Machine learning module for Fink Broker

Marco LEONI, in collaboration with E. ISHIDA, J. PELOTON

Orsay

3 Feb 2020

Acknowledgments

E. ISHIDA

J. PELOTON

ZTF simulated data : *courtesy of Daniel Muthukrishna*

Machines : Cloud@VirtualData

Funding :

References

Ishida et al., Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning, *MNRAS 2019*

Muthukrishna et al., RAPID: Early Classification of Explosive Transients using Deep Learning, https://arxiv.org/abs/1904.00014

MENU :

- Fink Broker and “Active” Machine Learning (ML) for SuperNova (SN) classification
- ML on simulated data
- Applying models to the observations
- Open issues/future directions

Introduction

G. Galilei, 1604

DIALOGO
DE CECCO DI RONCHITTI
DA BRVZENE.
IN PERPVOSITO
DE LA STELLA NVOVA.
Al Lostrio e Rebelendo Segnor Antuogno Squerengo degnetissemo Calonego de Pava, sò Paròn.
Con alcune ottave d'Incerto, per la medesima Stella,
IN PADOVA,
Appresso Pietro Paulo Tozzi. M.DC.V.

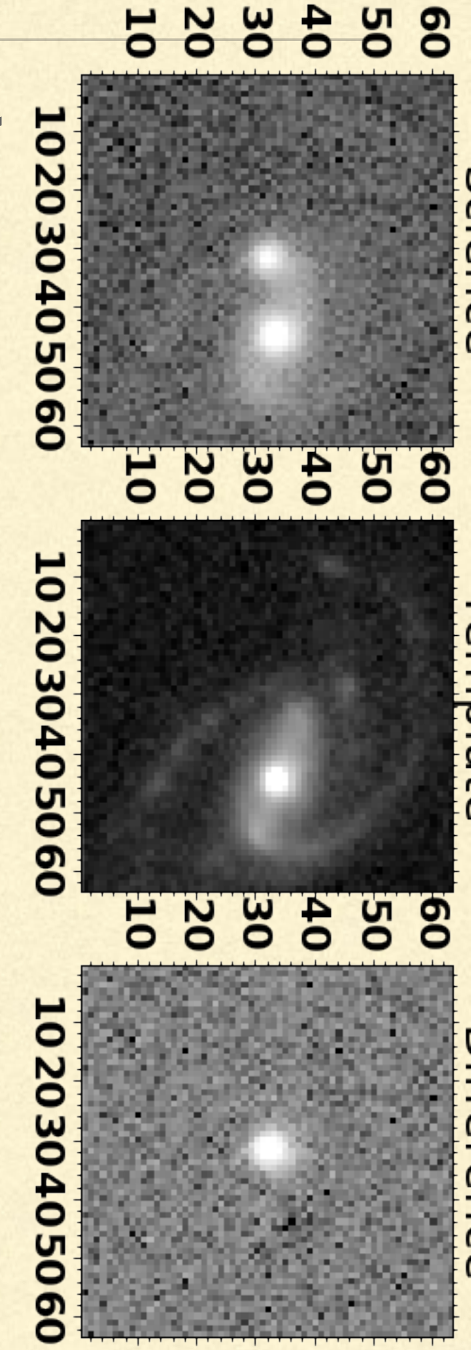
- Today known a few thousands *type Ia SN* (up to 2015 http://www.cbat.eps.harvard.edu/lists/Supernovae.html : 3000 *type Ia SN* out of 6500 *SN*)
- LSST Telescope data : approx. *15Tb* per night

