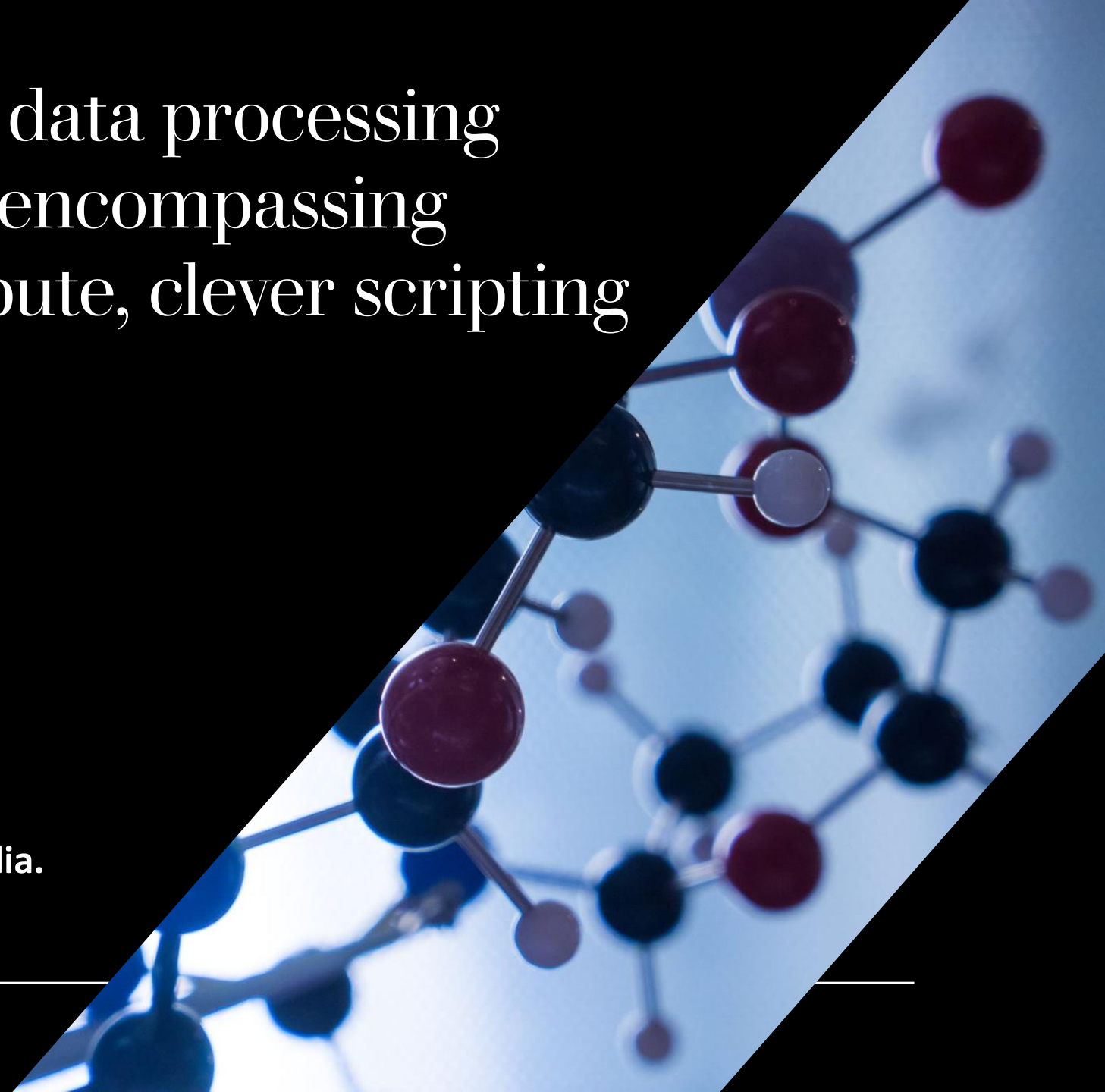
A seamless CryoEM data processing pipeline utilizing an encompassing network, supercompute, clever scripting and virtual desktops



Dr. Farrah Blades

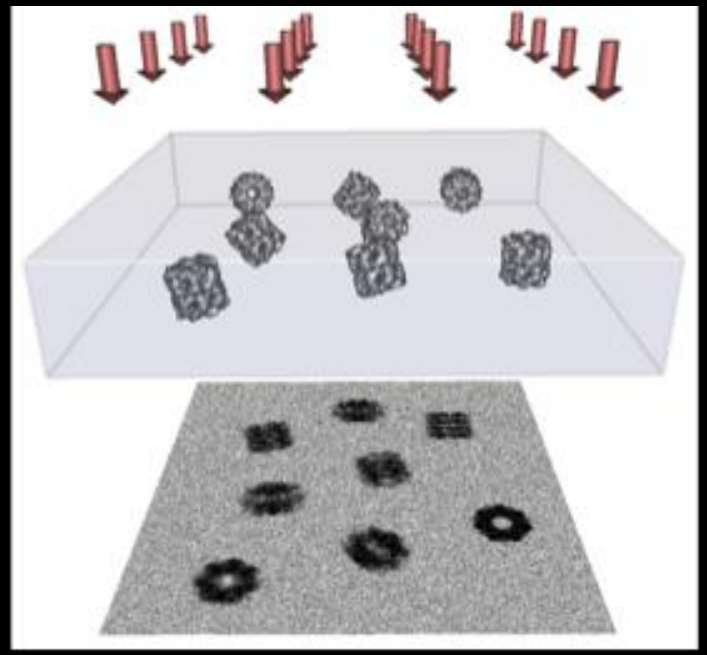
Cater group

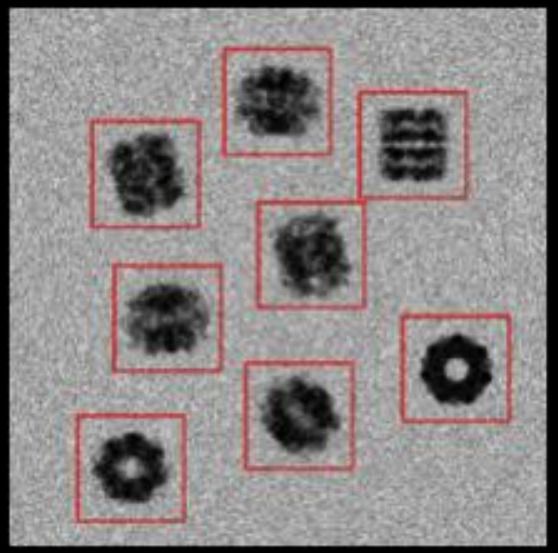
Institute for Molecular Biosciences,

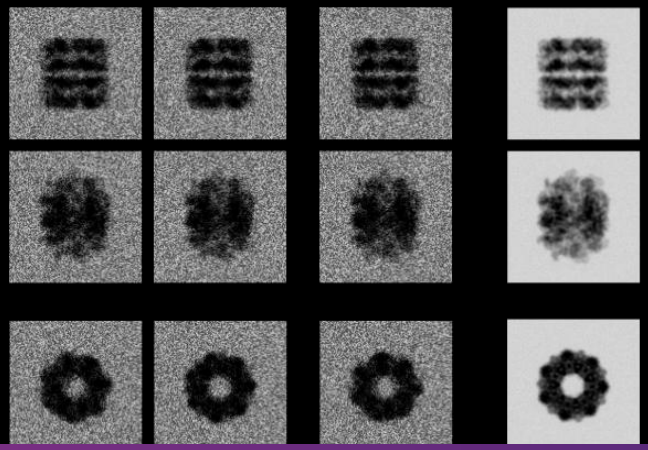
The University of Queensland, Australia.

