“Cold” research data storage at the University of Sydney

Stephen Kolmann
Peter Ceiley
Christopher Albone
Andrew Janke

THE UNIVERSITY OF SYDNEY

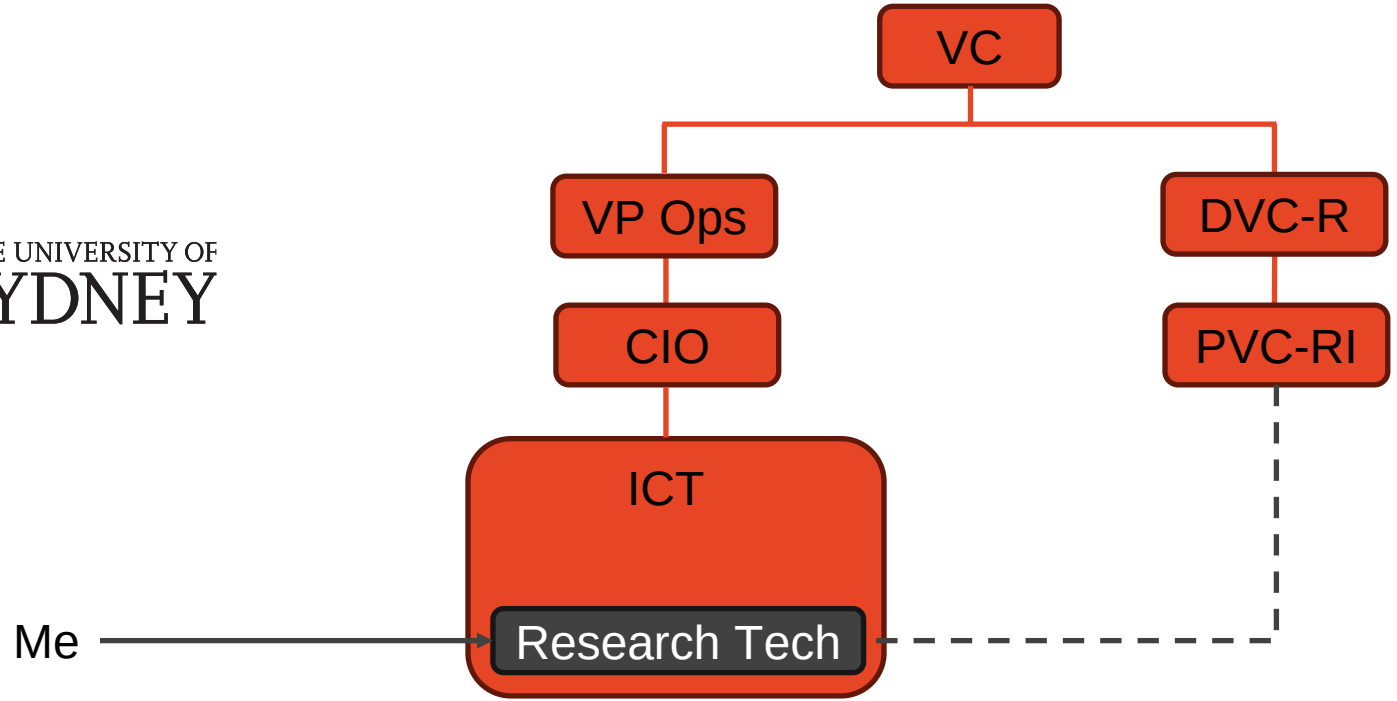
CRICOS 00026A TEQSA PRV12057

We recognise and pay respect to the Elders and communities – past, present, and emerging – of the lands that the University of Sydney's campuses stand on. For thousands of years they have shared and exchanged knowledges across innumerable generations for the benefit of all.

