# Deep learning approach to predict photometric redshifts of galaxies in the Sloan Digital Sky Survey DR12

Johanna Pasquet

Centre de Physique des Particules de Marseille

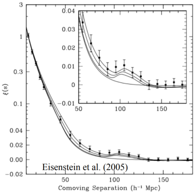
LSST France

7 November, 2018

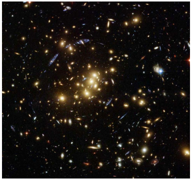
# Need accurate redshits for cosmology

Reliable redshifts are necessary to constrain the dark energy equation-of-state and to study the large scale structure of the universe

- Baryonic Acoustic Oscillations



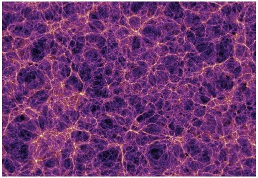

- Weak lensing



Strong gravitational lensing around galaxy cluster CL0024+17

Credit : NASA/ESA/M.J. Jee (John Hopkins University)

- Cosmic web

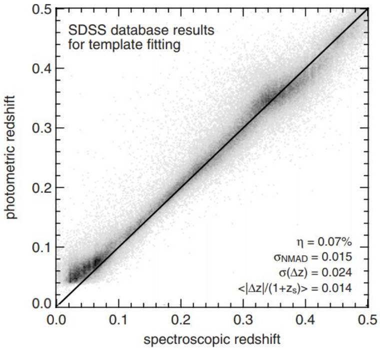


Results of a digital simulation showing the large-scale distribution of matter, with filaments and knots.

Credit: V.Springel, Max-Planck Institut für Astrophysik, Garching bei München

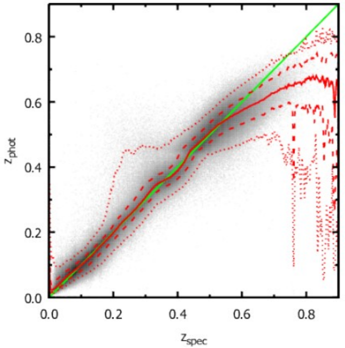
# Existing methods

A template fitting method




From Greisel et al. 2013

A machine learning method (KNN)




From Beck et al. 2016

Preliminary results with Deep Learning methods (Hoyle 2016, D'Isanto 2018)

# Photometric redshifts with Deep Learning

*Photometric redshifts from SDSS images using a Convolutional Neural Network* (**J. Pasquet**, E. Bertin, M. Treyer, S. Arnouts and D. Fouchez)

**arxiv**: 1806.06607, **code available at**: https://github.com/jpasquet/Photoz

## Key elements :

1. A representative and a complete training database with r-band magnitude $\leq 17.8$ and redshift, $z \leq 0.4$ (516,525 galaxies)
2. Photoz values + associated Probability Distribution Functions
3. Photoz immune to IQ variations and neighbours contamination
4. A dedicated Neural Network architecture

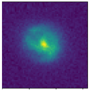
## Results obtained :

Clear improvements compared to other methods!

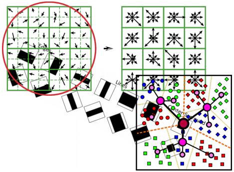
# Why use Deep Learning methods?

