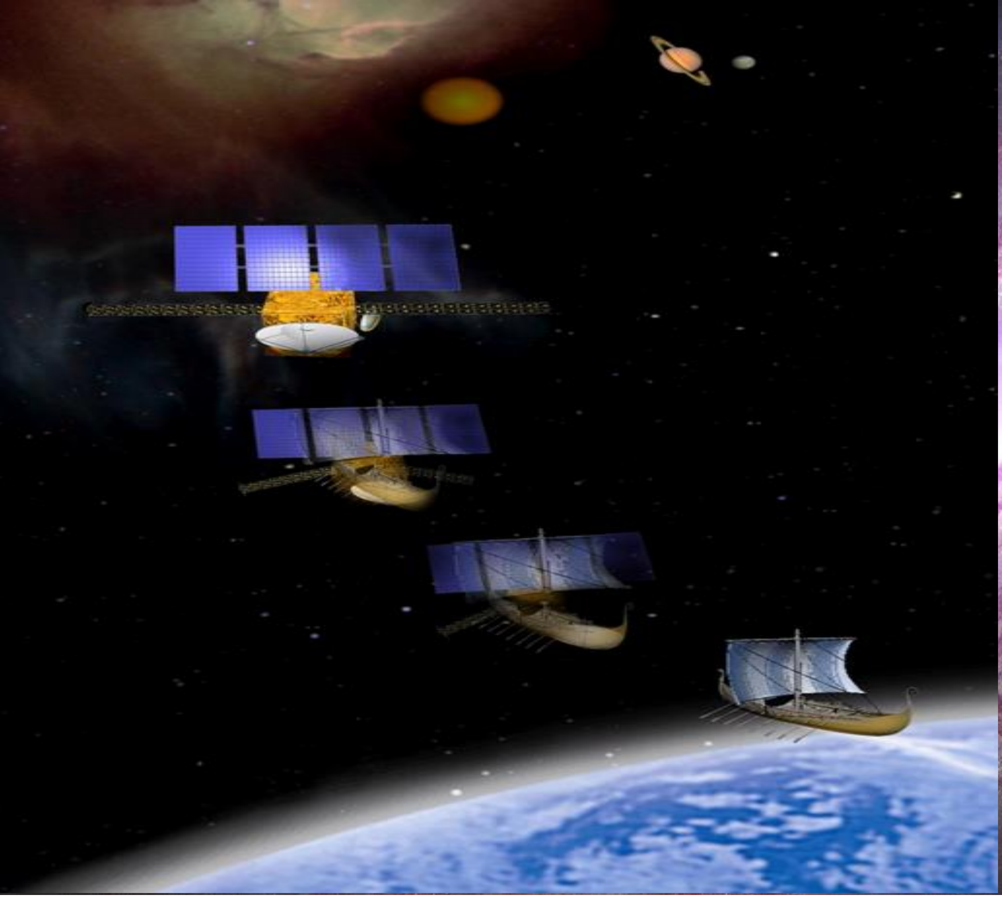


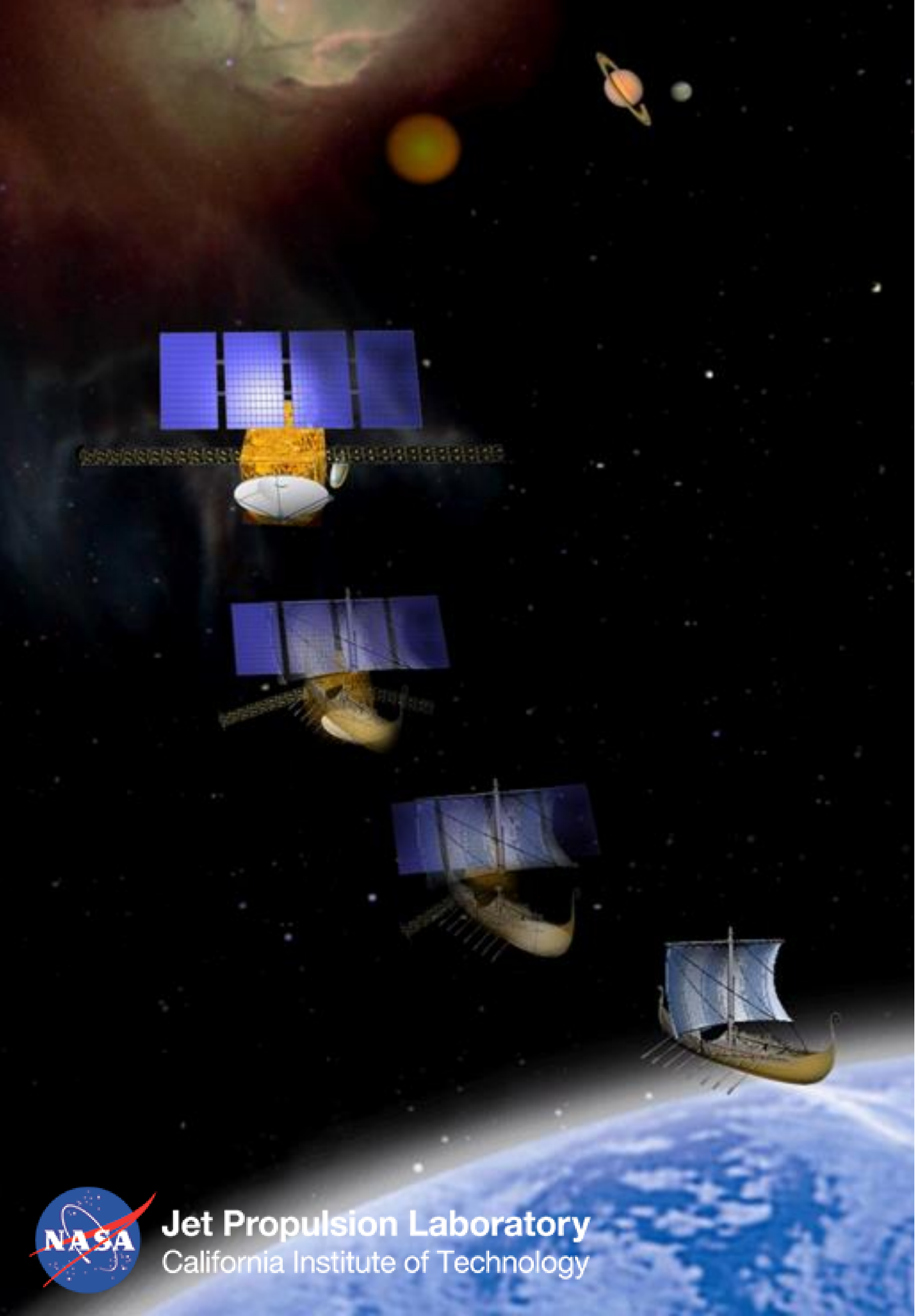

Enabling Data-Driven Methodologies across the Data Lifecycle and Ecosystem

Jet Propulsion Laboratory
California Institute of Technology

Richard J. Doyle and Daniel J. Crichton

AGU Fall Meeting 2017

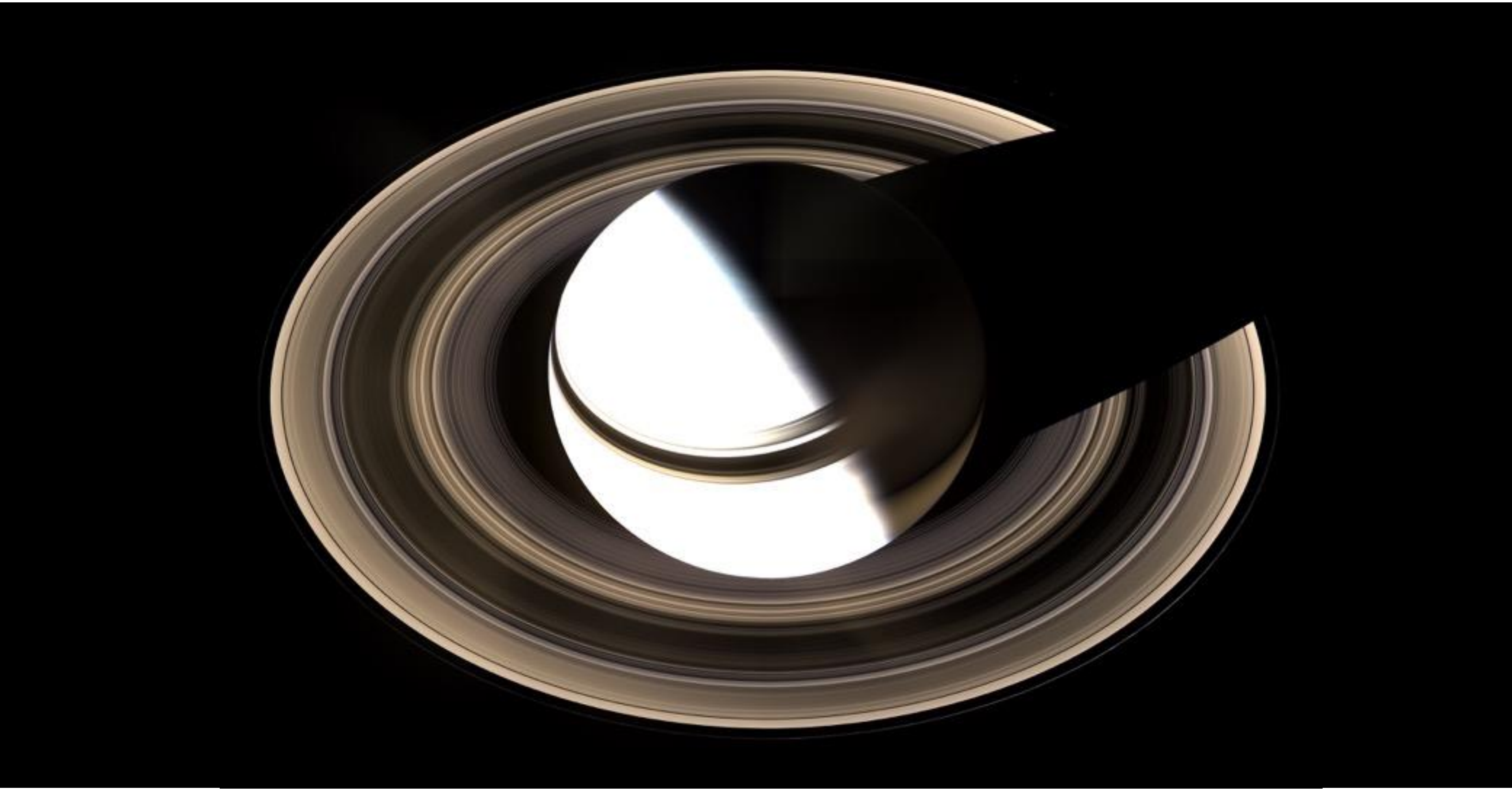
December 12, 2017


NASA
Jet Propulsion Laboratory
California Institute of Technology

Context

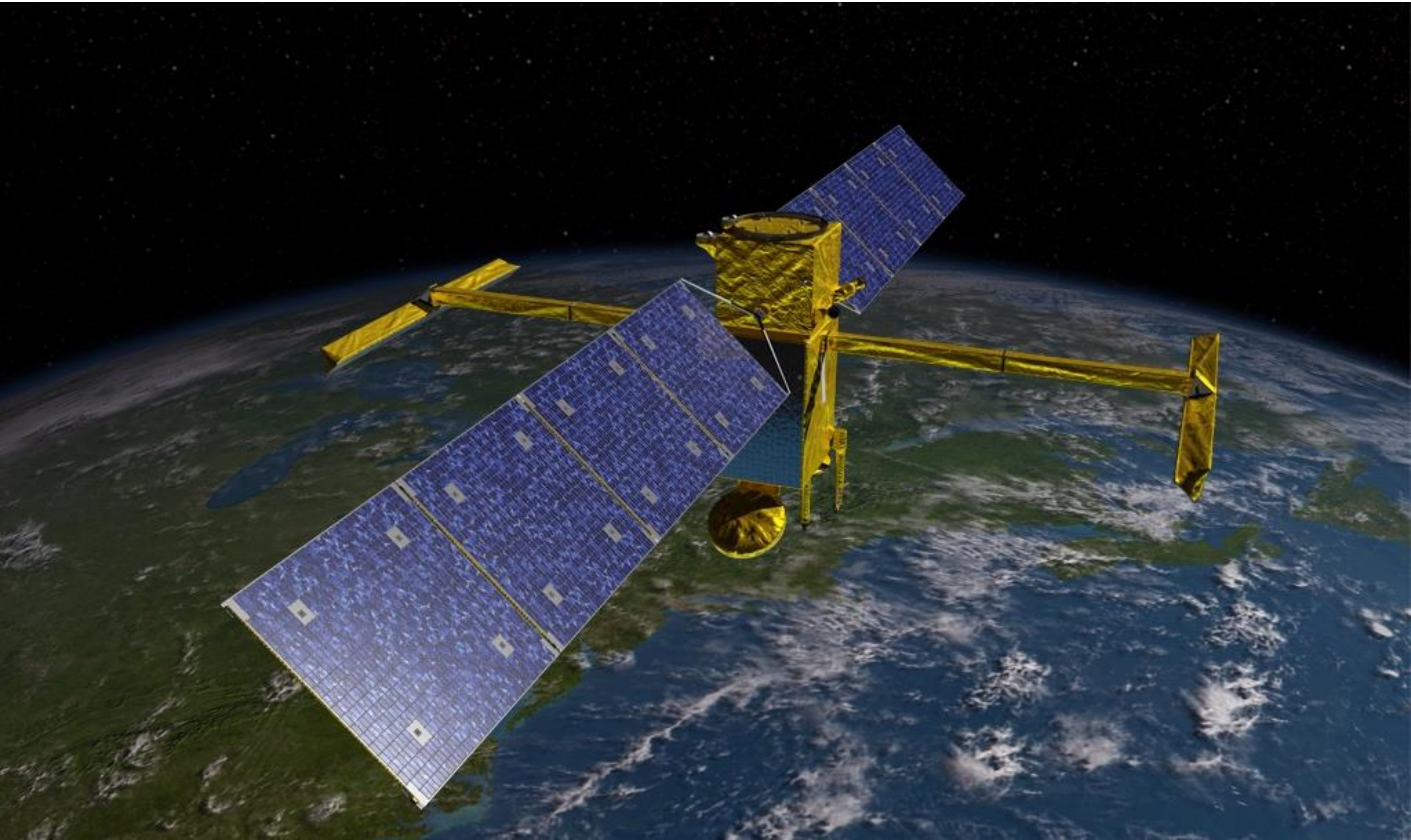
Cassini

Orbit Insertion June 28, 2004 Grand
Finale September 15, 2017



Surface Water Ocean Topography (SWOT)

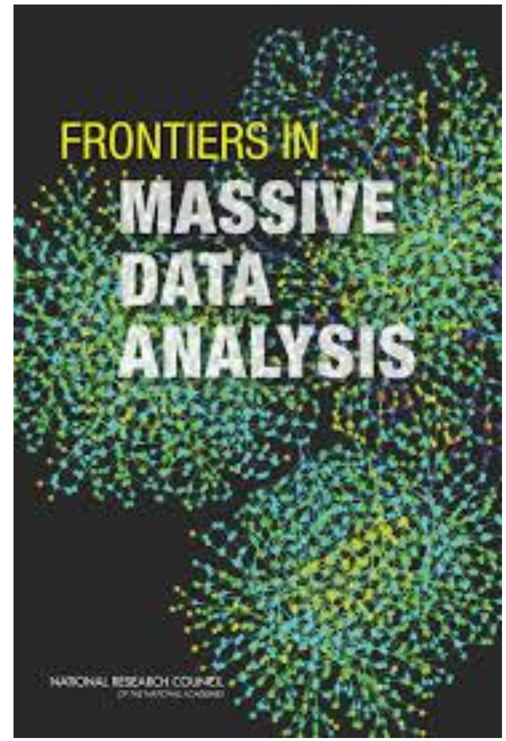
2020 Launch



Jet Propulsion Laboratory
California Institute of Technology

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

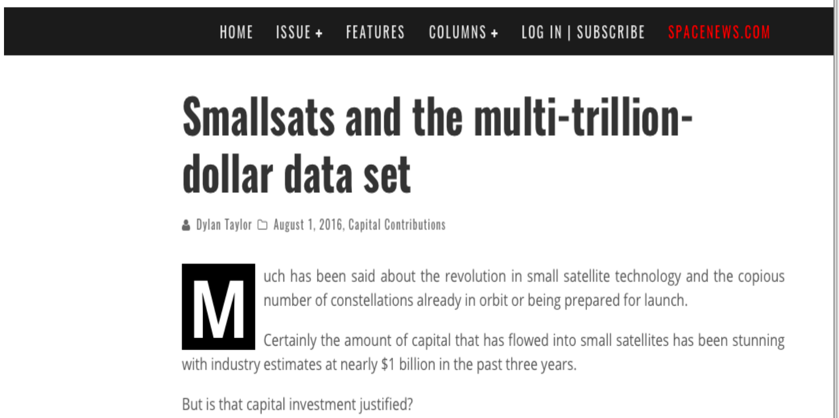


2013

Jet Propulsion Laboratory
California Institute of Technology

The Growing Need for Data Science

