Data and IT Technology Roadmapping

PDS Management Council, St Louis

Dan Crichton

April 2015

Topics

- Introduction
- OCT TA-11 Roadmapping
- PDS Technology Challenges and Roadmapping

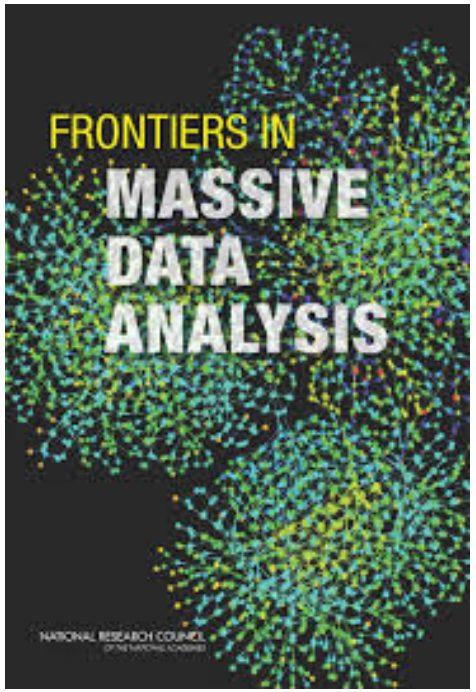
Introduction

- Multiple agencies are roadmapping and investing in future data technologies
- NASA, with its petabytes of data, has been a leader in developing scalable archives for science data
- NASA OCT has recently been working on technology roadmaps
- PDS is well positioned to take advantage of PDS4 and to plan its next technology roadmap

NRC Report:

Frontiers in the Analysis of Massive Data

- Chartered in 2010 by the National Research Council
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- Importance of systematizing the capture and analysis of data
  - Planetary, Astronomy, Earth Science drivers and capabilities identified
- Need for end-to-end lifecycle: from point of capture to analysis
- Integration of multiple disciplines experts
- Application of novel statistical and machine learning approaches for data discovery



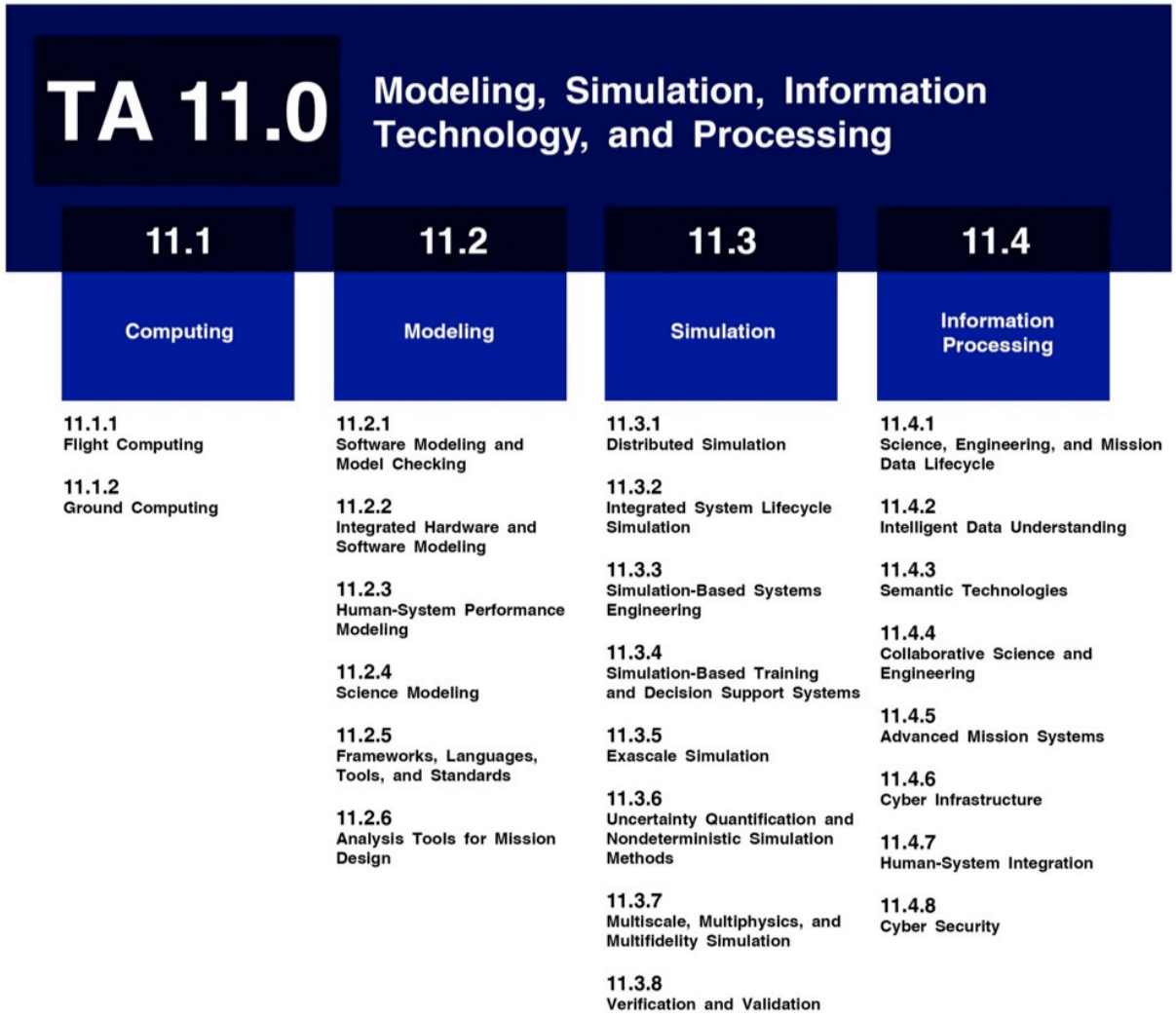
2013

OCT Roadmap

- NASA Office of the Chief Technologist Space Technology Roadmap
  - Multiple *Technical Areas* from launch propulsion to EDL to information systems
  - Released in 2010; reviewed by NRC
  - Currently being updated; center reviews and HQ have taken place
  - TA-11: Modeling, Simulation, Information Processing and Technology
    - Significant overhaul to 11.4: Information Processing

OCT/TA-11 Roadmap

11.0 Modeling, Simulation, Information Technology, and Processing	Goals:	Develop computing, modeling and simulation, and information technologies that are the basis of new solution paradigms across the breadth of NASA's missions. Enable the NASA mission through development of *virtual* technologies that increase our understanding and mastery of the *physical* world.
11.1 Computing	Sub-Goals:	Develop scalable, radiation-hardened flight processors, memory management and flight software to support more autonomous operations and data triage at the point of data collection. Exploit exascale supercomputing, data storage, and software development capabilities to enable 1,000 times larger mission-driven computations.
11.2 Modeling	Sub-Goals:	Develop autonomous, integrated, and interoperable approaches for models and model development. Increase productivity, improve performance, and manage risk through improvements in autonomy and integration in modeling for NASA's future missions.
11.3 Simulation	Sub-Goals:	Develop best-physics simulations of operative mechanisms that enable increases in system performance and management of uncertainty and risk across the entire lifecycle of NASA's distributed, heterogeneous, and long-lived mission systems.
11.4 Information Processing	Sub-Goals:	Develop software frameworks and toolsets that efficiently and reliably manage greatly increased volume, variety, and velocity of data across the science, engineering and mission data lifecycle while maintaining security of data. Enable advanced missions, effective remote and human-system collaboration, and greater system and crew autonomy through advanced software.


