Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC)

a Kaggle Competition

kaggle

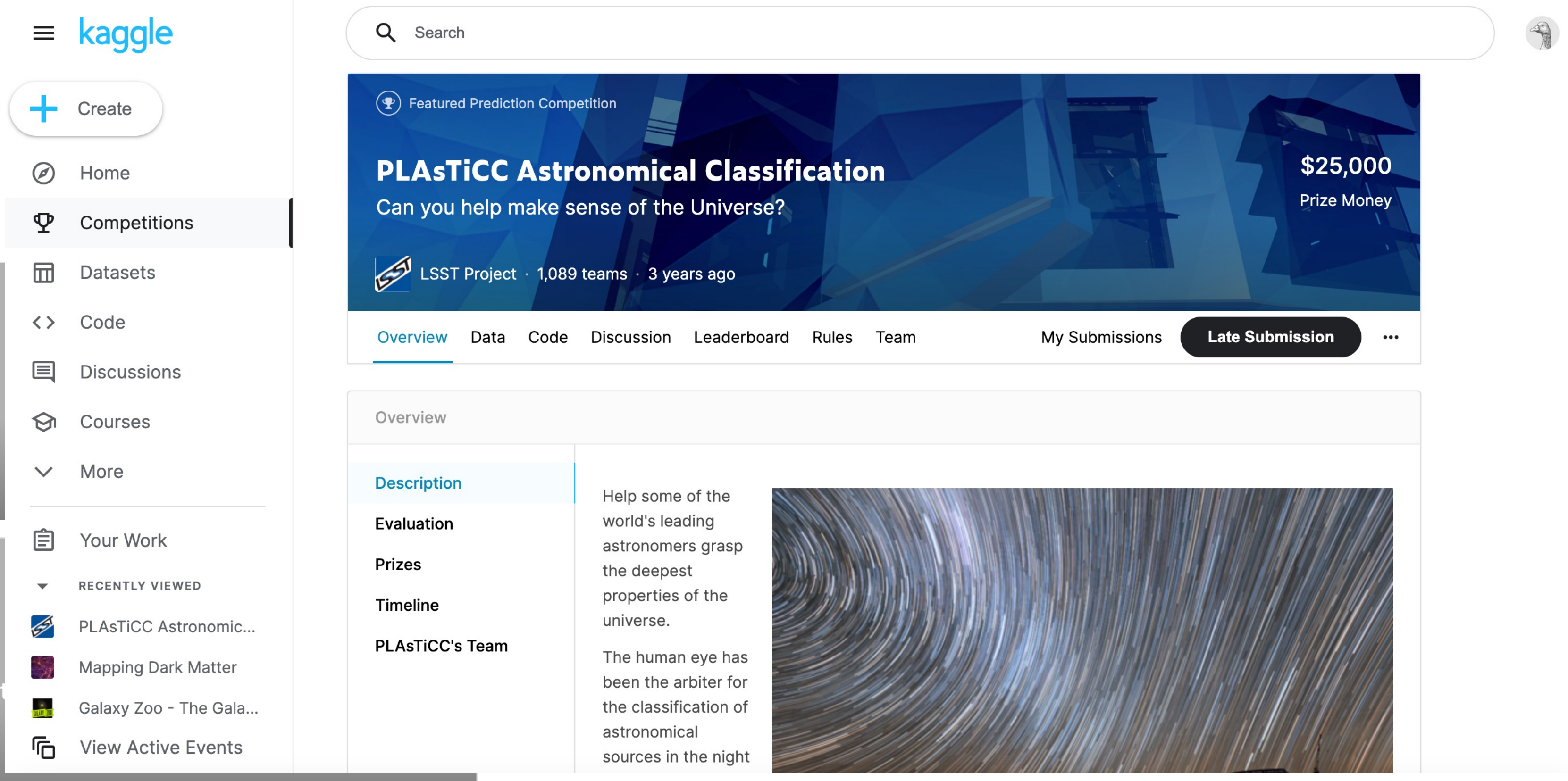
Outline

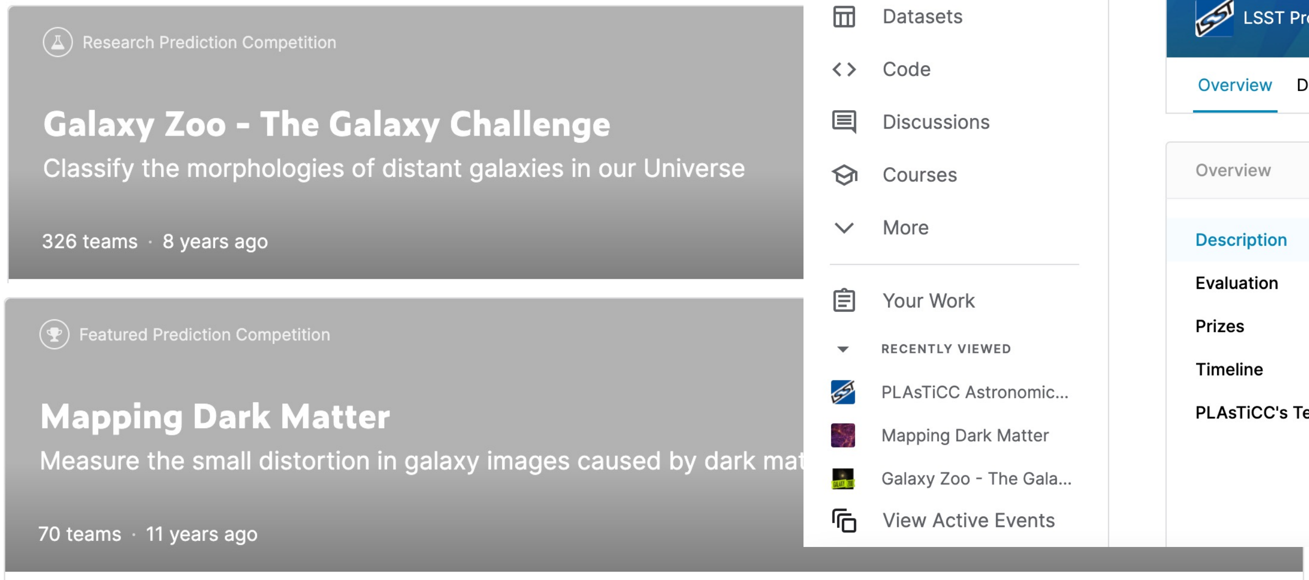
- Introduction & Motivation
  - What is a Kaggle competition?
  - Why PLAsTiCC?
  - How PLAsTiCC mimic the LSST observation?
- Methods
  - How to evaluate the performance of different classifiers?
  - What is the Classification strategy for 1st, 2nd, and 3rd place?
- Results
  - What's the results for the 1st, 2nd, and 3rd place?
- Conclusion

PLAsTiCC is a Kaggle Competition

What is Kaggle?

