ESCAPE WP2/WP5: SKA Use Cases

ESCAPE WP2/WP5 Integration Workshop 06-04-2021

SQUARE KILOMETRE ARRAY

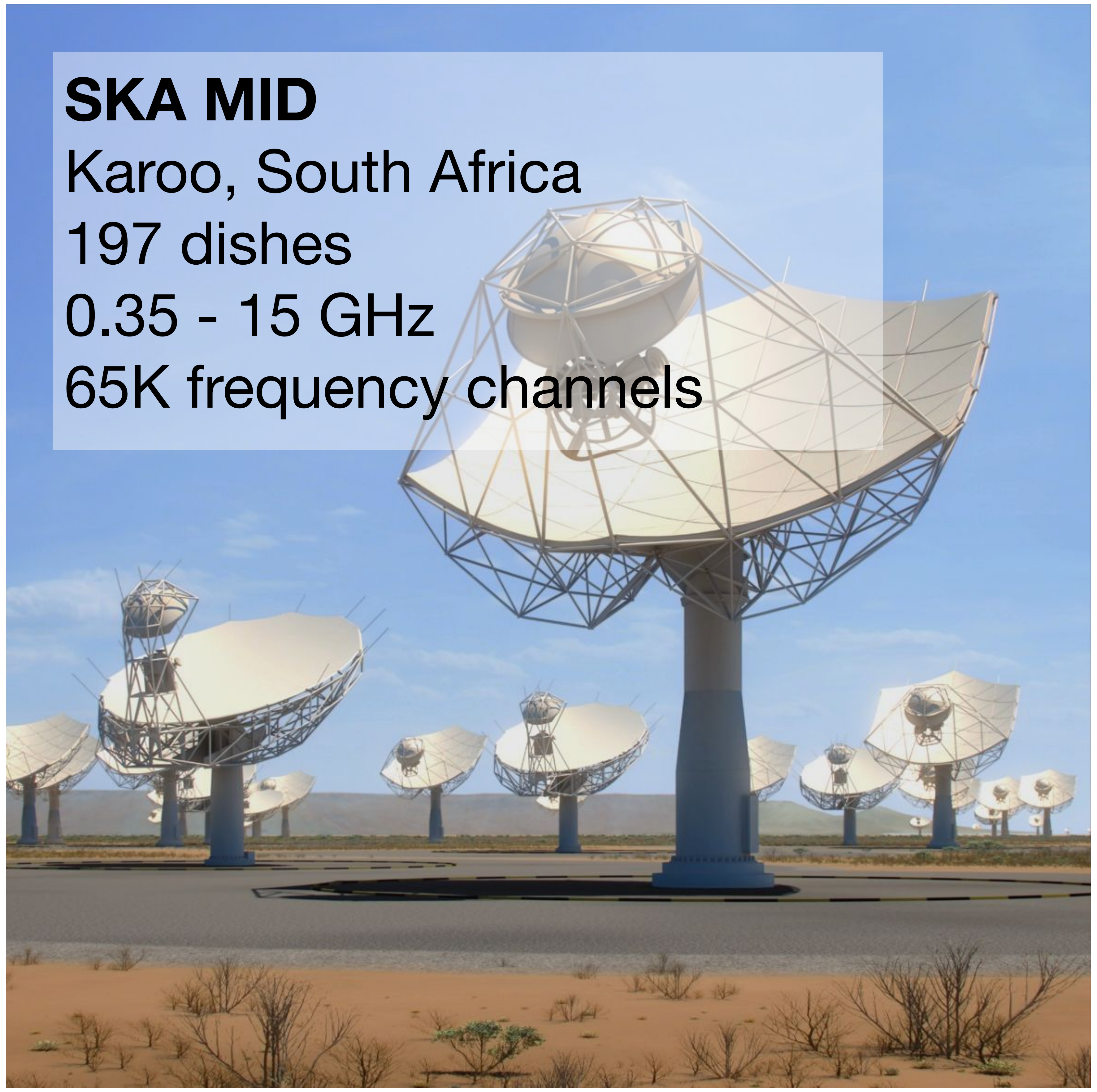
Exploring the Universe with the world's largest radio telescope

