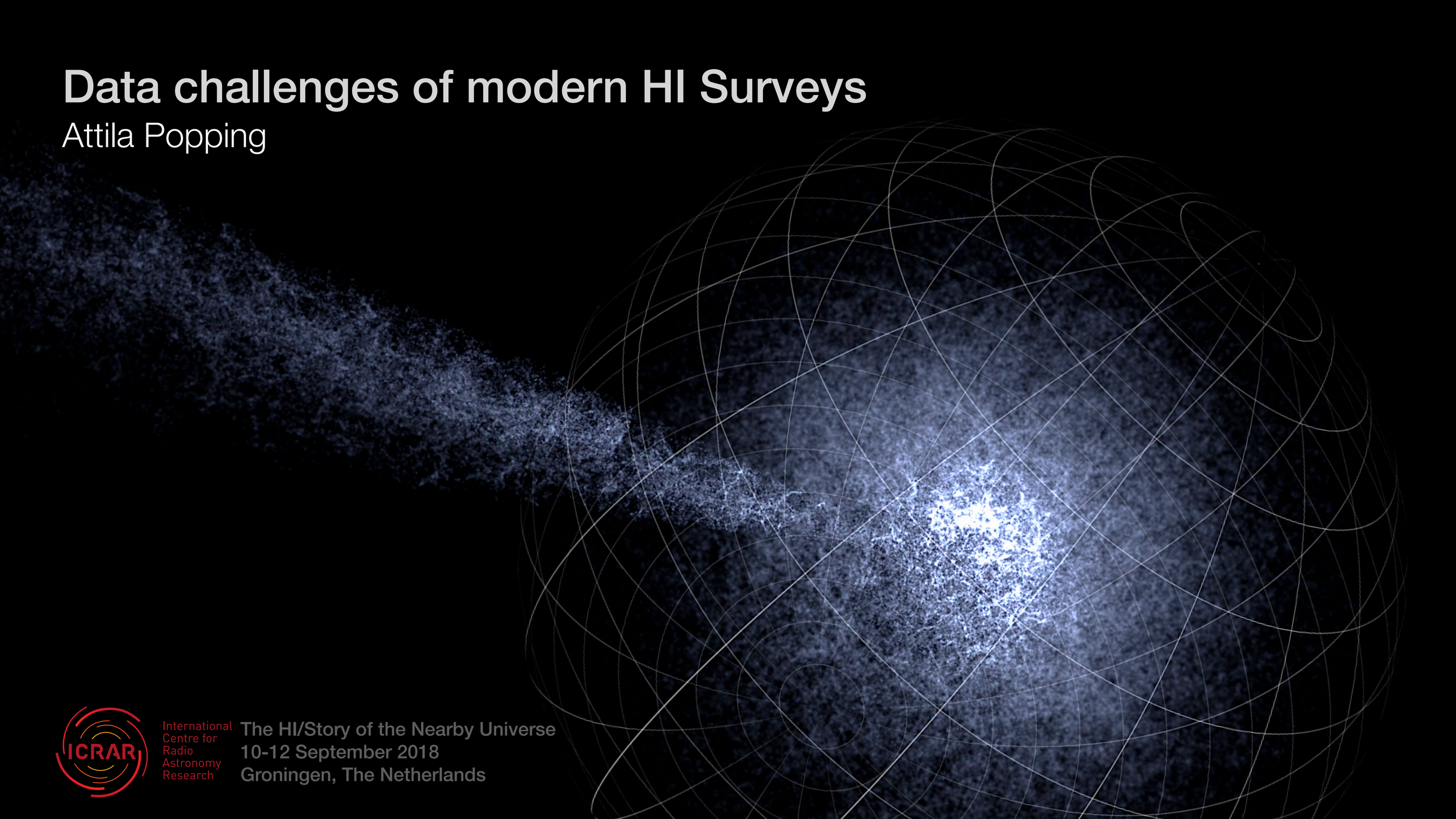
Data challenges of modern HI Surveys

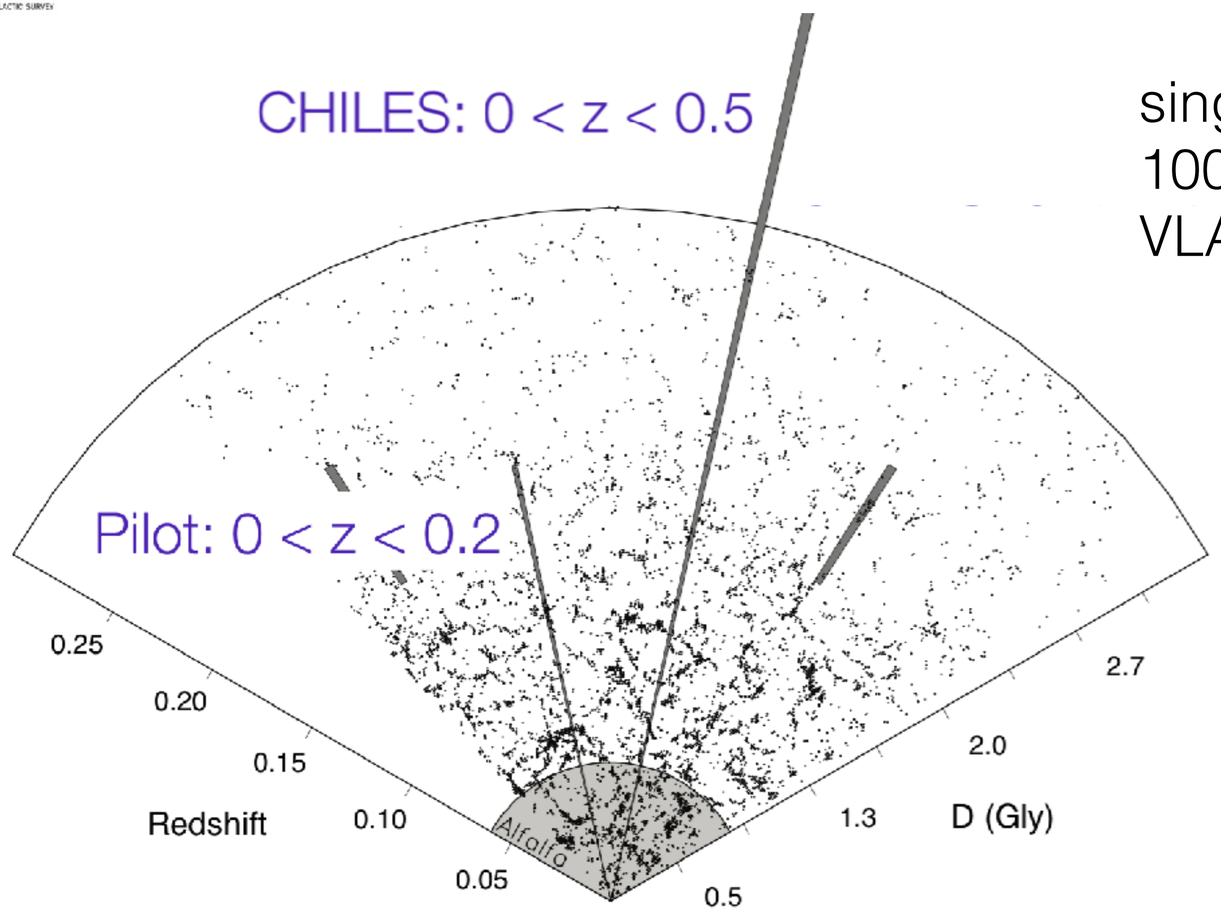
Attila Popping



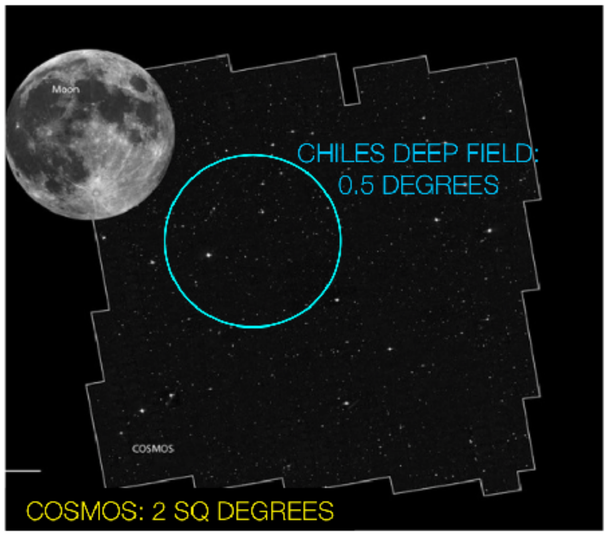
International Centre for Radio Astronomy Research

The HI/Story of the Nearby Universe
10-12 September 2018
Groningen, The Netherlands

COSMOS HI Large Extragalactic Survey



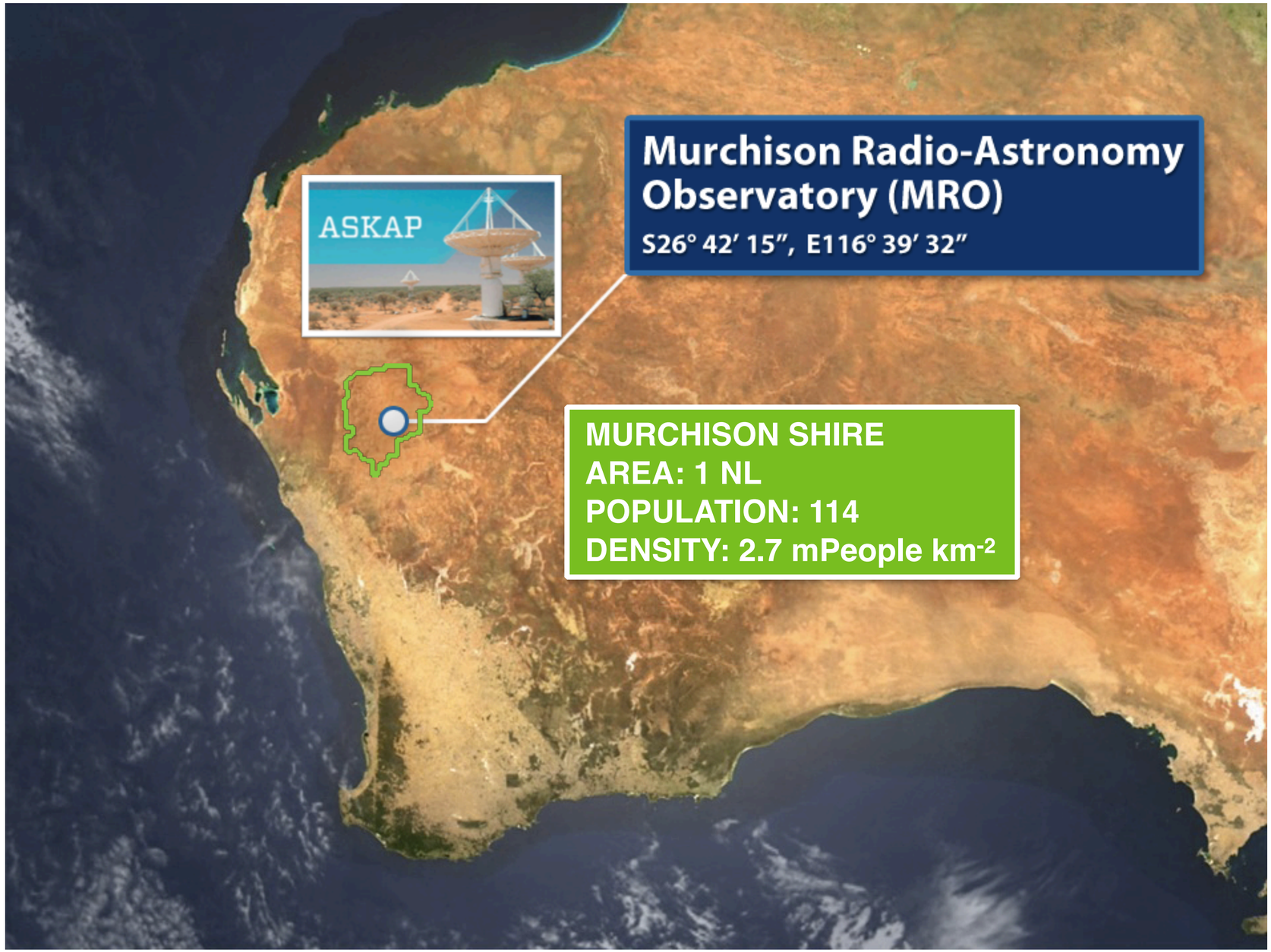
single pointing in COSMOS
1000 hours integration
VLA B-configuration



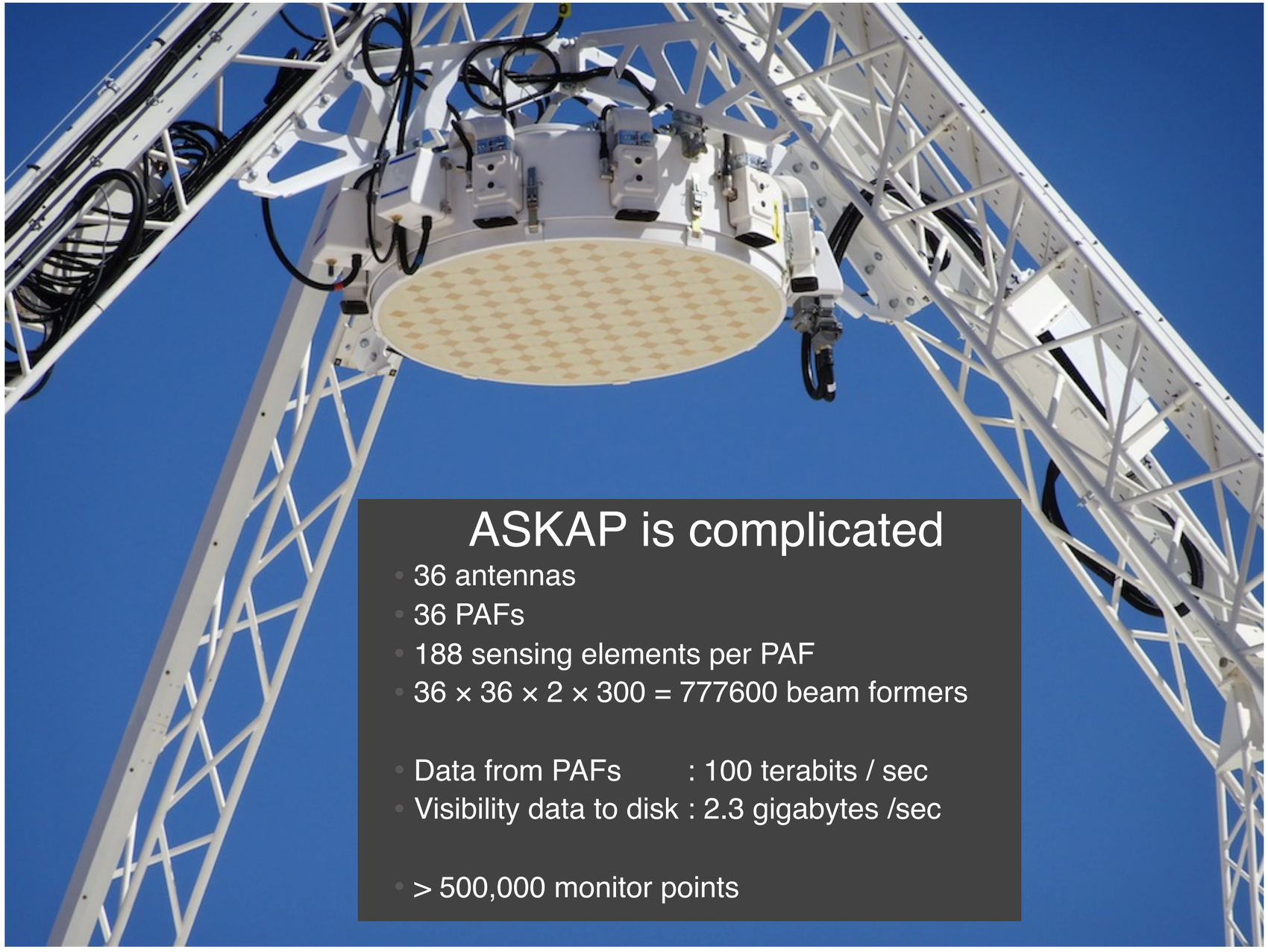
CHILES

	OLD	PILOT	NEW
Bandwidth (MHz)	6.25	240	480
Channels	31	16384	30720
Velocity resolution (km/s)	40	3.5	3.5
Instantaneous z coverage	0<z<0.004	0<z<0.193	0<z<0.5

ASKAP


ASKAP
Murchison Radio-Astronomy Observatory (MRO)
S26° 42′ 15″, E116° 39′ 32″
MURCHISON SHIRE
AREA: 1 NL
POPULATION: 114
DENSITY: 2.7 mPeople km-2

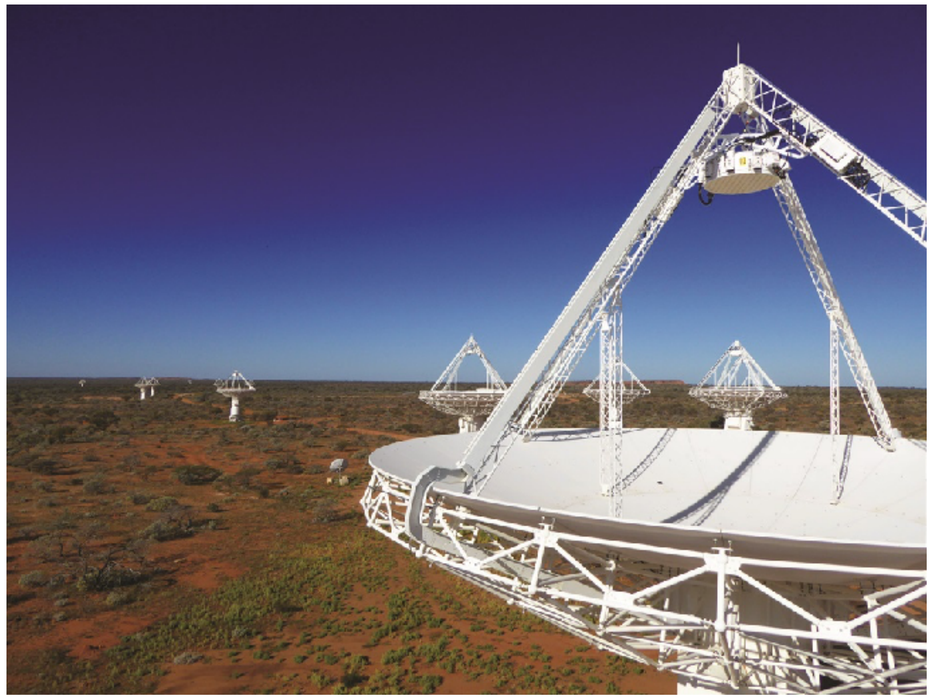
ASKAP


ASKAP is complicated
36 antennas
36 PAFs
188 sensing elements per PAF
36 × 36 × 2 × 300 = 777600 beam formers
Data from PAFs : 100 terabits / sec
Visibility data to disk : 2.3 gigabytes /sec
> 500,000 monitor points

Are we in a different Era?

People have been doing (HI) Surveys for decades.

