Richard Doyle, *Information and Data Science Program Office*
Daniel Crichton, *Center for Data Science and Technology*

NASA Jet Propulsion Laboratory

Welcome and Introductory Remarks

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Jet Propulsion Laboratory
California Institute of Technology

Objectives

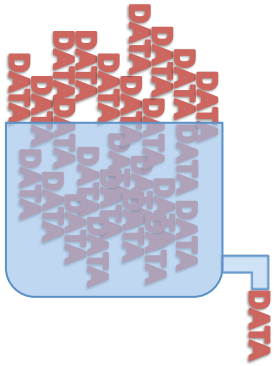
- Covers topics in the area of **data-intensive science** ranging from programming to computational methods for analyzing massive data
- Oriented towards computational science rather than computer science
- Familiarizes students with concepts and support learning through hands-on labs
- Delivered entirely as a MOOC

Jet Propulsion Laboratory
California Institute of Technology

Introduction
What is Big Data? What to Do About It?

The CHALLENGE: ***Big Data***

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

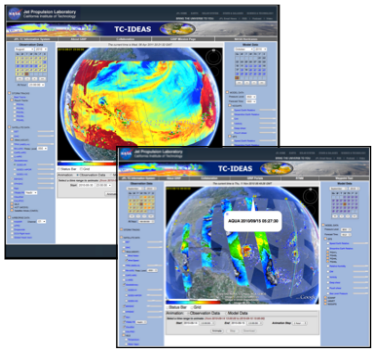


The SOLUTION: ***Data Science***

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

The RELEVANCE:

- Addressing the challenge of Big Data is on the critical path for many organizations
  - **the size and distribution of data sets and predictive models continues to burgeon**
  - **core objectives such as providing the basis of an analytic result are becoming compromised**

