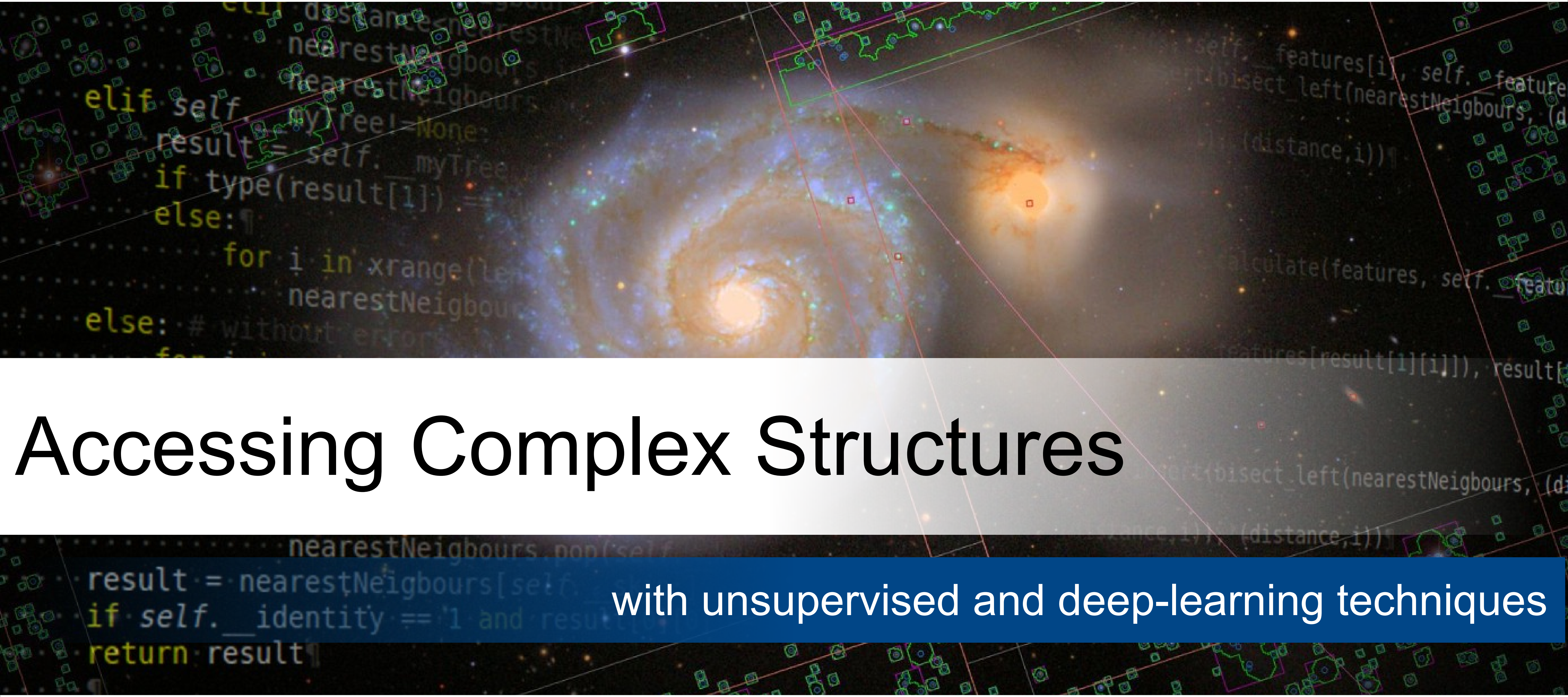
Heidelberg Institute for
Theoretical Studies



Accessing Complex Structures

with unsupervised and deep-learning techniques

Story from the past

Computer-science meets astronomy

- matching 3 lists, 200k each
- 3 nested for-loops, without break statement → $o(n^3)$
- 12 days compute time, 7 days remaining, but only 5 days until observation run

vs.

- scanning version on presorted lists → $o(n)$
- results after 4 seconds

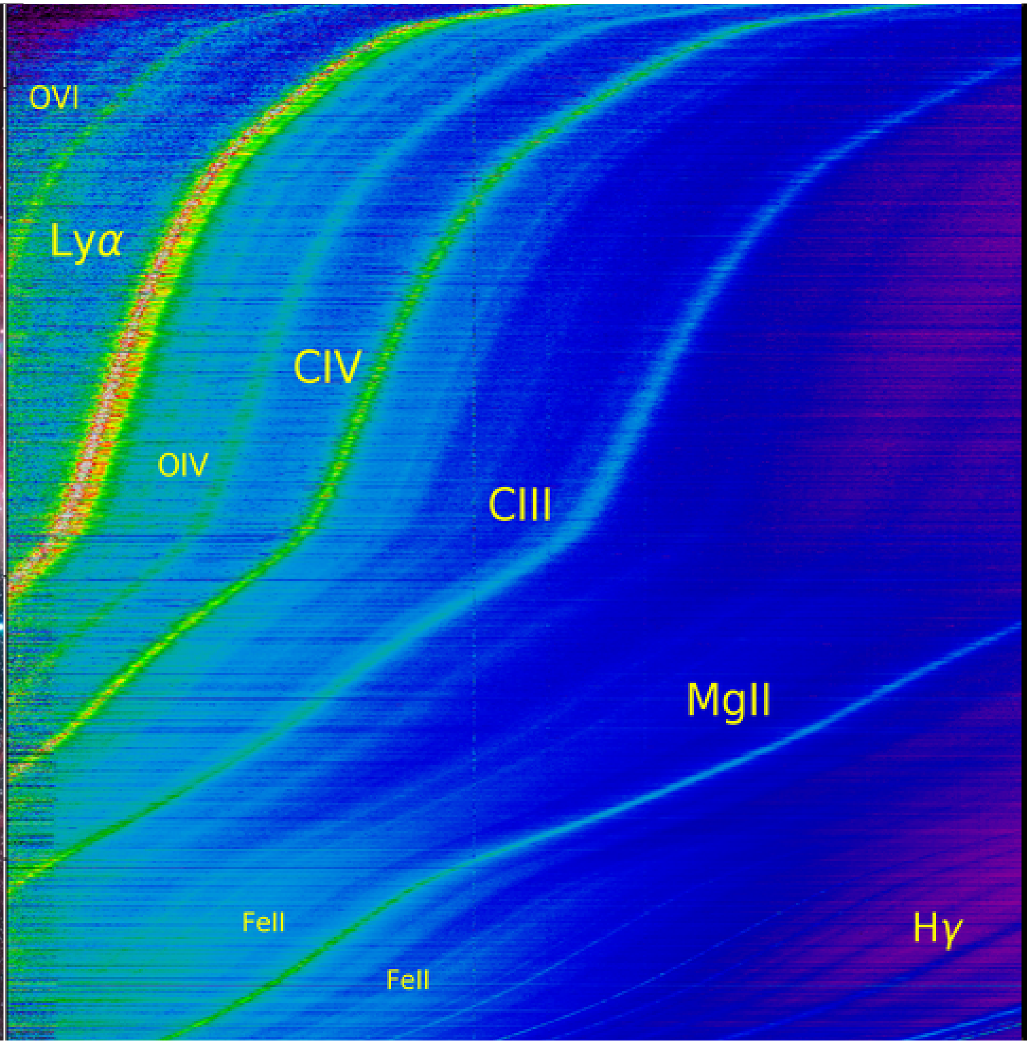
It’s not just about big-data

spatial

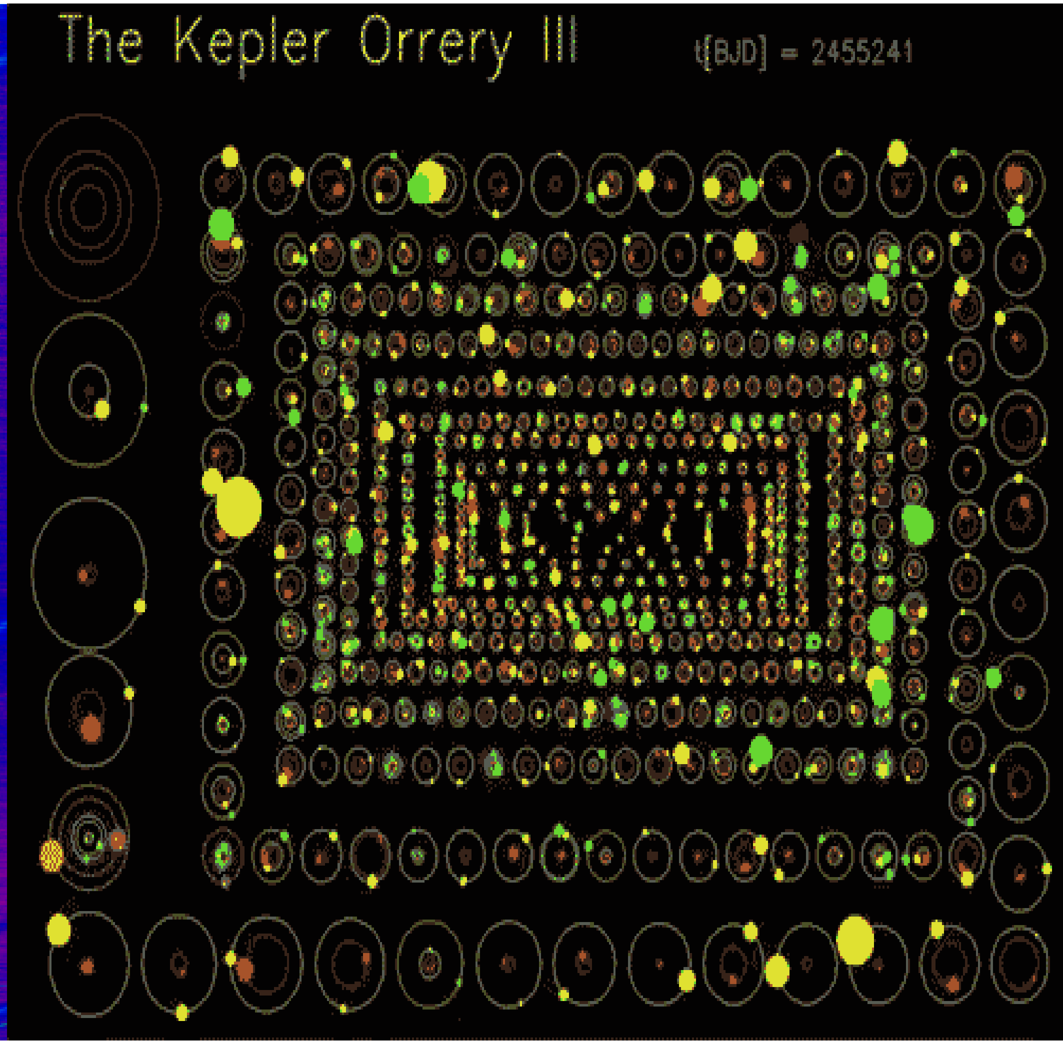
spectral

temporal

ESO



SDSS



NASA

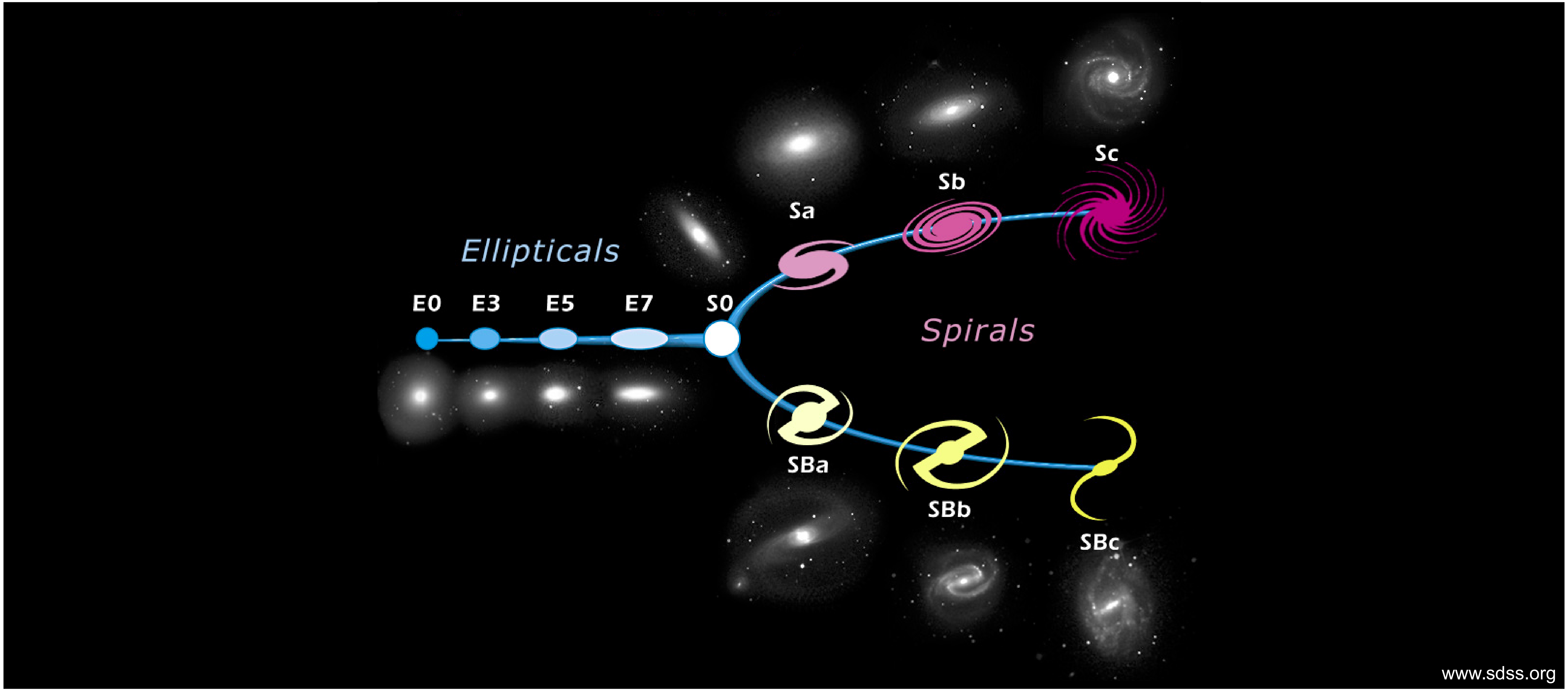
Heidelberg Institute for
Theoretical Studies
HITS

Morphology of Radio Galaxies

how to deal with complex shapes

Simplifying life / putting things in boxes

Morphology of galaxies / Edwin Hubble's classification scheme



Cheap humans

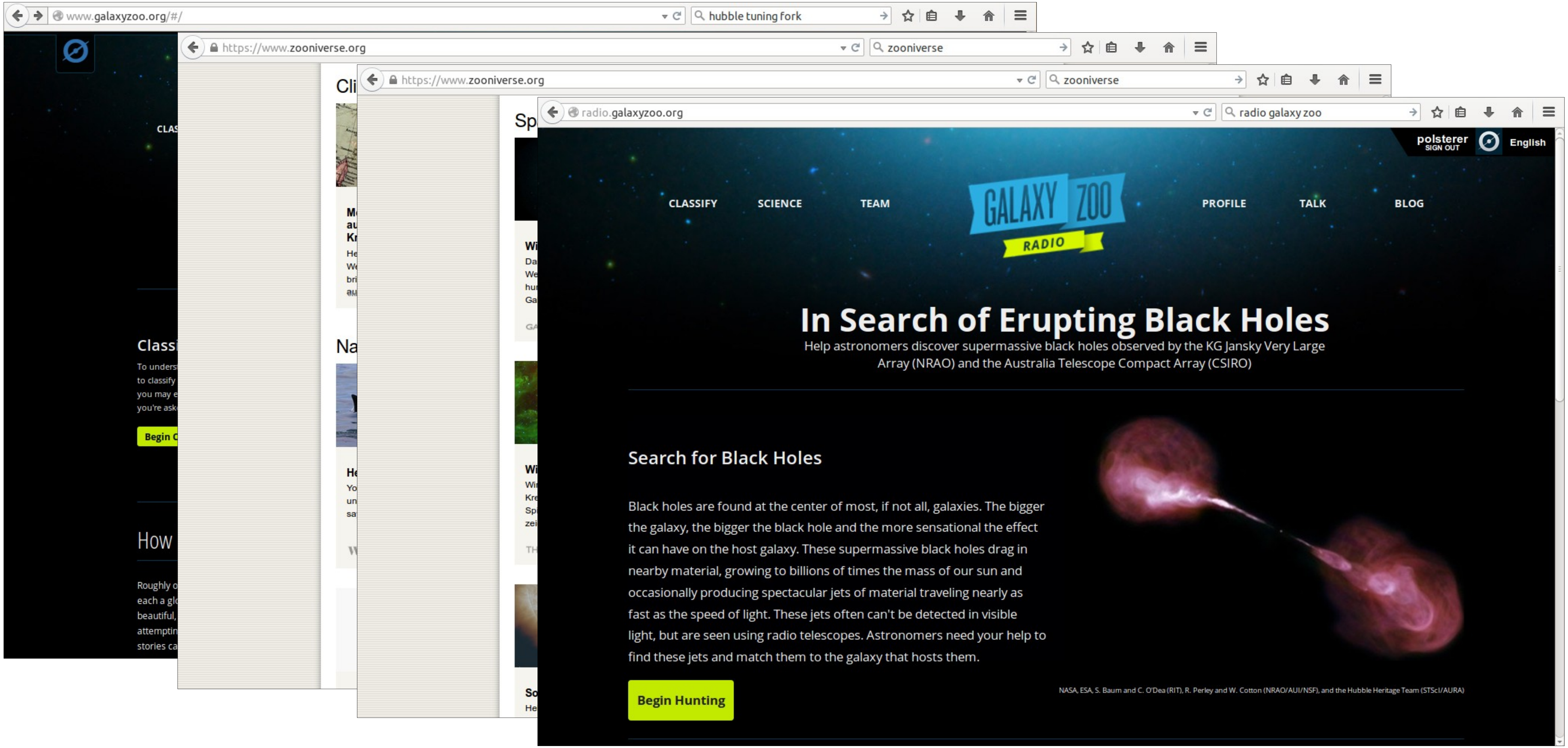
Analyzing 200.000 stellar spectra

- Annie Jump Cannon
  aka “Pickering’s Computers”
- Turning A, B, C, D, E, F, G, …
  into O, B, A, F, G, K, M

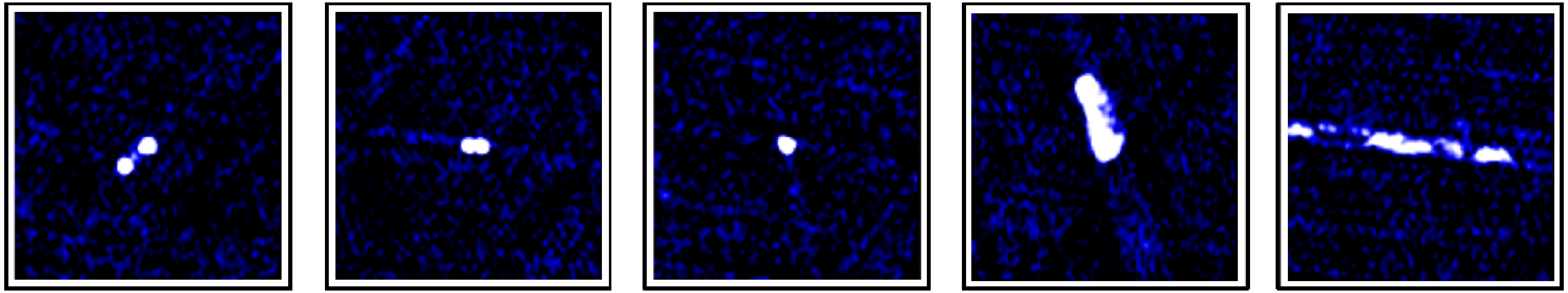
Smithsonian Institution -
Annie Jump Cannon (1863-1941), sitting at desk

What about 50.000.000 images of galaxies?

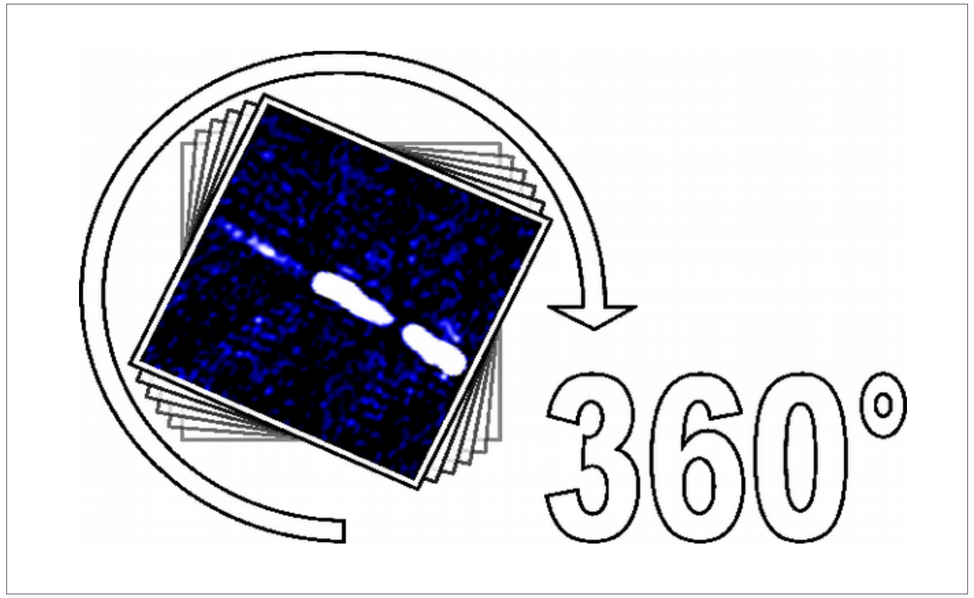
Outsourcing the work / citizen science

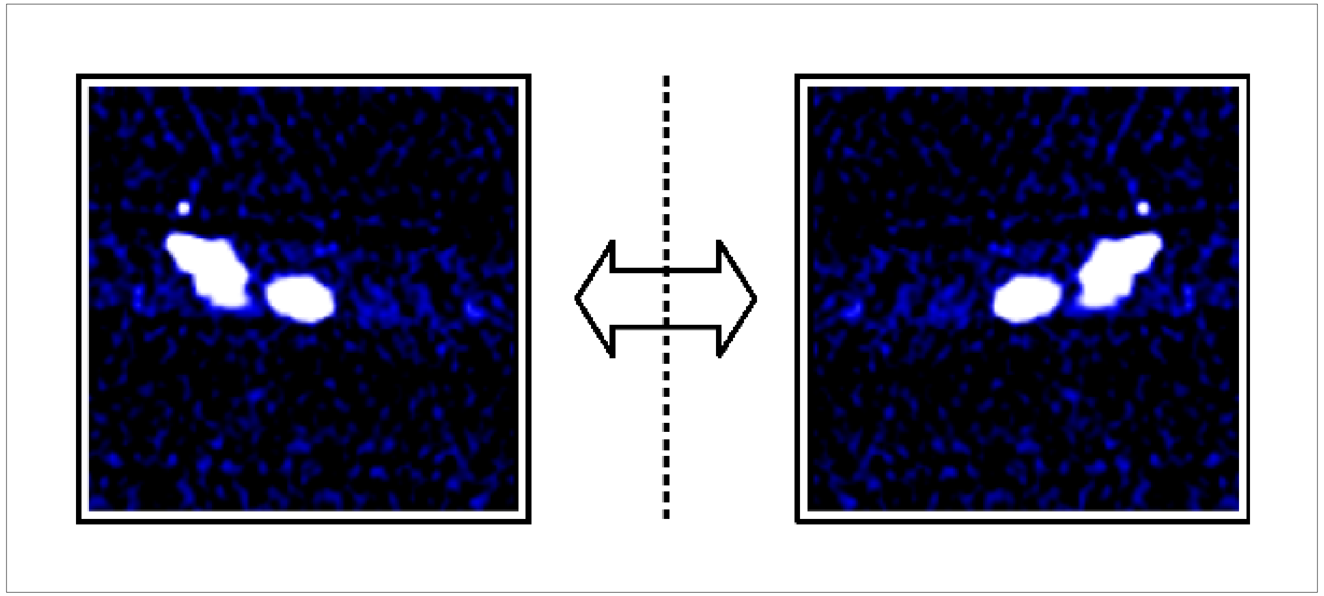


The challenge / Radio GalaxyZoo



rotation / flipping invariant





The solution

have an expert inspect every data-item

vs.