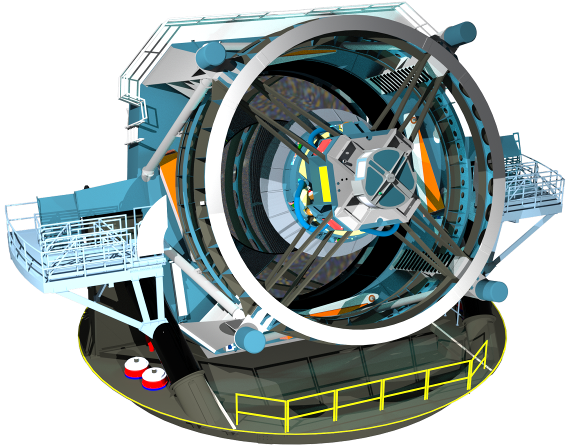
Kaggle is an online community platform for data scientists and machine learning enthusiasts.





Why PLAsTiCC?

- Motivation:
  - How well can we **classify** astronomical **transients** and **variables** from a **simulated light curve** data set designed to mimic the data from **LSST**?
- Challenge:
  - Classify the data into 15 classes



The Legacy Survey of Space and Time (LSST)

- image the entire Southern sky roughly every few nights and over a ten-year duration

Credit: NASA, LSST

How PLAsTiCC mimic the LSST observation?

1. LSST may well discover **new** transients

What types of objects?

Transients & variables stars

- Extragalactic models
  1-5. Supernovae (SNe)
    1. SNIa
    2. SNIa-91bg
    3. SNIax
    4. SNII
    5. SNIbc
  6. Super-Luminous Supernovae (SLSN)
  7. Tidal Disruption Event (TDE)
  8. Kilonova (KN)
  9. Active Galactic Nucleus (AGN)

- Galactic models
  10. RR Lyrae
  11. Mira variables
  12. Eclipsing binaries (EB)
  13. M-dwarf flares
  14. single microlensing events (μlens-Single)

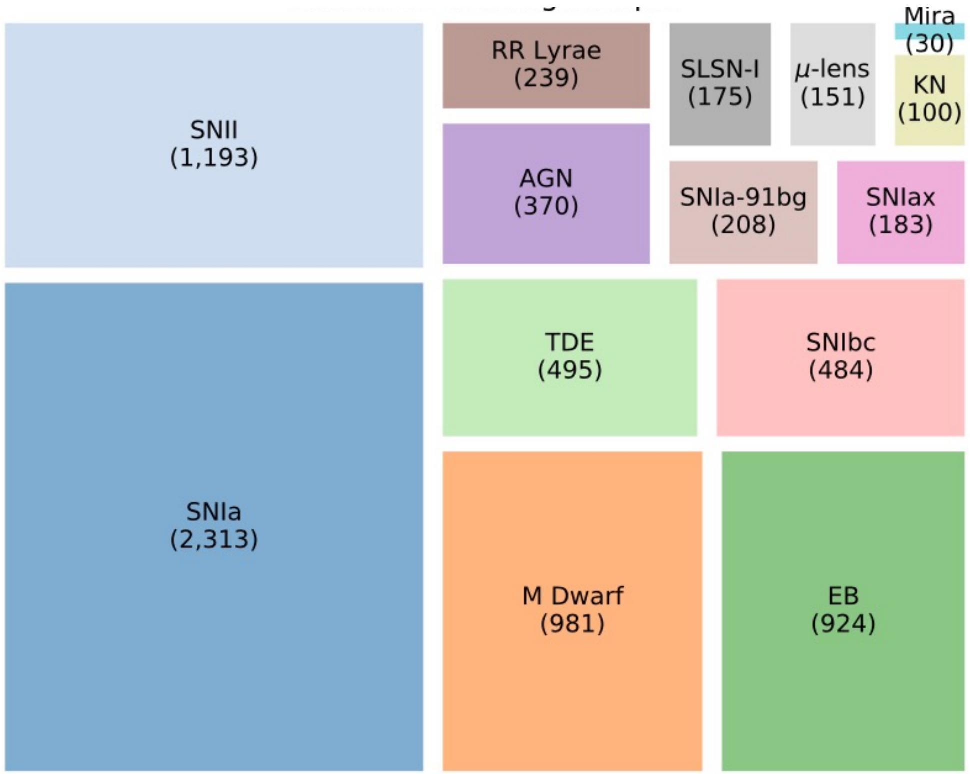
- "Novel" (Other) objects
  15. Pair Instability Supernovae (PISNs)
  16. Calcium-rich transients (CaRTs)
  17. Intermediate Luminosity Optical Transients (ILOTs)
  18. binary micro-lensing events (μlens-Binary)

How PLAsTiCC mimic the LSST observation?

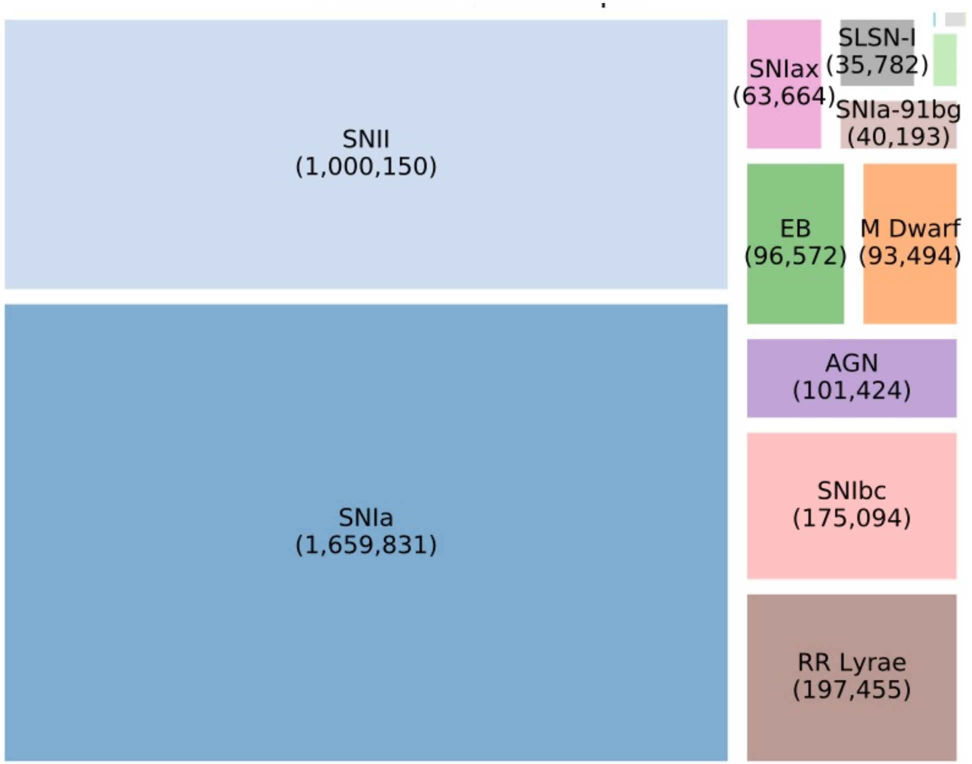
1. LSST may well discover **new** transients
   → **“Other ” object**: 4 models are not provided in the training sets
2. **Less available data** from current and near-term spectroscopic surveys that will be available data at the start LSST (Huge data)

What data?

