Some lessons learnt from the SKA precursor large survey projects

Filippo Maccagni and Julia Healy

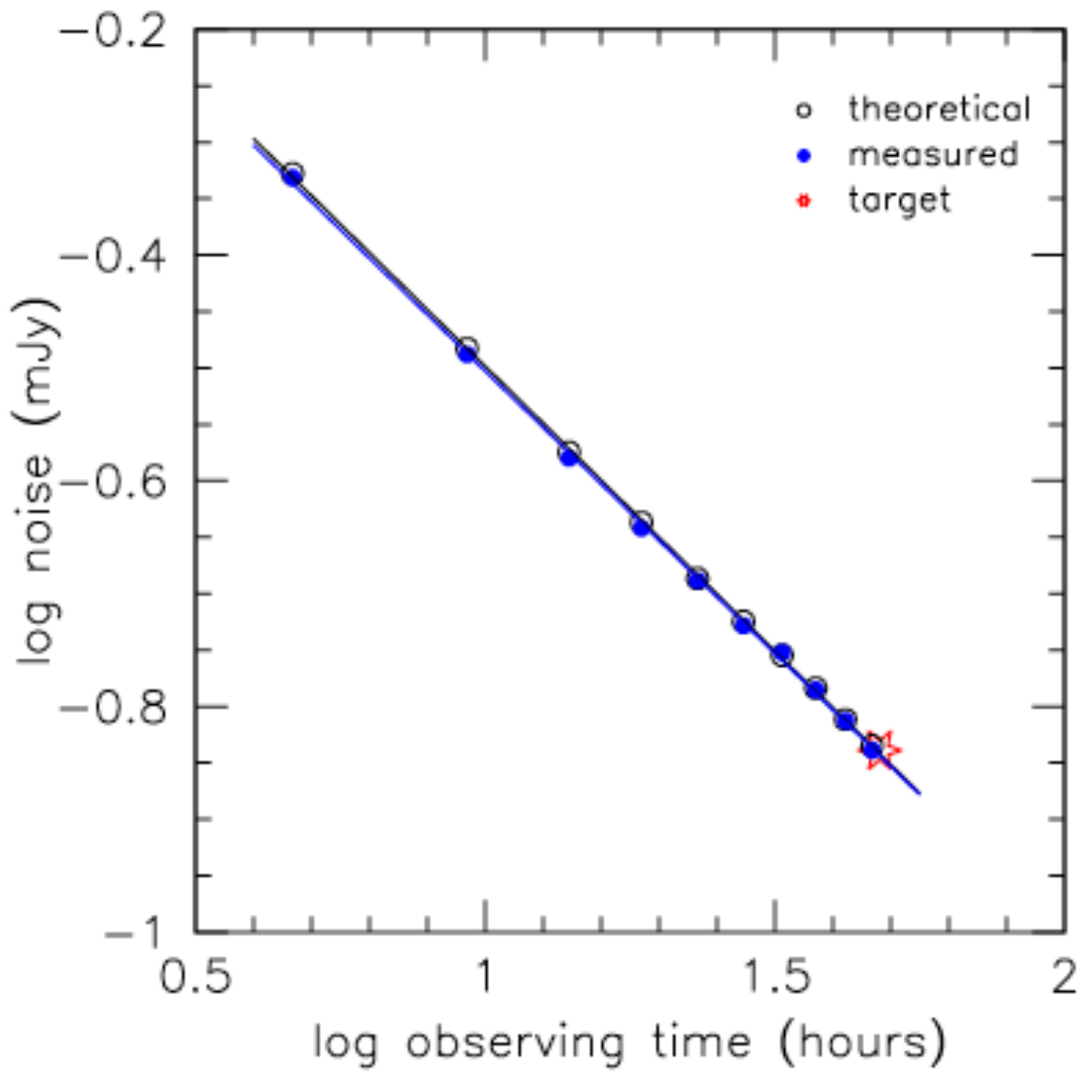
What data products we want for spectral line science

- **Calibrated visibilities, (non-gridded, not averaged as much as possible)**
  - keeping calibrated visibilities for a SKA MHONGOOSE survey (similar for MFS): 1500h, ~197 dishes, 32k channels at 1.6kHz resolution
    - Total observed data: ~50PB
    - Data to be stored: ~5PB
    - Manageable still on MeerGAS cluster located at ASTRON

- <u>At least for a limited time</u>: 1 year after observing?
  - As with every telescope
    - first years of observations will not provide the best quality data
    - it may be necessary to reprocess data
      - Re-observe means re-schedule projects delays
      - Conflict with 3-year contract projects/fellowships

