



The ALMA Re-Imaging (ARI) Development study



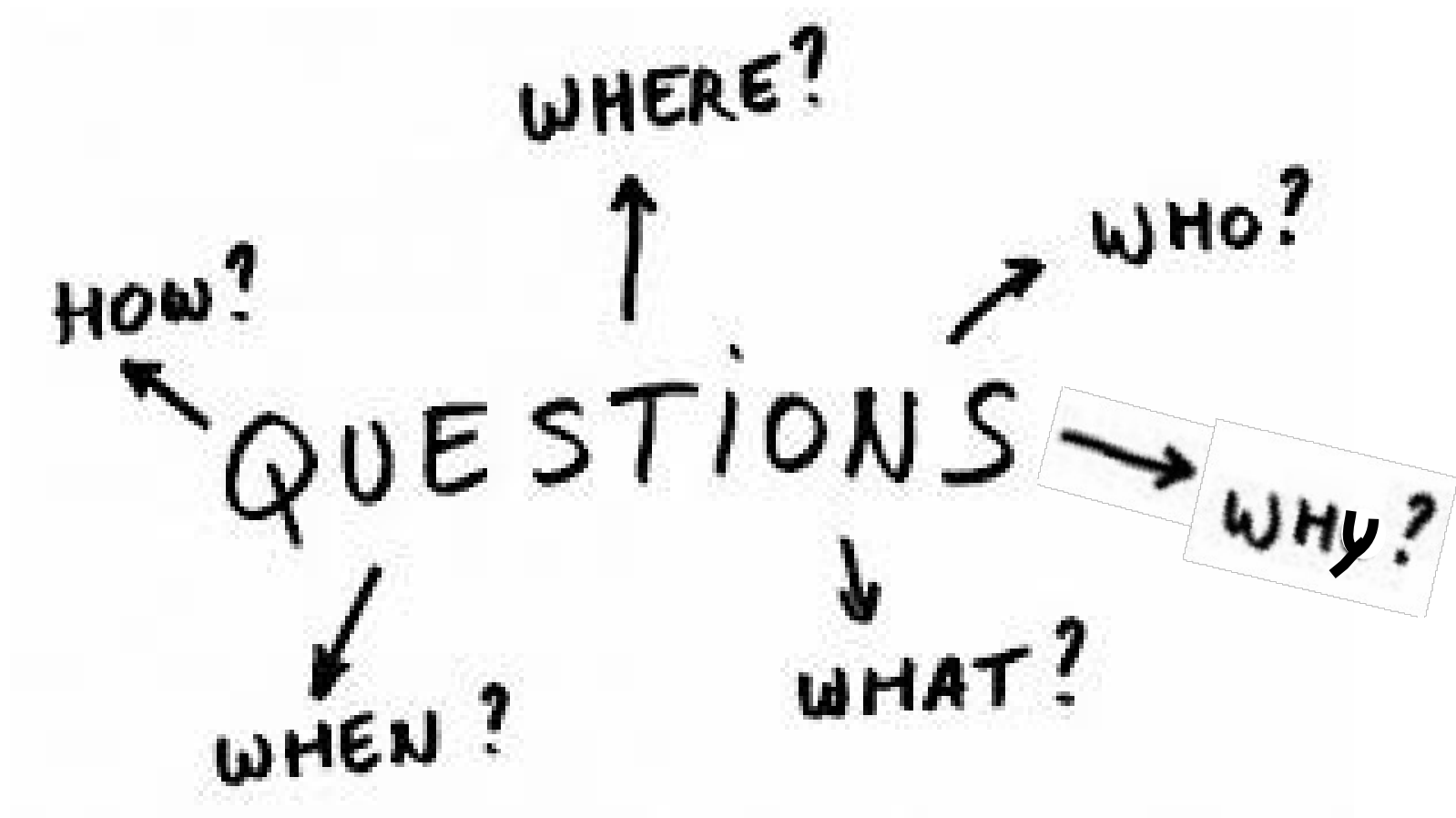
EUROPEAN ARC
ALMA Regional Centre || Italian

Marcella Massardi

*ALMA Archive and
Pipeline workshop
24-25 January 2017*



Re-Imaging the ALMA archive



Current status of the Archive... from a miner perspective

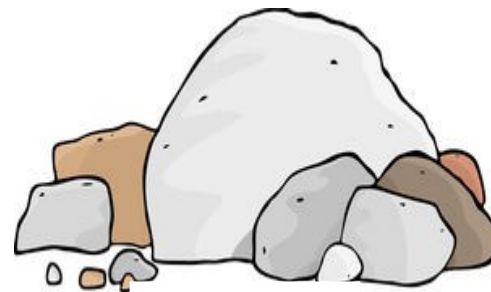
Why?



