Les activités deep learning en cosmologie au CPPM

Johanna Pasquet

Centre de Physique des Particules de Marseille

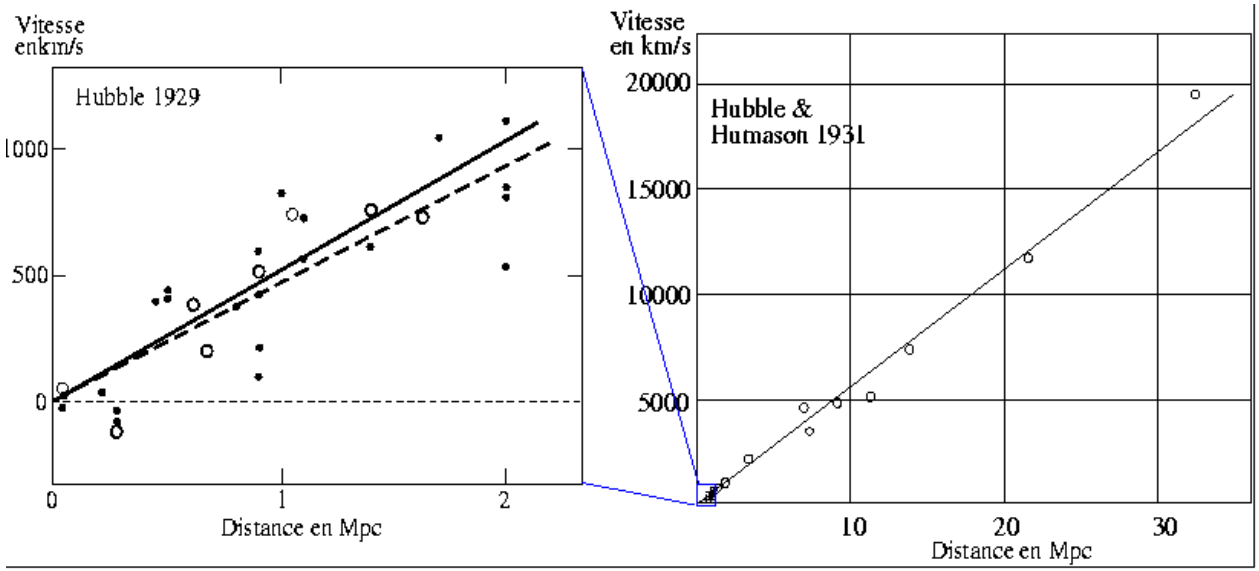
Méthodes IA / Data Science pour la physique

29 Janvier 2019

Définition

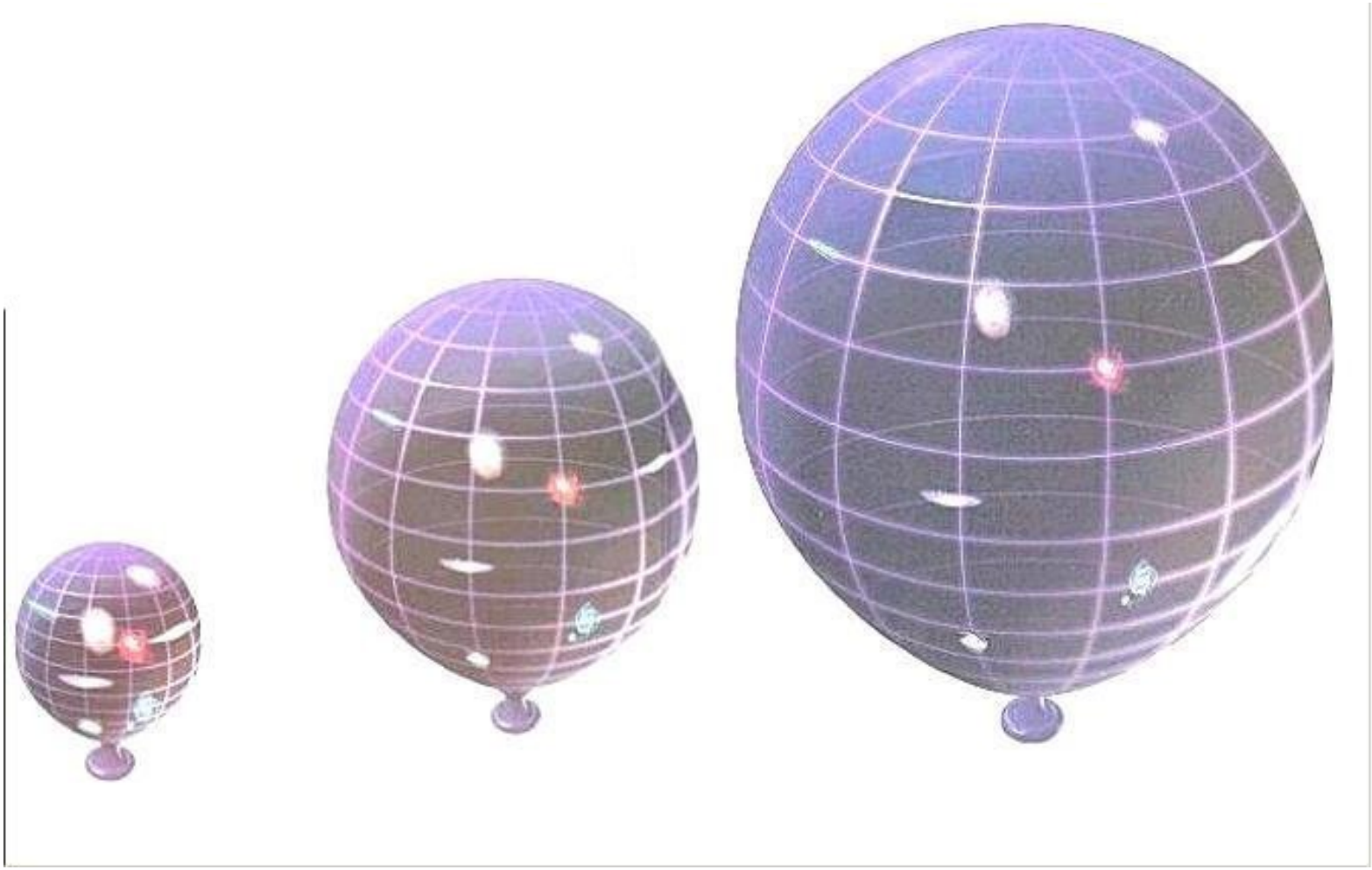
Cosmologie: branche de la science qui étudie l'univers comme un tout, son origine et sa possible évolution future.

Loi de Hubble



Crédit : Figure de gauche : Hubble E.P. 1929, ApJ 69, 103 ; figure de droite : Hubble E.P. & Humason M.L. 1931, ApJ 74, 43

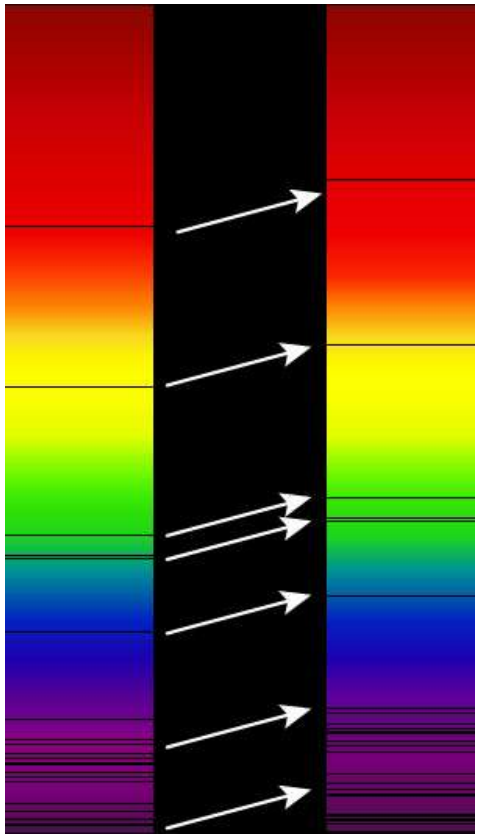
L'univers est en expansion!



Crédit : James N. Imamura of U. of Oregon

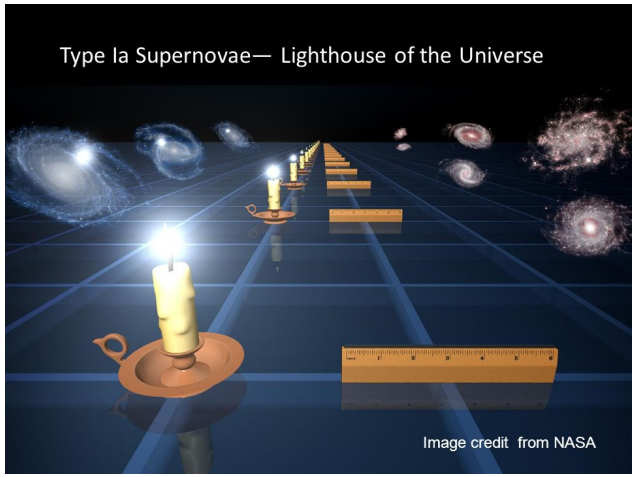
Le décalage vers le rouge ou *redshift*

Lorsque l'on reçoit la lumière d'une galaxie lointaine, la longueur d'onde de cette dernière, du fait de l'expansion de l'univers, entre le moment où elle a été émise et le moment où elle est observée par un télescope, s'est dilatée. La lumière apparaît alors plus rouge.



Découverte majeure avec les supernovae Ia

Etoiles très massives en fin de vie qui s'effondrent brutalement sur elles-mêmes et explosent en libérant d'énormes flashes de lumière.



En étudiant plusieurs supernovae Ia, Perlmutter, Riess et Schmidt s'aperçoivent qu'elles sont moins brillantes qu'en théorie car elles sont plus éloignées que prévu.

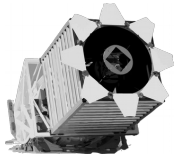
L'expansion de l'univers s'accélère sous l'effet d'une mystérieuse "énergie noire" qui s'opposerait à la gravitation et constituerait plus de 70% de l'univers.

L'ère du Big Data en Cosmologie

1924 Henry Drapper Catalog (0.2 Million)

1989 Guide Star Catalog (20 Million)

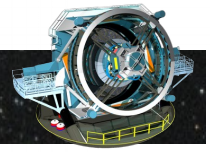
2008 SDSS (230 Million)



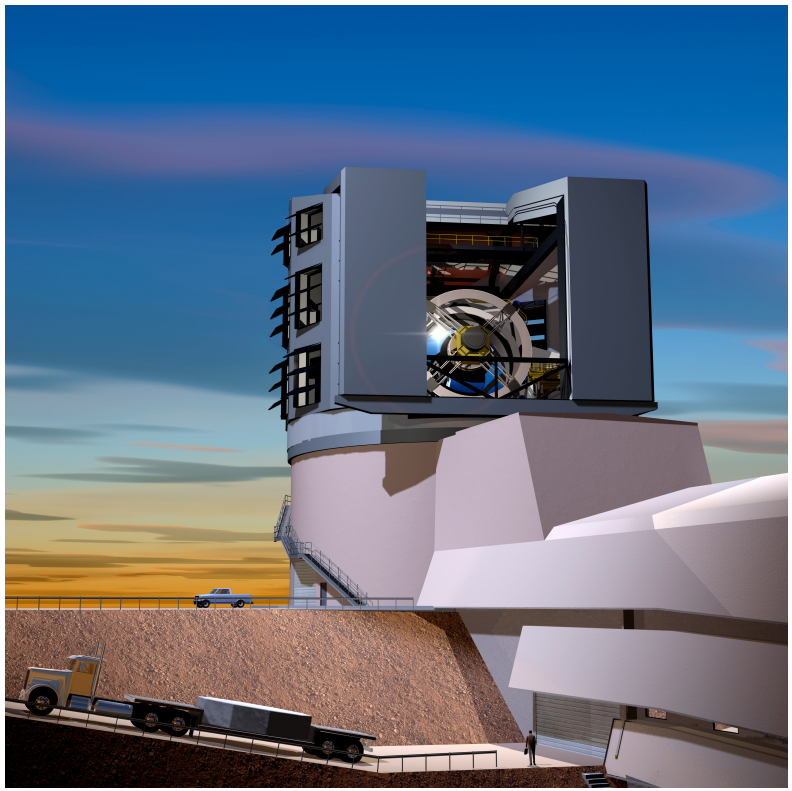
2018 Dark Energy Survey (400 Million)

