Opportunities and Challenges of Data Sharing in an Academic Research Computing Environment

Thomas Hauser
Director of Research Computing
University of Colorado Boulder
thomas.hauser@colorado.edu

The Conduct of Science is Changing

Revolution in the scientific workflow: many interfaces to shared CI services

Large Facilities

Software

Shared Data/Software Gateway Resources

Researcher

Collaboration Networks

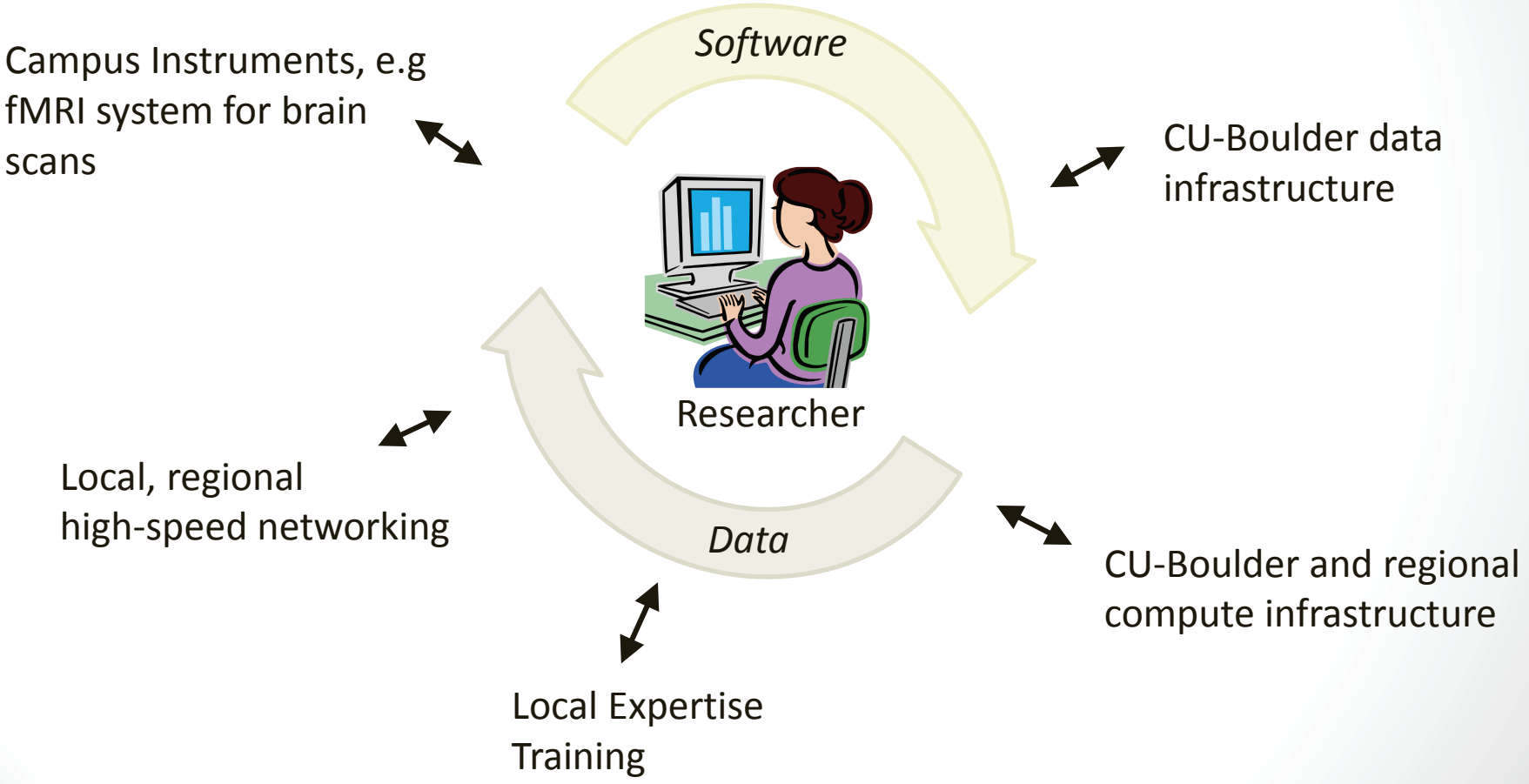
Data

Cloud Services

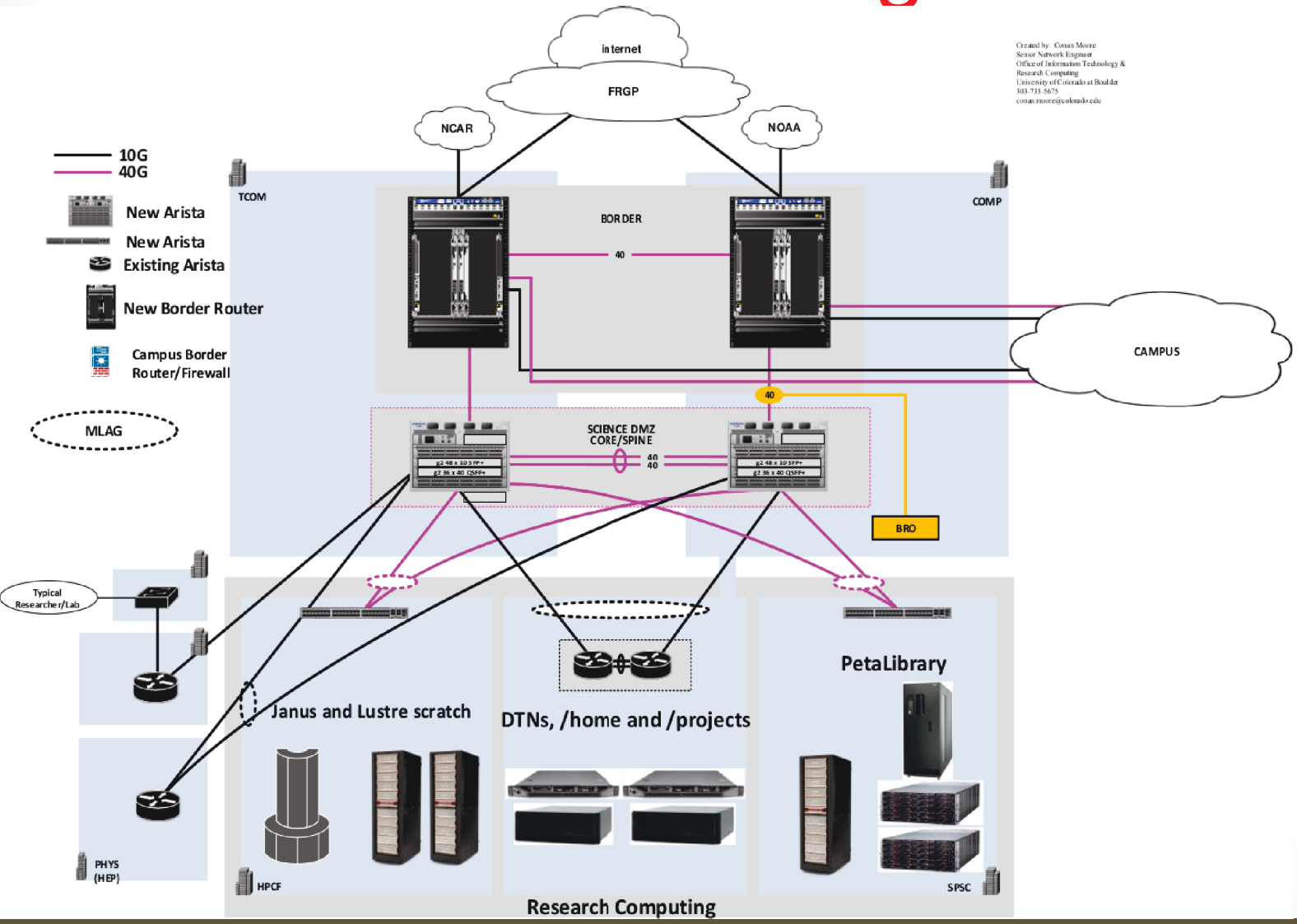
