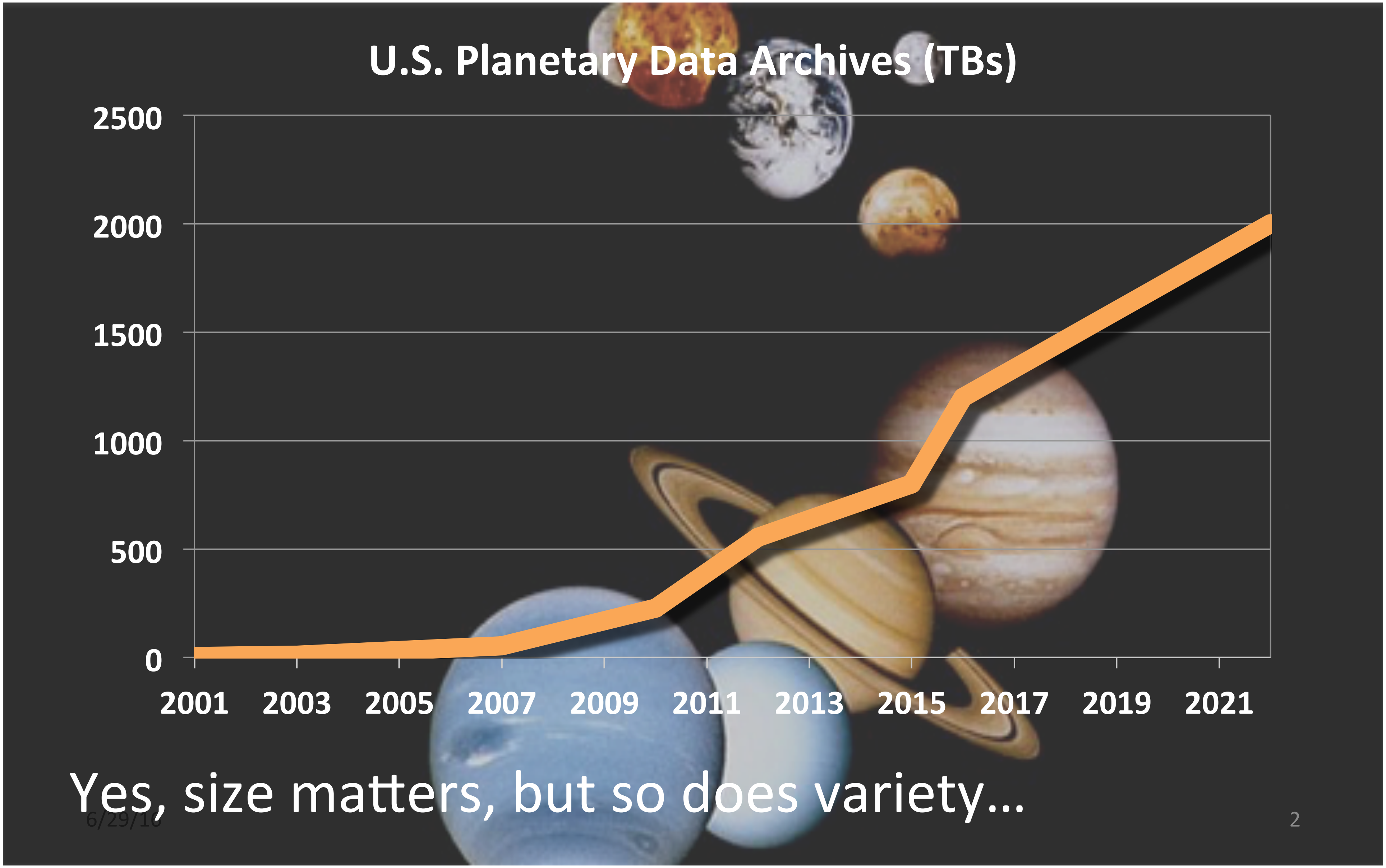
The Planetary Data System and “Big Data”

