NASA

Science
Mission
Directorate

From Data Curation to Data Analytics: A Big Data Experiment in NASA Earth Science

Tsengdar J. Lee, Ph.D.
Weather Data Analysis Program Manager
High-End Computing Manager

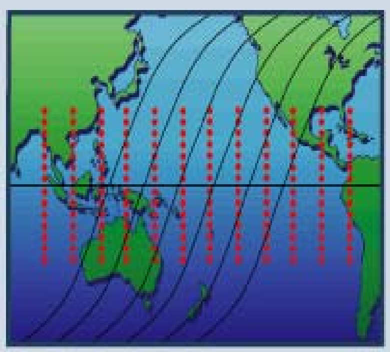
Turning Observations into Knowledge Products

Downlink Speed

Petabytes 10^{15}

Multi-platform, multiparameter, high spatial and temporal resolution, remote & in-situ sensing

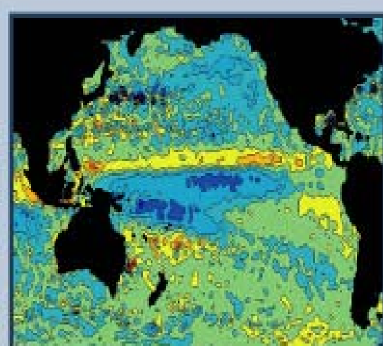
Advanced Sensors



Terabytes 10^{12}

Calibration, Transformation To Characterized Geo-physical Parameters

