A Square Kilometer Array Regional Centre: Scaling Digital Research Infrastructure for Astronomy in Canada

Stephen Gwyn
Science Lead, Canadian Astronomy Data Centre

Canadian Astronomy Data Centre

Preserving and distributing astronomical data

- The CADC archives and distributes the data from all Canadian telescopes, including JWST, HST, and NEOSSat in space, and CFHT, Gemini on the ground
- Holdings:
  - 2.3Pb
  - 300 million files
- 3 copies:
  - 2 on hardware provided by Digital Research Alliance of Canada
  - 1 on NRC hardware
- Annual downloads:
  - 100 million files
  - 4.9 Petabytes
  - ~10,000 users worldwide

By promoting data reuse, a good telescope archive can double to triple the impact of a facility for minimal cost

Canadian Astronomy Data Centre

Preserving and distributing astronomical data

- The CADC archives and distributes the data from all Canadian telescopes, including JWST, HST, and NEOSSat in space, and CFHT, Gemini on the ground
- Holdings:
  - 2.3Pb
  - 300 million files
- **3 copies:**
  - **2 on hardware provided by Digital Research Alliance of Canada**
  - **1 on NRC hardware**
- Annual downloads:
  - 100 million files
  - 4.9 Petabytes
  - ~10,000 users worldwide

By promoting data reuse, a good telescope archive can double to triple the impact of a facility for minimal cost

Storage:

- Ceph Object Store
- Storage Inventory ≈ dCache
  - syncs data between the three sites
  - files are retrieved from any of the sites for redundancy and availability
  - allows seeks within a file: image sub-raster cutouts very important
- Comprehensive metadata system:
  - CAOM2 (Common Archive Observation Model)
  - Allows users to find images from 216 different telescopes and instruments

CORE TRUST SEAL

Astronomy Data Management

Workflow #1: VERTICO (Virgo Environment Traced in CO)

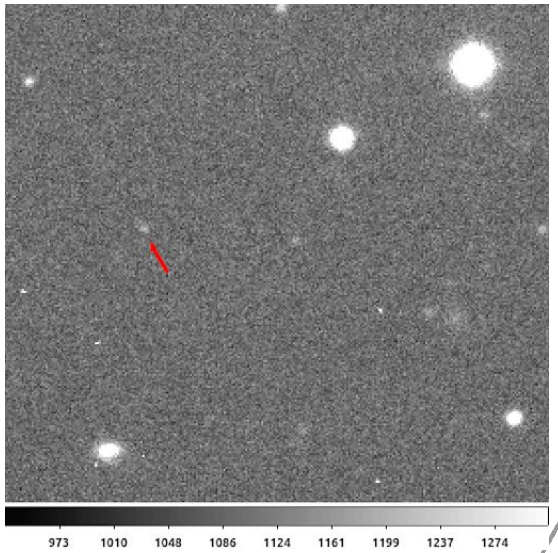
- Raw data transferred to data centre
- Data downloaded to the CADC and processed
- generic software (CASA) but specialized parameter choices
- generic hardware
- relatively small data sets (10s of Tb)

Nearby galaxy as seen in a carbon monoxide (a tracer of star formation

Astronomy Data Management

Workflow #2: OSSOS (Outer Solar System Origins Survey)

- Telescope observes sky at night
- Images are transferred automatically in minutes to CADC
- Basic calibration is done by astronomer A, using software specific to instrument
- Images are searched for moving objects by astronomers B,C and D using software specific to the project
- Generic hardware
- Selected objects are queued for follow up observation that evening

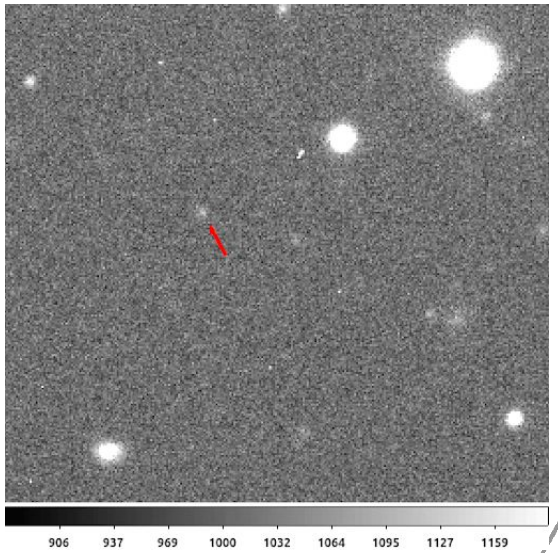


Detection triplet of Kuiper Belt Object

Astronomy Data Management

Workflow #2: OSSOS (Outer Solar System Origins Survey)

- Telescope observes sky at night
- Images are transferred automatically in minutes to CADC
- Basic calibration is done by astronomer A, using software specific to instrument
- Images are searched for moving objects by astronomers B,C and D using software specific to the project
- Generic hardware
- Selected objects are queued for follow up observation that evening



