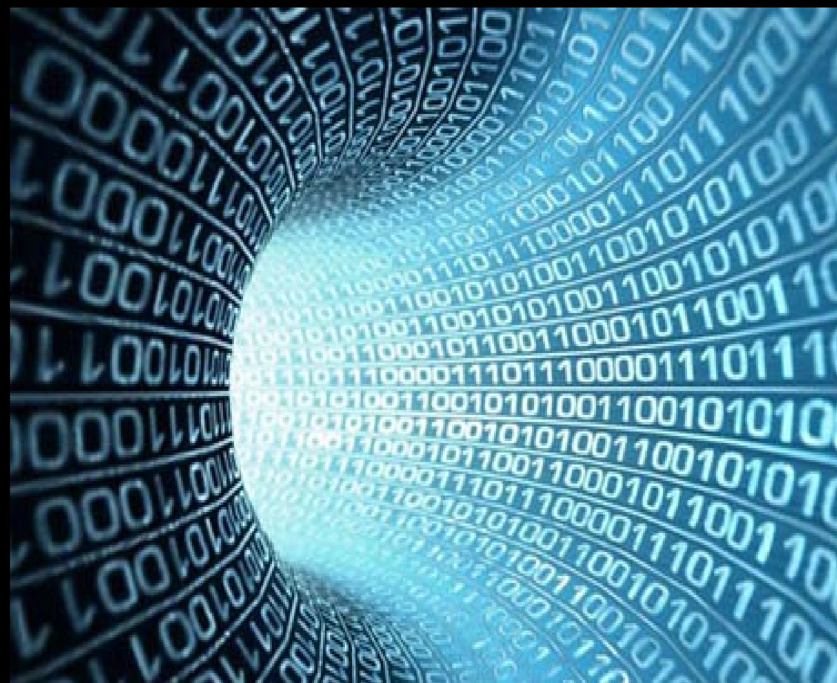
Big Data, Data Science and AI: Architectural Considerations



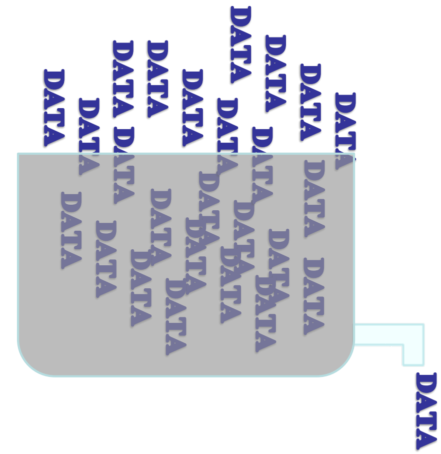
Daniel Crichton, Program Manager, Principal Investigator, Principal Computer Scientist
Leader, Center for Data Science and Technology

NASA Jet Propulsion Laboratory, California Institute of Technology

March 2020

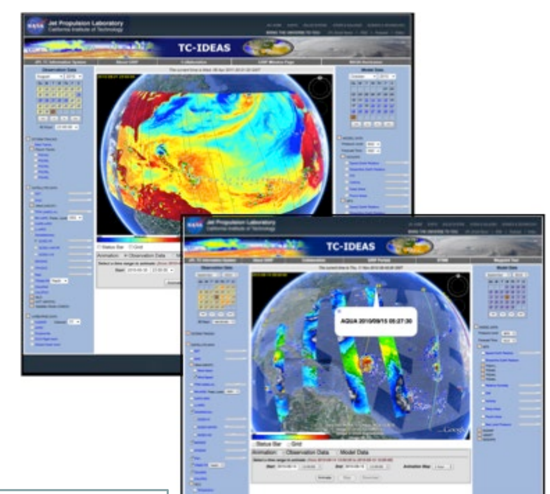
Terms: *Big Data and Data Science*

<u>Big Data</u>



When needs for data collection, processing, management and analysis go beyond the capacity and capability of available methods and software systems

<u>Data Science</u>



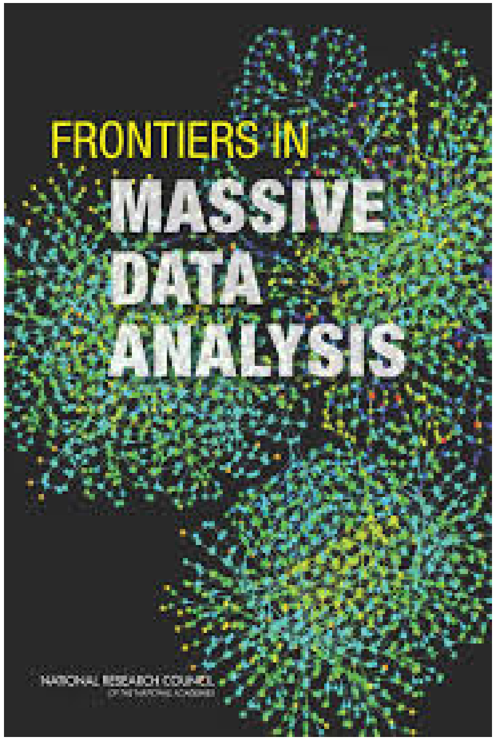
Scalable architectural approaches, techniques, software and algorithms which alter the paradigm by which data is collected, managed and analyzed

The opportunities to use data are immense!

