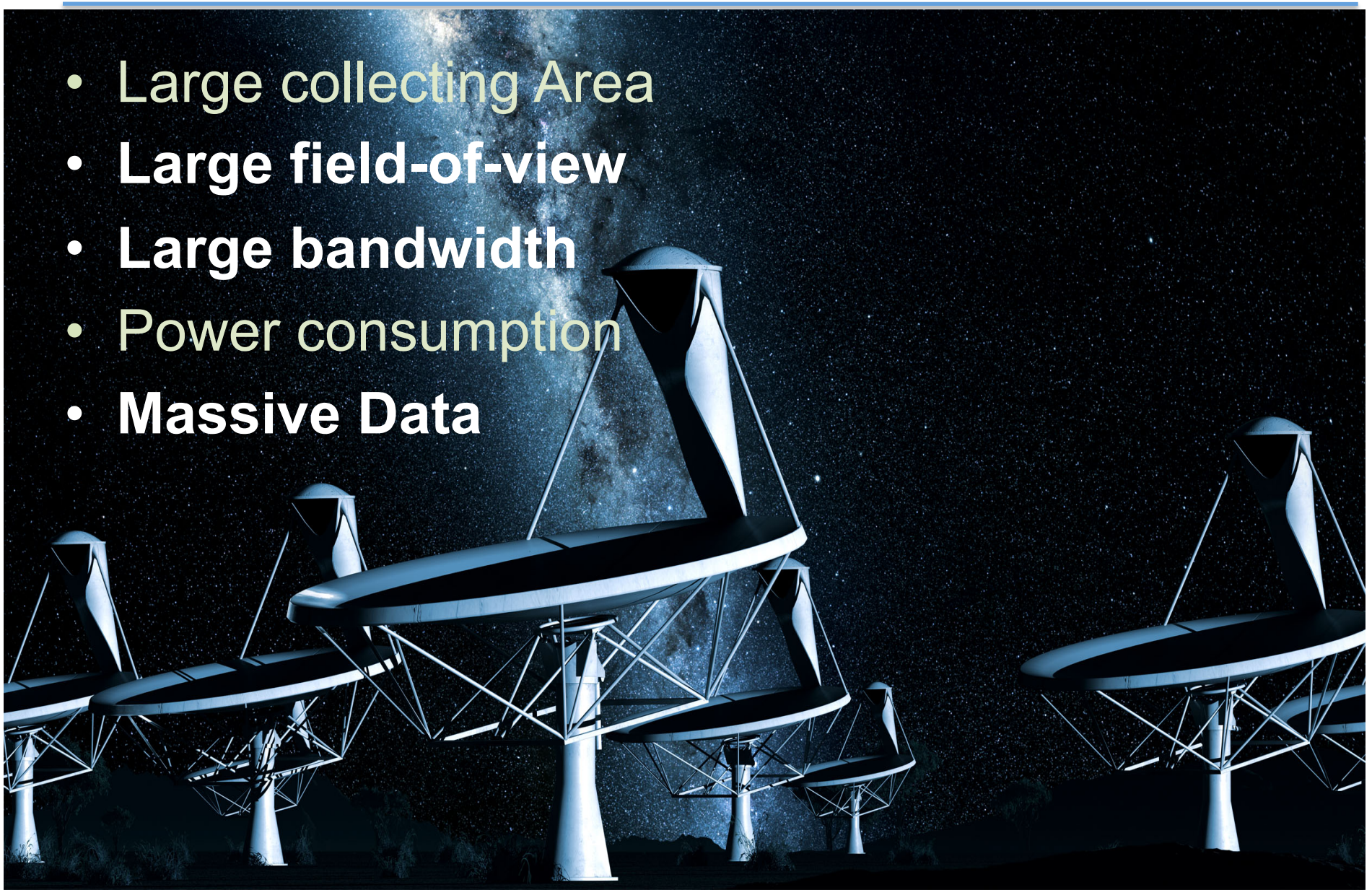
# Data Intensive Radio Astronomy en route to the SKA

## The rise of

# BIG RADIO DATA

Russ Taylor
Institute for Space Imaging Science
University of Calgary

# Technology Challenges en route to SKA

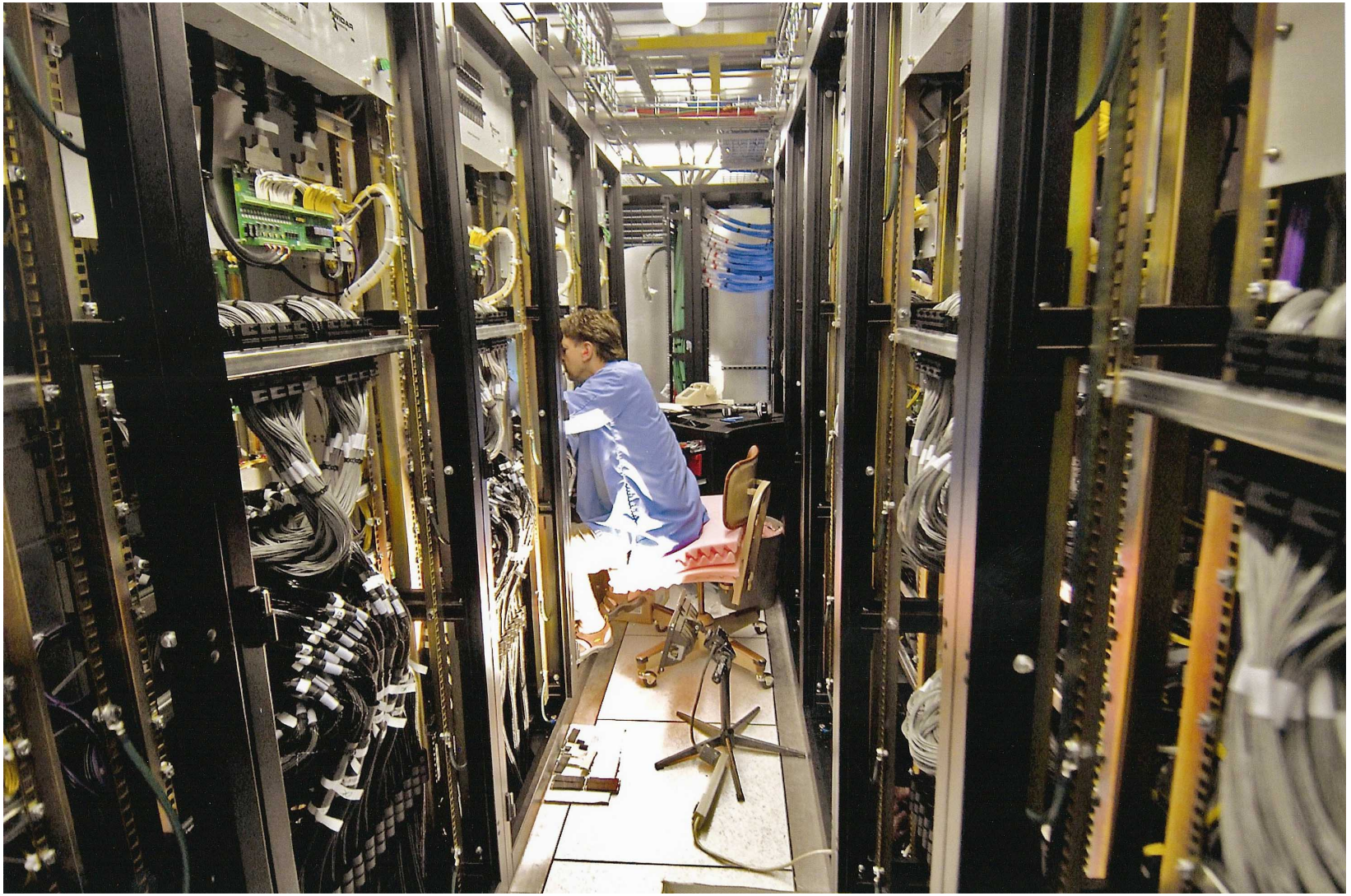


- Large collecting Area
- **Large field-of-view**
- **Large bandwidth**
- Power consumption
- **Massive Data**

# Bandwidth

- Until recently bandwidths of a few 10's of MHz were standard
  - VLA: 50 MHz per IF
- New telescopes and upgrades dramatically increase bandwidth
  - Arecibo: 300 MHz
  - ASKAP: 300 MHz
  - MeerKAT: 770 MHz
  - JVLA: 8 GHz
- At the same time observing in "spectral-line" mode
  - for RFI detection and excision (BW extends beyond protected regions)
  - Avoid signal degradation for wide-field imaging.
  - Use of multiple frequencies for visibility sampling (multi-frequency synthesis)

