Continuing Challenges in the Exploration and Stewardship of Data

Committee on Earth Science and Applications From Space

Earth Science Data Symposium

October 4-5, 2016

Sara J. Graves, PhD

Director, Information Technology and Systems Center
UA System Board of Trustees University Professor
Professor, Computer Science Department
256-824-6064
sgraves@itsc.uah.edu

http://www.itsc.uah.edu

Outline

- **Looking Back to Look Forward**
- Data Systems and Technologies
- Methodologies for Exploration and Stewardship
- Rethinking Approaches for Exploration and Stewardship

Data Access, Integration and Stewardship Challenges for the Future

NASA Earth System Science at 20 Symposium
June 22-24, 2009

Sara J. Graves

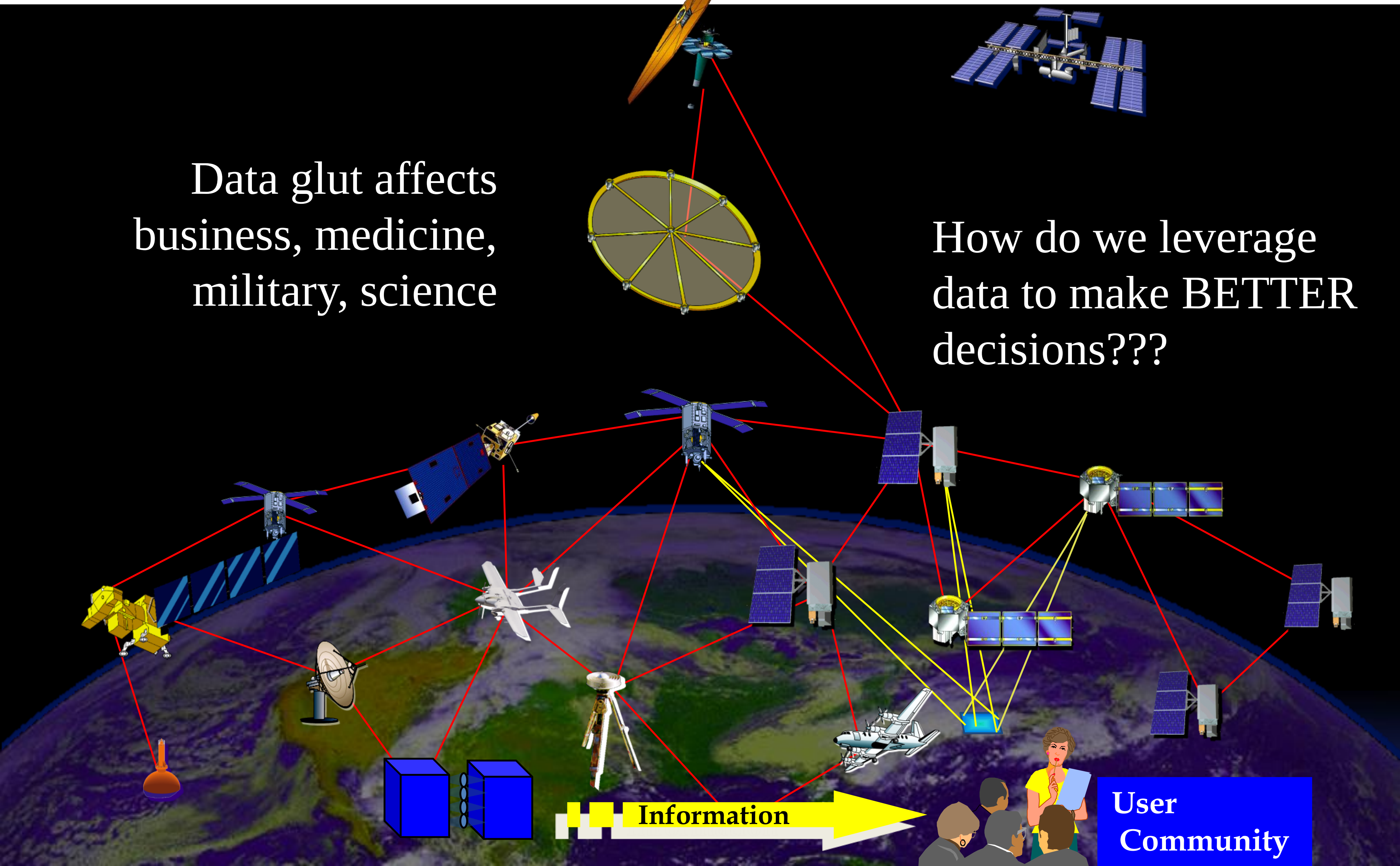
Director, Information Technology and Systems Center
Board of Trustees University Professor
Professor of Computer Science
University of Alabama in Huntsville
256-824-6064
sgraves@itsc.uah.edu

UAHuntsville THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

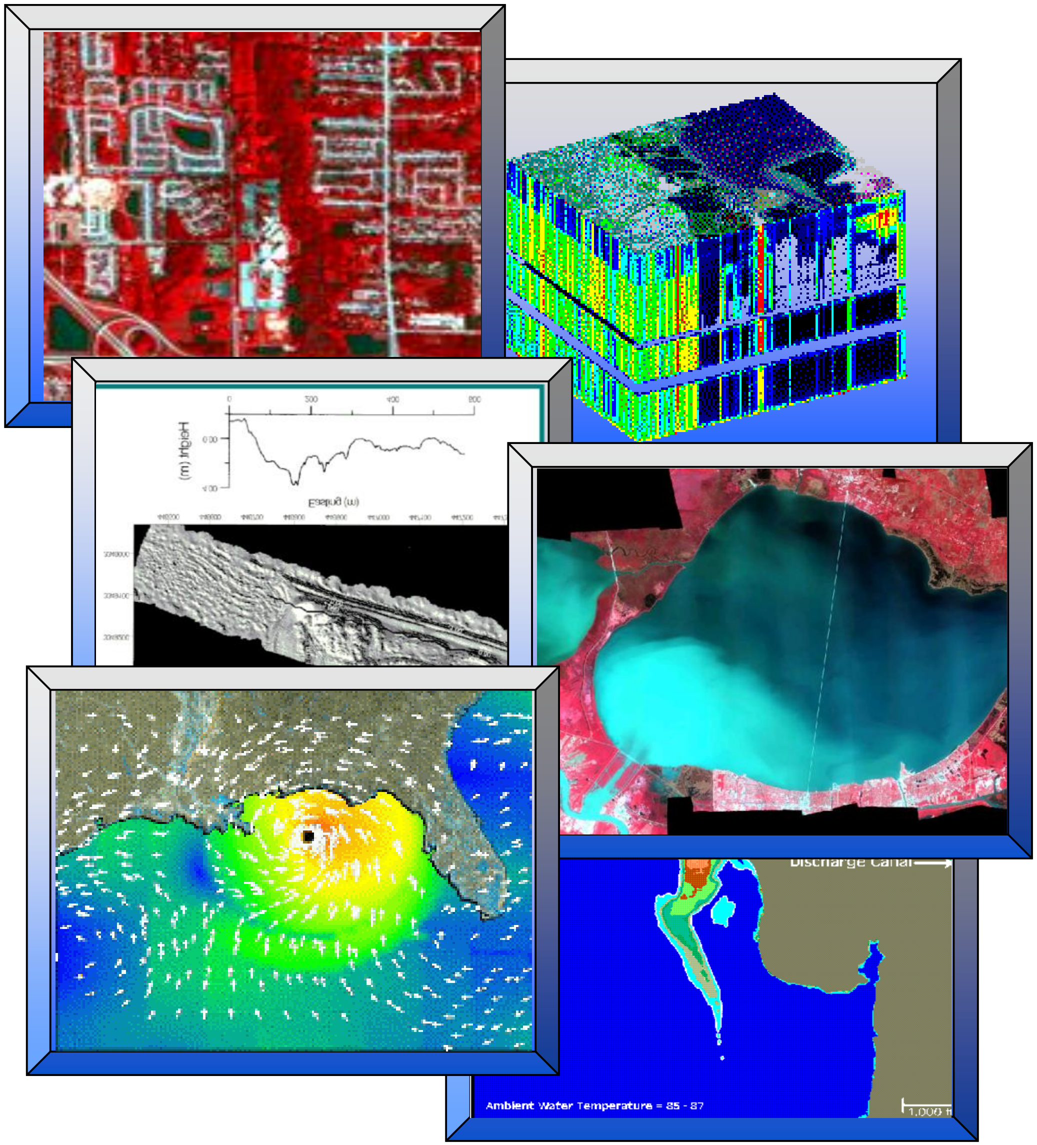
http://www.itsc.uah.edu

"…drowning in data but starving for knowledge"

2009



Heterogeneity Leads to Data Usability Problems



Data Characteristics

- Many different formats, types and structures
- Different states of processing (raw, calibrated, derived, modeled or interpreted)
- Enormous volumes

2009

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

2009

Characteristics of Adaptable Data Systems and Services

- Heterogeneous
  - participants (investigators & institutions)
  - data and services
  - technological approaches (many capabilities exist and many more to be developed)
- Distributed, adaptable and flexible, responsive systems
- Smaller, more manageable pieces
- Establish a framework to integrate activities.
  - define a core set of interface standards and practices
  - utilize community-wide interface standards

2009

Outline

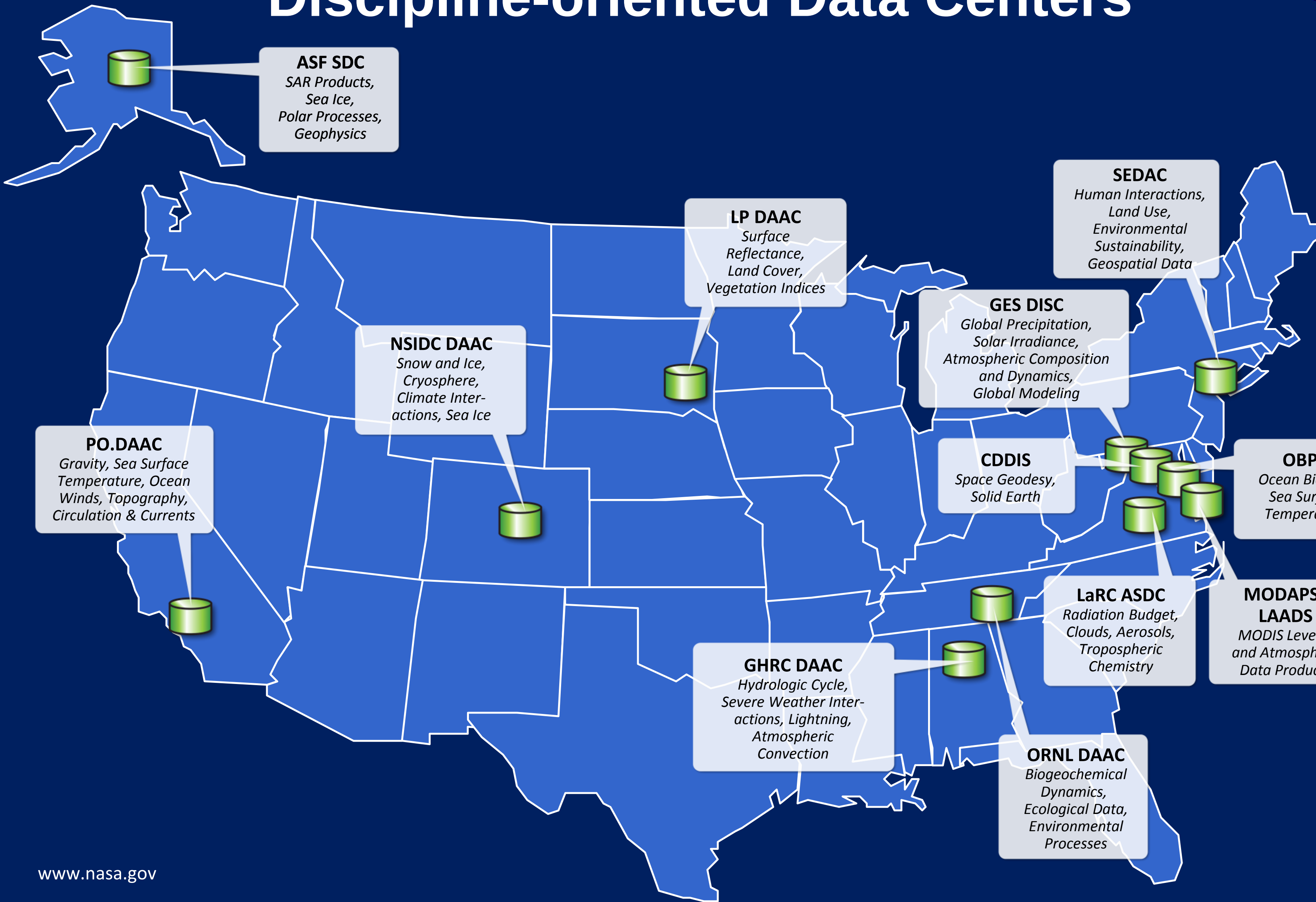
- Looking Back to Look Forward
- **Data Systems and Technologies**
- Methodologies for Exploration and Stewardship
- Rethinking Approaches for Exploration and Stewardship

NASA’s Earth Science Data Systems (ESDS)

