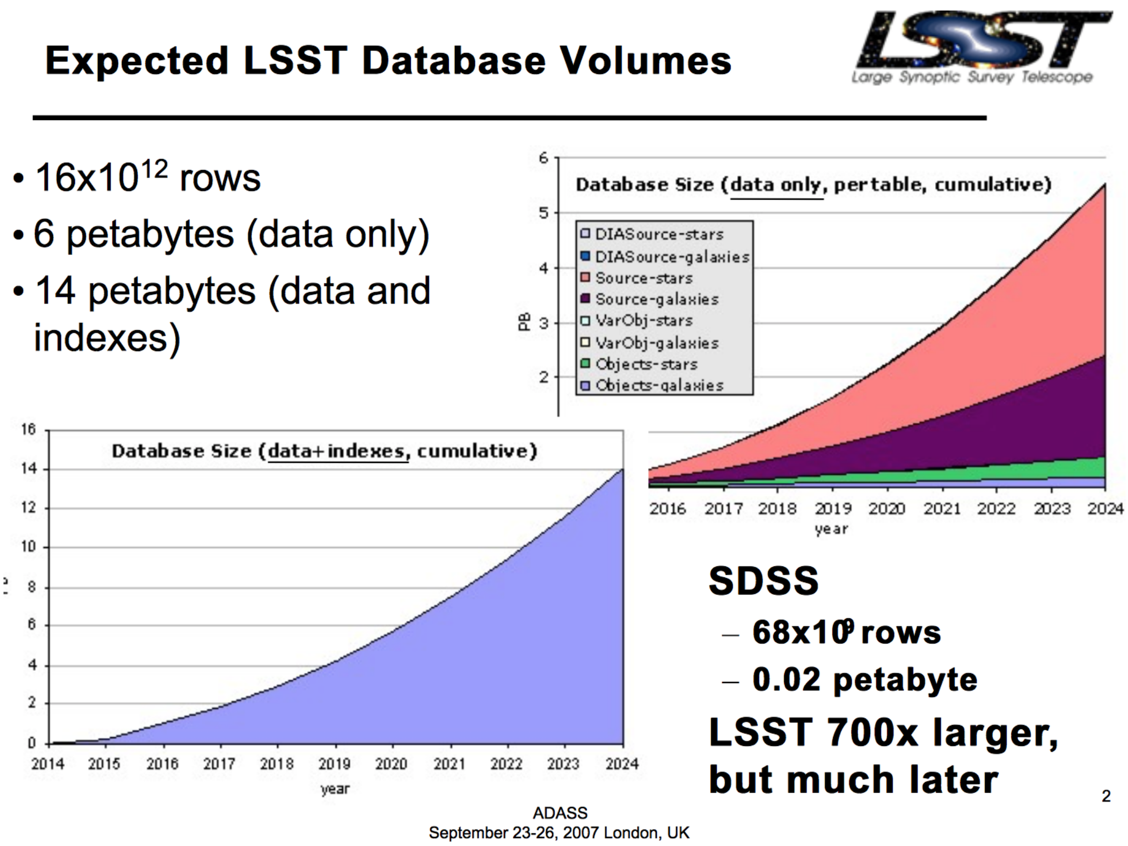
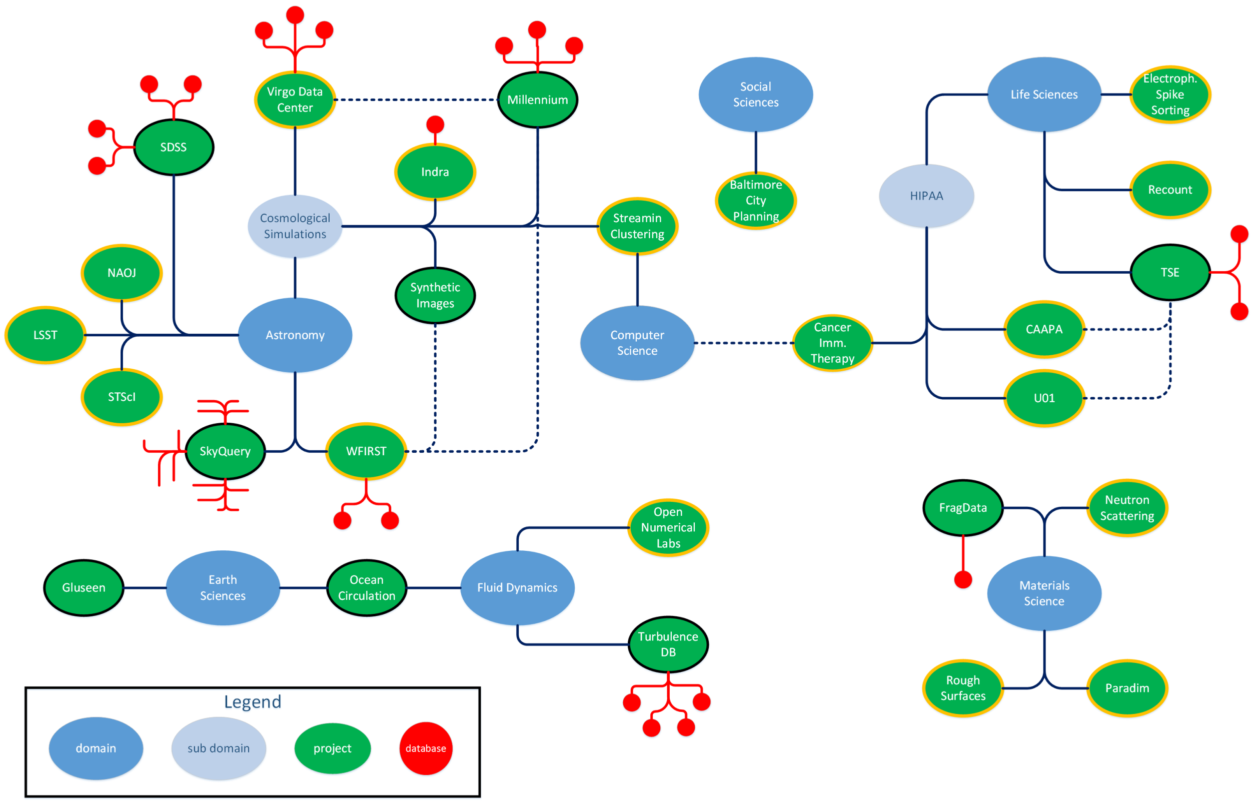
Galaxy classification and redshift determination using neural networks and the OTELO survey

