Big Data Challenges and Deep Learning Applications to Astronomy

Prof. Pablo Estévez, Ph.D., IEEE Fellow
Department of Electrical Engineering, Universidad de Chile
& Millennium Institute of Astrophysics, Chile

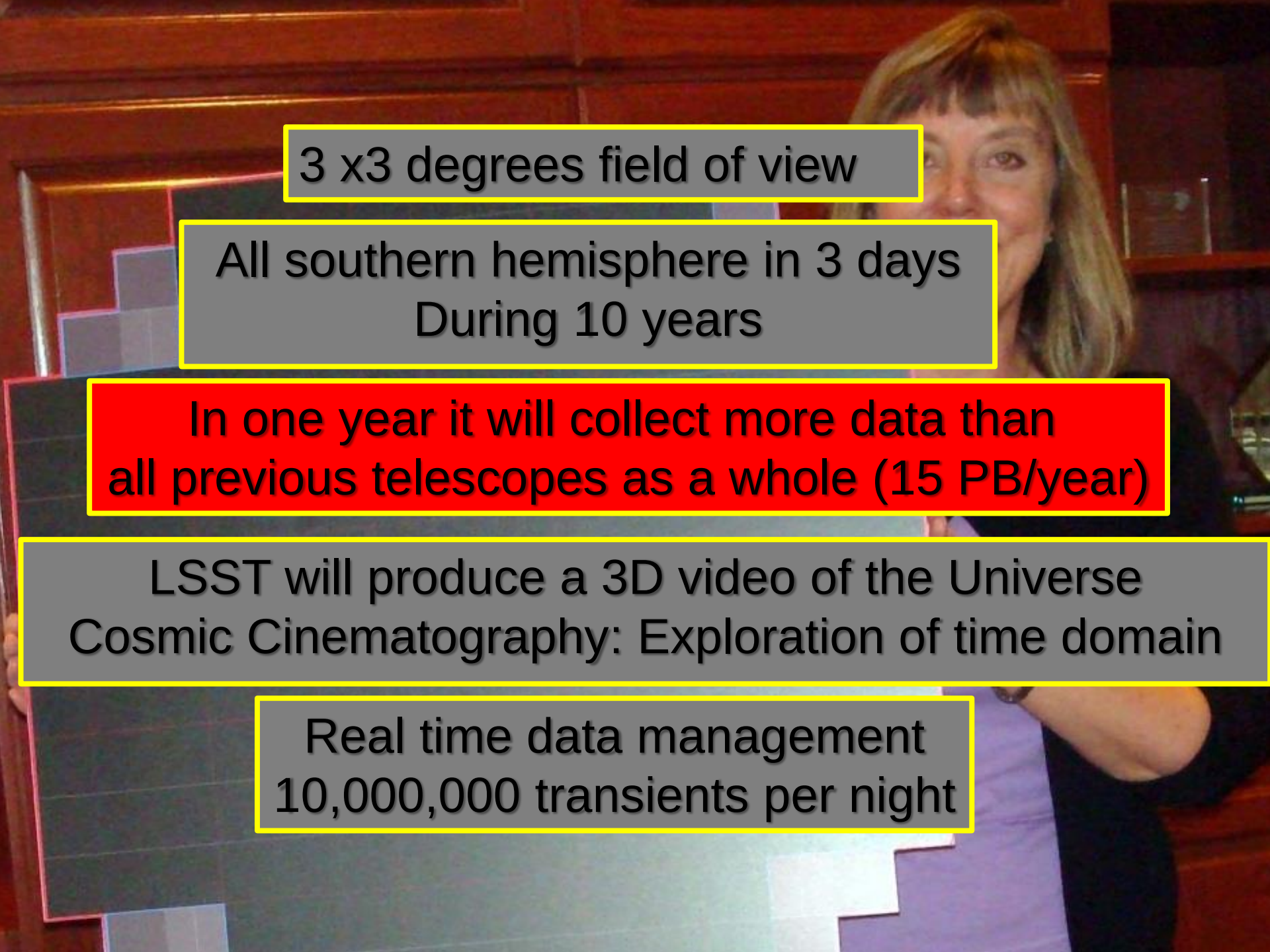
ITS, ULISBOA

Lisbon, Portugal, March 25, 2019

Contents

- Astronomy Context / Big Data Challenges
- Deep-HiTS Real/Bogus Clasiffier
- Clustering using Deep Variational Embedding Autoencoders
- Recurrent Convolutional Neural Network Sequential Classifier
- Conclusions

Large Synoptic Survey
Telescope (LSST)
Cerro Pachón, Chile, 2022
8.4m telescope +
3.2 gigapixel camera
AURA: Association of Universities for Research in Astronomy (Director: Chris Smith)



3 x3 degrees field of view

All southern hemisphere in 3 days
During 10 years

In one year it will collect more data than
all previous telescopes as a whole (15 PB/year)

LSST will produce a 3D video of the Universe
Cosmic Cinematography: Exploration of time domain

Real time data management
10,000,000 transients per night

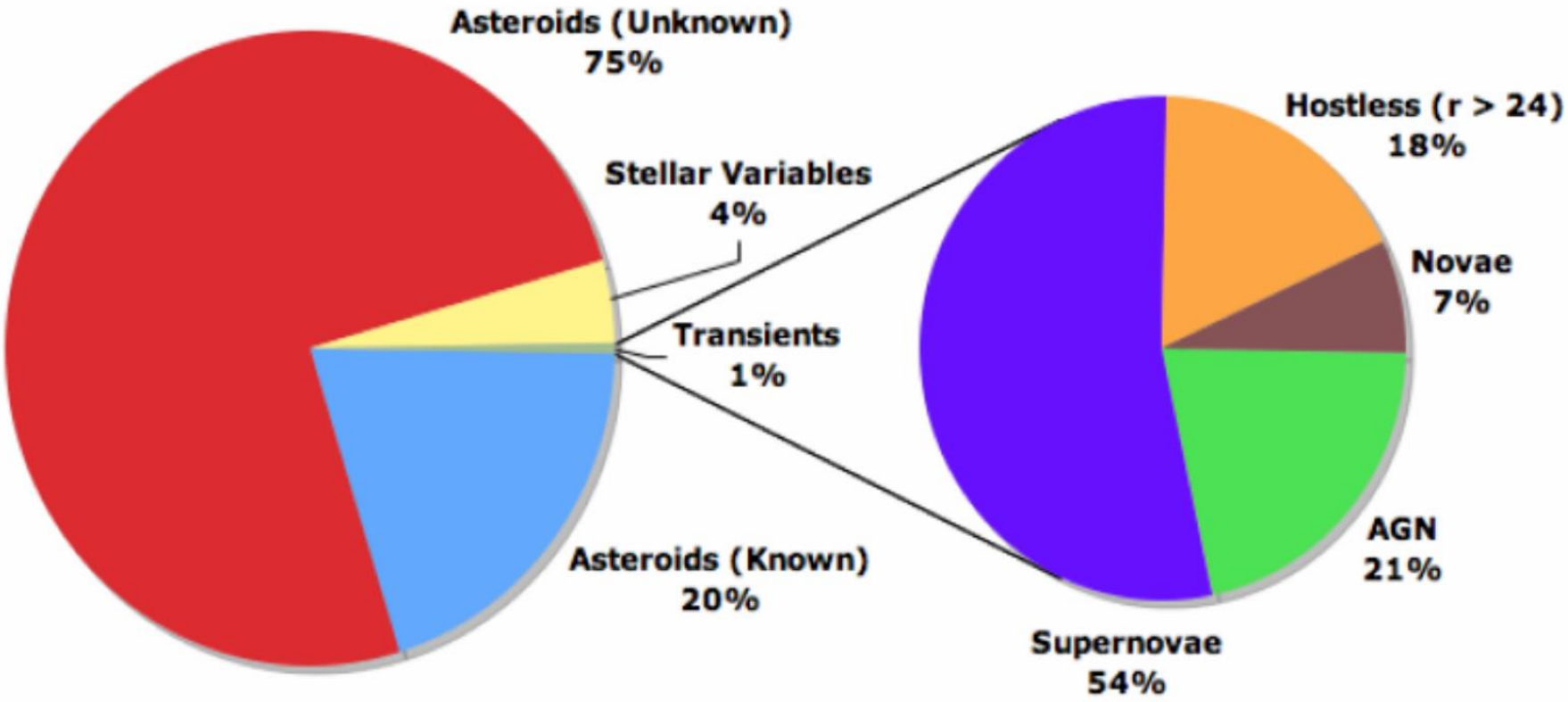
LSST: Big Data Challenges

- Mining in real time a massive data stream of ~2 Terabytes per hour for 10 years
- Classify more than 50 billion objects and follow up many of these events in real time
- Extracting knowledge in real time for ~1 to10 million events per night
- Broker System ALERCE: Automatic Learning for the Rapid Classification of Events ALeRCE.science
- Discovering the unknown unknowns (serendipity): the things that we do not even know that we don´t know!

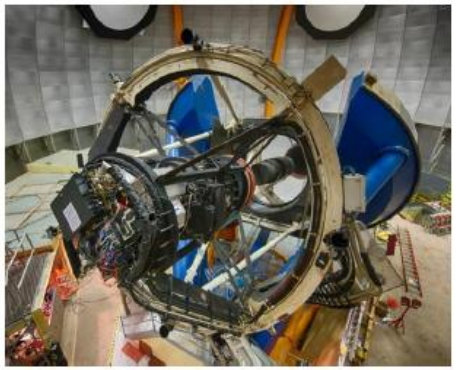
DATA PER YEAR

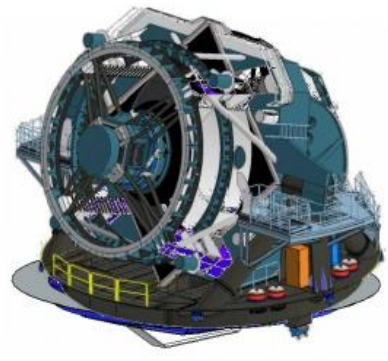
Credits: ALMA, Maccarena Gonzalez:

Composition of LSST streams



Astronomical Alerts





HiTS

2013-2015 (~3 weeks)

0.2 TB per night

~20 million objects

~100 million measurements

~0.1 million alerts per night

→ **10x**

ZTF

2018-2020

1.4 TB per night

~1 billion objects

~1 trillion measurements

~1 million alerts per night

→ **10x**

LSST

2022-2032

15 TB per night

~37 billion objects

~7 trillion measurements

~10 million alerts per night

ALeRCE Broker

- A broker is an alert management system
- We are building the first Chilean astronomical alert broker, ALeRCE
- There are only 5 broker systems under development in the world
- ALeRCE challenges: classify alerts with characteristic timescale of a few milliseconds!
- Serve the community to enable sensible follow up (friendly access tools)