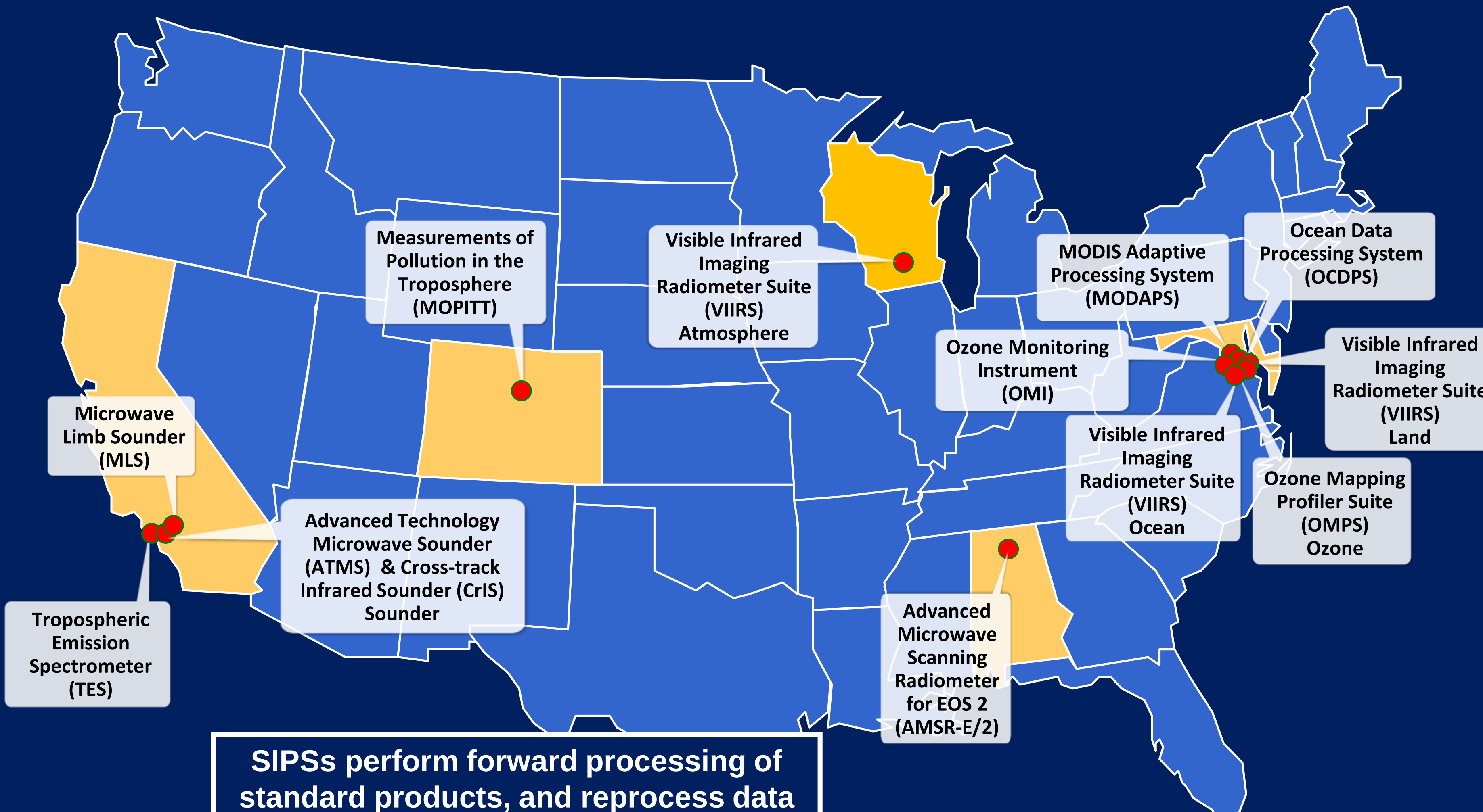
- NASA’s Earth Science Data Systems provide end-to-end capabilities to deliver data and information products to users using an “open data policy”
  - Data available to all users with no period of exclusive access
  - Users obtain most of the data with no charge for distribution
    - NASA waives the cost of dissemination for NASA-generated products so as not to inhibit use.
    - NASA restricts access and/or charges distribution fees only to the extent required by the appropriate MoU for data products supplied from an international partner or other agency
  - Distributed, heterogeneous system – a “virtual data system” or “System of Systems”
  - Architecture is discipline, measurement, and NASA research programmatically based.

National Aeronautics and Space Administration

Discipline-oriented Data Centers

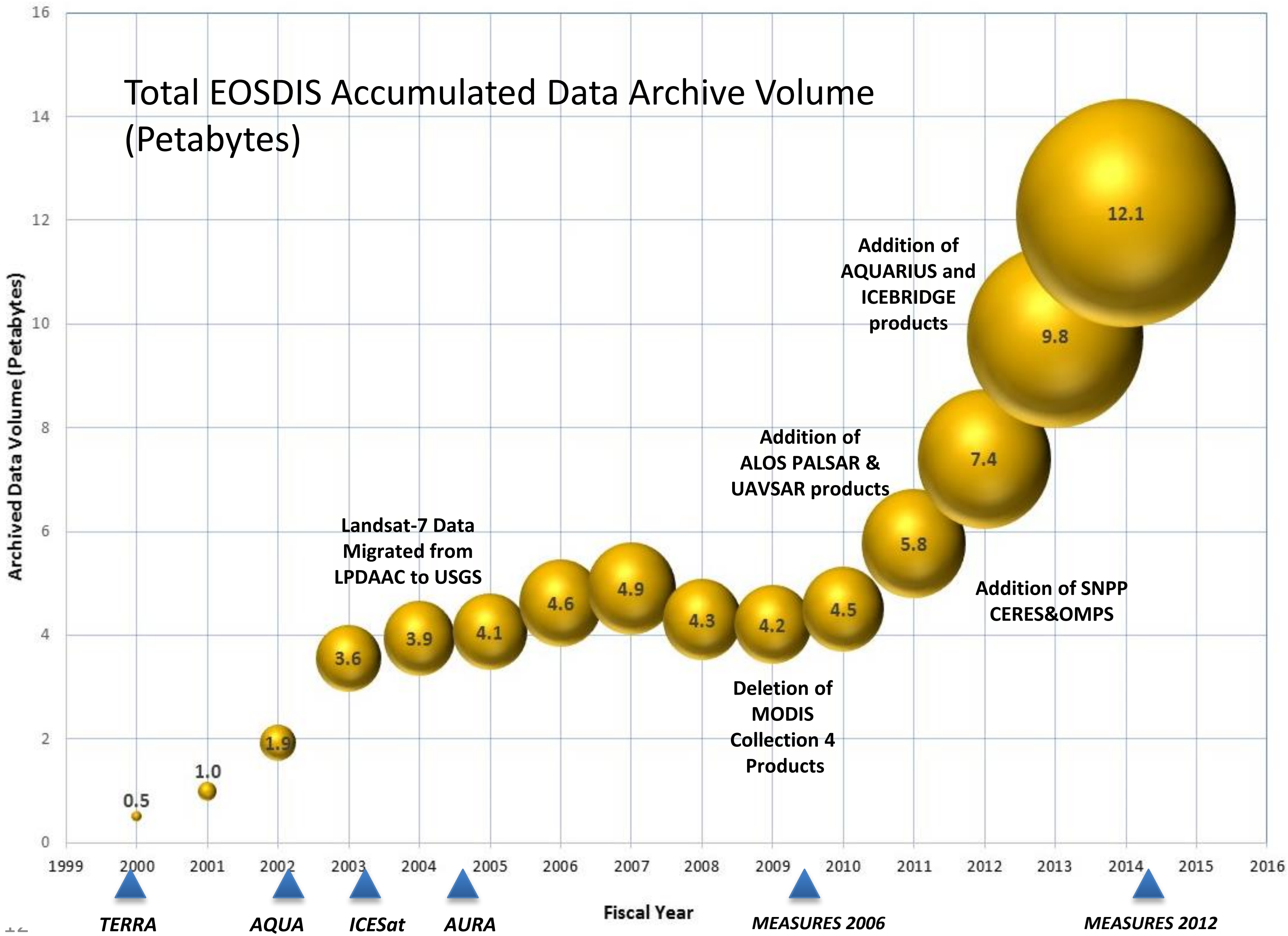


Science Investigator-led Processing Systems

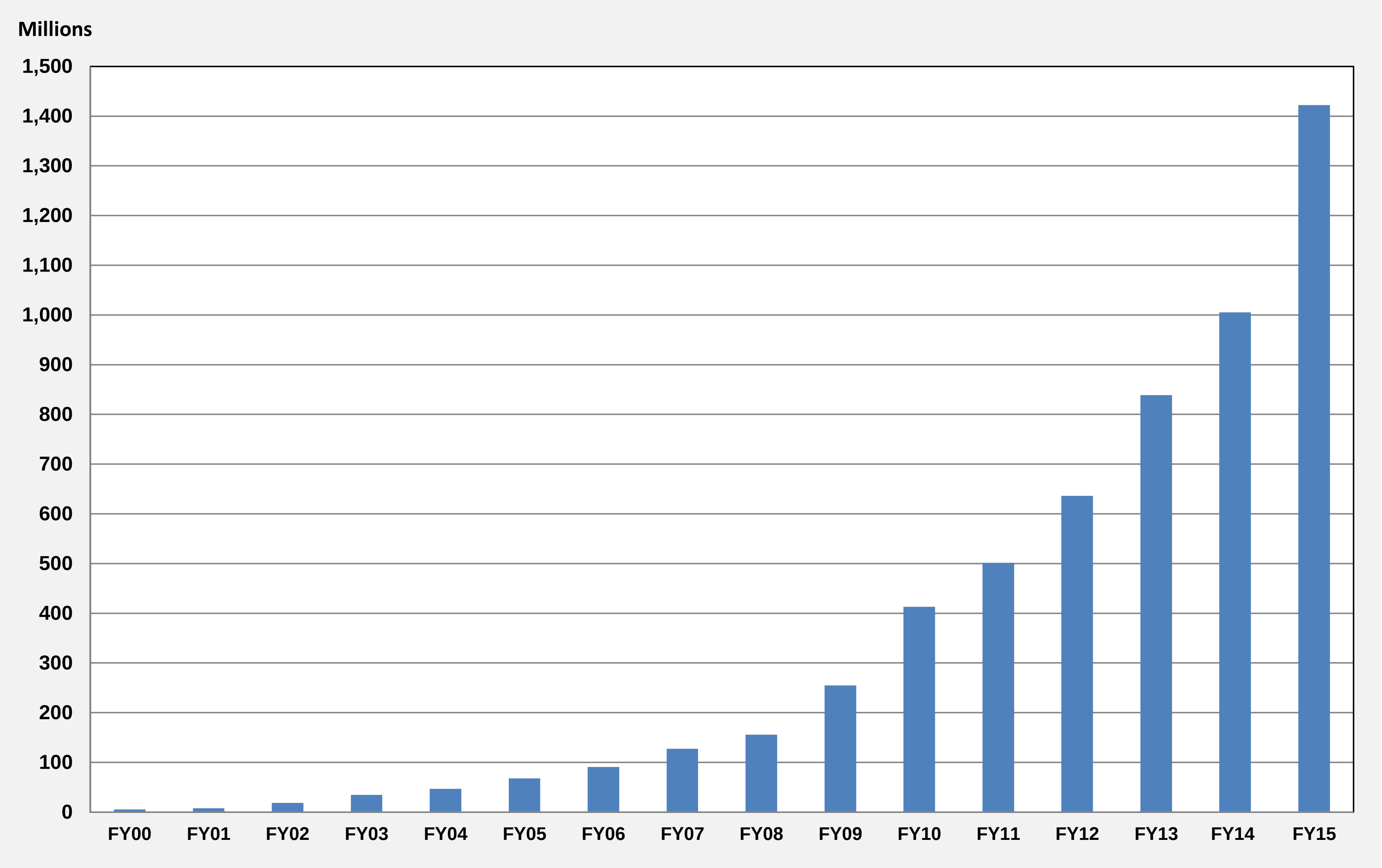


SIPSs perform forward processing of standard products, and reprocess data to incorporate algorithm improvements.