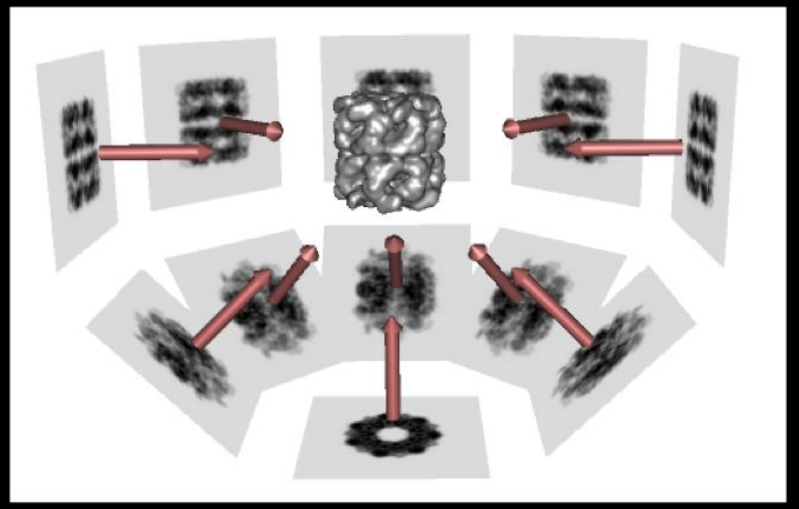
f.blades@imb.uq.edu.au

Single particle cryogenic electron microscopy (Cryo-EM)





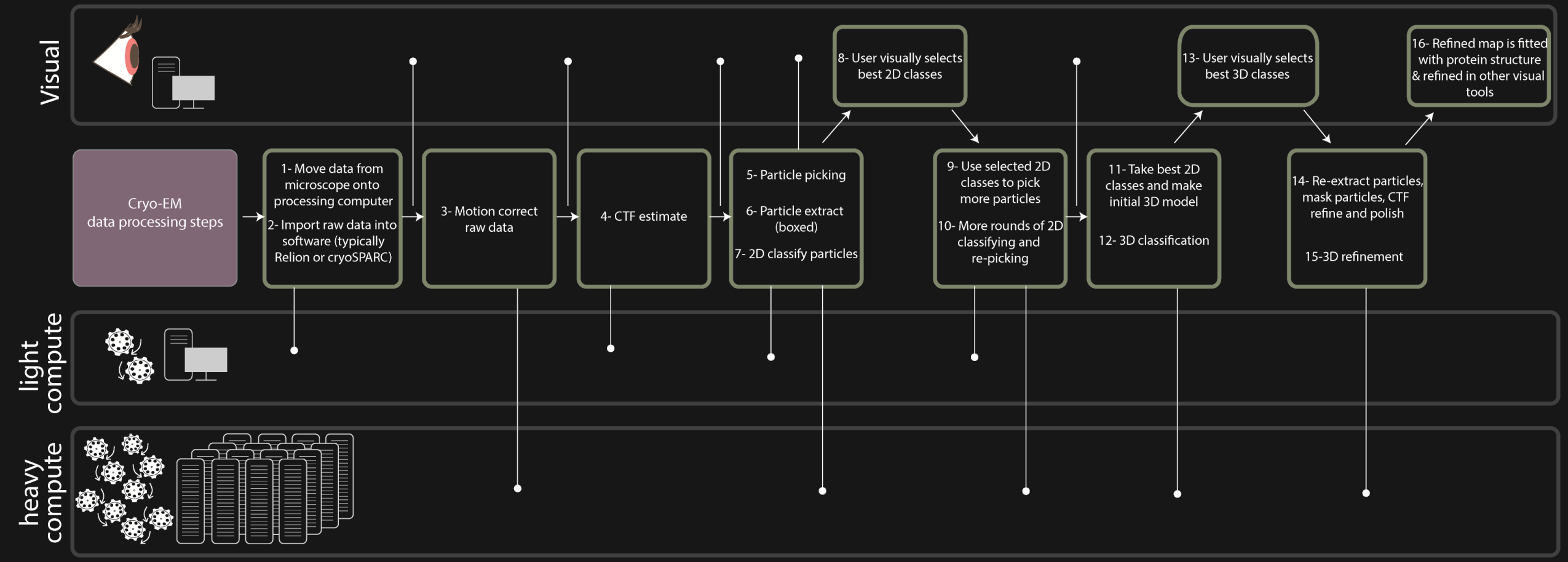




Thousands of micrographs

Millions of particles

Typical Cryo-EM data processing workflow



How do we process these huge data sets?

cryoSPARC

RELION

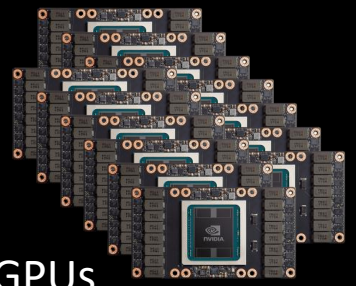
Command line skills

Writing scripts

Logging into HPC by command line

Submitting jobs by command line

Wiener GPU HPC

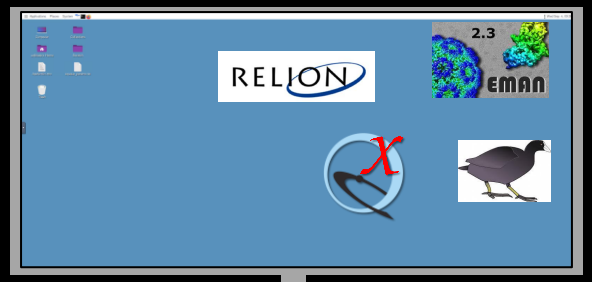


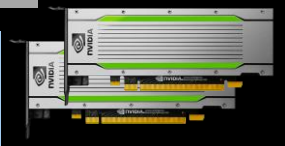
Nvidia V100 GPUs

GPU clusters
1) Set up by an institute or lab group
2) Pay for service

First attempts to solve the issue at UQ-Characterisation Virtual Laboratory (CVL)

