
# NASA and Apache Spark

Chris Mattmann

*Chief Architect, Instrument and Science Data Systems Section, NASA JPL*

*Adjunct Associate Professor, USC*

*Director, Apache Software Foundation*

Spark Summit 2015

# And you are?

- Chief Architect, Instrument and Science Data Systems Section at NASA JPL in Pasadena, CA USA
- Software Architecture/ Engineering Prof at Univ. of Southern California

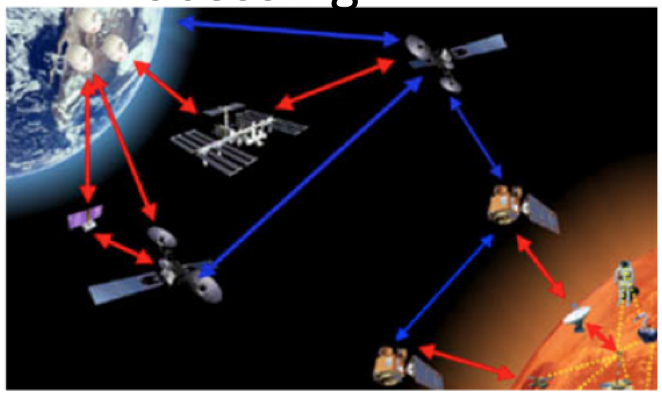
- Apache Board of Directors involved in
  - OODT (VP, PMC), Tika (PMC), Nutch (PMC), Incubator (PMC), SIS (PMC), Gora (PMC), Airavata (PMC)

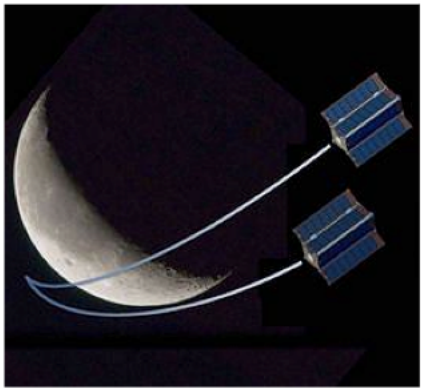

I work here

# Instrument & Ground Data Systems

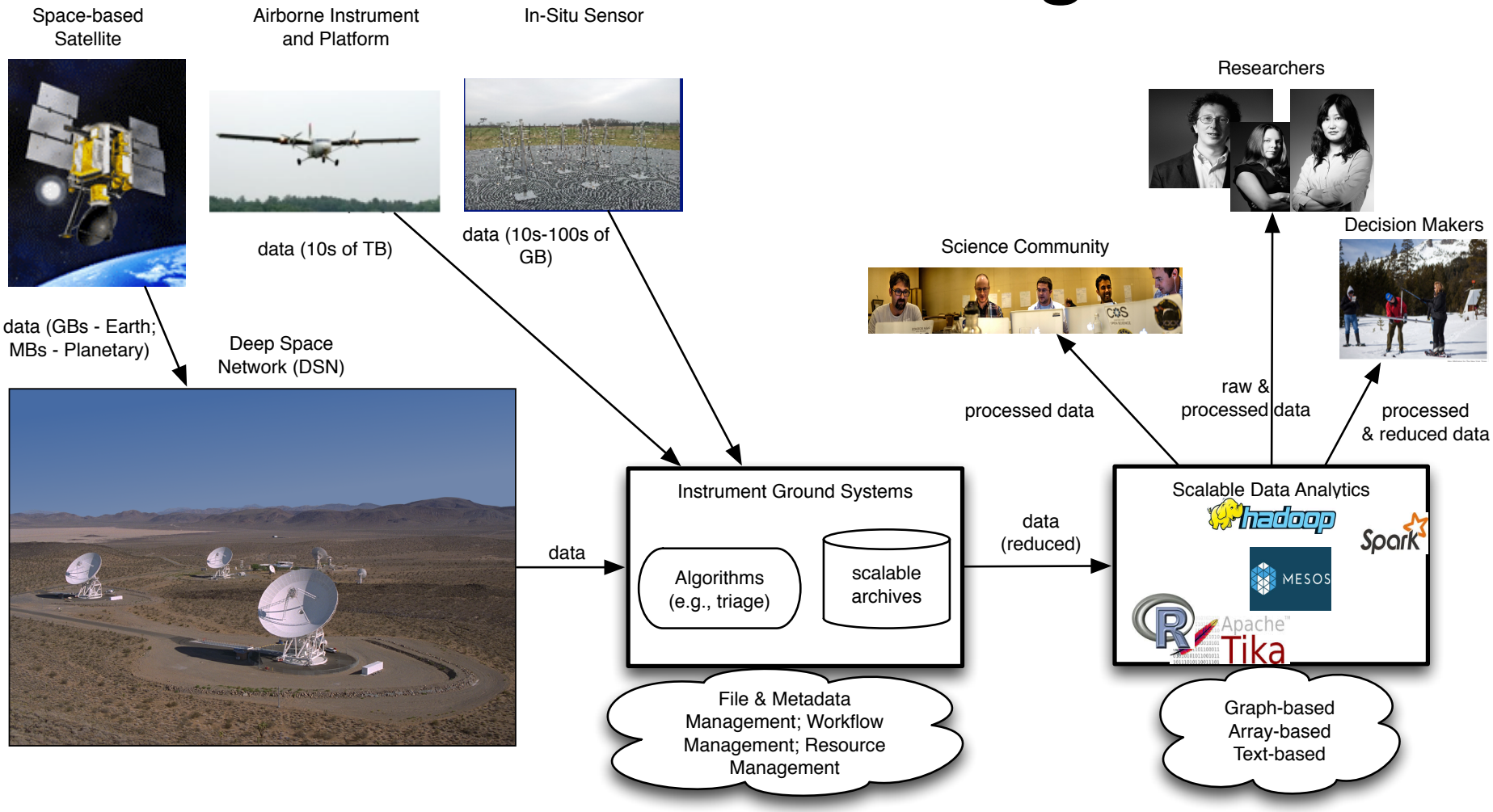
*(Section 398)*

- Largest Section on Lab
- 250+ people
- Data Science, Machine Learning, Visualization, Operations groups
- OCO-2, NPP Sounder PEATE, SMAP, MER, MSL, Mars 2020, Image Processing

