ALMA archive reimaging study

Andrea Giannetti

Italian ARC

1 December 2017

The ALMA interferometer

Largest and most sensitive radio telescope in the world

Built in the Chajnantor Plateau in the Atacama Desert

Composed of 50 12m antennas + 12 7m + 4 12m TP

The ALMA interferometer

Largest and most sensitive radio telescope in the world

Built in the Chajnantor Plateau in the Atacama Desert

Composed of 50 12m antennas + 12 7m + 4 12m TP

The ALMA interferometer

Largest and most sensitive radio telescope in the world

Built in the Chajnantor Plateau in the Atacama Desert

Composed of 50 12m antennas + 12 7m + 4 12m TP

Products in the ALMA archive

ALMA delivers quality controlled science-grade data to PIs and archival researchers

During the Early Science manual data processing

Pipeline has taken more and more over that process over time

All data are imaged by pipeline

Only 10% of the raw data converted to science products in early science and included in the archive

Products in the ALMA archive

ALMA delivers quality controlled science-grade data to PIs and archival researchers

During the Early Science manual data processing

Pipeline has taken more and more over that process over time

All data are imaged by pipeline

Only 10% of the raw data converted to science products in early science and included in the archive

Products in the ALMA archive

ALMA delivers quality controlled science-grade data to PIs and archival researchers

During the Early Science manual data processing

Pipeline has taken more and more over that process over time

All data are imaged by pipeline

Only 10% of the raw data converted to science products in early science and included in the archive

Products in the ALMA archive

ALMA delivers quality controlled science-grade data to PIs and archival researchers

During the Early Science manual data processing

Pipeline has taken more and more over that process over time

All data are imaged by pipeline

Only 10% of the raw data converted to science products in early science and included in the archive

Products in the ALMA archive

ALMA delivers quality controlled science-grade data to PIs and archival researchers

During the Early Science manual data processing

Pipeline has taken more and more over that process over time

All data are imaged by pipeline

Only 10% of the raw data converted to science products in early science and included in the archive

ALMA Re-Imaging

Reprocess entire ES archive with pipeline

Archive is inhomogenous and incomplete

Improve the user-experience providing first-look products

Integrate it with Virtual Observatory services

ALMA Re-Imaging

Reprocess entire ES archive with pipeline

Archive is inhomogenous and incomplete

Improve the user-experience providing first-look products

Integrate it with Virtual Observatory services

ALMA Re-Imaging

Reprocess entire ES archive with pipeline

Archive is inhomogenous and incomplete

Improve the user-experience providing first-look products

Integrate it with Virtual Observatory services

ALMA Re-Imaging

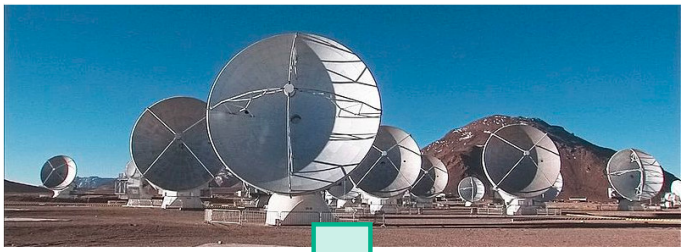
Reprocess entire ES archive with pipeline

