Big Data and NASA Space Sciences

Daniel Crichton
Leader, Center for Data Science and Technology
Manager, Data Science Office
Jet Propulsion Laboratory
California Institute of Technology

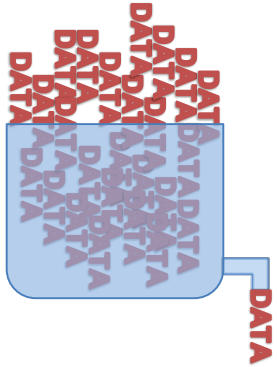
November 2, 2017

Jet Propulsion Laboratory
California Institute of Technology

Terms: Big Data and Data Science

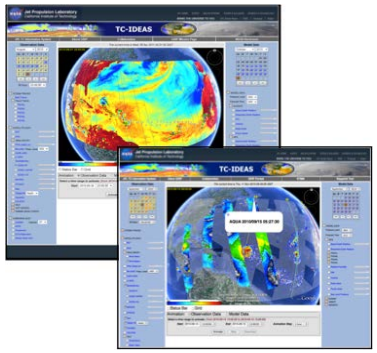
Big Data – Statement of the problem

When needs for data collection, processing, management, use, and analysis go beyond the capacity and capability of available methods and software systems. These constraints are often defined by volume, variety, velocity, veracity, etc.



Data Science – Statement of the techniques

S*calable* architectural approaches, software, and algorithms which alter the paradigm by which data is collected, managed and used.

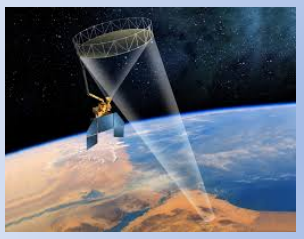


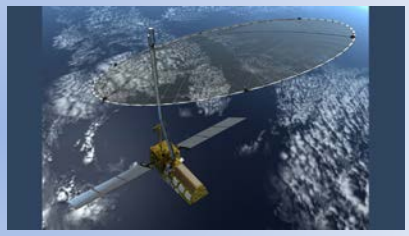
Jet Propulsion Laboratory
California Institute of Technology

NASA Big Data Landscape

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

Observational Platforms and Flight Computing

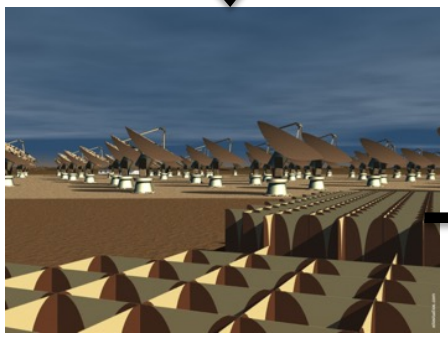




SMAP (Today): 485 GB/day NI-SAR (2020): 86 TB/day

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

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



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

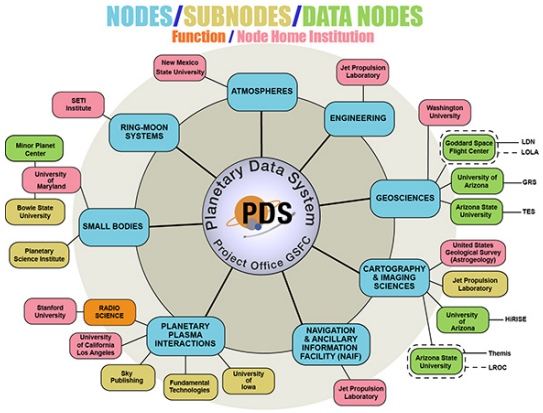
(3) Data distributed in massive archives; many different types of measurements and observations

Ground-based Mission Systems

NASA OCT Roadmap TA-11

Jet Propulsion Laboratory
California Institute of Technology

NASA Archives: Access to Data*



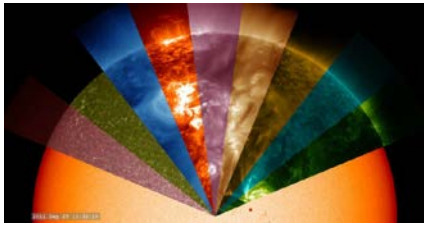
Planetary Science

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

Multiple Data Centers

Astronomy

Earth Observation



Multiple Data Centers

Heliophysics

*~20 PB of data observational data available today

Jet Propulsion Laboratory
California Institute of Technology

NASA Science and Big Data Today

On Board Processing

Science Teams

How do these connect?

Comm Network

Data Acquisition and Command

Mission Operations

Instrument Operations

Processing

Ground Data Systems

Science Data Processing
L0B
L1
L2
L3
L4

Science Data Systems

Science Data Management Archive & Distribution

Big Data Infrastructure (Data, Algorithms, Machines)

?

