Big Data and Data Science at NASA/JPL: Methodology Transfer From Space Science to Biomedicine

Daniel Crichton

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

Dr. Richard Doyle

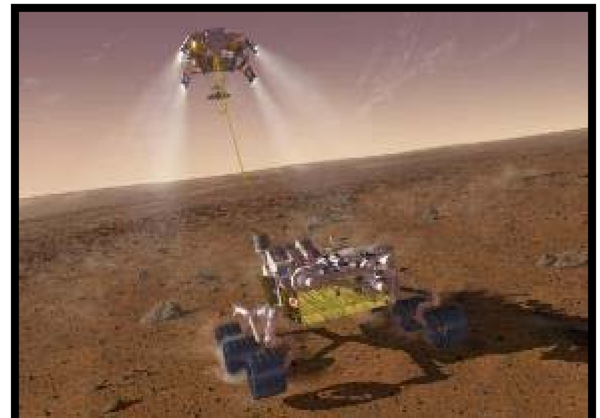
Prog. Manager, Information and Data Science
Proj. Manager, High Performance Spaceflight Computing

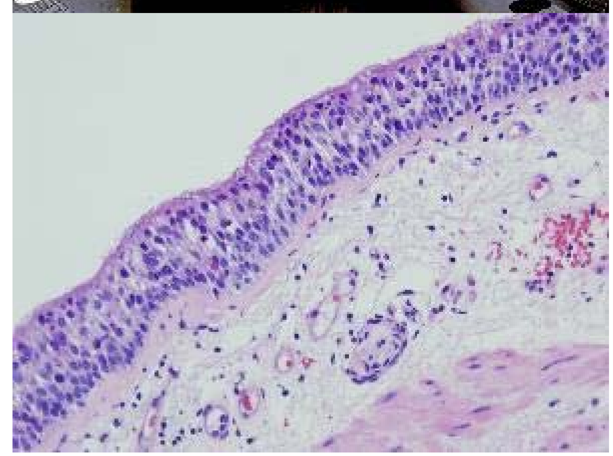
Jet Propulsion Laboratory
California Institute of Technology

March 6, 2017

Context

- 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.



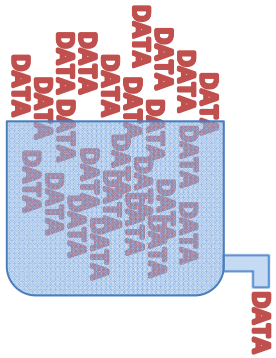


Jet Propulsion Laboratory
California Institute of Technology

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



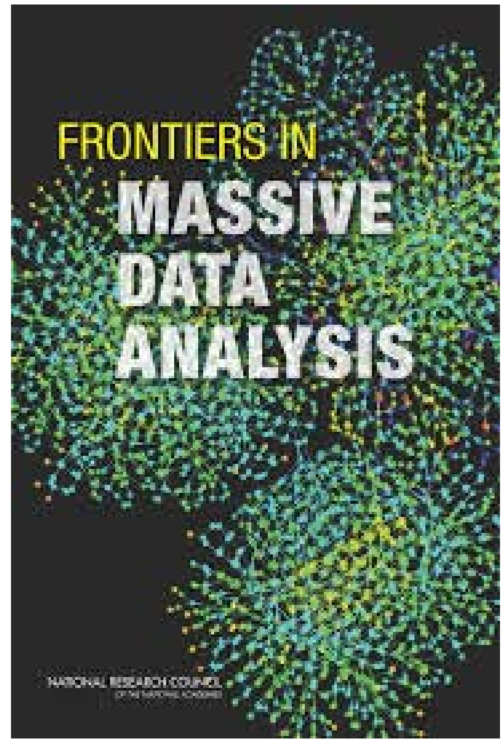
Data Science

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

Jet Propulsion Laboratory
California Institute of Technology

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

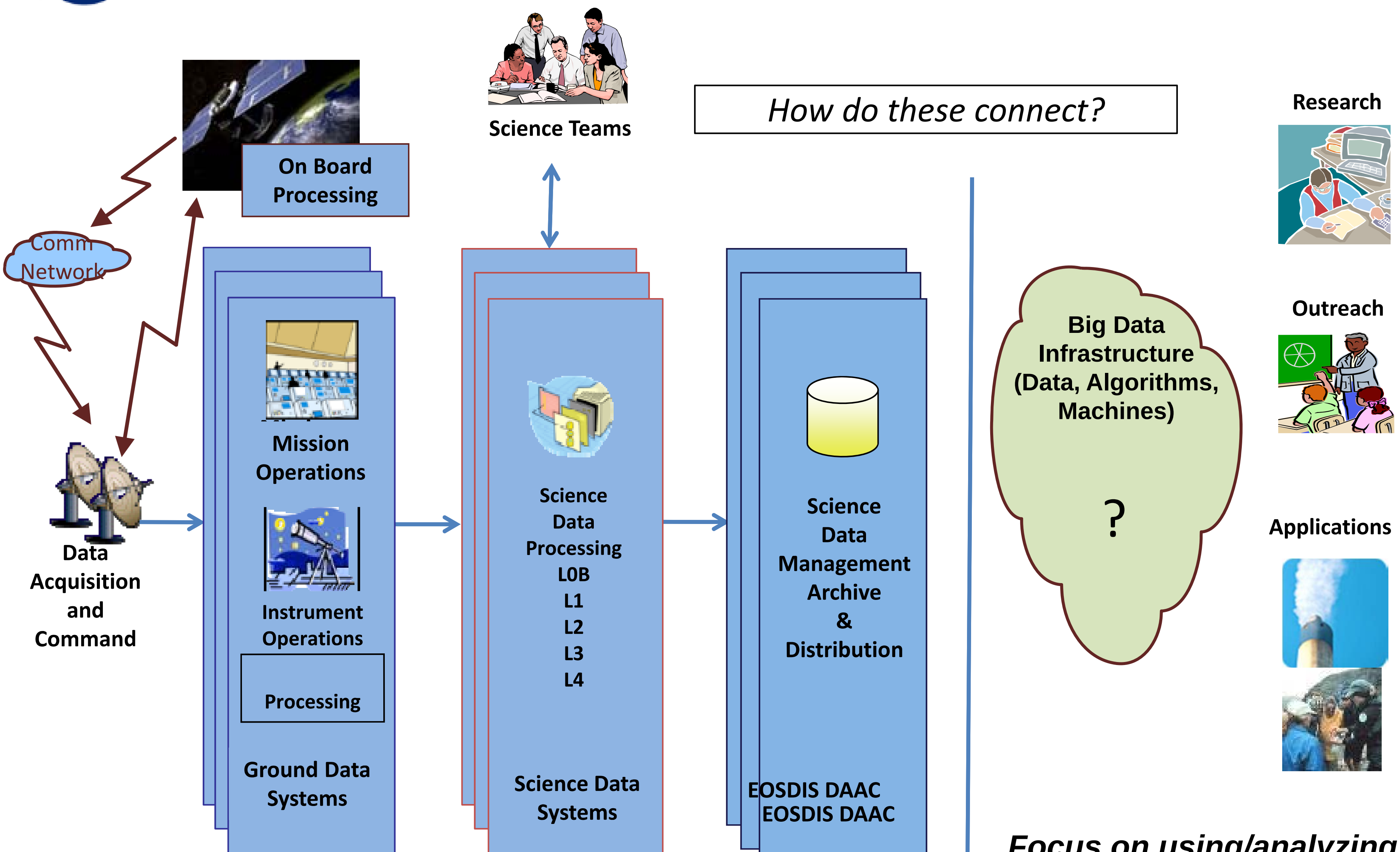
- Chartered in 2010 by the U.S. National Research Council, National Academies
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- NASA/JPL served on the committee covering systems architecture for big data management and analysis
- **Importance of more systematic approaches 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

NASA Science and Big Data Today

Jet Propulsion Laboratory
California Institute of Technology



Focus on using/analyzing big data

Focus on generating, capturing, managing big data

JPL Data Science Working Group

- Established in 2014 to explore big data use cases and challenges in science and to make a recommendation to JPL senior management.
  - Launched internal investments: planetary science (onboard agile science), earth science (distributed data analytics), and astronomy (machine learning and data collection methods).
  - Engaged cross disciplinary expertise (science, computer science – systems and machine learning, statistics, program management)
  - Partnered with Caltech to bring in research perspectives.

- In November 2016 a chartered Data Science WG reporting to JPL's Leadership Management Council (LMC), chaired by Deputy Director Larry James, was established in data science covering all aspects of the Lab operations.

