



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
- 12 days compute time, 7 days remaining, but only 5 days until observation run

vs.

- scanning version on presorted lists
- results after 4 seconds



$o(n^3)$



$o(n)$



It's not just about big-data



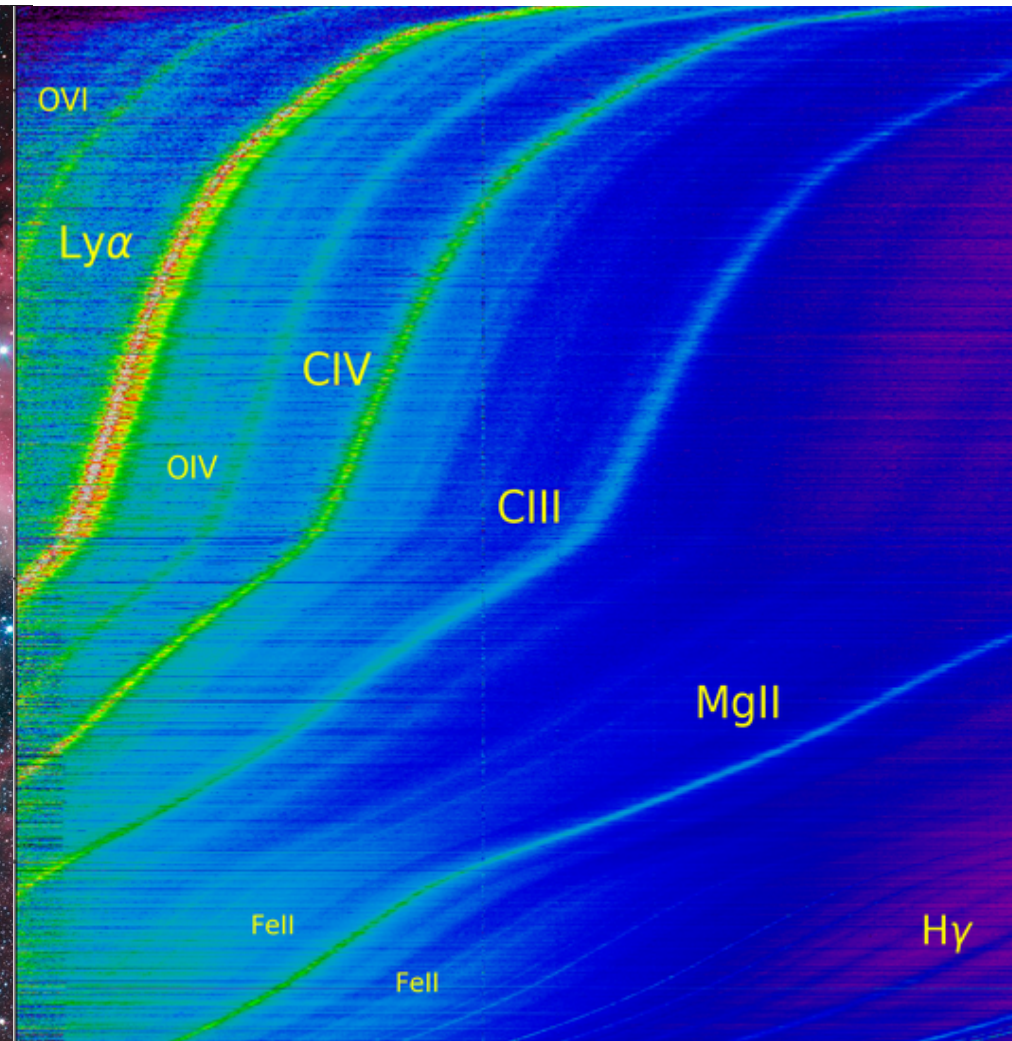
spatial

spectral

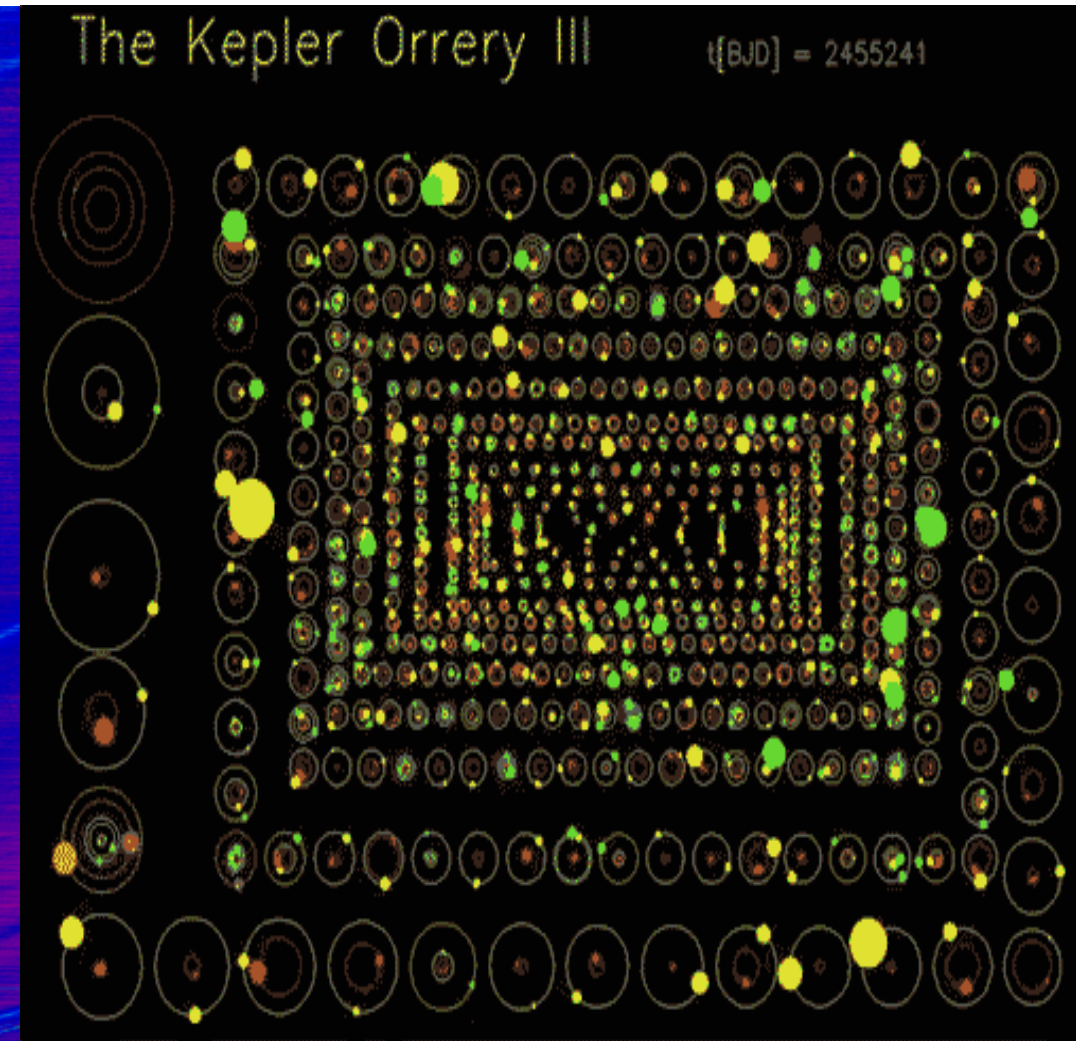
temporal



ESO



SDSS



NASA

metric. $\Delta(A, \phi(B)) = 0$ and thus $A = \phi(B)$ (i.e., $A \sim B$) because d is a

$\Delta(A, B) = \min\{d(A, \phi(B)) \mid \phi \in \Phi\}$

$= \min\{d(\phi(B), A) \mid \phi \in \Phi\}$ by symmetry of d

$= \min\{d(B, \phi^{-1}(A)) \mid \phi \in \Phi\}$ because of [A3]

$\geq \Delta(B, A)$

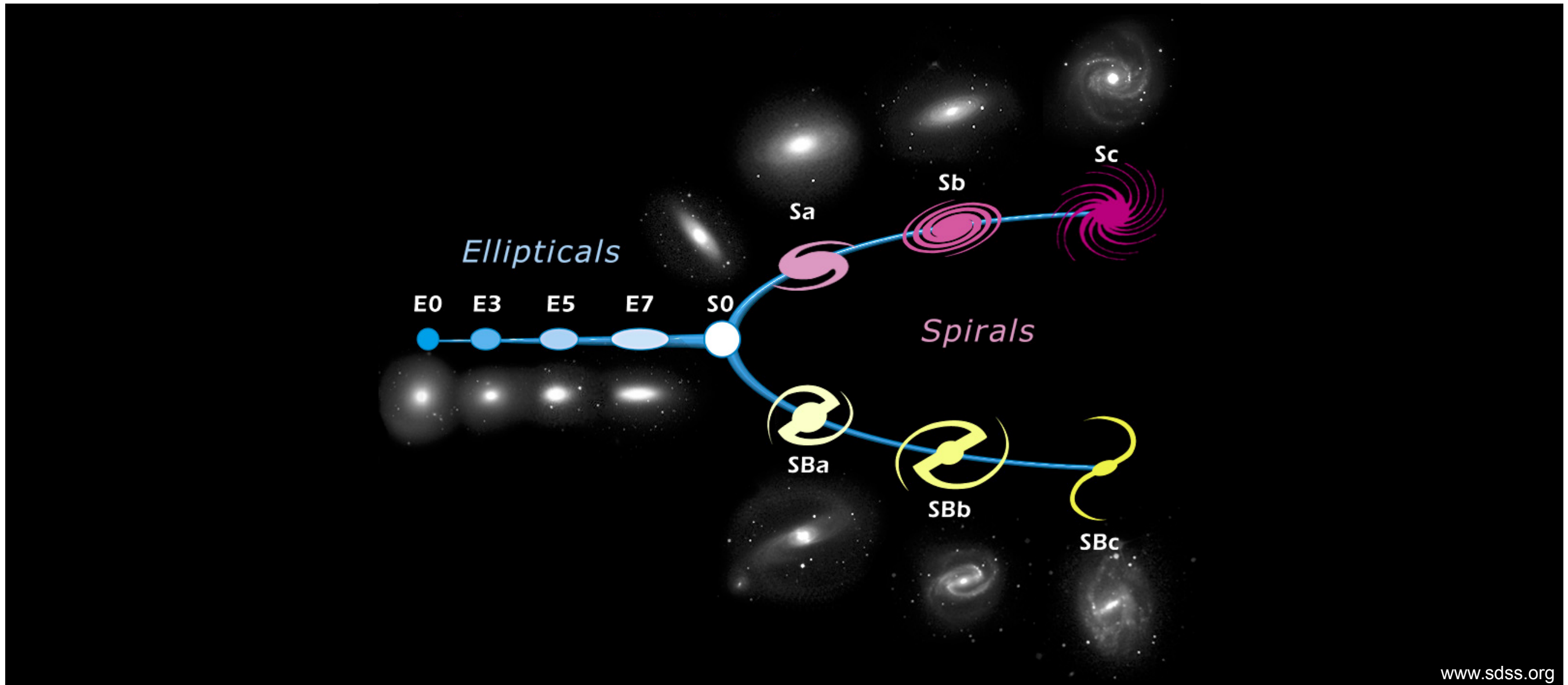
Morphology of Radio Galaxies

how to deal with complex shapes

Simplifying life / putting things in boxes



Morphology of galaxies / Edwin Hubble's classification scheme



www.sdss.org

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

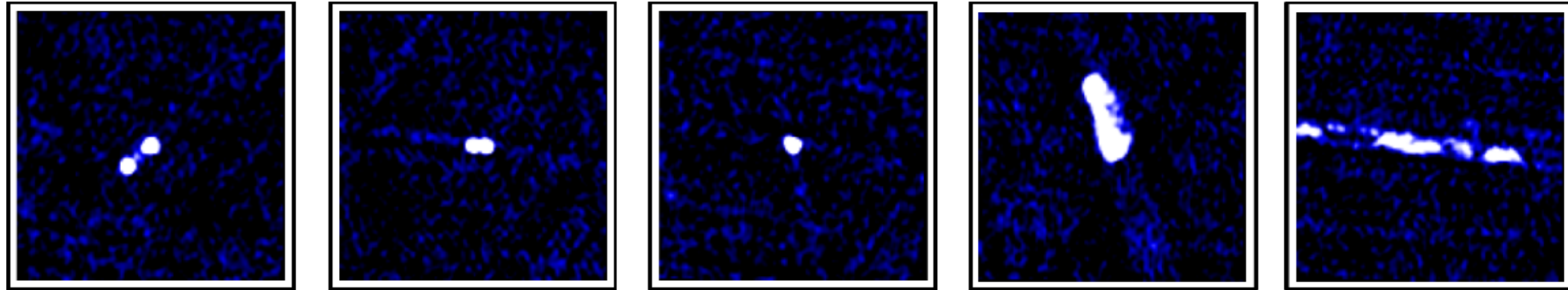
What about 50.000.000 images of galaxies?

Outsourcing the work / citizen science

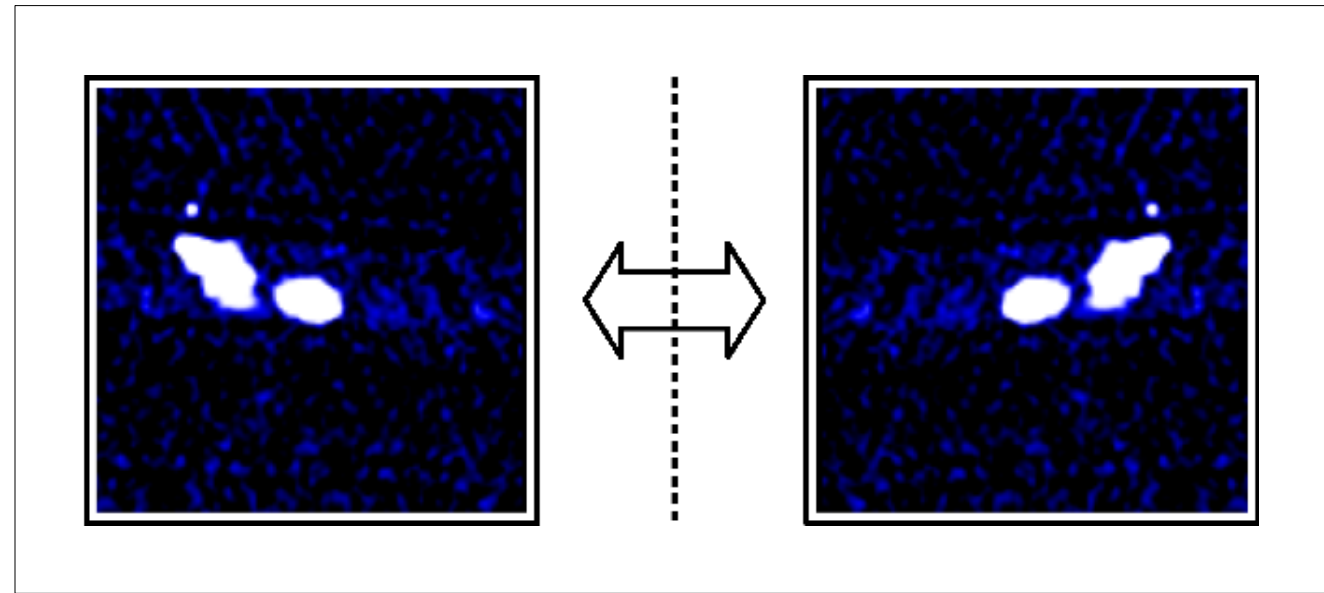
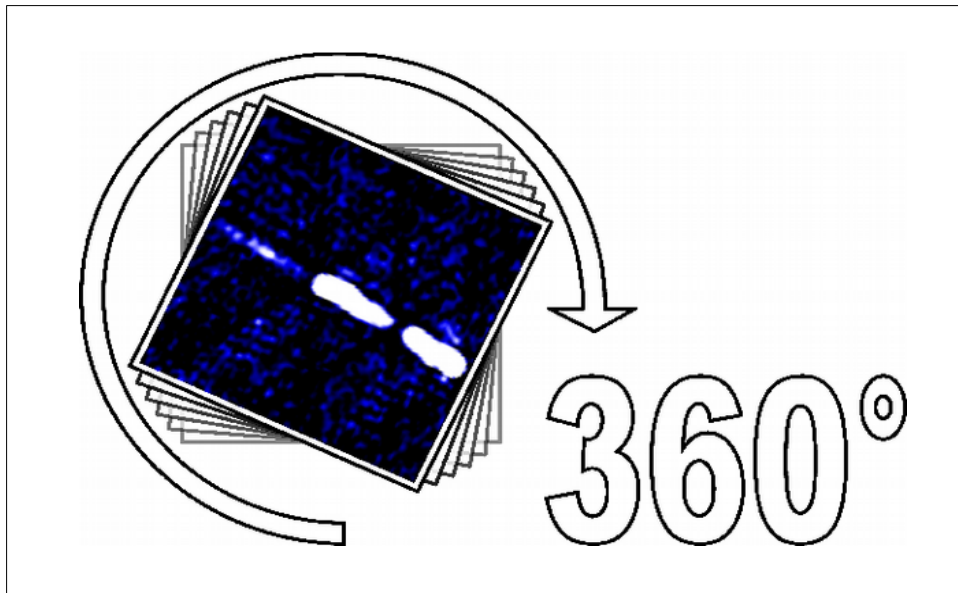


The image shows a browser window displaying the Galaxy Zoo website. The browser's address bar shows 'radio.galaxyzoo.org' and the search bar contains 'radio galaxy zoo'. The website header includes navigation links for 'CLASSIFY', 'SCIENCE', 'TEAM', 'PROFILE', 'TALK', and 'BLOG'. The main content area features a large heading 'In Search of Erupting Black Holes' with a sub-heading 'Help astronomers discover supermassive black holes observed by the KG Jansky Very Large Array (NRAO) and the Australia Telescope Compact Array (CSIRO)'. Below this is a section titled 'Search for Black Holes' with a paragraph of text: 'Black holes are found at the center of most, if not all, galaxies. The bigger the galaxy, the bigger the black hole and the more sensational the effect it can have on the host galaxy. These supermassive black holes drag in nearby material, growing to billions of times the mass of our sun and occasionally producing spectacular jets of material traveling nearly as fast as the speed of light. These jets often can't be detected in visible light, but are seen using radio telescopes. Astronomers need your help to find these jets and match them to the galaxy that hosts them.' A yellow 'Begin Hunting' button is located at the bottom of this section. To the right of the text is an image of two galaxies with jets of material being ejected from their centers. At the bottom right of the page, there is a small credit line: 'NASA, ESA, S. Baum and C. O'Dea (RIT), R. Perley and W. Cotton (NRAO/AUI/NSF), and the Hubble Heritage Team (STScI/AURA)'. On the left side of the browser window, there are overlapping panels from other parts of the website, including a 'Classify' section with a 'Begin C' button and a 'How' section.

The challenge / Radio GalaxyZoo



rotation / flipping invariant



The solution

have an expert inspect every data-item

vs.

machine learning

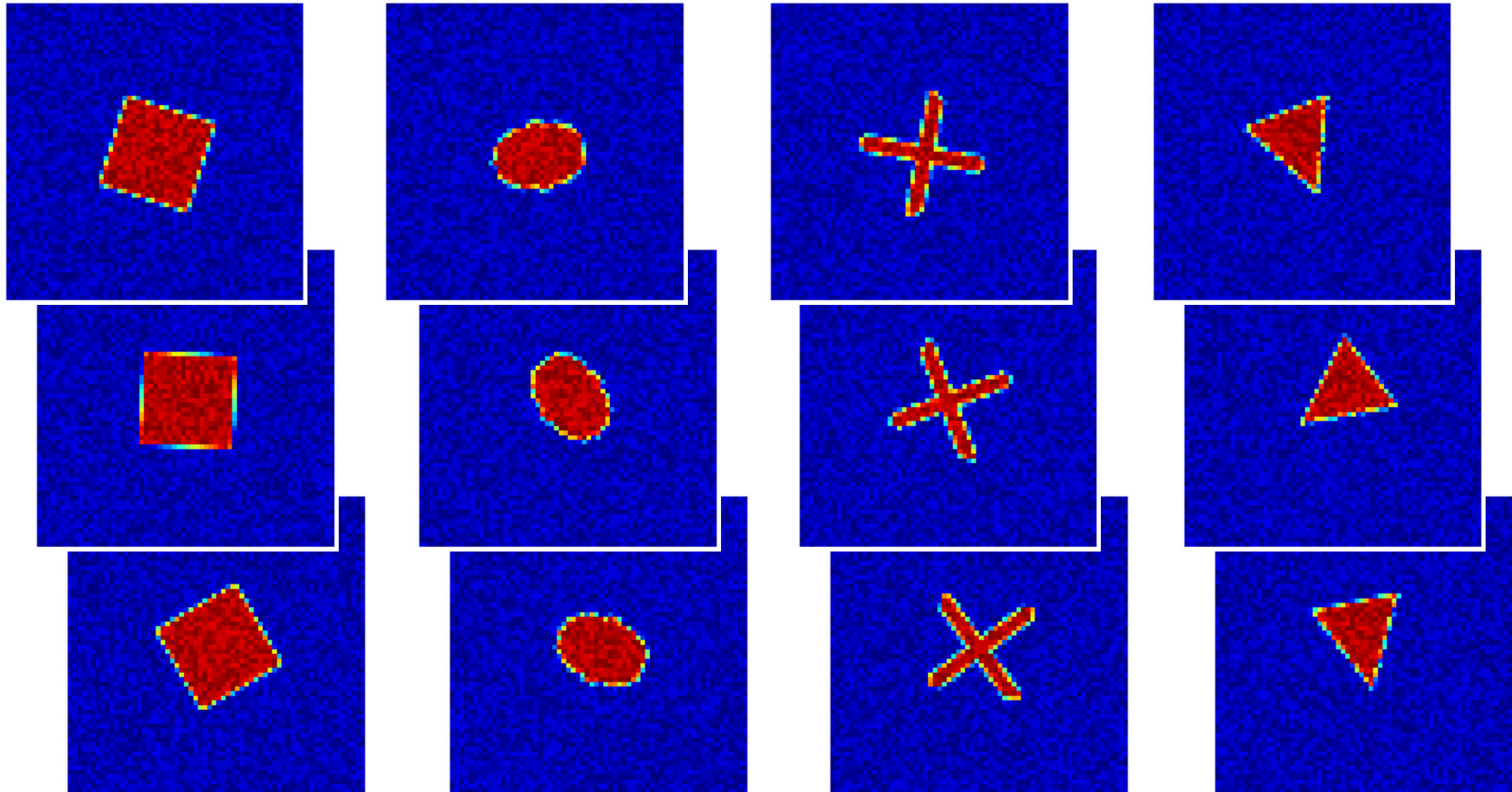
- don't ask scientific questions directly
- ask computers, to structures/sort the data
- do your individual interpretation/analysis
→ use dimensionality reduction

