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

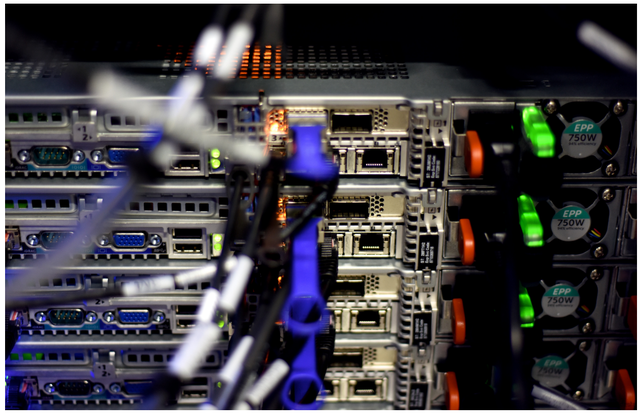
The SKA data journey

SKA LOW

SKA MID

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

SKA data processing



SKA data processor prototype, Cambridge, UK

600 PB/yr

SRC: SKA Regional Centre

SRC

SRC

SRC

SRC

SRC

SKA portal

Distributed facilities

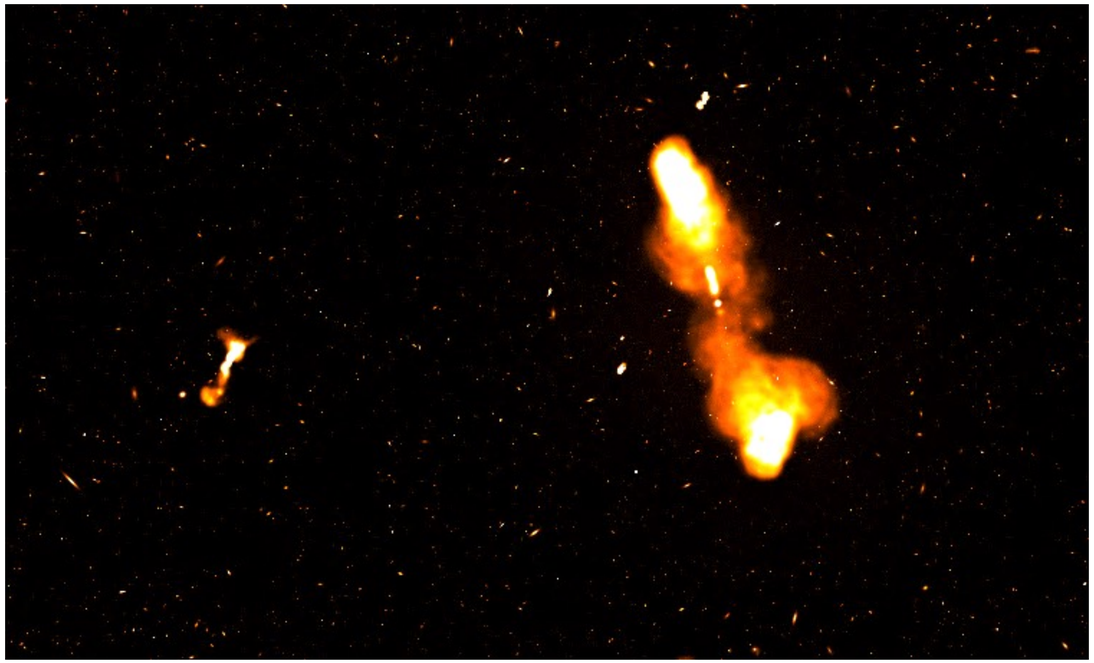
User data products up to **TB**s in size

Key Science Projects

Scientific analysis

SKA Science Data Challenges (SDCs)