- Things were new at the time
- Computers were not up to spec
- Working with a new instrument is always challenging

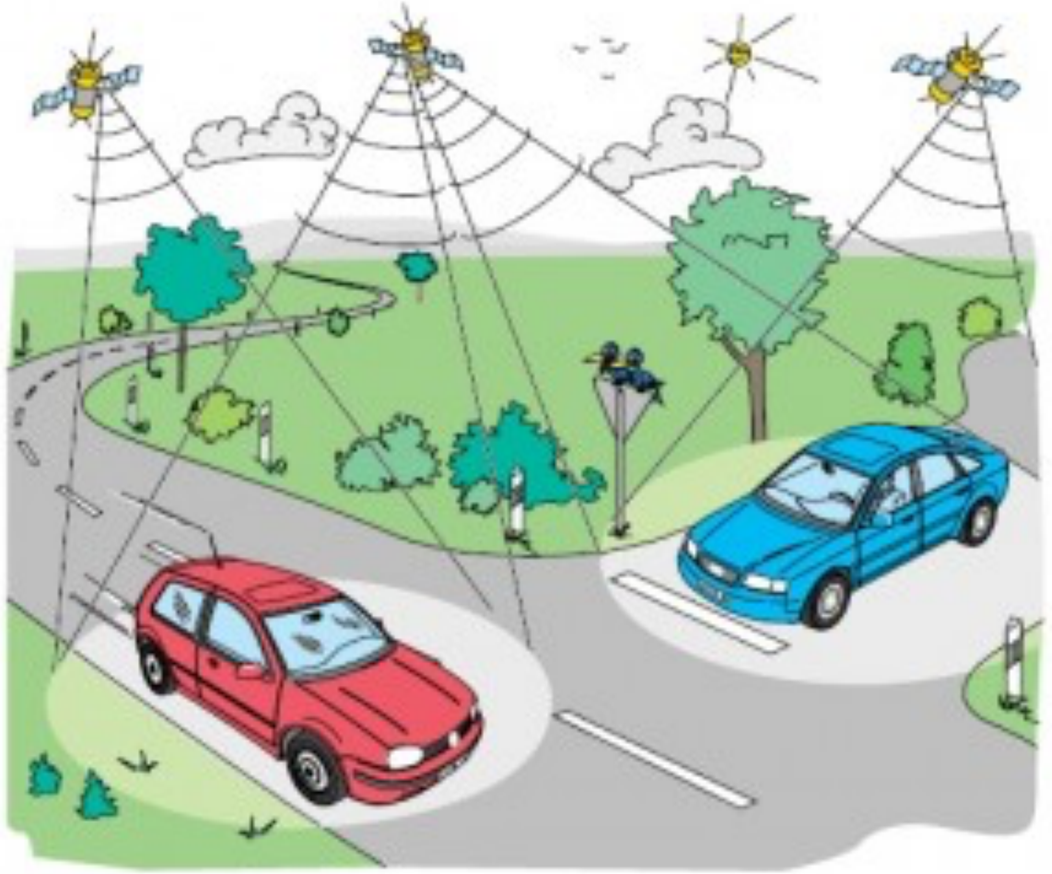


Are the challenges of today bigger ?

The challenges of today are bigger !

- RFI Enviroment is much more challenging
- Going from small (~10MHz) to large (>300MHz) bandwidth adds many complications
- Data volumes are unprecedented
- Data processing requirements are unprecedented
- Workload is larger than humanly possible
- We have more stringent requirements as new surveys go larger/wider/deeper/higher
- We go from ~1 instrument/survey per decade to N where $N > 7$

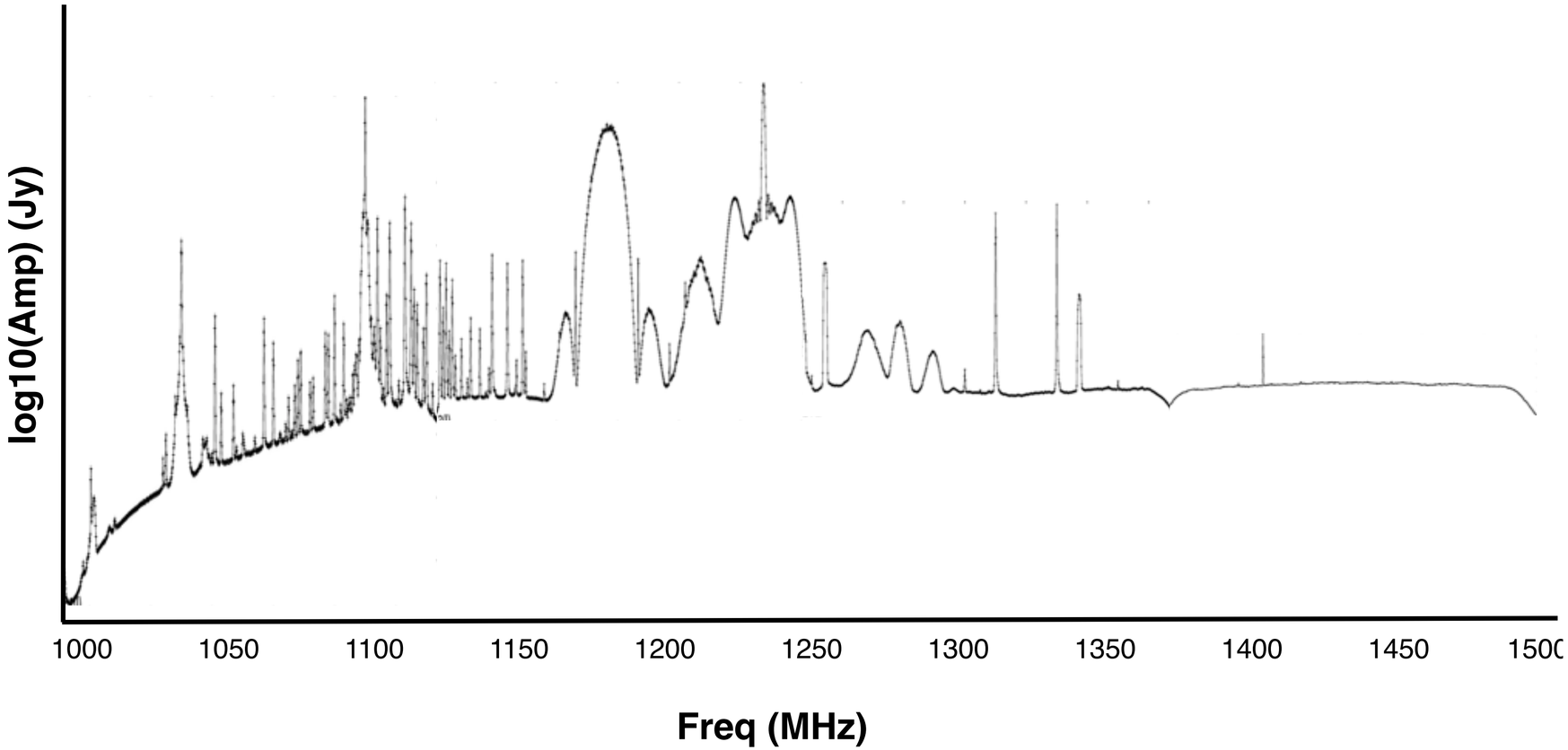
Challenge: Radio Frequency Interference



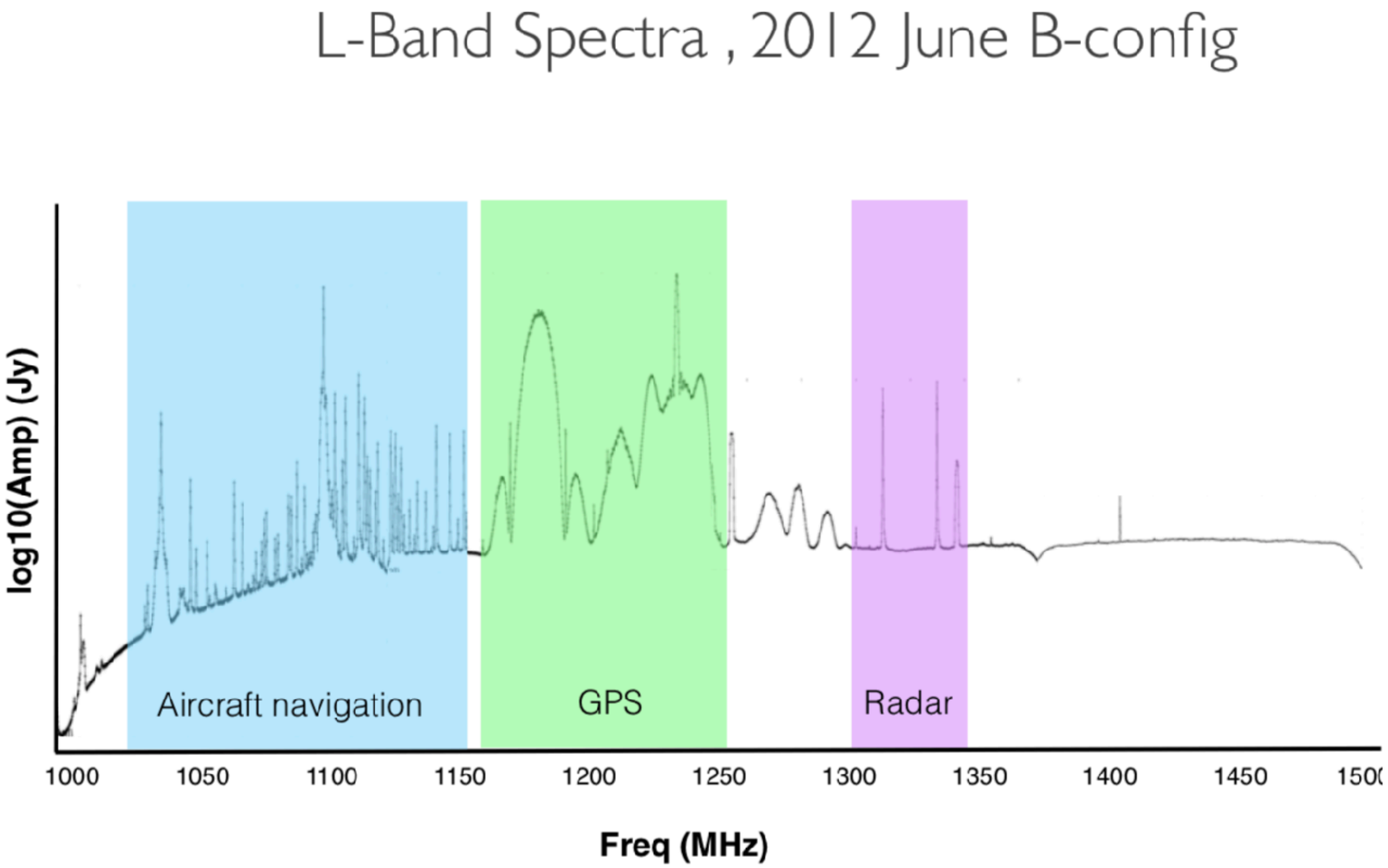
RFI environment:
In the Netherlands from the occasional tv tower to ~17 million people with ~17 million mobile phones, being stuck in electric cars, while getting information from 7 positioning satellites simultaneously.