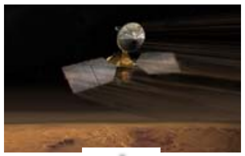
Jet Propulsion Laboratory
California Institute of Technology

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



Data Lifecycle

Extreme Data Volumes – Data Triage

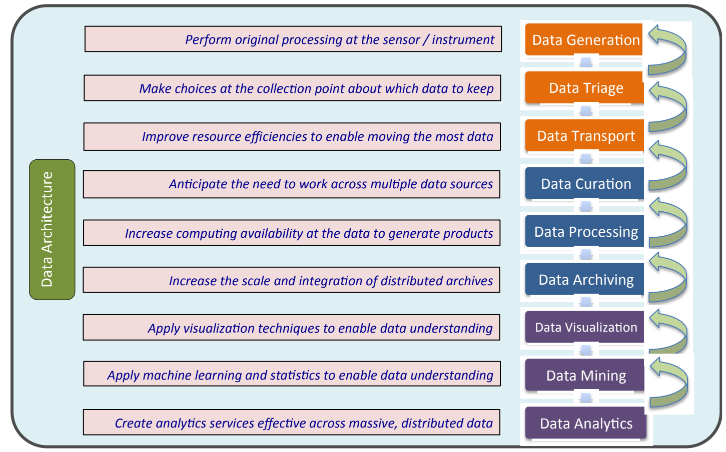
Challenge: Data collection capacity at the instrument outstrips data transport and data storage capacity

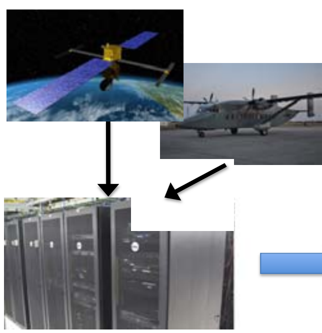
Future Solutions: Dynamic architectures to scale data processing and triage exascale data streams

Distributed Data Analytics

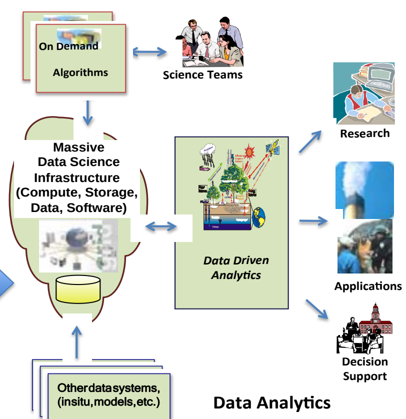
Challenge: Data distributed in massive archives; many different types of measurements

Future Solutions: Distributed data analytics; uncertainty quantification



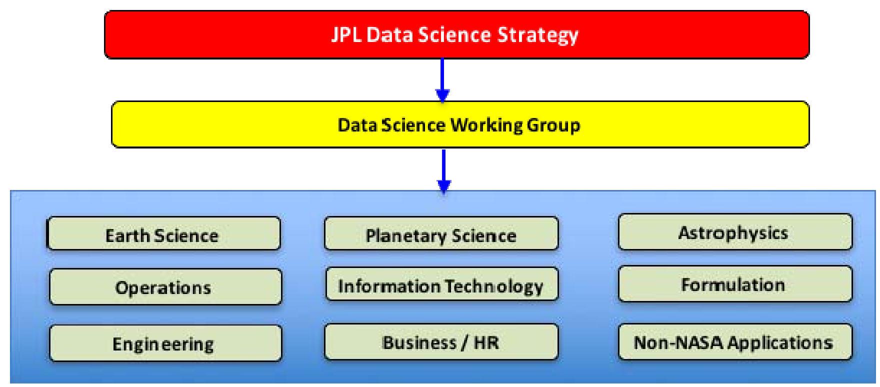


Data Stewardship
Today



Data Analytics
Future

Data Ecosystem



Cross-Cutting

Jet Propulsion Laboratory
California Institute of Technology

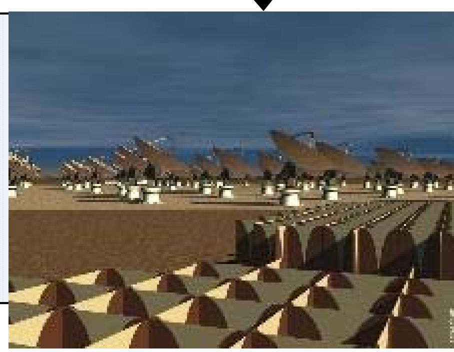
Data Lifecycle Model
for NASA Space Missions

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

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

Observational Platforms /Flight Computing

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



Massive Data Archives and Big Data Analytics

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

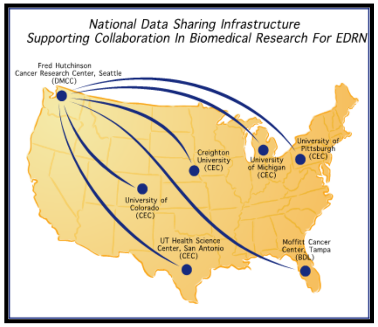
(2) Data collection capacity at the instrument continually outstrips data transport (downlink) capacity

Ground-based Mission Systems

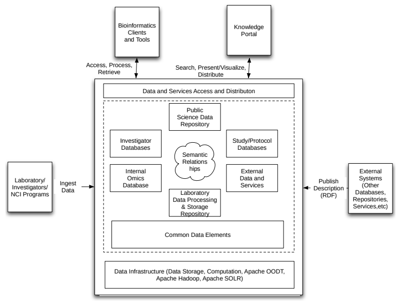
(3) Data distributed in massive archives; many different types of measurements and observations

Cross-Cutting Capabilities

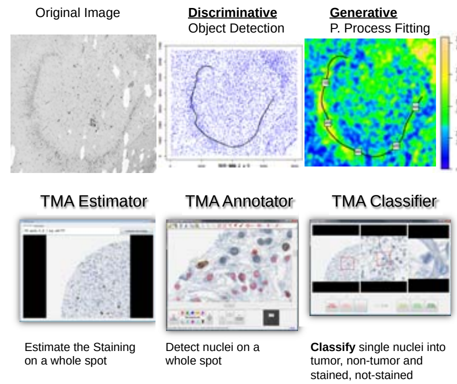
Jet Propulsion Laboratory
California Institute of Technology



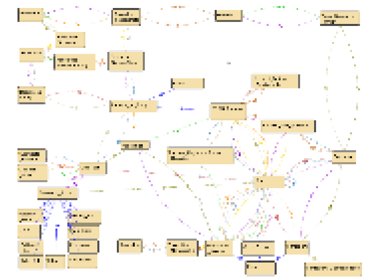
International Data Archive and Sharing Architectures



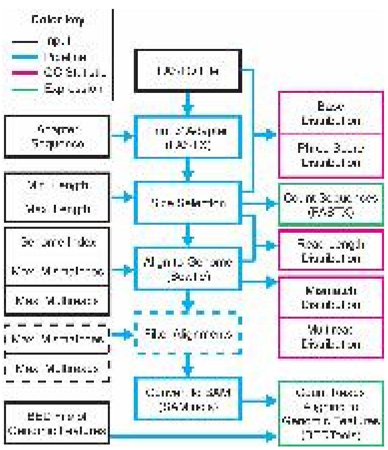
Big Data Infrastructures
(from open source to cloud computing and scalable compute infrastructures)



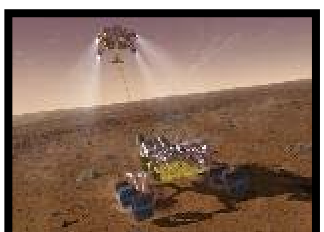
Intelligent Data Algorithms
(Machine Learning, Deep Learning)

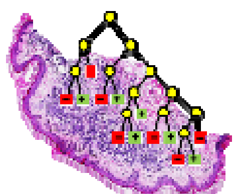


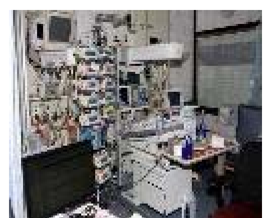
Common Data Elements & Information Models
(discipline and common)



Analytical Data Pipelines


Planetary


Biology


Healthcare

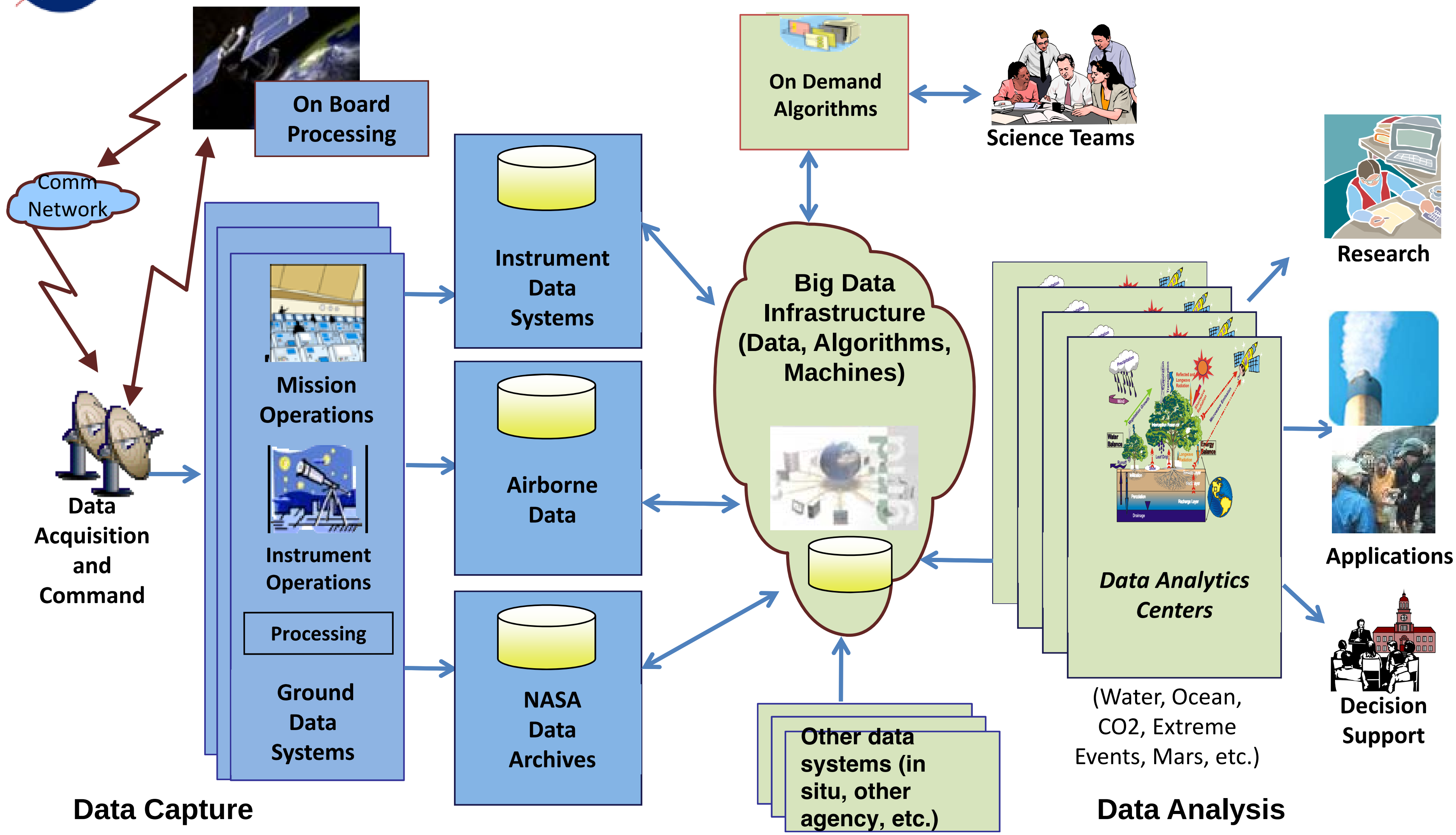
Visualization Techniques

Great Opportunities for Methodology Transfer and Collaboration

Future of Data Science at NASA

Enabling a Big Data Research Environment

Jet Propulsion Laboratory
California Institute of Technology



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.

