Advanced Information Systems Big Data Study for Earth Science

Daniel Crichton, NASA Jet Propulsion Laboratory
Michael Little, NASA Headquarters

October 29, 2015

Background

- NASA has historically focused on systematic capture and stewardship of data for observational systems
  - Limited use of advance computational technologies to support scientific inferences

- Increasing “big data” era is driving needs to
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computational and data science across the lifecycle

- NASA Advanced Information Systems Technology (AIST) program initiated a study of needed data and computational science techniques across the data lifecycle and have made some key recommendations
  - Leverages the NASA Office of the Chief Technologist Roadmap for Modeling, Simulation and Information Technology (2015)

Data and Computational Science Across the Data Lifecycle

- Architectural considerations/tradeoffs for integrating the entire data lifecycle

- Onboard
  - Enable data reduction and triage close to the sensor/instrument
  - Manage bandwidth for communicating results

- Scalable Data Management
  - Capturing well-architected and curated data repositories based on well-defined data/information architectures
  - Architecting automated pipelines for data capture

- Scalable Data Analytics
  - Access and integration of highly distributed, heterogeneous data
  - Novel statistical approaches for data integration and fusion
    - Including sampling strategies
  - Computation applied at the data sources
  - Algorithms for identifying and extracting interesting features and patterns

Jet Propulsion Laboratory
California Institute of Technology

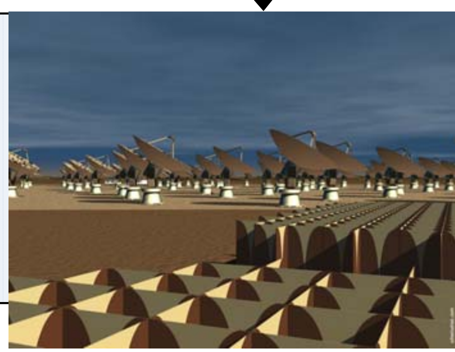
Data Lifecycle Model
for NASA Space Missions

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

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



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

Ground-based Mission Systems

Massive Data Archives and Big Data Analytics

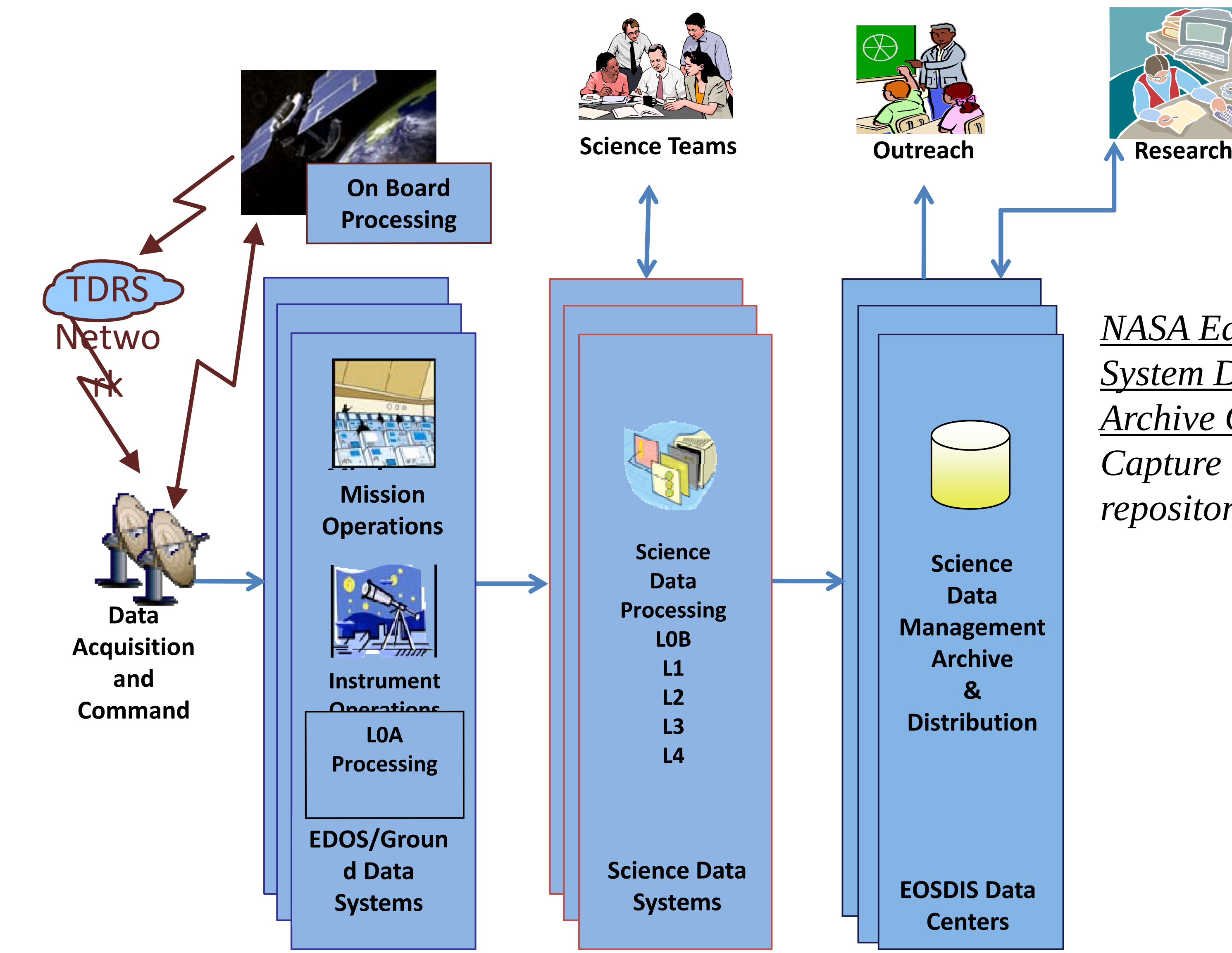
Emerging Solutions
- *Distributed Data Analytics*
- *Advanced Data Science Methods*
- *Scalable Computation and Storage*