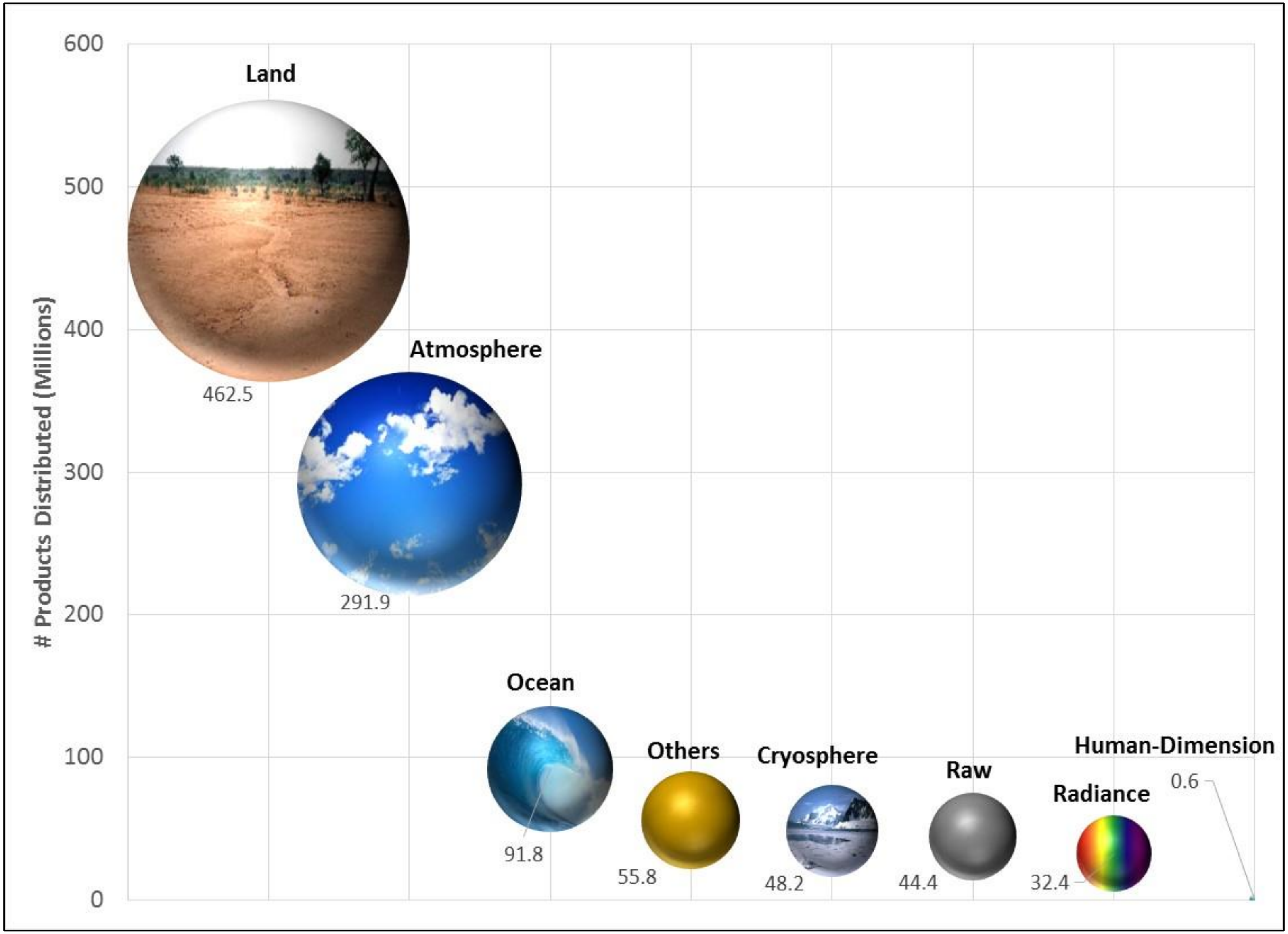
15+ years of Earth Science Data



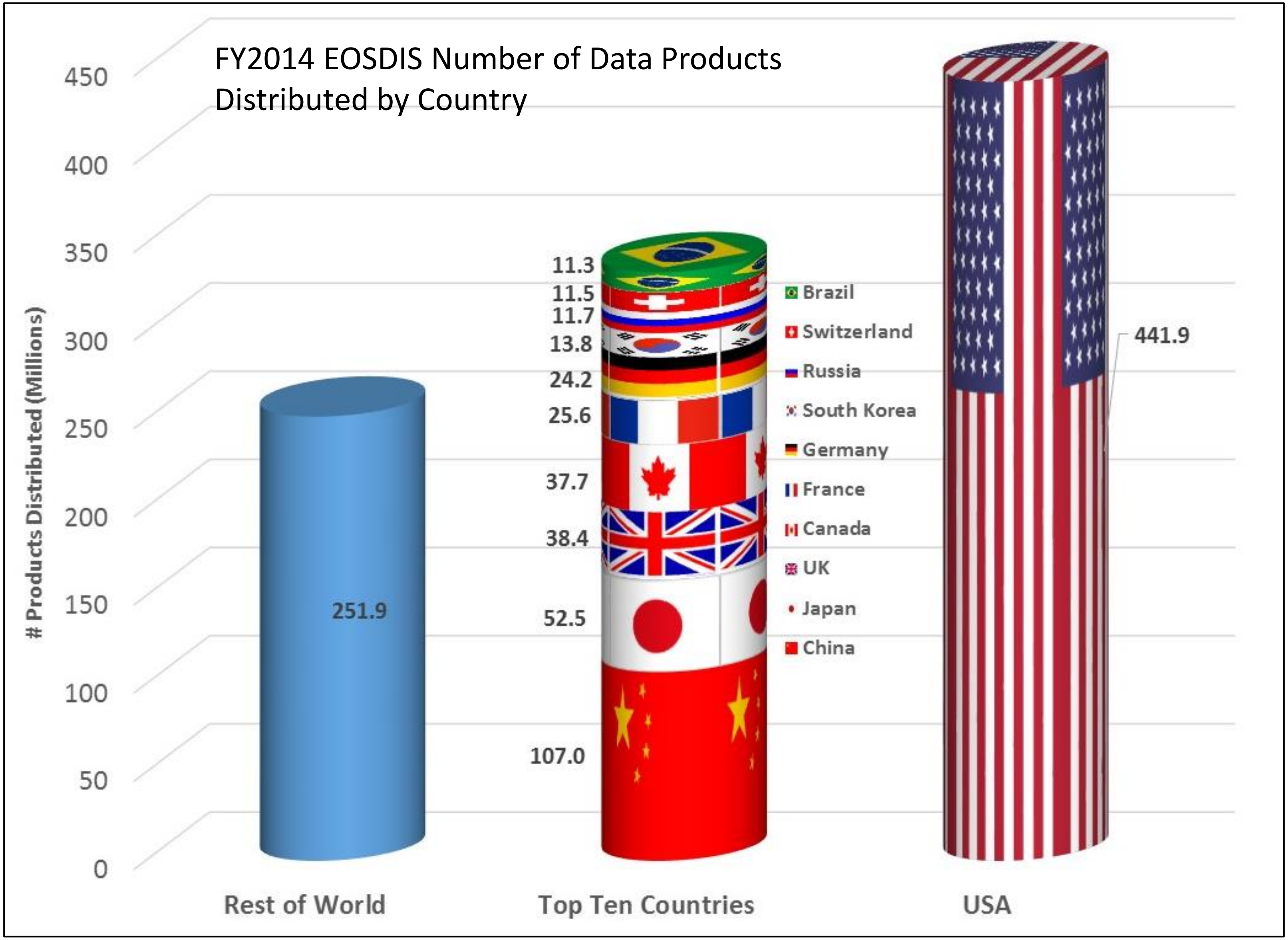
EOSDIS Products Delivered: FY2000 thru FY2015



FY2014 Number of Files Distributed by Discipline

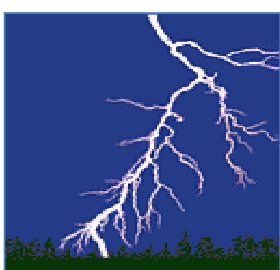


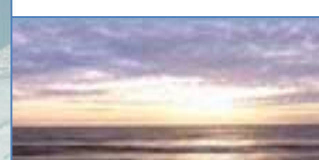
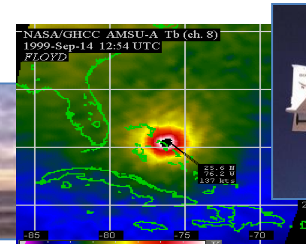
EOSDIS Distribution is World-wide



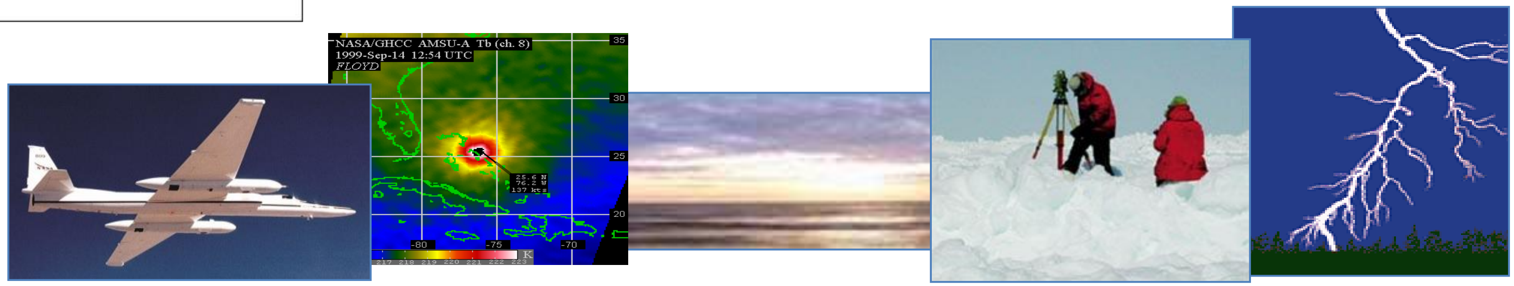
GHRC
Global Hydrology Resource Center
One of NASA's Earth Science Data Centers

- **Full service data center** providing data ingest, routine and custom processing, archive, distribution, user support, and science data services
- **Collaboration** between NASA and the University of Alabama in Huntsville to apply **advanced information technologies** to a variety of **science data projects**
- Global **lightning data from space**, airborne and ground based observations from **hurricane science field campaigns** and Global Precipitation Mission **(GPM) ground validation** experiments, and **satellite passive microwave** products



Global Hydrology Resource Center:
Data Systems and Services for Earth Science

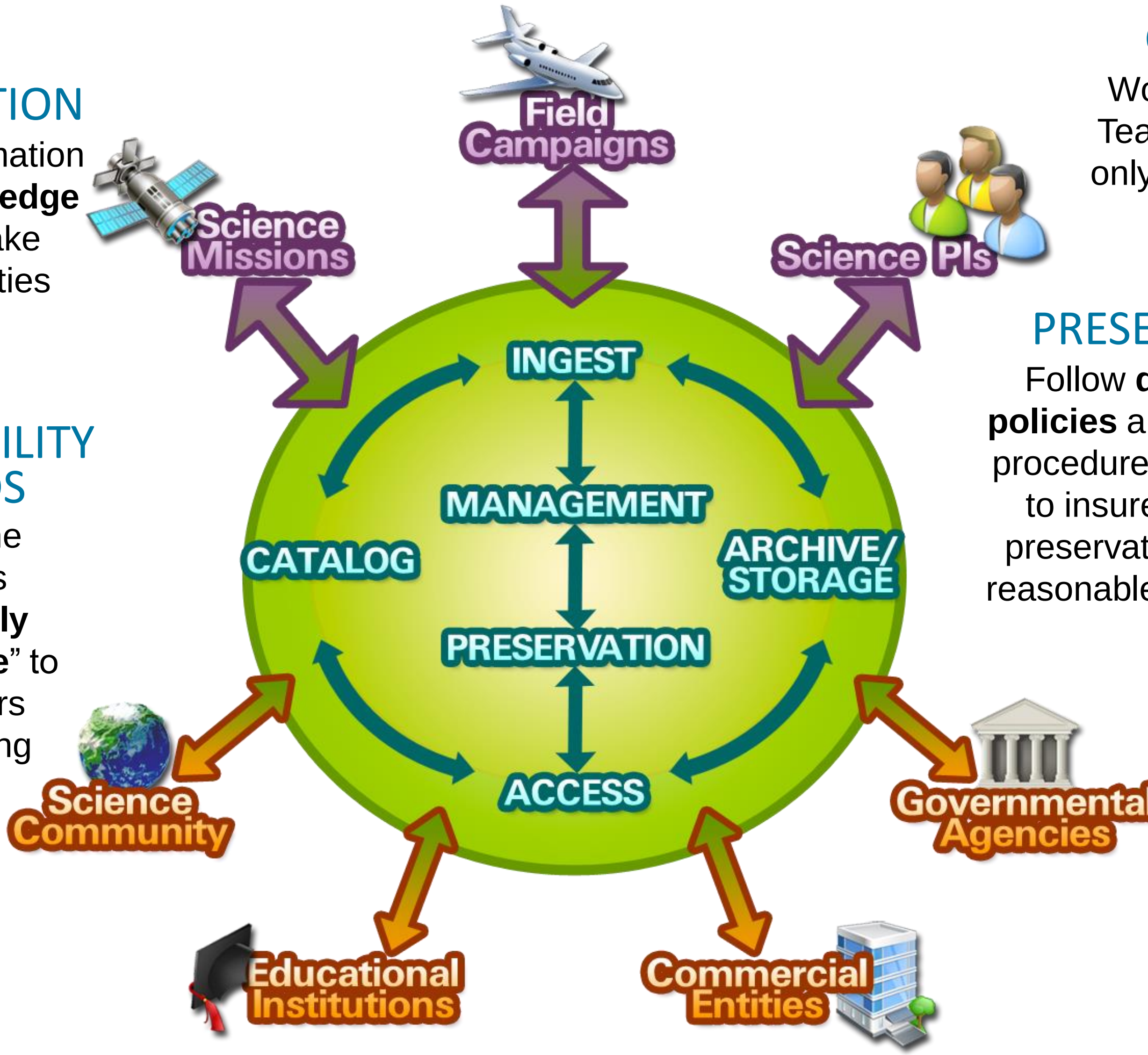


- **LIS SCF** with **TRMM LIS (**1997-2015), **ISS LIS** (2017 -) *PI-led data center, for intra cloud and cloud to ground lightning products from satellite and surface observations.*
- **Hurricane field campaigns HS3** (Fall 2012, 13, 14)**, GRIP** (Fall 2010), **TC4** (Summer 2007) **NAMMA** (Fall 2006), **TCSP** (Summer 2005), **ACES** (Fall 2002), **CAMEX 3** and **4** (Fall 1998 and 2001) : *Web-based collaboration for intra-project communications before, during, and after campaigns. Real-time mission monitoring. Data acquisition and integration from multiple instruments.*
- **GPM Ground Validation** (2010 - 2016) : *Collaboration tools and archive for a variety of ground validation datasets related to Global Precipitation Mission, including recent* **LPVEx** (Fall 2010*)*, **MC3E** (Spring 2011), **GCPEx** (Winter 2012), **IFloodS** (Spring 2013) *and upcoming* **IPHEX** (Spring 2014), **OLYMPEX** (Fall/Winter 2015-16)
- **DISCOVER** (2003 – 2015) and **Passive Microwave ESIP** (1998-2003) : *With Remote Sensing Systems, providing long-term climate data records via easy-to-use display and access tools.*
- **AMSR SIPS** (1999 –) : *Fully automated generation and delivery of research quality standard data products; on-demand subsetting for ground validation sites for* ***AMSR-E*** *and* ***AMSR2***
- **LANCE AMSR2** (2015 -) : *Near-real-time product generation to support low-latency applications*

