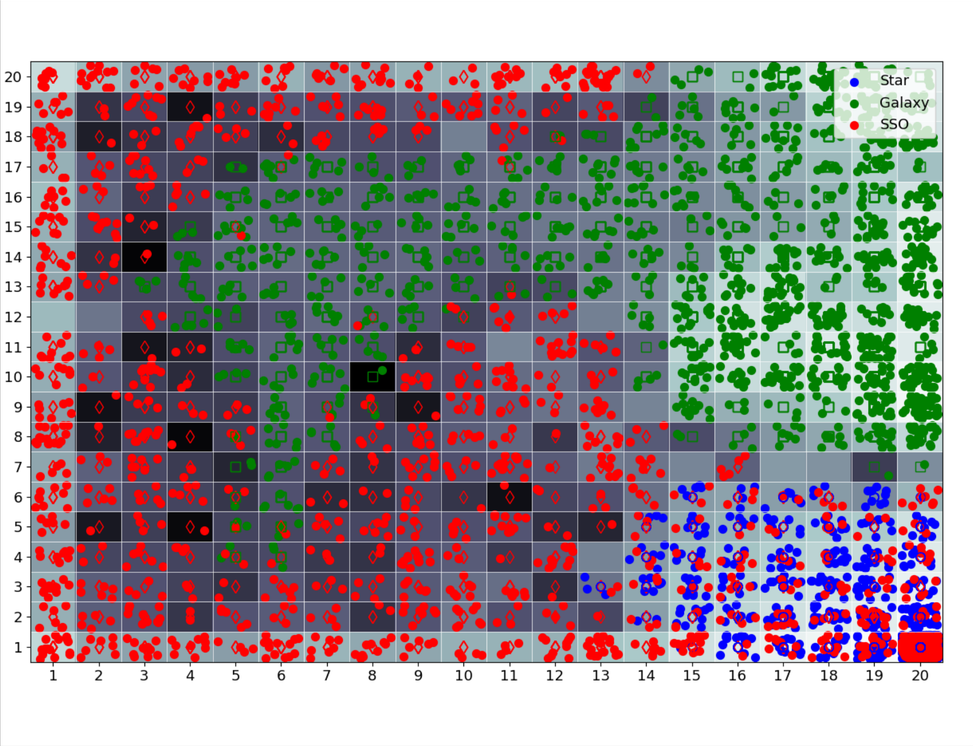
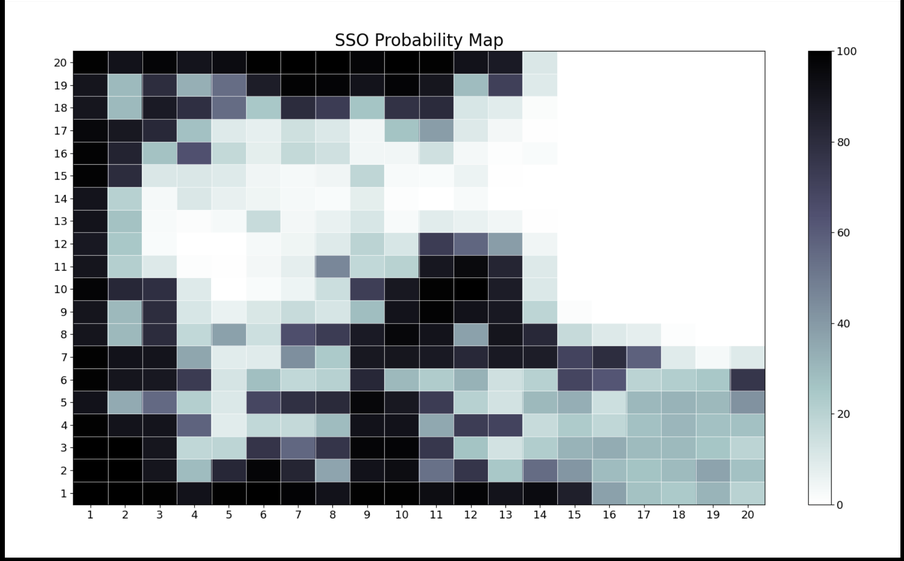



# Early-stopping SOMs as a tool to classify SSOs in space surveys

S. Sacquegna, A. A. Nucita, A. Franco, L. Conversi, A. Verdier, M. Pöntinen, B. Altieri, B. Carry, F. De Paolis, F. Strafella, G. Ingrosso, V. Orofino, M. Maiorano, V. Kansal, R. Vavrek, M. Miluzio, M. Granvik, V. Testa

International Conference on Machine Learning for Astrophysics
Catania, 10th July 2024