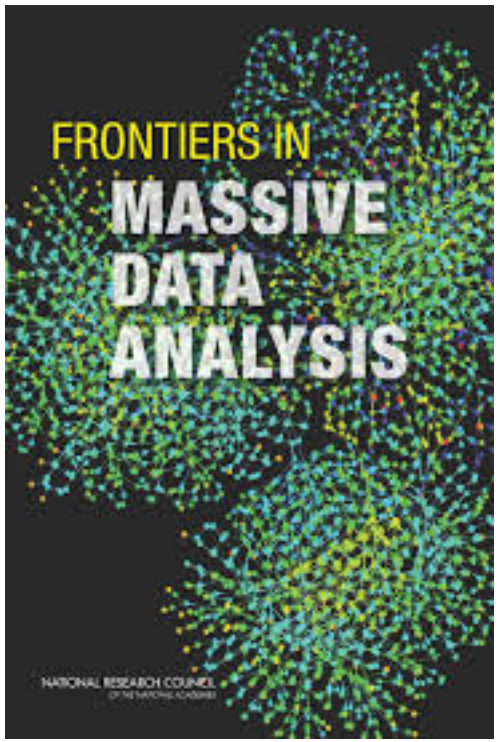


Jet Propulsion Laboratory
California Institute of Technology

NRC Report

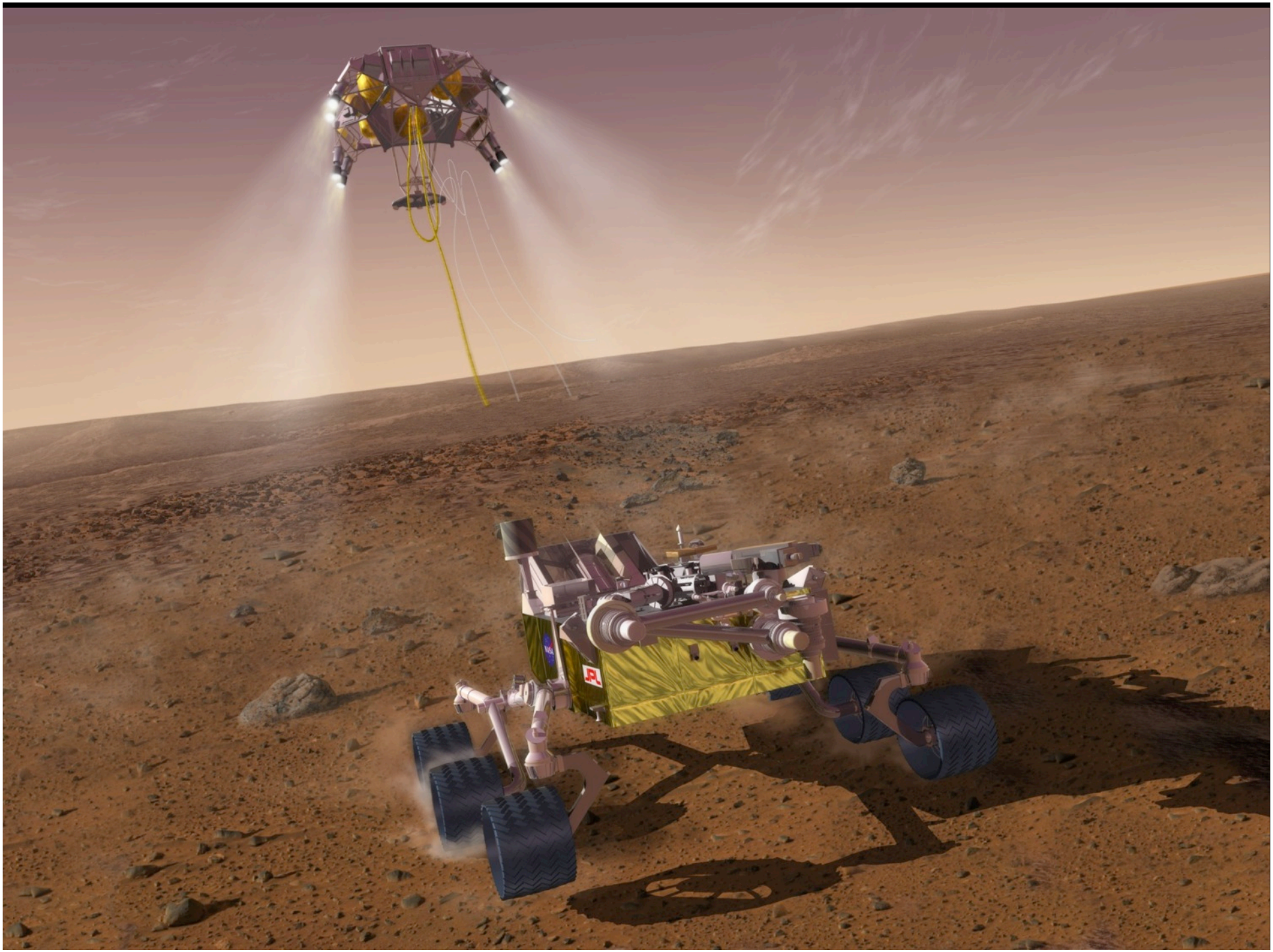
Frontiers in the Analysis of Massive Data

- Chartered in 2010 by the National Research Council
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- 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