ALeRCE
Automatic Learning for the Rapid Classification of Events

Guillermo Cabrera (UdeC / MAS)
Andrew Connelly (Dirac-UW)
Pablo Estévez (DIE-UCh / MAS)
Susana Ehyeramendy (UC / MAS)
Francisco Förster (CMM-UCh / MAS)
Matthew Graham (Caltech / ZTF)
Pablo Huijse (UACH / MAS)
Ashish Mahabal (Caltech / ZTF)
Juan Carlos Maureira (CMM-UCh / MAS)
Karim Pichara (UC / MAS)
Giuliano Pignata (UNAB / MAS)
Pavlos Protopapas (IACS-Harvard)
Andrea Rodríguez (UdeC)

In a nutshell

Real-time classification of large etendue telescope streams (e.g. ZTF, HATPI, LSST)

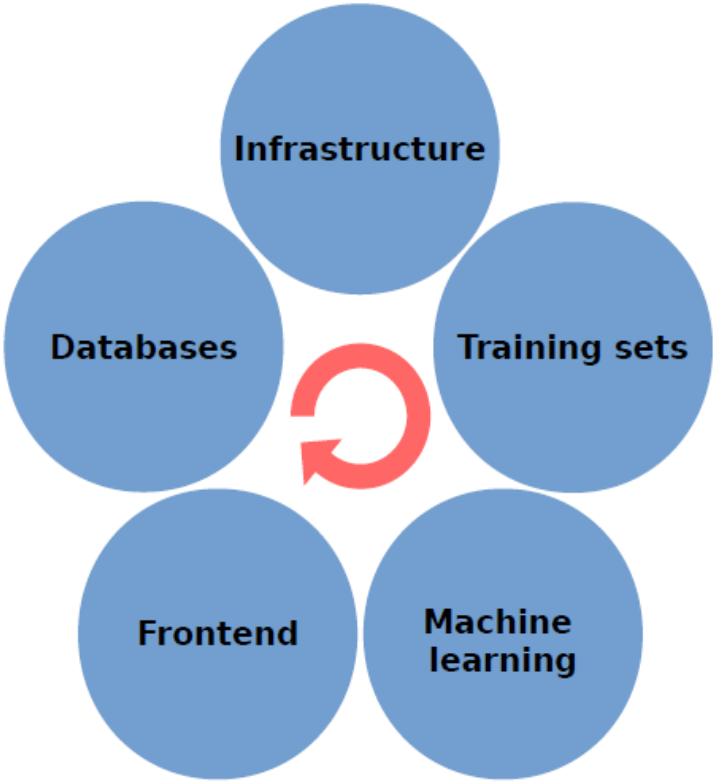
Federated & hierarchical classification model. Central node connected to Chilean national academic network (100 Gbps). Multiple classifiers downstream.

Facilitate **follow up** for:

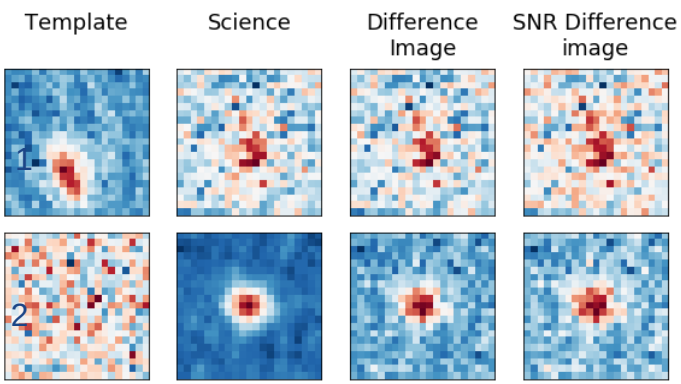
1. real-time reaction
2. classification
3. characterization

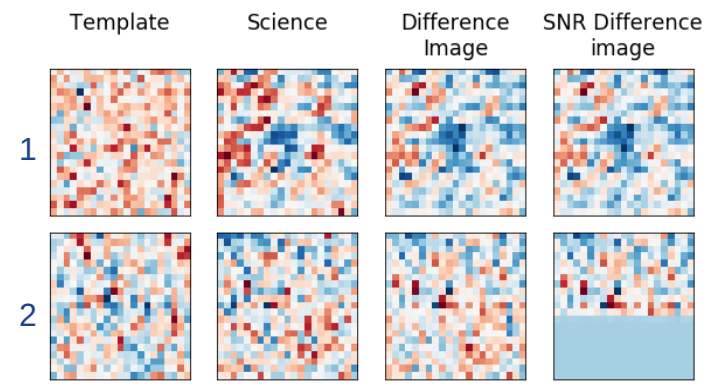
Output stream + visualization tools for full stream and individual alerts

Interdisciplinary teams

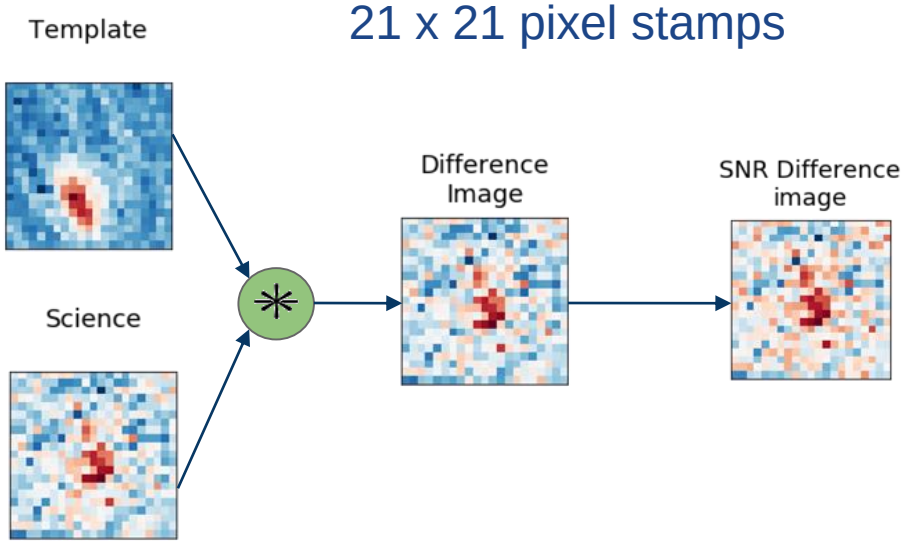


2013 HiTS dataset





21 x 21 pixel stamps



- Training set -> 1,250,000 samples
- Validation set -> 100,000 samples
- Test set -> 100,000 samples