Research

Outreach

Applications

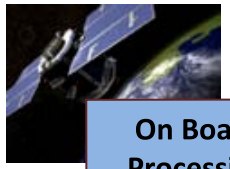
Focus on generating, capturing, managing big data

Focus on using/analyzing big data

Future of Data Science at NASA

Enabling a Big Data Research Environment

Jet Propulsion Laboratory
California Institute of Technology



On Board Processing

Comm Network

Data Acquisition and Command

Mission Operations

Instrument Operations

Processing

Ground Data Systems

Data Capture

Instrument Data Systems

Airborne Data

NASA Data Archives

On Demand Algorithms

Science Teams

Big Data Infrastructure (Data, Algorithms, Machines)

Other Data Systems (In-Situ, Other Agency, etc.)

Data Analytics Centers

Data Analysis

Research

Applications

Decision Support

<u>Reducing Data Wrangling</u>: "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.**

Jet Propulsion Laboratory
California Institute of Technology

Key Big Data Opportunities

- Support scalability to capture and analyze NASA observational data
  - All sciences in the multi-PB range; Earth science in 100 PB range in 5 years.
  - Leverage commercial cloud computing, as appropriate
- Apply data-driven approaches across the entire data lifecycle
  - From onboard computing to data analysis
- Increase access, integration and use of archival data
  - Across missions, instruments, and data centers
- Increased data science services for on-demand, interactive visualization and analytics
  - Target specific analytics

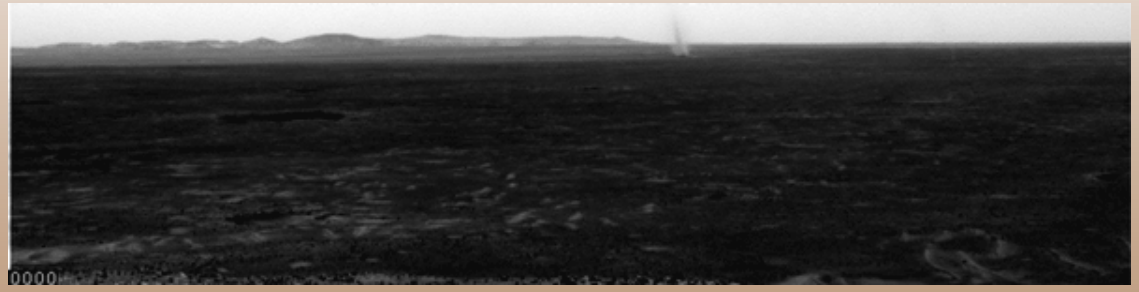
The NASA Advisory Council has created a Big Data Task Force under the Science Subcommittee (C. Holmes, Chair). Last meeting Nov 1-3 at JPL.

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



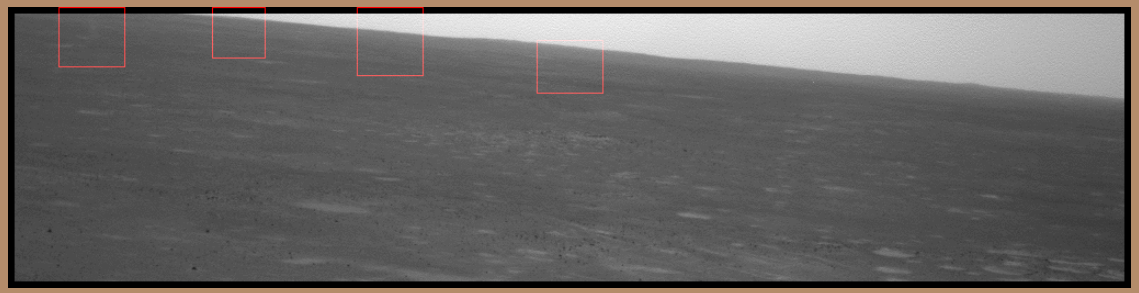
Spirit Sol 543 (July 13, 2005)

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***

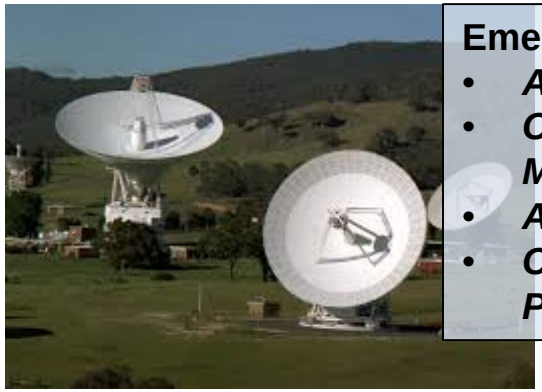


Credit: T. Estlin. B. Bornestein, A. Castano, J. Biesiadecki, L. Neakrase, P. Whelley, R. Greeley, M. Lemmon, R. Castano, S. Chien and MER project team

