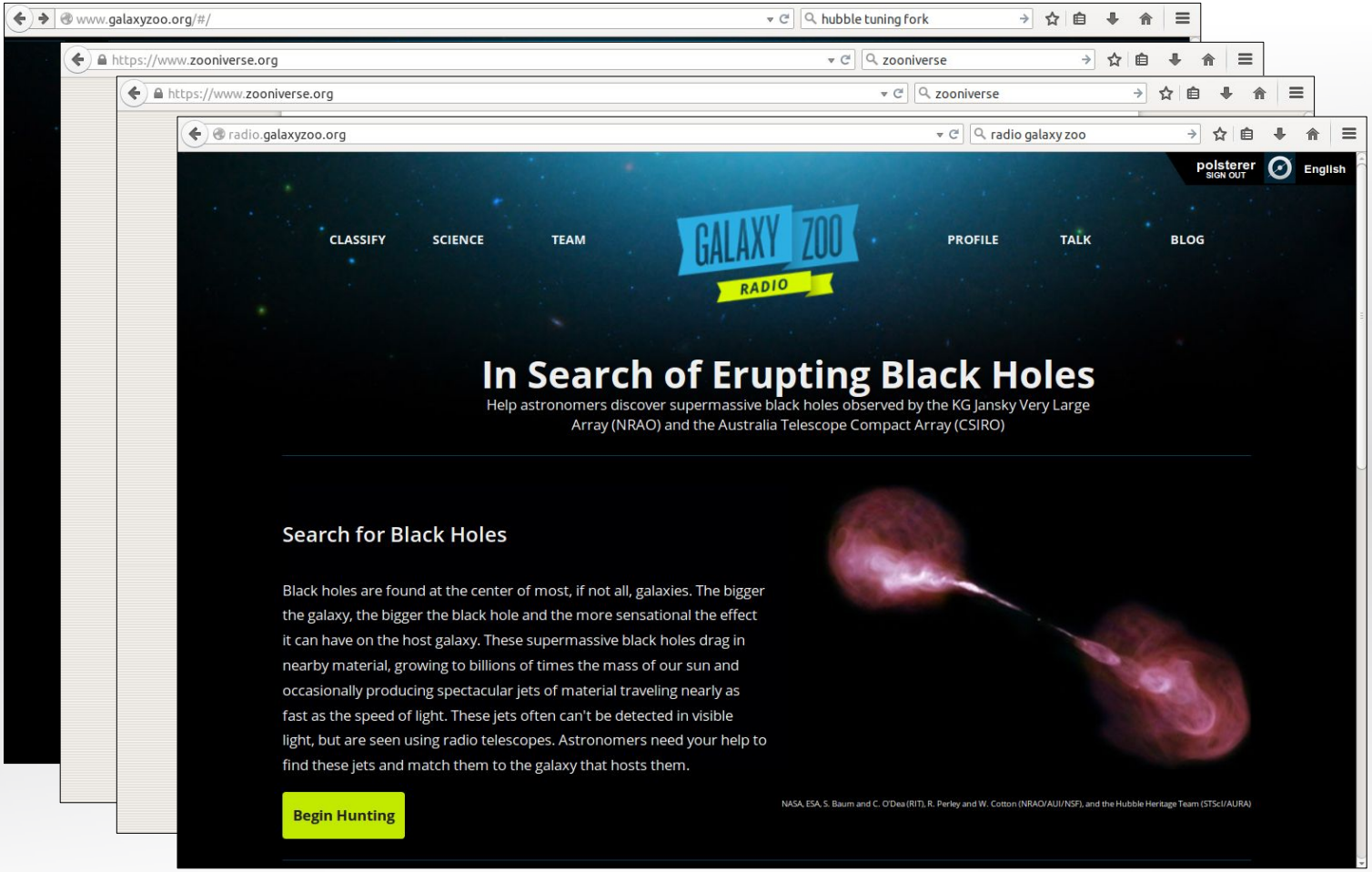
Heidelberg Institute for
Theoretical Studies

HITS

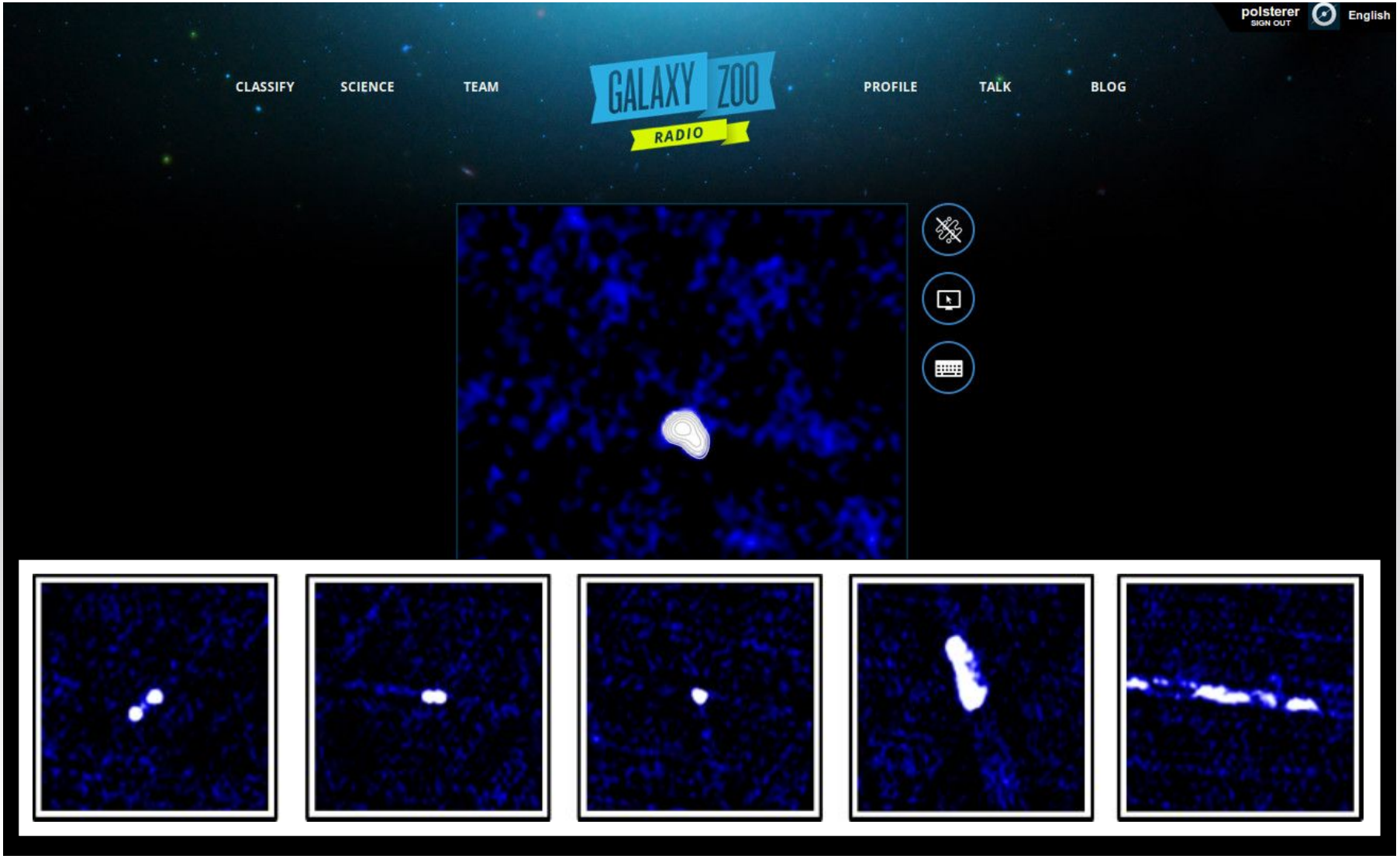
Wide Field Outlier Finder (WTF)