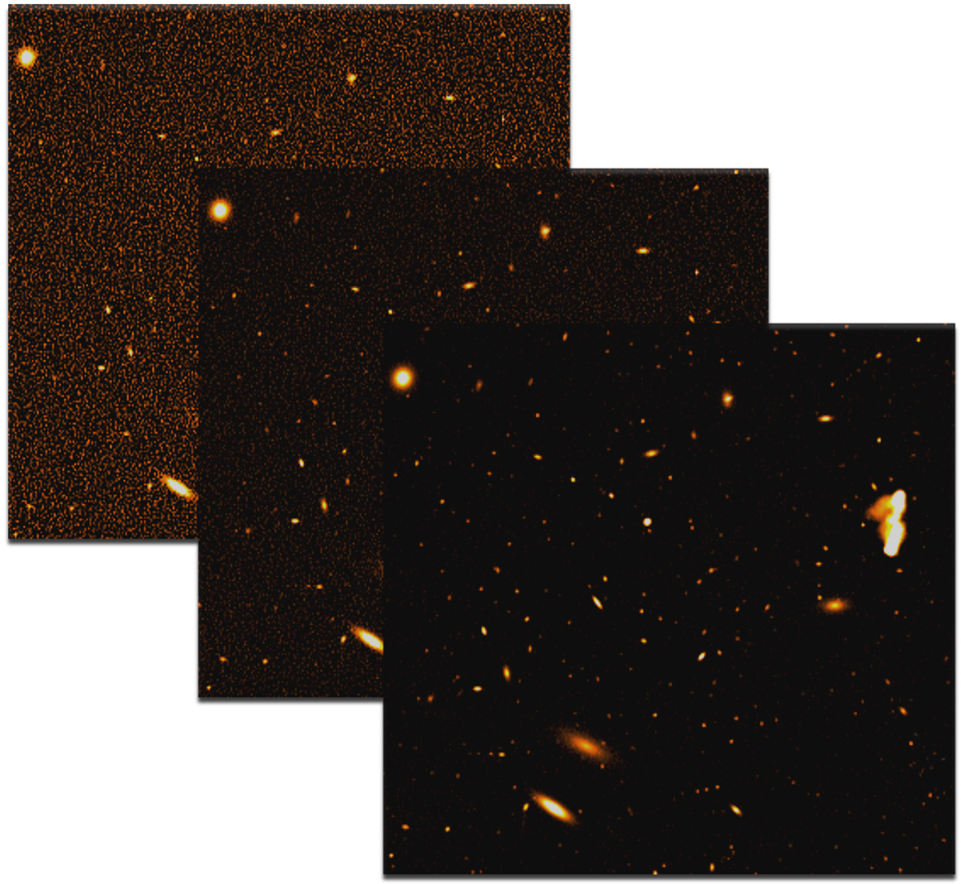
- Primary goals:
  - Familiarise the science community with size and complexity of SKA data
  - Drive the development of analysis techniques
  - Support the design of future SKA observations
- Additional benefits:
  - Familiarise the science community with data access models
  - Test SKA Regional Centre 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.