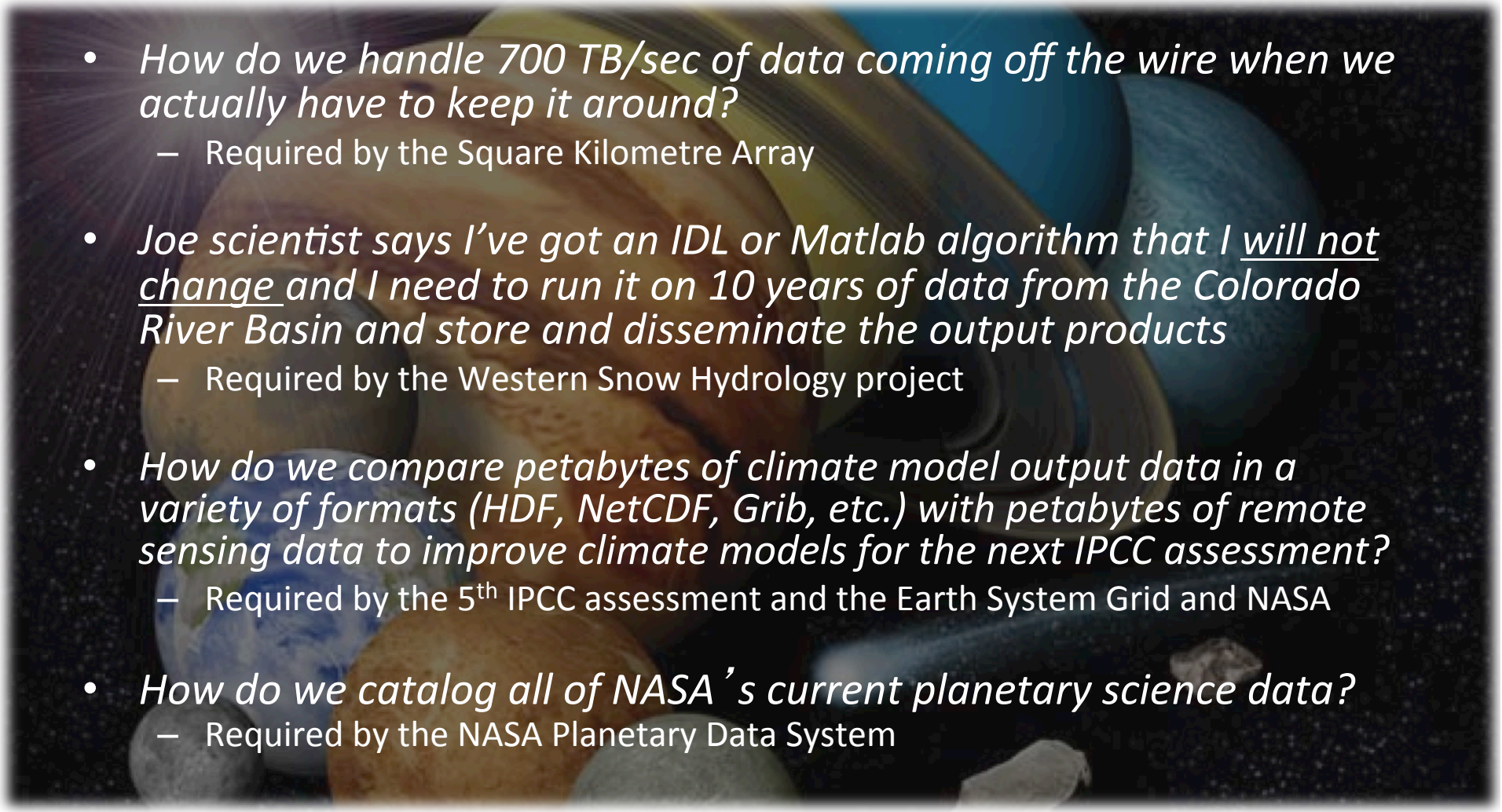




# Instrument and Ground Systems: Earth Monitoring



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

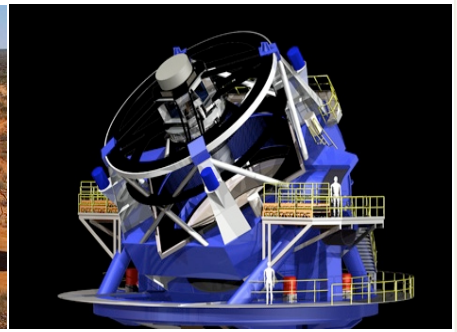


- *How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?*
  - Required by the Square Kilometre Array
- *Joe scientist says I've got an IDL or Matlab algorithm that I will not change and I need to run it on 10 years of data from the Colorado River Basin and store 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: Chris Mattmann**
**Steering Committee Leader: Joseph Lazio**

| 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

# The Square Kilometre Array


Credit: Andrew Hart

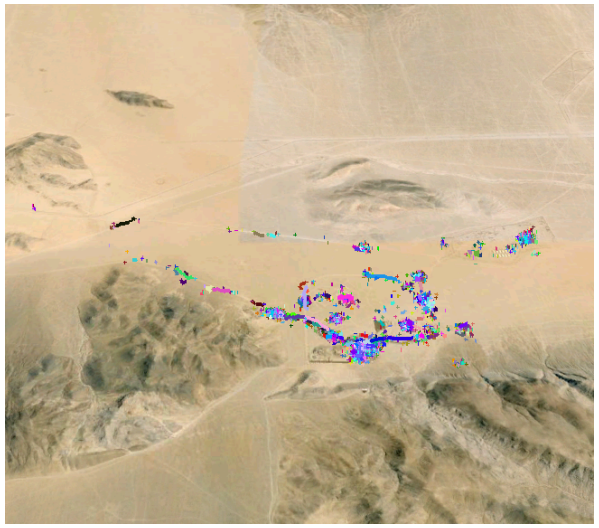


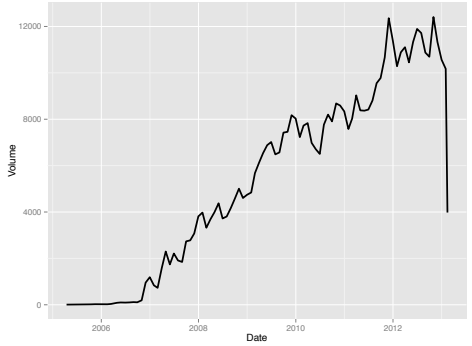
# How did I get involved with Spark?

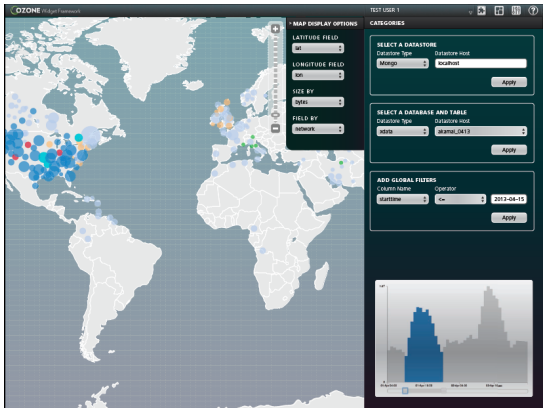
# DARPA XDATA

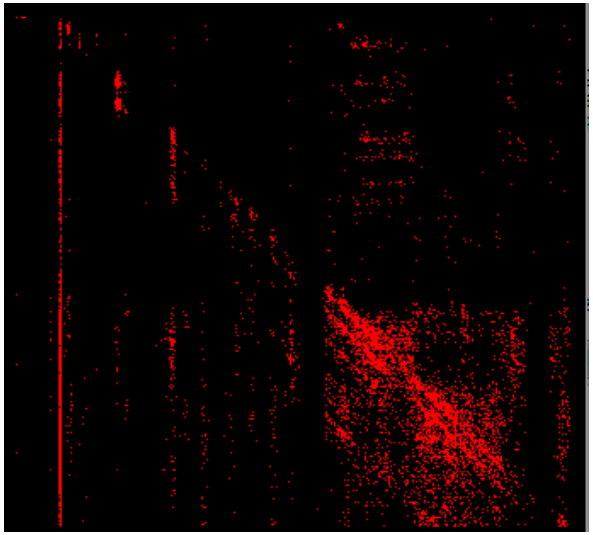
- Grab the principals behind the leading infrastructure/viz technologies
  - Shove them in a tight space
  - Provide ~~beer~~ coffee and snacks
  - Provide awesome data and challenges
  - Provide infrastructure and connectivity
- Check in every day and 1x a week
- Wall of Shame/Fame
- New Challenges Each Week
- Midterm Presentations
  - Peanut Gallery
- Make people talk/socialize
- Put that all together