# WIDAR Correlator Rack Installation at JVLA, Aug 2008



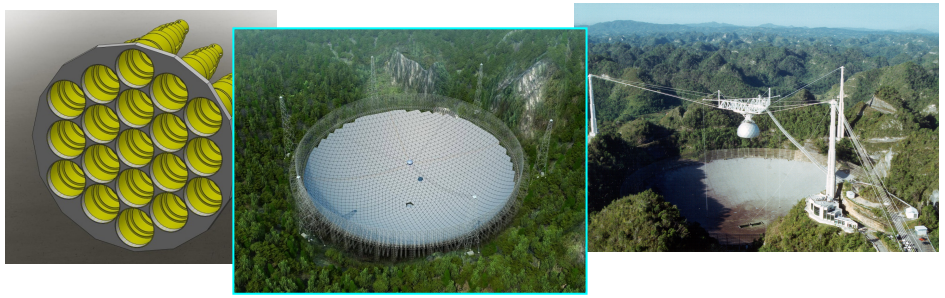
# Jansky VLA-VLA Comparison

| Parameter | VLA | JVLA | Factor | **Current** |
|---|---|---|---|---|
| Point Source Cont. Sensitivity (1σ,12hr.) | 10 μJy | 1 μJy | 10 | 2 μJy |
| **Maximum BW in each polarization** | **0.1 GHz** | **8 GHz** | **80** | **2 GHz** |
| # of frequency channels at max. BW | 16 | 16,384 | 1024 | 4096 |
| **Maximum number of freq. channels** | **512** | **4,194,304** | **8192** | **12,288** |
| Coarsest frequency resolution | 50 MHz | 2 MHz | 25 | 2 MHz |
| Finest frequency resolution | 381 Hz | 0.12 Hz | 3180 | .12 Hz |
| # of full-polarization spectral windows | 2 | 64 | 32 | 16 |
| (Log) Frequency Coverage (1 – 50 GHz) | 22% | 100% | 5 | 100% |

# Field of View technologies

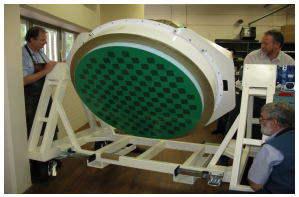
- Large N small D
  - 15 m antennas 1 sq deg at 1.4 GHz.

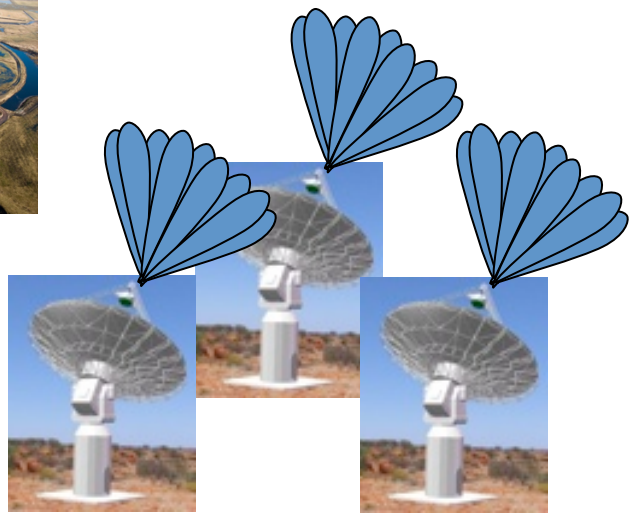
- Arecibo & FAST
  - Multiple horn receivers



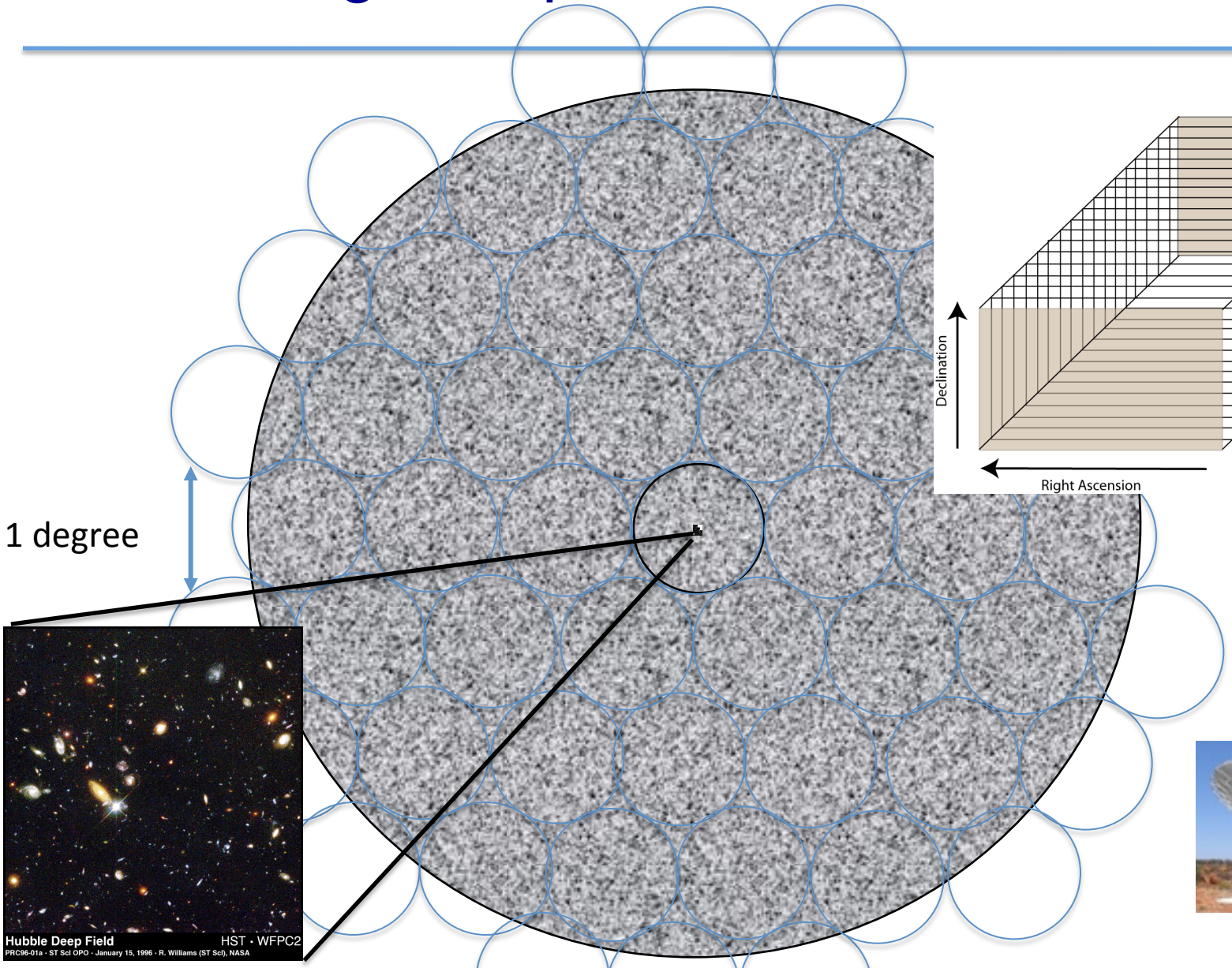
- Aperture plane arrays
  - Multiple independent beams over a hemisphere

- Phased Array Feeds
  - Multiple beams providing large FOV



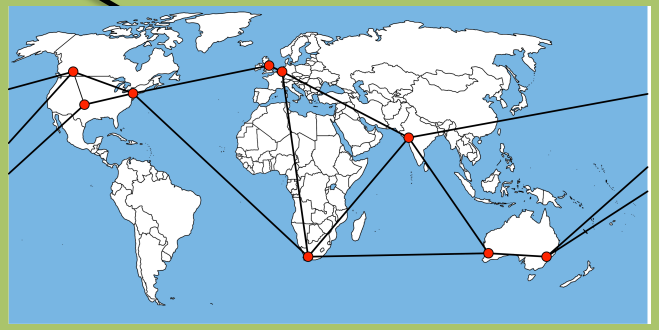


# FPA Image Outputs



# Sociology of Radio Astronomy

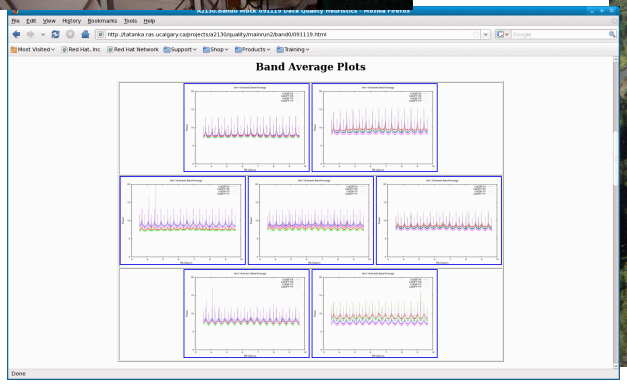
- Much of the key science en route to the SKA will be achieved via large-scale survey mode observing programs executed by globally distributed teams of researchers