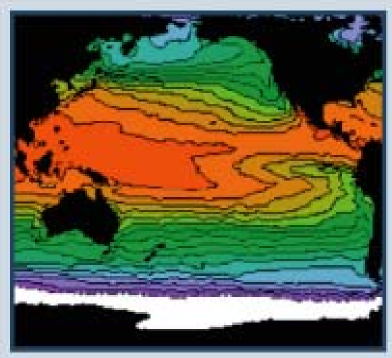
Data Processing & Analysis



Gigabytes 10^{9}

Interaction Between Modeling/Forecasting and Observation Systems

Information Synthesis

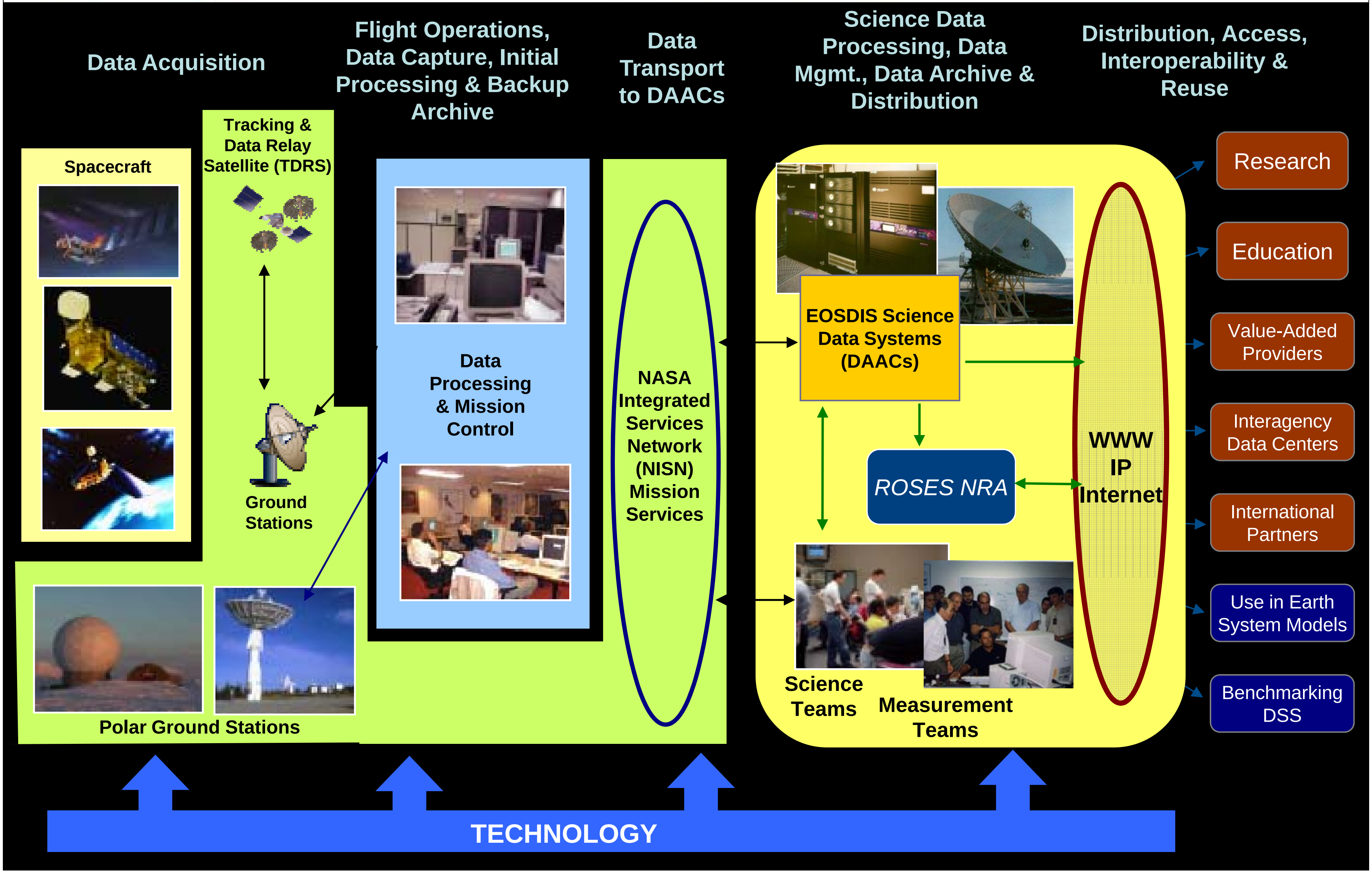


Megabytes 10^{6}

Interactive Dissemination and Predictions

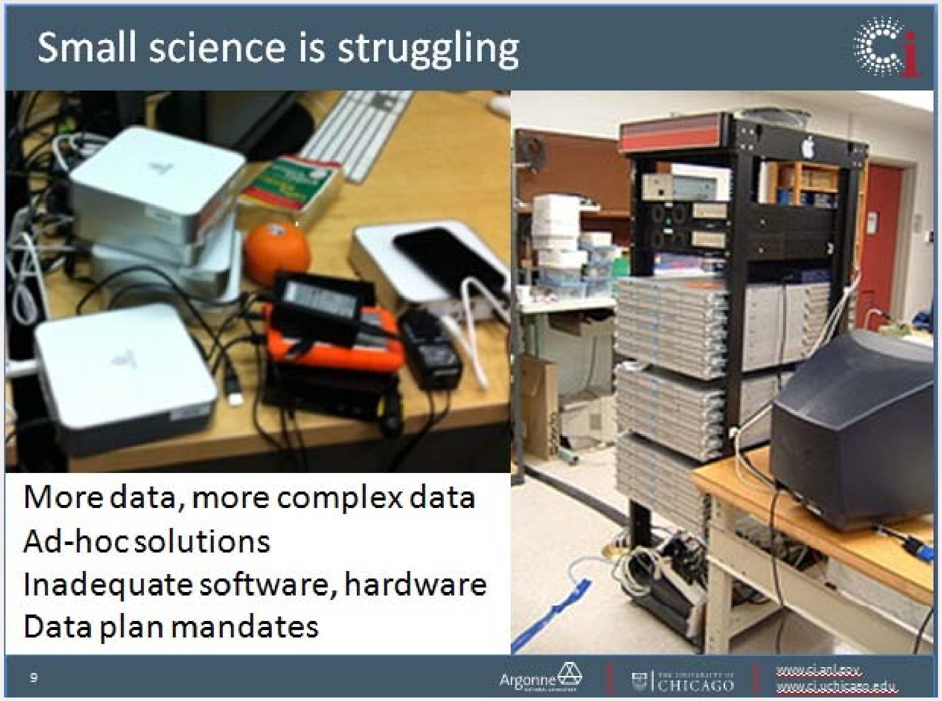
Access to Knowledge

Data Acquisition to Data Access



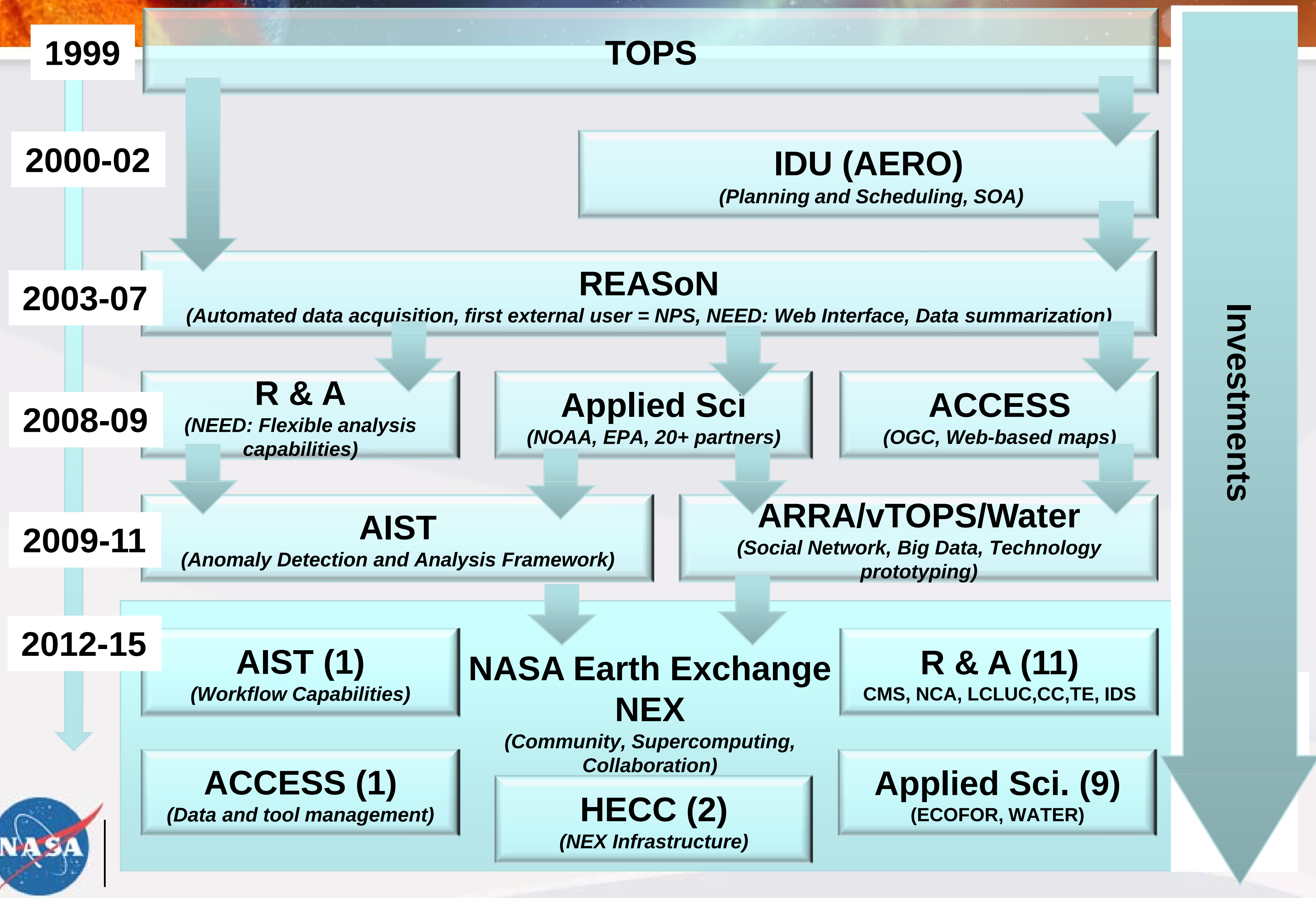
Where Can We Research/Play with Big Data Problems?

- Ian Foster talked about data processing workflow all the way to data distribution but why stop there?