James Collinson
Operations Data Scientist

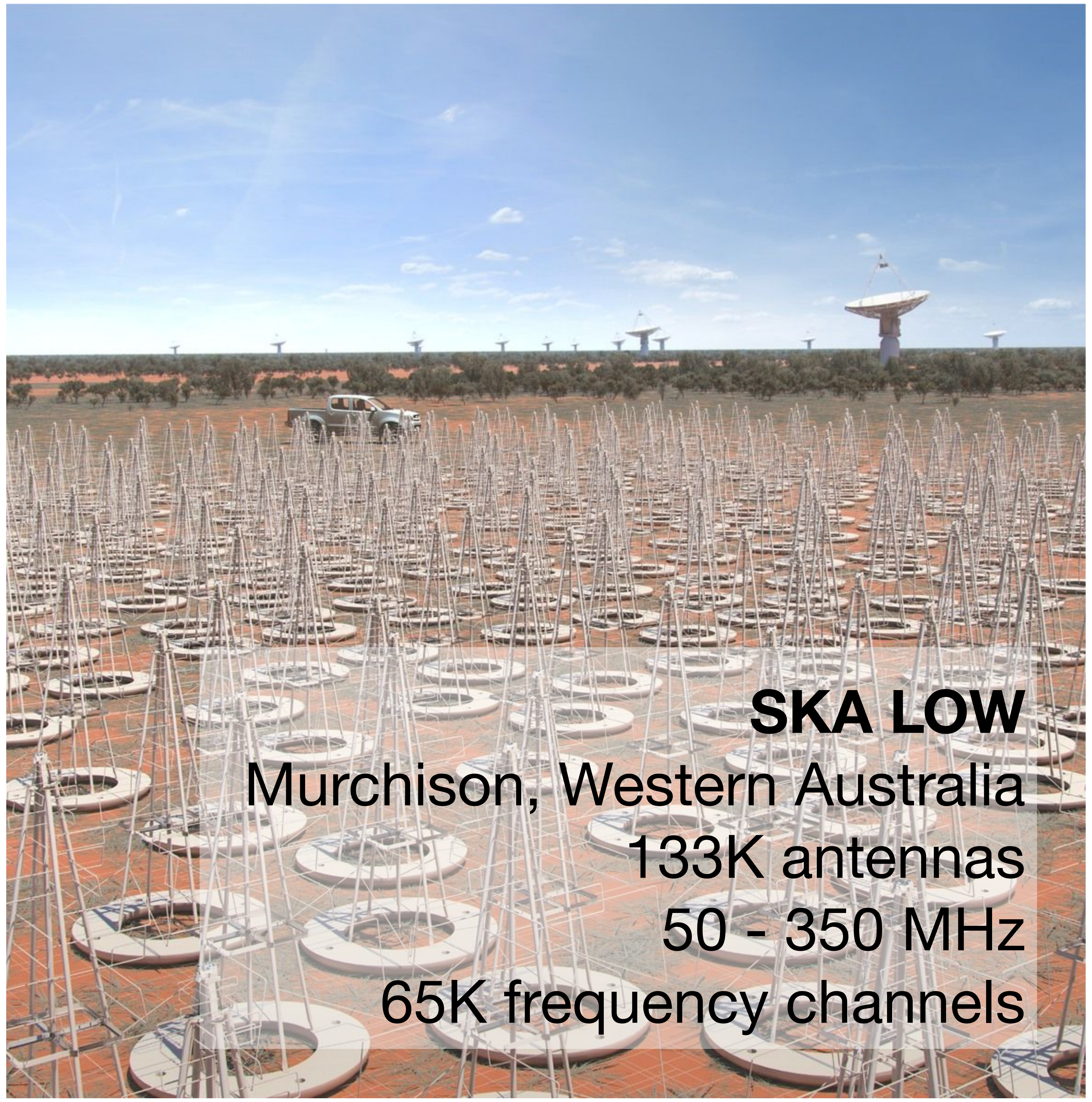
Overview

- SKA Observatory
  - Observatory data lifecycle and distribution model
- SRC Capabilities
  - Data archiving and management (WP2)
  - Data processing (WP5)
  - Astronomical data operations (WP4)

