Globus for Research Data Management

Rachana Ananthakrishnan
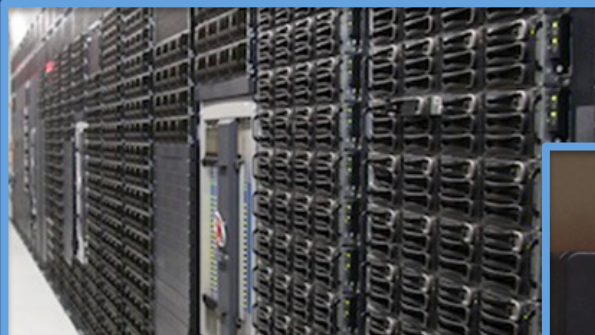Product Manager, Globus
ranantha@uchicago.edu

August 13 2015

globus

Globus delivers…

(big) data transfer, sharing, publication, and discovery…
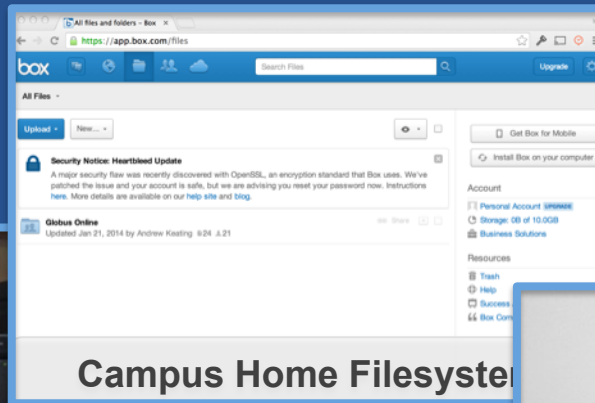
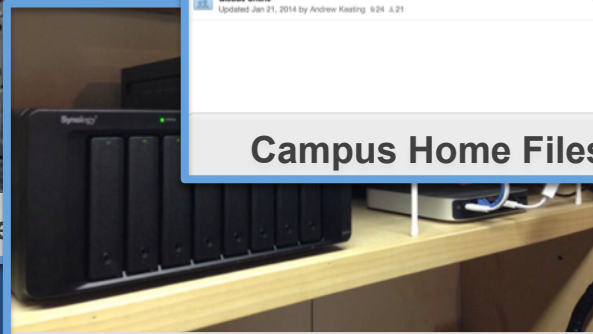…directly from your own storage systems

"I need to easily, quickly, & reliably move or replicate my data to other systems."

Research Computing HPC Cluster

Campus Home Filesystem

Lab Server

Personal Laptop

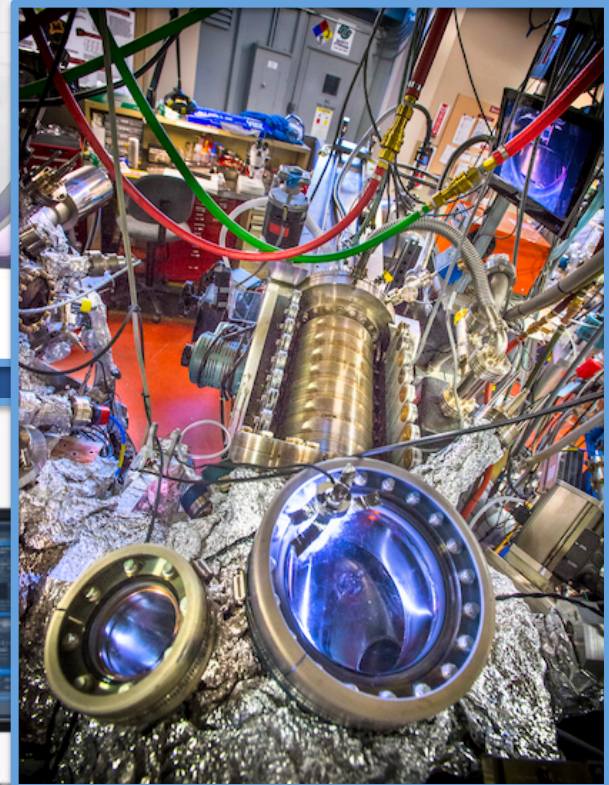**Desktop Workstation**

XSEDE Resource

Public Cloud

"I need to get data from a scientific instrument to my analysis system."

MRI

Advanced Light Source

Next Gen Sequencer

Light Sheet Microscope
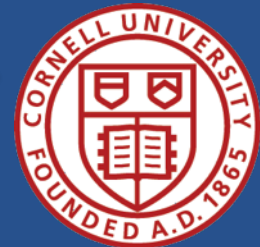
"I need to easily and securely share my data with my colleagues at other institutions."