To discriminate between real transients and bogus events with low FPR, FNR

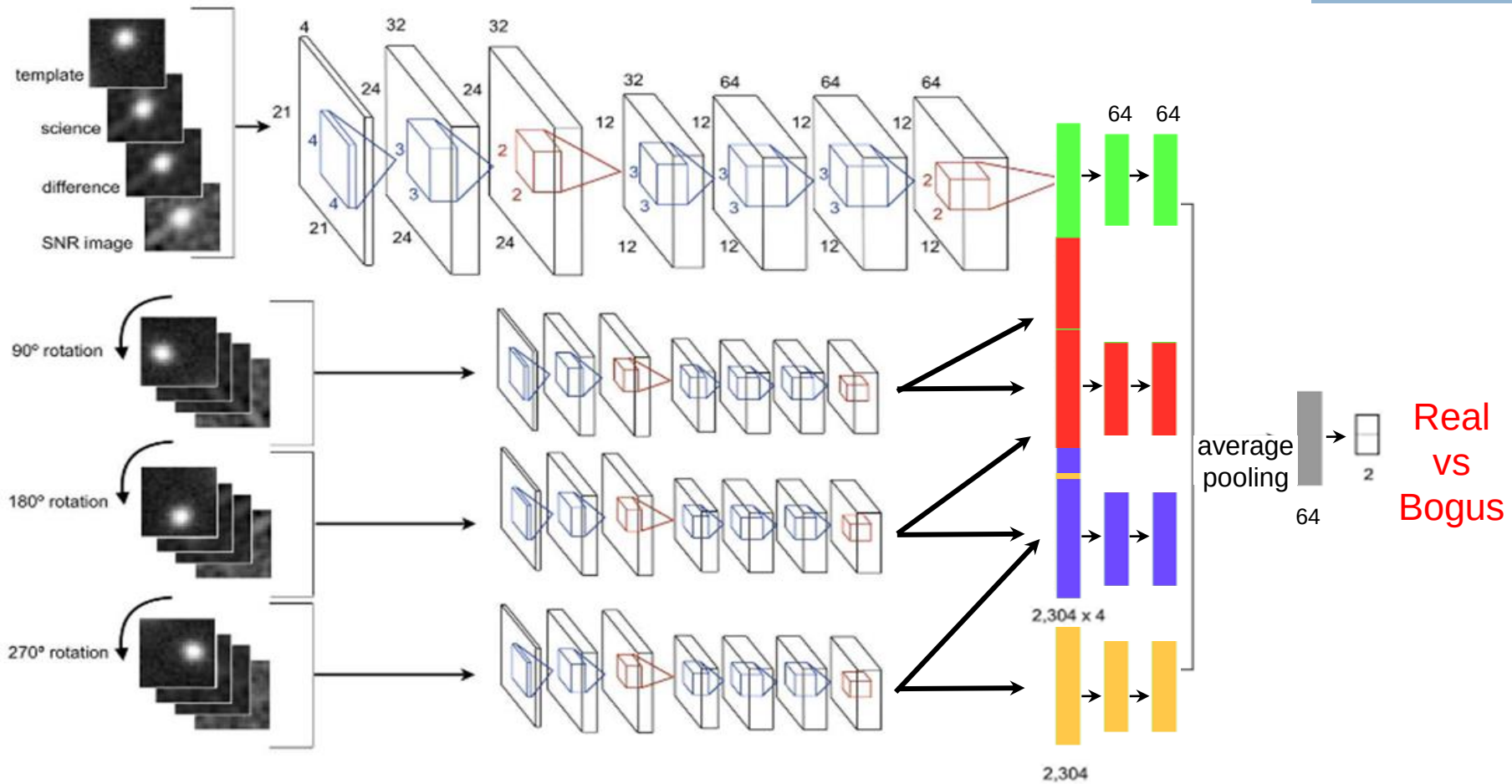
Inception Movie

Inception: Our version

Francisco Forster

Guillermo Cabrera

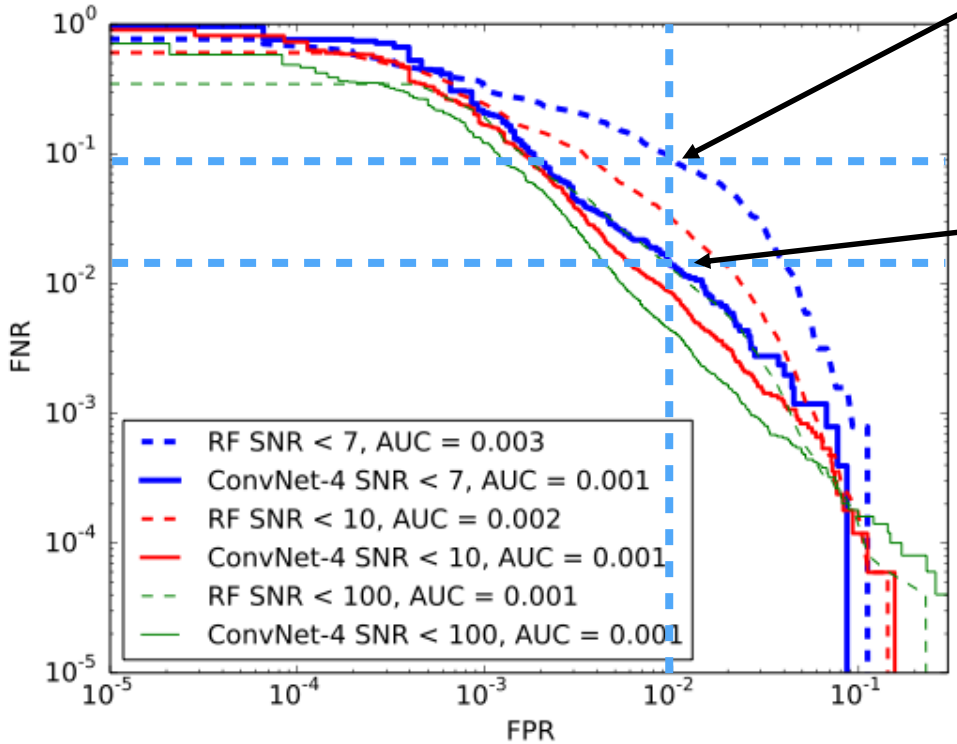
Deep-HiTS Real/Bogus Classifier: Supervised Approach



Cabrera-Vives, G., Reyes, I., Forster, F., Estevez, P.A., Maureira, J.C., “Deep-HiTS: Rotation Invariant Convolutional Neural Network for Transient Detection”, Astrophysical Journal, Feb. 2017.

Reyes, E., Estevez, P.A., Cabrera-Vives, G., Forster, F. “Enhanced Deep Neural Network Model”, 2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, July 2018.

Detection Error Tradeoff (DET)

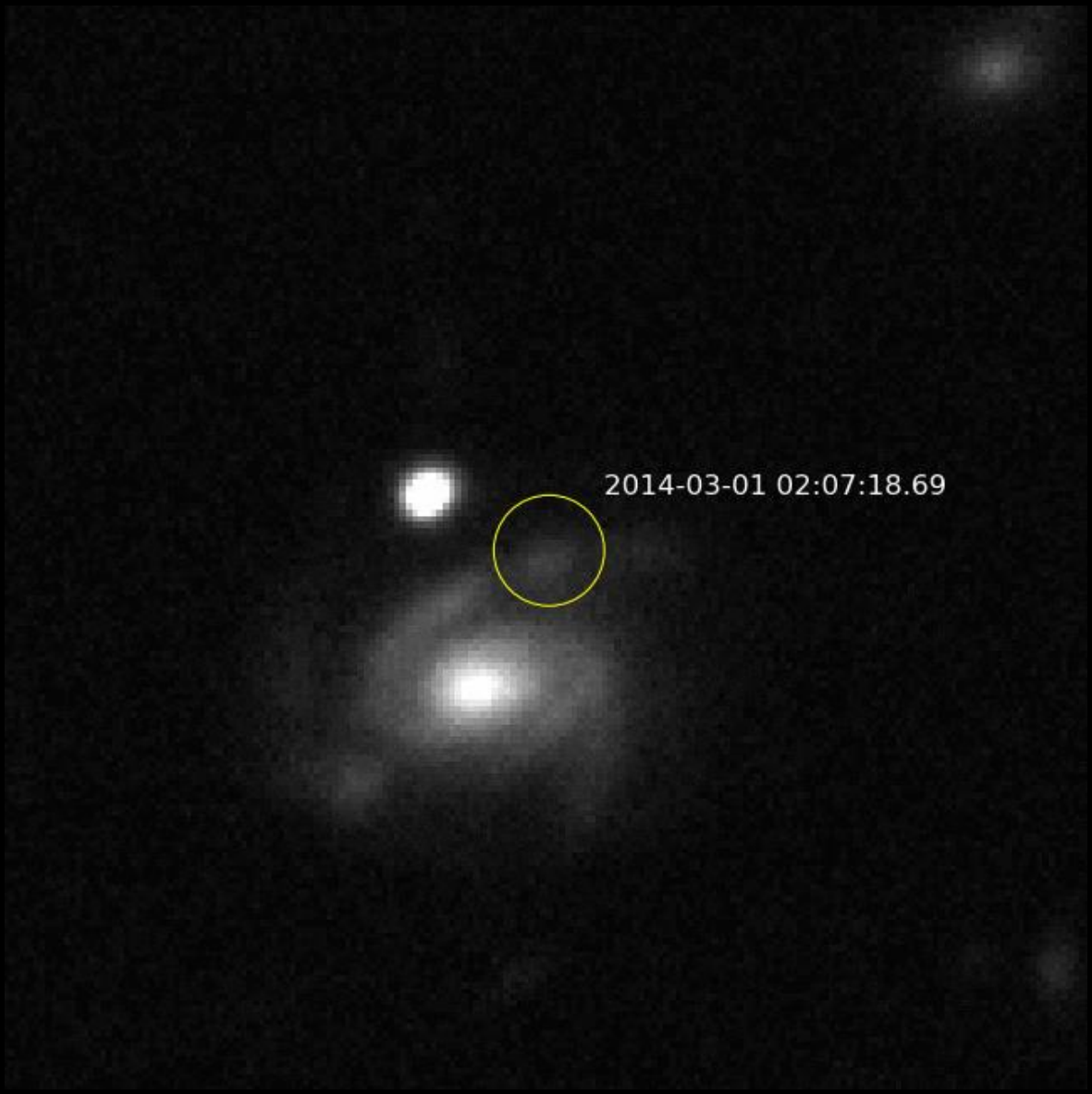


RF: random forests +feature engineering approach (56 handmade features)

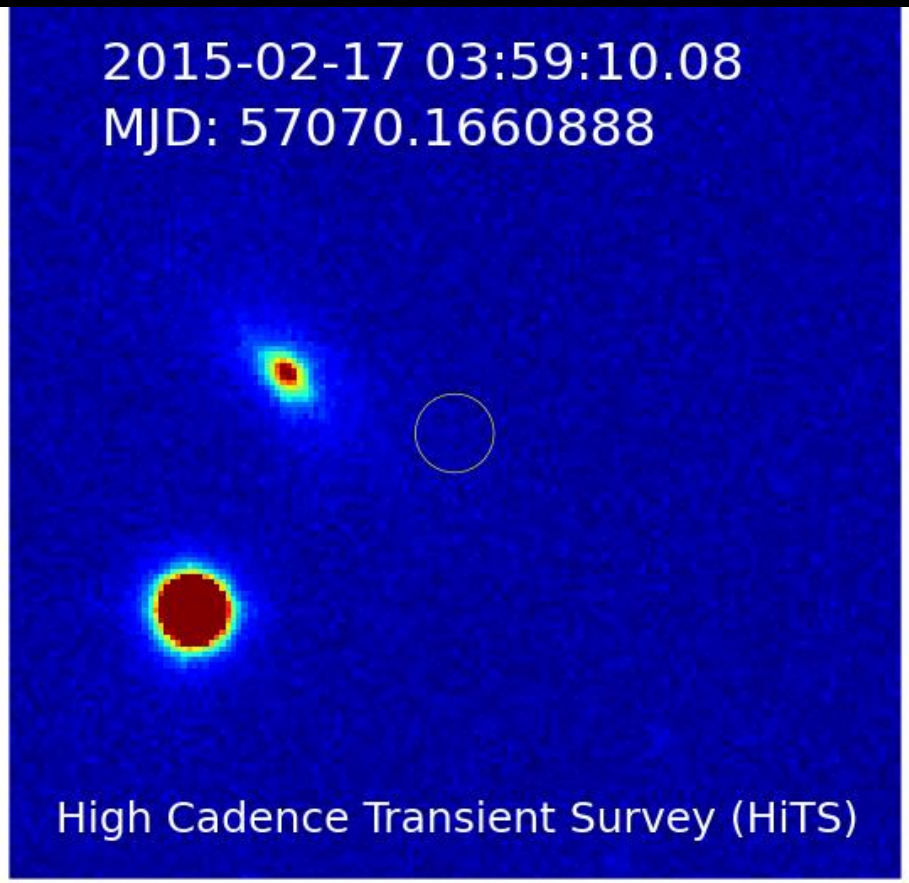
ConvNet-4: using all 4 images.

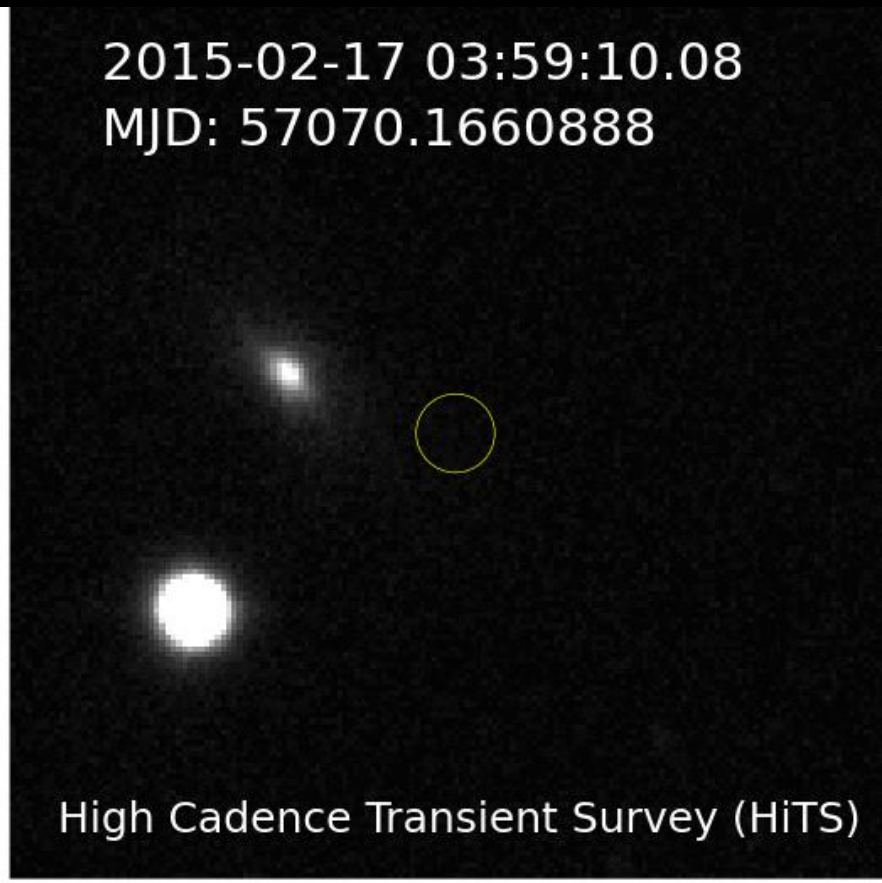
Model	Accuracy	Precision	Recall	F1-score
Deep-Hits (DH)	99.45±0.02	99.37±0.04	99.55±0.06	99.45±0.02
Deep-Hits + ReLU	99.47±0.02	99.44±0.04	99.52±0.06	99.47±0.02
Cyclic Avg. Pool (CAP)	99.52±0.01	99.45±0.01	99.61±0.02	99.52±0.01
Cyclic Avg. Pool Ensemble (CAPE)	**99.53±0.01**	**99.45±0.02**	**99.63±0.03**	**99.53±0.01**

Movies!



