Summary of ML activities in Capodimonte

S. Cavuoti
INAF - Astronomical Observatory of Capodimonte

DaME (Data Mining & Exploration) started as a collaboration between:

- Astronomical Observatory of Capodimonte
- University of Naples, Federico II
- Caltech

created in **2007** order to deep dive into the field of Astroinformatics by applying ML techniques to Astrophysics

A Band of Fools

Stefano Cavuoti, Giuseppe Riccio, Giuseppe Angora, Ylenia Maruccia, Giuseppe Sarracino, Natale De Bonis, Simone Vaccaro **- INAF OACN**

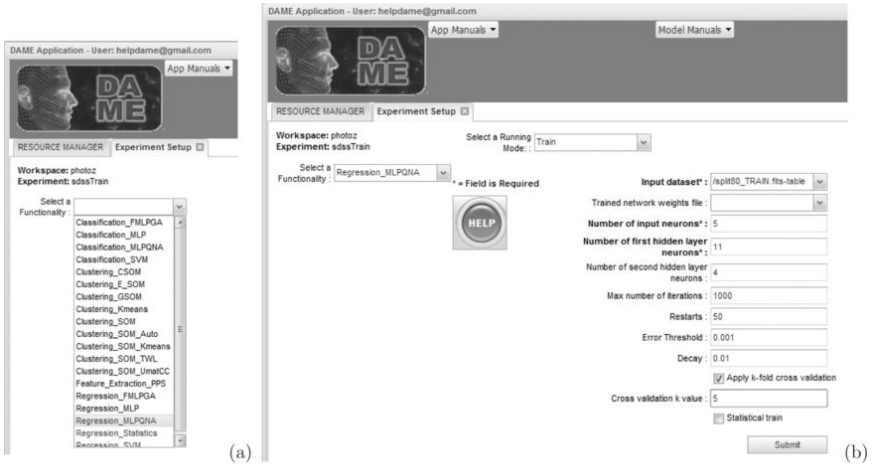
Giuseppe Longo, Massimo Brescia, Demetra De Cicco, Maurizio Paolillo **- Unina, Federico II**

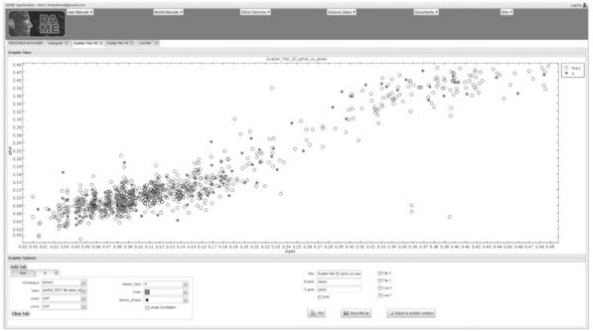
Past members (Fixed Term, Fellowship, PhD, Master Students...):
Giovanni Albano, Valeria Amaro, Marianna Annunziatella, Massimo Benedetto, Sabrina Checola, Raffaele D'Abrusco, Maurizio D'Addona, Pierluigi D'Andrea, Giovanni D'Angelo, Michele Delli Veneri, Virgilio De Stefano, Luna Di Colandrea, Alessandro Di Guido, Antonio D'Isanto, Lars Doorenbos, Francesco Esposito, Pamela Esposito, Michelangelo Fiore, Mauro Garofalo, Domenico Guarino, Marisa Guglielmo, Omar Laurino, Francesco Manna, Alfonso Nocella, Luca Pellecchia, Carlo Enrico Petrillo, Oleksandra Razim, Sandro Riccardi, Bojan Skordovski, Andrea Solla, Olena Torbaniuk, Gianluca Tutino, Civita Vellucci, Giovanni Vebber **- Around the World**

**We few, we happy few, we band of brothers;
For he to-day that sheds his blood with me shall be my brother**

**Henry V
William Shakespeare**

DAMEWARE





(c)

Publications of the Astronomical Society of the Pacific

INSTRUMENTATION • FREE ARTICLE

DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining

Massimo Brescia, Stefano Cavuoti, Giuseppe Longo, Alfonso Nocella, Mauro Garofalo, Francesco Manna, Francesco Esposito, Giovanni Albano, Marisa Guglielmo, Giovanni D'Angelo, Alessandro Di Guido, S. George Djorgovski, Ciro Donalek, Ashish A. Mahabal, Matthew J. Graham, Michelangelo Fiore, and Raffaele D'Abrusco ▲ Hide full author list

Publications of the Astronomical Society of the Pacific, Volume 126, Number 942

Citation Massimo Brescia *et al* 2014 *PASP* **126** 783

DOI 10.1086/677725

TABLE 1

DATA MINING MODELS AND FUNCTIONALITIES AVAILABLE IN THE DAMEWARE FRAMEWORK

Model	Name	Category	Functionality
MLPBP	Multi Layer Perceptron with Back Propagation	Supervised	Classification, regression
FMLPGA	Fast MLP trained by Genetic Algorithm	Supervised	Classification, regression
MLPQNA	MLP with Quasi Newton Approximation	Supervised	Classification, regression
MLPLEMON	MLP with Levenberg-Marquardt Optimization Network	Supervised	Classification, regression
SVM	Support Vector Machine	Supervised	Classification, regression
ESOM	Evolving Self Organizing Maps	Unsupervised	Clustering
K-Means		Unsupervised	Clustering
SOFM	Self Organizing Feature Maps	Unsupervised	Clustering
SOM	Self Organizing Maps	Unsupervised	Clustering
PPS	Probabilistic Principal Surfaces	Unsupervised	Feature Extraction

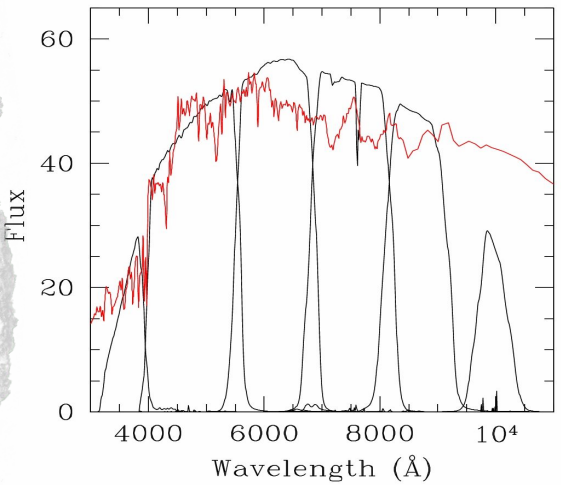
Photo-z are crucial to:

- Studies of large scale structure
- Weak lensing (hence dark matter and dark energy distribution)
- Tests of cosmological models
- Galaxy evolution and mass assembly
- Classification of galaxies
- etc....

Surveys:

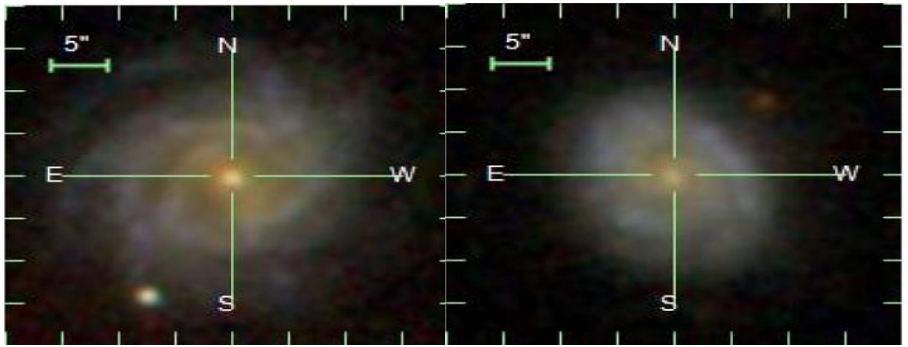
- SDSS (+ UKIDSS/ Galex/ Wise)
- KiDS (official)
- VST - VOICE
- EUCLID challenge
- LSST challenge
- eROSITA pre-launch
- EMU challenge
- Possible forgets

How can we derive photo-z?



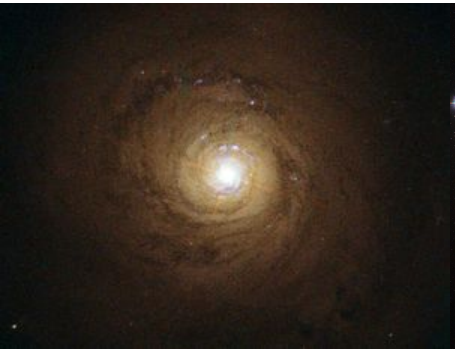
Problems & Data: AGNs

1) The same object seen from different orientation appear in different ways (this depends of what your line of sight intercept)
2) The AGNs are changing in time (with different time scales)



This is an AGN

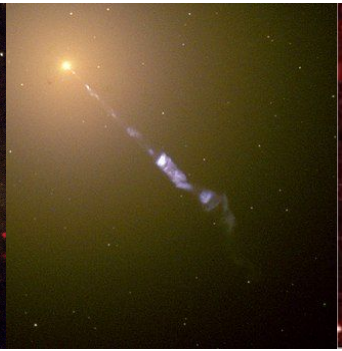
This is **NOT** an AGN



Seyfert Type 1

Seyfert Type 1.5

Seyfert Type 2

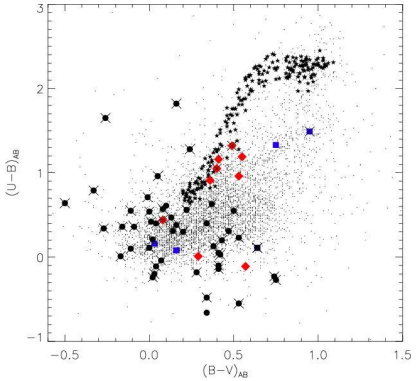


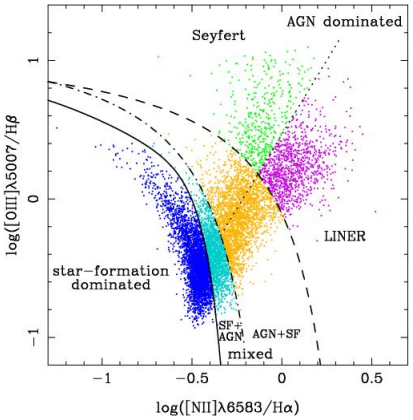
Blazar

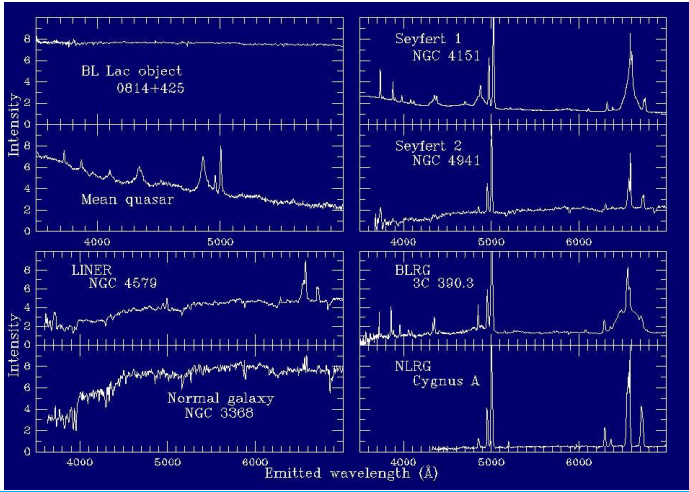
Quasar

How to Detect AGNs

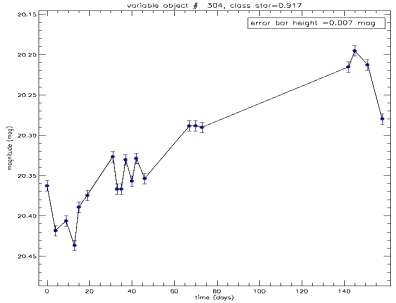
1) Through Photometry (**UV/Optical/IR**) there are few examples in literature that are able to detect with a reasonable efficiency one kind of AGNs: Quasars but practically no way to identify anything else
2) Through **X-ray** emission, very efficient BUT:
   - not all the AGNs have X-ray emission
   - we have a little X-ray coverage of the sky
3) Through **Radio** emission, quite efficient but the percentage of AGNs radio emitters is not that high
4) Through **Spectral** Observation, this is efficient but really time consuming, for example in one of the most famous catalogues of Astronomical Objects, the SDSS we have 1 billion of photometric observation and about 3 million of spectra.
   - we need reasonable candidates to be spectroscopically observed
5) Through **Variability**, this require to observe the same object again and again for years with the same instruments, probably we need decades of observations to raise the efficiency