TA 11.0
Modeling, Simulation, Information Technology, and Processing
11.1
Computing
11.1.1
Flight Computing
11.1.2
Ground Computing
11.2
Modeling
11.2.1
Software Modeling and Model Checking
11.2.2
Integrated Hardware and Software Modeling
11.2.3
Human-System Performance Modeling
11.2.4
Science Modeling
11.2.5
Frameworks, Languages, Tools, and Standards
11.2.6
Analysis Tools for Mission Design
11.3
Simulation
11.3.1
Distributed Simulation
11.3.2
Integrated System Lifecycle Simulation
11.3.3
Simulation-Based Systems Engineering
11.3.4
Simulation-Based Training and Decision Support Systems
11.3.5
Exascale Simulation
11.3.6
Uncertainty Quantification and Nondeterministic Simulation Methods
11.3.7
Multiscale, Multiphysics, and Multifidelity Simulation
11.3.8
Verification and Validation
11.4
Information Processing
11.4.1
Science, Engineering, and Mission Data Lifecycle
11.4.2
Intelligent Data Understanding
11.4.3
Semantic Technologies
11.4.4
Collaborative Science and Engineering
11.4.5
Advanced Mission Systems
11.4.6
Cyber Infrastructure
11.4.7
Human-System Integration
11.4.8
Cyber Security

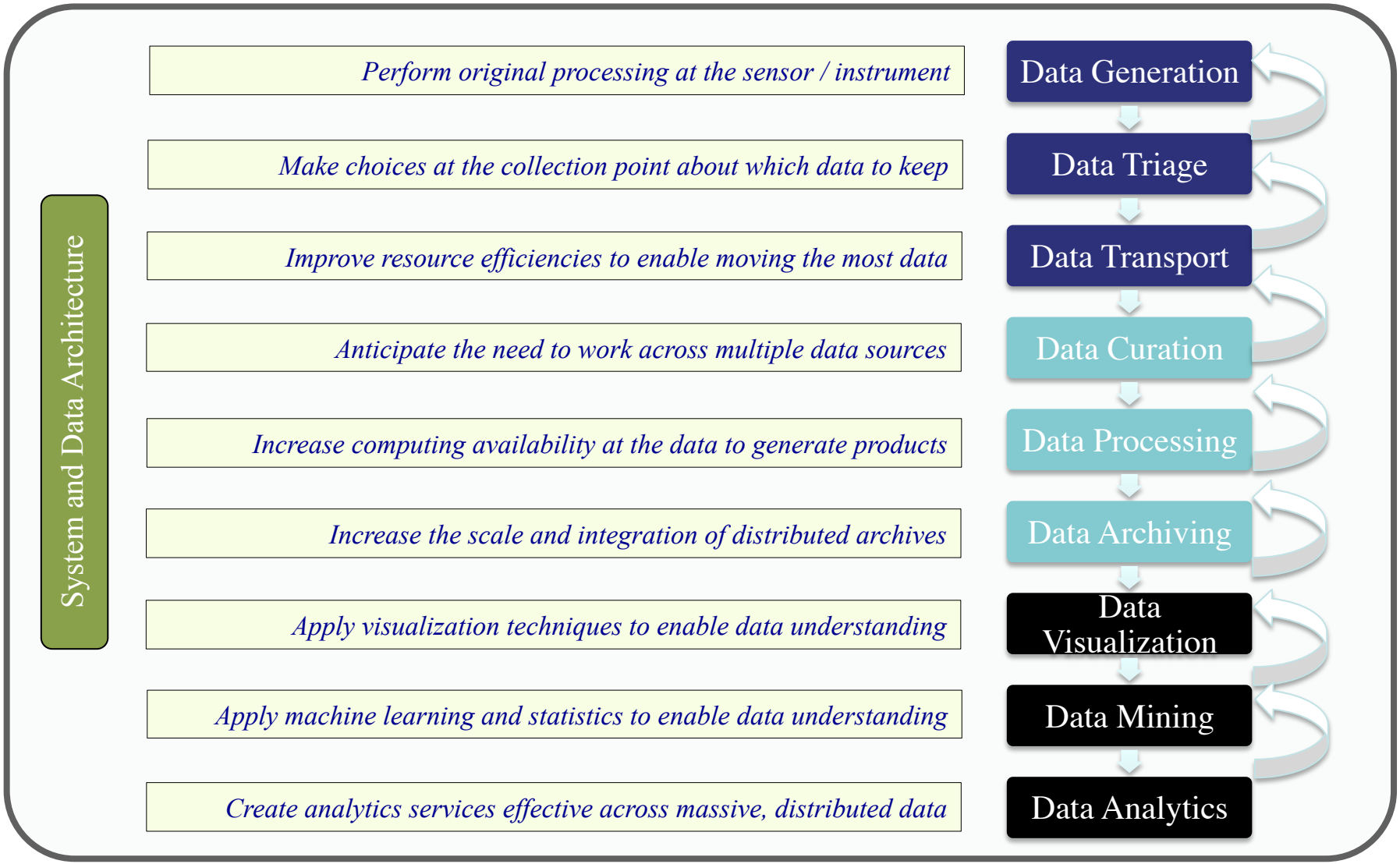
11.4: Information Processing

- **Science, Engineering, and Mission Data Lifecycle:** As the data-intensive nature of NASA science and exploration missions increases, there is an increasing need to consider the data lifecycle from the point of collection all the way to the application and use of the data.
- **Intelligent Data Understanding:** Intelligent data understanding (IDU) refers to the capability to automatically mine and analyze large datasets that are large, noisy, and of varying modalities (discrete, continuous, text, graph, etc.), in order to extract or discover information that can be used for further analysis or in decision making, on the ground or onboard. It is closely coupled to the capability to detect and respond to interesting events and/or to generate alerts.
- **Semantic Technologies:** Technologies that enable data understanding, analysis, and automated consulting and operations.

11.4: Cont…

- **Collaborative Science and Engineering:** Collaborative technology environments will allow distributed teams with disparate expertise and resources, including those of partner agencies and contractors, to work in a unified manner.
- **Cyber Infrastructure:** Includes storage and computation, data management services, distributed deployments, cross-cutting application to engineering, science and mission needs, cyber-security and assurance, and the lifecycle of data archiving and preservation.
- **Human-System Interaction:** Advances in information systems and interface design are needed to streamline access to mission systems and information to enhance mission capabilities and enable increased on-board autonomy.
- **Cyber Security:** Involves protecting information systems and data from attack, damage, or unauthorized access, and requires technologies for assurance of full-lifecycle information integrity and cyber security situational awareness and analysis.

