Jet Propulsion Laboratory
California Institute of Technology

Data Science Program Overview

Daniel Crichton
Leader, Center for Data Science and Technology
Proj. Manager, Planetary Data System Engineering
Prog. Manager, Data Science Office

Richard Doyle
Prog. Manager, Information and Data Science
Proj. Manager, High Performance Spaceflight Computing

Jet Propulsion Laboratory
California Institute of Technology

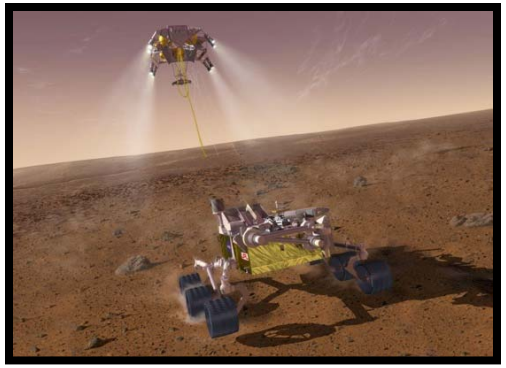
Jet Propulsion Laboratory
California Institute of Technology

Agenda

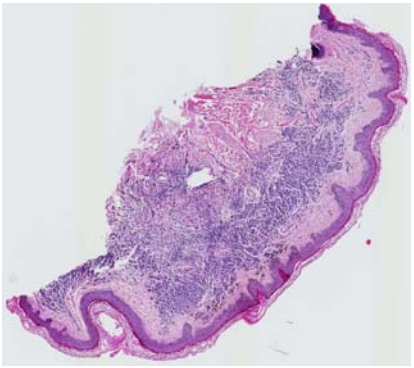
- JPL Overview
  - JPL Data Science Programs: Dan Crichton/Rich Doyle
  - Caltech Center for Data-Driven Discovery: George Djorgovski
- Machine Learning
  - Applications in Earth Science: Lukas Mandrake
  - Applications in Astronomy: Umaa Rebbapragada
  - Applications in Planetary Science: Kiri Wagstaff
- Data Analytics and Visualization
  - Sea Level Rise: Thomas Huang
  - Planetary Science: Emily Law/Shan Maholtra
  - Hydrology: Jay Famiglietti/Shan Maholtra
- Science Data Processing and Infrastructure
  - Commercial Partnering/Cloud Computing: Jim Rinaldi
  - SDS Plans for SWOT and NISAR: Hook Hua

Jet Propulsion Laboratory
California Institute of Technology

Data Science Projects at JPL



Planetary Science



Biology

Defense and Intelligence

Earth Science

Medicine

Radio Astronomy

Jet Propulsion Laboratory
California Institute of Technology

NASA Science and Big Data Today

Science Teams

How do these connect?