2027 Euclid (10 billion)

2032 Large Synoptic Survey Telescope (37 billion)



Large Synoptic Survey Telescope (LSST)

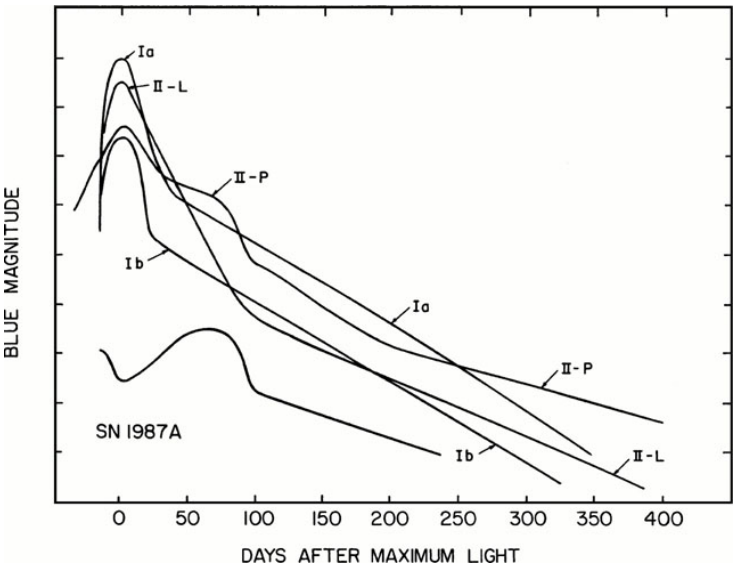


Artist view, Credit : Todd Mason, Mason Productions Inc. / LSST Corporation

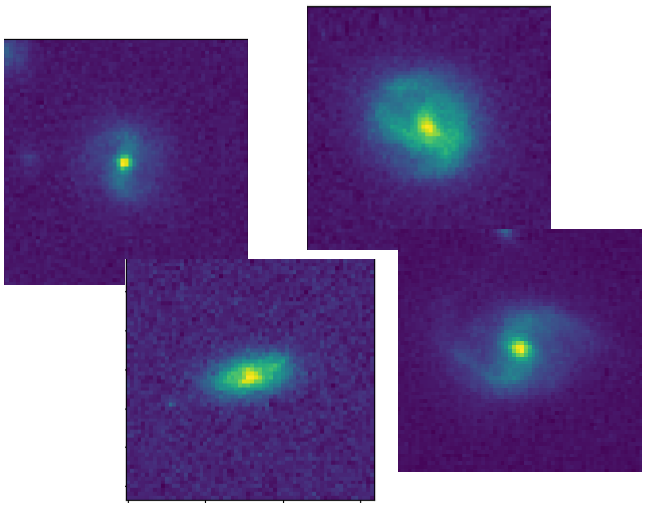
- relevé pendant 10 ans
- première lumière en 2020
- image du ciel entier en trois nuits
- 200 petabyte d'images et de données produites !
- 0.5 Exabytes de disque de stockage
- Un centre de calcul aux USA (+NERSC) et un au CC à Lyon

Méthodologie pour LSST

Séries temporelles

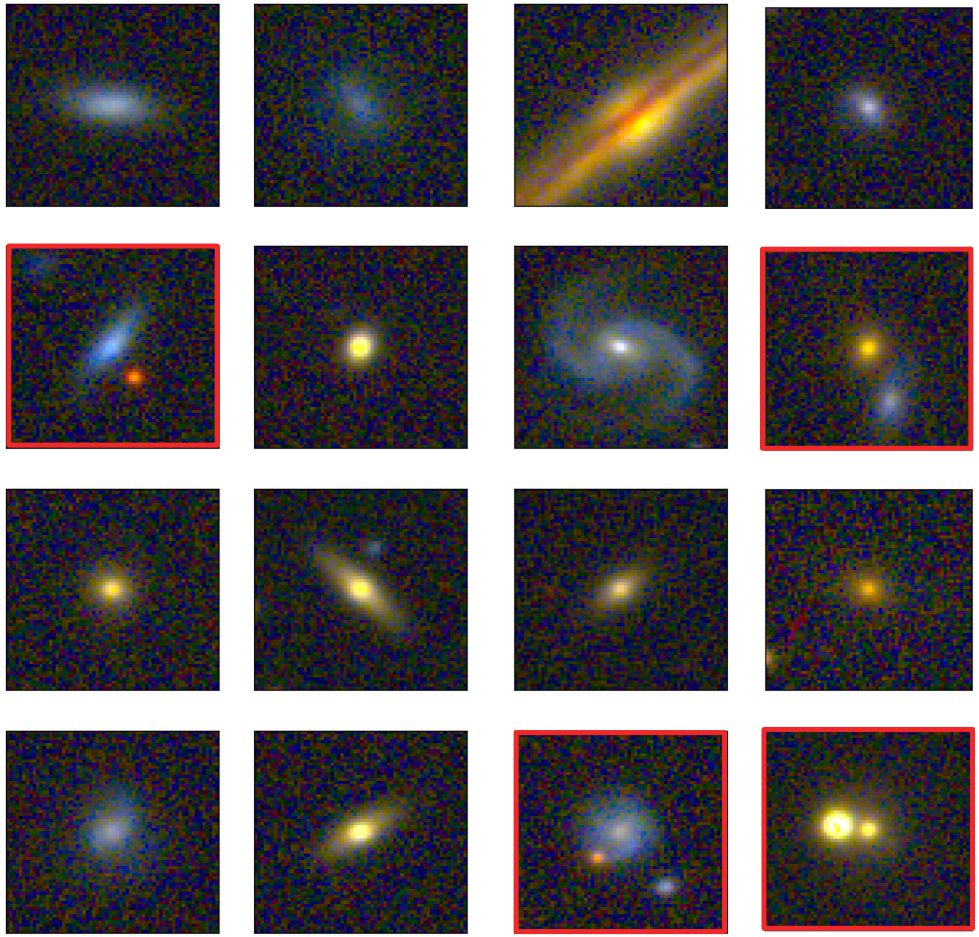


Images



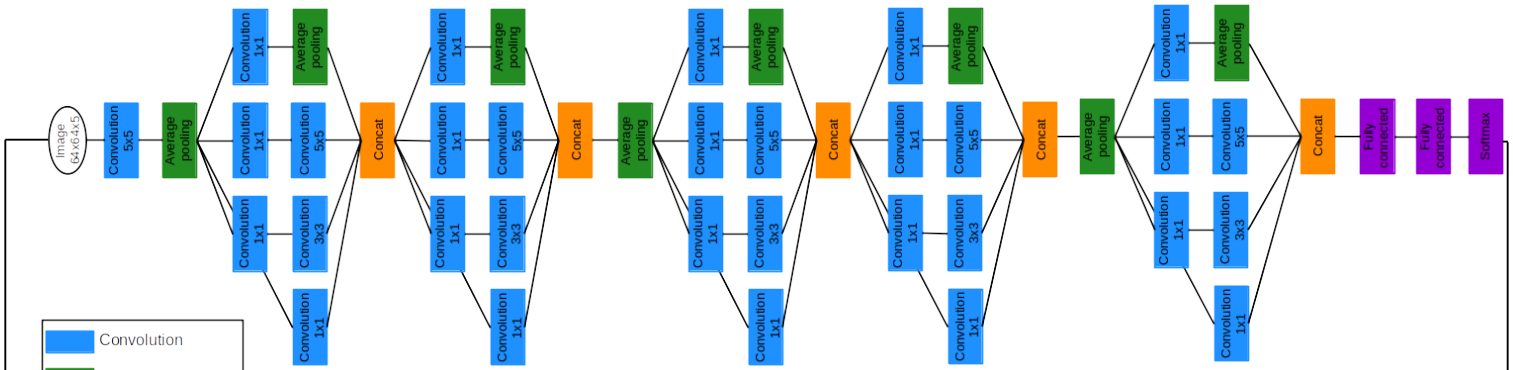
Problématique pour les redshifts

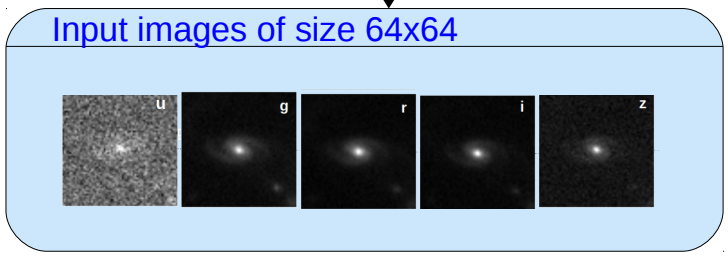
Déterminer le redshift des galaxies avec des images dans plusieurs bandes spectrales

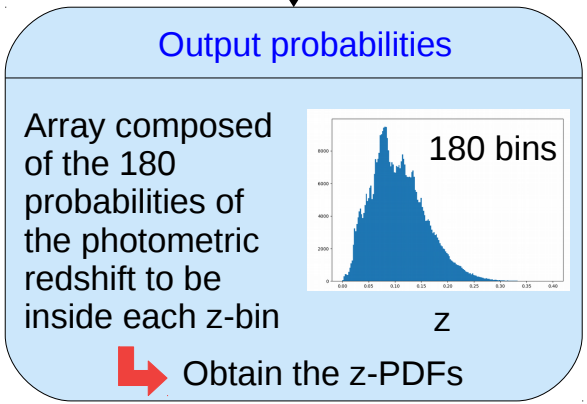


Un réseau de neurones convolutif

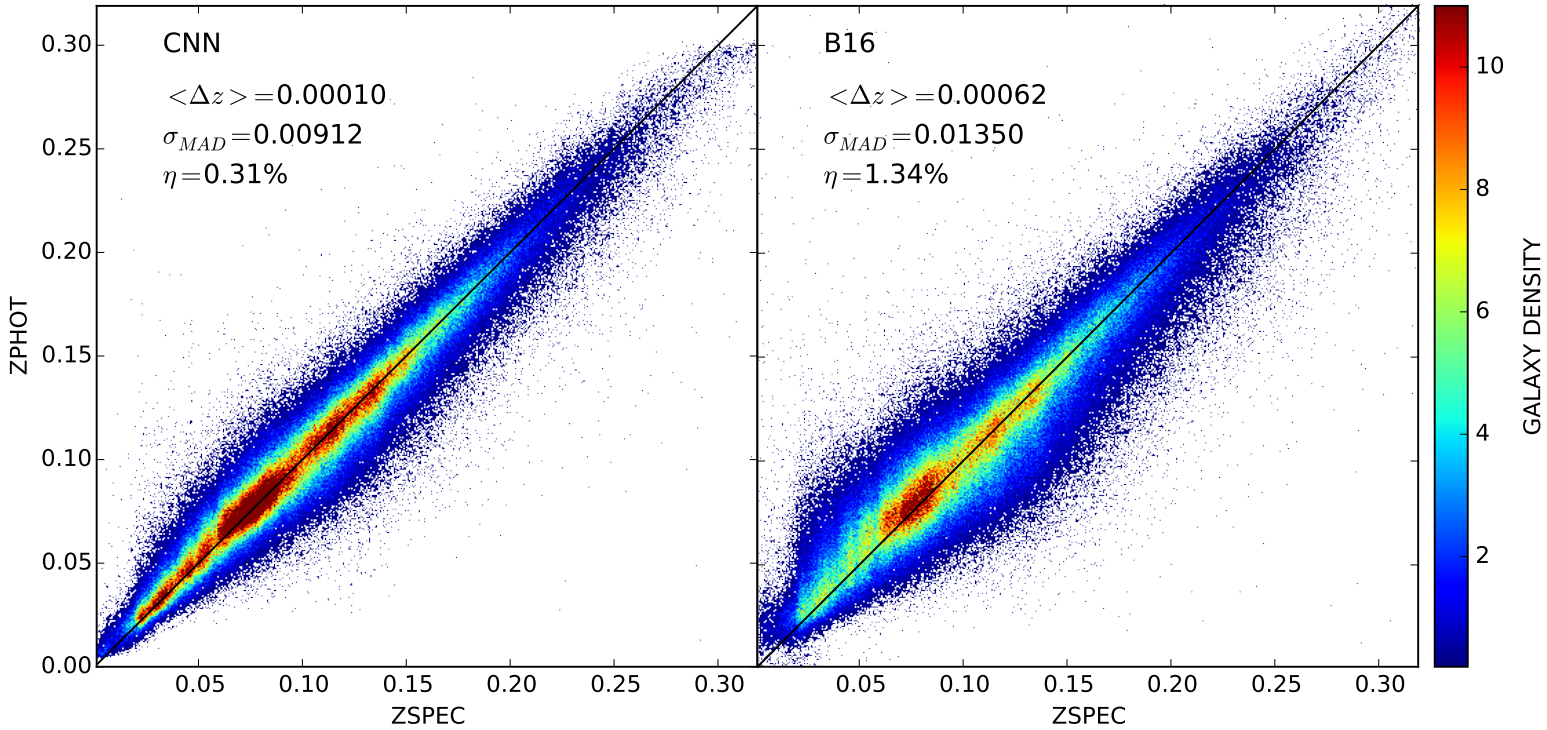
Johanna Pasquet(CPPM), Emmanuel Bertin (IAP), Marie Treyer (LAM), Stephane Arnouts (LAM) and Dominique Fouchez (CPPM) (A&A, 611 :A97, 2018, **arxiv**: 1806.06607, **code disponible**: https://github.com/jpasquet/Photoz)

