DATA STORAGE AND SHARING FOR THE LONG TAIL OF SCIENCE

PURDUE UNIVERSITY

Line Pouchard
Boyu Zhang
Preston Smith
Amanda Gasc
Bryan Pijanowski

NEW YORK SCIENTIFIC DATA SUMMIT

08/15/2016

OUTLINE

- The long tail of science
- Data management plan requirements
- Data Depot at Purdue
  - Motivation
  - Aims and objectives
  - Deployment
- Use case: Soundscapes
  - Data ingestion and sharing
  - Data indexing and analysis
  - Scaling out Soundscape analysis
- Transition to preservation
- Summary

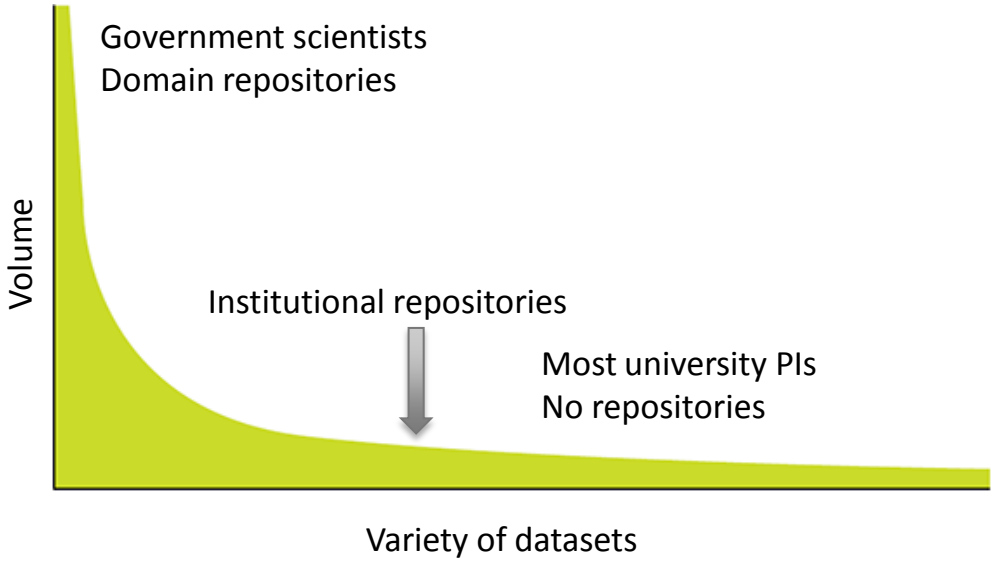
THE LONG TAIL OF SCIENCE

Big Data

- Big science
- Big data
- Large, multi-institution collaborations
- Agency-sponsored data collection
- Long-term perspective
- Remote sensing
- Well-curated and preserved
- Expensive

Small Data

- Small science
- Small collaborations
- Individual labs
- In-labs collection
- Poorly curated and preserved



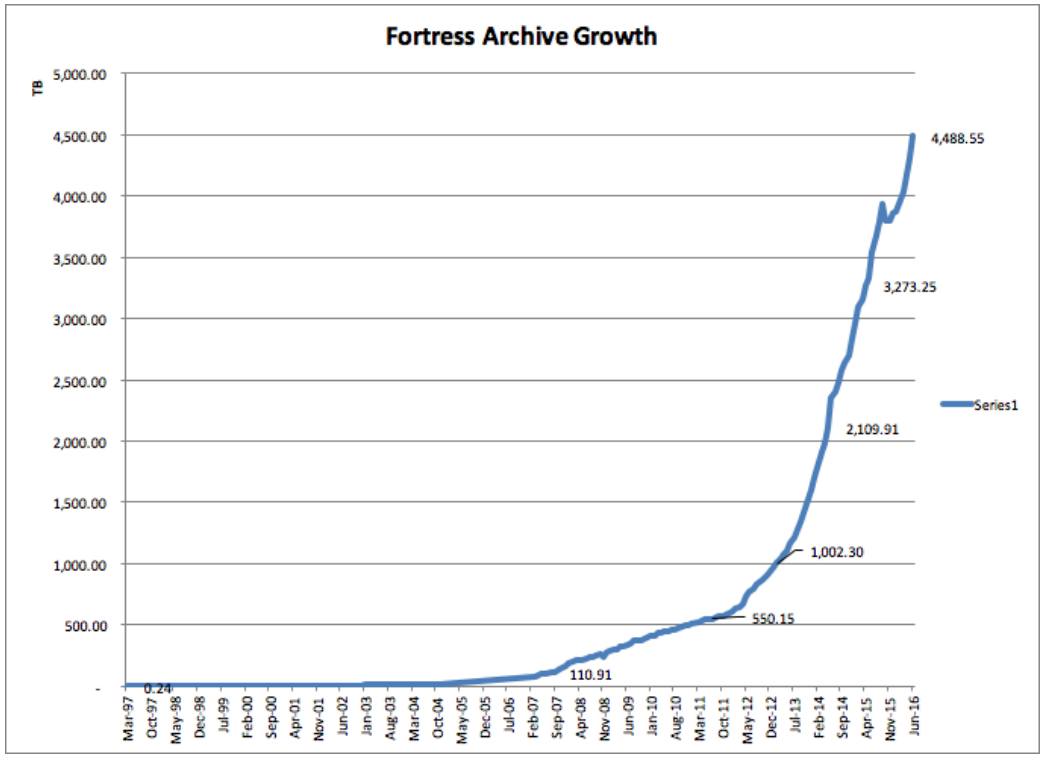
- Poor access
- Shor-term projects

Graphic: Bryan Heidorn, 2008, Shedding light on the dark data in the long tail of science