Jet Propulsion Laboratory
California Institute of Technology

Opportunities and Use Cases Across the Ground Environment

Intelligent Ground Stations



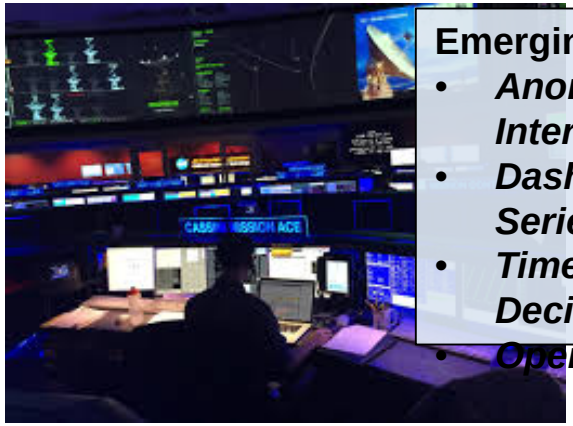
Emerging Solutions
- *Anomaly Detection*
- *Combining DSN & Mission Data*
- *Attention Focusing*
- *Controlling False Positives*

Data-Driven Discovery from Archives

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

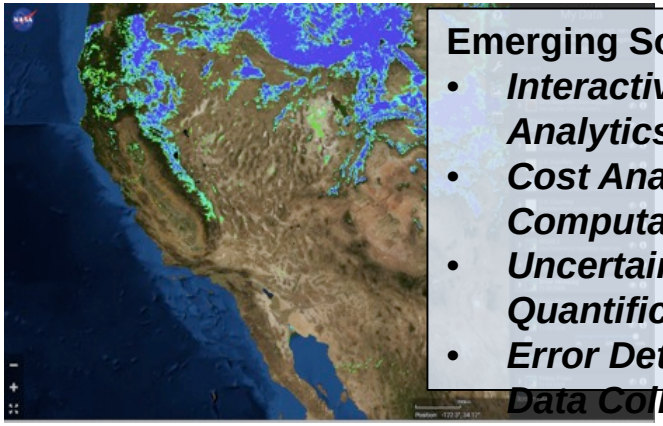
Technologies: Machine Learning, Deep Learning, Intelligent Search, Data Fusion, Interactive Visualization and Analytics

Agile MOS-GDS



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

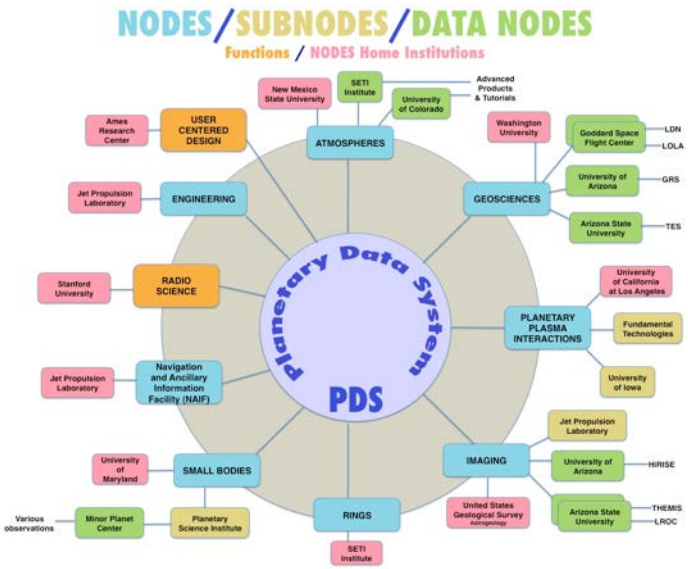
Data Analytics and Decision Support



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

Jet Propulsion Laboratory
California Institute of Technology

International Adoption of an Open Planetary Data Approach



NASA Planetary Data System

Big Data Challenges

Variety of planetary science disciplines, moving targets, and data
Volume of data returned from missions (~1.3 PB)
Federation of systems and data internationally

Big Data Solution (PDS4)

Governed by a well defined planetary science data model governing the variety of data types
Co-developed with international partners
Adoption occurring on all upcoming international planetary missions
Movement towards international interoperability across archives

International Planetary Data Alliance

Formed in 2006 to support the development of compatible archives world-wide (common metadata)
Steering Committee with participation from 20 agencies
Adoption of PDS4 by NASA, ESA, JAXA, ISRO, UAE, and others for upcoming missions