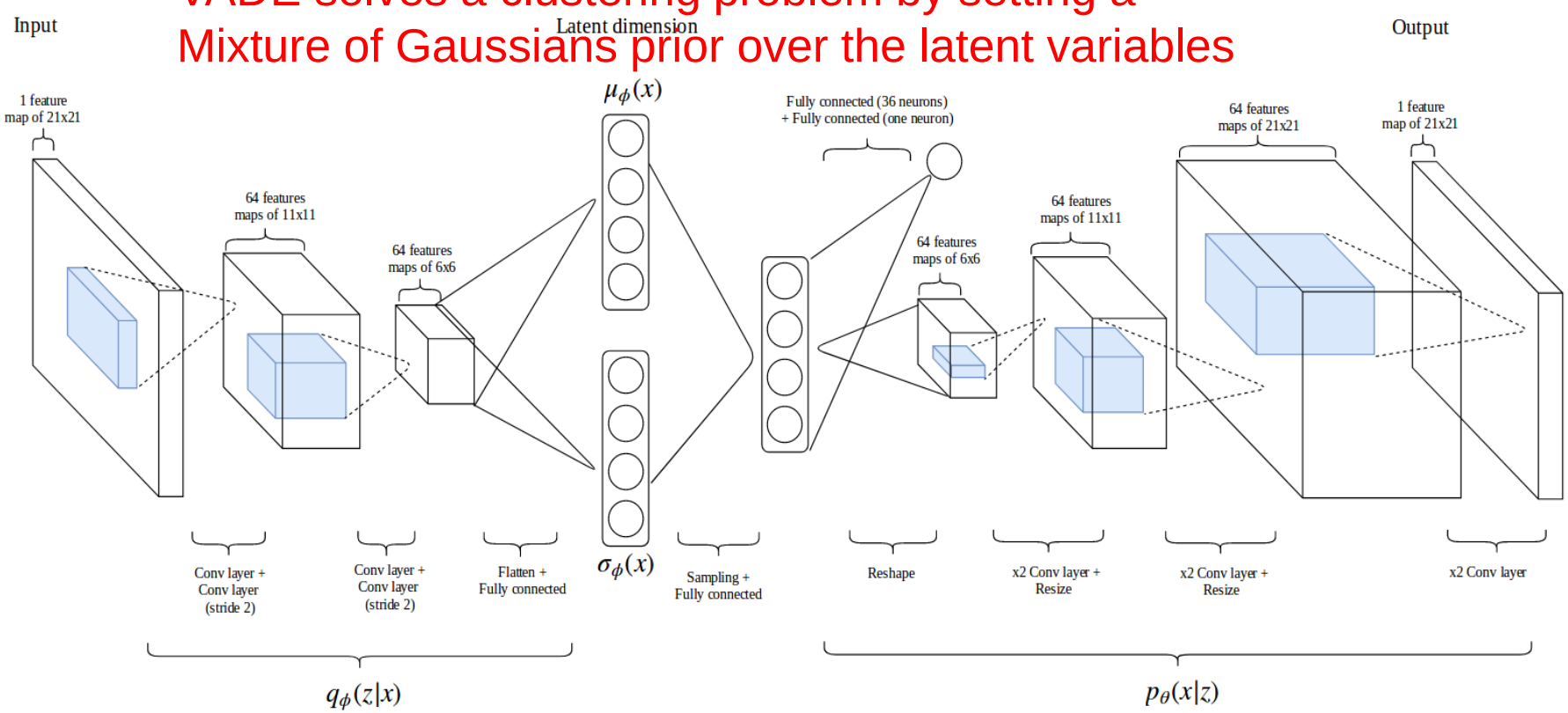
http://www.das.uchile.cl/~fforster/ATEL/summary_das.html


2015-02-17 03:59:10.08
MJD: 57070.1660888
High Cadence Transient Survey (HiTS)


2015-02-17 03:59:10.08
MJD: 57070.1660888
High Cadence Transient Survey (HiTS)

Deep Variational Embedding (VADE) Autoencoder: Unsupervised Approach

VADE solves a clustering problem by setting a Mixture of Gaussians prior over the latent variables



Original SNR diff image 21x21 pixels

Encoder: Encodes data X into latent variables

neurons: dimensionality of latent space

Decoder: generative model to reconstruct data X

Reconstructed SNR diff image 21x21 pixels

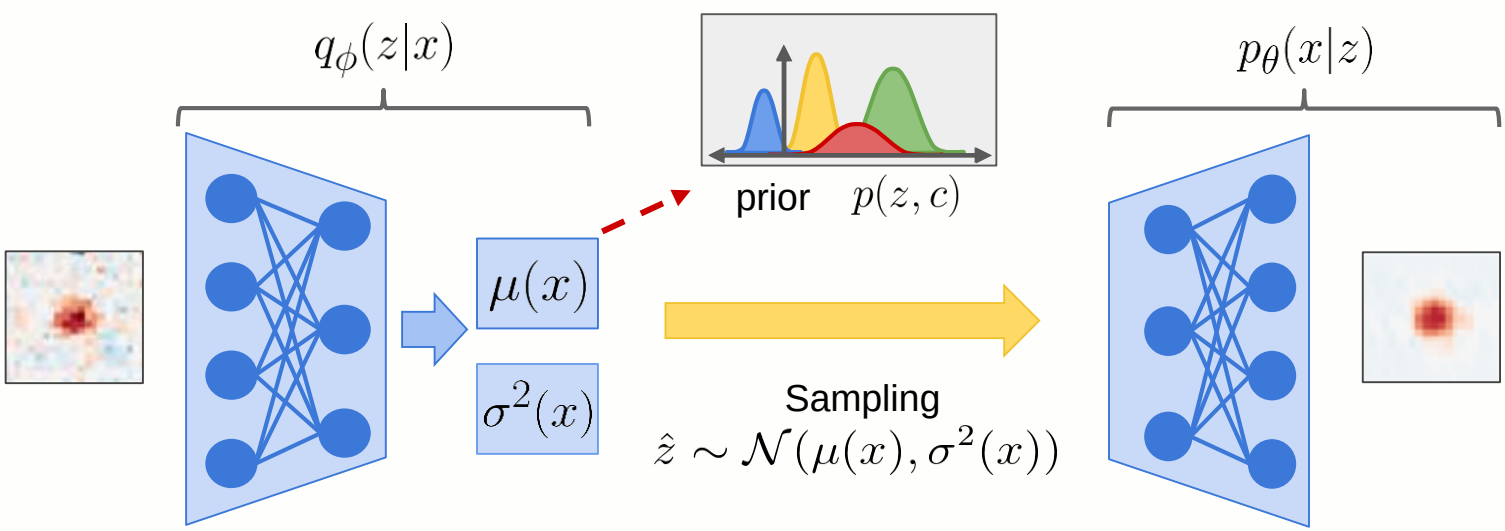
Deep Variational Embedding (VADE)

- Decoder is a generative model, latent variables are stochastic.
- Prior p(z, c) is a mixture of Gaussians.

$$\pi_c \mathcal{N}(\mu_c, \sigma_c^2)$$

$$\log p(x) \geq \text{ELBO}$$
$$\equiv \mathbb{E}_{q(z,c|x)} \left[\log p(x|z) + \log p(c) + \log p(z|c) - \log q(z|x) - \log q(c|x)\right]$$

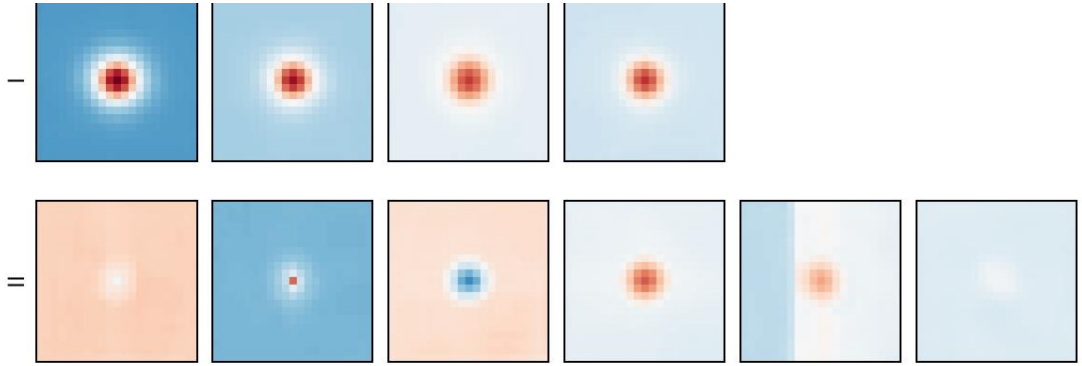
$q_\phi(z|x)$ Inference network
$p_\theta(x|z)$ Generative network



VADE Results on Real/Bogus

Method	Accuracy
Incremental PCA + GMM	93.43 ± 1.25 %
Online DL + GMM	90.79 ± 0.22 %
VAE + GMM	87.06 ± 0.69 %
VADE	95.33 ± 0.11 %

Cluster Centroids



Positive class

Negative class

Fig. 6. Reconstruction obtained by sampling from the VADE centroids μ_c. The first and second rows correspond to clusters associated to the positive class (stellar transient) and the negative class (artifacts), respectively.