On Board Processing

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

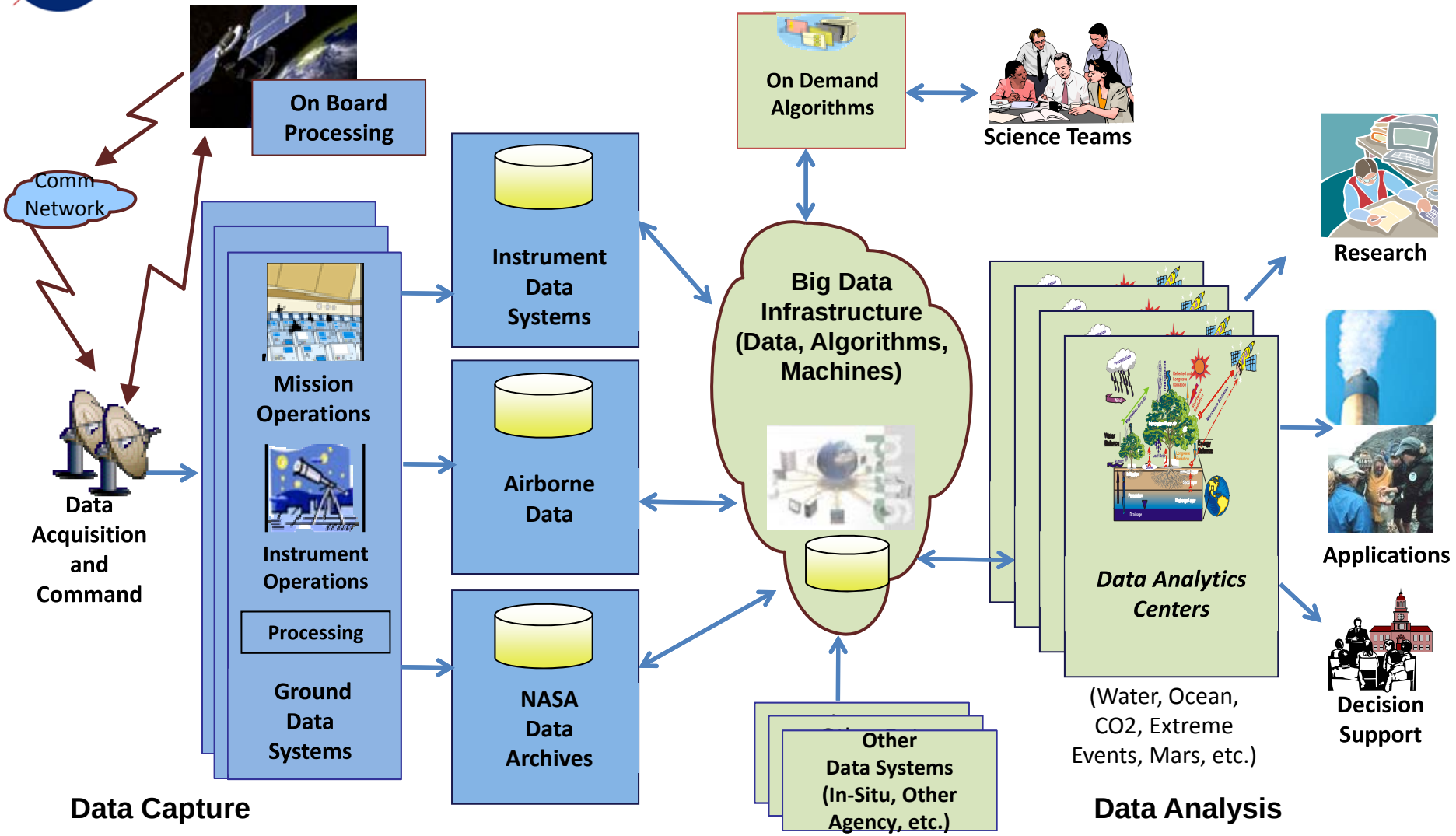
Focus on using/analyzing big data

Focus on generating, capturing, managing big data

Future of Data Science at NASA

Enabling a Big Data Research Environment

Jet Propulsion Laboratory
California Institute of Technology



Evolution towards an integrated data and computational environment

Jet Propulsion Laboratory
California Institute of Technology

JPL Data Science Strategy

Guiding Principles

Agile Science – Onboard Analysis

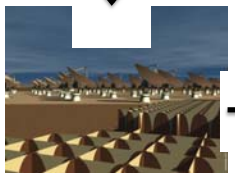
Challenge:
Too much data, too fast; cannot transport data efficiently enough