National Computing Resources

Irene Qualters NSF: http://slideplayer.com/slide/8972645/

Local Cyberinfrastructure Ecosystem



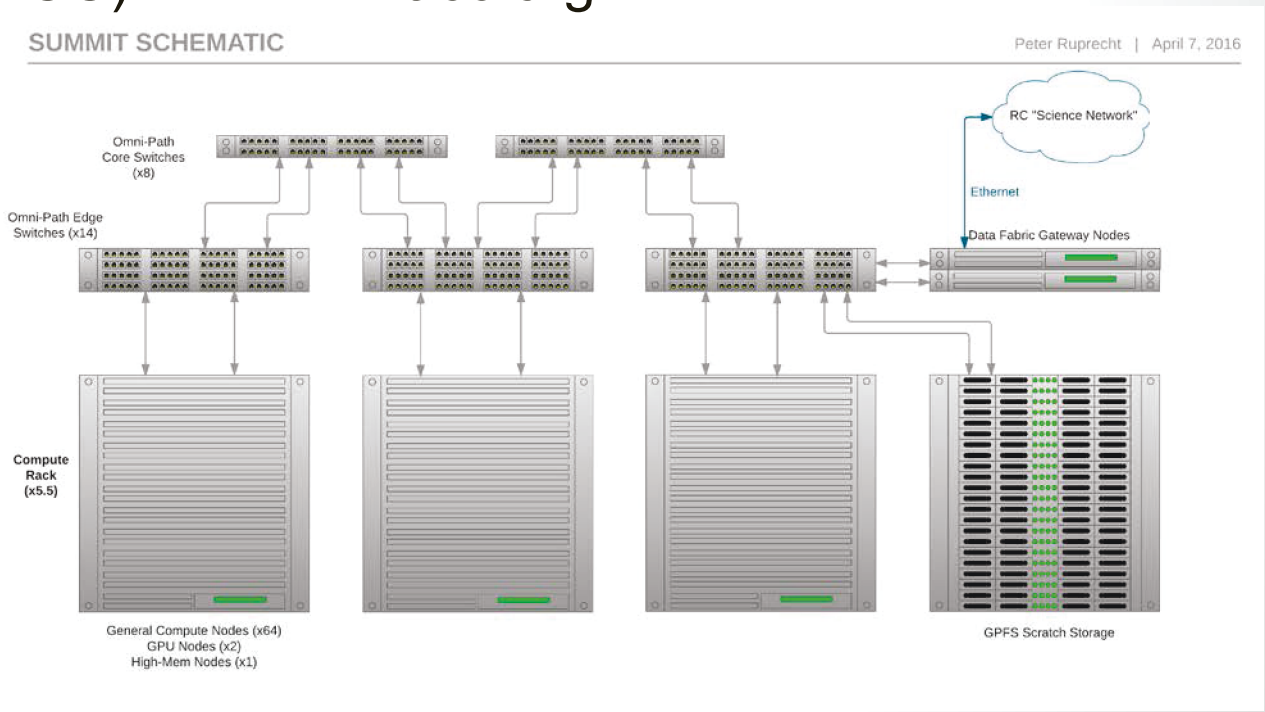
Frictionless Networking – ScienceDMZ



Graphic:
Dan Milroy
Conan Moore

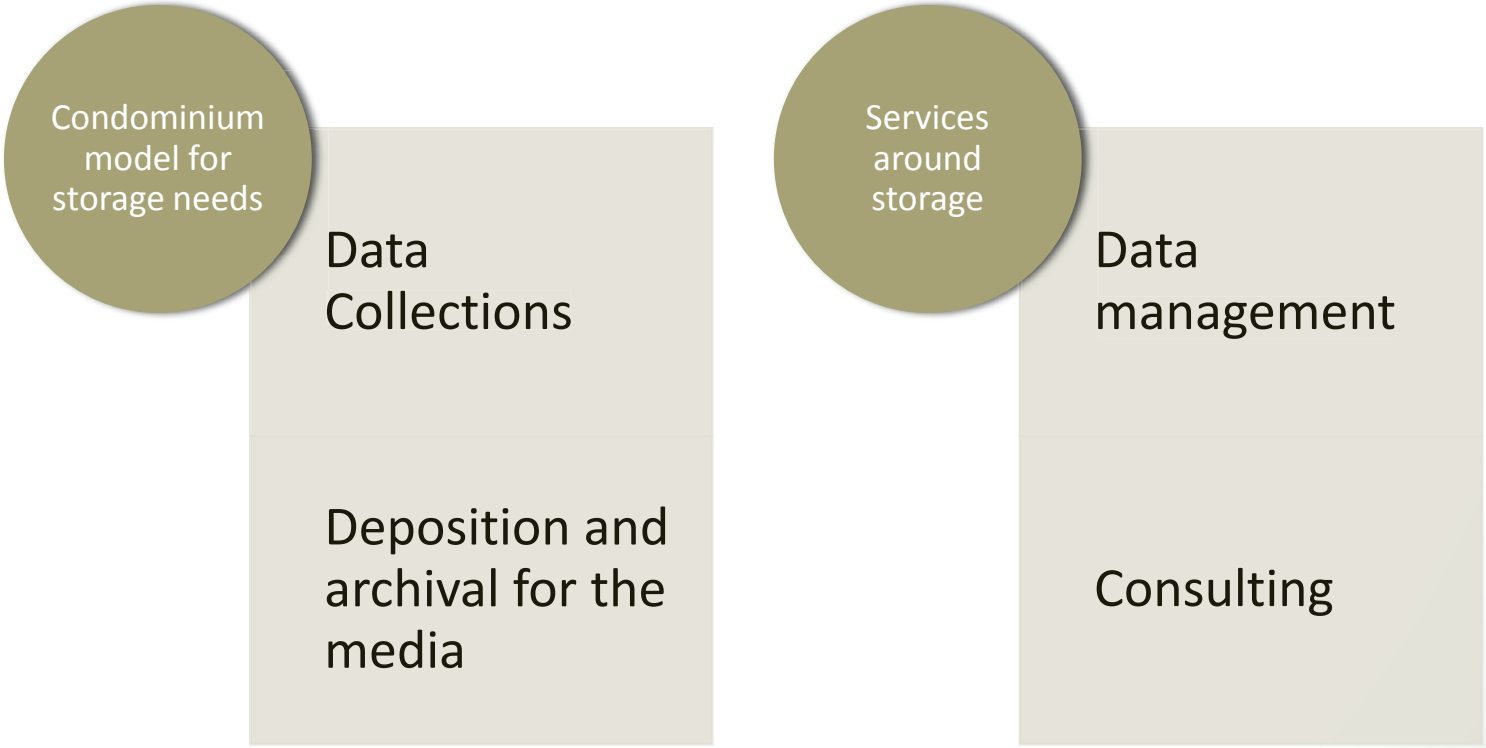
Shared Compute @ CU

- New regional Summit supercomputer funded by NSF
  - Joint project between CSU and CU-Boulder
  - Access for members of Rocky Advanced Computing Consortium (RMACC) – www.rmacc.org
- Condo model