CVL





24x Nvidia
T4 GPUs

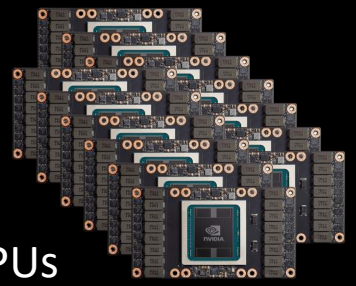
Command line skills

Writing scripts

Logging into HPC by command line

Submitting jobs by command line

Wiener GPU HPC



Nvidia V100 GPUs

GPU clusters
1) Set up by an institute or lab group
2) Pay for service

Easiest solution

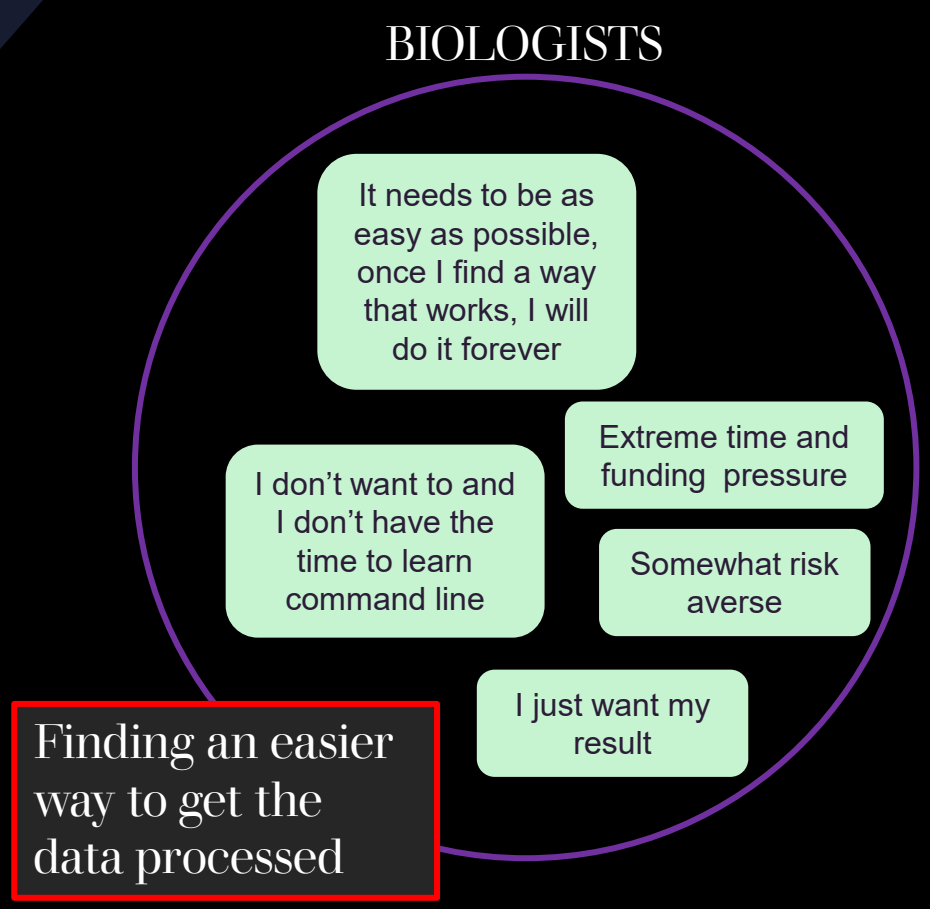
- Expiration countdown begins as soon as its built
- Maintenance
- Cannot store or process our data set sizes
- New techniques pushing capacity even further
- Moving data on and off
- Arduous script writing skills to run big jobs still required

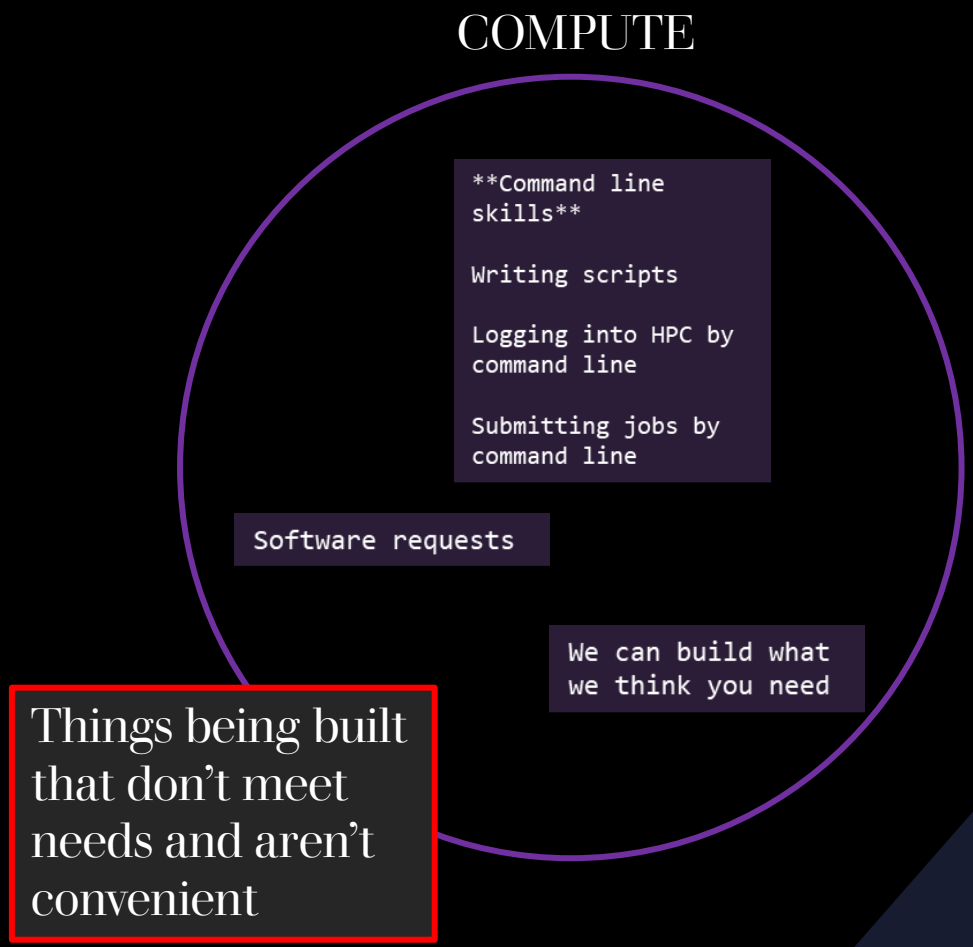
Data movement and processing

=

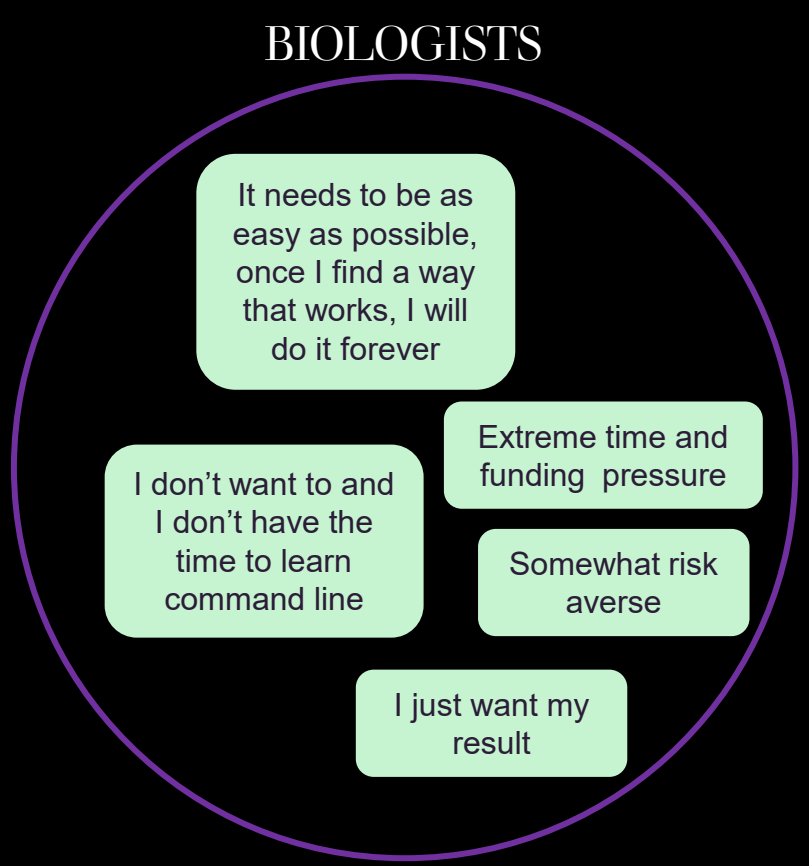
Can be one of the largest bottlenecks for CryoEM groups

