



one observatory
two telescopes
three continents

SKA Science Data Challenges

Preparing the community for a new era in astronomy

Philippa Hartley and Anna Bonaldi
SKA Regional Centre Training Event



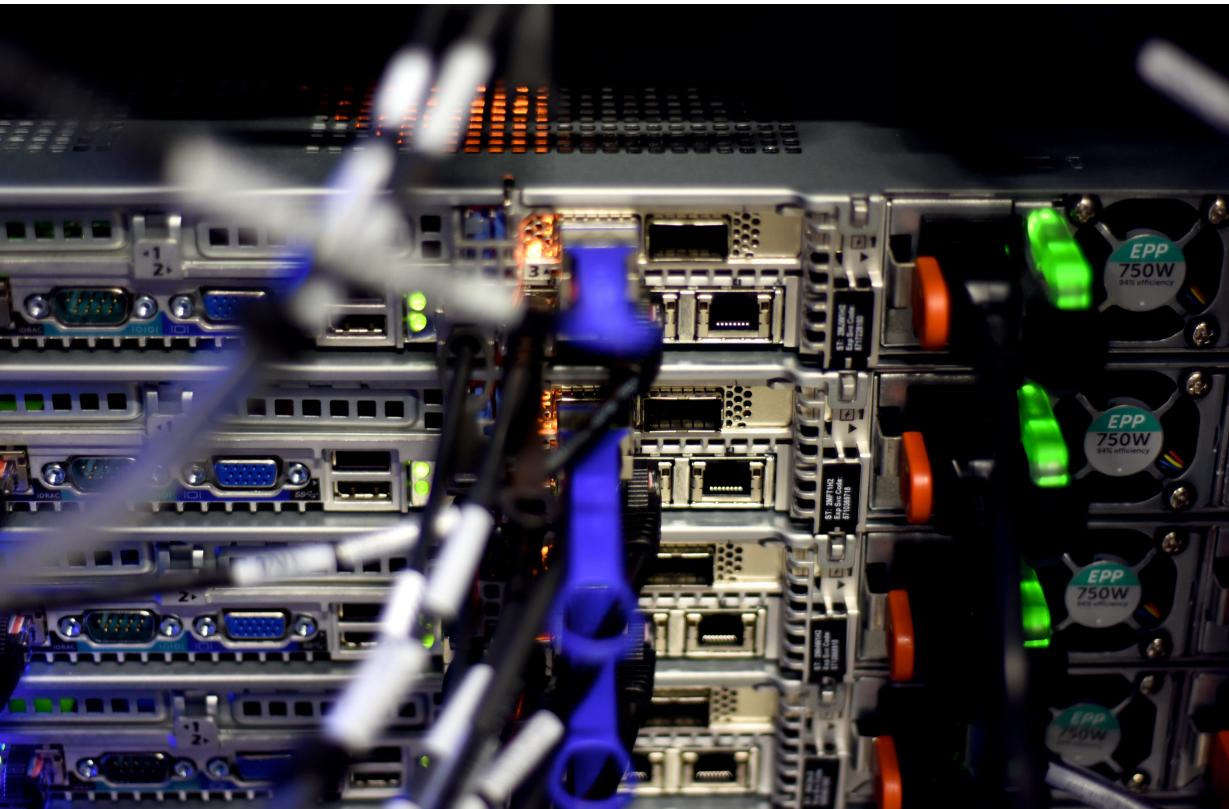
The SKA data journey

SKA LOW



SKA MID

SKA data processing

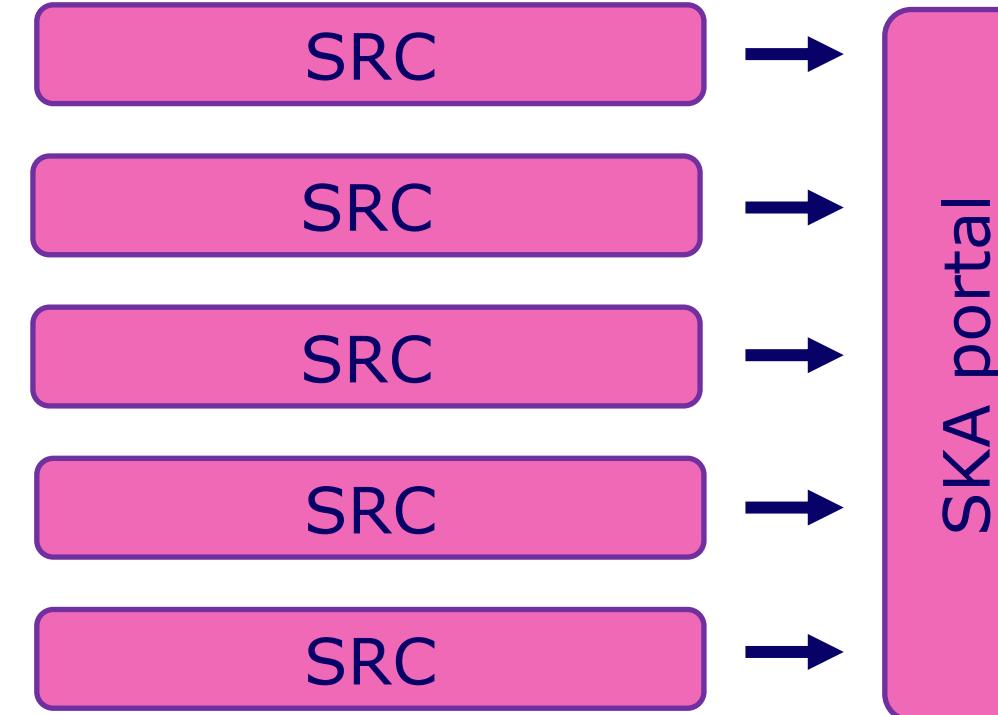


SKA data processor prototype,
Cambridge, UK

5 + 9 Tb/s
Approx 300
high definition
movies per
second!

