NASA's Earth and Space Science Data and Opportunities

NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation

Panel: Vision, Resources, and Opportunities for Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation

H. K. "Rama" Ramapriyan
Assistant Project Manager, ESDIS Project
NASA/Goddard Space Flight Center
October 11, 2007
Rama.Ramapriyan@nasa.gov

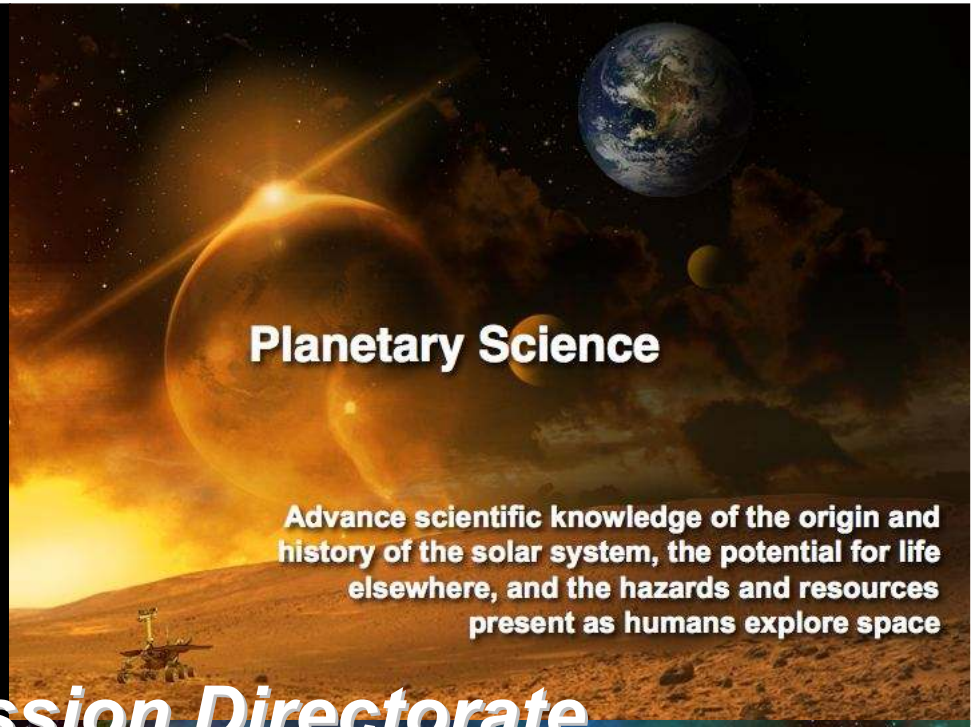
Agenda

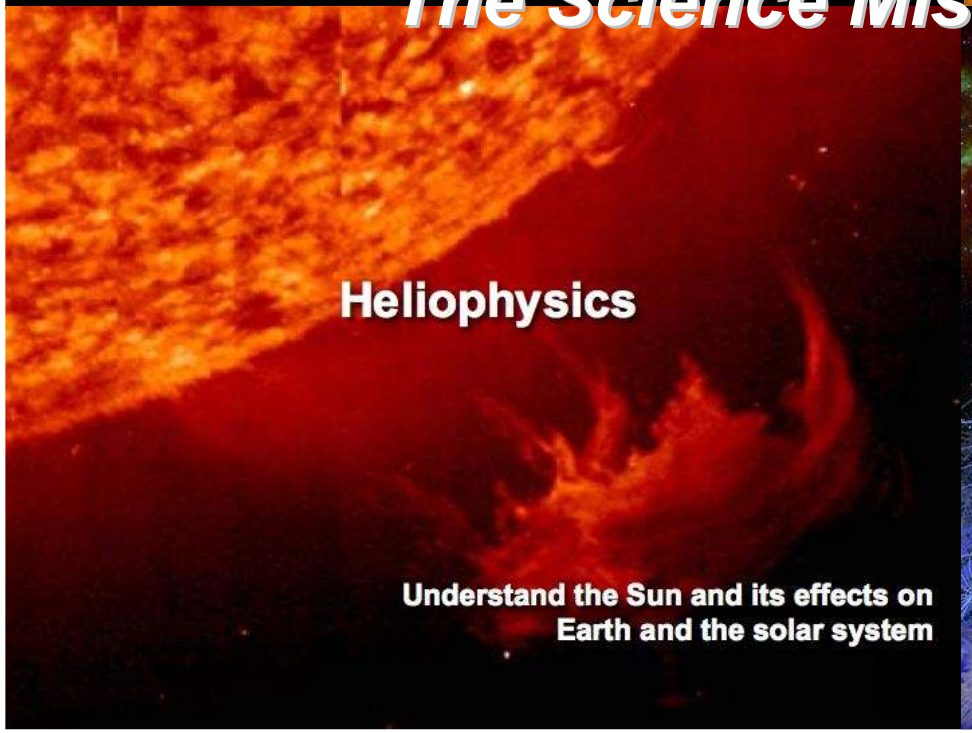
- NASA Strategic Goals/Sub-goals in Earth and Space Sciences
- Data Systems Context and Vision
  - Earth Science
  - Space Science
- Common Themes
- Opportunities