Detection triplet of Kuiper Belt Object

Astronomy Data Management

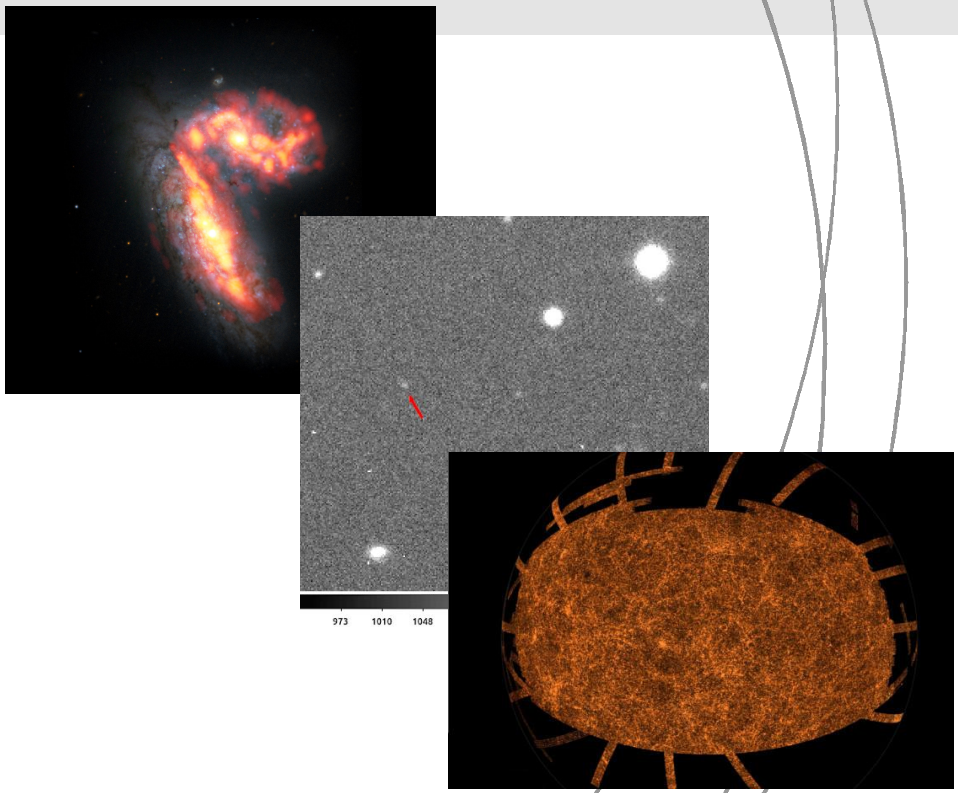
Workflow #3: SDSS (Sloan Digital Sky Survey)

- Telescope spends multiple years observing northern sky with imager and spectrograph
- Data is processed by a team of astronomers, developers and operations staff
- Specialized data pipeline
- Dedicated hardware
- Individual astronomers (or small teams) interact with processed data as database queries plus (maybe) spot checking the original images/spectra

Distribution of galaxies in the Northern Galactic Cap, showing large scale structure

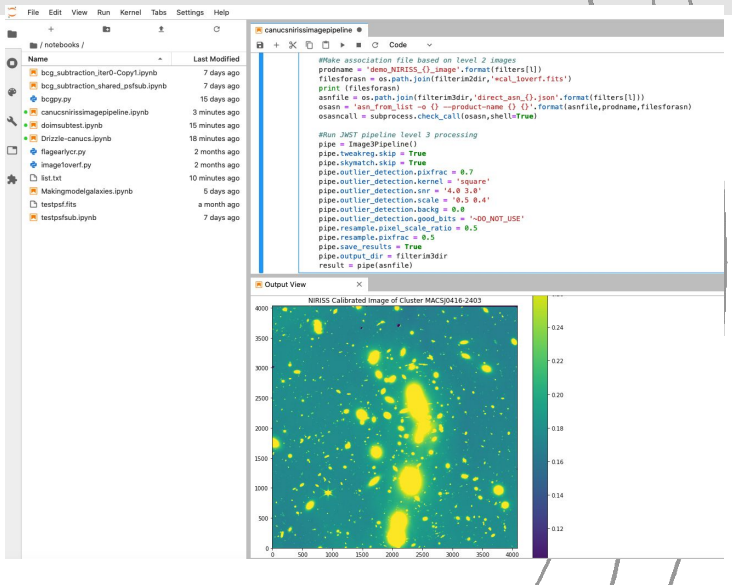
Astronomy Data Management

- Astronomy data is fairly heterogeneous
- Astronomy use cases are also heterogeneous
- Astronomy data management is similarly heterogeneous
- There are a few large, multi-purpose/ generic software packages, but a lot of astronomy software is more “artisanal”, tuned to a specific instrument or a specific use case
- While computer scientists are extensively employed in the infrastructure development, they are extremely rare in the scientific software development
    - Code is optimal for scientific analysis
    - Less than optimal for performance and maintainability



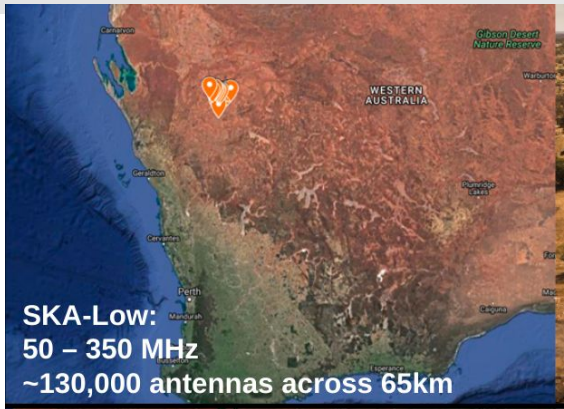
CANFAR: Canadian Advanced Network for Astronomical Research

