



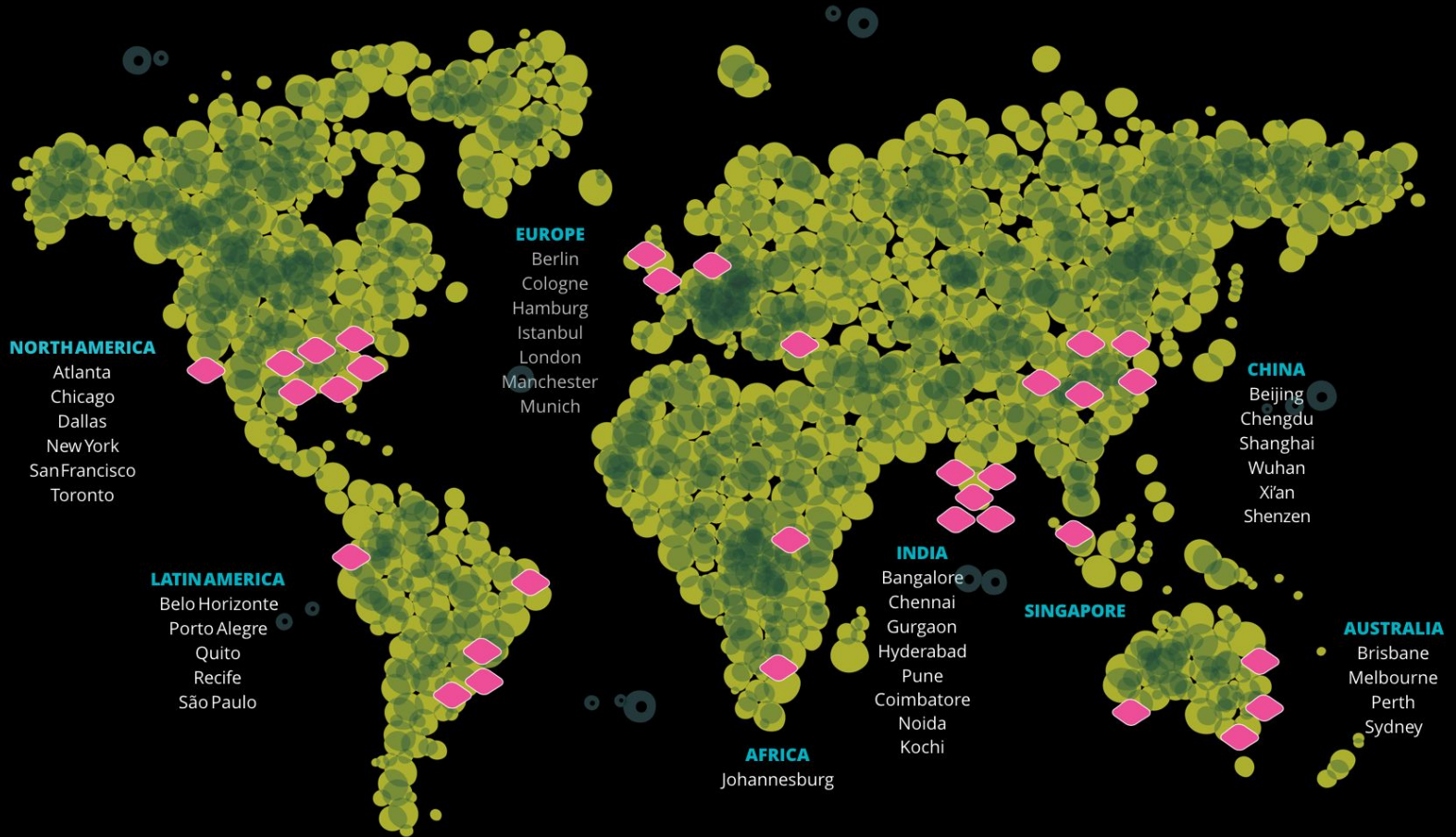
RADIO TELESCOPE DATA PROCESSING

Building software to trace evolution of the Universe

Dolly Gyanchandani | ThoughtWorks | @iam_dollyg



 #GHCI17



5000+

Passionate ThoughtWorkers

15

Countries

42

Offices

ENGINEERING FOR RESEARCH (E4R)

“Solving complex software engineering problems for the scientific & research world”



Thirty Meter Telescope (TMT)

Indian Institute of Astrophysics (IIA)



**Automated Radio Telescope
Imaging Pipeline (ARTIP)**

*The Inter-University Centre for
Astronomy and Astrophysics*

THE SQUARE KILOMETRE ARRAY

World's largest radio telescope being built in Australia and South Africa

First Results Expected

2020

How did stars and galaxies form & evolve

How were magnetic fields of stars and galaxies created?

Do we need a new and more powerful theory of gravity?

Member Countries

11

How did life come to be in the universe?

Was Einstein right about the theory of relativity?

Are we alone in this universe?

*SKA will produce more data in a day than the size of today's global internet:
a massive data storage and processing challenge !*