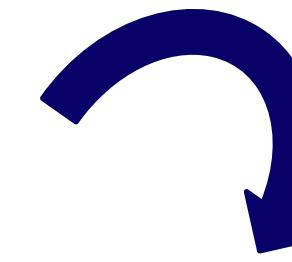
600 PB/yr

SRC: SKA Regional Centre



Distributed facilities

User data
products up to
TBs in size



Key Science Projects

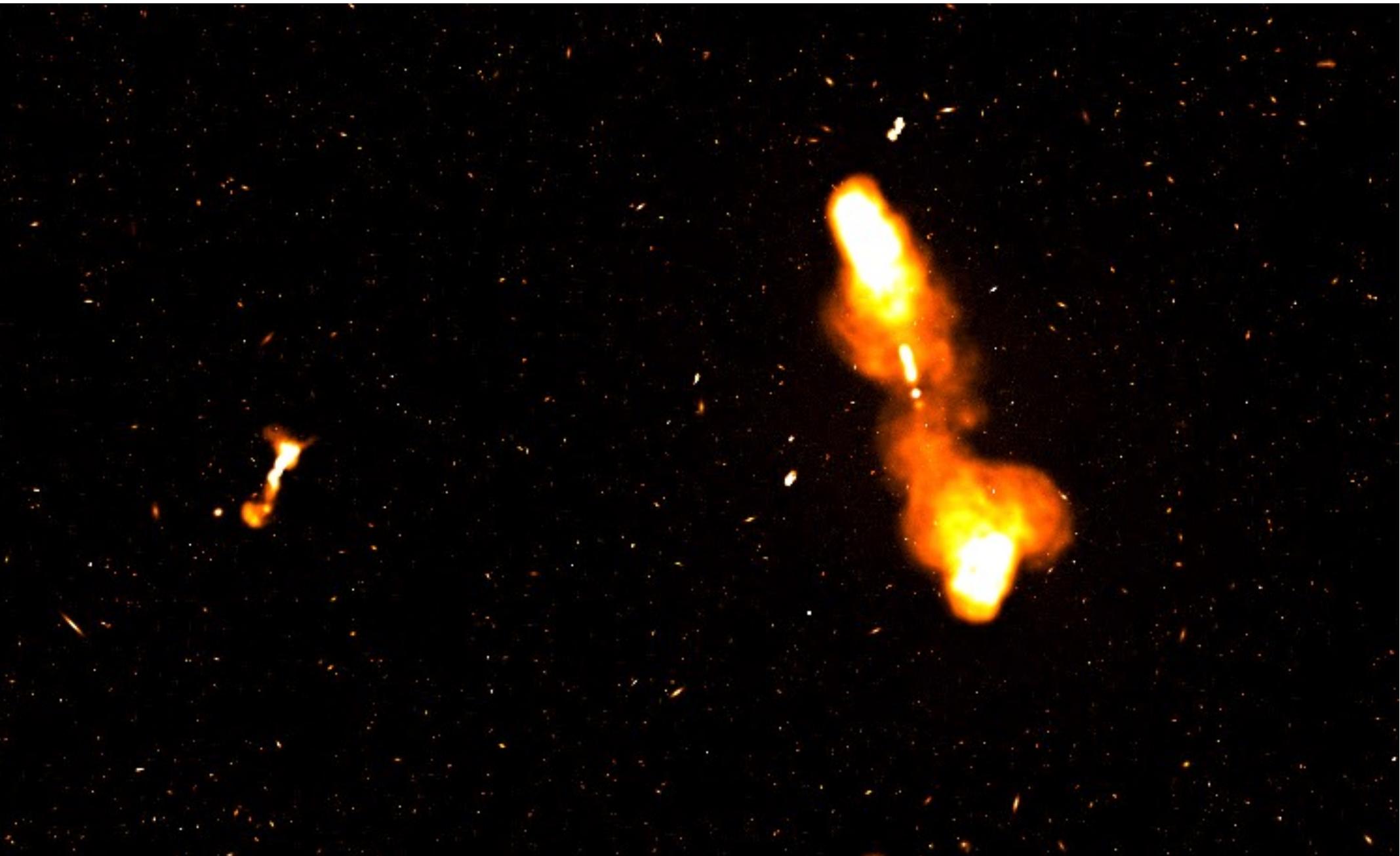


Scientific analysis



SKA Science Data Challenges (SDCs)

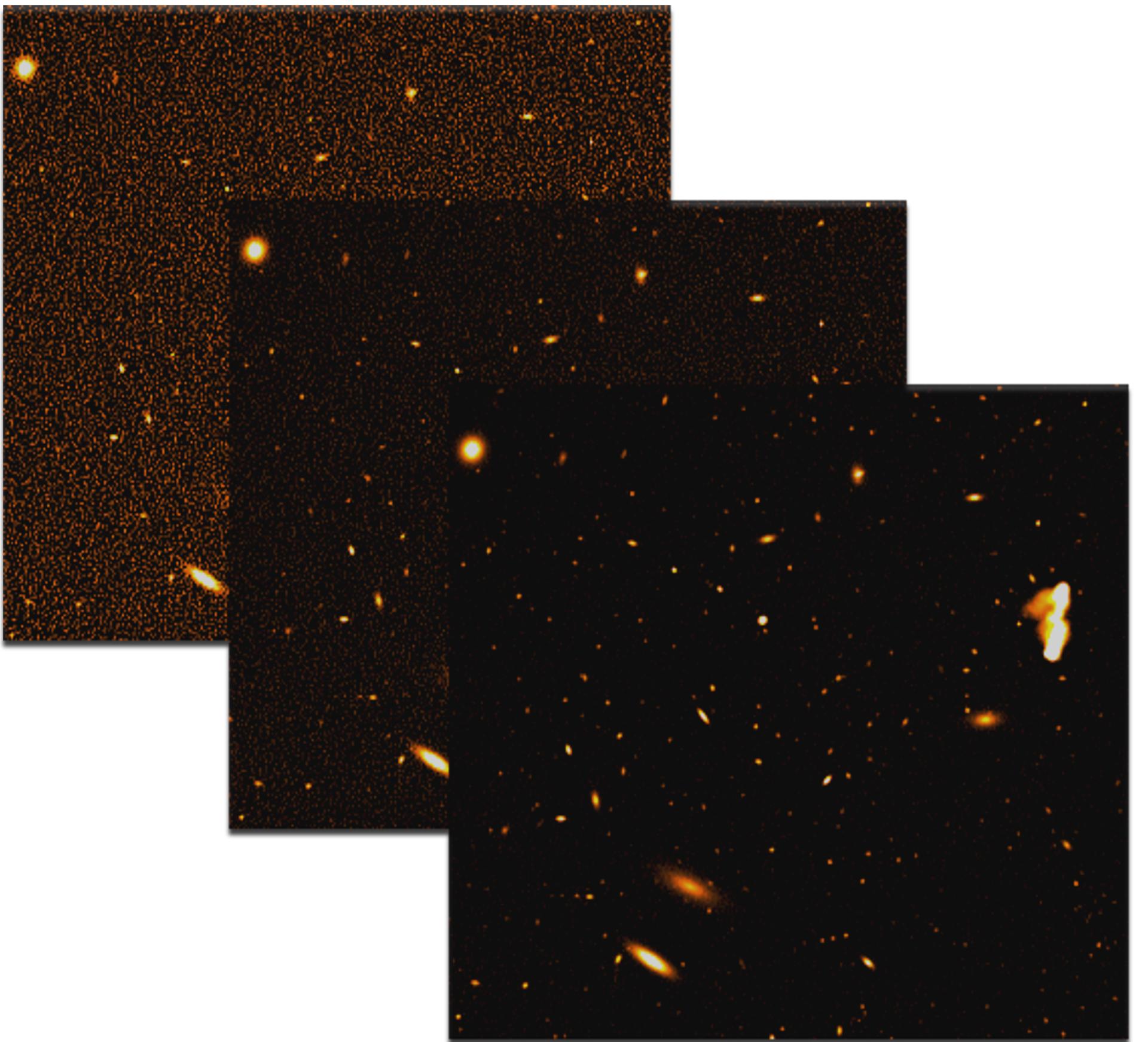
- Primary goals:
 - Familiarise the science community with size and complexity of SKA data
 - Drive the development of analysis techniques
 - Support the design of future SKA observations
- Additional benefits:
 - Familiarise the science community with data access models
 - Test SKA Regional Centre prototyping
 - Encourage best practices for Open Science and reproducibility



Science Data Challenge 1

Continuum emission

- **Continuum emission** images, simulating observations for SKA MID Bands 1, 2 and 5
- Images populated by star forming galaxies (**SFGs**) and active galactic nuclei (**AGN**)
- 3 telescope **integrations** each: 8, 100 and 1000h
- High telescope sensitivity → highly **crowded** images
- **The challenge:** to **find and characterise sources**



Zoom-in of the 1.4 GHz maps, showing the same region of the sky with different telescope integrations: 8, 100, 1000 h from left.