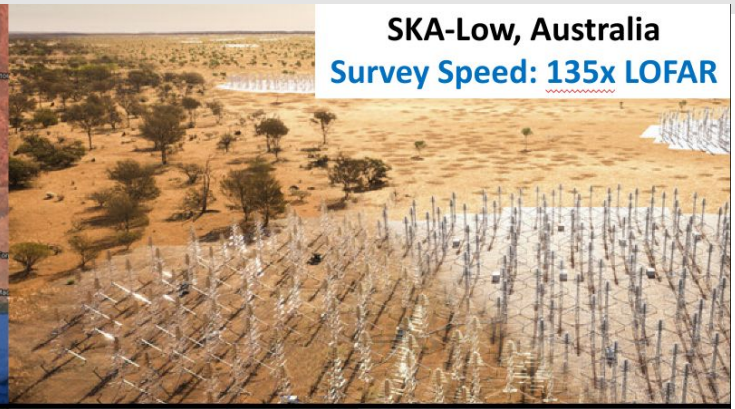
- Runs on Digital Research Alliance of Canada infrastructure
- User data storage
  - Tb-scale storage for astronomy
  - Fine-grained access control
  - Browser and python client interaction
- User controlled computing infrastructure
  - Version 1: (2008)
    - VM-based
    - powerful, but steep learning curve for users
  - Version 2: (2021)
    - Container-based
    - Jupyter notebooks
    - Browser-based VNC desktops
    - Data visualization
    - Ability to share containers
    - Very successful: 3x growth in 2021
    - ~⅓ of Canadian astronomers are users