Similarity measure

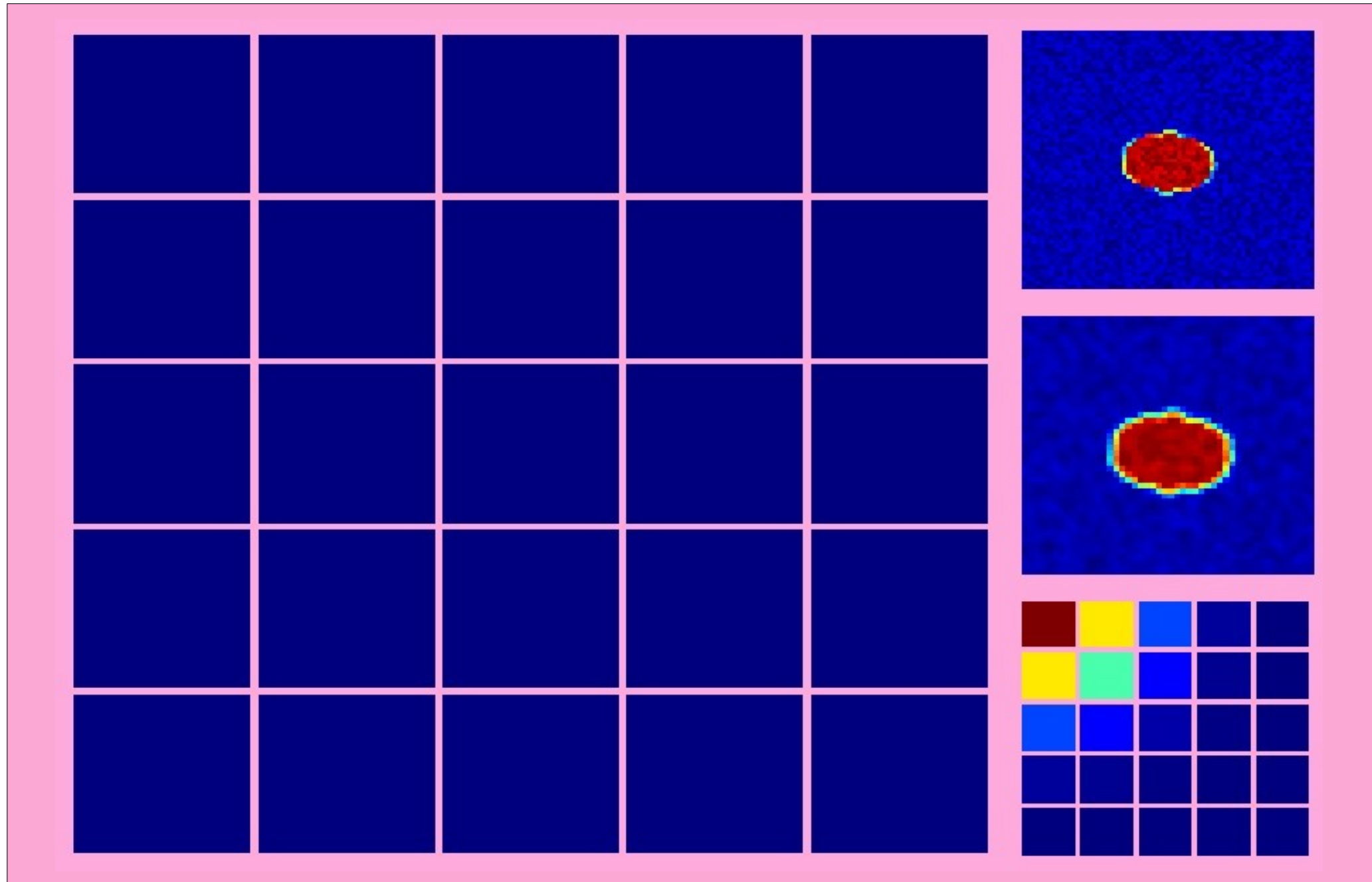
calculate ^{the} **Euclidean distance** ^{pixel based}

for all possible
rotations
and find
the best
matching angle
via
minimization

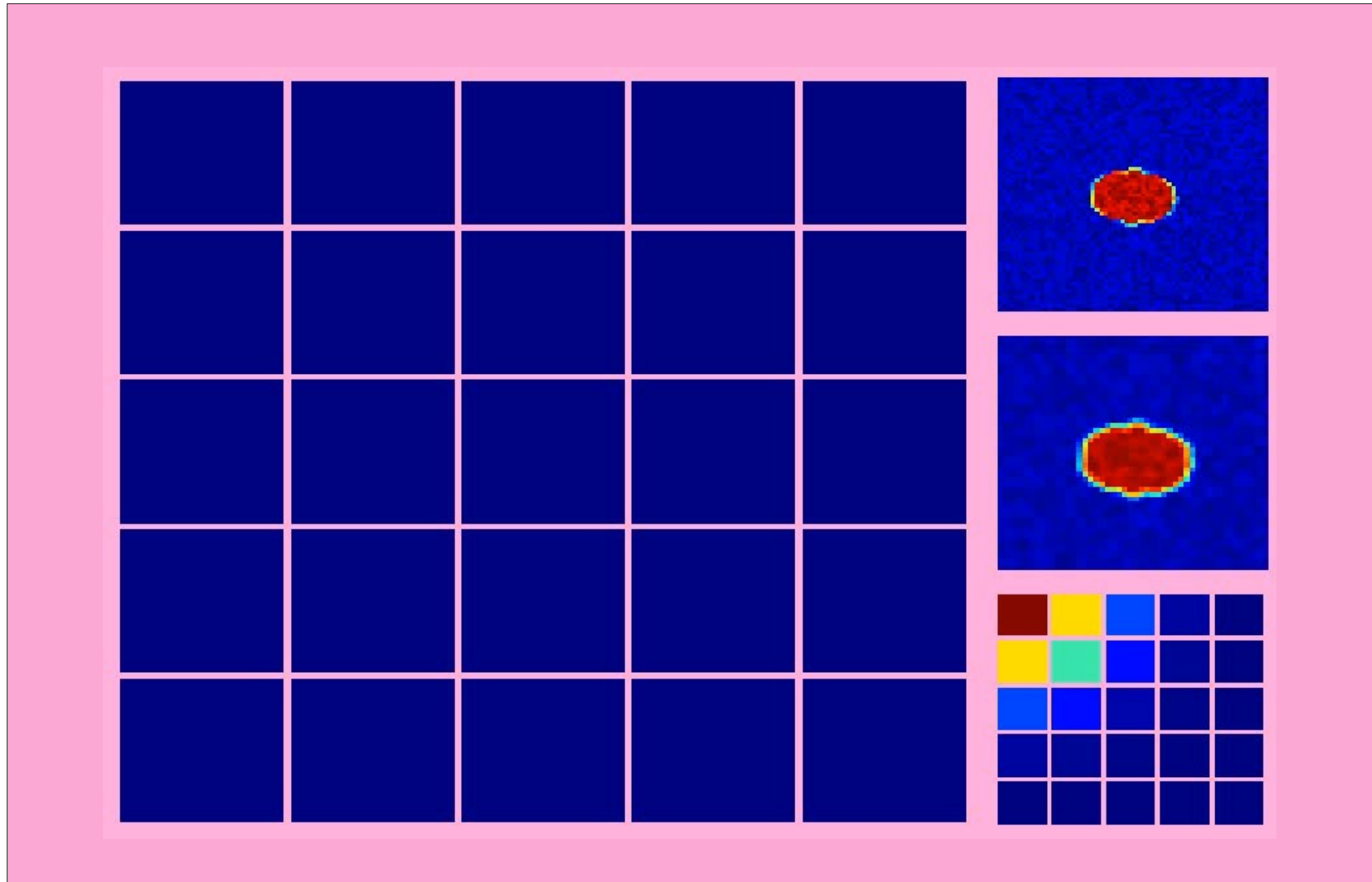
Example self-organizing map



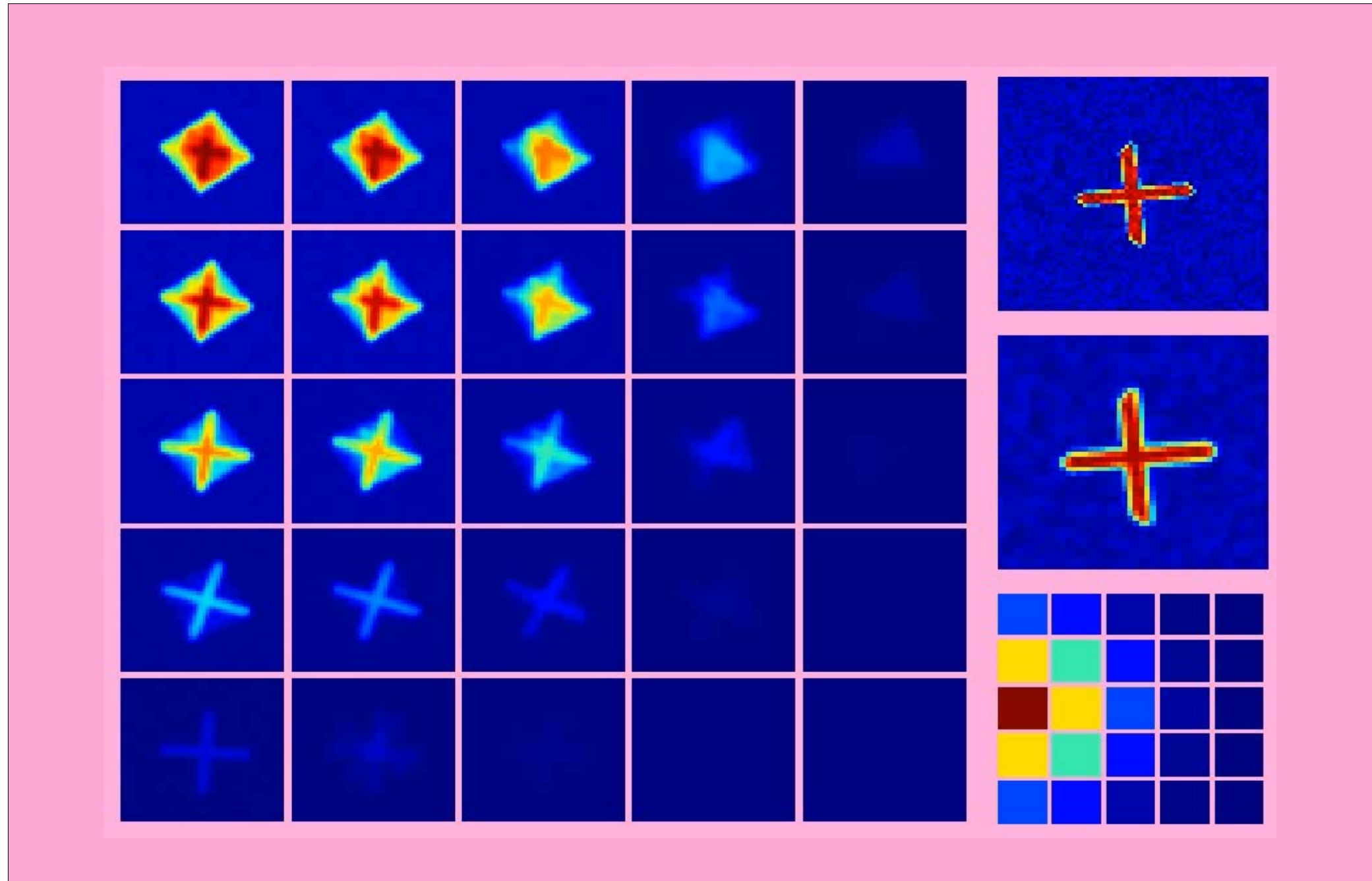
Example self-organizing map

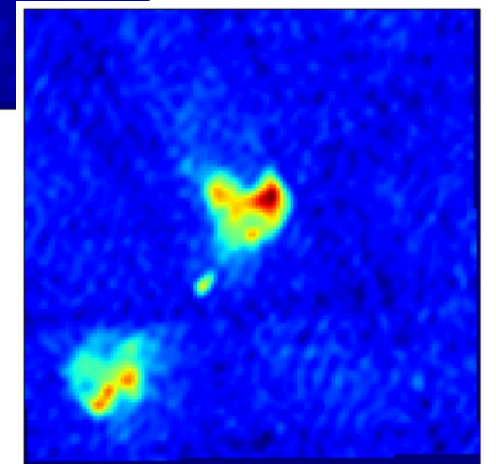
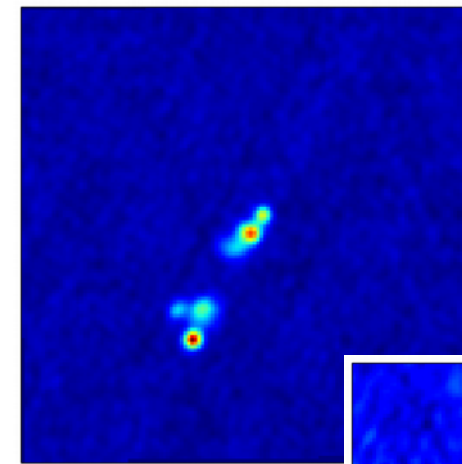
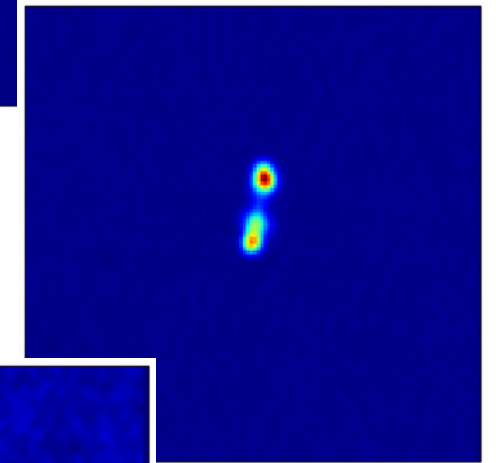
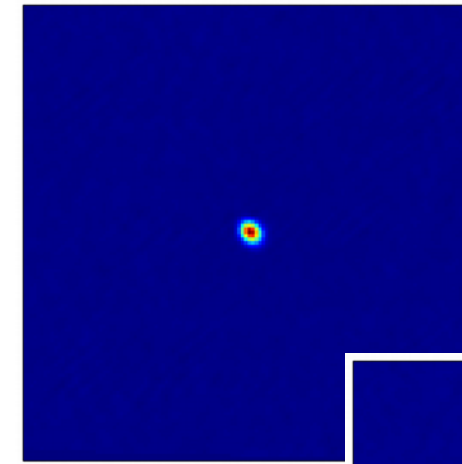
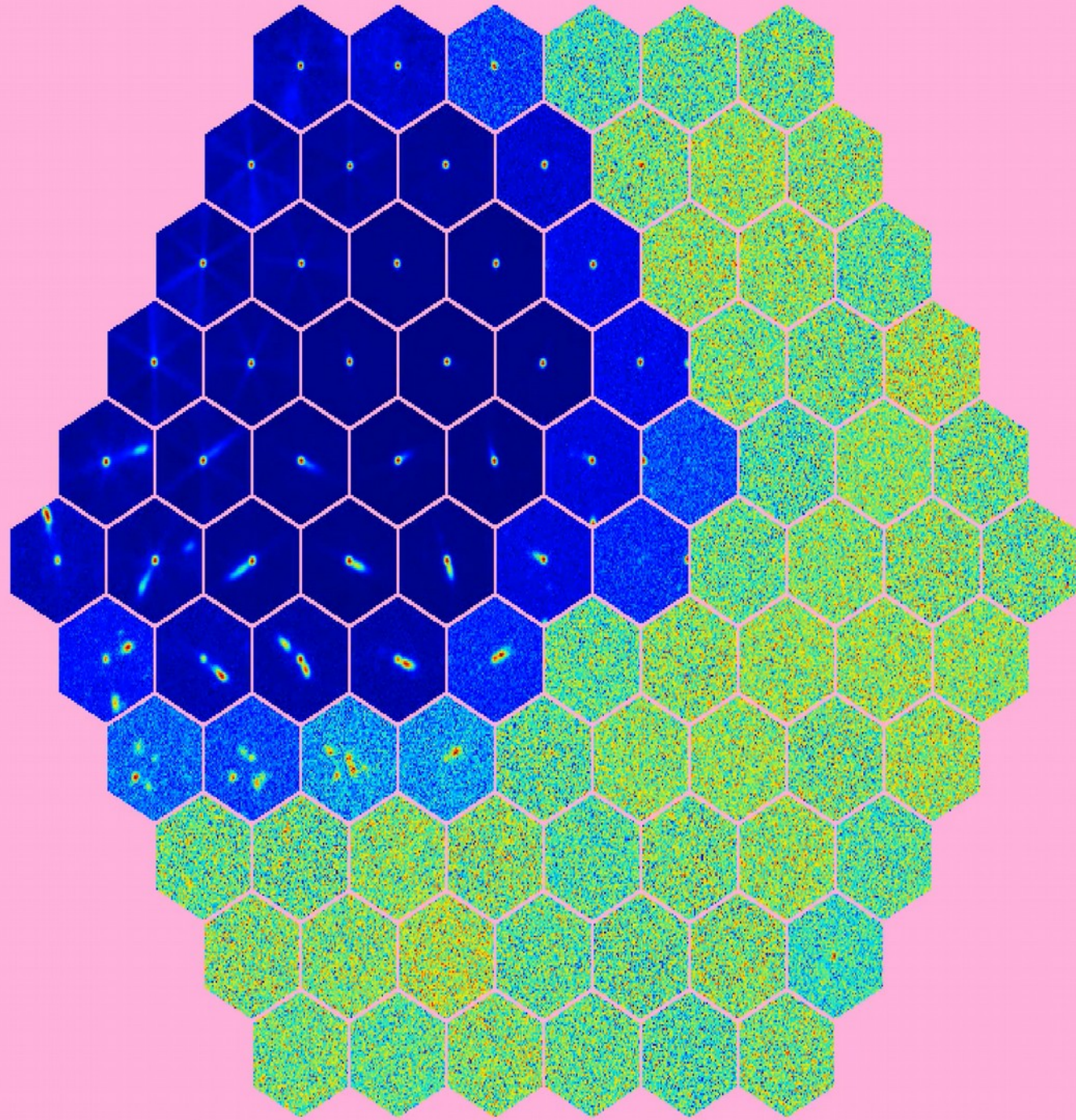