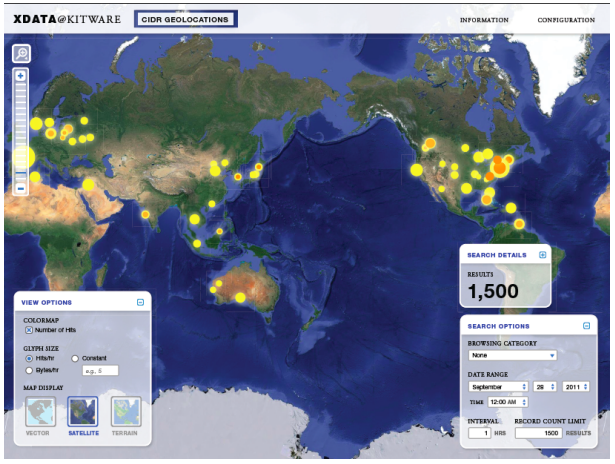
# DARPA XDATA: Analytics + Viz





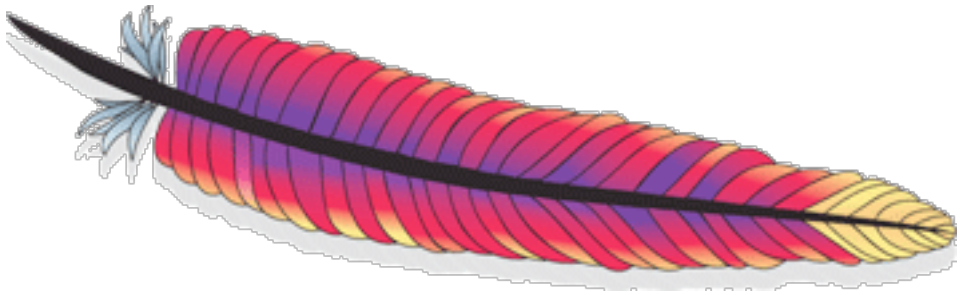


# Met these fine people

- Ion Stoica
  CEO, DataBricks
  Co-Director,
  AMP Lab
- Matt Massie
  Dev Manager,
  AMP Lab
- Dr. Chris White, DARPA
  XDATA PM


# The Apache Software Foundation

- Largest open source software development entity in the world
  - Over 2600+ committers
  - Over 4200+ contributors
  - Over 400+ members
- 100+ Top Level Projects
  - 57 Incubating
  - 32 Lab Projects
- 12 retired projects in the “Attic”
- Over 1.2 *million* revisions
- 501(c)3 non-profit organization incorporated in Delaware

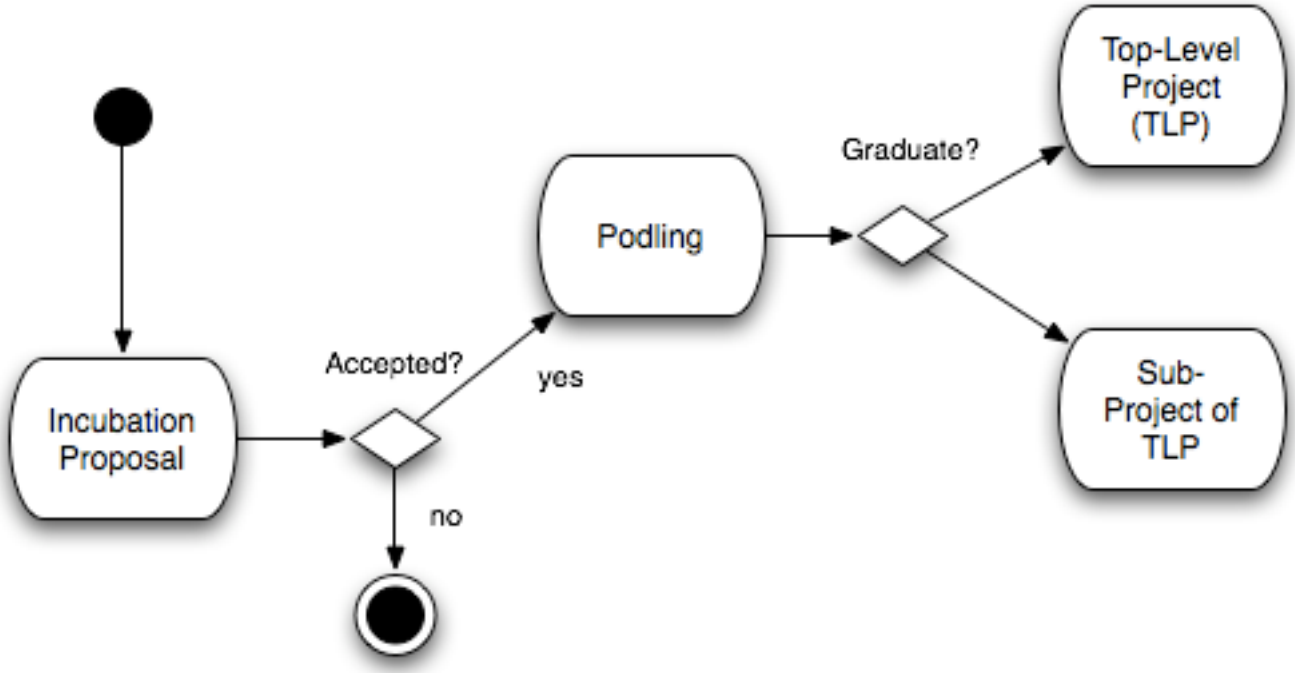


*-Over 10M successful requests served a day across the world*

*-HTTPD web server used on 100+ million web sites (52+% of the market)*

# Apache Maturity Model

- Start out with Incubation
- Grow community
- Make releases
- Gain interest
- Diversify




- When the project is ready, graduate into
  - Top-Level Project (TLP)
  - Sub-project of TLP
- Increasingly, Sub-projects are discouraged compared to TLPs

# Apache is a well recognized brand



GOVERNOR ARNOLD SCHWARZENEGGER

November 5, 2009

***Apache Software Foundation***

It is a great pleasure to extend my greetings to all those attending ApacheCon and congratulations on your tenth anniversary.

I applaud your incredible work over the past decade and appreciate you choosing California as the place to celebrate this fantastic milestone. Our state is a land of innovation, and you have likewise fostered great technological advancements that have touched the lives of millions of people around the world.

Whether managing financial systems, positioning satellites or powering websites through the Apache HTTP Server, your open source projects play key roles in making our information age possible. Thank you for your extraordinary accomplishments and commitment to discovery.

On behalf of all Californians, I send my gratitude to everyone in attendance for your participation, and I offer my best wishes for a rewarding conference and continued success.

Sincerely,

Arnold Schwarzenegger

STATE CAPITOL · SACRAMENTO, CALIFORNIA 95814 · (916) 445-2841



# Why Spark and NASA?

# Where does Spark fit into science?

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

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



# NASA Science & Architecture


OSTM/Jason 2
(NOAA)
Aquarius
QuikSCAT
TRMM
Terra
ISS-RapidScat,
CATS
EO-1
Landsat 7
(USGS)
SMAP
Aqua
Suomi NPP
(NOAA)
Landsat 8
(USGS)
SORCE,
TCTE (NOAA)
Aura
GPM
GRACE (2)
CALIPSO
CloudSat
OCO-2



