

Storage Backup Archive EPFL Services

Love Data Week
16.02.2022

Christian Cléménçon
SCITAS

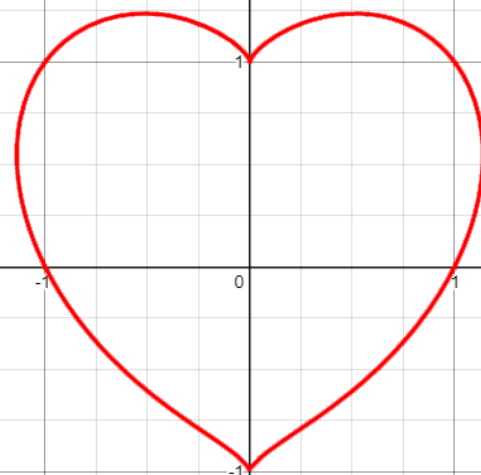
Micha D'Ans
Francesco Varrato
EPFL Library



CC BY 4.0

Love Data Week

EPFL 2♥22 edition



go.epfl.ch/quiz

Thu. Feb. 17 | 11:00-11:45

Online training "*Which platforms to publish my research data/code?*"

Fri. Feb. 18 | 10:00-12:00

Online speed-dating with EPFL Data Champions

go.epfl.ch/ldw22

Storage \neq Back-up \neq Archive

ACTIVE DATA

Data to be accessed *right away* during the research project



COLD DATA

Data with *low-frequency* access and not requiring fast access



STORAGE

A digital device or platform where data/code are saved and retrieved, usually active storage during a project

BACKUP

Supplementary **active** data storage, on different supports and/or locations, where copies of original data/code are made during the project

ARCHIVE

Cold data backup for *preservation*, where curated data/code (conversion, reformatting, metadata, licensing,...) stay accessible for the *long-term*



Needs / Uses (examples)



STORAGE

- Collect and analyze data
- Develop data analysis code
- Collaborate on shared files
- ...

BACKUP

- Rely on automated backup methods
- Synch your computer's storage on a server
- Entrust your work to the cloud
- ...

ARCHIVE

- Make a dataset reusable after 10 years
- Preserve your work when a project ends
- Keep your unpublished data (ex. negative results; personal data; etc.)
- ...



50 shades of **cold** data

Data Archive?

Long-term data preservation cold storage

Data repository?

Archiving systems allowing to publish and make datasets available online for the long-term, generic or discipline-specific, institutional or global

Data bank?

Collection of datasets for reuse, often mono thematic, normally institutional and requiring explicit requests for downloads

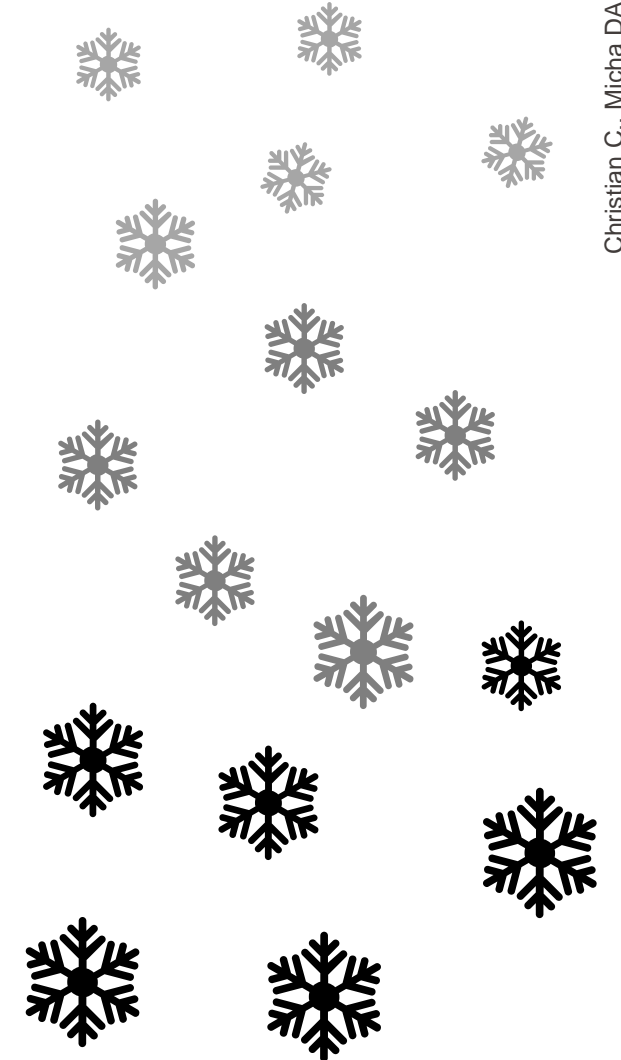
Cloud archive?