machine learning

- don’t ask scientific questions directly
- ask computers, to structures/sort the data
- do your individual interpretation/analysis

→ use dimensionality reduction

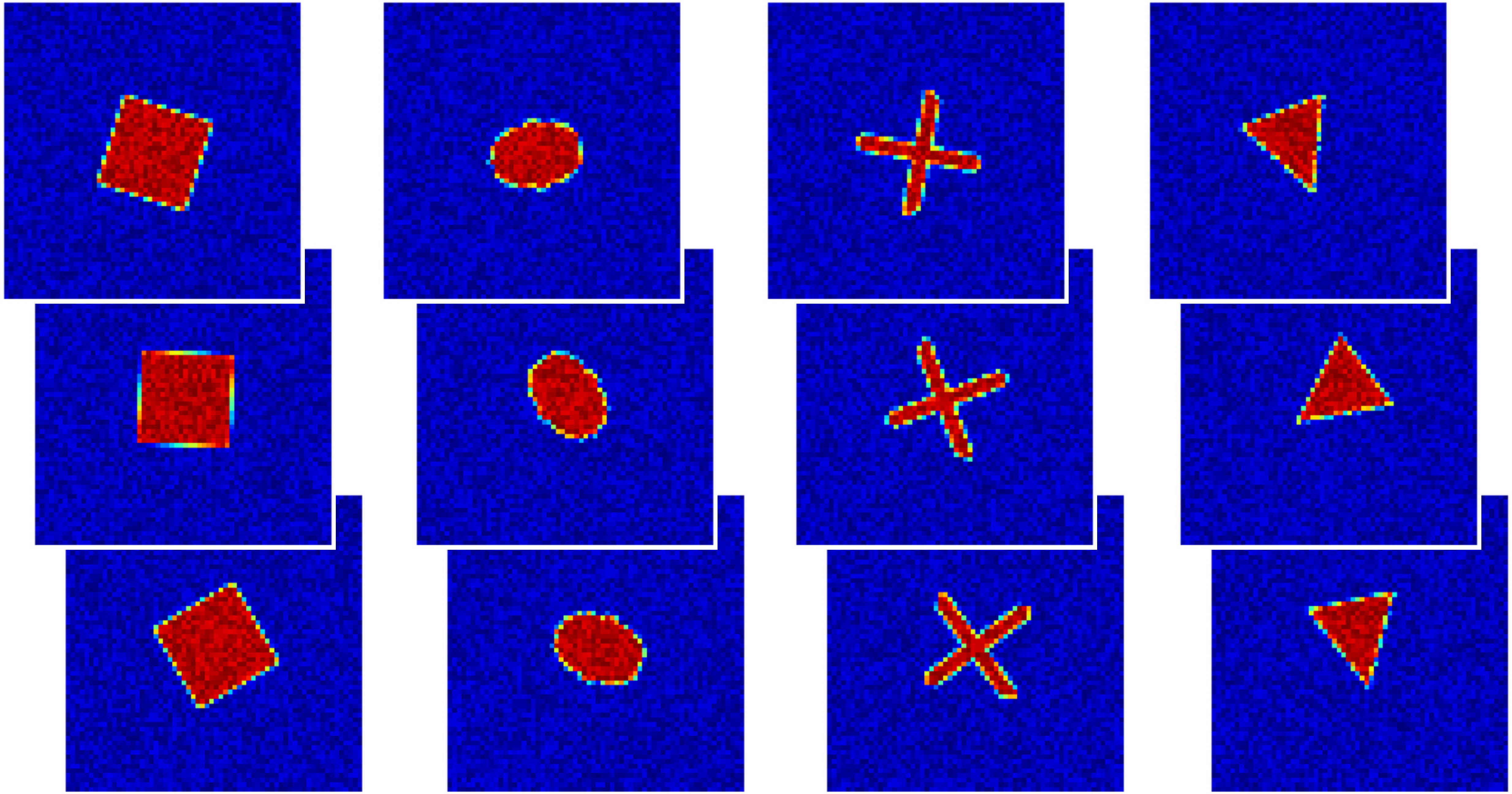
Similarity measure

calculate the pixel based **Euclidean distance**

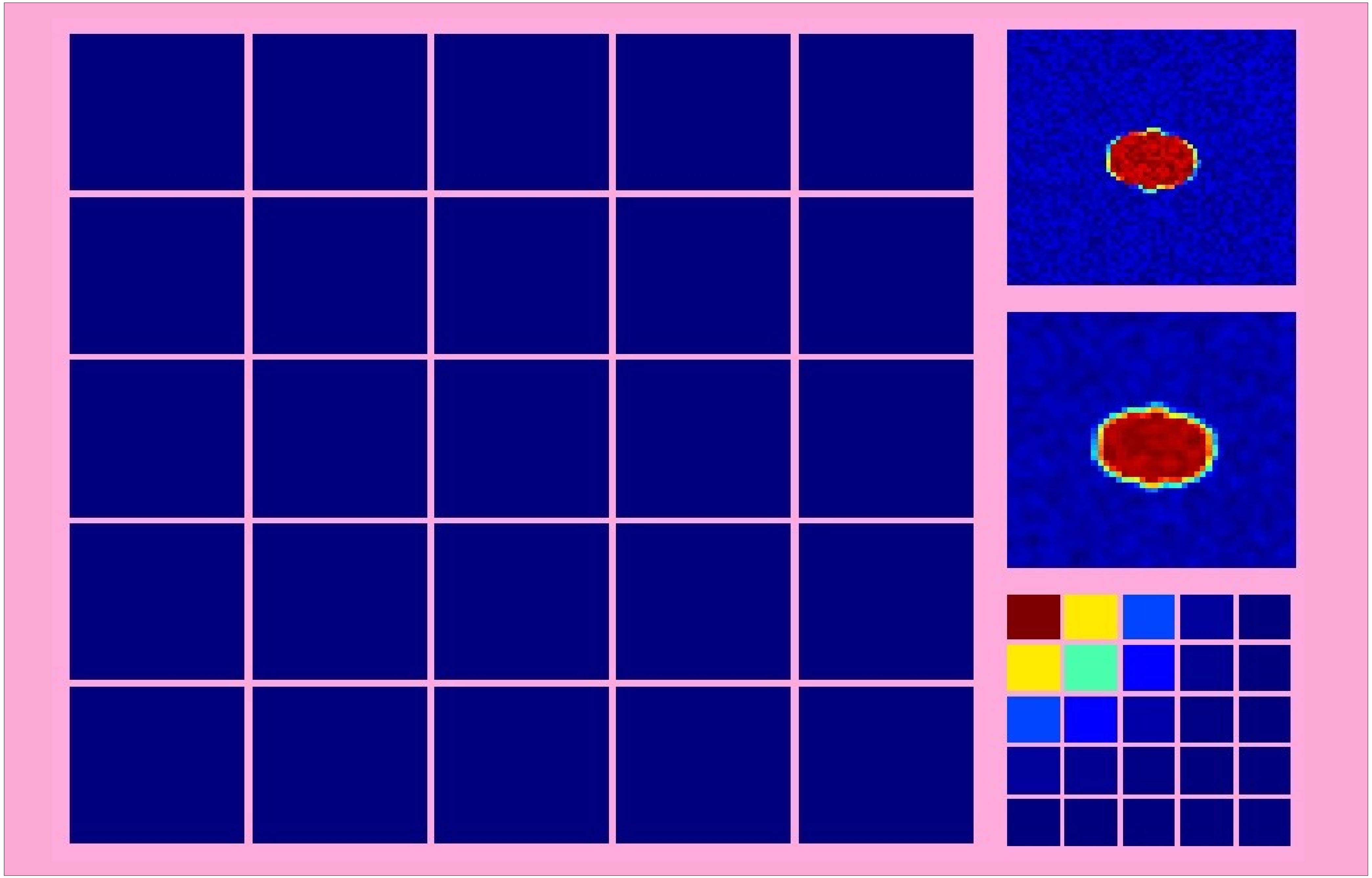
for all possible **rotations**

and find the best matching angle via **minimization**

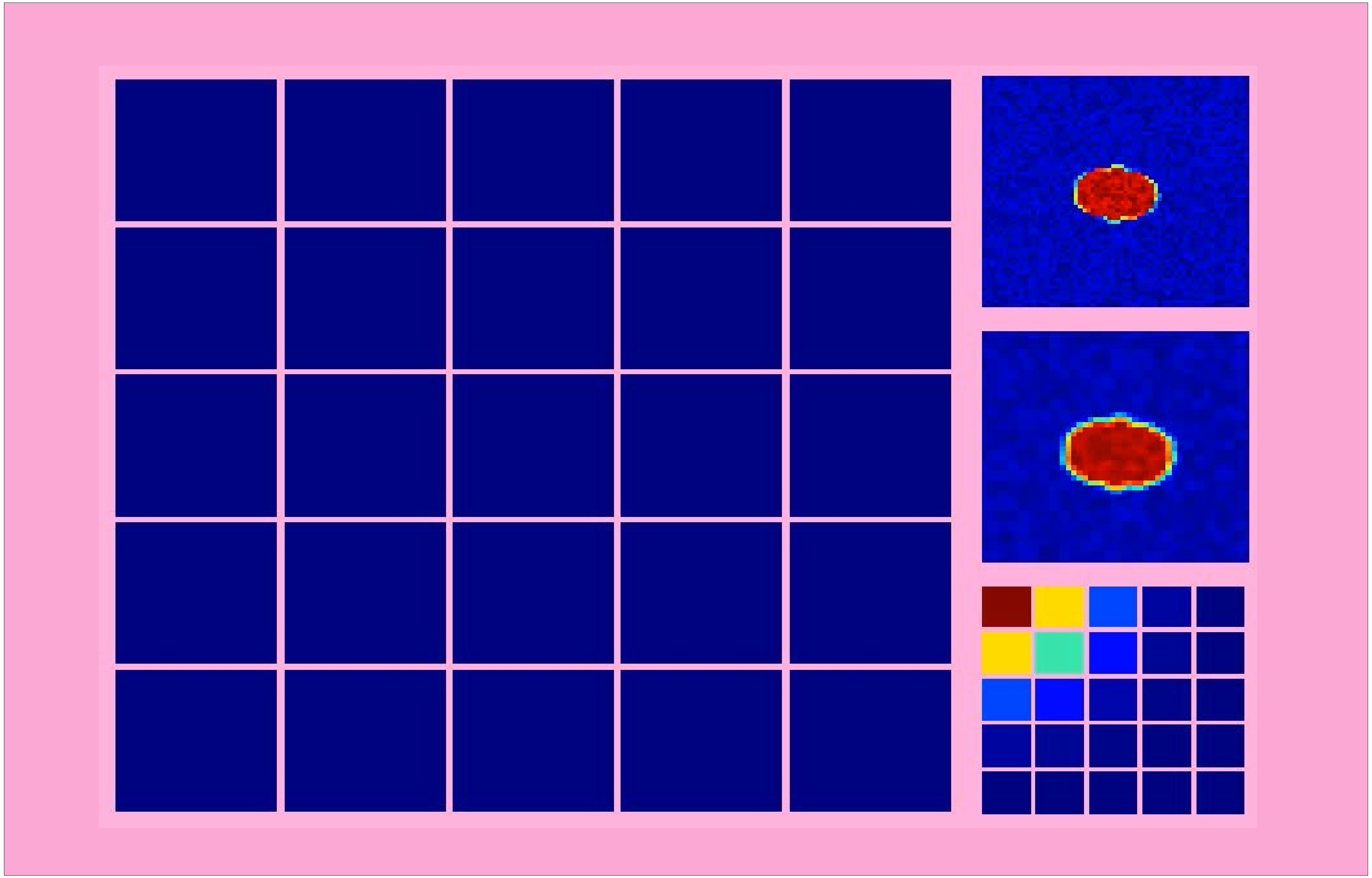
Example self-organizing map



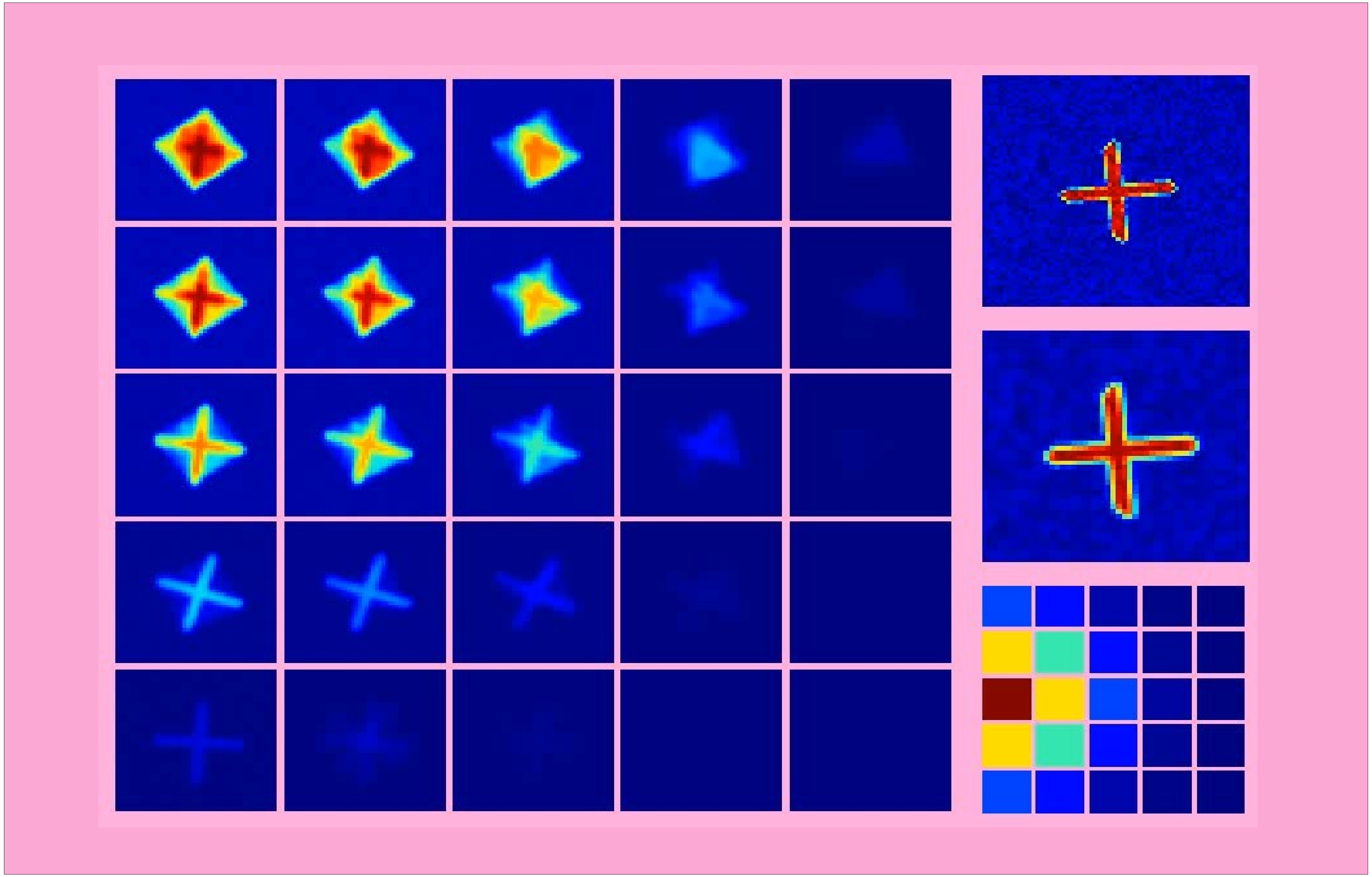
Example self-organizing map

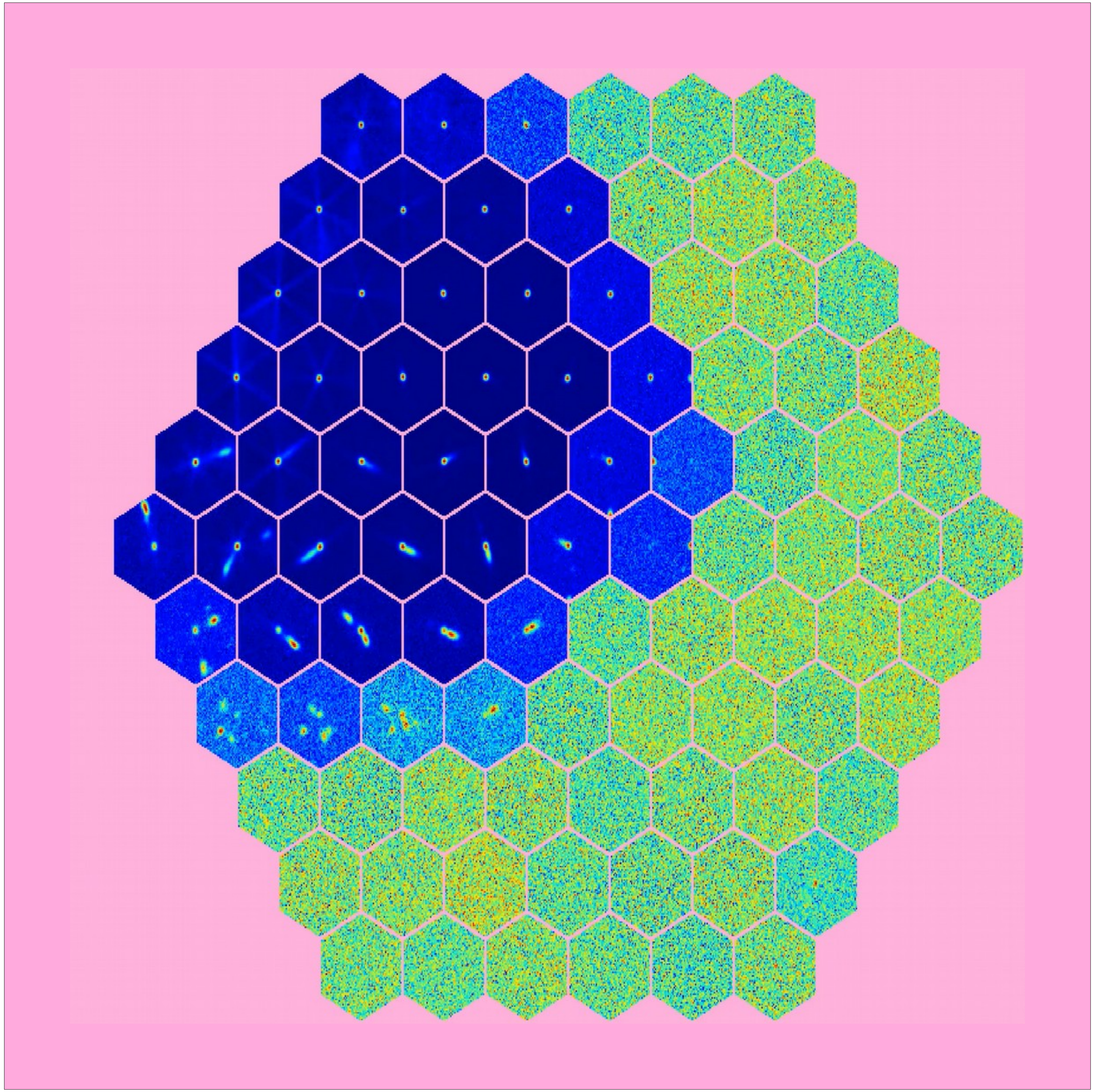


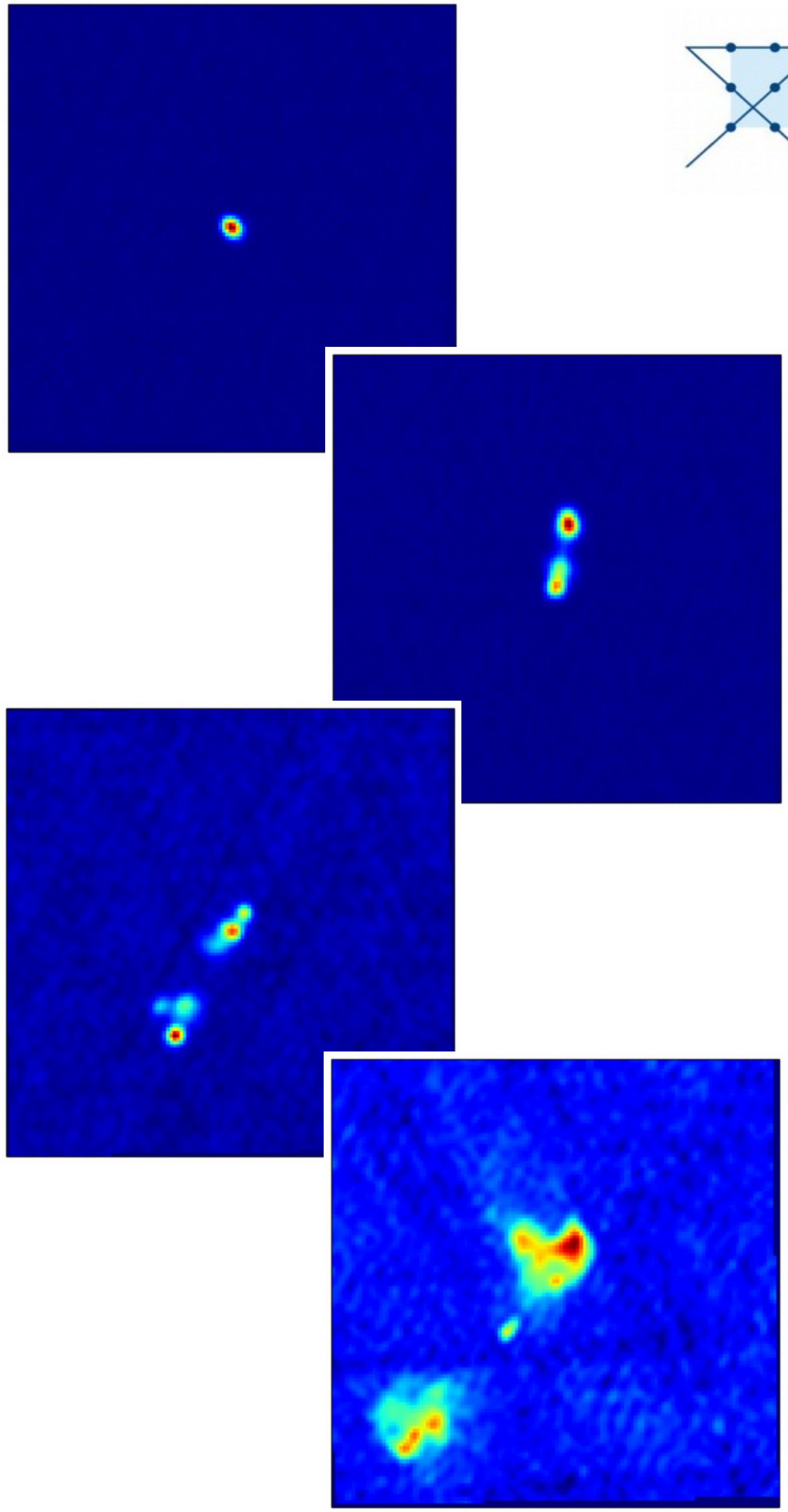
Example self-organizing map



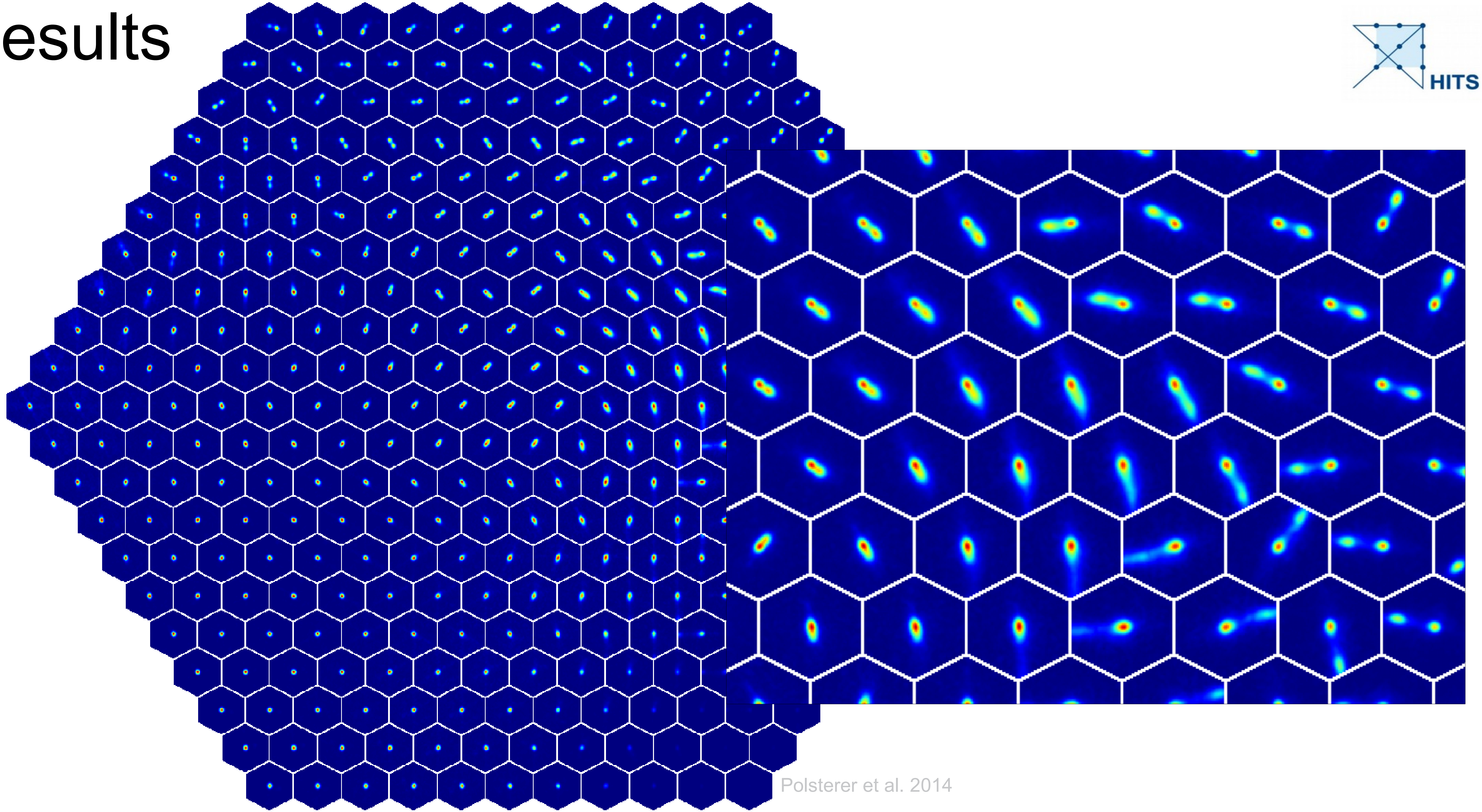
Example self-organizing map





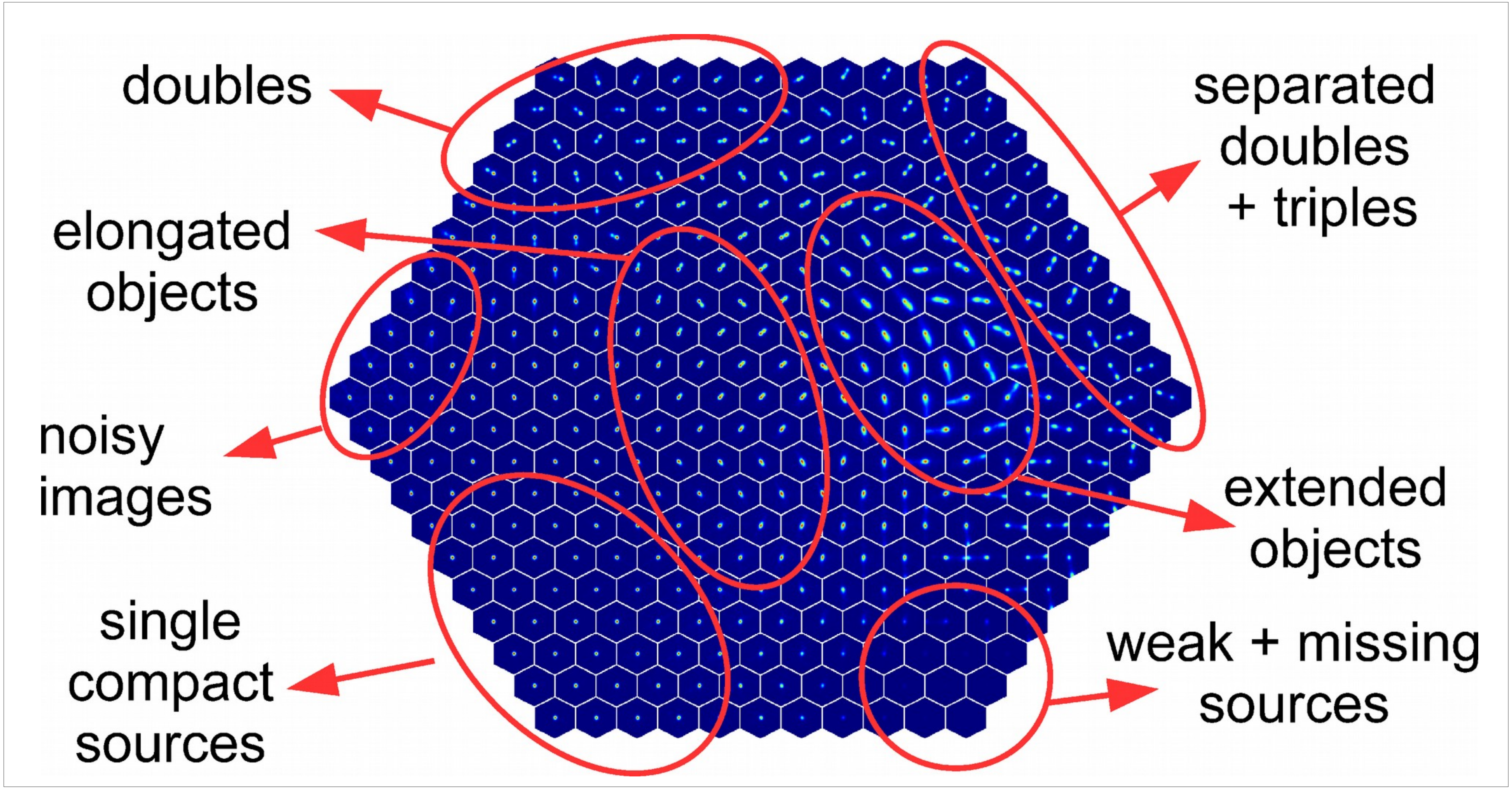


Results

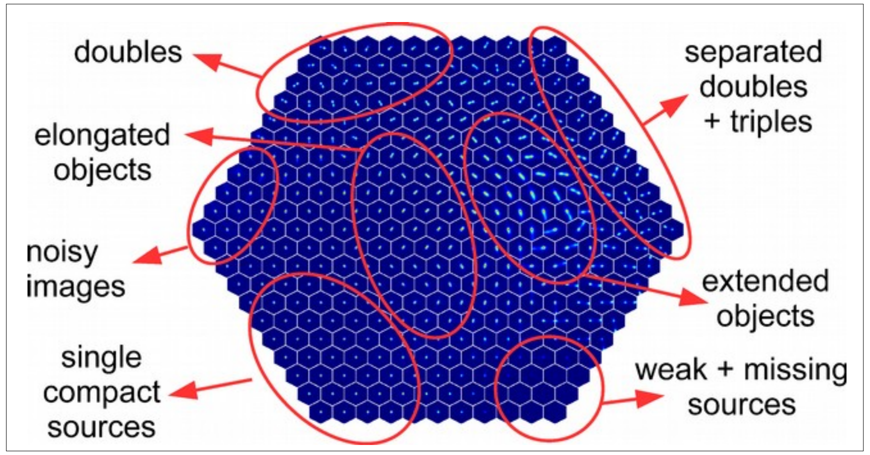


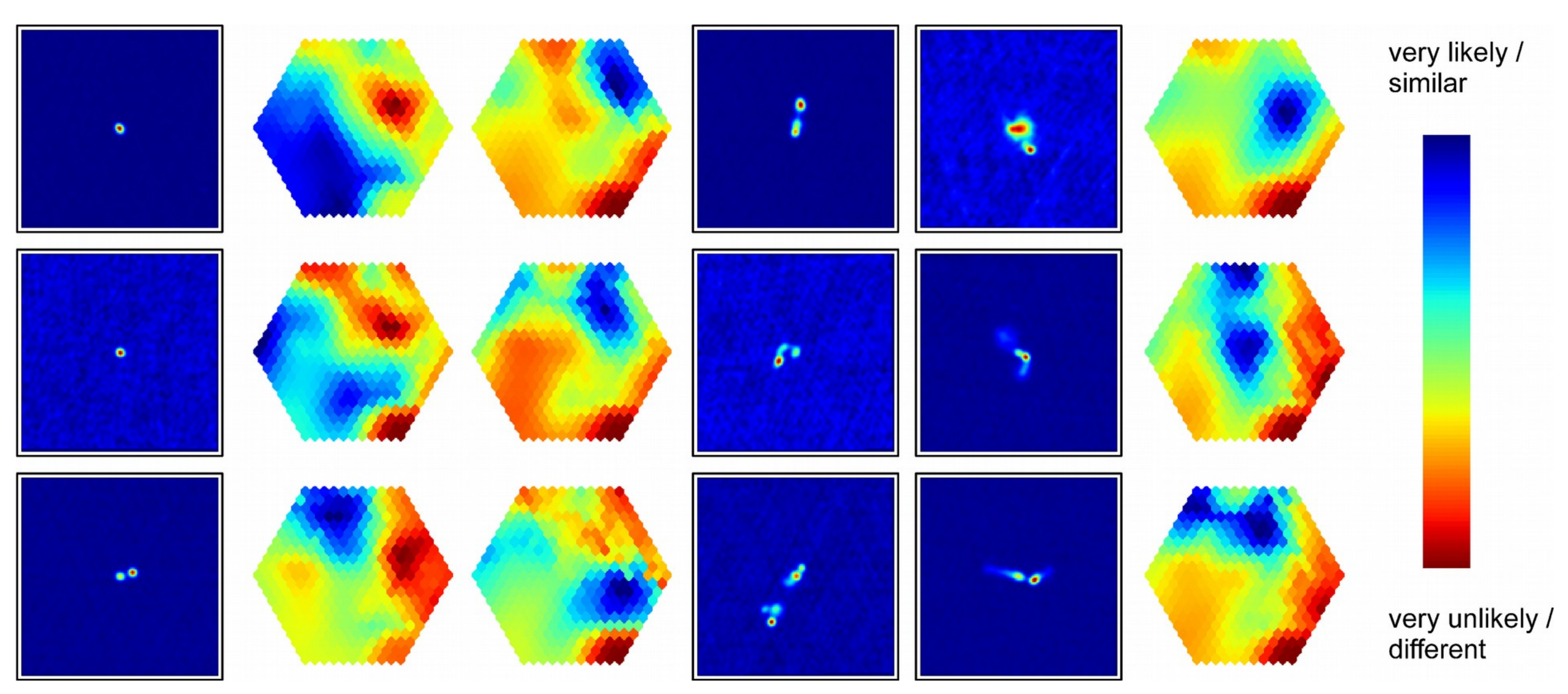
Polsterer et al. 2014

Results


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

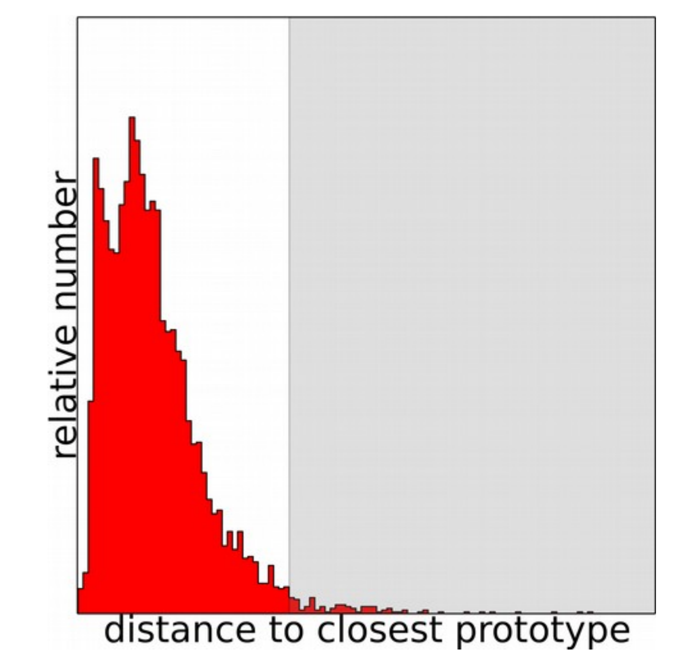