- Michael Stonebraker talked about moving compute to data but there is a tremendous challenge. *(I can't even do it at one NASA center).*

NEX Precursors and Development History



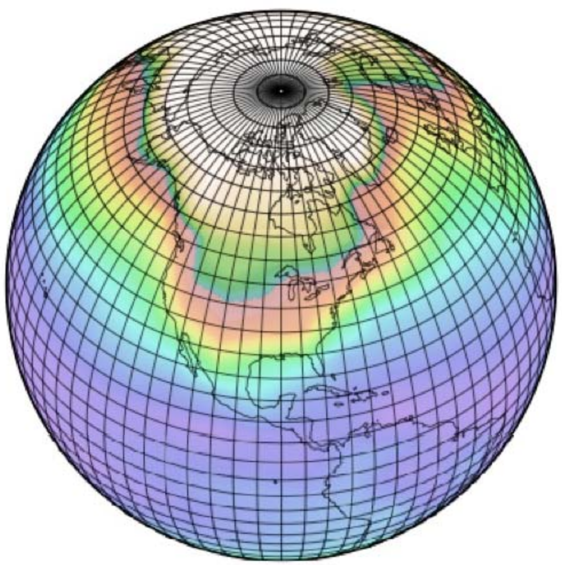
NEX: the Big Data experiment and work completed to date

Access to community/knowledge (240 members)

Access to ready-to-use data (24 products, 350TB)

Access to models/analysis tools

Climate, weather, carbon, hydrology

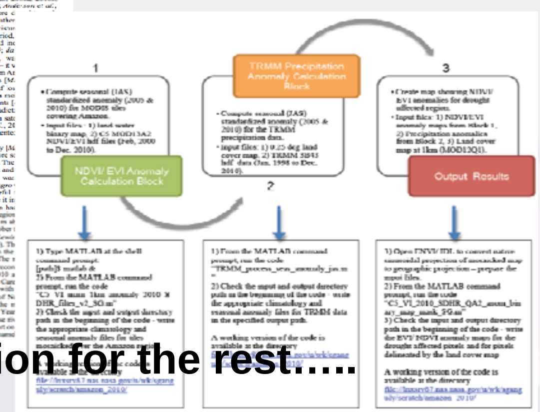


Access to workflows to build upon (VisTrails)