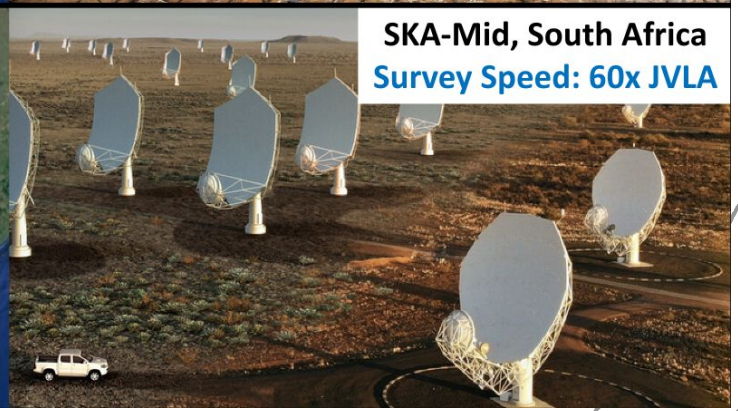


Screenshot from a JWST python notebook

Square Kilometer Array







SKAO Science Working Groups

Cosmology

Cradle of Life

Epoch of Reionization

Extragalactic Continuum

Extragalactic Spectral Line

Gravitational Waves

High Energy Cosmic Particles

HI Galaxy Science

Magnetism

Our Galaxy

Pulsars

Solar, Heliospheric & Ionospheric Physics

Transients

VLBI

Among the broadest science cases for observatories worldwide

Slide from Phil Diamond /

SKAO Science Working Groups

Cosmology

Cradle of Life

Epoch of Reionization

Extragalactic Continuum

Extragalactic Spectral Line

Gravitational Waves

High Energy Cosmic Particles

HI Galaxy Science

Magnetism

Our Galaxy

Pulsars

Solar, Heliospheric & Ionospheric Physics

Transients

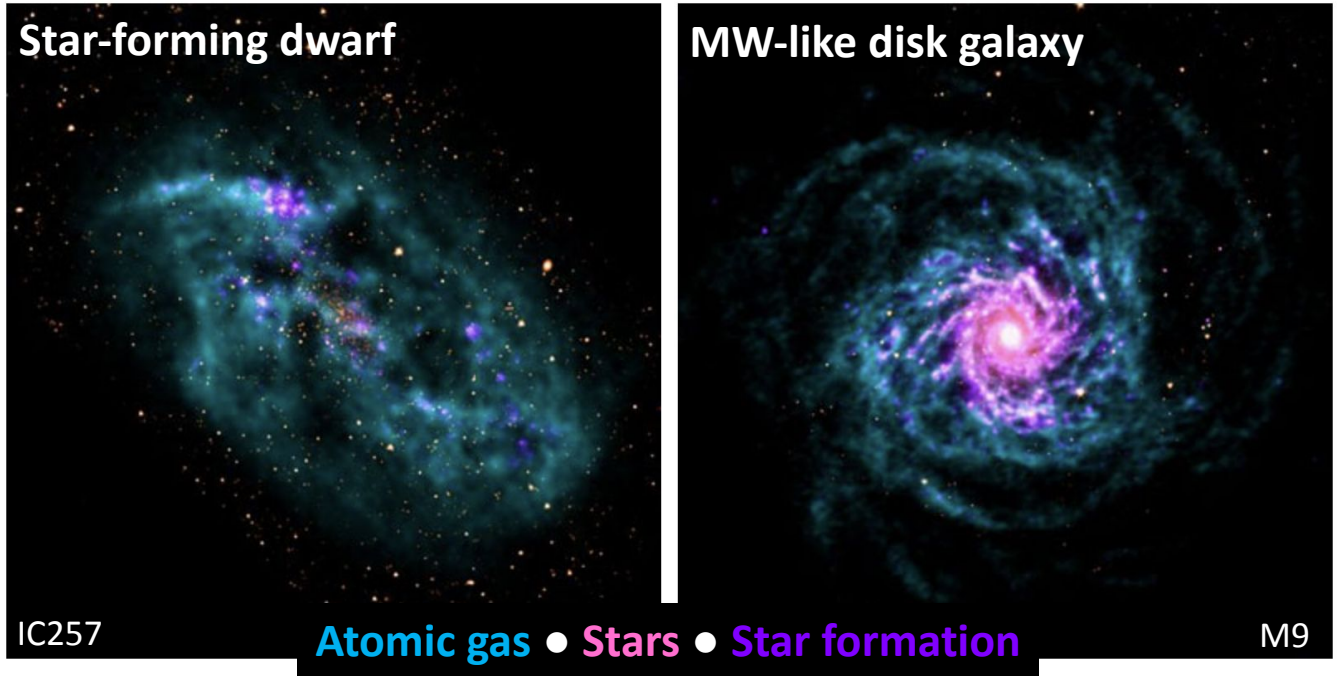
VLBI

Among the broadest science cases for observatories worldwide

Slide from Phil Diamond /

Atomic gas (HI) in disk galaxies

HI is more abundant than stars in dwarfs →



← HI disks extend farther into halos than the stars

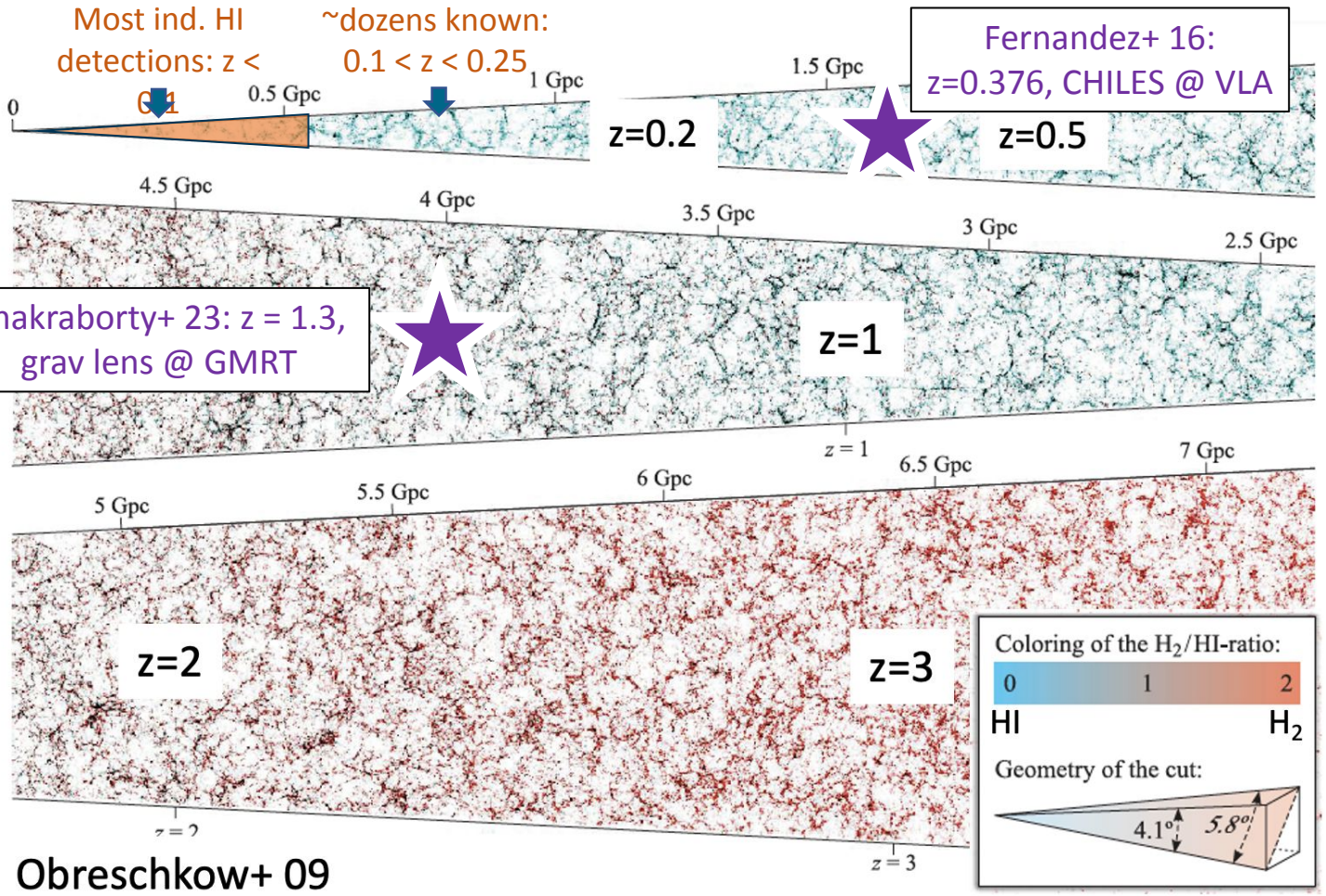
The HI content, morphology and kinematics of galaxy populations probe cosmological galaxy formation

The cosmic HI census

State of the art:

HI out to z~0.1, targeted maps within 100 Mpc.

Few individual galaxy detections at z > 0.1.



