NRC·CNRC

NRC.CANADA.CA •

The Canadian Astronomy Data Center: an Independent Data Access Centre for the LSST

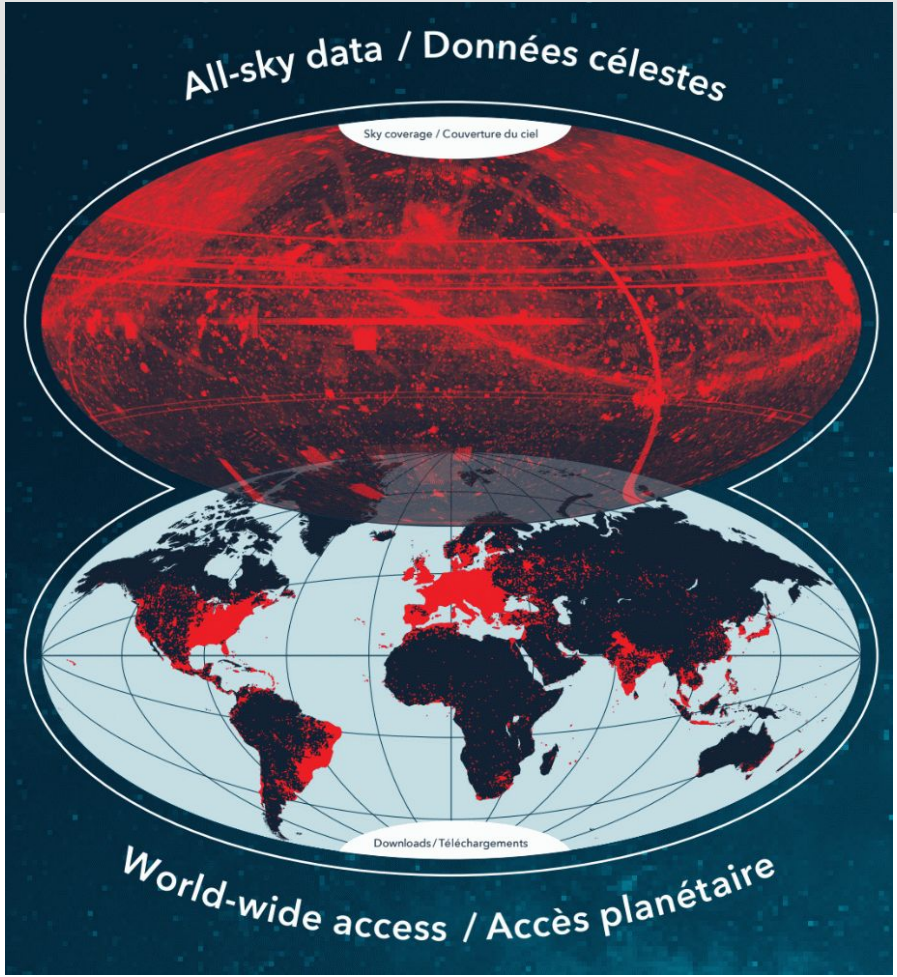
National Research Council Canada | Conseil national de recherches Canada

Stephen Gwyn Canadian Astronomy Data Centre

Canada

CADC:
Canadian Astronomy Data Centre

- Data distribution site for Canada's astronomical data
- We mirror some datasets (HST, JWST, Gemini, …)
- We are the sole repository for other datasets (CFHT, JWST, OMM, NEOSSat, …)
- 216 telescopes and instruments
- Current holdings:
  - 2.3 PB, 300 million files
  - Ceph ObjectStore back end
  - CADC layer on top: Storage Inventory
- Annual downloads:
  - 100 million files
  - 4.9 Petabytes
- Users are 15% Canadian, 85% international
- Upcoming datasets
  - LSST
  - Euclid
  - SKA