Jet Propulsion Laboratory
California Institute of Technology

Opportunities and Use Case Across the Ground Environment

Intelligent Ground Stations

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

Intelligent Archives and Knowledge-bases

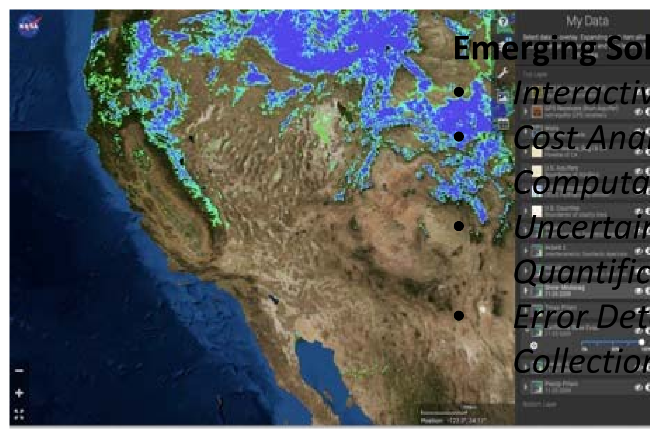
Emerging Solutions
- *Automated Machine Learning - Feature Extraction*
- *Intelligent Search*
- *Learning over time*
- *Integration of disparate data*

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

2015-2016 AIST Big Data Study

- Study led by JPL for the NASA Advanced Information Systems Technology Program (under Mike Little)
- Mapped technology and data needs against the mission-science data lifecycle
- Focuses on expansion from data stewardship to data use across the vast data ecosystem (satellite, airborne, in situ)
- Basis for 2015 IEEE Big Data workshop on Data and Computational Challenges in Earth Science Research
- Key input for 2016 ROSES AIST call (per Mike Little, NASA PM for AIST)

ESTO
Earth Science Technology Office

Data and Computational Science
Technologies for Earth Science Research

Jet Propulsion Laboratory
California Institute of Technology

AIST Big Data Study: 10 Year Capability Needs in Big Data

System	2015	2025	NASA Applications
Observational Platforms	Limited onboard computation including data triage and data reduction. Investments in new flight computing technologies for extreme environments.	Increase onboard autonomy and enable data triage (machine learning techniques) to support more capable instruments. Support reliable onboard processing in extreme environments to enable new exploration missions.	Onboard computation across all types of platforms; flight computing capabilities deployed for extreme environments; data triage for satellites and spacecraft.
Ground-based Mission 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/ML techniques, event and feature detection. Support 100 PB scale missions.	Future mission computational challenges; high bandwidth data volumes; more agile airborne, cube sat, multi-sensor campaigns; increase automated event detection across mission lifecycle.
Massive Data Archives	Support for 10 PB of archival data; limited automated event and feature detection.	Support exascale archives; automated event and feature detection/ML techniques; virtually integrated, distributed archives.	Turn archives into knowledge-bases to improve data discovery. Leverage massively scalable virtual data storage.
Distributed Data Analytics	Limited analytics services; generally tightly coupled to specific data centers; limited cross-archive/data center, cross-agency integration; limited capabilities in data fusion; statistical uncertainty; provenance of the results.	Computational techniques (ML, statistical methods) integrated into mission-science lifecycle; Integrated data, HPC, algorithms across archives; Support for cross product data fusion; capture of statistical uncertainty; virtual missions; specialized Analytics Centers.	Automated data analysis methods; integration of data across spacecraft, remote sensors, satellite, airborne, and ground-based sensors; systematic approaches to addressing uncertainty; complex scientific questions.

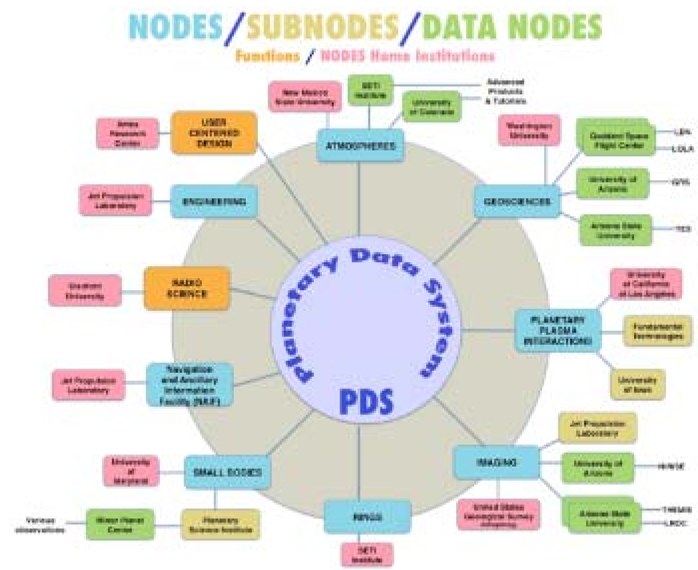
Derived from AIST Big Data Study & NASA Office of the Chief Technologist TA-11 Roadmap (2015)

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

Jet Propulsion Laboratory
California Institute of Technology

(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!

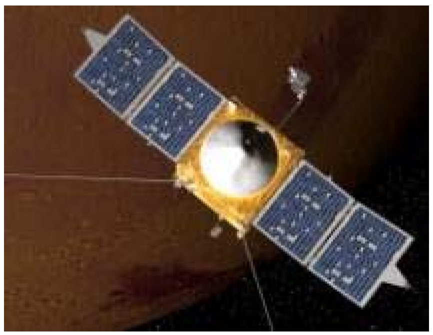
Jet Propulsion Laboratory
California Institute of Technology

PDS4: International Adoption of an Open Planetary Approach


LADEE (NASA)


InSight (NASA)


BepiColombo (ESA/JAXA)


MAVEN (NASA)


Osiris-REx (NASA)


ExoMars (ESA/Russia)

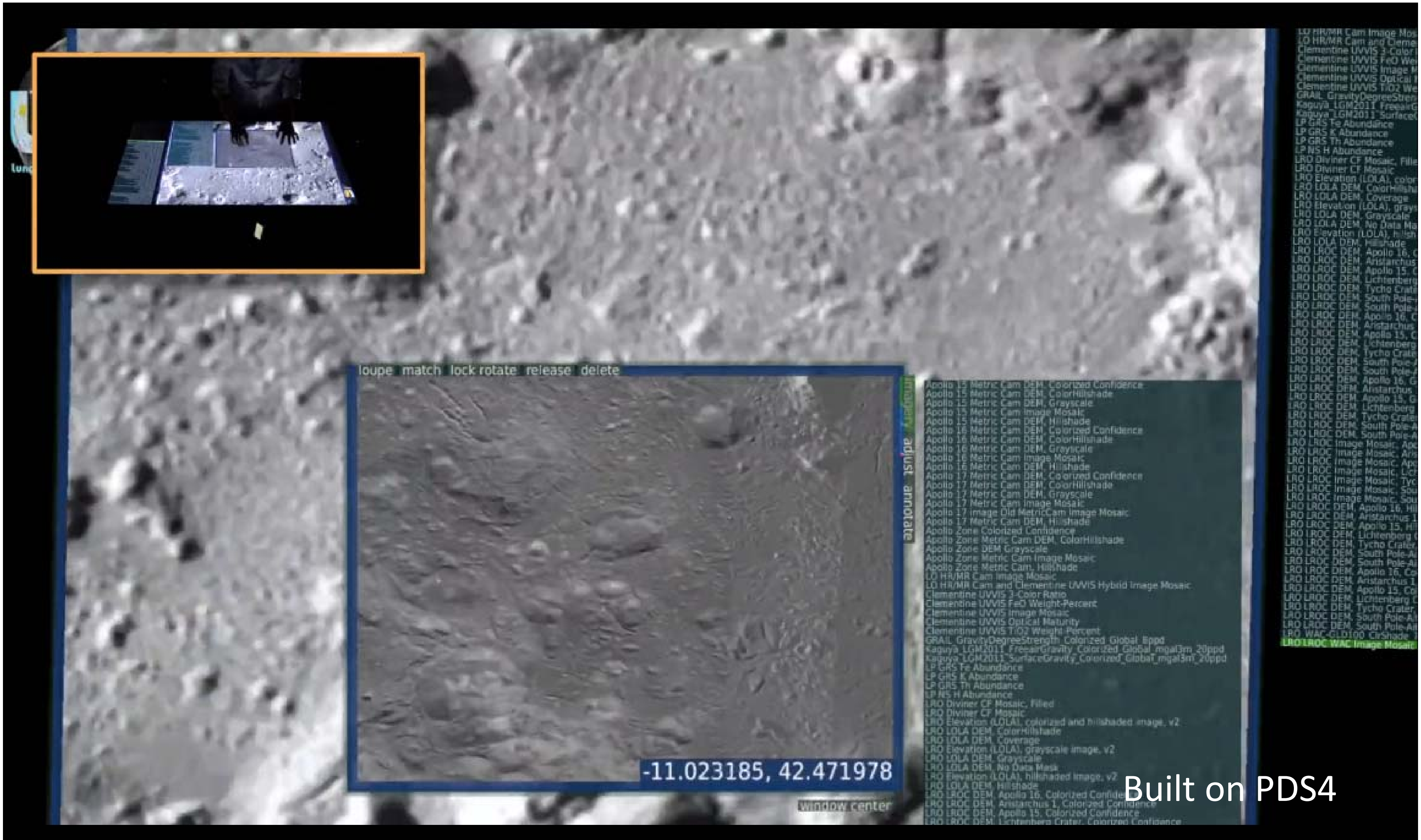

JUICE (ESA)