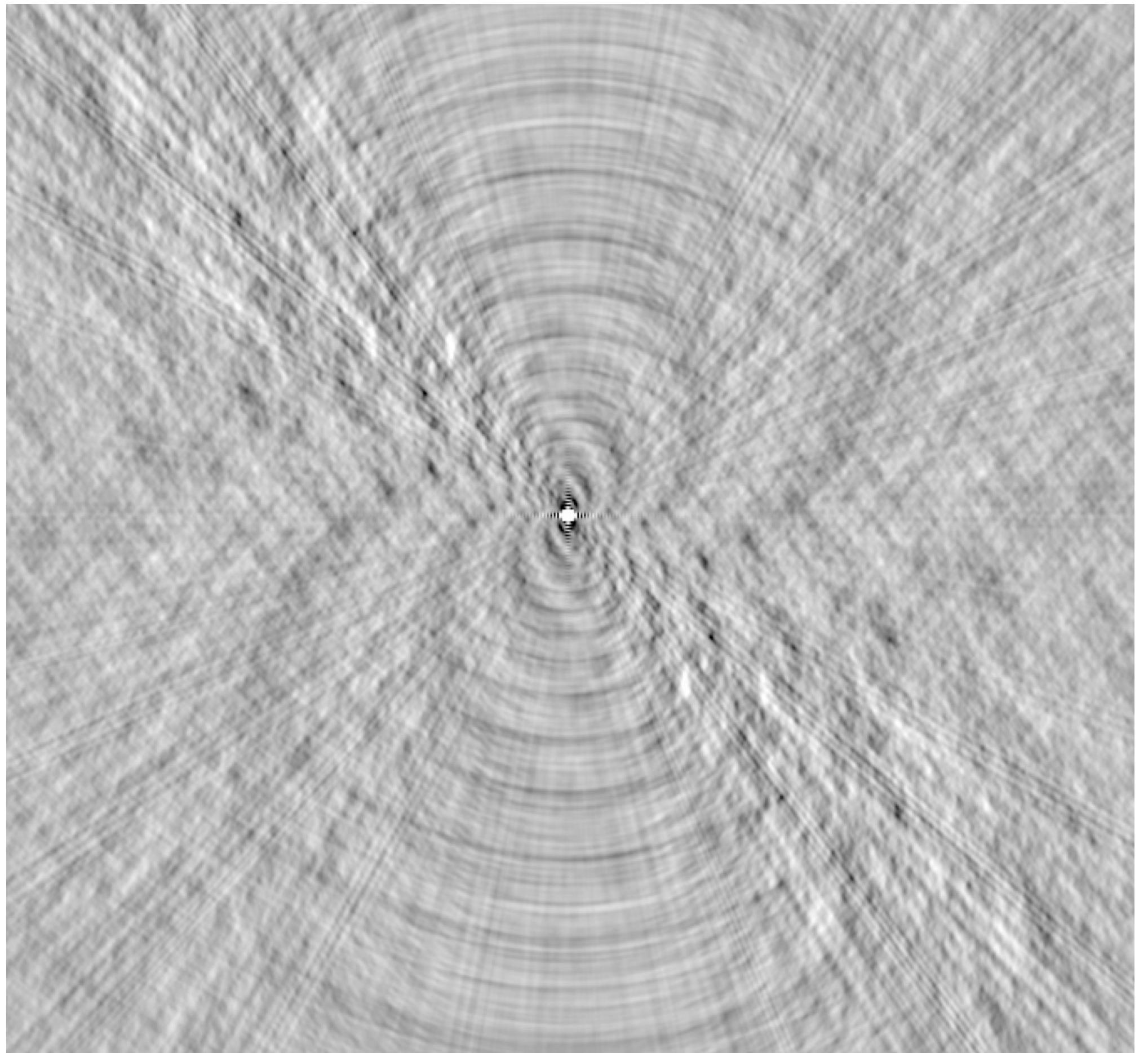
SKA characteristics

SKA-unique features of the data products:

- In the image plane, not visibilities
- “Benign” dirty beam
- Deconvolved down to 8h exposures
- Very deep -> towards confusion limit
- Very large number of sources to detect and classify

Data product specifications:

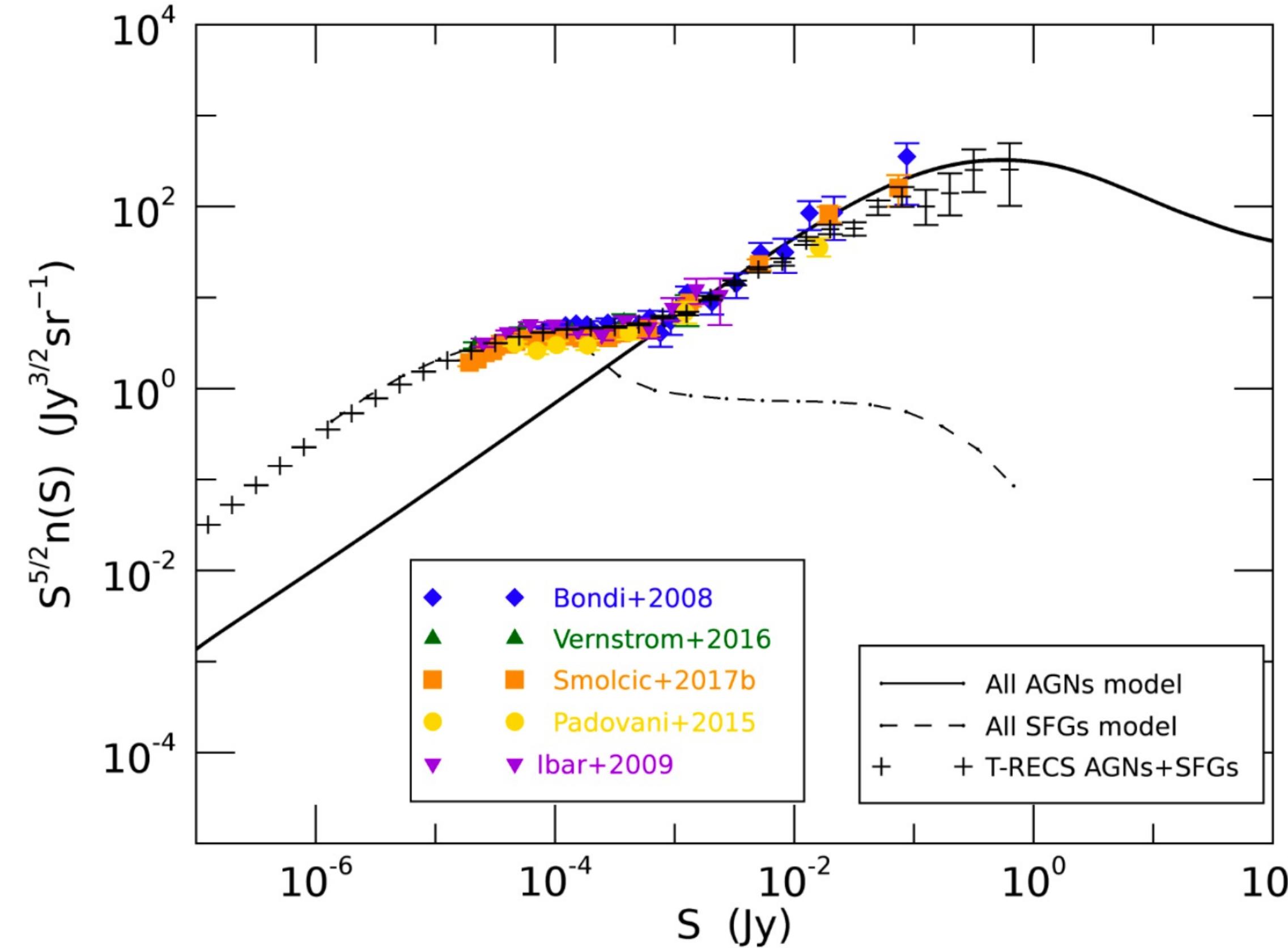
- 25 square degrees
- 1.5, 0.60 and 0.0913 arcsec FWHM
- Dirty beam sidelobes 4×10^{-4}