INTERNATIONAL PLANETARY DATA ALLIANCE
IPDA

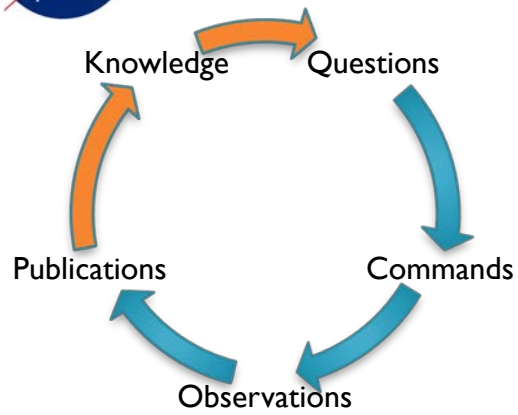
Started in 2006

Moving towards an international data platform for planetary science research

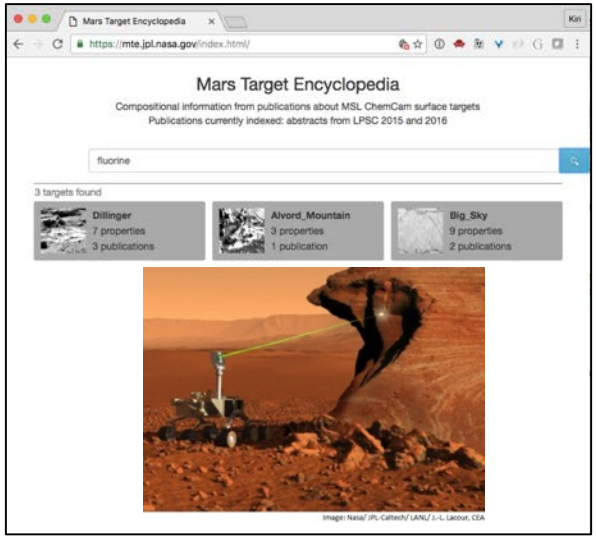
Facilitate Science: Mars Target Encyclopedia

Jet Propulsion Laboratory
California Institute of Technology

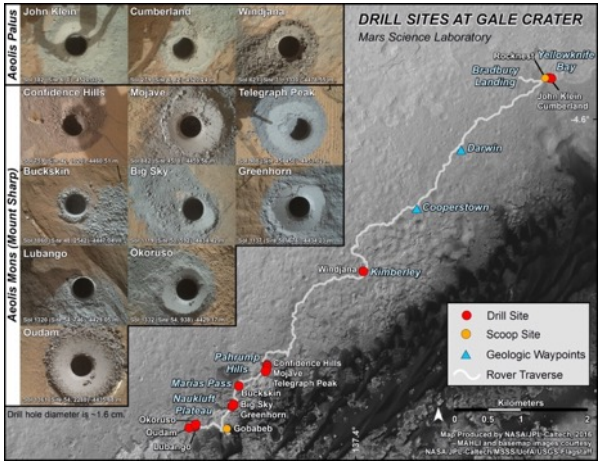
Dr. Kiri Wagstaff et al



Help Scientists Do What They Do Faster



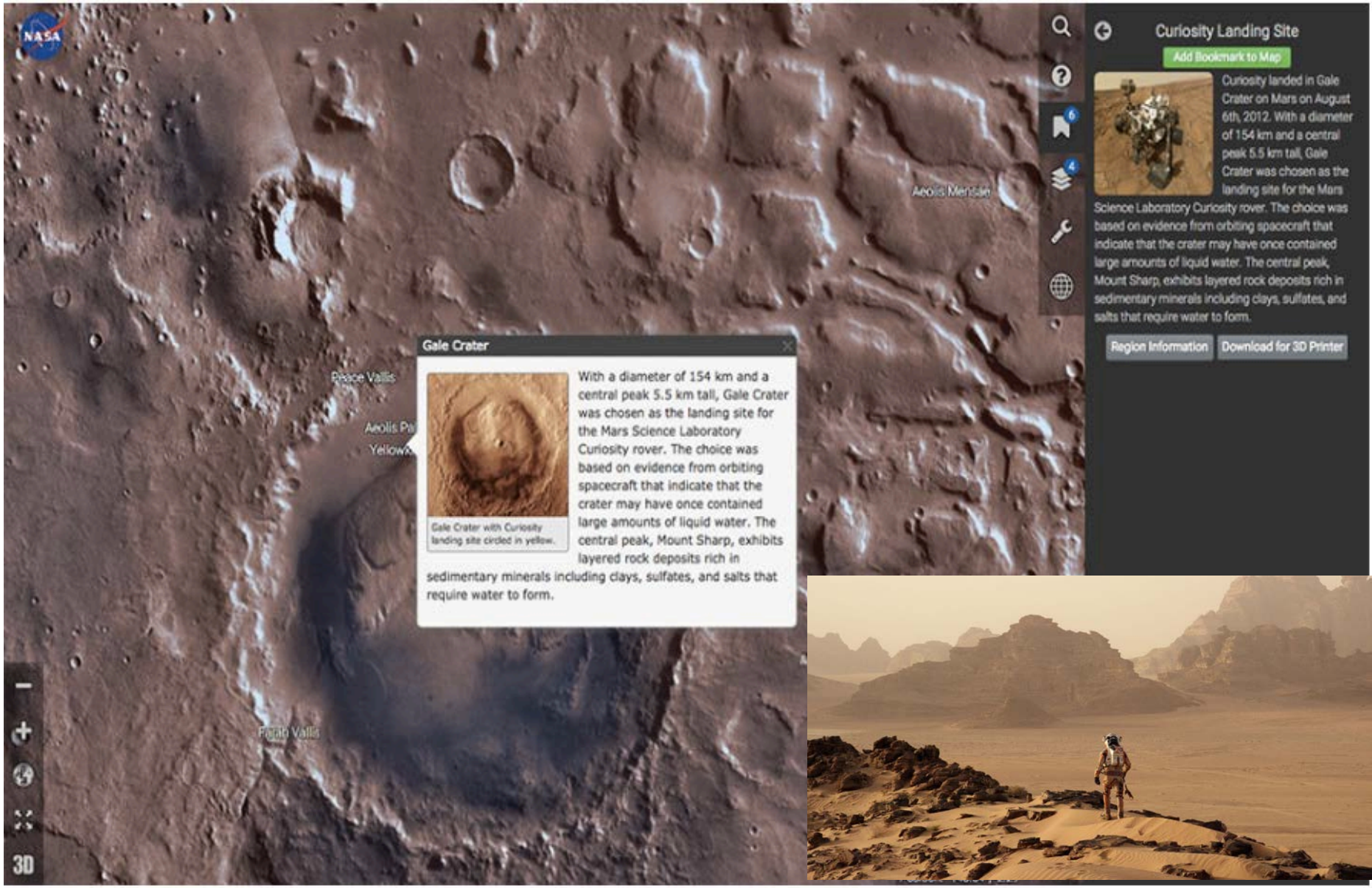
- Connects data to (published) knowledge & maps
- Enables new searches:
  - What is known about Target X?
  - Scientific consensus?
  - Others like Target X?
  - What Targets are most unusual?
  - Show me publications that support statements



