



# **More Than Data: Computing Services for Synchrotron Users**

Oscar Matilla – Computing Division Head @ ALBA

XI AUSE Conference and VI ALBA Users Meeting

05/09/2024



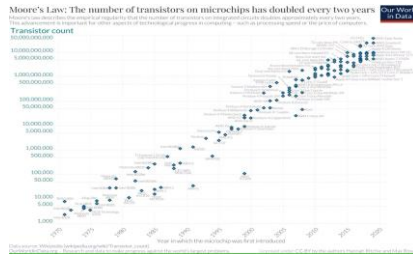
Generalitat de Catalunya  
**Departament de Recerca  
i Universitats**



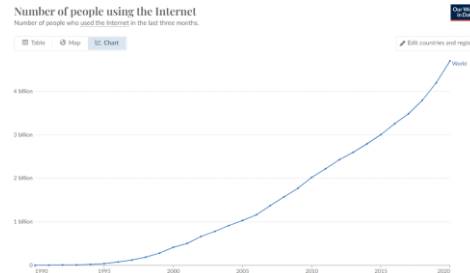
MINISTERIO  
DE CIENCIA, INNOVACIÓN  
Y UNIVERSIDADES

# Computing is experiencing exponential growth

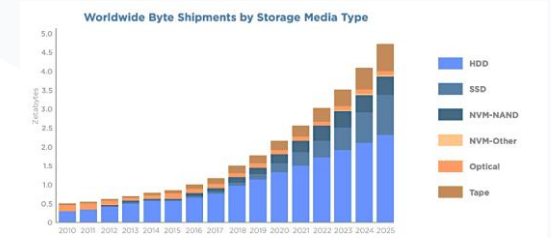
## Processing power



## Connectivity

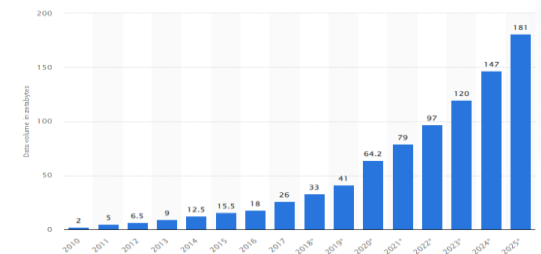
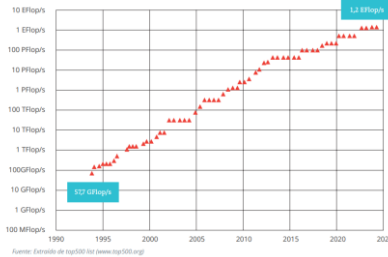


## Data Volume



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Ilustración 1. Desarrollo de la supercomputación entre 1993-2023.



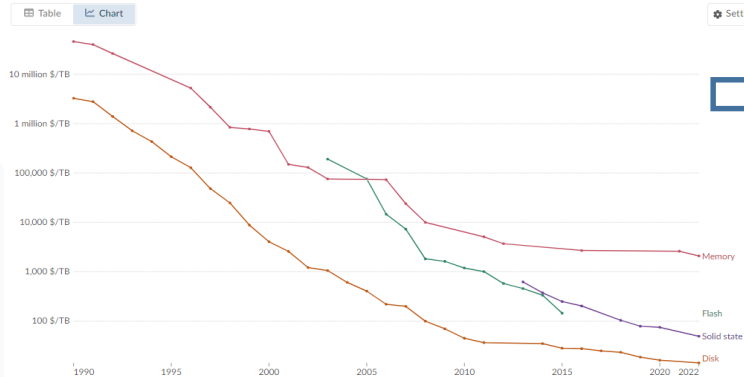
Data generated by year world-wide

# Computing is experiencing exponential growth

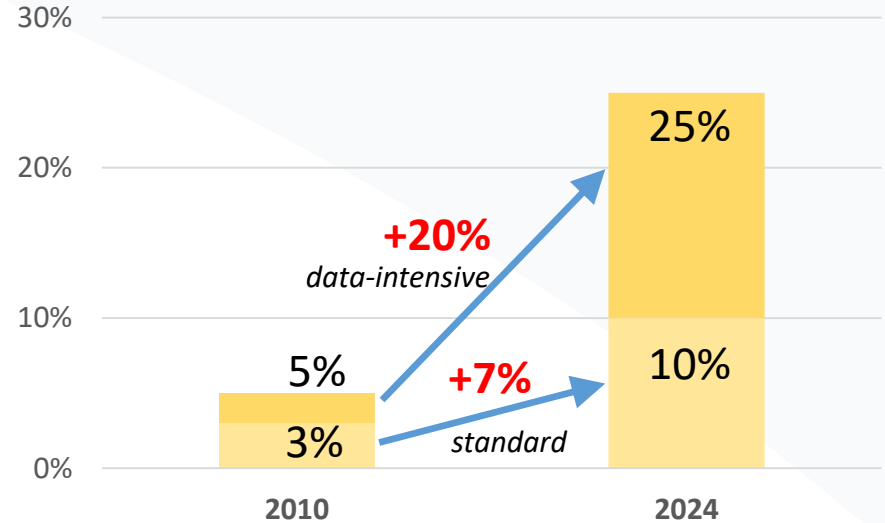
... what about the  
**costs?**

## Historical cost of computer memory and storage

This data is expressed in US dollars per terabyte (TB). It is not adjusted for inflation.

