Light curve classification with Deep Learning

Johanna Pasquet

Centre de Physique des Particules de Marseille

TransiXplore - Workshop

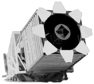
23 November, 2018

The era of Big Data

1924 Henry Drapper Catalog (0.2 Million)

1989 Guide Star Catalog (20 Million)

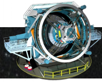
2008 SDSS (230 Million)



2018 Dark Energy Survey (400 Million)

2027 Euclid (10 billion)

2032 Large Synoptic Survey Telescope (37 billion)

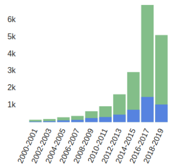


The emergence of artificial intelligence

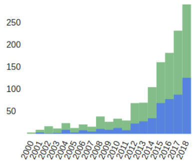
QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

machine learning year:2000-2019

General + physics + Astronomy



Astronomy

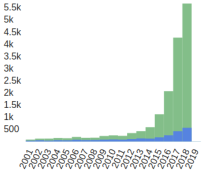


refereed
non refereed

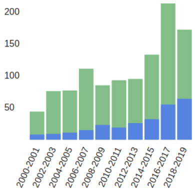
QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

deep learning year:2000-2019

General + physics + Astronomy

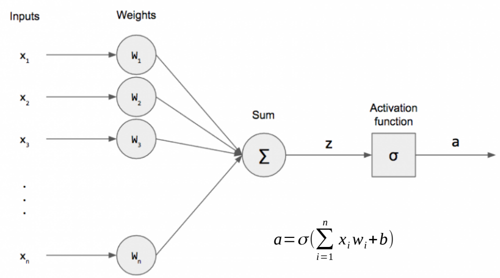


Astronomy



History

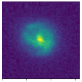
1957 Perceptron (Rosenblatt)
1986 MLP (Rumelhart et al.)
1998 LeNet (LeCun et al.)
2012 A CNN won ImageNet (Alexnet, Krizhevsky et al.)



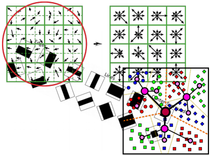
The main property of deep learning

Classical methods

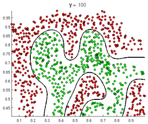
Input data



Feature crafting

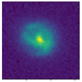


Separation with a classifier

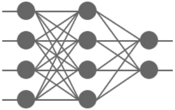


Deep learning

Input data

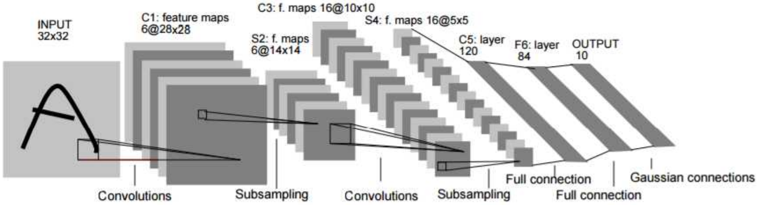


Feature learning



→ The best feature space representation is found by the network

LeNet5



Lecun et al. 1998

3 operations:

- Convolution + non linearity (feature extraction)
- Pooling
- Fully Connected (classification)

Convolutions

An image

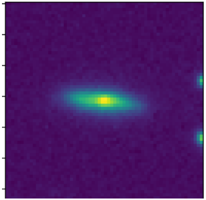
1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