Space observing data lifecycle framework



Data Challenges

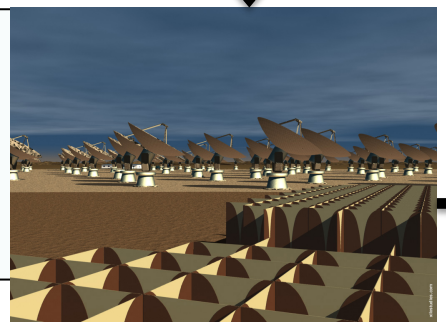
Emerging Solutions
- *Onboard Data Products*
- *Onboard Data Prioritization*
- *Flight Computing*

(1) Too much data, too fast; cannot transport data efficiently enough to store

Observational Platforms /Flight Computing

Science Archives and Analysis

Emerging Solutions
- *Low-Power Digital Signal Processing*
- *Data Triage*
- *Exa-scale Computing*



Emerging Solutions
- *Distributed Data Analytics*
- *Advanced Data Science Methods*
- *Scalable Computation and Storage*

(2) Data collection capacity at the instrument continually outstrips data transport (downlink) capacity

(3) Data distributed in massive archives; many different types of measurements

Ground-based Mission Systems

Sample Challenges by

Discipline	Volume	Variety	Velocity
Planetary Science	Data from on-board instruments will exceed communication capabilities. About 900 TB of data stored under new distributed PDS4 system.	Data is highly diverse supporting many sub-disciplines in planetary science. Its definition is based on a complex information model from PDS4.	Planetary data growing at about 200 TB/year. Newer instruments will require more upstream data triage.
Earth Science	Data archived in distributed repositories in the multi-PB range in NASA DAACs.	Data is highly diverse supporting many different instruments under HDF.	Data will continue to increase in multi-PB/ year. New missions such as NI-SAR will provide massive data (10s PB) to process and capture.
Astronomy	Data for new optical and radio instrument projects to grow substantially. Many PBs. Particularly LSST, but also SKA.	Observational data formatted by FITS standard.	Data from multiple observatories will be captured in real-time requiring triage algorithms to reduce data before archived.

Trends

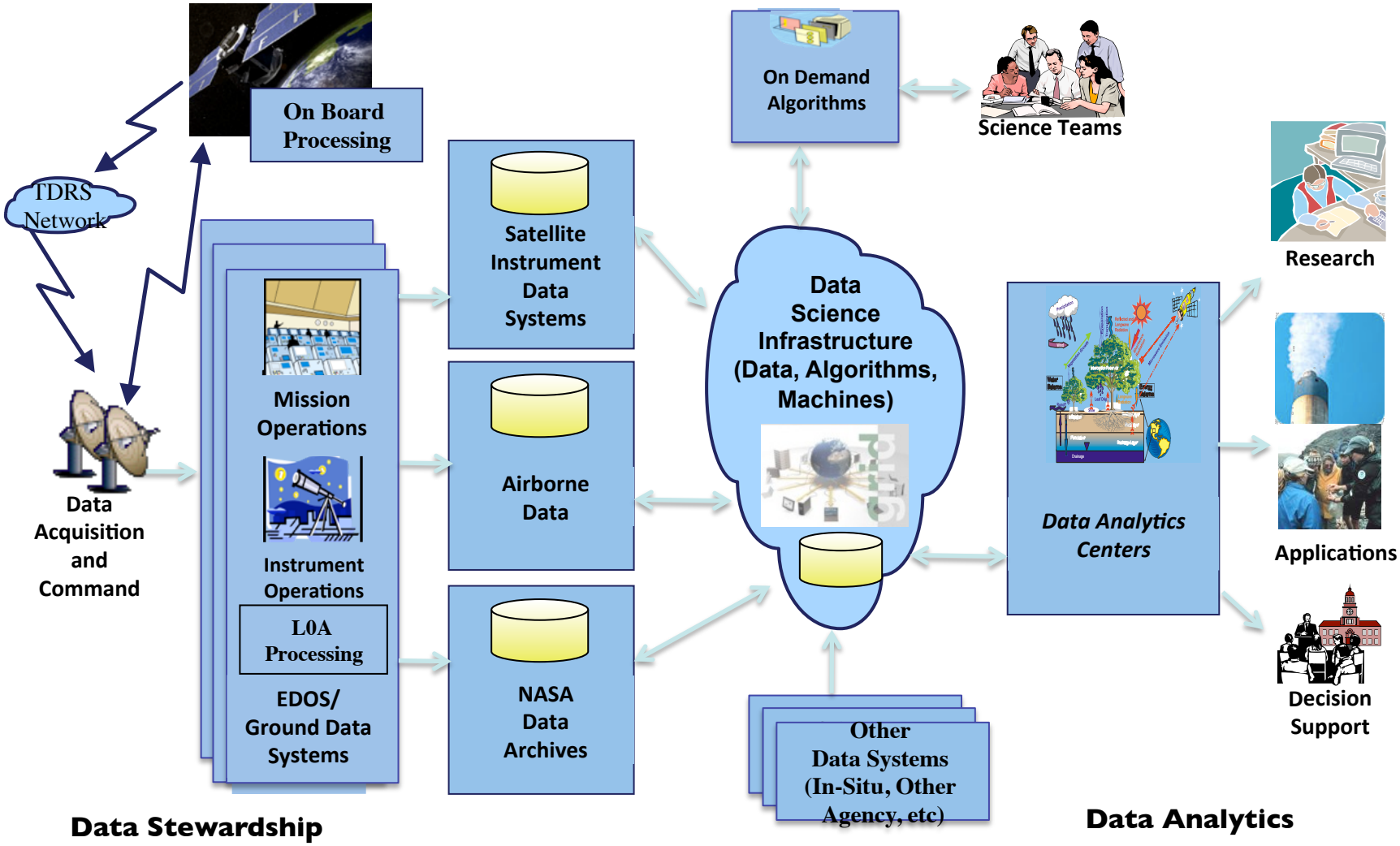
- Increasing capability on board
  - data reduction, planning, processing, etc, right at the sensor/ instrument
- Increasing data communications from space to ground
  - this could affect data processing pipelines and the need to scale the scale the archive (which includes not only storage, but also data movement challenges)
- Tradeoffs between static data pipelines and processing on the data itself
- Increasing data volumes: this could require new storage technologies, data movement, computation on the data, etc
- Analytics: intelligent algorithms in pipelines to derive automated metadata that can improve search; computation/data analytics co-located with data because it is massive? Methods to visualize massive data

Computational Capability Needs *

System	2015	2025	Application to Earth Science
Onboard	Limited onboard computation including data triage and data reduction. Investments in new flight computing technologies for extreme environments.	Increase onboard autonomy and enable large-scale data triage to support more capable instruments. Support reliable onboard processing in extreme environments to enable new exploration missions.	Onboard computation for airborne missions on aircraft; new flight computing capabilities deployed for extreme environments; use of data triage and reduction for high volume instruments on satellites.
Ground Systems	Rigid data processing pipelines; limited real-time event/feature detection. Support for 500 TB missions.	Increase computational processing capabilities for mission (100x); Enable ad hoc workflows and reduction of data; Enable realtime triage, event and feature detection. Support 100 PB scale missions.	Future mission computational challenges (e.g., NI-SAR); support more agile airborne campaigns; increase automated detection for massive data streams (e.g., automated tagging of data).
Archive Systems	Support for 10 PB of archival data; limited automated event and feature detection.	Support exascale archives; automated event and feature detection. Virtually integrated, distributed archives.	Turn archives into knowledge-bases to improve data discovery. Leverage massively scalable virtual data storage infrastructures.
Analytics	Limited analytics services; generally tightly coupled to DAACs; limited cross-archive, cross-agency integration; limited capabilities in data fusion; statistical uncertainty; provenance of the results	Analytics formalized as part of the mission-science lifecycle; Specialized Analytics Centers (separate from archives); Integrated data, HPC, algorithms across archives; Support for cross product data fusion; capture of statistical uncertainty; virtual missions.	Shift towards automated data analysis methods for massive data; integration of data across satellite, airborne, and ground-based sensors; systematic approaches to addressing uncertainty in scientific inferences; focus on answering specific science questions.