THE UNIVERSITY OF SYDNEY

Research tech + ICT

THE UNIVERSITY OF SYDNEY



Guide to storing and managing your projects research data

University supported and licenced platforms

Platform/Tool	eNotebook	REDCap	Research Data Store (RDS)	OneDrive (Enterprise)	Teams (Enterprise)	Highly Protected SharePoint (Enterprise)	Australian Imaging Service (AIS)	Unsuitable as primary storage for research data: Local storage, USB Drive	Prohibited for protected research data: Other cloud tools (e.g. Google Drive, Dropbox)
function	electronic notebook	survey and data capture, including Clinical trials	networked data storage, large files, HPC access	cloud storage	chat, collaboration, cloud storage	collaboration, cloud storage	imaging repository and analytics	removable media, local storage	cloud storage
suitable for data classification	●●●	●●●	●●●	○●●	○●●	●●●	●●●	●	●
stored in Australia	✓	✓	✓	✓	✓	✓	✓	various	✗
external collaborator access	✓	✓	✓	✓	✓	✓	✓	✗	✗
context and commentary supported	✓	✗	✗	✓	✗	✗	n/a	✗	✗
syncing with local copy	n/a	n/a	n/a	✓	✓	✓	n/a	✗	✗
available storage	unlimited	unlimited	unlimited (default 2TB)	1TB	2TB+	25TB max (default 2TB)	unlimited	✗	✗
backup and disaster recovery	✓	✓	✓	✓	✓	✓	✓	✗	✗
audit trail/version control	✓	✓	✓	✓	✓	✓	✓	✗	✗
versioning retained	✓	manual	up to 60 days	7 years	7 years	7 years	✗	✗	✗

Version 6.1
October 2024
Endorsed by CIO

THE UNIVERSITY OF SYDNEY

●	highly protected
○	highly protected data needs additional file encryption
●	protected
●	public

Highly Protected data may require additional encryption depending on some platforms. Protected data may benefit from encryption.

For more information about research data classifications, go to https://sydney.edu.au/research-data-classifications

For research data management enquiries, please contact digital.research@sydney.edu.au

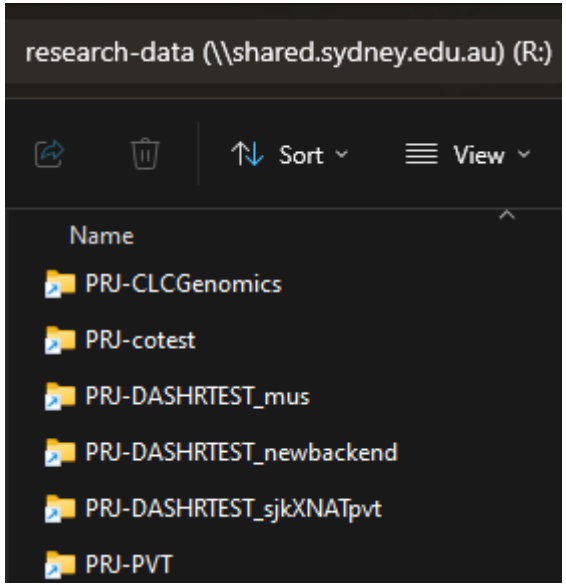
Research data store structure

- All RDS storage associated with a project in “Researcher Dashboard (DashR)”, with name PRJ-<project code>

- “Map a network drive” to \\shared.sydney.edu.au\research-data (SMB shares)

- SFTP access via SFTP servers, mounted under /rds

- Mounted on University HPC under /rds



Research data store (RDS) statistics

12.9 PiB

Total data stored

8680

Total projects

3557

Year-to-date active users

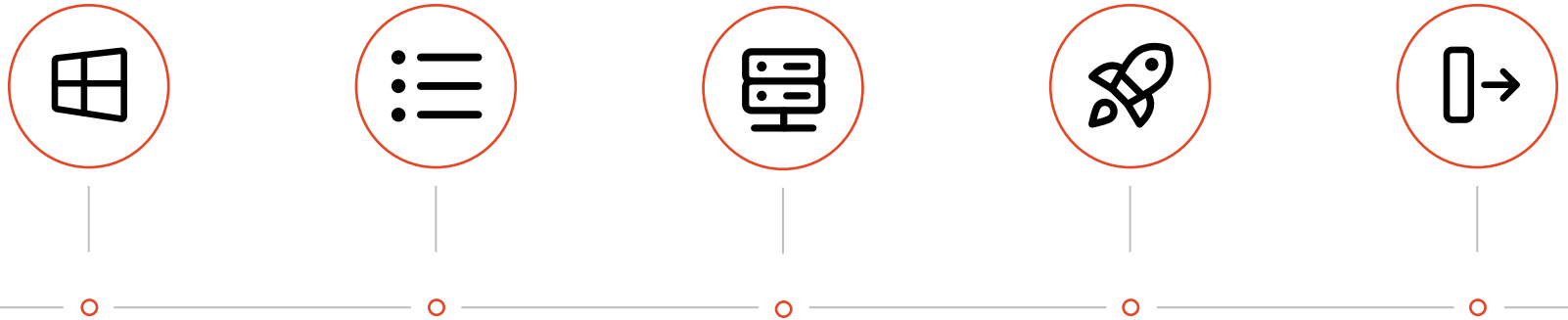
2.3 billon

Total files

30 PiB

Sum of quotas (demand)