Intro : from raw images to photometric data

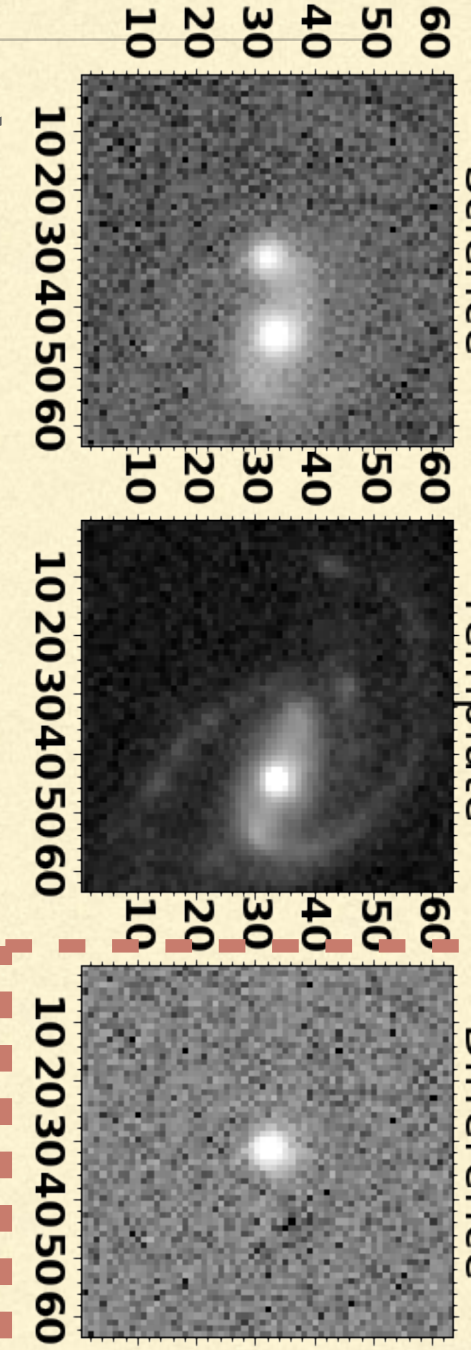
1. Data from telescope

2. Compare to existing database

3. Do these images contain **new** information?



Intro : from raw images to photometric data

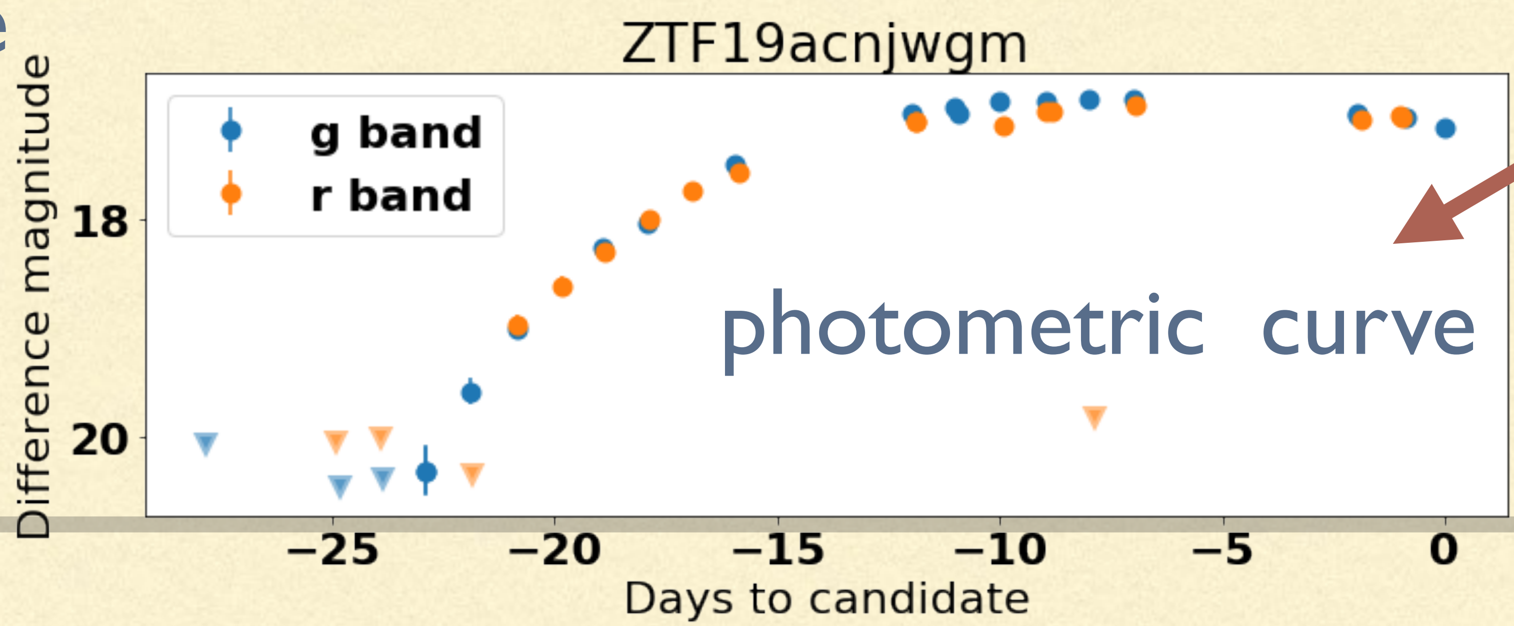


1. Data from telescope

2. Compare to existing database

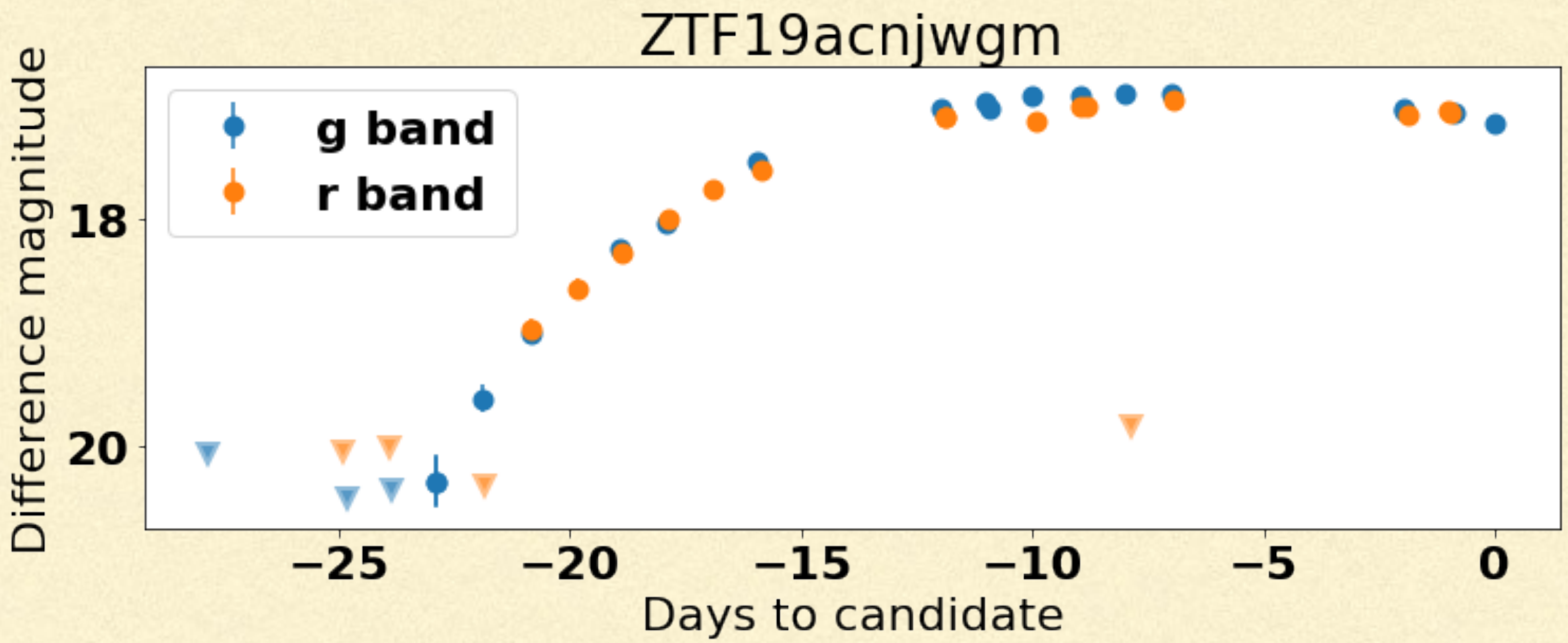
3. Do these images contain **new** information?

If yes it will be processed



photometric curve

Intro : focus on photometric data



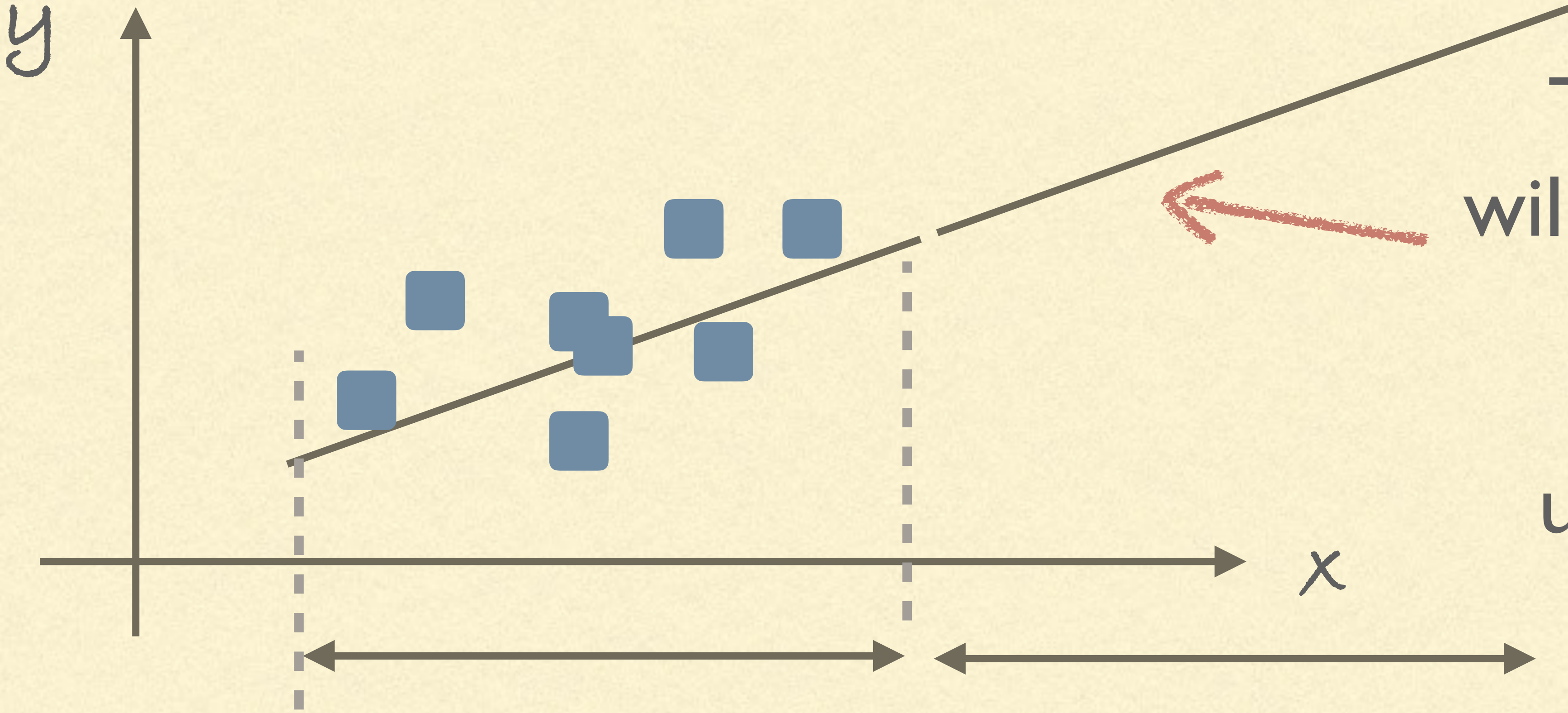
Question :

can one make
automated
predictions by looking at few first data points ?