ESTO Big Data Study; Derived from OCT TA-11 Roadmap (2015)

Future: Enabling Scalable, Data-Intensive Science



- Towards the systematic analysis of massive data -

OGA Data Efforts

Agency	Big Data Overview and Strategy
National Institutes of Health (NIH)	The NIH has initiated a new program in Data Science and appointed an Associate Director to lead the effort. They are establishing an NIH Commons (computation, software, standards,etc) through its Big Data to Knowledge (BD2K) initiative. The NIH commons will provide capabilities to various NIH institutes who support directed research efforts. Efforts focus on enabling data management and big data analytics capabilities.
National Science Foundation (NSF)	The NSF has several initiatives coordinating through the Office of Cyberinfrastructure (OCI). OCI coordinates with various disciplines within the NSF including the EarthCube program that seeks to build a national Geosciences Cyberinfrastructure. Goals of the NSF include: 1) Derive knowledge from data; 2) Develop new cyberinfrastructures to manage, curate and serve data; 3) Develop new approaches for education and workforce development; and 4) Enable new types of inter-disciplinary collaboration, community building
DARPA	DARPA has several programs in Big Data including the XDATA and Memex Programs that are developing data science frameworks for big data analytics and mechanisms to explore deep searching of the Internet. DARPA is working to explore the use of open source technologies and their application to these programs.
NOAA	NOAA is working to explore commercial opportunities to build cyberinfrastructures. This includes the use of cloud-based computing capabilities and support to scale data management and computation. NOAA is also participating in the Big Earth Data Initiative (BEDI) project.
Department of Energy (DOE)	DOE has been exploring programs in extreme scale science, particularly as it relates to high performance computing (HPC). The goal is to address the combined challenges of Big Data and Big Compute to develop a exascale computing environment for simulation and data analysis at scale cutting across various disciplines in energy, biology and climate.
USGS	USGS is exploring its role in big data, focusing on data capture and integration, as well as sharing and leveraging HPC infrastructure across the agency. Programs such as EROS are exploring new architectural approaches to scale for the future.
National Institutes of Standards and Technology (NIST)	NIST has established a program in Data Science focusing on the development of architectures, use cases, standards and interoperability. They are also focused on areas including measurement foundations/principles to increase the accuracy of derived inferences from massive data.

PDS 2016-2021

Engineering Node

April 2015

PDS Mission and Vision

Mission

Facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data and documentation produced by or relevant to NASA's planetary missions, research programs, and data analysis programs.

Vision

- To gather and preserve the data obtained from exploration of the Solar System by the U.S.
- To facilitate new and exciting discoveries by providing access to and ensuring usability of those data to the worldwide community
- To inspire the public through availability and distribution of the body of knowledge reflected in the PDS data collection

PDS is a federation of distributed discipline and service nodes.

Level 1 Requirements*

1. PDS will provide expertise to guide and assist missions, programs, and individuals to organize and document digital data supporting NASA's goals in planetary science and solar system exploration

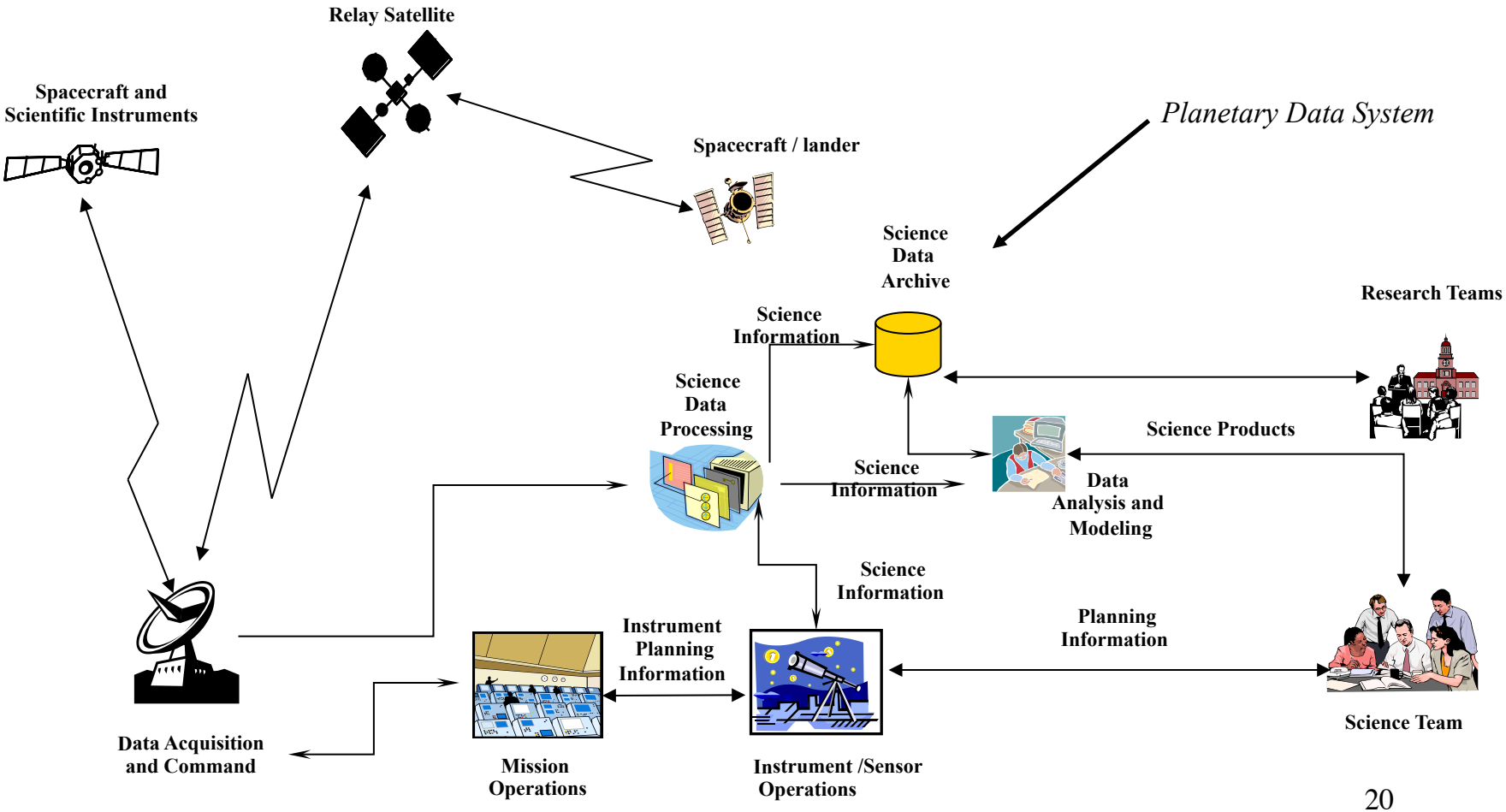
2. PDS will collect suitable and well-documented data into archives that are peer reviewed and maintained by members of the scientific community