VARIETY LEADS TO VOLUME

The long tail of science increasingly means Big Data

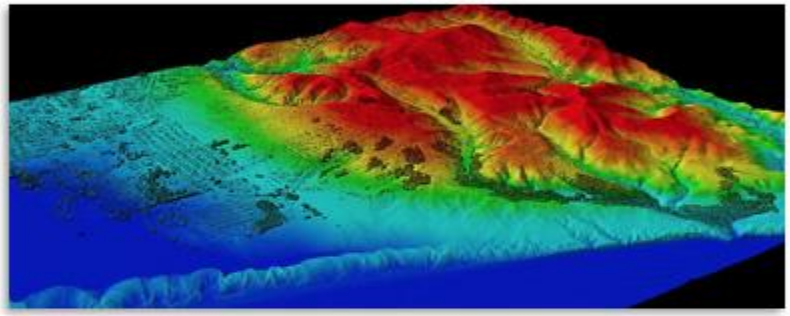
- Very heterogeneous data (the Variety V of Big Data)
- New problems increasingly require HPC resources
- Data volumes increase



Archive usage is skyrocketing

What is big data at Purdue?

- **Not just Facebook-style analytics!**
- 3.5 PB of high-energy physics detector data
- 1 PB of climate model data
  - 90 TB in an active workflow!
- 200 TB of astrophysics simulations
- 150 TB of CFD model output
- 120 TB of audio files
- 100 TB of actively-used next-gen sequencing data
  - Millions of files used in an active workflow
- 10s of TB of video files
- 5 TB of electron microscope images generated per day
- ..to the 75% of users on Conte using less than 1TB
- ... and to the social science researcher with stacks of excel sheets



Big data: A data set that is larger/faster/more complex than one feels comfortable dealing with.

WHY SHARE DATA?

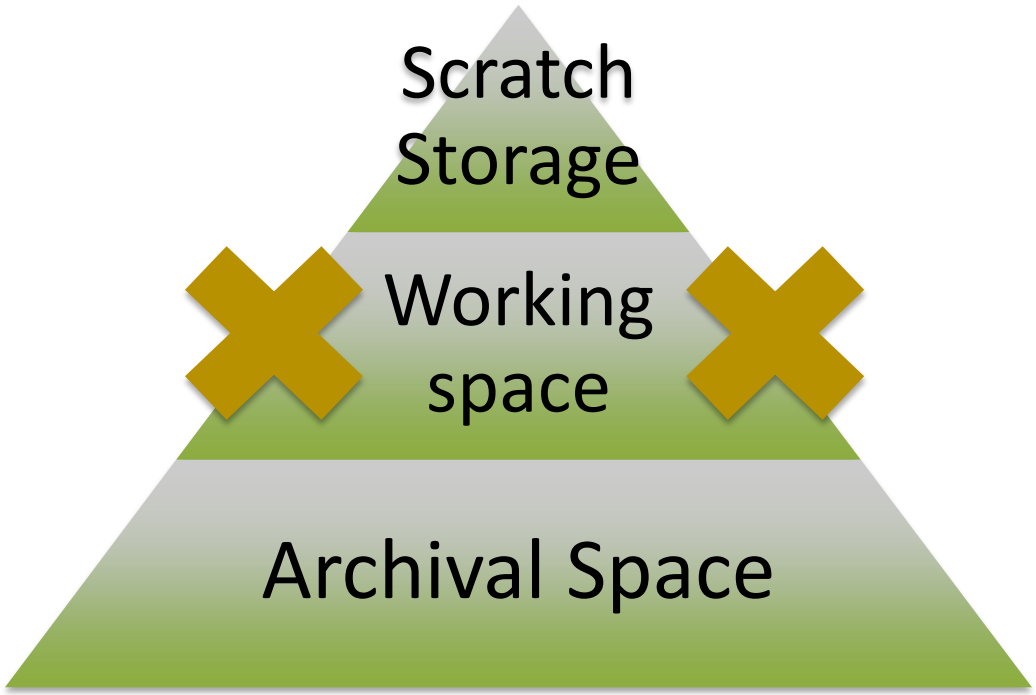
- **Re-use of data** for new research, including collection-based research to generate new science
- **Retention** of unique observational data which is impossible to re-create
- More data is **available for research** projects
- Ability to **validate** research results
- Use of data in **teaching**
- For the **public good**
- Compliance with legal/funder requirements

e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. Philip Lord and Alison Macdonald. 2003

AGENCY DATA MANAGEMENT PLANS

- The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
  - Including **RAW** Data
  - Timing vary by community and nature of data
- Ensure the timely release and sharing of **FINAL** research data
  - While respecting privacy
  - At the time of publication
- DOE Public Access Plan to scientific publications and to data
  - Peer-reviewed articles and accepted manuscripts – unclassified or unrestricted (PAGES)
  - All **research data** displayed in publications at the time of publication