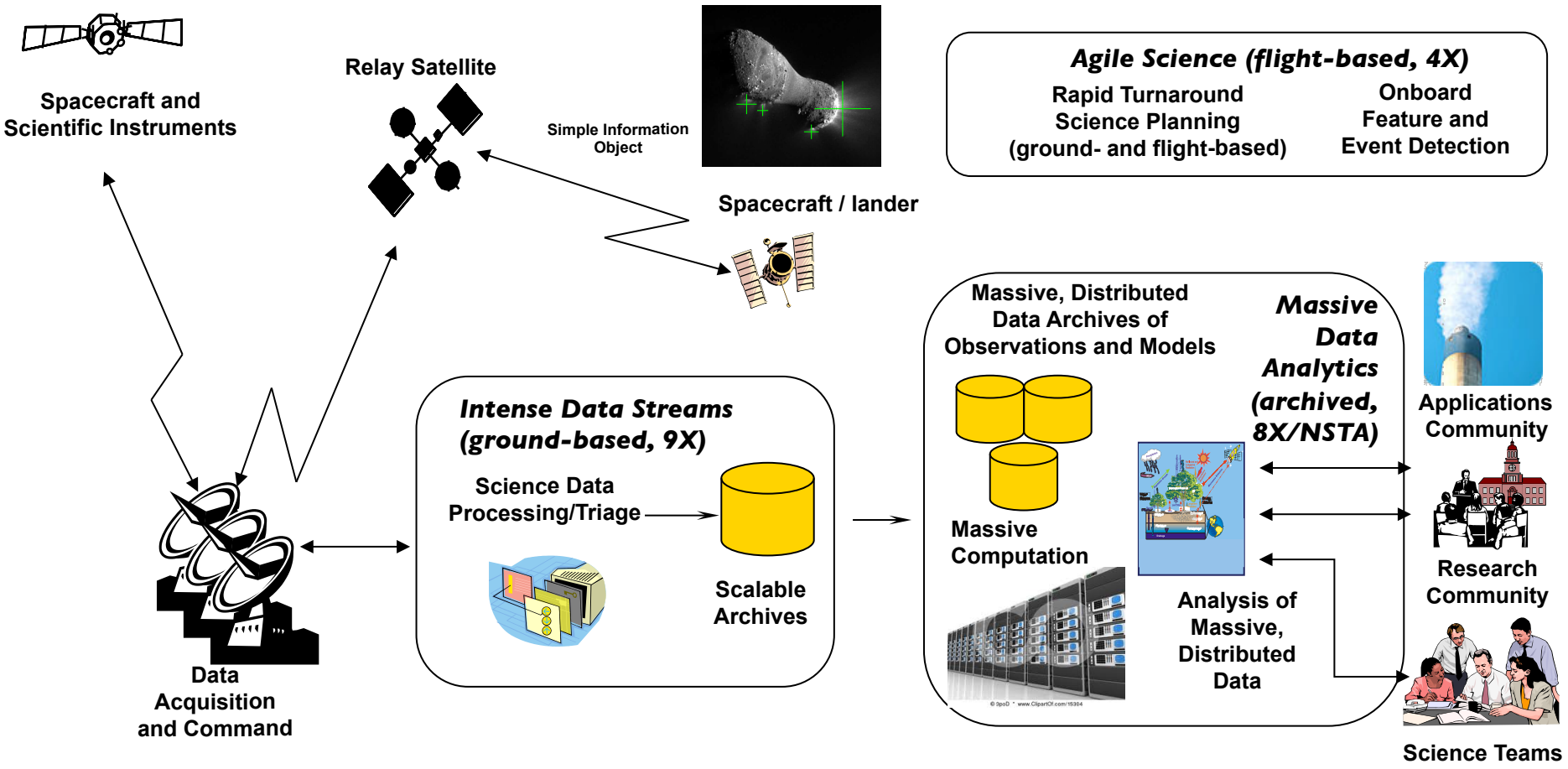
Published in 2013

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



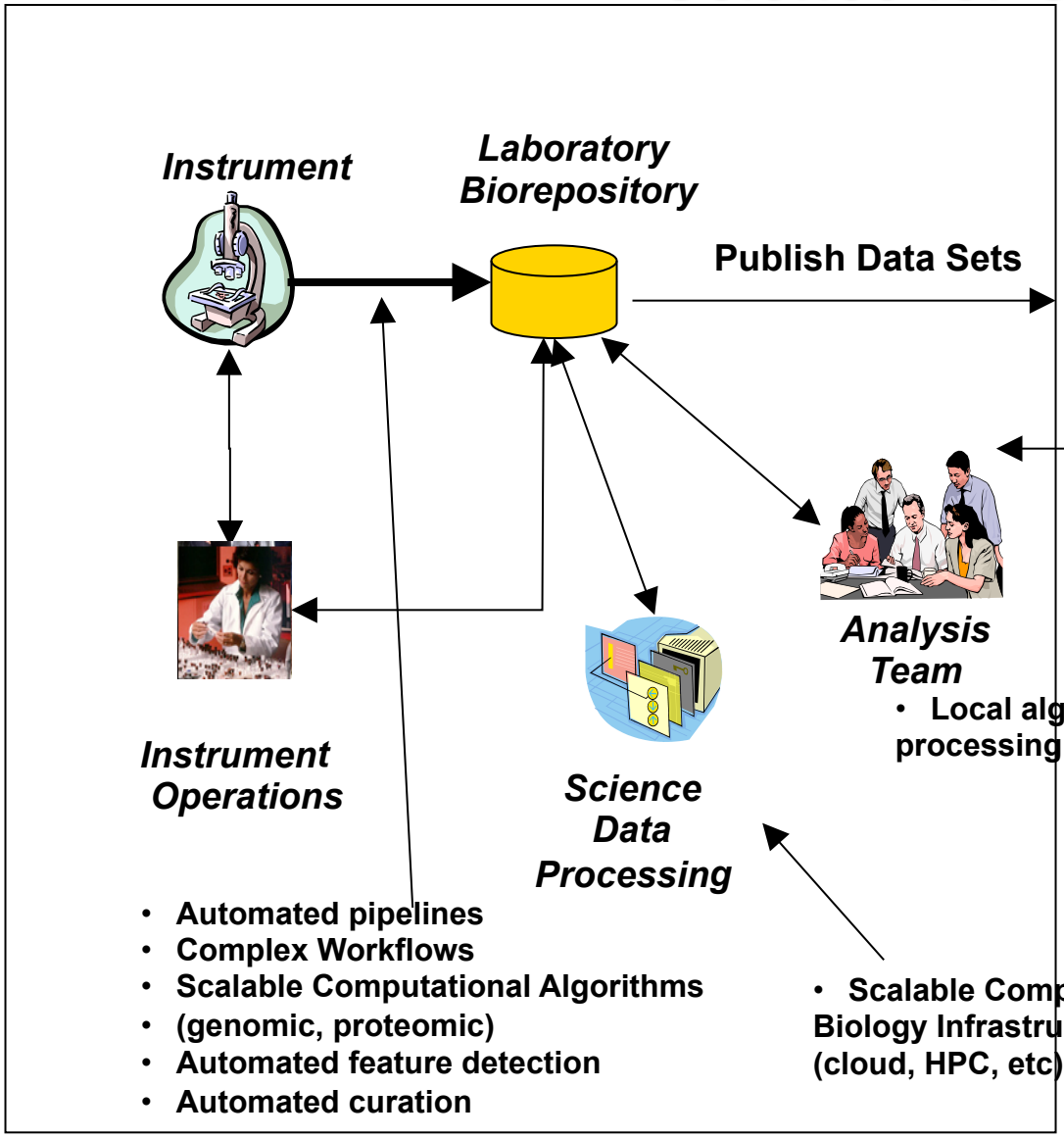
Jet Propulsion Laboratory
California Institute of Technology

