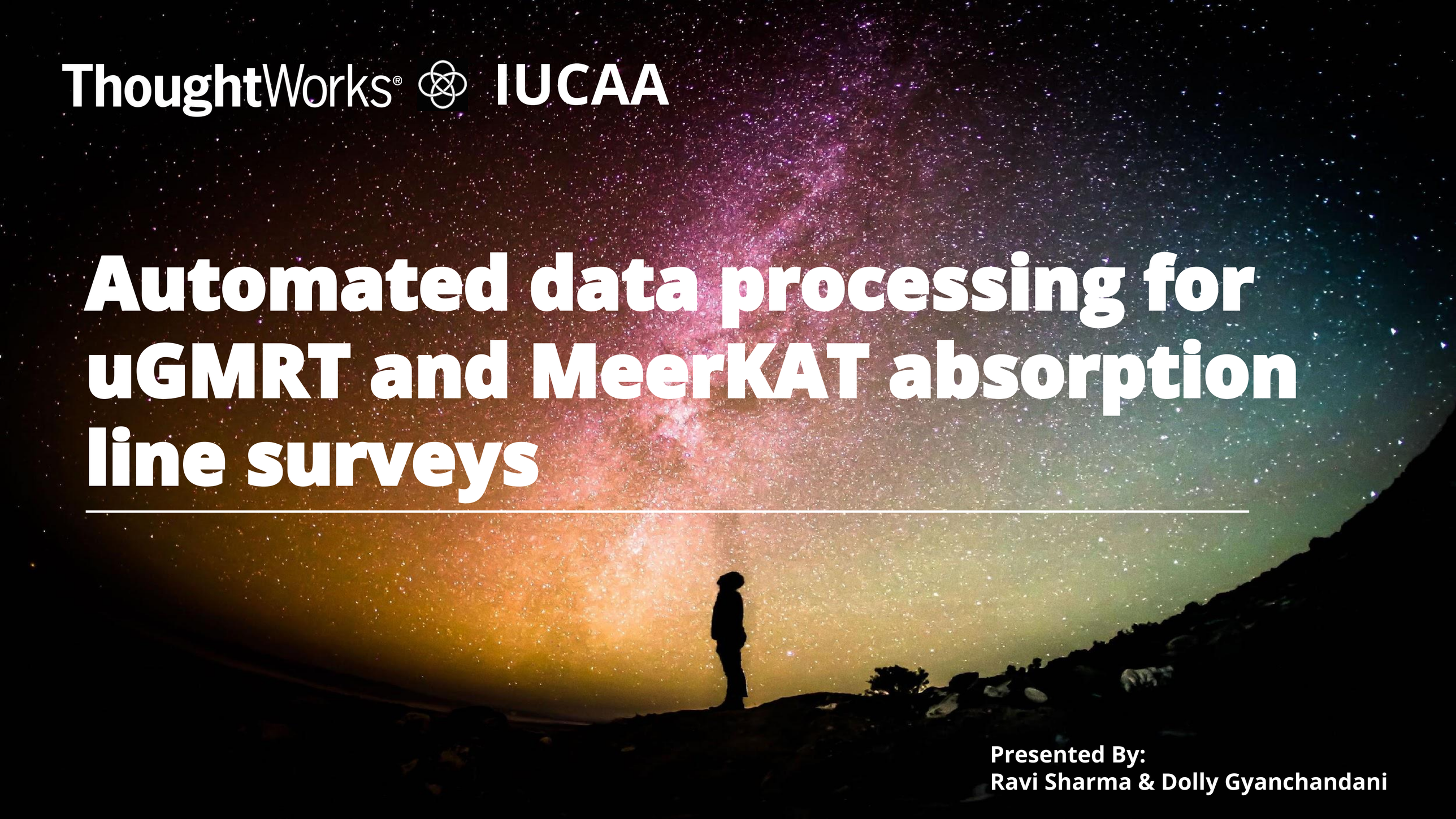
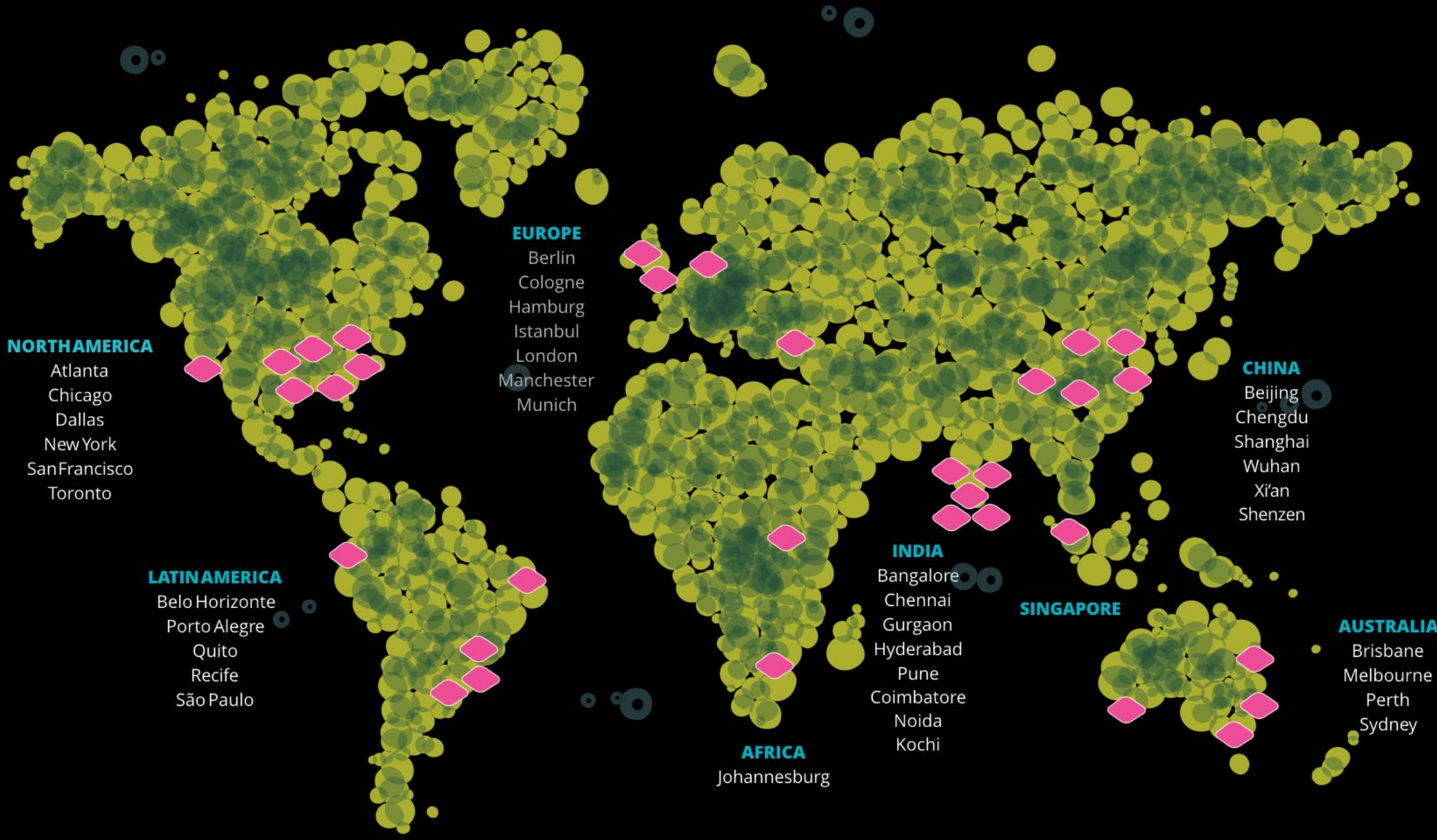


ThoughtWorks® IUCAA

Automated data processing for uGMRT and MeerKAT absorption line surveys

Presented By:
Ravi Sharma & Dolly Gyanchandani



5000+

Passionate ThoughtWorkers

15

Countries

42

Offices

ENGINEERING FOR RESEARCH (E4R)

Thirty Meter Telescope (TMT)

Indian Institute of Astrophysics (IIA)

Automated Radio Telescope Imaging Pipeline (ARTIP)

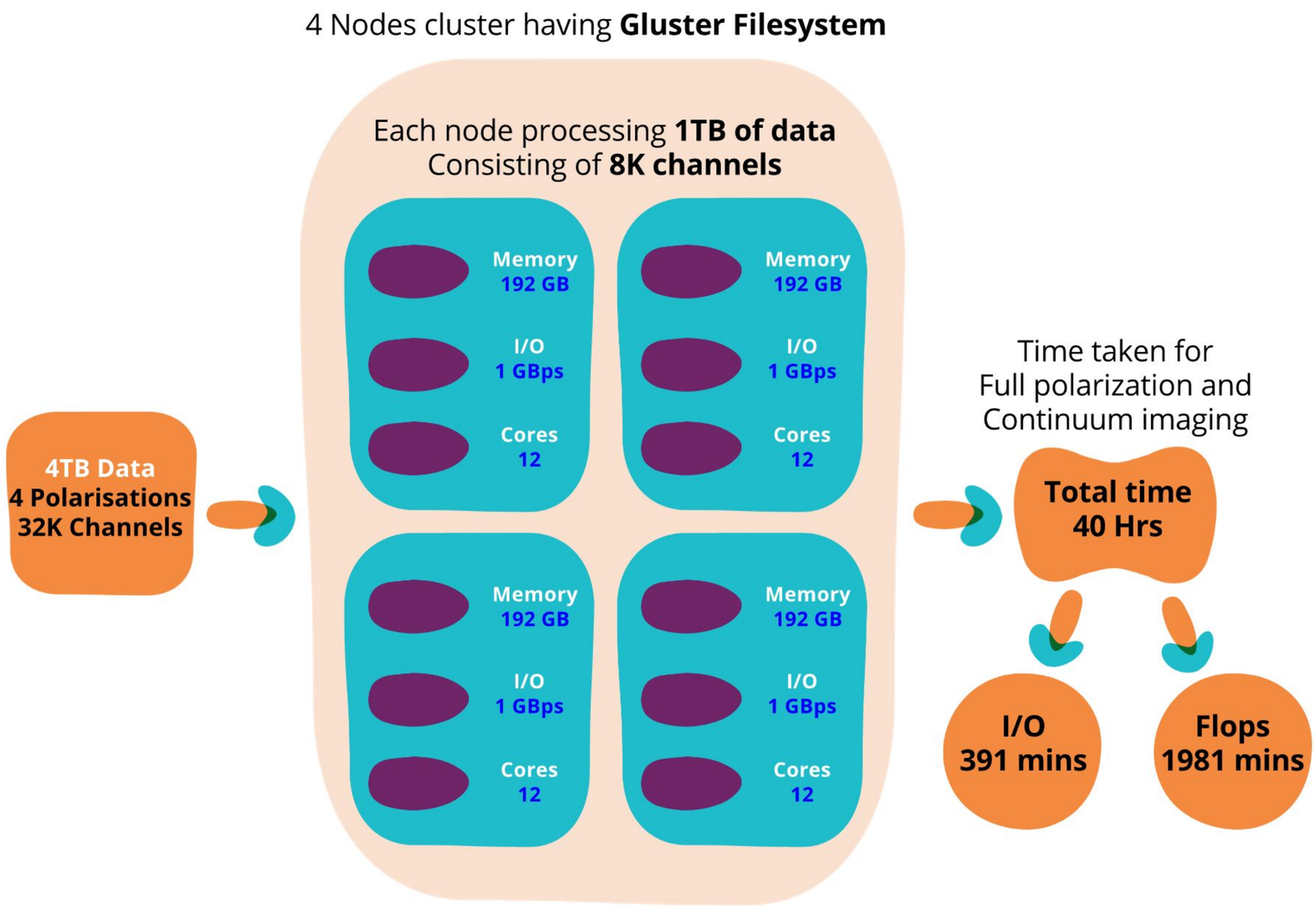
The Inter-University Centre for Astronomy and Astrophysics

