



# The Canadian CyberSKA Project

A. G. Willis (on behalf of the CyberSKA Project Team)

National Research Council of Canada

Herzberg Institute of Astrophysics

Dominion Radio Astrophysical Observatory

May 24, 2011

# The CyberSKA Project Team



## University of Calgary

- Russ Taylor (Professor, Lead PI)
- Eric Donovan (Associate Professor)
- Robert A. Este (Project Manager)
- Cameron Kiddie (Technical Coordinator)
- Mircea Andreucut (Developer)
- Roger Curry (Developer - Grid Research Centre)
- Pavoï Federi (Developer)
- Arne Grimstrup (Developer)
- Sukhpreet Guram (PhD Student)
- Andrey Mirtchovski (Developer - Grid Research Centre)
- Paolo Pragides (Developer)
- Dina Said (PhD Student)
- Christian Smith (System Administrator)
- Tingxi Tan (Developer - Grid Research Centre)



McGill

## McGill University

- Victoria Kaspi (Professor)
- Rafal Kiodzinski (Developer - Sequence Factory)
- Patrick Lazarus (MSc Student)
- Atallah Mourad (President - Sequence Factory)
- Alex Samoilov (Developer - Sequence Factory)



## University of British Columbia

- Ingrid Stairs (Associate Professor)
- Mark Tan (Developer)



OKANAGAN

## University of British Columbia, Okanagan

- Erik Rosolowsky (Assistant Professor)
- Venkat Mahadevan (Developer)



Cornell University

## Cornell University

- Jim Cordes (Professor)
- Adam Brazier (Research Associate)
- Shami Chatterjee (Research Associate)
- Eric Chen (Analyst Consultant)



## IBM Canada

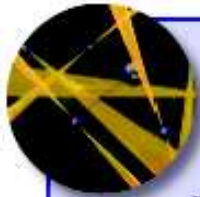
- Don Aldridge (General Manager, Research & Life Sciences)
- Olivier Eymere (IT Architect)



## National Research Council Canada

- Tom Landecker (Principal Research Officer)
- Tony Willis (Senior Research Council Officer)

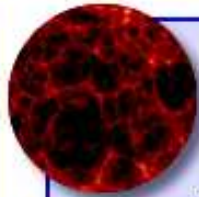
- SKA Overview
- GALFACTS - example of 'new-style' survey
- CyberSKA
- CANARIE
- CyberSKA Requirements
- CyberSKA Solutions
  - Social Networking
  - Visualization
  - Data Management
  - 3rd Party Applications
- Next Steps



## Was Einstein right?

The SKA will investigate the nature of gravity and challenge the theory of general relativity. Pulsars, the collapsed spinning cores of dead stars, will be monitored to study gravitational waves and black holes.

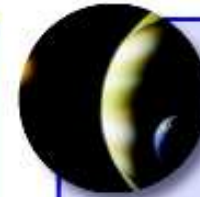
*Picture: D. Champion, M. Kramer*



## How were the first black holes and stars formed?

The SKA will look back to the Dark Ages, a time before the Universe lit up. It will provide detailed pictures of the cosmic web of the neutral gas to discover how the very first black holes and stars formed.

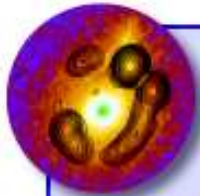
*Picture: S. Furlanetto*



## Are we alone?

The SKA will be able to detect extremely weak extraterrestrial signals and may even spot other planets capable of supporting life. Astrobiologists will use the SKA to search for amino acids, the building blocks of life, by identifying spectral lines at specific frequencies.

*Picture: NASA*



## How do galaxies evolve and what is dark energy?

Mysterious dark energy is thought to cause the increasing rate of expansion of the Universe. The SKA will investigate this expansion by mapping the cosmic distribution of hydrogen. This map will track young galaxies and help to identify the nature of dark energy.

*Picture: HST/STScI, MERLIN*



## What generates giant magnetic fields in space?

By measuring the radio emissions of millions of distant galaxies, the SKA will create three-dimensional maps of cosmic magnets throughout the Universe and reveal their role in its evolution.

*Picture: Hubble Heritage/NASA/STScI, R. Beck/MPfR*

# SKA Technical Requirements



## 1 The ambition and scale of the SKA projects demand innovative technical developments.

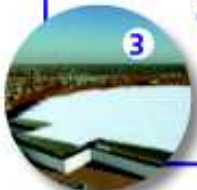
The SKA will use 3,000 dishes, each about 15 m wide. Two other types of receptor, known as aperture arrays, will also be used to observe very large areas of the sky simultaneously. The antennas will cover the frequency range 70 MHz to 10 GHz (4 m to 3 cm wavelength).

### Technical developments

- Phased array antenna technologies.
- Renewable energy generation and distribution options.
- Wideband optic fibre signal transport.
- Data storage and innovative retrieval.
- Fast, high-resolution analogue to digital converters.
- Software development.
- High-performance computing engines.

### Spin off technologies

- Radio technology for use in satellite communications and navigation systems.
- High-tech electronics for use in security systems, medical equipment and automotive applications.
- Materials research applicable to the aerospace industry.
- IT systems for remote regions worldwide.



*Picture 1: Dishes   Picture 2: Sparse aperture arrays   Picture 3: Dense aperture arrays*

