

# Data Science at NASA/JPL

*Richard Doyle*

*Program Manager, Information and Data Science  
Project Manager, High Performance Spaceflight Computing*

*Daniel Crichton*

*Leader, Center for Data Science and Technology  
Project Manager, Planetary Data System Engineering  
Program Manager, Data Science Office*

*leaving the  
safe harbor  
to **explore**  
uncharted waters*

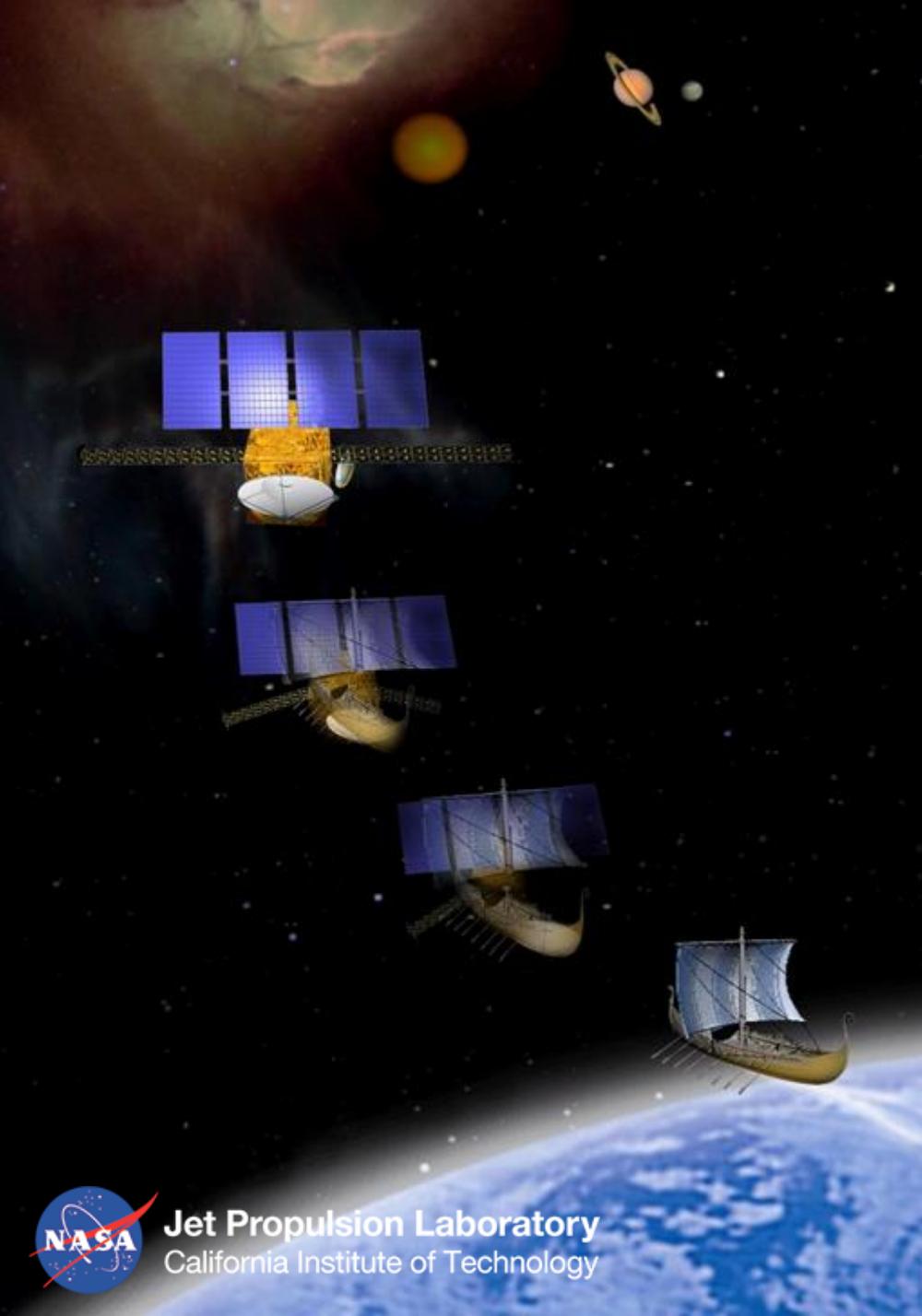
Jet Propulsion Laboratory  
California Institute of Technology

*November 14, 2017*



Jet Propulsion Laboratory  
California Institute of Technology

© 2017 California Institute of Technology.  
Government sponsorship acknowledged.



# Context and Strategy

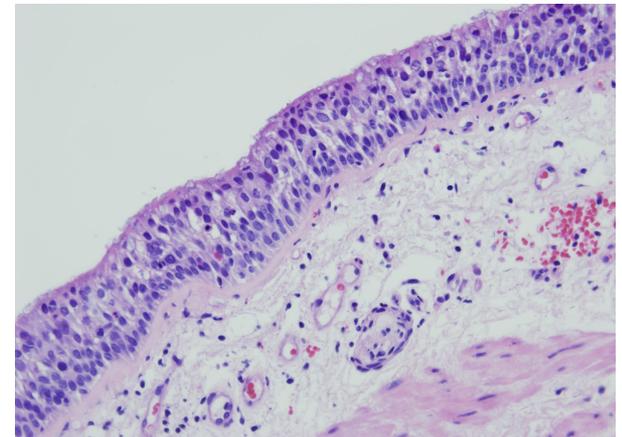
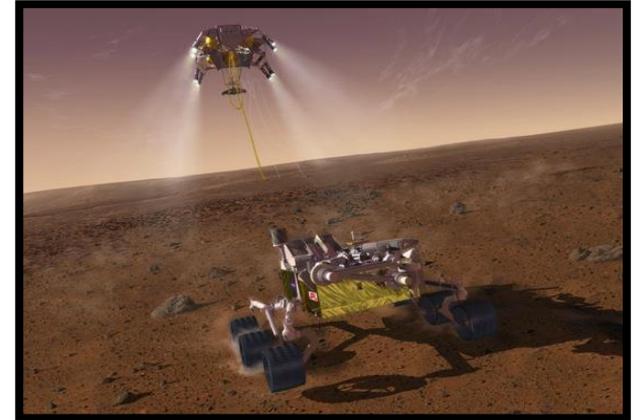


**Jet Propulsion Laboratory**  
California Institute of Technology



# Tackling Big Data

- JPL is involved in the research and development of technologies, methodologies in science, mission operations, engineering, and other non-NASA applications.
  - Includes onboard computing to scalable archives to analytics
- JPL and Caltech formed a joint initiative in Data Science and Technology to support fundamental research all the way to operational systems.
  - Methodology transfer across applications is a major goal.



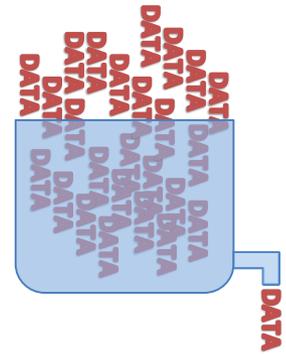


# Terms

## *Big Data and Data Science*

### Big Data

- When needs for data collection, processing, management and analysis go beyond the capacity and capability of available methods and software systems



### Data Science

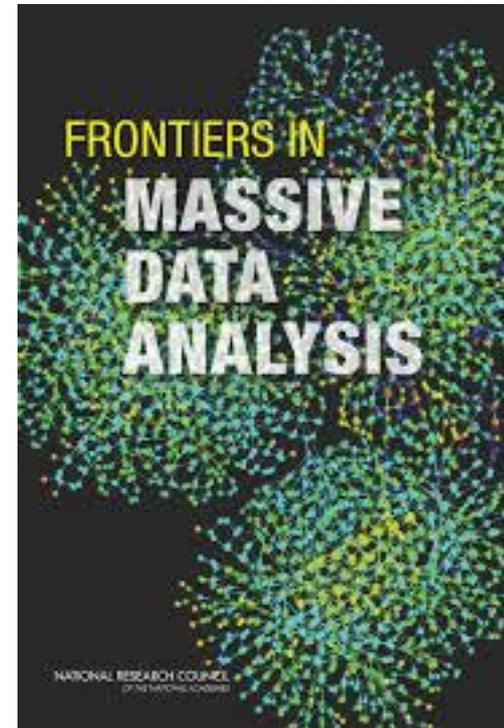
- *Scalable* architectural approaches, techniques, software and algorithms which alter the paradigm by which data is collected, managed and analyzed





# U.S. National Research Council Report: *Frontiers in the Analysis of Massive Data*

- Chartered in 2010 by the U.S. National Research Council, National Academies
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- NASA/JPL served on the committee covering systems architecture for big data management and analysis
- **Importance of more systematic approaches for analysis of data**
- **Need for end-to-end data lifecycle: from point of capture to analysis**
- **Integration of multiple discipline experts**
- Application of novel statistical and machine learning approaches for data discovery



2013



# The Growing Need for Data Science

from Space News

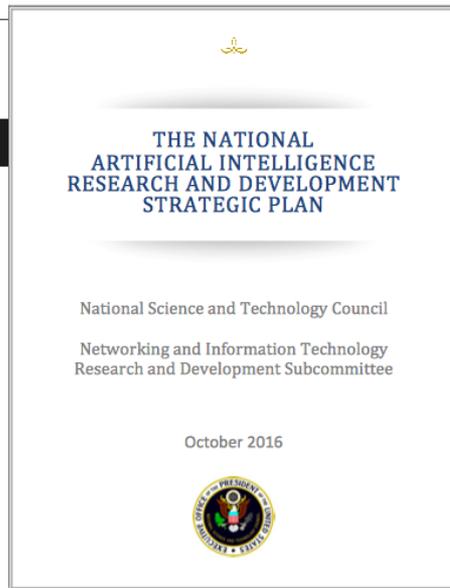
HOME ISSUE + FEATURES COLUMNS + LOG IN | SUBSCRIBE SPACENEWS.COM

## Smallsats and the multi-trillion-dollar data set

by Dylan Taylor August 1, 2016, Capital Contributions

**M**uch has been said about the revolution in small satellite technology and the copious number of constellations already in orbit or being prepared for launch. Certainly the amount of capital that has flowed into small satellites has been stunning with industry estimates at nearly \$1 billion in the past three years.

