Large-scale Data Systems at the National Center for Atmospheric Research (NCAR)

Anke Kamrath (anke@ucar.edu)
Director, Computing Operations and Services , NCAR

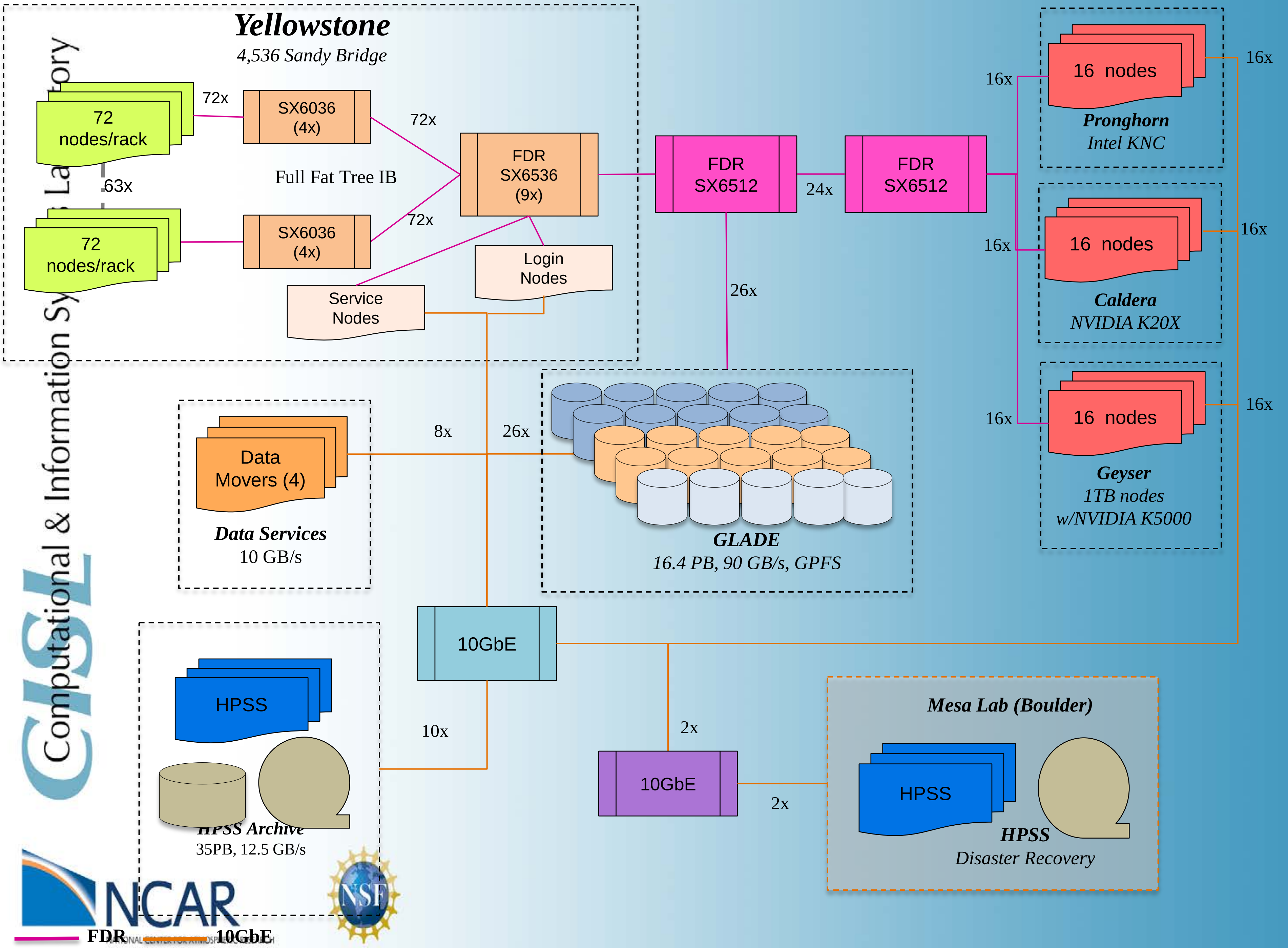
CISL Computational & Information Systems Laboratory

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NSF

Overview

- **Yellowstone - NCAR Data Intensive Computing Environment**
- **Globus Online**
  - Data Transfer Service
  - NCAR Cloud - Data Sharing Service
- **Large-scale Data Collections**
  - RDA – NCAR Research Data Archives
  - ESGF and CMIP – Earth System Grid Framework
  - CMIP Data Analysis Platform
- **“Herding” data from the NCAR researchers and Labs**
  - DSET – Data Stewardship Engineering Team



Yellowstone

4,536 Sandy Bridge

72 nodes/rack

72x

SX6036 (4x)

72x

63x

Full Fat Tree IB

72 nodes/rack

SX6036 (4x)

72x

FDR SX6536 (9x)

Login Nodes

Service Nodes

FDR SX6512

24x

FDR SX6512

26x

16x

16 nodes

Pronghorn

Intel KNC

16x

16x

16 nodes

Caldera

NVIDIA K20X

16x

16x

16 nodes

Geyser

1TB nodes w/NVIDIA K5000

16x

16x

Data Movers (4)

Data Services

10 GB/s

8x

26x

GLADE

16.4 PB, 90 GB/s, GPFS

10GbE

HPSS

HPSS Archive

35PB, 12.5 GB/s

10x

2x

10GbE

2x

Mesa Lab (Boulder)