José A. de Diego

6/9/2018

Why Deep Neural Networks?





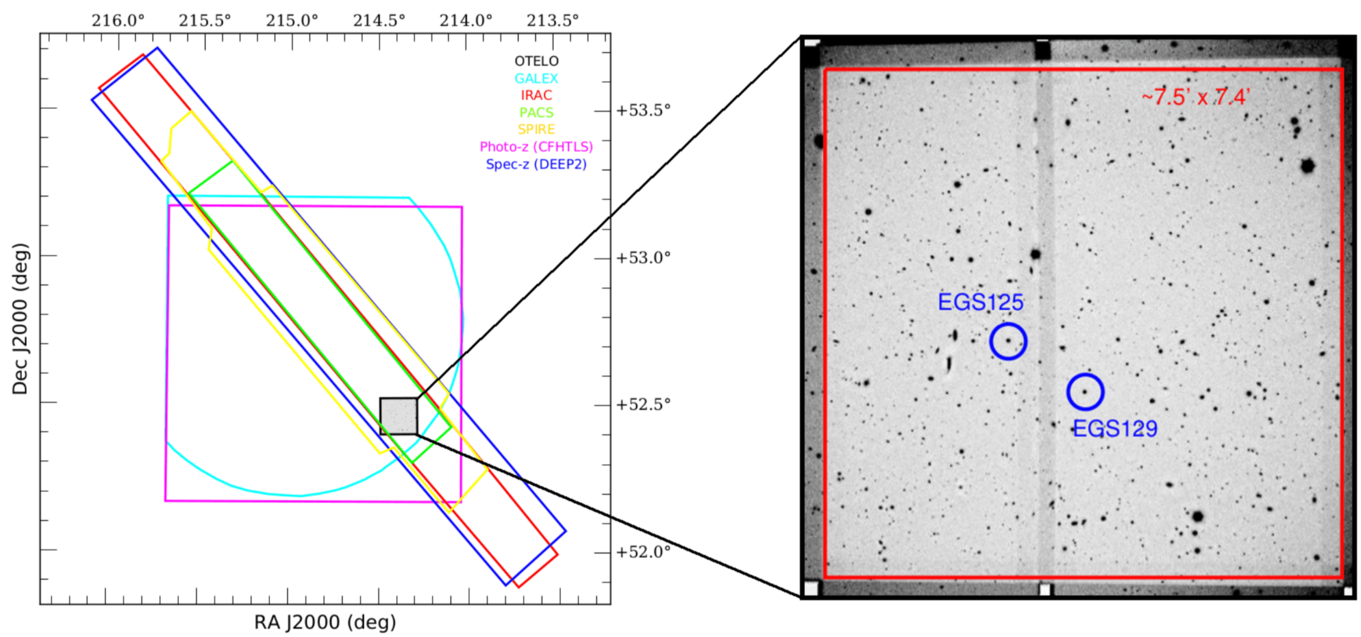
Domínguez Sánchez H, Huertas-Company M, Bernardi M, Tuccillo D, Fischer JL. 2018.*Improving galaxy morphologies for SDSS with Deep Learning.* MNRAS 476: 3661-76

Table 2. Precision and recall (TPR) values for different P_{thr} and average accuracy for the questions which have two possible answers in GZ2 classification scheme.

Question	Meaning	P_{thr}	TPR	Prec.	Acc.
		0.2	0.97	0.91	
Q1	Disc/features	0.5	0.95	0.96	0.98
		0.8	0.90	0.99	
		0.2	1.00	0.67	
Q2	Edge-on	0.5	0.99	0.83	0.97
		0.8	0.92	0.95	
		0.2	0.93	0.48	
Q3	Bar sign	0.5	0.79	0.80	0.97
		0.8	0.58	0.92	
		0.2	0.98	0.54	
Q6	Merger signature	0.5	0.96	0.82	0.97
		0.8	0.90	0.97	

The OTELO Survey | OSIRIS Tunable Emission Line Object Survey

- GTC → OSIRIS-TF → OTELO
- Deepest emission line survey
  - Over 11,000 galaxies
  - Field of $7.5' \times 7.4'$
  - 2D-spectroscopy
    - * R $\sim$ 700
    - * $\lambda_c = 9175\text{Å},\ \ \Delta\lambda = 210\text{Å}.$
  - Ancillary data:
    - * Extended Groth Field
    - * CFHT-Legacy Survey
    - * GALEX…



Deep Neural Networks

- DNNs consist of a number of interconnected layers of *neurons* (processors) that perform calculations.
- Information is stored in the *weights* on the connections.