But is that capital investment justified?



“...traditional data analytics infrastructure will start to give way to strategic investments in data systems that are broad in scope (embracing all enterprise silos), provide distributed data infrastructures, use open source software...” - [Tamr](#)

“2016 will be the year where Artificial Intelligence (AI) technologies...are applied to ordinary data processing challenges...the new shift will include widespread applications of these technologies in ... tools that support applications, real-time analytics and data science. “ - Oracle

“Today’s operations centers struggle with an extremely high volume of events coming in requiring human analysis, which is unsustainable...in 2016 we will see organizations focus on using machine learning to significantly reduce the number of events requiring analysis down to the most critical.” - Snehal Antani, [Splunk](#)’s CTO

“...data itself is no longer the number one problem; connected data is the problem. To overcome this challenge, organizations need to add edge analytics to their existing strategy, analyzing data close to its source instead of sending it to a central place for analysis. “ - Mike Flannagan, Vice President, Data and Analytics, Cisco



# NASA Data Lifecycle Model



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

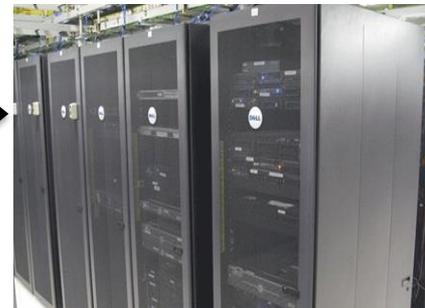
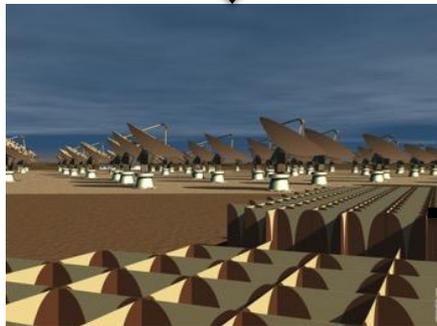
Observational Platforms and Flight Computing



SMAP (Today): 485 GB/day NI-SAR (2020): 86 TB/day

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

Massive Data Archives and Big Data Analytics



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

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

Ground-based Mission Systems

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

# Increasing Computing Capability Onboard

Heading Toward Multicore in Space



## Voyager computer

- 8,000 instructions/sec and kilobytes of memory

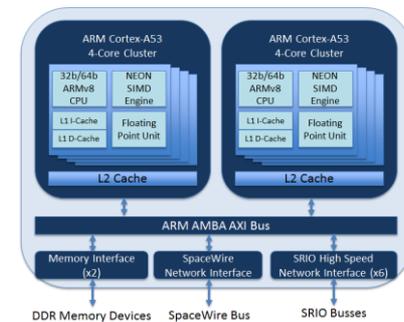


## iPhone

- 14 billion instructions/sec and gigabytes of memory

## Curiosity (Mars Science Laboratory)

Processor: 200 MOPS BAE RAD750



## HPSC (NASA STMD / AFRL)

Processor: 15 GOPS, extensible



# NASA Science and Big Data Today



Science Teams

How do these connect?

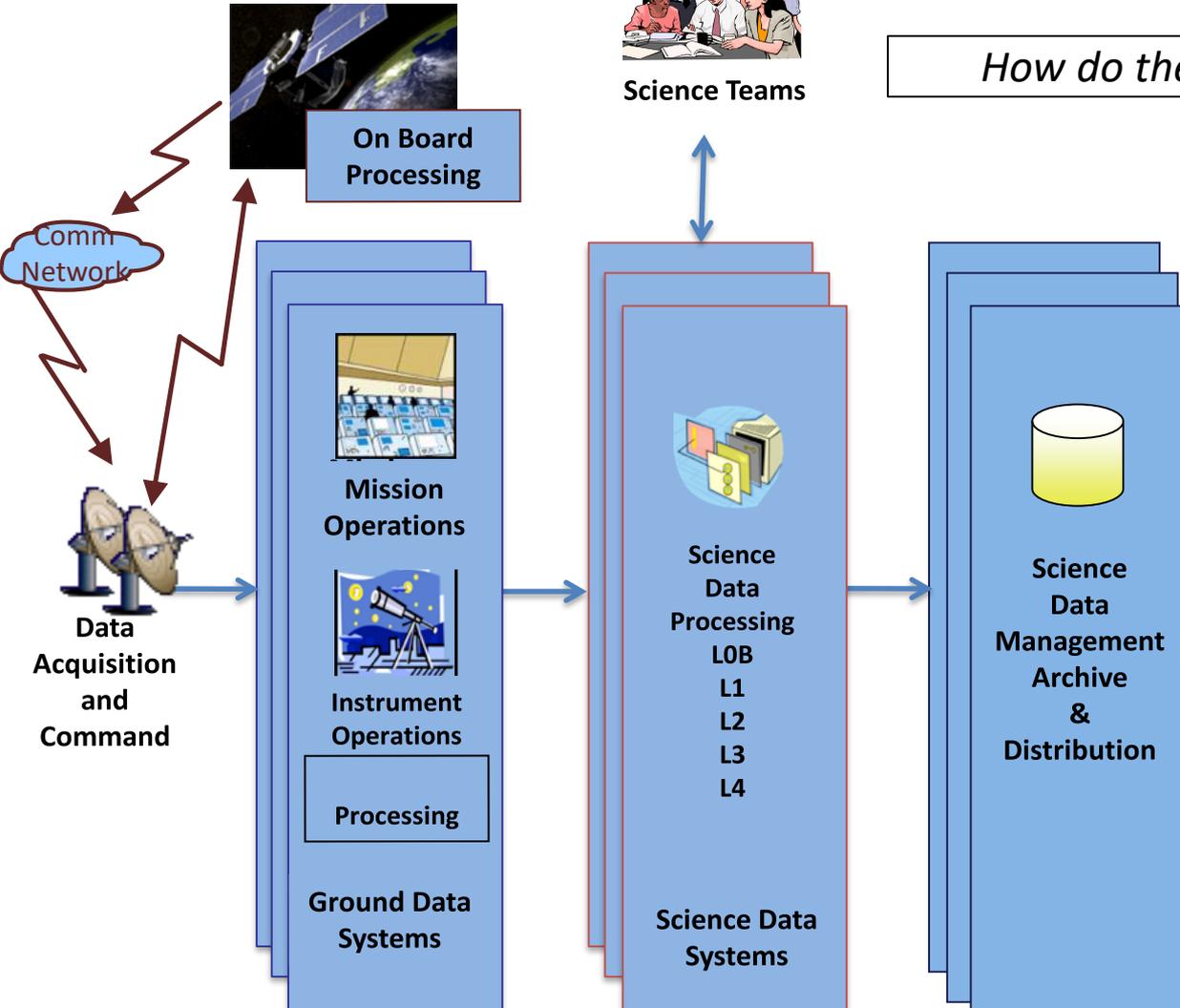
Research



Outreach



Applications



**Big Data Infrastructure (Data, Algorithms, Machines)**

?

*Focus on generating, capturing, managing big data*

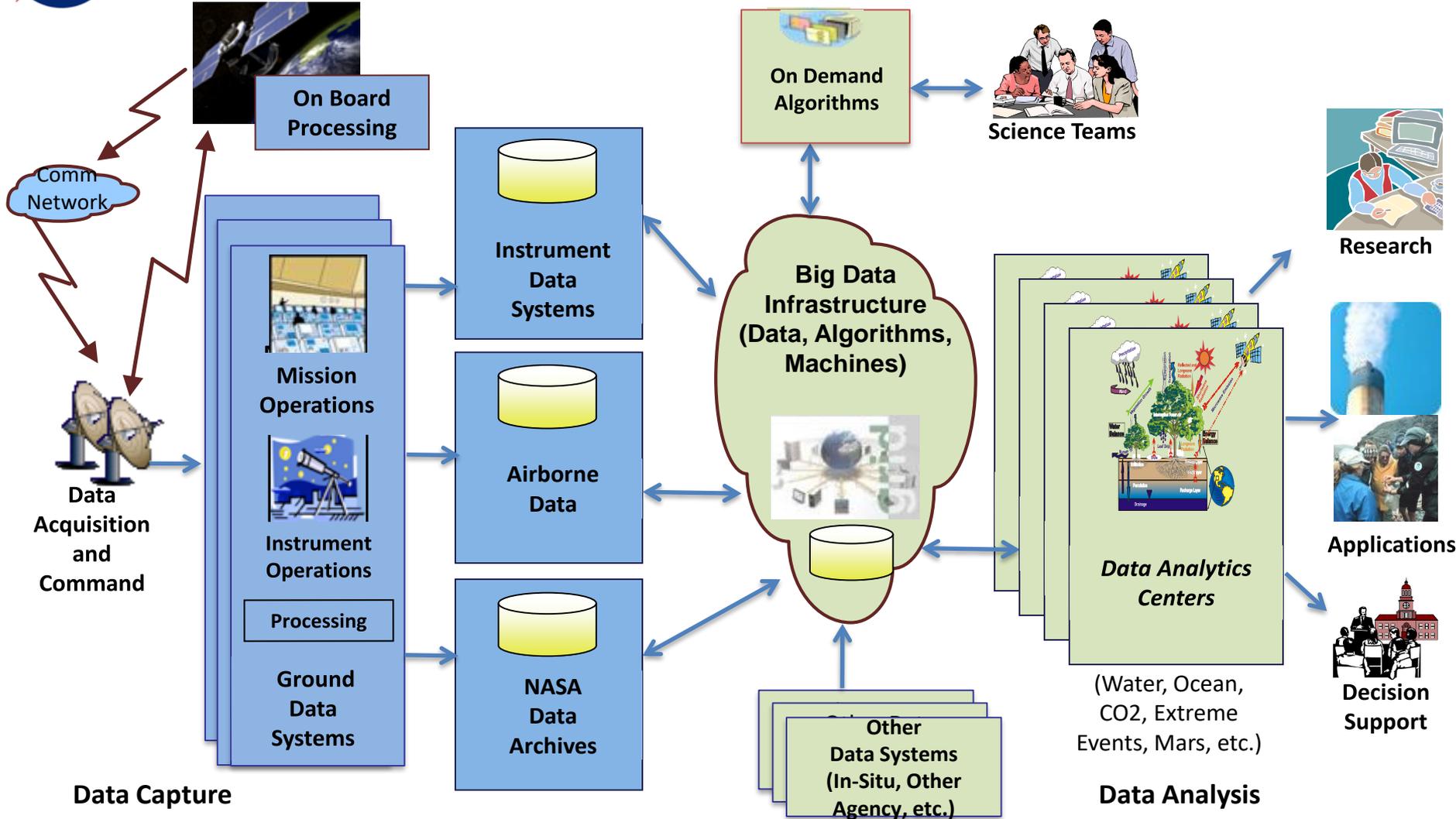
*Focus on using/analyzing big data*



Jet Propulsion Laboratory  
California Institute of Technology

# Future of Data Science at NASA

## Enabling a Big Data Research Environment



Reducing Data Wrangling: "There is a major need for the development of software components... that link high-level data analysis-specifications with low-level distributed systems architectures."