Data Stewardship

DOCUMENTATION
Capture this information to **create a knowledge base** for our stake holder communities

INTEROPERABILITY STANDARDS
Ensure that the information is "**independently understandable**" to all stakeholders without requiring experts



CURATION
Work with Science Teams to gather not only data but also **all relevant information**

PRESERVATION
Follow **documented policies** and **engineered** procedures at every step to insure information preservation against all reasonable contingencies

PROCESSING
include **science product generation** and reformatting, algorithm integration and test, interfaces with external providers

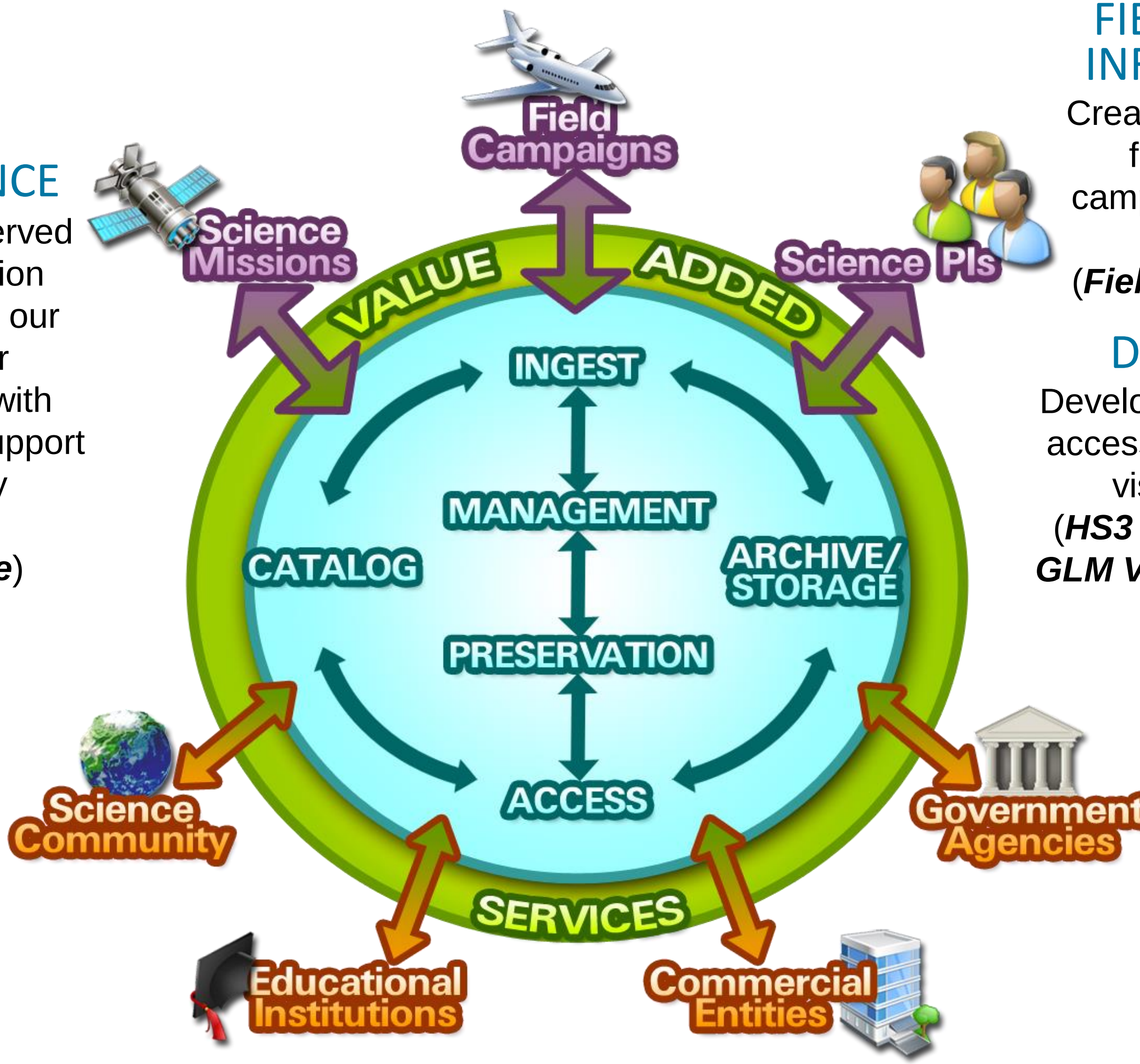
Data Stewardship in the Global Hydrology Resource Center (GHRC)

itsc
http://www.itsc.uah.edu

Knowledge Augmentation Services

PROVENANCE
Make the preserved data/information available to all our stakeholder communities with traceability to support authenticity (***AMSR-E Provenance***)

DATA DISCOVERY
Develop new tools for data discovery, curation and aggregation (***LIS Interactive Browse***)



FIELD CAMPAIGN INFRASTRUCTURE
Create specialized portals for managing field campaigns and collecting data (***Field Campaign Portal***)

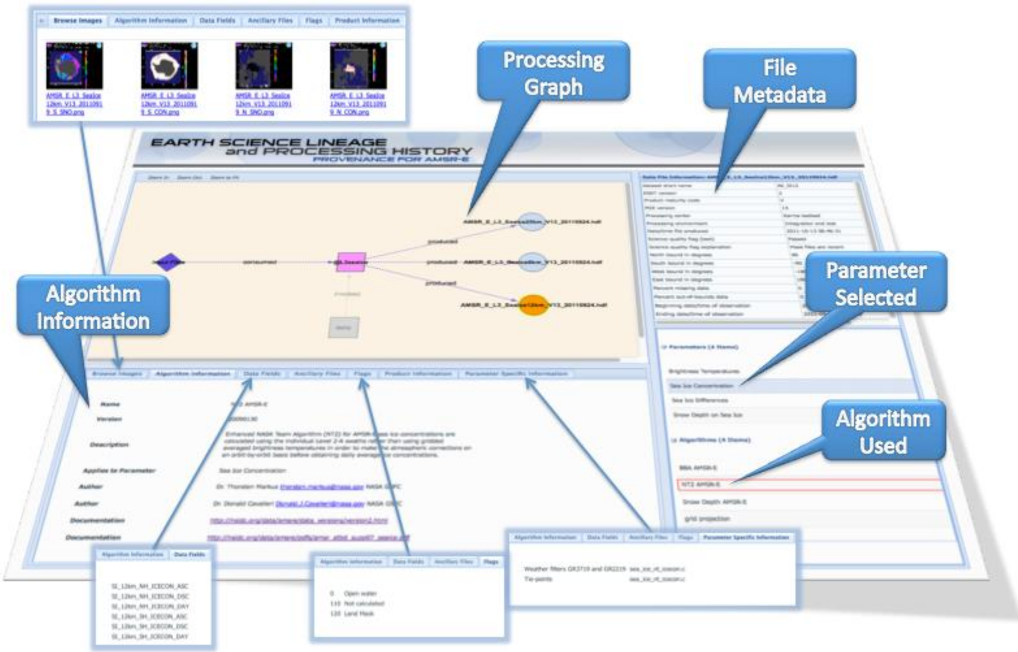
DATA USE
Develop new tools for access, analysis and visualization (***HS3 Data System, GLM Validation Tool, RASI***)

INFUSING CUTTING EDGE INFORMATICS
Research new approaches and technologies and infuse them into operational processes

GHRC provides knowledge augmentation services encompassing tools, infrastructure, user support, and expertise to our stakeholders

itsc
http://www.itsc.uah.edu

Provenance



- Provenance of hardware, software, and data best captured at **point of origin**
- Provenance **standards** and interoperability are desirable for models
- Provenance or "**chain of custody**" is used to verify trustworthiness, reliability, reproducibility, and security

- Provenance is important in **representation**, **management**, **presentation**; also system engineering and legal, policy and economic issues
- Provenance and context information, such as pedigree and tracking, may include **imagery, data fields, flag values and other information**, in addition to a provenance graph and metadata

LANCE: NASA Near Real-Time Data and Imagery

Land, Atmosphere Near real-time Capability for EOS (LANCE)

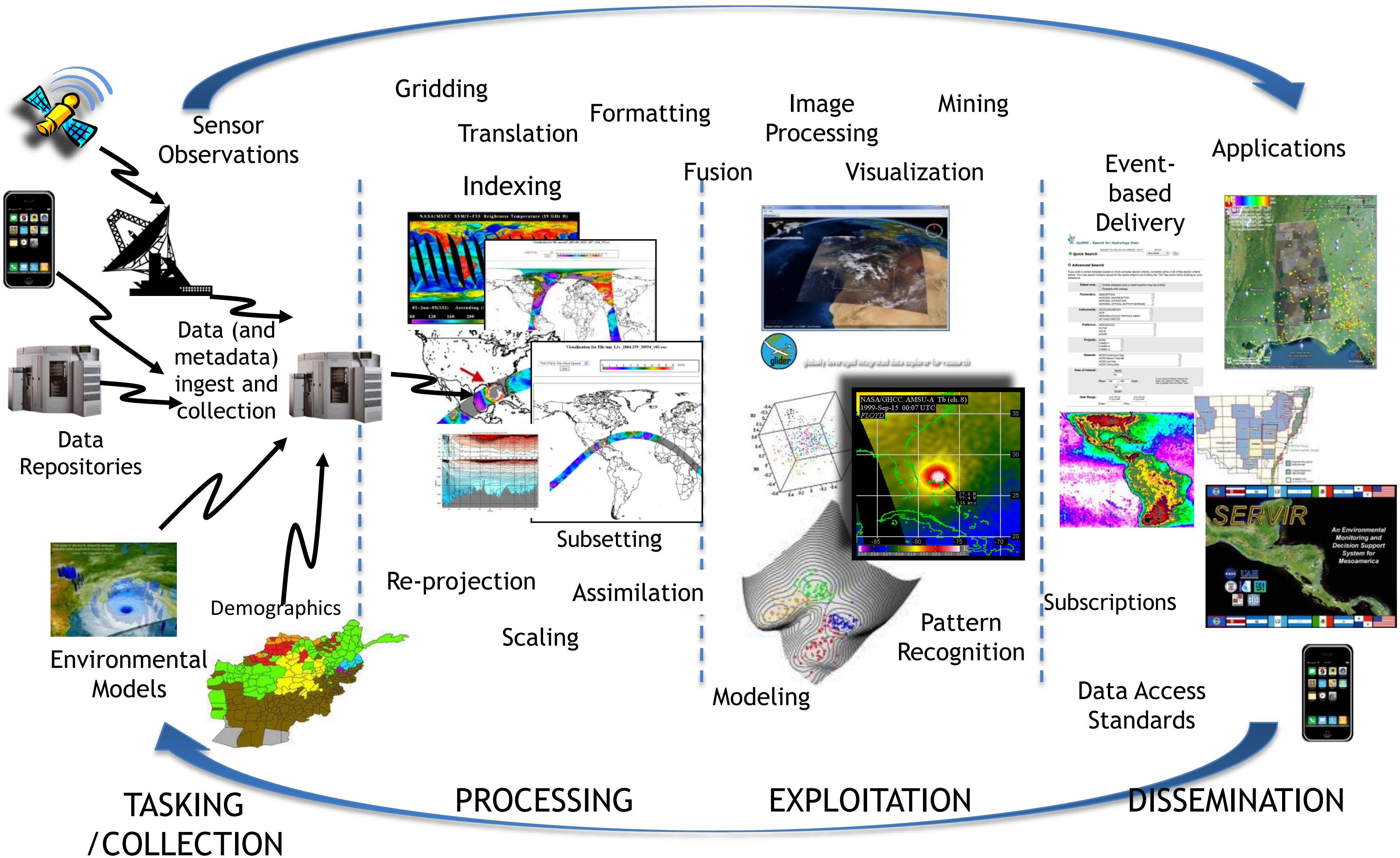
- Near real-time data and imagery from AIRS, AMSR2, MISR, MLS, MODIS, OMI and VIIRS instruments
- Most data products are available within 3 hours from satellite observation.
- NRT imagery are generally available 3-5 hours after observation.
- Tailored for application users interested in monitoring a wide variety of natural and man-made phenomena.

https://earthdata.nasa.gov/earth-observation-data/near-real-time

Outline

- Looking Back to Look Forward
- Data Systems and Technologies
- **Methodologies for Exploration and Stewardship**
- Rethinking Approaches for Exploration and Stewardship

Data-intensive Science Conceptual Framework for Multi-source, Multi-function Analysis