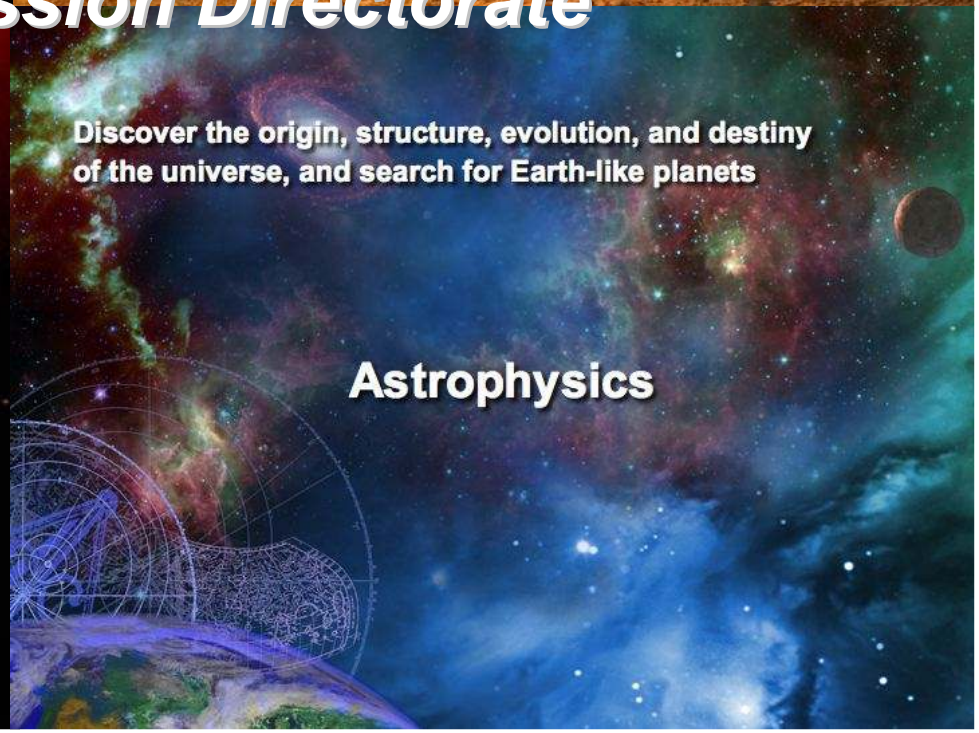
NASA Strategic Goal

- NASA's Strategic Plan – 2006 (http://www.nasa.gov/about/reports/index.html)
  - Strategic Goal 3: Develop a balanced overall program of science, exploration, and aeronautics consistent with the redirection of the human spaceflight program to focus on exploration.
    - Sub-goals 3A-3D: 4 major science disciplines (shown on next chart)