using cloud infrastructure to cooperate, scale, and democratise e-science

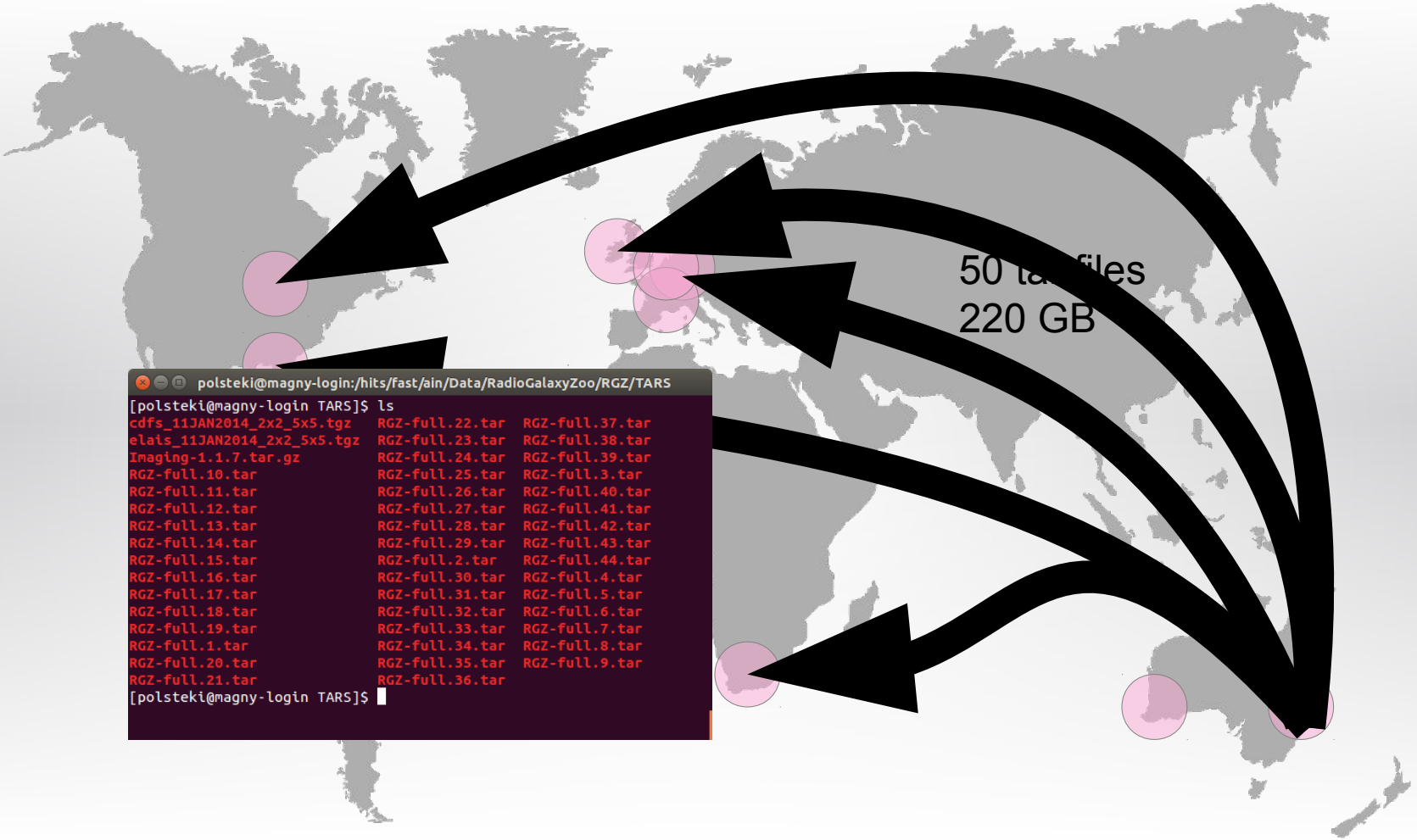
Crowd-sourcing / Citizen Science



Radio Galaxy Zoo



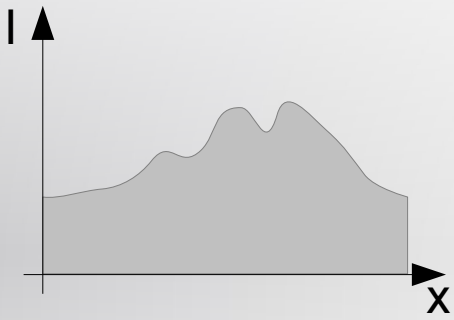
Starting the Project



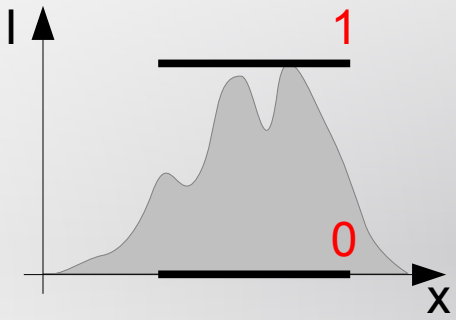
Preprocessing

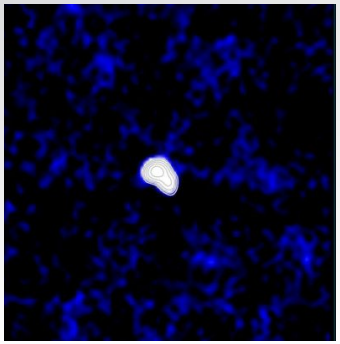
extract
matrix from fits

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & A_{2n} \\ \vdots & & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

