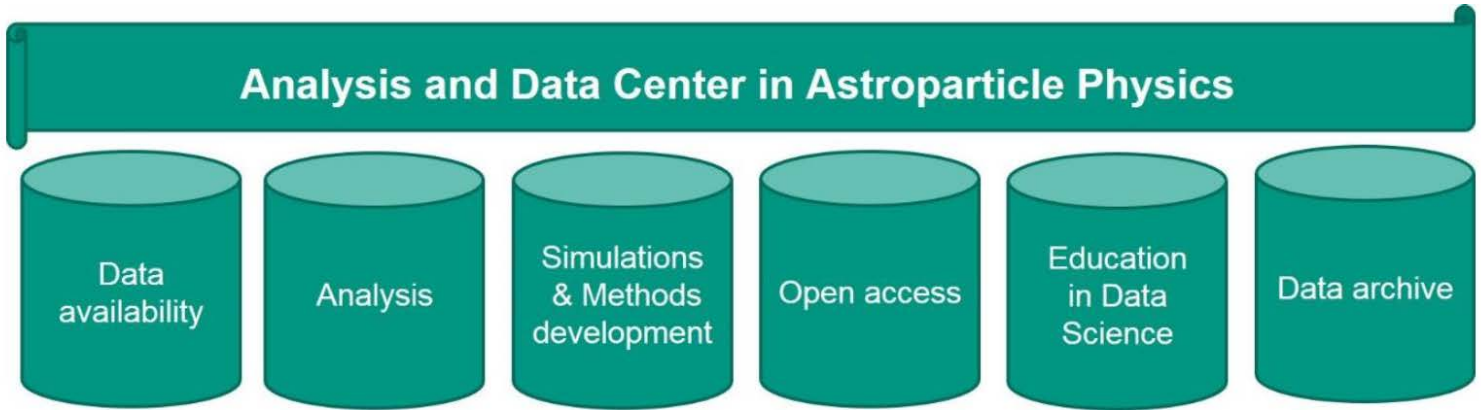


The six pillars in astronomy

Karl Mannheim

Contribution to: Initiative for a Data and Analysis Centre for Astroparticle Physics

Nov 2nd, KIT

General

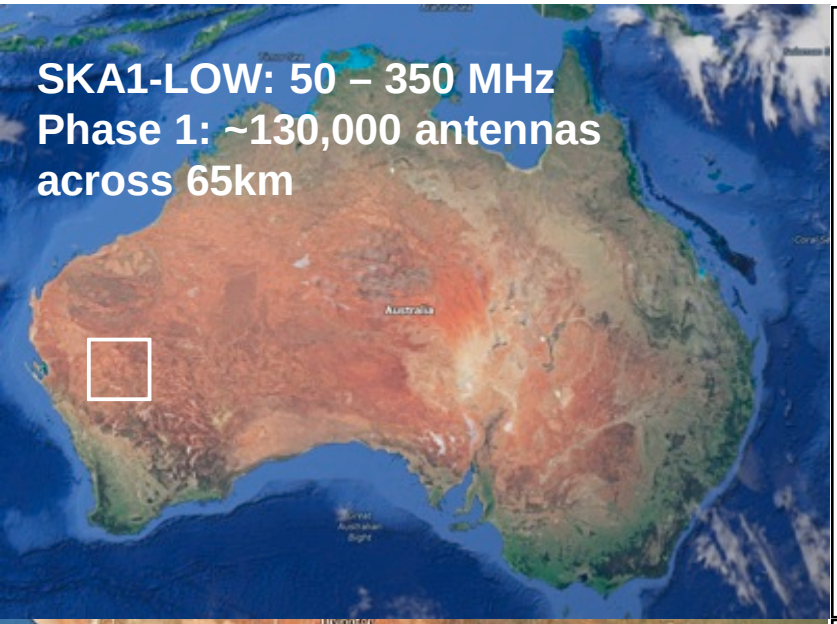
- **ROOTS:** Define data products transmitted via gateway from **central computing facilities** of hosted telescopes
- **FLEXIBILITY**: **diversity of data pipelines**
- **MANPOWER:** developing and maintaining data pipelines
- **COMPUTING:** **processing of raw data** to generate high-level data
- **ARCHIVE:** Long-term cost-efficient **raw data storage**
- **DISSEMINATION:** Storage and secure external access to **high-level data**
- **HPC:** Monte Carlo/numerical **simulations** (interface to specialized HPCs?)
- **R&D**: evolutionary approach to optimize *modus operandi*
- **SUSTAINABILITY**
- **TRAINING**
- **VIRTUAL OBSERVATORY:** (→ external partners such as CDS or GAVO)
  - Overlay images and c ross-identifications
  - Multi-wavelength spectra (correlations in energy space)
  - Statistical analysis (spatial correlations)
  - Time-domain studies (temporal correlations, transients)
  - Source classification (machine learning)
  - Pretty pictures and movies (public outreach)