Research Data Storage - PetaLibrary

- NSF Major Research Instrumentation grant
- No HIPAA, FERPA, ITAR data



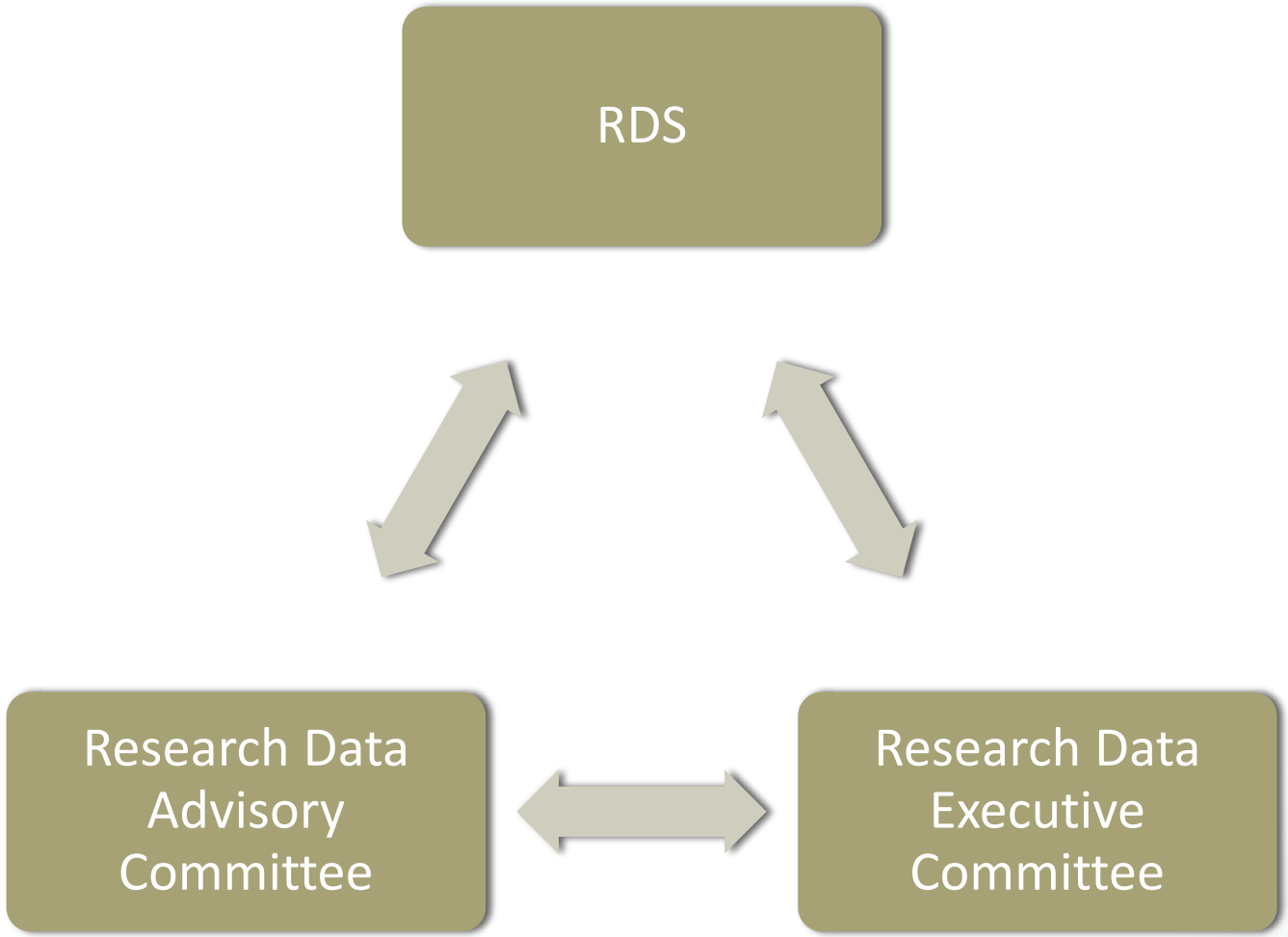
Quick Facts about RC at CU

- Nearly 2000 user accounts (including storage-only users)
- Between 150-200 are actively computing every month
- 50 in a given day
- About 700,000 compute jobs each year
- About 100,000,000 core-hours each year
- Over 400 TB in PetaLibrary storage projects
- Meetup every week during the spring and fall semester
  - 643 members of meetup group
- Software and data carpentry

Expertise - What is RDS?

- Research Data Services @ CU-Boulder
- Collaborative effort
  - Research Computing (Shelley Knuth)
  - Libraries (Andrew Johnson)
- Data management assistance
  - Data management plans (DMPs)
  - Advice on storage
  - Globus Online data sharing
  - Processing (forthcoming)
  - …

Research Data Governance



Data Education and Outreach

- Driven by Shelley Knuth (shelley.knuth@colorado.edu)
- Training the next generation
  - Data management boot camps
    - General campus
    - Campus admins
  - Specific topics:
    - Data publishing
    - Research Funder Requirements
    - HDF5
    - Python
    - Globus Online
  - Tutorial assessment to improve delivery and content
- Visit department faculty meetings

Shelley Knuth

Campus Competitions - Data

- Data management as part of internal competitions
