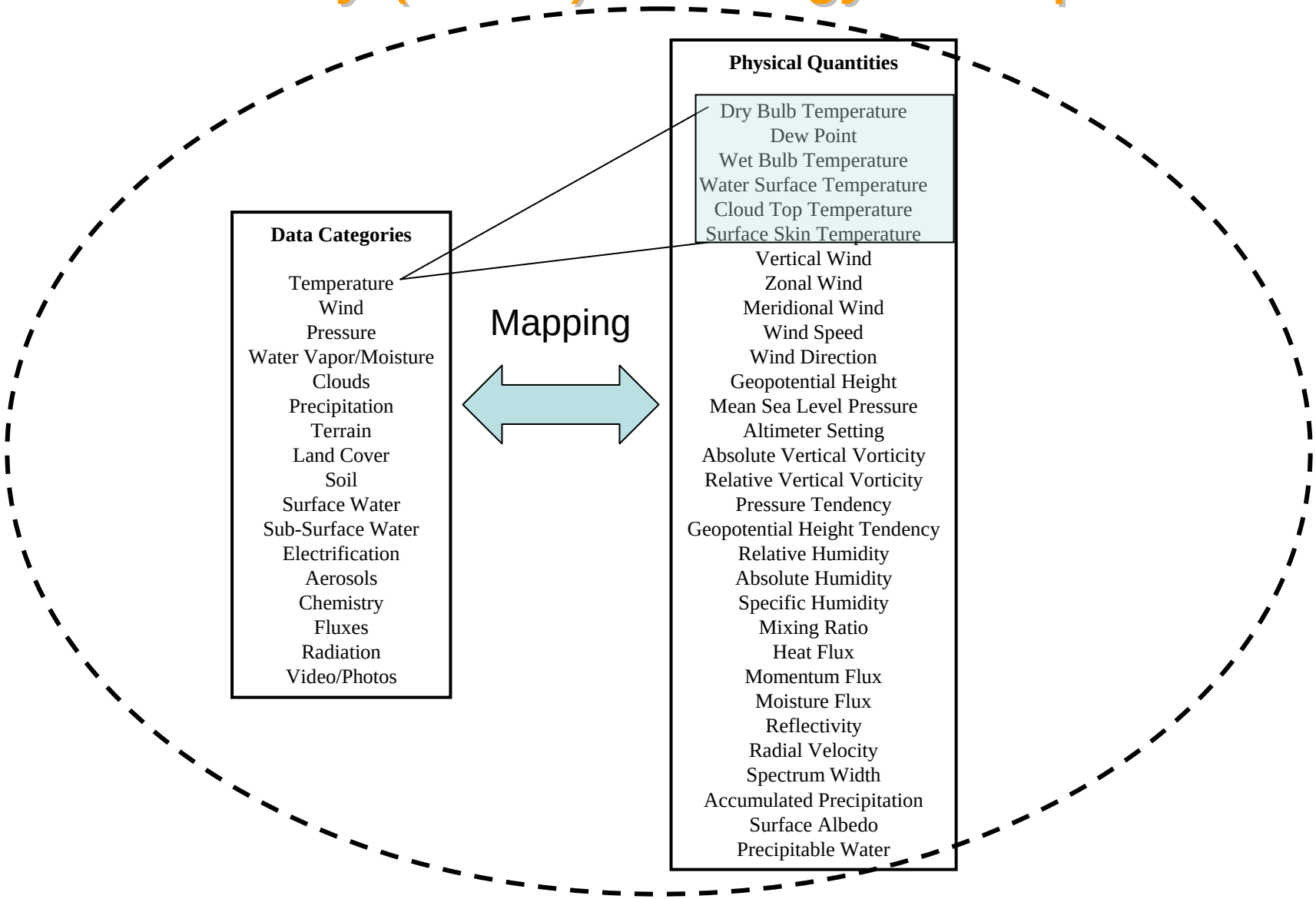
- To provide a vocabulary of terms for a domain instead of relying on an external agreement – such as yellow page services for:
  - Data
    - Search for data sets that contain “temperature”
  - *Services*
    - Search for mining capabililties
- Meaning of terms is formally specified
  - *Definitions and relationships between various phenomena*
- Specify relationships between terms in multiple catalogs
  - *Semantic metadata for data sets*

Smart Applications/Services using ESML and Ontologies

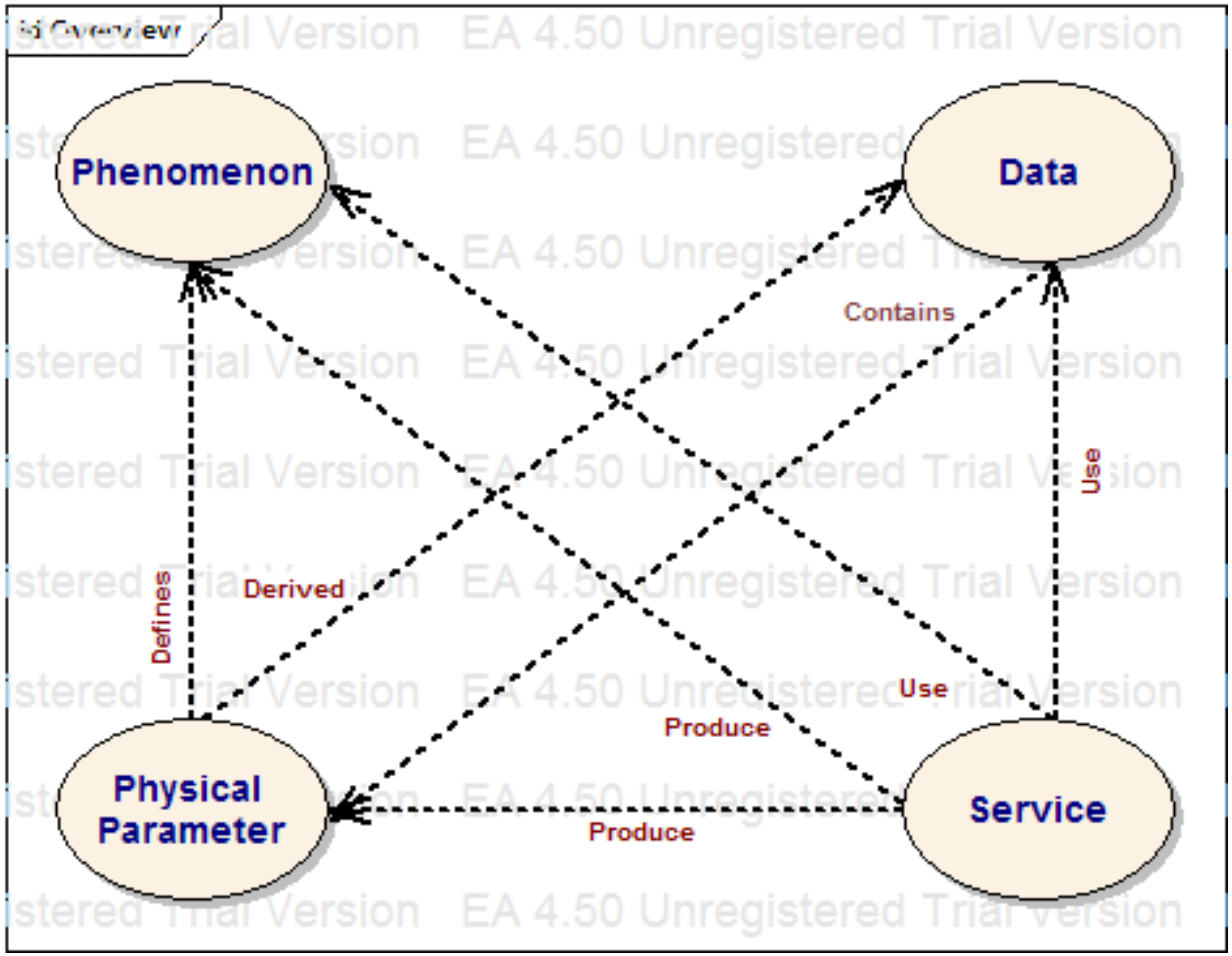


- ESML Schema's focus is on providing structural data interoperability between data/application
- However, ESML ***allows embedding semantic terms*** for data fields in the Description File to provide a complete structural and semantic description of the data
- Various science communities can create their own ontologies (for example, SWEET) and link them with ESML Description Files for their data
- Application developers can add semantic parsers on top of the core ESML Library to build ***"smart" applications or services***

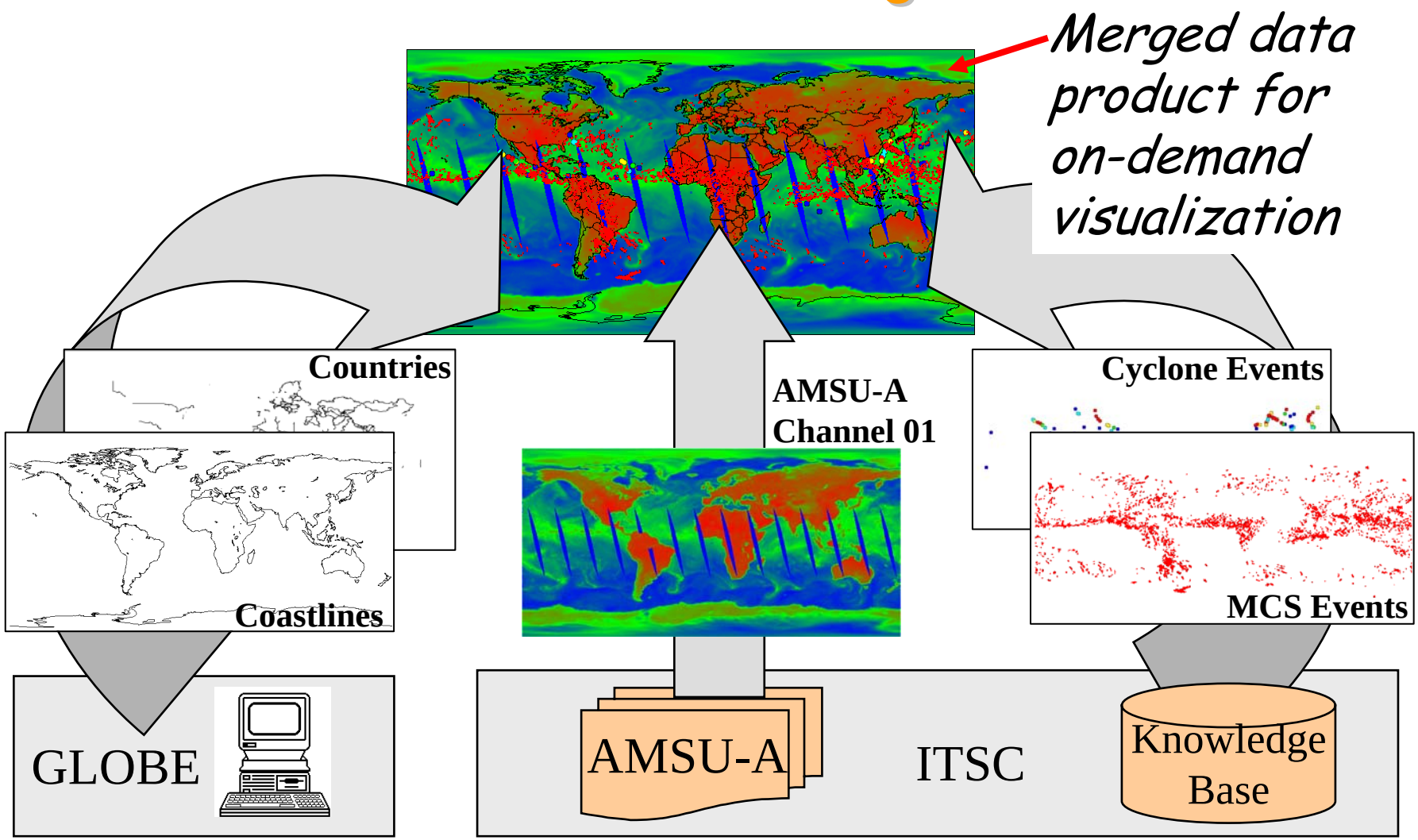
Linked Environments for Atmospheric Discovery (LEAD) Ontology Example