TIERS OF DATA STORAGE



Fast, large, purged, coupled with clusters, per-user – **for running jobs**

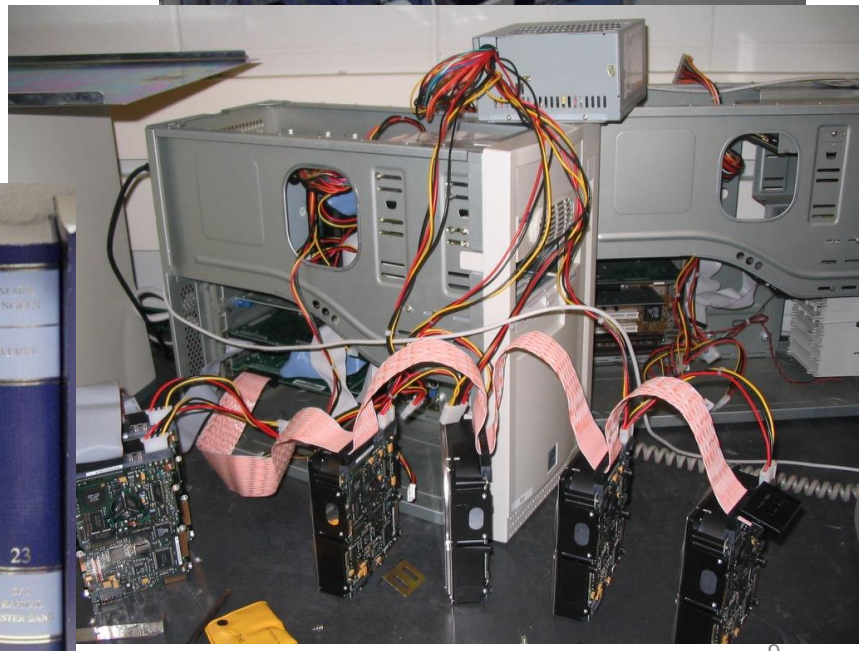
Medium speed, large, persistent, data protected, purchased, per research lab – **for shared data and apps**

High speed, high capacity, well protected, available to all researchers – **for permanent storage**

RESEARCH STORAGE GAPS

Many central storage options have not met all the needs that researchers care about

- Departmental or lab resources are islands and not accessible from HPC clusters
- Most are individually-oriented, rather than built around the notion of a research lab
- Scratch filesystems are also limited in scope to a single HPC system



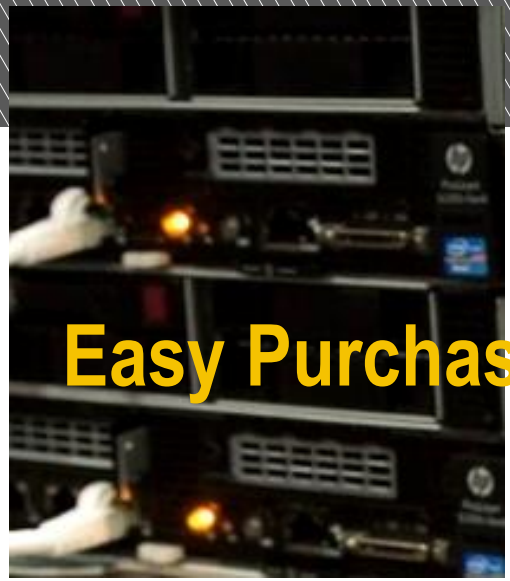
Back-ups?
Access?
Preservation?

SOLUTION: DATA DEPOT

- Approximately 2.25 PB of IBM GPFS
- 160 Gb/sec to each datacenter
- 5x Dell R620 servers in each datacenter
- Accessible to groups inside and outside Purdue
- Suitable for group sharing, group editing, developing and building software
- Hardware provided by a pair of Data Direct Networks SFA12k arrays, one in each of MATH and FREH datacenters
- 100G available at no charge to research groups
- Mounted on all clusters and exported via CIFS to labs
- Not scratch: Backed up via snapshots, with DR coverage

SUCCESS: A DATA SERVICE

It is not enough to just provide infrastructure
"Here's a mountpoint, have fun"



In just over a year, 280 research groups are participating

Many are long tail of science users (not HPC)

0.75 PB used since 2014

A research group purchasing space has purchased on average 8.6 TB

By comparison, 12 year old cluster program has 180 PI + workers

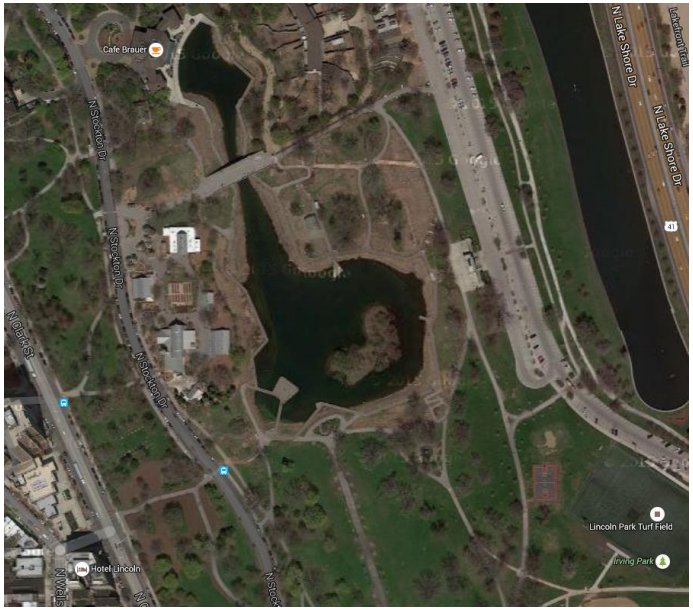
USE CASE: GLOBAL SOUNDSCAPE ANALYSIS

MOTIVATION AND CHALLENGES

A nonintrusive way to study global environment changes and effects on biodiversity