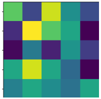
A kernel

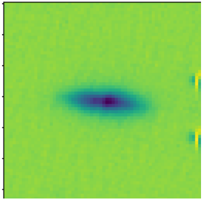
1	1	1
0	1	1
0	0	1

A convolved image

4	3	4
2	4	3
2	3	4

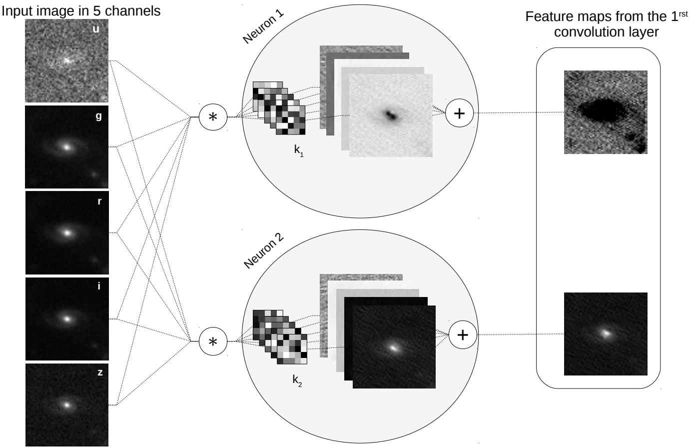






Convolution operation is followed by a non linear function (tanh, ReLu...)

Convolutions



Pooling

A feature map

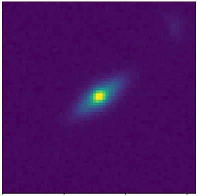
5	1	3	0
0	1	2	7
2	1	1	4
3	1	1	2

Pooling operation

Max in a 2x2 sliding window with a stride of 2

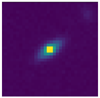
A subsampled feature map

5	7
3	4



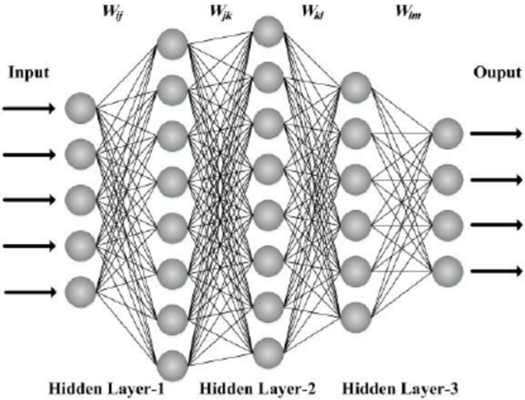
64x64

Max in a 2x2 sliding window with a stride of 2



32x32

Fully connected



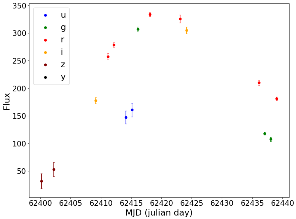
Difficulties for the classification

Many factors degrade the performance of machine learning algorithms:

Small training databases

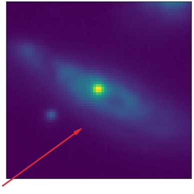
Data can be sparse with an irregular sampling

Non-representativeness between the training and the test databases



The spectroscopic follow-up

Identify and measure the redshift of a galaxy



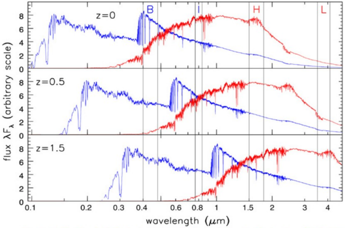
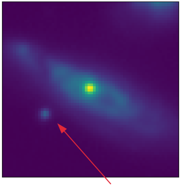
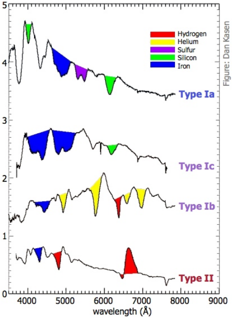


Fig 8.12 (S. Charlot) 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

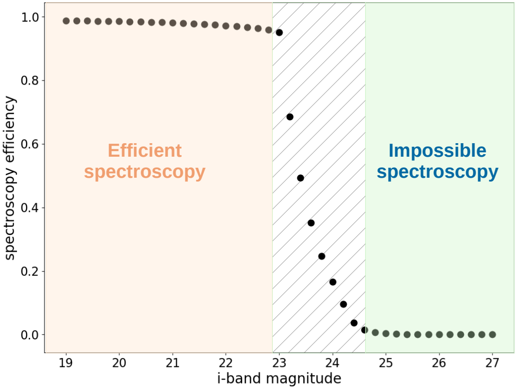
Determine the nature of an observed object



