Early SNIa classification using active learning

Marco Leoni

Rubin LSST-France, LPNHE
29 November 2022
Paris

Introduction

Who :

E. Ishida
LPC, Clermont

A. Moller
ANU, Swinburne

J. Peloton
IJCLAB, UPSaclay

What : **Early discovery of supernovae (no need to say why SNIa are relevant !)**

Why Machine Learning : **huge amount of data from LSST hence the need for making 'automatic' reliable predictions**

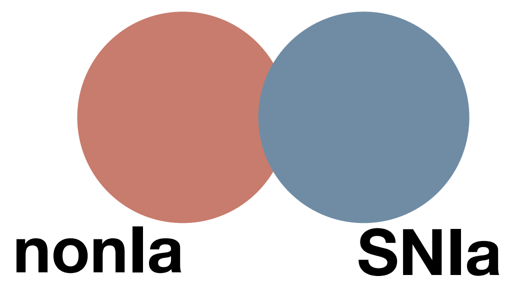
When : **focus on 2021-2022 using ZTF data**

Where : **https://doi.org/10.1051/0004-6361/202142715**

A&A, **Volume** 663, July 2022

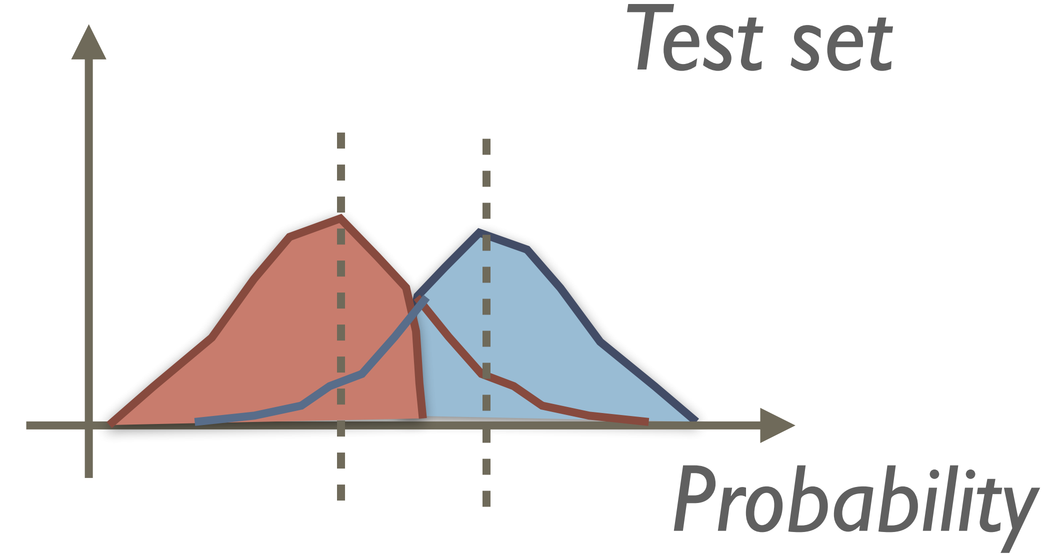
What is active ML ?

Train set



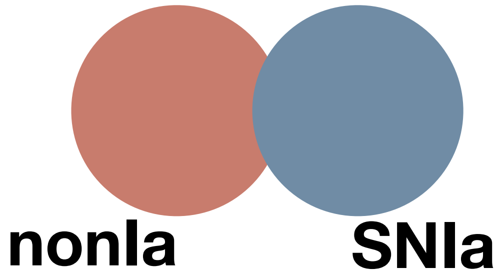
Model
Predictions

Test set

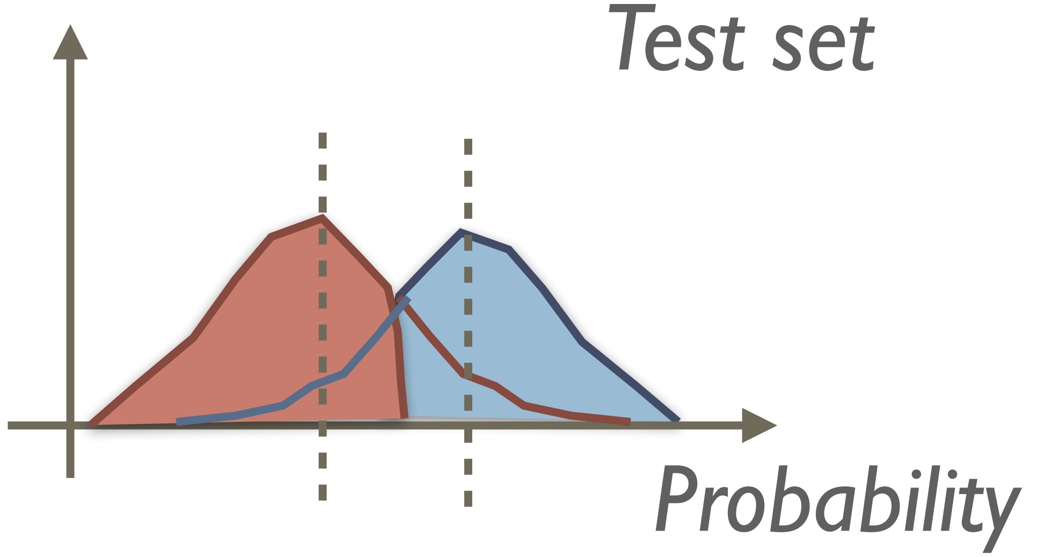


What is active ML ?

