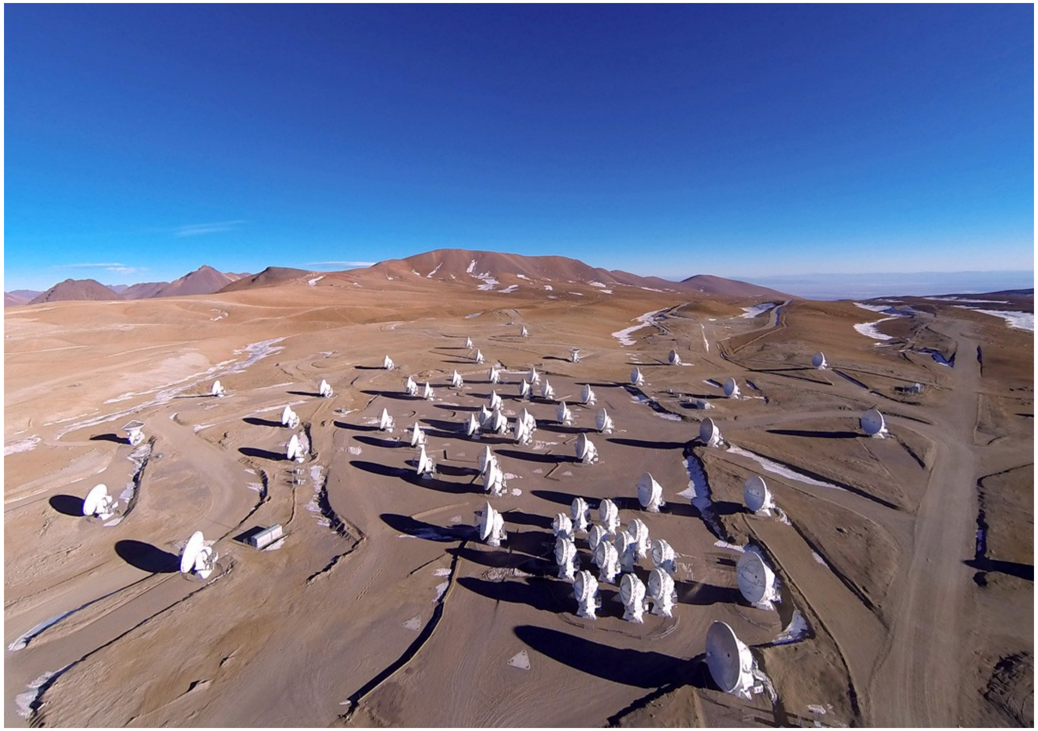
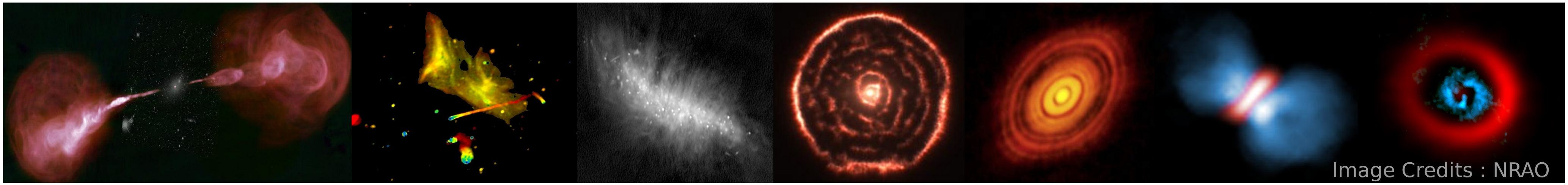
Computing and Algorithms in Radio Interferometry