Example self-organizing map

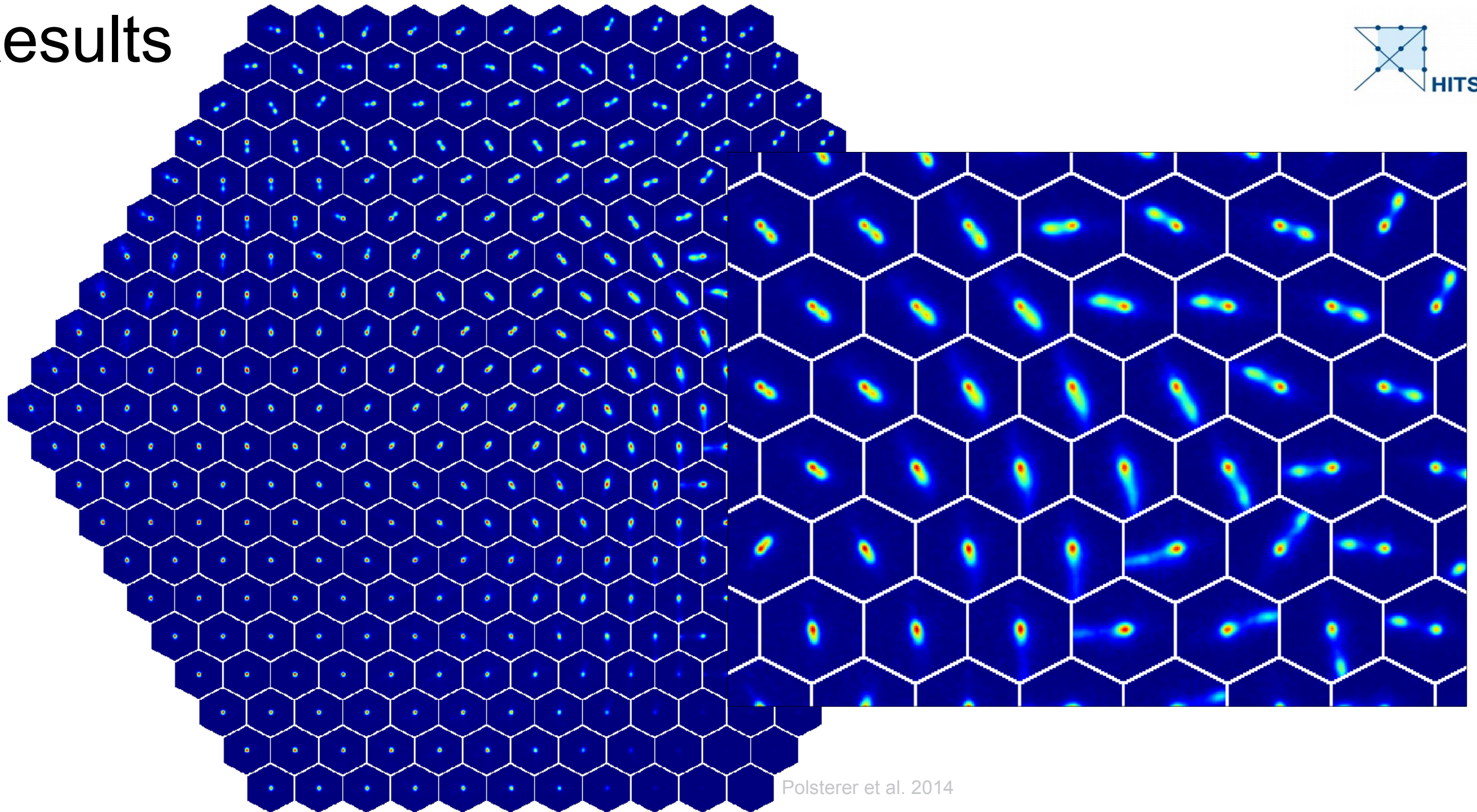


Example self-organizing map



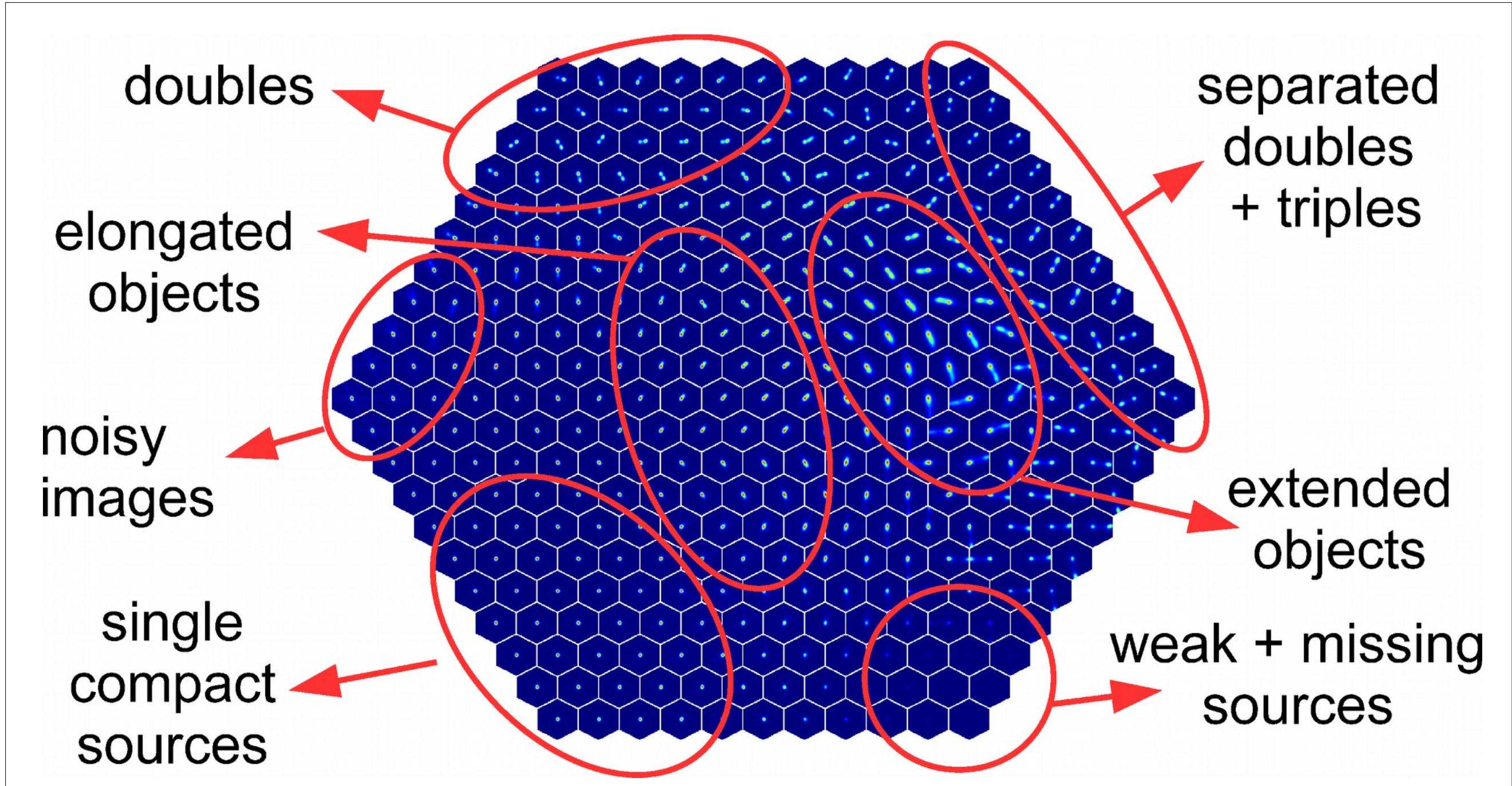


Results

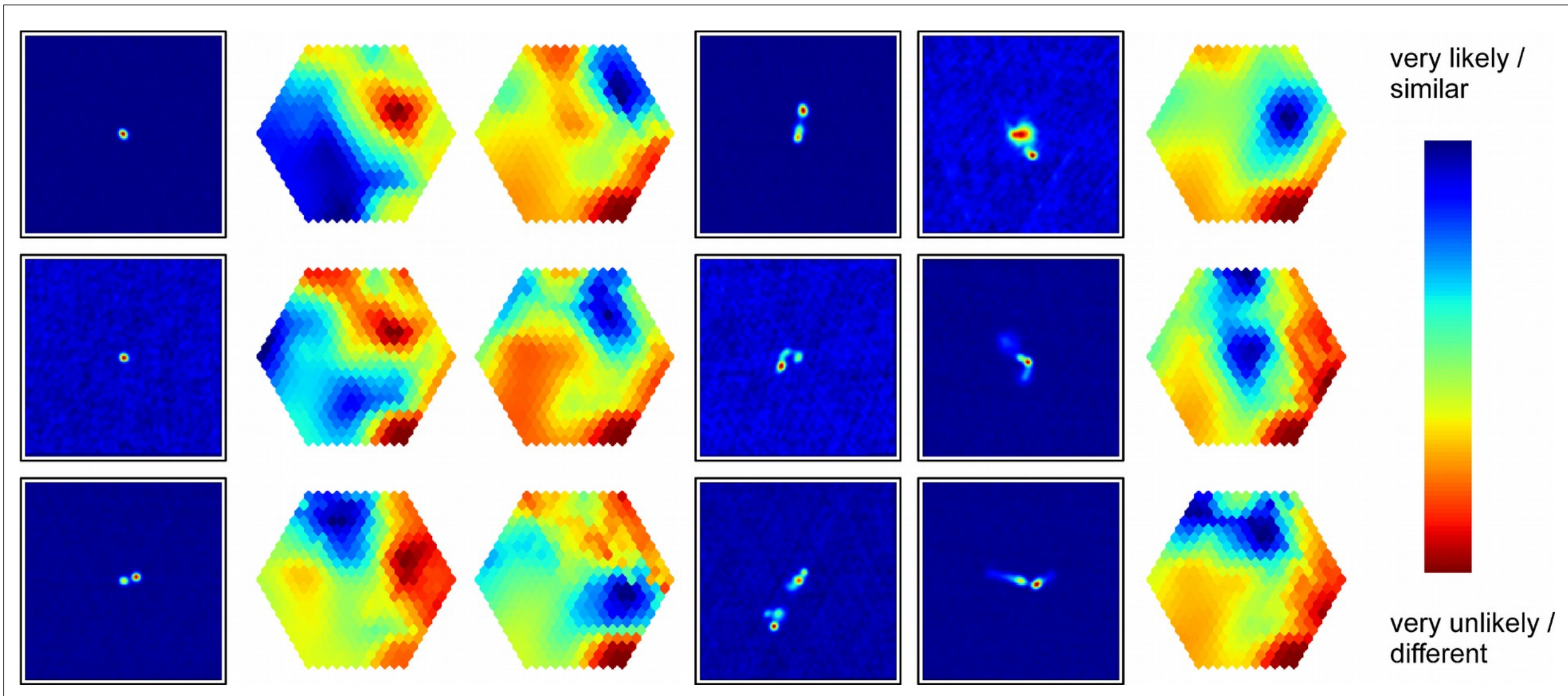
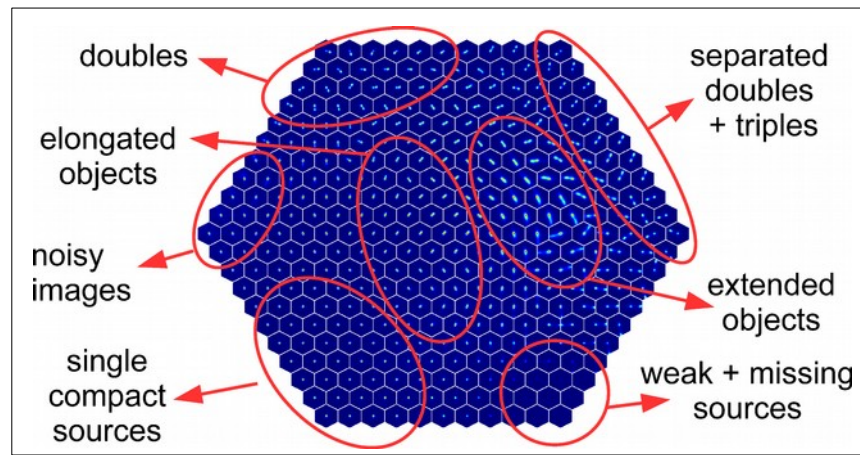


Polsterer et al. 2014

Results

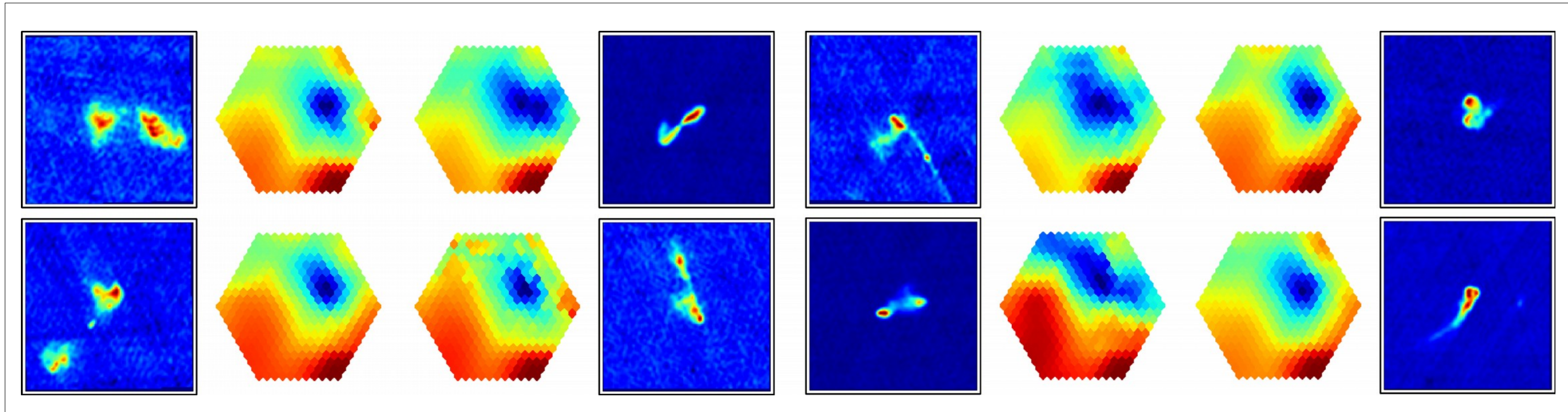
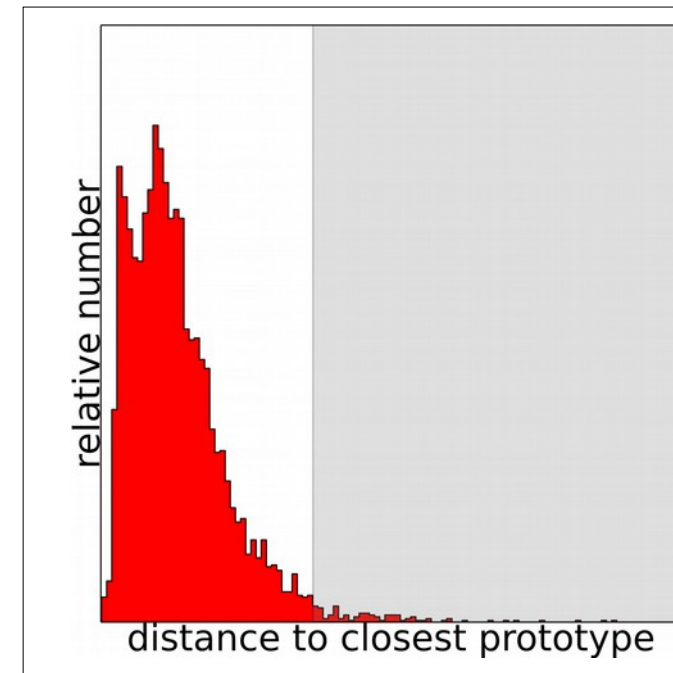


Results



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



ASTRON Netherlands Institute for Radio Astronomy Home Morphological outliers Downloads Acknowledgements

LOFAR-PINK Visualization Tool by Rafaël Mostert

11 06 46.893 +47 53 15.17

FoV: 11.98'

On the left you can see where the radio source you clicked on is located on the sky. The source might be accompanied or interacting with other sources or be part of some larger structure!

Show heatmap SOM properties

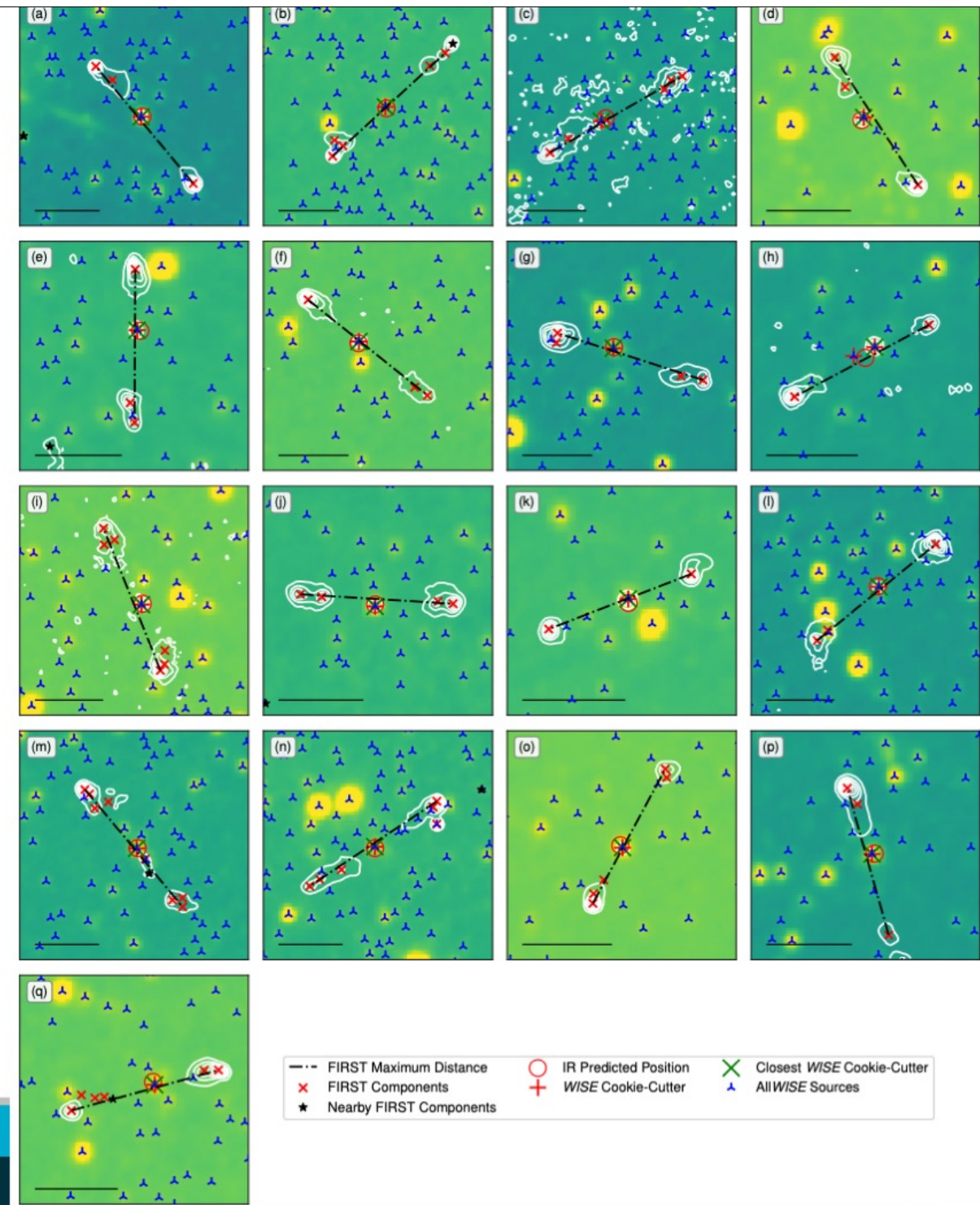
Radio sources from LOFAR survey that resemble the selected prototype (19,17):

The screenshot shows the LOFAR web-tool interface. On the left is a heatmap of the sky with a grid of small arrows. A red box highlights a specific location. Below the heatmap are two buttons: 'Show heatmap' and 'SOM properties'. To the right is a zoomed-in image of the selected radio source, showing a bright, elongated structure. Above the zoomed image is the coordinate '11 06 46.893 +47 53 15.17' and a field of view indicator 'FoV: 11.98''. Below the zoomed image is a caption 'Radio sources from LOFAR survey that resemble the selected prototype (19,17):' followed by five small thumbnail images of similar radio sources. The second thumbnail from the left is highlighted with a red border.

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



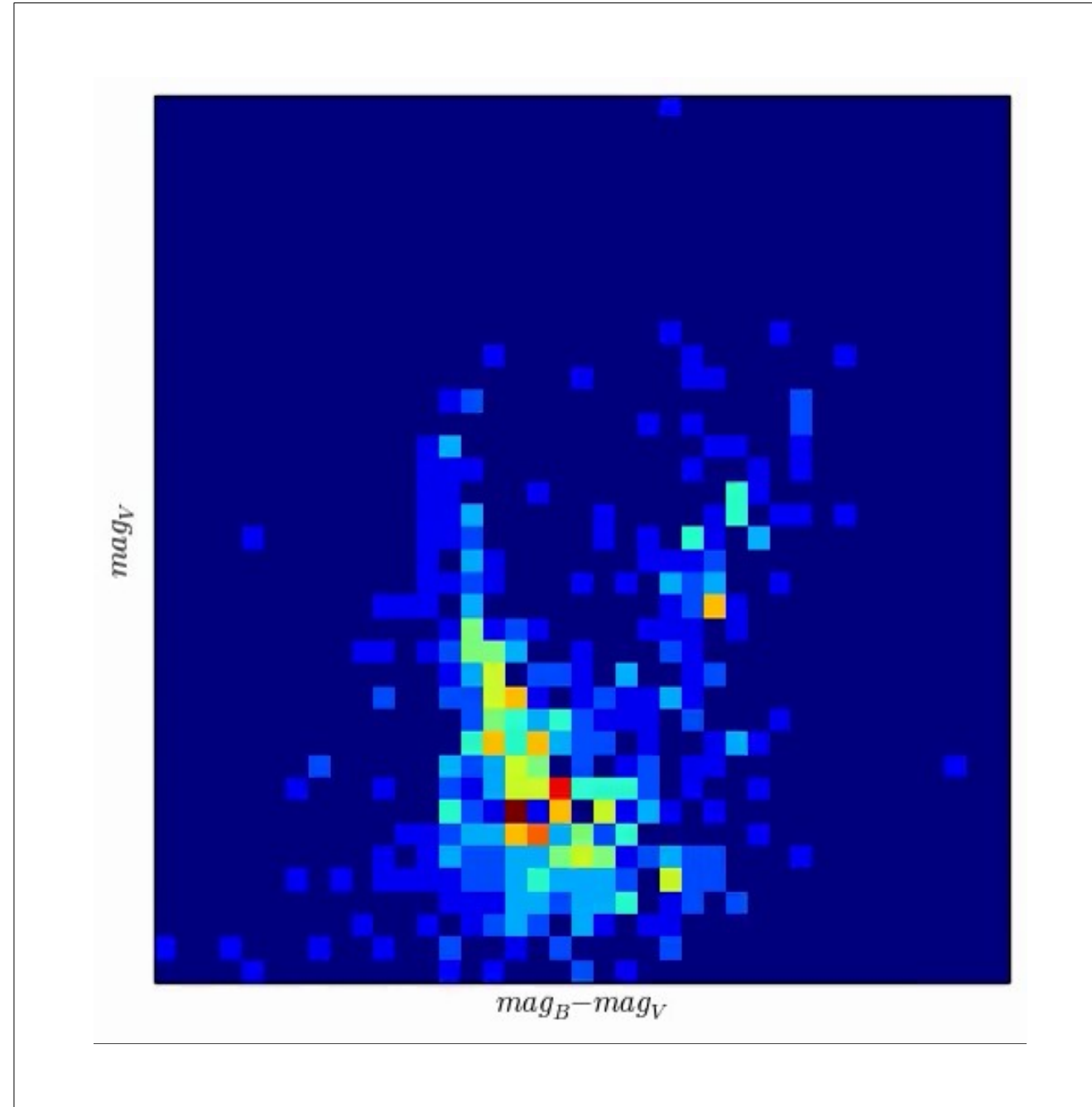
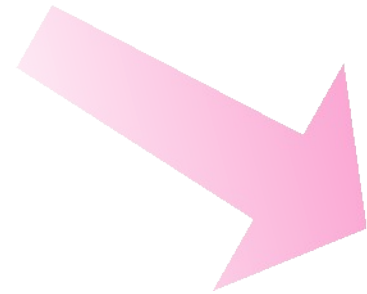
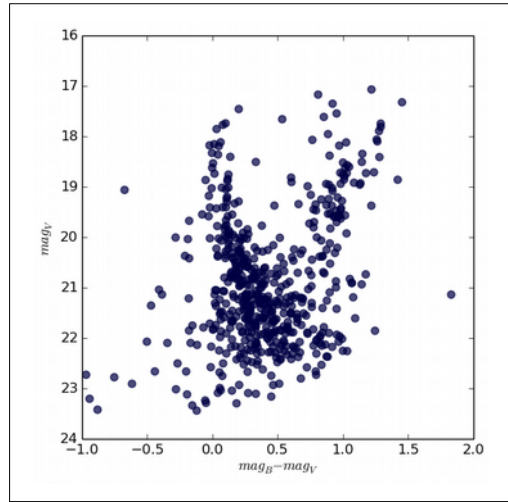
Tim Galvin CSIRO

