Data Science: Unlocking Scientific Research from Space to Biology

2018 Jean Golding Institute Showcase
University of Bristol

Dan Crichton, Leader, Center for Data Science and Technology

July 3, 2018

Jet Propulsion Laboratory
California Institute of Technology

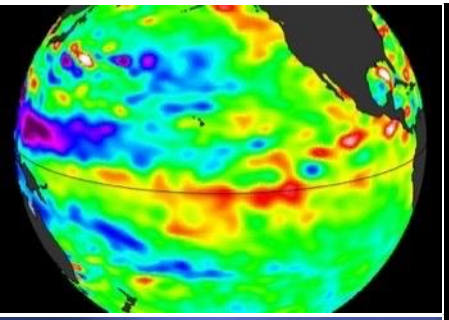
Overview

- JPL's Unique Mission
- The Data Challenge
- Autonomy
- Data Systems
- Data Analytics
- Methodology Transfer

JPL's Mission for NASA

Robotic Space Exploration

Earth Science • Mars • Solar System • Astrophysics • Exoplanets • Interplanetary network






Our mission has introduced unique challenges for protecting space system assets and information

JPL: From Caltech students testing rockets to the planets and universe in our lifetimes

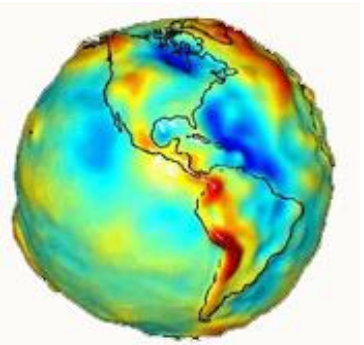
Caltech students (1936)

Missiles (1940s)

Explorer 1 (1958)

Mars Exploration Rovers
(2004 – present)

Spitzer Space Telescope
(2004 – present)



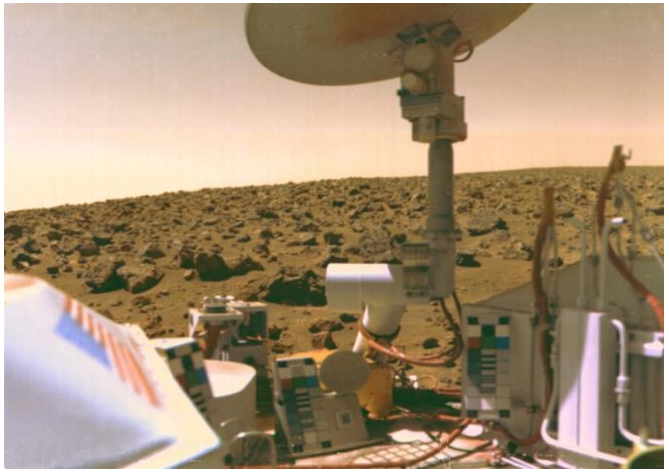
Earth Science
(1978 – now)

JPL is a NASA FFRDC and a Division of Caltech

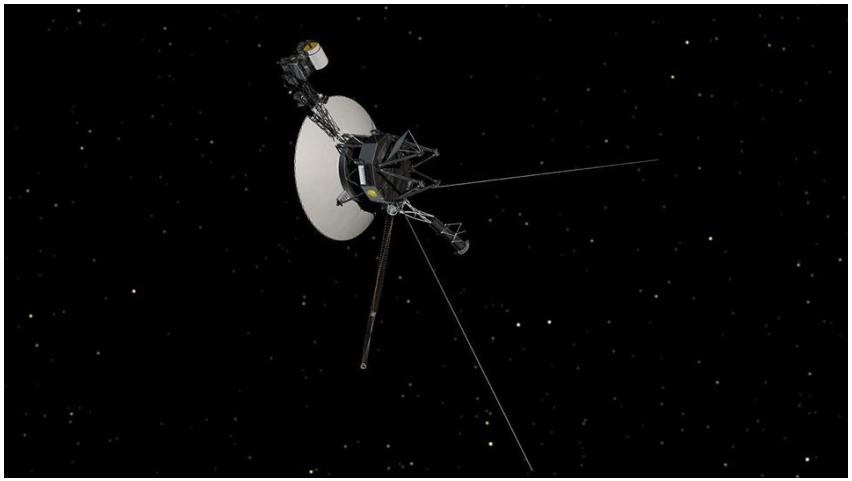
A History of Firsts…

Surveyor 1, First soft landing on the moon

Viking, first landing on another planet

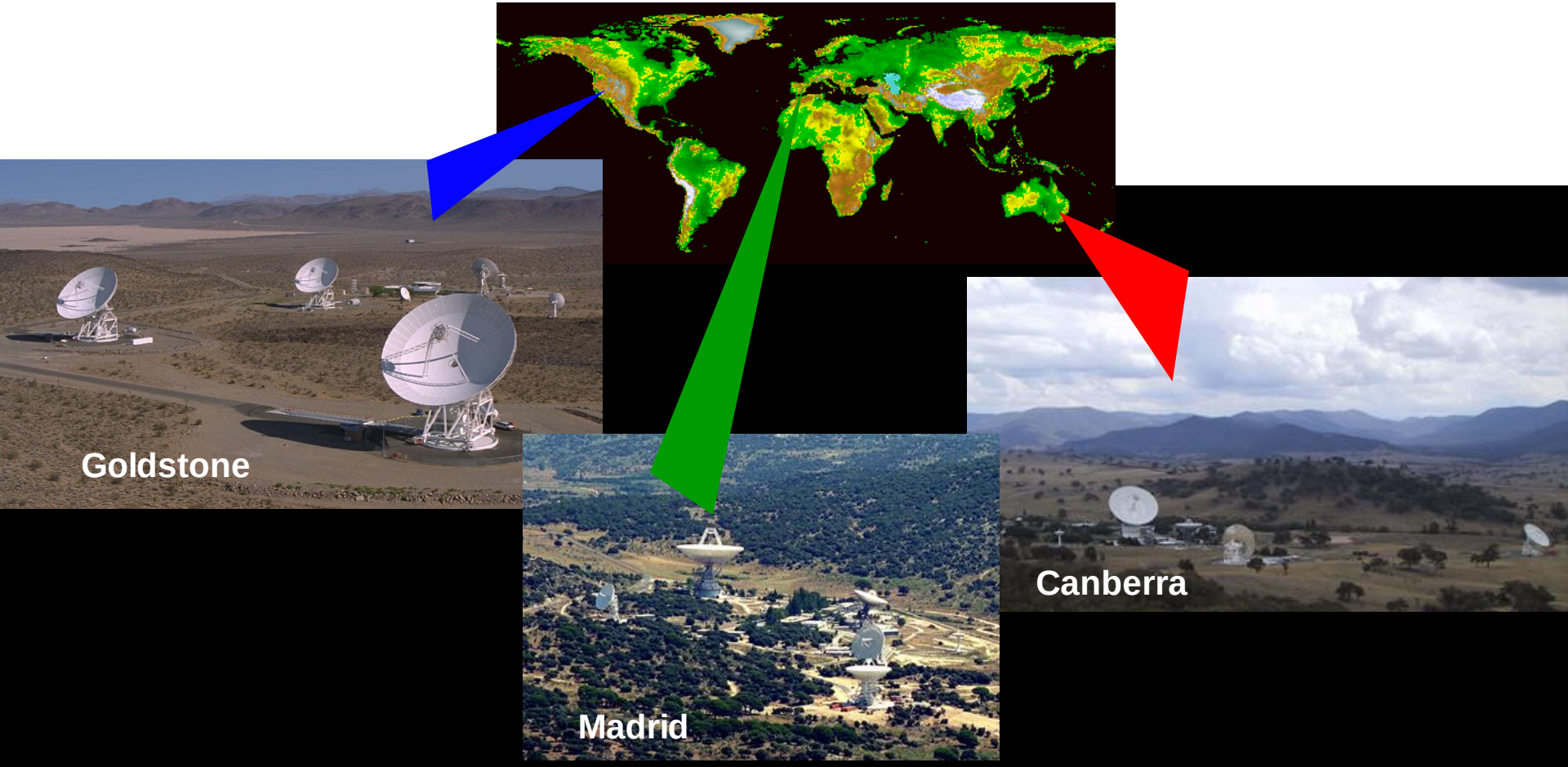


Voyager 1, First interstellar traveler



Continuous presence on Mars since 1997

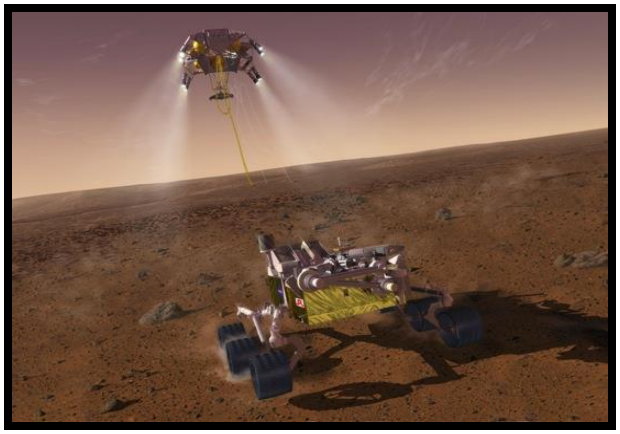
Deep space exploration enabled by NASA's Deep Space Network (DSN)



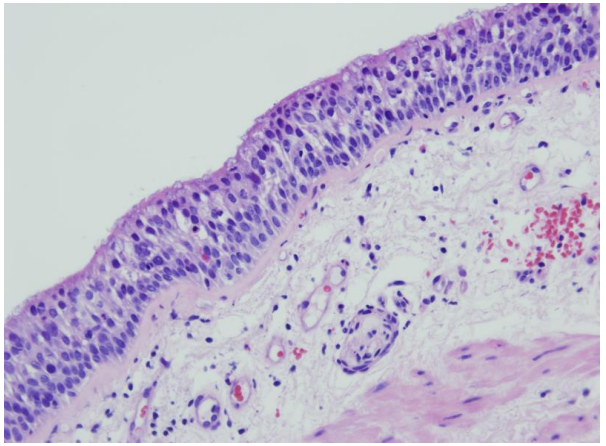
The Data Challenge

Tackling Big Data

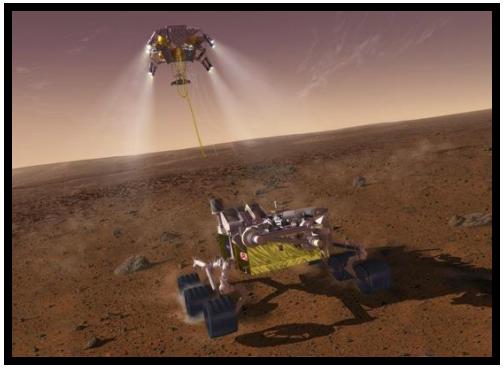
- JPL is involved in the research and development of technologies, methodologies in science, mission operations, engineering, and other non-NASA applications.
  - Includes onboard computing to scalable archives to analytics



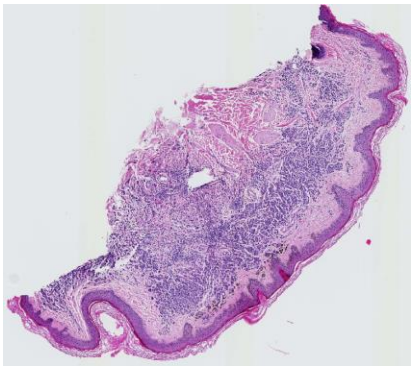
- JPL and Caltech formed a joint initiative in Data Science and Technology to support fundamental research all the way to operational systems.
  - Methodology transfer across applications is a major goal.



Data Science Projects at JPL



Planetary Science



Biology

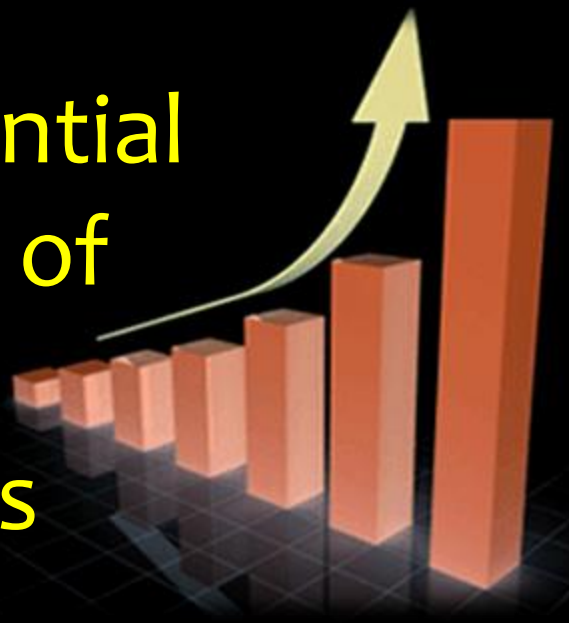
Defense and Intelligence

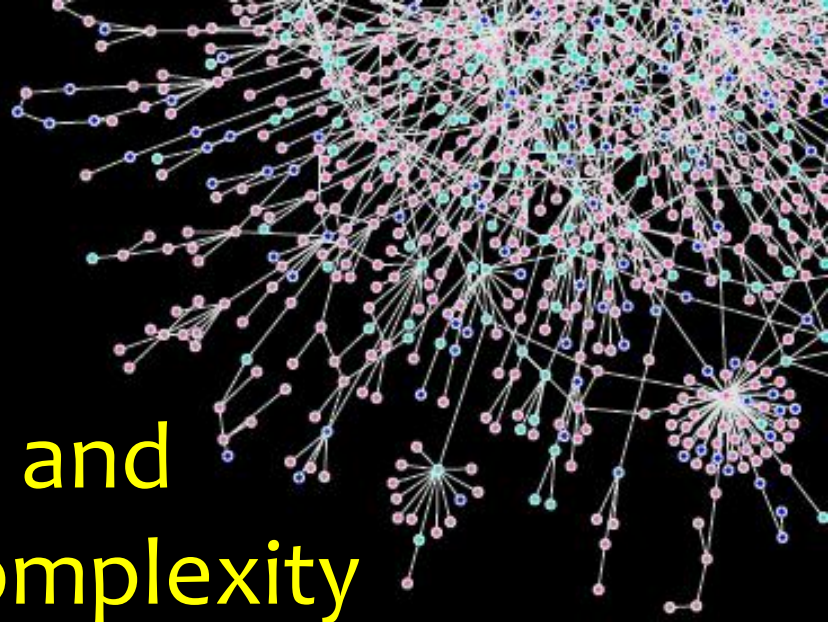
Earth Science

Medicine

Radio Astronomy

Exponential Growth of Data Volumes





... and Complexity

on Moore's law time scales

Understanding of complex phenomena requires complex data!