Ed Grayzeck

June 29 2016

Growth of Planetary Data Archived from U.S. Solar System Research



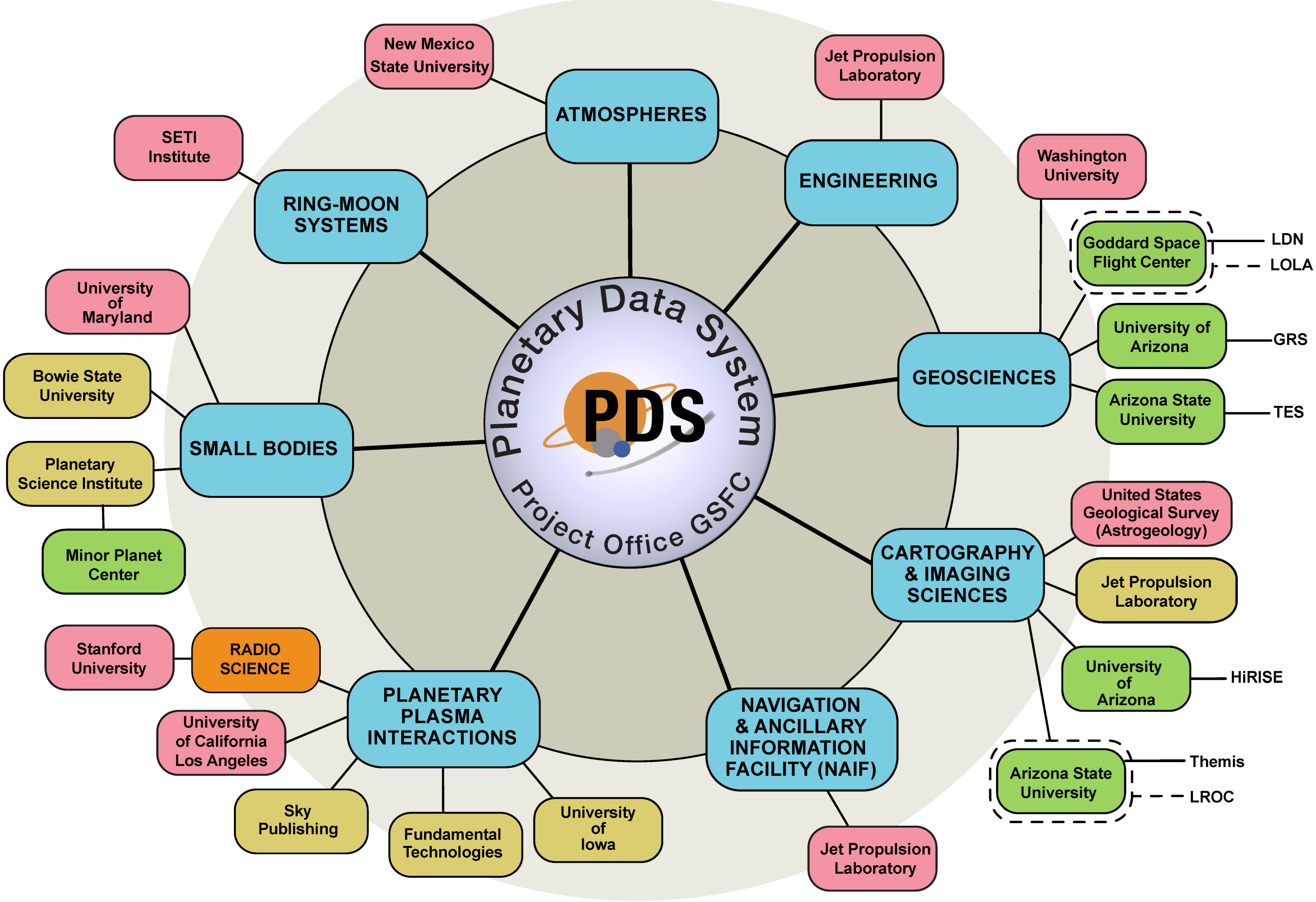
Yes, size matters, but so does variety…

Planetary Data System

- Purpose: Collect, archive and make accessible the digital data and documentation produced from NASA's exploration of the solar system from the 1960s to the present.
- Infrastructure: The federated system includes two technical support nodes and six science discipline nodes with sub-nodes as well as temporary data nodes often as part of mission archiving.
  - Diverse set of science disciplines
  - System driven by a well defined planetary science information model
  - Movement towards international interoperability

NODES/SUBNODES/DATA NODES

Function / Node Home Institution



Planetary Data System
PDS
Project Office GSFC

ATMOSPHERES
New Mexico State University

ENGINEERING
Jet Propulsion Laboratory

GEOSCIENCES
Washington University
Goddard Space Flight Center
LDN
LOLA
University of Arizona
GRS
Arizona State University
TES