Eventual LEAD Ontology Goal:
Create a topology linking these key concepts

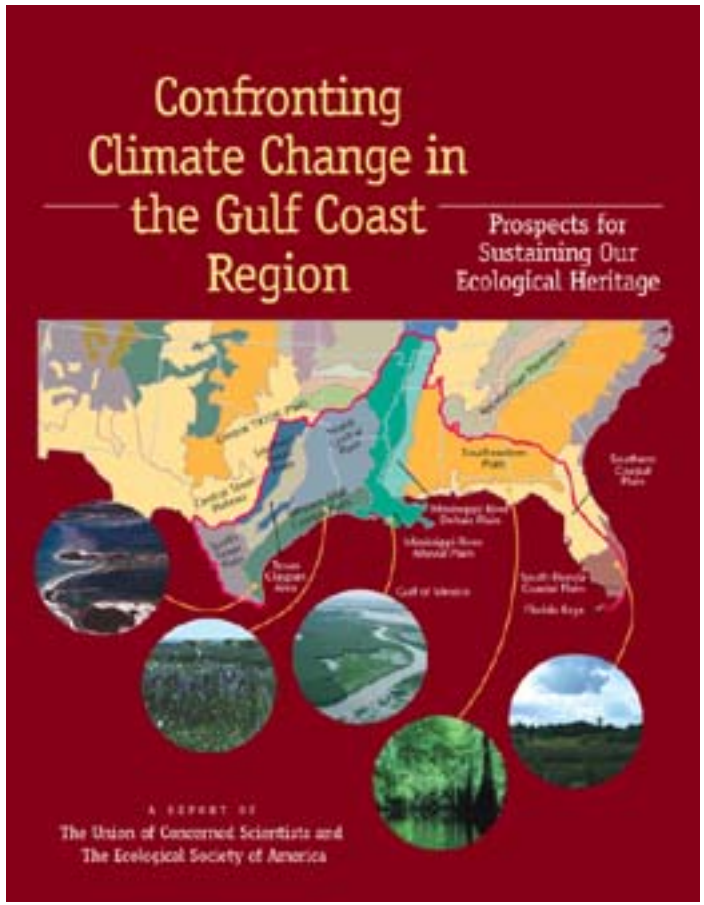


Distributed Data Integration



AMSU-A data overlaid with MCS and Cyclone events, merged with world boundaries from GLOBE.

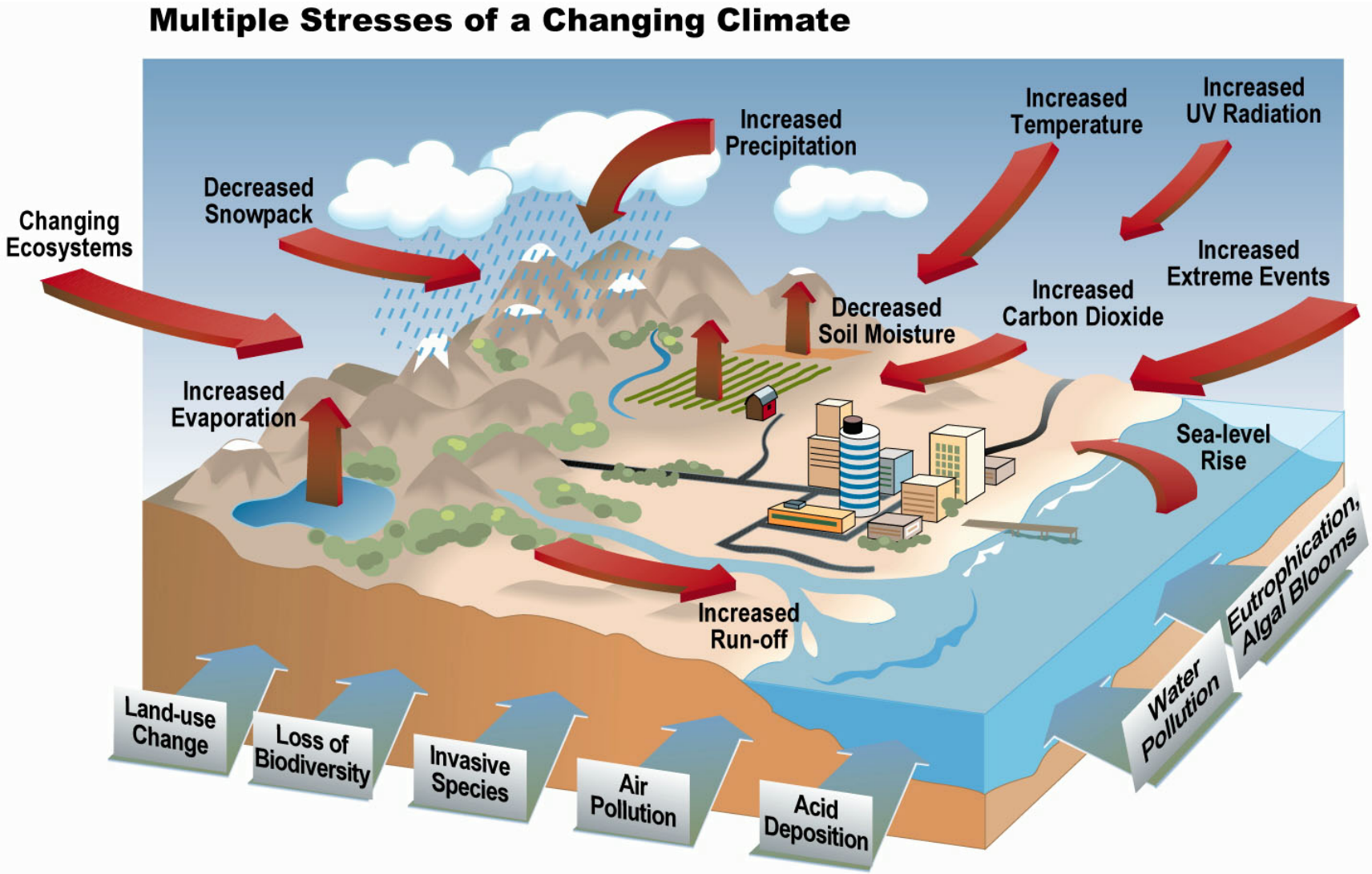
Confronting Climate Change in the Gulf Coast Region



- **The Gulf Coast region supports a diversity of ecosystems.**

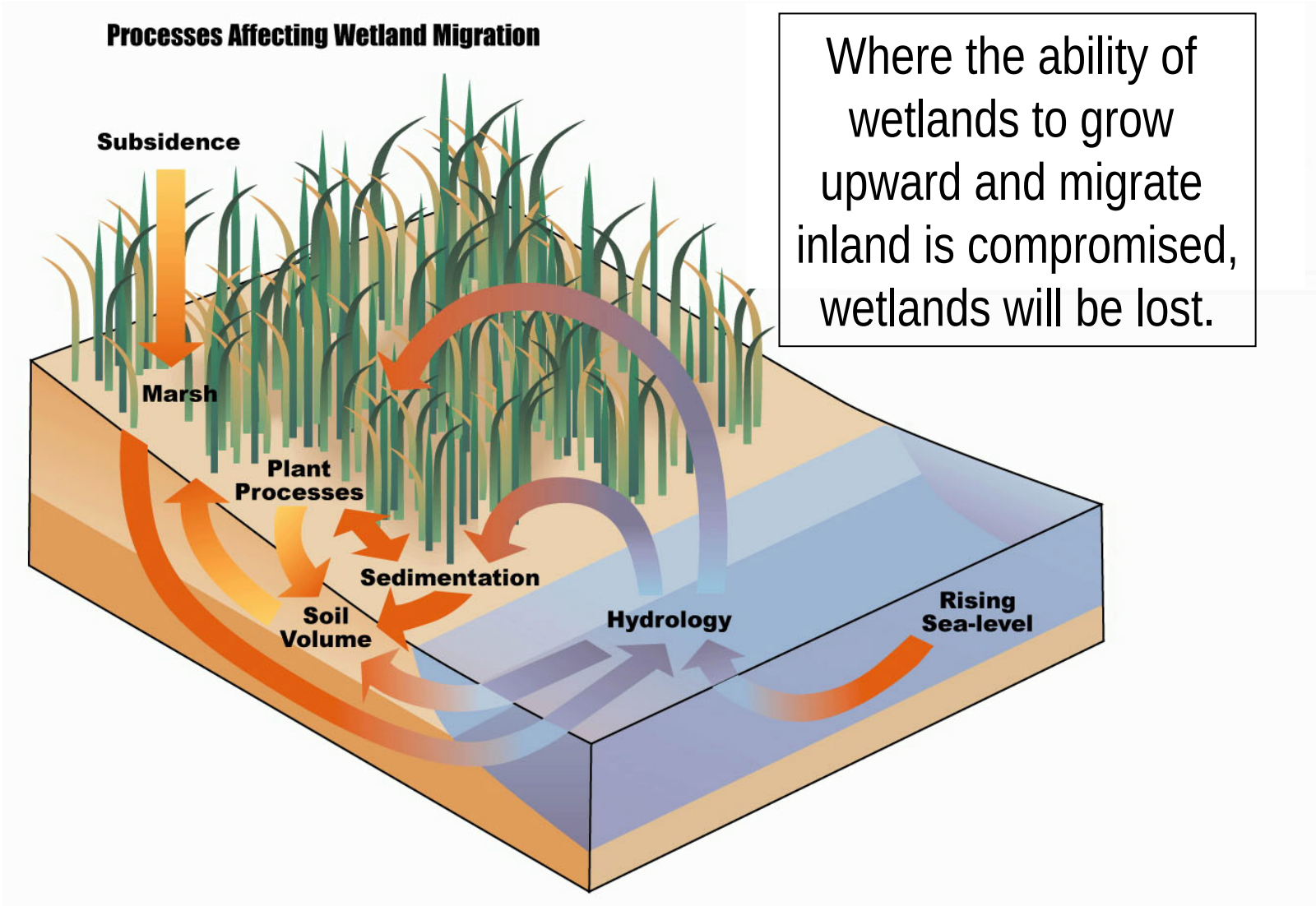
- **These ecosystems are at risk from human pressures and climate change.**