Starformation history



APOD, Roger Smith

Analysis of stellar cluster

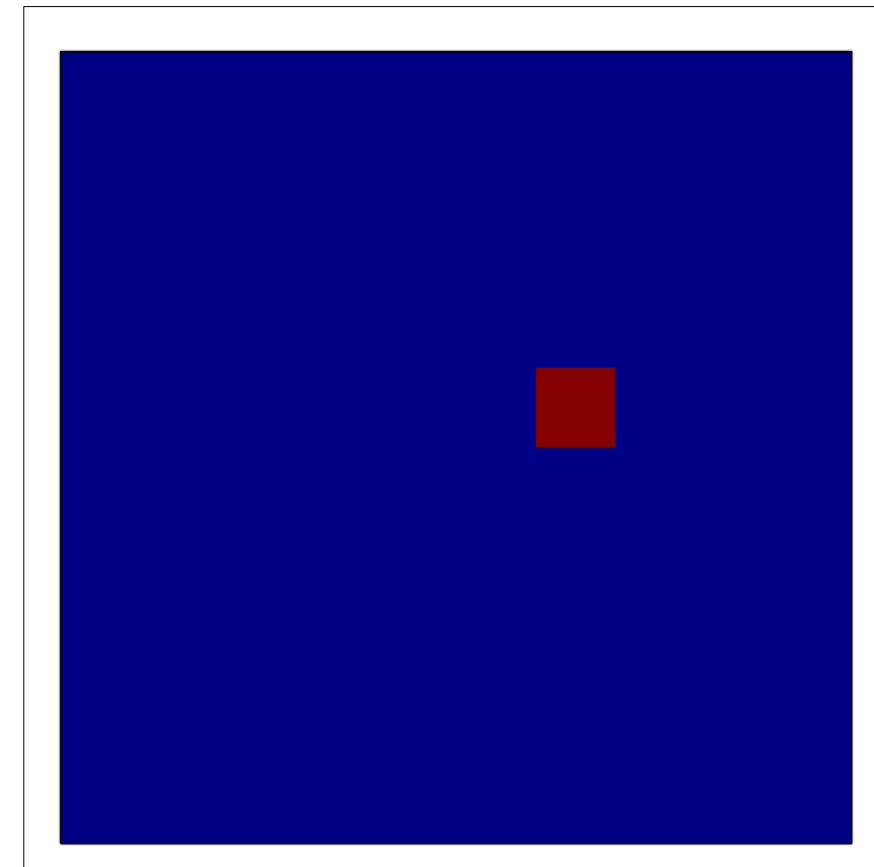
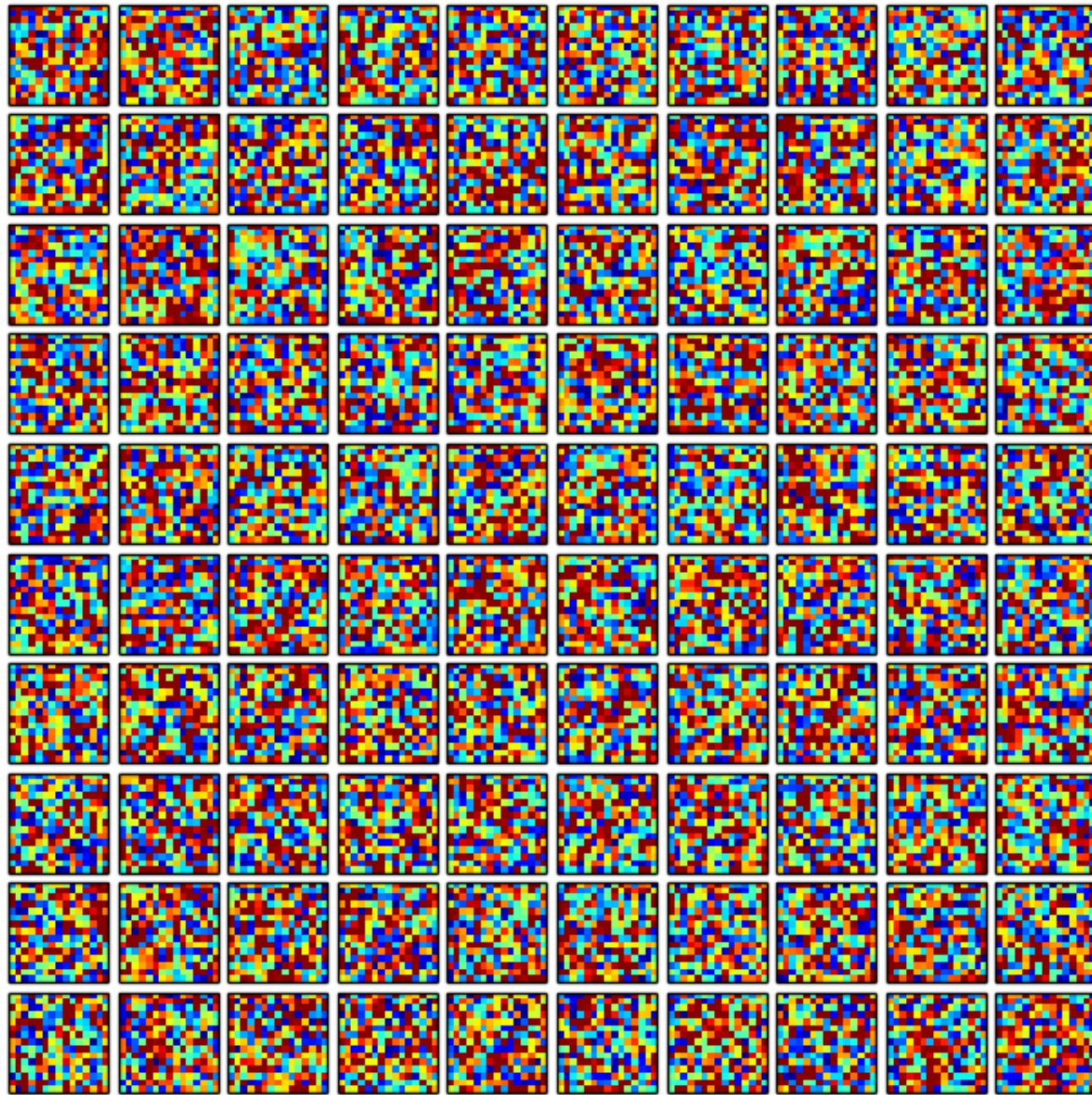


Dimensionality reduction

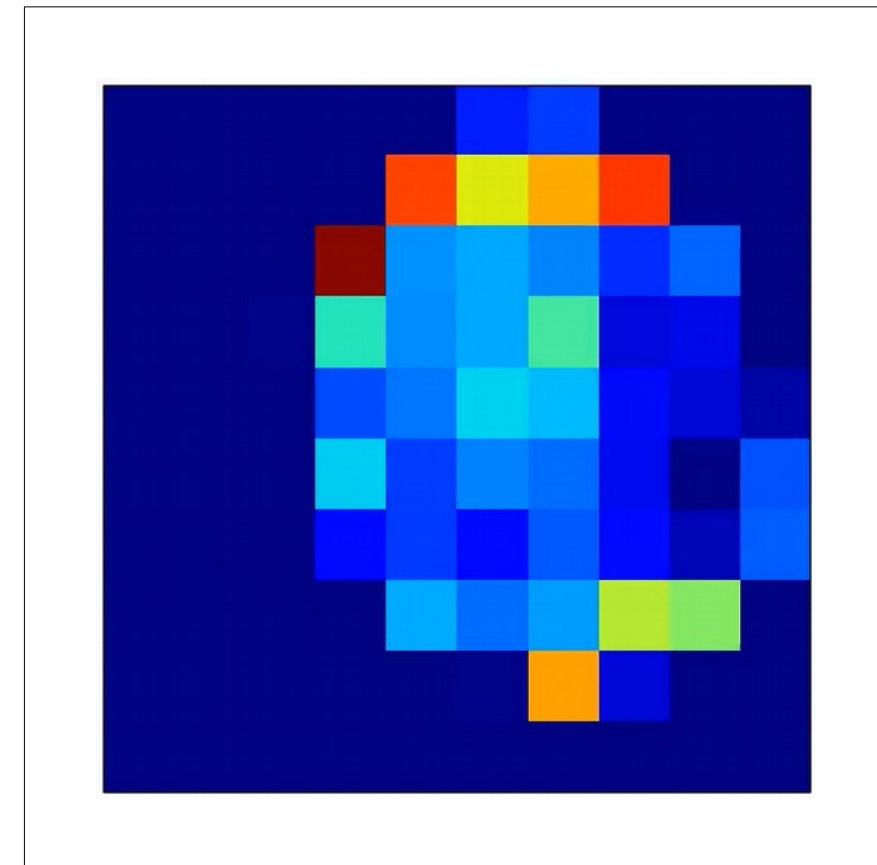
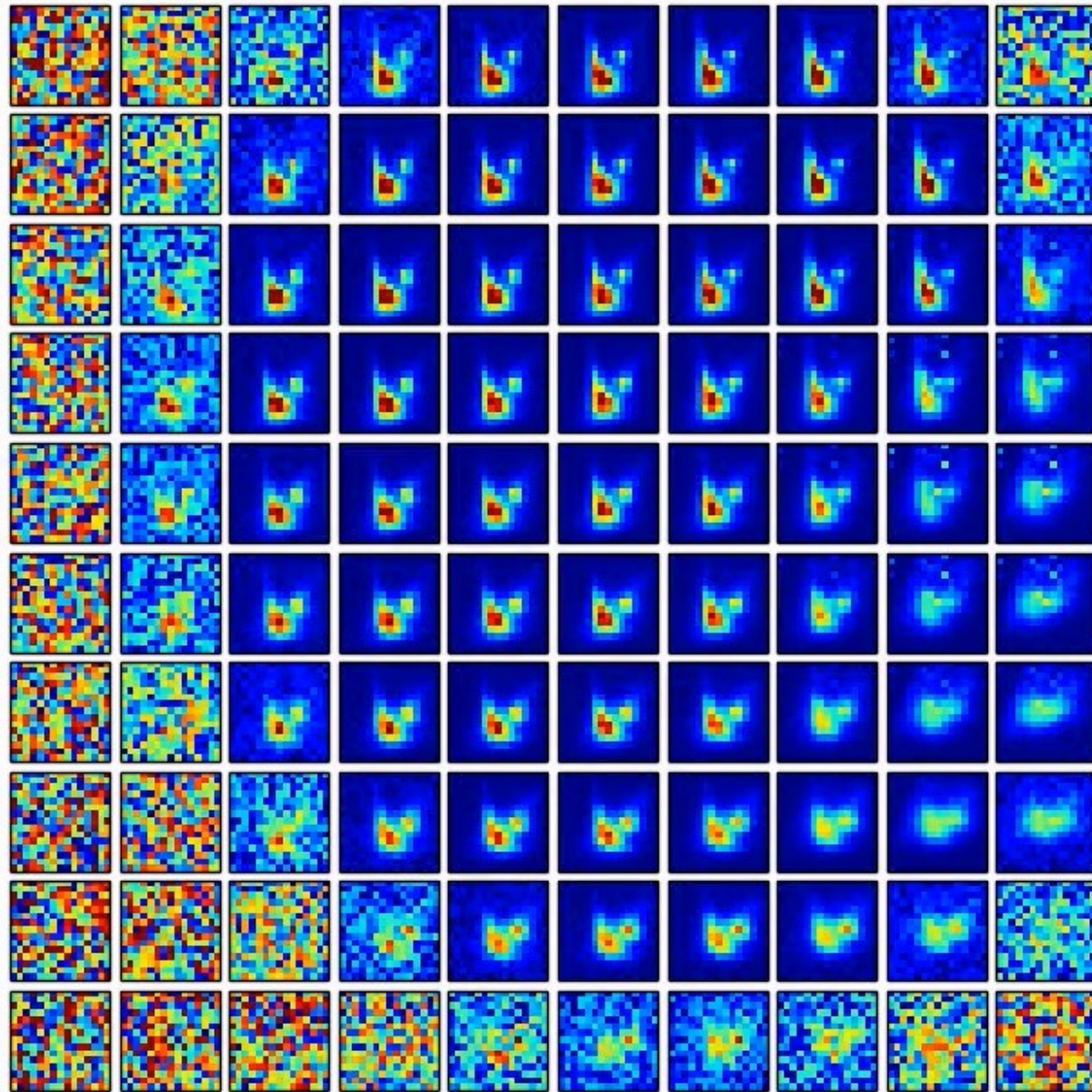


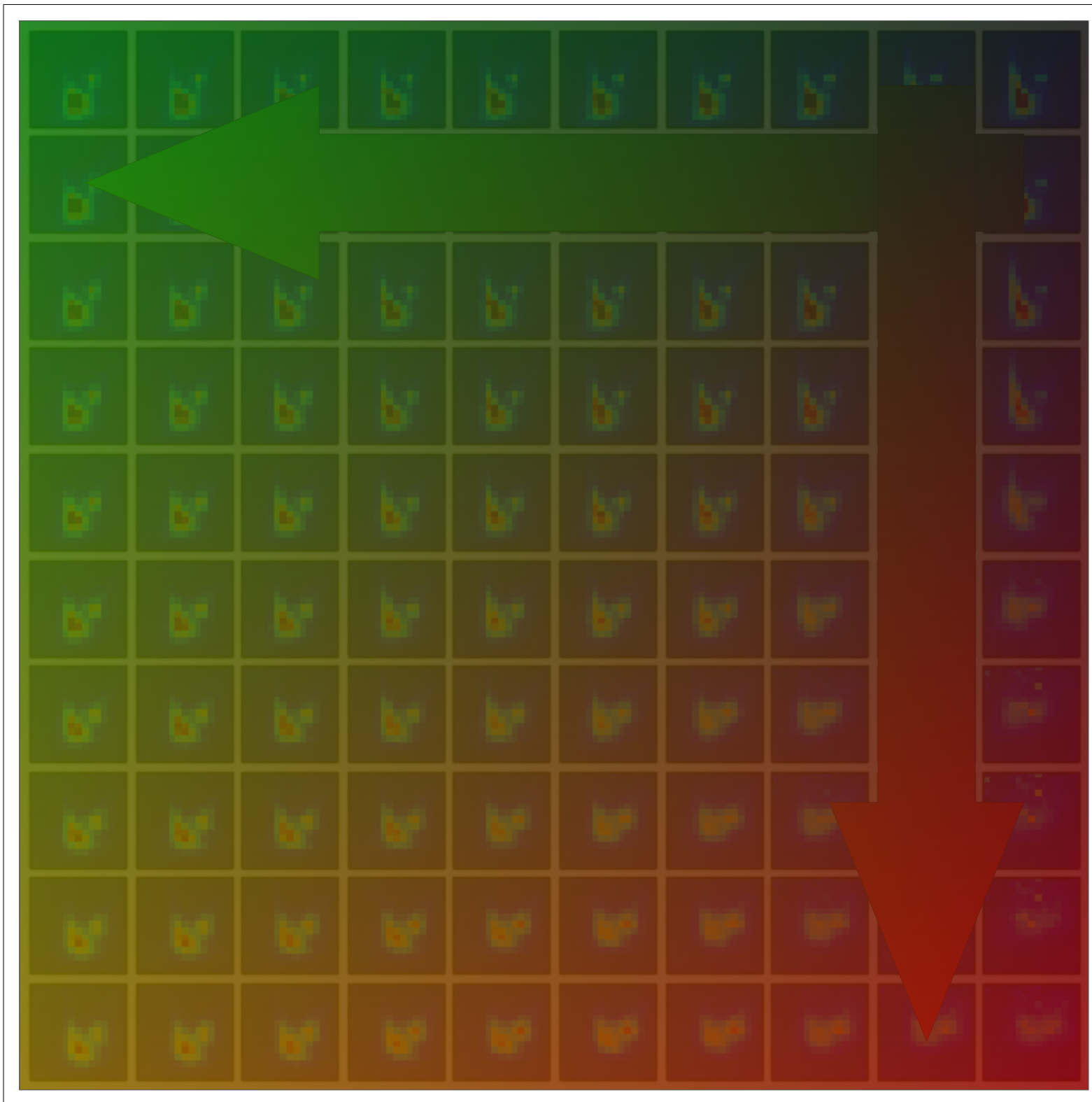
$$\Delta(A, B) = \sqrt{\sum_{x=1}^{D_x} \sum_{y=1}^{D_y} (A_{xy} - B_{xy})^2}$$

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



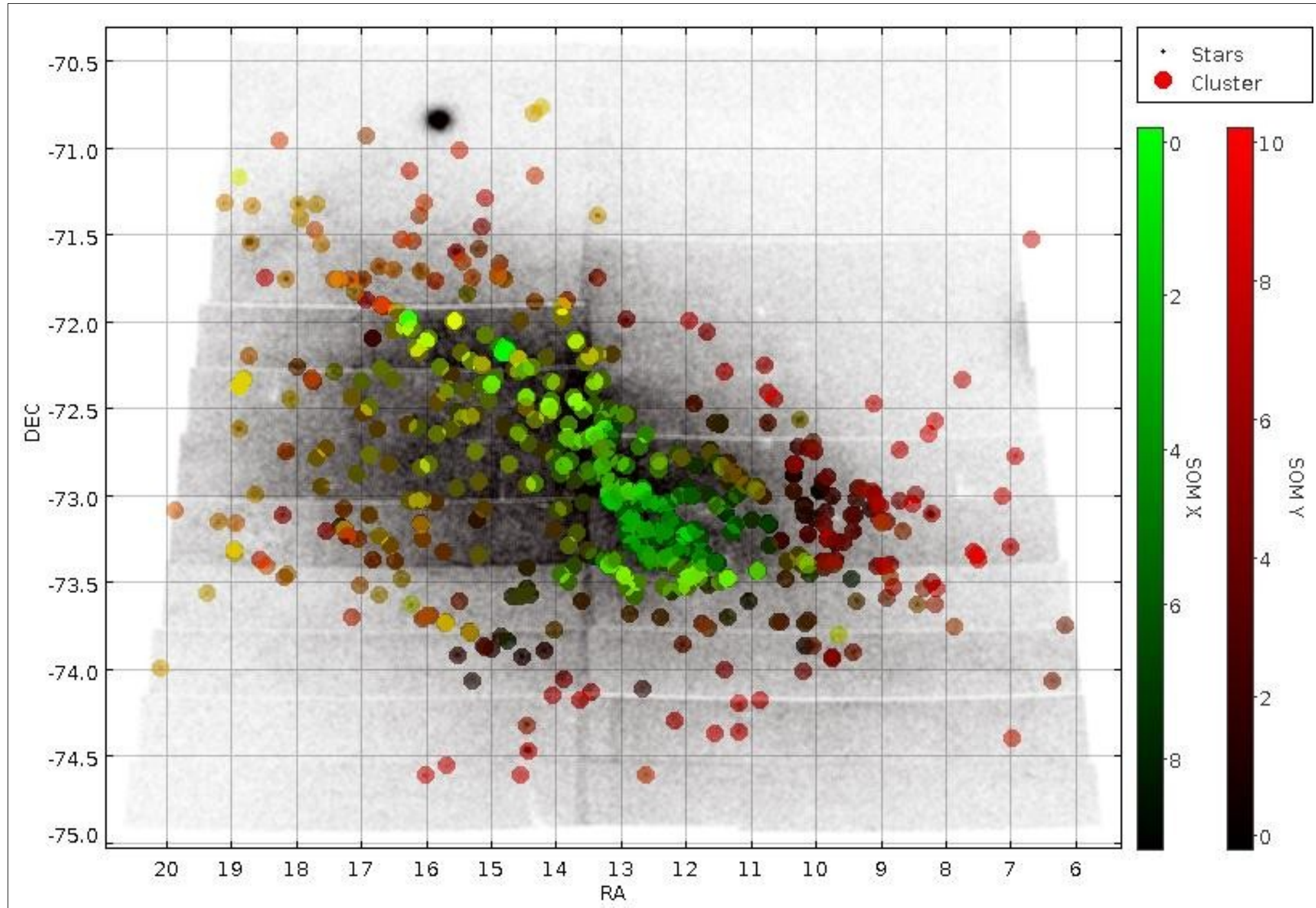
Dimensionality reduction





what is it
good for?

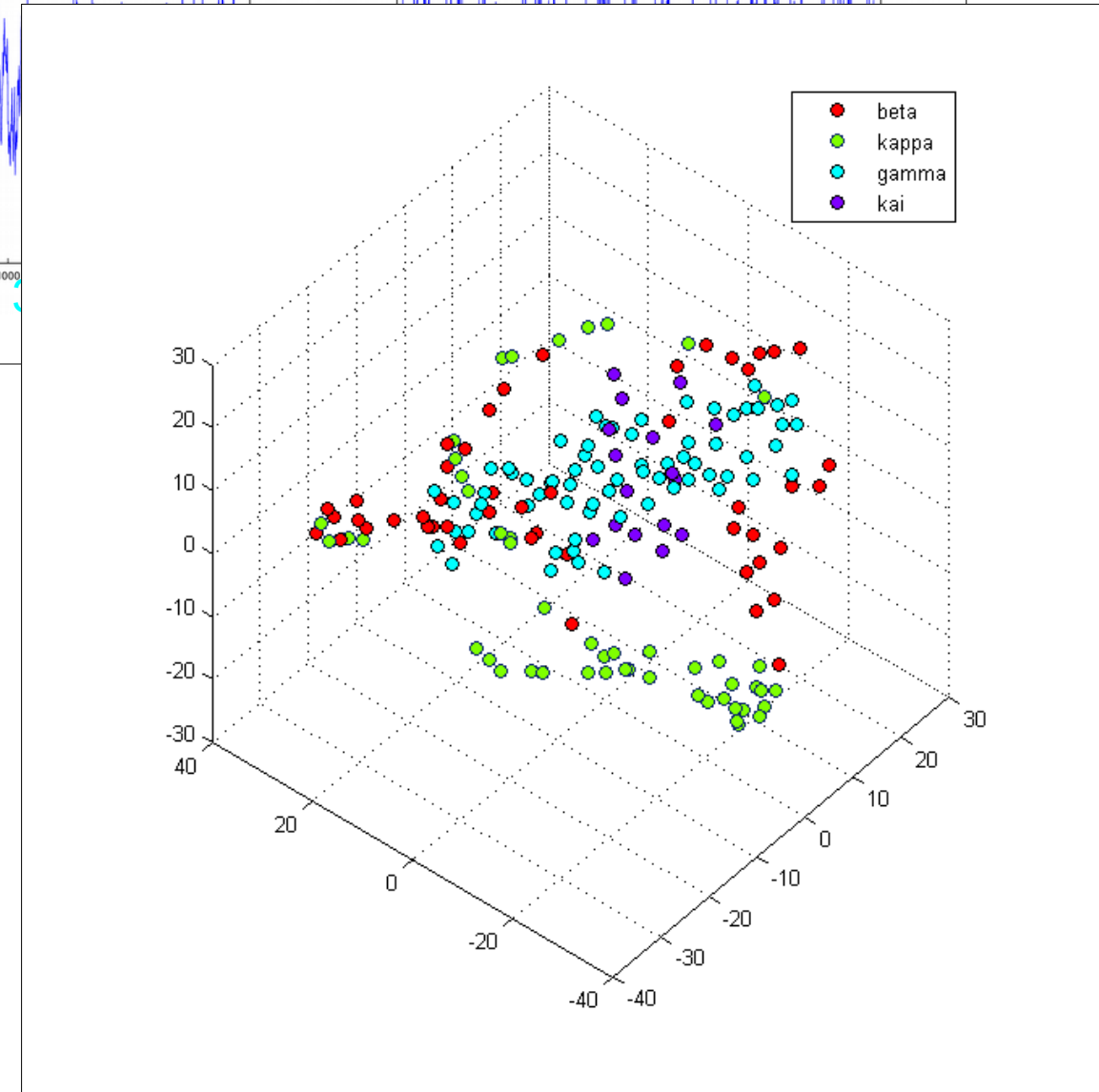
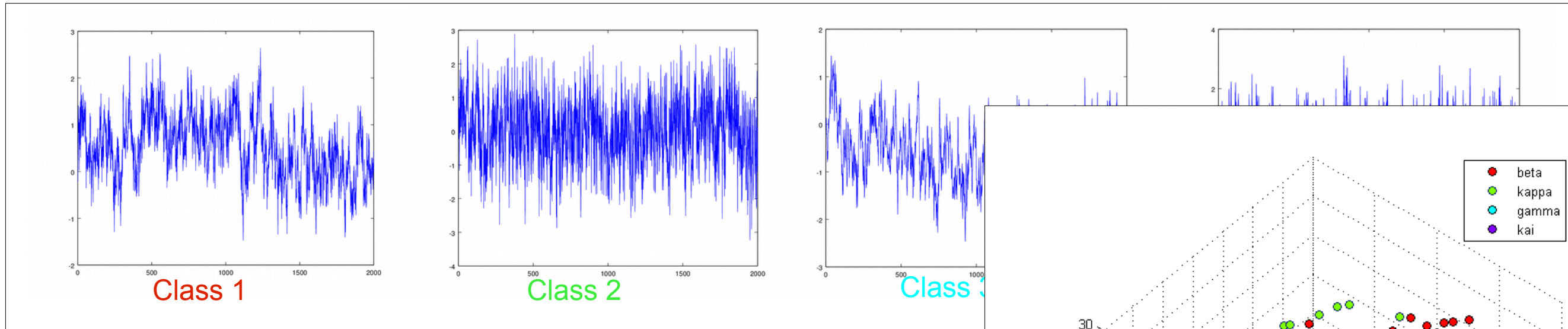
Results