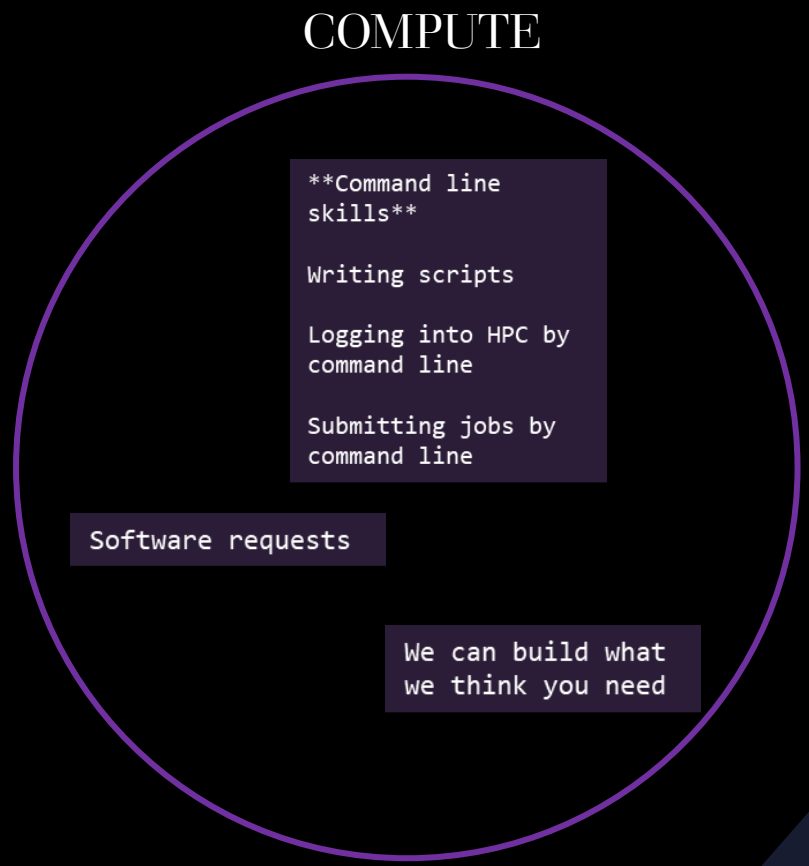
How did we get here?



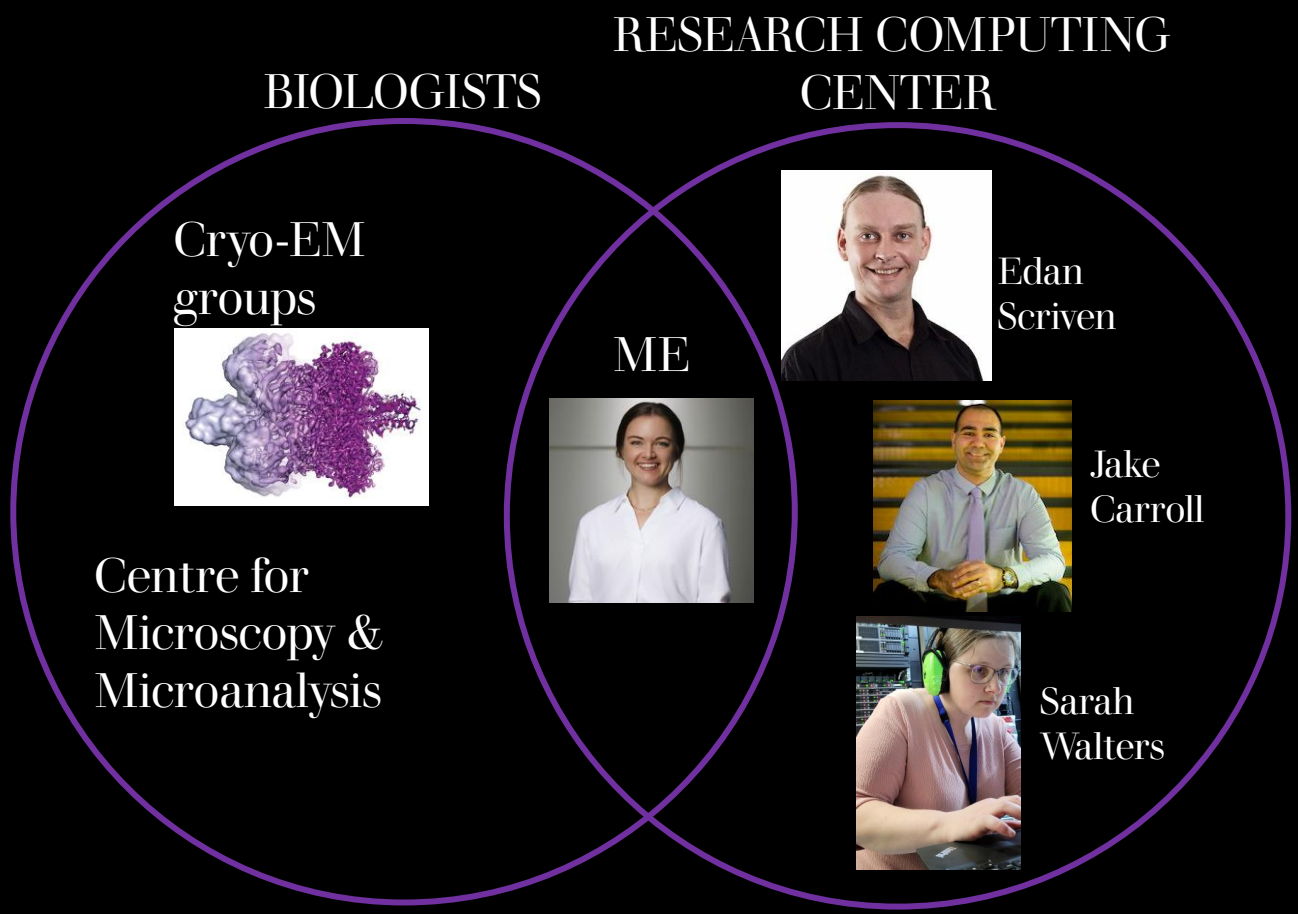


Fixing a university (and nation?) wide issue





Right people, right place, right time, right funding



HPC Bunya: UQ's new supercomputer, grass roots build

Phase 1 & 2

- RCC took many lessons from Wiener HPC, CVL and the feedback from the CryoEM community.
- ~10,000 cores of AMD Milan and Genoa CPUs.
- 2TB of DDR5 memory per node.
- 200Gbit/sec RDMA networking.
- SLURM entry but also cloud desktop.
- GPUs: Nvidia H100s, A100s, L40s and AMD's MI210s.
    - **21x H100s 80GB GPU RAM**
    - **8x A100s 80GB " "**
    - **3x A100 MIG 10GB (21 slices)**
    - **18x L40s 48GB " "**
    - **6x MI210 64GB " "**

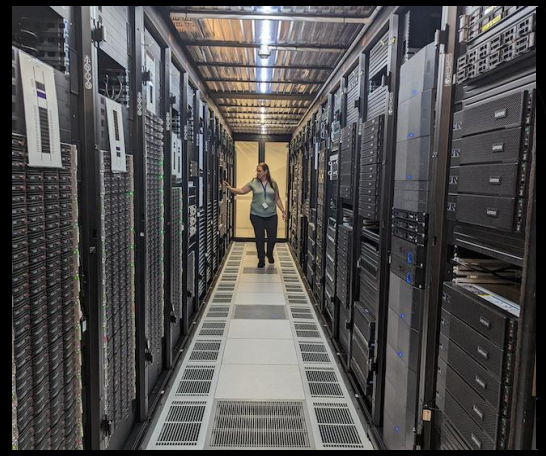
Phase 3

- RCC took more lessons from trench 1 & 2 and the viz community
- ~2,000 additional cores of AMD Milan and Genoa CPUs
    - **16x H100 SXMs 320GB GPU RAM**
    - **9x A16s 16GB " "**