from Space News



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

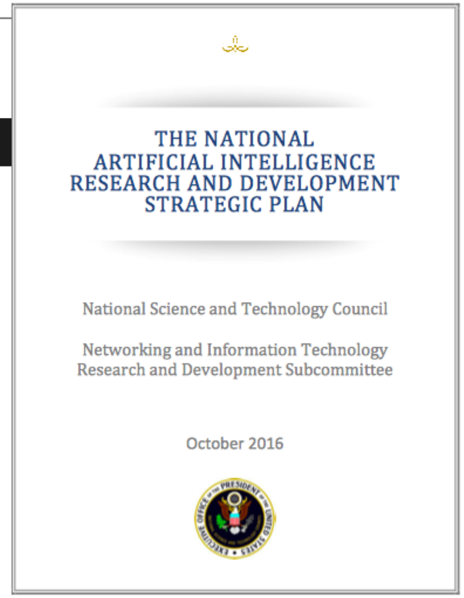
Smallsats and the multi-trillion-dollar data set

Dylan Taylor · August 1, 2016, Capital Contributions

Much 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?



THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

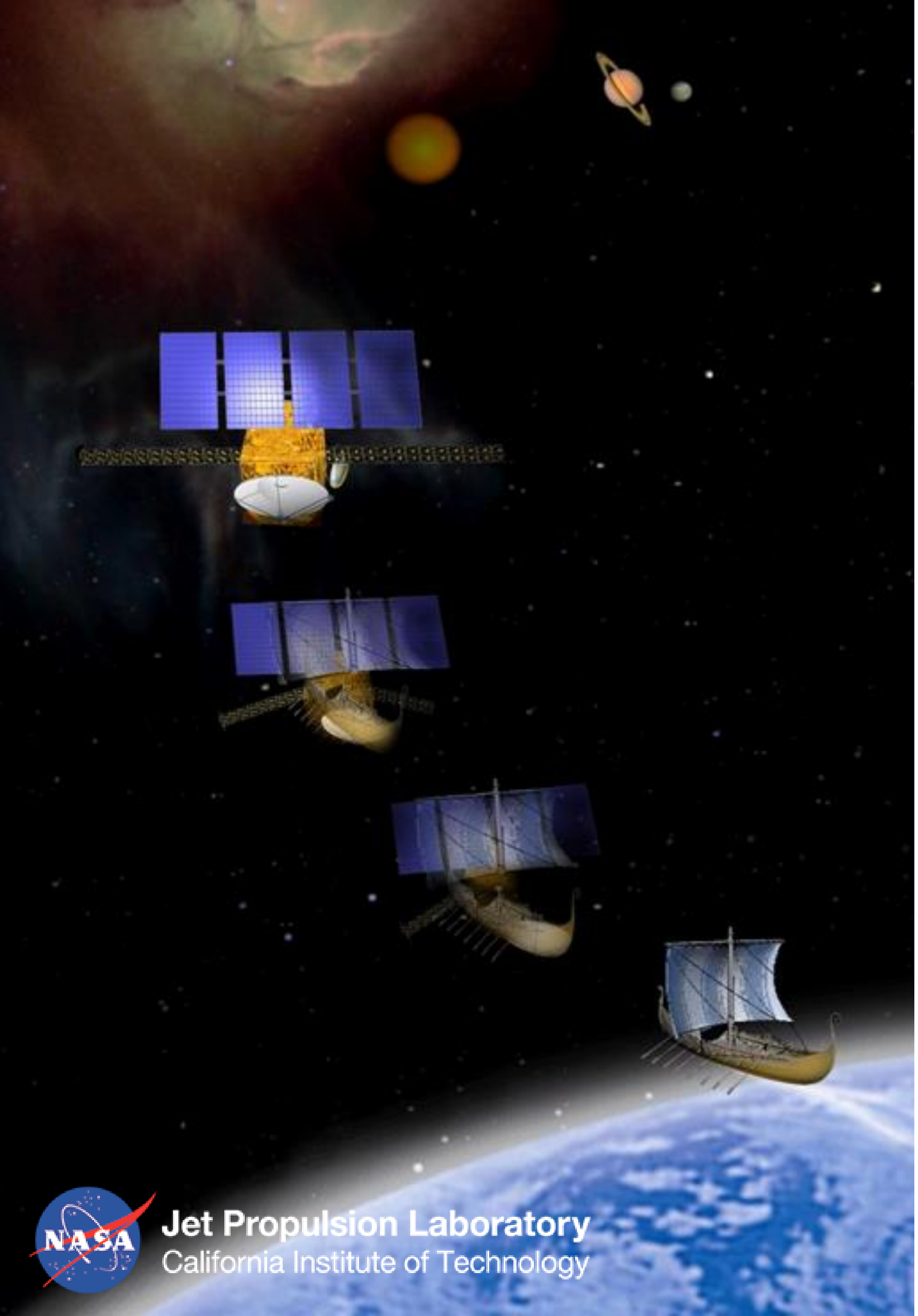
October 2016

"…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



The Challenge

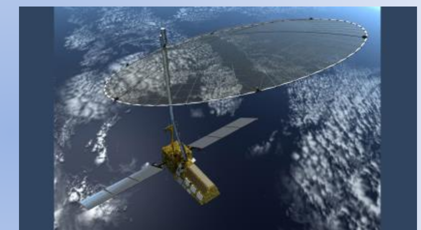
Enabling Data-Driven Exploration

NASA Data Lifecycle Model

Jet Propulsion Laboratory
California Institute of Technology

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

Observational Platforms and Flight Computing

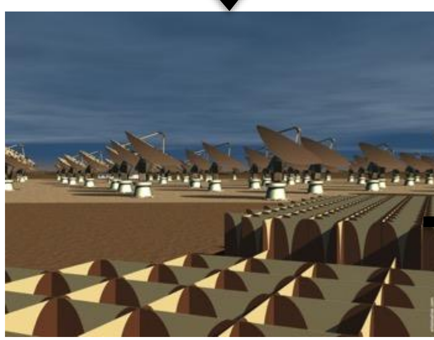


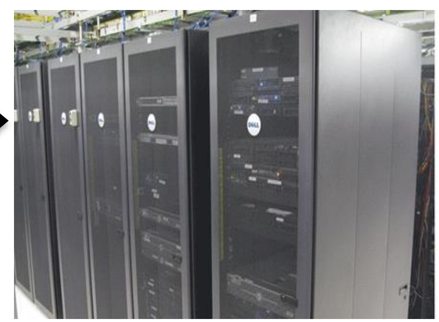
SMAP (Today): 485 GB/day NI-SAR (2020): 86 TB/day

