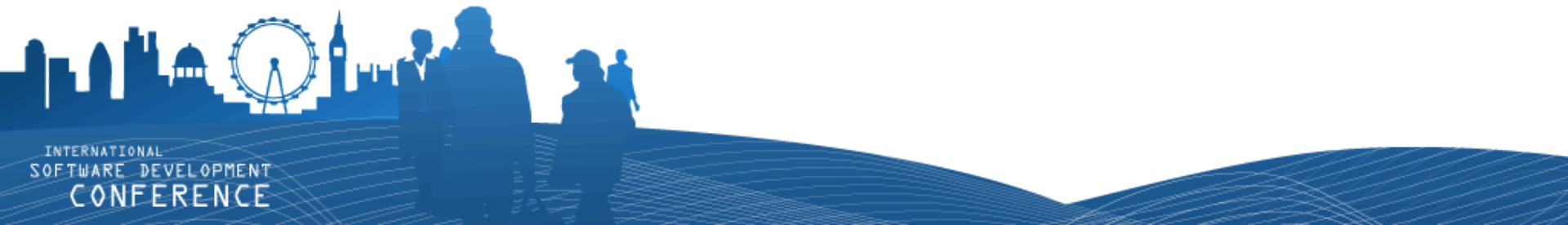
QCon

Please evaluate my talk via the mobile app!



Real Data Science at NASA

Chris Mattmann
Chief Architect, Instrument and Science Data Systems,
Jet Propulsion Laboratory, California Institute of Technology
Adjunct Associate Professor, USC
Director, Apache Software Foundation

Agenda

- Big Data – JPL's Initiative
- JPL's Big Data: ASO, RCMES, CMAC, SKA
  - Rapid Algorithm Integration, Smart Data Movement, Transient Archives, Automated text/metadata extraction and MIME identification
- Efforts with other agencies (DARPA, NSF)
- Big Data Opportunities

And you are?

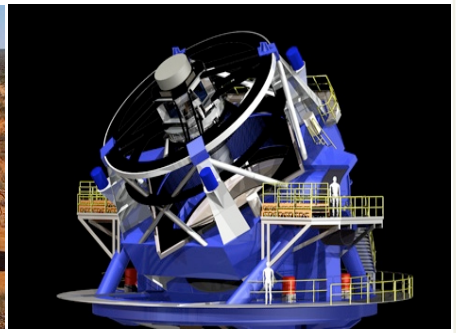
- Chief Architect at NASA JPL in Pasadena, CA USA
- Software Architecture/ Engineering Prof at Univ. of Southern California
- One of original PMC members for Apache Nutch
  - predecessor to Hadoop
- Apache Board of Directors involved in
  - OODT (VP, PMC), Tika (PMC), Nutch (PMC), Incubator (PMC), SIS (PMC), Gora (PMC), Airavata (PMC)

Some "Big Data" Grand Challenges I'm interested in

- *How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?*
  - Required by the Square Kilometre Array
- *Joe scientist says I've got an IDL or Matlab algorithm that I will not change and I need to run it on 10 years of data from the Colorado River Basin and store and disseminate the output products*
  - Required by the Western Snow Hydrology project
- *How do we compare petabytes of climate model output data in a variety of formats (HDF, NetCDF, Grib, etc.) with petabytes of remote sensing data to improve climate models for the next IPCC assessment?*
  - Required by the 5th IPCC assessment and the Earth System Grid and NASA
- *How do we catalog all of NASA's current planetary science data?*
  - Required by the NASA Planetary Data System

Image Credit: http://www.jpl.nasa.gov/news/news.cfm?release=2011-295

Big Data Strategic Initiative



Future Opportunities: Mission and instrument competitions, data-intensive industries, LSST, future radio observatories.

JPL Concept: Big data technology for data triage, archiving, etc.

Key Challenges this work enables: Broaden JPL business base (relevant to 1X, 3X, 4X, 7X, 8X, 9X Directorates)

Initiative Long Term Objectives

- Apply lower-efficient digital architectures to future JPL flight instrument developments and proposals.
- Expand and promote JPL expertise with machine learning algorithm development for real-time triage.
- Utilize intelligent anomaly classification algorithms in other fields, including data-intensive industry.
- Build on JPL investments in large data archive systems to capture role in future science facilities.
- Enhance the efficiency and impact of JPL's data visualization and knowledge extraction programs.

Initiative Leader: Dayton Jones
Steering Committee Leader: Robert Preston

Task Title	PI	Section
1 Power Minimization in Signal Processing for Data-Intensive Science	Larry D'Addario	335
2 Machine Learning for Smart Triage of Big Data	Kiri Wagstaff	388
3 Archiving, Processing and Dissemination for the Big Date Era	Chris Mattmann	388
4 Knowledge driven Automated Movie Production Environment distribution and Display (AMPED) Pipeline	Eric De Jong	3223

Initial Major Milestones for FY13	Date
Report on end-to-end power optimization of instruments	Jun 2013
Hierarchical classification method for VAST and ChemCam	Jan 2013
Demonstrate smart compression for Hyperion and CRISM	Mar 2013
Cloud computing research and scalability experiments	Feb 2013
Data formats and text, metadata extraction in big data sys.	Aug 2013
Develop AMPED pipeline and install in VIP Center	Dec 2012

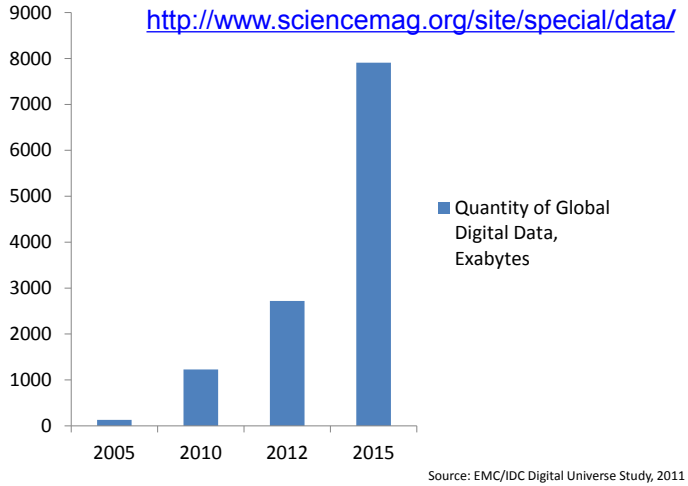
Credit: Dayton Jones

Archiving, Processing and Dissemination for the Big Data Era
Overview

- The Big Picture
  - Astronomy, Earth science, planetary science, life/physical science all *drowning in data*
  - Fundamental technologies and emerging techniques in archiving and data science
    - Largely center around *open source* communities and related systems

- Research challenges (adapted from NSF)
  - *More data is being collected than we can store*
  - *Many data sets are too large to download*
  - *Many data sets are too poorly organized to be useful*
  - *Many data sets are heterogeneous in type, structure*
  - *Data utility is limited by our ability to use it*

http://www.sciencemag.org/site/special/data/



Source: EMC/IDC Digital Universe Study, 2011

- Proposal focus: Big Data Archiving
  - *Research methods for integrating intelligent algorithms for data triage, subsetting, summarization*
  - *Construct technologies for smart data movement*
  - *Evaluate cloud computing for storage/processing*
  - *Construct data/metadata translators "Babel Fish"*

Archiving, Processing and Dissemination for the Big Data Era
Objectives vs Technical Challenge

- **Task Objectives vs Identified Challenge**
  - Infuse specific Big Data technologies and output from this effort into 8X Earth Science spaceborne and airborne missions (DESDynI, SMAP, ASO); into the National Climate Assessment; and 9x Astronomy work in future radio array precursors, e.g., HERA, KAT-7, LOFAR

U.S. National Climate Assessment (pic credit: Dr. Tom Painter)

  - Lead JPL participation in NSF Earth Cube Big Data Effort.
  - Work with the National Radio Astronomy Observatory (NRAO) and future radio observatories to explore Big Data technologies and techniques

SKA South Africa: Square Kilometre Array (pic credit: Dr. Jasper Horrell, Simon Ratcliffe)

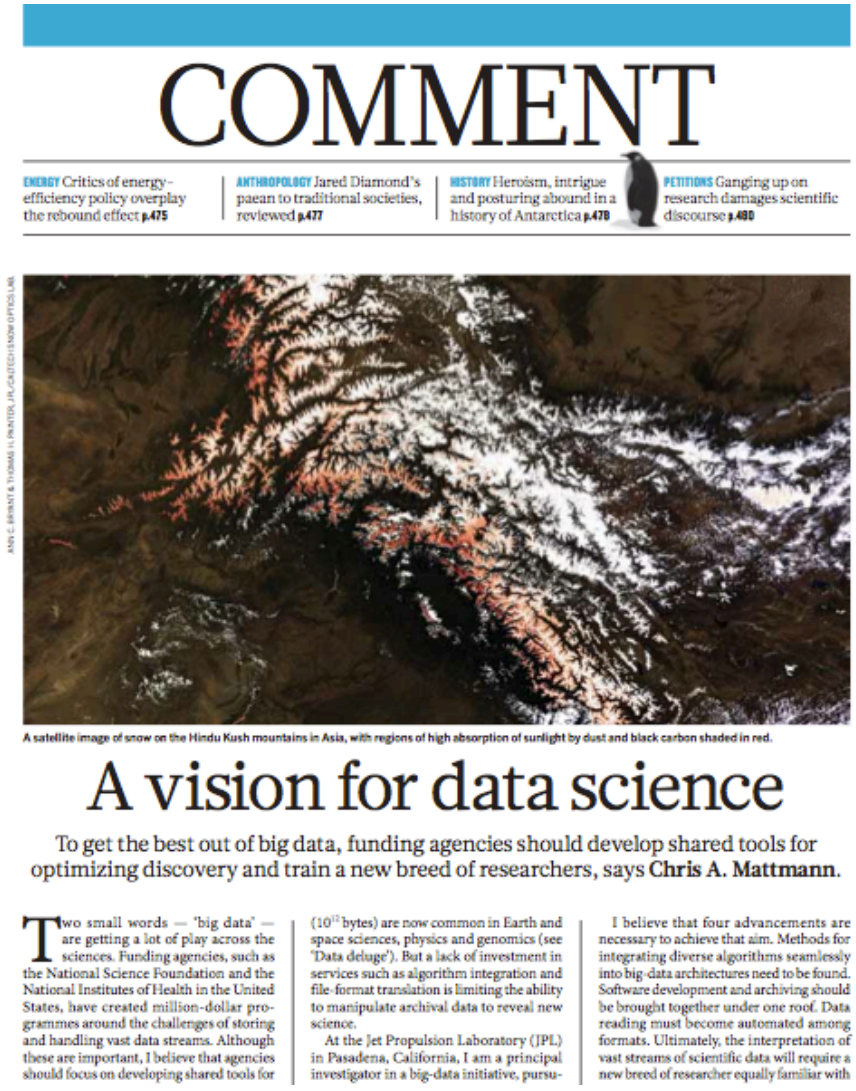
- **Technical Approach Highlights**
  - Rapid Science Algorithm Integration, data movement technology research, cloud computing, automatic data/metadata format and model extraction and classification

Archiving, Processing and Dissemination for the Big Data Era
Deliverables, Milestones and Schedule

- **Deliverables**
  - Framework for rapidly and unobtrusively integrating science algorithms for National Climate Assessment, RCMES, and ASO (8x) and for the CASA system (9x)
  - Research study identifying set of appropriate/inappropriate data movement technologies for 8x and 9x big data systems
    - Consideration of national and international collaborators
  - Cloud computing research study for DESDynI, ASO and other spaceborne and airborne missions – also for KAT-7 (int') work.
  - Extensions to JPL-led Apache Tika and Apache OODT technologies to handle data formats required by NCA, RCMES, ASO, and radio observatories (HERA, etc.)