Future Solutions:
Onboard computation and data science

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



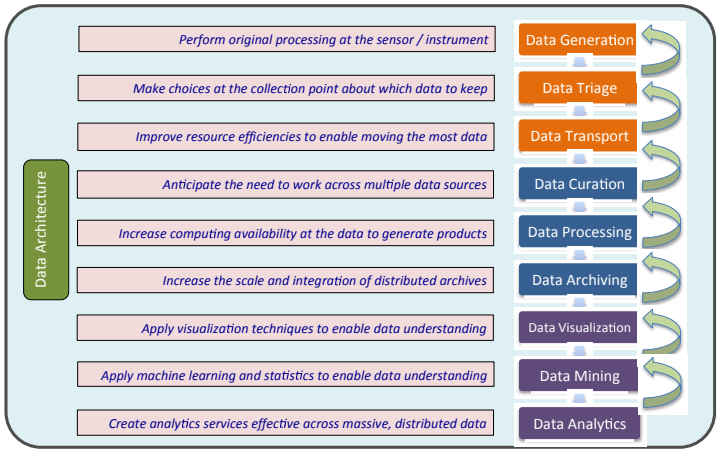
Distributed Data Analytics

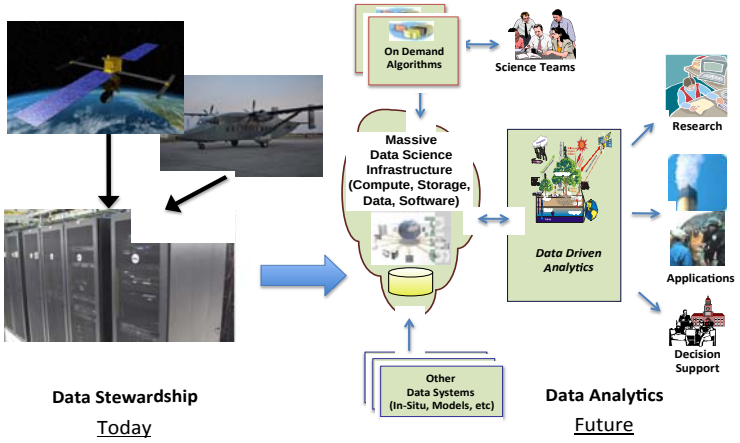
Challenge: Data distributed in massive archives; many different types of measurements

Future Solutions: Distributed data analytics; uncertainty quantification

Data Lifecycle

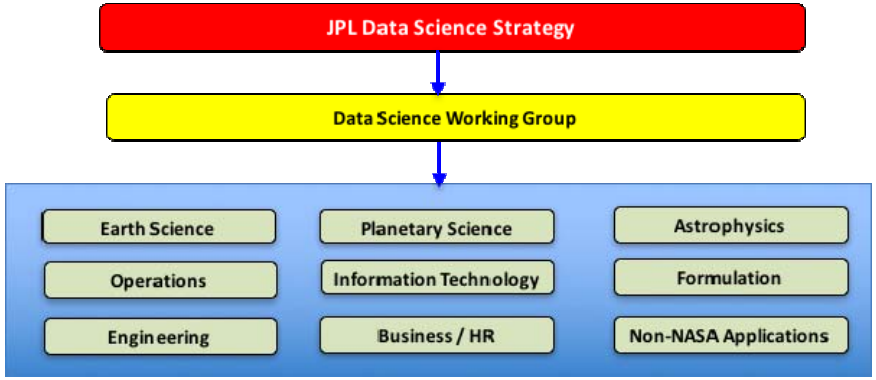
Organize data and computing end-to-end





Data Ecosystem

Integrated computing capabilities from HPC to data repositories to on-demand analytics.

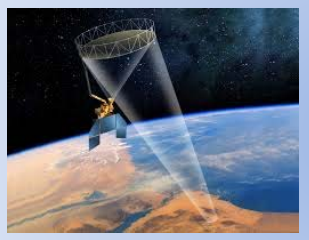


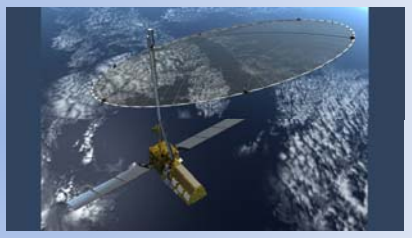
Cross-Cutting

Cut across disciplines to collaborate and share methodologies

Jet Propulsion Laboratory
California Institute of Technology

NASA Big Data Landscape



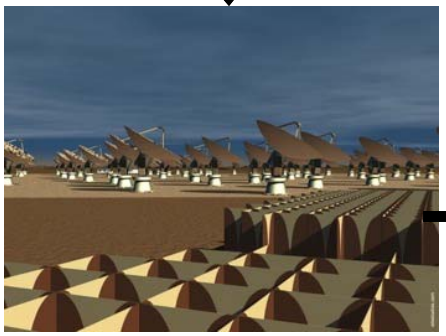


SMAP (Today): 485 GB/day NI-SAR (2020): 86 TB/day

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

Observational Platforms and Flight Computing

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



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

Ground-based Mission Systems

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

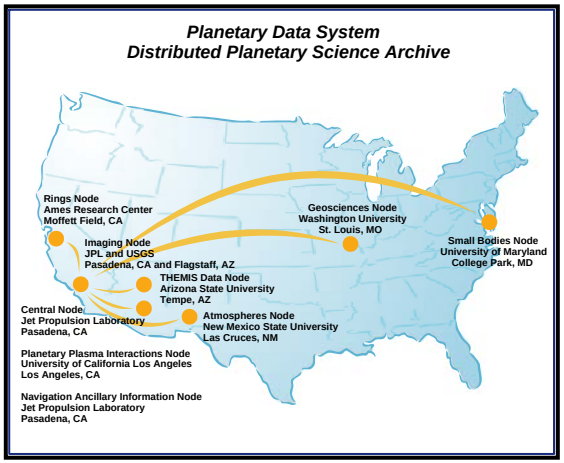
Massive Data Archives and Big Data Analytics