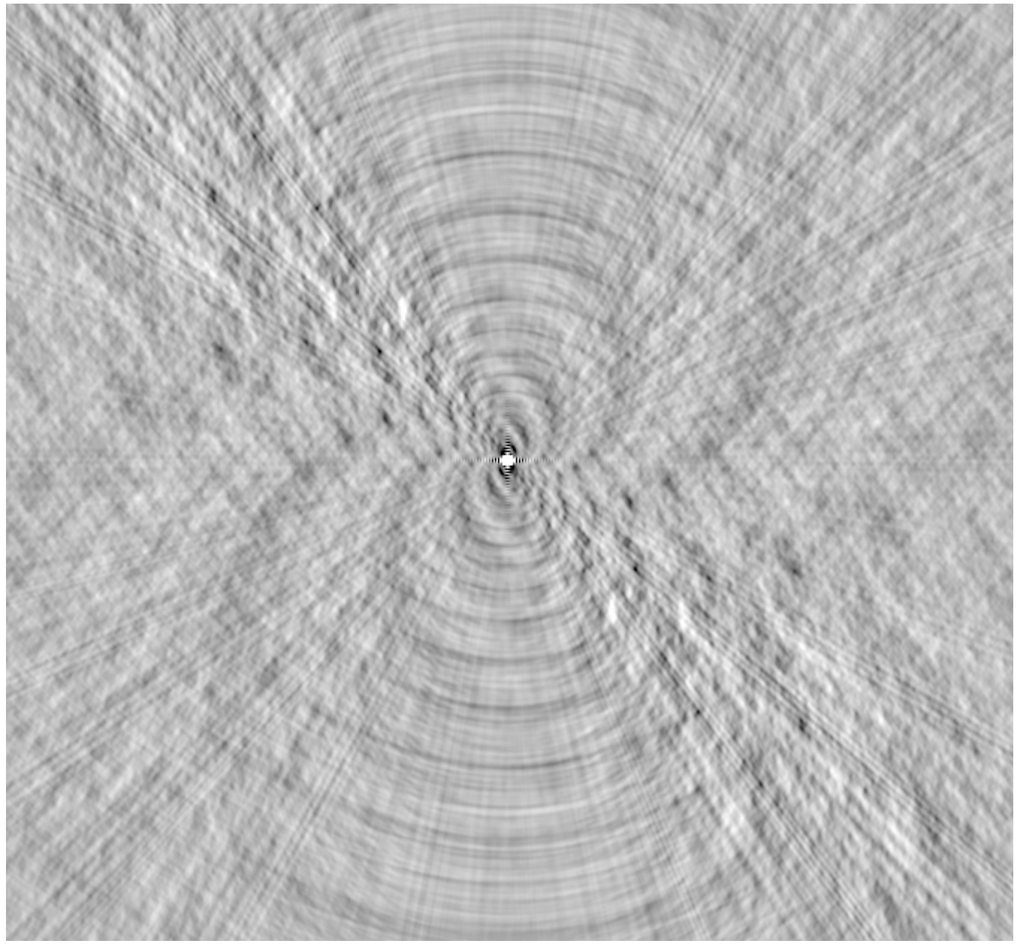
SKA characteristics

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:

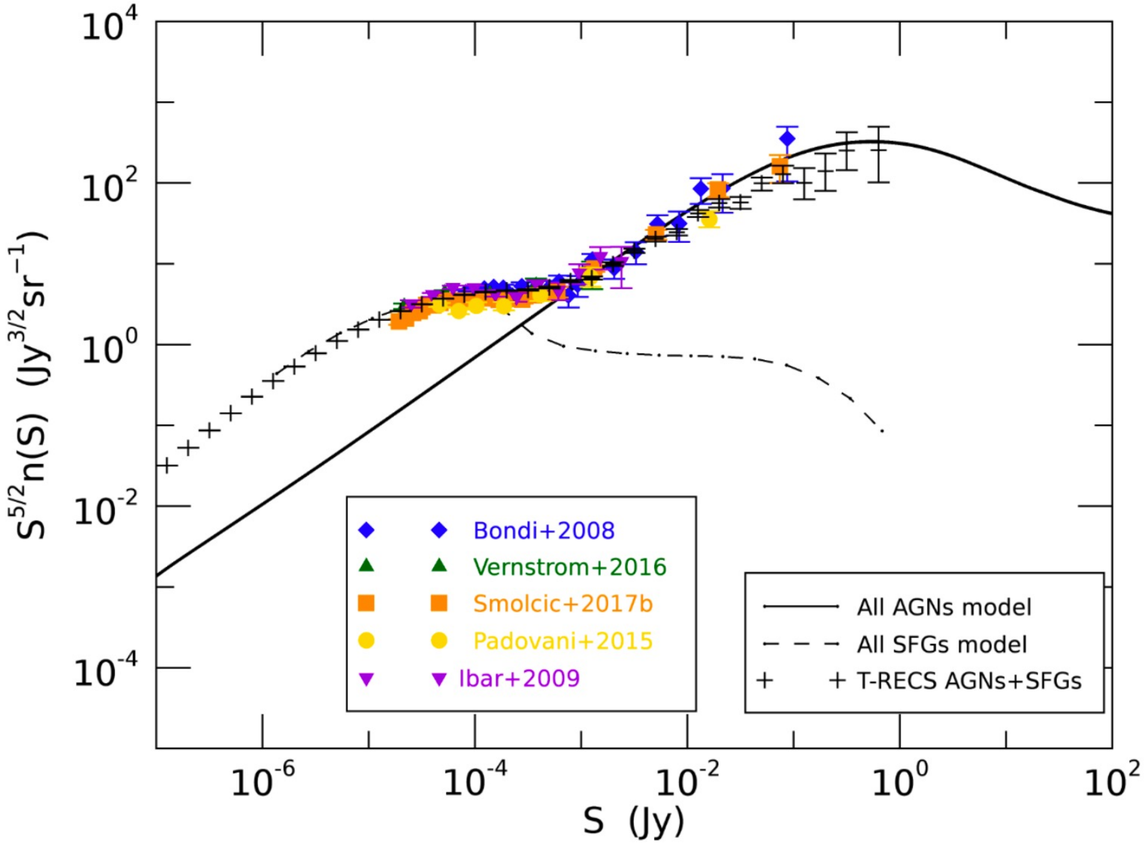
- *25 square degrees*
- *1.5, 0.60 and 0.0913 arcsec FWHM*
- *Dirty beam sidelobes $4x10^{-4}$*