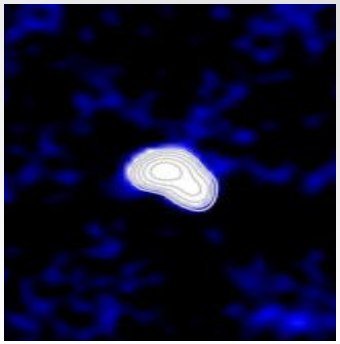


normalize
flux relative to the maximum





cutout
interesting region



Speeding up Preprocessing

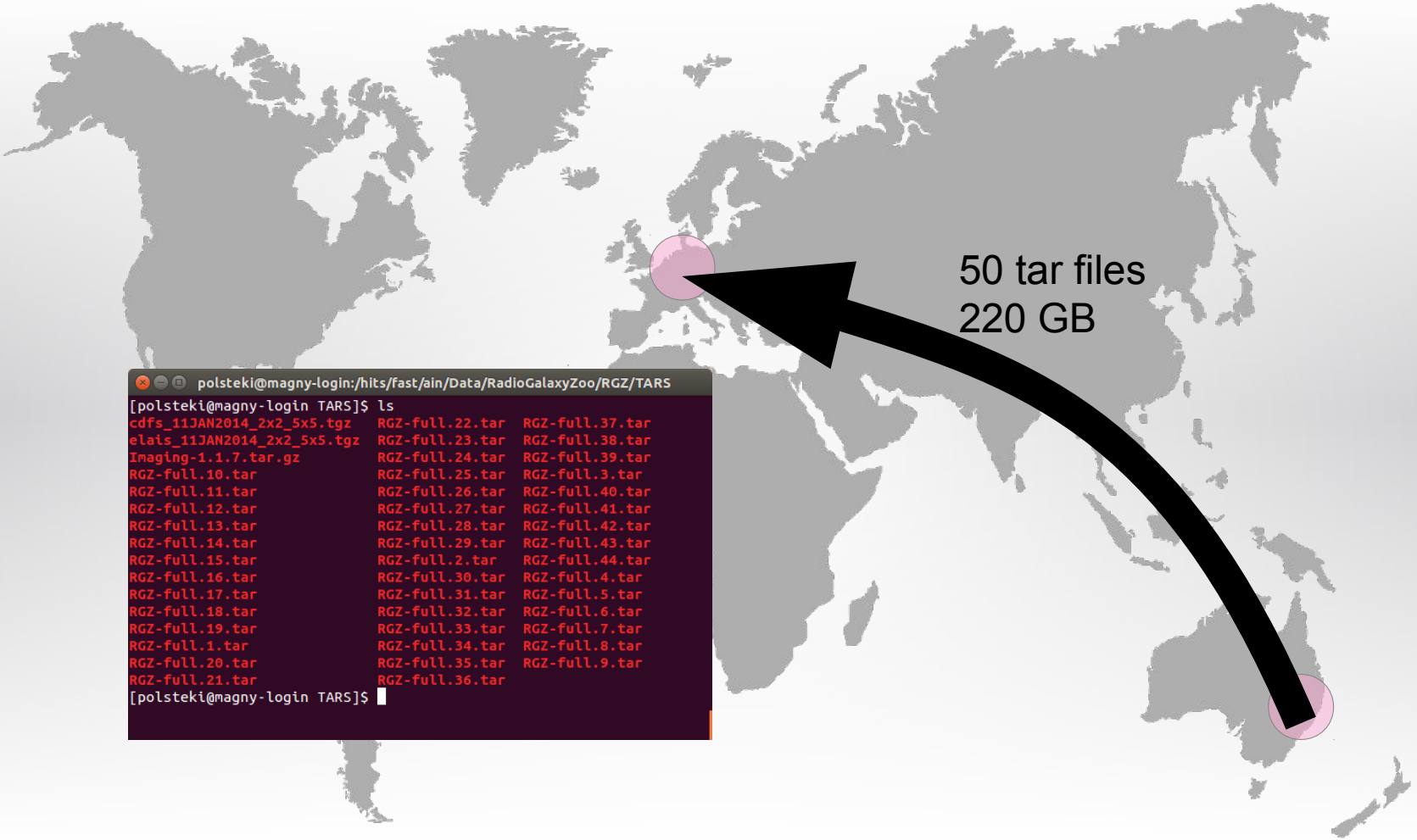
single core python = **48** hours

with hadoop

on **4x16** cores = **4** hours

file **access** is still the bottleneck!

