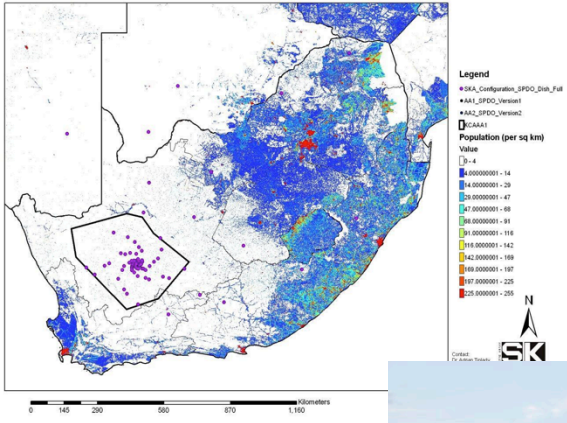
Inter-University Institute for Data Intensive Astronomy: Developments toward a Precursor SKA Regional Science Centre

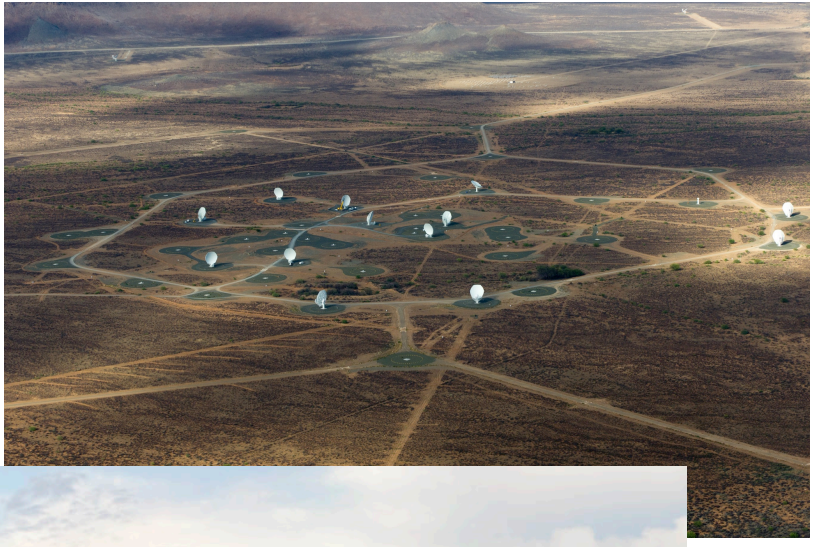
Russ Taylor

SKA-Link Kick-off, April 3 2017

MeerKAT - phase 0 of SKA-mid

MeerKAT Early science end of 2017
Large survey programs begin mid-2018



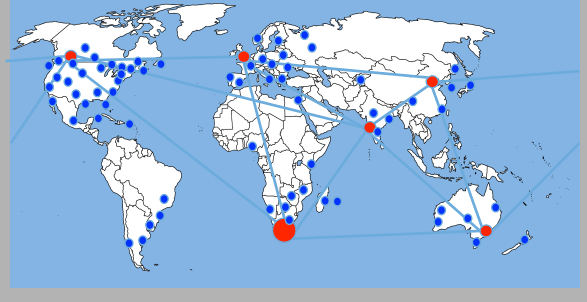


Changing Sociology of Radio Astronomy

- 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

old

new



MeerKAT Large Survey Projects

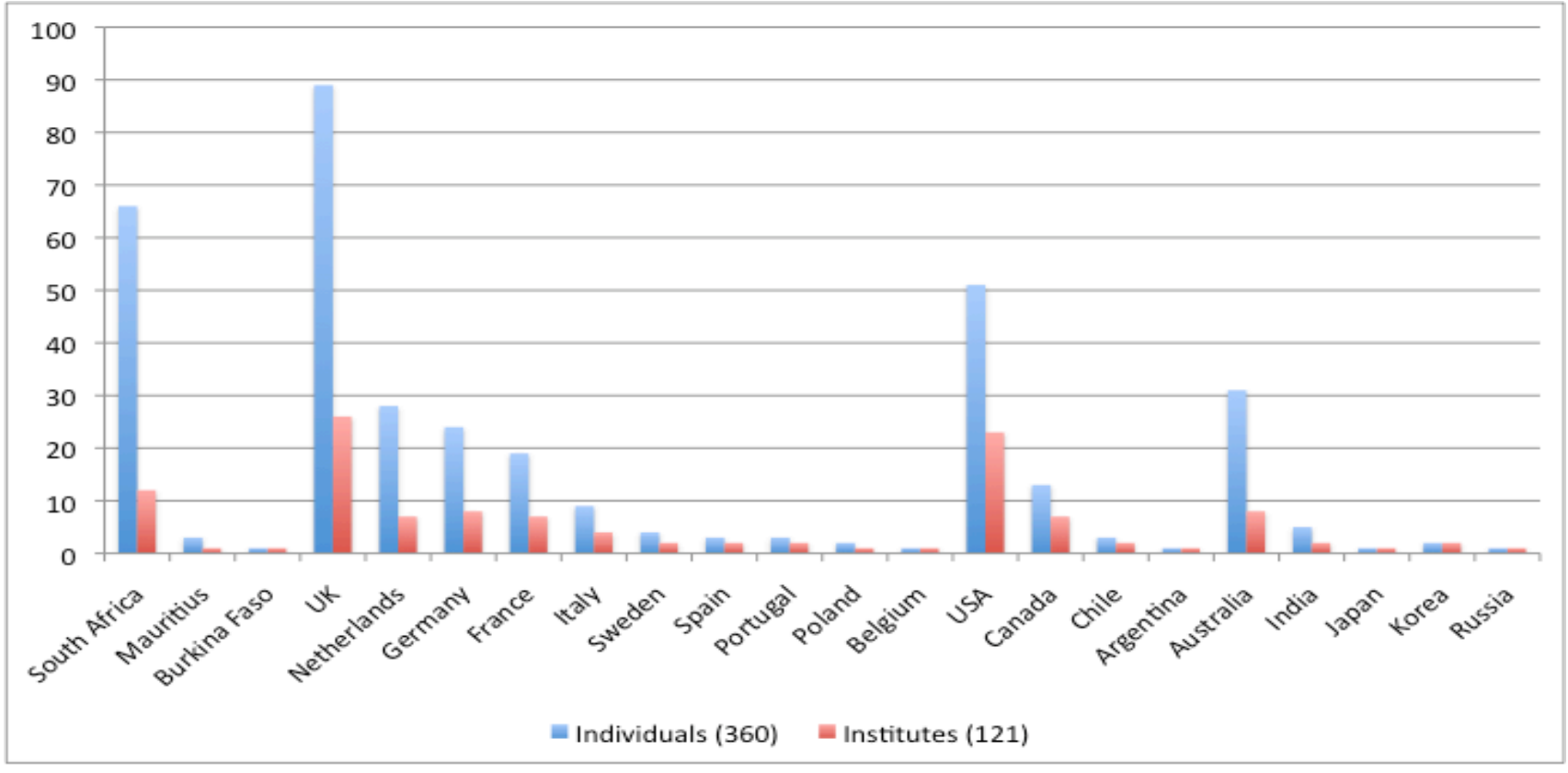
imaging