# Obligatory SKA Site Simulation

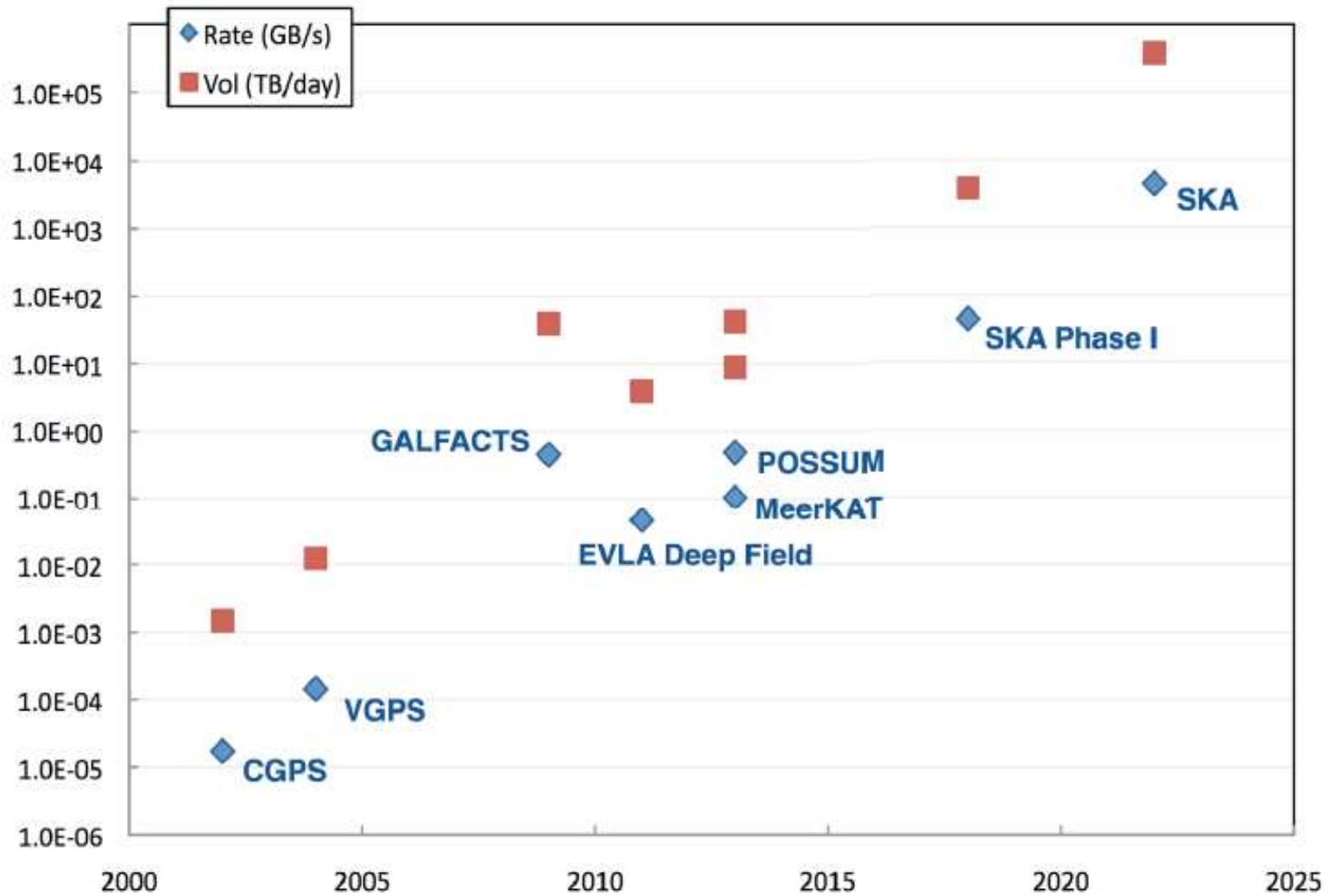
- Who's driving that vehicle? RTS? PED?



# Motivation for CyberSKA

- Most SKA key science goals will be achieved via large-scale survey type observing programs
  - Very high data rates and volumes
  - Complex multi-purpose processing and analysis
  - Executed by globally distributed teams of researchers
  
- Drives the need for cyber-infrastructure solutions for
  - Collaboration tools
  - Data storage, management and distribution methods
  - Distributed data processing, analysis and visualization

# Radio Imaging Survey Data Rates

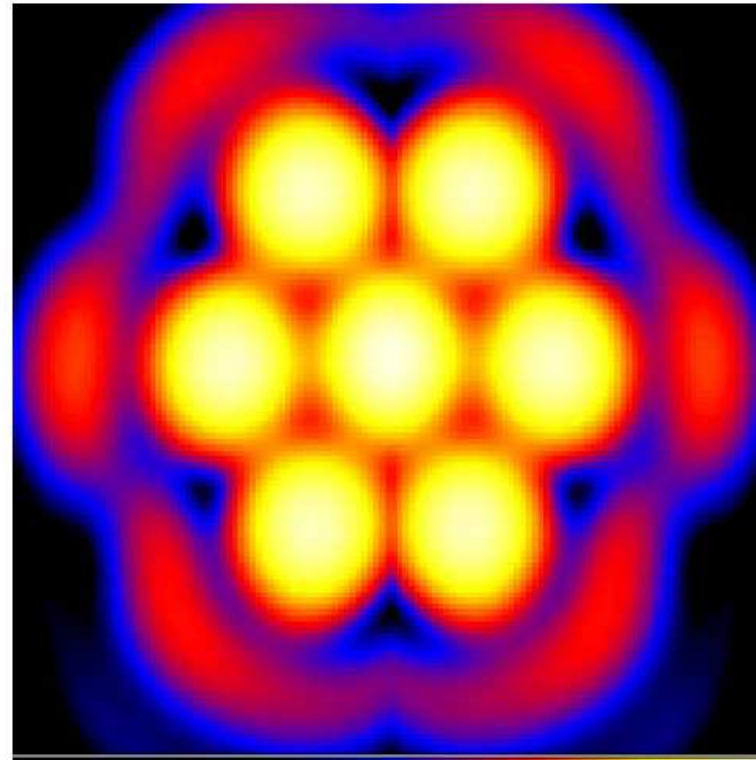
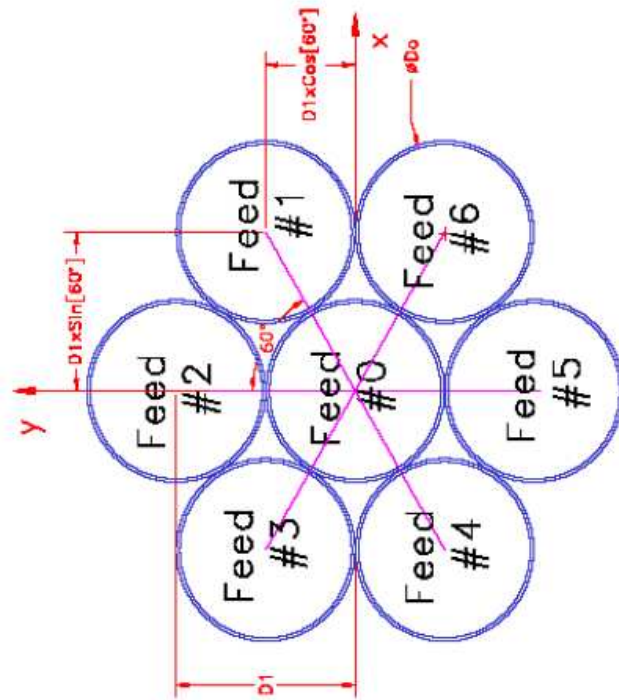