HPSS

HPSS

Disaster Recovery

FDR

10GbE

Computational & Information Systems Laboratory

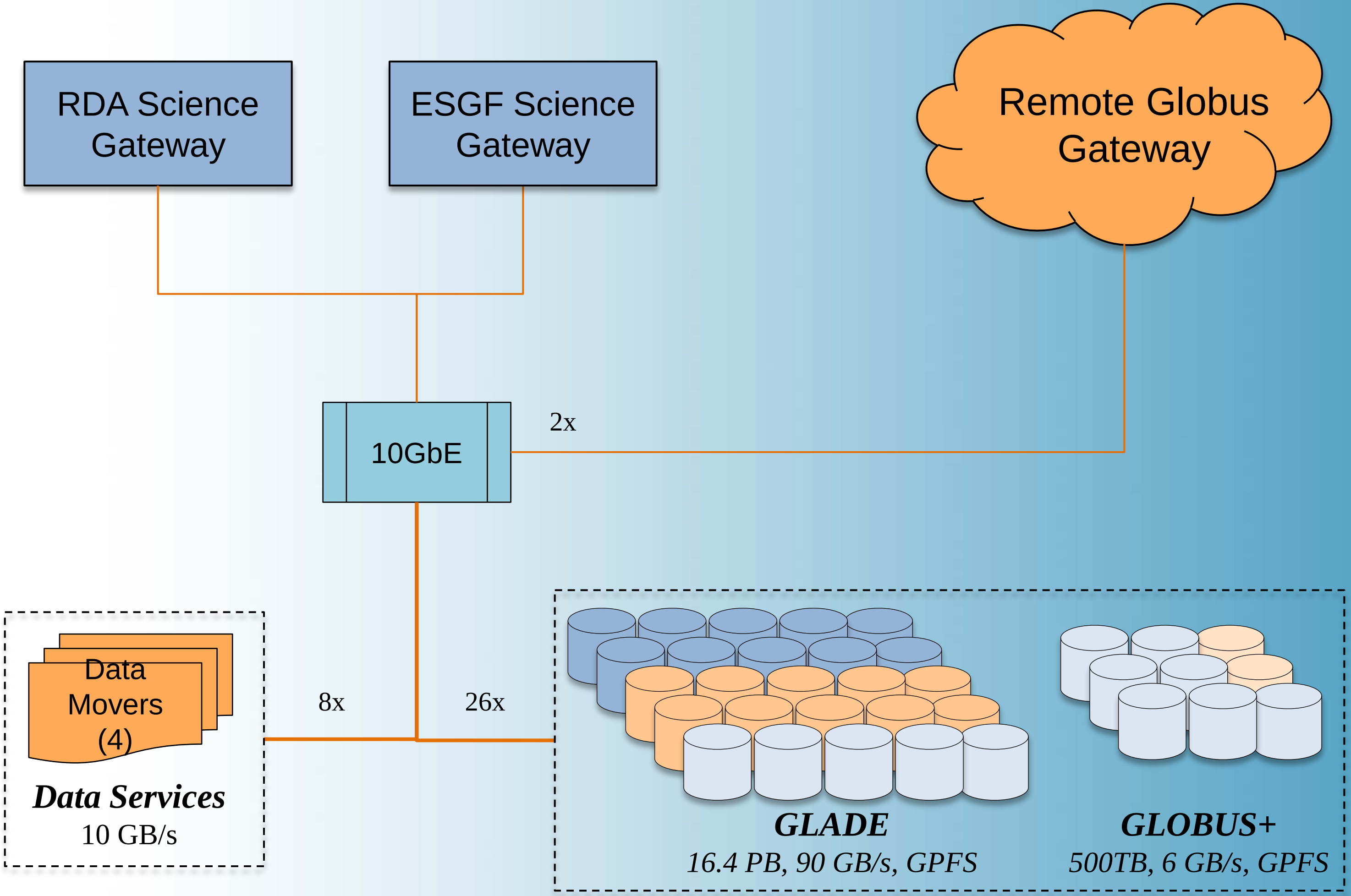
CISL

NCAR

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

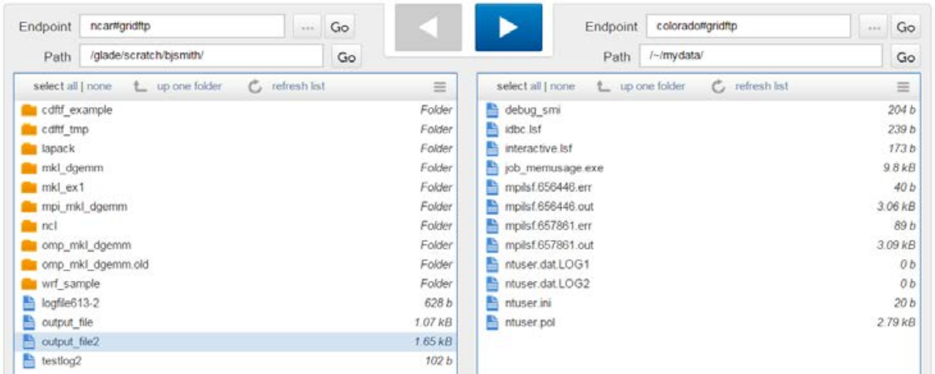
NSF

Data Transfer Gateway



Data Transfer Services

- **Globus Online Endpoints**
  - launch and forget data transfers
  - Access with users UCAS account and token
    - ncar#gridftp
  - Access with users XSEDE Account
    - xsede#ncar
    - xsede#glade
  - Web UI, CLI, REST API
  - Globus Connect for transfer to/from your desktop
- **gridftp, globus-url copy, scp/sft, bbcp**
- **HSI/HTAR for HPSS access through LSF**


Endpoint: ncar#gridftp

Path: /glade/scratch/bjsmith/

cdftf_example	Folder
cdftf_tmp	Folder
lapack	Folder
mkl_dgemm	Folder
mkl_ex1	Folder
mpi_mkl_dgemm	Folder
ncl	Folder
omp_mkl_dgemm	Folder
omp_mkl_dgemm.old	Folder
wrf_sample	Folder
logfile613-2	628 b
output_file	1.07 kB
output_file2	1.65 kB
testlog2	102 b

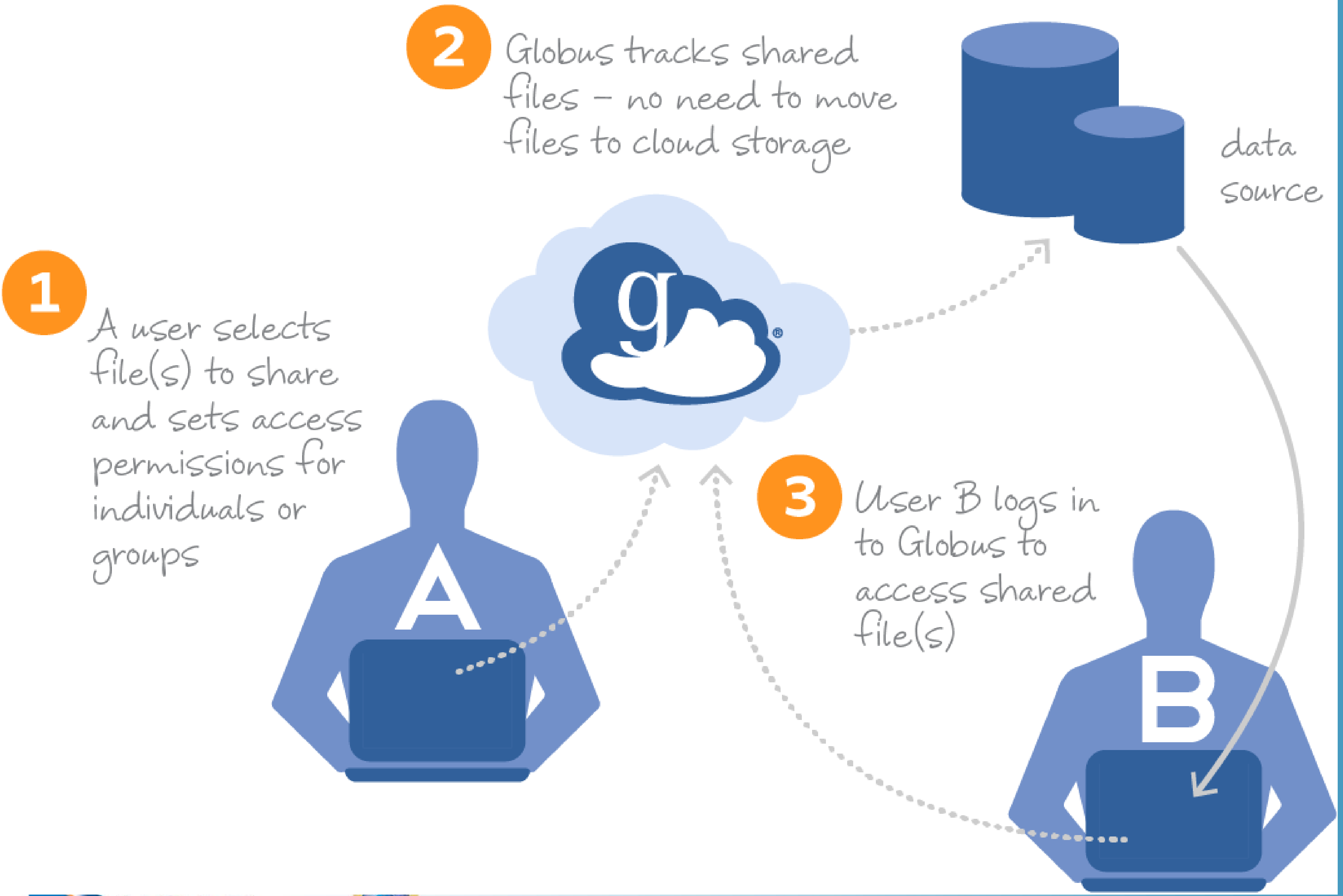
Endpoint: colorado#gridftp

Path: /~/mydata/

debug_smi	204 b
idbc.lsf	239 b
interactive.lsf	173 b
job_memusage.exe	9.8 kB
mpilsf.656446.err	40 b
mpilsf.656446.out	3.06 kB
mpilsf.657861.err	89 b
mpilsf.657861.out	3.09 kB
ntuser.dat.LOG1	0 b
ntuser.dat.LOG2	0 b
ntuser.ini	20 b
ntuser.pol	2.79 kB

Data Sharing Services

- ***ncar#datashare***
- **Globus Plus implementation**
- **data sharing allocations for self-publishing or data delivery**
- **data owner controls access**
  - can create groups for access control
  - can share ‘read-only’ or ‘read-write’
- **user can create custom access interfaces**
  - CLI or REST API
- **500 TB of useable space**
- **> 6 GB/s bandwidth available**
- **1/2014-4/2015: >160TB transferred. (average filesize >3GB)**