Train set



Model Predictions

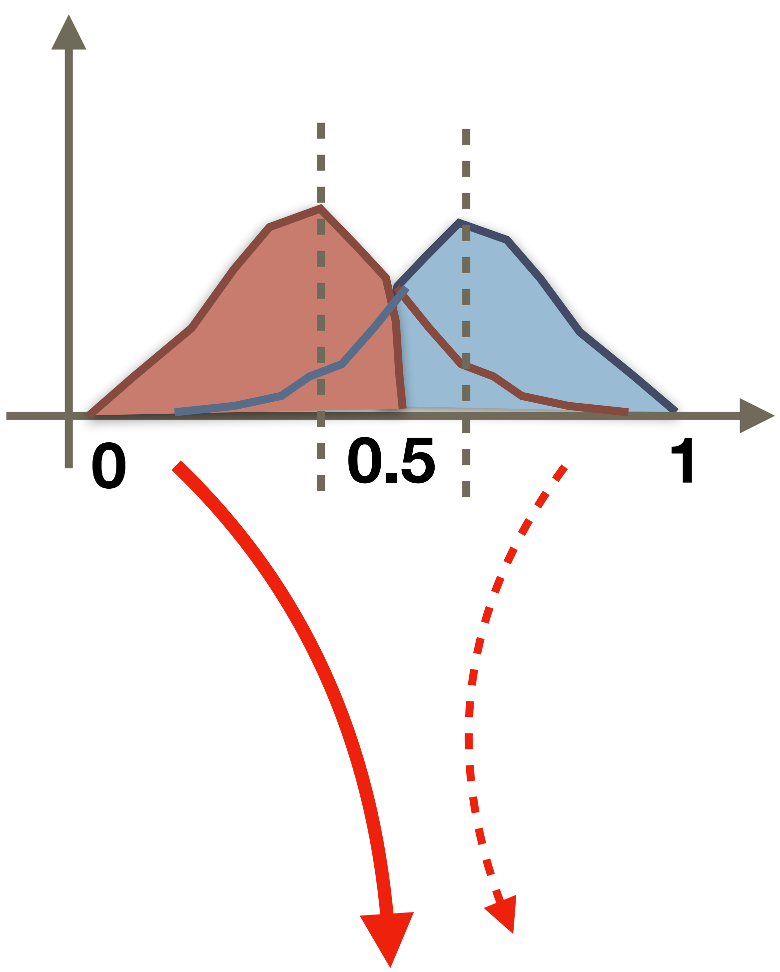


*Add to training set & **retrain***

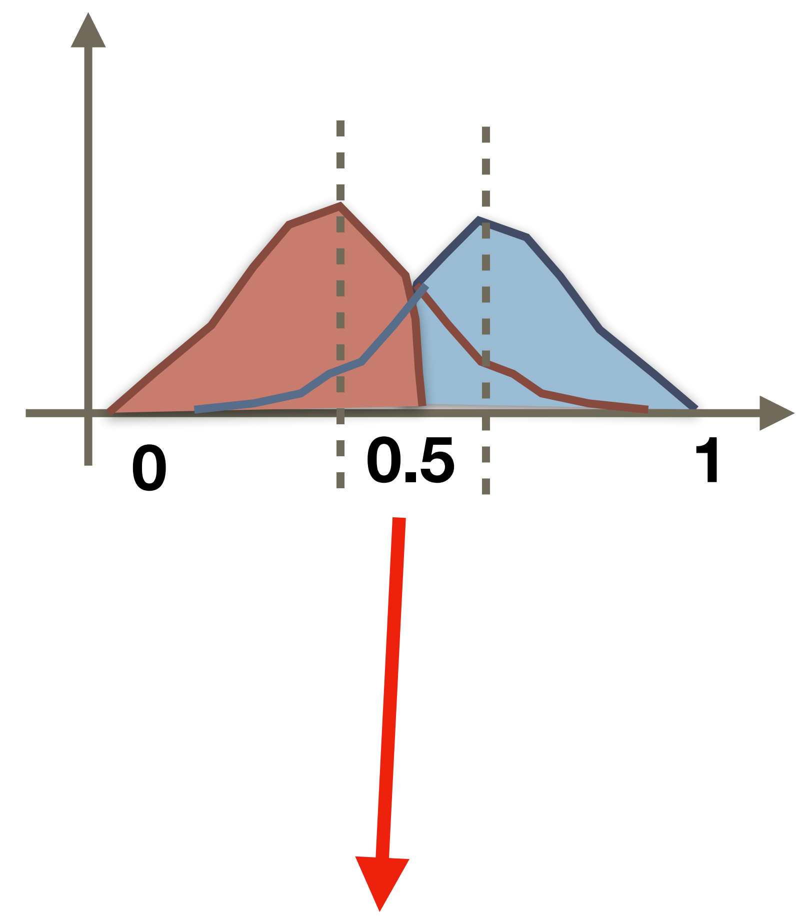
Remove data from test set

Get the actual label

Uncertainty vs Random Sampling

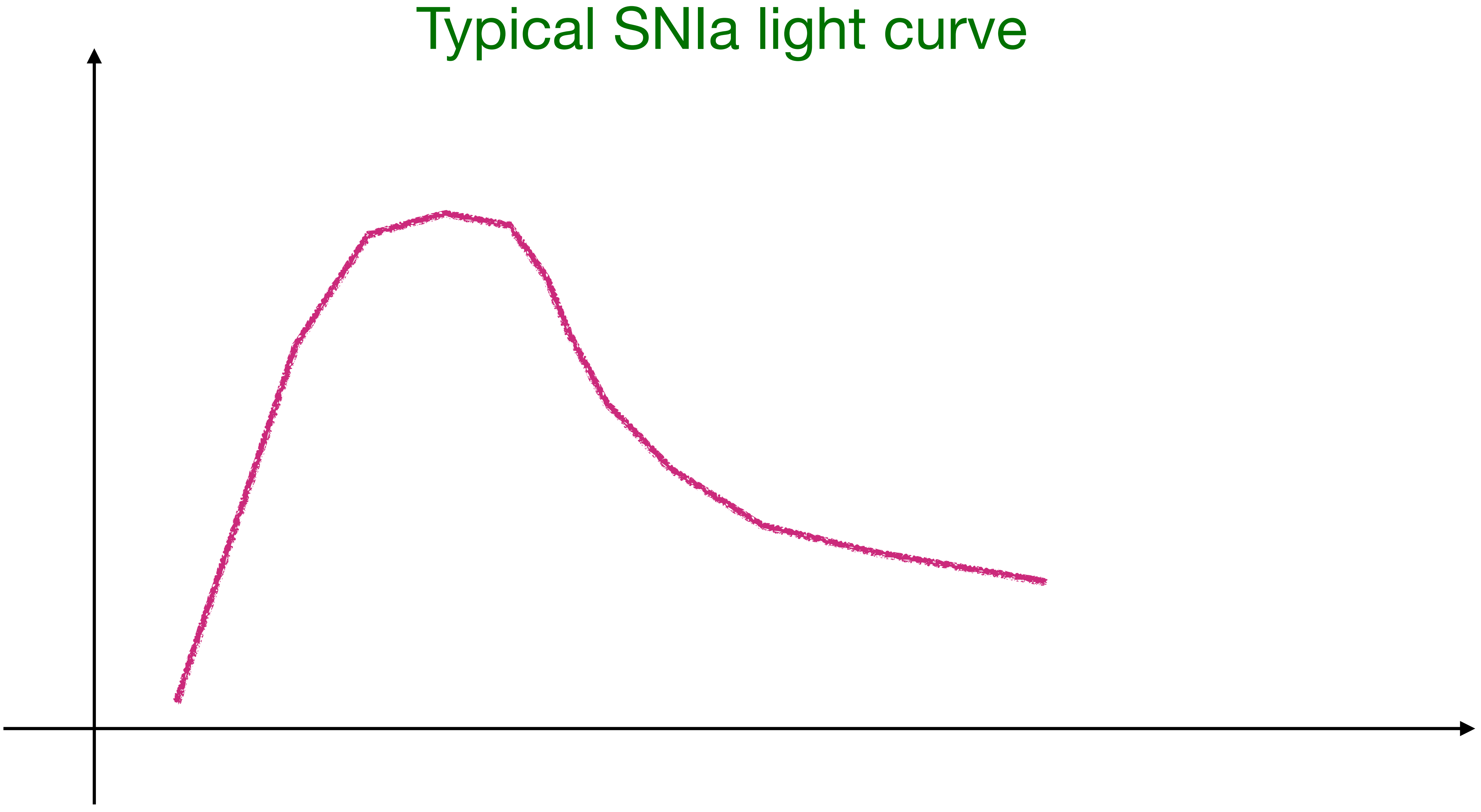


Random sampling