Training Sets (8000)



Test sets (3.5 M)



How PLAsTiCC mimic the LSST observation?

1. LSST may well discover **new** transients

   → **"Other " object**: 4 models are not provided in the training sets

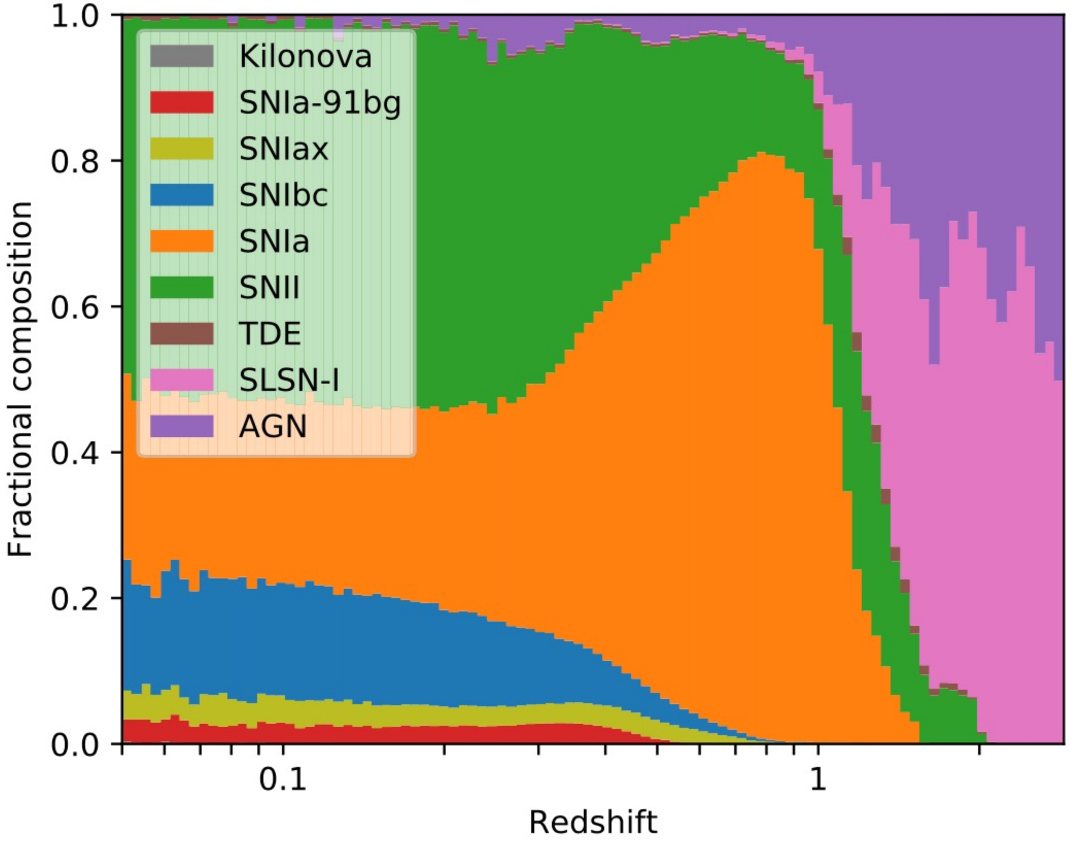
2. **Less available data** from current and near-term spectroscopic surveys that will be available data at the start LSST (Huge data)

   → **Non-representativity**: training sets (8000) vs. test sets (3.5 M)

   → **Class imbalance**

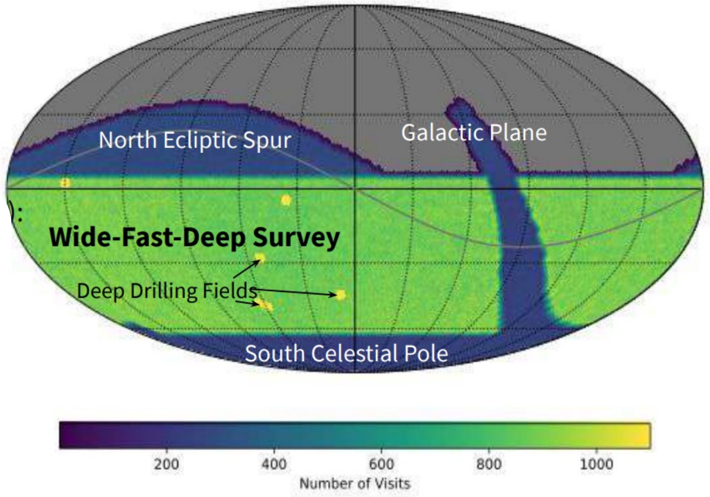
How PLAsTiCC mimic the LSST observation?

1. LSST may well discover **new** tran
   → **“Other ” object**: 4 models are not
2. **Less available data** from current
   surveys that will be available dat
   → **Non-representativity**: training sets
   → **Class imbalance**



How PLAsTiCC mimic the LSST observation?

1. LSST may well discover **new** transients
   → **"Other " object**: 4 models are not provided in the training sets
2. **Less available data** from current and near-term spectroscopic surveys that will be available data at the start LSST (Huge data)
   → **Non-representativity**: training sets (8000) vs. test sets (3.5 M)
   → **Class imbalance**
3. Observations in both
   - Deep Drilling Fields (DDF): Fewer objects with high SNR
   - Wide-Fast-Deep (WFD): More objects with low SNR