Multiple Stresses of a Changing Climate



Source: U.S. National Assessment, NAST, 2001

Wetland migration and losses



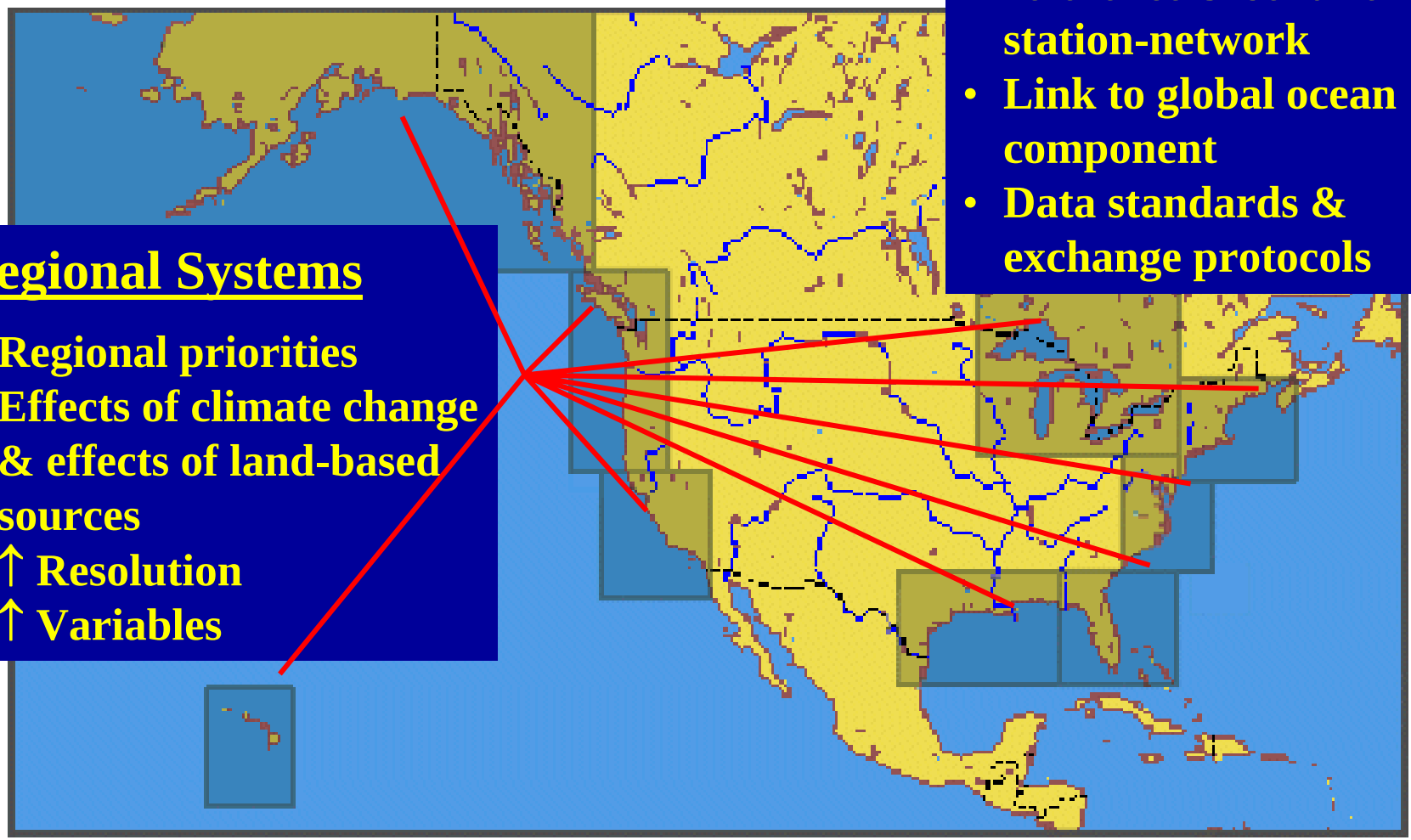
Where the ability of wetlands to grow upward and migrate inland is compromised, wetlands will be lost.

Source: Reed, 1995

National Federation of Regional Systems

National Backbone

- Satellite remote sensing
- In situ sensing reference & sentinel station-network
- Link to global ocean component
- Data standards & exchange protocols



Regional Systems

- Regional priorities
- Effects of climate change & effects of land-based sources
- ↑ Resolution
- ↑ Variables

Integrated Ocean Observing System (IOOS)

Serves national needs for:

Detecting and forecasting oceanic components of climate variability

- Facilitating safe and efficient marine operations
- Ensuring national security
- Managing resources for sustainable use
- Preserving and restoring healthy marine ecosystems
- Mitigating natural hazards
- Ensuring public health

SURA Coastal Ocean Observing and Prediction Program (SCOOP)

- Southeastern Universities Research Association (SURA) program to create an infrastructure for a distributed Integrated Ocean Observing System (IOOS) in the southeastern region
  - Provides a shared means for the acquisition of observational data
  - Enables modeling, analysis and delivery of information in real-time
- SCOOP will serve as a model for the national effort in integrating ocean observing systems

SCOOP Goals

- Assess and predict the coastal response to extreme atmospheric events – focus on storm surge, flooding & waves
- Modular modeling tools for regional issues (wave coupling, sediment suspension, etc.)
- Standardized interfaces for data and (coupled) model interoperability
- Ensemble prediction – forecasts based on many independent models runs
- Measure, understand and predict environmental conditions
- Provide R&D support for operational agencies including NOAA, the U.S. Navy, and others
- Include outreach and education components that assure relevance of their observing activities

SCOOP Research Outcomes

- SCOOP can help us understand:
  - Ecological issues such as the Louisiana coastal zone land loss problem
  - Societal issues such as hurricane threats and damage assessment

Coordination is Key

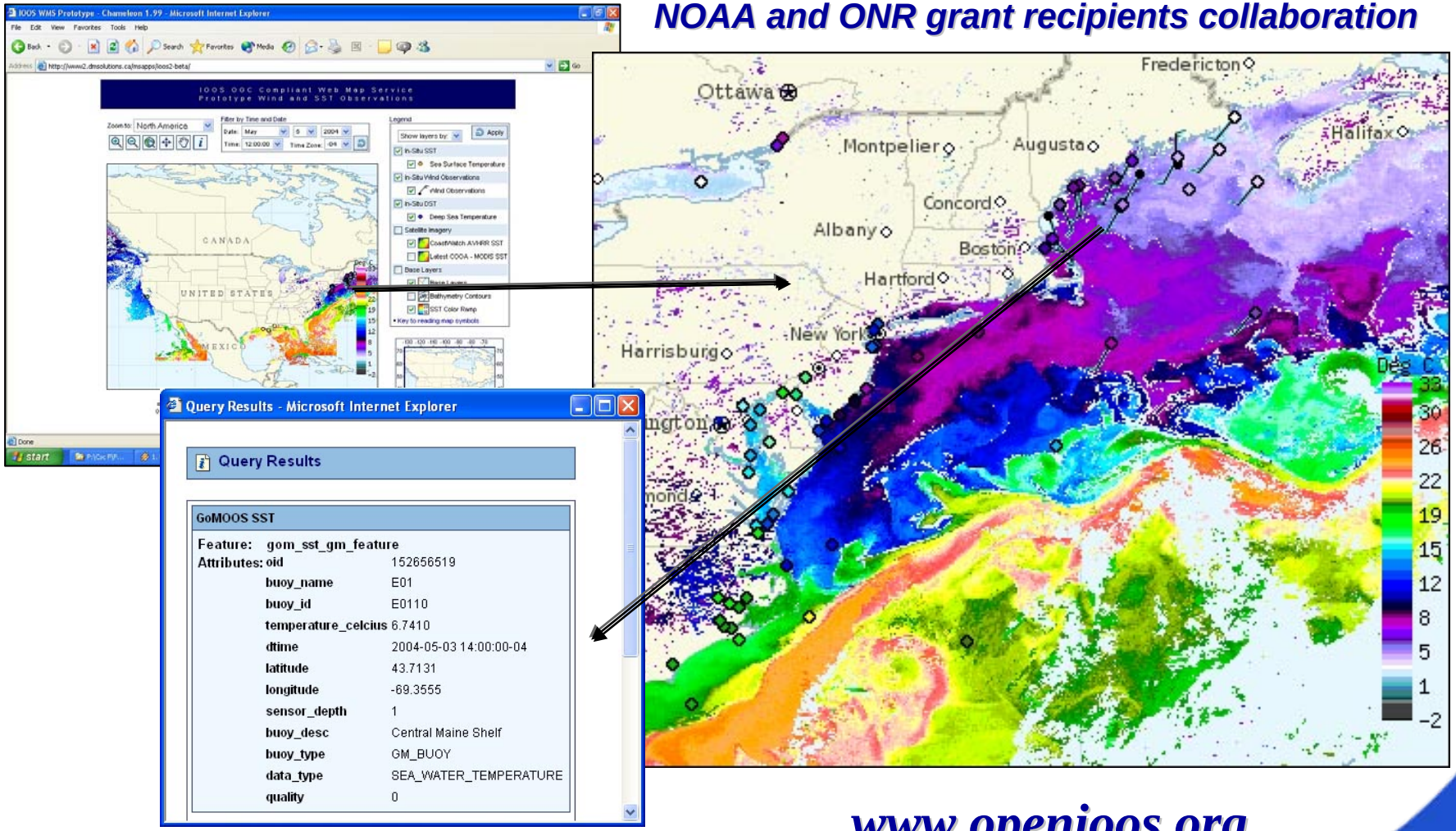
- **Ocean.US - National Office for Integrated and Sustained Ocean Observations** *coordinates development of an operational, integrated and sustained Ocean Observing System (created by NOPP)* http://www.ocean.us/
- **Integrated Ocean Observation System (IOOS)** *a national effort to create an Integrated Ocean Observing System* http://www.openioos.org/
- **National Oceanographic Partnership Program (NOPP)** *15 federal agencies providing leadership and coordination of national research and education programs* http://www.nopp.org/
- **National Federation of Regional Associations** *provide a framework for orchestrating regional collaborations* http://www.usnfra.org/
- **NSF Ocean Research Interactive Observatory Networks (ORION)** *an emerging network of science-driven ocean observing systems* http://www.orionprogram.org/default.html

