

Data Access, Integration and Stewardship Challenges for the Future

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

Sara J. Graves

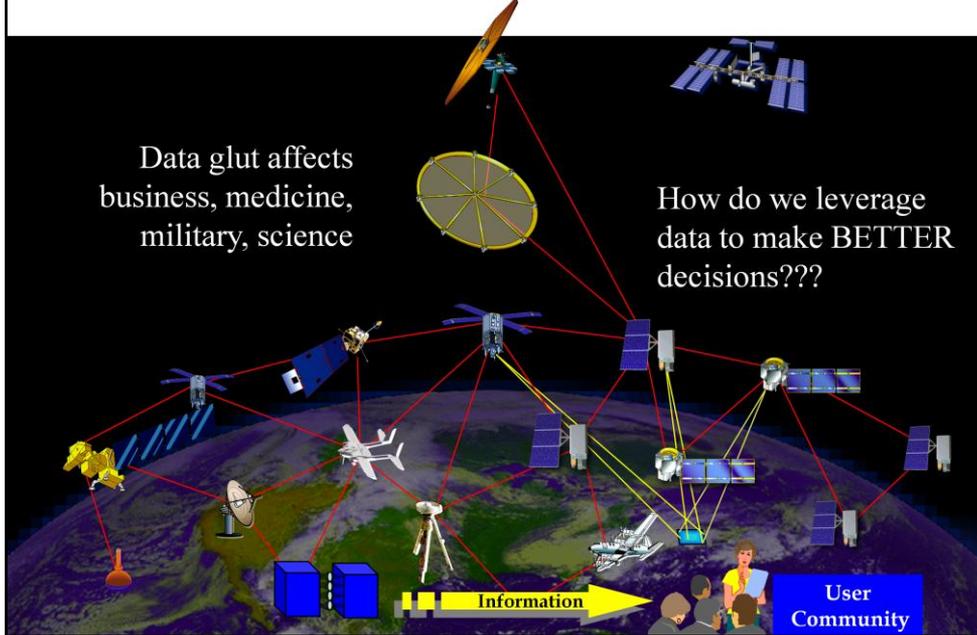
Director, Information Technology and Systems Center
Board of Trustees University Professor
Professor of Computer Science
University of Alabama in Huntsville
256-824-6064

sgraves@itsc.uah.edu

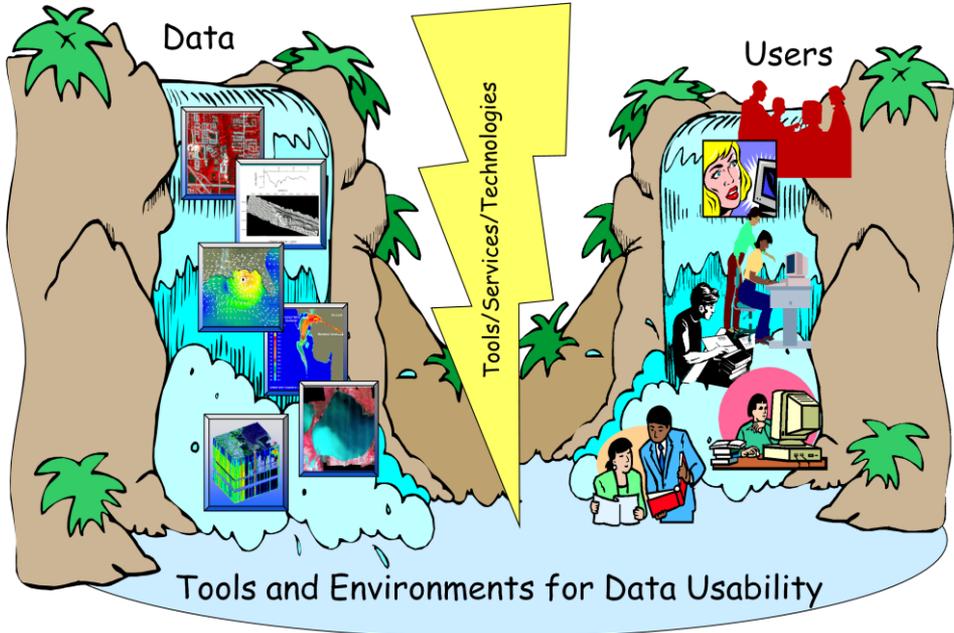


<http://www.itsc.uah.edu>

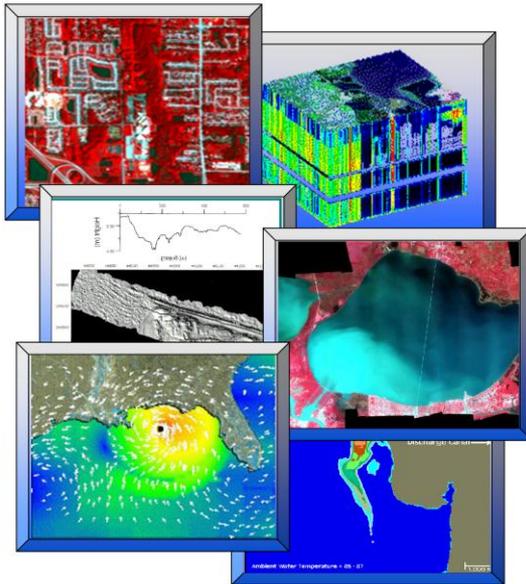
“...drowning in data but starving for knowledge”



Challenge: Increase usability of data and technologies to address the diverse needs of the flood of users.



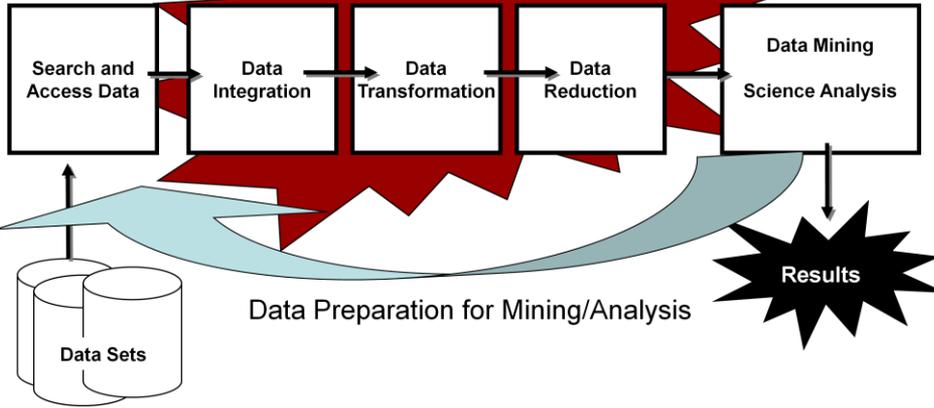
Heterogeneity Leads to Data Usability Problems