Time domain

- LADUMA (Deep atomic hydrogen)
- MIGHTEE (Deep continuum imaging of the early universe)
- Fornax (Deep HI Survey of the Fornax cluster)
- MHONGOOSE (targeted nearby galaxies HI)
- MeerKAT Absorption Line Survey (extagalactic HI absorption)
- ThunderKAT (exotic phenomena, variables and transients)
- TRAPUM (pulsar search)
- Pulsar Timing (no acronym)
- MESMER (High-z CO)
- MeerGAL (Galactic Plane Survey)

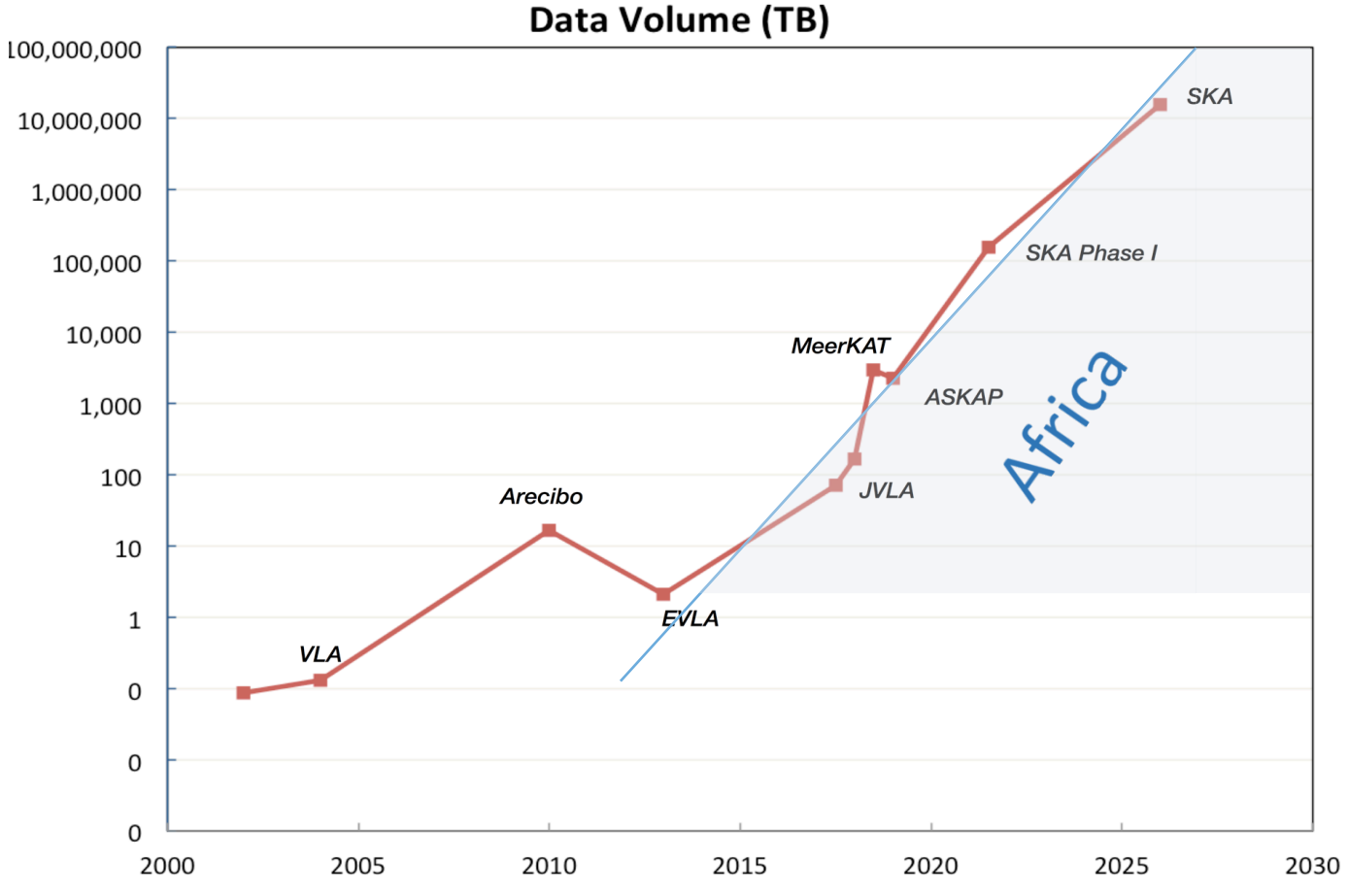
http://public.ska.ac.za/meerkat/meerkat-large-survey-projects

MeerKAT Large Surveys (43,000 hours allocated)