MEERKAT ABSORPTION LINE SURVEY

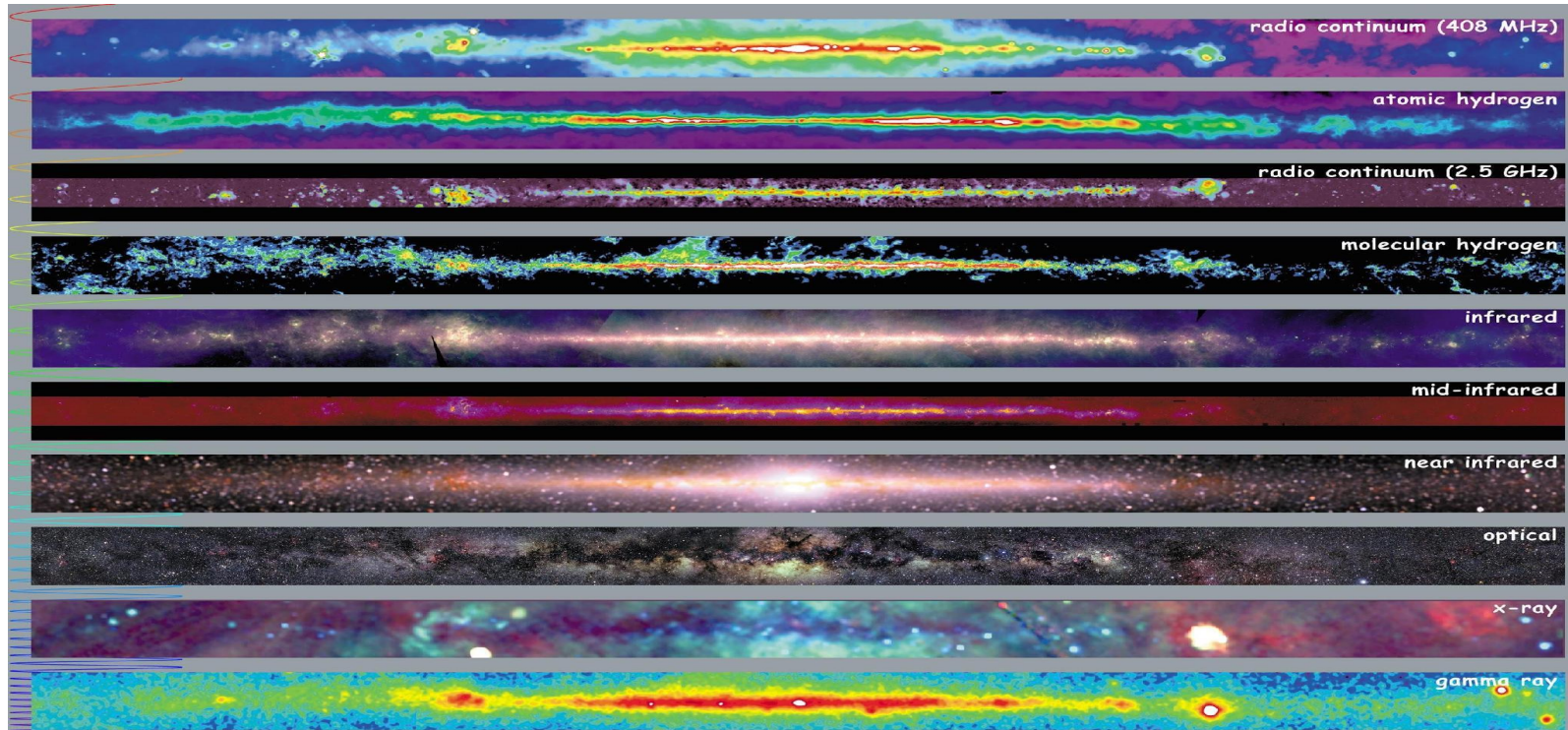


$$2 = 4\text{TB} \times \text{RAW DATA \& IMAGES}$$

Hours of
Observation

$$2\text{TB} \times 1700 \text{ hours} = 3.4 \text{ PB}$$

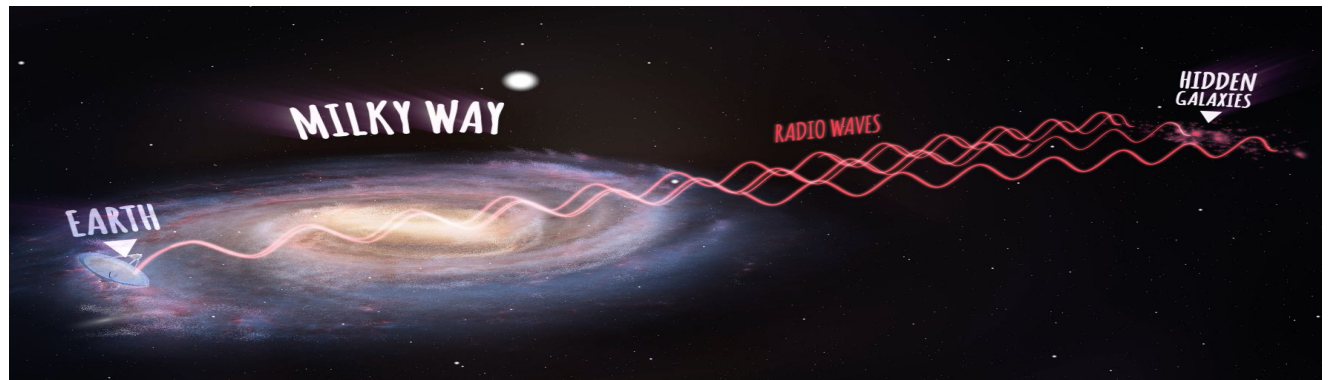
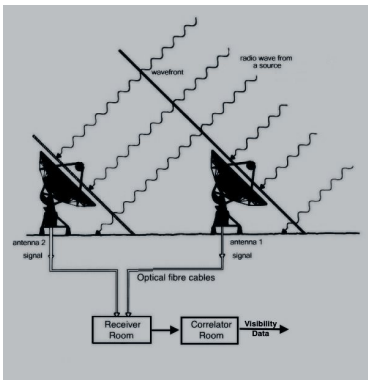
MULTIWAVELENGTH VIEW OF MILKY WAY