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

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

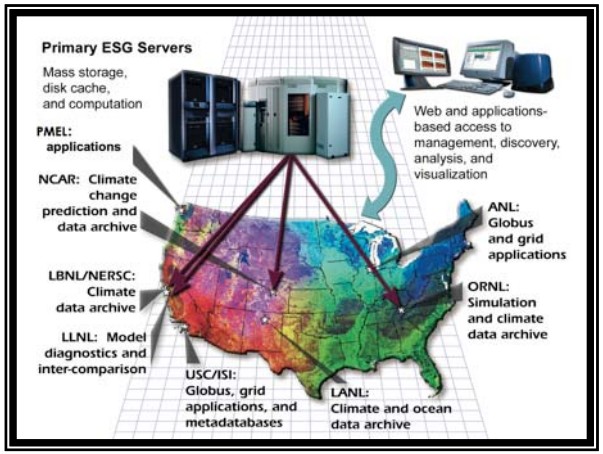
Jet Propulsion Laboratory
California Institute of Technology

Data System Scientific Research Networks: Access to Observations and Models



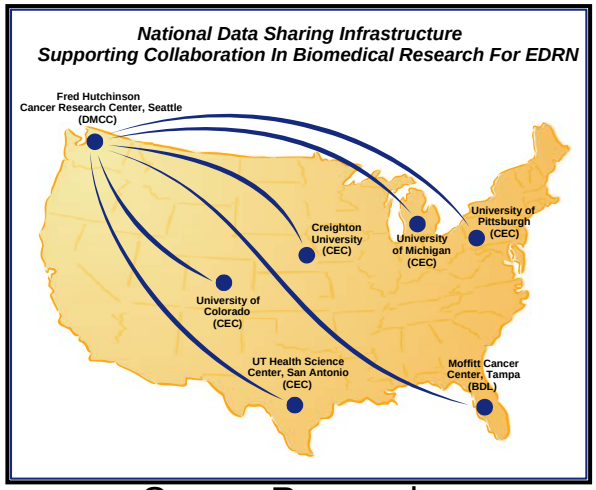
Solar System Exploration

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



Climate Research

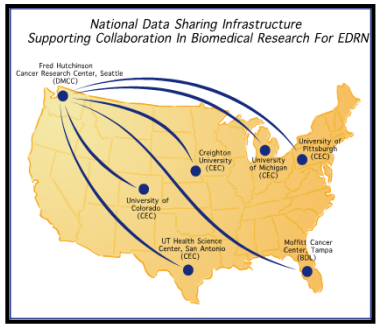
Earth Observation



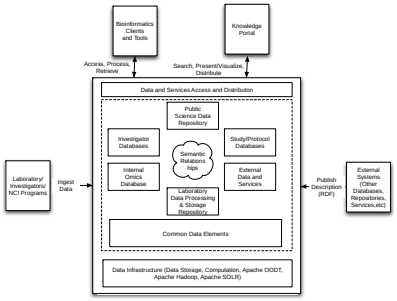
Cancer Research

Jet Propulsion Laboratory
California Institute of Technology

Cross-Cutting Capabilities

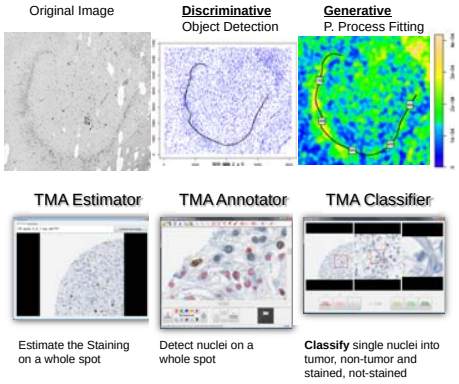


International Data Archive and Sharing Architectures



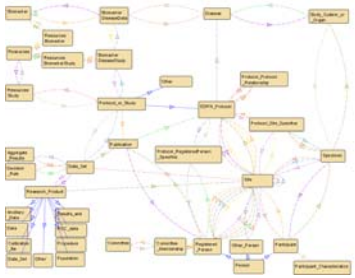
Big Data Infrastructures

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



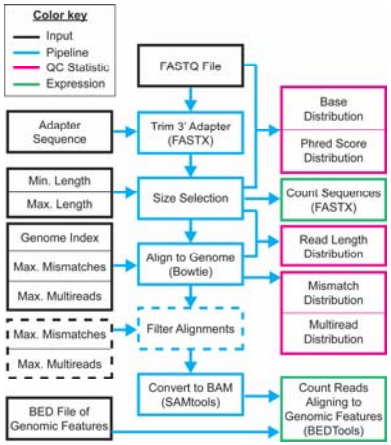
Intelligent Data Algorithms

(Machine Learning, Deep Learning)