CARTOGRAPHY & IMAGING SCIENCES
United States Geological Survey (Astrogeology)
Jet Propulsion Laboratory
University of Arizona
HiRISE
Arizona State University
Themis
LROC

NAVIGATION & ANCILLARY INFORMATION FACILITY (NAIF)
Jet Propulsion Laboratory

PLANETARY PLASMA INTERACTIONS
University of California Los Angeles
RADIO SCIENCE
Stanford University
Sky Publishing
Fundamental Technologies
University of Iowa

SMALL BODIES
University of Maryland
Bowie State University
Planetary Science Institute
Minor Planet Center

RING-MOON SYSTEMS
SETI Institute

Scale and Diversity of the PDS

Type of Data/Metadata	Distinct Entities
Data Sets	2151
Instrument Hosts	199
Instruments	625
Targets	4231
Missions/Investigations	71

- Total volume is currently ~1PB
- Represents 40M data products from 625 unique instruments
- The current MAVEN mission has a compliment of 8 diverse instruments with 300K data products at the current time
- Some missions have few instruments but many data products, e.g., LADEE

PDS Data Products - LADEE

- LADEE was complex and short lived mission with over 2M data products from 3 instruments but only a small volume
- Research includes comparison to Apollo era data (DTREM, LEAM, LACE, UVS) digitized from analog archives by NSSDCA
- Note: Mission included Lunar Laser Communication Demo that sent data from the Moon to Earth at 622 megabits per second or 1000-fold increase

(Some) Big Data Challenges in Planetary Science

- Variety of planetary science disciplines, moving targets, and data
- Volume of data returned from missions including provenance
- Federation of disciplines and international interoperability

- These factors can affect choices in:
  - Data Consistency
  - Data Storage
  - Computation
  - Movement of Data
  - Data Discovery
  - Data Distribution

Ultimately, having a planetary science information architectural strategy that can scale to support the size, distribution, and heterogeneity of the data is critical

PDS4: The Next Generation

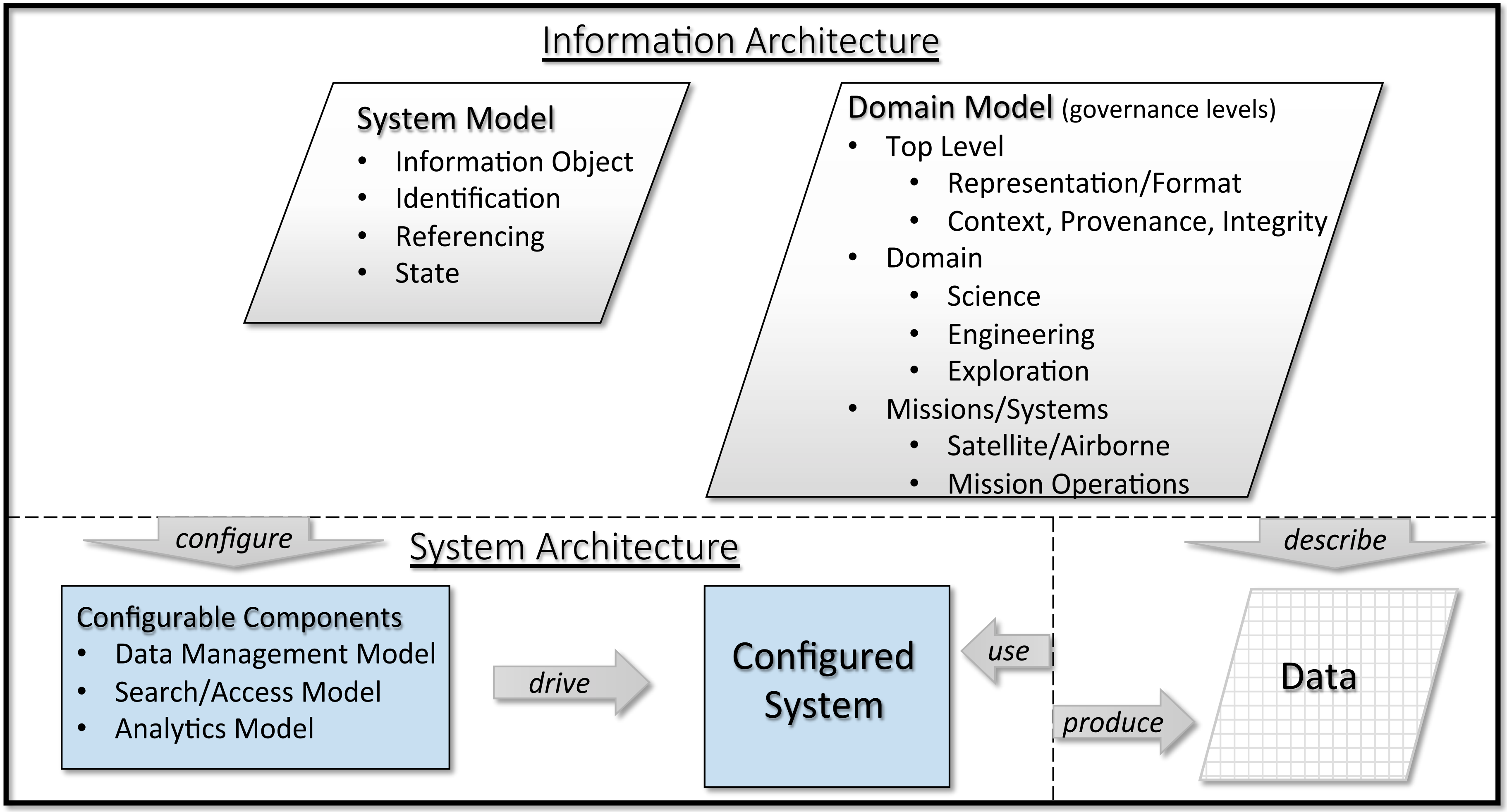
- PDS4 is a PDS-wide project to upgrade from PDS version 3 (PDS3) to address many of the big data challenges of a large-scale, distributed, international system
- An explicit information architecture
  - All products are tied to a common model for validation and discovery
  - Use of XML, a well-supported international standard, for labeling, validation, and searching
  - A hierarchy of dictionaries built to the ISO 11179 standard, designed to increase flexibility, enable complex searches, and make it easier to share data internationally
- Distributed services both within PDS and at international partners
  - 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
  - A distributed search infrastructure
  - Configured by the Information Architecture