*Frontiers in the Analysis of Massive Data*, National Research Council, 2013.



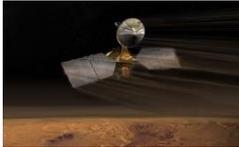
# Data Science Strategy

## Guiding Principles

### Agile Science – Onboard Analysis

**Challenge:** Too much data, too fast, cannot transport data efficiently enough

**Future Solutions:** Onboard computation and data science

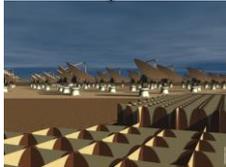


## Data Lifecycle

### Extreme Data Volumes – Data Triage

**Challenge:** Data collection capacity at the instrument, outstrips data transport and data storage capacity

**Future Solutions:** Dynamic architectures to scale data processing and triage, exascale data streams

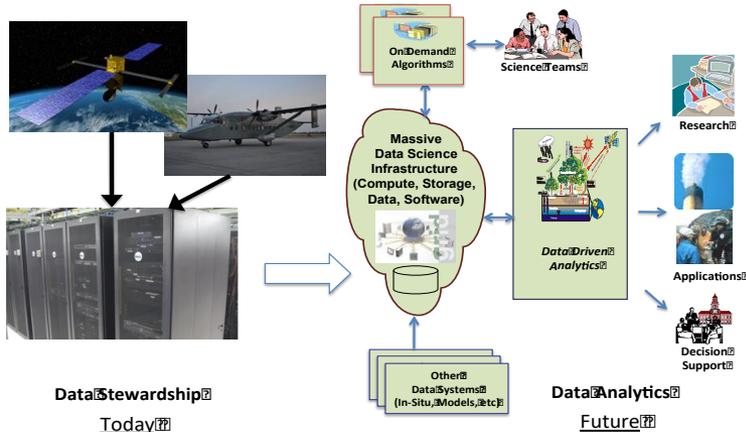
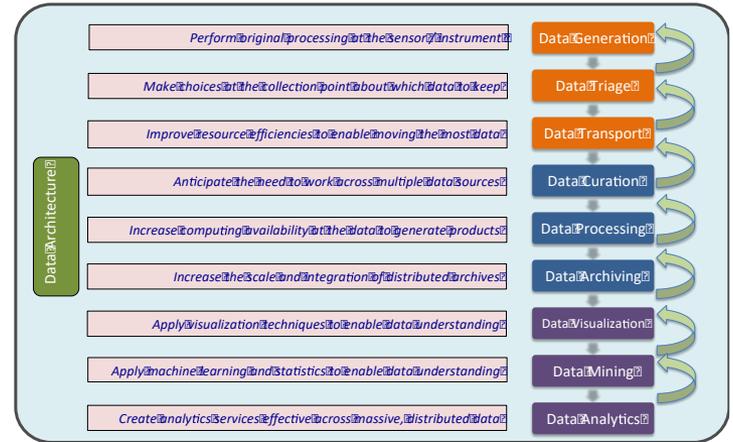


### Distributed Data Analytics



**Challenge:** Data distributed in massive archives, many different types of measurements

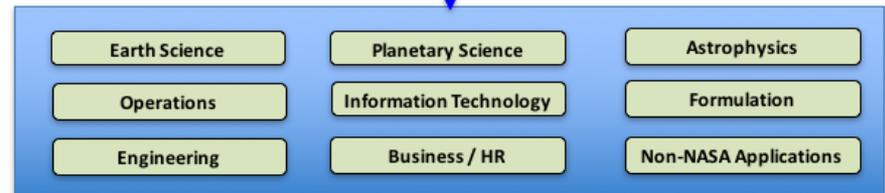
**Future Solutions:** Distributed data analytics, uncertainty quantification



## Data Ecosystem

## JPL Data Science Strategy

### Data Science Working Group



## Cross-Cutting



# Development Strategy

- In November 2016, JPL chartered a **Data Science Working Group** reporting to the Laboratory Management Council (LMC) and JPL Deputy Director, Larry James
- **Data Science Pilots** – Seed concepts and drive data science into the fabric of JPL
  - In 2017, JPL launched 12 funded pilots across science, mission operations, DSN, formulation, and business
  - In 2018, this is expanding to engage a Lab-wide data science community
- **Data Science Services** – Mature institutional capabilities to integrate data science into “how we do work”
- **Data Science Projects** – Drive to “grand challenge” topics such as reproducibility

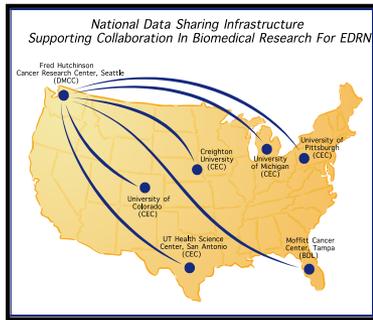


# Partnering Strategy

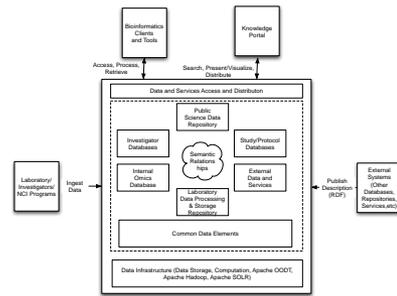
- Universities
  - Caltech joint center in data science
  - Increasing curriculum in data science
  - Opportunities for NASA and JPL investment in internships and research
  - Collaborations with UC, CMU, MIT (Lincoln Lab and CSAIL)
- International Partners
  - Interoperability of archives
  - Engagement of technologies and data scientists across agencies
- Commercial and Open Source
  - Leverage mature technologies in cloud computing
  - Leverage and collaboration on big data technologies
  - Form public-private research partnerships
  - Collaborations with Amazon Web Services and the Apache Foundation



# Cross-Cutting Capabilities



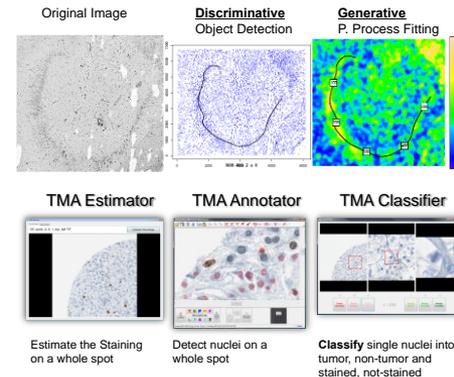
## International Data Archive and Sharing Architectures



## Big Data

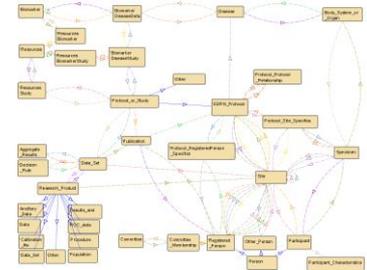
## Infrastructures

(from open source to cloud computing and scalable compute infrastructures)



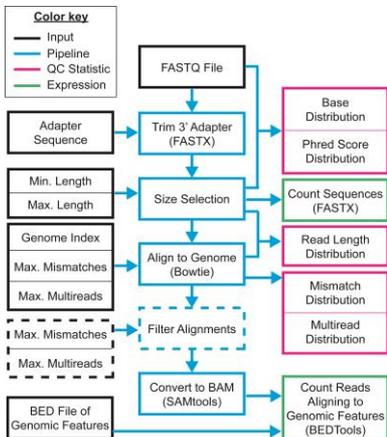
## Intelligent Data Algorithms

(Machine Learning, Deep Learning)

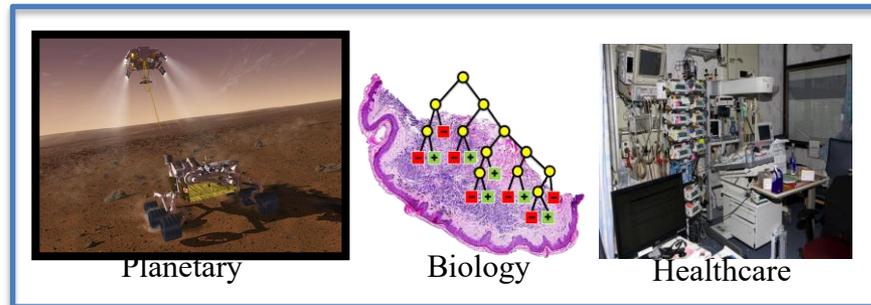


## Common Data Elements & Information Models

(discipline and common)



## Analytical Data Pipelines



## Great Opportunities for Methodology Transfer and Collaboration



## Visualization Techniques



# Opportunities Across the Ground System Environment

## Intelligent Ground Stations



### Emerging Solutions

- *Anomaly Detection*
- *Combining DSN & Mission Data*
- *Attention Focusing*
- *Controlling False Positives*