Storage as a service for long-term data retention

Code repository?

Digital platforms that simplifies the collaborative version control of code

Others...?



50 shades of **active** data @ EPFL

Christian
Cléménçon



- DSI central File storage service **NAS** (Network Attached Storage)
- MyNAS individual storage for students, staff and guests
- **SCITAS** work storage and c4science storage for coding/simulations (w/ LFS)
- gitlab.epfl.ch version control, open alternative to GitHub using EPFL servers
- Renku jupyterLab-based coding platform for reproducible data science
- Noto jupyterLab-based coding platform for testing and training
- eln.epfl.ch chemistry-oriented, web-based ELN with archiving support
- SLIMS (now Genohm SA) life sciences oriented ELN + LIMS
- ^{Lina} ~~Druva inSync~~ service to backup workstations / personal computers
- rsync utility to transfer / synchronize data across computer systems
- SWITCHDrive hosted by SWITCH servers (CH)
- gdrive.epfl.ch hosted on Google (Alphabet) servers

Not even in
Switzerland

File storage services (**NAS**)

COST – Was 165 Chf/TB/y... then no charges for 2021... then not indicated (90 Chf/TB/y?)

BACKUPS – Automated daily snapshot

VPN – Accessibility and network integration across the campus, plus remotely

PROTECTION – Local redundancy of data + synchronous replication on a secondary site

REQUEST FOR VOLUME – Will be on portal-xaas.epfl.ch (now by ticket in Service Now)

REF.: https://support.epfl.ch/epfl?id=epfl_service_status&service=49a363acdb34c700ef64731b8c96191f

= **go.epfl.ch/epfl-nas** 😊

- Faculty IT servers
- Stockage Object S3 based on AWS protocol
- **ACOUA** (Academic Output Archive) for data preservation:
 - Data curation
 - Long-term integrity
 - Security

Micha
D'Ans
😊

... launched during Love Data Week 2021

Christian
Cléménçon
😊

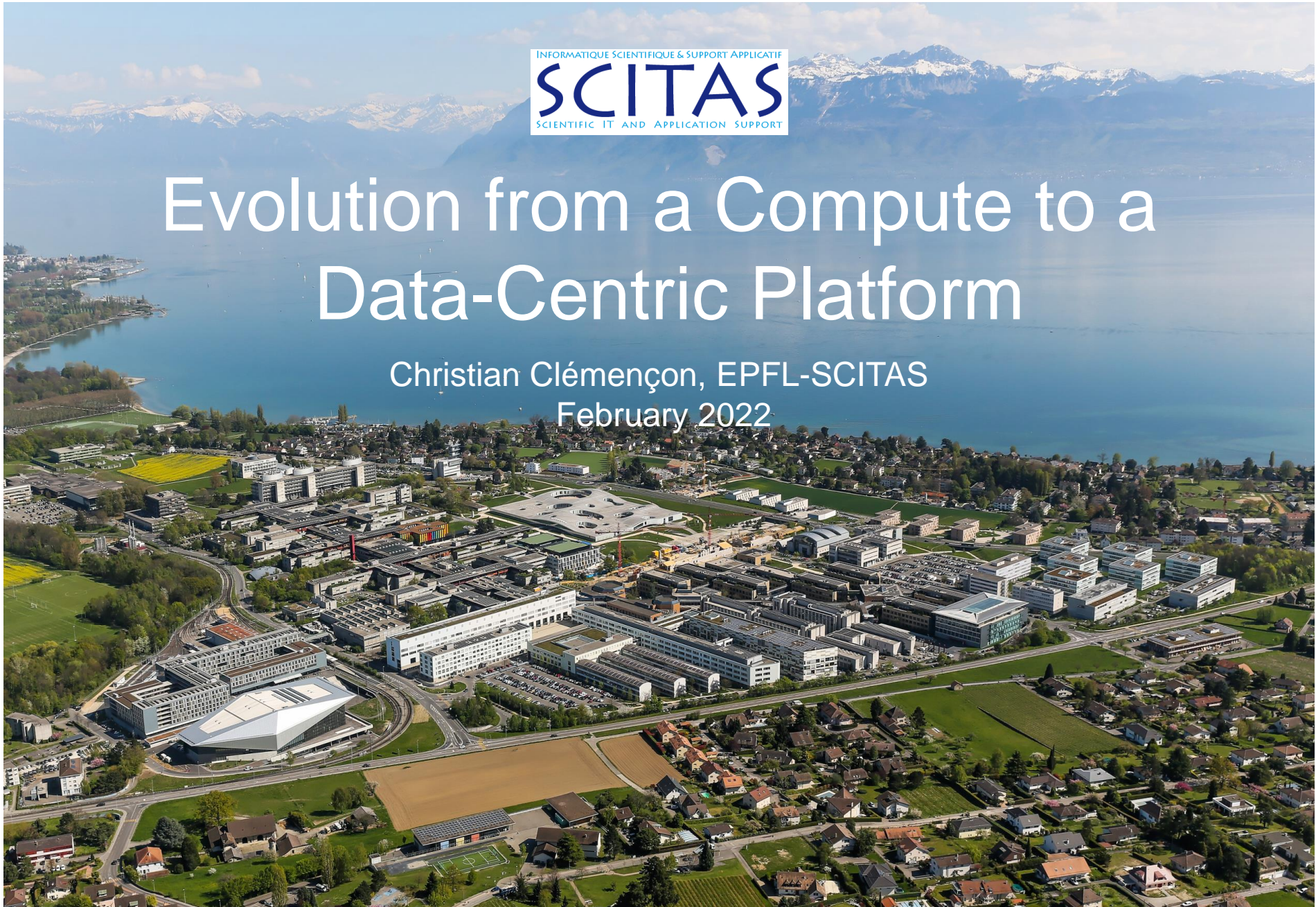
- **SCITAS** work storage and c4science storage for coding/simulations (w/ LFS)

