**Introduction**

This document aims to provide information on how the SKA data challenge #1 has been produced and to set the challenge for the community.

Small tip: Google the words that you are unfamiliar with.

**What are the SKA Data Challenges**

SKA Science Data Challenges will be regularly issued to the community as part of the science preparatory activities. The purpose of these challenges is to inform the development of the data reduction workflows, to allow the science community to get familiar with the standard products the SKA will deliver, and optimize their analyses to extract science from them.

These challenges may include real data from operating radio facilities or simulated SKA data. Data at different stages along a data reduction workflow have been broadly categorized into four main Data Layers (DL). So far, four data challenges have been planned, two of which have happened already. The focus of this document will be on the first data challenge (Data challenge #1).

**SKA Science Data Challenge #1 Dataset Description**

The SKA Science Data Challenge #1 (SDC1) release consists of 9 image files in FITS format. Each file is a simulated SKA continuum image in total intensity at three frequencies:

1. 560 MHz, representative of SKA Mid Band 1
2. 1.4 GHz, representative of SKA Mid Band 2
3. 9.2 GHz, representative of SKA Mid Band 5

Notice that the original data challenge has nine sets, but we will work only on the above three sets.

Ancillary Data also accompany the datasets above; The Primary Beam and Synthesized Beam for each frequency band are provided as FITS format images. Further, it has a training set with specific coordinates to train the machine learning models combined with a “truth catalogue”.

**Simulation Model Description**

The list of objects that populate the simulated fields was generated with the T-RECS simulation code (RD1), yielding catalogs of AGNs and SFGs, with integrated flux densities, sky coordinates, and size and shape information.

Sources from these catalogs were injected into the simulated fields either as:

* an extended source, if its major axis size (defined below) is larger than 3 pixels;
* a compact source if its size is smaller than 3 pixels.

As for the source type, SFGs have a morphological model with an exponential Sersic profile. On the other hand, AGNs can be split into two populations, which can be described as follows:

* Steep-spectrum AGNs (AGNs-SS): exhibit the typical double-lobe FRI/FRII morphology.
* Flat-spectrum AGNs (AGNs-FS): exhibit a compact core component with a single lobe viewed end-on.

The three catalogs (SFGs, steep spectrum AGN, and flat spectrum AGN) of compact objects were added to the image as elliptical Gaussian components with their primary beam fits files. Noise and dirty points are also added to the simulated image to mirror the actual images that will come for the SKA.

**The Challenge Defined**

The SKA community is invited to retrieve the SDC1 images and undertake:

1. Source finding (RA, Dec) to locate the centroids and where appropriate the core positions
2. Source property characterisation:
   1. integrated primary-beam corrected flux density (the primary beams are provided)
3. Source population identification (one of AGN-SS, AGN-FS, SFG)

The accuracy of submitted results will be graded on:

1. Reliability and completeness of sources found
2. Accuracy of property characterisation
3. Accuracy of population identification
4. Overall response score based on the total number of real sources (less any false detections) found in the three 1000h images (the sum over the three frequency bands, while excluding the Training Area for each band defined in Section 4) multiplied by the fractional accuracy of the property characterisation and the population identification