Credit: LSST

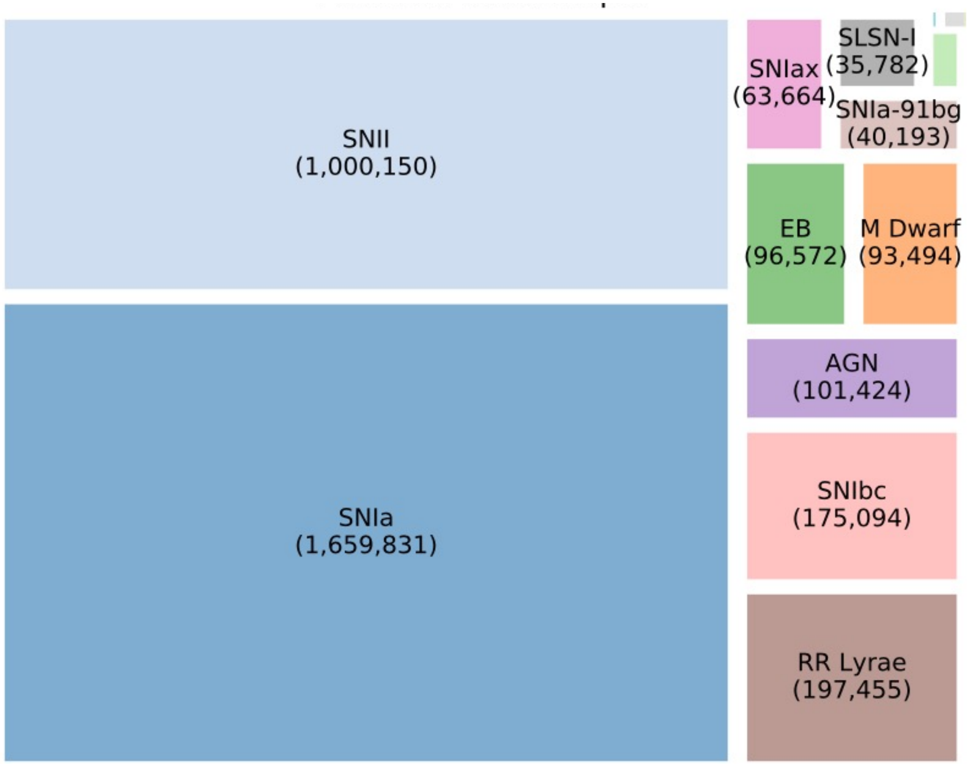
What data?

Training Sets (8000)

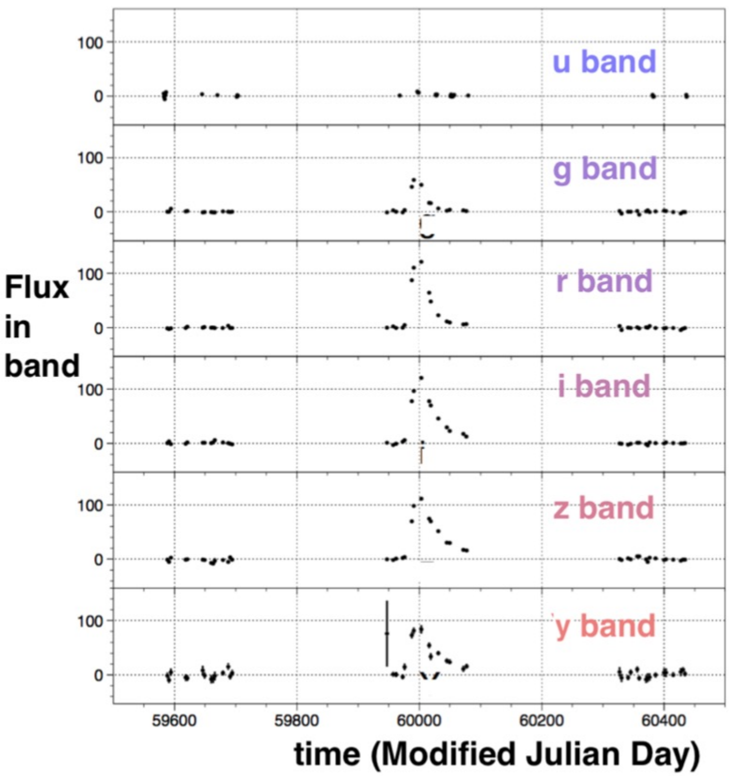
Ra & Dec
Gal long & Lat.
(73%) DDF (1%)
(27%) WFD (99%)
host galaxy spec-z (3.6%)
host galaxy photo-z
Distance module
MW Extinction
Detection
Class (0%)

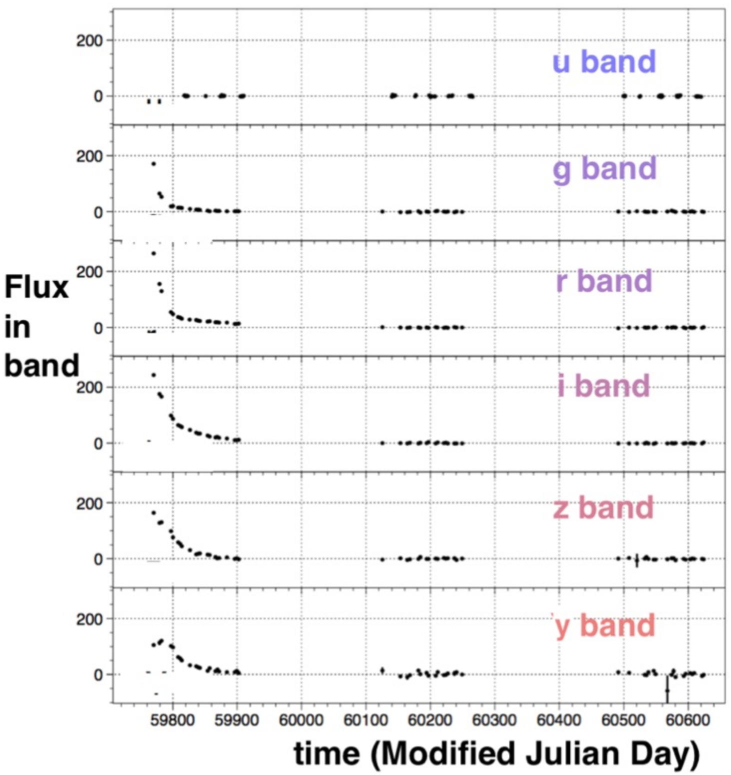
Light curve
Modified Julian Date
LSST passband (ugrizY)
Flux (err)

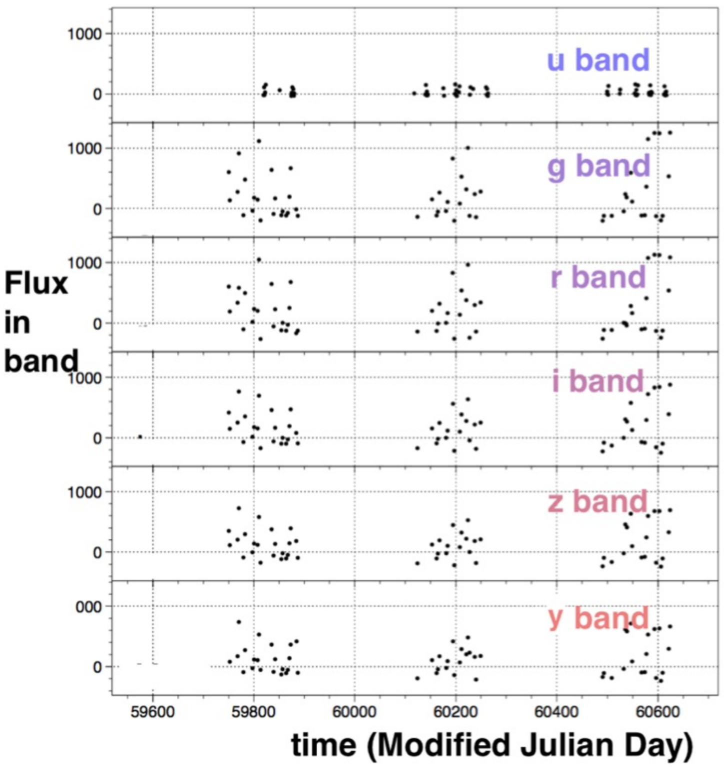
Test sets (3.5 M)



Light curve data







How PLAsTiCC mimic the LSST observation?