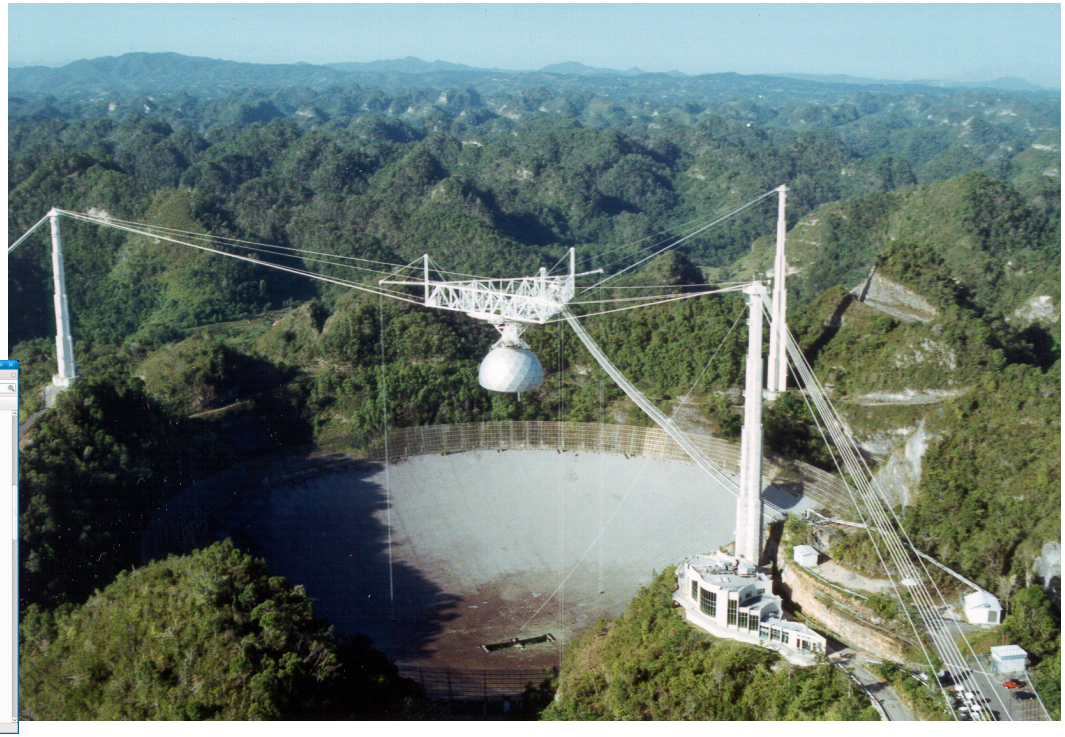


# Arecibo ALFA Surveys

- ALFALFA (HI)
- PALFA (Pulsars)
- GALFACTS (Spectro-polarimetry)

GALFACTS and PALFA
Aggregate rate 500 MB/s



# LOFAR Survey Science

- Sky surveys at  15, 30, 60, 120, 200 MHz
  - Galaxy formation
  - Intergalactic magnetic fields
  - Star formation in early universe
  - Expansion of discovery parameter space

# ASKAP Survey Science

- WALLABY (HI emission)
- EMU (continuum)
- POSSUM (polarization)
- FLASH (HI absorption)
- VAST (slow transients and variables)
- GASKAP (Galactic HI)
- CRAFT (fast transients)
- DINGO (Deep HI)
- COAST (pulsar and timing survey)
- VLBI (high resolution science)

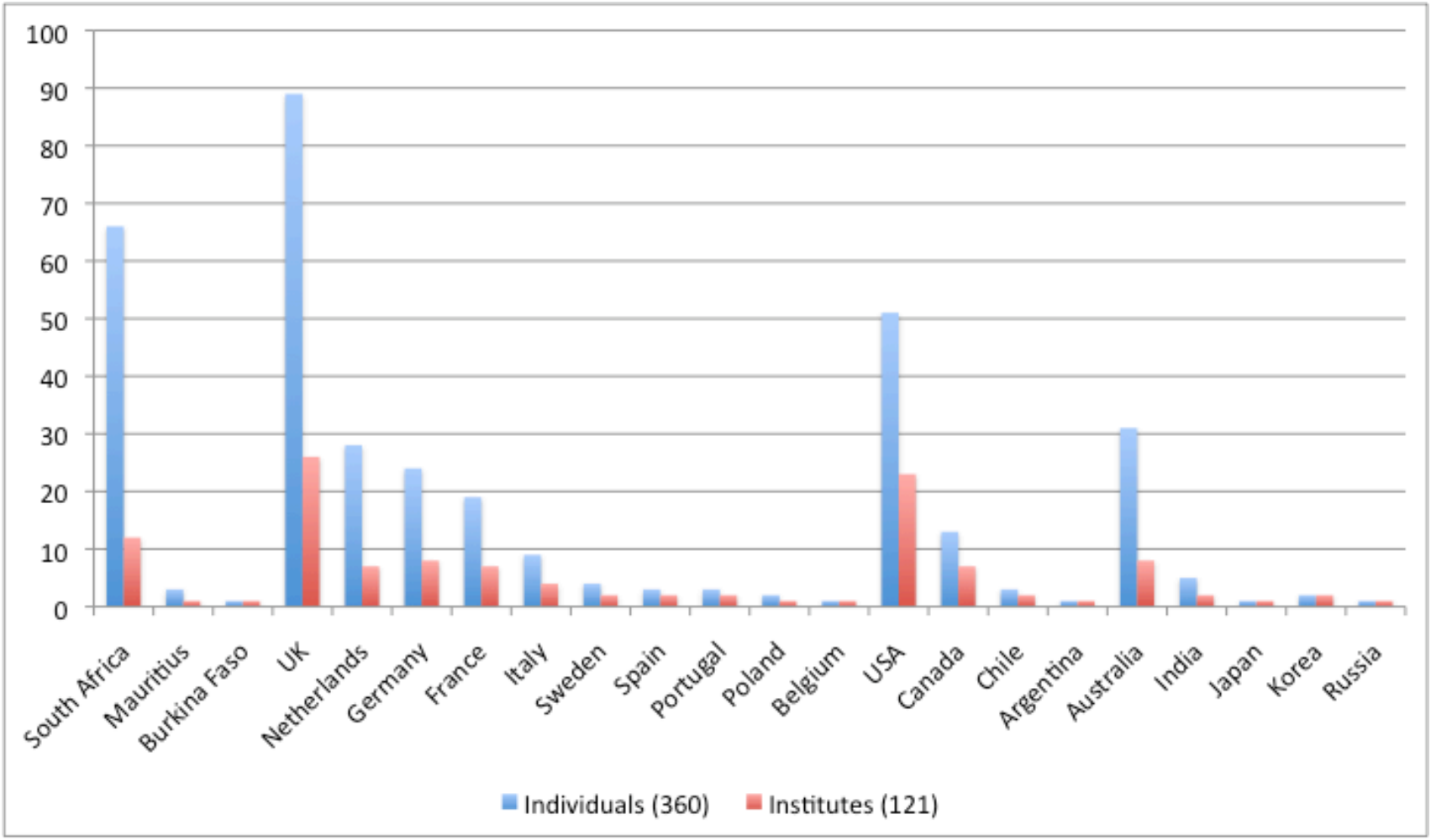
Some project will be commensal

# MeerKAT Survey Science

- Pulsar Timing
- LADUMA (Deep HI)
- MESMER (High-z CO)
- MeerKAT Absorption Line Survey
- MHONGOOSE (Nearby HI)
- TRAPUM (pulsar search)
- MeerKAT HI Survey of Fornax
- MeerGAL (Galactic Plane Survey)
- MIGHTEE (Deep continuum and polarization)
- ThunderKAT (variables and transients)