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)
- Co-author: Dan Crichton, JPL
- Consideration of the 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 discipline experts
- Application of novel statistical and machine learning approaches for data discovery



Published in 2013

- A Major Shift from Compute-Intensive to Data-Intensive -

Using Data as a Strategic Asset to Transform “How We Work”

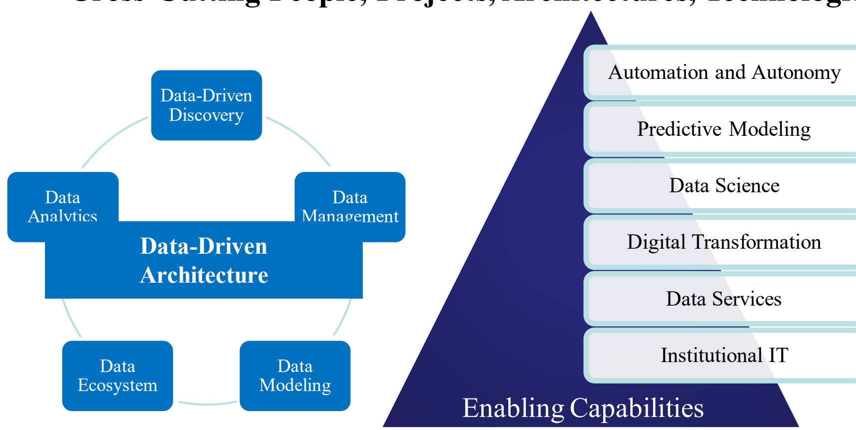
Vision

Establish a data-centric culture and competency for JPL using data and analytics to innovate and create the Lab of the future, transforming “what we do” and “how we do it.”

Institutional, Business Data Lifecycles

- Data-driven business and project decisions
- Analytics and decision support leveraging all operational data
- Best practice management of JPL institutional data repositories

Cross-Cutting People, Projects, Architectures, Technologies



Data is FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)

Mission, Science Data Lifecycles

- Data-Driven Discovery from Archives
- Intelligent Ground Systems and Automated Mission Operations
- Integrated Model- and Data-Driven Methodologies

Drive Decisions * Create Knowledge * Increase Efficiency * Connect Community

From Data to Models to Enable Automation and Autonomy: An Enterprise View

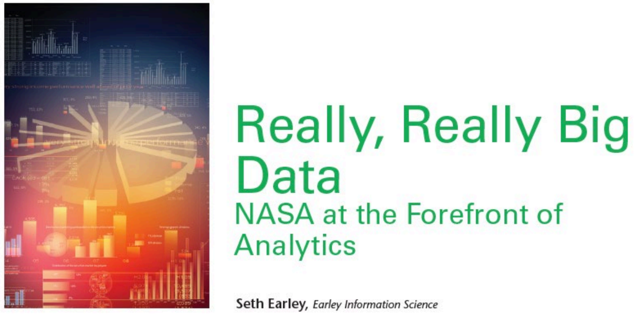
Automation and Autonomy – Use robust models and data-driven methods to enable autonomous decisions and automated operations in all types of environments

Predictive Modeling – Embrace modeling across JPL for science, missions, engineering, and institutional activities