"I need a good place to store / backup / archive my (big) research data, at a reasonable price."

Campus Store

Mass Store

Public Cloud Archive

"I need to publish my data so that others can find it and use it."

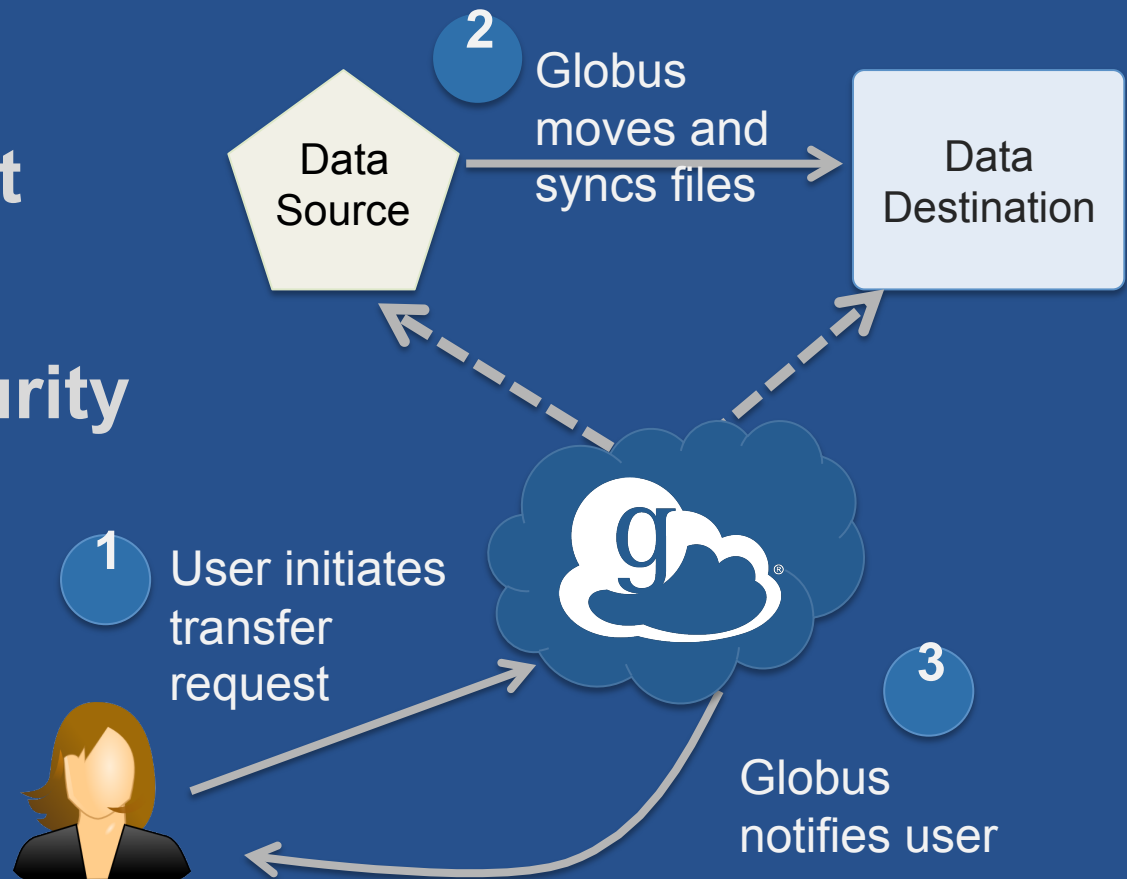Reference Dataset

Scholarly Publication

Active Research Collaboration

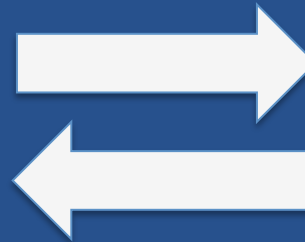# Reliable, secure, high-performance *file transfer and replication*