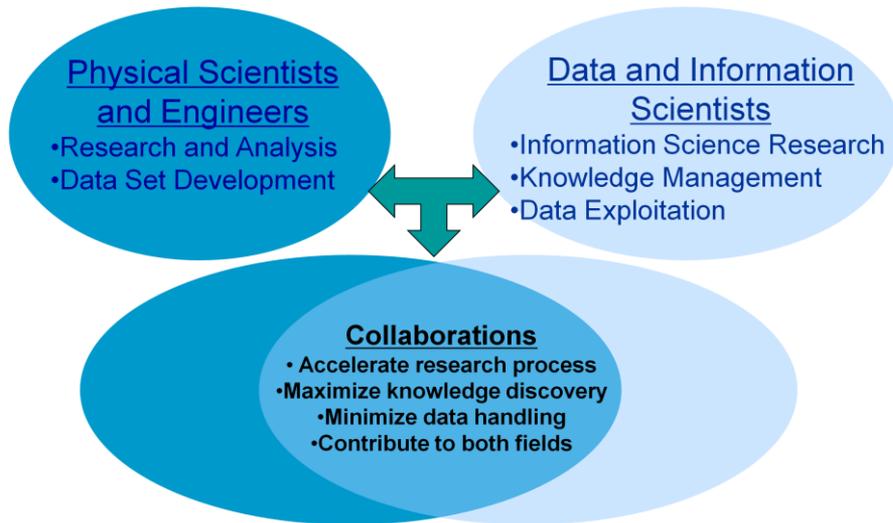
Data Characteristics

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

Data Challenge



Success Builds on the Integration of Science Domains and Disciplines



Applications have increasingly demanding requirements

- ✓ Data Mining and Knowledge Discovery
- ✓ Sensor Networks
- ✓ Modeling and Simulation
- ✓ Visual Analytics
- ✓ Advanced Processing Environments
- ✓ On-line Data Repositories

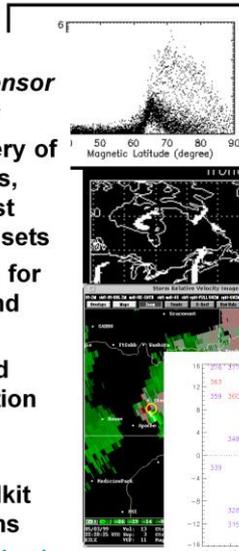


Data Mining

UAH has been at the forefront of mining sensor data for over 15 years

- Automated discovery of patterns, signatures, anomalies from vast observational data sets
- Derived knowledge for decision making and response
- Allows learning and training for adaptation
- ADaM – Algorithm Development and Mining System toolkit with 100+ algorithms

<http://datamining.itsc.uah.edu>



ADaM Documentation
Below are links to a tutorial and an overview of data mining, with examples using ADaM modules.

Data Mining Overview>>
Please refer to this overview document for information on what image processing and data mining components are available in the 4.0.2 release of ADaM. Tutorial>>

ADaM 4.0.2 Overview>>
Below is a link to a document containing API details about the individual ADaM 4.0.2 components. This is a compiled listing of header documentation that is also available when running the component executables interactively. Overview>>

ADaM 4.0.2 Components

Pattern Recognition	Image Processing
Classification Techniques <ul style="list-style-type: none">• Bayes Classifier• Naive Bayes Classifier• Bayes Network Classifier• CBEA Classifier• Decision Tree Classifier• SEA Classifier• Very Fast Decision Tree Classifier• Back Propagation Neural Network• k-Nearest Neighbor Classifier• Multiple Prototype Minimum Distance Classifier• Recursively Splitting Neural Network	Basic Image Operations <ul style="list-style-type: none">• Arithmetic Operations(+*?)• Collating• Cropping• Image Difference• Image Normalization• Image Moments• Equalization• Inverse• Quantization• Relative Level Quantization• Resampling• Rotation• Scaling• Statistics• Thresholding• Vector Plot
Clustering Techniques <ul style="list-style-type: none">• DBSCAN• Hierarchical Clustering• Isodata• k-Means• k-Medoids• Maximin	Segmentation/Edge and Shape Detection <ul style="list-style-type: none">• Boundary Detection• Polygon Circumscription• Making Region• Marking Region
Feature Selection Techniques	Filtering <ul style="list-style-type: none">• Dilatation

The Algorithm Development and Mining (ADaM) System is an excellent example of a tool that was originally funded by a NASA Research Announcement that is currently being used worldwide by many communities, including Earth Science.

NSF
Cyberinfrastructure
Report

2003

Advanced Cyber Infrastructure

Testimony for the
**NSF Advisory Committee on Cyber
Infrastructure**

January 22, 2002

Sara J. Graves

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

**Revolutionizing Science and Engineering
Through Cyberinfrastructure:**

Report of the National Science Foundation
Blue-Ribbon Advisory Panel on
Cyberinfrastructure

January 2003

Daniel E. Atkins, Chair
University of Michigan

Kelvin K. Droegemeier
University of Oklahoma

Stuart I. Feldman
IBM

Hector Garcia-Molina
Stanford University

Michael L. Klein
University of Pennsylvania

David G. Messerschmitt
University of California at Berkeley

Paul Messina
California Institute of Technology

Jeremiah P. Ostriker
Princeton University

Margaret H. Wright
New York University

New Computing Environments for Research and Education

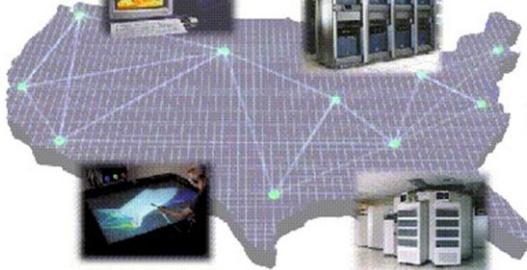
Computers



- Supercomputers
- Experimental Facilities



Distributed/heterogenous researchers, data, and computational resources



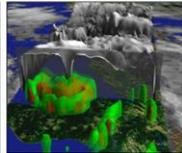
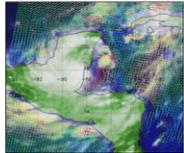
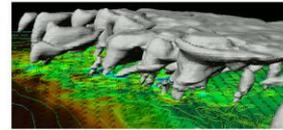
High-Speed Networks