New Images Extracted



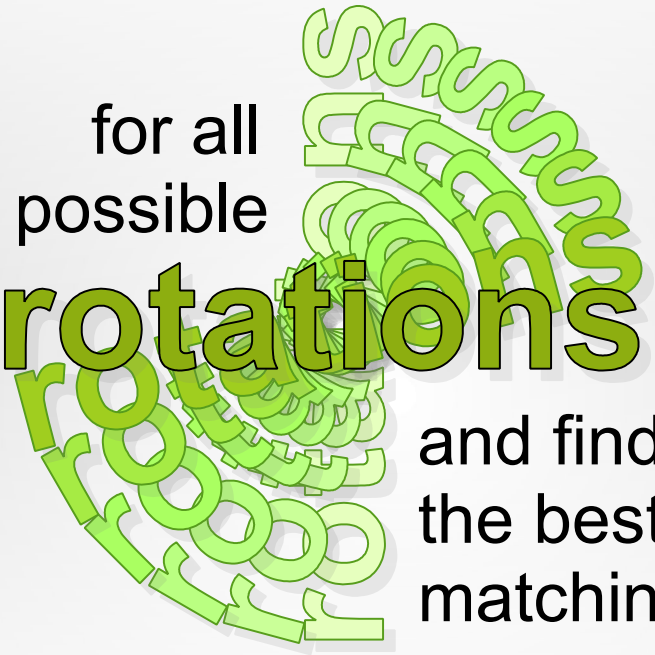
Similarity Measure

calculate the pixel based

Euclidean distance

for all possible

rotations

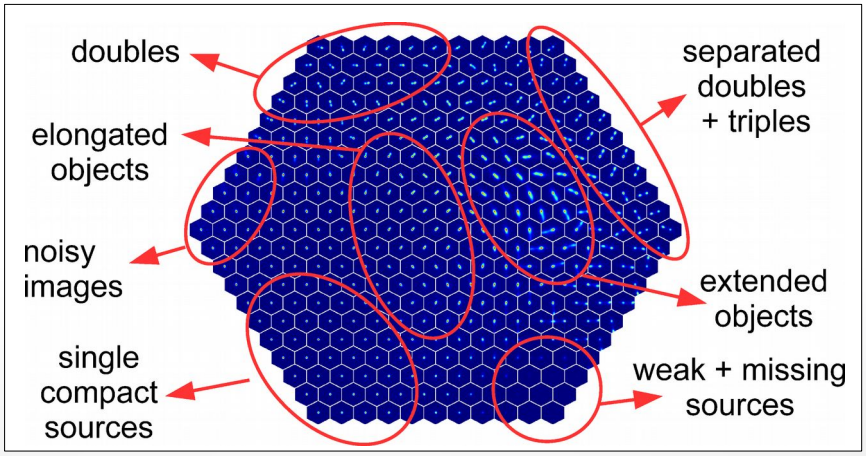


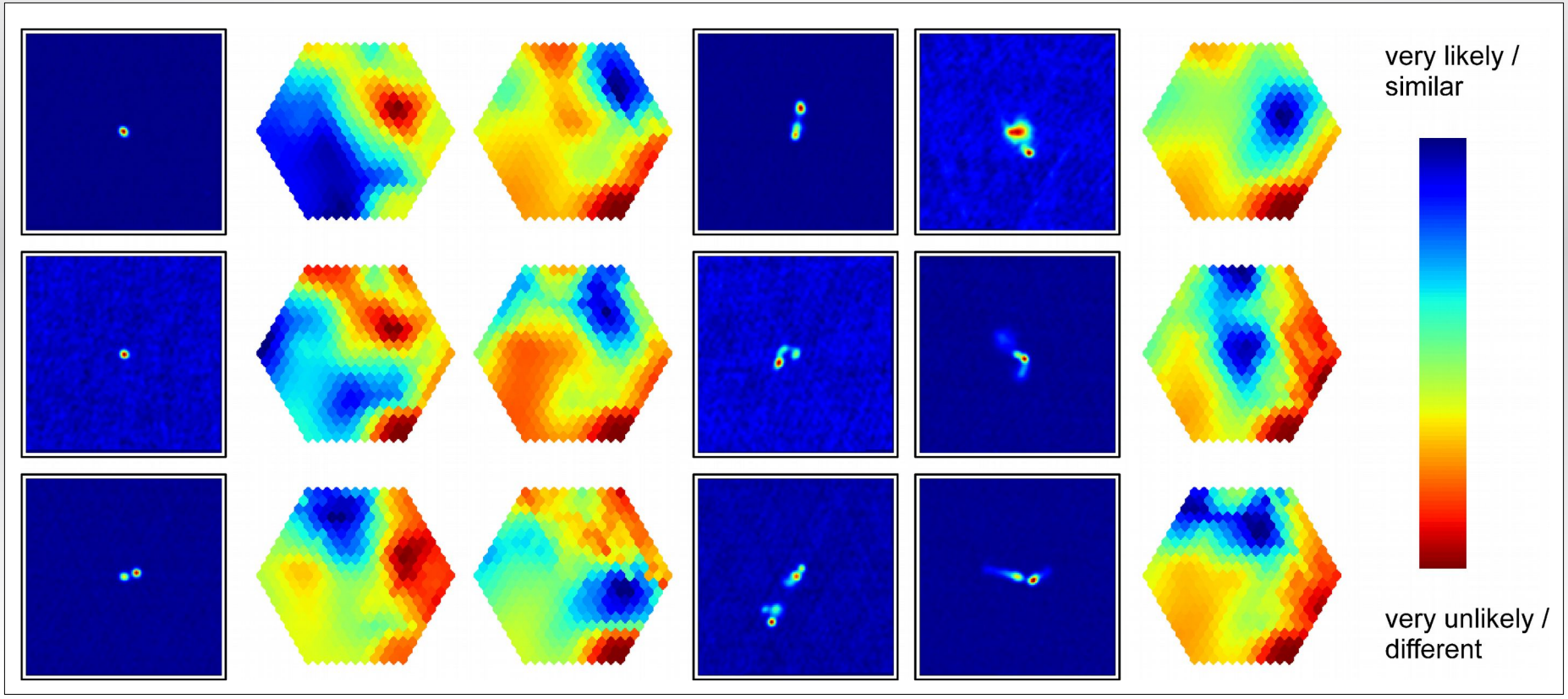
and find the best matching angle

via

minimization → GPU

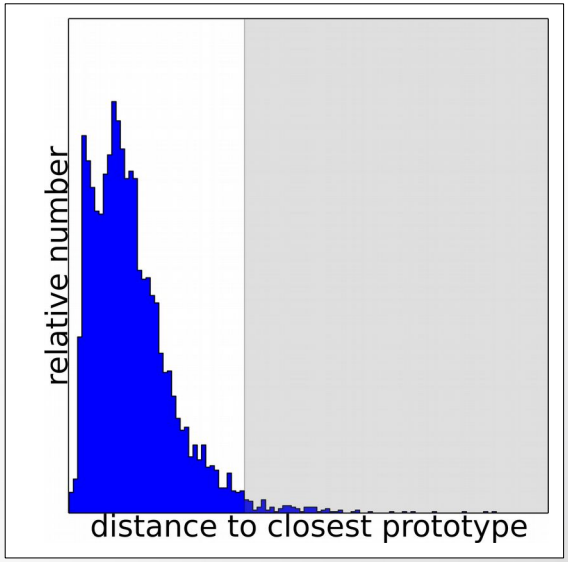
Processing

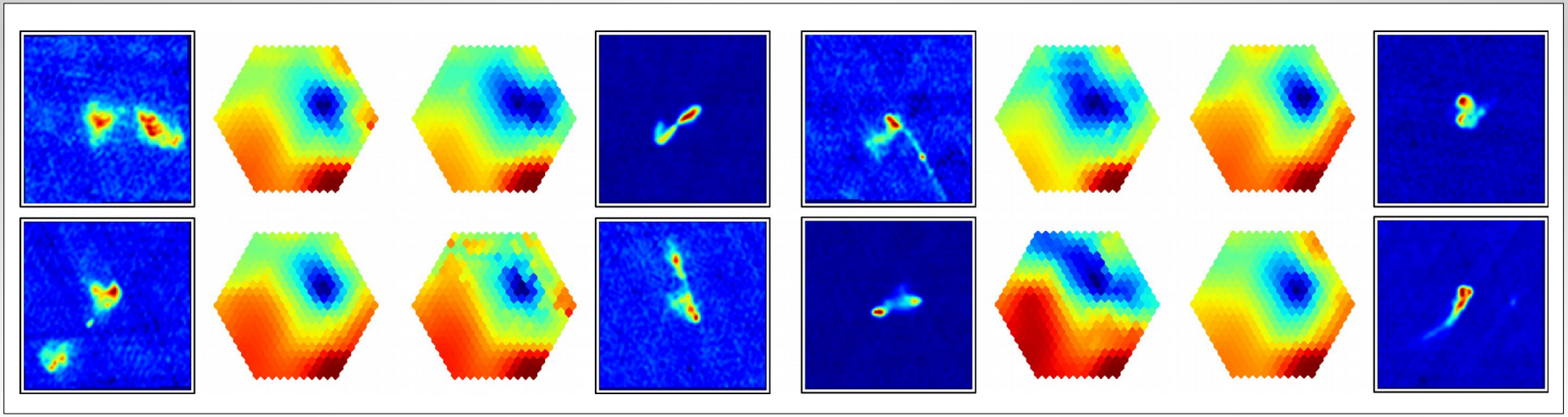

doubles
elongated objects
noisy images
single compact sources
separated doubles + triples
extended objects
weak + missing sources