Research data store history



2009
Central RDS service launches with group drives

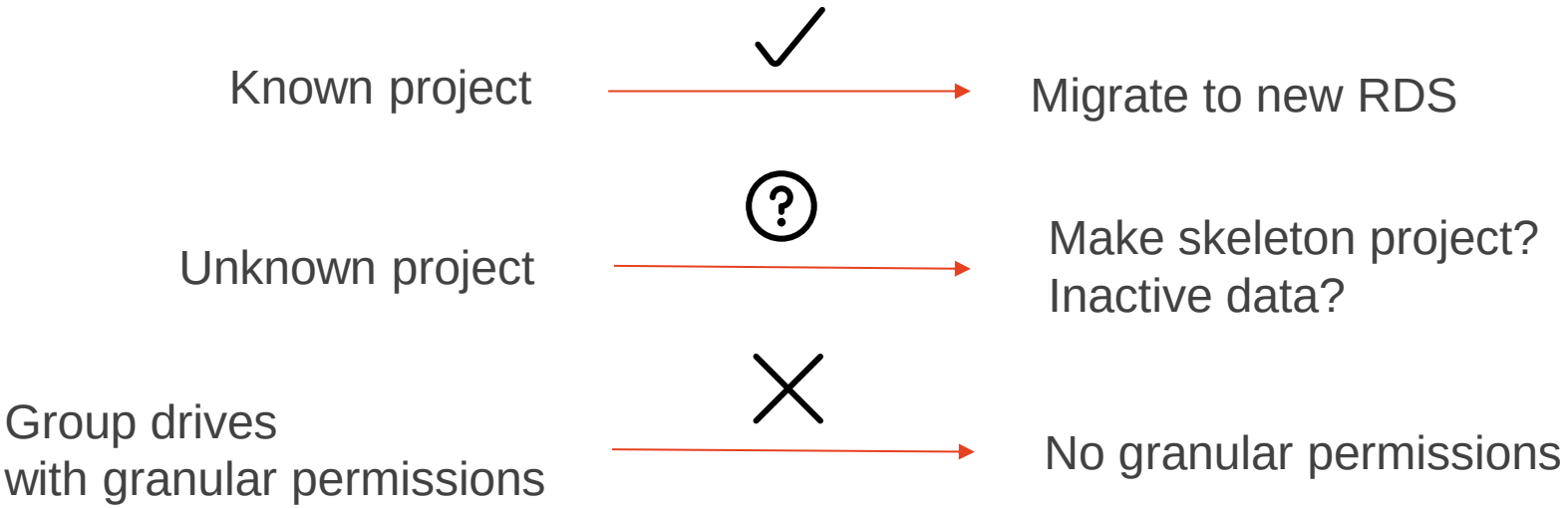
2012
Launch RDMP tool

2014
Centralised group drives and RDS on central NAS

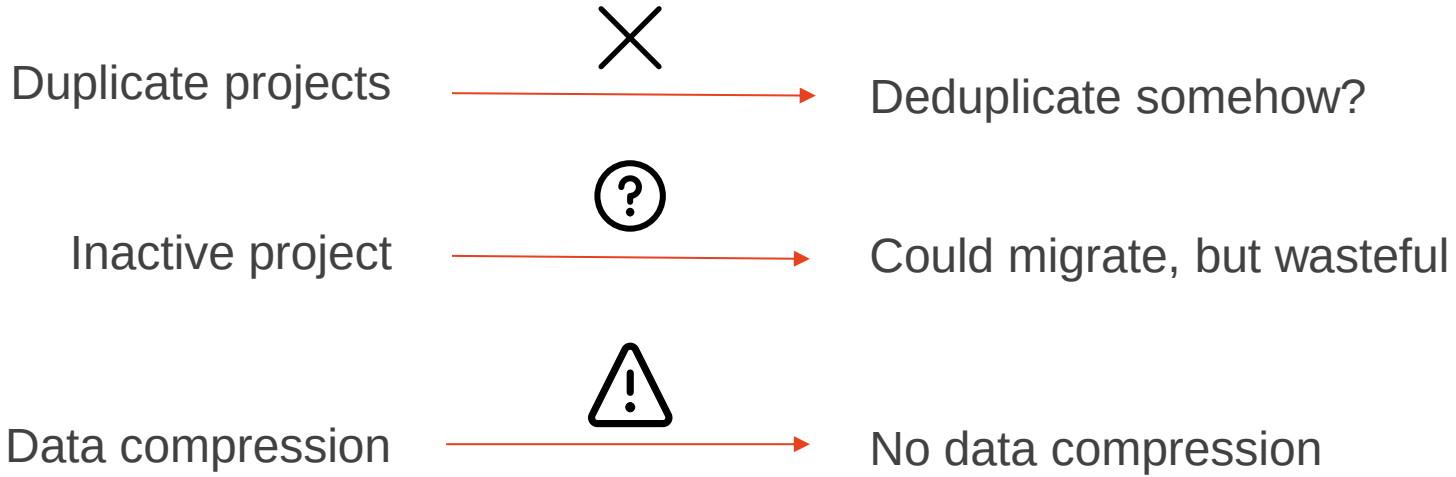
2015
Automated provisioning of RDS project directories in our RDMP tool

2020
Launch of new RDS clusters

New RDS migration challenges



New RDS migration challenges



How to tackle these migration challenges?

- We want visible/transparent data lifecycle management

Avoid opaque backend storage

No granular permissions

Cope with the “data tsunami”

Solution: cold storage

- Store inactive whole project directories in an inactive location
- A project's storage is either in active storage, or in cold storage, but never both simultaneously

Challenge	Cold storage solution
