Biosciences Working Group Final Update for PRAGMA 25

Wilfred W. Li, Ph.D., UCSD, USA

Habibah Wahab, Ph.D., USM, Malaysia

Hosted by CNIC, CAS

Beijing, China Oct 16-18, 2013

Breakout Sessions

- Presentations (Day 1, 2:40 – 4:15 pm, Rm 514)
  - Active Folder
    - Daeyoung Heo, Kookmin University
  - PRAGMA workshop activities from Konkuk University
    - Jaebum Kim, Konkuk University
  - IDigBio: Integrated Digitized Biocollections
    - Andrea Matsunaga, U Florida
- Planning (Day 2, 11:10 am – 12:30 pm)
  - UCSD Research Cyberinfrastructure Program
    - Wilfred Li, UCSD
- Join Sessions (Day 2, 3:50 – 4:30 pm, Cyber Learning)
  - Group discussion

Breakout Session 1

- Introduction
- Registration
  - https://groups.google.com/forum/#!forum/pragma-biosciences-working-group
  - Search for "PRAGMA Biosciences Working Group" at groups.google.com
  - 36 members
- Other participants
  - Ngai Shing Mok, University of Hong Kong
  - Dong Hwan Lee, Kookmin University
  - Gyu Yeun Choi, Kookmin Univeristy

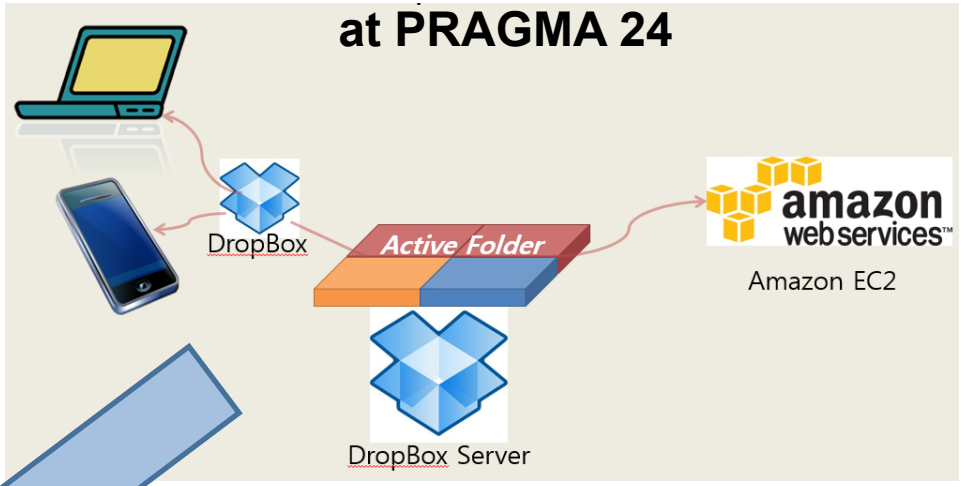
Active Folder: Integrating All Activities of Simulation on File System

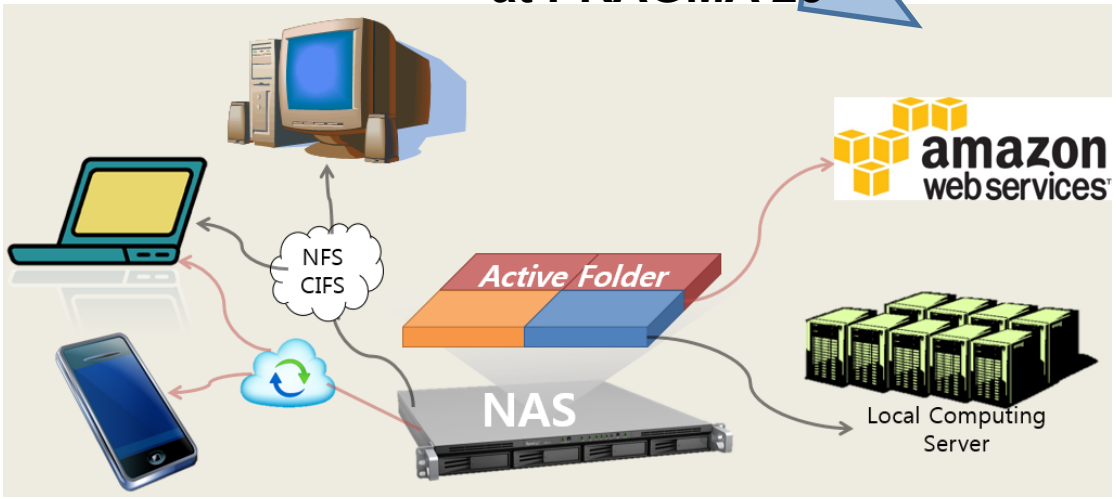
- Active Folder – good for case comparative study
  - Tasks
    - Described as regular folders and files
  - Product
    - Input or output of simulation
    - Can be handled like regular file by using legacy software
    - Contains provenance information (meta data, task info, etc)
    - Can be reproduced by the task which is extracted from the provenance information
  - Resource
    - Computing server(Local, Grid, Cloud, what ever, …) is registered as regular folders and files
    - To submit a Job(task), just Drag & Drop the task folder to the folder which represents computing server

