Big Data Orchestration in the Australian Characterisation Commons at Scale

Prof Wojtek James Goscinski

This project is supported by the Australian Research Data Commons (ARDC) and the following partners. The ARDC is enabled by NCRIS

The Australian Characterisation Commons at Scale

Challenges

Characterisation refers to the general process of probing and measuring the structures and properties of materials at the micro, nano and atomic scales. It is essential across natural, agricultural, physical, life and biomedical sciences and engineering. **Includes capabilities across NCRIS: Microscopy Australia, National Imaging Facility, ANSTO, and Institutions.**

Challenge:	Solution:
Scale and complexity "the science is being affected by compute"	**A national infrastructure program** of accessible tools, seamless access and integrated instruments
Working with digital objects is challenging	**Make characterisation digital objects FAIR** Requires community, coordination and commitment
Expertise is rare Digital expertise coupled with applied characterisation knowledge is rare	**A national program to spread knowledge and underpin change** National training and a national network

Australian Characterisation Commons at Scale

Program of Work

Australian Research Data Commons

It has co-investment from 10 Universities, three NCRIS facilities, alignment with two Medical Research Future Fund initiatives, two ARC CoEs and flagship proposals, and numerous flagship instruments.

Develop a coherent and accessible informatics landscape that promotes collaboration, increases ROI, and delivers value to researchers.

Deploy a **Characterisation Commons** for thousands of researchers who use characterisation techniques, facility scientists who run instruments, and researchers using imaging collections, and will uplift the research capability offered to industry.

The outcome will be a rich ecosystem of computing systems, data repositories, workflows, and services, connected with instruments.

3 specialised programs:

- Big Data Electron and Correlative Microscopy from Instrument to Publication
- Biomedical Imaging Collections and Analysis
- National Tools for Scattering and Beyond

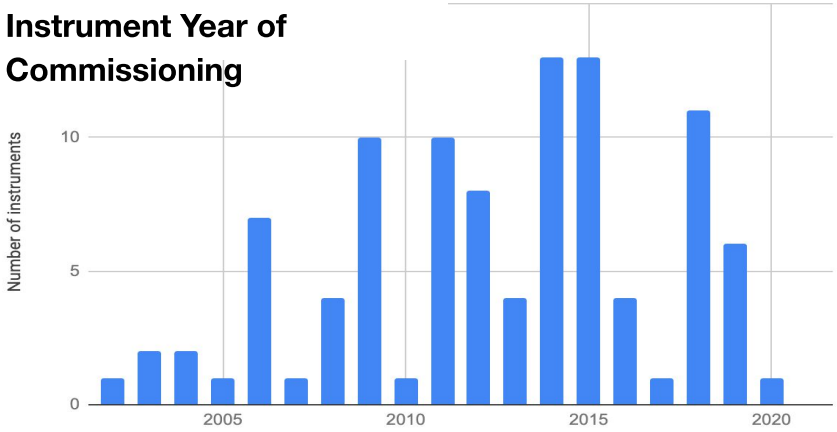
Characterisation Instrument Survey Overview (8th Oct 2019)

Responding
111 Instruments
17 Facilities
9 Universities and Institutes

Identified
399 Instruments
29 Facilities
9 Universities and Institutes

Estimated number of projects per year	Total research projects	**2505**
	Average no. per instrument	**22.6**
Estimated number of users per year	Total users	**3007**
	Average no. per instrument	**27.1**

Instrument Year of Commissioning



Classified into **22 characterisation modalities**

25 petabytes of data per year produced by responding instruments

Eight modalities produce over 1TB per week
However, the amount of cumulative data produced by these instruments is 95% of the total.

The most significant data producing instruments are:
Transmission Electron Microscopy, **Cryo Electron Microscopy**, **Lightsheet**, **Hybrid imaging -human**

42% of instruments reported feed data into a data management system which is capable of providing global data identifiers.

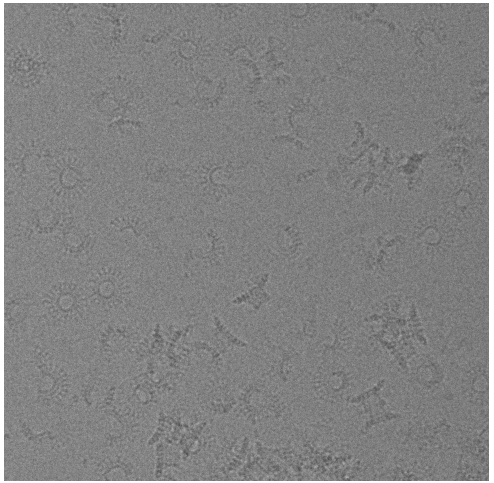
On average users required **between ½ year to 1 year to process their data**

Data Formats

TIF	45	DM3	4	LOG	1
DICOM	25	ND2	3	MMF	1
JPG	11	TXT	3	NII	1
LIF	9	HDR	2	SER	1
BMP	8	BIMG	1	TXRM	1
MRC	5	IMG	1	XLS	1
RAW	5	LM	1	XML (imzML)	1