Each can run 4x desktops = <u>32 desktops</u>

- **6x L40s 24GB " "**
- **AMD MI300 server access**

- **2x MI300x boxes with 8xGPUs each**

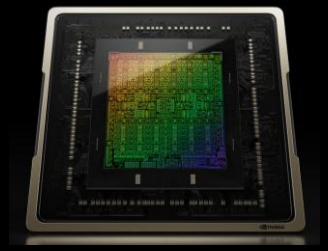


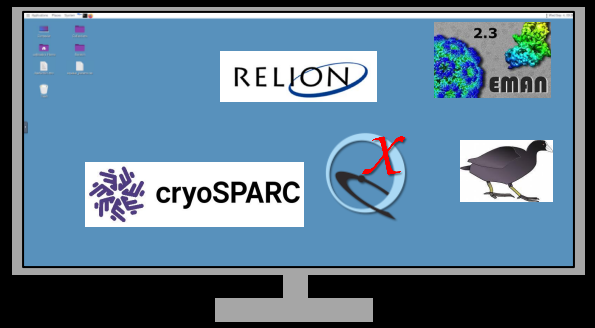

HPC Bunya: UQ's new GPU and CPU supercomputer

Virtual Desktop = OnBunya

OPEN OnDemand

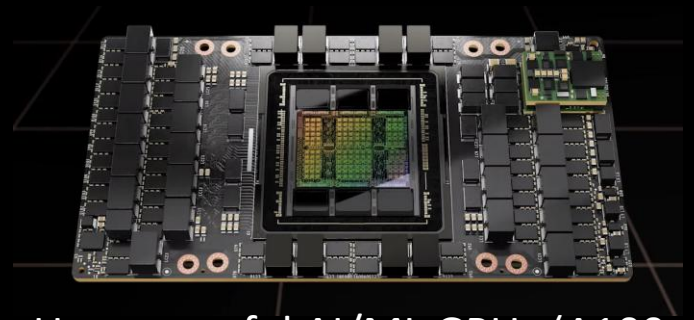
OnDemand provides an integrated, single access point for all of your HPC resources.