Collaborative Environments



- Databases
- Mass Storage

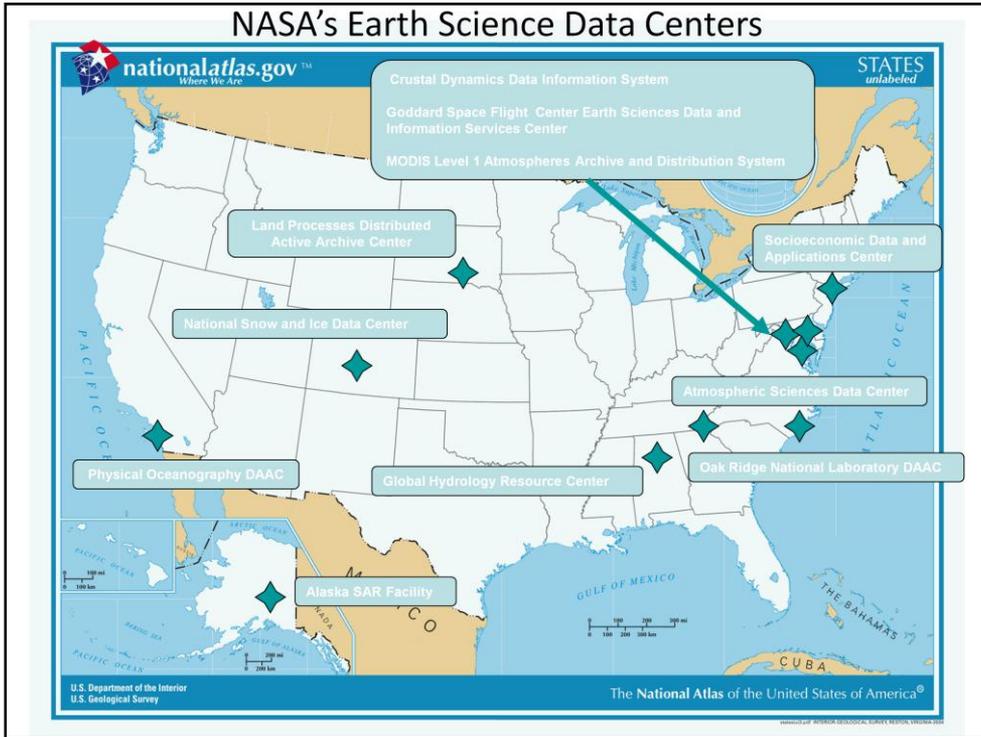


Observing systems require dynamic and powerful computing environments

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

NASA's Earth Science Data Centers



SERVIR

An Environmental Monitoring and Decision Support System for Central America

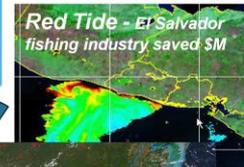
Earth Observatories



Central American Commission for Environment and Development

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

Environmental Monitoring & Decision Support Products



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



Rapid Response
ftp, e-mail, etc.

SERVIR Node in Panama

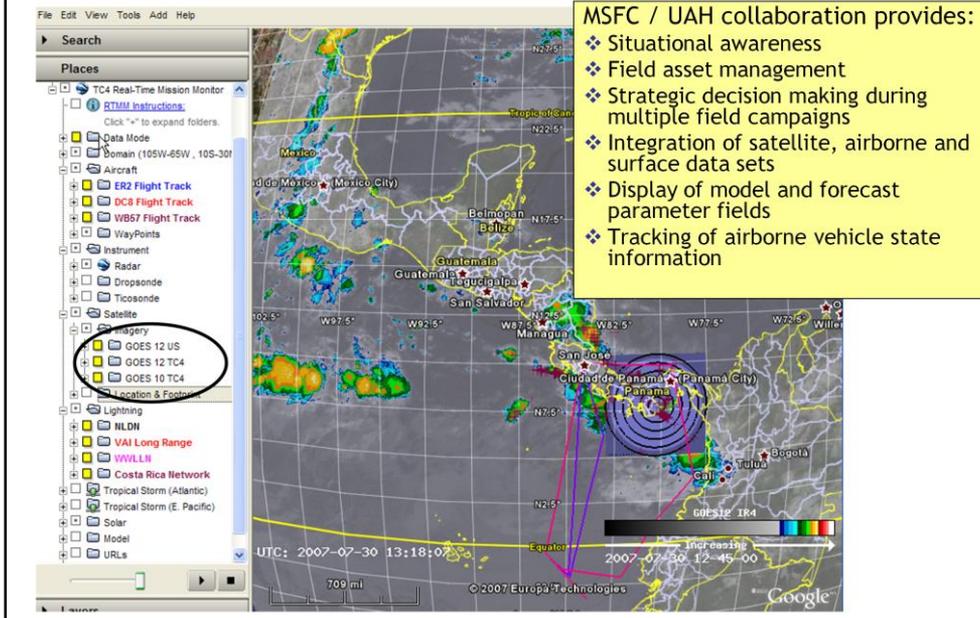
University of Arkansas
(World Bank Funding)