NEED FOR PIPELINE

MEERKAT ABSORPTION LINE SURVEY

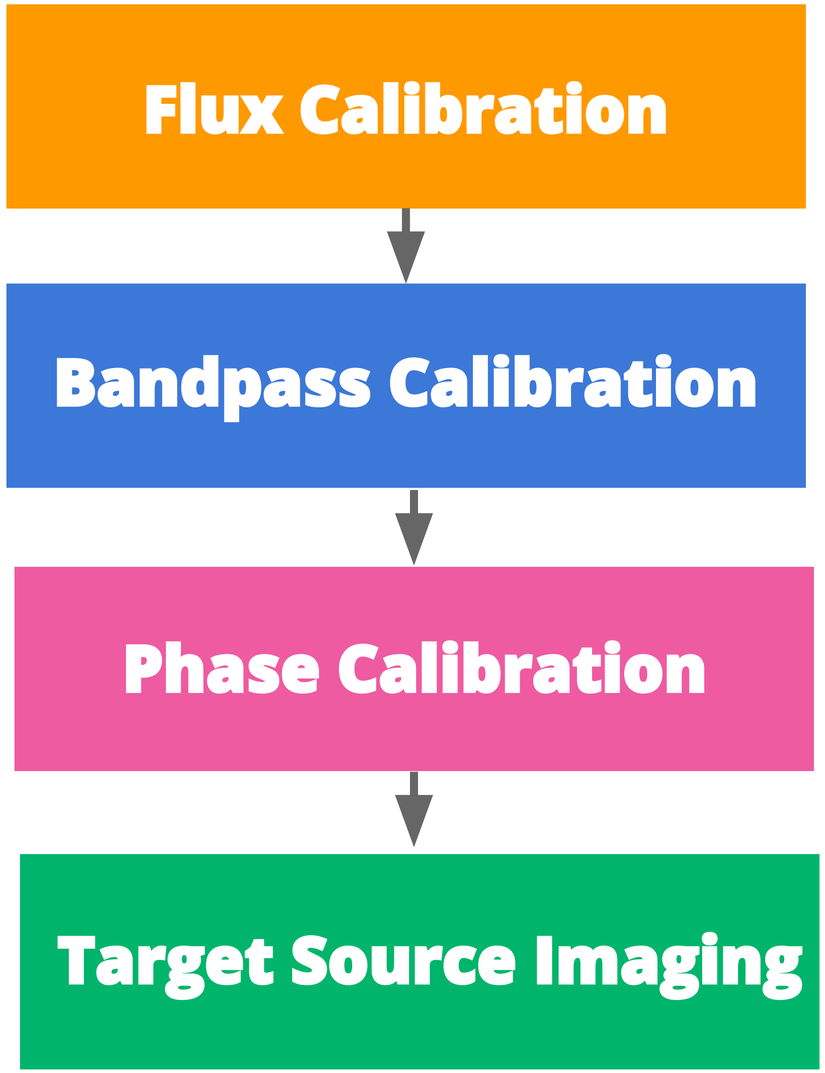
2 Hours of Observation = **4TB** x **RAW DATA IMAGES**

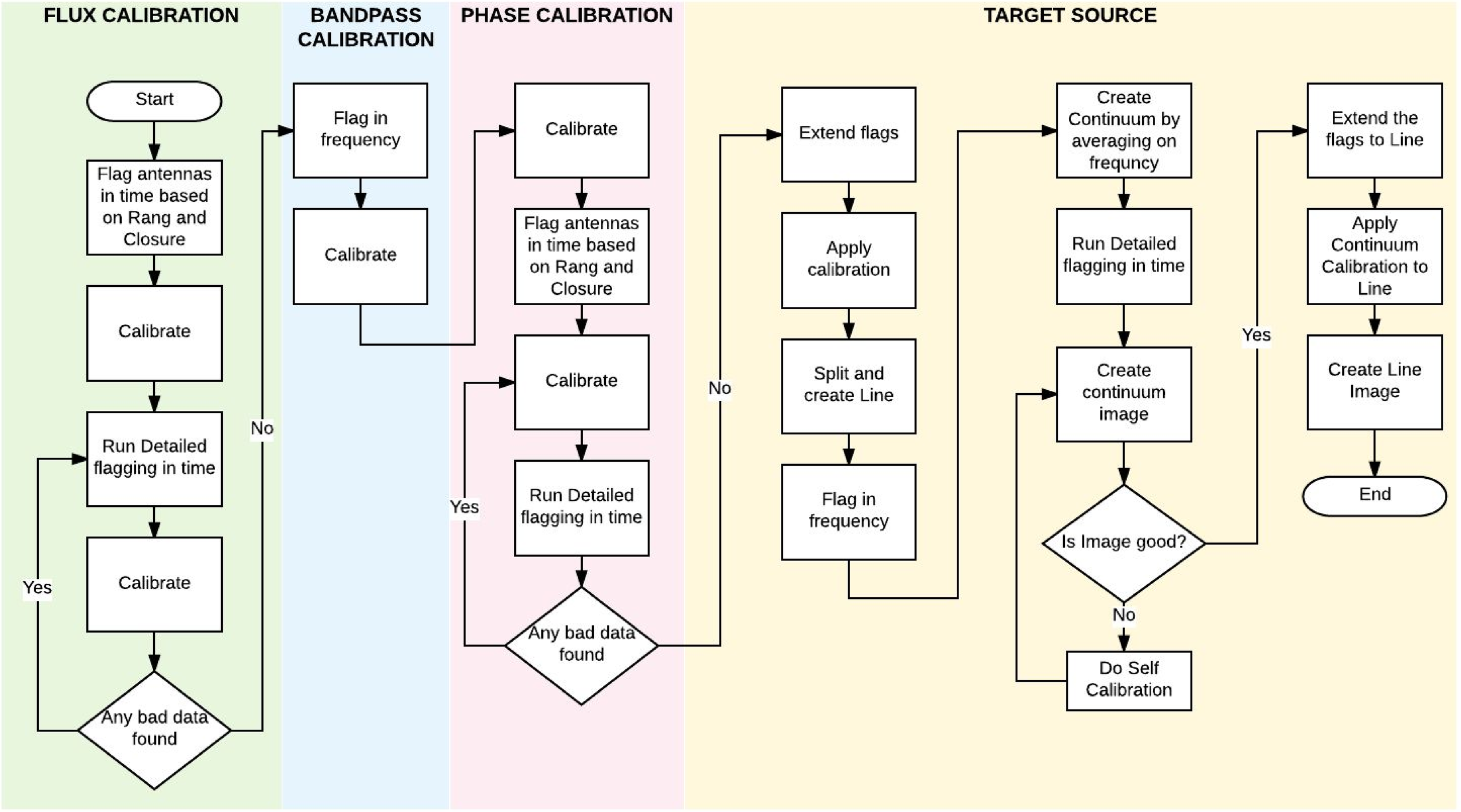
NEED FOR PARALLELIZATION



Processing same data on a single high class machine will take several weeks

PIPELINE STAGES

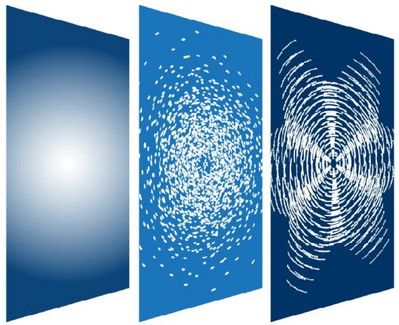



FLUX CALIBRATION
BANDPASS CALIBRATION
PHASE CALIBRATION
TARGET SOURCE
Start
Flag antennas in time based on Rang and Closure
Calibrate
Run Detailed flagging in time
Calibrate
Any bad data found
Yes
No
Flag in frequency
Calibrate
Calibrate
Flag antennas in time based on Rang and Closure
Calibrate
Run Detailed flagging in time
Any bad data found
Yes
No
Extend flags
Apply calibration
Split and create Line
Flag in frequency
Create Continuum by averaging on frequncy
Run Detailed flagging in time
Create continuum image
Is Image good?
No
Do Self Calibration
Yes
Extend the flags to Line
Apply Continuum Calibration to Line
Create Line Image
End