Results

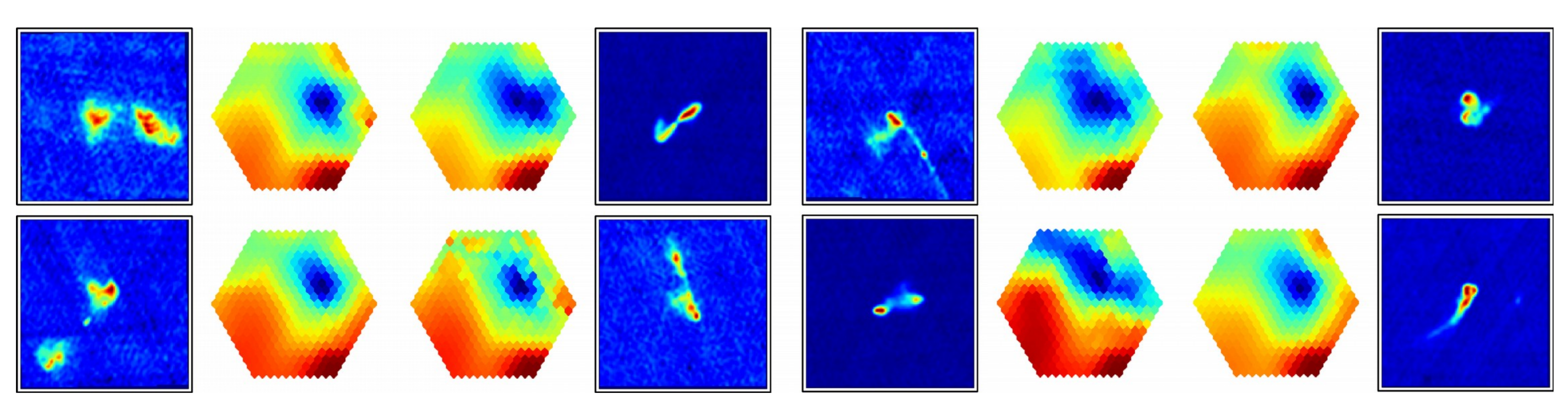

doubles
elongated objects
noisy images
single compact sources
separated doubles + triples
extended objects
weak + missing sources


very likely / similar
very unlikely / different

Results

select **outliers** based on distribution of distances





Result

1 master-student + 4 GPUs [6 month]

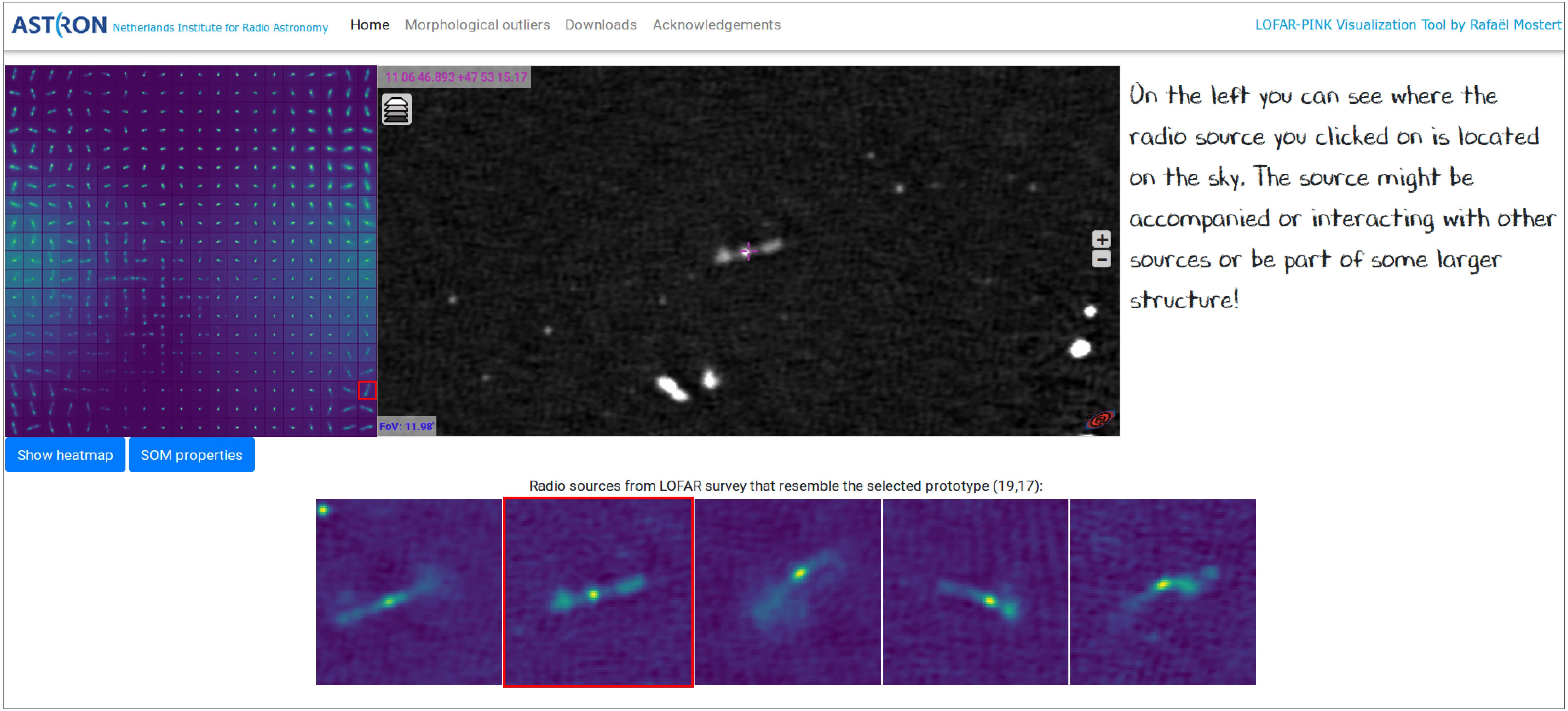
- → catalog of morphologies for 1,000,000 sources in FIRST + IR counterpart analysis

vs.

10,000 Volunteers + 4 PostDocs [4 years]

- → catalog of morphologies for 200,000 sources in FIRST + IR counterpart analysis

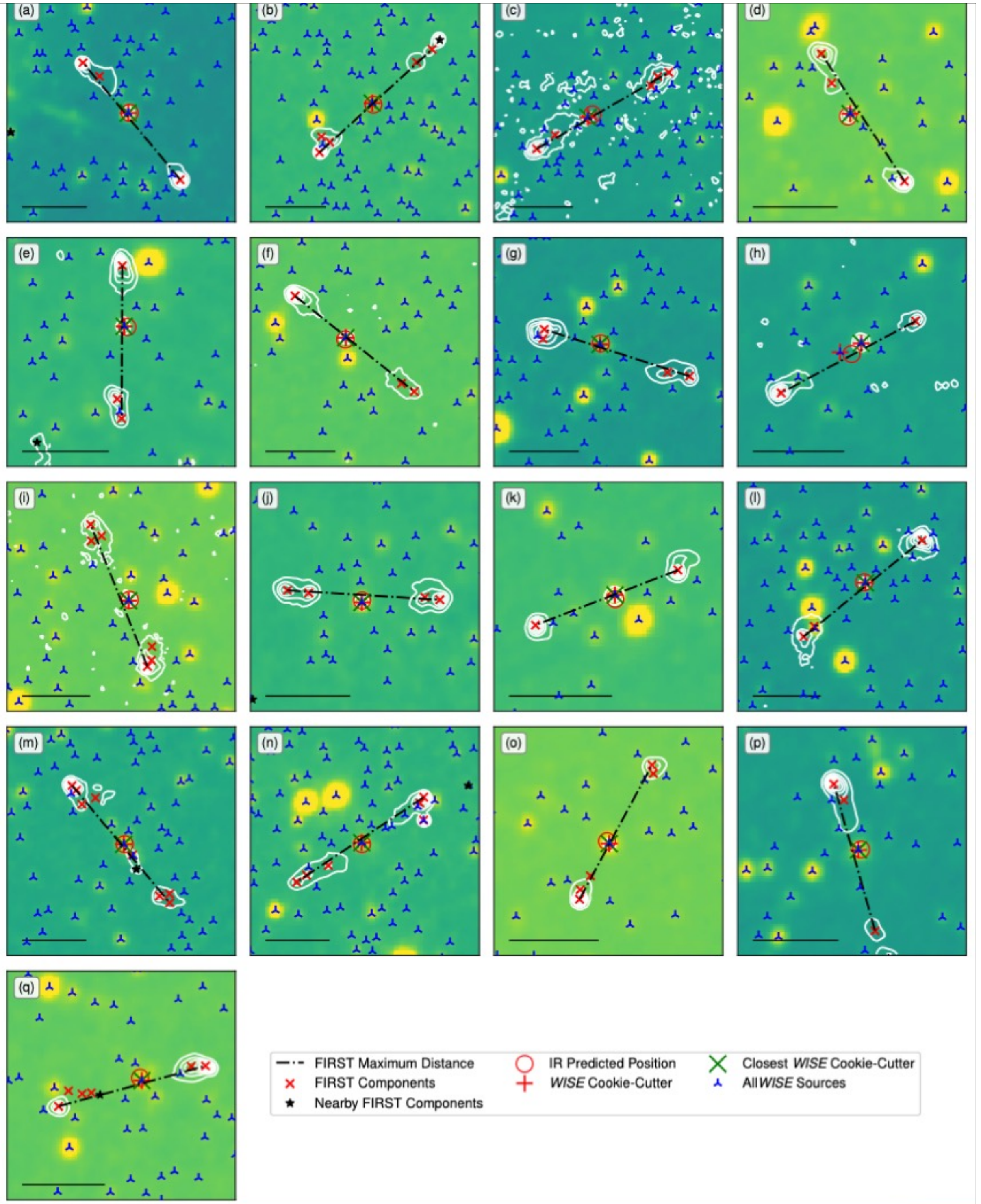
LOFAR web-tool



Mostert 2017

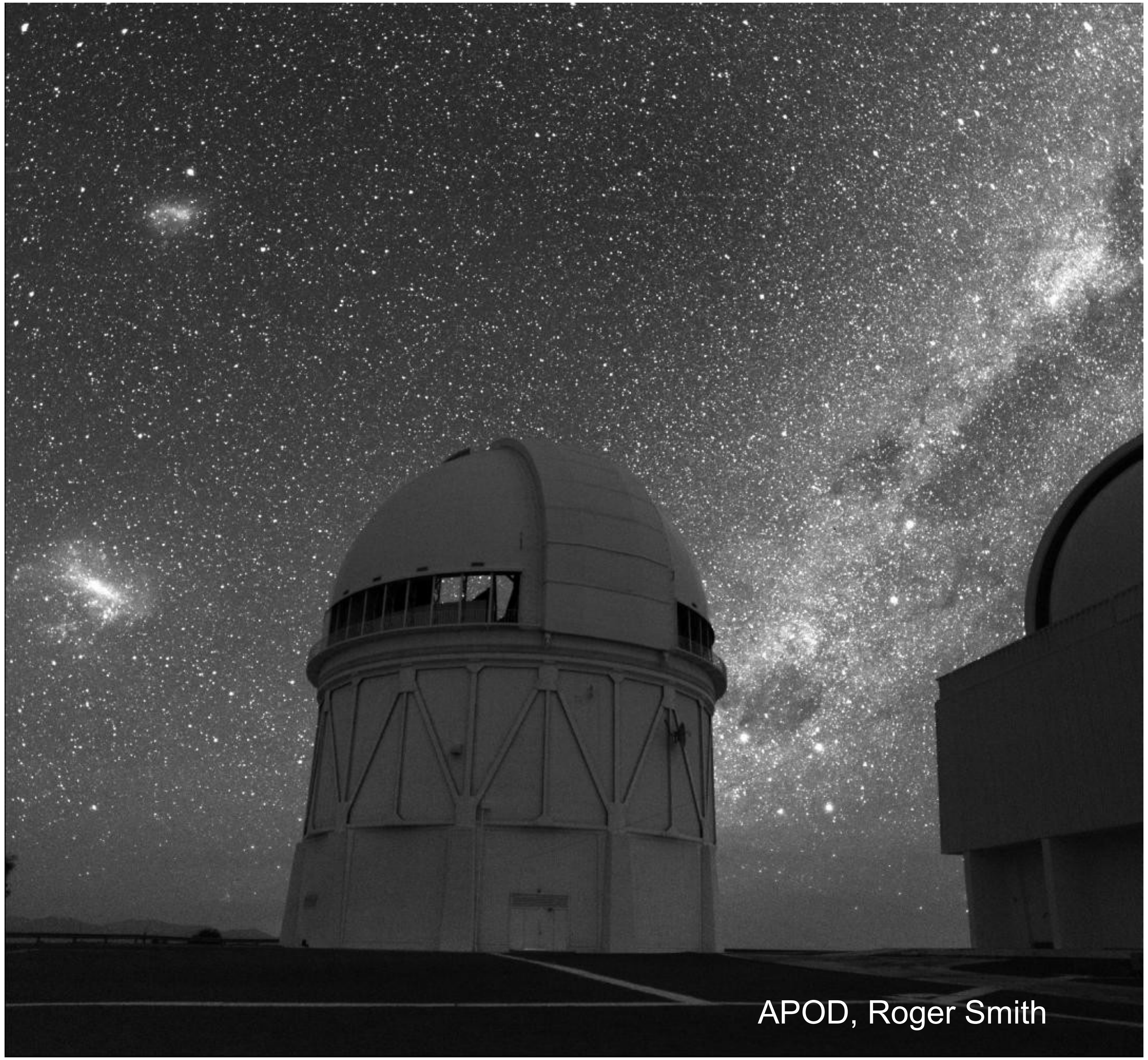
Start to go GRG hunting

- Cross reference to SDSS for redshifts
  - Only using spec-z, have photo-zs as well
- 17 GRGs between 0.7 – 1.5 Mpc
  - Neuron FoV comes into play
- Model was not *trained* for GRGs
  - Just a product of model understanding the structure of data