- 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

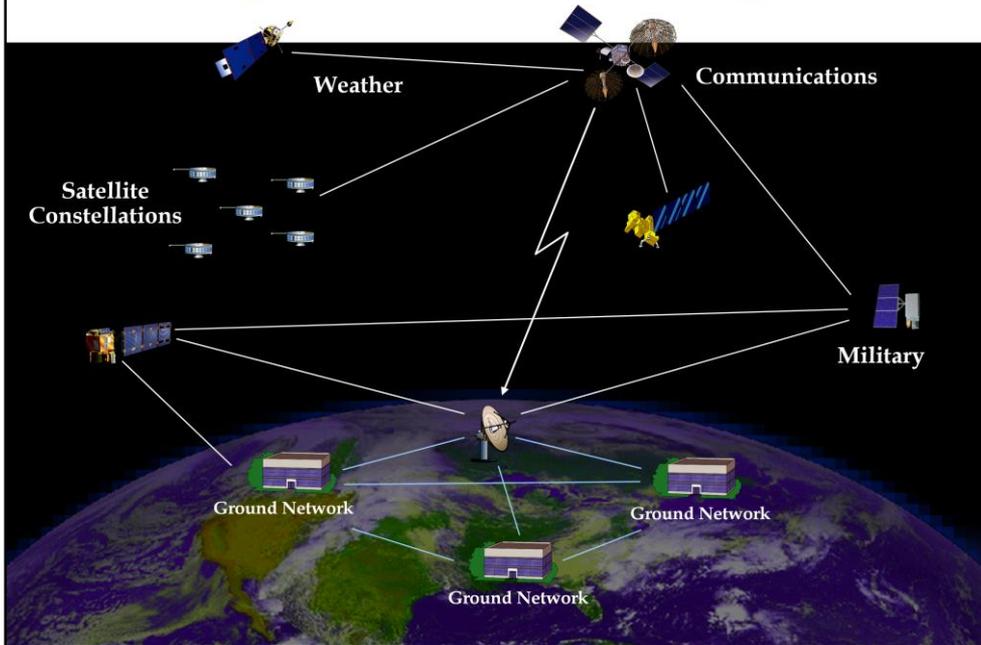
Data & Algorithms
SERVIR Partners

Real-Time Mission Monitor: Interactive visualization of remote sensing data

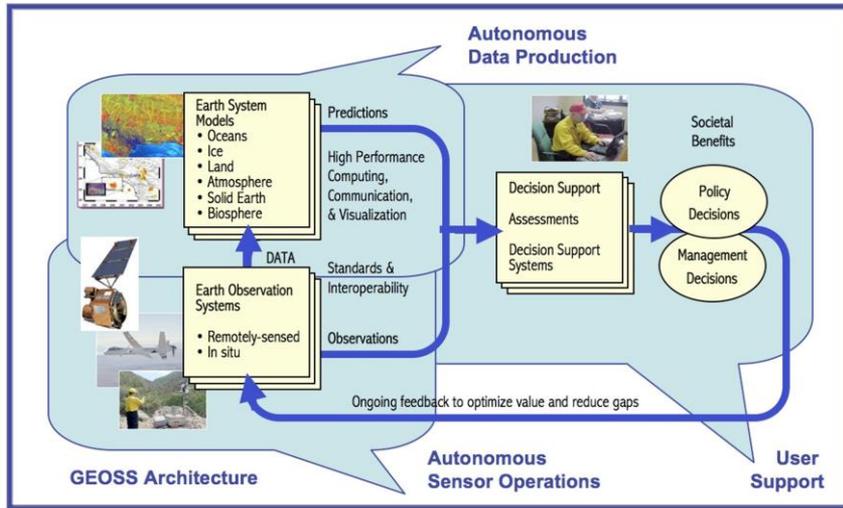


- MSFC / UAH collaboration provides:
- ❖ Situational awareness
 - ❖ Field asset management
 - ❖ Strategic decision making during multiple field campaigns
 - ❖ Integration of satellite, airborne and surface data sets
 - ❖ Display of model and forecast parameter fields
 - ❖ Tracking of airborne vehicle state information

A Reconfigurable Web of Interacting Sensors



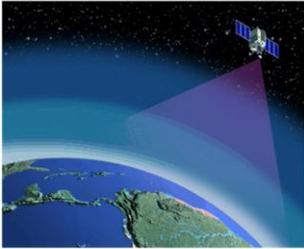
Sensor Web Themes in the GOESS Architecture



2008 Report of the NASA ESTO AIST Sensor Web Technology Meeting, April 2-3, 2008

On-Board Real -Time Processing Sensor Control/Targeting

EVE – Environment for On-board Processing (NASA)



EVE Capabilities

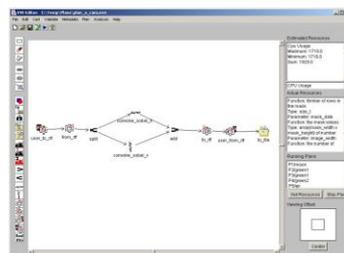
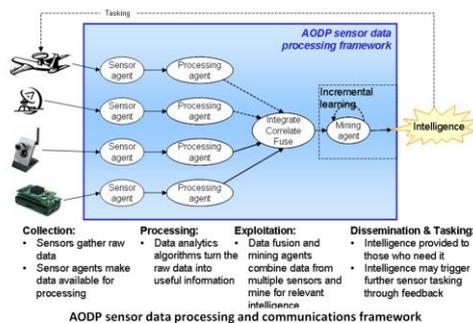
- Anomaly detection
- Autonomous Decision Making
- Immediate response
- Direct satellite to Earth delivery of results

Project Goals

- Prototype a processing framework for the on-board satellite environment.
- Provide specific capabilities within the framework
 - Data Mining
 - Classification
 - Feature Extraction
- Support research applications
 - Multi-sensor fusion
 - Intelligent sensor control
 - Real-time customized data products
- Create a laboratory testbed

Intelligent Sensor Networking

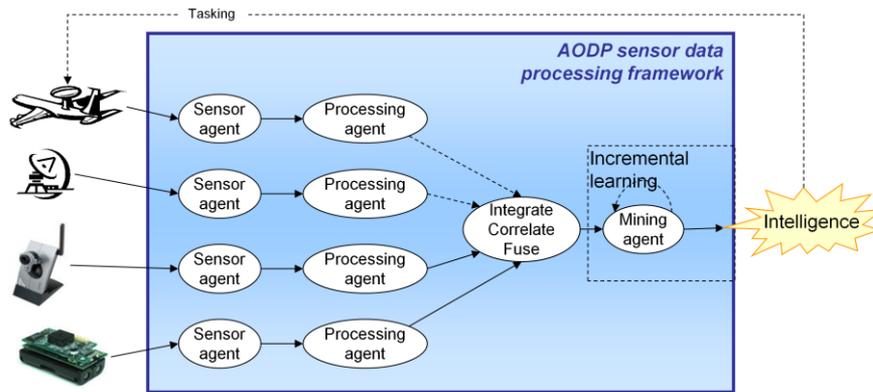
- **Who:** NASA and Defense Intelligence Agency – Measurements and Signal Intelligence (MASINT)
- **What:** EVE and The Adaptive On-demand Data Processing (AODP)
 - EVE and AODP take a unique approach to processing, integration and mining of sensor data in an on-board sensor network environment
 - The two primary research areas are
 - a *sensor data processing and communications framework*, and
 - *data analytics and mining algorithms* targeted to the needs and constraints of the on-board environment
 - The outcomes of this research will include:
 - a *framework* that can be used to correlate, integrate and fuse data for autonomous, intelligent sensor networks for applications in a wide range of environments
 - *data analytics and mining* components that can be utilized for processing of data in intelligent sensor networks
 - *transitioning of the technologies in follow-on related projects both inside and outside of the NASA and DIA community*
- **Why:** Together, the processing framework and advanced data analytics can enable a sensor network to provide as much intelligence as possible from sensor and related data, while reducing the time between data collection and dissemination of actionable intelligence.



Eve provides a workflow editor for on-board processing

An example of transitioning a NASA funded technology into a Department of Defense technology that enables better analysis of sensor data from some NASA satellites along with other data sources.

Sensor Data in the Intelligence Process



Collection:

- Sensors gather raw data
- Sensor agents make data available for processing

Processing:

- Data analytics algorithms turn the raw data into useful information

Exploitation:

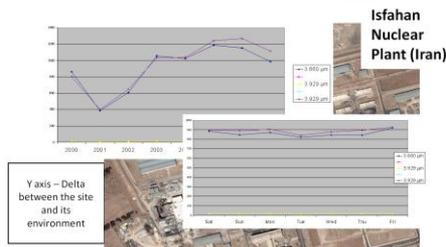
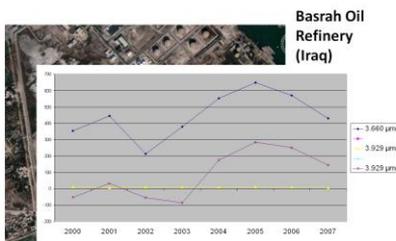
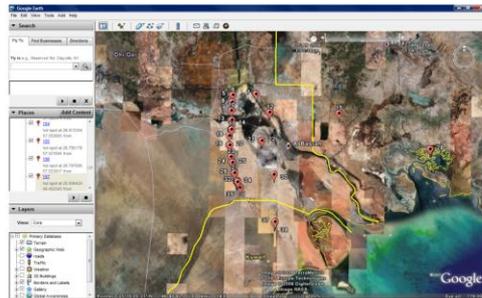
- Data fusion and mining agents combine data from multiple sensors and mine for relevant intelligence