Canadian Astronomy Data Centre

An alternate view of the archive

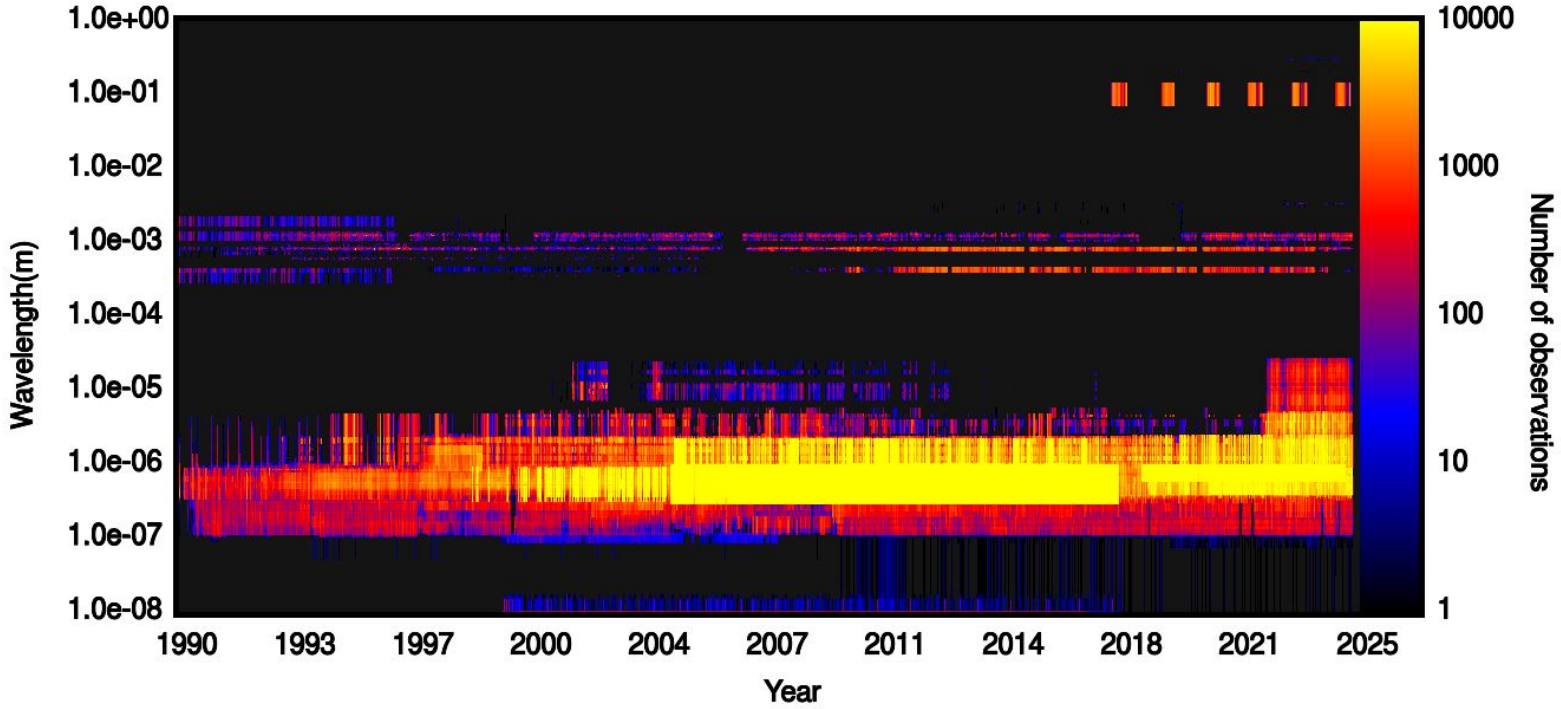
Radio

Microwave

Infrared

Optical

Ultraviolet



A brief history of the CADC

- Started in 1986 to host a copy of HST data. The idea was scientists would travel to Victoria to process data.
- Expanded its mandate to include other observatories
- Starting providing access to catalogs, both standard and user-provided
- Started processing data for users, providing science ready data products
- Evolved to 3 main activities:
  - Archives
  - Data expertise
  - Providing compute facilities

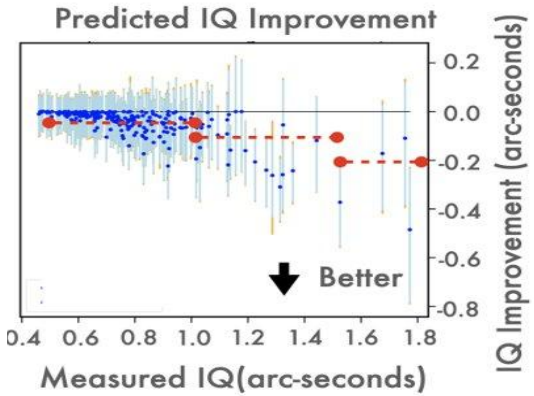
Canadian Astronomy Data Centre:
Processing data and providing machine learning expertise

The CADC processes raw data into science-ready data products

- Makes the data significantly easier for scientists to use
- Increases data re-use
- Two modes
  - Off-the-shelf / prêt-à-porter processing of entire archives
  - Bespoke / tailored processing for particular projects
- Started with HST data in the early 1990s
- Expanded to CFHT in the early 2000**s**