- **Milestones/Schedule**
  - **M1 +2 months** Research study on science algorithm integration framework – apply to CASA and ASO/NCAs
  - **M2 +6 months** Cloud computing study for ASO and KAT-7 and HERA
  - **M3 +9 months** Data movement technology study
  - **M4 +12 months** study on data/formats to support ASO, HERA, RCMES, KAT-7 and HERA

Recent pub highlights



COMMENT

ENERGY Critics of energy-efficiency policy overplay the rebound effect p.475 | ANTHROPOLOGY Jared Diamond's paean to traditional societies, reviewed p.477 | HISTORY Heroism, intrigue and posturing abound in a history of Antarctica p.478 | PETITIONS Ganging up on research damages scientific discourse p.480

A satellite image of snow on the Hindu Kush mountains in Asia, with regions of high absorption of sunlight by dust and black carbon shaded in red.

A vision for data science

To get the best out of big data, funding agencies should develop shared tools for optimizing discovery and train a new breed of researchers, says **Chris A. Mattmann**.

Two small words — 'big data' — are getting a lot of play across the sciences. Funding agencies, such as the National Science Foundation and the National Institutes of Health in the United States, have created million-dollar programmes around the challenges of storing and handling vast data streams. Although these are important, I believe that agencies should focus on developing shared tools for (10^{12} bytes) are now common in Earth and space sciences, physics and genomics (see 'Data deluge'). But a lack of investment in services such as algorithm integration and file-format translation is limiting the ability to manipulate archival data to reveal new science.

At the Jet Propulsion Laboratory (JPL) in Pasadena, California, I am a principal investigator in a big-data initiative, pursu-

I believe that four advancements are necessary to achieve that aim. Methods for integrating diverse algorithms seamlessly into big-data architectures need to be found. Software development and archiving should be brought together under one roof. Data reading must become automated among formats. Ultimately, the interpretation of vast streams of scientific data will require a new breed of researcher equally familiar with

- Nature magazine piece on "A Vision for Data Science" in Jan. 24th issue
  - Big Data Initiative highlighted
- *Outline algorithm integration (regridding, metrics); automatic understanding of data metadata formats and open source as "key issues"*

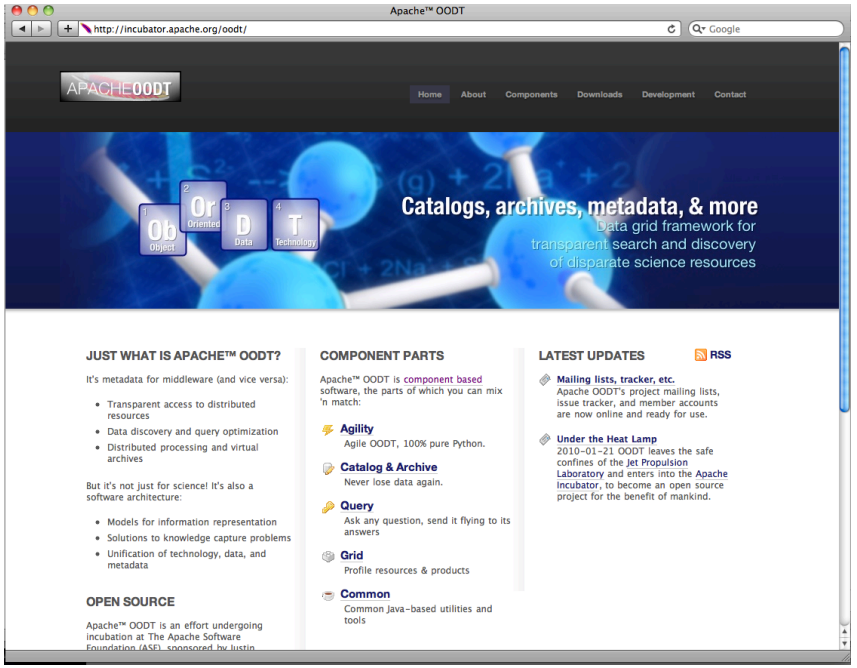
Data Science/Big Data progress

- Named to Editorial Board of Springer Journal of Big Data
- Helping to define USC's M.S. in Data Science program
- Won/Submitted several Big Data proposals for direct funding
  - DARPA Open Source Program Office XDATA
  - NSF Major Research Instrumentation (RAPID)
  - NSF Polar Cyberinfrastructure, NSF EarthCube (both via USC)
  - President's/Director's Fund for Cosmic Dawn/OVRO
  - National Science Foundation: High Performance Computing System Acquisition (submitted)

Open Source + some projects in Big Data/Data Science near and dear to my heart

Apache OODT

- Entered “incubation” at the Apache Software Foundation in 2010
- Selected as a top level Apache Software Foundation project in January 2011
- Developed by a community of participants from many companies, universities, and organizations
- Used for a diverse set of science data system activities in planetary science, earth science, radio astronomy, biomedicine, astrophysics, and more



OODT Development & user community includes:

Apache OODT: OSS "big data" platform originally pioneered at NASA

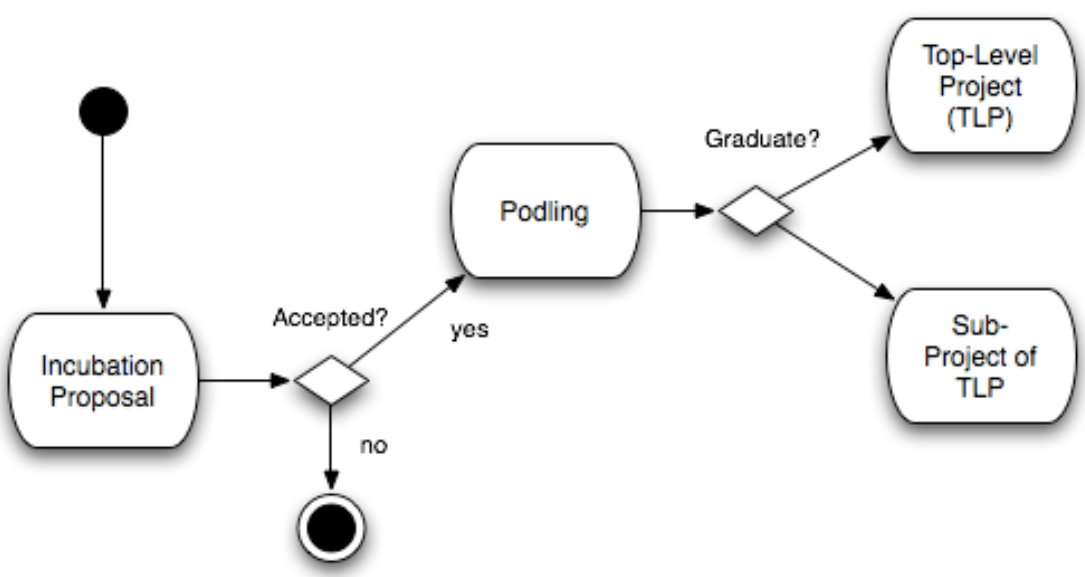
- OODT is meant to be a set of tools to help build data systems
  - It's not meant to be "turn key"
  - It attempts to exploit the boundary between bringing in capability vs. being overly rigid in science
  - Each discipline/project extends
- Projects that are deploying it operationally at
  - Decadal-survey recommended NASA Earth science missions, NIH, and NCI, CHLA, USC, South African SKA project
- Why Apache?
  - Less than 100 projects have been promoted to top level (Apache Web Server, Tomcat, Solr, Hadoop)
  - Differs from other open source communities; it provides a governance and management structure

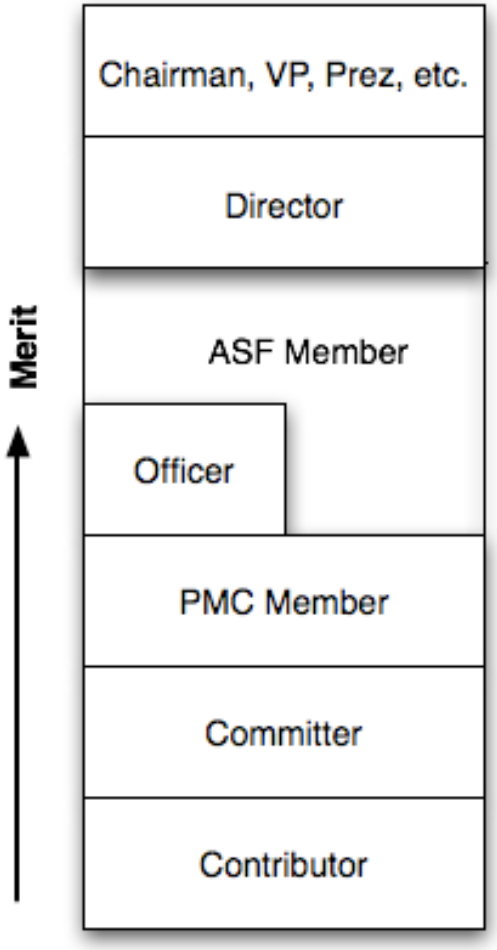
Why Apache and OODT?

- OODT is meant to be a set of tools to help build data systems
  - It's not meant to be "turn key"
  - It attempts to exploit the boundary between bringing in capability vs. being overly rigid in science
  - Each discipline/project extends

- Apache is the elite open source community for software developers
  - Less than 100 projects have been promoted to top level (Apache Web Server, Tomcat, Solr, Hadoop)
  - Differs from other open source communities; it provides a governance and management structure

Governance Model+NASA=♥





- NASA and other government agencies have tons of process
  - They like that

Apache Open Climate Workbench..OCW

Apache Open Climate Workbench (Incubating)
http://t.co/i4gQilDSwZ this is fing epic. well done apache! cc @tomraftery