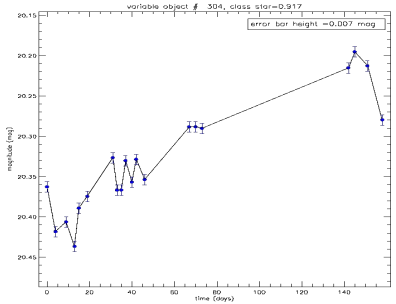


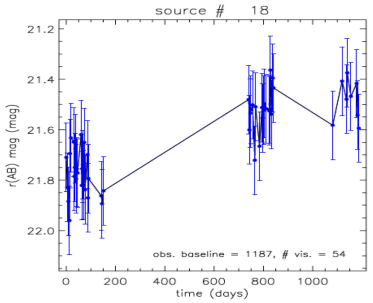




Extract Features From LC







Astronomical Features
Morphology, Colors, Excess Variance…

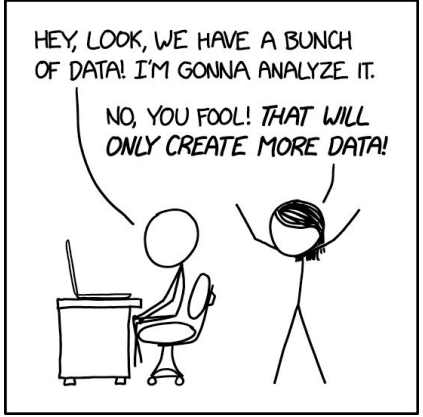
Measures of Variability
Standard Deviation, Amplitude, NMAD…

Measures of Position
Percentiles…

Measures of Shape
Skew, Kurtosis…

Trend Features
Maximum Slope, Pair Slope Trend …

…



We Convert each LC in a series of statistical indicators that will be our feature space

Selected Features (De Cicco+21)

"classic" variability features

variability features from the Python FATS (Feature Analysis for Time Series) library

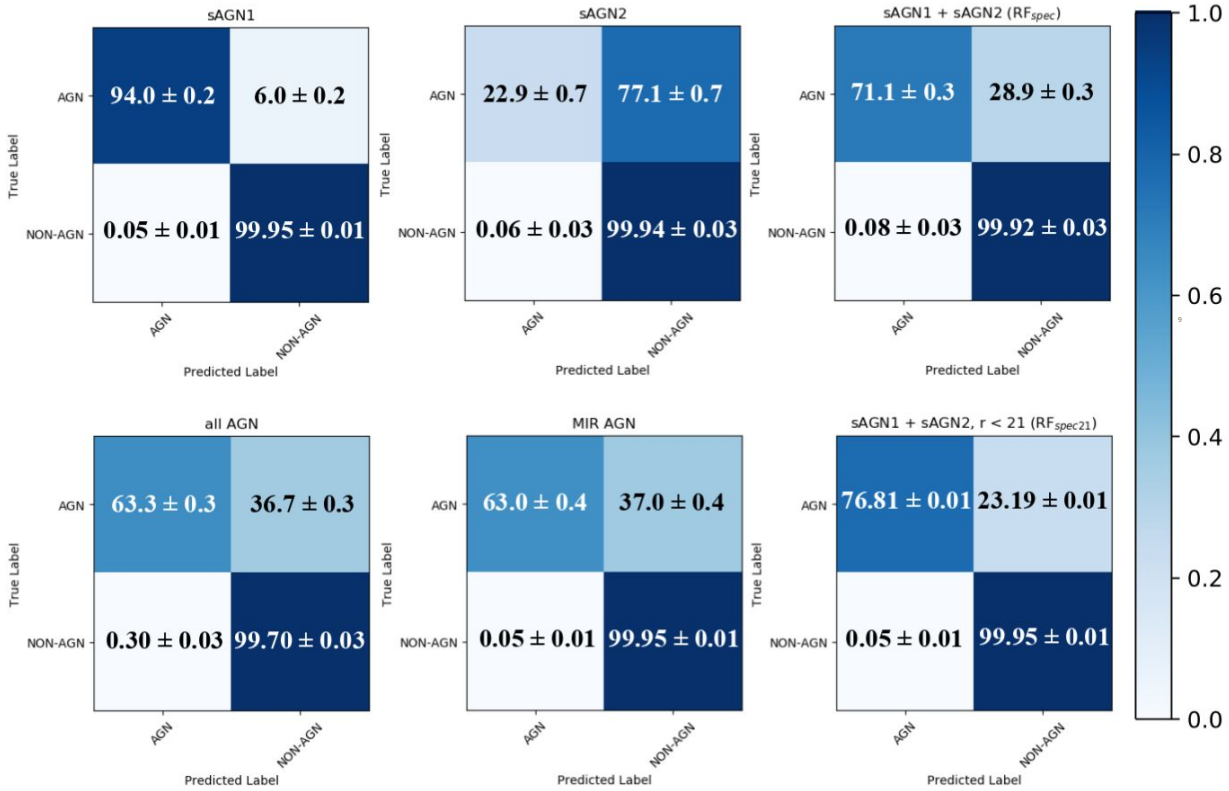
morphology feature

color features

Feature	Description	Reference
A_{SF}	rms magnitude difference of the SF, computed over a 1 yr timescale	Schmidt et al. (2010)
γ_{SF}	Logarithmic gradient of the mean change in magnitude	Schmidt et al. (2010)
GP_DRW_τ	Relaxation time τ (i.e., time necessary for the time series to become uncorrelated), from a DRW model for the light curve	Graham et al. (2017)
GP_DRW_σ	Variability of the time series at short timescales ($t \ll \tau$), from a DRW model for the light curve	Graham et al. (2017)
ExcessVar	Measure of the intrinsic variability amplitude	Allevato et al. (2013)
P_{var}	Probability that the source is intrinsically variable	McLaughlin et al. (1996)
IAR_{ϕ}	Level of autocorrelation using a discrete-time representation of a DRW model	Eyheramendy et al. (2018)
Amplitude	Half of the difference between the median of the maximum 5% and of the minimum 5% magnitudes	Richards et al. (2011)
AndersonDarling	Test of whether a sample of data comes from a population with a specific distribution	Nun et al. (2015)
Autocor_length	Lag value where the autocorrelation function becomes smaller than η^e	Kim et al. (2011)
Beyond1Std	Percentage of points with photometric mag that lie beyond 1σ from the mean	Richards et al. (2011)
η^e	Ratio of the mean of the squares of successive mag differences to the variance of the light curve	Kim et al. (2014)
Gskew	Median-based measure of the skew	–
LinearTrend	Slope of a linear fit to the light curve	Richards et al. (2011)
MaxSlope	Maximum absolute magnitude slope between two consecutive observations	Richards et al. (2011)
Meanvariance	Ratio of the standard deviation to the mean magnitude	Nun et al. (2015)
MedianAbsDev	Median discrepancy of the data from the median data	Richards et al. (2011)
MedianBRP	Fraction of photometric points within amplitude/10 of the median mag	Richards et al. (2011)
MHAOV Period	Period obtained using the P4J Python package (https://github.com/phuijse/P4J)	Huijse et al. (2018)
PairSlopeTrend	Fraction of increasing first differences minus the fraction of decreasing first differences over the last 30 time-sorted mag measures	Richards et al. (2011)
PercentAmplitude	Largest percentage difference between either max or min mag and median mag	Richards et al. (2011)
Q31	Difference between the third and the first quartile of the light curve	Kim et al. (2014)
Period_fit	False-alarm probability of the largest periodogram value obtained with LS	Kim et al. (2011)
Ψ_{cs}	Range of a cumulative sum applied to the phase-folded light curve	Kim et al. (2011)
Ψ_{η}	η^e index calculated from the folded light curve	Kim et al. (2014)
R_{cs}	Range of a cumulative sum	Kim et al. (2011)
Skew	Skewness measure	Richards et al. (2011)
Std	Standard deviation of the light curve	Nun et al. (2015)
StetsonK	Robust kurtosis measure	Kim et al. (2011)
class_star	HST stellarity index	Koekemoer et al. (2007), Scoville et al. (2007b)
$u - B$	CFHT u magnitude – Subaru B magnitude	Laigle et al. (2016)
$B - r$	Subaru Suprime-Cam B mag – Subaru Suprime-Cam r+ mag	Laigle et al. (2016)
$r - i$	Subaru Suprime-Cam r+ mag – Subaru Suprime-Cam i+ mag	Laigle et al. (2016)
$i - z$	Subaru Suprime-Cam i+ mag – Subaru Suprime-Cam z++ mag	Laigle et al. (2016)
$z - y$	Subaru Suprime-Cam z++ mag – Subaru Hyper-Suprime-Cam y mag	Laigle et al. (2016)
ch21	*Spitzer* 4.5 μm (*channel2*) mag – 3.6 μm (*channel1*) mag	Laigle et al. (2016)

Looking for the Optimal AGN LS:

Confusion Matrices



AGN & SOM

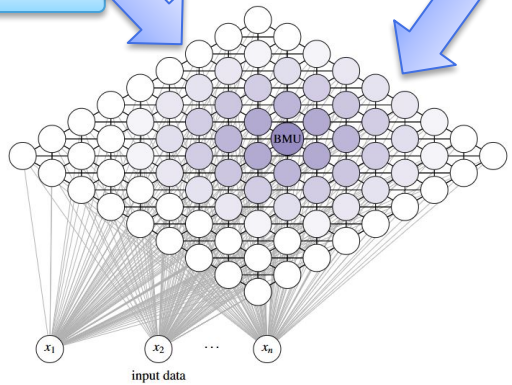
Features:

- Variability
- Morphology
- Colors

Label Set:

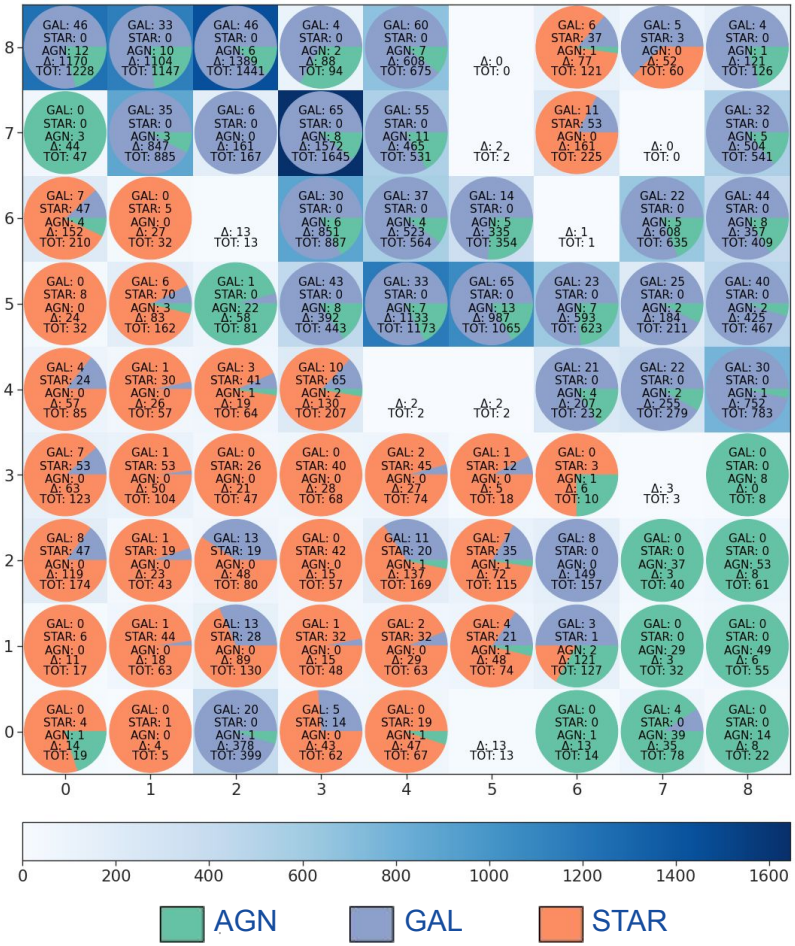
AGN, GAL, STAR

Unlabeled Set



Finding Similar Objects

Activation Map with the distribution of the labels



AGN & SOM

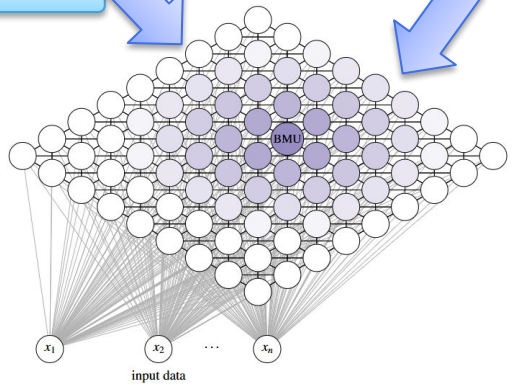
Features:

- Variability
- Morphology
- Colors

Label Set:

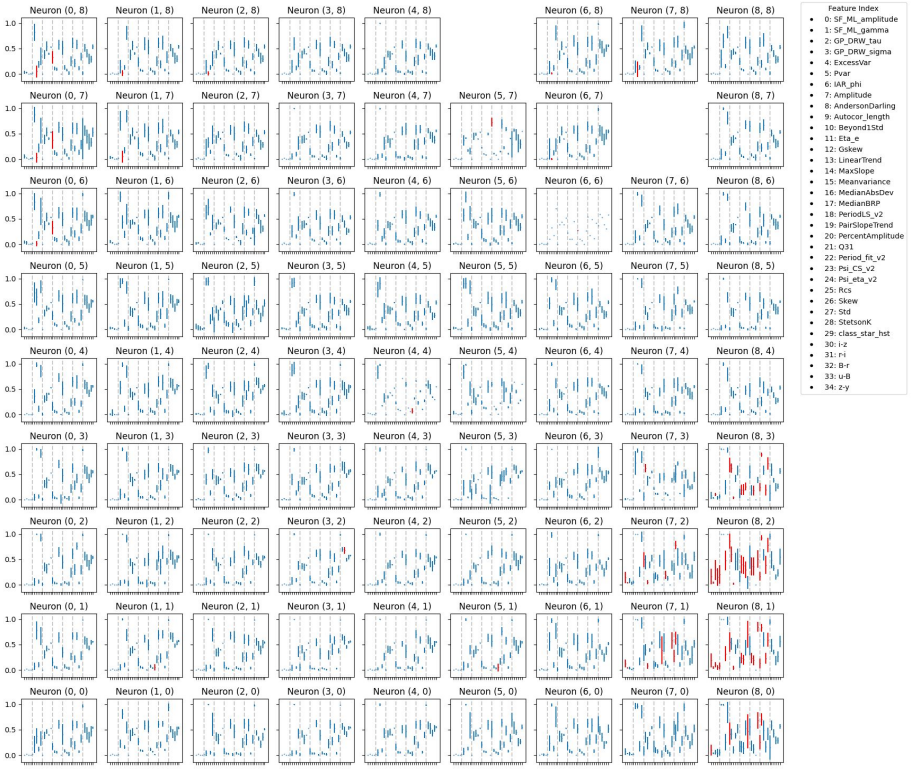
AGN, GAL, STAR

Unlabeled Set



Finding Similar Objects

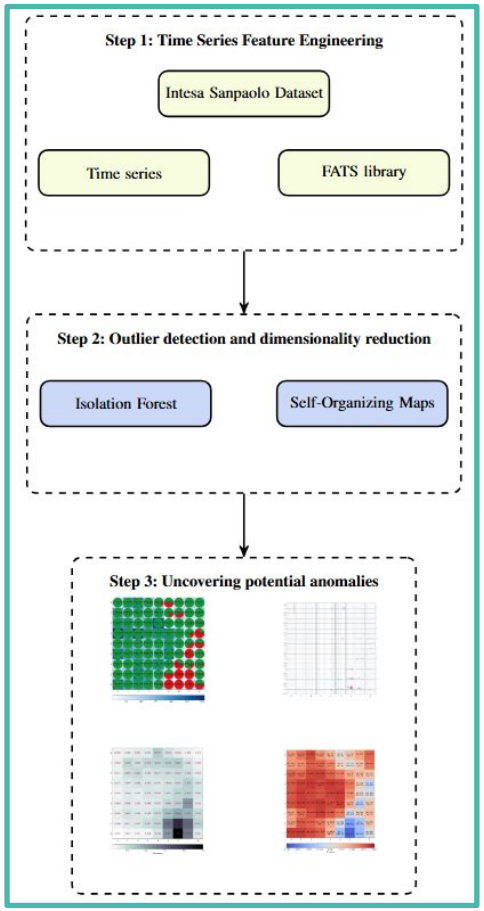
Distributions of the Feature Means

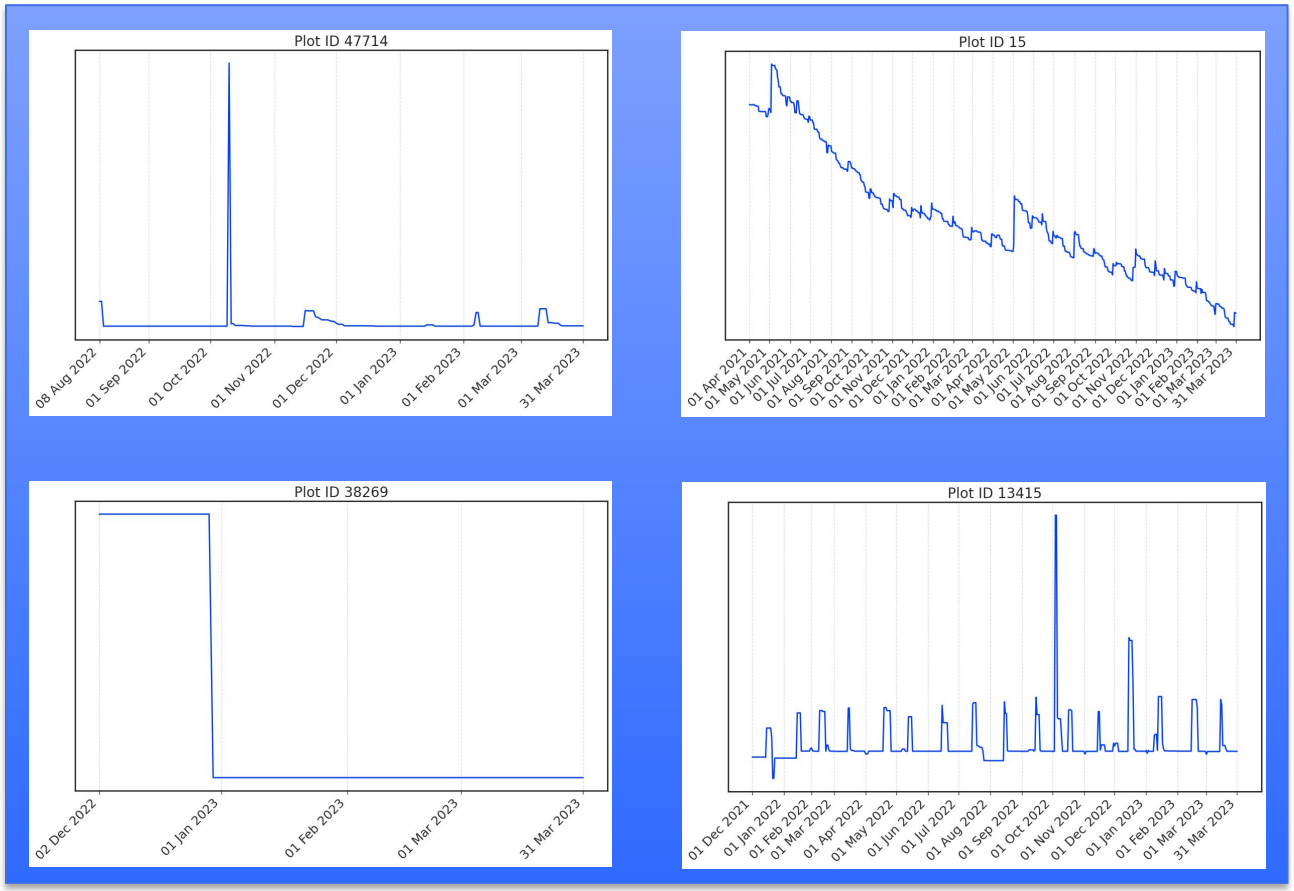


For each neuron, we calculated the global mean and standard deviation. Features are in red if:

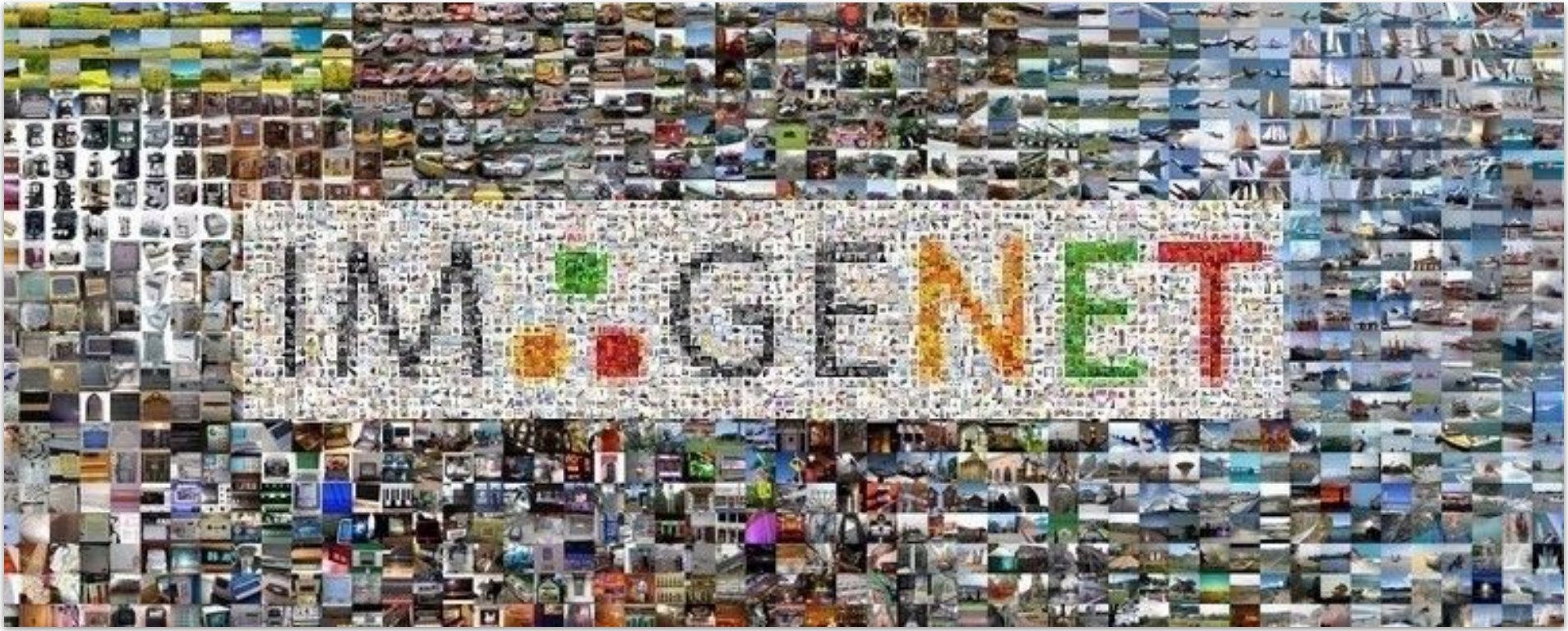
mean < global_mean - 2*global_std
or
mean > global_mean + 2*global_std