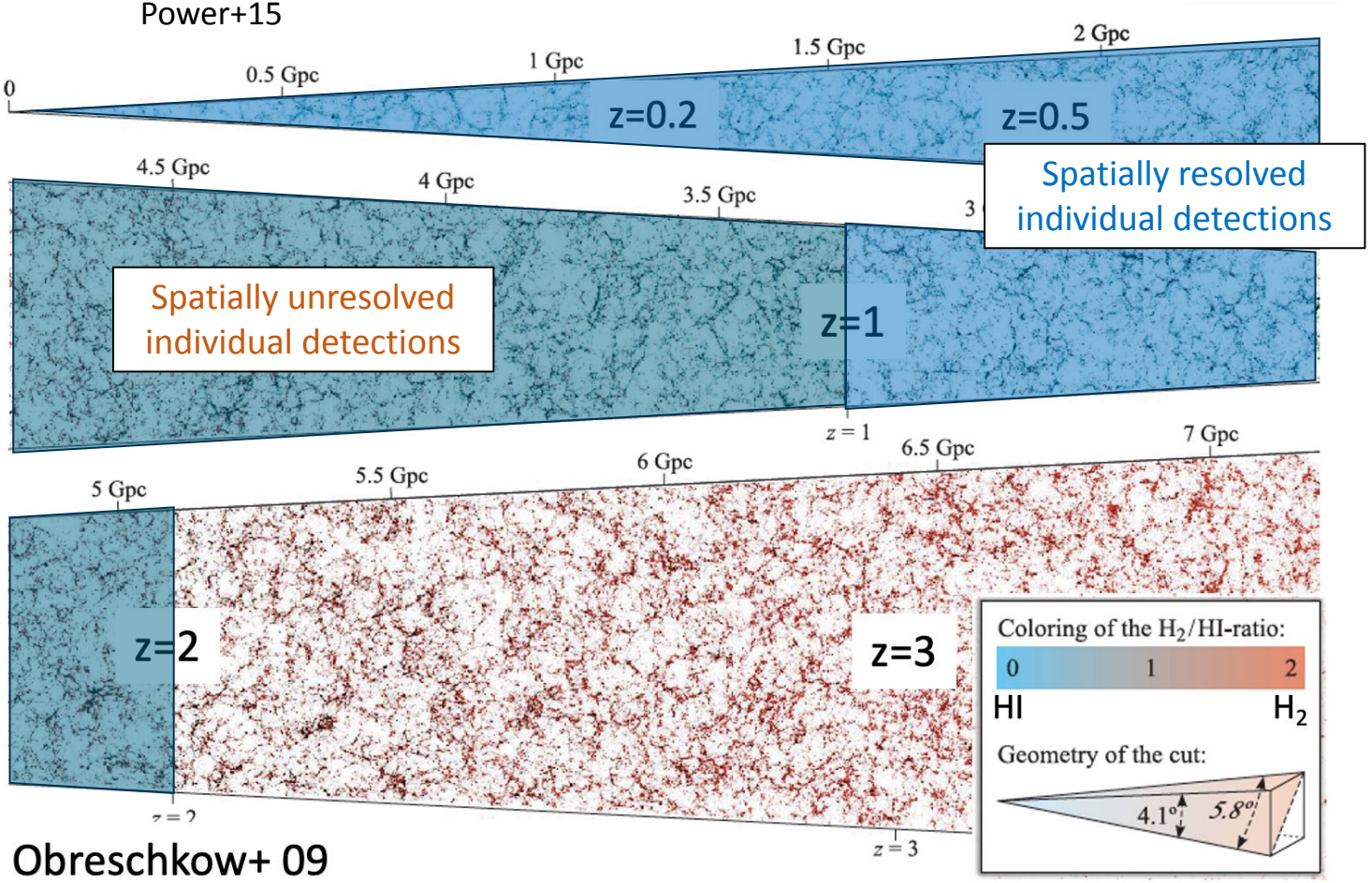
Obreschkow+ 09

The cosmic HI census

SKA KSP, 10,000 hrs (~2030-35): AM buildup across cosmic time

HI out to z ~ 2, map HI disks [$\log(M_{HI}/M_o) > 10$] to z ~ 1.