  - Gets campus ready for DMPs
  - Brings awareness of services

1. Innovative Seed Grants
   - Research money for faculty
   - Added DMPs in 2014 and 2015
   - Held workshops to assist
2. DMP competition
   - Submit just DMP
   - Agree to publication of plan on web
   - Winners in 5 categories
   - $2000 in unrestricted funds

Collaboration with Libraries

- Center for Research Data and Digital Scholarship
  - Under review by campus leadership
- Joint leadership
  - Senior Associate Dean of the Libraries
  - Director of Research Computing
- Faculty appointments for Research Computing staff
- Centered around Initiatives
  - Education and Training – Shelley Knuth (RC)
  - Cyberinfrastructure – Debora Weiss (Libraries)
  - Digital Scholarship – Thea Lundquist (Libraries)
  - Data Management – Andrew Johnson (Libraries)

Regional Collaboration

- Rocky Mountain Advance Computing Consortium (RMACC)
  - www.rmacc.org
- Yearly HPC Symposium
  - August 9-11 at CSU this year
- Coordinate and collaborate on proposals
- Share infrastructure
- Create a community of system administrators and computational scientists
  - Quarterly HPC system administrator retreat
  - Planning to replicate with computational scientists

Globus Online for Researchers

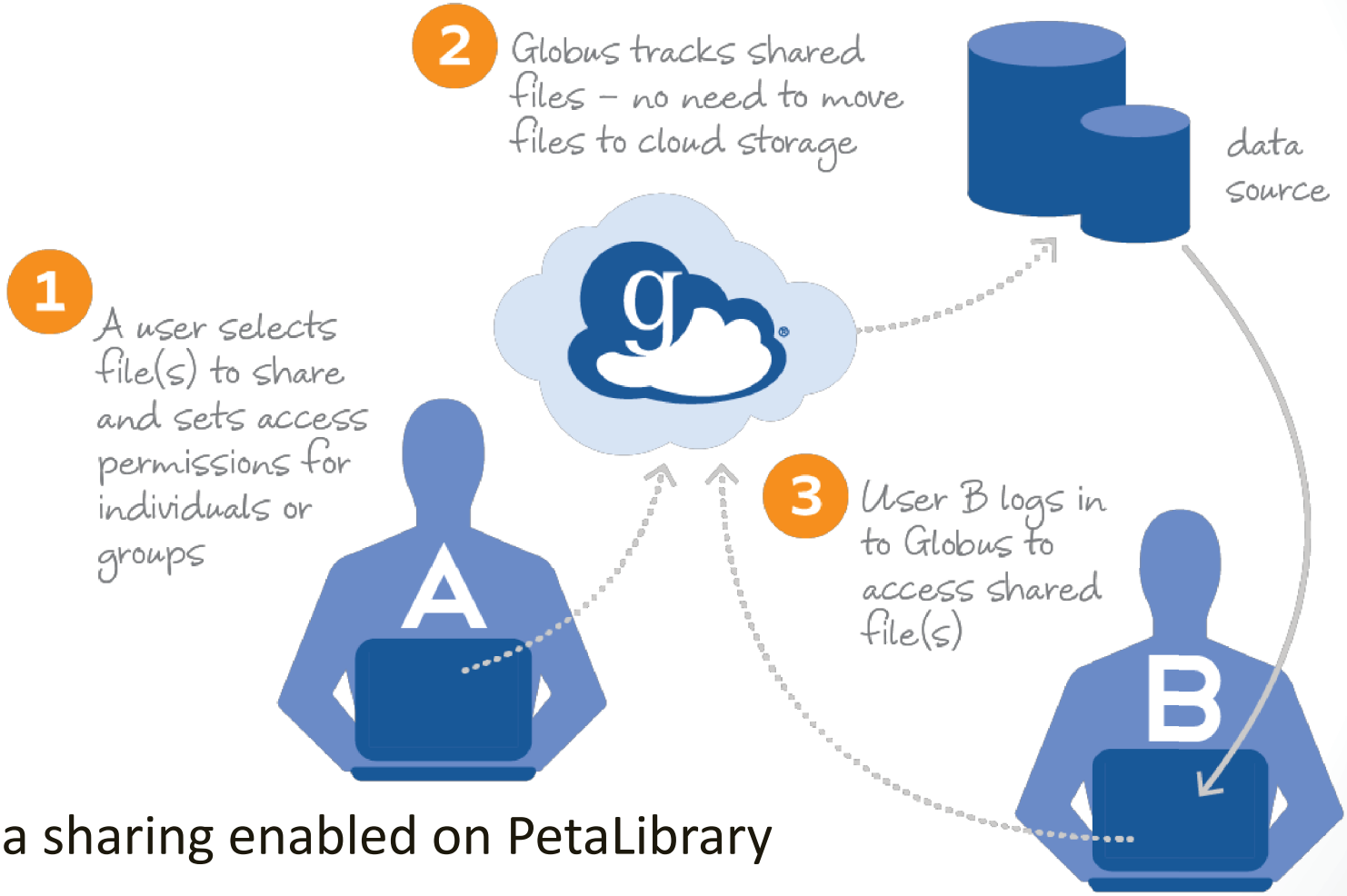
Image from https://www.globus.org/sites/default/files/landing_page/illustration-globus-for-researchers002%402x_0.png