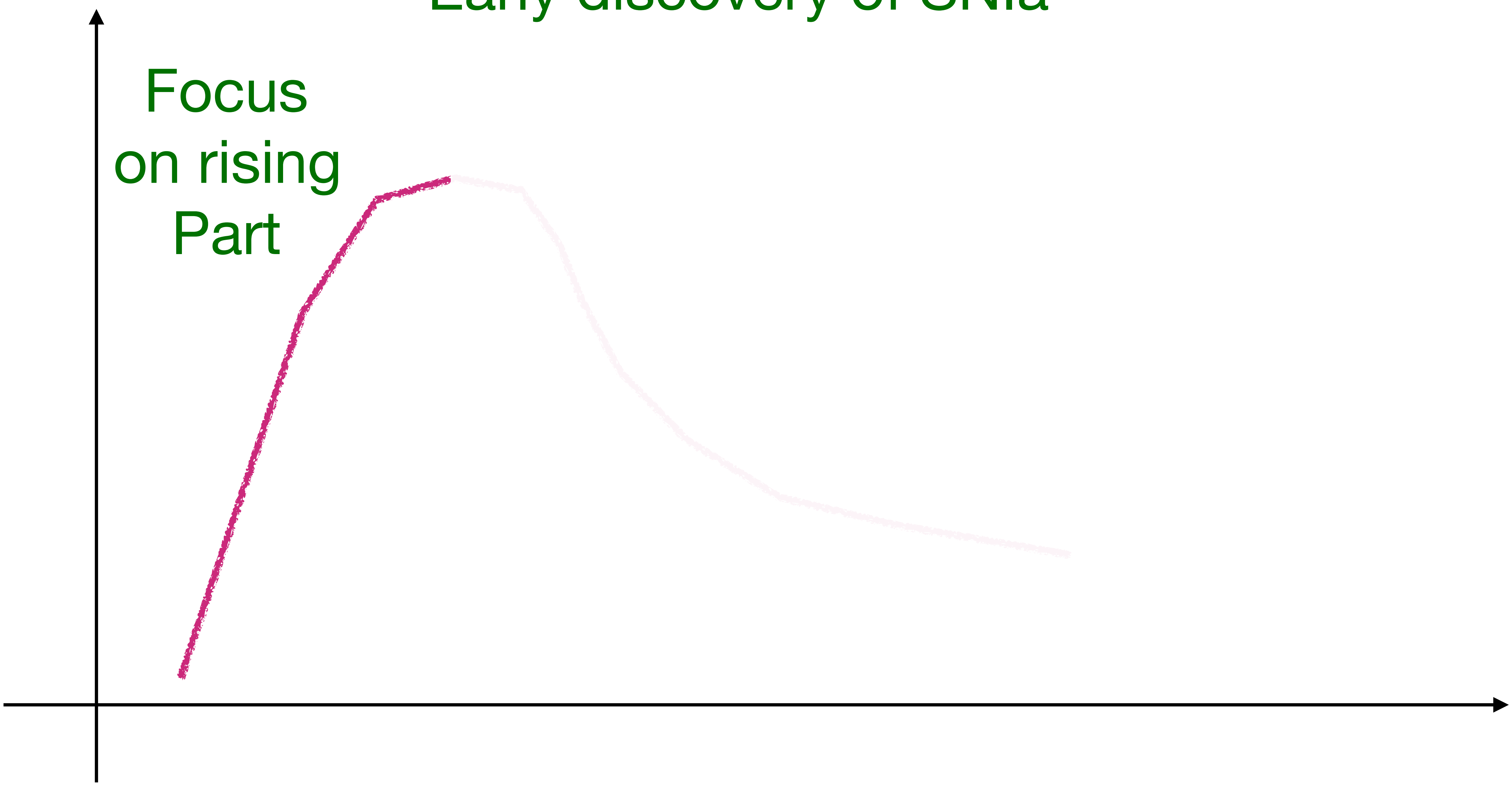
Uncertainty sampling

Learning from light curves



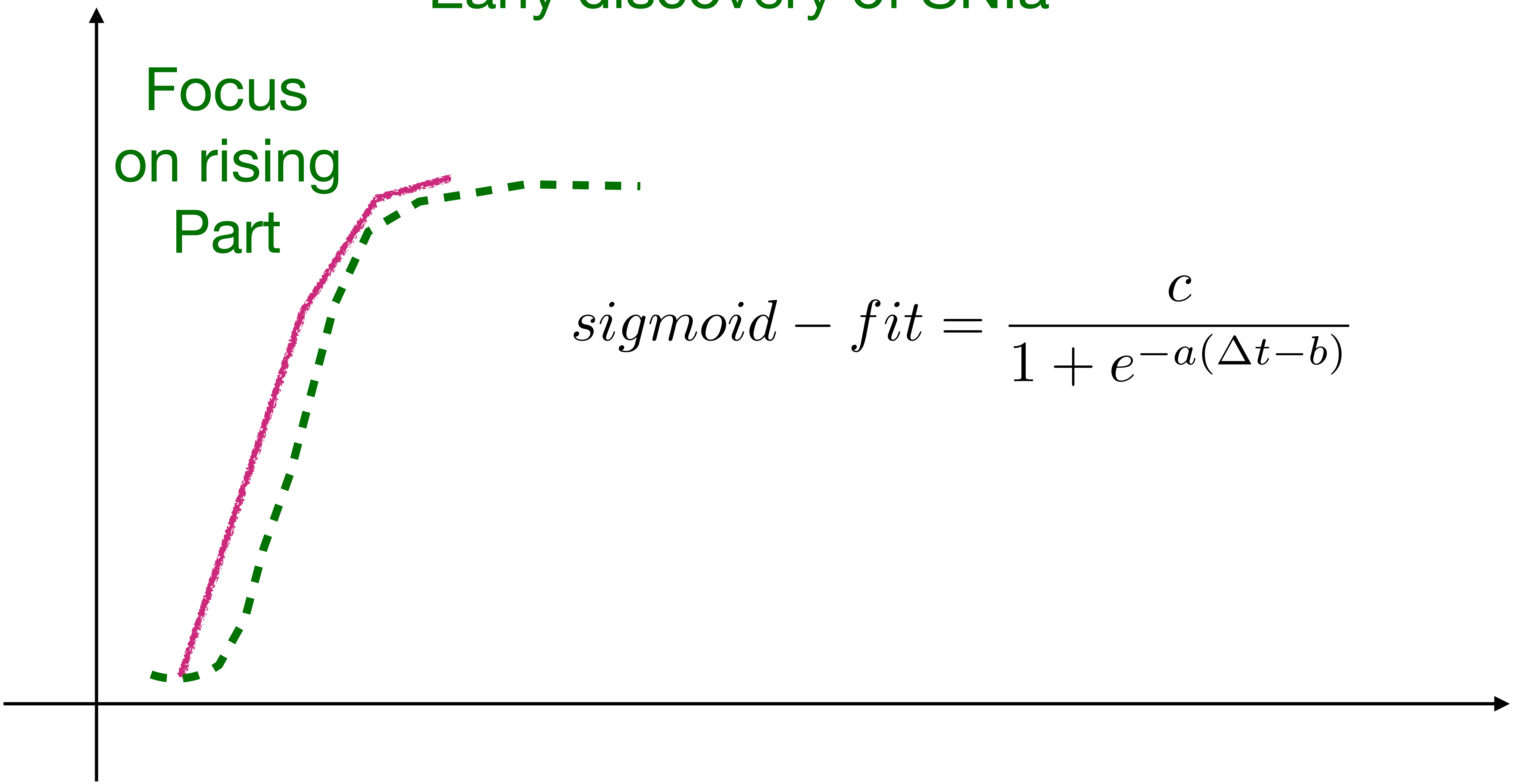
Learning from light curves

Early discovery of SNIa



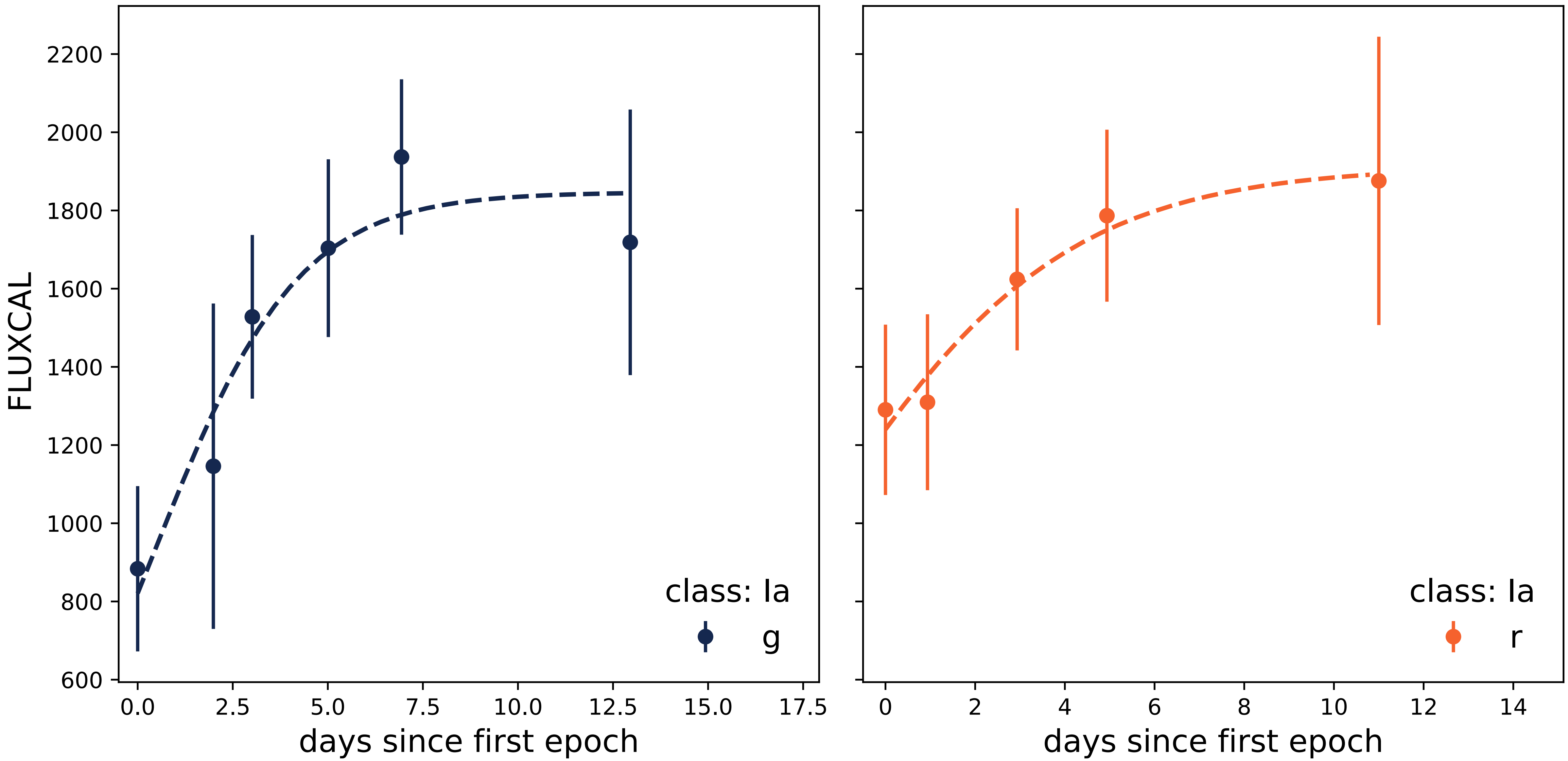
Learning from light curves

Early discovery of SNIa

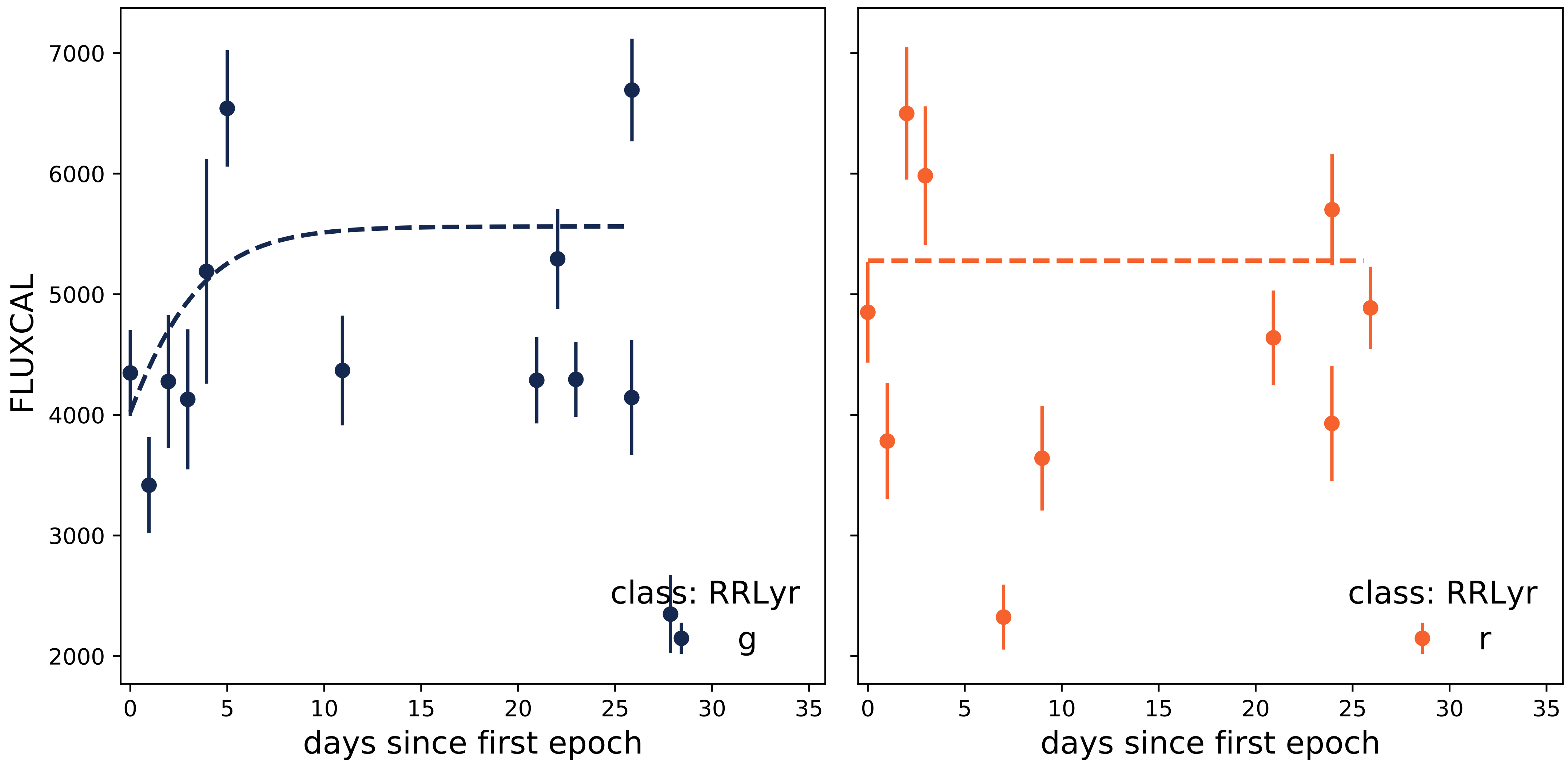


$$sigmoid - fit = \frac{c}{1 + e^{-a(\Delta t - b)}}$$

Actual data for SNIa



...and for nonIa



Luckily nonIa objects are very diverse hence the fit of such light curves is pretty bad.

A feature $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$ describes the goodness of the fit