The Science Mission Directorate





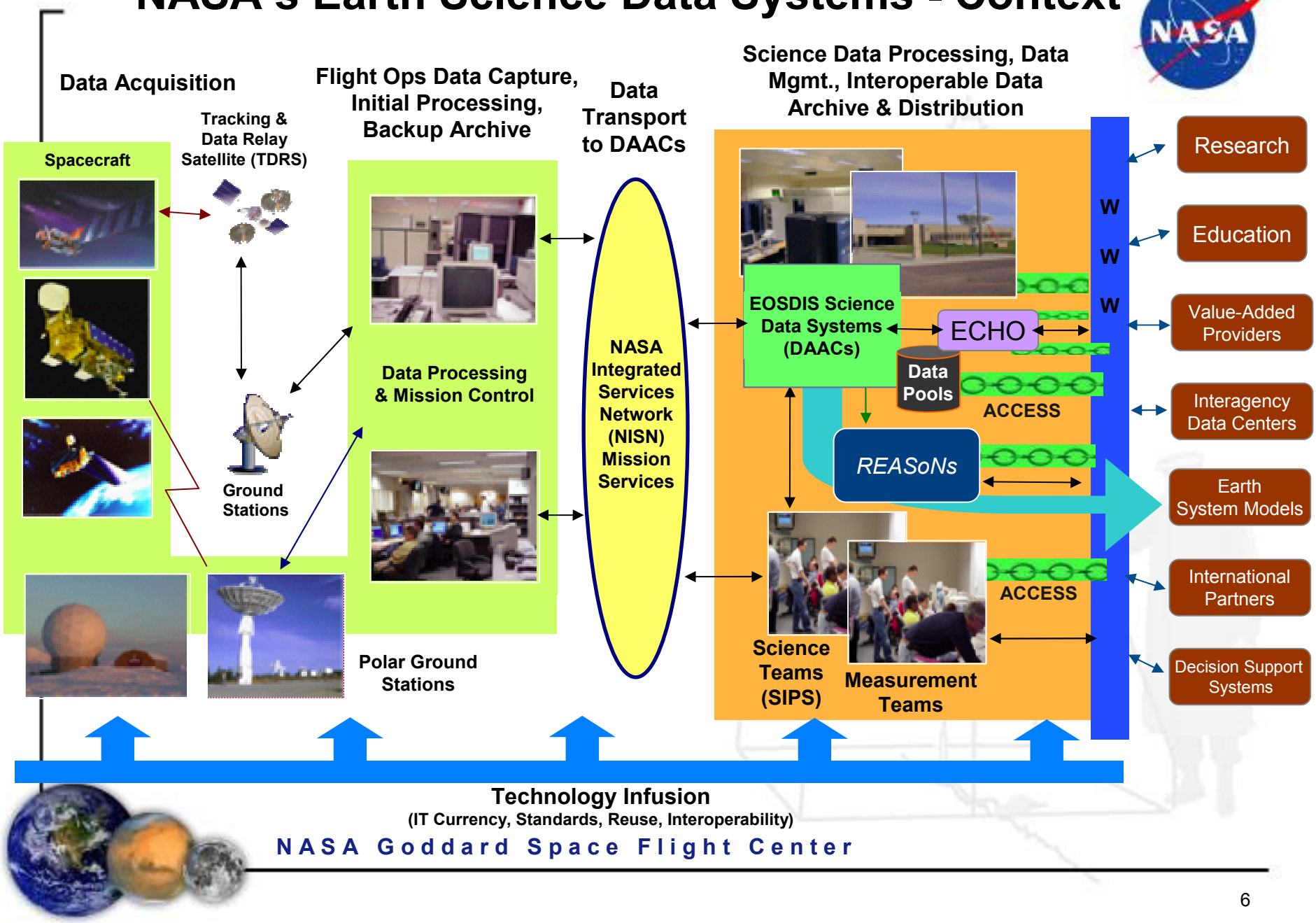


Earth Science Data Systems – Present State

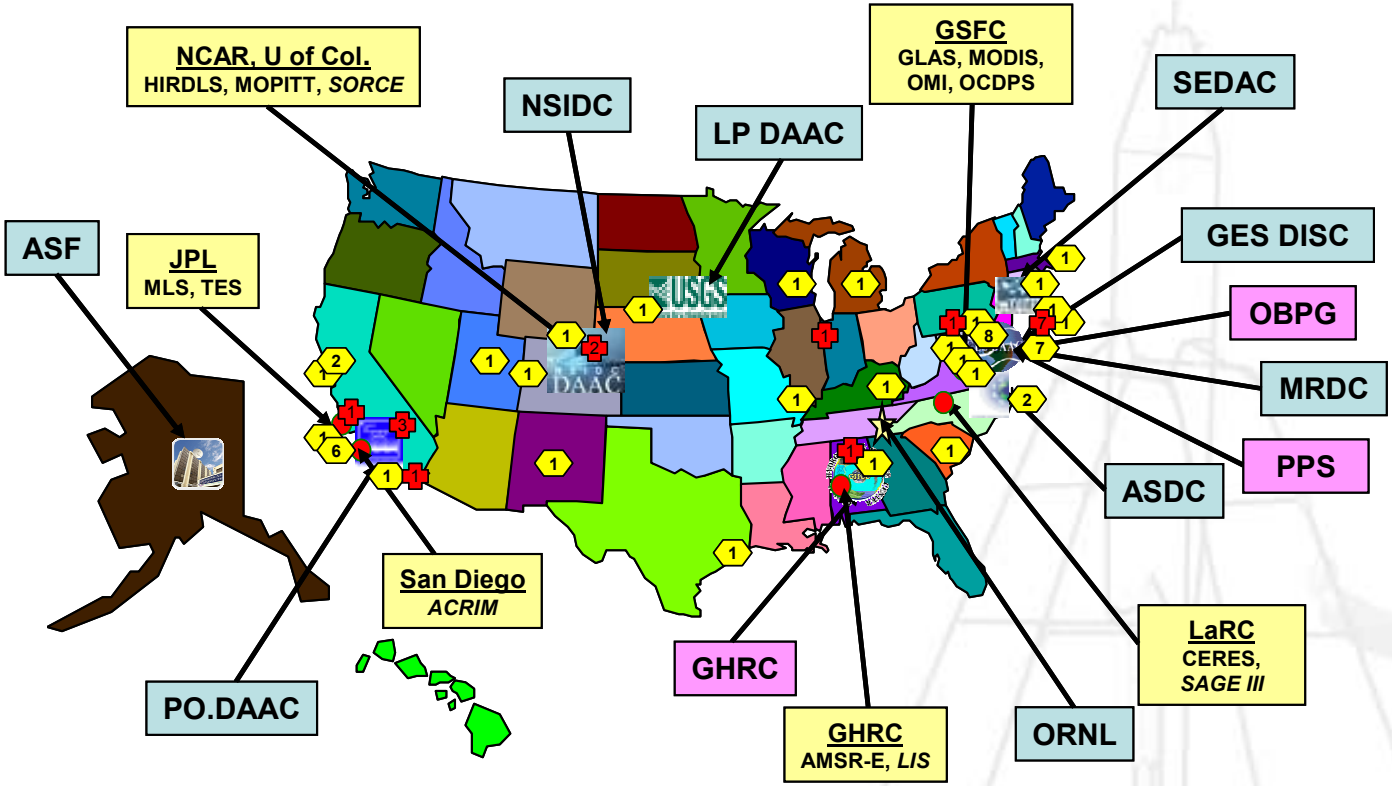
- Numerous (100s of) data systems are in place providing remotely sensed data and derived products to users
- NASA plays a very active role in this area
  - Core Capabilities (Robust basic infrastructure - e.g., EOSDIS with its DAACs)
  - Community Capabilities (specialized and innovative services to data users and/or research products - e.g., REASoN and ACCESS Projects)

- EOSDIS = Earth Observing System Data and Information System; DAAC = Distributed Active Archive Center; REASoN = Research, Education and Applications Solution Network; ACCESS = Advancing Collaborative Connections for Earth System Science