- Transfer
- Share
- Publish – not a service for CU-Boulder's researchers yet

Data Corruption during Data Transfer

- Using naïve data transfer tools like FTP
  - 1 in 65,536 packets will be erroneously accepted as correct
  - Data corruption during transfer
  - Example: NOAA Big Data project transfer to the cloud
- Globus Online
  - 128-bit checksum
  - One in 2×10^{13} will be accepted as correct

Globus Data Sharing



Data sharing enabled on PetaLibrary

Image from https://www.globus.org/sites/default/files/how_it_works-sharing@2x.png

Globus Data Publishing

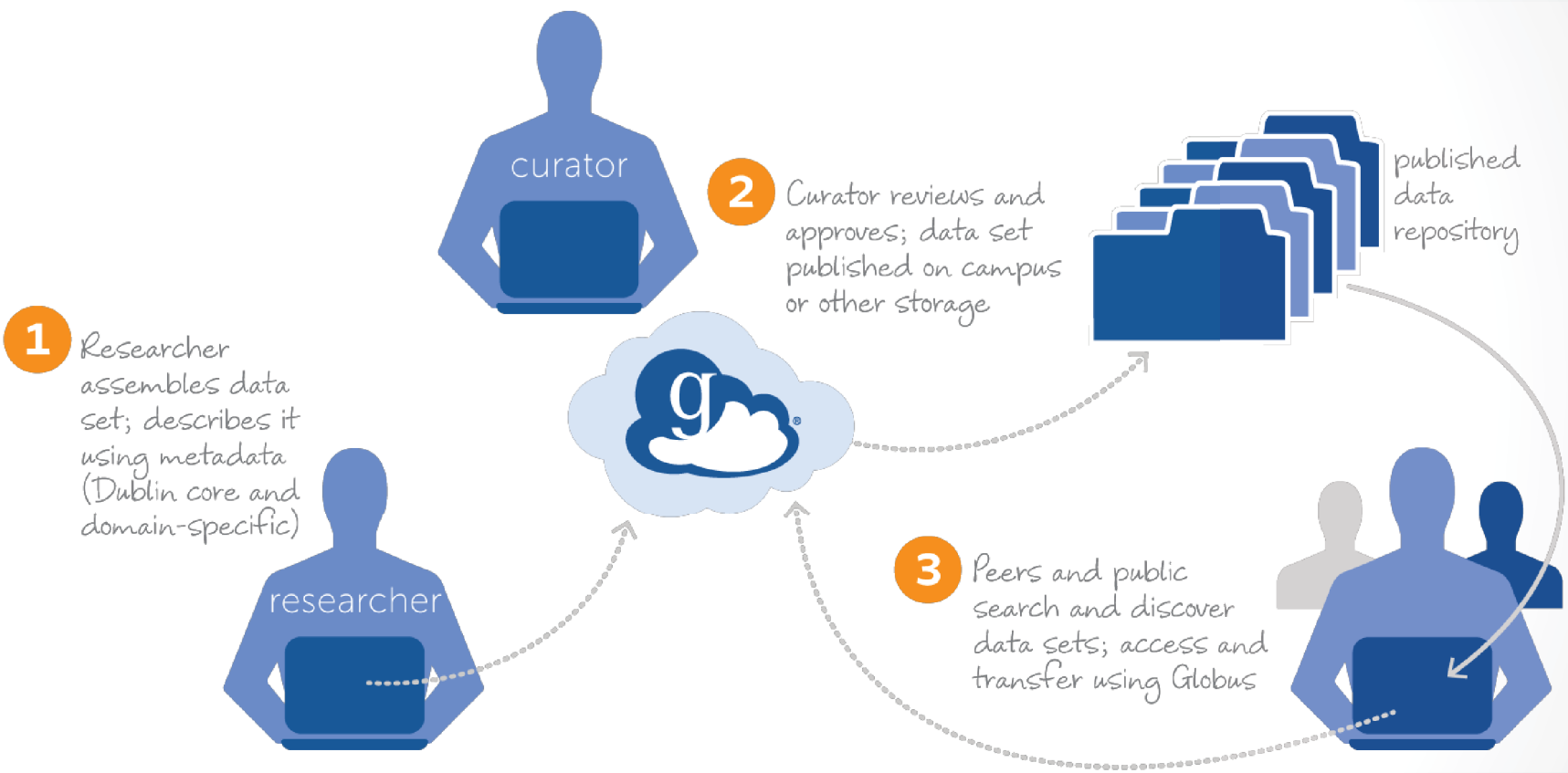


Image from https://www.globus.org/sites/default/files/how_it_works-publishing2%402x_0.png