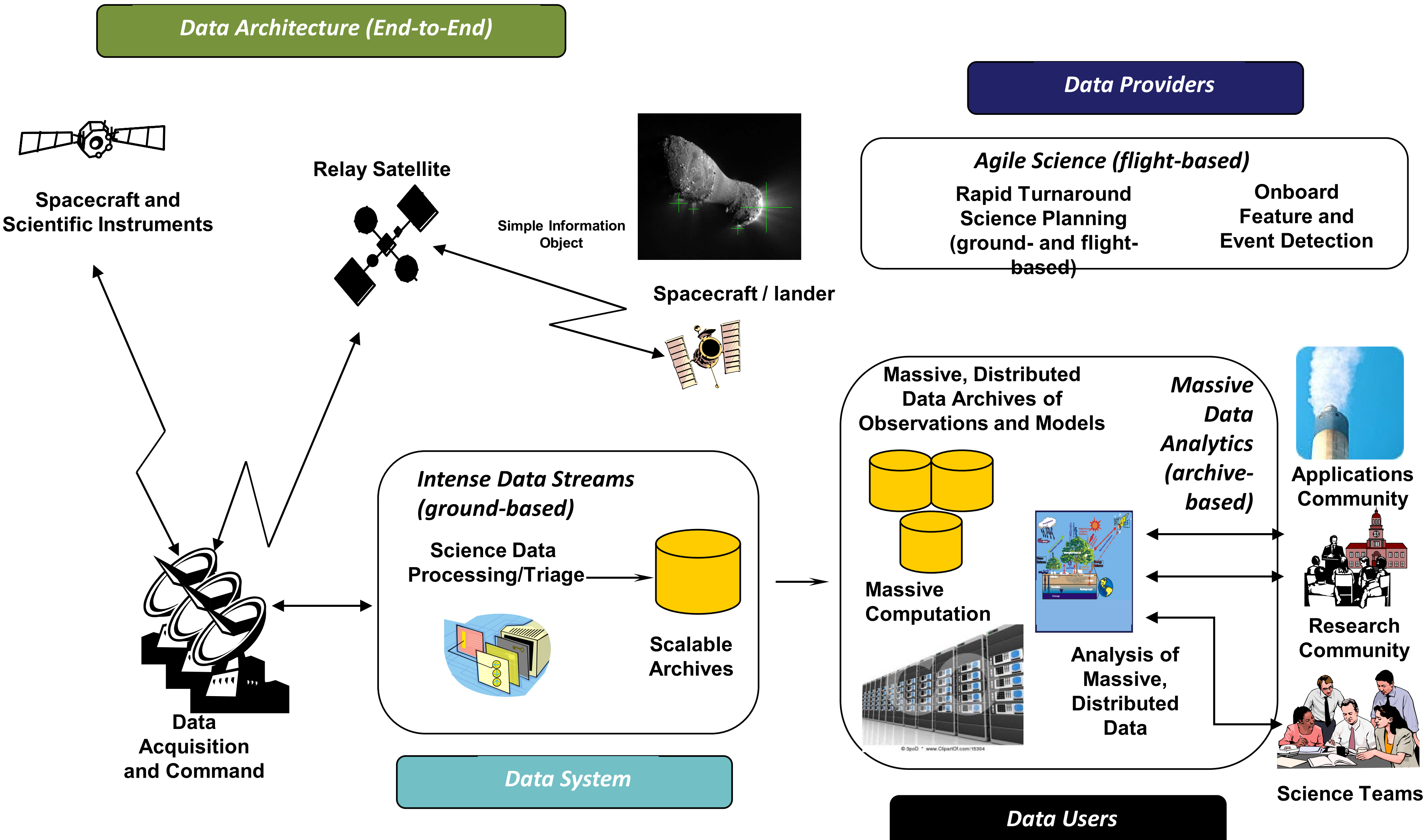
Data Science – Embrace AI, ML and data analytics for JPL science, missions and other areas

Digital Transformation – Capture the Laboratory's Digital Data Assets

Institutional IT – Provide a foundational set of services to support scalability in storage, computation, and networking



Data Lifecycle Model for Space Missions



Data Architecture

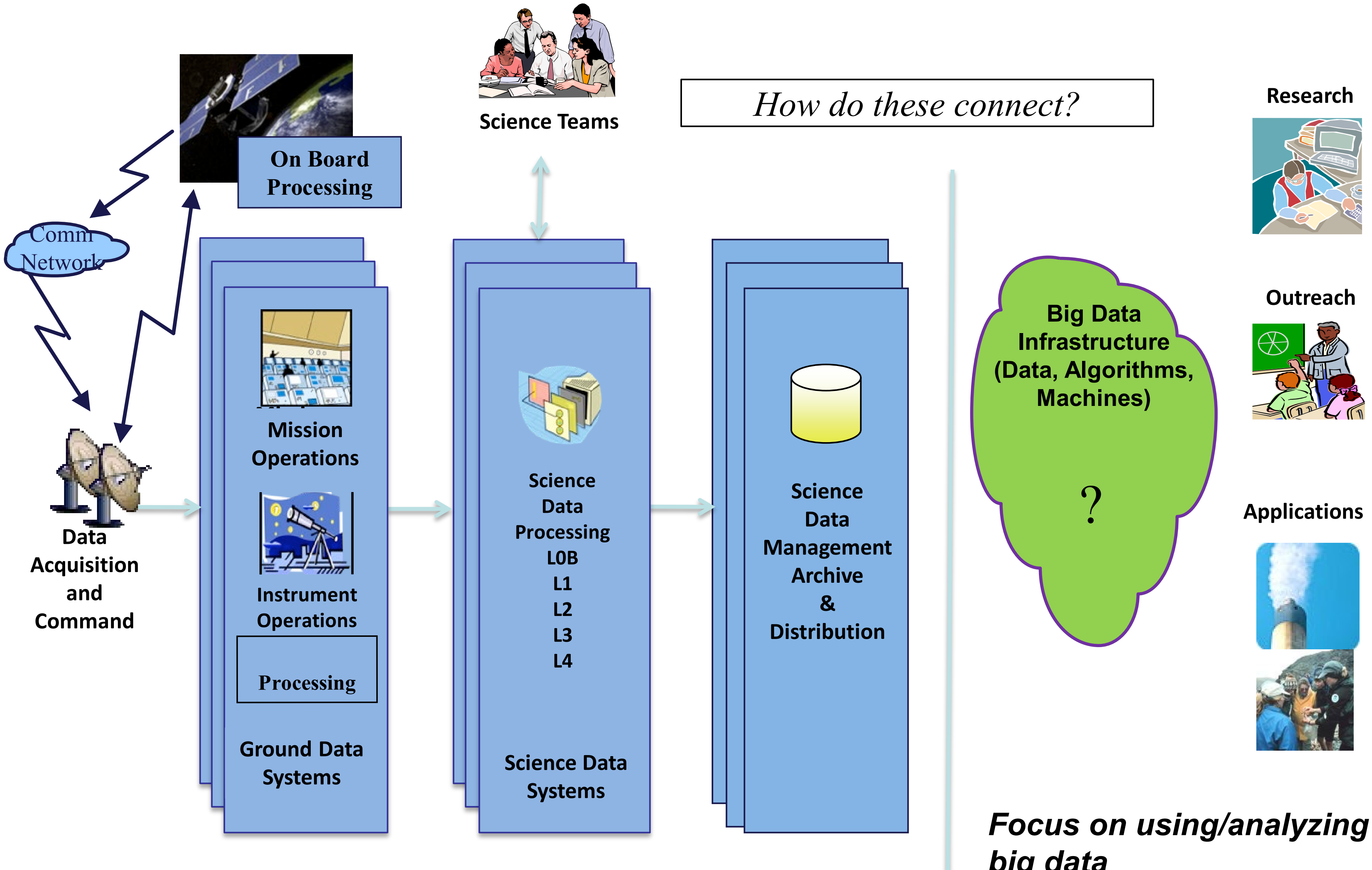
Address Big Data Challenges Across the Full Data Lifecycle