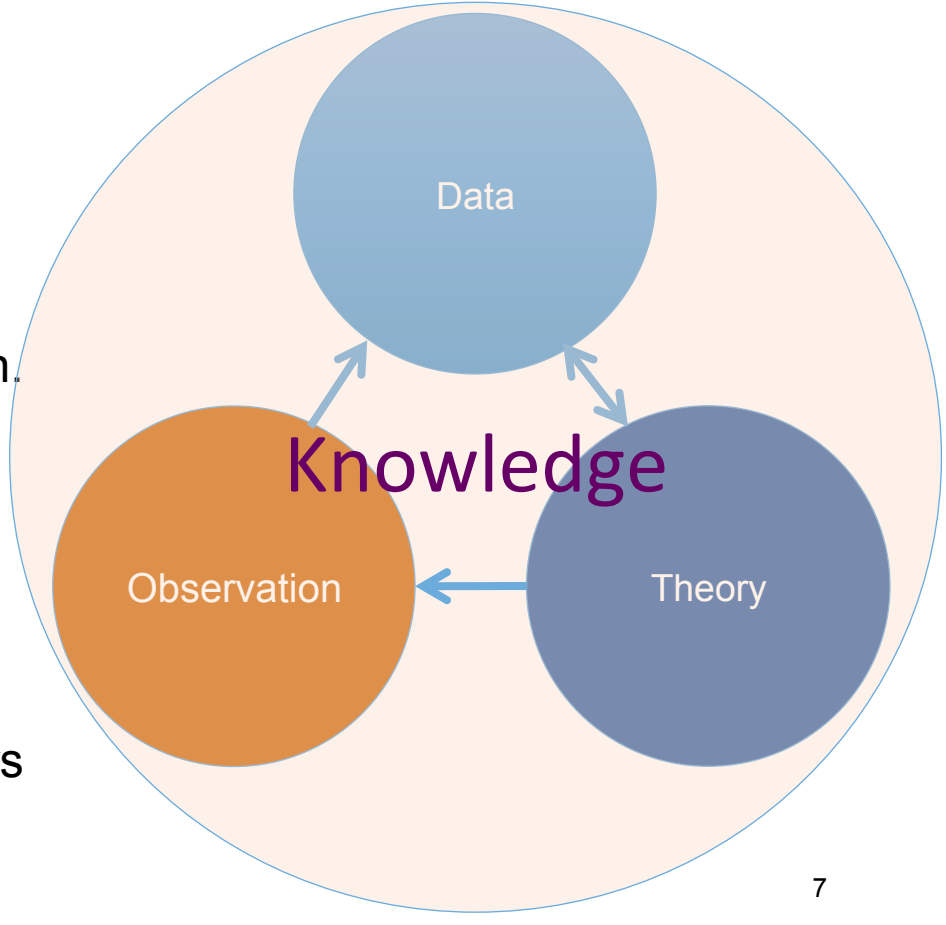
22 countries

Growth of Data Volumes to Radio Astronomers



The Challenges: data to knowledge

- Managing exponential increase in rates and volumes.
- Empowering the end user for multi-purpose processing and analysis to create science data and for post processing, data mining and exploration.
- Data Fusion with big multi-wavelength data and big simulations
- Collaboration, sharing and joint execution of data intensive projects by globally distributed teams of researchers



IDIA Inter-University Institute for Data Intensive Astronomy

from big data to big ideas

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

UNIVERSITY of the WESTERN CAPE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SAP®

SKA Precursor Regional Science Data Centres

MeerKAT

MeerKAT Telescope (SKA-SA)

- generate and manage telescope data
- First Stage processing
  - flagging
  - Near-real time calibration and imaging
  - Data validation

T1 data store
- calibrated and averaged visibilities
- Image repository

Science products

T1 processed

IDIA Tier 2 Facility

- Project-based data extraction from T1 data store
- Processing aggregate data to scientific image data sets
- Post-processing, analytics
- Visualization and data mining
- Platform co-development for global data intensive project collaboration and data sharing.

Global nodes

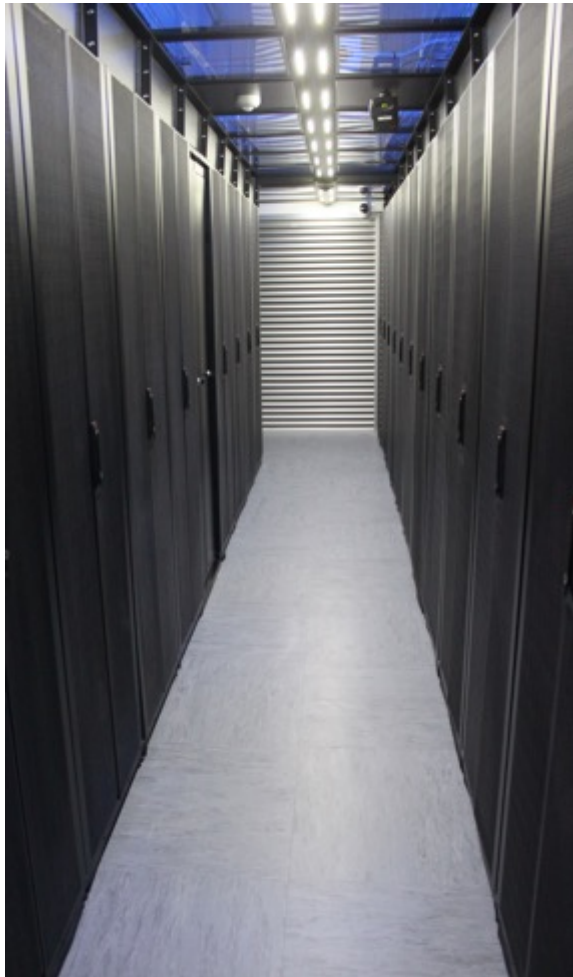
LOFAR

Research and Development Collaboration

DIRISA Tier 2 Data Intensive Research Facility

- University of Cape Town
- University of the Western Cape
- Cape Peninsula University of Technology
- Stellenbosch University
- Sol Plaatje University
- SKA-South Africa

R60M IDIA + CSIR
Astronomy and Bioinformatics



“Dome” Collaboration

- Collaborate on Precursor SKA Regional Science and Data Centre technologies bringing together MeerKAT and LOFAR science
- Partners: ASTRON, IDIA, SKA-SA and IBM
- Two initial components:
  - Data transport for moving data to Tier 2 centres
  - Processing pipelines for executing at Tier 2 centres
- Data transport implement key parts of SDP DELIV data transfer components
  - Request, Prepare and Deliver
  - Query component not currently part of project, but will be added to system in collaboration with CADC