## Intelligent Archives and Knowledge-bases



### Emerging Solutions

- *Automated Machine Learning - Feature Extraction*
- *Intelligent Search*
- *Learning over time*
- *Integration of disparate data*

**Technologies: Machine Learning, Deep Learning, Intelligent Search, Data Integration, Interactive Visualization and Analytics**

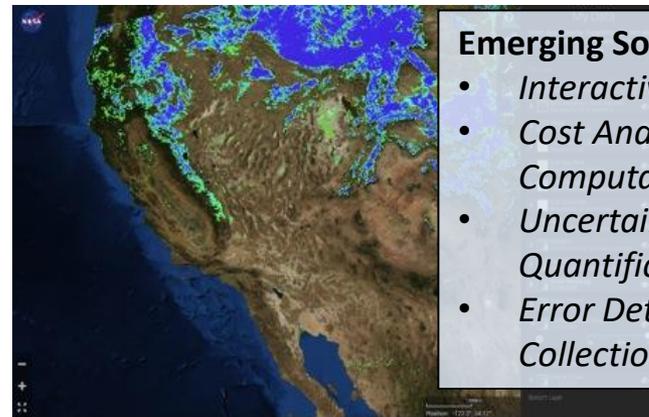
## Intelligent MOS-GDS



### Emerging Solutions

- *Anomaly Interpretation*
- *Dashboard for Time Series Data*
- *Time-Scalable Decision Support*
- *Operator Training*

## Data Analytics and Decision Support



### Emerging Solutions

- *Interactive Data Analytics*
- *Cost Analysis of Computation*
- *Uncertainty Quantification*
- *Error Detection in Data Collection*



Jet Propulsion Laboratory  
California Institute of Technology

# Highly Scalable Data-Driven Ground Systems

## Intelligent Ground Stations



## Data-Driven Discovery from Archives



Machine Learning, Deep Learning, Intelligent Search, Data Fusion, Uncertainty Quantification, Attention Focusing, Decision Support, Interactive Visualization and Analytics

## Mission Operations



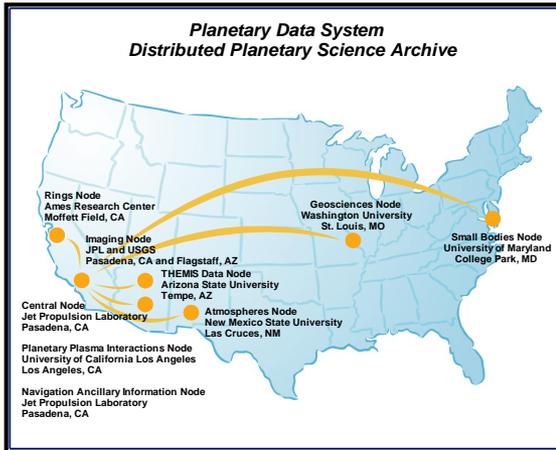
## Data Analytics and Decision Support





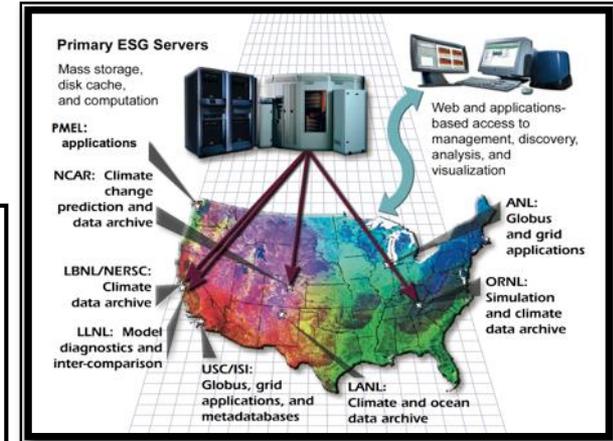
Jet Propulsion Laboratory  
California Institute of Technology

# Scientific Research Networks: Access to Observations and Models



Solar System Exploration

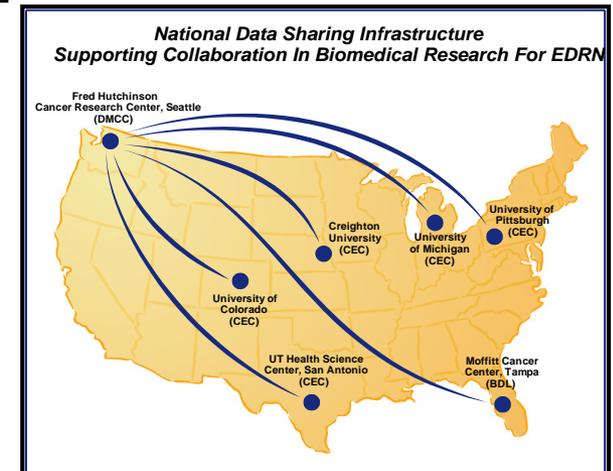
Highly distributed/federated  
Collaborative  
Information-centric  
Discipline-specific  
Growing/evolving  
Heterogeneous  
International



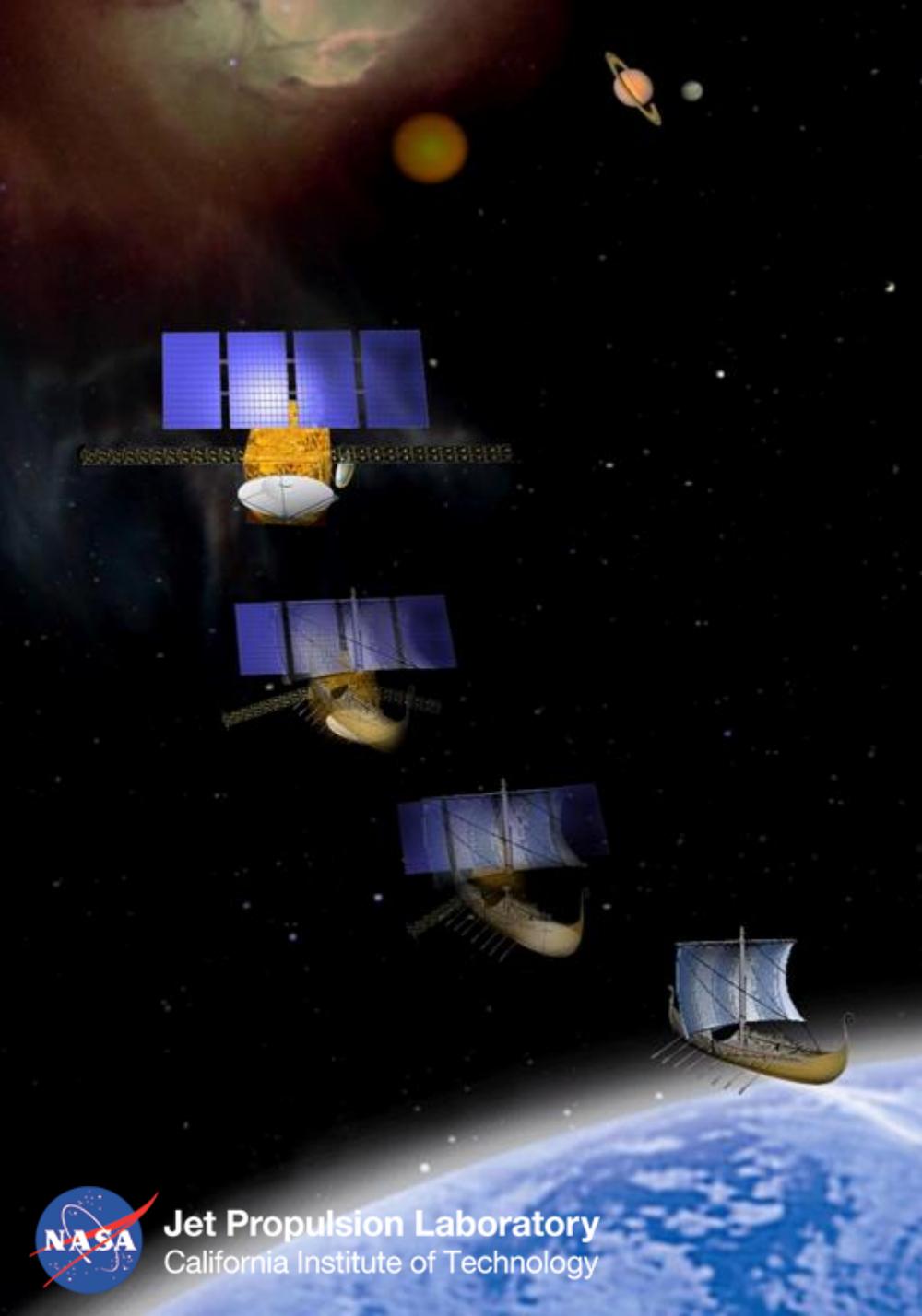
Climate Research



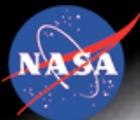
Earth Observation



Cancer Research



# Planetary Data System

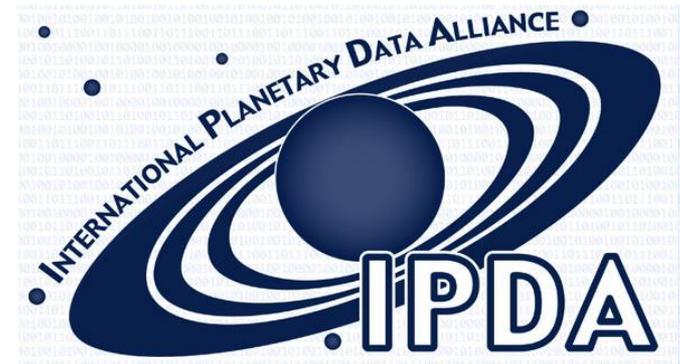
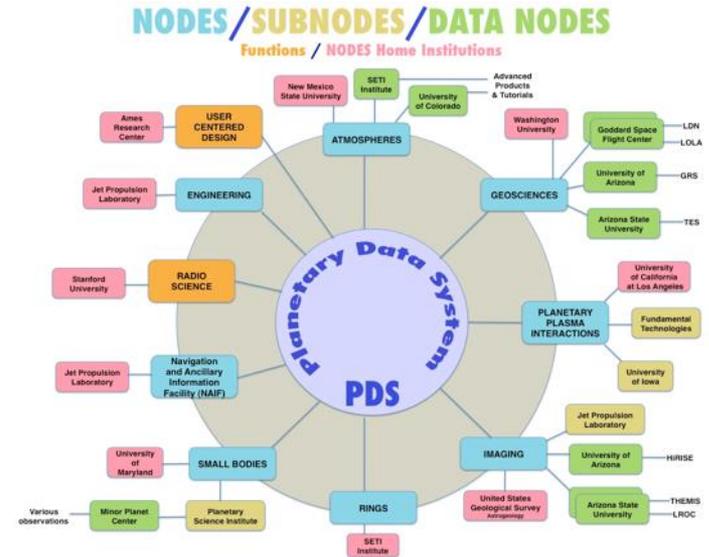