Goal :

early
discovery of *type Ia SN* (fundamental for cosmology)

Machine Learning :
a generalisation of regression (data fit)

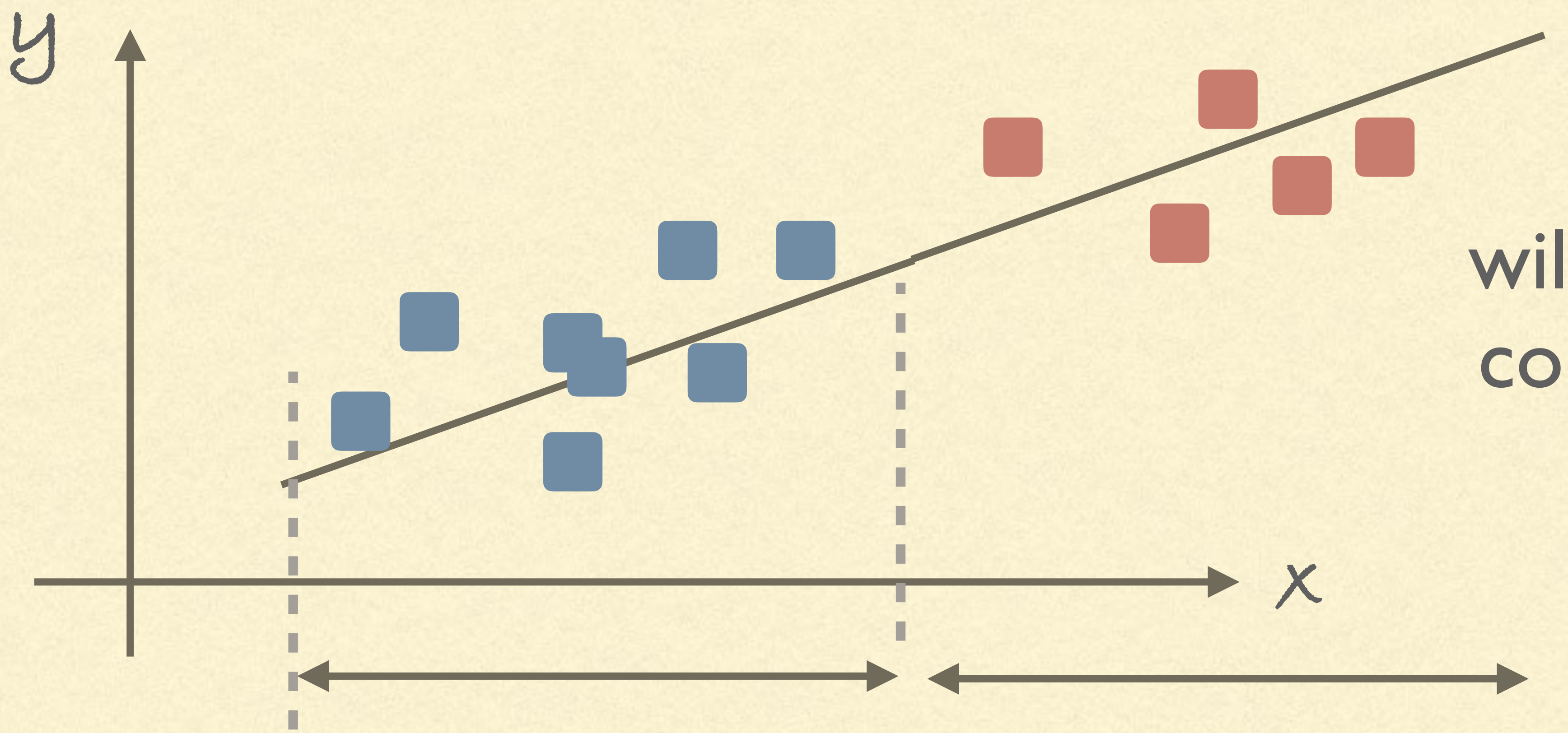


The fitted slope
will be used to make
predictions
concerning
unseen data sets

Training set
(or in sample)

Test set
(or out of sample)

Machine Learning : Prediction

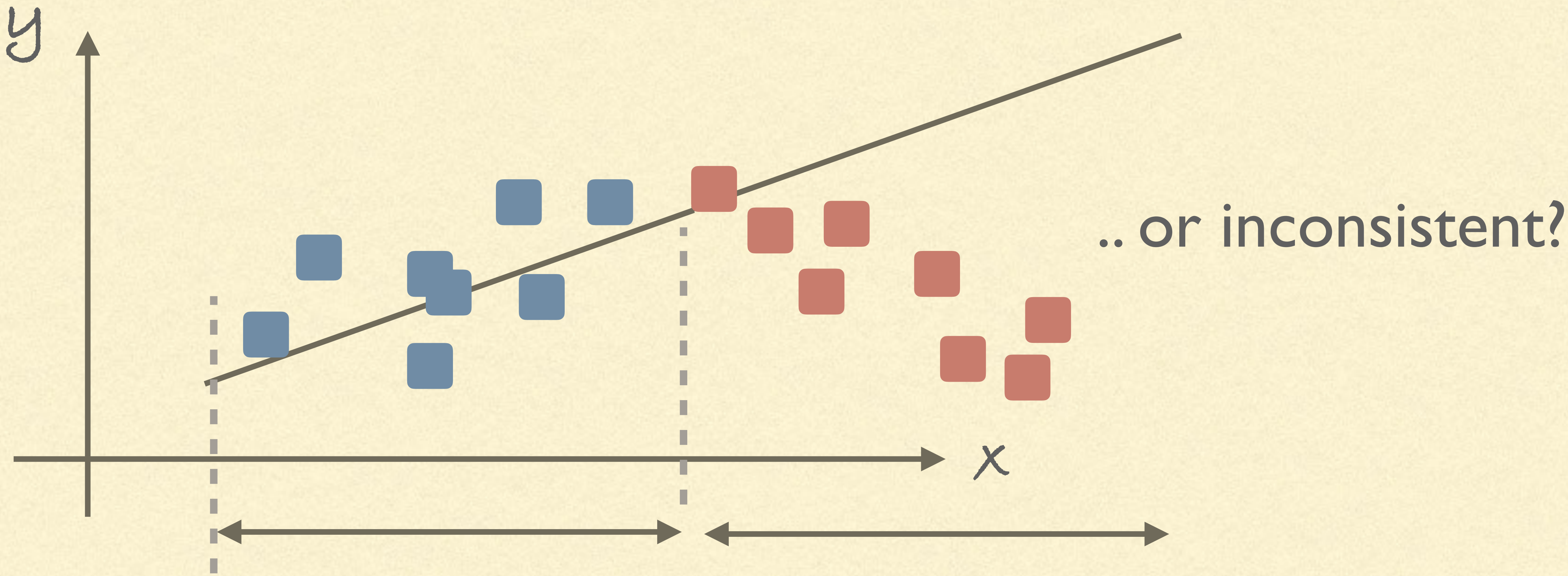


Question:
will the new data be consistent with this *prediction* ?

Training set
(or in sample)

Test set
(or out of sample)

Machine Learning : Prediction

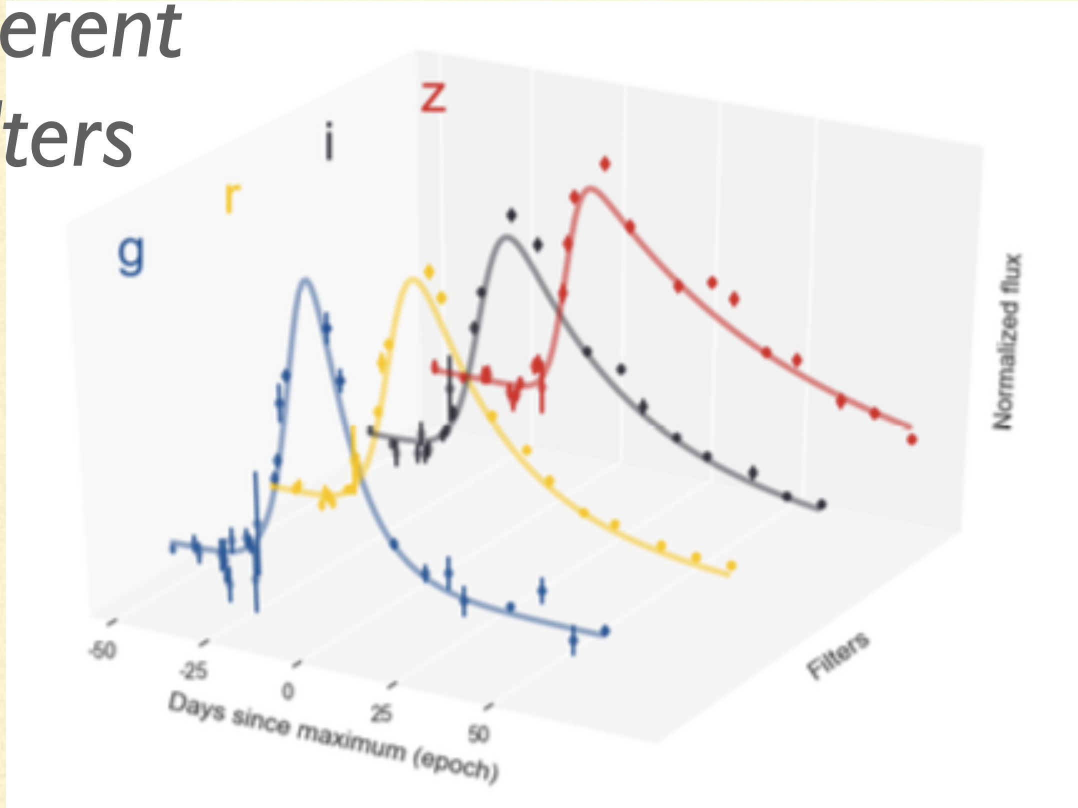


Training set
(or in sample)

