Conceptual Study of Intelligent Data Archives of the Future

H. K. Ramapriyan, Steve Kempler, Chris Lynnes, Gail McConaughy, Ken McDonald, Richard Kiang
NASA Goddard Space Flight Center

Sherri Calvo, Robert Harberts, Larry Roelofs,
Global Science and Technology, Inc.

Donglian Sun
George Mason University

Study Objectives

- Formulate concepts and architectures that support data archiving for NASA science research in the 10 to 20 year time frame
- Focus on architectural strategies that will support intelligent processes and functions
- Identify and characterize science research scenarios that drive intelligent archive requirements
- Assess technologies and research that will be needed for the development of an intelligent archive

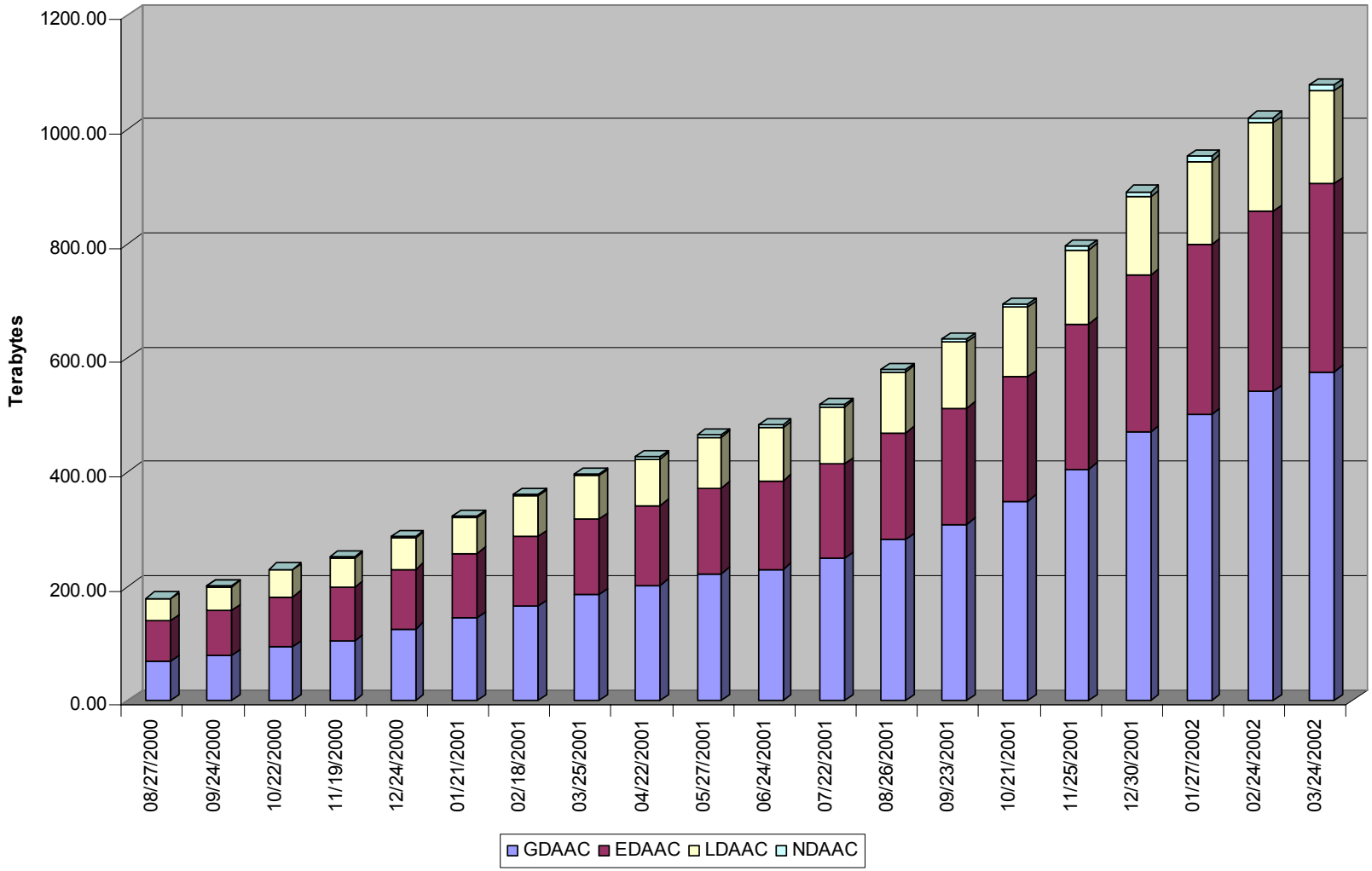
Problem: Very Difficult to Find and Use Right Data, Information, & Knowledge