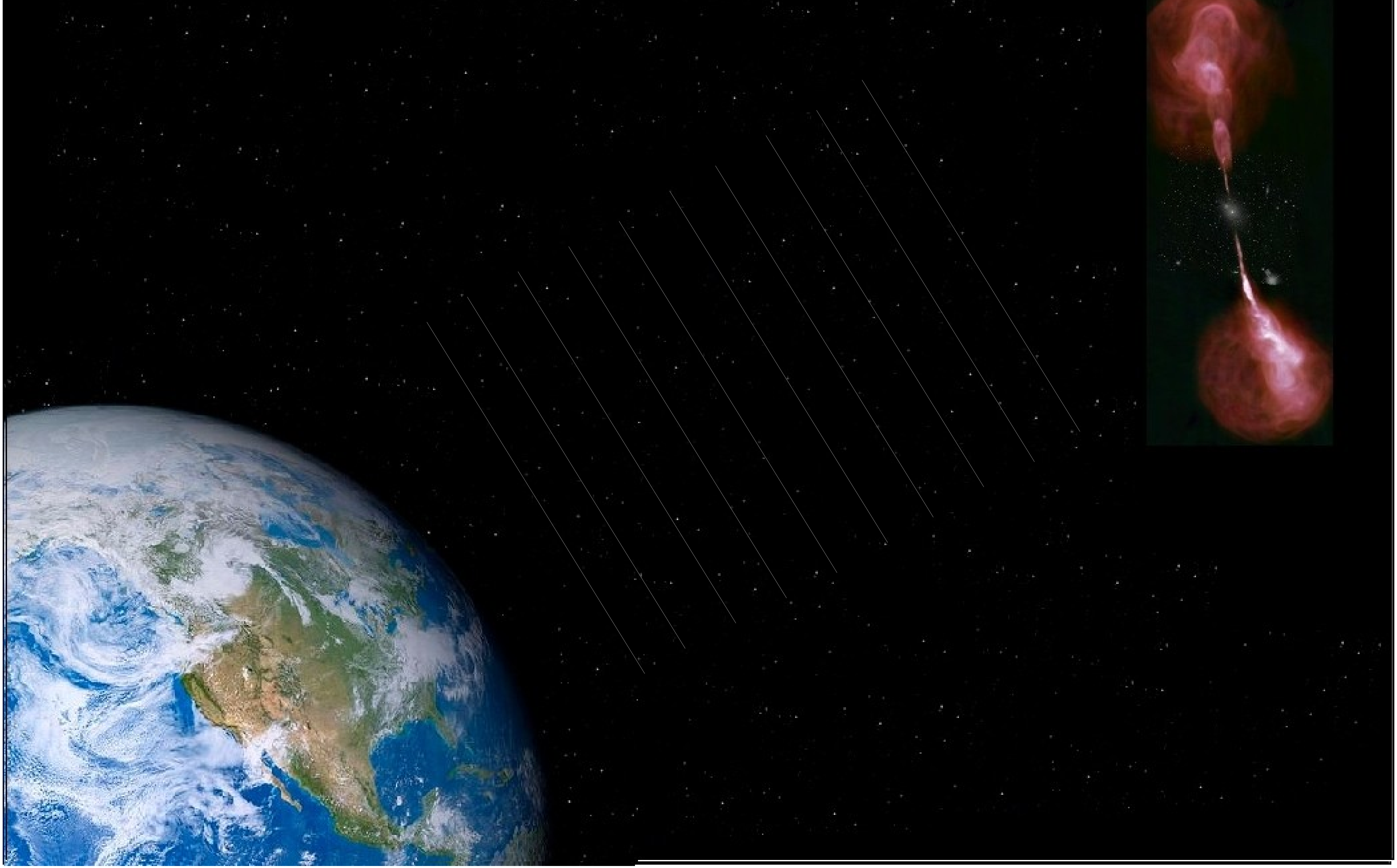


Urvashi Rau
National Radio Astronomy Observatory

20 April 2018
Computer Science & Engineering Dept, New Mexico Tech

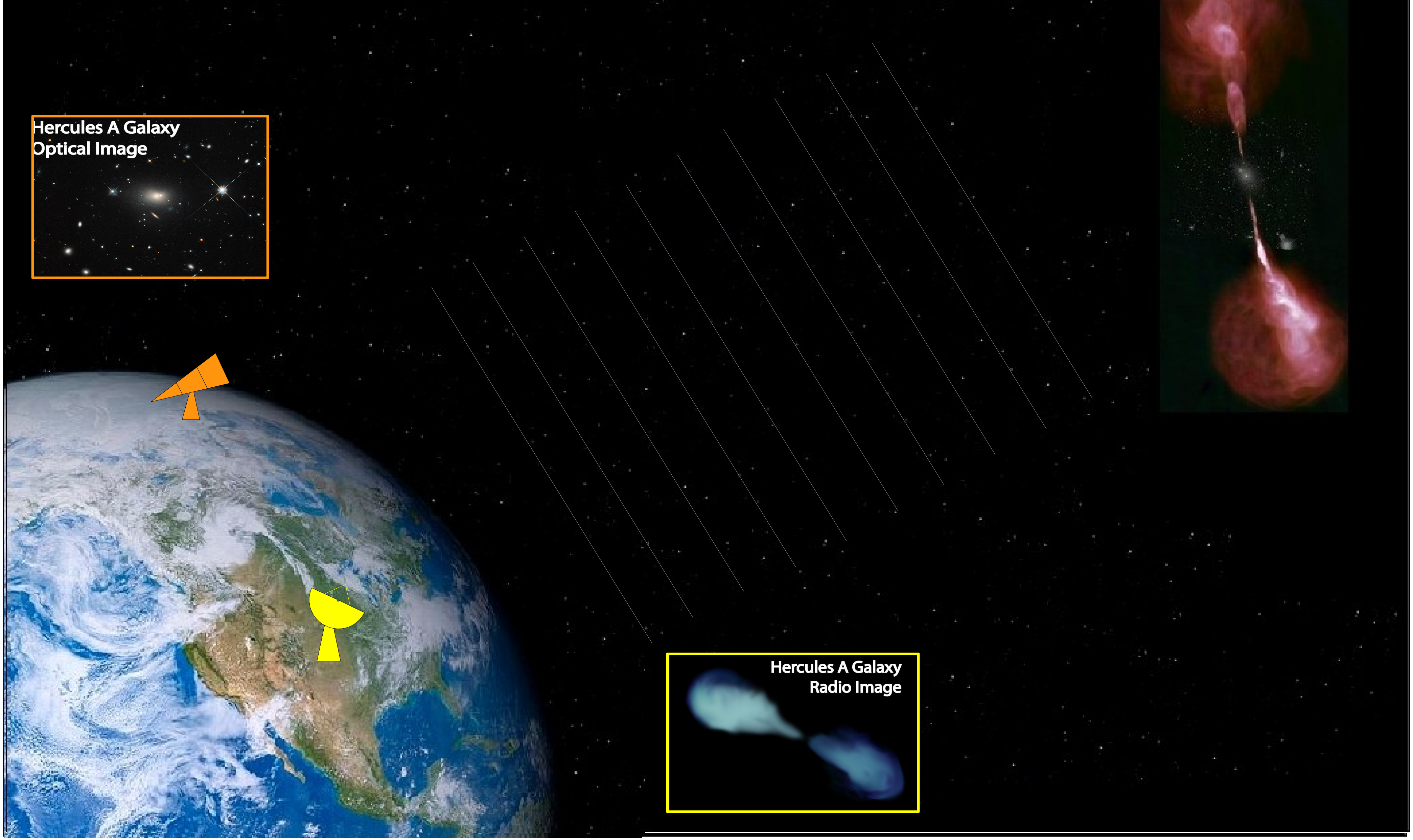
Observational Astrophysics

Space is a unique laboratory for extreme Physics



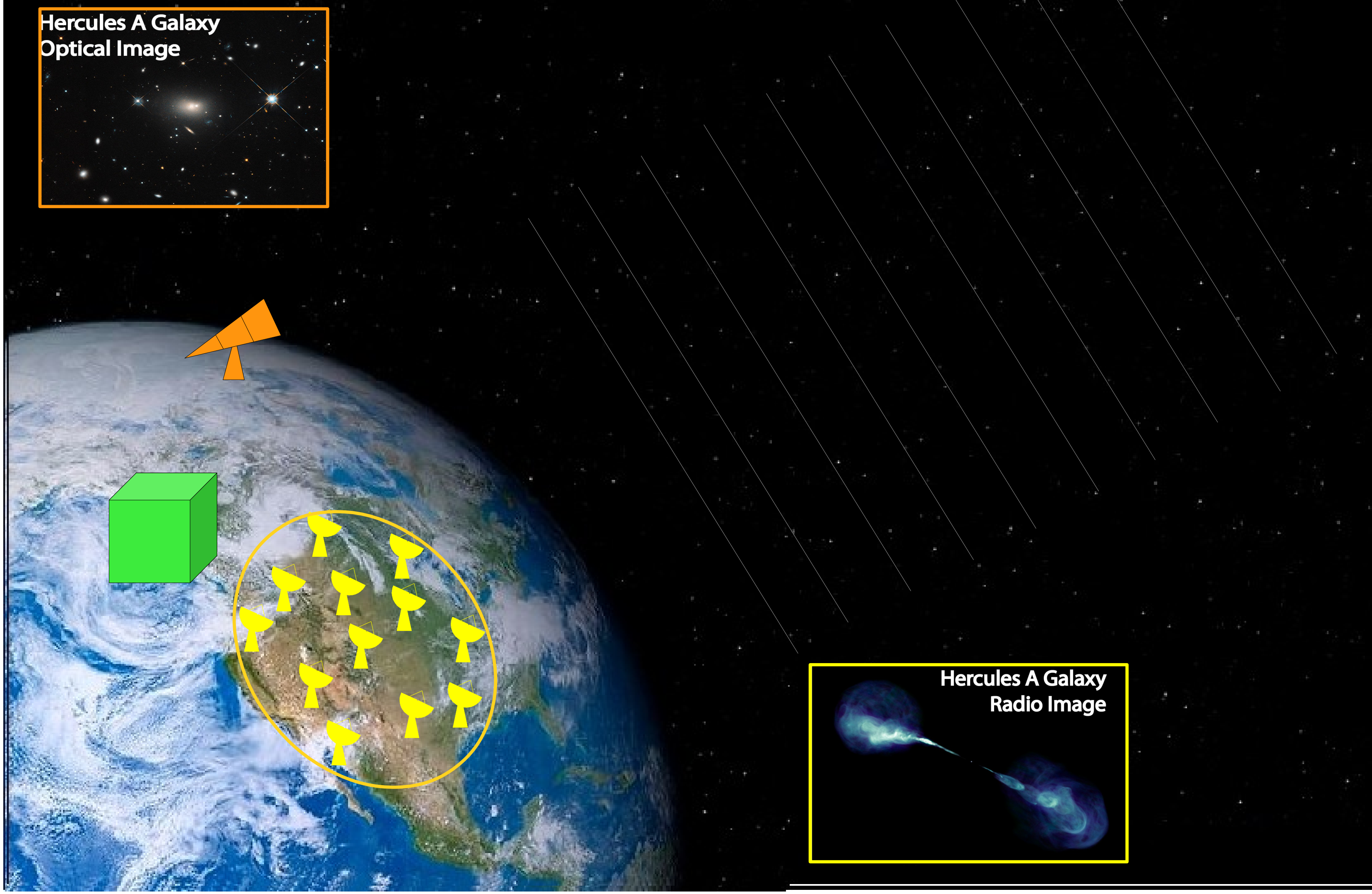
Observational Astrophysics

Make images at different parts of the electromagnetic spectrum

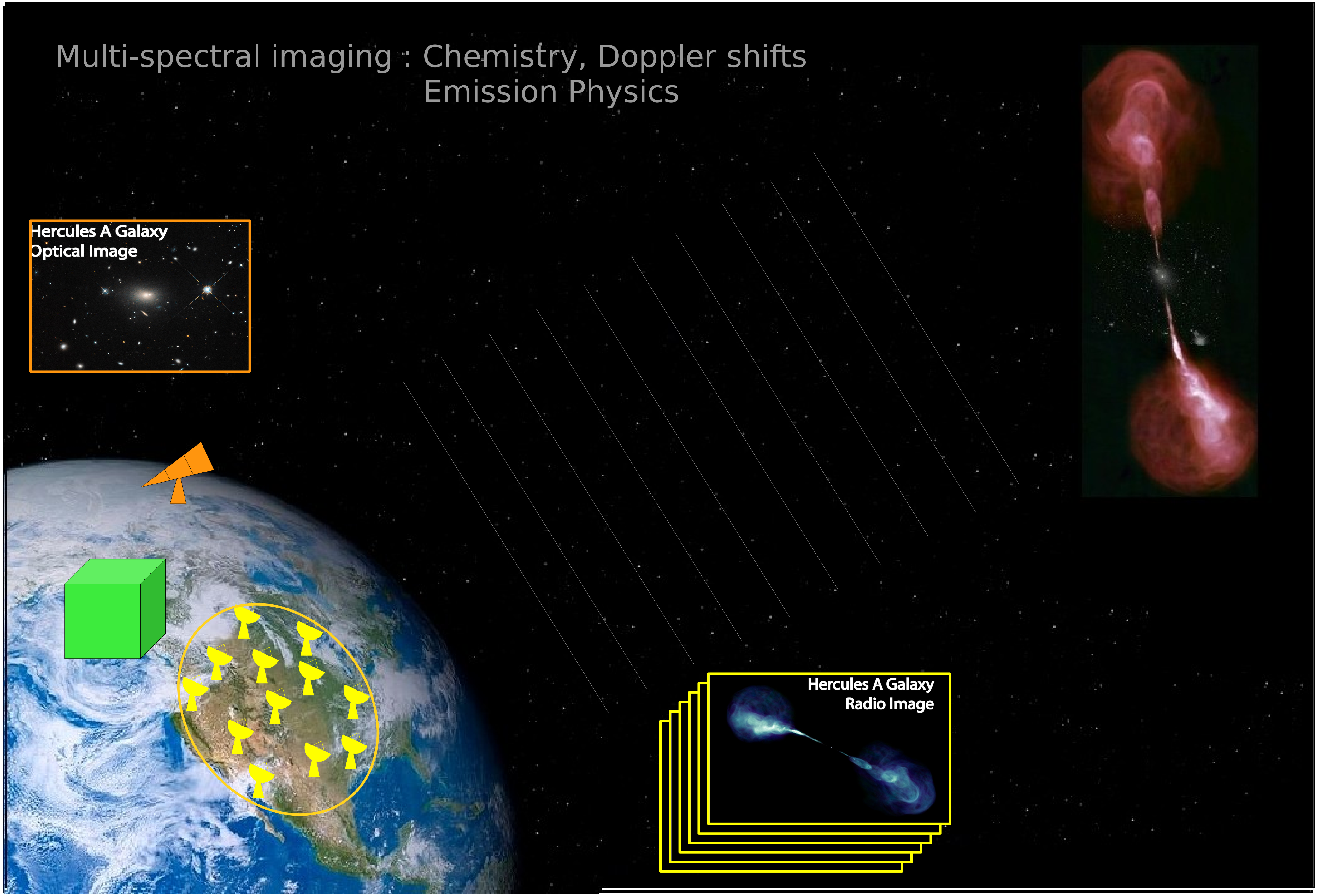


Observational Astrophysics

Use a radio interferometer to
"synthesize" a very large dish

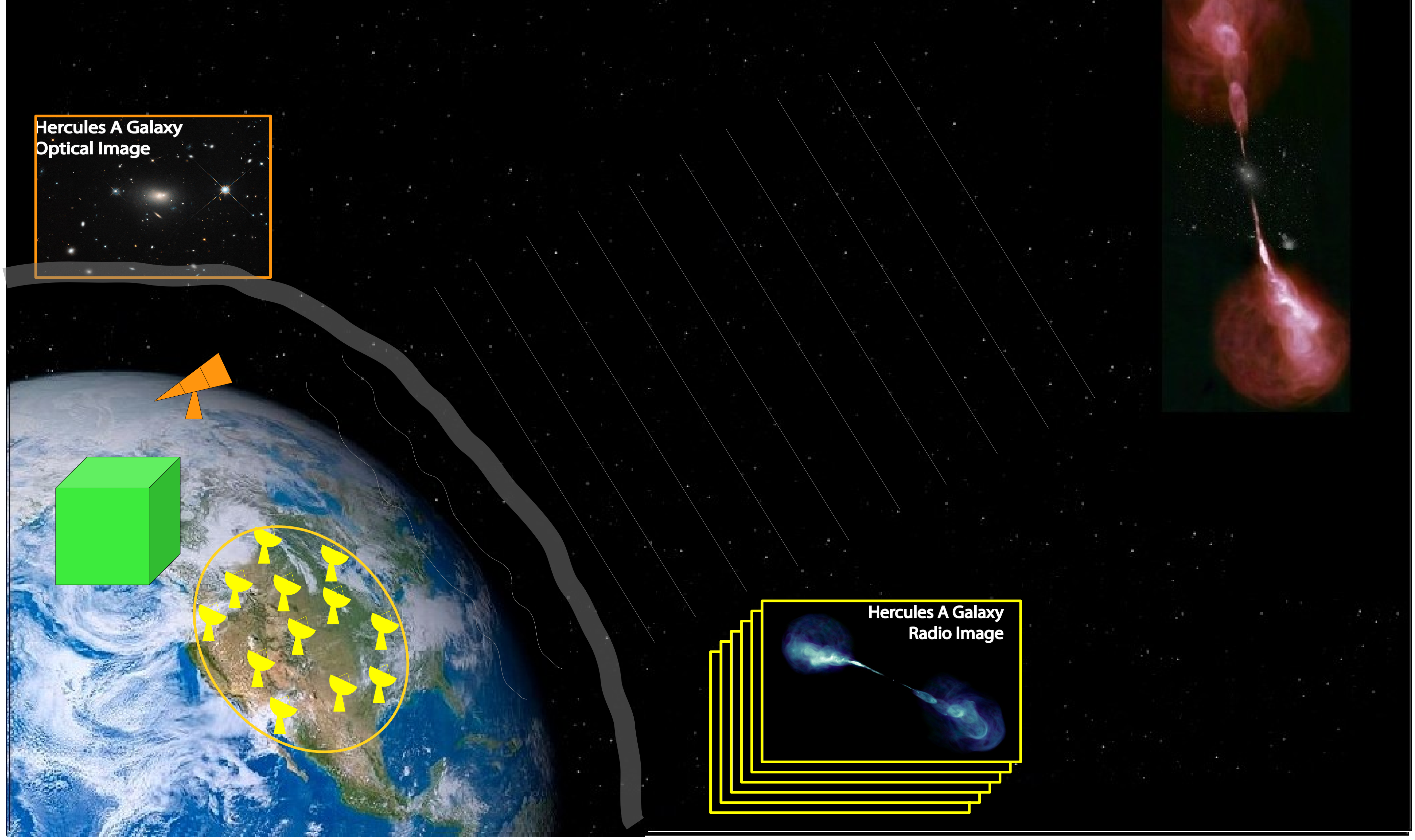


Observational Astrophysics



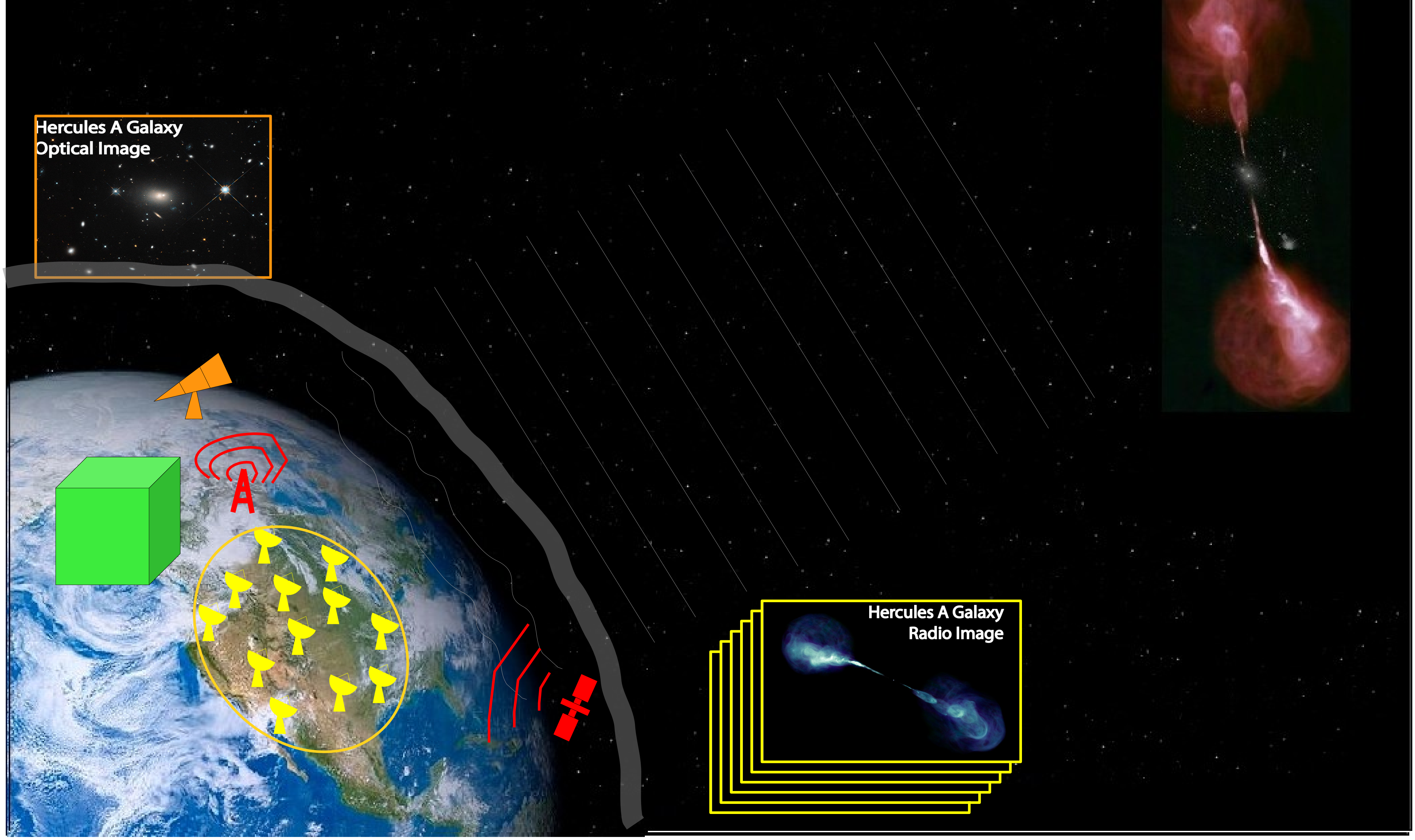
Observational Astrophysics

Received and measured signals are non-ideal
=> Use algorithms to reconstruct sky signals



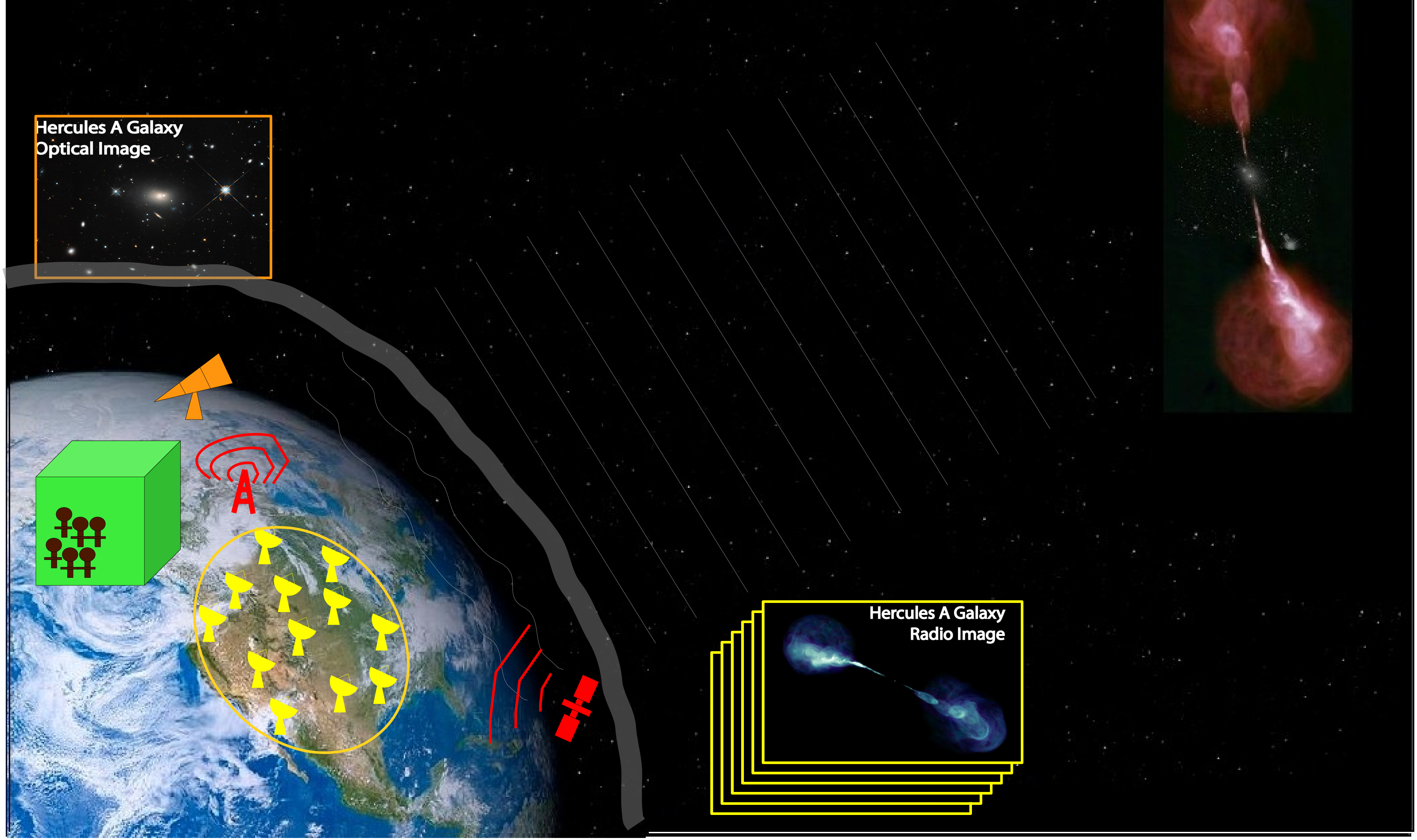
Observational Astrophysics

Received and measured signals are non-ideal
=> Use algorithms to reconstruct sky signals

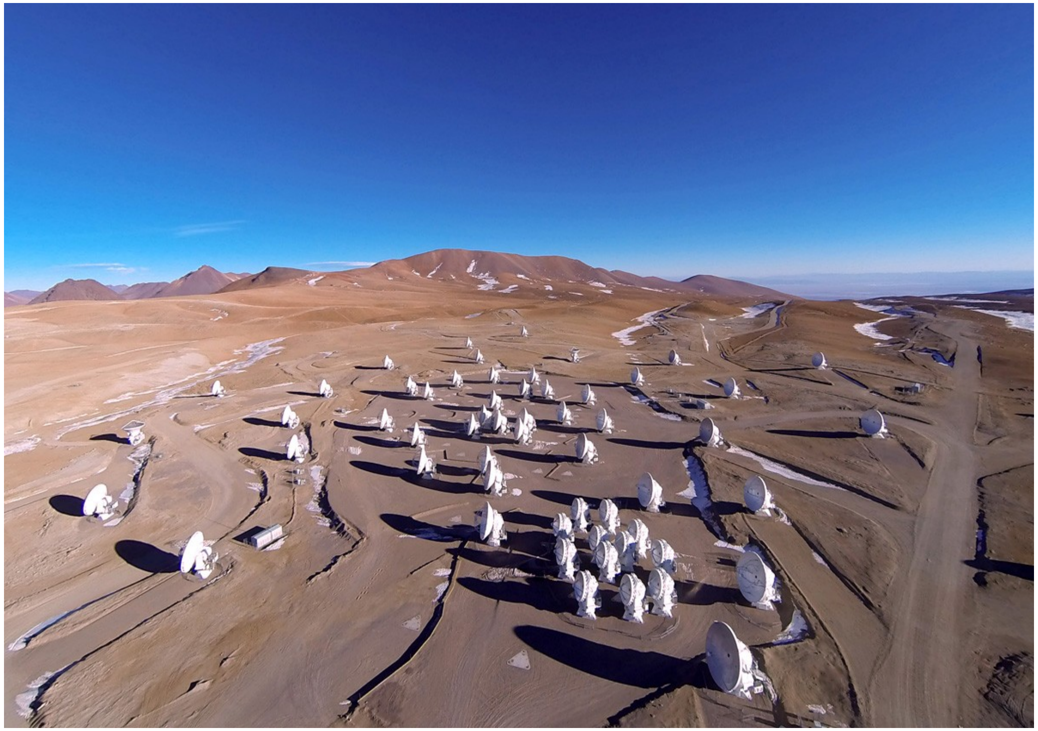


Observational Astrophysics

Many data analysis strategies
=> Algorithms operated by trained humans



Current NRAO radio interferometers



Very Large Array (1975+)

27 dishes (25m each)
1-50 GHz
4 configs (1,3,10,30 km)

Typical data rate : 1 TB / day

Manual data analysis ==>
Assisted pipeline processing

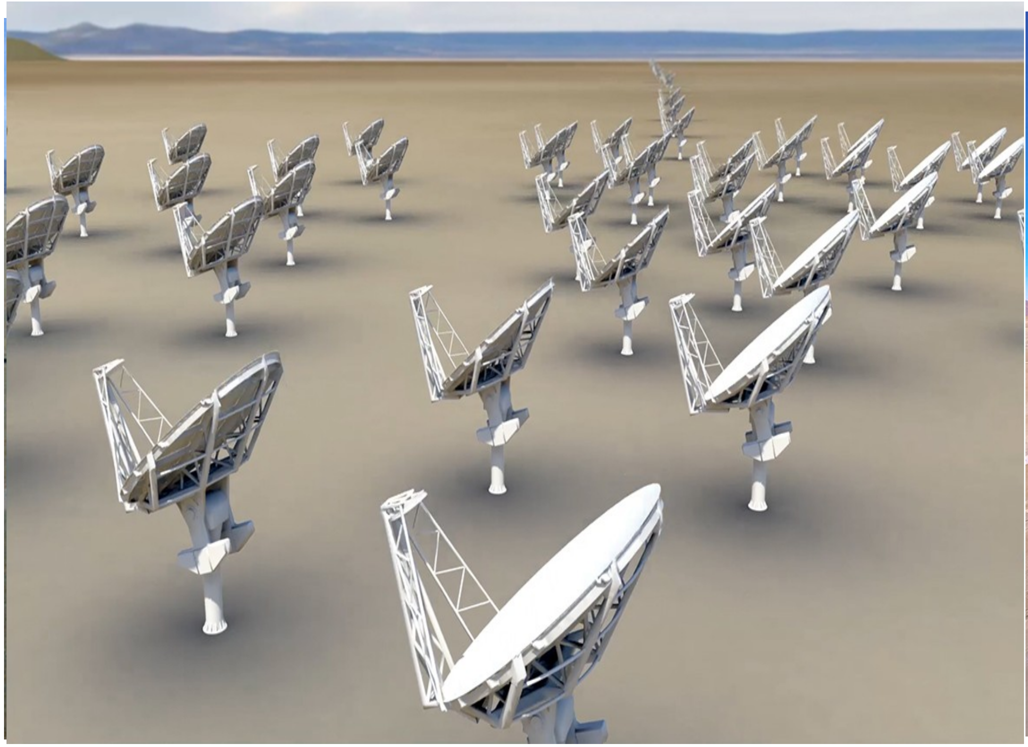
Atacama Large Millimeter Array (2011+)
(partners are ESO, NAOJ)

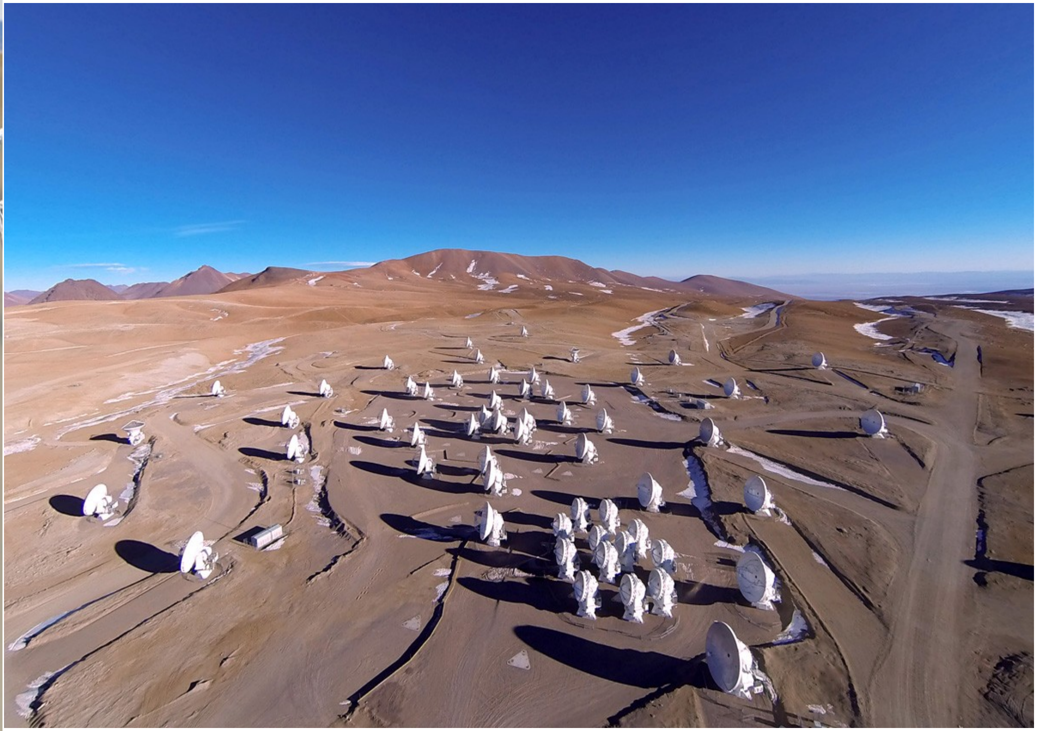
60 dishes (12m + 7m)
35-950 GHz
150m – 16km + Short spacing array

Typical data rate : 700 GB / day

Assisted Pipeline processing

Future NRAO radio interferometers





Next Generation VLA (2030 – if funded)

214 dishes (18m each + short spacing)
1.2 – 50.5, 70-116 GHz GHz
Fixed config : 1000km baselines

Expected data rate : ~ 1 PB / day

Science Ready Data Products

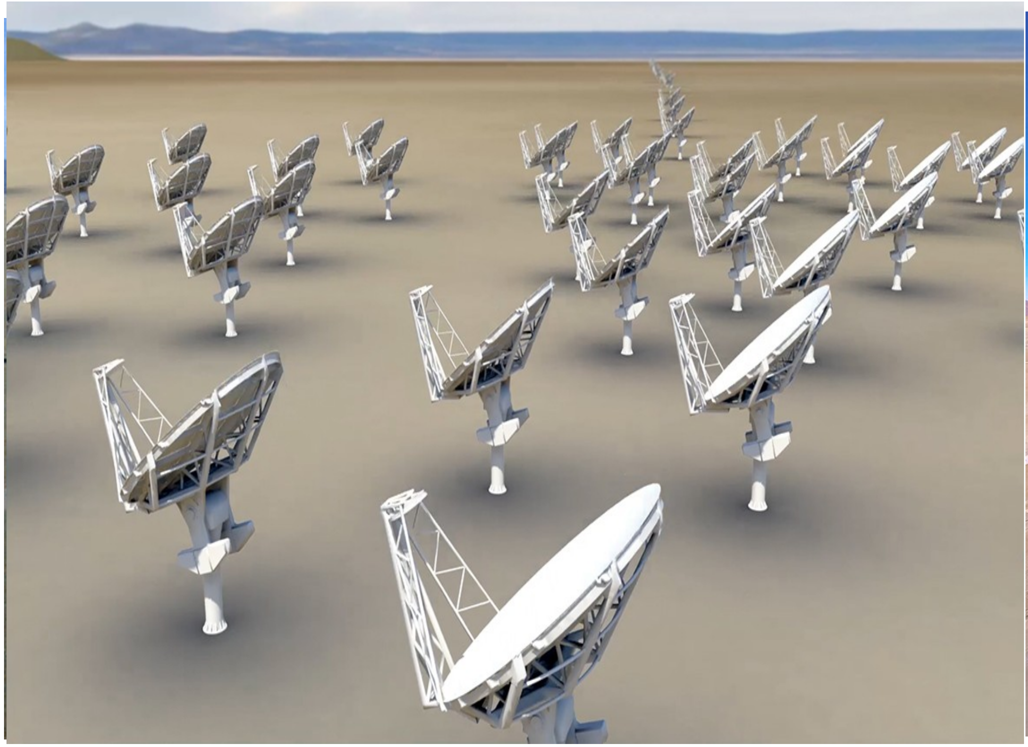
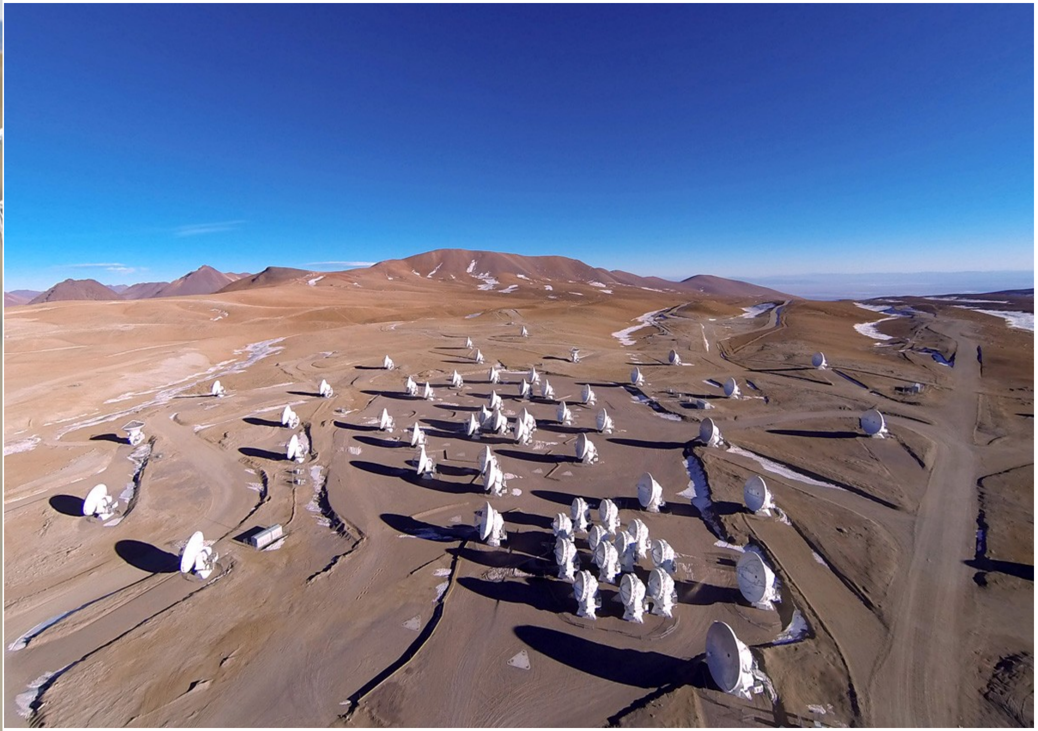
ALMA upgrades (2020+)

60 dishes (12m + 7m short spacings)
35-950 GHz
Baselines of 150m – 16km

Expected data rate : 2.5 TB / day

Assisted pipelines ==> Science Ready Data Products

Future NRAO radio interferometers

Major areas of focus (related to Computer Science)

(1) High Performance Computing
(2) End-to-end Automation

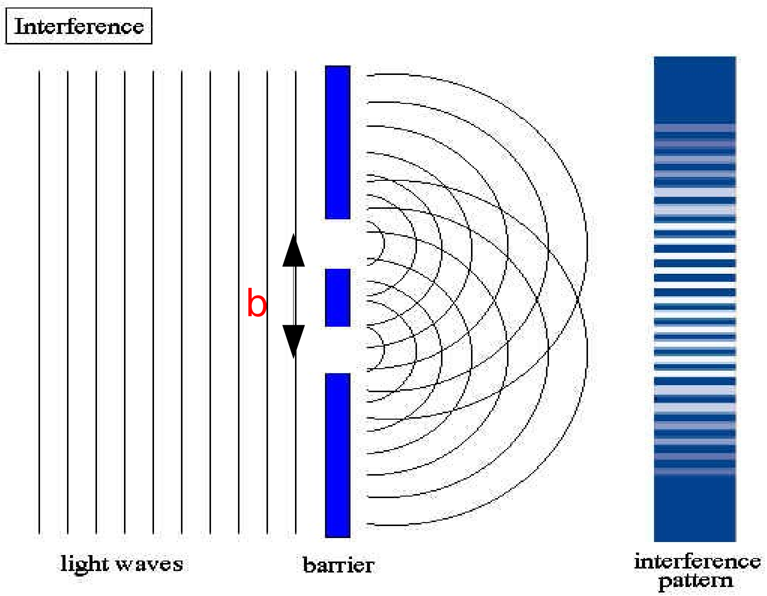
Technical projections show that both problems are tractable

But, current practices are not optimal.