Test set
(or out of sample)

Machine Learning for supernovae (SN) classification :

Different filters



Ishida et al., MNRAS 2019

A fit with Bazin's function f(t)

$$f(t) = A\frac{e^{-(t-t_0)/\tau_f}}{1+e^{(t-t_0)/\tau_r}} + B$$

a nonlinear mapping

$A, B, t_0, \tau_f, \tau_r$

5 features for each light curve

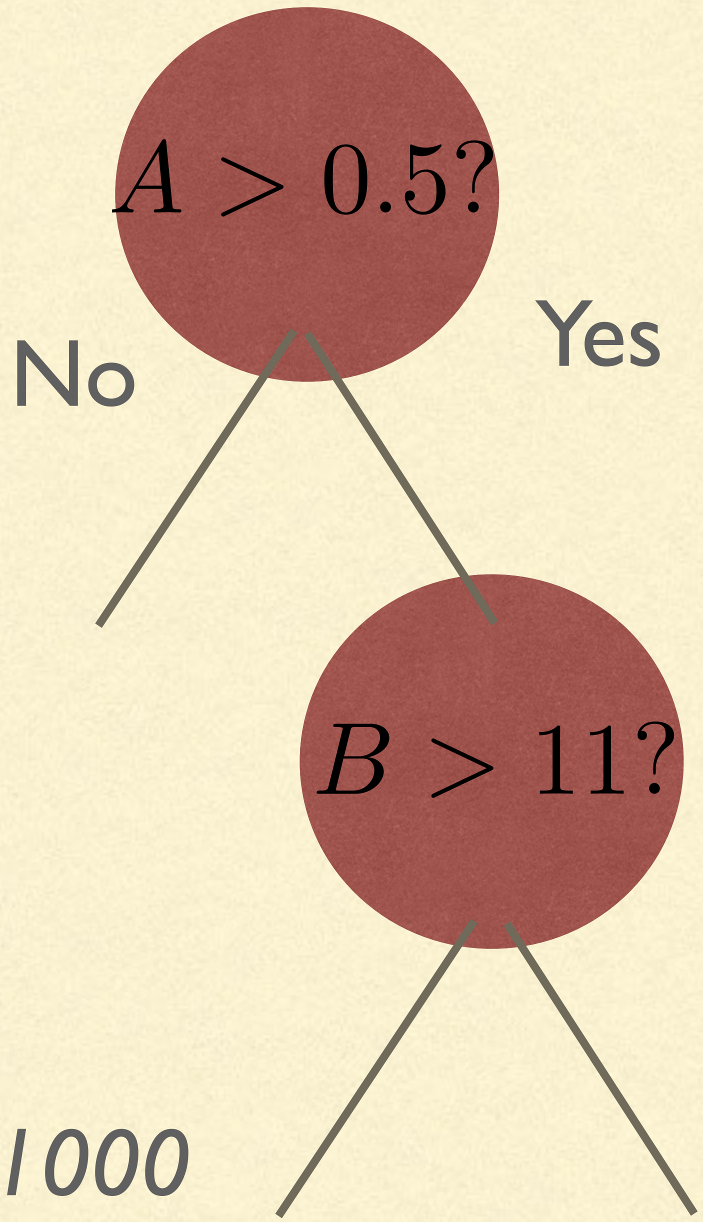
=>

0 (SN Ia),

1 (all the others)

Recipe :
Random forest

$$A, B, t_0, \tau_f, \tau_r$$



i. a series of trees

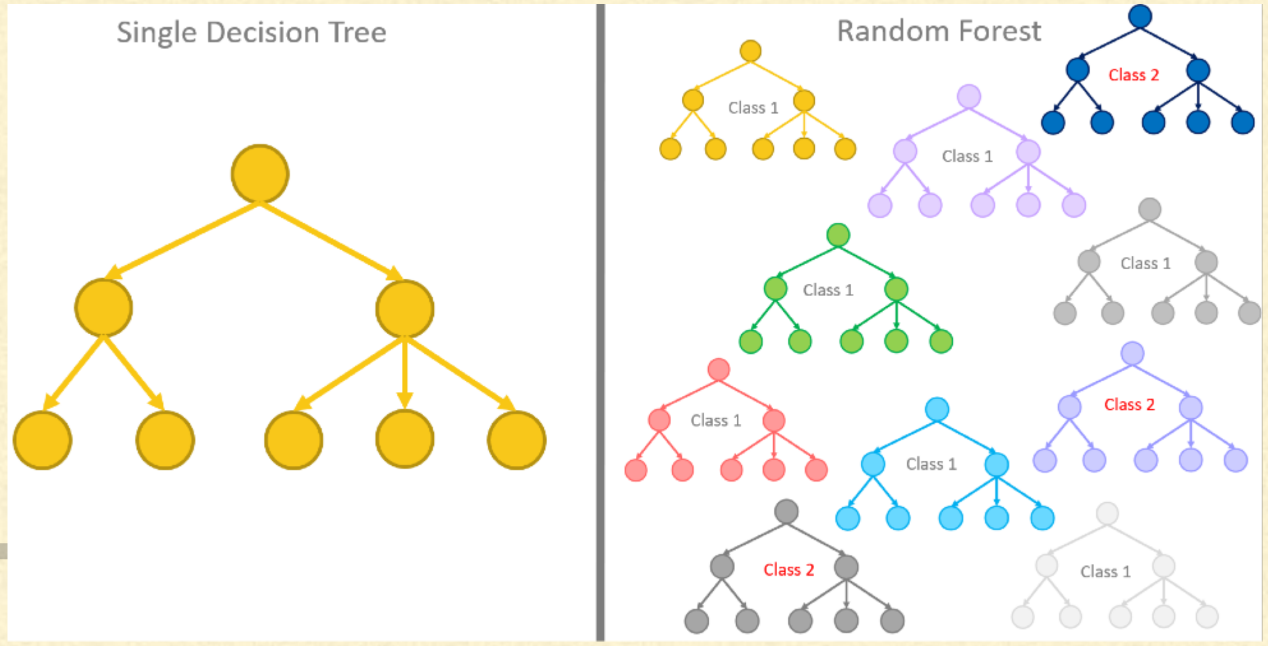
ii.Each tree enquires on the features getting down