very likely / similar
very unlikely / different

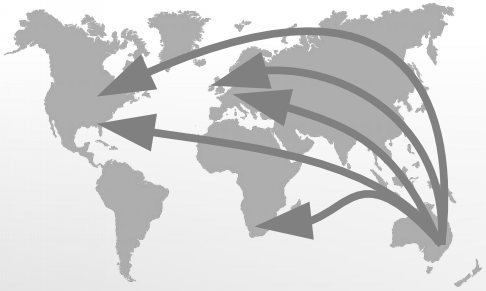
Interesting Objects

select **outliers** based on distribution of distances

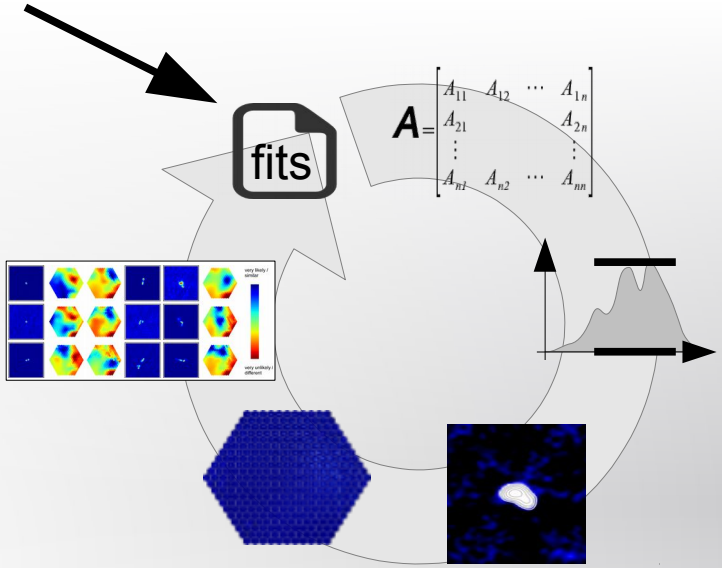




The Downsides of this Approach



a lot of local copies



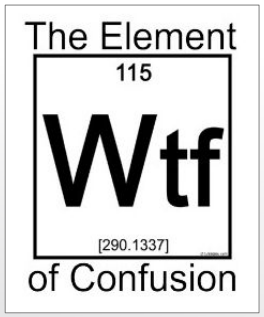
no orchestration of work-flow

bad exchange of intermediate results

very exclusive concerning hardware requirements

NVIDIA Tesla K40

A new Project



Widefield ou**T**lier **F**inder
(Ray Norris)