**Jet Propulsion Laboratory**  
California Institute of Technology



# Planetary Data System

- Purpose: To collect, archive and make accessible digital data and documentation produced from NASA's exploration of the solar system from the 1960s to the present.
- Infrastructure: A highly distributed infrastructure with planetary science data repositories implemented at major government labs and academic institutions
  - System driven by a well defined planetary science information model
  - Over 1 PB of data
  - Movement towards international interoperability
  - Distributed federation of US nodes and international archives
- Being realized through PDS4





# (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*

***A well formed model that drives the software is something that many groups have struggled with!***

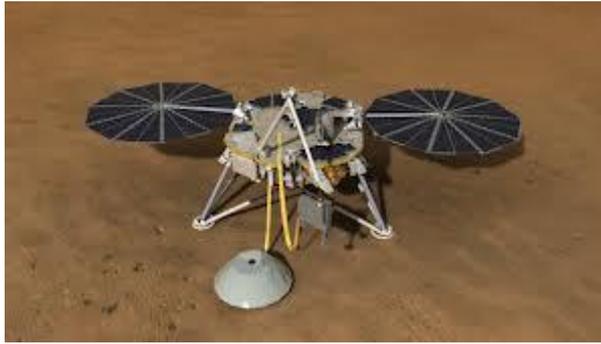


Jet Propulsion Laboratory  
California Institute of Technology

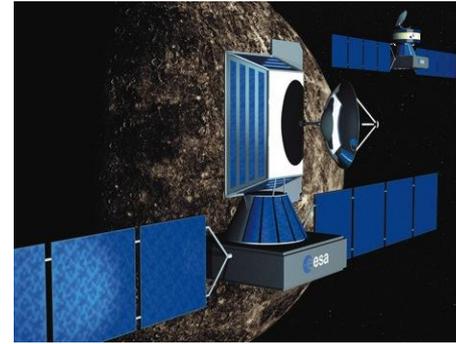
# PDS4: International Adoption of an Open Planetary Approach



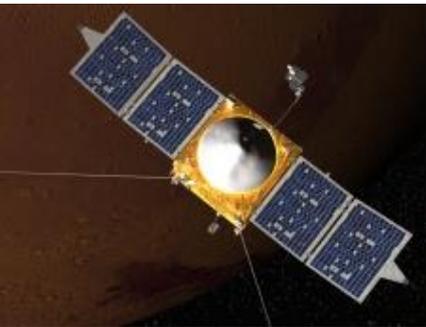
LADEE (NASA)



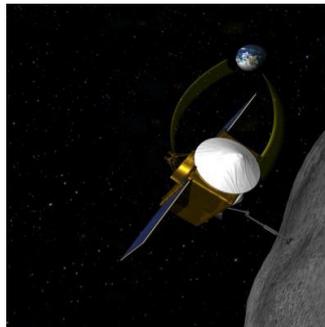
InSight (NASA)



BepiColombo (ESA/JAXA)



MAVEN (NASA)



Osiris-Rex (NASA)



ExoMars



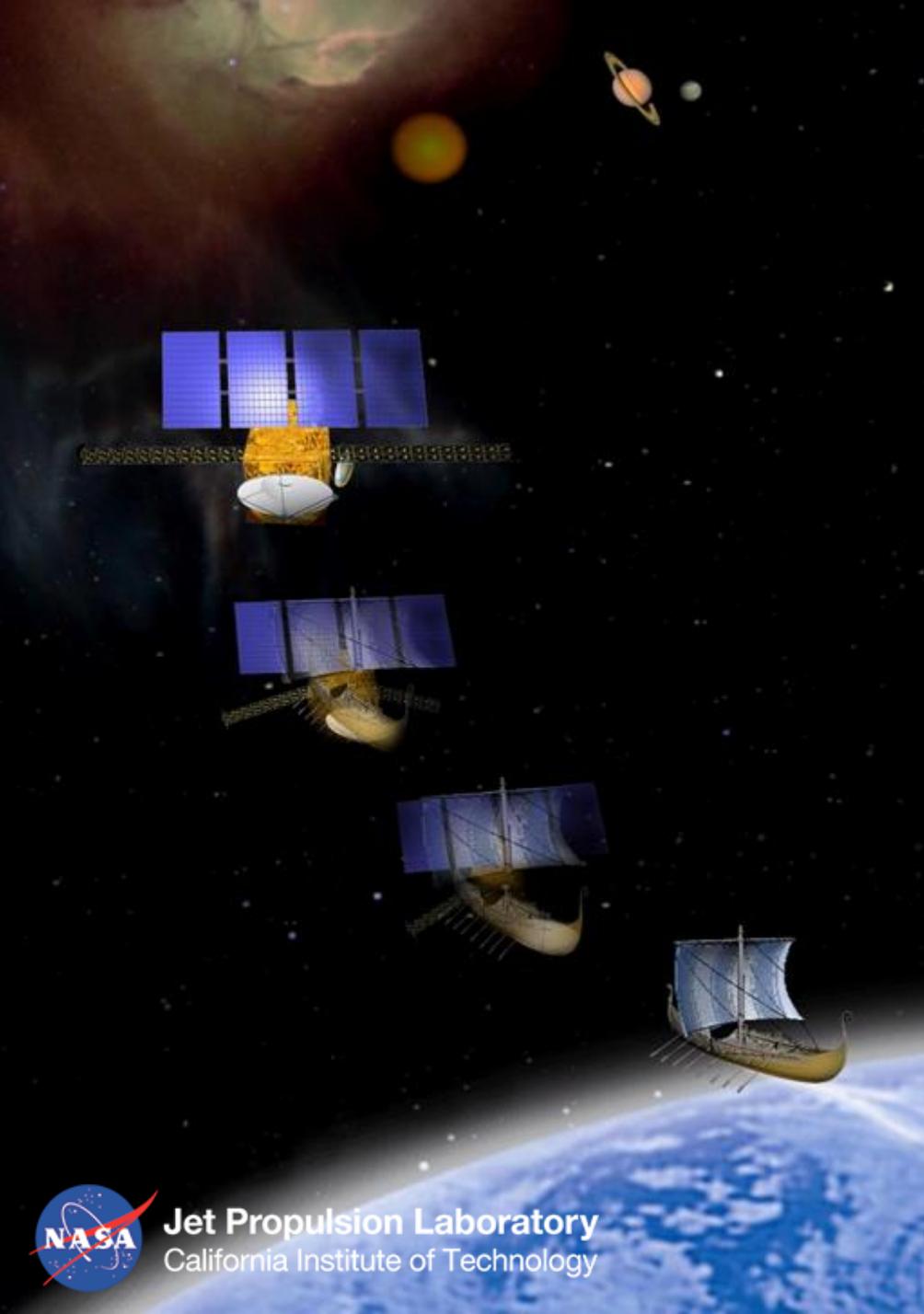
JUICE

(ESA/Russia)(ESA)

...also Hayabussa-2, Chandrayaan-2



Mars 2020 (NASA)



# Some Applications



# Onboard Analysis

## *Dust Devils on Mars*

Dust devils are scientific phenomena of a transient nature that occur on Mars

- They occur year-round, with seasonally variable frequency
- They are challenging to reliably capture in images due to their dynamic nature
- Scientists accepted for decades that such phenomena could not be studied in real-time



*Spirit Sol 543  
(July 13, 2005)*

New onboard Mars rover capability (as of 2006)

- Collect images more frequently, analyze onboard to detect events, and only downlink images containing events of interest

Benefit

- < 100% accuracy can dramatically increase science event data returned to Earth
- *First notification includes a complete data product*



6/10/2020



# Mars Trek: The Google Earth of Mars

The image displays the Mars Trek interface, which provides a 3D perspective of the Martian surface. The main view shows a topographic map of Mars with various features labeled, including Peace Vallis, Aeolis Mensae, and Fretted Vallis. A central popup window titled "Gale Crater" provides detailed information about the landing site. The popup includes a small image of the crater and a text box that reads: "With a diameter of 154 km and a central peak 5.5 km tall, Gale Crater was chosen as the landing site for the Mars Science Laboratory Curiosity rover. The choice was based on evidence from orbiting spacecraft that indicate that the crater may have once contained large amounts of liquid water. The central peak, Mount Sharp, exhibits layered rock deposits rich in sedimentary minerals including clays, sulfates, and salts that require water to form." To the right of the main map, a sidebar titled "Curiosity Landing Site" features a small image of the rover and a text box that reads: "Curiosity landed in Gale Crater on Mars on August 6th, 2012. With a diameter of 154 km and a central peak 5.5 km tall, Gale Crater was chosen as the landing site for the Mars Science Laboratory Curiosity rover. The choice was based on evidence from orbiting spacecraft that indicate that the crater may have once contained large amounts of liquid water. The central peak, Mount Sharp, exhibits layered rock deposits rich in sedimentary minerals including clays, sulfates, and salts that require water to form." Below the sidebar, there are buttons for "Region Information" and "Download for 3D Printer". In the bottom right corner, a 3D rendering shows a rover standing on a rocky ridge, looking out over the vast, hazy Martian landscape.