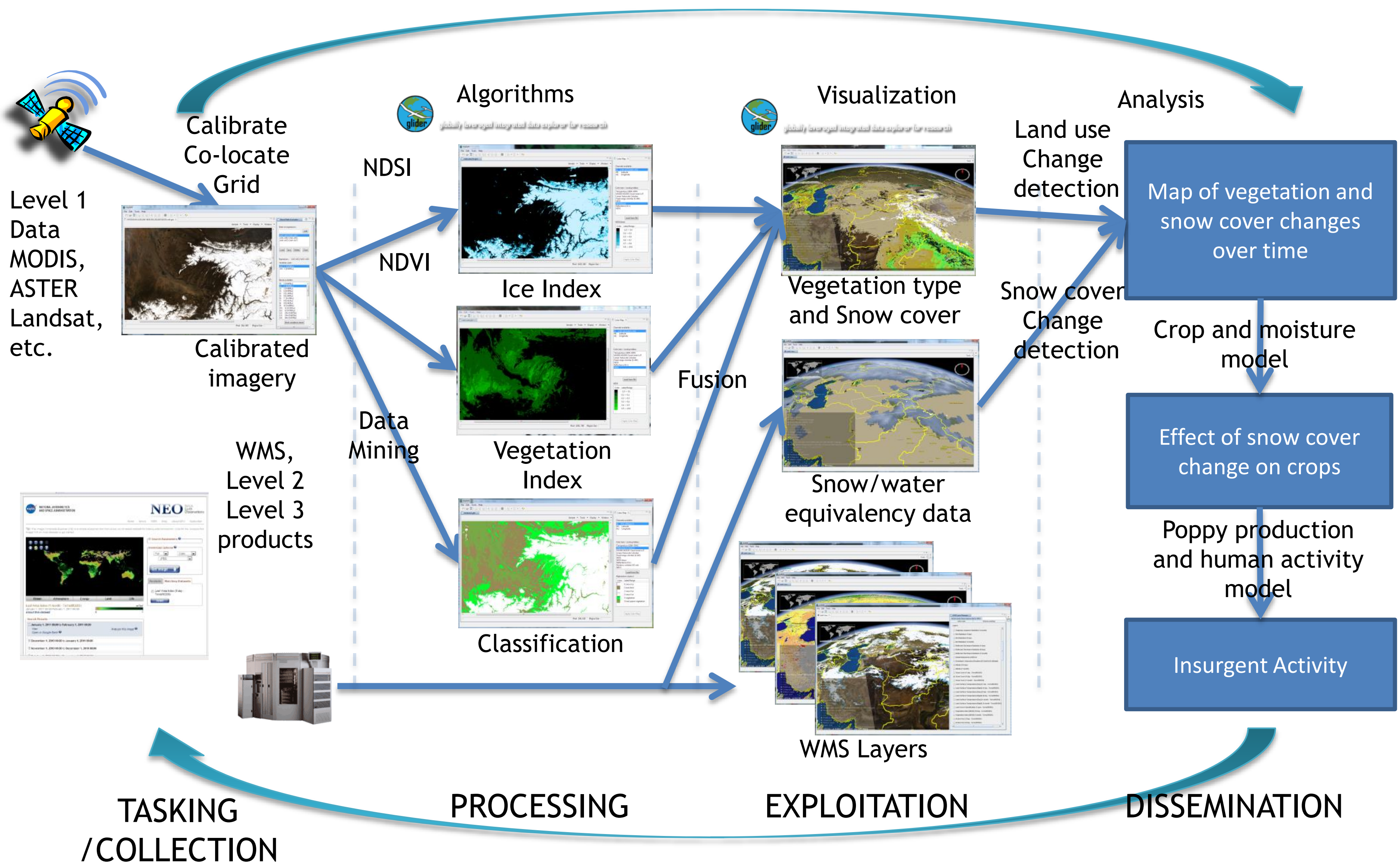


Environmental Impacts on National Security

Mountain Snow Cover in Afghanistan

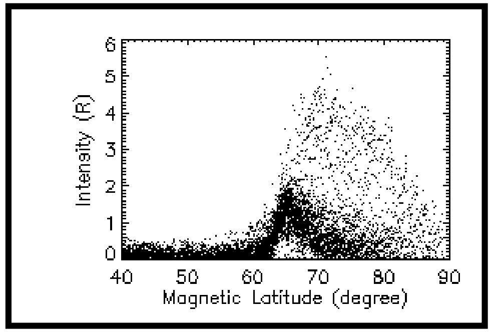
- Coupled System Hypothesis: Winter mountain snow cover and poppy crop production
- Assumed human impact: Increase in snow cover leads to increased poppy and food crop yields
  - Poppy crop production funds insurgent activity
  - Increase in poppy crop leads to increased insurgent activity
  - Decrease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

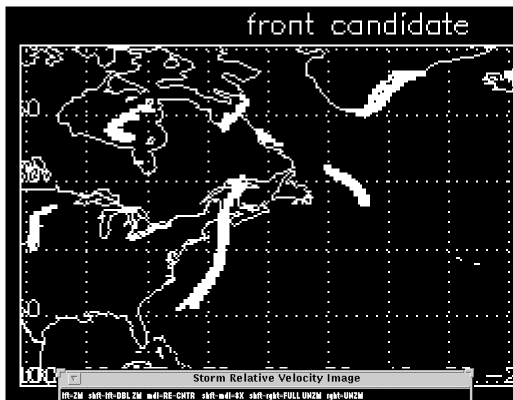
Mountain Snow Cover Process

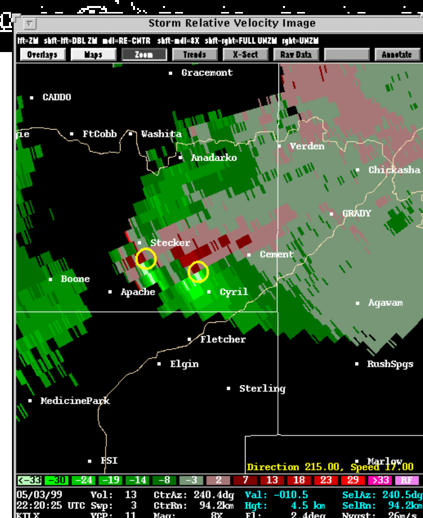


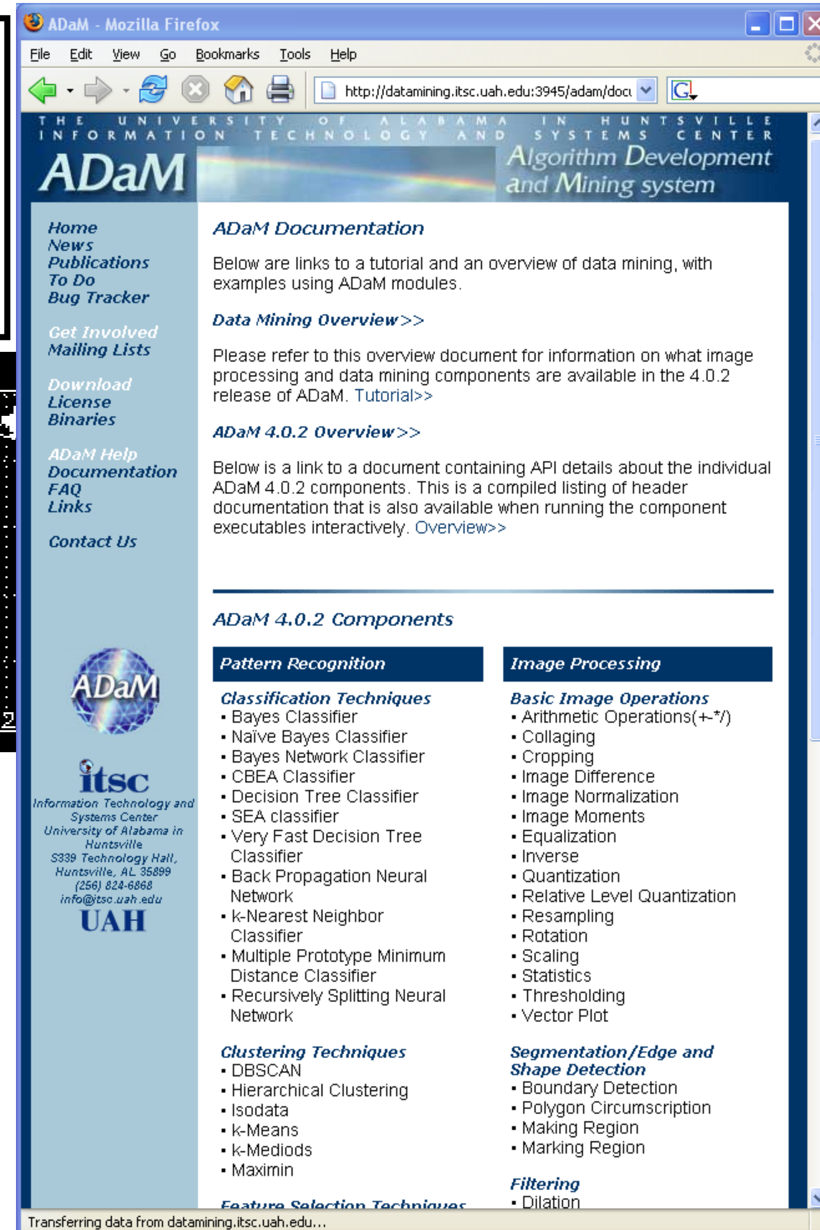
Data Mining: Algorithm Development and Mining (ADaM) Toolkit

- UAHuntsville has been at the forefront of mining sensor data for over 20 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







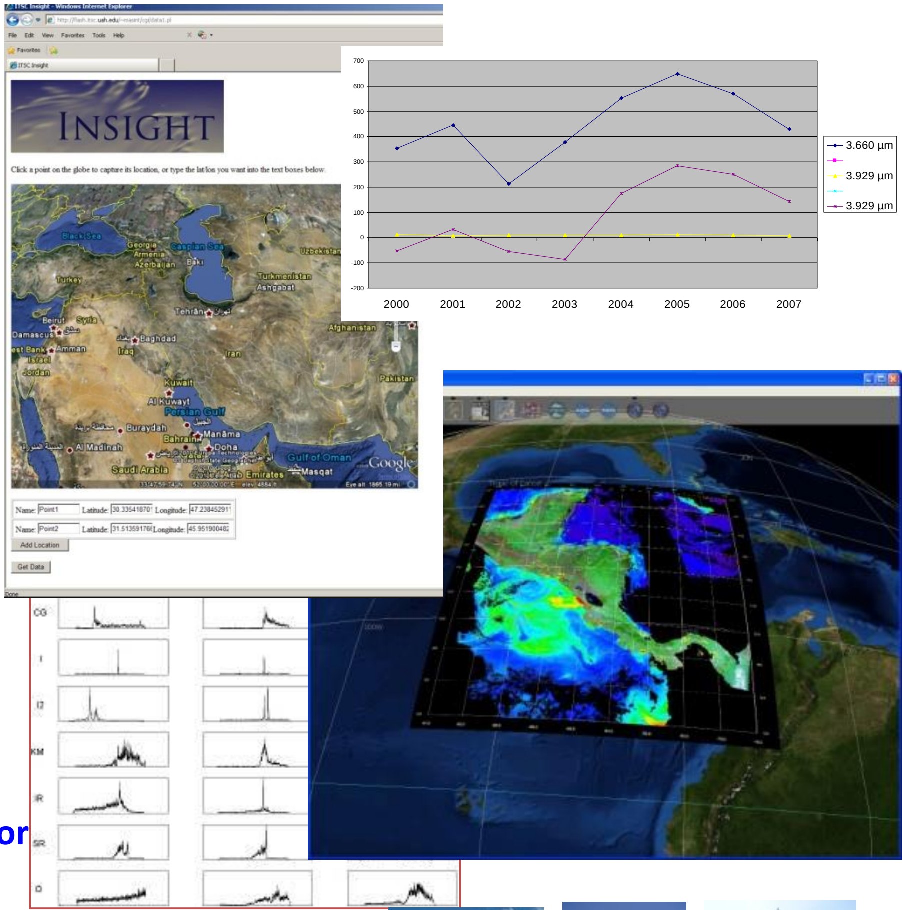


Data Mining: Situational Awareness and Analysis

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

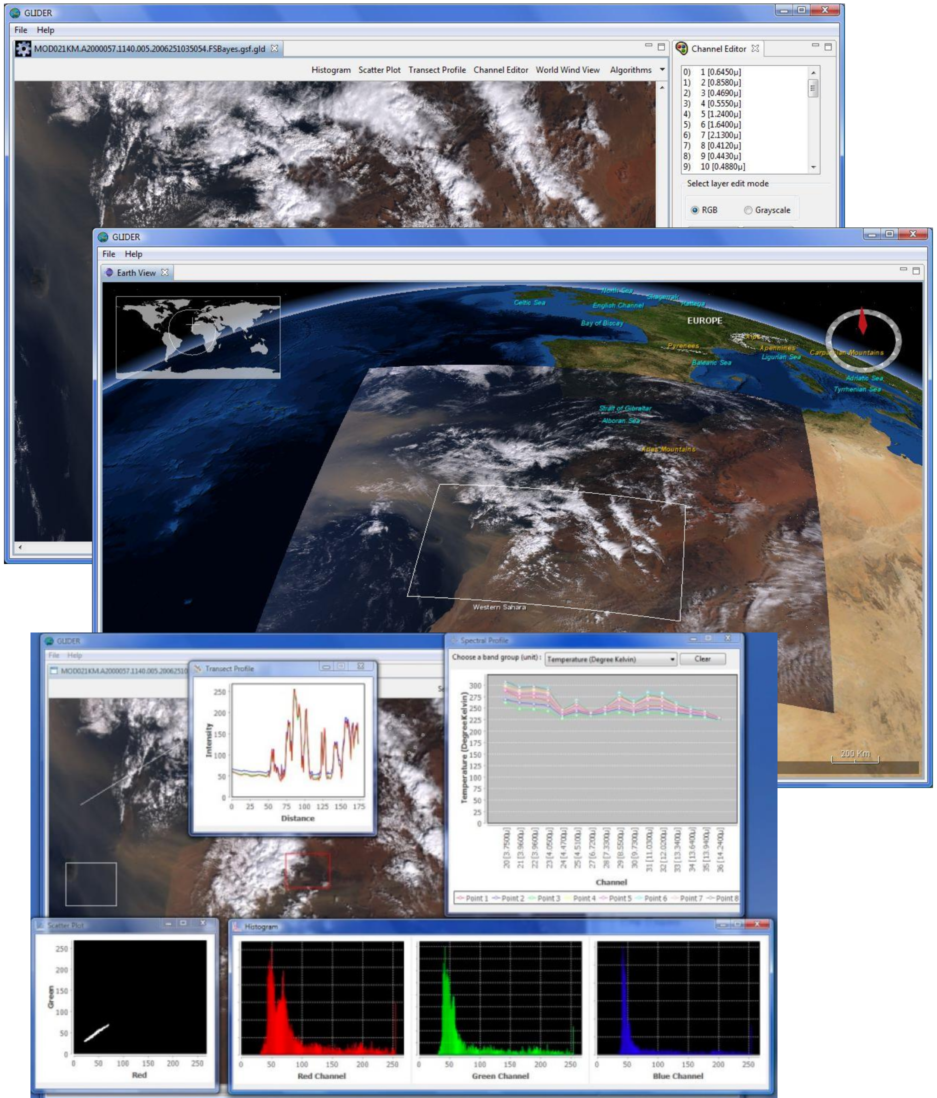
•Sensor Data Integration/Fusion
•Signature Analysis
•Pattern Recognition
•Real-Time Data Analysis