Staveley-Smith + Oosterloo 15; Blyth+ 15; Meyer + 15; Obreschkow+ 15; Power+15



Obreschkow+ 09

Square Kilometre Array Observatory

Intergovernmental organization

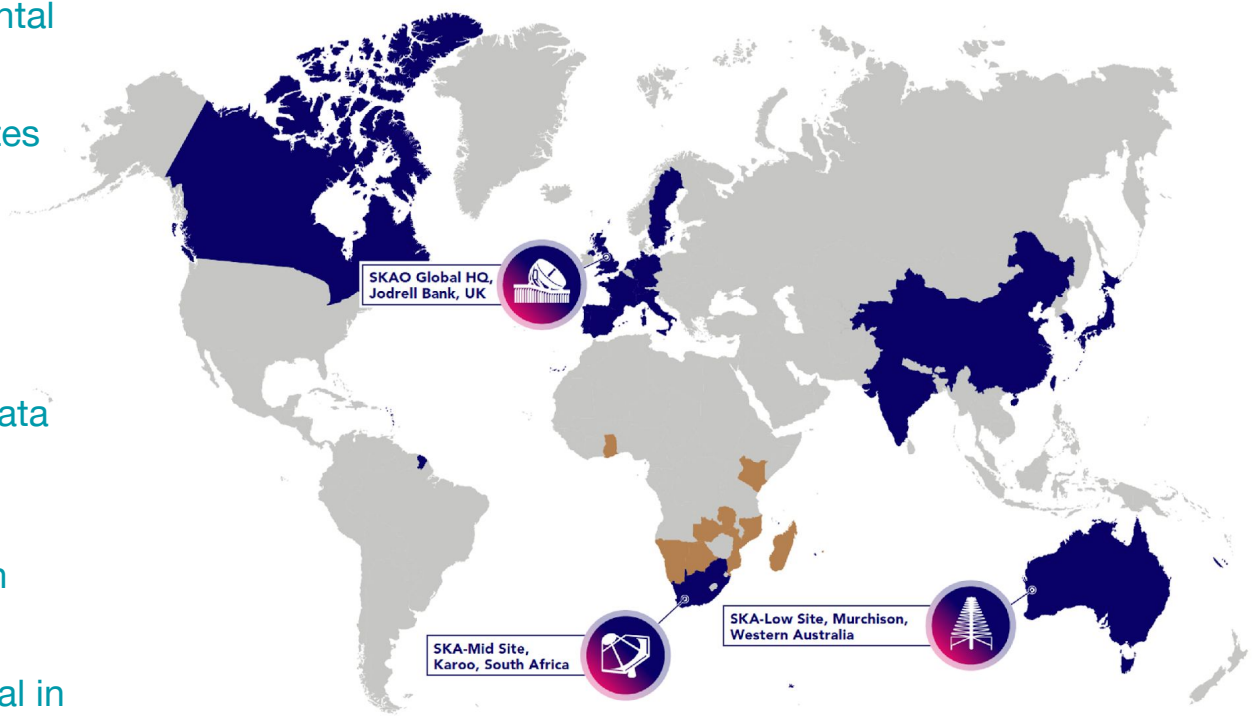
9 Member States
2 Acceding
5 Observers

Construction underway

First science data in 2026

Shared-risk observations in 2028

Fully operational in 2029/30



January 2023: Canada announces intention to join the SKA Observatory with a 6% share

May 2023: Treasury Board approves budget to NRC ($270 million)

SKA Regional data Centres (SRCs)

- Raw data is processed at Science Data Processors (SDPs) in South Africa and Australia
- Data is transferred to SRCs for user-driven science analysis
- Long term archiving will be federated; no single site will have a copy of all the data
- At the end of proprietary period, other astronomers should be able to schedule processing
- Code-to-data: Avoid unnecessary data movement -> A job should be able to run everywhere