- Voluminous scientific data archives
  - Acquisition and accumulation rates continue to outpace our ability to manage, discover, and exploit scientifically meaningful data, information and knowledge
  - Human-based strategies for managing, searching, identifying, and creating required data and information for research purposes are time-consuming and cost-prohibitive
  - Increasing numbers and kinds of data sources (sensors, models, users, etc.) are generating large quantities of data
  - Model outputs are expected to be even more voluminous than satellite data

The Problem: Very Difficult to Find The Right Data, Information, & Knowledge (continued)

- Data and service providers are distributed geographically and belong to diverse institutions with their own data organization and access mechanisms
- Diverse sources produce heterogeneous data, information and knowledge
- Existing archive architectural strategies do not easily support intelligent understanding
- Current scientific archive strategies do not allow for/exploit adequate intelligence
  - On-demand tailoring of data, creation of fused products, conversion from data into information targeted to a end-user, etc. are not generally supported
- Next two charts show examples in Earth sciences. Space science data are expected to be similarly voluminous and distributed

Cumulative volume of part of EOSDIS holdings



NASA Earth Science Enterprise "Data Center" Locations

ESE supports 68 data centers (some of which at the same location), widely distributed geographically. Additional data centers, including NOAA's NCDC and Unidata, are networked through Membership in the ESIP Federation.