- **ADaM Algorithm Development and Mining toolkit**
- **Data Analysis for Studying Environmental Impacts**
  - **Thermal analysis of human activity**
  - **Measuring nuclear, chemical and oil facility usage and production**
  - **Evaluating environmental impacts on national security**
- **Multi-source Data Analysis**
  - **Algorithm Development**
  - **Multi-source integration and fusion**
  - **System signatures**
- **NASA/USAID sponsored SERVIR Environmental Data Products for Central America, Kenya and Nepal**
  - **Decision Support System for environmental analysis**
- **ARCTIC Climate Change Impacts**
  - **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

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

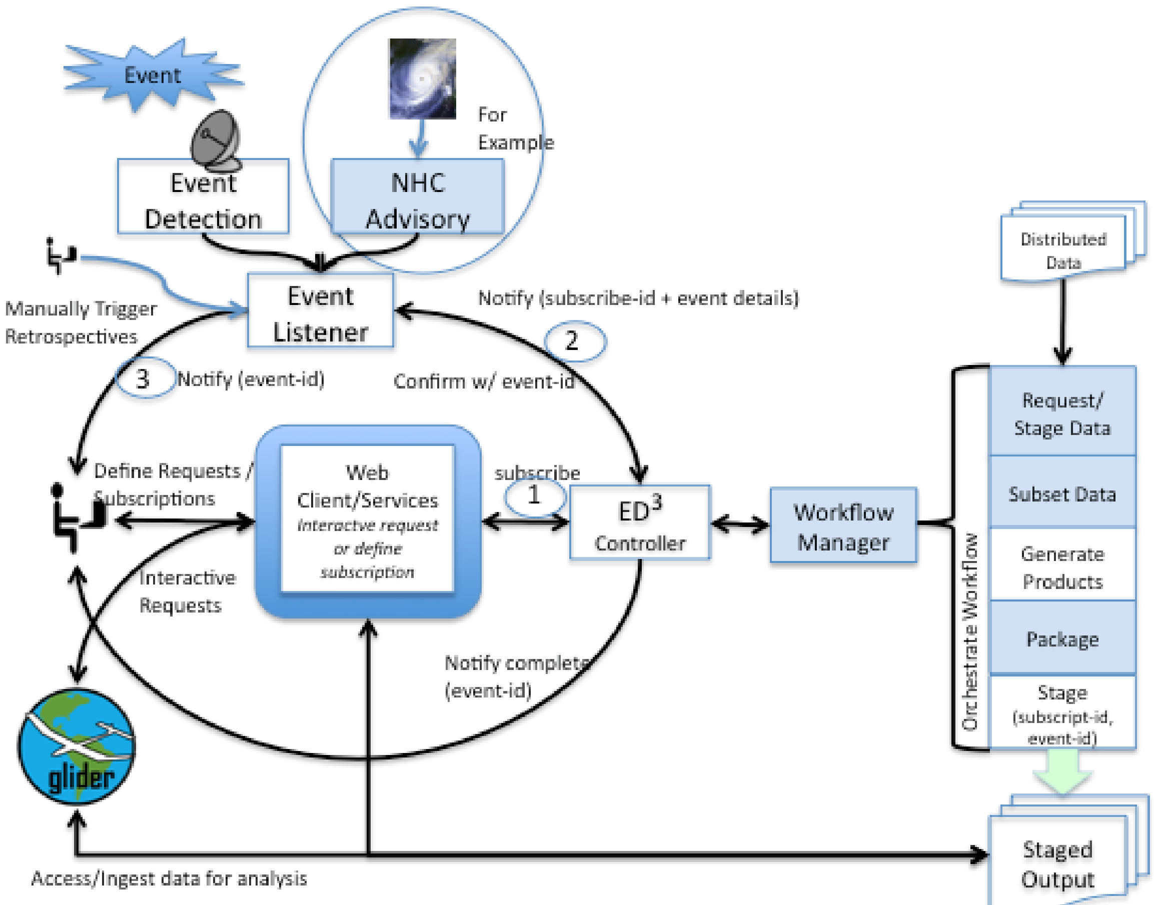
Integrates existing tools:

- 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 and other geolocation systems

2010 winner NASA ESDSWG Software Reuse Award and also used by defense community

Event-Driven Data Delivery (ED^3)

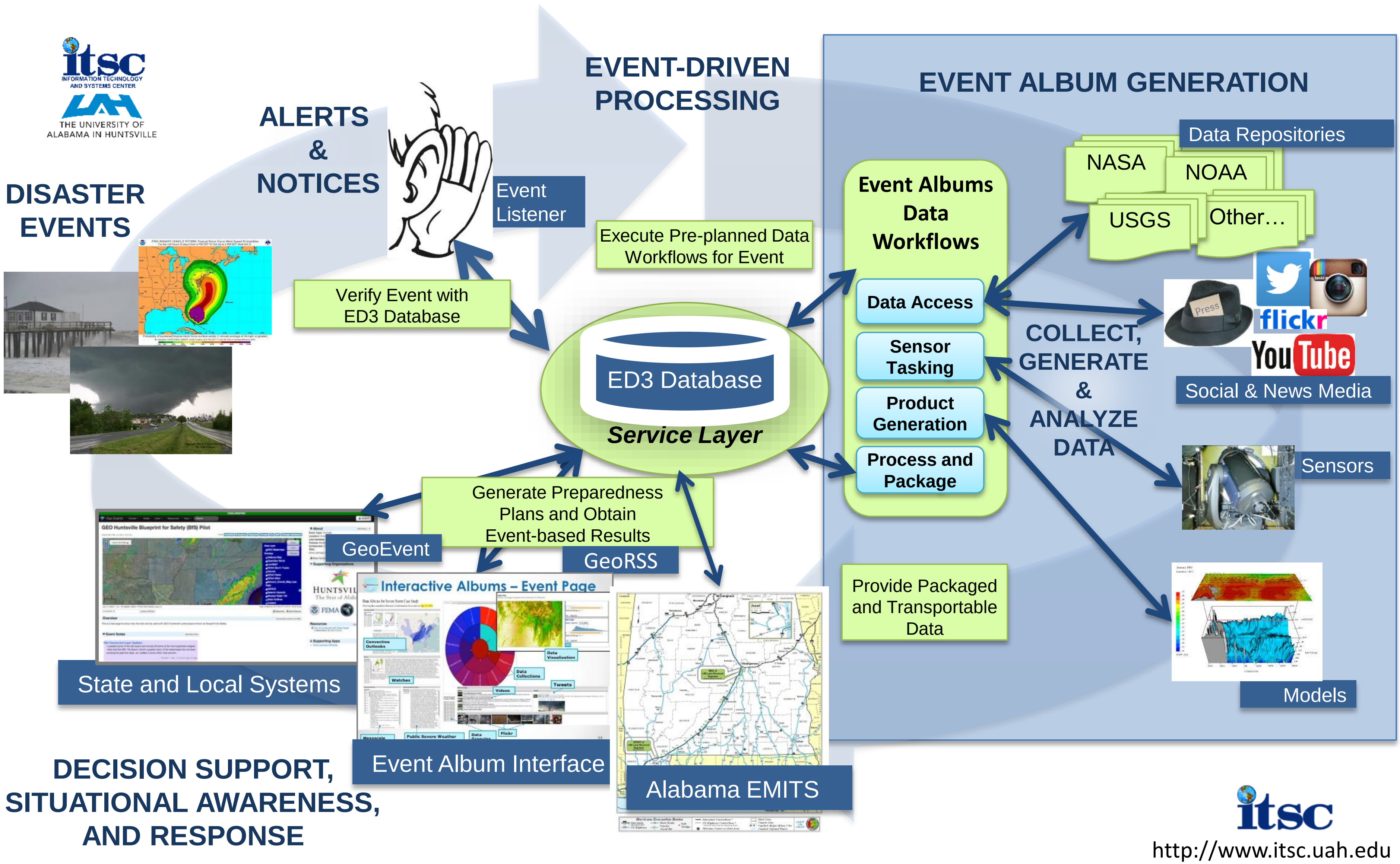
Data Management and Dissemination for Analysis and Visualization