Time Series Analysis

taking temporal nature into account

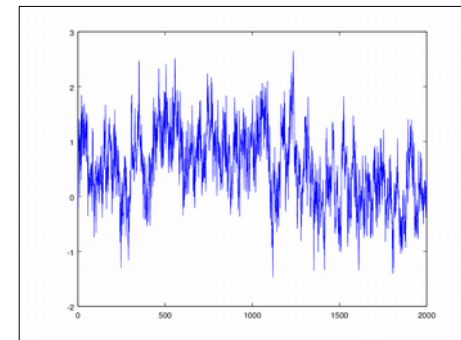
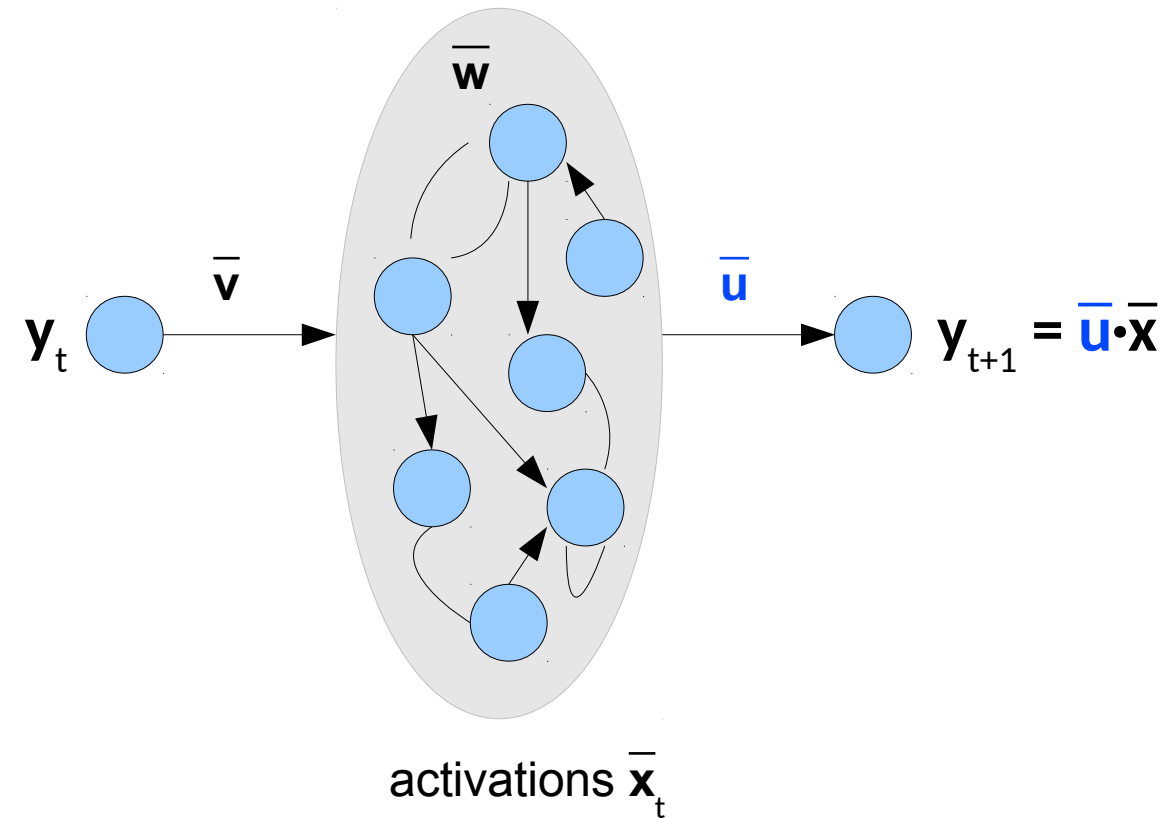
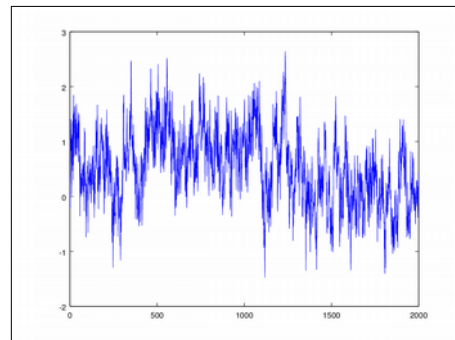
Time series analysis



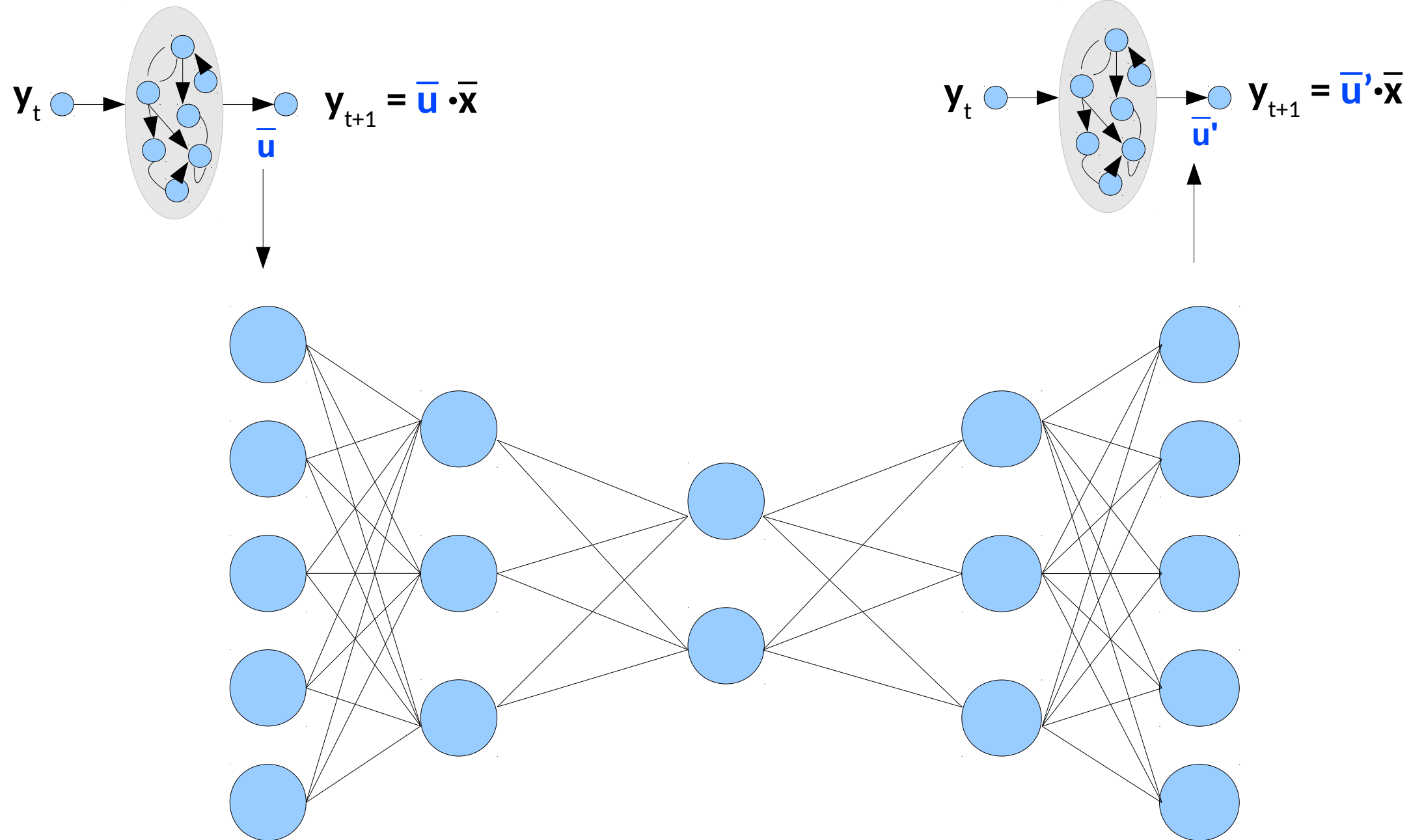
treat merely
as vectors

Recurrent neural network

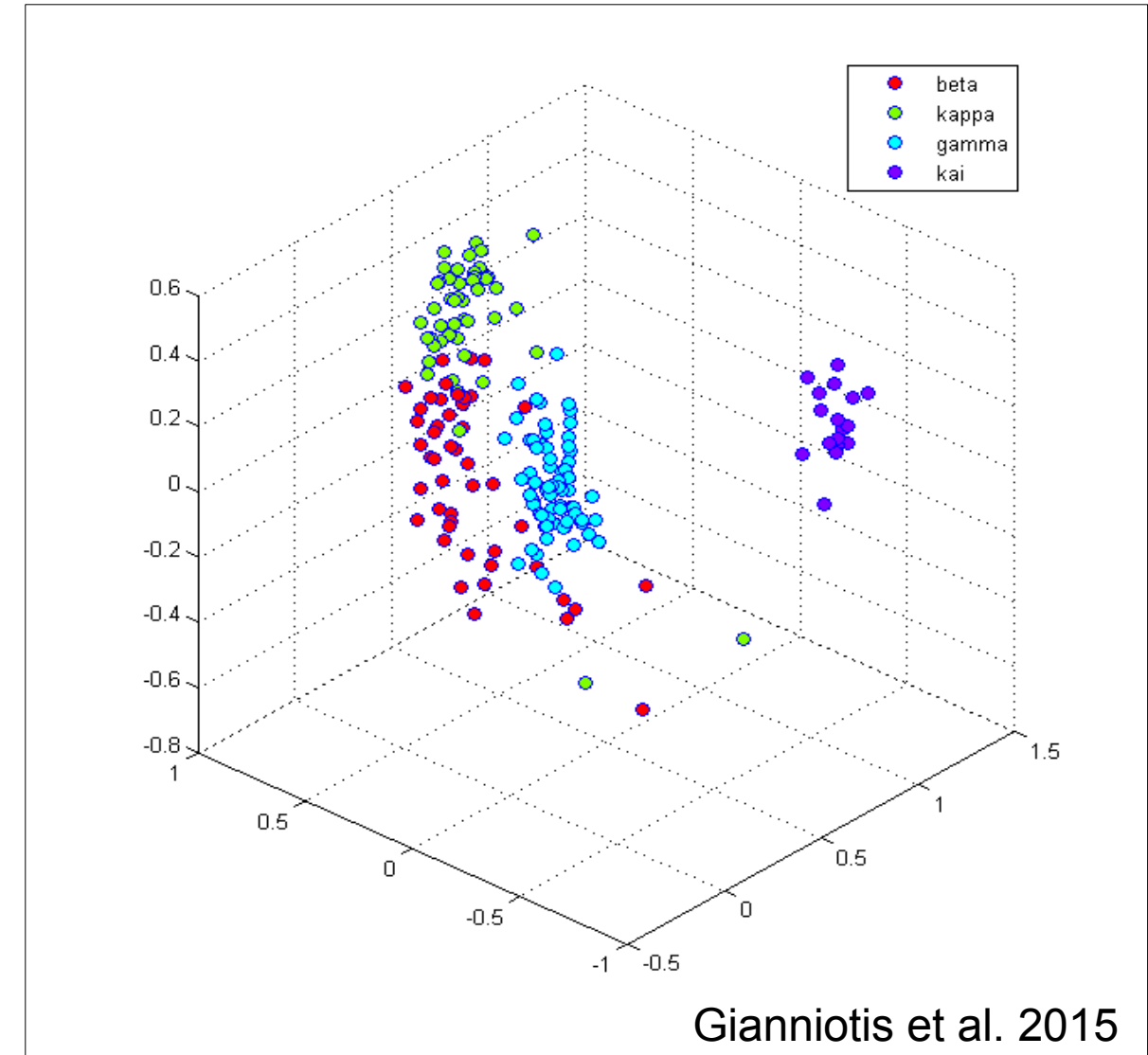
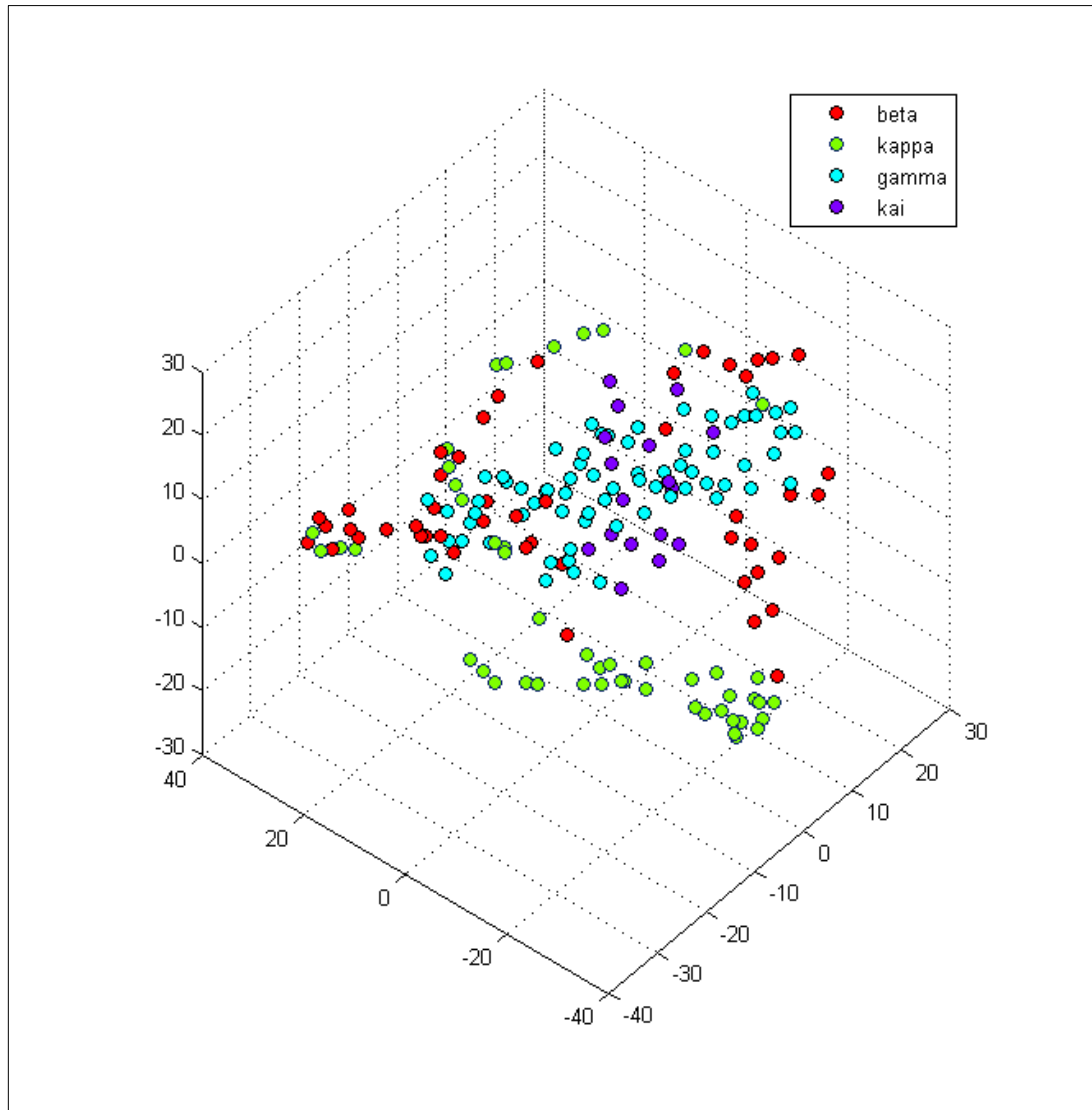
take temporal nature into account



ECN + Autoencoder



Time series analysis

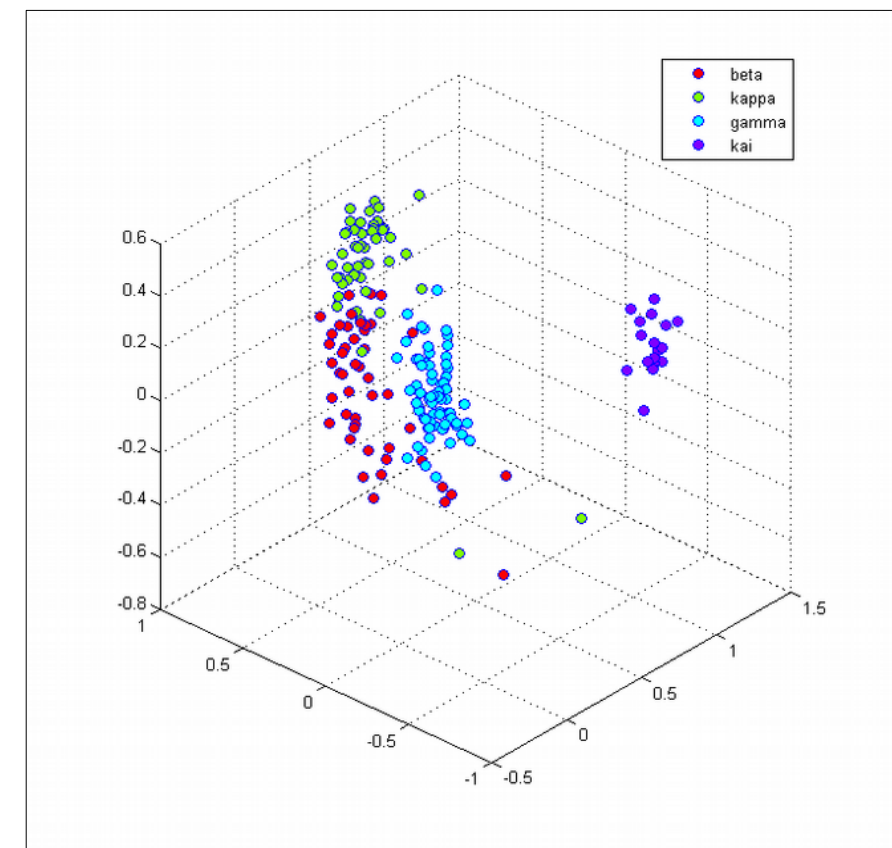
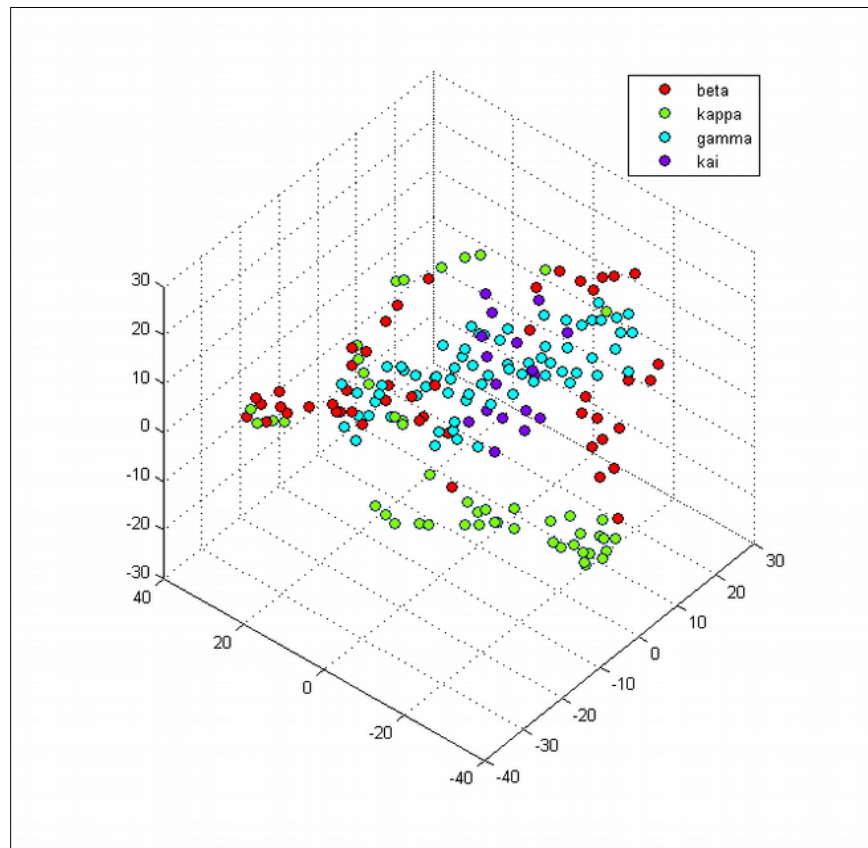