Data Architecture

Perform original processing at the sensor / instrument	Data Generation
Make choices at the collection point about which data to keep	Data Triage
Improve resource efficiencies to enable moving the most data	Data Transport
Anticipate the need to work across multiple data sources	Data Curation
Increase computing availability at the data to generate products	Data Processing
Increase the scale and integration of distributed archives	Data Archiving
Apply visualization techniques to enable data understanding	Data Visualization
Apply machine learning and statistics to enable data understanding	Data Mining
Create analytics services effective across massive, distributed data	Data Analytics

Develop principled techniques and scalable architectures to address challenges across the entire Data Lifecycle

Unifying Steward and Analytics



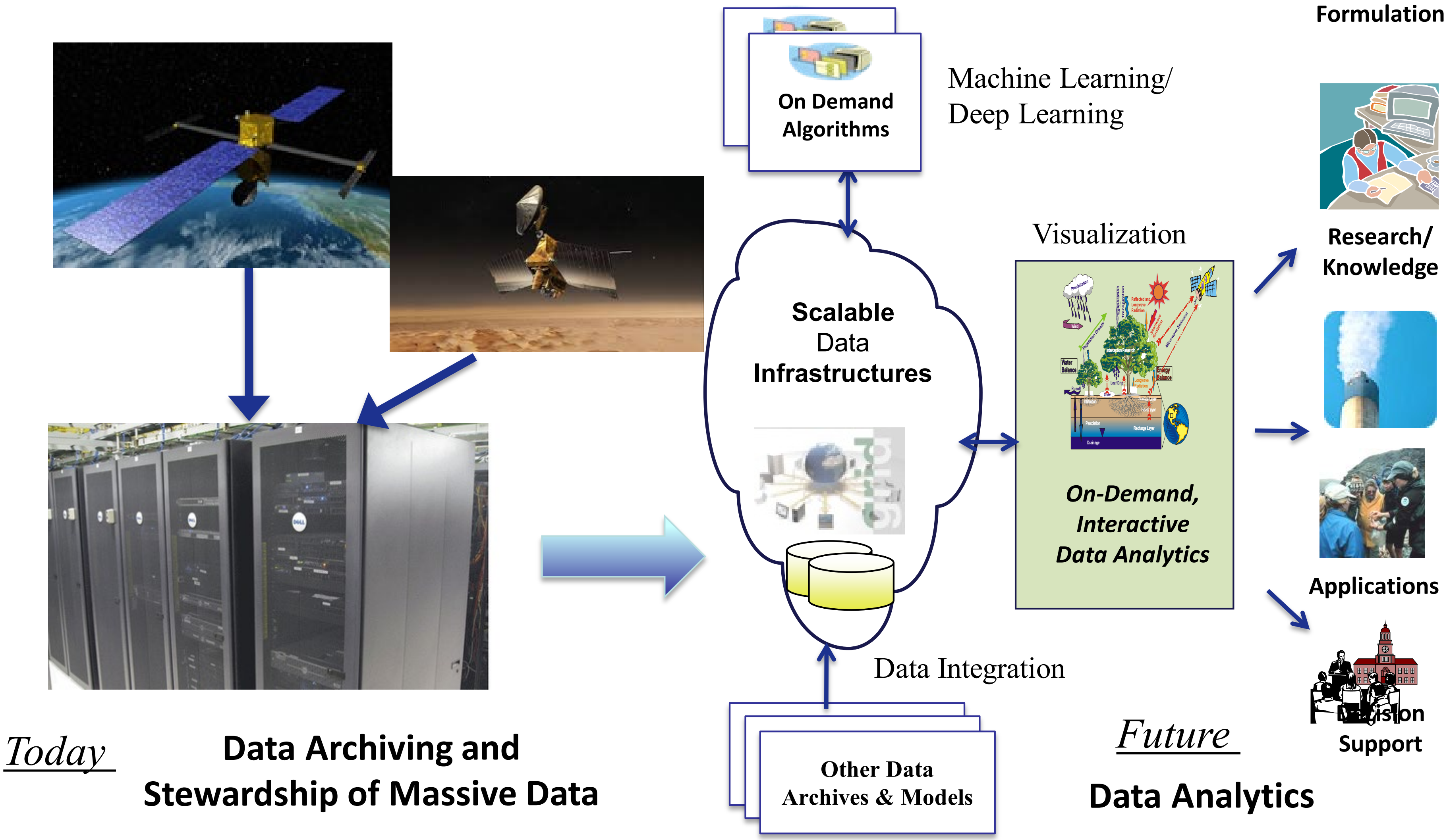
Focus on generating, capturing, managing big data