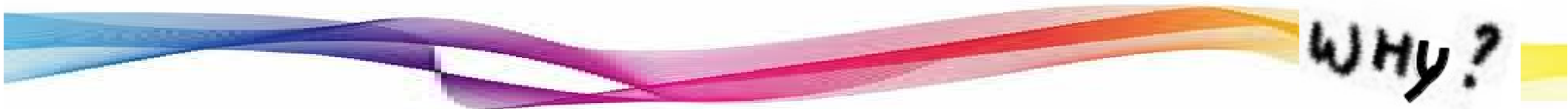
- more than 1500 project observed and archived
- data are already calibrated
- not all the possibilities have been exploited
- archive will keep growing

BUT

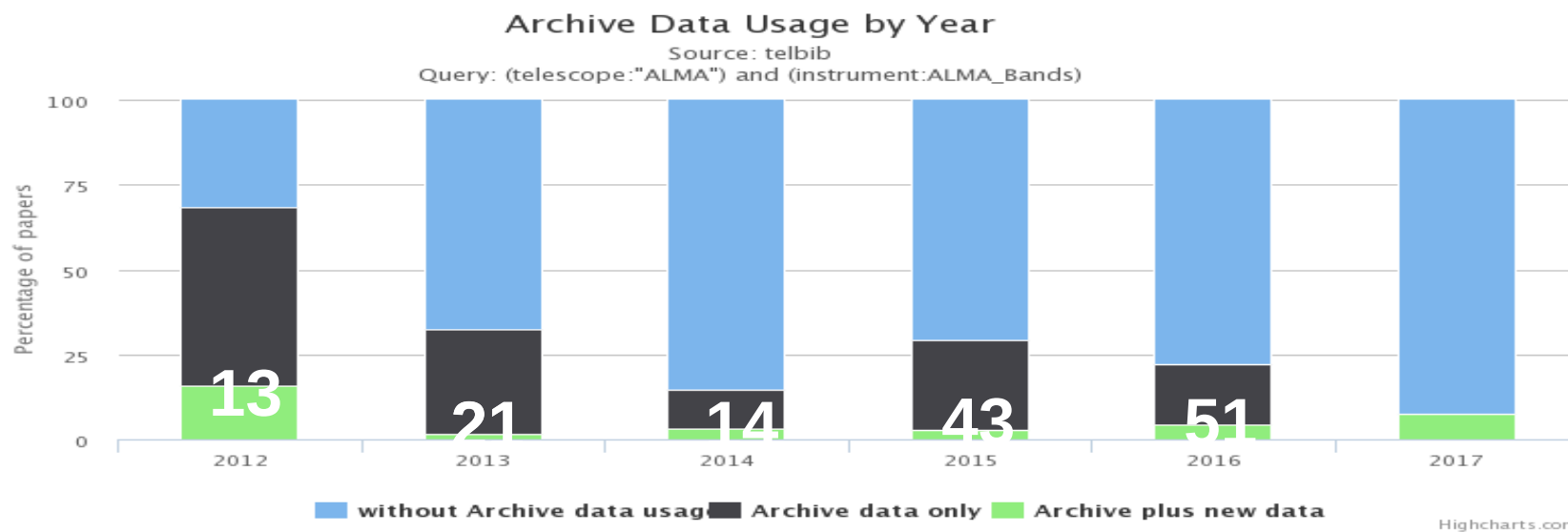
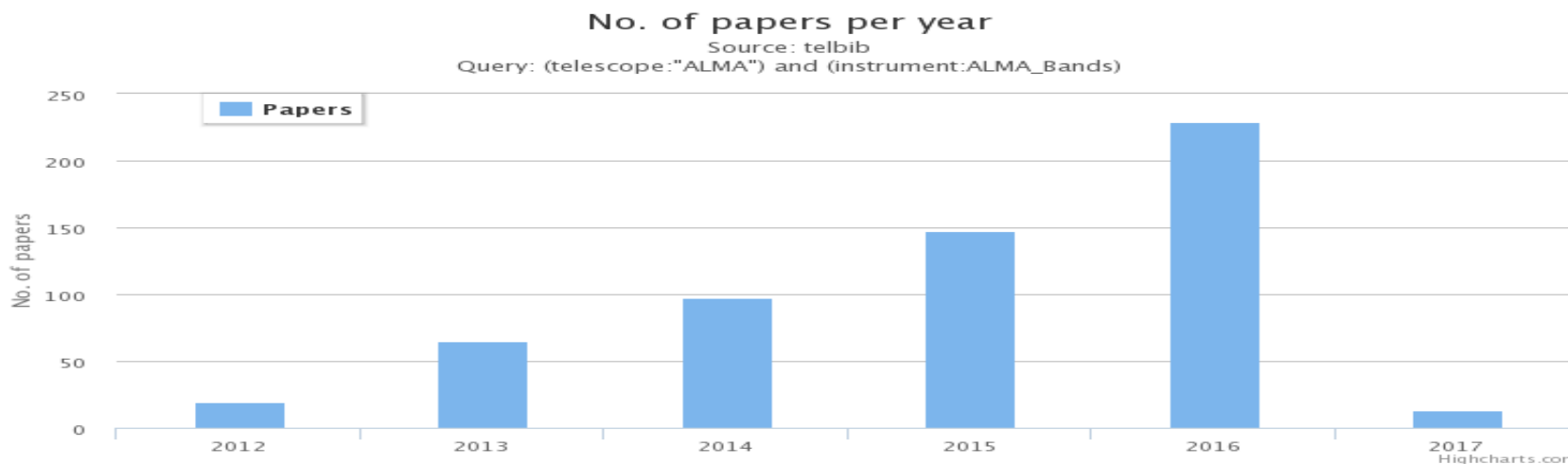
- dataset are huge
- images available are made for QA2
 - = incomplete
 - = inhomogeneous
 - = not easily comparable
- to understand scripts I need interferometry induction/skill (luckily I have the ARC nodes)
- to run scripts I need CASA (which version? How? ...)



