General Introduction

Issues for the classification
- The problem of representativeness

Classification of light curves
- Architecture and data
- Results
  - SPCC
  - LSST
  - SDSS

Conclusion

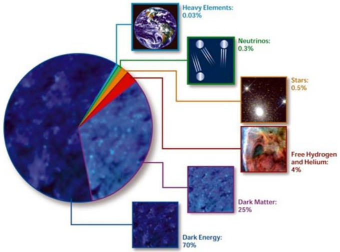
deeP architecturE for the LIght Curve ANalysis

Johanna Pasquet (CPPM)

LSST France

8 November, 2018

Current cosmology questions

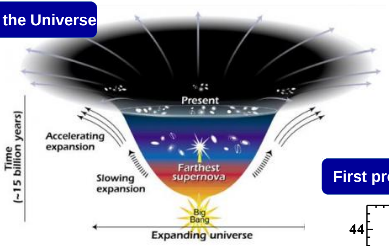


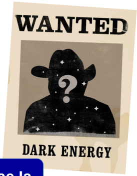
Credit : NASA

- What is the nature of dark matter ?
- What is the nature of dark energy ?
- Is it "dark energy" arising from quantum fluctuations in the vacuum, or is it new gravitational physics ?

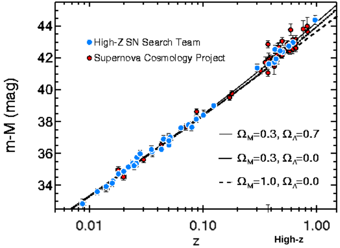
Supernovae Ia as cosmological probe

History of the Universe





First proof with supernovae Ia



- Dark energy causes the universal expansion to accelerate
- Recent observations of supernovae have produced a value for an acceleration that implies a universe that is about 70 % dark energy

General Introduction

Issues for the classification

The problem of representativeness

Classification of light curves

Architecture and data

Results

SPCC

LSST

SDSS

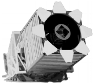
Conclusion

The era of Big Data

1924 Henry Drapper Catalog (0.2 Million)

1989 Guide Star Catalog (20 Million)

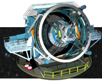
2008 **SDSS (230 Million)**



2018 **Dark Energy Survey (400 Million)**

2027 **Euclid (10 billion)**

2032 **Large Synoptic Survey Telescope (37 billion)**



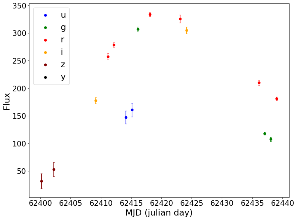
Difficulties for the classification

Many factors degrade the performance of machine learning algorithms:

Small training databases

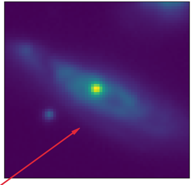
Data can be sparse with an irregular sampling

Non-representativeness between the training and the test databases



The spectroscopic follow-up

Identify and measure the redshift of a galaxy



galaxy

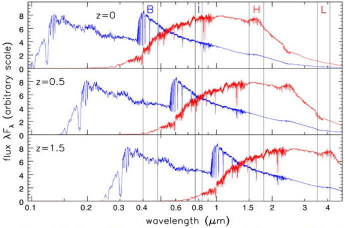
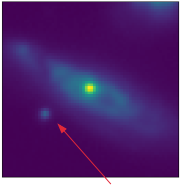
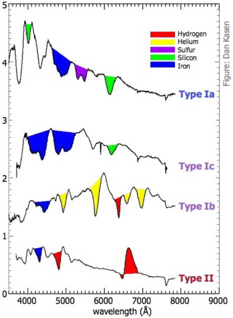