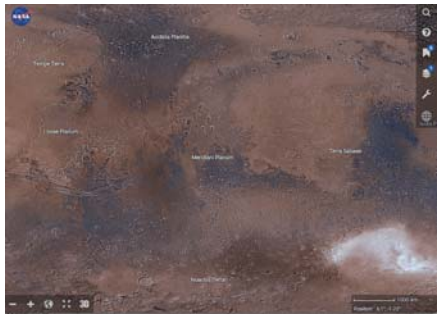
Common Data Elements & Information Models

(discipline and common)



Analytical Data Pipelines

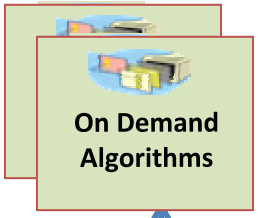
Great Opportunities for Methodology Transfer and Collaboration

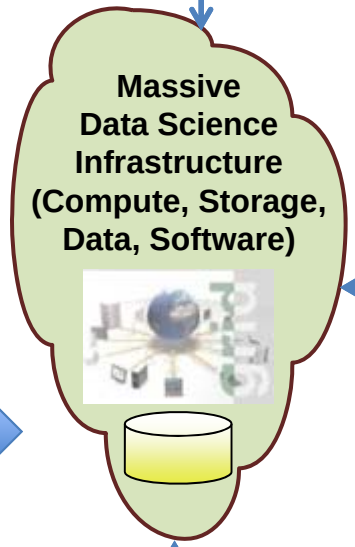


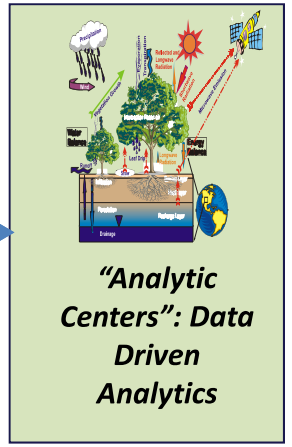
Visualization Techniques

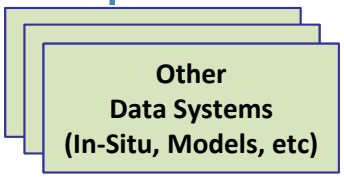
Jet Propulsion Laboratory
California Institute of Technology

Evolving Towards Data Analytics









Data Stewardship

Today

Data Analytics

Future

Jet Propulsion Laboratory
California Institute of Technology

The Growing Need for Data Science


SPACENEWS magazine

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

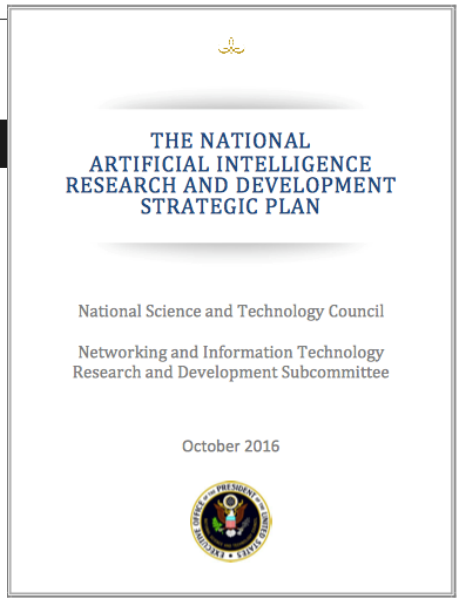
Smallsats and the multi-trillion-dollar data set

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

Jet Propulsion Laboratory
California Institute of Technology

Data Science Growth Strategy

- In November 2016 a chartered Data Science WG reporting to JPL's Leadership Management Council (LMC), chaired by Deputy Director Larry James, was established. We have and are launching:

- 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.