# WaterTrek

User Defined Polygon

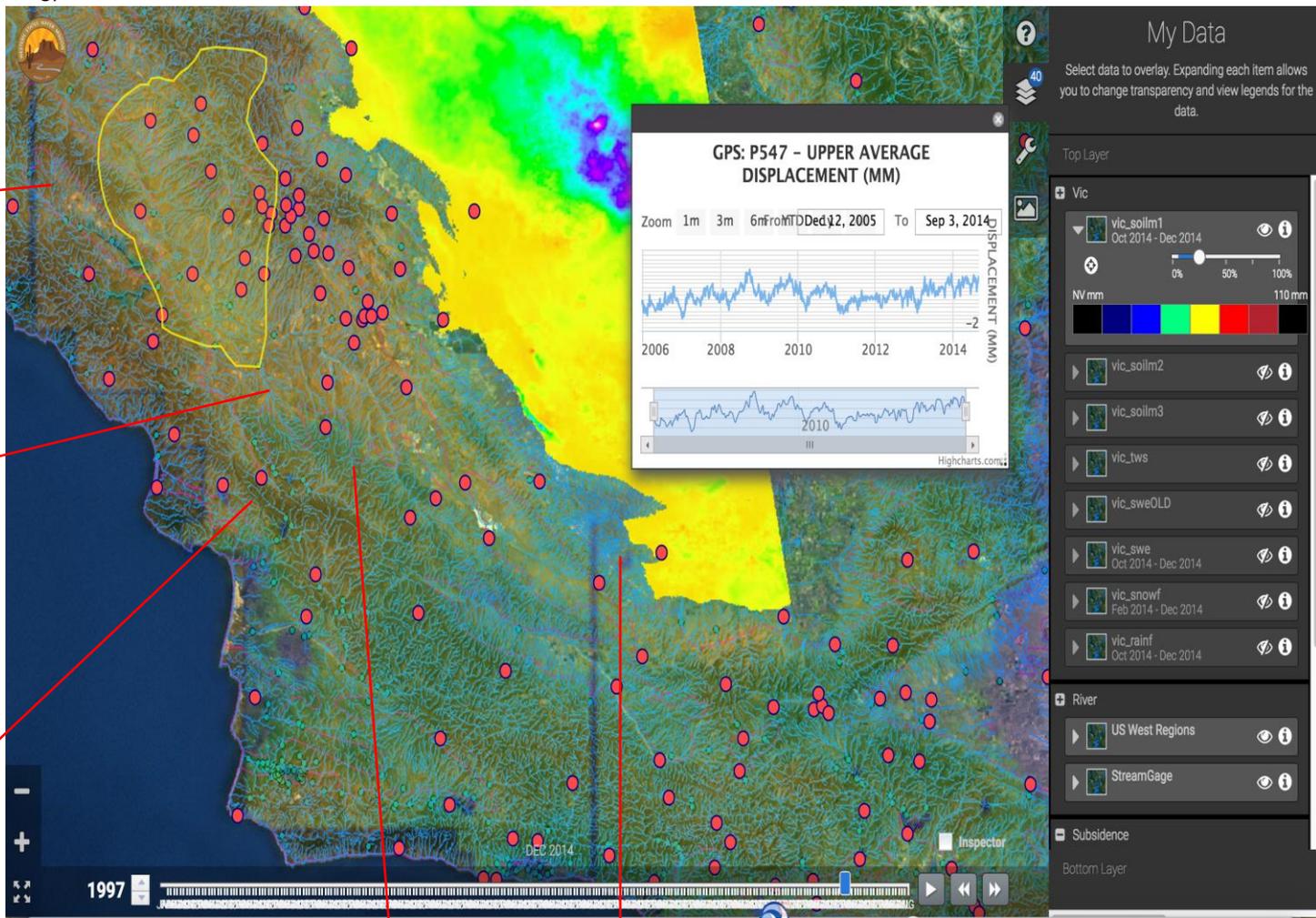
GPS

In-Situ: Stream Gage Sensors

River Network

SAR derived Subsidence

Model Output  
Soil Moisture



Fusing In-situ, Air-borne, Space-borne and model generated data using visualization and a big data analytics engine



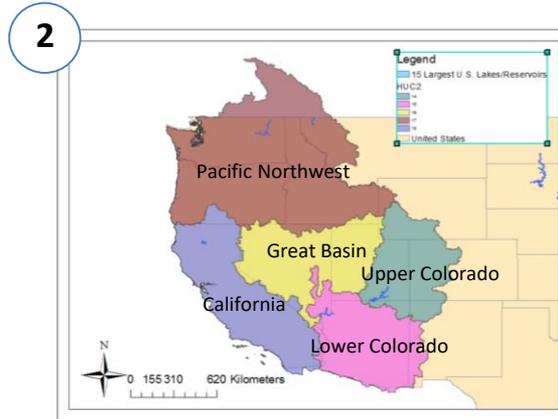
Jet Propulsion Laboratory  
California Institute of Technology

# Western States Water Mission

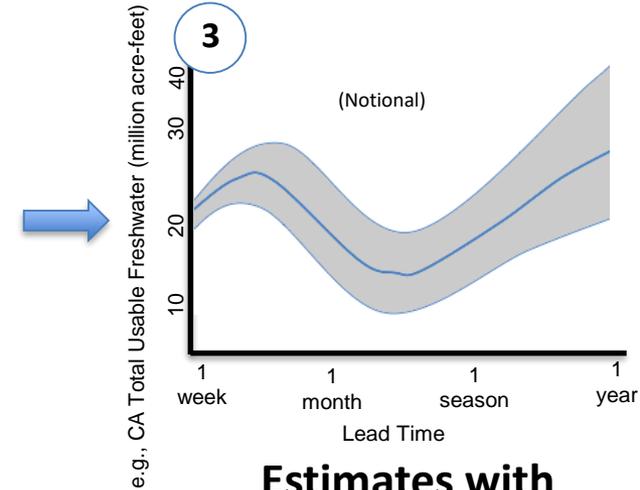
## Understanding Water Availability



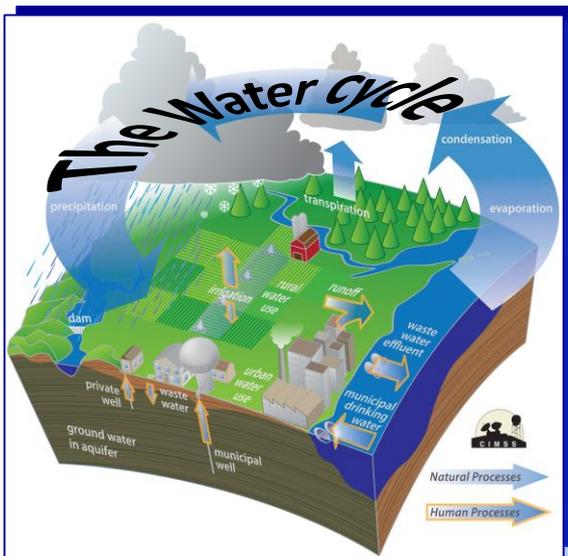
**Observations**



**Coupled and Validated  
Computer Models**



**Estimates with  
Uncertainties**



12  
January  
2016



(Prospective customers)



Colorado River Basin

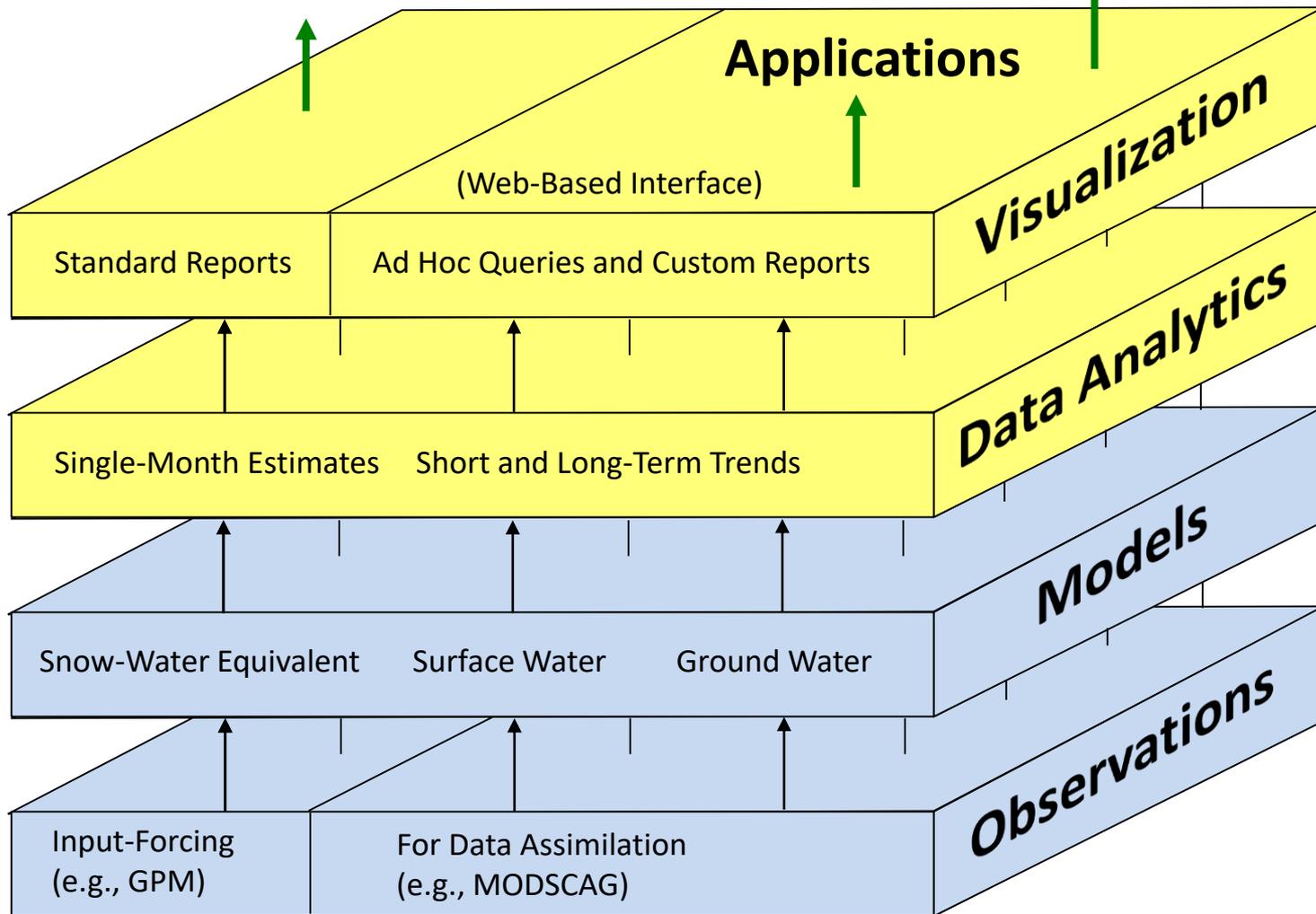
**Stakeholders and Customers**



# Western States Water Mission (WSWM): A Science/Data Science Collaboration

**Decision Support**

**Research**



Data Science Infrastructure  
(Tools, Services, Methods for Massive Data Analysis)