Some project will be commensal

# MeerKAT Large Surveys (43,000 hours allocated)




22 countries

# SKA Timeline

2008 | 2012 | 2016 | 2020

€ 200 M

Preparatory Phase

Site Selection

€ 90 M

**SKA1**

Pre-construction

€ 350 M

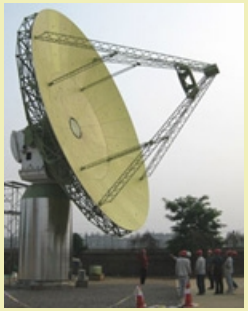
**SKA1**

250 15-m Dishes +
f=0.1 Aperture Array

€ 1.5 B (1 & 2)

**SKA2**

2500 15-m Dishes +
f = 0.5 Aperture Array



ASKAP & MeerKAT

0.5 – 10 GHz

MWA, LOFAR, PAPER

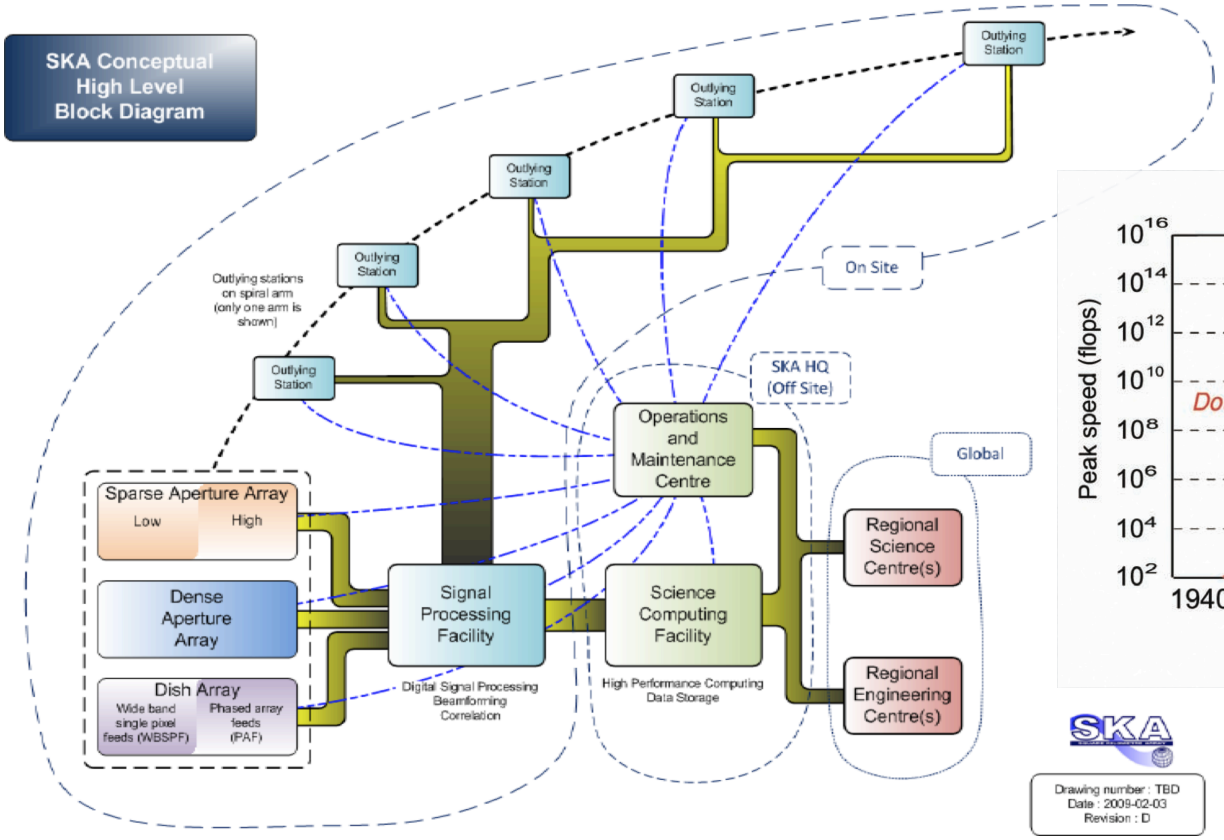
0.1 – 0.5 GHz

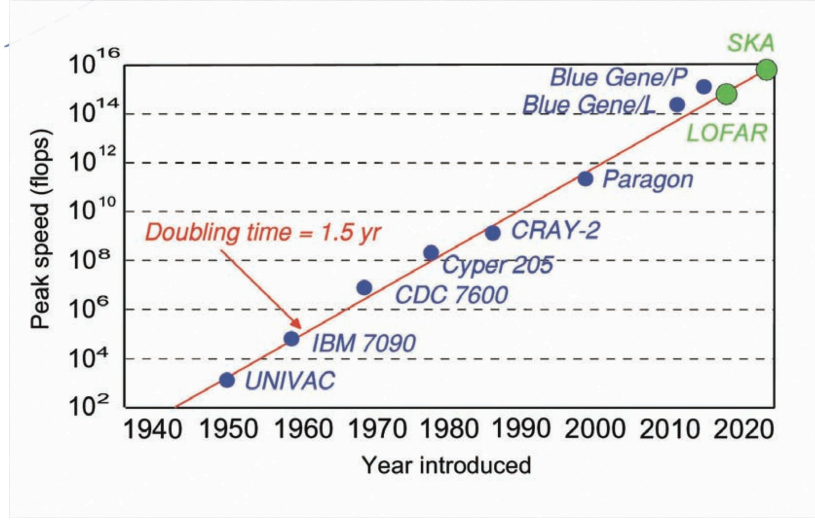
1%

10%

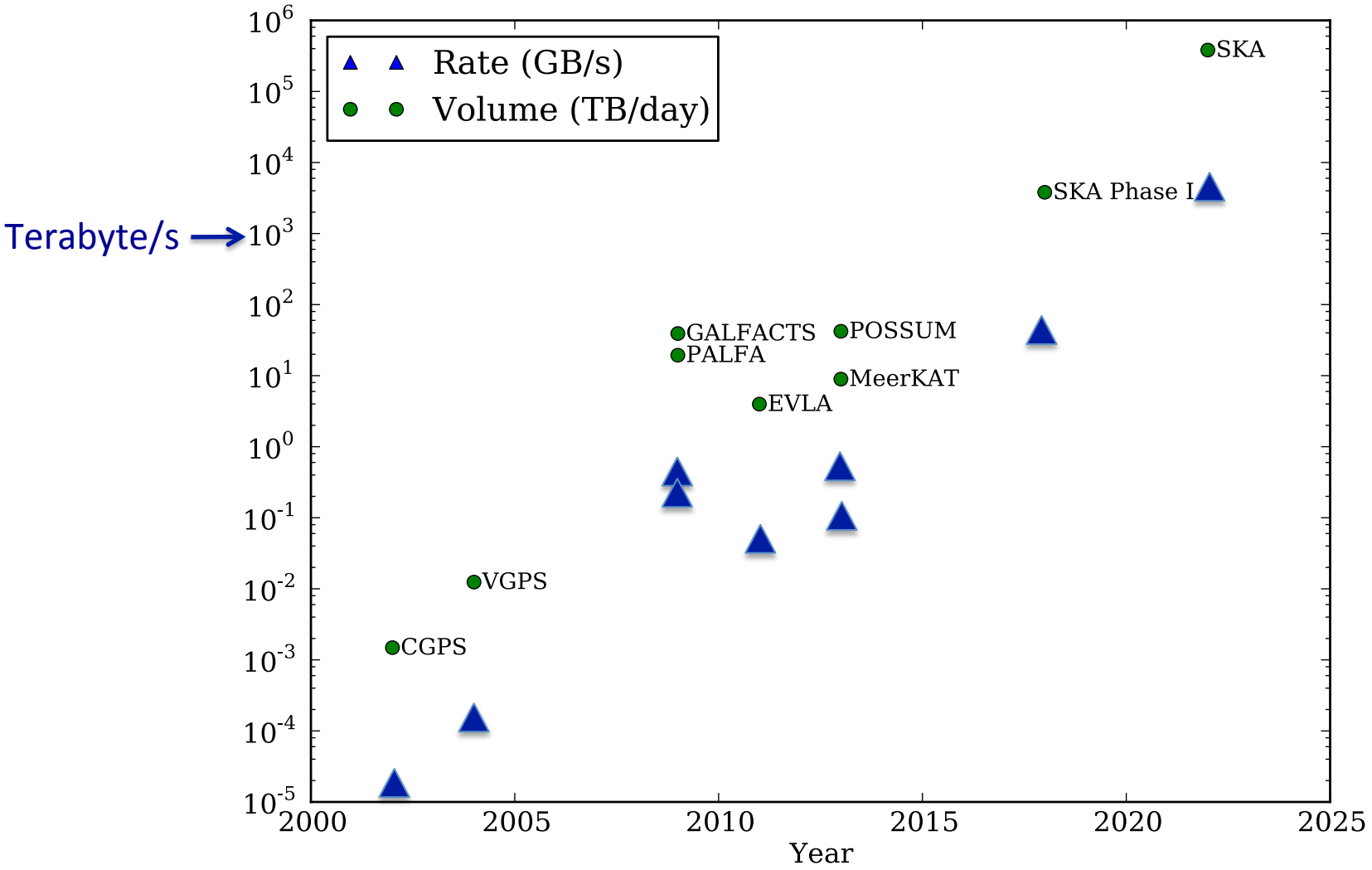
100%

# Data, data, data, informa…..(gulp).



# Survey Raw Data Rates



# The Challenge

- Survey mode observations drive:
  - Very high data rates and volumes
    - Storage, transfer, access
    - Delivery of data to end users not practical
  - Complex, multi-purpose, processing and analysis
    - Processing, analysis, visualization, data mining
    - Multiple processing and analysis chains
  - Collaborative execution by globally distributed teams of researchers
    - Distributed and remote science community
    - Distributed colllaboration in data processing, analysis and science