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

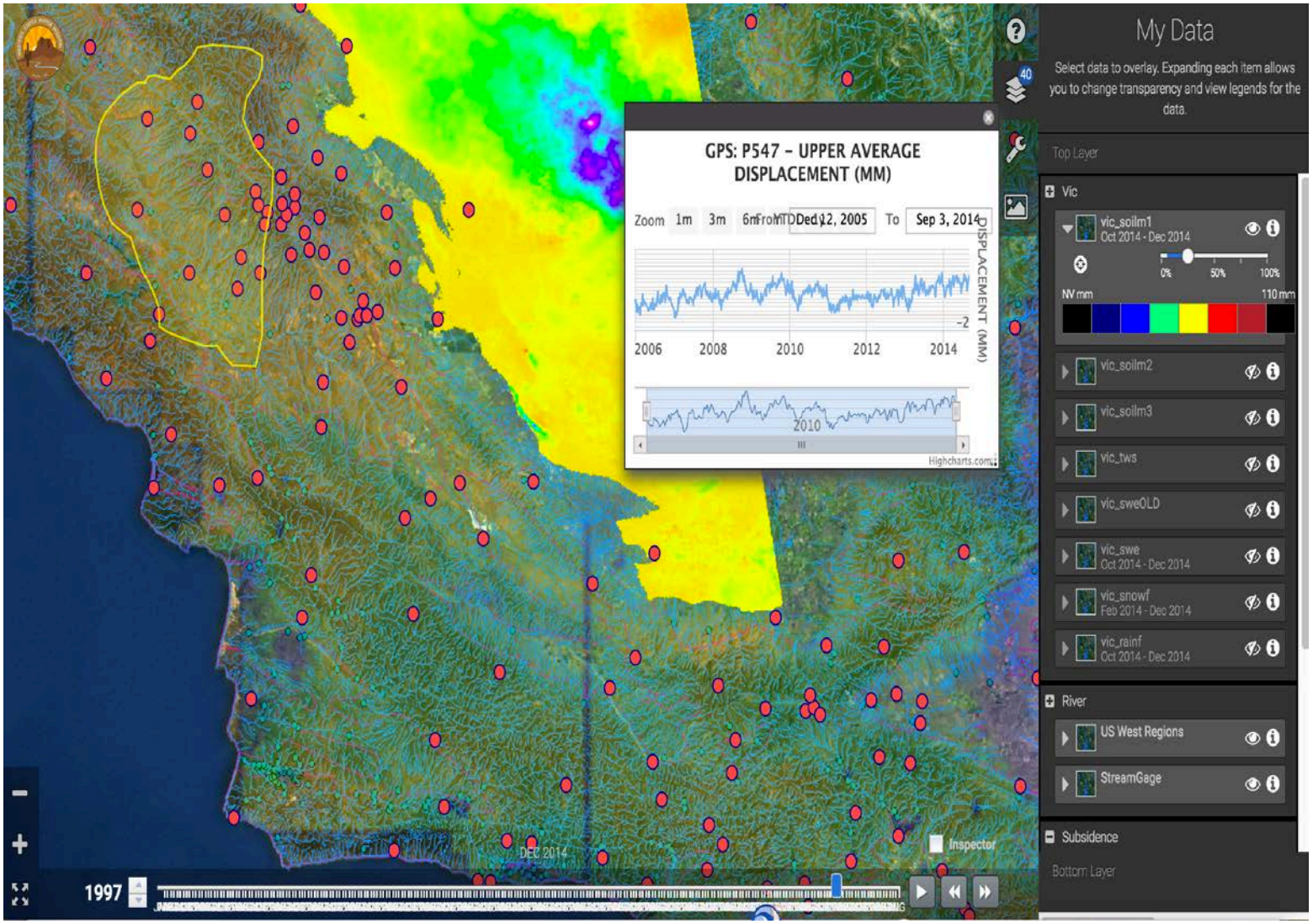
Jet Propulsion Laboratory
California Institute of Technology