Systemizing the Analysis: Integrating Data Archiving and Analytics/AI/ML

- **Scalable Data Management**
  - Define the data lifecycle for different domains in science, engineering, business
  - Capture well-architected and curated curated data repositories based on well-defined data/information architectures
  - Architect automated pipelines for data generation and capture

- **Scalable Data Analytics**
  - Create analytics ready data sources; new data results
  - Develop computational capabilities at the data sources
  - Develop analytical methods
    - Novel statistical approaches for data integration and fusion
    - Machine Learning/AI for data extraction, prioritization, reduction, pattern recognition, etc

Enabling Data-Driven Analysis

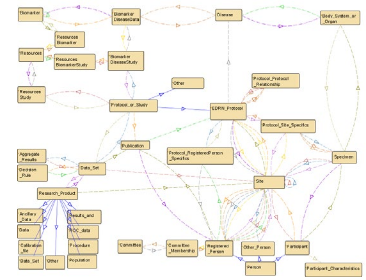


Separate analysis ready data from archive formatted data for data-driven approaches

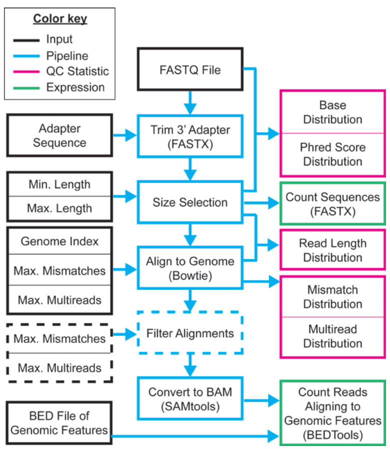
Enabling Technical Capabilities Exist Today