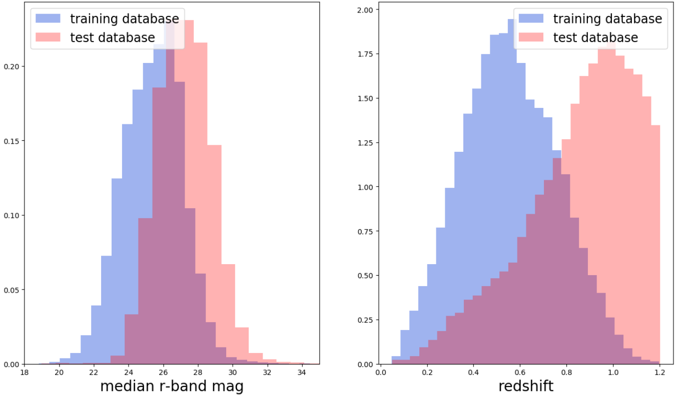


Limitation of the spectroscopic follow-up

Observation with an hypothetic 8 m class telescope with a limiting i-band magnitude of 23.5

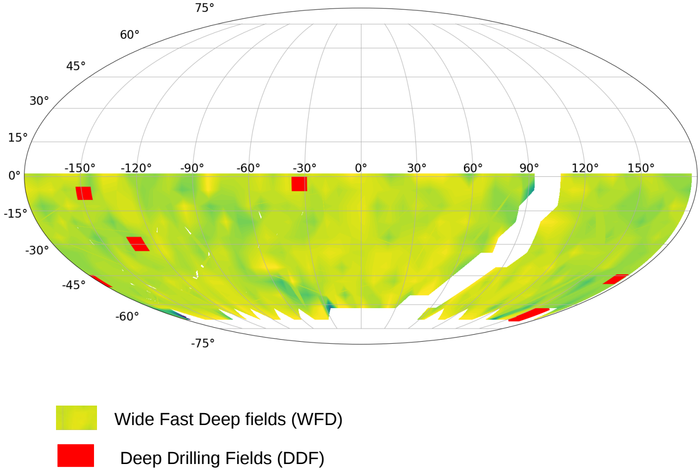


Non-representativeness between the training and test databases



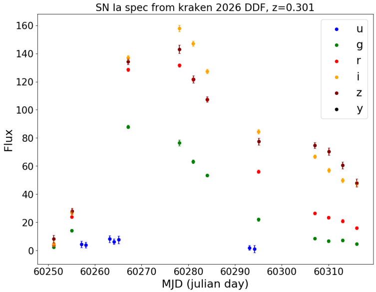
The non-representativeness of the databases, which is a problem of mismatch, is critical for machine learning process.

The main survey and the deep fields of LSST

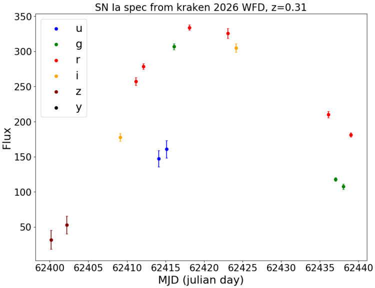


Comparison of light curves

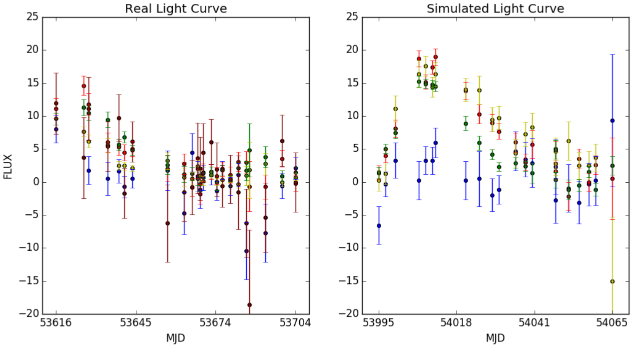
DDF light curve



WFD light curve



A training on simulated data and a testing on real data



Analyze the behaviour of the deep architecture

- Control the behaviour of the model with physical parameters (e.g. EBV, redshift...)
- Control the behaviour of the model with observational conditions (e.g. SNR, cadence,magnitudes)
- Understand the limit of the model in redshift, magnitude...