SKA MID 1.4 GHz beam

Continuum catalogue

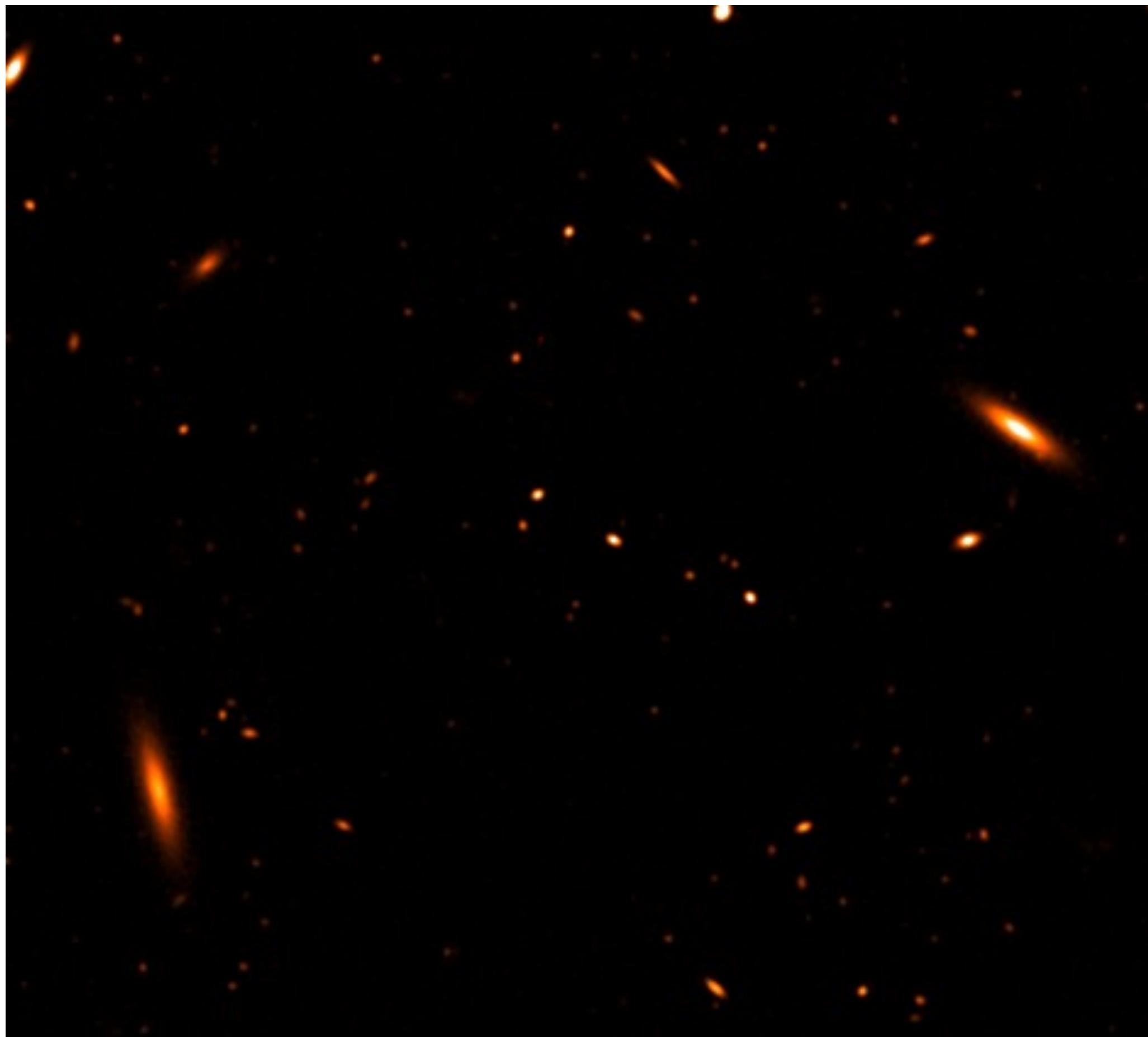
The Tiered Radio Extragalactic Continuum Simulation (T-RECS) Bonaldi+ 2019



1.4 GHz differential source counts

- Good agreement with source counts at the sub-mJy level
- Polarisation
- Realistic clustering