Mars 2020 (NASA)

…also Hayabussa-2, Chandrayaan-2

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

Lunar Mapping and Modeling Portal as Data Analytics and Visualization



E. Law, S. Malhotra, G. Chang

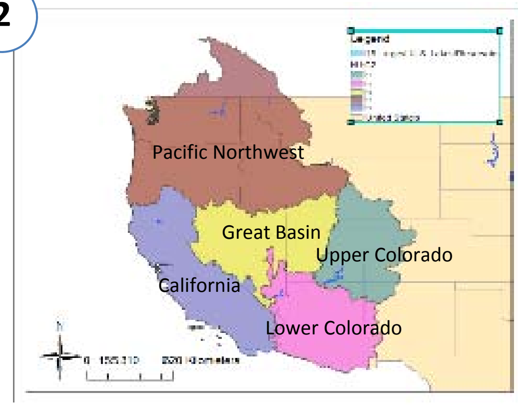
Jet Propulsion Laboratory
California Institute of Technology

Western States Water Mission – Understanding Water Availability

1

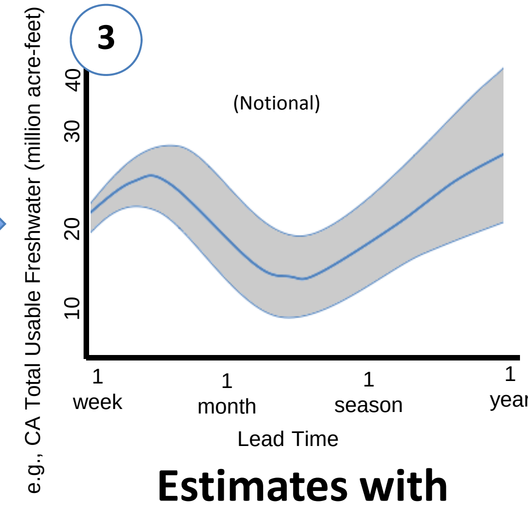
Observations

2

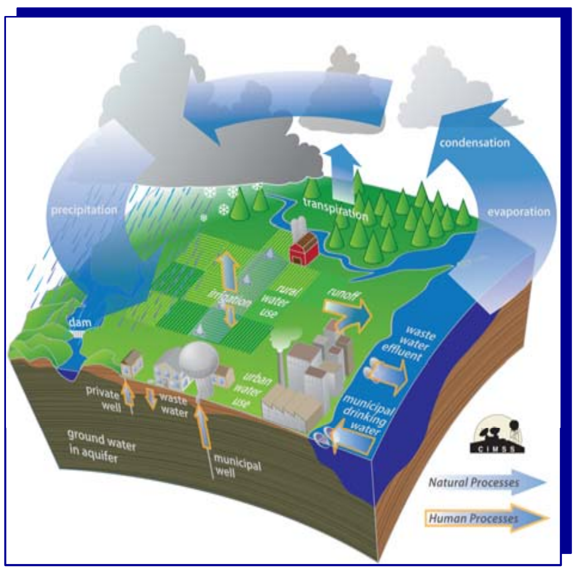


Coupled and Validated Computer Models

3



Estimates with Uncertainties



4

DEPARTMENT OF WATER RESOURCES

western states water council

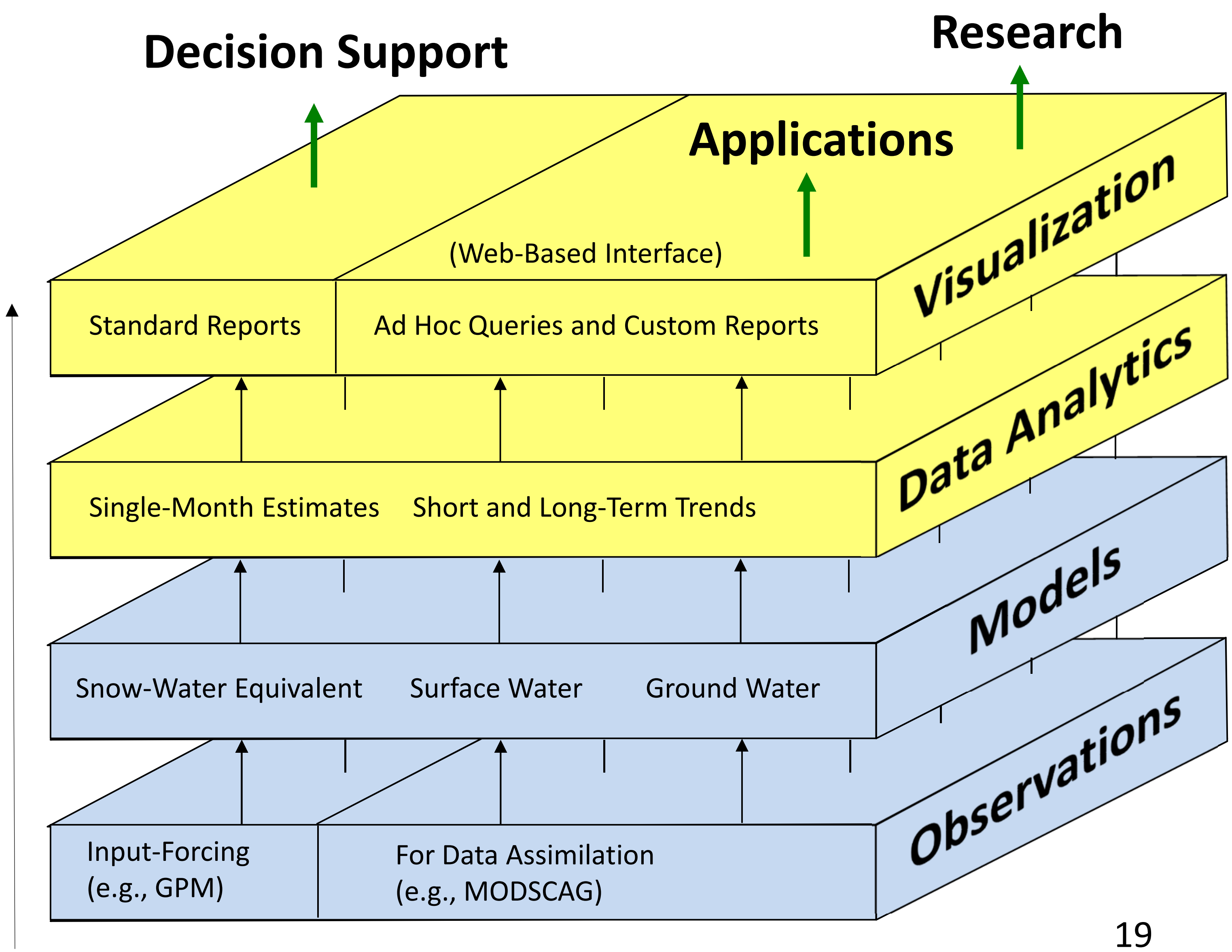
(Prospective customers)

U.S. DEPARTMENT OF THE INTERIOR BUREAU OF RECLAMATION

Colorado River Basin

Stakeholders and Customers

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



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

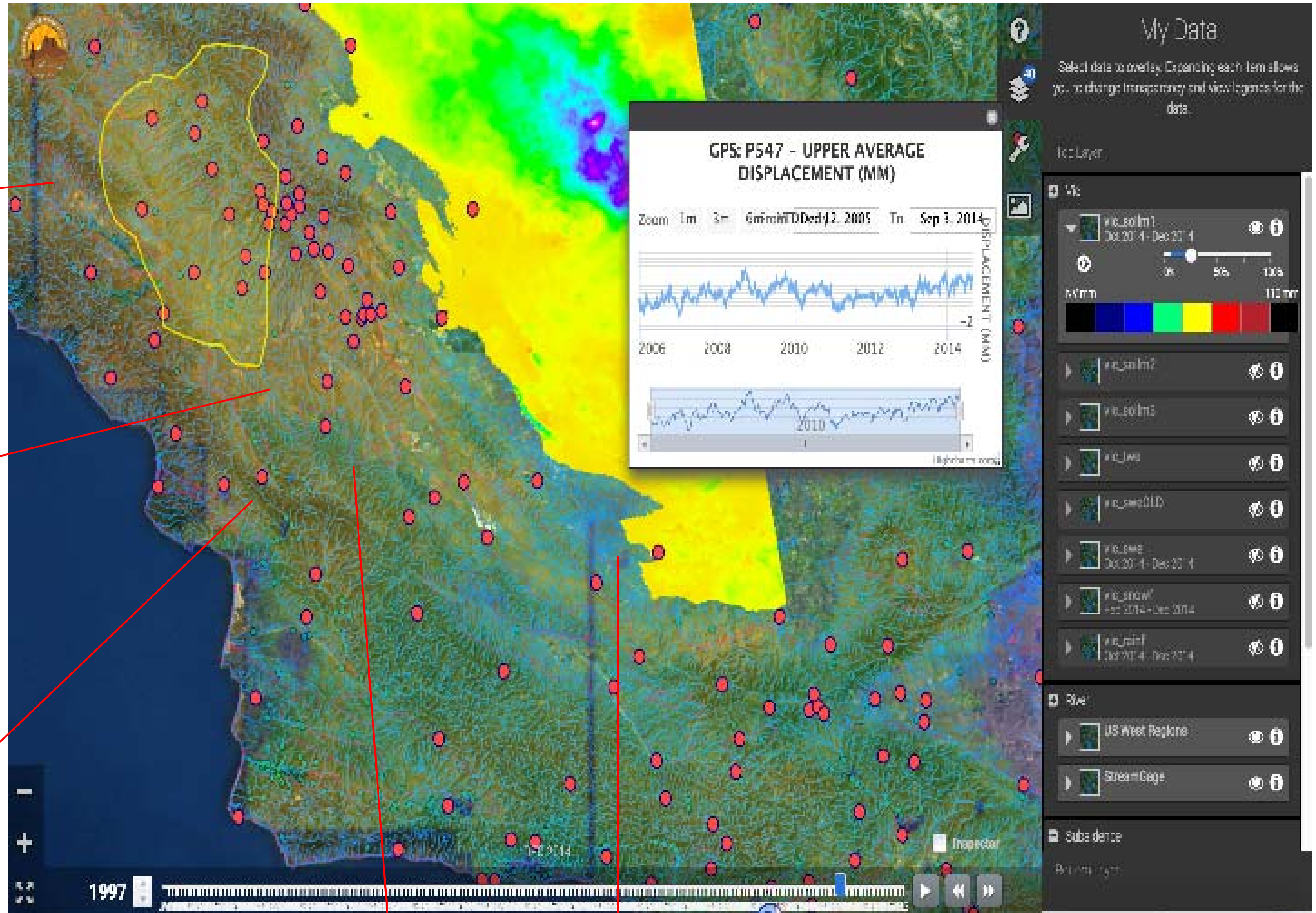
A Scalable Data Processing System for Hydrological Science