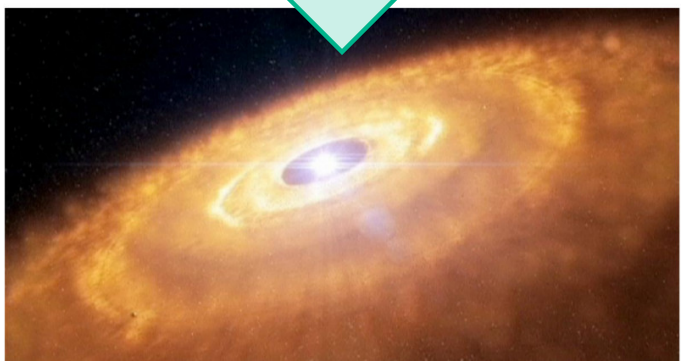
Archive is inhomogenous and incomplete

Improve the user-experience providing first-look products

Integrate it with Virtual Observatory services

Objectives





Develop automated procedure for complete imaging based on pipeline

Evaluate its unsupervised success rate

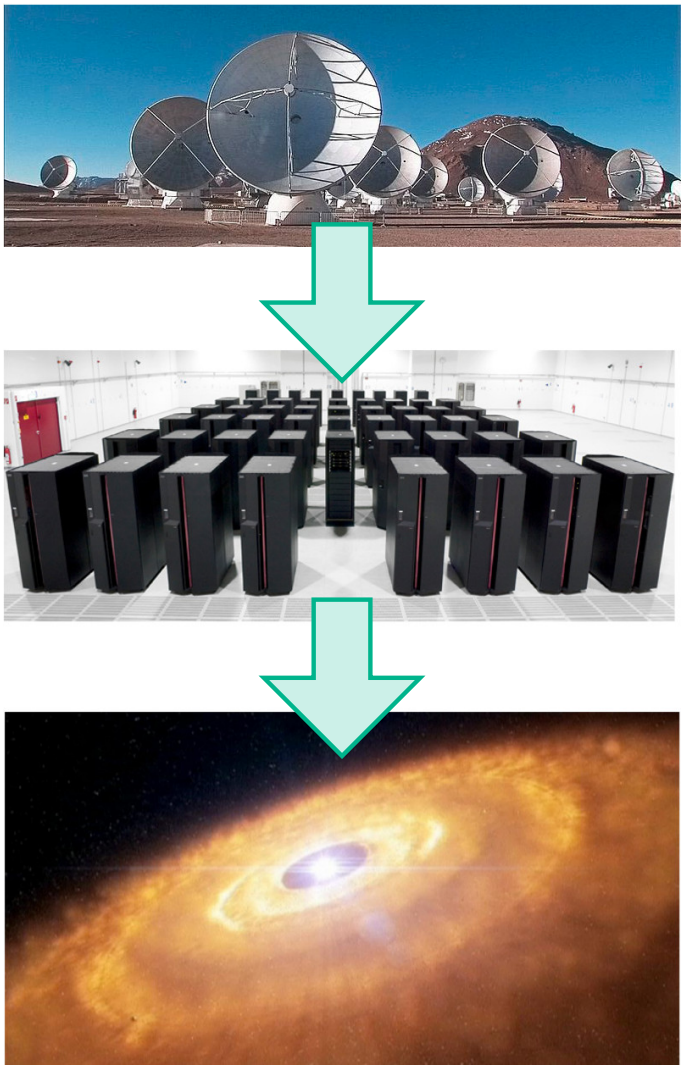
Evaluate performance wrt QA2

Estimate time and resources needed for completion

Evaluate possibility to run code on different machines

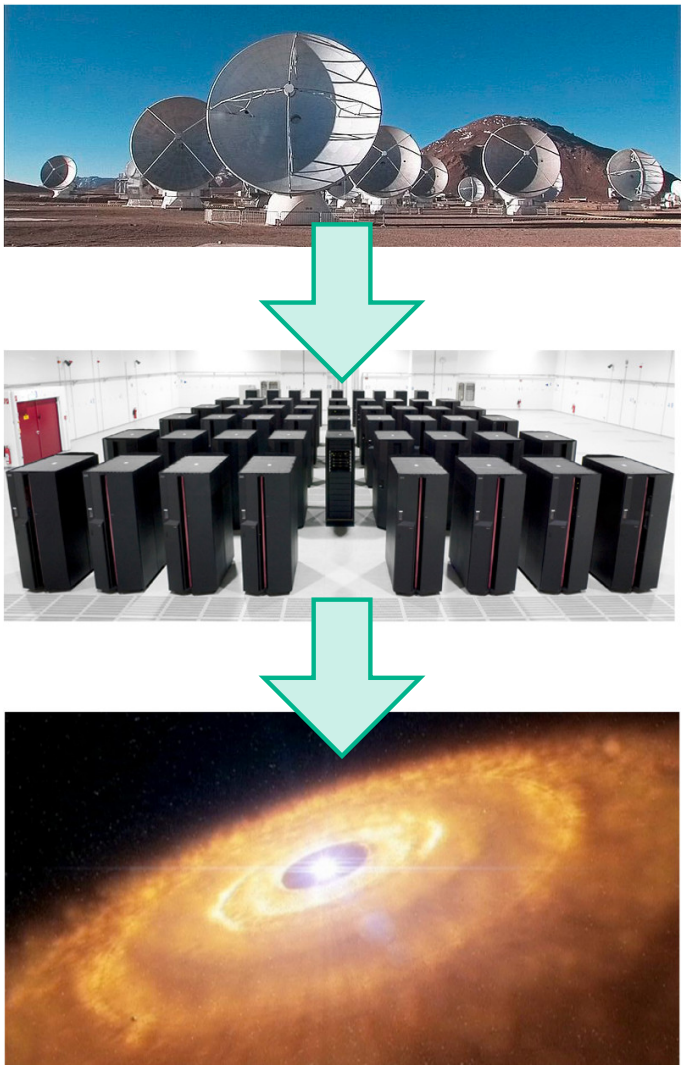
Assess feasibility of complete reimaging

Objectives



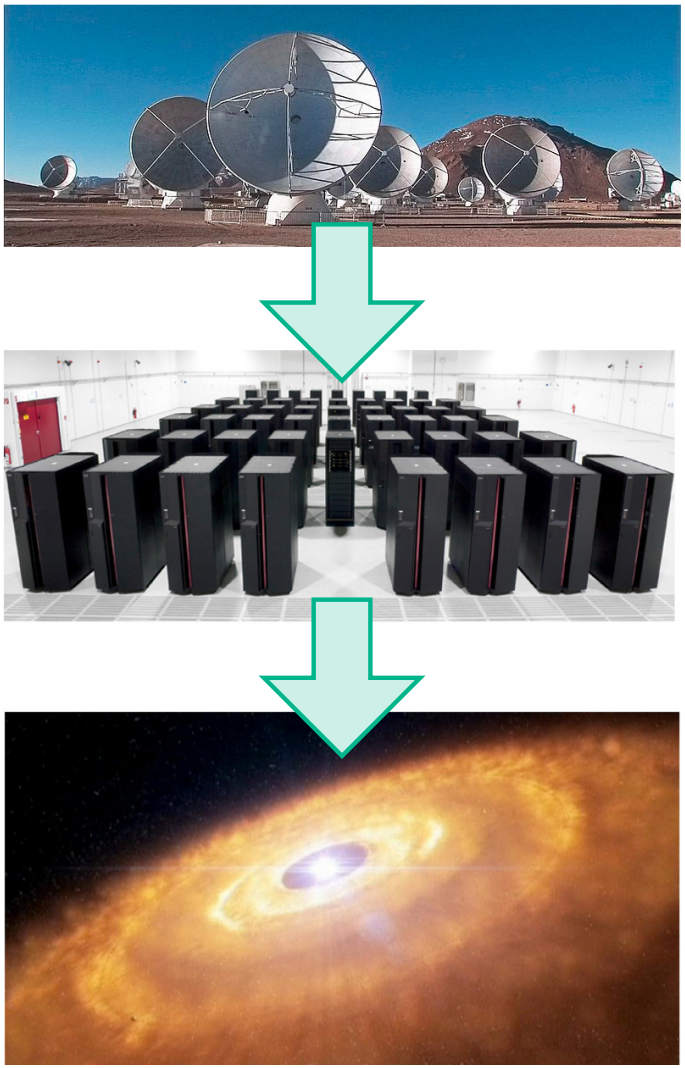
- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

Objectives



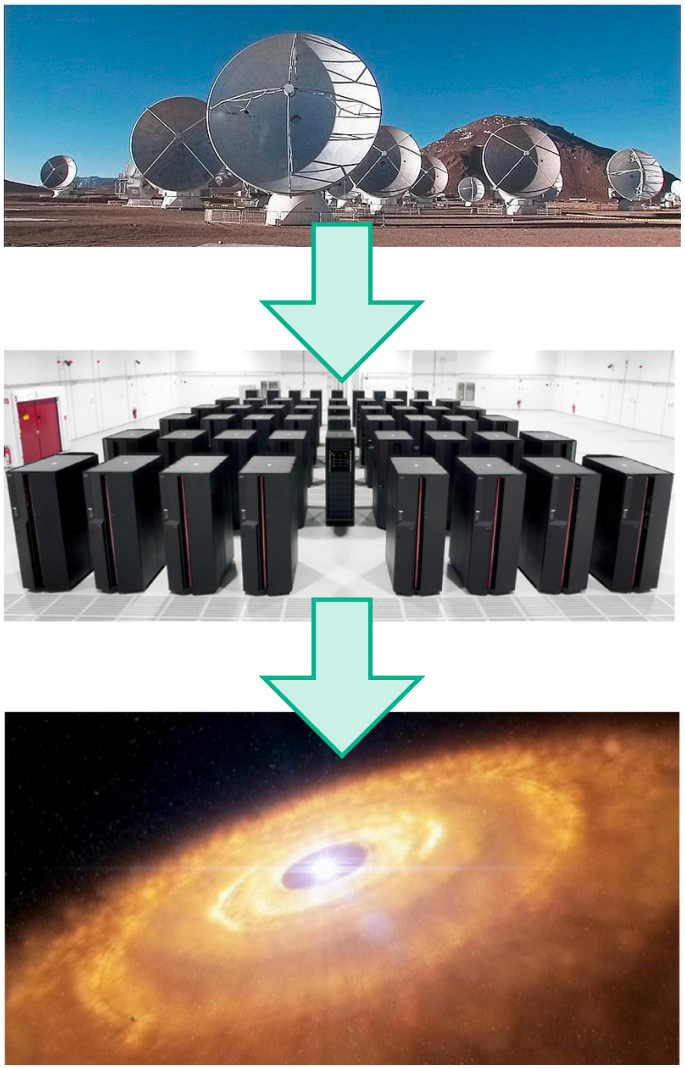
- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

Objectives



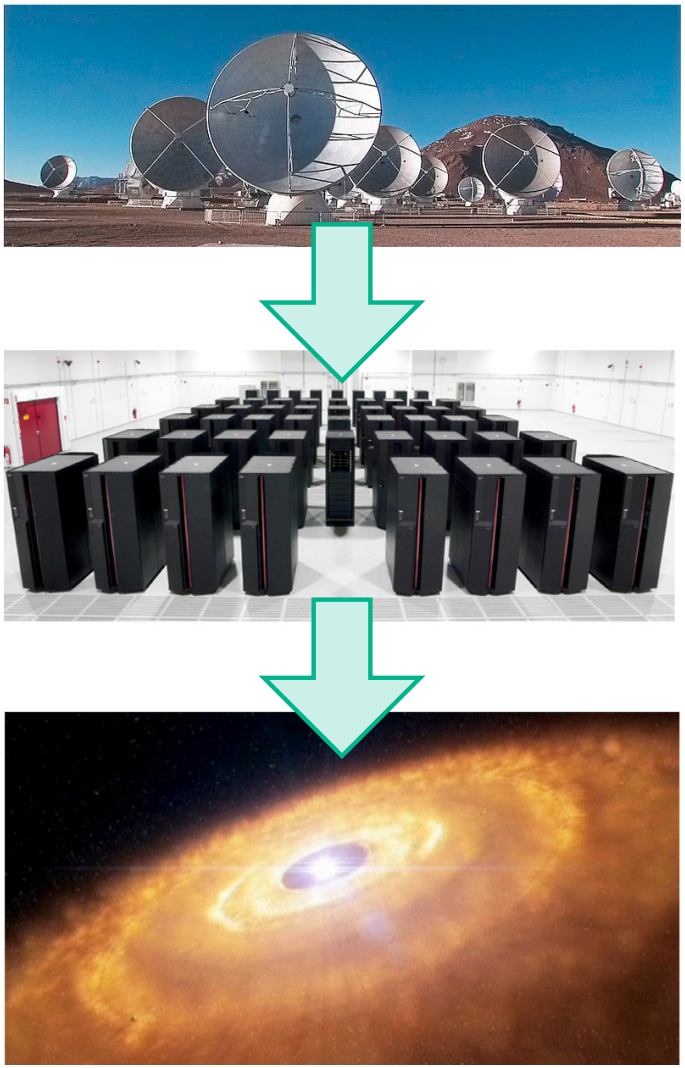
- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