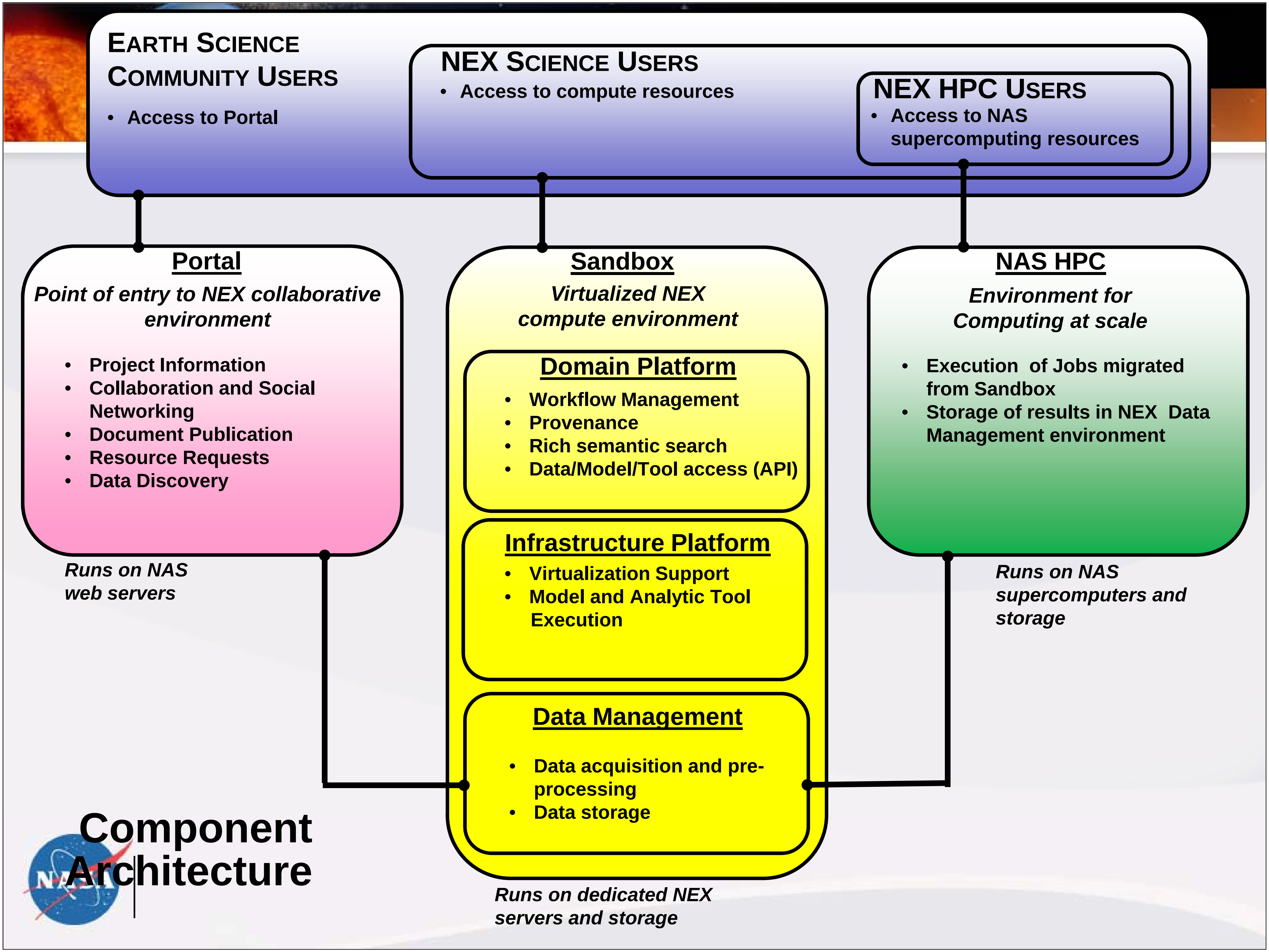

GEOPHYSICAL RESEARCH LETTERS, VOL. 38, L07402, doi:10.1029/2011GL046824, 2011

Widespread decline in greenness of Amazonian vegetation due to the 2010 drought

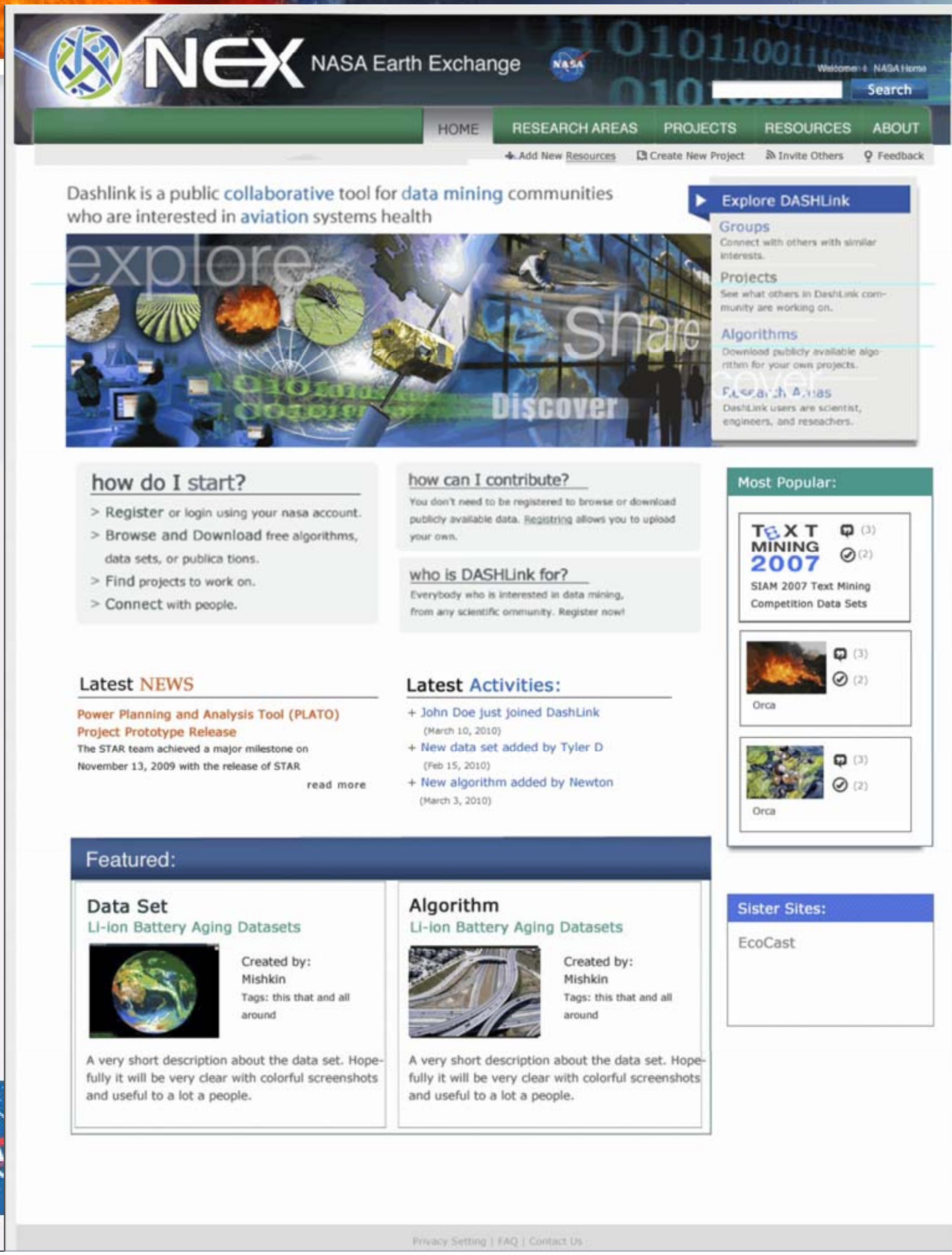
Liang Xu, Arindam Samanta, Marcos H. Costa, Sangram Ganguly, Ramakrishna R. Nemani, and Ranga B. Myneni



Ready for sophisticated users, needs proper integration for the rest…..


EARTH SCIENCE COMMUNITY USERS
• Access to Portal
NEX SCIENCE USERS
• Access to compute resources
NEX HPC USERS
• Access to NAS supercomputing resources
Portal
Point of entry to NEX collaborative environment
• Project Information
• Collaboration and Social Networking
• Document Publication
• Resource Requests
• Data Discovery
Runs on NAS web servers
Sandbox
Virtualized NEX compute environment
Domain Platform
• Workflow Management
• Provenance
• Rich semantic search
• Data/Model/Tool access (API)
Infrastructure Platform
• Virtualization Support
• Model and Analytic Tool Execution
Data Management
• Data acquisition and pre-processing
• Data storage
Runs on dedicated NEX servers and storage
NAS HPC
Environment for Computing at scale
• Execution of Jobs migrated from Sandbox
• Storage of results in NEX Data Management environment
Runs on NAS supercomputers and storage
Component Architecture
NASA

NEX Portal



Search capabilities

Who is doing what where
Science network through abstracts and papers
Who's the expert
Workflows
Archived seminars

Reporting capabilities

Annual reports
Highlights
Publications
Spatial distribution of funding

Virtual Institute

- Following the lead from Astrobiology and NASA Lunar Science Institutes to create a virtual institute, and offer
  - Summer short courses
  - Seminars
  - Conferences
  - Presentations
  - Have each funded project do one or two seminars that can be archived