Dissemination & Tasking:

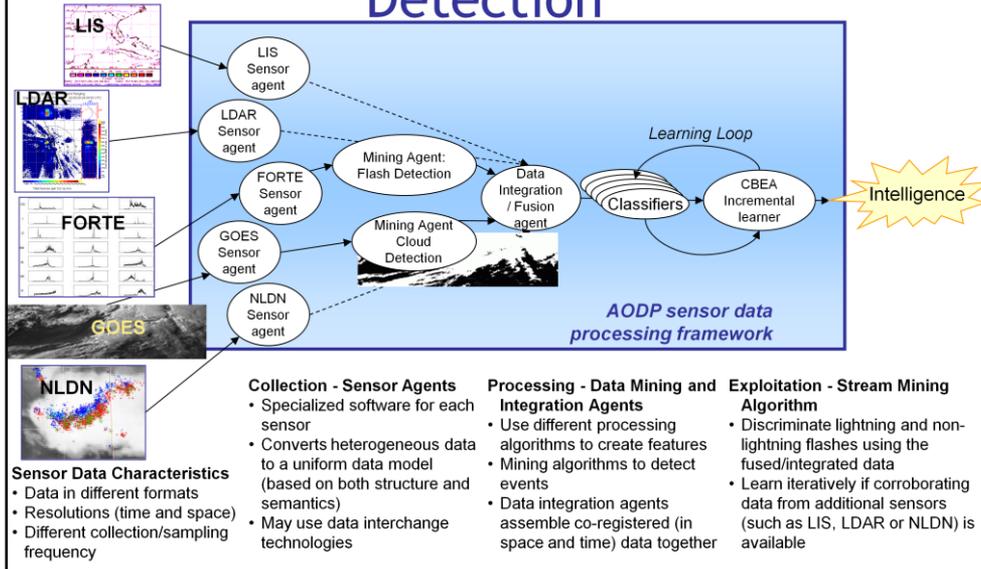
- Intelligence provided to those who need it
- Intelligence may trigger further sensor tasking through feedback

Global Monitoring using Multiple Data Sources for Automated Alerts of Changing Events and Trend Analysis for MASINT

- ❖ Compiled data for 83 sites in Iraq and Iran for all MODIS emissive bands, 2000-2007; will continue collecting data for these sites
- ❖ Types of facilities include oil wells, refineries, steel mills, aluminum processing, power plants, nuclear reactors, factories, and large kilns
- ❖ Also monitoring 28 sites in the U.S. Great Lakes area for comparison
- ❖ In progress: compiling data for 60 U.S. nuclear facilities with ground truth available



Electromagnetic Signature Classification and Anomaly Detection

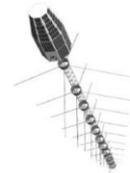


An example of the integration of NASA and other data sources for DOE signature classification and anomaly detection.

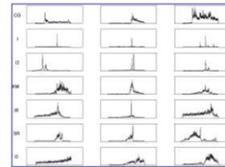
Anomaly Detection and Characterization of Radio Frequency Signals



- Detection of nuclear proliferation is difficult
- FORTE Radio frequency (RF) data provides an ideal information source for uncovering nuclear proliferation activity
- Proliferation signatures must be separated from known background RF signals (tv, radio, etc.) and natural events (e.g. lightning)
- Anomalous (unknown) signals must be detected and analyzed
- Apply data mining and visualization tools developed at UAHuntsville
 - **Algorithm Development and Mining Toolkit (ADaM)**
 - Supervised and unsupervised classification methods
 - Feature selection and optimization
 - **Globally Leveraged Integrated Data Explorer for Research (GLIDER)**
 - Data fusion of geospatial data
 - Validation of results



FORTE Satellite



Time Series RF Signals

Homeland Security and Law Enforcement



PREPARE (DOJ Byrne grant)

- Incident management
- SafeSchool
- ACJIC information system evaluation



AL Fusion Center (DHS)

- Prepared and vetted Privacy Policy
- Prepared and vetted Concept of Operations
- Intelligence analyst



SafeSchool (DOJ Byrne grant)

- Document management
- Interactive web map of schools, with analysis capabilities
- Integration of plume events for school identification
- Web access to detailed school information and emergency plans



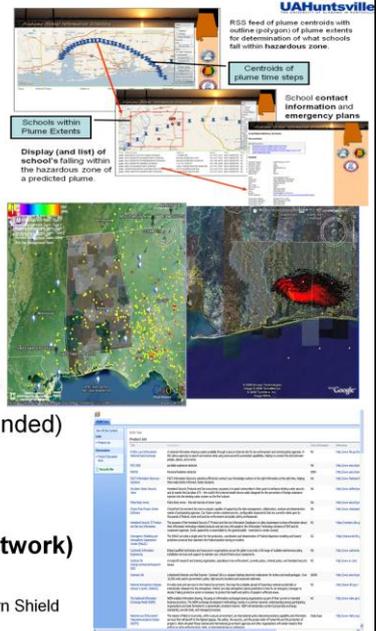
Plume Modeling (AL DHS collab - NASA funded)

- Web interface for easy plume modeling for decision support
- improved plume model
- integration with SafeSchool
- integration with Virtual Alabama



HSIN (Homeland Security Information Network)

- DHS Fusion Center Product trade study
- Product review application to allow user input and participation
- Integration and testing with TN fusion center and maybe Southern Shield



Examples of NASA data and tools that are used in some Homeland Security and Law Enforcement applications.

Integrated Use of NASA Data to Aid in Disaster Response and Decision Support (DHS)

Use current atmospheric model predictions to generate more accurate plume models for localized events.

Automated feed of plume path with outline (polygon) of plume extents for determination of what schools fall within hazardous zone.

Time-sequenced plume path

School contact information and emergency plans

Display (and list) of schools falling within the hazardous zone of a predicted plume.