Why

- Optimise the output of science projects, for which responsibility falls upon the astronomers and not the observatory or the SRC
- UV-Continuum subtraction of unknown HI sources
- Possibility of additional flagging
- Recover mistakes (i.e. ALMA pipeline) and/or additional processing
  - **Lesson learnt from MHONGOOSE and MeerKAT Fornax Survey**

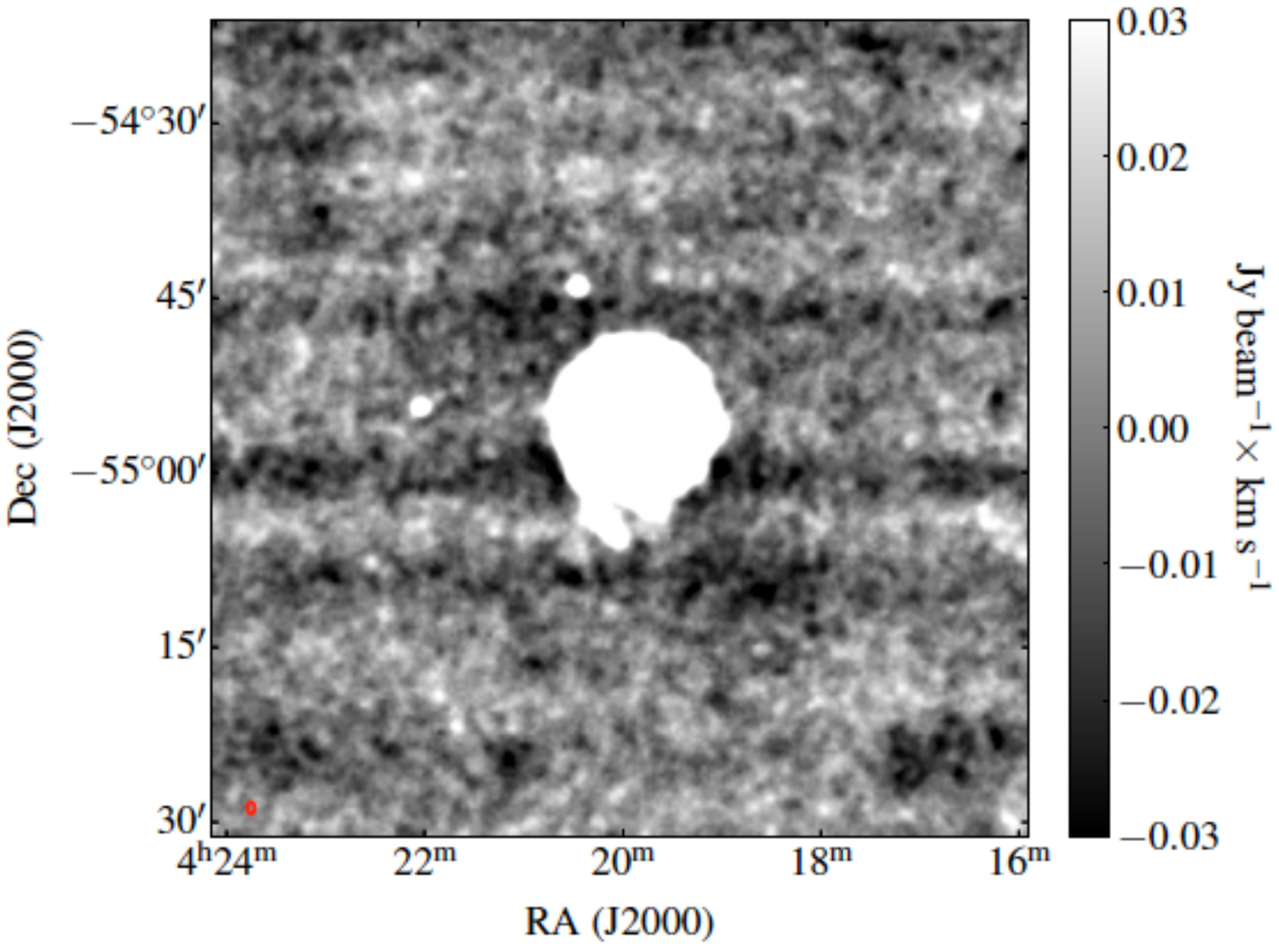


Noise in the datacubes has always been as expected

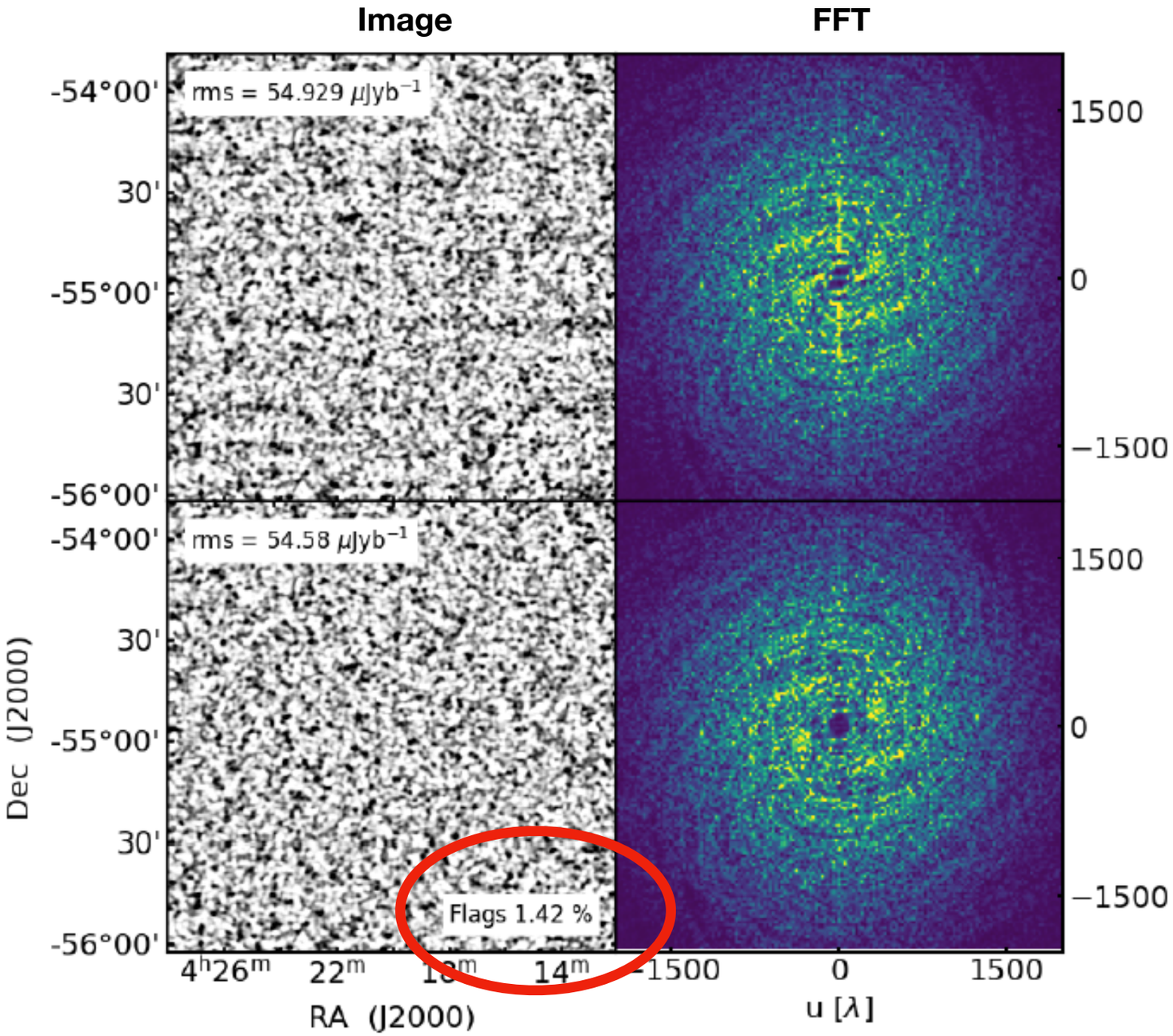
BUT

u=0 problem

once reached low noise (<0.5 mJy) or rebinning to wide channels **horizontal artefacts** appeared in the datacubes



u=0 problem



Reprocess of 2 years of survey:

3 months

Re-observe 2 years of survey:

2 years (if SKA has nothing to do in the mean time)

… bulk of these surveys are ERC-funded …

!!! !!!! KAT-7 (MeerKAT precursor) also noticed this problem (known in radio-interferometry)

But the lesson learnt was lost

Maccagni et al (2022) - https://doi.org/10.48479/bhpj-nz95

What we would like the SRC to do for spectral imaging

- Continuum subtraction, spectral cleaning with masks
- Generate datacubes with the requested specs for the project's science goals
    - Different tapering / robustness
    - Adjust continuum subtraction in the uv-plane
- Most HI science goals will require datacubes at multiple resolutions.
- Field of view (from a fraction to several deg^2) and bandwidth depends on the science goals of the different projects,
    - **Maximum versatility**