to the known label = 0 or 1

averaging over many trees

=> probability

N estimators 1000



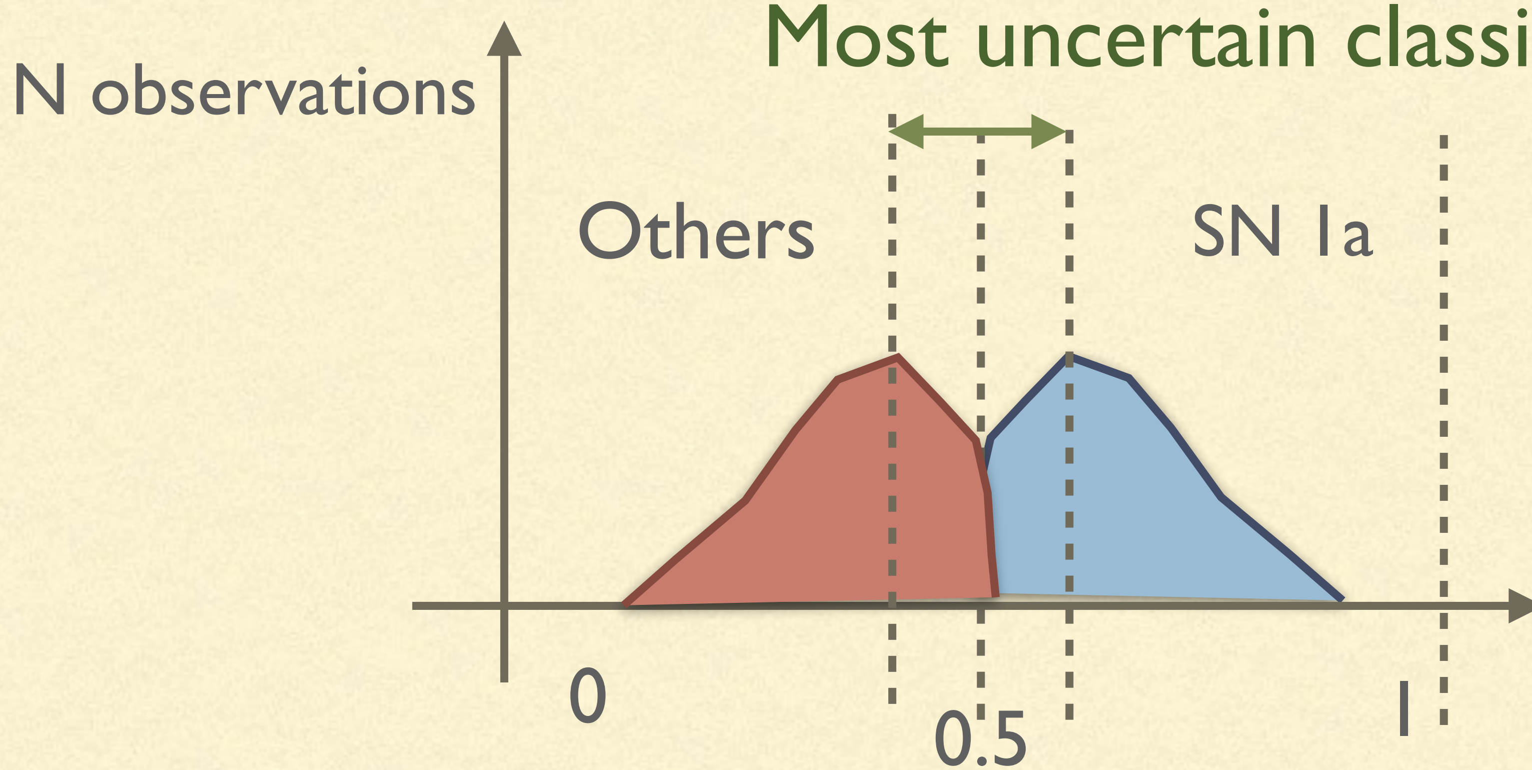
Recipe: “Active learning”

“Active learning is a branch of machine learning that deals with problems where *unlabeled data* is abundant yet obtaining labels is *expensive* (computationally or otherwise). The learning algorithm has the possibility of querying a limited number of samples to obtain the corresponding labels, subsequently used for supervised learning”

Cui et al., arXiv:1912.03927

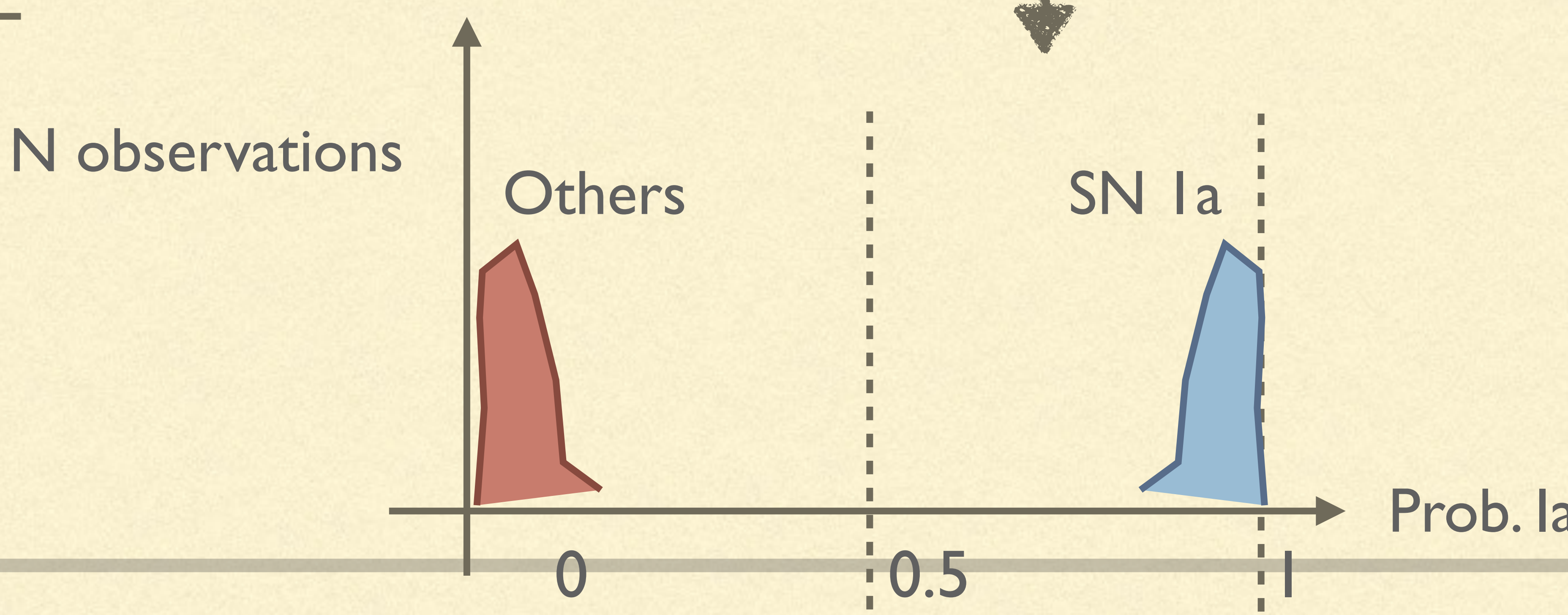
Active learning:

Most uncertain classification



1. Remove one of these objects at each active learning step,
2. Ask an oracle to label It
3. Use it to retrain mode

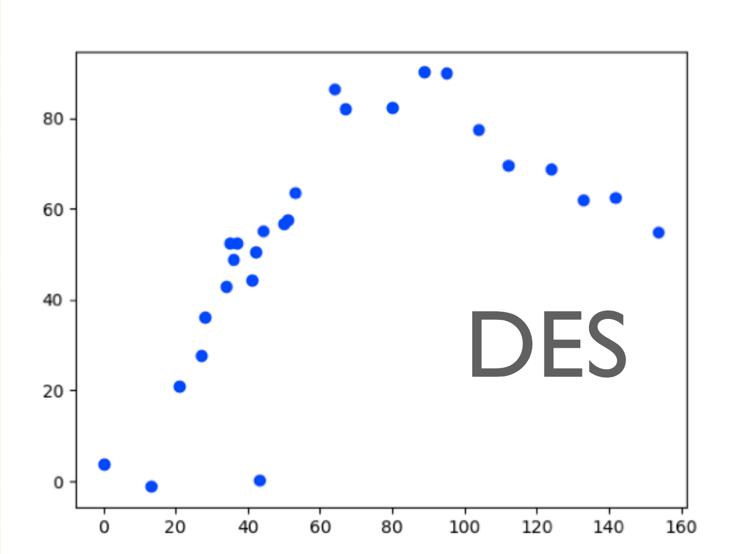
ideally, after M learning steps



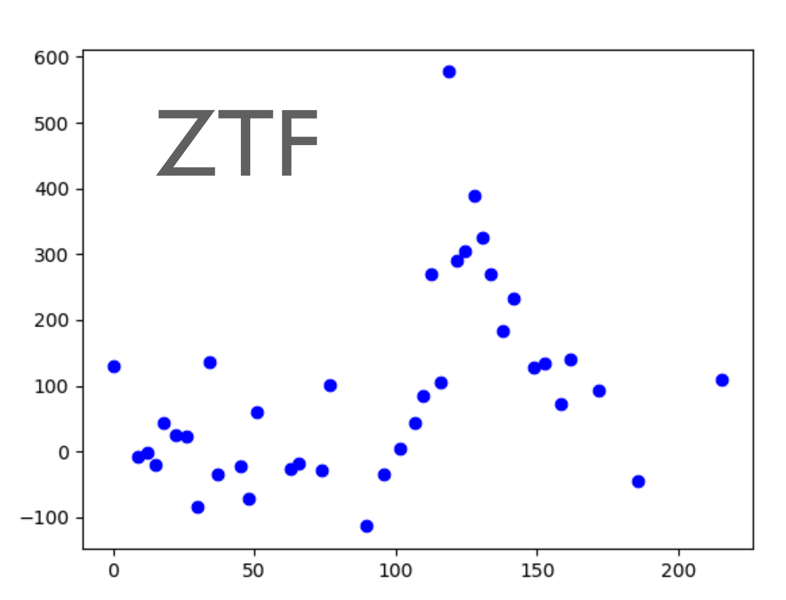
Some theoretical results :