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

Concept included in OCT TA-11 Roadmap (2015): 11.4.1 – Mission, Science, Engineering Data Lifecycle

Jet Propulsion Laboratory
California Institute of Technology

NASA Earth Science Data Pipeline Today: Constructing Scientific Archives



NASA Earth Observation System Distributed Active Archive Centers

Capture of well-curated repositories > 10 PB

Emerging Challenges as Data Increases

- Reproducibility
- Uncertainty management
- Data fusion (including distributed data)
- Data reduction
- Data movement
- Data visualization
- Cost
- Performance

Jet Propulsion Laboratory
California Institute of Technology

Driving Use Cases

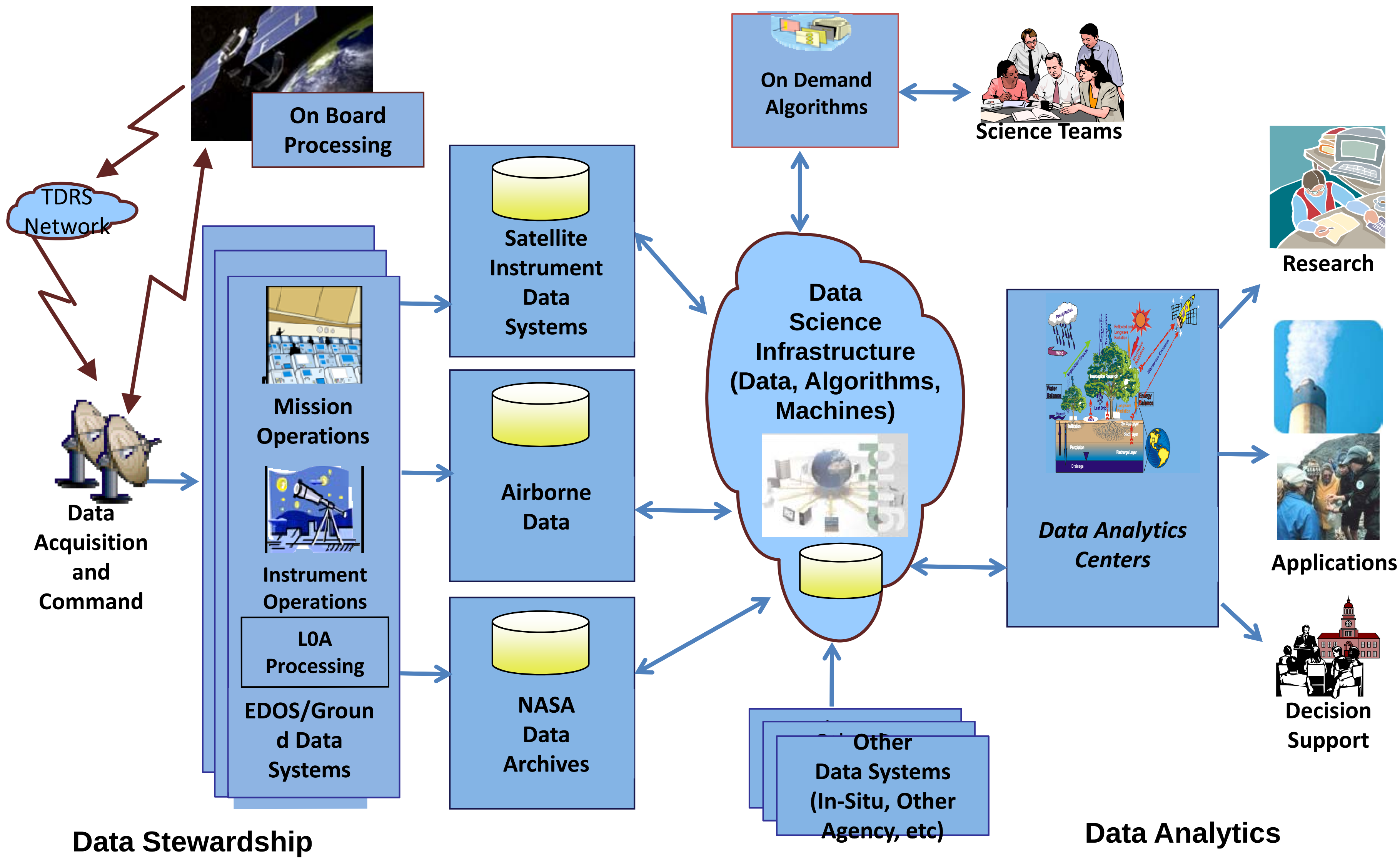
Use Case	Description	Data Science Challenge	Enabling Mission/ Capability
Climate Modeling	Formulate hypotheses from observed empirical relationships; Simulate current and past conditions under those hypotheses using climate models; Test hypotheses by comparing simulations to observations; Evaluate uncertainty of predictions originated from statistical sampling of models and observations.	Highly distributed data sources; fusion of different observations; moving computation to the data; data reduction	CMIP6 will move towards exascale archives requiring new approaches to evaluating models relative to observational data.
Satellite Missions	Missions such as NI-SAR and SWOT will generate massive observational data. However, they are have different architectural patterns including compute intensive, data intensive, heterogeneous, etc.	Massive data rates, data movement challenges, computational scalability, archiving and distribution; onboard processing for data reduction/analysis; high-volume data transfer for ground processing	NI-SAR and SWOT require new approaches for computation, data movement, data archiving and distribution, analytics.