Source morphology



SFGs: Exponential Sersic profiles
Flat spectrum AGN: Gaussians and points



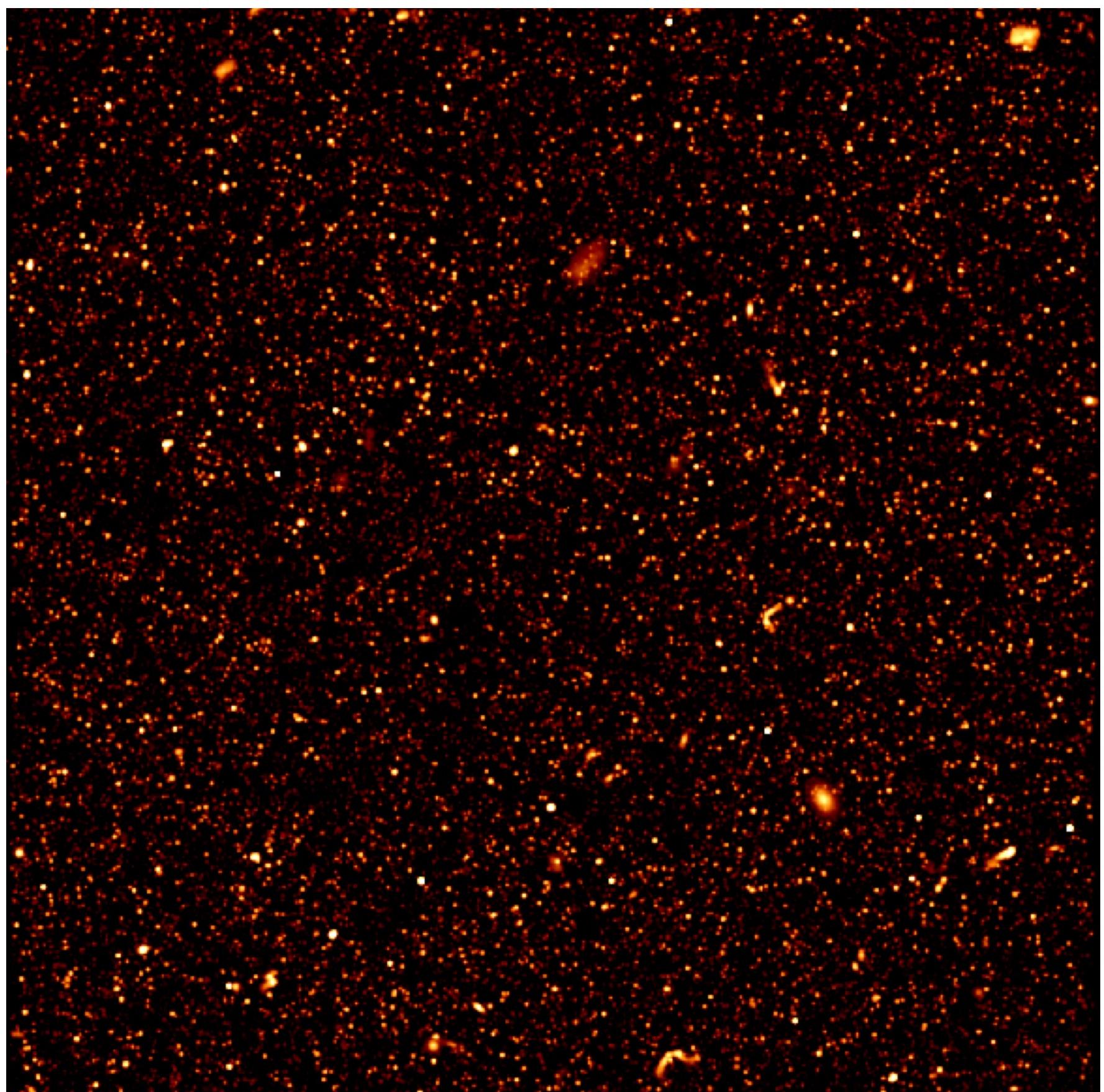
Steep spectrum AGN: DRAGNs
database, University of Manchester



Scoring method

- **Cross-matching** of source positions between submission catalogue with truth catalogue
- **Crowded field** necessitates use of size and flux information in cross-matching
- Final overall score: $G_{\text{tot}} = \frac{B_{\nu 1}}{\text{FoV}_{\nu 1}} + \frac{B_{\nu 2}}{\text{FoV}_{\nu 2}} + \frac{B_{\nu 3}}{\text{FoV}_{\nu 3}}$
- Weighted matches minus **false positives**: $B = \tilde{N}_m - N_f$
- Weighted matches are determined from **property accuracies** such that:

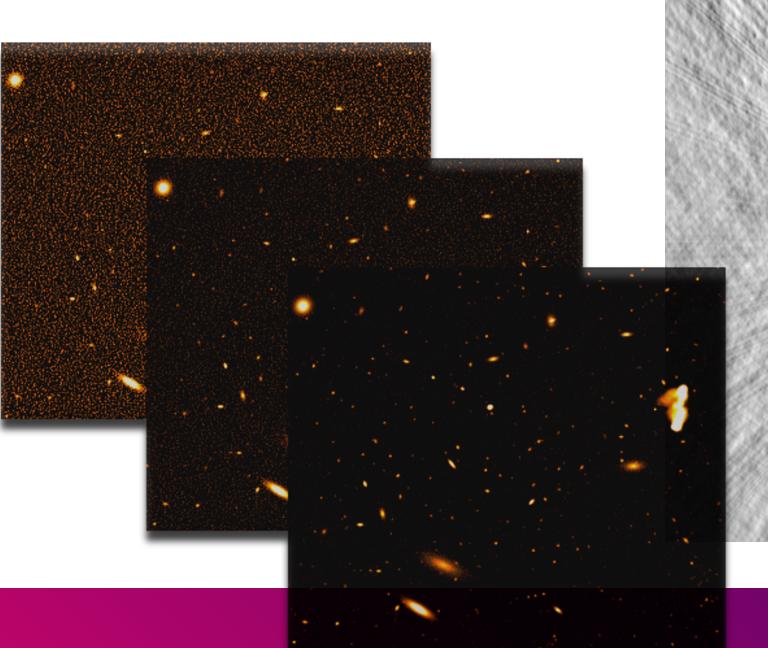
$$\tilde{N}_m = \sum_{i=1}^{N_m} w_i \leq N_m,$$



Portion of the 1.4 GHz continuum data product

Data products and scoring package

- Pip-installable package available from
<https://pypi.org/project/ska-sdc/>
- Data products for download at
<https://astronomers.skatelescope.org/>
- 9 continuum sky images
- Synthesized beam and primary beam
- Truth catalogues



8685975	-1.23265946	-29.8048365	-1.23118271	-29.5984961	2.66418e-07	0.89942311	316.865	218.684	134.438	1	1		
14851711	-0.18805235	-30.1882153	-0.18781235	-30.1882153	1.21204e-06	0.89942311	27.978	16.491	289.730	1	1		
15232796	-0.24582253	-29.5991458	-0.24582253	-29.5991458	1.24582253	0.81220083	127.777	56.439	39.345	1	1		
2134326	0.80672383	-29.62682746	0.80672383	-29.62682746	0.88840969	-0.398446e-07	0.39704477	144.553	84.979	39.344	1	1	
21810208	-0.24582253	-29.5991458	-0.24582253	-29.5991458	1.24582253	0.81220083	127.777	56.439	39.345	1	1		
24748488	1.-2.29823392	-29.59832587	1.-2.29823392	-29.59832587	1.22237675	0.58159836	1.23247e-07	0.89856582	138.952	99.088	323.566	1	1
29858881	1.-2.29823392	-29.59832587	1.-2.29823392	-29.59832587	1.24916764	-0.36138234	1.62734e-08	0.81191398	143.985	99.837	157.783	1	1
30000000	1.-2.29823392	-29.59832587	1.-2.29823392	-29.59832587	1.24916764	-0.36138234	1.62734e-08	0.81191398	143.985	99.837	157.783	1	1
4422538	0.14182825	-29.34311282	0.14206512	-29.34366989	5.5189e-06	0.89749287	141.822	31.848	265.385	1	1		
13987171	0.83484249	-38.2352432	0.82354994	-38.23537766	1.26124e-08	0.89937579	174.685	129.588	177.747	1	1		
15527955	0.76481771	-38.2352432	0.76481771	-38.2352432	0.82354994	-0.2352432	209.144	191.730	191.452	1	1		
15983868	-0.82718248	-29.8869386	-0.82718248	-29.8869386	0.89852217	-0.89852217	0.89852217	0.89852217	242.875	124.447	19.617	1	1
22617986	0.76481771	-38.2352432	0.76481771	-38.2352432	0.82354994	-0.2352432	209.144	191.730	191.452	1	1		
29587911	0.76481771	-38.2352432	0.76481771	-38.2352432	0.82354994	-0.2352432	209.144	191.730	191.452	1	1		
30000000	0.76481771	-38.2352432	0.76481771	-38.2352432	0.82354994	-0.2352432	209.144	191.730	191.452	1	1		

ska-sdc 2.0.0

pip install ska-sdc

Latest version

Released: Sep 8, 2021

A package providing tools for the SKA Science Data Challenges.

