

*ALBA-II day*



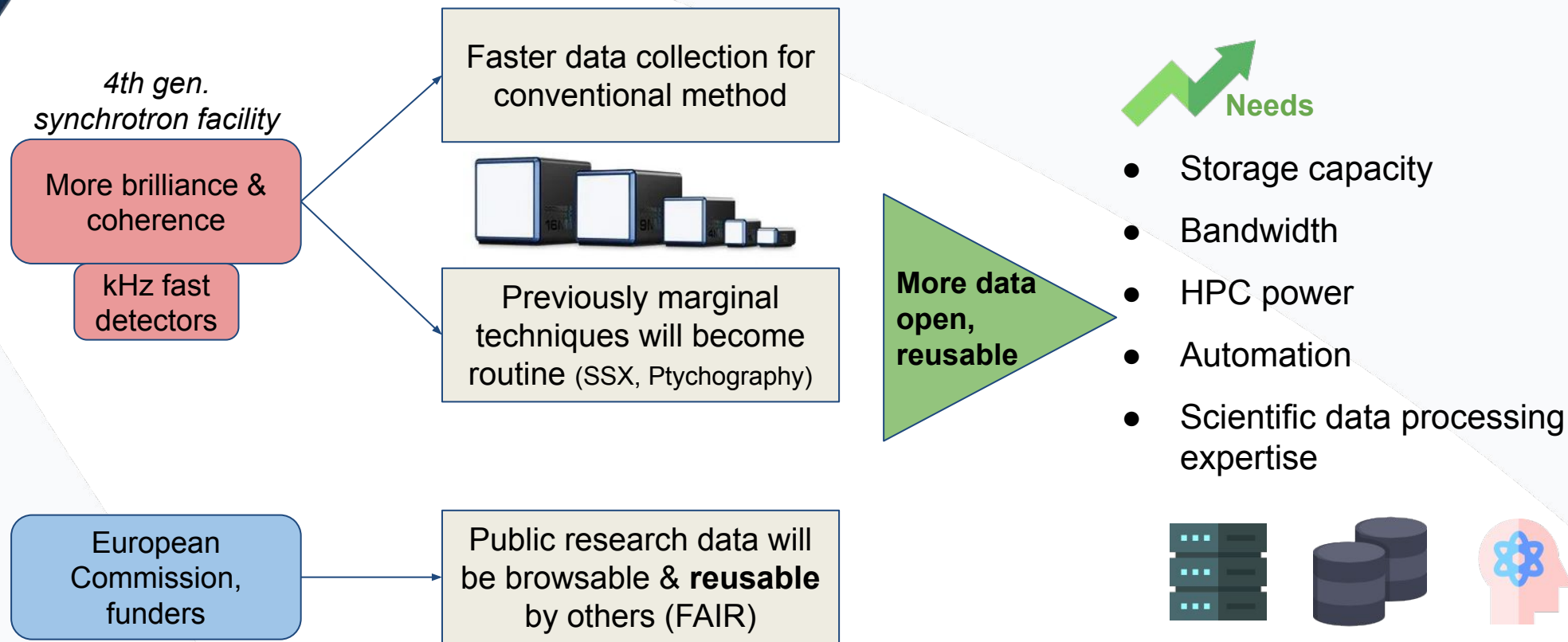
# Scientific Computing in 4th gen synchrotrons