TECHNOLOGY STACK

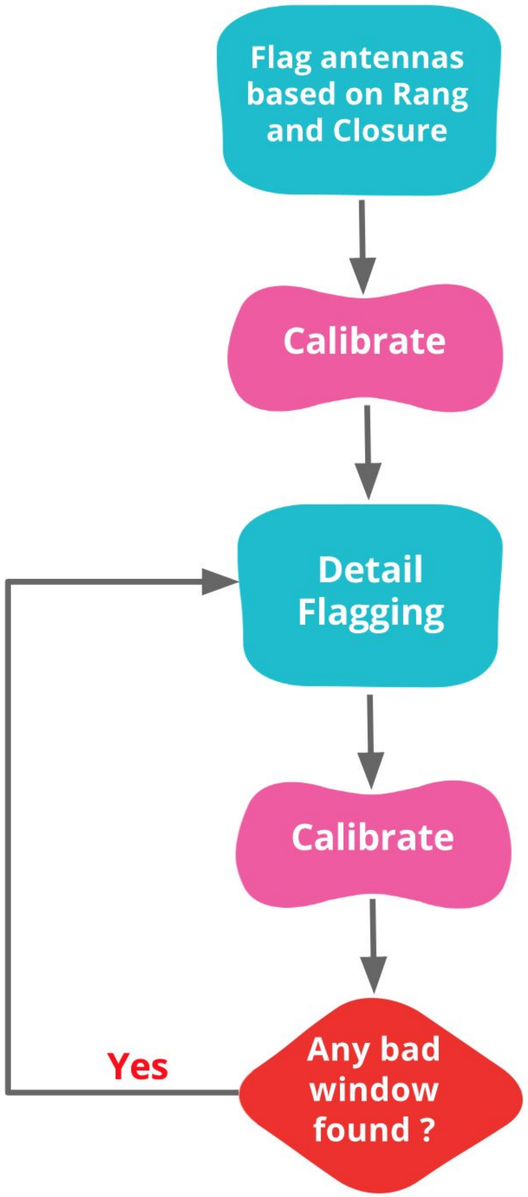
ANACONDA®

Peter Williams conda packages
https://github.com/pkgw/conda-recipes



CASA
Common Astronomy Software Applications

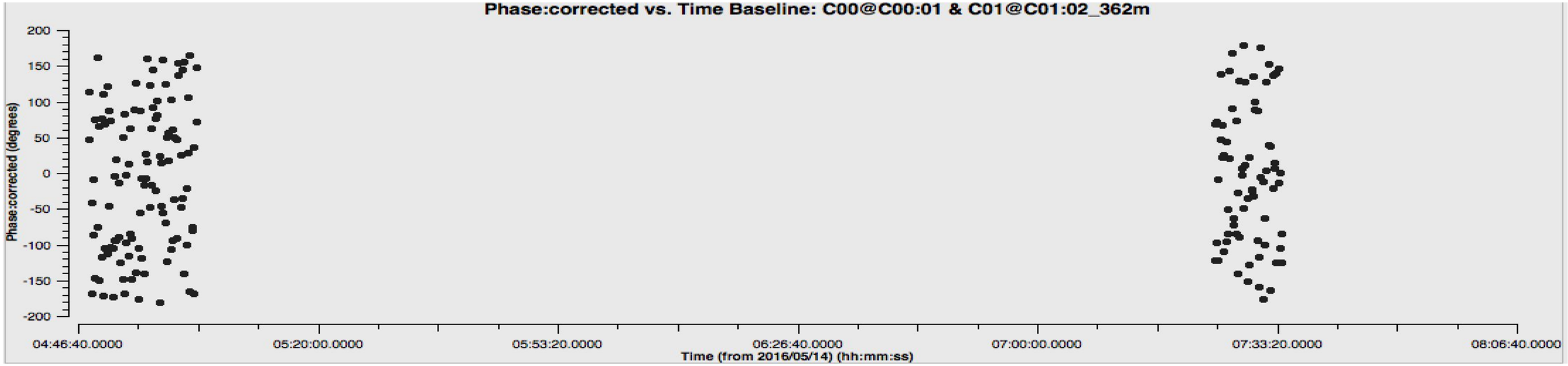
FLUX CALIBRATION

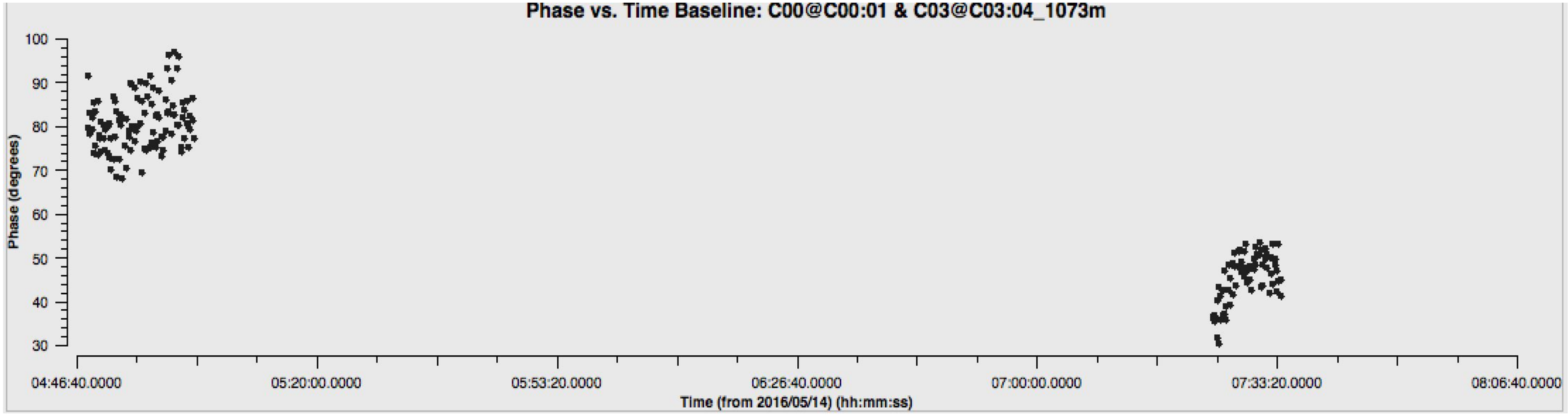


- Flagging
  - Initial screening of antennas
  - Detailed analysis in time
  - Keep track of flags:
    - Extension to other sources, frequencies, etc
    - Flagging statistics to user

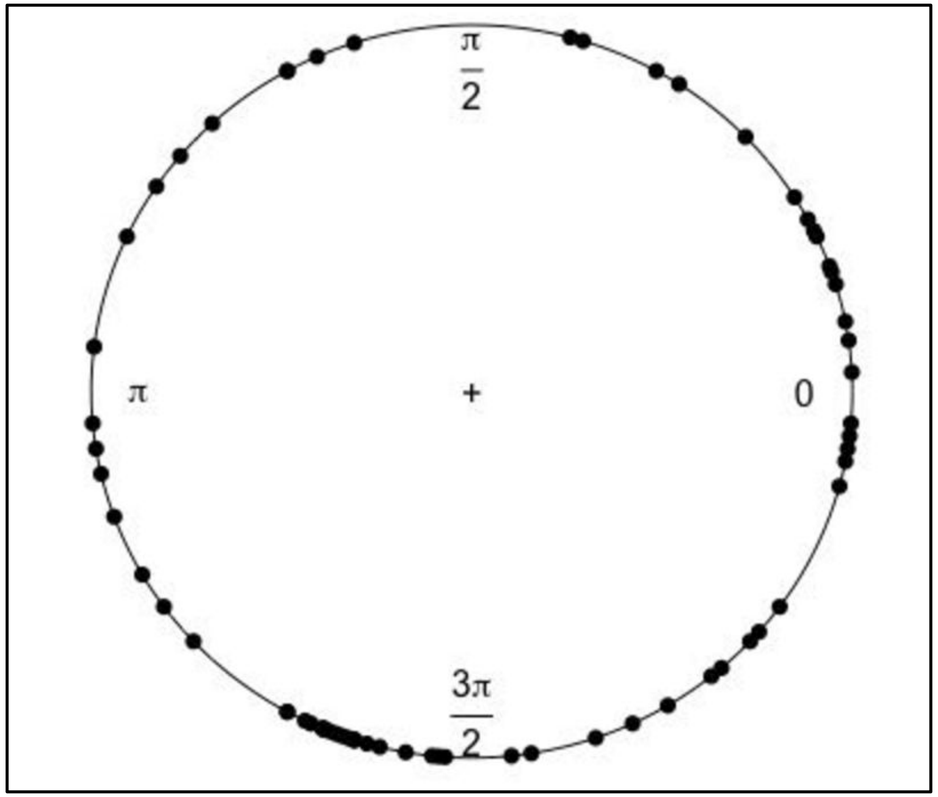
- Done on **single** channel on flux calibrator