NRAO-IDIA-Taiwan Collaboration

- Data processing for data-intensive radio astronomy projects and for visualization for large radio astronomy data cubes
- Software system used for processing for Jansky Very Large Array and Atacama Large Millimetre Array and MeerKAT

Signed 17 January 2016

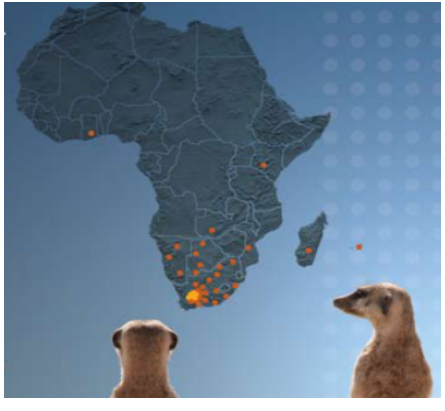
The African Research Cloud Project

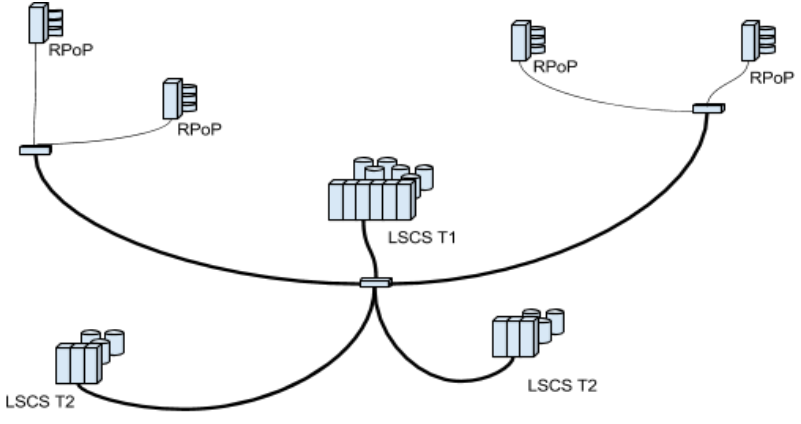
- Deploy federated cloud resources among South African institutions and African SKA partner countries, and link to EU collaborators.
- Tiered distributed e-infrastructure to create a data intensive cyber-infrastructure commons with a mix of system sizes
- Prototype and test bed under development by IDIA and collaborators.

IST-Africa
IST-Africa 2016 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2016
ISBN: 978-1-905824-54-0

The African Data Intensive Research Cloud

Rob SIMMONDS[1], Russ TAYLOR[2,3], Jasper HORRELL[4], Bernie FANAROFF[5], Happy SITHOLE[6], Sakkie JANSE VAN RENSBURG[7], Boeta PRETORIUS[8]

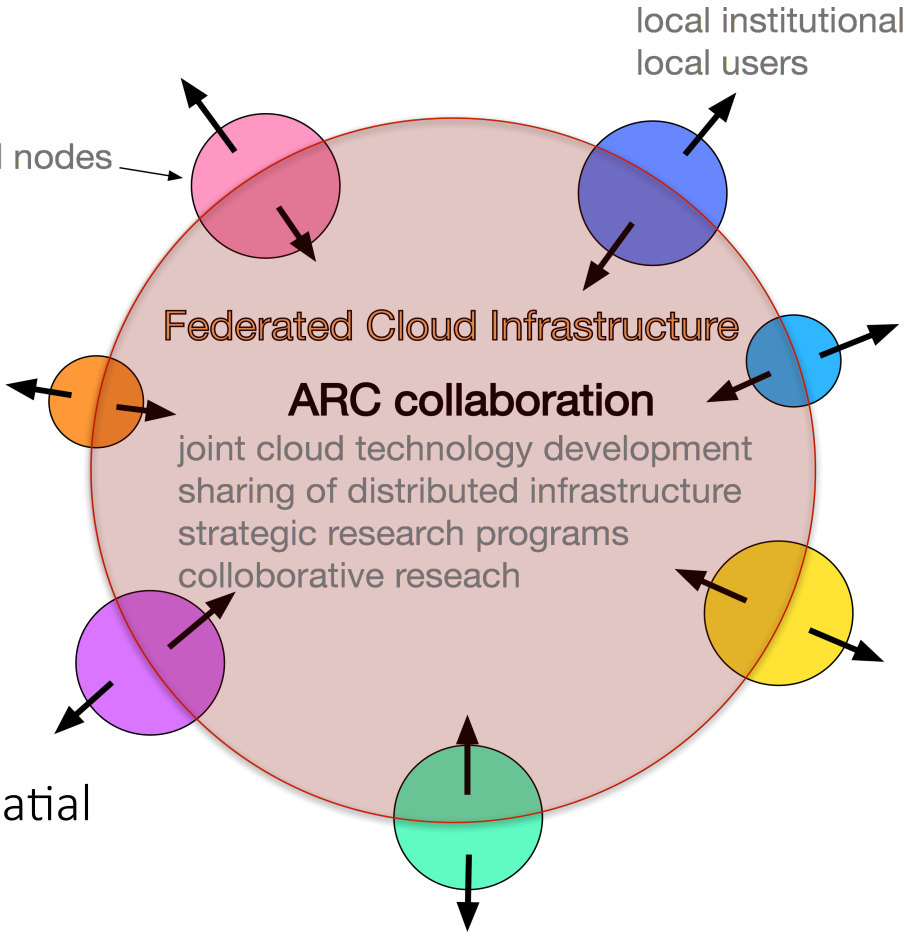




The African Research Cloud Project

- University of Cape Town
- University of the Western Cape
- Wits University
- University of Pretoria
- North-West University
- Sol Plaatje University