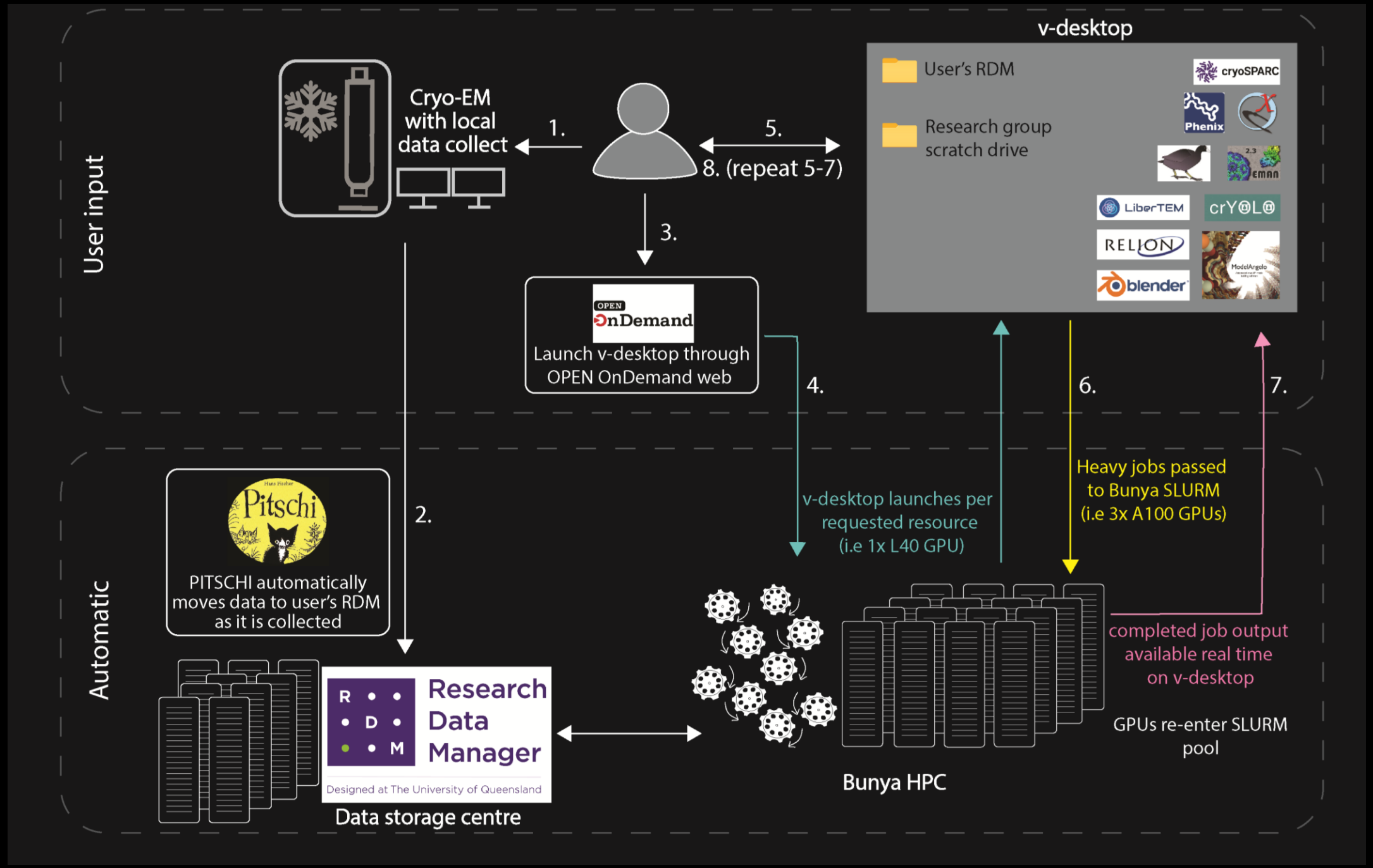


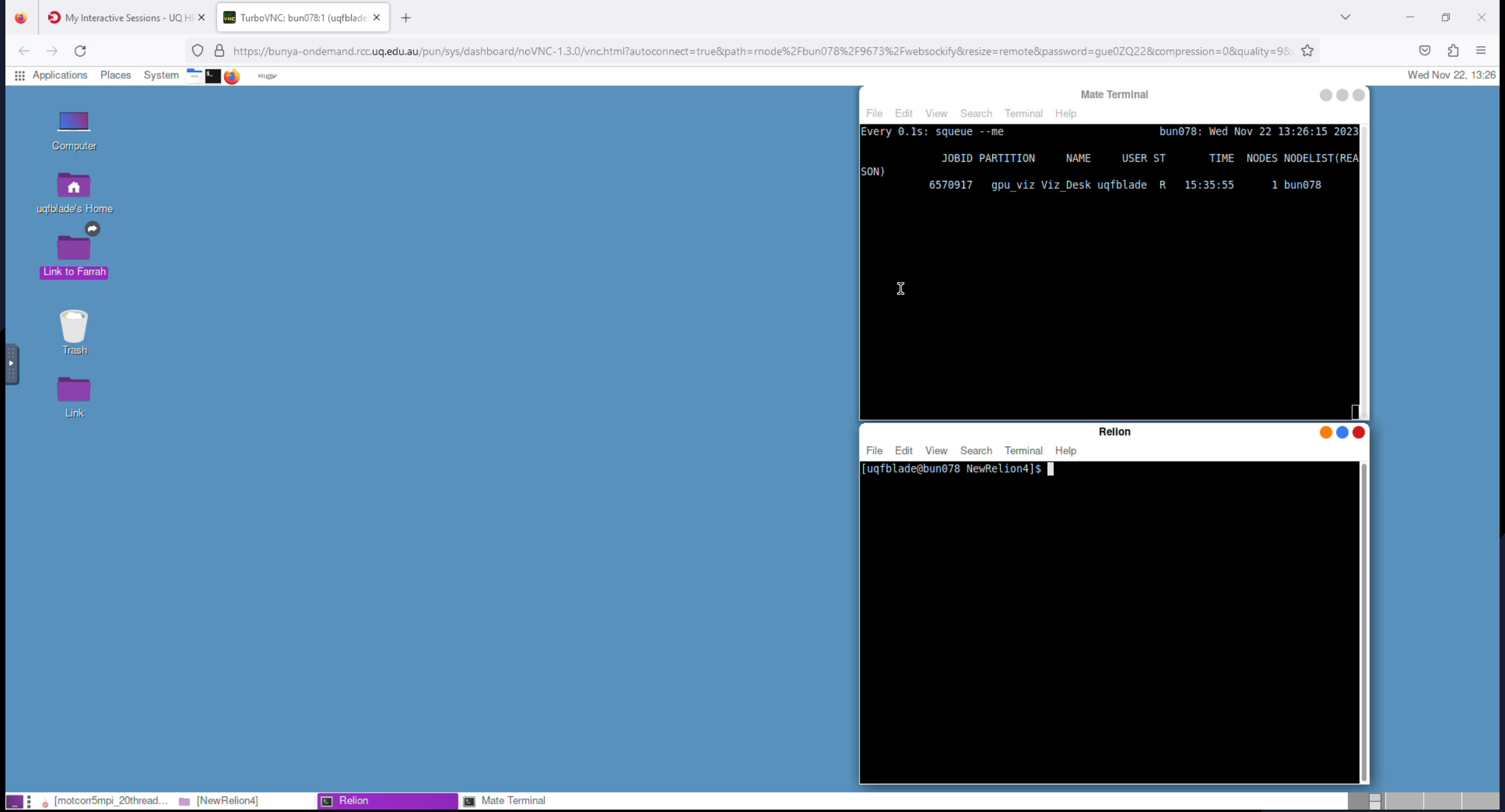
Use powerful visual L40 GPUs for the desktop workloads

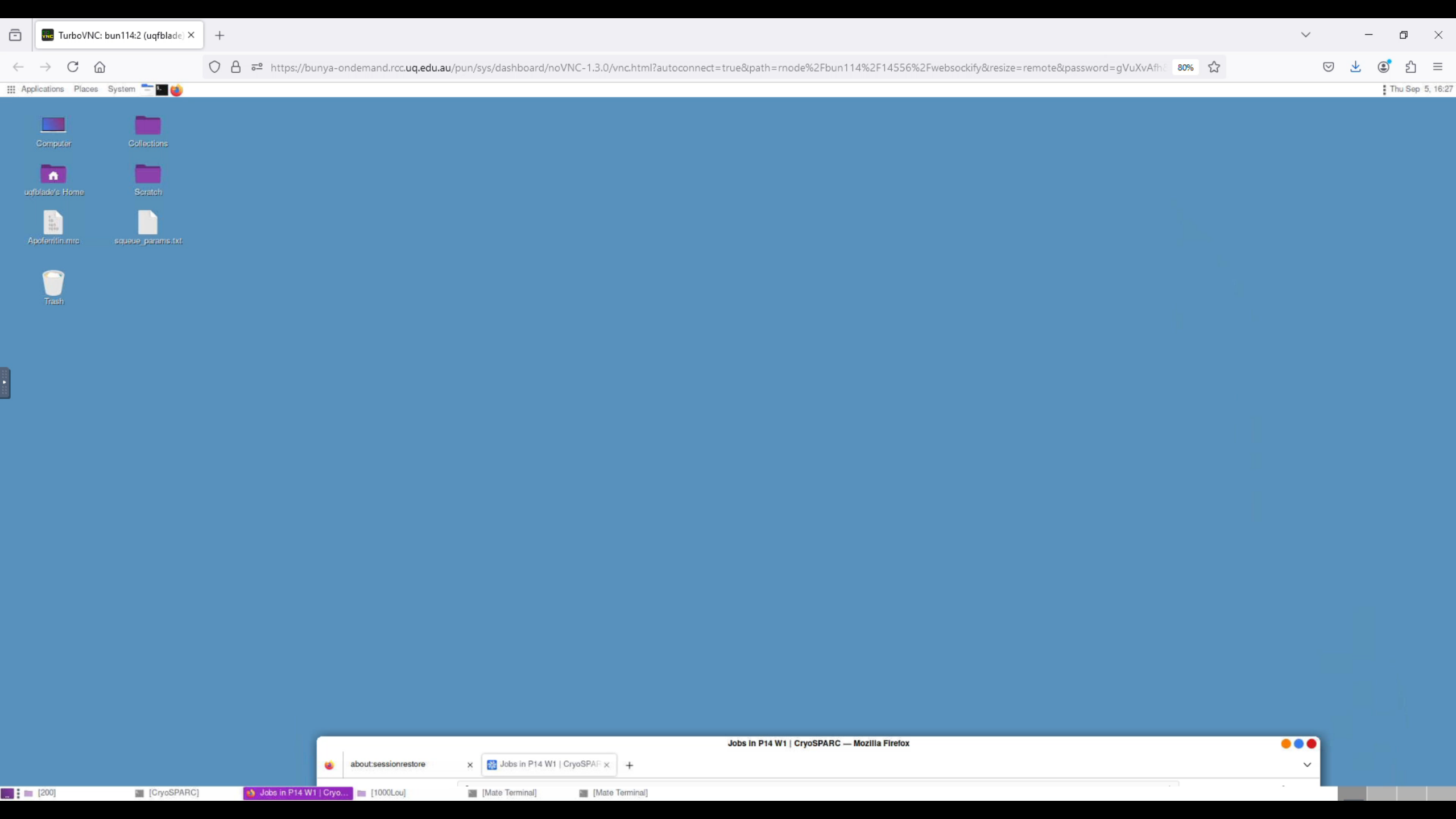
Working nodes



Use powerful AI/ML GPUs (A100, H100, MI210) or L40 for compute jobs


User input
Automatic
Cryo-EM with local data collect
1.
5.
8. (repeat 5-7)
3.
v-desktop
User's RDM
Research group scratch drive
cryoSPARC
Phenix
EMAN
LiberTEM
crYOLO
RELION
blender
OPEN OnDemand
Launch v-desktop through OPEN OnDemand web
4.
6.
7.
2.
Pitschi
PITSCHI automatically moves data to user's RDM as it is collected
v-desktop launches per requested resource (i.e 1x L40 GPU)
Heavy jobs passed to Bunya SLURM (i.e 3x A100 GPUs)
completed job output available real time on v-desktop
GPUs re-enter SLURM pool
Research Data Manager
Designed at The University of Queensland
Data storage centre
Bunya HPC


My Interactive Sessions - UQ HI
TurboVNC: bun078:1 (uqfblade
https://bunya-ondemand.rcc.uq.edu.au/pun/sys/dashboard/noVNC-1.3.0/vnc.html?autoconnect=true&path=rnode%2Fbun078%2F9673%2Fwebsockify&resize=remote&password=gue0ZQ22&compression=0&quality=9
Applications
Places
System
Wed Nov 22, 13:26
Computer
uqfblade's Home
Link to Farrah
Trash
Link
Mate Terminal
File Edit View Search Terminal Help
Every 0.1s: squeue --me bun078: Wed Nov 22 13:26:15 2023
JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON)
6570917 gpu_viz Viz_Desk uqfblade R 15:35:55 1 bun078
Relion
File Edit View Search Terminal Help
[uqfblade@bun078 NewRelion4]$
[motcorr5mpi_20thread...
[NewRelion4]
Relion
Mate Terminal