Advances in automated acoustic recording result in big collections of **audio data** that keep growing

Challenges in data storage, sharing, and processing



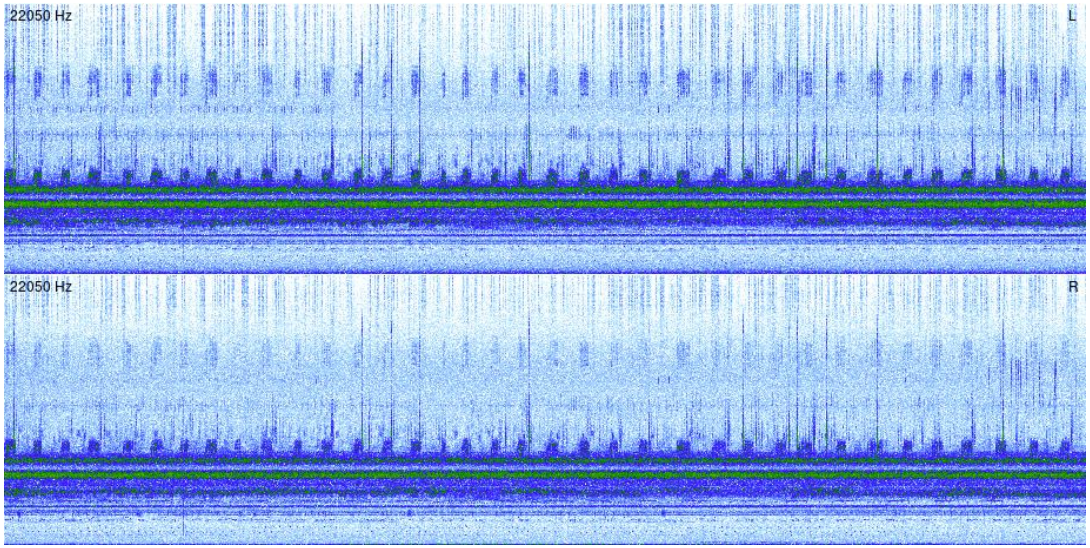
THE DATA COLLECTIONS

Started in 2008, the recordings are organized in 13 collections across hundreds of different sites around the world, total size of 120 TB and growing

Different location and environment: tropical (La Selva), urban (Chicago), desert (Arizona), estuary (wells), wet land (West Lafayette)

The standard recording protocol is to deploy automated recorders on a site and record 10 minutes every half hour, for a duration that can vary from 10 days to several years

Analyzing these recordings over large spatial and temporal scales can produce insights such as the natural and human disturbance impacts on animal communities