Mars Trek: The Google Earth of Mars



Jet Propulsion Laboratory
California Institute of Technology

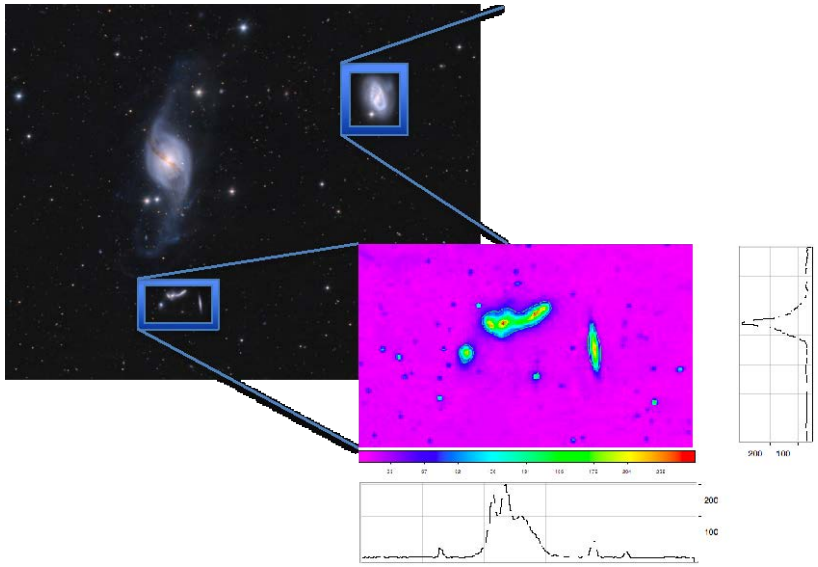
Western States Water Mission at JPL

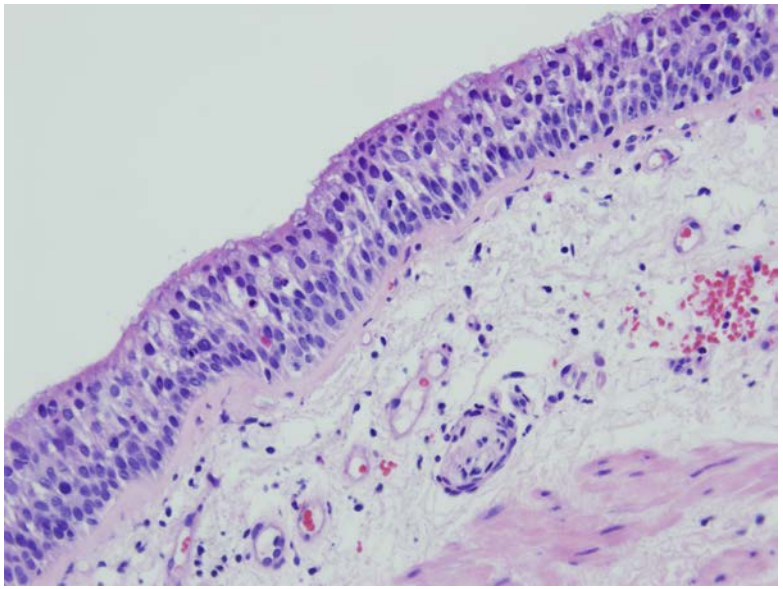


Jet Propulsion Laboratory
California Institute of Technology

Crossing Disciplines to Support Scientific Research through Big Data

- Development of an advanced Knowledge System to *capture*, *share* and support *reproducible analysis* for biomarker research
  - Genomics, Proteomics, Imaging, etc data types of data
- NASA-NCI partnership, leveraging informatics and data science technologies from space science
  - From modeling and managing massive data systems to machine learning