RADIO INTERFEROMETRY



Giant Metrewave Radio Telescope , near Narayangaon, Pune

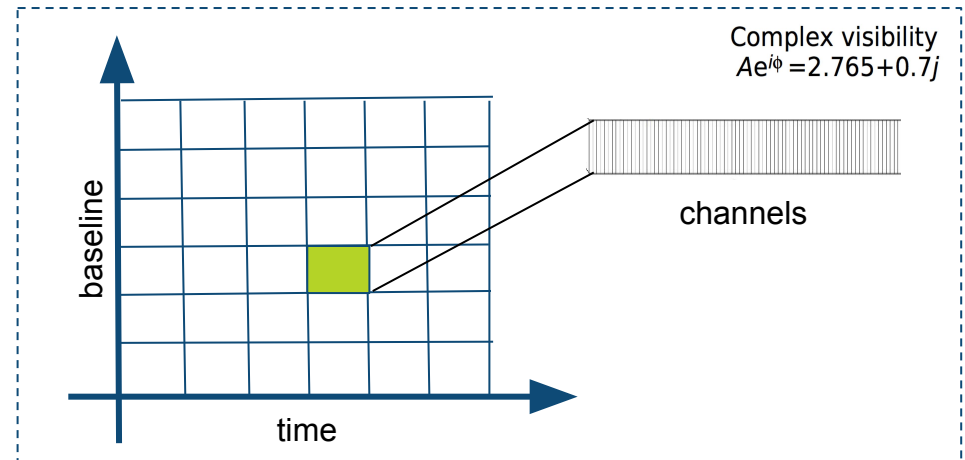


DATA FORMATS

Measurement Sets

-- ANTENNA	-- FIELD	-- POINTING	-- SPECTRAL_WINDOW	-- table.f1_TSM1
-- table.dat	-- table.dat	-- table.dat	-- table.dat	-- table.f2
-- table.f0	-- table.f0	-- table.f0	-- table.f0	-- table.f2_TSM1
-- table.info	-- table.f0i	-- table.f0i	-- table.f0i	-- table.f3
-- table.lock	-- table.info	-- table.f1	-- table.info	-- table.f3_TSM1
-- DATA_DESCRIPTION	-- table.lock	-- table.info	-- table.lock	-- table.f4
-- table.dat	-- FLAG_CMD	-- table.lock	-- STATE	-- table.f4_TSM1
-- table.f0	-- table.dat	-- POLARIZATION	-- table.dat	-- table.f5
-- table.info	-- table.f0	-- table.dat	-- table.f0	-- table.f5_TSM1
-- table.lock	-- table.info	-- table.f0	-- table.info	-- table.f6
-- DOPPLER	-- table.lock	-- table.f0i	-- table.lock	-- table.f6_TSM1
-- table.dat	-- HISTORY	-- table.info	-- table.dat	-- table.f7
-- table.f0	-- table.dat	-- table.lock	-- table.f0	-- table.f7_TSM1
-- table.info	-- table.f0	-- PROCESSOR	-- table.f1	-- table.f8
-- table.lock	-- table.info	-- table.dat	-- table.f10	-- table.f8_TSM1
-- FEED	-- table.lock	-- table.f0	-- table.f9	-- table.f9
-- table.dat	-- OBSERVATION	-- table.info	-- table.f9_TSM0	-- table.f9_TSM0
-- table.f0	-- table.dat	-- table.lock	-- table.info	-- table.info
-- table.f0i	-- table.f0	-- SORTED_TABLE	-- table.lock	-- table.lock
-- table.info	-- table.info	-- table.dat	-- table.dat	-- table.dat
-- table.lock	-- table.lock	-- table.info	-- table.info	-- table.info