3. PDS will make these data accessible to users seeking to achieve NASA's goals for exploration and science

4. PDS will ensure the long-term preservation of the data and their usability

** Level 1/2/3 requirements agreed by PDS MC*

PDS in Context



Credit: CCSDS Reference Architecture for Space Information Management

Key Drivers *

- More Data
- More Complexity (instruments, data)
- More Producer Interfaces
- Greater User Expectations
- Limited Funding
- Creating a system from the federation
- Internationalization
- Increasing IT security threats
- Failover capabilities across the PDS

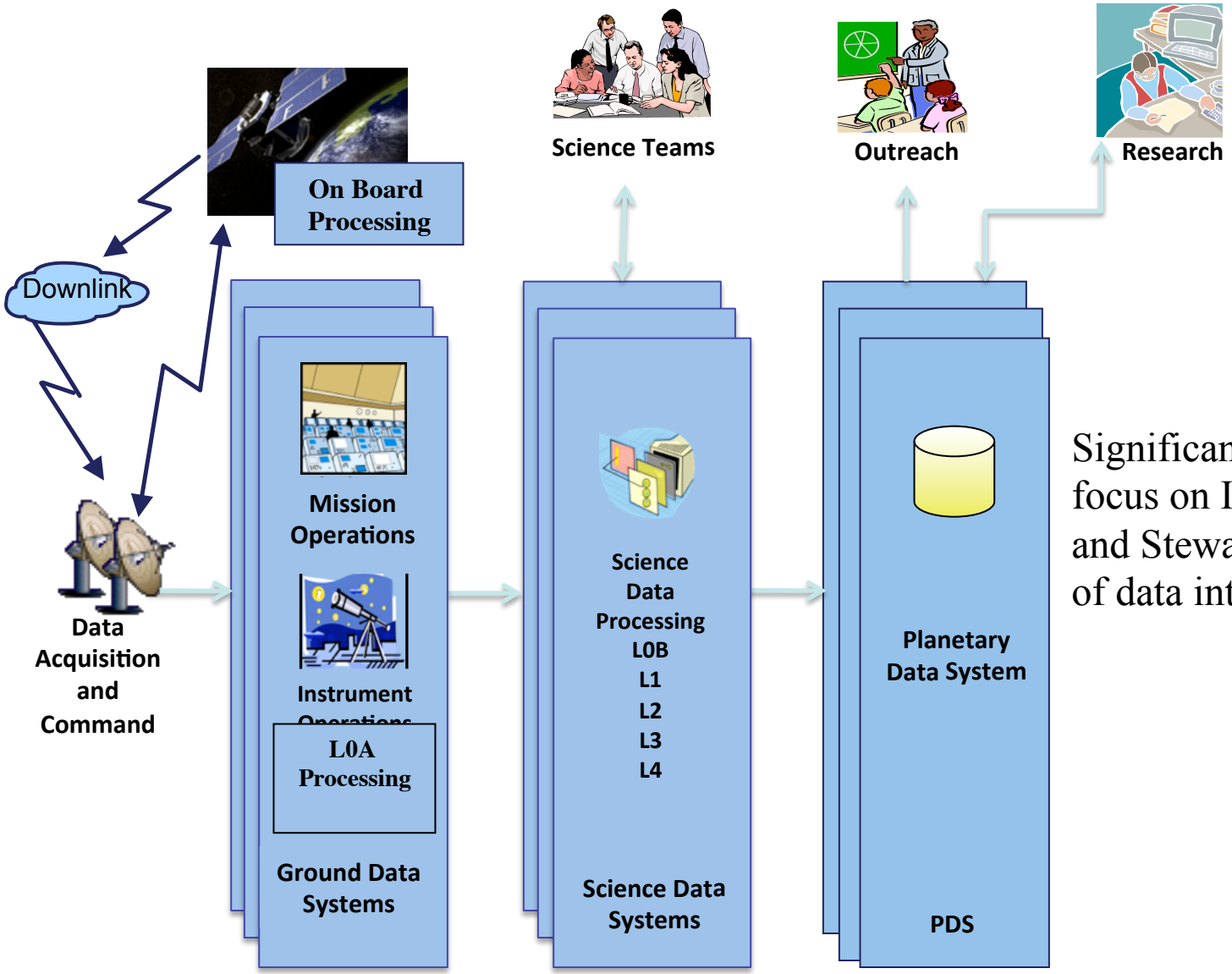
* Derived from PDS Study on Drivers for PDS4

"*Support the ongoing effort to evolve the Planetary Data System from an archiving facility to an effective online resource for the NASA and international communities.*" -- Planetary Science Decadal Survey, NRC, 2013-2022

PDS4: The Next Generation PDS

- PDS4 is a PDS-wide project to upgrade from PDS version 3 to version 4 to address many of these challenges
- An **explicit information architecture**
  - All PDS data tied to a common model to improve validation and discovery
  - Use of XML, a well-supported international standard, for data product labeling, validation, and searching.
  - A hierarchy of data dictionaries built to the ISO 11179 standard, designed to increase flexibility, enable complex searches, and make it easier to share data internationally.
- An **explicit software/technical architecture**
  - Distributed services both within PDS and at international partners
  - Consistent protocols for access to the data and services
  - Deployment of an open source registry infrastructure to track and manage every product in PDS
  - A distributed search infrastructure