Current status of the Archive... from a miner perspective



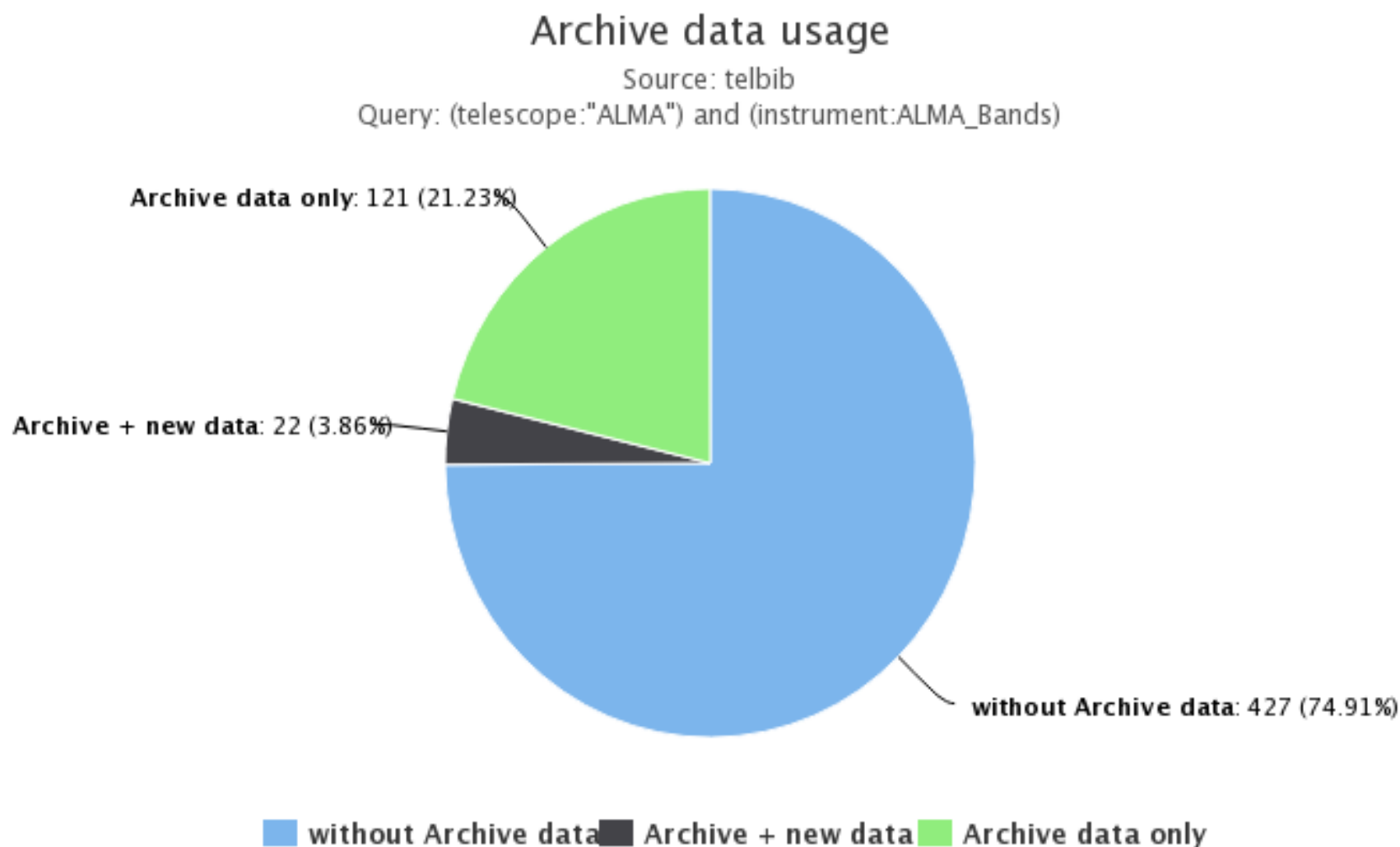
...NEVERTHELESS >143/570 papers include archival and SV data