PDS4 Information Model: Addressing Variety in Big Data Systems

- PDS4 Information Model plays a key role in defining the data and its relationships
  - Defines explicit relationships between major entities of the PDS
  - Establishes an overarching governance model for PDS data
- The PDS4 system is enabled by an “information model-driven” approach where the information model is the corner-stone of the system
  - Handles the diversity of different disciplines
  - Enables federated governance
  - New instruments, observation types and data can be accommodated
  - Allows the system to be configured by the information model
  - Ensures updates to the model do not break the software
  - Provides metadata definitions that are tied to the model to increase consistency

Model-Driven PDS

Information System Architecture



Information Architecture

System Model
- Information Object
- Identification
- Referencing
- State

Domain Model (governance levels)
- Top Level
  - Representation/Format
  - Context, Provenance, Integrity
- Domain
  - Science
  - Engineering
  - Exploration
- Missions/Systems
  - Satellite/Airborne
  - Mission Operations

System Architecture

Configurable Components
- Data Management Model
- Search/Access Model
- Analytics Model

Configured System

Data

Crichton, D. Hughes, J.S. ; Hardman, S. ; Law, E. ; Beebe, R. ; Morgan, T.; Grayzeck, E.
A Scalable Planetary Science Information Architecture for Big Science Data.
IEEE 10th International Conference on e-Science, October 2014.

Characteristics of the PDS4 IM

- Multiple disciplines (Atmospheres, Geosciences, Plasma, etc) supported
- Multi-level governance enabled (independent extensions)
- Multiple models integrated into an overarching ontology
  - A core model that describe the missions, instruments, targets, observations, etc
  - Models that describe disciplines
  - Models for registries, data dictionaries, etc
- Active Data Design Working Group to accommodate updates
- Maintained by a Change Control Board with representatives both across the Planetary Data System and Internationally