Active Folder: Integrating All Activities of Simulation on File System

⊙Active Folder on DropBox+EC2

– Cost & Performance Problem with very large files





⊙**Active Folder on NAS**

⊙**NAS(Network Attached Storage)**

- **Large Volume Storage**
- **Network File System (NFS, SMB/CIFS, AFP …)**
- **Most vendors support Cloud solution like DropBox**

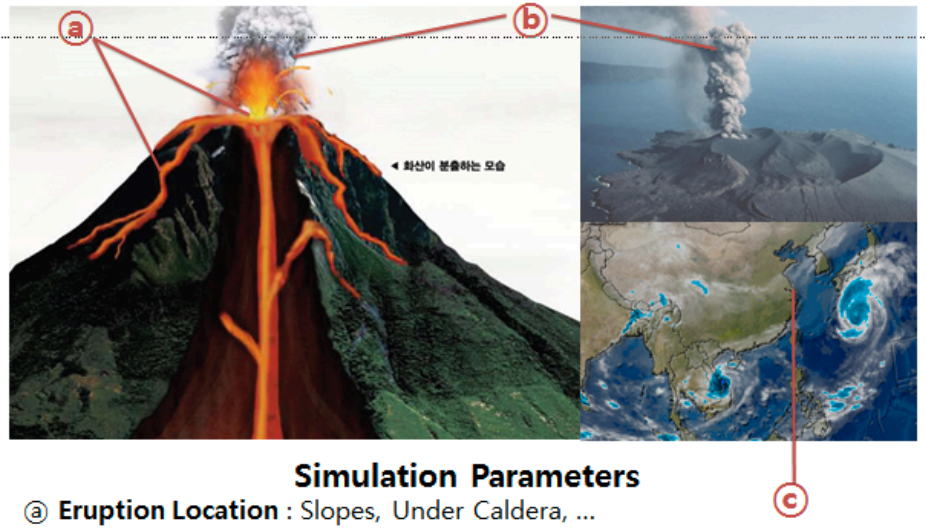
Case Study : Volcano Eruption Simulation using Active Folder

Volcanic Disaster Preparedness Research Center

Development of Preparedness Procedure and Technology for Volcanic Disaster in Korea.

Project #3:
Development of IT-based Response System for Volcanic Disaster

http://www.volcano.re.kr



Simulation Parameters

ⓐ **Eruption Location** : Slopes, Under Caldera, ...
ⓑ **Ash**
- Plume Style : Suzuki, Point, Plume or Resuspension
  - Height, Mass Flow Rate, ... (physics parameters)
- Ash Size : Granulometry, Distribution, ... or Density

ⓒ **Weather Condition**
- Wind field, Air temperature, ...

Available Actions

1. Search simulation results by parameters
2. Compare Results

Or

Reproducing by parameter sweeping

Simulation Procedure

WPS → WRF → Fall3D

Weather Simulations (WPS, WRF)

Volcano Eruption Simulations (Fall3D)