Current status of the Archive... from a miner perspective

Why?

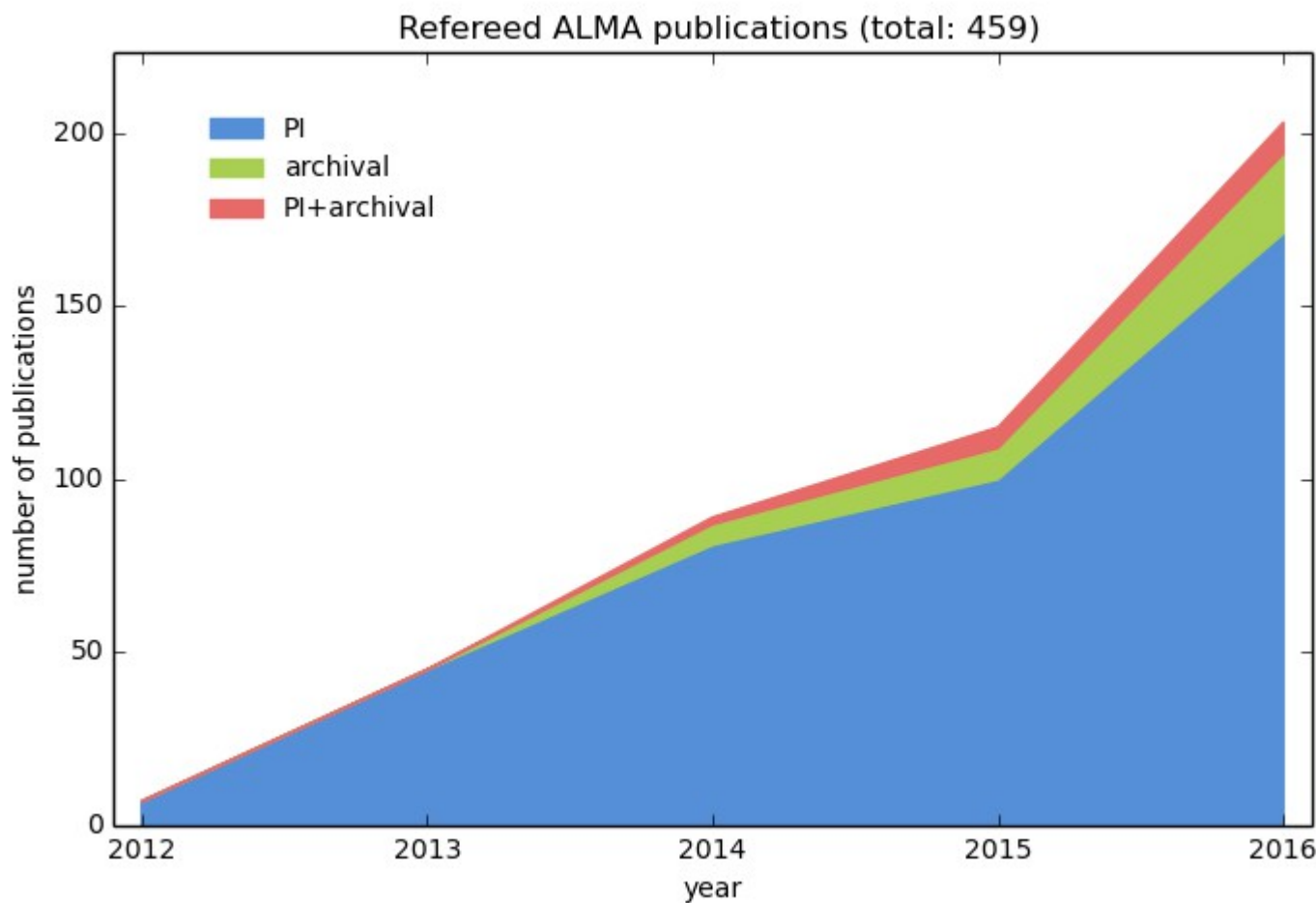
...NEVERTHELESS >143/570 papers include archival and SV data