@monkchips Yesterday

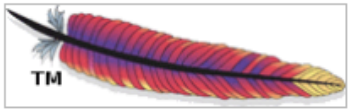
Follow @monkchips

3 retweets

Apache Open Climate Gets Top-Level Project Status

by Sudarshana Banerjee • March 3, 2014

Like Share 1

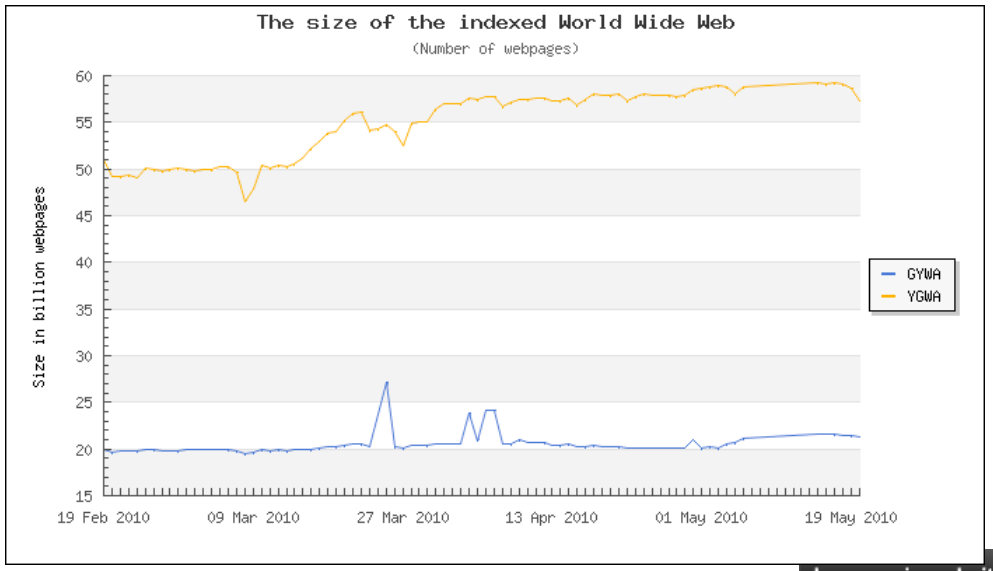


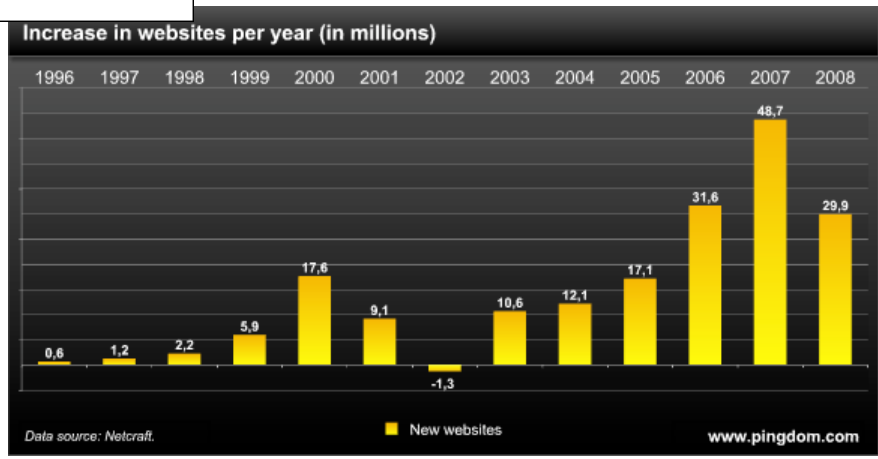
The Apache Open Climate Workbench Project has graduated from the Apache Incubator to become a Top-Level Project (TLP), says the Apache Software Foundation. The Apache Open Climate Workbench software is released under the Apache License v2.0.

Apache Climate is a climate evaluation toolkit used to leverage model outputs from organizations such as the Earth System Grid Federation (ESGF), the Coordinated Regiona Downscaling Experiment (CORDEX), the U.S. National Climate Assessment (NCA), and the North American Regional Climate Change Assessment Program (NARCCAP), coupled with remote sensing data from the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and other agencies.

[Image courtesy: Apache Software Foundation]

The Information Landscape





Proliferation of content types available

- By some accounts, 16K to 51K content types*
- What to do with content types?
  - Parse them
    - How?
    - Extract their text and structure
  - Index their metadata
    - In an indexing technology like Lucene, Solr, or in Google Appliance
  - Identify what language they belong to
    - Ngrams

*http://filext.com/

Apache™ Tika is…

- A content analysis and detection toolkit
- A set of Java APIs providing MIME type detection, language identification, integration of various parsing libraries
- A rich Metadata API for representing different Metadata models
- A command line interface to the underlying Java code
- A GUI interface to the Java code

Science Data File Formats

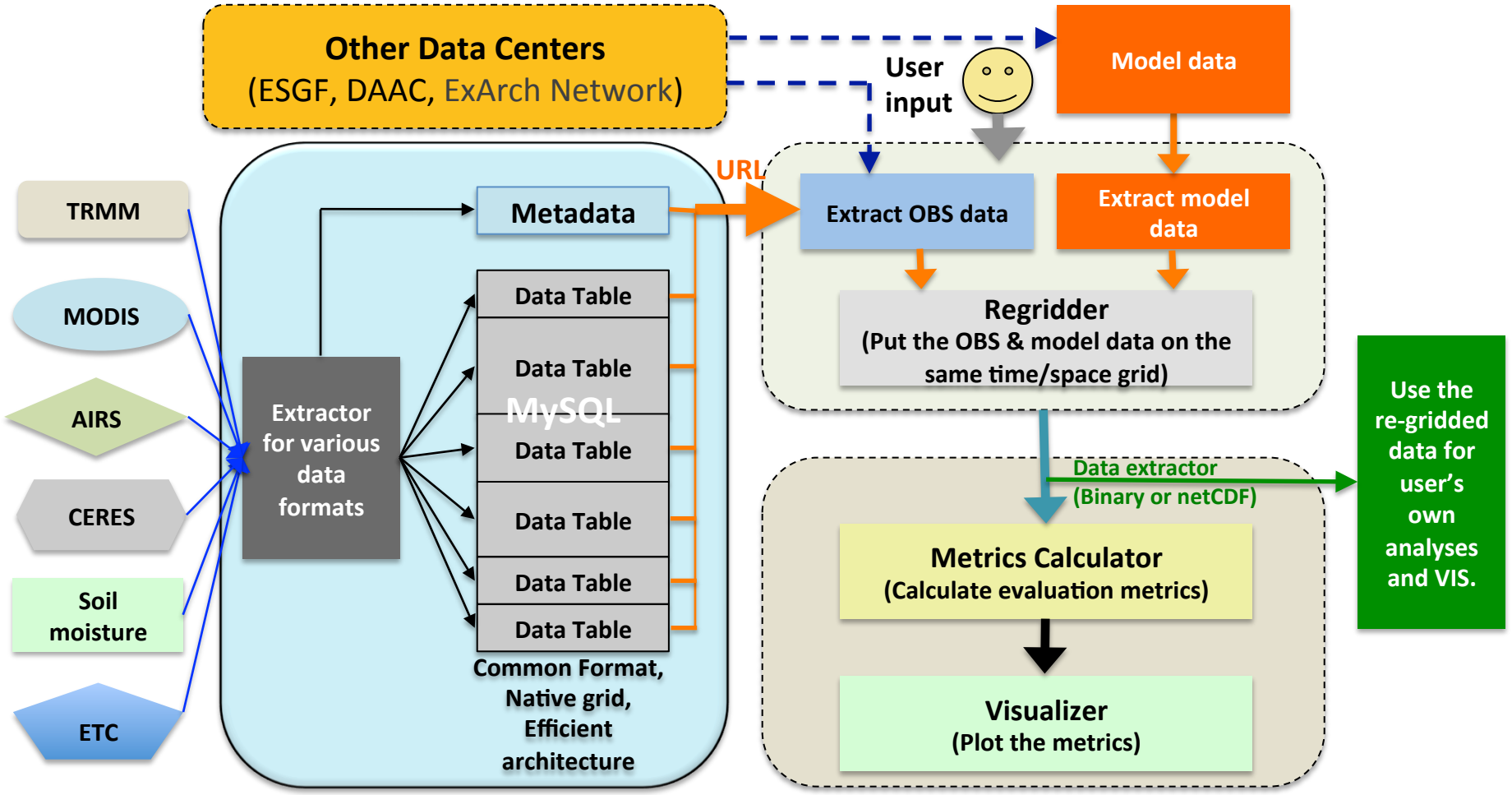
- Hierarchical Data Format (HDF)
  - http://www.hdfgroup.org
  - Versions 4 and 5
  - Lots of NASA data is in 4, newer NASA data in 5
  - Encapsulates
    - Observation (Scalars, Vectors, Matrices, NxMxZ…)
    - Metadata (Summary info, date/time ranges, spatial ranges)
  - Custom readers/writers/APIs in many languages
    - C/C++, Python, Java

Science Data File Formats

- network Common Data Form (netCDF)
  - www.unidata.ucar.edu/software/**netcdf**/
  - Versions 3 and 4
  - Heavily used in DOE, NOAA, etc.
  - Encapsulates
    - Observation (Scalars, Vectors, Matrices, NxMxZ…)
    - Metadata (Summary info, date/time ranges, spatial ranges)
  - Custom readers/writers/APIs in many languages
    - C/C++, Python, Java
  - Not Hierarchical representation: all flat

Now some specific NASA/JPL project examples

RCMES2.0
(http://rcmes.jpl.nasa.gov)



Raw Data:
Various sources, formats, Resolutions, Coverage

RCMED
(Regional Climate Model Evaluation Database)
A large scalable database to store data from variety of sources in a common format

RCMET
(Regional Climate Model Evaluation Tool)
A library of codes for extracting data from RCMED and model and for calculating evaluation metrics

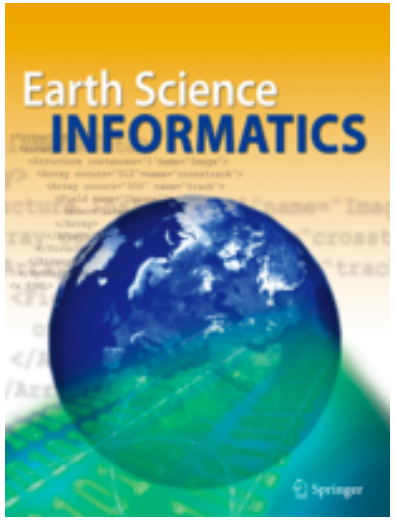
Evaluation of Cloud Computing for Storage & Application of NASA Observations

Challenge

- Regional climate model evaluation with daily temporal resolution to assess representation of extreme events.
- More voluminous, requires scalability in web services, system throughput, and also elasticity based on study demands

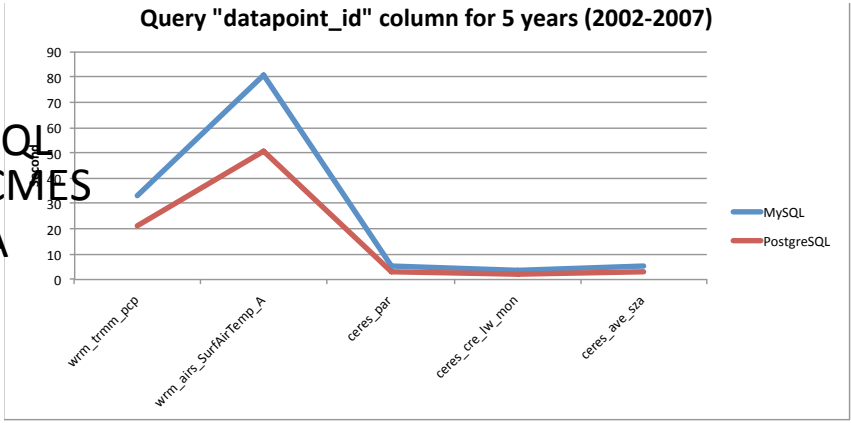
Objective

- Understand and evaluate popular cloud computing technologies, and provide a framework for selecting the best one for supporting Regional Climate Model Evaluation System (RCMES) & applications such as the National Climate Assessment and IPCC's CORDEX regional model evaluations.