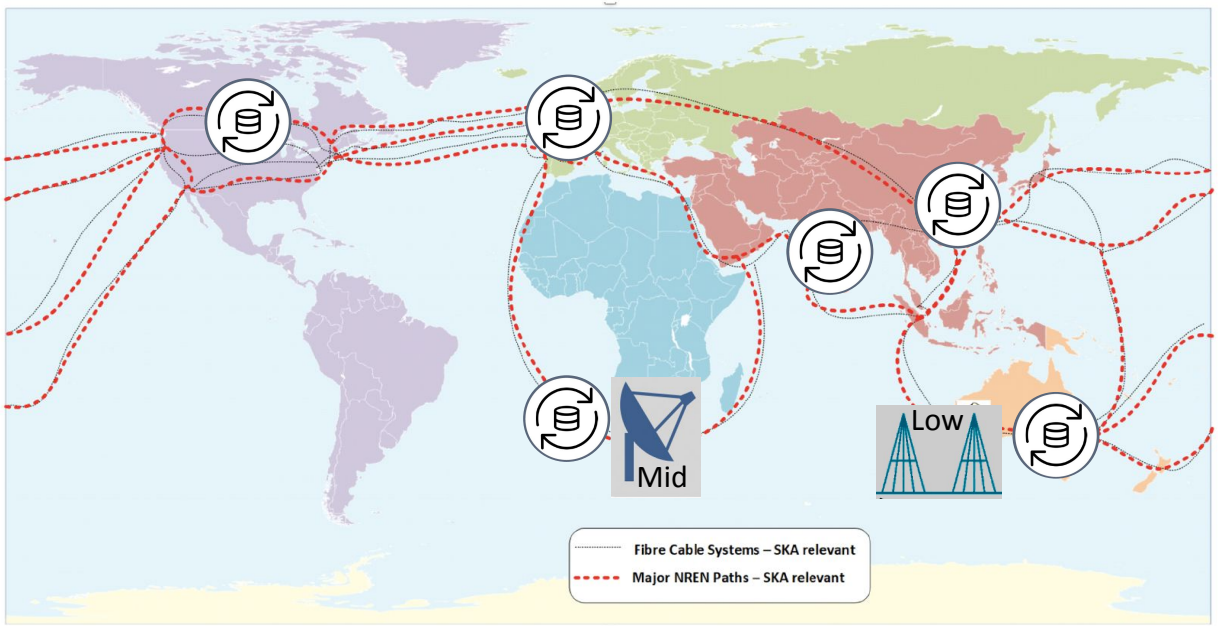


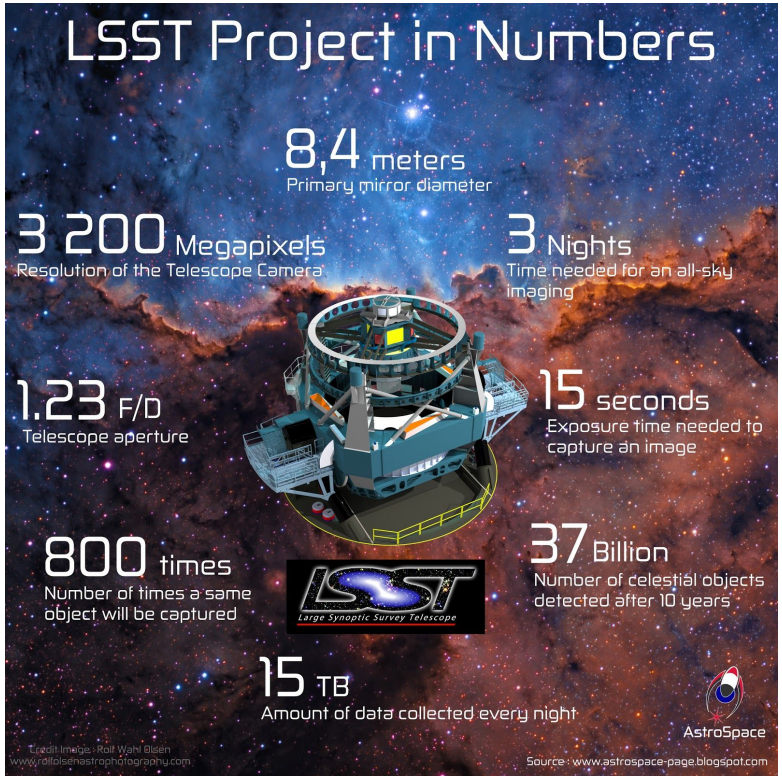
Image credit: SRCSC and SKAO

CADC and SKA

The CADC has been making a strategic investment in SKA since 2012

- **2012-2017**: Design of Delivery Subsystem of the Science Data Processor
- **2016–2018**: Member of SRC Consultation Group
- **2017–2019**: Participant in the AENEAS Horizon 2020 project on SRCs
- **2018–**: Member of the SRC Steering Committee
- **2018–: Supporting CIRADA (Canadian Initiative for Radio Astronomy Data Analysis)**
- **2020**–: Participation in the SRC Requirements, Design and Prototyping phases
  - Leading the SRCNet demonstrator epic being deployed at 3+ sites (3 FTE)
  - Contributing software for data management, metadata management, science platform, user storage
- **2023**: Official SKA Canada including SRC Canada budget approved

Vera Rubin Observatory
Legacy Survey of Space and Time

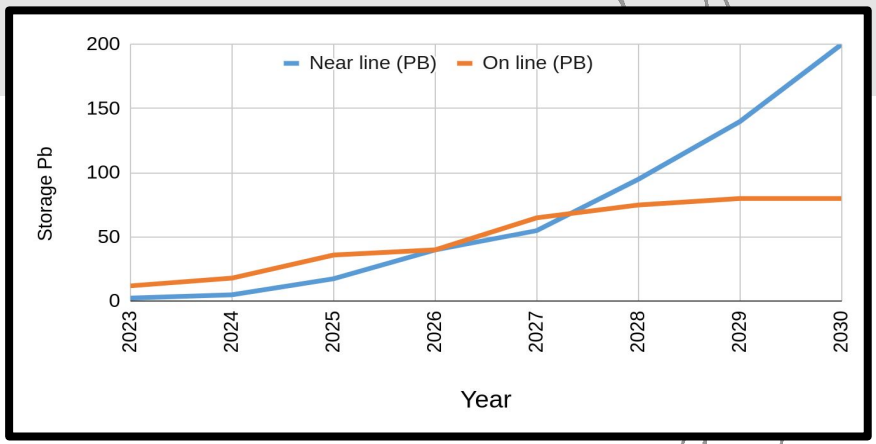


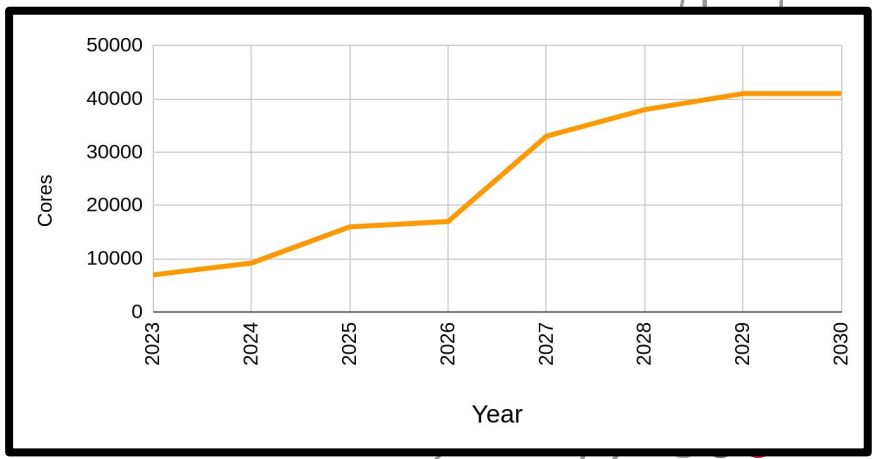
- Time variable sky: observe the sky repeatedly and note what changed in brightness or moved
- Add all the images together to produce a deep map of the sky on an annual cadence
- Canada will host an alerts database
- We will host 2 copies of the static sky: the latest data release, plus the latest public data release
- The CADC is currently the only group planning a public interface
- Total archive data volumes:
  - 3Pb database
  - 10Pb of images
- Extension to science platform:
  - 6000 cores
  - 4Pb of user storage

CADC expansion

Hardware expansion

- CADC currently:
  - 2.3Pb (x3 replication) on-line
  - 3000 cores
- CADC in 2030 and beyond
  - 100Pb on-line
  - 60Pb/year growth of nearline
  - 42000 cores
- All new hardware will be managed by the Digital Research Alliance of Canada
- Most (all?) will be here in Victoria