## Classical methods

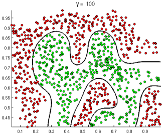
Input data



Feature crafting

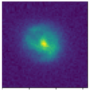


Separation with a classifier

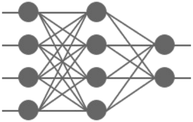


## Deep learning

Input data

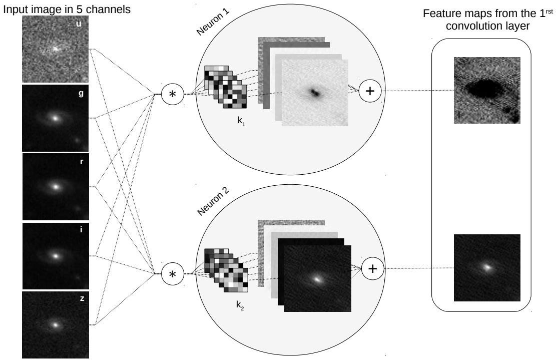


Feature learning

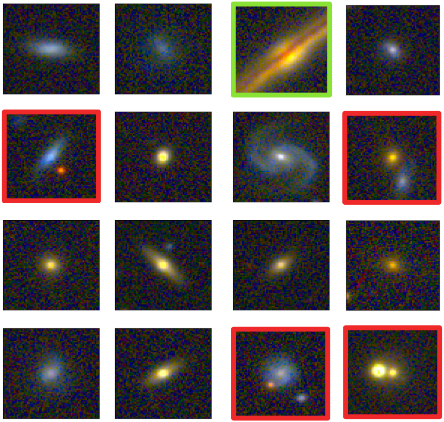


→ The best feature space representation is found by the network

# The Convolutional neural network (CNN)



# Input SDSS galaxy images transmitted to the CNN

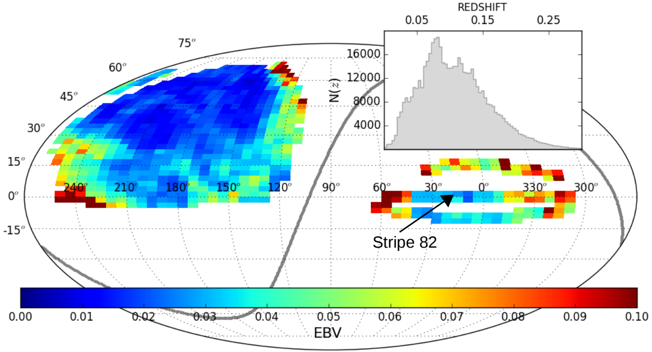


– large galaxies

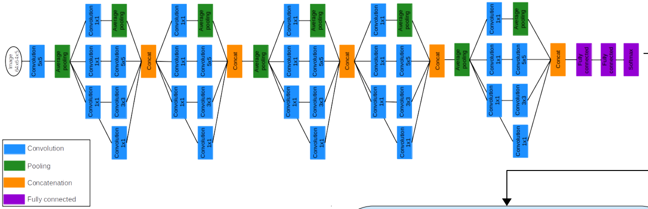
— crowded images

# Main Galaxy Sample SDSS

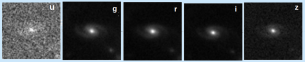
A multi-band imaging and spectroscopic redshift survey



# Our architecture

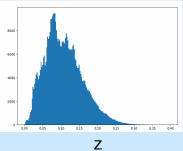



## Input images of size 64x64



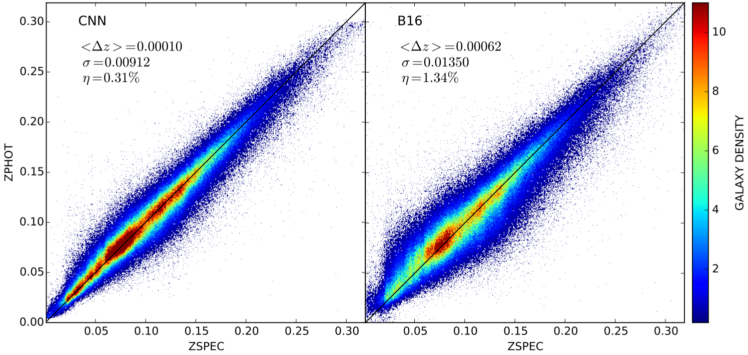

## Output probabilities

Vector of posterior probabilities for the galaxy redshift to be in a z-bin




The estimator for z-phot is the centroid of the zPDF

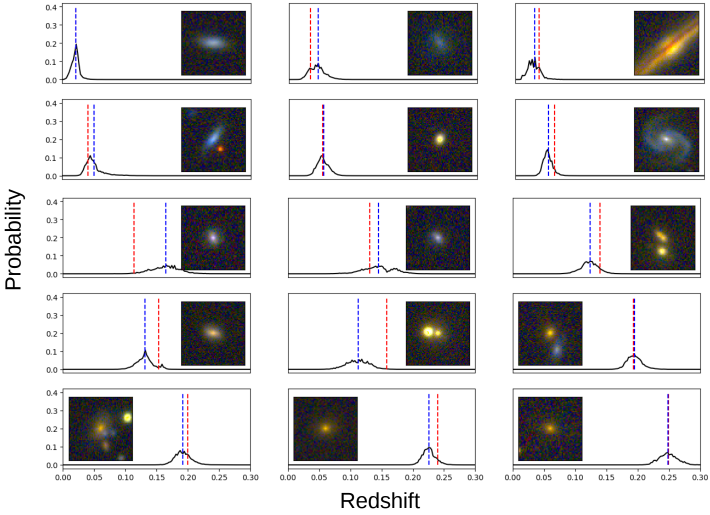
# Results of the method




| | | |
|---|---|---|
| $\langle \Delta z \rangle = 1.0\times10^{-4}$ | ← Factor of 6 improvement | $\langle \Delta z \rangle = 6\times10^{-4}$ |
| $\sigma = 9.1\times10^{-3}$ | ← 30 % improvement | $\sigma = 1.3\times10^{-2}$ |
| $\eta = 0.31\,\%$ | ← Factor of 4 improvement | $\eta = 1.35\,\%$ |