NASA's Earth Science Data Systems - Context



NASA Earth Science Data Systems - Distribution



EOSDIS Data Centers

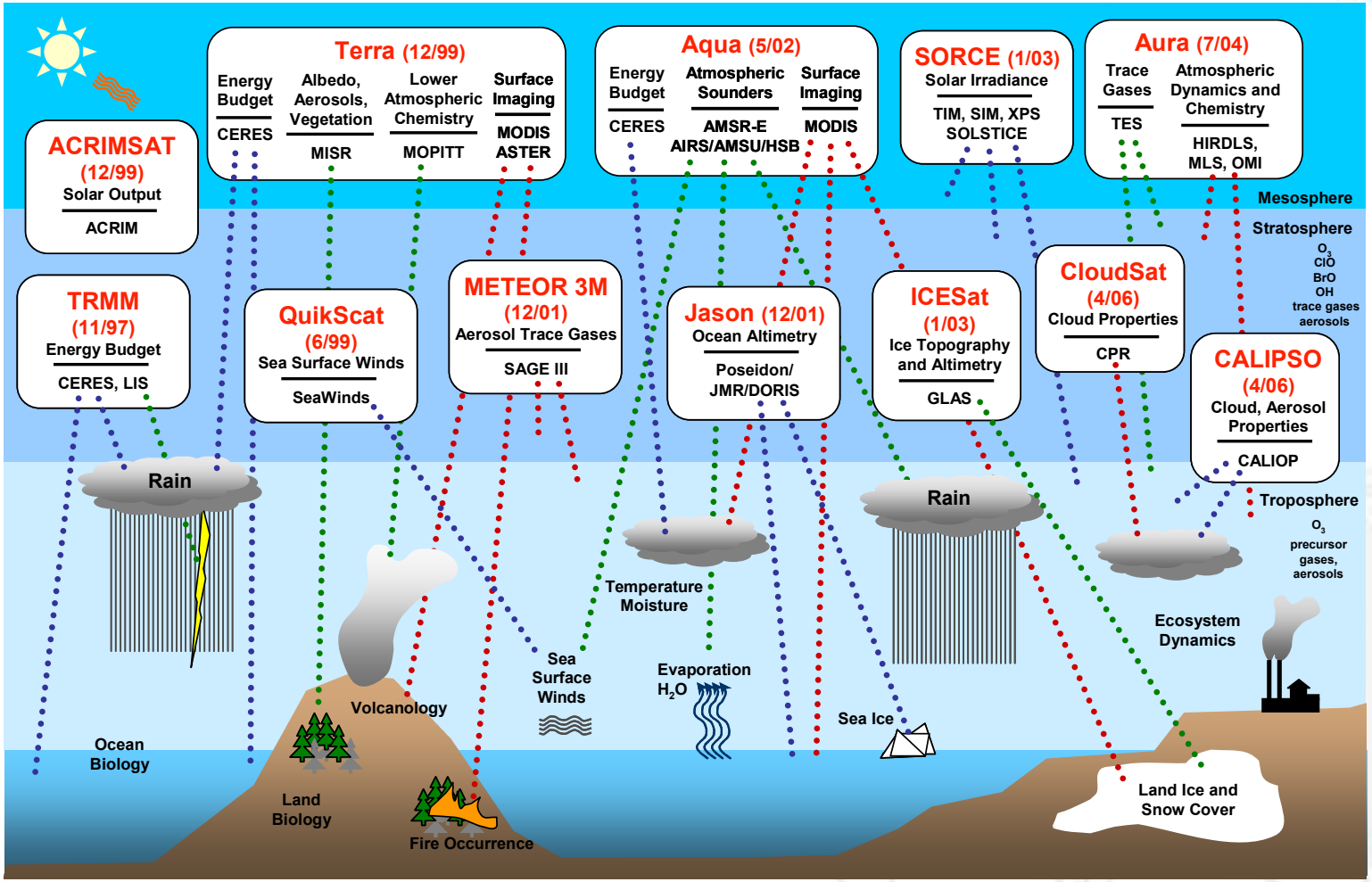
Related Data Providers

Science Investigator-led Processing Systems (SIPSs)

41 REASoN

17 ACCESS

EOSDIS Manages Data For All 24 EOS Measurements



Mission & Science 04102007.ppt

EOSDIS – A few key metrics

EOSDIS Science Elements	No.
EOSDIS Data Centers (DAACs)	9
Science Investigator-led processing Systems (SIPSs)	14

Partnerships	No.
U.S.	8
International	16

EOSDIS Overall Metrics (Oct 1, 06 to July 31, 07)	EOSDIS Systems
System Interface Control Docs (ICDs)	50
Unique Data Products	2746
Number of Accesses at Data Centers	>8.0M
Distinct Users over FY07 at Data Centers	>2.8M

Missions	No.
Science Data Processing	7
Archiving and Distribution	35
Instruments Supported	75

EOSDIS Metrics for Oct 1, 06 to July 31, 07	Total
Daily Archive Growth without Deletion (TB/day)	2.13
Daily Archive Growth with Deletion (TB/day)	2.08
End User Daily Distribution Volume (in TB)	3.9
End User Distribution Products (in millions)	50.4
Total Archive Volume at the End of FY07 (in PB)	4.8