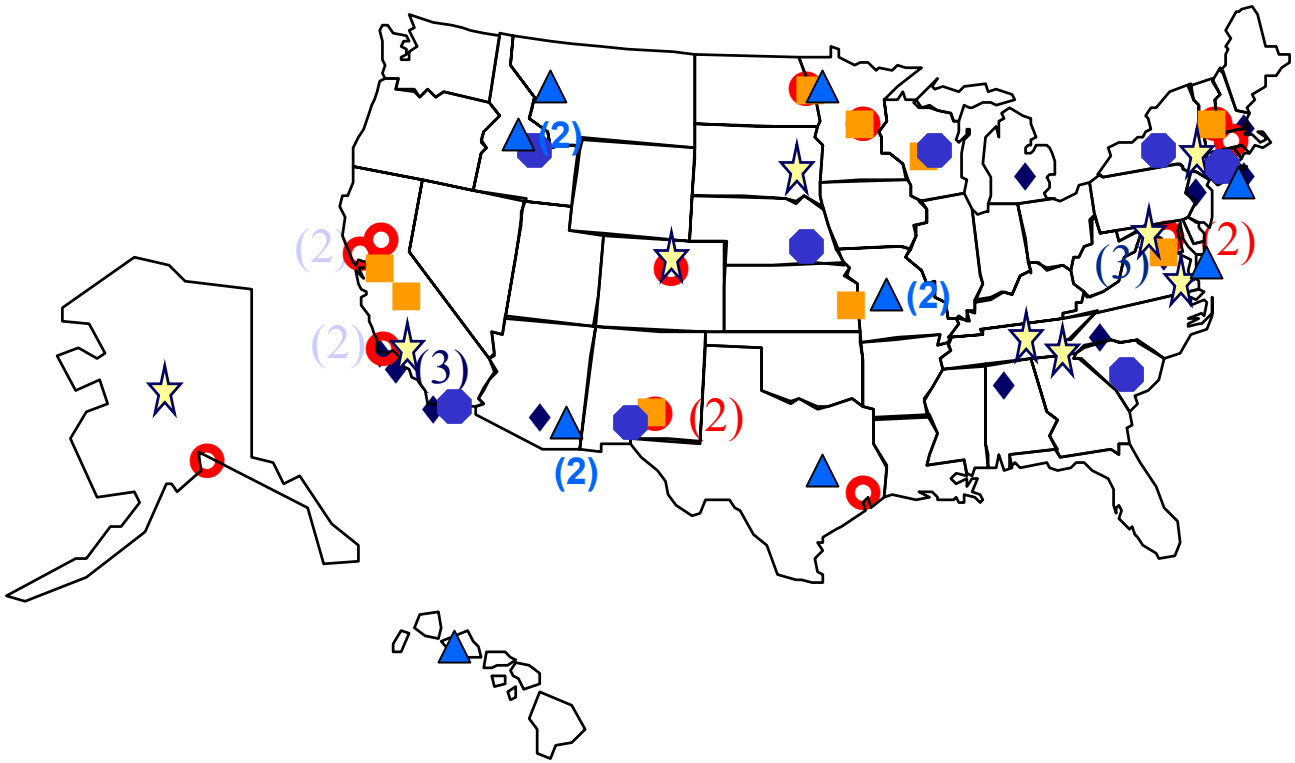


Chart from Martha Maiden, NASA HQ

An Intelligent Archive (IA)

- An IA includes all items stored to support “end-to-end” research and applications scenarios
  - Stored items include:
    - **Data, information and knowledge**
    - **Software needed to manage holdings**
    - **Interfaces to algorithms and physical resources to support acquisition of data and their transformation into information and knowledge**
- Architecture expected to be highly distributed so that it can easily adapt to include new elements as data and service providers
- Will have evolved functions beyond that of a traditional archive
  - The “borders” of an intelligent archive are intrinsically fuzzy, but may be determined in practice by institutional structure and expectations
- Will be based on and exploit technologies in the 10 to 20 year time range
- Will be highly adaptable so as to meet the evolving needs of science research and applications in terms of data, information and knowledge