$\Rightarrow$ Be able to use the ouput probabilities in a confident interval

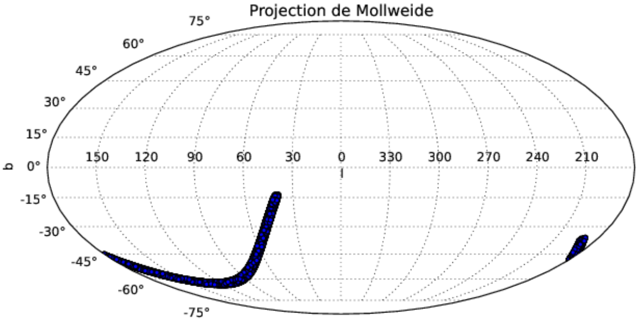
The classification of light curves of quasars

Johanna Pasquet and Jérôme Pasquet

A&A 611, A97 (2018)

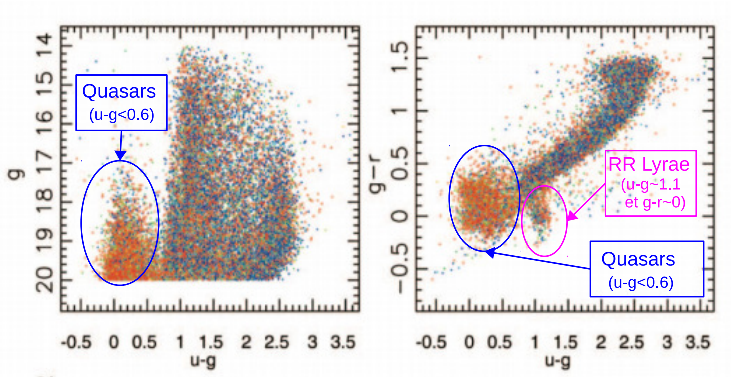
The data

- a 2.5 degree wide stripe along the Celestial Equator in the Southern Galactic Cap
- Coordinates : -60° $\leq \alpha \leq$ +60° et $-1.26° \leq \delta \leq 1.26°$,
- Observations in five bands (u, g, r, i et z)during nine years,
- In 2007 : catalog of 67 507 transients coming from Stripe 82 (Ivezić et al.)

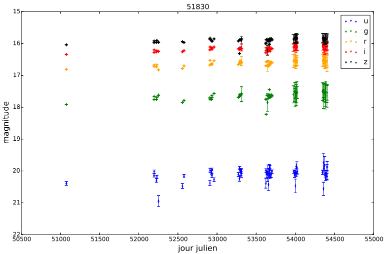


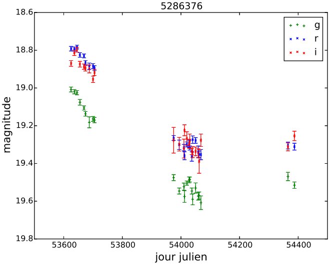
Known objects

$\rightarrow$ quasars ($\sim$ 8000 quasars, variation time scale de variation from day to year), RR Lyrae and δ Scuti ($\sim$ 500 , $0.1 \leq T \leq 1$ day)...