The CADC provides machine learning expertise to astronomers

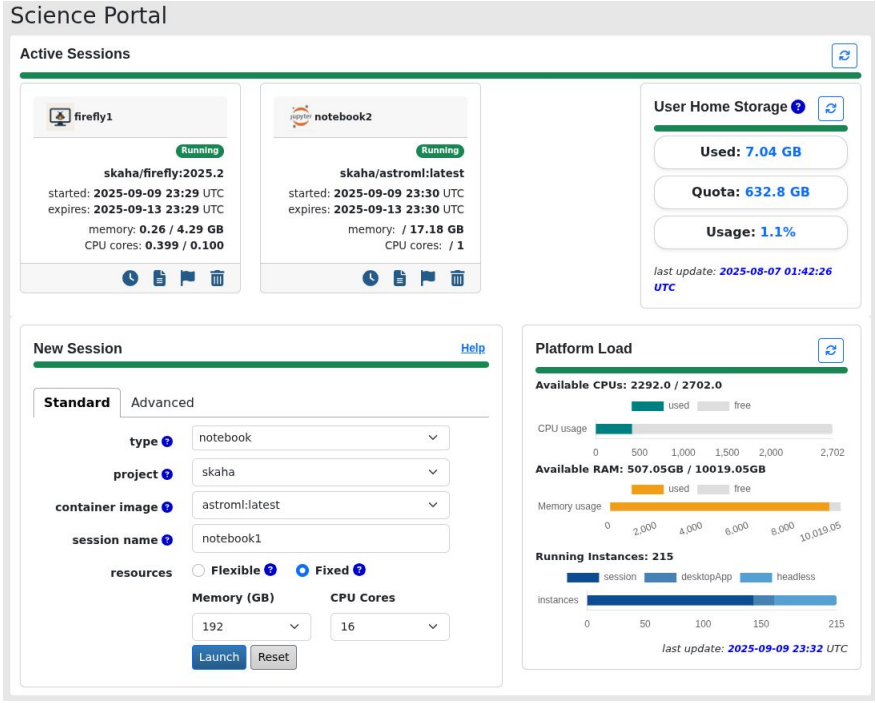
- Detection of post-merger galaxies in (CFHT)
- CFHT image quality forecasting
- Extragalactic star-formation studies (SDSS)
- Detection of gravitational lenses (HST)
- Image assessment and recommendation (CFHT)
- Photometric redshifts (HSC)
- Adaptive optics predictive controls
- Extragalactic metalicity studies
- Moving object detection (CFHT, HSC)
- Spectral analysis of nearby galaxies (MaNGA IFU)
- Delirium problem in COVID-19 patients



CANFAR:
Canadian Advanced Network For Astronomical Research

The CANFAR Science Platform:

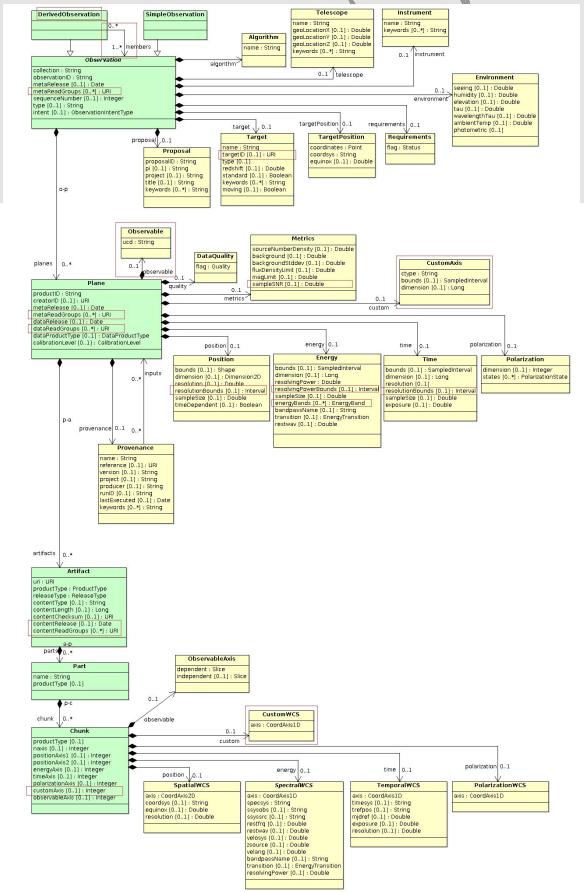
- Kubernetes-based
- Notebooks
- Browser-based VNC desktops
- Data visualization
- Scalable resources
  - up to 192 GB RAM / 16 cores per instance
- Multiple software container options (key!)
  - ALMA / JWST / LSST / …
  - user generated
  - ability to share containers
- Batch processing (now uses Kueue)
- Very successful: most Canadian astronomers are users