$$\Delta z = (z_{phot} - z_{spec})/(1+z_{spec})$$
$$\sigma = 1.4826 \times MAD$$
$$MAD = Median(|\Delta z - Median(\Delta z)|)$$
$$\eta = |\Delta z| > 0.05$$

# Examples of PDFs



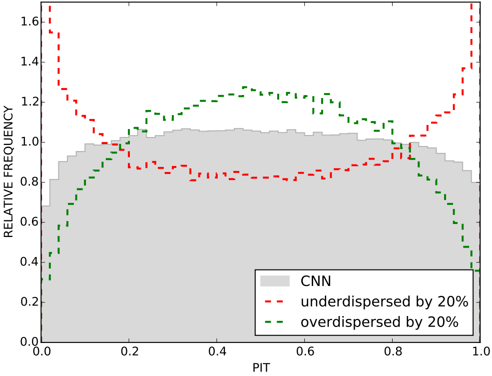

-- Spectroscopic redshift

-- Photometric redshift

# Assess the prediction quality of our PDFs

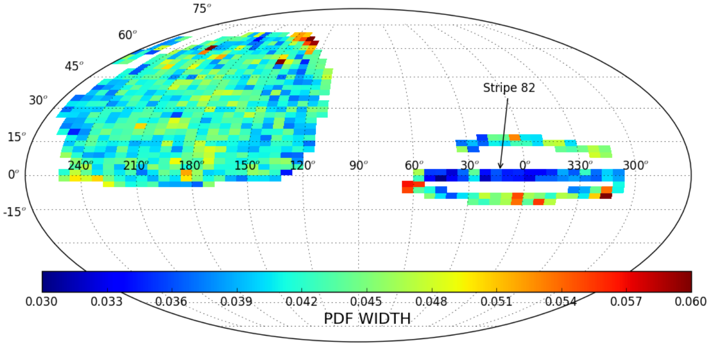
The PIT statistic (Dawid 1984) is based on the histogram of the cumulative probabilities at the true value. For galaxy $i$ with spectroscopic redshift $z_i$ in the test sample :

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



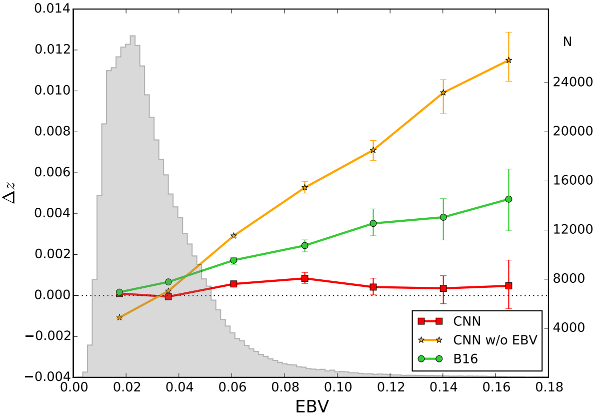
# Impact of Signal-to-Noise Ratio (SNR) on widths of PDFs

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



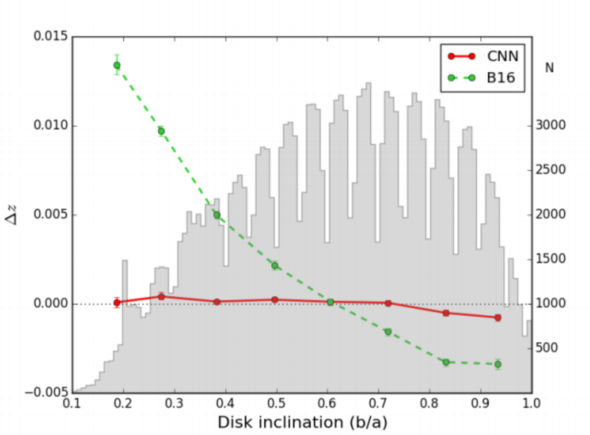
# Impact of the extinction of our Galaxy on photometric redshifts