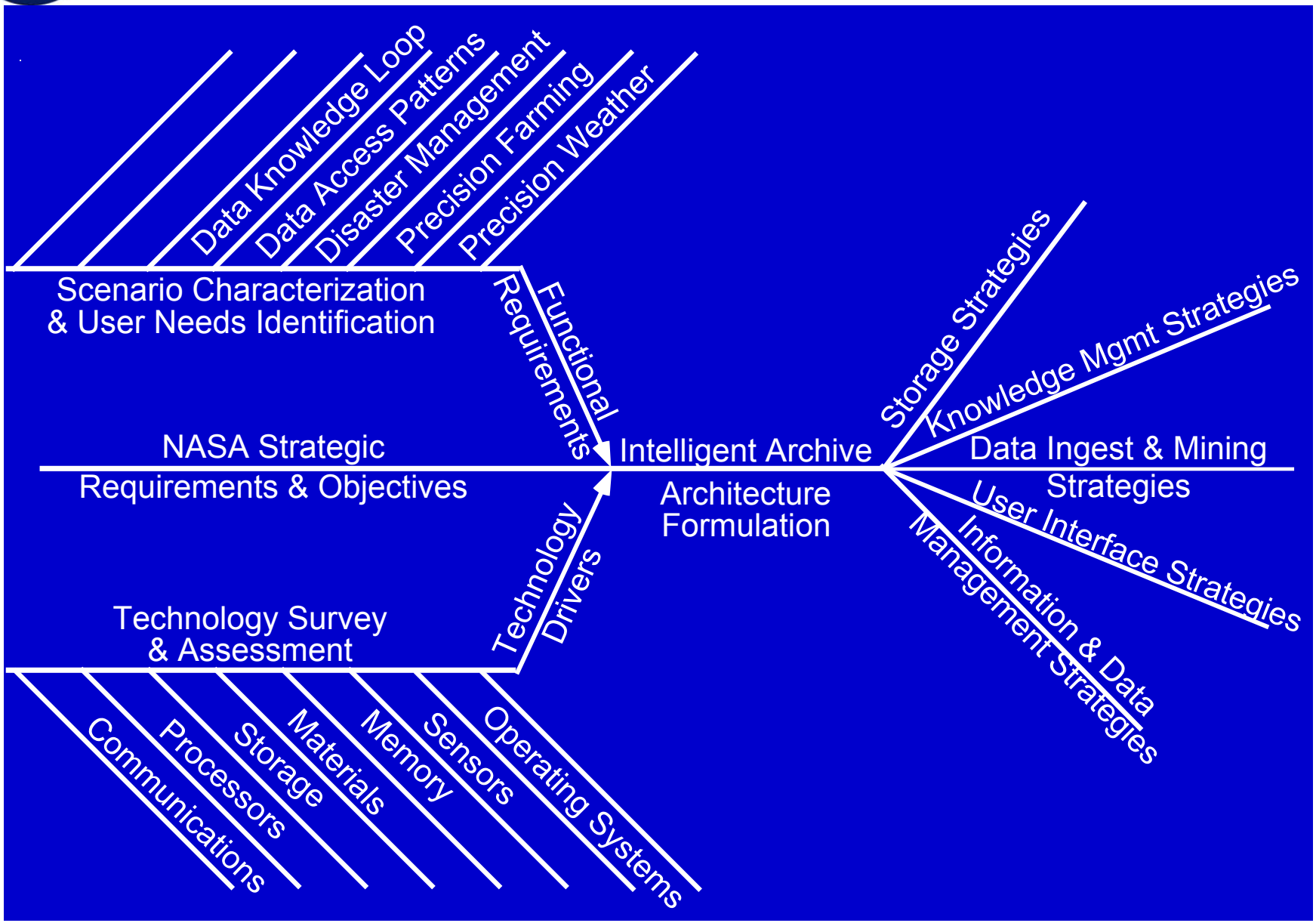
Data, Information and Knowledge

- Data: output from a sensor, with little or no interpretation applied
  - Examples: Scientific instrument measurements, market past performance

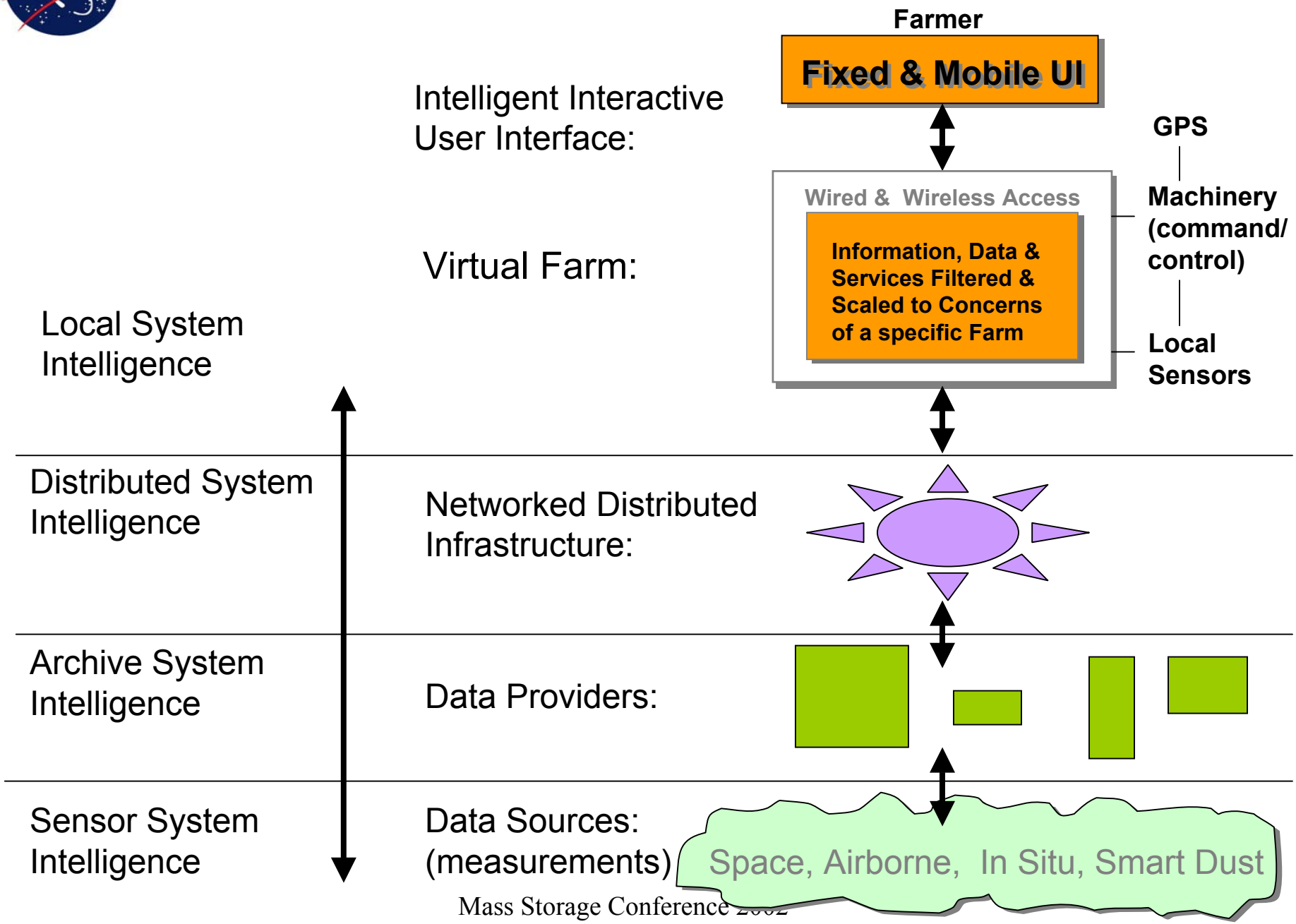
- Information: a summarization, abstraction or transformation of data into a more readily interpretable form
  - Examples: Results after performing transformations by data mining, segmentation, classification, etc., such as a Landsat scene spatially indexed based on content , assigned a “class” value and subset for an application, or National Weather Service storm monitoring fused with a GIS of the spatial location of the Beltway.