1. LSST may well discover **new** transients
   → **"Other " object**: 4 models are not provided in the training sets
2. **Less available data** from current and near-term spectroscopic surveys that will be available data at the start LSST (Huge data)
   → **Non-representativity**: training sets (8000) vs. test sets (3.5 M)
   → **Class imbalance**
3. Observations in both
   - Deep Drilling Fields (DDF): Fewer objects with high SNR
   - Wide-Fast-Deep (WFD): More objects with **low SNR**

   → **Non-representativity**: test data are more noise and sparse

How to evaluate the performance of diff. classifiers?

- The PLAsTiCC metric
  - weighted multi-class logarithmic loss metric

$$\text{Log Loss} = -\left(\frac{\sum_{i=1}^{M} w_i \cdot \sum_{j=1}^{N} \frac{y_{ij}}{N_i} \cdot \ln p_{ij}}{\sum_{i=1}^{M} w_i}\right)$$

N is the number of objects in the class set
M is the number of classes
y_{ij} is 1 if observation i belongs to class j and 0 otherwise
p_{ij} predicted probability that observation i belongs to class j.

Class Name	Class number	Weight
Point source μ-lensing	6	2
Tidal disruption event (TDE)	15	2
Eclipsing binary event (EBE)	16	1
Core-collapse supernova Type II (SNII)	42	1
Supernova Type Ia-x (SNIax)	52	1
Mira Variable	53	1
Core-collapse Supernova Type Ibc (SNIbc)	62	1
Kilonova (KN)	64	2
M-dwarf	65	1
Supernova Type Ia-91bg (SNIa-91bg)	67	1
Active galactic nucleus (AGN)	88	1
Supernova Type Ia (SNIa)	90	1
RR Lyrae	92	1
Super Luminous Supernova (SLSN)	95	2
'Other' class	99	2

Rare or interesting objects

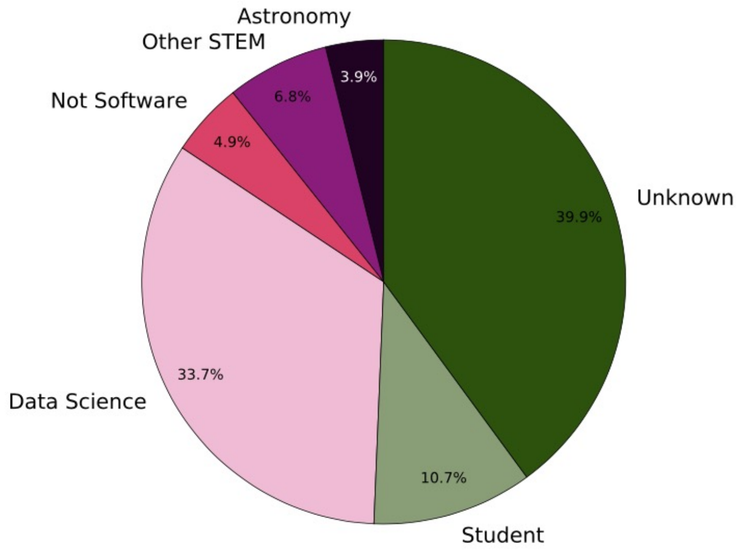
The PLAsTiCC results

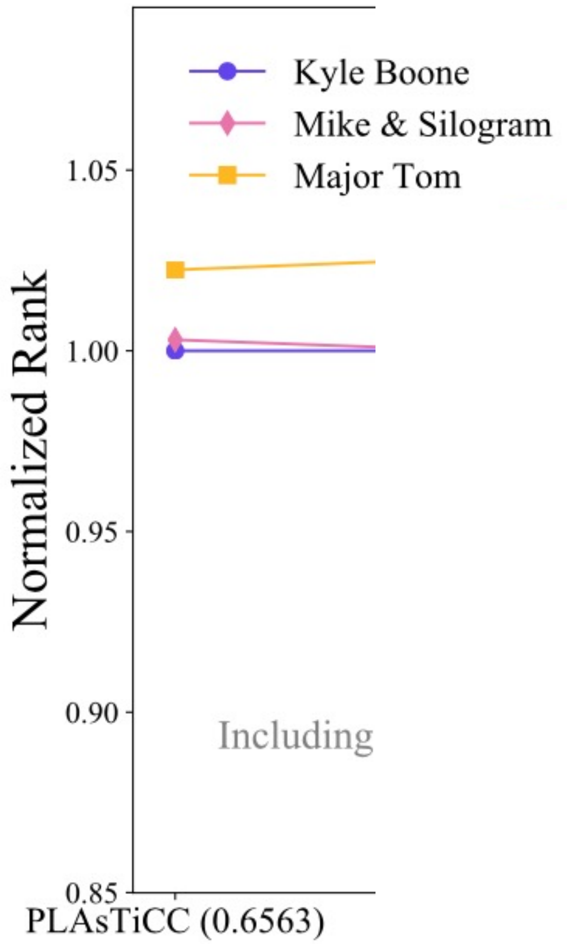
- from Sep 28, 2018 until Dec 17, 2018
- received entries from 1094 teams around the world from a variety of different groups.

(1) Kyle Boone

(2) Mike & Silogram

(3) Major Tom, mamas & nyanp





Summary of method used

Name	Boosted Decision Trees			Neural Nets			
	LightGBM	CatBoost	XGBoost	NN	CNN	RNN	MLP
Kyle Boone	✓	✗	✗	✗	✗	✗	✗
Mike & Silogram	✓	✗	✗	✗	✗	✓	✗
Major Tom, mamas & nyanp	✓	✓	✗	✗	✓	✗	✗
Ahmet Erdem	✓	✗	✗	✓	✗	✗	✗
SKZ Lost in Translation	✓	✗	✗	✗	✗	✓	✓
Stefan Stefanov	✗	✗	✗	✓	✗	✗	✗
rapids.ai	✓	✗	✗	✗	✗	✓	✓
Three Musketeers	✓	✓	✓	✗	✓	✗	✗
Simon Chen	✓	✗	✗	✗	✗	✗	✗
Go Spartans!	✓	✗	✓	✗	✗	✗	✗

- **no strong preference** for any one machine learning architecture was found
- successful classification procedures often **combine a range of methods** from neural networks to tree-based approaches (**ensemble)**