Data kept forever	Cost-effective storage of inactive data
No visible data lifecycle management	Provides a visible offline storage tier
Data volumes growing quickly	Makes more storage available for active data
Unknown old project ownership	Cut-off access until owner identified
Granular permissions	Cut-off access while "flattening" permissions

Cold storage

Move whole projects to cold storage

Store data in Bagit* (v0.97) format

Split zip file if the file is bigger than 25 GB

* https://datatracker.ietf.org/doc/html/draft-kunze-bagit-14
* https://libraryofcongress.github.io/bagit-python/

BagIt bag structure

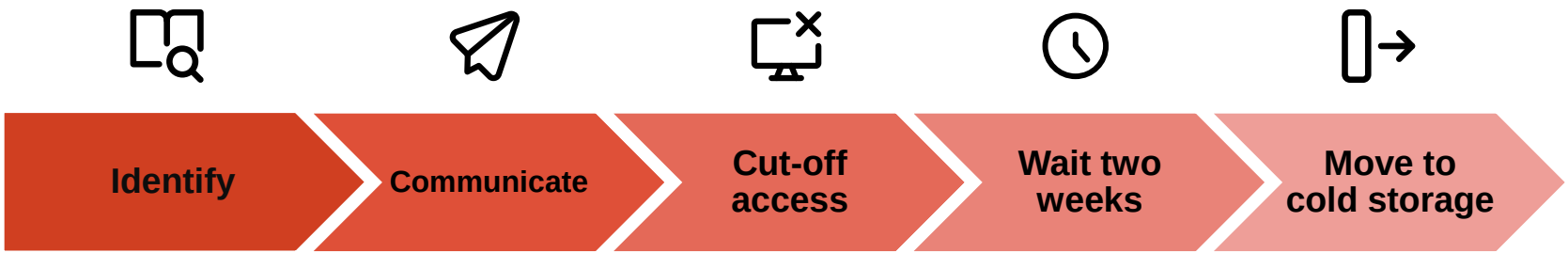
```
PRJ-Test
|    bagit.txt
|    bag-info.txt
|    manifest-sha512.txt
|    tagmanifest-sha512.txt
\--- data/
     |   [project files]
```

```
PRJ-Test.zip.001
PRJ-Test.zip.002
PRJ-Test.zip.003
…
PRJ-Test.zip.nnn
```