[1] Astorga, Huijse, **Estévez, Förster**, "Clustering of Astronomical Transient Candidates using Deep Variational Embedding", IJCNN 2018
[2] Huijse, **Estévez, Forster,** Pignata, "Latent representations of transients from an astronomical image difference pipeline using VAE", ESANN 2018

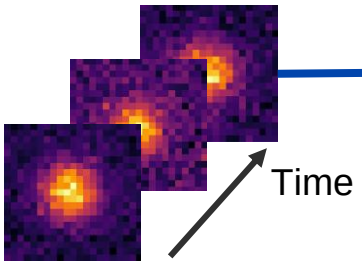
Deep Learning for Image Sequence Classification of Astronomical Events

- We propose a sequential classifier based on a recurrent convolutional neural network (RCNN)
- It uses sequences of images as inputs
- This approach avoids the computation of light curves or difference images
- A basic assumption is that there is more information in the original images than in the light curves, e.g. spatial information.
- We use synthetic datasets to train the models, and real data from the HiTS survey to test them.

Carrasco-Davis, R.; Cabrera-Vives, G.; Forster, F., Estevez, P.A., Huijse, P., Protopapas, P., Reyes, I., Martinez-Palomera, J., and Donoso, C. “Deep Learning for Image Sequence Classification of Astronomical Events”, accepted to PASP, 2018.

Basic Idea

Synthetic

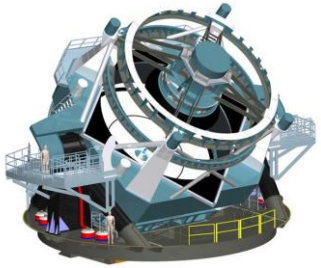


Train with synthetic data

Machine Learning Classifier Model

- Sequence of Images > Images > Light Curve
- Without differencing images

Real Data



After training
Predict class with real data

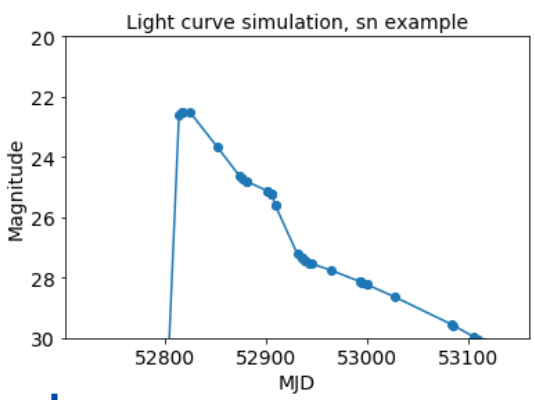
Machine Learning Classifier Model

List of supernovae candidates

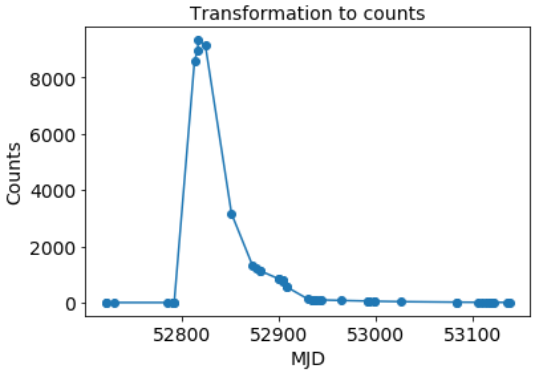
Synthetic Dataset

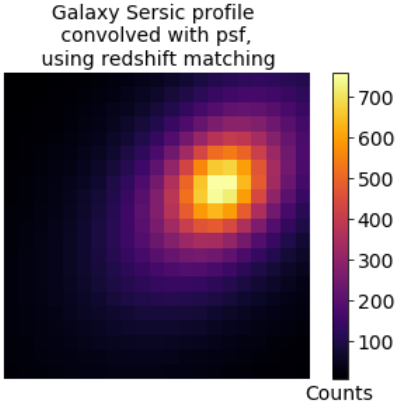
- First we simulate light-curves based on physical and empirical models, and sample them using the observation times.
- Each point in a light curve is transformed to an image by taking into account:
  - Instrument specifications
  - Exposure times
  - Atmospheric conditions
- A given PSF (point spread function) is assumed, which is sampled from a collection of empirical PSFs.

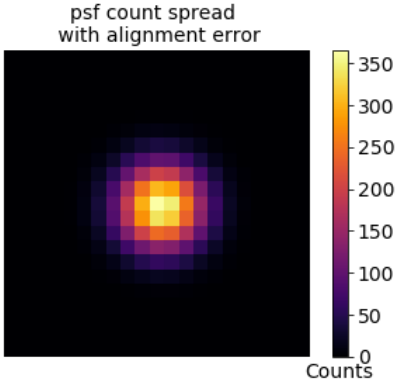
Example of Image Simulation

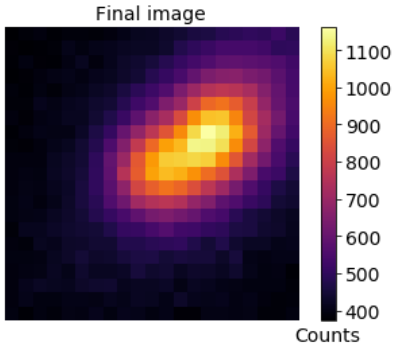


$$m = Z_{p,g} - 2.5\log_{10}(counts/t_{exp}) - k_{air}(airmass - 1)$$







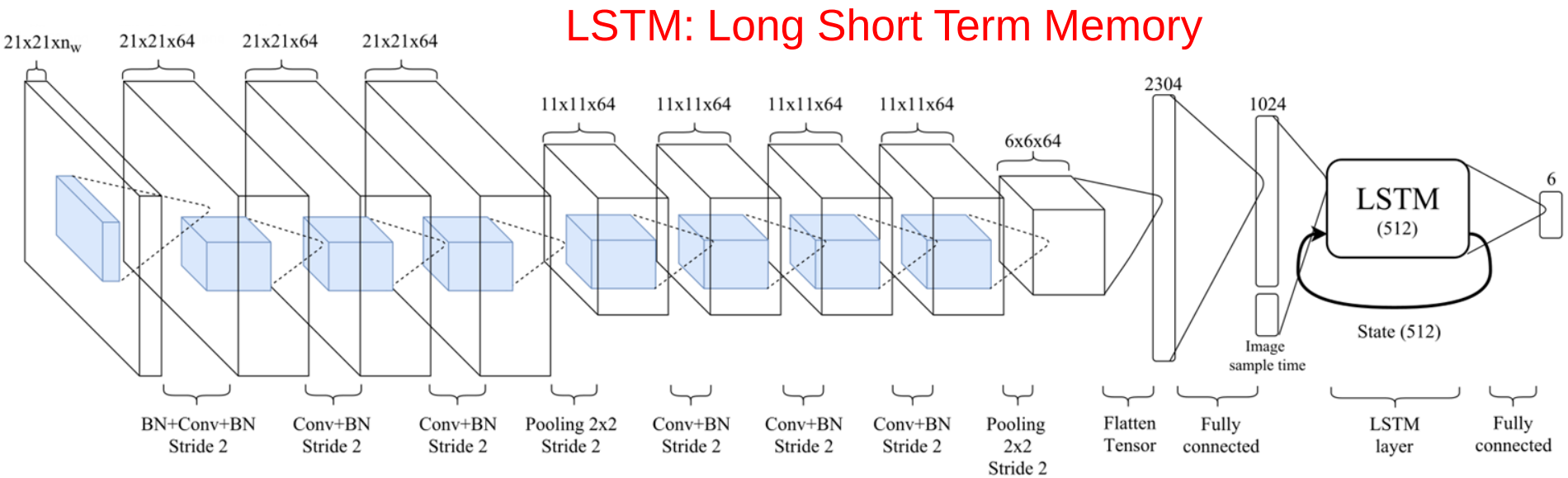


Sky brightness estimation and Poisson noise

Synthetic and Real Datasets

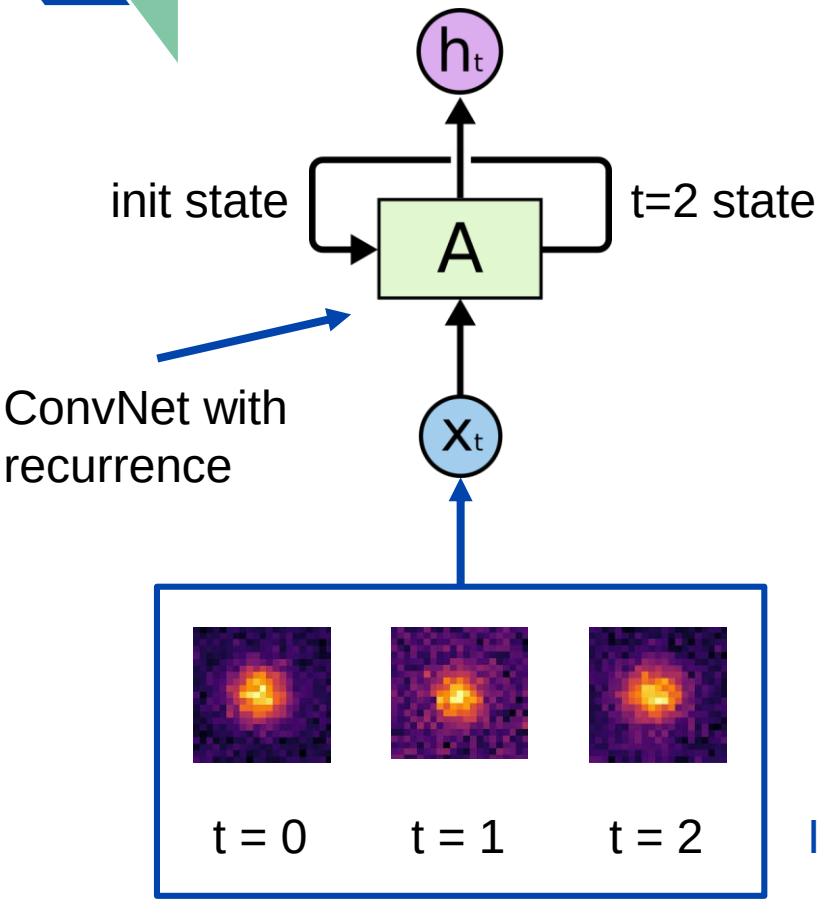
- We simulated 686,000 objects for the training set, 85,750 for the validation set and 85,750 for the test set.
- In this work, the observing conditions were sampled from real observations from the 2015 HiTS survey.
- Five classes are available for real image sequences: supernovae (73), RR Lyrae (111), eclipsing binaries (76), non-variables (255) and asteroids (51).