Software and Tool Collaborations

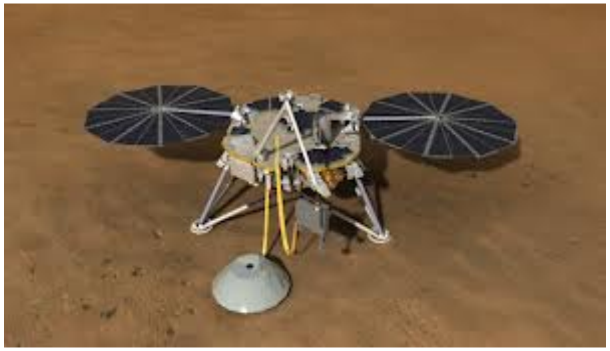
- PDS4 is enabled by a set of core software services for registration, search, and distribution
  - Major open source software products used for registration, search, and distribution
- PDS-wide tools are provided for design, validation, and transformation of PDS4 data products
  - Use of XML provides significant leveraging for using common libraries
- Regular software builds and releases integrate software and information model, and released for use by data providers, nodes, and international partners
- Each node builds search and support services tailored for their community; inventory has 20 such tools to support PDS3 and PDS4
- PDS has an increasing desire to distribute software via open source channels
  - Continue to look for avenues to increase coordination and collaboration in tool and software development

International Planetary Data Alliance

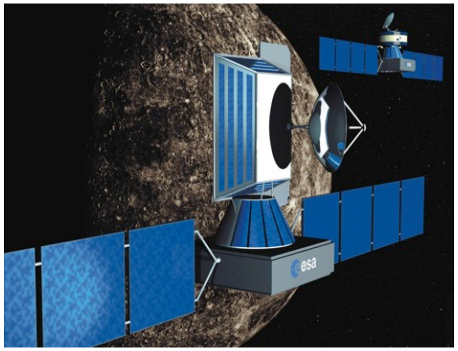
- Founded in 2006
  - Resulted from meeting between the ESA Planetary Science Archive and the PDS at ESAC
- Includes all major space agencies involved in planetary science data archiving
- Mission is to build compatible, international planetary data archives for the purpose of interoperability
- Major investment and buy-in in PDS4
  - Leveraging both the PDS4 Information Model and core software tools and services

PDS4: Support for an Era of US and International Missions

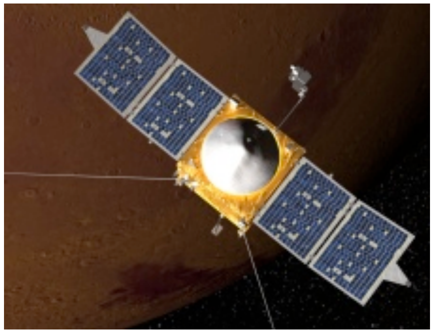
LADEE (NASA)



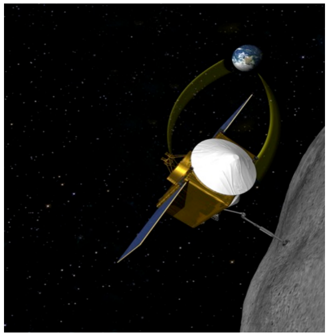
InSight (NASA)



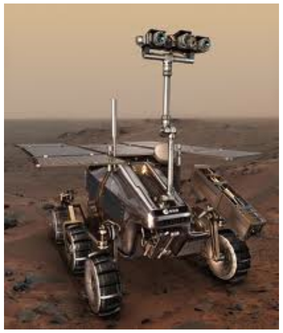
BepiColumbo (ESA/JAXA)



MAVEN (NASA)



OSIRIS-REx (NASA)



ExoMars (ESA/Russia)

JUICE (ESA)