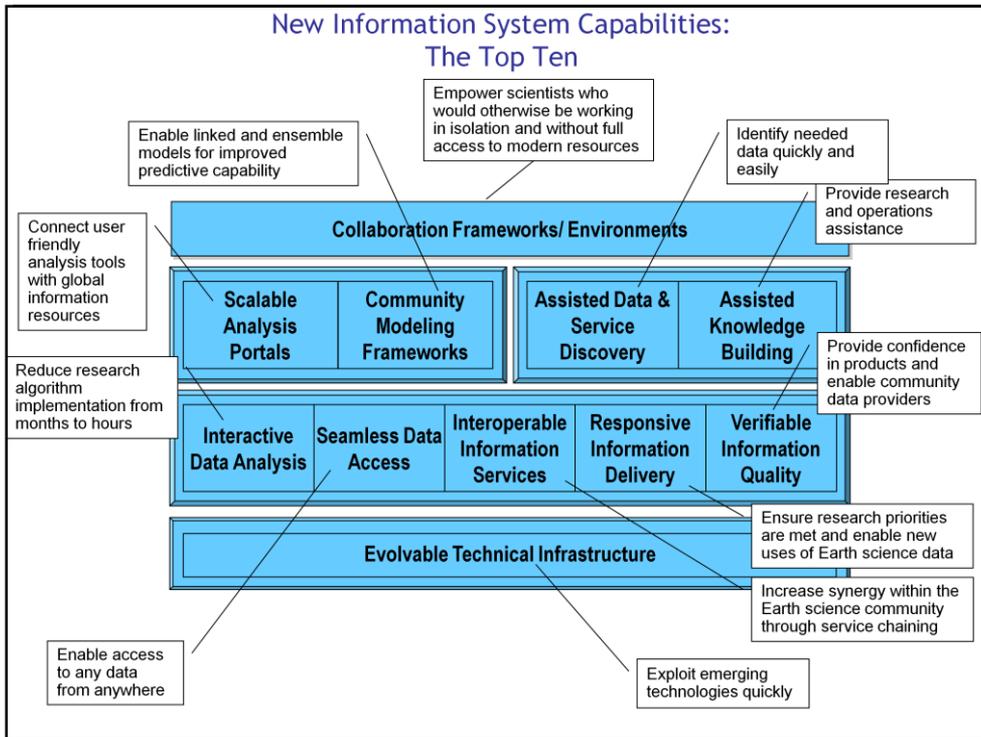
Schools within Plume-affected area

Type	School Name	Address	Phone	City
Elementary	ALABAMA COUNTY BRIDGE SCHOOL	17000 BR	251-897-2101	WAY NINETEEN AL
Elementary	WAY NINETEEN ELEMENTARY SCHOOL	600 BLACKBURN AVE	251-897-7943	WAY NINETEEN AL
Elementary	WAY NINETEEN INTERMEDIATE SCHOOL	600 BLACKBURN AVENUE	251-898-9893	WAY NINETEEN AL
Elementary	WAY NINETEEN MIDDLE SCHOOL	600 BLACKBURN AVENUE	251-898-9893	WAY NINETEEN AL
Elementary	WAY NINETEEN ELEMENTARY SCHOOL	600 BLACKBURN AVENUE	251-898-9893	WAY NINETEEN AL
Elementary	WAY NINETEEN CENTRAL HIGH TECHNOLOGY	600 W. WILSON LANE RD	251-897-9779	WAY NINETEEN AL
Elementary	WAY NINETEEN ELEMENTARY SCHOOL	6000 POND CROSSLAND	251-897-6945	WAY NINETEEN AL

Example of a collaboration between the Information Technology and Systems Center (ITSC) and the Atmospheric Science Dept of UAH with the Alabama Criminal Justice Information Center and the Alabama Fusion Center of DHS in merging the NASA-funded RAMSEASE and Department of Education SafeSchool efforts to provide the capability of determining schools that might fall in the predicted path of a hazardous air-borne release.

Trends for Transformative Science

- ❖ Exponential growth in volume of data collected
- ❖ Rapid expansion in the capability of computational hardware (multicore architectures)
- ❖ Advanced networking technologies
- ❖ Virtualization
- ❖ Software for knowledge representation, discovery, and manipulation
- ❖ Contextually aware applications
- ❖ Tools for electronic publication, digital libraries, and virtual collaboration
- ❖ Data assimilation - ability to blend models and data
- ❖ Adaptive computing
- ❖ Visualization



Slides 26-28 are from the NASA ESDSWG Technology Infusion Working Group

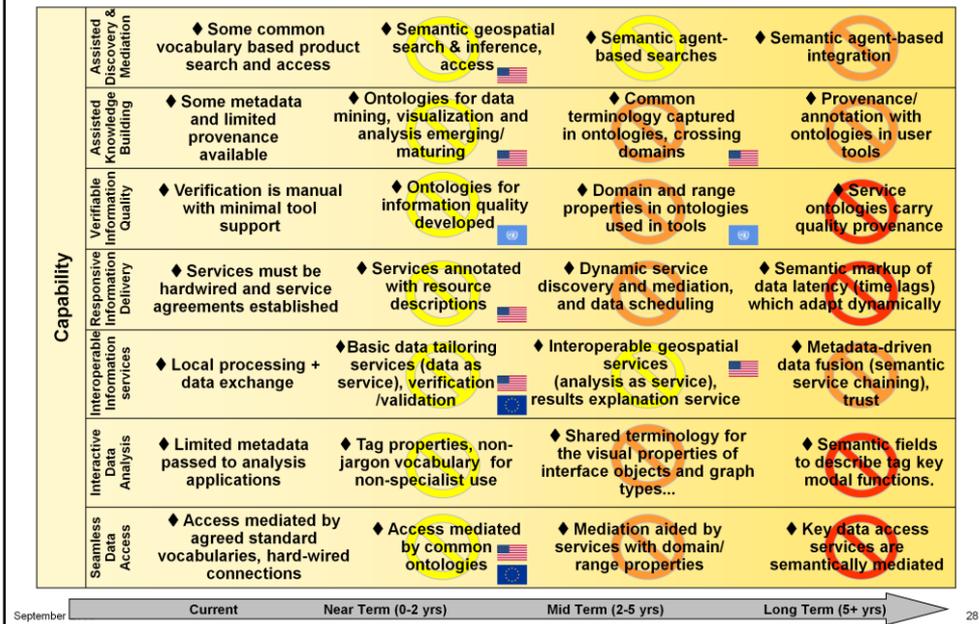
Web Services Roadmap 2008 Update



Results	Outcome	◆ Improved Information Sharing	◆ Accelerated Research & System Cost Savings	◆ Increased Collaboration & Interdisciplinary Science	◆ Increased PI Participation in Information Production	◆ Automated Data Utilization
	Output	◆ Geospatial services established	◆ Open geospatial services proliferate	◆ Widespread production quality geospatial services	◆ Intelligent Services	
Capability	Assisted Discovery & Mediation	◆ Parameter-based product searches and access	◆ Full geospatial logical searches and access	◆ Semantic geospatial search & access	◆ Automatic service mediation	
	Interoperable Information Infrastructure	◆ Local processing + data exchange	◆ Basic data tailoring services (data as service)	◆ Interoperable geospatial services (analysis as service)	◆ Metadata-driven data fusion (semantic service chaining)	
Technology	Data	◆ Open data access established (OpenDAP, OGC)	◆ Common geospatial schema adopted (GML)	◆ Geospatial service catalog established (WSDL, UDDI, ???)	◆ Open geospatial ontology converges (OWL)	
	Messaging	◆ Open service protocols established (HTTP, REST)	◆ Common service protocol, description adopted (SOAP, WSDL)	◆ Standard-workflow languages infused (e.g. BPEL)	◆ Unified security & identity management (WS-Security, SAML)	



Semantic Web Roadmap (expanded capability)



This is an expanded version of the previous slide where the middle capability layers include all seven of the components that semantic web influences, these are the

Same components from slide 5 and slides 6-12.

The roadmap is to be read left to right within a row (building on the previous element).