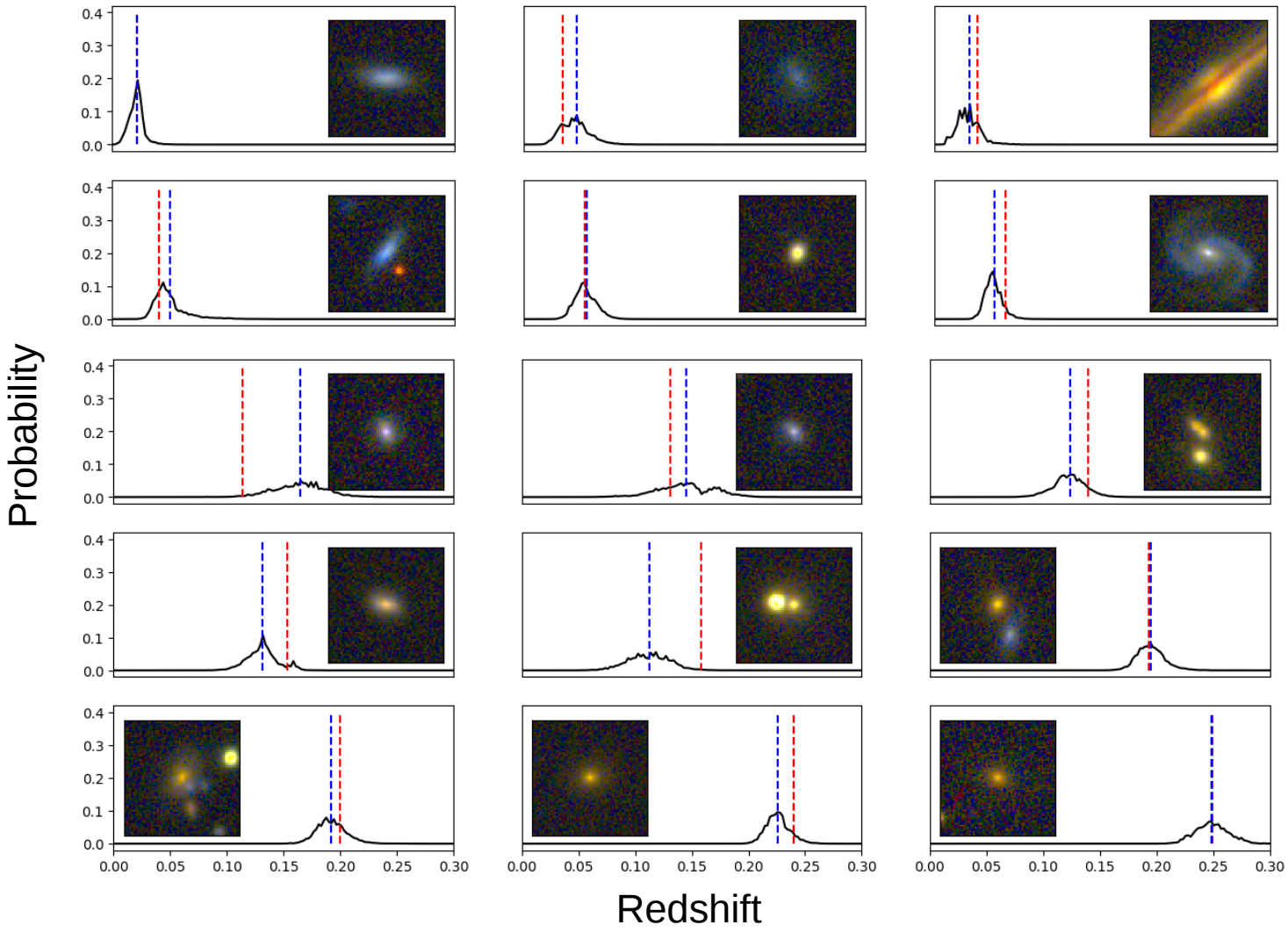






Des performances jamais atteintes!





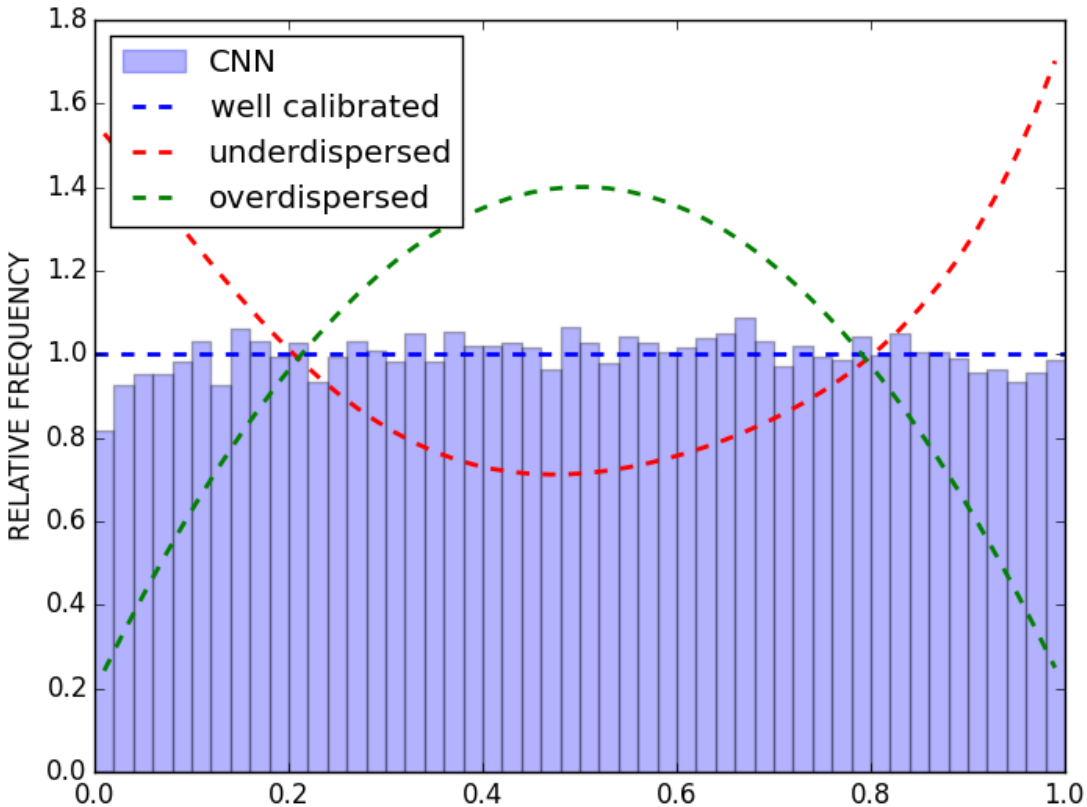
-- Spectroscopic redshift

-- Photometric redshift

Evaluation des PDFs

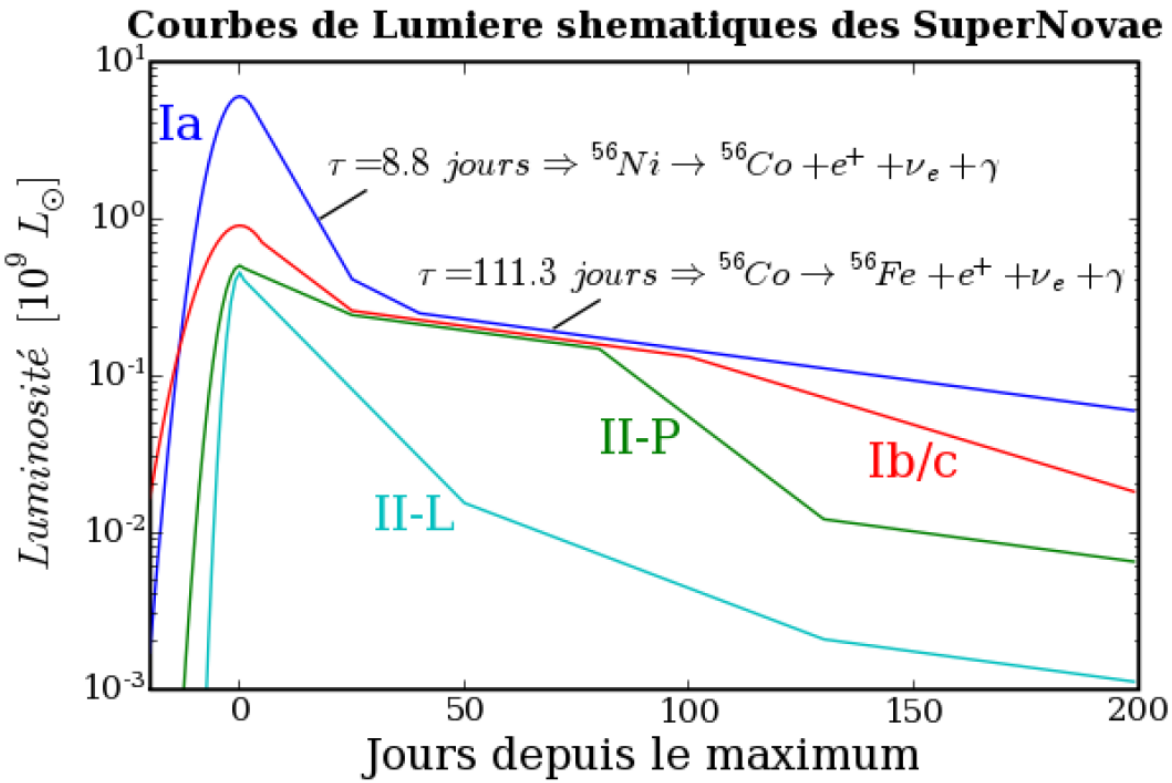
Le test du PIT (Probability Integral Transform, Dawid 1984) calcule l'histogramme des probabilités cumulées à la vraie valeur. Pour une galaxie i avec un redshift de z_i le PIT est de:

$$\mathrm{PIT}_i = \int_{-\infty}^{z_i} PDF_i(z)dz$$

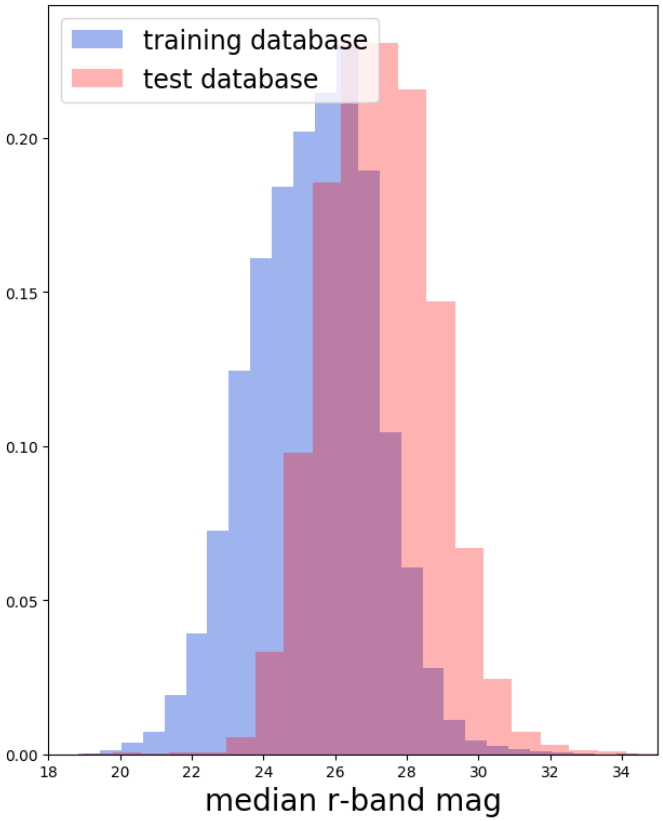


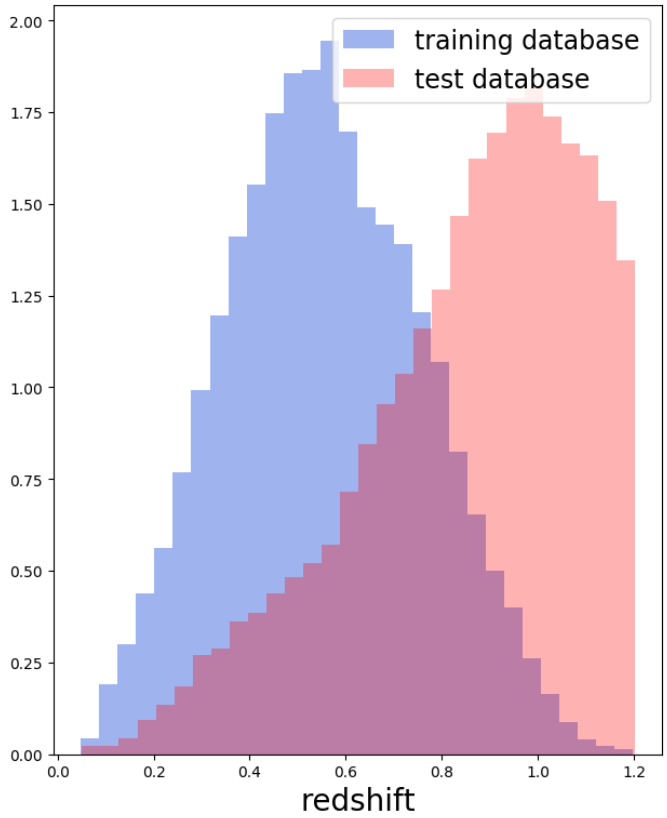
Problématique pour les supernovae

Classer des séries temporelles de supernovae pour séparer les supernovae Ia des autres



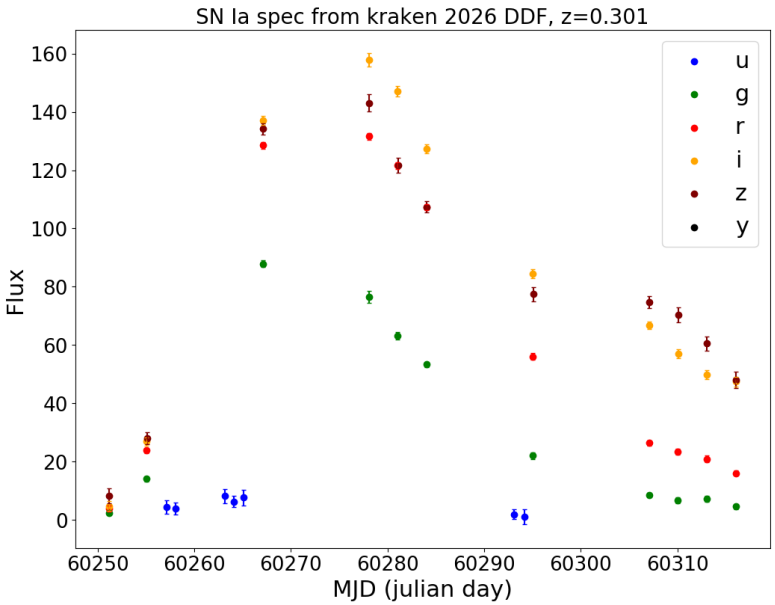
Non-representativité entre les bases d'apprentissage et de test



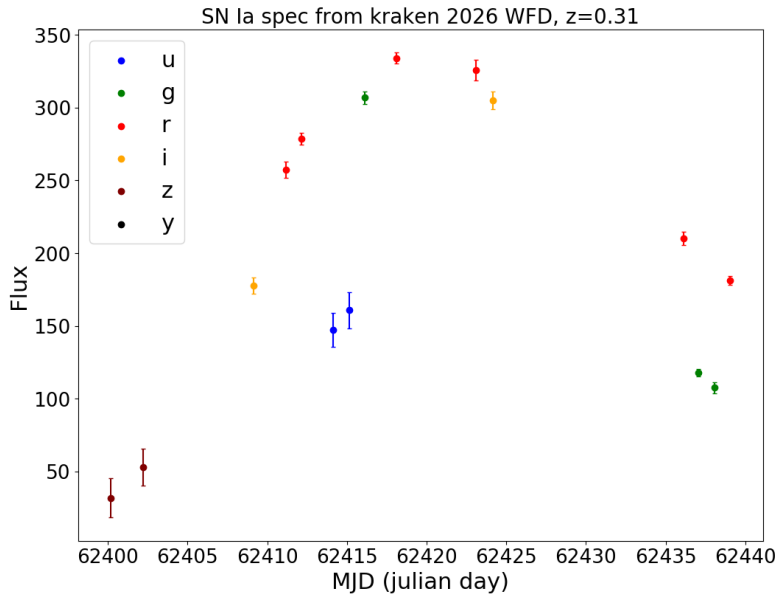


Stratégie observationnelle variable

Stratégie 1

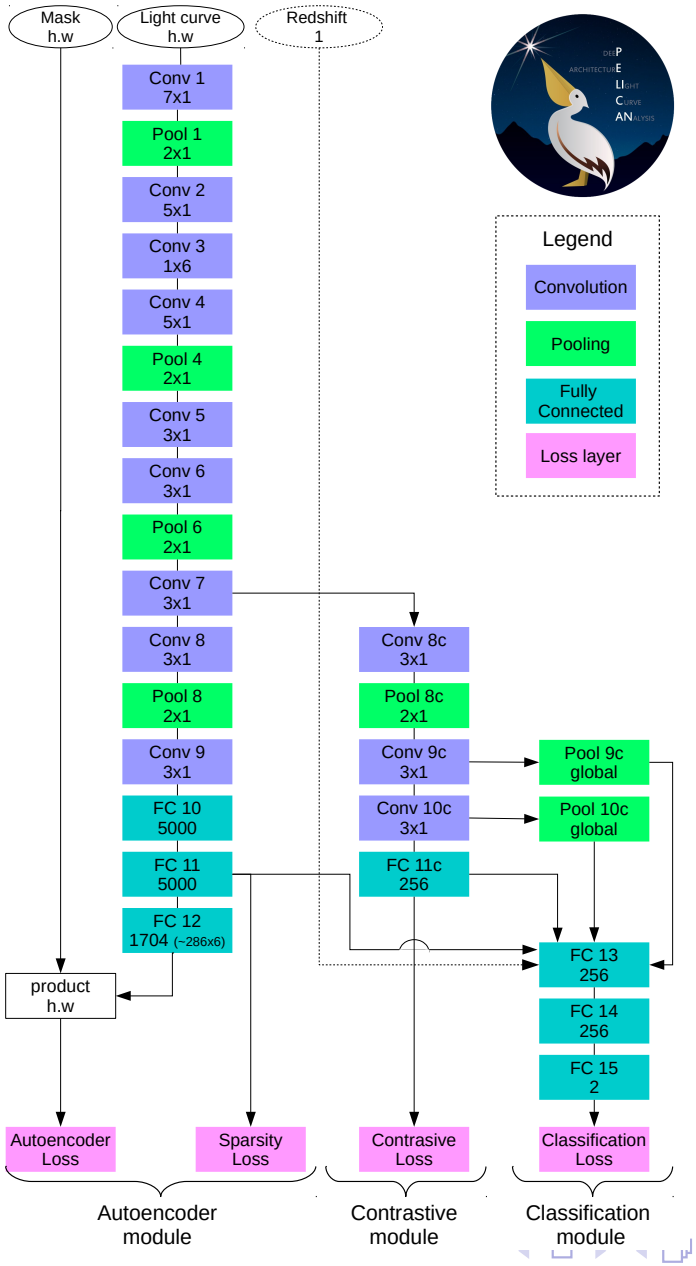


Stratégie 2

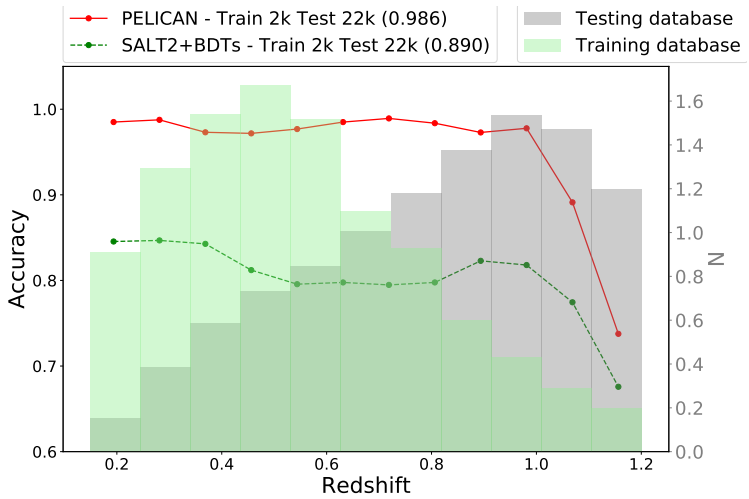


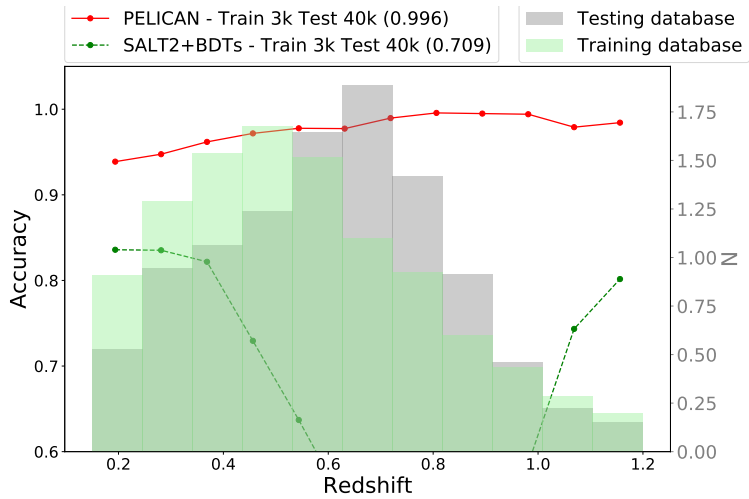
PELICAN: a deeP architecturE for the LIght Curve ANalysis

Johanna Pasquet (CPPM), Jérôme Pasquet (Tetis, Montpellier), Marc Chaumont (LIRMM, Montpellier) and Dominique Fouchez (CPPM) (**arxiv**: 1901.01298)