# GALFACTS: The G-ALFA Continuum Transit Survey

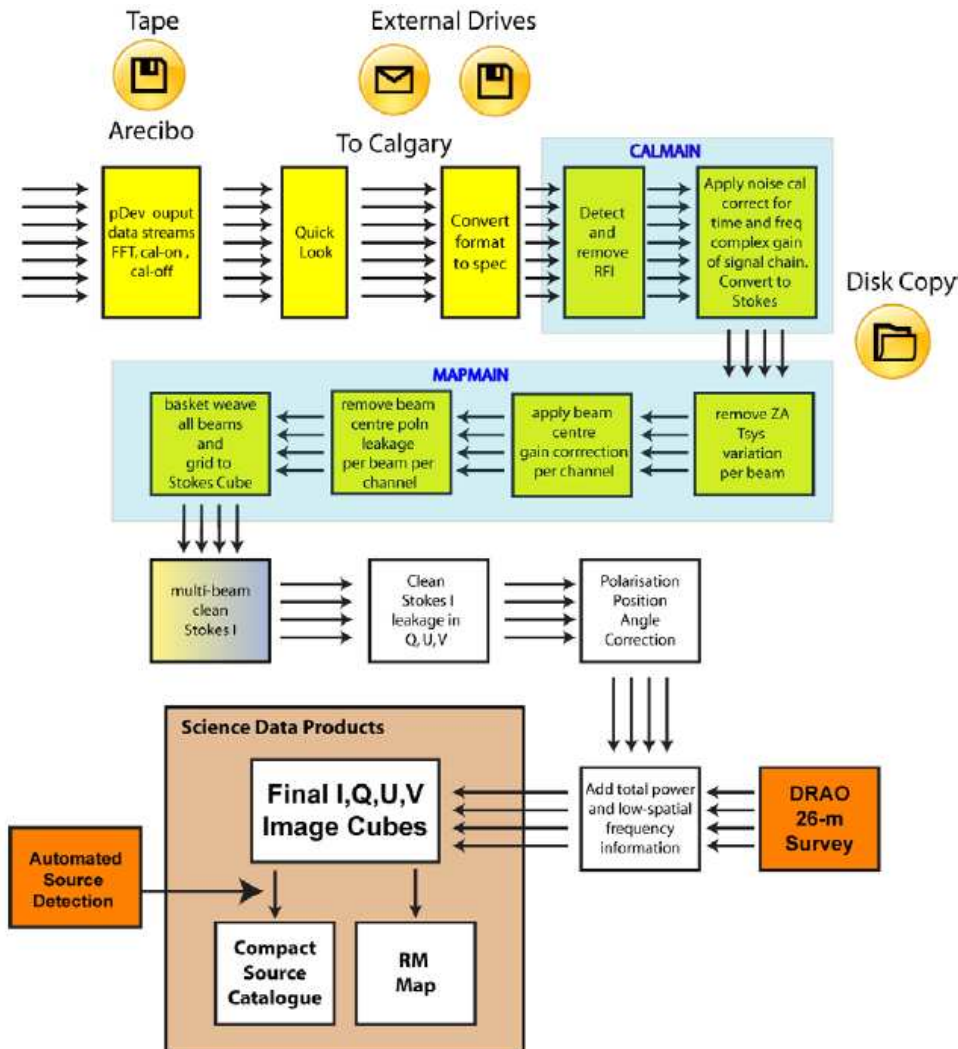
- GALFACTS - example of a 'new-style' survey with data that you cannot reduce on your laptop
- GALFACTS Data Rate
  - 7 beams, 2 bands, 4 Stokes, 4098 channels per band gives 460 MB / sec
    - 6.5 hrs per night gives 10.5 TB
  - Near real-time processing at Arecibo
    - high time resolution, low spectral resolution (HTLS) 1.5 TB / day
    - low time resolution, high spectral resolution (LTHS) 53 GB / day
    - these data sets transferred to University of Calgary
  - For 28 night observing session
    - HTLS 40 TB
    - LTHS 1.5 TB
  - Total observing time for project - 1800 hours
    - correlator produces 2.9 PB
    - 250 TB transferred to Calgary

# GALFACTS Beam Pattern





# GALFACTS Data processing Pipeline



Issues:  
 processing speed  
 data I/O  
 algorithms and software

- Data base design
- Computing resources
- parallelization

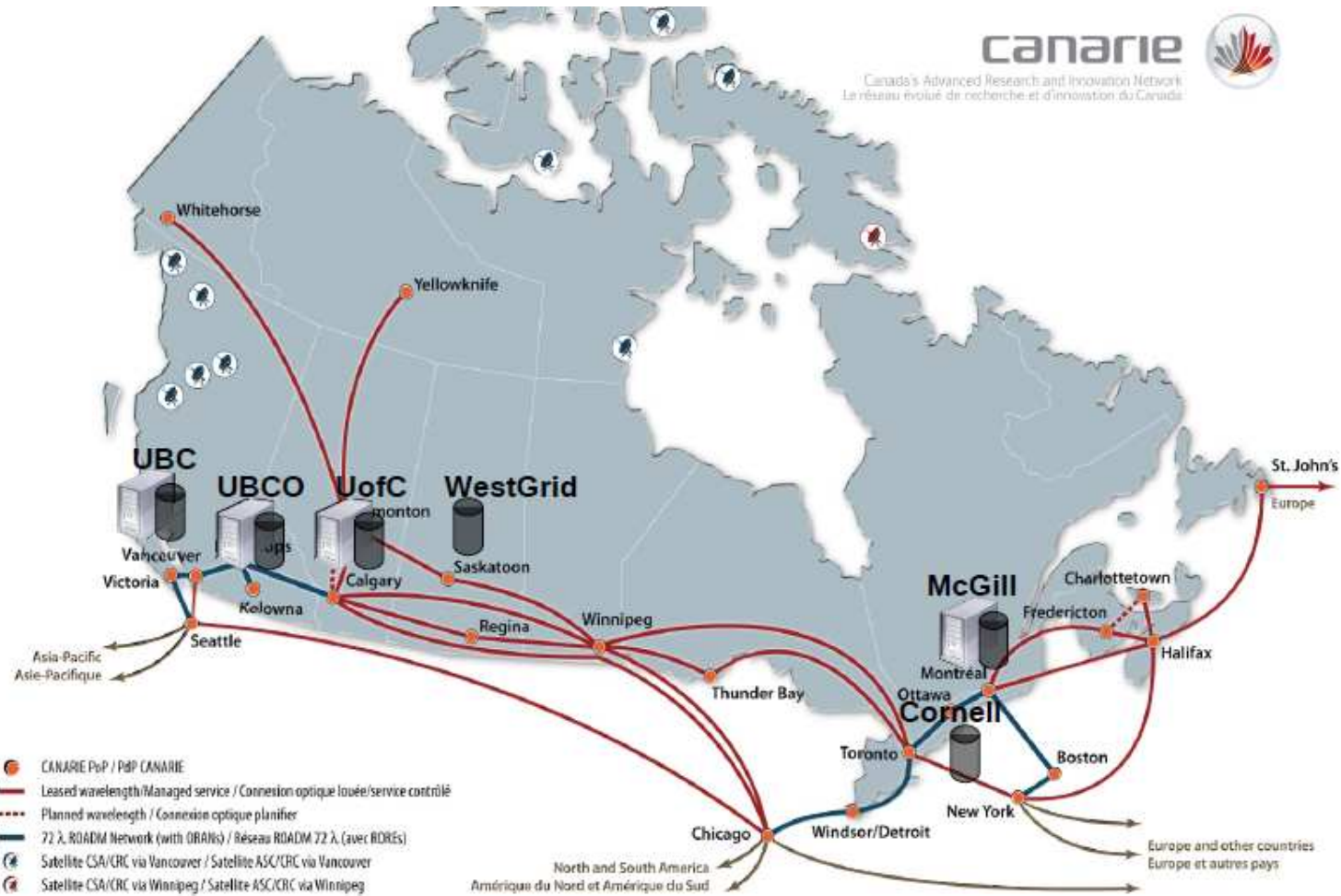
- An initiative to develop a scalable and distributed infrastructure platform to meet evolving science needs of the SKA
- Led by the University of Calgary (Russ Taylor - project lead) with several partner institutions (currently) from North America
- Canadian funding for CyberSka provided by CANARIE as part of their Network Enabled Platforms (NEP) program, and Cybera
- NEP funding two Canadian Astronomy-related programs
  - CyberSKA - led by University of Calgary
  - CANFAR (Canadian Advanced Network for Astronomical Research) - led by University of Victoria
- Cybera - Alberta Cyberinfrastructure for Innovation
- Start by establishing cyberinfrastructure to support current large-scale astrophysical data needs generated by GALFACTS, PALFA and other high data volume SKA Pathfinder projects.