From data poverty to data glut
From data sets to data streams
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise

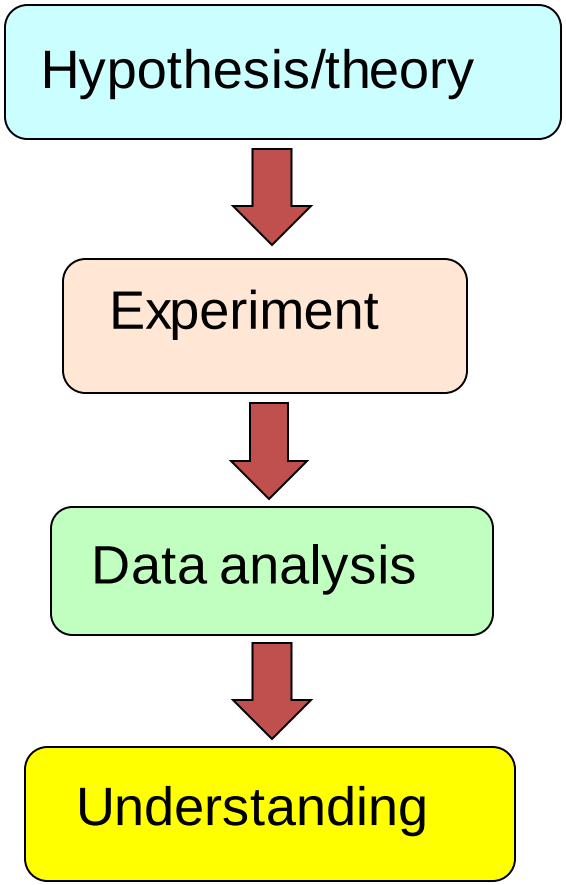
Transformation and Synergy: Big Data to Data Science

- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) - and with this change comes the need for ***a new scientific methodology***

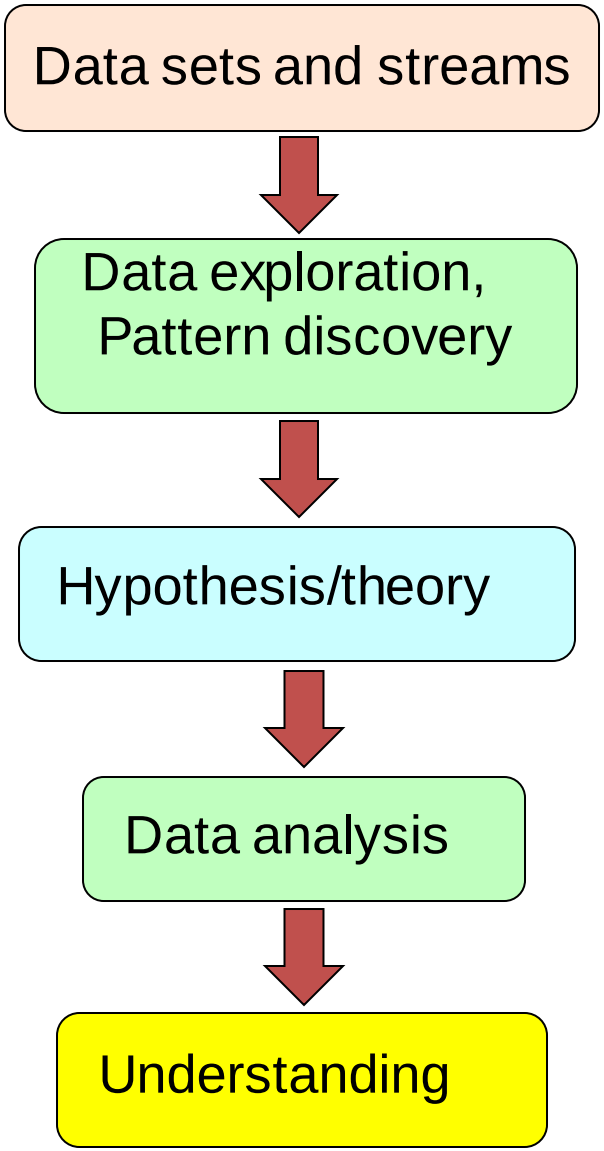
- The challenges we are tackling:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data new knowledge
  - **These challenges are universal**

- A great synergy of the computationally enabled science, and the science-driven informatics

Hypothesis-driven science



Data-driven science



The two approaches are complementary

A Modern Scientific Discovery Process

Data Gathering (instruments, sensor networks, their pipelines…)

↳ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability

} Databases, Data grids, Data commons

↳ **Data Mining**

Pattern or correlation search
Clustering analysis, classification
Outlier / anomaly searches
Hyperdimensional visualization

↳ **Data Understanding**

↳ **New Knowledge**

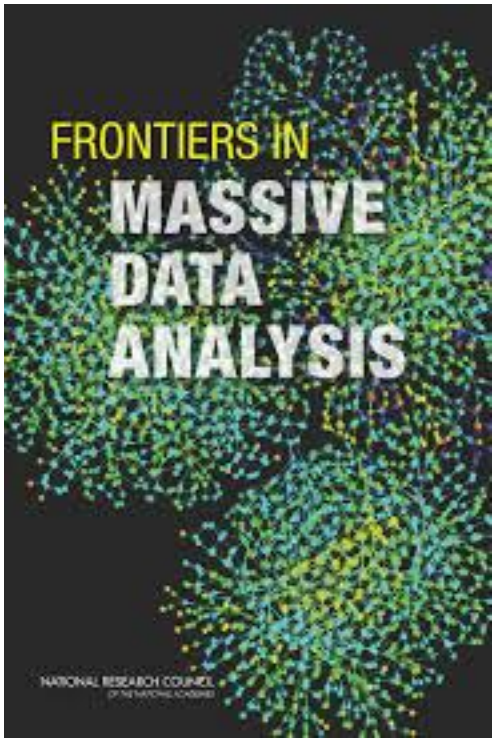
+feedback




U.S. National Research Council Report:

Frontiers in the Analysis of Massive Data

- Chartered 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

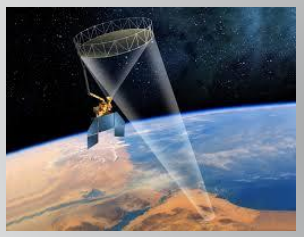


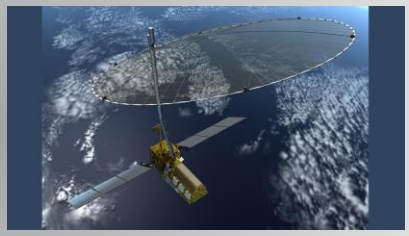
2013

NASA Big Data Lifecycle Model

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

Observational Platforms and Flight Computing

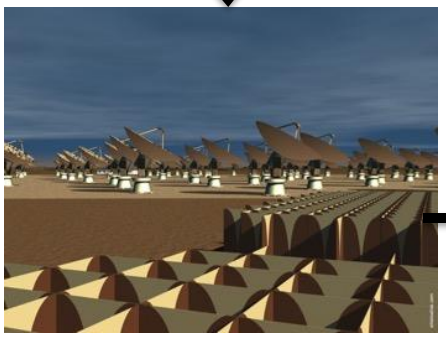




SMAP (Today): 485 GB/day **NI-SAR (2020): 86 TB/day**

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

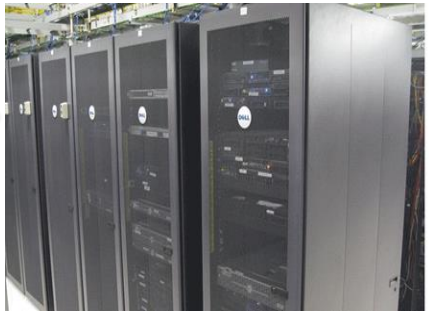
Emerging Solutions
- *Intelligent Ground Stations*
- *Agile MOS-GDS*



Ground-based Mission Systems

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

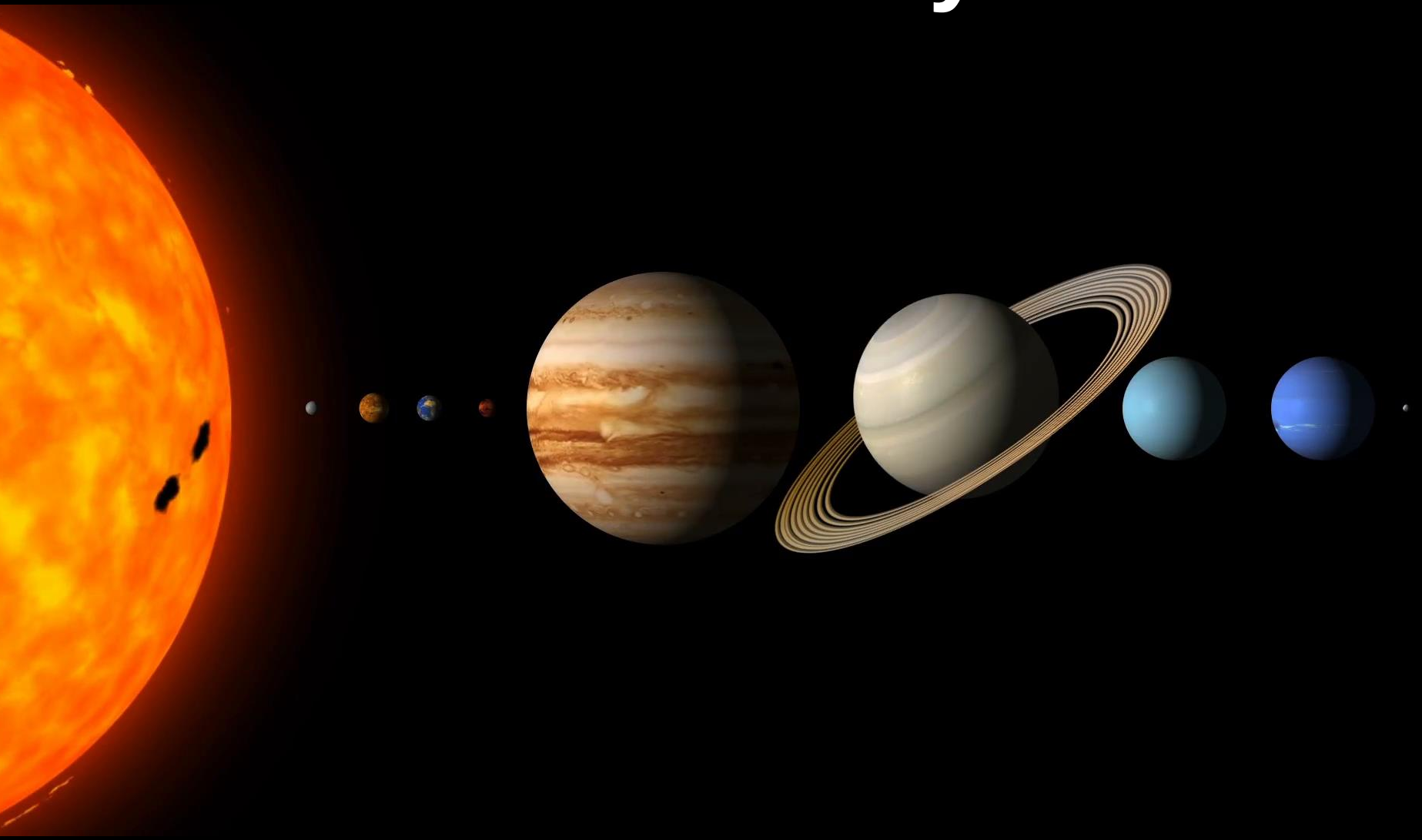
Massive Data Archives and Big Data Analytics



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

Autonomy



Increasing Computing Capability Onboard

Heading Toward Multicore in Space

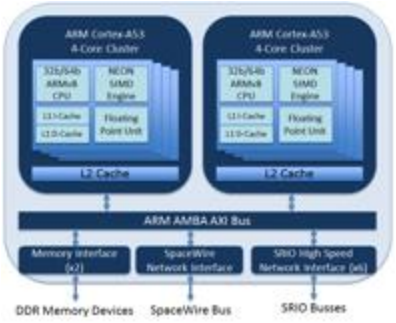
Voyager computer

- 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 / USAF)
Processor: 15 GOPS, extensible

Surface Mobility

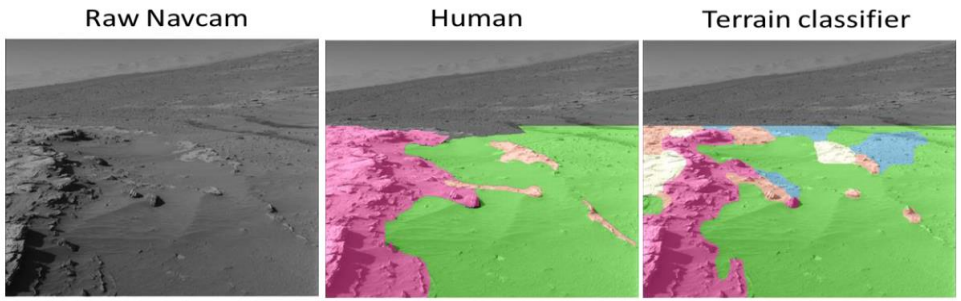
Mars Rover Navigation

Flight Deployed

- **1996 Mars Pathfinder:** obstacle avoidance with structured light
- **2003 Mars Exploration Rover:** obstacle avoidance with stereo vision; pose estimation and slip detection with visual odometry; goal tracking
- **2011 Mars Science Laboratory:** enhanced obstacle avoidance, visual odometry and goal tracking