- Services – mature capabilities to grow data science

- Projects – drive maturing services and pilots to address specific use cases

Jet Propulsion Laboratory
California Institute of Technology

Partnering Strategy

- Universities
  - Early collaborations with UC, CMU, MIT
  - Increasing curriculum in data science
  - Opportunities for NASA and JPL investment in internships and research
- 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

bigdata@CSAIL
MIT BIG DATA INITIATIVE

MIT Lincoln Laboratory

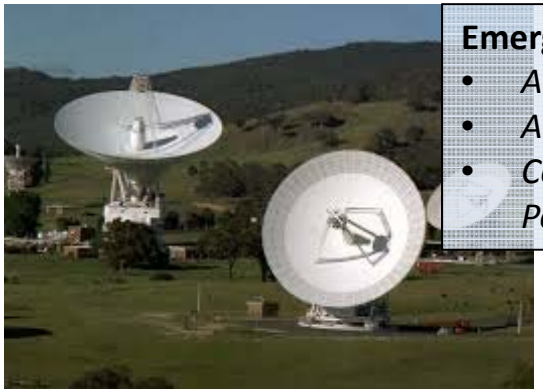
amazon web services

THE APACHE SOFTWARE FOUNDATION

Jet Propulsion Laboratory
California Institute of Technology

Applying Data Technologies Across Ground Environment

Intelligent Ground Stations



Emerging Solutions
- *Anomaly Detection*
- *Attention Focusing*
- *Controlling False Positives*

Data-Driven Discovery from peta-scale Archives

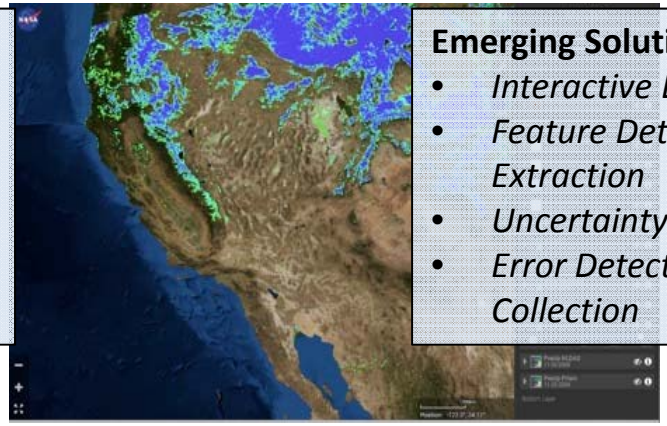
Emerging Solutions
- *Machine Learning - Feature Extraction and Classification*
- *Intelligent Search*
- *Data Fusion*