Navigation

Project description

Release history

Download files

Project description

Science Data Challenge Scoring API

This repository contains the code used to score submissions for SKA's Science Data Challenges (SDCs). To date there are two such challenges, SDC1 (run in 2019) and SDC2 (running February-July 2021), and with each having similar methods of evaluating submissions. Both SDCs challenged participants to identify and characterise sources in synthetic radio images. The source catalogues the participants produced (called the submission catalogues) can then be compared to the real source properties used in creating the synthetic images (called the truth catalogues) to determine which solutions achieve the best result.

Science Data Challenge 1

Main findings

- Very **crowded** skies demand new approaches
- Variety of methods including **latest machine learning** techniques
- **Complementarity** of methods: tendency to score well either on finding galaxies or measuring them

MNRAS, Volume 500, Issue 3, January 2021, Pages 3821–3837

Square Kilometre Array Science Data Challenge 1: analysis and results

A. Bonaldi,^{1,2*} T. An³, M. Brüggen⁴, S. Burkutean⁵, B. Coelho⁶, H. Goodarzi⁷, P. Hartley¹, P. K. Sandhu⁸, C. Wu⁹, L. Yu¹⁰, M. H. Zholideh Haghghi⁷, S. Antón^{11,6}, Z. Bagheri^{7,12}, D. Barbosa⁶, J. P. Barraca^{6,13}, D. Bartashevich⁶, M. Bergano⁶, M. Bonato⁵, J. Brand⁵, F. de Gasperin⁴, A. Giannetti⁵, R. Dodson⁹, P. Jain⁸, S. Jaiswal³, B. Lao³, B. Liu¹⁰, E. Liuzzo⁵, Y. Lu³, V. Lukic⁴, D. Maia¹⁴, N. Marchili⁵, M. Massardi⁵, P. Mohan³, J. B. Morgado¹⁴, M. Panwar⁸, Prabhakar⁸, V. A. R. M. Ribeiro^{6,15}, K. L. J. Rygl⁵, V. Sabz Ali⁷, E. Saremi⁷, E. Schisano¹⁶, S. Sheikhnezami^{17,7}, A. Vafaei Sadr¹⁸ A. Wong¹⁹, O. I. Wong^{9,21,20}

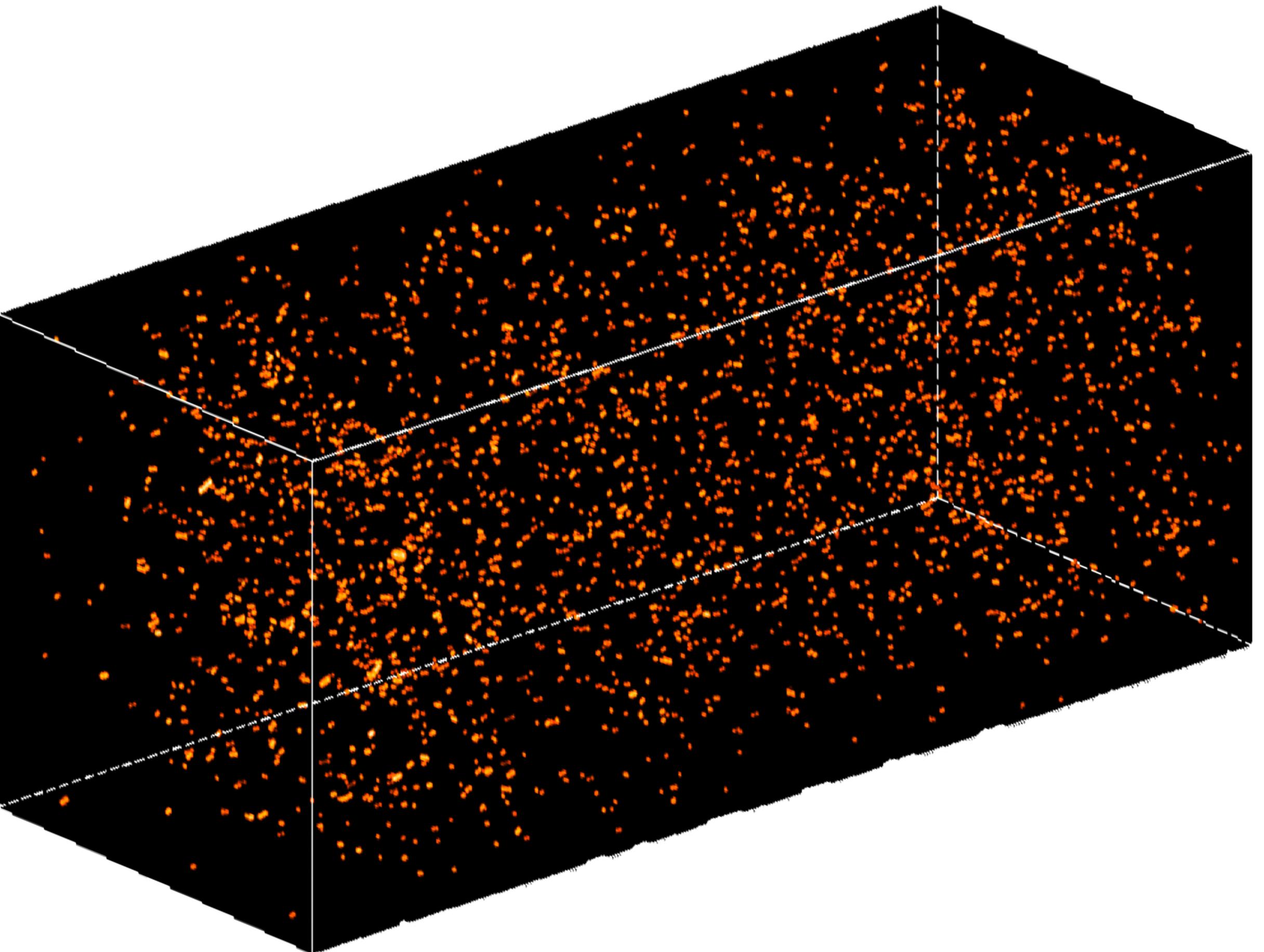
Affiliations are at the end of the paper



Science Data Challenge 2

HI emission

- A simulated 'datacube' of stacked images, each image representing a different frequency of light.
- Simulating galaxies up to **redshift 0.5**
- **The challenge:** finding galaxies and measuring them
- **Data volume = 1 TB**
- [SDC2 website](#)