PDS: Today

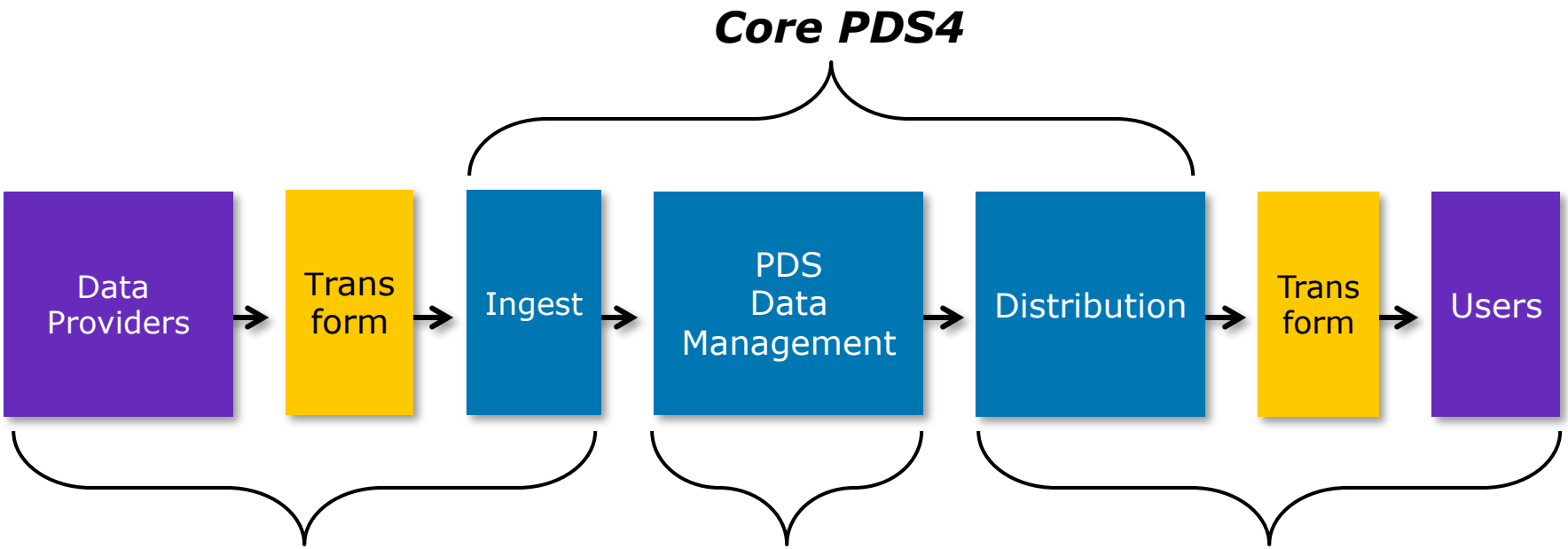


Significant focus on Ingestion and Stewardship of data into/in the PDS

PDS4 Capabilities: Today

- New information model captured in a modern modeling environment
- Model-driven software services (registry, search, etc)
- Core software tools and services
  - Heavy emphasis on data ingestion and validation
- Full deployment at Engineering Node
  - Integrated with PDS3 and PDS4
- Software installed at all nodes
- Integration underway at nodes
  - LADEE delivering under PDS4
- All new U.S. and International planetary missions using PDS4

Challenge: End-to-End System and Data Integration



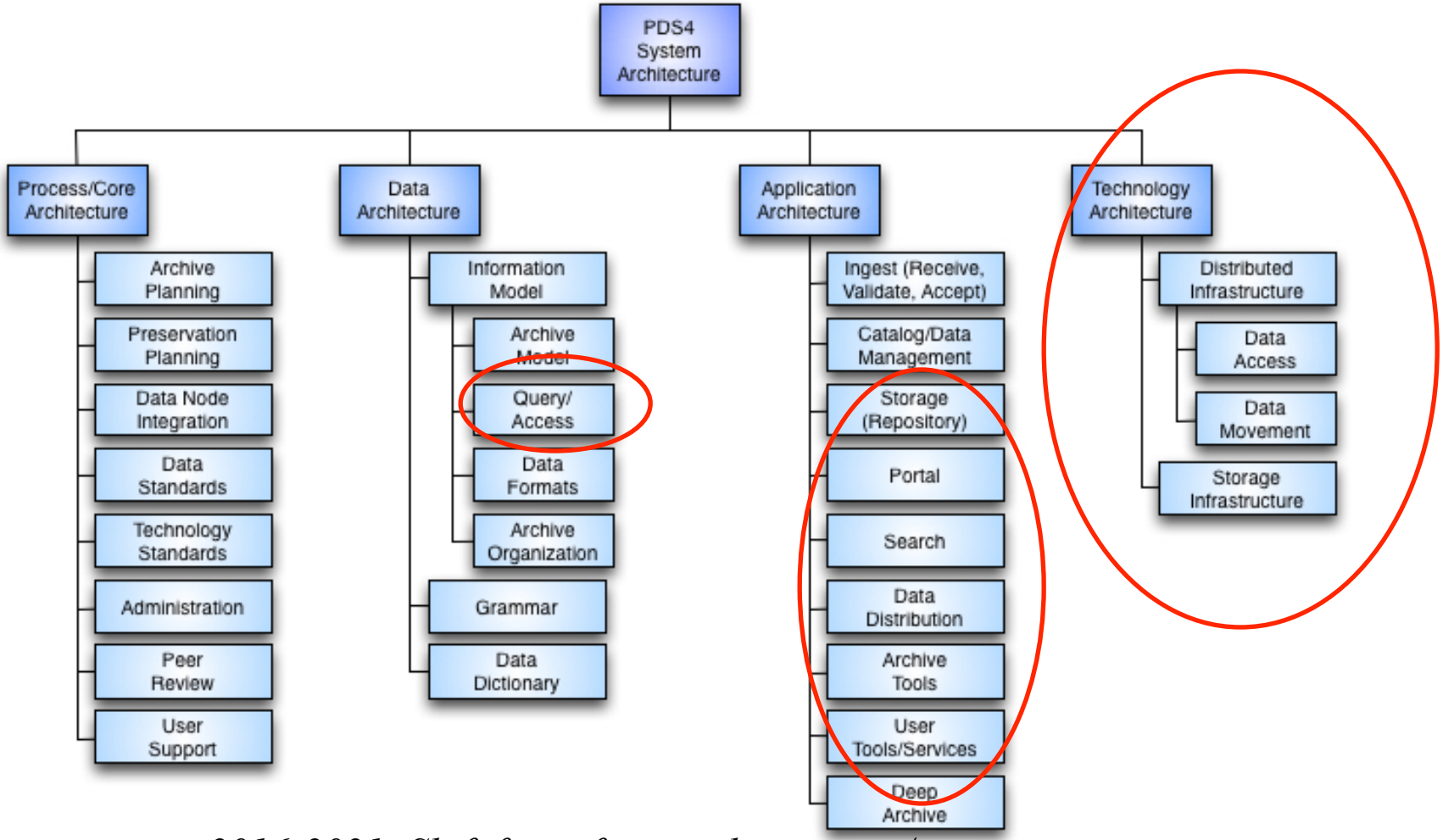
1) Improve efficiency and support to deliver high quality science products to PDS

2) Preserve and ensure the stability and integrity of PDS data

3) Improve user support and usability of the data in the archive

2010-2015: Implementation of common PDS4 information model, tools, and stewardship of data. ⇨ ***2016-2021: Increasing integration of data, tools, services across PDS***

PDS4 System Architecture Decomposition



2016-2021: Shift focus from archive ingest/management to data/node integration, search, distribution and discipline/user tools/support

PDS4 Technical/Software Roadmap

Function	2010-2015	2016-2021
Ingestion	Manual process for submission; tools based on PDS4 standards for design/validation	Automated ingestion/submission of data; include increased support for capturing mission information.
Data Management	Independent data management systems across PDS; initial PDS4 software installed and registration beginning.	Integrated data registries with PDS3 and PDS4 data across the PDS to allow for end-to-end tracking and search; interoperability with international partners.
Storage Management	Data stored online in independent storage repositories; backup/ failover unique at each node.	Virtualization/commodity storage services to increase integration and reduce cost; PDS-wide disaster recovery and failover in place.
Preservation Planning	Data maintained in a few simple formats	Transformation services to transform from archive formats to contemporary formats.
Distribution/ Access	Data distributed in archival format	Enhanced portal for access to data/services/tools; Data distributed in user formats; user services and tools to better facilitate and meet user analysis needs.