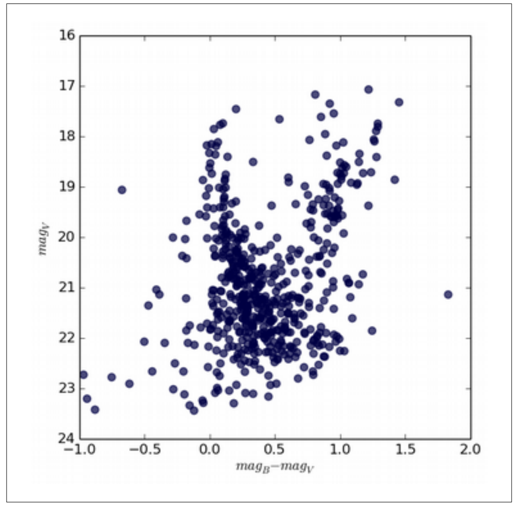


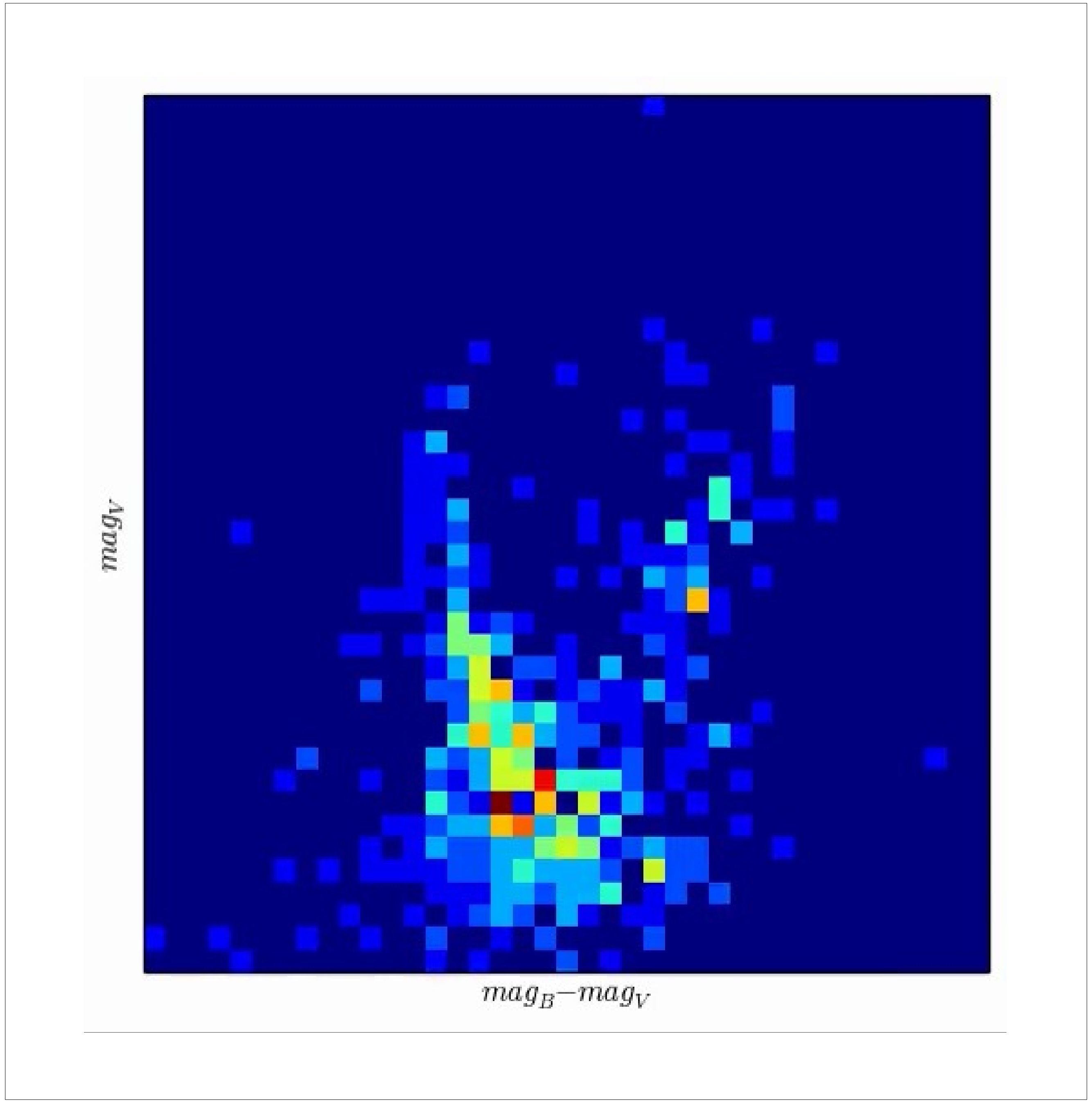
Tim Galvin CSIRO

Starformation history

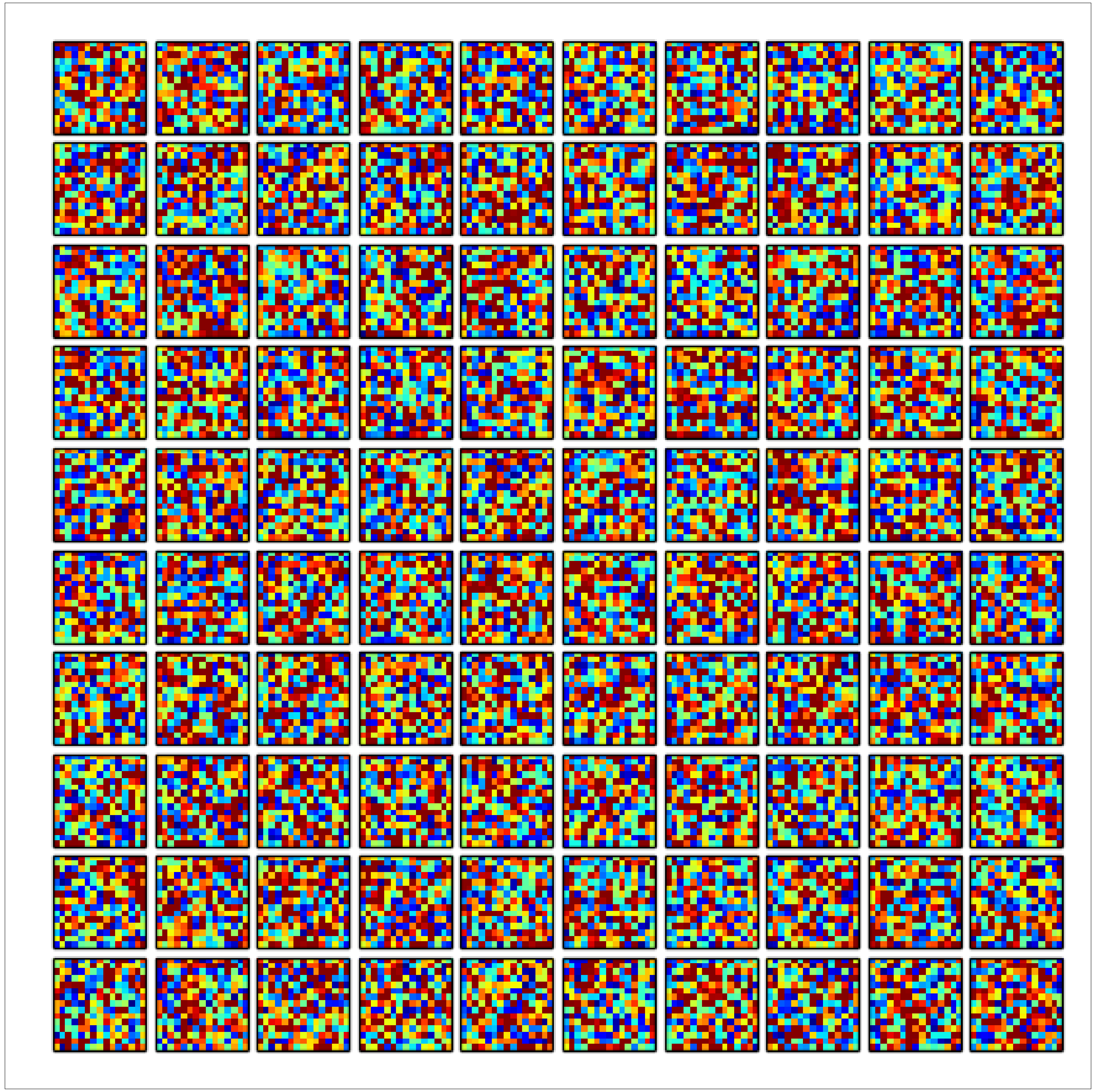


Analysis of stellar cluster



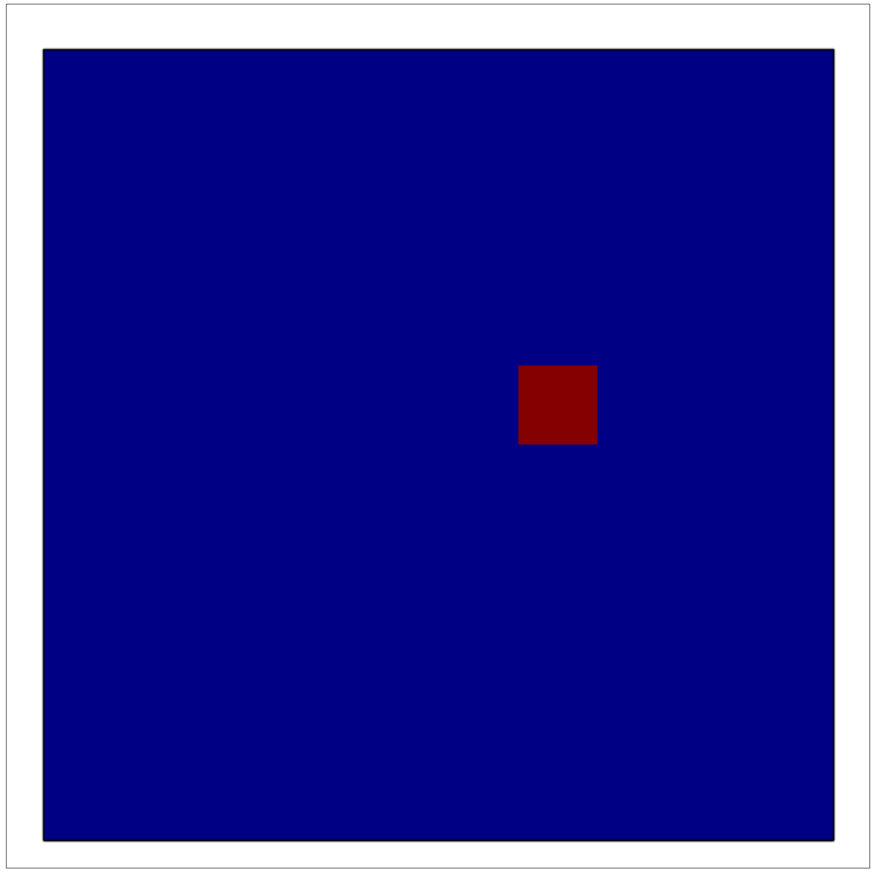


Dimensionality reduction

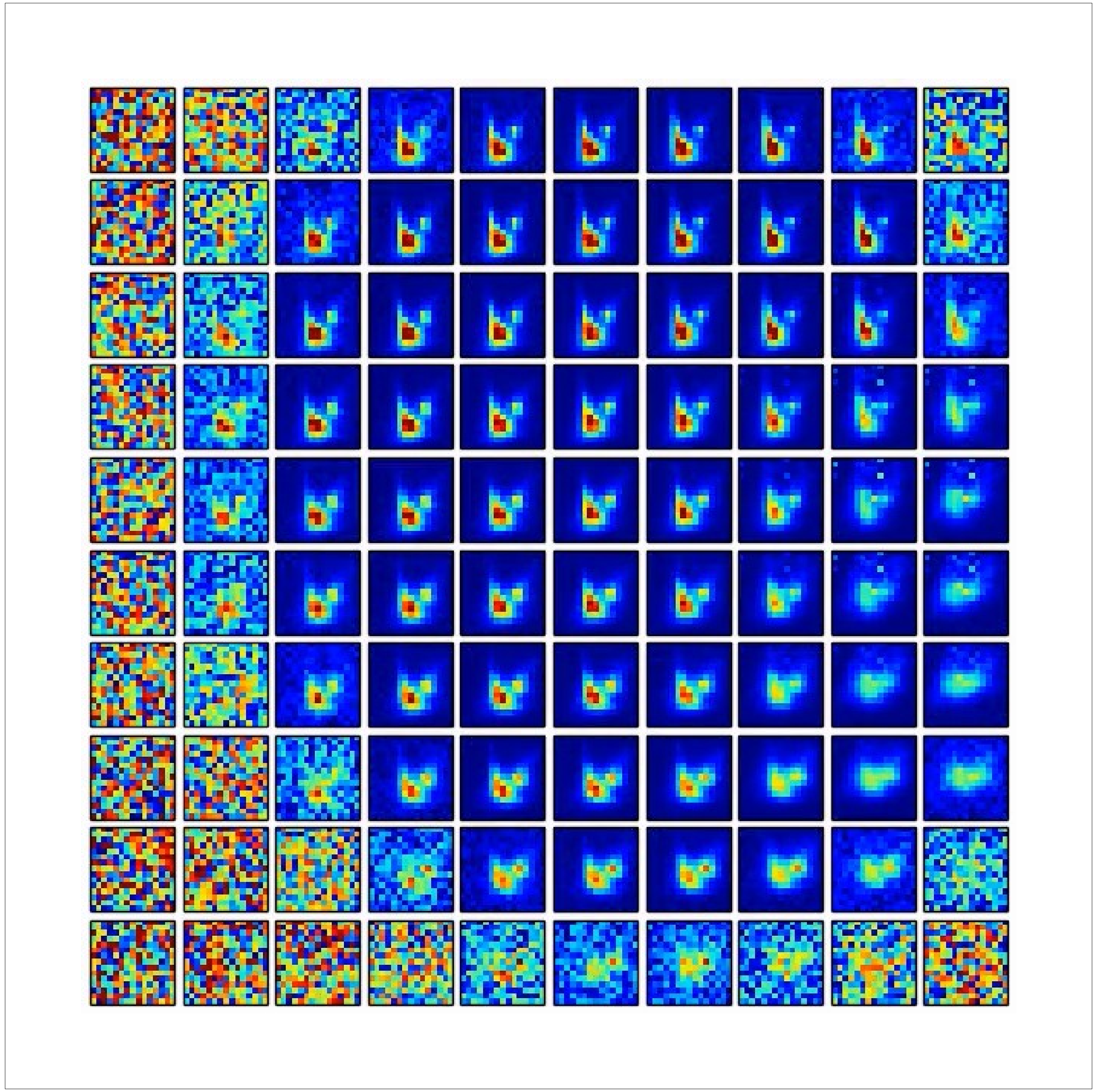


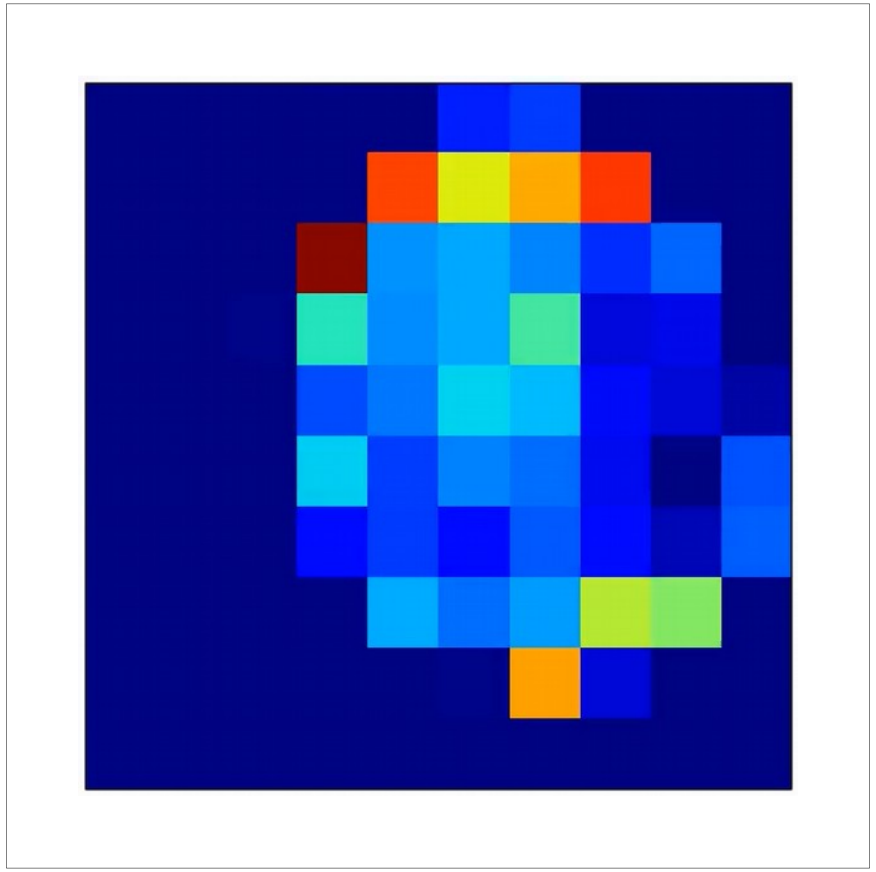
$$\Delta(A,B)=\sqrt{\sum_{x=1}^{D_x}\sum_{y=1}^{D_y}(A_{xy}-B_{xy})^2}$$

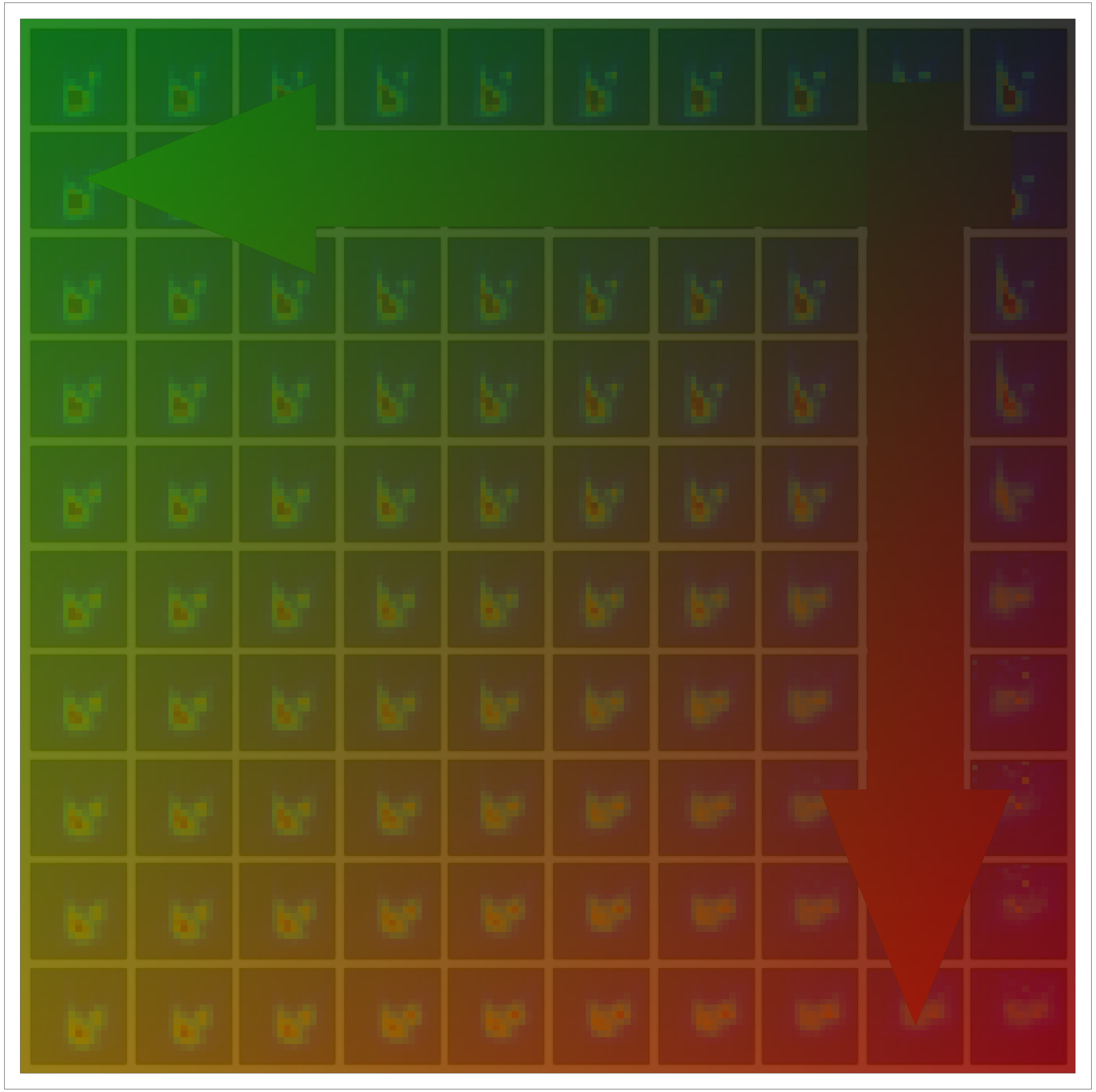
$$\Delta(A,B)=\sqrt{\sum_{x=1}^{D_x}\sum_{y=1}^{D_y}\frac{\left(\frac{A_{xy}}{N_A}-\frac{B_{xy}}{N_B}\right)^2}{\frac{A_{xy}}{N_A^2}}}$$