EOSDIS_Today_10052007. xls

EOSDIS Evolution 2015 Vision Tenets

Vision Tenet	Vision 2015 Goals
Archive Management	▪ NASA will ensure safe stewardship of the data through its lifetime. ▪ The EOS archive holdings are regularly peer reviewed for scientific merit.
EOS Data Interoperability	▪ **Multiple data and metadata streams can be seamlessly combined.** ▪ **Research and value added communities use EOS data interoperably with other relevant data and systems.** ▪ **Processing and data are mobile.**
Future Data Access and Processing	▪ **Data access latency is no longer an impediment.** ▪ **Physical location of data storage is irrelevant.** ▪ **Finding data is based on common search engines.** ▪ **Services invoked by machine-machine interfaces.** ▪ **Custom processing provides only the data needed, the way needed.** ▪ **Open interfaces and best practice standard protocols universally employed.**
Data Pedigree	▪ Mechanisms to collect and preserve the pedigree of derived data products are readily available.
Cost Control	▪ Data systems evolve into components that allow a fine-grained control over cost drivers.
User Community Support	▪ Expert knowledge is readily accessible to enable researchers to understand and use the data. ▪ Community feedback directly to those responsible for a given system element.
IT Currency	▪ **Access to all EOS data through services at least as rich as any contemporary science information system.**

Existing Astronomy & Space Science Data Infrastructure

- The Recent Past: many independent distributed heterogeneous data archives
- Today: NVO = the National Virtual Observatory (http://www.us-vo.org/)
  - Web Services-enabled: e-Science paradigm (middleware, standards, protocols)**
  - Part of the IVOA (International VO Alliance @ IVOA.net)
  - Precursor to VAO = Virtual Astronomical Observatory (NSF+NASA co-funded)
  - Provides seamless uniform access to distributed heterogeneous data sources
    - *"Find the right data, right now"*
    - *"One-stop shopping for all of your data needs"*
  - One of many VxO's – for example:
    - VSO = Virtual Solar Observatory
    - VSPO = Virtual Space Physics Observatory
    - NVAO = National Virtual Aeronomical Observatory
    - VITMO = Virtual Ionospheric, Thermospheric, Magnetospheric Observatory
    - VHO = Virtual Heliospheric Observatory
- ** IVOA-approved standards for data formats, data/metadata exchange, data models, registries, Web Services, VO queries, query results, semantic astronomical catalog table headings (UCDs = Uniform Content Descriptors)
- ** And of course: The Grid, Web Services, Semantic Web, etc. ...