# How to handle the Big Data Challenge

- Dedicated observatory pipelines to
  - create science ready output at the observatory and send finished products to users
  - Being adopted by LOFAR, EVLA, ASKAP,…
- Cons
  - Monolothic system under control of central “authority” of experts
  - One-pass through data.  No room for iteration and improvement
  - Disconnects the user community from intellectual development processing techniques and technologies

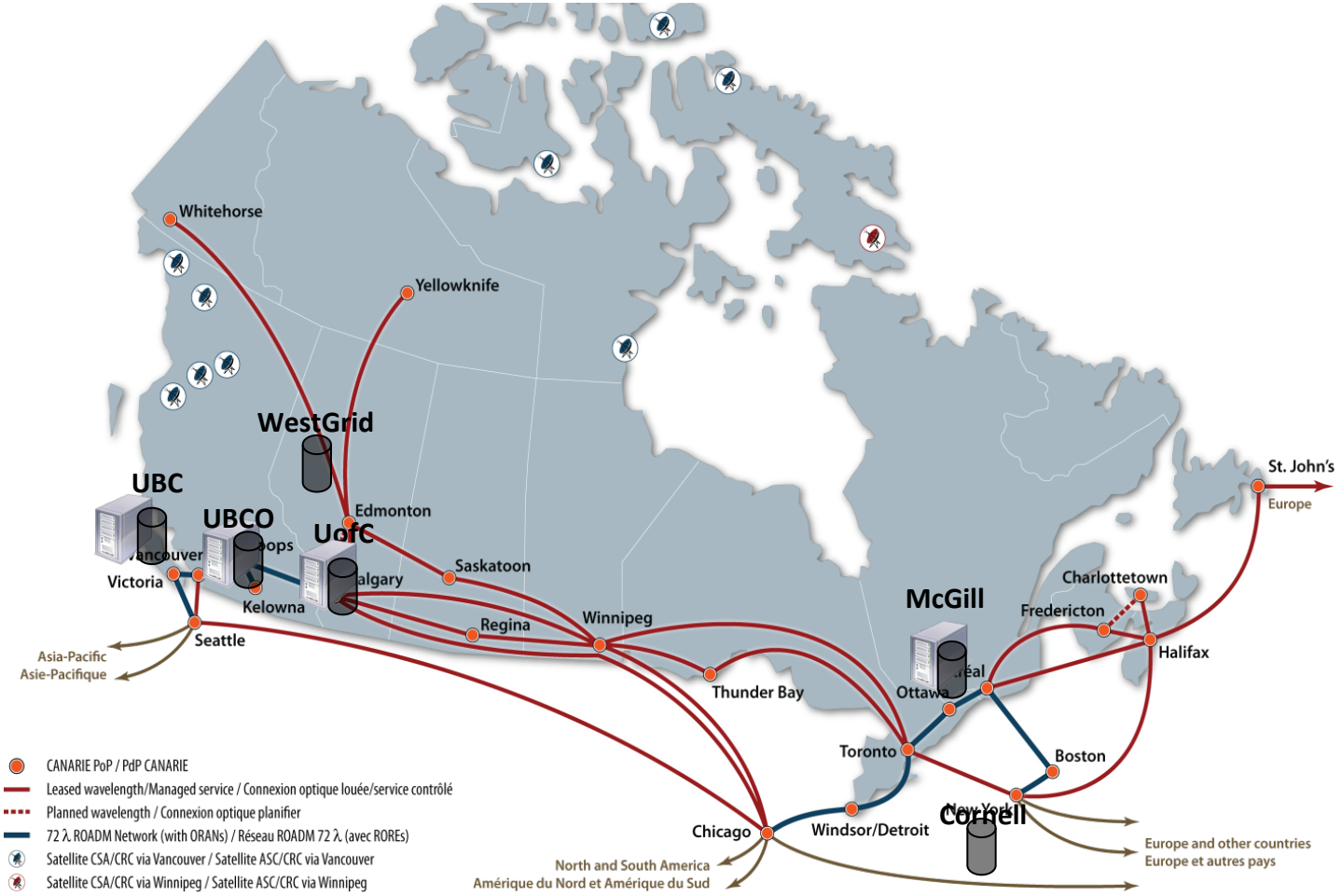
# How to handle the Big Data Challenge

## An alternate approach: cyberinfrastructure platform

- Use cloud and web 2.0 technologies to empower the end user
- Turn global resources into the solution
  - On-line interactive access to a global HPC cloud and data
  - Collaborative development of novel pipelines processes
  - Visualization and visual analytics of very large data
  - Creation and ingestion of derived and ancilliary data products
  - E-science tools for distributed collaboration and analysis

# The CyberSKA Project

Initiative to develop a scalable and distributed cyberinfrastructure platform to meet evolving needs of data-intensive radio astronomy en route to the SKA