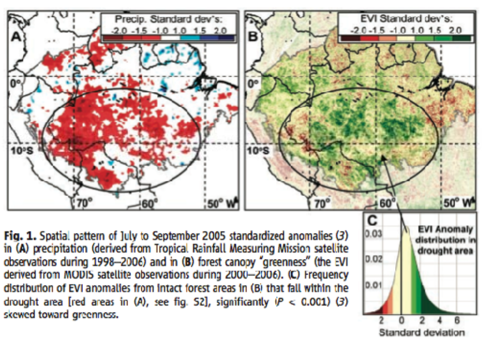
Benefits

Science 2007

BREVIA

Amazon Forests Green-Up During 2005 Drought

Scott R. Saleska,[1*†] Kamel Didan,[2*] Alfredo R. Huete,[2] Humberto R. da Rocha[3]



GRL 2010

GEOPHYSICAL RESEARCH LETTERS, VOL. 37, L05401, doi:10.1029/2009GL042154, 2010

Amazon forests did not green-up during the 2005 drought

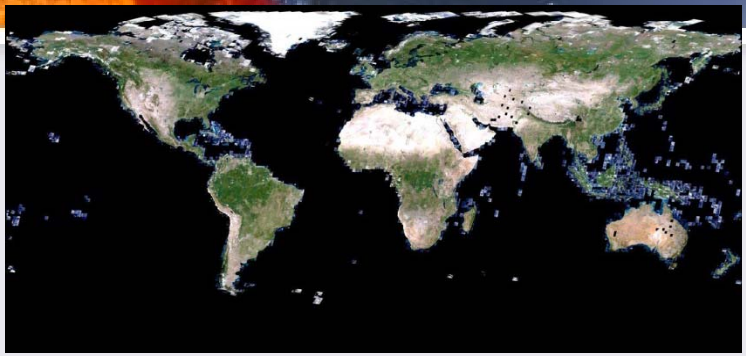
Arindam Samanta,[1] Sangram Ganguly,[2] Hirofumi Hashimoto,[3] Sadashiva Devadiga,[4] Eric Vermote,[5] Yuri Knyazikhin,[1] Ramakrishna R. Nemani,[6] and Ranga B. Myneni[1]

Would have taken less than a week in NEX, Instead of one year it took to repeat the analysis

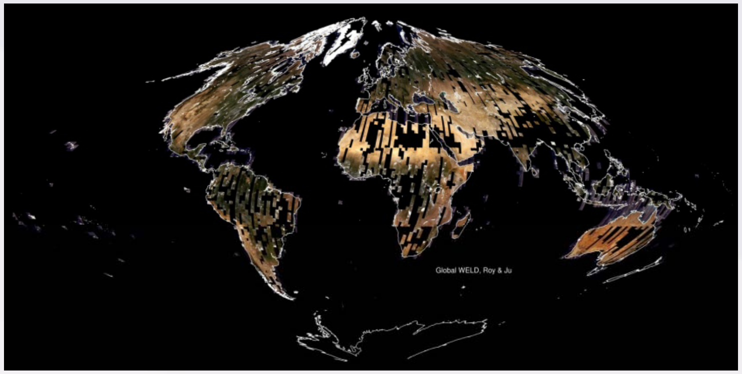
- ***Efficient use of resources***
- ***Lowering barriers to entry***
- ***Interdisciplinary work***
- ***Transparency, repeatability, extensibility***
- ***Cost reduction***
  - **Travel**
  - **Hardware**
  - **IT personnel**
  - **Data acquisition**
  - **Network costs**

Benefits (Unprecedented Opportunities)

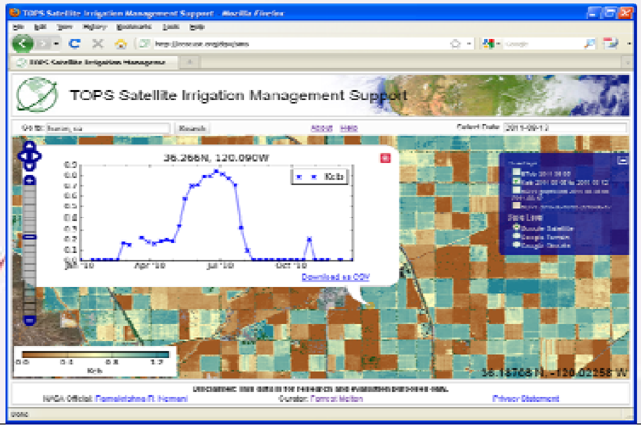
Global Land Survey from Landsat (LCLUC)