Astronomical metadata

Metadata indexing

- 219 telescopes and instruments / 1 metadata system
  - CAOM2: Common Astronomical Observation Model
  - Very comprehensive
- Reduces effort needed to manage existing archives
- Low overhead to ingest new archives
- Not all data findable at via CADC search resides at the CADC
- Allows "one stop shopping" for astronomical data
- Initial development lasted ~1 year
  - multiple stakeholders
- Significant changes will be required as we pivot towards radio astronomy
- Each new telescope/instrument requires a period of data engineering (DE)
- Initial model: Astronomers (domain experts) drive DE
- Current model: Developers (model experts) drive DE



Unified Modeling Language diagram of CAOM2

CANFAR:
User storage

Cavern / arc:

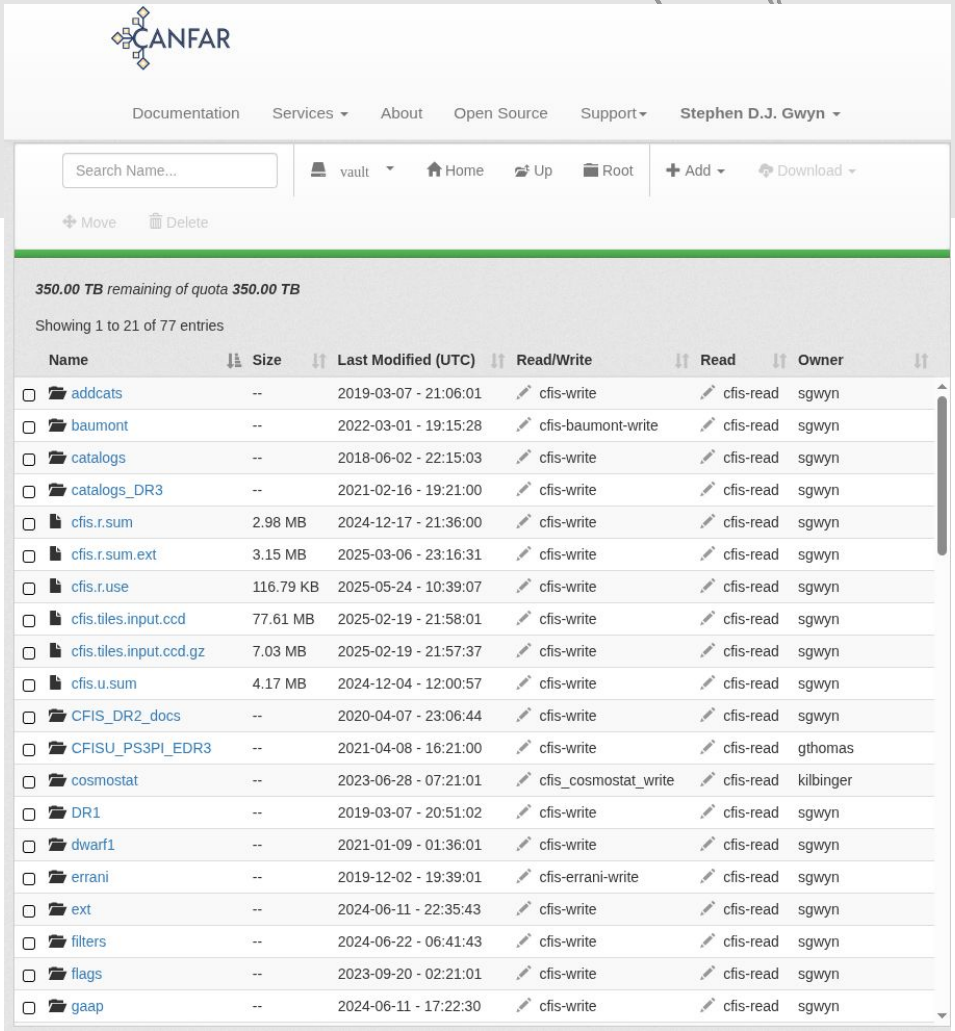
- Suitable for processing
- Each user gets a small home directory
- Project directory on request, quota can be expanded
- Persistent, stable, but not suitable for archival
- Looks like a POSIX file system
- Backed by CephFS

Vault / vos:

- Suitable for data publication and long term curation
- Data mirrored at two sites
- Backed by Ceph Object Store
- Available by request
- Self-serve DOIs also available

Both:

- Accessible via
  - VOSpace API
  - Python Client
  - Web UI
- Fine grained access control