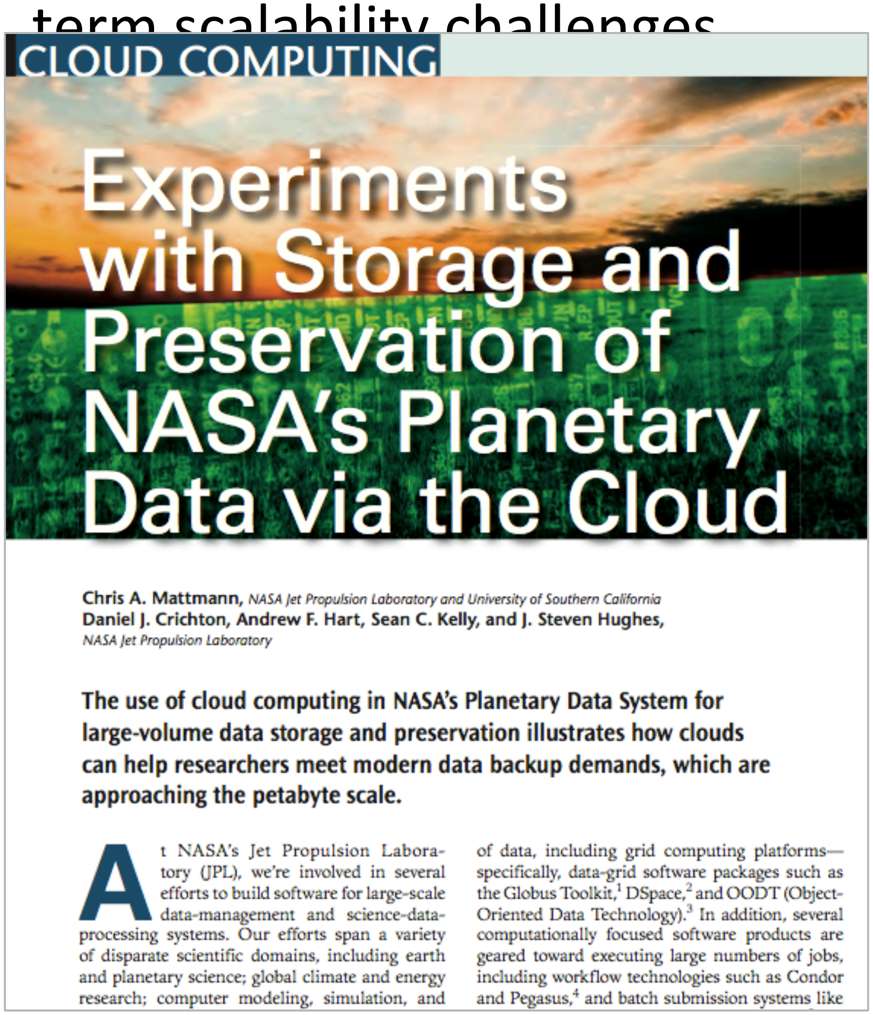
…also Hayabusa-2, Chandrayaan-2, Mars 2020…

Endorsed by the **International Planetary Data Alliance** in July 2012:
https://planetarydata.org/documents/steering-committee/ipda-endorsements-recommendations-and-actions

The Planetary Cloud Experiment

- Can fit into the PDS4 architecture
- Data movement challenges can be an issue (e.g., data to/from cloud)
- Different clouds (Amazon, Azure, Hybrid, …) tested as a secondary storage option for large data, e.g., HIRISE images
- EN has procurement path to AWS S3
- Long-term costs remain a concern since downloads are not constrained

- Focus on addressing long-term scalability challenges



CLOUD COMPUTING

Experiments with Storage and Preservation of NASA's Planetary Data via the Cloud

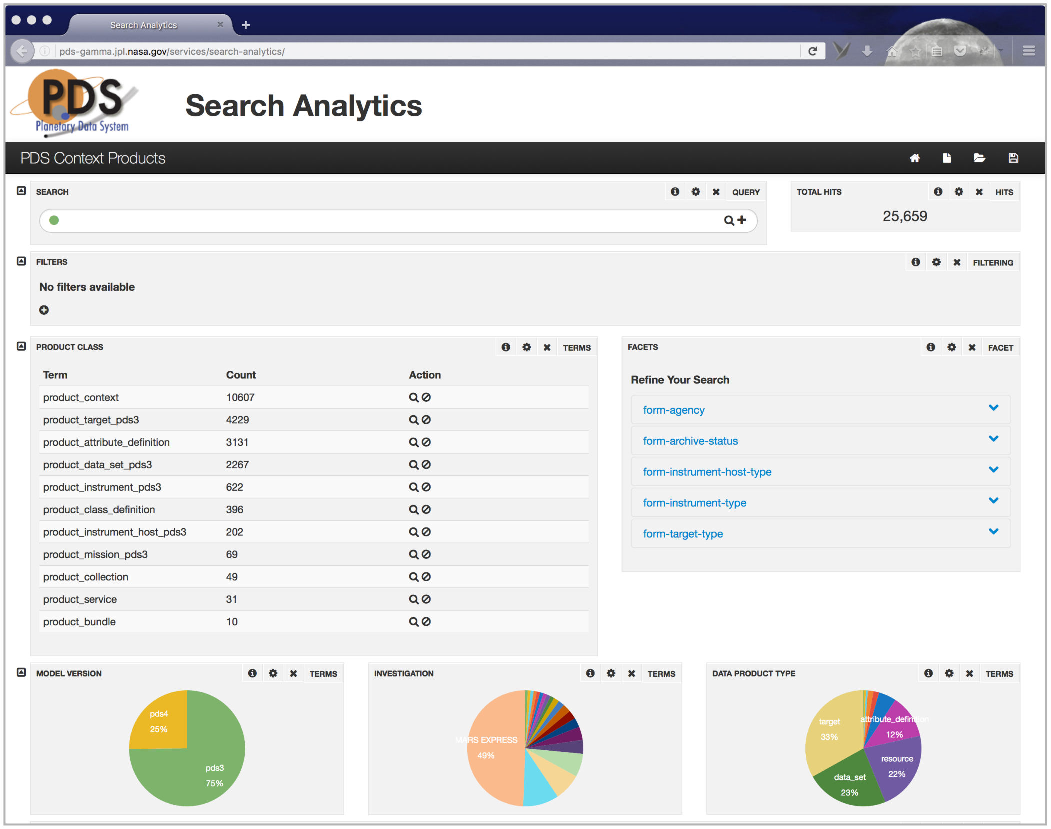
Chris A. Mattmann, *NASA Jet Propulsion Laboratory and University of Southern California*
Daniel J. Crichton, Andrew F. Hart, Sean C. Kelly, and J. Steven Hughes, *NASA Jet Propulsion Laboratory*