Current status of the Archive... from a miner perspective



...ONLY 50/459 papers include archival data (no SV up to 2016)

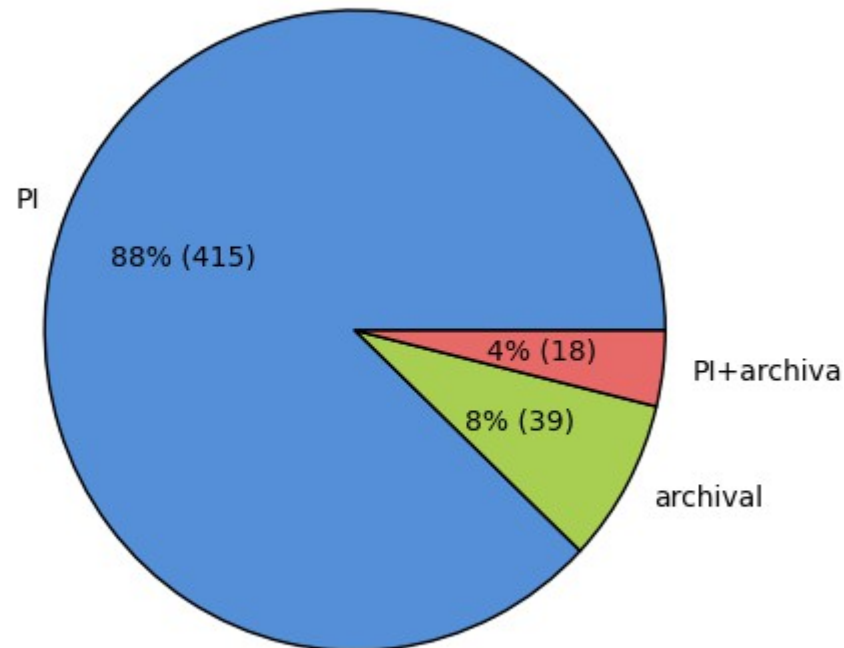


Current status of the Archive... from a miner perspective

Why?

...ONLY 58/472 papers include archival data (no SV as today)

Refereed ALMA publications (total: 472)