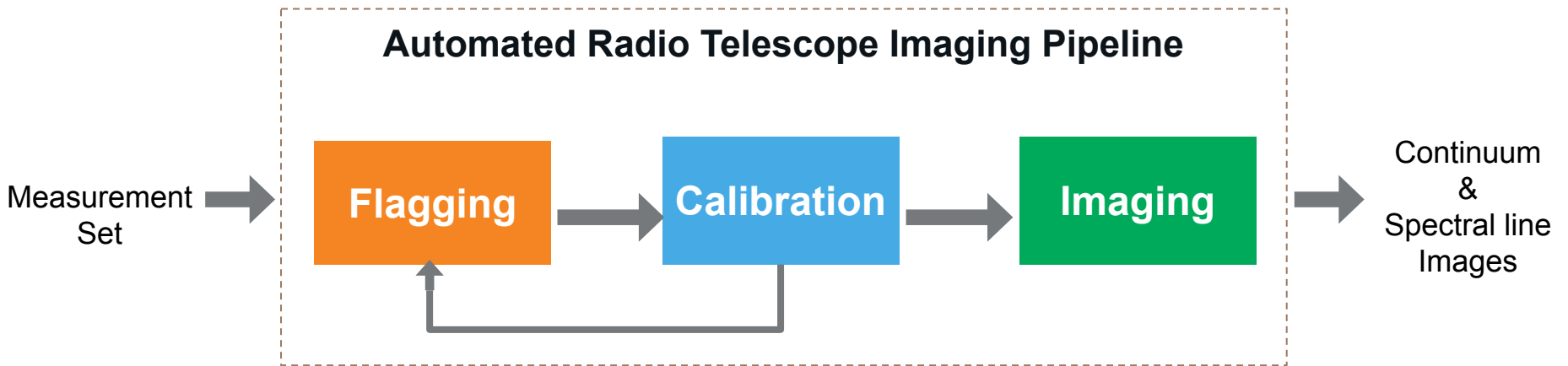
15 directories, 88 files



X 4 Polarizations

60 X 2016 X 32000 X 4 = 15 Billion data-points

ARTIP

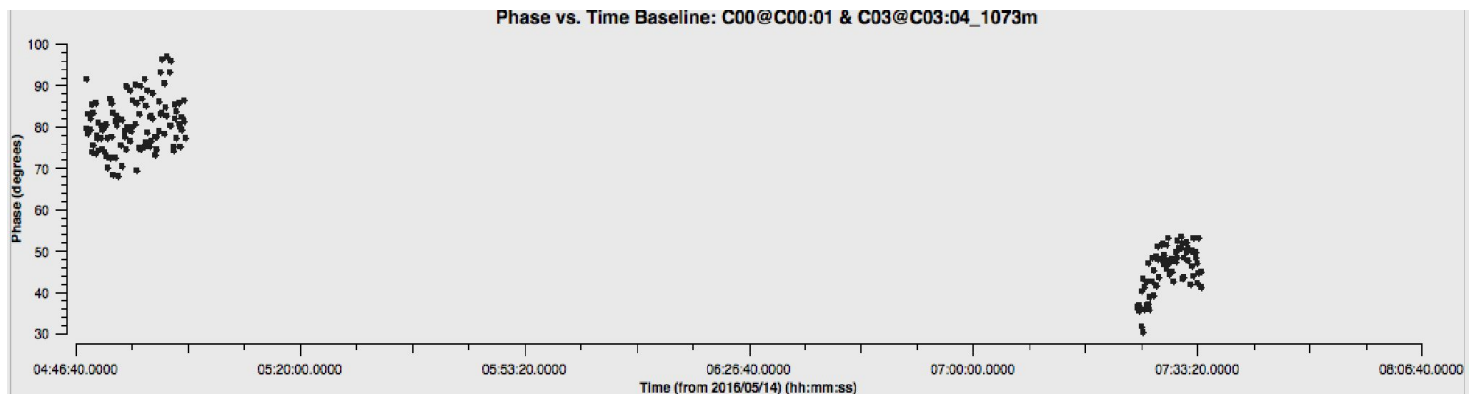
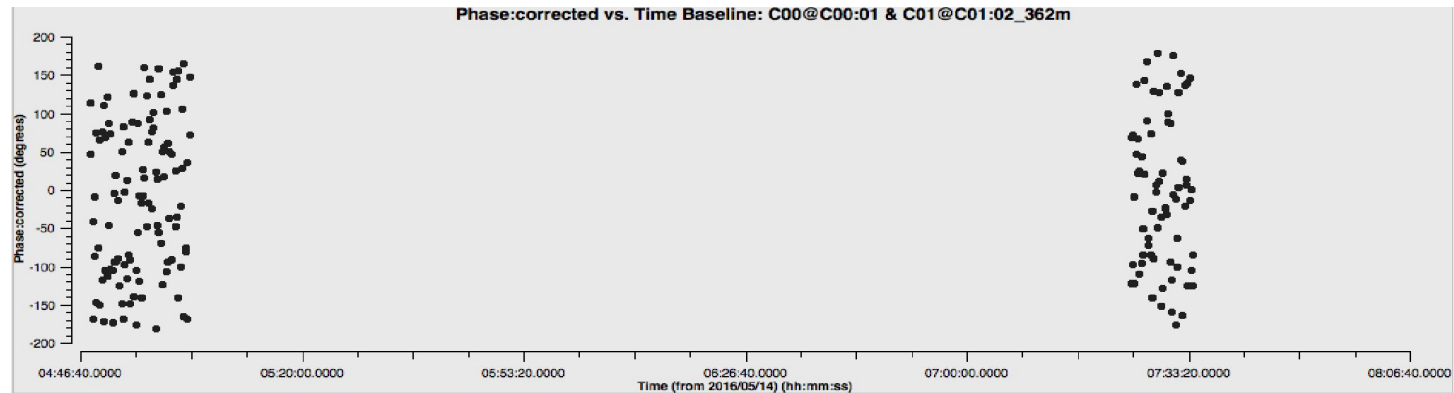


Speed

Objectivity

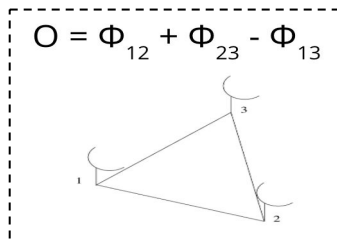
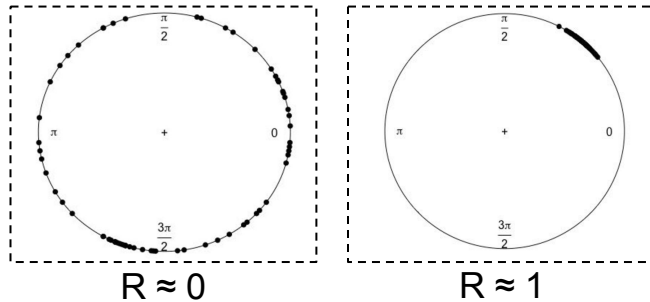
Repeatability