Interoperability is Key

- **Ocean.US Data Management and Communications (DMAC) Plan** *provides the framework for interoperability* http://dmac.ocean.us/dacsc/imp_plan.jsp
- **Open Geographic Information Systems (GIS) Consortium (OGC)** *an open consortium of industry, government, and academia developing interface specifications to support interoperability* http://www.opengis.org
- **Marine Metadata Interoperability** *a community effort to make marine metadata easier to find, access and use* http://www.marinemetadata.org/

Interoperability Demonstration

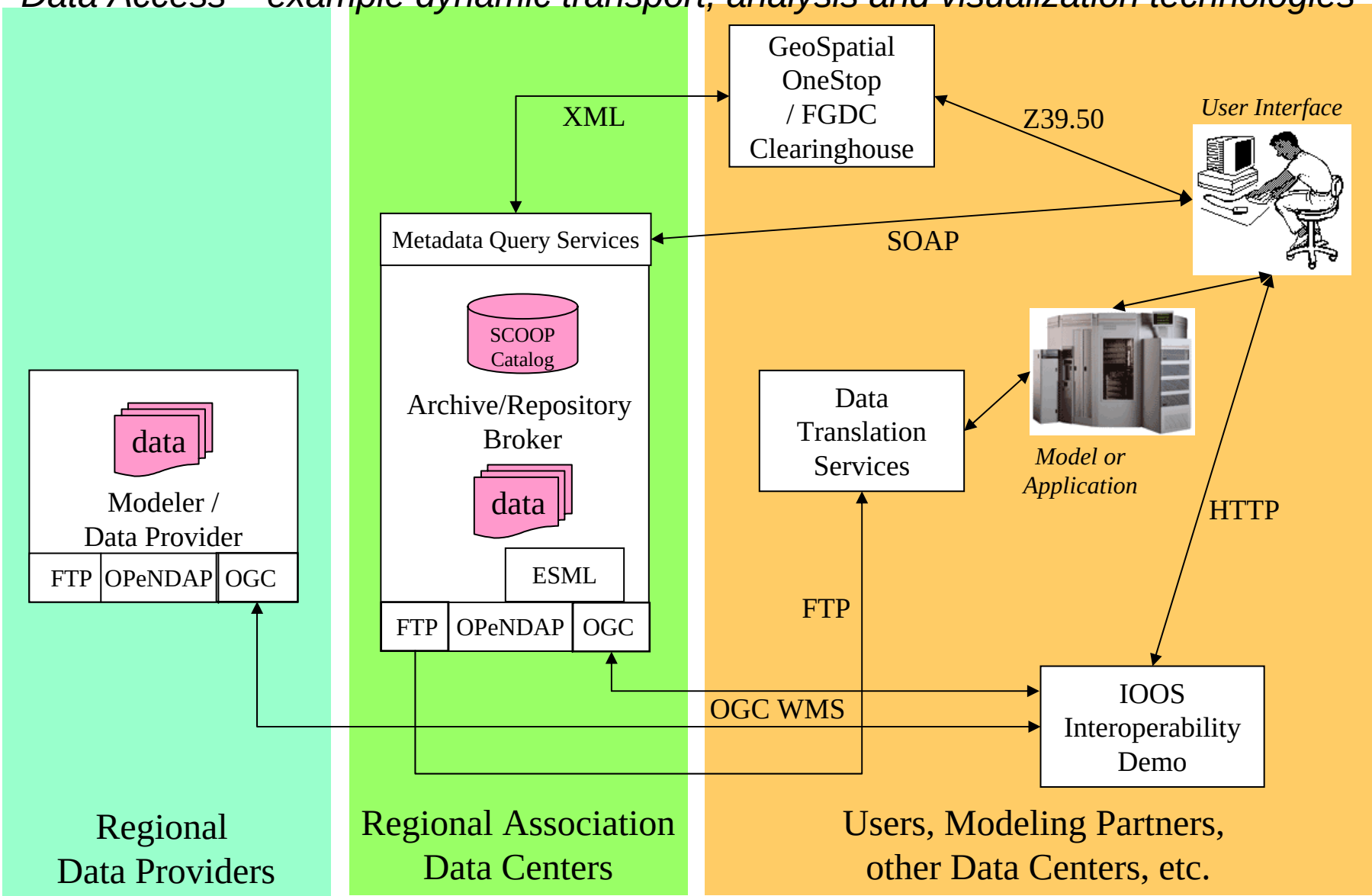
NOAA and ONR grant recipients collaboration



www.openioos.org

SCOOP Data Architecture Specifics:

Data Access – example dynamic transport, analysis and visualization technologies



SCOOP System Development

- **Funding provided by ONR, NOAA**
- **SCOOP Partners:**

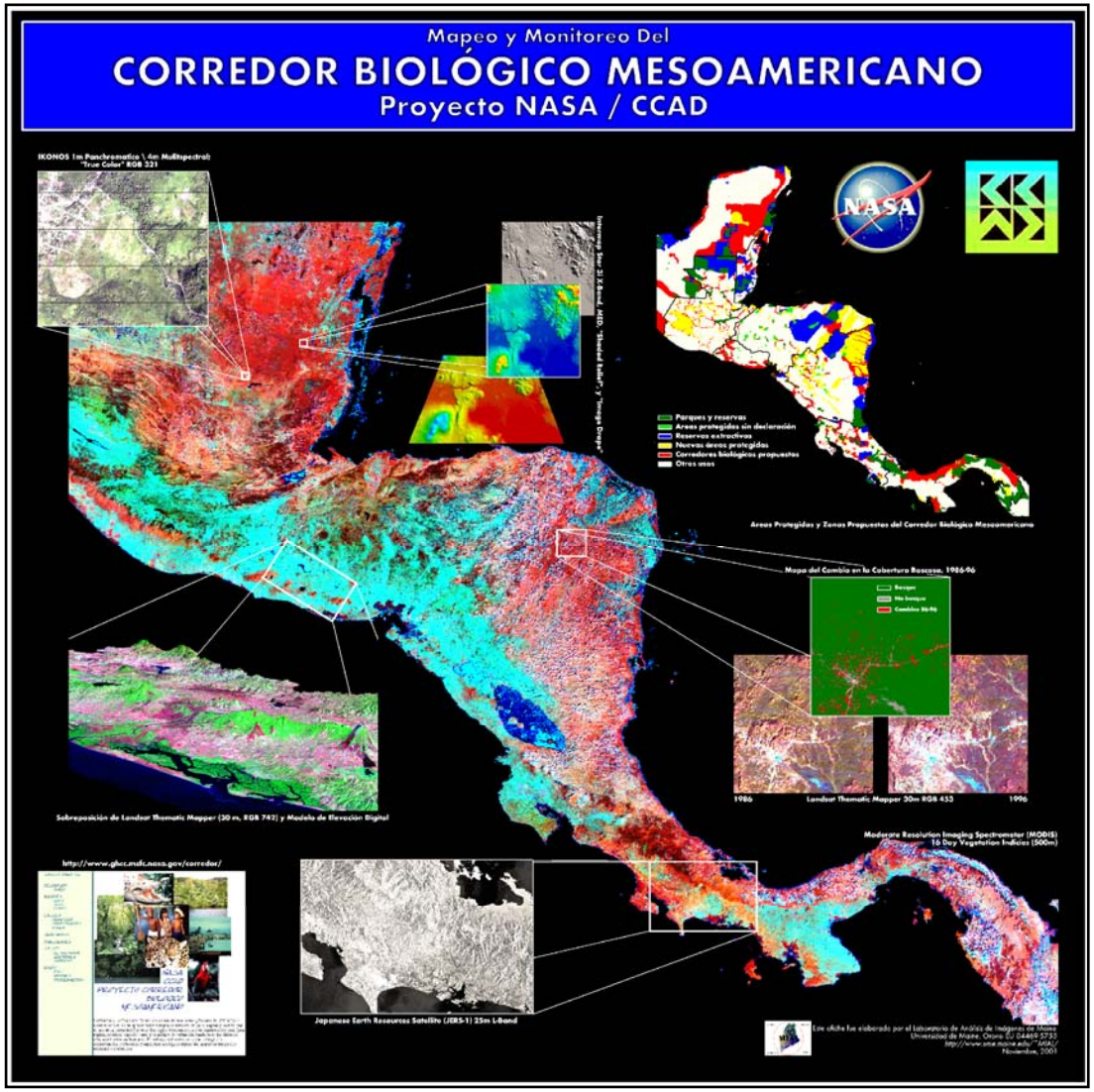
  - Consortium for Oceanographic Research and Education
  - Gulf of Maine Ocean Observation System (GoMOOS)
  - Louisiana State University, Center for Computation & Technology
  - Louisiana State University, Coastal Studies Institute
  - Southeast Atlantic Coastal Ocean Observing System (SEACOOS)
  - Texas A&M University
  - University of Alabama in Huntsville
  - University of Delaware (Mid-Atlantic Regional Association (MARA)
  - University of Florida
  - University of Miami, Center for Southeastern Tropical Advanced Remote Sensing
  - University of North Carolina
  - Virginia Institute of Marine Science