Summary of method used

Final Rank	Name	Public score	Private score	Class imbalance mitigation	Non-representativity mitigation	'Other' class	Ensemble	Light curve fit
1	Kyle Boone	0.6706	0.6850	Light curve augmentation	Data degradation	Weighted Average	No	GP
2	Mike & Silogram	0.6937	0.6993	-	Model spec-z, Data degradation	Probability Scaling	Yes	-
3	Major Tom, mamas & nyanp	0.6804	0.7002	-	Model spec-z, Data degradation	Tuned Combination	Yes	SALT2
4	Ahmet Erdem	0.6913	0.7042	-	Model spec-z	Weight based on top prediction	Yes	Bazin
5	SKZ Lost in Translation	0.7397	0.7523	Flux error augmentation	Removed photo-z	Probability Scaling	Yes	-
6	Stefan Stefanov	0.7933	0.8017	Flux augmentation	Dropped light curve points, photo-z augmentation	-	Yes	-
8	rapids.ai	0.7922	0.8091	Pseudo label on test data	-	-	Yes	- -
9	Three Musketeers	0.7921	0.8131	-	Dropped light curve p oints	Probability Scaling	Yes	-
11	Simon Chen	0.7948	0.8225	-	-	-	No	SALT2
12	Go Spartans!	0.8112	0.8265	-	-	Probability Scaling	Yes	Bazin

What is the Classification strategy for 1st place?

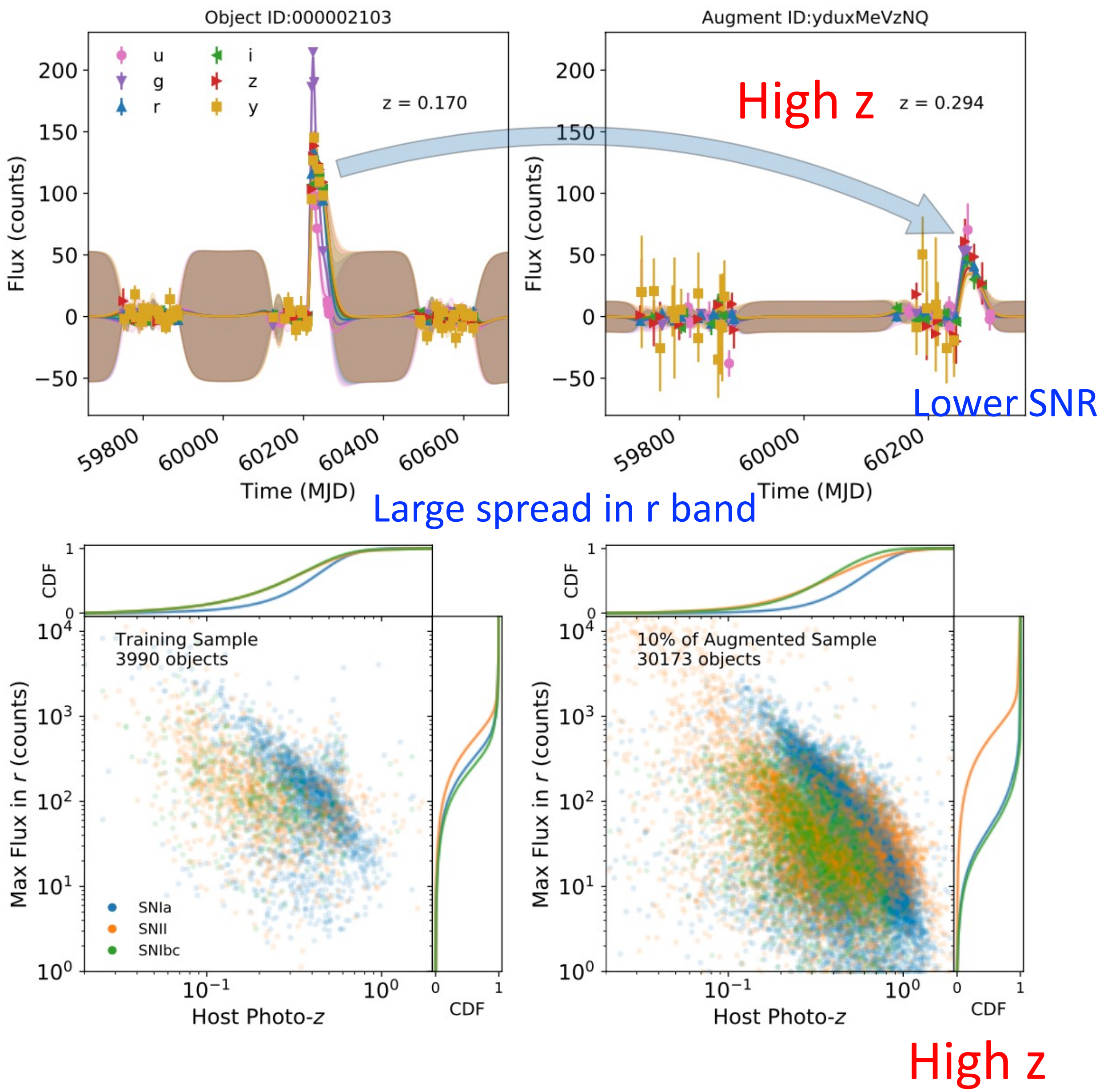
- **Feature**
  - **Light curve fit**: Gaussian process (GP) fitting
  - Feature selection
    - Color & time for SNIa-like objects
    - Lack of pre-maximum observations
- **Classification algorithms**
  - Boosted Decision Trees: LightGBM

Data augmentation

- Class imbalance
  - Light curve augmentation (shift data in time, randomly dropping time position)
- Non-representativity
  - Data degradation (increase SNR)
  - redshift augmentation (low-z -> high z)
- ‘Other’ class
  - probing the leader board:
    - weighted average of the SN class probabilities (SNIa & SNIbc & SNII)

How the 1st place team do data augmentation?

- light curve space
  - interpolated using GP interpolation
  - the object is moved to higher z
  - the errors on the light curve points are increased
- feature space
  - has a larger spread in the maximum r−band flux
  - extends each class to higher redshift



What is the confusion matrix?