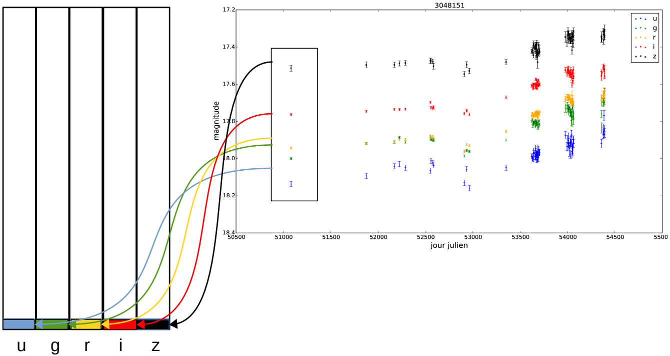


Example of light curves

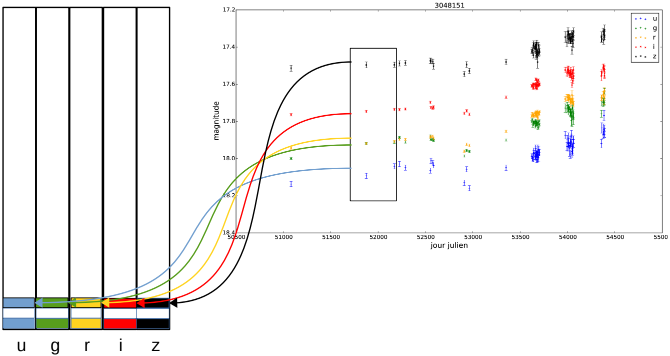




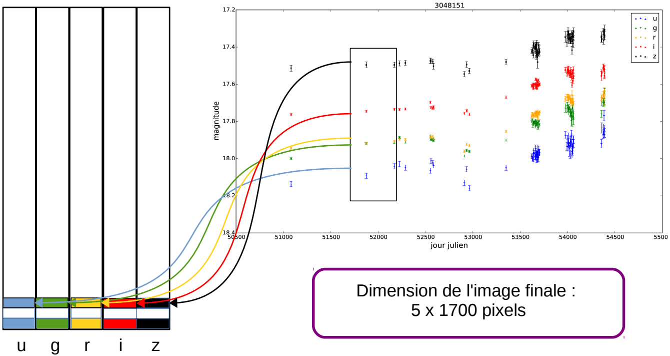
Light Curve Images (LCI)



Light Curve Images (LCI)

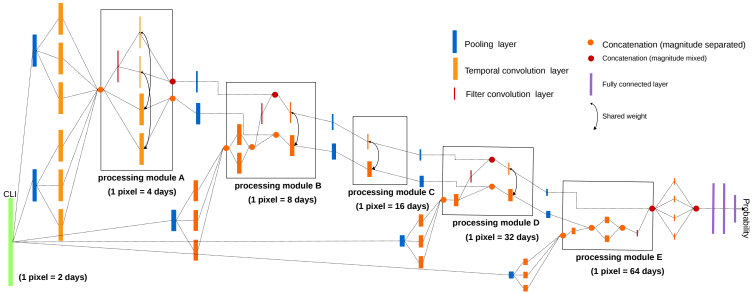


Light Curve Images (LCI)



Dimension de l'image finale :
5 x 1700 pixels

The architecture



Results

	Detectability threshod	Recall	Precision
Random Forest (+ FATS)	0.816	0.900	0.986
	0.684	0.950	0.978
	0.56	0.970	0.974
CNN	0.993	0.900	0.986
	0.739	0.950	0.974
	0.190	0.970	0.956

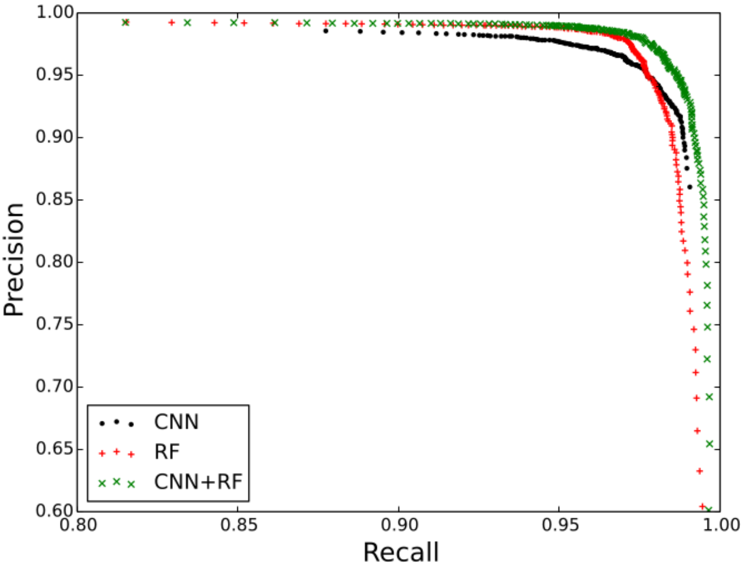
- Precision (P)