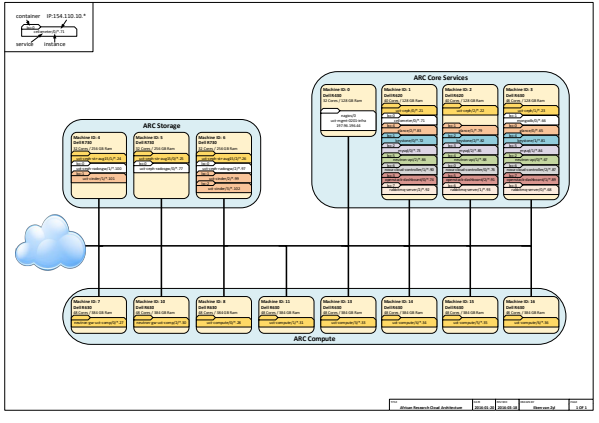
- SKA-South Africa
- South African National Space Agency



Strategic research programs
Astronomy, Bioinformatics, Geospatial

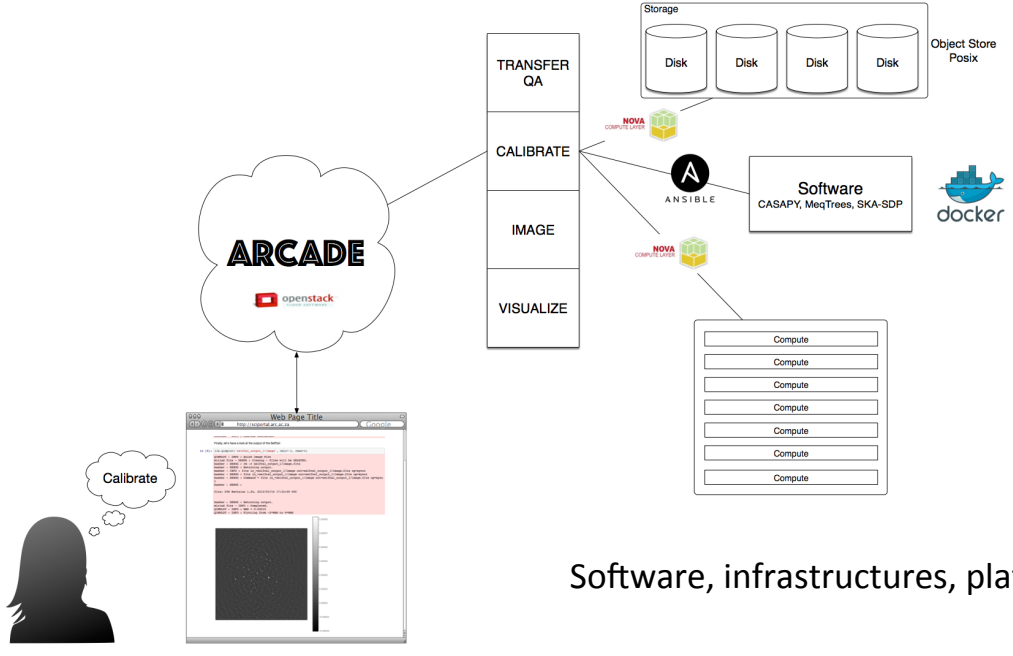
The African Research Cloud Project

- OpenStack virtualized environment with docker deployment support
- Mix of Swift object store, CEPH block and object store and BeeGFS parallel file-systems
- Interoperating between sites using OpenStack services
- Adding OCCI for federating beyond ARC
- Access services include cyberSKA and Jupyter notebooks
- CASA production and development environments in place
- Currently using keystone for AuthN and AuthZ
  - In process of moving AuthN to SAFIRE/eduGAIN
  - Exploring approaches to role based AuthZ



Cloud-based Pipelines

- Democratization of processing of big data and end user pipeline development
- Docker container with:
  - CASA, drive-casa, python libraries, Jupyter Notebook & Hub.
- Software framework for running pipelines in Notebook



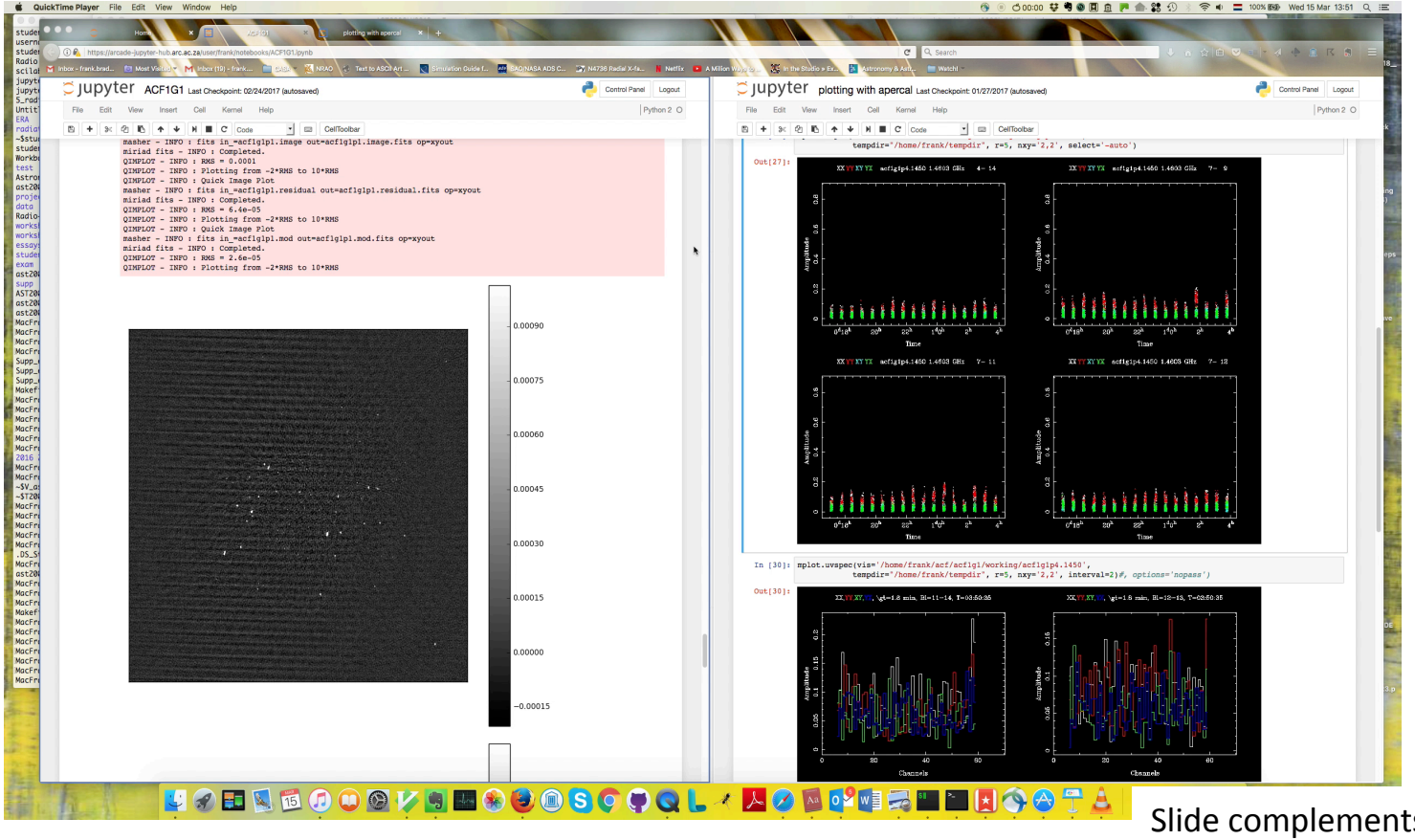
Software, infrastructures, platforms and data as service