SOM & IF: Anomaly Detection and Outlier Identification for ISP

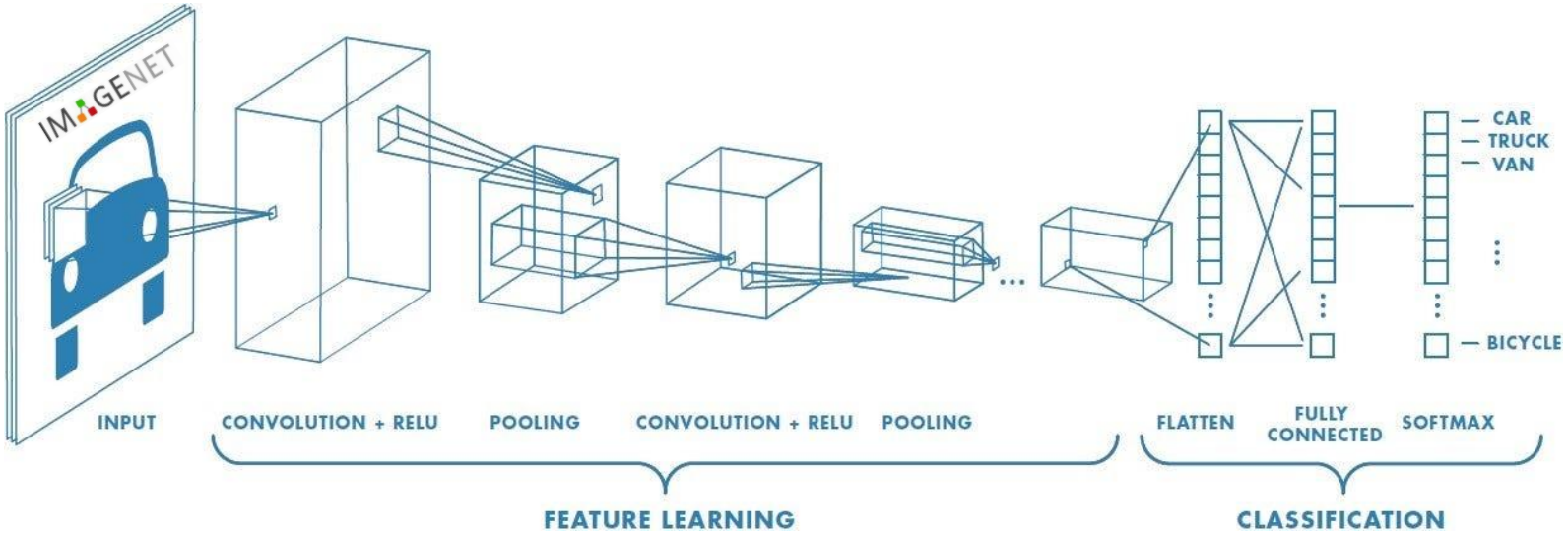




Transfer Learning Applications



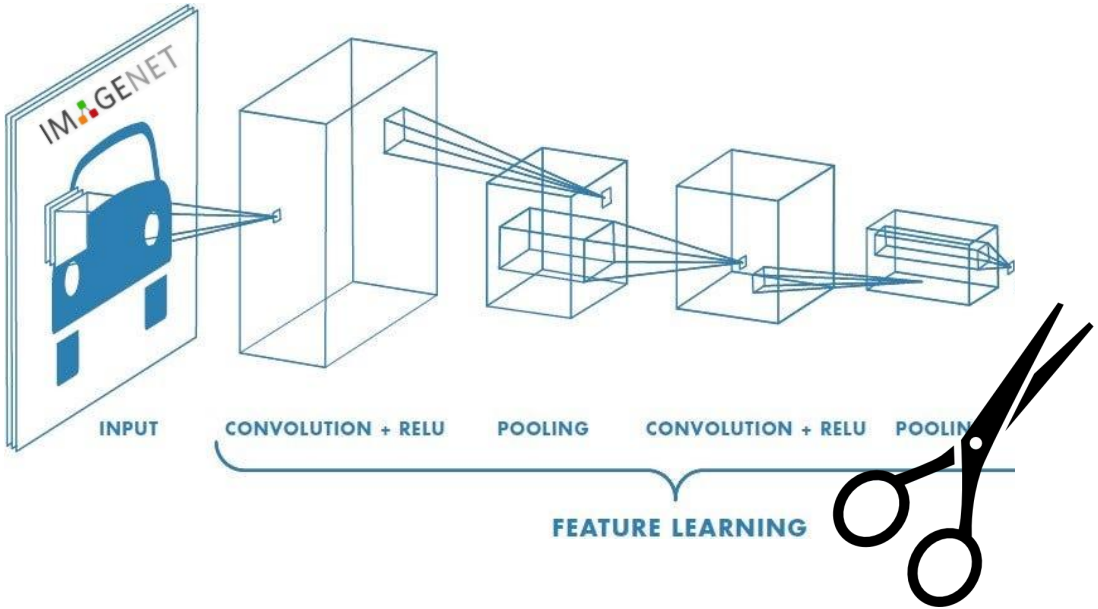
Convolutional Neural Networks



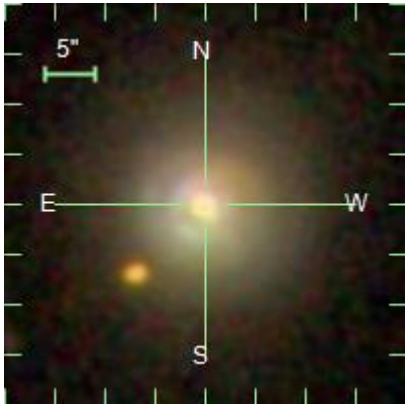
Convolutional Neural Networks

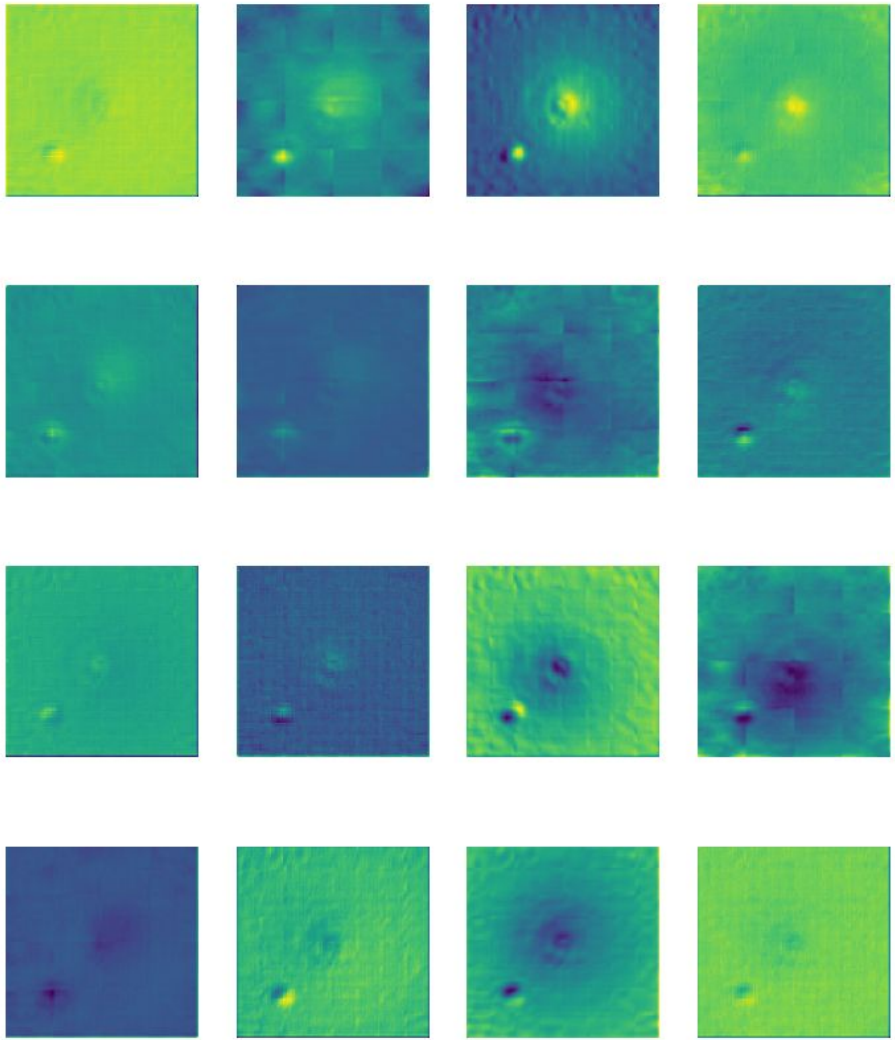
One man's trash is another man's treasure.