Results

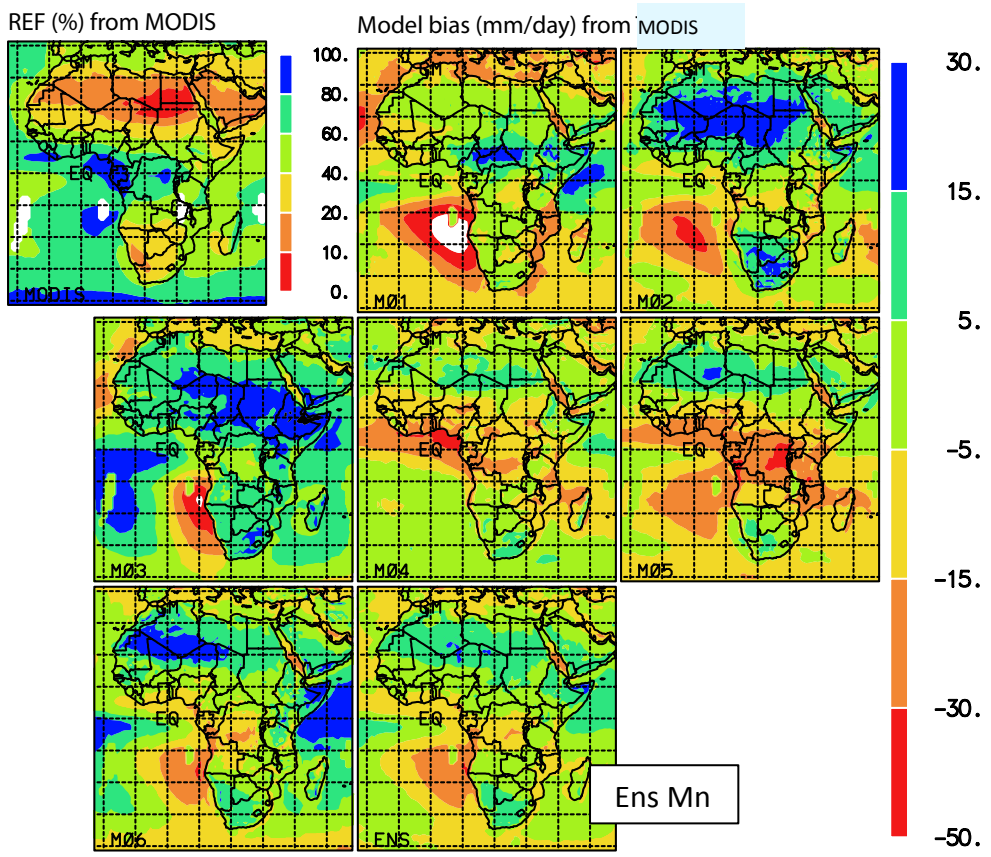
- Conducted evaluation demonstrating 44 % avg query time speedup of PostGIS over MySQL for 5 years of 5 parameters of obs data in RCMES
- Will incorporate into RCMES to facilitate NCA and CORDEX regioal model evaluations.



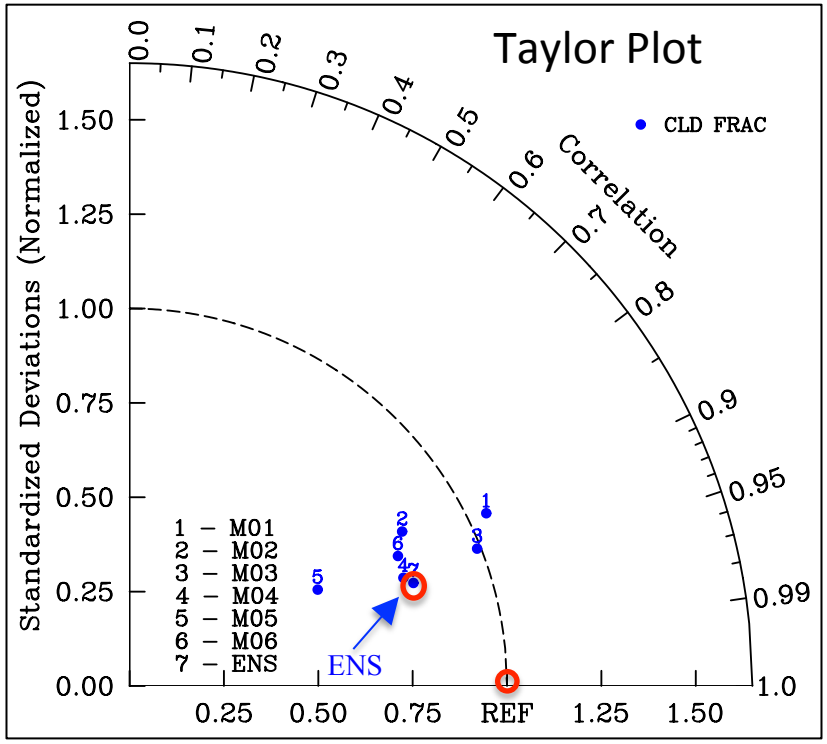
C. Mattmann, D. Waliser, J. Kim, C. Goodale, A. Hart, P. Ramirez, D. Crichton, P. Zimdars, M. Boustani, H. Lee, P. Loikith, K. Whitehall, C. Jack, B. Hewitson. Cloud Computing and Virtualization Within the Regional Climate Model and Evaluation System. *Earth Science Informatics*, 2013.

Example application for CORDEX-Africa

Annual Cloudiness Climatology Against MODIS; 2001-2008



Kim et al. 2013, In press



NOTE: The blank areas in the REF (MODIS) data are due to missing values.

NARCCAP Multi-decadal Hindcast Evaluation Result

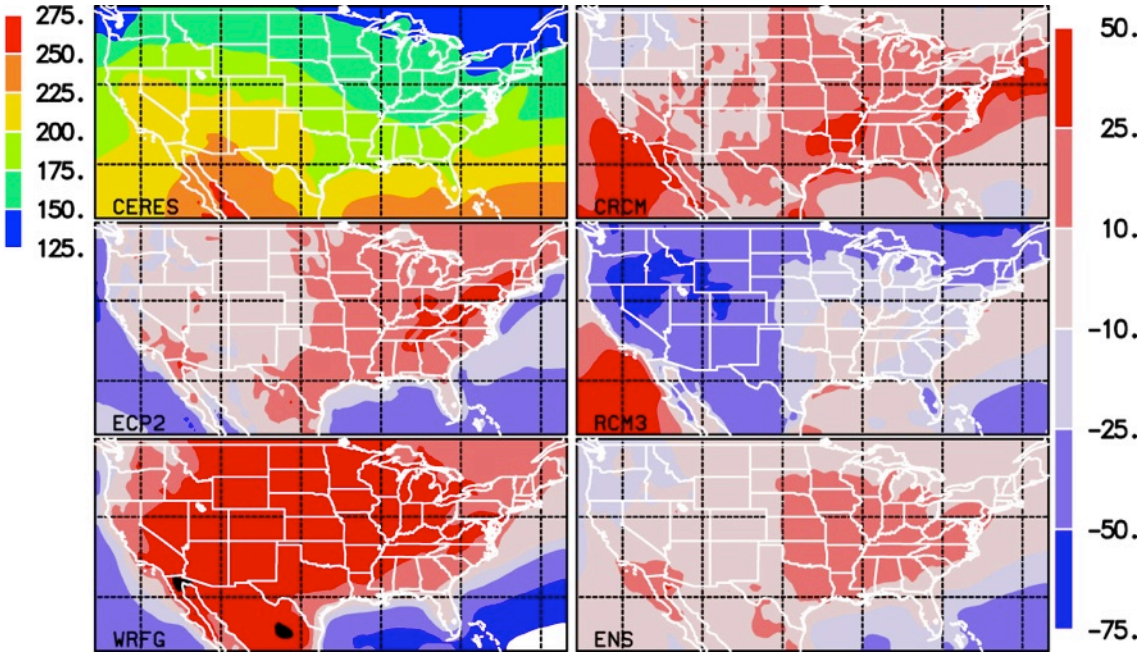


Figure. RCM biases in surface insolation against CERES

Considerable biases exist in surface insolation fields, a not so typical variable scrutinized in RCMs.

Kim, J., D.E. Waliser, C.A. Mattmann, L.O. Mearns, C.E. Goodale, A.F. Hart, D.J. Crichton, and S. McGinnis, 2013: Evaluations of the surface air temperature, precipitation, and insolation over the conterminous U.S. in the NARCCAP multi-RCM hindcast experiments using RCMES. J. Climate, In press.

Table 2. The relationship between precip & insolation biases.

Model	Land-mean bias – Precipitation (mm/d)	Land-mean bias - Insolation (Wm^{-2})	Bias pattern Correlation
CRCM	0.33	10.2	-0.47
ECP2	0.41	9.0	-0.28
RCM3	0.54	-29.9	-0.50
WRFG	-0.08	30.4	-0.18
ENS	0.25	4.9	-0.62

Table 1. The RCMs evaluated in this study.

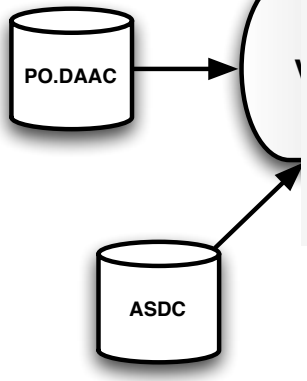
Model ID	Model Name
M01	CRCM (Canadian Regional Climate Model)
M02	ECP2 (NCEP Regional Spectral Model)
M03	MM5I (MM5 – run by Iowa State Univ.)
M04	RCM3
M05	WRFG (WRF – run by PNNL)
ENS	Model Ensemble (Uniform weighting)

CMAC NextGen Infra

CMAC-2: Deployment and connection to NASA data centers

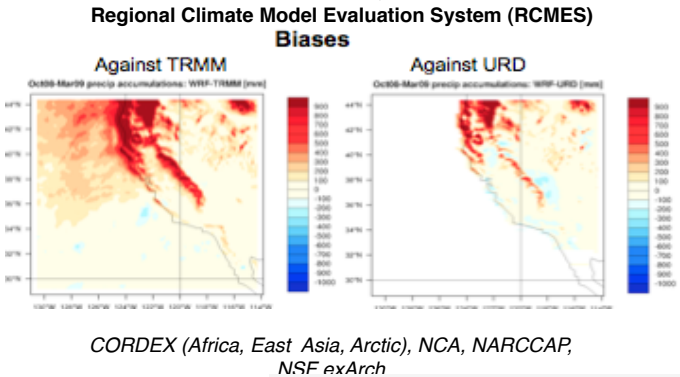
CMAC-1: Automated Conversion, Validation and Publication Toolkit