Cloud-based Pipelines

Calibration of WSRT data

Pipeline Control & Data Exploration

Visualization



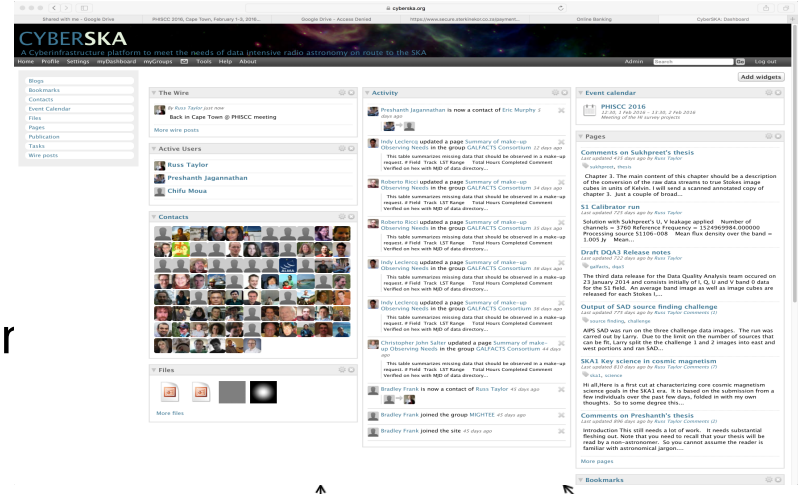
Slide complements of Brad Frank

Cloud-base Visualization

- Democratization
  - interactive user access for visualization and analytics of large remote sets
- Collaboration and sharing
  - Saving and sharing of visualization states
  - Interactive sharing of sessions among multiple users
  - User and group related annotations, notes
- Big data
  - Large data sets and compute intensive visual analytics
  - Server side compute abstracts compute and memory resources from the client/user
  - Flexible, dynamic allocation of resources to match the scale of the task
- Independence of processing and rendering from client
  - Serve various clients from mobile devices to digital planetaria.

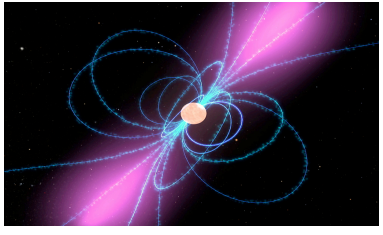
CyberSKA: A cloud-enabled Big Data Research Platform

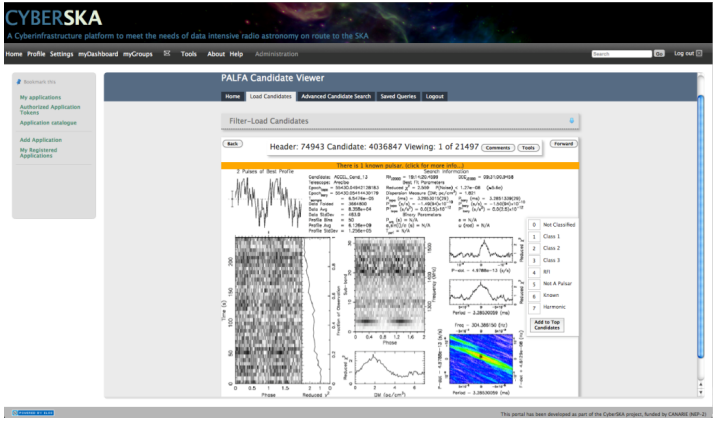
- Collaboration
  - Portal built on social networking and sharing technologies
- Data Management
  - Scalable collaborative access, sharing ar searching of distributed (BIG) data sets
- Data Visualization and Visual analytics
  - On-line interactive visualization of remote Big Data
- Third Party Applications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