Micha
D'Ans
😊

- **ACOUA** (Academic Output Archive) for data preservation:
 - Data curation
 - Long-term integrity
 - Security

Evolution from a Compute to a Data-Centric Platform

Christian Clémenton, EPFL-SCITAS
February 2022



SCITAS : what is it ?

SCIENTIFIC IT & APPLICATION SYSTEMS

- Check <https://www.epfl.ch/research/facilities/scitas/>
- SCITAS provides Scientific Computing resources and High Performance Computing (HPC) expertise to the EPFL community.
- Five systems specialists, fourteen application specialists

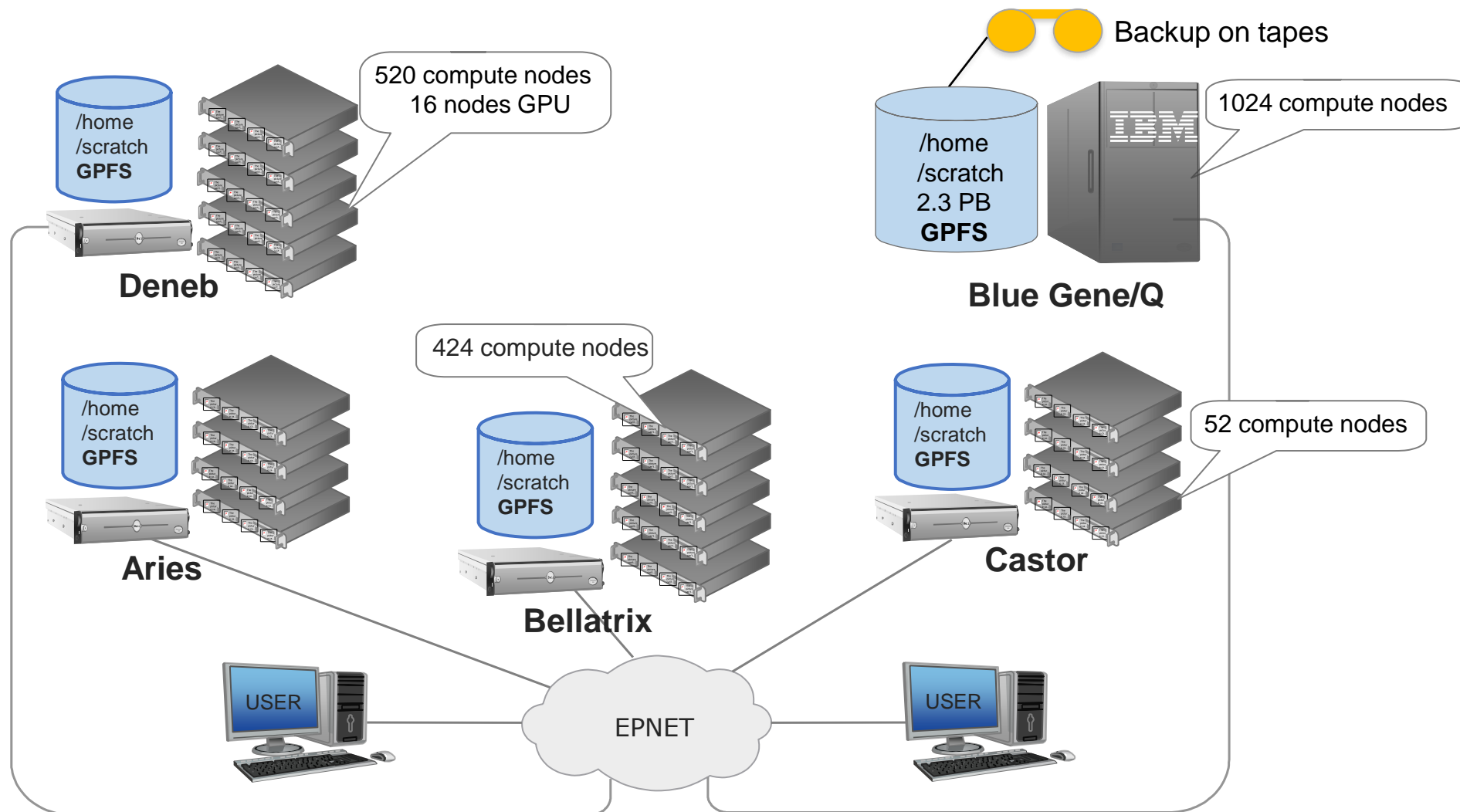
COMPUTING RESOURCES

- Several compute and GPU clusters
- More than thousand compute nodes
- About 5 Pflops of computing power