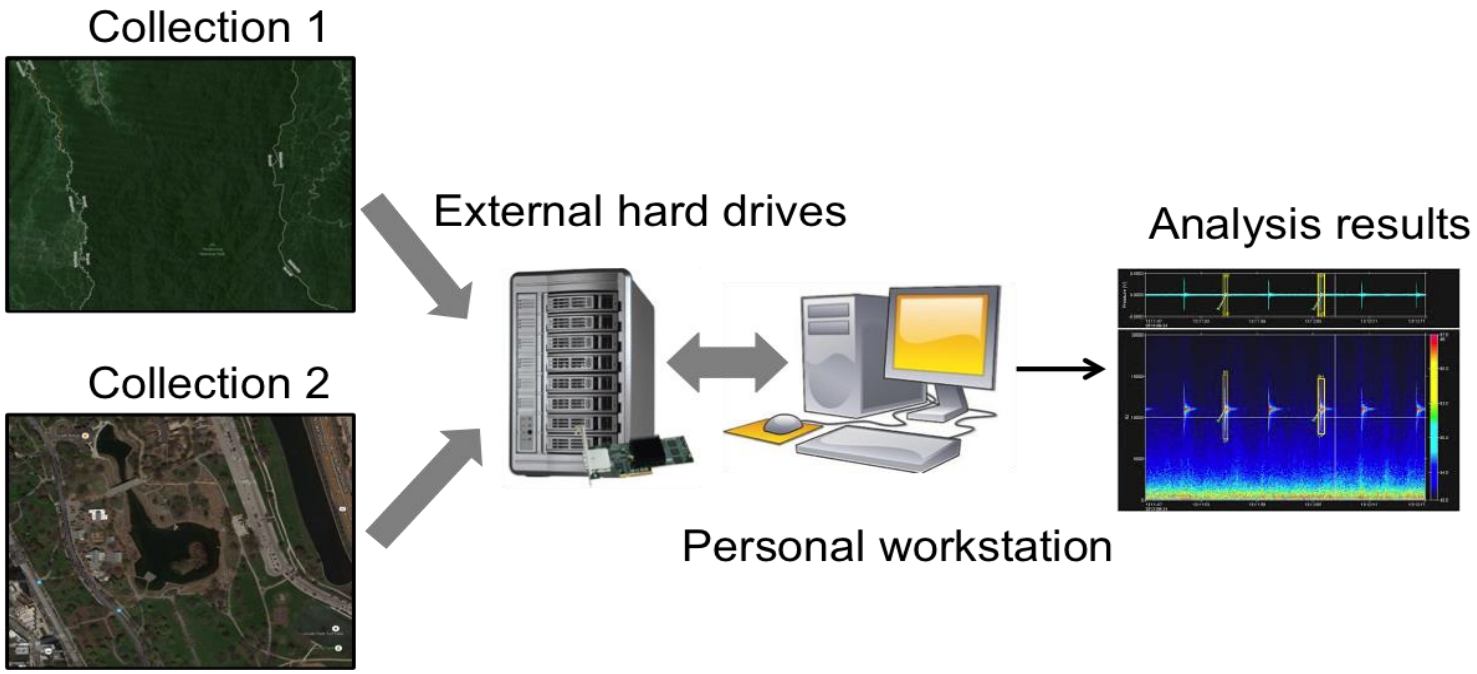
DATA STORAGE IN SOUNDSCAPE

WORKFLOW – BEFORE DATA DEPOT

Audio data were stored in a set of external drives

A powerful workstation is connected to the external drives

Scientists perform computation and analysis on the workstation



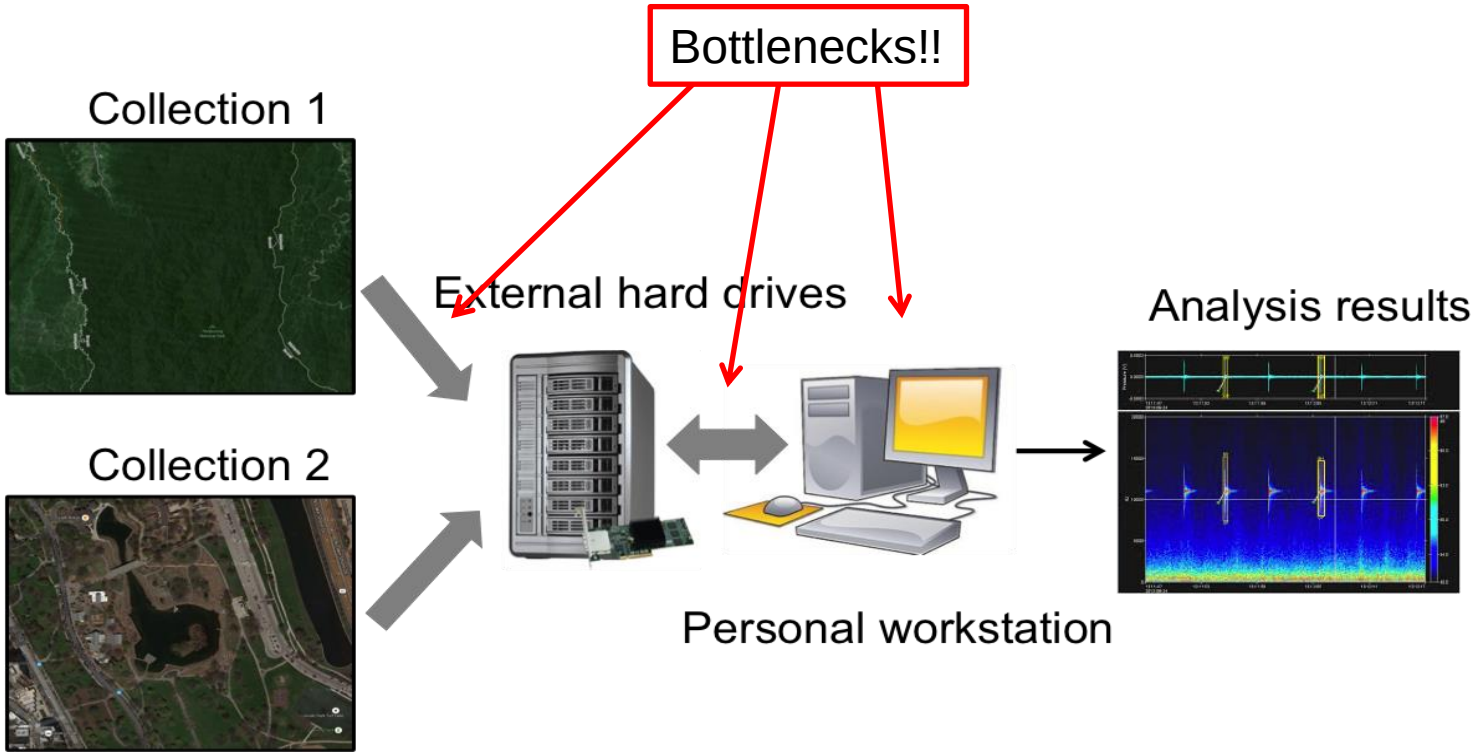
DATA STORAGE IN SOUNDSCAPE

WORKFLOW – BEFORE DATA DEPOT

Audio data were stored in a set of external drives

A powerful workstation is connected to the external drives

Scientists perform computation and analysis on the workstation

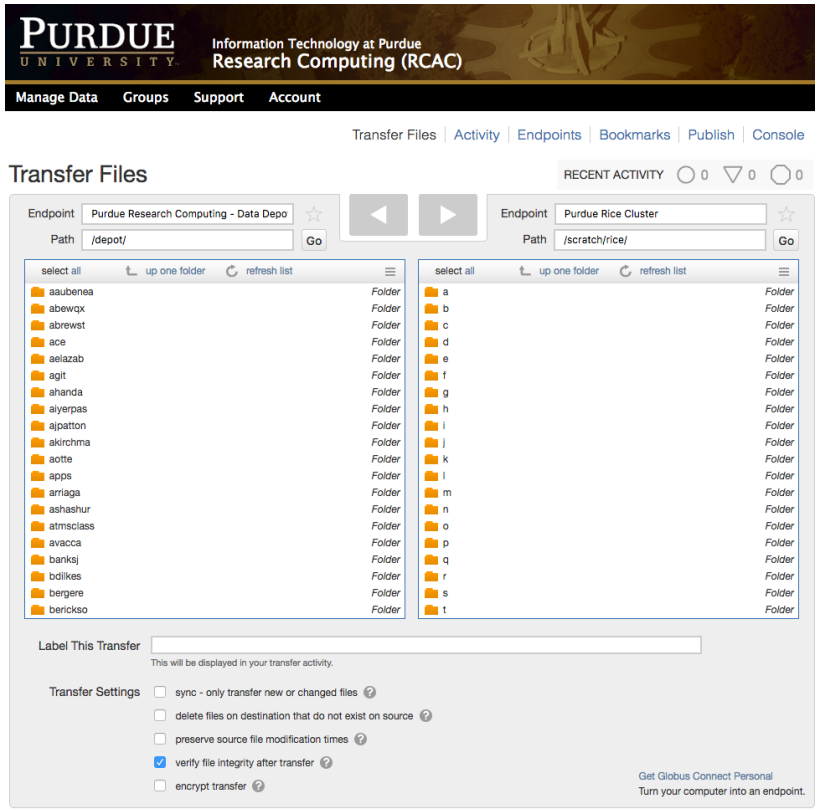


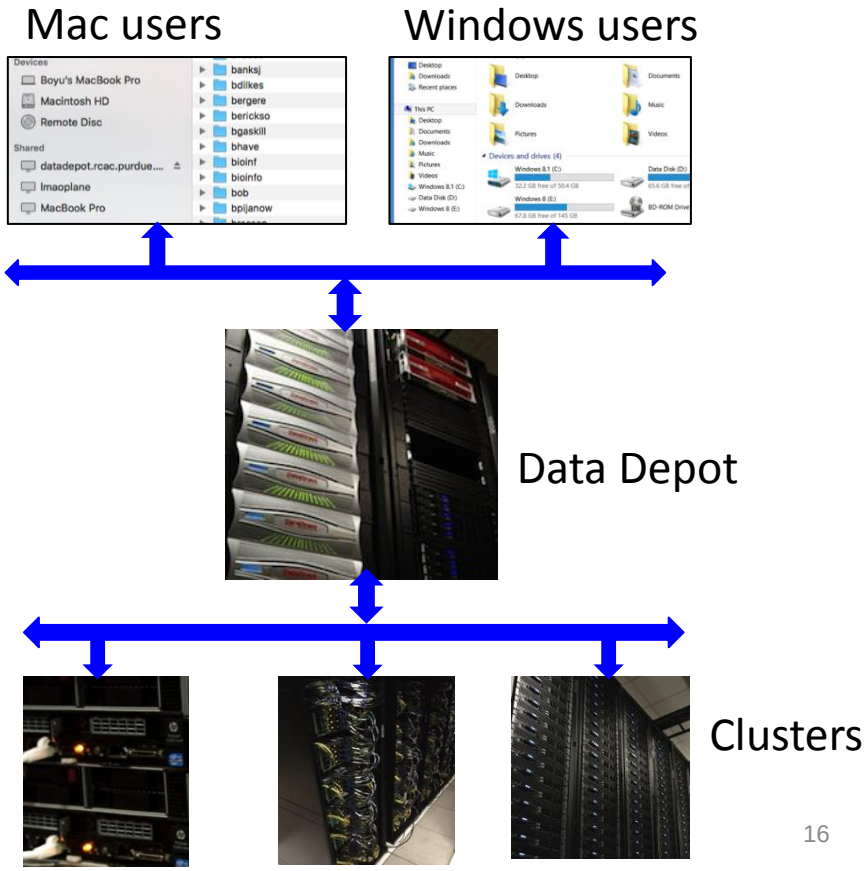
DATA DEPOT SERVICES

EASY DATA INGESTION AND SHARING

Data transfer using Globus

Accessible from community clusters, user laptops, workstations





DATA DEPOT SERVICES

LARGE STORAGE CAPACITY

Stores all raw audio data in lossless format

/depot/mylab/

```
+--/apps/
|
+--/data/
|               +--colID1/
|                           +--siteID1/
|                           +--siteID2/
|               +--colID2/
|
+--/etc/
|               +--bashrc
|               +--cshrc
+--/repo/
|
+--/www/
|
+--/users
|
+--/users/ijones
```

120TB collected since 2008 have been moved to Depot. Collections keep growing!