FLAGGING: SIGNATURES OF BAD DATA

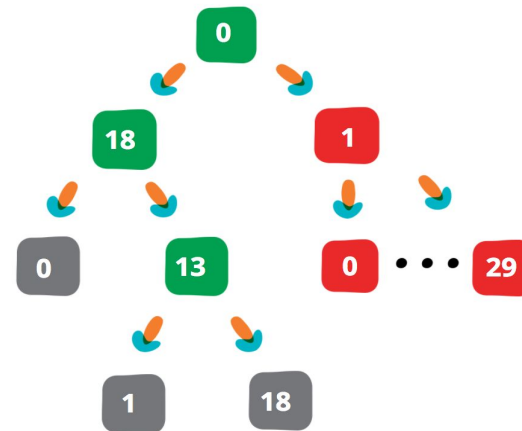


FLAGGING: INITIAL SCREENING

Angular Dispersion



Closure Phases

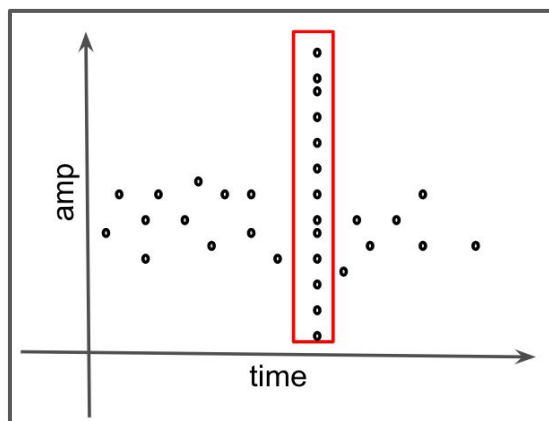


Depth First Tree Traversal

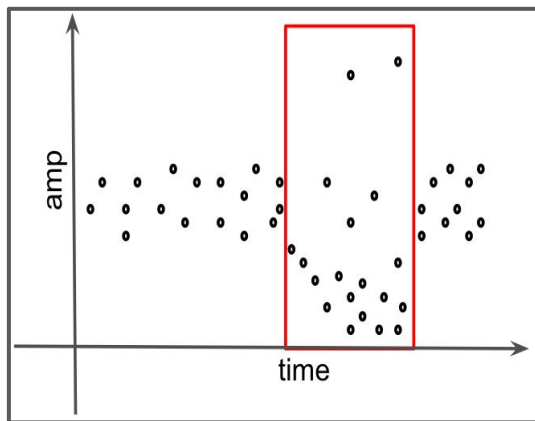
FLAGGING: DETAILED IN TIME

Window based Median and MAD Statistics

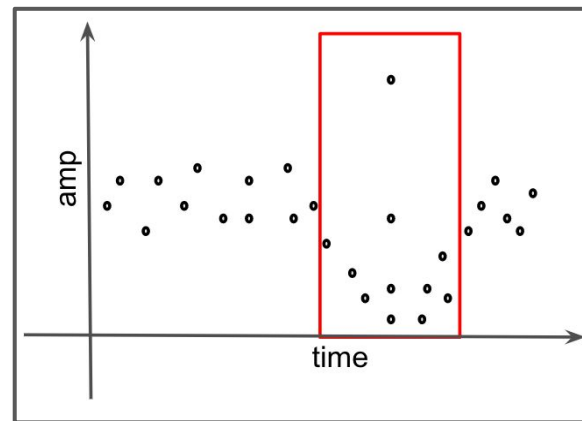
Bad Time



Bad Antenna Time



Bad Baseline Time



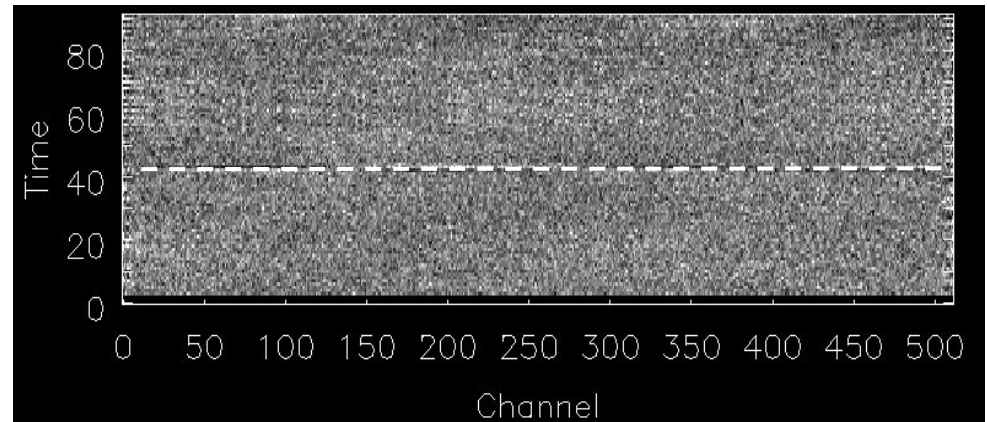
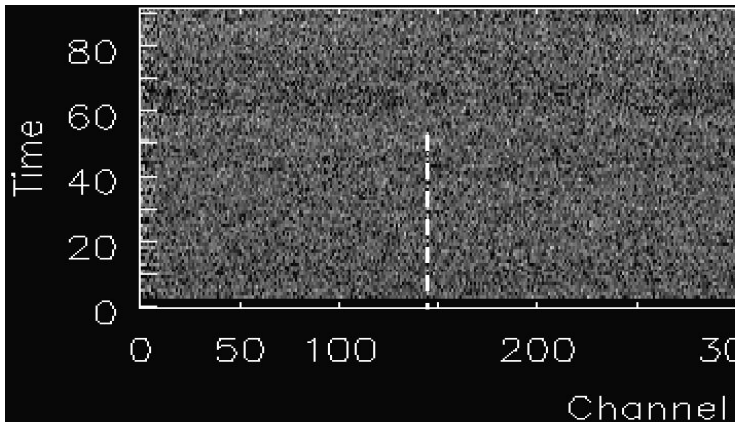
Initial and detailed analysis both run on single reference channel