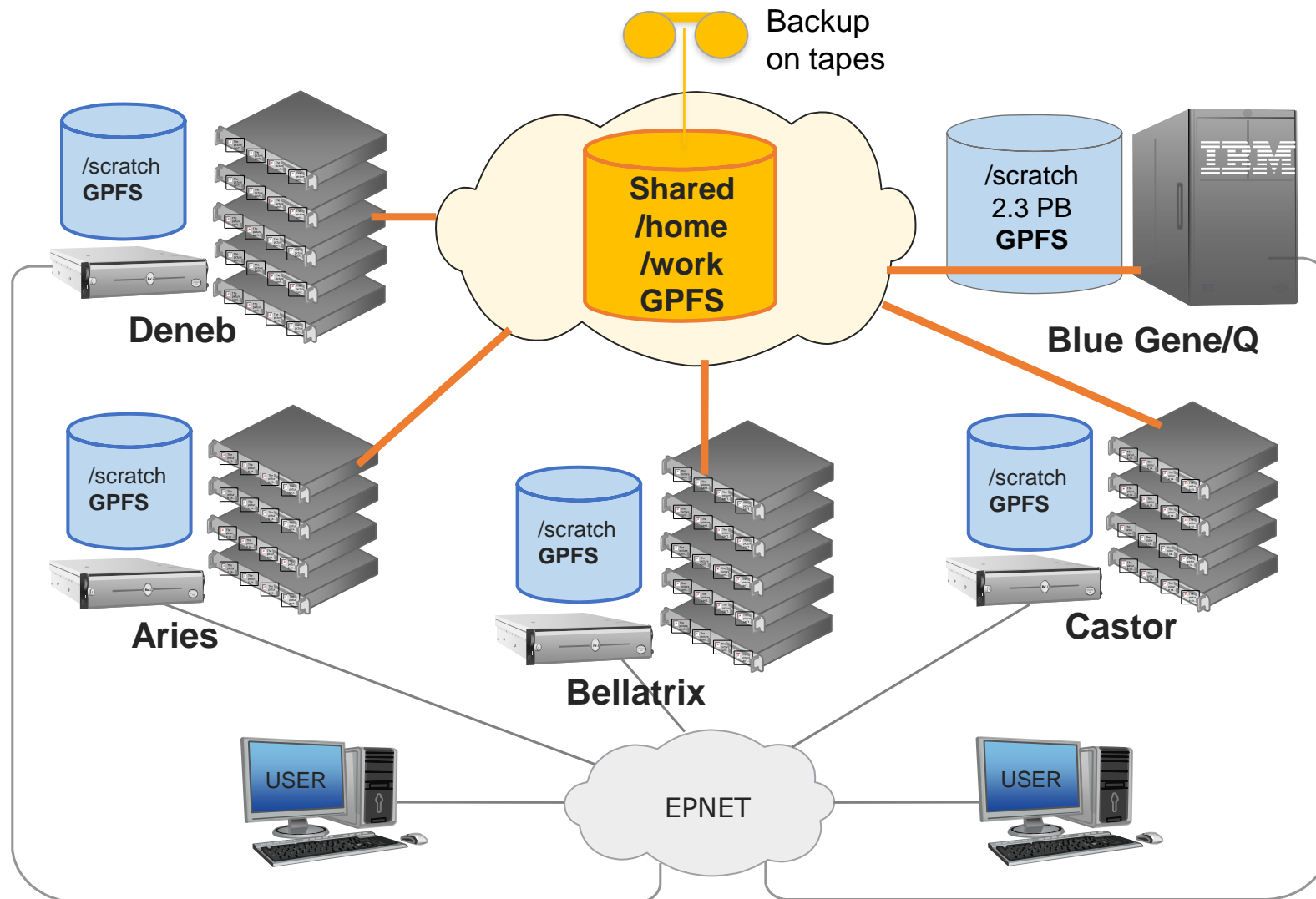
STORAGE SYSTEMS

- Parallel, distributed file systems based on Spectrum Scale IBM technology (aka GPFS)
- Fast, local storage systems for short-term, hot data