Capability levels can be assessed by accessibility as well i.e. - available, easily available, widespread

Arrows between boxes to show what it takes to get from one to another...

Education/ outreach, briefing, etc. to NASA and within their available fora

Semantic Web & Ontologies

- ❖ **Noesis** - meta search engine and a resource aggregator

Uses ontologies to guide users to refine their search query producing better search results, reducing the user's burden to experiment with different search strings

Serves as a tool to allow users to browse and traverse the different concepts in the ontology



With the broadening communities that are using NASA data, there is increasing need for semantic mediation technologies such as Noesis.

Noesis: Ontology Based Search and Resource Aggregation Tool

The screenshot displays the Noesis web application interface. At the top, there is a navigation bar with links for Home, About, FAQs, Contacts, ITSC, and Disclaimer. The main search area shows a search bar with the query 'Pressure' and a 'Search' button. Below the search bar, the number of results is shown as 177. The search results are displayed in a table format with columns for Refine Search, Search Results, and Filter by Engine. The search results include a definition of pressure, a list of related terms, and a list of search results from various sources like Wikipedia, Google, and NASA. A blue banner at the bottom of the screenshot contains the text: 'Noesis, developed for LEAD, is also being used by NASA ESIP Federation, Gulf of Mexico Regional Collaboration Project and others'. To the right of the screenshot, there is a list of bullet points describing the capabilities of Noesis.

noesis Home | About | FAQs | Contacts | ITSC | Disclaimer

Search Stop

Number of Results: 177

Definition:
 1. A type of stress characterized by uniformity in all directions. As a measurable on a surface, the net force per unit area normal to that surface exerted by molecules rebounding from it. In dynamics, it is that part of the stress tensor that is independent of viscosity and depends only upon the molecular motion appropriate to the local temperature and density. It is the negative of the mean of the three normal stresses. The concept of pressure as employed in thermodynamics is based upon an equilibrium system, where tangential forces vanish and normal forces are equal. 2. In meteorology, commonly used for atmospheric pressure. 3. In mechanics, same as stress. 4. See radiation pressure.
 Source: <http://anaglossary.allenpress.com/glossary>

Refine Search: Pressure Related Qu... Pressure Static Pressure Hydrostatic Pre... Atmospheric Pre... Total Pressure Partial Pressur...

Search Results: Pressure - Wikipedia, the free encyclopedia
 Google
 This article is about pressure in the physical sciences. For the psychological meaning, see Peer pressure. For other uses, see Pressure (disambiguation) ...
<http://en.wikipedia.org/wiki/Pressure>
 Atmospheric pressure - Wikipedia, the free encyclopedia
 Google
 Atmospheric pressure is the pressure at any point in the Earth's atmosphere. In most circumstances atmospheric pressure is closely approximated by the ...
http://en.wikipedia.org/wiki/Atmospheric_pressure
 Online Conversion - Pressure Conversion
 Google
 Convert between many different pressure equivalents.
<http://www.onlineconversion.com/pressure.htm>
 Pressure
 Google
 Pressure is defined as force per unit area. It is usually more convenient to use pressure rather than force to describe the influences upon fluid behavior.
<http://hyperphysics.phy-astr.gsu.edu/hbase/press.html>

Filter by Engine: All Web Google - 10 Yahoo - 10 Data LEAD - 10 NCDC - 8 NASA OCMO - 50 Publications AMS - 20 Elsevier - 20 Springer - 20 RMS - 25 Education ELSE - 4

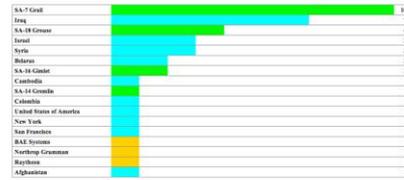
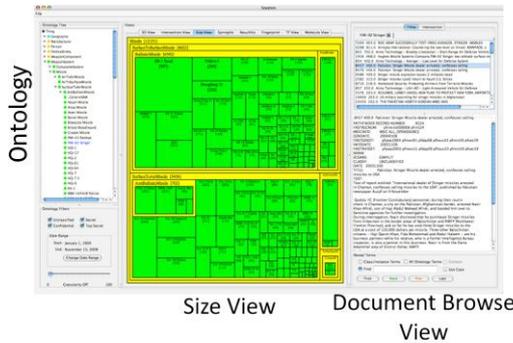
Noesis, developed for LEAD, is also being used by NASA ESIP Federation, Gulf of Mexico Regional Collaboration Project and others

<http://noesis.itsc.uah.edu/>

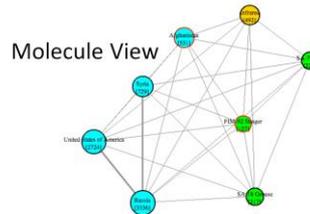
- Information searching can be enhanced considerably through the **integration of ontologies into search systems** – Noesis is one example
- Ontologies allow searches to be conducted in terms of concepts rather than words, to reduce the ambiguity problem
- The use of ontologies while searching has only 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.)



Spyglass: Ontology-based Indexing and Retrieval of Documents



Document Fingerprint View



Visual Analytics such as

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

Note that these views are representative – others (including 3D) are implemented for better visual analysis of large numbers of documents. Spyglass is a very powerful text mining system that can be valuable in searching many of the documents generated about NASA research, etc.

Standards Process Group: Current Standards and Technical Notes

ESDS-RFC	Title	Rev	Class	Status
ESDS-RFC-001	Charter of the ESDS Standards Process Group	2	Tech Note	Final
ESDS-RFC-002	The ESDS Standards Process	2	Tech Note	Final
ESDS-RFC-003	Instructions to Authors	2	Tech Note	Final
ESDS-RFC-004	The Data Access Protocol - DAP 2.0	1.1	Standard	Final
ESDS-RFC-005	OpenGIS Web Map Service Version 1.3	1	Tech Note	Final
ESDS-RFC-006	OpenGIS Web Map Service Version 1.1.1	1	Standard	Final
ESDS-RFC-007	HDF 5	1	Standard	Final
ESDS-RFC-008	HDF EOS 5	1	Standard	Final
ESDS-RFC-009	Aura Guidelines Technical Note	1	Tech Note	Final
ESDS-RFC-010	Backtrack Orbit Search Algorithm Technical Note	1	Tech Note	Final
ESDS-RFC-011	NetCDF Classic	1	Standard	Final
ESDS-RFC-012	GCMD Directory Interchange Format (DIF)	1	Standard	Final
ESDS-RFC-013	GCMD Service Entry Resource Format (SERF)	0.1	Stds Track	Draft
ESDS-RFC-014	Interoperability between OGC CSW and WCS Protocols	0.1	TN Track	Draft
ESDS-RFC-015	Provenance within Data Interoperability Standards	0.1	TN Track	Draft
ESDS-RFC-016	Lessons Learned Re: WCS Server Design / Implementation	0.1	TN Track	Draft
ESDS-RFC-017	Mapping HDF 5 to DAP 2	0.1	TN Track	Draft
ESDS-RFC-018	Creating File Format Guidelines: The Aura Experience	0.1	TN Track	Draft
ESDS-RFC-019	ICARTT File Format Standards	0.1	TN Track	Draft

From the NASA ESDSWG Standards Process Group
(<http://www.esdswg.org/spg>)

Looking Back to Look Forward

The following four charts have been used over the past years during discussions of the evolution of NASA Earth Science data systems. They are included in this presentation to illustrate that some of these concepts continue to be as pertinent today as they were in the past.