Cloud, Open Source, and Big Data Infrastructures

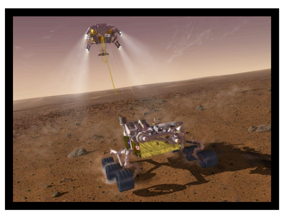
Machine Learning and Deep Learning



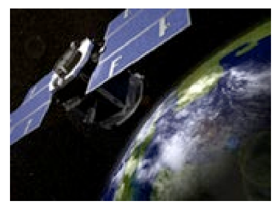
Ontologies and Information Models



Computational Pipelines/HPC



Planetary



Earth Science

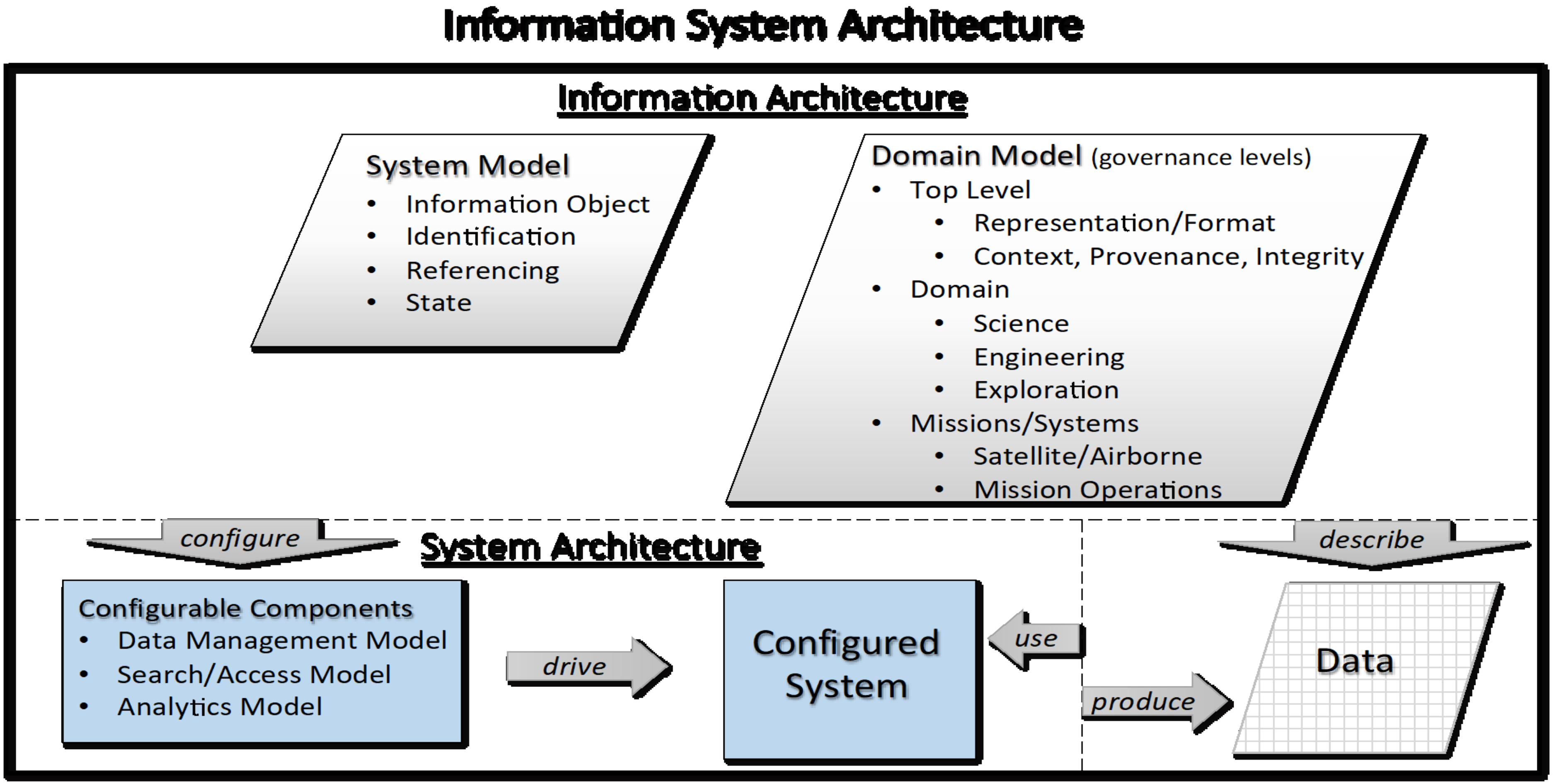
Astronomy

Great Opportunities for Methodology Transfer and Collaboration

Visualization and HCI Techniques

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Models of Data: An information model-driven approach

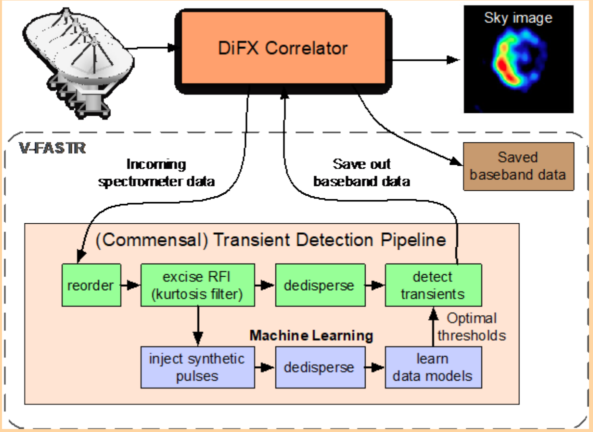


Crichton, D. Hughes, J.S. ; Hardman, S. ; Law, E. ; Beebe, R. ; Morgan, T.; Grayzeck, E.
A Scalable Planetary Science Information Architecture for Big Science Data.
IEEE 10th International Conference on e-Science, October 2014.

Triage, Analysis, and Understanding of Massive Data using Machine Learning

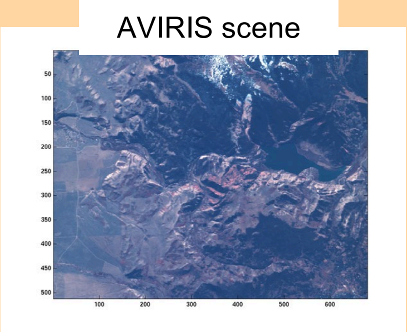
- Detection: fast identification of signals of interest (triage)

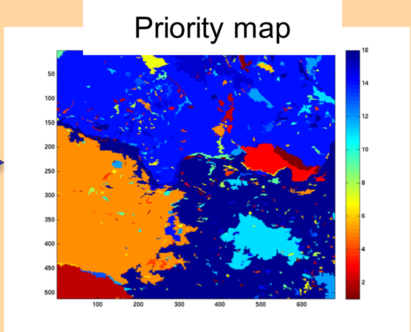
Radio astronomy:
V-FASTR realtime system at the VLBA



- Prioritization: use triage decisions to inform adaptive data compression

Earth science:
Onboard content-sensitive data compression

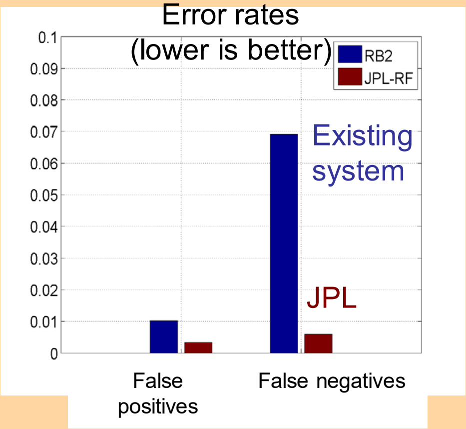




- Classification: online, real-time source type classification

Optical astronomy:
Reducing false positives for the Palomar Transient Factory

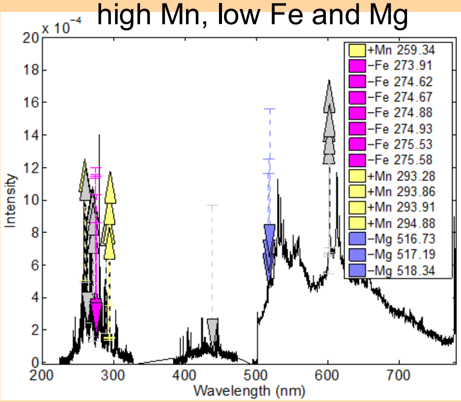
Real or spurious?