Descriptive statistics

Sérsic indices

$$I(R) = I_e\, e^{-b\left[\left(\frac{R}{R_e}\right)^{1/n} - 1\right]},$$

```
Gal. Type   Sersic Index   Sd
```

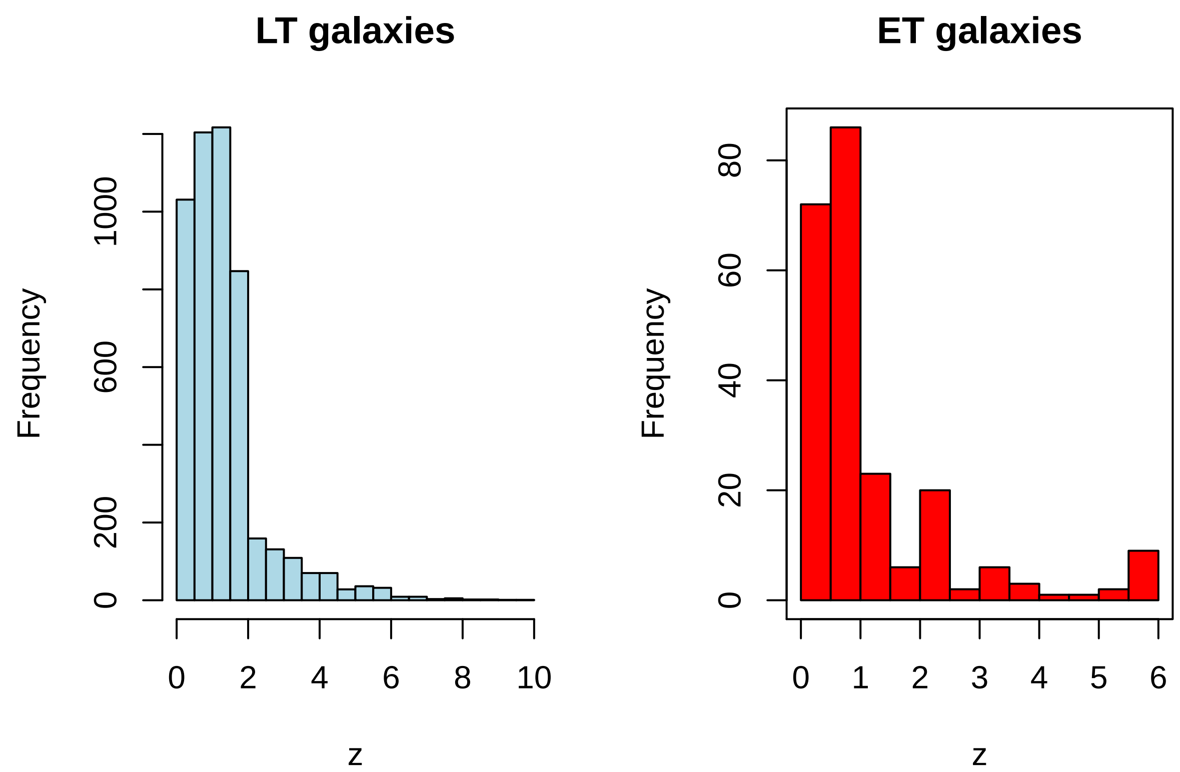


Figure 1: Histograms for lone galaxies in OTELO.

```
LT          1.11            1.93
ET          3.72            1.36

            Proportion 0.10 ET
```

Training data:

Best color separation choice

```
        g-i train
OTELO   ET  LT Sum
  ET    57  19  76
  LT    33 851 884
  Sum   90 870 960


g-i classification:
Error rate  0.0542
Accuracy    0.9458


Best ET/LT separation:  g-i = 2.15
```

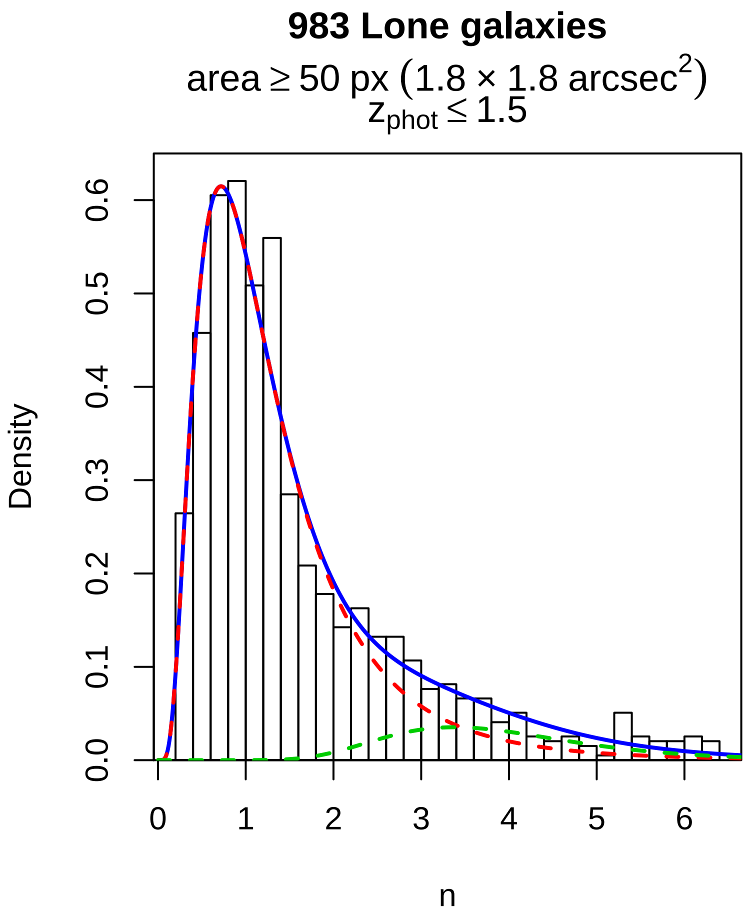


Figure 2: Lone ET & LT galaxies: Lognormal fit.