Recurrent Convolutional Neural Network

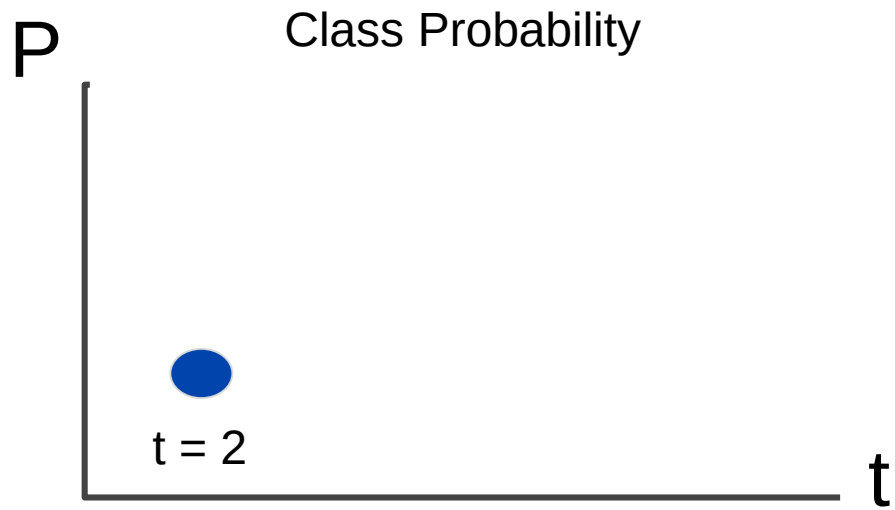


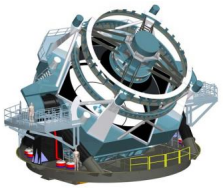
- Convolutional layers are able to automatically learn the spatial correlation among pixels in the input images and extract high-level features.
- A LSTM recurrent layer is used to learn time dependencies among images at irregular times.