Nicolas Soler

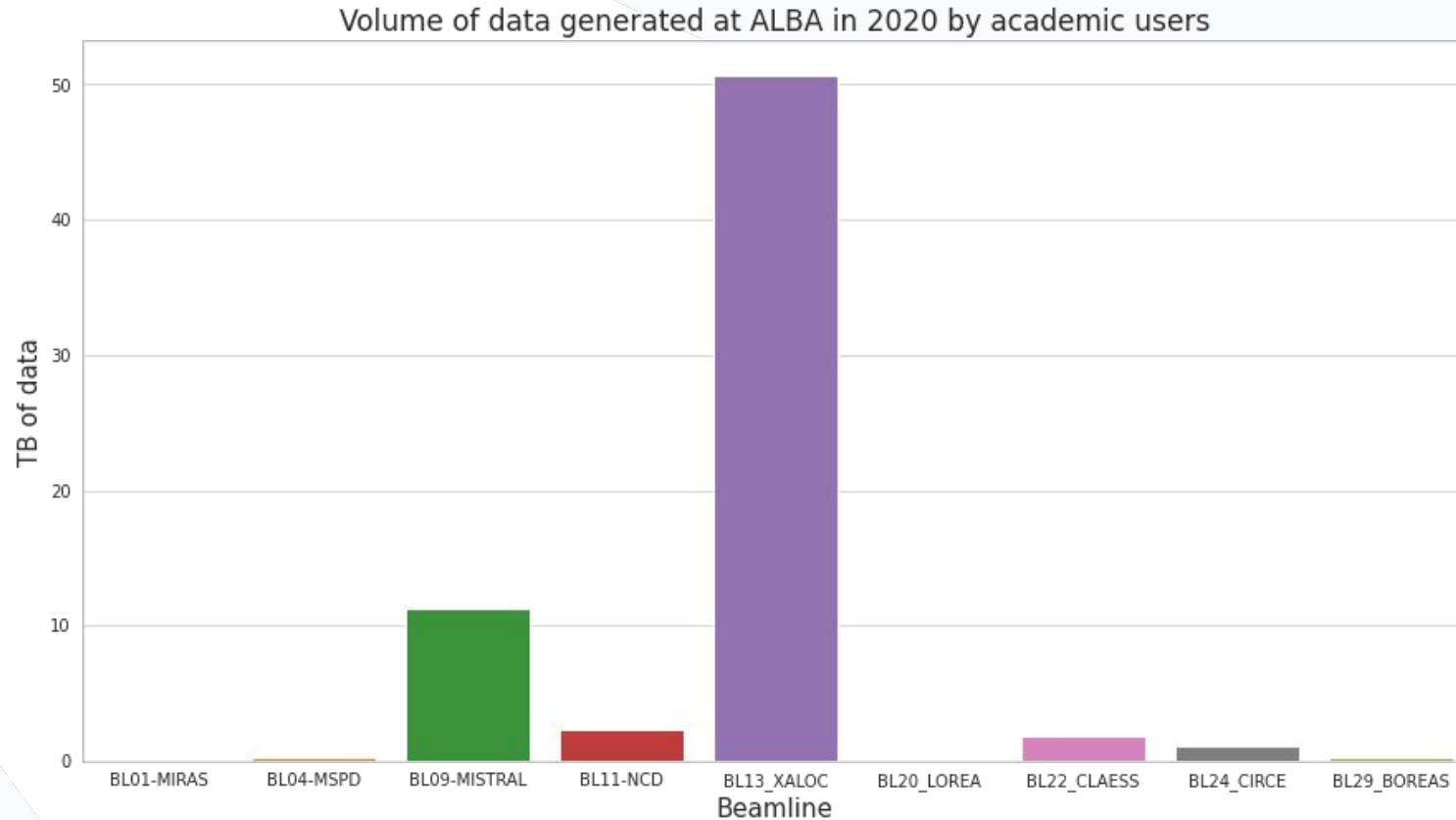
Scientific Data Management section (Computing division)