Data Lifecycle Model for Space Missions



Jet Propulsion Laboratory
California Institute of Technology

NCI Early Detection Research Network: Moving towards Data-Driven Science for Cancer Biomarkers



-Cross-cutting Data/Information Architectures -

Jet Propulsion Laboratory
California Institute of Technology

The Need for Data Science

The Data Lifecycle: Too Fast, Too Much, Not Enough

Perform original processing at the sensor / instrument

Make choices at the collection point about which data to keep

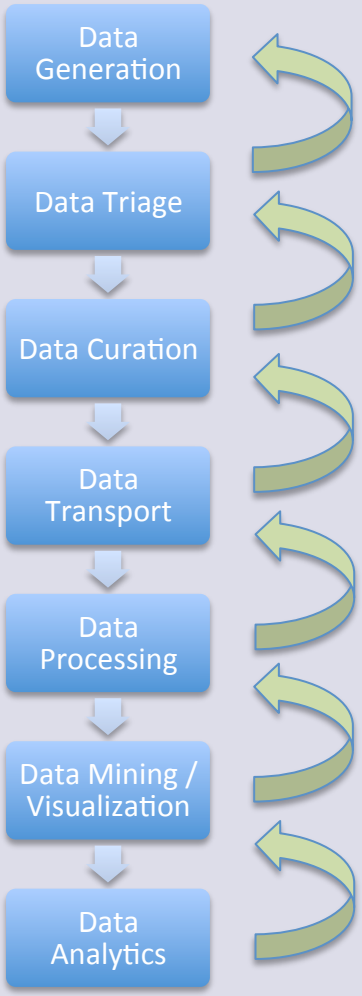
Anticipate the need to work across multiple data sources

Improve resource efficiencies to enable moving the most data

Increase computing availability at the data to generate products

Develop and apply analysis techniques to enable data understanding

Create analytics services effective across massive, distributed data



Data Science is the focused research to develop principled techniques and scalable architectures to address challenges across the entire Data Lifecycle

Jet Propulsion Laboratory
California Institute of Technology

Common Challenges in Massive, Distributed, Heterogeneous Data

- 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
- Developing mechanisms for identifying and extracting interesting features / patterns
- Developing methodologies for reconciling predictive models vs. measurements
- Methods for visualizing massive data
- Providing a trusted basis for actionable results of data analytics