Dimensionality reduction

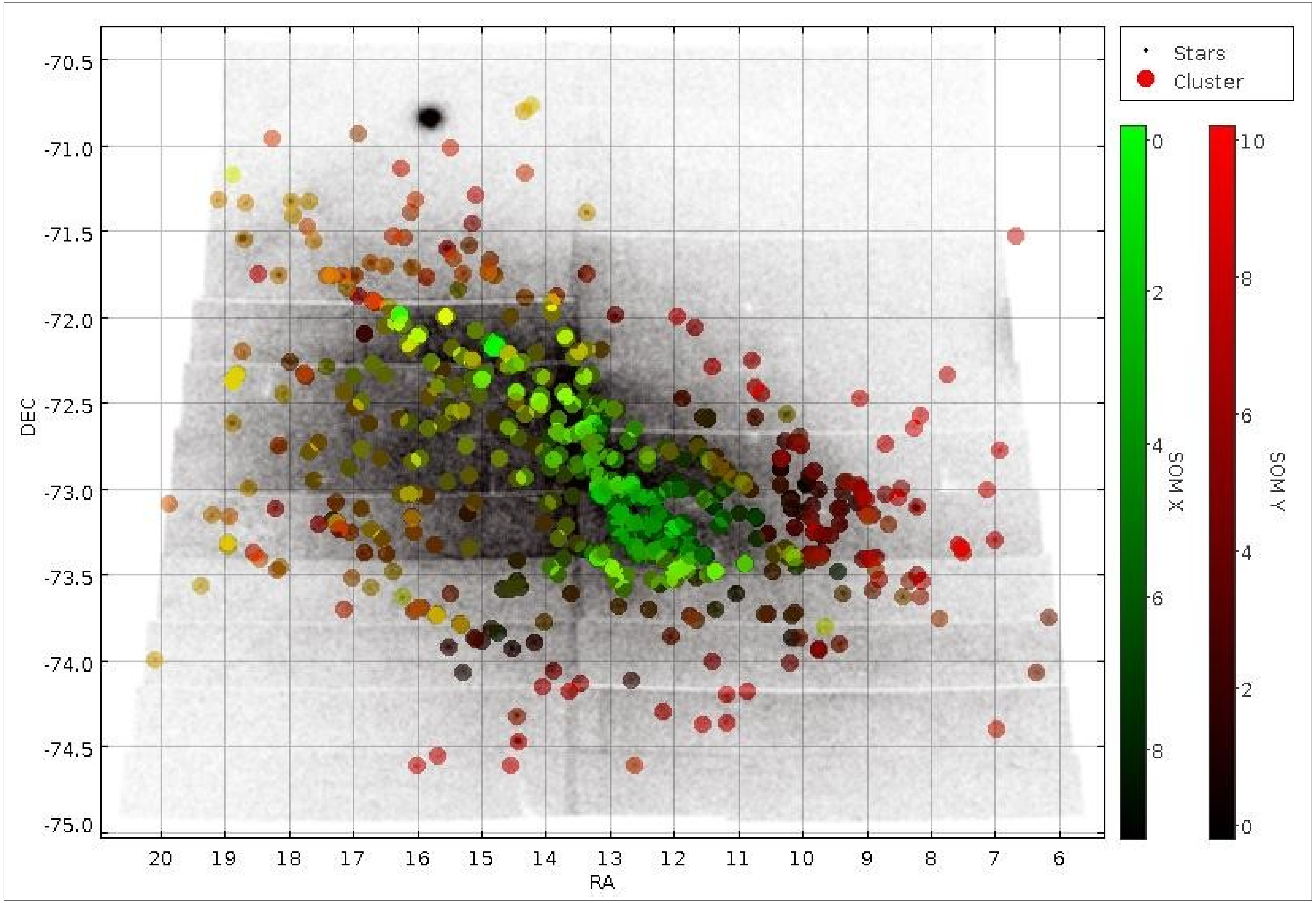




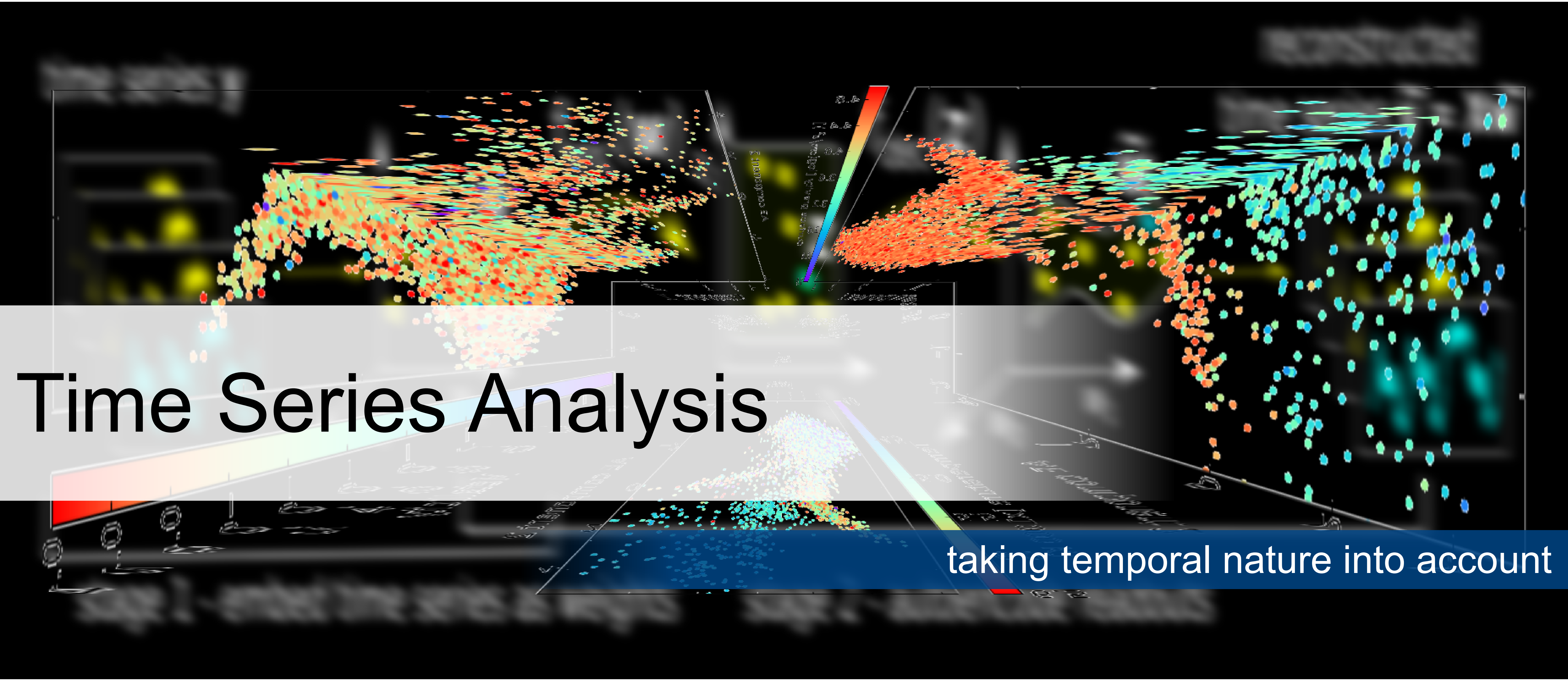


what is it
good for?

Results



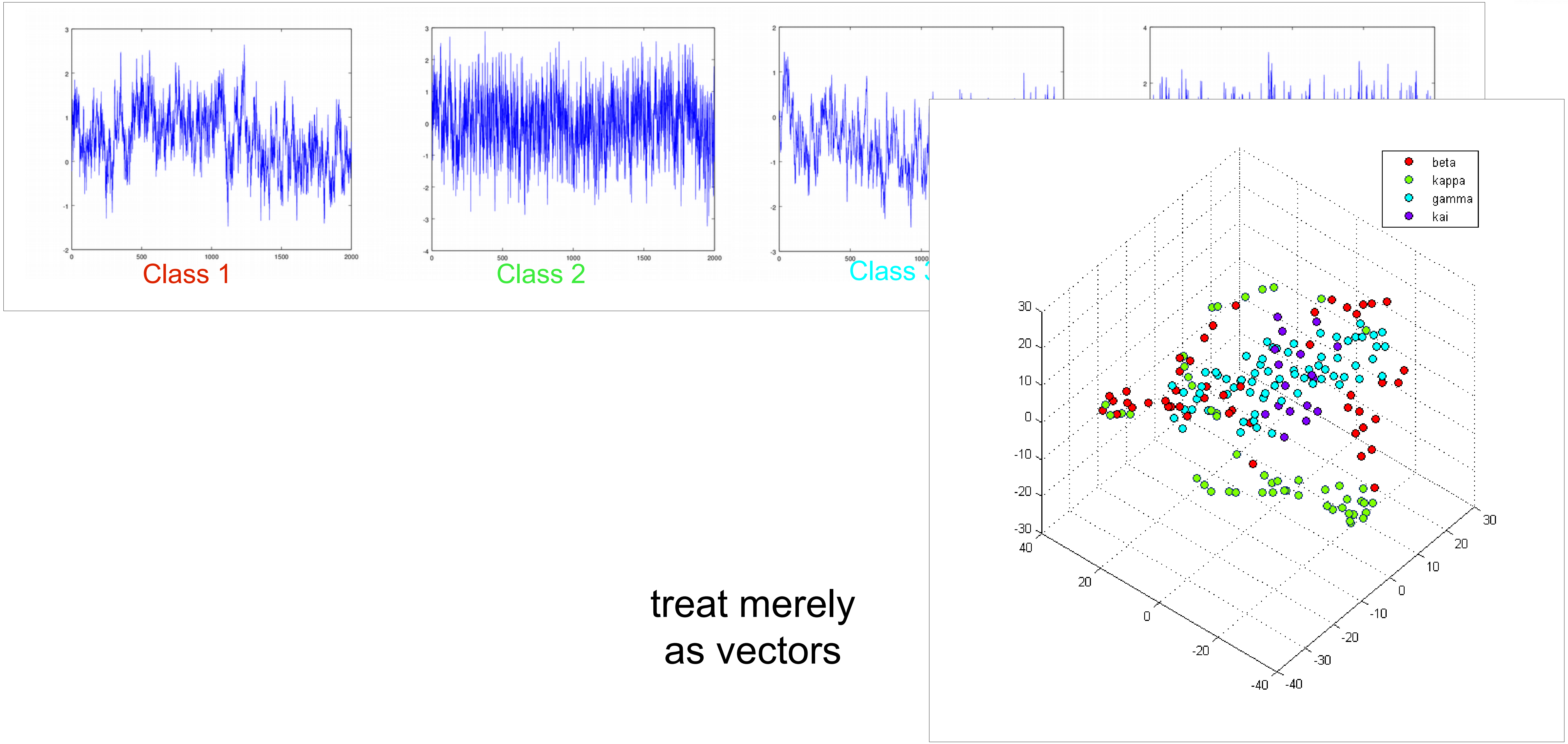
Stars
Cluster
DEC
RA
SOM X
SOM Y
-70.5
-71.0
-71.5
-72.0
-72.5
-73.0
-73.5
-74.0
-74.5
-75.0
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6
0 2 4 6 8
10 8 6 4 2 0