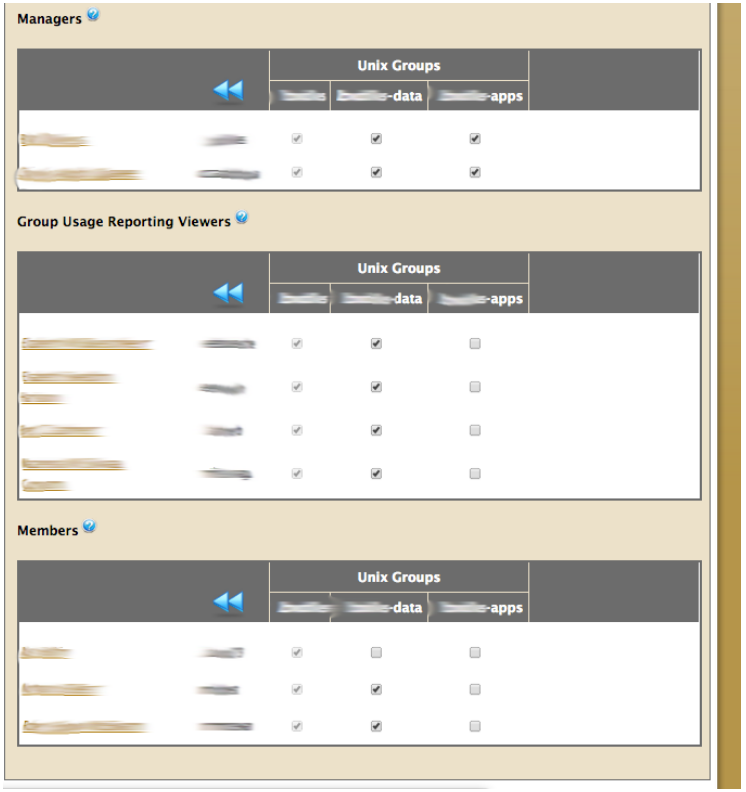
DATA DEPOT SERVICES

FLEXIBLE ACCESS CONTROL

PIs manage access control through a web portal by clicking buttons

/depot/mylab/

```
+--/apps/
|
+--/data/
|
+--/etc/
|          +--bashrc
|          +--cshrc
+--/repo/
|
+--/www/
|
+--/users
|
+--/users/ijones
```



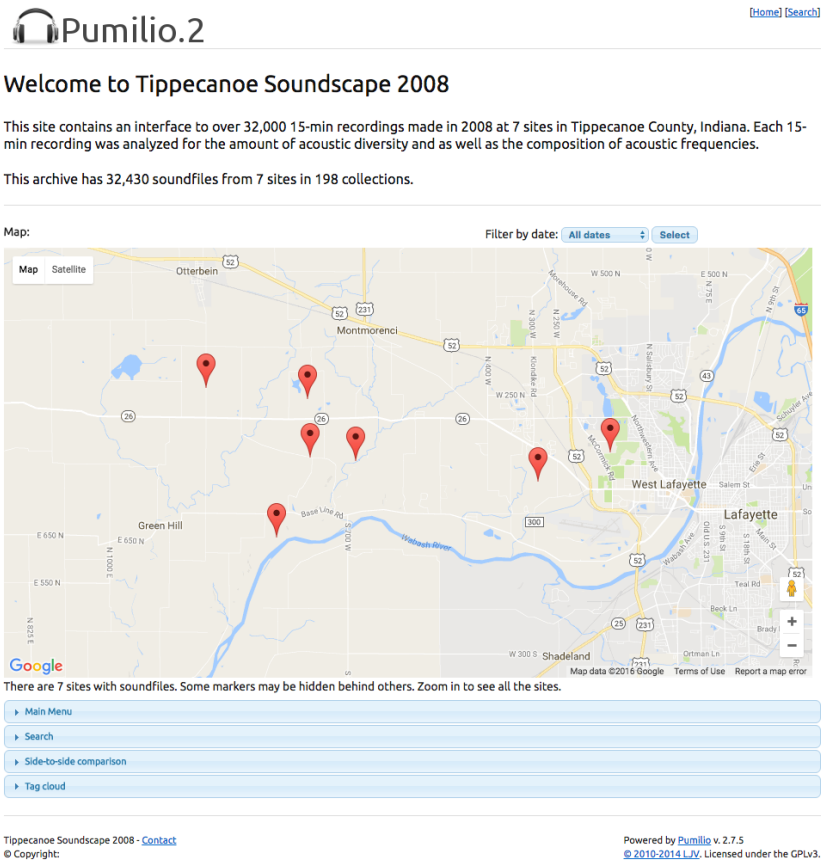
DATA DEPOT SERVICES

SUPPORT FOR OPEN DATA SHARING

Pumilio - a SQL database with metadata about the recording collections with web-based user interface

/depot/mylab/

```
+--/apps/
|
+--/data/
|
+--/etc/
|          +--bashrc
|          +--cshrc
+--/repo/
|
+--/www/
|
+--/users
|
+--/users/ijones
```