PHASE DISPERSION

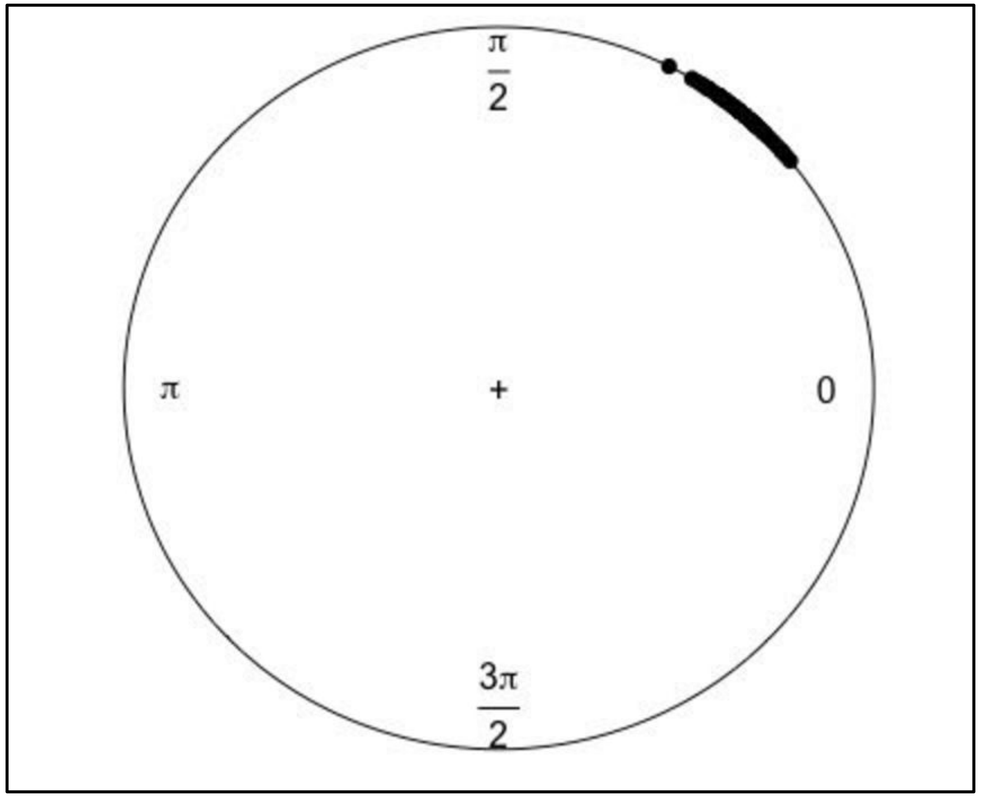




CIRCULAR STATISTICS: ANGULAR DISPERSION



R ≈ 0

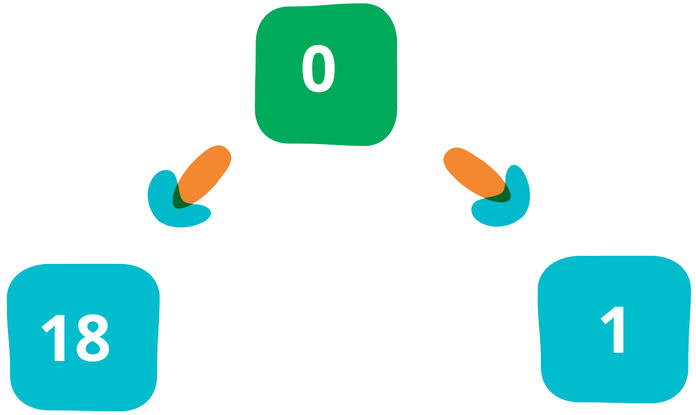


R ≈ 1

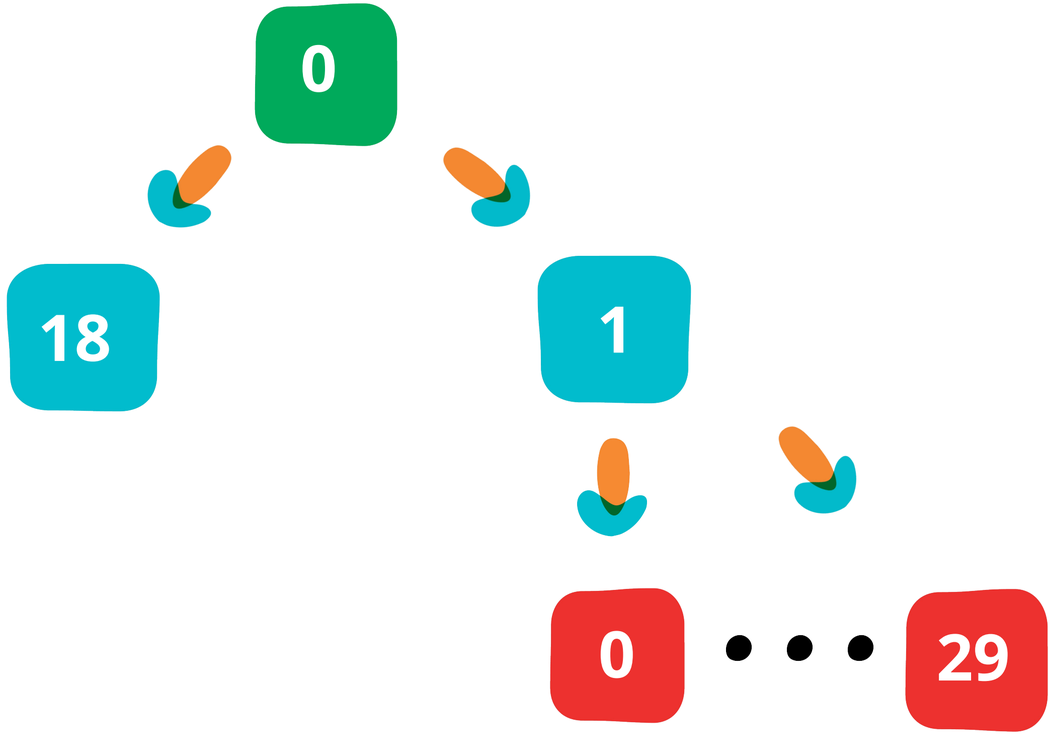
ANGULAR DISPERSION

0

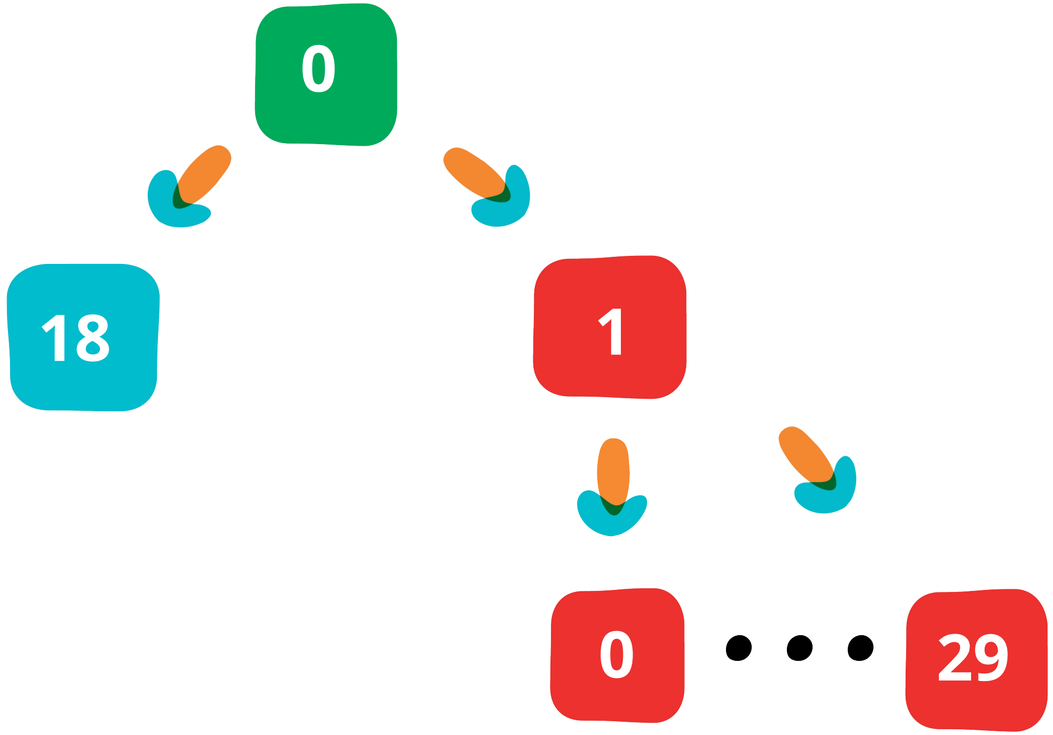
ANGULAR DISPERSION



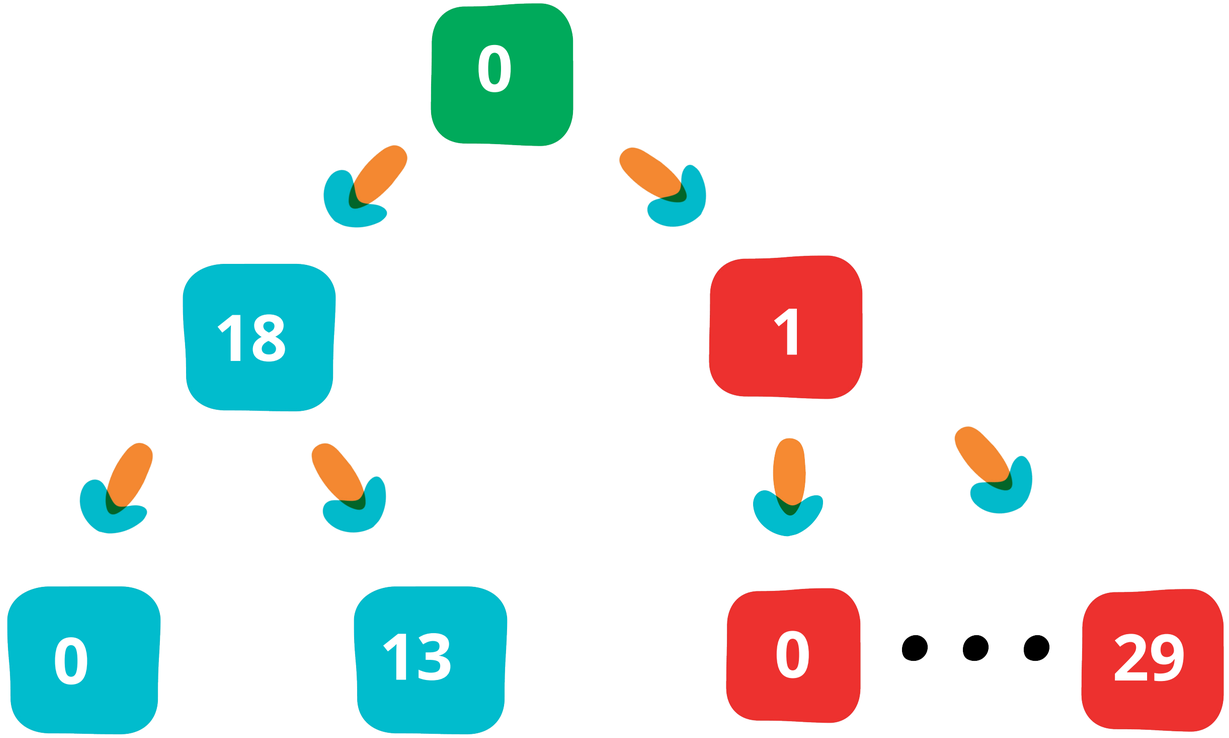
ANGULAR DISPERSION



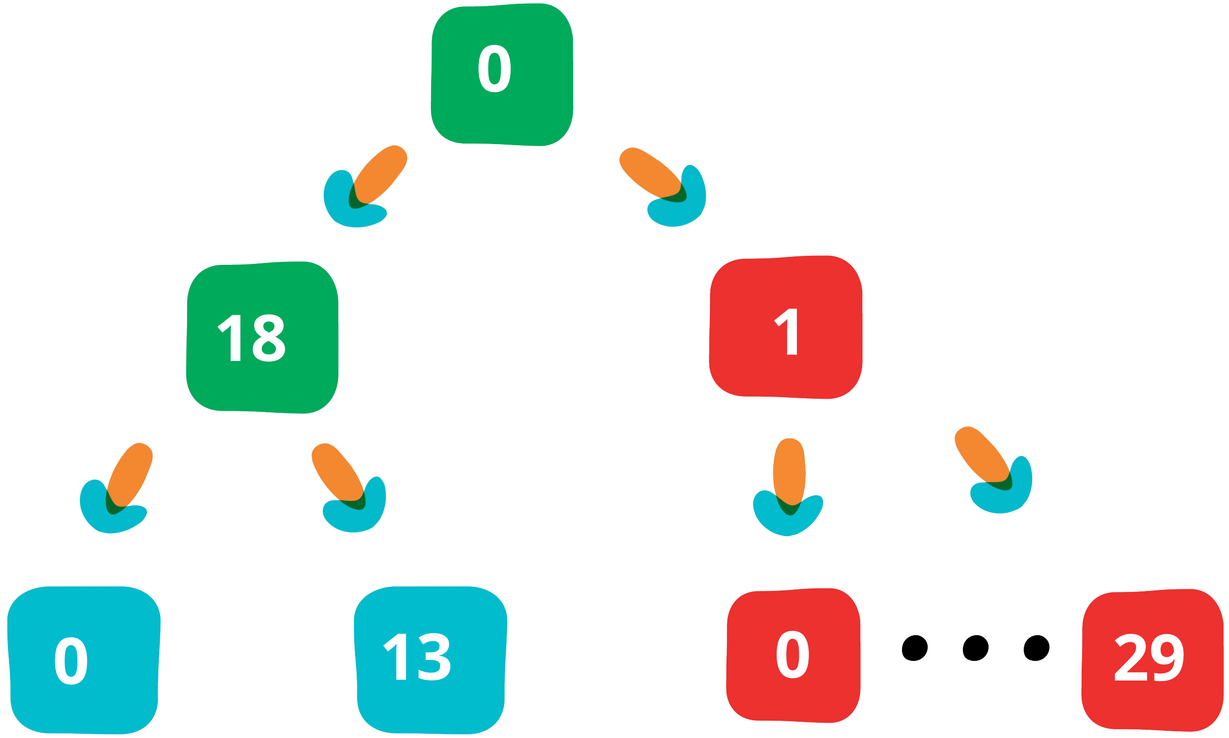
ANGULAR DISPERSION



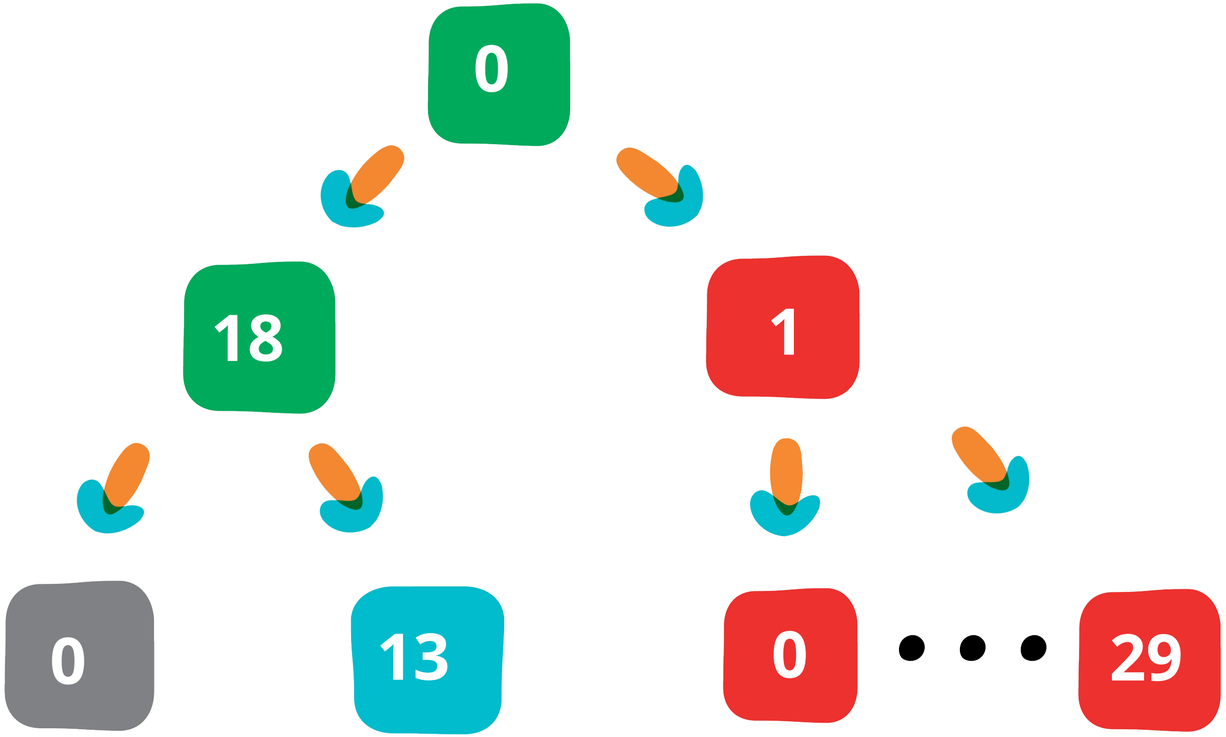
ANGULAR DISPERSION



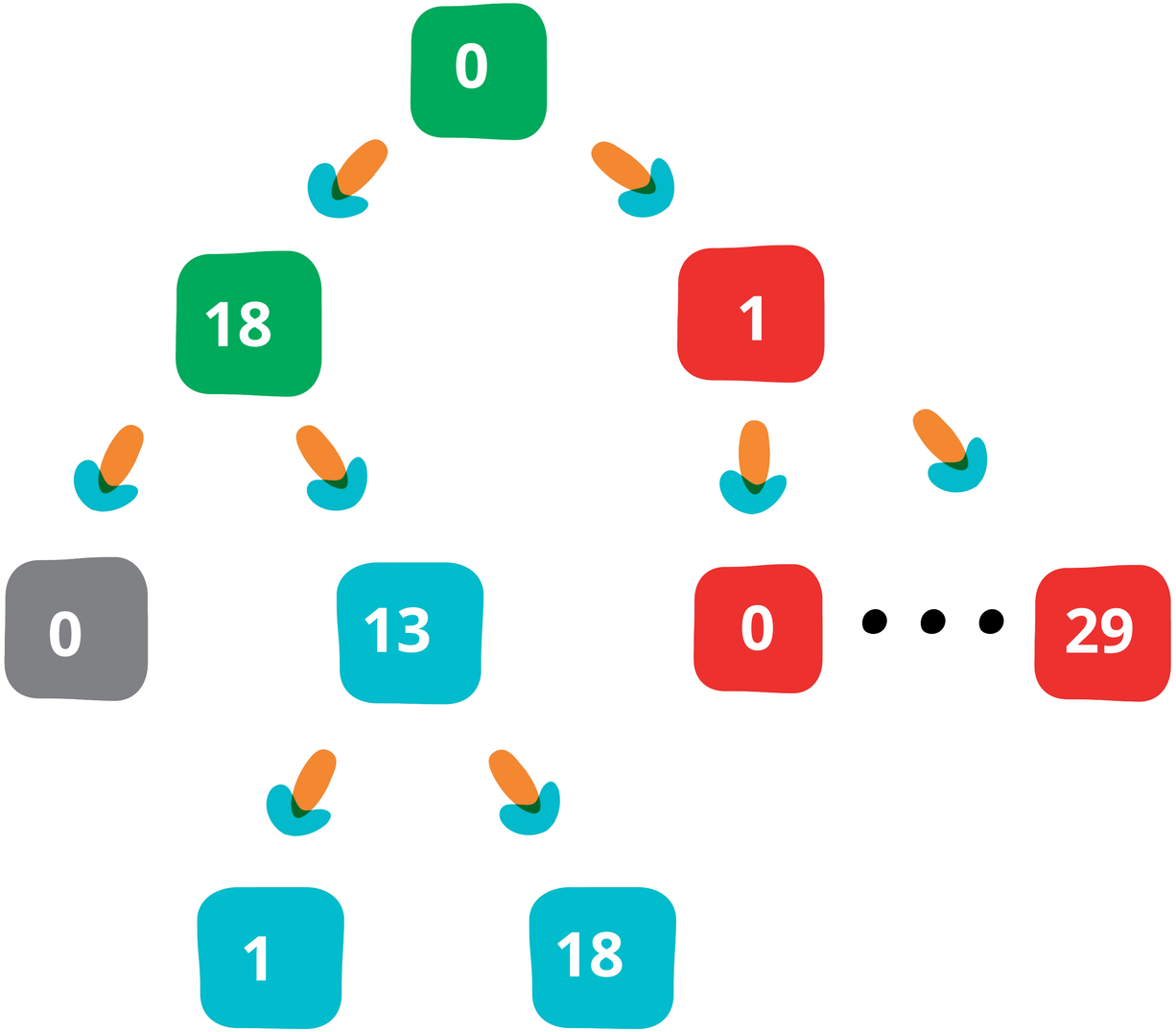
ANGULAR DISPERSION



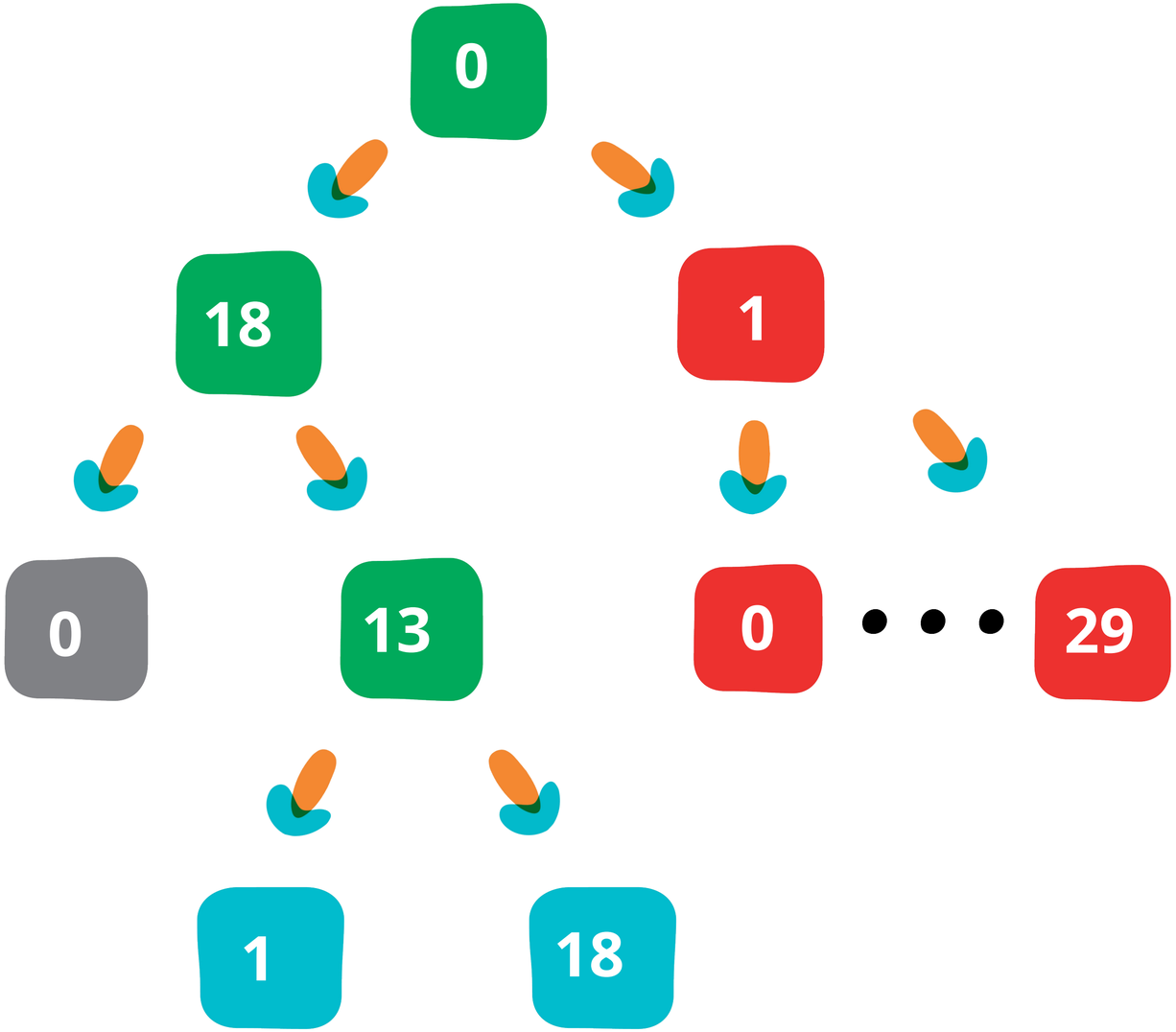
ANGULAR DISPERSION



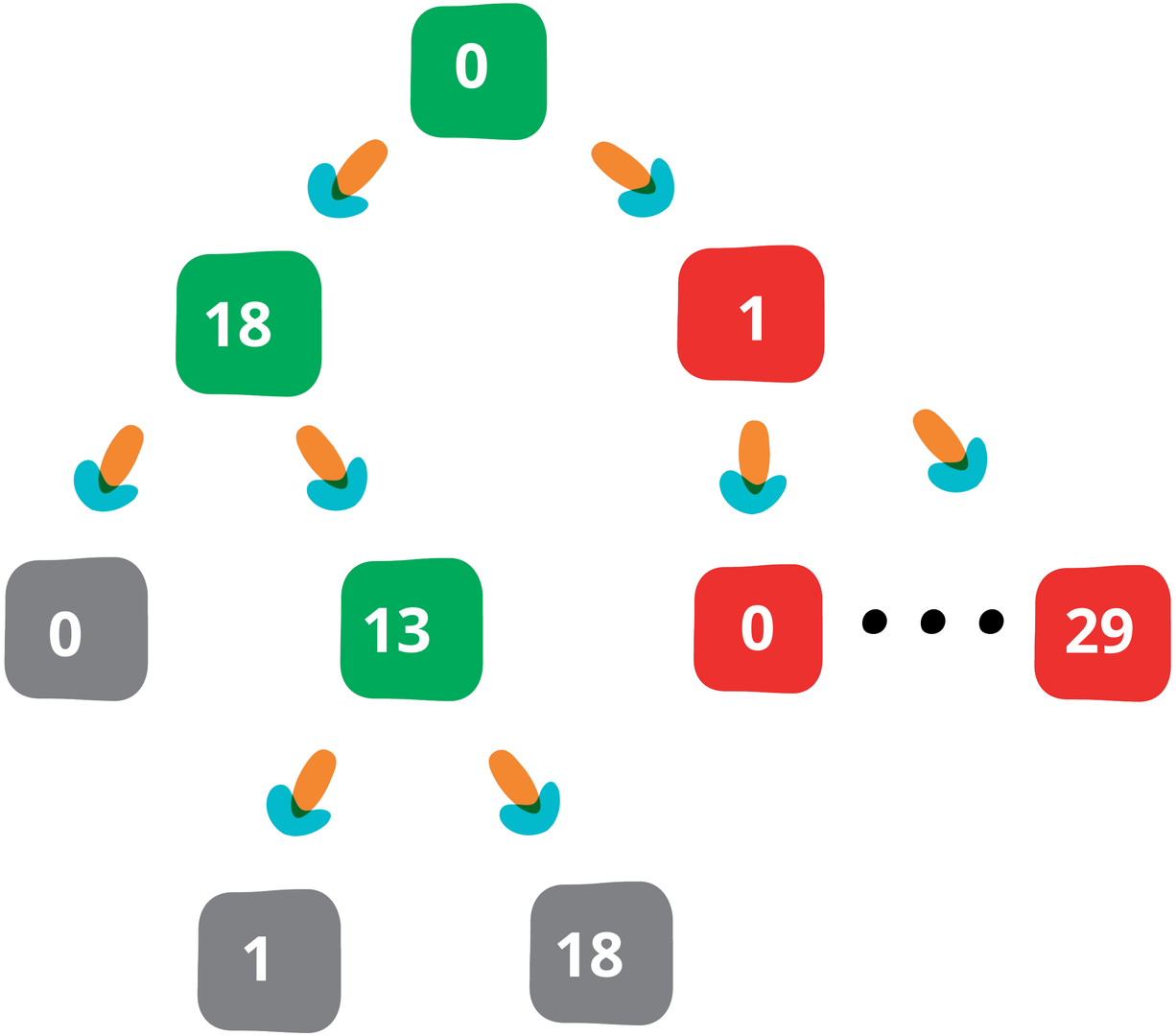
ANGULAR DISPERSION



ANGULAR DISPERSION

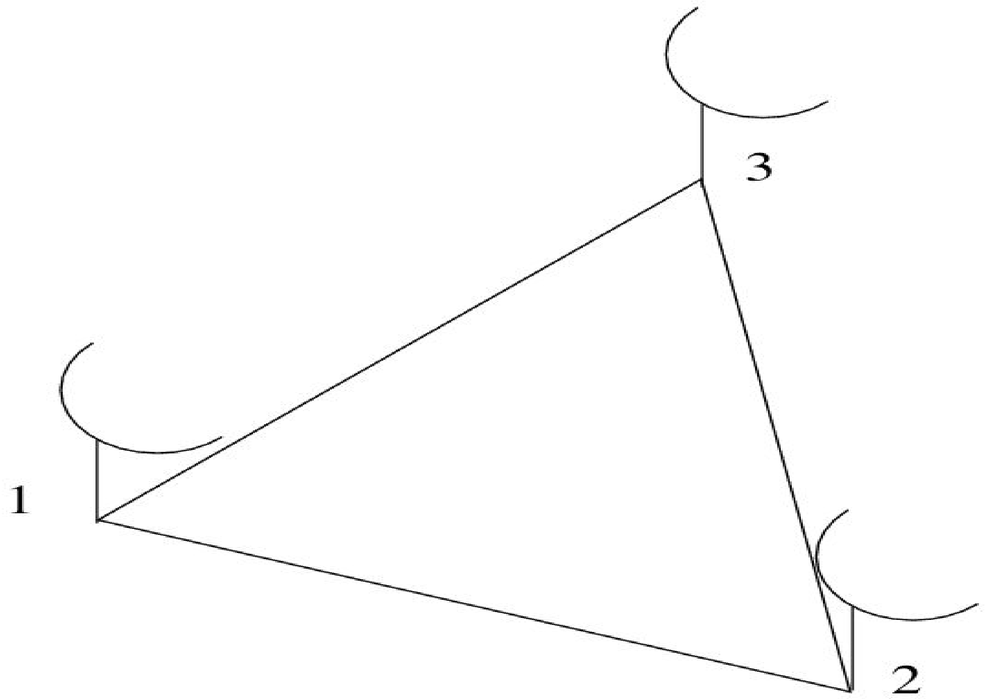


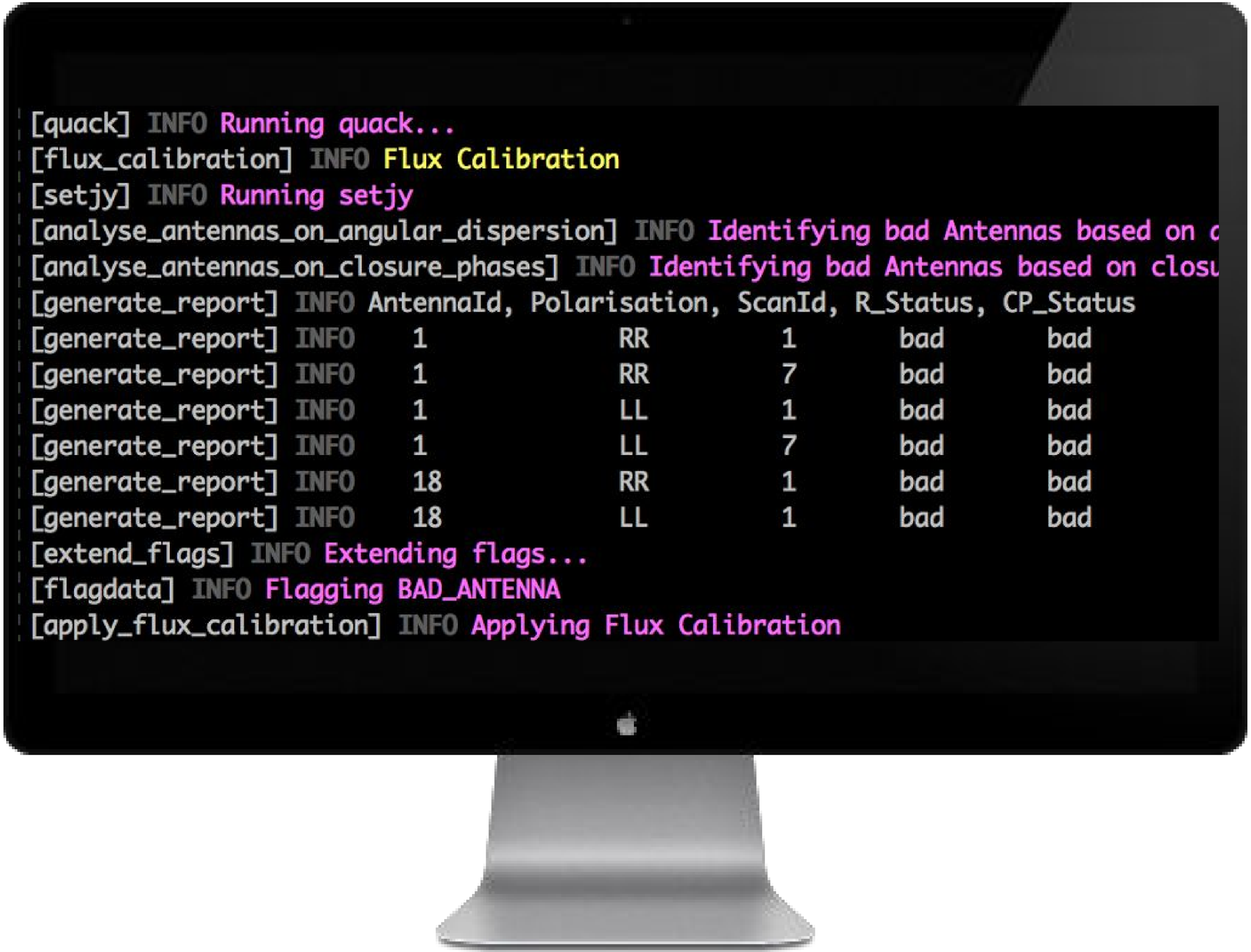
ANGULAR DISPERSION



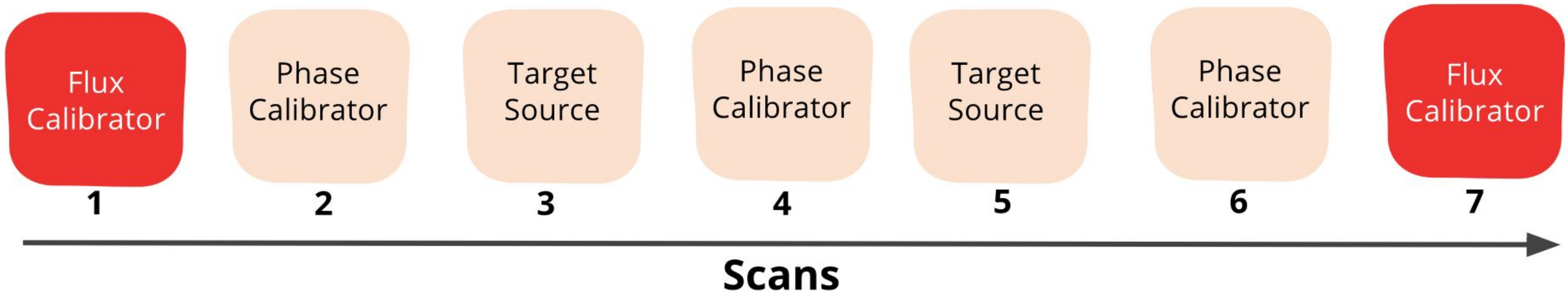
CLOSURE PHASES

$$O = \Phi_{12} + \Phi_{23} - \Phi_{13}$$




[quack] INFO Running quack...