1. Assuming the objects are only supernovae (type *Ia SN* or others *SN*)

2. Data are from different surveys :
DES vs. ZTF (higher levels of noise + only “r” and “g" filter)



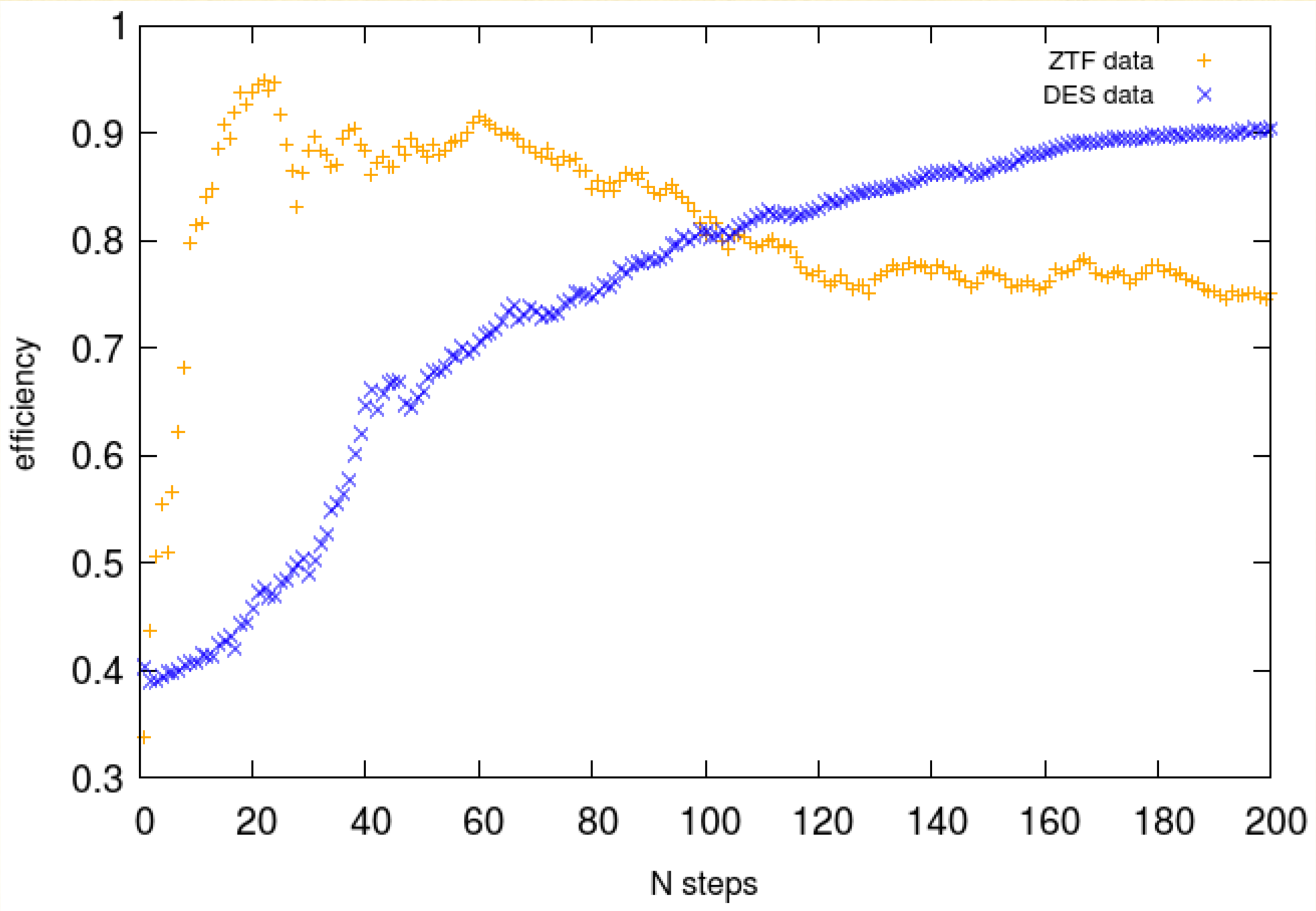
Ishida et al., MNRAS 2019

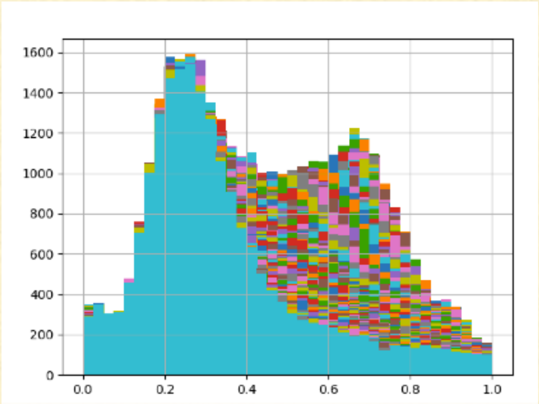


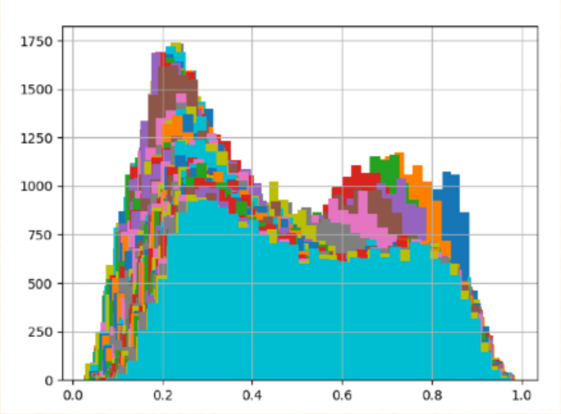
courtesy of Daniel Muthukrishna

Why theoretical results ?
Advantage is that labels are known beforehand also in the test set => can compute metrics quantify performance

