Data Science Program Overview

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

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

Jet Propulsion Laboratory
California Institute of Technology
Agenda

• JPL Overview
  – JPL Data Science Programs: Dan Crichton/Rich Doyle
  – Caltech Center for Data-Driven Discovery: George Djorgovski

• Machine Learning
  – Applications in Earth Science: Lukas Mandrake
  – Applications in Astronomy: Umaa Rebbapragada
  – Applications in Planetary Science: Kiri Wagstaff

• Data Analytics and Visualization
  – Sea Level Rise: Thomas Huang
  – Planetary Science: Emily Law/Shan Maholtra
  – Hydrology: Jay Famiglietti/Shan Maholtra

• Science Data Processing and Infrastructure
  – Commercial Partnering/Cloud Computing: Jim Rinaldi
  – SDS Plans for SWOT and NISAR: Hook Hua
Data Science Projects at JPL

Planetary Science

Biology

Defense and Intelligence

Earth Science

Medicine

Radio Astronomy
Focus on generating, capturing, managing big data

How do these connect?

Focus on using/analyzing big data
Future of Data Science at NASA
Enabling a Big Data Research Environment

Evolution towards an integrated data and computational environment
JPL Data Science Strategy

Guiding Principles

Agile Science – Onboard Analysis

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

Future Solutions: Onboard computation and data science

Extreme Data Volumes – Data Triage

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

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

Data Lifecycle

Organize data and computing end-to-end

Distributed Data Analytics

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

Future Solutions: Distributed data analytics; uncertainty quantification

Data Architecture

Perform original processing at the sensor / instrument

Make choices at the collection point about which data to keep

Improve resource efficiencies to enable moving the most data

Agglomerate the need to work across multiple data sources

Increase computing availability at the data to generate products

Increase the scale and integration of distributed archives

Apply visualization techniques to enable data understanding

Apply machine learning and statistics to enable data understanding

Create analytics services effective across massive, distributed data

Data Stewardship

Today

Data Analytics

Future

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

Cross-Cutting

Cut across disciplines to collaborate and share methodologies

JPL Data Science Strategy

Data Science Working Group

Earth Science

Planetary Science

Astrophysics

Operations

Information Technology

Formulation

Engineering

Business / HR

Non-NASA Applications

Data Ecosystem

Cut across disciplines to collaborate and share methodologies
NASA Big Data Landscape

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

Observational Platforms and Flight Computing

Emerging Solutions
- Intelligent Ground Stations
- Agile MOS-GDS

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

Massive Data Archives and Big Data Analytics

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

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

(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
Data System Scientific Research Networks: Access to Observations and Models

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

Solar System Exploration

Climate Research

Earth Observation

Cancer Research

National Data Sharing Infrastructure
Supporting Collaboration in Biomedical Research for EDRN
Cross-Cutting Capabilities

International Data Archive and Sharing Architectures
(from open source to cloud computing and scalable compute infrastructures)

Big Data Infrastructures

Intelligent Data Algorithms
(Machine Learning, Deep Learning)

Analytical Data Pipelines

Common Data Elements & Information Models
(discipline and common)

Great Opportunities for Methodology Transfer and Collaboration

Visualization Techniques
Evolving Towards Data Analytics

Data Stewardship
- Today

Massive Data Science Infrastructure (Compute, Storage, Data, Software)

On Demand Algorithms

Science Teams

Other Data Systems (e.g. NOAA)

Other Data Systems (In-Situ, Models, etc)

"Analytic Centers": Data Driven Analytics

Research

Applications

Decision Support

Data Analytics
- Future
The Growing Need for Data Science

“…data itself is no longer the number one problem; connected data is the problem. To overcome this challenge, organizations need to add edge analytics to their existing strategy, analyzing data close to its source instead of sending it to a central place for analysis.” - Mike Flannagan, Vice President, Data and Analytics, Cisco

“…traditional data analytics infrastructure will start to give way to strategic investments in data systems that are broad in scope (embracing all enterprise silos), provide distributed data infrastructures, use open source software…” - Tamr

“2016 will be the year where Artificial Intelligence (AI) technologies...are applied to ordinary data processing challenges...the new shift will include widespread applications of these technologies in ... tools that support applications, real-time analytics and data science. “ - Oracle

“Today’s operations centers struggle with an extremely high volume of events coming in requiring human analysis, which is unsustainable...in 2016 we will see organizations focus on using machine learning to significantly reduce the number of events requiring analysis down to the most critical.” - Snehal Antani, Splunk’s CTO

“Smallsats and the multi-trillion-dollar data set

But is that capital investment justified?
Data Science Growth Strategy

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

• Pilots – seed concepts and drive data science into the fabric of JPL
  – In 2017, JPL launched 12 funded pilots across science, mission operations, DSN, formulation, and business.
  – In 2018, this is expanding to engage a Lab-wide data science community.

• Services – mature capabilities to grow data science

• Projects – drive maturing services and pilots to address specific use cases
Partnering Strategy

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

• International Partners
  – Interoperability of archives
  – Engagement of technologies and data scientists across agencies

• Commercial and Open Source
  – Leverage mature technologies in cloud computing
  – Leverage and collaboration on big data technologies
  – Form public-private research partnerships
Applying Data Technologies Across Ground Environment

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

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

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

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

Data Analytics and Decision Support
Emerging Solutions
- Interactive Data Analytics
- Feature Detection and Extraction
- Uncertainty Quantification
- Error Detection in Data Collection
Visualization, Analytics and Applications

Interactive visualization of heterogeneous planetary objects

Examples: Hydrology and sea level rise
Integration of multiple earth observing remote sensing instruments; comparison against models

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

Real-time feature extraction and classification in astronomy
Recommendations

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

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

What do we do with all this data?

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

- Frontiers on Massive Data Analysis, NRC, 2013
- NASA OCT Technology Roadmap, NASA, 2015
- Planetary Science Informatics and Data Analytics Conference, April 2018
Questions?