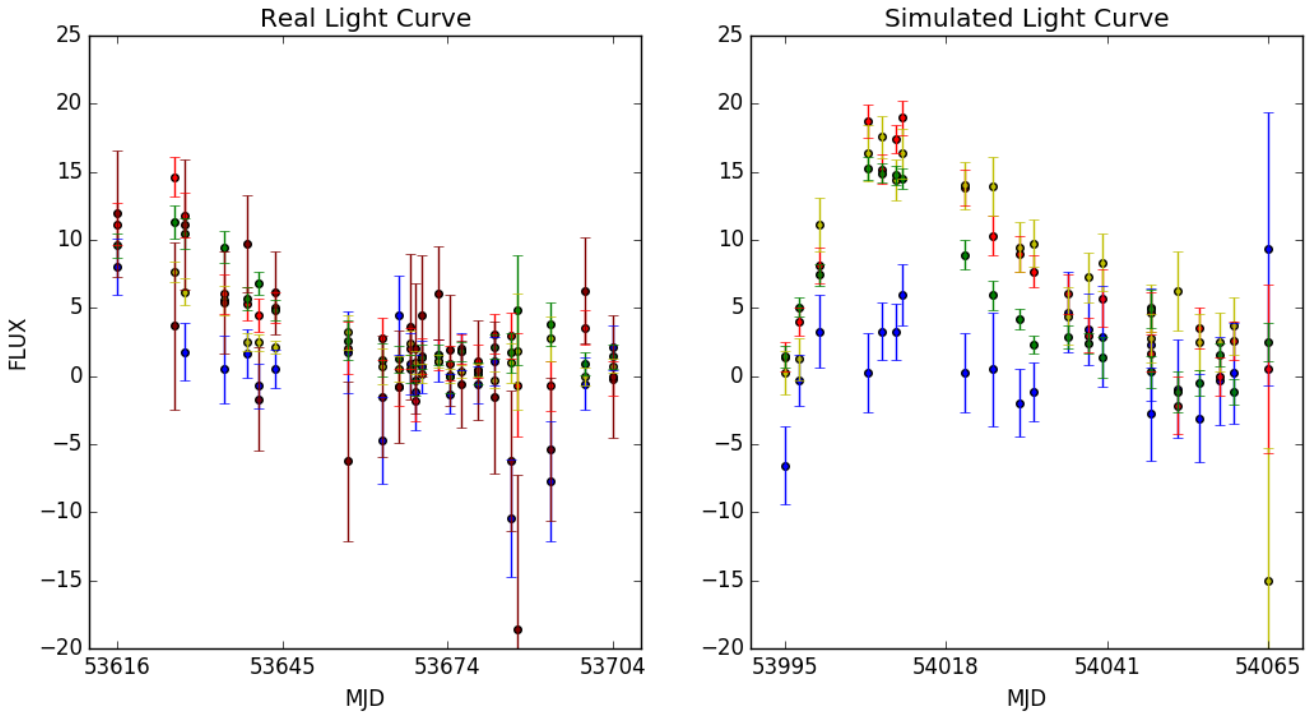

Mask
h.w
Light curve
h.w
Redshift
1
Conv 1
7x1
Pool 1
2x1
Conv 2
5x1
Conv 3
1x6
Conv 4
5x1
Pool 4
2x1
Conv 5
3x1
Conv 6
3x1
Pool 6
2x1
Conv 7
3x1
Conv 8
3x1
Pool 8
2x1
Conv 9
3x1
FC 10
5000
FC 11
5000
FC 12
1704 (~286x6)
product
h.w
Conv 8c
3x1
Pool 8c
2x1
Conv 9c
3x1
Conv 10c
3x1
FC 11c
256
Pool 9c
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

Resultats sur des données simulées de séries temporelles





Apprentissage de transfert (*transfer learning*)



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

In the money Gold Silver Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Kyle Boone			0.68503	104	1mo
2	▲ 2	Mike & Silogram			0.69933	176	1mo
3	▼ 1	Major Tom			0.70016	366	1mo
4	▼ 1	AhmetErdem			0.70423	233	1mo
5	—	SKZ Lost in Translation			0.75229	343	1mo
6	▲ 2	Stefan Stefanov			0.80173	28	1mo
7	▲ 3	hklee			0.80836	63	2mo
8	▼ 1	rapids.ai			0.80905	133	1mo
9	▼ 3	Three Musketeers			0.81312	313	1mo
10	▲ 3	J&J			0.81901	246	1mo

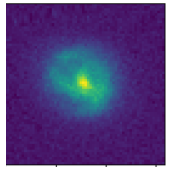
Résumé

- Les futurs grands relevés en Cosmologie delivreront plus de données qu'il ne sera possible de labelliser
- Les méthodes de Deep Learning que nous avons dévelopées apportent une solution automatique dans un contexte réaliste (non-représentativité, faible nombre de données d'apprentissage, stratégie observationnelle variable...)
- Ces méthodes pourraient également être adaptées à d'autres domaines de recherche (e.g. thèse en co-tutelle sur une thématique d'images sous-marines)

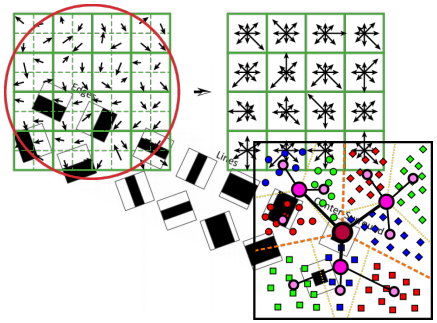
The main property of deep learning

Classical methods

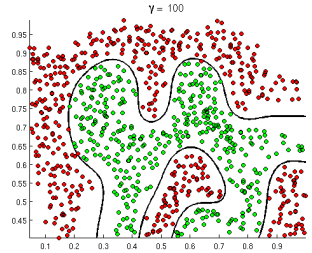
Input data



Feature crafting

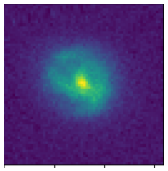


Separation with a classifier

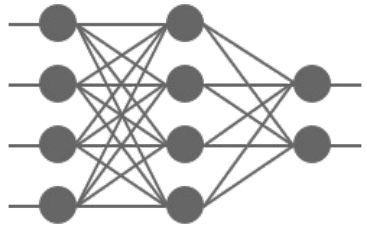


Deep learning

Input data



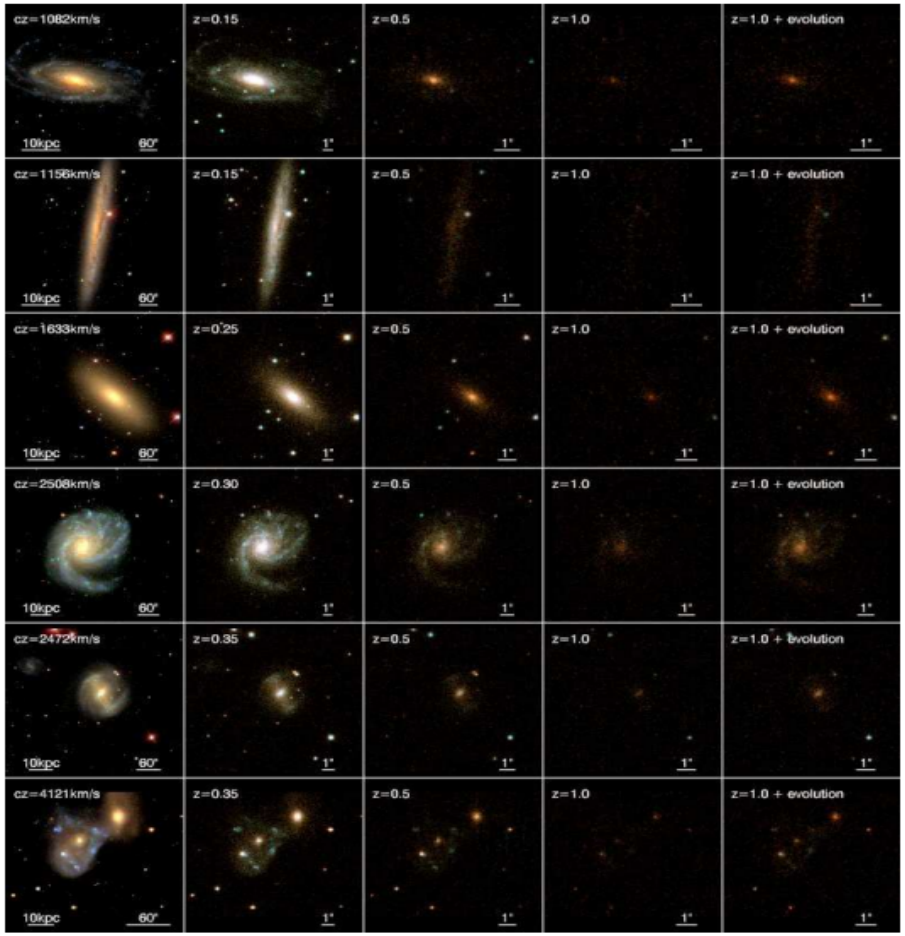
Feature learning



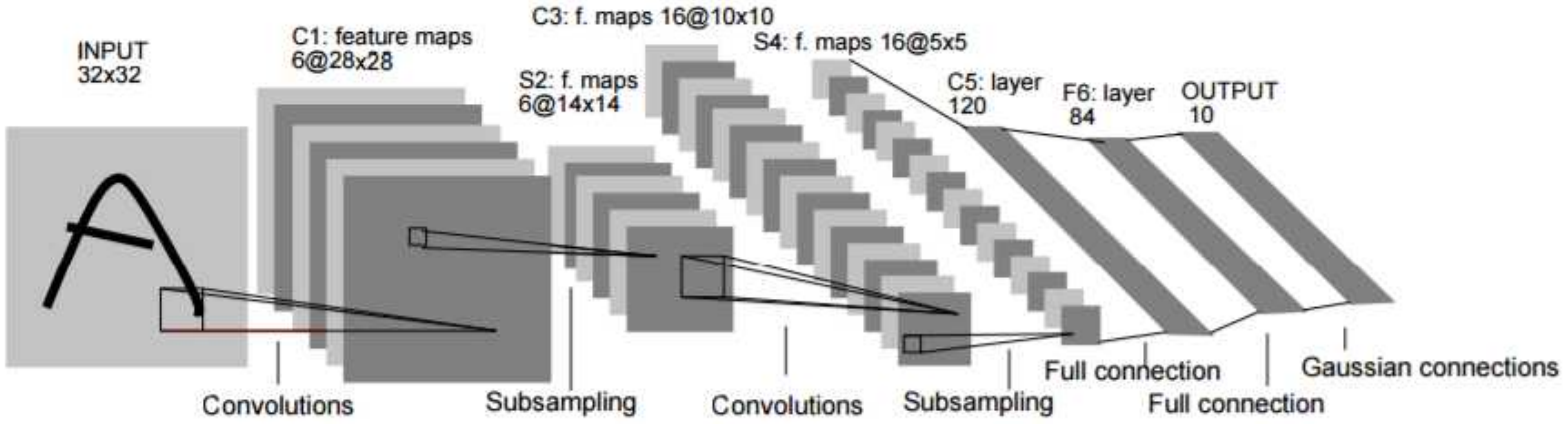
→ The best feature space representation is found by the network

The convolutional neural network in astronomy

Kaggle challenge with the goal to build an algorithm to classify the different morphologies of galaxies from JPEG images : a CNN won the challenge (Dieleman et al. 2015)



LeNet5

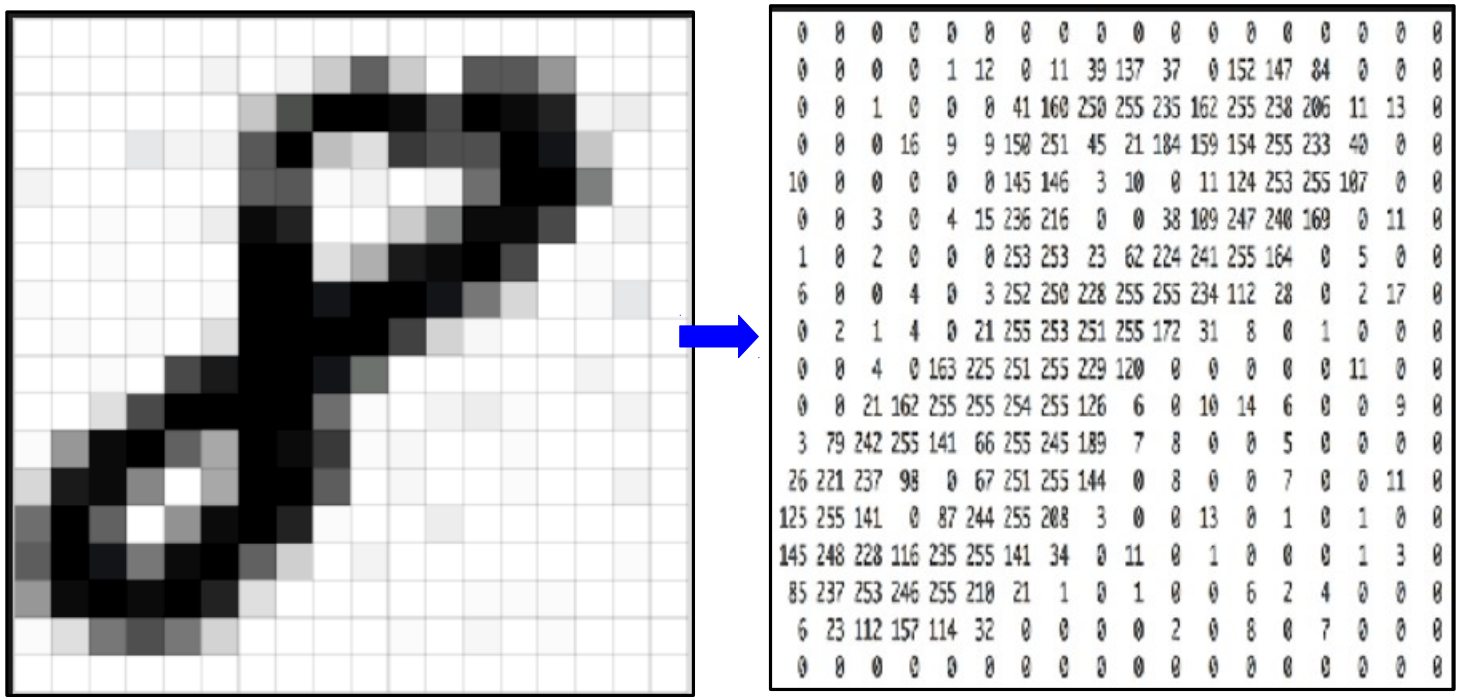


Lecun et al. 1998

3 operations:

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

Convolutions



Every image is a matrix of pixel values.

