Fink broker

The ELAsTiCC data challenge: preparing the Fink broker for LSST
LSST France, LPNHE, 29 November 2022

Emille Ishida (FR), Julien Peloton (FR) and Anais Möller (Australia)
Andre Santos (BR), Bernardo Fraga (BR), Clecio de Bom (BR),
Etienne Russeil (FR), Marco Leoni (FR), Tarek Alam (UK)
on behalf of the Fink Team

Now and then

ZTF

2018 - now

1.4 TB/night

Primary mirror: 1.2 m
2 filters
~ 300k alerts/night

VERA C. RUBIN OBSERVATORY

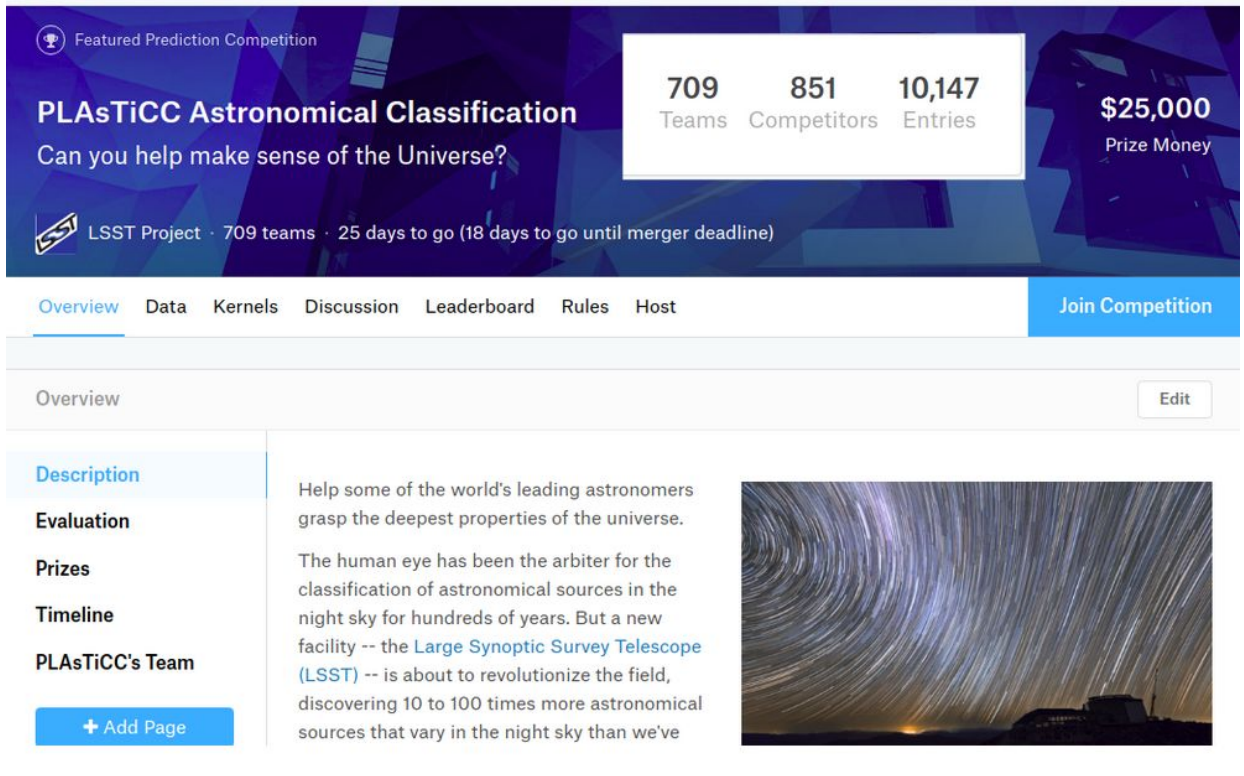
15 TB/night

From 2024

Primary mirror: 8.4 m
6 filters
1 million alerts/night

Remembering PLAsTiCC

https://www.kaggle.com/c/PLAsTiCC-2018



ELAsTiCC (PLAsTiCC v2)

Main goal

Prepare and test broker infrastructure under LSST-like requirements

1 - **July - September 2022**
10% of the entire test set streamed every 2 weeks

2 - **October - December 2022**
Full stream sent during 3 months + gap period + another 3 months stream