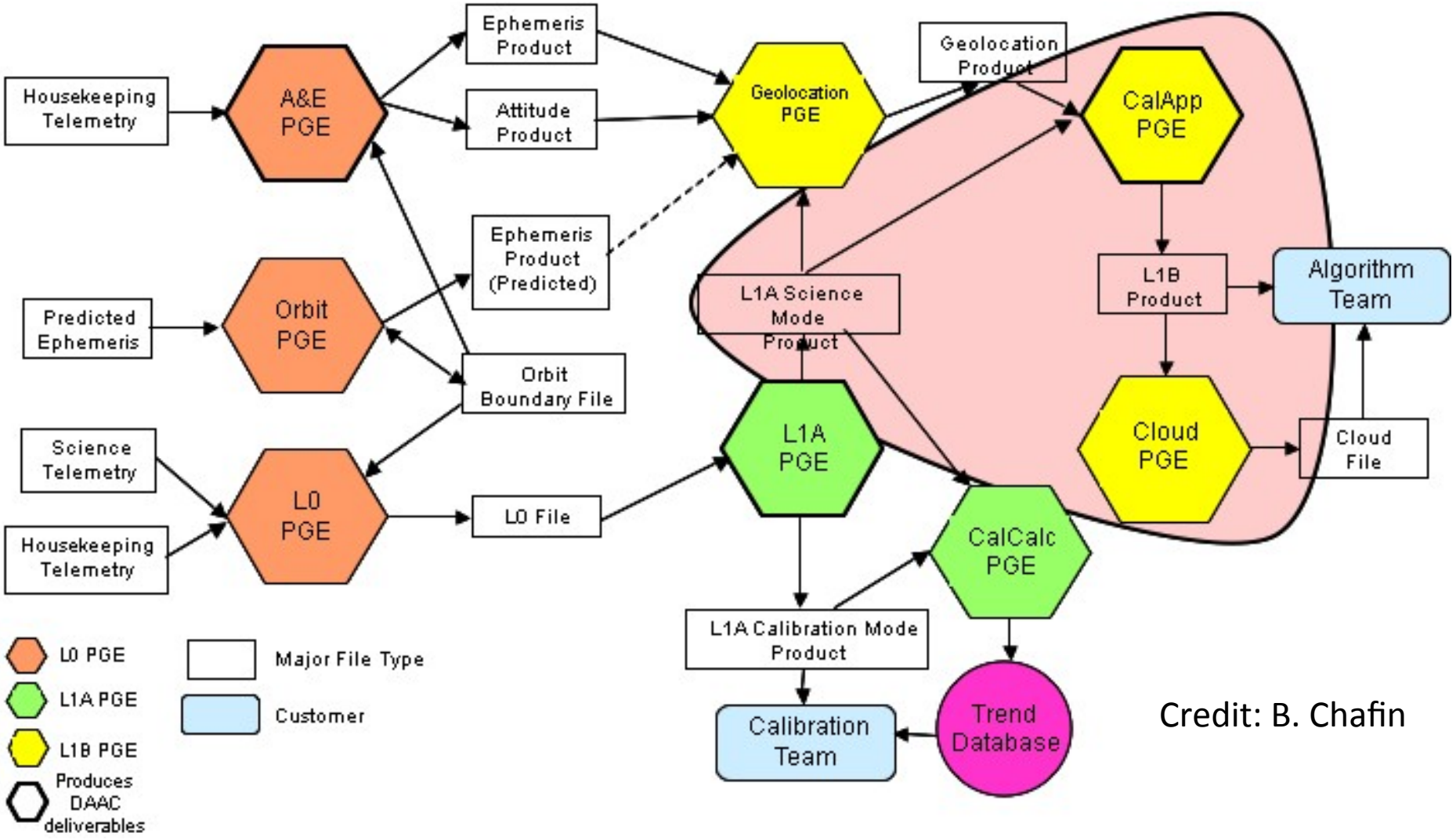
# Science Data File Formats

- 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

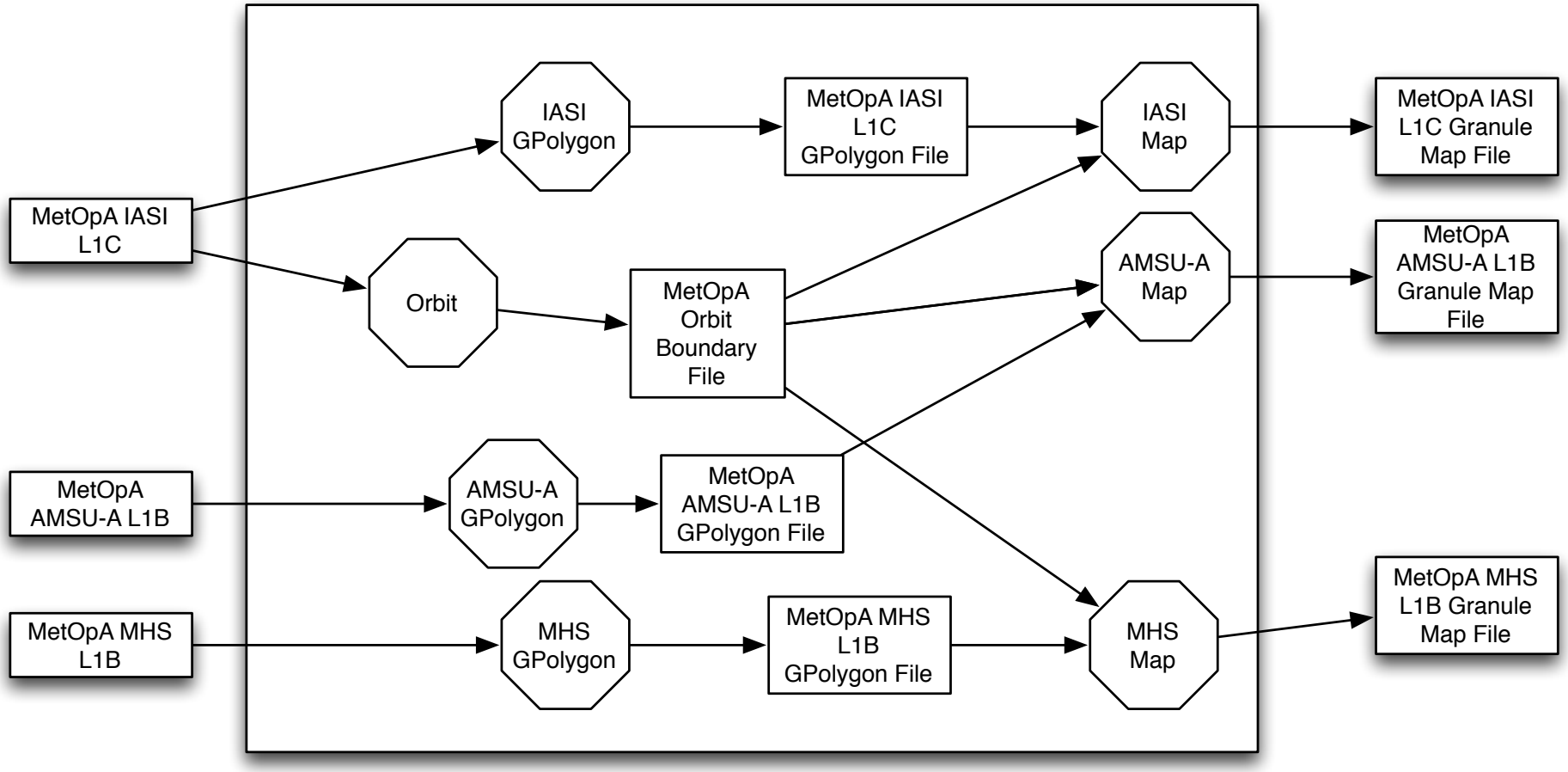
- network Common Data Form (netCDF)
  - www.unidata.ucar.edu/software/**netcdf**/
  - Versions 3 and 4
  - Heavily used in DOE, NOAA, etc.
  - Encapsulates
    - Observation (Scalars, Vectors, Matrices, NxMxZ…)
    - Metadata (Summary info, date/time ranges, spatial ranges)
  - Custom readers/writers/APIs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