Gianniotis et al. 2015

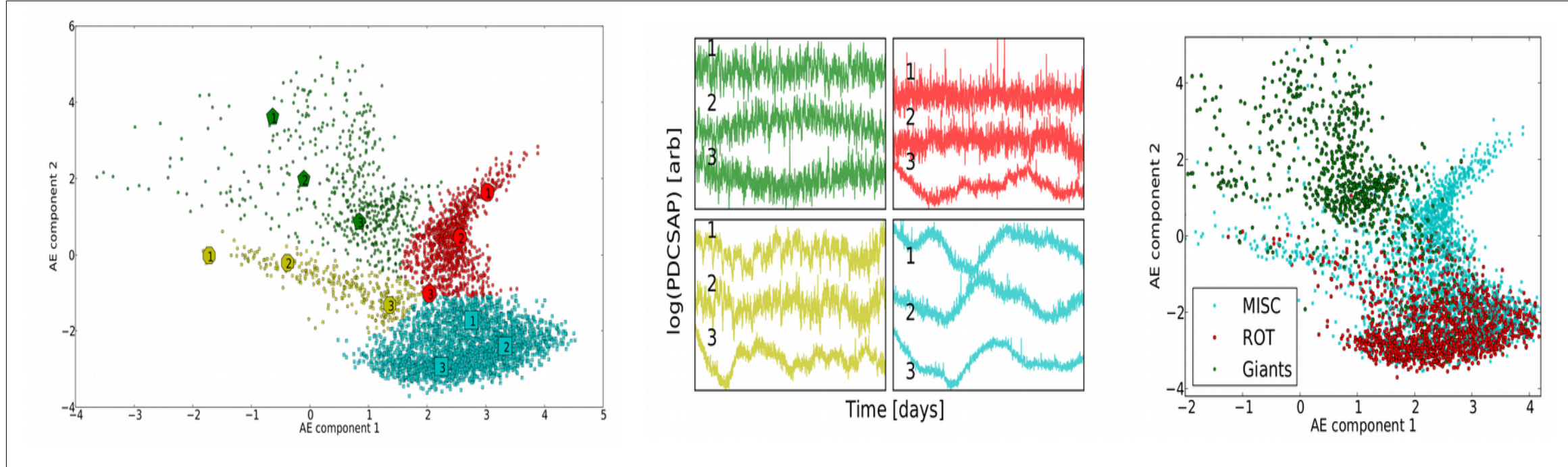
Lessons learned



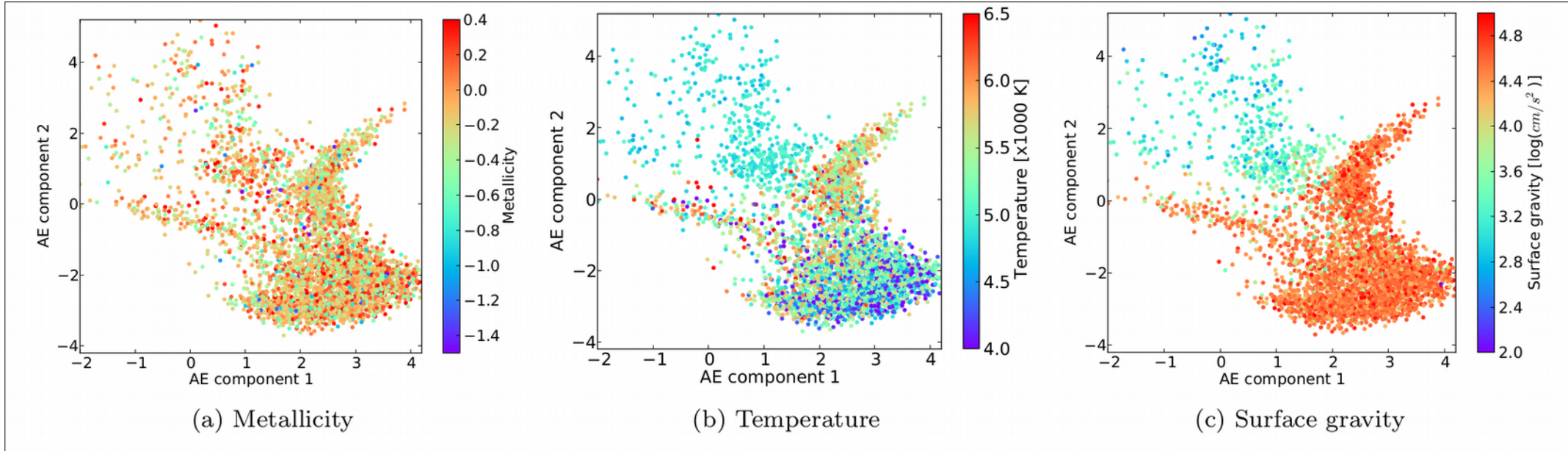
temporal behavior is **more than** a simple **vector**



Kepler data / stellar objects

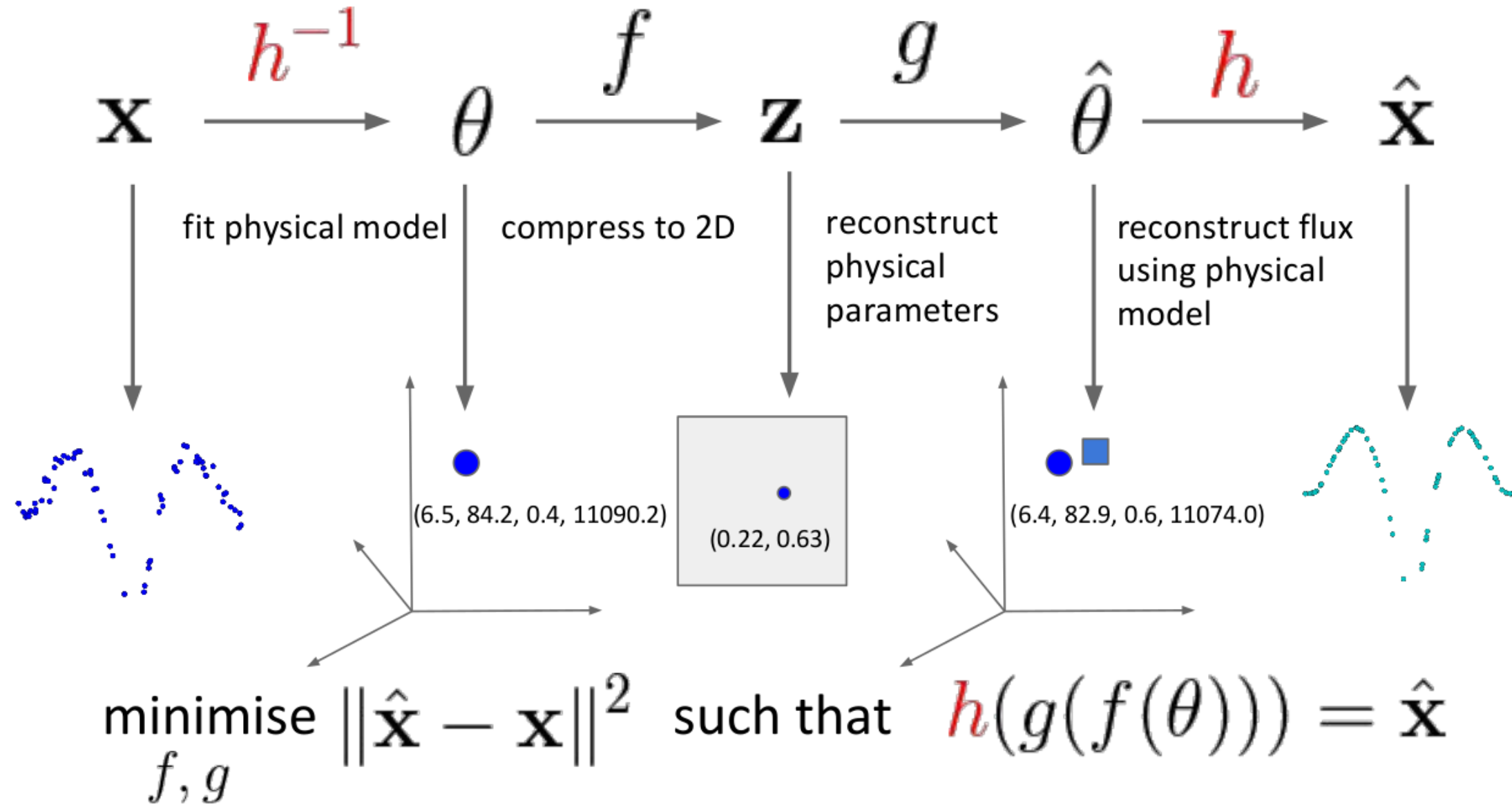


Kuegler et al. 2015



Physical model & autoencoder

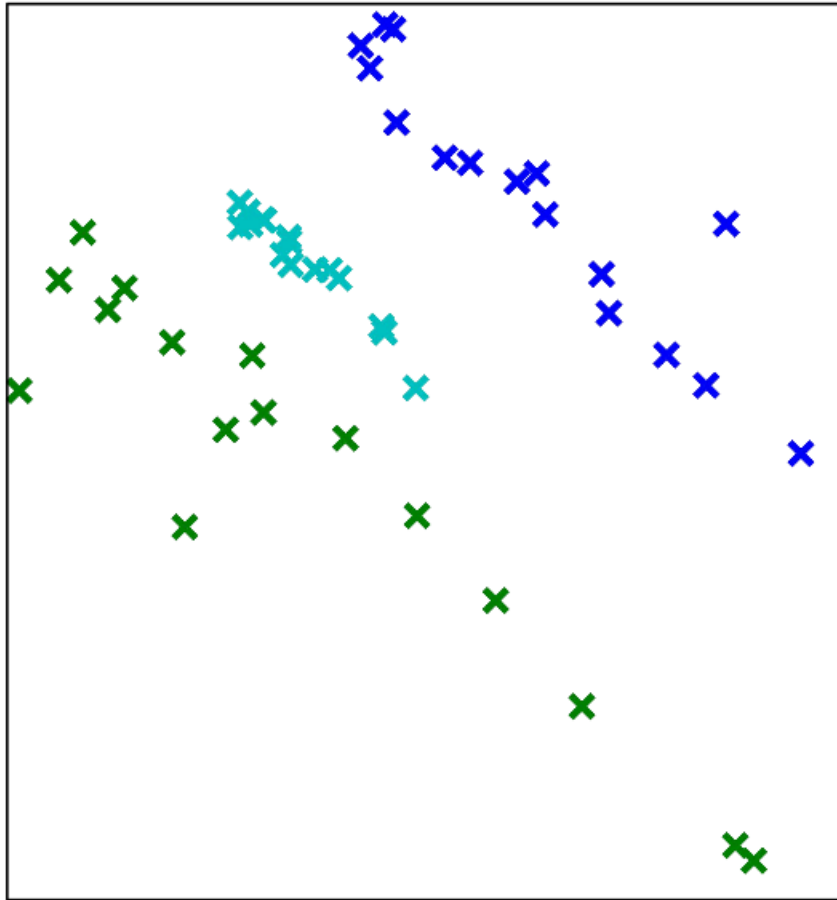
Timeseries of stellar binary systems



Project data



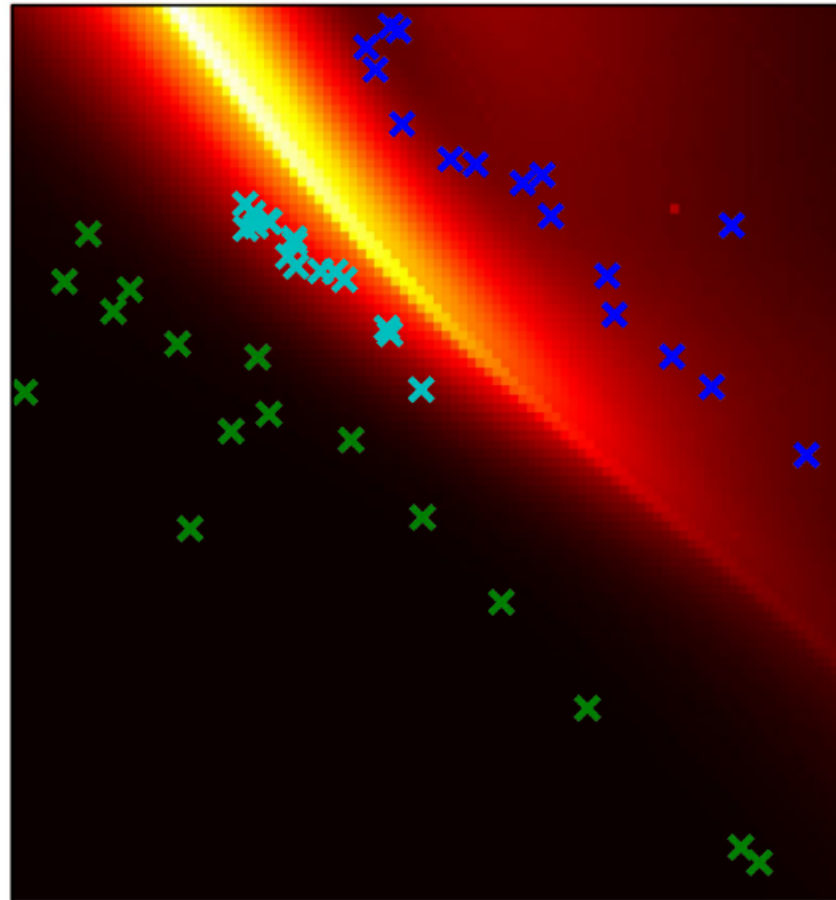
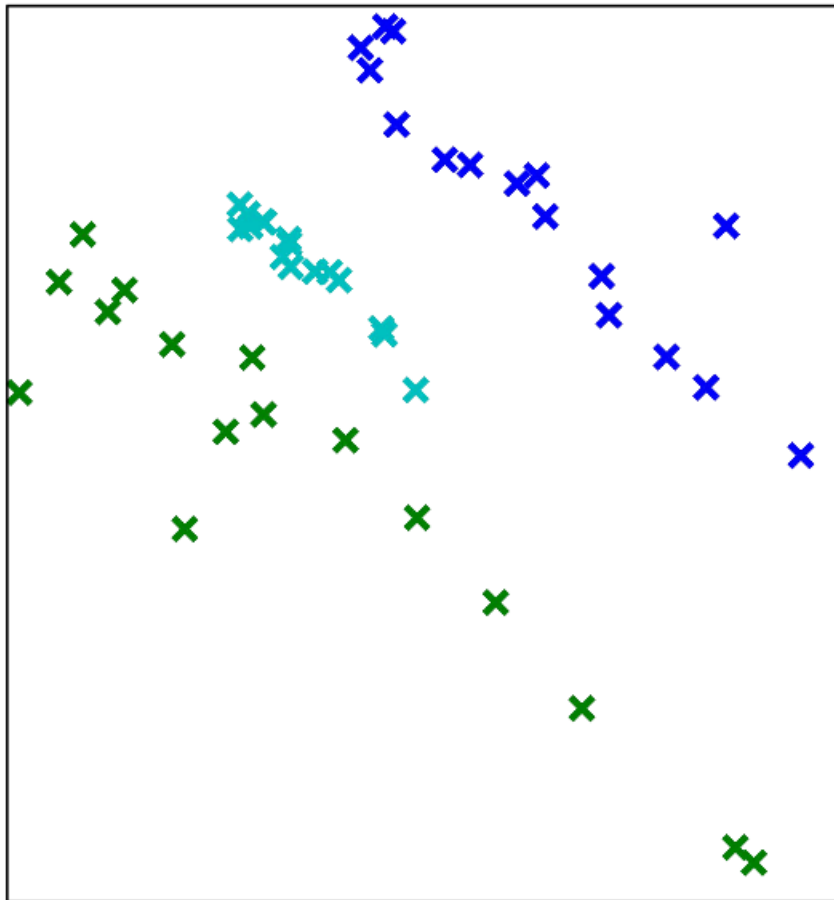
$$\mathbf{x} \xrightarrow{h^{-1}} \theta \xrightarrow{f} \mathbf{z} \xrightarrow{g} \hat{\theta} \xrightarrow{h} \hat{\mathbf{x}}$$



Analyze distance

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

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

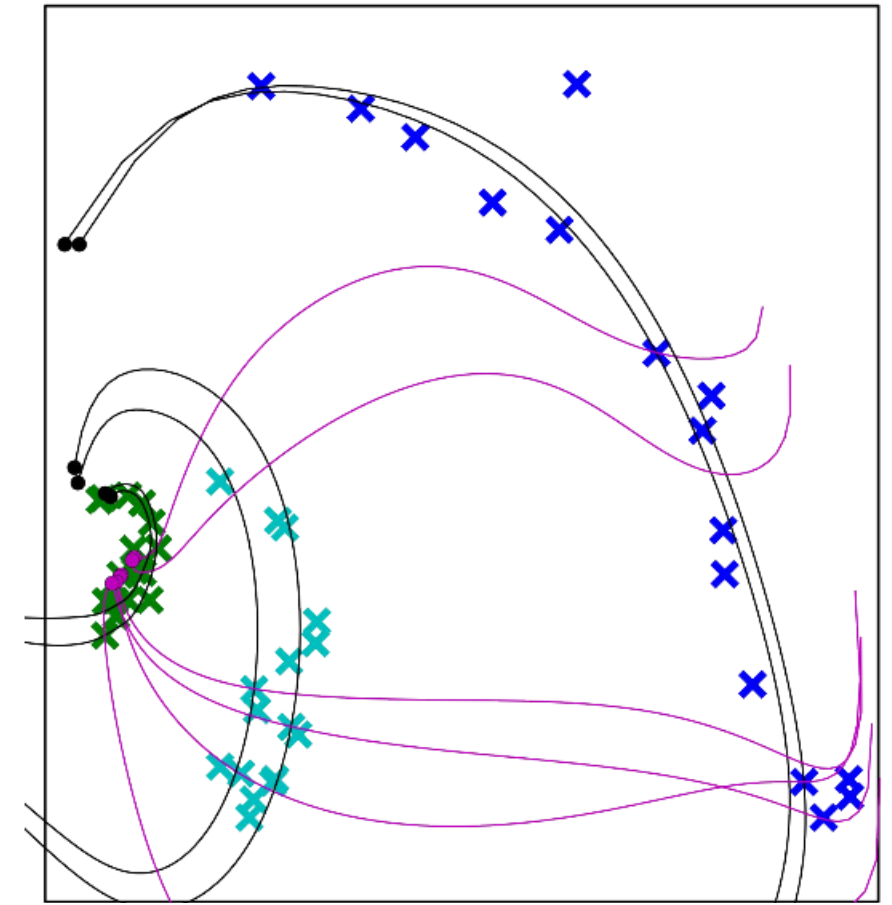
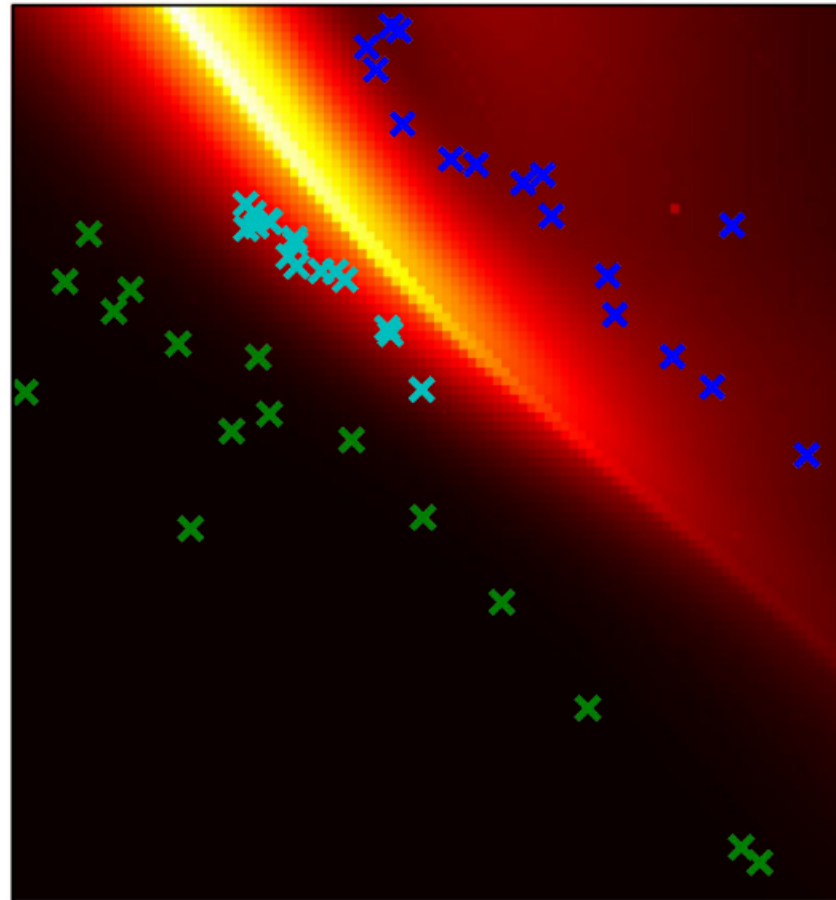
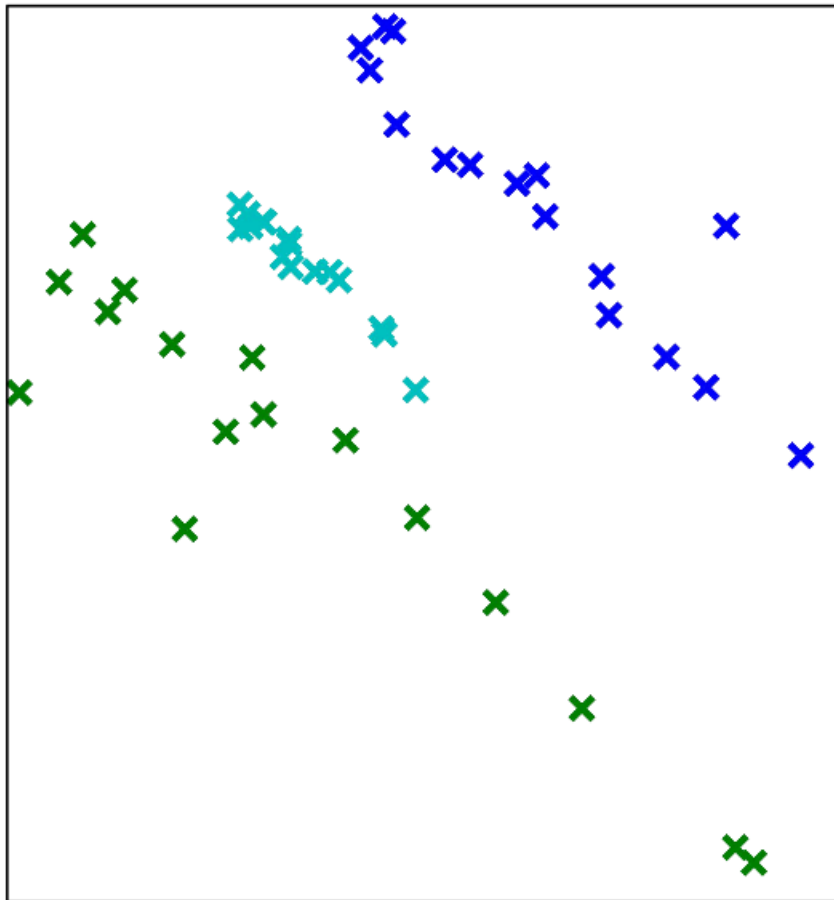