# CyberSKA Development Partners





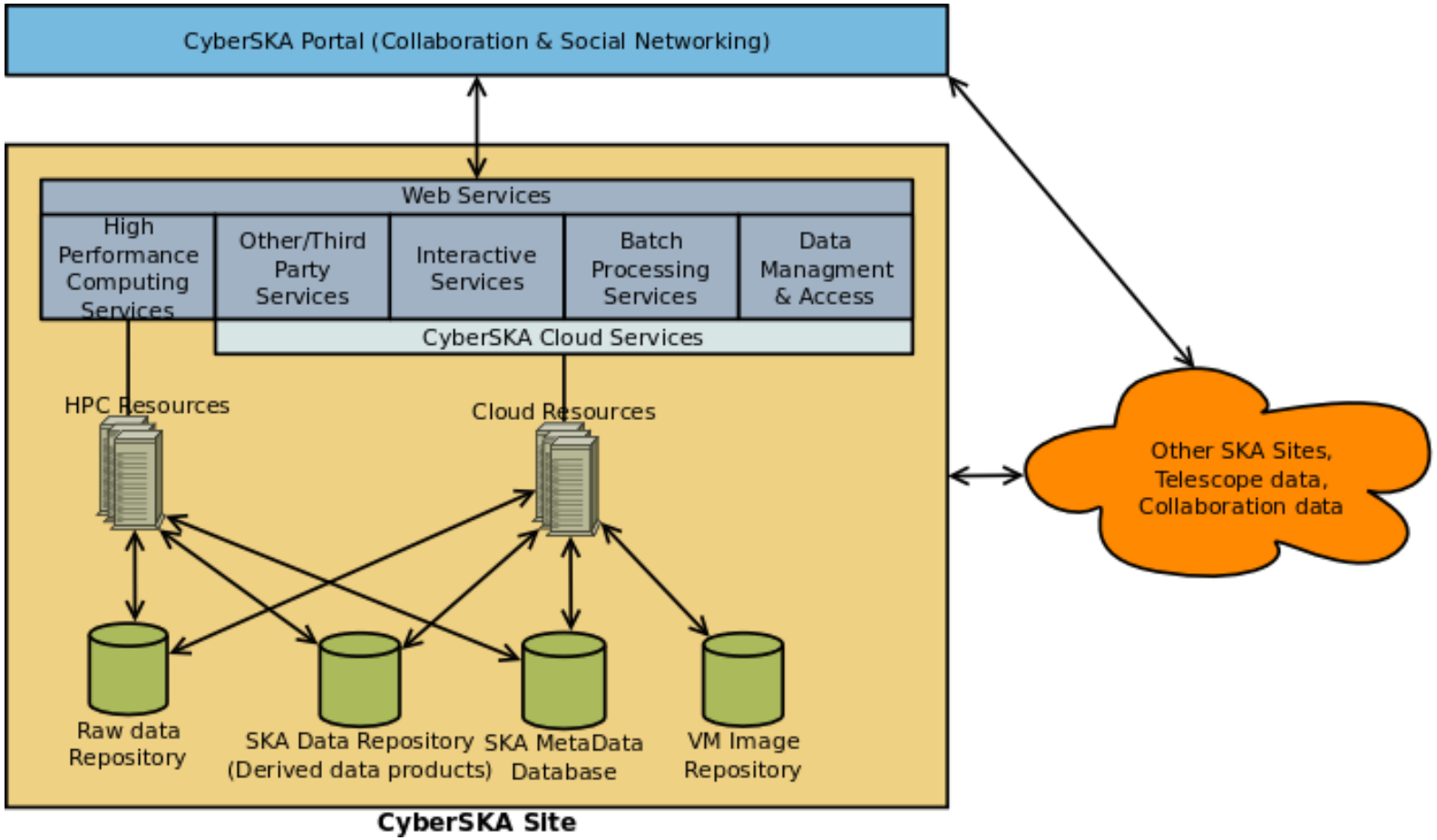




# Focus Areas

- Collaboration
  - Portal built on social networking technologies
- Data Management
  - Scalable collaborative access, sharing and searching of distributed (BIG) data sets
- Data Processing
  - Framework for executing algorithms and workflows
- Data Visualization and Visual analytics
  - On-line interactive visualization of remote Big Data
- Third Party Applications
  - Community driven site with common API

# CyberSKA High Level Architecture

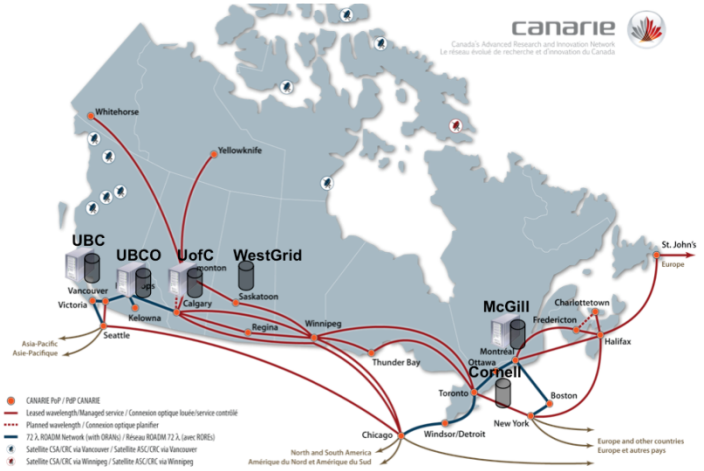


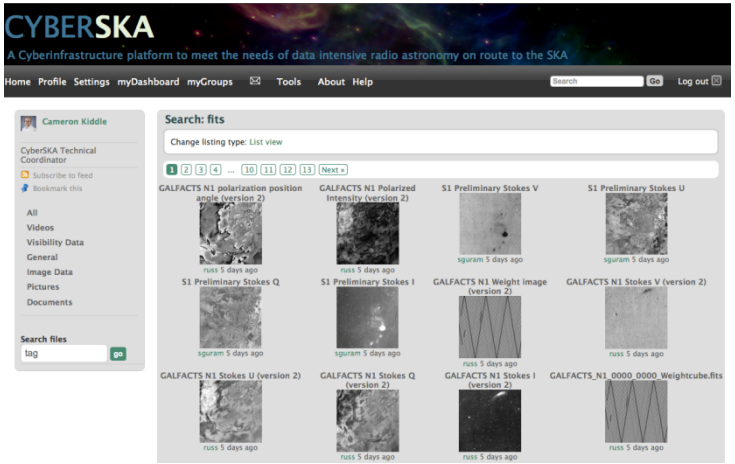
# Collaborative Portal

- Portal built on top of the Elgg open source social networking platform
  - Provides many features including: tags, bookmarks, profiles, blogs, wikis, contacts, groups, document sharing, discussions, messaging, calendars, status, activity feeds

# Distributed Data System

- Based on iRODS (Integrated Rule-Oriented Data System)
  - Abstracts data location
  - Supports data replication / cross-site backup
  - Efficient WAN data transfer
  - Rule engine to automate various tasks
- Upload/download tools
  - Java Applet / Java Web Start based
  - Supports “large” data uploads/downloads
- Automated mime type recognition
  - For many common file types
  - FITS and Measurement Set (CASA) image data or visibility data
- Automated header extraction and thumbnail generation