Classification Process

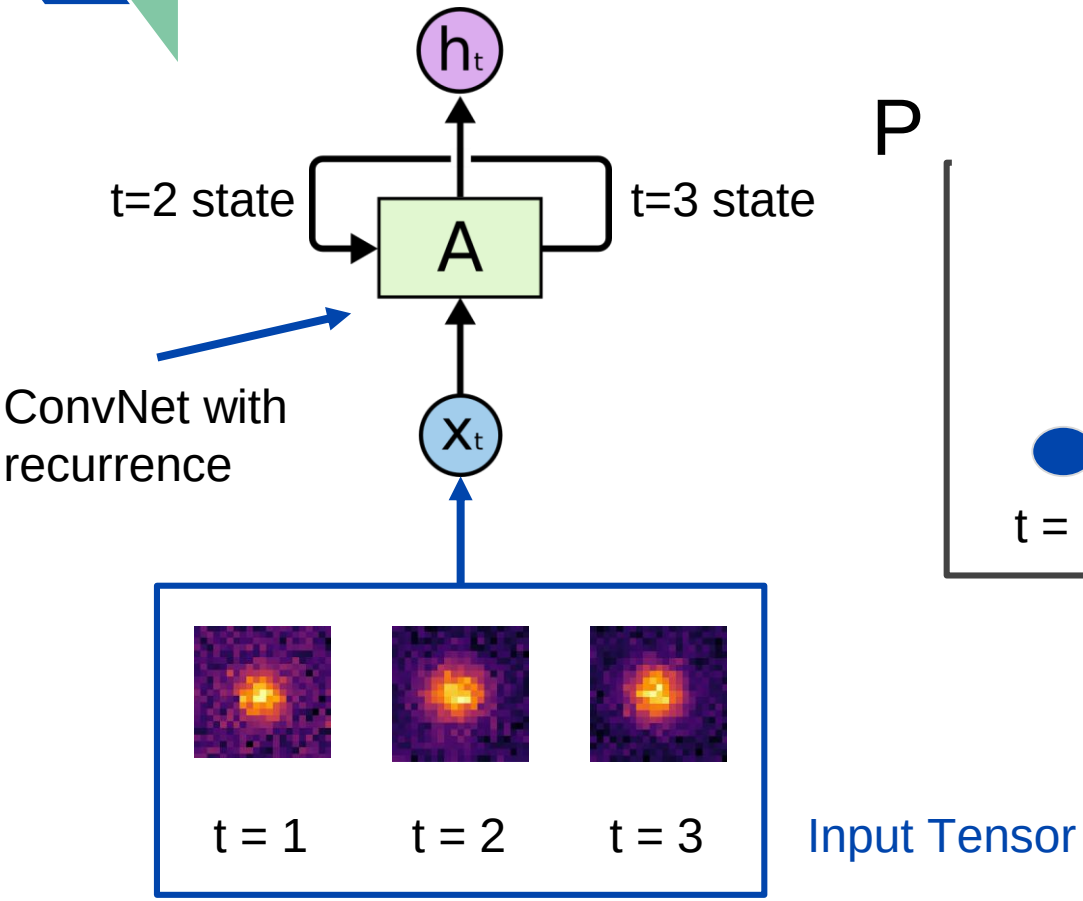


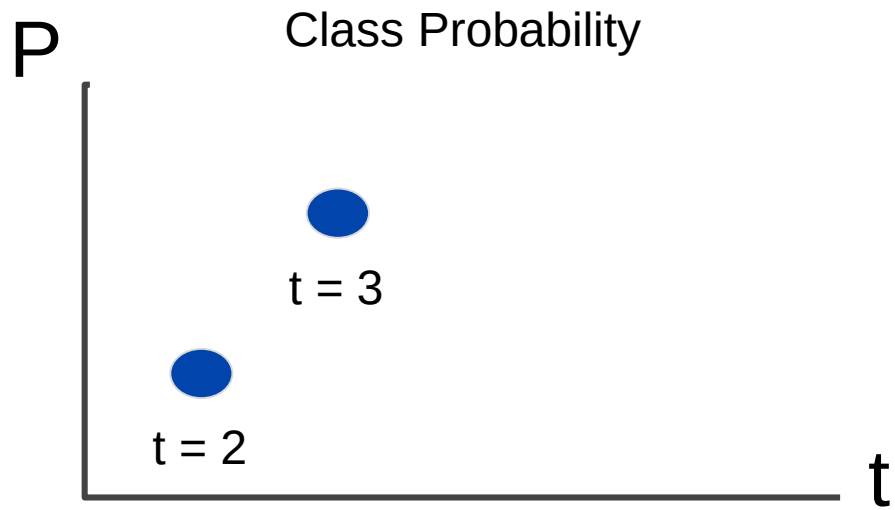
Input Tensor

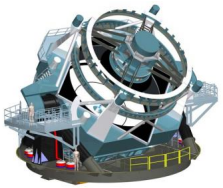




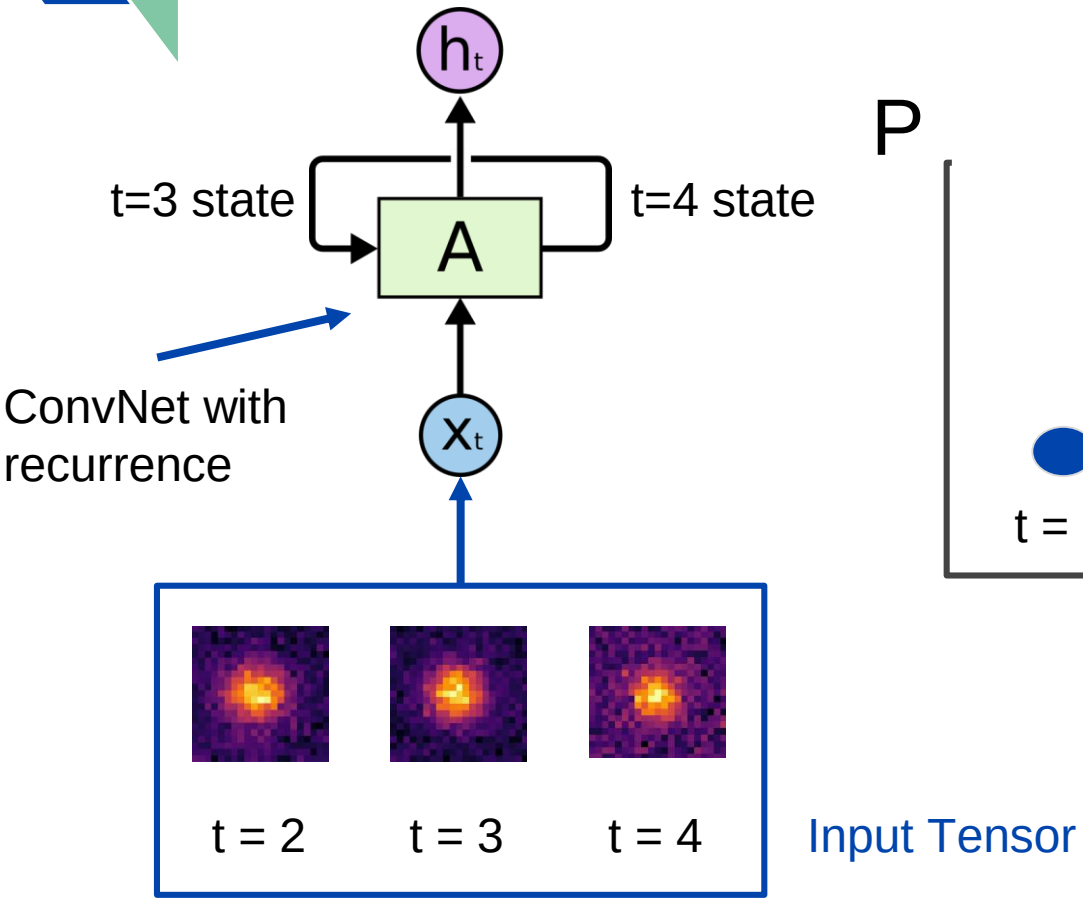
Classification Process

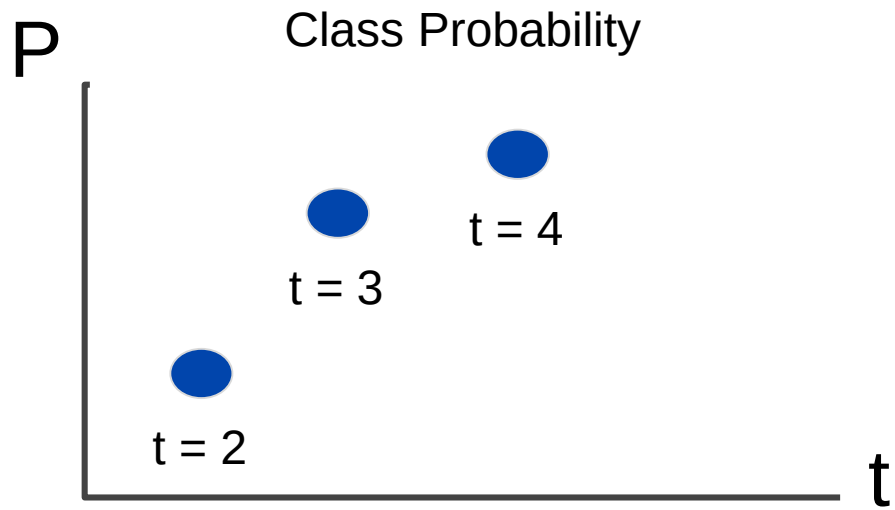


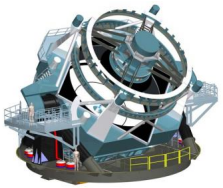




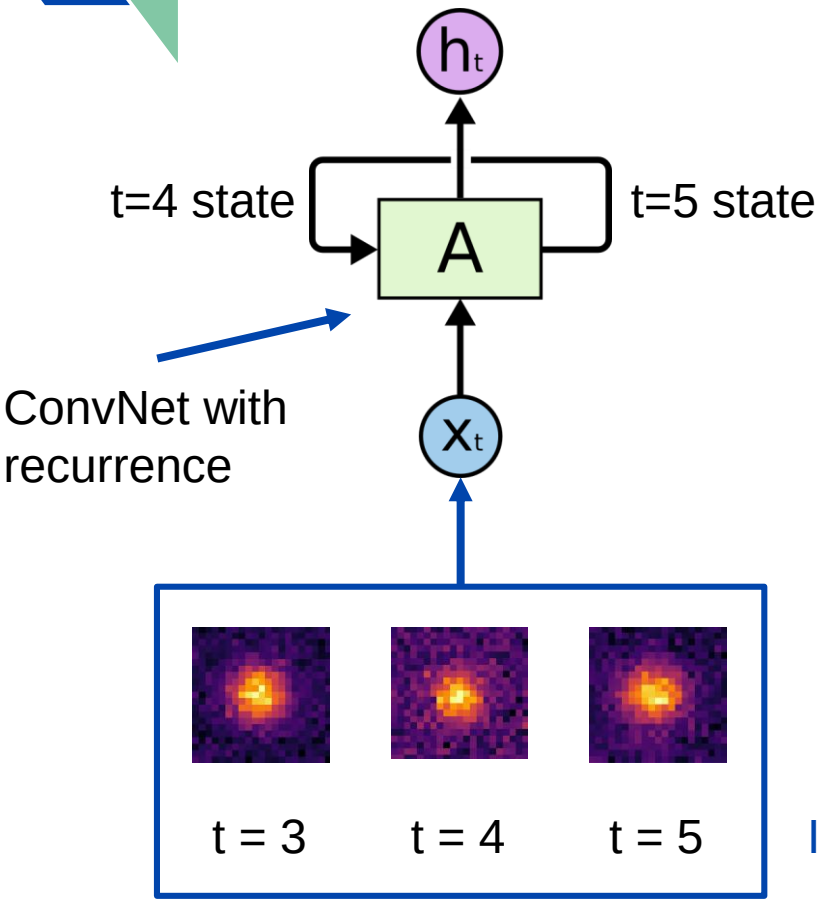
Classification Process



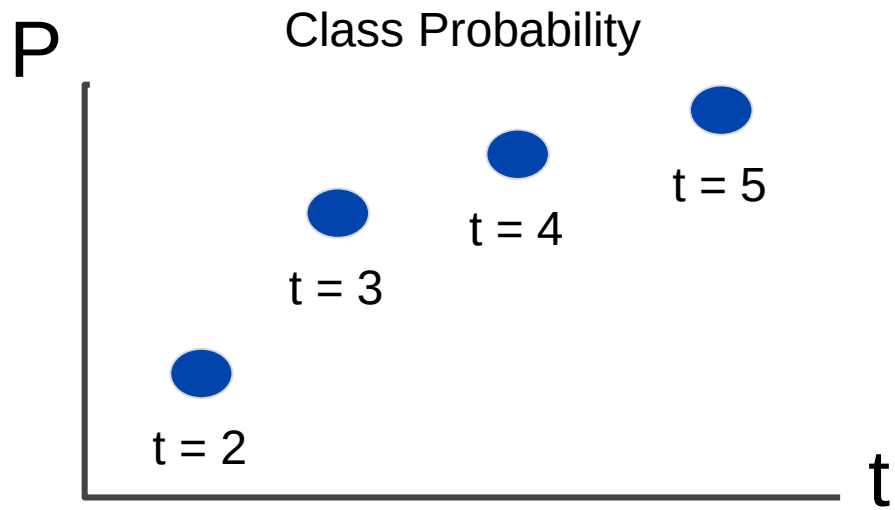


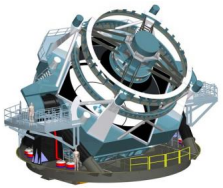


Classification Process



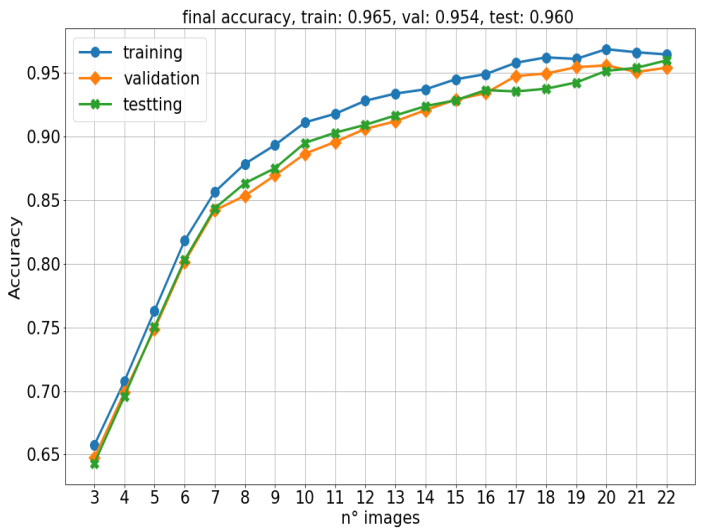
Input Tensor





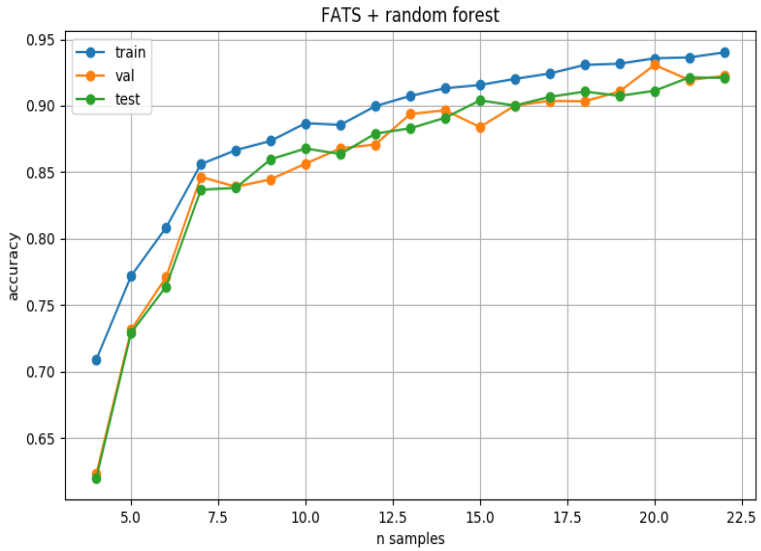
Model performance on synthetic data

- RCNN model



Our model uses sequences of images as inputs directly

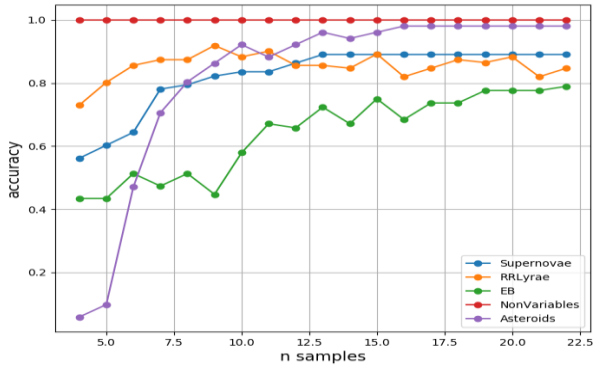
FATS (58 features) + RF (Random Forests)

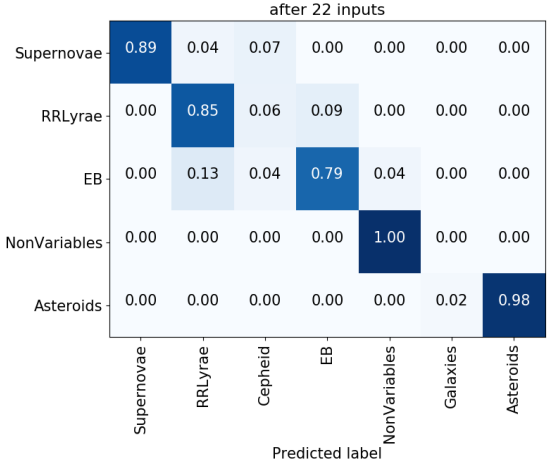


Light curves in count units were extracted from the simulated images using optimal photometry.