Secondary goal

Test classifiers

- Training set is not even close to representative
- Cadence of the survey is different from the test sample

enrich, ***select***, *distribute to maximise science*

Machine Learning

- Early SNIa: Random Forest (RF) and Active Learning (Ishida+2019, Leoni+2022)
- Supernova: RNNs (Möller+2019)
- Microlensing: RF (Godines, Bachelet+2019)
- Fast transients: RF (Biswas+2022)
- Multi-class: LSTM (CATS, Fraga+ in prep.)
- AGN and PISN: Summary Statistics + Symbolic Regression (Russeil+2022)
- [In progress] Multi-class: Transformers (Allam+2021)

These algorithms + cuts select candidates! Möller, Peloton, Ishida+2020

This is different in ELAsTiCC as we can't query catalogues nor have additional info

Slide by Anais Moller

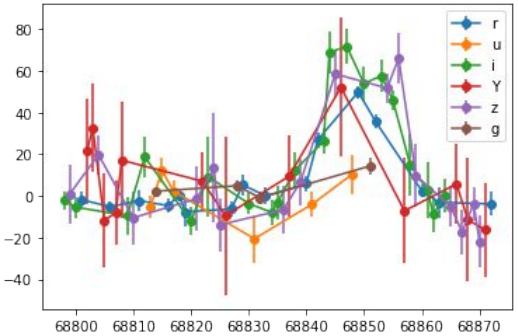
Machine Learning for ELAsTiCC

- Size and balance of subclasses affect the potential performance of algorithms when using only training set (no augmentation)
- Taxonomy
- Baselines
- Magnitude limits and detections

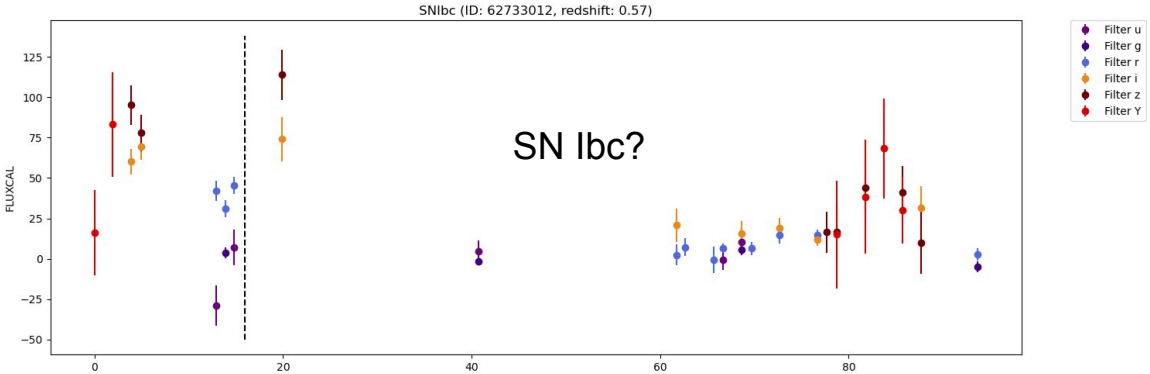
```
Alert/Real/Static
├── 0 Static/Other
├── 1 Non-Recurring
│   ├── 10 Non-Recurring/Other
│   ├── 11 SN-like
│   │   ├── 110 SN-like/Other
│   │   ├── 111 Ia
│   │   ├── 112 Ib/c
│   │   ├── 113 II
│   │   ├── 114 Iax
│   │   └── 115 91bg
│   ├── 12 Fast
│   │   ├── 120 Fast/Other
│   │   ├── 121 KN
│   │   ├── 122 M-dwarf Flare
│   │   ├── 123 Dwarf Novae
│   │   └── 124 uLens
│   └── 13 Long
│       ├── 130 Long/Other
│       ├── 131 SLSN
│       ├── 132 TDE
│       ├── 133 ILOT
│       ├── 134 CART
│       └── 135 PISN
└── 2 Recurring
    ├── 20 Recurring/Other
    ├── 21 Periodic
    │   ├── 210 Periodic/Other
    │   ├── 211 Cepheid
    │   ├── 212 RR Lyrae
    │   ├── 213 Delta Scuti
    │   ├── 214 EB
    │   └── 215 LPV/Mira
    └── 22 Non-Periodic
        ├── 220 Non-Periodic/Other
        └── 221 AGN
```