- English Proverb

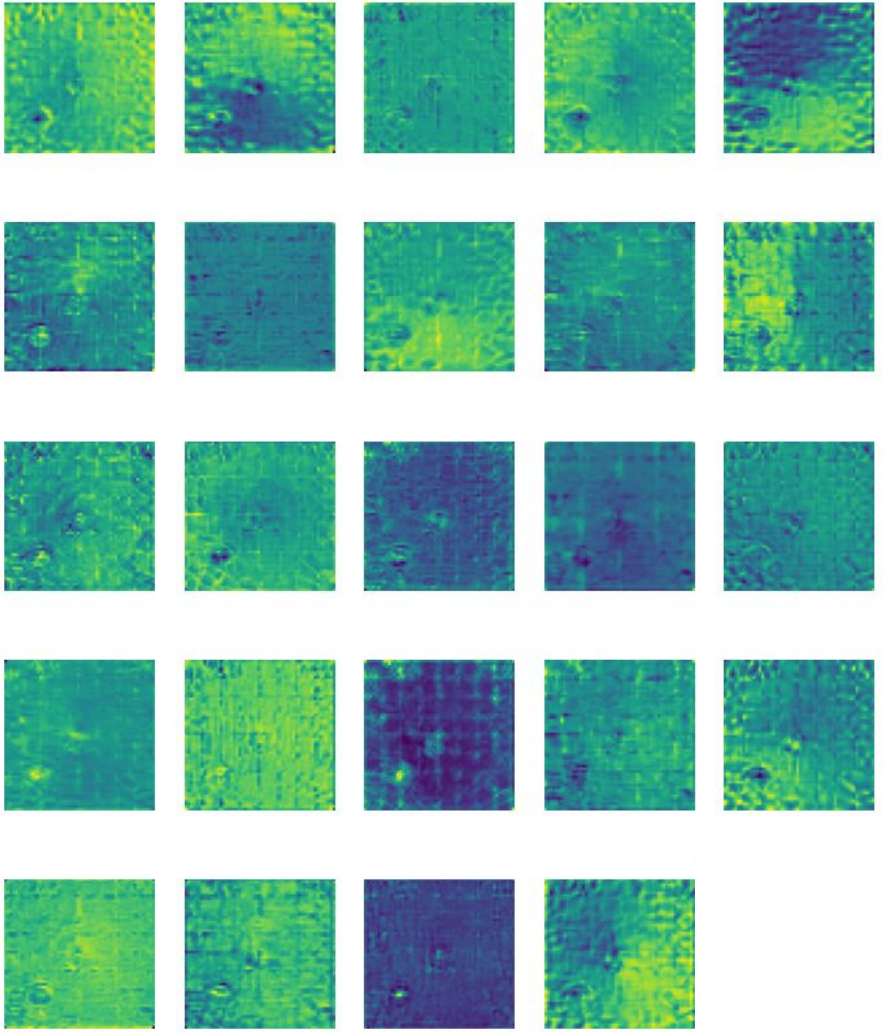


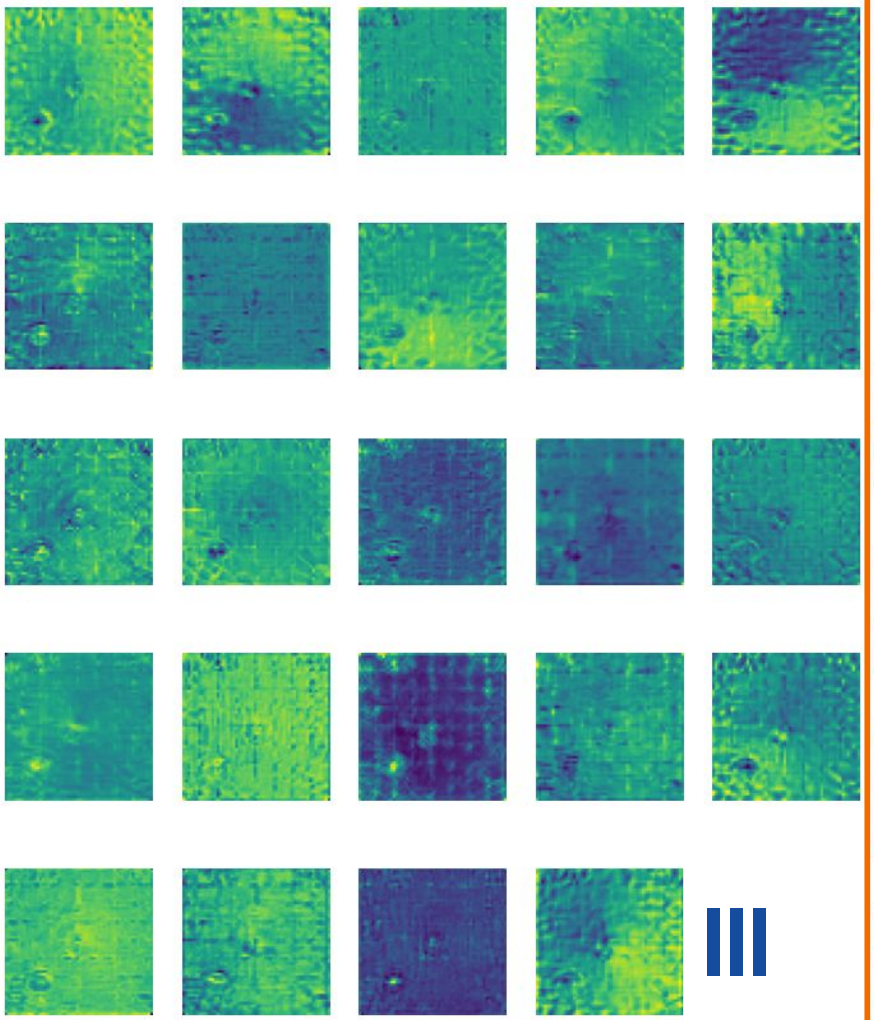
Application to an astronomical object

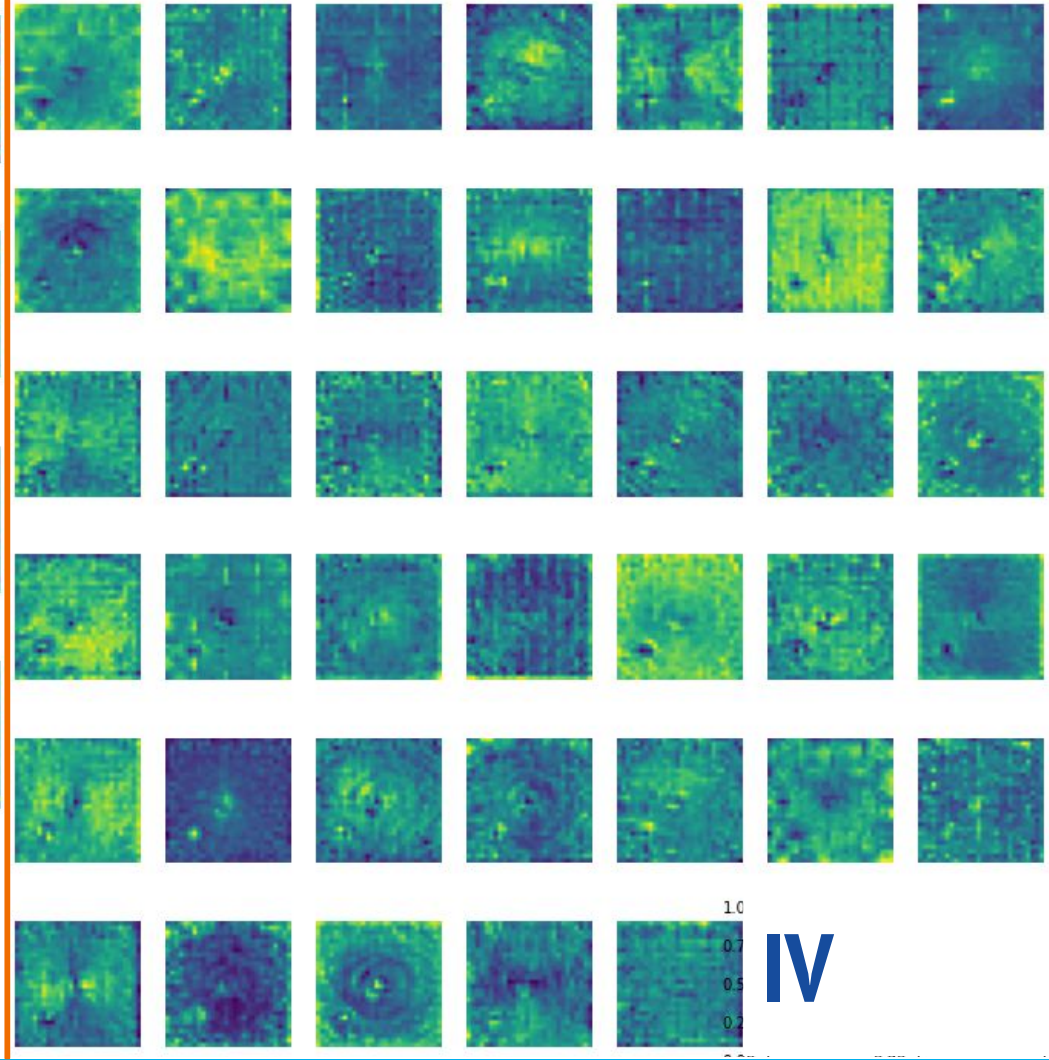




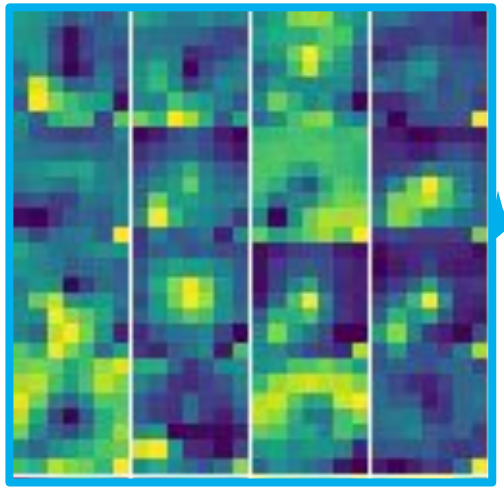
2nd layer

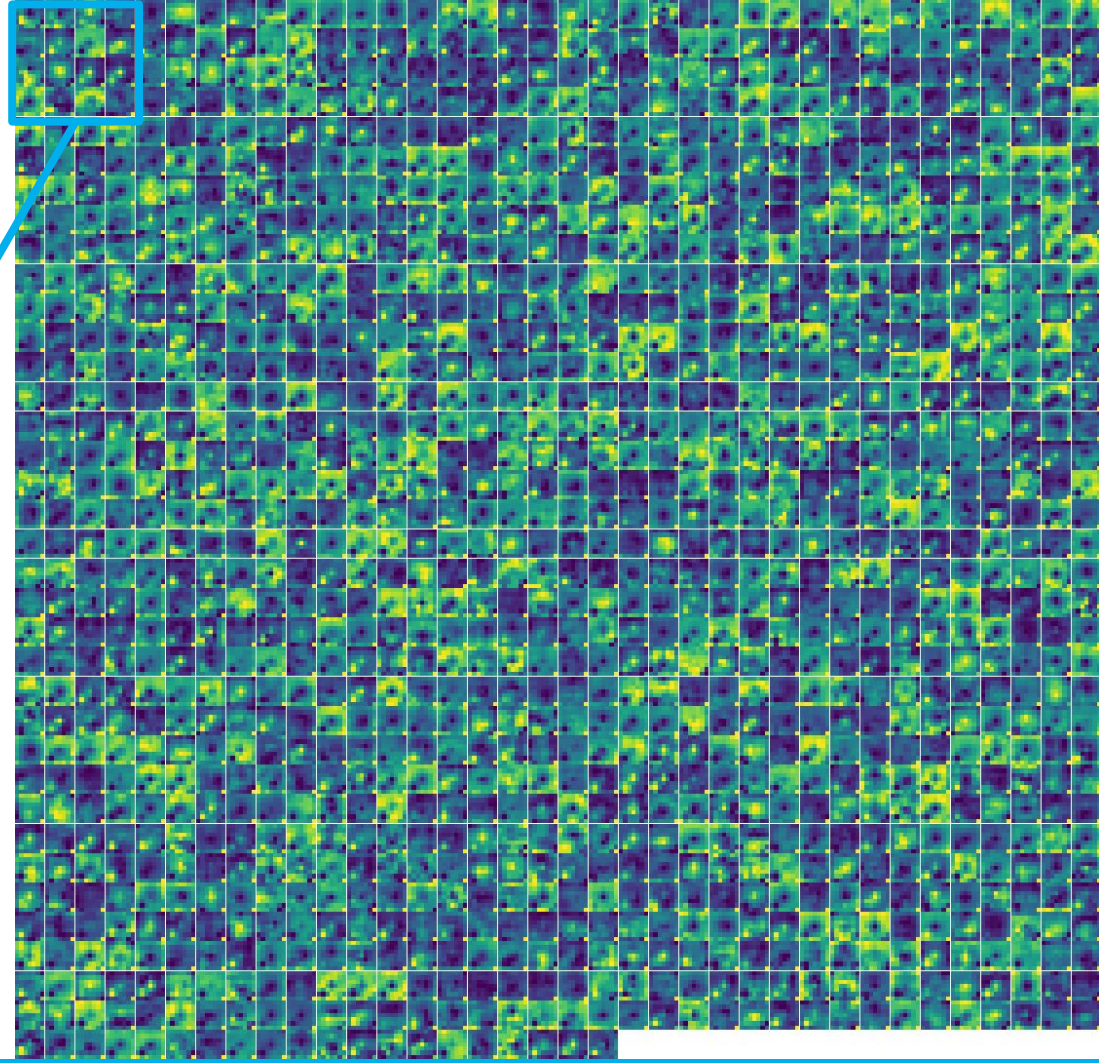



III

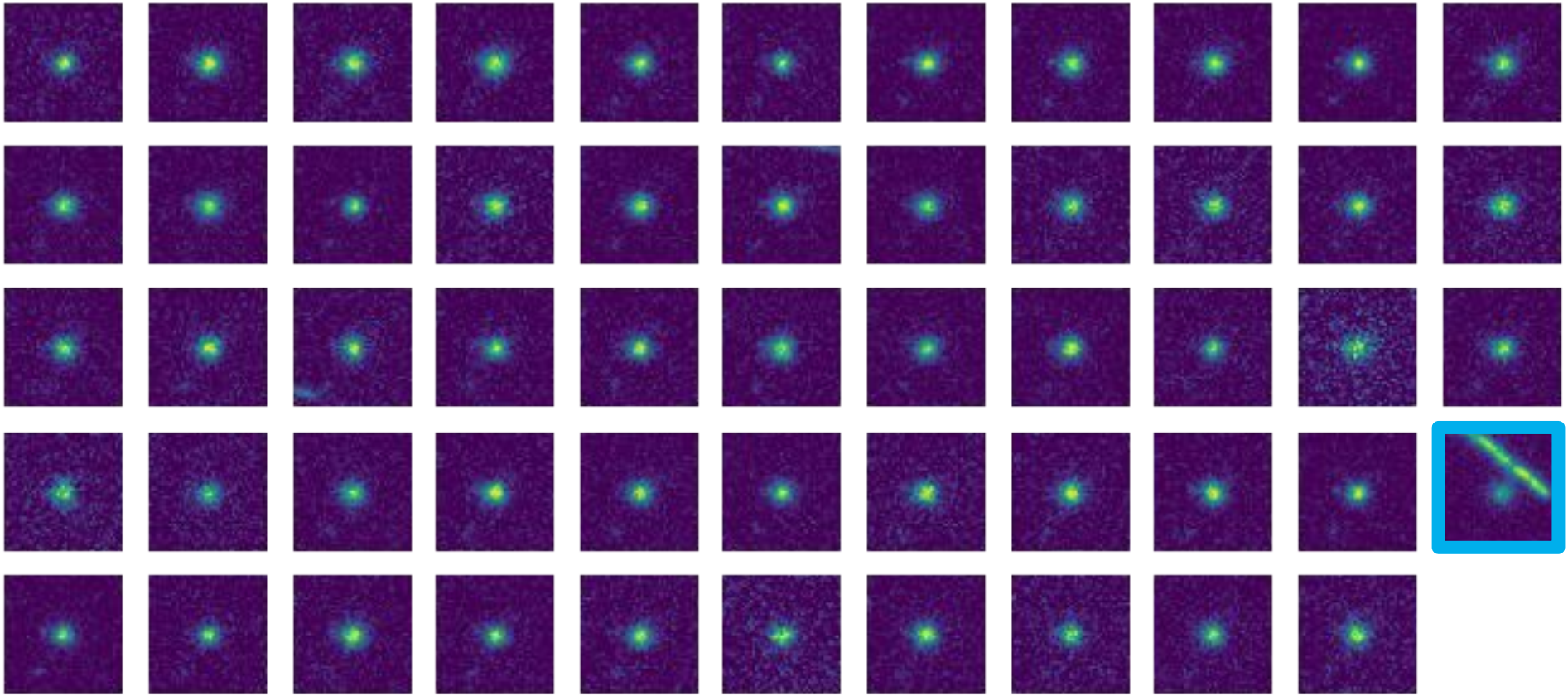

IV

Last layer (~1300 7x7 maps)

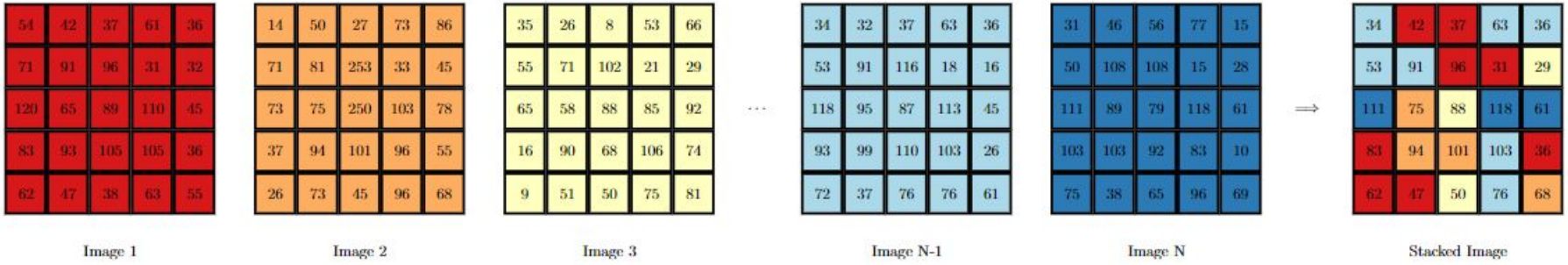


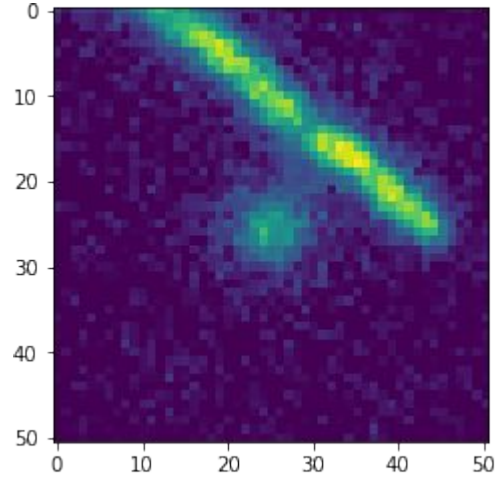


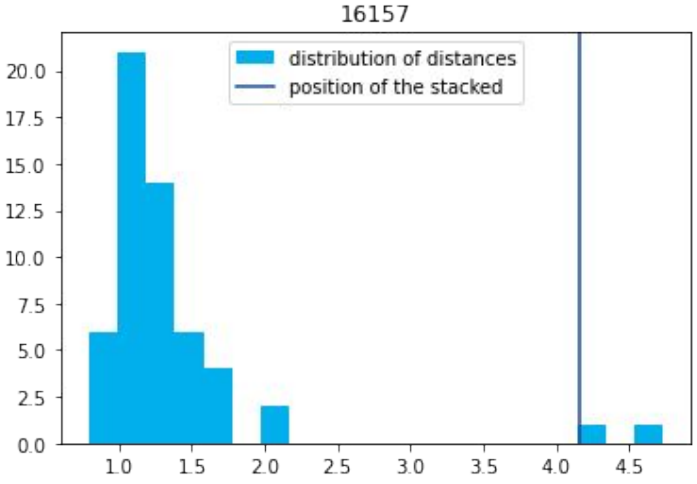
Example of a Problematic Image in Time Series