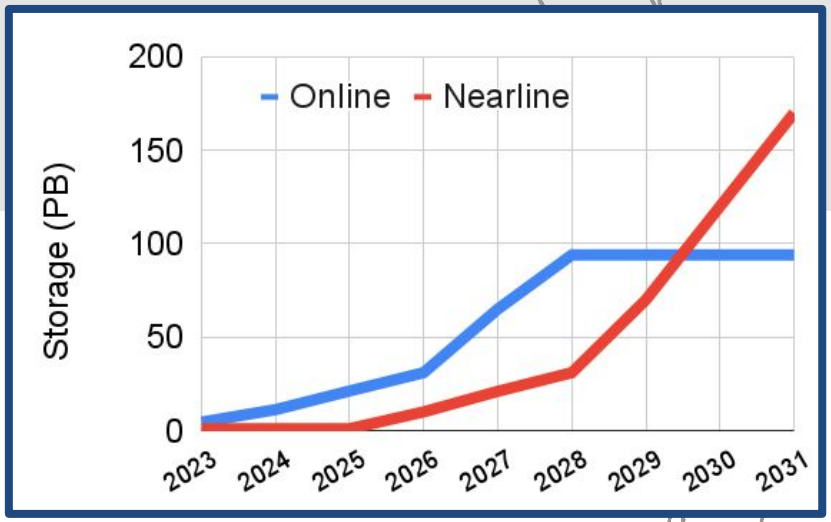


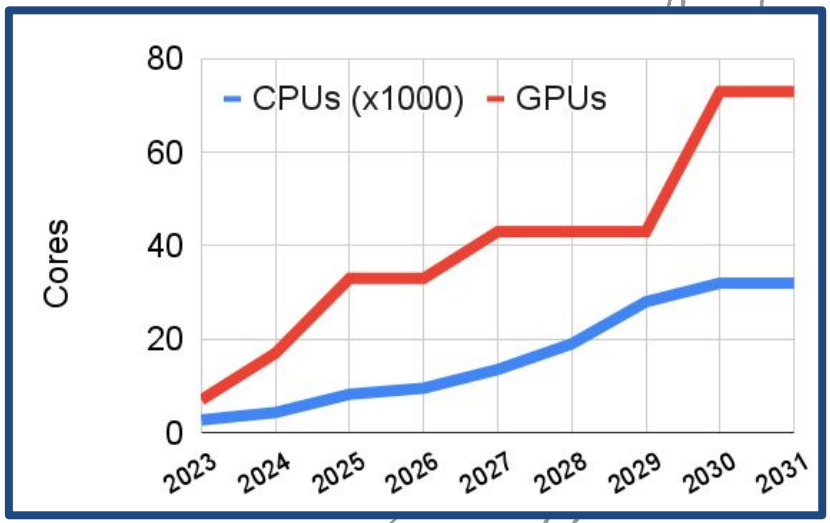
Enhancements to infrastructure

Currently:
- 10PB
- 3000 CPUs
- 10 V100 + 1 A100 GPUs
- 20TB database

Arrived for 2025
- 8PB
- 3900 CPUs
- 16 H100 GPUs
- 200TB federated database

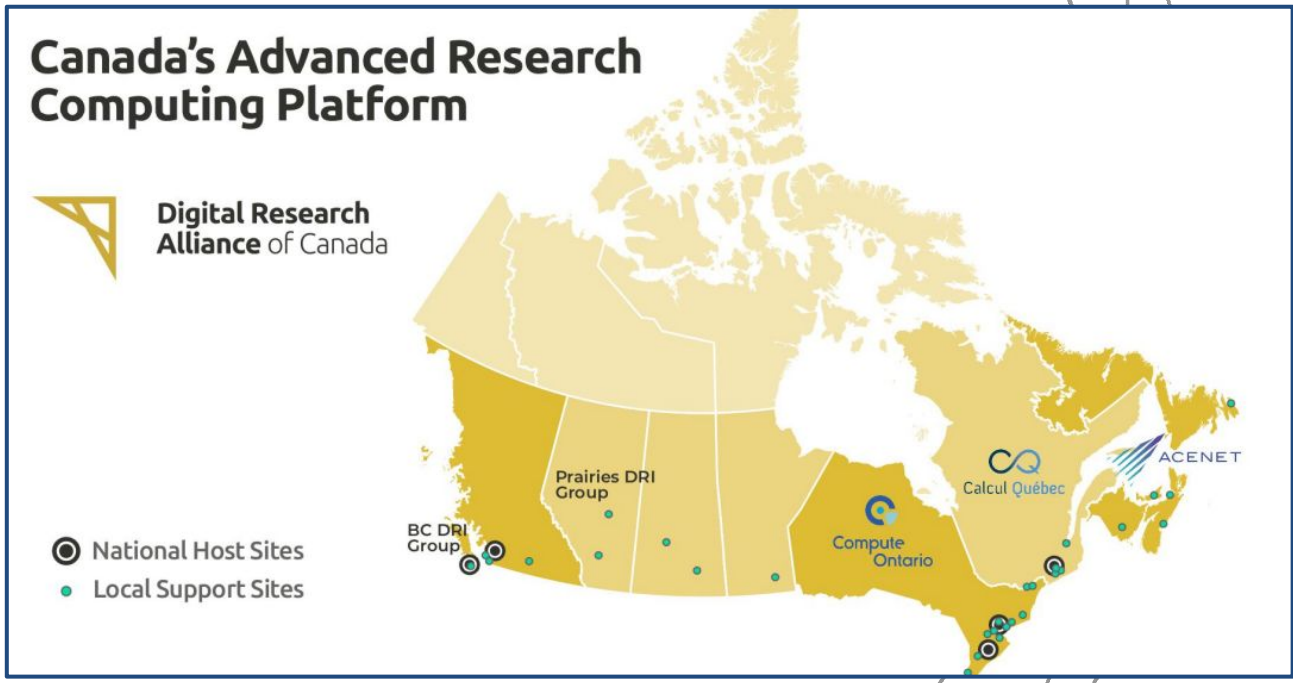
Steady state 2030
- 32000 CPUs
- 72 GPUs
- 90 PB online storage
- 170 PB nearline storage
    - (growing 50PB/year)