2010 to 2015: Focus on shifting missions and nodes to support PDS4
2016-2021: Future plans on integrating data, nodes, services, etc, together to improve user experience

PDS4 Information Architecture/Model Roadmap

Function	2010-2015	2016-2021
Data Model	Entire PDS model captured as an explicit model (ontology) defining all aspects including data, missions, instruments, etc	Expanded discipline node and mission models to provide increased capture of mission/science information and provide more tailored user support (search, tools, analysis, etc). Improve integration between software and model.
Data Dictionary	Captured using a rigorous, well-defined structure based on the ISO/IEC 11179 standard; elements organized into namespaces to allow for international coordination	Online data dictionary registries for mission and user use.
Grammar	Extensible Markup Language (XML) used to capture PDS metadata; Standard XML tools used	In addition to XML-based support, support for multiple standards for expressing the PDS model (RDF, JSON, etc) to increase use of information model in tools.

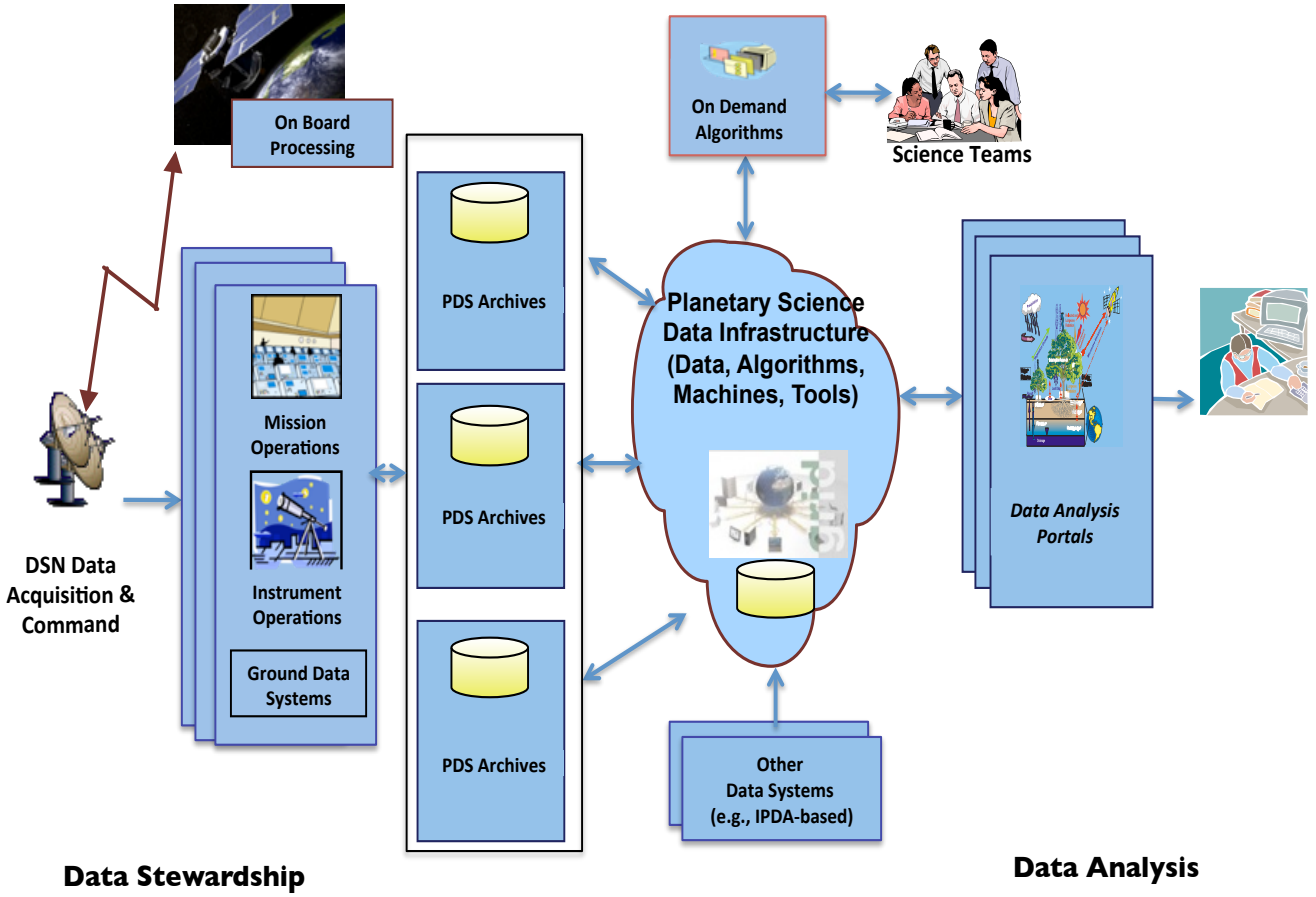
2010 to 2015: PDS4 IM stable and released with data dictionary and grammar.
2016-2021: Future plans in the information architecture/model focus on discipline node extensions to improve users support.

Proposed Goals (2016-2021)

1. Automate ingestion of data
   - Automated Design; improve label design
   - Automated Ingest (including harvesting)
   - Integrated registry, tracking
2. Capture well-formed, high quality PDS4 data collections
   - Data validated, harvested, registered, tracked
3. Shift Towards User Services
   - New portal design; PDS website as world-wide site for planetary data and tools
   - Search service deployed at every node
   - Leverage SOA-based architecture for new services, both for Engineering and DN
   - Every product discoverable and accessible
   - Ensure all data products can be visually inspected for peer review
   - *Registration and sharing of tools and data across planetary science community*
4. Establish Virtualization
   - 2nd copy of all data in storage service
   - Ensure PDS can fail over to secondary storage
   - Scaling online/computation services and support for data analysis
5. Upgrade legacy software tools to PDS4
   - Begin decommissioning PDS3 tools
   - Upgrade PDS3 tool to PDS4 ensuring backward compatibility to PDS3