The use of cloud computing in NASA's Planetary Data System for large-volume data storage and preservation illustrates how clouds can help researchers meet modern data backup demands, which are approaching the petabyte scale.

At NASA's Jet Propulsion Laboratory (JPL), we're involved in several efforts to build software for large-scale data-management and science-data-processing systems. Our efforts span a variety of disparate scientific domains, including earth and planetary science; global climate and energy research; computer modeling, simulation, and of data, including grid computing platforms—specifically, data-grid software packages such as the Globus Toolkit,[1] DSpace,[2] and OODT (Object-Oriented Data Technology).[3] In addition, several computationally focused software products are geared toward executing large numbers of jobs, including workflow technologies such as Condor and Pegasus,[4] and batch submission systems like

Using Analytics to Understand PDS Data Trends

- Use of open source software and XML coupled with the PDS4 Information Model enables opportunities to explore PDS data holdings
  - Data Classification (missions, instruments, targets, etc)
  - Trend Analysis



Search Analytics

PDS Context Products

SEARCH

TOTAL HITS

25,659

FILTERS

No filters available

PRODUCT CLASS

Term	Count	Action
product_context	10607	🔍⊘
product_target_pds3	4229	🔍⊘
product_attribute_definition	3131	🔍⊘
product_data_set_pds3	2267	🔍⊘
product_instrument_pds3	622	🔍⊘
product_class_definition	396	🔍⊘
product_instrument_host_pds3	202	🔍⊘
product_mission_pds3	69	🔍⊘
product_collection	49	🔍⊘
product_service	31	🔍⊘
product_bundle	10	🔍⊘