TurboVNC: bun114:2 (uqfblade
https://bunya-ondemand.rcc.uq.edu.au/pun/sys/dashboard/noVNC-1.3.0/vnc.html?autoconnect=true&path=rnode%2Fbun114%2F14556%2Fwebsockify&resize=remote&password=gVuXvAfh
80%
Applications
Places
System
Thu Sep 5, 16:27
Computer
Collections
uqfblade's Home
Scratch
Apoferritin.mrc
squeue_params.txt
Trash
Jobs in P14 W1 | CryoSPARC — Mozilla Firefox
about:sessionrestore
Jobs in P14 W1 | CryoSPAR
[200]
[CryoSPARC]
Jobs in P14 W1 | Cryo...
[1000Lou]
[Mate Terminal]
[Mate Terminal]

cryoSPARC

RELION

2.3 EMAN

crYOLO

RFdiffusion: A generative model for protein design

ModelAngelo

Automated cryo-EM model building software

Alphafold

LiberTEM

blender

Mate Terminal

```
Every 0.1s: squeue --me                          bun078: Wed Nov 22 13:26:15 2023

             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           6570917   gpu_viz Viz_Desk uqfblade  R   15:35:55      1 bun078
```

Relion

```
[uqfblade@bun078 NewRelion4]$
```

Users (so far)

Beta-testing Dec 2023

Official launch May 2024

Users of OnBunya = 595

~43% of Bunya HPC users are using OnBunya

“OnBunya has proven to be extremely useful, helpful, and convenient for our need”

Testimonies

"Without OnBunya more than half of our people wouldn’t have [means] to do their processing at all and would have to wait”

“for the future, OnBunya should be the main place people do their processing, it is a no-brainer”

“It will help with publications and future grant applications”

“I extensively use the OnBunya desktop”

Why is this platform so successful?

User designed

User tested

Trust across disciplines

Grass roots HPC build

Workshops to increase user confidence

Data access

Storage and scratch space

Everything user needs in one place

It is at least 5x better than the next best alternative

Where to from here?

- We have found a scalable solution to moving large data sets and processing them in house and we hope to help other universities do the same

- A national service as smooth as OnBunya…

Acknowledgements

My email: f.blades@imb.uq.edu.au
www.linkedin.com/in/farrah-blades

HANKAMER GROUP
Prof Ben Hankamer

Research Computing Centre

Centre for Microscopy & Microanalysis UQ
Dr Matthias Floetenmeyer
Dr Lou Brillault
Dr Na'ama Koifman
Professor Roger Wepf


Edan Scriven


Jake Carroll


Sarah Walters

Marlies Hankel
Irek Porebski

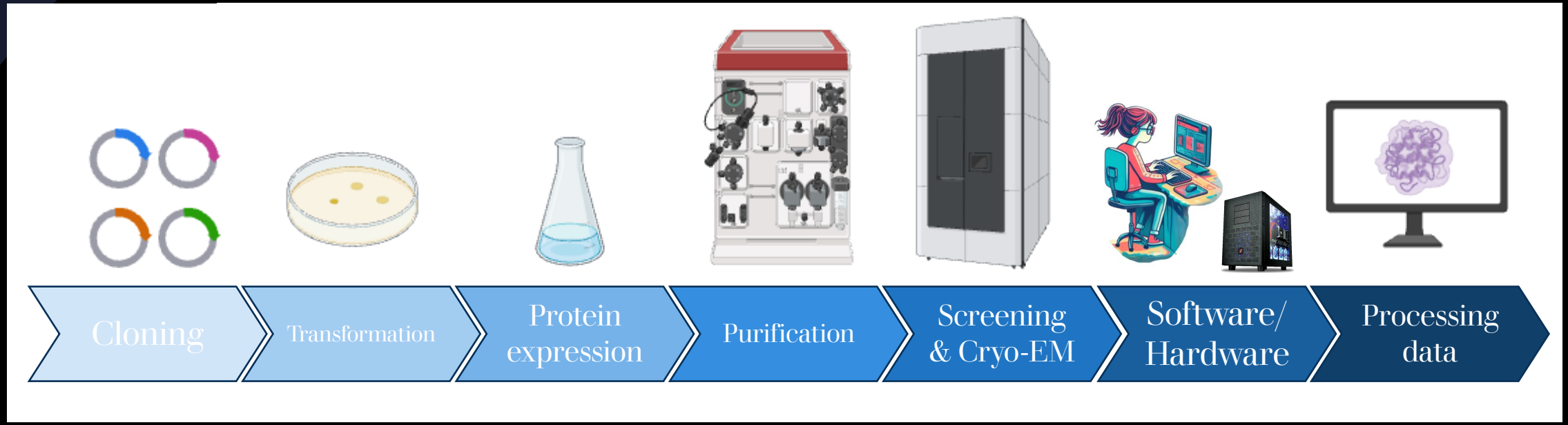
Code sharing

All open access: https://github.com/UQ-RCC/hpc-docs

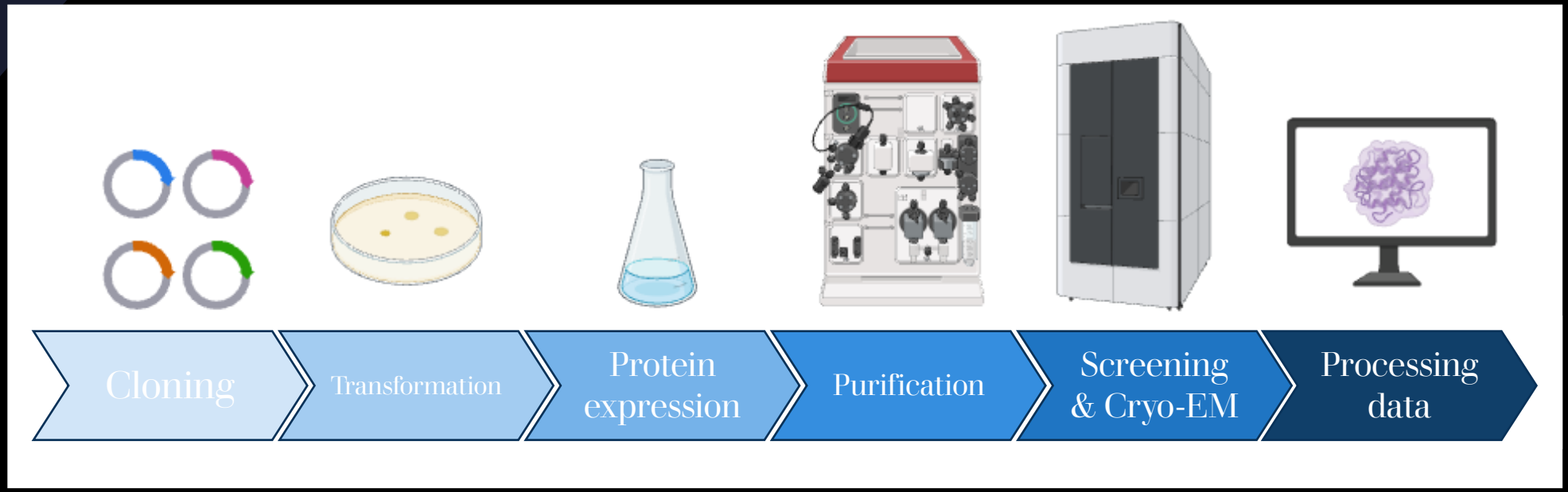
we hope to show others how to tackle this huge data/ big viz and big compute problem