- Knowledge: a summarization, abstraction or transformation of information that increases our understanding of the physical world
  - Examples: Predictions from model forward runs, published papers, output of heuristics or other techniques applied to information to answer a “what if” question such as “What will the accident rate be if an ice storm hits the Beltway between Chevy Chase and the Potomac crossing at 7 a.m.?”

Overview of Approach



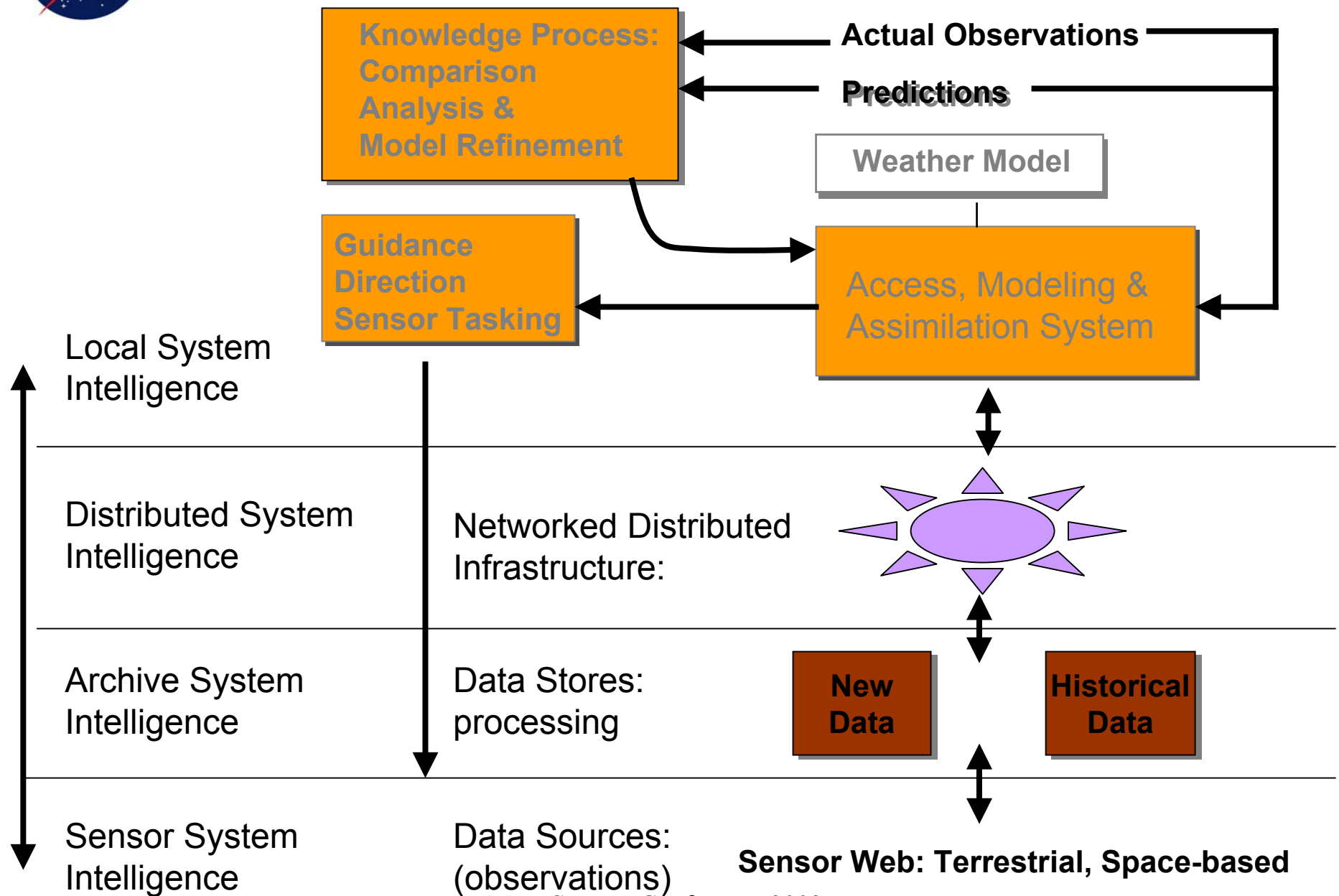
Precision Agriculture Scenario



Precision Agriculture Scenario

- **Data Volumes**
  - **Estimated 1.5 TB per year for a 1000-acre farm, including satellite and airborne remote sensing data, in situ data, visualizations, modeling, etc.**
  - **In 2001 there were 2,158,000 farms in the U.S. averaging 436 acres each, for a total of 940,000,000 acres of farmland (source: Wisconsin Agricultural Statistics Service via USDA). Extrapolating, we have a data stream of 1,411 PB per year for all U.S. farms.**
- **Data Sources**
  - **Remote sensing at spatial resolutions as small as 1 foot and temporal resolutions as fine as 1 hour; precision weather forecasts from three hours ahead to longer-range climate predictions for months and years; land profiles including soils, moisture, elevations, drainage patterns and water sources, digital ortho-photo quadrangles (images with integrated USGS topography maps), digital elevation models, geo-rectified spatial data, ecological zone profiles, biodiversity inventory, local calibrations including ground truth.**
- **Types of analysis needed**
  - **Planting conditions indicators, crop monitoring (growth, maturity, health and stress indicators), weed and pest identification and tracking, chemical and other intervention impact prediction and analysis, weather conditions, advance forecasts, environmental alerts, microclimate surveys, soil types and depths.**