Challenge: Radio Frequency Interference

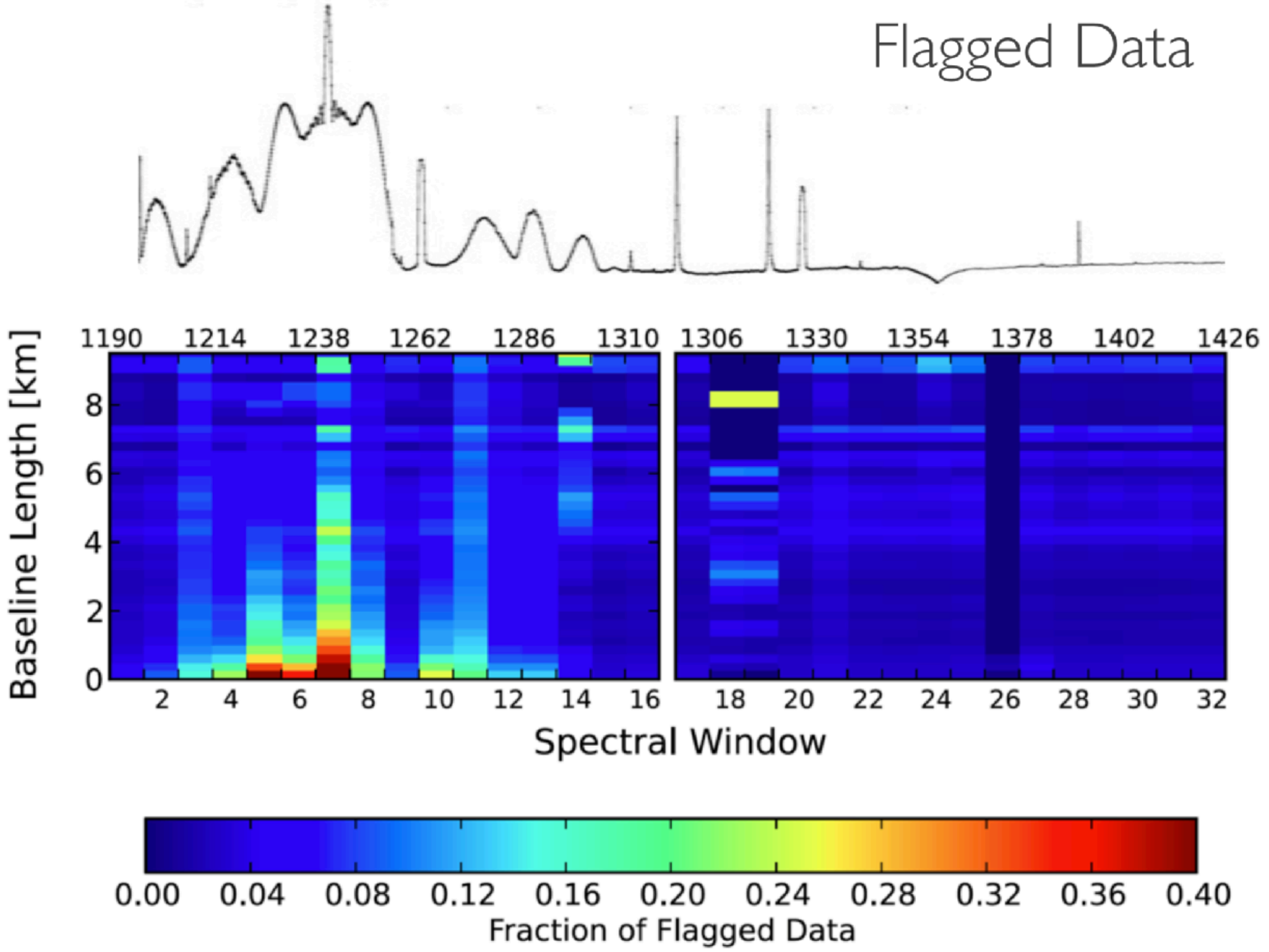
L-Band Spectra , 2012 June B-config



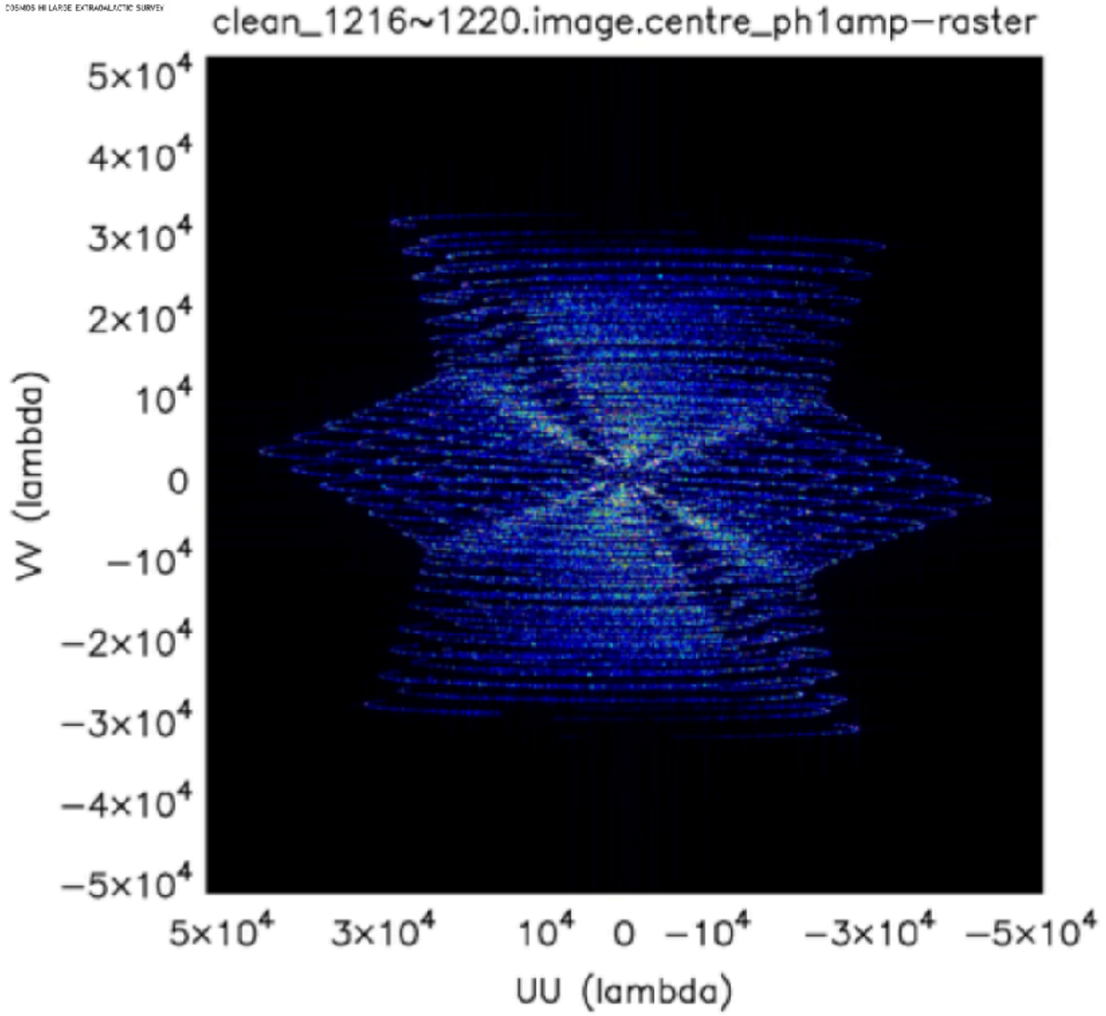
Challenge: Radio Frequency Interference

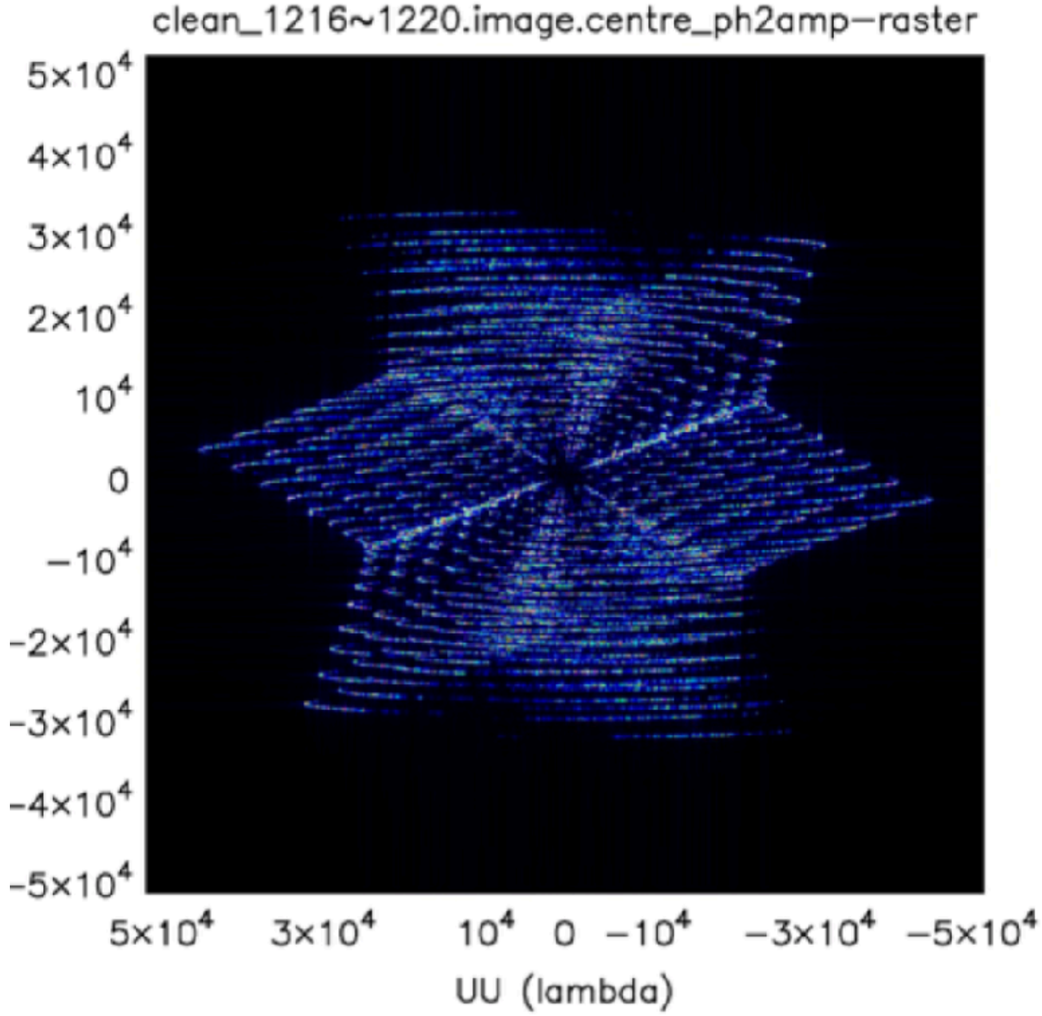


Challenge: Radio Frequency Interference

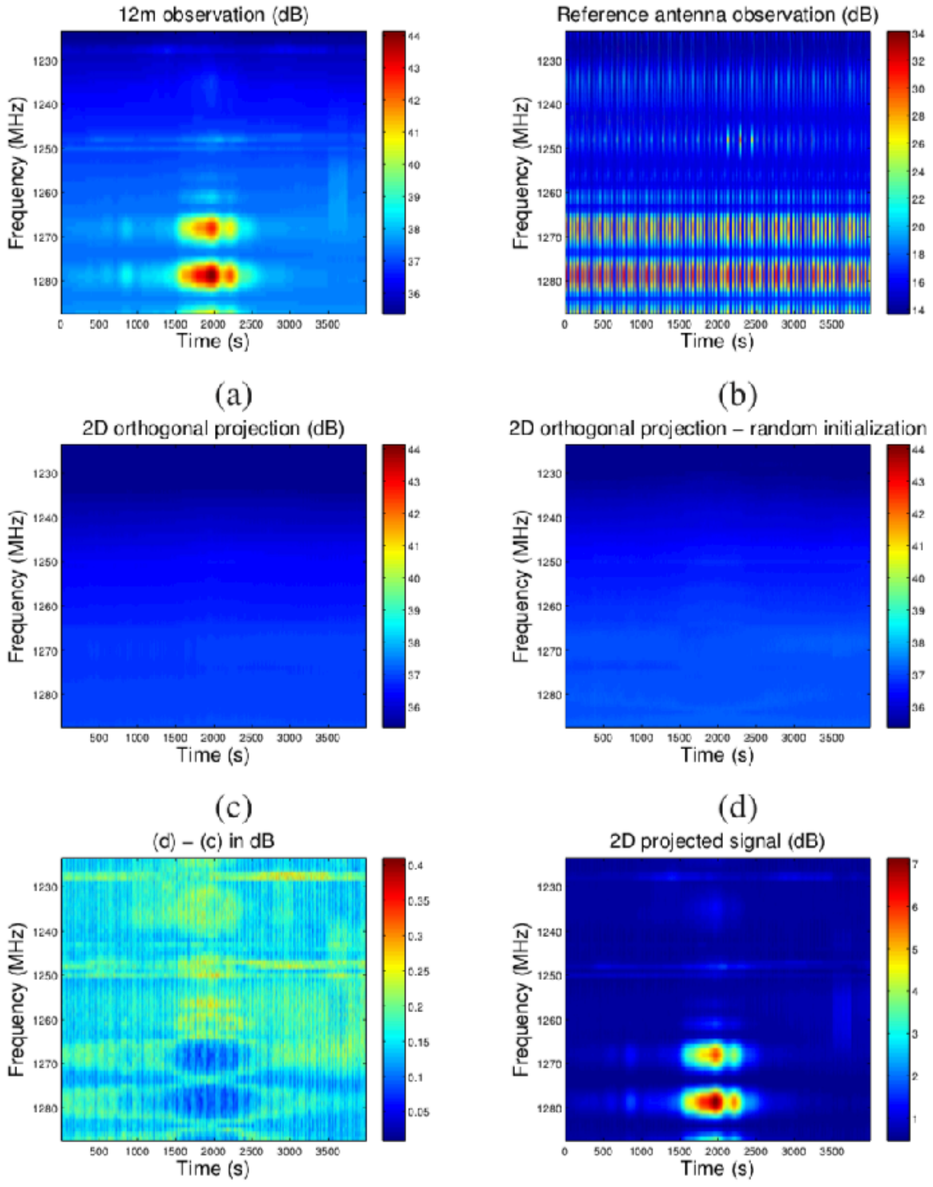


It is hard to find good flagging parameters





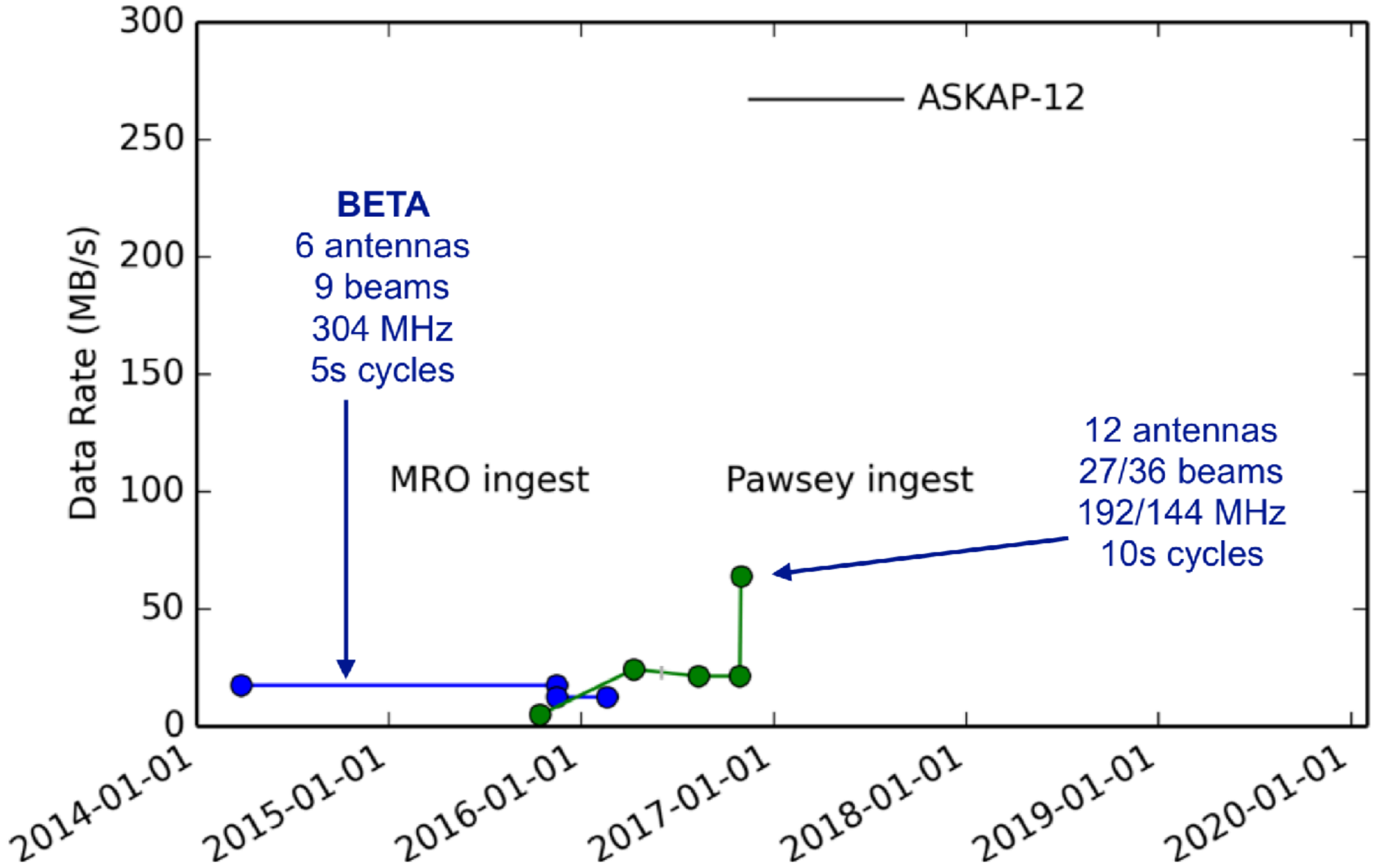
Reference antenna based RFI tracking



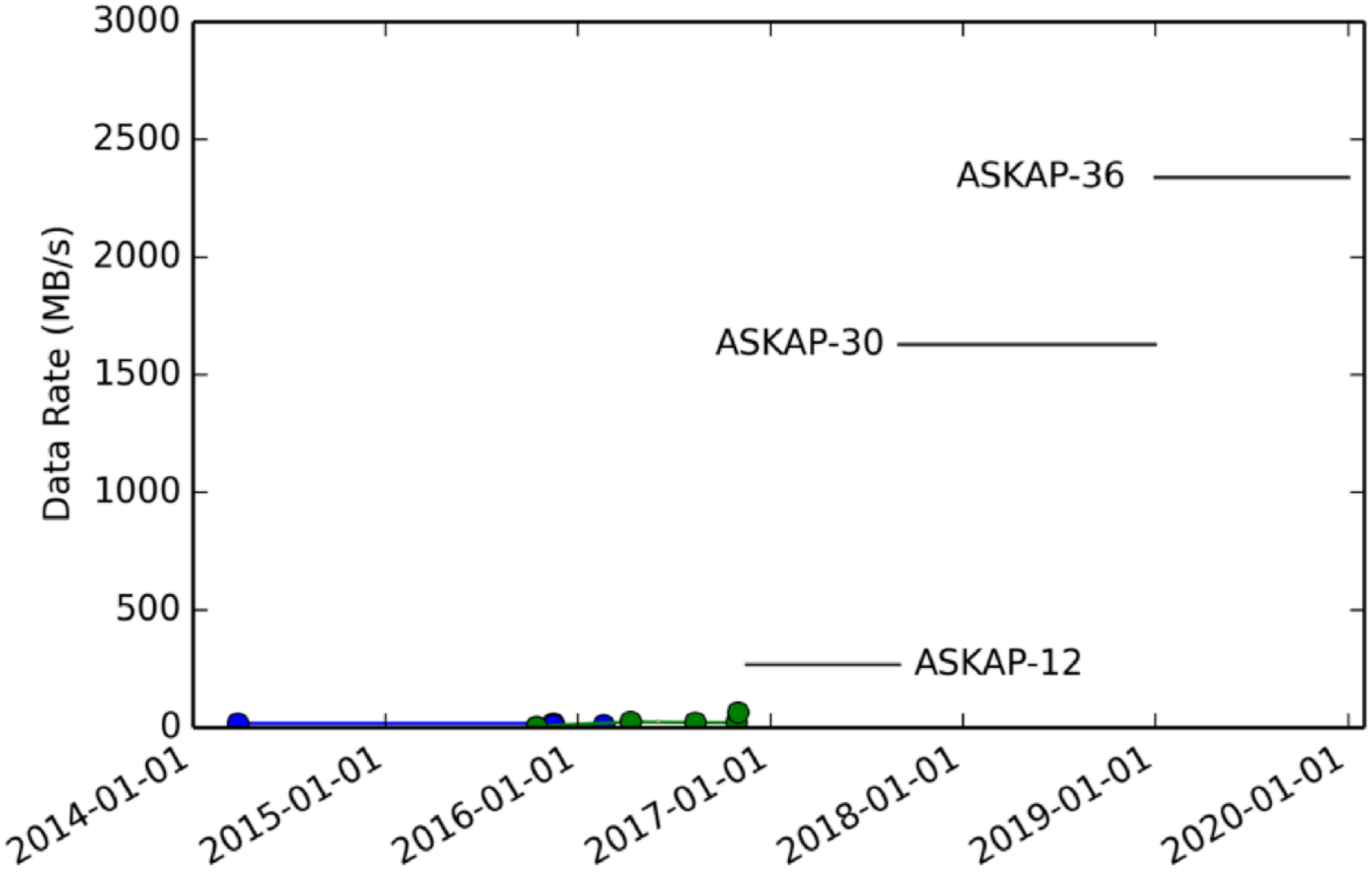
- Many instruments are limited by the GPS bands.
- RFI is the hot potato of many surveys
- Current methods are masking effected data rather than taking out the bad data.
- Use a reference antenna to track the RFI source and subtract the power from observations
- Potentially powerful method, especially for a telescope with a phased array feed
- Needs a serious commitment and hence investment to make it work

G. Hellbourg et al.

Data volumes are uprecedented

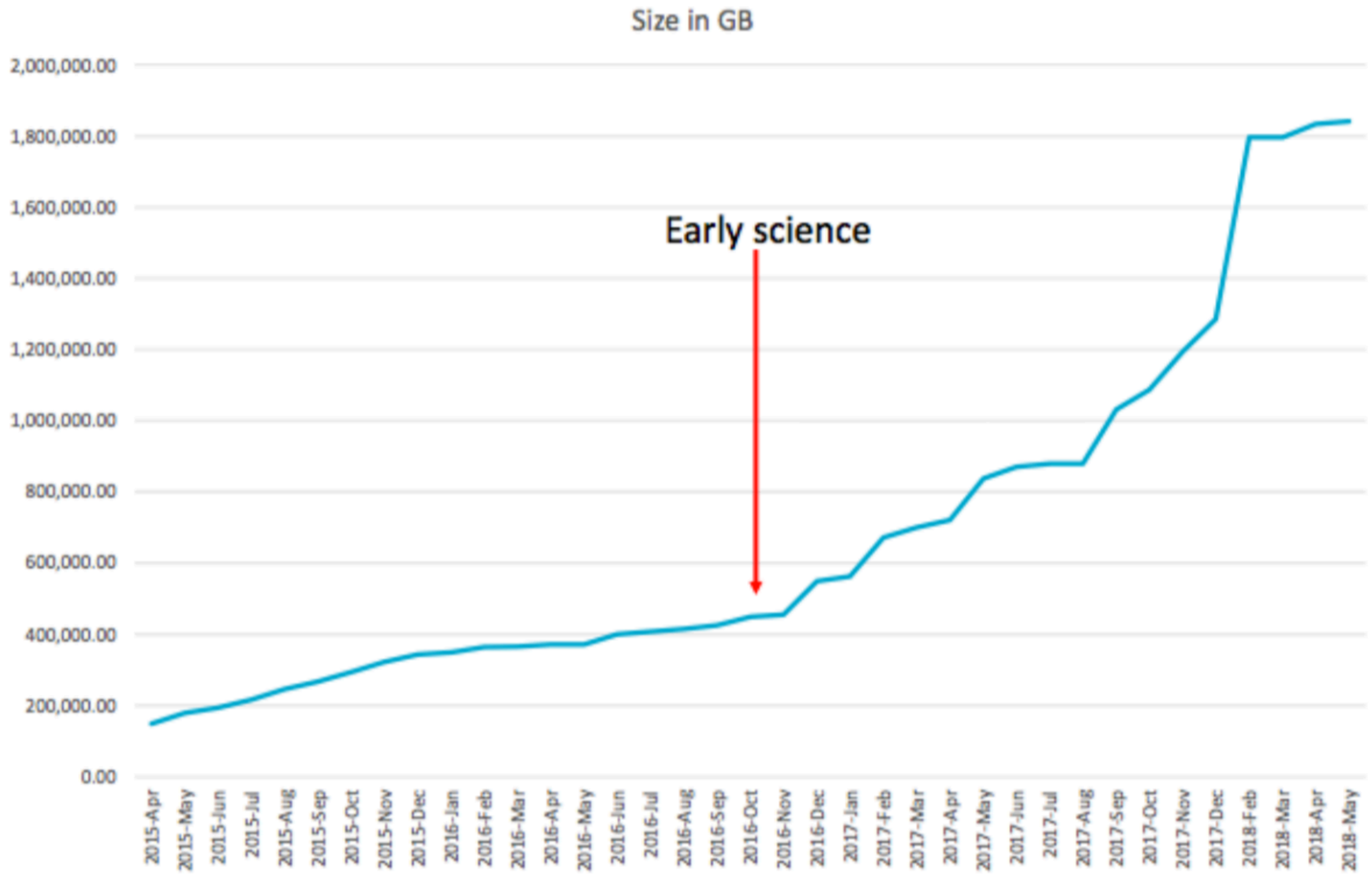


Data volumes are uprecedented



Data volumes are uprecedented

Commissioning archive use



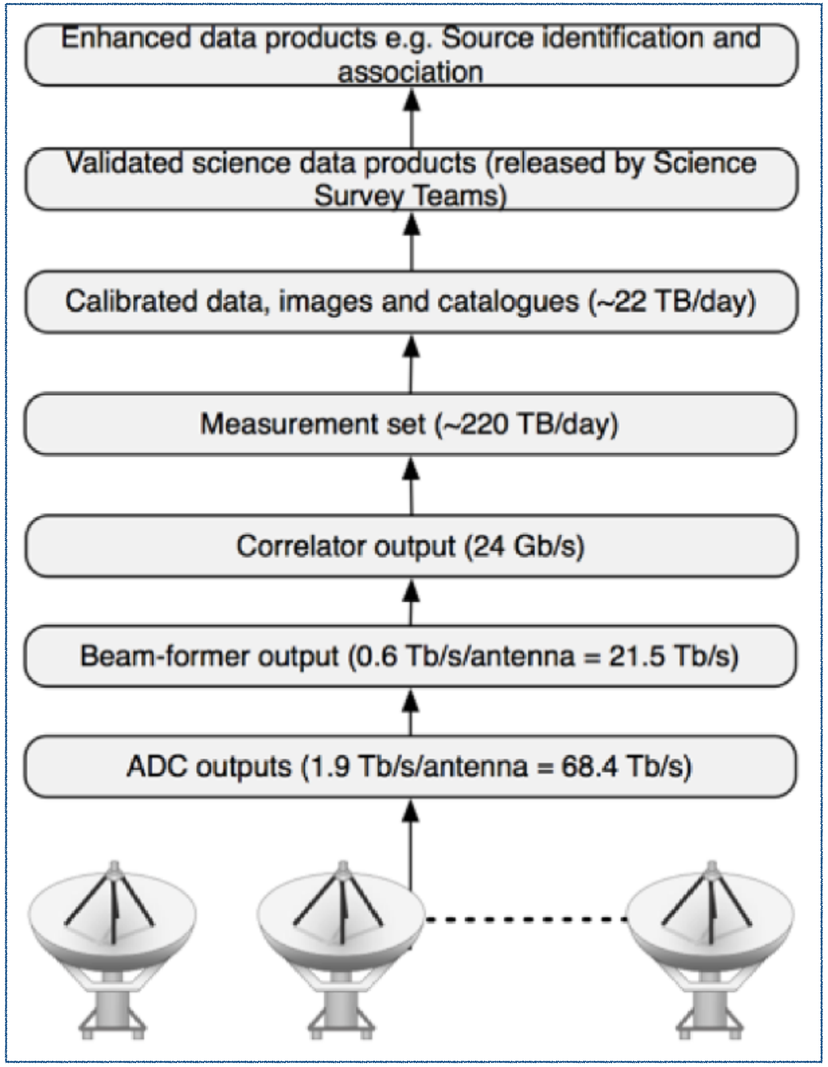
A. Hotan

No more data storage

- ASKAP is the first telescope that will not be able to hold on to the raw visibilities
- We will require real time processing
- We will require very robust pipelines
- We need high quality pipelines that we trust