Jet Propulsion Laboratory
California Institute of Technology

Caltech-JPL
Partnership in Data Science

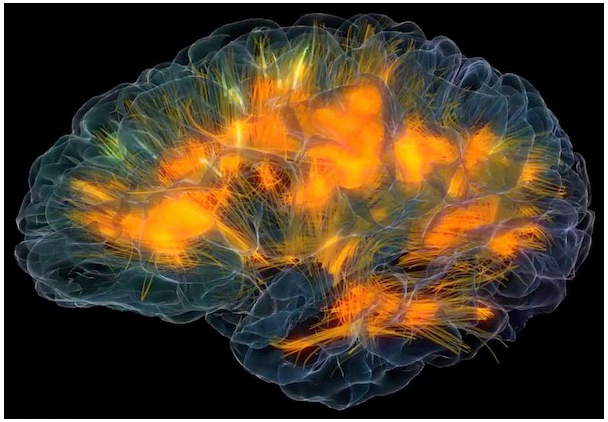
- Center for Data-Driven Discovery on campus/Center for Data Science and Technology at JPL

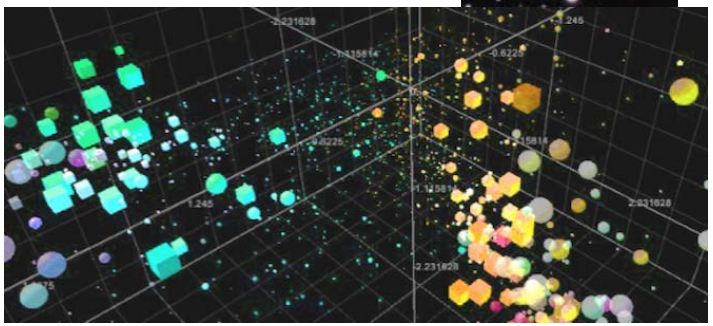
- From basic research to deployed systems ~10 collaborations
  - Leveraged funding from JPL to Caltech; from Caltech to JPL

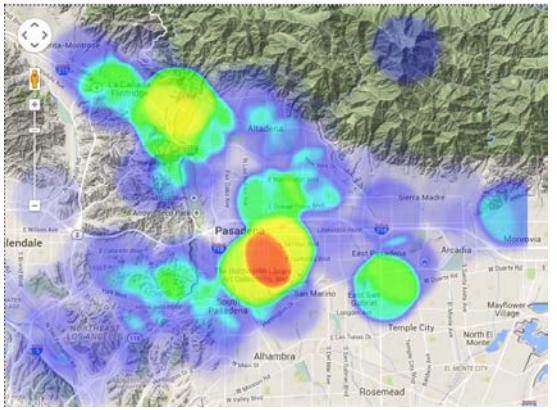
- Virtual Summer School (2014) has seen over 25,000 students

CENTER FOR DATA-DRIVEN DISCOVERY

CDST





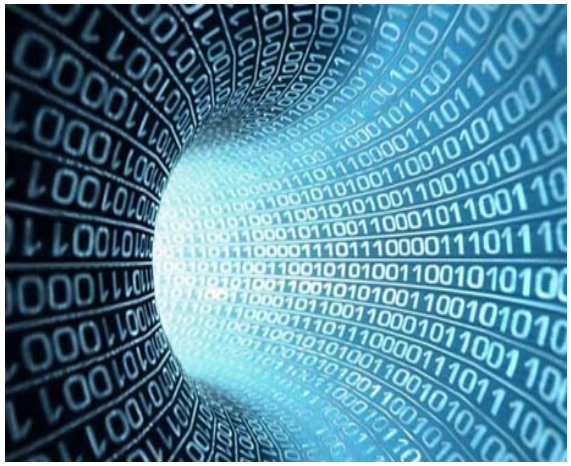


Recommendations

- Rich opportunities to leverage data science
  - Use the Mission-Science Data Lifecycle to organize a vision for data and computing
  - Great opportunities for methodology transfer across disciplines
  - Increasing need for training and investments

What do we do with all this data?

- Evolve from archiving to data-driven approaches and infrastructure
  - Drive broad, international data ecosystems
  - Increase use of data-driven approaches to gain insight and understanding
  - Develop sustainability models for data, computing, and software



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

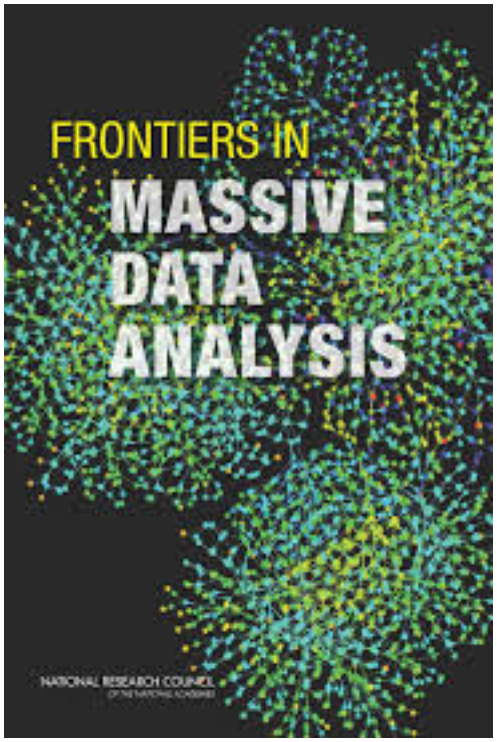
References

- 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

Jet Propulsion Laboratory
California Institute of Technology

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)
- JPL (Dan Crichton) served on the committee covering systems architecture for big data management and analysis
- Importance of systematizing the analysis of data
- 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