SERVIR

An Environmental Monitoring and Decision Support System for Mesoamerica

NASA UAH
The University of Alabama in Huntsville

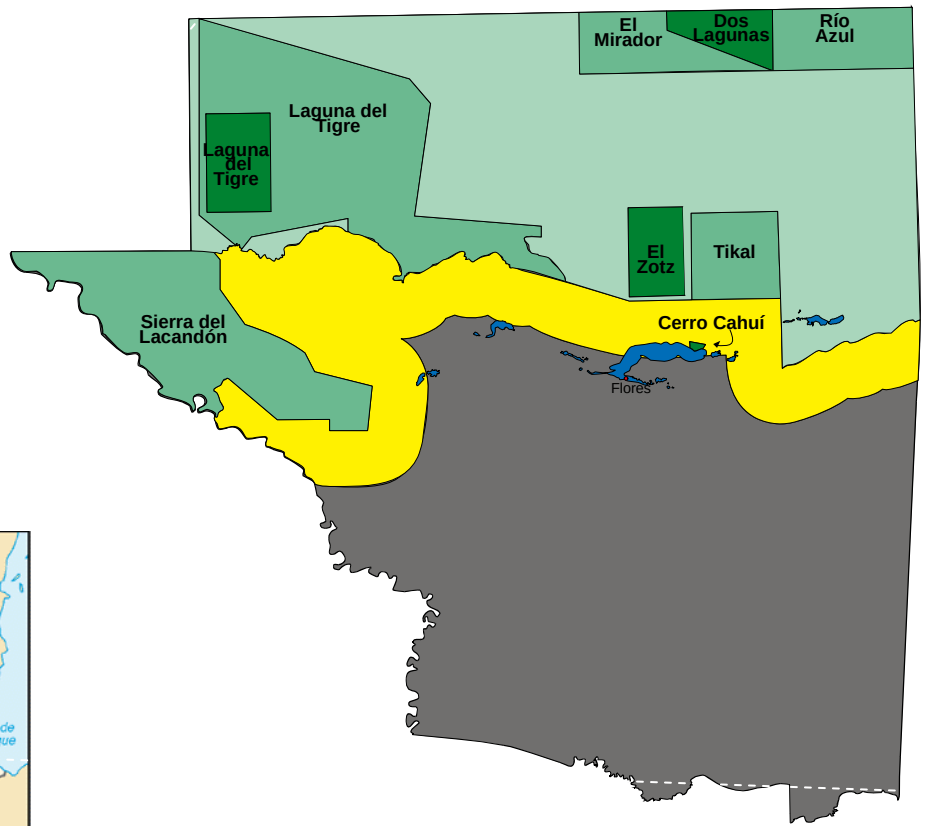
NASA/CCAD Mesoamerican Biological Corridor (MBC)



MBC Diversity

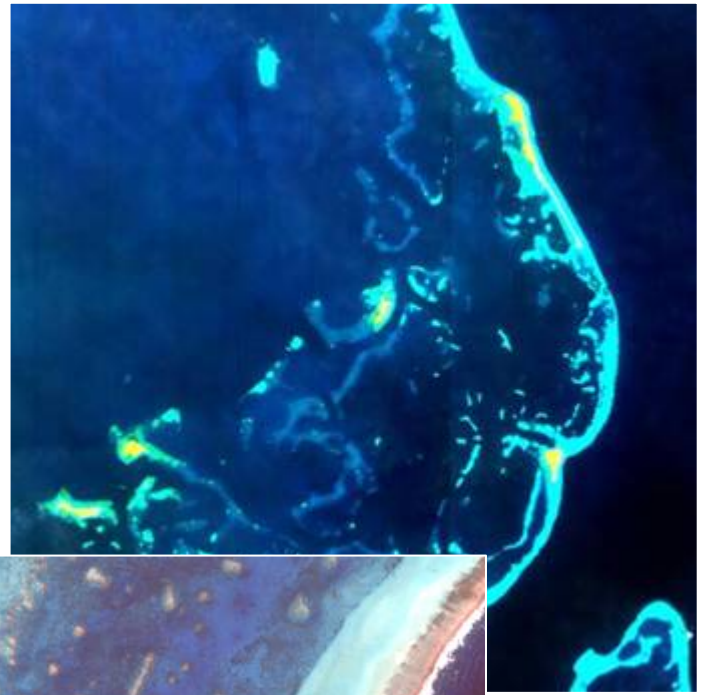
- Less than 1% of Earth's landmass, but contains 7-8% of world's plant and animal species

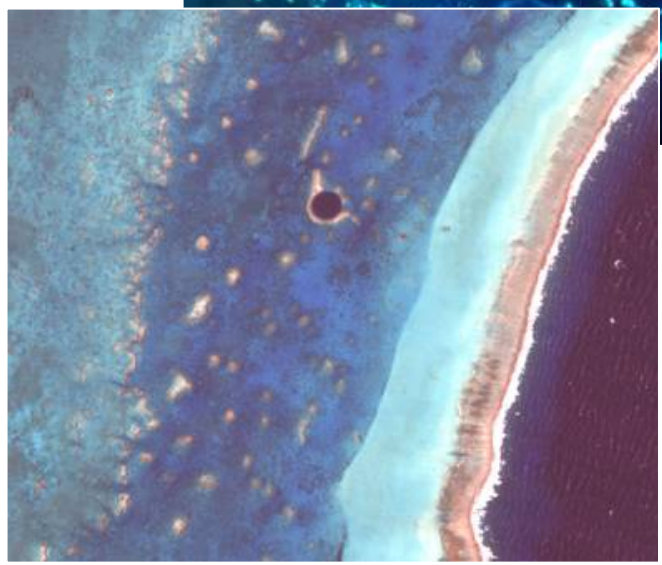
The Maya Biosphere Reserve



- Parques Nacionales
- Biotopos
- Zona de Usos Múltiples
- Zona de Amortiguamiento