Work is required to optimize throughput and algorithmic reliability.

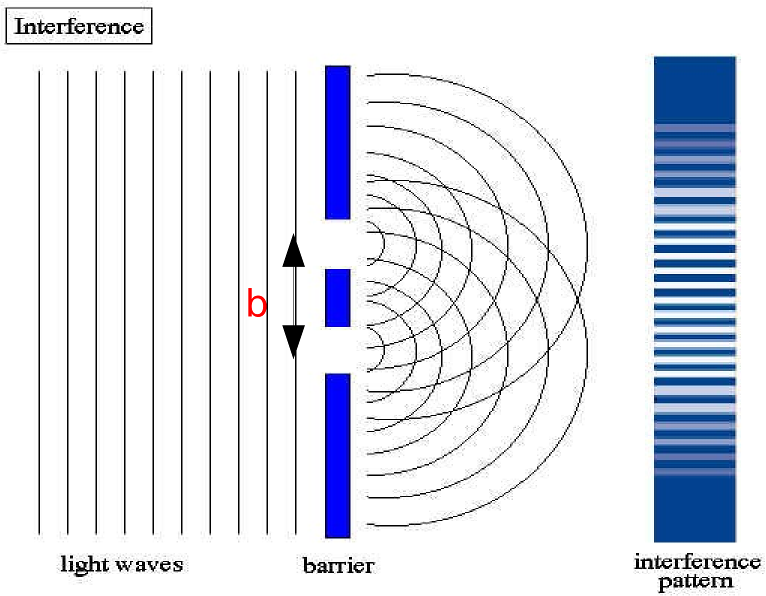
An interferometer is an indirect imaging device

Young's double slit experiment

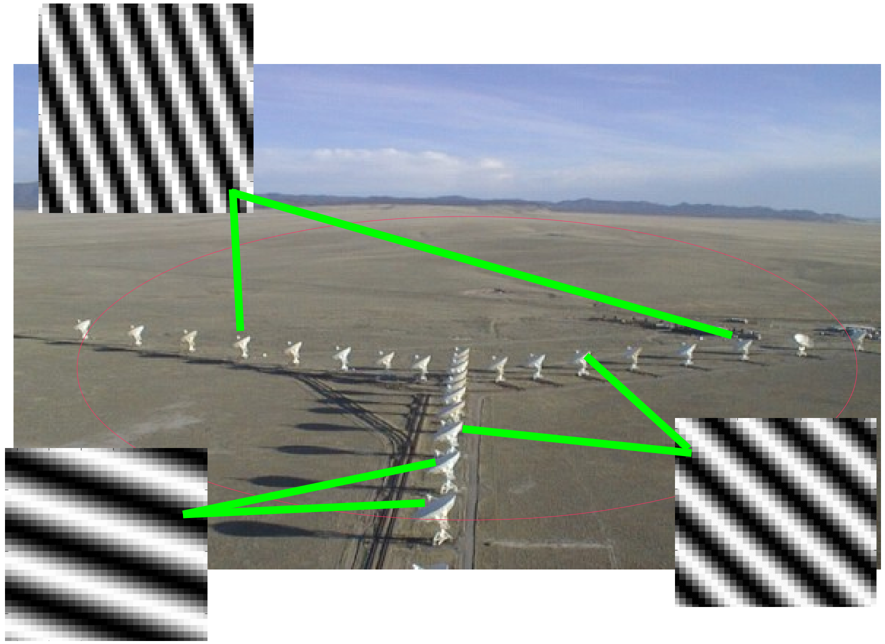


An interferometer is an indirect imaging device

Young's double slit experiment

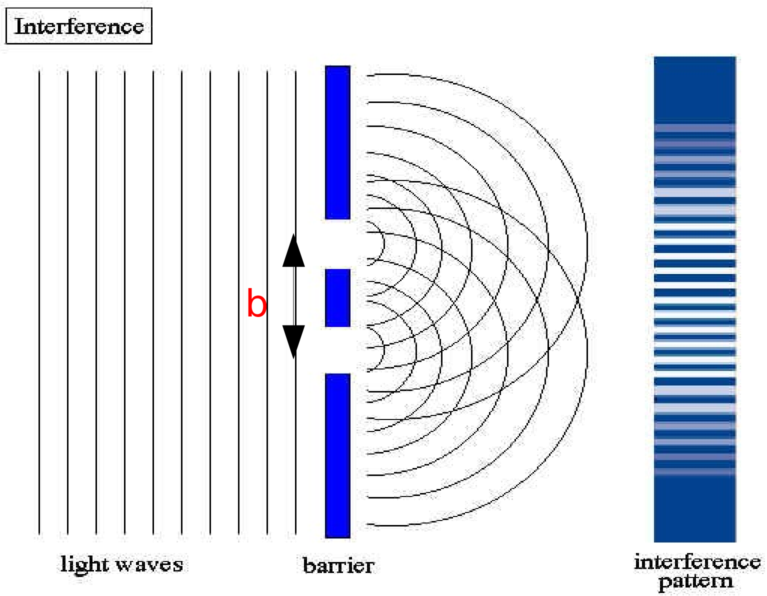


Each antenna-pair => one 2D fringe

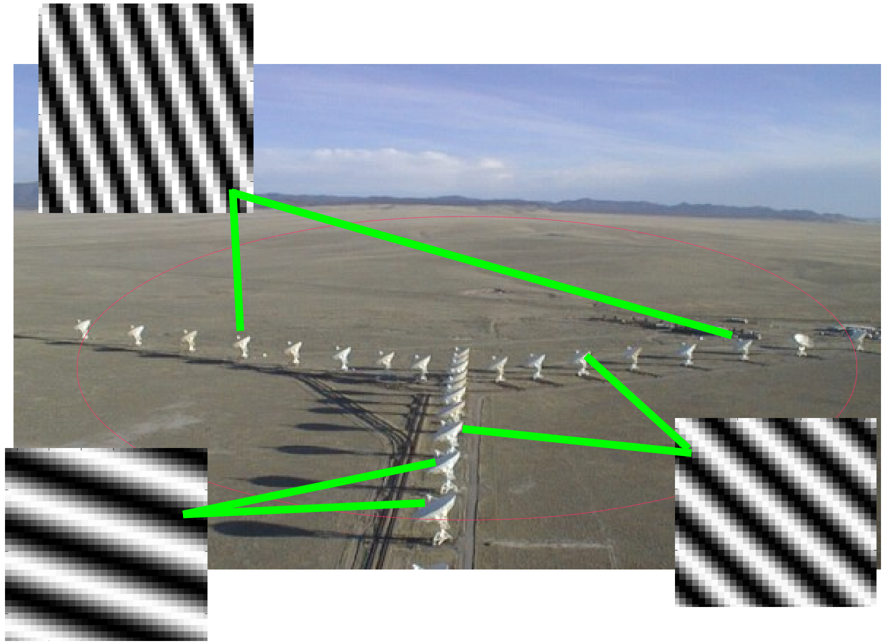


An interferometer is an indirect imaging device

Young's double slit experiment



Each antenna-pair => one 2D fringe



2D Fourier transform :

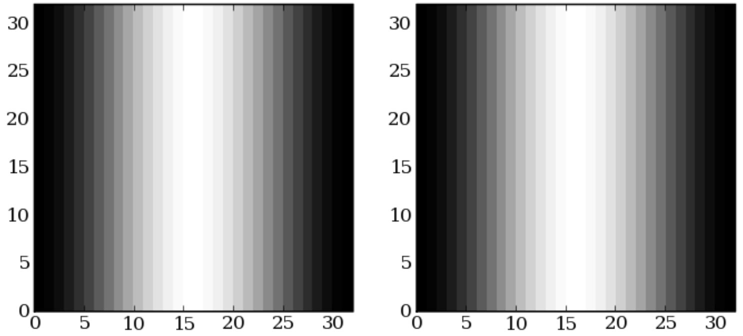
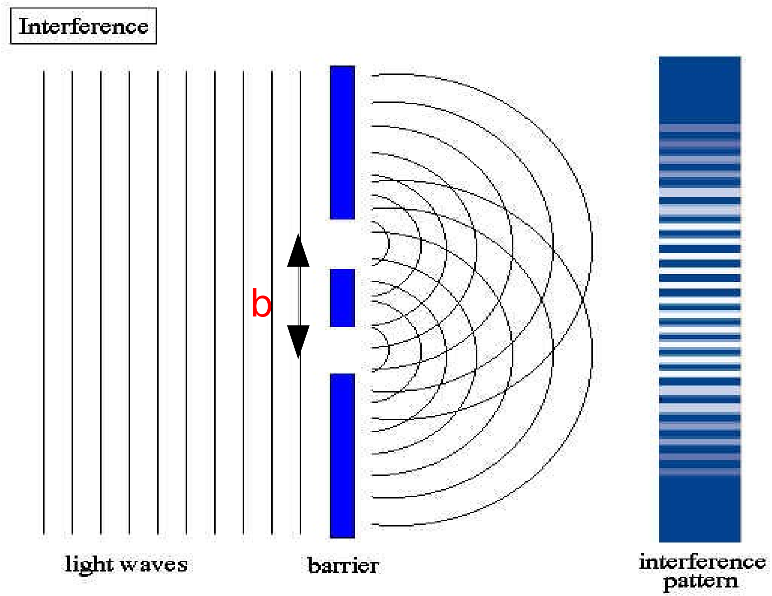


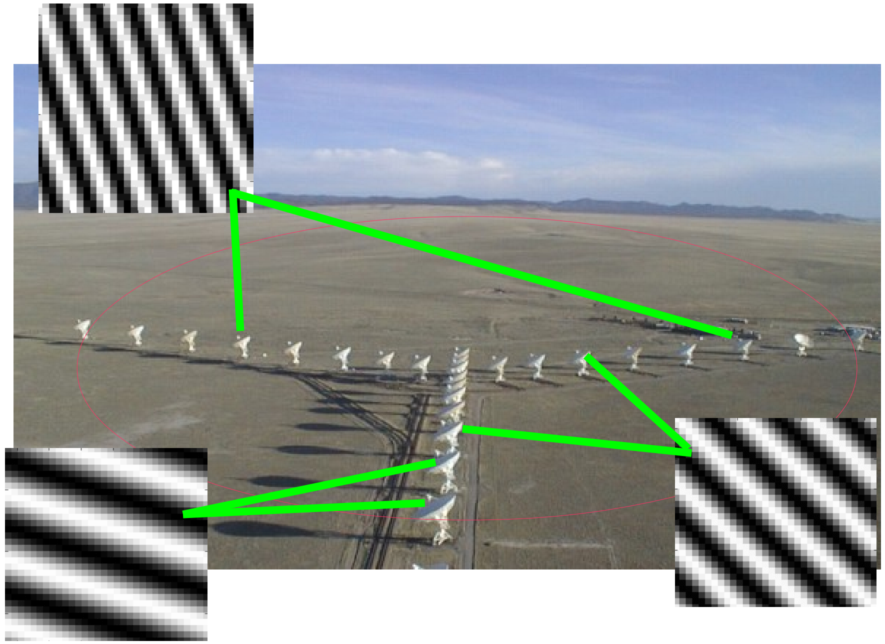
Image = sum of cosine 'fringes'.

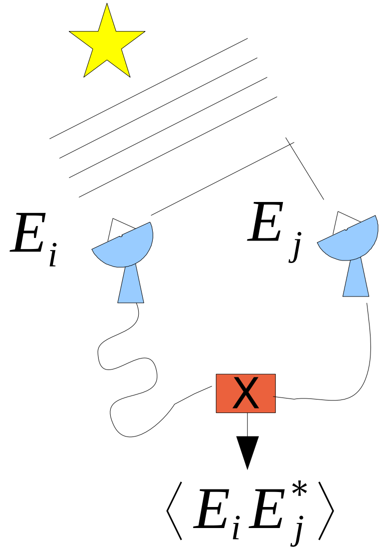
An interferometer is an indirect imaging device

Young's double slit experiment



Each antenna-pair => one 2D fringe





Measuring fringe parameters

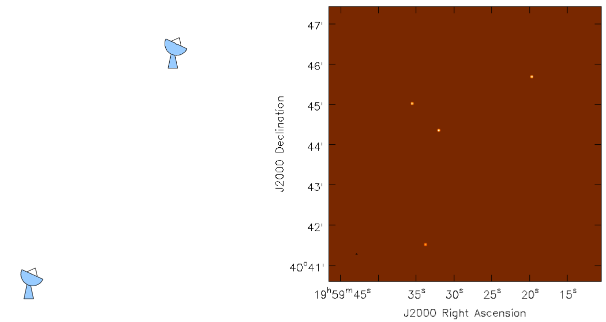
Amplitude, Phase : $\langle E_i E_j^* \rangle$ is a complex number

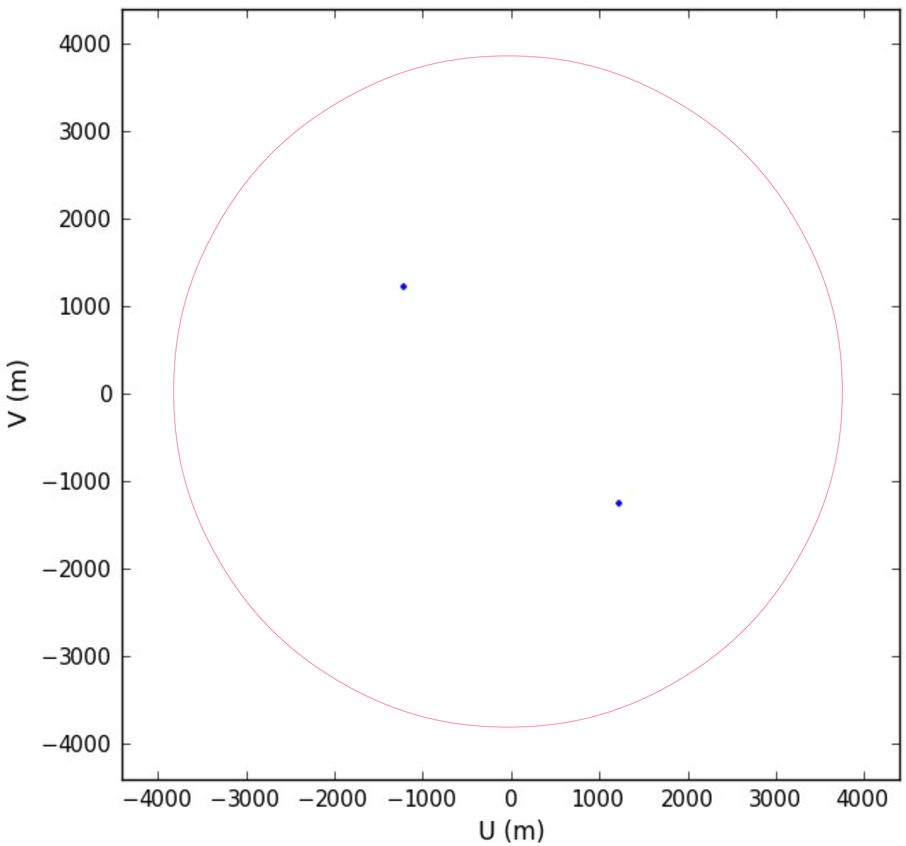
Orientation, Wavelength : Vector between each pair of antennas