# OCO-1 Workflow




Credit: B. Chafin

# NPP Sounder PEATE



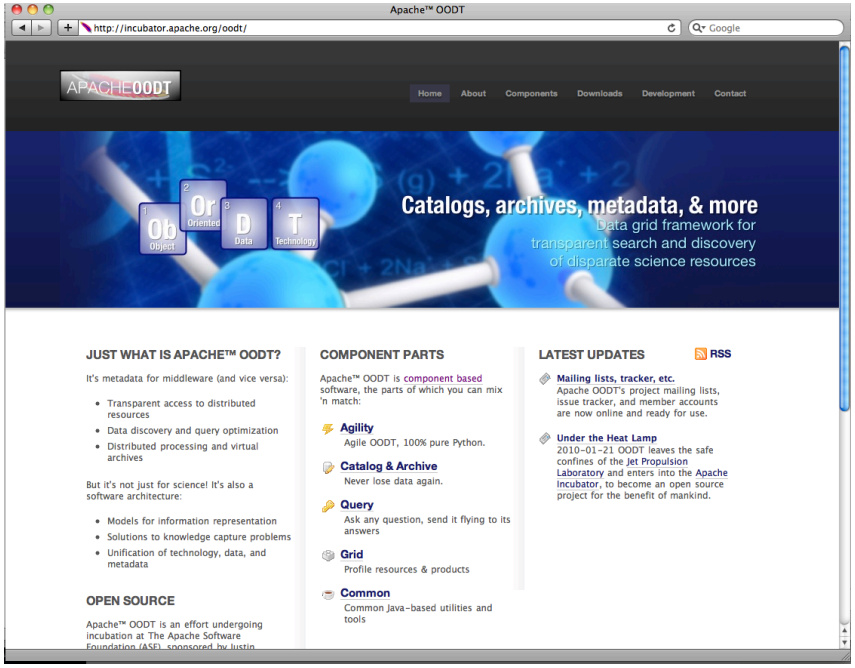

Credit: B. Foster


# Data-Reuse Between Stages

- All of these science data pipelines
  - Read/Write NetCDF, HDF files
  - Write to distributed file systems (only recently HDFS, GlusterFS, etc.)
- Have timing constraints
- Include jobs with varying timing
  - Some early completing jobs (<1ms)
  - Some long running jobs
- What does this sound like? SPARK

# Apache OODT

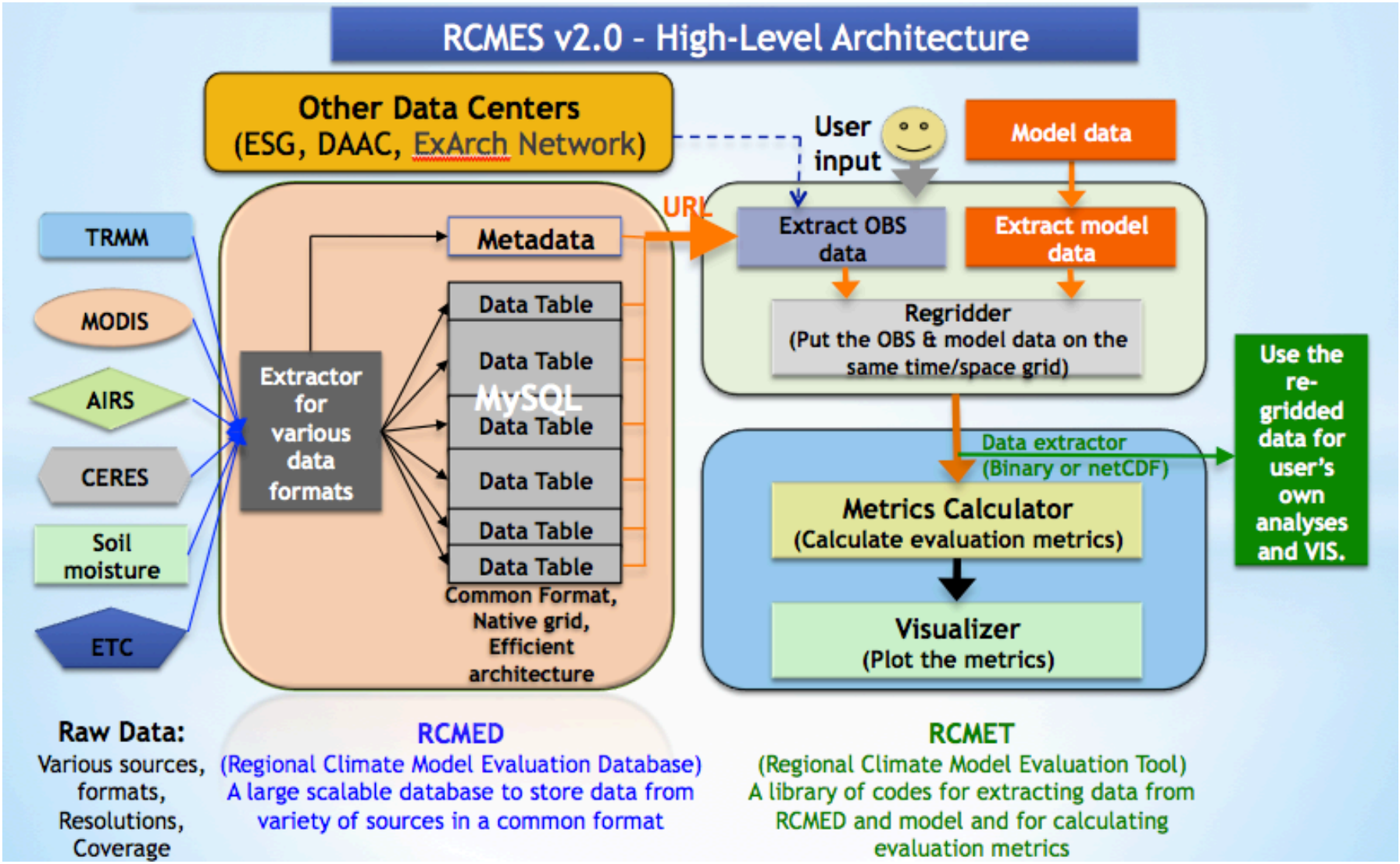
- Entered “incubation” at the Apache Software Foundation in 2010
- Selected as a top level Apache Software Foundation project in January 2011
- Developed by a community of participants from many companies, universities, and organizations
- Used for a diverse set of science data system activities in planetary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




OODT Development & user community includes:



# SciSpark

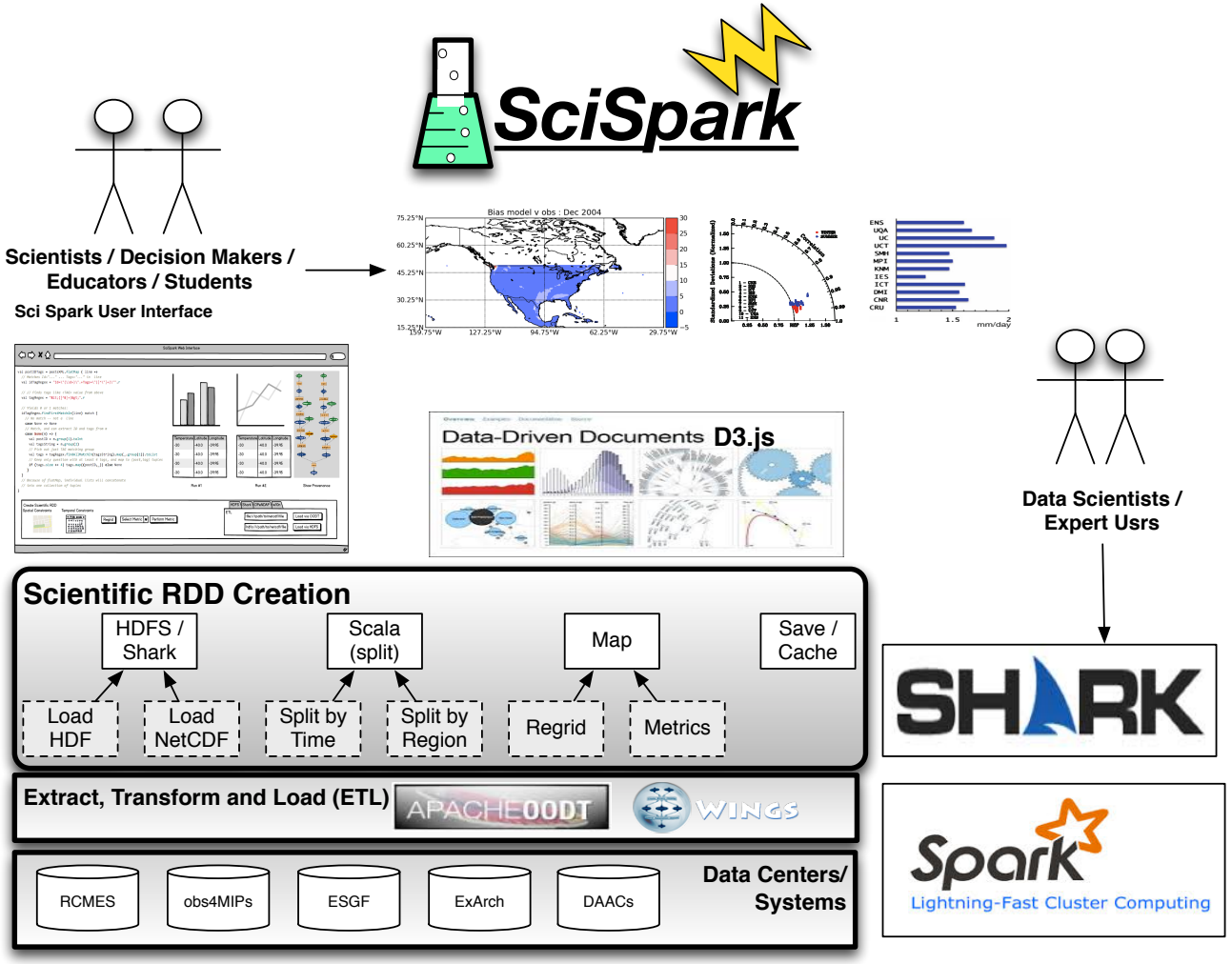

RCMES v2.0 - High-Level Architecture
Other Data Centers
(ESG, DAAC, ExArch Network)
User input
Model data
URL
TRMM
MODIS
AIRS
CERES
Soil moisture
ETC
Extractor for various data formats
Metadata
Data Table
Data Table
MySQL
Data Table
Data Table
Data Table
Data Table
Common Format, Native grid, Efficient architecture
Extract OBS data
Extract model data
Regridder
(Put the OBS & model data on the same time/space grid)
Data extractor
(Binary or netCDF)
Metrics Calculator
(Calculate evaluation metrics)
Visualizer
(Plot the metrics)
Use the re-gridded data for user's own analyses and VIS.
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



# Motivation for SciSpark

- Experiment with *in memory and frequent data reuse* operations
  - Regridding, Interactive Analytics such as MCC search, and variable clustering (min/max) over decadal datasets could benefit from in-memory testing (rather than frequent disk I/O)
  - Data Ingestion (preparation, formatting)

# Architecture of *SciSpark*



# Sci Spark – Visualization (D3 and friends)

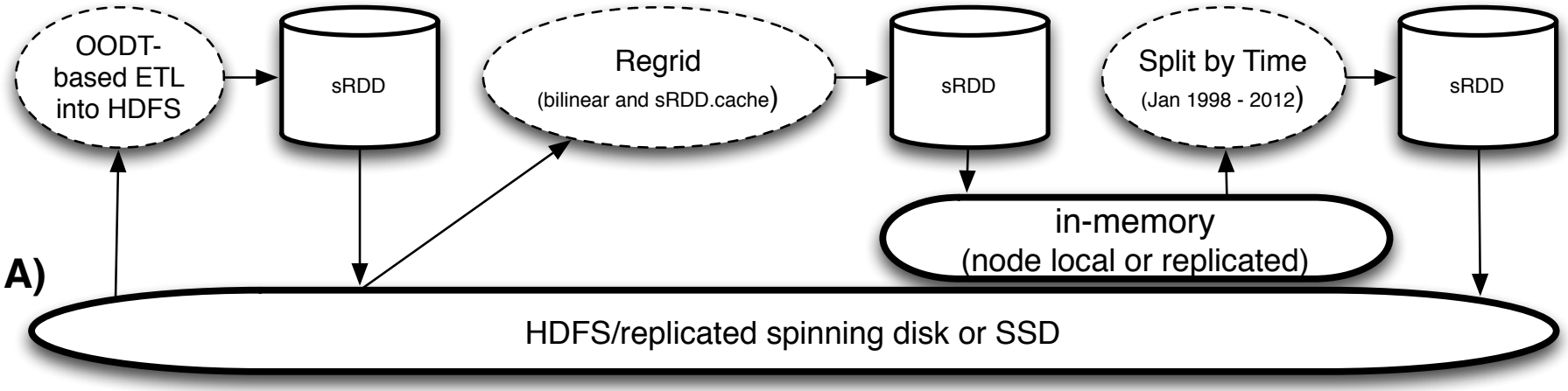
- Tika, Nutch, Blaze, Bokeh, Solr, Tangelo

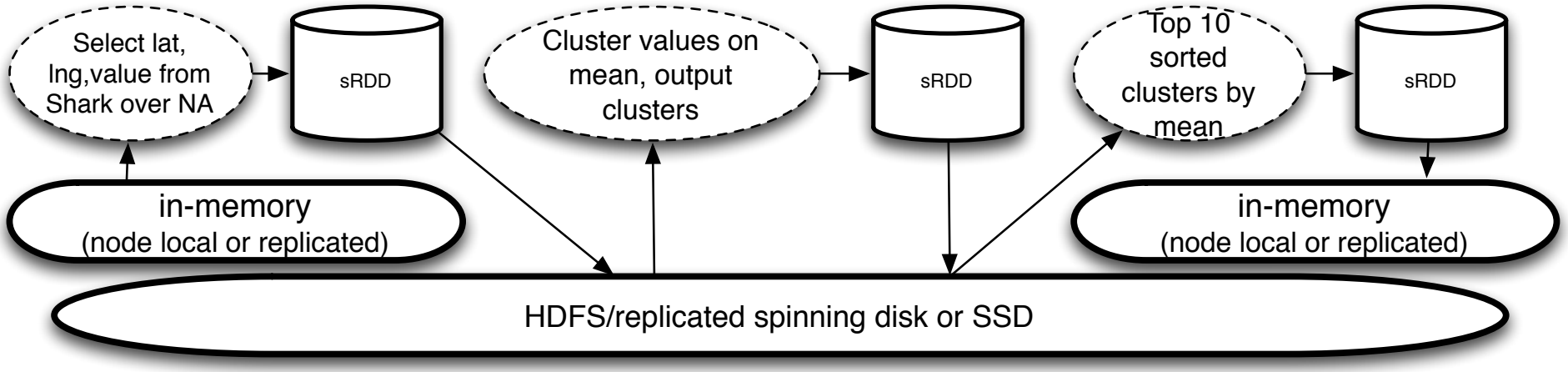

# Use Cases

- (A) Multi-stage generation to generate time-split data

- (B) Multi-stage operation to select data from Shark, and cluster by deviation from mean