Binary Classification

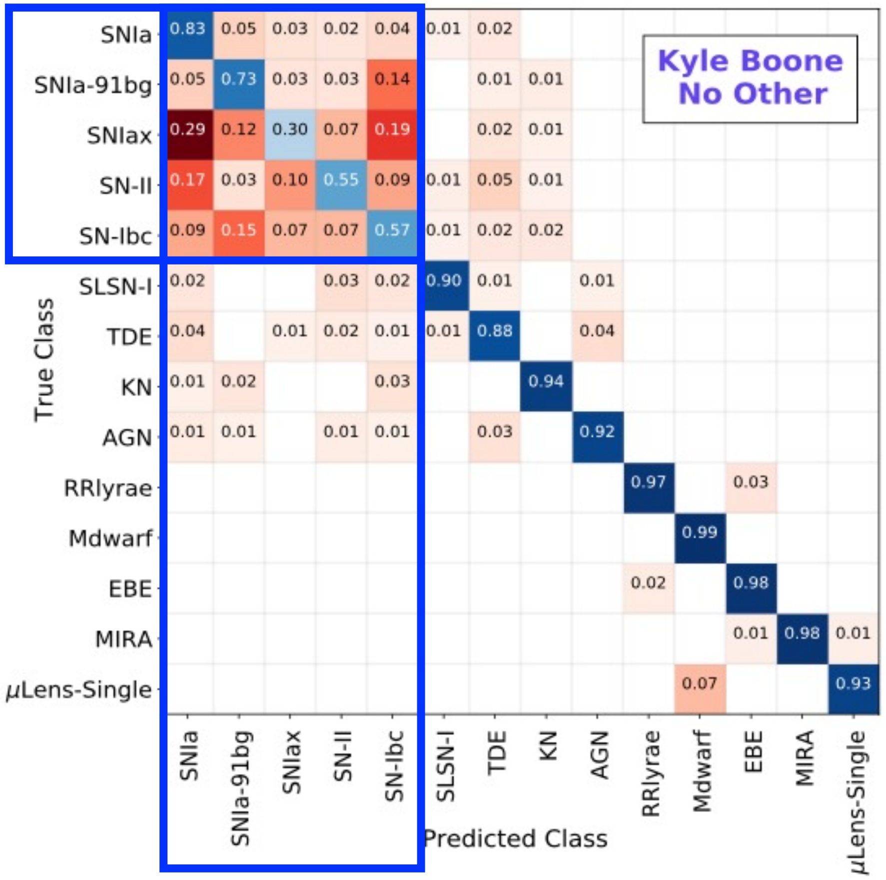
Predicted Class \ True Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN

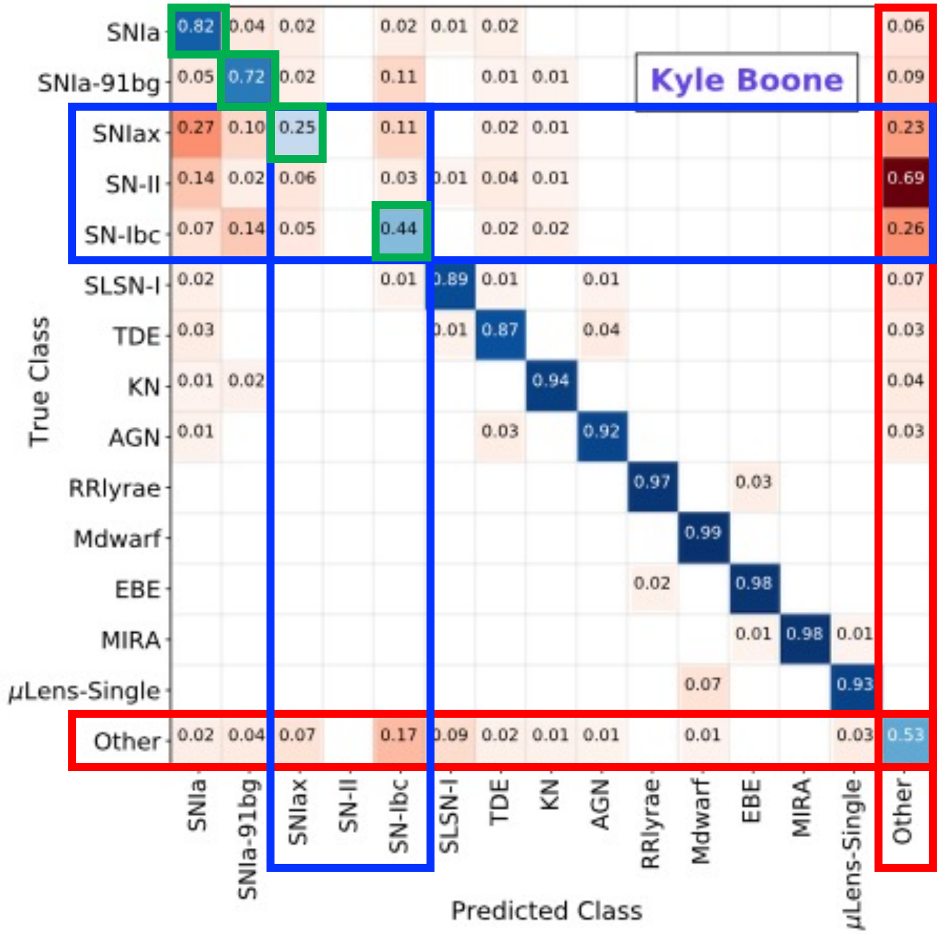
Multi-Class Classification

Predicted Class \ True Class	Apple	Orange	Mango
Apple	7	8	9
Orange	1	2	3
Mango	3	2	1