[flux_calibration] INFO Flux Calibration
[setjy] INFO Running setjy
[analyse_antennas_on_angular_dispersion] INFO Identifying bad Antennas based on a
[analyse_antennas_on_closure_phases] INFO Identifying bad Antennas based on closu
[generate_report] INFO AntennaId, Polarisation, ScanId, R_Status, CP_Status
[generate_report] INFO 1 RR 1 bad bad
[generate_report] INFO 1 RR 7 bad bad
[generate_report] INFO 1 LL 1 bad bad
[generate_report] INFO 1 LL 7 bad bad
[generate_report] INFO 18 RR 1 bad bad
[generate_report] INFO 18 LL 1 bad bad
[extend_flags] INFO Extending flags...
[flagdata] INFO Flagging BAD_ANTENNA
[apply_flux_calibration] INFO Applying Flux Calibration

EXTENSION OF FLAGS: FLUX CALIBRATION



RESULTS OF INITIAL SCREENING

All Antennas

Antenna 0

Antenna 1

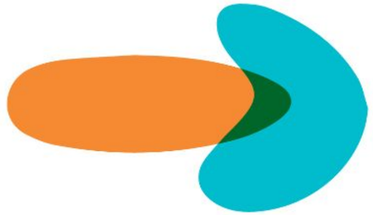
Antenna 2

•
•
•

Antenna 18

•
•
•

Antenna 29



Bad Antennas

Antenna 1

Antenna 18