Time Series Analysis

taking temporal nature into account

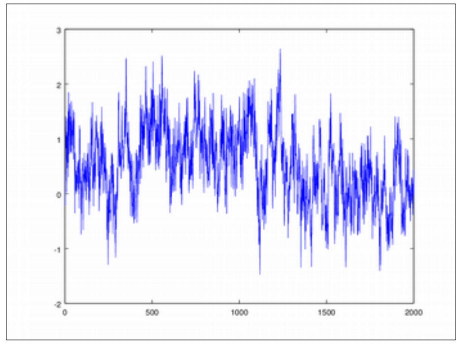
Time series analysis

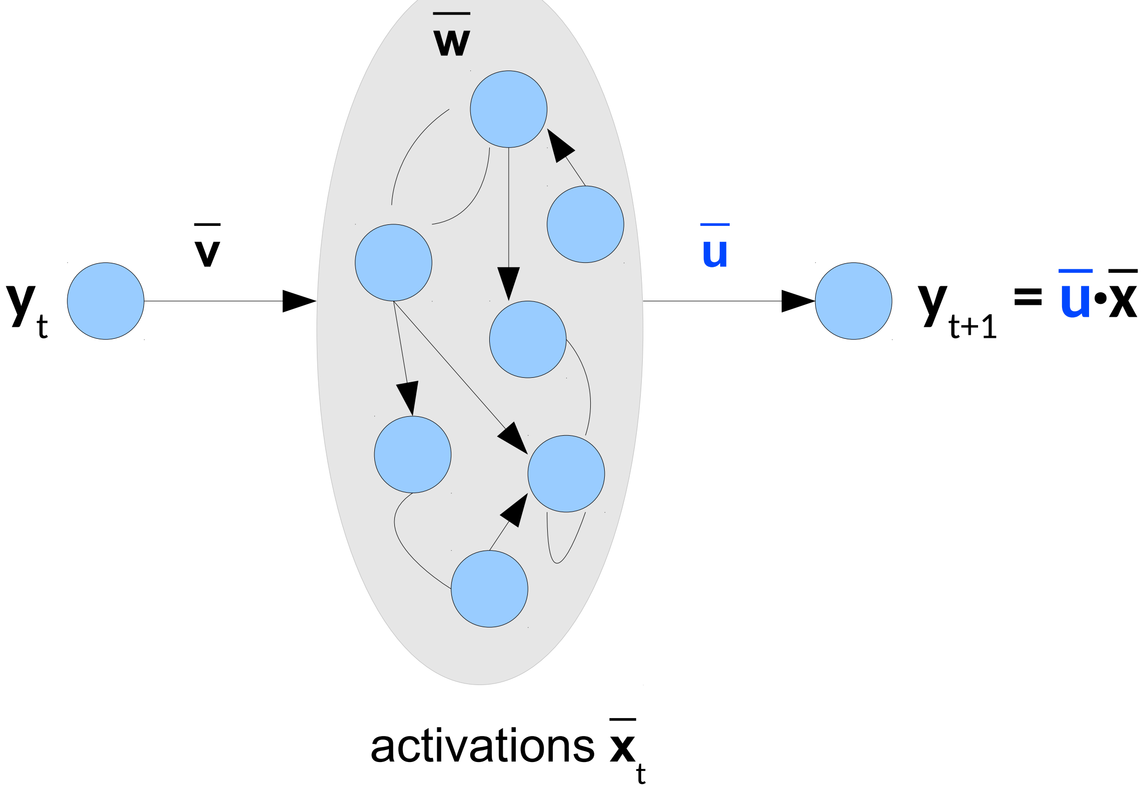


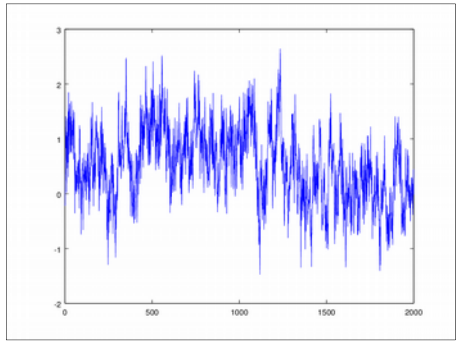
treat merely
as vectors

Recurrent neural network

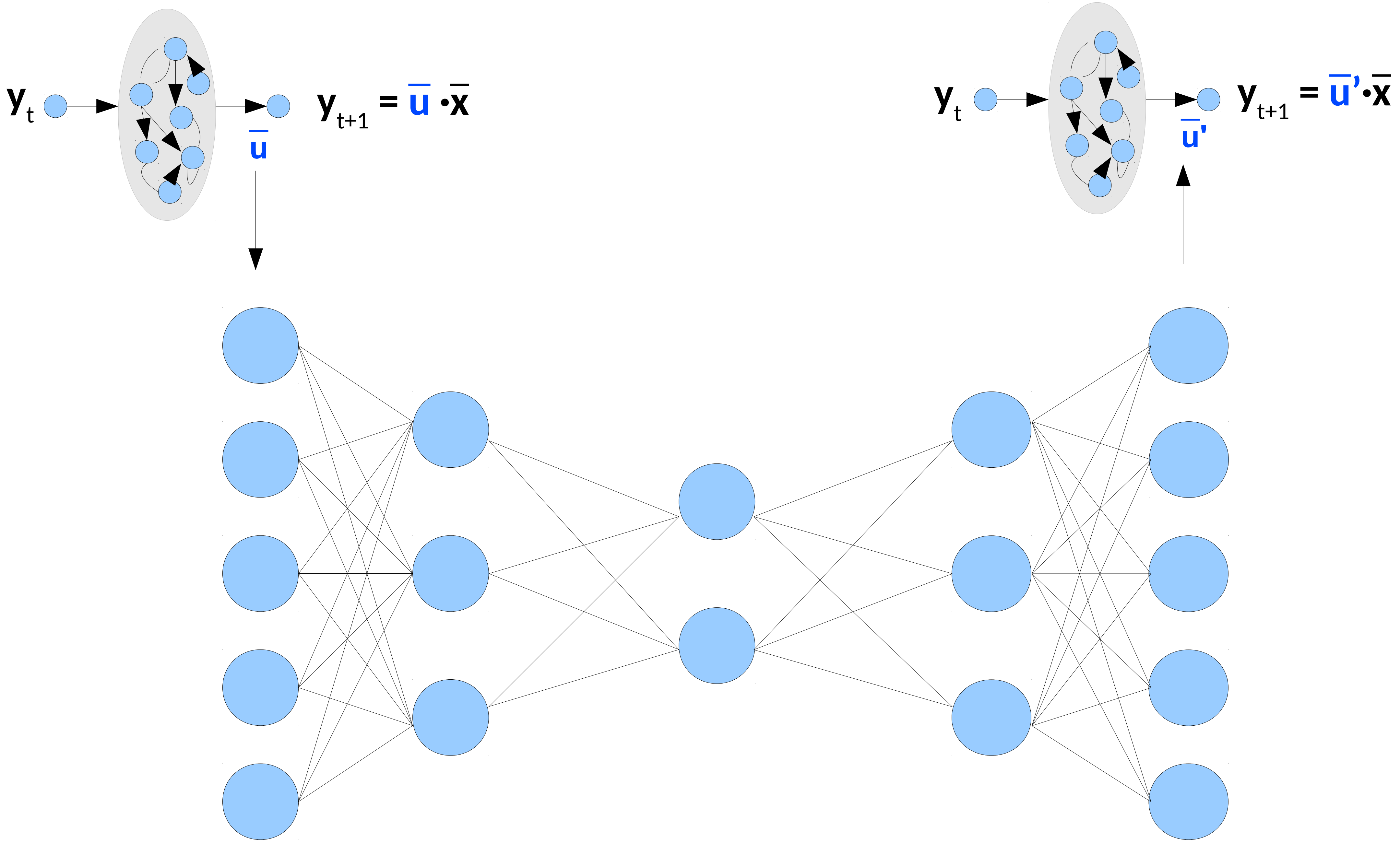
take temporal nature into account



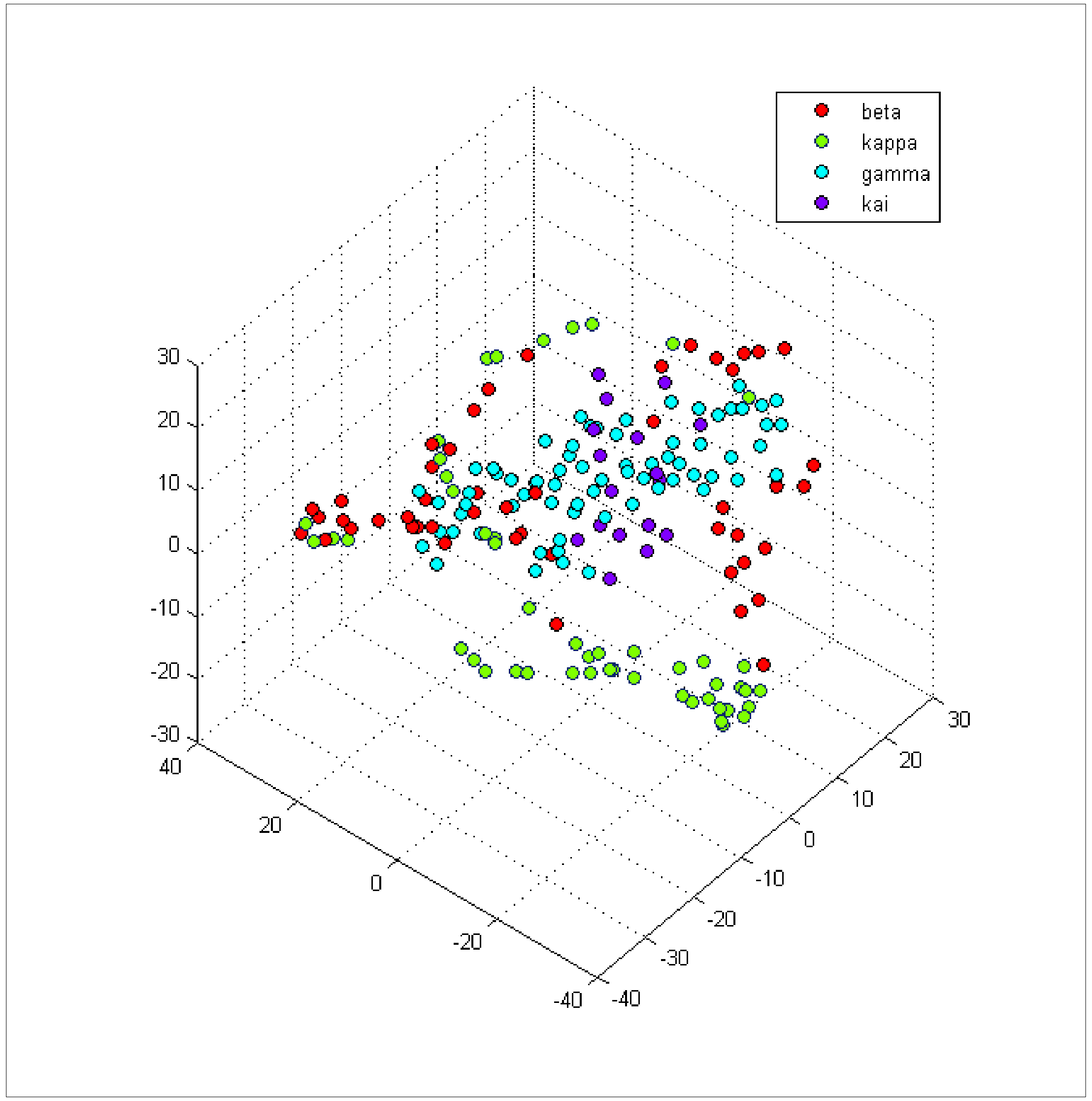


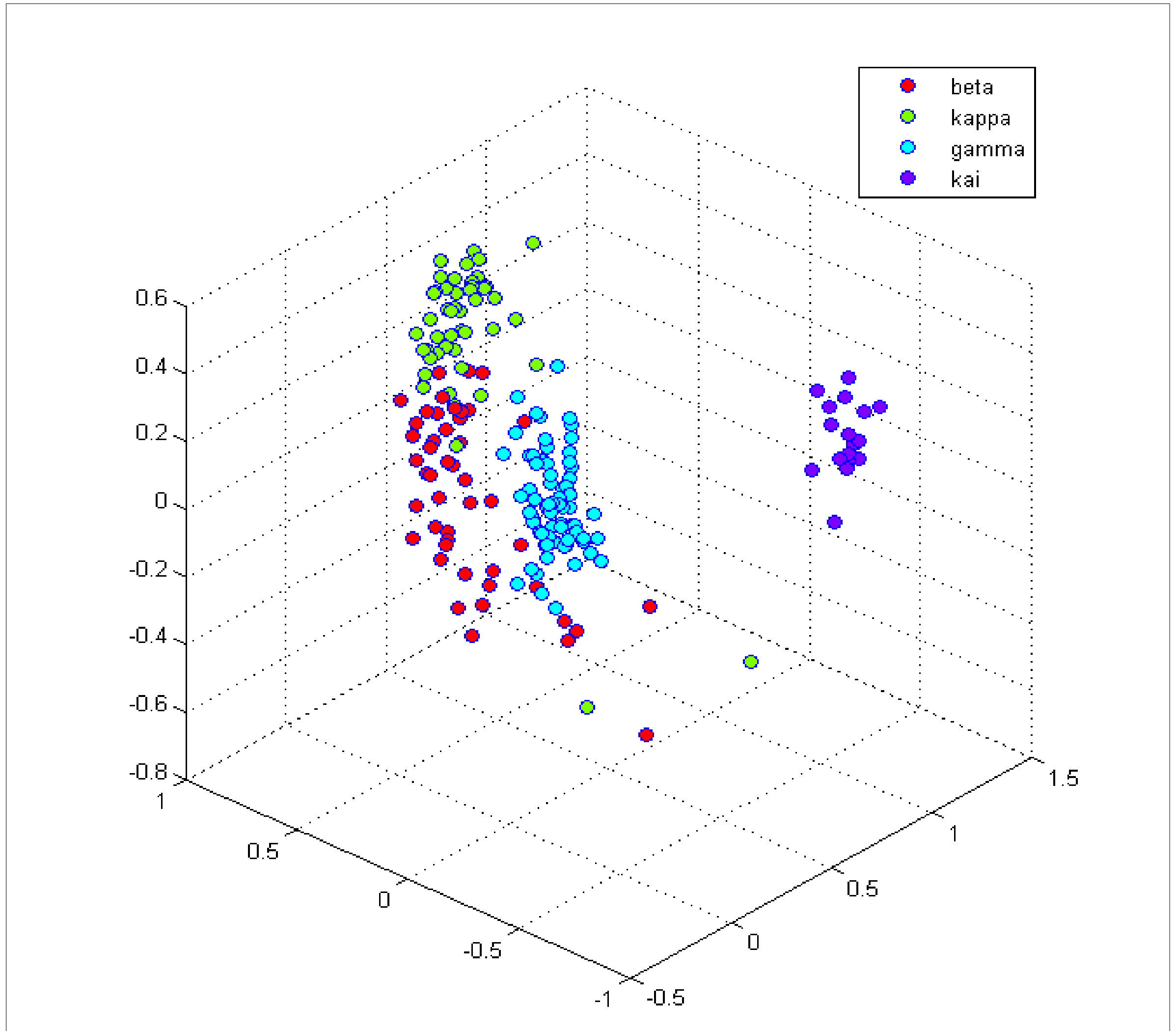


ECN + Autoencoder



Time series analysis

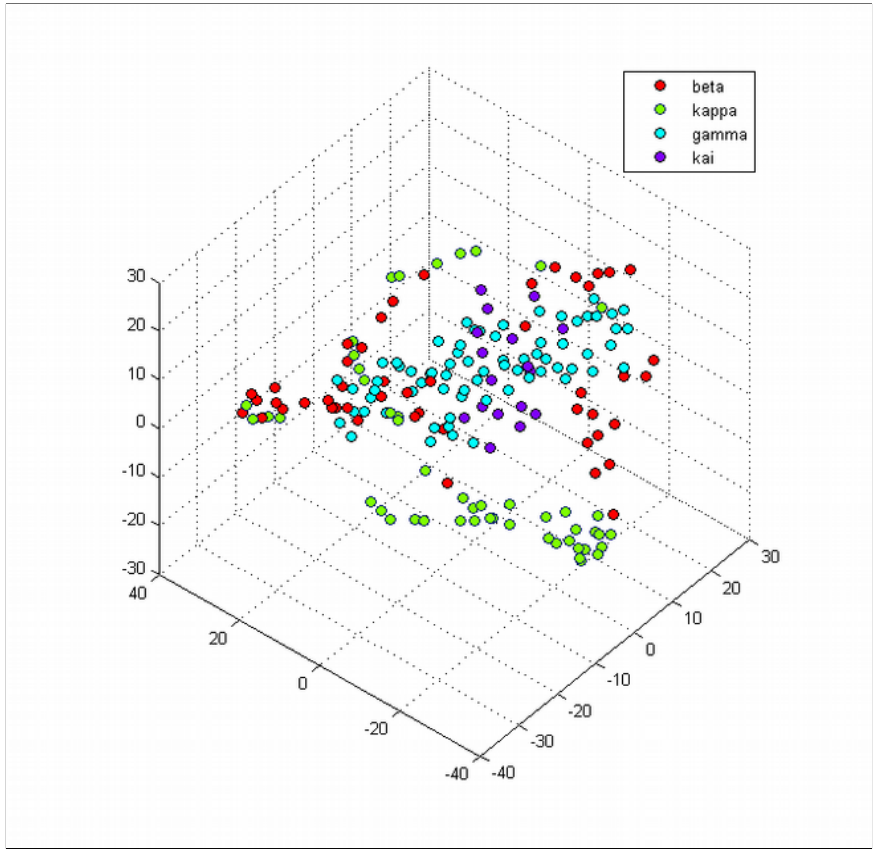


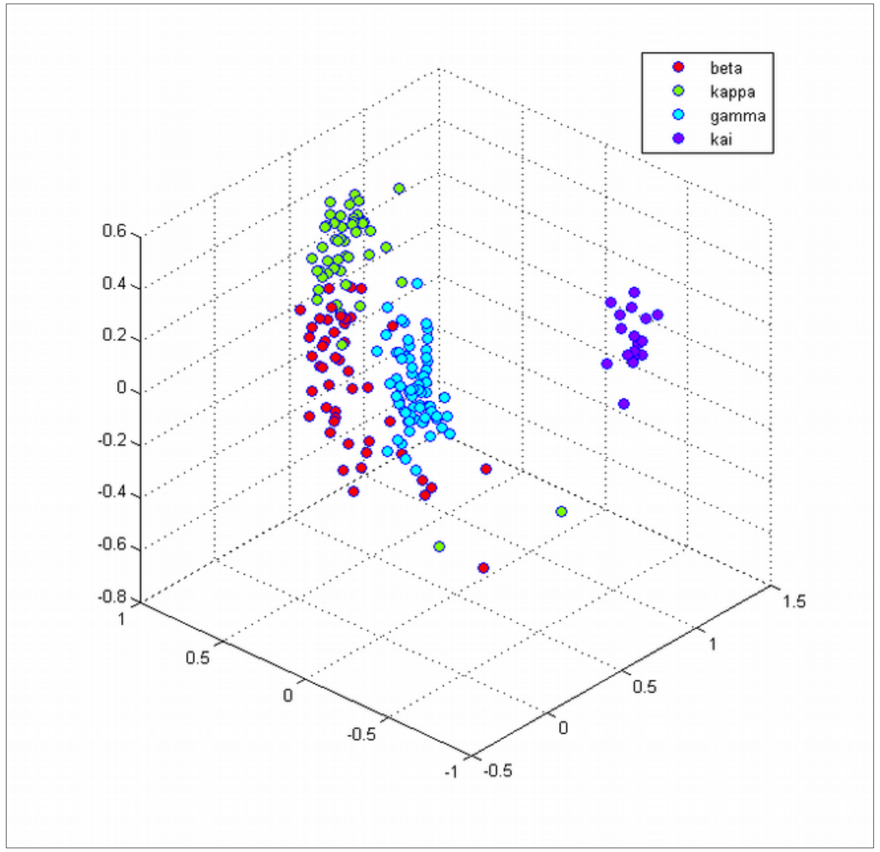


Gianniotis et al. 2015

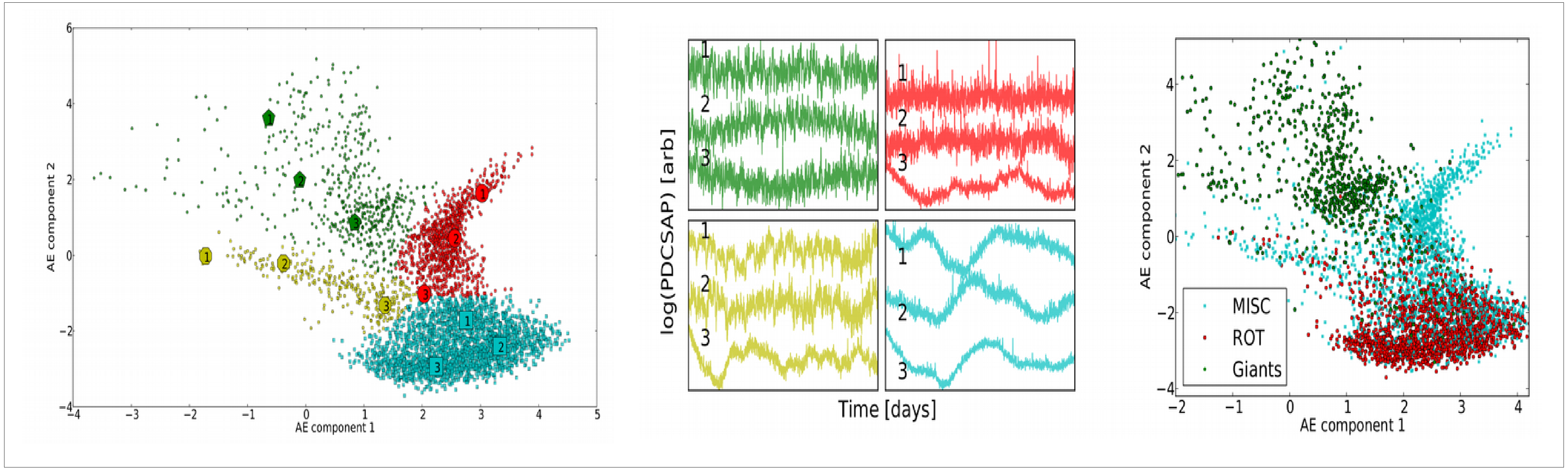
Lessons learned

temporal behavior is **more than** a simple **vector**

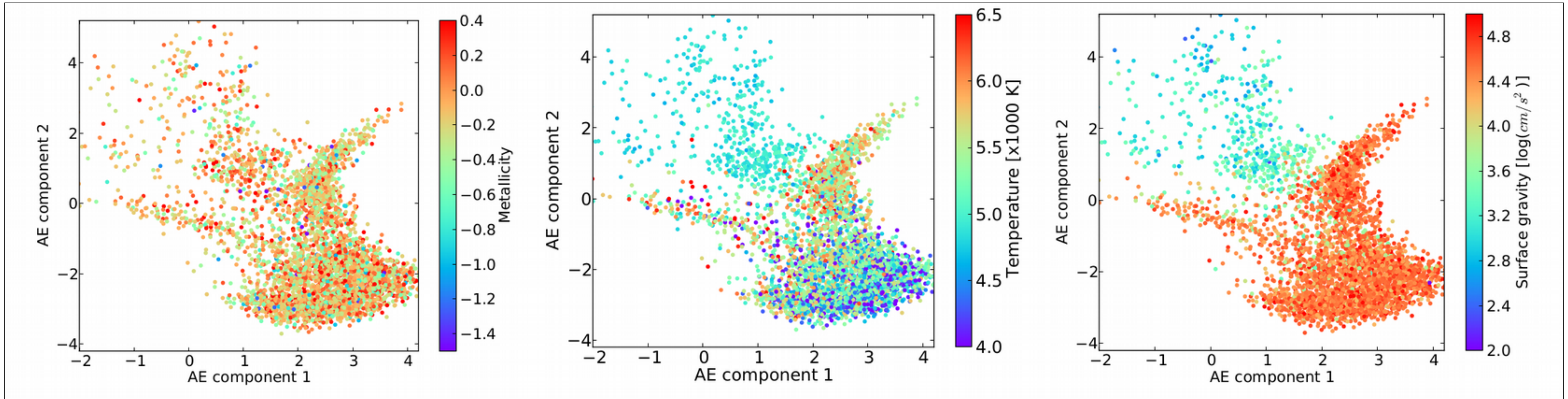




Kepler data / stellar objects



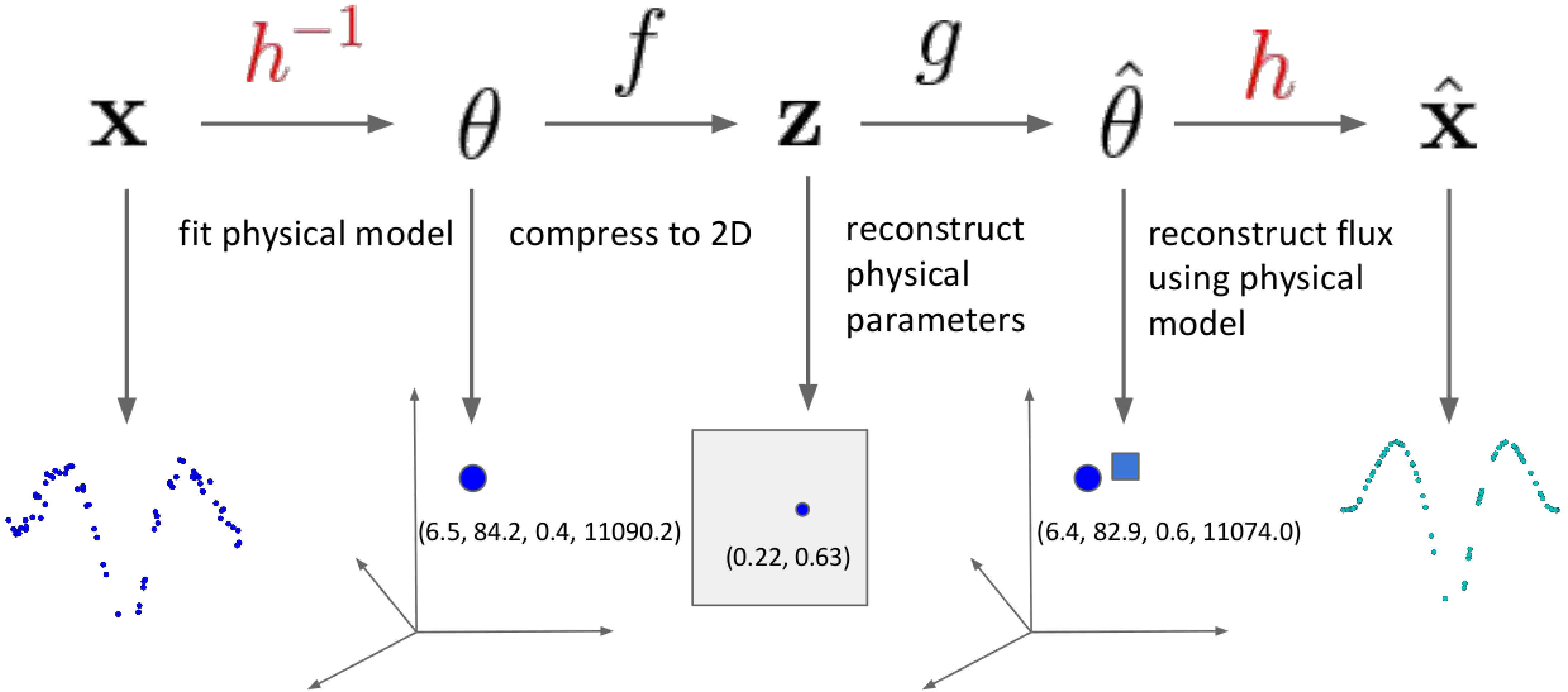
Kuegler et al. 2015



(a) Metallicity (b) Temperature (c) Surface gravity

Physical model & autoencoder

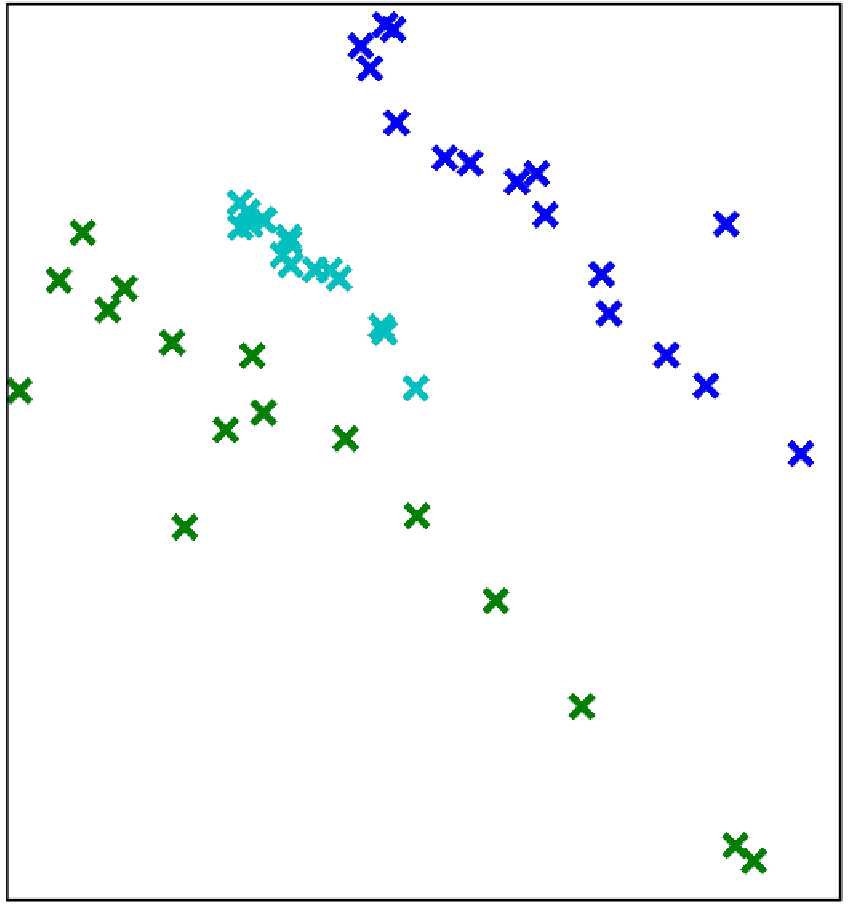
Timeseries of stellar binary systems



$$\min_{f,g} ||\hat{\mathbf{x}} - \mathbf{x}||^2 \text{ such that } h(g(f(\theta))) = \hat{\mathbf{x}}$$

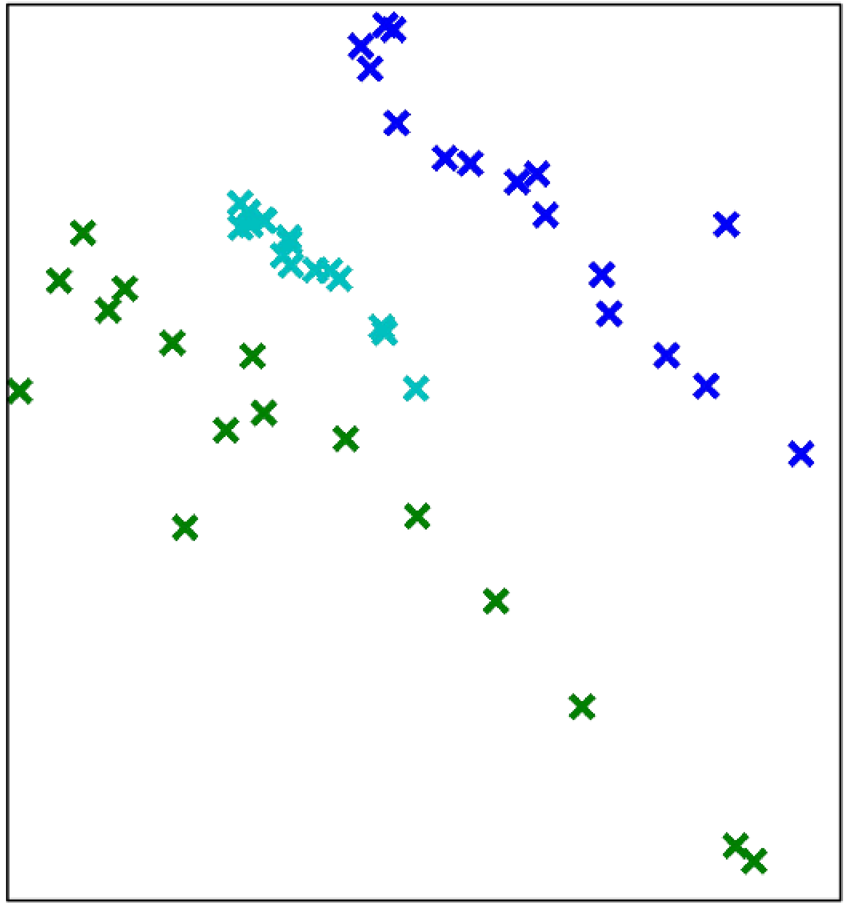
Project data

$$\mathbf{x} \xrightarrow{h^{-1}} \theta \xrightarrow{f} \mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$

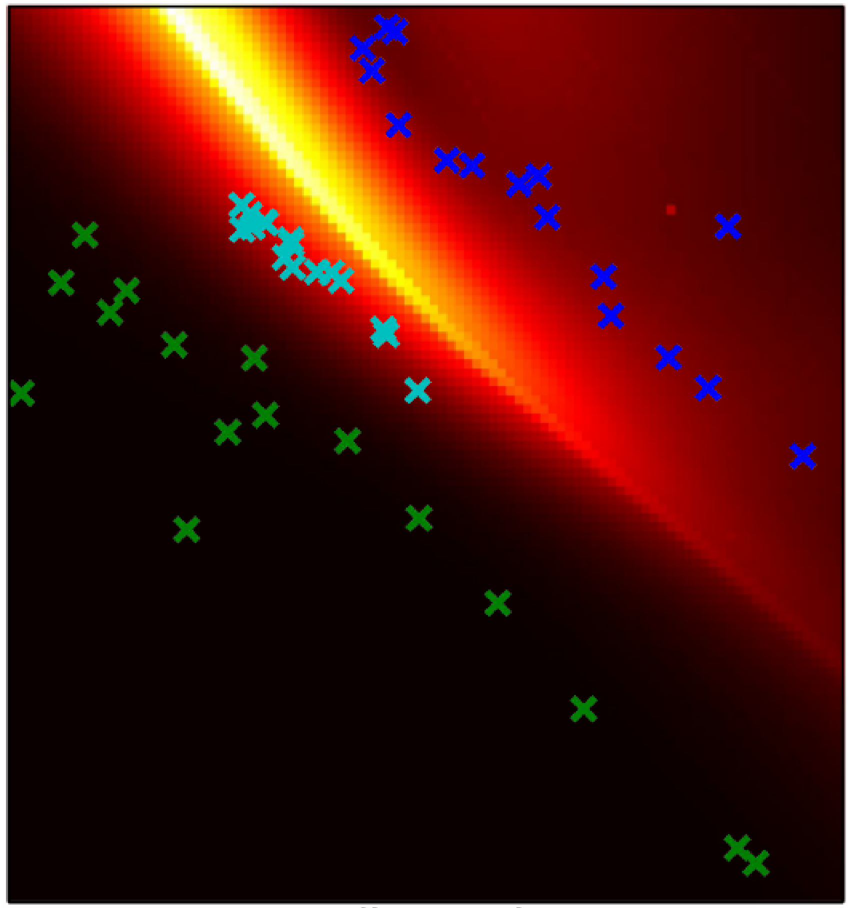


Analyze distance

$$\mathbf{x} \xrightarrow{h^{-1}} \theta \xrightarrow{f} \mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$