CMAC-3: Facilitate CORDEX, NCA, and obs4MIPs comparisons

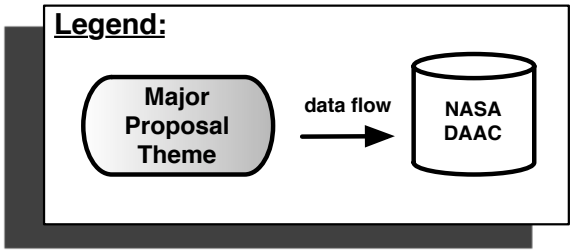


Earth System Grid Federation ESGF

Model to Data Comparison

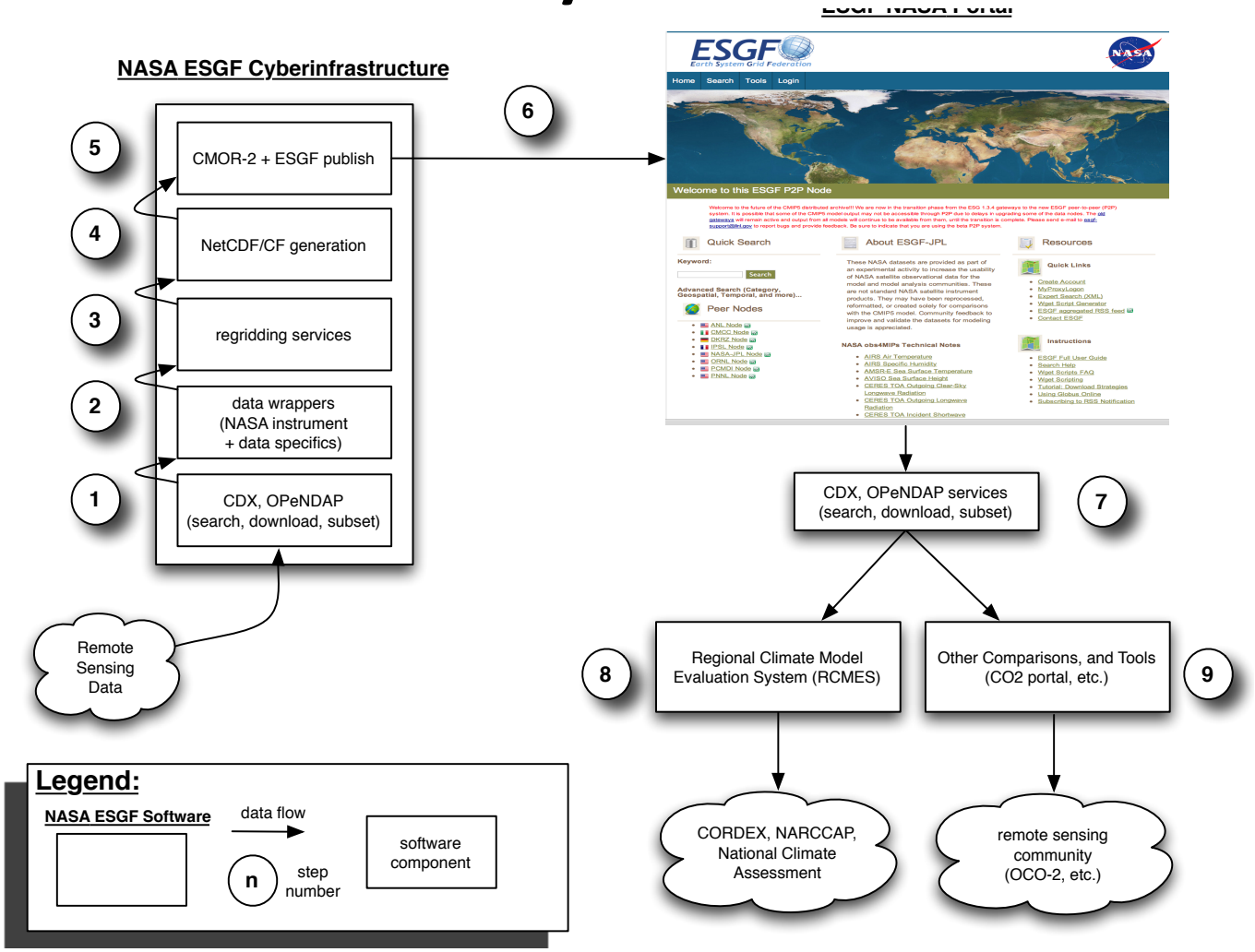


CORDEX (Africa, East Asia, Arctic), NCA, NARCCAP, NSF exArch



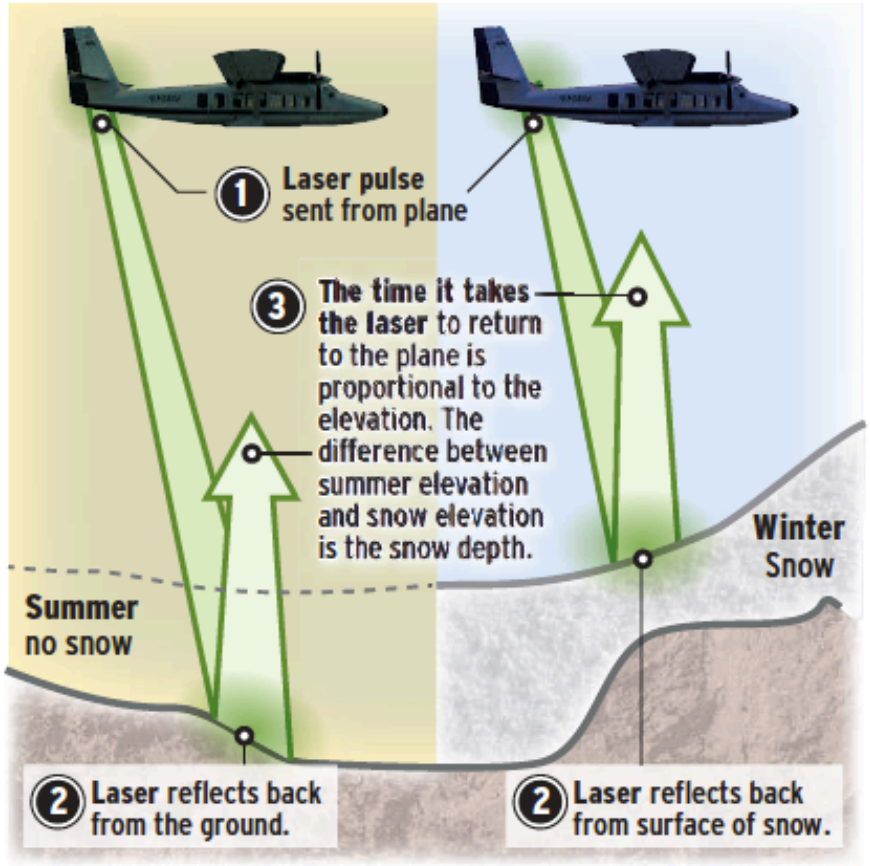
CMAC-4: Open Climate Workbench

NASA ESGF CyberInfastructure



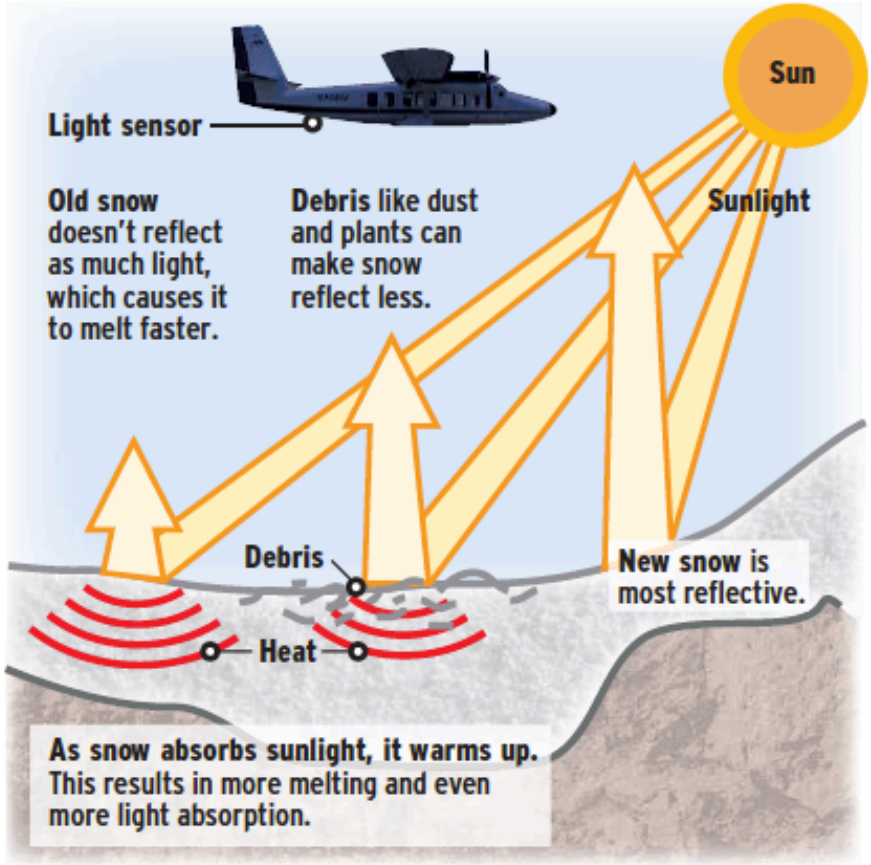
How much snow?

Using laser radar, known as Lidar, researchers measure the depth of snowpack in California.



How will it melt?

With an advanced light sensor, scientists measure snow's reflectivity – an indicator of how it will melt.

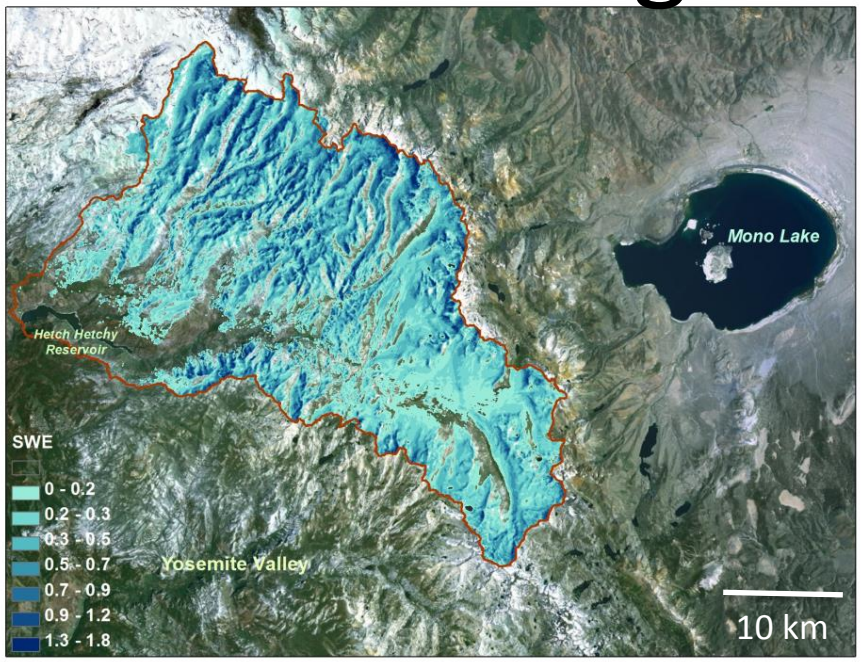


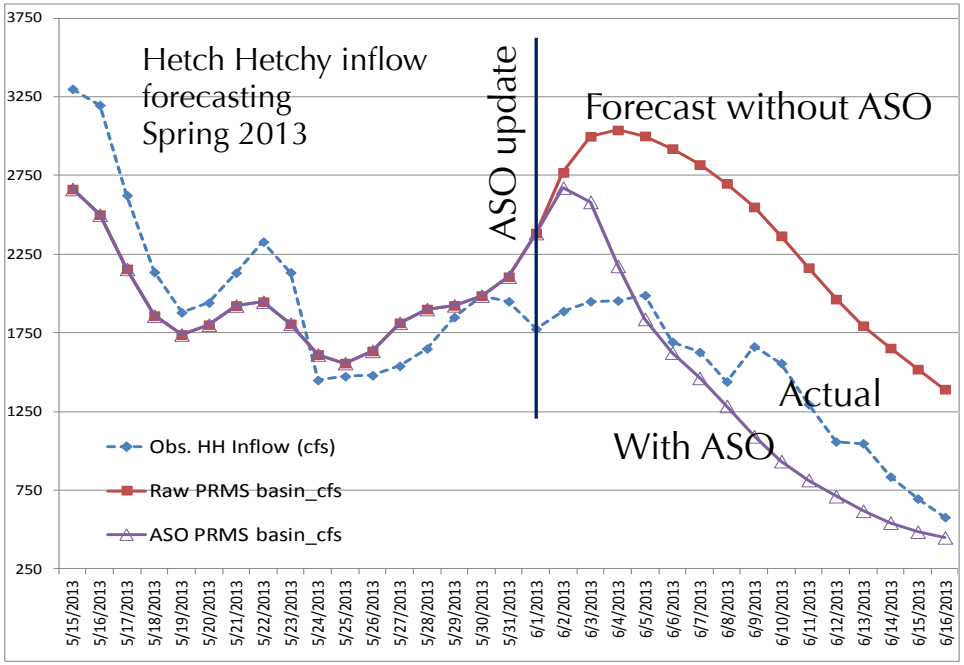
Sources: Thomas Painter, Frank Gehrke, Optech Inc.