2
Globus tracks shared files – no need to move files to cloud storage
data source
1
A user selects file(s) to share and sets access permissions for individuals or groups
A
3
User B logs in to Globus to access shared file(s)
B

Data Sharing Use Cases

- **Delivery of data from non NCAR users for publication in a Science Gateway**
  - Periodic delivery of data sets from NCAR collaborators for publication through the ESGF Science Gateway
- **Delivery of 3D visualization to non NCAR users**
- **Publication of supporting data associated with publication**
- **Share a file or data set with a non NCAR collaborator**

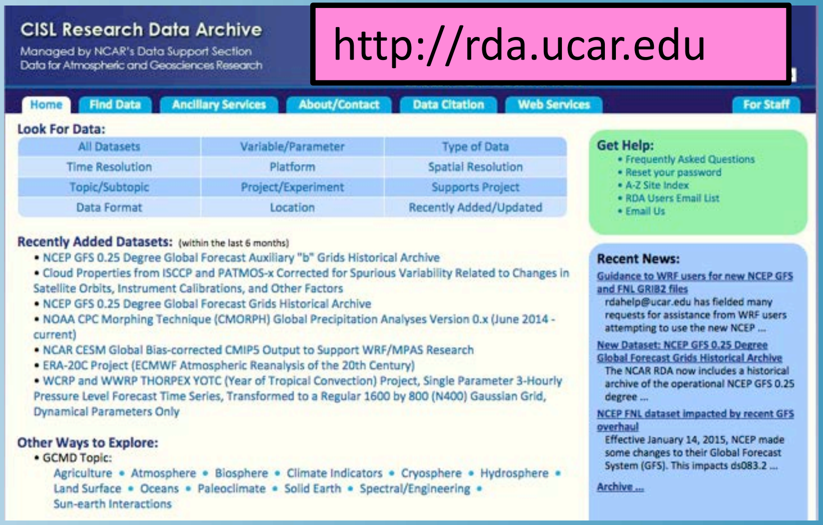
Large NCAR Data Collections

- **RDA – NCAR Research Data Archives**
- **ESGF and CMIP – Earth System Grid Framework**
  - CMIP Data Analysis Platform *(Coming Soon)*

What is the RDA?

ation Systems Laborato

- 600+ distinct datasets for climate and weather research, 8M Files, 1.8 Pb
- Collections: ocean & atmosphere observations, analyses, reanalyses, operational NWP outputs
- Science educated staff
- Free and open access



http://rda.ucar.edu

CISL Research Data Archive
Managed by NCAR's Data Support Section
Data for Atmospheric and Geosciences Research

Home | Find Data | Ancillary Services | About/Contact | Data Citation | Web Services | For Staff

Look For Data:

All Datasets	Variable/Parameter	Type of Data
Time Resolution	Platform	Spatial Resolution
Topic/Subtopic	Project/Experiment	Supports Project
Data Format	Location	Recently Added/Updated

Recently Added Datasets: (within the last 6 months)
- NCEP GFS 0.25 Degree Global Forecast Auxiliary "b" Grids Historical Archive
- Cloud Properties from ISCCP and PATMOS-x Corrected for Spurious Variability Related to Changes in Satellite Orbits, Instrument Calibrations, and Other Factors
- NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive
- NOAA CPC Morphing Technique (CMORPH) Global Precipitation Analyses Version 0.x (June 2014 - current)
- NCAR CESM Global Bias-corrected CMIP5 Output to Support WRF/MPAS Research
- ERA-20C Project (ECMWF Atmospheric Reanalysis of the 20th Century)
- WCRP and WWRP THORPEX YOTC (Year of Tropical Convection) Project, Single Parameter 3-Hourly Pressure Level Forecast Time Series, Transformed to a Regular 1600 by 800 (N400) Gaussian Grid, Dynamical Parameters Only

Other Ways to Explore:
- GCMD Topic:
  Agriculture • Atmosphere • Biosphere • Climate Indicators • Cryosphere • Hydrosphere • Land Surface • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • Sun-earth Interactions

Get Help:
- Frequently Asked Questions
- Reset your password
- A-Z Site Index
- RDA Users Email List
- Email Us

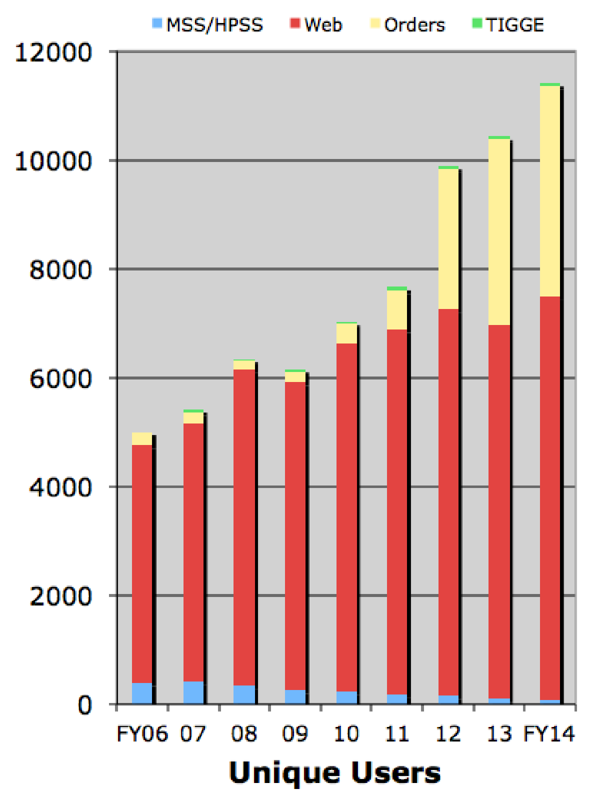
Recent News:

Guidance to WRF users for new NCEP GFS and FNL GRIB2 files
rdahelp@ucar.edu has fielded many requests for assistance from WRF users attempting to use the new NCEP ...

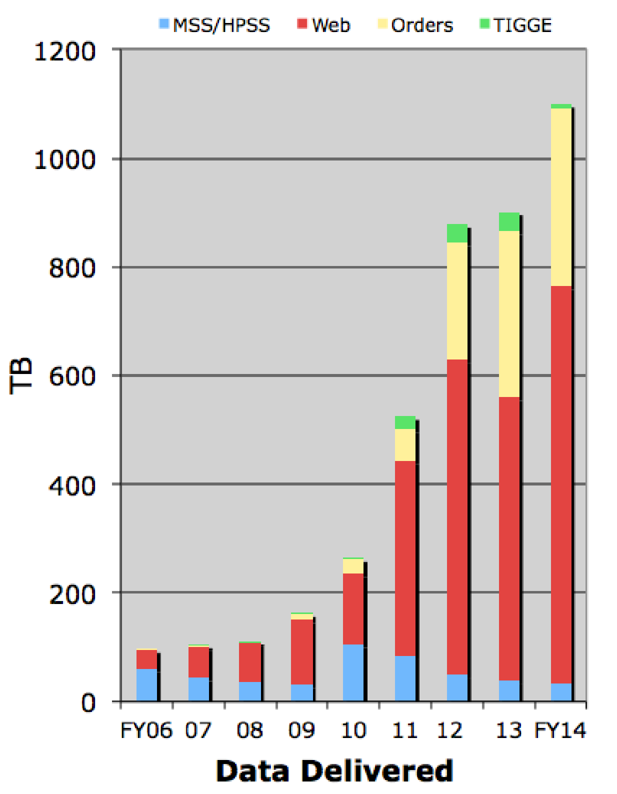
New Dataset: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive
The NCAR RDA now includes a historical archive of the operational NCEP GFS 0.25 degree ...

NCEP FNL dataset impacted by recent GFS overhaul
Effective January 14, 2015, NCEP made some changes to their Global Forecast System (GFS). This impacts ds083.2 ...

