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

Daniel J. Crichton
Leader, Center for Data Science and Technology
Proj. Mgr., Planetary Data System Engineering
PI, NCI/EDRN Informatics Center
Program Manager, Data Science
Program Manager, Data Systems and Technology

leaving the safe harbor to explore uncharted waters

Jet Propulsion Laboratory
California Institute of Technology

October 2016

The Data Lifecycle Perspective
Context
Observational Systems

What Do These Have in Common?

Data Lifecycle Challenges
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 data storage capacity

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

Preparing for exascale computing...

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

Distributed Data Analytics

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

Future Solutions: Distributed data analytics; uncertainty quantification
The Need for Data Science
Address Big Data Challenges Across the Data Lifecycle

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

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*
Shifting to Data Architecture and Data Analytics
Notional Science Data Pipeline
Constructing Scientific Archives

Major goal of capturing data as primary requirements
“There is a major need for the development of software components… that link high-level data analysis specifications with low-level distributed systems architectures.”

Enabling a Data-Driven Strategy

National Data Sharing Infrastructure

Big Data Infrastructure

Intelligent Data Algorithms

Common Data Elements & Models

Analytical Data Pipelines

Leveraged Capabilities for Methodology Transfer

Visualization Techniques
• An open source data science framework
  • Developed at NASA/JPL
  • Top Level Project at the Apache Software Foundation (2011)
  • Used across NASA centers (JPL, GSFC, LaRC)
  • Used across multiple agencies (NASA, NIH, NSF, DARPA, NOAA)
  • Framework to enable data science at across multiple centers to capture, integrate and analyze data

• Applied to Earth science, planetary science, astronomy, biomedicine, defense

• NASA’s first Open Source project at Apache!

http://oodt.apache.org
Data Triage, Analysis, and Understanding of Massive Data

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

- Prioritization: use triage decisions to inform adaptive data compression

- Classification: online, real-time source type classification

- Understanding: generate human-understandable explanations for decisions

Radio astronomy: V-FASTR realtime system at the VLBA

Optical astronomy: Reducing false positives for the Palomar Transient Factory

Real or spurious?

AVIRIS scene

Priority map

Earth science: Onboard content-sensitive data compression

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

Credits: K. Wagstaff, U. Rebbapragada, D. Thompson, B. Tang, H. Xie
Methodology Transfer in Data Science from Planetary & Earth to Biomedicine
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 ontology
  - Approximately 1 PB of data
  - Movement towards international interoperability
  - Implemented Apache OODT in 2002 to share data; scaled to a service-oriented architecture (SOA) implemented in 2010
Enabling a Model-Driven Data System

Information System Architecture

System Model
- Information Object
- Identification
- Referencing
- State

Domain Model (governance levels)
- Top Level
  - Representation/Format
  - Context, Provenance, Integrity
- Domain
  - Science
  - Engineering
  - Exploration
- Missions/Systems
  - Satellite/Airborne
  - Mission Operations

Configure System

Configurable Components
- Data Management Model
- Search/Access Model
- Analytics Model

Describe System

Data

Use

Produce

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.
Adoption of Re-Architected Planetary Data System (PDS4)

Planetary Data System Version 4
International, distributed, model-driven data architecture for capturing, managing and distributing planetary science data results to the world-wide science community.*

2000: 4 TBs; 2015: 1 PB

Data Science Challenges
Understanding Underground Water in the California Central Valley
WaterTrek: A platform for interactive data analytics for hydrology

- Shift from archives towards a scalable, automated, interactive platform integrating many different distributed sources of data
  - Scale with massive data processing and databases
  - Shift from archive “data containers” to analytic “data containers”
  - Fuse multiple data sources for visualization
  - Enable plug-in analytics
  - Ensure platform can apply to other disciplines

- Serve as a distribution point to support science and decision-maker needs in hydrologic research
Fusing In-situ, Air-borne, Space-borne and model generated data using visualization and a big data analytics engine.
Development of an advanced Knowledge System to *capture, share* and support *reproducible analysis* from the biomarker data results

- Genomics, Proteomics, Imaging, etc data types of data

NASA-NCl 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 NCI/EDRN and NASA/JPL informatics collaboration at a congressional briefing in October 2015