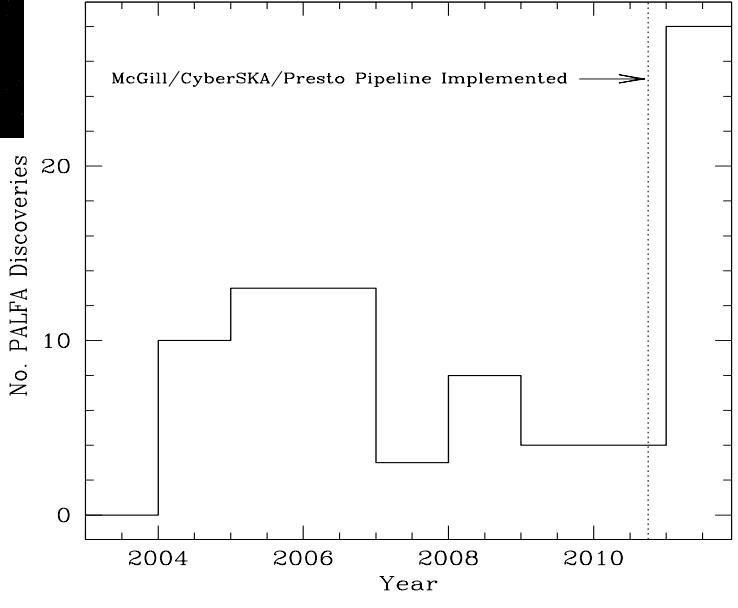


CyberSKA: Enabling Discovery

- PALFA: Millisecond Pulsar Search – 126 global collaborators





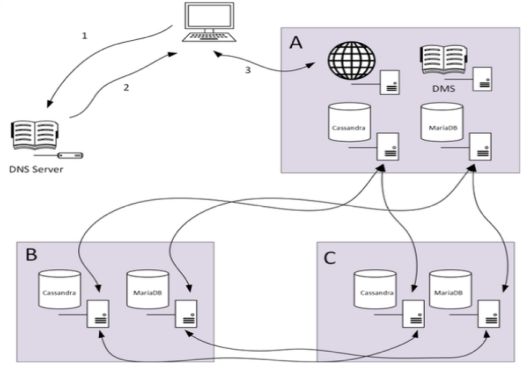


Federated CyberSKA Platform

Integrating Globally Distributed Resources into the CyberSKA Platform

David Aikema, Rob Simmonds and Russ Taylor – info@cyberska.org

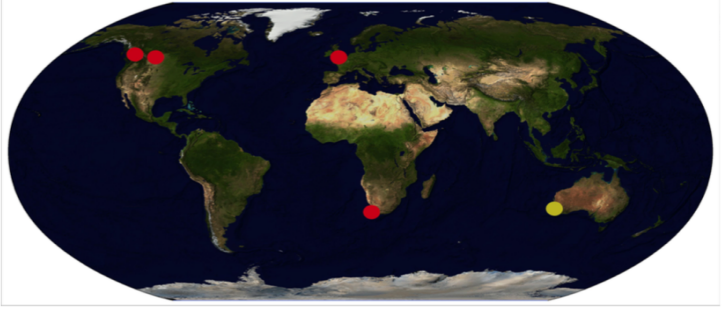
Restructuring for a globally-distributed system



- DNS based geolocation used to connect to the nearest CyberSKA instance – Region A in this case.
- Metadata describing files in the DMS instances in each region is stored in Apache Cassandra database that provides eventually consistent replication of this metadata between regions.
- MariaDB using multi-master replication with global transaction ID support is used to distribute portal configuration and account information.

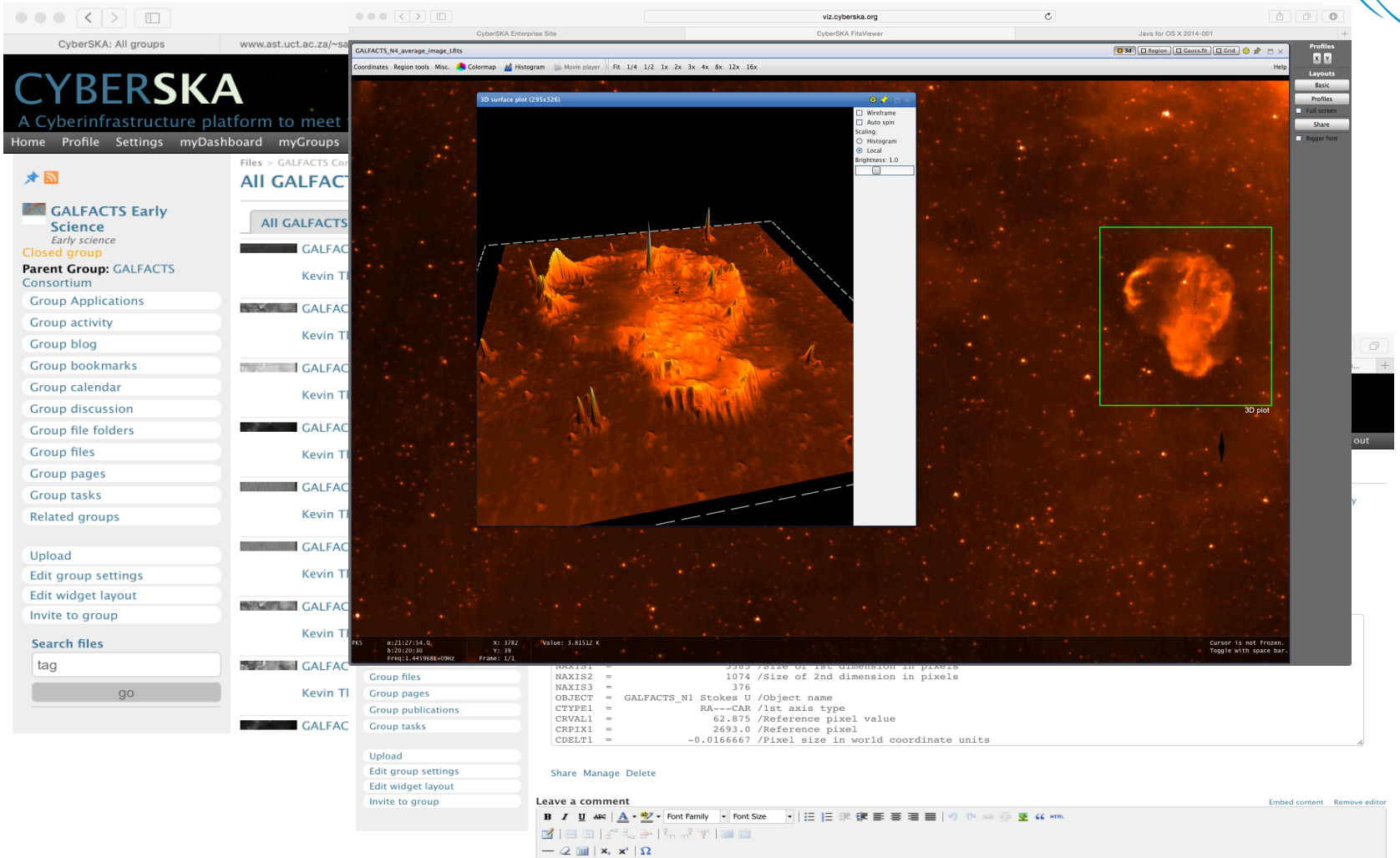
CyberSKA operates two platforms - one for production use and another for experimental purposes. This poster outlines some of the developments in our experimental testbed.