- Large, shared storage for mid-term project data
- C4science infrastructure for scientific code co-creation, preservation, sharing and testing

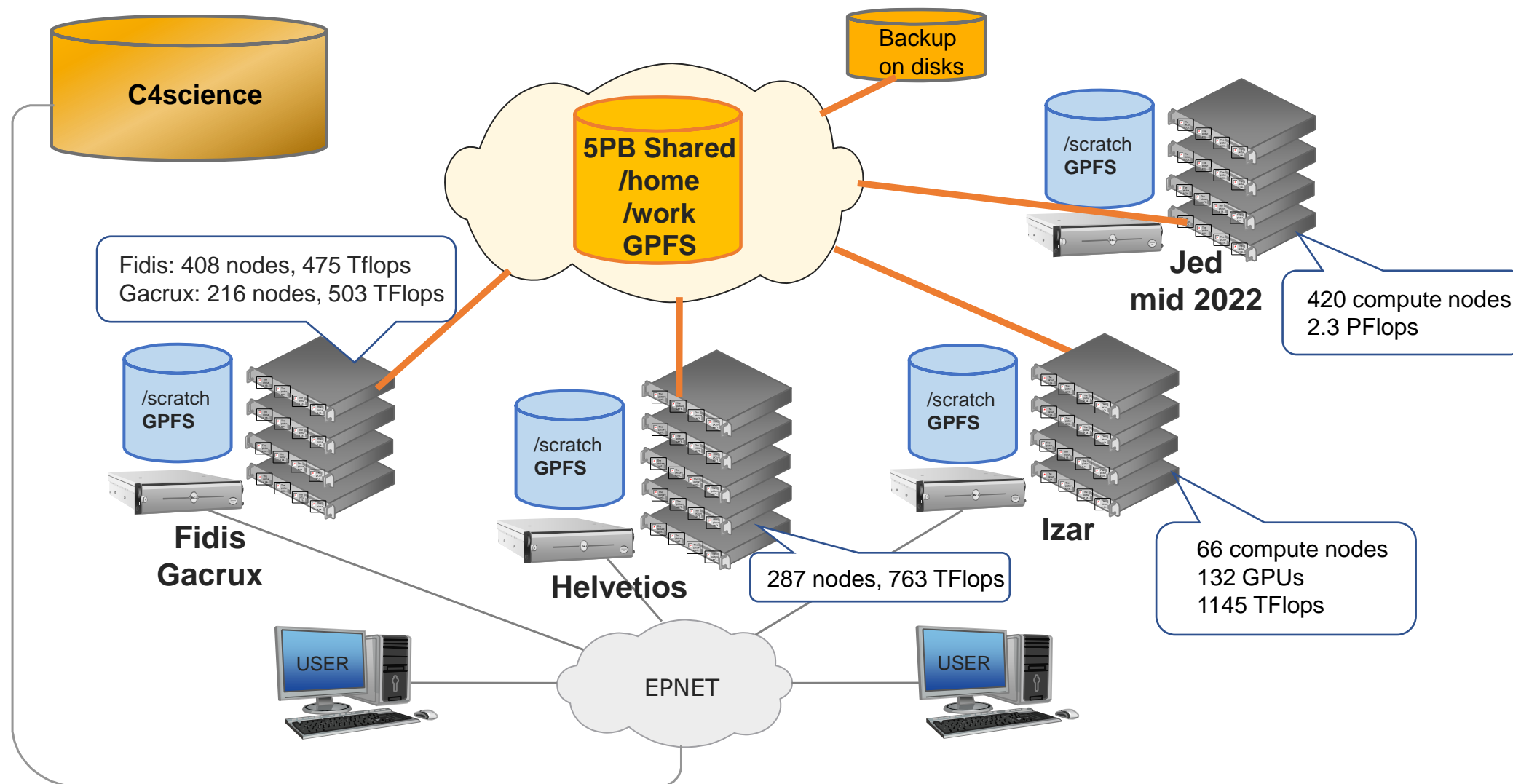
2014 - SCITAS clusters with local storage systems