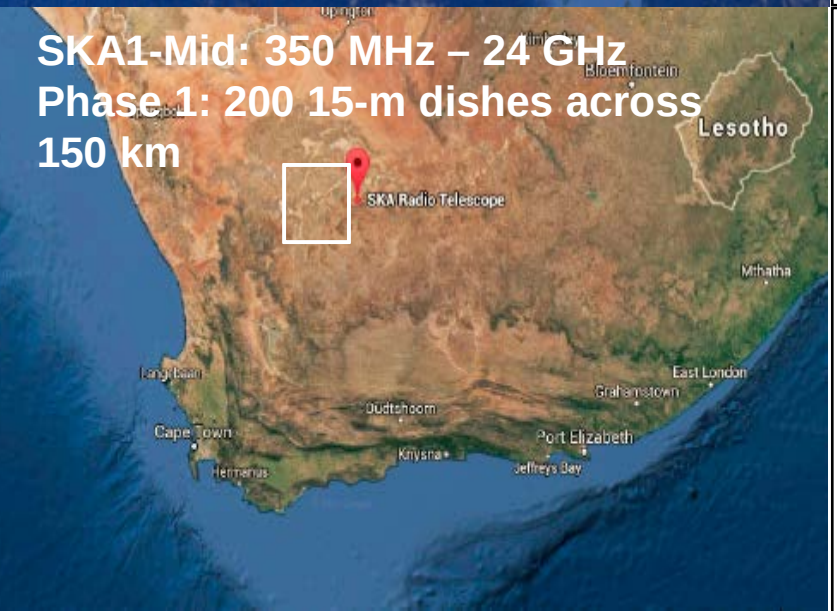
Examples

- **Examples of data/HPC center support:**
  - *Jülich Supercomputing Centre*: LOFAR/SKA data-pipeline support
  - *Rechenzentrum Garching*: EUCLID collects >500,000 images during 6 years of operation (30 Pbytes). Combine with hybrid data from ground-based spectroscopy (e.g. DES, KIDS). RZG provides: 36 compute nodes with 576 kernels and 500 TB storage realized as GPFS, external data access using gridFTP or Globus Online and external access to computing by globus GRAM.
  - LRZ: SuperMUC (2 Pflops/s) for numerical simulations (MHD, PIC)
- **Education**: Hands-on-Universe, robotic telescopes (MONET), citizen science (SETI, Panstarrs, galaxy zoo of SDSS)
- **Upcoming**: Euclid, eROSITA, Athena, LSST, ELT, CTA, LOFAR, **SKA**
- **Denkschrift 2017 & Strategiepapiere (RDS)**
  - Information Science and E-Infrastructure Challenges in Astronomy (Polsterer et al.)
  - Computational Astrophysics (Röpke et al.)