Plot iso-lines



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

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



Mass, Temperature

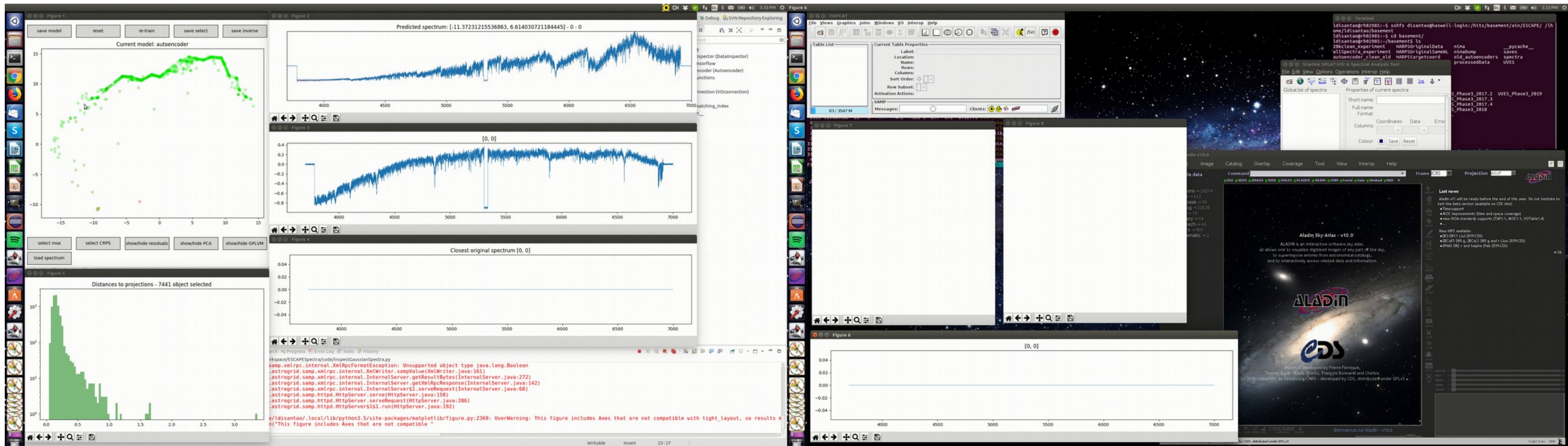
Spectral Data Analysis

dealing with spectral data

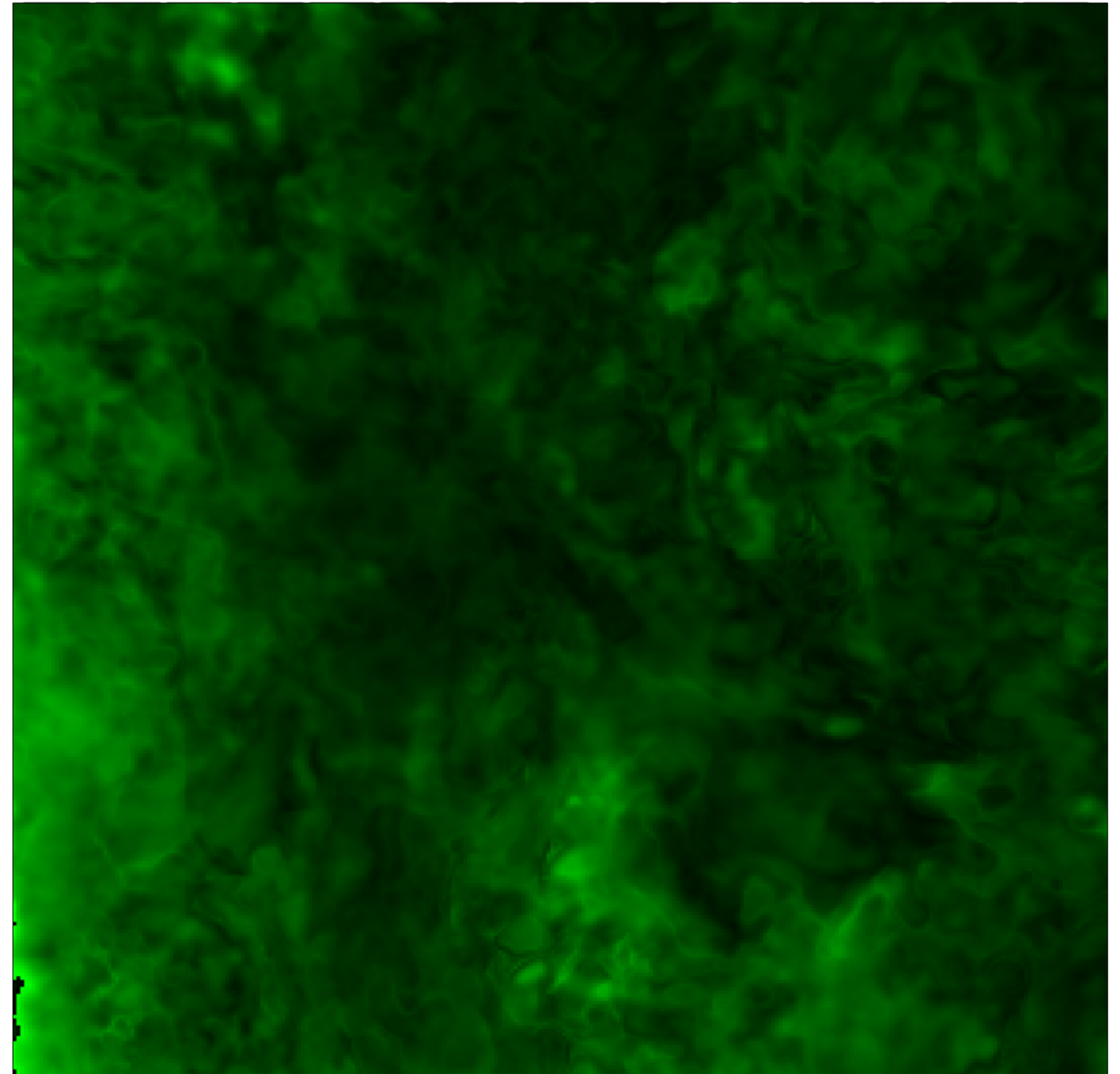
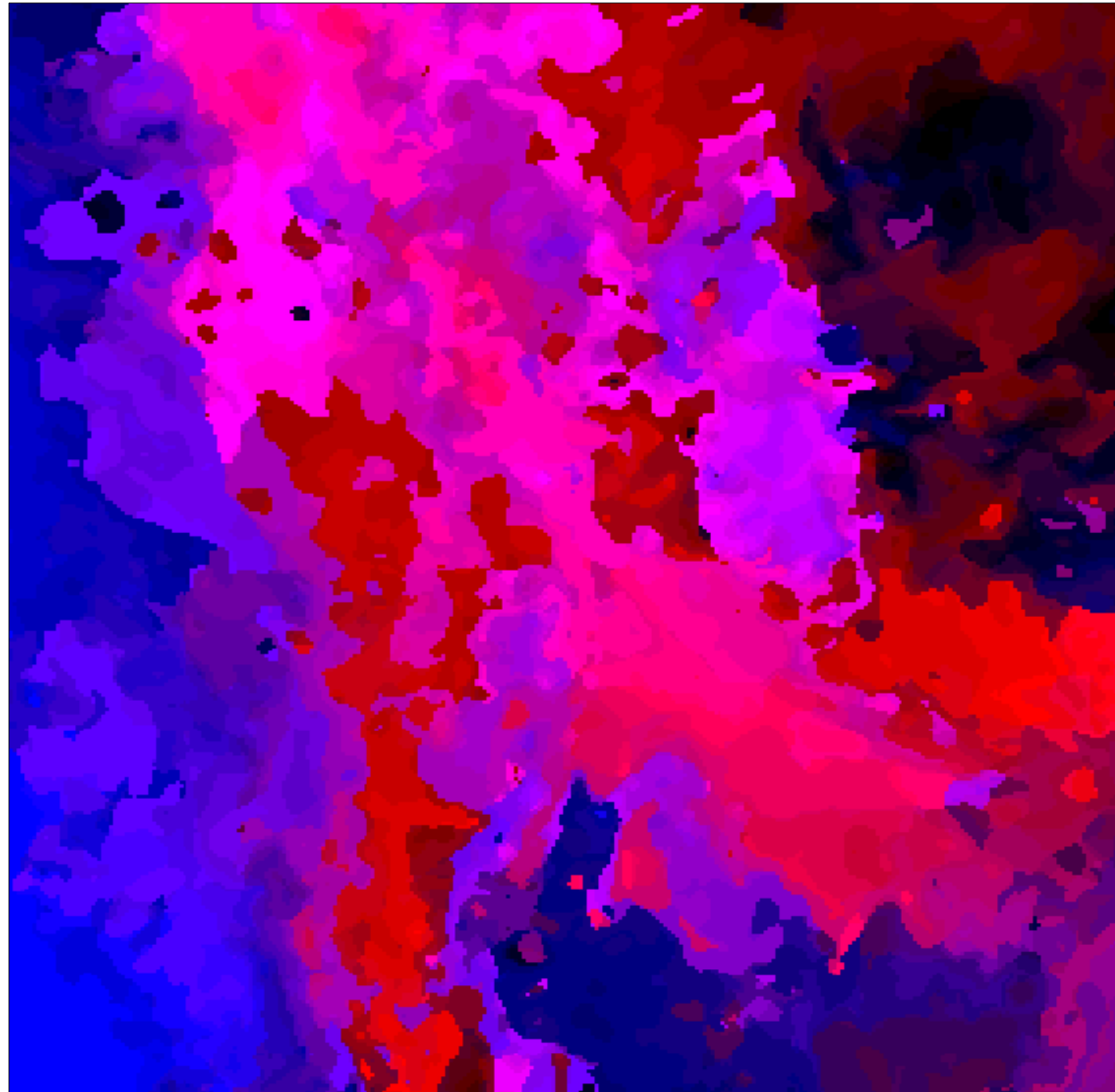
Playing with spectra

ESCAPE project ESO/CDS/HITS

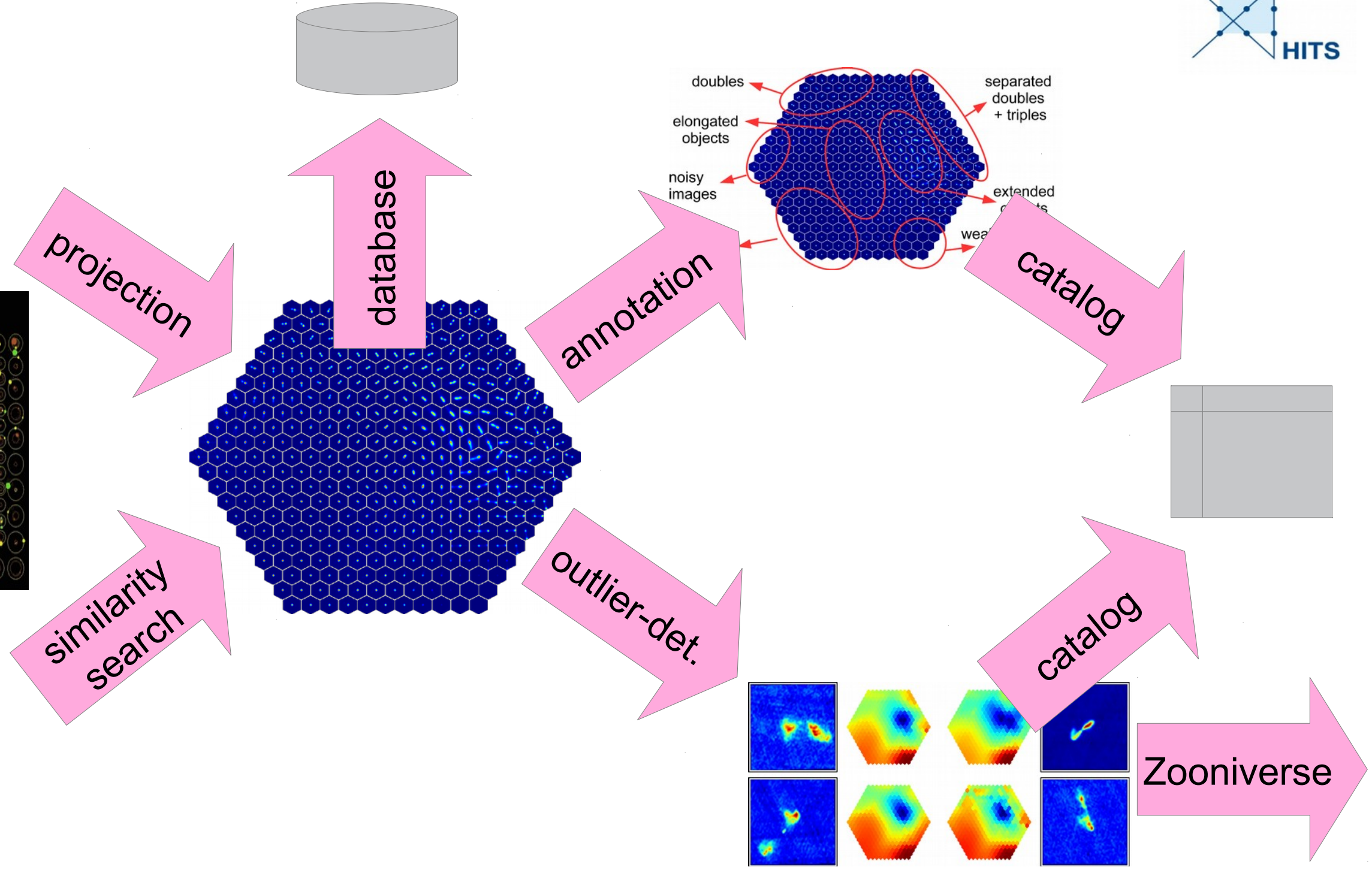
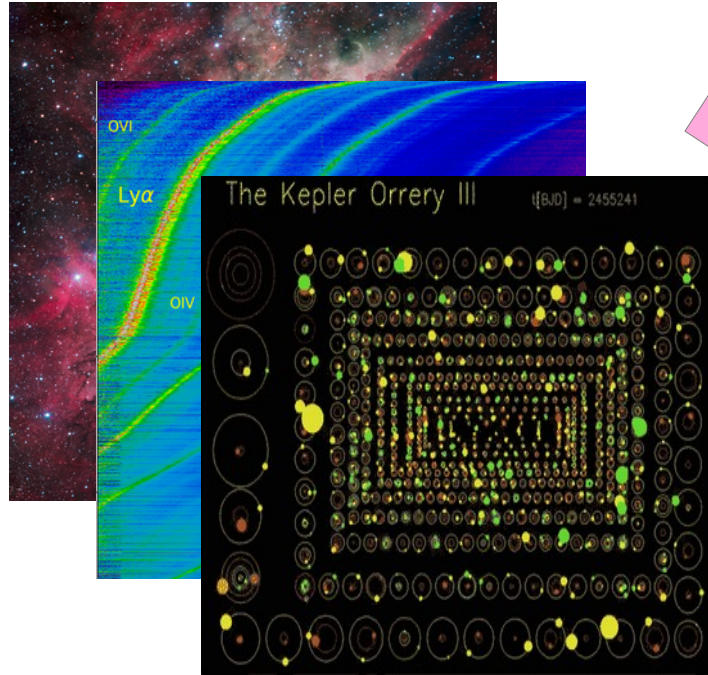
- exploring 300k spectra in realtime on a laptop



Data cubes / ppv-data



Schema



Lessons learned



don't **bias**
your system

use stupid, but fast
computers
for the boring tasks

do the
creative
interpretation of the data



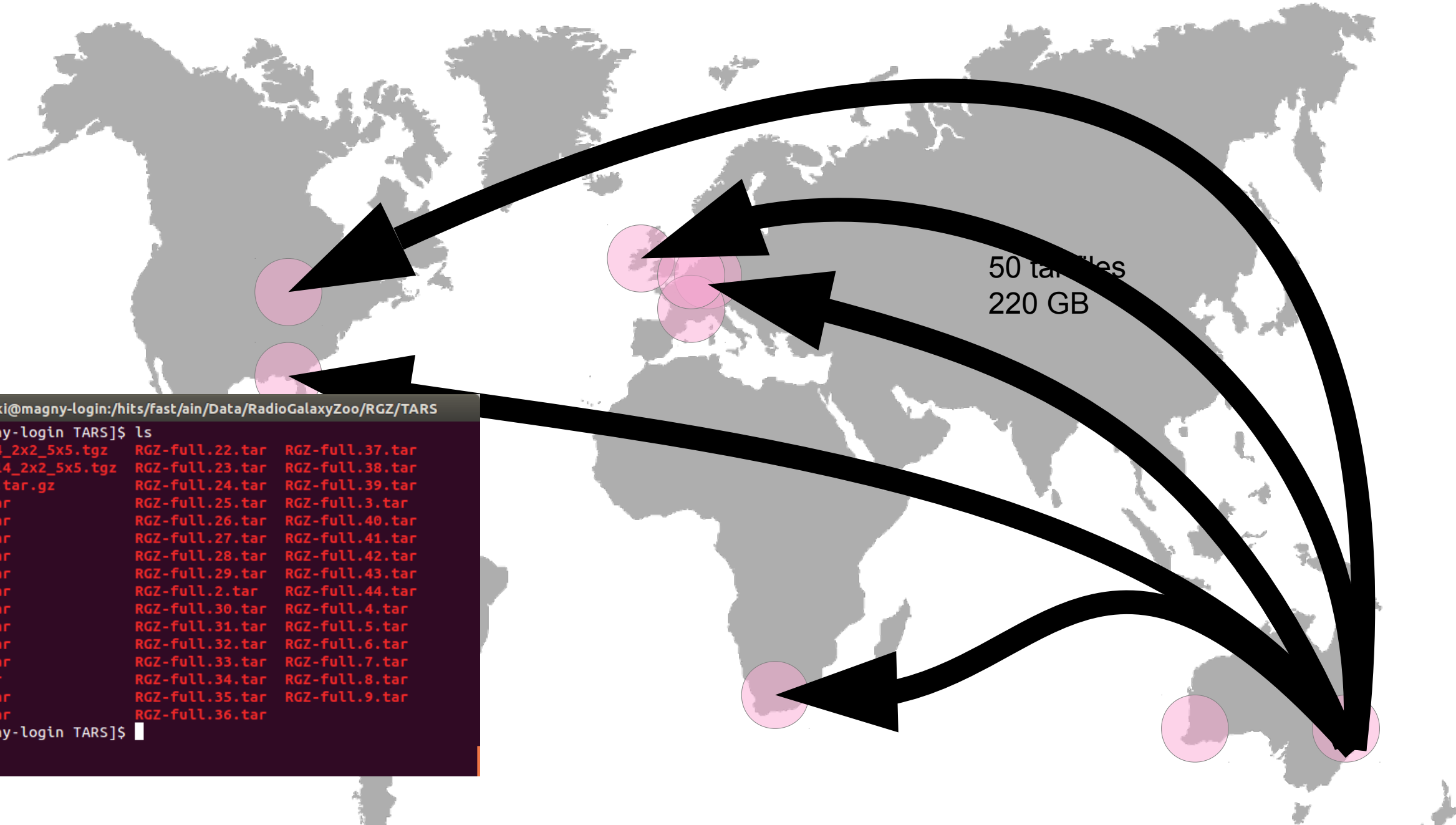


```
/**  
 * CUDA Kernel Device code for combined rotation and cropping of a list of images.  
 */  
template <unsigned int block_size>  
global void  
rotateAndCropTexture_kernel(float *rotatedImages, float *image, int neuron_size,  
int neuron_dim, int image_dim, float *cosAlpha, float *sinAlpha)  
{  
    int x2 = blockIdx.x * blockDim.x + threadIdx.x;  
    float cosAlpha_local = cosAlpha[blockIdx.z];  
    float sinAlpha_local = sinAlpha[blockIdx.z];  
    if ((inputData.verbose) cout << "\n Size of SOM = " << som.getSize() << endl;  
    float *d_som = cuda_alloc_float(som.getSize());  
    cuda_copyHostToDevice_float(d_som, som.getDataPointer(), som.getSize());  
    // Memory allocation  
    int rotatedImagesSize = inputData.numberOfChannels * inputData.numberOfRotationsAndFlip * inputData.neuron_size;  
    if (inputData.verbose) cout << "\n Size of rotated images = " << rotatedImagesSize * sizeof(float) << "\n bytes" << endl;  
    float *d_rotatedImages = cuda_alloc_float(rotatedImagesSize);  
    if (inputData.verbose) cout << "\n Size of euclidean distance matrix = " << inputData.som_size * sizeof(float) << "\n bytes" << endl;  
    int *d_euclideanMatrix = cuda_alloc_float(inputData.som_size * inputData.som_size * sizeof(int));  
    // Prepare trigonometric values  
    float *d_cosAlpha = NULL, *d_sinAlpha = NULL;  
    int *d_metricValues = NULL;  
    if (inputData.verbose) cout << "\n Size of best rotation metric values = " << inputData.som_size * sizeof(int) << "\n bytes" << endl;  
    rotateAndCropTexture_kernel(float *rotatedImages, float *image, int neuron_size,  
int neuron_dim, int image_dim, float *cosAlpha, float *sinAlpha)  
{  
    int x2 = blockIdx.x * blockDim.x + threadIdx.x;  
    int y2 = blockIdx.y * blockDim.y + threadIdx.y;  
    int x0margin = x0 - margin;  
    int y0margin = y0 - margin;  
    float cosAlpha_local = cosAlpha[blockIdx.z];  
    float sinAlpha_local = sinAlpha[blockIdx.z];  
    int x1 = (x2 - x0margin) * cosAlpha_local - (y2 - y0margin) * sinAlpha_local;  
    int y1 = (y2 - y0margin) * cosAlpha_local + (x2 - x0margin) * sinAlpha_local;  
    atomicAdd(pCurRot + x2 * neuron_dim + y2 * neuron_dim * neuron_dim, d_som[x1 * neuron_dim + y1 * neuron_dim]);  
}
```