- **"Fire-and-forget" transfers**

- **Automatic fault recovery**

- **Seamless security integration**

- **Powerful GUI and APIs**

2 Globus moves and syncs files

Data Source → Data Destination

1 User initiates transfer request
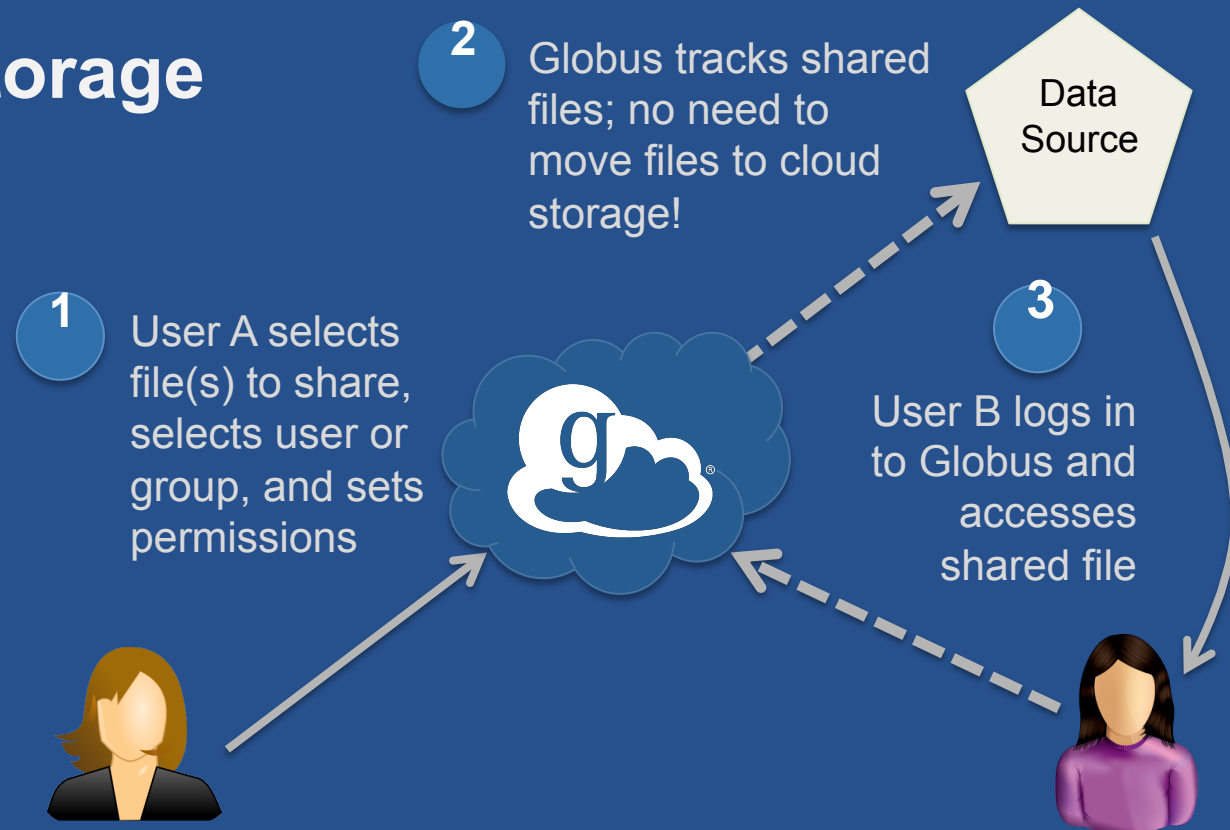
3 Globus notifies user

Amazon S3 Endpoints

Demonstration:
Sign Up
File Transfer

Simple, secure *sharing* off existing storage systems

- **Easily share large data with any user or group**

- **No cloud storage required**

2 Globus tracks shared files; no need to move files to cloud storage!

Data Source

1 User A selects file(s) to share, selects user or group, and sets permissions