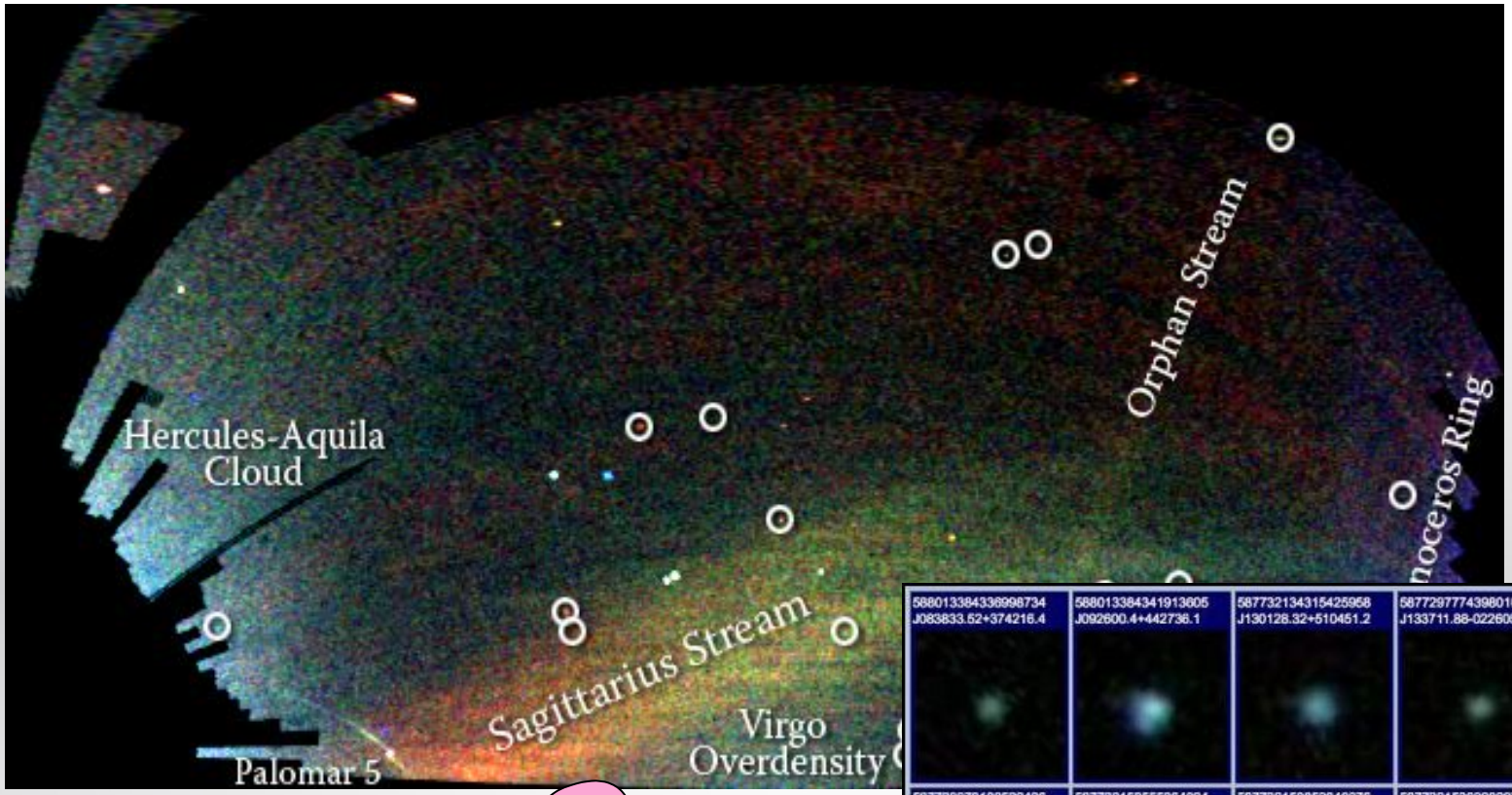
ASKAP-EMU project

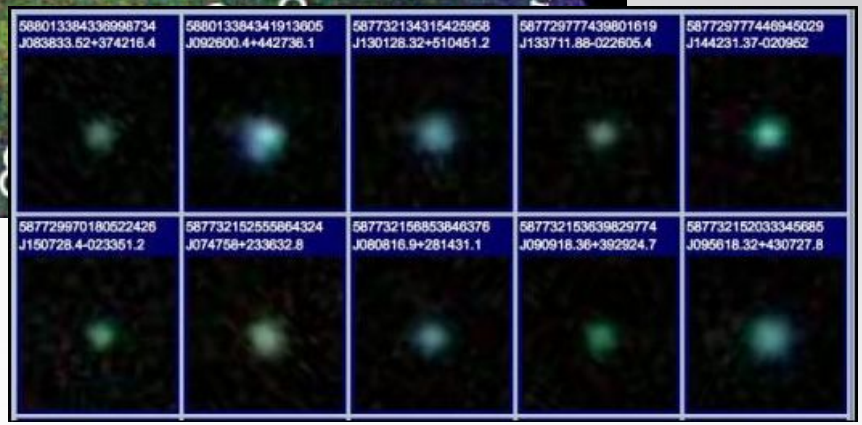
from 2.5 mill. → **70 mill.** known radio sources

develop methodology to discover **unexpected**

science in large data sets

What are Outliers?

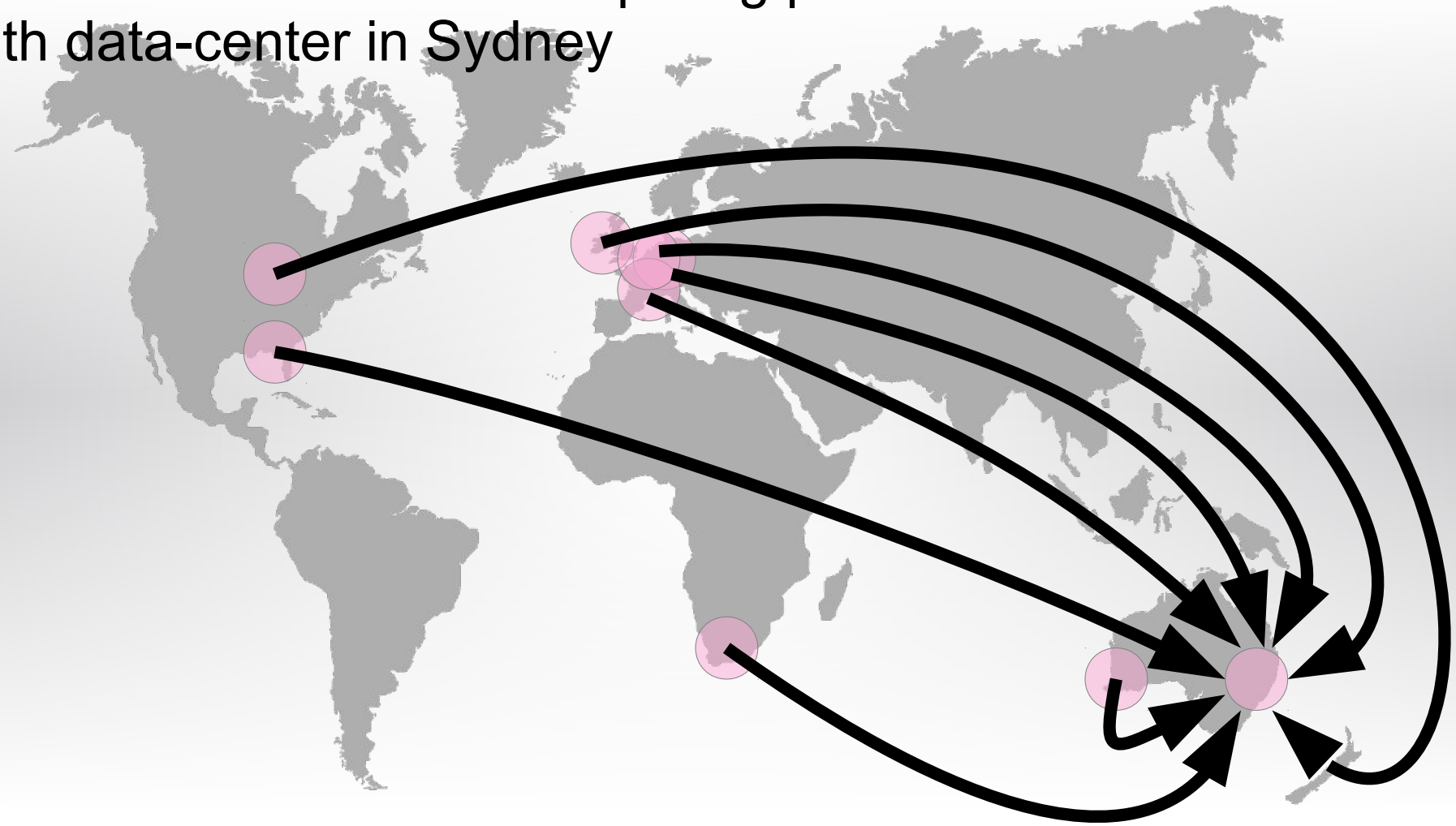




IFRS

A new Concept

use commercial cloud computing provider
with data-center in Sydney



Why use a Company?

HITS

international project:

- who is providing resources?
- who has the infrastructure?
- who provides the services?

why not run our own science cloud?