# SOM: How Does It Work

- $Q = N\,M$ neurons organized in a rectangular lattice, with coordinates $(i, j)$ with $(i = 0, \dots, N - 1, j = 0, \dots, M - 1)$
- Each neuron associated to a reference vector $r_k$ with $K$ values and is exposed to an input vector $w_k$ coming from a training dataset with $L$ inputs
- How does a neuron win? By minimizing the Euclidean distance:

$$D^l_{\text{min}} = \min_{i,j}\left(\sqrt{\sum_{k=0}^{K-1} m^l_k\,(r^{i,j}_k - w^l_k)^2}\right)$$

- After a full epoch, iterative process by updating neuron components:

$$r'^{i,j}_k = r^{i,j}_k + \alpha\left(\frac{t}{N_e}\right) H\left(\frac{t}{N_e}, d_{win} - d\right)\left(w^l_k - r^{i,j}_k\right)$$

- The parameters $\alpha(t)$ and $\sigma(t)$ can either be linearly or exponentially decreasing with the passing epochs.

# Early stopping

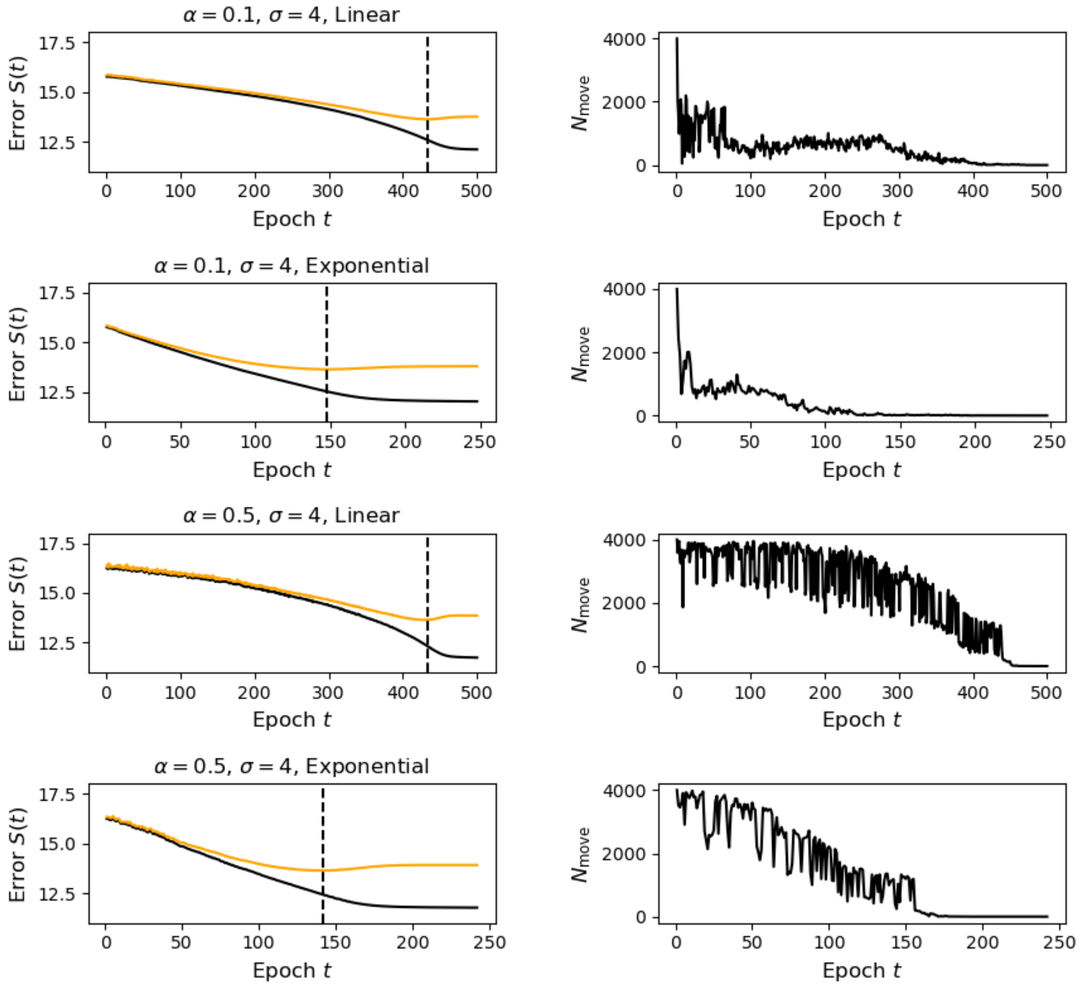
- Using $N_e$, process can be long and time (and memory!) consuming, and eventually produce over-fitting
- Solution: check performance of the SOM on an independent, randomly chosen validation dataset
- Metric:

$$S(t) = \frac{\sum_{l=0}^{L-1} D_{\min}^{l}}{L}$$

- Expected behavior: $S(t)$ will continuously decrease, $S_{val}(t)$ will reach a minimum and trend upward (overfitting!)
- By stopping the learning process before the overfitting stage, the SOM will maintain its capability to classify new data.