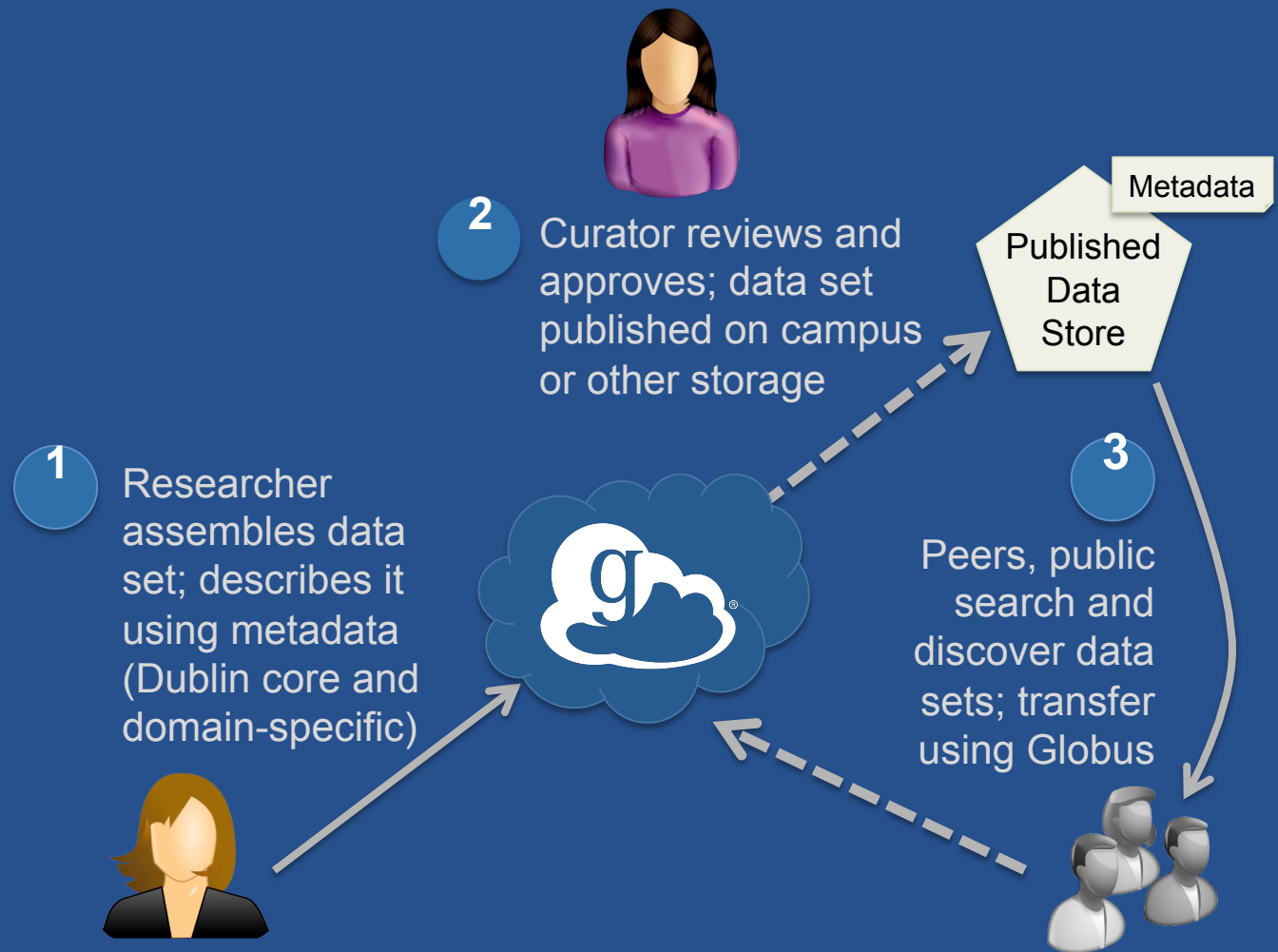**3** User B logs in to Globus and accesses shared file

11

Demonstration:
Globus sharing
Group Management

Curated *publication* of data, with relevant metadata for *discovery*

- **Identify**

- **Describe**

- **Curate**

- **Verify**

- **Access**

- **Preserve**

2 Curator reviews and approves; data set published on campus or other storage

Metadata

Published Data Store

1 Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

3 Peers, public search and discover data sets; transfer using Globus

Typical publication workflow

Globus Publish

Globus Publish enables easy and efficient publishing of large datasets. It uses Globus as a mechanism to transfer and share data.