Simulation examples

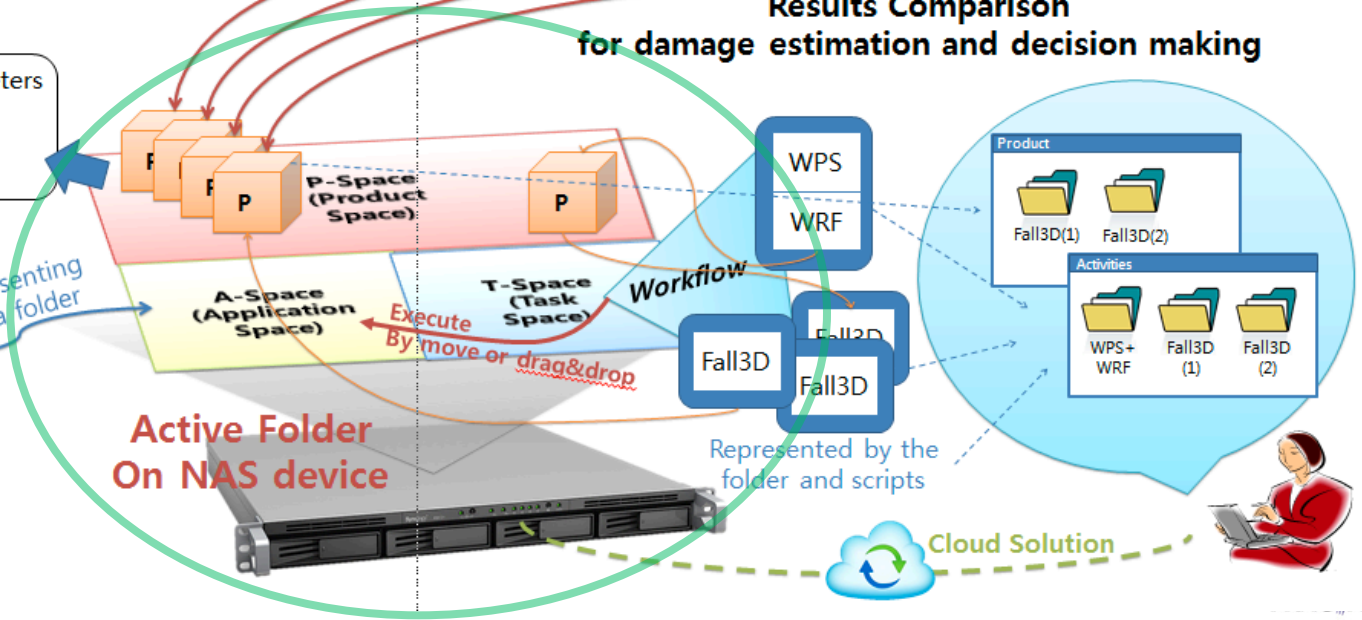
Parameters

Fall3D (1)	Fall3D (2)	Fall3D (3)	Fall3D (4) ...
ⓐ East slope ⓑ Plume: Suzuki Height: 3,000m MFR: 2 ~ 3 kg/s Ash size: **0.5um**	ⓐ East slope ⓑ Plume: Suzuki Height: 3,000m MFR: **3 ~ 4 kg/s** Ash size: 0.1um	ⓐ East slope ⓑ Plume: Suzuki Height: **8,000m** MFR: 2 ~ 3 kg /s Ash size: 0.1um	ⓐ **Under Caldera** ⓑ Plume: Suzuki Height: 8,000m MFR: 3 ~ 4 kg/s Ash size: 0.1um

Results Comparison for damage estimation and decision making

High Performance Computing

Representing as a folder



Update from Konkuk University

- Prepared a proposal for a government grant
  - Development of novel technologies for studying metagenomics based on cloud computing (Institutes: CBRU and BDRC at Konkuk University, SDSC and Calit2 at UCSD)
- Workshop proposal for PRAGMA26
  - Theme: NGS, Metagenomics, HPC, Clouds and Collaboration, CFP out early next year.

Jaebum Kim Ph.D. (jbkim@konkuk.ac.kr)

Update from Konkuk University

- Plan for an international consortium
  - Time: Jan. 2014 (tentative)
  - Place: Konkuk University, Korea
  - Topic: Environment- and toxicity-related microorganism and bioinformatics
  - Institutes: UW-METC (Dr. Yu), Konkuk University, and more (tentative)
  - More information will be out soon
  - If you are interested let us know. We can invite you (jbkim@konkuk.ac.kr)

A Computational- and Storage-Cloud for Integration of Biodiversity Collections

Andréa Matsunaga,
Alex Thompson,
Renato Figueiredo,
Charlotte Germain-Aubrey,
Matthew Collins,
Reed Beaman,
Bruce MacFadden,
Greg Riccardi,
Pamela Soltis,
Lawrence Page,
José A.B. Fortes