FACETS

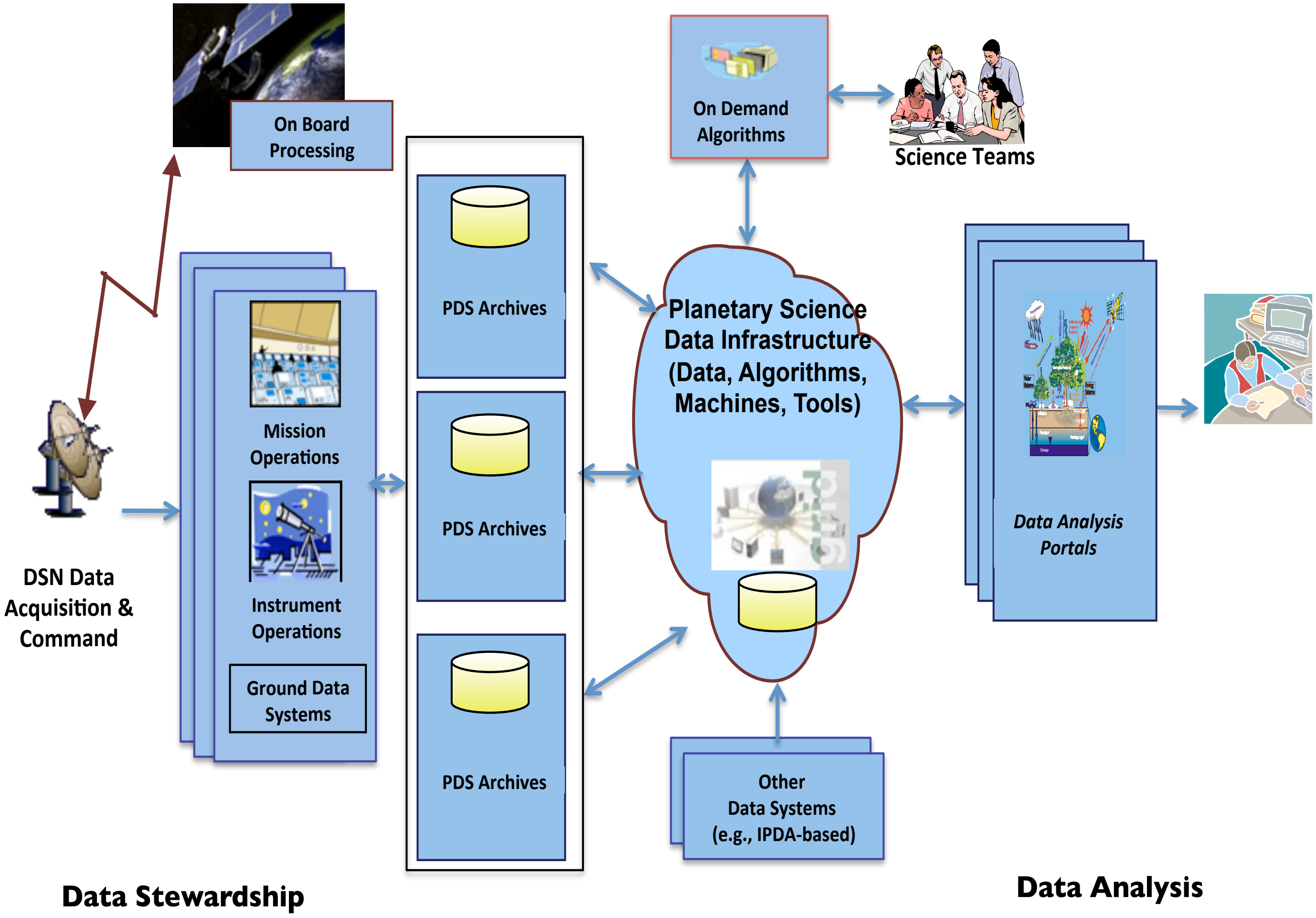
Refine Your Search

MODEL VERSION

INVESTIGATION

DATA PRODUCT TYPE

Towards 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

PDS community Roadmap Update

- PDS Roadmap process to outline scope for next 10 years
- Identified community members through meeting workshops, self nomination, and direct solicitation of expert help
- Initial meeting summer of 2016 with final draft due summer of 2017
- Identify areas of improvement such as mission pipelines, search capabilities, tool improvement, and metrics of node and system

Next Step: 2016 IEEE International Big Data Conference

- Workshop on Big Data Challenges, Research, and Technologies in the Earth and Planetary Sciences
- Location: Washington DC, December 5-9, 2016
- Topics: Architectures, Onboard/Sensor-based Computing, Scalable Data Analytics for Massively Distributed Data
- http://geo-bigdata.github.io
- Follows two successful workshops in 2015
- Workshop Chairs: Dan Crichton (JPL), Tom Narock (Marymount University)

PDS Big Data Presentations

Upcoming meetings

- IPDA (July 27-29)
- COSPAR? (July 30-August 7)
- OAGS (July 31-August 5)
- Planetary Interoperability workshop at DPS
- DPS exhibit area with focus on IPDA and SPICE
- 2016 AGU special session (IN023)
- LPSC workshop
- Planetary Data Workshop (tools, PDS4)

Background

BDTF Questions

Planning for the future

- Community based roadmap looks out 10 years
- Each Discipline node has an assessment group
- Priority set by PDS Management Council

What feature could be stopped

- PDS3 tool could be deprecated once legacy missions and node data migration are complete
- Would need mission and community input
- Missions keep getting extended

BDTF Questions continued

What steps to make data interoperable - NASA

- Work with SPDF to define CDF/A for PDS4
- Successful usage to describe and archive MAVEN data
- Project with MAST to provide pointers to HST data

What steps to make data interoperable – non-NASA

- Founding and active member of IPDA
- Share tools for PDS3 and PDS4 with IPDA
- ESA/PSA built on PDS standards
- Direct searching of PSA data
- ESA a voting member of CCB