Jet Propulsion Laboratory
California Institute of Technology

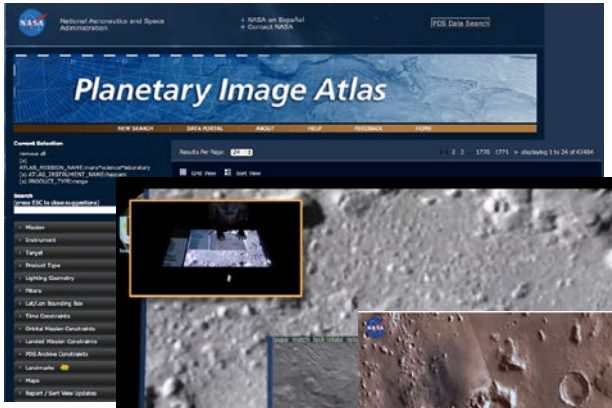
Increasing Data and Computational Science Needs

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.

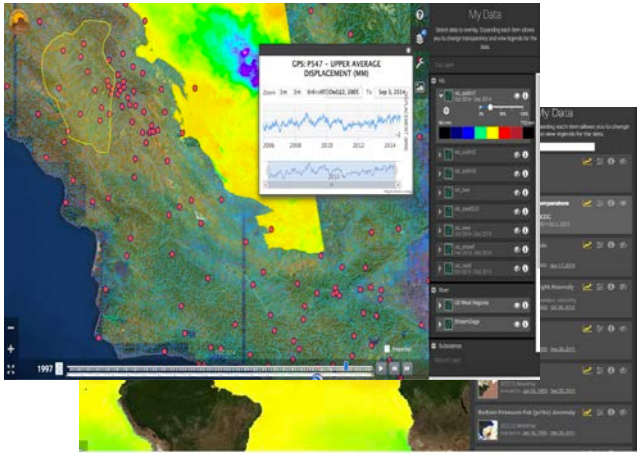
NASA AIST Big Data Study (2015/2016)

Jet Propulsion Laboratory
California Institute of Technology

Visualization, Analytics and Applications

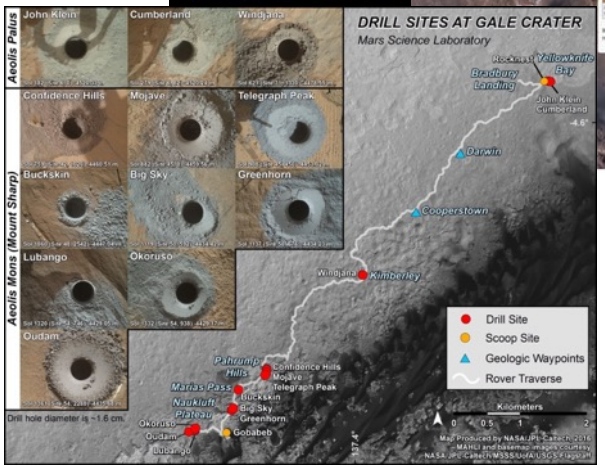


Interactive visualization of heterogeneous planetary objects

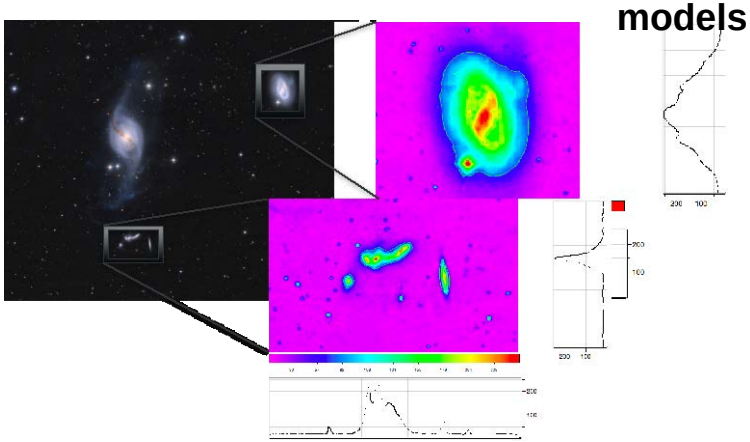


Examples: Hydrology and sea level rise

Integration of multiple earth observing remote sensing instruments; comparison against models



Examples: Planetary Image search, Mars and Moon surface navigation, feature extraction from Planetary images.



Real-time feature extraction and classification in astronomy