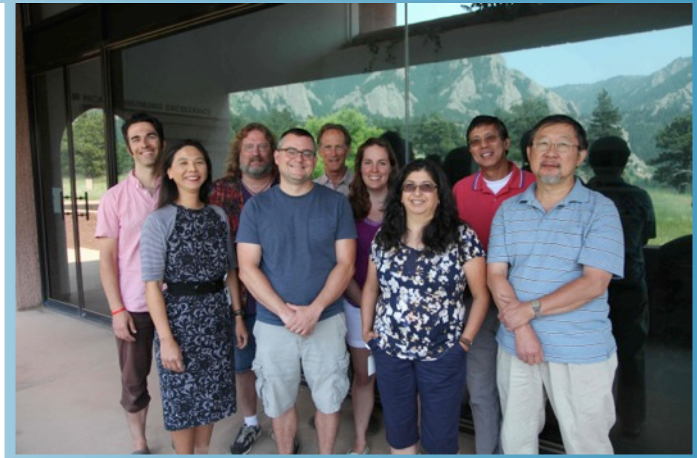
Archive ...



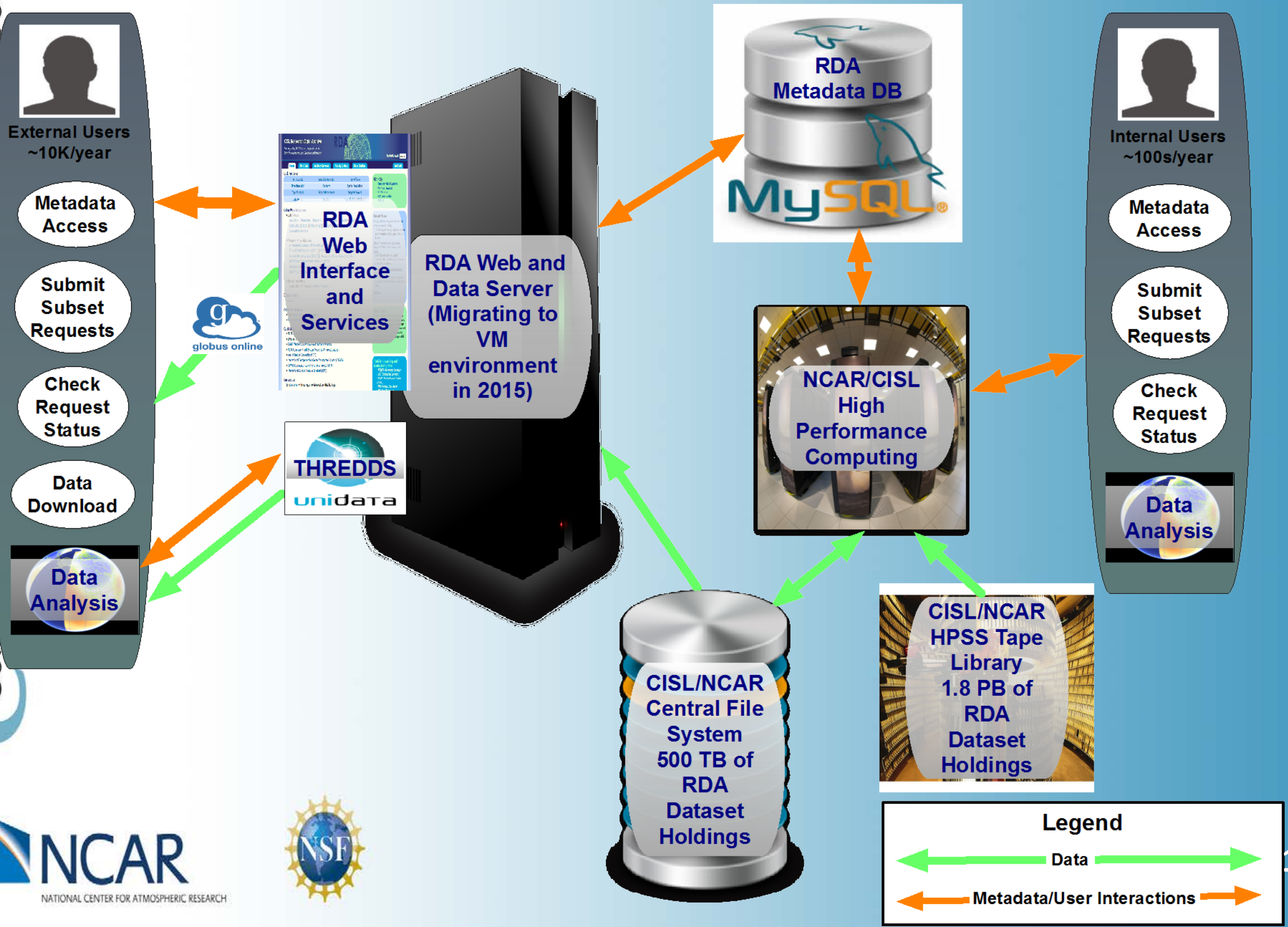
Unique Users



Data Delivered



Current RDA architecture



External Users ~10K/year

Metadata Access

Submit Subset Requests

Check Request Status

Data Download

Data Analysis

RDA Web Interface and Services

globus online

THREDDS

unidata

RDA Web and Data Server (Migrating to VM environment in 2015)