Applications - Hydrology (Central Valley of California)	Understanding groundwater dynamics on a regional scale using measurements from satellite, airborne and in-situ measurements. Compare against predictive models.	Distributed computation; highly distributed data sources; data fusion of multiple products; massive new satellite observations.	Integration of data from PALSAR-2, Sentinel, Grace-FO, ASO, and SMAP. Scale to support NI-SAR and SWOT. Comparison against models. Requires new architectural approaches for distributed data analytics. .
Airborne Missions	Airborne missions tend to be much more agile and on-demand. Integrating this into a data ecosystem provides new opportunities to quickly generate and understand various measurements.	On-demand architectures; distributed data sources; on-the-fly data processing; onboard processing for data reduction/analysis; high-volume data transfer for ground processing	Current missions such as CARVE and Airborne Snow Observatory; Future such as proposed EVI-3 and ASO follow-on missions

Limitations and Gaps Across the Space Data Lifecycle

- Flight Computing, Ground Systems, Archiving/Distribution, and Analytics are not architected into a scalable big data system…

- Problems across the data lifecycle:
  - Data Generation: Limited onboard computing (or computing at sensor) for planning
  - Data Triage: Limited onboard triage and processing
  - Data Compression: Limited intelligent data reduction
  - Data Transport: Dependent on bandwidth capabilities; challanges in moving and distributing massive data
  - Data Processing: Ground systems and ground processing have limited support for dynamic workflows, scaling to large-scale environments (clouds, HPC), integrating intelligent discovery algorithms, etc. Processing disconnected from science analysis.
  - Data Archiving: Scaling the capture, management and distribution of data; distributed archives; limited computational capabilities; different models, formats, representations of data.
  - Visualization: Limited visualization capabilities for massive data; challenges in presenting massive data to users
  - Data 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

Future: Enabling Scalable, Data-Intensive Science



- Towards the systematic analysis of massive data -

Computational Capability Needs and Gaps Across Lifecycle*

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.