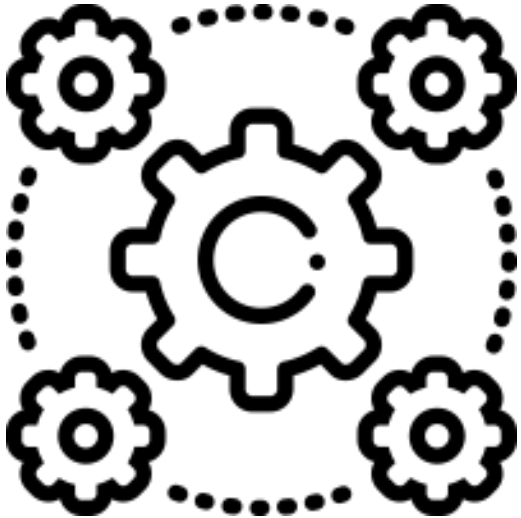




Interoperability

Need to interoperate with other SRCs

- “A job should be able to run everywhere”
- Interoperability but at what level?
- Hardware? OS? Software/Middleware? API?



Need to interoperate with other CADC archives

- CADC data volumes:
  - SKA 90%
  - LSST 9%
  - Everything else 1%
- CADC users:
  - SKA 20%
  - LSST 20%
  - Everything else 60%

The CADC’s goal is to build and run one system to serve all communities

Summary

- The CADC is a Pb scale astronomy data centre serving Canada and the world
- Canada is joining SKA at the 6% level
- The CADC is building an SKA Regional Centre
- To support the SRC and LSST, the CADC will have to grow by a factor of x100 by 2030
- Our aim is to build a single system that will support a variety of astronomy computation
- The CADC is here to learn from the HEP community

~~Any questions?~~ Answers please

extra slides

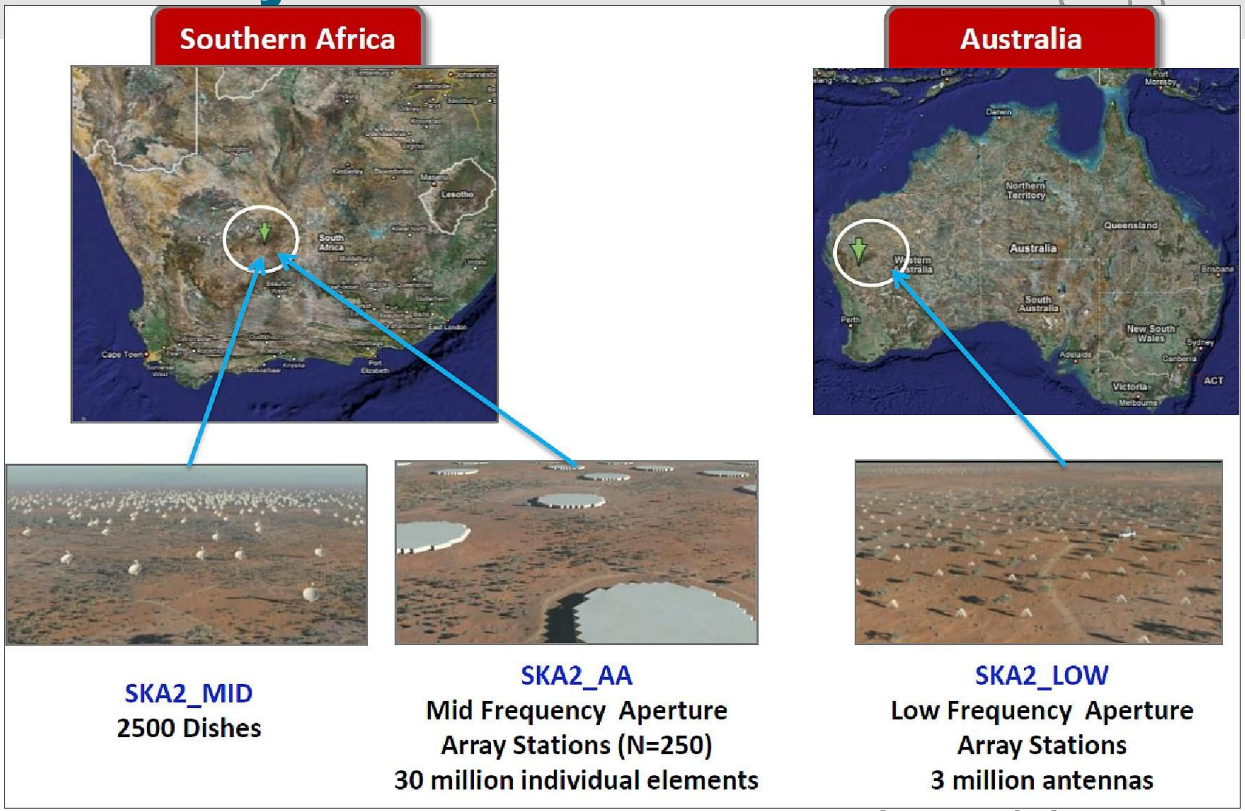
Square Kilometer Array

SKA-Mid

- South Africa
- 197 dishes (15m across)
- 350Mhz-15.4Ghz
- 150 km across

SKA Low

- Australia
- 130 000 dipoles
- 50-350Mhz
- 75km across



Square Kilometer Array