Supported by NSF Award EF-1115210

iDigBio
Integrated Digitized Biocollections

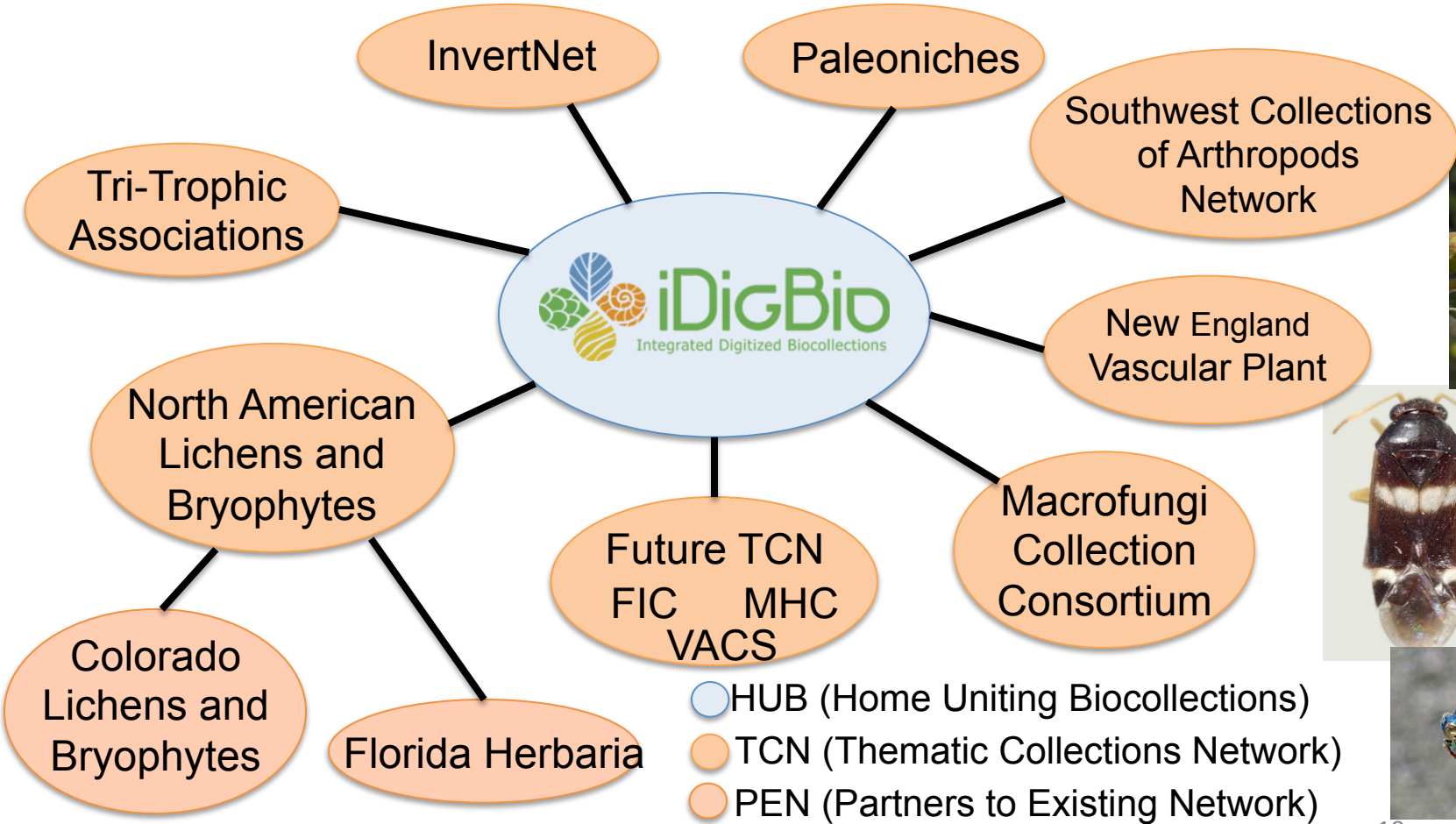
NSF

PRAGMA 25
Beijing, China
16-18, October, 2013

PRAGMA

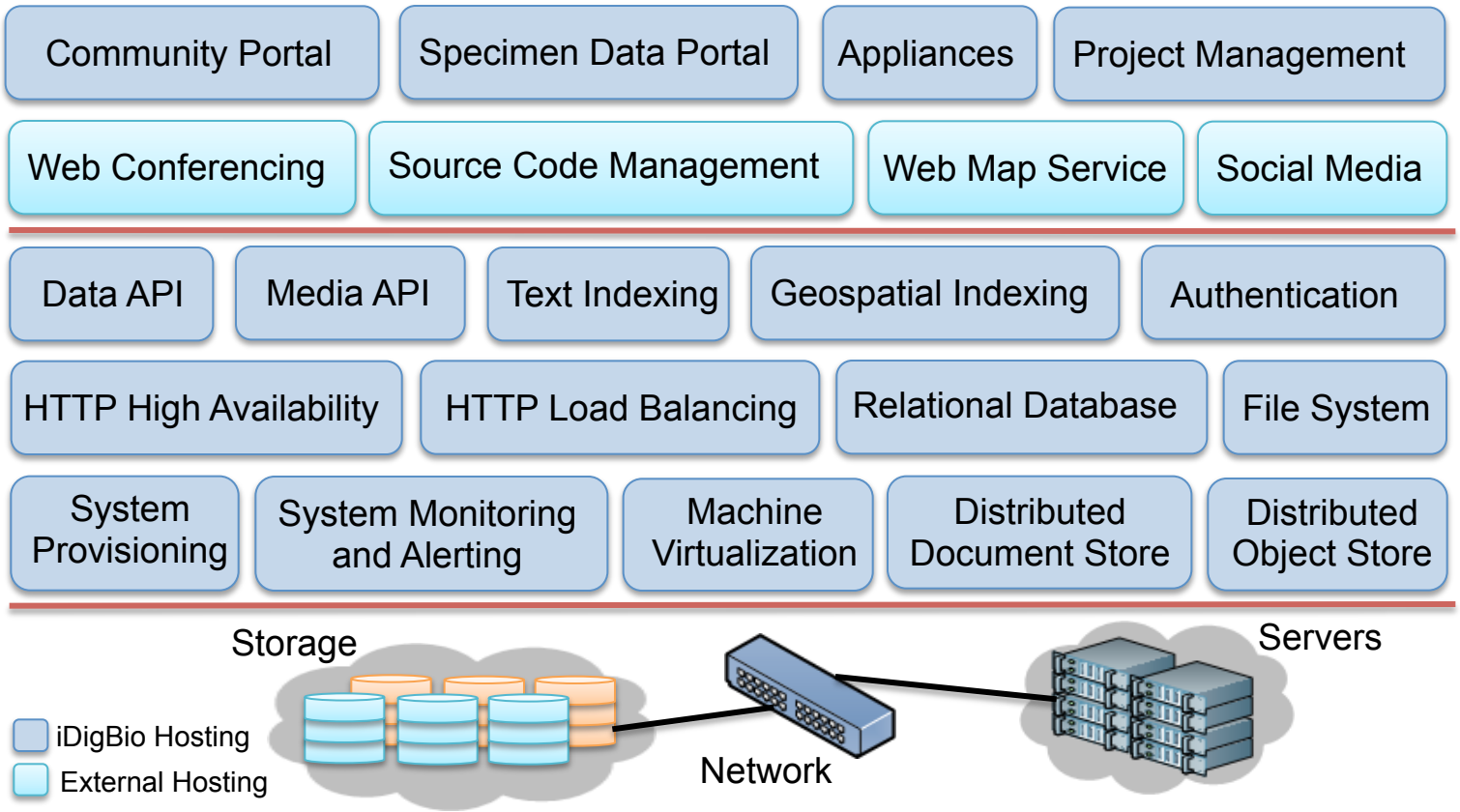
Integrated Digitized Biocollections (iDigBio)

- 10-year effort to digitize and mobilize the scientific information associated with vouchered specimens held in U.S. neontological and paleontological research collections



iDigBio Cyberinfrastructure

- Flexible to meet the diverse needs of TCNs
- Horizontally scalable to meet future demands to access the data
- Agile in taking advantage of and integrating proven open-source technologies, thus minimizing system development and maintenance risk
- Resilient to certain types of failures
- Based on standards where they exist to enable interoperability and reuse of tools, libraries, and services



Breakout Session 2

- Developing Sustainable Data Services in Cyberinfrastructure for Higher Education

How do You Handle Data Storage/Backup?