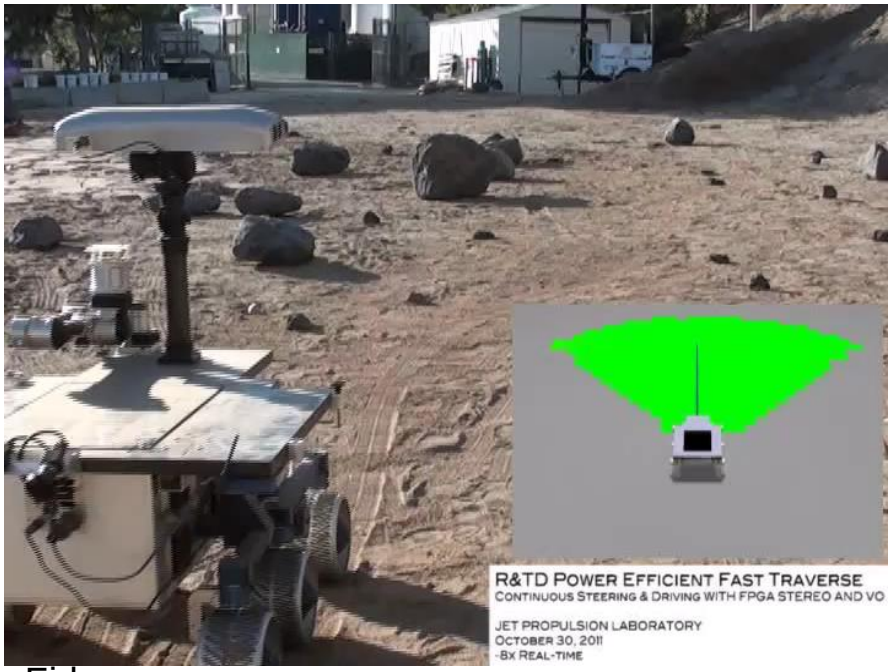
Research and Development

– Enhanced hazard detection, traversability analysis and motion planning for Mars 2020 and beyond



Terrain Classification

Athena



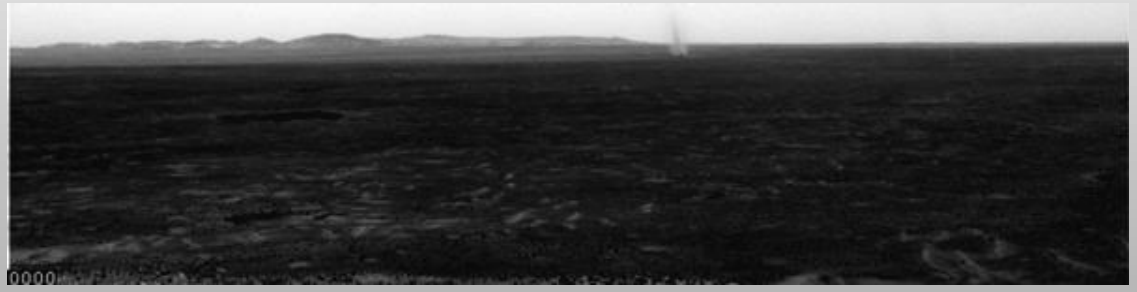
Fido

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



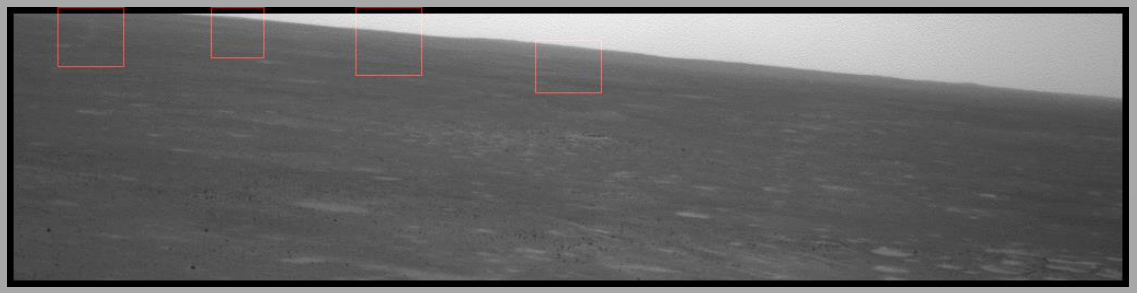
Spirit Sol 543 (July 13, 2005)

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*

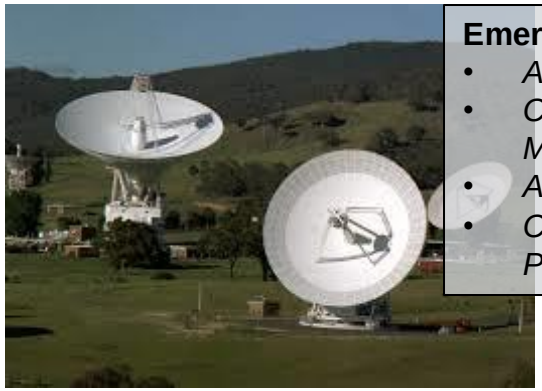


Credit: T. Estlin. B. Bornestein, A. Castano, J. Biesiadecki, L. Neakrase, P. Whelley, R. Greeley, M. Lemmon, R. Castano, S. Chien and MER project team

Data Systems

Data-Driven Capabilities 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

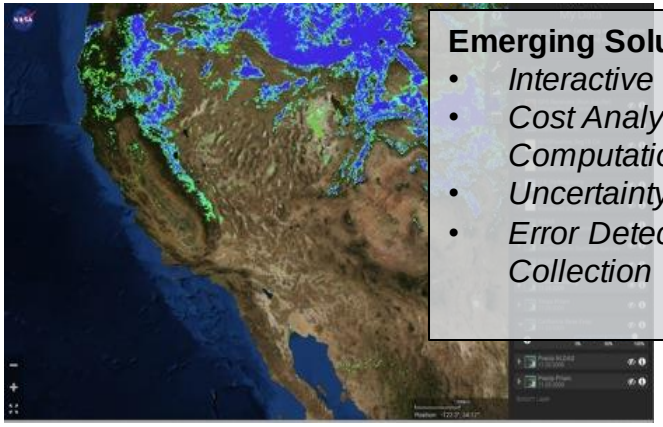
Emerging Solutions
- *Automated Machine Learning - Feature Extraction*
- *Intelligent Search*
- *Integration of disparate data*

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*

Data-Driven Approaches for Deep Space Communication: Detecting Anomalies

Current Inputs: DSN operationally relevant data

Real Time Monitor Data

Real Time Operator Logs

Track Predicts

DSN Software Quality Assessment (SQA) Data Archive

- Relational database
- 10 years of data
- 1.3+ billion records

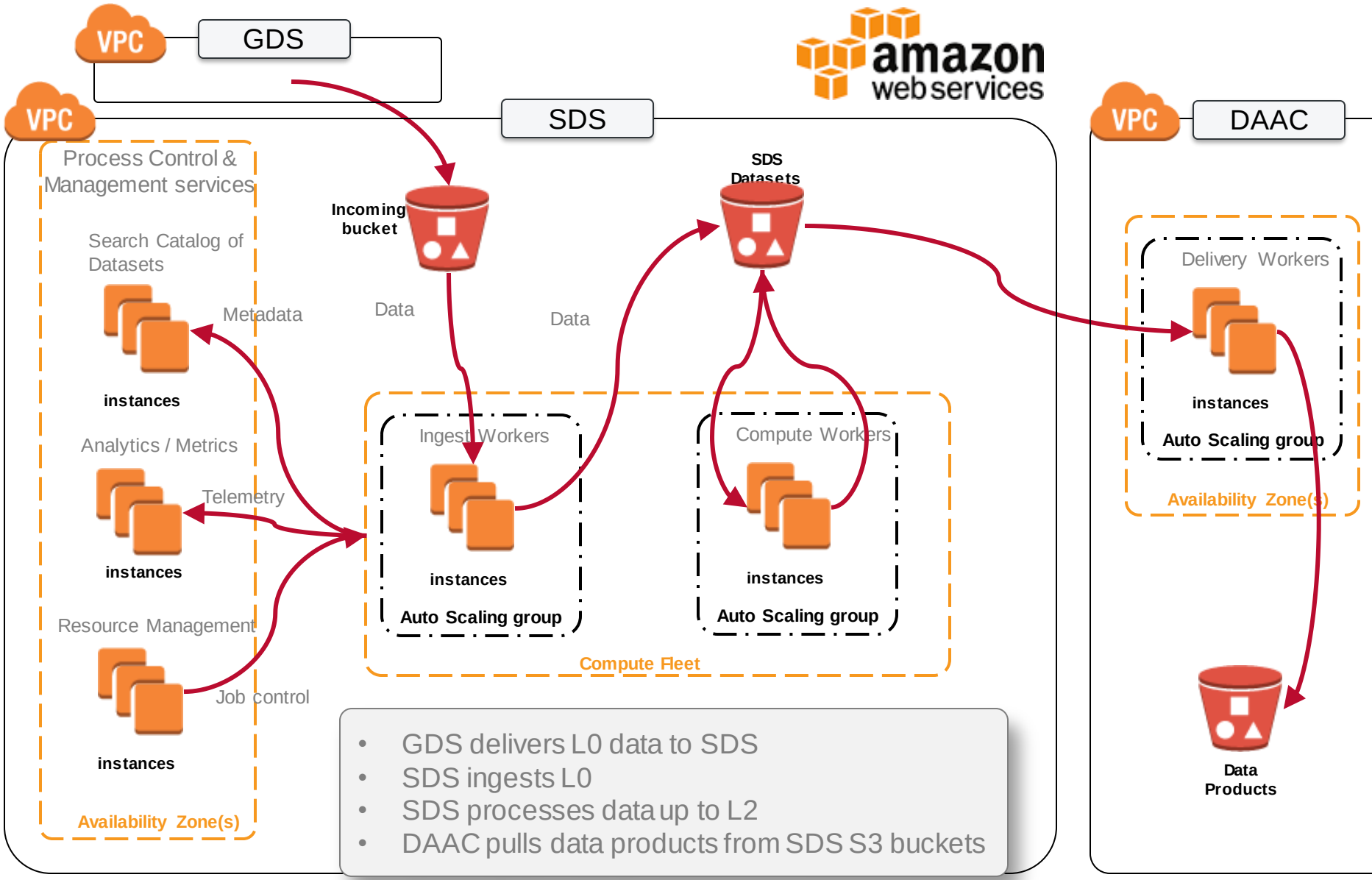
Desired Output: Better Fault Detection and Diagnosis

Real-Time, Historically Informed Alerts

Real-Time Insight into Data Points' Criticality and Relationships

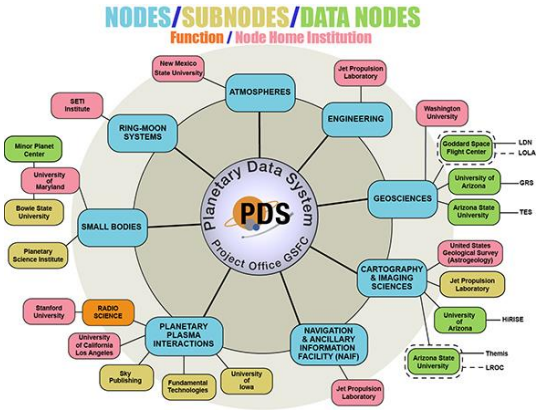
Credit: Rishi Verma, JPL

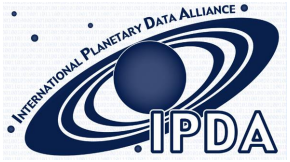
Scaling Processing of NISAR in AWS Cloud



POC: Hook Hua

NASA Archives: Access to Data*



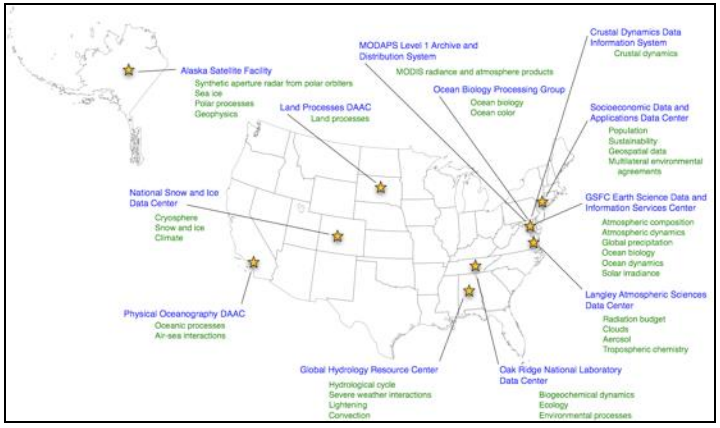


Planetary Science

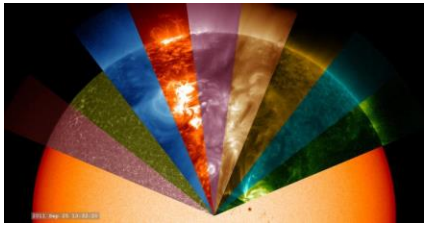
Highly distributed/federated
Collaborative
Information-centric
Discipline-specific
Growing/evolving
Heterogeneous
International Standards & Interoperability