- CANARIE - Canada's Advanced Research and Innovation Network
- 98% of CANARIE's funding goes toward improving the effectiveness of research in Canada
  - Network capacity improvements and new services
  - Programs to simplify researcher access
  - Support for provincial partner networks
- major funding of its programs and activities provided by the Government of Canada
- Annual cost about 25 million dollars
- Underpins \$3.5 billion spent per year on research in Canadian universities and government labs
- 10 billion bits per second across the core network
- 100 billion bits per second in key corridors

# Network-enabled Platforms (NEP)

- This program provides funding for the ICT infrastructure needs of each research community and provides for the development of such things as:
  - Web portals aggregating large data sets
  - Sophisticated software tools for modelling and visualization
  - Sophisticated software tools enabling collaboration
  
- Goals
  - Accelerate development and implementation of research platforms
  - Facilitate collaboration
  - Increase International Connectedness
  
- 20 NEP research domains including Transportation, High Energy Physics, Ocean Science, Space Science, Health Science

# CANARIE and CyberSKA Sites





# CyberSKA Experience/Background

- Leverage knowledge and experience of the Grid Research Centre at the University of Calgary, IBM, and a large technical team
- Adapt, customize and extend technologies used by GeoChronos (<http://geochromos.org>) - another CANARIE NEP funded project
  - Platform developed by the Grid Research Centre
  - Enables Earth observation scientists to access and share data and applications and collaborate more effectively.
  - Employs social networking, cloud computing and data management technologies
- Make use of other existing tools and technologies where possible



# Requirements for CyberSKA Platform

## ■ Distributed and transparent

- Provide transparent access to distributed data, computing resources and services

## ■ Scalable

- Must scale to support increasing data and processing needs

## ■ Deployable

- Different sites should be able to deploy developed tools and participate in CyberSKA relatively easily.

## ■ Heterogeneous

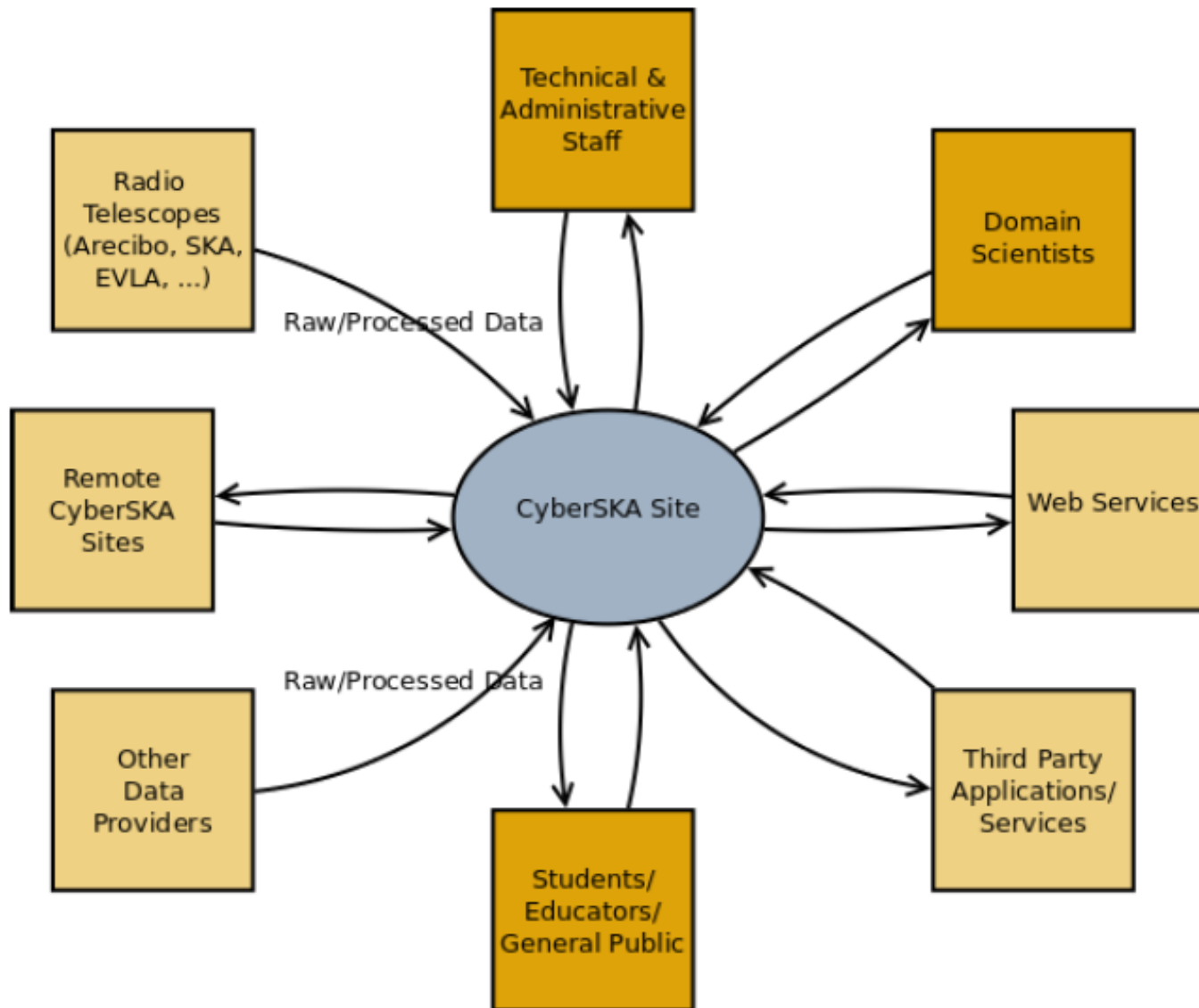
- Provide a framework to enable interaction with different types of data, computing resources and services and to add/execute different processing algorithms and workflows.