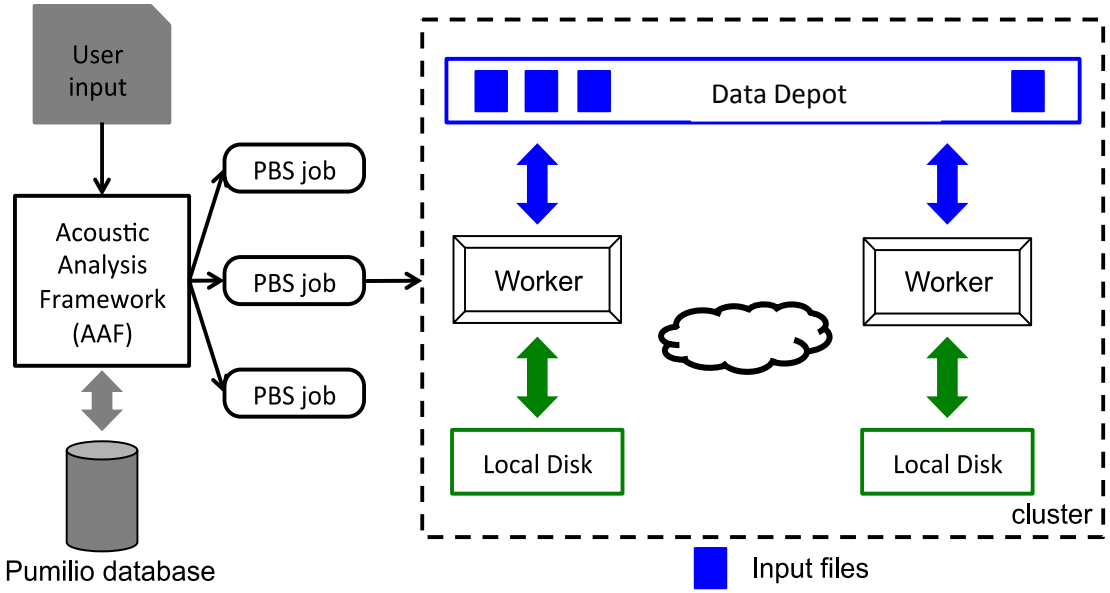
SOUNDSCAPE ANALYSIS

WORKFLOW – WITH DATA DEPOT SUPPORT

Audio data are stored in Data Depot

Our Acoustic Analysis Framework takes user specified options (e.g. locations and time periods), queries Pumilio, and submits jobs to a cluster

Each job runs R Workers that perform acoustic analysis in parallel



Hardware: Rice, a 576 nodes cluster, two 10-core Intel Xeon-E5 (20 cores/node); 64 GB memory per node

Software: R - 3.2.2

SOUNDSCAPE ANALYSIS

QUANTIFY LEVEL OF ACOUSTIC ACTIVITIES – ALPHA-DIVERSITY INDICES

Our framework computes 11 different alpha-diversity indices 13 collections of audio recordings

- Different alpha-diversity index measures different aspects of acoustic activities
- For example, the table shows 3 representative alpha-diversity indices

Theses analysis could not be done before due to computation and storage limitations

Index name	Calculation principle
Bioacoustic index (Bi)	Area under the frequency spectrum
Temporal entropy (Ht)	Entropy of the amplitude envelope over the time units
Spectral entropy (Hf)	Entropy of the amplitude values over the frequency units

SOUNDSCAPE ANALYSIS

QUANTIFY LEVEL OF ACOUSTIC ACTIVITIES – ALPHA-DIVERSITY INDICES

Different collections of audio recordings are of different size

For example, the table shows three representative collections and the resources used to process them

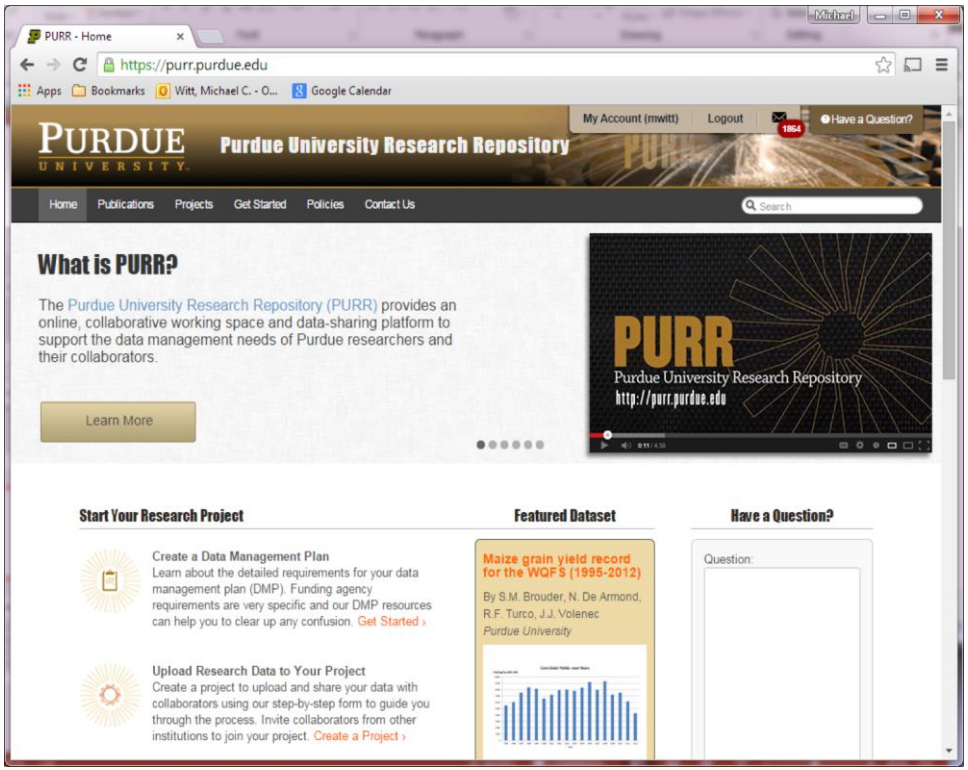
Collection name	Short description	Num. of files	Size (TB)	Num. of nodes	Walltime (hh:mm:ss)
arizona2013	Long-term wildfire disturbance survey, desert ecosystems	121739	10	24	2:18:05
tip2009	Long-term habitats description, Temperate ecosystems	32425	6.5	8	2:23:39
tip2010	Long-term habitats description, Temperate ecosystems	46937	5.8	8	2:30:48

TRANSITION TO PRESERVATION

NSF requires access and preservation of datasets

Soundcapes data is accessible via Pumilio ----- Depot does not have preservation

We are evaluating migrations of selected datasets to the Purdue University Research Repository



select

SUMMARY

- We presented the success of Data Depot as a service and storage platform at Purdue University for researchers campus-wide
- The Use case with Soundscapes illustrates both the need for Depot services and how they are used
- The ability to share data easily
  - to comply with NSF data sharing requirements
  - and makes perishable data available to ecologists worldwide
- The combination of Depot services in use by Soundscapes illustrates how to support the long tail of science at a land-grant university