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

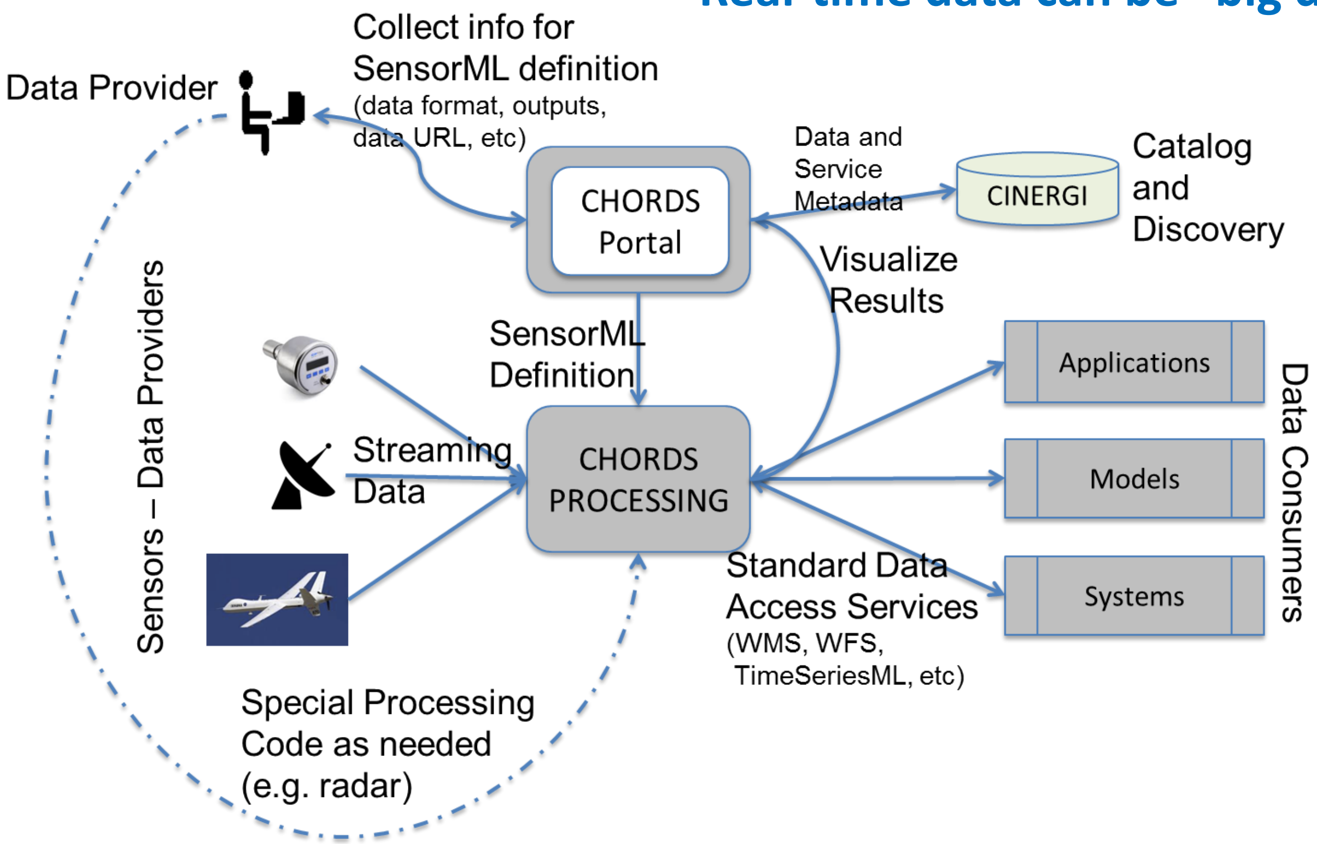
itsc

Event Driven Architecture



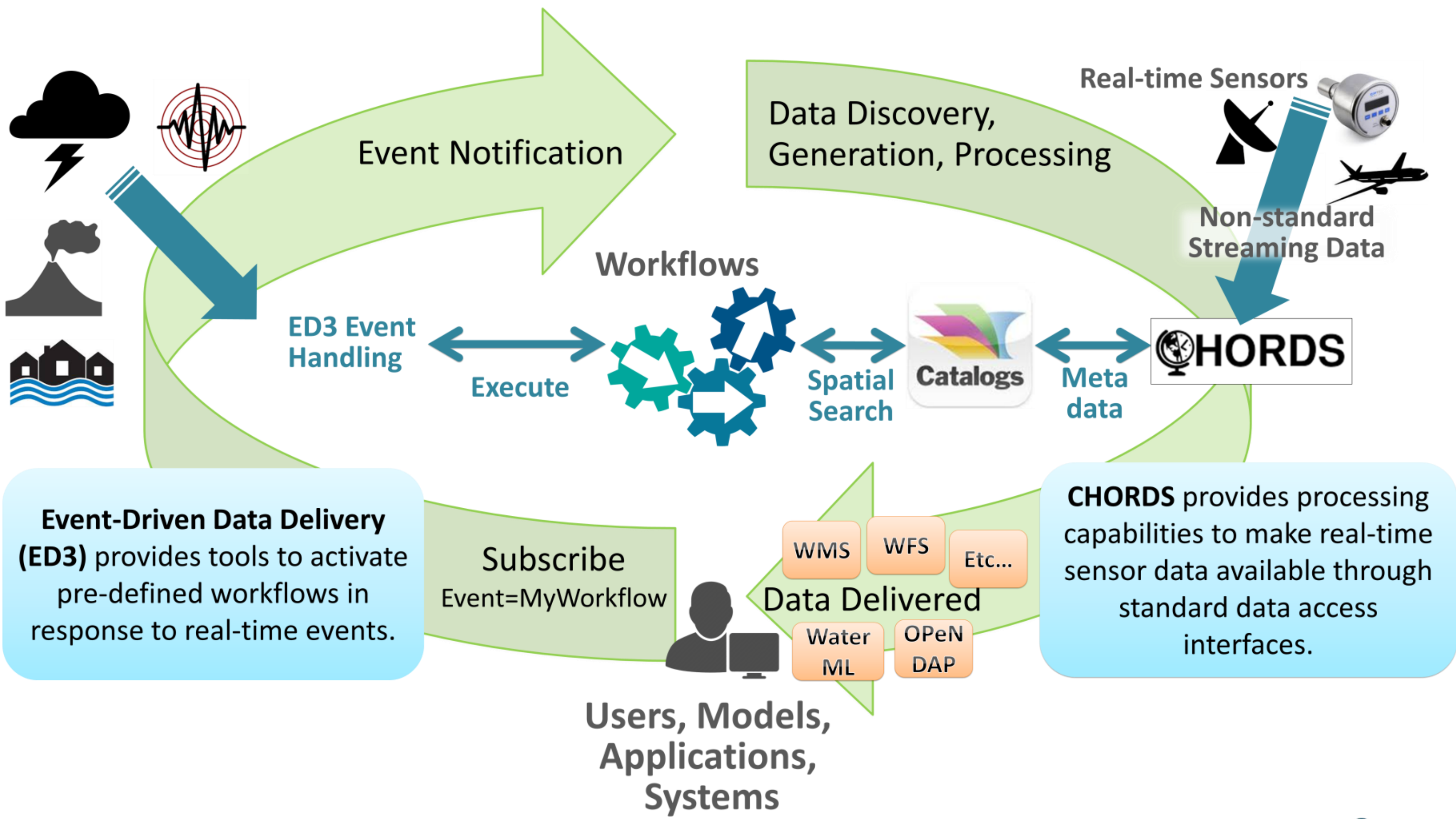
Cloud-Hosted Real-time Data Services (CHORDS)

Real-time data can be "big data" for users



CHORDS provides a system to ingest, navigate and distribute real-time data streams, and employ data and metadata formats that adhere to standards, which simplify the user experience.

Event-Driven Real-Time Data

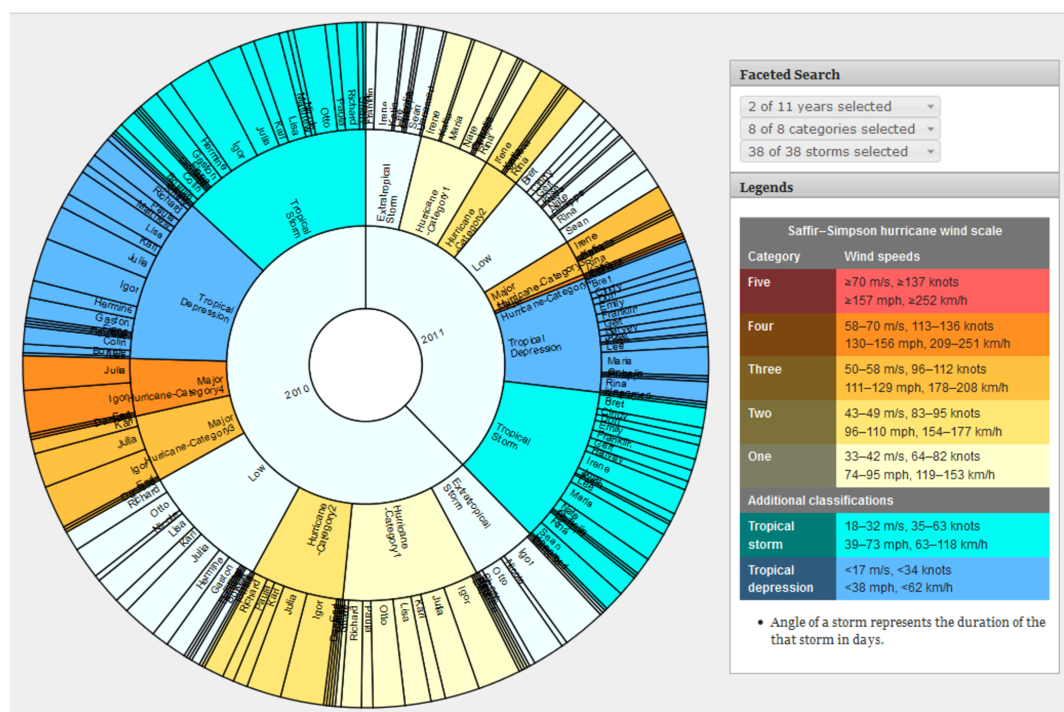


itsc
http://www.itsc.uah.edu

Event Albums

- Supports **search, aggregation and access** of data and online resources around events
- Automates the gathering of online resources with **information filtering**
- Displaying search results as:
  - Infographics – graphic visual representation of information, data or knowledge intended to present complex information quickly and clearly
  - Results enriched with additional information
- **Analytics dashboard** on the gathered information for events
- Use of **semantic technology** for relevancy ranking

- **Compiled collections** of information related to a specific topic or event with links to relevant data
  - Tools and services for visualization and analysis
  - News reports, social media, images or videos to supplement analysis
- **Curation** provides the author of a Event Album the means to select the aggregated information



Outline

- Looking Back to Look Forward
- Data Systems and Technologies
- Methodologies for Exploration and Stewardship
- **Rethinking Approaches for Exploration and Stewardship**

Data Science - An Emerging Field

National Science Board
Governing Board of the U.S. National Science Foundation
And Policy Advisors to the U.S. President and Congress

2005 National Science Board report, **“Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century”** introduced “Data Science” as an emerging field of science

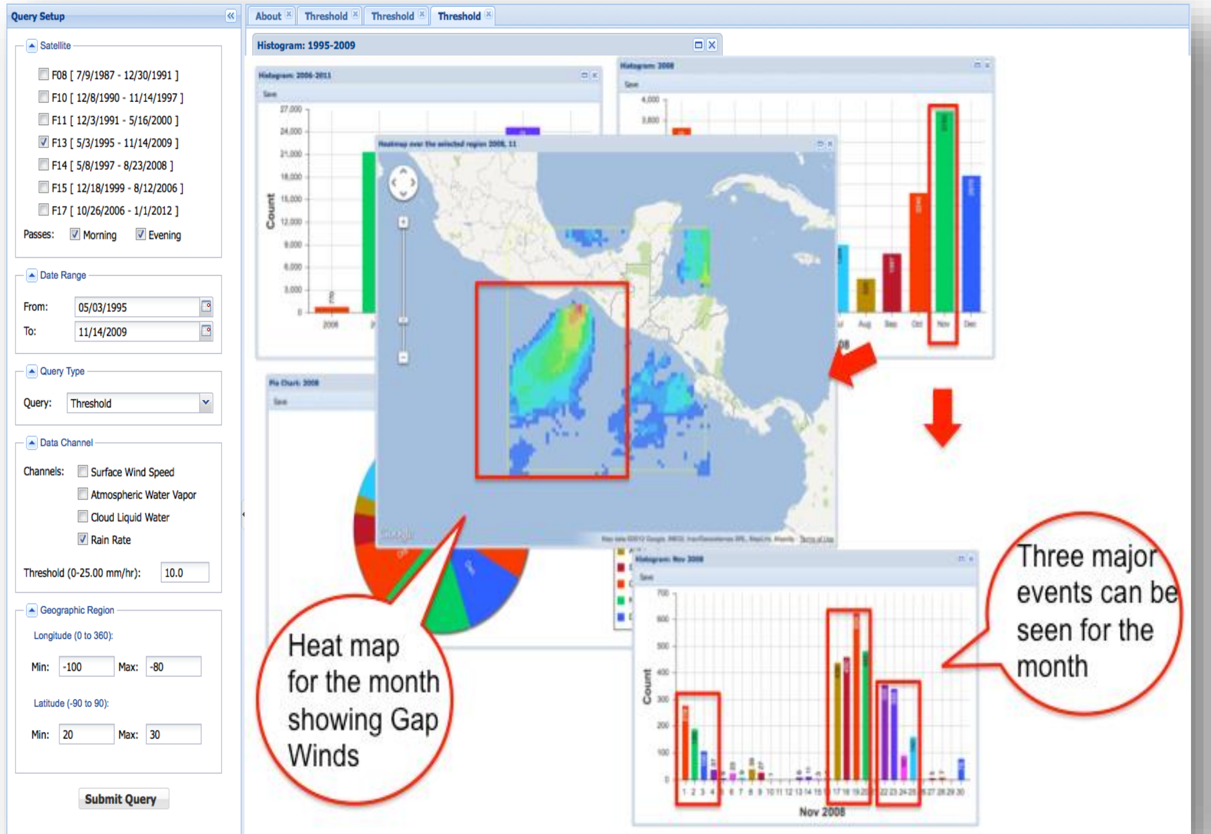
- Scale, scope and complexity of science and engineering data collections expanded by the ease data is gathered, processed, analyzed, and disseminated
- Many types of data; text, image, maps, etc.
- New methods needed to exploit value of data and unlock new discoveries
- Data can be used in disciplines not originally intended
- Workforce development is needed; university curricula introduced

*The **Data Scientist** brings together expertise from multiple disciplines, such as Data Mining, Engineering, Math and Statistics, Information Technology, Visual Analytics, Library Science, and Domain Science, to use data in new ways for developing new knowledge and understanding*

To transform data into knowledge and accelerate scientific innovation

Polaris: Big Data Exploration Engine

- Supports **data driven interactive exploration** of large amounts of data
- Utilizes high performance computing and new techniques for **efficient distributed file access**
- **Content based search** focuses the retrieval of the desired data stored in the file



Event Queries

- Detection
- Segmentation
- Characterization
- Correlation
- Statistics
- Tracking

http://www.itsc.uah.edu

Event Driven Exploration

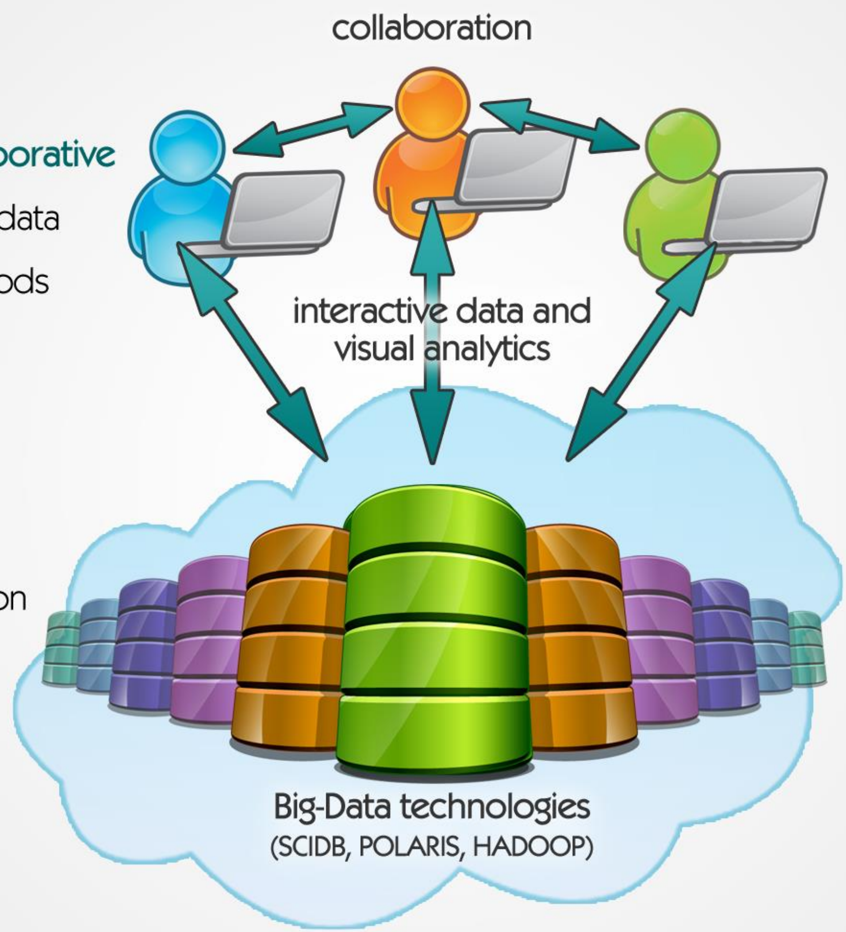
Automated Event Services

Utilize **Big-Data** technologies to…

- Enable **interactive** and **collaborative** scientific data analysis on big data
- Share data and analysis methods seamlessly,

…in order to…

- Relieve scientists from data management,
- Empower scientists to focus on science, and
- Boost science productivity.