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

Mission Operations

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

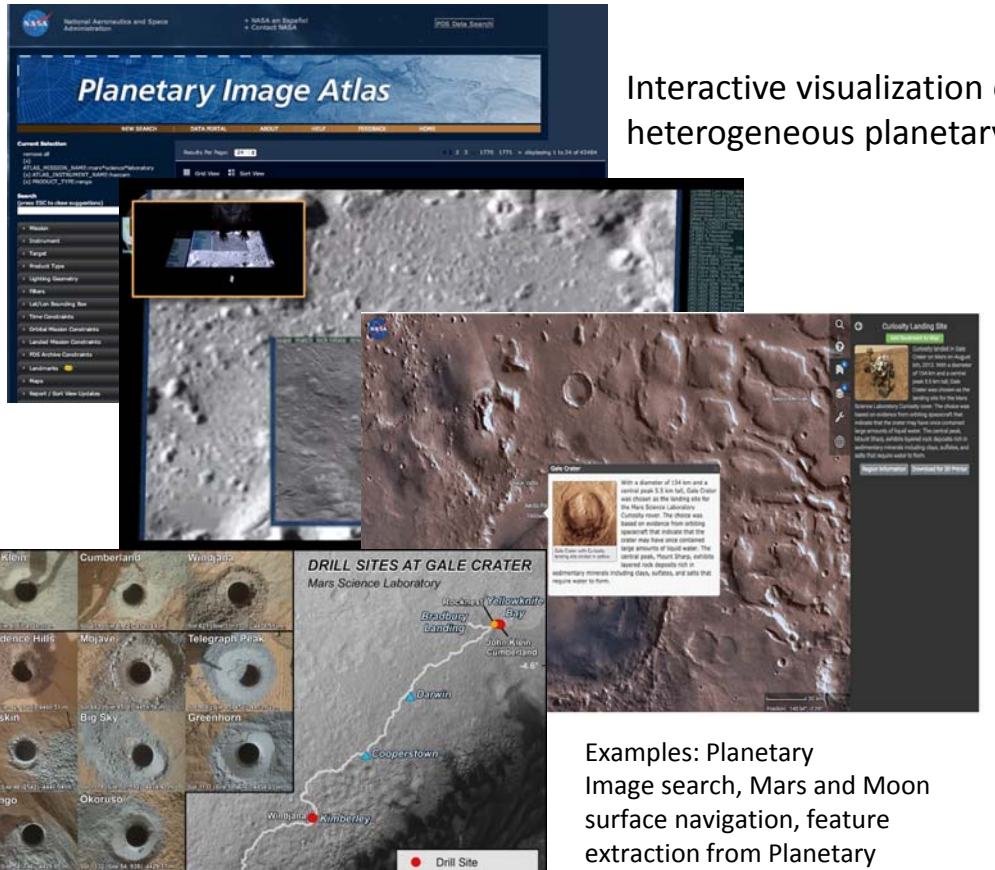
Data Analytics and Decision Support



Emerging Solutions
- *Interactive Data Analytics*
- *Feature Detection and Extraction*
- *Uncertainty Quantification*
- *Error Detection in Data Collection*

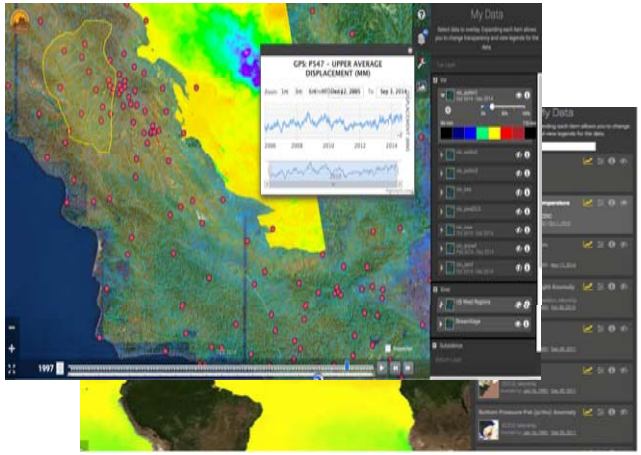
Jet Propulsion Laboratory
California Institute of Technology

Visualization, Analytics and Applications



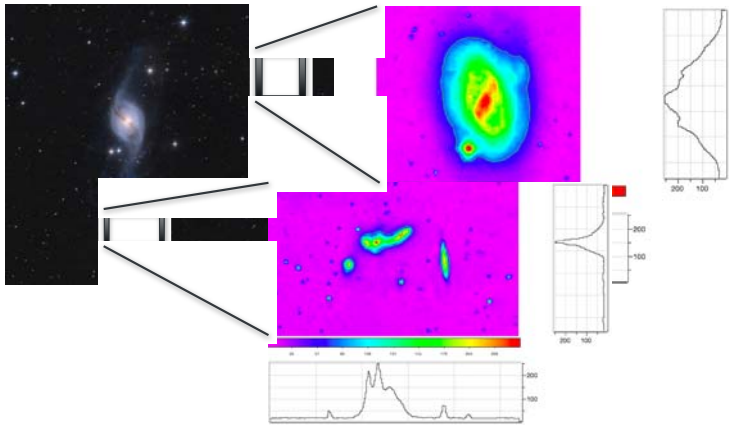
Interactive visualization of heterogeneous planetary objects

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



Examples: Hydrology and sea level rise

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



Real-time feature extraction and classification in astronomy

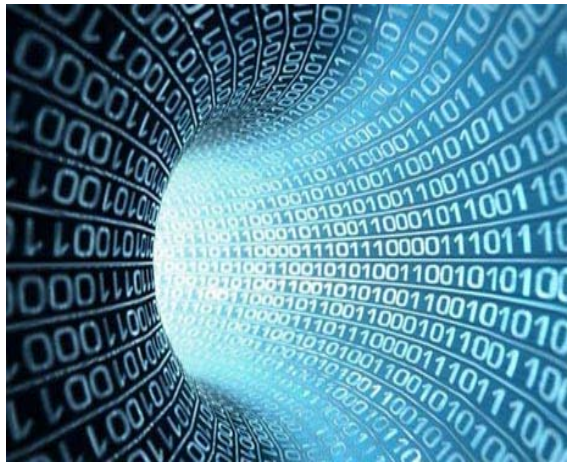
Jet Propulsion Laboratory
California Institute of Technology

Recommendations

- Great opportunities to leverage data science
  - Use the Mission-Science Data Lifecycle to organize a vision for data and computing
  - Partner across SMD and with other agencies; explore opportunities for methodology transfer
  - Launch investments (technology to operational capabilities)
    - Expand use of ROSES to support data science technologies across all science disciplines in SMD
  - Support open source and industry collaborations

- Evolve to support use and data analytics for the community.
  - Drive broad, international data ecosystems
  - Increase use of data-driven approaches to gain insight and understanding
  - Develop sustainability models for data, computing, and software

What do we do with all this data?