Possibly Good Antennas

Antenna 0

•
•
•

Antenna 29

DETAILED FLAGGING

- Good Antennas can be bad for some part of time
- Baseline combinations can be bad for some part of time

Amplitude data of a single channel

	A	B	C	D	E	F	G	H
1	Amplitude Matrix							
2	**Baseline**	**T1**	**T2**	**T3**	**T4**	**....**	**....**	**Tn**
3	(0,1)	5	2	7	8	....	....	6
4	(0,2)	6	4	5	4	....	....	9
5	(1,3)	7	4	7	**Nan**	....	....	5
6	(1,4)	8	5	6	9	....	....	7

Global Median
Global Mad

Filter

Antenna

Amplitude matrix for Antenna 1

	A	B	C	D	E	F	G	H
1	Antenna Matrix							
2	**Baseline**	**T1**	**T2**	**T3**	**T4**	**....**	**....**	**Tn**
3	(0,1)	5	2	7	8	....	....	6
4	(1,3)	7	4	7	**Nan**	....	....	5
5	(1,4)	8	5	6	9	....	....	7

Filter

Baseline

Amplitude matrix for Baseline (0,1)

	A	B	C	D	E	F	G	H
1	Baseline Matrix							
2	**Baseline**	**T1**	**T2**	**T3**	**T4**	**....**	**....**	**Tn**
3	(0,1)	5	2	7	8	....	....	6