Convolutions

An image

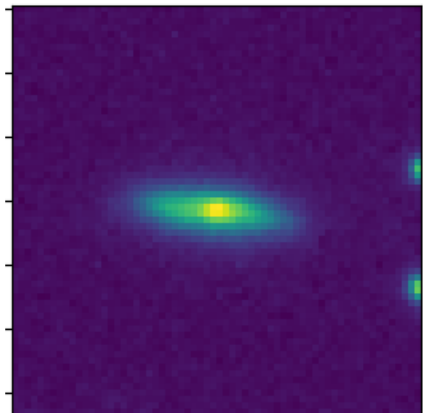
1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

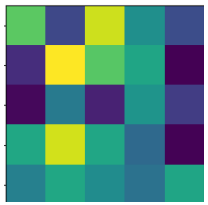
A kernel

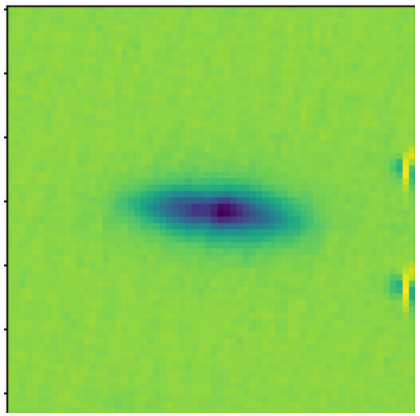
1	1	1
0	1	1
0	0	1

A convolved image

4	3	4
2	4	3
2	3	4

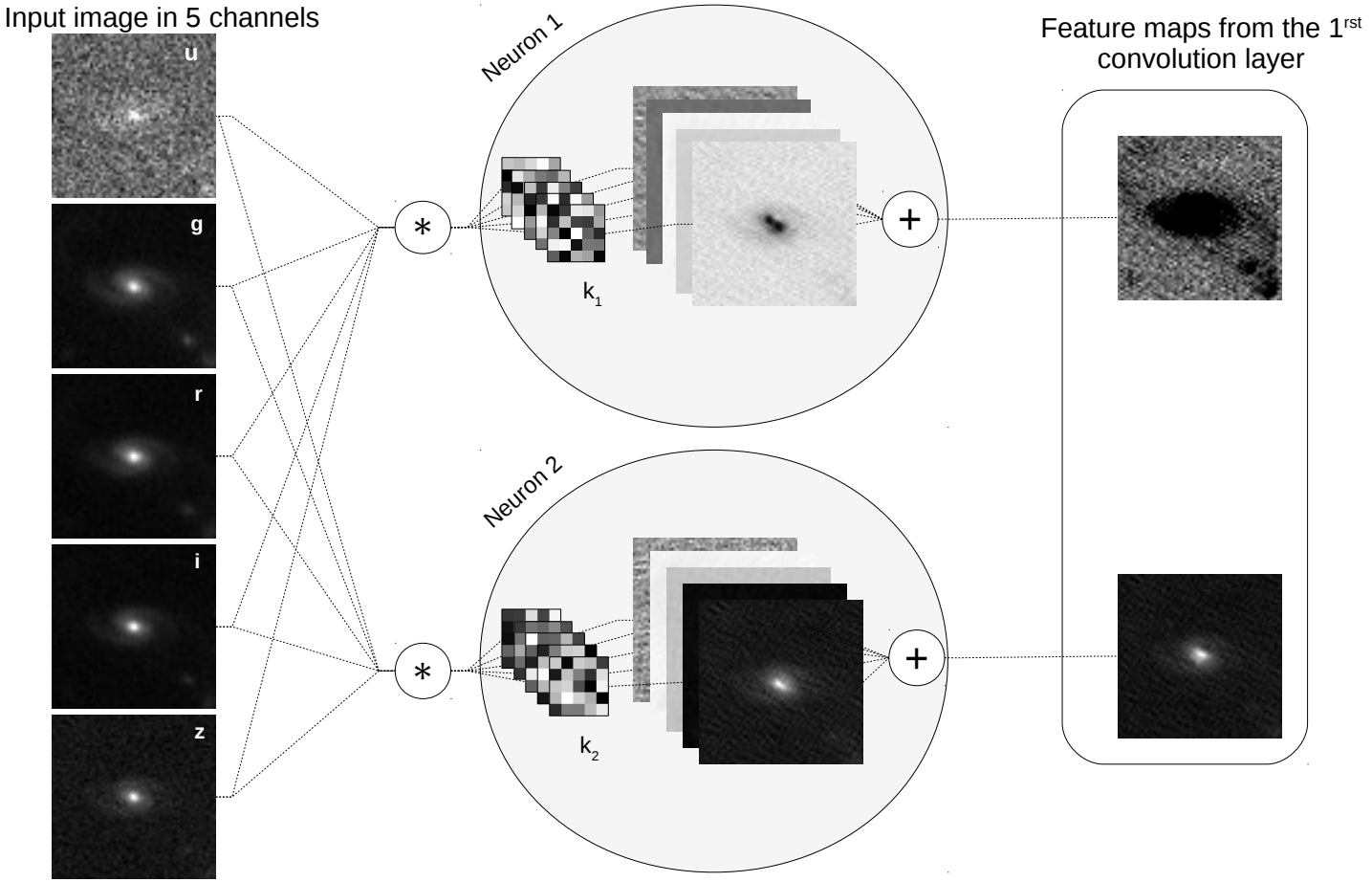






Then introduce non-linearity (tanh, ReLu...)