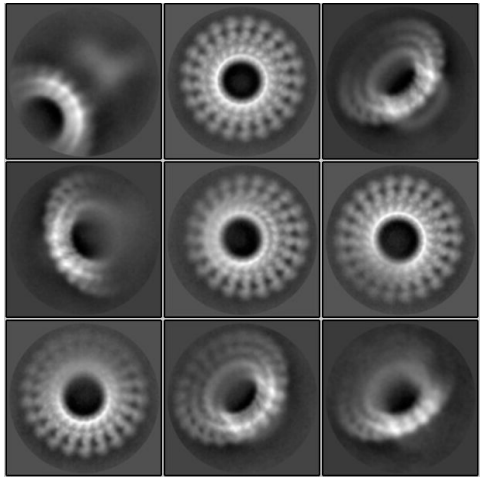
Cryo Transmission Electron Microscopy
Single-particle analysis

Capture



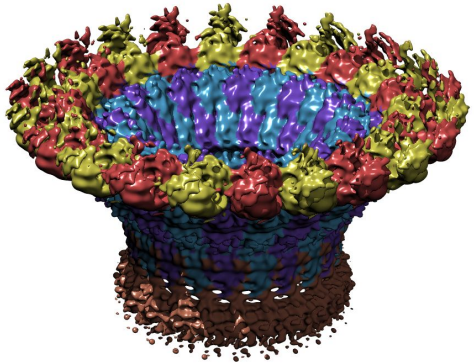
1 Gigabyte - 20 Terabyte

Processing

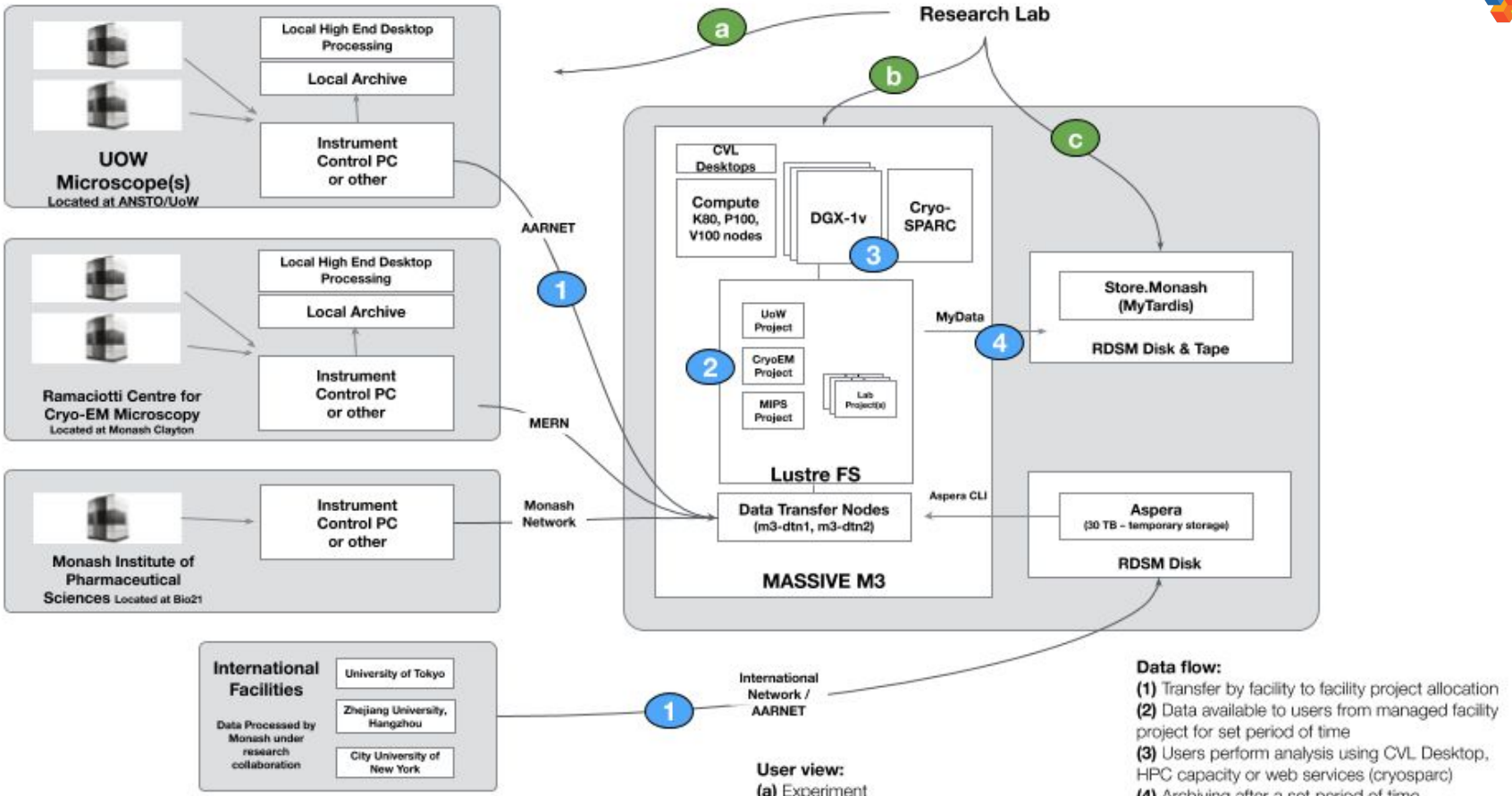


Hours to Weeks across various configurations of HPC nodes / GPUs

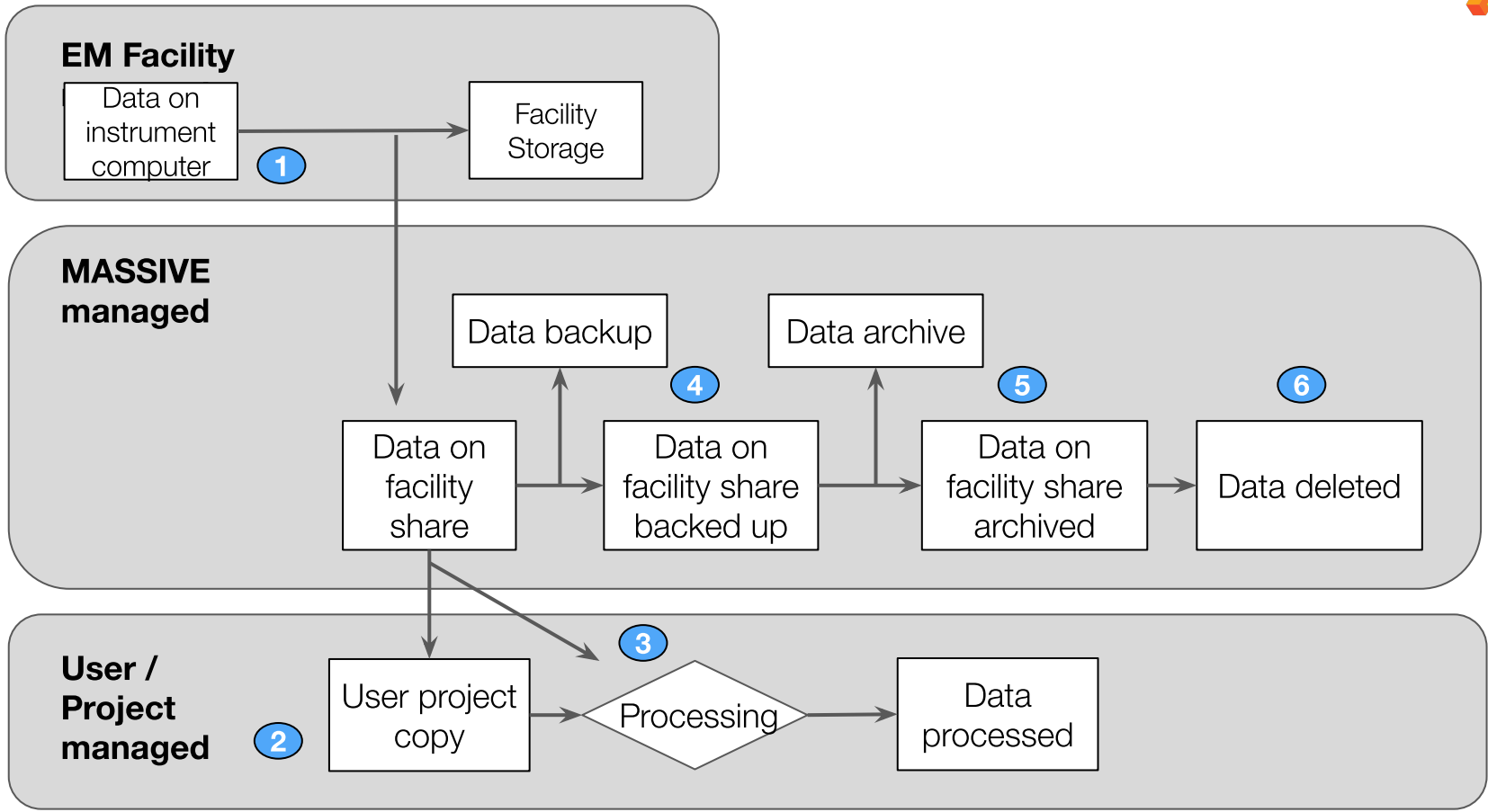
Structural biology …



CryoEM workflow at MASSIVE



CryoEM data flow state diagram



Big Data Orchestration in the Australian Characterisation Commons at Scale

Prof Wojtek James Goscinski

This project is supported by the Australian Research Data Commons (ARDC) and the following partners. The ARDC is enabled by NCRIS