RDA Metadata DB

MySQL

NCAR/CISL High Performance Computing

CISL/NCAR Central File System 500 TB of RDA Dataset Holdings

CISL/NCAR HPSS Tape Library 1.8 PB of RDA Dataset Holdings

Internal Users ~100s/year

Metadata Access

Submit Subset Requests

Check Request Status

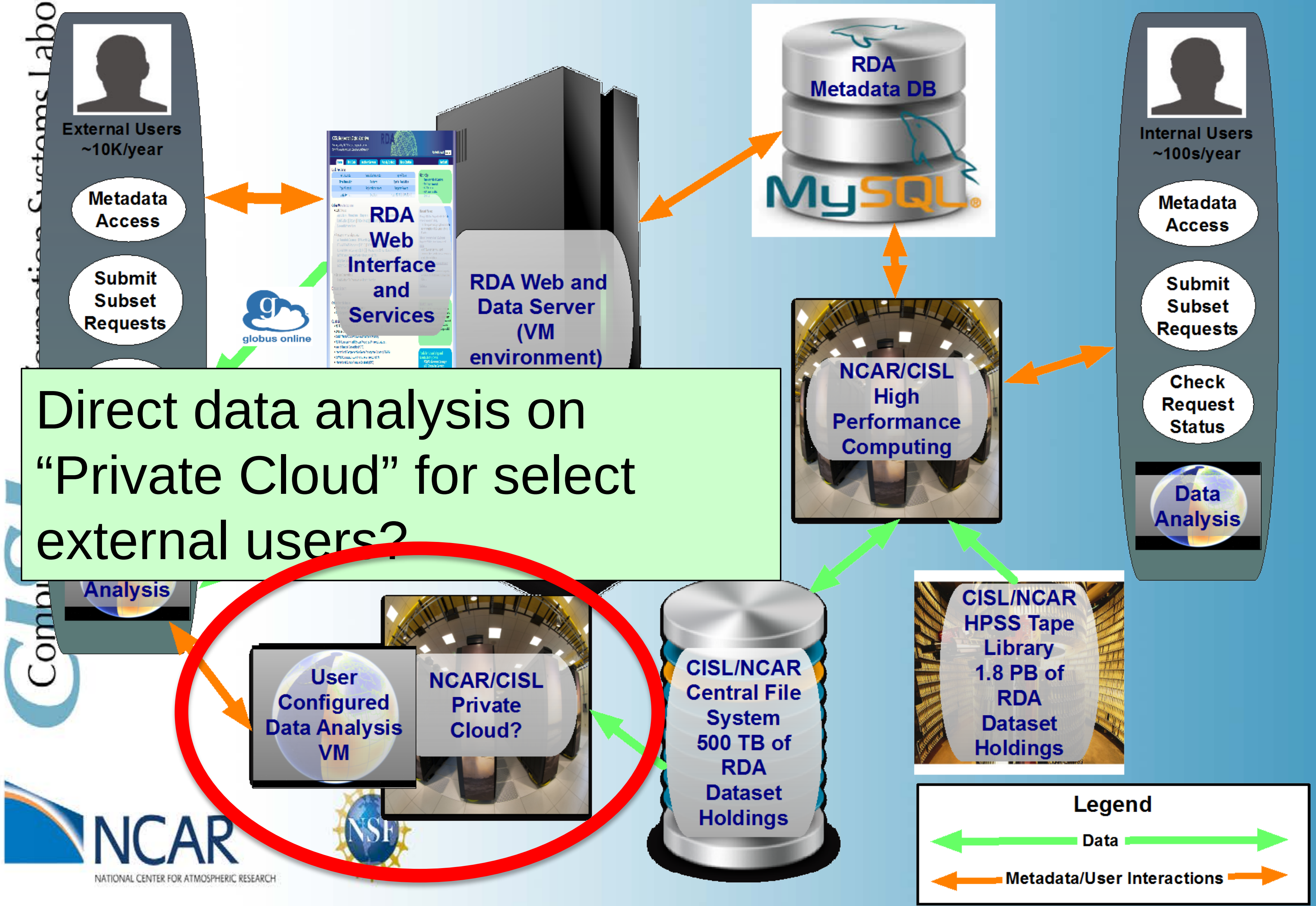
Data Analysis

Legend

Data

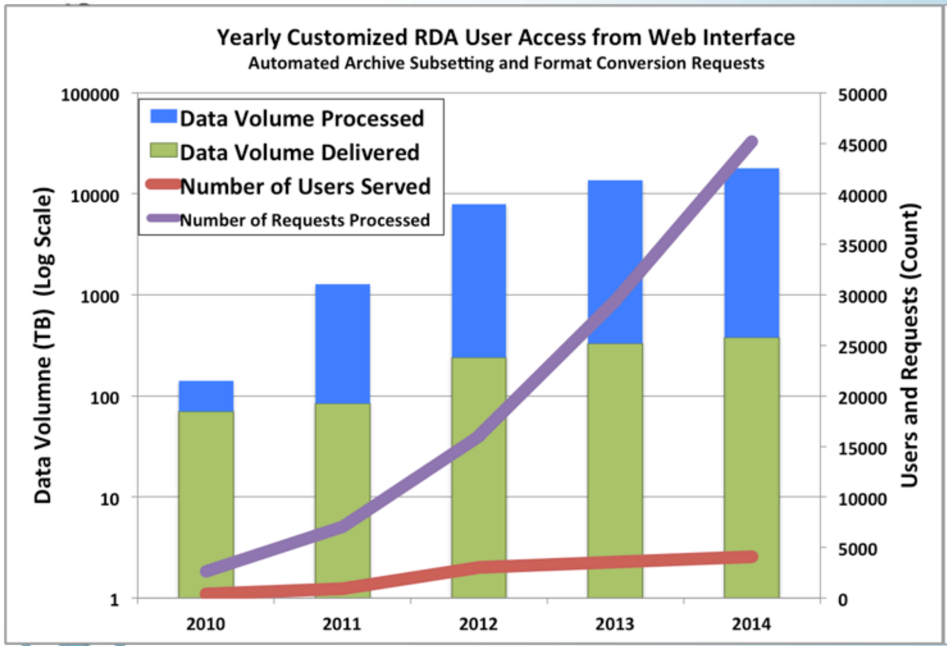
Metadata/User Interactions

Private Cloud: Future RDA/NCAR architecture?

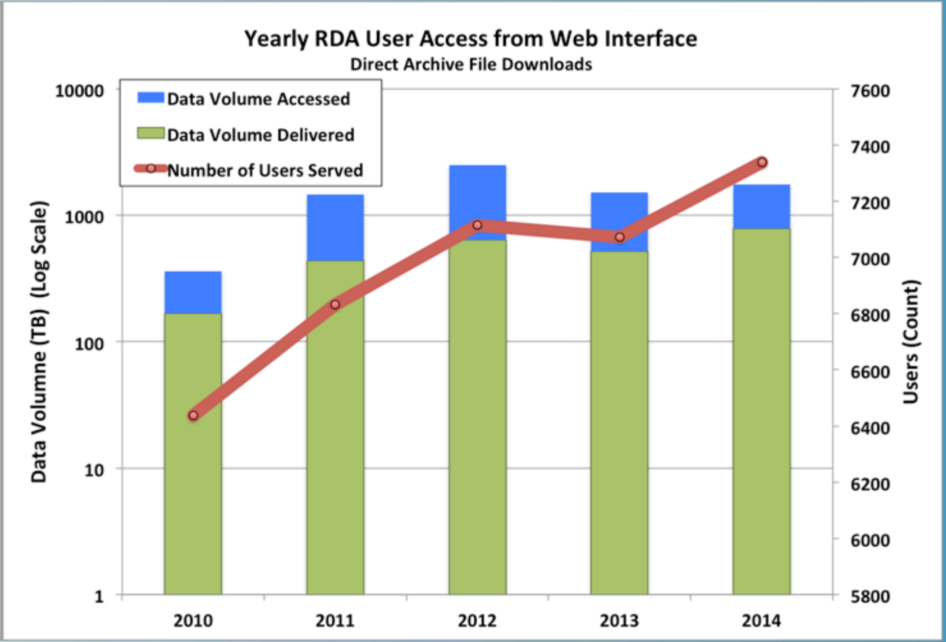

External Users ~10K/year
Metadata Access
Submit Subset Requests
Analysis
RDA Web Interface and Services
globus online
RDA Web and Data Server (VM environment)
RDA Metadata DB
MySQL
Internal Users ~100s/year
Metadata Access
Submit Subset Requests
Check Request Status
Data Analysis
NCAR/CISL High Performance Computing
CISL/NCAR HPSS Tape Library 1.8 PB of RDA Dataset Holdings
CISL/NCAR Central File System 500 TB of RDA Dataset Holdings
User Configured Data Analysis VM
NCAR/CISL Private Cloud?
Legend
Data
Metadata/User Interactions

Direct data analysis on “Private Cloud” for select external users?

RDA Usage metrics (1/2)


Yearly Customized RDA User Access from Web Interface

Automated Archive Subsetting and Format Conversion Requests


Yearly RDA User Access from Web Interface

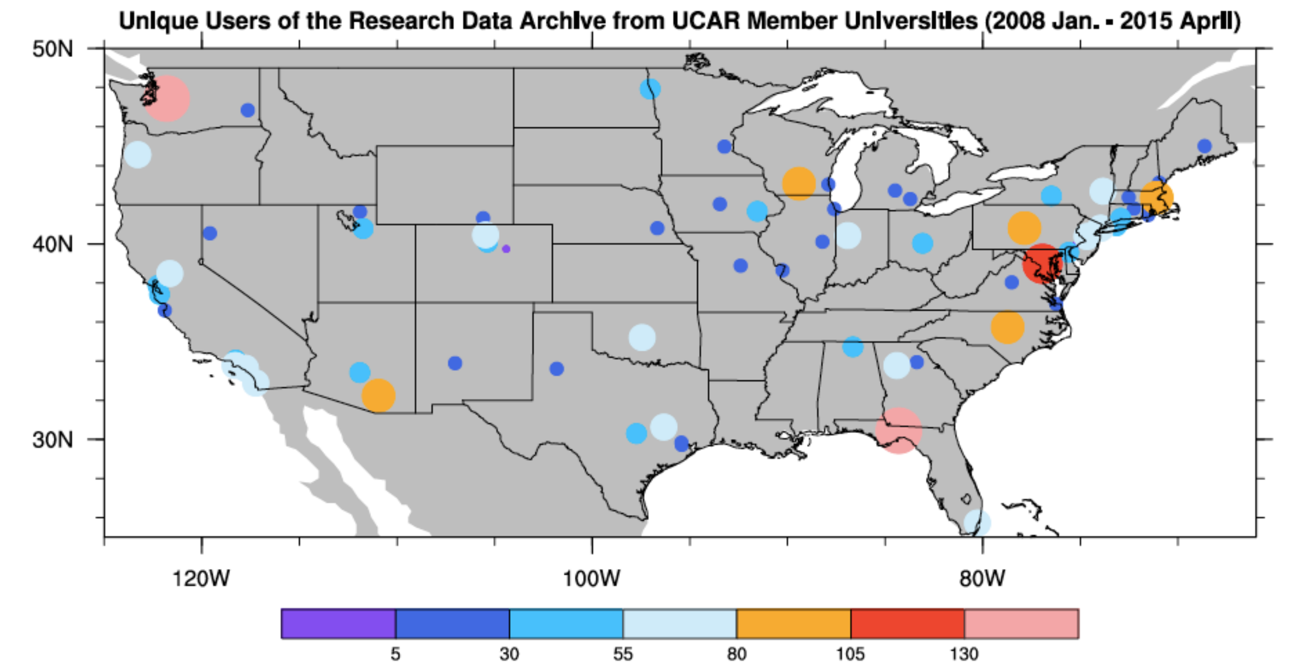
Direct Archive File Downloads

In 2014

- 17+ PB (virtual) processing
- 4000 users received data subsetting and format conversion services
- 45K requests were processed and 380 TB of data delivered
- 7300 users downloaded 750TB of data from archive files