Fig 8.12 (S. Charlot) 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

Determine the nature of an observed object

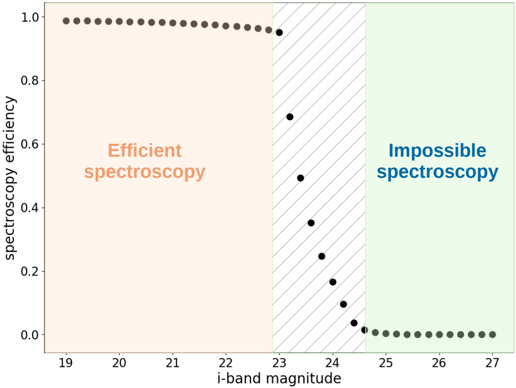


Supernovae

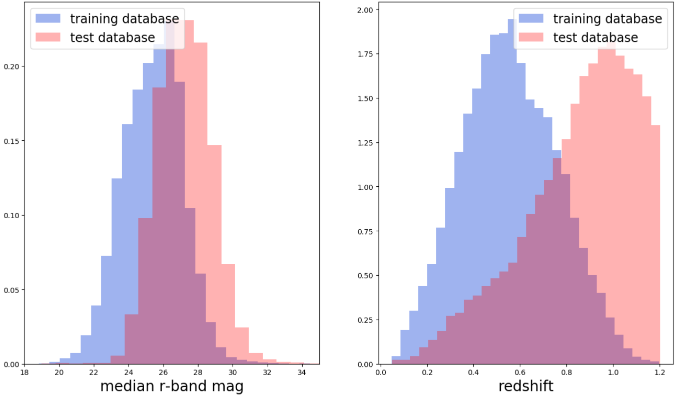


Limitation of the spectroscopic follow-up

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.

General Introduction

Issues for the classification

The problem of representativeness

Classification of light curves

Architecture and data

Results

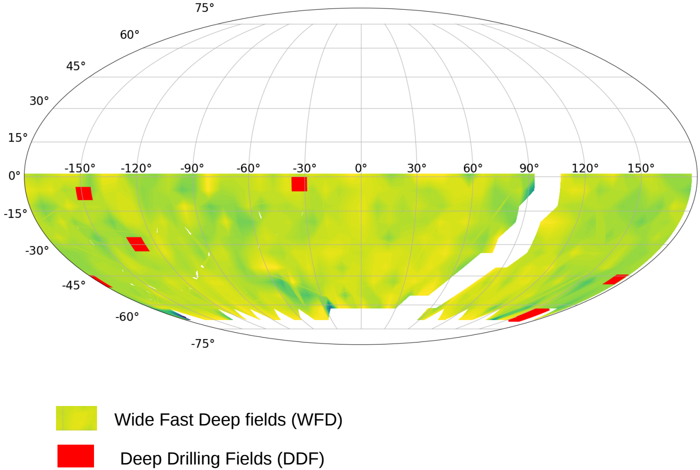
SPCC

LSST

SDSS

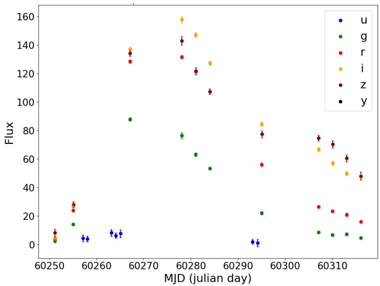
Conclusion

The main survey and the deep fields of LSST

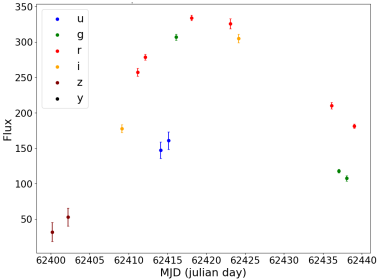


Comparison of light curves

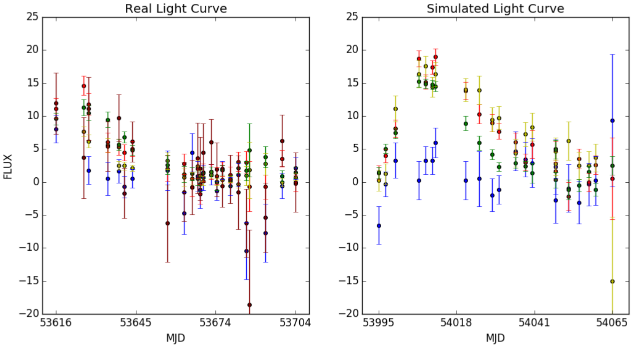
DDF light curve



WFD light curve



A training on simulated data and a testing on real data

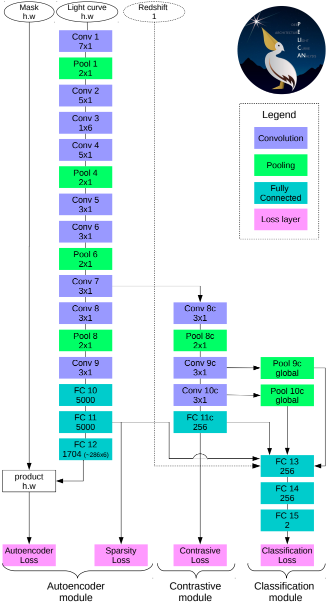


PELICAN: a deeP architecturE for the LIght Curve ANalysis
(Johanna Pasquet, Jérôme Pasquet, Marc Chaumont and Dominique Fouchez)

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.