Goal : Measure as many distinct fringes as possible

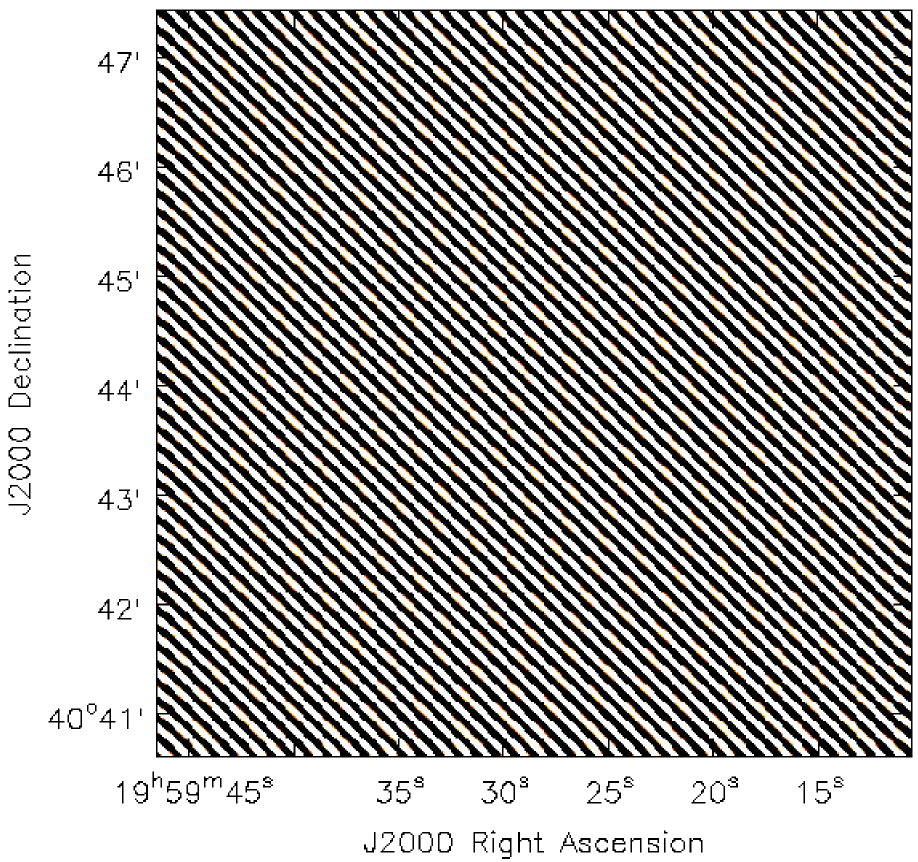
Aperture Synthesis Imaging

Image with 2 antennas => 1 fringe





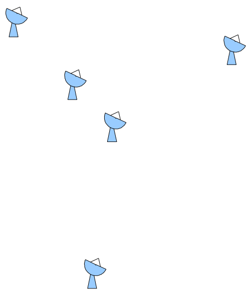
Synthesized Aperture

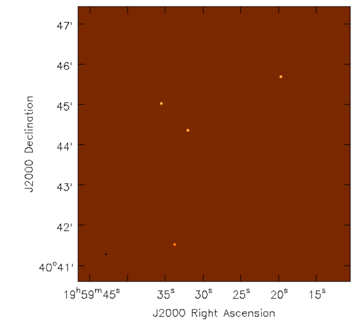


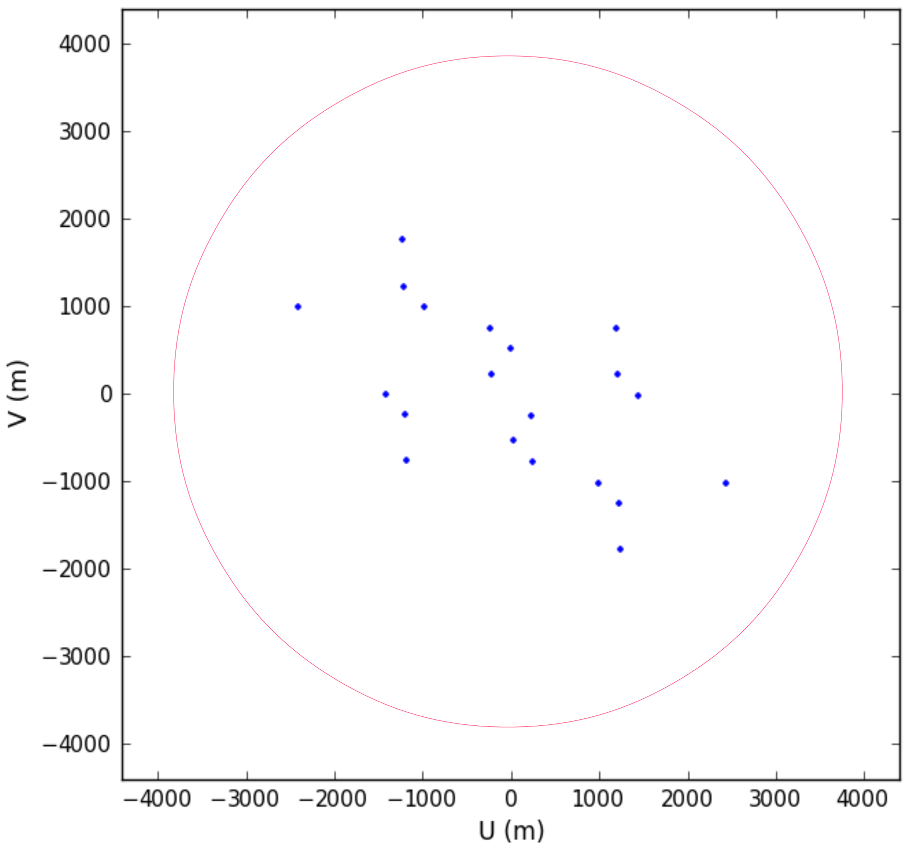
Observed Image

Aperture Synthesis Imaging

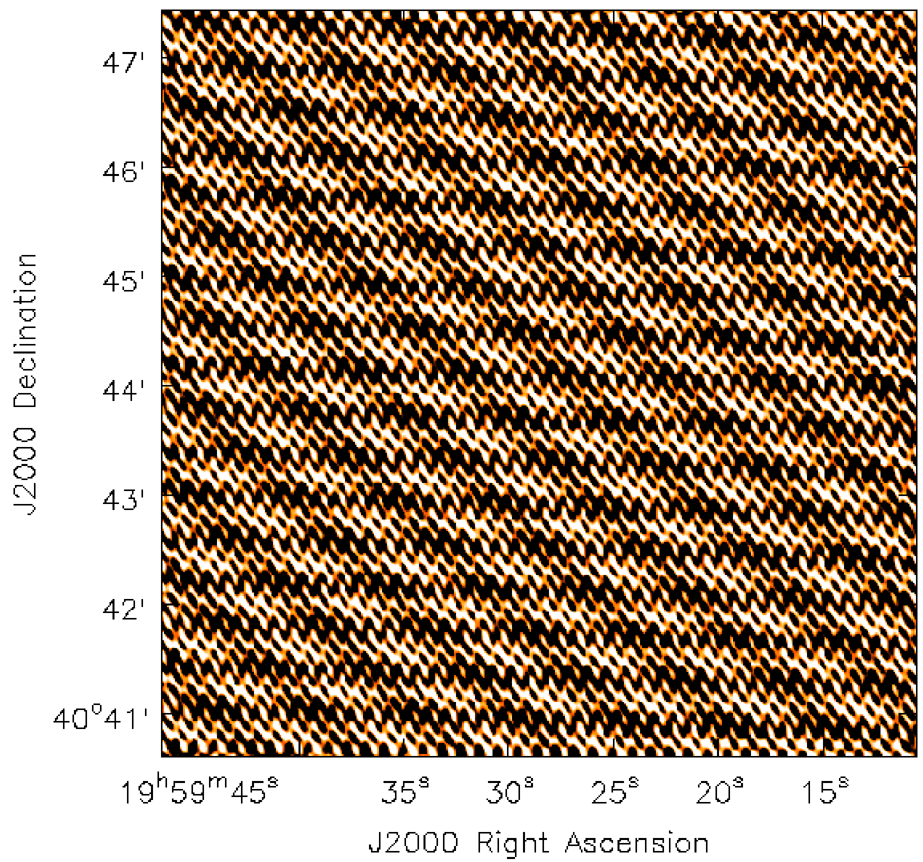
Image with 5 antennas => 10 fringes







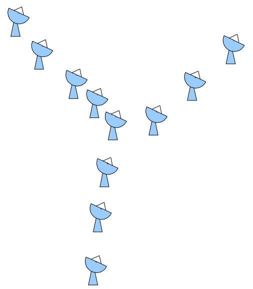
Synthesized Aperture

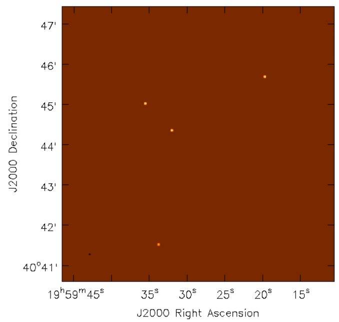


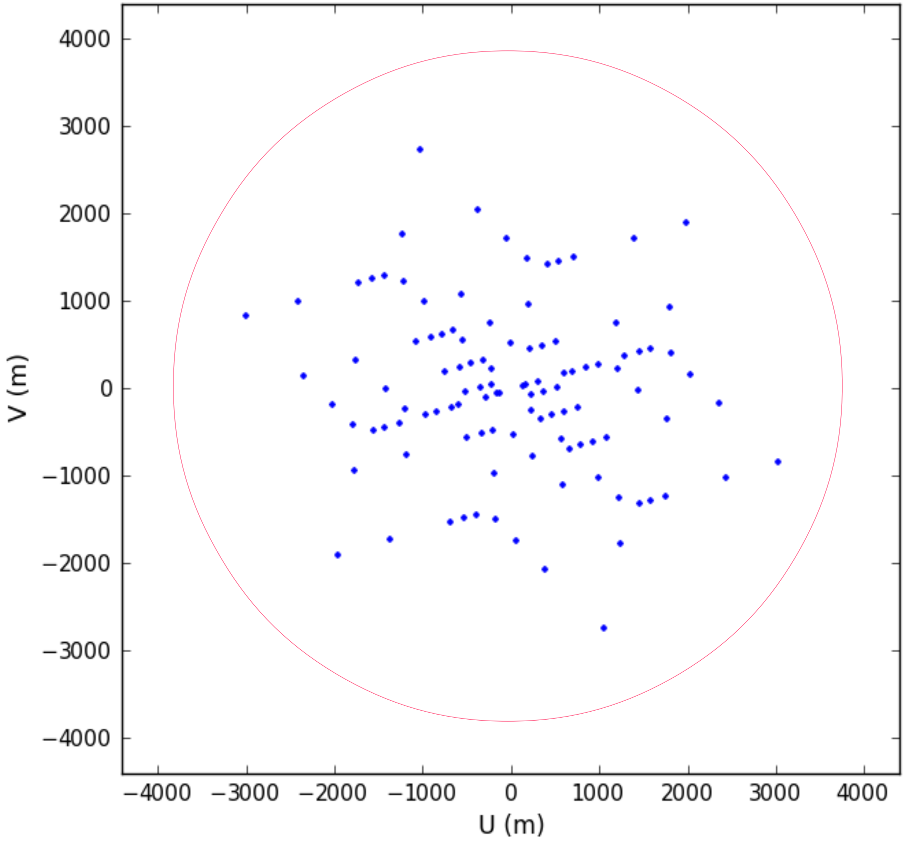
Observed Image

Aperture Synthesis Imaging

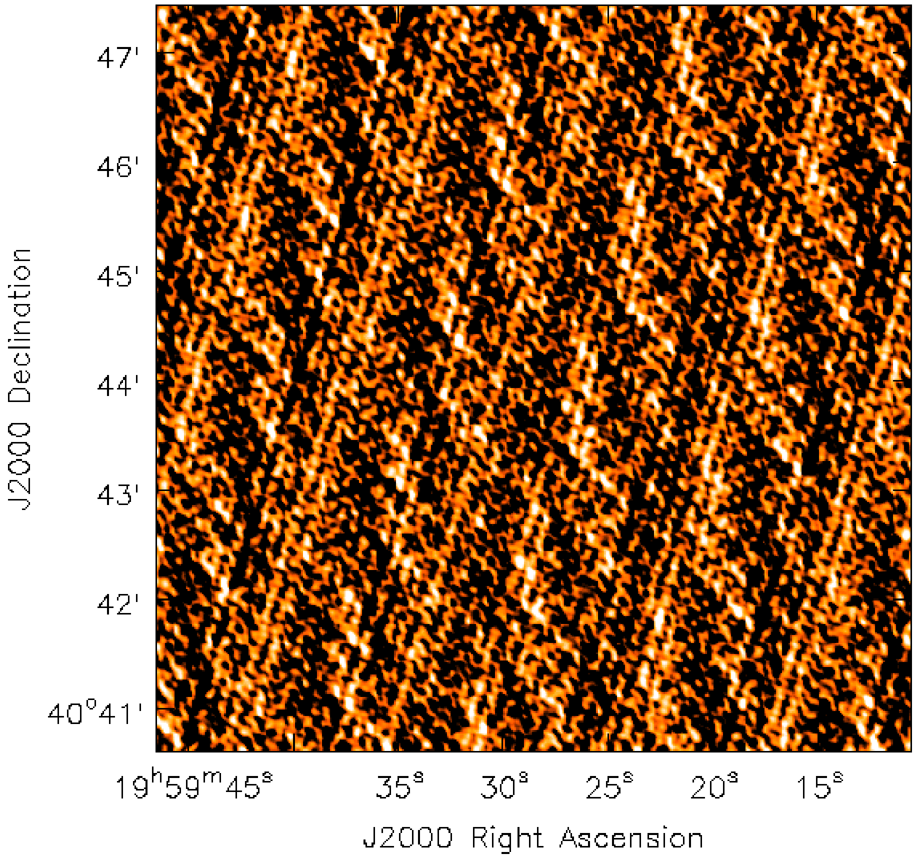
Image with 11 antennas => 55 fringes







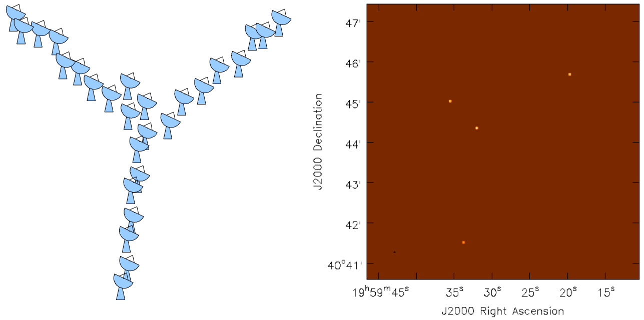
Synthesized Aperture

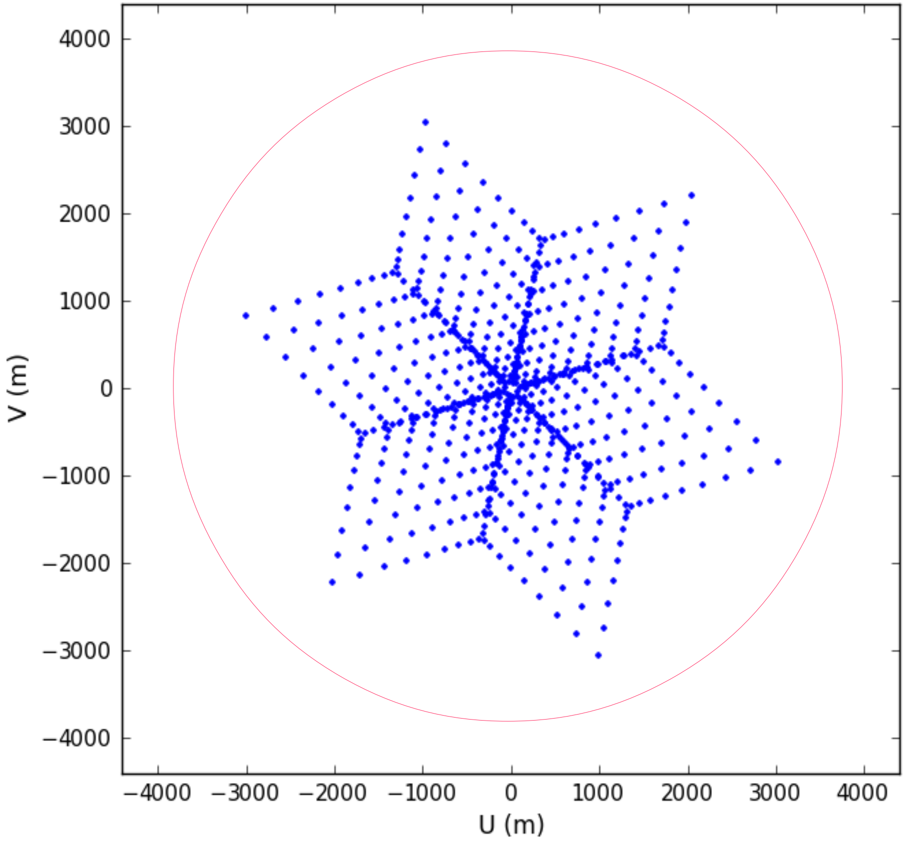


Observed Image

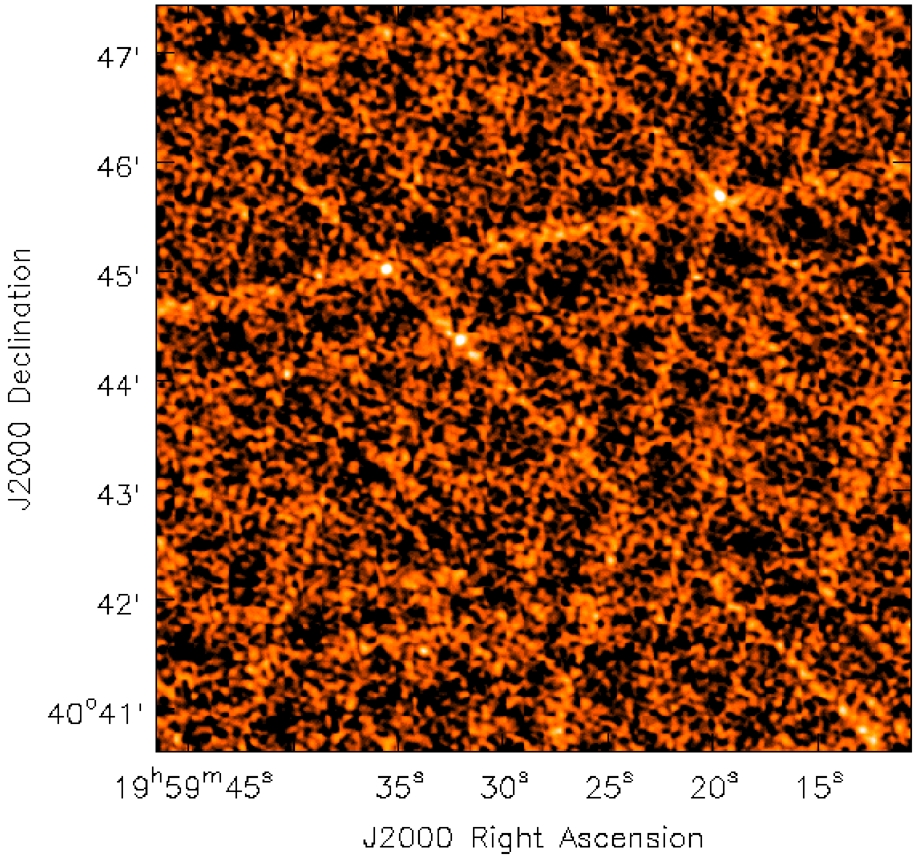
Aperture Synthesis Imaging

Image with 27 antennas => 351 fringes





Synthesized Aperture

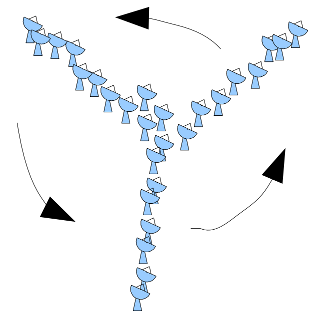


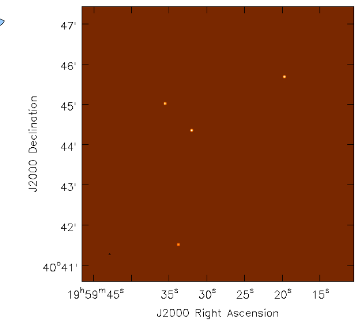
Observed Image

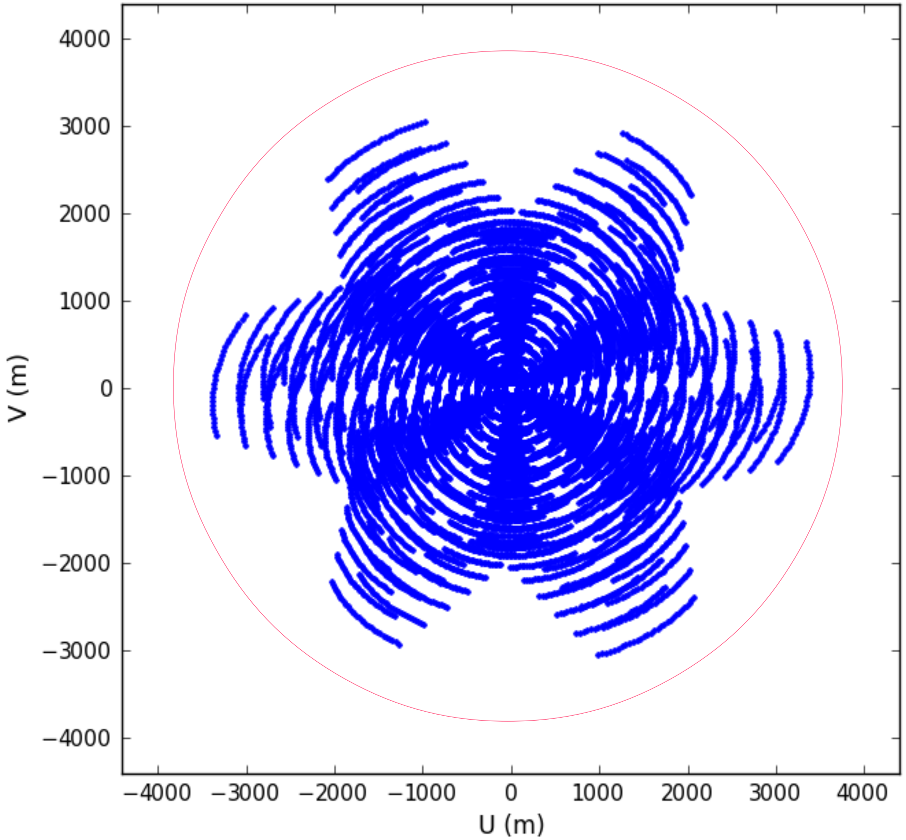
Aperture Synthesis Imaging

Image with 27 antennas over 2 hours

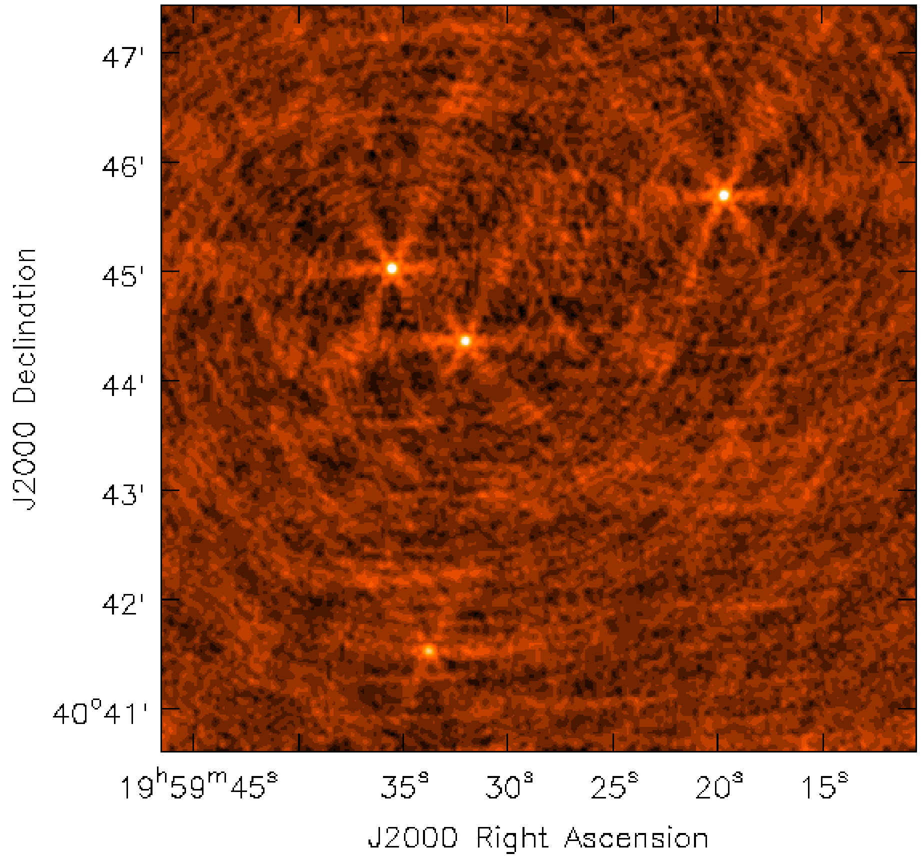
" Earth Rotation Synthesis "