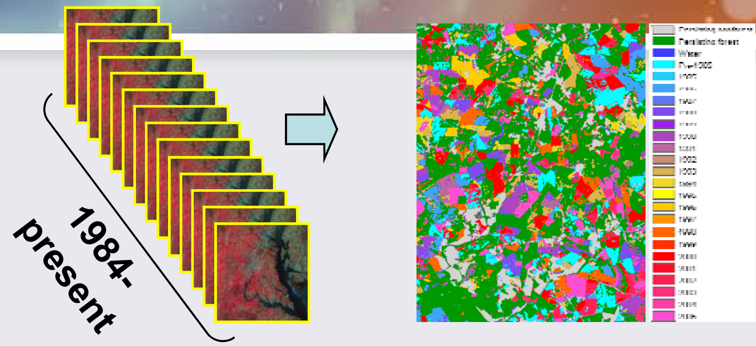
Global monthly Landsat (WELD, MEASURES)



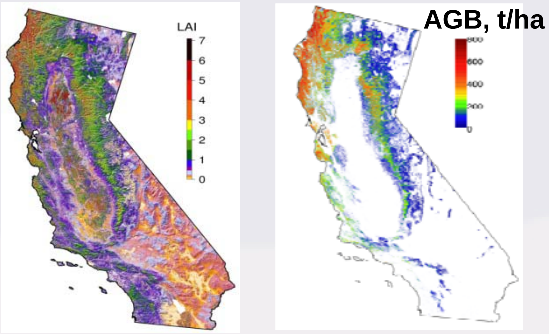
Crop water management with Landsat



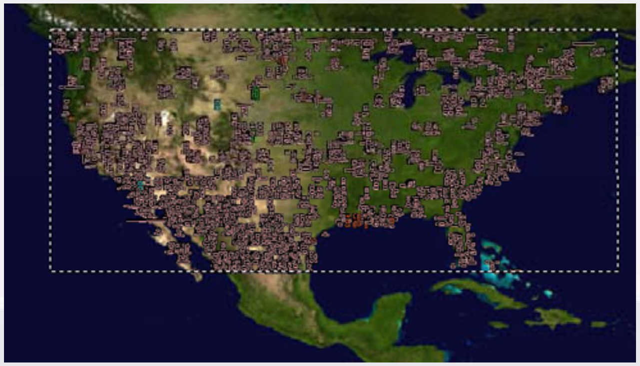
30 years of change analysis (CC)



Biophysical products from Landsat (CMS)