SKAO - An Observatory



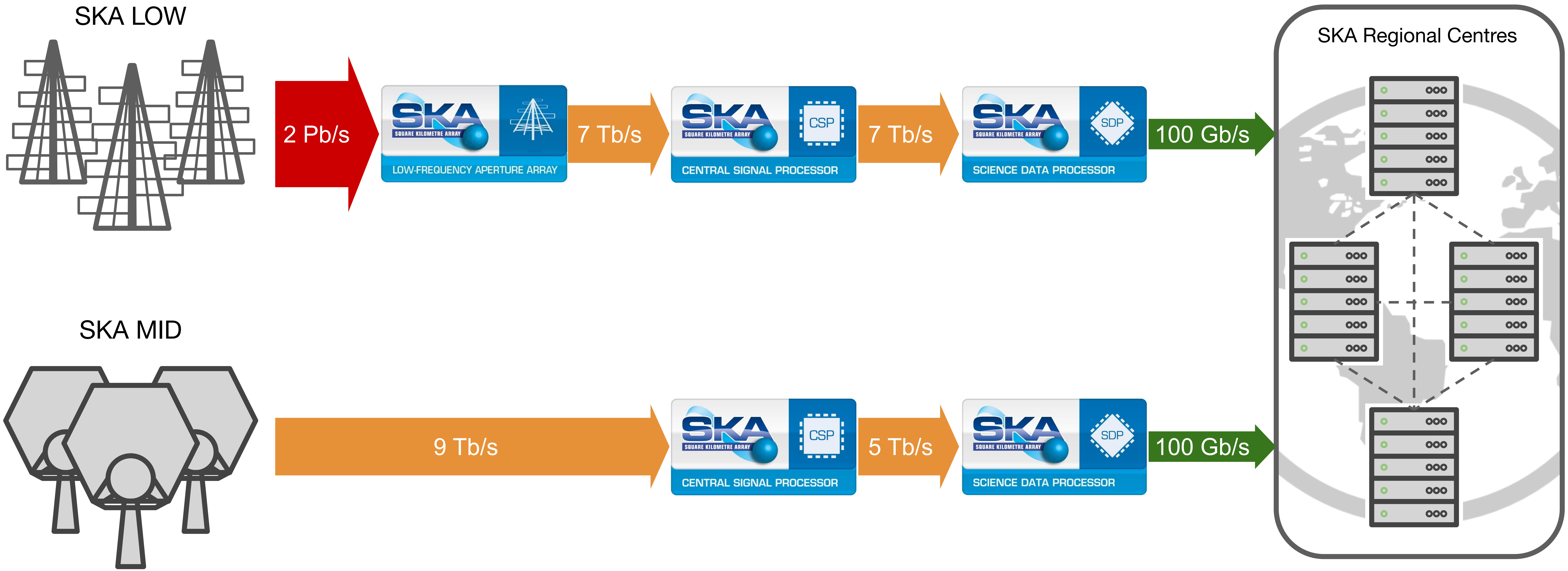
SKA MID
Karoo, South Africa
197 dishes
0.35 - 15 GHz
65K frequency channels



SKA LOW
Murchison, Western Australia
133K antennas
50 - 350 MHz
65K frequency channels

Test systems already taking data. Main science programmes from ~2028/9. 50 year lifetime.

SKA Observatory Data Flow



* Data rates approximate

SKA Regional Centres

ARCHIVE

Archival of the observatory data products. Once scientific results are published, outputs of analyses are made available.

DISTRIBUTED DATA PROCESSING

Use cases are made to be reproducible. Compute comes to the data (high data volume).

DATA DISCOVERY

Once SDP has pushed the data to the regional centres, how will users find/peruse their data? How will data from published results be easily found?

USER SUPPORT

SRCs must support the key science project teams as well as general users. This will mean user ability will be varied.

INTEROPERABILITY

Multiple regional SRCs, locally resourced but interoperable. SRCs may be heterogeneous in nature but with common core functionality.

Credit: Rohini Joshi

SKA Regional Centres

ARCHIVE

Archival of the observatory data products. Once scientific results are published, outputs of analyses are made available.

DISTRIBUTED DATA PROCESSING

Use cases are made to be reproducible. Compute comes to the data (high data volume).