## ■ Automated

- Automation and dynamic reconfiguration of services and data workflows in response to user demand, changing user objectives, available data and resource availability

- Web-enabled
  - Web-based platform that users can access from anywhere with Internet access
- Collaborative
  - Enable international/distributed teams to collaborate and communicate effectively
- Interactive
  - Enable on-line interactive visualization of data
- Auditable
  - Be able to track where data has come from and processes applied to it (data provenance)
- Interoperable
  - Compliant with existing standards such as the Virtual Observatory (VOE)

# System Context Model



## ■ Radio Telescopes (Arecibo, EVLA, ASKAP, SKA)

- Raw telescope data, monitoring data, control messages and commands
- Owner - Telescope providers

## ■ Remote CyberSKA Sites

- Raw and processed data transferred between sites, user access, virtual machines, system services, collaboration services
- Owner - Cyber SKA community

## ■ Other Data Providers

- Content not defined yet. CyberSKA will provide a series of APIs and utilities to allow for integration of other data providers
- Owner - various sources

## ■ Web Services

- Method calls to execute defined services
- Owner - CyberSKA community

# System Context Model III

## ■ technical and administrative staff

- Applications, services, documents, Web pages, profiles, discussions, messages, publications, events and many other resources
- Owner - Cyber SKA community

## ■ Domain scientists (astronomers, physicists)

- Raw and processed data, documents, Web pages, profiles, discussions, messages, publications, events, and many other resources
- Owner - individual researchers and teams

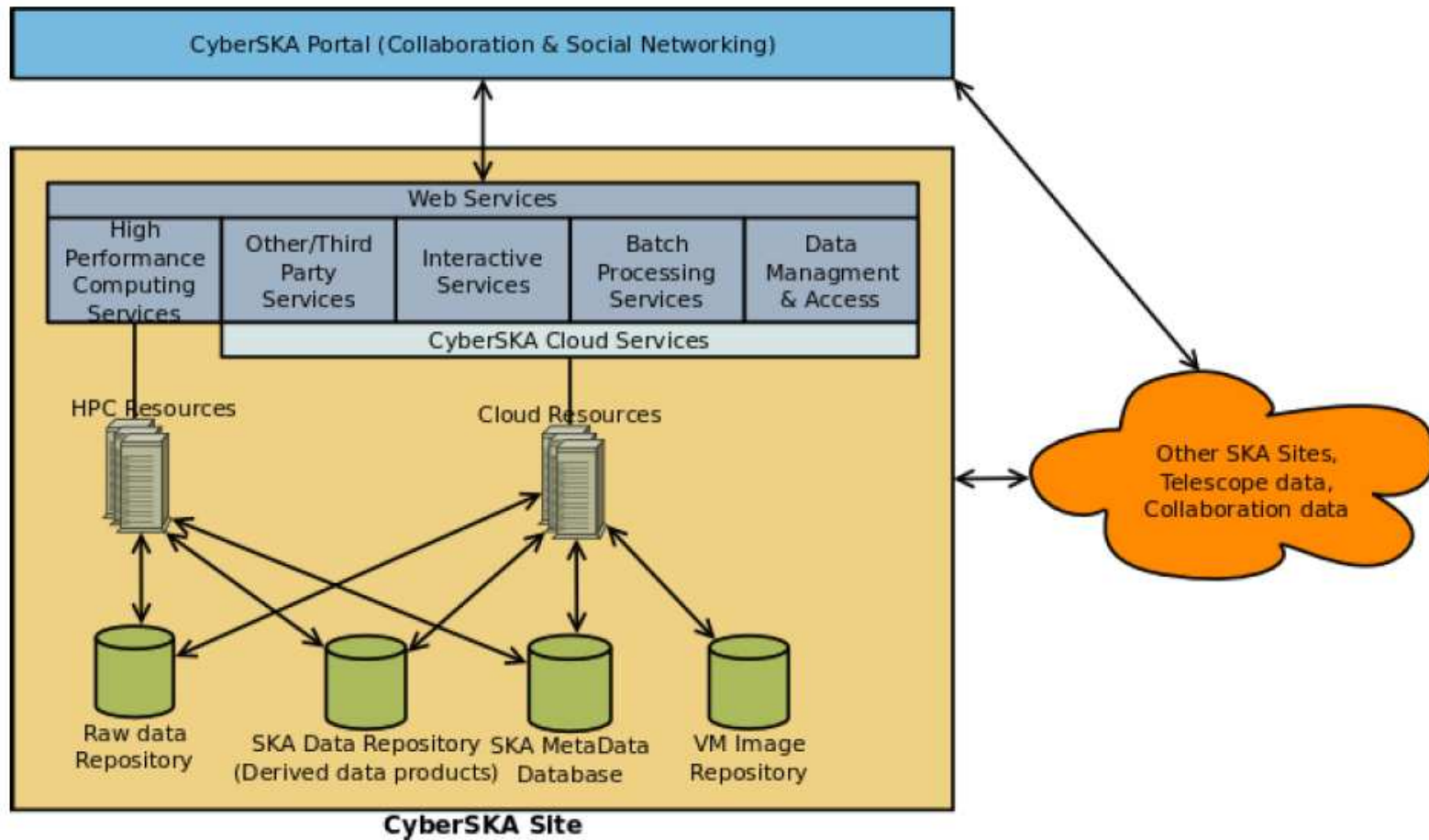
## ■ Third Party Applications / Services

- Links and interfaces to tools and applications provided outside of the standard CyberSKA site. Applications may be hosted outside of CyberSKA site or may be hosted on CyberSKA resources. These applications are maintained and managed separately from CyberSKA regardless of where they are stored.
- Owner - various sources

## ■ Educators, students and general public

- Information, crowd sourcing (identification of pulsars and extragalactic radio sources)
- Owner - individuals and schools

# High Level Architecture



# High Level Architecture II

- The core of CyberSKA is cloud based. Virtual machines are created and removed based on user and application needs
- A site may also have high performance computing or other specialized services that are not as well suited to vitalization.
- Collaboration and social networking are deployed outside of the core CyberSKA sites.
  - This allows greater flexibility and ease in adding new sites while providing a single portal to access all of CyberSKA
- Access to the CyberSKA data and functionality is primarily through the web services layer.
  - A common services definition allows new sites to join CyberSKA relatively easily while providing a common experience to all users



# Solution - Use Social Networking

- Can enhance collaboration capabilities around data and applications
  - Facebook for Scientists
  
- Facebook analogy
  - Platform dealing with large scale in terms of users, data and applications
    - more than 500 millions users, of whom about 50% log on to Facebook on any given day
    - more than 30 billion pieces of content shared each month
    - more than 550 thousand applications on Facebook platform

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

The screenshot shows the CyberSKA portal dashboard, which is built on the Elgg open source social networking platform. The page features a navigation bar with links for Home, Profile, Settings, myDashboard, myGroups, Tools, About, and Help. The main content area is divided into several sections:

- Left Sidebar:** A user profile for Cameron Kiddle with options to subscribe to feeds, bookmark this, and view various activity feeds (Your files, Your pages, Your bookmarks, Your blogs, Your tasks, Your event calendar, Your activity, Contact's activity, Site activity).
- Contacts:** A grid of user profile pictures.
- Event calendar:** A list of upcoming events:
  - Innovations in Data-Intensive Astronomy:** A workshop to encourage new ideas for the effective processing, analysis, and interpretation of Tera- to Peta-scale data sets and to promote the establishment of collaborations to develop those ideas. 2 May 2011 - 6 May 2011.
  - TERENA Networking Conference 2011:** research networking conference - <http://tnc2011.terena.org/> 15 May 2011 - 18 May 2011.
  - Workshop: The Growing Demands on Connectivity and Information Processing in Radio Astronomy from VLBI to the SKA:** The workshop intends to review current R&D trends in radio-astronomical data analysis. 24 May 2011 - 25 May 2011.
- Group membership:** A list of groups:
  - DMS support of Measurement Sets:** This subgroup is for discussing how to support Measurement sets in the CyberSKA Data Management System.
  - CyberSKA Sys Admins:**
  - Application Developers:** Group for developers working/creating portal applications.
  - Portal Support:** Group for portal support - documentation, comments and forums.
- Recent Astro-ph Eprints:** A list of recent eprints with titles and PDF links, such as "Testing the cosmological evolution of magnetic fields in galaxies with the SKA" and "A generalised Measurement Equation and van Cittert-Zernike theorem for wide-field radio astronomical interferometry".
- Activity:** A section for recent activity, including a "Contacts" sub-section with recent uploads and file sharing.
- Active Users:** A section for active users.

CyberSKA: CASA Pipeline

Home Profile Settings myDashboard myGroups Tools About Help Search Go Log out

### CASA Pipeline

All those using the CASA Pipeline

- Subscribe to feed
- Bookmark this


[Leave group](#)  
[Email group members](#)

[Group discussion](#)

[Group file folders](#)

[Group pages](#)  
[Group bookmarks](#)  
[Group files](#)  
[Group blog](#)  
[Group calendar](#)  
[Group tasks](#)  
[Group applications](#)

**Group members**



[View more members](#)

**Description:**  
This group includes the developers and testers of the CyberSKA CASA pipeline for the time being.

**Tags:** casa, work flow tool, pipeline

**Website:**

**Membership Criteria:**

**Parent Group:** CASA Users

**Owner:** Shannon Jaeger  
Group members: 6

#### Group pages

- Weekly Reports**  
Last updated 6 days ago by Shannon Jaeger (CASA Pipeline)
- ASKAP beam files**  
Last updated 69 days ago by Shannon Jaeger (CASA Pipeline)

[New page](#) [More pages](#)

#### Group Applications

- CASA Pipeline (Run Application)**  
Registered by Shannon Jaeger
- CASA Pipeline - Test Application**  
Registered by Shannon Jaeger

[More applications](#)

#### Group bookmarks

This group does not have any bookmarks yet

[Add bookmark](#)

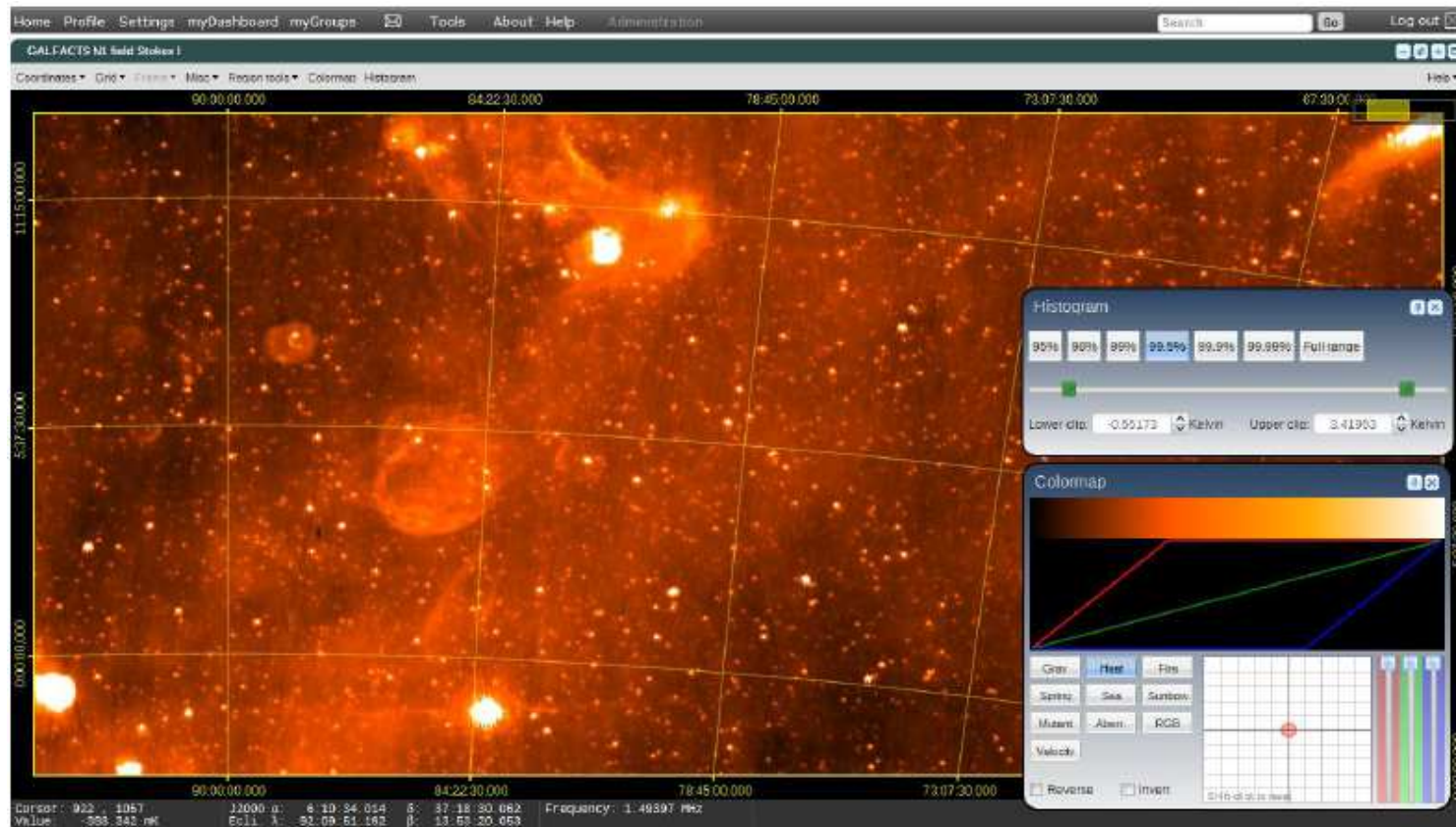
#### Group activity

- Shannon Jaeger updated a page titled Weekly Reports (6 days ago)
- Shannon Jaeger updated a page titled Weekly