Convolutions



Pooling

A feature map

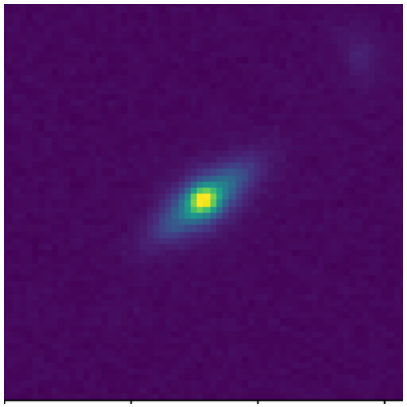
5	1	3	0
0	1	2	7
2	1	1	4
3	1	1	2

Pooling operation

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

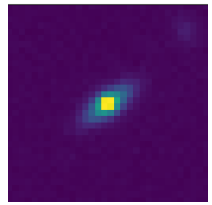
A subsampled feature map

5	7
3	4



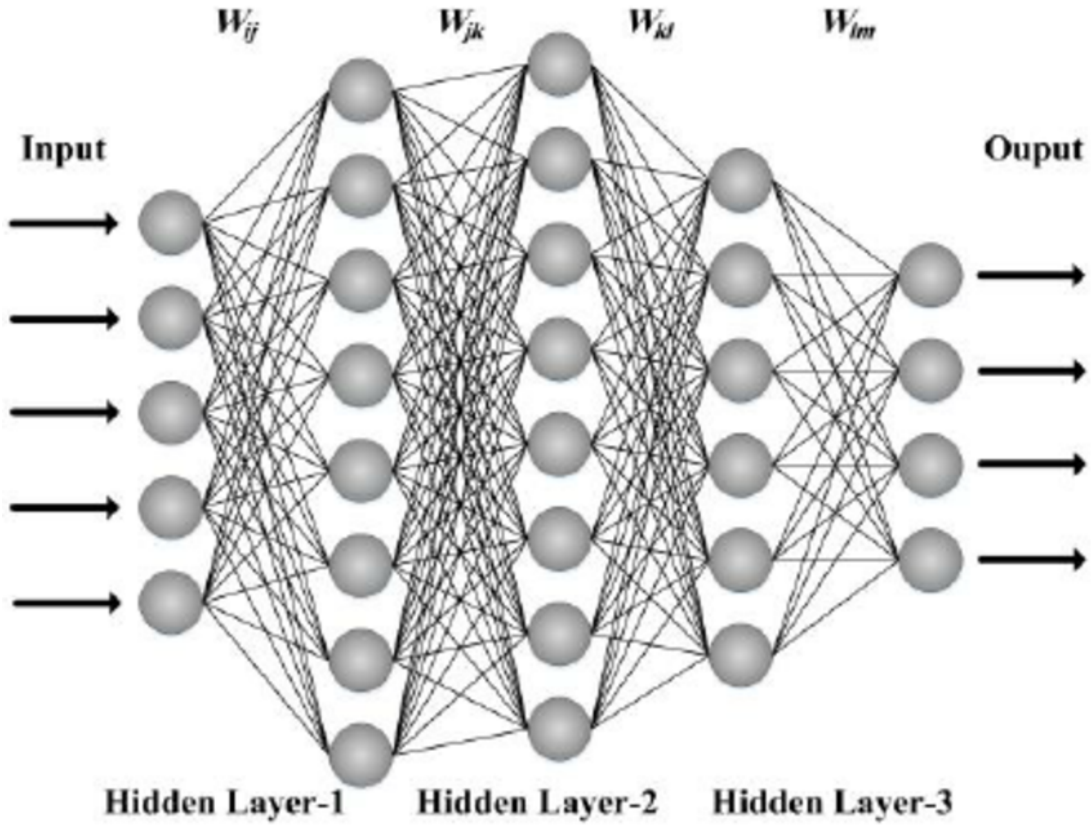
64x64

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



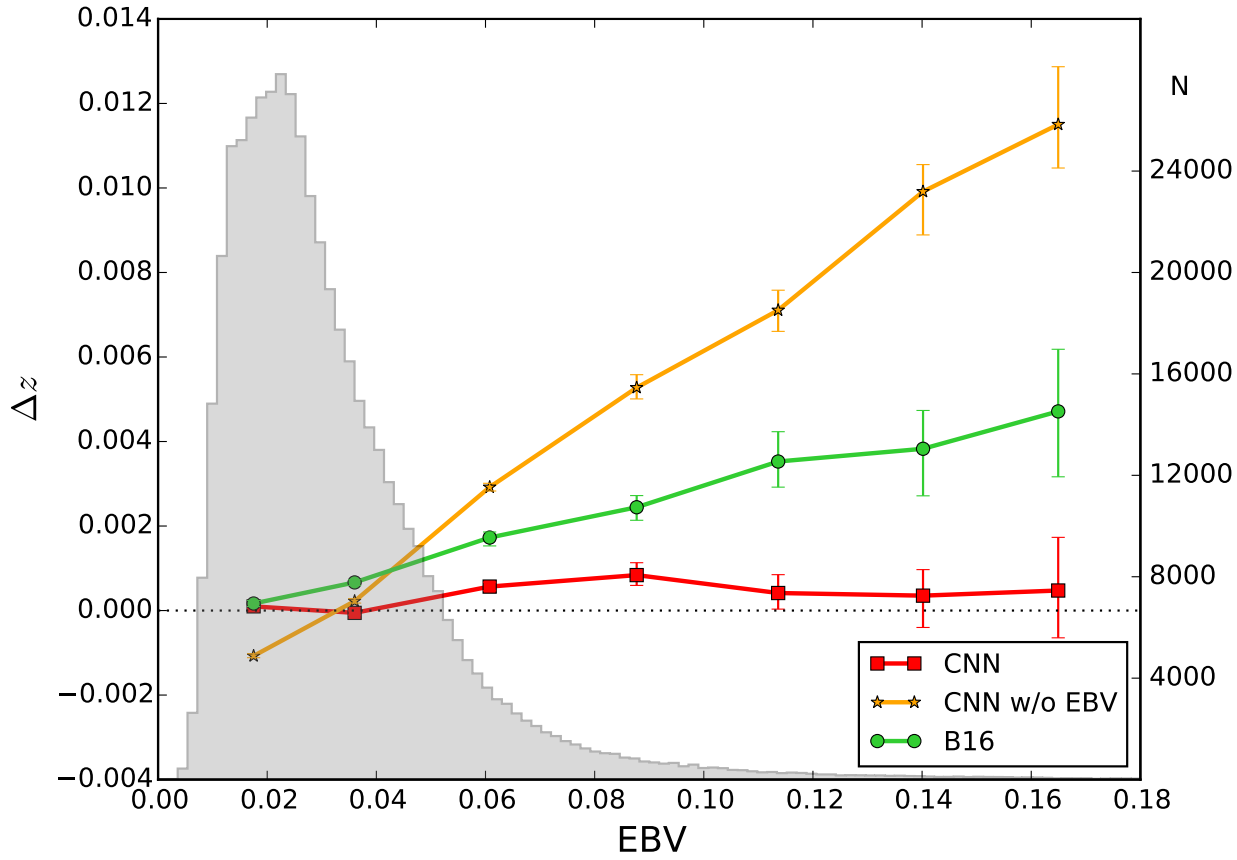
32x32

Fully connected



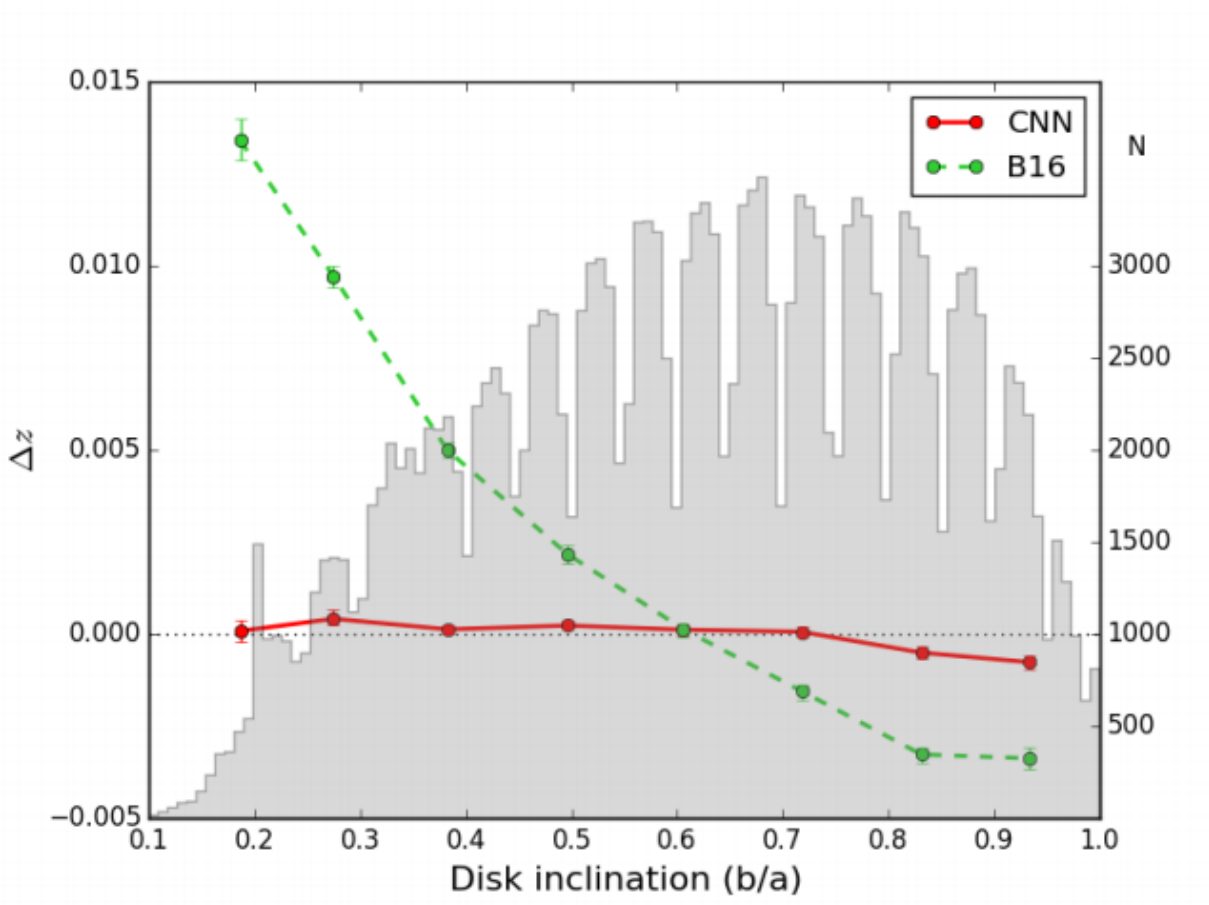
Impact of the extinction of our Galaxy on photometric redshifts

Our method tends to overestimate redshifts in obscured regions (confusing galactic dust attenuation with redshift dimming), unless $E_{(B-V)}$ is used for training



Impact of the disk inclination of galaxies on photometric redshifts

Our method automatically corrects for galactic dust reddening which increases with disk inclination



Summary results

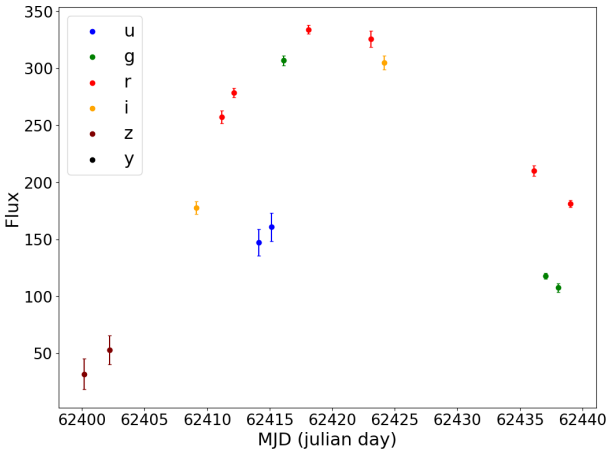
Trial	training sample size	bias	σ	η
Training with 80% of the dataset	393,219			
Full test sample		**0.00010**	**0.00912**	**0.31**
(B16)		(0.00062)	(0.01350)	(1.34)
Widest 20% of PDFs		**0.00005**	**0.00789**	**0.06**
Stripe 82 only		-0.00009	0.00727	0.34
Stripe 82 with widest 20% of PDFs removed		**0.00004**	**0.00635**	**0.09**
Training with 50% of the dataset*	250,000	0.00007	0.00910	0.29
Training with 20% of the dataset	**99,001**	**-0.00001**	**0.00914**	0.30
Training with 2% of the dataset	10,100	-0.00017	0.01433	1.26
Training and testing on Stripe 82	**15,771**	-0.00002	**0.00795**	0.38

Difficulties for the classification

Many factors degrade the performance of machine learning algorithms:

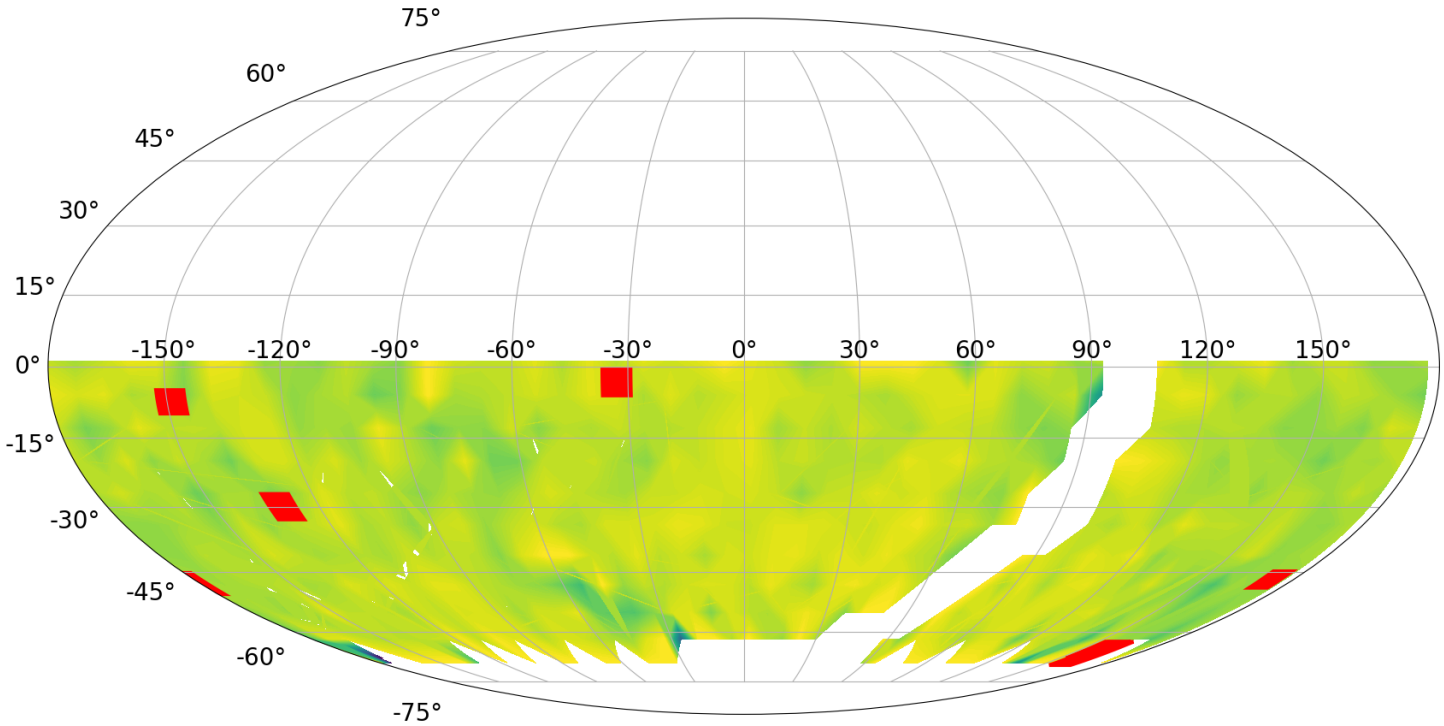
Small training databases

Data can be sparse with an irregular sampling



Non-representativeness between the training and the test databases

The main survey and the deep fields of LSST

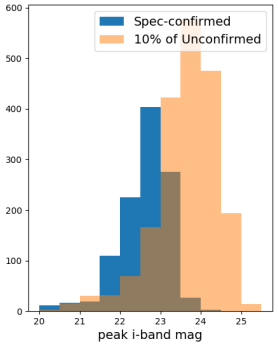


Wide Fast Deep fields (WFD)

Deep Drilling Fields (DDF)

Different databases

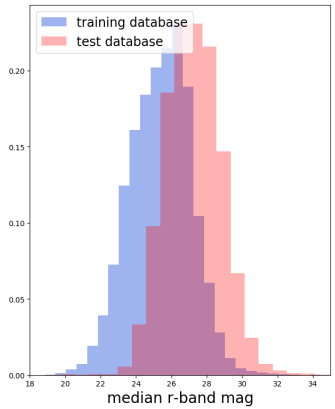
1. The Supernova Photometric Classification Challenge in 2010 (SPCC, Kessler et al.)