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

Massive Data Archives and Big Data Analytics

Emerging Solutions
- *Intelligent Ground Stations*
- *Agile MOS-GDS*





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

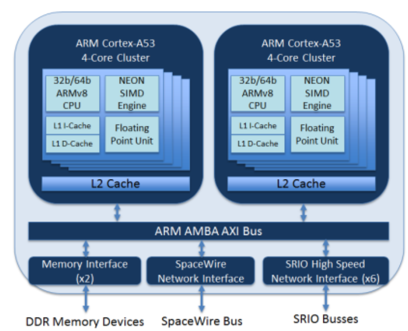
Voyager computer

- 8,000 instructions/sec and kilobytes of memory

iPhone

- 14 billion instructions/sec and gigabytes of memory

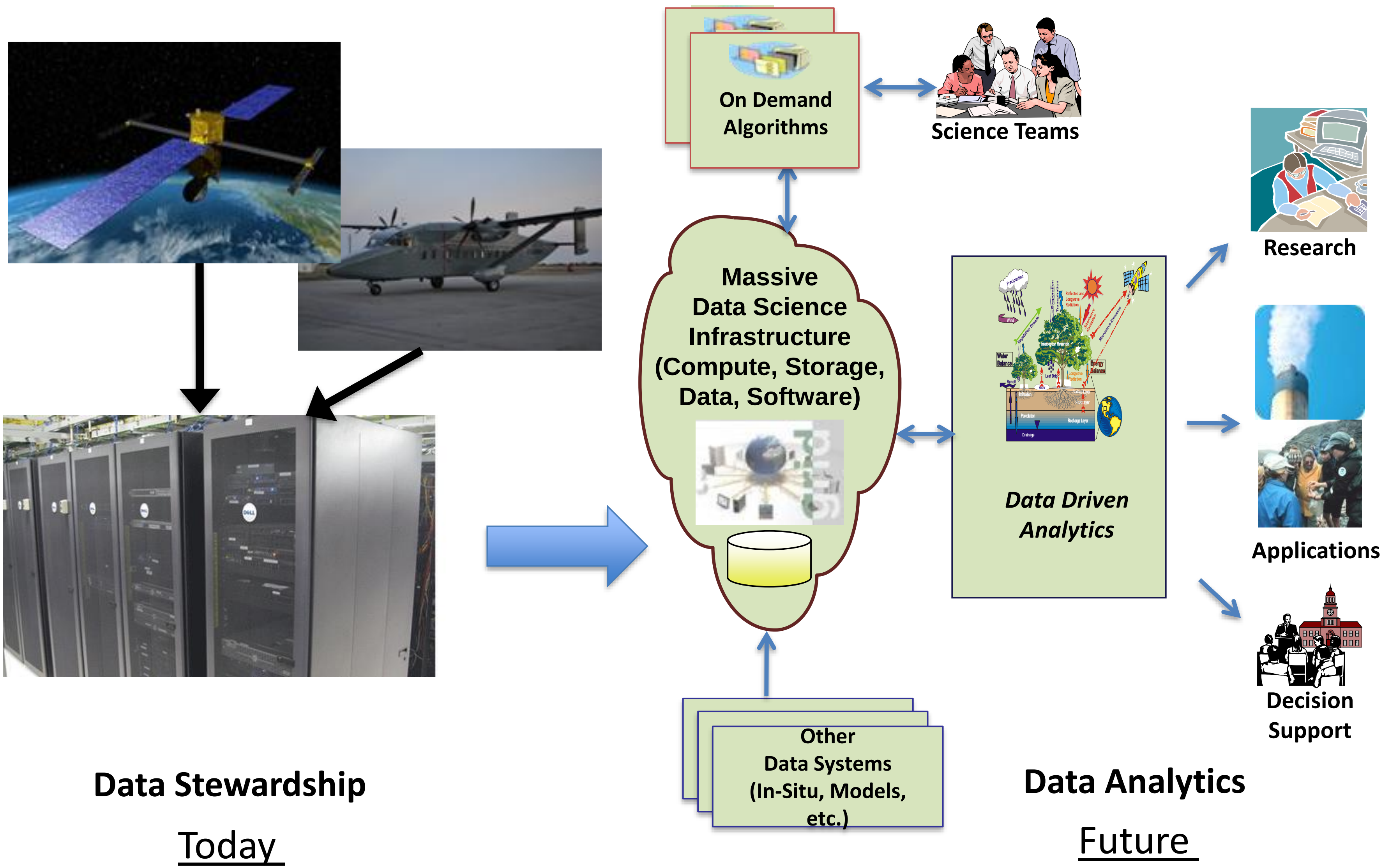
Curiosity (Mars Science Laboratory)
Processor: 200 MOPS BAE RAD750



HPSC (NASA STMD / AFRL)
Processor: 15 GOPS, extensible

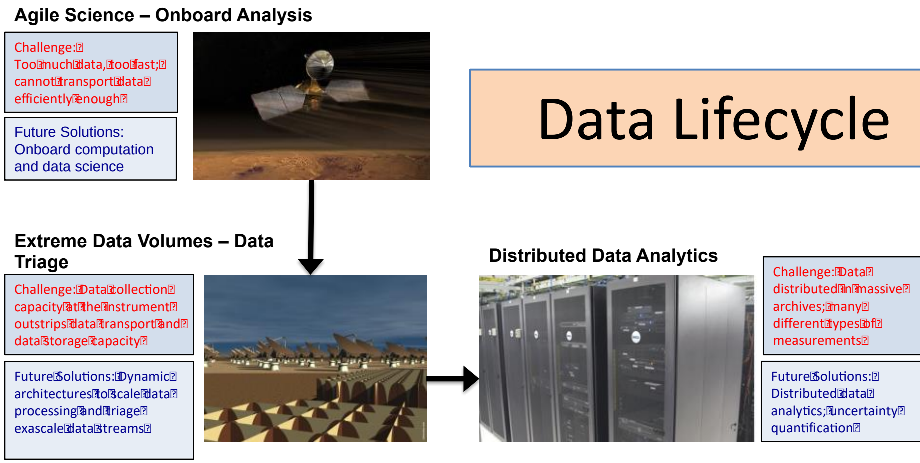
Future of Data Science at NASA

Shift Toward Data Analytics

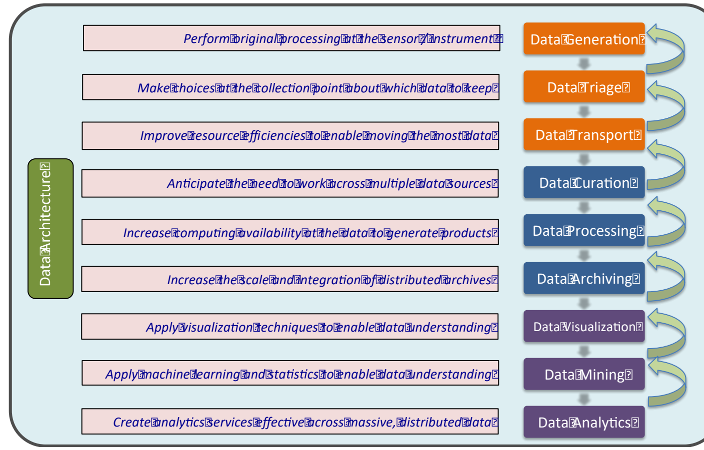


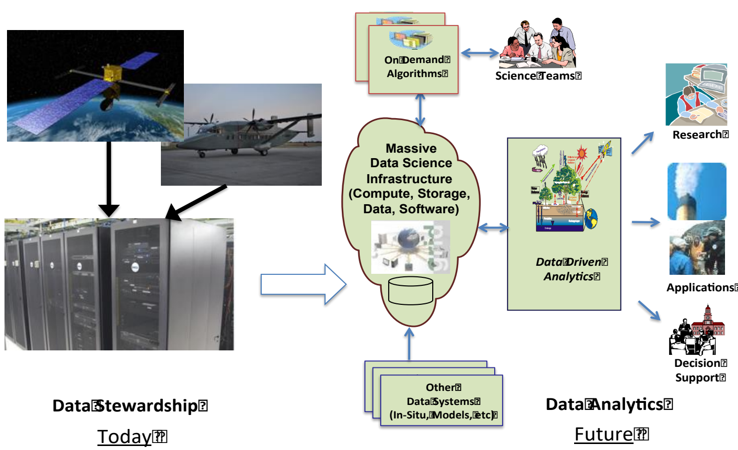
Credit: D. Crichton

Data Science Strategy
Guiding Principles

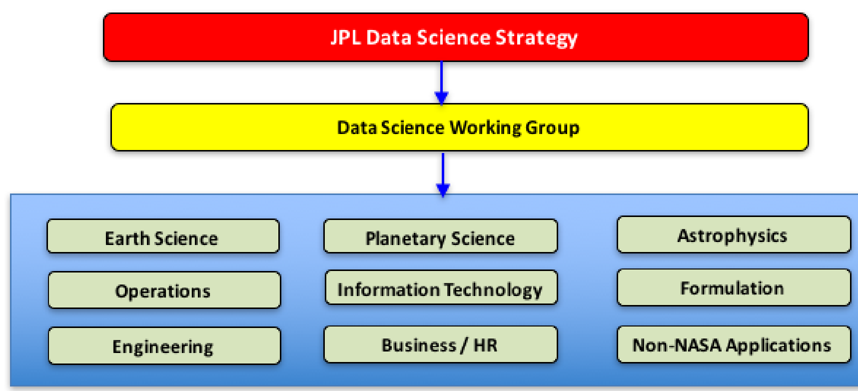


Data Lifecycle



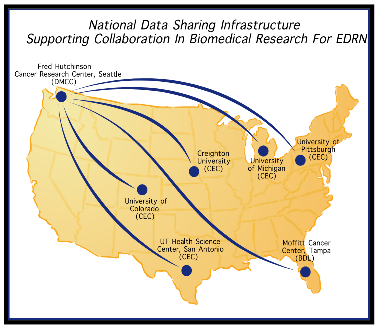


Data Ecosystem

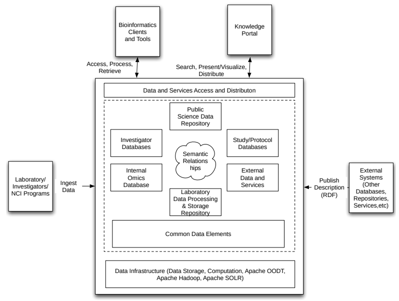


Cross-Cutting

Cross-Cutting Capabilities

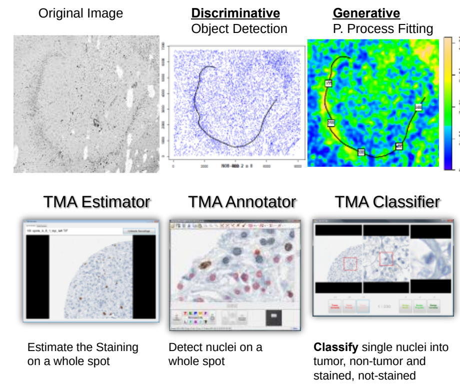


International Data Archive and Sharing Architectures



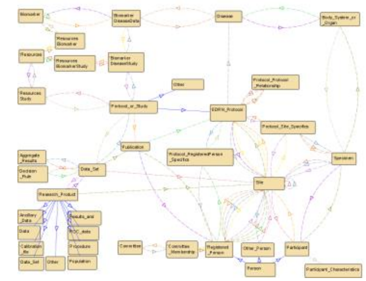
Big Data Infrastructures

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



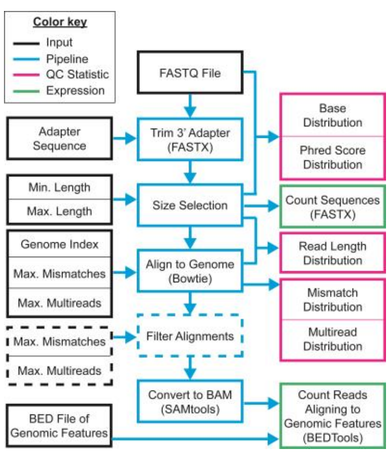
Intelligent Data Algorithms

(Machine Learning, Deep Learning)