SKA MID 1.4 GHz beam

Continuum catalogue

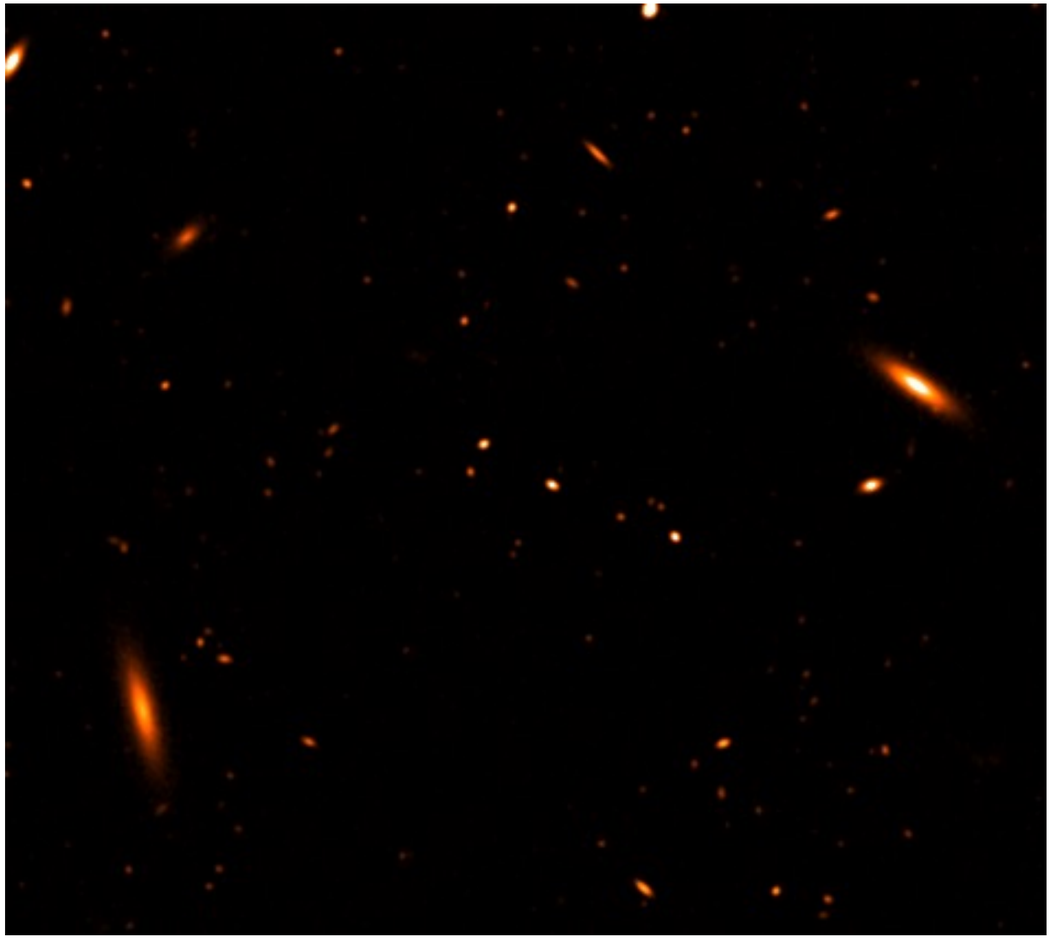
The Tiered Radio Extragalactic Continuum Simulation (T-RECS) *Bonaldi+ 2019*