  → **www.denkschrift2017.de**

SKA: HQ in UK; telescopes in AUS & RSA





Data Flow through the SKA

SKA1-MID

8.8 Tb/s

SIGNAL AND DATA TRANSPORT

~50 PFLOPS

CENTRAL SIGNAL PROCESSOR

~5 Tb/s

7.2 Tb/s

~2 Pb/s

LOW-FREQUENCY APERTURE ARRAY

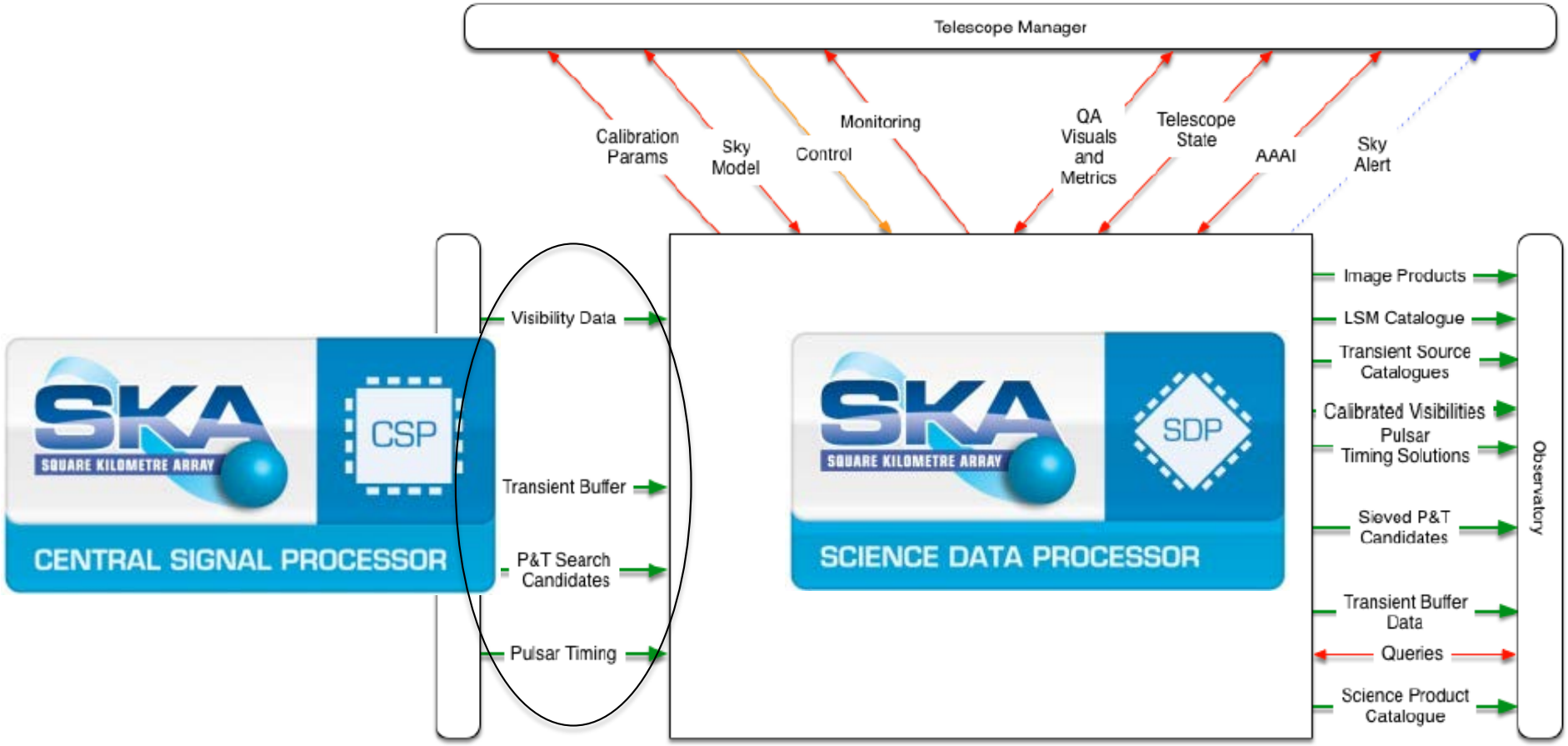
beamforming

~250 PFLOPS

SCIENCE DATA PROCESSOR

SKA1-LOW

~300 PB / year



Raw data processing is extremely challenging