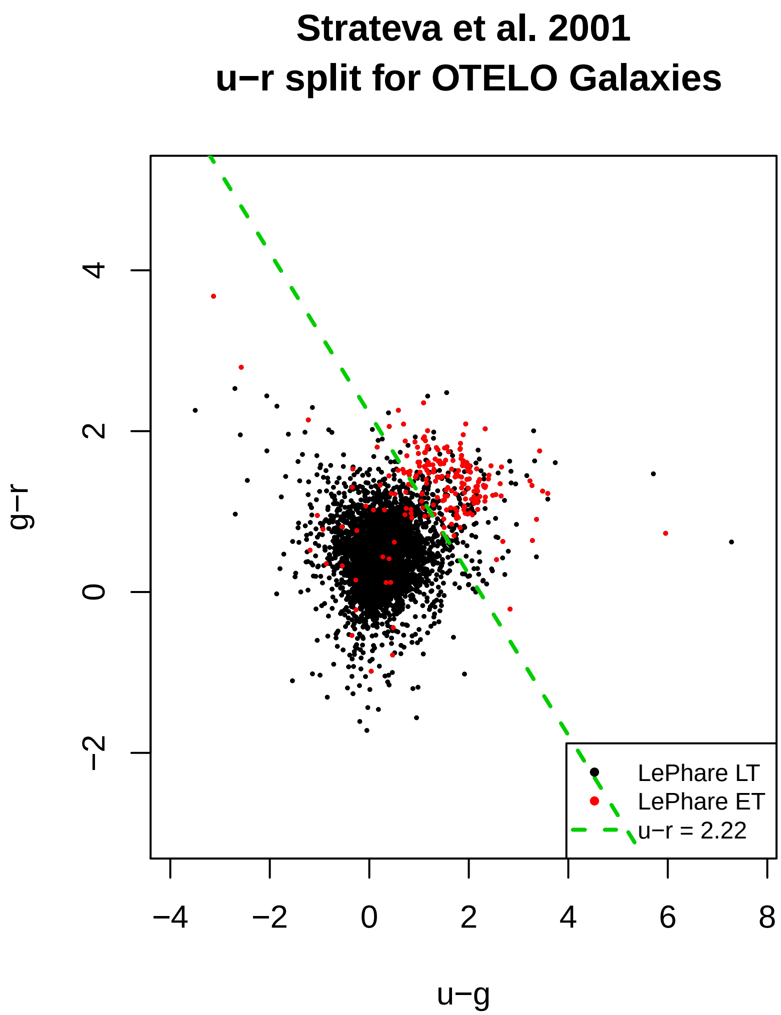


Figure 3: Strateva et al. 2001 plot.

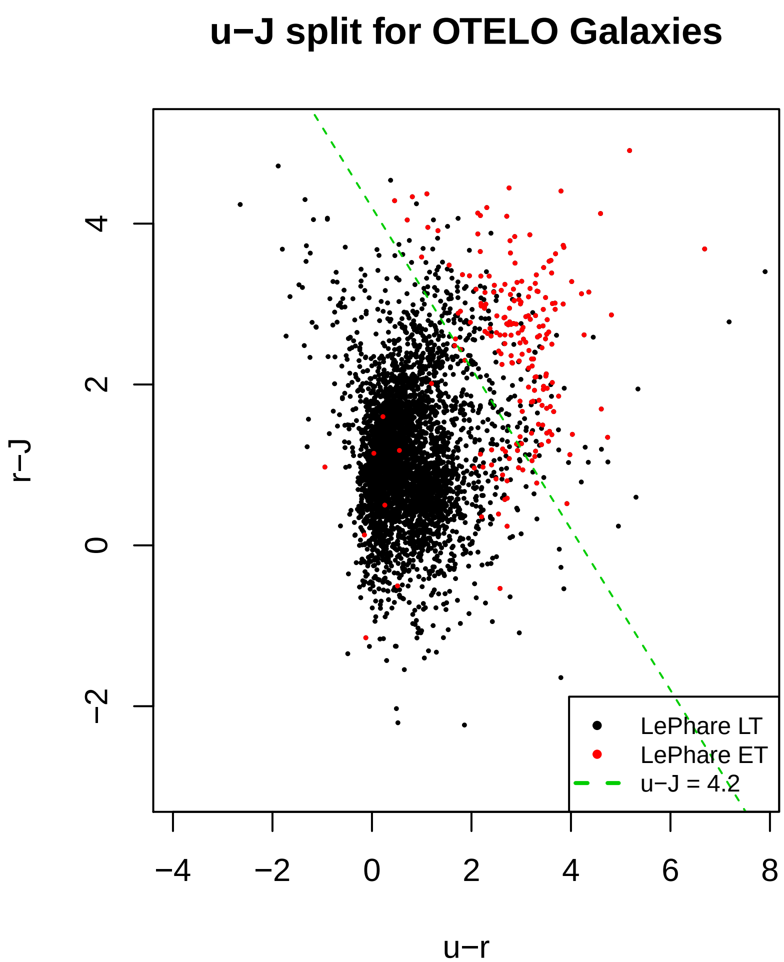


Figure 4: u-J plot.