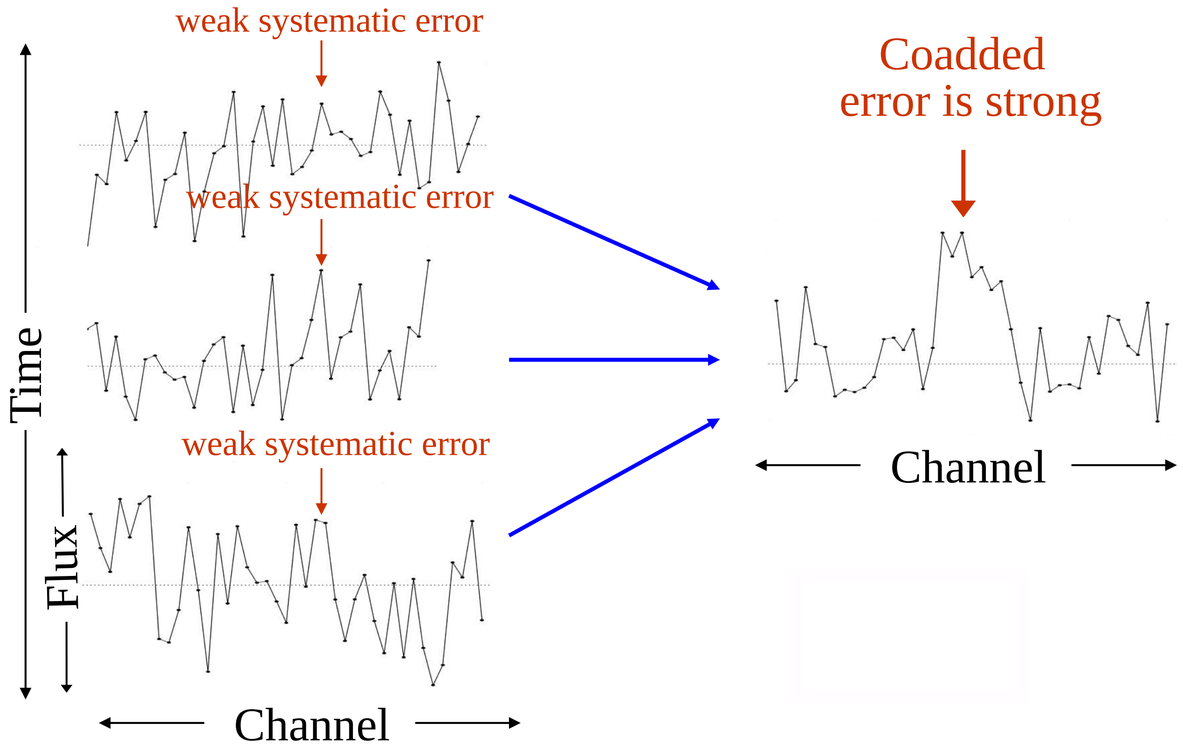
Will we ever reach a state where single-pass processing will be sufficient ?

ASKAP data rates



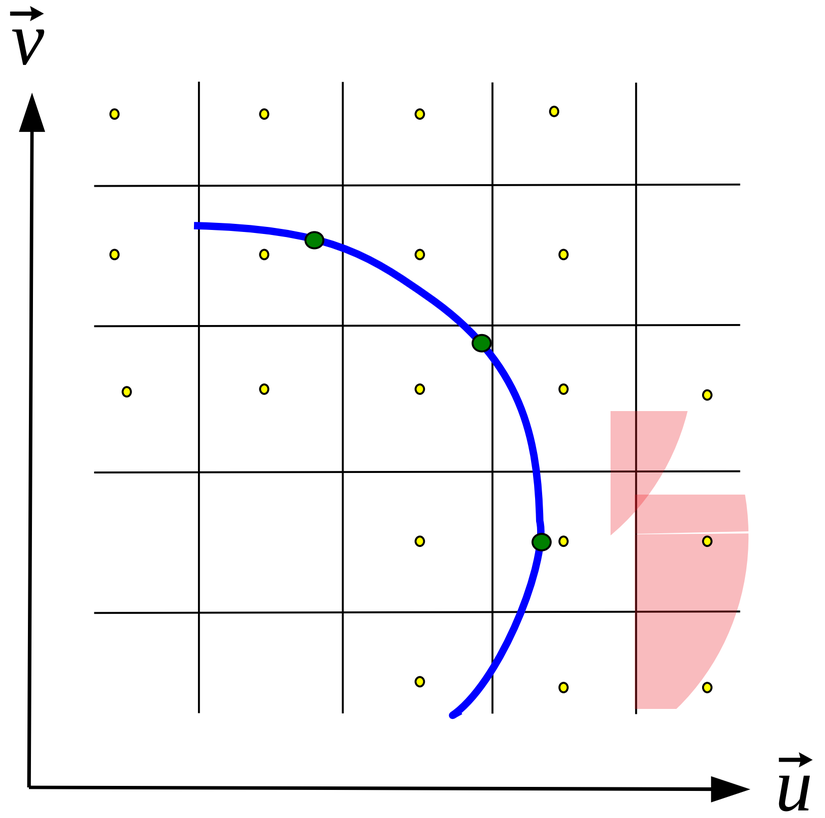
Visibility storage using UV-grids

- By performing one-pass data processing we lose the calibration versatility
- Deep surveys are extremely sensitive to small systematic uncentairtinies
  - Low-level RFI and weak continuum sources
  - Poorly removed sidelobes and other systematic calibration errors



UV grid storage

- There are ways to compress visibility for storage, e.g average by time, baseline. . . etc. (e.g Wijnholds, S. J. et al 2018, arxiv: 1802.09321)
- However visibilities evaluated to a uniform grid for FFT
- Compressed uv data is naturally available because grids are sparse
- By storing some cells several times, we can hold on to baseline information
- Time is present on the grid



uv gridding using convolutional resampling

DINGO storage requirements

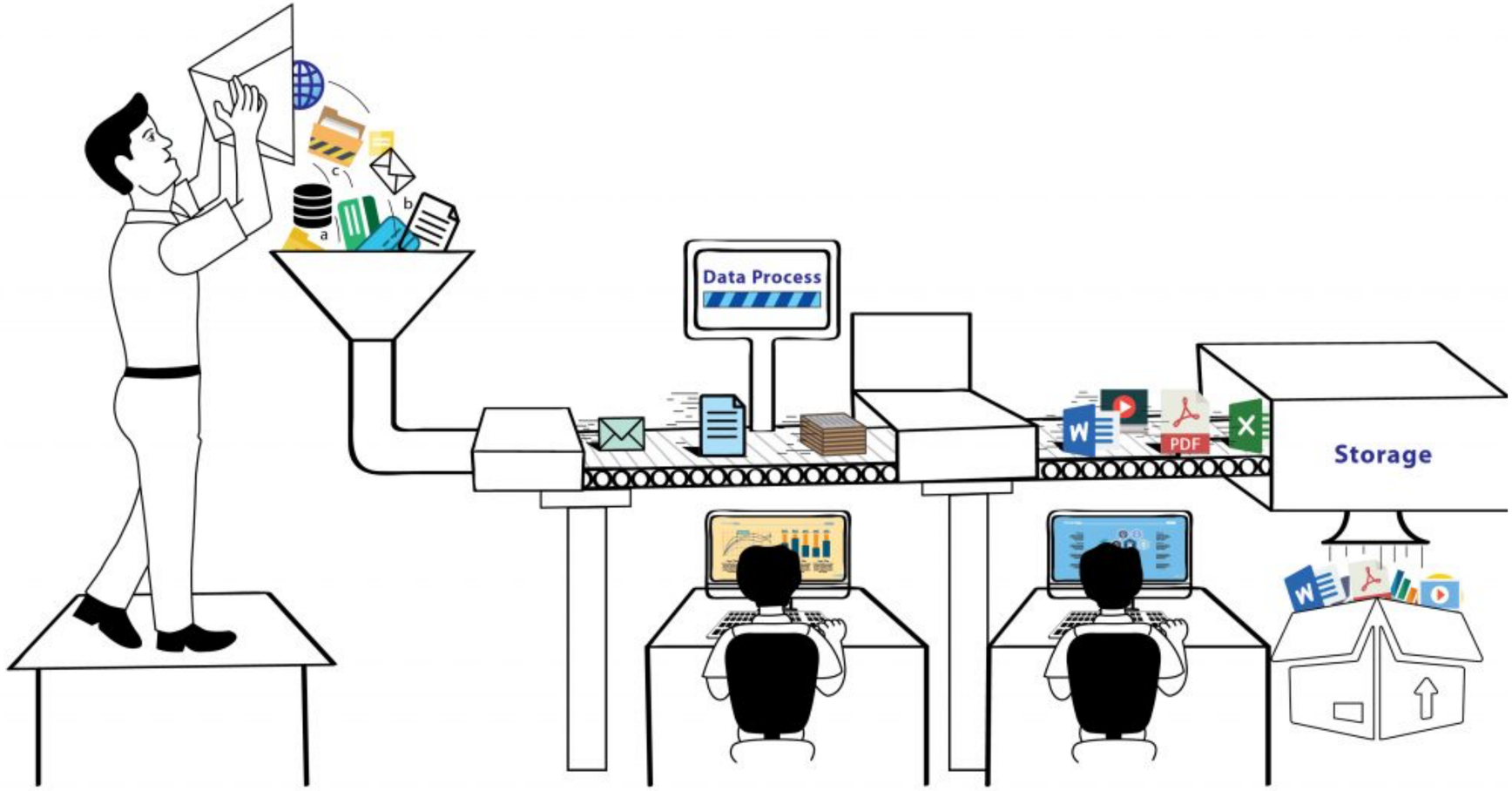
- Holding on to the >2km baselines don't increase data rate dramatically on the grids ⇒ method scales well with additional long baselines
- However mosaicking reduces the required storage the most, we lose a lot of – even the beam – information

	ASKAP – core (2km max baseline)		**ASKAP – full (6km max baseline)**	
	$T_{obs} = 0.5$ h	$T_{obs} = 8$ h	$T_{obs} = 0.5$ h	$T_{obs} = 8$ h
Quick & Dirty	64.14 PB	**4.01 PB**	577.26 PB	**36.08 PB**
Precise	81.85 PB	**5.12 PB**	736.69 PB	**46.04 PB**
Imaging pipeline	434.52 PB	**27.16 PB**	1.84 EB	**117.87 PB**
Mosaicking	34.58 PB	**2.16 PB**	311.25 PB	**19.45 PB**
Sparse uv grid	10.00 PB	**4.54 PB**	17.36 PB	**9.33 PB**
#1, 8h mask	69.75 PB	**4.36 PB**	143.24 PB	**8.95 PB**
#16, 0.5h mask	8.63 PB	**4.36 PB**	15.15 PB	**8.95 PB**
On the fly	7.97 PB	**3.63 PB**	13.81 PB	**7.46 PB**

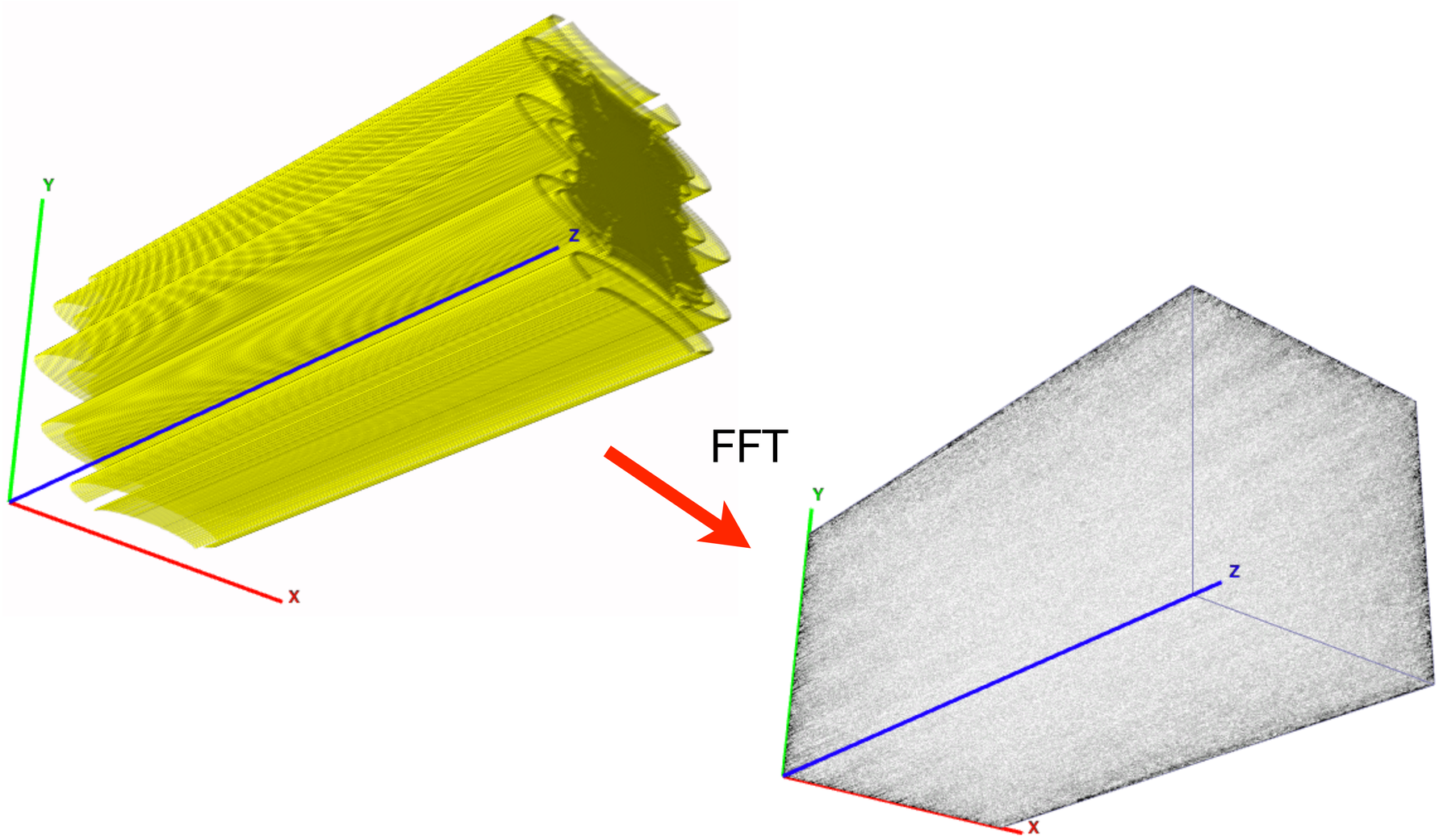
Would enable new processing methodologies with improved: RFI mitigation, continuum subtraction, cleaning, resolution

Data processing requirements

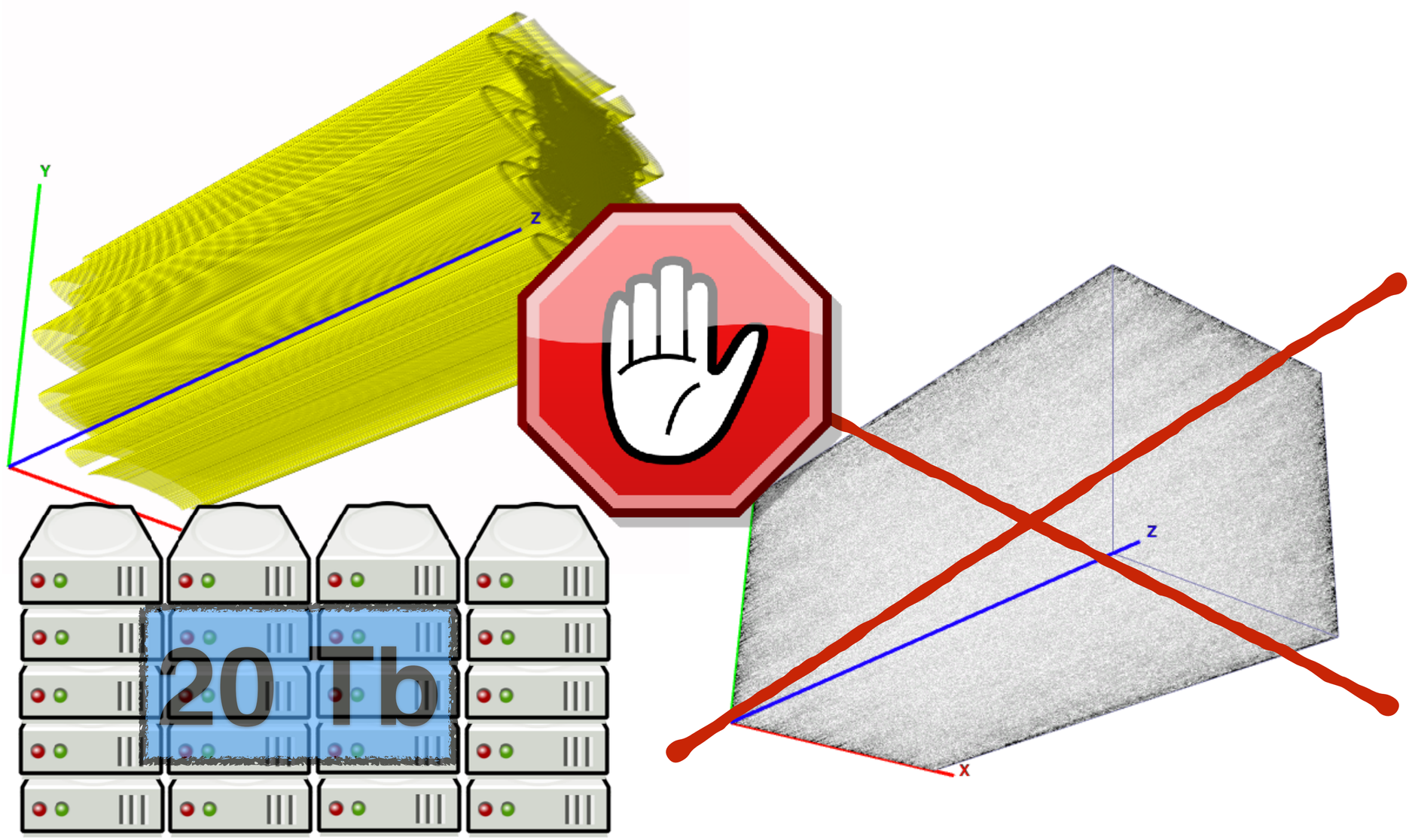
With the increased amount of data, the processing is also more demanding