Model performance on real data: an example

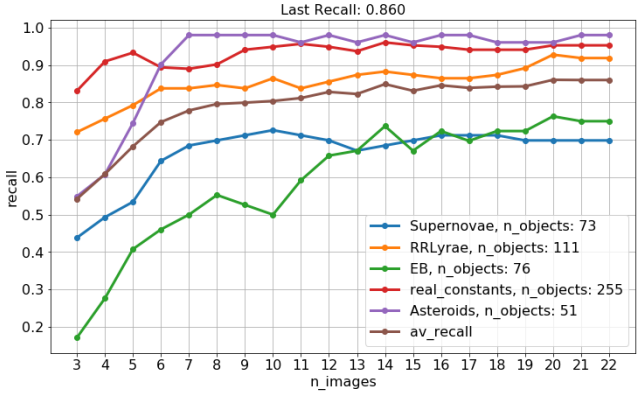
Random Forest

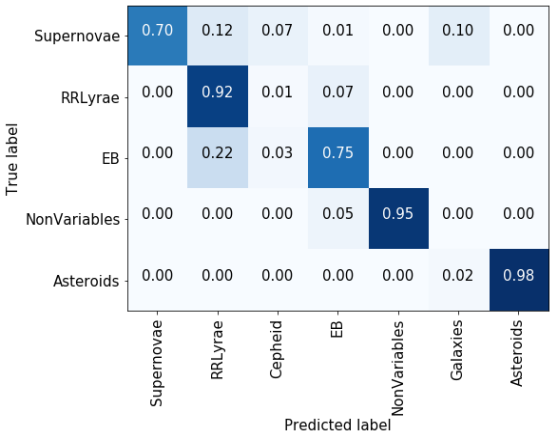




Recall: 90%

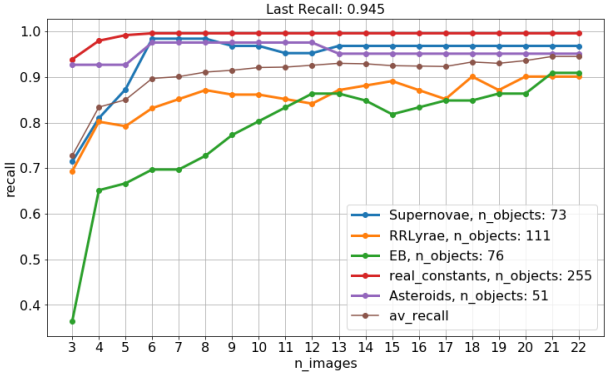
RCNN

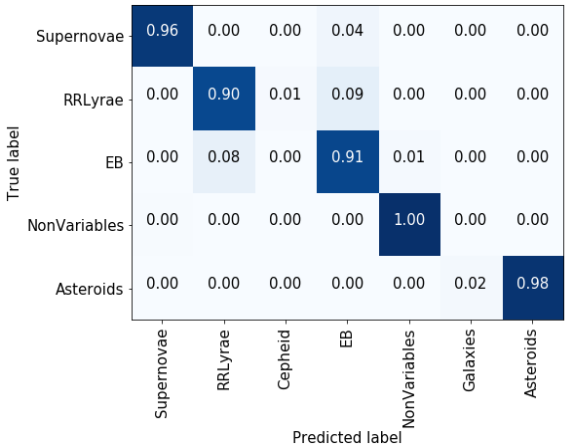




Recall: 86%

RCNN with fine tuning





Recall: 95%

Model evaluation on real data: average results

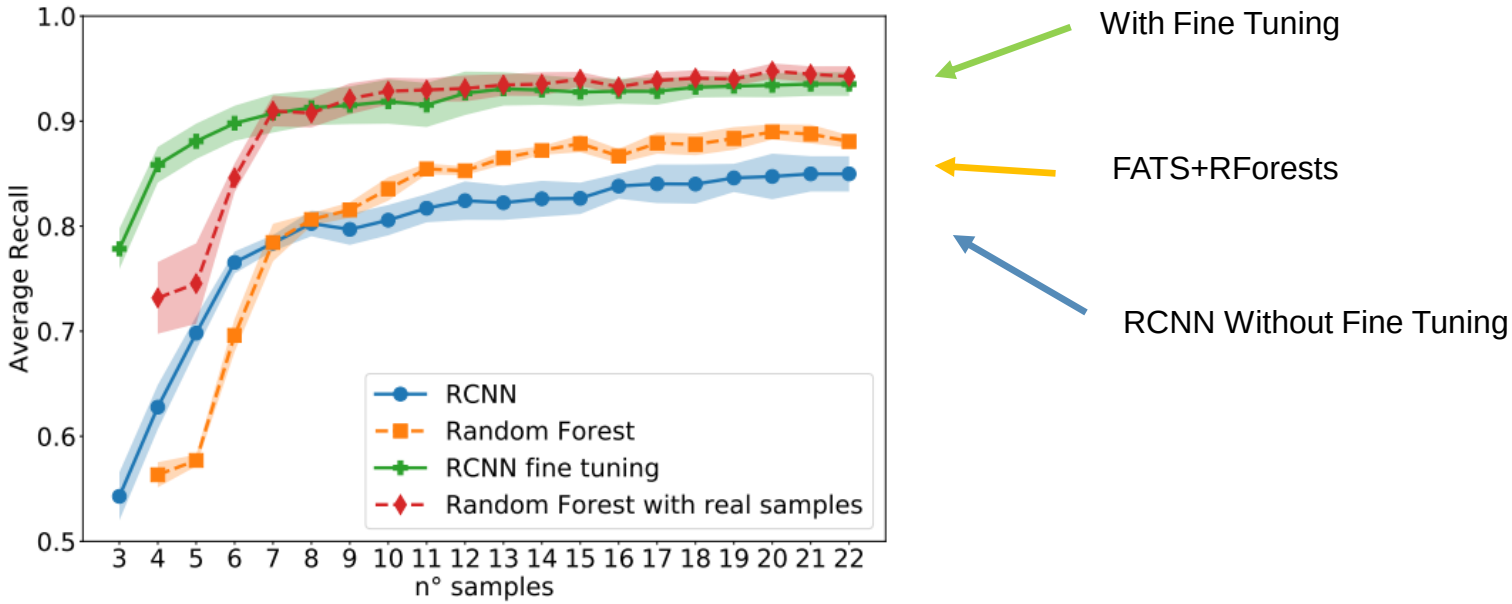
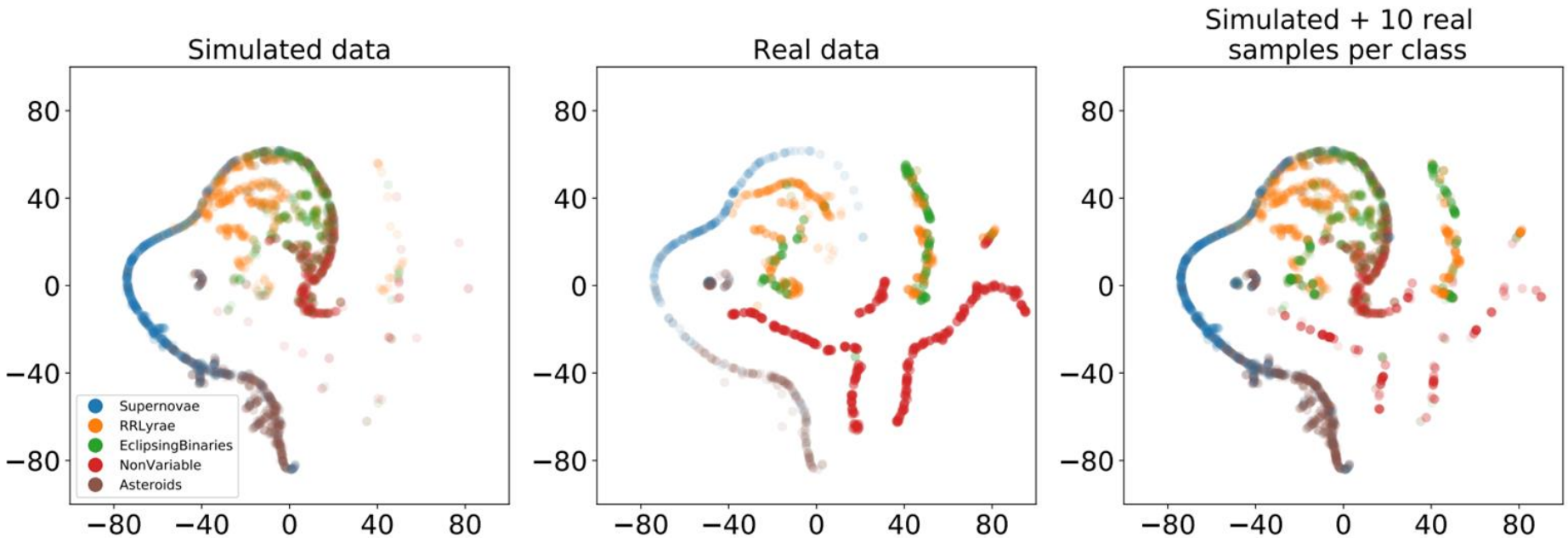


Figure 6. Average recall on the real dataset without and with fine tuning as a function of the number of samples of the sequence available to the classifiers. The average recall obtained at the last point (after 22 samples) for each curve is: RCNN: 0.85 ± 0.02, Random Forest: 0.88 ± 0.01, RCNN with fine tuning: 0.94 ± 0.01 and Random Forest trained with augmented training set: 0.94 ± 0.01. For the last to cases, the evaluation set does not include samples used to fine tune the model.

Why fine tuning does work?

- t-SNE projection from 6 most important features used by random forest to 2D



Conclusions

- The proposed RCNN model uses sequences of images. Scales well since it has a constant computational cost.
- Domain adaptation is an important area of research. Our model gets a very high performance on the real dataset after fine tuning. Just after presenting a few real samples.
- The proposed approach allows us to generate datasets to train and test our RCNN model for different astronomical surveys and telescopes.

Conclusions

Big Data is here to stay

Deep learning is a computational intelligence/machine learning technique that allows us to extract features automatically

The importance of multidisciplinary teams

To do list: My anti-library keeps growing!,
Larger sets and input sequences of images to Convnets, semi-supervised learning

Please come to Chile!

Any Questions?

Here
We love
Computational Intelligence
to death!

References

- Carrasco-Davis, R.; Cabrera-Vives, G.; Forster, F., Estevez, P.A., Huijse, P., Protopapas, P., Reyes, I., Martinez-Palomera, J., and Donoso, C. “Deep Learning for Image Sequence Classification of Astronomical Events”, accepted to PASP, 2018.
- Guillermo Cabrera-Vives, Ignacio Reyes, Francisco Forster, Pablo Estevez, Juan C. Maureira, “Supernovae Detection by Using Convolutional Neural Networks”, IJCNN 2016.
- Cabrera-Vives, G., Reyes, I., Forster, F., Estevez, P.A., Maureira, J.C., “Deep-HiTS: Rotation Invariant Convolutional Neural Network for Transient Detection”, Astrophysical Journal, Feb. 2017.
- Esteban Reyes, Pablo Estevez, Ignacio Reyes, Guillermo Cabrera-Vives, Pablo Huijse, Rodrigo Carrasco and Francisco Forster, “Enhanced Rotational Invariant Convolutional Neural Network for Supernovae Detection”, IJCNN 2018.

References

- Astorga, Huijse, Estévez, Förster, “Clustering of Astronomical Transient Candidates using Deep Variational Embedding”, IJCNN 2018
- Huijse, Estévez, Forster, Pignata, “Latent representations of transients from an astronomical image difference pipeline using VAE”, ESANN 2018
- Huijse, P., Estevez. P.A., Forster, F., Daniel, S., Connolly A., Protopapas, P., Carrasco, R., Principe, J., “Robust Period Estimation Using Mutual Information for Multi-band Light Curves in the Synoptic Survey Era”, Astrophysical Journal Supplement Series, 2018.
- P. Huijse, P. Estevez, P. Protopapas, JC Principe, P. Zegers, “Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases”, IEEE Computational Intelligence Magazine, August 2014.