Objectives



- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

Objectives



- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

Re-Imaging

Develop automated procedure for complete imaging based on pipeline

Evaluate its unsupervised success rate

Evaluate performance wrt QA2

Estimate time and resources needed for completion

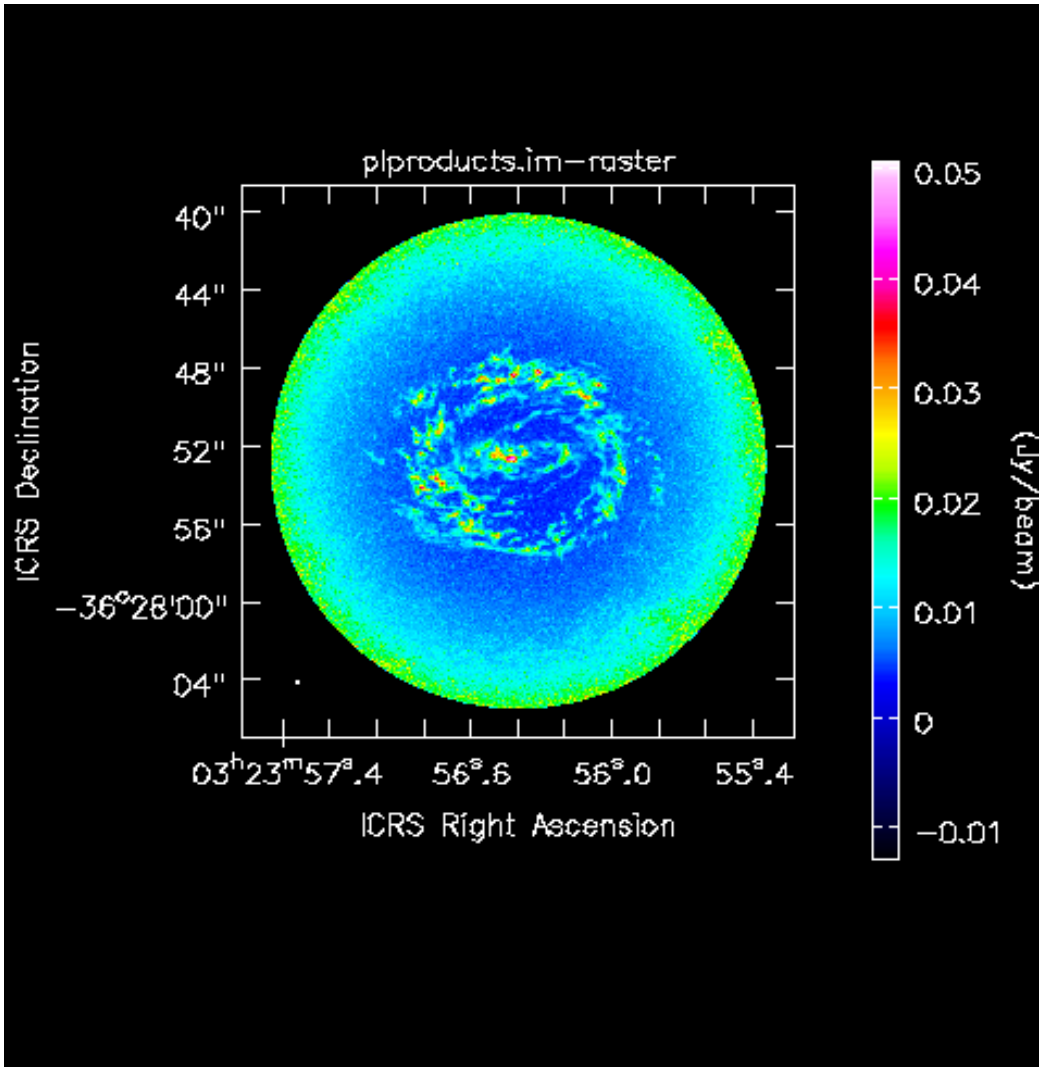
Evaluate possibility to run code on different machines

Assess feasibility of complete reimaging

- Performed on single observation
- Download of raw data products and scripts
- Calibration is restored from manual processing
- Imaging is done via the imaging pipeline
- On failure errors are captured

Re-Imaging

Develop automated procedure for complete imaging based on pipeline



Error statistics

Develop automated procedure for complete imaging based on pipeline

Evaluate its unsupervised success rate

Evaluate performance wrt QA2

Estimate time and resources needed for completion

Evaluate possibility to run code on different machines

Assess feasibility of complete reimaging

- Approx. 50% went through
- Solved problem and executed again: 10%
- UVcontfit errors: 9%
- Probable memory issues: 7%
- PL error (importdata; makeimlist; findcont): 8%
- Crash during calibration: 2%

- We expect that $\sim$ 80% of all data sets can be processed without incidents for late cycles

Comparison with QA2

- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

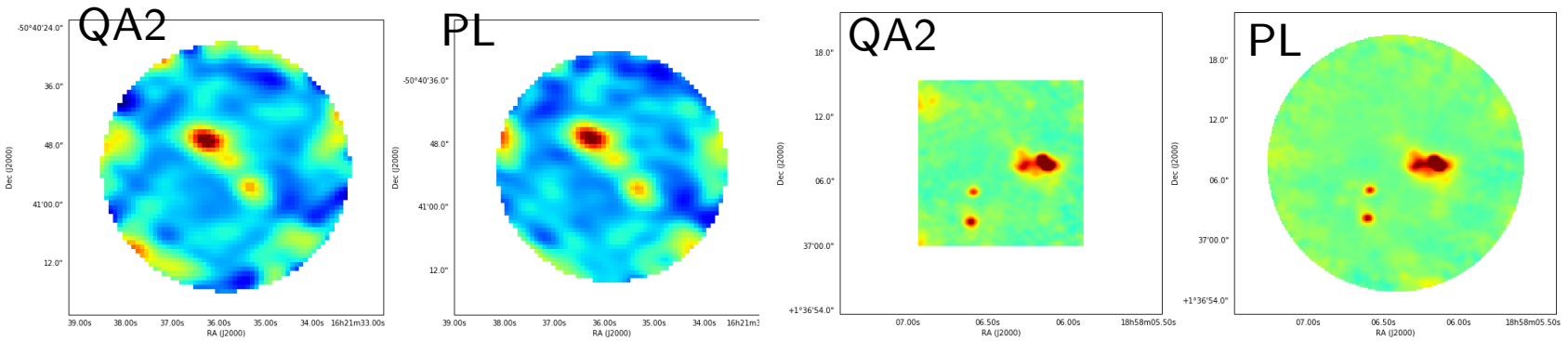
- By eye comparison of PL-QA2 products
- Automated comparison for all projects processed
  - Code to associate QA2-PL images
  - Development of a routine to compute fluxes
- Performance comparable in vast majority of images