Multiple Data Centers

Astronomy



Earth Observation

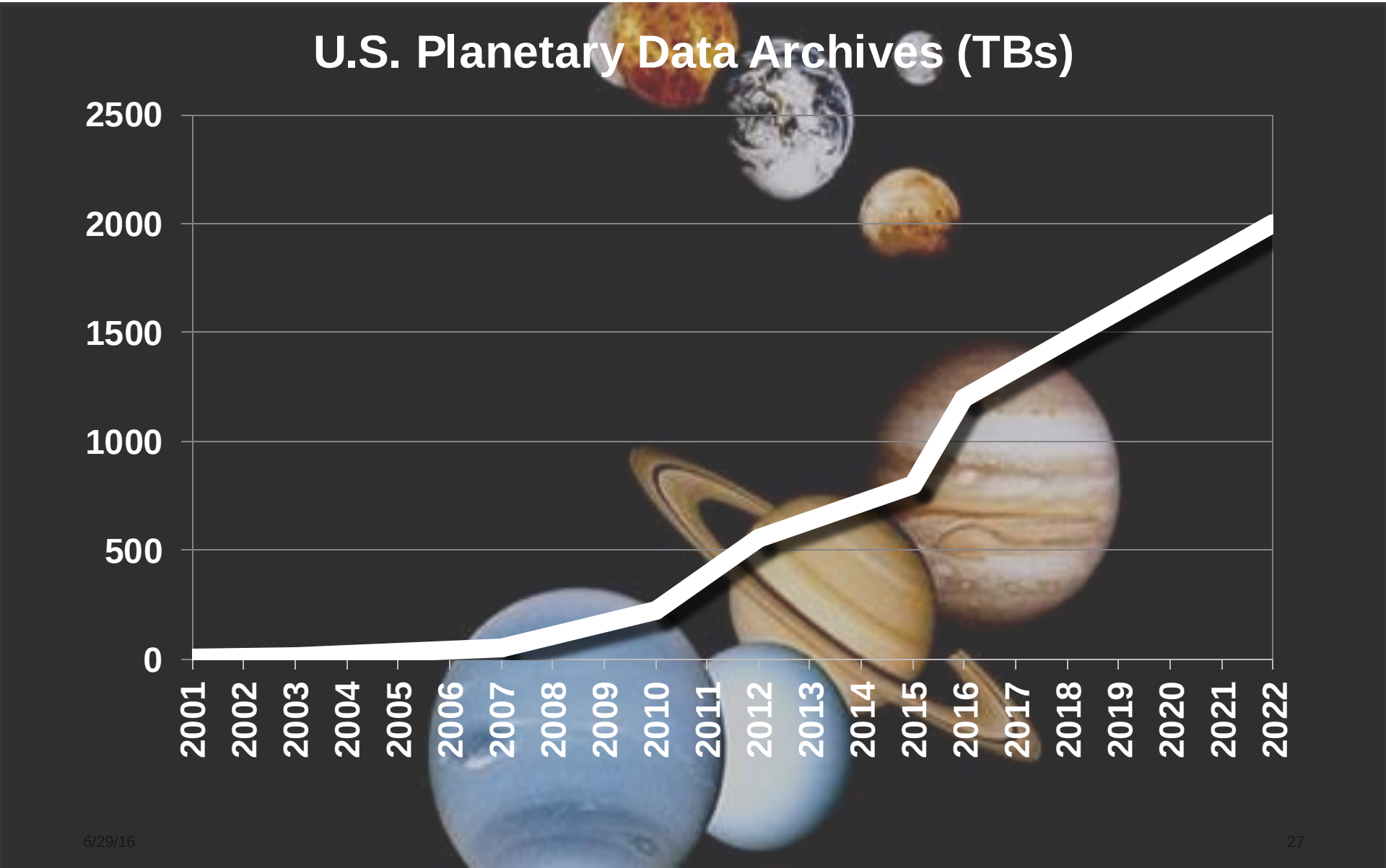


Multiple Data Centers

Heliophysics

* Petascale environment that is moving to an exascale requirement

Growth of Planetary Data Archived from U.S. Solar System Research

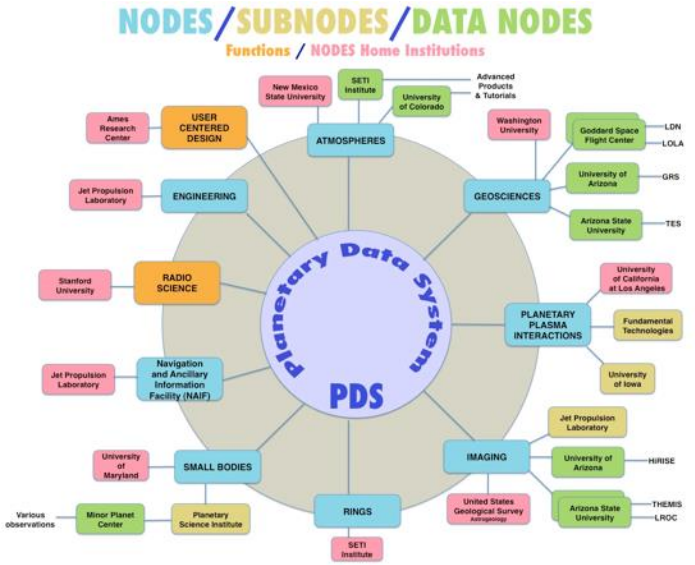


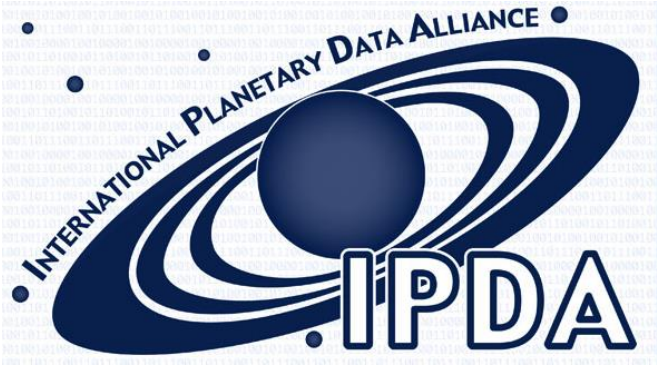
Planetary Data System

- 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
  - 4000 different types of data
  - Over 1.4 PB of data
  - International interoperability
  - Distributed federation of US nodes and international archives

- **Realized through an international data science platform!**

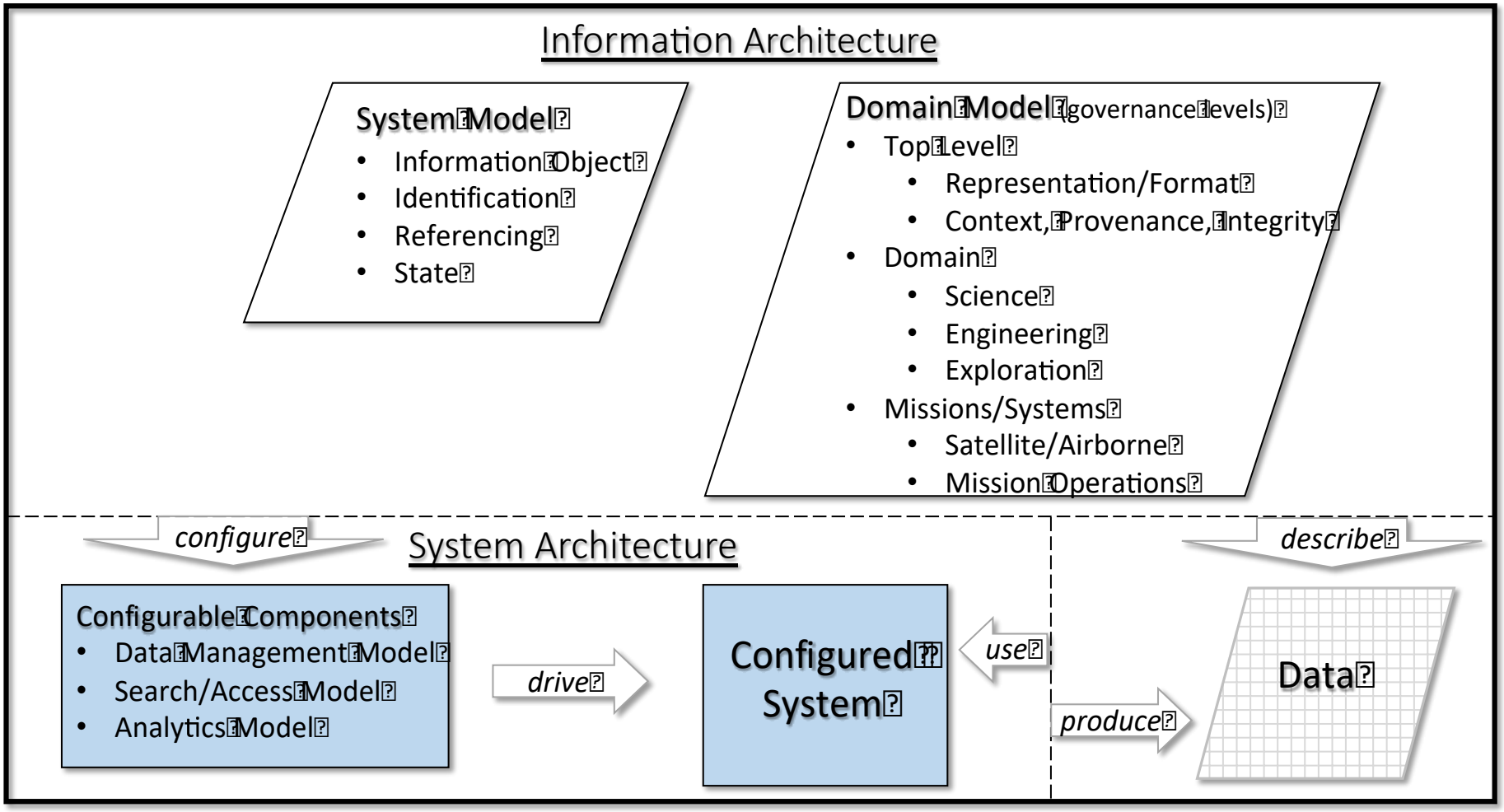




28

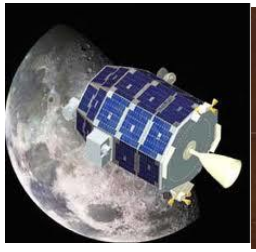
Enabling a *Metadata Model-Driven Data System*

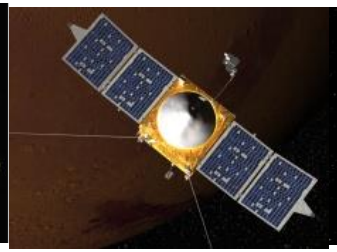
Information System Architecture

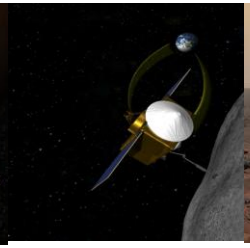


Crichton, D. Hughes, J.S. ; Hardman, S. ; Law, E. ; Beebe, R. ; Morgan, T.; Grayzeck, E. A Scalable Planetary Science Information Architecture for Big Science Data. IEEE 10th International Conference on e-Science, October 2014.

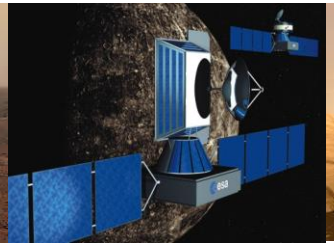
International Planetary Data Alliance: Collaboration and Access to Data Archives


LADEE
(NASA)


MAVEN
(NASA)


Osiris-REx
(NASA)

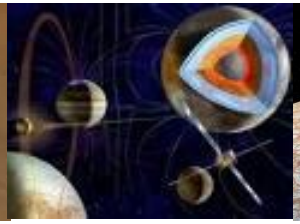

ExoMars
(ESA/Russia)

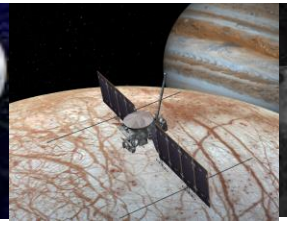

BepiColombo
(ESA/JAXA)


Mars 2020
(NASA)

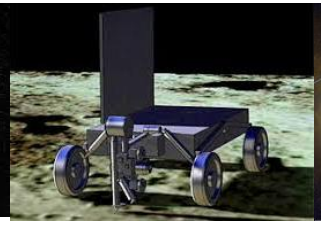

Psyche
(NASA)

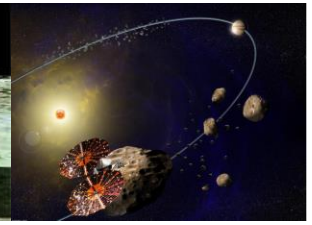

InSight
(NASA)


JUICE
(ESA)


Europa
(NASA)


Hyabusa-2
(JAXA)

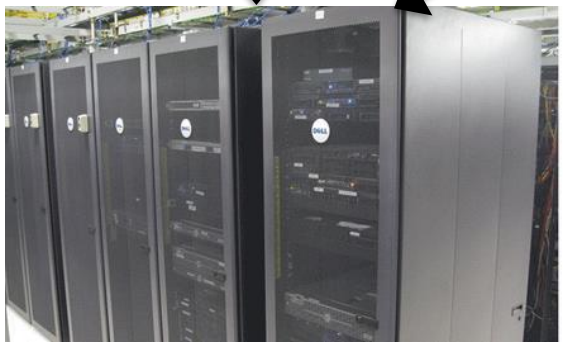

Chandrayaan-2
(ISRO)


Lucy
NASA

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

Data Analytics

Increasing Need for Data-Driven Analysis



Data Archiving and Stewardship of Massive Data

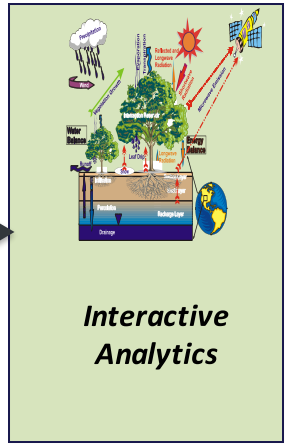
Today

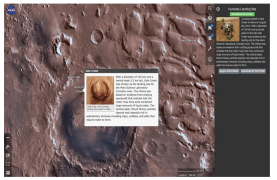
Machine Learning/ Deep Learning

Research/ Knowledge

Visualization

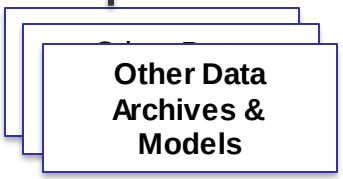
Scalable Cloud Infrastructures





Public Engagement

Data Integration

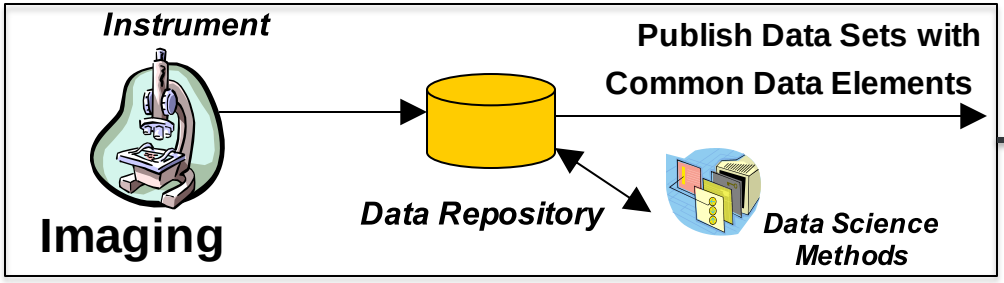


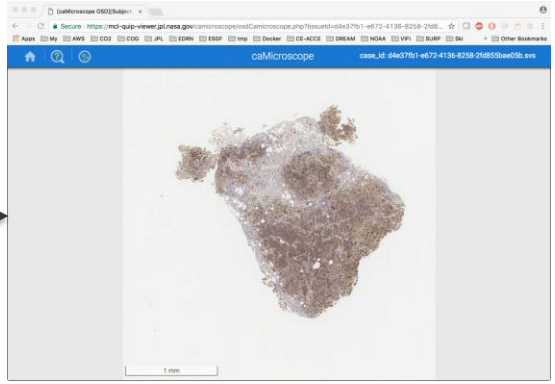
Data Analytics

Future

Setup Automated Data Analytic Pipelines (from archives to analytics)