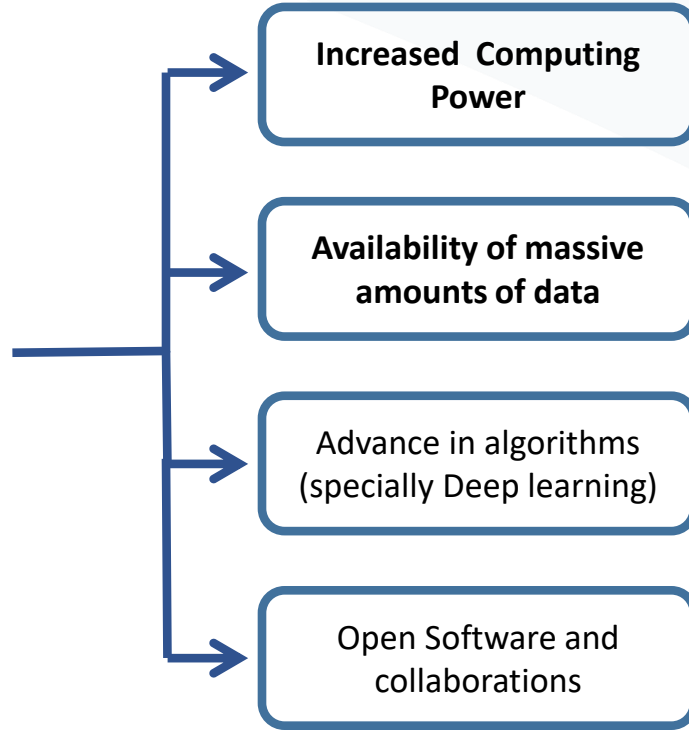
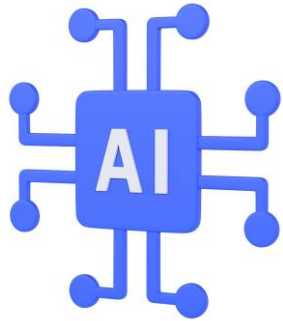


## % IT investment respect total budget @ new beamlines



- Computing architecture design must be part of the beamline design from its conception.

# Artificial Intelligence expansion



## European Strategy for Data

A common European data space, a single market for data

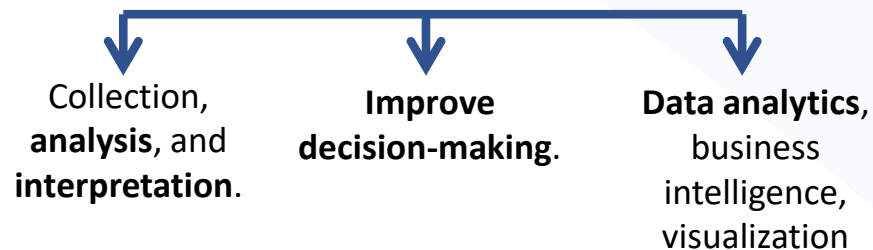


## • European Data Space (EDS):

- Establish a **common framework** for sharing and utilizing **data** while respecting data protection and privacy.
- Improve **accessibility**, **interoperability**, and **reusability**.



## Data-drivenness



# Data Challenges On the Horizon

**Data  
Challenges**

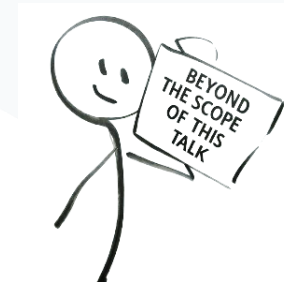


**Data Volume**



# Data Challenges - Data volume

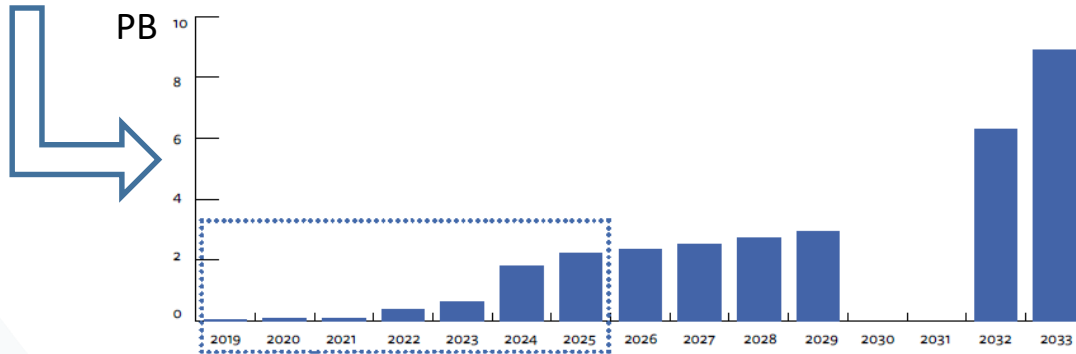
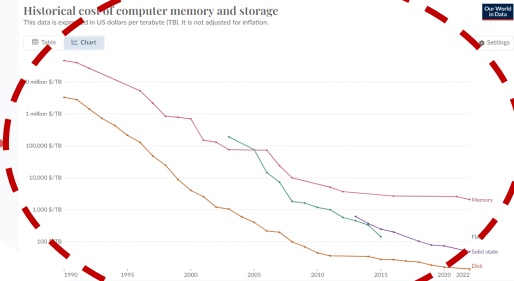
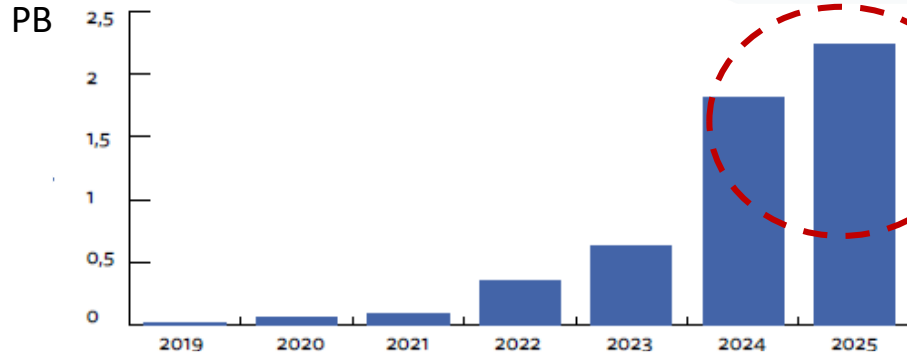
- The volume of data generated is soaring dramatically due to various factors:
  - The evolution of X-Ray detectors



- Real data storage figures of Eiger 16M falls into 30 TB/day.
  - New integrating detectors (as CITIUS) are already generating 5 PB/h raw data (\* which must be dramatically reduced).
- The increased automatization of the sample environment.
- The increased brilliance of the sources.