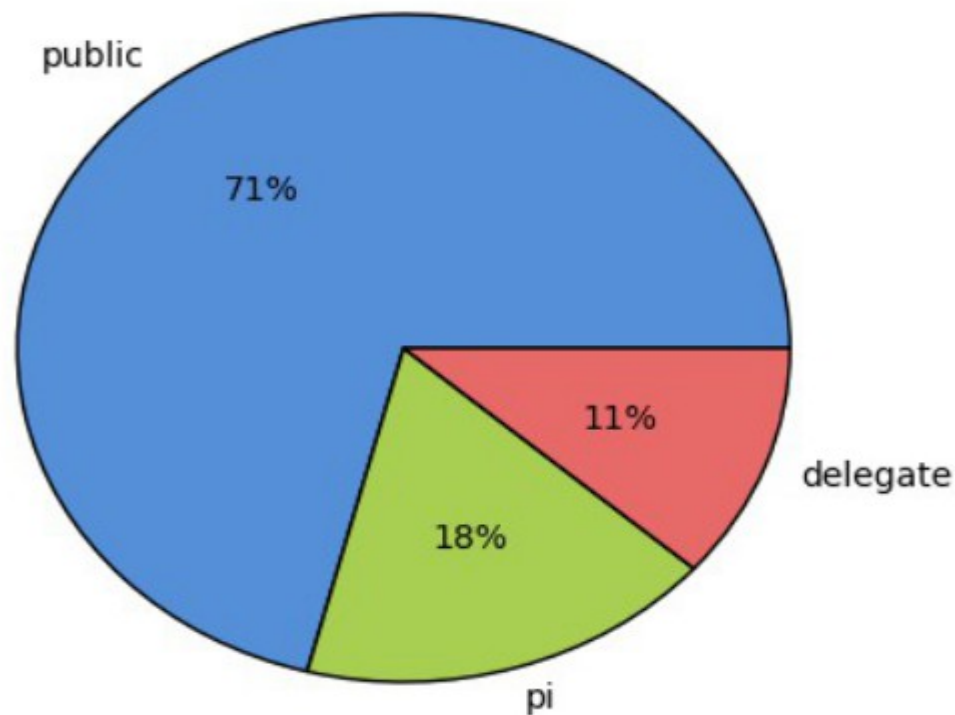


Current status of the Archive... from a miner perspective

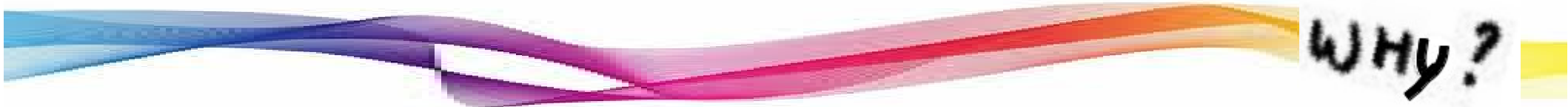
Why?

...and in most of the cases they have been downloaded outside the proprietary period

Downloaded ALMA data (total: 259 TB)



Current status of the Archive... from a miner perspective



WHY?

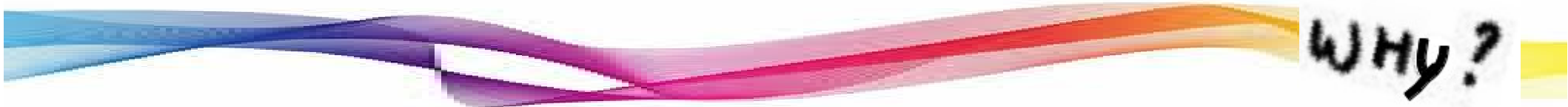
But in the helpdesk users require

- **products representative of the data potential**
- **easy way to access the info about the data**

Currently

- **only a small fraction (<10%) of science channels are converted in cubes (<3% if cals are included)**
- **imaging requires downloading all the data and running all the scripts**
- **products are downloaded twice than raw data**

Current status of the Archive... from in the ALMA dev plans



WHY?

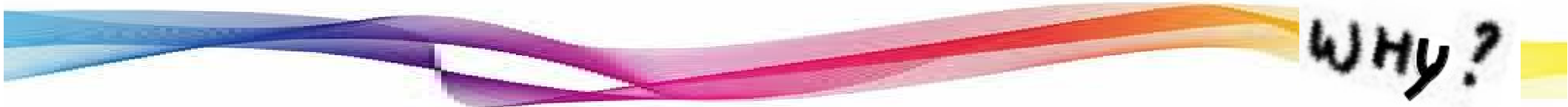
ASAC view in “Road map for developing ALMA” towards ALMA 2030

In order for the archive to be productive, it needs to be public, searchable, easy to mine, and it needs to contain fully reduced science-grade data products.

Analysis of the productivity of mature facilities shows that publications using archival data can rapidly overtake the publications from the original proposers acquiring the dataset, as is the case for the Hubble Space Telescope and other facilities. Thus the archive may be what ultimately determines the productivity of ALMA.¹

Developing the ALMA archive into a fully-fledged science-grade minable archive, however, requires significant further development into pipelines and automated analysis.

Current status of the Archive... from in the ALMA dev plans



WHY?

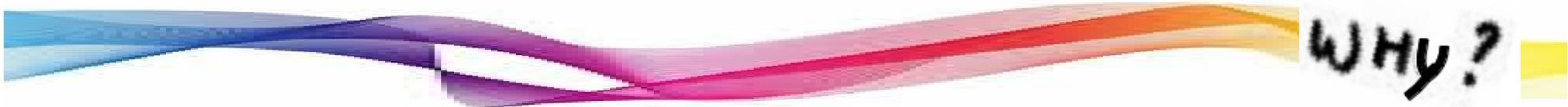
ALMA Development group in “Pathways to Developing ALMA”

ALMA needs to provide the tools to find data in an enriched and forever expanding data and software archive... It is envisioned that as pipeline heuristics improve, reprocessing of archived data is essential and will be supported. The content of the archive will therefore improve as heuristics improve... It is anticipated that pipeline improvements to perfect imaging and calibration and to add new observing modes will be ongoing throughout the lifetime of the observatory...