Learn More

Search   Publish   Manage Data   Groups   Support   Log In

Univ. of Chicago   Argonne   IIT   UIUC

5. Search

6. Download

1. Publish Data

2. Describe Submission

Scientist

3. Assemble Dataset (Transfer Data)

Shared Endpoint

4. Curate Dataset ✔

Argonne Storage System

Argonne Curator

14

Publication GA release

Identifier

URL   Handle   DOI

Description

None   Standard   Domain-specific   Custom

Curation

None   Acceptance   Human-validated   Machine-validated

Access

Anonymous   Public   Embargoed   Collaborators
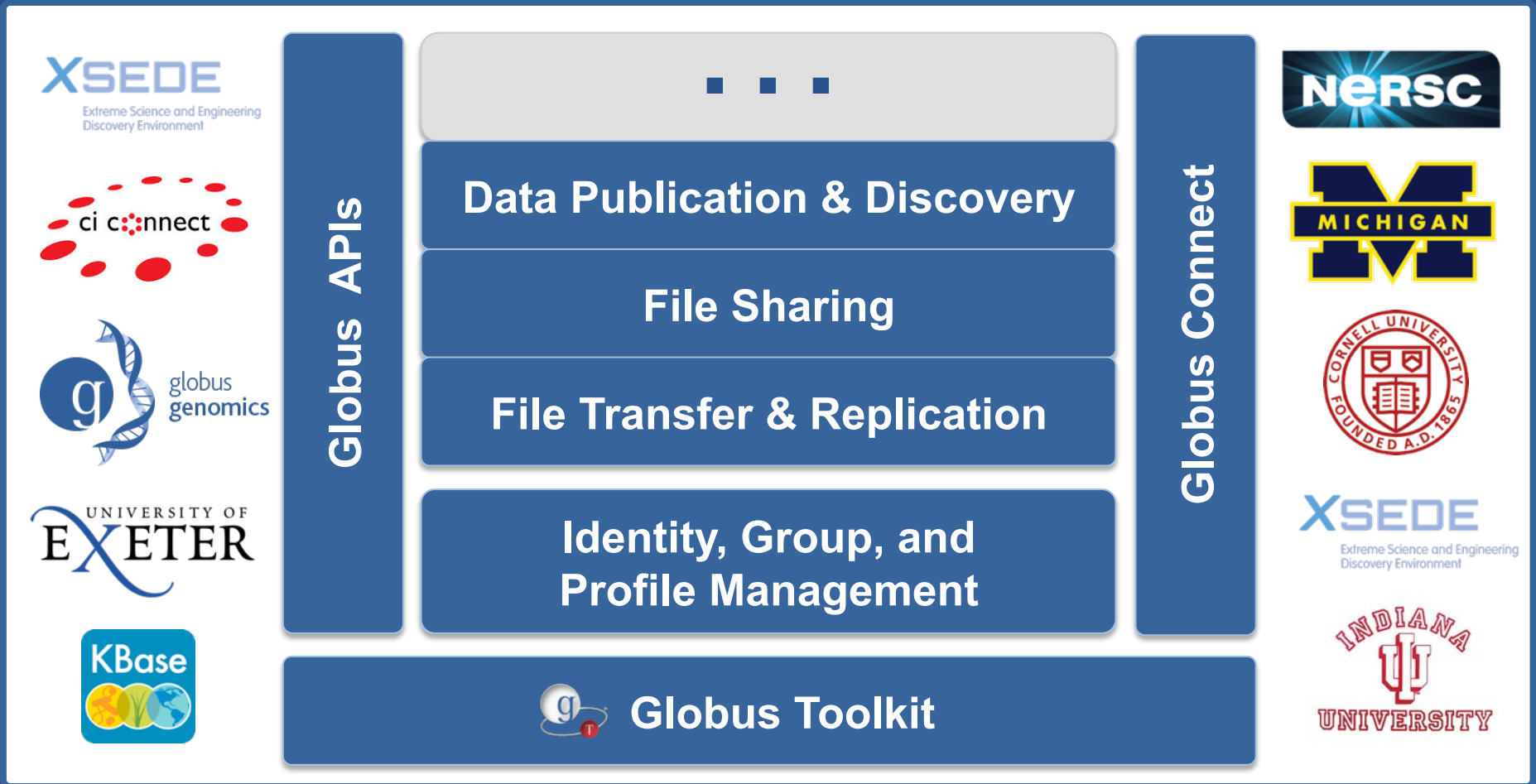
Preservation

Transient   Project Lifetime   Archive   "forever"

Globus Platform-as-a-Service

Globus APIs

Data Publication & Discovery

File Sharing

File Transfer & Replication

Identity, Group, and Profile Management

Globus Toolkit

Globus Connect

Platform adopters

Globus at a glance

|			
4 major services	**100 PB** transferred	**20 billion** files processed	**25,000** registered users
13 national labs use Globus services	**8,000** active endpoints	**>300** users are active daily	**99.95%** uptime over the past two years
>30 institutions subscribed	**1 PB** largest single transfer to date	**3 months** longest continuously managed transfer	**157** federated campus identities