1.4 GHz differential source counts

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

Source morphology

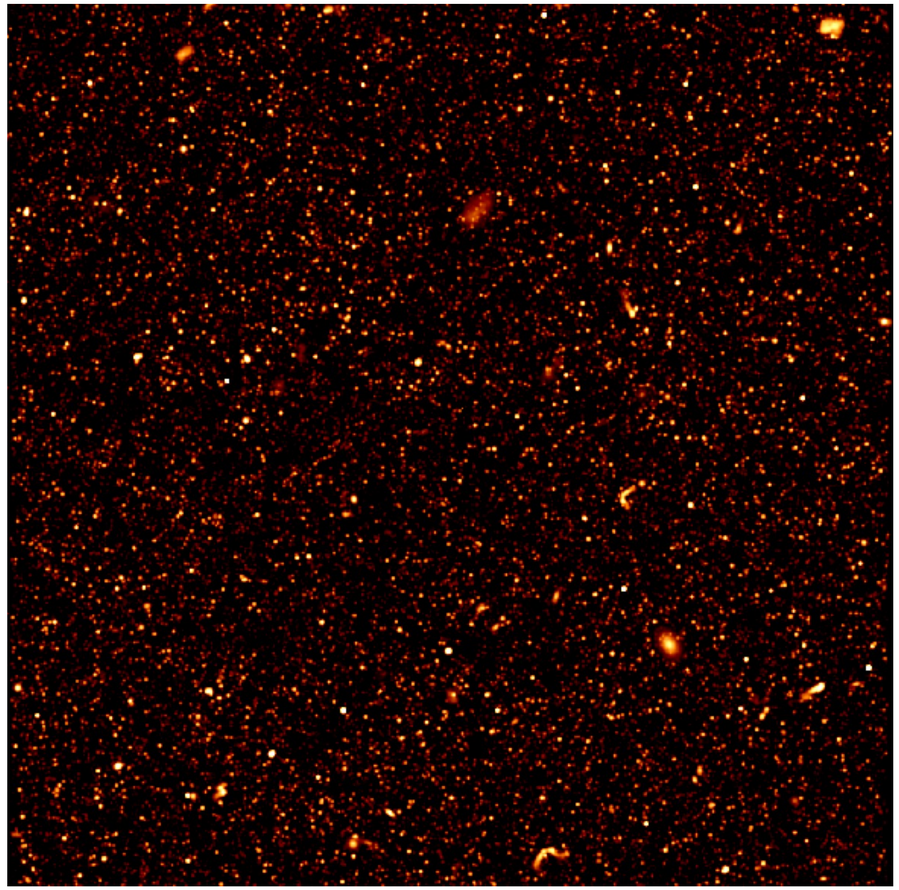


SFGs: Exponential Sersic profiles
Flat spectrum AGN: Gaussians and points

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: $$G_{\rm tot} = \frac{B_{\nu 1}}{{\rm FoV}_{\nu 1}} + \frac{B_{\nu 2}}{{\rm FoV}_{\nu 2}} + \frac{B_{\nu 3}}{{\rm FoV}_{\nu 3}}$$
- Weighted matches minus **false positives**: $B = \tilde{N}_{\rm m} - N_{\rm f}$
- Weighted matches are determined from **property accuracies** such that: $$\tilde{N}_{\rm m} = \sum_{i=1}^{N_{\rm m}} w_i \leq N_{\rm m},$$