Jet Propulsion Laboratory
California Institute of Technology

WaterTrek

User Defined Polygon

GPS



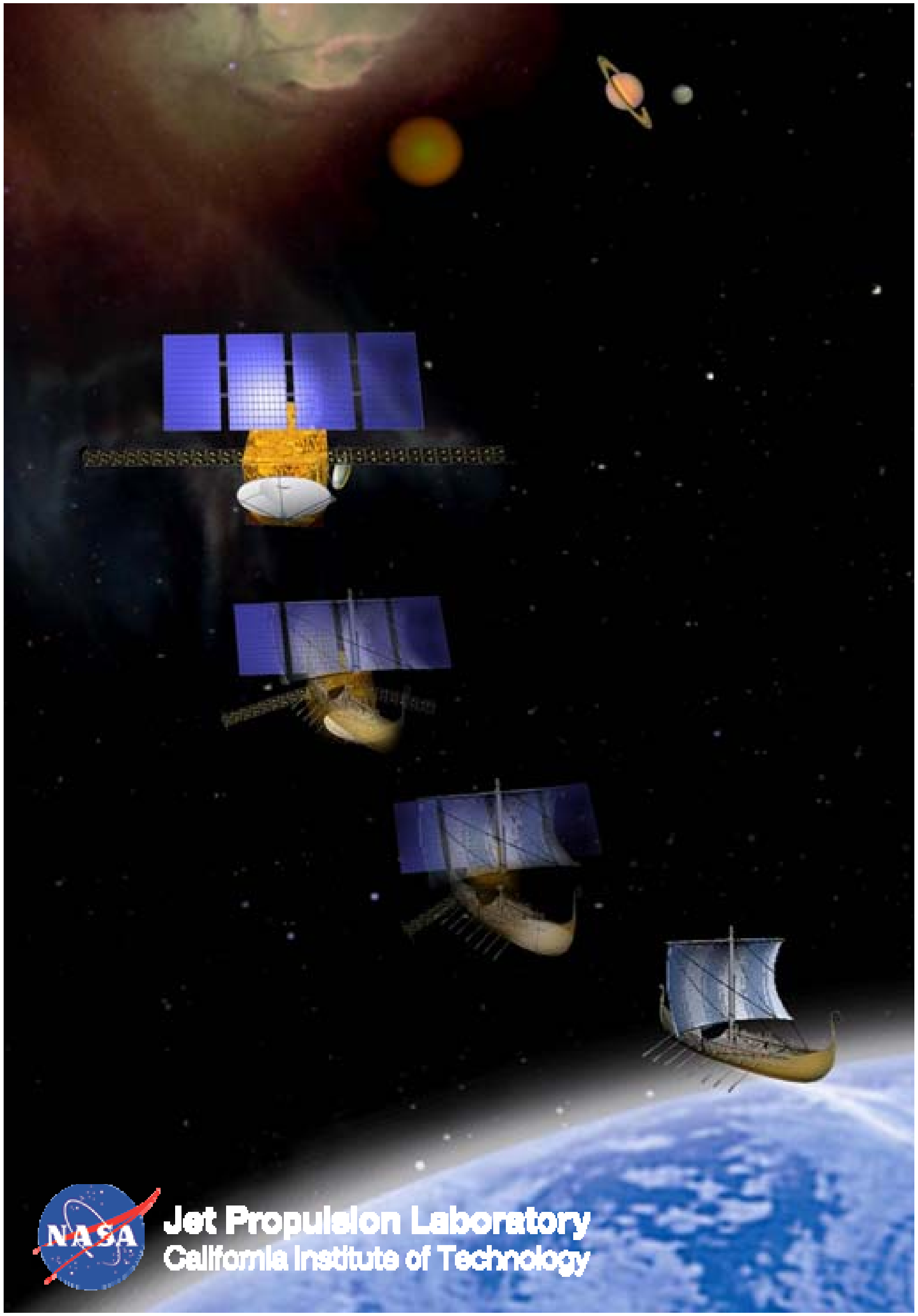
In-Situ: Stream Gage Sensors

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

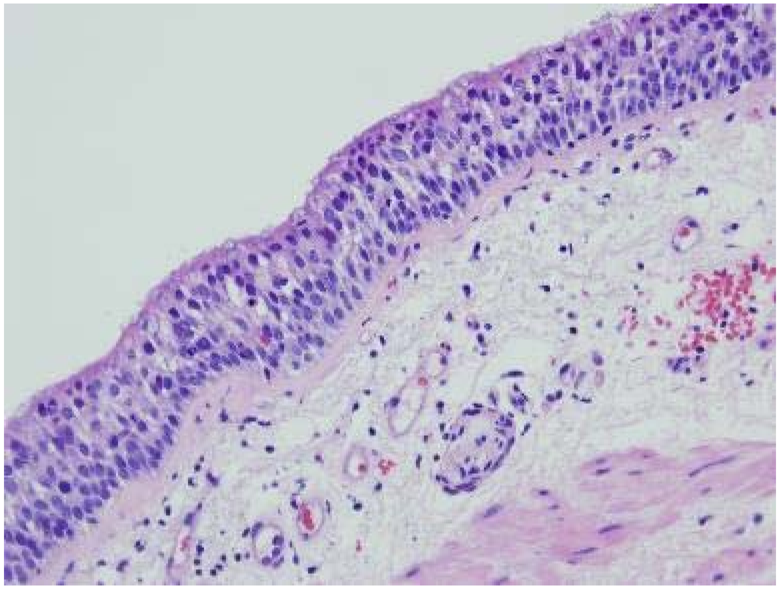


Methodology Transfer in Data Science from Planetary & Earth to Biomedicine

NASA/JPL Informatics Center: Crossing Disciplines to Support Scientific Research

- Development of an advanced Knowledge System to *capture*, *shar* and support *reproducible analysis* for biomarker research
  - Genomics, Proteomics, Imaging, etc data types of data

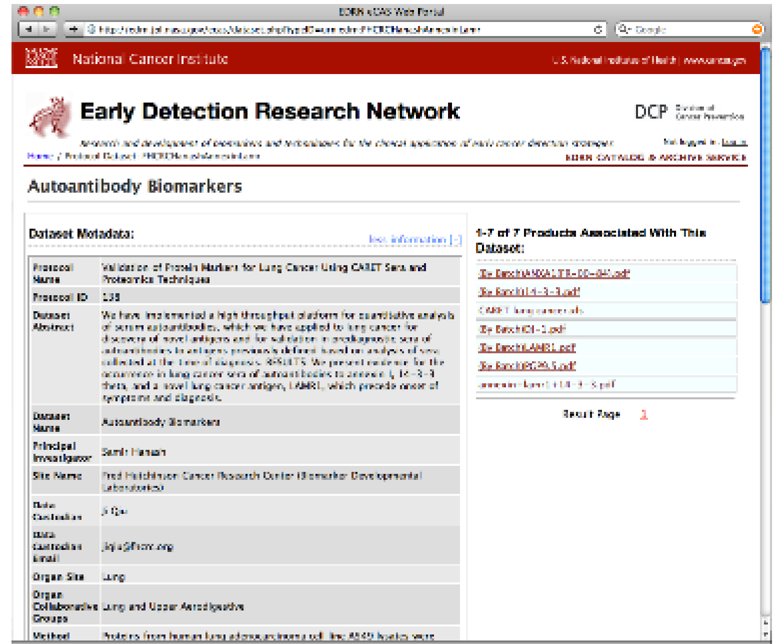
- NASA-NCI partnership, leveraging informatics and data science technologies from planetary and Earth science
  - Reproducible, Big Data Systems for exploring the universe
  - Software and data science methodology transfer
  - Presented informatics collaboration at a congressional briefing in October 2015



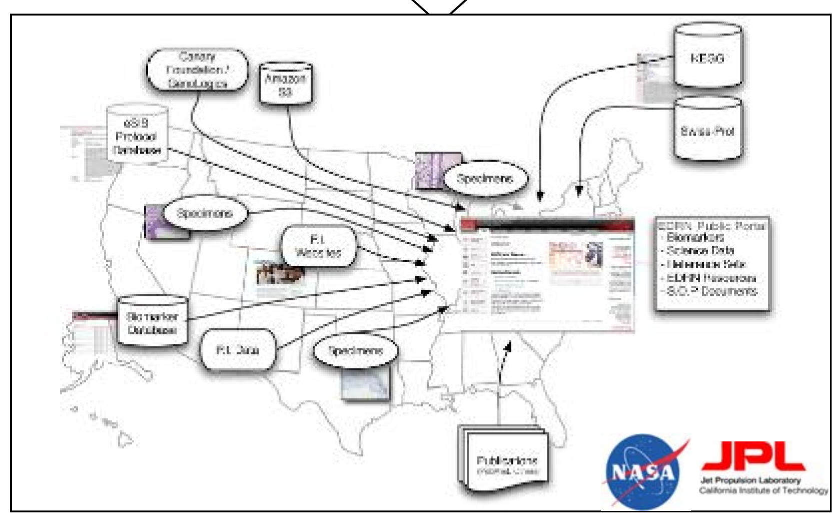
Early Detection Research Network: Finding Cancer Biomarkers