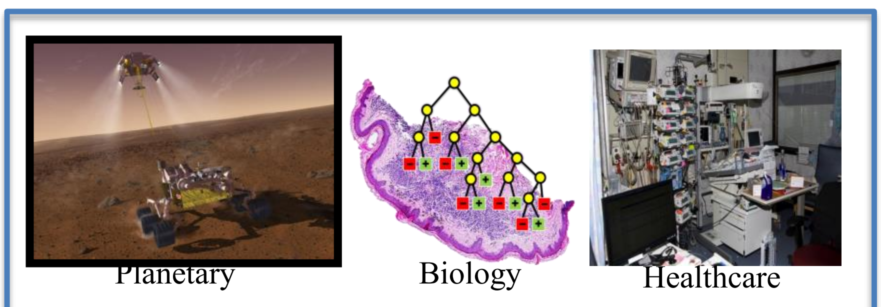


Common Data Elements & Information Models

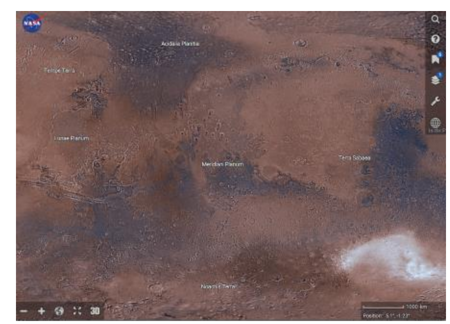
(discipline and common)



Analytical Data Pipelines



Great Opportunities for
Methodology Transfer and Collaboration



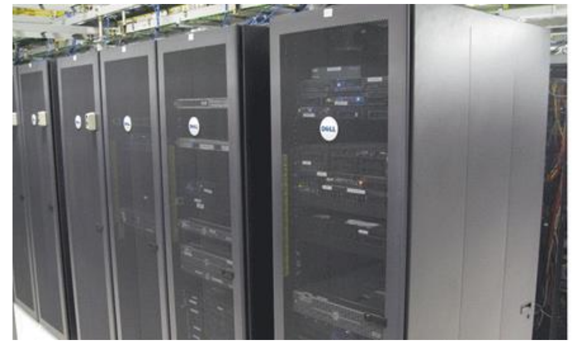
Visualization Techniques

Jet Propulsion Laboratory
California Institute of Technology

Highly Scalable Data-Driven Ground Systems

Intelligent Ground Stations

Data-Driven Discovery from Archives



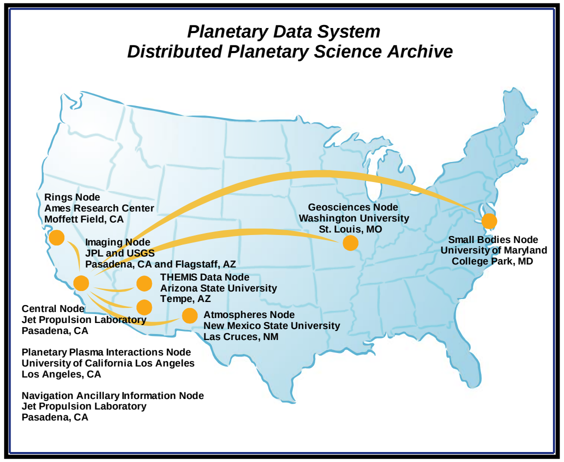
Machine Learning, Deep Learning, Intelligent Search, Data Fusion, Uncertainty Quantification, Attention Focusing, Decision Support, Interactive Visualization and Analytics

Mission Operations

Data Analytics and Decision Support

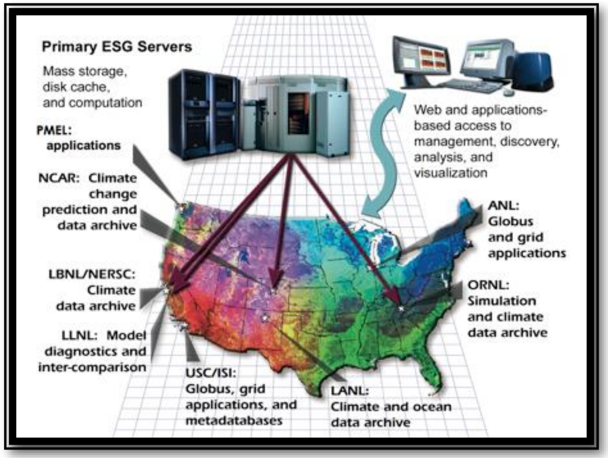
Jet Propulsion Laboratory
California Institute of Technology

Scientific Research Networks: *Access to Observations and Models*



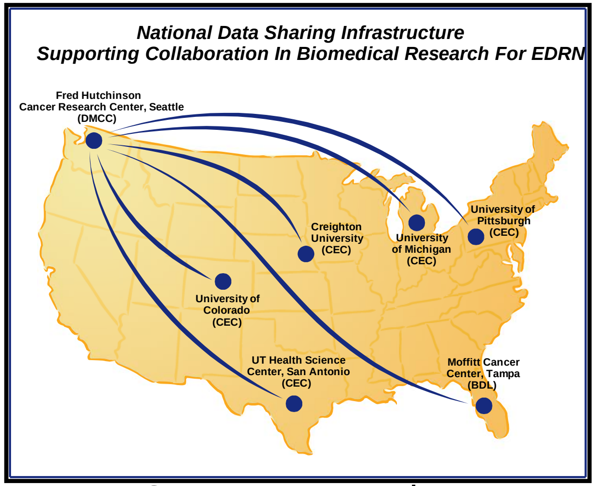
Solar System Exploration

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

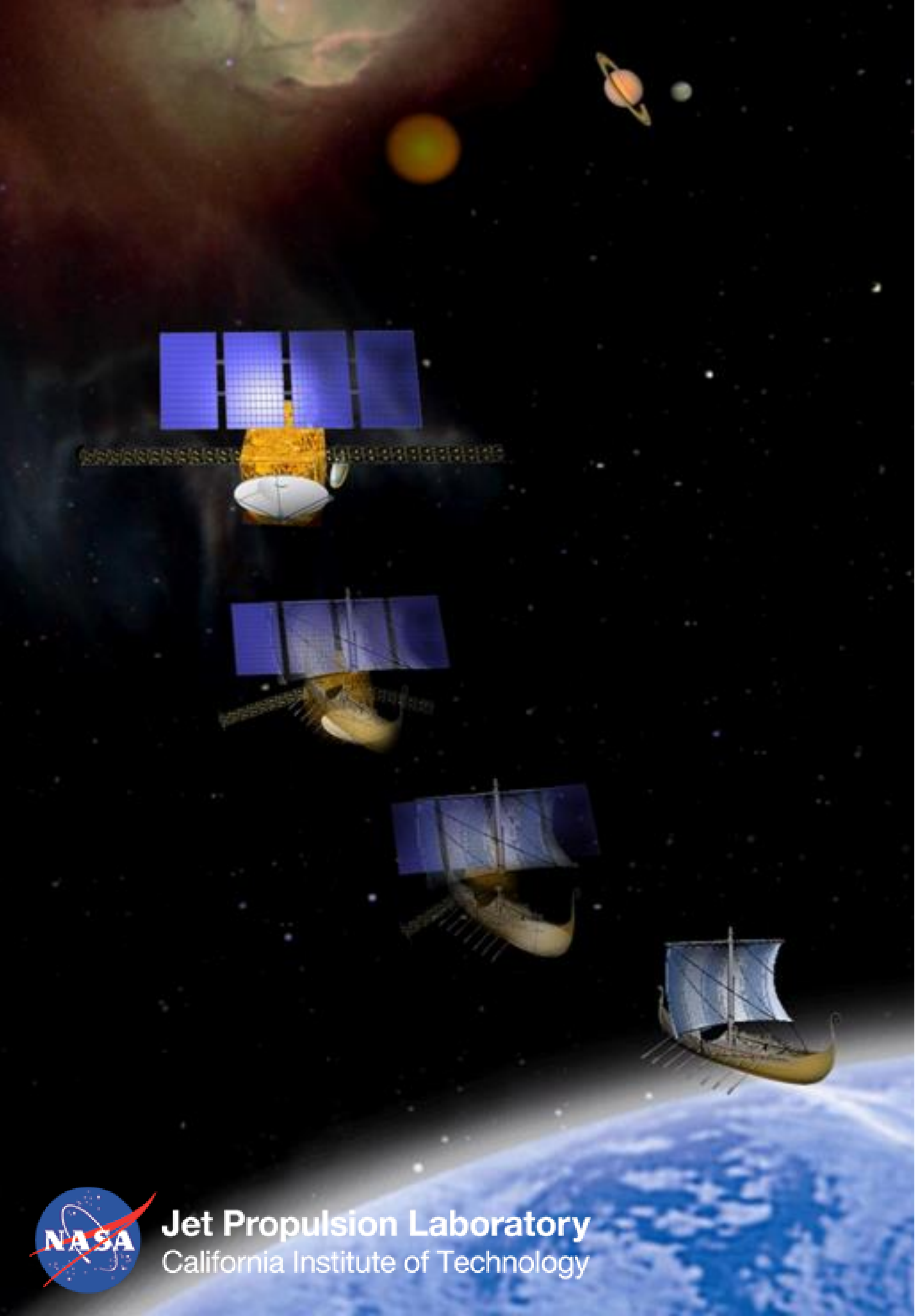


Climate Research

Earth Observation



Cancer Research



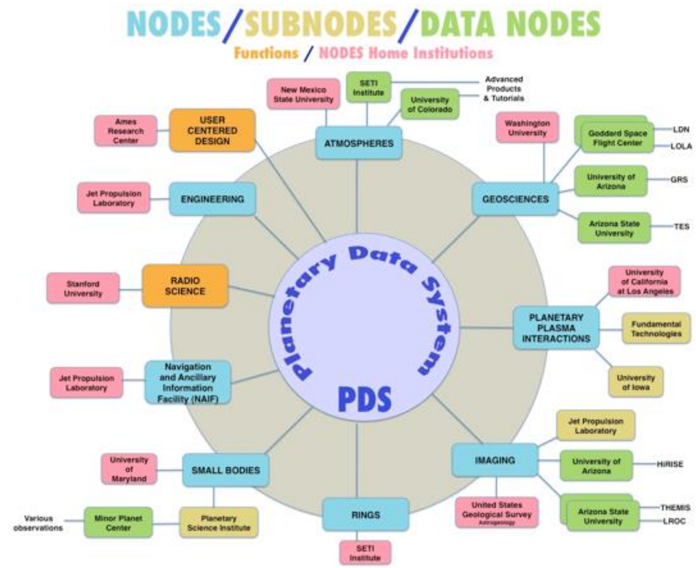
Planetary Data System

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.