Example simulated HI datacube



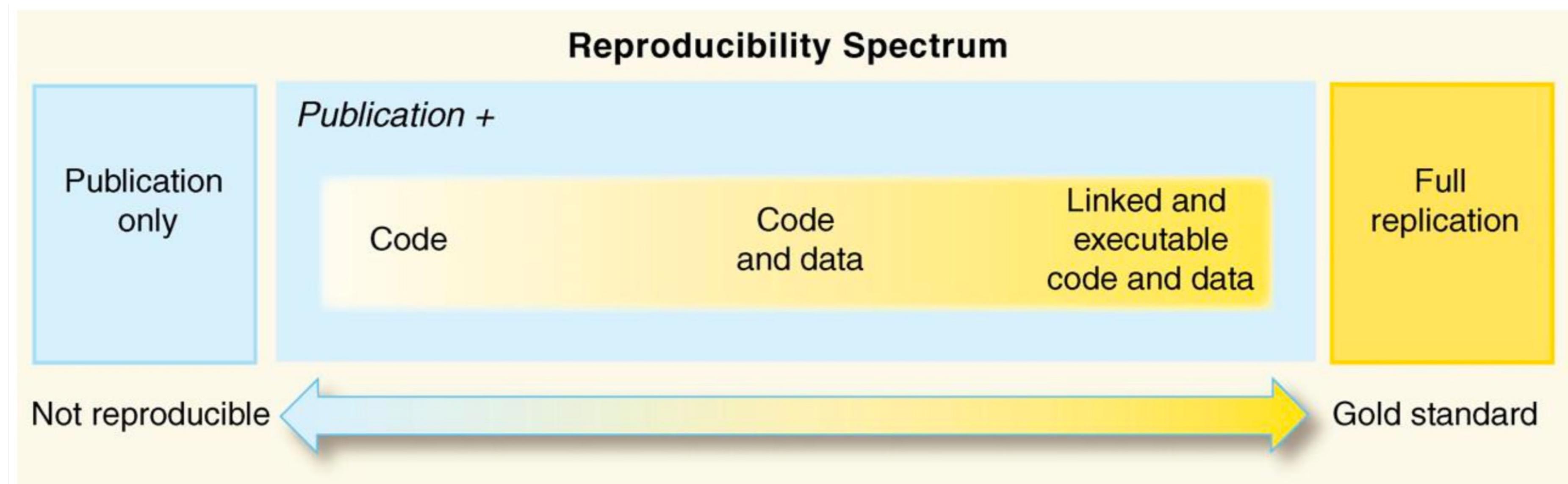
Reproducibility

In partnership
with the Software
Sustainability
Institute



www.software.ac.uk

- An essential part of the scientific method, **reproducibility** leads to better, more efficient science.
- **Reusability** generalises this principle to create software that can be adapted by others, allowing previous work to be built upon for the future: a key feature of Open Science
- SKA is committed to delivering on the **FAIR** principles for scientific data management



Credit: Rachael Ainsworth



SDC1 solution

- **Fully reproducible** solution by Alex Clarke (SKAO)
- Utilises **containerization** in order to package up all software and environment
- Demonstrates **best practices** in research and software development
- **Hands-on containers (beginners)** session later today

The screenshot shows a GitHub project page for "SDC1 Solution". The header includes the organization "ska-telescope", the repository "Science Data Challenges", and the specific branch "SDC1 Solution". The project ID is 19657157. It has 181 commits, 6 branches, 0 tags, 5.5 MB files, and 5.5 MB storage. A star icon indicates 0 stars. The description is "A portable solution to SKA's first Science Data Challenge". Below the header, there is a "README.md" file section and a "Science Data Challenge 1 Solution Workflow" section. The workflow section describes the challenge, provides instructions for setting up the environment, and notes that the document contains instructions for running a workflow script.

ska-telescope > Science Data Challenges > SDC1 Solution > Details

SDC1 Solution Project ID: 19657157

181 Commits 6 Branches 0 Tags 5.5 MB Files 5.5 MB Storage

A portable solution to SKA's first Science Data Challenge

README.md

Science Data Challenge 1 Solution Workflow

The SKA Science Data Challenge 1 (SDC1, <https://astronomers.skatelescope.org/ska-science-data-challenge-1/>) tasked participants with identifying and classifying sources in synthetic radio images.

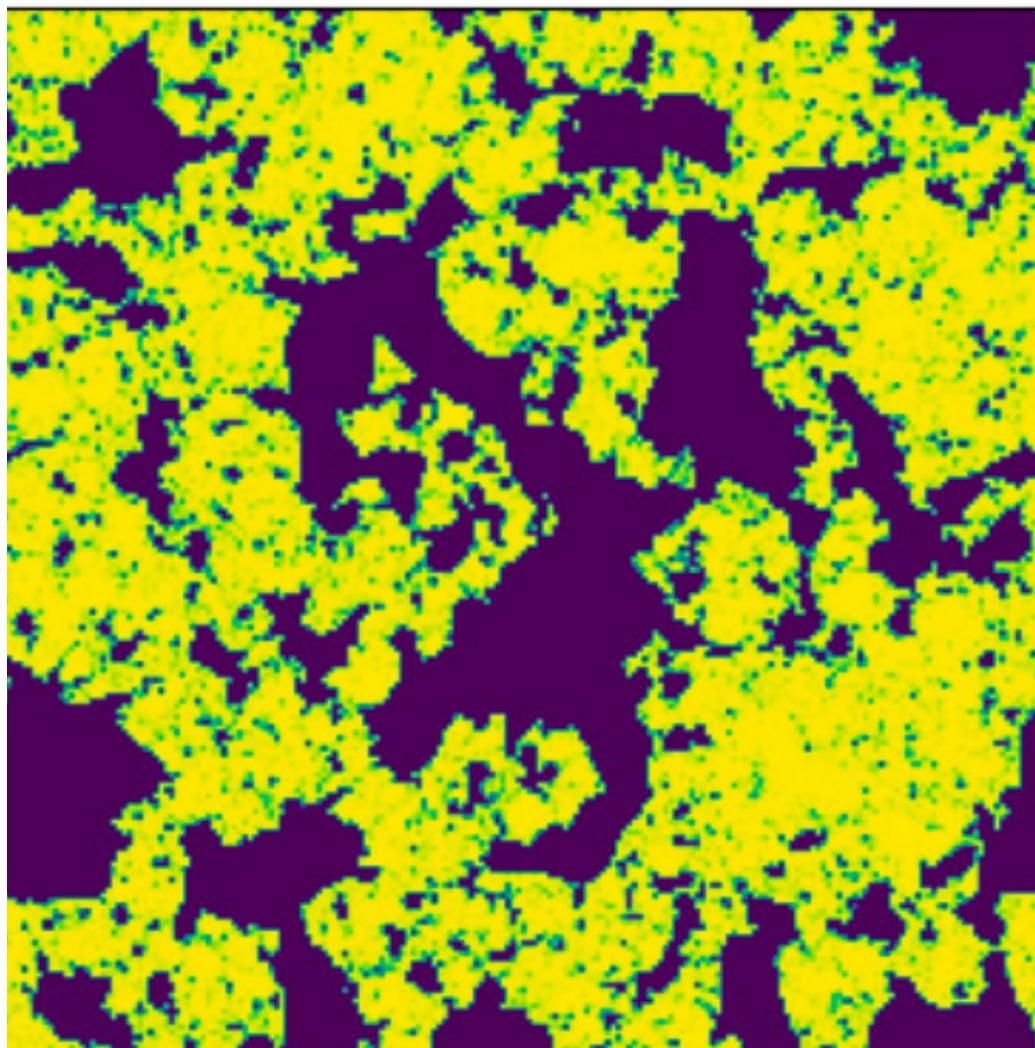
Here we present an environment and workflow for producing a solution to this challenge that can easily be reproduced and developed further.

Instructions for setting up the (containerised) environment and running a simple workflow script using some Python helper modules are provided in this document.