Jet Propulsion Laboratory
California Institute of Technology

OODT: Object-Oriented Data Technologies

An Open Source Data Science Framework

- An open source data science framework
  - Developed at NASA/JPL
  - Top Level Project at the Apache Software Foundation (2011)
  - Used across multiple NASA centers (JPL, GSFC, LaRC)
  - Used across multiple agencies (NASA, NIH, NSF, DARPA, NOAA)
  - Integrates with an information architecture – metadata and ontology (e.g., earth science, biomedicine, etc.)
  - Significantly reduces cost and increases performance of data processing and management
  - Integrates with data analytics
- Applied to multiple Earth Science missions
  - Seawinds, OCO-2, SMAP, NPP Sounder Peate, JPSS
  - CARVE, Airborne Snow Observatory
- Applied to Earth science, planetary science, astronomy, biomedicine, defense

http://oodt.apache.org

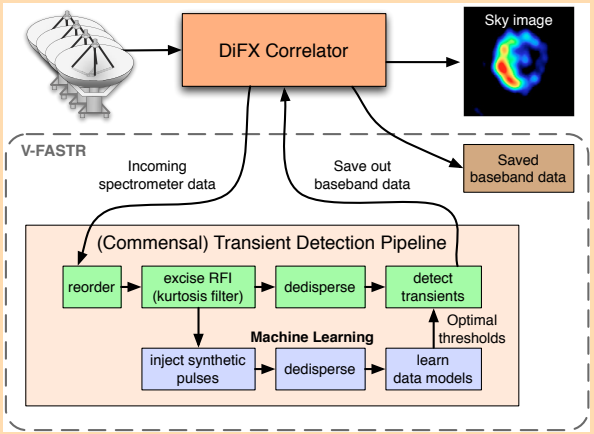
POC: Dan Crichton, Chris Mattmann

Jet Propulsion Laboratory
California Institute of Technology

Triage, Analysis, and Understanding of Massive Data

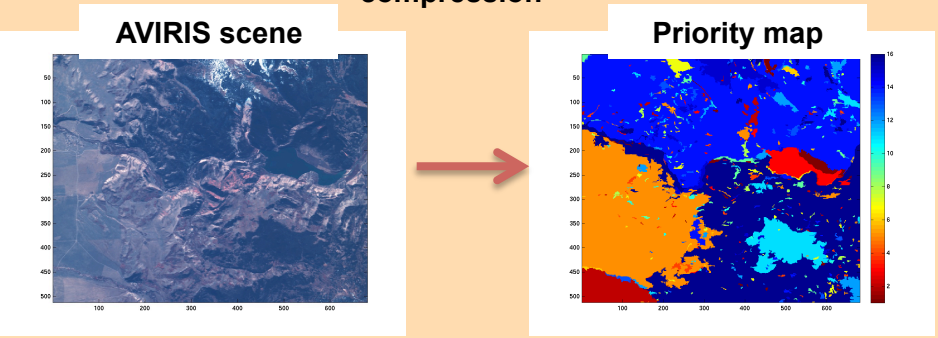
• Detection: fast identification of signals of interest (triage)

Radio astronomy:
V-FASTR realtime system at the VLBA



• Prioritization: use triage decisions to inform adaptive data compression

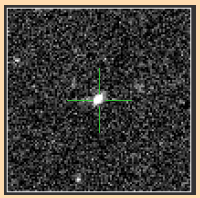
Earth science:
Onboard content-sensitive data compression

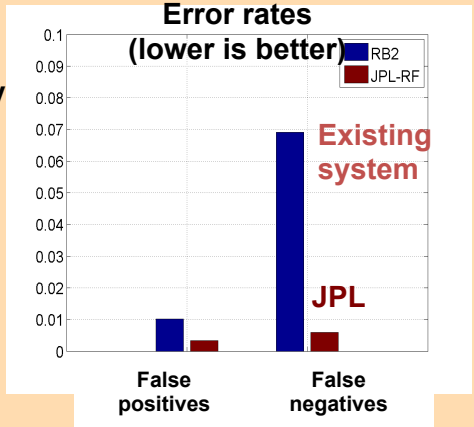


• Classification: online, real-time source type classification

Optical astronomy:
Reducing false positives for the Palomar Transient Factory

Real or spurious?