$$P = \frac{TP}{TP + FP}$$

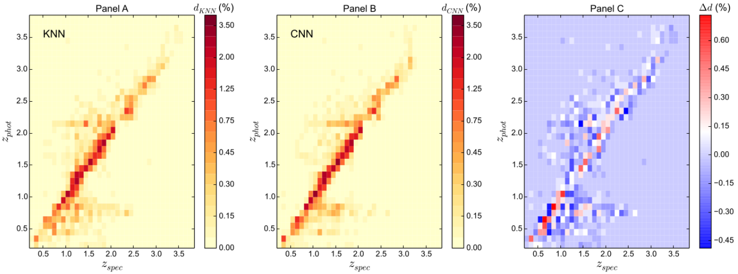
- Recall (R)

$$R = \frac{TP}{TP + FN}$$

Combination of the two classifiers



Estimation of photometric redshifts



The classification of light curves of supernovae (SN Ia/ SN Non-Ia)

Johanna Pasquet, Jérôme Pasquet, Marc Chaumont and Dominique Fouchez

PELICAN: a deeP architecturE for the LIght Curve ANalysis
(**Johanna Pasquet**, Jérôme Pasquet, Marc Chaumont and Dominique Fouchez, just submitted)

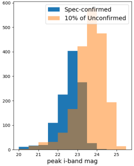
Key elements :

1. a complex Deep Learning architecture to classify light curves of supernovae
2. trained on a small and biased training database
3. overcome the problem of non-representativeness between the training and the test databases
4. deal with the sparsity of data and the difference of sampling and noise

The ability of PELICAN to deal with the different causes of non-representativeness between the training and test databases, and its robustness against survey properties and observational conditions, put it on the forefront of the light curves classification tools for the LSST era.

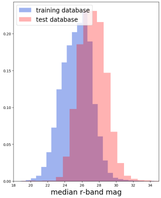
Different databases

1. The Supernova Photometric Classification Challenge in 2010 (SPCC, Kessler et al.)



- Small training database (1,103 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up

2. LSST simulated data

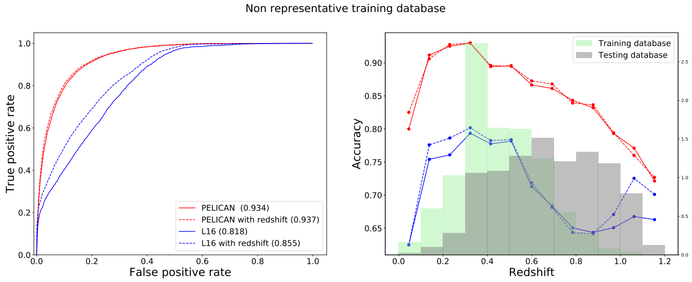


- Small training database (until 500 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up
- Non-representativeness of the sampling and noise between main survey and deep fields

3. SDSS-II Supernova Survey Data (Frieman et al. 2008; Sako et al. 2008)

- Non-representativeness between the training (simulated data) and the test databases (real data)

The SPCC challenge



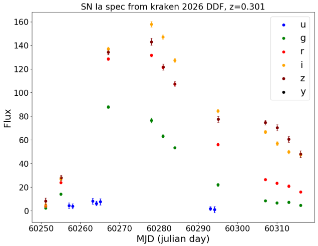
- We compared our results to one of the best current supernova classifier from Lochner et al. (2016, noted L16 hereafter) whose code and features used are available.
- PELICAN obtains an accuracy of 0.856 and an AUC of 0.934 which outperforms L16 method which reaches 0.705 and 0.818