DETAILED FLAGGING: SLIDING WINDOW

Filtered Matrix : Antenna 1

	A	B	C	D	E	F	G	H
1	Antenna Matrix							
2	**Baseline**	**T1**	**T2**	**T3**	**T4**	**....**	**....**	**Tn**
3	(0,1)	5	2	7	8	....	....	6
4	(1,3)	7	4	7	**Nan**	....	....	5
5	(1,4)	8	5	6	9	....	....	7

Decide
window size and

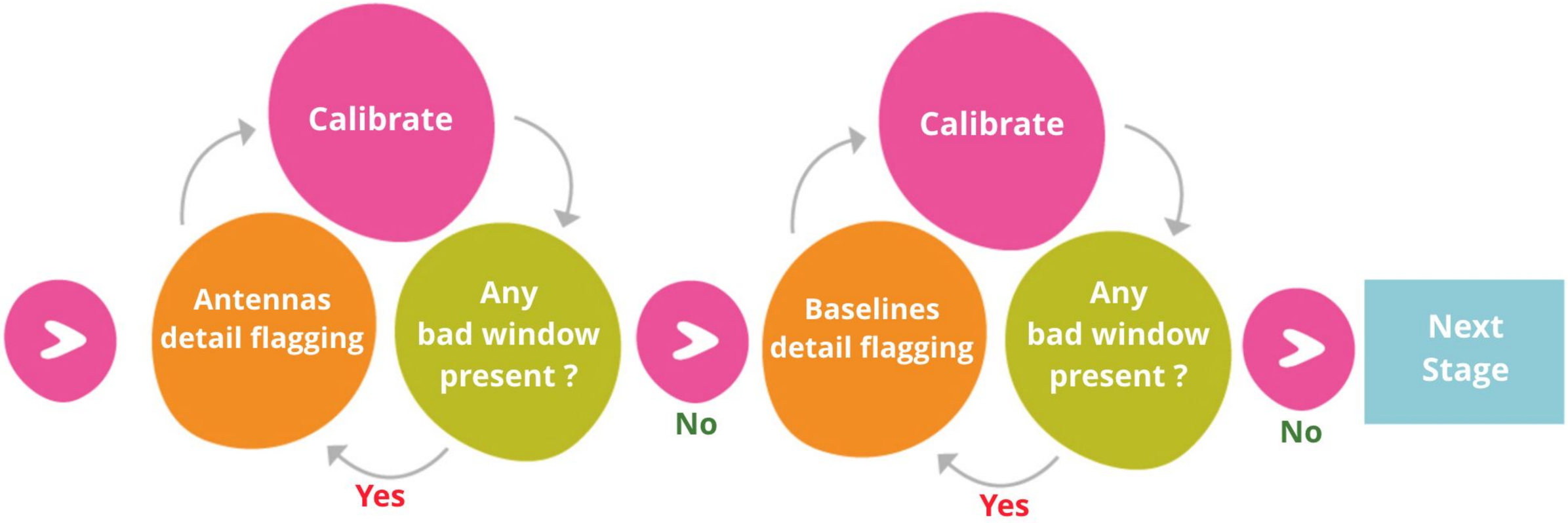
overlap count

Sliding window on Antenna Matrix

	A	B	C	D	E	F	G	H
1	Antenna Matrix							
2	**Baseline**	**T1**	**T2**	**T3**	**T4**	**....**	**....**	**Tn**
3	(0,1)	5	2	7	8	....	....	6
4	(1,3)	7	4	7	**Nan**	....	....	5
5	(1,4)	8	5	6	9	....	....	7

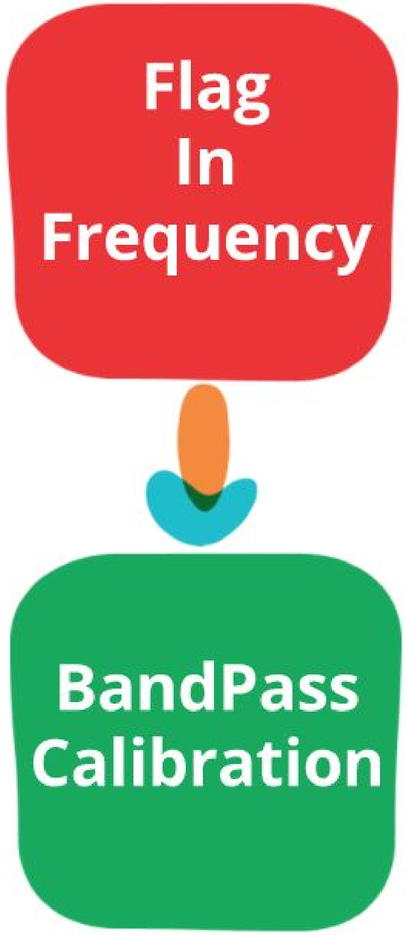
- Compare window median with Global Median to check deviated median
- Compare window MAD with Global MAD to check scattered amplitude

DETAILED FLAGGING: LOOPS



Track the flagged windows data

BANDPASS CALIBRATION



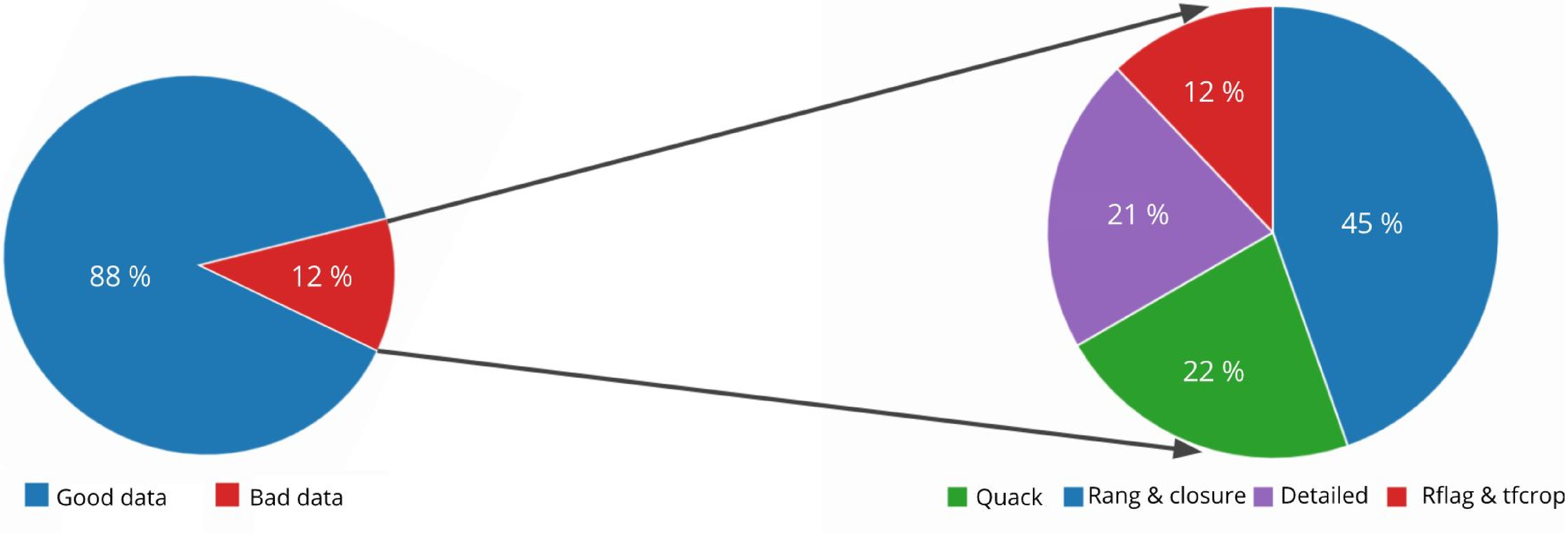
CASA
Common Astronomy
Software Applications

Flagging Algorithms

1. RFlag
2. TFCrop

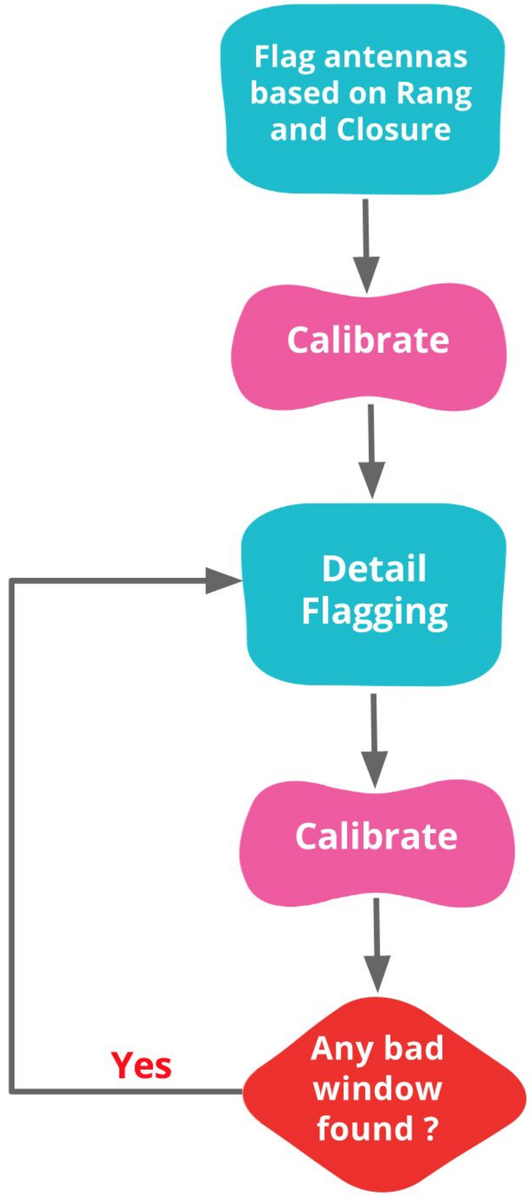
Bandpass gain tables are extended to flux and phase calibrators

PERCENTAGES OF FLAGGED DATA



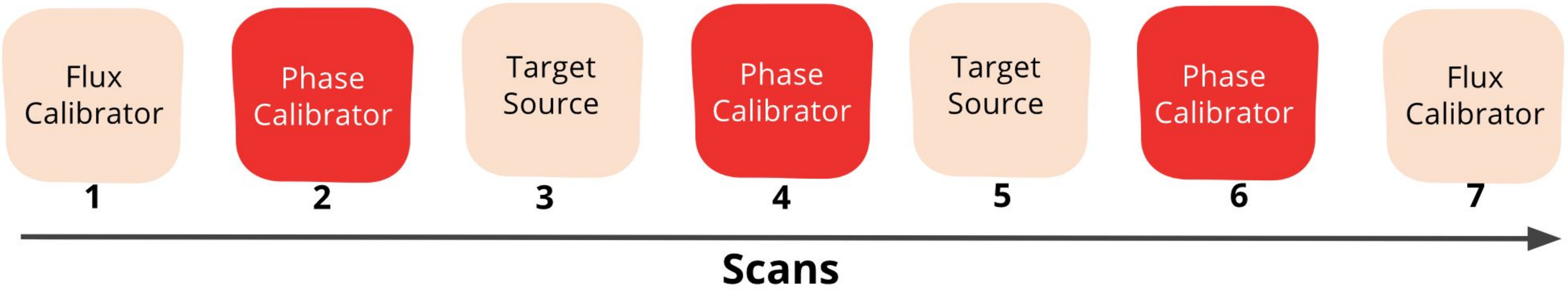
Flagging statistics available at different stages of the pipeline: useful for reliability and quality check.