Slide by Anais Moller

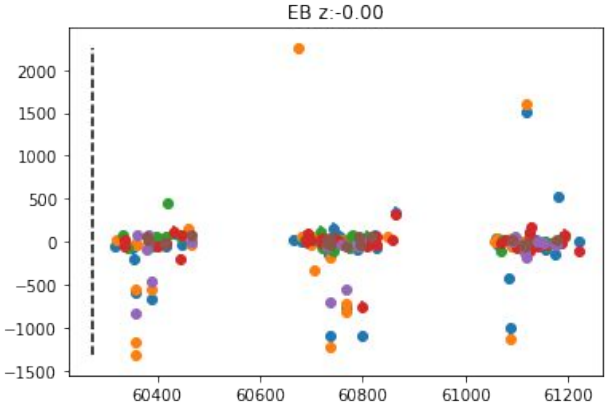
Machine Learning for ELAsTiCC



SN Ia



SN Ibc?



EB

SN Ia!

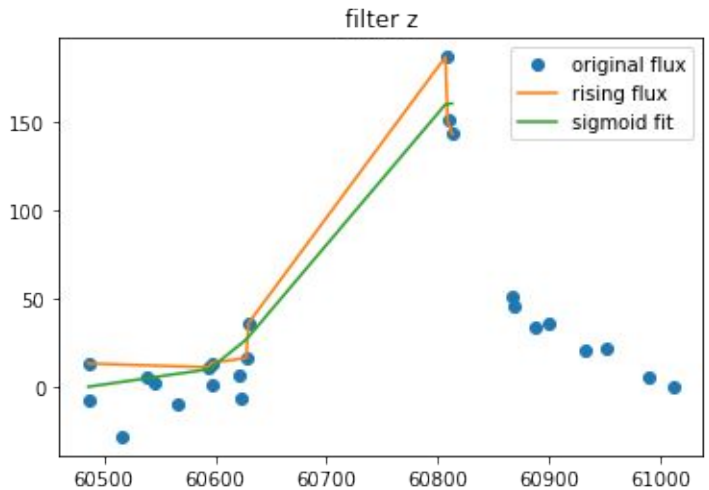
Slide by Anais Moller

Machine Learning for ELAsTiCC

Early SNIa with Active Learning Leoni+2022

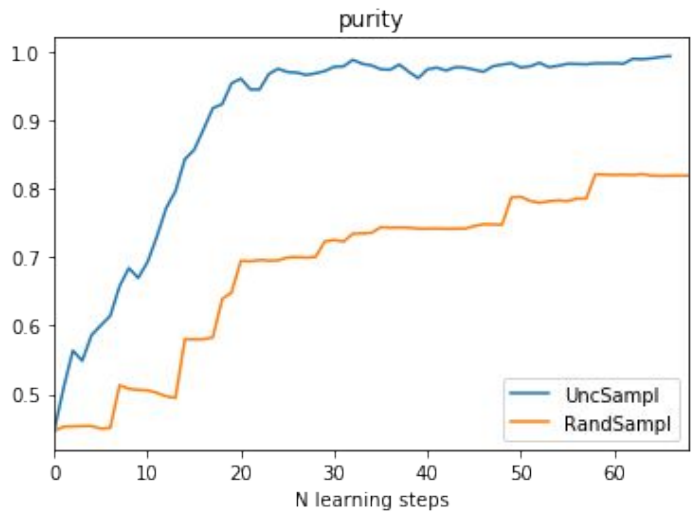
Only rising events and FLUXCAL>200

Feature extraction



See Marco's talk

AL loop



Purity : TP/TP+FN

Slide by Anais Moller

Machine Learning for ELAsTiCC

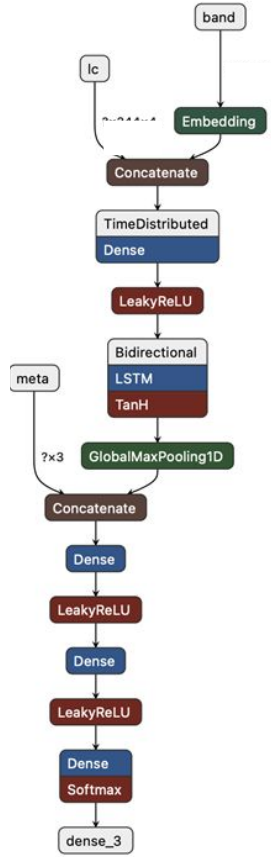
Very very preliminary!