DATA DISCOVERY

Once SDP has pushed the data to the regional centres, how will users find/peruse their data? How will data from published results be easily found?

USER SUPPORT

SRCs must support the key science project teams as well as general users. This will mean user ability will be varied.

INTEROPERABILITY

Multiple regional SRCs, locally resourced but interoperable. SRCs may be heterogeneous in nature but with common core functionality.

Credit: Rohini Joshi

SKA Use Case: Data Management

ARCHIVE

Archival of the observatory data products. Once scientific results are published, outputs of analyses are made available.

- ESCAPE
  - Production Rucio version on CERN infrastructure
  - Multi-experiment prototyping data transfers
- SKA
  - SKA-specific operations (policies, member collaboration...)
  - Build internal experience
  - Long distance transfers

SKA Use Case: Data Management

ARCHIVE

Archival of the observatory data products. Once scientific results are published, outputs of analyses are made available.

- SKA RUCIO
  - Use cases:
    - PI-led observation policies (embargoed data)
    - Data life cycles to serve observatory policies, e.g.
      - Migration from ‘site’ to SRC network
      - Maintenance of archive (number of copies etc)
      - Staging for advanced data product generation and processing

SKA Use Case: Data Processing

DISTRIBUTED DATA PROCESSING

Use cases are made to be reproducible. Compute comes to the data (high data volume).

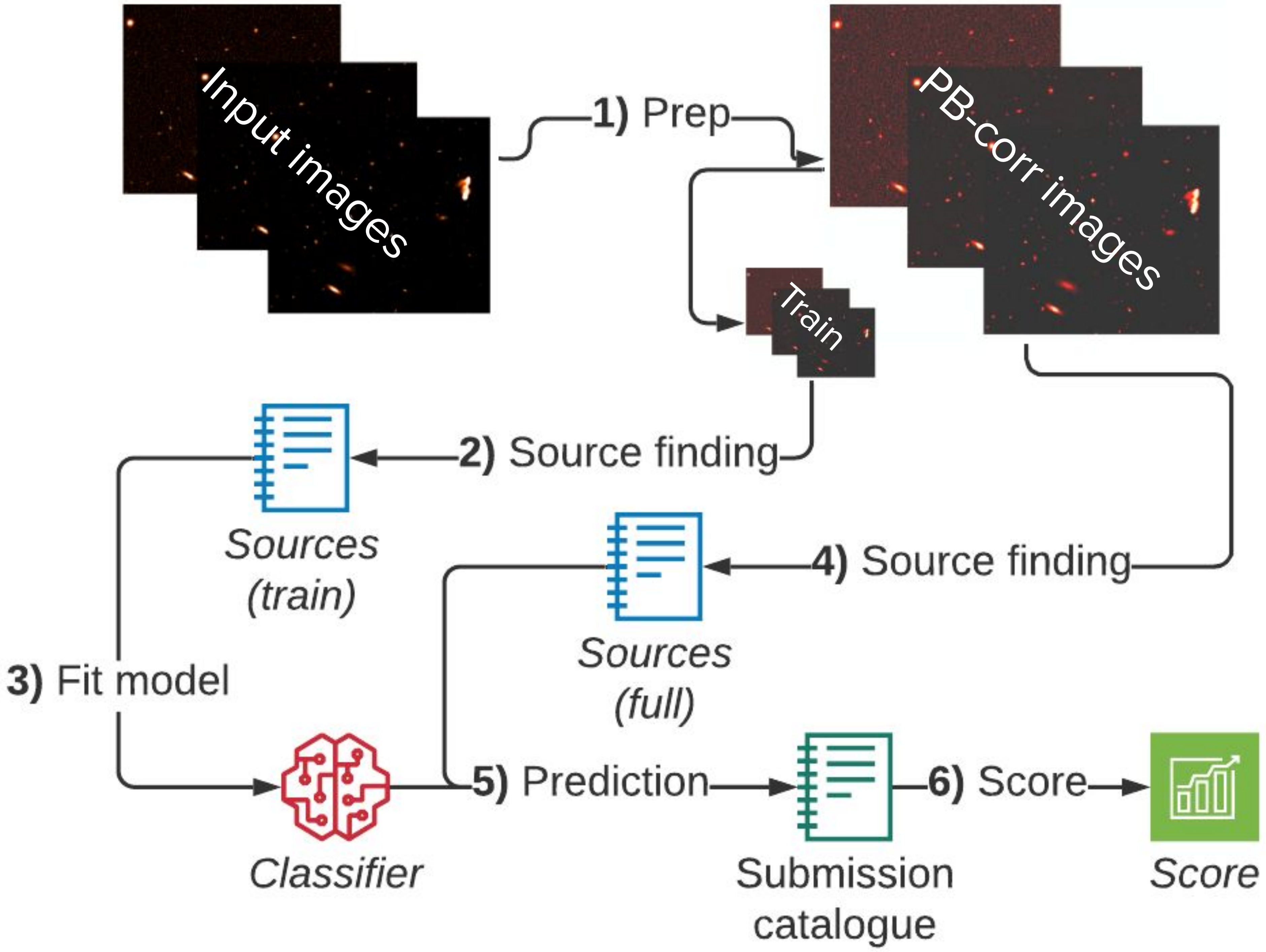
- Homogeneous operations - batch processing
  - Generation of Project-level Data Products from ODPs
- Interactive Data Analysis
  - Astronomer-led data operations
  - Heterogeneous operations
    - Development and deployment of analyses, both novel and proven