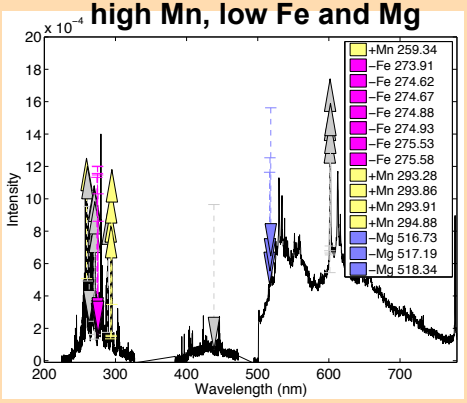




• Understanding: generate human-understandable explanations for decisions

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

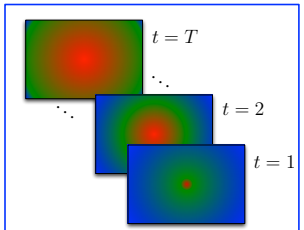
Jet Propulsion Laboratory
California Institute of Technology

Distributed Data Analytics

Model-to-Data Reconciliation for Carbon Cycles

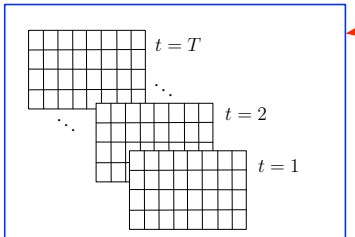
Science Goal:
- Understand processes that control CO2 flux

Strategy:
- Experiment with models to increase agreement observations / inferences



True fluxes: $s(u_1, u_2, t)$

(unknown)

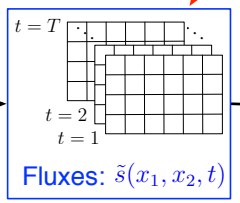


Inferred fluxes: $\hat{s}(u_1, u_2, t)$

Inference
(Inverse problem)

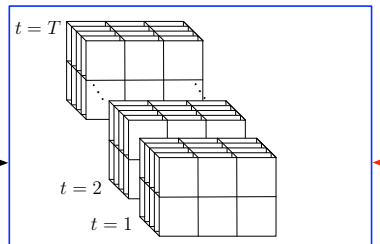
Compare

Flux model



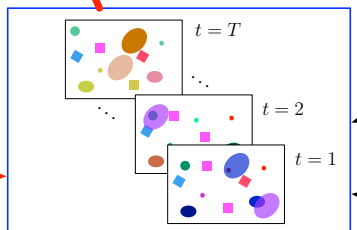
Fluxes: $\tilde{s}(x_1, x_2, t)$

Transport model



CO2 concentrations: $\tilde{x}(u_1, u_2, u_3, t)$

Compare

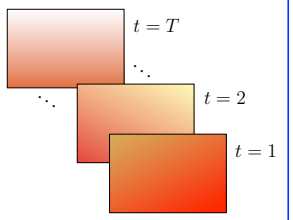


Heterogeneous, incomplete observations: $\hat{x}(u_1, u_2, t)$, $\hat{x}(u_1, u_2, u_3, t)$

In situ sampling

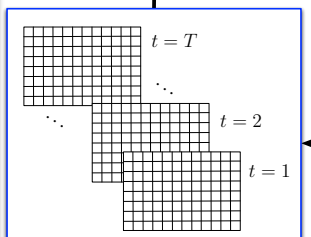
Ground-based retrieval

(unknown)

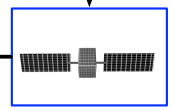


True CO2 concentrations: $x(u_1, u_2, u_3, t)$

Retrieval algorithm



Radiances: $y(u_1, u_2, t)$



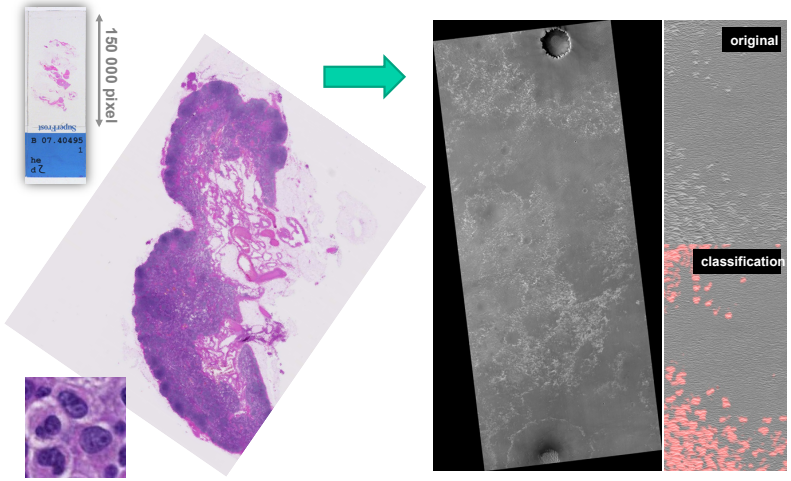
Analysis Challenges:
- Models and data reside at different locations
- Data are heterogeneous and must be reconciled as to format, scope, fidelity, resolution, etc.
- Meaningful comparison requires uncertainty estimation on both observational data and model output