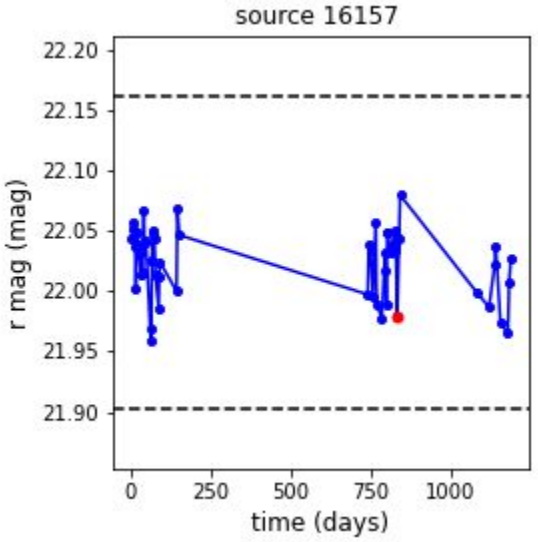


Stacked Image as Reference



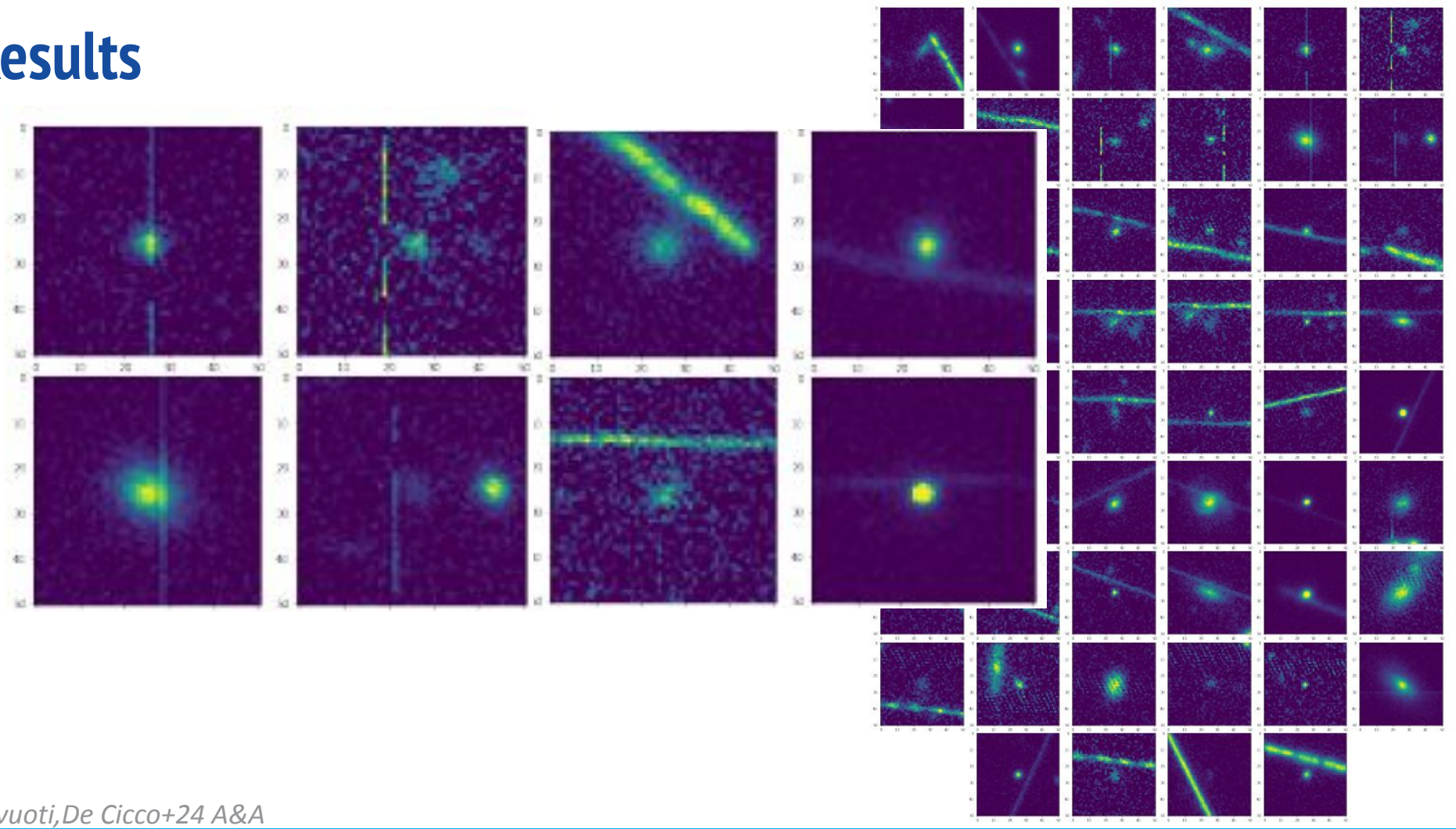






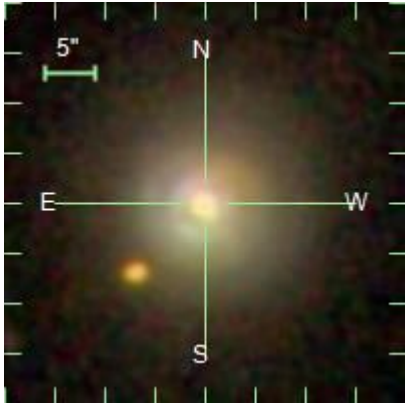
Cavuoti,De Cicco+24 A&A

Results

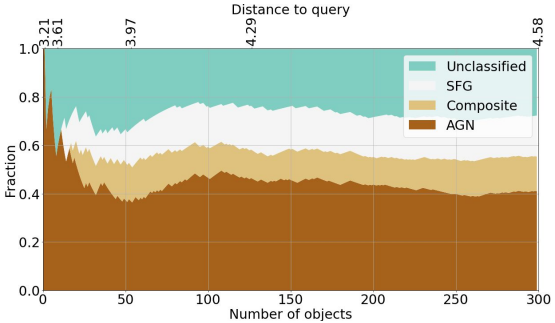


Cavuoti,De Cicco+24 A&A

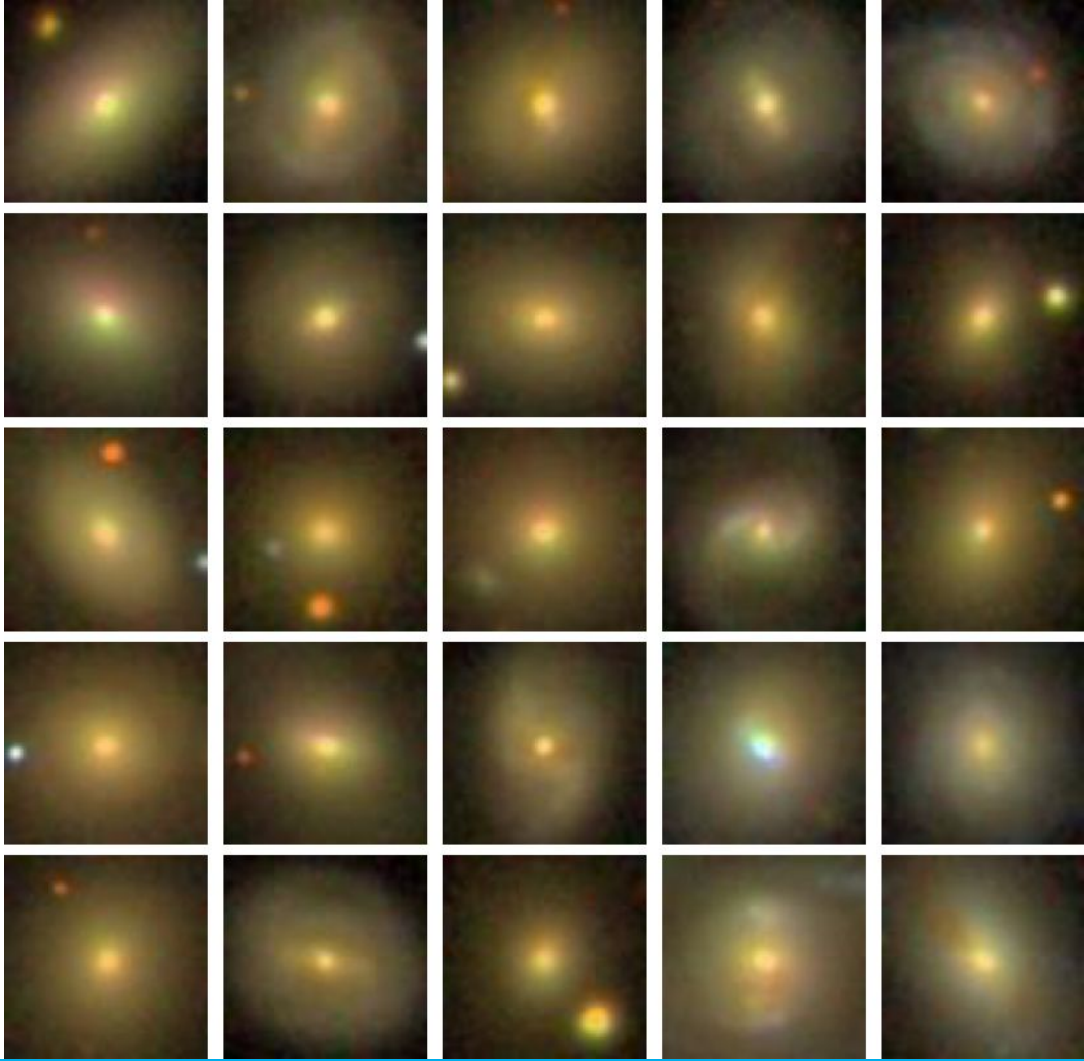
AGN 1



SDSS classification: **Galaxy - AGN**
BPT classification **AGN**
x-ray classification: **AGN**



Doorenbos et al. A&A, 666 (2022) A171

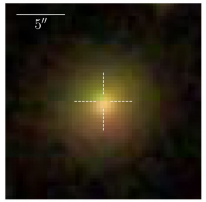


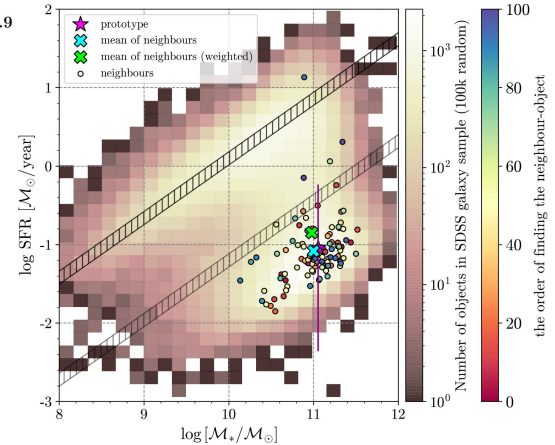
Star Formation Rate and Stellar Mass in Local Galaxies

SDSS J113337.09+114955.9

objid: 1237661812269908042
RA = 173.40455, **Dec** = 11.83220
Redshift = 0.084

SFR-$\mathcal{M}_*$ type: Quiescent (QG)
BPT class: Unclassified (non-AGN)
GZ2 morphology: E

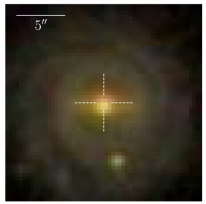


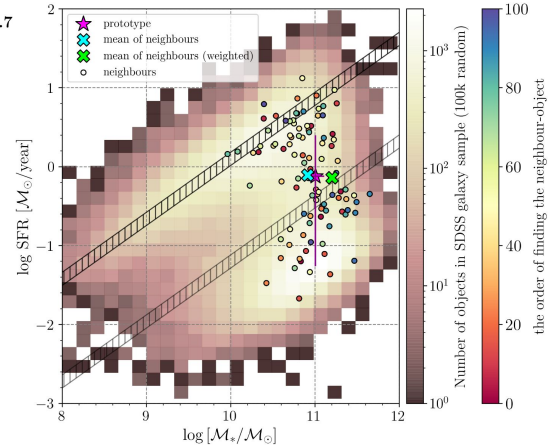


SDSS J093227.84+110253.7

objid: 1237661069245284654
RA = 143.11604, **Dec** = 11.04826
Redshift = 0.087

SFR-$\mathcal{M}_*$ type: Star-forming (SFG)
BPT class: low S/N AGN (AGN)
GZ2 morphology: SB

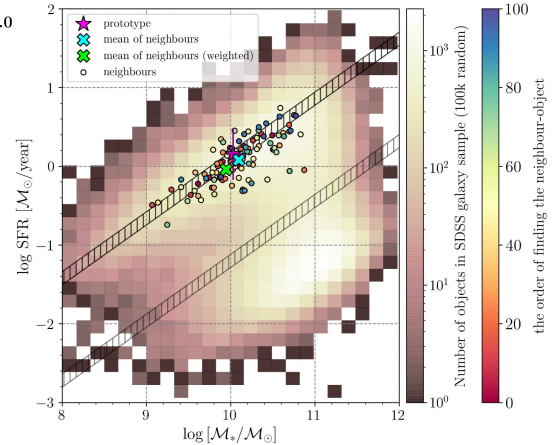




SDSS J111250.51+094316.0

objid: 1237658492811149399
RA = 168.21046, **Dec** = 9.72111
Redshift = 0.047

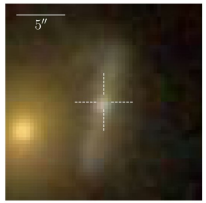
SFR-$\mathcal{M}_*$ type: Star-forming (SFG)
BPT class: SFG (non-AGN)
GZ2 morphology: S