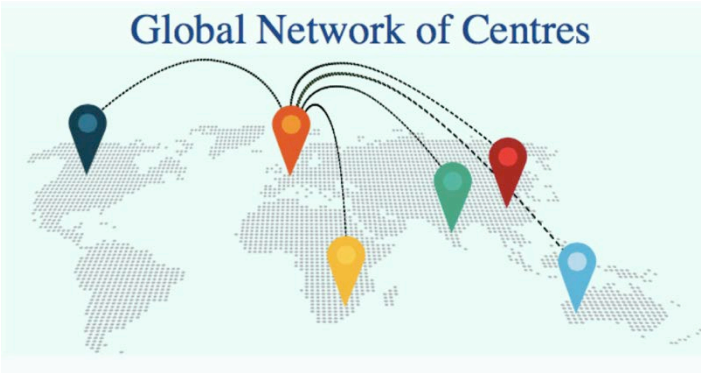
- Pre-analysis in near-realtime and strong data reduction (~ 10^5)

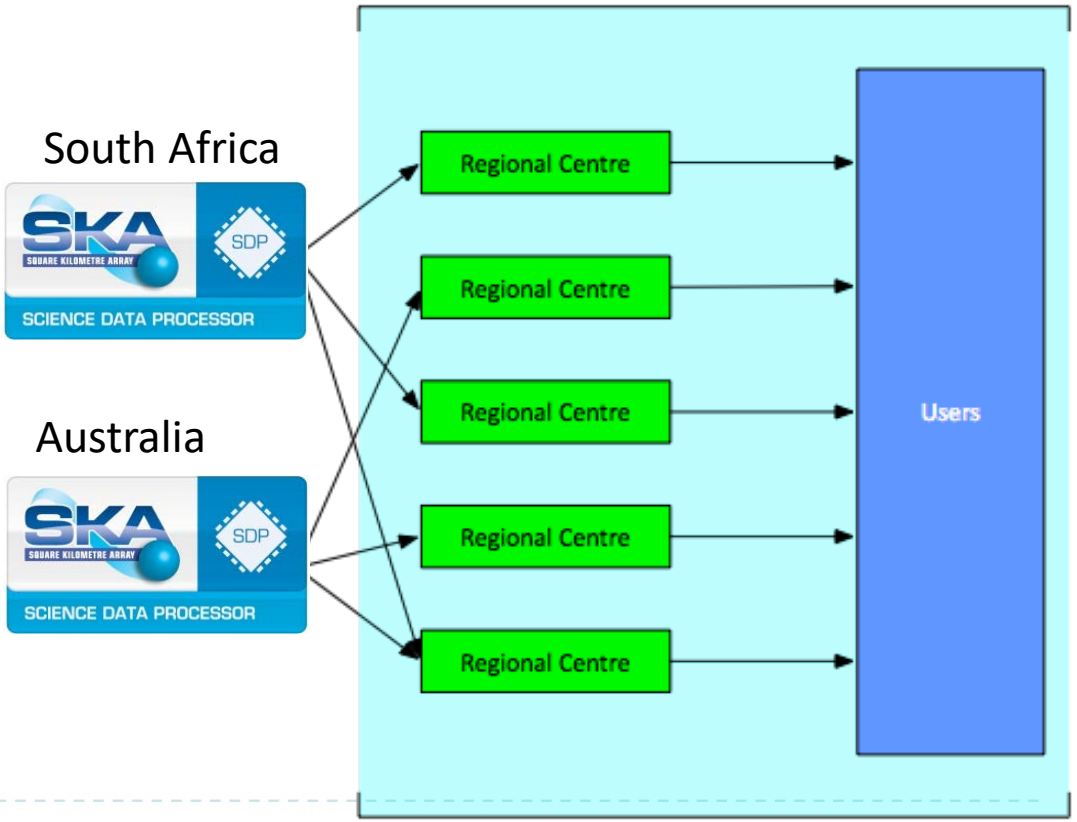
- Science data products = large objects (up to ~ 1 Petabyte / 3D image)
  - To be “improved” in Regional Data Centers

The science data products that emerge from the SKA observatory are not in the final state required for science analysis