Features

I) $$sigmoid - fit = \frac{c}{1 + e^{-a(\Delta t - b)}}$$

3 features +

II) $$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

1 feature +

III) $$SNR = \frac{1}{N} \sum_{i=1}^{N} \frac{o_i}{\Delta o_i}$$

1 feature+

IV) *N points in the raising part*

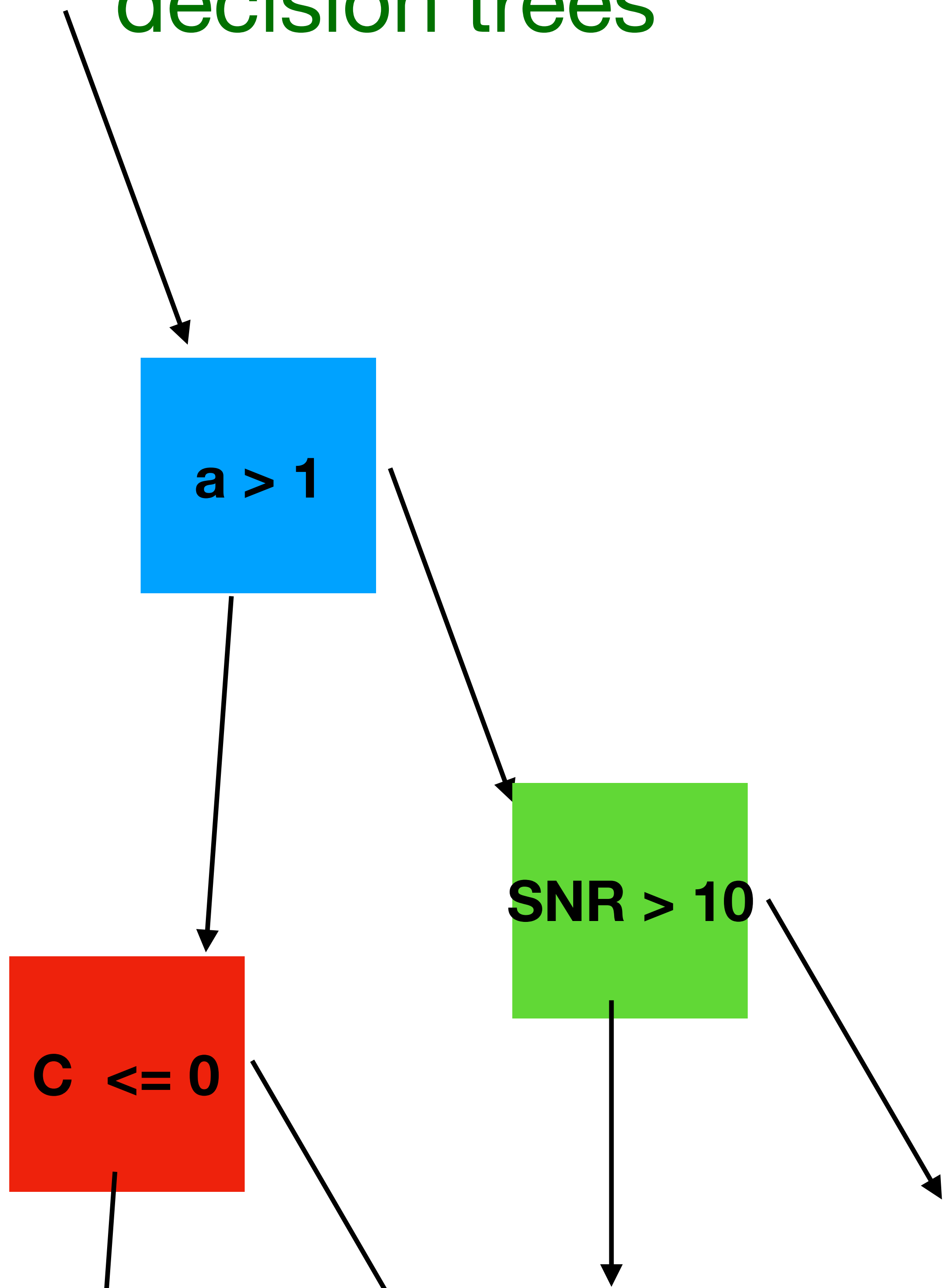
1 feature =

---

6 features which we use to train RF model

Training using random forests

an ensemble of
decision trees



Parameters for our training

N-trees = 1000

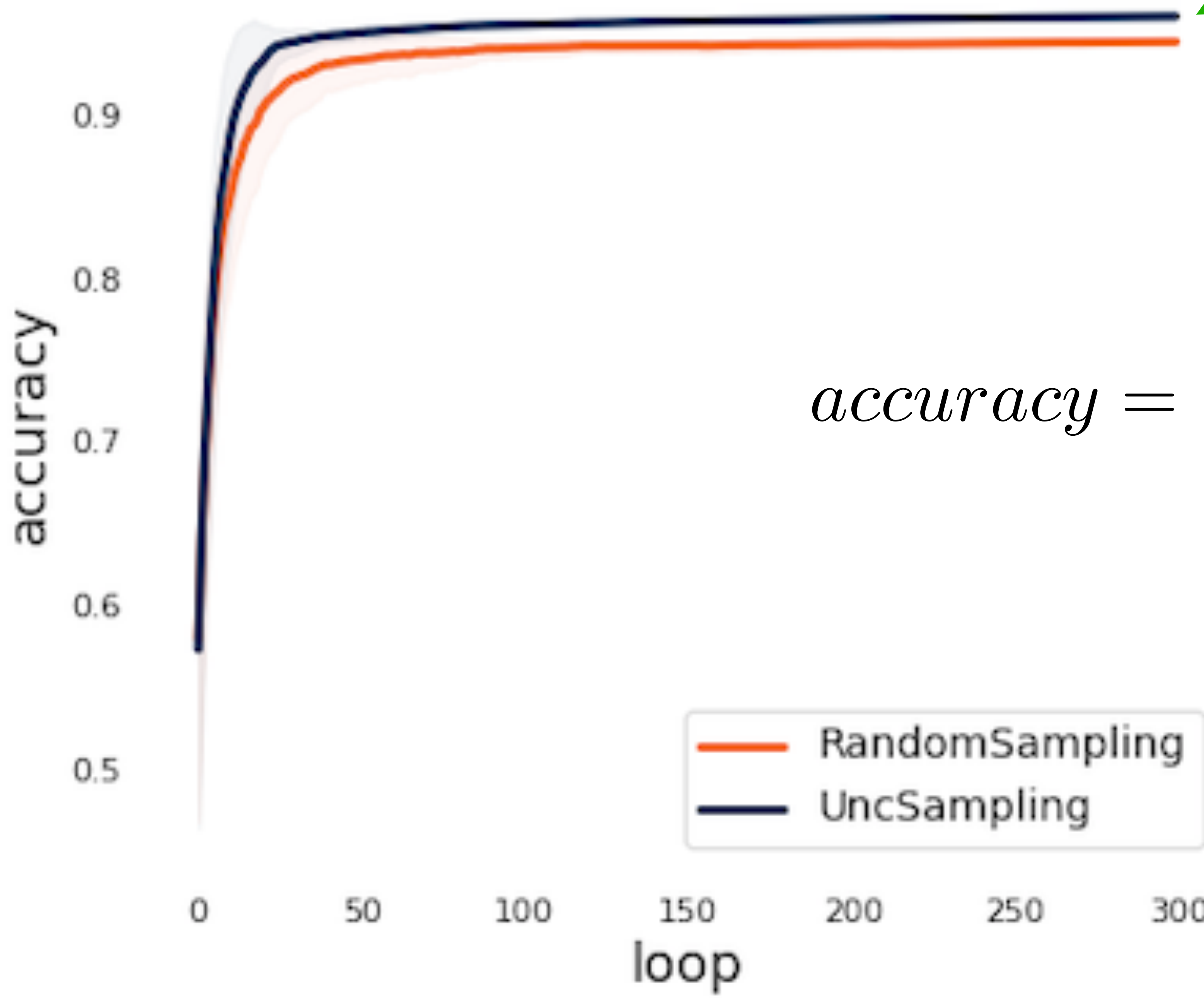
Train Sample ~ 10 alerts
(half of which SNIa)

Test Sample ~ 23000 alerts
(vast majority of nonIa)