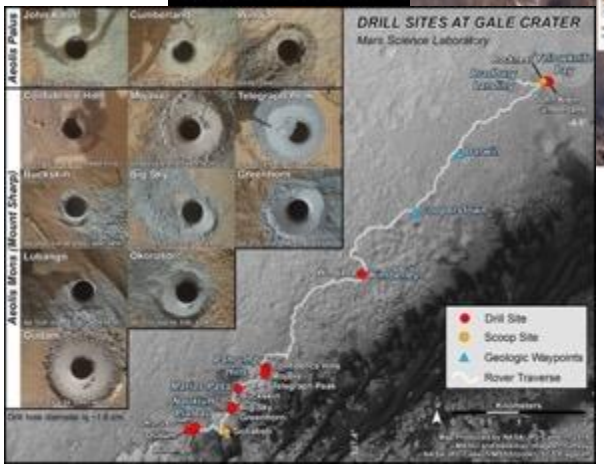
- Support construction of online data analytic capabilities from archives
  - Ad hoc workflows and data pipelines
  - Rapid integration of different methodologies (e.g., feature detection, classification, etc)
  - Establishment of highly scalable databases for analytics
  - Derive additional metadata to support analysis





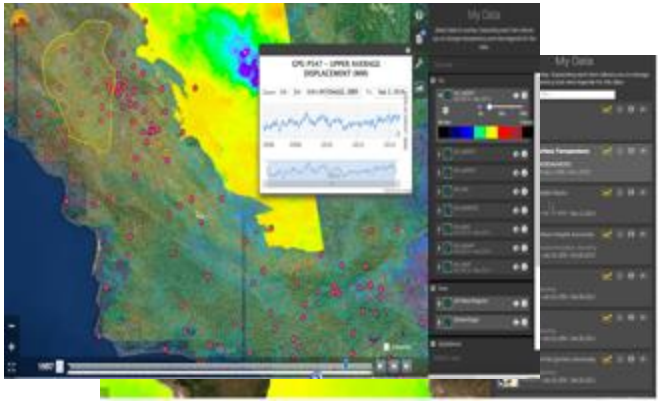
Visualization and Analytics

Interactive visualization of heterogeneous planetary objects

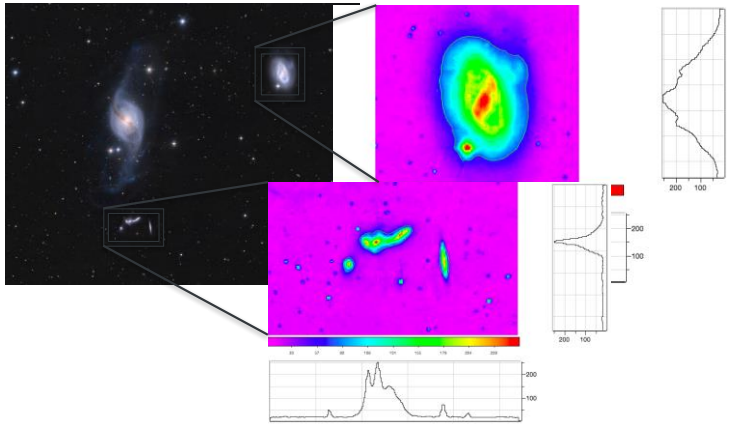


Examples: Planetary Image search, Mars and Moon surface navigation, feature extraction from Planetary images.

Examples: Hydrology and sea level rise

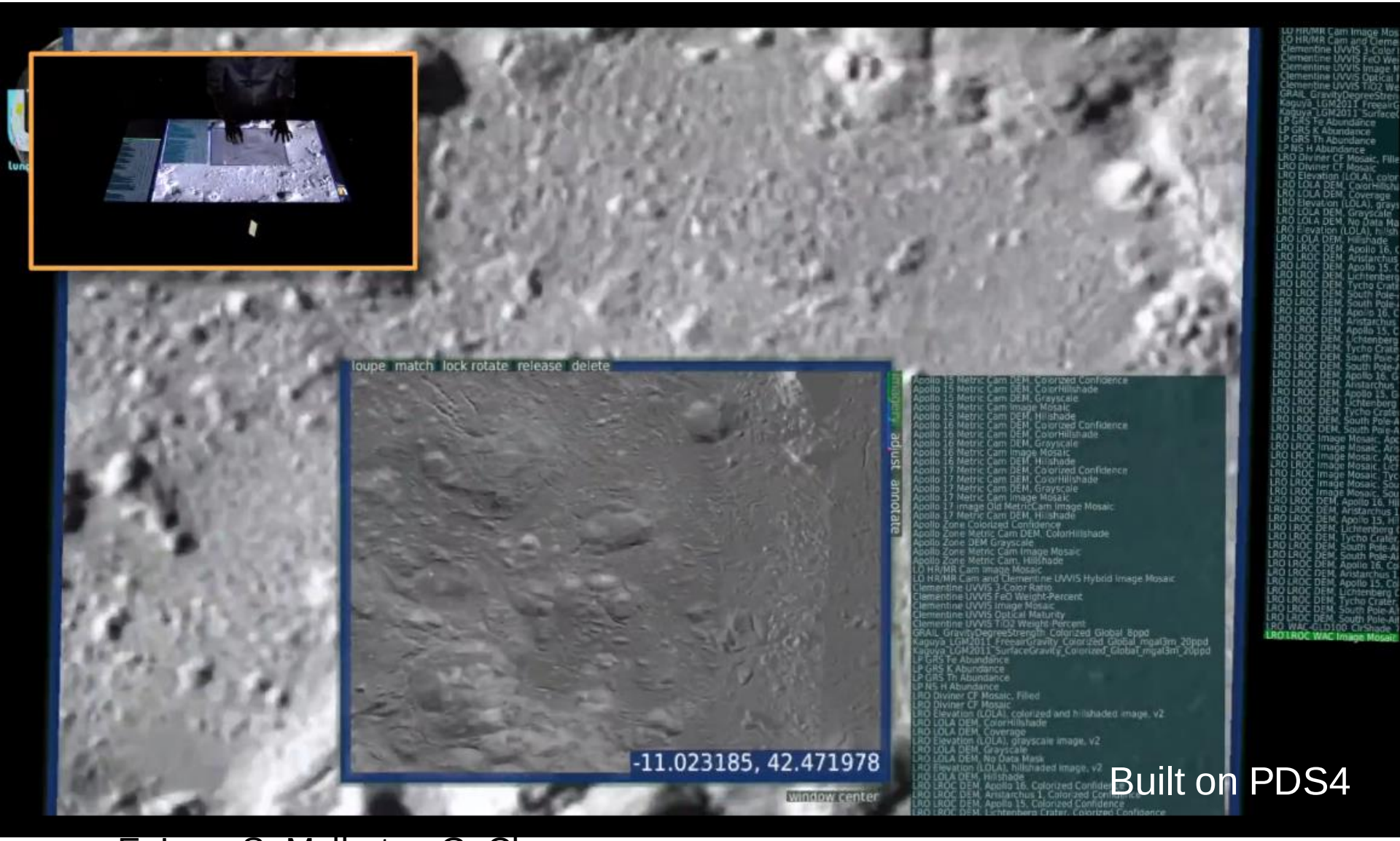


Integration of *multiple* Earth observing remote sensing instruments; comparison against models



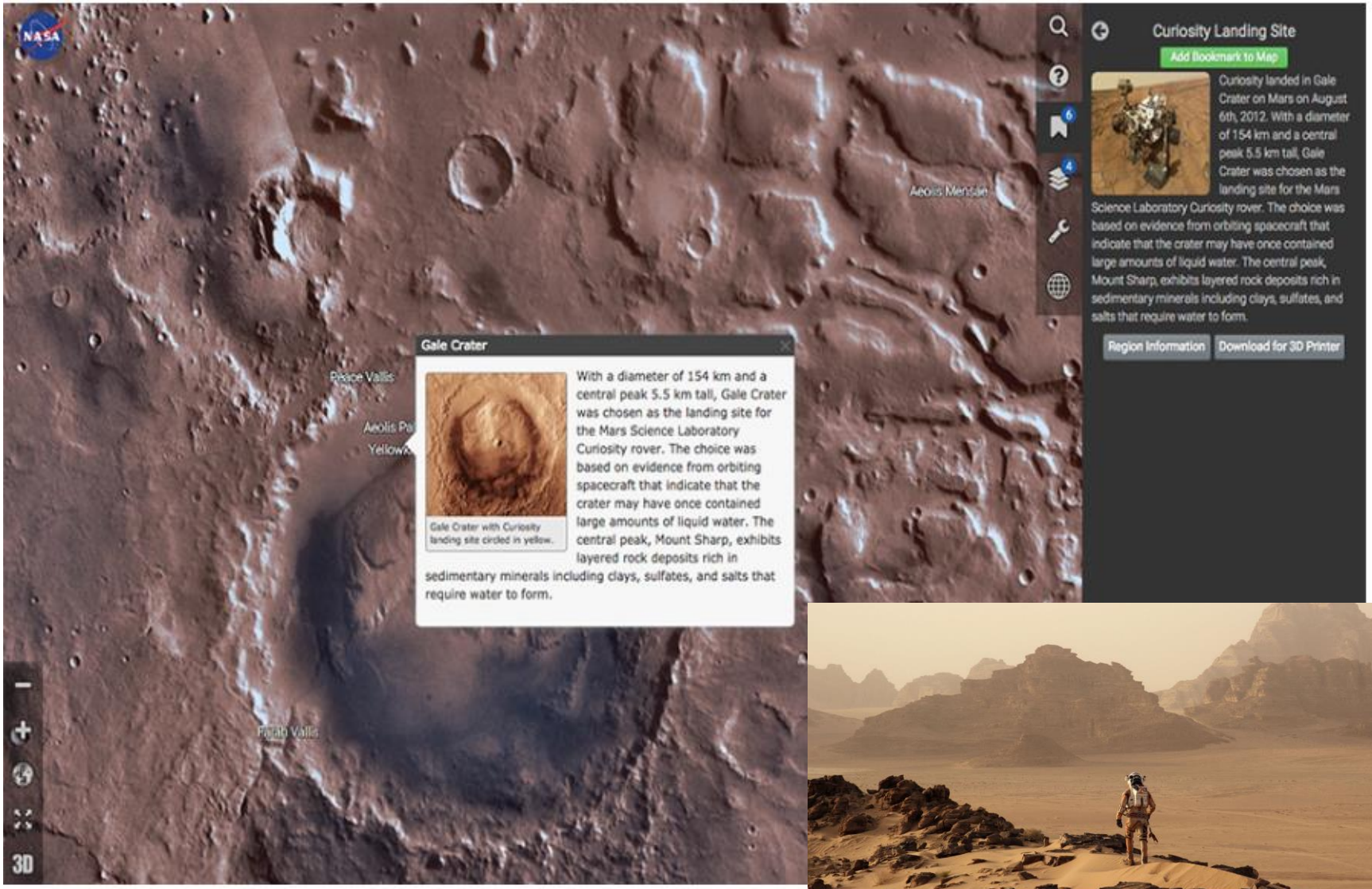
Real-time feature extraction and classification in astronomy

Exploring the Moon through Data Analytics



E. Law, S. Malhotra, G. Chang

Mars Trek: The Google Earth of Mars

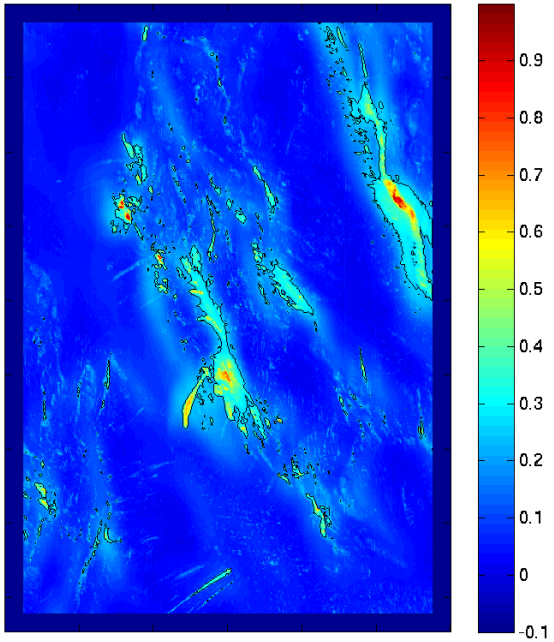


Credit: Emily Law, Shan Malhotra

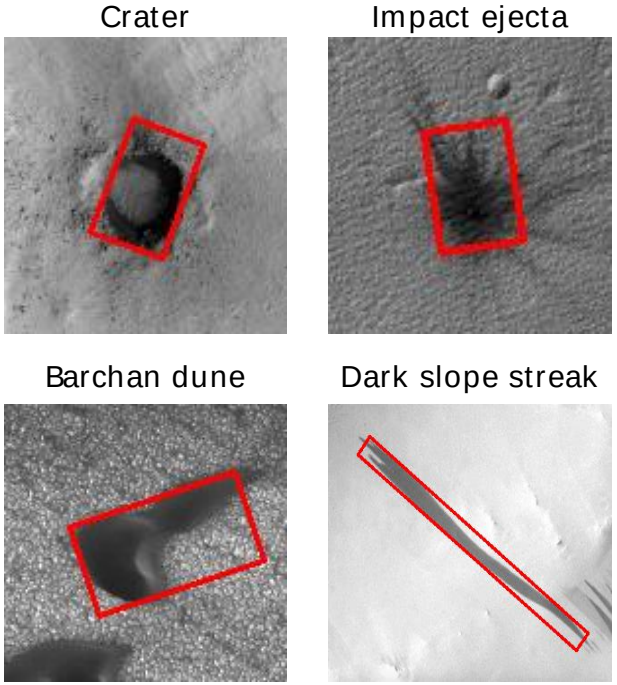
Content-Based Image Classification

- About ~1.3M images from MRO Mission HiRISE instrument
- Previously no way to easily find images with certain landmarks (e.g., craters)
- New Approach:
  1) Determine high salience (i.e., distinctive) regions by computing statistical differences between pixel and surrounding context
  2) Classify landmarks using *machine learning model* and user *training data*