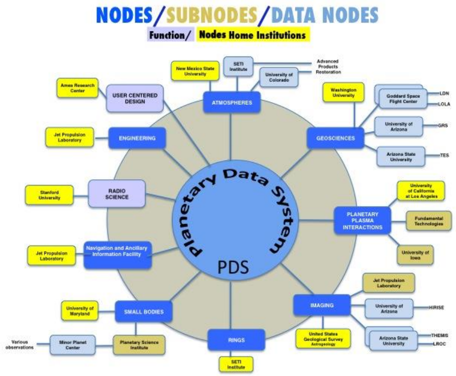
- Infrastructure: A highly distributed infrastructure with planetary science data repositories implemented at major government labs and academic institutions
  - System driven by a well defined planetary science information model
  - Over 1.3 PB of data
  - Movement towards international interoperability
  - Distributed federation of US nodes and international archives

- Being realized through PDS4



Jet Propulsion Laboratory
California Institute of Technology

Developing an Information-Model-Driven Strategy



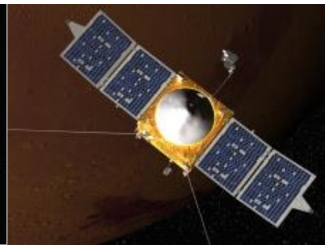
- An information model identifies and defines:
  - the elements to be processed
  - the entities that provide context for the elements
  - the relationships that provide meaning for the elements
- An information model increases productivity by
  - maximizing compatibility between systems (via reuse of common knowledge)
  - simplifying the process of design (via recurring patterns in the data)
  - promoting communication between individuals and teams working on the system (via common terminology and semantics)
- An information model is effective because
  - models are developed through extensive communication between domain experts, users, and model developers
  - models capture the points of view of users and experts who are most familiar with the domain
  - models serve as the basis for implementing the system software and processes

Credit: D. Crichton

Jet Propulsion Laboratory
California Institute of Technology

International Collaboration on Planetary Data Science through PDS4

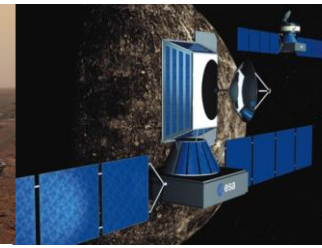
LADEE
(NASA)



MAVEN
(NASA)

Osiris-REx
(NASA)

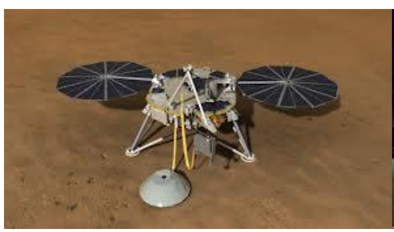
ExoMars
(ESA/Russia)



BepiColombo
(ESA/JAXA)

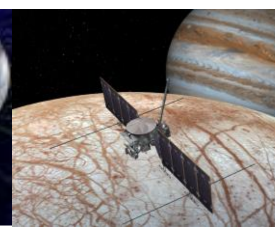
Mars 2020
(NASA)

Psyche
(NASA)



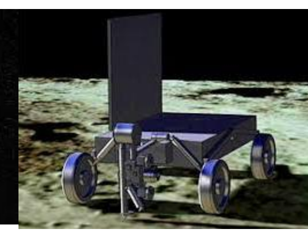
InSight
(NASA)

JUICE
(ESA)

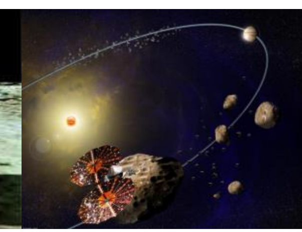


Europa
(NASA)

Hyabussa-2
(JAXA)

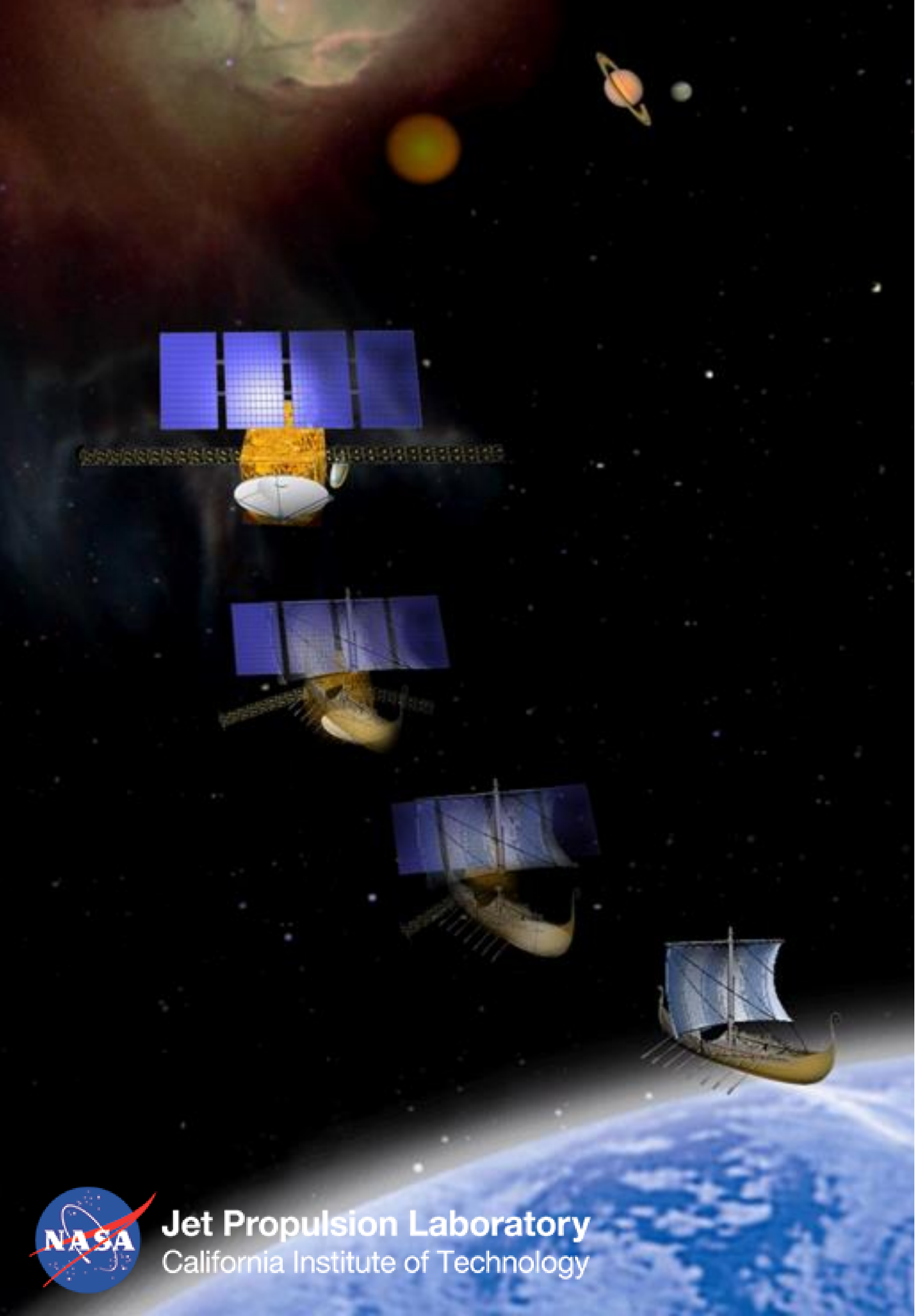


Chandrayaan-2
(ISRO)



Lucy
NASA

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

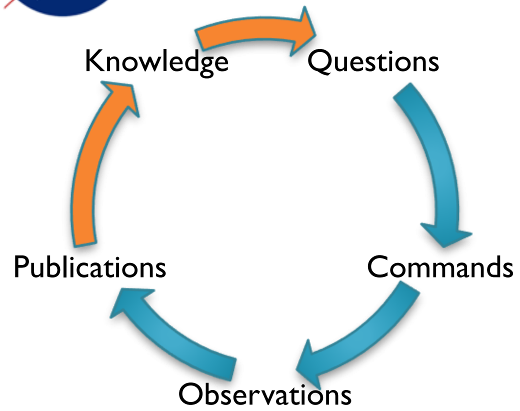


Looking Ahead

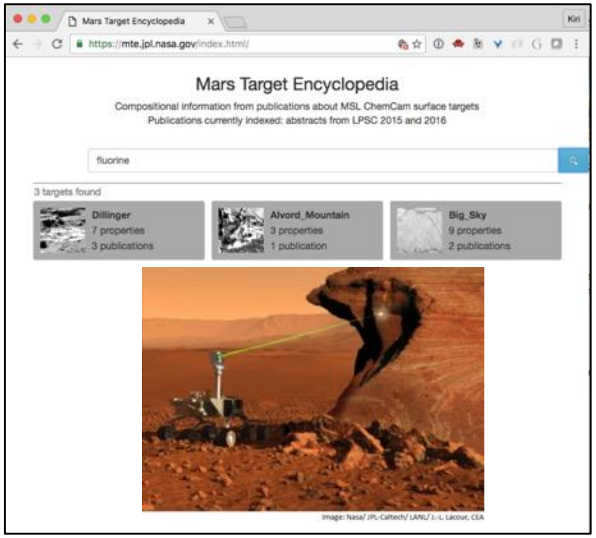
The Future of Analytics

Facilitate Science: Mars Target Encyclopedia

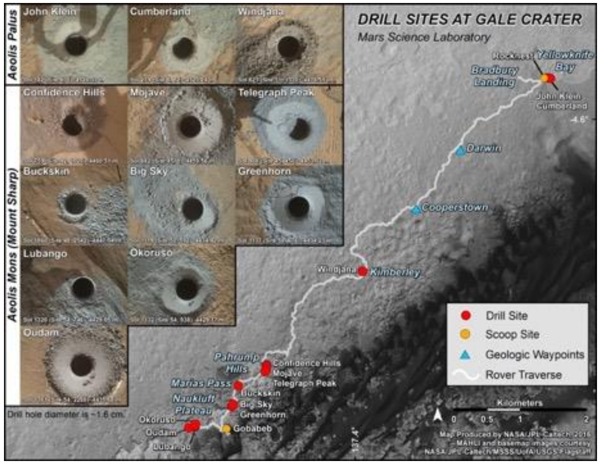
Jet Propulsion Laboratory
California Institute of Technology



Help Scientists Do What They Do Faster



- Connects data to (published) knowledge & maps
- Enables new searches:
  - What is known about Target X?
  - Scientific consensus?
  - Others like Target X?
  - What Targets are most unusual?
  - Show me publications that support statements



- Explore by geography, topic, publications, authors
- Speed scientific progress and exploration

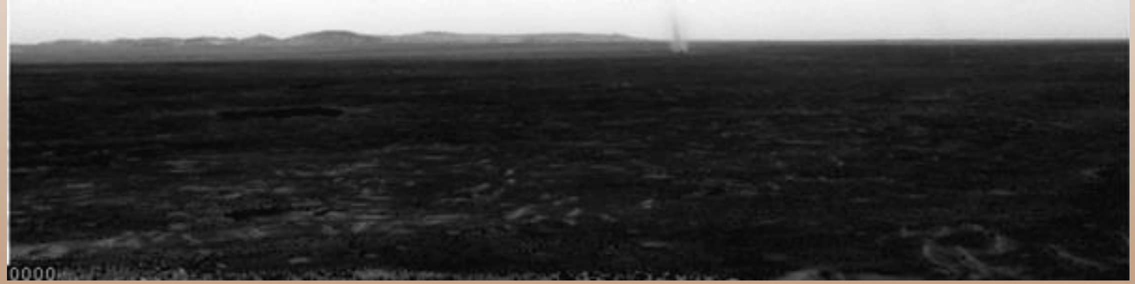
Credit: K. Wagstaff et al

Jet Propulsion Laboratory
California Institute of Technology

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



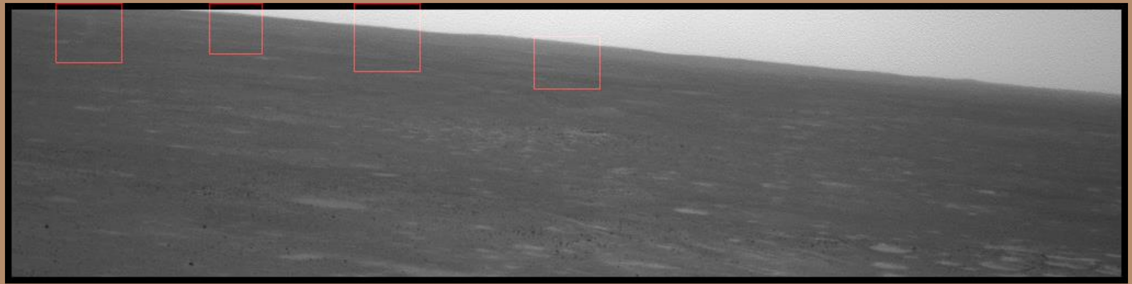
Spirit Sol 543 (July 13, 2005)