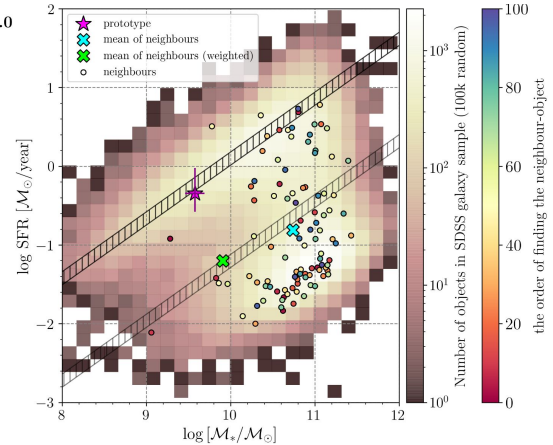


SDSS J110545.22+194705.0

objid: 1237667916486475792
RA = 166.43843, **Dec** = 19.78475
Redshift = 0.031

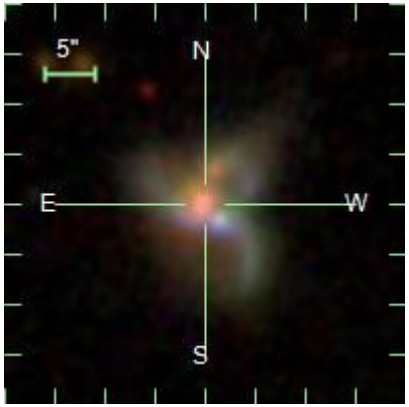
SFR-$\mathcal{M}_*$ type: Star-forming (SFG)
BPT class: SFG (non-AGN)
GZ2 morphology: Se(d)

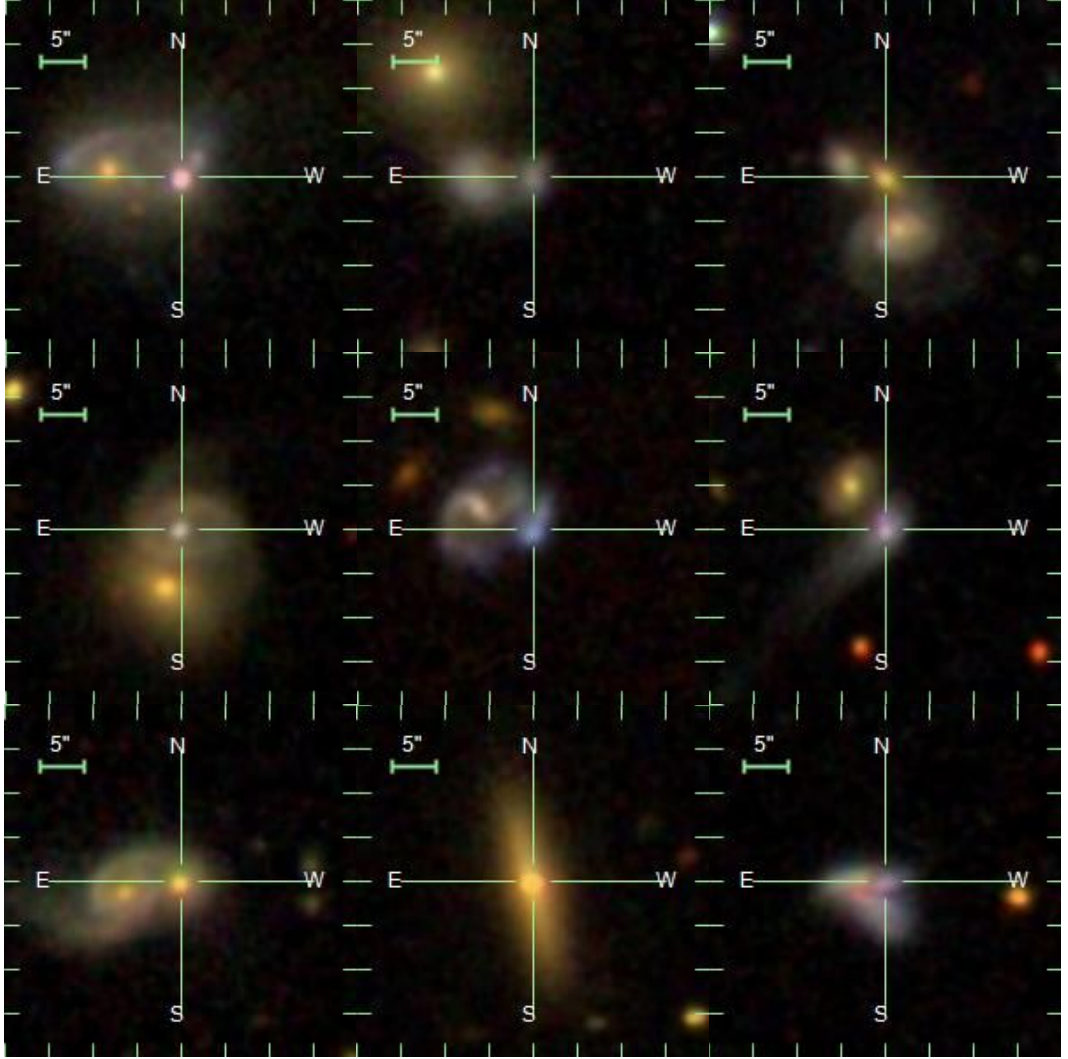




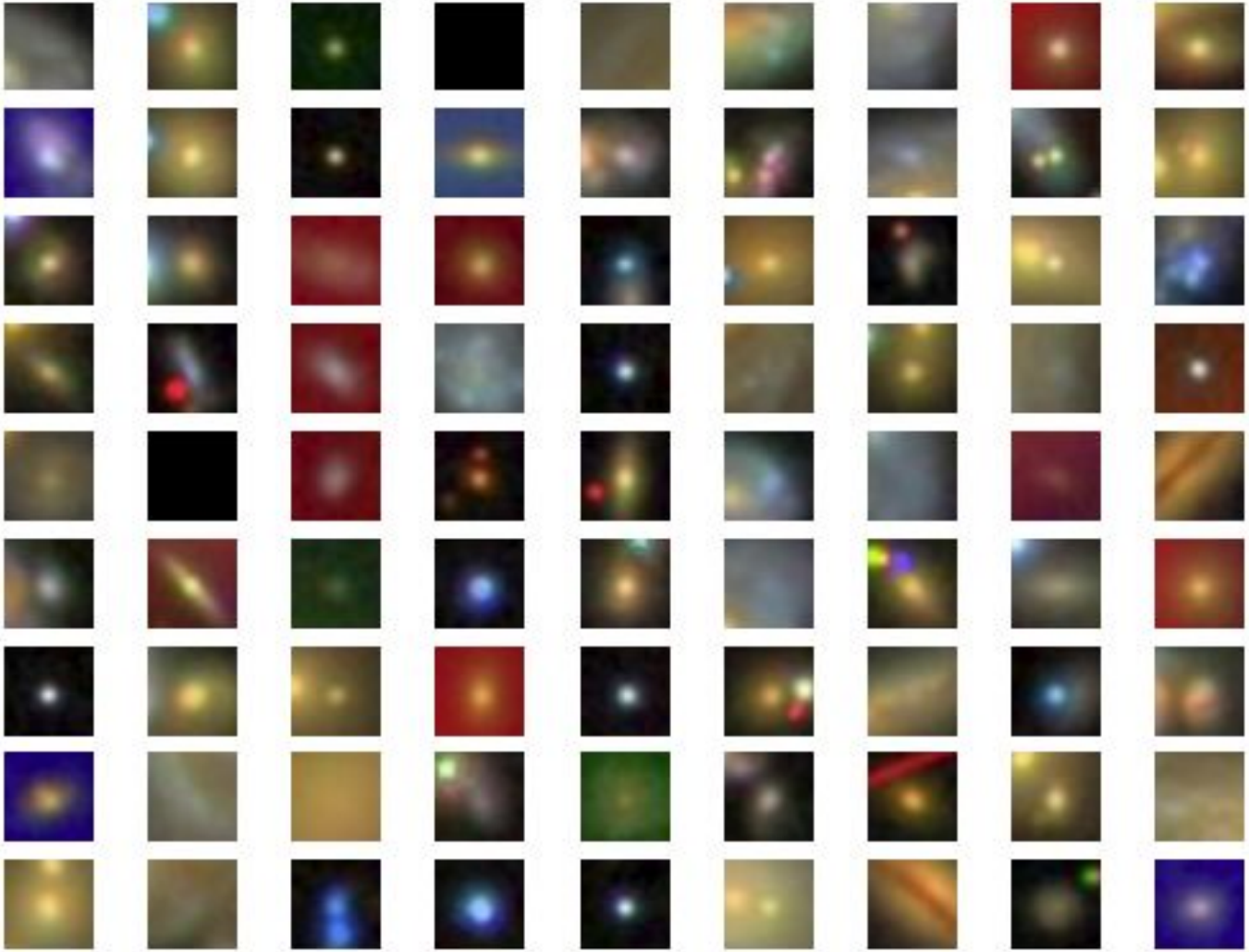
Torbaniuk et al. in prep.

Interacting Galaxies





What if we have no reference images?



“When you have eliminated the impossible, whatever remains, however improbable, must be the truth”
Sir Arthur Conan Doyle - The Sign of the Four

Galaxy-Galaxy Strong Lenses (GGSLs) and Cluster Members (CLMs) detection in Galaxy Clusters with CNN

(**Angora**, Rosati, Meneghetti, Grillo, Mercurio, **Brescia**, Bergamini, Acerbron, Vanzella)

Training and test set (for both GGSLs and CLMs identification):

Euclidization of HST (CLASH + HFF + HFF parallel) images (HST2Euclid tool, Bergamini+2025)

Labelling of samples:

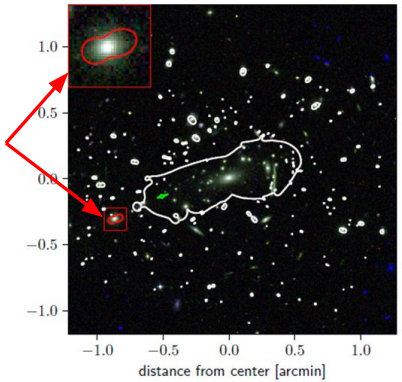
CLMs selected through their cluster rest-frame velocity separation: $|v| \leq 3000$ km/s

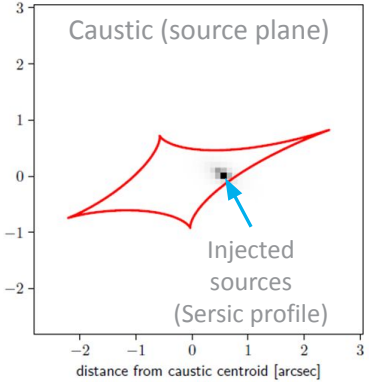
Selection of non-GGSLs by visually inspecting thousands of cutouts

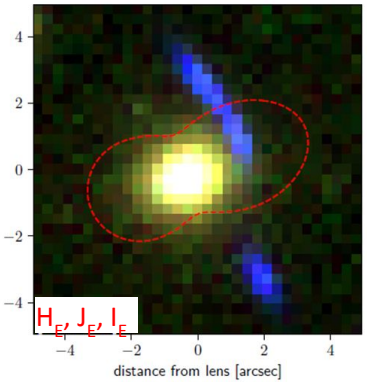
GGSLs injected around cluster galaxies by exploiting high-precision cluster lens models (Bergamini+ papers, Caminha+2016)

GGSL injection example

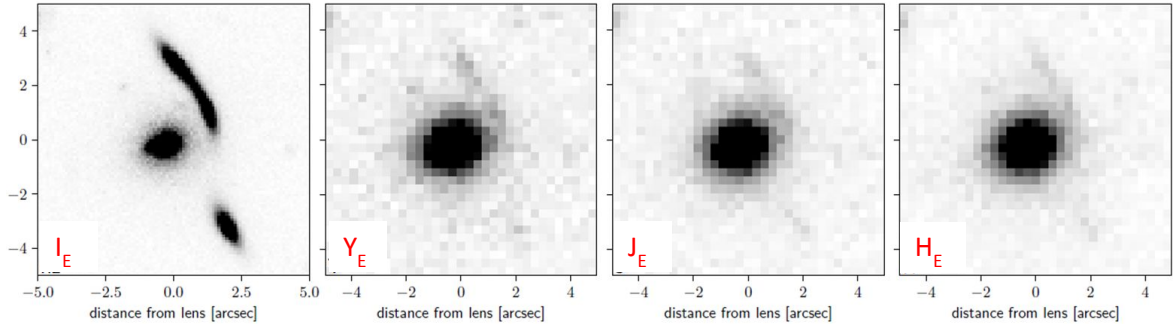
Critical lines (lens plane)