SKA Regional Centres – outside SKAO scope

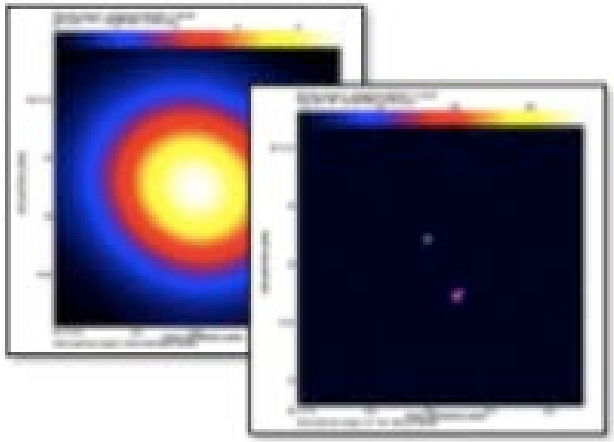
- Required
  - capacity for reprocessing data and their analysis
  - storage for a long-term archive
  - local user support
- Intent
  - SKA partner countries planning SKA regional Centres
  - National super-computing centres
  - Provide local support to scientists
  - Development of new techniques, new algorithms
  - Deliver SKA science





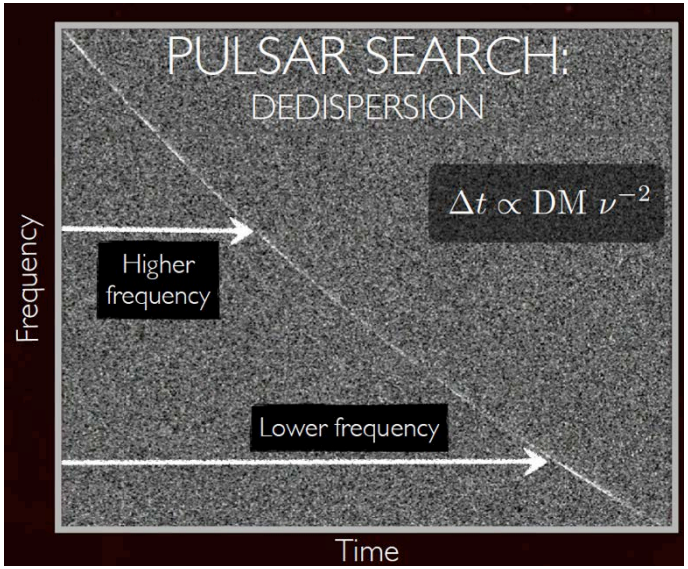
SKA Regional Centre

Data Processing



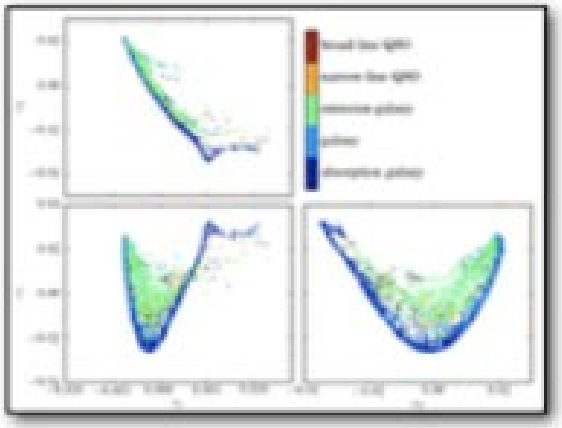
- Reprocessing
- Calibration and imaging
- Source extraction
- Catalog (re-)creation
- DM searches