30 June 2021

# In a nutshell

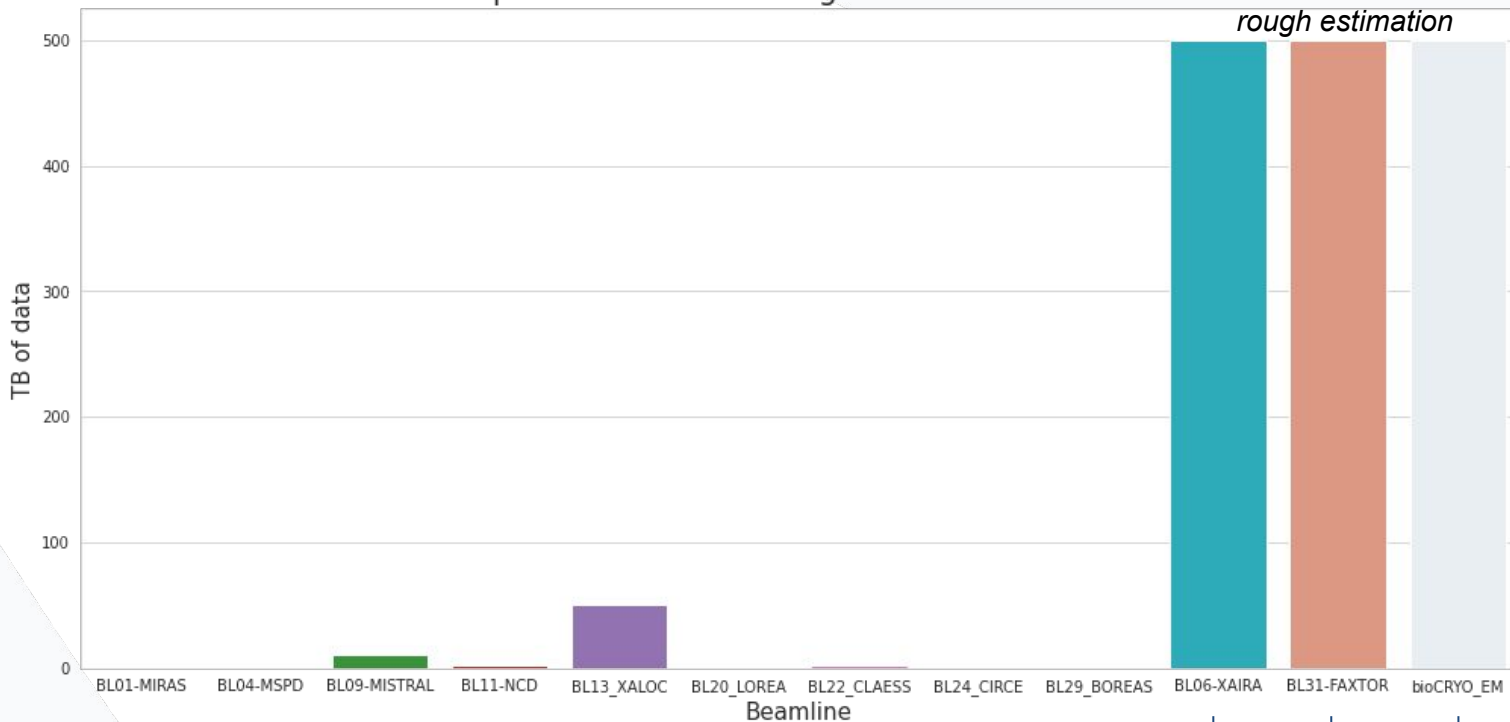


# Annual data volume generated



# But this will change soon:

Comparison with future high data-rate beamlines



MX  
SSX

$\mu$ -com  
puted  
tomo


CryoEM  
SPA,  
tomo

- + 2 other new beamlines (NOTOS, MINERVA)
- + 2 material science microscopes.

# Data @ 4th generation synchrotrons



## *Changes*

- Annual data **volume will increase** by an order of magnitude, at least.
- Users won't be able anymore to **download, process and analyze** their data at their home laboratory.
- Public research data **must be FAIR**. (Reusable) 

## *Evolution*

- Provide our public experiment users with all the necessary tools **infrastructure to store, browse and process** their data in collaboration with other European photon & neutron (PaN) sources.
- This will **extend our role beyond custody** without altering the **ownership** and restricted access of the data during the embargo (future data policy).

# 2017 Data policy for public users



To be updated in the near future

<b>Data &amp; metadata accessibility</b>	Kept online for <b>1 year</b> , then on tape (total: min. 5yrs). Read only. <b>On-line catalogue</b> of metadata. ALBA compromises to <b>best effort</b> to capture as complete as possible metadata.
<b>Privacy</b>	Restricted to the team for <b>3yrs</b> , then <b>opened to the public</b> . Access rights transferable. High level metadata public.
<b>Ownership</b>	Determined by contractual obligation of the person performing the analysis.
<b>Curation / analysis</b>	Optional but encouraged on site. <b>Results can be stored on-site</b> .
<b>Loss of data</b>	ALBA non liable.
<b>Publication</b>	Obligation to cite the experiment and data <b>PIDs</b> and to deposit the reference in ALBA's publication database.
<b>Commercial users</b>	Confidential data, owned by the client, not stored unless agreed otherwise