Example bag-info.txt

Bag-Size: <estimated human-readable bag size (if known)>
Bag-Software-Agent: bagit.py v1.7.0 <https://github.com/LibraryOfCongress/bagit-python>
Bagging-Date: yyyy-mm-dd
Contact-Email: lead.investigator@sydney.edu.au
Contact-Name: <Lead investigator name>
External-Description: <DashR project title>
External-Identifier: <DashR project code>
Organization-Address: The University of Sydney, NSW, 2006, Australia
Payload-Oxum: <exact bag size>
Source-Organization: The University of Sydney

Cold storage migration process



Find projects with no files accessed or modified for 2 years

Email project members that their project will be moved to cold storage

Update DFS target to point to read-only cold storage README.txt file

In case researchers need the data

Bag the project, copy to cold storage location, and delete from active storage.

Cold storage restore process

Researcher requests restore

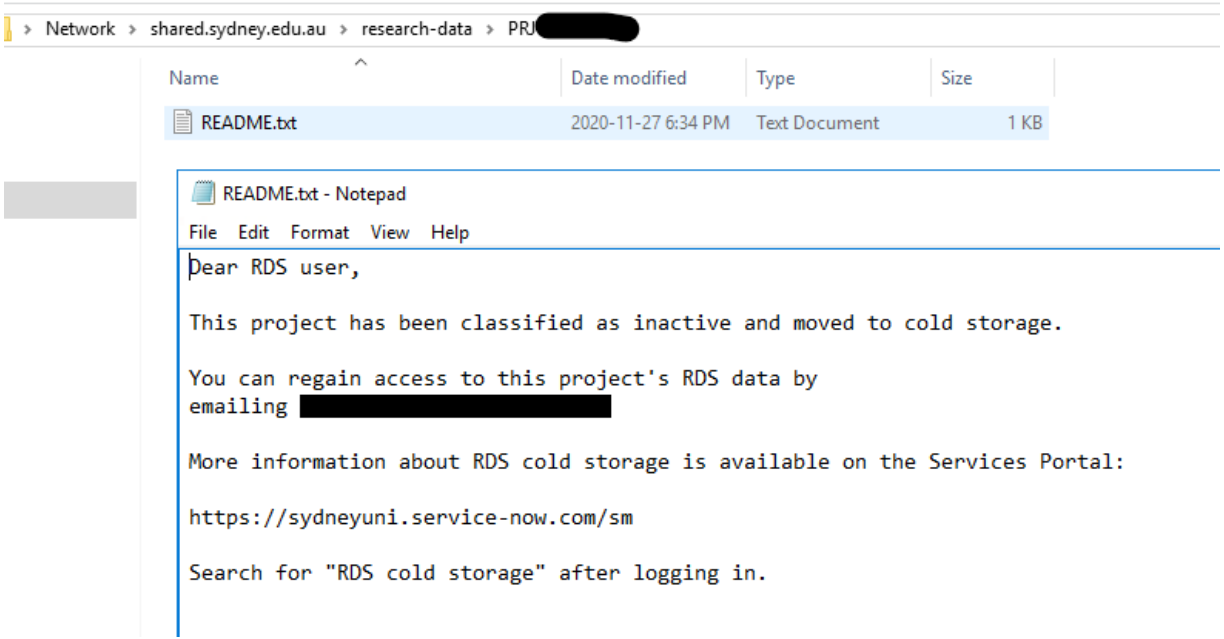
Restore to active storage initiated

Restore takes time to download, unzip, and check bag

Update DFS to point to active storage location

Inform researcher that restore is complete

Example cold stored project



RDS cold storage key statistics

1.2 PiB

Data stored in RDS cold storage in October 2024