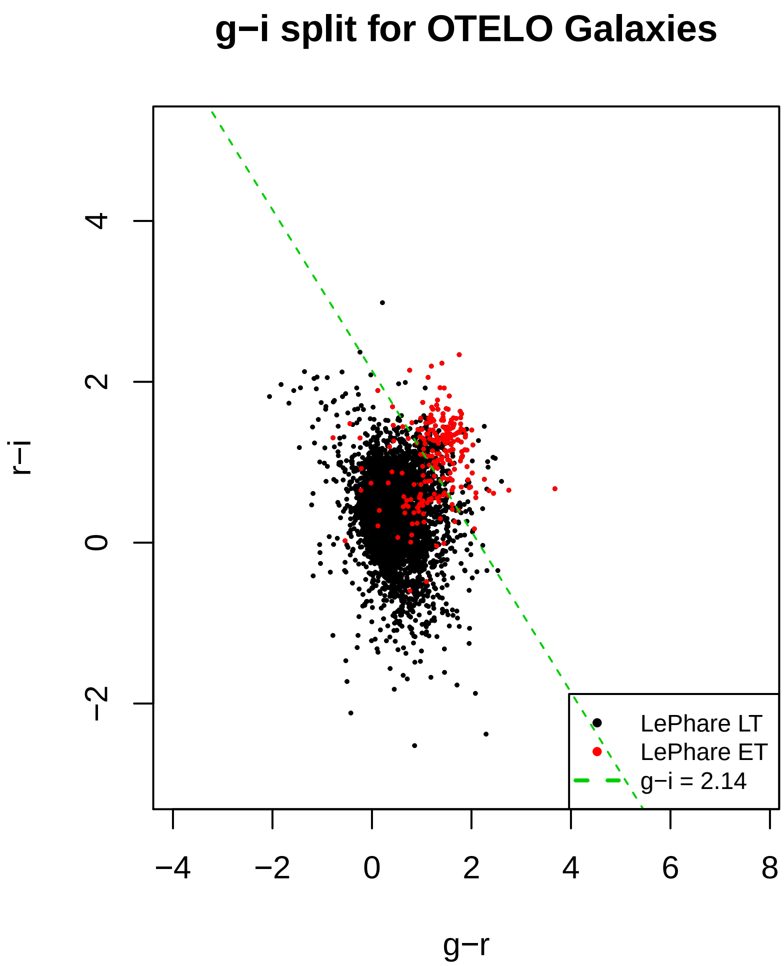


Figure 5: g-i plot.

Evaluate model for g-i test data

```
Confusion matrix (absolute):
          Actual
Prediction  ET  LT Sum
       ET   13   9  22
       LT    4 214 218
       Sum  17 223 240

Confusion matrix (relative):
          Actual
Prediction   ET   LT  Sum
       ET  0.05 0.04 0.09
       LT  0.02 0.89 0.91
       Sum 0.07 0.93 1.00

Accuracy:
0.9458 (227/240)

Error rate:
0.0542 (13/240)

Error rate reduction (vs. base rate):
0.2353 (p-value = 0.1912)


Base Rate: 0.9292
```

Deep Neural Network:Galaxy Classification

Objectives of DNN classification

- Get as close as possible to OTELO classification
- Improve baselines classifications
  - Base rate classification 0.9292
  - Color g-i based classification 0.9458

Data set properties

- OTELO subset
  - Lone galaxies
  - Iso-area image ≥ 50
- Data variables
  - dmodelu, dmodelg, dmodelr, dmodeli, dmodelz, dmodelj, dmodelh, dmodelks, N_GALFIT_BAND_1, galclass
- Data sample size: 4241
  - Data training size: 2968
    - Data validation size: 20% training
  - Data testing size: 1273