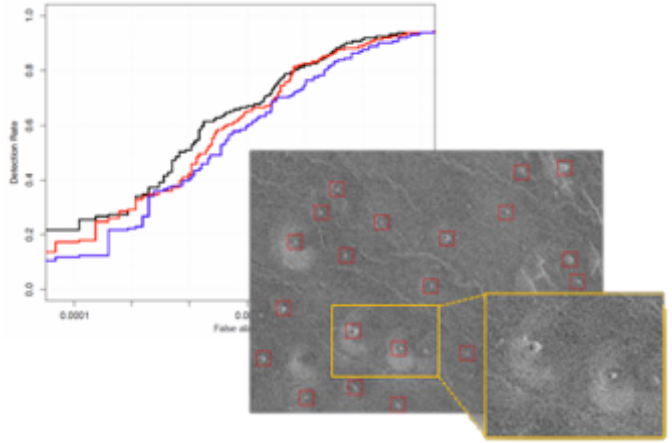
Data Science Contribution:
Address the analysis challenges to minimize both uncertainty and data movement as a dynamic optimization problem

POC: Amy Braverman

Jet Propulsion Laboratory
California Institute of Technology

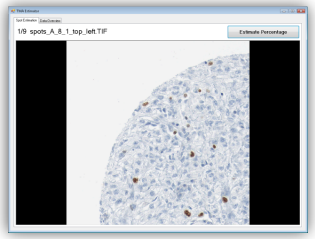
Application of Machine Learning Methodologies to Cancer Biomarker Research





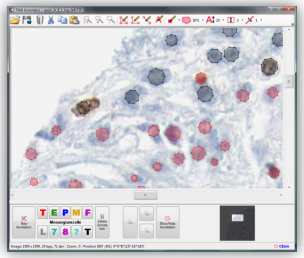
Volcanoes on Venus

TMA Estimator



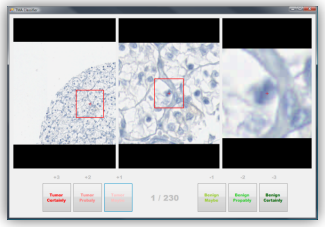
Estimate the Staining on a whole spot

TMA Annotator



Detect nuclei on a whole spot

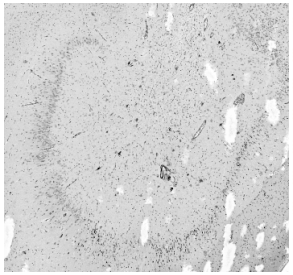
TMA Classifier



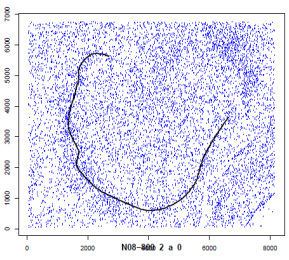
Classify single nuclei into tumor, non-tumor and stained, not-stained

Automated Classification

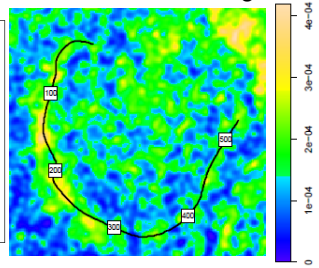
Original Image



Discriminative
Object Detection



Generative
P. Process Fitting



Feature/Object Detection

POC: Thomas Fuchs

Jet Propulsion Laboratory
California Institute of Technology

Summary

- **Data Science** is a growing area that requires new thinking across the data lifecycle, in software/system architectures, in the application of intelligent algorithms, in addressing uncertainty
- JPL and Caltech have both established Centers to respond to this important growing area
- Caltech and JPL are also partnering to maximize the value of combining our collective expertise
- JPL is working with NASA, other government agencies, academia and industry to bring together solutions