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

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 links

- Homepage

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

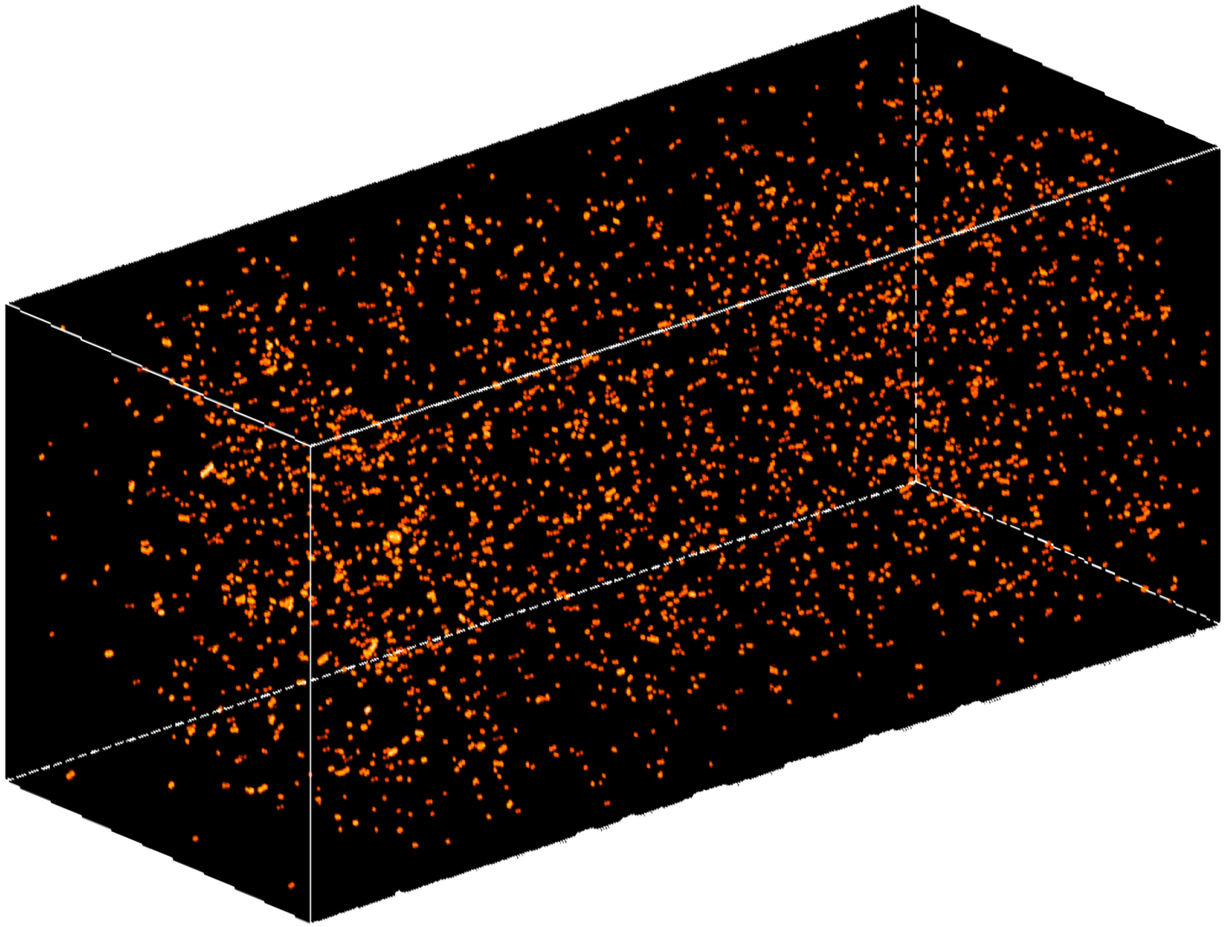
Square Kilometre Array Science Data Challenge 1: analysis and results

A. Bonaldi,[1,2⋆] T. An[3], M. Brüggen[4], S. Burkutean[5], B. Coelho[6], H. Goodarzi[7], P. Hartley[1], P. K. Sandhu[8], C. Wu[9], L. Yu[10], M. H. Zhoolideh Haghighi[7], S. Antón[11,6], Z. Bagheri[7,12], D. Barbosa[6], J. P. Barraca[6,13], D. Bartashevich[6], M. Bergano[6], M. Bonato[5], J. Brand[5], F. de Gasperin[4], A. Giannetti[5], R. Dodson[9], P. Jain[8], S. Jaiswal[3], B. Lao[3], B. Liu[10], E. Liuzzo[5], Y. Lu[3], V. Lukic[4], D. Maia[14], N. Marchili[5], M. Massardi[5], P. Mohan[3], J. B. Morgado[14], M. Panwar[8], Prabhakar[8], V. A. R. M. Ribeiro[6,15], K. L. J. Rygl[5], V. Sabz Ali[7], E. Saremi[7], E. Schisano[16], S. Sheikhnezami[17,7], A. Vafaei Sadr [18] A. Wong[19], 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