LSST simulated data

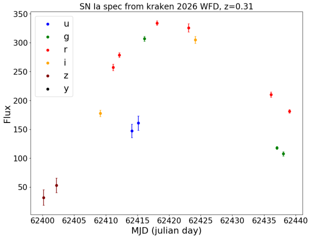
Two methodologies:

1. A training and a test on deep fields (DDF)
2. A training on deep fields and a test on the main survey (WFD)

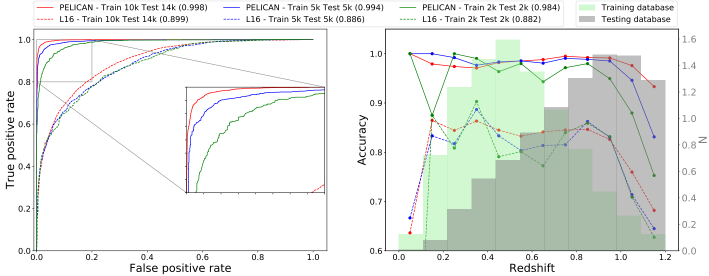
DDF light curve



WFD light curve

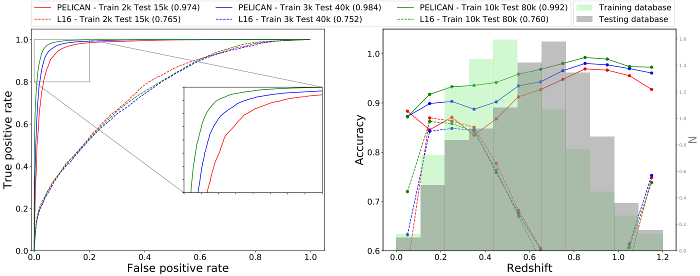


Results on DDF



	Training database (spec only)	Test database (phot only)	Accuracy	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} > 0.95$	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} > 0.98$	AUC
DDF	500	1,500	0.849 (0.746)	0.617 (0.309)	0.479 (0.162)	0.937 (0.848)
	2,000	2,000	0.925 (0.783)	0.895 (0.482)	0.818 (0.299)	0.984 (0.882)
	2,000	**22,000**	**0.934 (0.793)**	**0.926 (0.436)**	**0.851 (0.187)**	**0.986 (0.880)**
	10,000	14,000	0.979 (0.888)	0.992 (0.456)	0.978 (0.261)	0.998 (0.899)

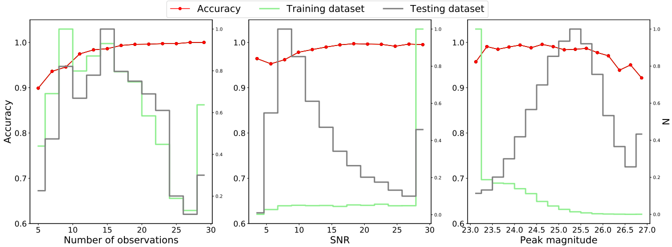
Results on WFD



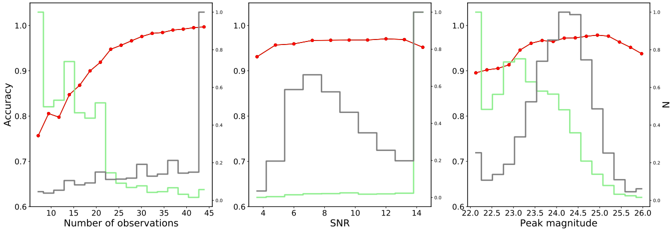
	Training database (spec only)	Test database (phot only)	Accuracy	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} > 0.95$	$\text{Recall}_{\text{Ia}}$ $\text{Precision}_{\text{Ia}} > 0.98$	AUC
WFD	DDF Spec : 2, 000	WFD : 15, 000	0.917 (0.650)	0.857 (0.066)	0.485 (0.000)	0.974 (0.765)
	DDF Spec : 3, 000	**WFD : 40, 000**	**0.940 (0.650)**	**0.939 (0.111)**	**0.729 (0.000)**	**0.984 (0.752)**
	DDF Spec : 10, 000	WFD : 80, 000	0.962 (0.651)	0.977 (0.121)	0.889 (0.010)	0.992 (0.760)