# Happening now:

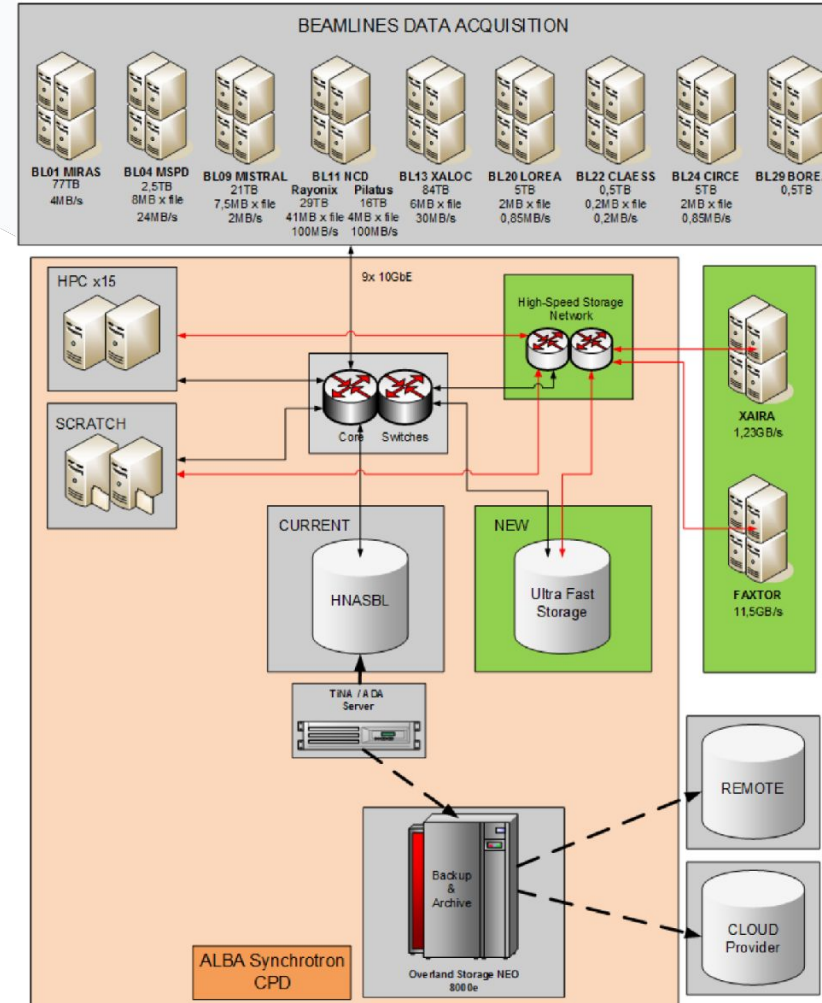
## *Increase of our storage capacities*



- XAIRA, FAXTOR and XALOC will be working with kHz data rates.
- MX Typical size 8h shift: ~**4-25 GB**, (SX: ~**30-50TB**)
- An **ultra-fast storage** system is planned to be installed and shared with the FAXTOR tomography beamline (burst-buffer 24h + storage) , ideally around 2PB. °

Beamline	Data Generation (PB/year)	DataSet Size (MB)	W Performance Base (GB/s)
BL XAIRA	1.26 PB/year	200 MB x 8 files	1,23
BL FAXTOR	2 PB/year	4 MB x 4000 files	11,5

Tabla 1: Previsión de la generación de datos de los detectores para XAIRA y FAXTOR