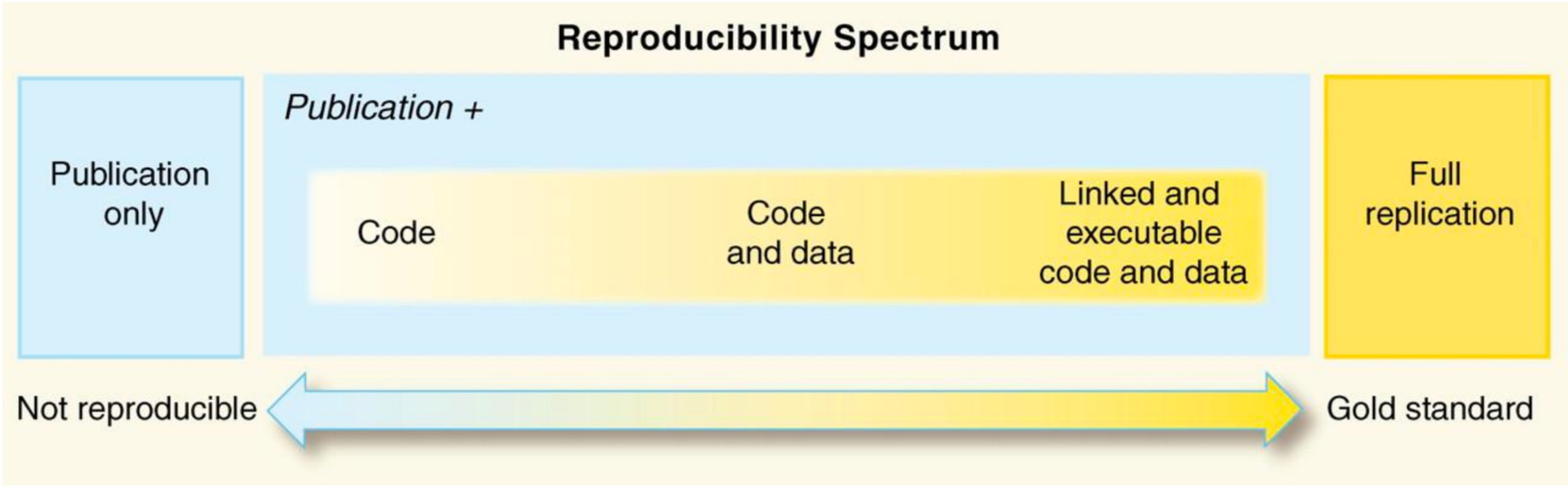
Example simulated HI datacube

Reproducibility

In partnership with the Software Sustainability Institute

www.software.ac.uk

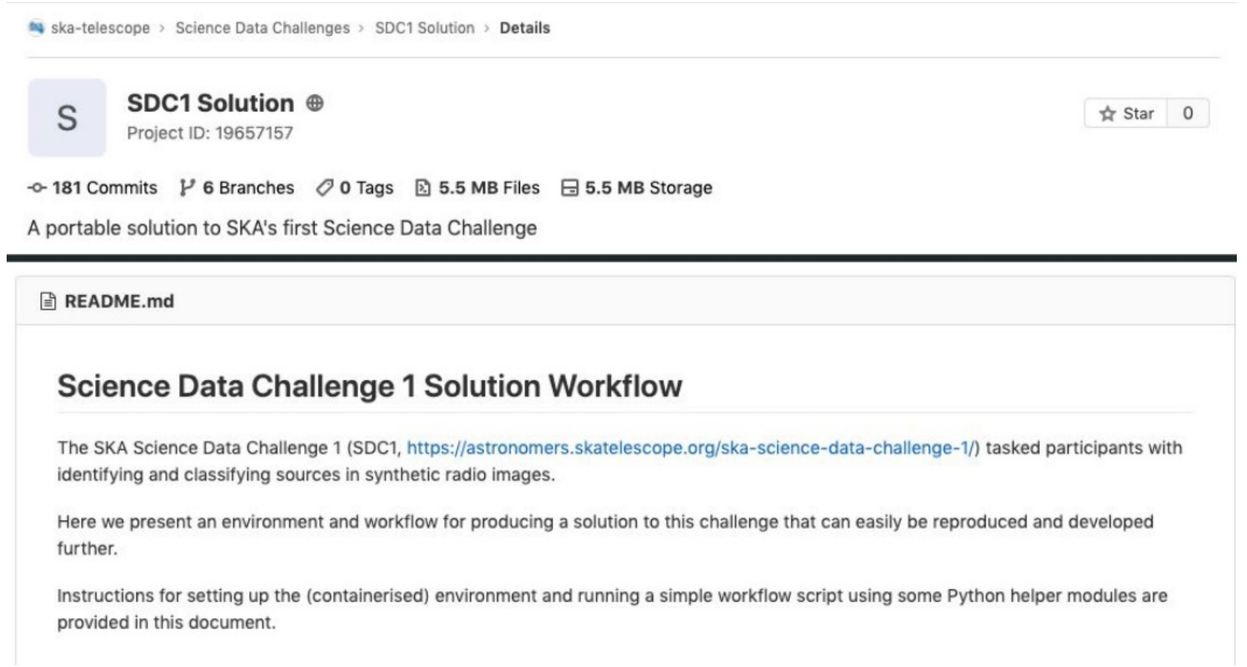
- An essential part of the scientific method, **reproducibility** leads to better, more efficient science.
- **Reusability** generalises this principle to create software that can be adapted by others, allowing