2015 - SCITAS shared storage system












2022 - SCITAS resources



2022 - SCITAS file systems

File Systems	Purpose	Scope	Size Bandwidth	Backup Snapshots	Lifetime
/home	Login directories – personal files	Global, shared	100 TB ~1GB/s	Backup to disks, 14 x daily snapshots	User account
/work	Project data sets	Global, shared	2 PB 1–4GB/s	Backup upon request, 14 x daily snapshots	~3 years
/scratch	Result files, checkpoints	Local to cluster	~400 TB per cluster 20–40 GB/s	No	2 weeks, automatic deletion using a policy rule
/tmp	Temporary job files	Local to node	~ 1 TB 0.1-1GB/s	No	Job lifetime, automatic deletion

Shared Storage: pros and cons

Data uniformity		Global, consistent /home and /work file name spaces
Data accessibility		Mounted on all cluster's front-end and compute nodes
Data lifetime		Independent of individual cluster's lifetime
Data recoverability		Backups and GPFS snapshots
Manageability		One single point of administration for home directories, software repositories, quotas, backup/restore, monitoring
Extensibility		Scale-out architecture, i.e., to increase storage capacity
Scalability		~1000 GPFS client nodes and ~1000 users
Performance predictability		I/O performance is subject to variations due to the shared nature of the storage
Complexity		More complex to set-up, especially the shared data transfer network

C4science

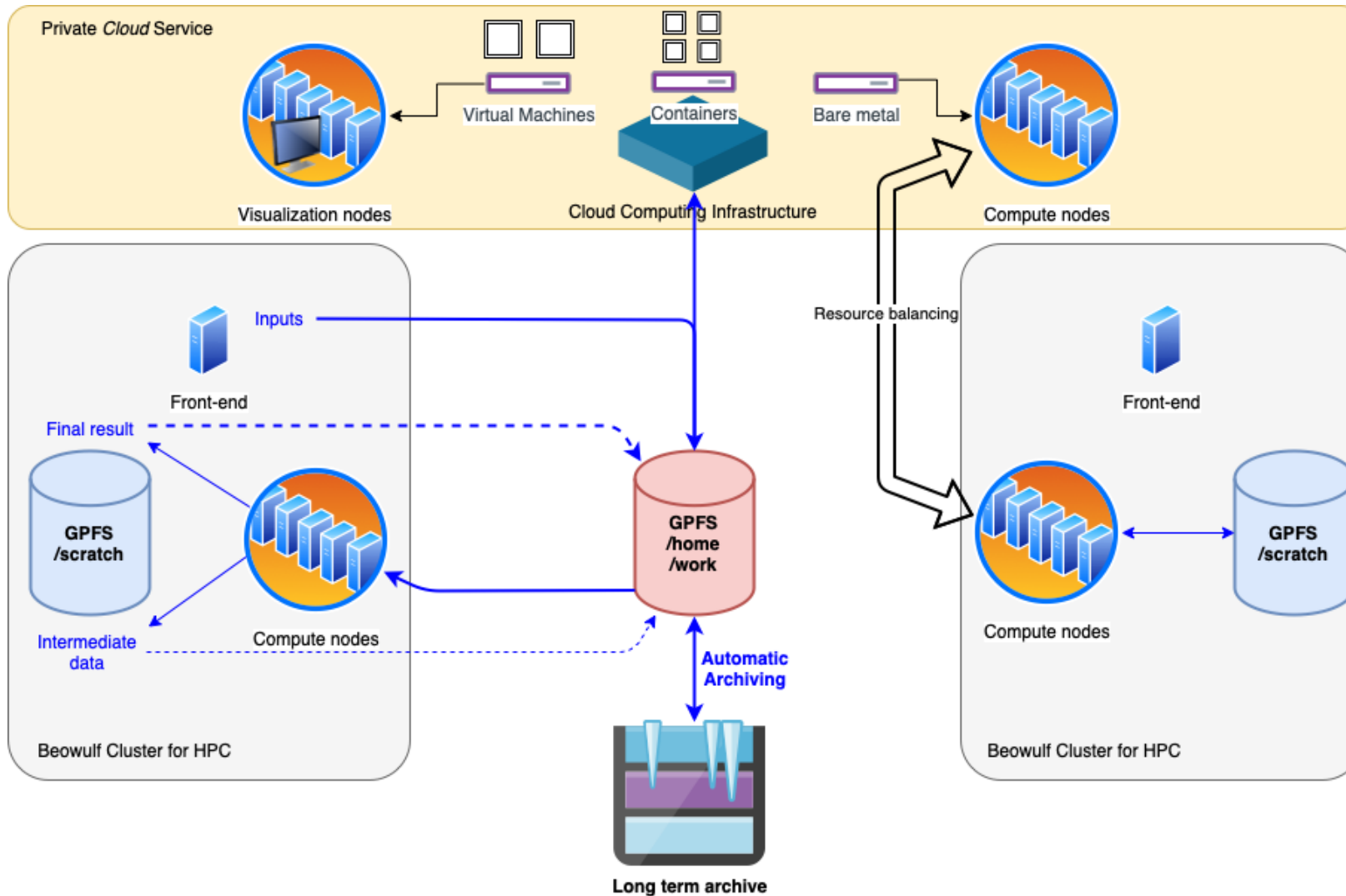
Phabricator (c4science engine)