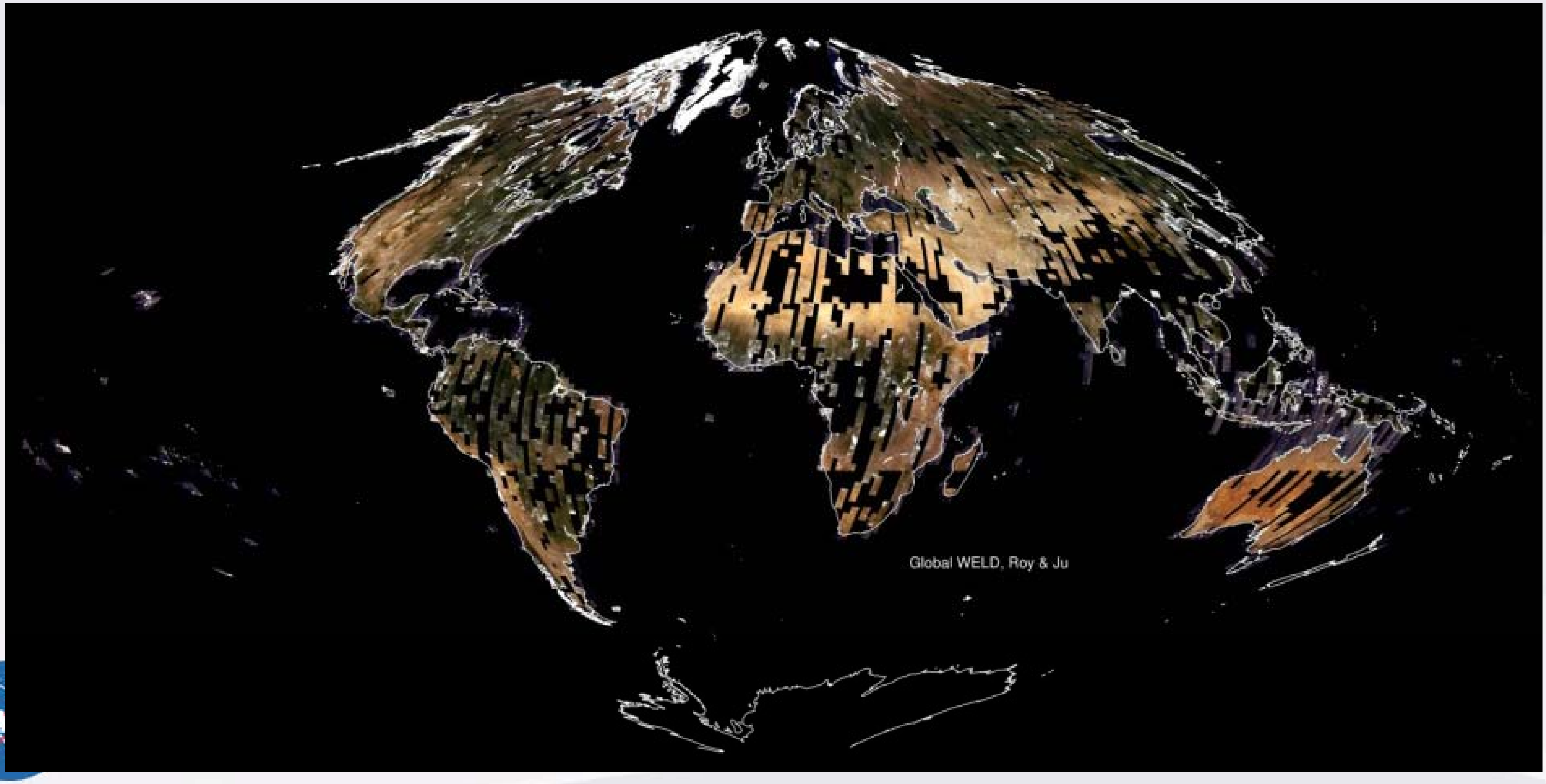


WorldView-2, 50cm, 8 bands (NGA)



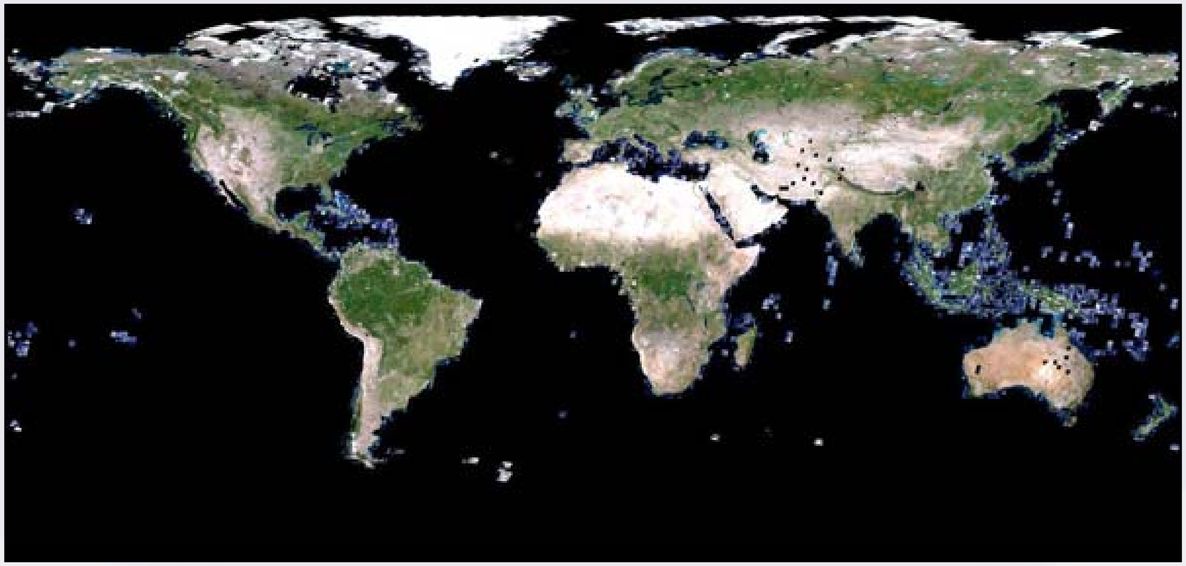
Big Data Metric?

Space-time-resolution metric (higher spatial resolution, larger extent, longer time-series studies)



NEX Experiments Demonstrate Value of Collaborative Environment

- In a first application of NEX, a research team from around the U.S. used the environment to adjoin and atmospherically correct a mosaic of 9,000 Landsat Thematic Mapper scenes and retrieve global vegetation density at a 30-meter resolution.
- The entire processing of the nearly 340 billion pixels in the composite took just a few hours on the Pleiades supercomputer, allowing the team to experiment with new algorithms and products within just a few days.



- NEX's collaboration and knowledge-sharing platform for the Earth science community combines supercomputing, Earth system modeling, workflow management, and NASA remote sensing data feeds to deliver a complete work environment for users to explore/analyze large datasets, run modeling codes, collaborate, and share results.

POC: *Petr Votava, petr.votava@nasa.gov, (650) 604-4675****;*** *Ramakrishna Nemani, ramakrishna.nemani@nasa.gov****,*** *(650) 604-6185, NASA Ames Research Center*

Future Work

- Data Management
  - How to load and offload data in a multi-tier storage environment?
  - Distributed Multi-site Analytics?
- Workflow Reuse
  - Why create your own if a similar analytics was done before?
  - Use workflow to capture the data provenance?
- Workflow Discovery
  - Why not mine the literature to uncover the workflow?