Weather Prediction Scenario

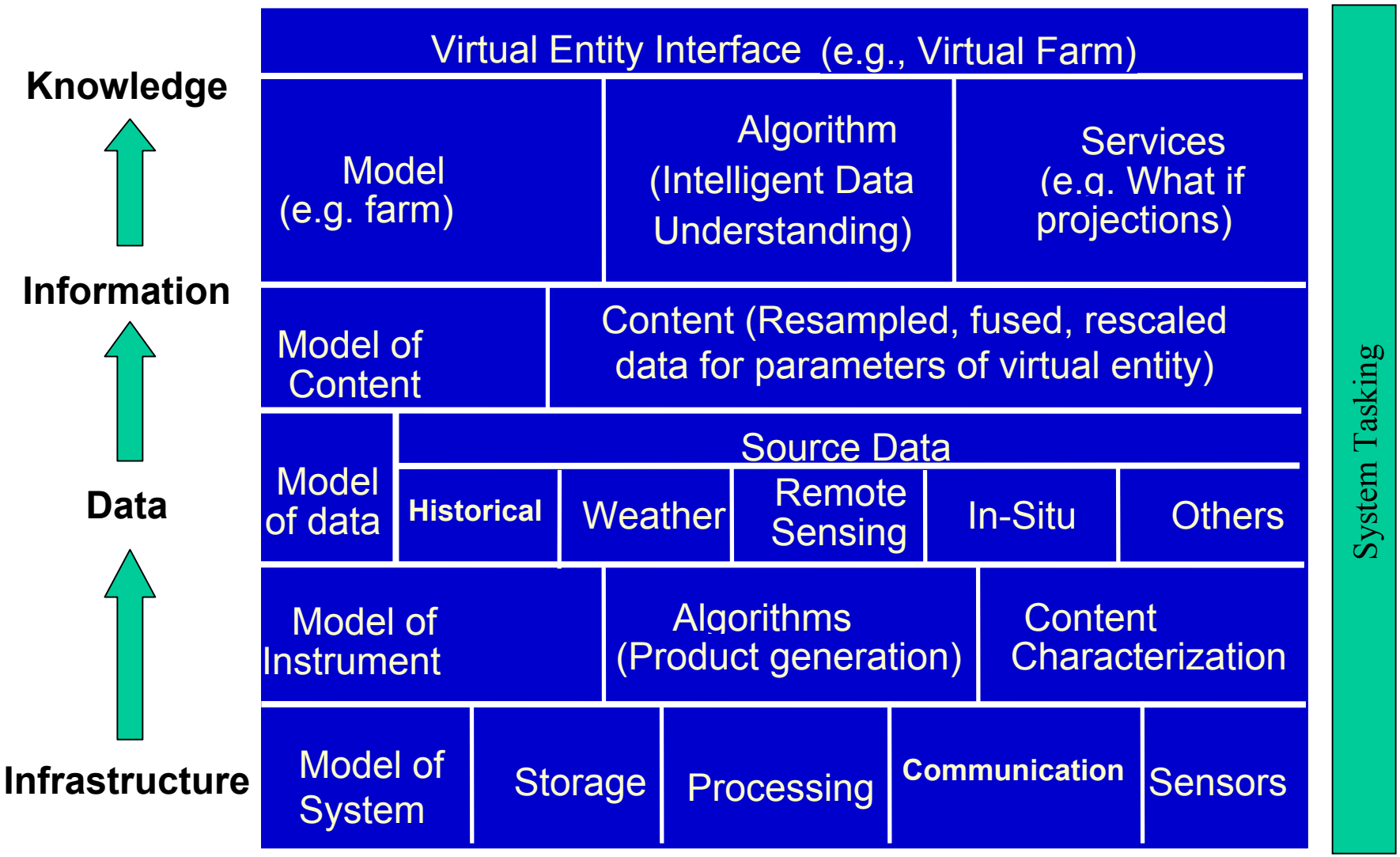


Weather Scenario

- Data Volumes
  - Approximately 7.5 TB/day by 2025 (source: ESTO* Weather Prediction Technology Investment Study)
- Data Sources
  - Space-based, airborne and terrestrial sensors, models and analytical tools, the models will have about 90 million unique grid points with a million observations)
- Types of analysis needed
  - Structure information in the free atmosphere every 3 hours, every 25 km globally, and vertically from the surface to 80 km altitude; global 3D distribution of cloud height, cloud depth, aerosols, water/ice, and suspended precipitation rates; land and sea surface temperature, land surface moisture, albedo, vegetation type; planetary boundary layer depth