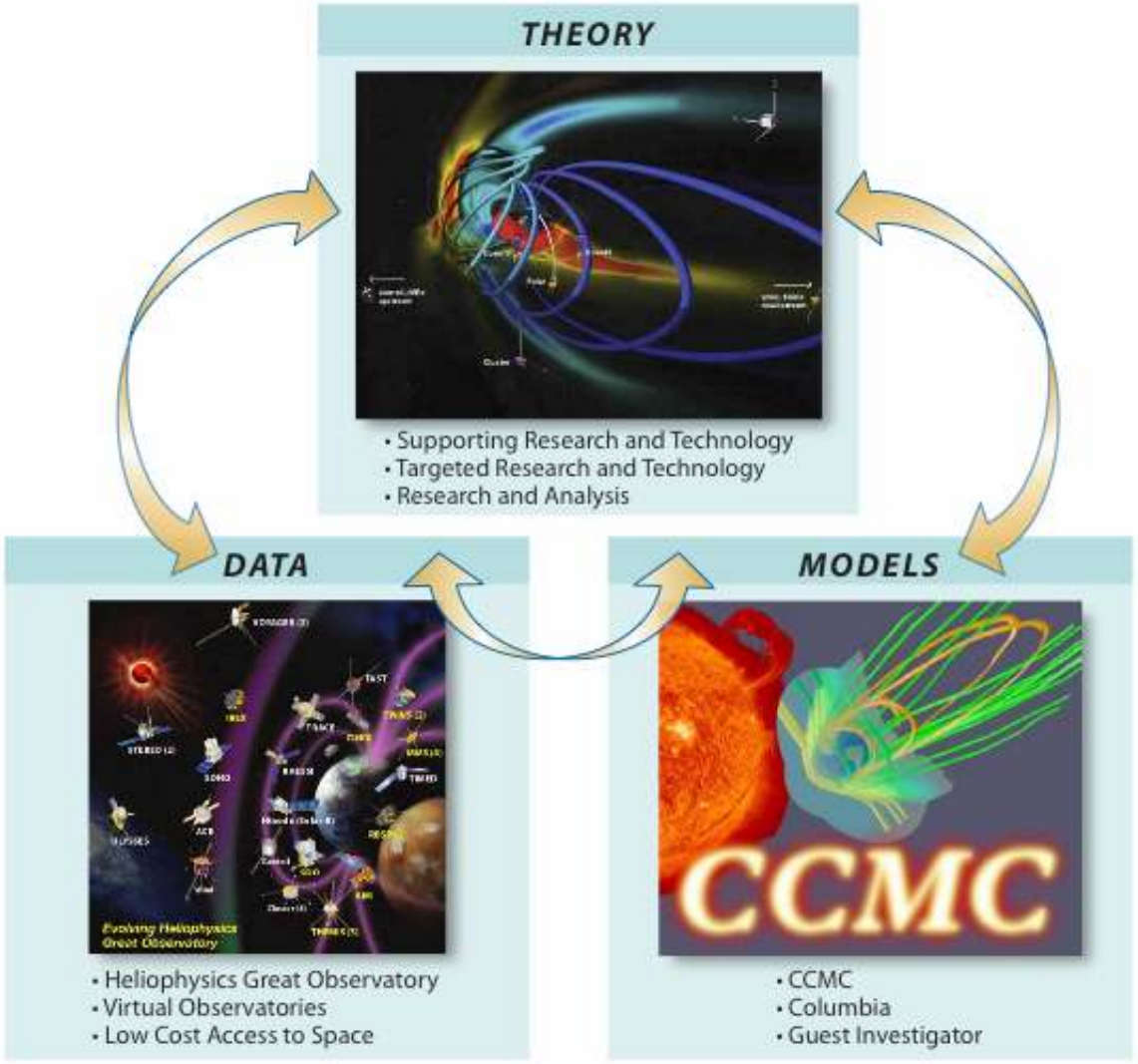
Massive Astronomy Data Collections

- NVO (IVOA) registry of ~14,000 data resources (collections, repositories, services)
- Large Astronomy Data Archives (including NASA space astronomy missions)
- Large Astronomy Sky Surveys (past and present) = uniform data sets, including:
  - MACHO and related surveys for dark matter objects: ~ 1 Terabyte
  - DPOSS (Digital Palomar Observatory Sky Survey): 3 Terabytes
  - 2MASS (2-Micron All Sky Survey): 10 Terabytes
  - GALEX (Ultraviolet Space Telescope): 30 Terabytes
  - SDSS (Sloan Digital Sky Survey): 40 Terabytes
- Future: Massive Astronomy Sky Surveys, including:
  - PanSTARRS: 10 Terabytes per night, 40 Petabyte final archive anticipated
  - LSST (Large Synoptic Survey Telescope @ http://www.lssto.org/):
    - Begin operations in 2012, with 3-Gigapixel camera
    - 10 Gigabytes every 30 seconds
    - 30 Terabytes every night for 10 years
    - 100 Petabyte final image data archive anticipated – all data are public!!!
    - 30 Petabyte final catalog anticipated
    - Real-Time Event Mining: 10,000-100,000 events per night, every night, for 10 yrs
    - Repeat images of the entire night sky every 3 nights: *Celestial Cinematography*

Common Themes (Earth and Space Sciences)

- Multitude and diversity of missions
- Volume, richness, complexity, and breadth of types of data
- Highly distributed and heterogeneous data sources
- Increasingly interdisciplinary nature of research
- Science cultural diversity
- Issues associated with requisite robust infrastructure
  - Standards, interoperability, commonality, etc.
  - New technology infusion
  - Balancing competing constraints for open access vs. protection (e.g., IT Security)

Toward an Integrated Agency Program



"…, the research and analysis mission of the Agency is extended to the utilization of new knowledge for societal benefit by forming partnerships with industry, academia, and other governmental agencies that are also engaged in the similar endeavors."

Opportunities

- Data-rich environment helps both NASA and non-NASA funded investigators
  - Publicly available data
  - Free-of-charge
- NASA's Information Technology R&D programs (mid-TRL)
  - Advanced Information Systems Technology (AIST) Program
  - Applied Information Systems Research Program (AISRP)
  - Research, Education and Applications Solutions Network (REASoN)
  - Advancing Collaborative Connections for Earth System Science (ACCESS) Program

AIST Data Mining Solicitation (Mini-NRA) 2004

- Goals of Proposed Research
  - **Tools for warehousing, data mining, and knowledge discovery**
  - **Technologies to facilitate queries/access of multi-disciplinary data**
  - **Techniques to facilitate customized data services**
- This mini-NRA addressed data mining technologies for two challenge areas
  - **Ocean Biology and Biogeochemistry Data Mining**
  - **Data Mining for Climate and Weather Models**
- Development of component, subsystem or system data mining technologies with entry Technology Readiness Levels (TRLs) of 4 or greater, with at least one TRL advancement over the duration of research (up to 2 years)
- 6 awards were made