- Understanding: generate human-understandable explanations for decisions

Planetary science:
Anomaly detection in ChemCam emission spectra from Mars, with content-sensitive "explanations" indicated with arrows (higher than expected vs. lower than expected)

Rhodochrosite explanations:
high Mn, low Fe and Mg

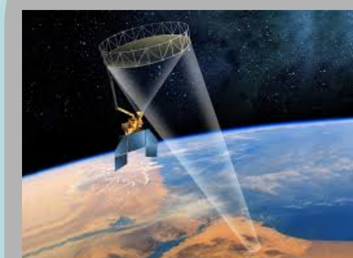


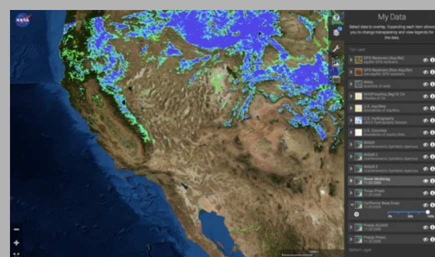
Credits: Kiri Wagstaff, Umaa Rebbapragada, David Thompson, Benyang Tang, Hua Xie

Driving Data Science into JPL’s Fabric

- ~50 pilots launched 2017-2020
  - Spanning science, mission and DSN operations, and formulation
  - Building towards a data science vision of full utilization of data and agile application of analytics