Not a service for CU-Boulder's researchers yet

Data Guiding Principles

- FAIR Guiding Principles: https://www.force11.org/group/fairgroup/fairprinciples

- Findable
- Accessible
- Interoperable
- Reusable

Reality Check of Globus Publishing again Fair Data Guiding Principles

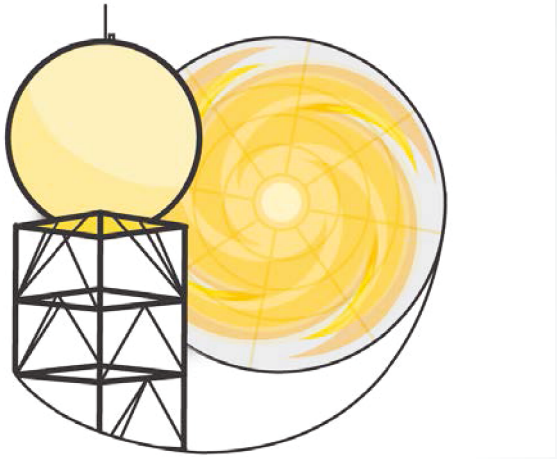
- Spot check the following data set
  - Jang, Hyejin; Wood, Joshua D.; Ryder, Christopher R.; Hersam, Mark C.; Cahill, David G., "Anisotropic Thermal Conductivity of Exfoliated Black Phosphorus," 2016, http://dx.doi.org/doi:10.18126/M2BC7S
- Findable
  - F1. (meta)data are assigned a globally unique and eternally persistent identifier.
    - Data set has DOI assigned
  - F2. data are described with rich metadata.
    - Excel and jason file for metadata
  - F3. (meta)data are registered or indexed in a searchable resource.
    - I had difficult finding the data set - could find the publication
- Accessible through Globus
- Data might be interoperable
- No clear license for the data

Earth Lab Initiative at CU-Boulder

- Earth Lab
  - Capitalize on the data deluge from Space and other platforms to accelerate science
  - Reduce environmental risk by using this wealth of data to better understand and predict Earth System change
  - Train a new generation of data scientists who can address outstanding Earth Science questions
- Creating an analytics hub for data integration, visualization and analysis
- Proving ground for data approaches
  - Data standards
  - Expectations around data for a member of Earth Lab
- Training success
  - Globus Online adoption for data transfers by Earth Lab members

NOAA Data Pilot

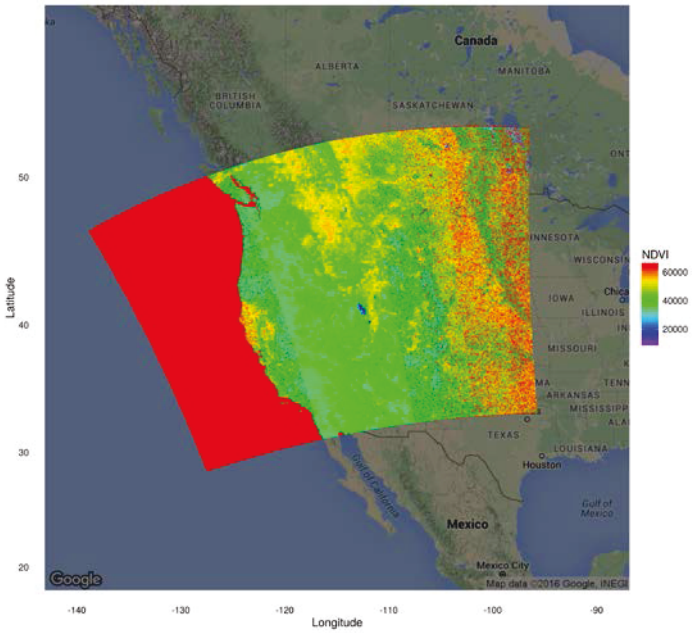
- Earth observation data meeting at CU-Boulder
- Integrate several data sets
  - Low resolution soil moisture data
  - High resolution vegetation data
  - Expected outcome
    - Soil moisture increases
    - Vegetation gets greener
- Evaluate relationships between
  - Soil moisture
  - Normalized difference vegetation index (NDVI)
  - Ecoregions in Colorado

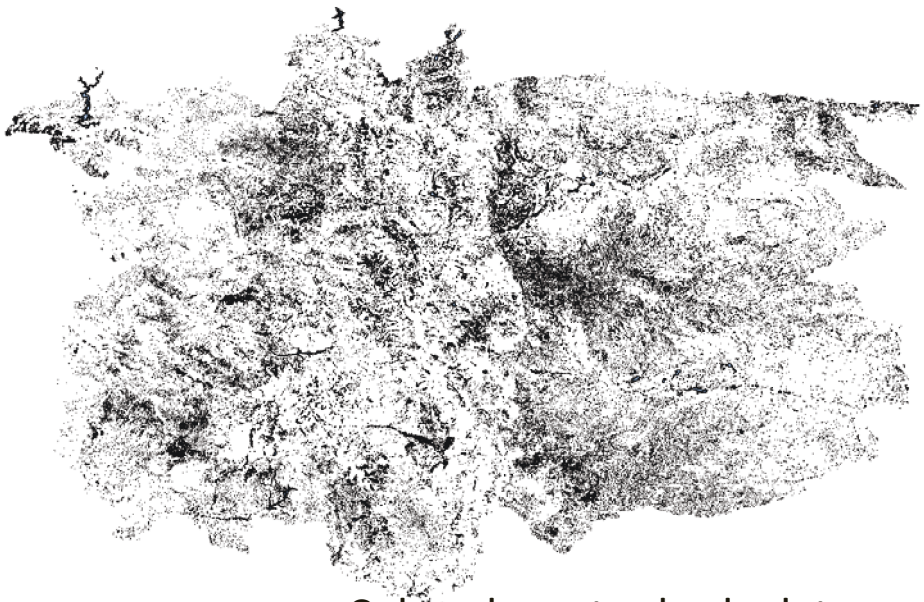


http://d0.awsstatic.com/events/aws-hosted-events/2015/WWPS/Miscellaneous/440x303_NEXRAD-graphic_Yellow-v2.png