1. Identify occurrences (events) of phenomena
   - Entities in the 4D spatiotemporal space
2. Associate additional relevant data with events.
3. Characterize phenomena with defining features extracted from data.
4. Correlate defining features of various phenomena in both space and time.
5. Improve predictions of future events using correlations among phenomena for **better decision making**.

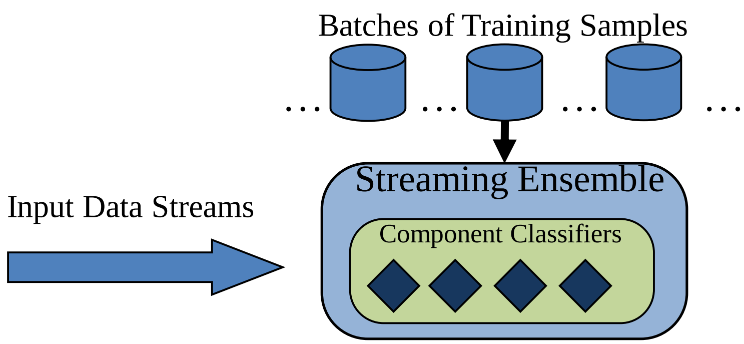
Big-Data Vision → **Technology Infrastructure** → **Science Enablement**

Signature Identification/Characterization Coverage Based Ensemble Algorithm (CBEA)

A 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

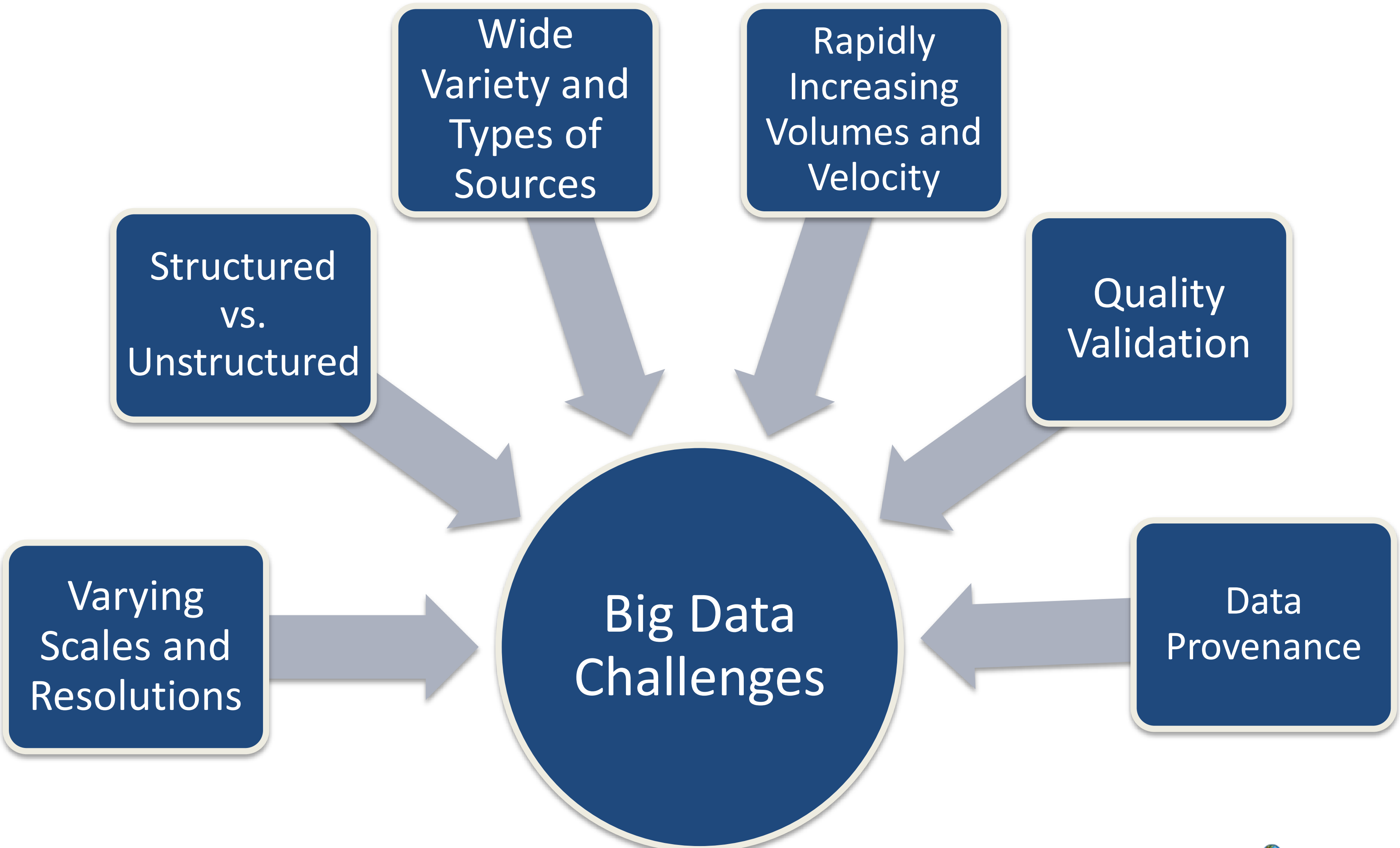
Characteristics of CBEA

- 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.

itsc
http://www.itsc.uah.edu

Big Data Challenges



itsc

http://www.itsc.uah.edu

Data Semantics

- Innovative methods for **representing** and **reasoning** about the meaning of data are key for data integration, exploration, and analysis.
- Unstructured data such as **text** (documents, social media, etc.) are particularly challenging.
- **Ontologies** are widely used to represent data semantics.
- **Ontology development is driven by the user's view** and should be based on early engagement with domain experts.
- New methods are needed for **ontology-based rapid search, analysis, and visualization** of unstructured data.

Text Mining using Ontologies

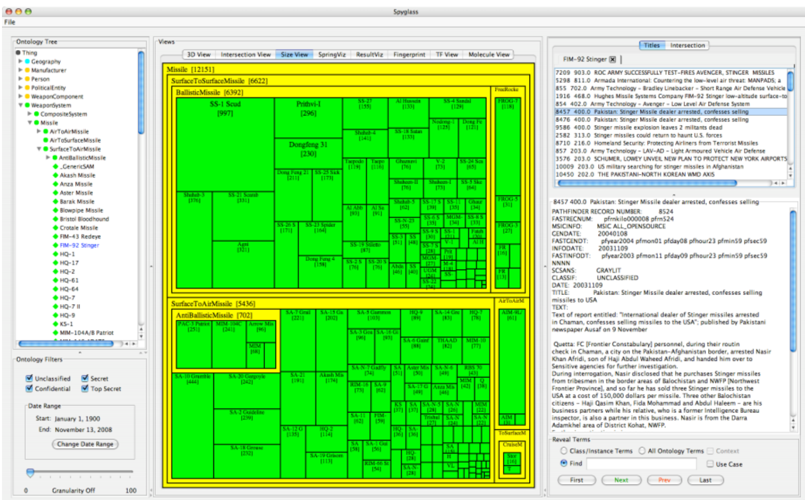
analyze and search large text sources (databases, on-line resources, etc.)

spyglass

Multiple Visual Analytics

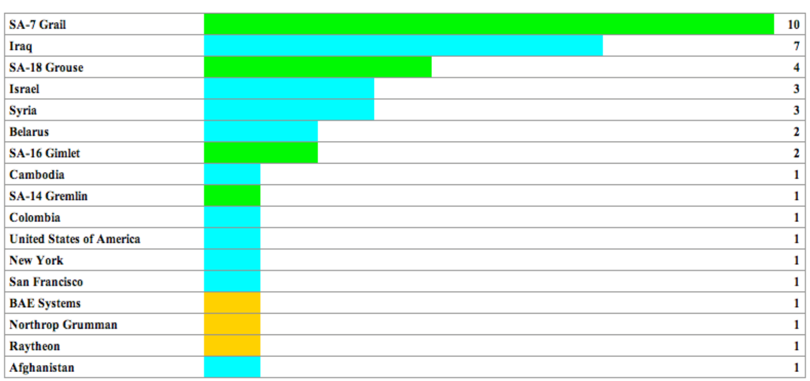
– Show distribution of documents across categories
– Show relationships between documents / categories

Ontology



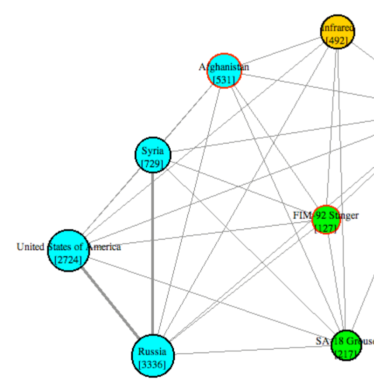
Document Browse View

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
- Support for **multiple languages**
- **Scores documents** by number of relevant terms
- **More powerful** than simple keyword queries
- **Possible to reason** over ontological structures



Molecule View of Concept Associations

itsc
http://www.itsc.uah.edu

Critical Data Processes

Collaborative workbench for cyberinfrastructure to accelerate science algorithm development

Manil Maskey
University of Alabama in Huntsville
Rahul Ramachandran
NASA/MSFC
Kwo-sen Kuo and Chris Lynnes
NASA/GSFC

NASA Grant #NNX13AB37G

mmaskey@itsc.uah.edu

Problem Statement

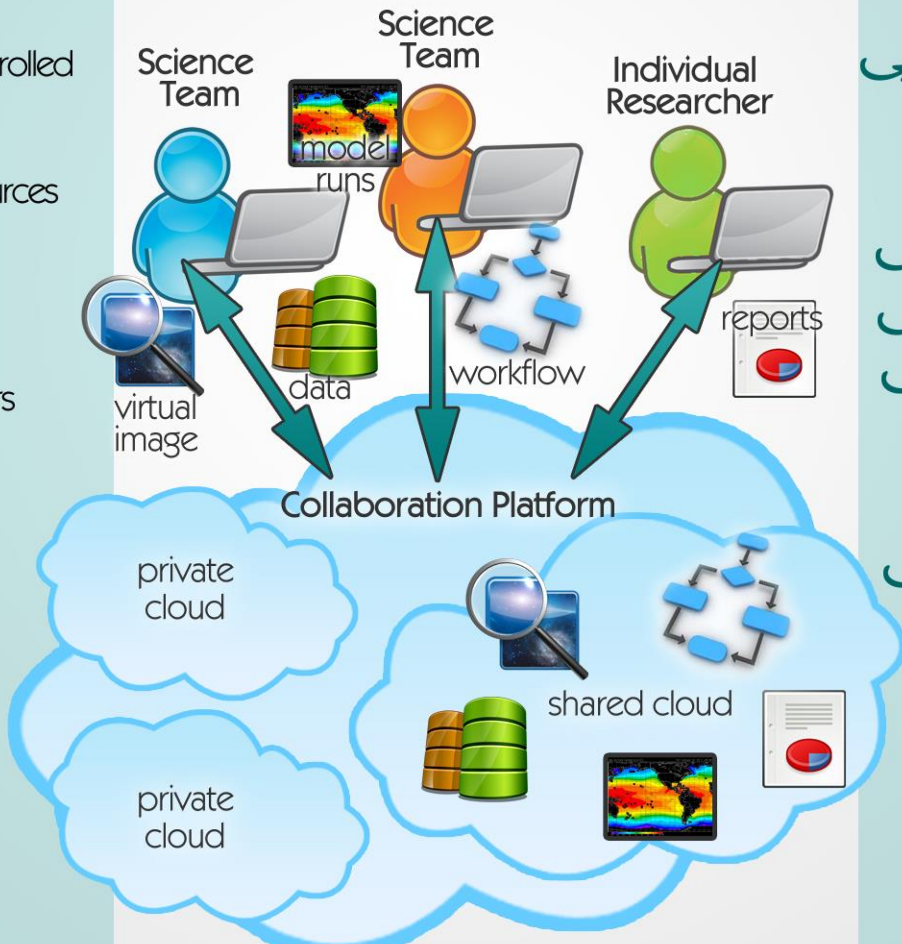
- There are significant untapped resources for information and knowledge creation in the science community.
  - Data
  - Algorithms and services
  - Analysis workflows or scripts
  - Related knowledge about these resources
- Resources often reside on an investigator's workstation or laboratory server and are rarely shared.
- One obstacle is lack of incentive, Antunes [Ref] gives 3Rs – Recognition, Reputation and Reward.
- Another obstacle is technological, infusing technologies into a researcher's existing analysis environment.
  - Few scientific tools support collaboration via sharing
  - Those that do often mean learning a new analysis framework and paradigm

Collaborative Workbench (CWB)
to Accelerate Science Algorithm Development

Sharing Knowledge is at the heart of science, yet it is challenging for researchers to effectively share information and tools

Goals

- An architecture for scalable, controlled collaboration
- Selective sharing of science resources
  - among individuals
  - within science teams
  - with the entire science community.
- Software that fits how researchers currently do scientific analysis to promote adoption



Benefits

- Accelerate science algorithm development by distributed science teams
- Reduce redundancy
- Improve productivity
- Securely share all science artifacts (data, information, workflow, virtual machines)
- Generalizable to support collaborative science algorithm development for other mission and model enterprises

Collaboration Terminology

Scientific Collaboration

Provisioning

Computation, storage, data and tools provided by federal agency or similarly trusted institution

Sharing

Exchange of data, programs, and knowledge among individuals (often strangers) and groups

Collaboration

Sharing with others (usually known colleagues) in pursuit of a common goal

Synchronous

Asynchronous

Collaboration can occur at different levels and at different speeds!!

CWB Benefits

- Accelerate science algorithm development by distributed science teams
- Reduce redundancy and improve productivity
- Securely share all science artifacts (data, information, workflow, virtual machines)
- Can be generalized to support collaborative science algorithm development for other mission and model enterprises