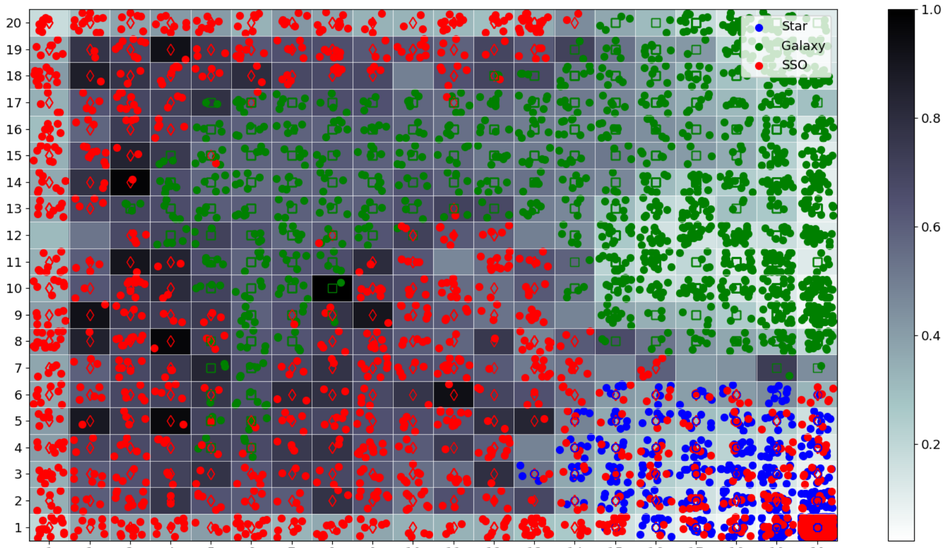
# Training the SOM

- 4000 random 101x101 images → 20x20 neurons
- Comparing errors and numbers of movements for different models and parameters:
- Results: exponential, $\alpha = 0.5$ , $\sigma = 4$

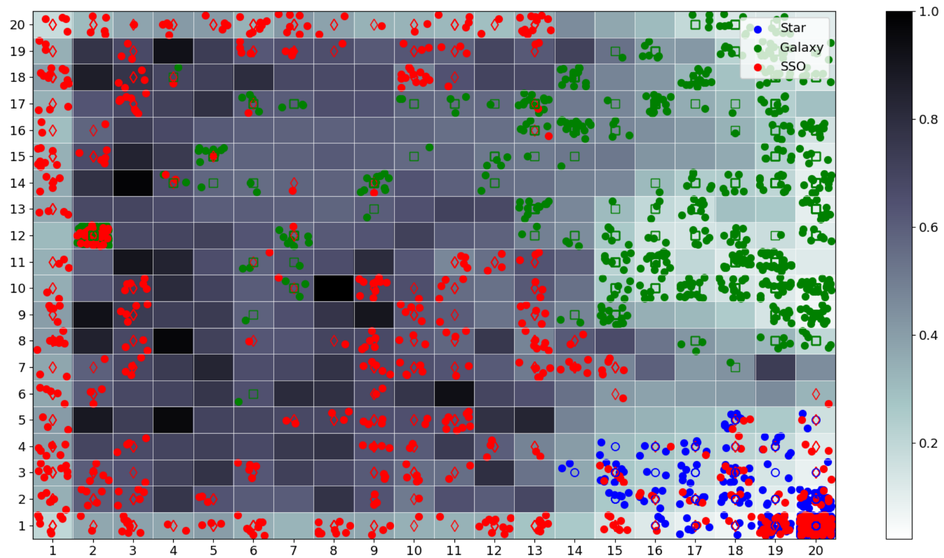


# Training the SOM

- U-matrix generated from SOM training:

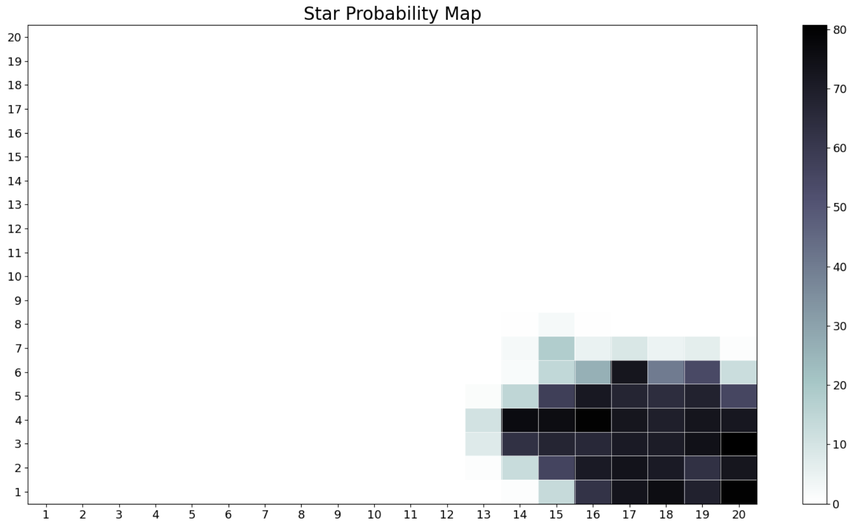


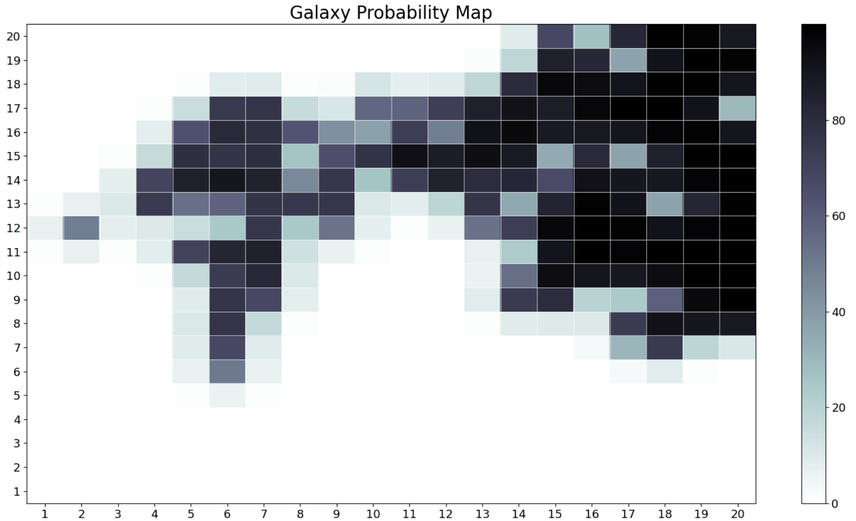

- U-matrix with 2000 images from test dataset:

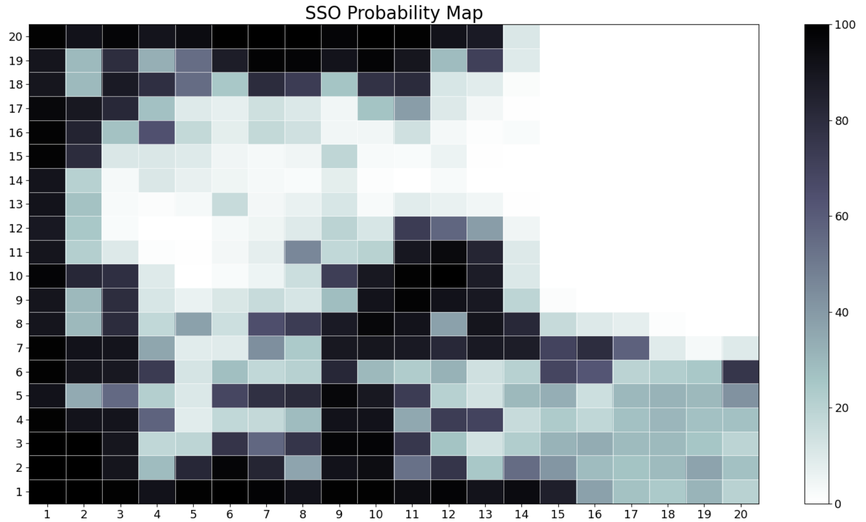


# Results

- Quality assessment with $10^4$ images, bins 0-3’’/h, 3-6’’/h, 6-10’’/h
- Purity: 11%, 100%, 100%
- Completeness: 51%, 48%, 60%



# Conclusions

- Early-stopping is a powerful technique to improve SOMs
- With proper parameters, stopping time is more than halved
- Relatively high accuracy above 1’’/h (slow moving SSOs are undistinguishable from stars!)
- Difference between simulated and authentic images: more training is needed for real survey data