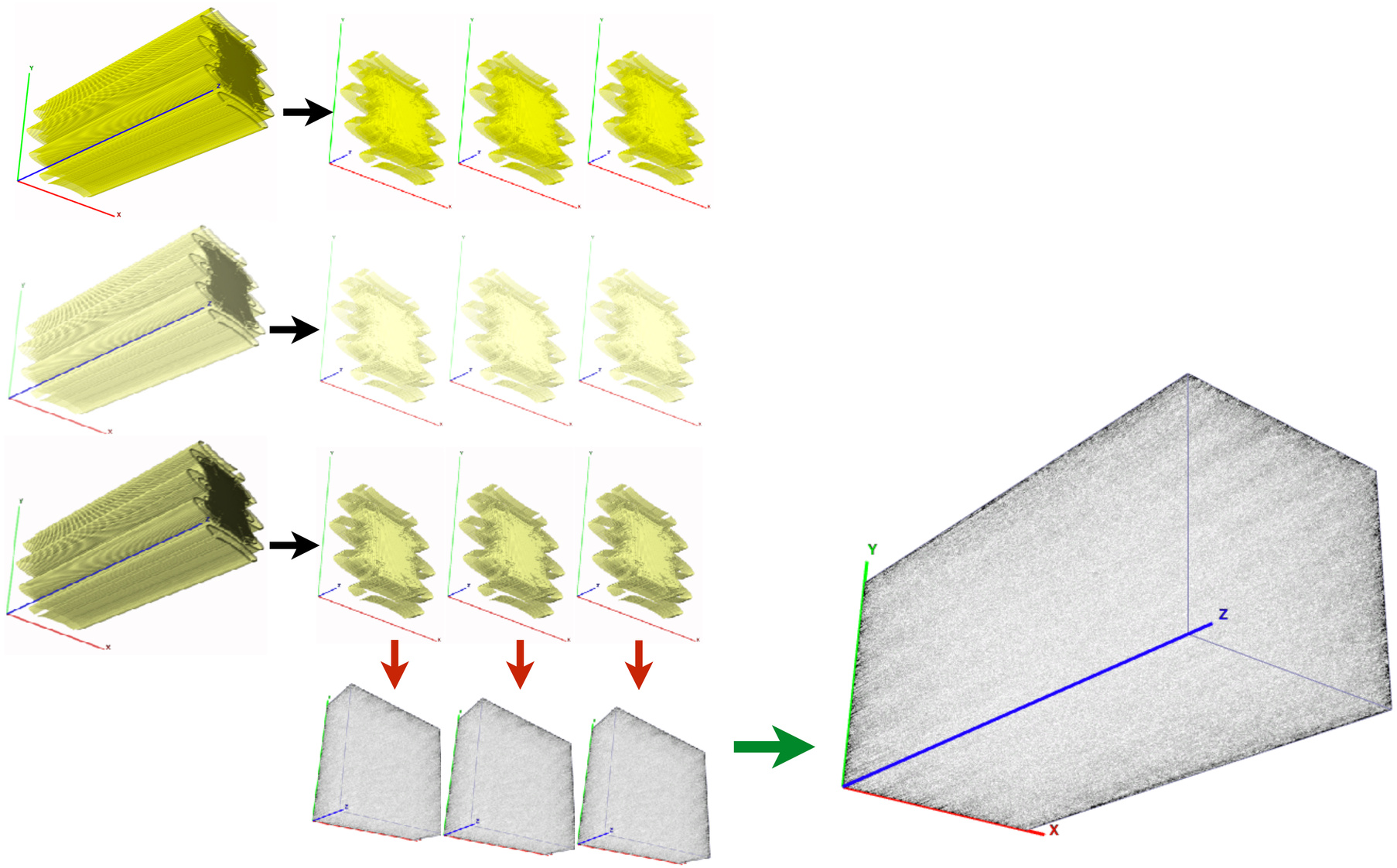
Imaging



Imaging



Imaging



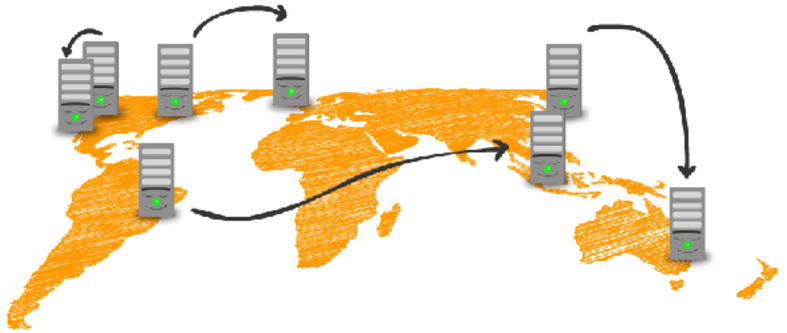
Computing efforts

Conventional Cluster (pleiades) 5 nodes each node has 2x Intel Xeon X5650 2.66GHz CPUs (6 cores / 12 HTs) with 64-192 GB of RAM	Enough computing power, however it would take weeks
Super computer (MAGNUS) Cray XC40 - 24 cores per node 2.6GHz Intel Xeon E5-2690V3 64GB per Node 35,712 cores available 3PB of storage #58 in the world	
AWS Whatever we wanted r3.xlarge 16 cores 122GB Ram	amazon web services™

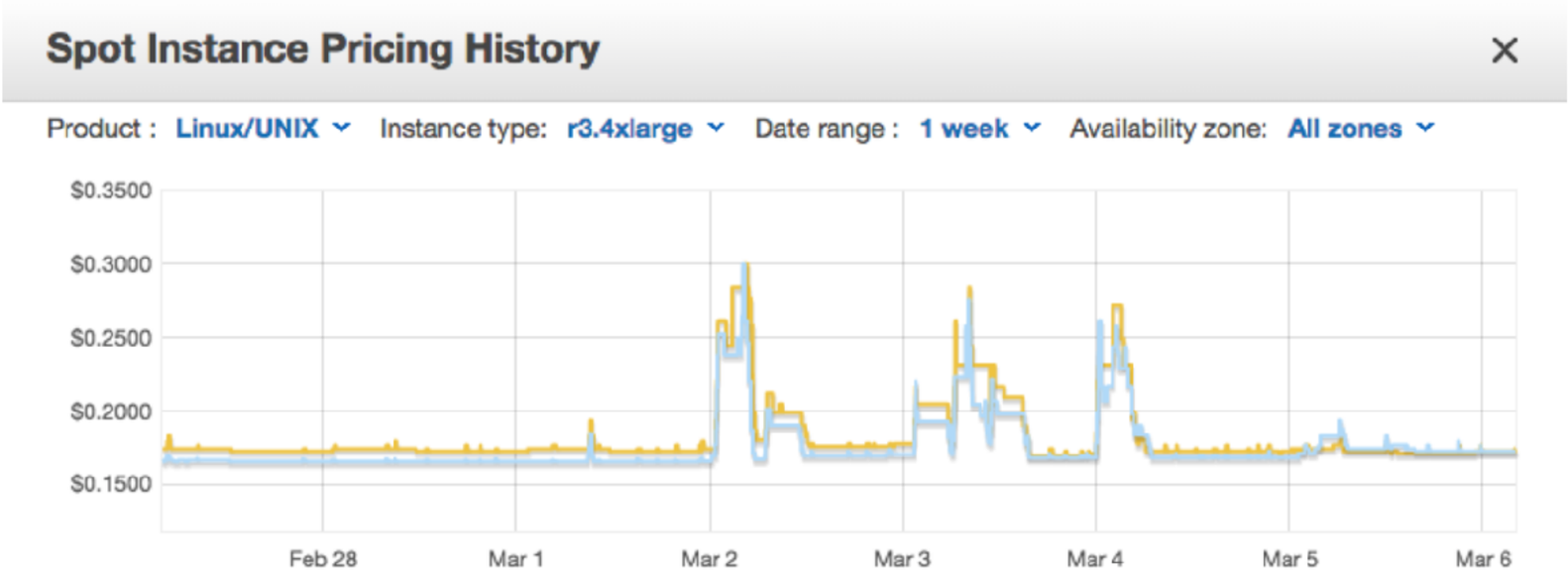
AWS pricing

amazon web services™

	On demand	Spot Price
r3.4xlarge	$1.68	$0.20
r3.2xlarge	$0.840	$0.09
m3.xlarge	$0.392	$0.04
m3.medium	$0.098	$0.01



Works!
costs so far : ~$2000

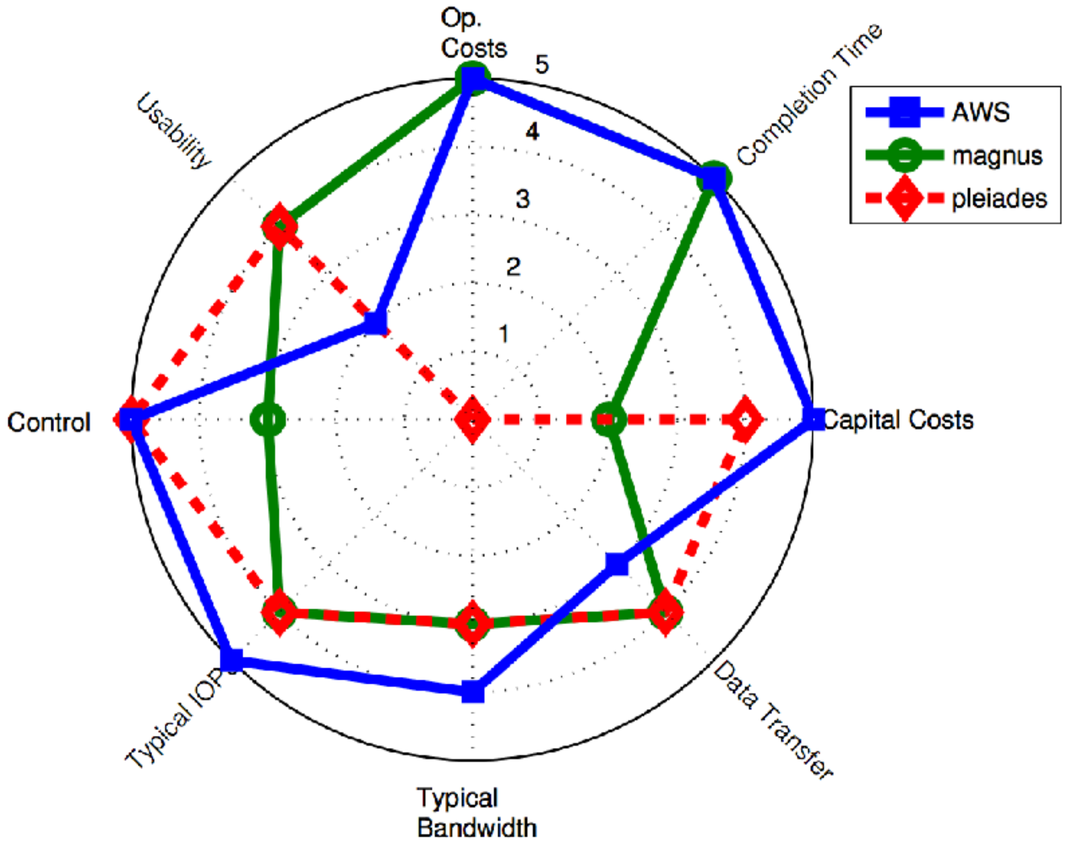


Computing efforts

	AWS	Magnus (HPC)	Pleiades
Completion Time	96hr	110hr	1,060hr (est)
Capital Costs	AUD$0	AUD$12,000,000	AUD$50,000
Operational Costs	AUD$2,000	AUD$3,240 (free)	-
Control	Root	Limited	Root
Usability	Complex	Good	Good

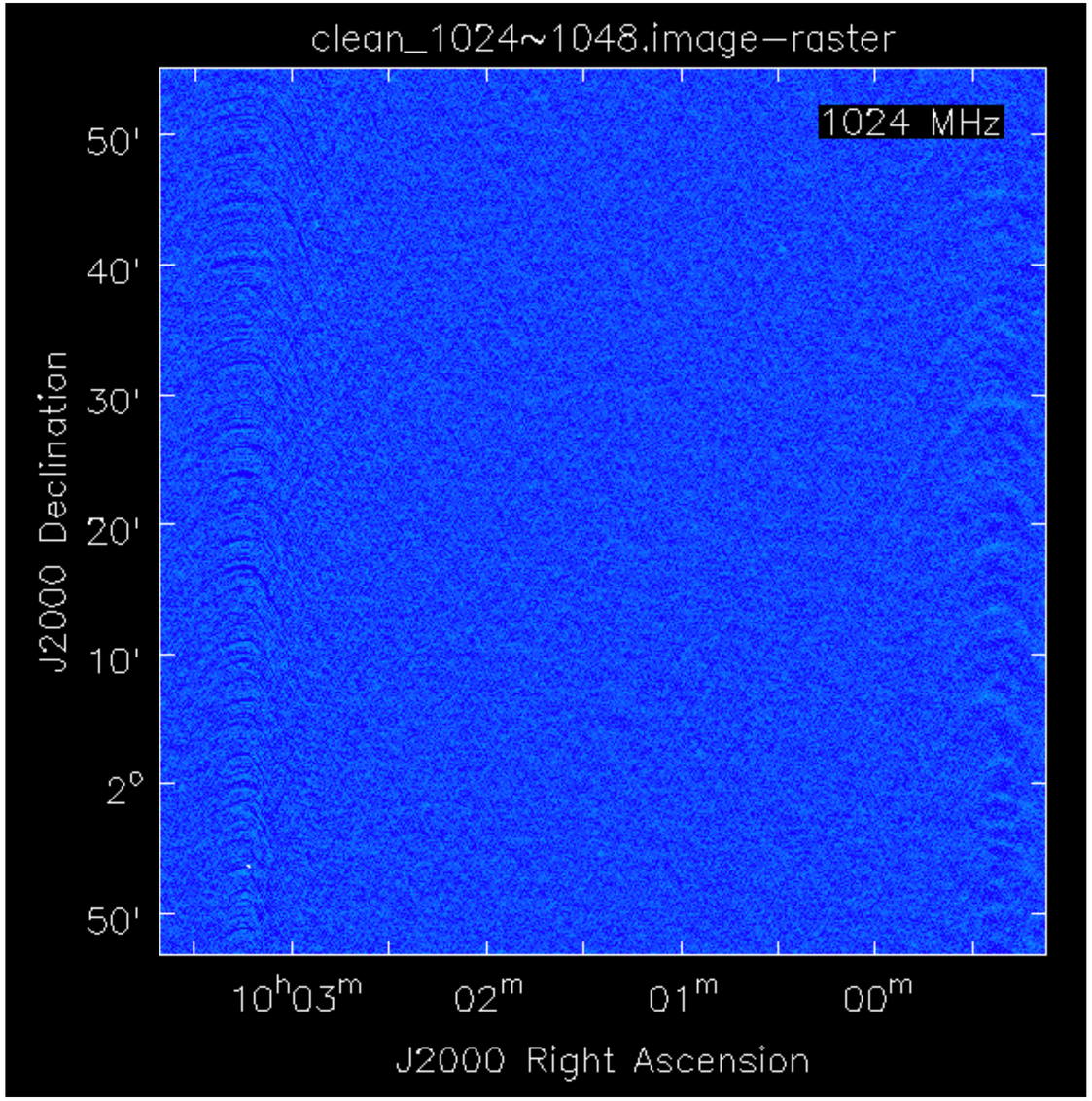
Compute conclusion

- Department Cluster
  - Not very satisfactory
- HPC
  - Very fast
  - You can't have root access
  - Installing new software is done by the admin's
  - In WA it is effectively free
- Cloud
  - You can do what you like (a good and a bad thing)
  - EBS volumes are slow
  - Directed attached SSDs are very quick
  - You pay for what you use… And if you forget to turn it off you are still paying



Cloud computing can provide an excellent alternative, however will require a change in mindset

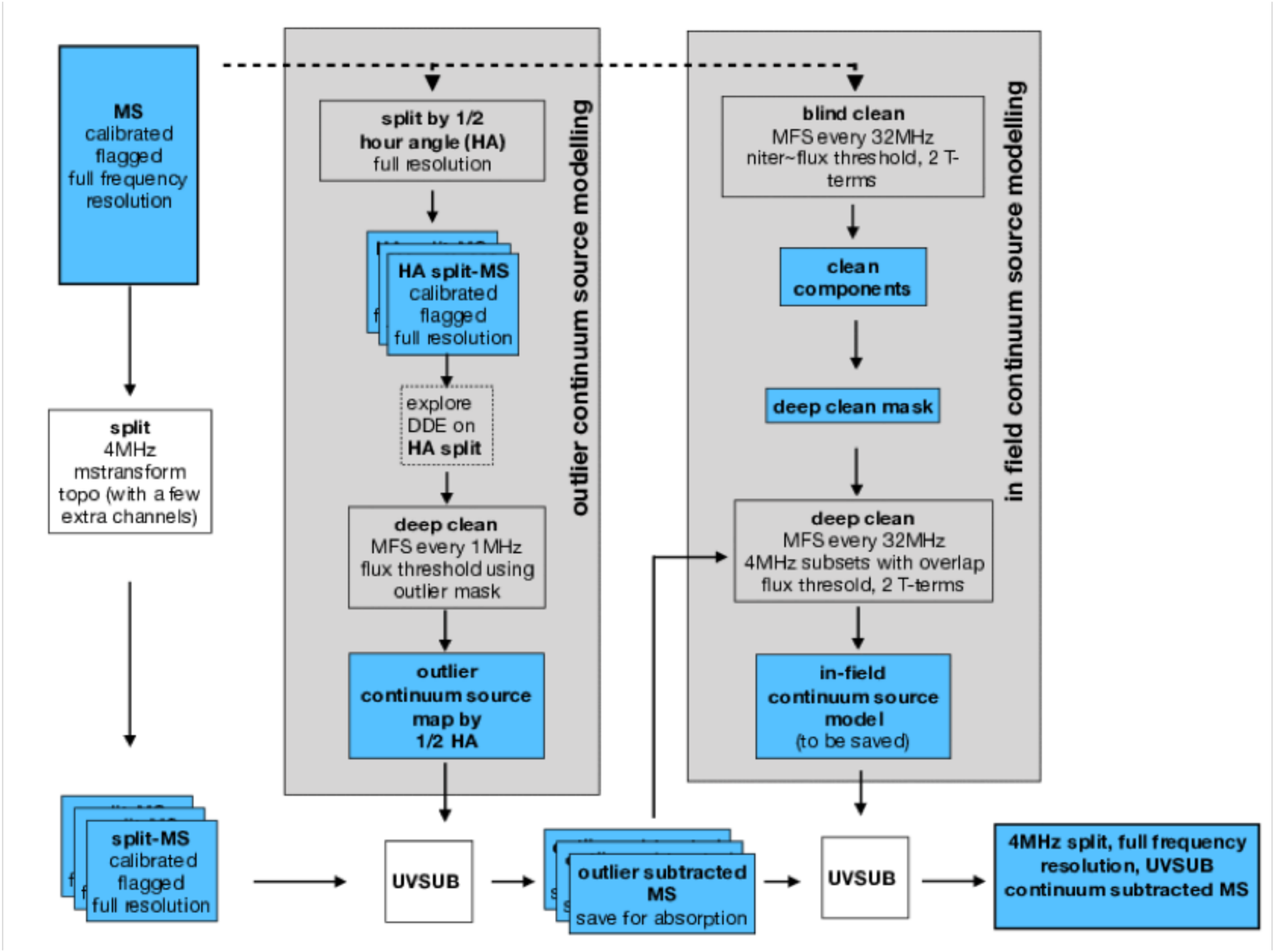
We have more stringent requirements



- Low level continuum artefacts after combining data
- Not visible in individual observations
- Need better continuum model and go back to raw data to subtract.

Note: we may have deleted the raw visibilities at this stage ….