Comparison with QA2

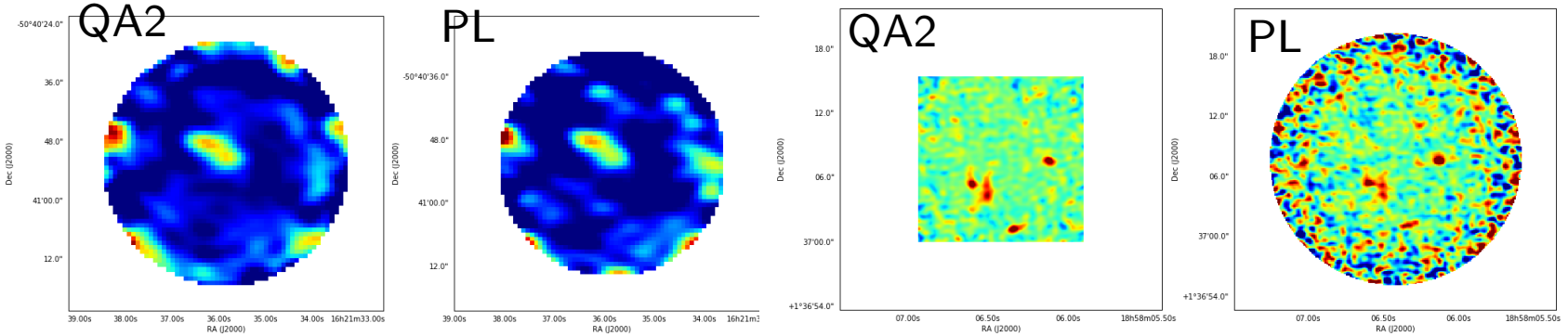
Manual comparison

Evaluate performance wrt QA2

Continuum



Lines



Comparison with QA2

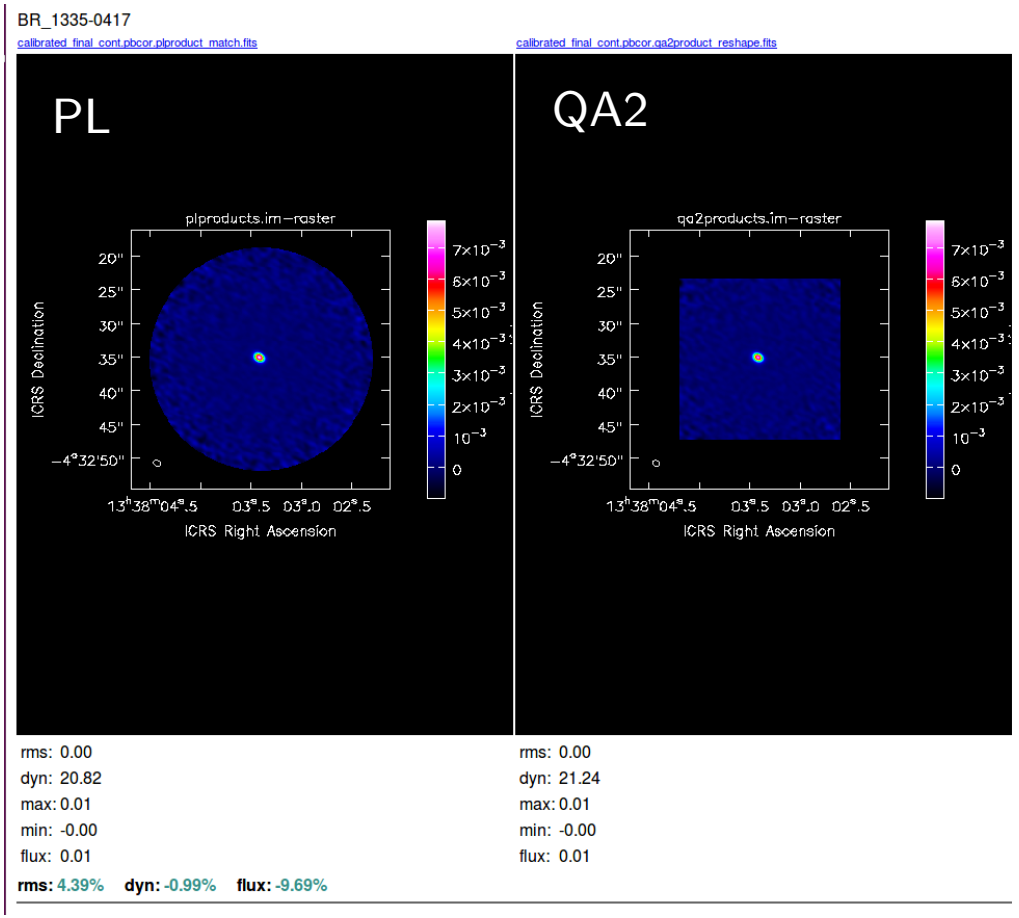
- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

- By eye comparison of PL-QA2 products
- Automated comparison for all projects processed
  - Code to associate QA2-PL images
  - Development of a routine to compute fluxes
- Performance comparable in vast majority of images