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



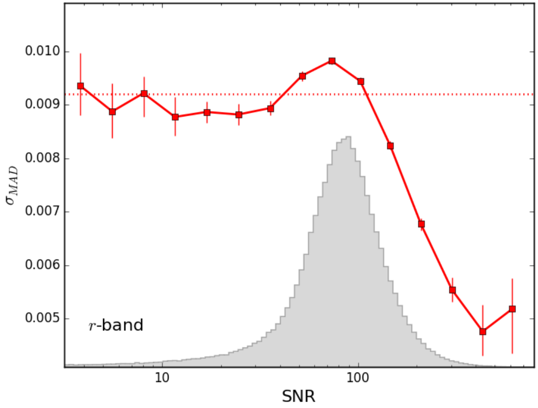
# Impact of the disk inclination of galaxies on photometric redshifts

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



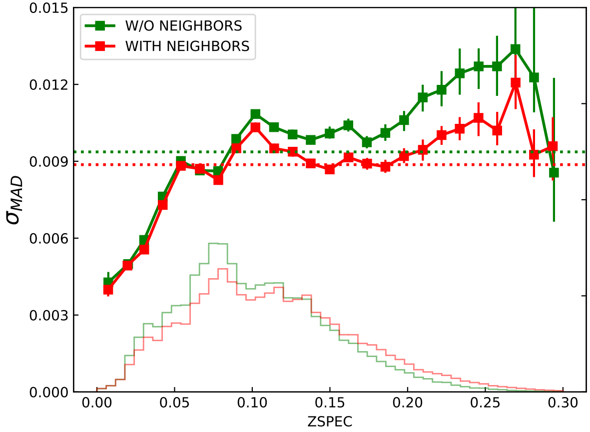
# Impact of the SNR on the performance

$\sigma_{MAD}$ decreases with the signal-to-noise ratio (SNR), achieving values below 0.007 for SNR $>$ 100, as in the deep stacked region of Stripe 82



# Neighboring galaxies

The MAD deviation is significantly improved for galaxies with fainter neighbors (43%) compared to those without

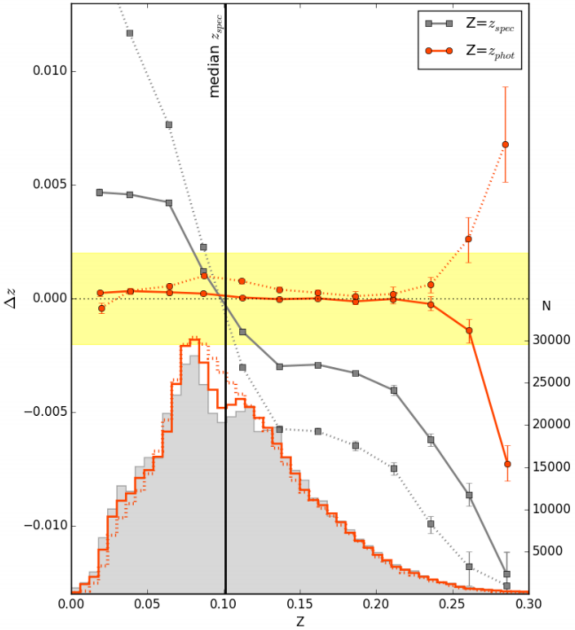


# Summary results

| Trial | training sample size | bias | $\sigma$ | $\eta$ |
|---|---|---|---|---|
| **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$^\star$ | 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 |

# Summary

- We developed a **Deep Convolutional Neural Network** (CNN) used as a classifier to estimate photometric redshifts and their associated PDFs.
- Our work shows **significant improvements** for:
  - the **dispersion** of photometric redshifts, $\sigma_{MAD}$
  - the **PDFs** that are **well calibrated**
  - **no measurable bias** with the reddening and the inclination of galaxies
- A high SNR tends to improve the results
- This work opens very **promising perspectives** for the exploitation of large and deep photometric surveys which encompass a larger redshift range and where spectroscopic follow-up is necessarily limited
- Next work: extend this method to higher redshifts and consider a non-representative training database **(see my talk tomorrow!)**


median $z_{spec}$
Z=$z_{spec}$
Z=$z_{phot}$
0.010
0.005
0.000
−0.005
−0.010
$\Delta z$
N
30000
25000
20000
15000
10000
5000
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Z