- Provides “github-like” software repository for all EPFL users
- Continuous integration and deployment (CI/CD)
- Currently **6'000+ users** and **10'000 repositories**,
- **5 TB storage**

A new project for migration to another service

- Migration of all user data to a new gitlab solution
- Merge with gitlab.epfl.ch

Towards a platform for the production, exchange and archiving of scientific data

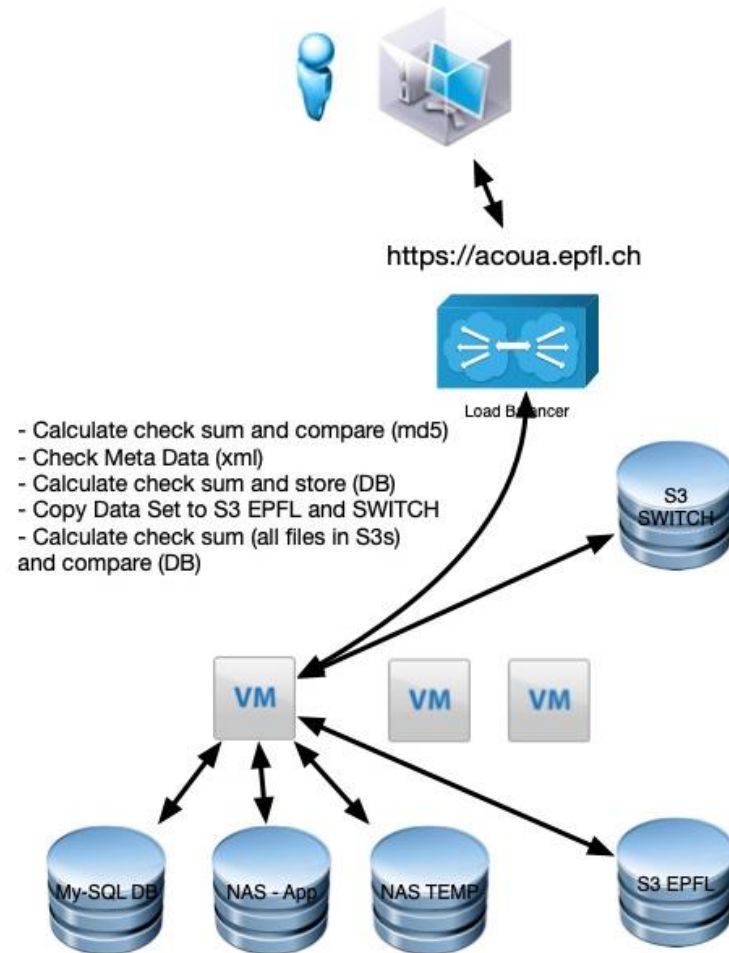


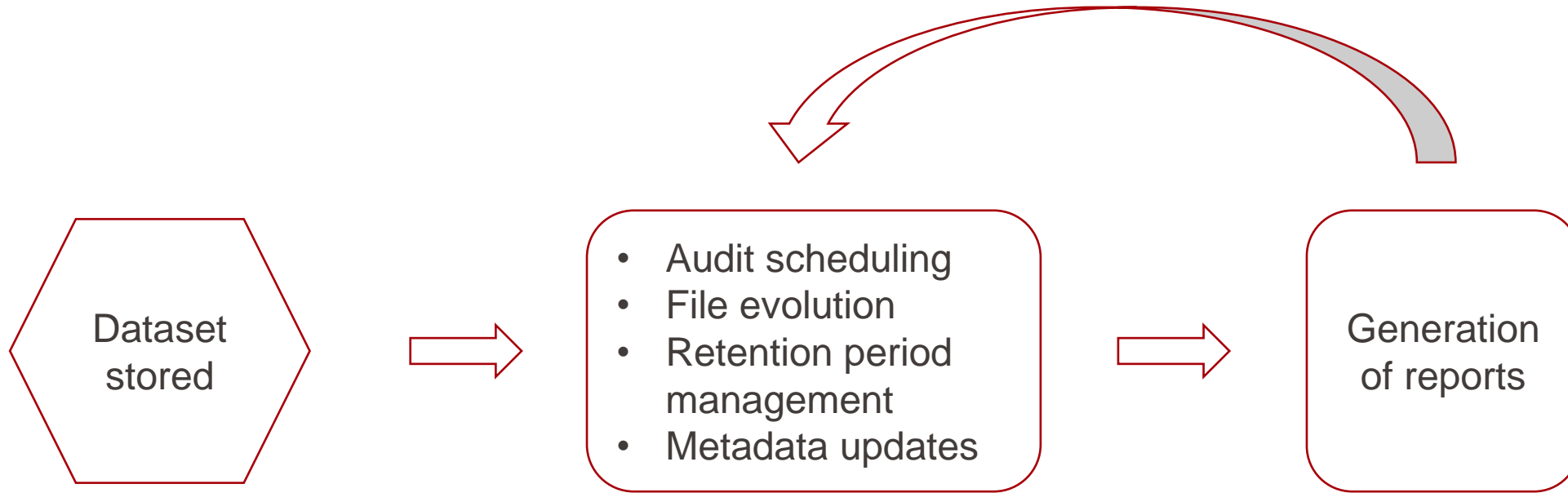
ACOUA = **AC**ademic **O**utput **A**rchive

- Long-term preservation of research data
 - Datasets linked to publications
 - Data from labs about to close
- Active preservation with customizable parameters
- Assistance and data curation by a dedicated team