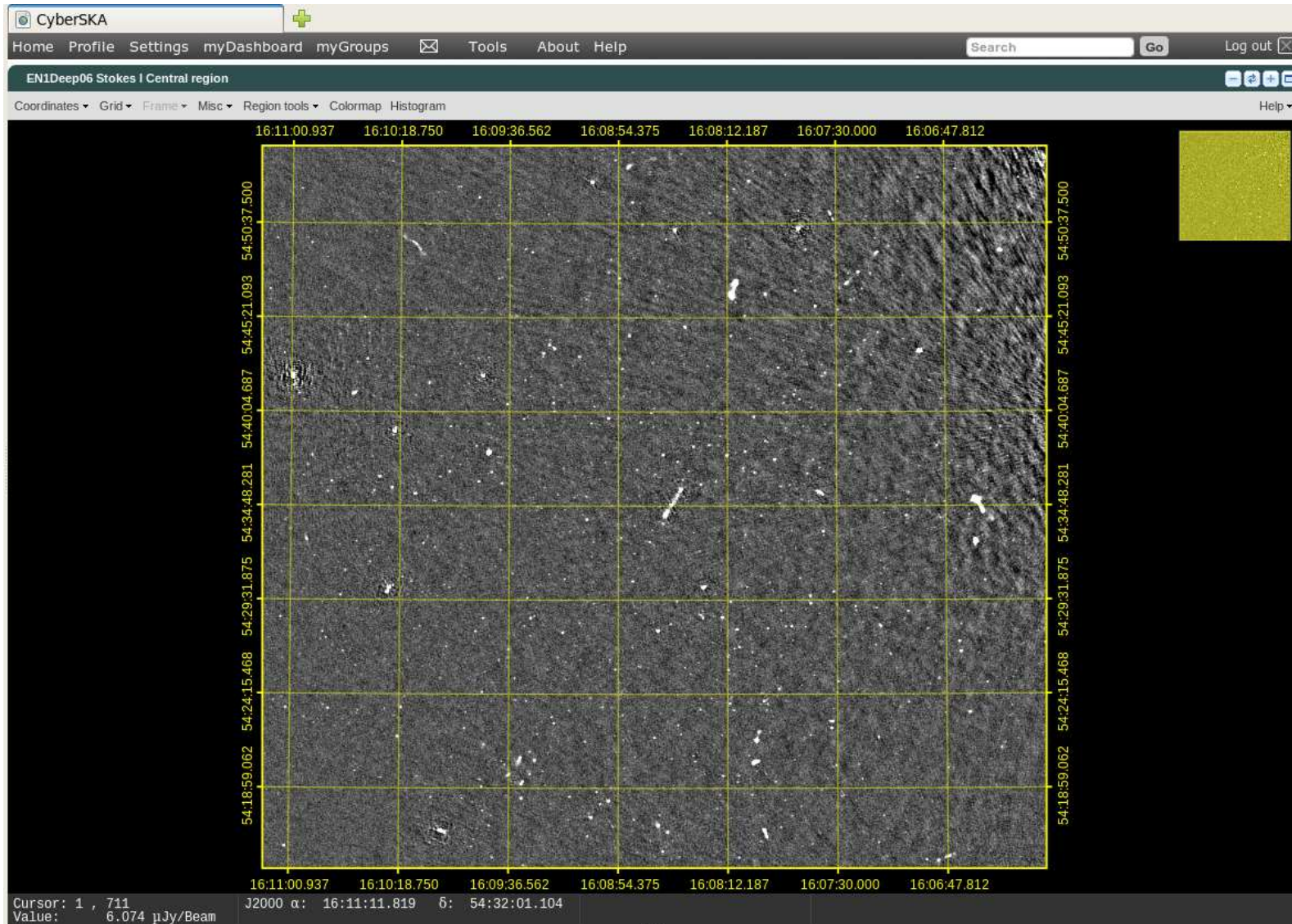
#### Group files

- FTFS** ASKAP Simulation of Beam 1 - Stokes I,Q,U,V for Channel 0  
CASA Pipeline 73 days ago (CASA Pipeline)
- png created from CASA Viewer**  
Shannon Jaeger 77 days ago (CASA Pipeline)
- MPEG Video of Stokes I - zoomed in**  
CASA Pipeline 77 days ago (CASA Pipeline), Comments (2)

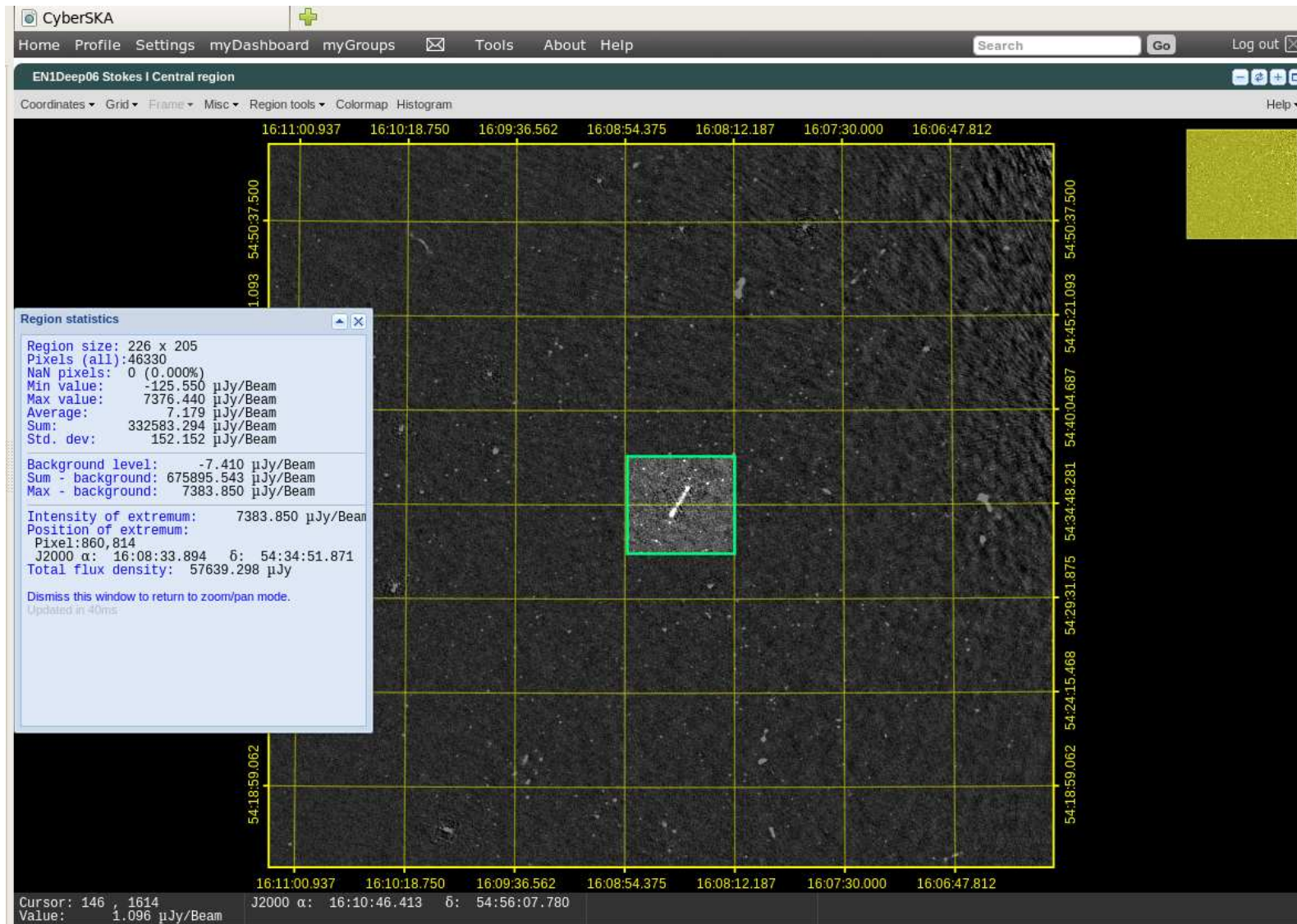
- On-line visualization of multi-dimensional FITS files
  - Supports interactive panning and zooming, histogram correction, colour map adjustments, display of pixel data value, region statistics, multiple coordinate systems, grids, selection of frame for multi-dimensional images, 2D Gaussian fitting, permalink, screenshots