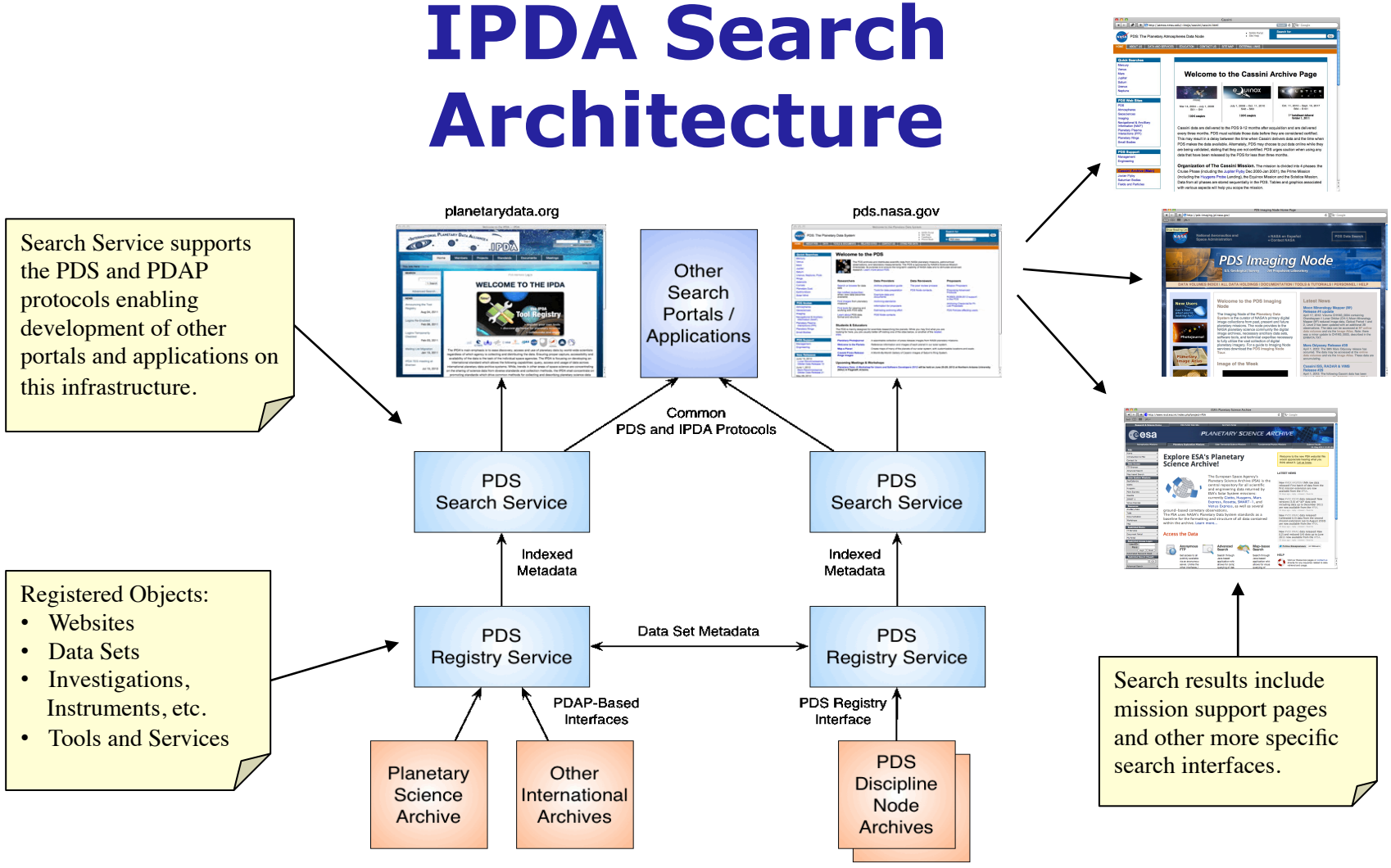
Future Vision

"An International Platform for Planetary Data Archiving, Management and Research"



"*Support the ongoing effort to evolve the Planetary Data System from an archiving facility to an effective online resource for the NASA and international communities.*" -- Planetary Science Decadal Survey, NRC, 2013-2022

IPDA Search Architecture



Conclusion

- PDS is well positioned to take advantage of PDS4 and to plan its next technology roadmap
  - The PDS4 architecture enables PDS to integrate the data and tools for the community
- The technology trends around “big data”, can help us scale and improve the user community
- There are opportunities to partner and leverage roadmaps

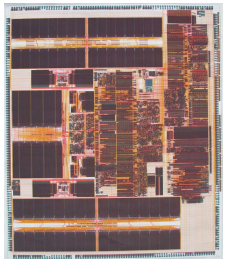
Backup

HPSC
High Performance Spaceflight Computing

PROBLEM / NEED BEING ADDRESSED

Space-based computing has not kept up with the needs of current / future NASA missions
NASA has extreme requirements for low power and energy management, and fault tolerance

STATUS QUO



RAD750
- *133 MHz*
- *200 MOPS*
- *200 Krad*
- *5 watts*

- *Power PC single core architecture*
- *Flown on most SMD missions*

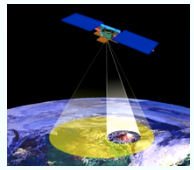
NEW INSIGHTS

- Rad Hard By Design (RHBD) methods realize >100x improved floating point performance and power efficiency over RAD750
- Multicore architectures
- Provide both general purpose and some DSP capability as well as interoperability with co-processors
- Are conducive to power scaling at core level and thread-based fault tolerance
- Allow flexible operation; dynamic trades for computational performance, energy management and fault tolerance

PROJECT DESCRIPTION/ APPROACH

- Issue a BAA for hardware architecture designs in FY13-14
  - Solicit flight computing system concepts
  - Prepare NASA requirements and benchmarks for evaluation of architectures
  - A competitive initial phase, seeking innovative solutions and early risk retirement
- Product of the investment
  - Multi-core hardware chip designs evaluated against NASA-provided benchmarks
- Development phase to follow in FY15-17
  - Multi-core hardware chip with bundled real-time operating system (RTOS), FSW development environment, and middleware elements, integrated on evaluation board
  - Middleware elements for allocating/managing cores for varying operational objectives, working closely with the FSW community, driven by knowledge of the NASA applications

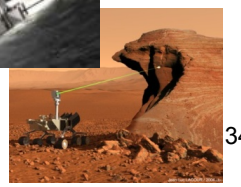
QUANTITATIVE IMPACT



- 100X performance of RAD750
- <7W power budget, scalable
- Support for a range of fault tolerance methods
- Interoperable with co-processors

PROJECT GOAL

- Deliver NASA's next-generation flight computing system
- Enabler for human-robotic teaming, onboard science data processing, real-time precision control, autonomous operations in uncertain environments



Capability Areas (1)

- Big Data Architectures – Big data architectures consider software, methodologies and data end-to-end and provide an integrated framework for implementing scalable, and cross-disciplinary solutions.

- Cyberinfrastructures – Scalable cyberinfrastructures are critical in capturing, generating, managing, and distributing massive data from remote sensing instruments and in supporting exploration and analysis across the data lifecycle for geographically distributed, multi-agency and multi-institutional environments.

Capability Areas (2)

- Statistics – Statistical science provides a theoretical foundation for the development of new methodologies to interrogate data and draw scientific inferences *with quantifiable uncertainties*

- Machine Learning – Machine learning (ML) provides automated approaches for detecting features and events of interest in data. These techniques have been applied successfully to ground-based and onboard science data to, e.g., detect and classify astronomical objects in sky surveys, detect volcanic eruptions, forest fires, sea ice break-up events, and the like from Earth-observing platforms, and detect and track dust devils on the surface of Mars

Capability Areas (3)

- Data Modeling/Information Architecture – The definition and relationships of data is critical for addressing discovery, integration, and other data science functions across the data lifecycle. These models of the data need to be in place, particularly as data assets are distributed.

- Visualization – Visualizing massive data is an increasingly critical part of the data science lifecycle. It is also a very active research area determining effective mechanisms for research and decision support

Capability Areas (4)

- Provenance – Reproducibility and integrity of results is a foundational consideration in science. In the Big Data era, scientists find themselves grappling with the difficult realities of 1) being able to utilize only vanishingly small slices of the available data, and 2) often pursuing highly exploratory, ad-hoc approaches to analysis. Provenance techniques that capture workflows and log other details of analytics sessions provide a basic capability, but ultimately a more sophisticated approach to provenance that extracts the relevant methodological steps for reproducibility (as opposed to the fully fine-grained messy meanderings of exploration) will be needed.