Mask
h.w
Light curve
h.w
Redshift
1
Conv 1
7x1
Pool 1
2x1
Conv 2
5x1
Conv 3
1x6
Conv 4
5x1
Pool 4
2x1
Conv 5
3x1
Conv 6
3x1
Pool 6
2x1
Conv 7
3x1
Conv 8
3x1
Pool 8
2x1
Conv 9
3x1
FC 10
5000
FC 11
5000
FC 12
1704 (~286x6)
product
h.w
Conv 8c
3x1
Pool 8c
2x1
Conv 9c
3x1
Conv 10c
3x1
FC 11c
256
Pool 9c
global
Pool 10c
global
FC 13
256
FC 14
256
FC 15
2
Autoencoder
Loss
Sparsity
Loss
Contrasive
Loss
Classification
Loss
Autoencoder
module
Contrastive
module
Classification
module
Legend
Convolution
Pooling
Fully
Connected
Loss layer

General Introduction

Issues for the classification

The problem of representativeness

Classification of light curves

Architecture and data

Results

SPCC

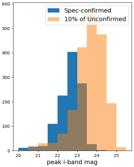
LSST

SDSS

Conclusion

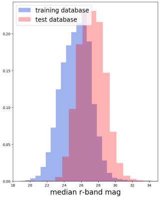
Different databases

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

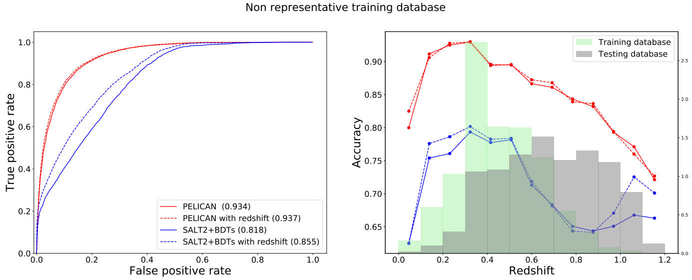


- Small training database (until 500 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up
- Non-representativeness of the sampling and noise between main survey and deep fields

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



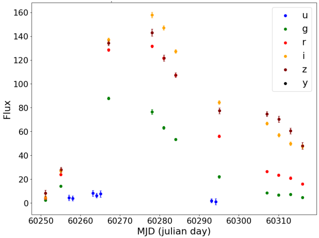
- We compared our results to BDTs classifier + SALT2 features as it is the best combination in Lochner et al. (2016)
- PELICAN obtains an accuracy of 0.856 and an AUC of 0.934 which outperforms BDTs+SALT2 method which reaches 0.705 and 0.818

LSST simulated data

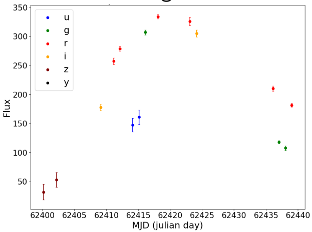
Two methodologies:

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

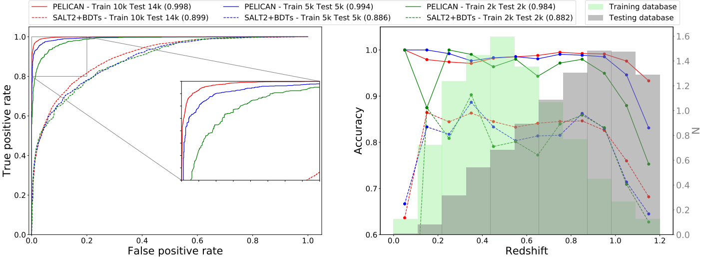
DDF light curve



WFD light curve

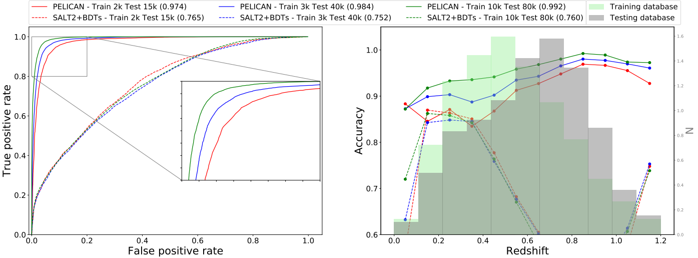


Results on DDF



	Training database (spec only)	Test database (phot only)	Accuracy	Recall_{Ia} $\text{Precision}_{Ia}>0.95$	Recall_{Ia} $\text{Precision}_{Ia}>0.98$	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)