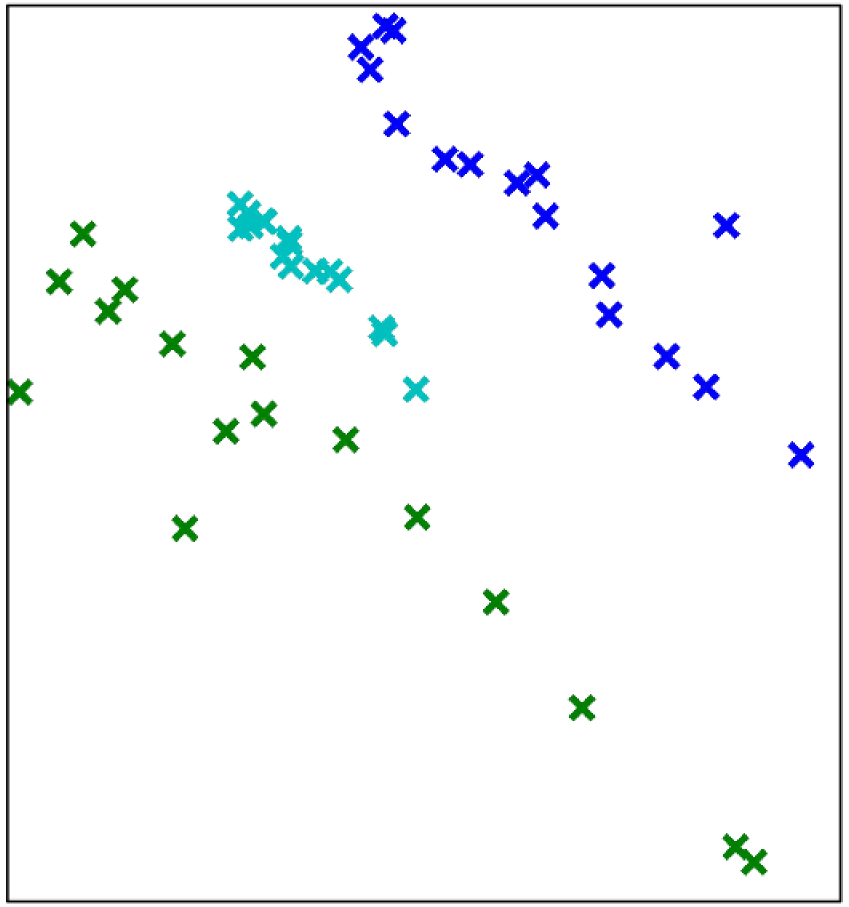
$$\mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$

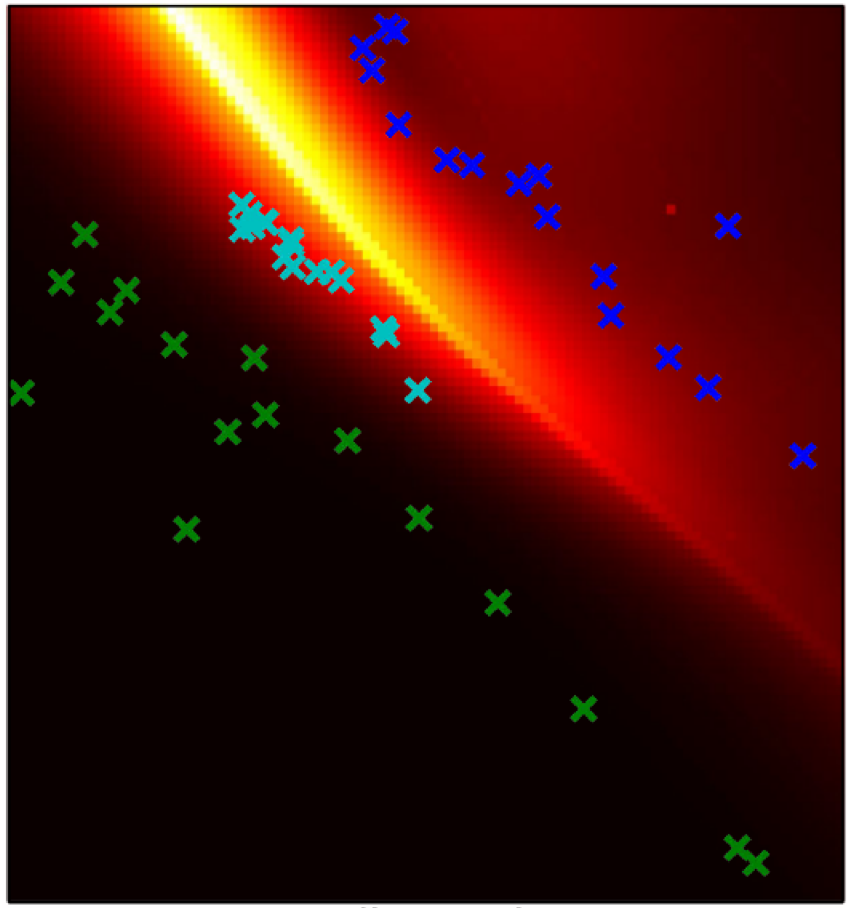


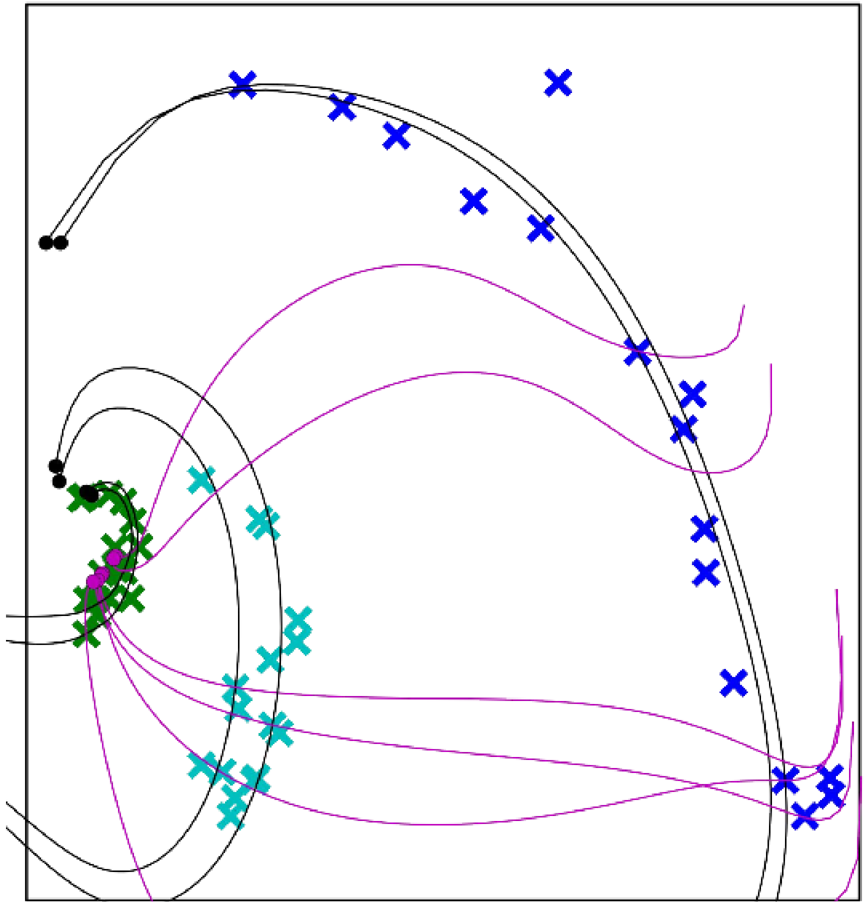
Plot iso-lines

$$\mathbf{x} \xrightarrow{h^{-1}} \theta \xrightarrow{f} \mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$

$$\mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$







Mass, Temperature

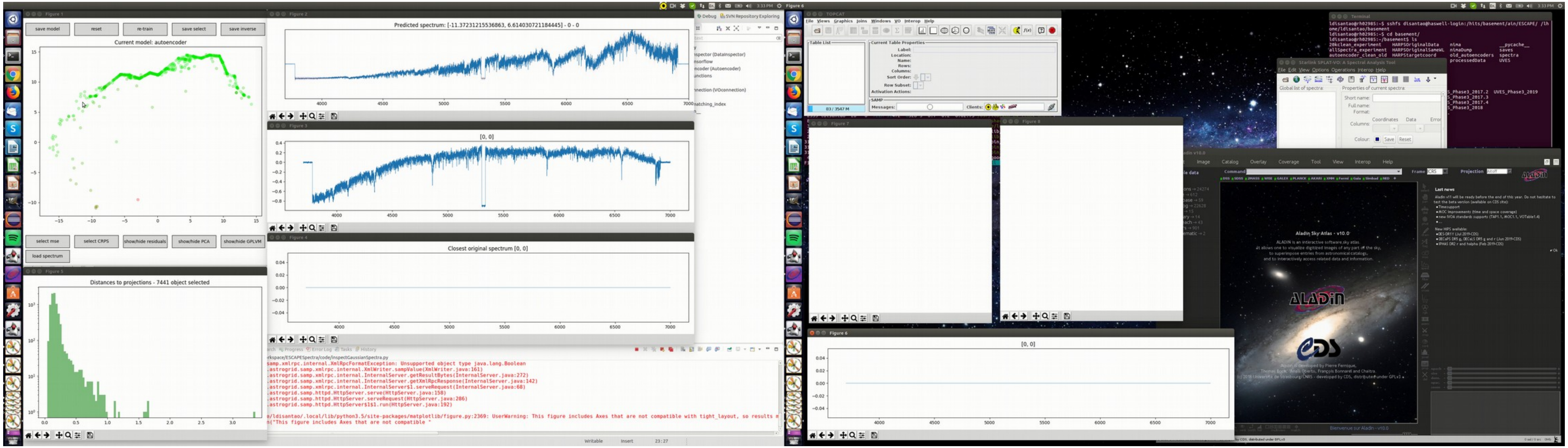
Spectral Data Analysis

dealing with spectral data

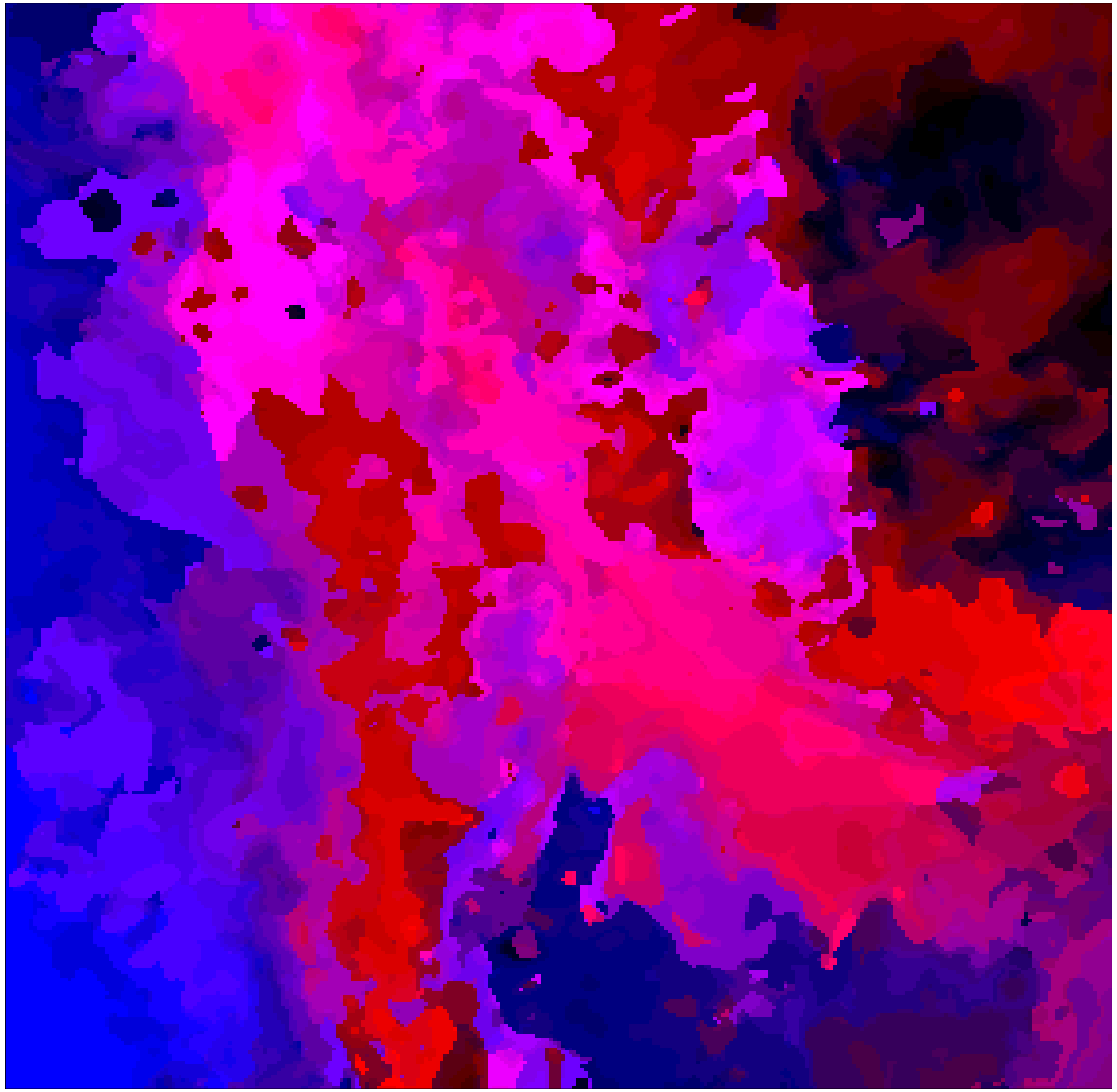
Playing with spectra

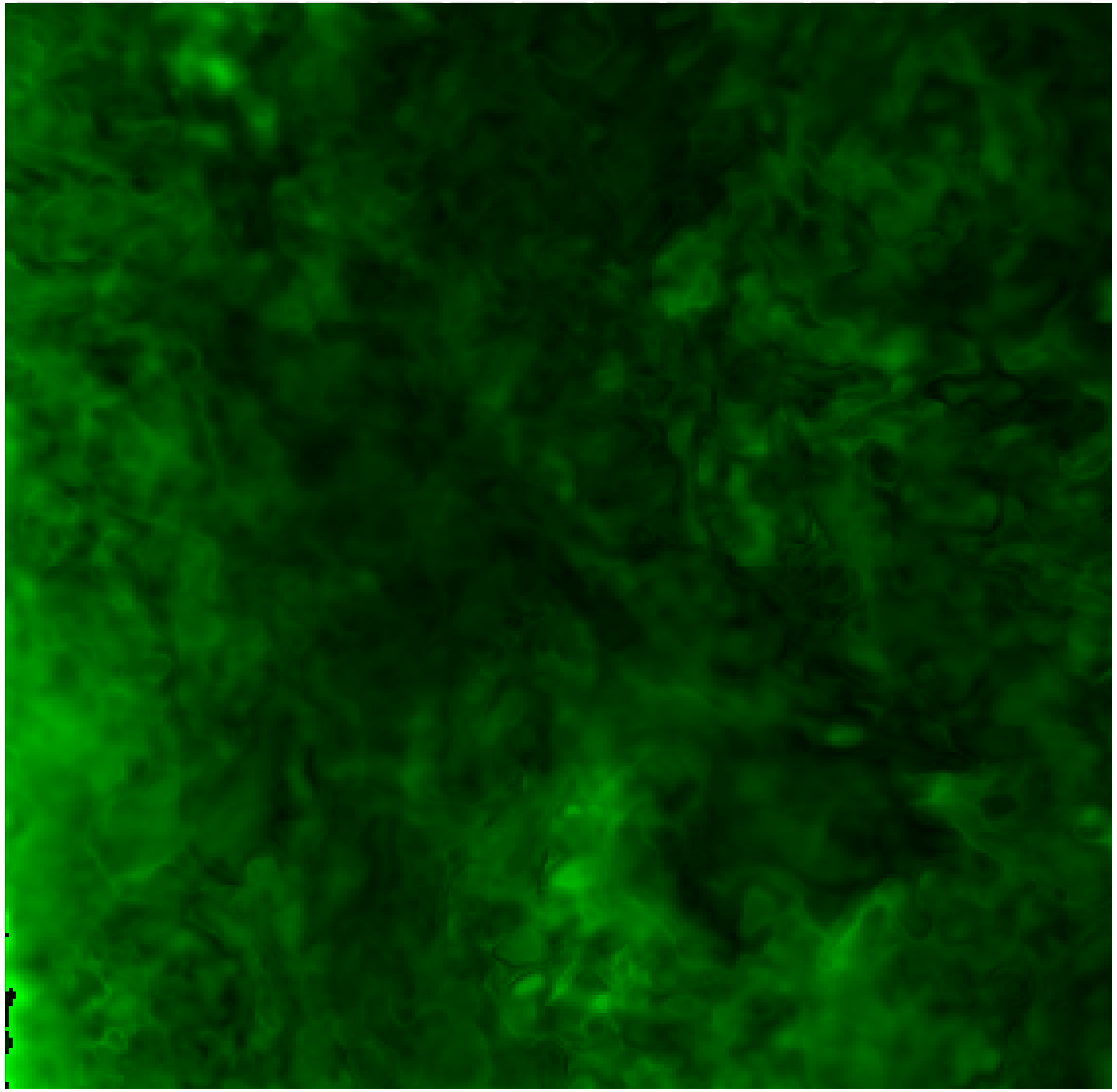
ESCAPE project ESO/CDS/HITS

- exploring 300k spectra in realtime on a laptop

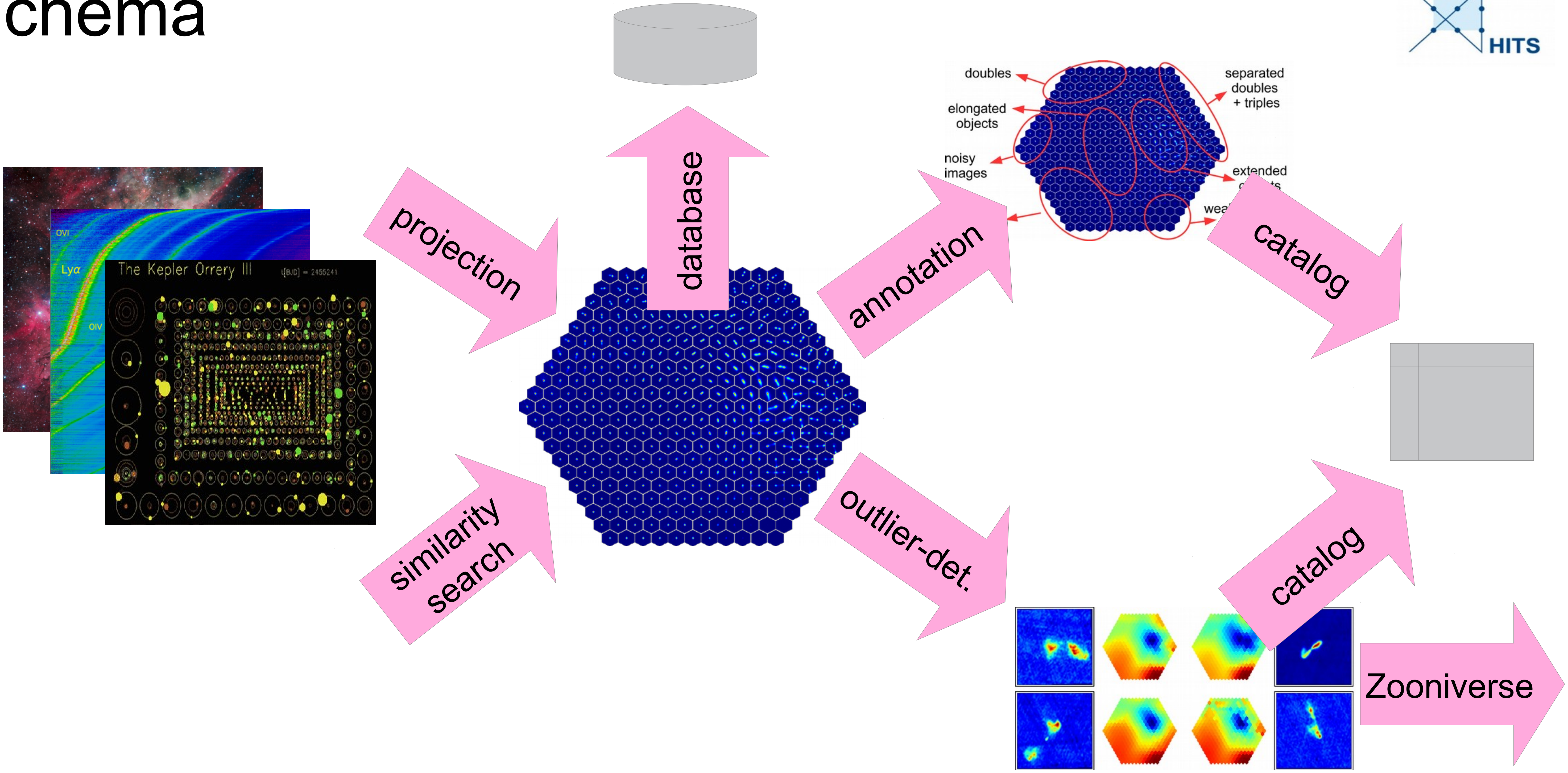


Data cubes / ppv-data





Schema



The Kepler Orrery III
Lyα
OVI
OIV
projection
database
similarity search
annotation
outlier-det.
doubles
elongated objects
noisy images
separated doubles + triples
extended
catalog
catalog
Zooniverse

Lessons learned

don't **bias** your system

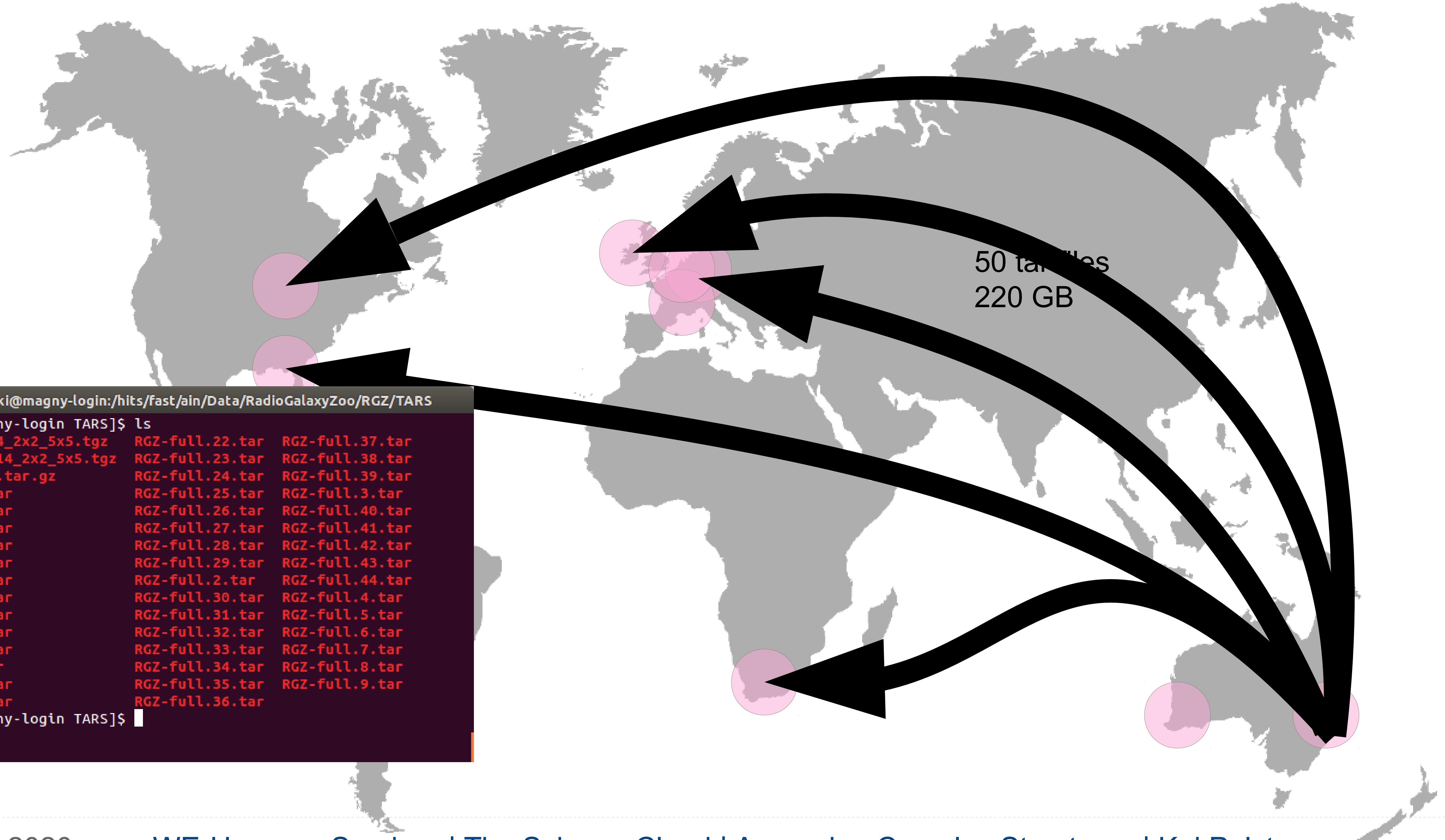
use stupid, but fast **computers** for the boring tasks

do the **creative** interpretation of the data

Accessing and Analyzing Data

uhhhhs, ooohhs, don’ts and lessons learned

Starting a project in 2015

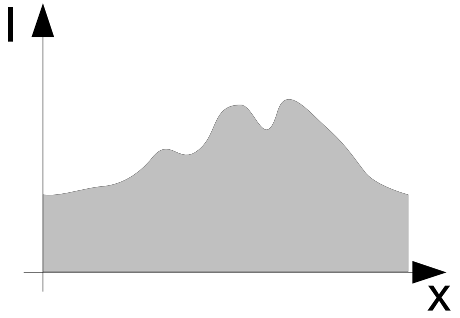



```
polsteki@magny-login:/hits/fast/ain/Data/RadioGalaxyZoo/RGZ/TARS
[polsteki@magny-login TARS]$ ls
cdfs_11JAN2014_2x2_5x5.tgz   RGZ-full.22.tar  RGZ-full.37.tar
elais_11JAN2014_2x2_5x5.tgz  RGZ-full.23.tar  RGZ-full.38.tar
Imaging-1.1.7.tar.gz         RGZ-full.24.tar  RGZ-full.39.tar
RGZ-full.10.tar              RGZ-full.25.tar  RGZ-full.3.tar
RGZ-full.11.tar              RGZ-full.26.tar  RGZ-full.40.tar
RGZ-full.12.tar              RGZ-full.27.tar  RGZ-full.41.tar
RGZ-full.13.tar              RGZ-full.28.tar  RGZ-full.42.tar
RGZ-full.14.tar              RGZ-full.29.tar  RGZ-full.43.tar
RGZ-full.15.tar              RGZ-full.2.tar   RGZ-full.44.tar
RGZ-full.16.tar              RGZ-full.30.tar  RGZ-full.4.tar
RGZ-full.17.tar              RGZ-full.31.tar  RGZ-full.5.tar
RGZ-full.18.tar              RGZ-full.32.tar  RGZ-full.6.tar
RGZ-full.19.tar              RGZ-full.33.tar  RGZ-full.7.tar
RGZ-full.1.tar               RGZ-full.34.tar  RGZ-full.8.tar
RGZ-full.20.tar              RGZ-full.35.tar  RGZ-full.9.tar
RGZ-full.21.tar              RGZ-full.36.tar
[polsteki@magny-login TARS]$
```