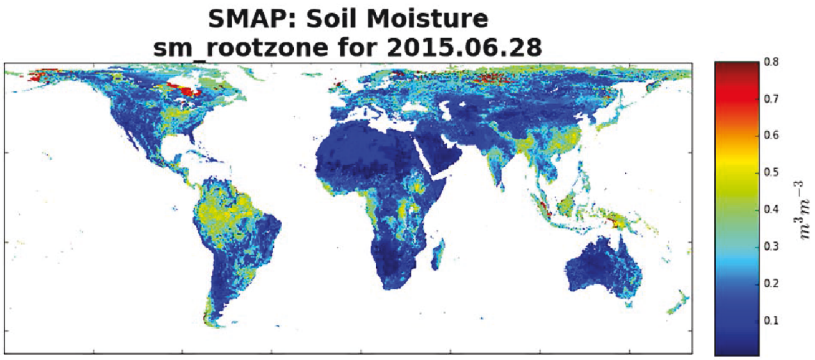
NOAA Big Data Initiative

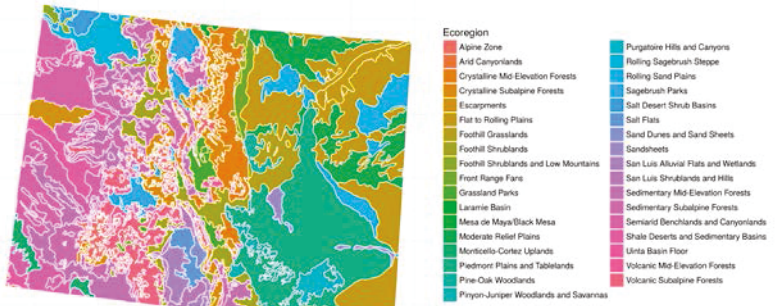
Data Sets





Colorado water body data

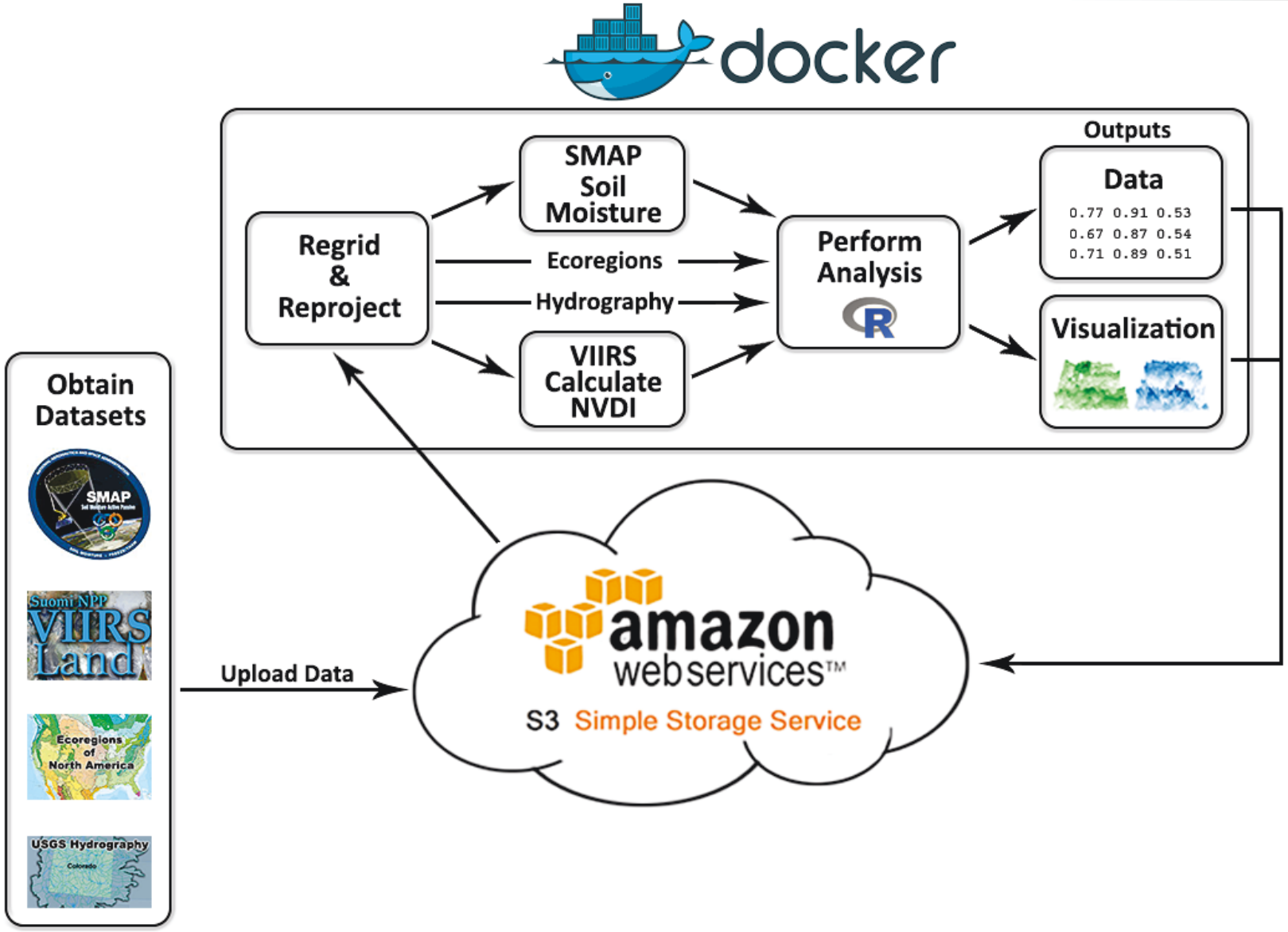




Ecoregion data

Paradigm Shift

- Move analytics to the data
  - Solve science and engineering problems at scale
  - Large scientific data sets
  - Difficult to move to local environment
- Create reproducible analytics work flows
  - Leverage virtualization
  - Package the whole workflow with Docker
  - Develop locally – deploy at scale in the cloud
- Central data repository
  - Even for 100 GB network download might not be possible
  - Manage and organize data only once
  - Provide access to the data/meta-data through interfaces