Table 2. Data Storage Devices and Services Utilized		
Type	***%***	**Primary purpose**
Network attached storage (NAS) devices	73	Standard performance network filesystem
USB Drives	70	Storage and backup
Local server hard disk drives	65	Storage and backup
Dropbox	33	Data sharing
SDSC Project Storage	13	Standard performance network filesystem
XSEDE Lustre Filesystem	10	Parallel filesystem
Google Drive	10	Storage and sharing
Amazon S3	8	Storage and sharing
SDSC Cloud Storage	8	Storage and sharing
Tape library	5	Storage and backup
Small Area Network Storage Array	3	Databases
CD/DVD	3	Storage and backup
Hadoop Filesystem	3	Replication and Map Reduce
iRODS	3	Metadata driven storage and sharing

Numbers reflect percentages of PIs surveyed that utilize each solution ; Individual PIs use multiple solutions, so %'s add up to >100%.

- Storage Devices
  - Network accessible storage (NAS), USB and server local drives dominate
  - Use of Dropbox for sharing
  - Others use Google Drive, Hadoop, XSEDE, SDSC co-location
  - Email attachment
- Backup modes
  - Replicated copies in two NAS
  - A copy in the NAS,
  - A copy in local hard drive (laptop/workstation),
  - And a copy in a USB drive
  - Maybe a copy in email/Dropbox
- Problems:
  - Out of sync
  - Lost track of its location
  - Lost version control
  - High cost of recovery
  - No metadata

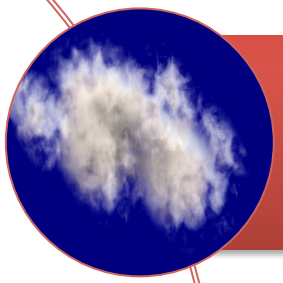
Table 4. Top 10 requirements for campus cyberinfrastructure			
Type	**%**	**Comments**	**Category**
Better CI with inimal direct cost	91	Least burden on research budget	Cost
Network Attached Storage	73	Shared POSIX compliant filesystem	Sharing
Data replication as backup	66	Keep a second copy somewhere safe	Recovery
Dropbox- or Google Drive-like service	43	Ease of access and worry free backup	Ease of use
10G network connection	38	High speed network bandwidth	Network bandwidth
Minimal cost beyond hardware cost	24	Little operating cost	Cost
Shared technical expertise	20	Infrastructure, software and application consulting	Expertise
Distributed multisite replication	18	Geographical safety	Recovery
Desktop backup	18	Routine research data safety	Backup
Compliant and secure storage for sensitive data	16	Personal and clinical data safety	Security
Tiered storage plans	16	Data retention and automatic removal	Cost

Top 10 Requirements for Campus Cyberinfrastructure

- Cost effectiveness tops the list
- Ease of use follows
- “Cost is King, Ease of Use Follows”
- Reliable, NFS/CIFS storage most common platform
- Many responses relate to data durability – backups/copies/ tiered storage
- High-speed networking enhances quality of service
- “Compliant” environment (storage/computing)
- Tiered storage options is desirable

PRAGMA

SDSC Data Services



Cloud Storage (OpenStack Swift)

- Purpose: Storage of Digital Data for Ubiquitous Access and High-Durability
- Access Mechanisms: S3/Rackspace API, Web Cloud Explorer, Clients, CLI

Traditional File Server Storage (NFS/CIFS)