Globus endpoints worldwide
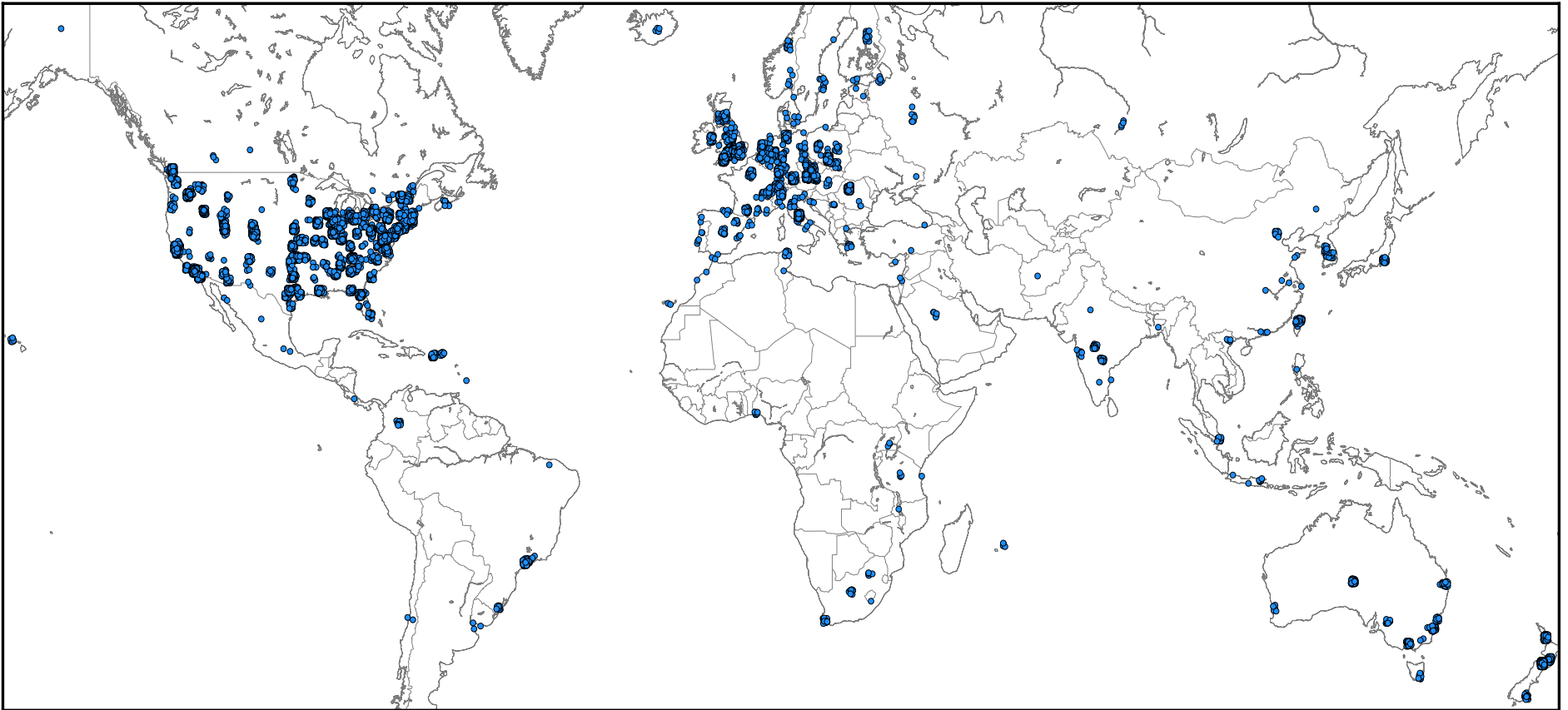
Globus is enabling…