merging individual observations of long term projects



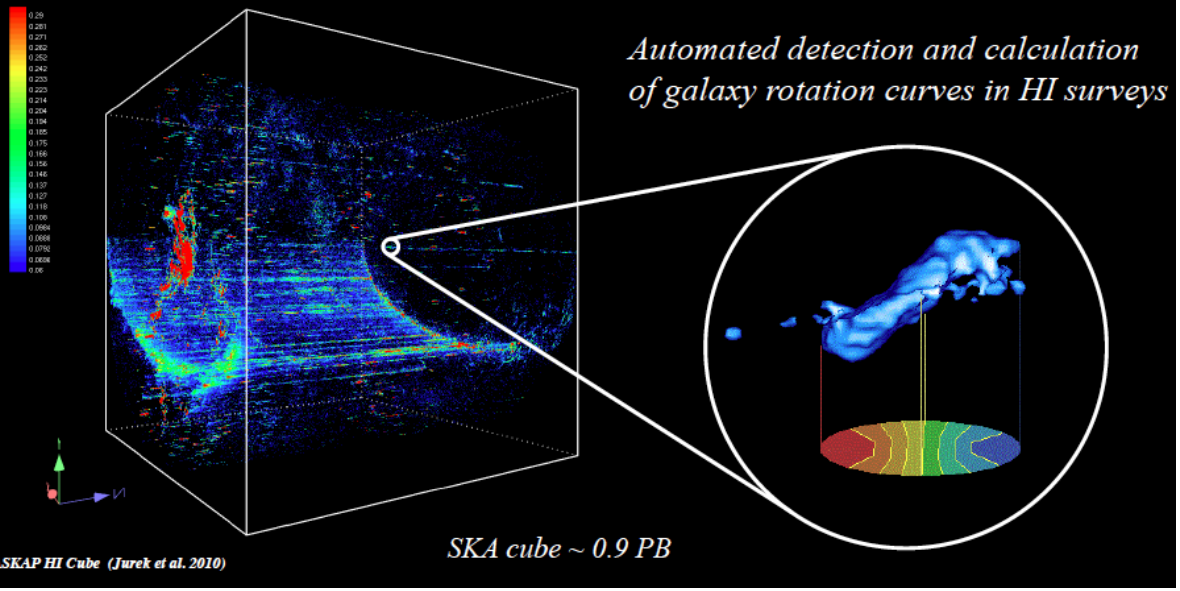
Data Mining

- Multi-wavelength studies
- Catalog cross-matching
- Transient classification
- Feature detection
- Visualization



SKA Regional Centre

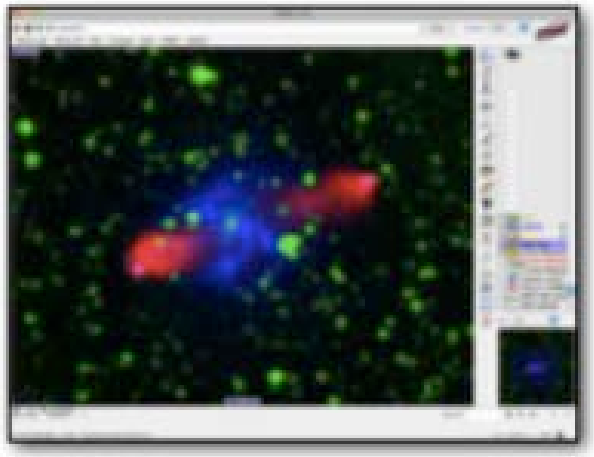
Visualisation, Detection, Classification, Inference