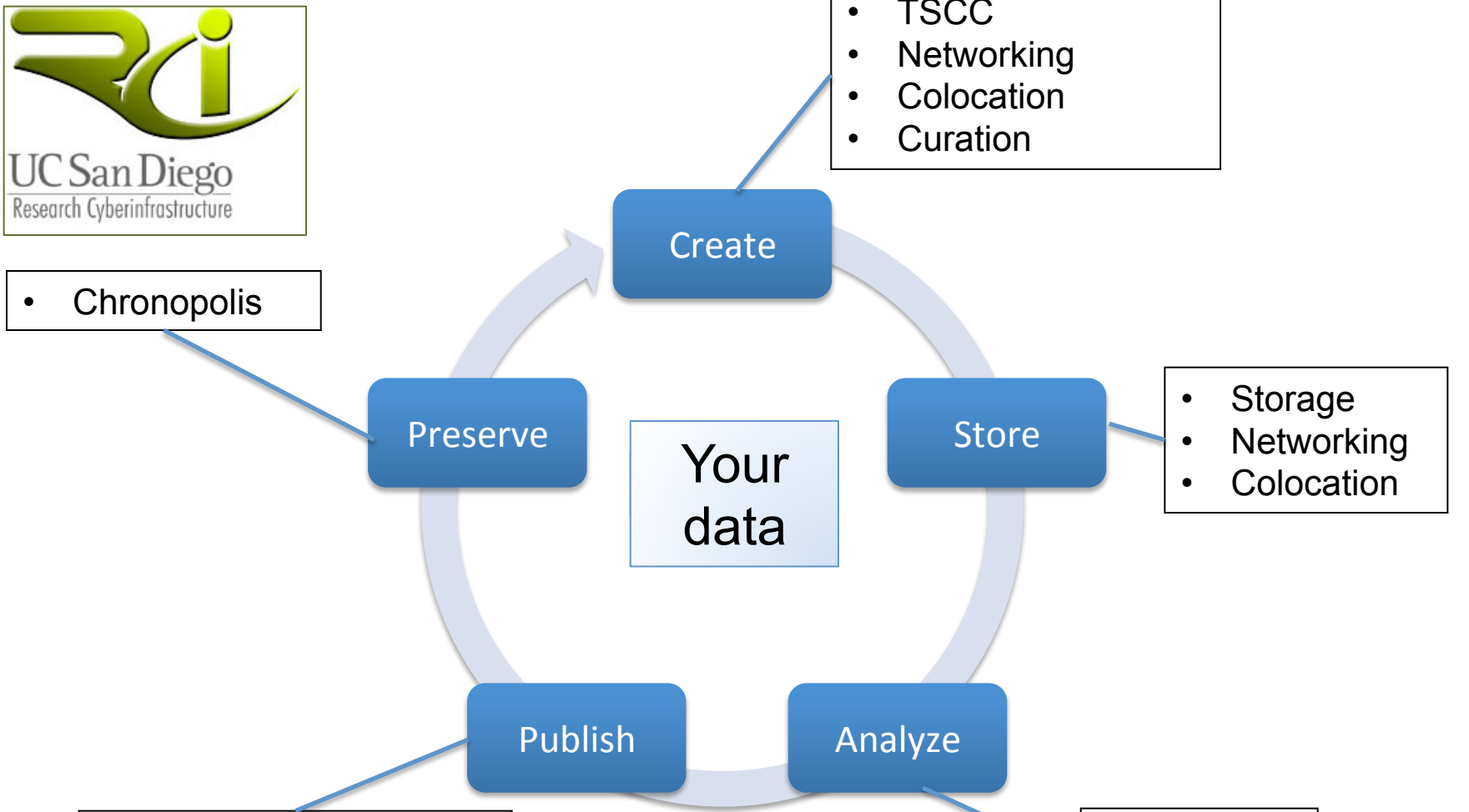
- Purpose: Typical Project / User Storage Needs
- Access Mechanisms: NFS/CIFS

High Performance Computing Storage (Lustre)

- Purpose: Transient Storage to Support HPC
- Access Mechanisms: Lustre on HPC Systems, NFS/CIFS for data migrations