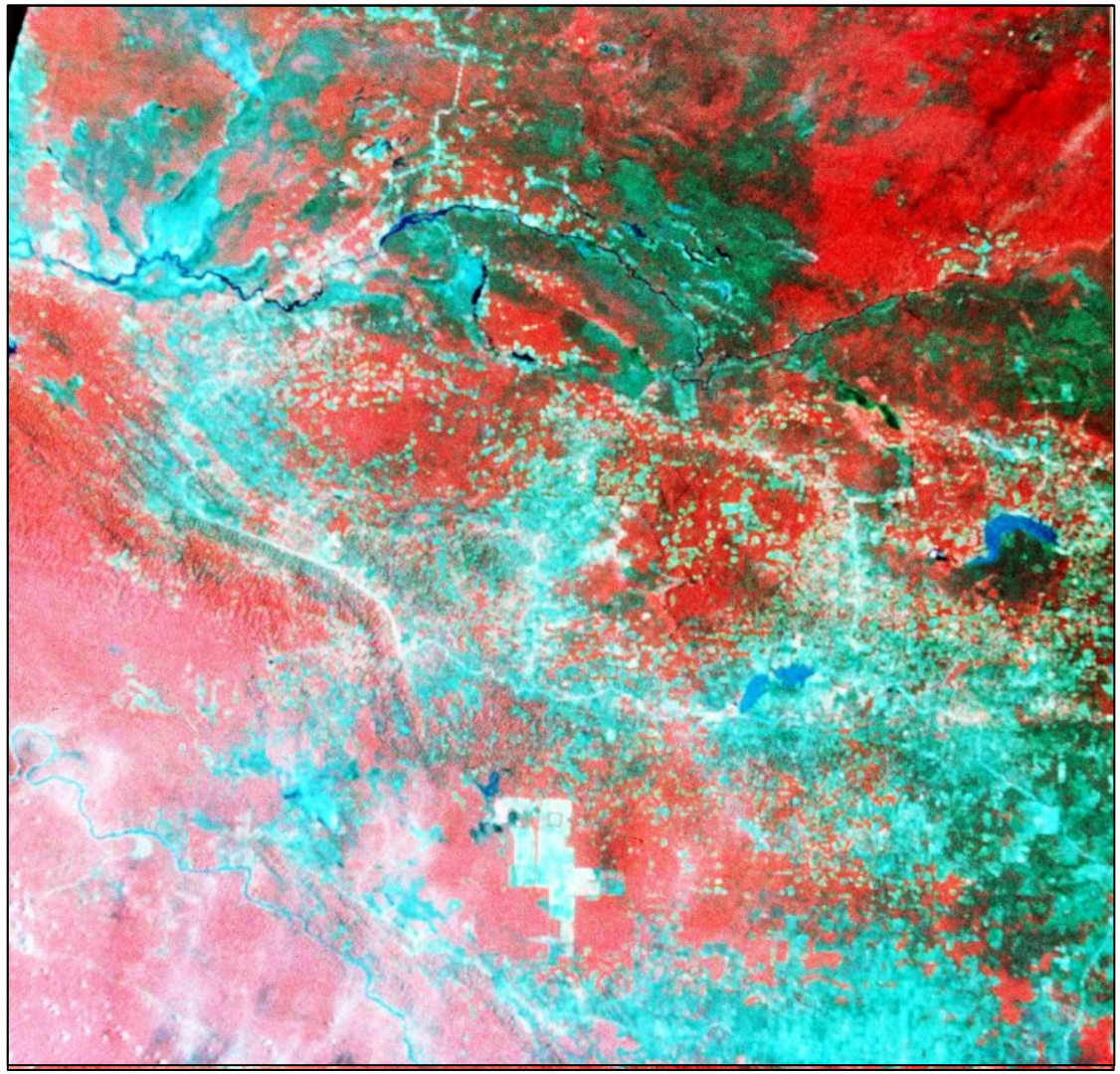
Coral Reef Monitoring





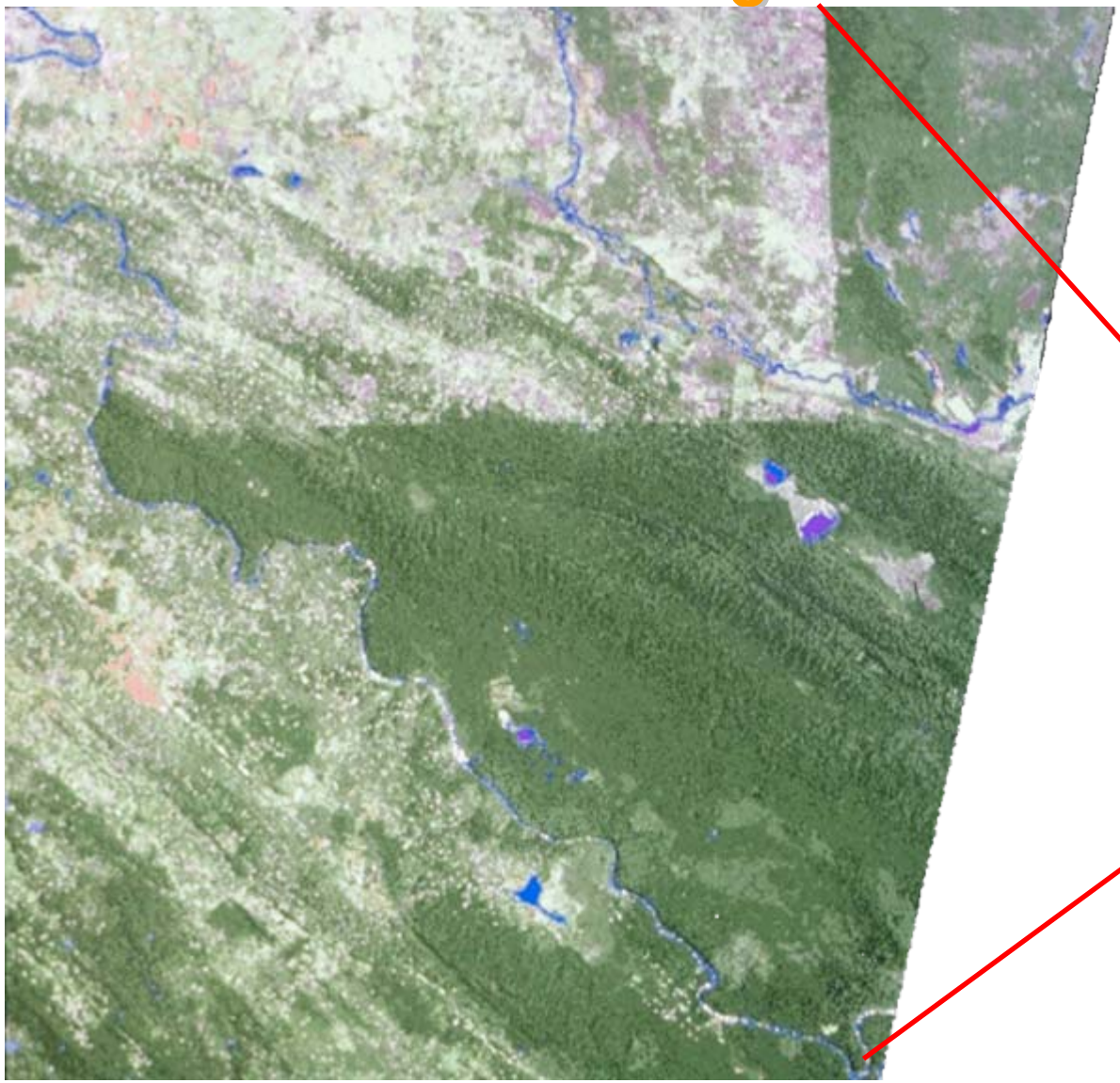
Land Cover and Land Use Change

- Conservation & Preservation
- Transportation Infrastructure
- Urban Growth Planning
- Human Impacts on the Land
- Infrastructure and Utilities

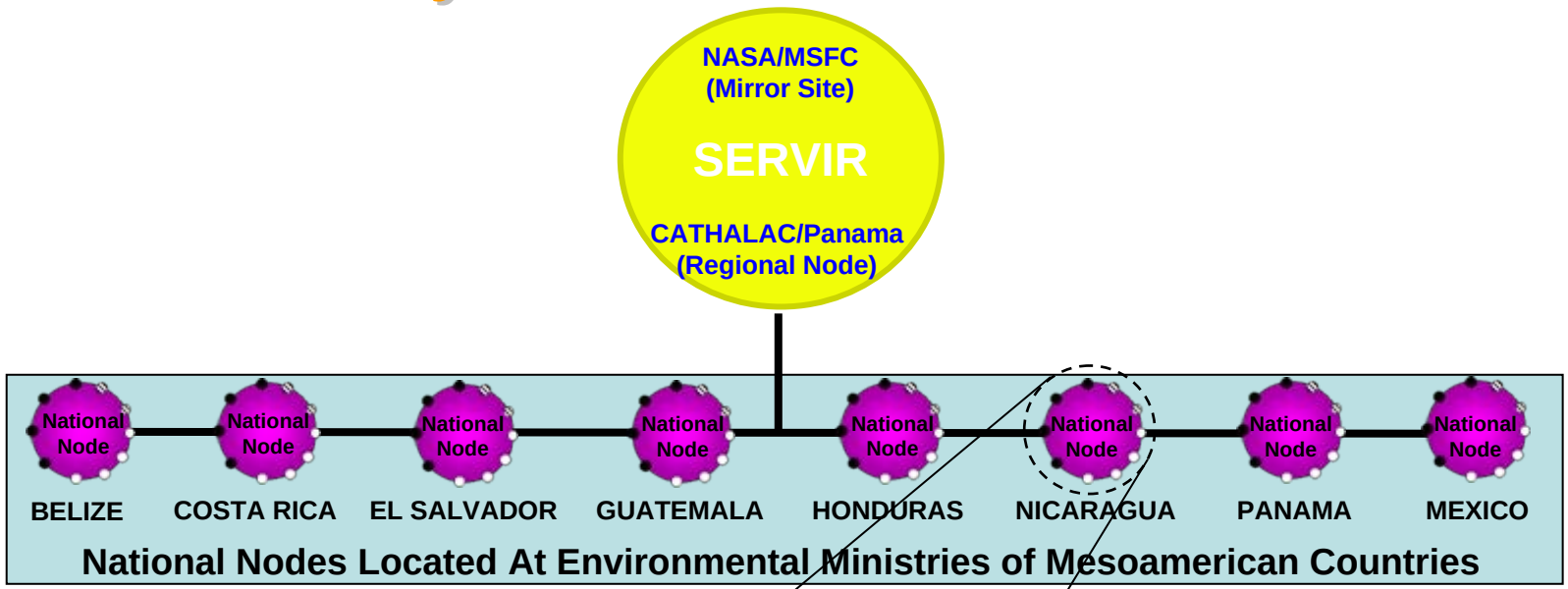


Peten, Guatemala (1986-1995)

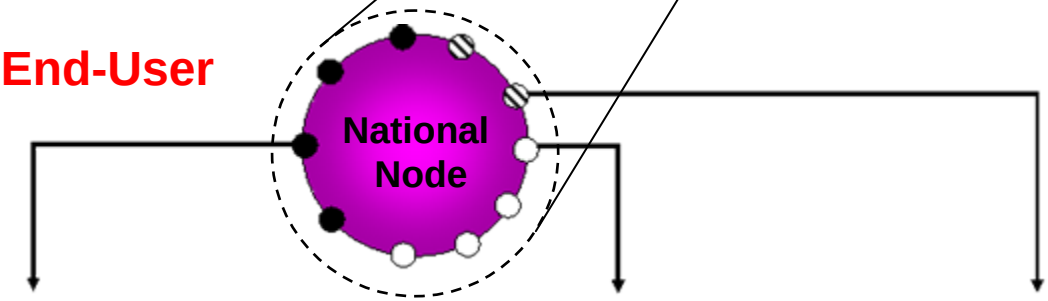
Remote Sensing over Central America



8 Country Institutional Infrastructure



National Node End-User Communities



PUBLIC INSTITUTIONS

- Geography/Mapping
- Climate/Meteorology/Hydrology
- Biodiversity (Forest, Protected Areas)
- Agriculture
- Disaster Prevention
- Other

