SKAO SRC Batch Use Cases
ESCAPE Batch Processing Use Cases
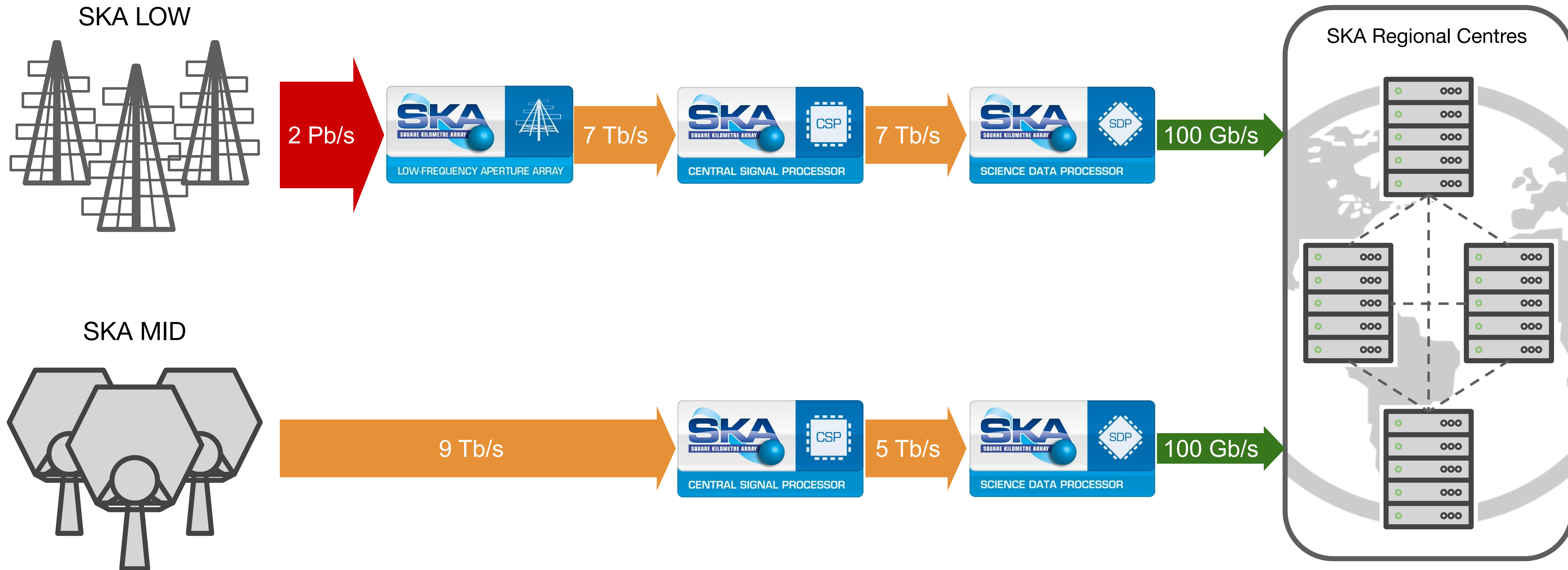
James Collinson

Overview

- SKAO data rates and models
- Observatory use cases
- Astronomer use case:
  Source identification/characterization
  - SoFiA 2
  - U-Net
  - R-CNN
- Consequences for SKAO batch processing

SKA Observatory Data Flow

SKA LOW

2 Pb/s — LOW-FREQUENCY APERTURE ARRAY — 7 Tb/s — CENTRAL SIGNAL PROCESSOR — 7 Tb/s — SCIENCE DATA PROCESSOR — 100 Gb/s — SKA Regional Centres

SKA MID

9 Tb/s — CENTRAL SIGNAL PROCESSOR — 5 Tb/s — SCIENCE DATA PROCESSOR — 100 Gb/s

* Data rates approximate

Observatory-led batch processing

- SKAO will output ODPs
  - Data model TBD, likely 3-5 dimensional data cubes
- ADPs will be created in SRCs
  - Combined observations (deeper/broader field)
  - Reduced data products (source catalogues, transient candidates, etc)