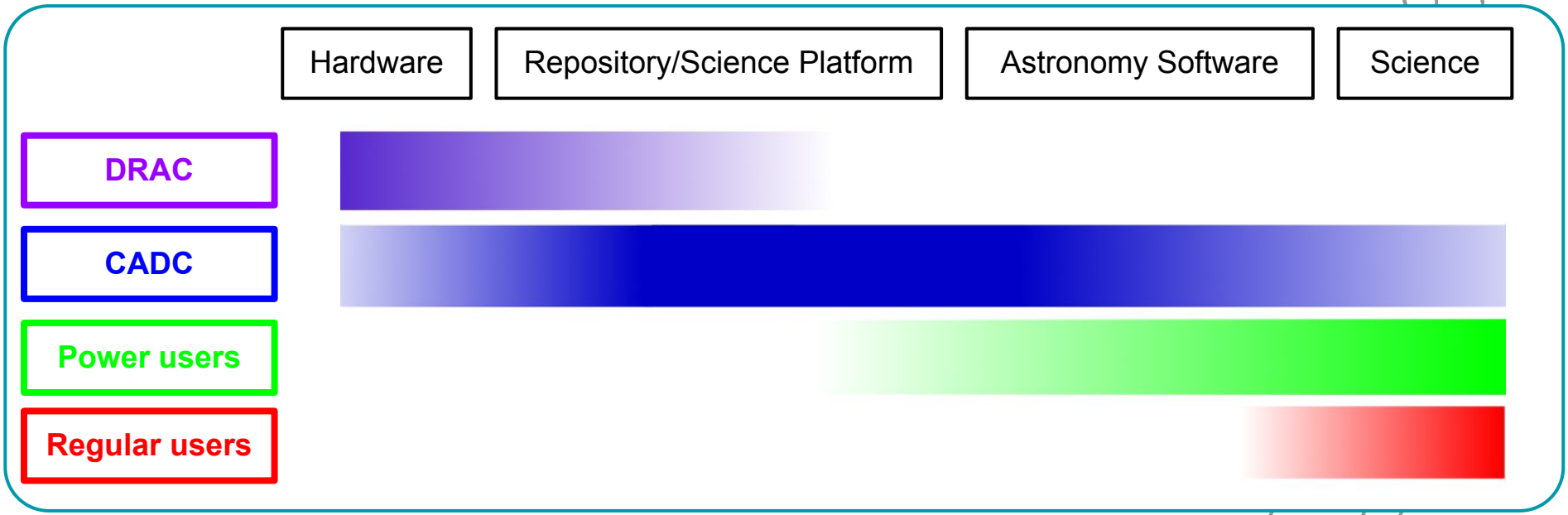


Digital Research Alliance of Canada

- Provide research IT infrastructure to Canadian academics
- Long-standing partnership with CADC
- Initially through a resource allocation through university-based astronomers
- Expansion driven through funds transferred from CADC