Jet Propulsion Laboratory
California Institute of Technology

- **A comprehensive infrastructure to support biomarker data management across EDRN's distributed cancer centers**
  - A national data sharing architecture
  - Data Integration
  - Information model for cancer biomarkers following the PDS4 approach
  - Development of data analytic pipelines
  - Shared open source software capabilities



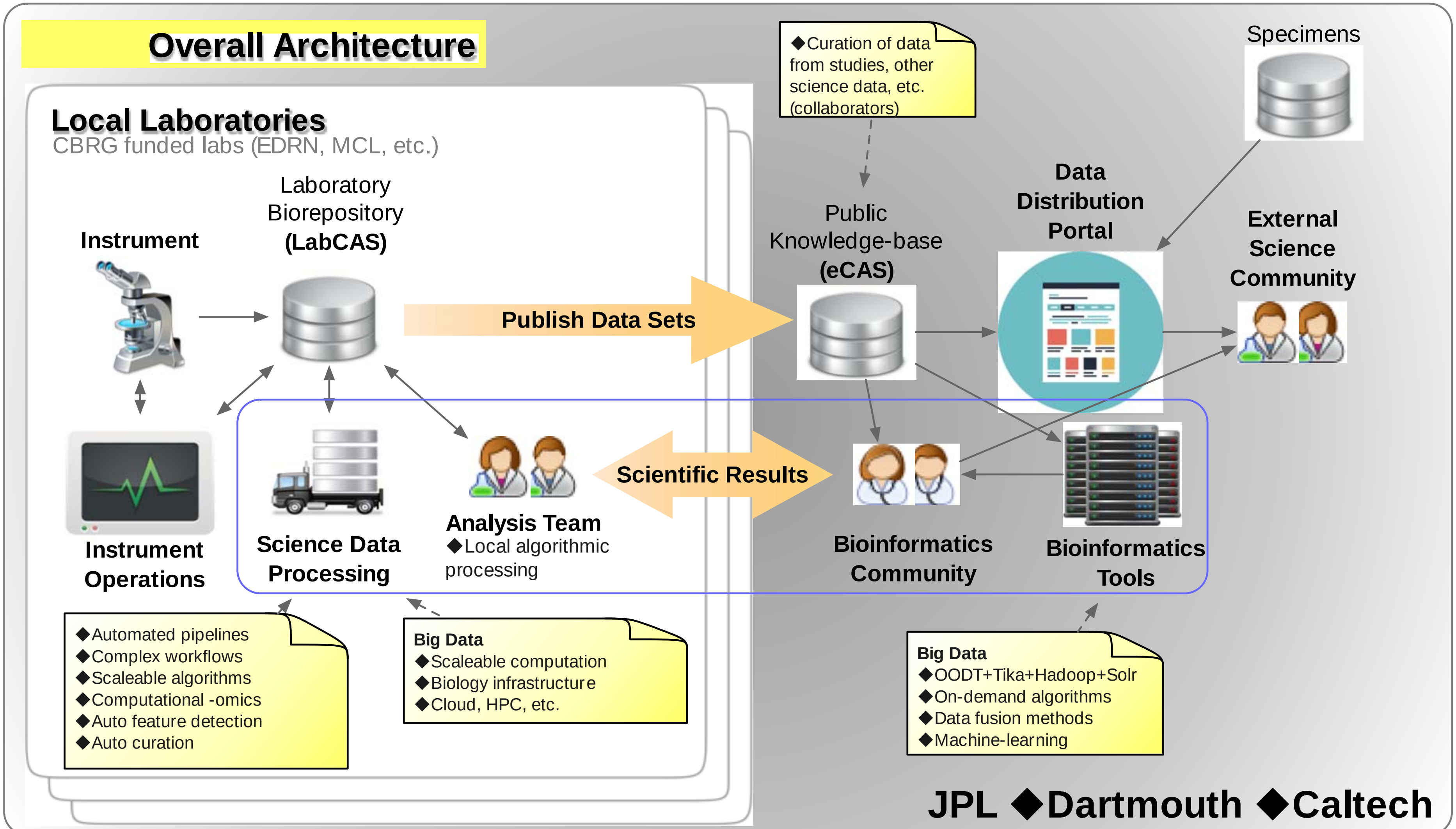
http://cancer.gov/edrn

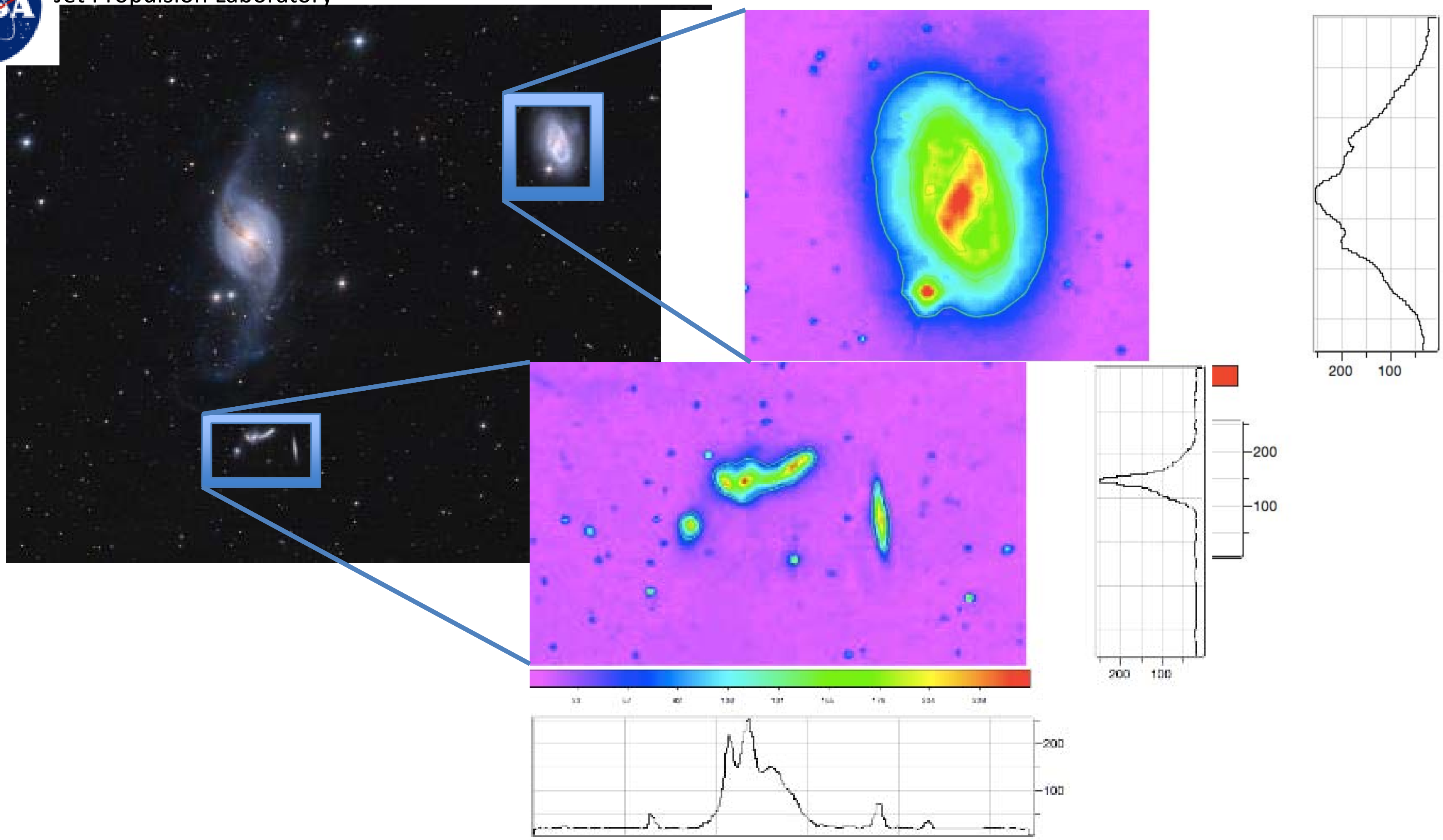


- **Integration of data across the EDRN (biomarkers, specimens, protocols, biomarker data, publications) including:**
  - Data from over 100 research labs; multiple organs
  - 800+ data elements
  - 900+ biomarkers captured
  - 200+ protocols of study
  - 1500+ publications
  - Multiple terabytes of data from biomarker studies