- 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

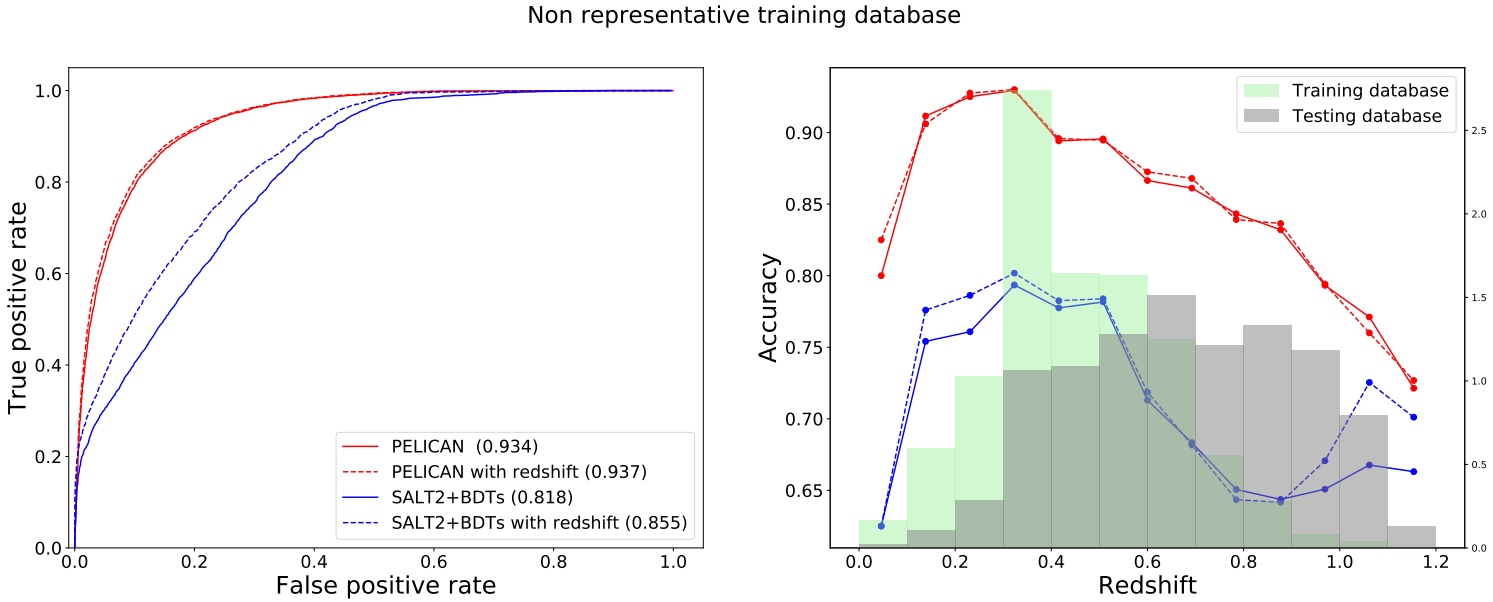
- Small training database (until 500 light curves)
- Non-representativeness between the training and the test databases due to the limitation of the spectroscopic follow-up
- Non-representativeness of the sampling and noise between main survey and deep fields



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

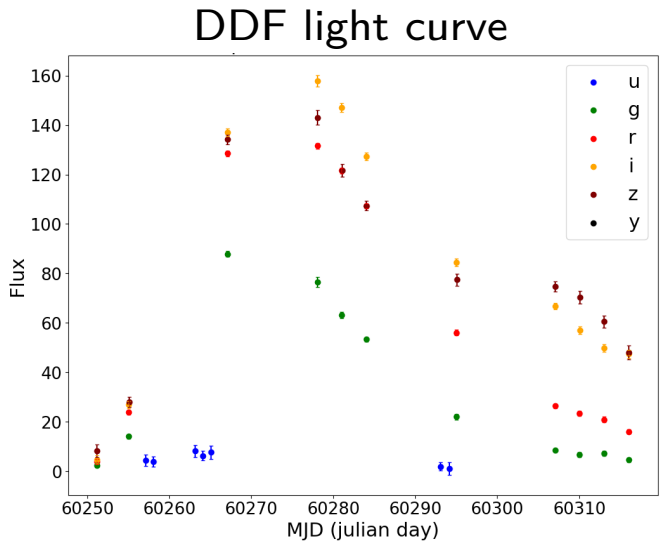


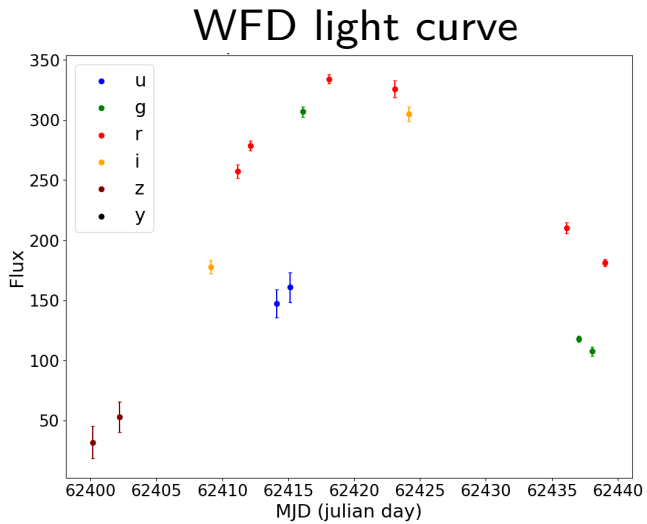
- We compared our results to BDTs classifier + SALT2 features as it is the best combination in Lochner et al. (2016)
- PELICAN obtains an accuracy of 0.856 and an AUC of 0.934 which outperforms BDTs+SALT2 method which reaches 0.705 and 0.818

LSST simulated data

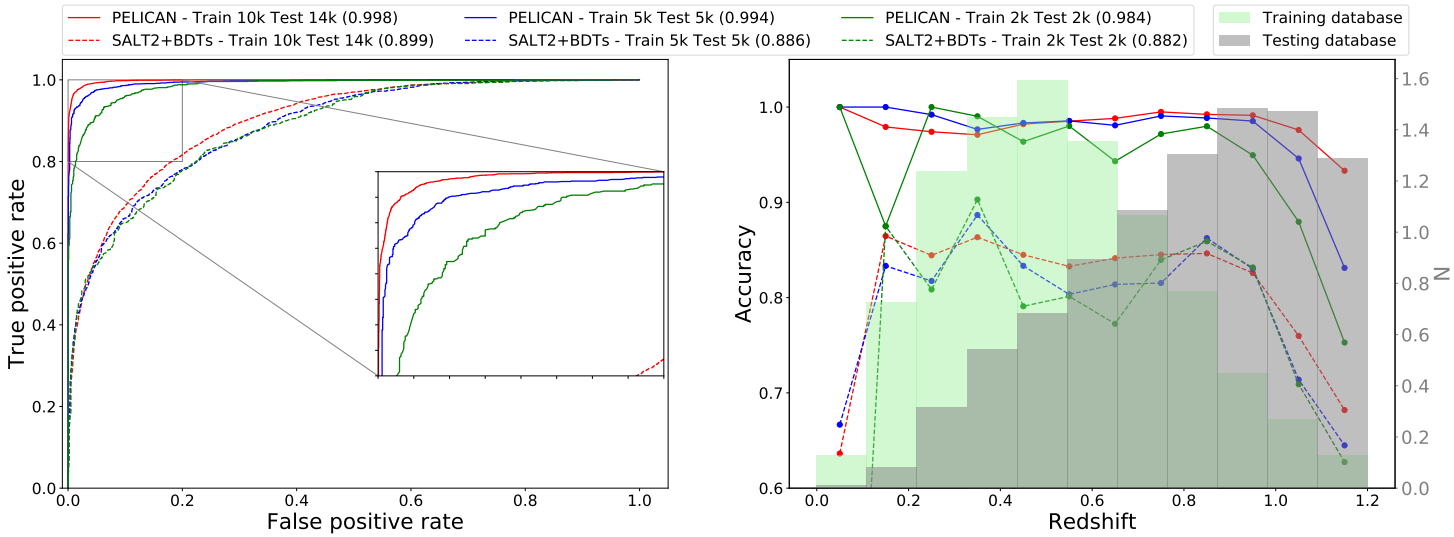
Two methodologies:

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



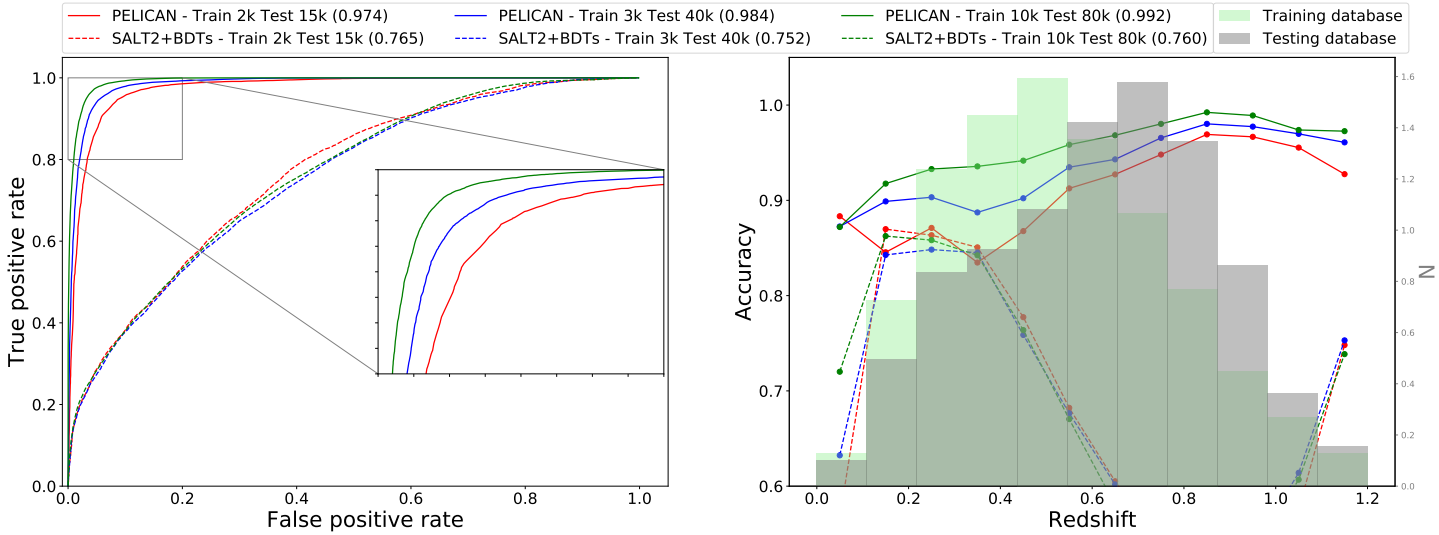


Results on DDF



	Training database (spec only)	Test database (phot only)	Accuracy	$Recall_{Ia}$ $Precision_{Ia}$> 0.95	$Recall_{Ia}$ $Precision_{Ia}$> 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)

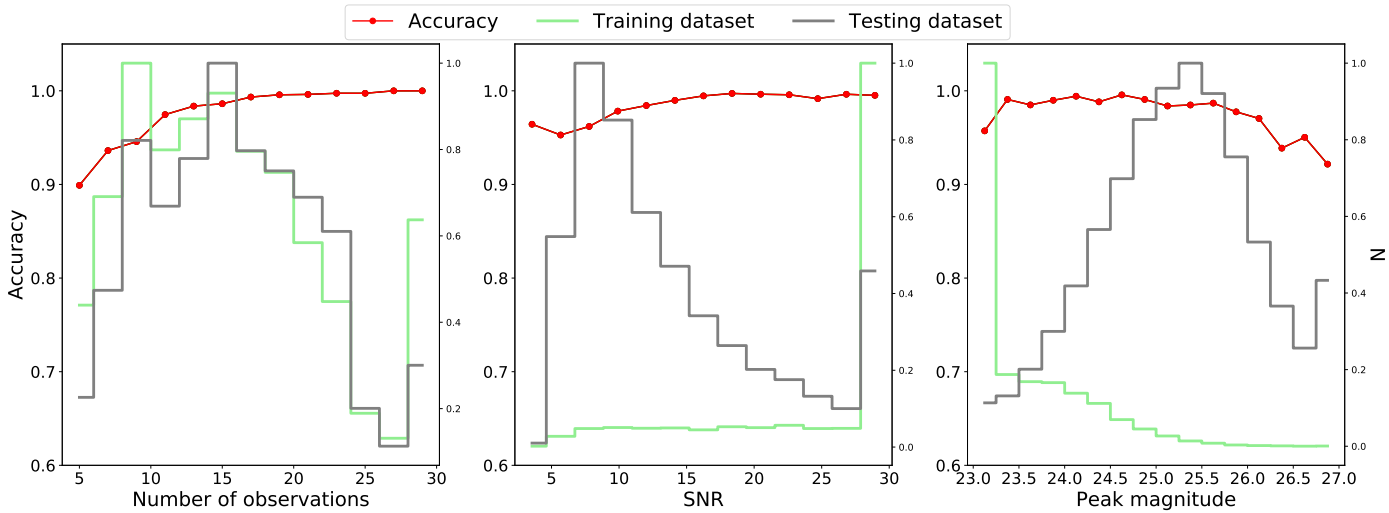
Results on WFD



	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