previous work to be built upon for the future: a key feature of Open Science
- SKA is committed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



Credit: Rachael Ainsworth

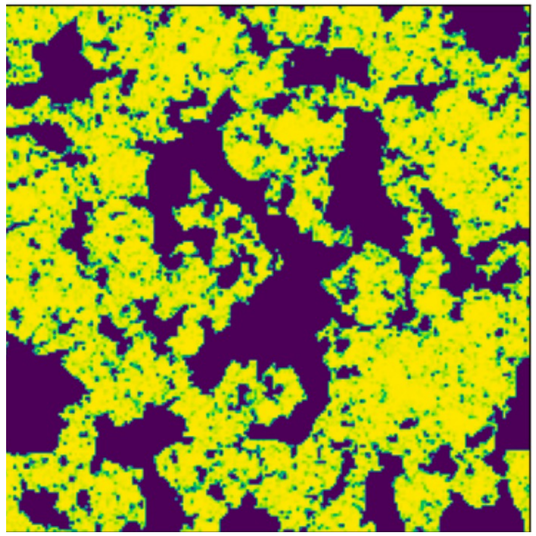
SDC1 solution

- **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

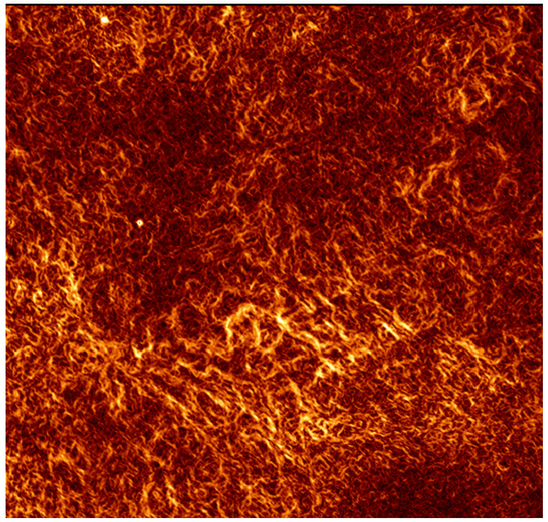


Future data challenges

- 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