Comparison with QA2

Automatic comparison

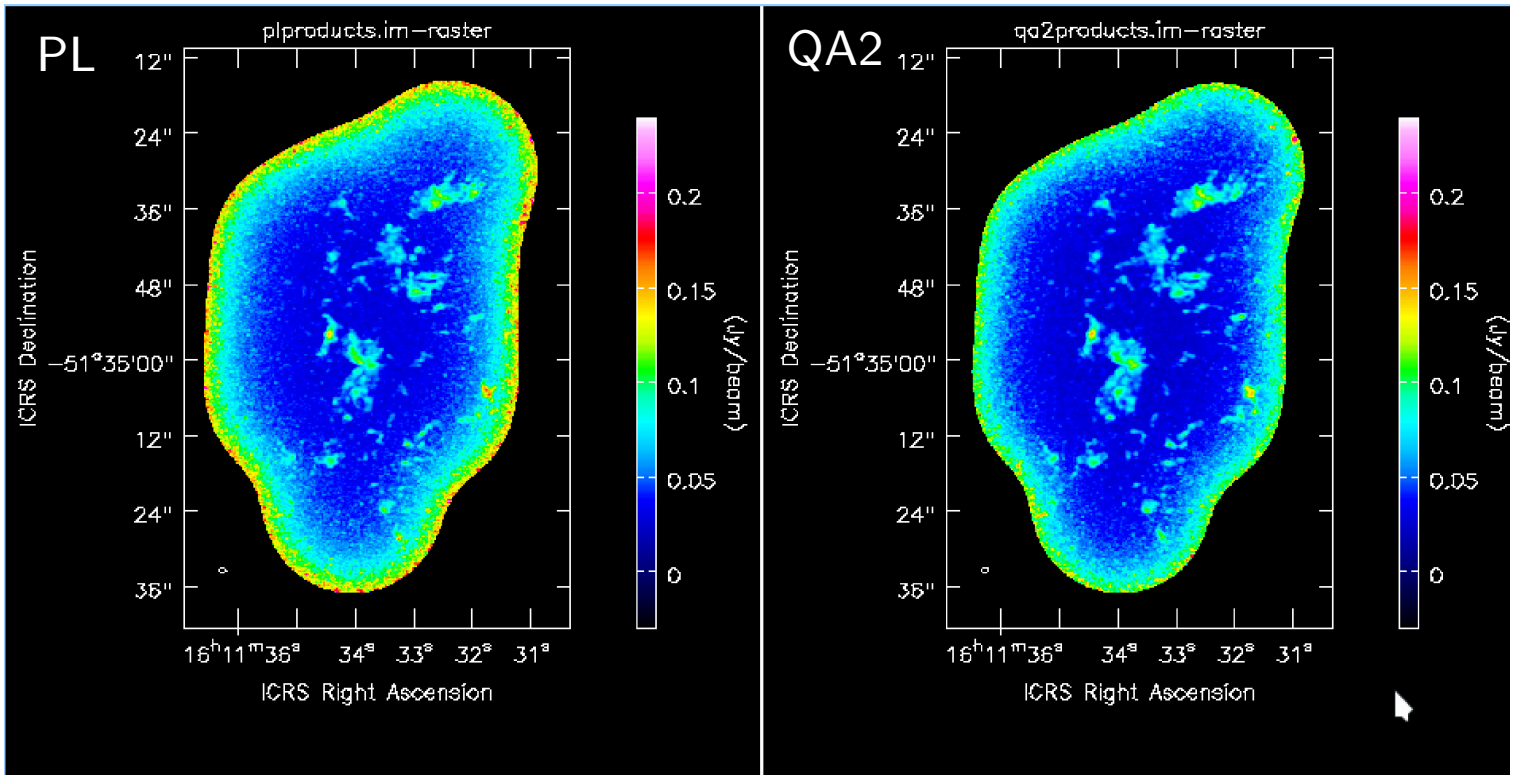
Evaluate performance wrt QA2



Comparison with QA2

Automatic comparison

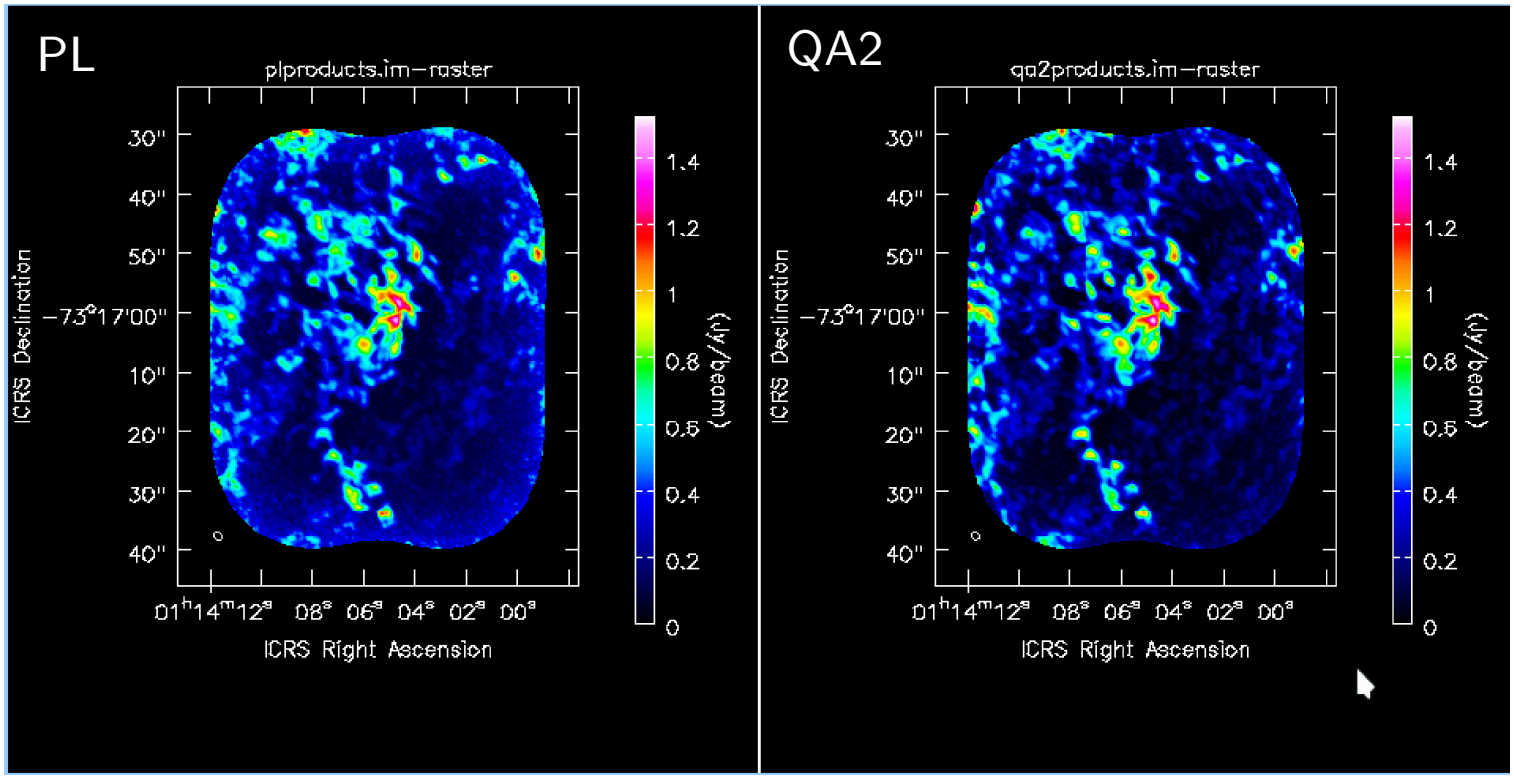
Evaluate performance wrt QA2



Comparison with QA2

Automatic comparison

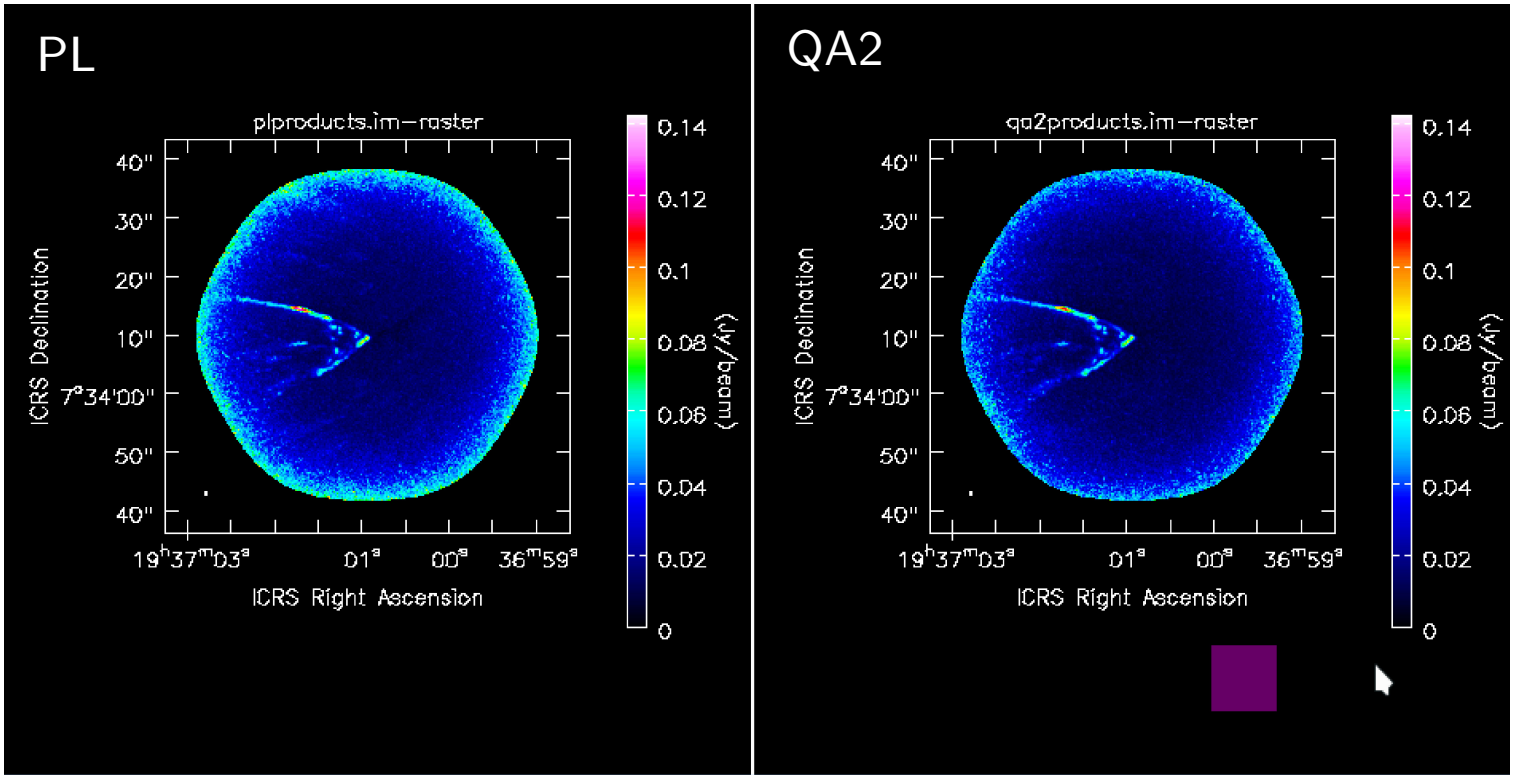
Evaluate performance wrt QA2



Comparison with QA2

Automatic comparison

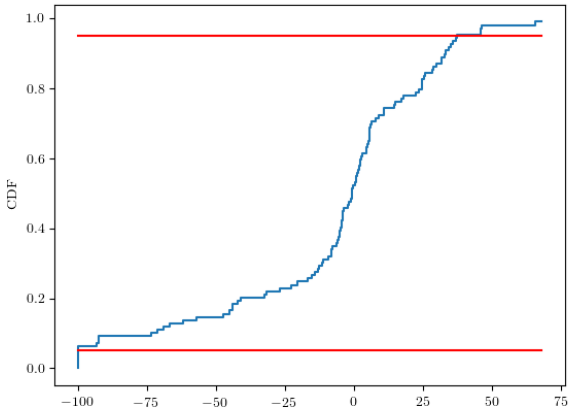
Evaluate performance wrt QA2



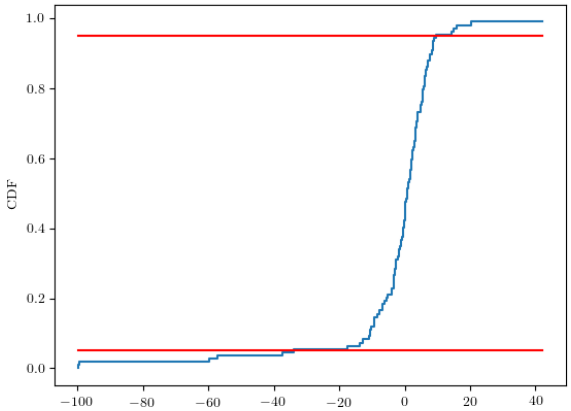
Comparison with QA2

Automatic comparison

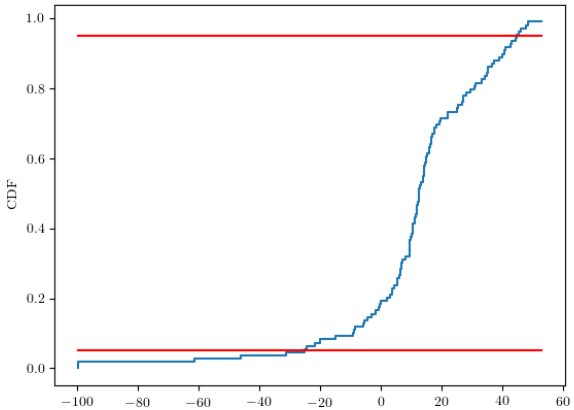
Evaluate performance wrt QA2



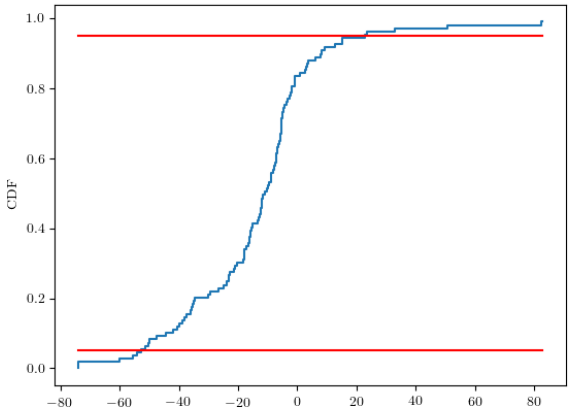
Flux



Max



RMS noise

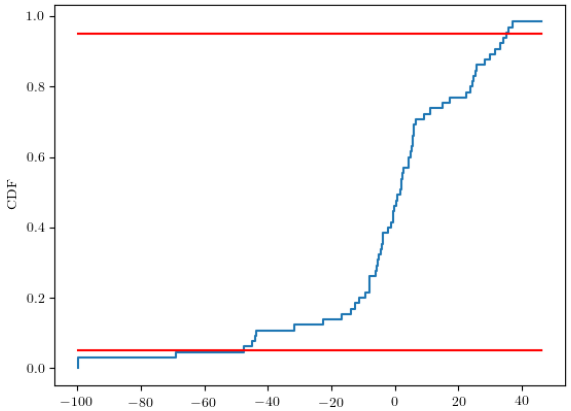


Dyn. range

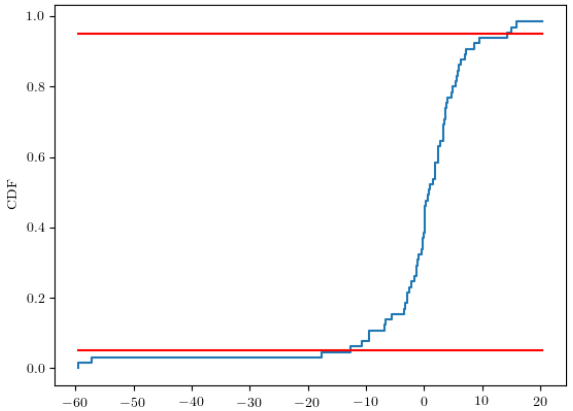
Comparison with QA2

Automatic comparison

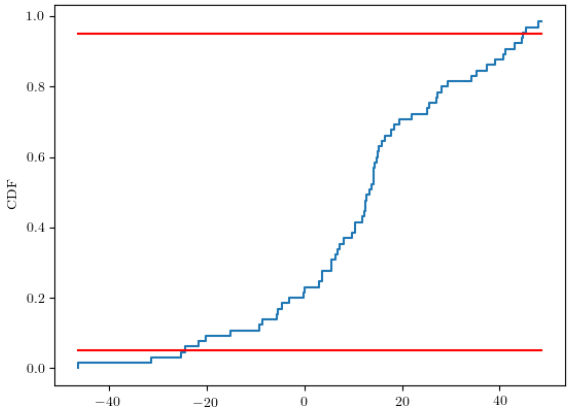
Evaluate performance wrt QA2



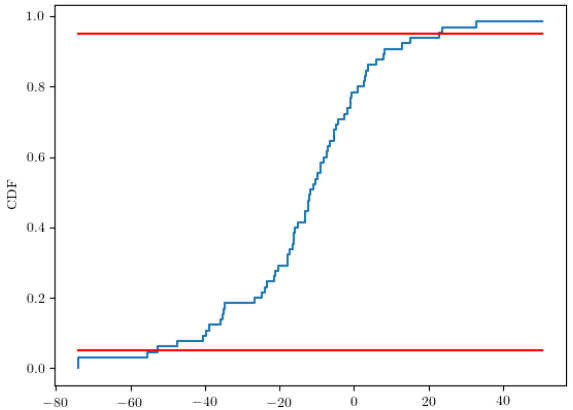