Synthesized Aperture

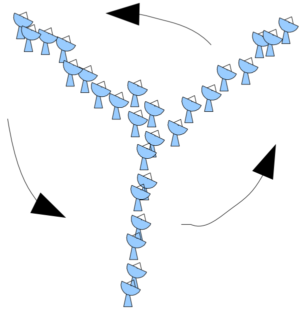


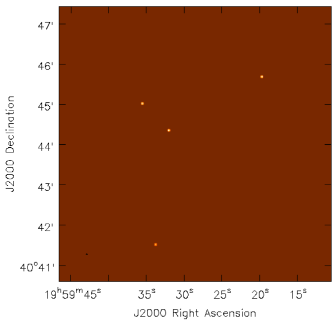
Observed Image

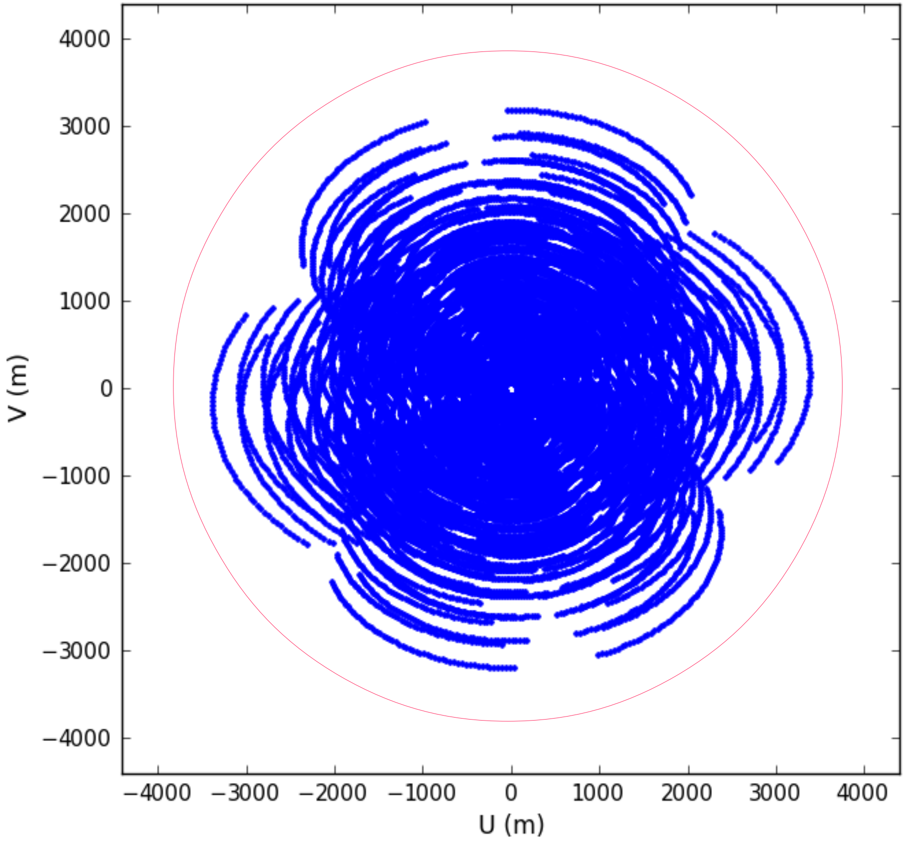
Aperture Synthesis Imaging

Image with 27 antennas over 4 hours

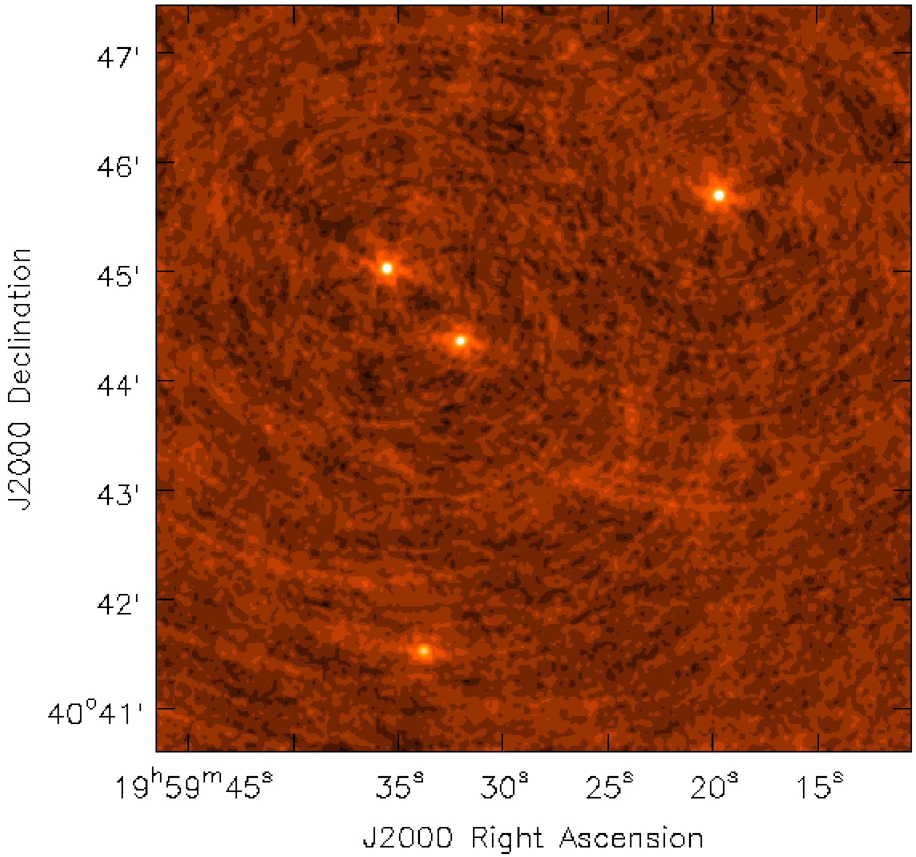
" Earth Rotation Synthesis "







Synthesized Aperture

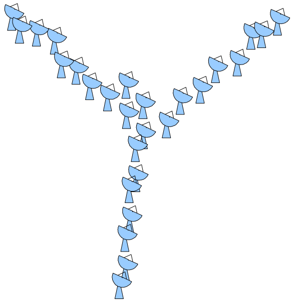


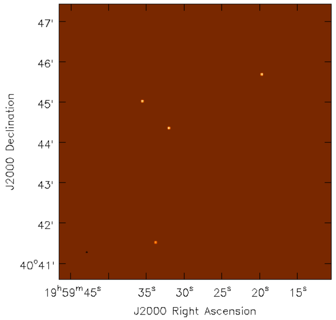
Observed Image

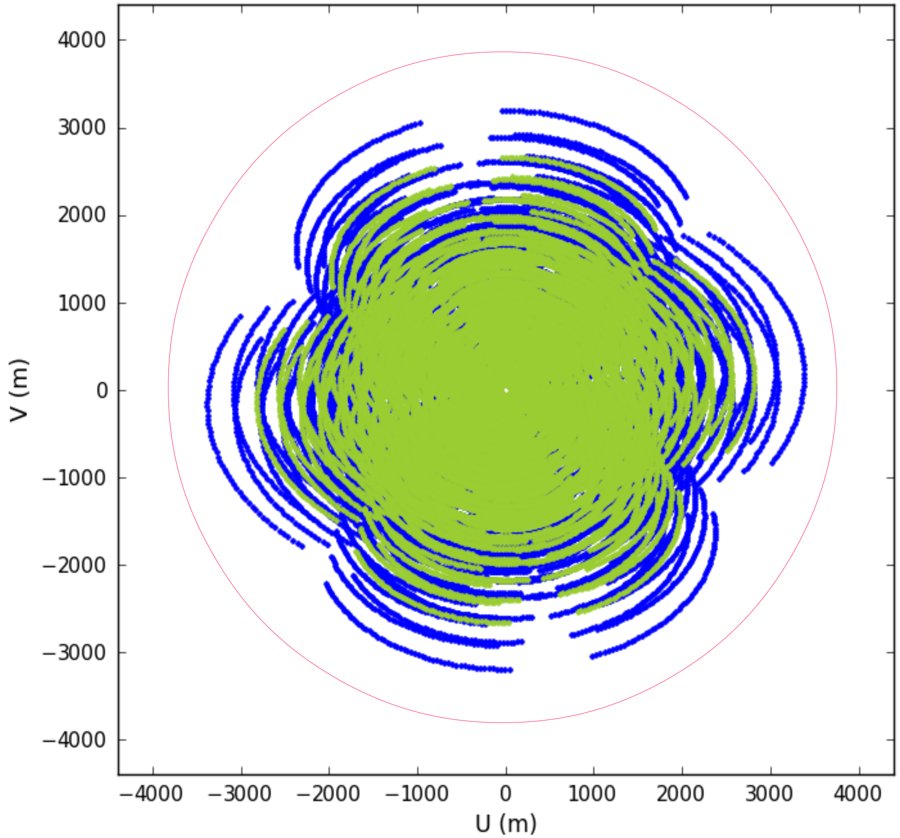
Aperture Synthesis Imaging

Image with 27 antennas over 4 hours
at 2 observing frequencies

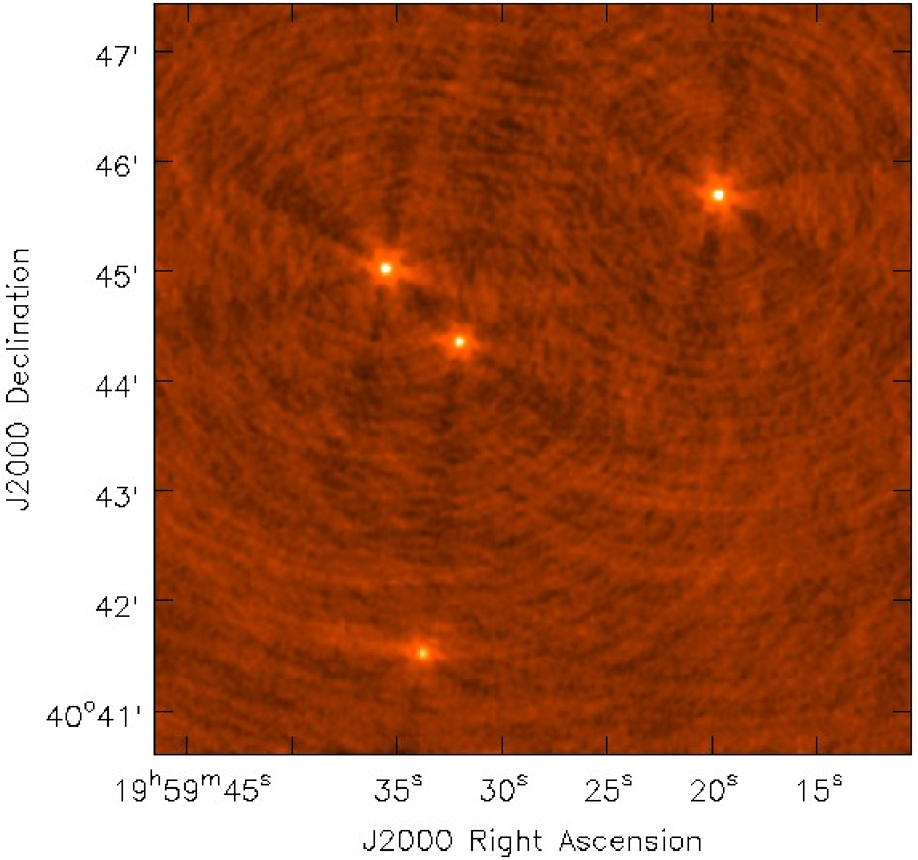
" Multi-frequency Synthesis "







Synthesized Aperture

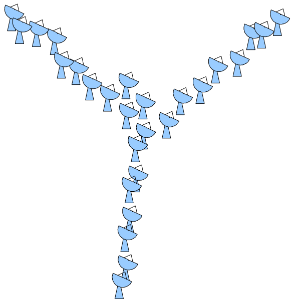


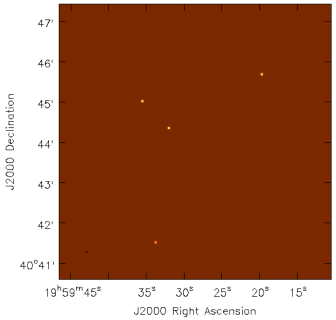
Observed Image

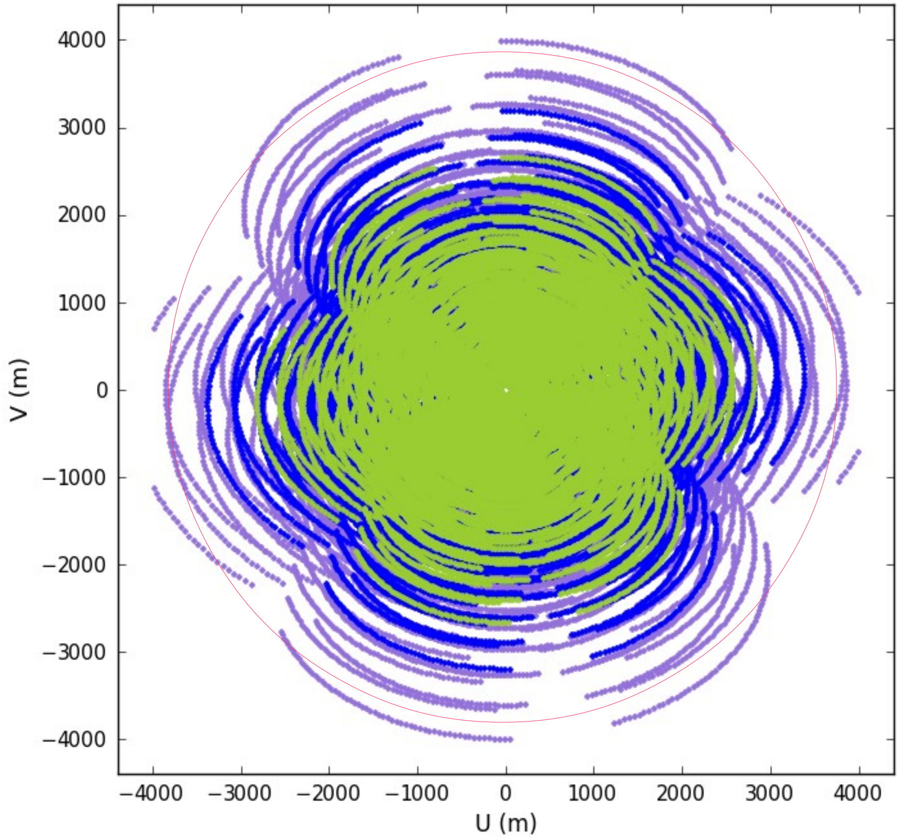
Aperture Synthesis Imaging

Image with 27 antennas over 4 hours
at 3 observing frequencies

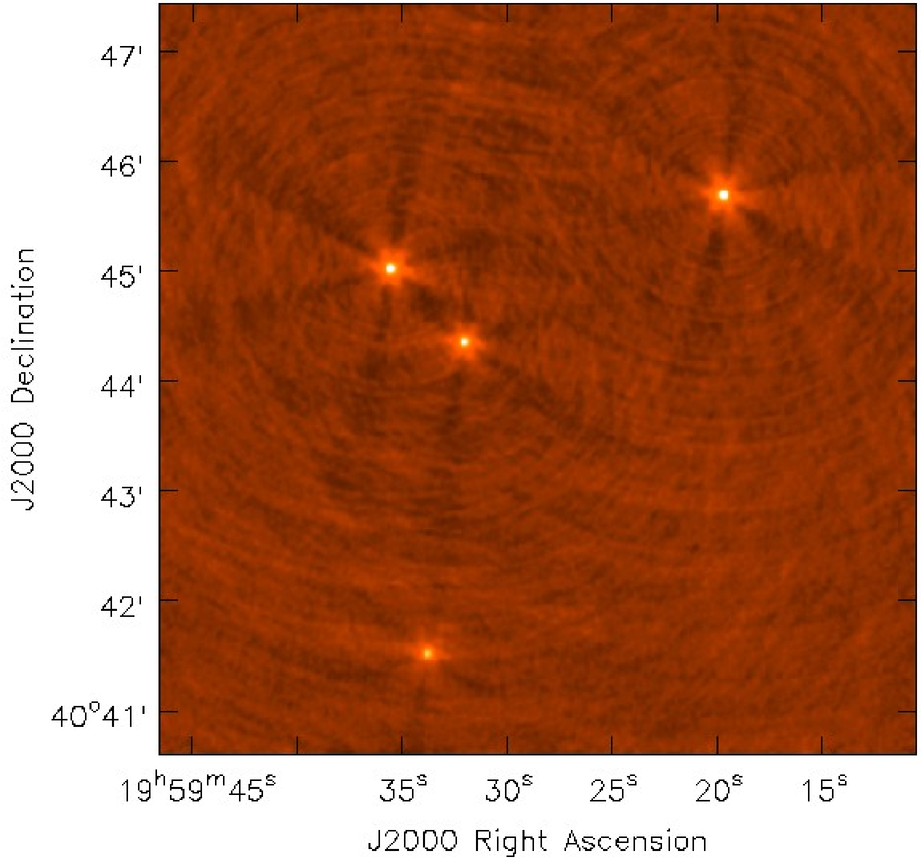
" Multi-frequency Synthesis "





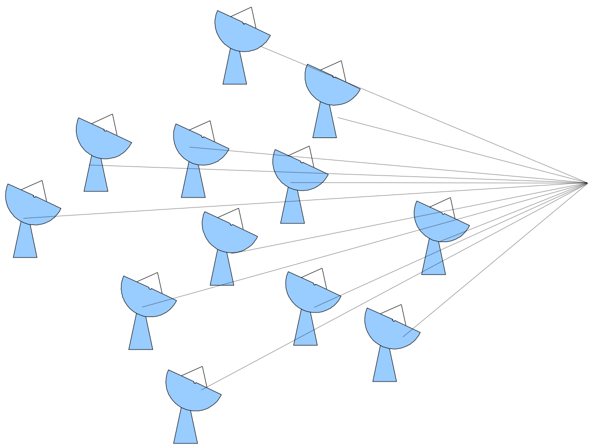


Synthesized Aperture



Observed Image

Data Acquisition and Analysis



Correlation (Real time system. FPGA/ASIC + backend cluster)

Time Series → Correlation → Spectral Channels → Integrate

Example Data rate : N(N-1)/2 * 1000 complex values per second

Data Archive (2.4 PB RAID storage)

Each observation is stored as a relational database

Example : VLA archive is 1.8 PB in size (+ 1 TB per day)

Post Processing – (1.6 PB Lustre FS, workstations, 90 node cluster, AWS)

Flagging → Calibration → Imaging

Flagging: Identify and mask corrupted data (RFI, Instrument errors, etc)

Calibration: Derive and apply corrections to undo the effects of complex valued antenna gains

Imaging: Reconstruct images by iterative model fitting while correcting for other instrumental effects

Thanks to J.Robnett for numbers

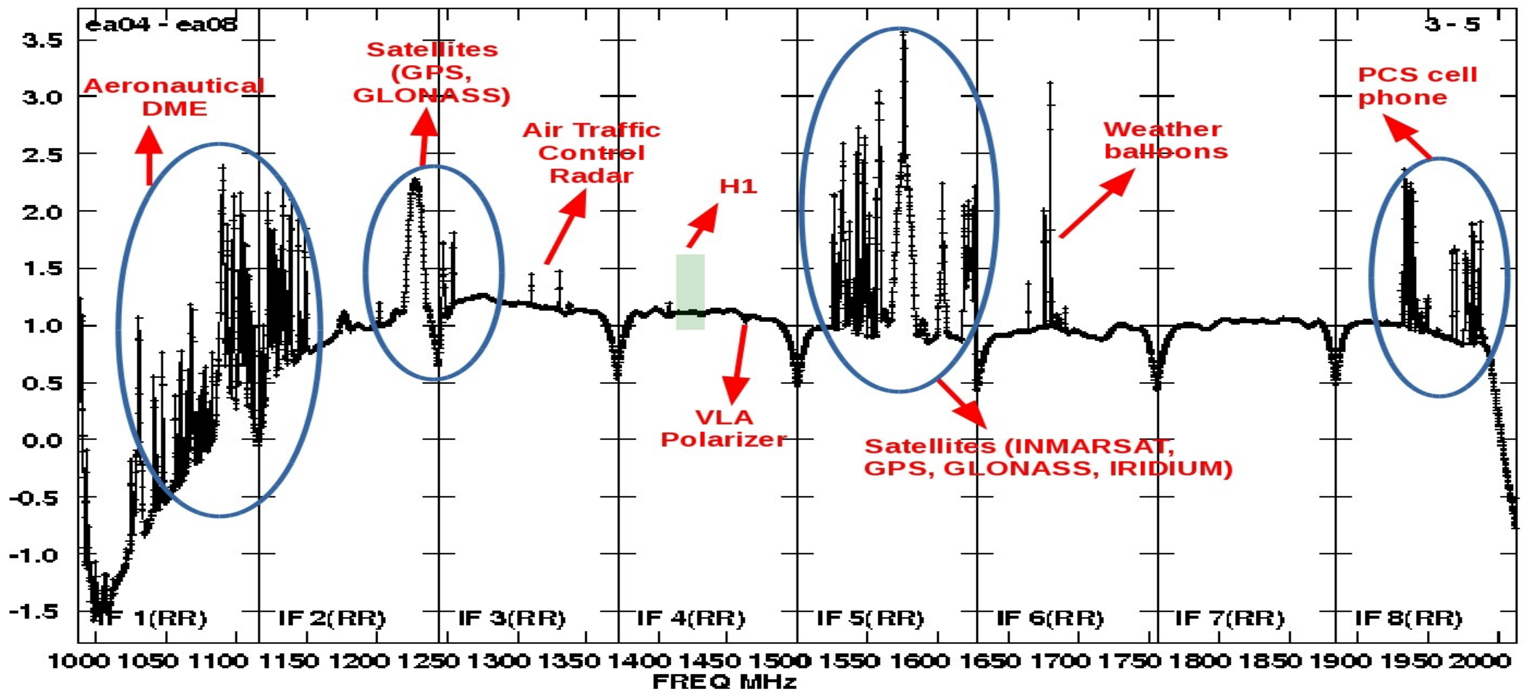
Flagging → Calibration → Imaging

Radio Frequency Interference

- Cellular phones, aircraf radar, satellite comms, military radar, car radars, etc...

Instrumental flags

- Antenna tracking delays, glitches in signal processing, antenna dropouts, shadowing...