AIST Data Mining Solicitation: Awards

Proposal Title	PI	Science Impact	Technology Category
Mining Massive Earth Science Data Sets for Climate and Weather Forecast Models	Amy Braverman / JPL	Creation of highly complex & dense summary data sets for validation of climate & weather models; **Maps to Climate & Weather**	Uses data reduction methods (thinning) to create summary data sets from observational data
Spatiotemporal Data Mining System for Tracking and Modeling Ocean Object Movement	Yang Cai / Carnegie Mellon University	Improve monitoring and prediction of ocean features (SeaWiFS data) & phenomena including sediment, blooms (differentiate natural and human induced features); **Maps to Carbon Cycle, esp. in coastal zone**	Development and implementation of a generalized spatiotemporal data mining system to track identified objects
Selection technique for thinning satellite data for numerical weather prediction	Ross Hoffman / AER, Inc.	Improve data selection for data assimilation of QuikSCAT, other data into weather models to better represent sub-grid features; **Maps to Weather**	Wavelet analysis for data selection to produce thinned satellite data; evaluate wavelet vs regular decimation (every nth observation)
Rapid Characterization of Causal Interactions among Climate/Weather System Variables: An Advanced Information-Theoretic Approach	Kevin Knuth / ARC	Enables understanding of causal relationships of forcing response & feedback among climate variables (e.g. clouds, evap, precip, ocean temp). **Maps to Global Climate Change**	Techniques to identify, characterize and quantify causal interactions & uncertainties, using tools from the field of information theory
Data Mining for Understanding the Dynamic Evolution of Land-Surface Variables: Technology Demonstration using the D2K Platform	Praveen Kumar / Univ of Illinois	Improve parameterization of vegetation processes in forecast models for weather & climate prediction; **Maps to climate & weather**	Heterogeneous data set integration extracting features and producing visualization using cluster and grid computing
Interactive Analysis of Heterogeneous Data to Determine the Impact of Weather on Crop Yield	Kiri Wagstaff / JPL	Prediction of crop yields based on 2-step automated classification techniques with improved performance (over neural nets) & prediction accuracy; **Maps to Weather impact on agriculture, & Carbon Cycle**	Statistical data mining toolkit integrating machine-learning techniques, Support Vector Machines, & models for feature classification, clustering and prediction

Opportunity:
A "discovery infrastructure"

A tool set to identify, classify, quantify, and catalog properties of features and events in large volumes of multi-dimensional data from all heliophysics observatories and models:

Component I: A portfolio of autonomous algorithms to extract contents from image sequences: find, classify, and quantify events throughout heliospace, ideally in 'pipeline processing' at archive site.

Status: needs development:

Feature-recognition algorithms, autonomous data-model comparison, data/model visualization tools, etc., currently see inadequate progress; explicit stimuli should change that:

Suggestion: "stimulate the community's activities in this area by creating one or more explicit proposal categories for the development, test, and validation of automated tools to identify events, find features, and otherwise exploit the much larger data sets"

Component II: A knowledge base of metadata with a system for compound queries.

Status: under development:

This system complements VxO-s, COSEC, and EGSO, by focusing on data contents. One version of this component is being developed by the TRACE (observer-guided entries), HINODE (planner-guided entries), and SDO (query functionality) teams.

Component III: A suite of tools for visual comparison and multi-parameter representation of a wide variety of observational and model data

Status: needs development:

To enable model-data overlays, customizable (web) displays, movie tools (multi-channel blending, interactive zoom/pan via the internet, …), applicable to the early 'browsing' phase to final analysis phase.

Potential tool classes:

- Autonomous feature/event-finding modules: solar flares, coronal loops, auroral events, Interplanetary Coronal Mass Ejections (ICMEs) (in stereo), …
- Model-enabled measurements: coronal currents and free energy, coronal thermal structure, ICME geometry, induced ring currents, …
- Visualization/presentation tools: model-data overlays, customizable web displays, movie tools (dozen-channel blending, interactive zoom/pan via the internet, …)
- Query tools: ontology-based inferences, correlation trawlers, …