Flux



Max



RMS noise

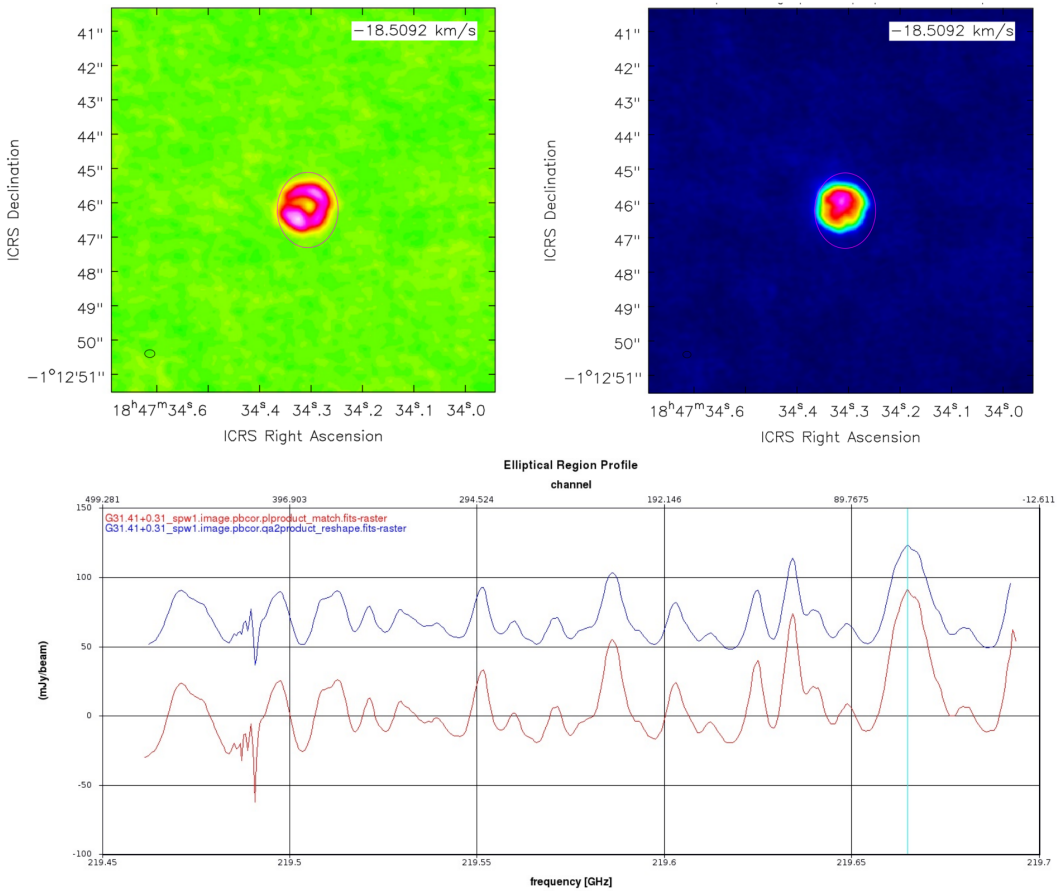


Dyn. range

Comparison with QA2

Automatic comparison

Evaluate performance wrt QA2



Comparison with QA2

- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

- By eye comparison of PL-QA2 products
- Automated comparison for all projects processed
  - Code to associate QA2-PL images
  - Development of a routine to compute fluxes

- Performance comparable in vast majority of images

Time and resources needed

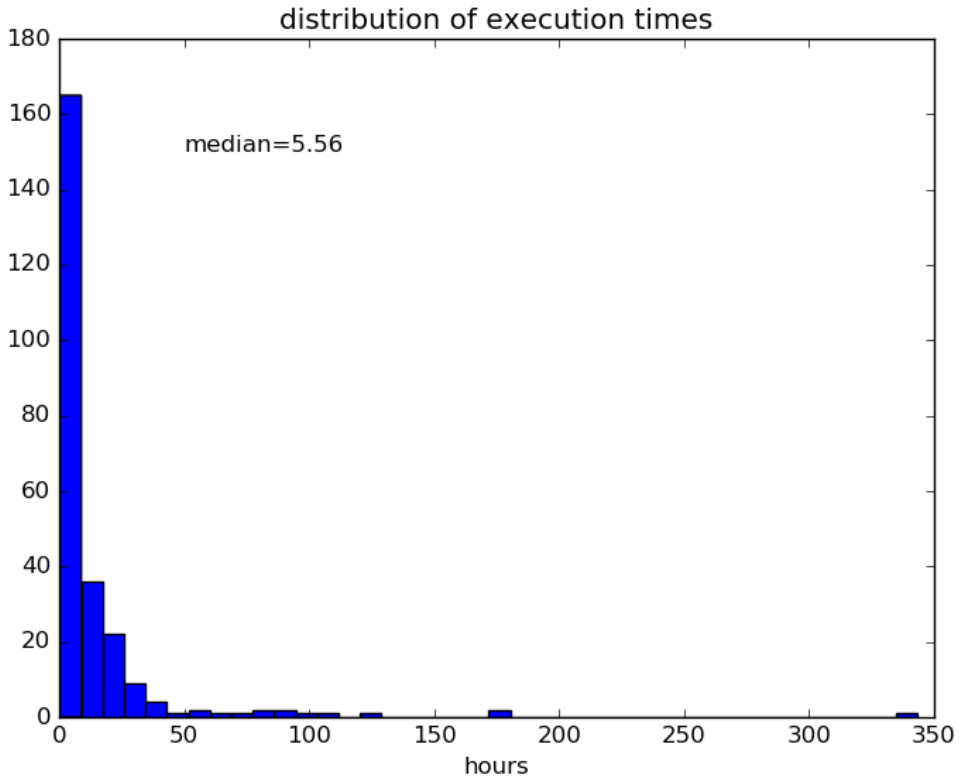
- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

- As of today $\sim$ 6000 data sets
- Median time needed per data set: $\approx$ 6 hours
- Complete reimaging would need $\sim 3-4$ years on a single machine
- Italian node can process small projects on older machines (11) + HPC @ INAF
- Evaluating possibility to expand cluster with low-cost machines

Time and resources needed