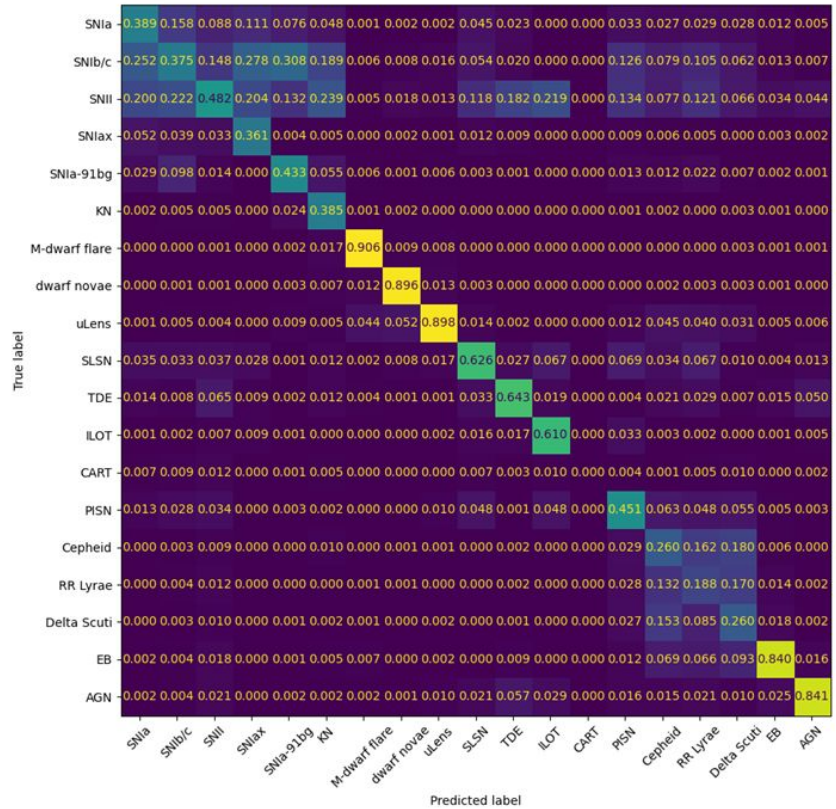
Long Short-term Memory Deep Network

Using only the first alert and forced photometry

Metadata used:

- redshift+error
- host galaxy redshift+error
- MW extinction





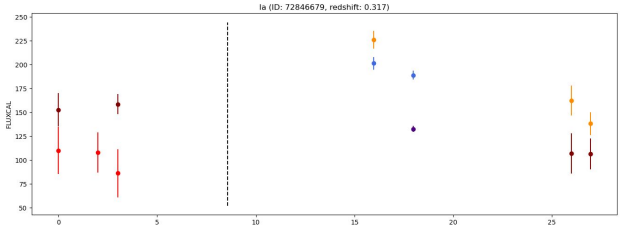
Work by Clecio, Bernardo and Andre

Slide by Anais Moller

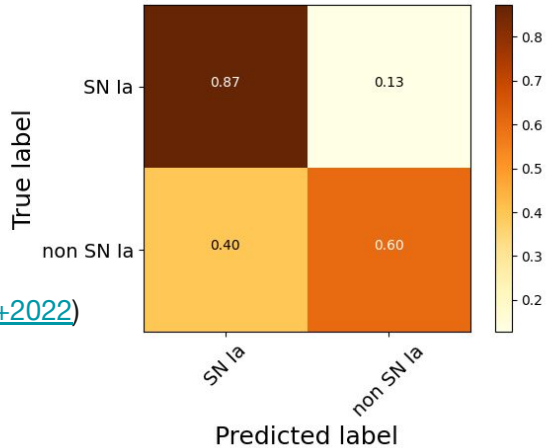
Machine Learning for ELAsTiCC

Very very preliminary!