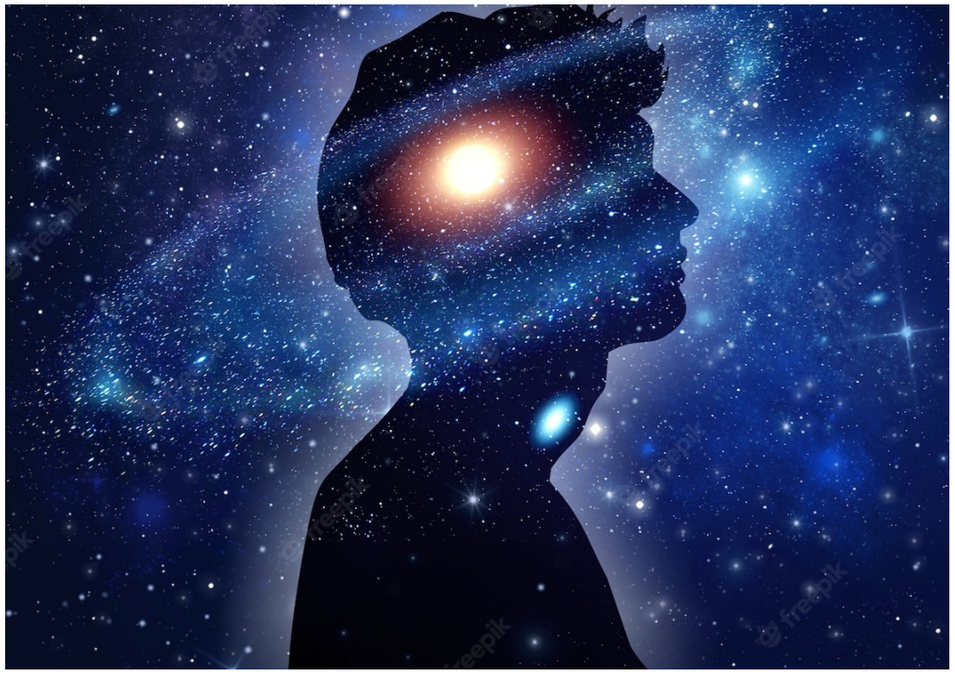
What do we need from the SRC

- Pipeline processing to obtain the best datacubes
  - a pipeline including all best data reduction algorithms and all lessons learnt from the SKA pathfinders and precursors

- Source finding:
  - enable different possibilities, i.e. ML based, SoFiA

- Archive
  - Easy access to HI datacubes and products - ESO user Portal ?
  - Possibility of smoothing/rebinning already existing cubes.
  - Possibility of re-running source-finders.

- Support scientists
  - LOFAR, ALMA-Allegro



Hardware and Software environments

- Ilifu: 110 x 32 cpu nodes each with 232GB RAM, allows for short (and longer) term storage of data
- MeerGAS cluster: 4 x 128 cpu 40TB nodes each with 1000GB RAM, 1PB storage disk
- Existing software, may need to be scaled:
  - Pipelines
    - CARAcal (https://caracal.readthedocs.io/en/latest/)
  - Data analysis
    - SoFiA — source finding (https://github.com/SoFiA-Admin/SoFiA-2)
    - Casatools/Casacore — “easy” interaction with the measurement sets
    - Python environments (some standard, but also user-created and maintained), Jupyter notebooks
  - Visualisation
    - Carta & Kvis
    - iDaVIE (https://arxiv.org/abs/2012.11553)

Data visualisation

- KVIS or CARTA
  - KVIS limited by RAM as entire cube is loaded into memory
  - CARTA not memory limited, but needs features added before it supersedes KVIS

- IDaVIE-v — Virtual Reality software for visualising spectral line (or other 3D) data
  - Being developed in Cape Town (https://idavie.readthedocs.io/en/latest/)
  - Requires Windows environment as software based on gaming engines
  - Need enough RAM to load cubes, and GPU to power rendering
  - In the process of being updated to hosted on a cluster for remote access from the headset, requires 5GHz wireless connection
  - Can currently handle HI data cubes ~80Gb (MeerKAT Fornax Survey, Apertif MDS field)

Some final thoughts

- Support astronomers at the SRC
  - What is the SRC's role in helping to solve identified issues in the data?
  - Will the SRC operate as a "middle person" between data users and observatory, particularly if there are issues with the data that affect the science exploitation?

- Collaborating with other Regional Centres that already have the experienced personnel to build necessary tools?

- Who is responsible for quality control, making the call that the calibrated data is good enough for the science proposed?