FLAGGING: IN FREQUENCY

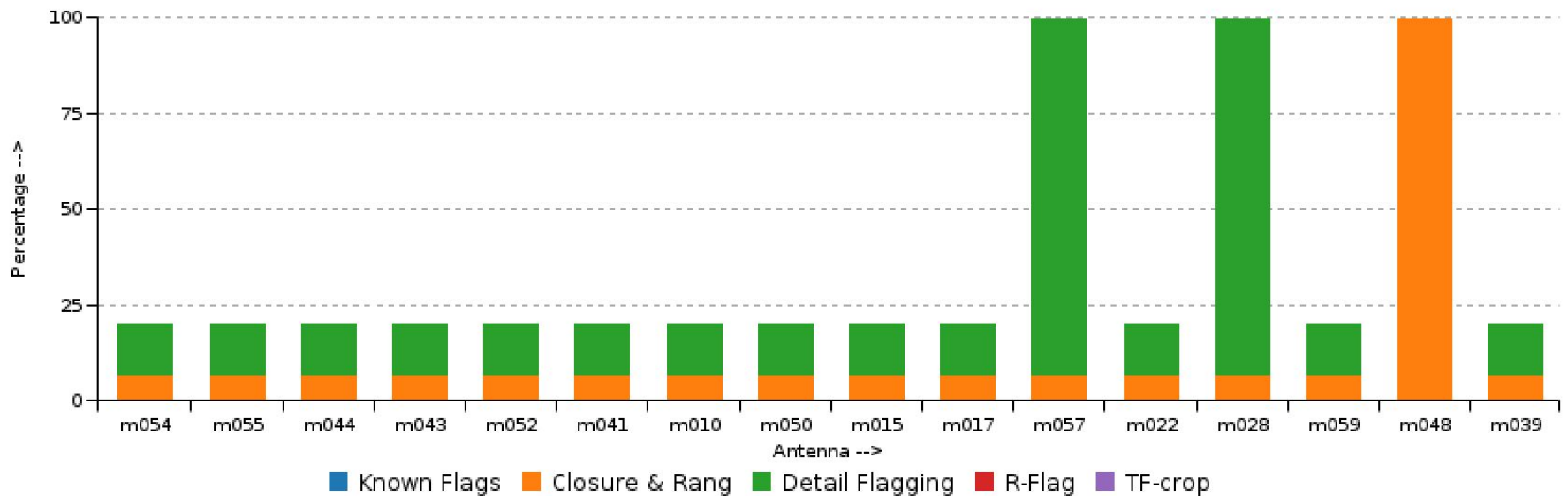
Algorithms Used



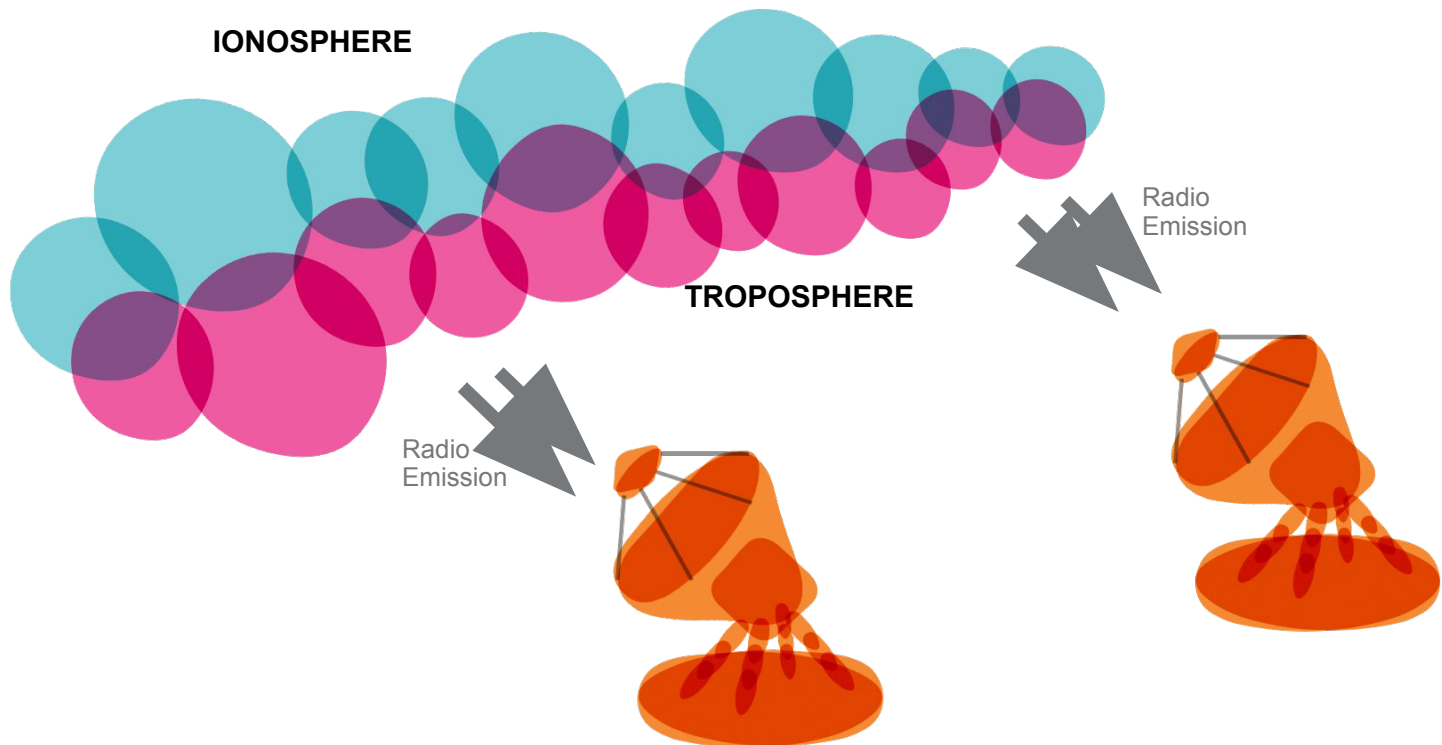
Tfcrop
Rflag



FLAGGING REPORTS

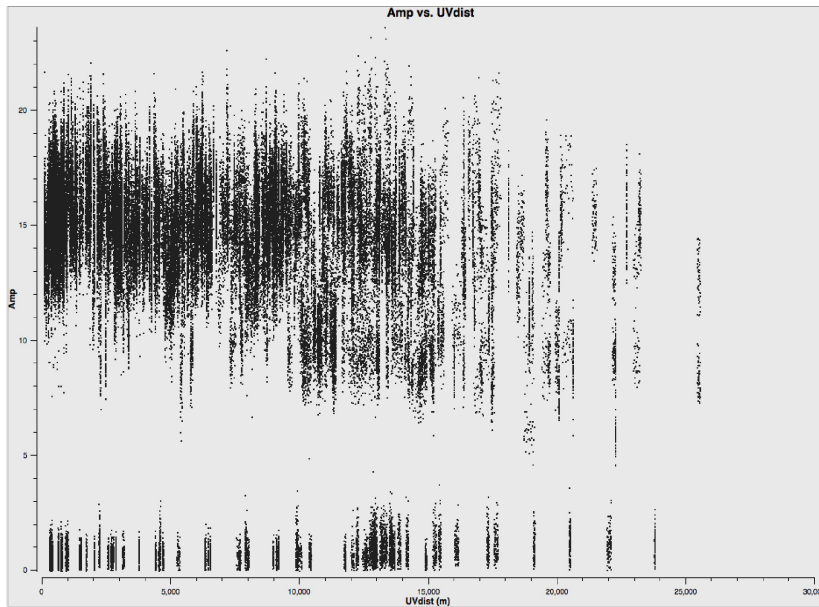


WHY CALIBRATION?