Complicated workflows

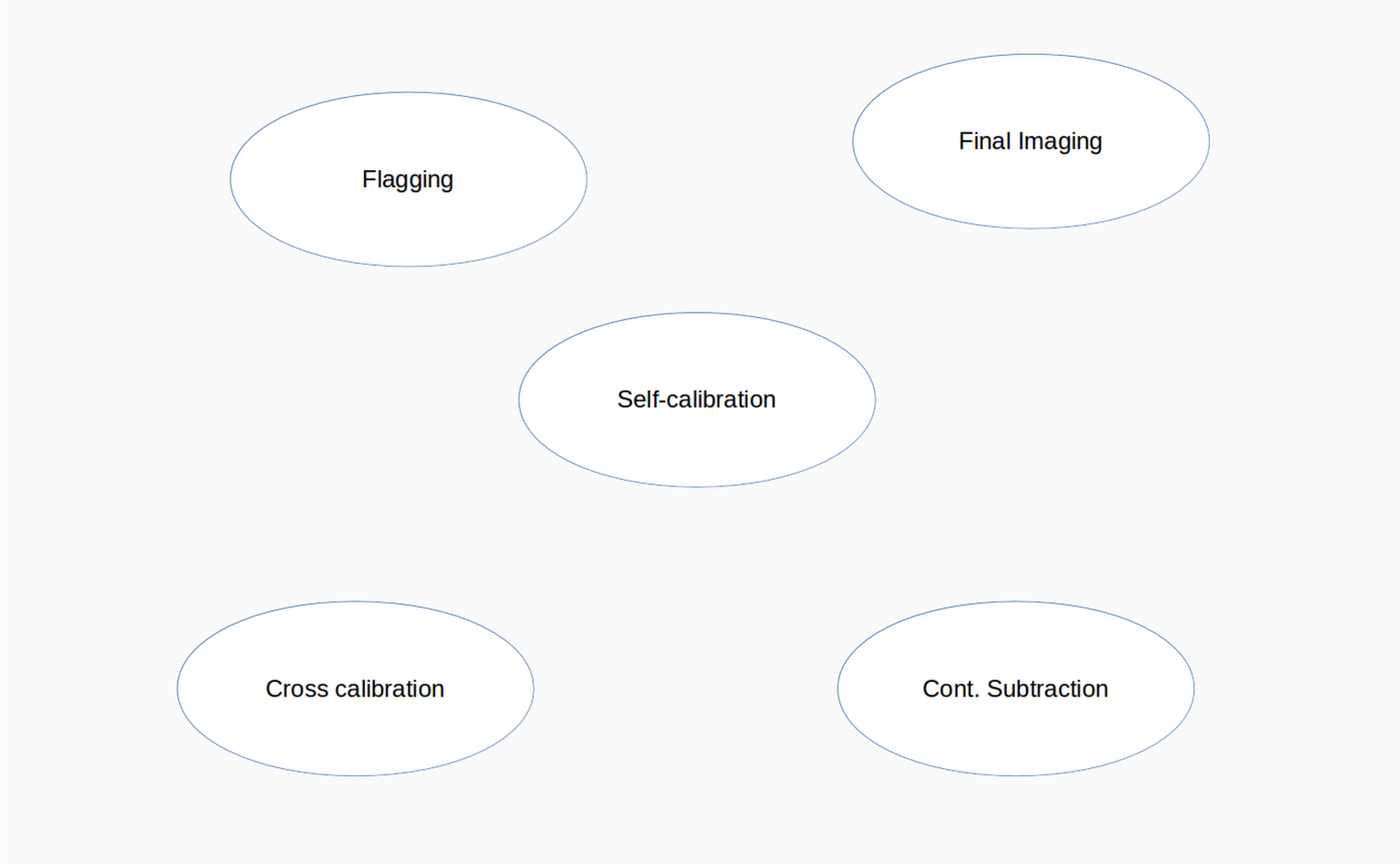


The ideal pipeline

Level 0	Level 1	Level 2
Reproducible results	Documented	Smart use of resources
Easy to build/install	Portable	Scalable
Easy to use	Good diagnostics	Cache intermediate results?
	Configurable	

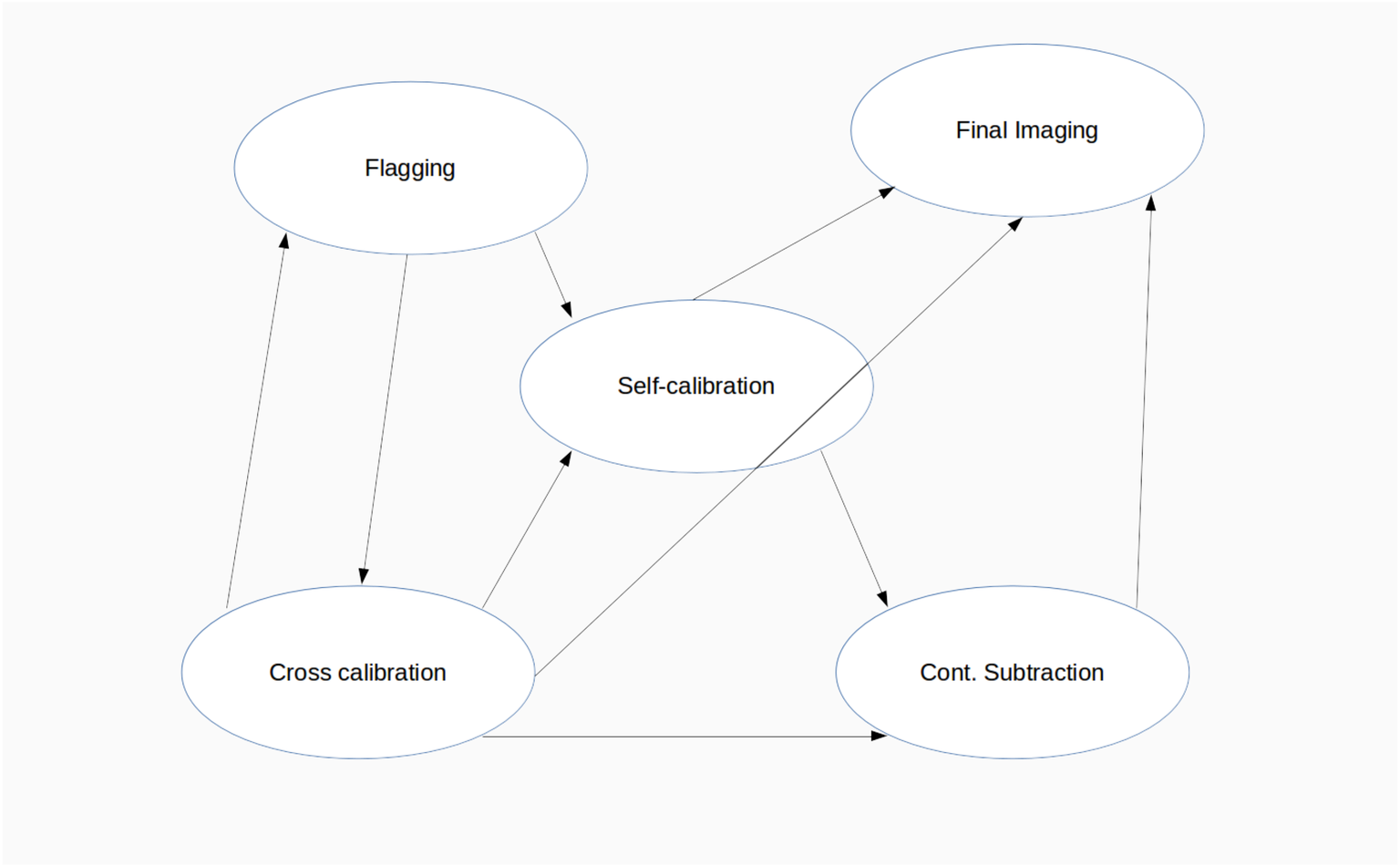
S. Makhathini

Typical data reduction



S. Makhathini

Typical data reduction



S. Makhathini

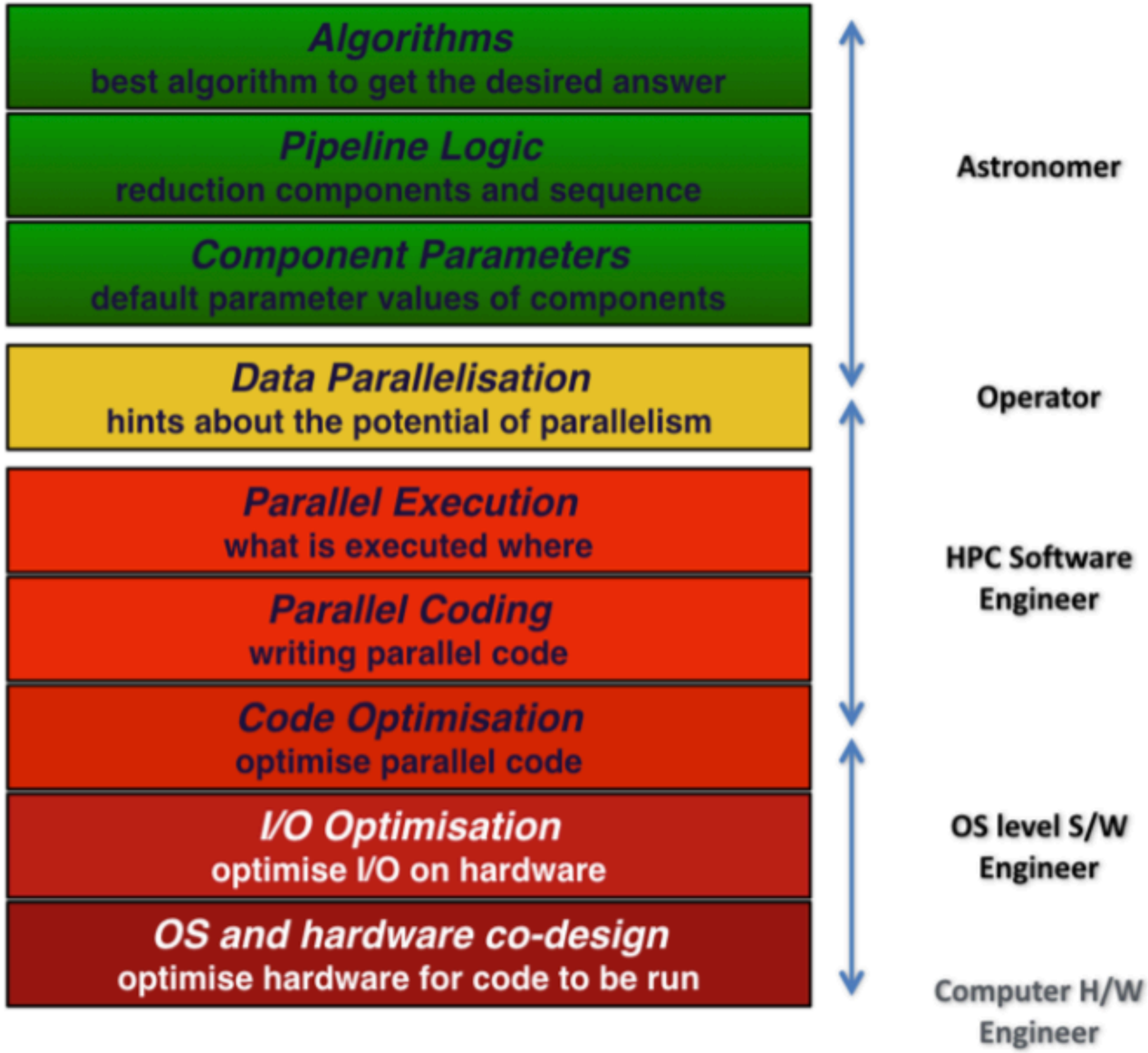
- Workflows have to be modular
- To make this effective, the input and output format of all modules has to be similar, so you can keep data in memory

DALiuGE 流

The Data Activated Liu 流 Graph Engine (DALiuGE) developed by ICRAR is an execution framework for processing large astronomical datasets at a scale required by the SKA1.

DALiuGE provides:

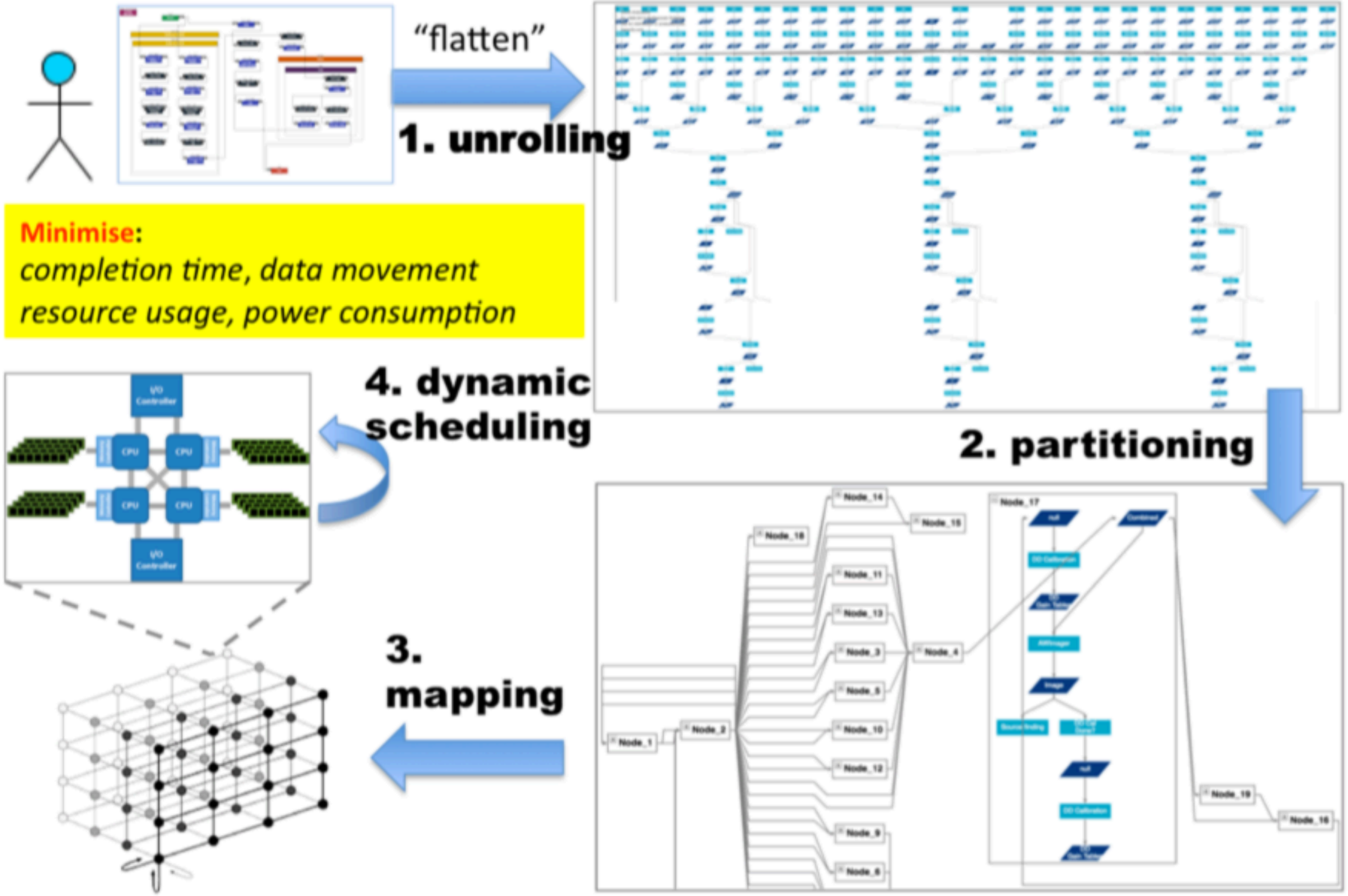
- an interface for expressing complex data reduction pipelines consisting of both data sets and algorithmic components; and
- an implementation run-time to execute pipelines on distributed resources.



Chen Wu

We require workflows and graph engines

DALiuGE execution



Chen Wu

ASKAP: telescope + computer

Cray XC30 Series Supercomputer

472 Compute Nodes:
- 2 x 3.0 GHz Intel Xeon CPUs
- 10 Cores per CPU
- 64 GB DDR3-1866
- Cray Aries Interconnect
- Cray Dragonfly Topology
- 200 TeraFLOPS

1.4 Petabytes Lustre Data Storage

ASKAPsoft Data Pipelines

What will be the model for the SKA ?

Workload larger than humanly possible

- Projects are not run by individuals but large (international teams)
  - project management and team interaction is very important
  - Should we use redmine / confluence / zoom / slack / pbworks / email / jira / git / svn to share information
  - Should we use all (e.g. ASKAP) to make sure no information gets lost, at the risk of losing track overall
- Human pipeline vs Automated pipeline
  - Do you spend your time working on the data or working on the pipeline
- We have to accept that the results will not be as good as theoretically possible
- Loss of expertise

The unwritten rule of a PhD is that you put in a large amount of hard work (i.e. flagging data, eyeballing sources etc.) The one condition here is that at least it is possible to flag or eyeball this data within the duration of a PhD. This condition is changing which is good and bad

We go from ~1 instrument/survey per decade to many

Are we going to many or too many … ?