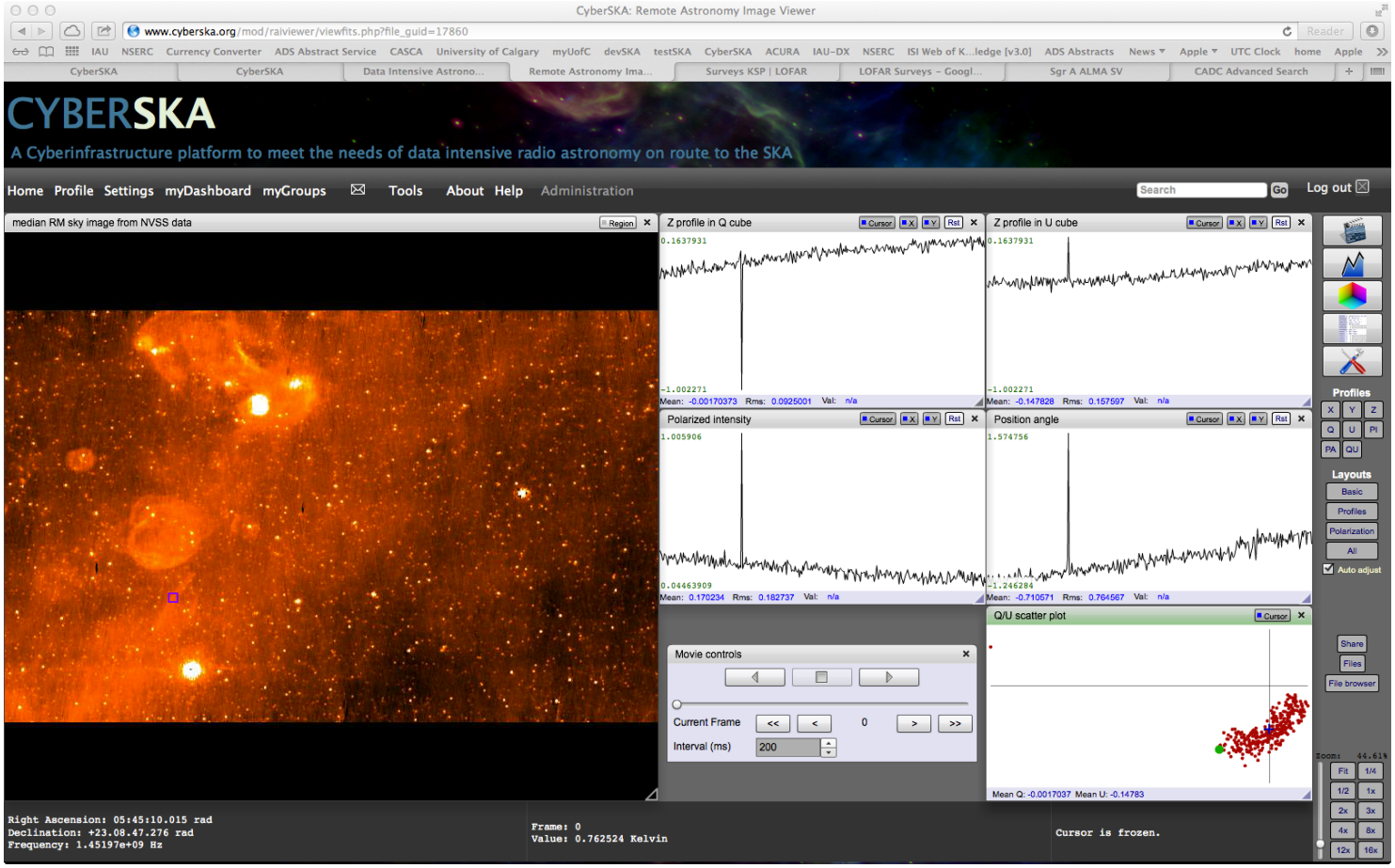


# On-line visualization of Big Data

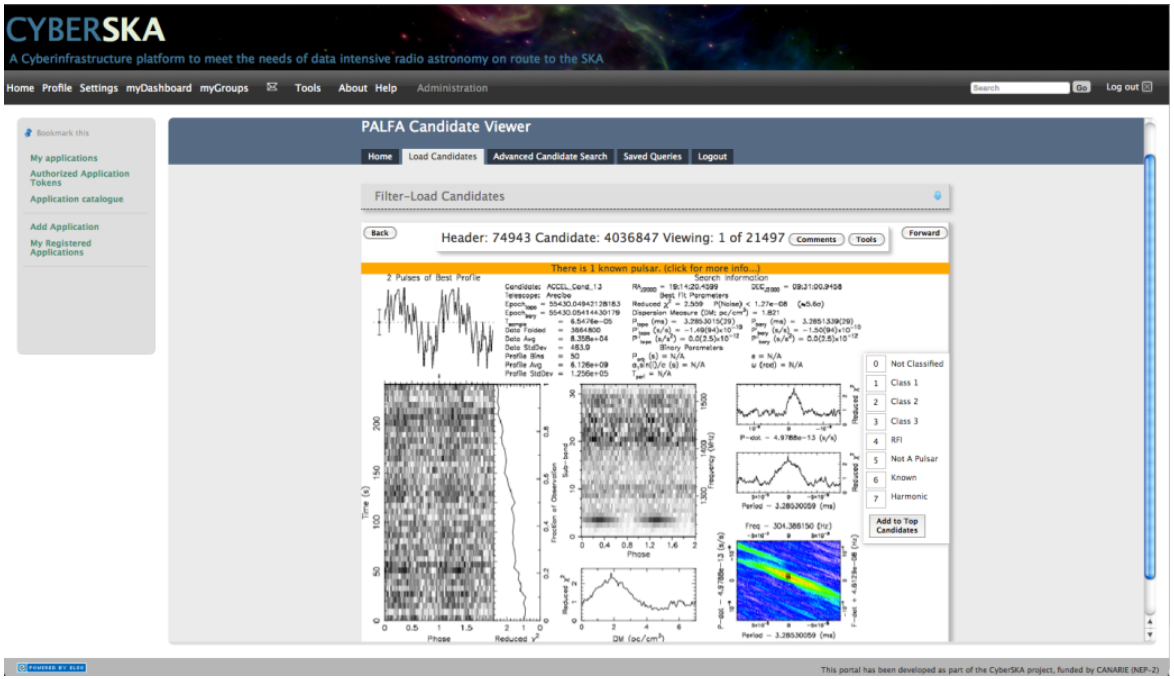
VM based on-line interactive visual analytics of large, multi-dimensional image cubes

- 0.5 TB  full Stokes I, Q, U image cube sets
- Collaboration, screen sharing, etc.