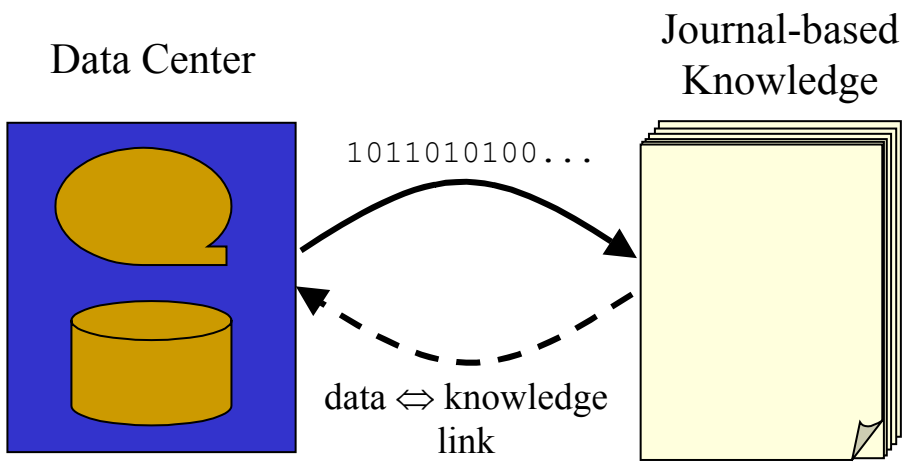
* NASA Earth Science Technology Office

An Example Architecture Abstraction



Understanding Advanced Knowledge Access:
The "Data-Knowledge Loop"

- A complete Intelligent Archive would serve data *plus* knowledge derived from that data
- Study explored knowledge feedback and its architectural relationship to the data it was derived from
- A proof-of-concept of a closed data-knowledge loop was developed
  - Datasets point to knowledge and vice versa
  - Mined on-line sources of abstracts and articles for references to EOSDIS data to develop web hyperlinks



Some Relevant Current Technology Efforts

- **Earth Observing System Data and Information System (EOSDIS)**
  - Manages data from NASA's Earth science research satellites and field measurement programs, providing data archiving, distribution, and information management services.
  - Distributed system with many interconnected nodes (Distributed Active Archive Centers (DAACs) and Science Investigator-led Processing Systems (SIPSs).
  - Now handling operationally extraordinary rates and volumes of scientific data (including processed data, ingest rates into the archive are over 2.5 TB per day)
  - Archive capacity passed the 1 PB mark in February 2002
- **Earth Science Information Partners' (ESIP) Federation**
  - Brings together government agencies, universities, non-profit organizations, and businesses to make Earth Science information available to a broader community.
  - Goal is to establish and continuously improve science-based processes to increase quality and value of Earth science products and services throughout their life-cycle for the benefit of stakeholder communities.
  - Maintains comprehensive searchable inventory; over 2,000 data sets currently available
  - Consists of both NASA-funded entities (including EOSDIS DAACs) and others
  - Together, in FY 2001, NASA-funded members provided over 15 million products to 2.4 million distinct users
- **GRID**
  - Information science activity to address distributed computing infrastructure for advanced science and engineering
  - It has emerged as an important architectural strategy, distinguished from conventional distributed computing, by its focus on large-scale resource sharing, innovative applications, and high-performance computing
  - A GRID architecture could easily evolve to adapt to changes in technologies and missions so as to incorporate intelligent archive elements and functions as they mature and become available
- **Intelligent Systems Program/Intelligent Data Understanding (IDU)**
  - Several funded NASA and university researchers working on NASA-relevant problems
  - Focuses on the development of methodologies for extracting meaningful information from large, diverse databases.
  - Work funded under IDU uses data and addresses questions that support multiple enterprises in NASA, ESE being one of the prime customers

Conclusion

- **Intelligent archives can evolve from present foundations**
- **However, more work is needed to make archives "intelligent"**
  - **Standard interfaces, terminology, protocols, representation of data, information and knowledge**
- **Some of the attributes of an intelligent archive**
  - It is aware of its data and knowledge holdings and is constantly searching new and existing data for unidentified objects, features or processes
  - Provides data to a science knowledge base in the context of research activities
  - Is able to exploit and use collected data in the context of a science enterprise
  - Is able to work autonomously to identify and characterize objects and events
  - Works with other autonomous information system functions to support research
  - Manages its activities and functions from sensor to user
  - Interacts with users in human language and visual imagery that can be easily understood by both people and machines