**The archive is envisioned as an evolving framework,
easy to access and mine,
open to new tools and improvements.**

**It is not the final stage of scientific analysis,
but a portal to begin the analysis**

The archive needs also images



WHY?

- 1) To easily download FITS product files instead of large dataset
- 2) To allow creating previews of the data content
- 3) To generate first-look accessible also to non-expert that do not use CASA
- 4) To quickly access large subset of data (e.g. for stacking)
- 5) To compare homogeneous products from different dataset
- 6) To use CARTA in the archive interface
- 7) To use ADMIT and run preliminary analysis
- 8) To add ALMA data in the VO in a coherent way
- 9) To use VO tools to manipulate ALMA images
- ...

Here comes ARI

WHAT?



The **ALMA Re-Imaging development study** will try to **define the feasibility, necessary efforts and cost of the re-imaging of the whole ALMA archive for cycles 0-4 using the ALMA Imaging pipeline.**

Purpose of re-imaging:

- provide homogeneous products in the form of image
- for all the data in cycles 0-4
- quality of the images will be as defined by the pipeline

Caveats and rules of the game:

- we do not aim at generating the best science image but only one image representative of the data content and comparable with other from similar projects
- we are not making pipeline quality checks or commissioning
- we will investigate the quality of the products to the only purpose to advise the archive user of its content
- new images will be added to the archive (not substitute to QA2) dataset trees and process will remain the same
- nothing will be changed in calibrations

Your proposal was well received, but we are unsure whether we can fit your study in the budget available, until all higher priority study agreements have been placed. We should be able to have a final answer by the end of the first quarter of 2017.

ESO Contract Officer

Toward the re-imaging feasibility assessment: the ARI study

How?



- * **Categorising the archive**
Projects will be included in categories according to science and technical requests for images
- * **Selecting samples from each category**
- * **Testing the Imaging Pipeline on each category**
- * **Verifying the feasibility for each category**
(i.e. amount of data to be images, success rate, running time, hardware requirements for storage and transfer quality of the data)
- * **Produce a prototype of the re-imaging software to automatically**
 - 1) extract a dataset from the ALMA archive
 - 2) detect the CASA version necessary
 - 3) execute the restoring of the calibrated data from the raw data
 - 4) execute the official ALMA Pipeline Imaging on the calibrated data
 - 5) create FITS products from the images
 - 6) execute post-imaging analysis like the creation of footprints and previews
 - 7) execute ADMIT (if delivered in time for the present study)

The ARI study: risks and reasons for a study

How?



Feasibility might be a function of cycle and category

It is difficult to categorize the archive

It is difficult to predict the sample size for testing

The pipeline may not work on old data

Are the product representative of the dataset?

What is the definition of science-grade?

The ARI study: action items and open questions



HOW?

- * **Categorising the archive**

- Categories : galactic/extragalactic
 - Extended/pointlike
 - Continuum/line

- ...
 - ?

- * **Selecting samples from each category**

- Is sample size a function of the category?

- * **Testing the Imaging Pipeline on each category**

- Quality tests (i.e. comparison with manual/previous QA2 products, Matching with science purposes?)

- * **Verifying the feasibility for each category**

- Definition of advises for the archive users

- * **Produce a prototype of the re-imaging software to automatically run the PL and other tools**

The ARI study: action items and open questions

HOW?



- * **Categorising the archive**

Categories : galactic/extragalactic
Extended/pointlike
Continuum/line

...
?

- * **Selecting samples from each category**

Is sample size a function of the category?



- * **Testing the Imaging Pipeline on each category**

Quality tests (i.e. comparison with manual/previous QA2 products,
Matching with science purposes?)



- * **Verifying the feasibility for each category**

Definition of advises for the archive users

- * **Produce a prototype of the re-imaging software to automatically run the PL and other tools**

(let's leave these questions to the discussion sessions...)