CyberSKA Testbed Resource Locations



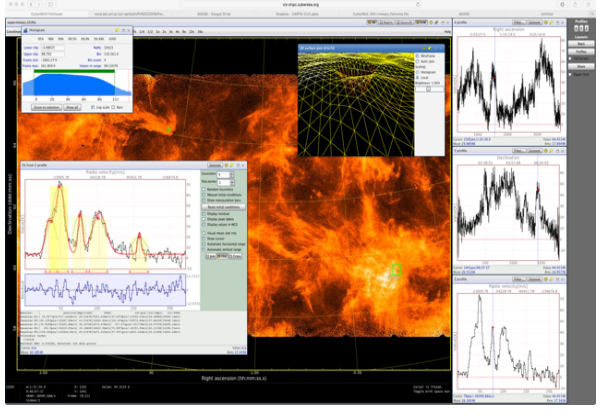
Red indicates currently active resources in the testbed whereas yellow indicates resources awaiting integration

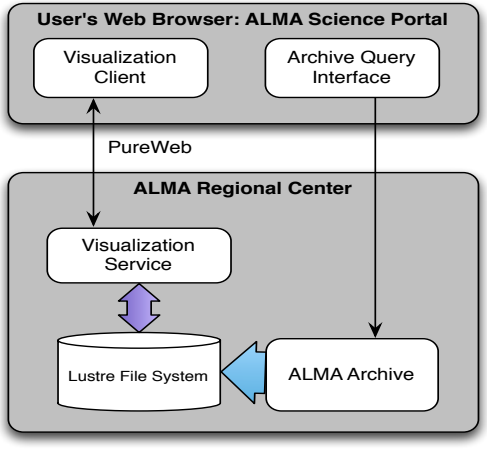
CyberSKA Project Portal



Next Generation Viewer (CARTA)

- Collaboration with IDIA, NRAO, ASIAA Taiwan, University of Alberta, SAP
- Next generation visualization system for radio astronomy to be implemented at ALMA Regional Science Centres
- For MeerKAT aim to scale analytics to multi-terabyte cubes
- Address use cases of MeerKAT Large Survey projects
- Considering HDF5 formats for efficient execution

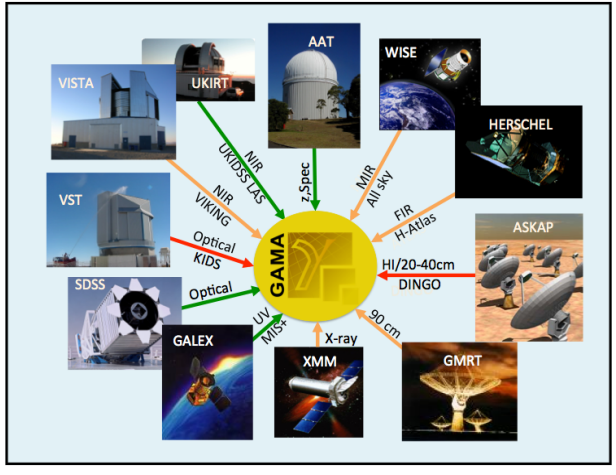




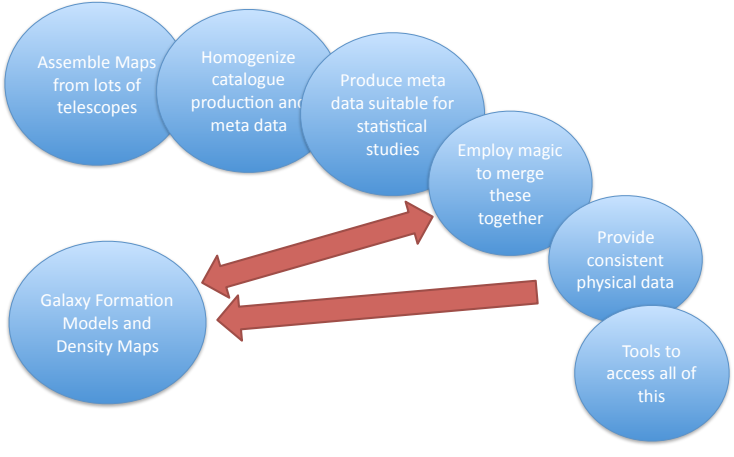
Fusion Large Multi-wavelength Data Sets

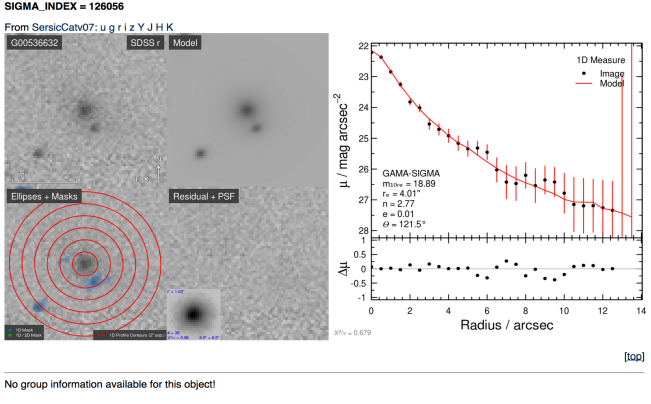
Planned integration of

- EU FP7 HELP project ending 2017
- South Africa GAMA multi-wavelength team
- African Research Cloud and cyberSKA platform
- IVOA interoperability with radio data (CADC)



Concept





IDIA leads: Mattia Vaccari, Michelle Cluver, Tom Jarret

IDIA Inter-University Institute for Data Intensive Astronomy

from big data to big ideas

www.idia.ac.za

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

UNIVERSITY of the WESTERN CAPE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SAP®