Derived from OCT TA-11 Roadmap (2015)

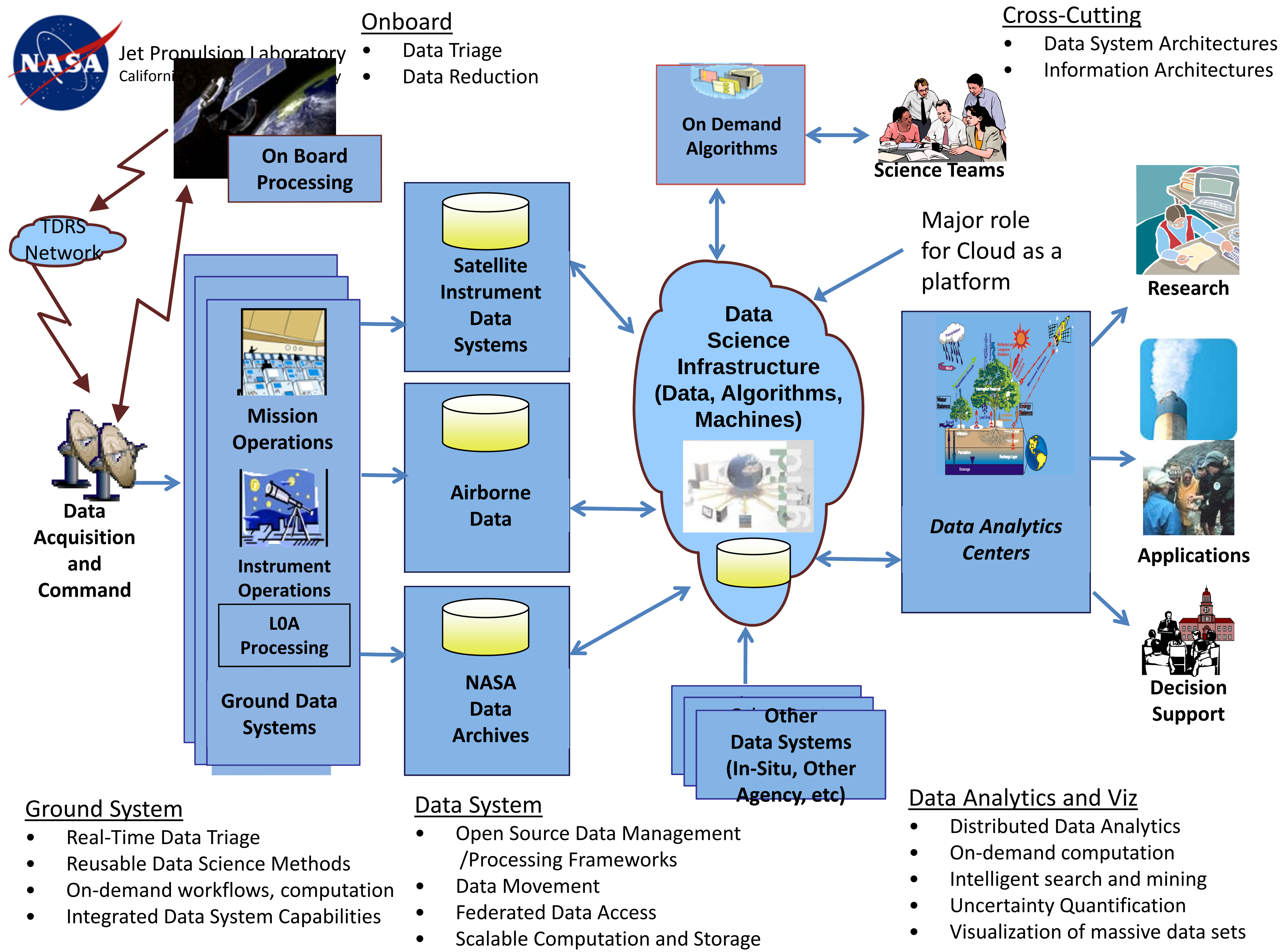
Proposed Technology Areas

Technology Name	Data Lifecycle Area (s)	Description
Big Data Architecture Earth Science Remote Sensing	Cross-Cutting	Definition of a scalable data big data lifecycle architecture for earth observing systems identifying how Big Data can scale from onboard computing to data analysis to increase science yield.
Big Data Information Models and Semantics	Cross-Cutting	Advanced semantic technologies for defining, deriving, and integrating heterogeneous ontologies and information models as applied across the entire data lifecycle (onboard, ground-based operations, archives, analysis)
Onboard data science methods for data triage	Data triage	Onboard data science methods for real-time event detection, and planning.
Onboard data science methods for data reduction	Data Compression	Onboard data science methods for data reduction.
Massive Data Movement Technologies	Data Transport	Massive data movement technologies for ground-based networks from operations through analysis
Real-time ground-based data science methods	Data Processing	Real-time ground-based data science methods for data reduction and real-time event detection for massive data streams as part of the data lifecycle architecture.
Open source data processing frameworks	Data Processing	Open source data processing and workflow frameworks that can massively scale to computational infrastructures (HPC, public cloud, etc.) handling large data streams, products, including near-real time constraints, as part of the data lifecycle architecture.
Reusable data science methodologies for missions and science	(1) Data Processing; (2) Data Analytics	Development of reusable data science methodologies for analysis of data on the ground as part of the data lifecycle architecture. This includes on-demand data analytics for massive data repositories.

Jet Propulsion Laboratory
California Institute of Technology

Proposed Technology Areas (2)

Federated data access	Data Archives	Federation of data access from distributed repositories as part of the data lifecycle architecture, moving towards on-demand distributed data analytics
Massive Data Distribution	Data Archives	Massive data distribution for large-scale repositories and archives including methods for data reduction, computation, etc., as integrated, on-demand data analytics.
Intelligent search and mining	(1) Data Archives; (2) Data Analytics	Provide methods for intelligent search and mining of massive data. This may include integration of on-demand analytics to perform deep searches.