Average processing time per data set

Estimate time and resources needed for completion



Time and resources needed

- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

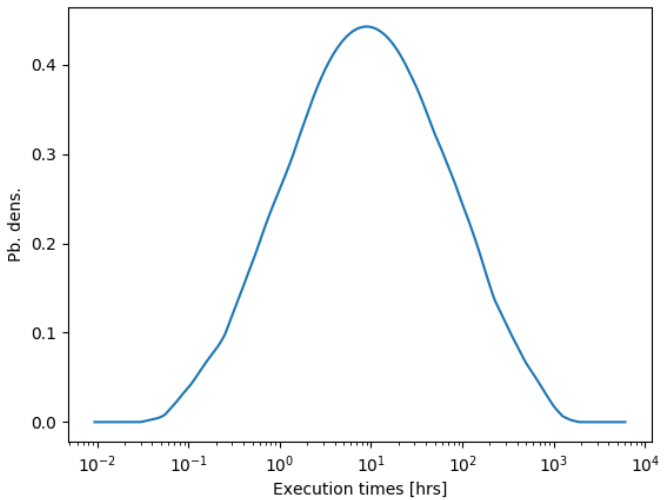
- As of today $\sim$ 6000 data sets
- Median time needed per data set: $\approx$ 6 hours
- Complete reimaging would need $\sim 3 - 4$ years on a single machine
- Italian node can process small projects on older machines (11) + HPC @ INAF
- Evaluating possibility to expand cluster with low-cost machines

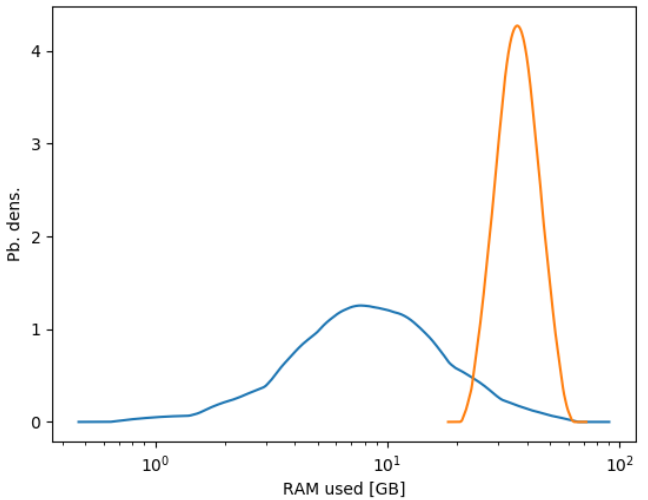
ARI outside ESO

- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Evaluate possibility to run code on different machines
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

- IRA ARC cluster
- MUP cluster
  - 16 nodes
  - 12 cores (24 hyper-threading) Intel Xeon E5-2620
  - 64 GB RAM
  - NFS storage: 70TB (RAID 5)
- Successfully running on both