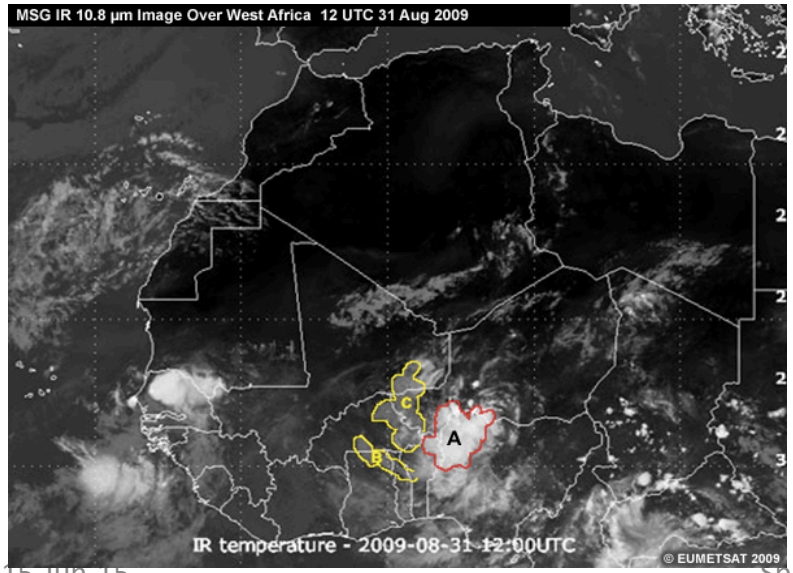
# Climate Metrics on SciSpark

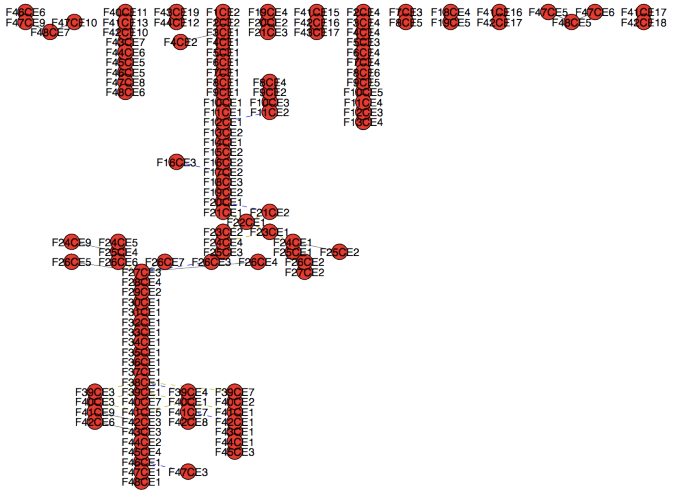


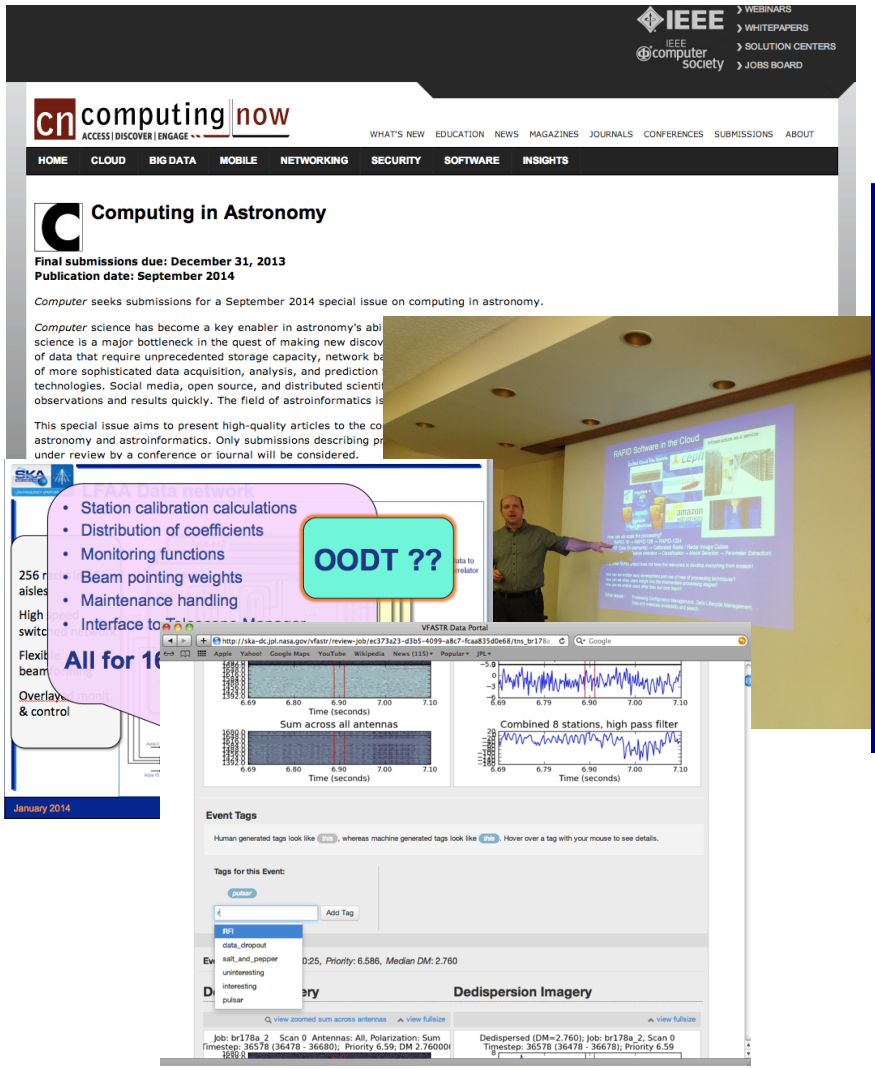
# SciSpark – just getting started

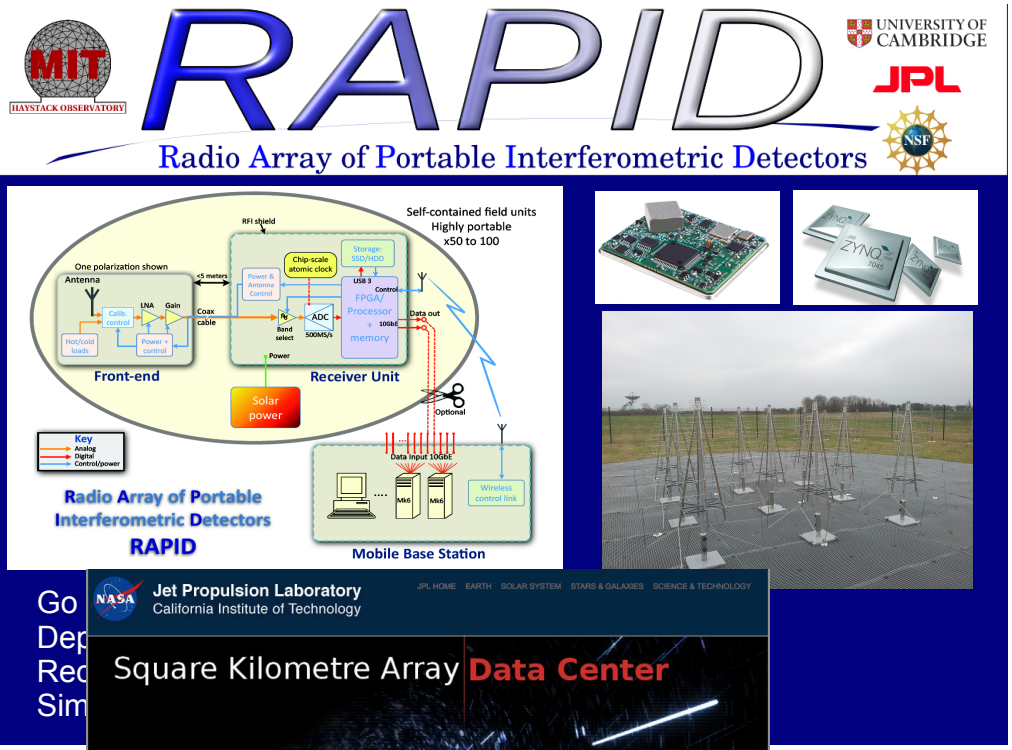
- Funded NASA AIST14 award to construct
- SciSpark  climate scenarios
  - Climate extremes / impact analysis and clustering
  - Mesocale Convective Complex Search