New $12 million dollar award to SDSC, online in 2015

UC San Diego
Research Cyberinfrastructure



- TSCC
- Networking
- Colocation
- Curation

- Storage
- Networking
- Colocation

- TSCC
- Curation

- Curation
- Storage
- Colocation
- SDSC Cloud

- Chronopolis

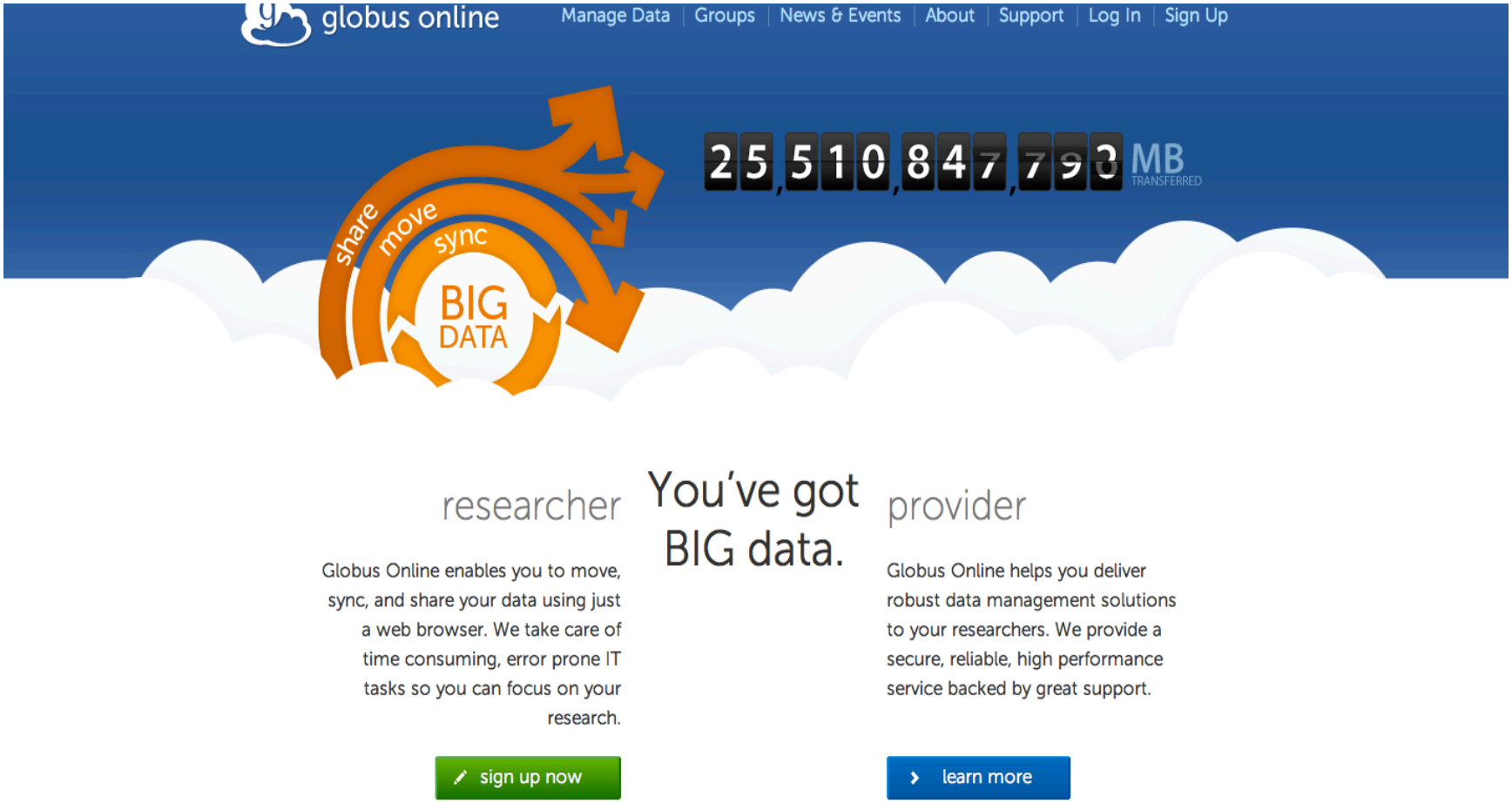
Network is everywhere

PRAGMA

Planned Activities

- Interested in using the virtual clusters in PRAGMA cloud for NGS and metagenomics analysis
  - Dr. Jaebum Kim, Konkuk University
- Using Globus Online for data transfer to virtual cluster and compute/storage resources
  - Dr. Jaebum Kim, Dr. Wilfred Li
- Using Active Folder for data synchronization and job execution
  - Dr. Daeyoung Heo, Dr. Wilfred Li
- Using PRAGMA virtual cluster comprising multiple sites using ViNE for Map Reduce Blast analysis
  - Dr. Andrea Matsunaga

Globus Online, Dropbox for Science



Joint Session

- Evaluate the Cyber Learning group resources and Data Synchronization for Research
  - Dr. Ruth, Dr. Daeyoung Heo
- Planning for next Cyber Learning workshop at next PRAGMA
  - Dr. Ruth Lee
- Share information on MOOC development in education and learning
  - Dr. Putchong Uthayopas, Dr. Wilfred Li