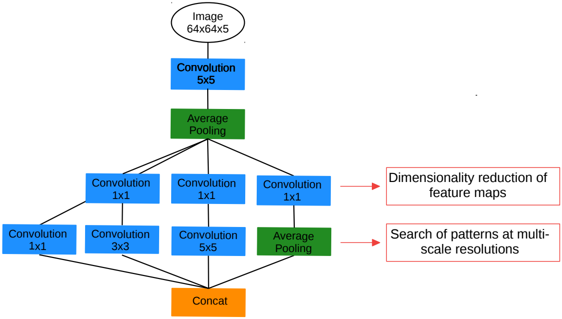
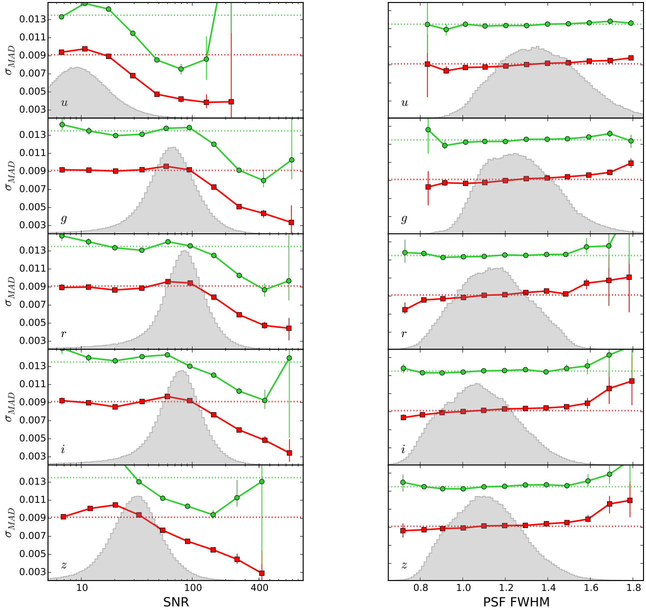


Image
64x64x5
Convolution
5x5
Average
Pooling
Convolution
1x1
Convolution
1x1
Convolution
1x1
Convolution
1x1
Convolution
3x3
Convolution
5x5
Average
Pooling
Concat
Dimensionality reduction of feature maps
Search of patterns at multi-scale resolutions

| Trial | training sample size | size of 1 test sample | bias | $\sigma_{\mathrm{MAD}}$ | $\eta$ | $< CRPS$ |
|---|---|---|---|---|---|---|
| **Training with 80% of the dataset** | 393,219 | | | | | |
| Full test sample | | 103,306 | 0.00010 | 0.00912 | 0.31 | 0.00674 |
| (B16) | | (103,306) | (0.00062) | (0.01350) | (1.34) | |
| Suspect zone (SZ) removed | | 101,499 | 0.00004 | 0.00908 | 0.31 | 0.00672 |
| Widest 10% of PDFs | | 91,543 | 0.00006 | 0.00848 | 0.09 | 0.00606 |
| Widest 20% of PDFs | | 79,897 | 0.00005 | 0.00789 | 0.06 | 0.00556 |
| Stripe 82 only | | 3,943 | -0.00009 | 0.00727 | 0.34 | 0.00574 |
| Stripe 82 with widest 20% of PDFs removed | | 3,131 | 0.00004 | 0.00635 | 0.09 | 0.00467 |
| Training with 50% of the dataset⋆ | 250,000 | 252,500 | 0.00007 | 0.00910 | 0.29 | 0.00672 |
| Training with 20% of the dataset | 99,001 | 385,970 | -0.00001 | 0.00914 | 0.30 | 0.00677 |
| Training with 2% of the dataset | 10,100 | 434,228 | -0.00017 | 0.01433 | 1.26 | 0.01009 |
| Training on Stripe 82 | 15,771 | | | | | |
| Stripe 82 removed⋆ | | 478,274 | 0.00194 | 0.01341 | 1.15 | 0.00988 |
| Stripe 82 only | | 3,942 | -0.00002 | 0.00795 | 0.38 | 0.00622 |
| Training w/o Stripe 82 | 486,560 | | | | | |
| Stripe 82 removed⋆ | | 97,607 | 0.00000 | 0.00914 | 0.33 | 0.00680 |
| Stripe 82 only⋆ | | 19,714 | -0.00077 | 0.00760 | 0.41 | 0.00606 |


$\sigma_{MAD}$
0.013
0.011
0.009
0.007
0.005
0.003
u
g
r
i
z
10
100
400
SNR
0.8
1.0
1.2
1.4
1.6
1.8
PSF FWHM