SKA Use Case: Data Processing

DISTRIBUTED DATA PROCESSING

Use cases are made to be reproducible. Compute comes to the data (high data volume).

- SKA JupyterHub prototype
  - Publicly visible (via ESCAPE IAM)
- SKA Science Data Challenge 1 solution workflow
  - Processing of small synthetic images
  - Notebook environment
  - Uses proven source-finding software (LOFAR) and ML classification



Exploring the Universe with the world's largest radio telescope

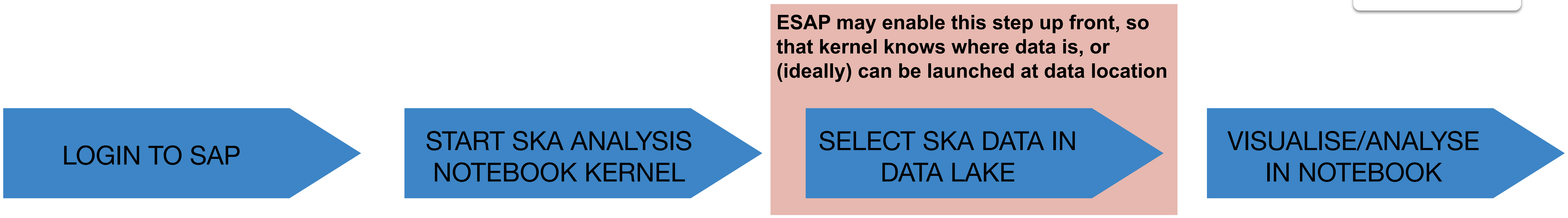
SKA Use Case: Astronomical Data Ops

DATA DISCOVERY

Once SDP has pushed the data to the regional centres, how will users find/peruse their data? How will data from published results be easily found?

- Visualisation of large (PB-scale) data cubes
- IVOA recommended schema (Hierarchical Progressive Surveys)
  - Currently unable to integrate into notebook environment with custom data
- CARTA (Cube Analysis and Rendering Tool for Astronomy) to be explored further

Current integration - ESAP Use Case?



ESAP may enable this step up front, so that kernel knows where data is, or (ideally) can be launched at data location

LOGIN TO SAP

User will authenticate with external service to login to Science Analysis Platform

START SKA ANALYSIS NOTEBOOK KERNEL

Choose from a selection of notebook server environments depending on requirement

SELECT SKA DATA IN DATA LAKE

In running kernel, user can download data from data lake prototype

VISUALISE/ANALYSE IN NOTEBOOK

Interactive visualisation and processing of chosen data using in Jupyter notebook

Summary

- SKA Observatory - 700 PB/yr of data
- Managed and maintained within SKA Regional Centres
- Use Case: Data management
  - Rucio data lake orchestrator is mature, still requires some policy functionality for SKA use case
- Use Case: Data processing
  - JupyterHub platform, requires further integration with data storage and visualization tools