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

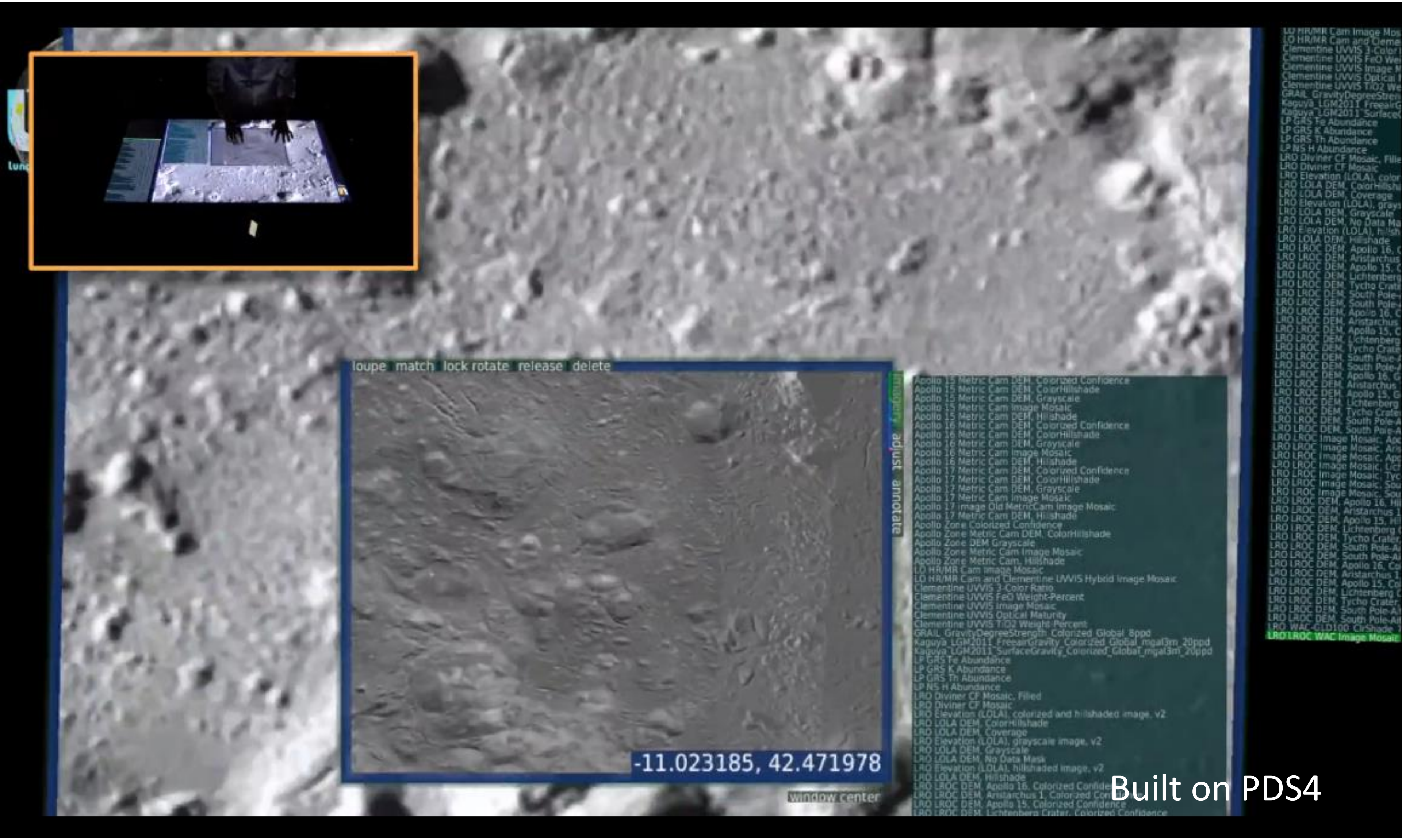
Benefit

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



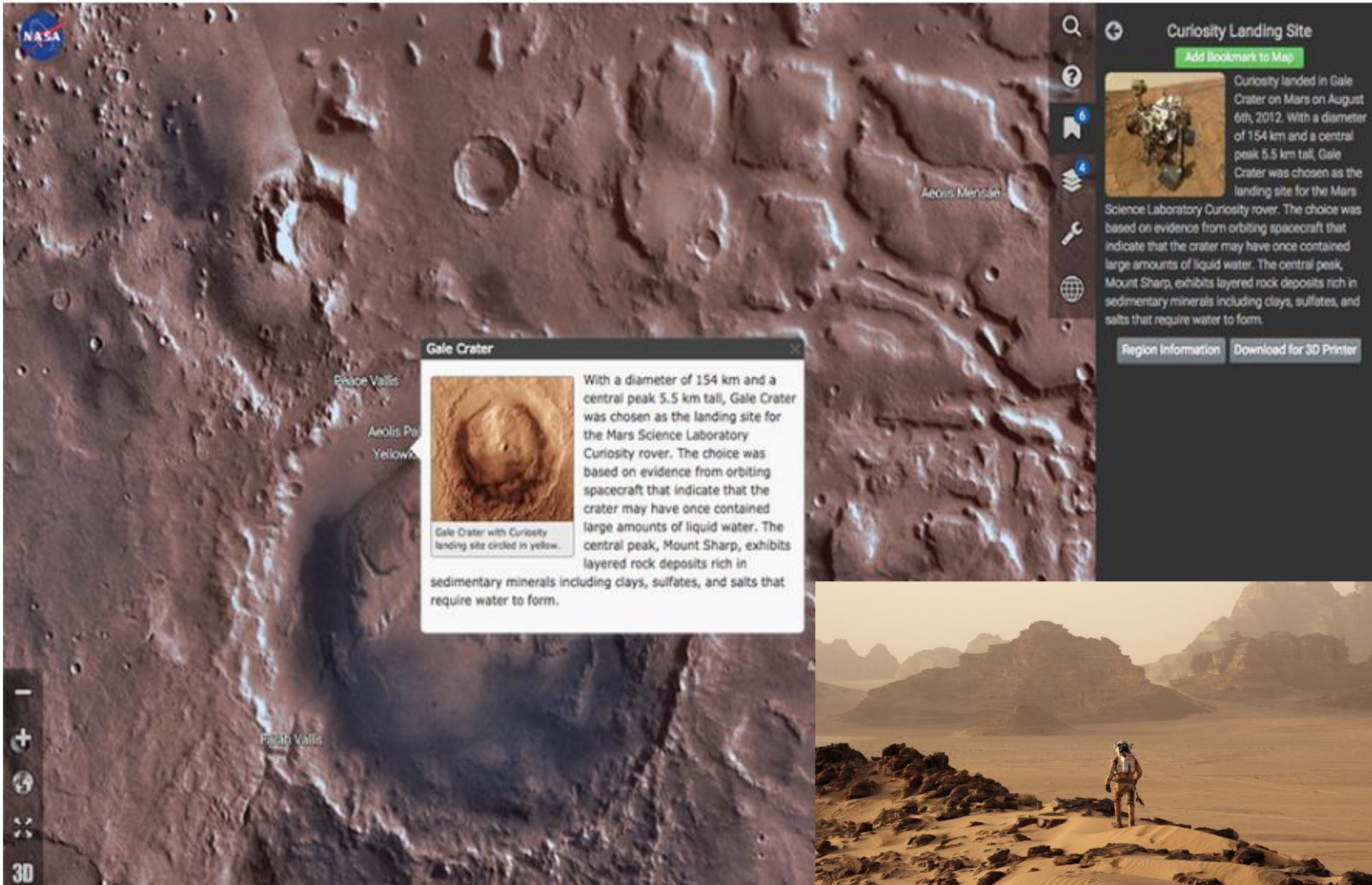
Credit: T. Estlin. B. Bornestein, A. Castano, J. Biesiadecki, L. Neakrase, P. Whelley, R. Greeley, M. Lemmon, R. Castano, S. Chien and MER project team