Study of the structure and evolution of galaxies, the nature of dark energy, and cosmological history of the universe
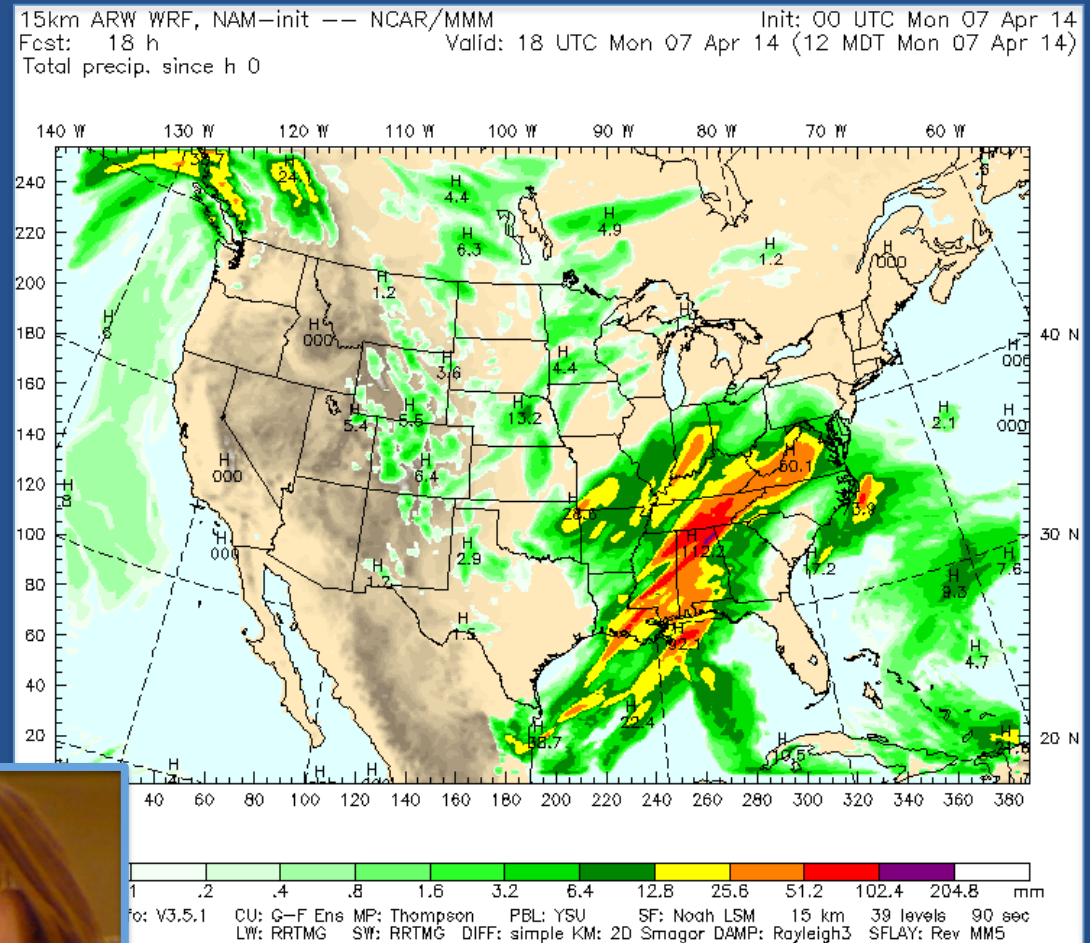
Joel Brownstein
University of Utah
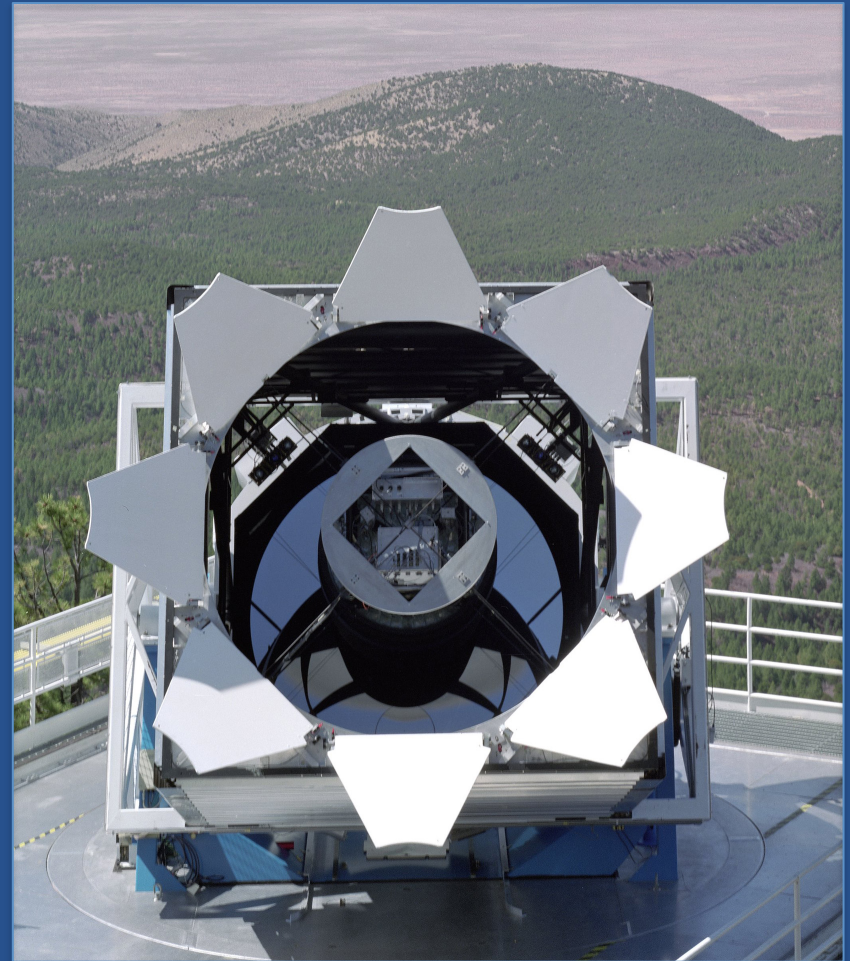
Sloan Digital Sky Survey
Source: University of Utah