SNIa vs non-Ia with SuperNNova (RNN) Möller+2019



- ✅ Adapted algorithm to LSST filters/inputs
- 🚧 Training set curation
  - Time window
  - Sampling + magnitude limits



Accuracies SNIa vs non-Ia:

~75% complete-lightcurve, partial are more challenging

(compared to DES SNIa <2% contamination w. ML scores+selection cuts Vincenzi+2021 Möller+2022)

Multi-class on the works…

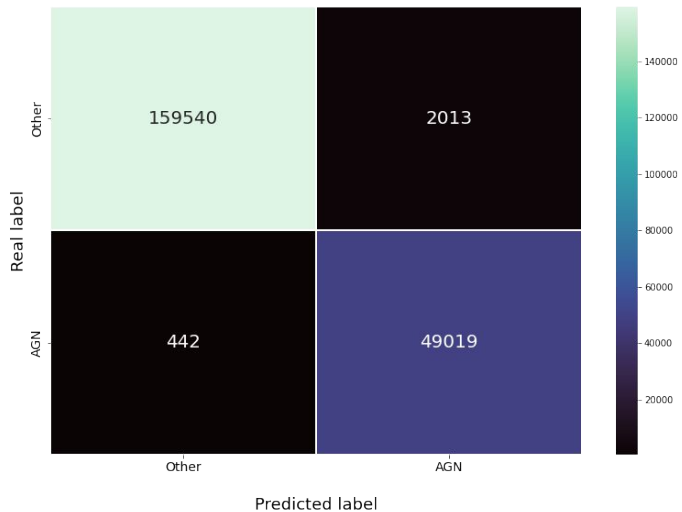
Slide by Anais Moller

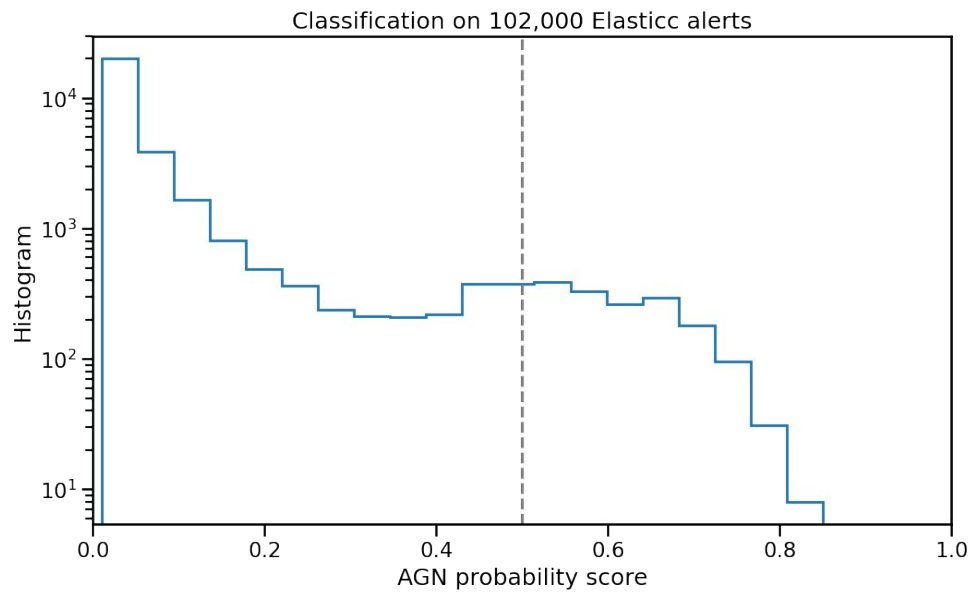
Machine Learning for ELAsTiCC

AGN

Summary statistics + colors from parametric function using symbolic regression

Accuracy, efficiency, purity > 95% *(in training)*





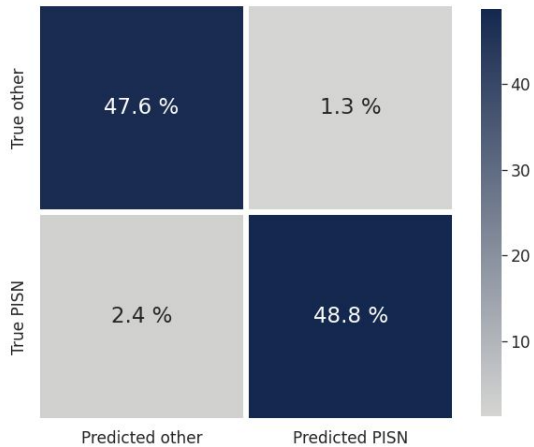
Work by Etienne and Julien

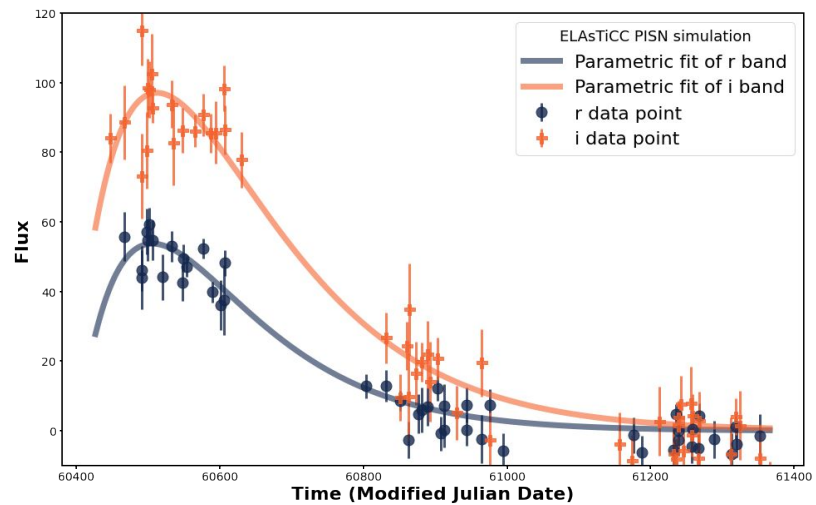
Machine Learning for ELAsTiCC

PISN