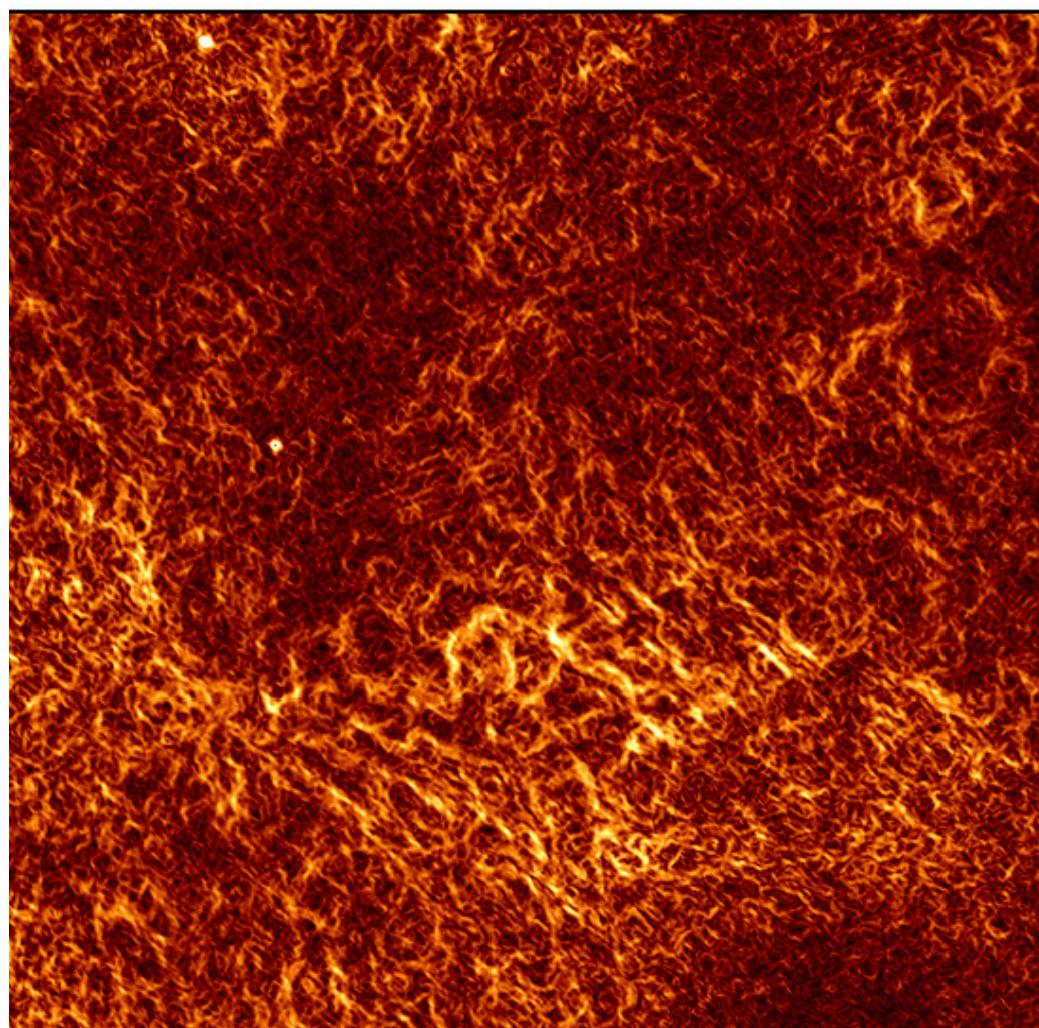


Future data challenges

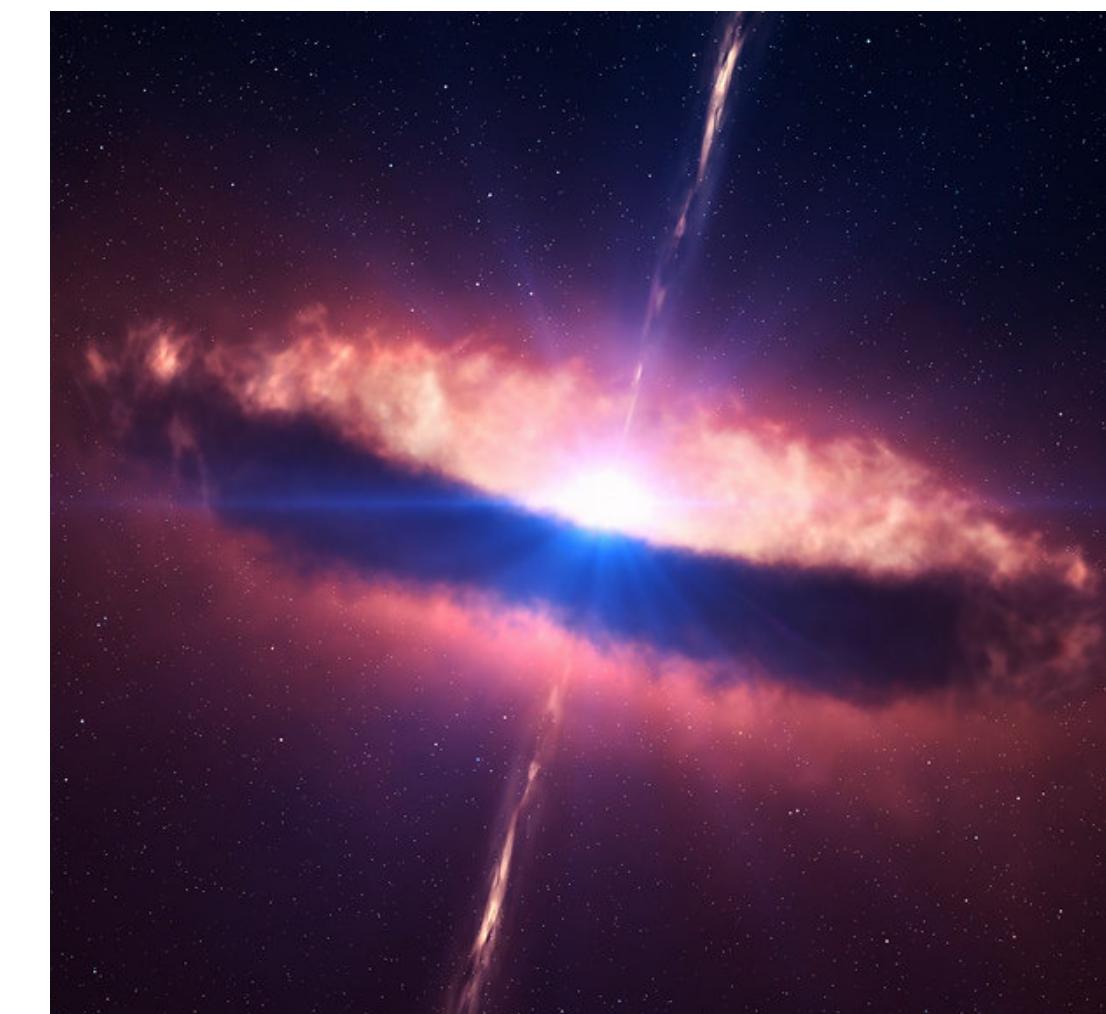
- Epoch of Reionisation, cosmic magnetism, pulsars and quasars
- Planning underway: stay tuned for news and updates!



Ionised EoR field at redshift 10.
Credit: Eunseong Lee



'Snakes' of cosmic magnetism.
Credit: B. Gaensler et al.



Quasar schematic. Credit: NASA