- **JVLA:**
- **ASKAP:**
- **APERTIF:**
- **MeerKAT:**
- **FAST:**
- **…**

Are we doing too much ? : new telescopes, new software, new projects etc …

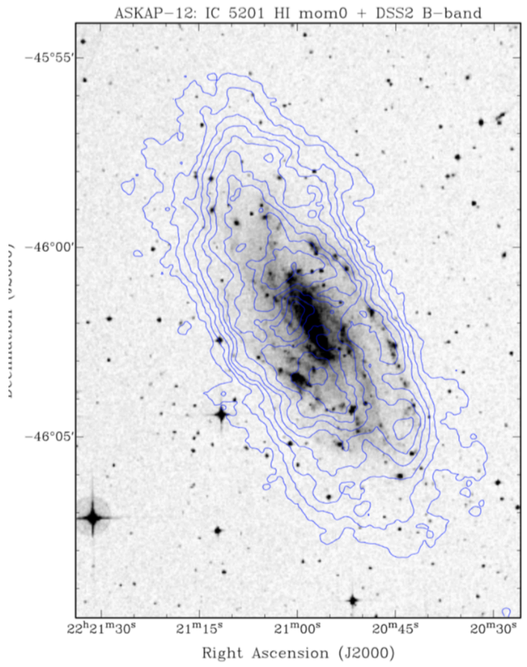
- Most people are involved in many large projects, realistically you can only fully commit to one
- There is a risk in using new software with a new telescope:
  - CHILES statement: We have found this problem, it is either a bug or feature of CASA
  - ASKAP: in early days it was very difficult to verify results
- If you use new software, make sure you have motivated developers in your team
- Keep things transparent, software should be modular and not a black box (e.g. CASA *clean* task)
- We need a collaborative approach to solve the same problems (e.g. SoFiA)

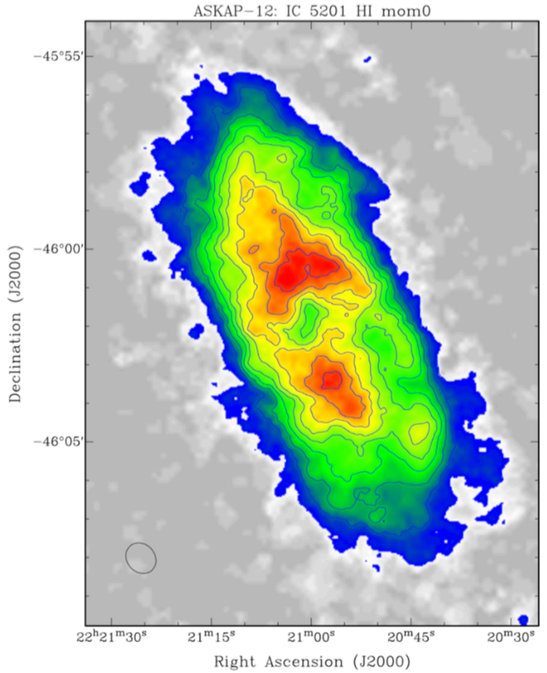
SoFiA - the Source Finding Application

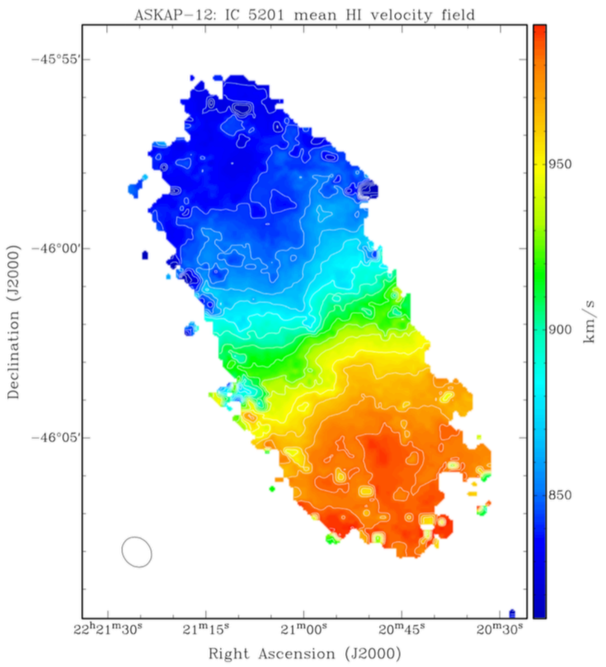
SoFiA
Source Finding Application

- H I source finding and parameterisation pipeline for extragalactic SKA precursor surveys
- Originally initiated by the WALLABY team
- International collaborative project

  T. Westmeier (chair), P. Serra, B. Koribalski, B. Winkel, R. Jurek, H. Courtois, A. Popping, N. Giese, L. Flöer, L. Staveley-Smith, T. van der Hulst, M. Meyer

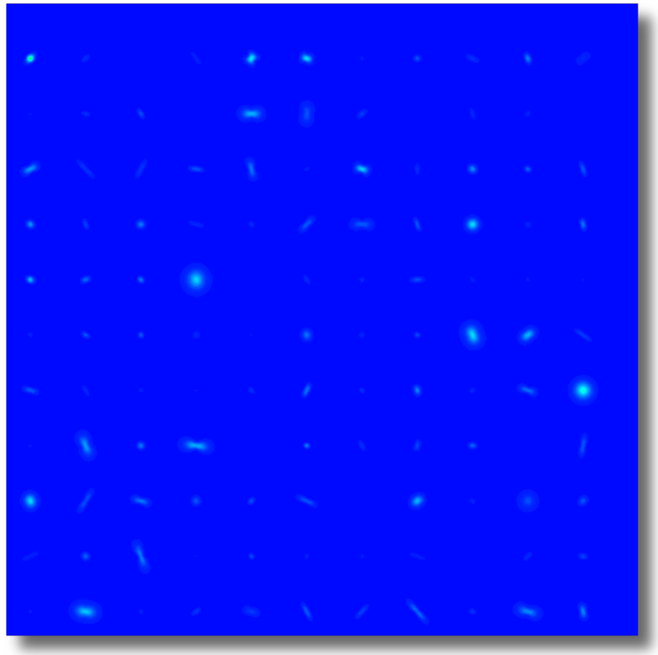




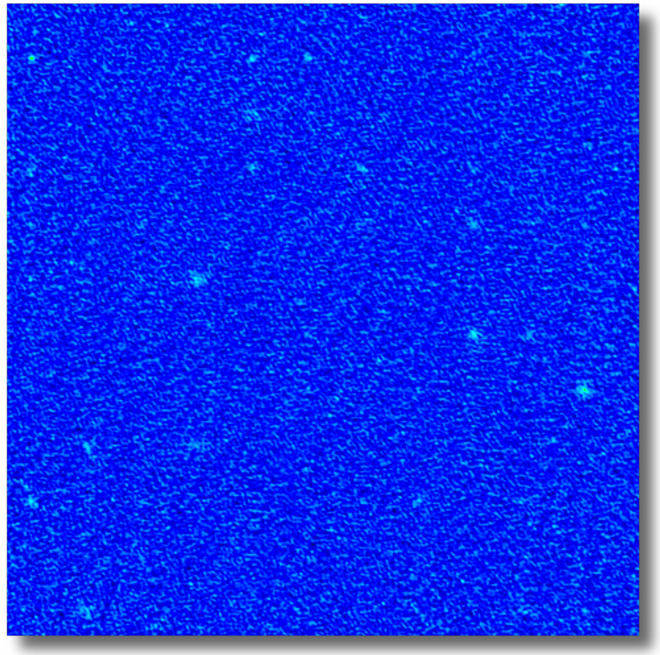


ASKAP Early Science Data

- ★ Model galaxies
  - ▸ Created in GIPSY using galmod
  - ▸ Wide range of
    - sizes
    - fluxes
    - inclinations
    - rotation velocities
  - ▸ Placed on a regular grid for efficiency
    - 600 galaxies
    - cube size: 800 × 800 pixels
      800 channels
    - 1.9 GB of data



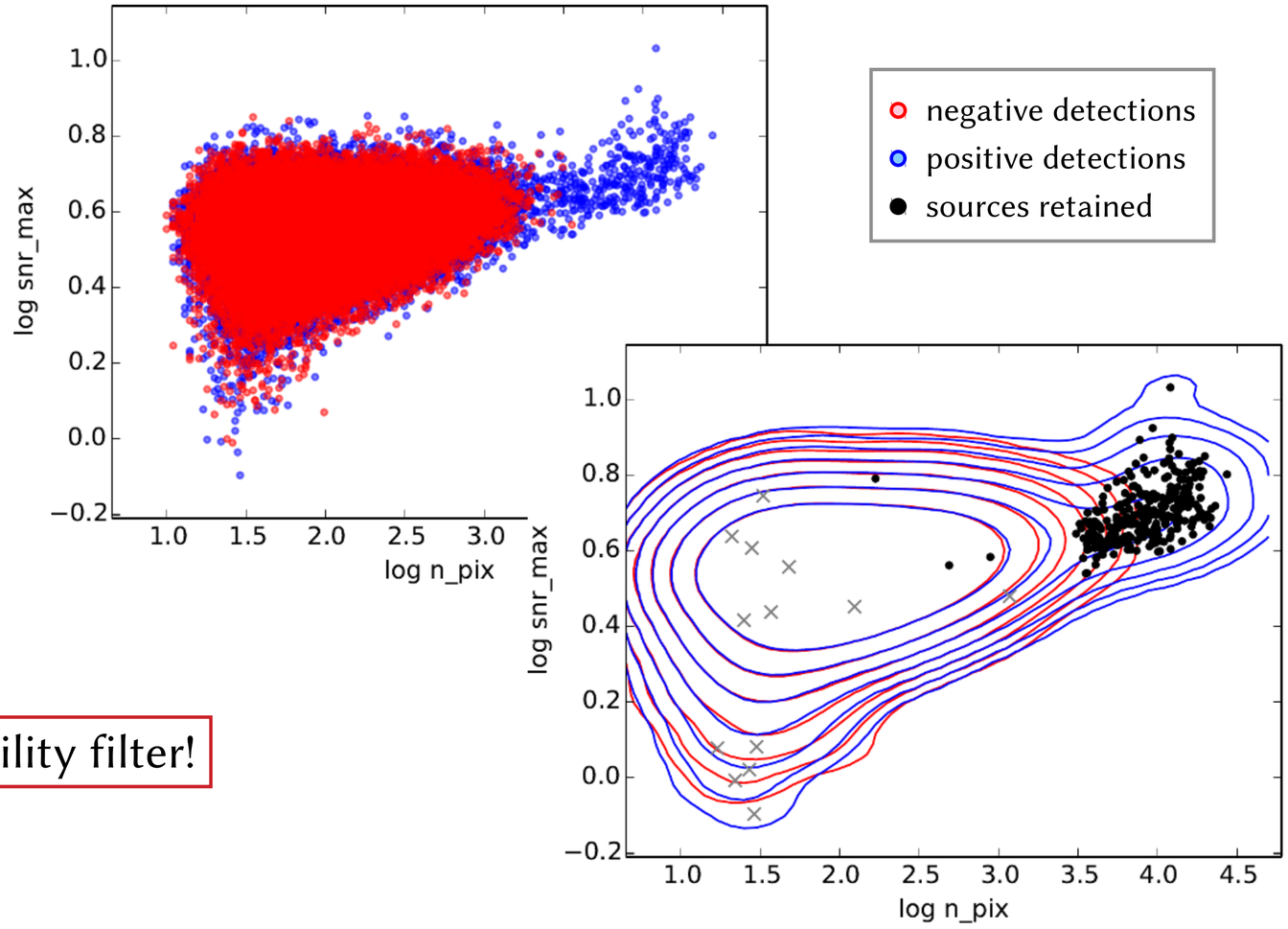
Model galaxies



Model galaxies + ASKAP noise

ASKAP Early Science Data

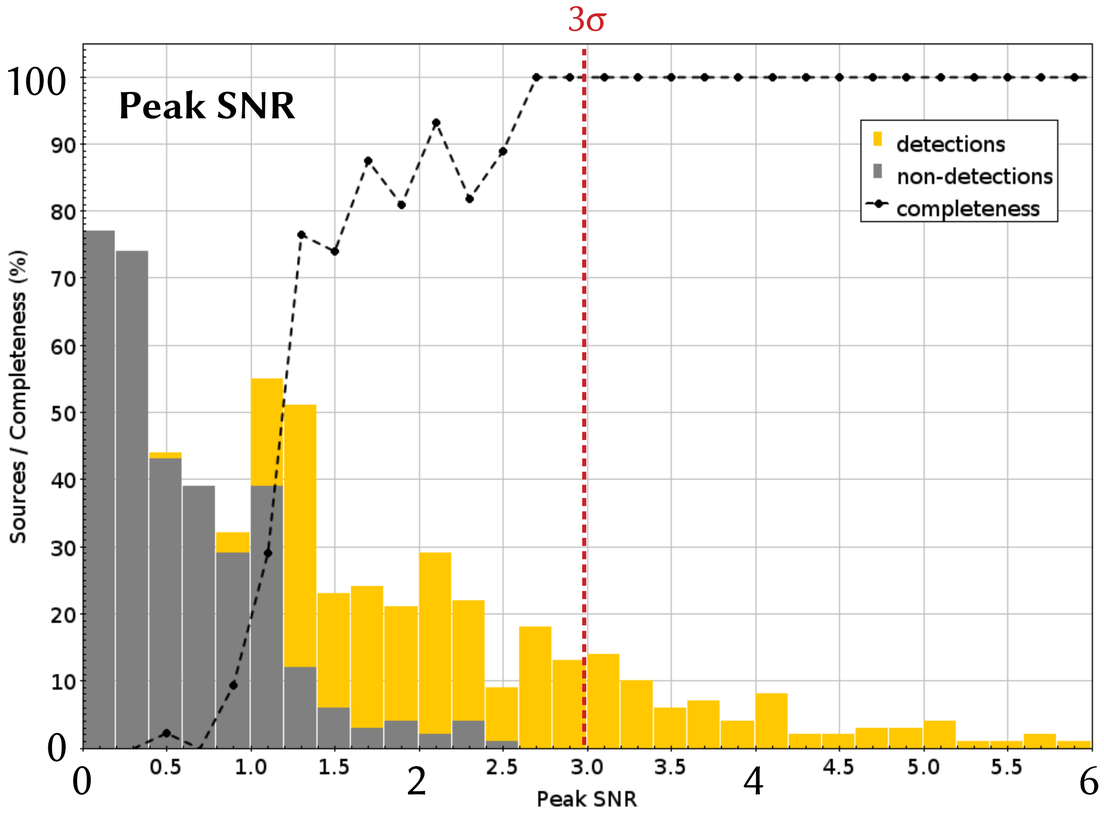
- ★ SoFiA settings
  - S+C finder with 3σ threshold
  - Reliability estimation enabled
- ★ Results
  - About 63,000 detections
  - Half of these are negative
  - 276 sources remain after reliability filtering

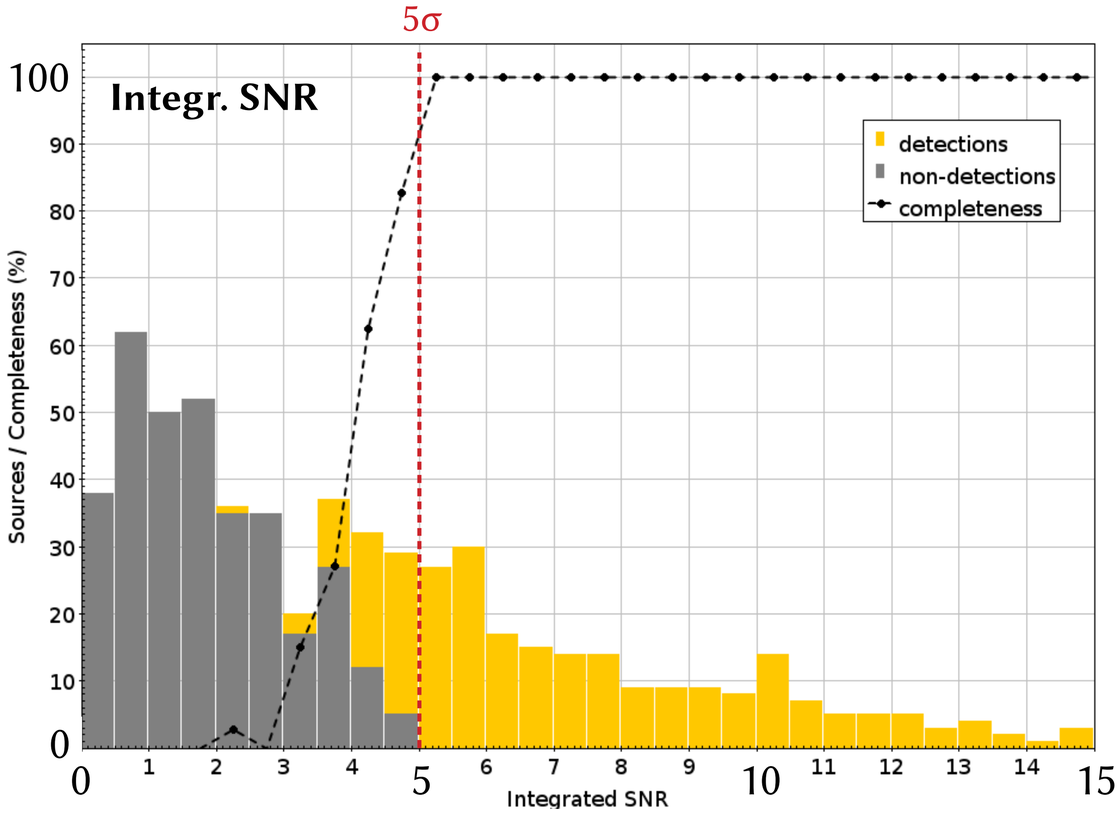


High reliability thanks to reliability filter!

T. Westmeier

★ Completeness





Peak SNR: 100% above 2.6σ
detections at ≳ 0.5σ

Integrated SNR: 100% above 5σ
detections at ≳ 2σ

ASKAP Early Science Data

- Data set II
  - ASKAP–12
  - H I data of the NGC 7232 group
  - WALLABY early science
  - Mosaic of multiple PAF beams
  - 12.6 GB in size
  - Thanks to Karen Lee-Waddell

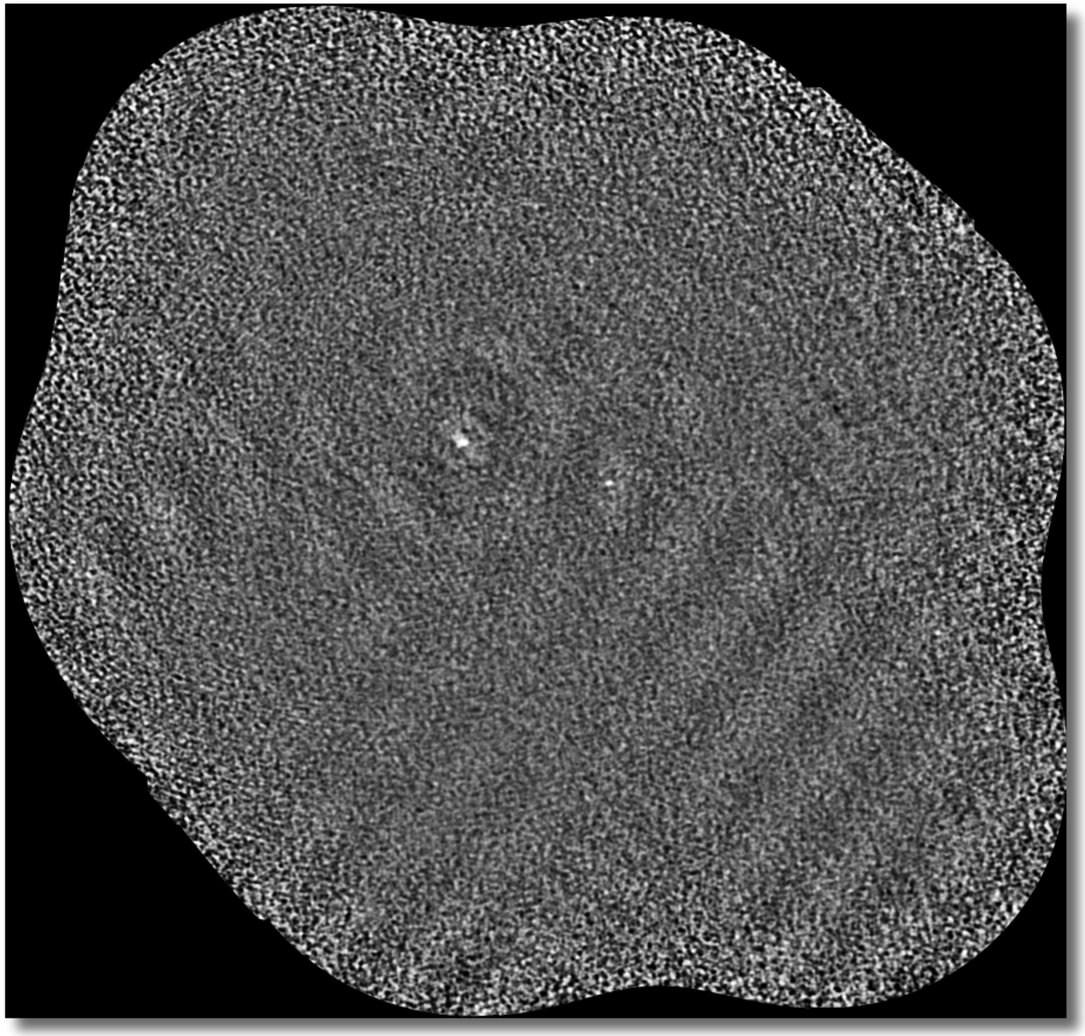
WALLABY

Wide-field ASKAP L-band Legacy All-sky Blind surveY
(Koribalski & Staveley-Smith)