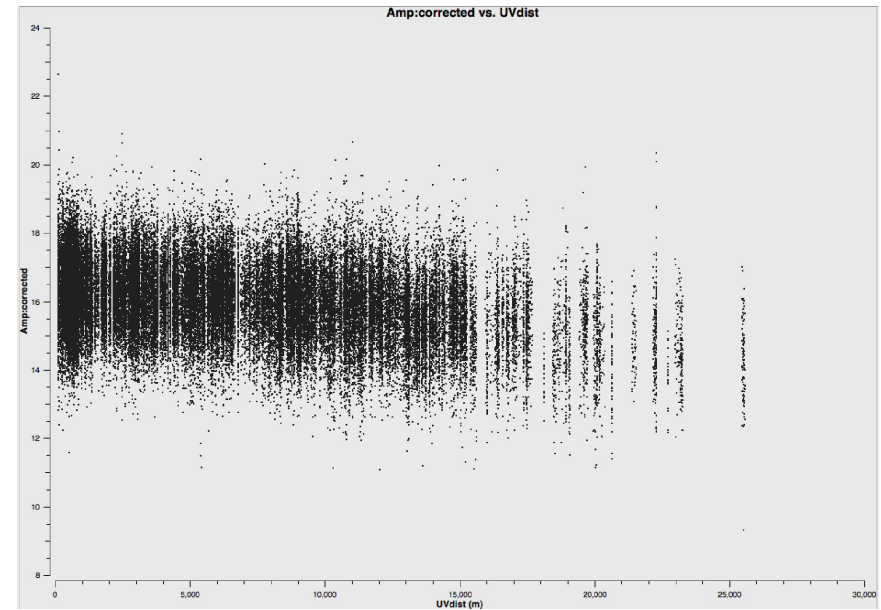


THE DIFFERENCE

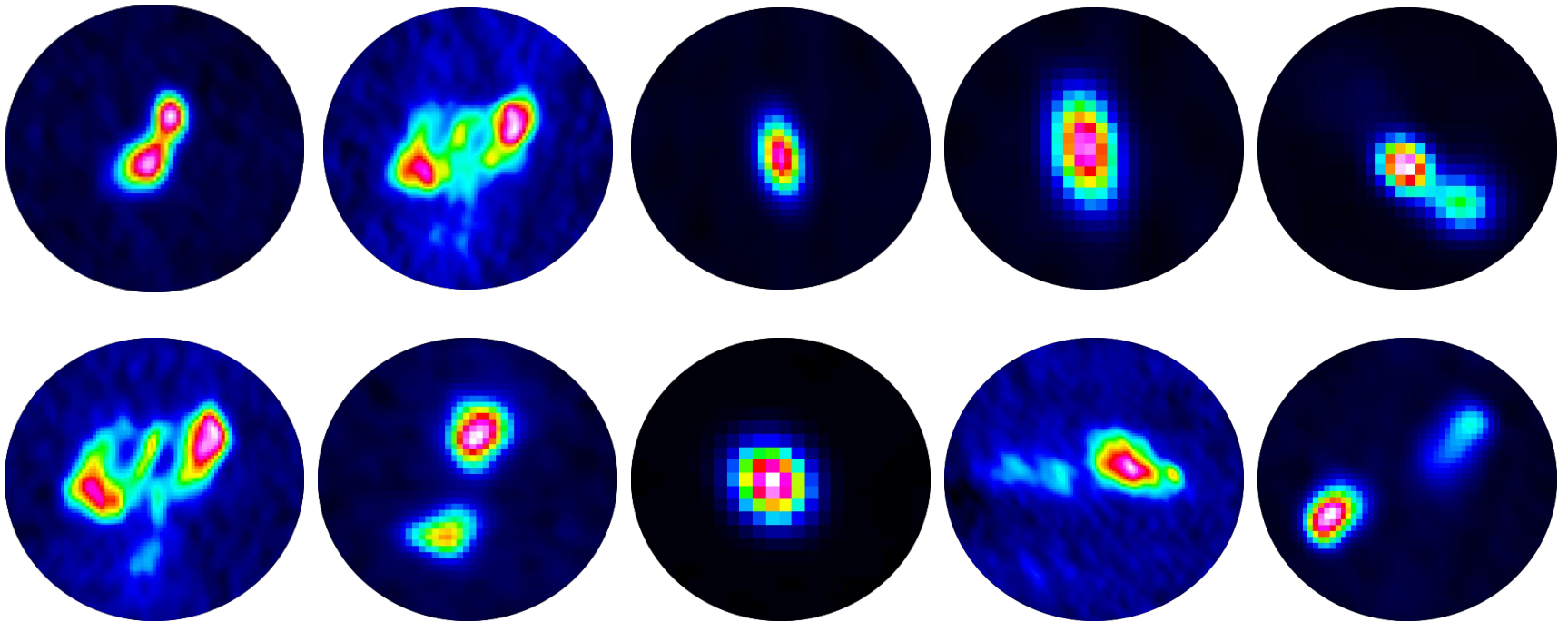
Uncalibrated



Calibrated



IMAGES GENERATED BY ARTIP



TECH STACK



C3.js D3-based reusable chart library

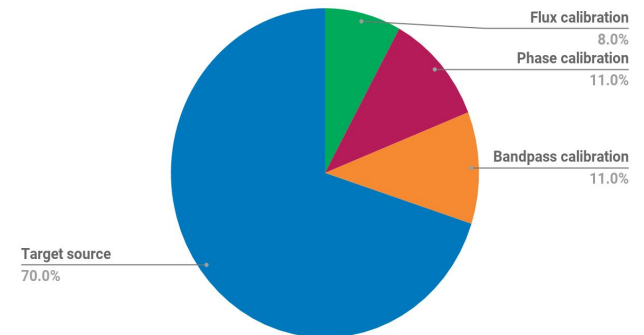
PIPELINE PERFORMANCE

Tested against GMRT and VLA telescopes datasets

RAM - 256 GB
Cores - 40
Storage - 18 TB

Server specs

Time taken by each stage



10GB in 20 mins

*For an expert astrophysicist, to manually reduce a dataset of 10 GB,
time taken is around 3-4 hours, the pipeline does it in 20mins*