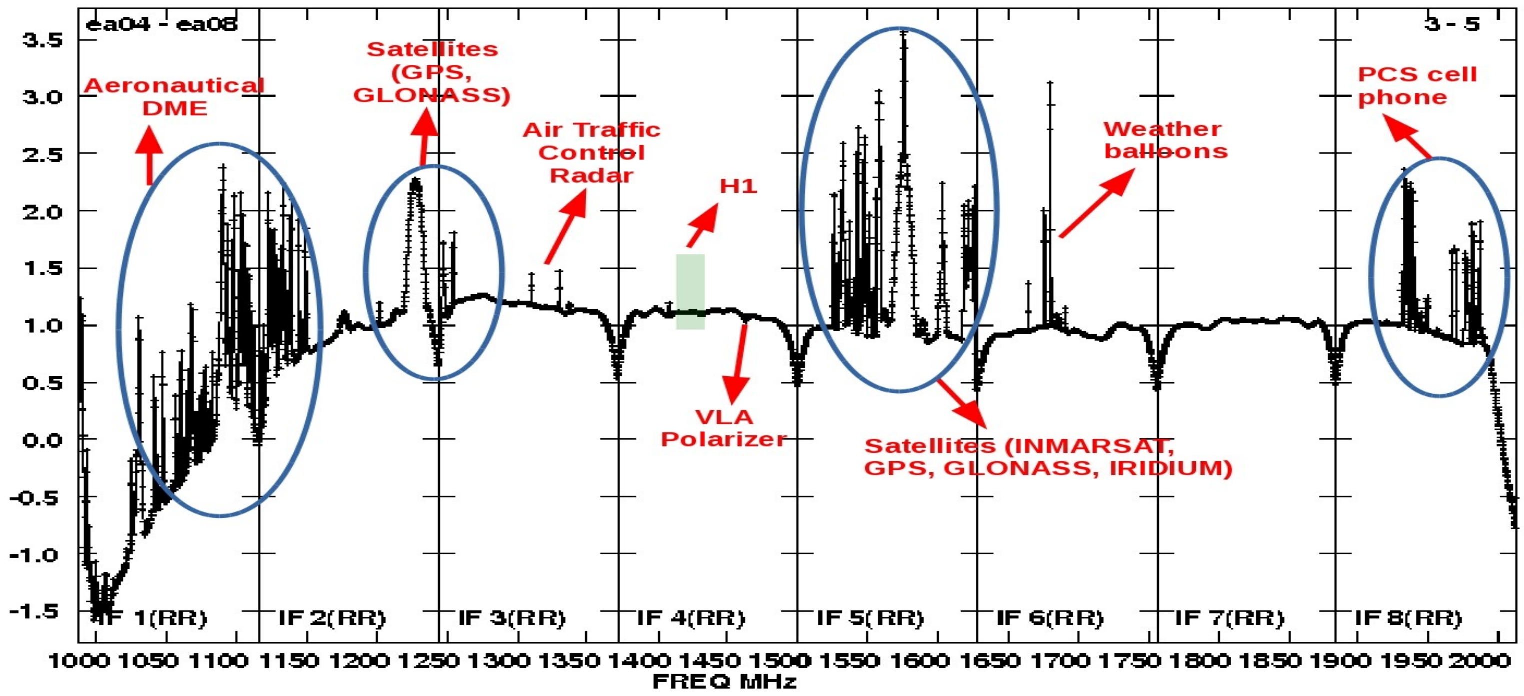
Flagging → Calibration → Imaging

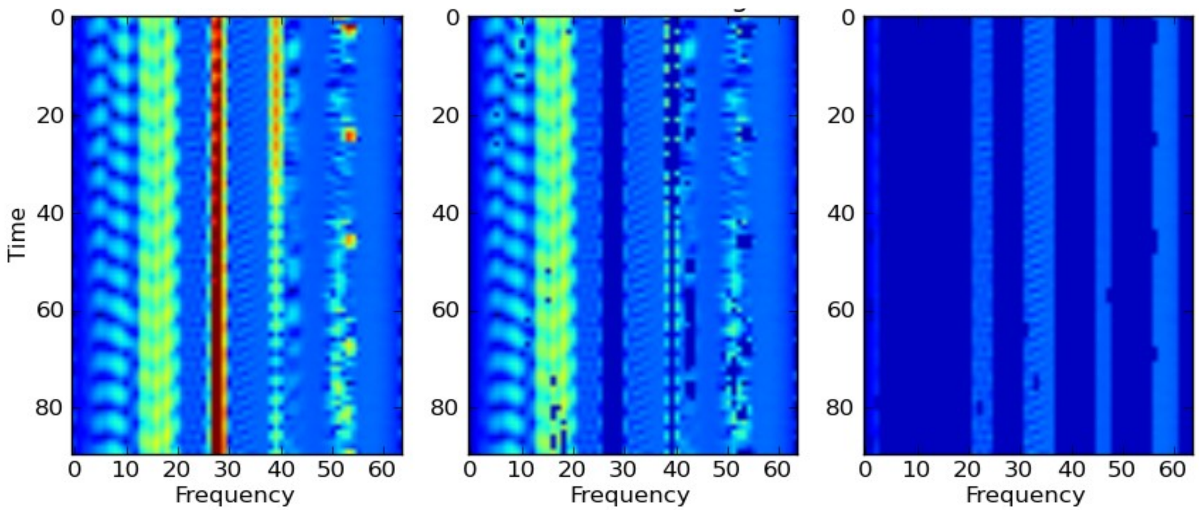
Radio Frequency Interference

- Cellular phones, aircraf radar, satellite comms, military radar, car radars, etc...

Instrumental flags

- Antenna tracking delays, glitches in signal processing, antenna dropouts, shadowing...





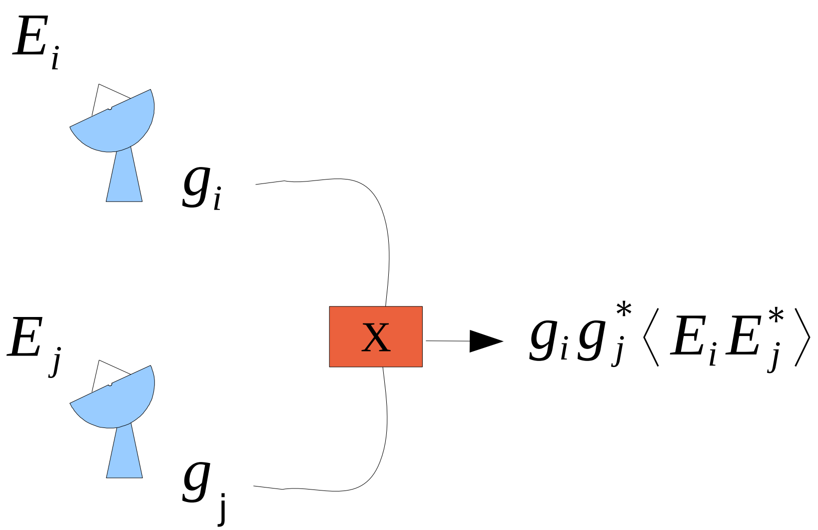
Automatic Flagging

- Model based and statistical outlier detectors

- Needs manual tuning
(Algorithm R&D ongoing)

- Parallelization via partitioning
(RFI type is classifiable)

The front-end electronics on each antenna introduces a multiplicative complex gain on the incoming signal. This must be removed.



(1) Observe a known source

$\langle E_i E_j^* \rangle$ is known

(2) Use data from all correlation pairs ij
Solve for complex gains g_i

(3) Apply corrections to target data :

$$\frac{g_i g_j^* \langle E_i E_j^* \rangle}{g_i g_j^*}$$

Calibration is usually a multi-stage process (different reasons, averaging, etc)

Gain solutions are a useful diagnostic of antenna-based instrumental errors
=> Detect outliers and apply flags to data.

Algorithms and their implementations can parallelize easily.

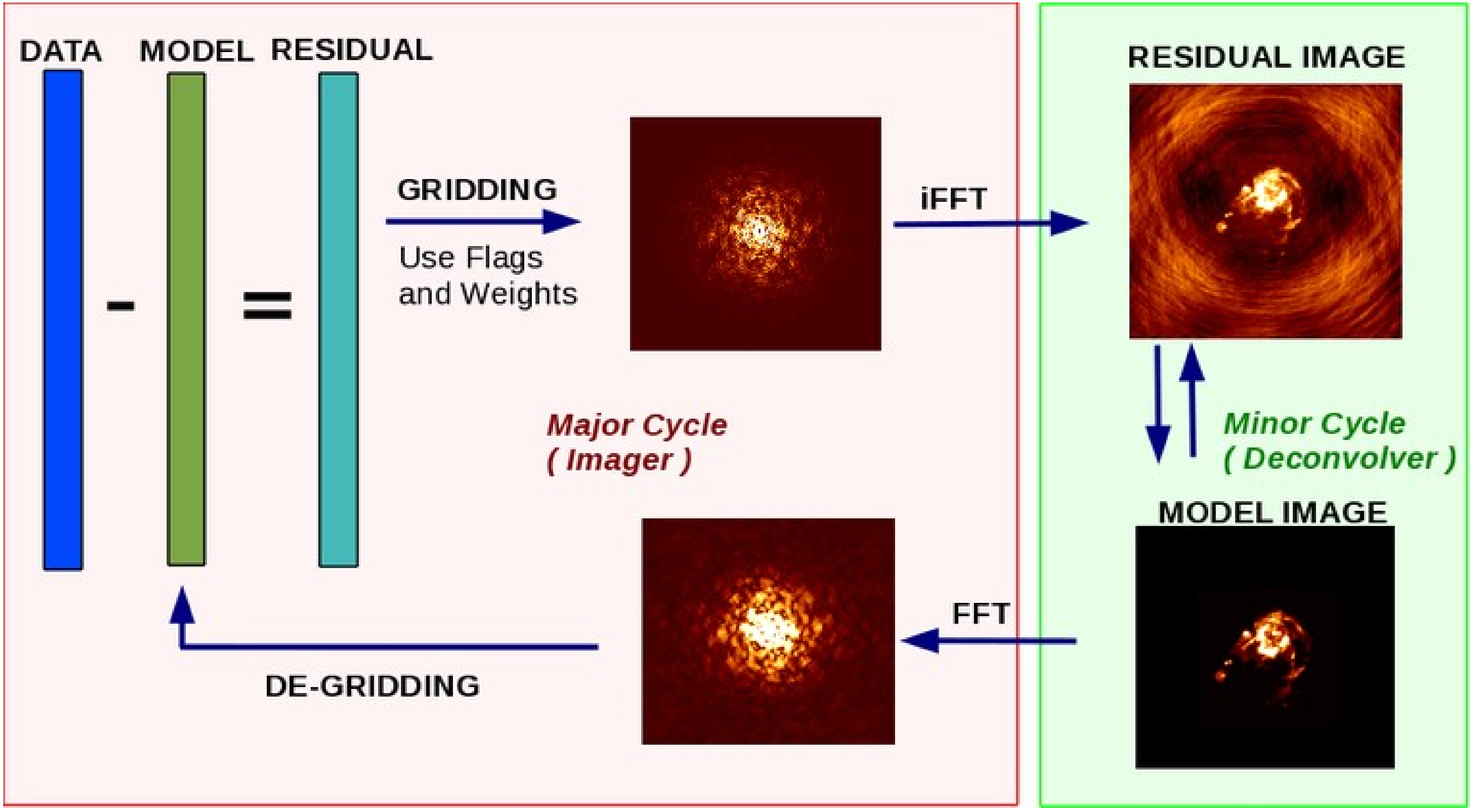
Iterative solvers : O(N_ant ^2)

Flagging → Calibration → **Imaging**

Image reconstruction is an iterative model-fitting / optimization problem

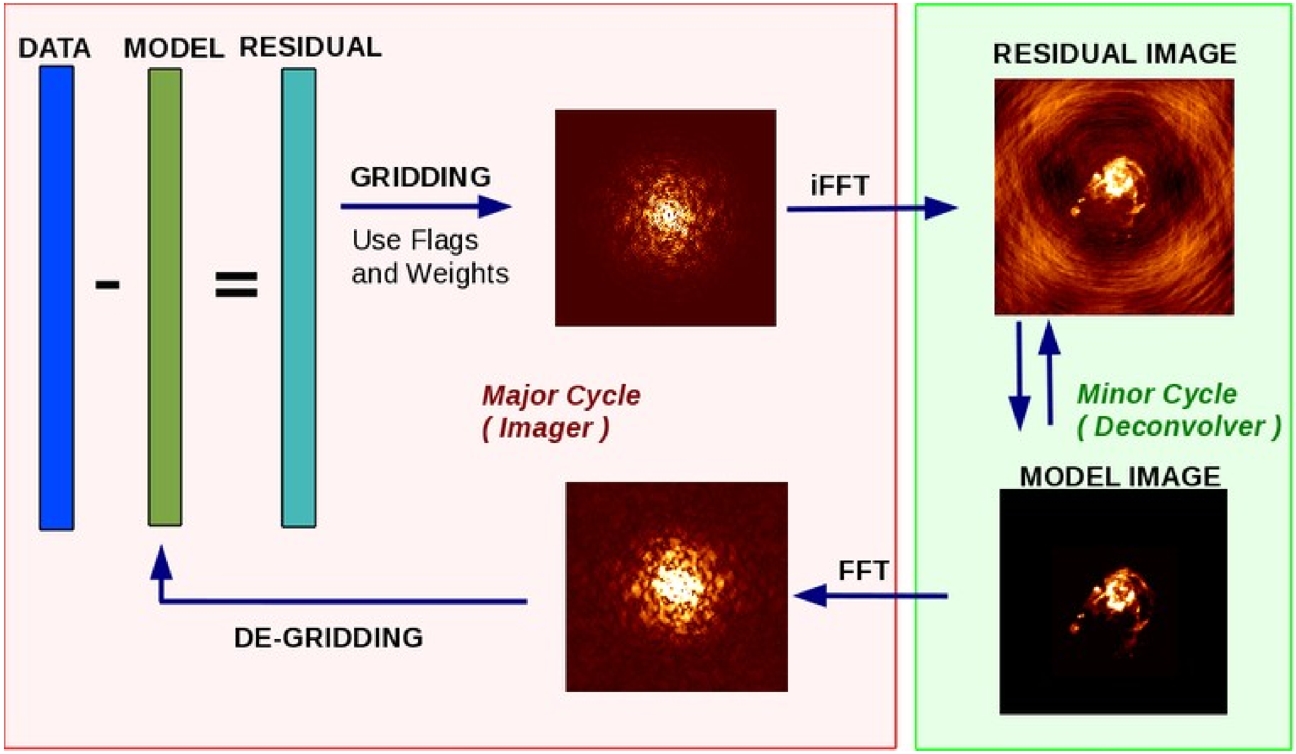
Measurement Eqn : $[A] I^m = V^{obs}$

Iterative solution : $I^m_{i+1} = I^m_i + g [A^T W A]^+ \left(A^T W \left(V^{obs} - A I^m_i \right) \right)$



High Performance Computing

Computing, Data Volume, Image Sizes, Memory Use,..



Flagging → Calibration → Imaging

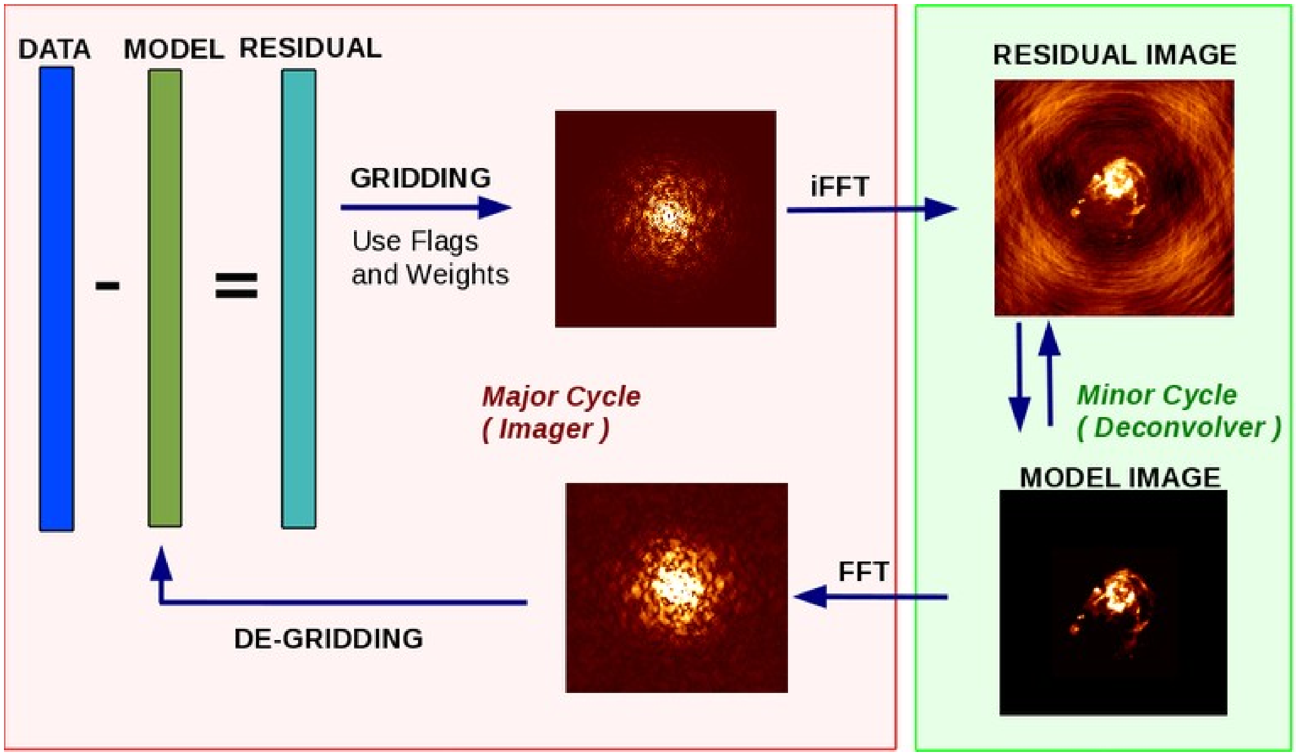
Data volume

N_data =
 N_ant^2 x
 N_chan x
 N_pol x
 N_time

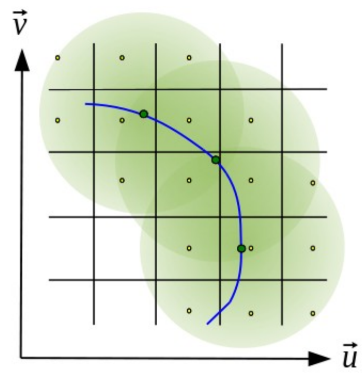
Complex numbers

Lustre I/O

Example :
 8hr data
 300 GB



Flagging → Calibration → **Imaging**



Gridding : Convolutional resampling

O(N_data) x (nxn) complex multiply/add (n=5 - 100)
=> Compute load : O(N_data) * 10^{2-5} flops

Data parallelization, Multi-threading, GPUs, etc...

Example : Major cycle : 1hr → 10 days (Diff Algorithms)