Globus is enabling…

Development of numerical simulations of severe storms for improved responsiveness to weather events

Ann Syrowski
University of Illinois

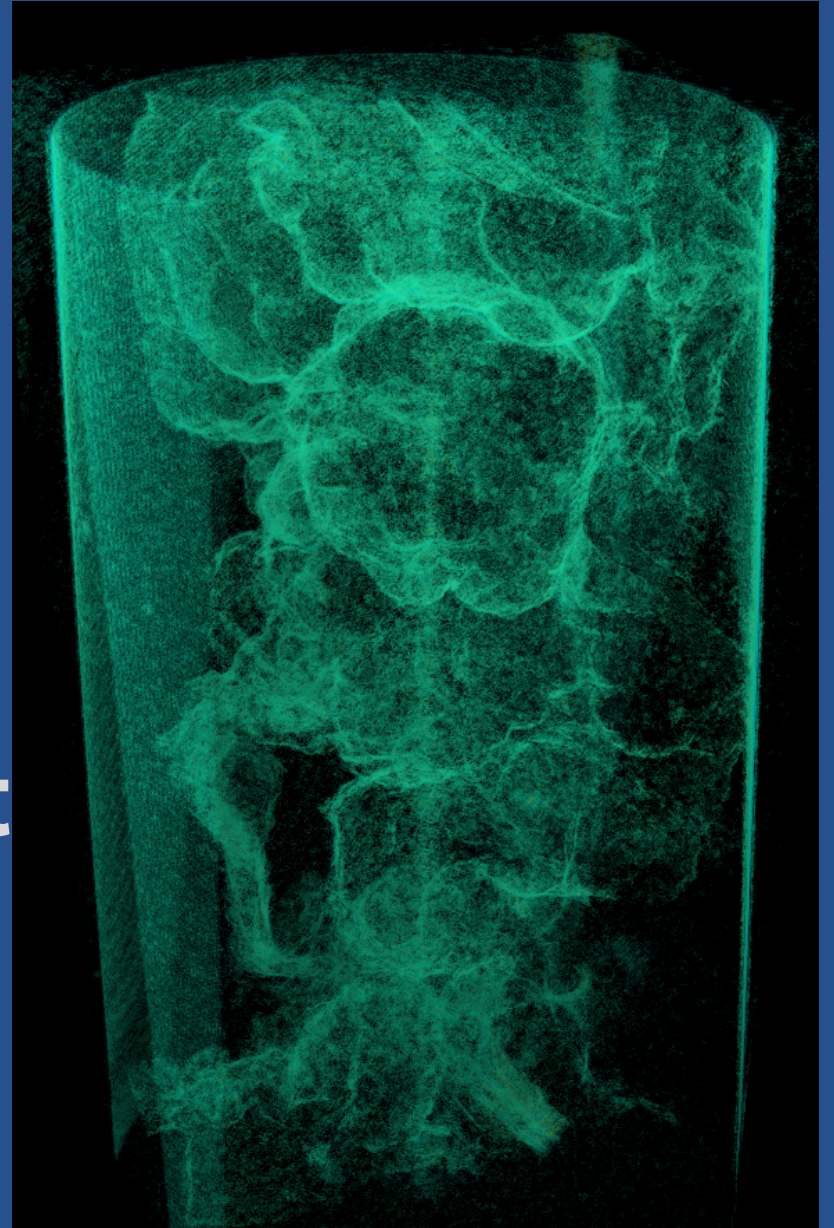Weather Research and Forecasting Model
Source: UCAR

Dan Kozak (Caltech) replicates 1 PB LIGO **astronomy** data for resilience

Erin Miller (PNNL) collects data at Advanced Photon Source, renders at PNNL, and views at ANL

Credit: Kerstin Kleese-van Dam

Globus Subscriptions

- **Provider**
  - Host shared endpoints
  - Management console
  - Data publication*
  - Amazon S3 endpoints
  - Usage reports
  - Priority support
  - Application integration

- **Branded Web Site**

- **Alternate Identity Provider (InCommon is standard)**

- **Mass Storage System optimization**

globus.org/provider-plans

Questions

More information:

www.globus.org

www.globus.org/data-publication

Contact:

support@globus.org

Hands-on exercise

http://tinyurl.com/globusatpesc