Maxwell Henderson / The Register

Credit: Tom Painter, JPL

Improved Estimates for Water Management in California



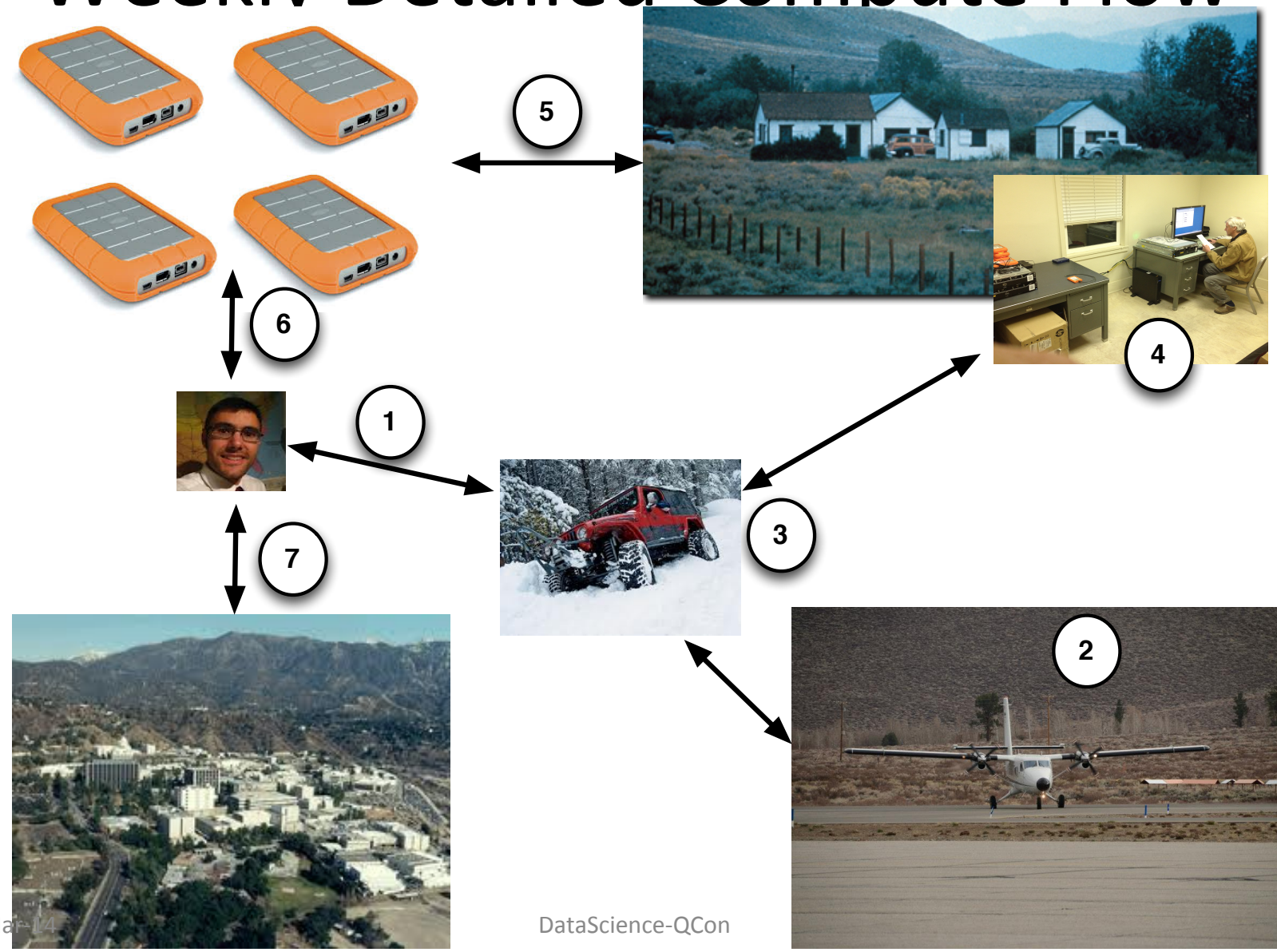


AIRBORNE SNOW OBSERVATORY

The JPL ASO team and California Dept. of Water Resources (DWR) prediction of water inflow into the Hetch Hetchy Reservoir in thousand acre feet (shown in red) was modified on June 1, 2013 based on snow water equivalent (SWE) data from the NASA/JPL Airborne Snow Observatory. The new forecast (shown in purple) provided a factor of 2 better estimate of the actual inflow (shown in blue) and enabled water managers to optimize reservoir operations in its first year.

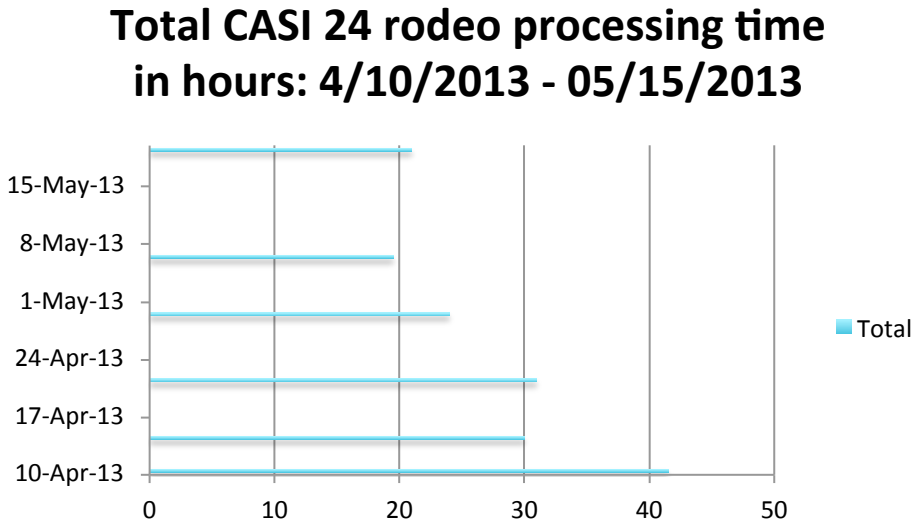
Tom Painter, JPL

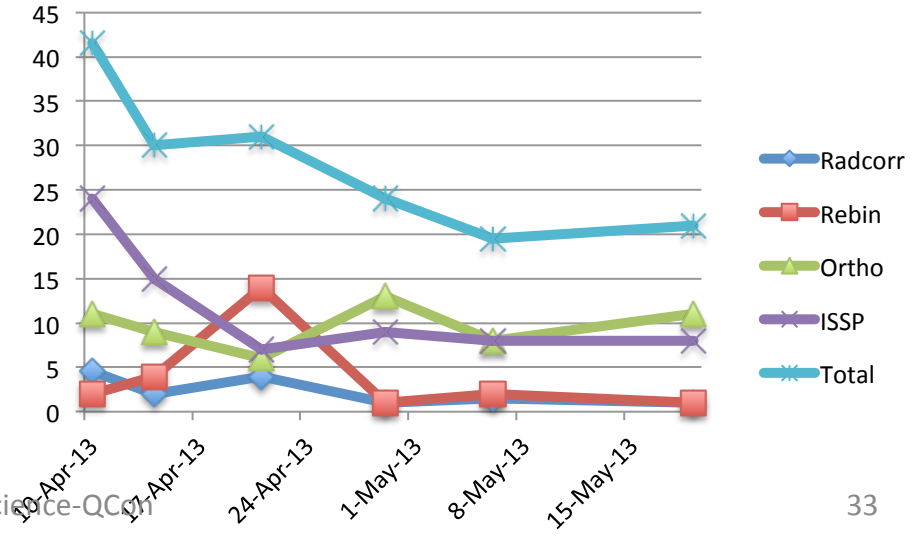
Weekly Detailed Compute Flow



Rodeo improvements over time (CASI/spectrometer)

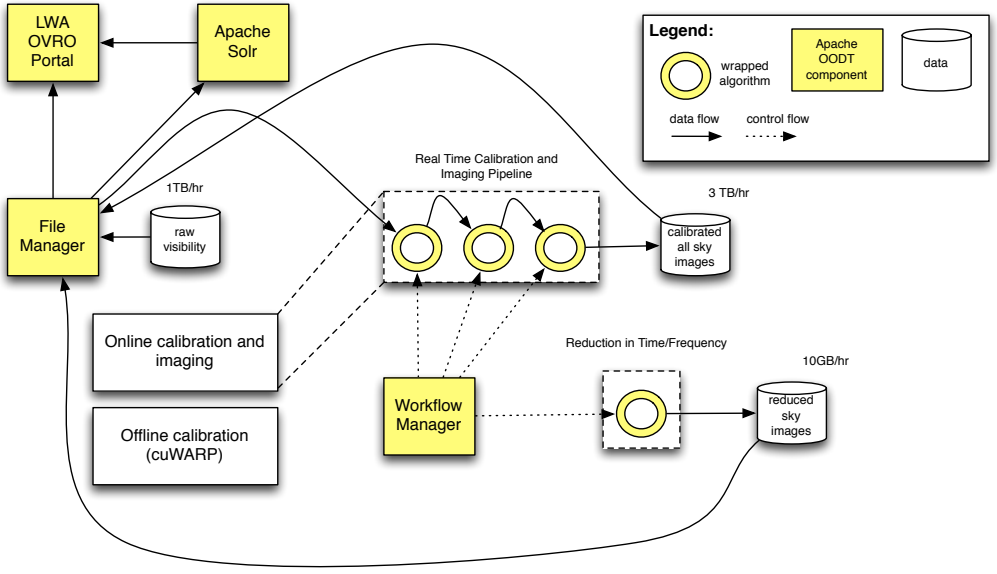
- Earlier, ISSP was dominant processing time in rodeo
  - Eventually Ortho became a problem too due to issues like flying off DEM; and/or discovery of resource contention at alg. Level
- Radcorr and Rebin processing time were equated to nil through paralllelism and automation
- Within a month of near automation, we were making 24 hrs on CASI side
- Updates to algs to make deadline





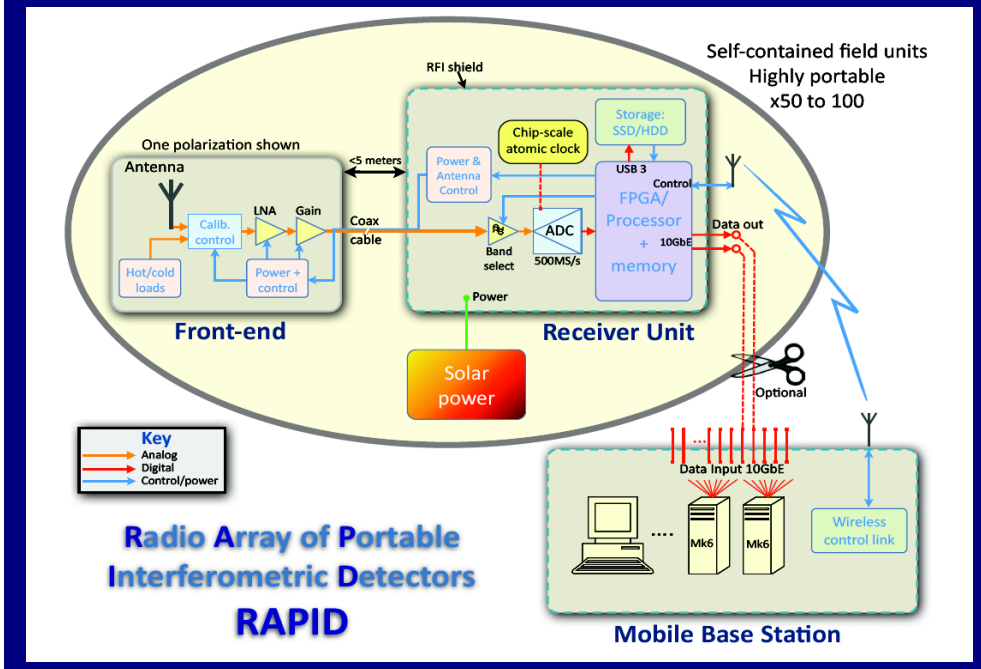
Owen's Valley Radio Observatory

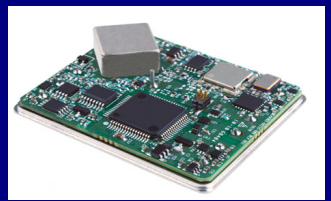
- LWA Owens Valley Radio Observatory – Probing for Cosmic Dawn
  - Joe Lazio, JPL co-PI, Gregg Hallinan, Caltech co-PI
  - Larry D'Addario, Chris Mattmann JPL co-Is
- Will lead the data management for OVRO