2124

Project directories and group drives in cold storage

112

Projects restored from cold storage

266 TiB

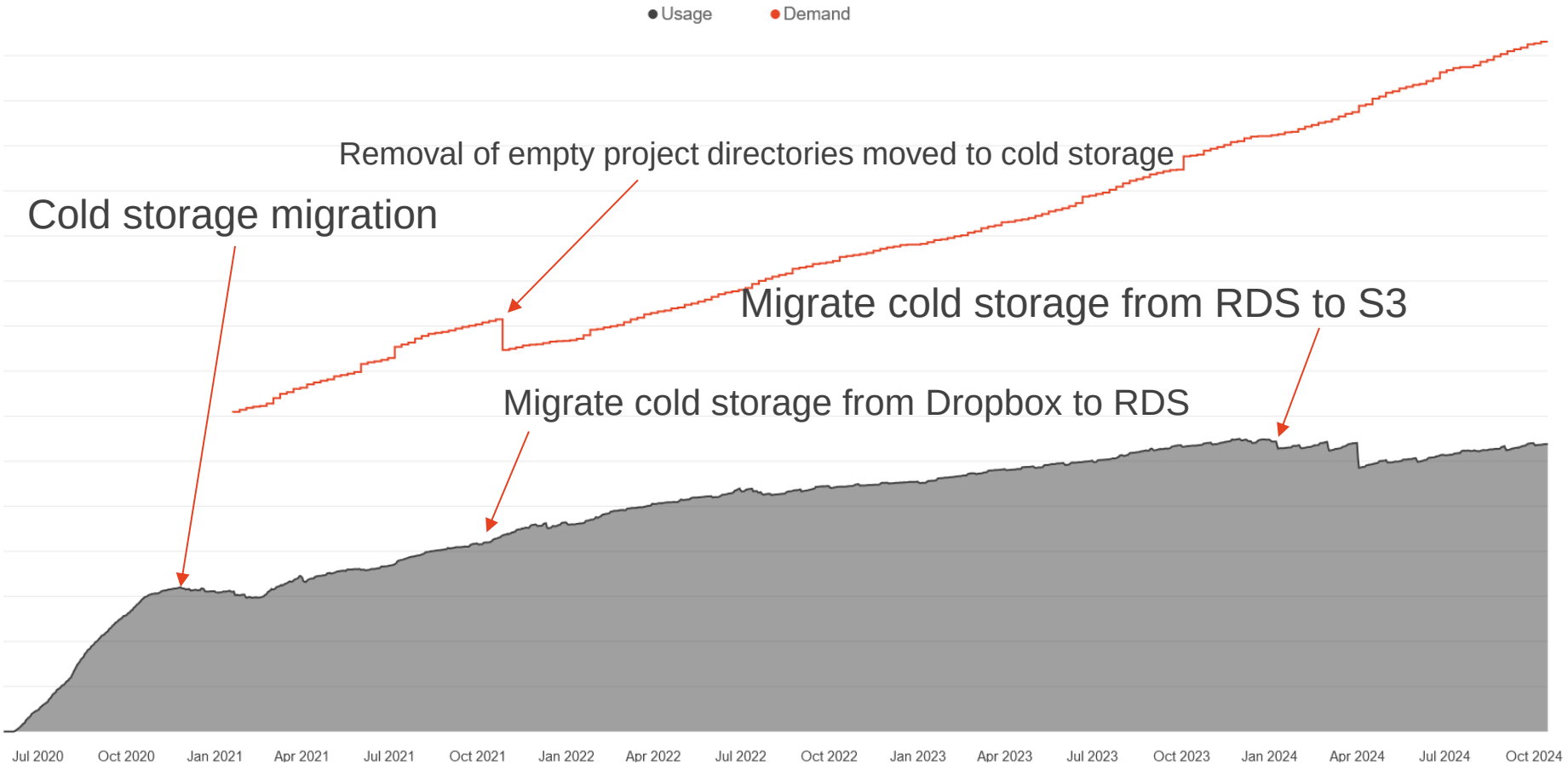
Data restored from cold storage

1

Formal complaint lodged

2

Cold storage data migrations as of October 2024


Usage
Demand
Removal of empty project directories moved to cold storage
Cold storage migration
Migrate cold storage from RDS to S3
Migrate cold storage from Dropbox to RDS
Jul 2020
Oct 2020
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Apr 2023
Jul 2023
Oct 2023
Jan 2024
Apr 2024
Jul 2024
Oct 2024

Cold storage challenges

Zipping/unzipping projects with lots of files in a single directory

Researchers who want to cold store sub-directories and not whole projects

Difficult to restore sub-sets of files from a project

Lessons learned

Effective communication was crucial

Researchers will pull things back from cold storage many years later

Cold storage service doesn’t support long-term data collections

Future directions

Automating moving data into and out of cold storage

Integrating cold storage into wider data lifecycle management processes

Provide data collection services

Thank you!

ICT Research Technology
sydney.edu.au/research-data-platforms