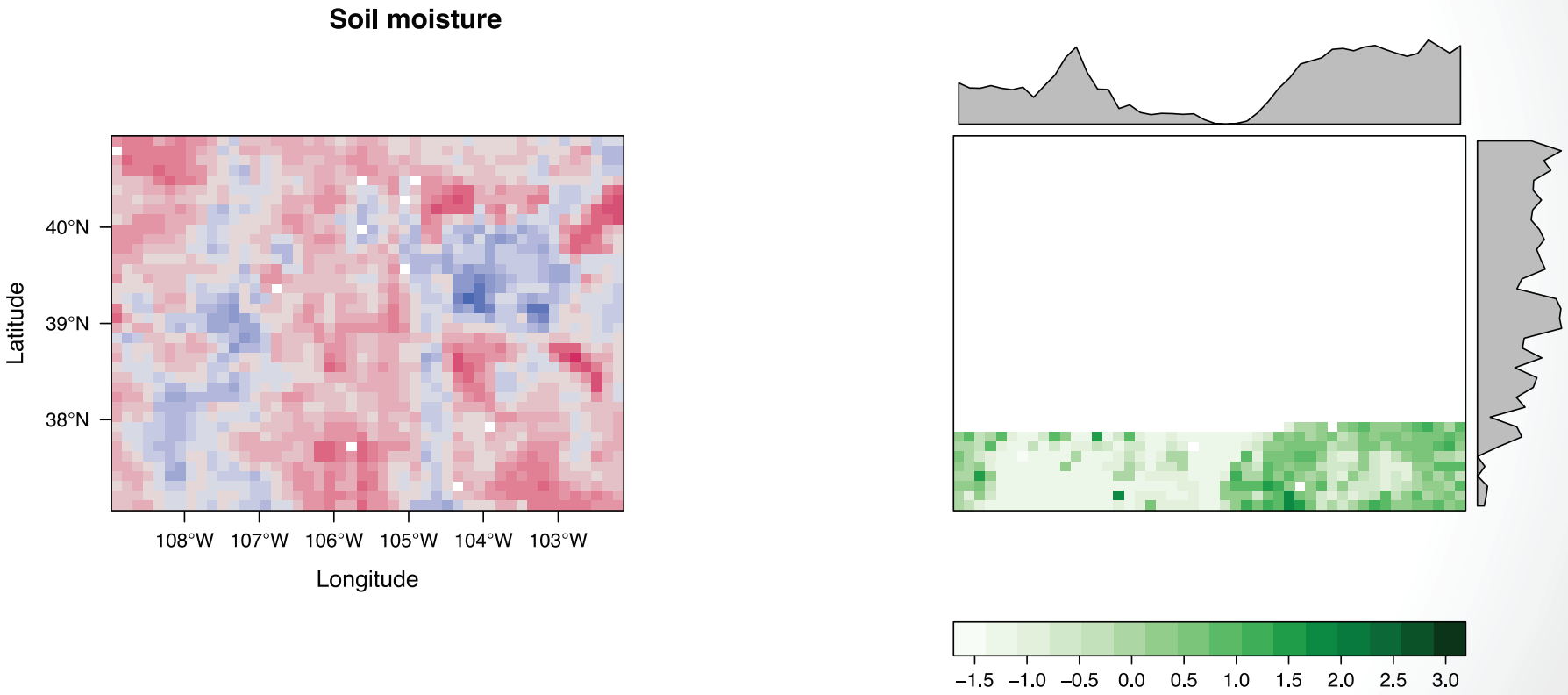

docker
Outputs
SMAP Soil Moisture
Regrid & Reproject
Ecoregions
Hydrography
VIIRS Calculate NVDI
Perform Analysis
Data
0.77 0.91 0.53
0.67 0.87 0.54
0.71 0.89 0.51
Visualization
Obtain Datasets
SMAP
Suomi NPP VIIRS Land
Ecoregions of North America
USGS Hydrography
Upload Data
amazon web services™
S3 Simple Storage Service

Analytics Command Line

- `docker run -e AWS_ACCESS_KEY_ID=$AWS_ACCESS_KEY_ID -e AWS_SECRET_ACCESS_KEY=$AWS_SECRET_ACCESS_KEY mbjoseph/noaa-demo`

- Code is available
  - `docker pull mbjoseph/noaa-demo`

Result of Analytics Workflow



Challenges & Opportunities

- Culture
- Researcher view data management as a burden and overhead
  - Publish and move on
- Peer reviewed data publication and software publication should count in the research reward system
- Data is not machine accessible
- Provenance of data
- Indexing and finding data
- Reducing time of data wrangling
  - From 80% wrangling to 20% wrangling data
- Career path for support staff

Questions

Join us for only $30 this summer to learn more about High Performance Computing with sessions such as:

- Introduction to High Performance Computing
- Parallel Programming for Beginners
- Introduction to Visualization
- Making the most of Grad School
- and many others...

All students are also invited to participate in our online student poster competition on Wednesday August 12th. Student poster submissions will include a digital poster and accompanying 5-minute video presentation. They will be accepted online through 5 pm MDT, August 5, 2016. Students must be affiliated with an RMACC member institution, but need not attend the RMACC Symposium to win.

The poster competition finalists will receive an all-expense paid trip to the IEEE Supercomputing Conference SC16, in Salt Lake City, Utah
More information is available at our website:
https://rmacc.org/hpcsymposium/posters

Travel funds are available to assist students planning to attend the RMACC Symposium

Organized and hosted by the
Rocky Mountain Advanced Computing Consortium
www.rmacc.org
Questions? Contact us at rmacc@colorado.edu

RMACC
Rocky Mountain Advanced Computing Consortium