# Visualization



# Visualization

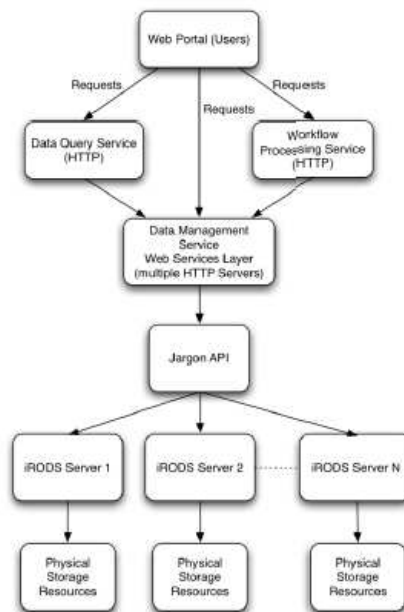


- Access/download data for selected parameters and region of interest
- Requested data generated in virtualized Condor pool on server side

The screenshot displays the GALFACTS Consortium web interface. At the top, there is a navigation bar with links: Home, Profile, Settings, myDashboard, myGroups, Tools, About, Help, a search box, and a Log out button. A left sidebar contains the GALFACTS Consortium logo and a menu with options like 'GALFACTS Consortium Group', 'Bookmark this', 'Create new download', 'Download requests', 'Group pages', 'Group bookmarks', 'Group files', 'Group blog', 'Group calendar', 'Group tasks', 'Group applications', and 'Group data'. The main content area features a large image of a star field with a green rectangular region of interest highlighted. Below the image is a control panel with several sections: 'Coordinates' (Bottom left, Center, Top right) with fields for Xl, Yl, Ol, Bl, Xc, Yc, Oc, Bc, Xr, Yr, Or, Br; 'Archive' (tar, gz, zip); 'Cubes' (I, Q, U, V, W); 'Download information' (Estimated download size: 769.55 MB, Cubes: 2, 912 x 316 x 350 (3500/16)); 'Frequency range' (Start channel #: 0, freq: 1523.374 MHz, End channel #: 3499, freq: 1376.410 MHz); and 'Spectral averaging' (Averaging width  $\Delta\nu$ : 10,  $\Delta f$ : 0.42 MHz). At the bottom, there are 'Submit' and 'Cancel' buttons. A footer at the bottom right states: 'This portal has been developed as part of the CybersKA project, funded by CANARIE (NEP-2)'.

## ■ Distributed data management service

- Built on iRODS (Integrated Rule-Oriented Data System)
- Used PostgreSQL database for image metadata (spatial, temporal, and spectral queries supported)
- Supports mosaicing, plane extraction, compression and staging of images returned by query
- Details in talk by Venkat Mahadevan



The screenshot shows the 'Data Management Service Workflow Process Setup' interface. At the top, there are three buttons: 'Create Pipeline', 'Execute Pipelines', and 'Clear All Pipelines'. Below these are tabs for 'Segment', 'Mask', 'Plane Extract', 'Compress', and 'Stage'. The main area is divided into two columns for 'Pipeline Number: 0' and 'Pipeline Number: 1'. Each column contains a 'file list' section with a 'Add files...' button, followed by a 'pipeline' section with parameters like 'Mask file', 'Plane start', and 'Plane end', and a 'stage' section with a 'Directory prefix' field. A 'Results' panel on the right shows a list of jobs with their status and links to view details.



# Solutions - Applications

- API for integrating third party / remotely hosted applications
- Single sign-on to applications enabled using OAuth

The screenshot displays the PALFA Applications Suite web interface. At the top, there is a navigation bar with links for Home, Profile, Settings, myDashboard, myGroups, Tools, About, Help, and Administration. A search bar and a 'Log out' button are also present. The main content area is titled 'PALFA Applications Suite' and is logged in as 'ita@cybaska.org'. The interface is divided into several sections:

- Applications:** A central section containing four main panels:
  - Candidate Viewer:** Displays a grid of plots including light curves, spectra, and other data visualizations for a selected candidate.
  - Observations Scheduler:** A table for scheduling observations, with columns for Date, Start, Stop, Filter, Location, and Comments. It includes a search function and a 'Batch' button.
  - Top Candidates:** A section for viewing the top candidates, with a 'Statistics' button.
  - Diagnostics:** A section for diagnostic tools, including 'Observation analysis' and 'M1 analysis'.
- Left Sidebar:** Contains navigation options such as 'Bookmark this', 'Authorized Application Tokens', 'My Applications', 'Application catalogue', 'Add Application', and 'My Registered Applications'.

# CyberSKA Portal Usage

- 140+ members from around the world
- 20+ groups - GALFACTS, PALFA, EVLA, GMRT, CASA Users, etc



## ■ Infrastructure

- Set up cloud computing environments and key services at each site

## ■ Collaboration

- Refinement and development of collaboration features based on user feedback

## ■ Data Management

- Expansion of distributed data management system to other sites
- Better integration of data management system with other CyberSka tools and services

## ■ Visualization

- Provide server side support and improve scalability

## ■ Data Processing

- Establish dynamic batch-based processing and interactive service environments on cloud platform
- Establish framework for adding and integrating different processing algorithms and workflows

## ■ Applications

- Extension of third-party application API to enable two-way interaction between portal and applications (i.e. pull data/information from portal, push news feeds to portal based on application activities)

# Contact Information and Acknowledgements

- Portal: <http://www.cyberska.org>
- e-mail: [info@cyberska.org](mailto:info@cyberska.org)
  
- Acknowledgements
  - Russ Taylor - project Principal Investigator
  - Cameron Kiddle - technical coordinator
  - Olivier Eymere - IT Architect (IBM)
  - CyberSKA project team