Evolution in ESE Data Management

*Less NASA control
in data management*

*More NASA control
in data management*

① **Pre-EOSDIS** (before 1994)

- Data held by science researchers or data centers; data difficult to locate
- Varying storage organization
- Long-term data preservation issues

④ **SIP Federation** (1998 to present)

- Experimental not operational
- Natural clustering
- More interest in individual research than federation evolution

③ **EOSDIS Core System** (1999 to present)

- Data held by DAACs
- Uniform data management environments
- Distributed data system

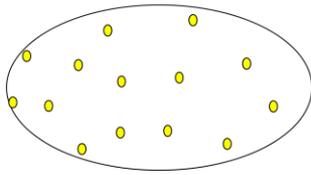
② **EOSDIS Version 0** (1994 to present)

- Hard to be "all things to all people"
- Data held by DAACs, affiliated with researchers
- Heterogeneous data management
- Locate data thru VO IMS, a federated data system
- Commitment to archive quality

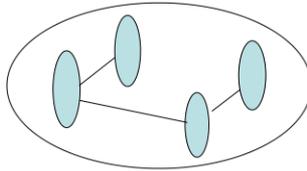
⑤ **NewDISS**

- **The intent of the NewDISS program is to focus ESE data system evolution to:**
 - increase NASA's flexibility to adapt the network of data systems & service providers;
 - enable access for NASA's Applications program and its educational programs;
 - improve cost effectiveness throughout the data system development and operational life cycle;
 - leverage the capabilities, expertise, and lessons learned from existing data systems; and
 - assure long-term data stewardship and continuity of services.

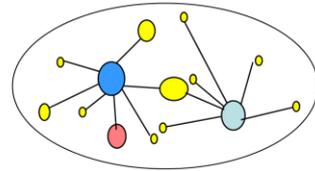
Data System Evolution



Individual researchers,
no data management
coordination

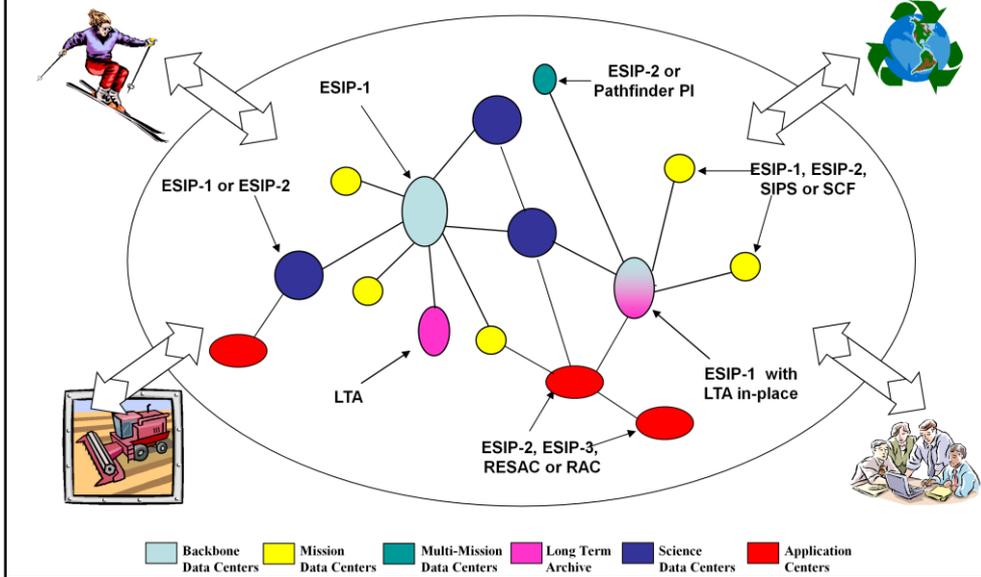


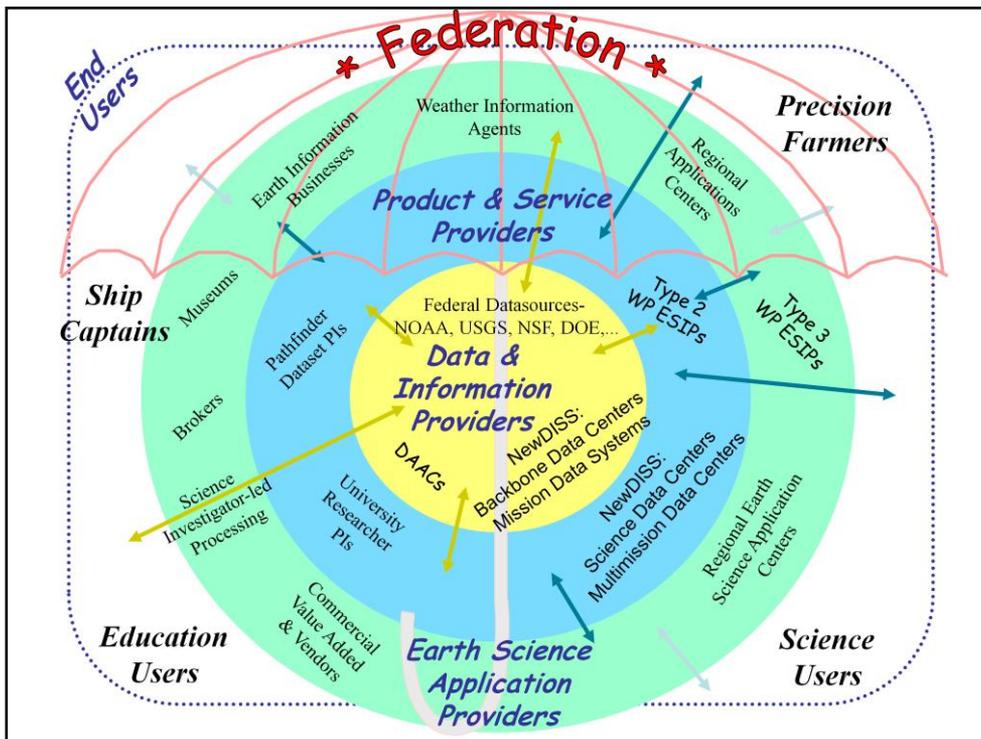
Uniform data system distributed to
various data repositories



PI-led science affiliated with
heterogeneous data centers for
processing, archive and distribution

SEEDS “Petri Dish” with ESIP Federation Mapping





Data Integration and Mining: From Global Information to Local Knowledge