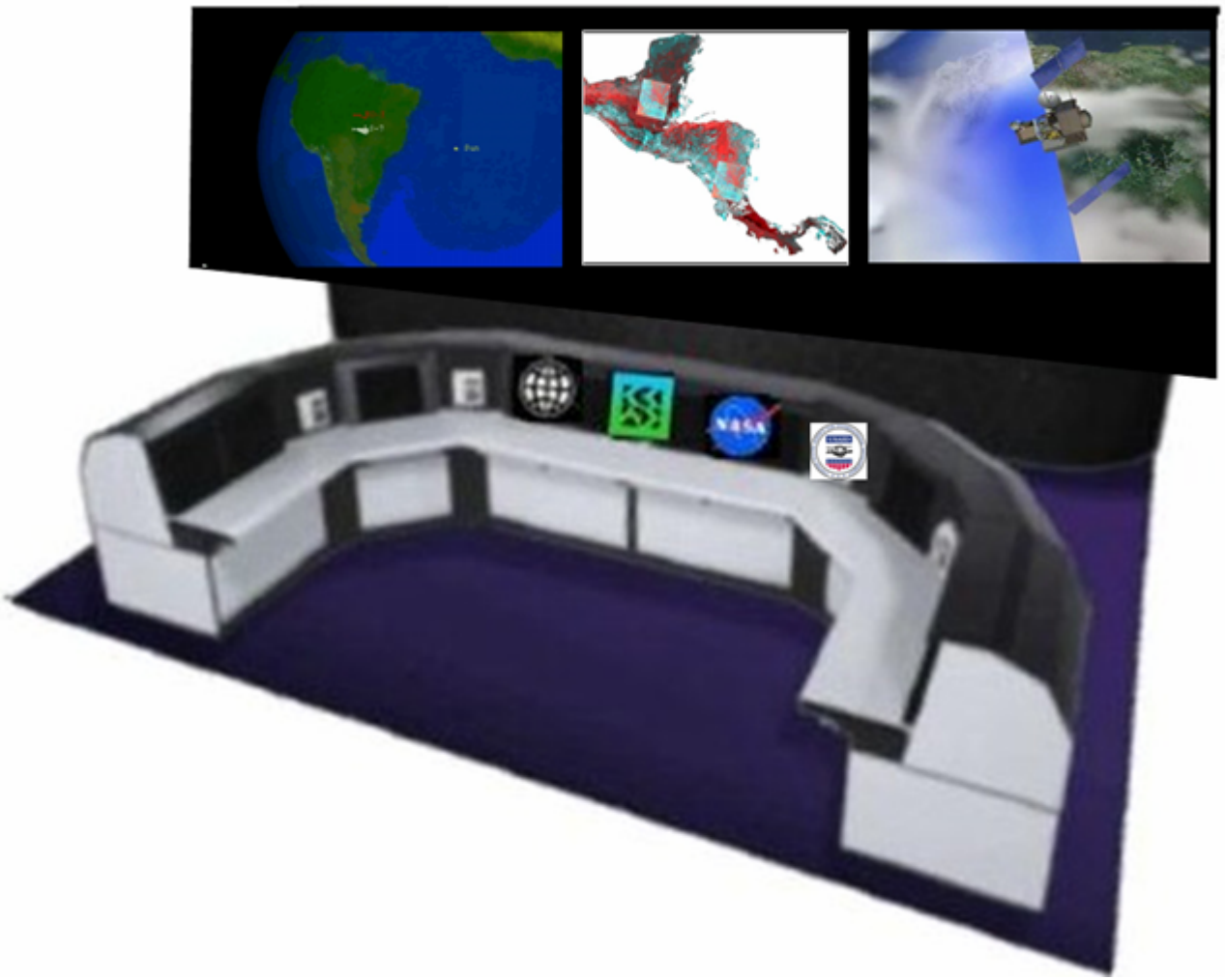
UNIVERSITIES & NON-GOVERNMENTAL

- Research/Educational
- Conservation
- Other

PRIVATE SECTOR

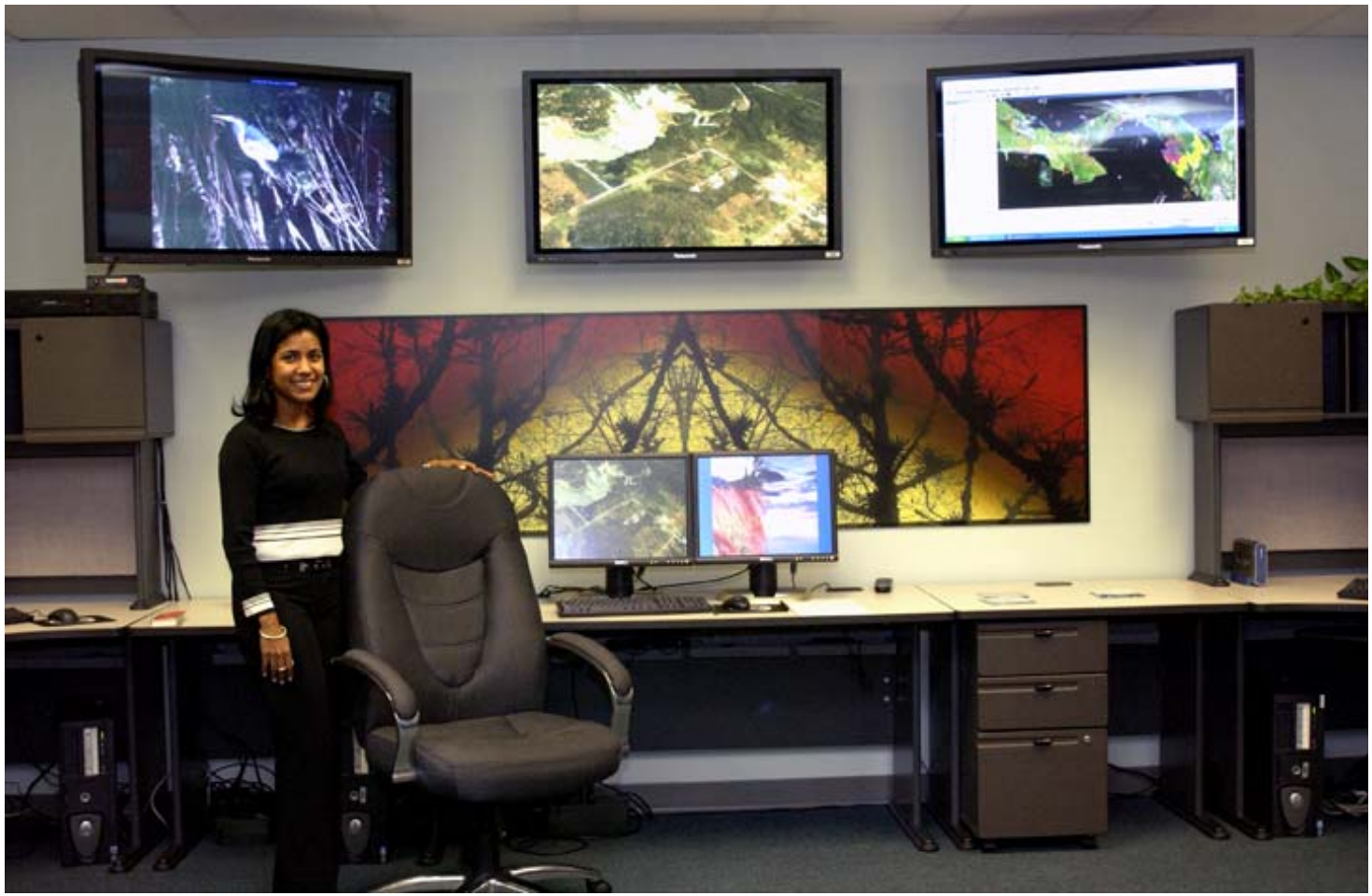
- Hydroelectric
- Aviation
- Other

SERVIR Concept Drawing - 2003



- Fire Detection
- Land Cover/Land Use
- Forest Monitoring
- Red Tides
- Climate Change
- Short Term Weather
- Drought Monitoring
- Inter. Scientific Research
- Carbon Flux
- Hot Spot Monitoring
- Disaster Mitigation
- Urban Studies

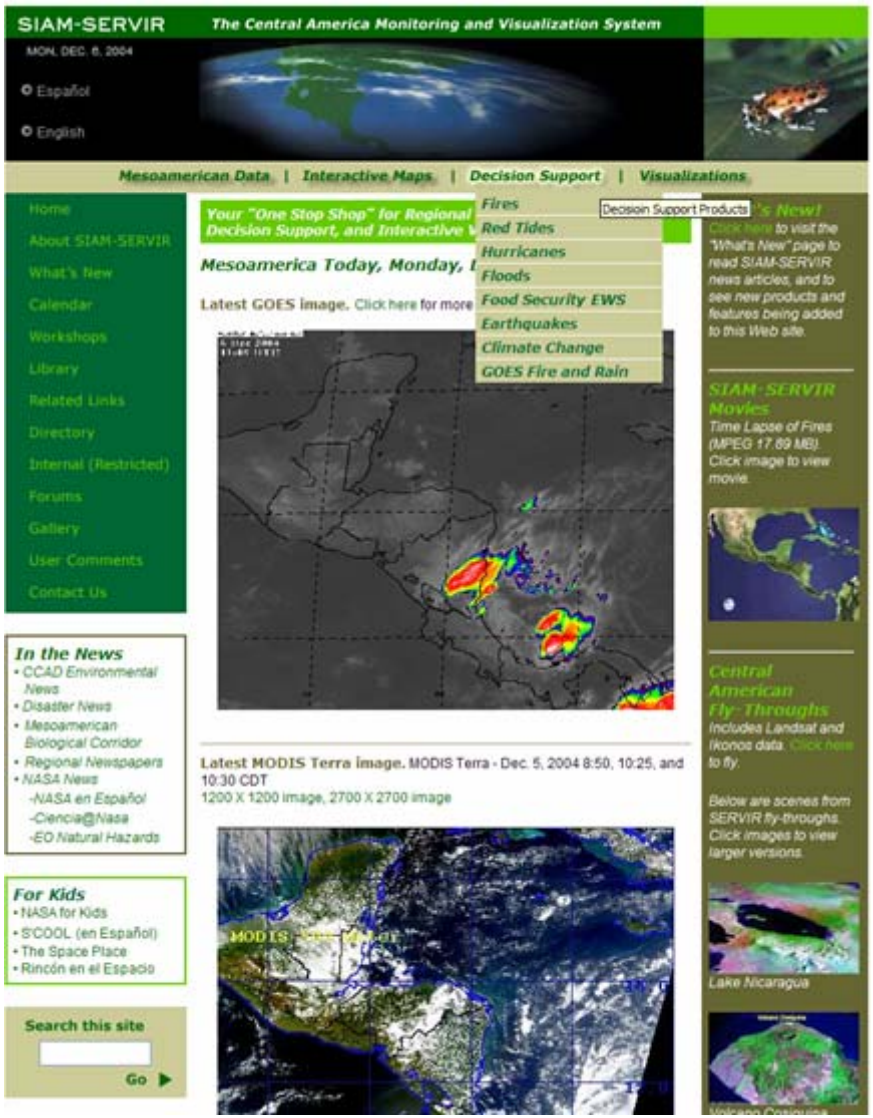
SERVIR Test Bed Node – NASA/MSFC in 2004



SERVIR Lab in Panama at City of Knowledge

SERVIR Dedication – February 3, 2005

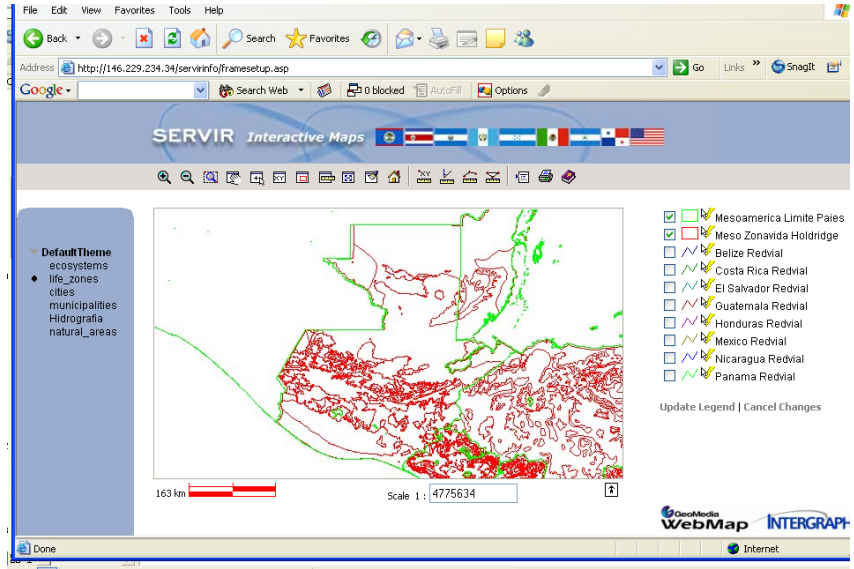
SERVIR Web Page
http://servir.nsstc.nasa.gov