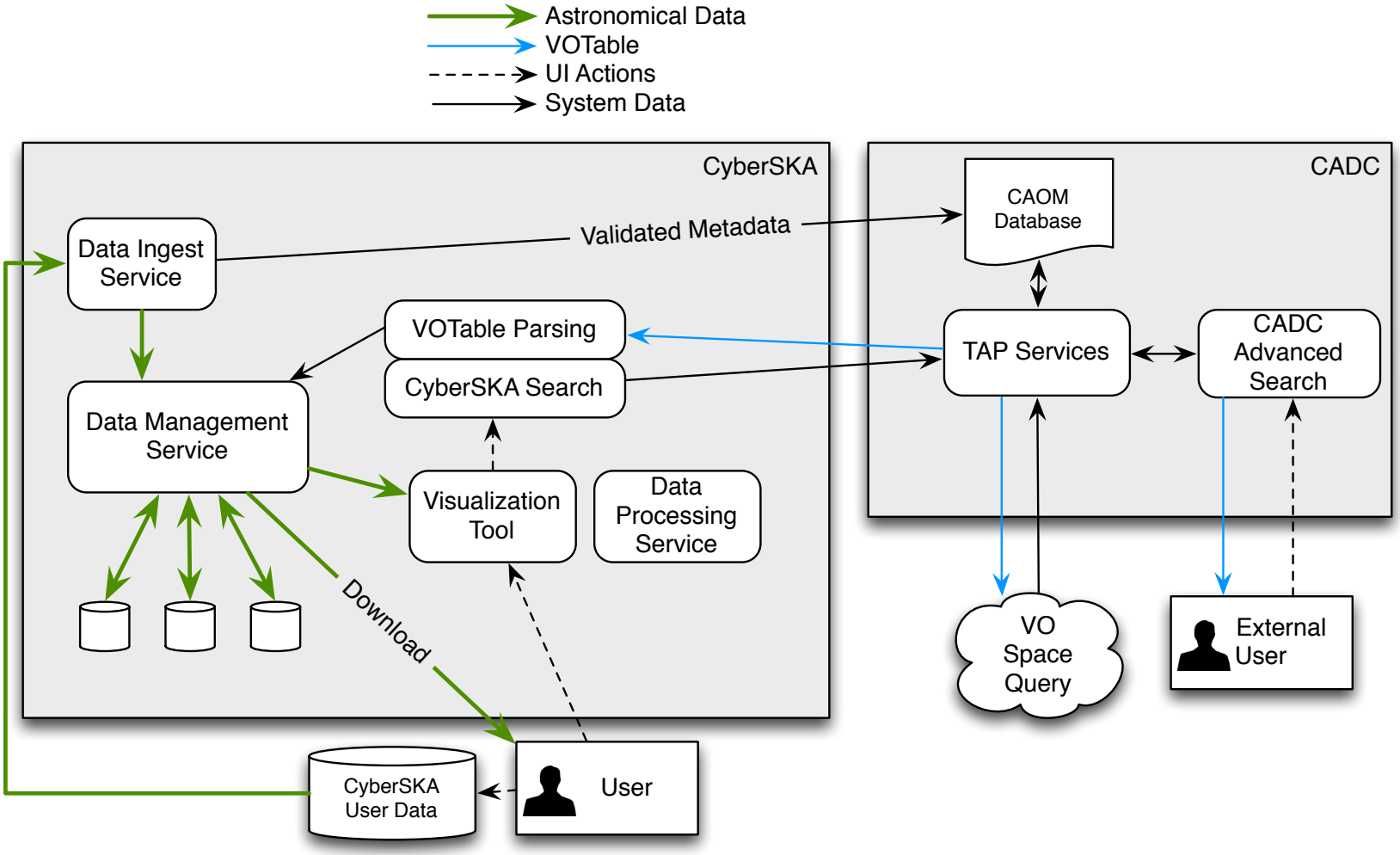


# Third Party Application Interface

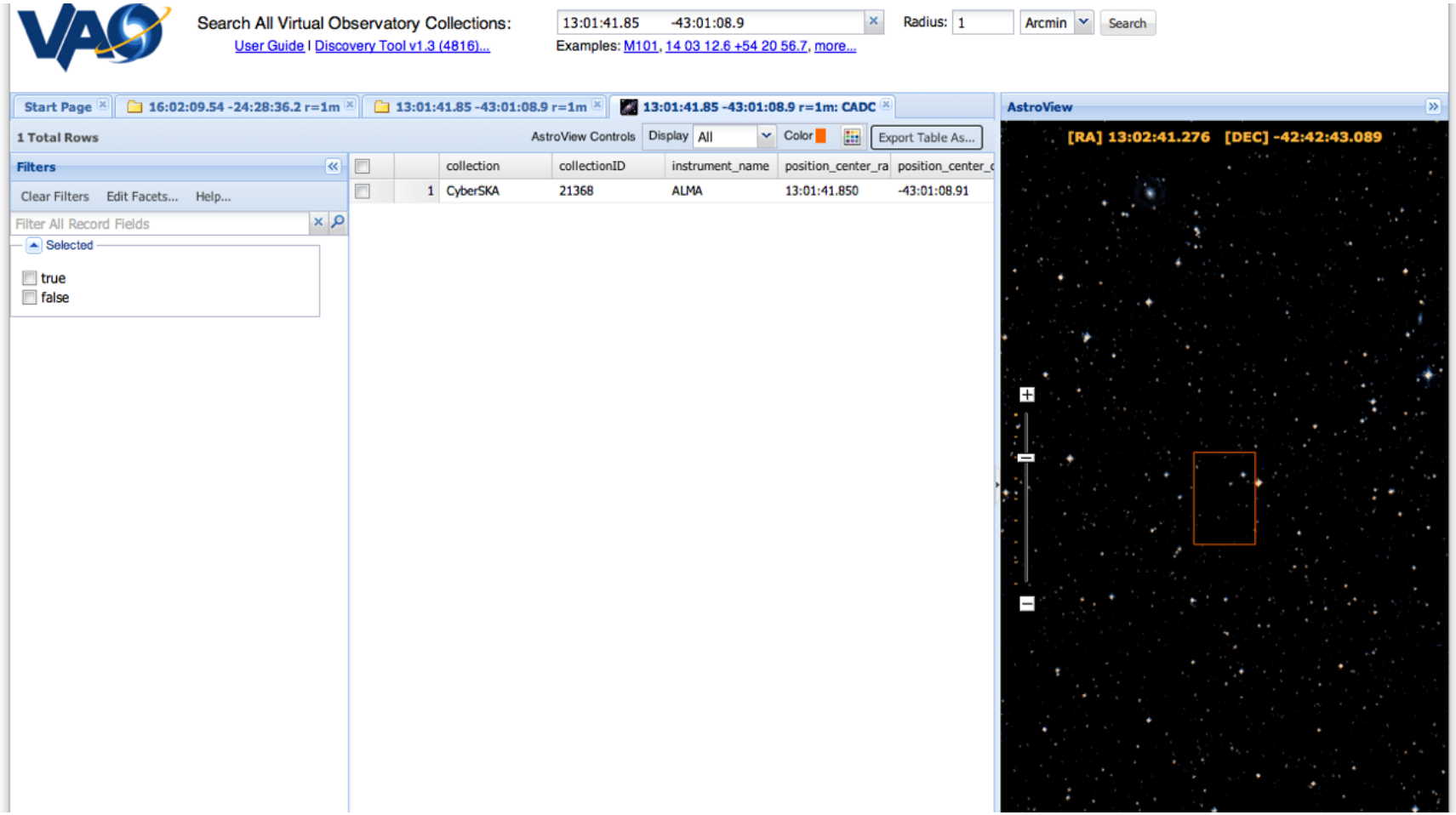
- API for integrating third party / “remote” server applications
- Single sign-on to applications enabled using Oauth
- Push/pull information and data to/from portal
- Current applications include PALFA Candidate Viewer, PALFA Top Candidates, PALFA Observation Scheduler, PALFA Diagnostics Tool, GALFACTS Processing Pipeline, Visibility Data Processing Pipeline, Source Counts, …



# IVOA Data Query Interface



# CyberSKA IVOA Collection



# CyberSKA Usage

- 258 members from around the world
- 40+ “groups” (GALFACTS, PALFA, RM Synthesis, EVLA Deep Polarization Field, GMRT Deep Polarization Field, CASA Users, …)

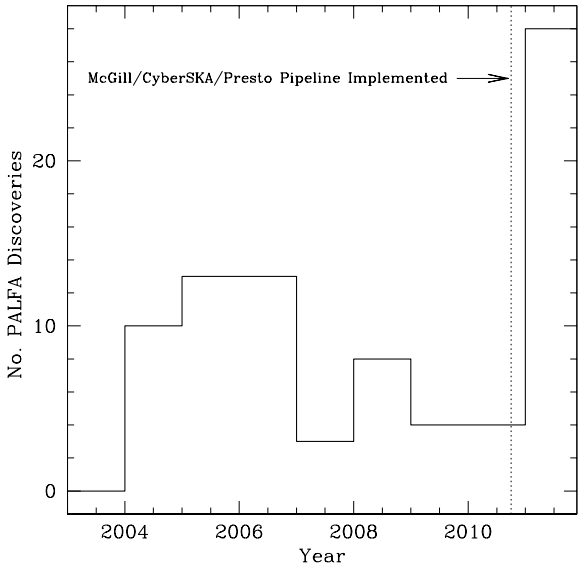
# Imaging Survey Co-development



- GALFACTS (Arecibo) - 49 collaborators on CyberSKA
- Deep Polarization Field Surveys (EVLA, GMRT)
- CyberSKA:
  - Used for sharing documents, creating wiki pages, having discussions and bookmarking resources
  - Enables on-line visualization of remote data sets
  - Access to GALFACTS survey data and third party applications for running data processing pipelines

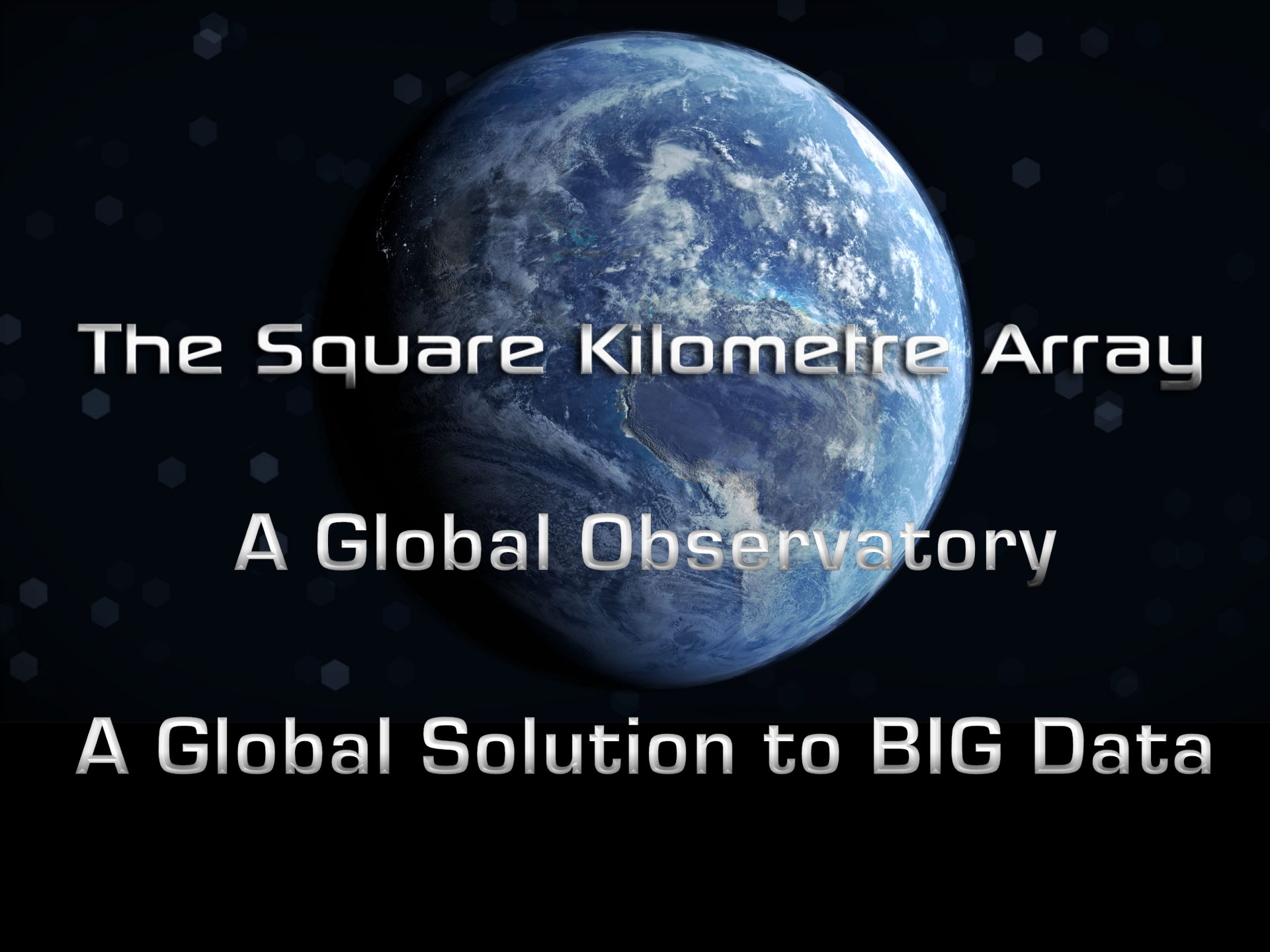
# Pulsar Survey Co-development

- PALFA (Arecibo) – 69 collaborators
- CyberSKA:
  - shared hub for documentation, meeting minutes, publications and task lists
  - Used as an on-line application centre for single sign-on access to a variety of third-party applications
  - Resulted in a significant increase in the rate of discoveries



# On-going and future work

- Study funded by North American ARC in collaboration with Harvard to adapt on-line visual analytics for ALMA on-line data system
- Completion of user interface for data pipeline tool to allow user developed pipeline processing
- Collaboration with CADC on next generation IVOA interface to include visibility data sets (CAOM2)
- Incorporation of distributed HPC-based cloud architecture for scalability to multi-site petascale (Compute Canada, IBM Watson).


The Square Kilometre Array
A Global Observatory
A Global Solution to BIG Data