ASKAP, Boolardy, Western Australia

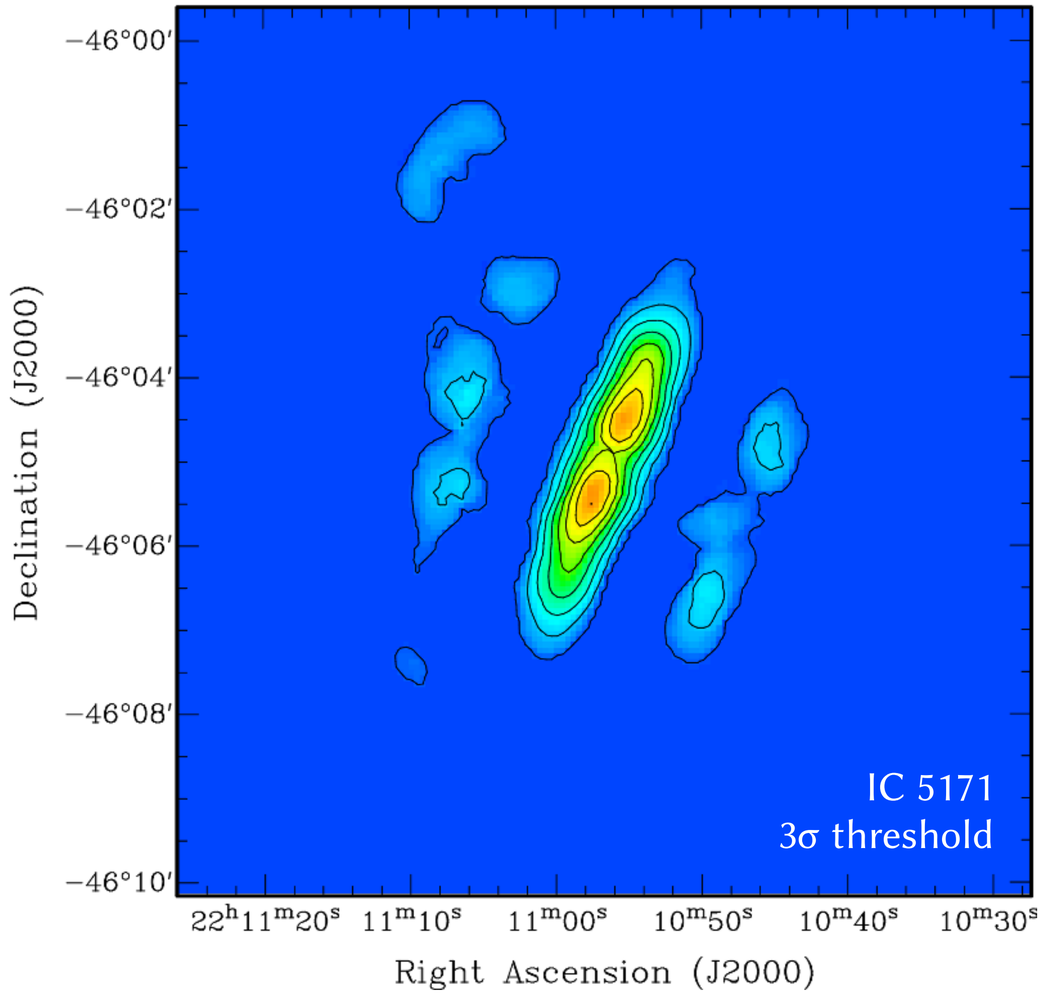
ASKAP Early Science Data

- Challenges for SoFiA
  - Strong noise variation across field
  - Residual sidelobes from shallow deconvolution
  - RFI stripes from insufficient flagging
  - Other imaging artefacts
- Note
  - A lot of these problems with ASKAPsoft are now under control or being resolved
  - Nevertheless a good test data set to throw at SoFiA

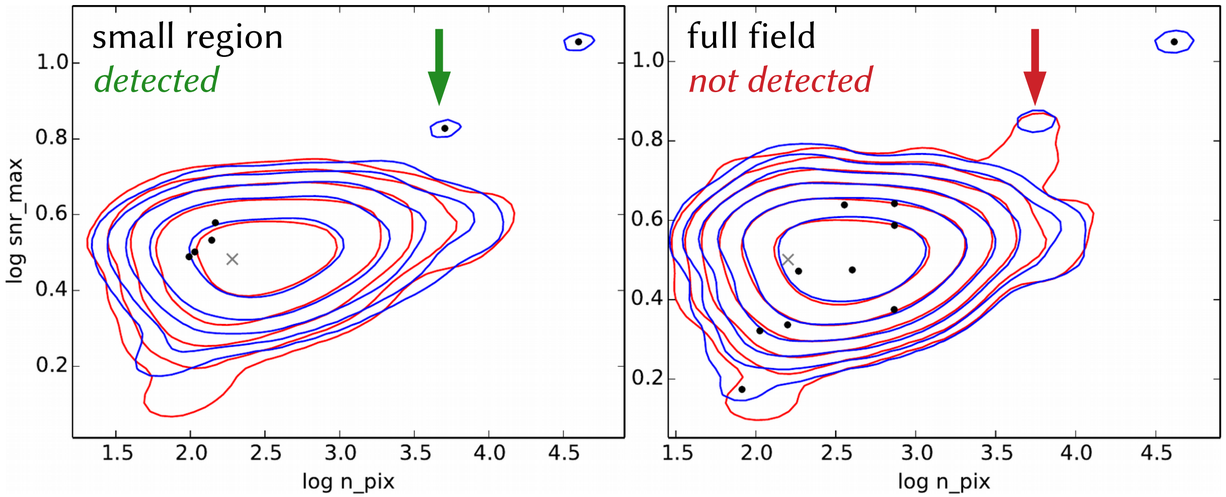


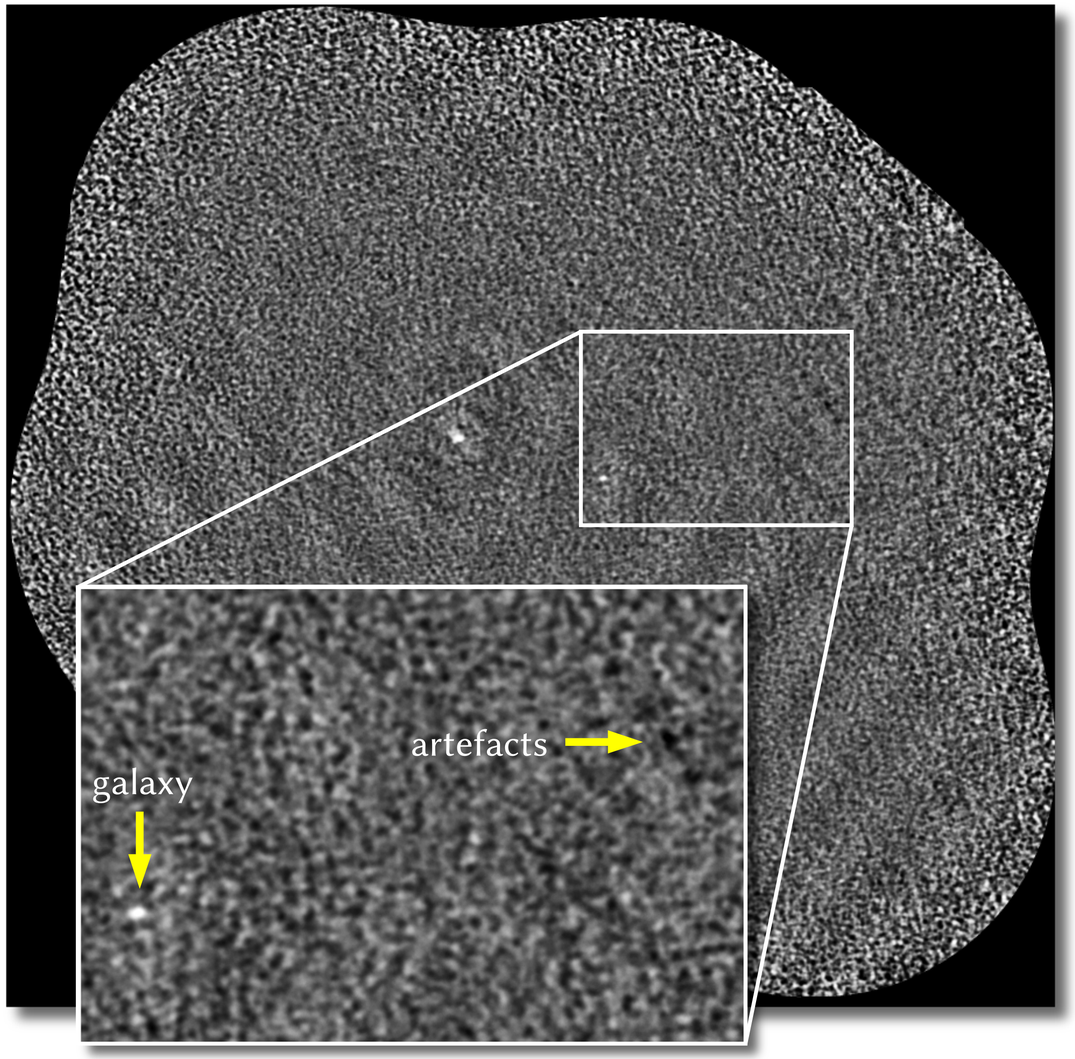
ASKAP Early Science Data

- ★ Issues and solutions
  - ▸ Low detection threshold would pick up too many artefacts
  - ▸ Use 5σ threshold



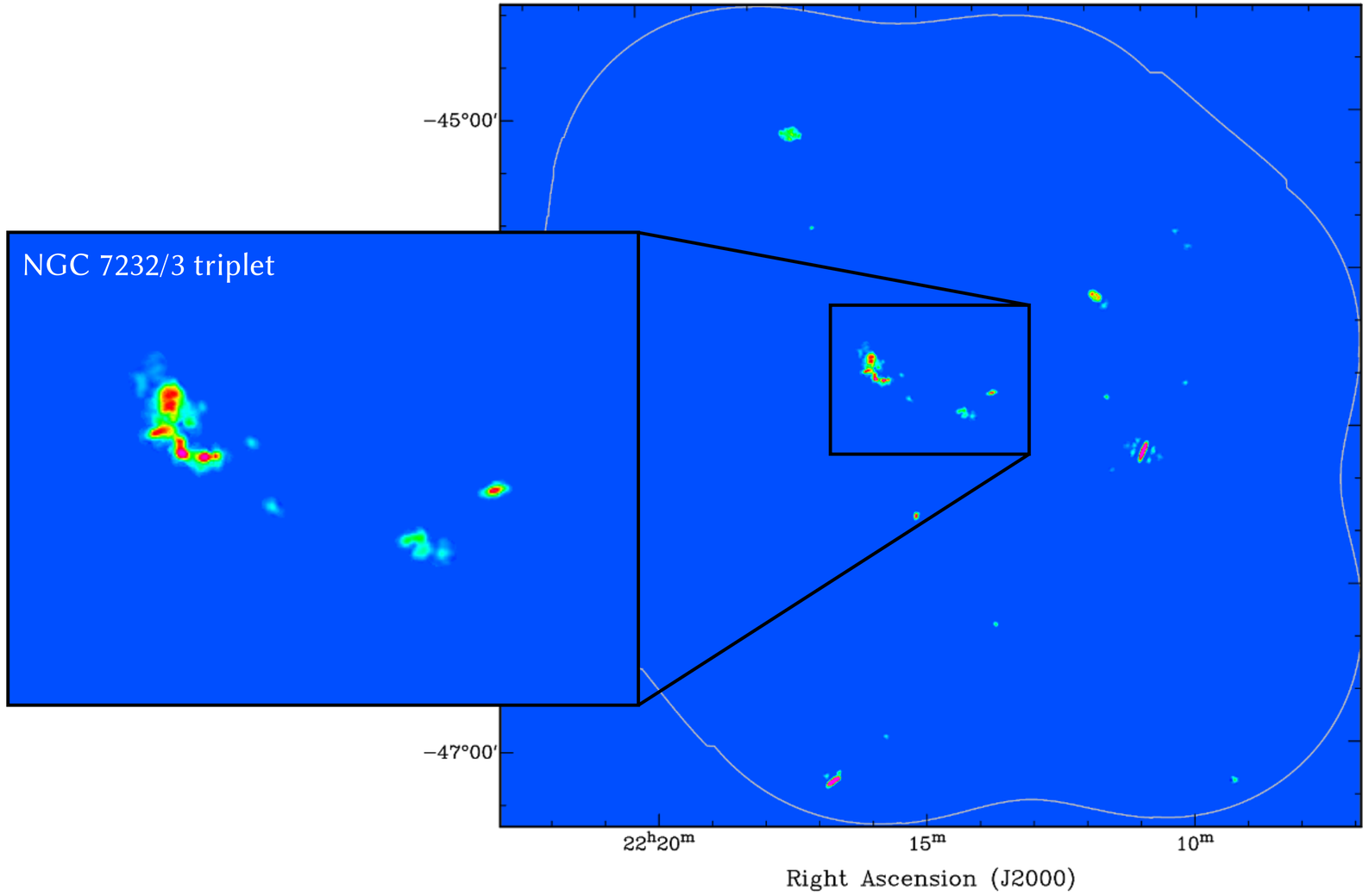
- Issues and solutions
  - Reliability estimator will discard some galaxies as their parameters are similar to those of artefacts
  - Assumption: *negative detections are noise*
  - Disable reliability filter





ASKAP Early Science Data

- ★ Result
  - ▸ Detailed H I map of NGC 7232 group
  - ▸ Lee-Waddell et al. (in prep.)



Source Finding

- ★ Source finding on ASKAP early science data using SoFiA works
- ★ Clean data free of artefacts
  - ▸ Low threshold of 3σ possible
  - ▸ Completeness of 100% at $SNR_{peak} \gtrsim 2.6$ and $SNR_{int} \gtrsim 5.0$
  - ▸ Reliability near 100%

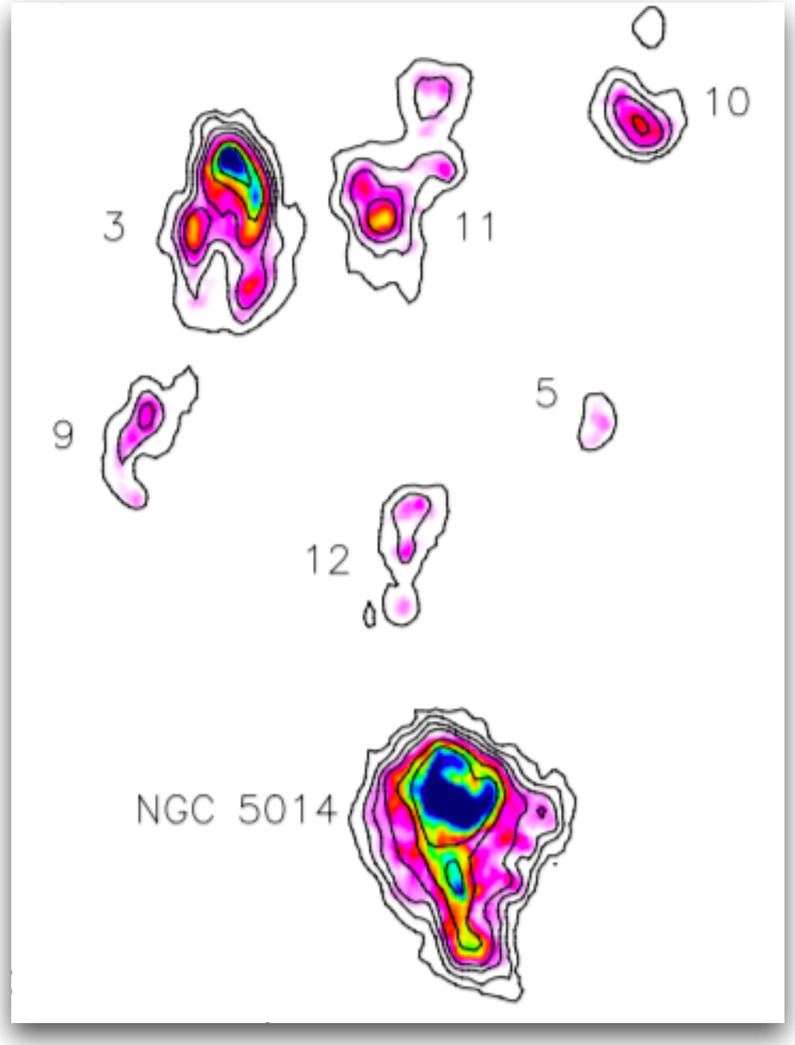
⇨ **BEST CASE SCENARIO**

- ★ Data affected by artefacts
  - ▸ Higher threshold of 5σ required
  - ▸ Limited completeness and reliability
  - ▸ Negative artefacts render reliability filter useless

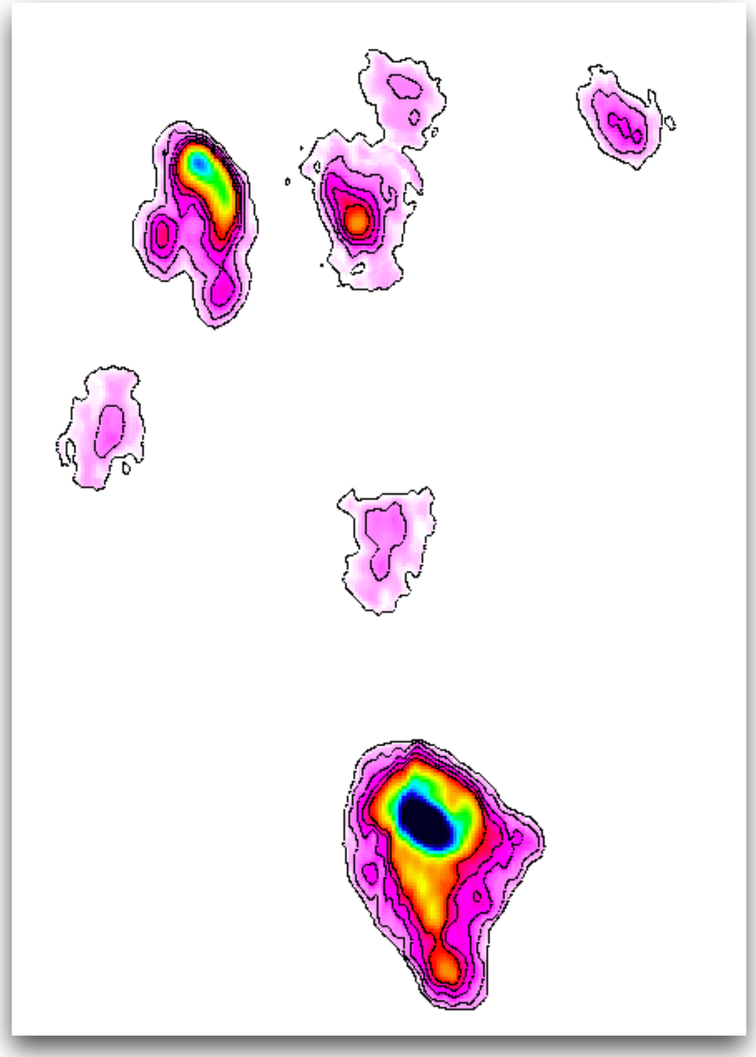
⇨ **WORST CASE SCENARIO**

LGG 334 then and now

AP's Masters thesis (with Thijs)
~ year



SoFiA (with Thijs)
~20 seconds




Did you solve the problem?
*“No, but I am very positive.
Otherwise why coming to work”*

Questions (to you)

- Can we afford to delete raw visibilities ?
- Will we ever reach a state where we can do single pass processing ?
- Is it still realistic to do data reduction by hand ?
- Should we put effort in decreasing data volume rather than more processing ?
- Who should be working on or take responsibility for RFI mitigation ?
- Have you implemented data verification and quality control ?
- How do we avoid reinventing the wheel for every telescope ?
- Do you have enough software engineers in your team ?