- Free Service
- Preservation of large data sets (up to 10 TB)
- Storage on two servers (EPFL and Switch)

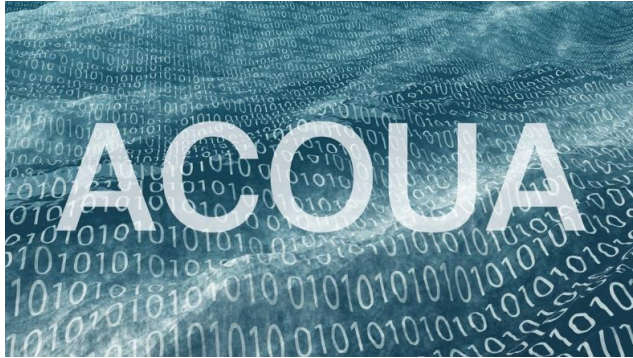
Planned and well-thought actions on the long-run
to mitigate:

- Obsolescence
 - Hardware
 - Software
 - File format
- Loss of the meaning of the content





Retrieve



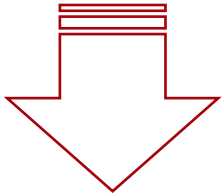
Preview & download
from platform

or

Pre-signed URL



Connector

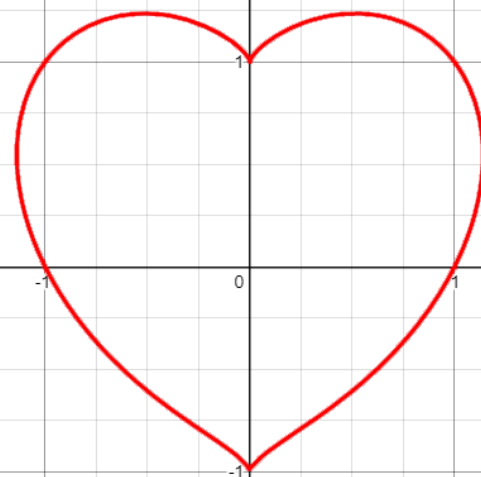


Publish

- **acoua@epfl.ch**
- **go.epfl.ch/acoua**

Love Data Week

EPFL 2♥22 edition



Thu. Feb. 17 | 11:00-11:45
Online training "*Which platforms to publish my research data/code?*"

Fri. Feb. 18 | 10:00-12:00
Online speed-dating with EPFL Data Champions

go.epfl.ch/ldw22