Metrics with full l.c. : efficiency

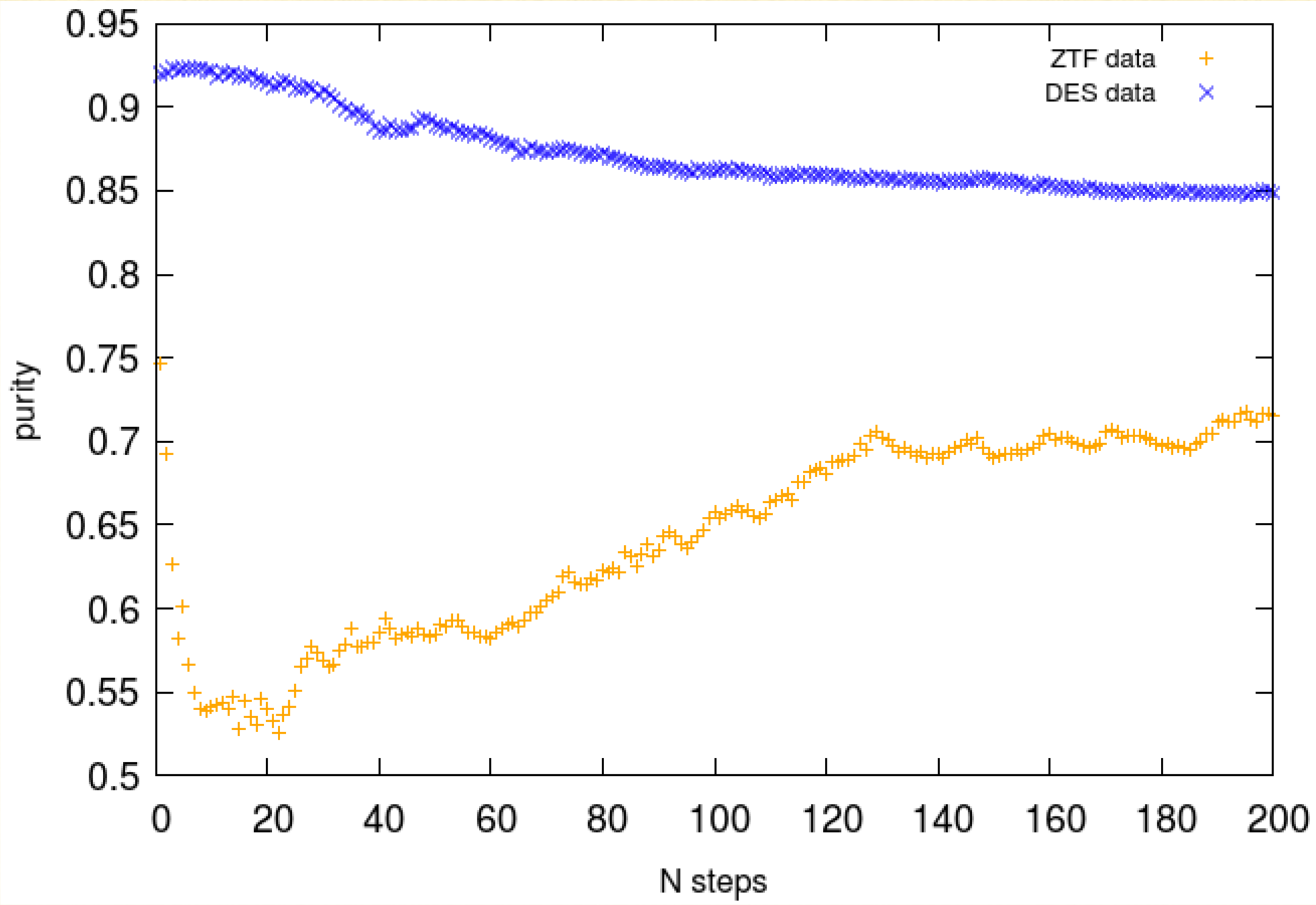


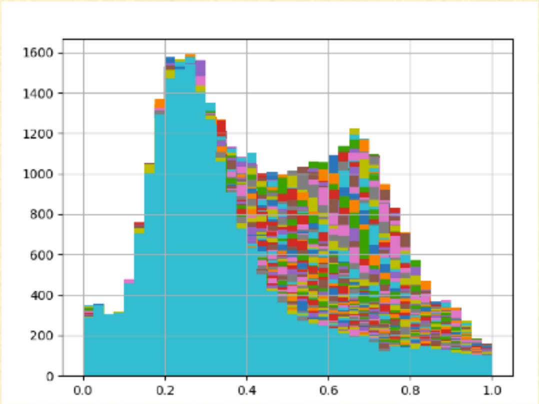


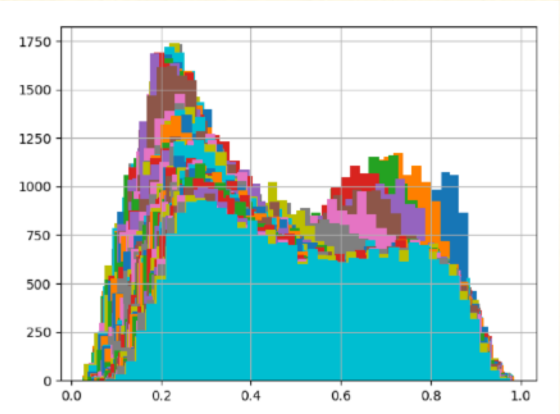


$$\text{efficiency} := N_{Ia,s.c.}/N_{Ia,tot.}$$

Metrics with full l.c. : efficiency

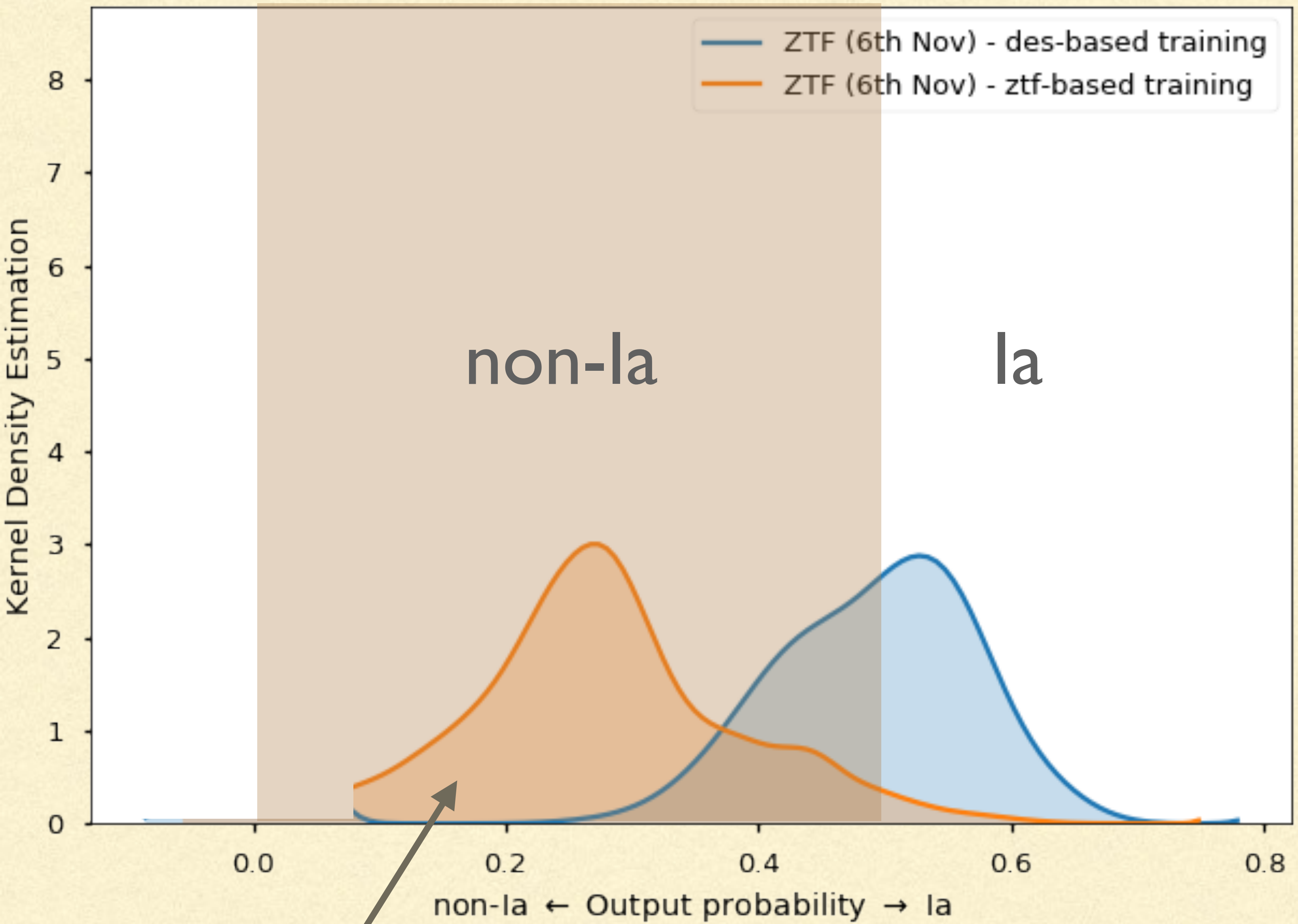






$$\text{purity} := N_{Ia,s.c.}/(N_{Ia,s.c.} + N_{Ia,w.c.})$$

Test observing real ZTF data



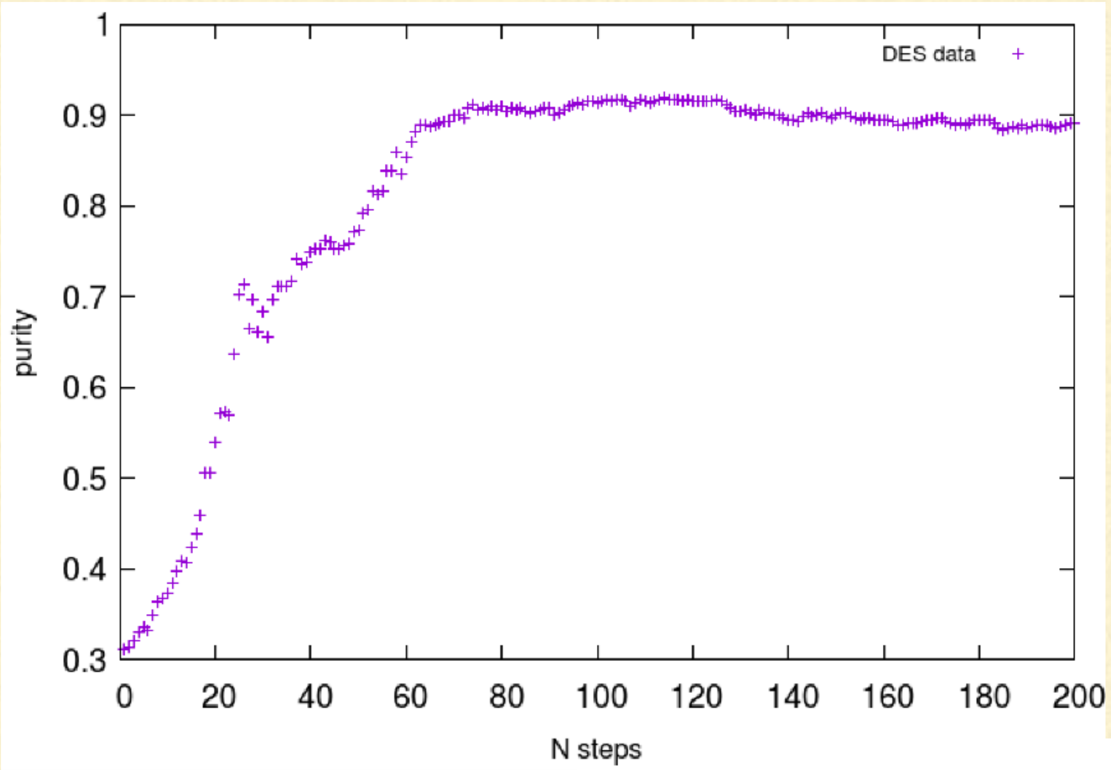
New ZTF simulations effective at classifying Ia…