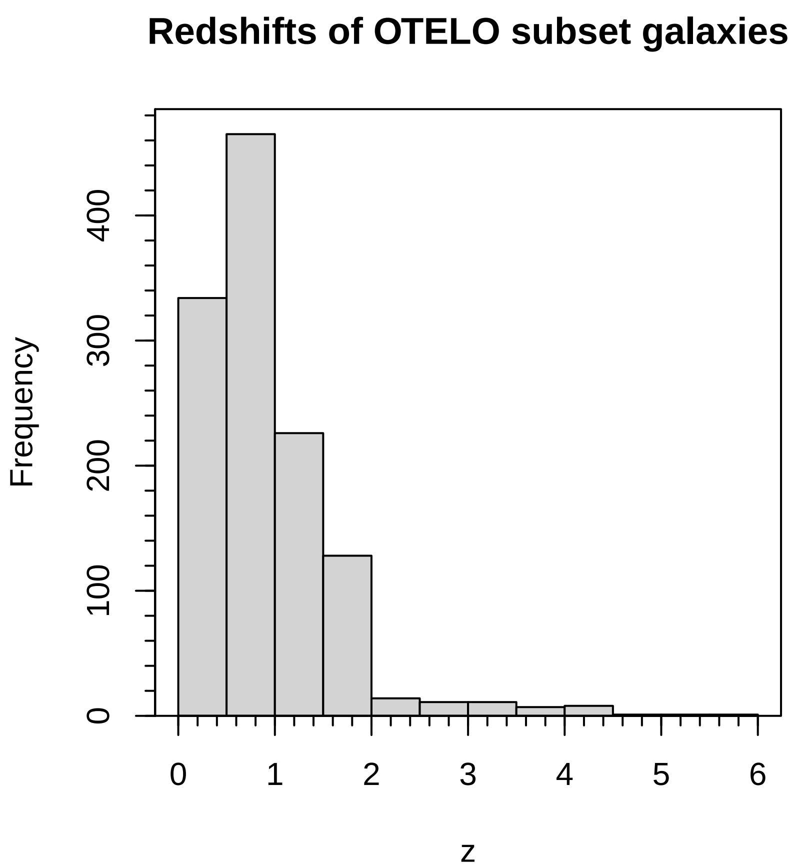


Figure 6: DNN Sample

DNN architecture

```
___________________________________________________________________________
Layer (type)                     Output Shape                  Param #
===========================================================================
dense_1 (Dense)                  (None, 32)                    320
___________________________________________________________________________
dense_2 (Dense)                  (None, 32)                    1056
___________________________________________________________________________
dense_3 (Dense)                  (None, 1)                     33
===========================================================================
Total params: 1,409
Trainable params: 1,409
Non-trainable params: 0
___________________________________________________________________________
```

Relu / Sigmoid activations

- DNN compilation
  - Loss function: binary_crossentropy
  - Optimizer: optimizer_rmsprop
  - Metrics: accuracy
- Fit
  - Epochs:150, Batch size: 128, Validation: 20% of training sample

---

Random sample from 11457 test galaxies

```
OTELO        DNN Prob

    1        1   0.986551
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    1        1   0.902308
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
    0        0   0.000000
```