Accessing and Analyzing Data

uhhhs, oohhs, don'ts and lessons learned

Starting a project in 2015



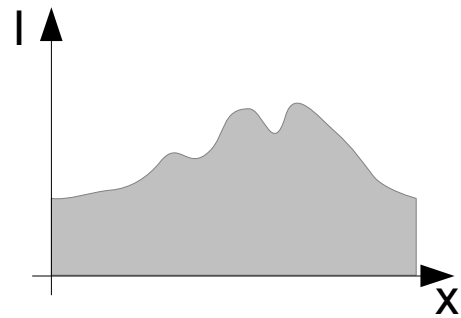
```
polsteki@magny-login:/hits/fast/ain/Data/RadioGalaxyZoo/RGZ/TARS
[polsteki@magny-login TARS]$ ls
cdfs_11JAN2014_2x2_5x5.tgz  RGZ-full.22.tar  RGZ-full.37.tar
elais_11JAN2014_2x2_5x5.tgz  RGZ-full.23.tar  RGZ-full.38.tar
Imaging-1.1.7.tar.gz       RGZ-full.24.tar  RGZ-full.39.tar
RGZ-full.10.tar            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



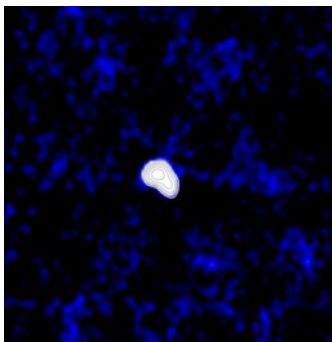
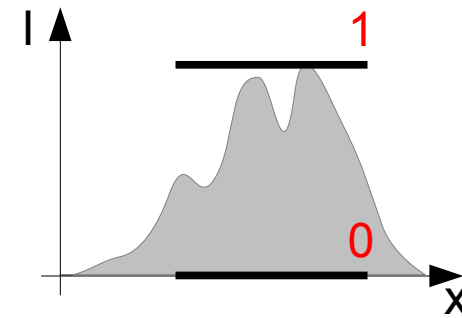
extract
matrix from fits

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



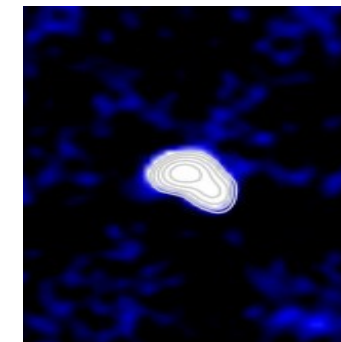
normalize

flux relative to the maximum



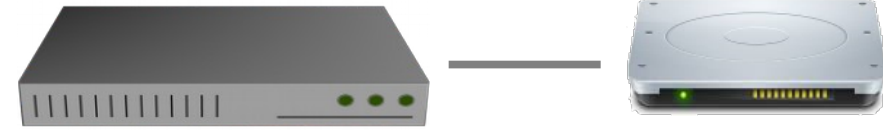
cutout

interesting region



Speeding up preprocessing

single core python = **48**
hours



with  **hadoop**

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



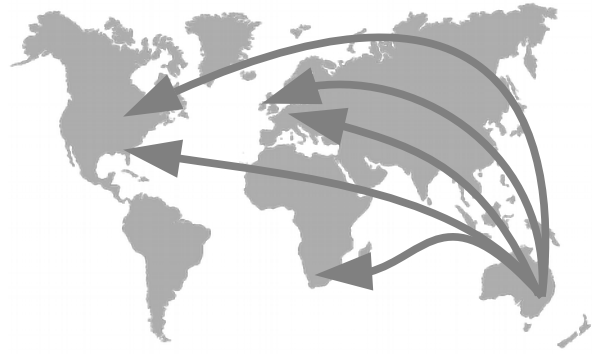
file **access** was still the bottleneck!

New images extracted



```
polsteki@magny-login:/hits/fast/ain/Data/RadioGalaxyZoo/RGZ/TARS
[polsteki@magny-login TARS]$ ls
cdf5_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]$
```

The downsides of this approach



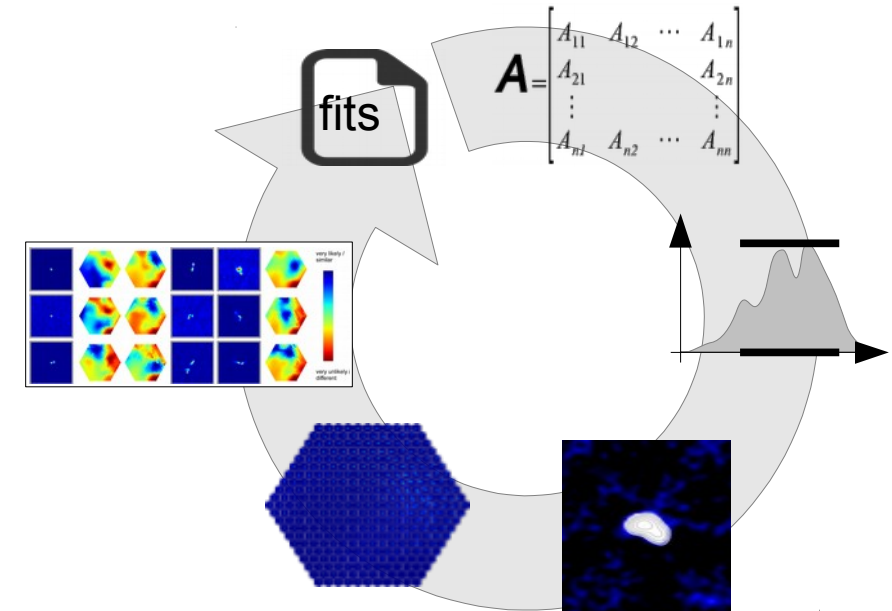
a lot of local copies

no orchestration of work-flow



bad exchange of intermediate results

very exclusive concerning hardware requirements

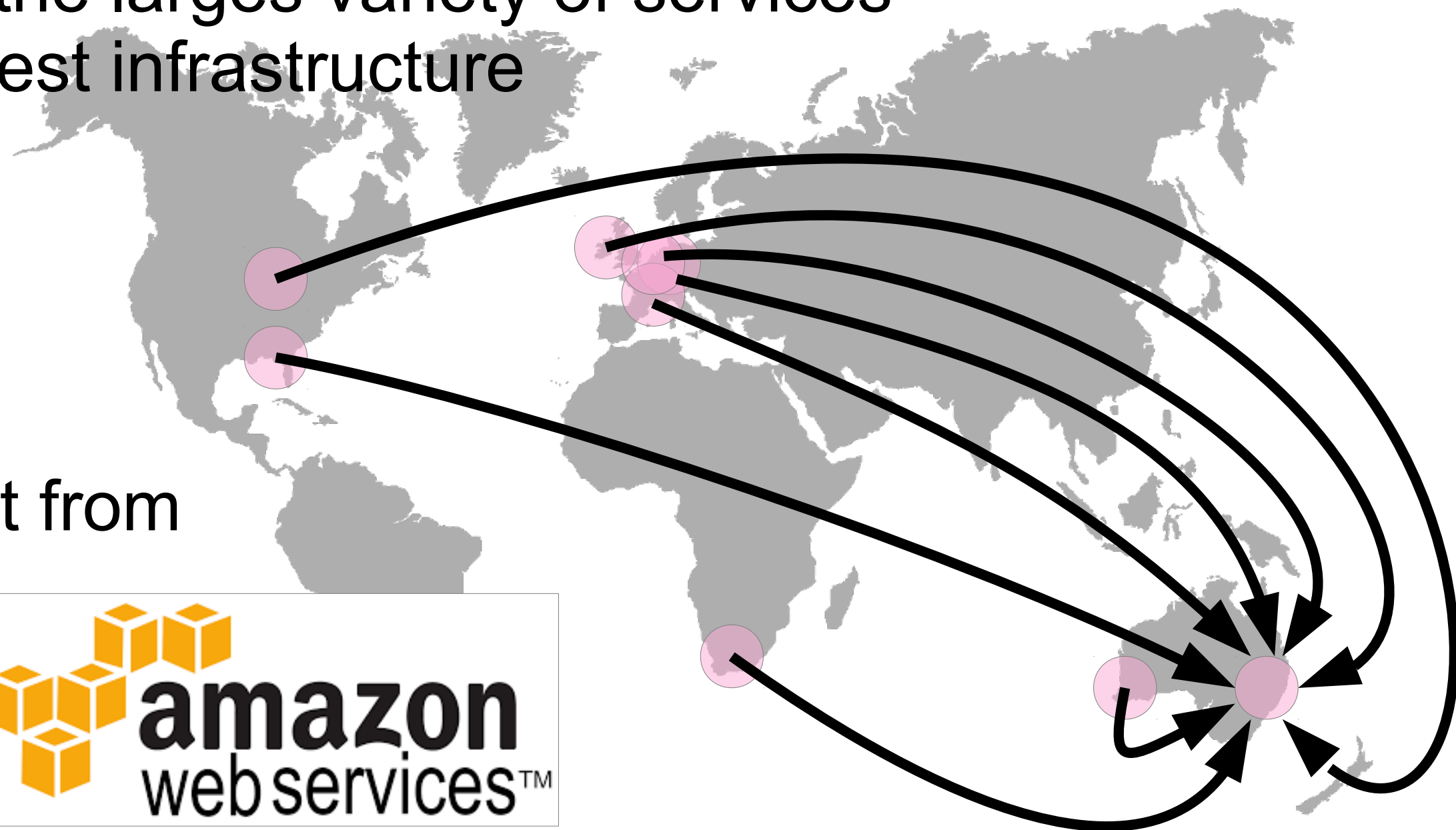


NVIDIA Tesla K40

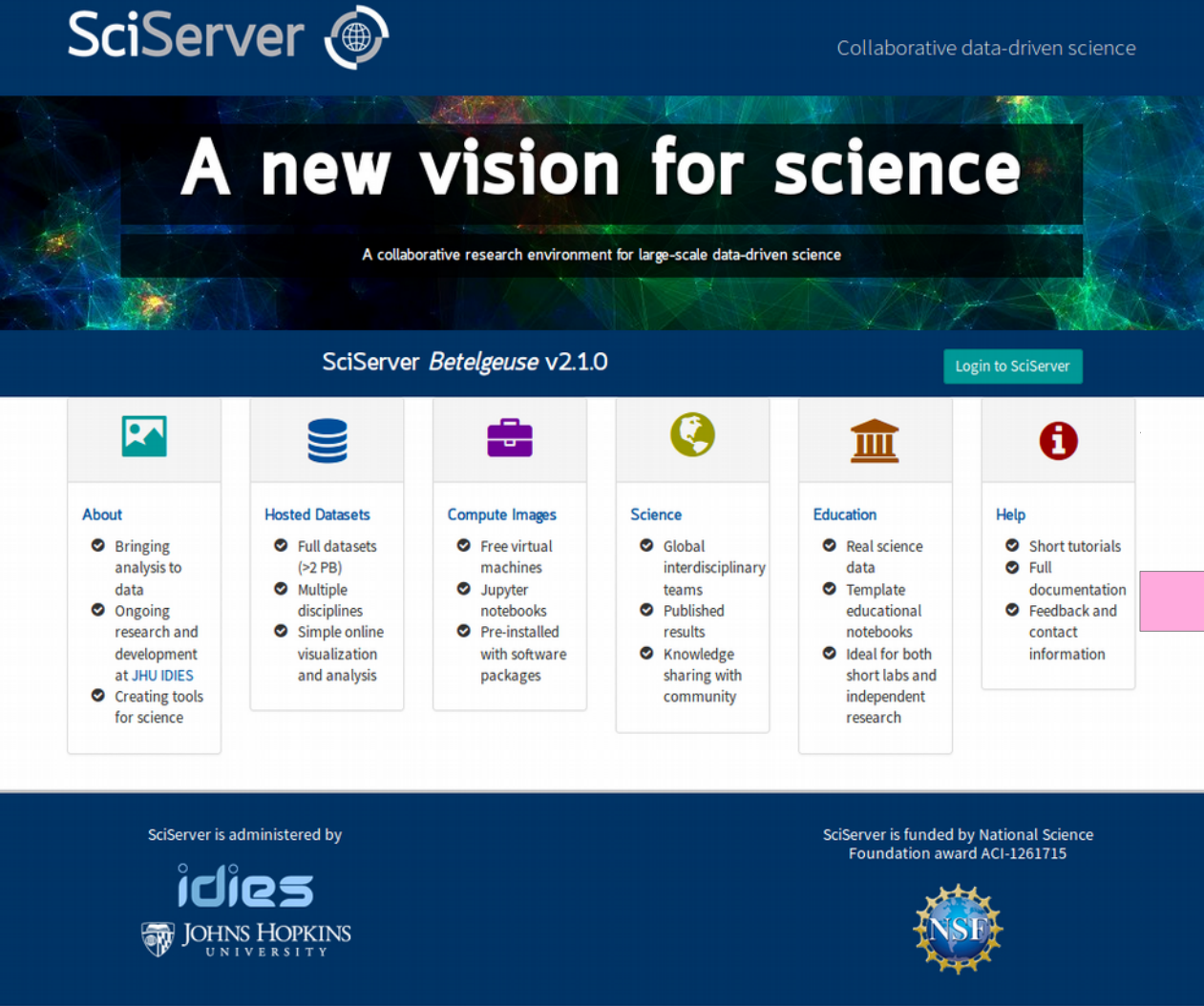
2015 / used Amazon Web Services

provided the largest variety of services
and the best infrastructure

SKA grant from



Today / bringing code to the data




SciServer Collaborative data-driven science


A new vision for science

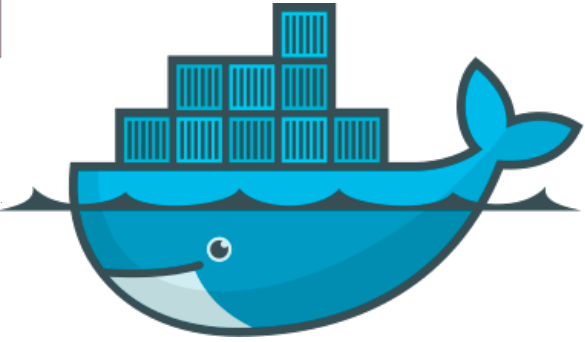
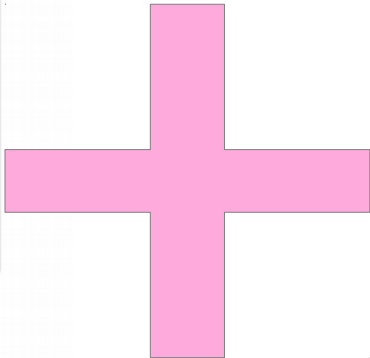
A collaborative research environment for large-scale data-driven science

SciServer *Betelgeuse* v2.1.0 [Login to SciServer](#)

- About**
 - Bringing analysis to data
 - Ongoing research and development at JHU IDIES
 - Creating tools for science
- Hosted Datasets**
 - Full datasets (>2 PB)
 - Multiple disciplines
 - Simple online visualization and analysis
- Compute Images**
 - Free virtual machines
 - Jupyter notebooks
 - Pre-installed with software packages
- Science**
 - Global interdisciplinary teams
 - Published results
 - Knowledge sharing with community
- Education**
 - Real science data
 - Template educational notebooks
 - Ideal for both short labs and independent research
- Help**
 - Short tutorials
 - Full documentation
 - Feedback and contact information

SciServer is administered by  IDIES JOHNS HOPKINS UNIVERSITY

SciServer is funded by National Science Foundation award ACI-1261715 



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

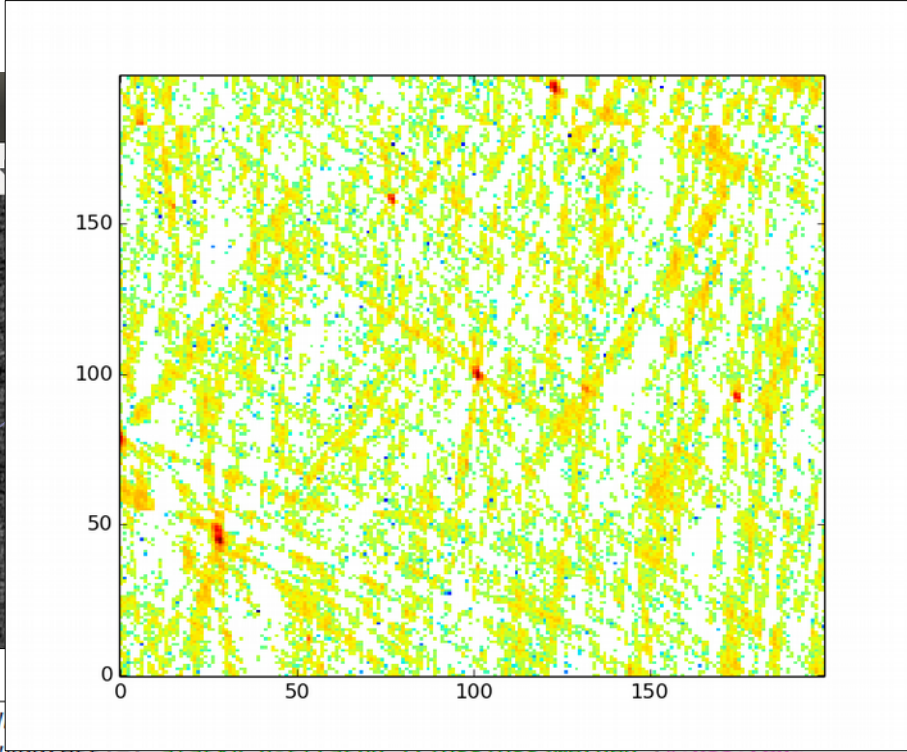
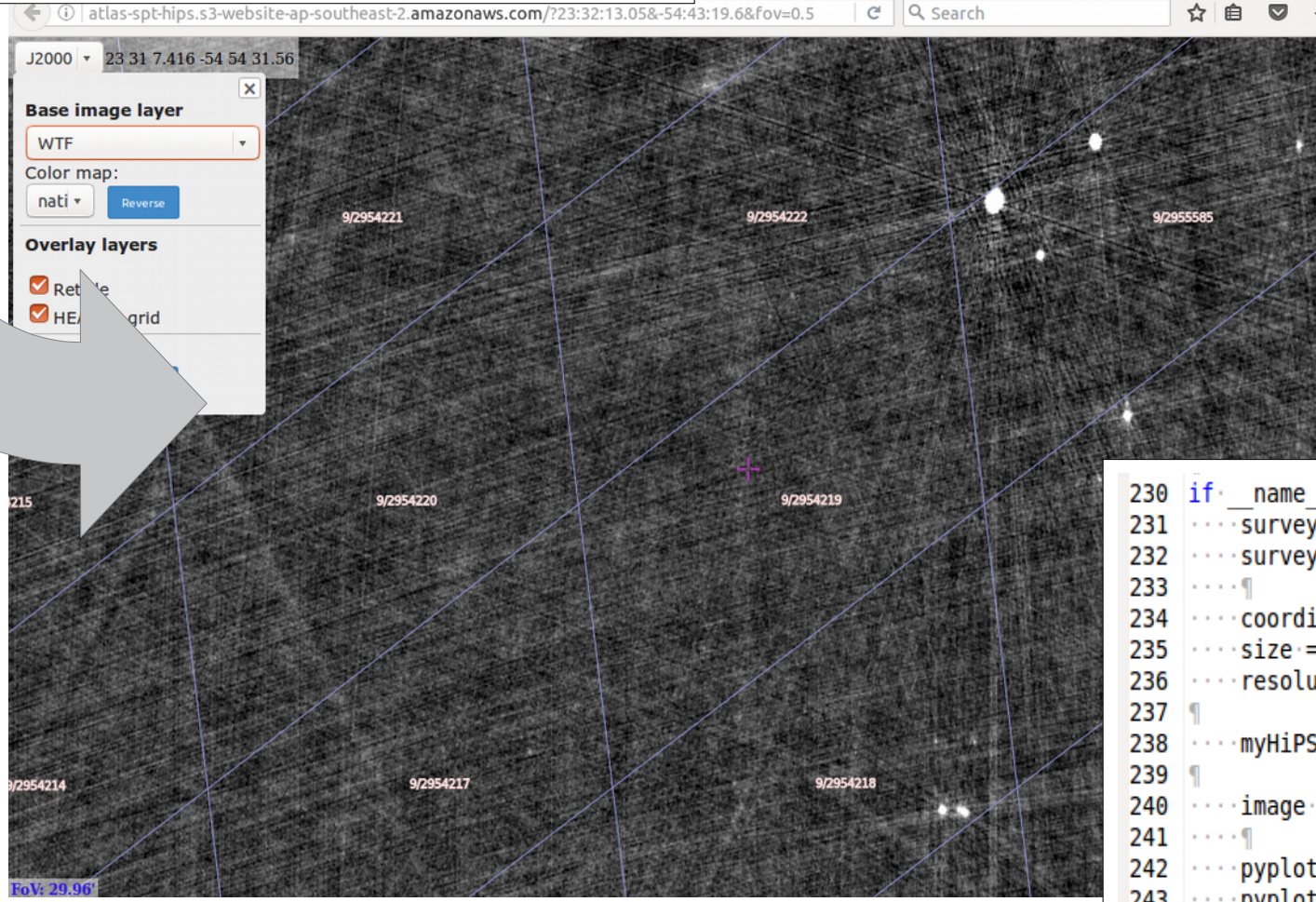
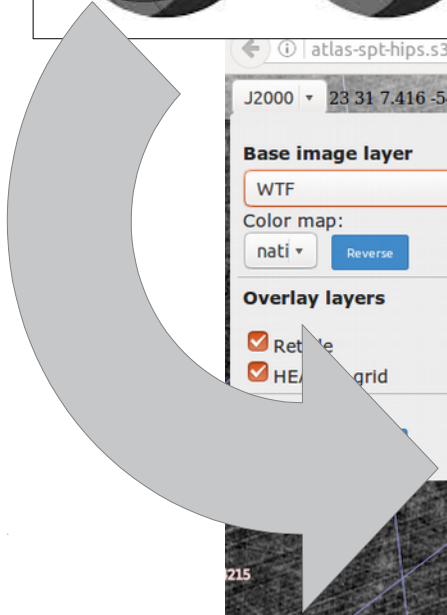
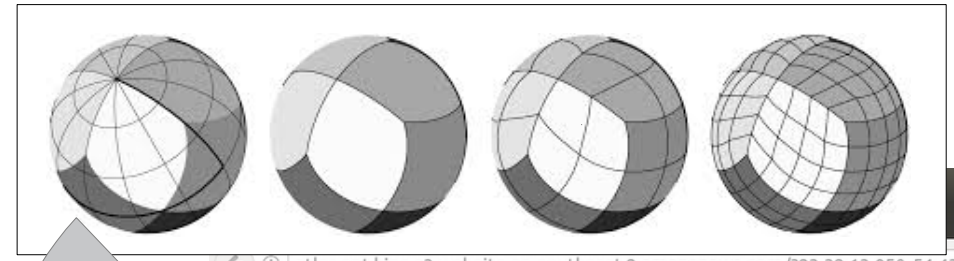
how to train a model without downloading the data

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

how to deal with distributed data-sets

- deal with radio and IR-data

Healpix / HiPS / IVOA



```

230 if __name__ == '__main__':
231     survey = Survey(surveyAddress)
232     surveyAddress = 'atasky.u-strasbg.fr/DSS/DSS2merged'
233     #
234     coordinate = [350.86, -55.225]
235     size = [200, 200]
236     resolution = 0.002
237     #
238     myHiPSfs = HiPSfs(surveyAddress) # create access
239     #
240     image = myHiPSfs.extractCoordinate(coordinate, size, resolution, nested=True) # extract data array
241     #
242     pyplot.figure()
243     pyplot.imshow(image, aspect='auto', interpolation="nearest")
244     pyplot.gca().invert_yaxis()
245     pyplot.show()

```

Conclusion



machine learning → accessing data

but, we have to make the data accessible to ML