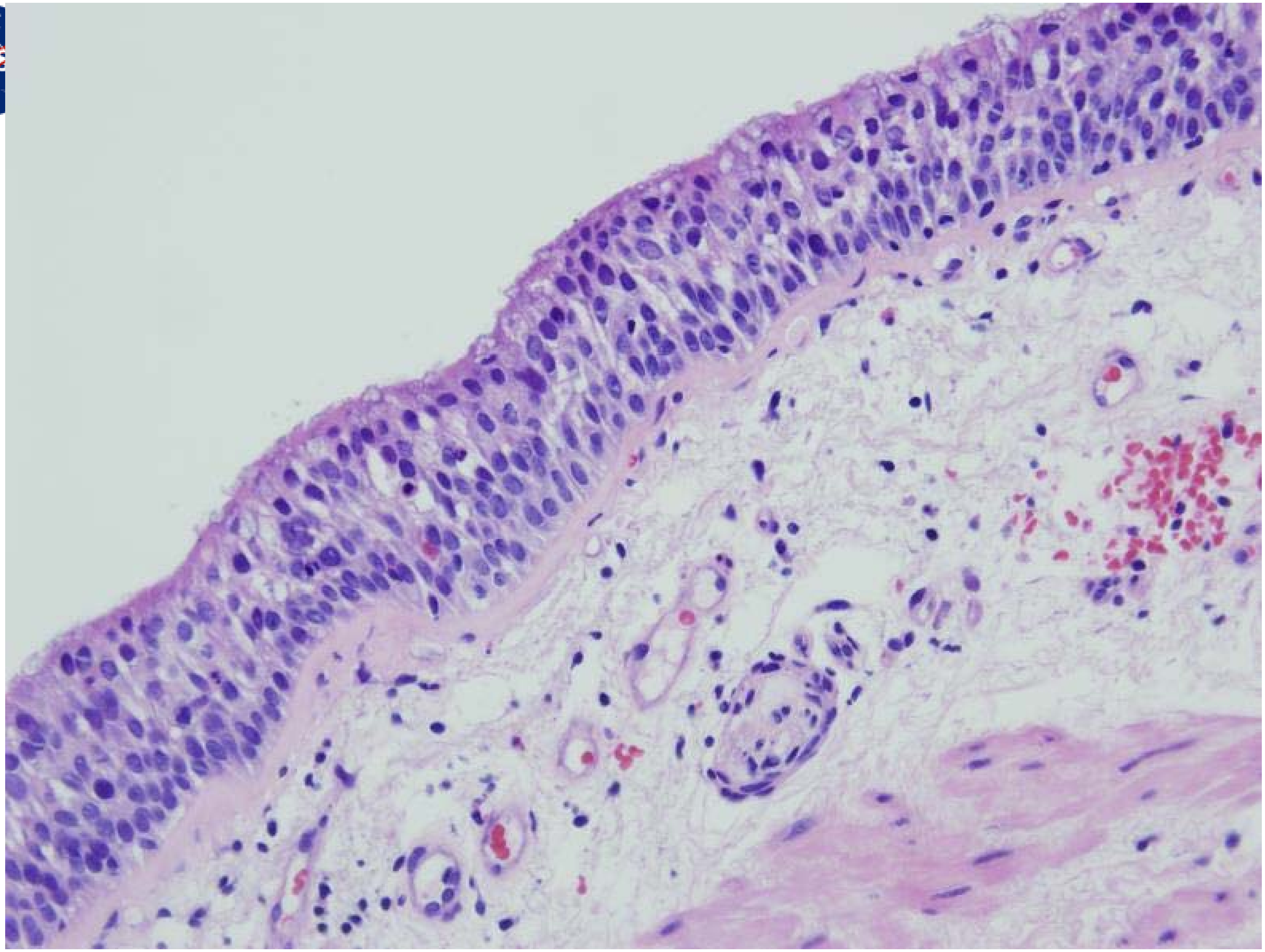
Example of Data Science Capabilities in Cancer Research from NASA

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

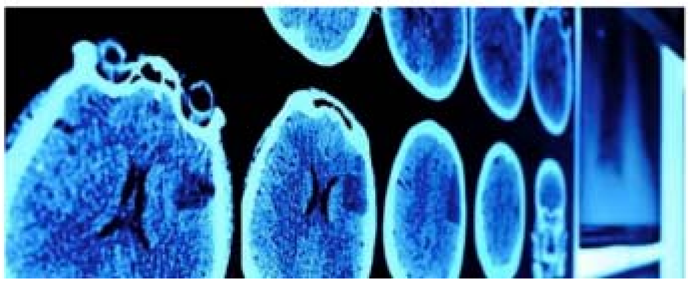
INNOVATION

10 ways tech is improving cancer research

New advances in cancer diagnosis and treatment leverage and even NASA tools to help detect and beat the disease.

By Alison DeNisco | September 22, 2016, 6:31 AM PST

0 | 114 | 21



Sep 22, 2016

2. NASA: Using space technology to find cancer markers

A NASA machine learning algorithm that identifies similarities between galaxies will now analyze tissue samples for signs of cancer. Earlier this month, NASA's Jet Propulsion Laboratory and the National Cancer Institute renewed a research partnership through 2021 to collect research on these biomarkers into one searchable network. This way, physicians can compare, for example, a CT scan with an archive of similar images to search for early signs of cancer, based on a patient's demographics. Ultimately, this could translate into new techniques for early diagnosis of cancer or cancer risk.

More about Innovation

- When your driverless car crashes, who will be responsible? The answer remains unclear
- GE makes $1.4B bet on 3D printing, acquires two firms to boost additive manufacturing
- IoT helping Tassie oyster farmers avoid unnecessary closures
- Subscribe to TechRepublic's Next Big Thing newsletter.

Dozens of institutions, including Dartmouth College's Geisel School of Medicine, Harvard Medical School's Massachusetts General Hospital, and Stanford's NIST Genome-Scale Measurements Group have joined the network. It is similar to NASA's Planetary Data System, in which all can share information.

Jet Propulsion Laboratory
California Institute of Technology

Other Partnerships

Searching deep and dark: Building a Google for the less visible parts of the web

January 8, 2017 8.33pm EST

A geographical map depicting hotbeds of dark web activity related to illegal products. Larger circles indicate more activity. Christian Mattmann, CC BY

DARPA/Memex, C. Mattmann, JPL

SPAWAR

SPAWAR/Data Science for C4CSI, L. Deforrest, JPL

Earth System Grid Federation

Earth System Grid Federation

An open source effort providing a robust, distributed data and computation platform, enabling world wide access to Peta/Exa-scale scientific data

Learn more »

DOE/ESGF, L. Cinquini, JPL

UNC CHARLOTTE WINS $4 MILLION NSF GRANT FOR BIG DATA RESEARCH

- Search News and Features
- Archive News and Features
- UNC Charlotte wins $4 million NSF grant for Big Data research
- Faculty Spotlights
- Student Spotlights
- Research

Tuesday, September 13, 2016

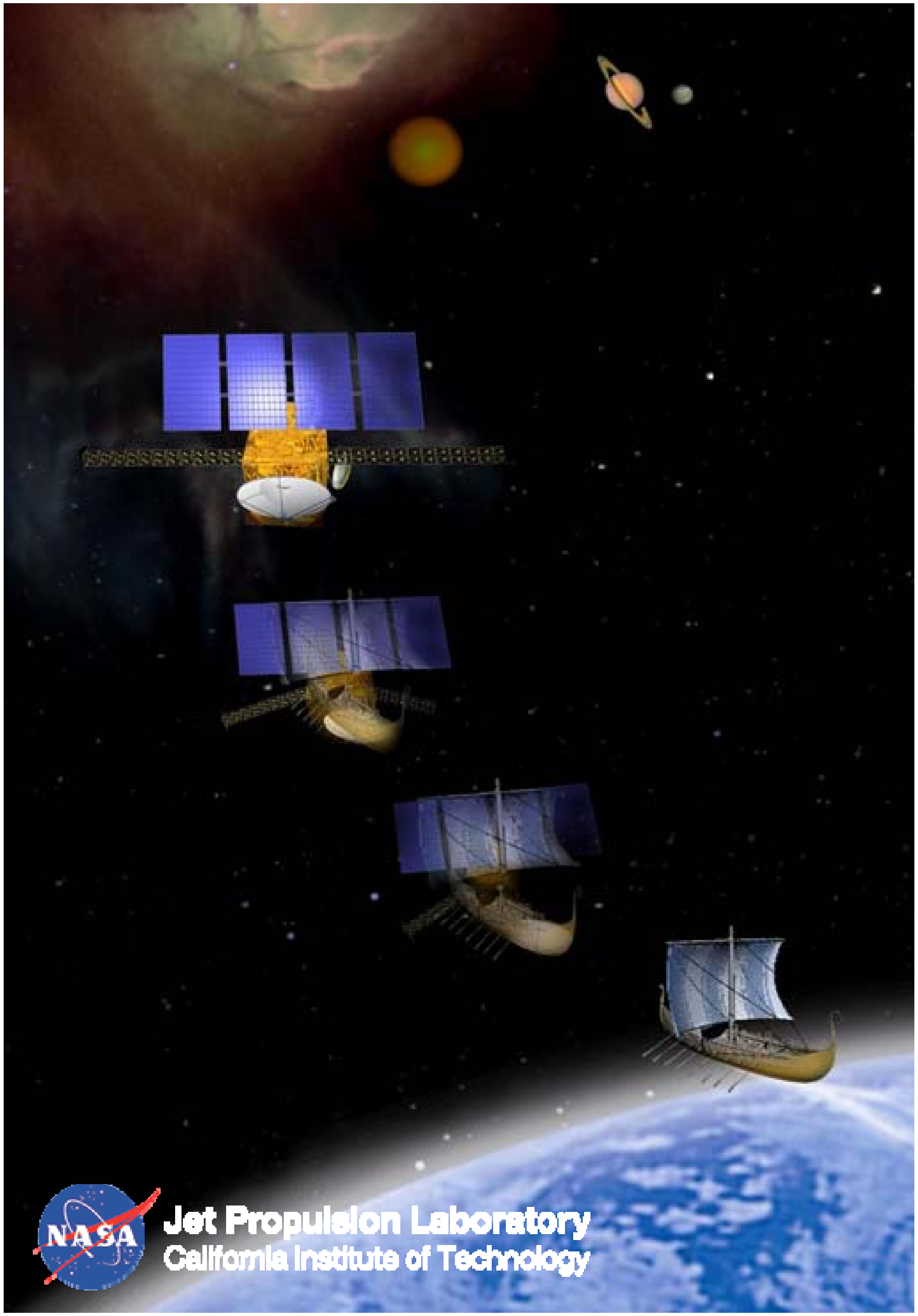
Ashit Talukder

The National Science Foundation has awarded a $4 million grant to UNC Charlotte researchers to develop a multidisciplinary research program called Virtual Information Fabric Infrastructure (VIFI) that will create new ways to manage, use and share Big Data and analytic results

Ashit Talukder, director of the Charlotte Data Visualization Center and the Bank of America Endowed Chair in Information Technology in the College of Computing and Informatics, is the principal investigator for the grant. The award was made under the NSF-CISE/ACI-Data Infrastructure Building Blocks (DIBBS) solicitation.