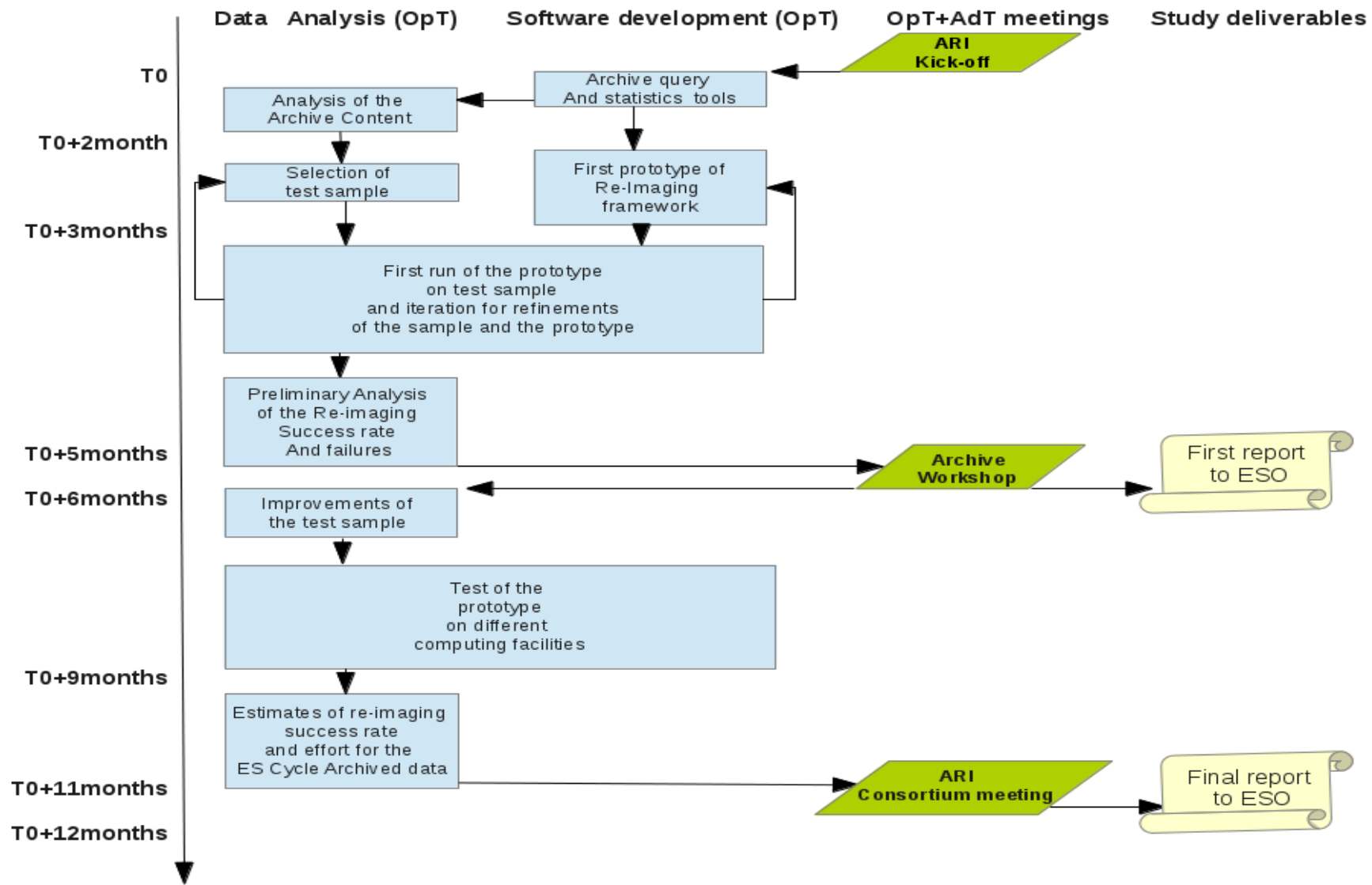
The ARI study people

WHO?

Main Activities	Institute/Staff
Management	INAF/Marcella Massardi
Prototype development	ESO/Felix Stoehr
Software development, Sample selection, Prototype testing, Archive Analysis	Operative Team
Quality control	Operative Team, Advisor Team

The ARI milestones

WHEN ?



The ARI study deliverables



The ARI development study's major deliverables will be a report detailing the technical feasibility, the achievable quality as well as the cost of the proposed project including the analysis of

- 1) the analysis of the re-imaging feasibility
- 2) the production of a prototype of the re-imaging software
- 3) the evaluation of the product quality