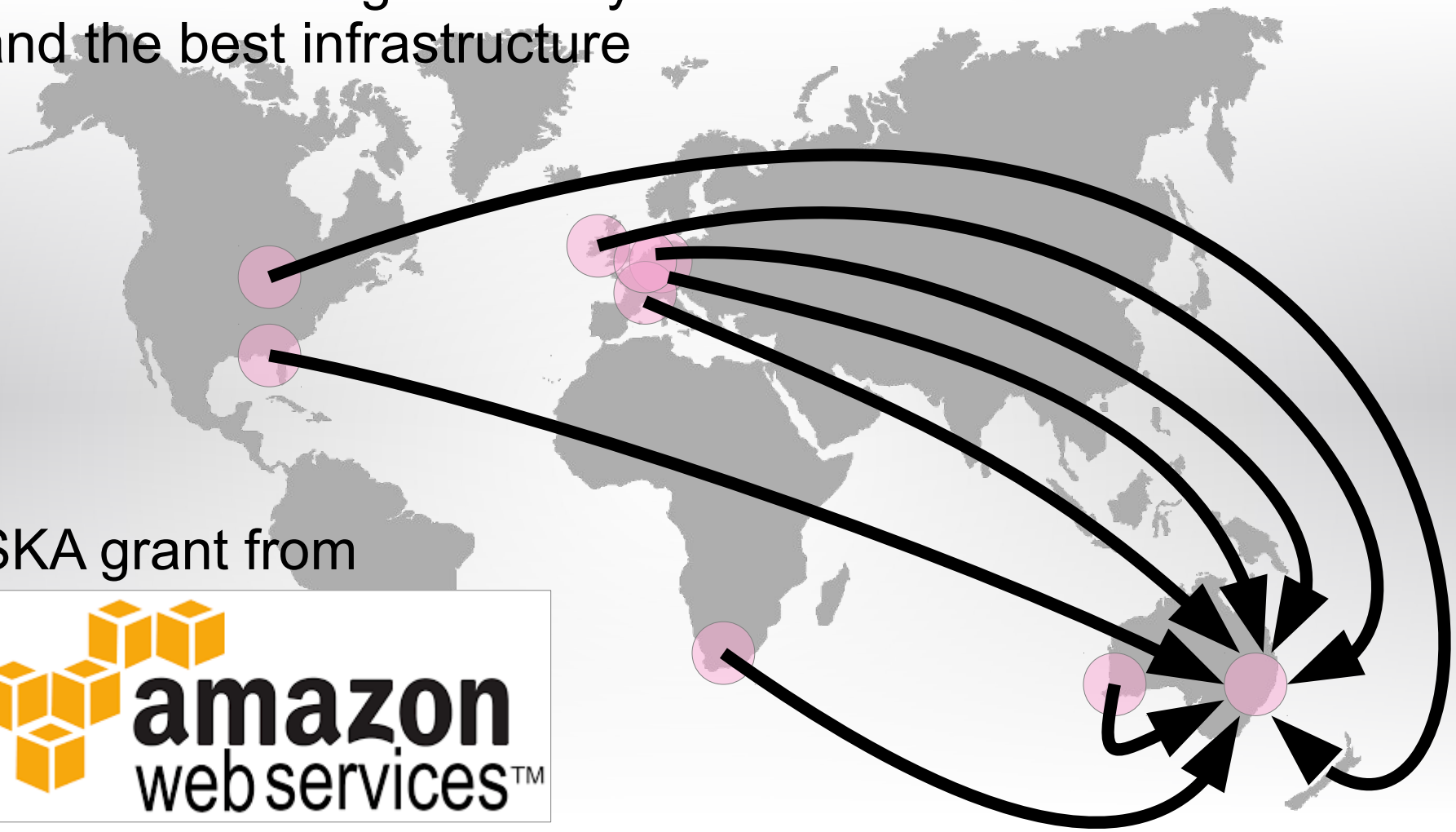
- using open stack

Commercial Cloud Provider

There a many Cloud Provider

Use Amazon Web Services

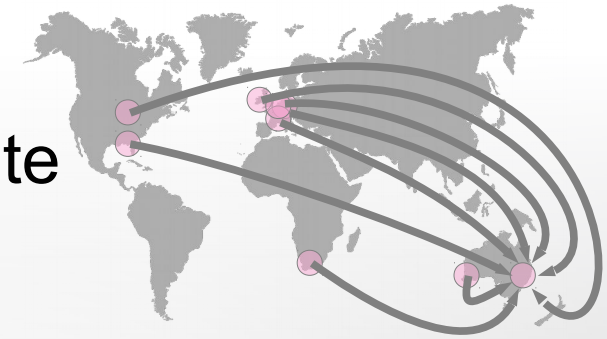
provides the larges variety of services
and the best infrastructure