300 steps of active learning

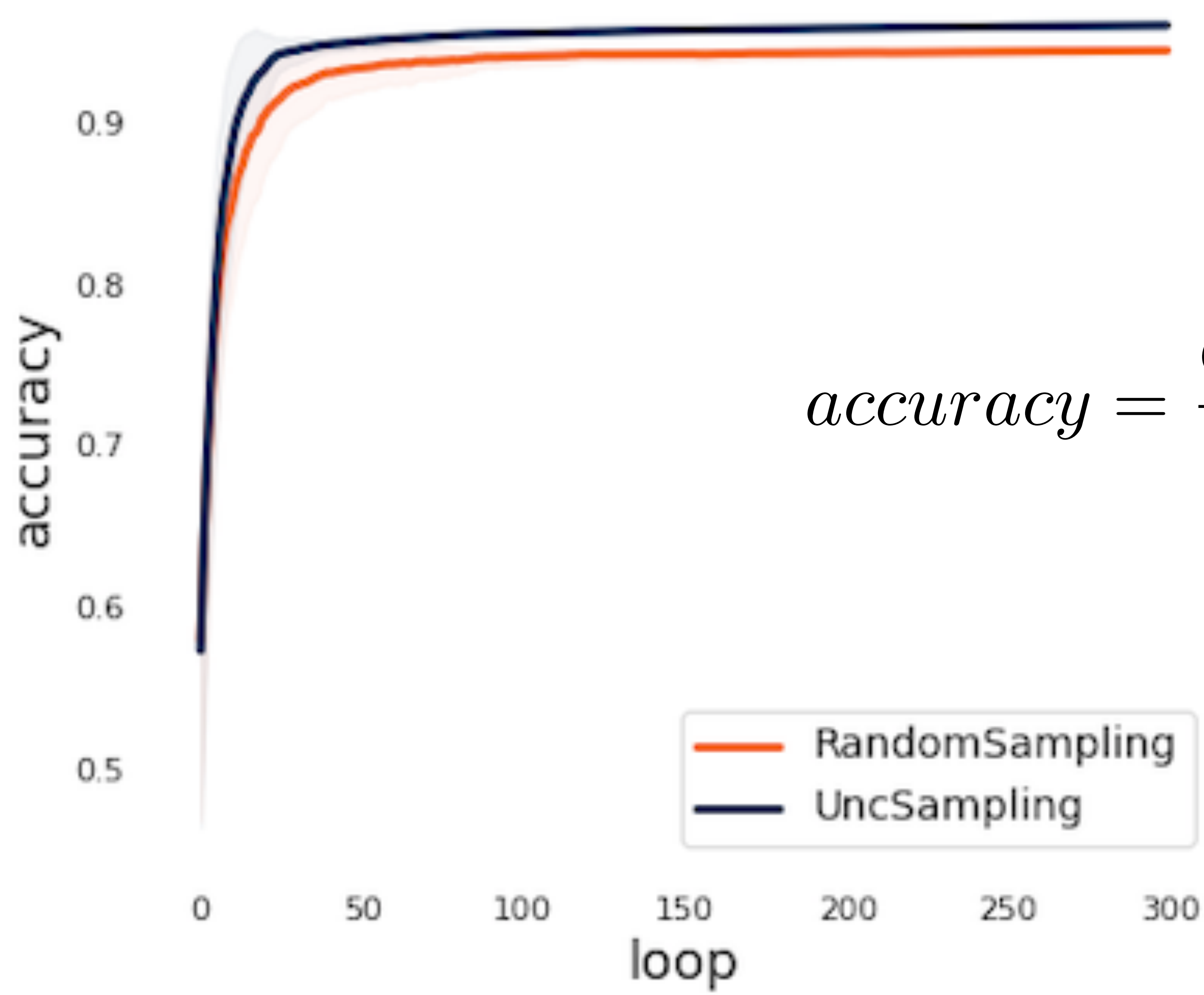
Results : accuracy

average over
100 different
realisations

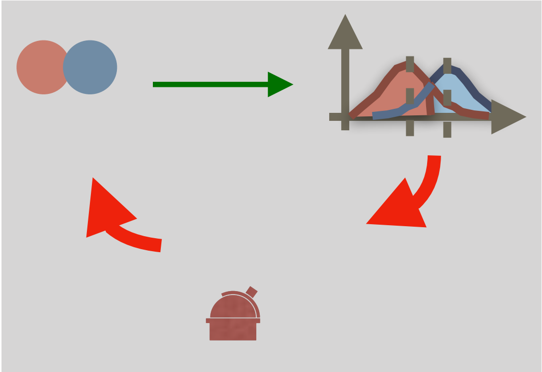


$$accuracy = \frac{C_{Ia} + C_{nonIa}}{N}$$

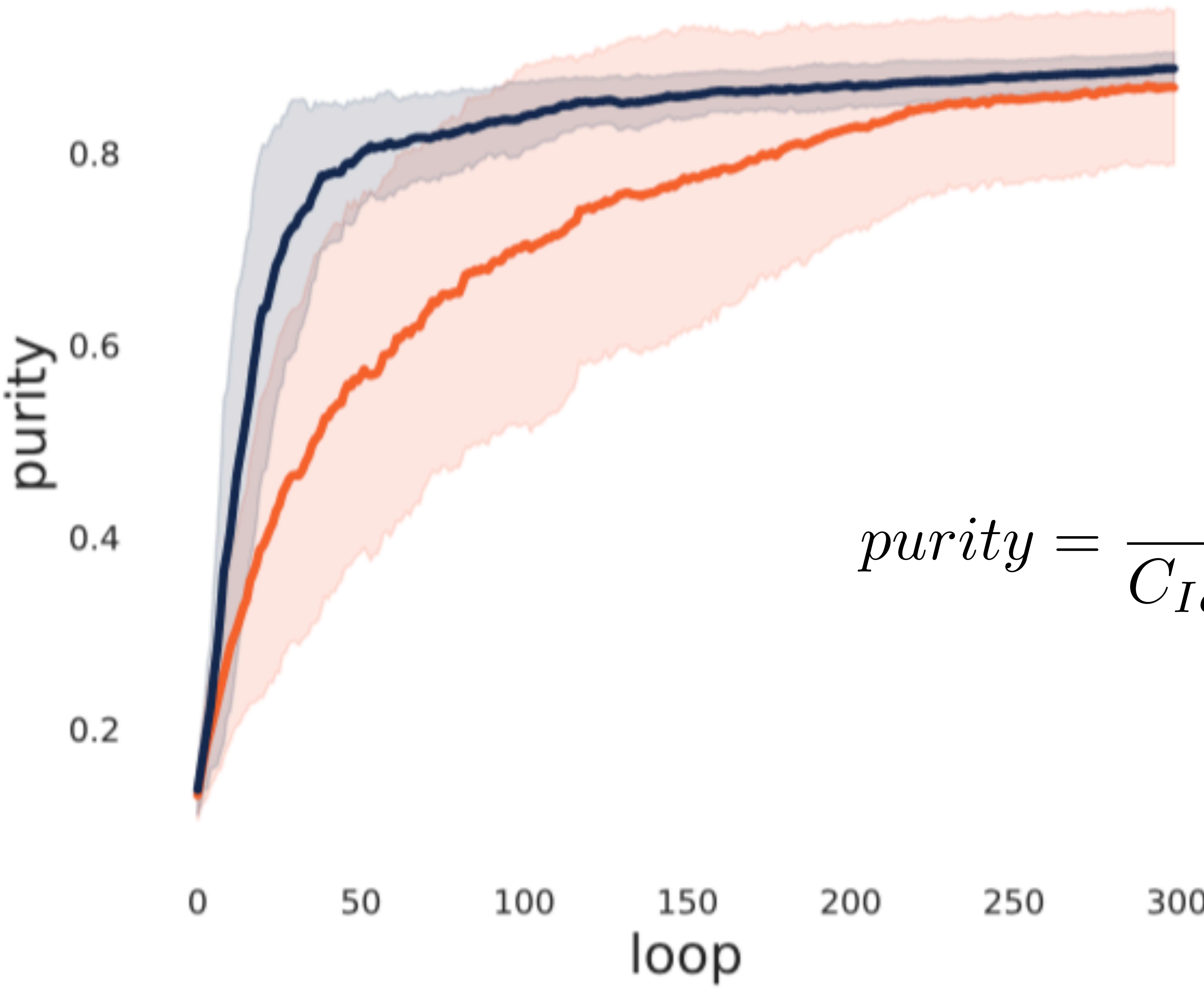
Results : accuracy



$$accuracy = \frac{C_{Ia} + C_{nonIa}}{N}$$

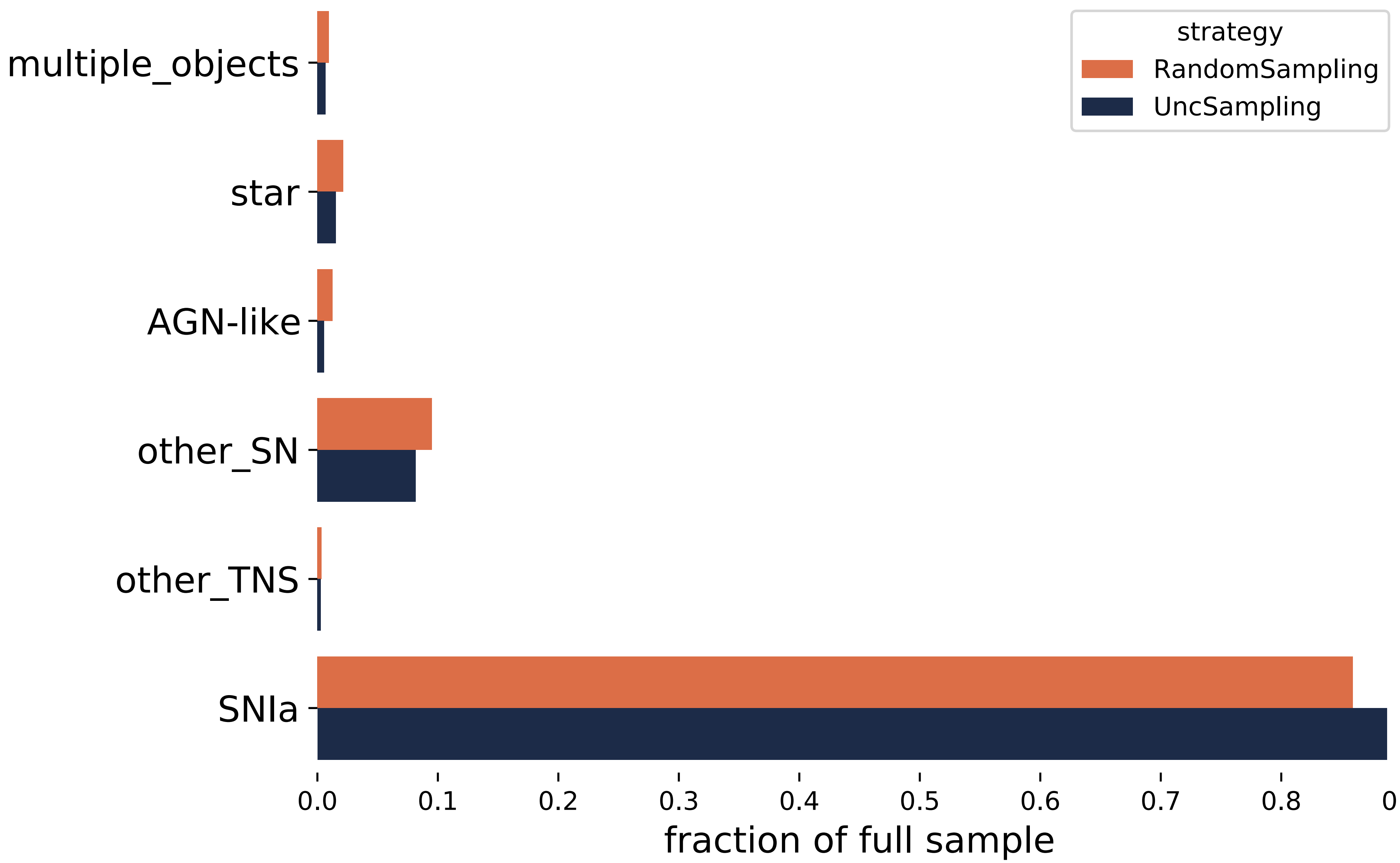


Results : purity

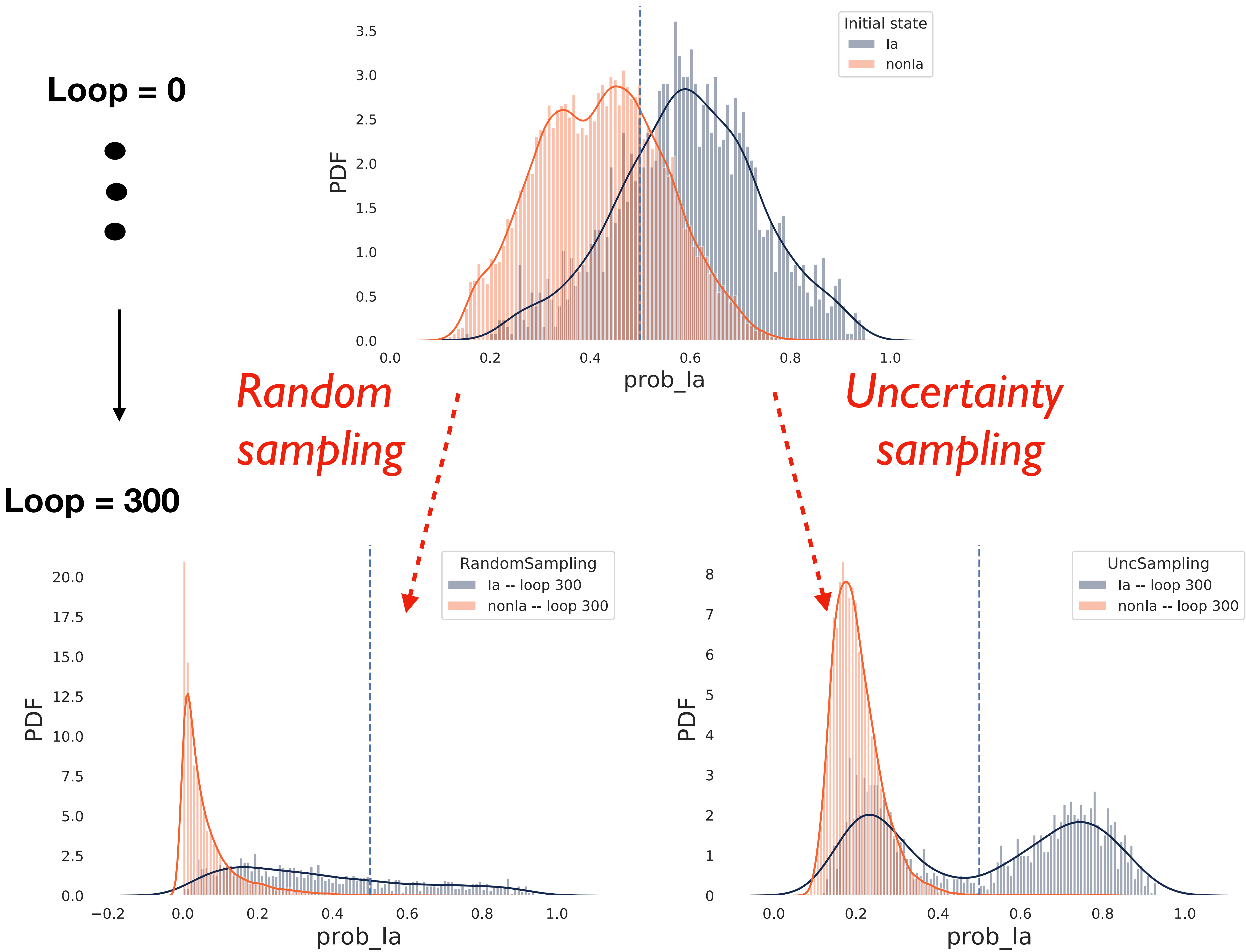


$$purity = \frac{C_{Ia}}{C_{Ia} + W_{nonIa}}$$

Results : actual classes



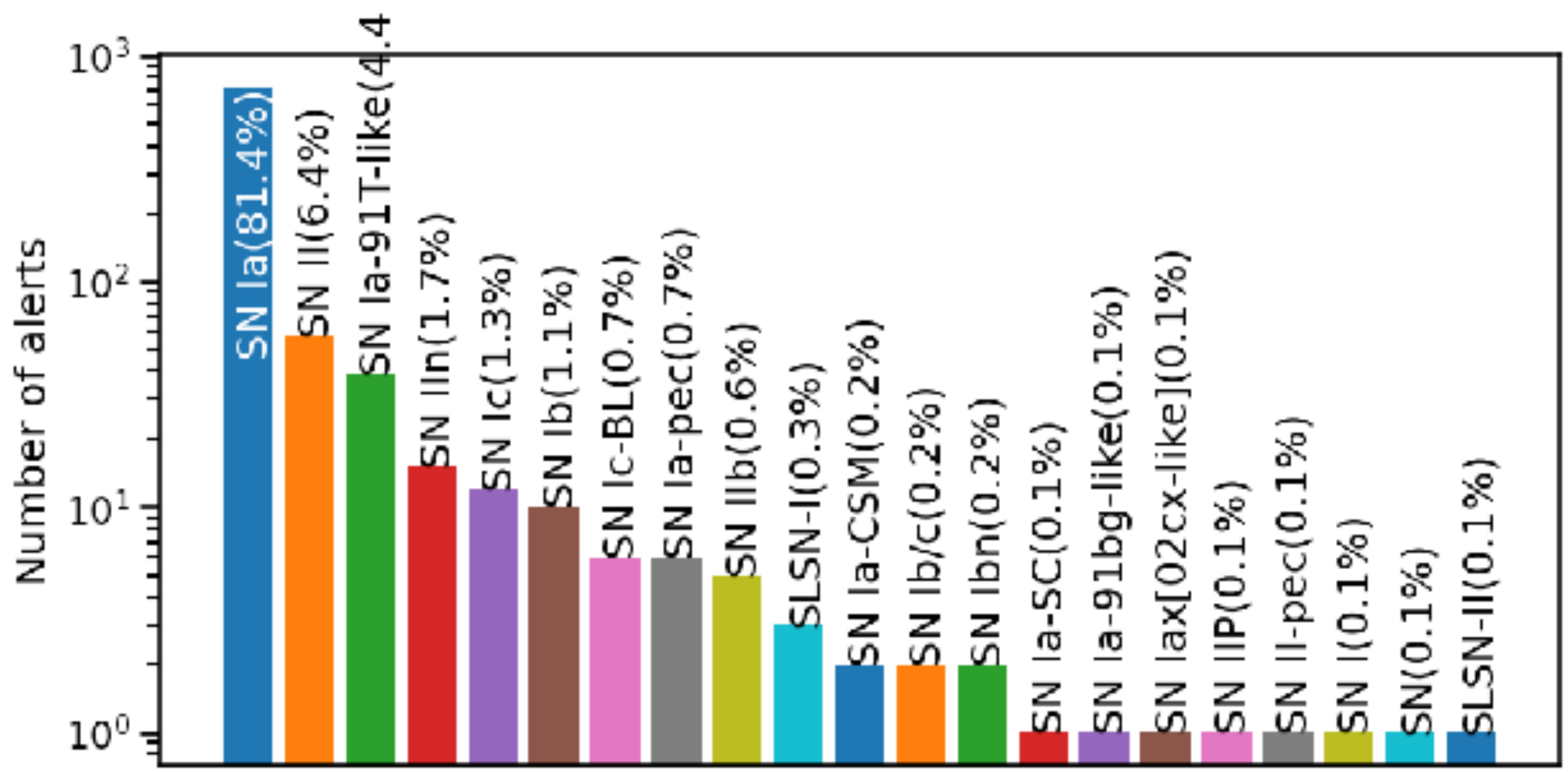
Results : probability distribution