Lunar Trek
Exploring the Moon through Data Analytics



Credit: E. Law, S. Malhotra, G. Chang

Mars Trek: The Google Earth of Mars



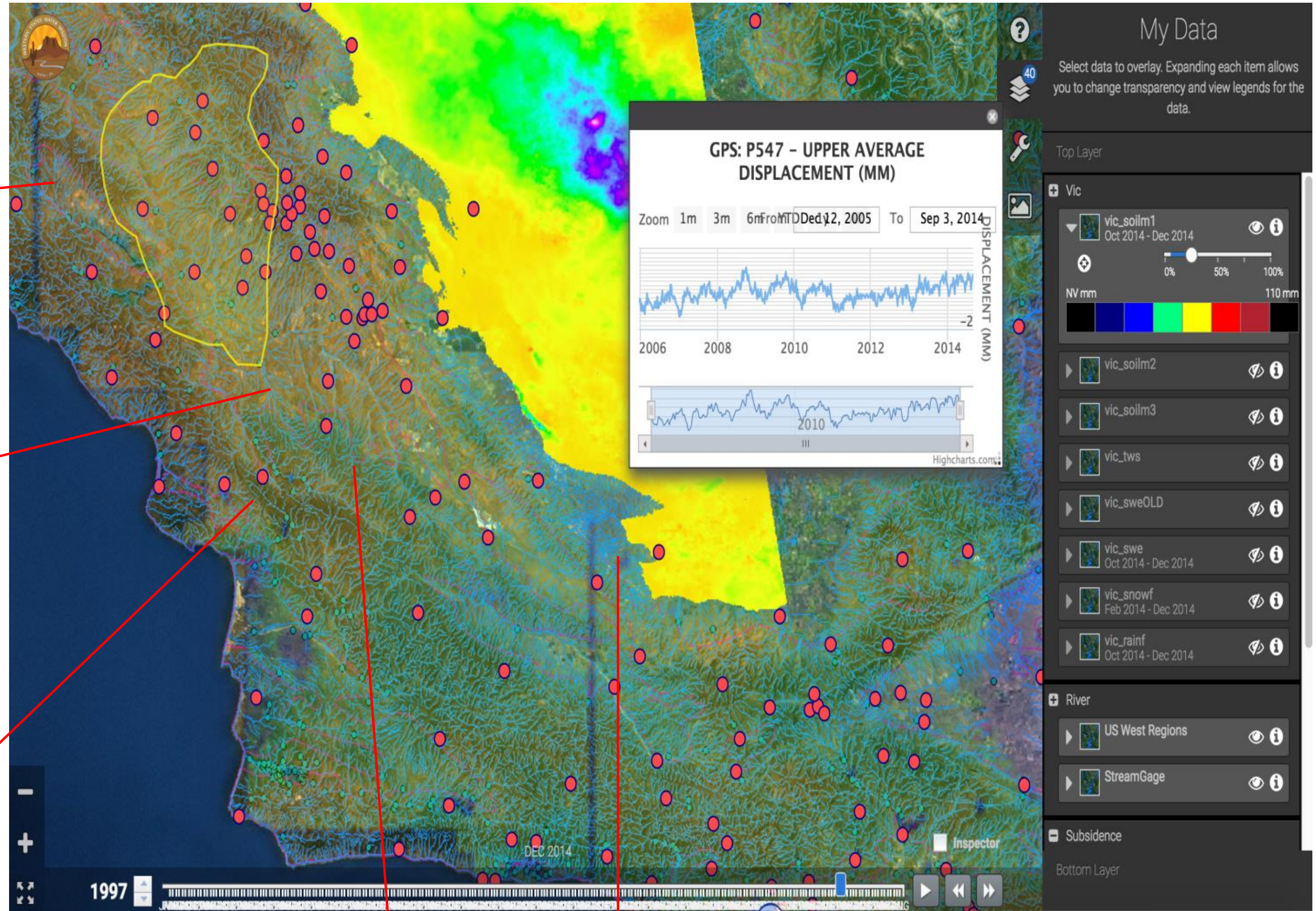
Credit: E. Law, S. Malhotra et al

Jet Propulsion Laboratory
California Institute of Technology

WaterTrek

User Defined Polygon

GPS



In-Situ: Stream Gage Sensors

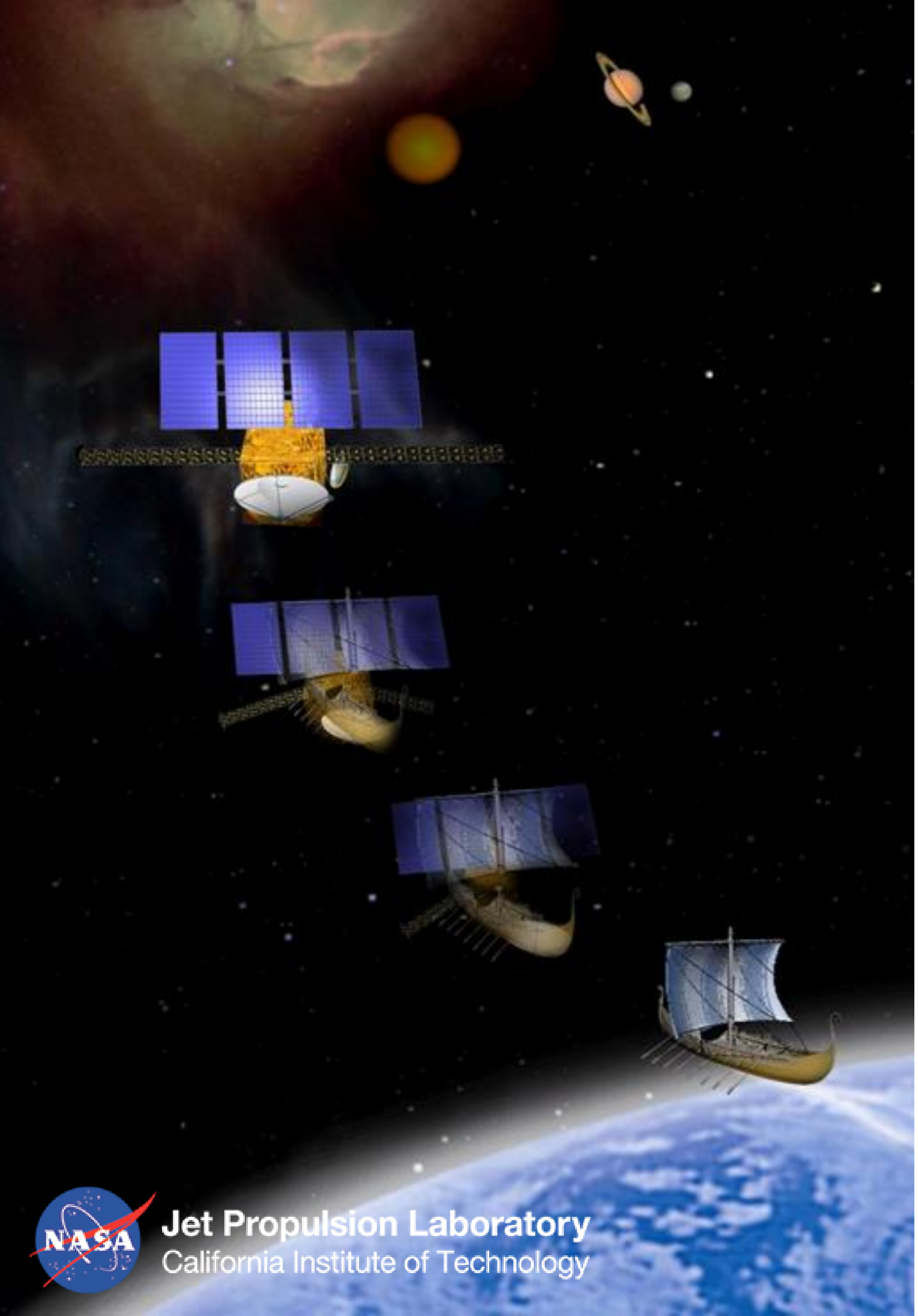
River Network

SAR derived Subsidence

Model Output Soil Moisture

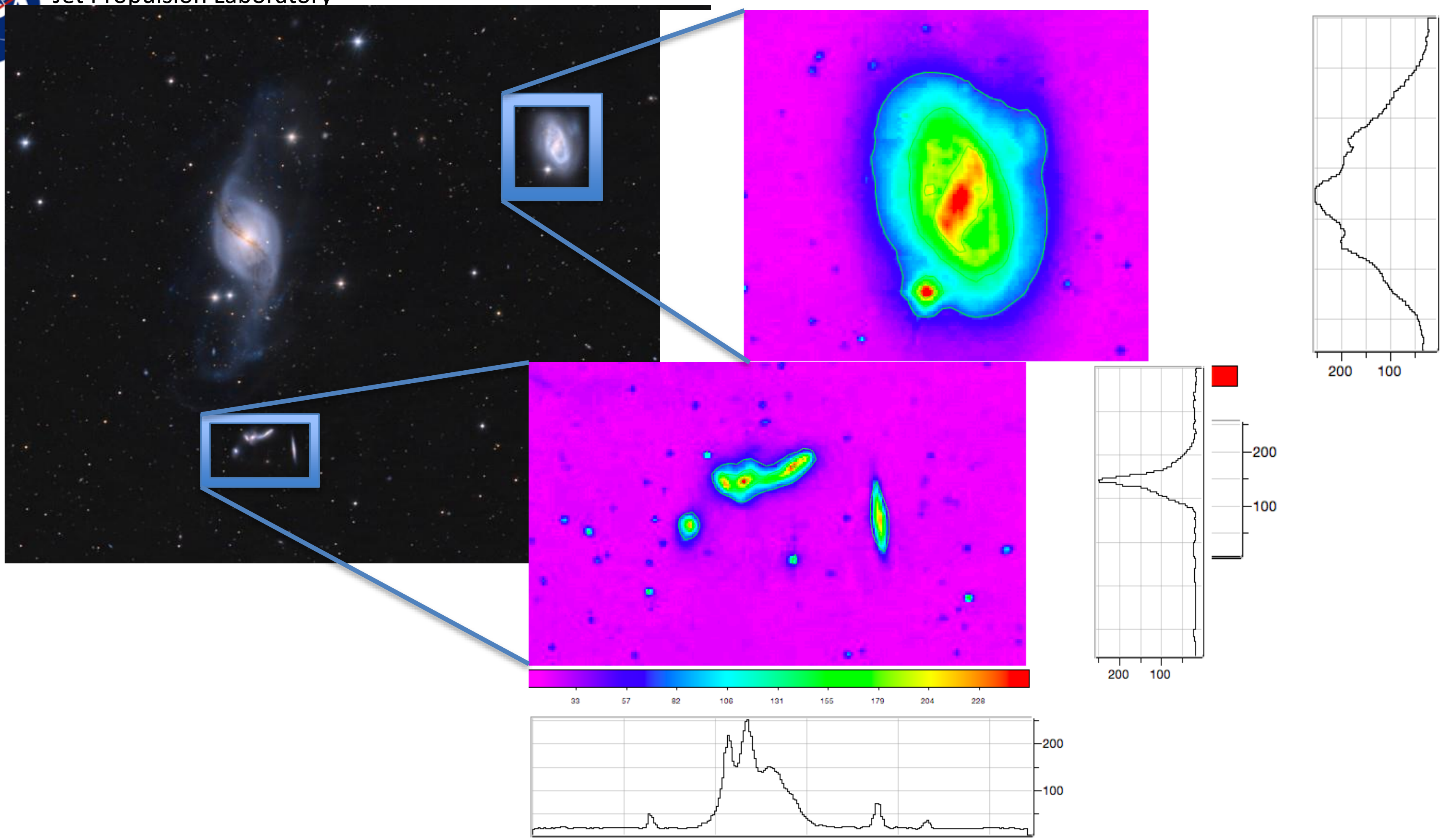
Fusing In-situ, Air-borne, Space-borne and model generated data using visualization and a big data analytics engine