Amr Hassan
Swinburne University of Technology
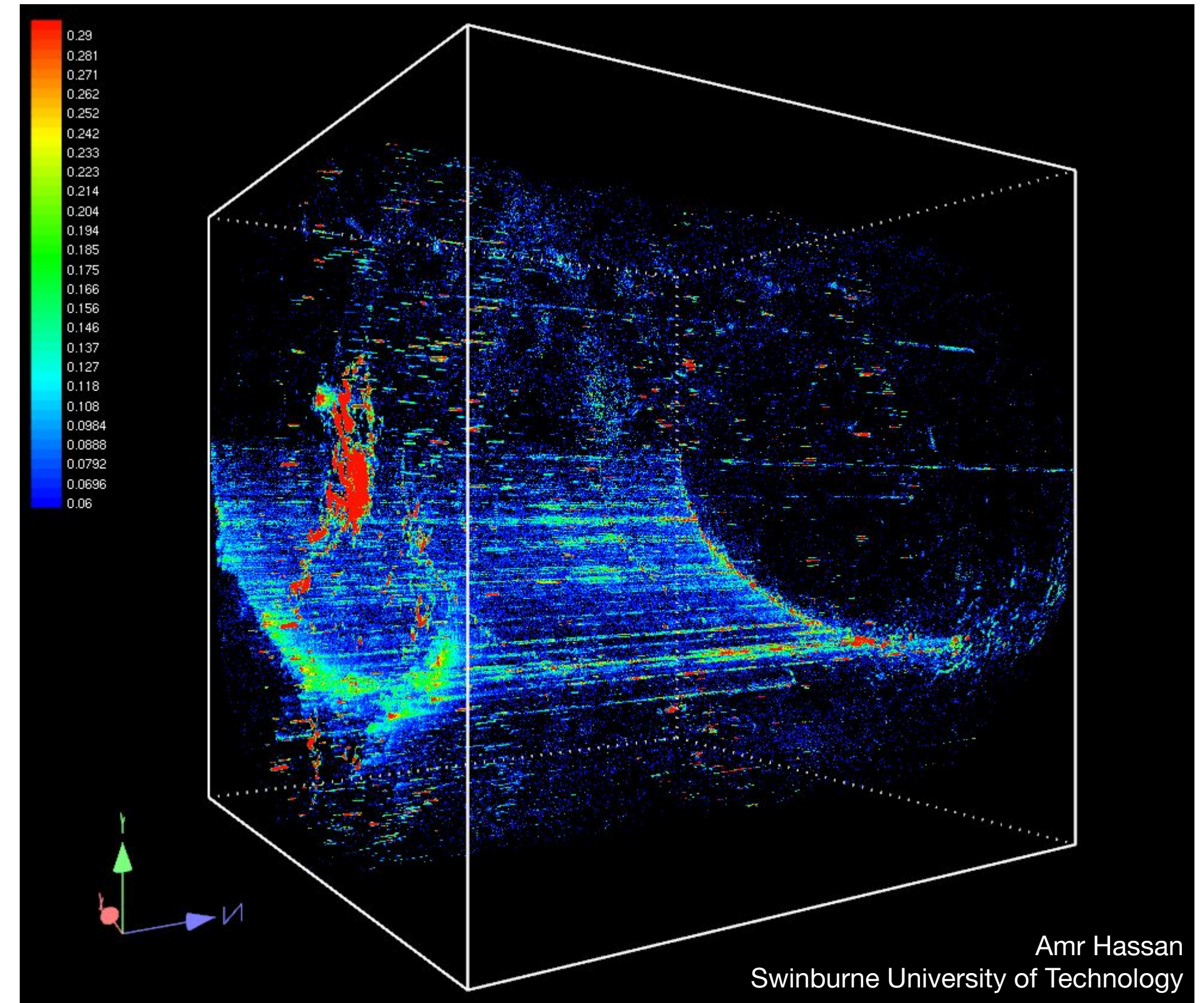
User-led batch processing

- Users will define pipelines to scale up from e.g. interactive analysis, or test new algorithms
  - Training models
  - Model validation
  - Tuning hyperparameters
- May wish to apply algorithms to large number of transient candidates

Use Case: Source Identification/Characterization

- SDC2 ran last year, publication in prep
  - 1TB simulated data cube
  - Teams challenged to find and characterise HI sources
- Identification
  - All teams segmented the 1 TB cube for training
  - Full SKAO cube would be ~1000 times larger

Use Case: Source Identification/Characterisation

- The [Source Finding Application (SoFiA)](#)
  - Parallel algorithm for identifying HI sources
  - CPU-based, more workers improve performance
  - Overhead from parallel file access (400% longer)
  - RAM required ~ 2×(input file size)
    - Break into small enough segments to process on available resource
- U-Net, R-CNN
  - Related means of detecting objects in images
  - Benefit from GPU acceleration for training
  - Also used for source characterisation (separate stage)

Consequences for batch processing

- Heavy-duty SKAO workflows likely to be highly parallelizable
- Homogenous workflows will be performed by the observatory
- More heterogeneous workflows developed by user community
  - E.g. SDC2: algorithmic source identification/characterization
  - Interactive analysis modes will inform development of these

We recognise and acknowledge the
Indigenous peoples and cultures that have
traditionally lived on the lands on which
our facilities are located.

SKAO

www.skao.int