Visualization of massive data sets	Visualization	Visualization of massive data sets including data reduction methods that are driven by domain.
On-demand distributed data analytics	Data Analytics	On-demand data analytics that can integrate data from archives, repositories, etc., applying data science methods (data reduction, fusion, feature detection, etc.) provided through a computational infrastructure
Distributed data analytics	Data Analytics	Analysis of data across distributed archives to support Earth system science
Uncertainty Quantification; Measurement Science	Data Analytics	Management of uncertainty in scientific inferences as part of a measurement science strategy for data fusion and data science
Open source data management/science frameworks	(1) Data Archives; (2) Data Analytics	Open source data management/science frameworks that can massively scale to handle and manage large data streams, products, including near-real time constraints, as part of the data lifecycle architecture, for archiving and analytics as part of a big data cyber-infrastructure.
Computational Infrastructures	(1) Data Processing; (2) Data Analytics	Computational Infrastructures to scale data analytics using HPC and public cloud. This includes on-demand massive HPC and storage for integration to drive analytics.


Jet Propulsion Laboratory
Californi
Onboard
• Data Triage
• Data Reduction
On Board Processing
TDRS Network
On Demand Algorithms
Science Teams
Cross-Cutting
• Data System Architectures
• Information Architectures
Major role for Cloud as a platform
Research
Satellite Instrument Data Systems
Mission Operations
Instrument Operations
L0A Processing
Ground Data Systems
Data Acquisition and Command
Airborne Data
NASA Data Archives
Data Science Infrastructure (Data, Algorithms, Machines)
Data Analytics Centers
Applications
Decision Support
Other Data Systems (In-Situ, Other Agency, etc)
Ground System
• Real-Time Data Triage
• Reusable Data Science Methods
• On-demand workflows, computation
• Integrated Data System Capabilities
Data System
• Open Source Data Management /Processing Frameworks
• Data Movement
• Federated Data Access
• Scalable Computation and Storage
Data Analytics and Viz
• Distributed Data Analytics
• On-demand computation
• Intelligent search and mining
• Uncertainty Quantification
• Visualization of massive data sets

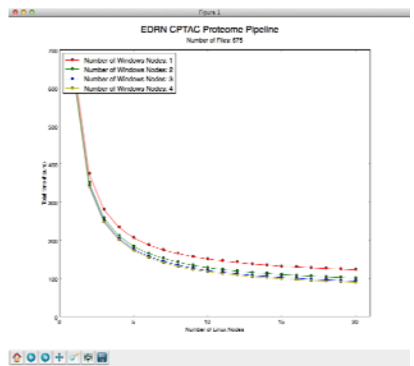
Cloud Computing: Enabling the Data Ecosystem