MOC, June 2000



Salience Map



Examples of classified landmarks

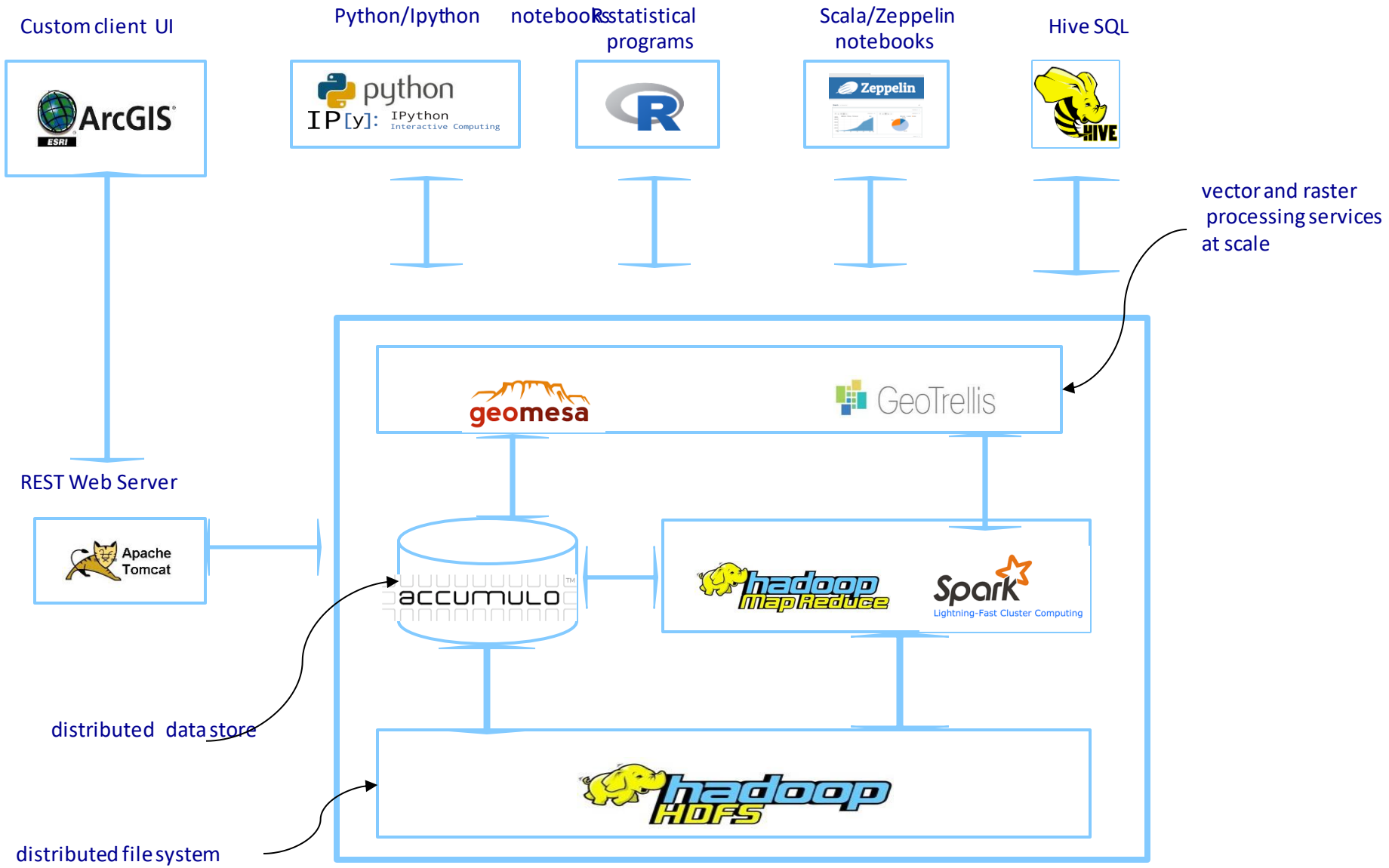
POC: K. Wagstaff

Ground-based Data Analysis

Understanding Underground Water in the California Central Valley

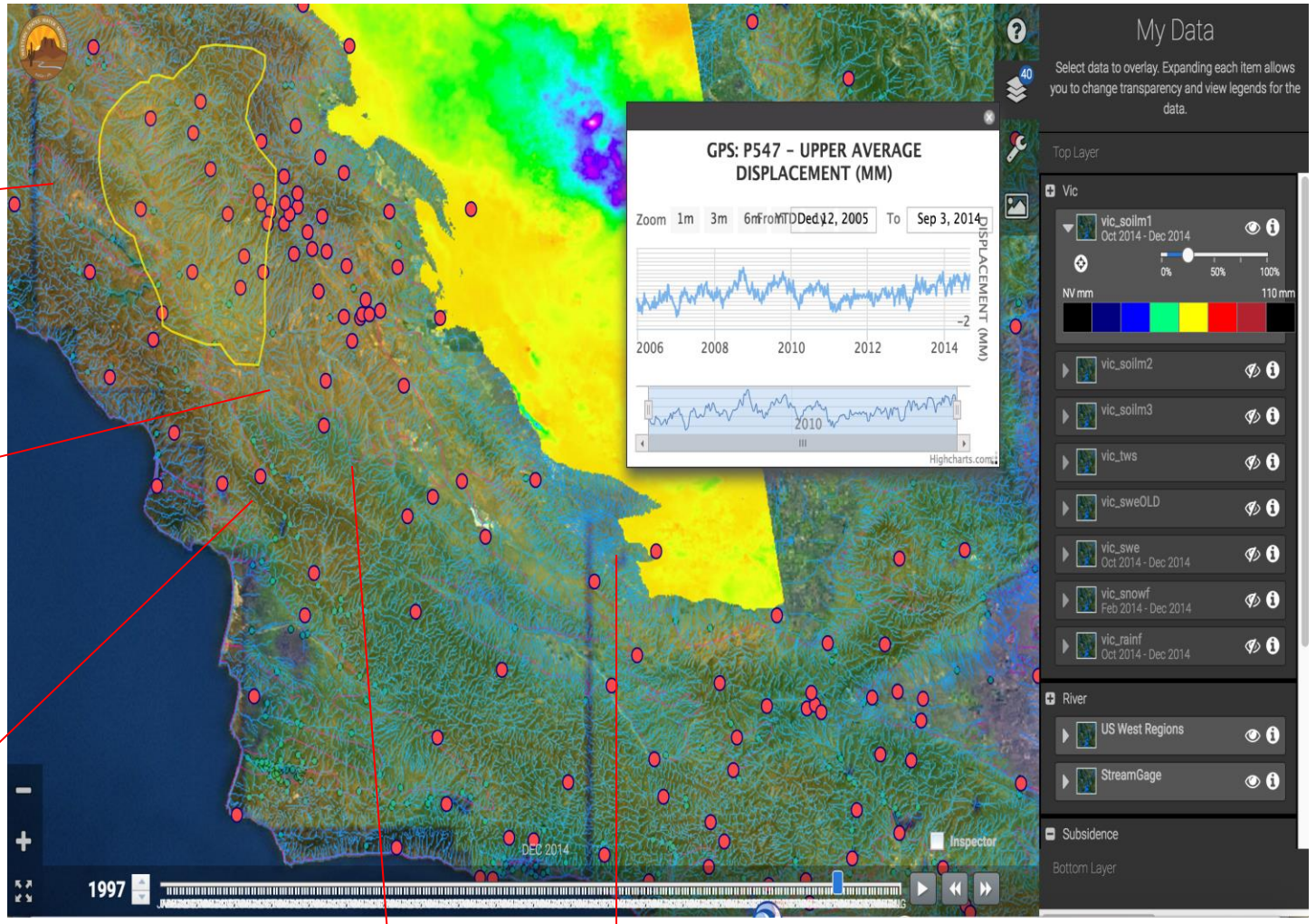
Credit: Tom Farr

Build Analytic Data Infrastructures



Open source and scalable to cloud; 180 billion data points accessible < 1 second

WaterTrek: Interactive Data Analytics for Hydrology



User Defined Polygon

GPS

In-Situ: Stream Gage Sensors

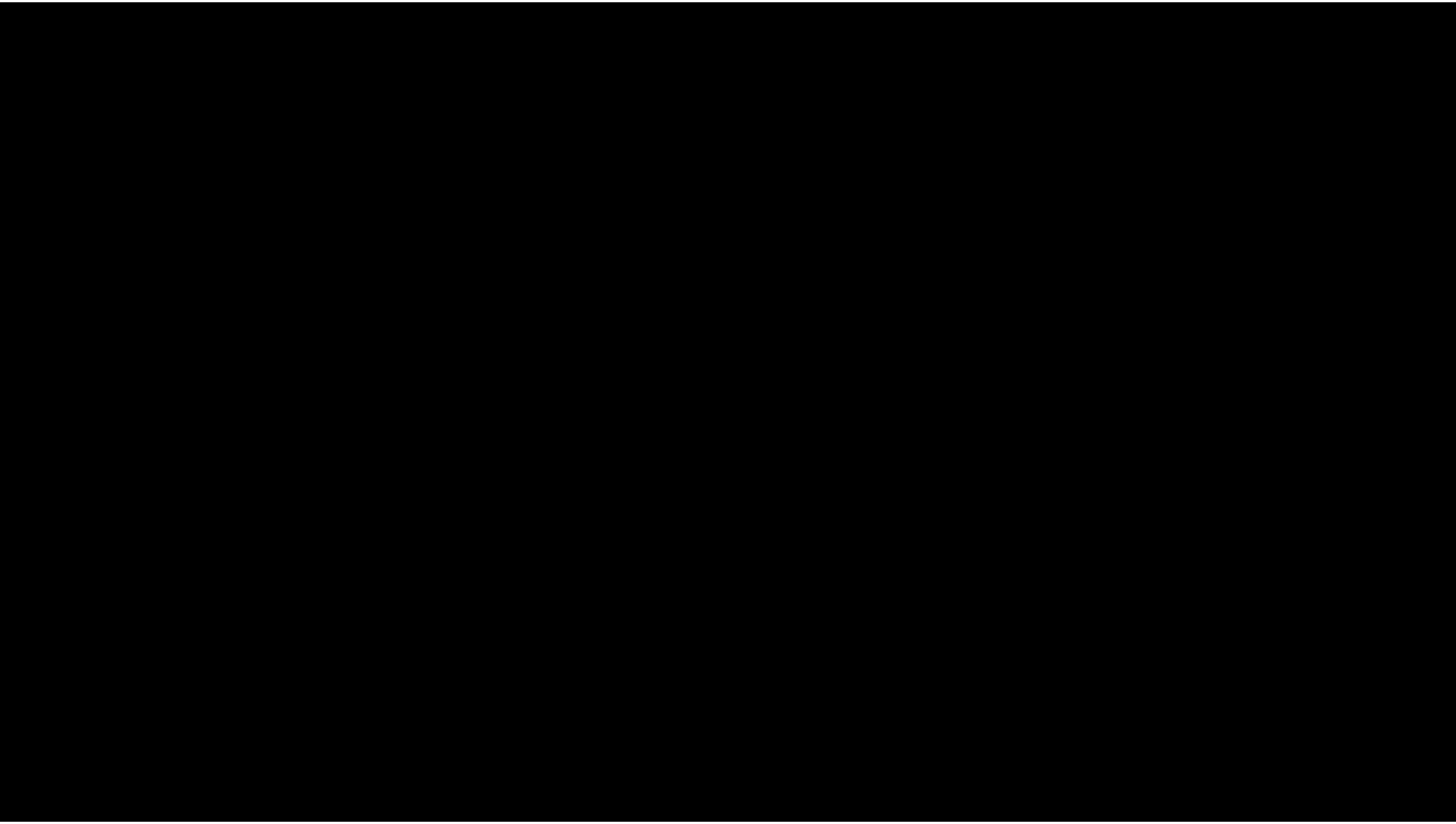
River Network

SAR derived Subsidence

Model Outpu
Soil Moisture

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

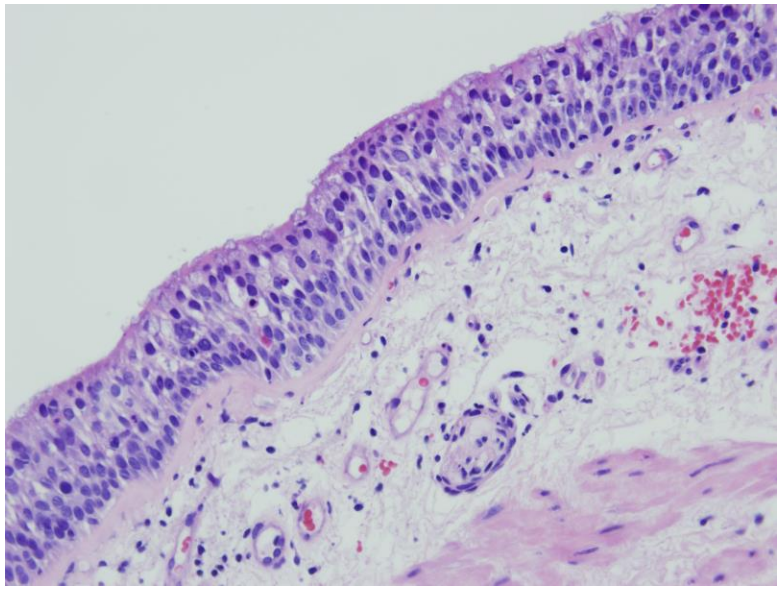
Virtual and Augmented Reality: Virtual Mars