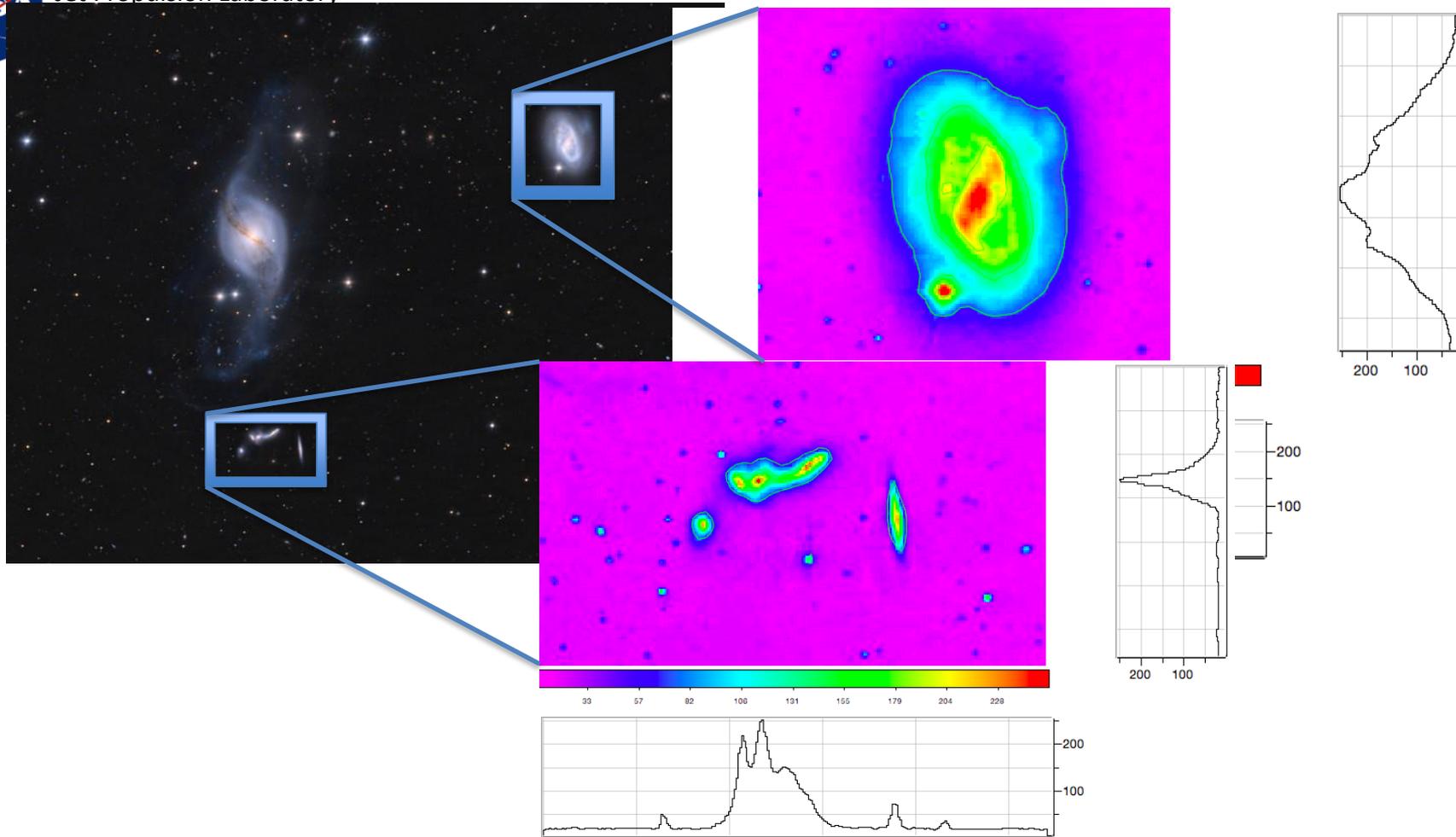
A Scalable Data Processing System for Hydrological Science



# Methodology Transfer from Planetary and Earth Science to Biomedicine

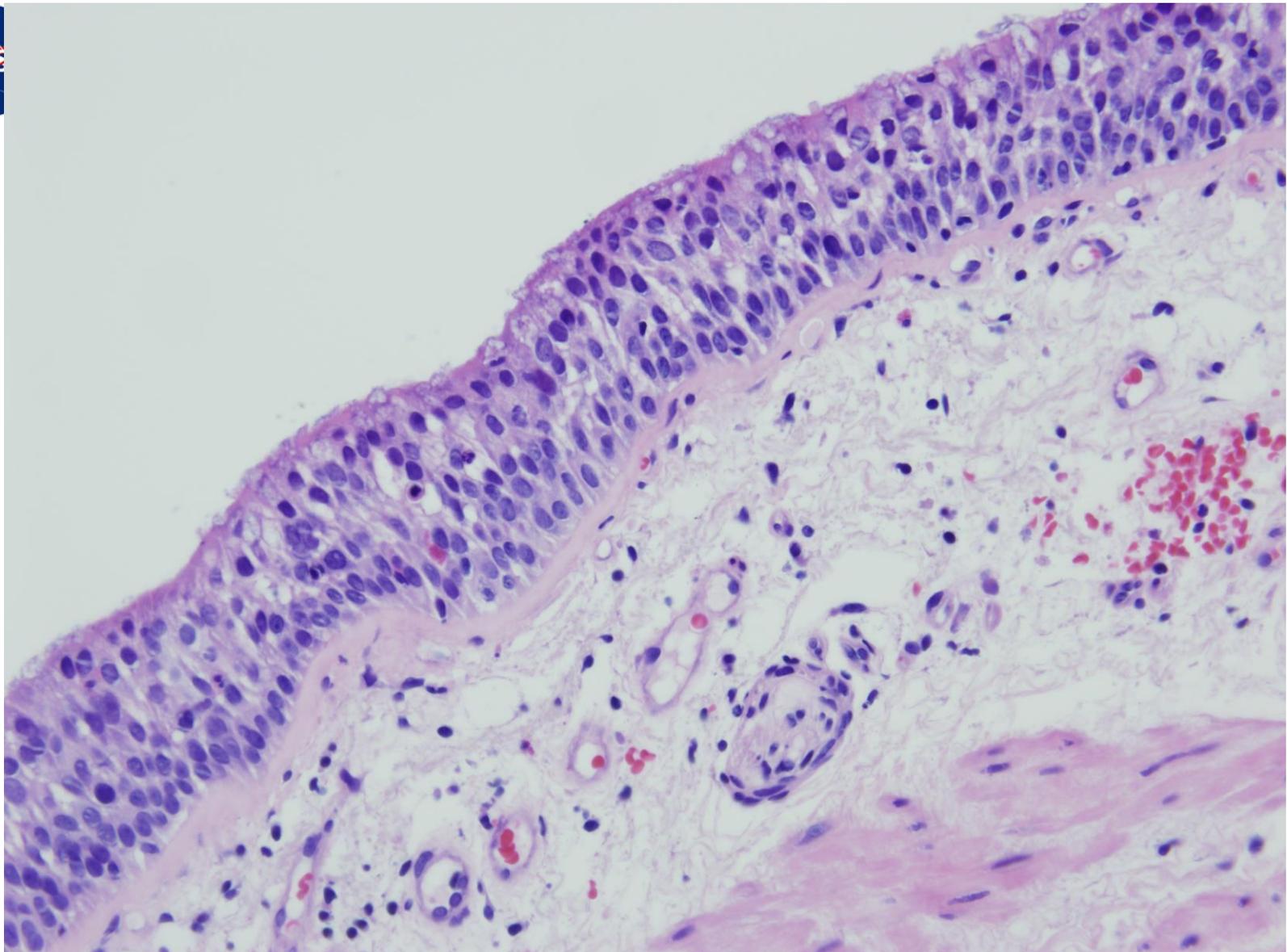


**Jet Propulsion Laboratory**  
California Institute of Technology



Description: Detecting objects from astronomical measurements by evaluating light measurements in pixels using intelligent software algorithms.

Image Credit: Catalina Sky Survey (CSS), of the Lunar and Planetary Laboratory, University of Arizona, and Catalina Realtime Transient Survey (CRTS), Center for Data-Driven Discovery, Caltech.

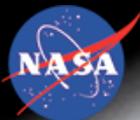


Description: Detecting objects from oncology images using intelligent software algorithms transferred to and from space science.