# Scientific Data Management (Feb. 2021)

*A new section inside the Computing division*



## Computing & Controls



Oscar Matilla  
Computing Division Head



Concepción Girbau  
Secretary

## Controls and DAQ



Guifré Cuni  
Controls Section Head

## Electronics



José Ávila  
Electronics Section Head

## IT Systems



Toni Pérez  
IT Systems Section Head

## MIS



Daniel Salvat  
MIS Section Head

## SDM



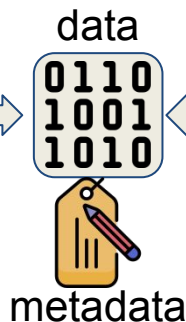
Nicolas Soler  
SDM Section Head

New



# Scientific Data Management (Feb. 2021)

*A new section inside the Computing division*

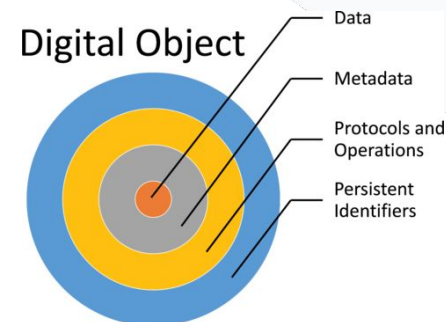


Data processing

Programming  
Optimization Java HPC  
Data C/C++ Analysis  
Pipelines Processing Visualization  
Scientific Computing  
Machine Learning Web scrapping  
Matlab Python Libraries  
API data reduction

Data reusability

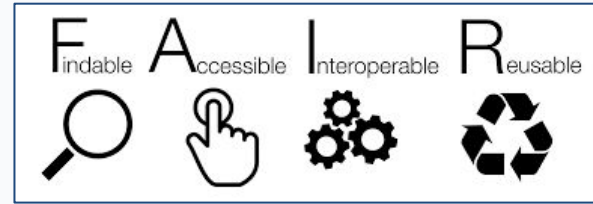
- Metadata ingestion
- Provenance
- Persistent identifiers
- Catalogue
- DAaaS



**4 people** (1 more in the beginning of 2022) to be hired during the summer with hybrid science / computing profiles

(will work on **all beamlines**, occasional support to new 4th generation synchrotron source)

# The future is open science



[Wilkinson, M., Dumontier, M., Aalbersberg, J. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 \(2016\).](#)



## Mission:

- Make data easily **reusable** for the scientific community
- Allow users to move and **remotely process** their data in between European PaN facilities

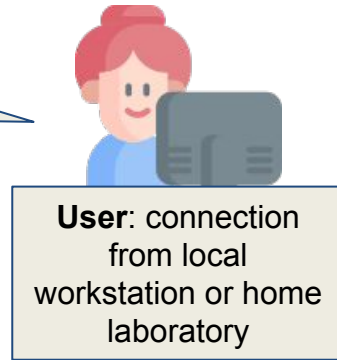
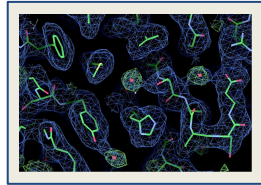
**WP2:** data policy and stewardship

**WP3:** data catalogues

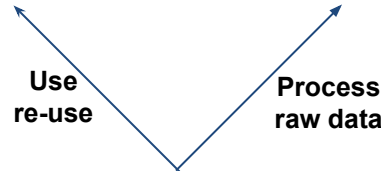
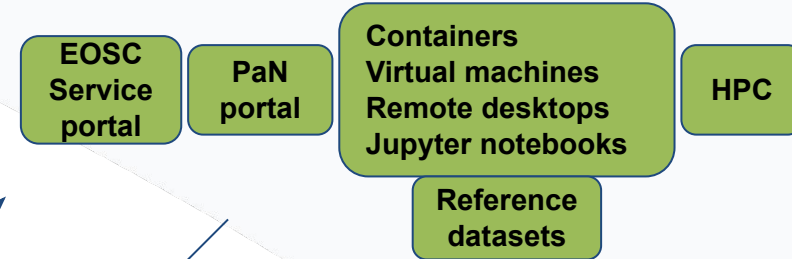
**WP4:** data analysis as a service