# Data Challenges - Data volume

## ALBA's Yearly Raw Data Output (Current and Projected)



Improved data  
reduction  
algorithms will be  
required



- The weighty data has consequences:

## Users

- Huge data volumes (**Terabytes**)
- Sample metadata
- Raw data quality
- Data processing
- Data exporting

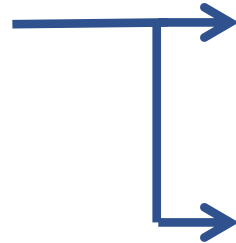


## Facilities

- Huge data volumes (**Petabytes**)
- Data acquisition
- Metadata collection
- Data curation
- Data archiving

# Data Challenges On the Horizon

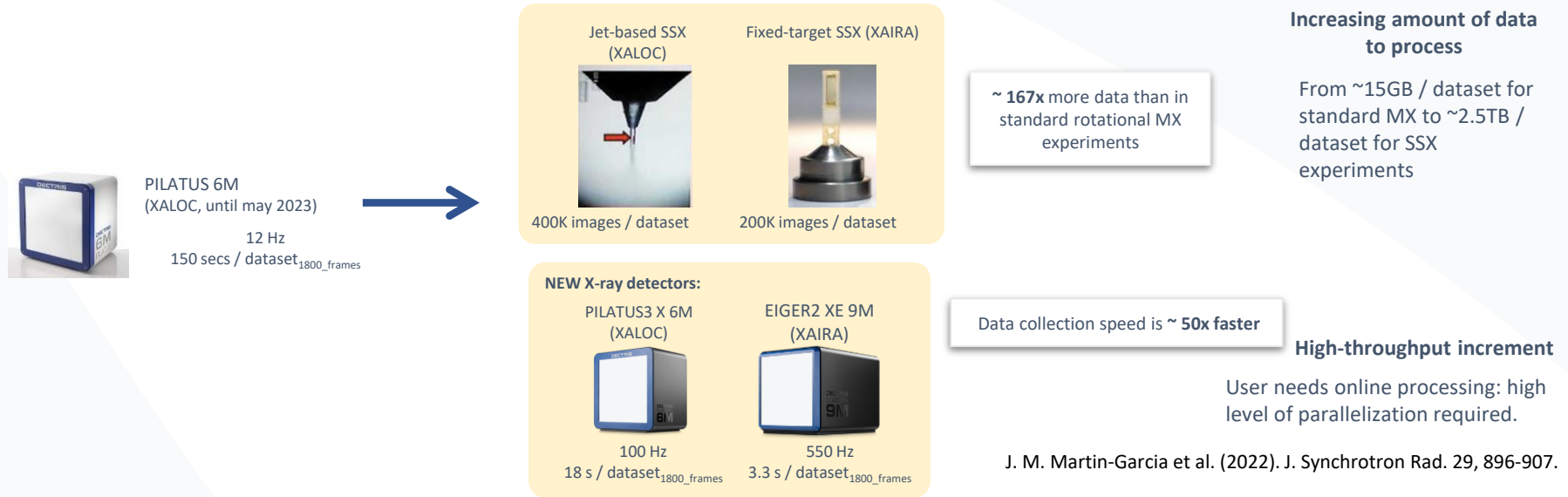
**Data  
Challenges**



**Data Volume**

**Data Processing**

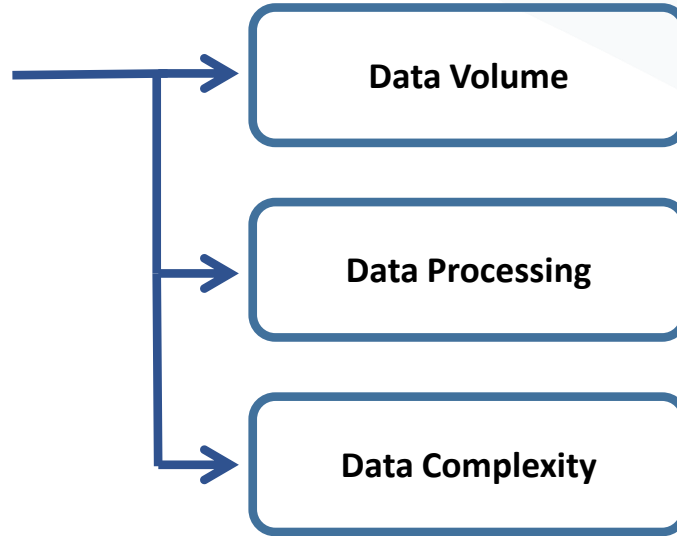
- **Optimized use of an HPC cluster becomes imperative** when handling extensive datasets and faster experiments.
  - Example of HPC demand for Serial Crystallography in MX beamlines:



\* **Extended insights Nicolas Soler (Section Head Scientific Data Management)**

# Data Challenges On the Horizon

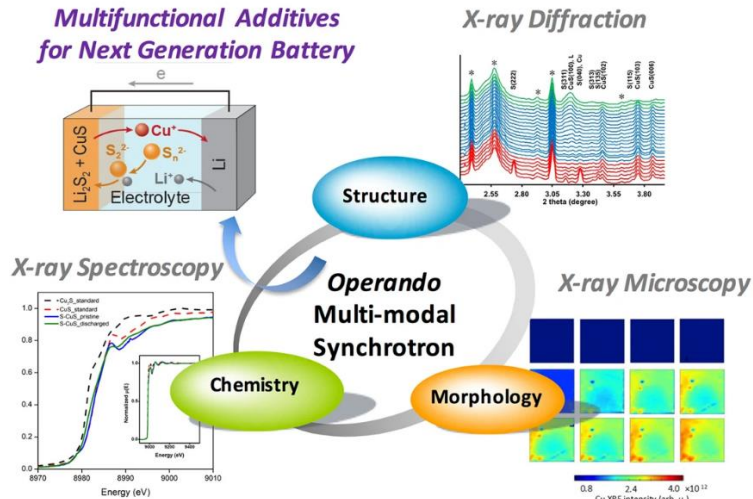
## Data Challenges



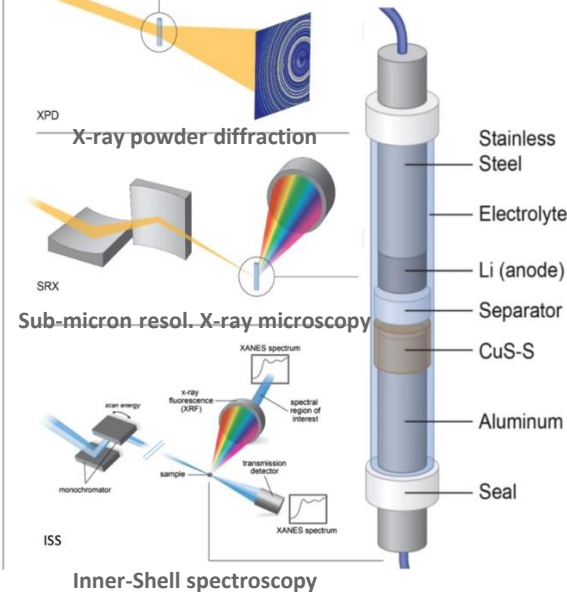
- In addition to data volume, the data will become **increasingly complex** due to:
  - Combining multiple techniques (multimodal)
    - Example: combination of X-ray tomography, spectroscopy, and diffraction
  - Assessing multiple scales and dimensions (increased spatial and temporal resolution)
    - Example: time-resolved SSX, SAX/WAX, tomography
  - Combining experiments with computer simulation
    - Example: Density Functional Theory (DFT), molecular dynamics (MD)

# Data Challenges – Data Complexity

A

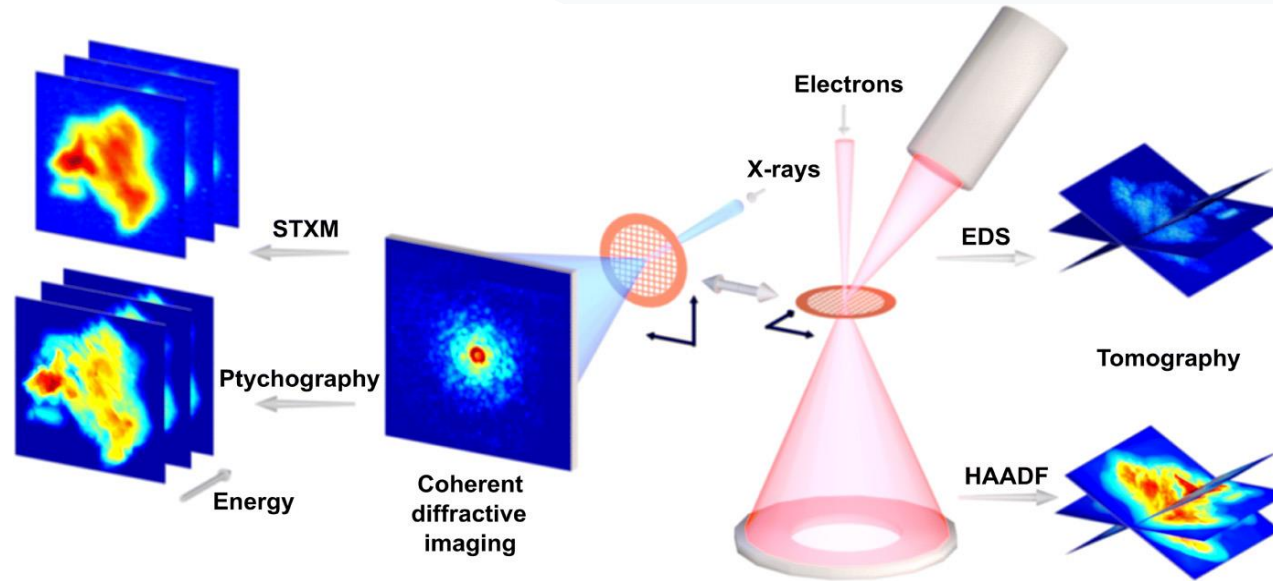


B



**Example:** Integrating data from different beamlines at NSLS-II for studying a new type of battery.  
*Sun. K. et al., Scientific Reports | 7: 12976 | DOI:10.1038/s41598-017-12738-0*

# Data Challenges – Data Complexity



**Example:** Correlative imaging on the Allende Meteorite at ALS and Lawrence Berkeley National Laboratory  
*Lo et al., Sci. Adv. 2019; 5 : 20 September 2019*