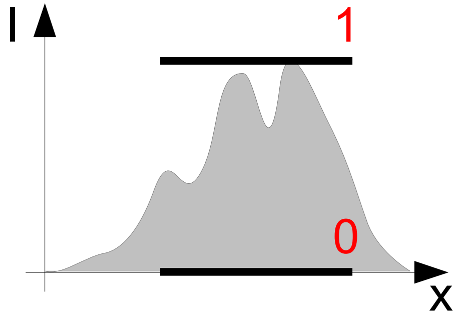
Preprocessing

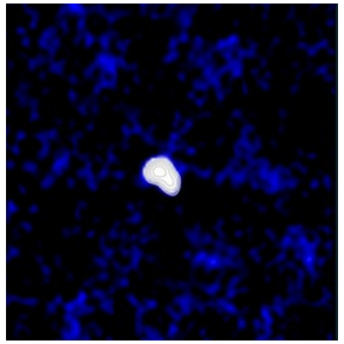
extract
matrix from fits

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

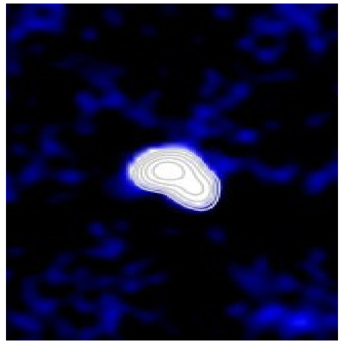


normalize
flux relative to the maximum



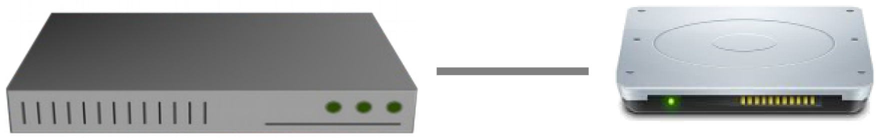


cutout
interesting region



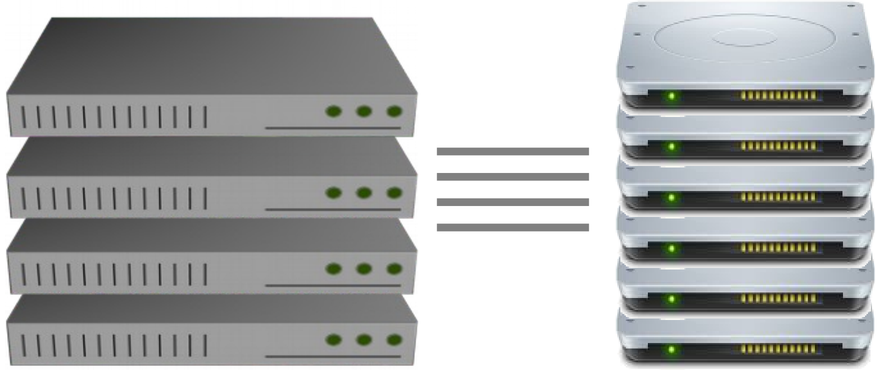
Speeding up preprocessing

single core python = **48** hours



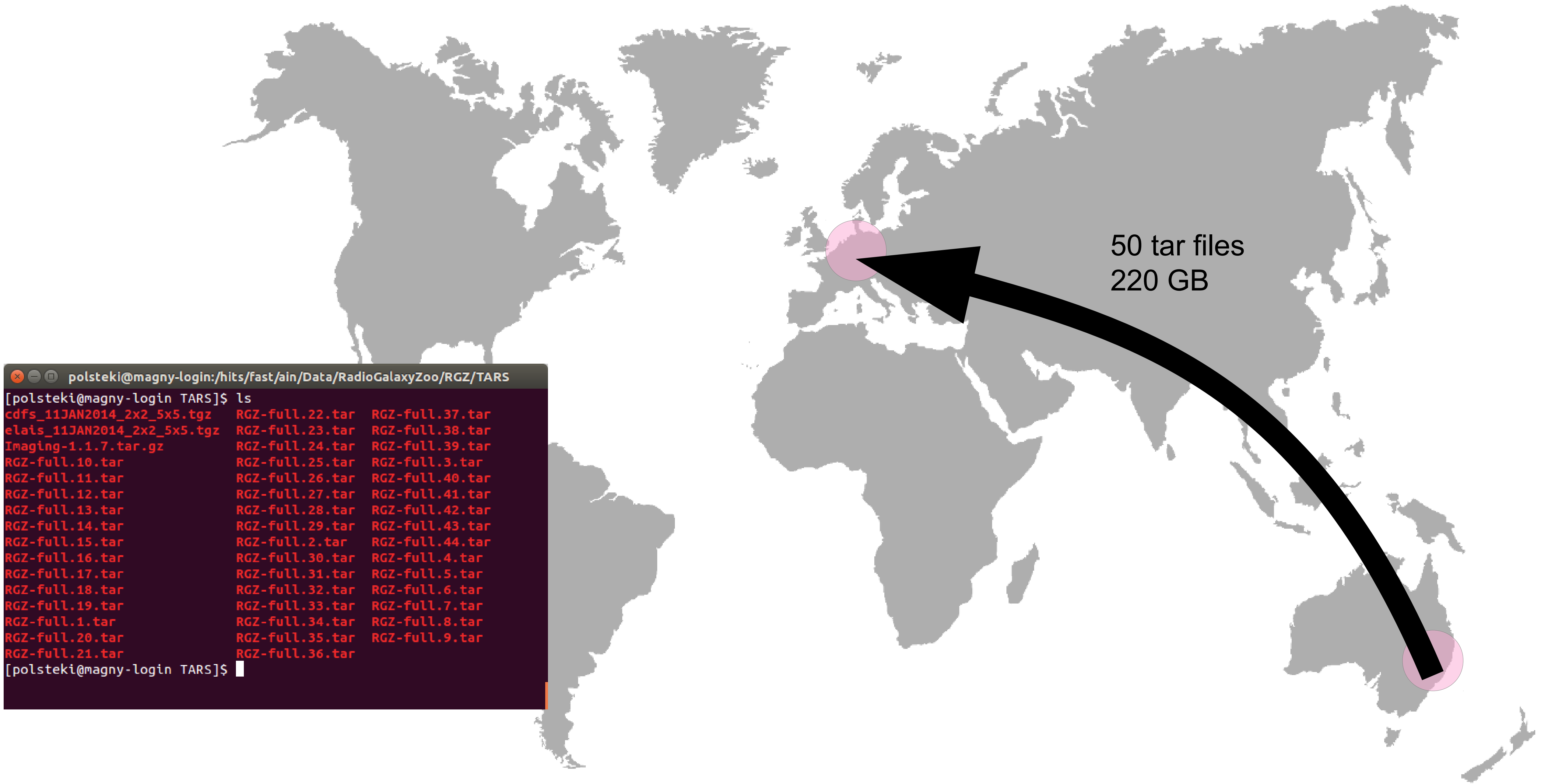
with hadoop

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

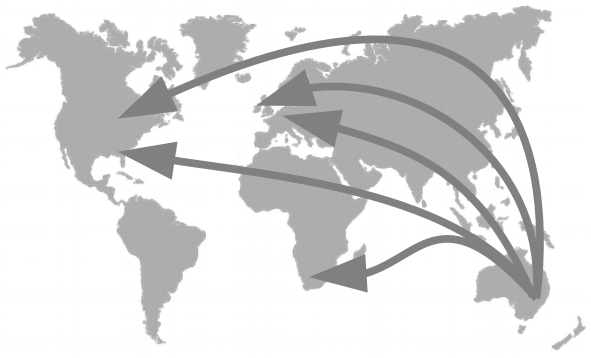


file **access** was still the bottleneck!

New images extracted

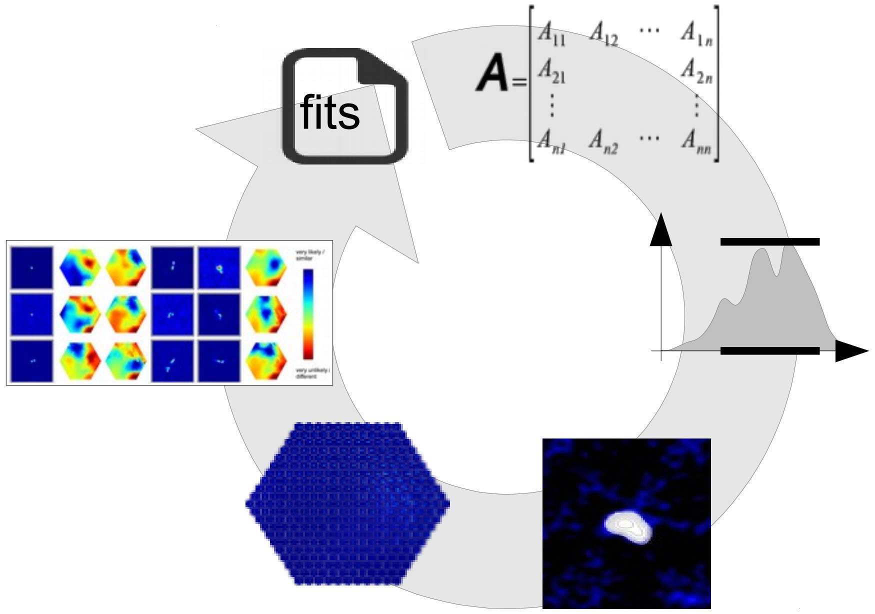


The downsides of this approach



a lot of local copies

no orchestration of work-flow



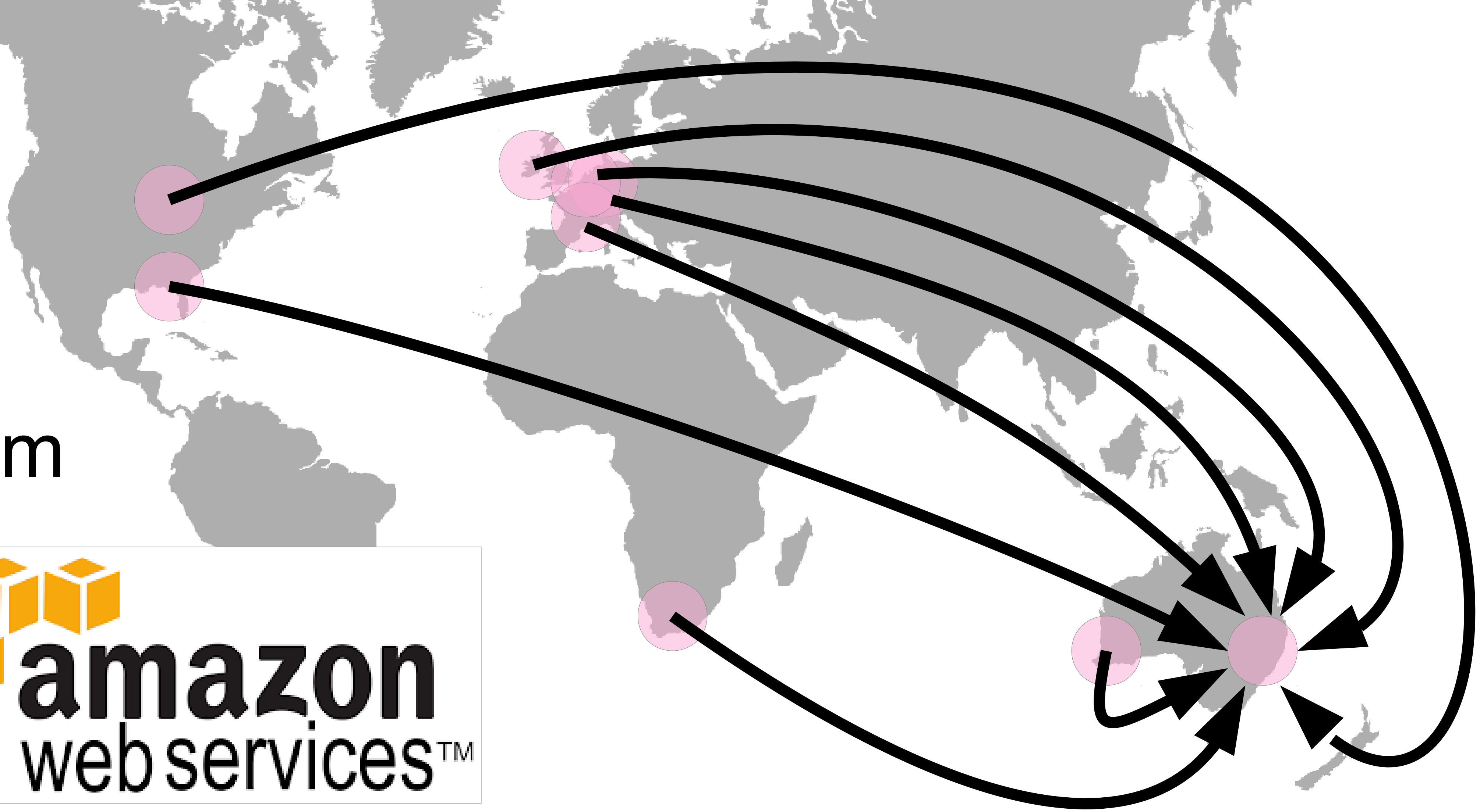
bad exchange of intermediate results

very exclusive concerning hardware requirements

NVIDIA Tesla K40

2015 / used Amazon Web Services

provided the larges variety of services
and the best infrastructure



SKA grant from

amazon web services™

Today / bringing code to the data

+ docker

Challenges with data that still exist

how to extract 1,000,000 thumbnails of 64x64 pixel²

- required for a lot of machine learning tasks
- standards exist but often not/partially implemented

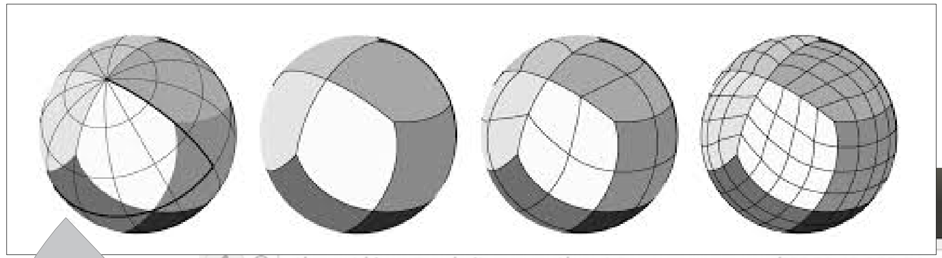
how to train a model without downloading the data

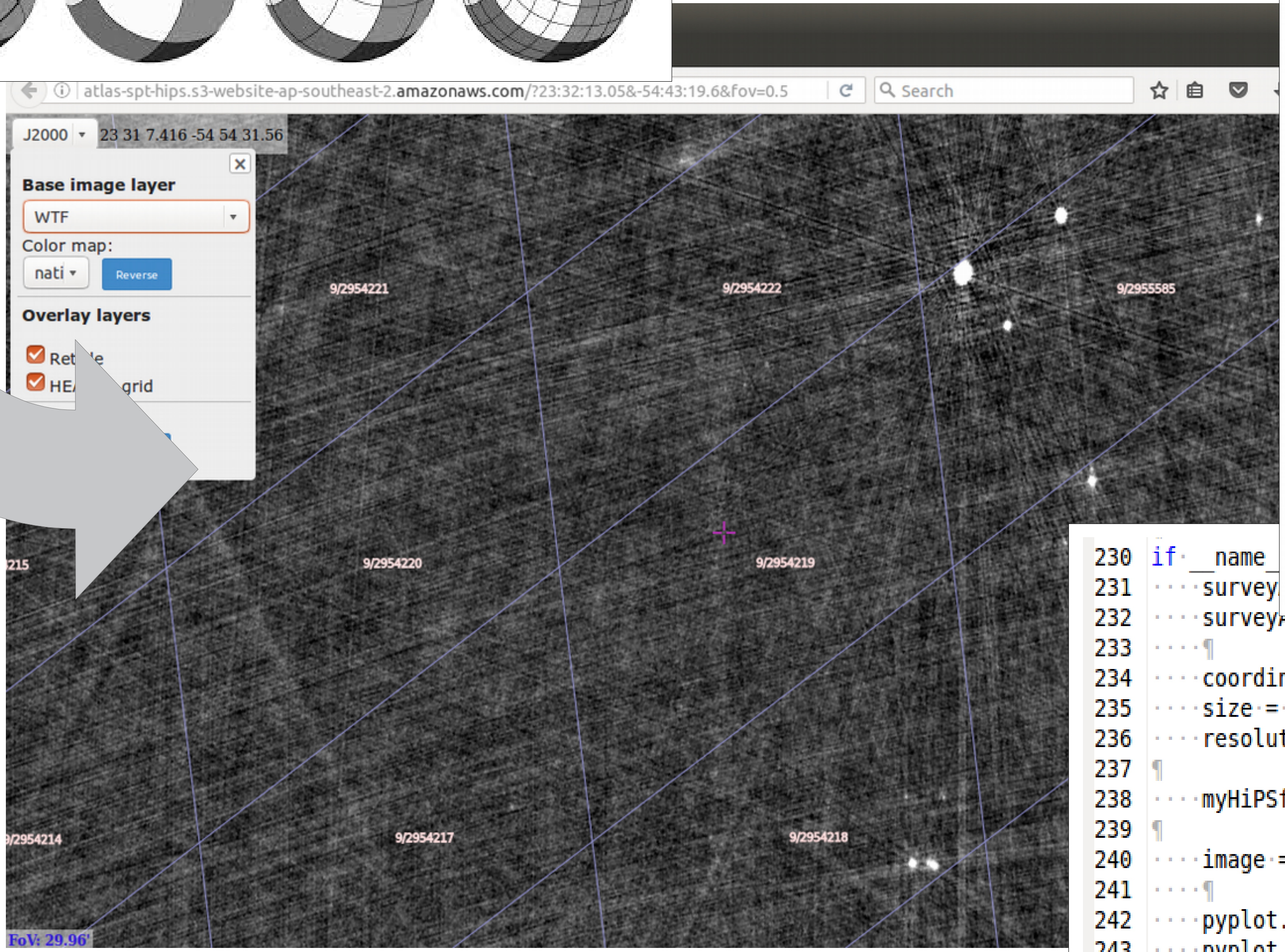
- extraction/pre-processing
- reproducibility/training+test data

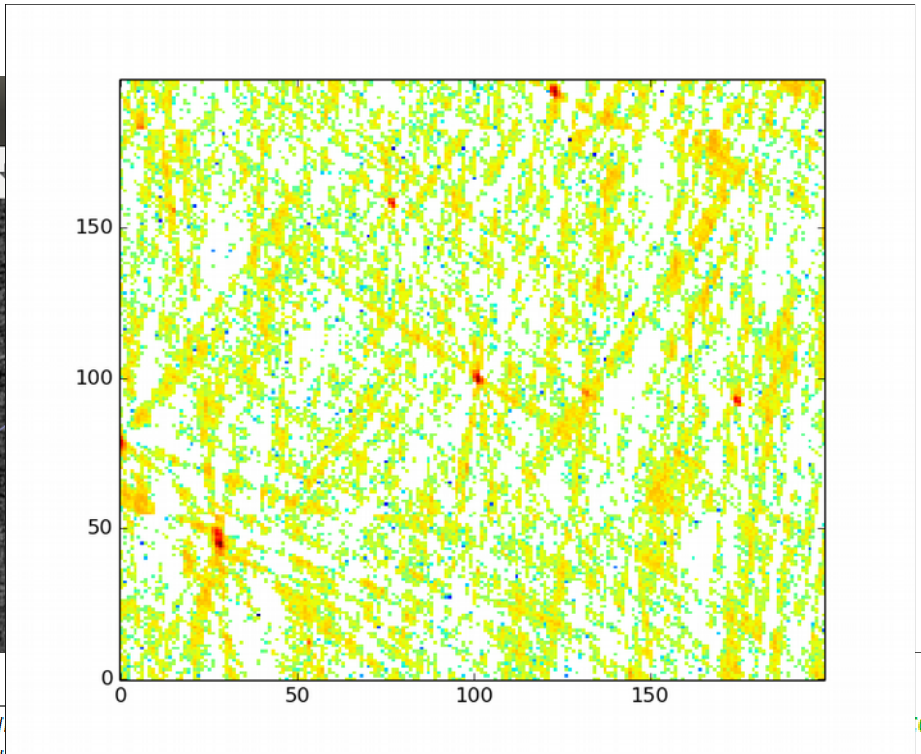
how to deal with distributed data-sets

- deal with radio and IR-data

Healpix / HiPS / IVOA








```
if __name_
    survey [illegible] om/ATLAS-SPT-64x64"
    surveyAddress = [illegible]alasky.u-strasbg.fr/DSS/DSS2Merged" # DSS red

    coordinate = [350.86, -55.225]
    size = [200,200]
    resolution = 0.002

    myHiPSfs = HiPSfs(surveyAddress) # create access

    image = myHiPSfs.extractCoordinate(coordinate, size, resolution, nested=True) # extract data array

    pyplot.figure()
    pyplot.imshow((image), aspect='auto', interpolation="nearest")
    pyplot.gca().invert_yaxis()
    pyplot.show()
```

Conclusion

machine learning → accessing data

but, we have to make the data accessible to ML