Use Cases: Science





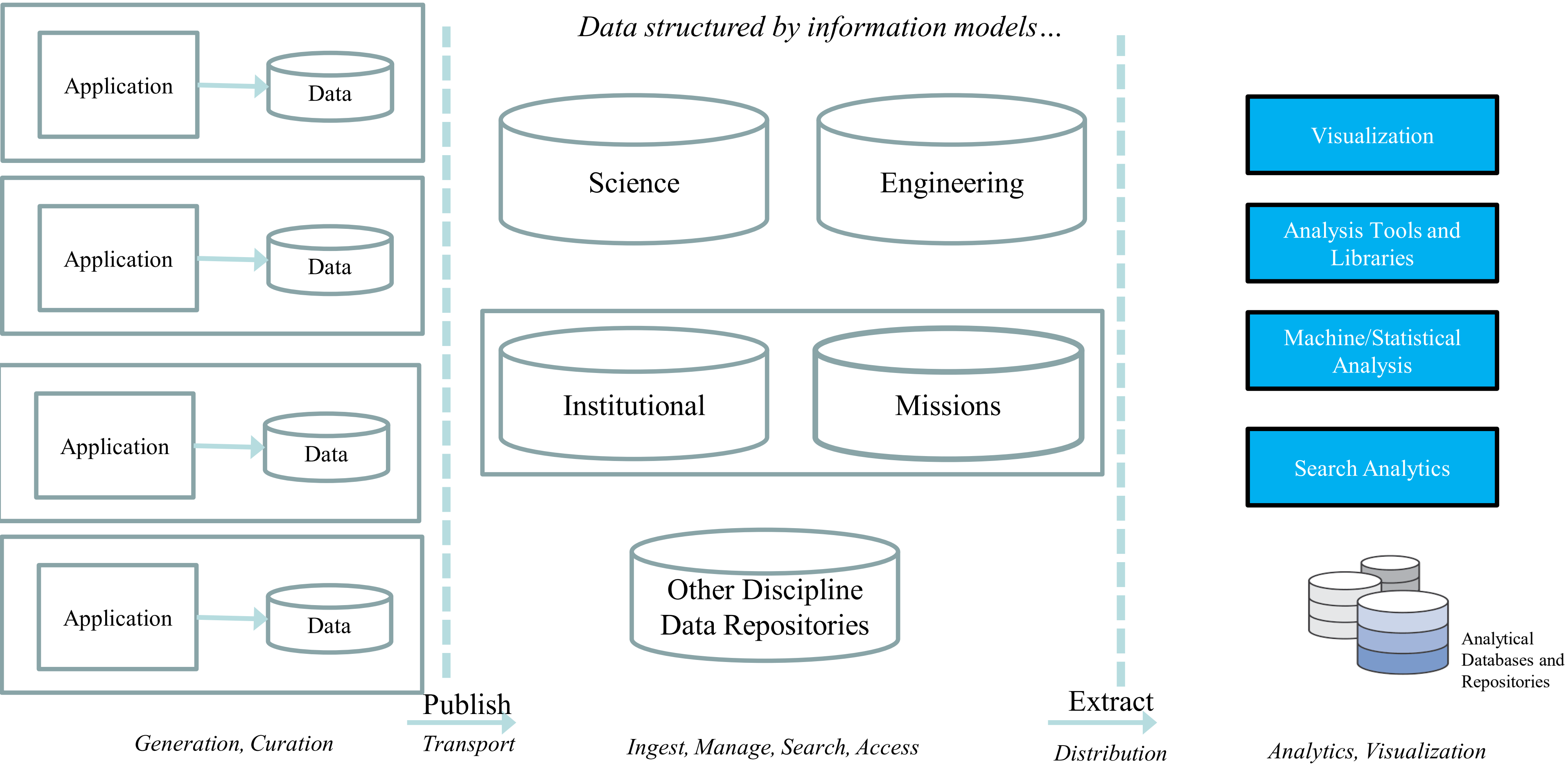
Use Cases: Mission Ops

Use Cases: Formulation

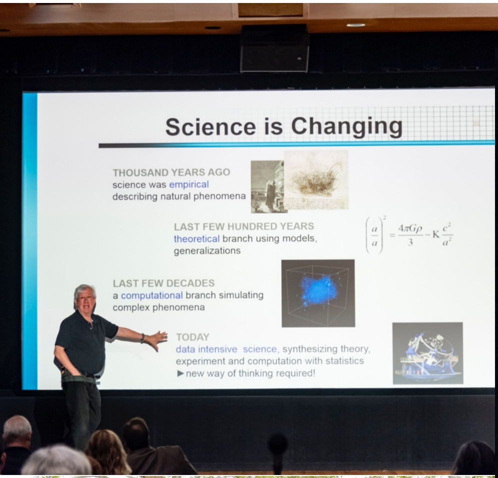
Use Cases: Institution

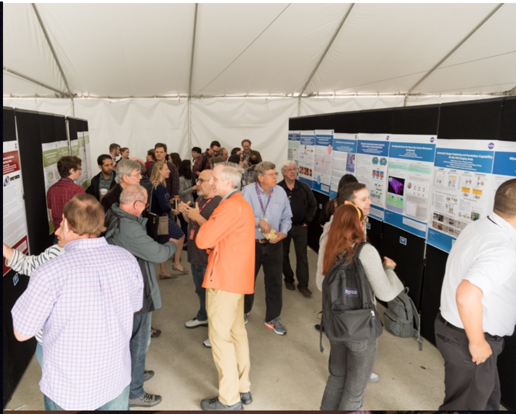
JPL's Emerging Enterprise Data and Analytics Strategy

From Applications to Data-Driven Discovery and Analytics



Capacity building across JPL: Driving a Lab-wide Data Strategy





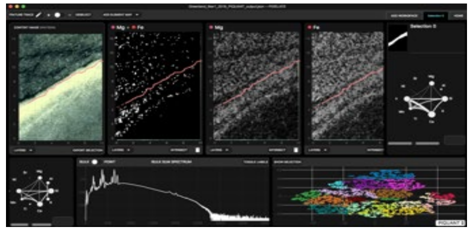
This is our future!

Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov

Data Science Pilots: Direct Infusion now into the Fabric of JPL

1) Astrobiology (Mars 2020) S. Davidoff



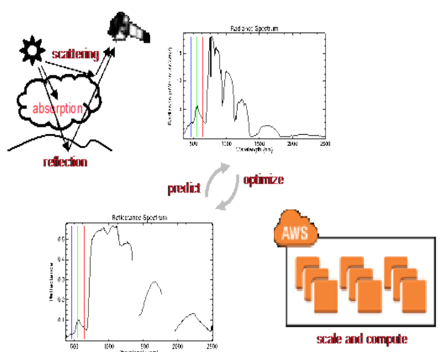
- **Machine Learning**: Increased performance for identifying geochemical similarity in images by 10,000% (days to seconds)
- **Missions**: Mars 2020 for PIXL Science Ops in FY2020

2) Autonomous Spectral Mapping Instrumentation (SOFIA) J. Pineda

- **Machine Learning:** techniques to identify data anomalies in spectral line mapping instruments in real time.
- **Missions:** SOFIA

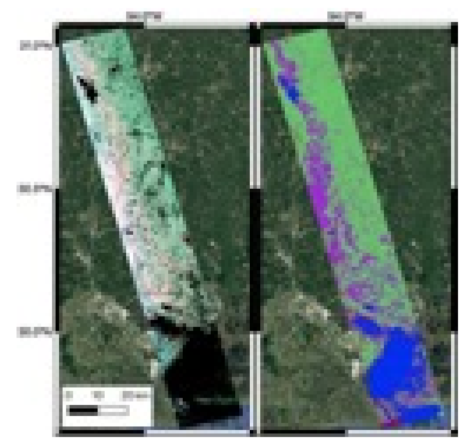
3) Mission-Ready Prototype Level 2 for SBG D. Thompson

- **Machine Learning:** multiple orders of magnitude improvement in analyzing atmospheric radiative transfer models (RTMs)
- **Missions**: EMIT, SBG, and Geology Decadal Observable



4) Automatic Per-Pixel Classification of UAVSAR Imagery M. Denbina

- **Machine Learning**: Increased automated flood detection accuracy from 76% to 87%
- **Missions**: JPL UAVSAR processing group for faster disaster response.



Direct quantitative advances and infusion of capabilities into NASA Projects