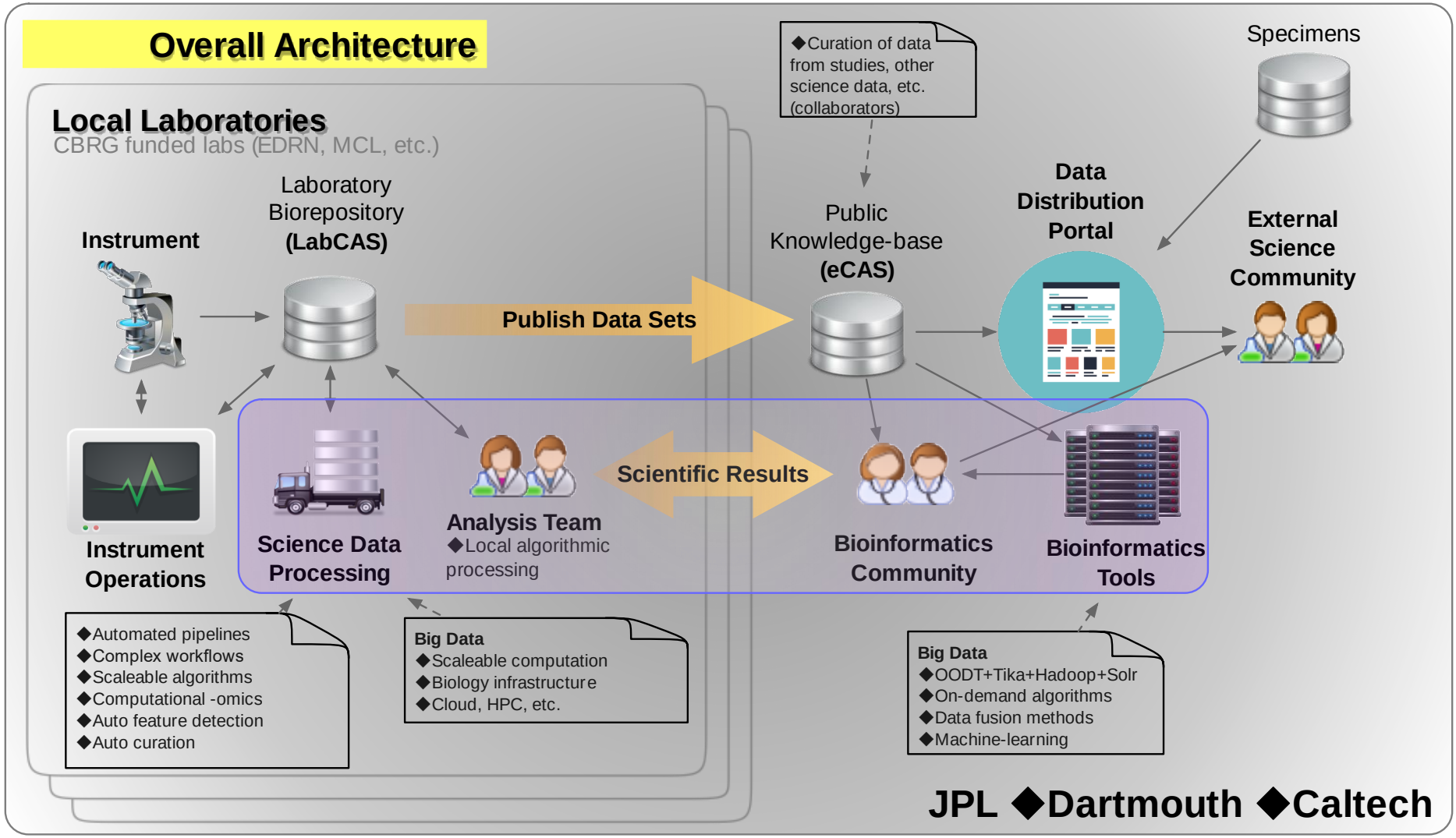
Methodology Transfer

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

- 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 planetary and Earth science
  - Reproducible, Big Data Systems for exploring the universe
  - Software and data science methodology transfer

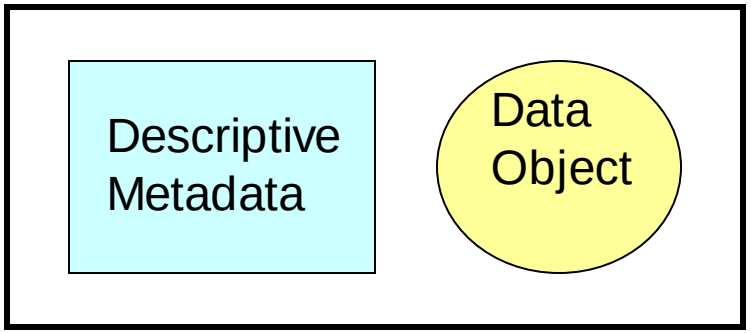


Data Science Architectures in Cancer Research

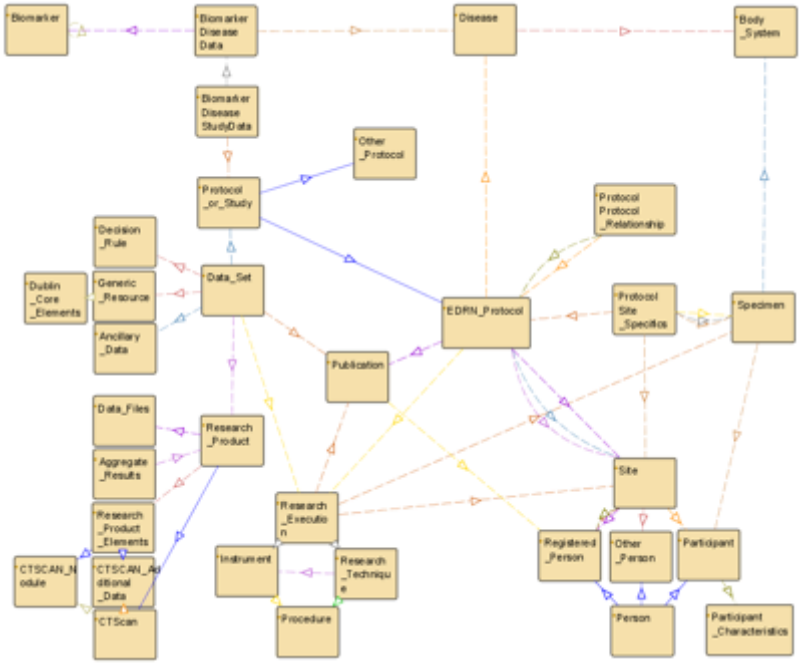


Common Data Elements and Information Models

- Provide standard data semantics to capture and share biomarker data
  - EDRN CDEs
  - MCL CDEs

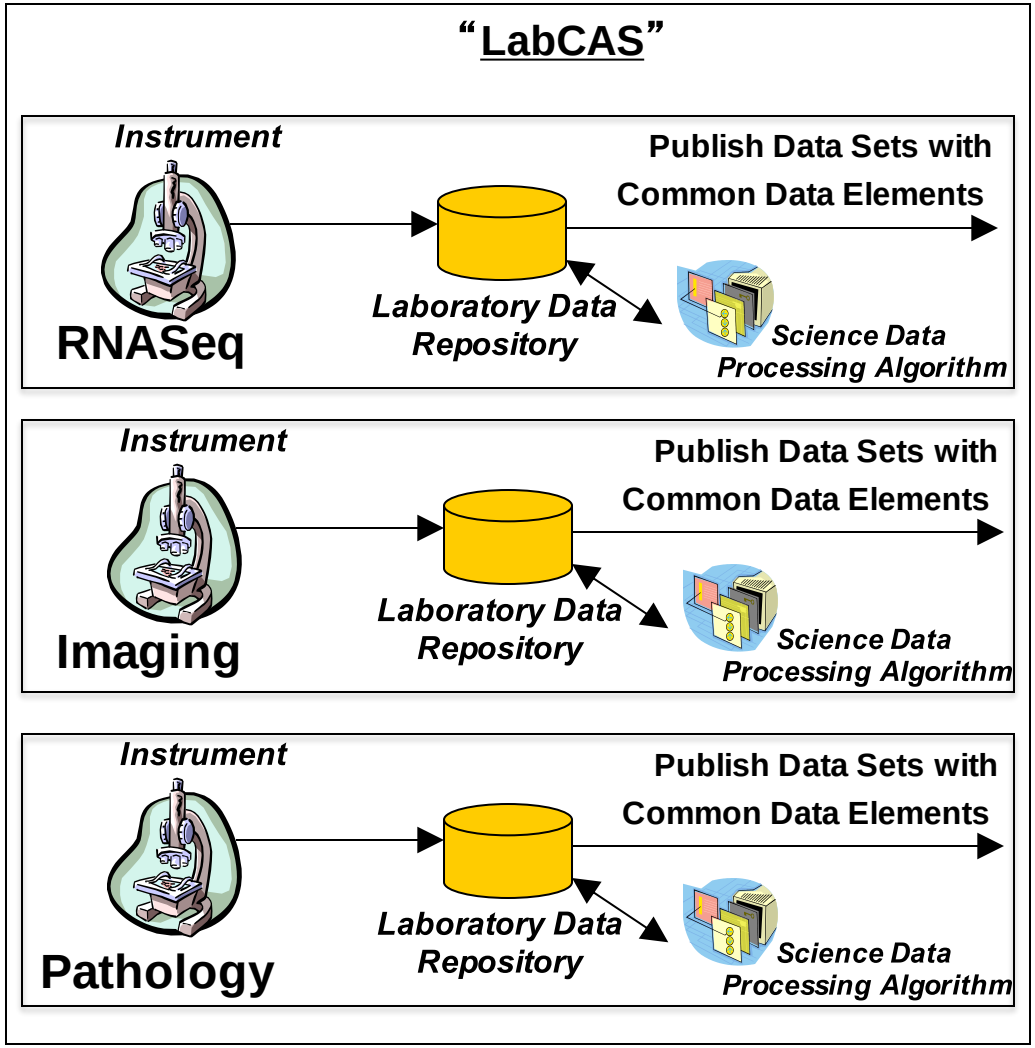


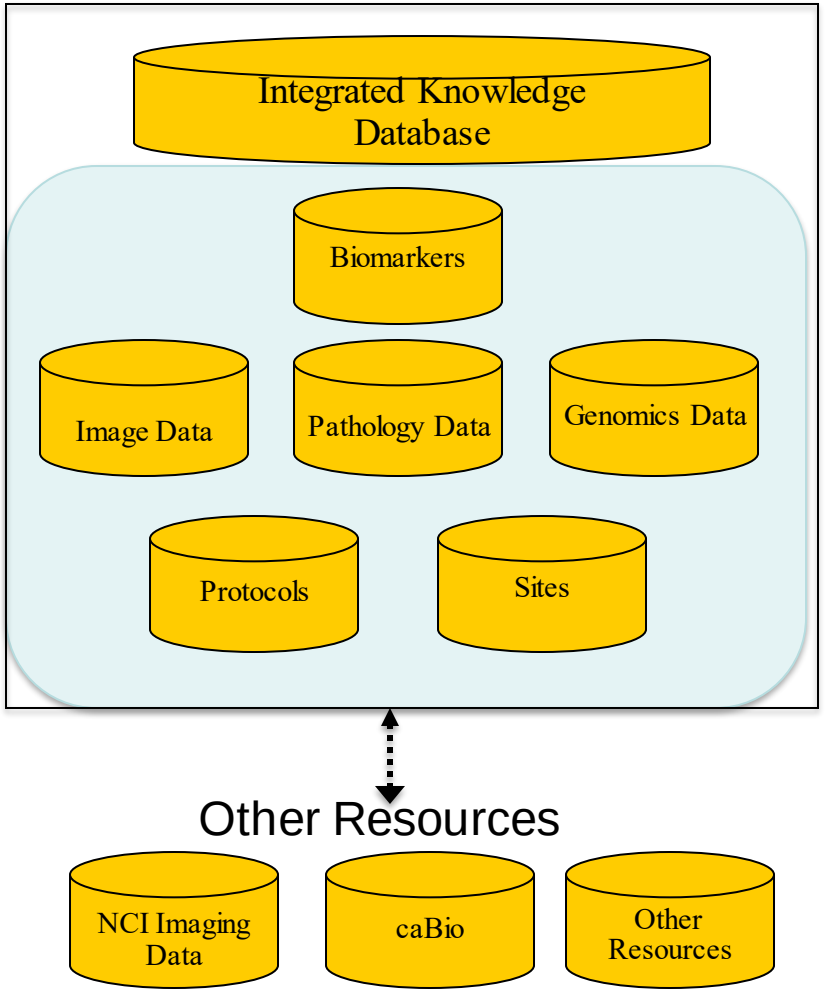
- Used as models to drive the knowledge system