Image Credit: EDRN Lung Specimen Pathology image example, University of Colorado



# Driving Forward



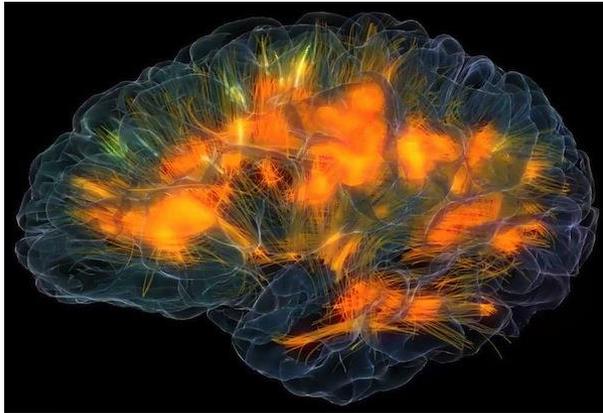
**Jet Propulsion Laboratory**  
California Institute of Technology



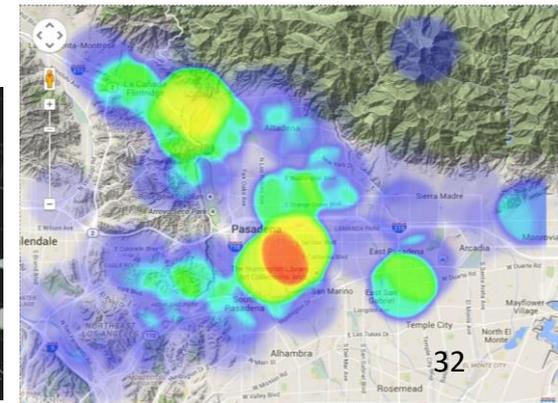
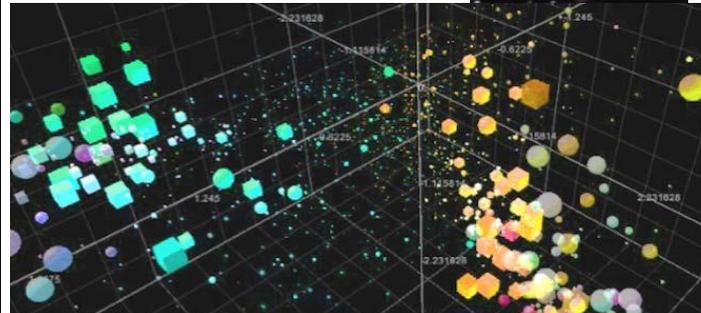
Jet Propulsion Laboratory  
California Institute of Technology

# Caltech-JPL Partnership in Data Science

- Center for Data-Driven Discovery on campus/Center for Data Science and Technology at JPL
- From basic research to deployed systems ~10 collaborations
  - Leveraged funding from JPL to Caltech; from Caltech to JPL
- Virtual Summer School (2014) has seen over 25,000 students



CENTER FOR DATA-DRIVEN DISCOVERY





# Example University Partnerships

## UC Riverside Students Training at NASA's Jet Propulsion Laboratory

Ten students from UC Riverside will have internships at JPL thanks to a \$4.5 million grant from NASA

By Sean Nealon On JUNE 10, 2016

SHARE THIS ARTICLE:



RIVERSIDE, Calif. (www.ucr.edu) — Ten University of California, Riverside students will have internships at NASA's Jet Propulsion Laboratory (JPL) this summer thanks to a \$4.5 million grant the university received last year from NASA.

The grant will also allow 22 high school students from Riverside Unified School District to take a STEM (Science, Technology, Engineering, Mathematics) class at UC Riverside this summer.

The University of California, Riverside received the NASA grant to develop research, education,



Big Data promises a better world. A world where data will be used to make better decisions, from how we invest money to how we manage our healthcare to how we educate our children and manage our cities and resources. These changes



The MIT Big Data Challenge Take me to the CITY OF



The next available Tackling the Challenges of Big



MIT BIG DATA LIVING LABA key issue today is that

### LATEST NEWS

SystemsThatLearn@CSAIL Lecture Series | Inaugural Event

March 23, 2017

Speakers:

**Daniel Crichton**, Program Manager, Principal Investigator and Principal Computer Scientist,

[NASA's Jet Propulsion Laboratory \(JPL\)](#)

**Richard Doyle**, Program Manager for Information and Data Science, [Jet Propulsion Laboratory \(JPL\)](#), California Institute of Technology

onal  
an  
data

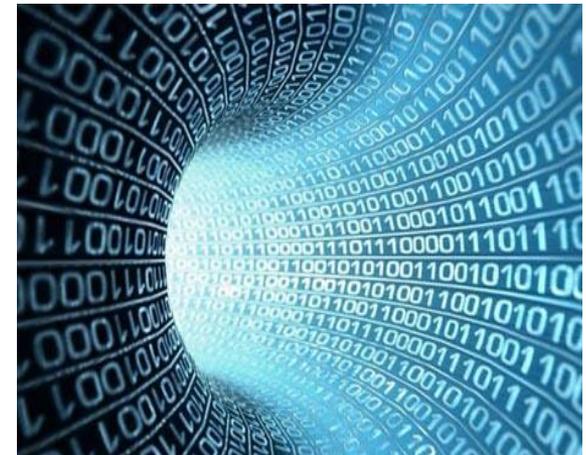


# Go Forward Strategy

- Use the Mission-Science Data Lifecycle to organize Big Data at NASA
  - From flight computing to data analytics
- Enable use of data analytics in the community
  - Promote data ecosystems for sharing data
  - Support international partnerships
- Explore opportunities for methodology transfer
  - Across science disciplines at NASA
  - With other agencies
  - Focused around open source
- Establish multi-disciplinary teams between science/discipline experts, and computer science/data science experts



*What do we do with all this data?*

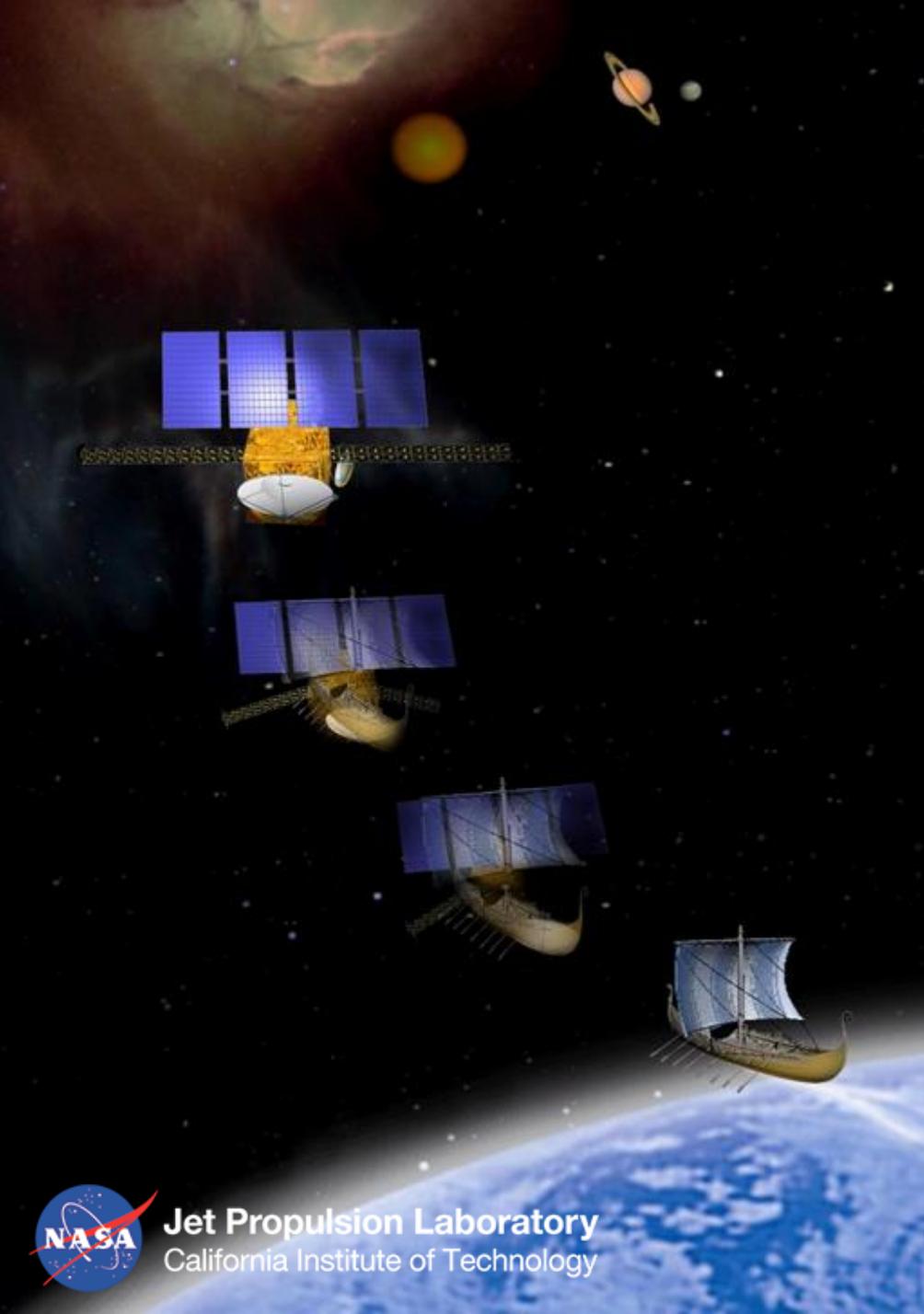


*This is looking like a black hole –  
but wait, there's light at the end of the tunnel!*



# References

- Frontiers on Massive Data Analysis, NRC, 2013
- NASA OCT Technology Roadmap, NASA, 2015
- NASA AIST Big Data Study, NASA/JPL 2016
- IEEE Big Data Conference, Data and Computational Science Big Data Challenges for Earth Science Research, IEEE, 2015
- IEEE Big Data Conference, Data and Computational Science Big Data Challenges for Earth and Planetary Science Research, IEEE, 2016
- Planetary Science Informatics and Data Analytics Conference, April 2018



**Questions?**



**Jet Propulsion Laboratory**  
California Institute of Technology