"Under this large-scale research program, a novel Virtual Information Fabric Infrastructure (VIFI) will be created, allowing scientists to search, access, manipulate and evaluate fragmented, distributed data in the information 'fabric' (the infrastructure to facilitate data sharing) without directly accessing or moving large amounts of data," said Talukder.

NSF/DIBBS, A Talukder, UNC, G. Djorgovski, Caltech, D. Crichton, JPL



Driving Forward

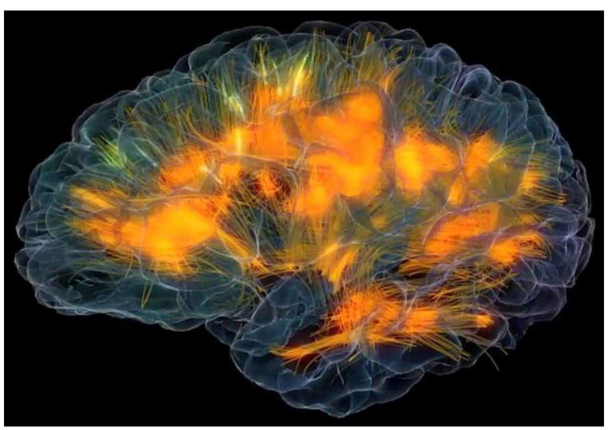
Jet Propulsion Laboratory
California Institute of Technology

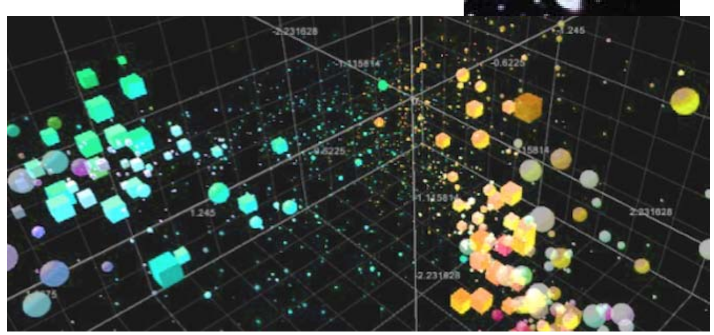
Caltech-JPL
Partnership in Data Science

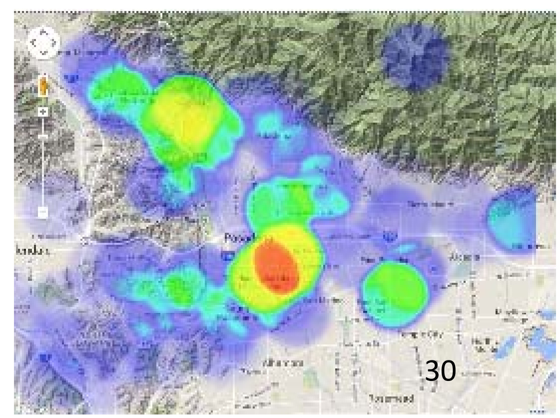
- Center for Data-Driven Discovery on campus/Center for Data Science and Technology at JPL
- From basic research to deployed systems ~10 collaborations
  - Leveraged funding from JPL to Caltech; from Caltech to JPL
- Virtual Summer School (2014) has seen over 25,000 students

CENTER FOR DATA-DRIVEN DISCOVERY

CDST







Jet Propulsion Laboratory
California Institute of Technology

Example University Partnerships

bigdata@CSAIL
MIT BIG DATA INITIATIVE

CSAIL

ABOUT PEOPLE EVENTS PARTNERS SPECIAL PROJECTS LOGIN

TACKLING BIG

MIT B | Tackl | MIT B

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, Prinicpal 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

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.

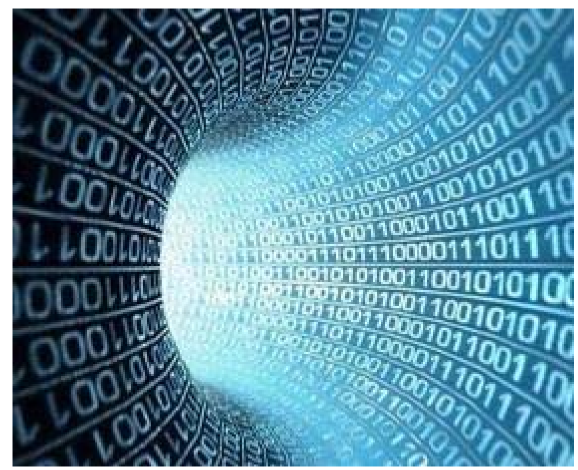
The grant will also allow 22 high school students from Riverside Unified School District to take a STEM (Science, Technology, Engineering, Mathematics) class at UC Riverside this summer.

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

Recommendations

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

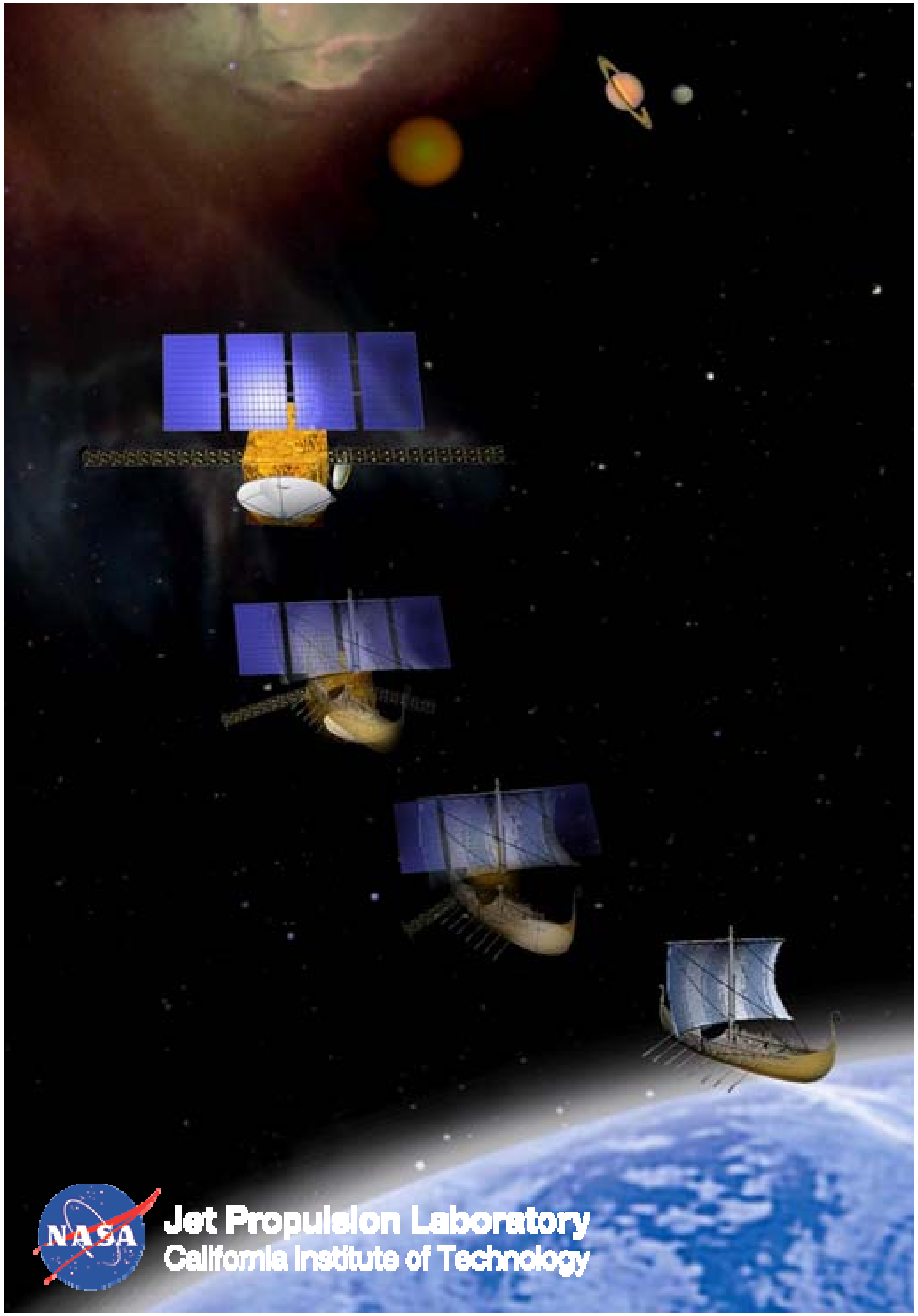
What do we do with all this data?



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


NASA
Jet Propulsion Laboratory
California Institute of Technology

Questions?

Jet Propulsion Laboratory
California Institute of Technology

The Role of Open Source in Big Data Infrastructures

- Open source is an excellent vehicle for collaborations in big data across the science community
  - Great opportunities for sharing software frameworks and tools

- JPL has been involved in the Apache Software Foundation for several years and helped launch Apache in Science.
  - JPLers are *committers* on several Apache projects

APACHE Spark

APACHE OPEN CLIMATE WORKBENCH

APACHE OODT

Apache Tika

hadoop

Common Big Data Challenges

- Defining the data lifecycle for different domains in science, engineering, business
- Capturing well-architected and curated data repositories
- Enabling access and integration of highly distributed, heterogeneous data
- Developing novel statistical approaches for data preparation, integration and fusion
- Supporting analysis and computation across highly distributed data environments and silos
- Developing mechanisms for identifying and extracting interesting features and patterns
- Developing methodologies for validating and comparing predictive models vs. measurements
- Methods for visualizing massive data