# Equipping Ourselves: The Human and Infrastructure Needs



## Computing Division

**IT Systems Section  
(12)**

**Controls and DAQ Section  
(19)**

**MIS Section  
(10)**

**Scientific Data  
Management Section\*  
(7)**

## Experiments Division

**Life Sciences Section**

**Electronics and Magnetic  
Structure of Matter  
Section**

**Chemistry and Material  
Science Section**

**Interdisciplinary and  
Multimodal Section\***

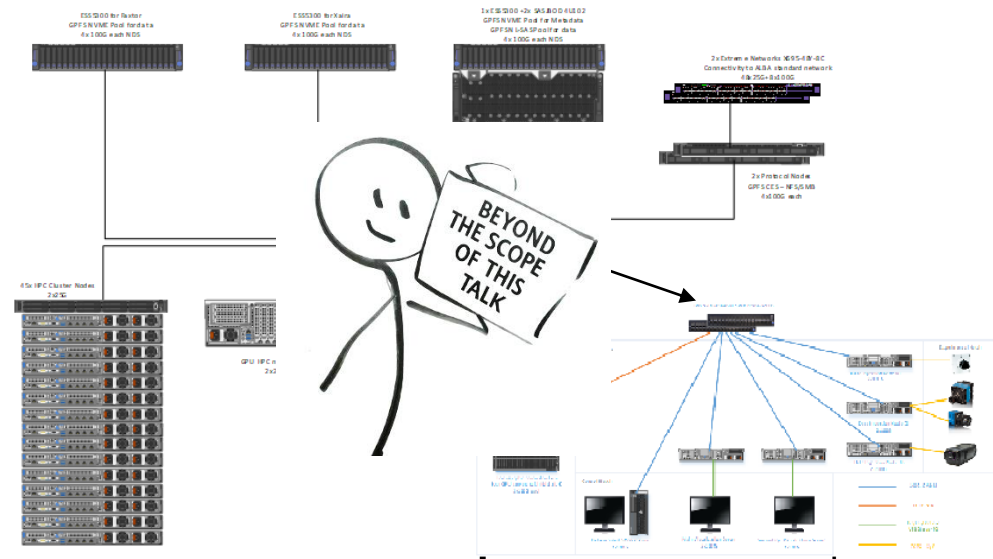
## Users





# Equipping Ourselves: The Human and Infrastructure Needs

- Such a dramatic need for an increase in performance necessitates **profound architectural changes**. From the **hardware** up to the **software** stack.

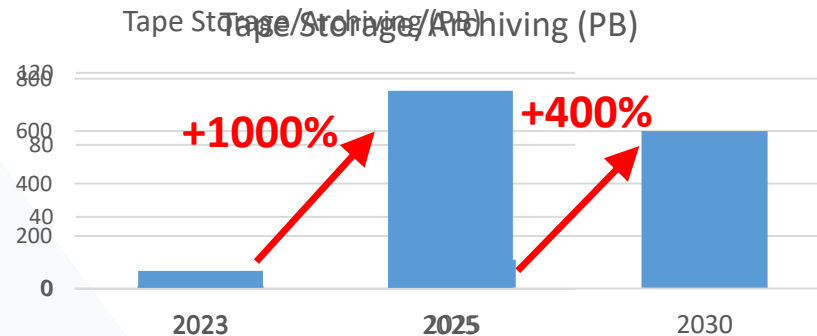
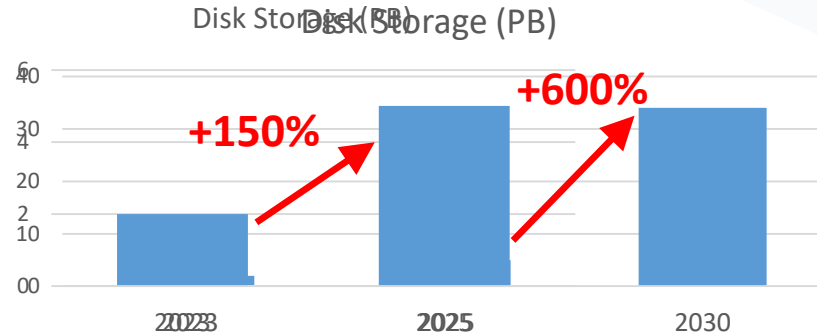


- An ambitious IT investment program is underway to support this plan.

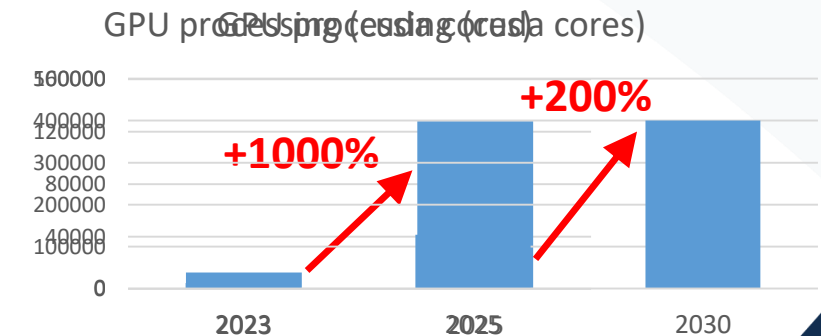
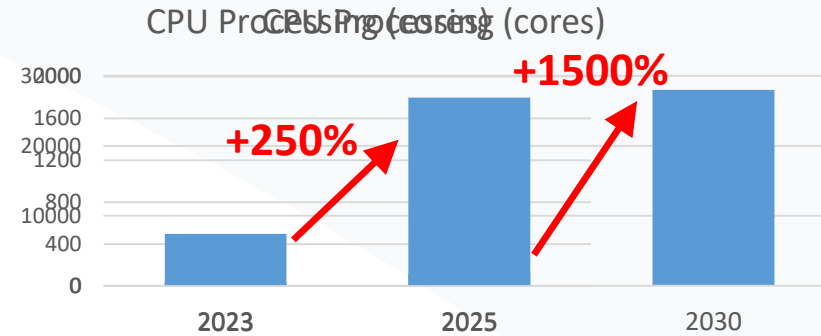
# Equipping Ourselves: The Human and Infrastructure Needs



## Data Storage



## Data Processing



- In times of trouble, love takes a backseat.