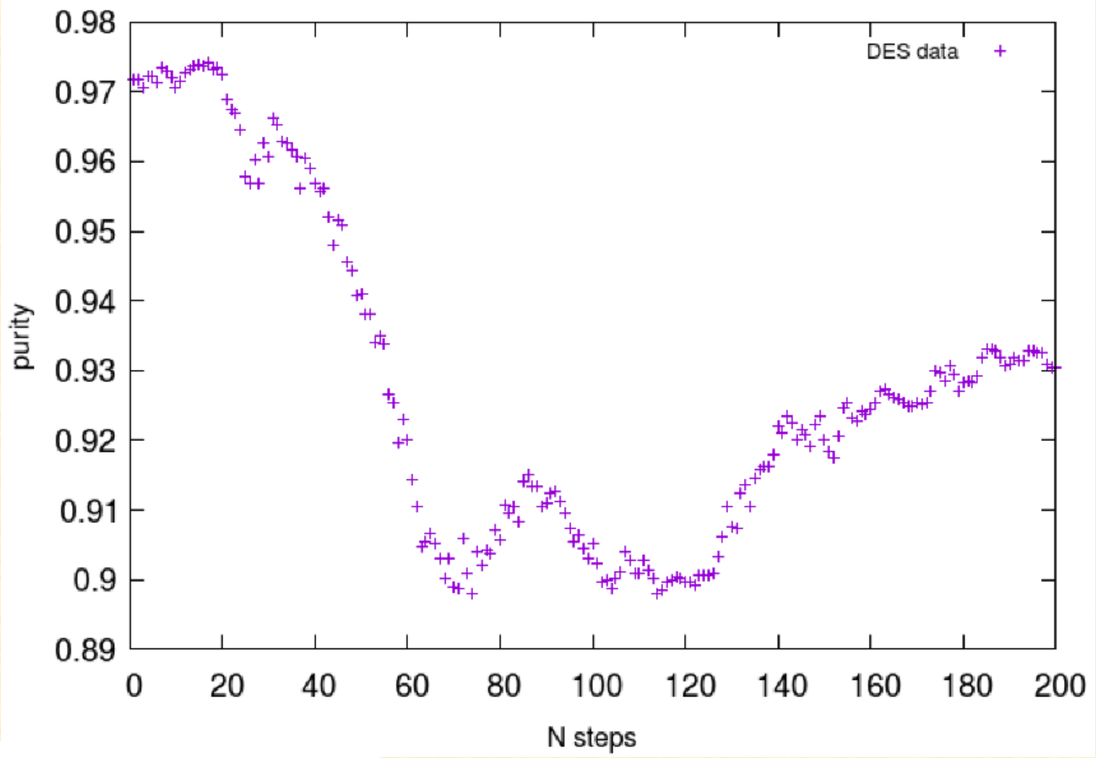
however…

ZTF-based model

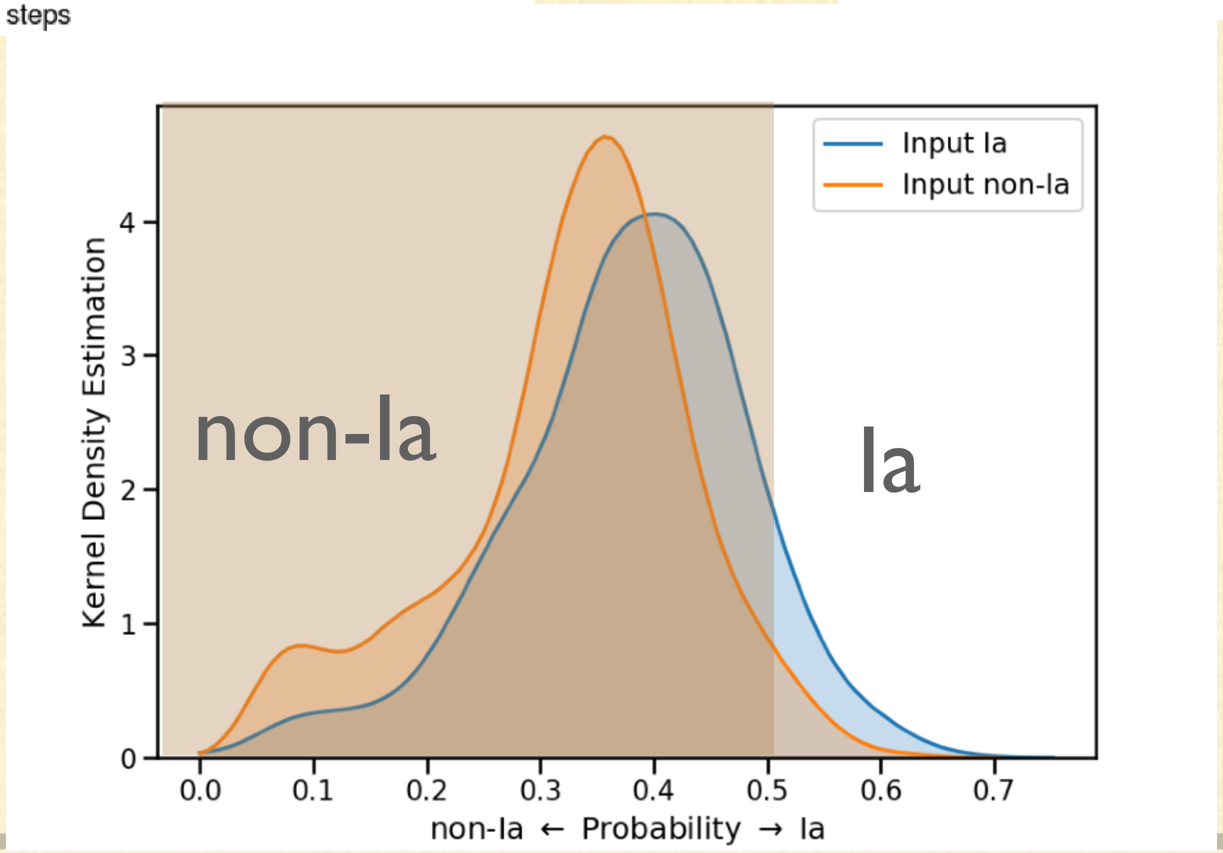
	objectId	score	tns
0	ZTF19achufoh	0.538	NaN
1	ZTF19achufoh	0.533	NaN
2	ZTF19achufoh	0.505	NaN
3	ZTF19aclnrkg	0.344	NaN
4	ZTF18acruoyq	0.335	NaN
5	ZTF19acmdpyr	0.330	SN Ia
6	ZTF19abgiwkt	0.322	SN II
7	ZTF18acmyprz	0.308	NaN
8	ZTF19acmdpyr	0.308	SN Ia
9	ZTF18aaznglt	0.306	NaN

Different set of features : "moments" of the photometric curves (for DES)





but looking at real data, hard to distinguish Type Ia from the others



Summary and future directions

- We have integrated into the broker the work of *Ishida et al.*, MNRAS *2019*
- We compared DES and ZTF simulations (higher levels of noise)
- Mostly looked at Bazin features

Scratched the surface so far - more challenges lie ahead :

1. Other systems of *features*, besides moments, perhaps also consider the error-bars
2. Which *classifier* works better ?
3. In real data the test set has a few points (at least 5 are needed with Bazin) which algorithm accounts for this in a optimal way?