Summary statistics + colors and fit from parametric function using *multiview* symbolic regression applied to SNAD160

Accuracy, efficiency, purity > 95% *(in training)*





Work by Etienne and Julien

FINK & ELAsTiCC

✅ Infrastructure tests successfully ongoing

🚧 ML phase

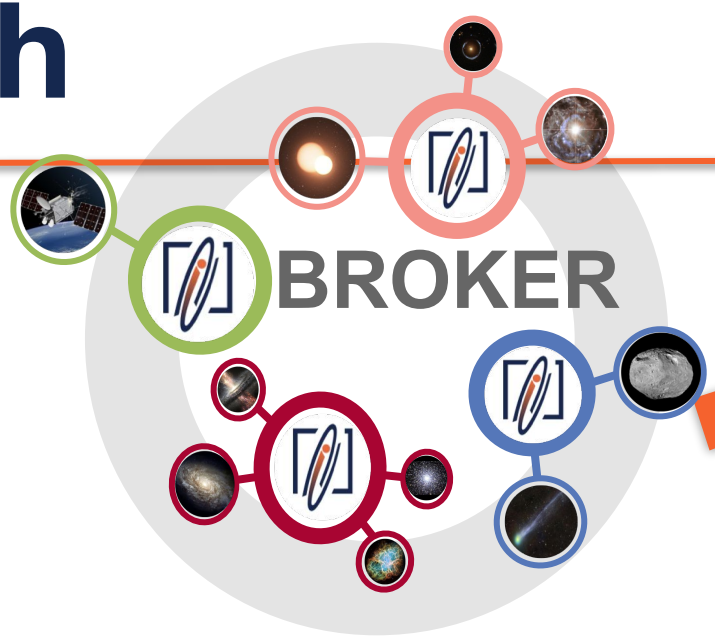
- ❗ Real data uses ML score**s** + selection cuts + catalogues + context
- ❗ Training set curation is non-trivial
- ❗ Different ML algorithms -> non-normalized scores…

From the 2022 Fink Hackathon … an ELAsTiCC paper is in preparation!

Stay tuned!

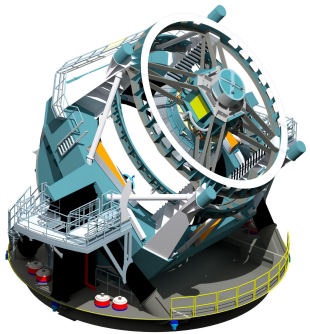
Original slide by Anais Moller

Data path



Machine learning
Catalog association
Streams join

every ~30 seconds down to mag ~24



10 million alerts per night…

We would like the interesting ones

Rubin broker landscape