↳ **A siloed approach to data management will hinder progress.**

## Data Journey from Collection to Analysis and Dissemination

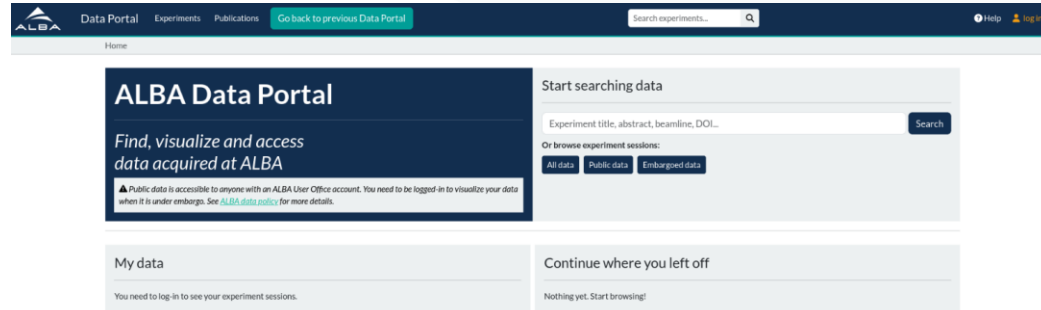


# ICAT Data Catalogue



- ICAT Data Catalogue **in production** since starting in 2024.

<https://data.cells.es>



The data catalogue allows users to browse data directly, which is a significant improvement over the previous sftp-only access method.

Experiment sessions

Search: name, title, abstract, DOI...

Filter by: Public data Embargoed data

My data

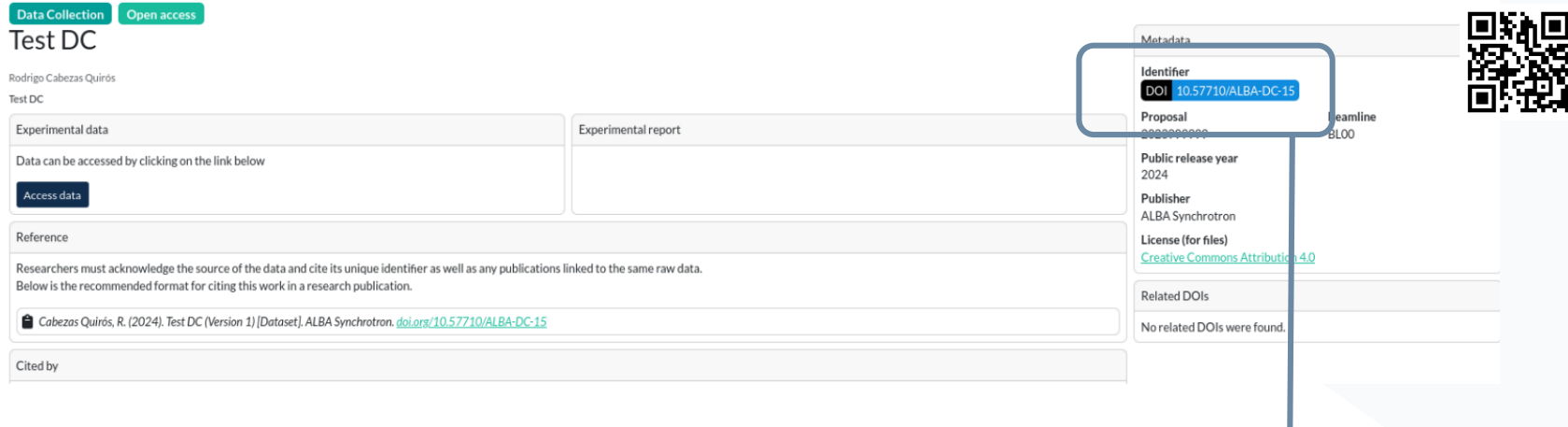
Beamline: any

Start date: 24/08/2024

End date: 24/08/2024

User: Select a user...

Beamline	Start	Title	A-Form	Samples	Datasets	Files	Release	DOI	Logbook
2024028206	BL11	18/07/2024	Operando SAXS/WAXS Investigation of Silicon Anode Morphology evolution upon cycling, and effect of binder selection for High-Energy Li-ion Batteries				21/07/2027		
2024028317	BL01	16/07/2024	Evaluation of lipid and protein alterations in the secondary brain lesion of a preclinical model of intracerebral hemorrhage by $\mu$ FTIR				20/07/2027		
2024028359	BL20	16/07/2024	Spin and Angle-Resolved photoemission investigations on the spin texture of magnetically modified topological surface states	14 265.58 MB	14 265.58 MB	14	20/07/2027	<a href="#">DOI: 10.57718/ALBA-ES-2024028359</a>	
2024028174	BL11	11/07/2024	Unveiling the mechanism of hydrometallurgical recycling of neodymium from permanent magnets using phosphonium ionic liquids				11/07/2027		
2024028232	BL20	10/07/2024	Atomic mechanisms of electron doping in perovskite nickelates	254 3.98 GB	270 3.98 GB	270	14/07/2027	<a href="#">DOI: 10.57718/ALBA-ES-2024028232</a>	
2024028286	BL01	09/07/2024	Structural changes in bioinspired thermoplastics during aging studies in the marine environment				11/07/2027		
2024046477	BL01	05/07/2024	Feasibility test for "Investigating the interaction between microstructure and mechanical performance in MEW Soft Robotic Actuators I (MIRAS part)"				13/07/2027		
2024028081	BL11	05/07/2024	Investigating coiled coil-based self-assembled structures				07/07/2027		
2024028144	BL11	03/07/2024	Correlation between nature, loading and spatial confinement of biomolecules within Metal-Organic Frameworks				05/07/2027		