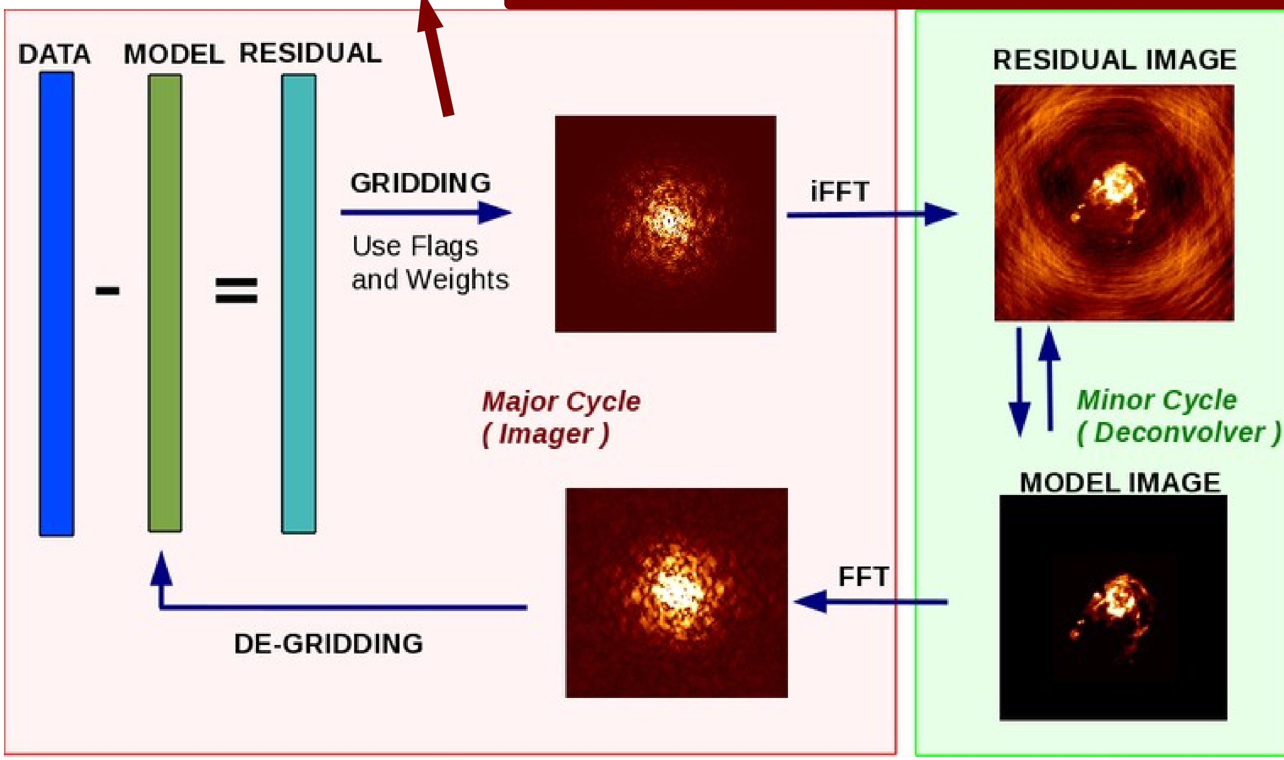
Data volume

N_data =
N_ant^2 x
N_chan x
N_pol x
N_time

Complex numbers

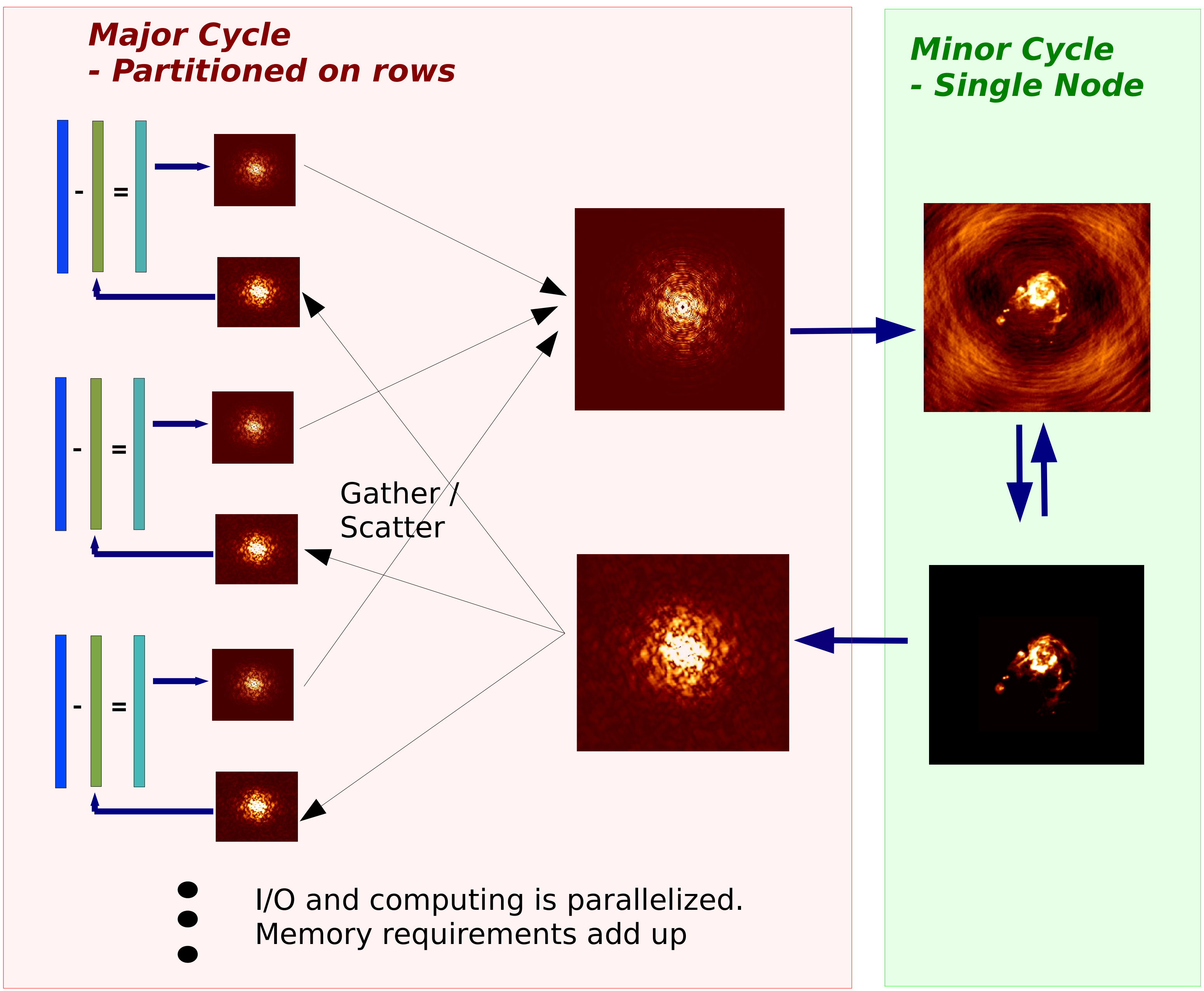
Lustre I/O

Example :
8hr data
300 GB



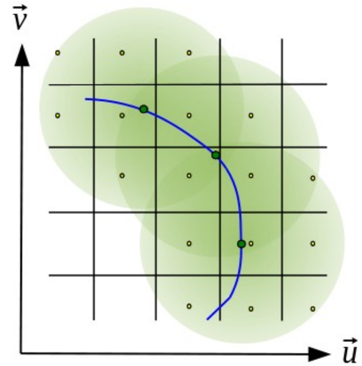
Flagging → Calibration → **Imaging**

Data partitioning to parallelize compute/IO intensive major cycles



Images are typically 2D, and pixel partitioning is not needed.

Flagging → Calibration → **Imaging**



Gridding : Convolutional resampling

O(N_data) x (nxn) complex multiply/add (n=5 - 100)
=> Compute load : O(N_data) * 10^{2-5} flops

Data parallelization, Multi-threading, GPUs, etc...

Example : Major cycle : 1hr → 10 days (Diff Algorithms)

Data volume

N_data =
N_ant^2 x
N_chan x
N_pol x
N_time

Complex numbers

Lustre I/O

Example :
8hr data
300 GB

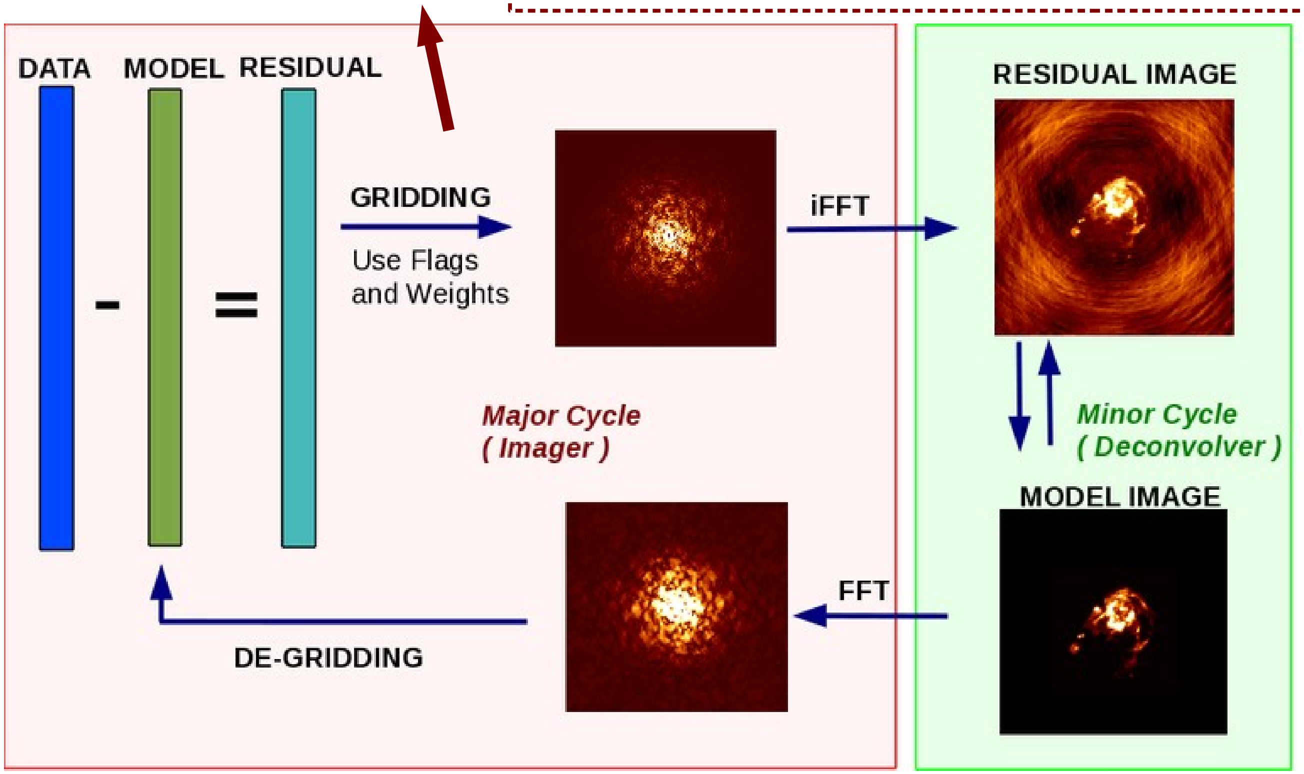


Image sizes

N_pix = Nx x Ny x
N_chan x N_pol

Real / Complex

FFTs : O(NlogN)
Pixel math: O(N^2)
Mem : ~8 copies

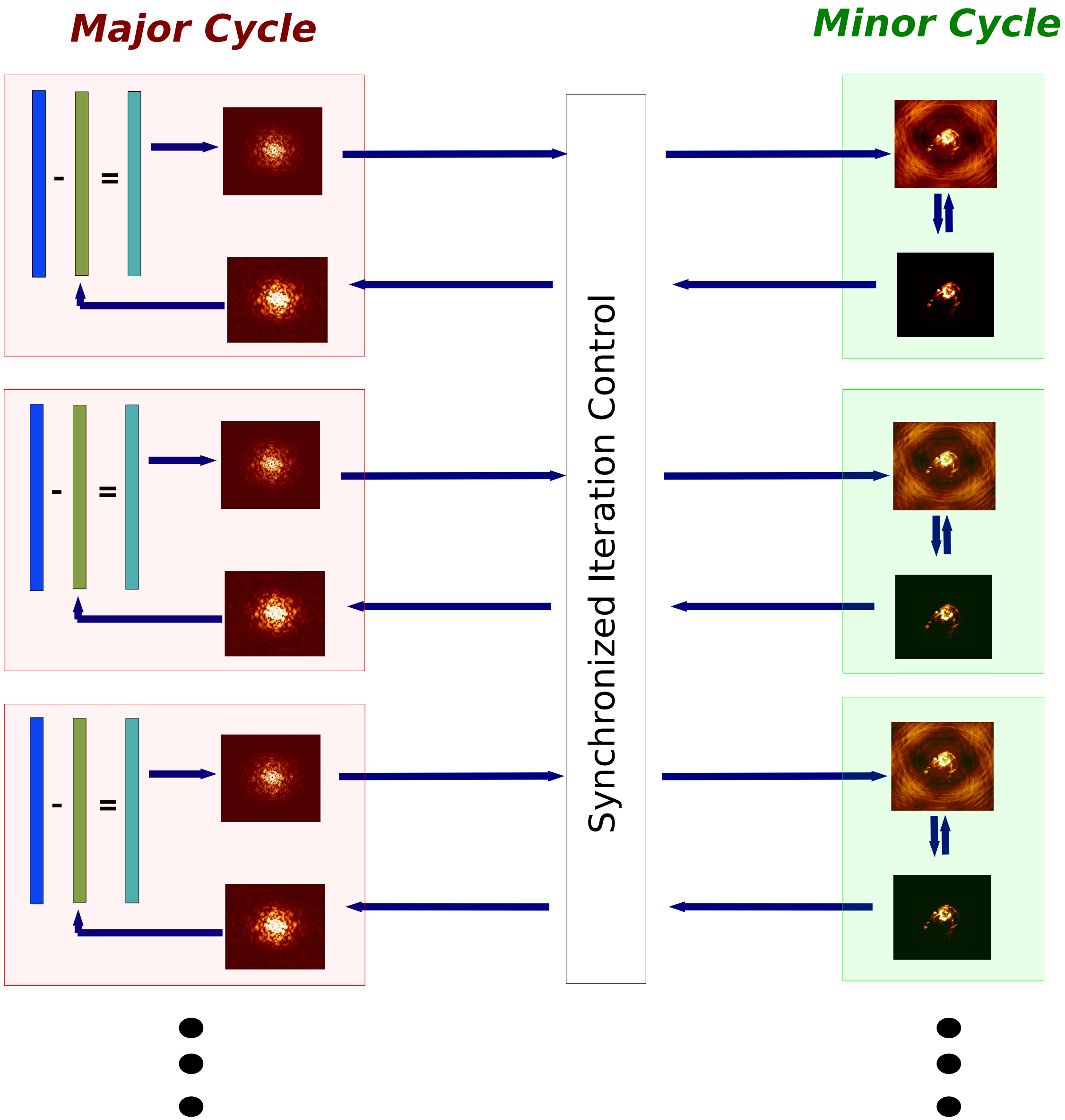
Multi-threading
Chan parallelization

Nx : 1k → 40k
N_chan : 200 - 16K

Example :
2K x 2K x 1000
~15 GB per image

Flagging → Calibration → **Imaging**

Data and Image partitioning to parallelize Spectral Cube imaging



Images are typically 3D, and partitioning along the frequency axis is required.

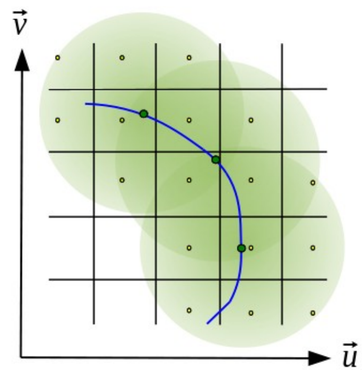
Image pixels :

1kx1k to 80kx80k

Spectral channels (per pixel) :

1k to 16K

Flagging → Calibration → **Imaging**



Gridding : Convolutional resampling

O(N_data) x (nxn) complex multiply/add (n=5 - 100)
=> Compute load : O(N_data) * 10^{2-5} flops

Data parallelization, Multi-threading, GPUs, etc...

Example : Major cycle : 1hr → 10 days (Diff Algorithms)

Data volume

N_data =
N_ant^2 x
N_chan x
N_pol x
N_time

Complex numbers

Lustre I/O

Example :
8hr data
300 GB

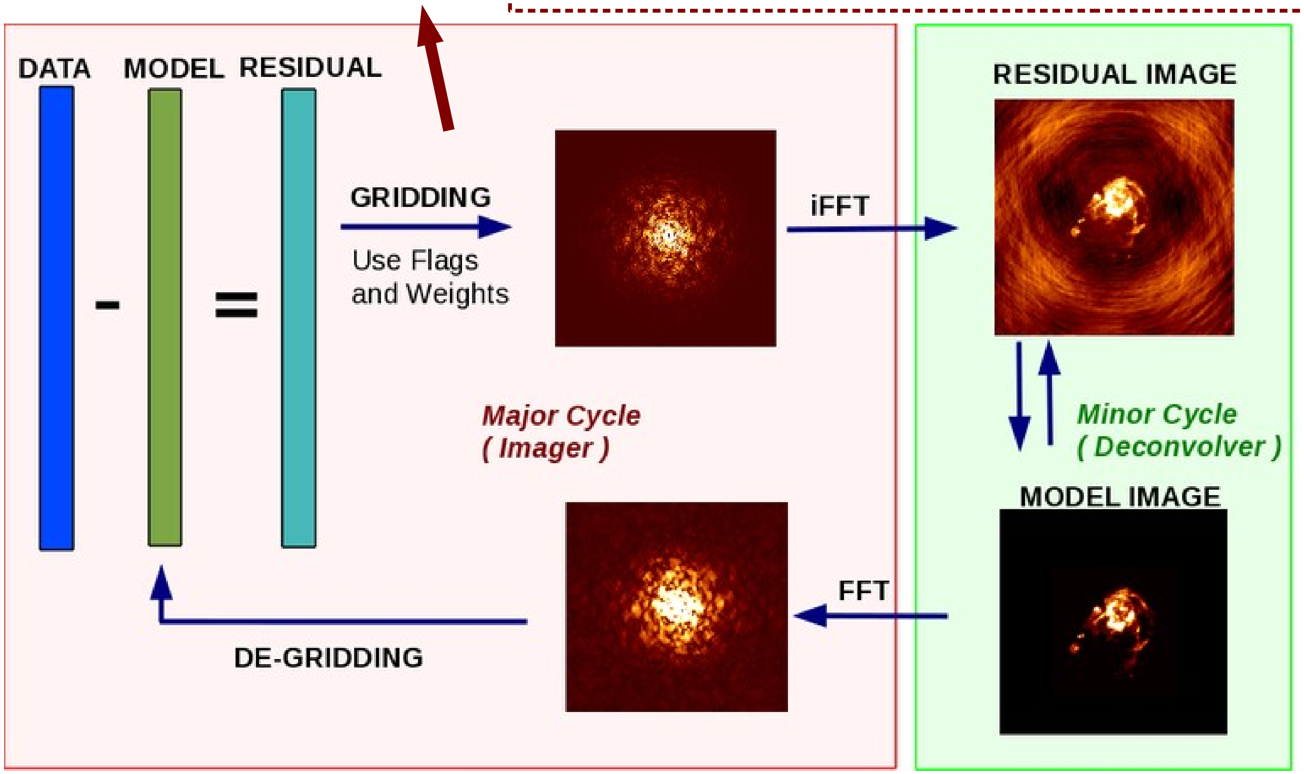


Image sizes

N_pix = Nx x Ny x N_chan x N_pol

Real / Complex

FFTs : O(NlogN)
Pixel math: O(N^2)
Mem : ~8 copies

Multi-threading
Chan parallelization

Nx : 1k → 40k
N_chan : 200 - 16K

Example :
1K x 1K x 256
~1 GB per image

Number of iterations : 5 – 10 major cycle loops
10^2 to 10^4 minor cycle steps

Runtime varies by 1-2 orders of magnitude. Depends on data.

Flagging → Calibration → **Imaging**

Observed image = Instrumental Point-Spread-Function convolved with the true sky

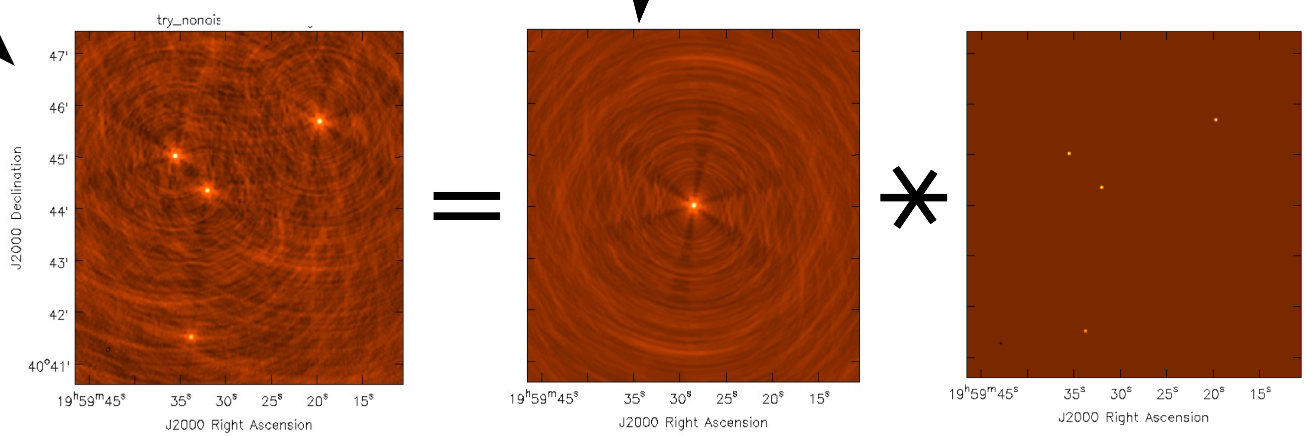
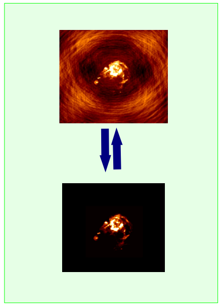


Image reconstruction is typically a "deconvolution" process.



Algorithms : Parameterized models + Iterative model fitting.
Feature extraction + classification, Mixed models

– **Basis functions** are : Delta functions, Gaussians, Wavelets, Shapelets, Polynomials to represent spectral structure or time-variability, 2D,3D,4D models

– **Metrics** being optimized : L2 or L1 or TV norms, weighted combination of norms and a-priori bias terms, etc..

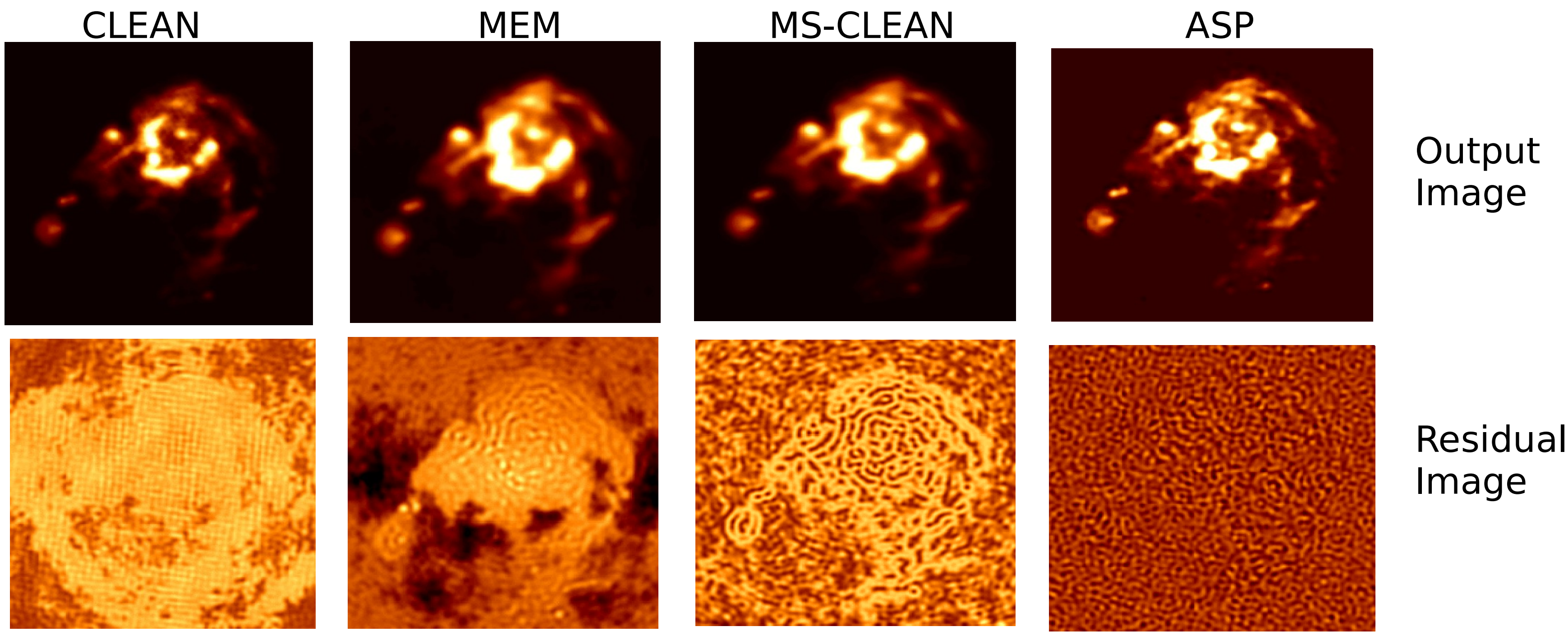
– **Optimization schemes** : Greedy algorithms + gradient descent, etc..