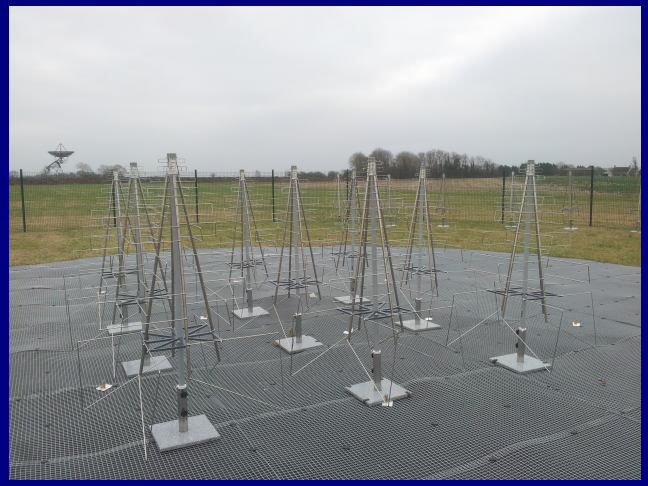


RAPID

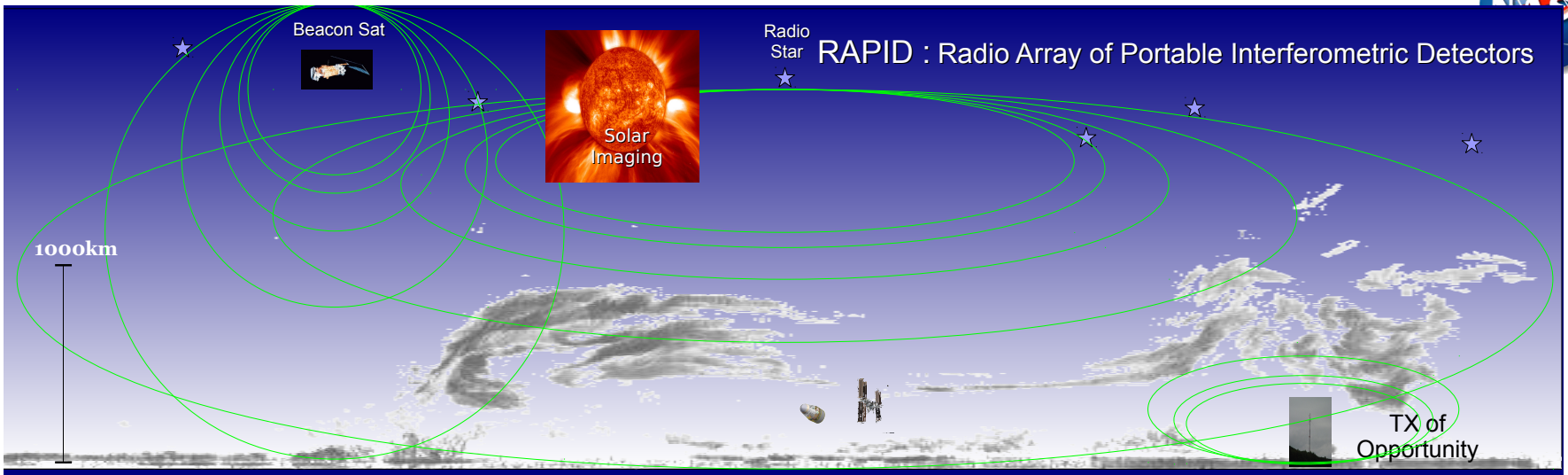
Radio Array of Portable Interferometric Detectors





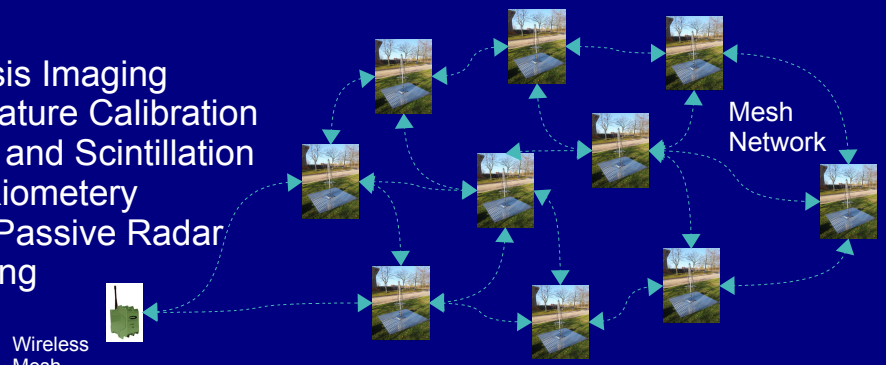


Go Where the Science is Best!
Deploy Where there is No Infrastructure
Reconfigure as Needed to Optimize Performance
Simplify by Using Raw Voltage Capture



Techniques

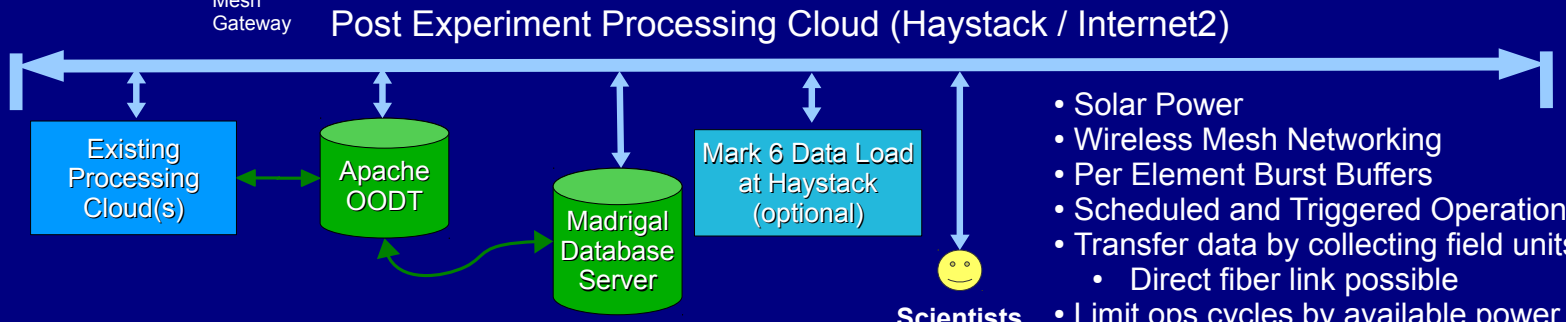
Aperture Synthesis Imaging
Absolute Temperature Calibration
Ionospheric TEC and Scintillation
Digital Imaging Riometery
Bi-static Active / Passive Radar
Spectral Monitoring



Science Targets

Solar Imaging
Galactic Synchrotron Emission
Cosmic Ray Air Showers
Ionospheric Irregularities
Ionospheric Scintillation

Go Where the Science is Best!

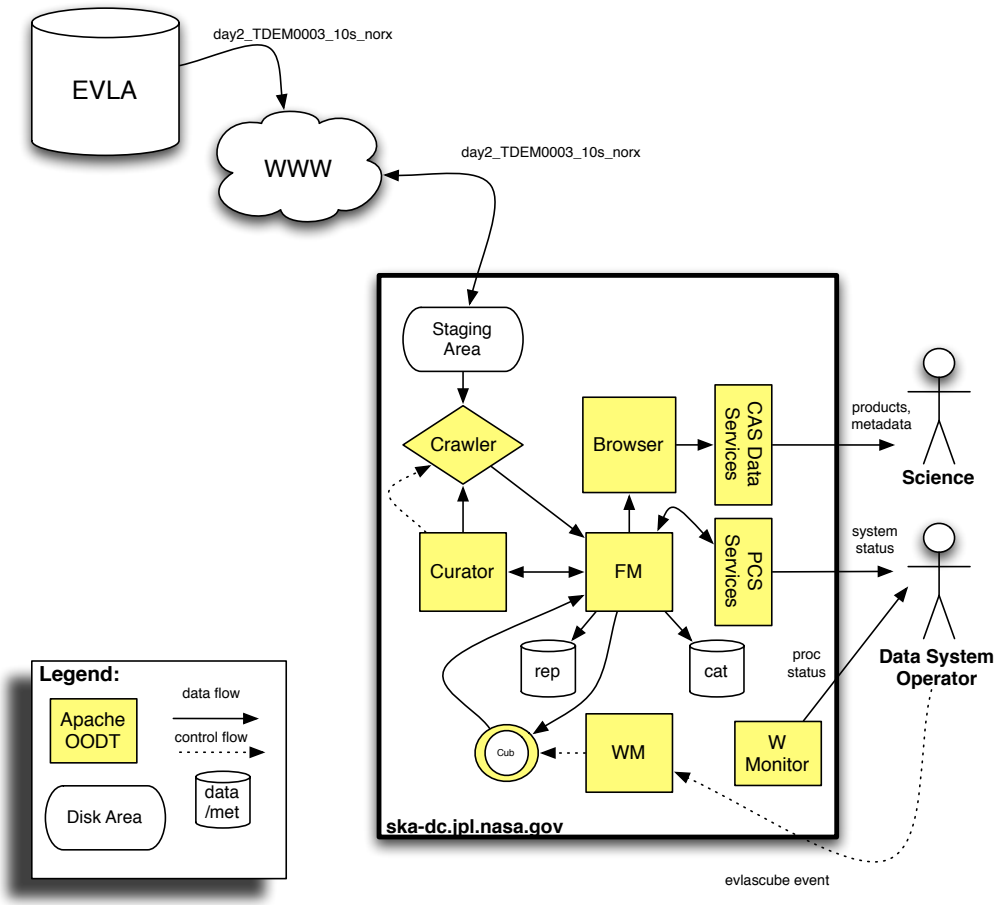


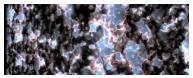
- Solar Power
- Wireless Mesh Networking
- Per Element Burst Buffers
- Scheduled and Triggered Operations
- Transfer data by collecting field units
  - Direct fiber link possible
- Limit ops cycles by available power
- Offline processing (post deployment)

U.S. National Radio Astronomy Observatory (NRAO)

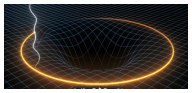
- Explore JPL data system expertise
  - Leverage Apache OODT
  - Leverage architecture experience
  - Build on NRAO Socorro F2F given in April 2011 and Innovations in Data-Intensive Astronomy meeting in May 2011
- Define achievable prototype
  - Focus on EVLA summer school pipeline
    - Heavy focus on CASApy, simple pipelining, metadata extraction, archiving of directory-based products
    - Ideal for OODT system

SKA/NRAO Architecture



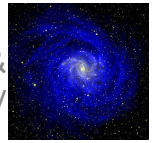


Emerging from the Dark Ages & the Epoch of Reionization

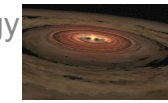


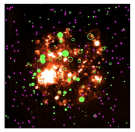
Strong-field Tests of Gravity with Pulsars and Black Holes