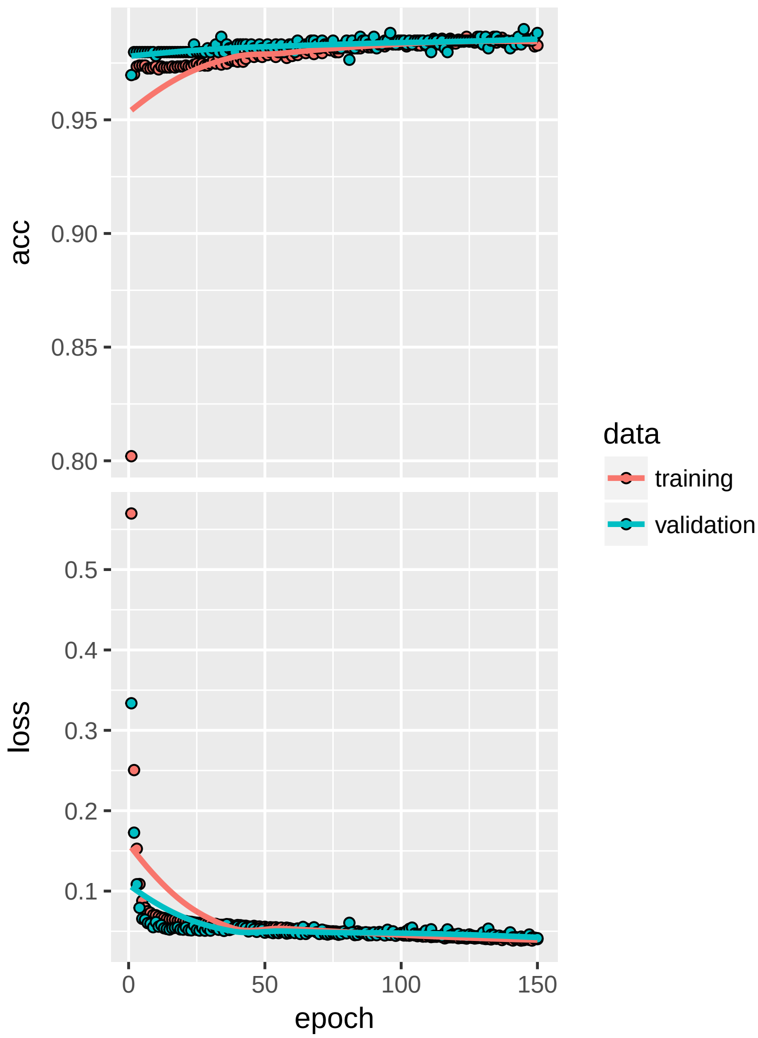


Figure 7: Accuracy & Loss

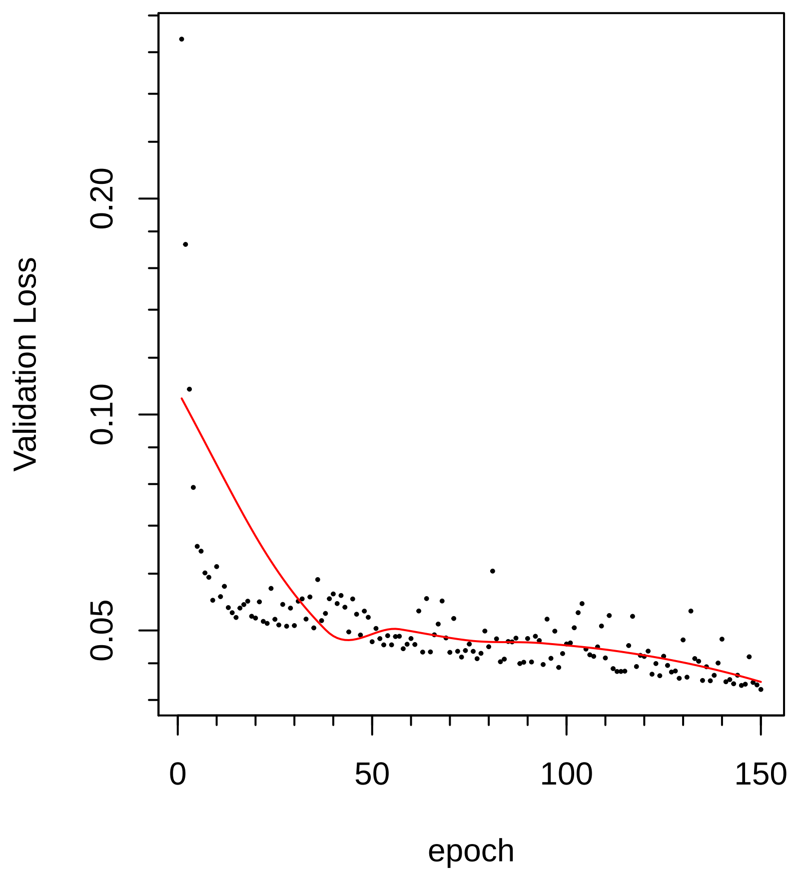


Figure 8: Validation Loss

0	0	0.000000
0	0	0.000000
1	1	0.997628
1	1	0.959253
1	1	0.995208
1	1	0.985048
0	0	0.000000
0	0	0.000000
0	0	0.000000
0	0	0.000000

---

21 Mismatches

OTELO	DNN	Prob
1	0	0.410606
0	1	0.525443
1	0	0.309931
0	1	0.984879
0	1	0.624495
0	1	0.802046
1	0	0.272197
0	1	0.906919
1	0	0.355655
1	0	0.324490
0	1	0.992975
0	1	0.902258
0	1	0.990199
0	1	0.954445
0	1	0.834345
0	1	0.899426
0	1	0.596614
1	0	0.426365
1	0	0.422188
0	1	0.955578
0	1	0.998744

	OTELO class		
DNN	ET	LT	Sum
ET	346	14	360
LT	7	906	913
Sum	353	920	1273

Comparison with baselines

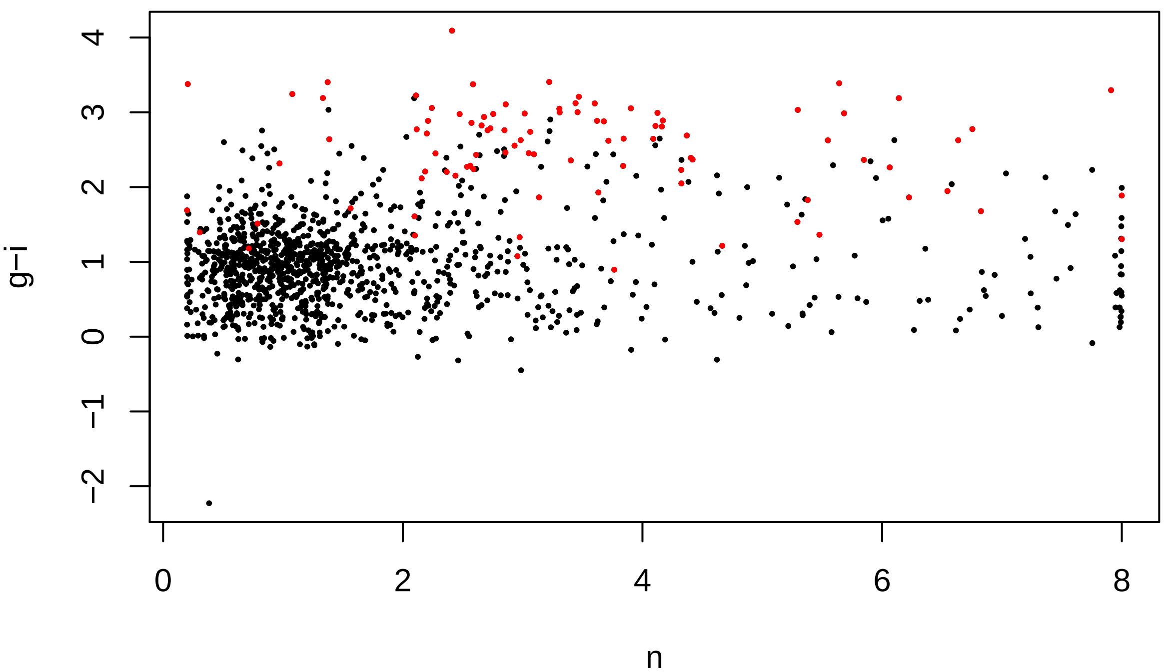
- Baselines classifications
  - Base rate classification 0.9292
  - Color g-i based classification 0.9458
  - DNN classification 0.9835

Similar to:

- Domínguez Sánchez H, Huertas-Company M, Bernardi M, Tuccillo D, Fischer JL. 2018. *Improving galaxy morphologies for SDSS with Deep Learning.* MNRAS 476: 3661-76

using a *convolutional neural network* on $\sim 670,000$ SDSS images.