Credit: S. Malhotra et al

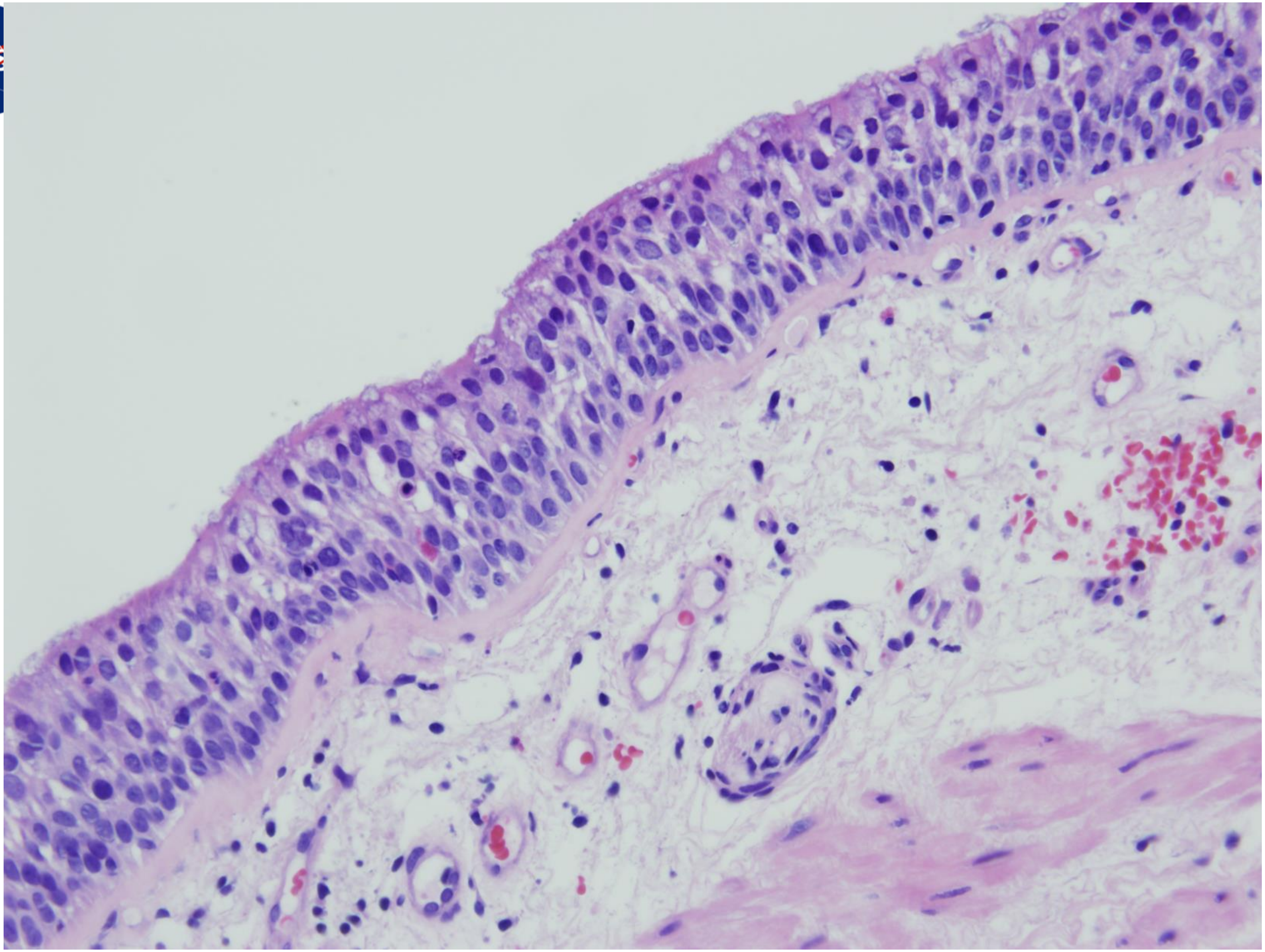


Looking Around

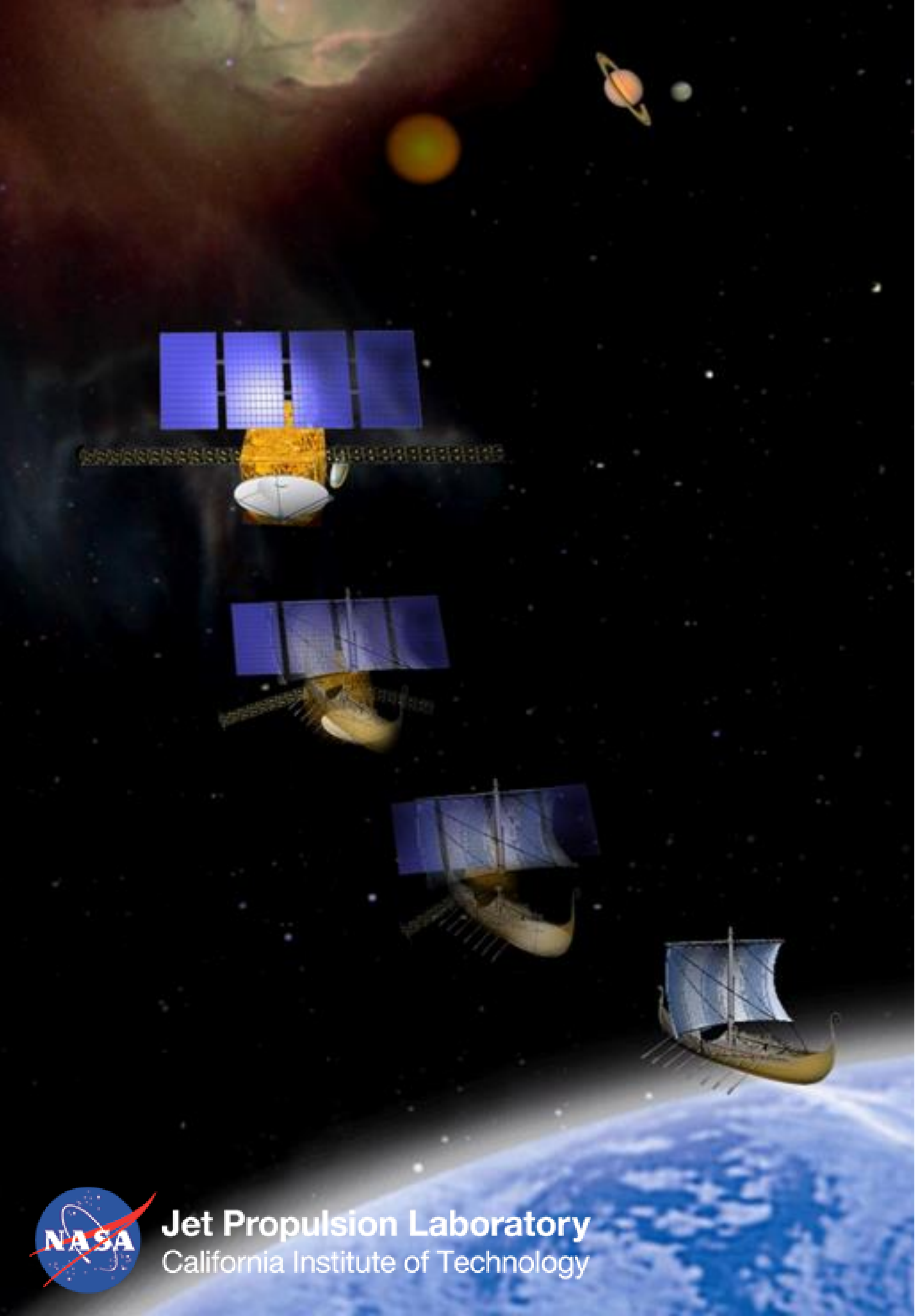
Methodology Transfer



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

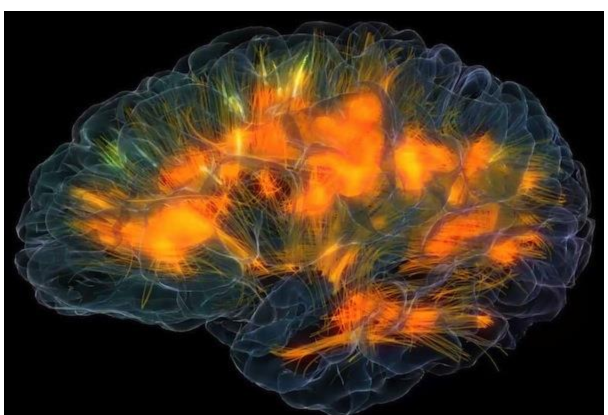
Caltech-JPL
Partnership in Data Science

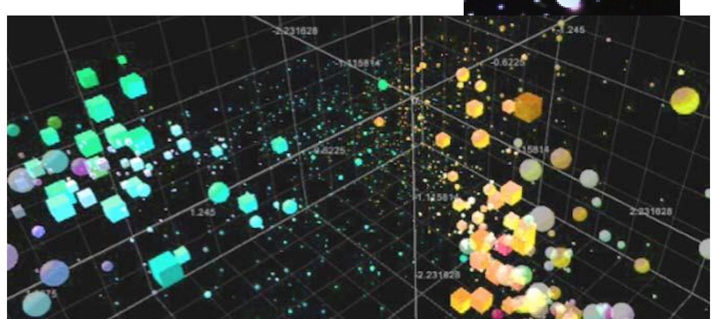
Jet Propulsion Laboratory
California Institute of Technology

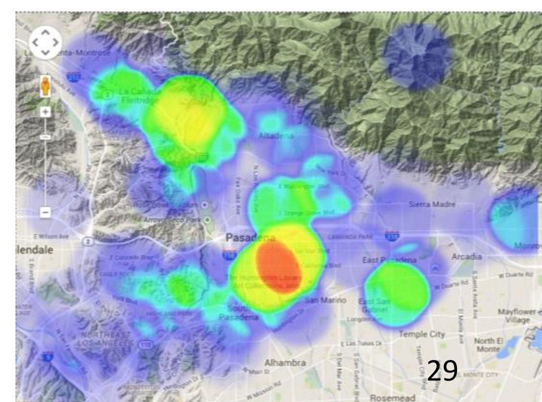
- Center for Data-Driven Discovery on campus/Center for Data Science and Technology at JPL
- From basic research to deployed systems ~10 collaborations
  - Leveraged funding from JPL to Caltech; from Caltech to JPL
- Virtual Summer School (2014) has seen over 25,000 students

CENTER FOR DATA-DRIVEN DISCOVERY

CDST





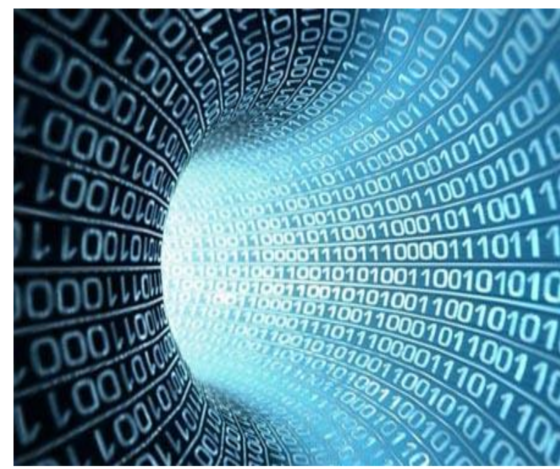


Jet Propulsion Laboratory
California Institute of Technology

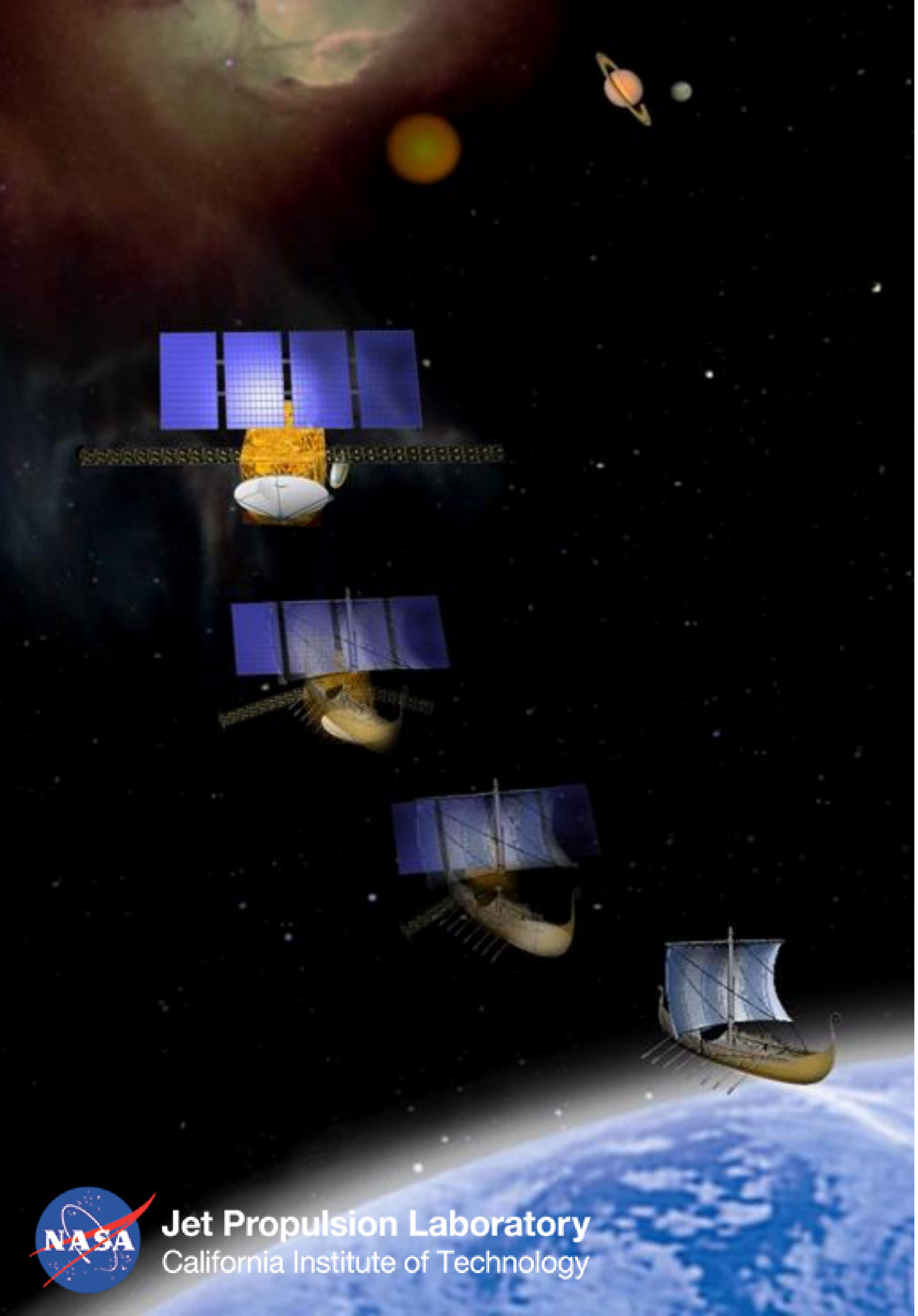
Go Forward Strategy

- 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!


Jet Propulsion Laboratory
California Institute of Technology

Questions?

Jet Propulsion Laboratory
California Institute of Technology

Dare Mighty Things