24 hour Run-a-thon for Lifeline

Community examples- Meihan Liu (Collins Group)



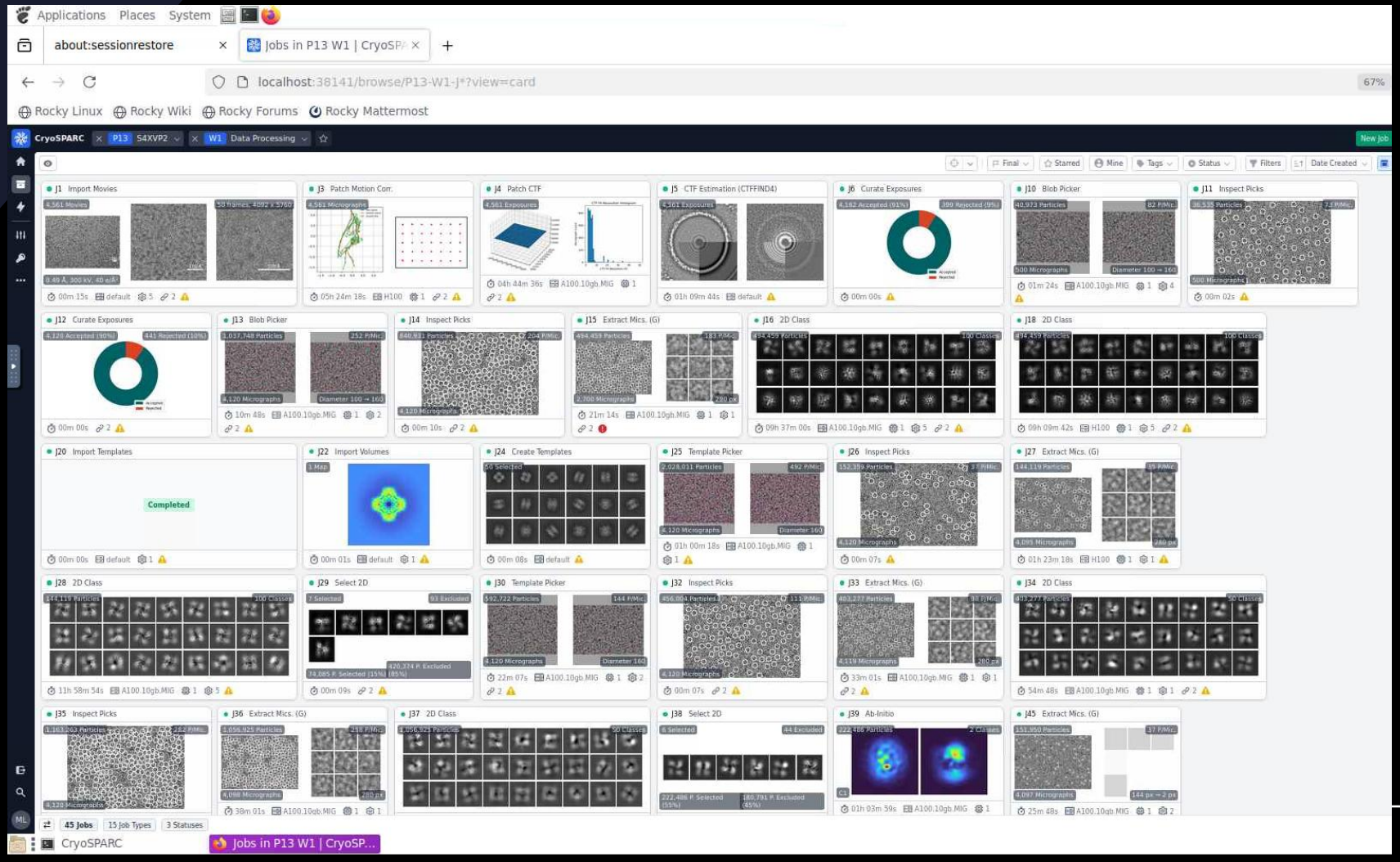
Community examples- Meihan Liu (Collins Group)



5 days if all goes well

3 days

Community examples- Meihan Liu (Collins Group)



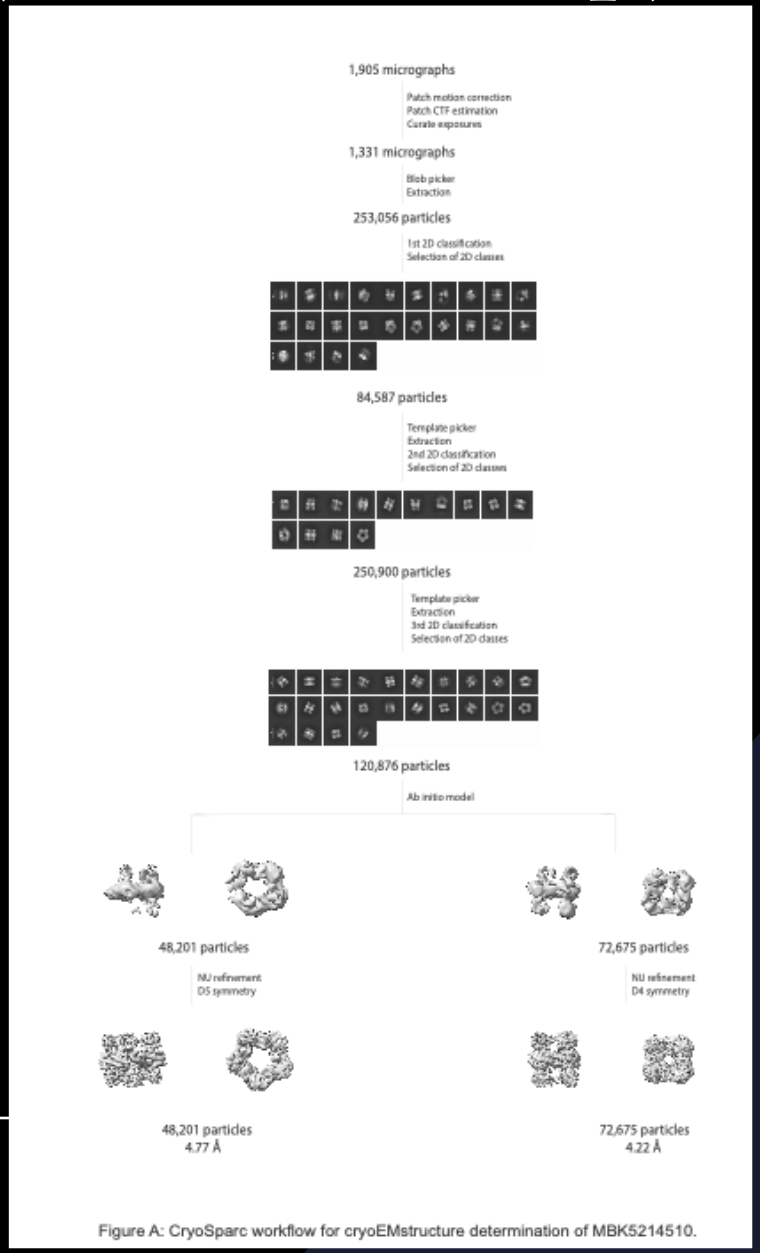


Figure A: CryoSparc workflow for cryoEMstructure determination of MBK5214510.