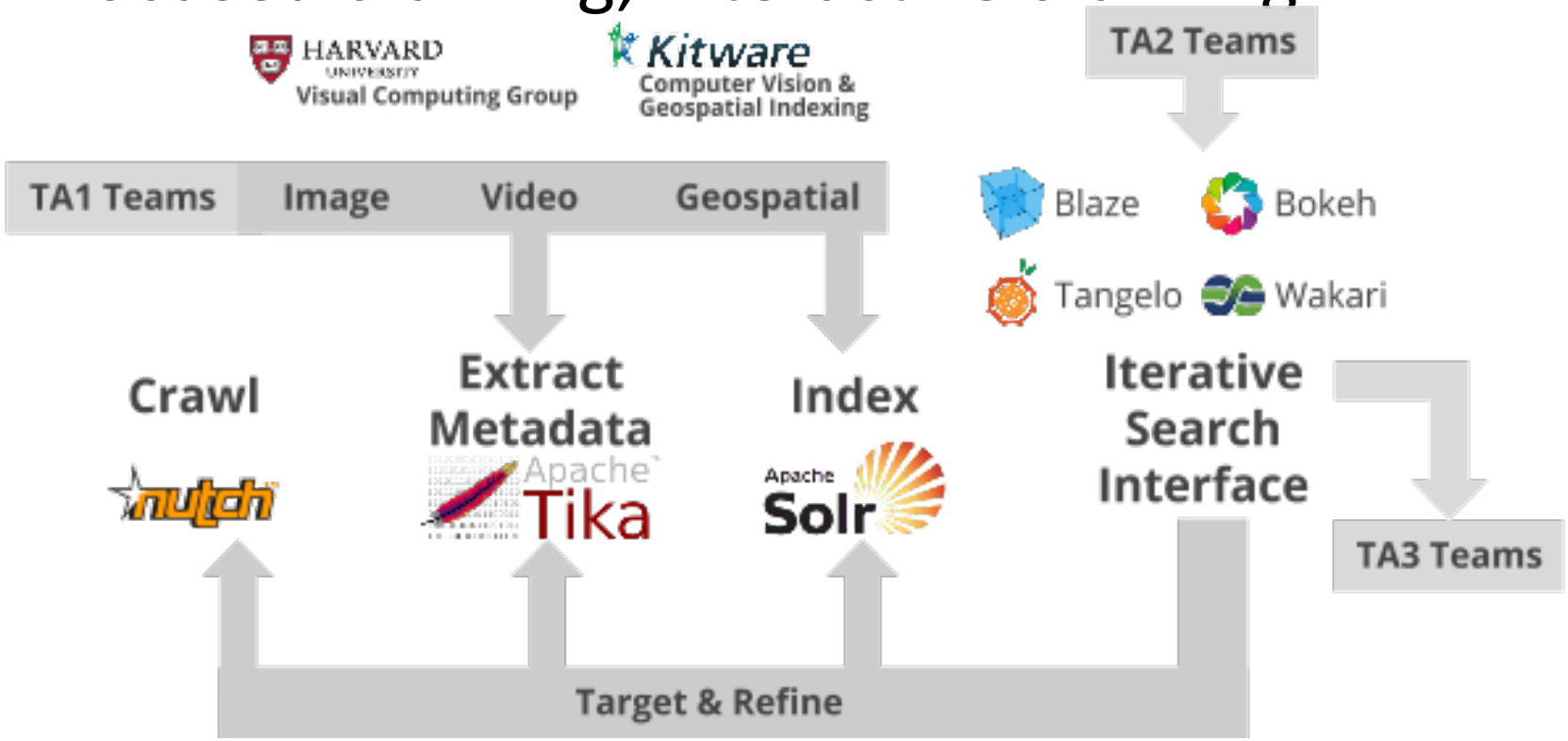
# SKA/Astronomy



# DARPA Memex

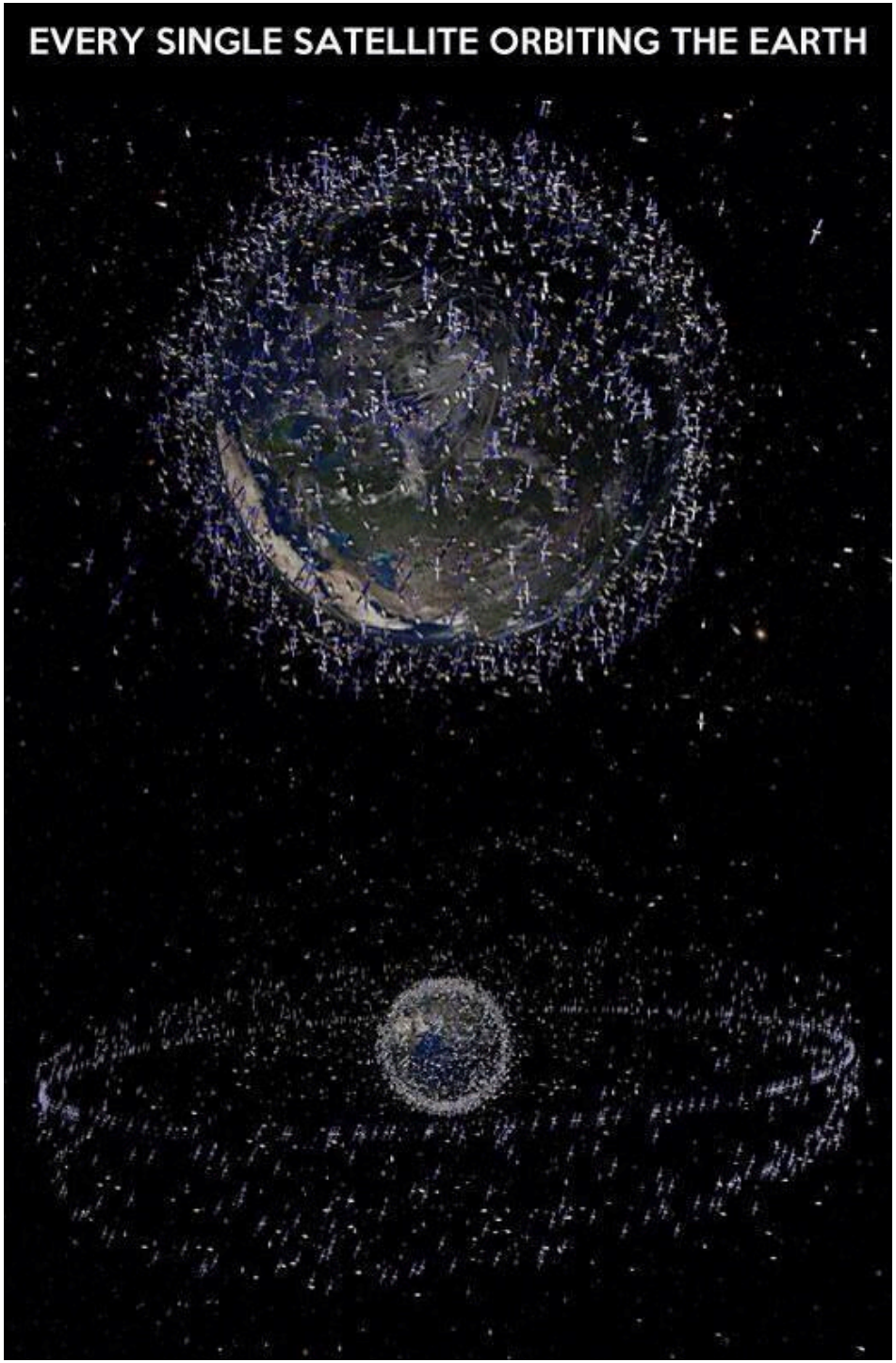
- Domain Specific search of audio/video/media
- Focused crawling; interactive crawling





# Conclusions

- Lots of places in science and NASA for Spark
- Great connections already
- Existing Apache projects to integrate upstream
- Downstream use cases
- Come chat with me today!




Credit: Vala Afshar, Extreme Networks

# Thank you!

chris.a.mattmann@nasa.gov
@chrismattmann/Twitter

http://sunset.usc.edu/~mattmann/