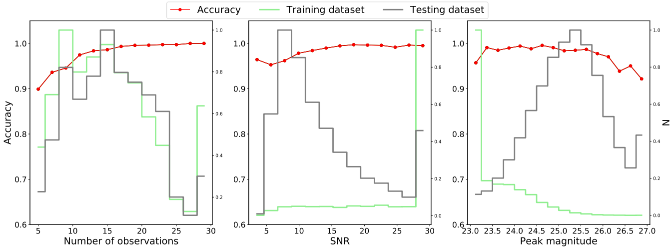
Results on WFD



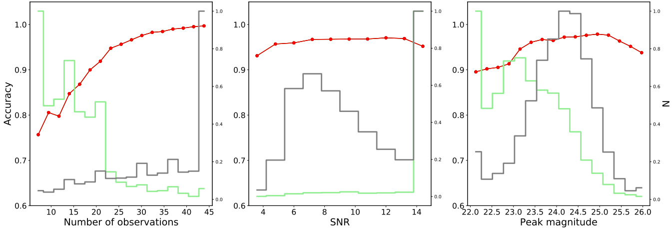
	Training database (spec only)	Test database (phot only)	Accuracy	$Recall_{Ia}$ $Precision_{Ia} > 0.95$	$Recall_{Ia}$ $Precision_{Ia} > 0.98$	AUC
WFD	DDF Spec : 2, 000	WFD : 15, 000	0.917 (0.650)	0.857 (0.066)	0.485 (0.000)	0.974 (0.765)
	DDF Spec : 3, 000	**WFD : 40, 000**	**0.940 (0.650)**	**0.939 (0.111)**	**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

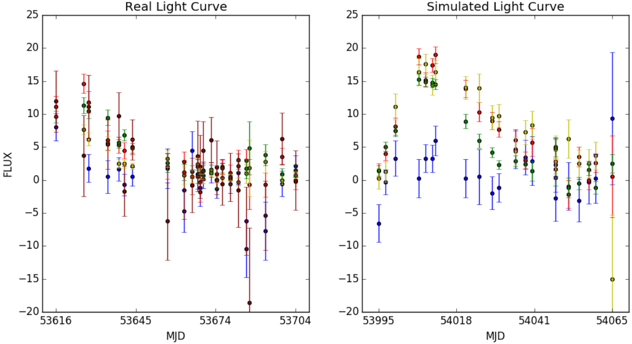
DDF



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

Era of Big data

The future surveys will deliver multi-band photometry for billions of sources

Many issues for the classification algorithms

- Small size of the training database due to the limitation of the spectroscopic follow-up
- Several problems of representativeness
- Nature of data : sparse with an irregular sampling

New solutions for the classification of light curves

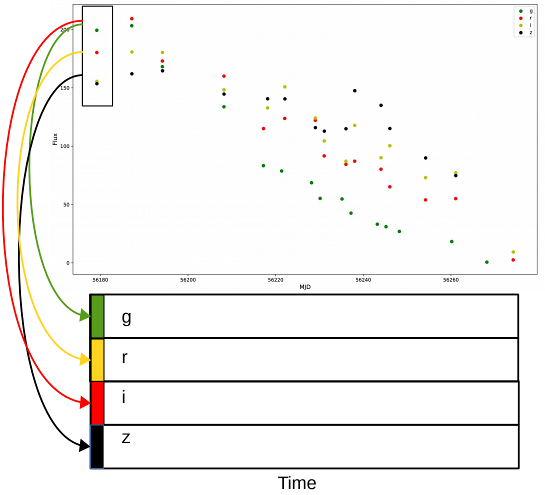
PELICAN obtained the best performance ever achieved with a non-representative training database of the SPCC challenge

PELICAN is able to significantly remove several types of non-representativeness between the training and the test databases due to :

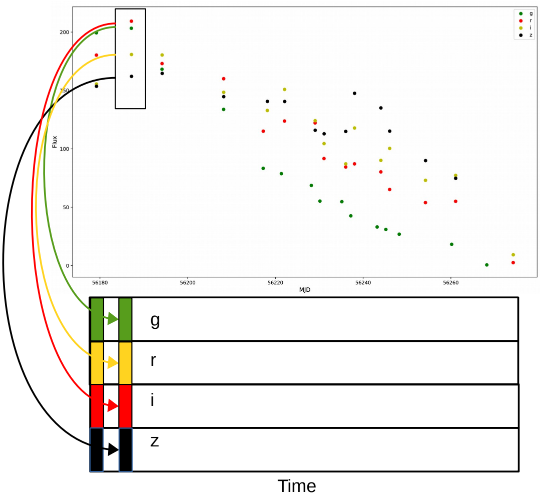
- the limit in brightness and redshift of the spectroscopically confirmed data
- the different observational strategies
- the difficulty of simulated data to reproduce perfectly real data

PELICAN can deal with the data that are sparse, with an irregular sampling

The Light Curve Image (LCI)



The Light Curve Image (LCI)



⚠ Overfitting of missing data (zero values)

Projection of features