PHASE CALIBRATION



- Run flagging algorithms with averaged data over channels
- Lenient flagging thresholds because completely bad antennas are removed in Flux calibration

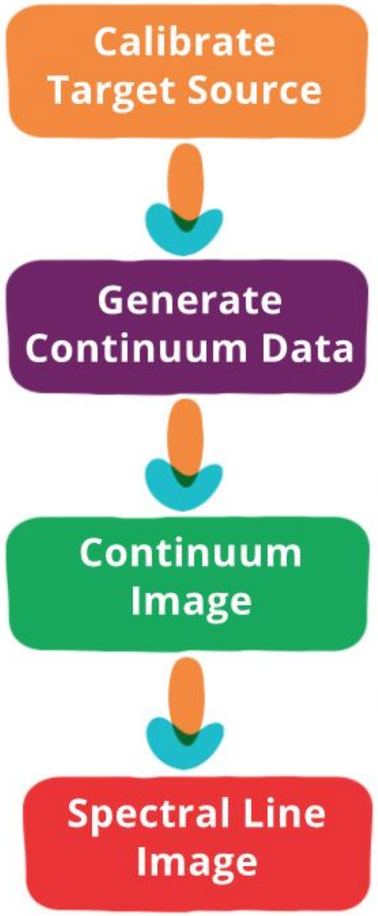
EXTENSION OF FLAGS: PHASE CALIBRATION



If an antenna is bad between two Phase calibrator scans,

it will also be bad in all the scans between them

IMAGING

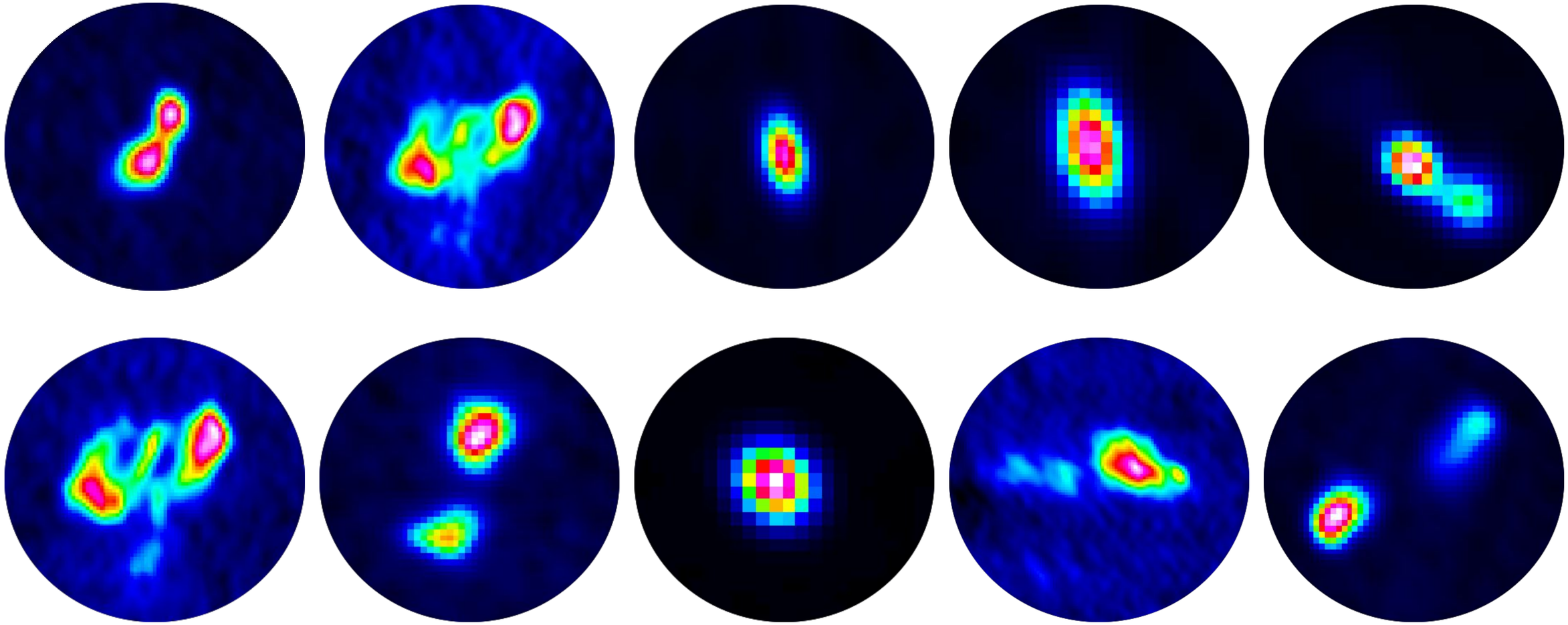


In continuum self calibration masks are constructed based on two criteria :

a).Flux density

b).Shape

IMAGES GENERATED BY THE PIPELINE



Data size = 10GB, Bandwidth = 4 MHz, Channels = 512;
Validated quality of data products and pipeline performance for standard GMRT modes.

PIPELINE PERFORMANCE

Specs:

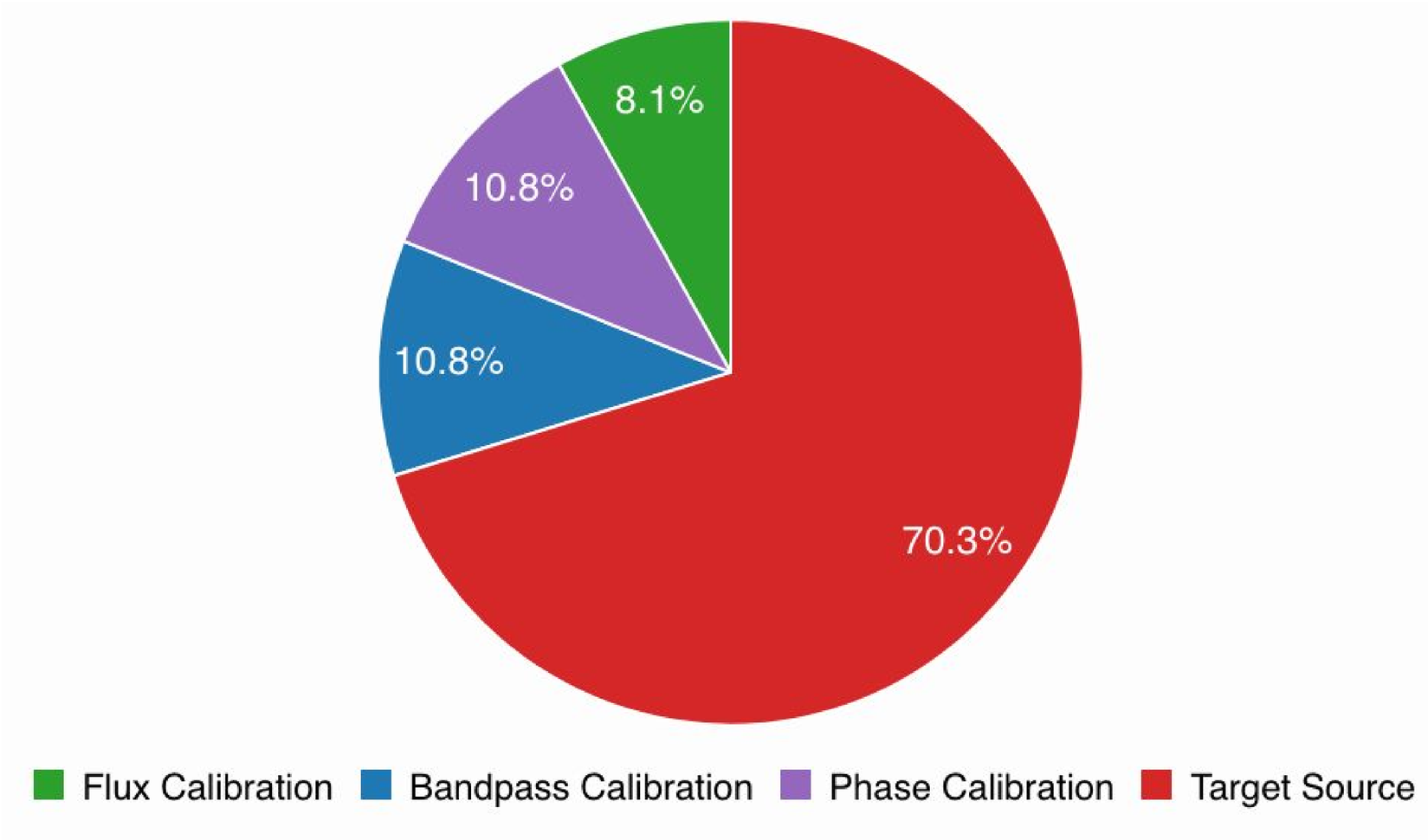
RAM - 256 GB

Cores - 40

Storage - 18 TB

Data volume: 8 GB

Time taken: 37 minutes



In progress

Profiling for wideband uGMRT datasets (200 MHz with 8 K channels; 100 GB): refer to Neeraj's talk

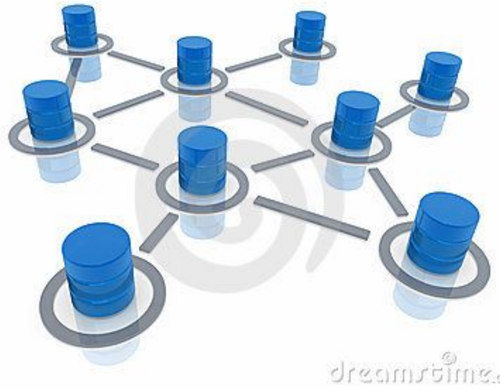
Simulated 1 TB dataset

PARALLELIZATION

1. Data Storage
2. Computing
    - mpicasa (works on mms partitioned by frequencies)

/ lustre

+



40 Nodes IUCAA cluster
- 128 GB RAM
- 16 CPU cores

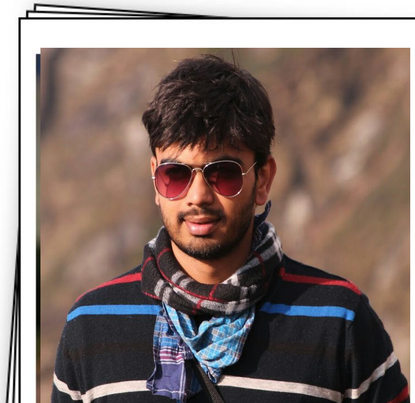
DISTRIBUTED COMPUTING ON SERVER CLASS MACHINES

CONTRIBUTORS

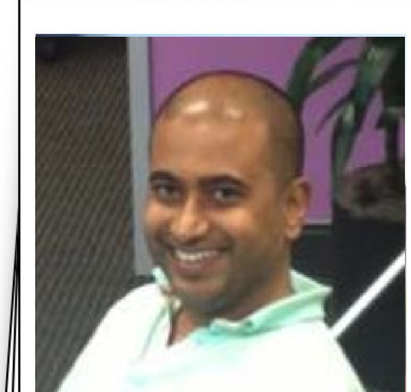
Neeraj Gupta

Dolly Gyanchandani

Unmesh Joshi



Sarang Kulkarni



Santosh Mahale

Arti Pande

Vineet Pathak

Ravi Sharma

Gunjan Shukla


THANK YOU
ThoughtWorks®
IUCAA