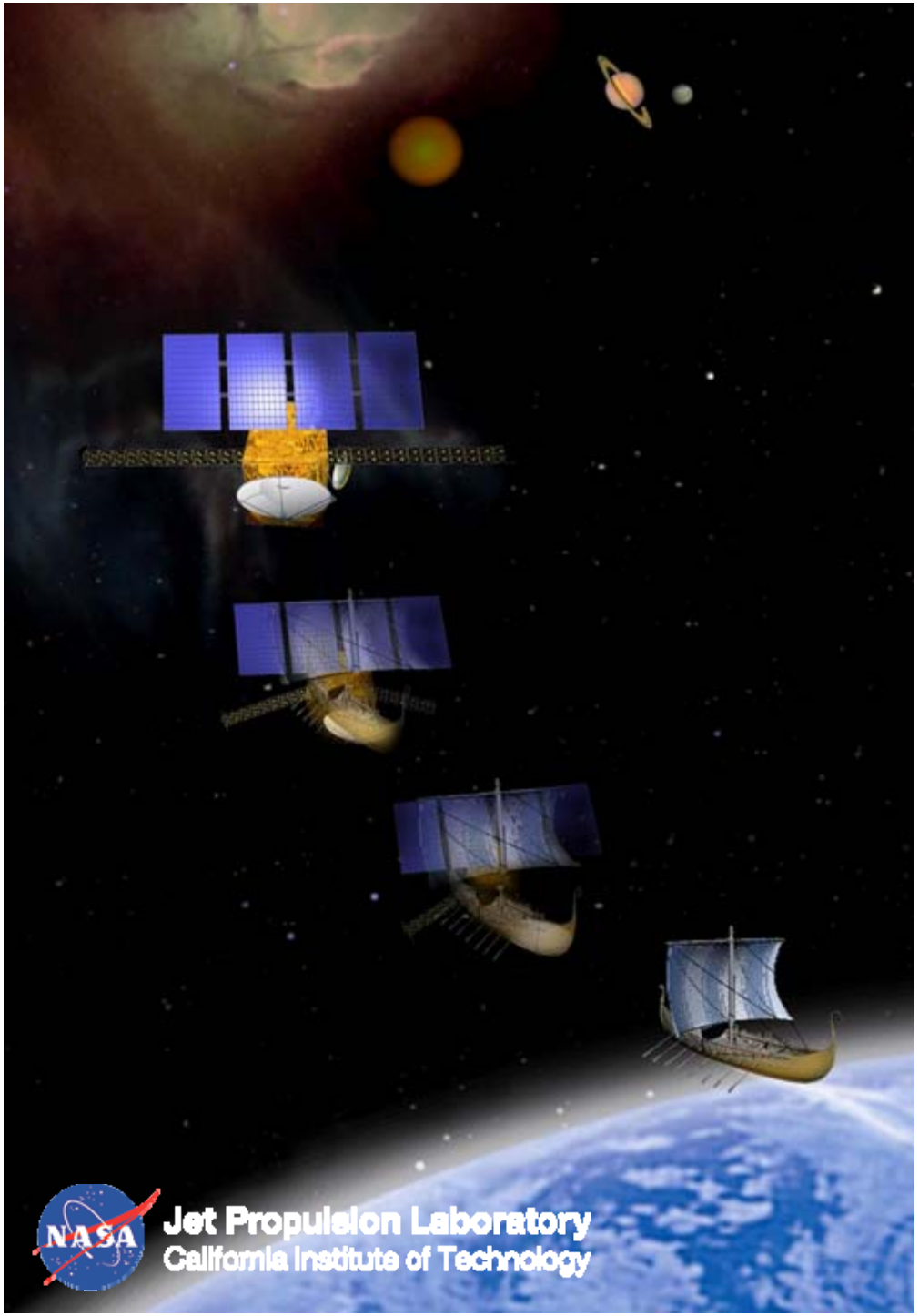


This is looking like a black hole –
but wait, there's light at the end of the tunnel!

Jet Propulsion Laboratory
California Institute of Technology

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


NASA
Jet Propulsion Laboratory
California Institute of Technology

Questions?