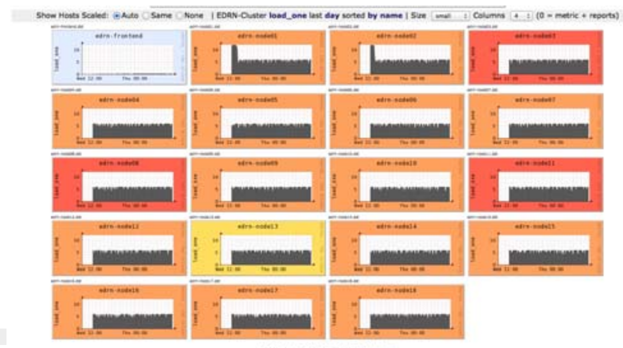
- A platform for Data and Computational Science
  - Ground systems
  - Archive systems
  - Data analytics
- Delivering data and computational services
  - APIs for data access
  - PGEs for data processing
  - Algorithms for data integration within and across systems
  - Algorithms for reduction, classification, event detection, etc
- Scalability on-demand

Selecting Cloud Topologies for Scalability

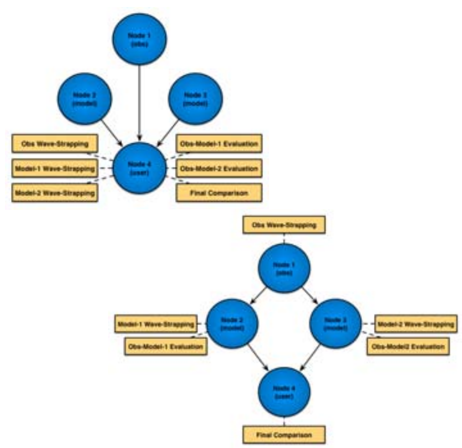
- DAWN (Distributed Analytics, Workflows and Numeric) is a model for simulation and optimization of system architectures for intensive data processing
- Particularly suited to analyze the deployment of a processing pipeline on the Cloud:
  - Can predict application performance as a function of allocated Cloud resources
  - Can "score" different Cloud topologies (for same resources) based on performance

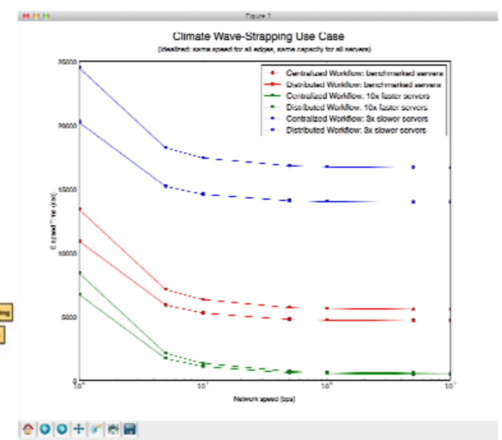
<u>EDRN Example</u>: duration of CPTAC data processing pipeline versus number of processing nodes
-> <u>conclusion</u>: allocate 18 nodes, no more gain after that





<u>Climate Example</u>: centralized vs distributed architecture for comparing models and observations as a function of network speed -> <u>conclusion</u>: distributed architecture is more efficient, more so for slower network and less powerful servers





Key Recommendations

- Shift from ad hoc investments across the mission and science data lifecycle to an *integrated architecture* where technology investments fit into a broader capability to enable earth system science.
  - Big Data Architectures should be *modeled and assessed* overall to address and plan technology capabilities and improvements to ensure that architectural support for science activities can scale and meet performance, cost, and uncertainty goals.
  - Architectures should enable *flexible and transparent tradeoffs* of where to compute including improved integration of HPC and data infrastructures.
- Formalize *data analytics* as a first class capability across the data lifecycle
  - Shift from a stewardship model to a *data-driven discovery model* where both stewardship and data discovery are enabled through a systematic computational infrastructure.
  - Data discovery methods should be applied *across the entire data lifecycle* to support scalable science activities at each point, sometimes automated, from onboard computing, to data processing and archive, to analysis and discovery.
- Computation and data science should play an important role in *planning new missions* including identification of how data, algorithms, and computation are to be integrated to improve overall data discovery, reproducibility and uncertainty management.
  - New capabilities should improve *reproducibility* of derived scientific results.
  - Derived scientific inferences should be *measurable and quantifiable.*

Acknowledgements

- NASA AIST Program
- JPL Data Science Working Group
- NASA OCT TA-11 Roadmap Team

See: http://ieee-bigdata-earthscience.jpl.nasa.gov/references for more details