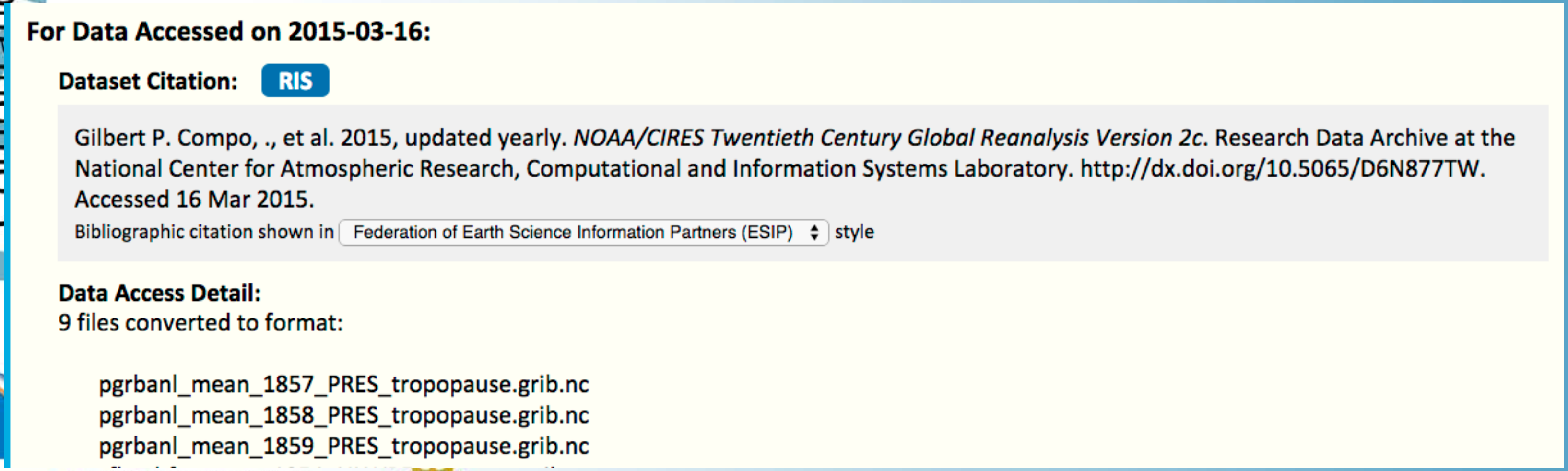
2014 top 10 RDA datasets, ranked by number of unique users.

Dataset Title	Rank	Unique Users	Amount (TB)
NCEP FNL Global Tropospheric Analyses, 1999 - ongoing	1	4530	146
NCEP Climate Forecast System Global Reanalysis, 1979 - ongoing	2	3156	505
NCEP Global Upper Air and Surface Weather Obs. 1997 - ongoing	3	943	33
ECMWF-Interim Global Reanalysis, 1979 - ongoing	4	490	137
NCEP/NCAR Global Reanalysis, 1948 - ongoing	5	473	6
Japanese 55-year Global Reanalysis, 1957 - ongoing	6	381	169
NCEP North American Regional Reanalysis, 1979 - ongoing	7	361	54
Historical Unidata (IDD) Gridded Model Data, 2002 - ongoing	8	348	29
International Comprehensive Ocean Atmosphere Data Set (ICOADS)	9	268	1
	10	239	3


Unique Users of the Research Data Archive from UCAR Member Universities (2008 Jan. - 2015 April)

RDA DOI support

- **Mint DOIs using EZID and managed by DataCite**
- **Management requirements set from 6 RDA use cases**
  - http://rda.ucar.edu/#!data-citation/use-cases
- **DOIs currently assigned to 49 RDA dataset collections**
- **Formal data citation in publications are supported with dashboard tool that generates citation text – see below.**



For Data Accessed on 2015-03-16:

Dataset Citation: RIS

Gilbert P. Compo, ., et al. 2015, updated yearly. *NOAA/CIRES Twentieth Century Global Reanalysis Version 2c*. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://dx.doi.org/10.5065/D6N877TW. Accessed 16 Mar 2015.

Bibliographic citation shown in Federation of Earth Science Information Partners (ESIP) style

Data Access Detail:
9 files converted to format:

pgrbanl_mean_1857_PRES_tropopause.grib.nc
pgrbanl_mean_1858_PRES_tropopause.grib.nc
pgrbanl_mean_1859_PRES_tropopause.grib.nc

RDA Globus integration

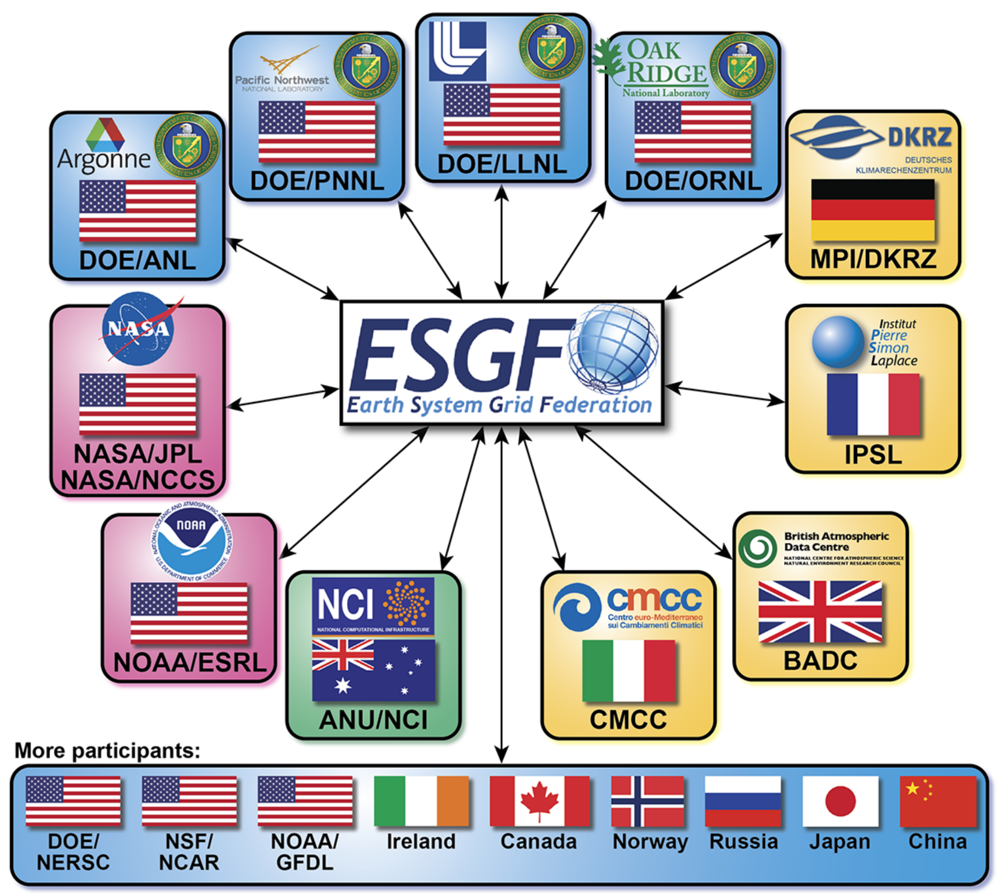
- Programmatic Globus endpoint shares
- Managed endpoints
  - general archive access (Starting 4/28/15)
  - custom data orders (e.g., subset requests)
- Access managed via Globus CLI and API
- Single sign-on to Globus via RDA account credentials