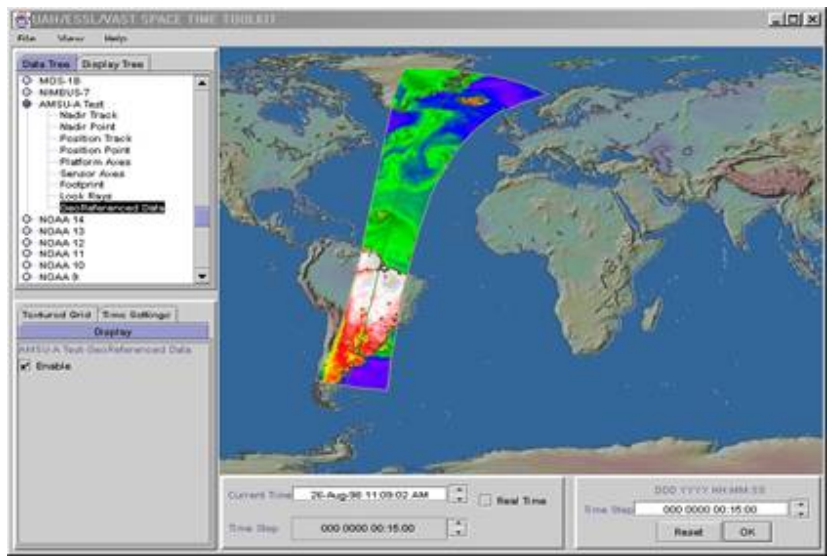


- Fully Operational by February 1, 2005
- Four Main Sections
  1. Mesoamerican Data
  2. Interactive Maps
  3. Decision Support
  4. Visualizations

Interactive On-Line Maps

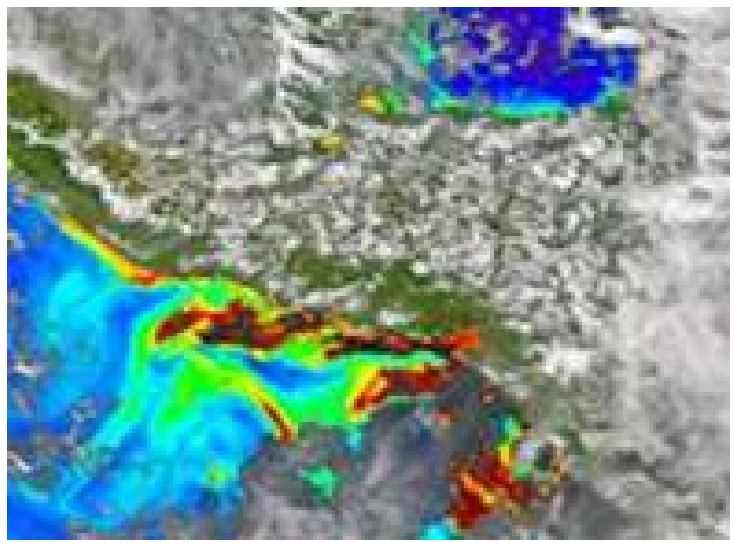
- Mapping for the “Casual User”
- Web Map Server at regional node
- GeoMedia WebMap
- Specialized Tools (UAH-Space Time Toolkit)
- Thematic Areas – Biological, Protected Areas, etc.
- User Friendly Thematic Queries





Decision Support Tools

- Fires
  - MODIS Rapid Response
  - Web Fire Mapper (UMD)
  - Fire Alerts (Pilot)
- Red Tides (Harmful Algae Blooms)
- Climate Change Scenarios
- Short Term Weather Prediction (SPORT)
- Land Cover and Land Use Change for Carbon Management
- Floods
- Other



Visualization Tools for Decision Support

- NASA and COTS integrated by UAH
  - World Wind - NASA
  - Skyline – Skyline Software
- Users
  - Decision Makers
  - Media
  - Educators
  - Students

Primary SERVIR Team Institutions

- NASA/MSFC
- CATHALAC
- USAID
- World Bank
- CCAD
- University of Alabama Huntsville
- University of Arkansas
- Science Systems and Applications Inc.
- Oak Ridge National Laboratory

SERVIR

An Environmental Monitoring and Decision Support System for Central America

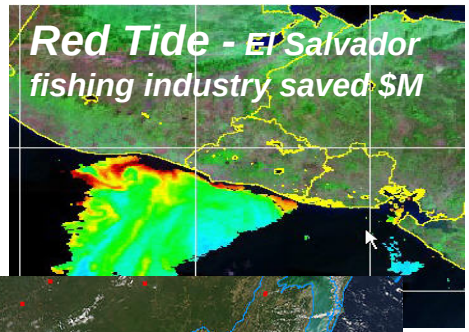
Earth Observatories

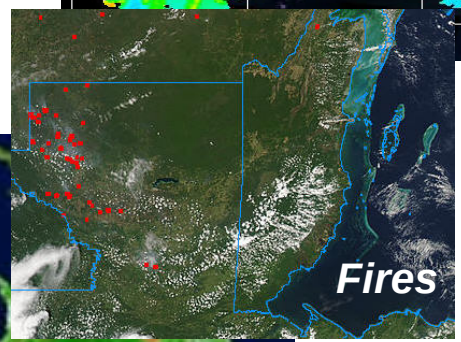
Electronic Transfer

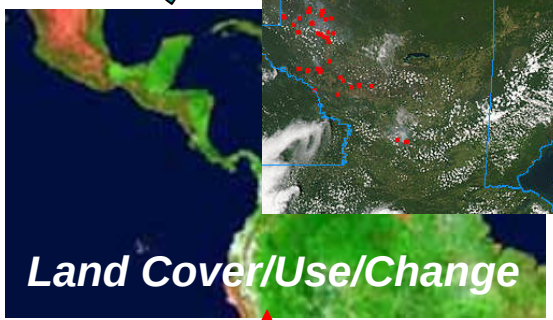
Central American Commission for Environment and Development

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

Environmental Monitoring & Decision Support Products







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

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

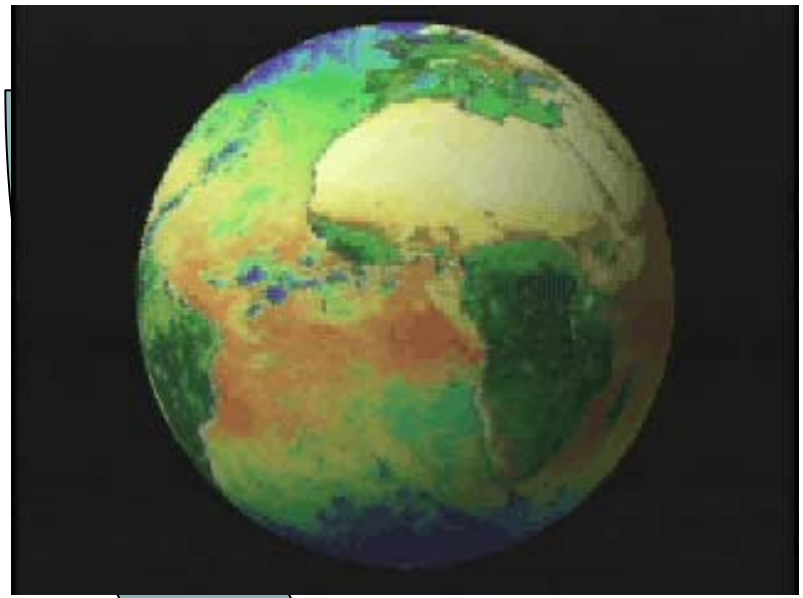
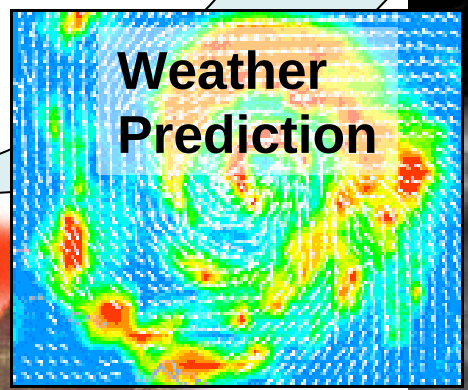
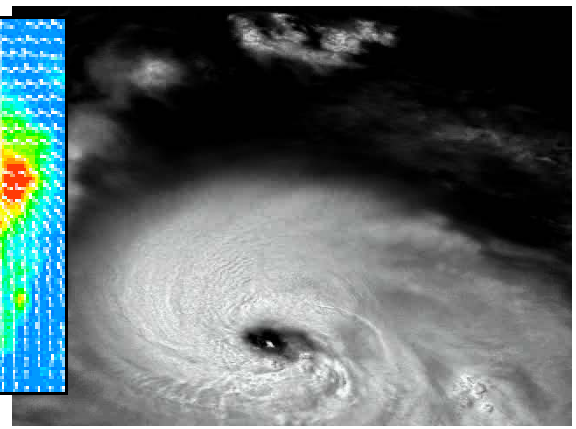
Goals

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

Data & Algorithms

SERVIR Partners

Challenges

- Develop and document common/standard interfaces for interoperability of data and services
- Design new data models for handling
  - real-time/streaming input
  - data fusion/integration
- Design and develop distributed standardized catalog capabilities
- Develop advanced resource allocation and load balancing techniques
- Exploit the Grid for enhanced data mining functionality
- Develop more intelligent and intuitive user interfaces
- Develop ontologies of scientific data, processes and data mining techniques for multiple domains
- Support language and system independent components
- Incorporate data mining into scientific curricula

Data Integration and Mining:
From Global Information to Local Knowledge