Flagging → Calibration → **Imaging**

No unique solution (theoretically) => Differences between algorithm outputs

=> Algorithm choice depends on sky structure, data quality, target science

=> Different algorithms and parameters (e.g. convergence criteria) could result in orders of magnitude differences in computing load.



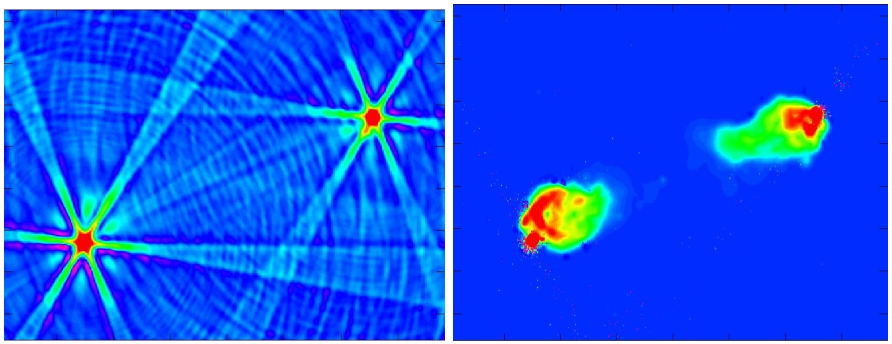
Metrics for Image Quality : Noise RMS, Peak residual, Dynamic Range, etc...

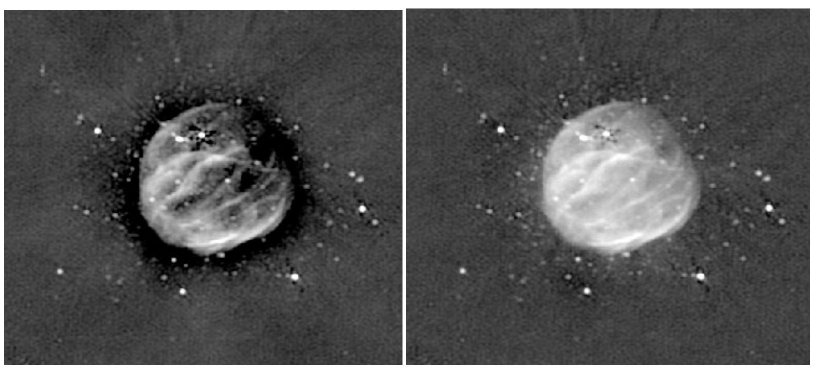
Recognizing characteristic error patterns, Knowing when to stop trying.

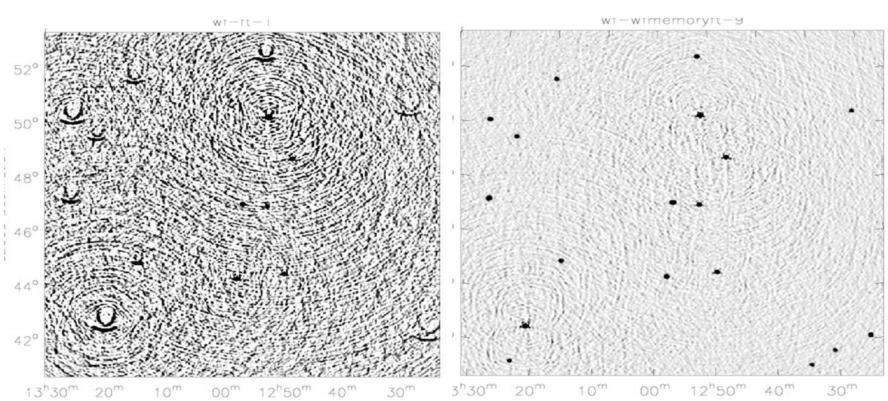
Flagging → Calibration → **Imaging**

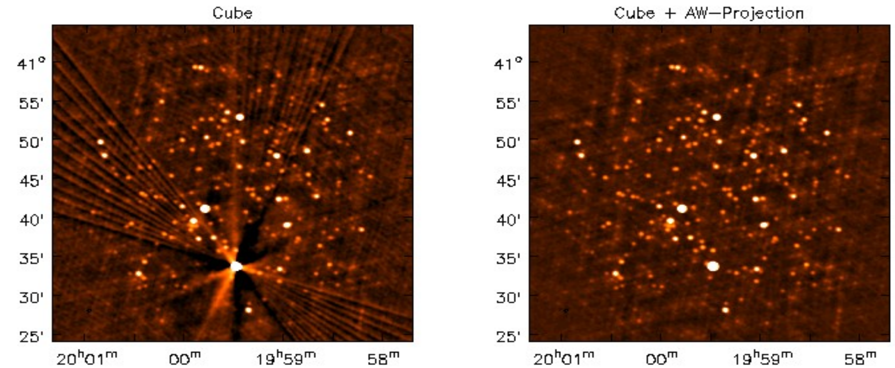
Recognizing imaging artifacts

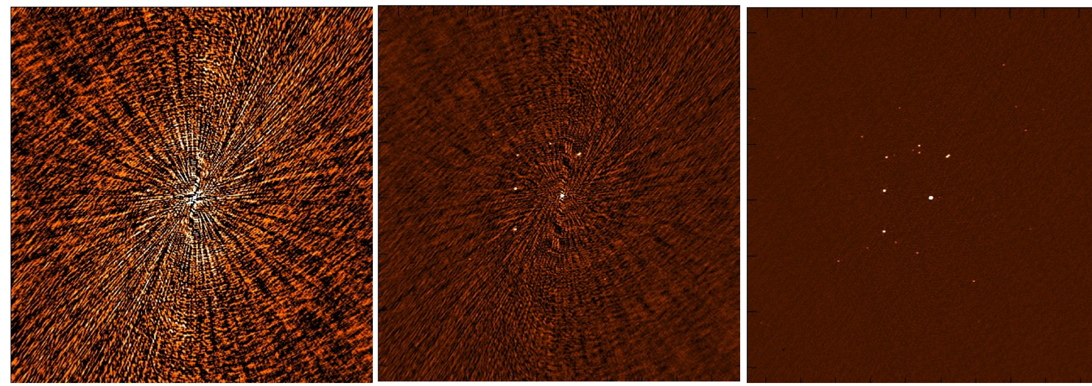
=> Choose appropriate next step
=> Recognise when to stop
=> Quantify remaining errors

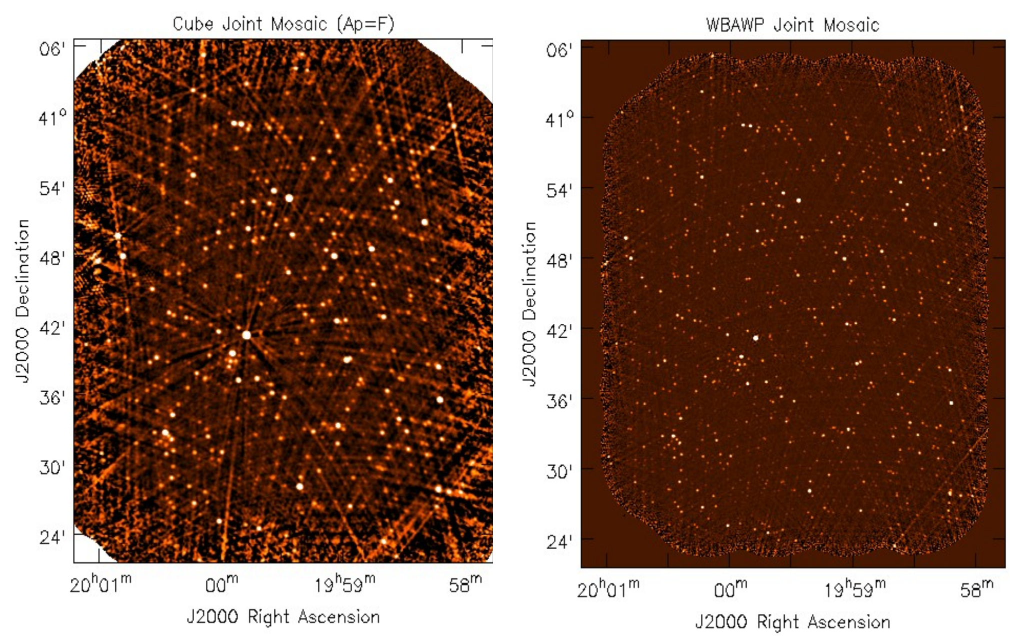




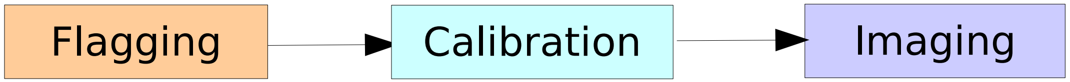




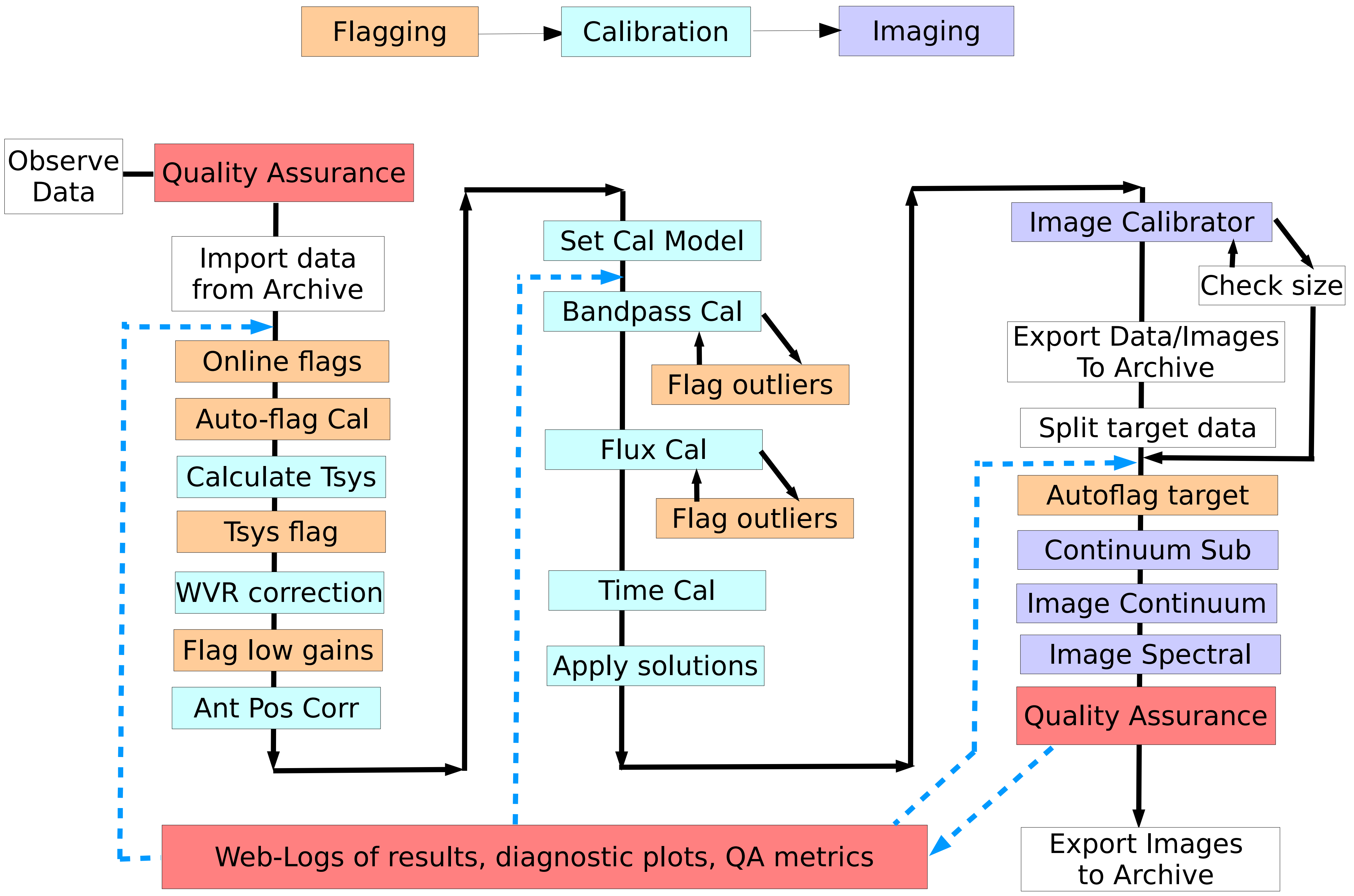




Science Ready Data Products – Automated Analysis Pipelines

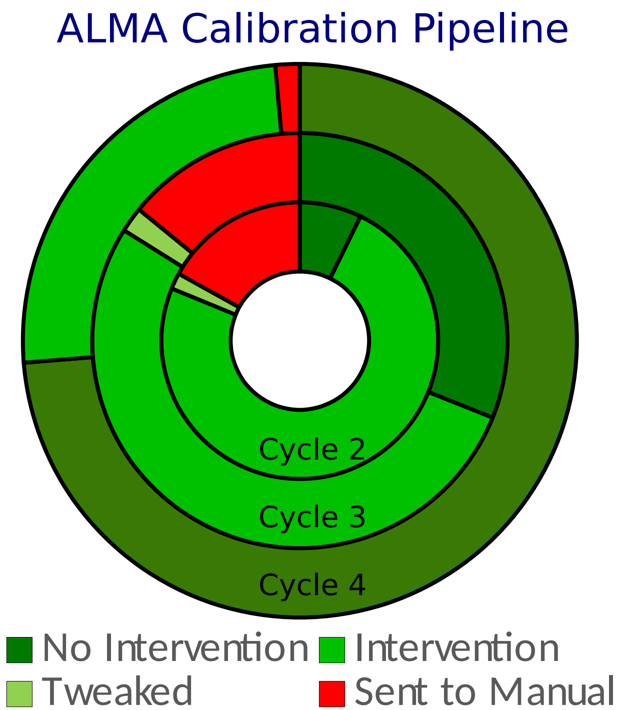


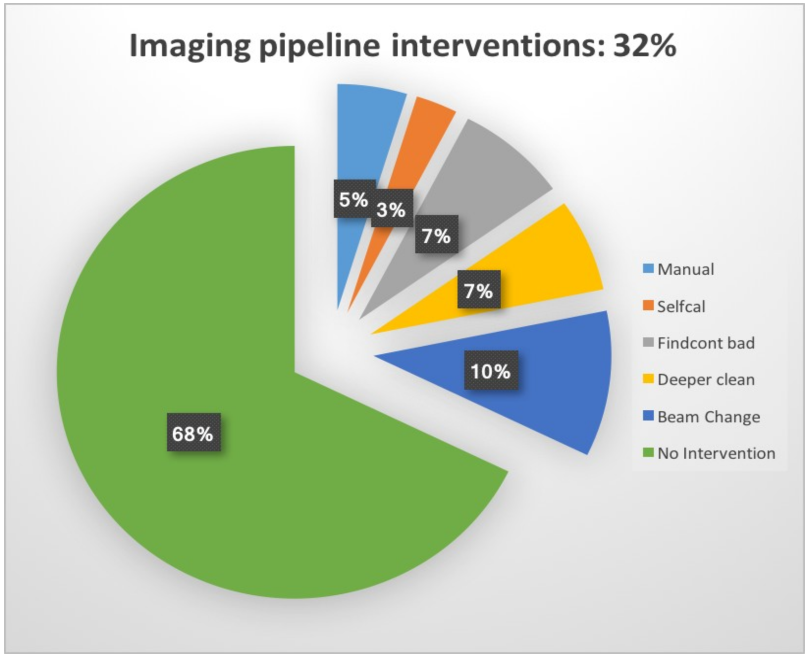
Science Ready Data Products – Automated Analysis Pipelines



Science Ready Data Products – Automated Analysis Pipelines

Our current pipeline steps are the result of hand-optimized manual tuning by a team of scientists, validated on ~100 datasets, for a few 'standard' usage modes.





Graphs from J.Kern

VLA Pipeline (Calibration only). VLA Sky Survey Pipeline (Calibration + Imaging)

=> Current practice works, but we would like to reduce manpower required both for heuristic development as well as Quality Assessment.

=> Limited plans to support complicated or experimental modes.

Areas where Machine Learning and AI may be useful

(1) Automating the data analysis decision tree :

It is possible to choose a sequence of steps and detailed parameter tunings that provides the best flagging, calibration and imaging outcome for any given dataset. This may differ between types of datasets and science goals.

(2) Error recognition :

Humans are adept at identifying RFI patterns in plots of recorded data, non-standard antenna behaviour from calibration solution plots, and artifacts and other tell-tale shapes in images.

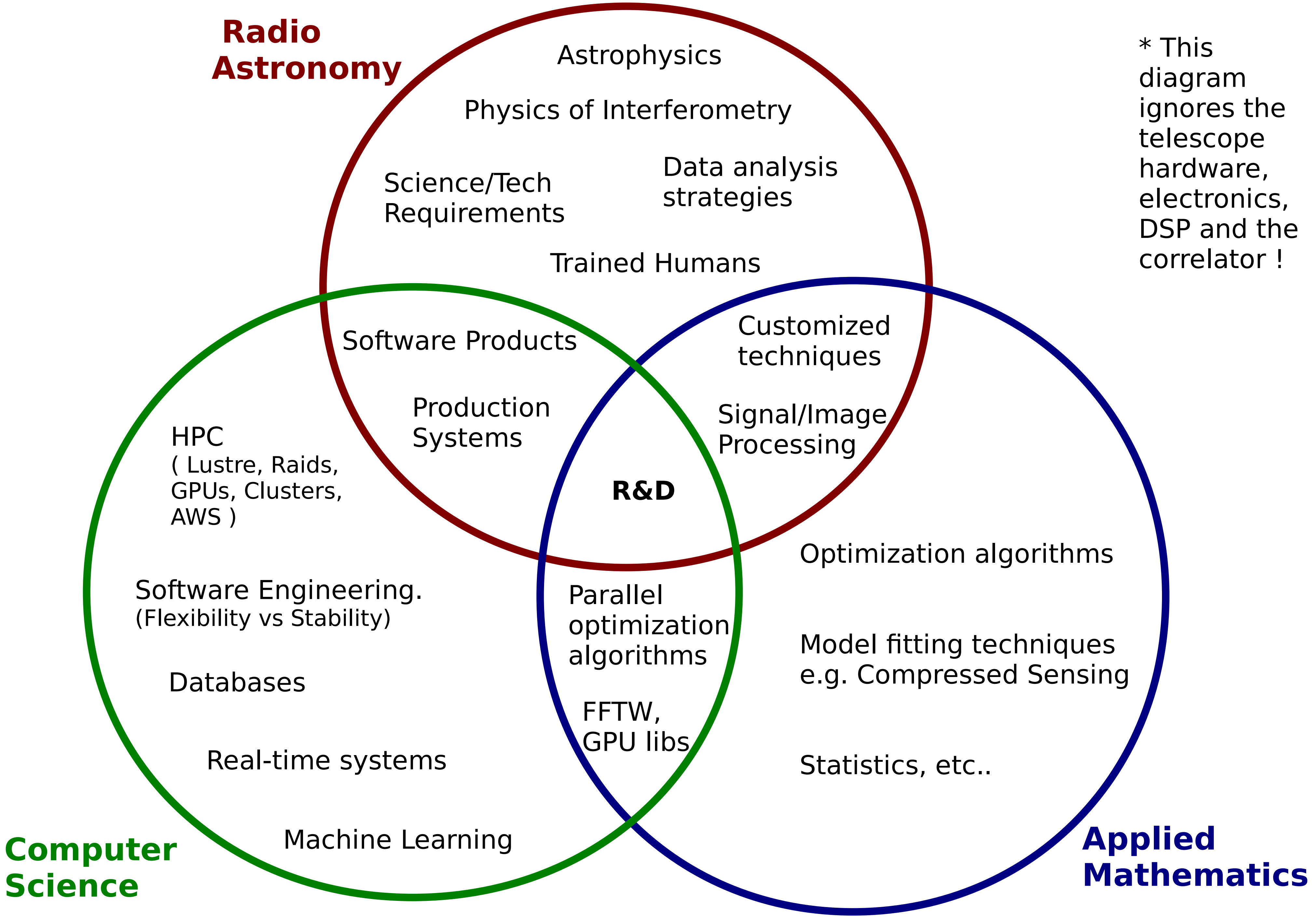
(3) Telescope monitoring and control :

Using telemetry and monitoring data to classify problems and their symptoms and perhaps predict failures. Use information about RFI sources, weather, to optimize the observation schedule and setup.

(4) Image and Spectrum analysis :

Feature detection/description and classification for surveys and catalogues
Spectral profile matching (mixed models)
Quality assessment : Have we gotten the best we can out of the data ?

This is an interdisciplinary field



* This diagram ignores the telescope hardware, electronics, DSP and the correlator !