Galaxy Evolution, Cosmology, & Dark Energy



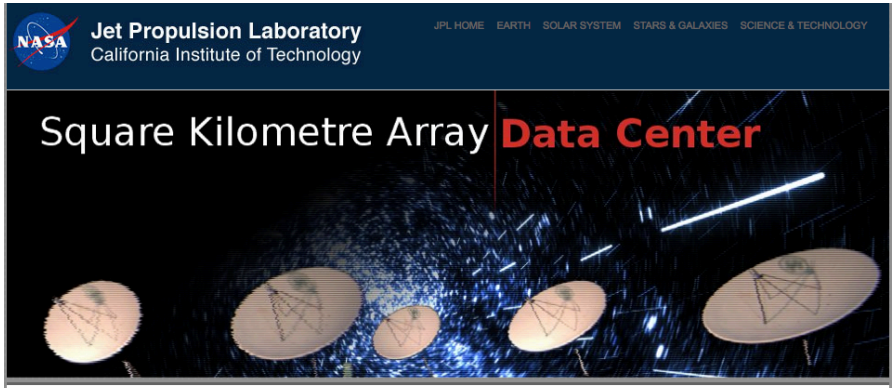
The Cradle of Life & Astrobiology

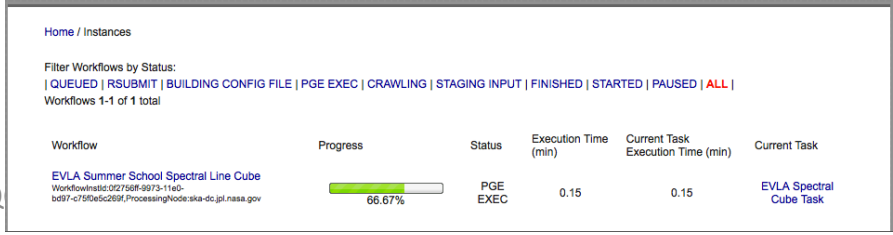




Origin & Evolution of Cosmic Magnetism

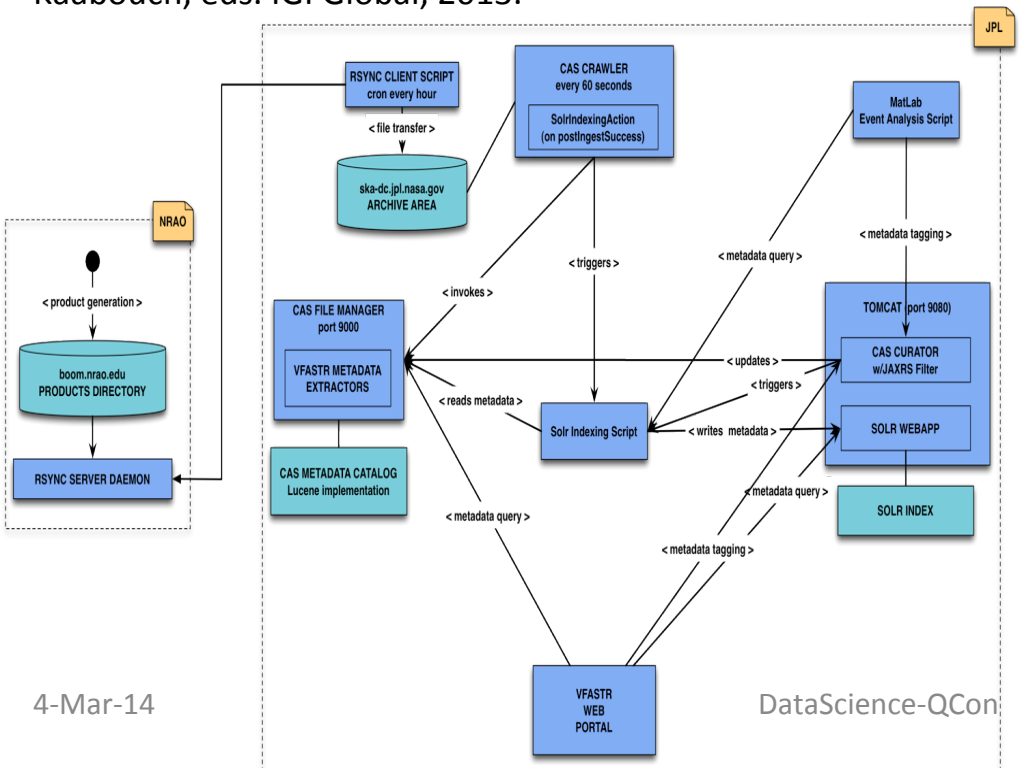
Exploring the Universe with the world's largest radio telescope

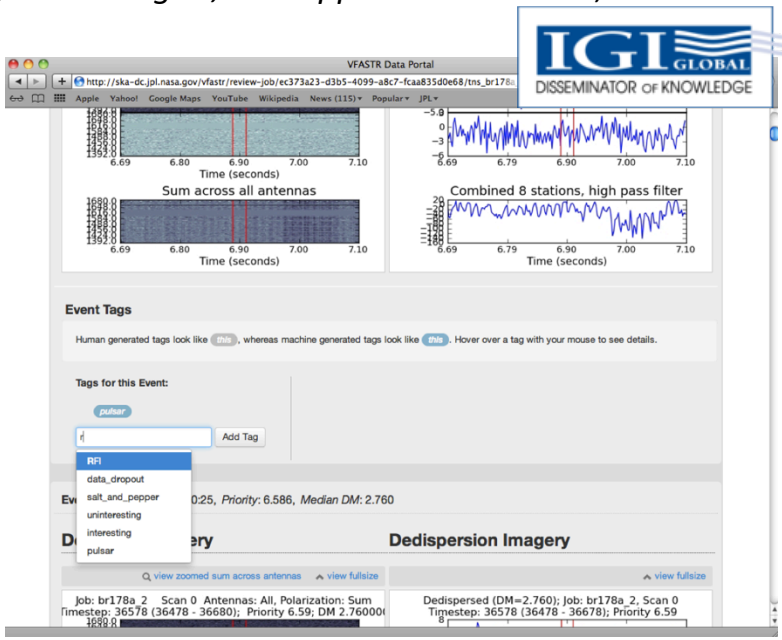




Fast Radio Transients

- VFASTR Transient Event Collaborative Review Portal – collaboration with Wagstaff/ Thompson
  - Web-based platform for easy and timely review of candidate events
  - Automatic identification of interesting events by a self-trained machine agent
  - **Demonstrates rapid science algorithm integration**
    - C. Mattmann, A. Hart, L. Cinquini, J. Lazio, S. Khudikyan, D. Jones, R. Preston, T. Bennett, B. Butler, D. Harland, K. Cummings, B. Glendenning, J. Kern, J. Robnett. Scalable Data Mining, Archiving and Big Data Management for the Next Generation Astronomical Telescopes. *Big Data Management, Technologies, and Applications*. W. Hu, N. Kaabouch, eds. IGI Global, 2013.

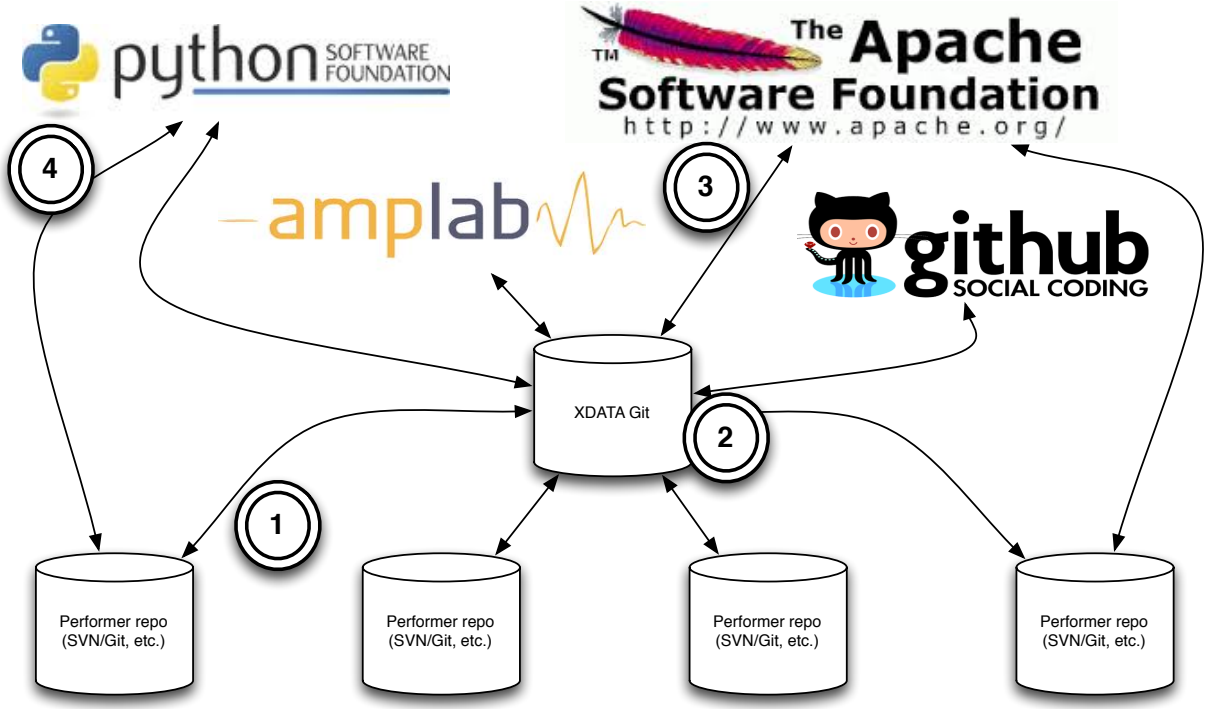




DARPA XDATA: HOW IS OPEN SOURCE PLAYING INTO THE INFRA?

Scenario 1

Performers dev code in their own local repos and occasionally “push” upstream and “pull” downstream



Scenario 2

Canonical repo is XNET Git repo

Scenario 3

Code can be “pushed” upstream to foundations and gold sources of code (and “pulled” periodically)

Scenario 4

Code can be “pushed” & “pulled” from performer repo to foundation

Data Scientist Training

- Supervised 4 current postdocs (co-supervise with Waliser and Painter) and 1 current PhD in Atmospheric Sciences (Whitehall)
  - *What am I doing on her dissertation committee?*
- USC and UCLA in-flow and outflow
  - Hired ~10 USC PhD and MS students at JPL
  - Attracting more all the time
  - Fielding them in courses at USC in Search Engines/Big Data, and in Software Architecture
  - $$$ at USC and UCLA from NSF to flow in and out

NASA: where from here?

- Agency framework for open source: we can't do it all on our own
- Strategic investments/opportunities
  - Rapid Science algorithm integration
    - Needed by next gen missions (Space/airborne), e.g., ASO, needed by next gen astronomical archives e.g., SKA, and existing NRAO work, and collaborations (MIT)
  - Smart data movement
    - Needed by next gen missions, climate science work (RCMES, ESGF), and next gen astronomy data transfer (S. Africa to US)
  - Transient/Persistent archives (Science Computing Facilities)
    - How do we get it done for cheaper, tear it down, stand it up quickly
  - Automatic text/metadata extraction from file formats
    - There will never be a "god" format, so we need Babel Fish – needed by ASO, RCMES, SKA, XDATA
- More data scientist training

Thank you!

chris.a.mattmann@nasa.gov
@chrismattmann/Twitter
http://sunset.usc.edu/~mattmann/

QCon

Please evaluate my talk via the mobile app!