Further analysis of the behaviour of PELICAN

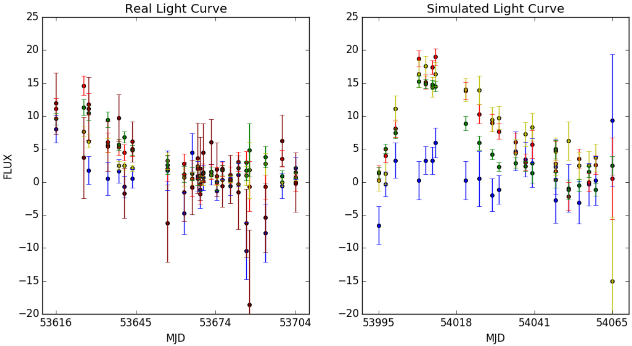
DDF



WFD



SDSS data

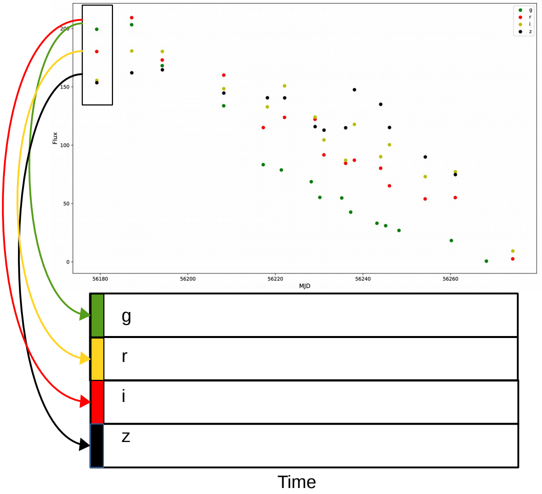


Training database	test database	Accuracy	AUC
SDSS simulations : 219,362	SDSS-II SN confirmed : 582	0.462	0.722
SDSS simulations : 219,362 SDSS-II SN confirmed : 80	SDSS-II SN confirmed : 582	0.868	0.850

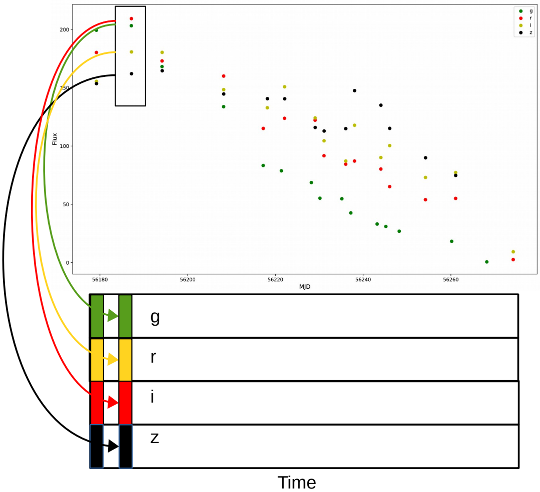
Summary

- The automatic classification of light curves has become a necessity in the context of the future large photometric surveys
- The problem of representativeness correponds to the real scenario and classification algorithms have to deal with it
- PELICAN brings a solution to different kind of non representativeness thanks to a dedicated architecture for the classification of supernovae light curves
- Perspectives for PLAsTiCC: the comptetition includes additional difficulties:
  - difference of fluxes
  - multiclass
  - class 99
  - asymetric data
- Therefore PELICAN has to be adapted to be applicable to the challenge: we are working on :)

The Light Curve Image (LCI)



The Light Curve Image (LCI)



⚠ Overfitting of missing data (zero values)

Impact of Signal-to-Noise Ratio (SNR) on widths of PDFs

The Stripe 82 region, which combines repeated observations of the same part of the sky, gives us the opportunity to look into the impact of SNR