Biomarker Ontology Information Model

Data Pipelines and Cancer Research

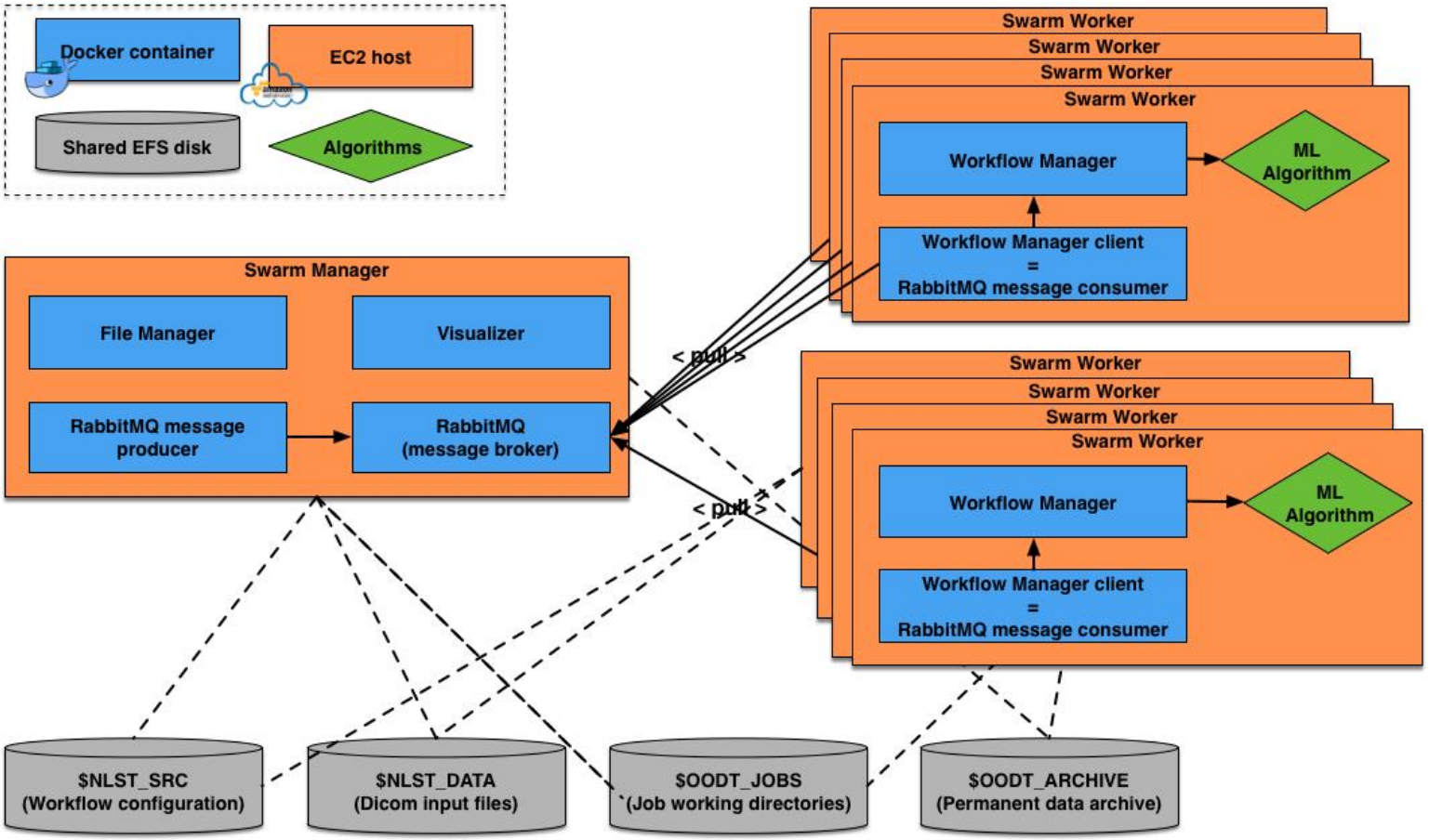




Scalable computing, common data elements, computational methods

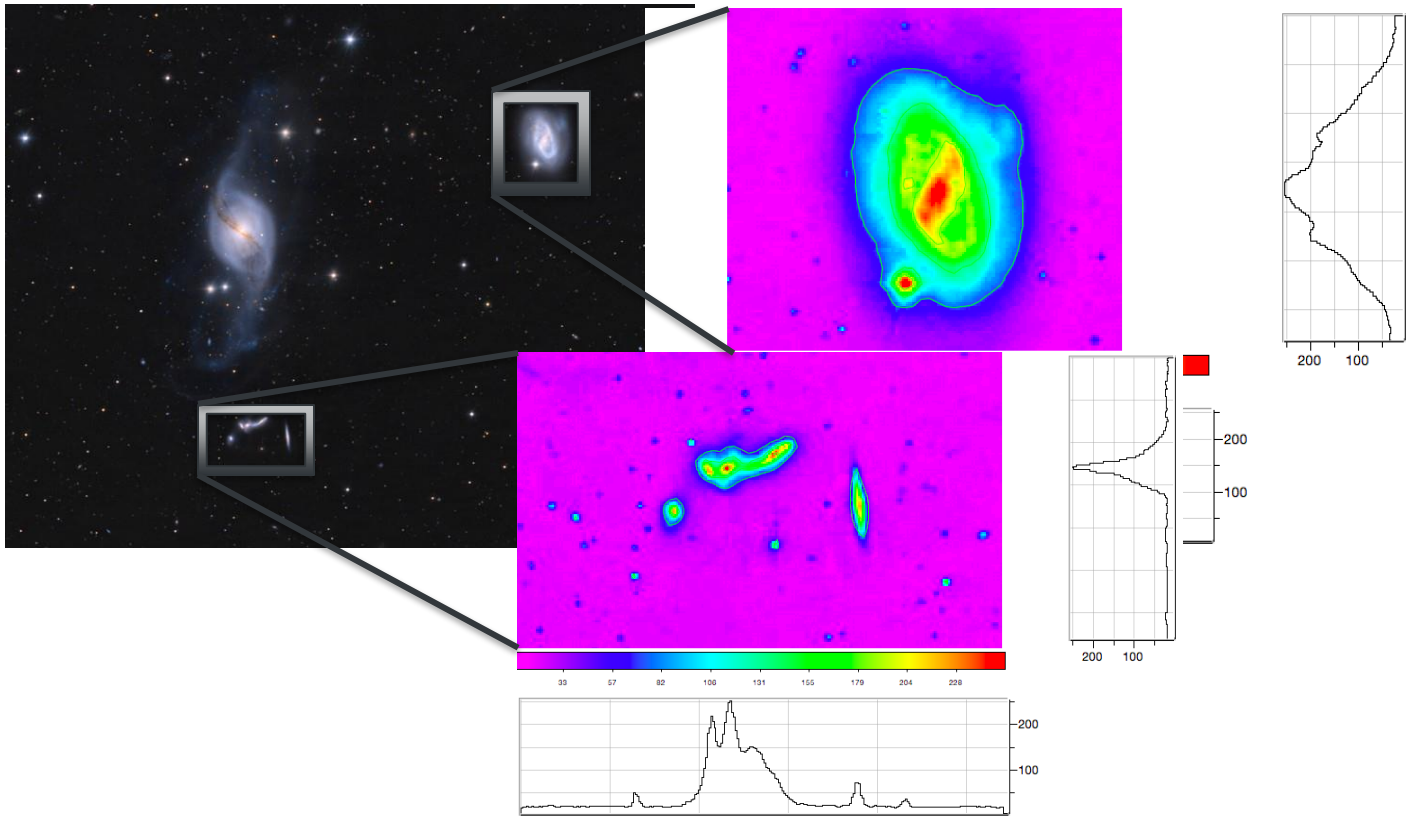
Scalable architecture for ML on AWS

- Uses Docker Swarm, Apache OODT workflows (from NASA/JPL), RabbitMQ messaging
- Can scale/auto-scale to any number of EC2 nodes



Methodology Transfer

From Astrophysics…



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.

Feature classification in images

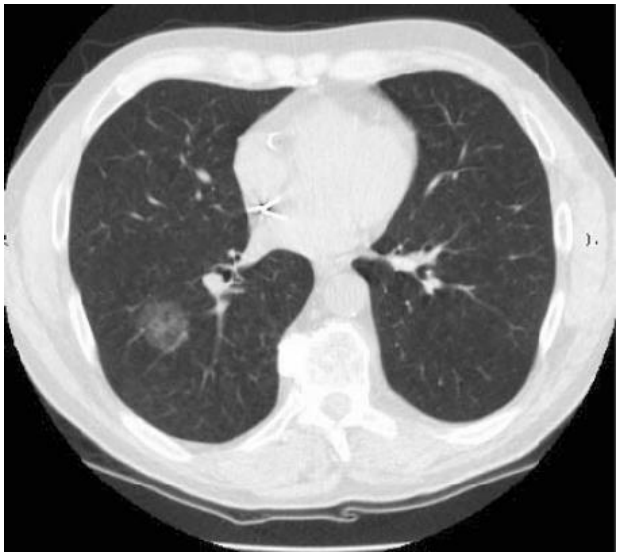
Pulmonary Nodule

Solid Nodule

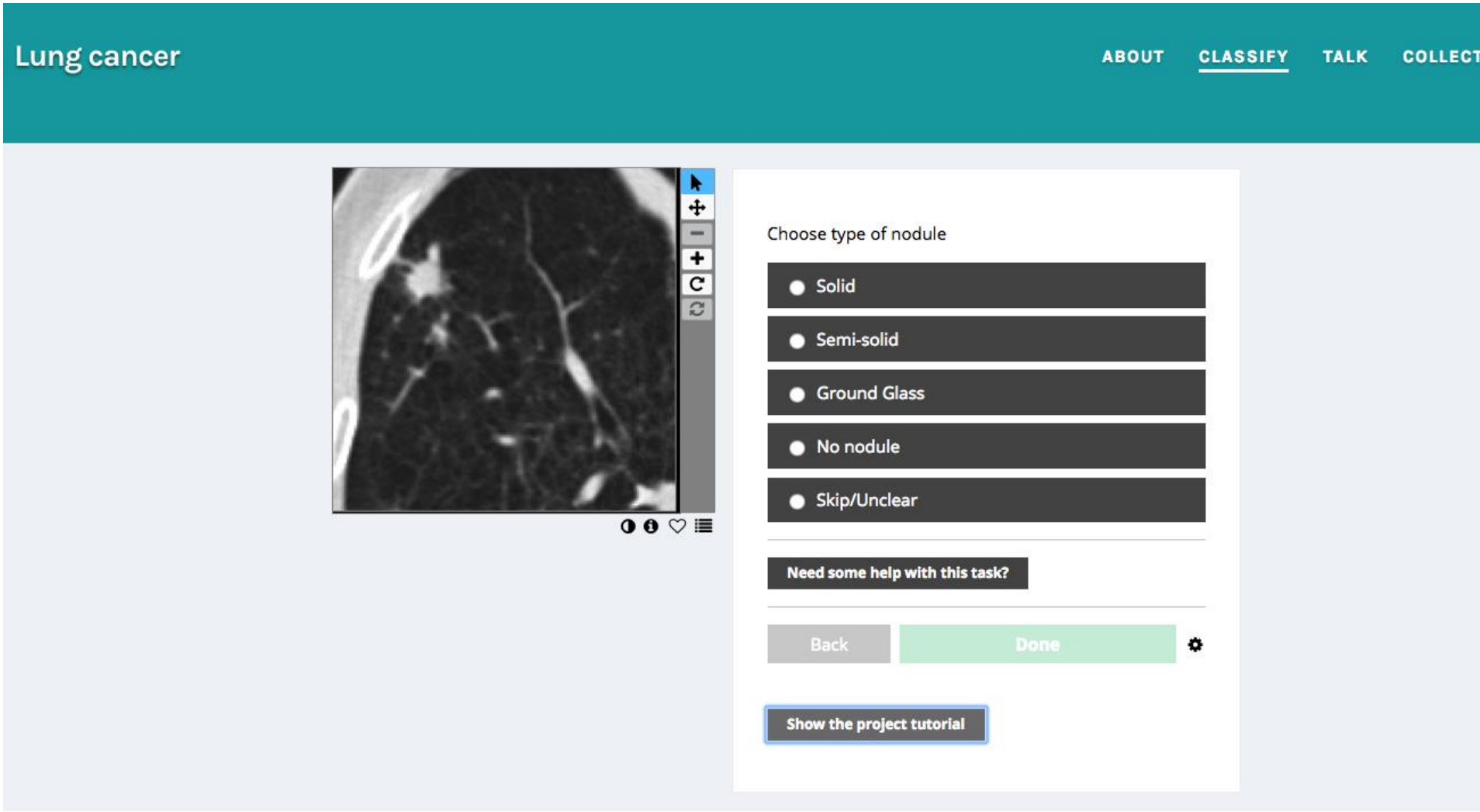
Sub-solid Nodule

Part-solid

Ground Glass

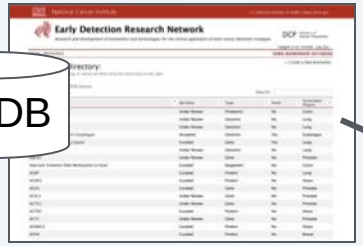


Crowd Sourcing Image Analysis



Integrated Knowledge Data Environment

Biomarker Database



BMDB

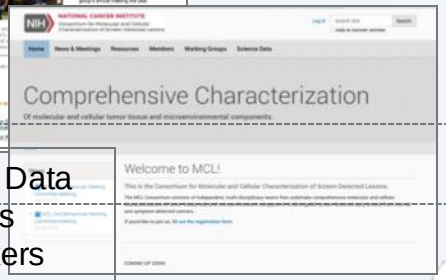
Informatics Center Knowledge Systems

KEGG

EDRN Portal

GEO

MCL Portal



Project Databases

PI Data

eSIS

KSDB

PubMed

PI WebSites

- Science Data
- Protocols
- Biomarkers
- Publications
- Signatures
- Sites
- SOPs
- Specimens
- Etc

BioMuta

Laboratory Catalog and Archive Service

LabCAS

eCAS

Specimens

Canary Foundation

Amazon S3

Conclusion

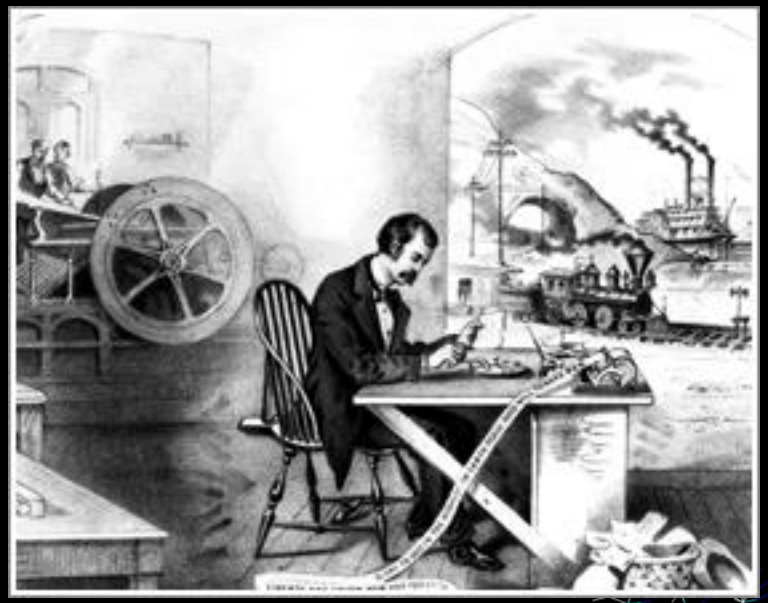
- Data Science is changing the paradigm for how we can do science
  - Science is increasingly data-rich and computationally enabled
  - Many new informatics fields are emerging
  - Good successes but more work to do!

- Much of the focus has been on building infrastructures for capturing data
  - Cloud infrastructures and other capabilities have paved the way

- However, great opportunity to consider how analytics is integrated end to end
  - From point of collection to analysis

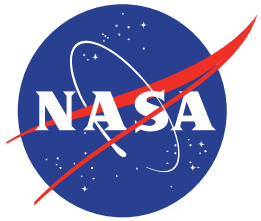
- New methods are needed to bring together and drive interactive analytics

These are Extraordinary Times



Changing the paradigm for how scientific data is captured, organized, and analyzed…

Djorgovski



JPL Caltech