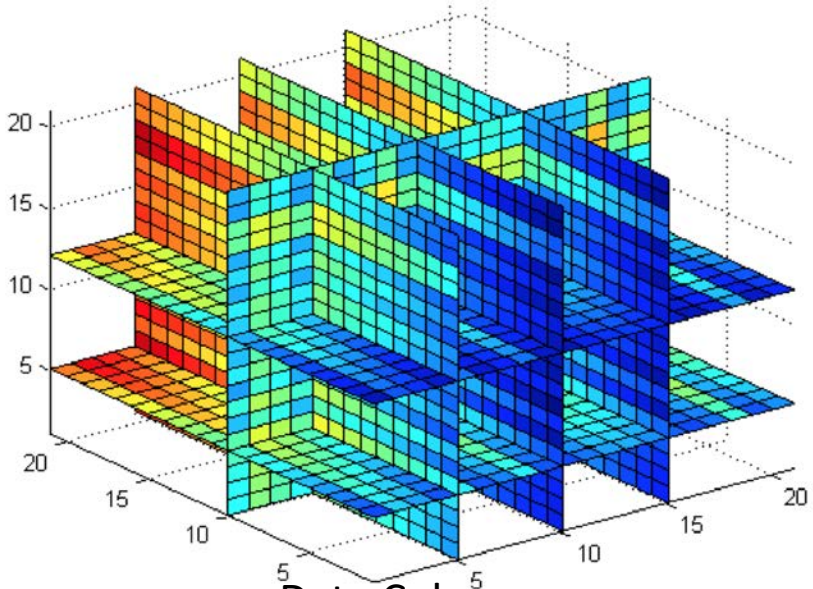
SKA Regional Centre

Data Discovery

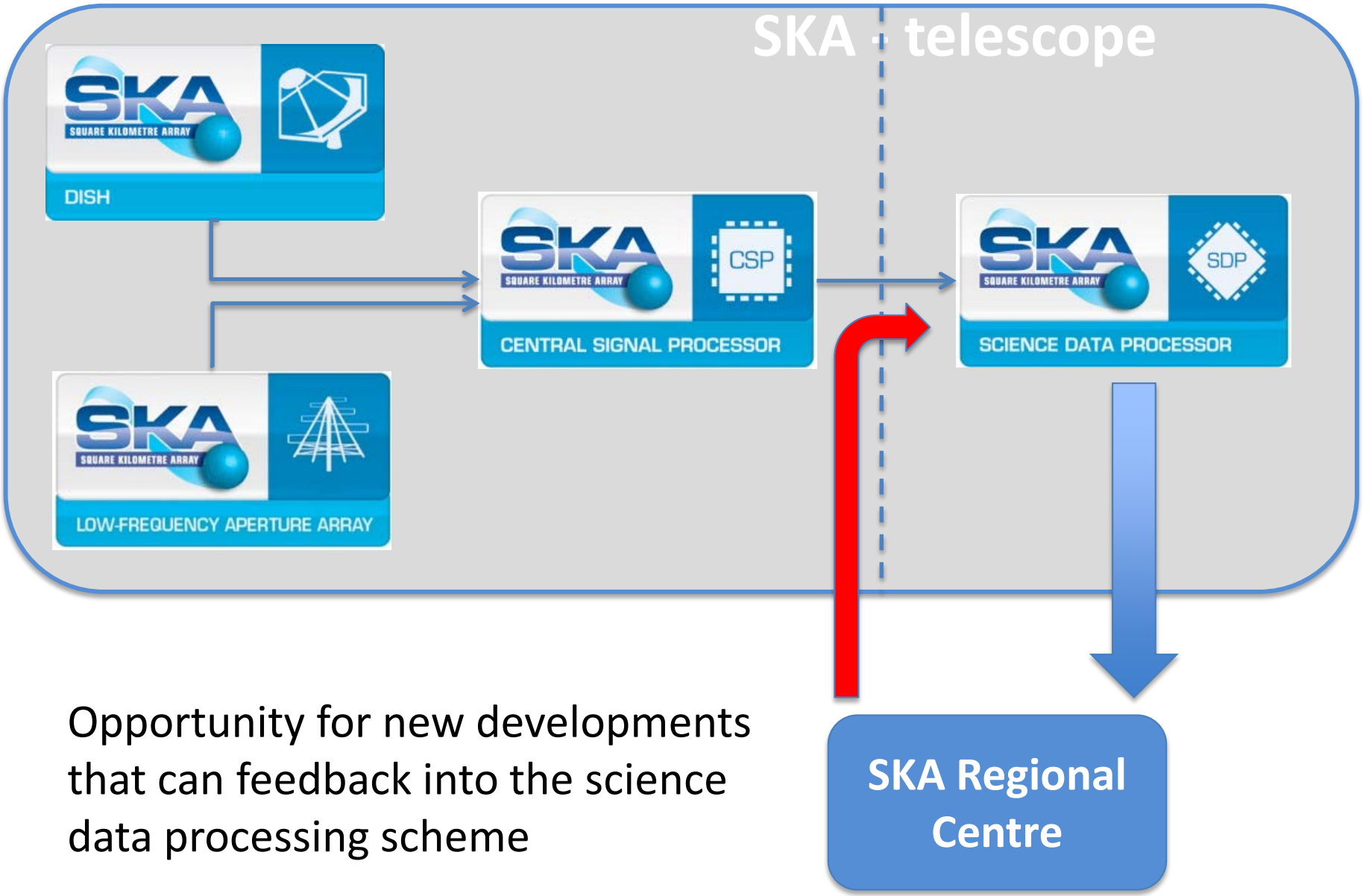
- Observation database
- Quick-look data products
- Flexible catalog queries
- Integration with VO tools
- Publish data to VO