**JPL Informatics Center Team:**

- Dan Crichton - PI
- Kristen Anton (Dartmouth)
- Luca Cinquini
- Maureen Colbert (Dartmouth)
- Sean Kelly (Sloan Kettering)
- Thomas Fuchs
- Heather Kincaid
- David Liu
- Chris Mattmann
- Ashish Mahabal
Data-Driven Science for Cancer Biomarkers

“LabCAS”

**Instrument**

- Automated pipelines
- Complex Workflows
- Scalable Computational Algorithms
- (genomic, proteomic)
- Automated feature detection
- Automated curation

**Laboratory Biorepository**

Publish Data Sets

**Analysis Team**

- Local algorithms & processing
- Scalable Computational Biology Infrastructures (cloud, HPC, etc)

**Public Biorepository**

“eCAS”

Data Distribution

External Science Community

Results

Bioinformatics Tools

- On-demand algorithms
- Algorithms
- Data fusion methods
- Machine learning techniques

**Bioinformatics Community**

10/24/2016
A Virtual, National Integration Biomarkers Knowledge System
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
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

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.

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.
Driving Forward
Caltech/JPL Joint Initiative for Data Science and Technology

Information and Data Science Program (JPL/815):
Program Strategy, Partnerships, Investments, Roadmaps

Center for Data Science and Technology (JPL/3902):
People, Projects, Technologies, Research

Center for Data-Driven Discovery (Caltech): People, Projects, Technologies, Research

People: JPL and Caltech Data Science Research Affiliates

Projects: Shared NASA, NSF, NIH, DARPA, etc projects

Technologies: Core Caltech and JPL Product Lines and Thrusts

Primary Target: NASA/FFRDC Community, Agencies
Technology Capabilities: Architectures, Cyberinfrastructure, Machine Learning, Visualization Large-Scale Projects

Primary Target: Education, Focused Research, Foundations
Technology Capabilities: Machine Learning, et al

Caltech-JPL Executive Advisory Board on Computational Sciences

10/24/2016
Summary

• JPL has a growing Data Science program working across many agencies
  – Leveraging architectures, technologies and expertise around data science for observational systems

• JPL is partnered with Caltech via joint Centers for a strong research element

• Excellent opportunities continue to emerge to partner on Data Science challenges that revolve around observational systems
  – Significant leveraging of technologies and methodologies across scientific disciplines, from space science to biomedicine to … !

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!
Questions?
# Computational and Data Science

## Future Capability Needs

<table>
<thead>
<tr>
<th>System</th>
<th>2015</th>
<th>2025</th>
<th>NASA Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observational Platforms</td>
<td>Limited onboard computation including data triage and data reduction. Investments in new flight computing technologies for extreme environments.</td>
<td>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.</td>
<td>Onboard computation across all types of platforms; flight computing capabilities deployed for extreme environments; data triage for satellites and spacecraft.</td>
</tr>
<tr>
<td>Ground-based Mission Systems</td>
<td>Rigid data processing pipelines; limited real-time event/feature detection. Support for 500 TB missions.</td>
<td>Increase computational processing capabilities for mission (100x); Enable ad hoc workflows and reduction of data; Enable real-time triage/ML techniques, event and feature detection. Support 100 PB scale missions.</td>
<td>Future mission computational challenges; high bandwidth data volumes; more agile airborne, cube sat, multi-sensor campaigns; increase automated event detection across mission lifecycle.</td>
</tr>
<tr>
<td>Massive Data Archives</td>
<td>Support for 10 PB of archival data; limited automated event and feature detection.</td>
<td>Support exascale archives; automated event and feature detection/ML techniques; virtually integrated, distributed archives.</td>
<td>Turn archives into knowledge-bases to improve data discovery. Leverage massively scalable virtual data storage.</td>
</tr>
<tr>
<td>Distributed Data Analytics</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>
Chartered in 2010 by the NRC
- Members include UC Berkeley, Stanford, MIT, Georgia Tech, JHU, Google, Yahoo, and JPL

Emerging imperative to embrace Data Science at a national level across disciplines and industries
- *Data architecture* is central
- Need for end-to-end *data lifecycle* treatment from point of capture to analysis
- Need to call on both discipline/domain and data/computer science *subject matter expertise*
- Great potential to apply novel statistical and machine learning approaches for *data analytics*
- *Reproducibility* of analytic results is emerging as a critical-path challenge