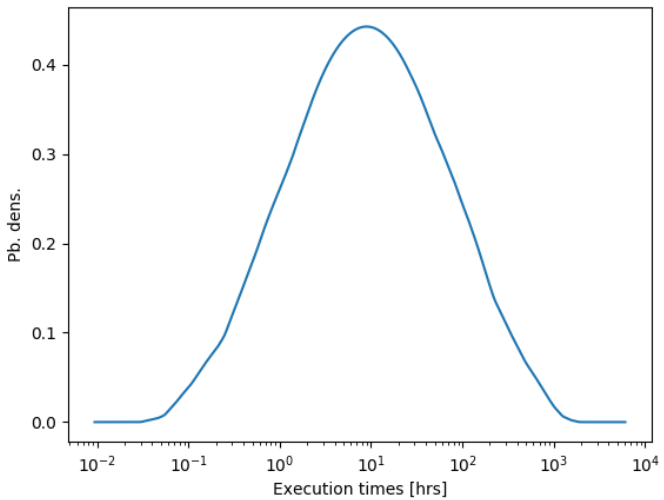
Results from MUP cluster

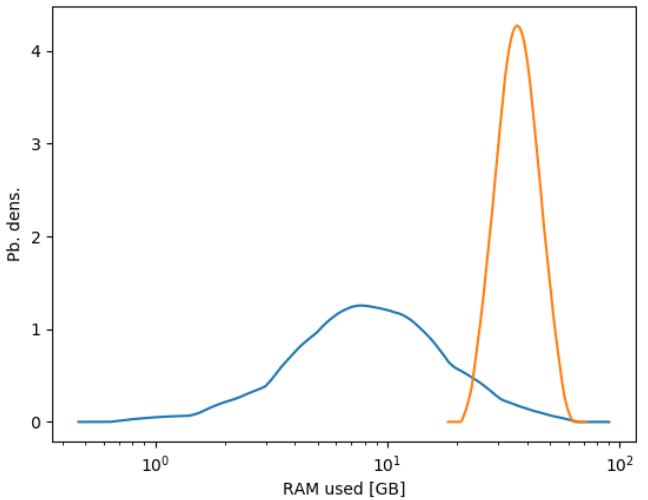




- Processed 79 data sets
  - Successful: 29
  - Walltime: 26
  - Download problem: 14
  - ARI error: 9
  - TP dataset: 1
- Median runtime: $\sim$ 10 hr (including download)
- All large datasets hit wall-time

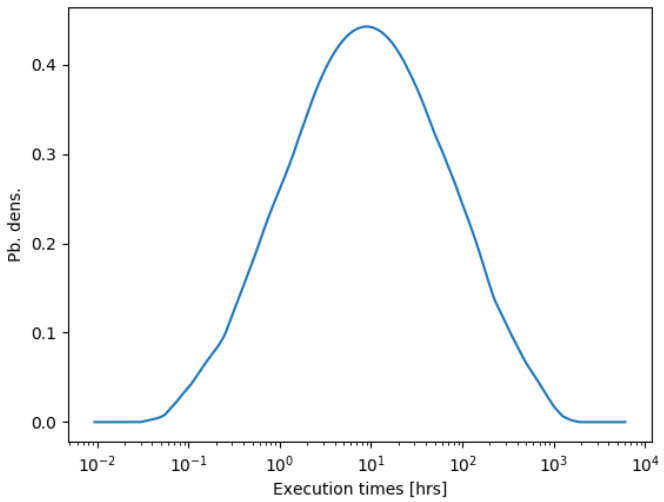
Results from MUP cluster

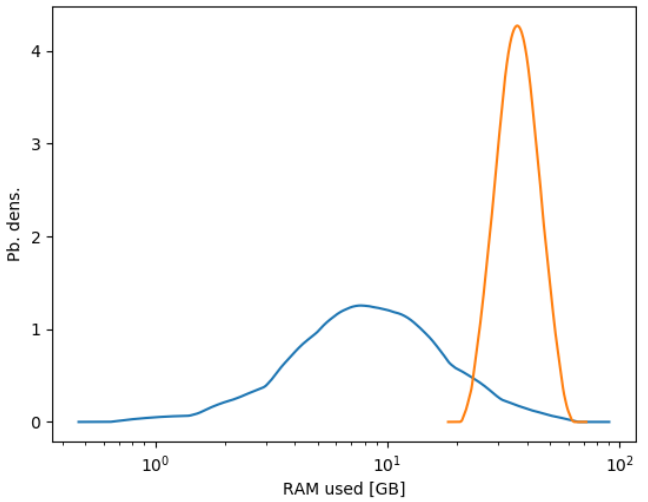




- Processed 79 data sets
  - Successful: 29
  - Walltime: 26
  - Download problem: 14
  - ARI error: 9
  - TP dataset: 1
- Median runtime: $\sim$ 10 hr (including download)
- All large datasets hit wall-time

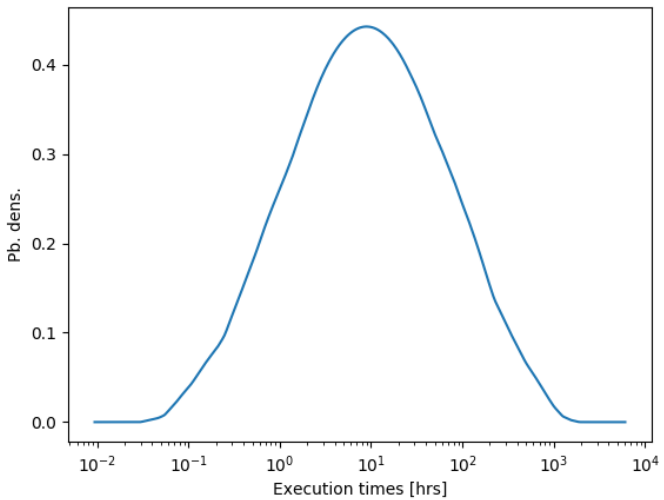
Results from MUP cluster

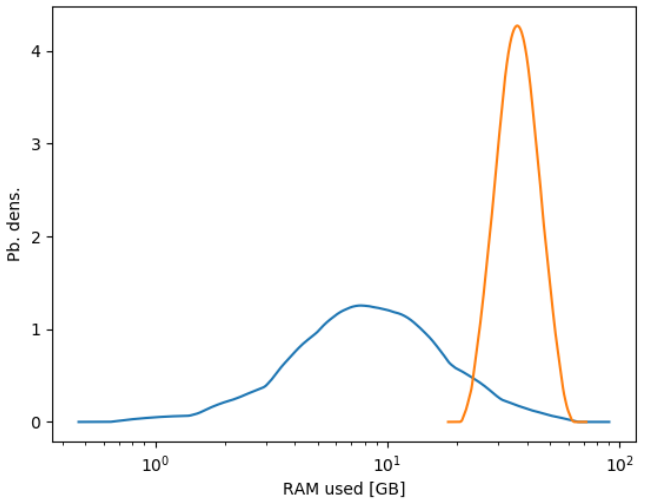




- Processed 79 data sets
  - Successful: 29
  - Walltime: 26
  - Download problem: 14
  - ARI error: 9
  - TP dataset: 1
- Median runtime: $\sim$ 10 hr (including download)
- All large datasets hit wall-time

Results from MUP cluster





- Processed 79 data sets
  - Successful: 29
  - Walltime: 26
  - Download problem: 14
  - ARI error: 9
  - TP dataset: 1
- Median runtime: $\sim$ 10 hr (including download)
- All large datasets hit wall-time

Assess feasibility

- Develop automated procedure for complete imaging based on pipeline
- Evaluate its unsupervised success rate
- Evaluate performance wrt QA2
- Estimate time and resources needed for completion
- Evaluate possibility to run code on different machines
- Assess feasibility of complete reimaging

- Time: with 5 machines $\sim$ 6 months
- Costs:
  - How many dedicated machines?
  - How much space do we need?
  - People (FTE)?