The screenshot shows the ICAT Data Catalogue interface for a dataset named 'Test DC'. At the top, there are two buttons: 'Data Collection' and 'Open access'. Below the dataset name, the user 'Rodrigo Cabezas Quirós' is listed. The main content area is divided into two columns. The left column contains 'Experimental data' and 'Reference' sections. The 'Experimental data' section has a text box stating 'Data can be accessed by clicking on the link below' and an 'Access data' button. The 'Reference' section contains a text box with instructions on how to cite the data and a citation example: 'Cabezas, Quirós, R. (2024). Test DC (Version 1) [Dataset]. ALBA Synchrotron. [doi.org/10.57710/ALBA-DC-15](https://doi.org/10.57710/ALBA-DC-15)'. The right column contains 'Experimental report' and 'Metadata' sections. The 'Metadata' section is highlighted with a blue box and a line pointing to the 'Identifier' field, which displays 'DOI: 10.57710/ALBA-DC-15'. Other metadata fields include 'Proposal', 'Public release year', 'Publisher', 'License (for files)', and 'Related DOIs'. A QR code is located to the right of the metadata section.

- Once the user's data are collected, their entire experiment is referenced by a **Digital Object Identifier (DOI)** and a **landing page**.
- The user can also select a **subset of datasets** they want and mint them with a **DOI** to accompany their **publication**.

- **Rich metadata capture**
  - As automatically as possible.
- **NeXus / HDF5**

## Encapsulating raw data and rich metadata:

NeXus is a convenient data format, providing **metadata recipes** for different techniques (called application definition).

Such a **self-contained format**, relying on **HDF5**, is becoming increasingly common among PaN facilities and is being implemented in our beamlines (e.g. default format for LOREA and XAIRA, implemented gradually in other beamlines, along ICAT implementation).

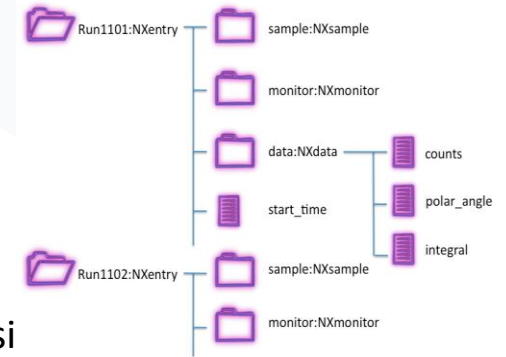
Introduction:



Applications definition:

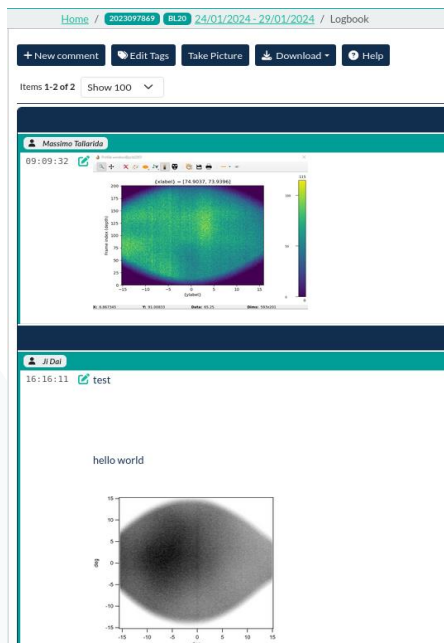


<https://data.cells.es>

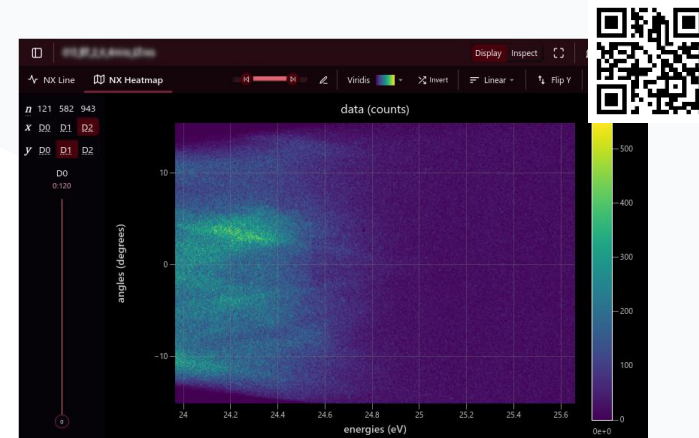


- Further functionalities:

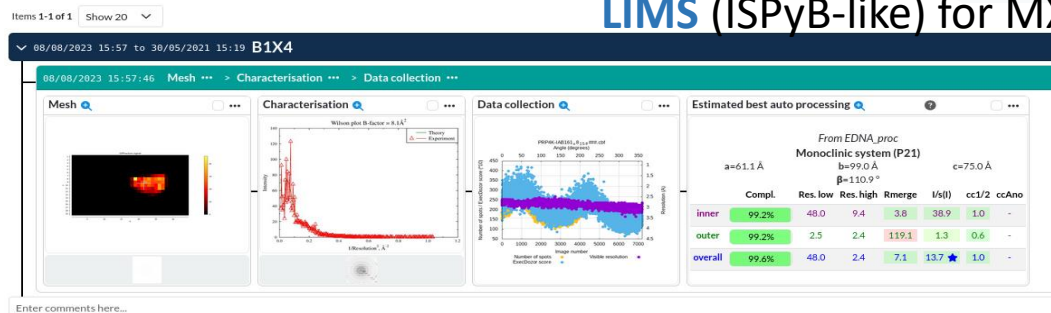
**e-logbook** to annotate the experiment while it is running



NeXus data  
visualiser



**LIMS** (ISPyB-like) for MX





- Implementation progress:

In  
production



End 2024/  
Begin. 2025

- BL20 LOREA
- BL06 XAIRA, BL09 MISTRAL, BL16 NOTOS, BL24 CIRCE, BL29 BOREAS
- All other beamlines to follow shortly after.