The importance of including Sérsic index



Deep Neural Network:Photometric redshifts

Data set properties

- Galaxies with spectroscopic redshift: 370
- Data variables:
  - idobj, dmodelu, dmodelg, dmodelr, dmodeli, dmodelz, dmodelj, dmodelh, dmodelks, ZSPEC
- Data trainging size: 259 (70%)
  - Data validation size: 20% training
- Data test size: 111(30%)

```
Metrics
Mean Absolute Error
MAE 0.1121

Loss
Root Mean Square for Proportions
RMS 0.0211

PHOTOMETRIC-SPECTROSCOPIC REDSHIFTS
Residuals    Sd      Sample Size
- - - - - - - - - - - - - - - - - -
OTELO        0.24    370
```

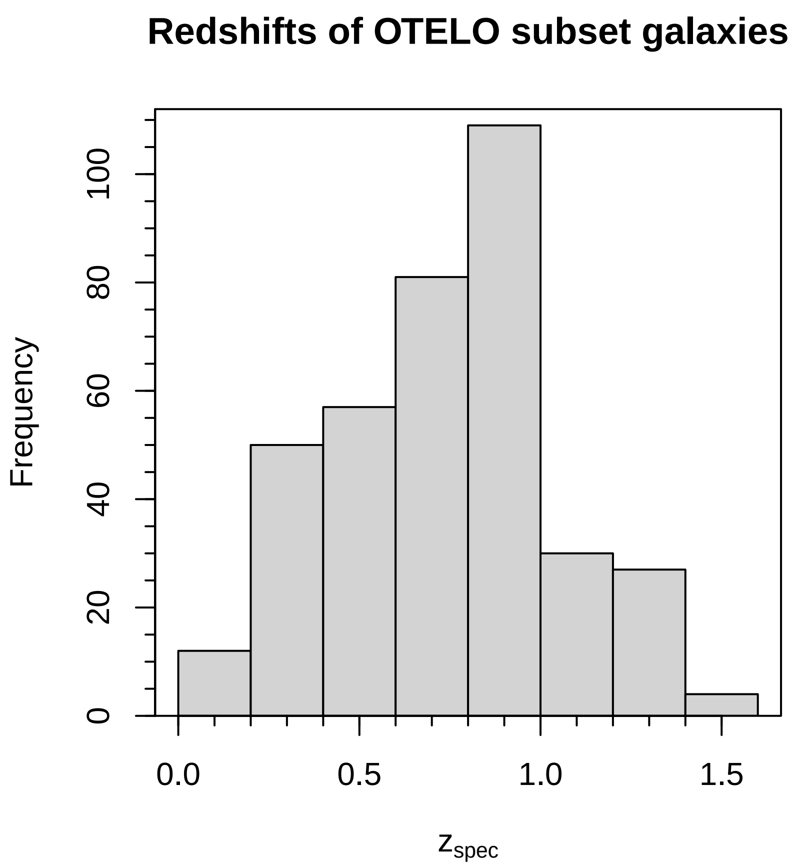


Figure 9: DNN Sample

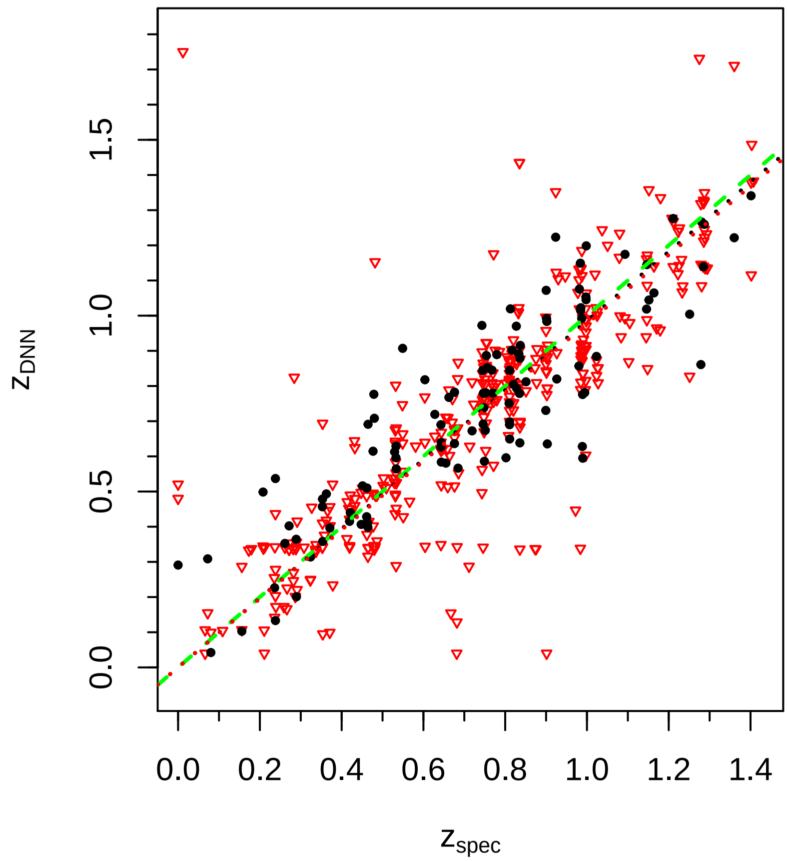


Figure 10: DNN vs Spectroscopic redshifts

```
DNN          0.15    111
```

Conclusions

- We are getting into the Petabyte Large Survey Epoch.
  - We need new analysis tools.
- DNNs offer simple and reliable classification / regression results.
  - Not only for images.
  - In oposite to parametric fitting techniques:
    - DNNs *learn* from previous training samples.

Open question

- How do we can use DNNs for GTC — GTM synergy?