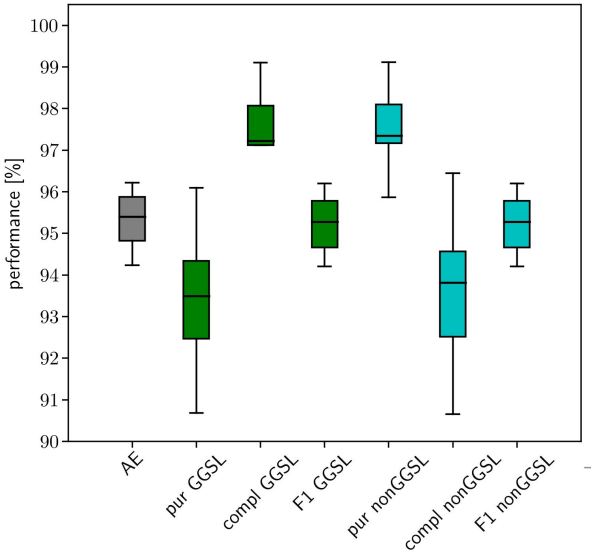


GGSLs

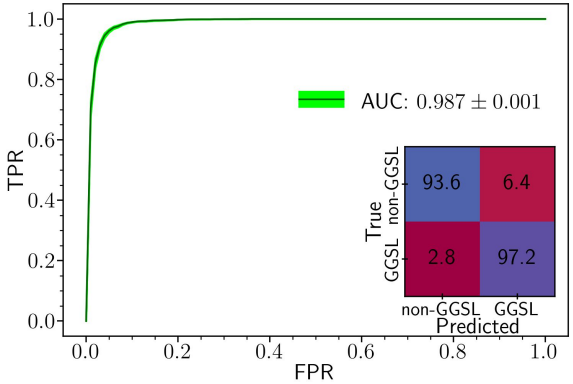


Performances on simulations (evaluated over 10 folds without intersection)

GGSLs detection

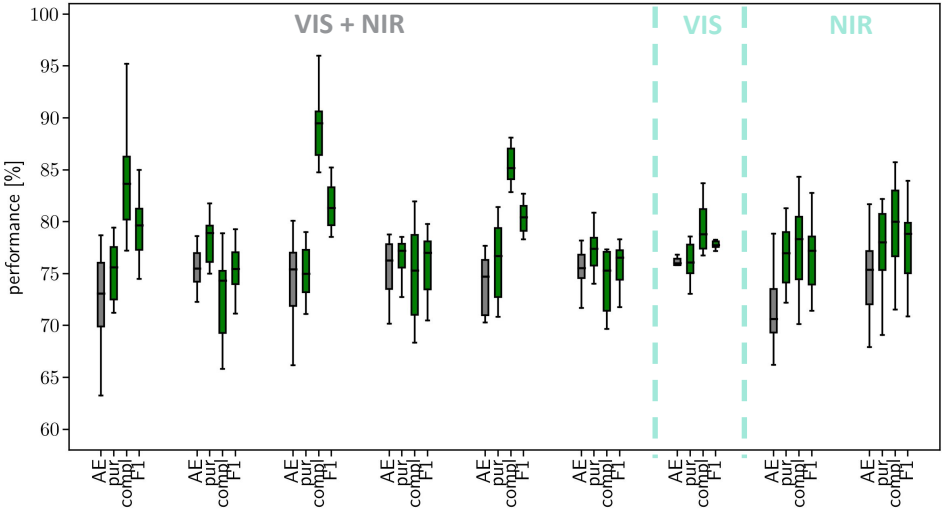


→ (Very) high performances purity-completeness ~ 93 – 97%



CLMs detection

By varying the network configuration and/or the filter combination



→ Acceptable performances on simulations: F1-score median ~ 80% (even if purity-completeness fluctuations ~ 75 – 90%)

Results on the ERO Abell 2390 imaging and Abell 2764

GGSLs detection

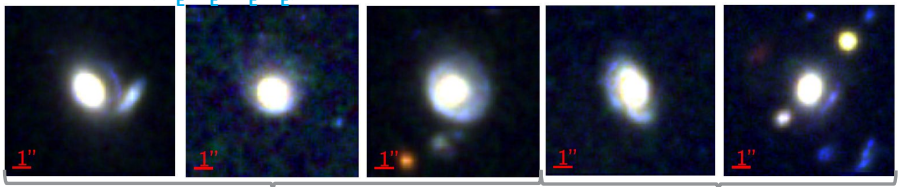
Abell 2390 (H_E+J_E, Y_E, I_E) candidates:

Already found (via visual inspection)

Small selection of other candidates

→ 44 candidates to be (visually) inspected through a voting mechanism

Abell 2764 (H_E+J_E, Y_E, I_E) candidates:



Already found (via visual inspection)

Small selection of other candidates

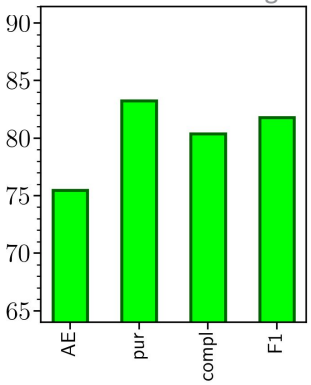
→ 34 candidates to be (visually) inspected through a voting mechanism

CLMs detection (only for Abell 2390)

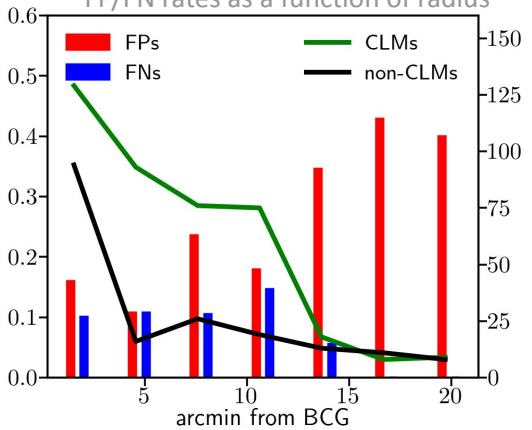
Using Available spectroscopy:

Abriola+(in prep.), Sohn+2020, Richard+2021, Abraham+1996

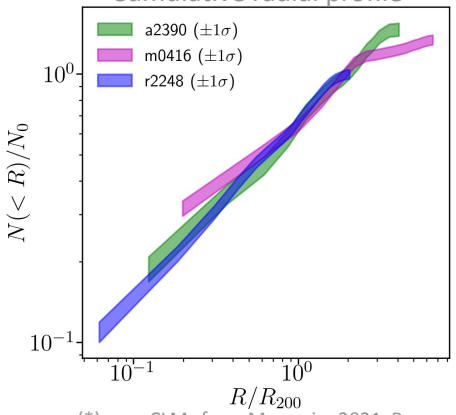
Ensemble learning



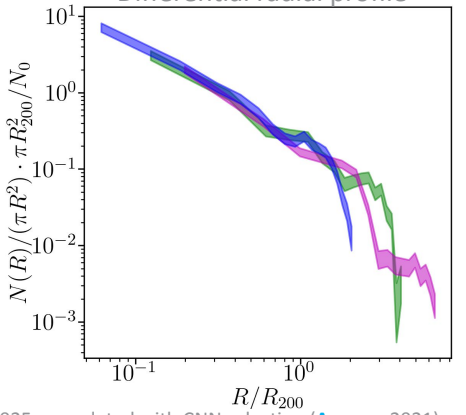
FP/FN rates as a function of radius



Cumulative radial profile*



Differential radial profile*

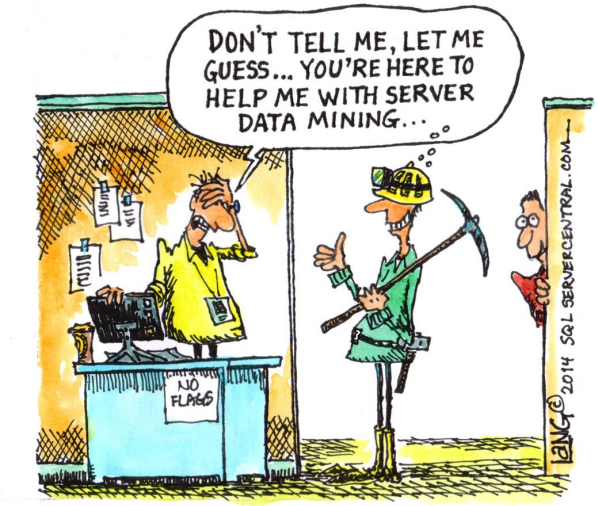


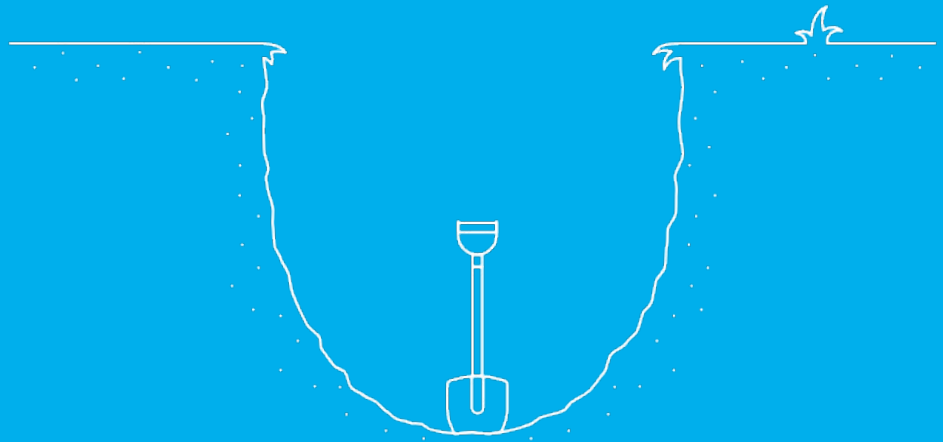
(*) spec CLMs from Mercurio+2021, Ragusa+2025; completed with CNN selection (**Angora**+2021)

Further Side-quests

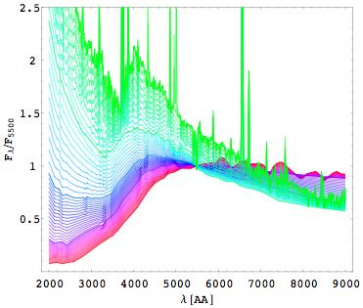
- **Medicine:**
  - Feature selection based on machine learning in MRIs for hippocampal segmentation (Tangaro+2015)
- **GeoMorphology:**
  - A novel approach to the classification of terrestrial drainage networks based on deep learning and preliminary results on solar system bodies (Donadio+2021)
  - Geomorphical Analysis Through Automatic Segmentation of Satellite and Topographic Images for Deep Learning Based Classification (D'Aniello MsC thesis)
- **Earth Science:**
  - AMBER -- Advanced SegFormer for Multi-Band Image Segmentation: an application to Hyperspectral Imaging (Dosi+2025)
- **Algorithms:**
  - Genetic algorithm modeling with GPU parallel computing technology (Cavuoti+ 2013)
  - HyCASTLE: A Hybrid ClAssification System based on Typicality, Labels and Entropy (Delli Veneri+2022)
- **Quantum Computing:**
  - A Quantum Genetic Algorithm for Cosmological Parameters Estimation (Sarracino+2025 Submitted)
  - Quantum Markov Chain Monte Carlo (Sarracino+2025 In Prep)
- **Neutrino Physics:**
  - Statistical analysis of the trigger algorithm for the NEMO project (Riccio+2006)
- **Dark Matter Direct Detection:**
  - DEAP-3600 experiment (Signal/Background classification)

Thank you for your attention!





Spectral Energy Distribution (SED) template fitting methods



Library of M template spectra (M<100)

Convolve with filter bandpasses for a specific survey

Stretch templates for redshift (z) assuming constant step Δz in an interval range z_{min} , z_{max}

$$SED(T_i, z_{min} + n\Delta z), i \in [1, M], n \in INT\left[\frac{z_{max} - z_{min}}{\Delta z}\right]$$

Find best fitting i,j using any optimization method

Templates: either synthetic or observed

Arbitrary choice of templates, lots of assumptions on physics, strong dependence on zero points, photometric calibrations, etc.

But they go very deep, well beyond the spectroscopic limit

Interpolative (empirical) methods

Let f be the complex (unknown) function which maps the input photometric parameter space onto the redshift space:

$$f: x \to z, where\ x = \{x_1, x_2, \ldots, x_n\}$$

Are the input parameters, (hereafter **features**)

Empirical methods use a subset of the objects (TRAINING SET) for which the spectroscopic redshifts (or in this case, target) are known, to infer the mapping function

Performances are then evaluated on a second disjoint dataset (TEST SET) for which the target is known and which has not been used during the training (BLIND TEST)

More accurate
No assumptions on physics, almost independent on zero points, photometric calibrations, etc.

They are bounded by the spectroscopic limit

Selected Labeled Set (LS)

Main sample: 16,490 sources with *uBrizy* counterparts.

Stars: 1,000; selected from the COSMOS ACS catalog (Koekemoer+07, Scoville+07a; a morphological classification is provided), belonging to the stellar sequence on an *r-z* vs. *z-k* diagram, and cross-matched with other COSMOS catalogs to avoid conflicting classification.

"Inactive" galaxies: 1,000; selected from our main sample (below the variability threshold defined in De Cicco+19), classified as "non active" based on the best-fit templates (Bruzual & Charlot 03) reported in the COSMOS2015 catalog (Laigle+16), and cross-matched with other COSMOS catalogs to avoid conflicting classification.

AGN: 414, consisting of:

- 225 sAGN1, i.e., spectroscopically confirmed unobscured (Type 1) AGN and
- 122 sAGN2, i.e., spectroscopically confirmed obscured (Type 2) AGN

from the Chandra-COSMOS Legacy Catalog (Marchesi+16, Civano+16);

- 67 sources classified as AGN according to the MIR selection criterion by Donley+12 and with no spectroscopic classification.