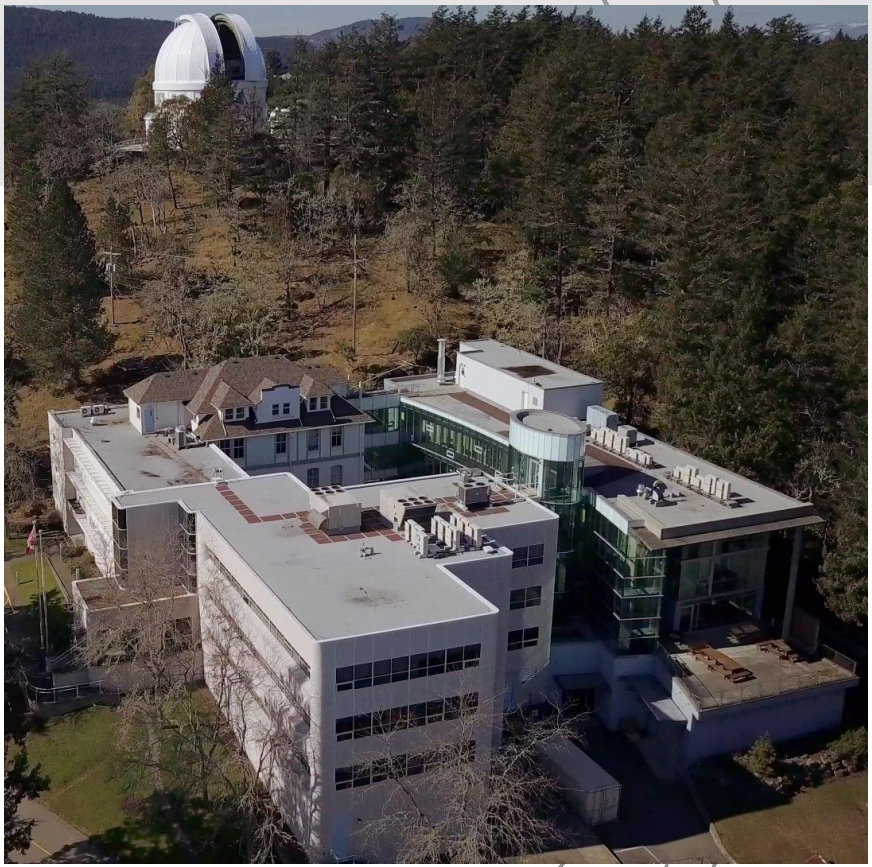


A spectrum of competancies



The Canadian IDAC

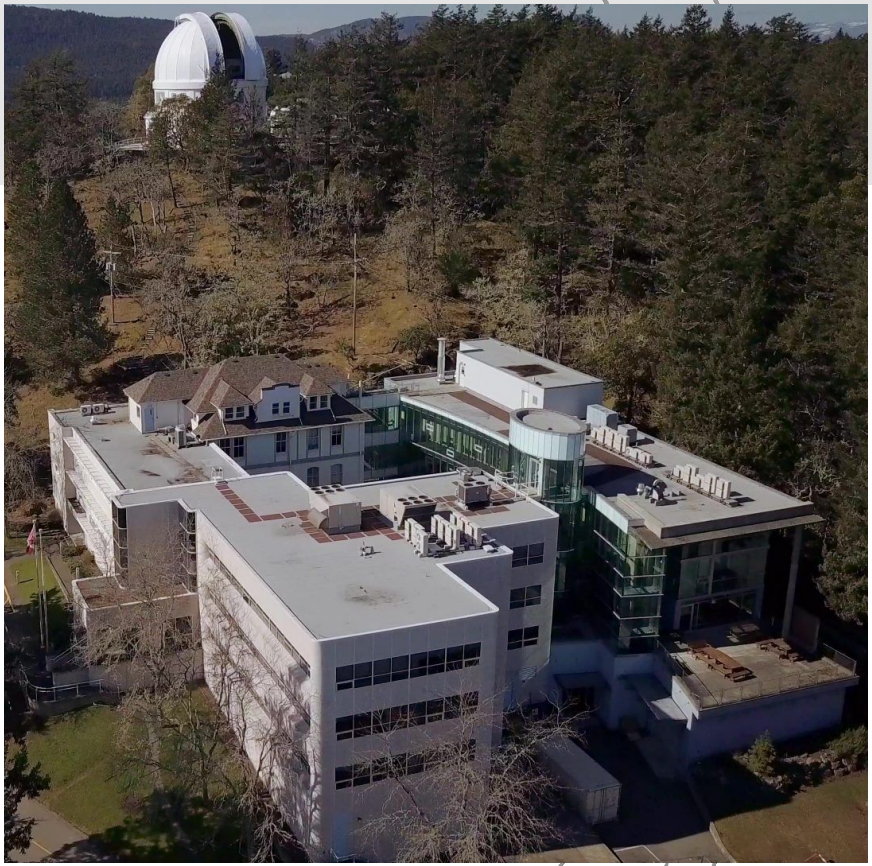
- The CADC is building an IDAC
- DP1 and DP2 will be available through CANFAR
- For the DRs we will host
  - coadded images: 2.7 PB
  - ObjectLite catalog
  - providing 3000 cores (may change some to GPUs)
  - providing 2PB of user storage
- One set of resources for Rubin data-rights holders
- One set of resources for the wider astronomical community when the data becomes public



The Canadian IDAC

Interfaces:

- CADC tools
  - Data storage / retrieval
    - StorageInventory
    - VOSpace
  - Database interface
    - CADC TAP service
  - Processing
    - CANFAR Science Platform
- LSST tools
  - DataButler
  - LSSTpipe container
- Data visualization
  - Rubin Science Portal

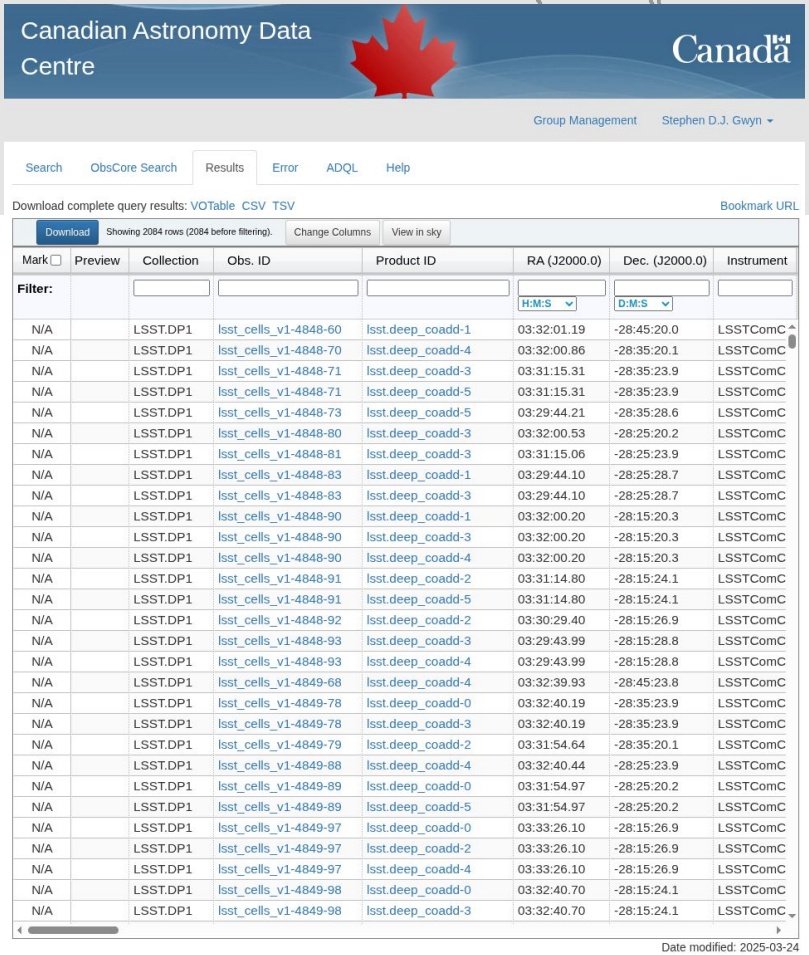


CADC's place in the LSST computing landscape

- Based on our experience with CANFAR, usage is evenly distributed in log space
  - most users require very few resources: Rubin Science Platform
  - some users require some resources
  - a few users require a lot of resources: IN2P3, NERSC, EPCC
- For the Rubin data rights holders
  - we best serve users who require a more power than RSP provides
  - not (currently) optimized for large single-user allocations
  - "First come, first serve" provides maximum flexibility
- For the public releases, the "first come, first serve" method is the only option
- Besides LSST, the CADC is hosting:
  - SKA
  - Euclid
  - CFHT (UNIONS = LSST of the North)
  - All available through a common set of tools and interfaces

Current status

- DP1 is ingested into the CADC
  - Access via CADC tools
  - Access via Rubin tools / DataButler
- Access control is by CADC username
  - Need to be manually linked to Rubin rights
- Working on a general solution:
  - Same login as RSP will get you access to the CADC
- Rubin users currently have full access to all CANFAR resources
  - Separate allocations not yet in place
- Test users welcome!



LSST activities in Canada

- Canadian LSST project lead: Renée Hložek, University of Toronto
- In-kind contributions
  - Directable development
  - Skilled software development (science pipelines)
    - Dark energy science
    - Galaxies science
    - Transients
    - Solar system science
      - RAFTS (Research Announcements For The Solar system)
  - CanDIAPL = Canadian Data-Intensive Astrophysics PLatform
  - 17 PI slots
- CADC IDAC: 30 PI slots
- Total 47 PI slots + 4x postdocs+students/PI
- ICRADA is (finally) signed

Summary

- The CADC is the astronomical data repository for Canada
- CANFAR provides a compute/user storage platform to support multiple projects through a common set of tools
- CADC/CANFAR is growing by a factor of 30 to support SKA+LSST
- The CADC supports LSST with
  - data hosting
  - user processing and storage
  - software development
- CANFAR provides a intermediate level of resources relative to RSP or IN2P3 / NERSC / EPCC