Usage: Nov 2014 – current, only data orders

- **49 users**
- **14 TB transferred**

ESGF and CMIP

- **ESGF – Earth System Grid Federation**
  - international collaboration for the software that powers most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC).
- **CMIP – Coupled ModeL Intercomparison**
  - protocols enable the periodic assessments carried out by the IPCC
  - Big Data!!!
    - CMIP5 – 2PB distributed across planet
    - CMIP6 (coming in ~2018) – 20PB



Pacific Northwest National Laboratory
DOE/PNNL
DOE/LLNL
Oak Ridge National Laboratory
DOE/ORNL
Argonne
DOE/ANL
DKRZ Deutsches Klimarechenzentrum
MPI/DKRZ
NASA
NASA/JPL
NASA/NCCS
ESGF
Earth System Grid Federation
Institut Pierre Simon Laplace
IPSL
NOAA
NOAA/ESRL
NCI National Computational Infrastructure
ANU/NCI
cmcc Centro euro-Mediterraneo sui Cambiamenti Climatici
CMCC
British Atmospheric Data Centre
BADC
More participants:
DOE/ NERSC
NSF/ NCAR
NOAA/ GFDL
Ireland
Canada
Norway
Russia
Japan
China

†Additional participants could not be illustrated in this figure.

CMIP Analysis Platform

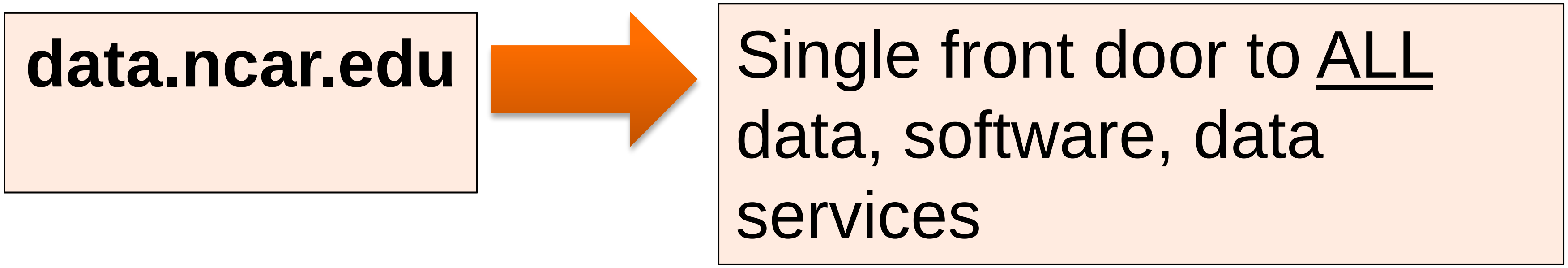
- **For big data – *"Bring compute to the data, not data to compute"***
- **Challenges:**
  - University researchers don't have infrastructure to download data and analyze
  - Takes weeks-to-months to download data.
- **Creating *"NCAR CMIP Analysis Platform"***
  - House high-value CMIP5 (and soon CMIP6) data on GLADE
  - Also download and manage *per-request* data on GLADE
  - Advertise to US researchers, make allocations, create user accounts
  - Provide Analysis platforms (with Matlab, IDL, NCL, etc) attached to high-speed disk.

Working at NCAR for Comprehensive Cross-organizational Data Management

The

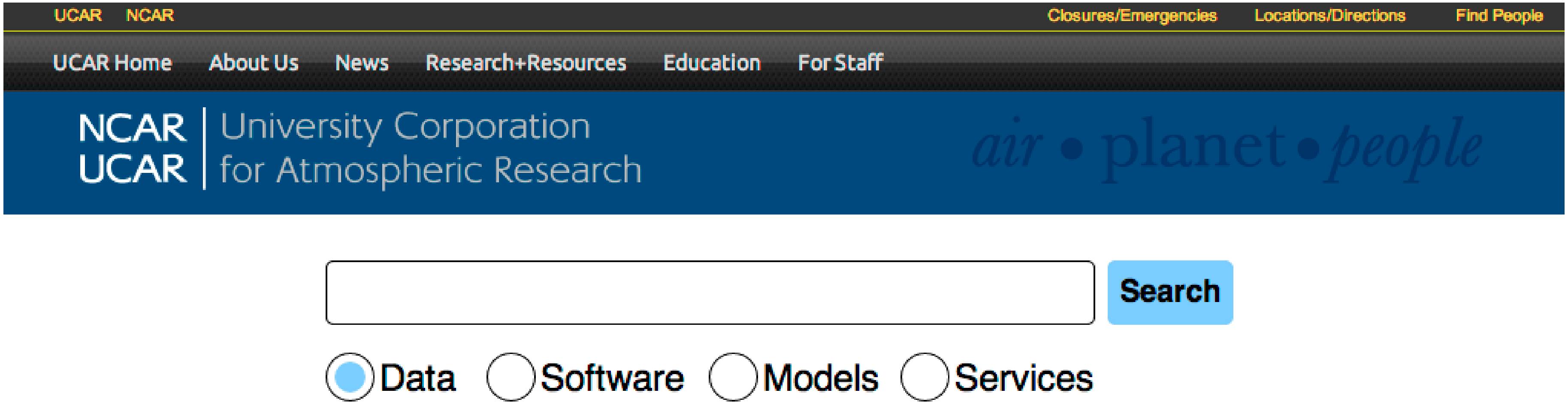
Data Stewardship Engineering Team (DSET)

The vision:



data.ncar.edu

Single front door to ALL data, software, data services

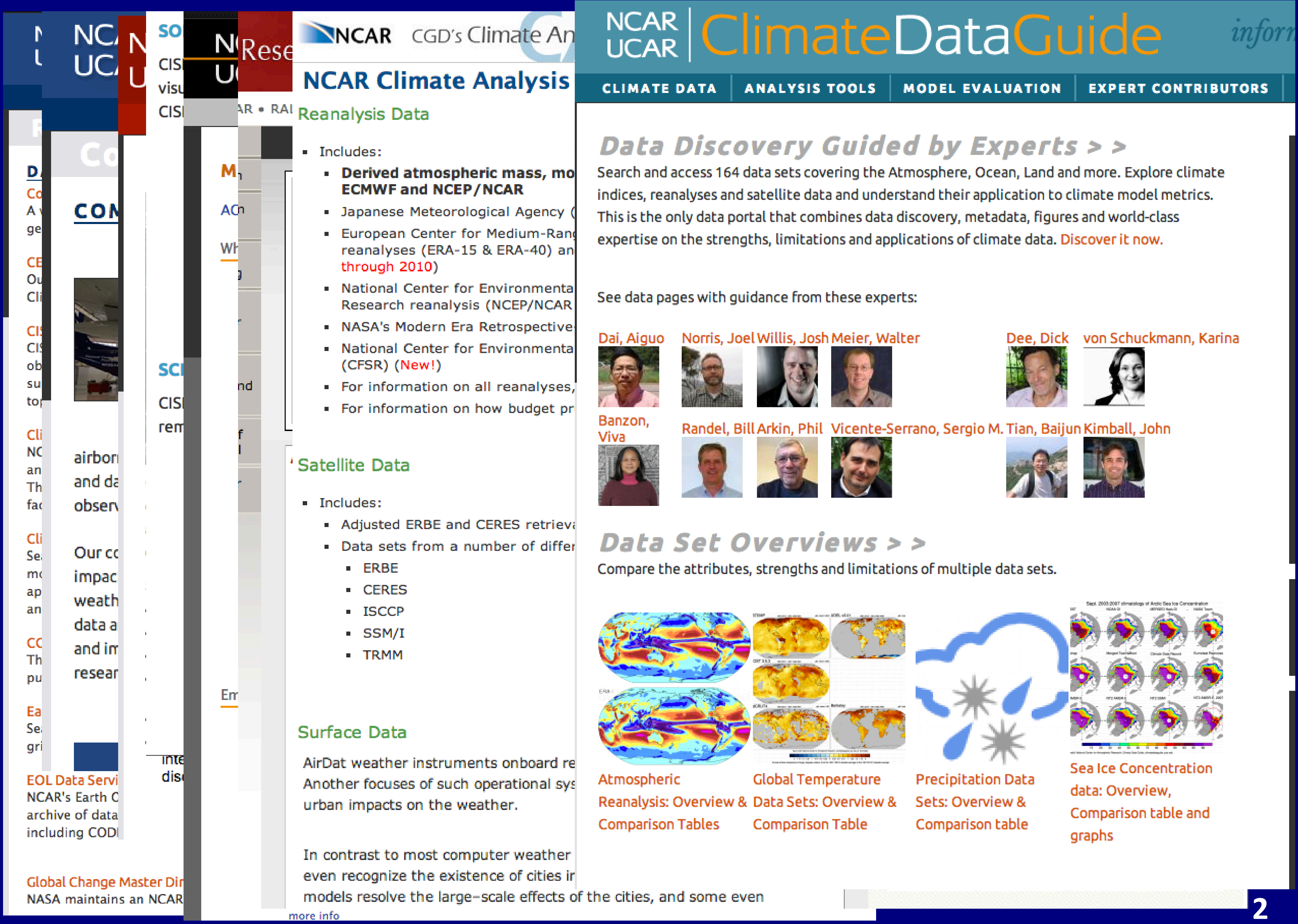


● Data ○ Software ○ Models ○ Services

Definition

Data

Digital assets intended for scientific community use, including files and metadata, publications, reports, images, software (visualization, analysis, model codes), and related data services.



NCAR UCAR ClimateDataGuide

CLIMATE DATA | ANALYSIS TOOLS | MODEL EVALUATION | EXPERT CONTRIBUTORS

Data Discovery Guided by Experts > >

Search and access 164 data sets covering the Atmosphere, Ocean, Land and more. Explore climate indices, reanalyses and satellite data and understand their application to climate model metrics. This is the only data portal that combines data discovery, metadata, figures and world-class expertise on the strengths, limitations and applications of climate data. Discover it now.

See data pages with guidance from these experts:

Dai, Aiguo · Norris, Joel · Willis, Josh · Meier, Walter · Dee, Dick · von Schuckmann, Karina · Banzon, Viva · Randel, Bill · Arkin, Phil · Vicente-Serrano, Sergio M. · Tian, Baijun · Kimball, John

Data Set Overviews > >

Compare the attributes, strengths and limitations of multiple data sets.

Atmospheric Reanalysis: Overview & Comparison Tables

Global Temperature Data Sets: Overview & Comparison Table

Precipitation Data Sets: Overview & Comparison table

Sea Ice Concentration data: Overview, Comparison table and graphs

NCAR CGD's Climate An

NCAR Climate Analysis

Reanalysis Data

- Includes:
  - **Derived atmospheric mass, mo ECMWF and NCEP/NCAR**
  - Japanese Meteorological Agency (
  - European Center for Medium-Rang reanalyses (ERA-15 & ERA-40) an through 2010)
  - National Center for Environmenta Research reanalysis (NCEP/NCAR
  - NASA's Modern Era Retrospective
  - National Center for Environmenta (CFSR) (New!)
  - For information on all reanalyses,
  - For information on how budget pr

Satellite Data

- Includes:
  - Adjusted ERBE and CERES retrieva
  - Data sets from a number of differ
    - ERBE
    - CERES
    - ISCCP
    - SSM/I
    - TRMM

Surface Data

AirDat weather instruments onboard re Another focuses of such operational sys urban impacts on the weather.

In contrast to most computer weather even recognize the existence of cities ir models resolve the large-scale effects of the cities, and some even

more info

EOL Data Servi
NCAR's Earth O
archive of data
including CODI

Global Change Master Dir
NASA maintains an NCAR

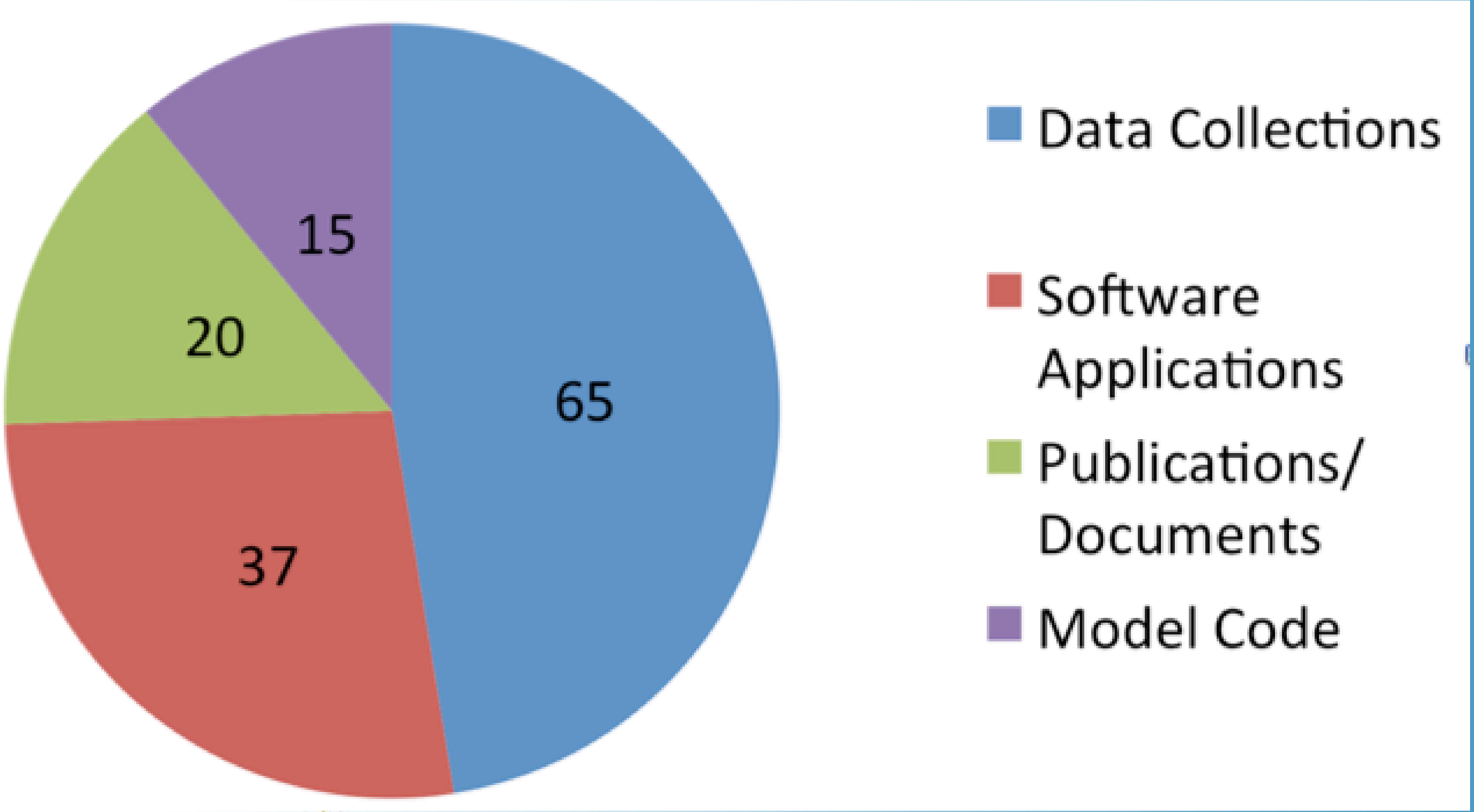
www2.ucar.edu/research-resources/data-archive-services

Goals

- Inventory all community-based digital assets
- Assess feasibility and resource requirements
- Establish the technical and organizational foundation

Inventory of NCAR digital collections

Collection = a group of assets managed by a system, e.g. RDA is one Data Collection



Data Collections: observations, images, model output
Software Applications: data processing and analysis

Challenges for success

- **Level of preparedness varies greatly across NCAR**
  - Especially, regarding metadata standards
- **Need more data engineering expertise**
- **Need organization-wide digital asset consulting**
- **Need sustainable and cost effective technical solutions**

Stepping Up to the Challenge

We are:

- Reviewing, testing, and prototyping available technologies
- Making a comprehensive metadata needs assessment, by experts, across NCAR

Computational & Information Systems Laboratory

Questions?

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NSF