## Users

- Huge data volumes (**Terabytes**)
- Sample metadata
- Raw data quality
- Data processing
- Data exporting



- We want to enable users to analyze their data **using ALBA's high-performance computing resources.**
  - Access methods:
    - Jupyterhub
    - VISA

**VISA** is a virtual compute instance infrastructure allowing to access experimental data and processing software **from anywhere**.

This is particularly useful in case the data is too big to move around or too CPU/GPU expensive to process at user's home institutes.



## Data Analysis, in the cloud

VISA (Virtual Infrastructure for Scientific Analysis) makes it simple to create compute instances on the data analysis infrastructure to analyse your experimental data using just your web browser

[Sign in with your user account](#)

### Analyse your data

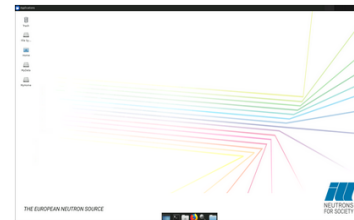
Create a new [compute instance](#) and use your web browser to access a Remote Desktop or JupyterLab to start analysing your experimental data

### Collaborate with your team

Share your compute instance with other members of your team to [collaborate together](#) in real time

### No need to install software

The compute instances come with pre-installed [data analysis software](#) so you can start analysing your experimental data immediately



### Questions or feedback?

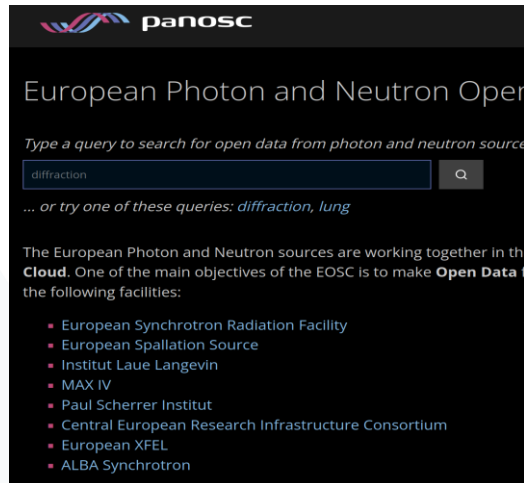
Please send the IT service an email to [mis@cells.es](mailto:mis@cells.es)

- Implementation progress:
  - **VISA** platform is **in production** but still **not available to users**.

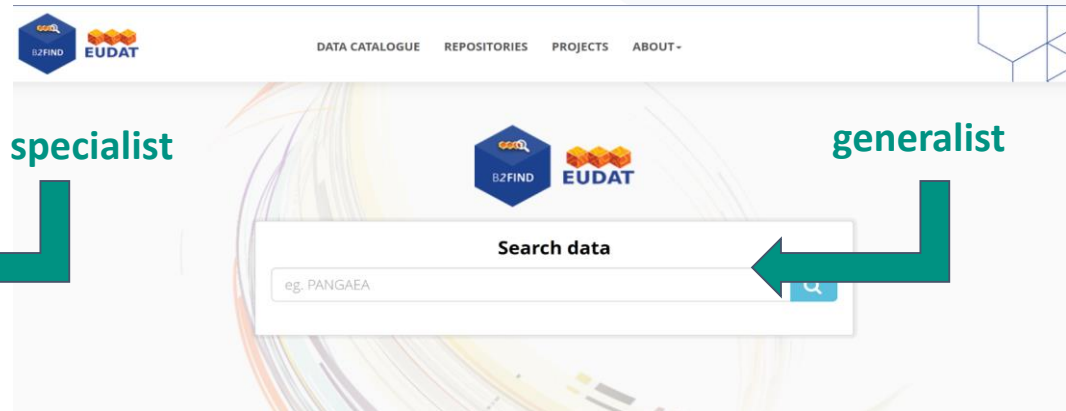
# Photon and Neutron Data Space



- We aim to foster the **creation of a European PaN Data Space** and to **facilitate Spanish users'** access to it.
- Our data catalogue seamlessly **integrates with the PanOSC data catalogue**, allowing for easy discovery through general search engines.



<https://data.panosc.eu/>



<https://b2find.eudat.eu>



- We are actively promoting the creation of a PaN Node to connect to the central EU-Node



- The current proposal involves DESY, Elettra, ESRF, HZBR, SOLEIL, and ALBA, with the possibility of others joining later.

- This marks the initial Build-up phase of the EOSC Federation.



- **How will this benefit our users?**
  - ALBA users, as part of this data space, **will gain access to all the open data stored across these facilities**, plus access to their computing infrastructure to **analyze the open data**.

- The **exploding volume** and **complexity** of synchrotron **data** necessitate architectural changes in data management and computing infrastructures.
- ALBA is proactively **addressing these challenges** through strategic **investments** in hardware, software, and human resources.
- Key steps towards efficient data access, processing and analysis are the **ICAT Data Catalog**, adopting the **Nexus/HDF5 data format**, and developing the **VISA platform**.
- ALBA is actively participating in the **Photon and Neutron Data Space** initiative under the **EOSC**, fostering data sharing within the European research community and providing the Spanish scientists community with access to this **valuable resource**.
- Synchrotron science and computing are deeply connected. By using a data-driven approach, **ALBA aims to give researchers the tools they need to make new discoveries and drive innovation**.