Discovery of the un-knowns in semi-raw visibility data

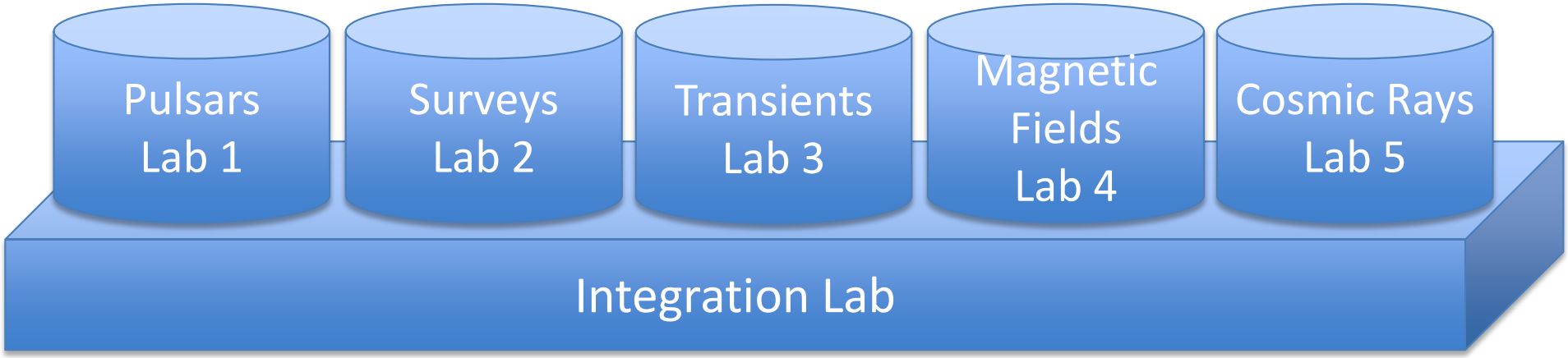


Data Cubes:
36.600 x 36.600 x 250.000 Pixel
up to PetaByte / Cube



Opportunity for new developments that can feedback into the science data processing scheme

National SKA Data Centre



5 Key Science Labs

- The SKAO has compiled a list of 13 High Priority Science Objectives (HPSO)
- HPSO require different approaches in data analytics due to different nature of data.

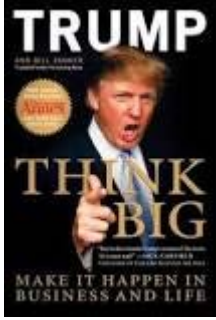
1 Integration Lab (general support)

- Basic services: Network, HPC, system integration
- Theory: Information theory, algorithms for Big Data analytics, Machine Learning, …

In total: approx. 50 people required

- **How to set this up properly ?**

Think big enough!



- **SKA national center could host all other experiments/observatories**
- Amazon-type of investment: civil playground for future big data analytics developers
- Part of European science cloud based on **CERN-SKA cooperation: Each of the current HGF centers is way too small!**
- Data archive
  - **Ten-ExaByte**-scale mass storage with fast access
  - Responsibility for part of SKA central computing facility
  - Sustainable design (power supply /waste heat for green houses)
- Analysis and data science competence center
  - **Ten-PFLOPS-scale** compute power
  - 50 data scientists
  - Software implementations of data pipelines
  - Algorithms for data analysis and visualisation
  - User support
- Master-supplement „**data scientist/engineer**“