SKA grant from

The Concept

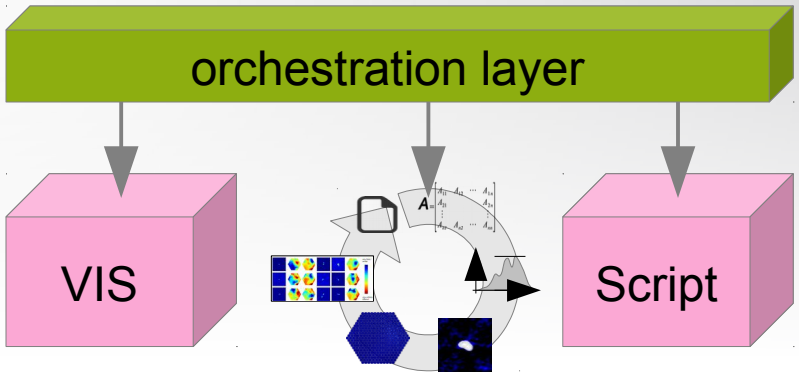
centralized storage and compute



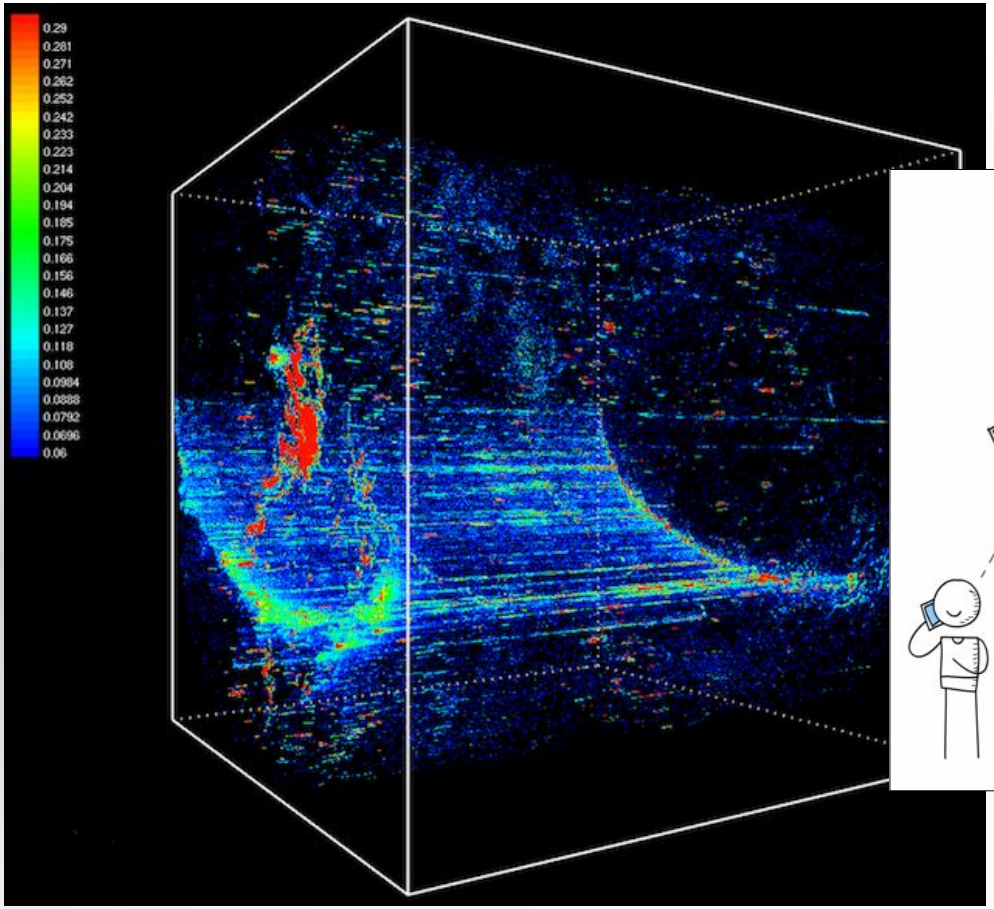
REDMINE
flexible project management

embedded repository and project management tool

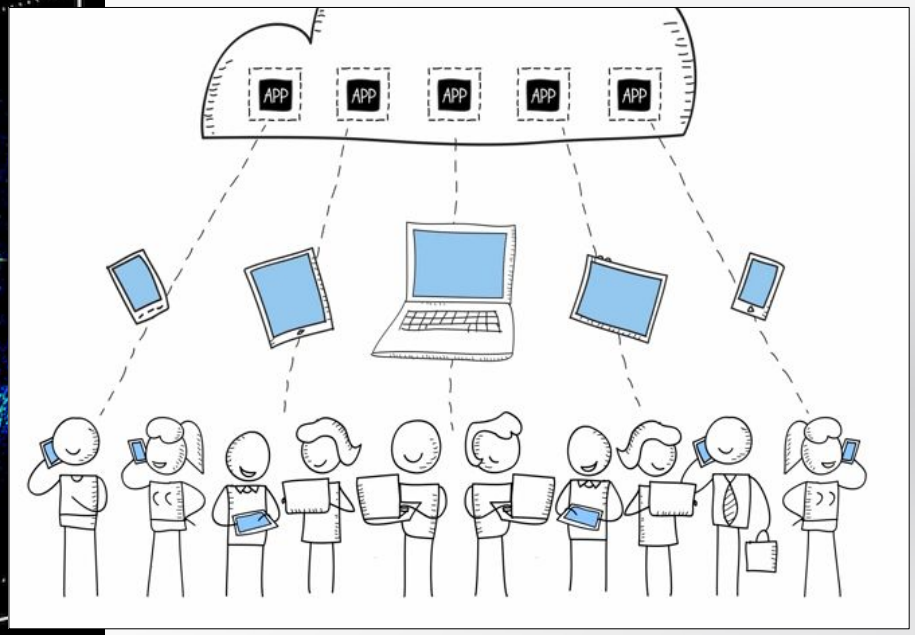
script all preprocessing and provide orchestration



Remote Visualization

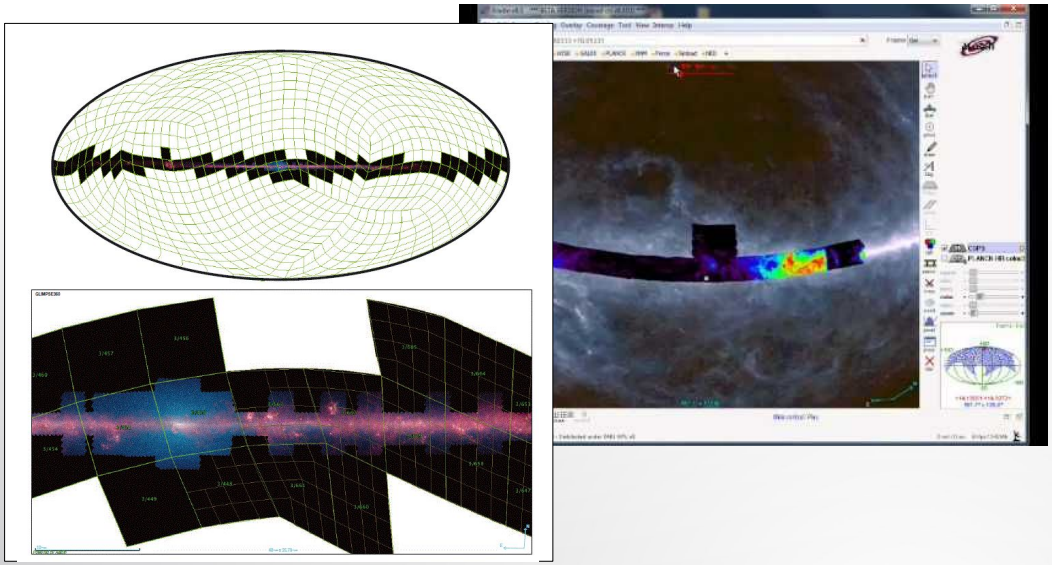


Amr Hassan



AWS appStream

HIPS is Underestimated



Fernique et al. 2015 A&A

don't miss underesti-
mate me

more . than . just .
visualization

it's a
hierarchical data-structure

What is really new?

scaling UP and DOWN immediately

no longer depending on exclu$ive hardware requirements

no difference between small university institutes and large research centers

data via credits challenges open to all

Conclusion

thanks to Virtual Observatory we know how to access data

we still need to learn how to uniformly process and analyze data

Thank you for your attention !

@AstroInformatix