IMAGING PARALLELIZATION USING CASAMPI

Simulated data of **12GB with 32768 channels for 8 seconds**

RAM - 256 GB
Cores - 40
Storage - 18 TB

Server specs

Imaging

without MPI

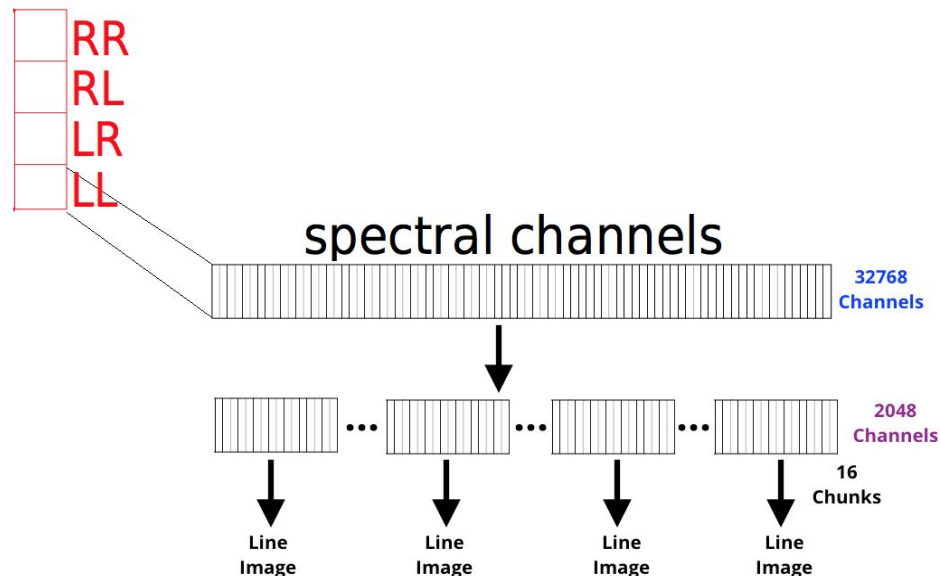
3:32:52 (h:m:s)

with MPI

8:56:08 (h:m:s)

>95% of time was spent in
imageconcat process and was
using only 1 core

IMAGING PARALLELIZATION USING DATA DIVISION



4 Polarizations X 16 chunks = 64 chunks

IMAGING PARALLELIZATION USING DATA DIVISION

RAM - 256 GB
Cores - 40
Storage - 18 TB

Server specs

13 Nodes, each node
with

RAM - 128 GB
Cores - 16
Backed up by **Lustre FS**

Cluster specs

	1 chunk	16 chunks	32 chunks	64 chunks
Size	23 GB	$23 * 16 = 368 \text{ GB}$	$23 * 32 = 736 \text{ GB}$	$23 * 64 = 1.472 \text{ TB}$
Success	All	All	All	Server - 18 succeeded, 46 failed
Server (40 Cores)	0:22:51	1:29:15	3:08:10	3:13:08
Cluster (208 Cores)	0:31:46	0:51:11	0:55:24	1:16:21
% Gain	-39.0226112 3	42.65172736	70.5580159 4	60.46772523

CHALLENGES

Domain

- Understanding the requirements
- Changing requirements
- Closed community

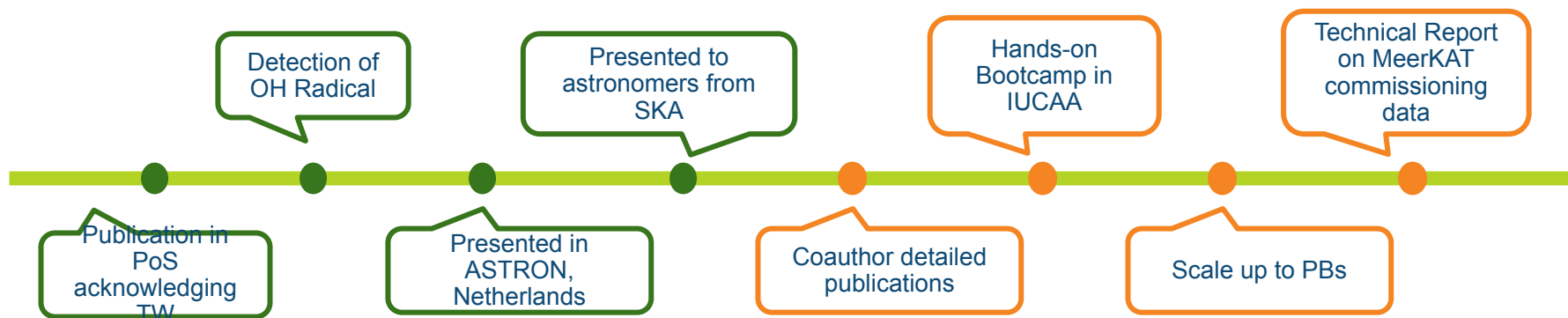
Tech

- Limited tools that are difficult to integrate
- Freedom of experimentation without increase in complexity

Data

- New aspects with each new dataset
- Uncommon data formats restricting use of big data tools

MILESTONES



CONTRIBUTORS



Dr. Neeraj Gupta



Dolly Gyanchandani



Unmesh Joshi



Sarang Kulkarni



Santosh Mahale



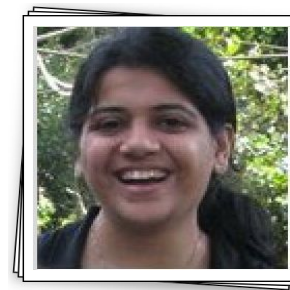
Arti Pande



Vineet Pathak



Ravi Sharma



Gunjan Shukla



Chhaya Yadav

THANK YOU