# European projects



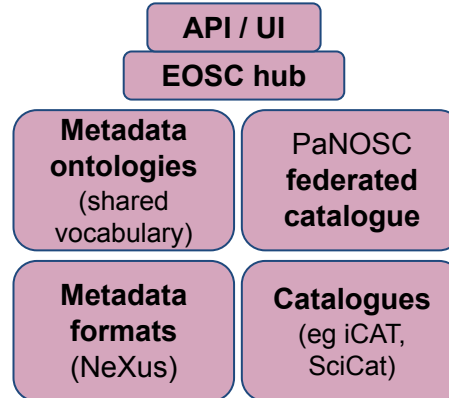
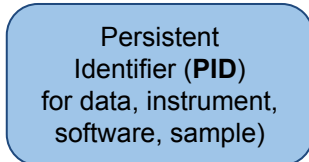
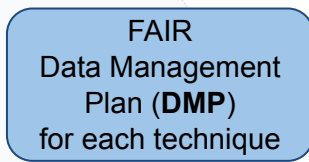
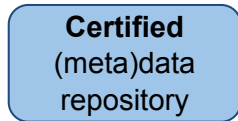
## Data analysis as a service (WP4)



Annotate & store processed data + provenance

## FAIR-ready data (WP2)

## Data catalogues (WP3)

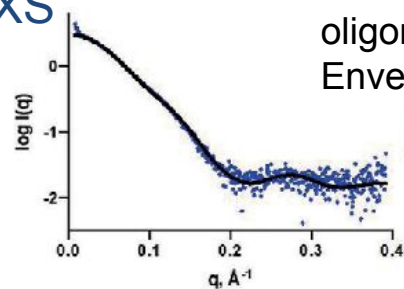


## Data reduction / compression

- Metrics
- New algorithms
- Software vs hardware
- Technique-specific
- Lossy vs non lossy
- Meta-compressors (ML)

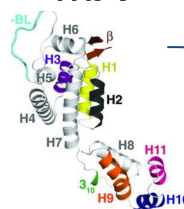
# Towards multimodal experiments

bioSAXS



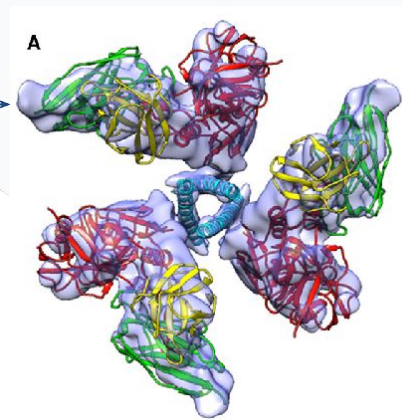
Solution  
oligomerization state,  
Envelope

MX

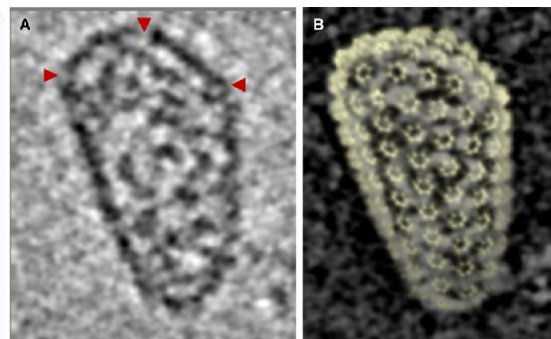


High. res.  
Individual subunits

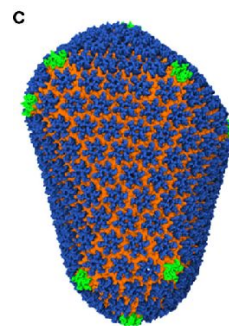
CryoEM (SPA)



MR models



CryoET/ X-ray  
microscope  
(MISTRAL)



Molecular  
Modelling  
(MD)

Zhao G. et al, Nature vol497, 2013

Bio tomography  
FAXTOR  
(phenotype)

# Summary: 4th gen.

- Faster acquisition time and bigger volumes of data generated will require more computational tools for data **sorting**, **processing**, **reduction** and **visualization**.
- Tools will also be developed to integrate data from different sources (**multimodal experiments**).
- **Data reusability** will be ensured by proper metadata ingestion at the beamline (as automated as possible) served by a federated data catalogue.
- Users will be able to keep and process their data **on-site** via a software portal.



Contact : [\*nsoler@cells.es\*](mailto:nsoler@cells.es)



# Thank you!