Results : in prod

predictions are matched with those from Bayesian Neural Network models (Anais Moller).
If both models agree predictions are sent out.

from November/2020
to November/2022
Fink communicated
1,533 early
SNIa candidates to TNS.



908 (59%) of which : followed-up & spectroscopically by facilities around the world
788 (86%) of which : correctly classified as SNIa

Thanks, for your attention!

Backup

Early discovery: difficulties

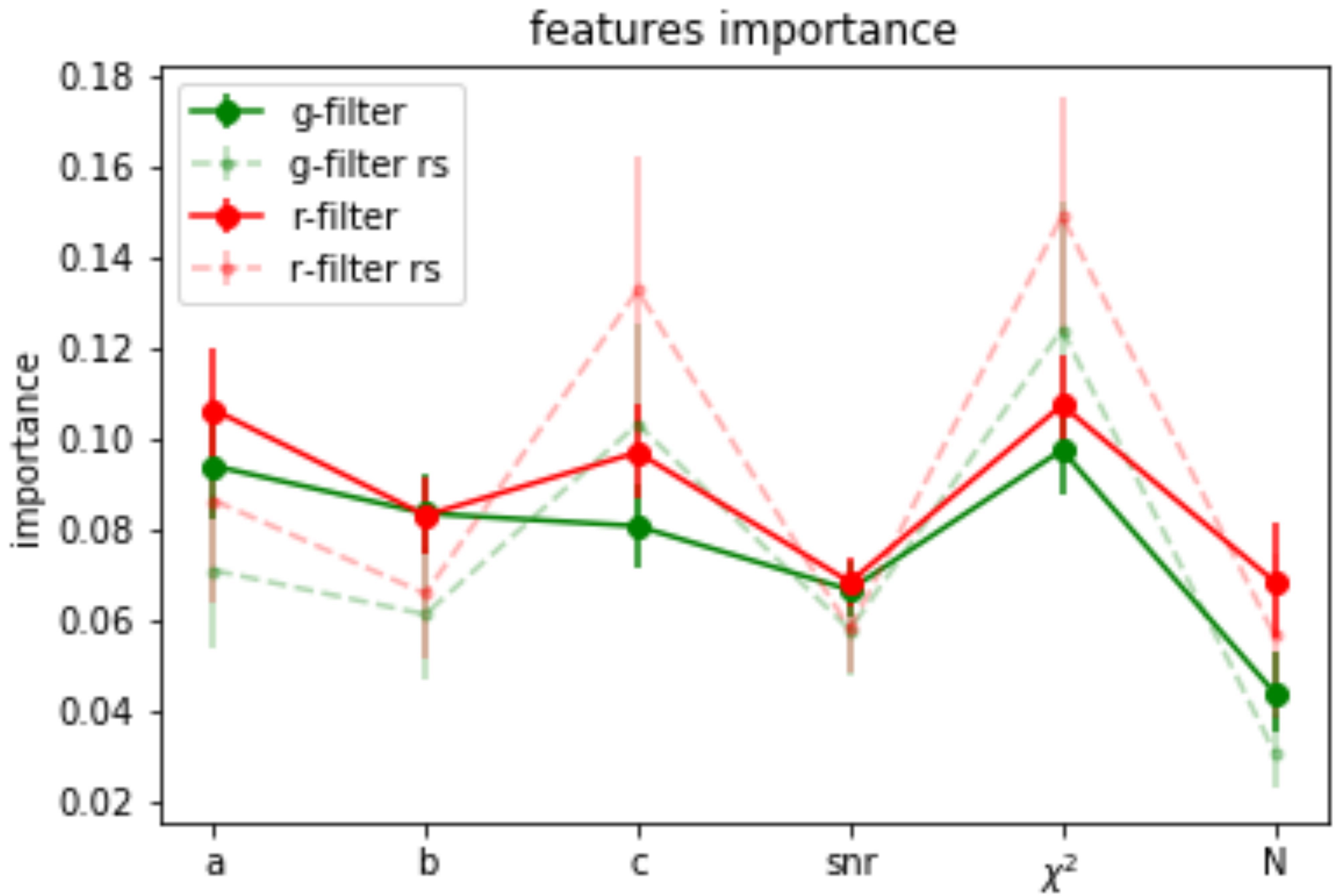
i) from ZTF simulations, we have to reconstruct a dataset of rising light curves.

ii) *Bazin's* fit does not work well with a few data points (more later)

iii) sometimes there are seasonal gaps in the time series (telescope off for a season etc.)

iv) some of the data points for the flux are negative (with values below -10)

Features importance



Estimate of how many days before max ?

1.typical length of a light curve last ? : 30 data points

2. typical length of rising part ? (1/3 of the light curve I.e. 10 points)

3. Typical prediction sent out ? (probably 5 points on average)

4. Typical N points before max ? (10 -5 = 5)

5. Confirm with predicted data from the broker ?