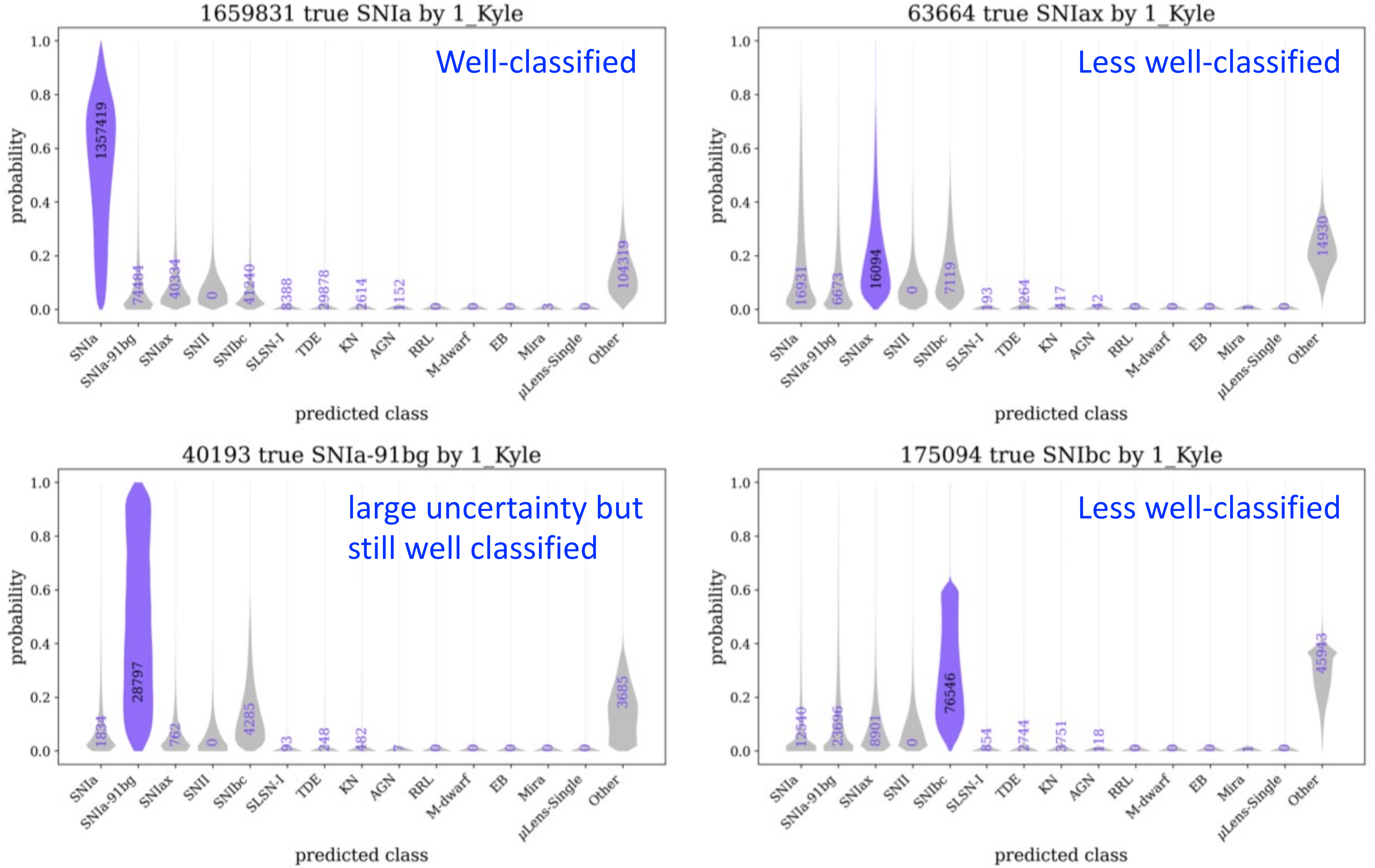
How good is the classification? confusion matrices

- Different kinds of **supernovae** are relatively **hard** to distinguish
- Degeneracy btw SN II and the "other"





How good is the classification for 1st place team ?



What is the Classification strategy for 2nd place?

- **Feature**
  - **Light curve fit**: (?!)
  - Feature selection
- **Classification algorithms: ensemble**
  - Boosted Decision Trees
    - LightGBM
  - Recurrent neural network (RNN)

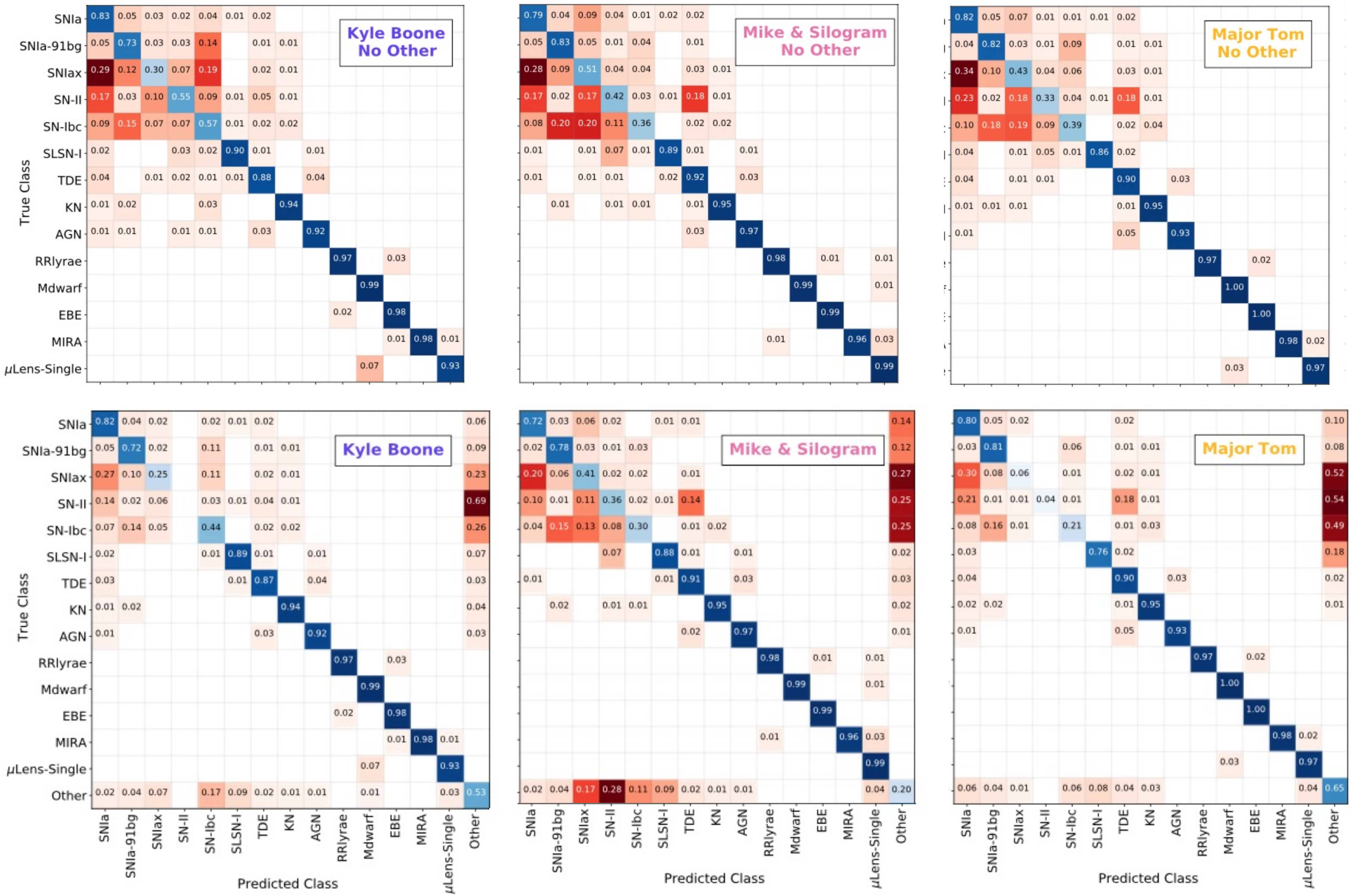
Data augmentation
  - dropped 30% of the measurements
- Class imbalance (?!)
- Non-representativity
  - Data degradation
  - Model spec-z
  - Address the redshift and the flux
- ‘Other’ class
  - probing the leader board:
    - Probability scaling: scaling all the class probabilities with weights based

What is the Classification strategy for 3rd place?

- **Feature**
  - Light curve fit: SALT2
  - Feature selection
- **Classification algorithms: ensemble**
  - Boosted Decision Trees:
    - LightGBM & CatBoost
  - Convolutional neural network (CNN)

Data augmentation

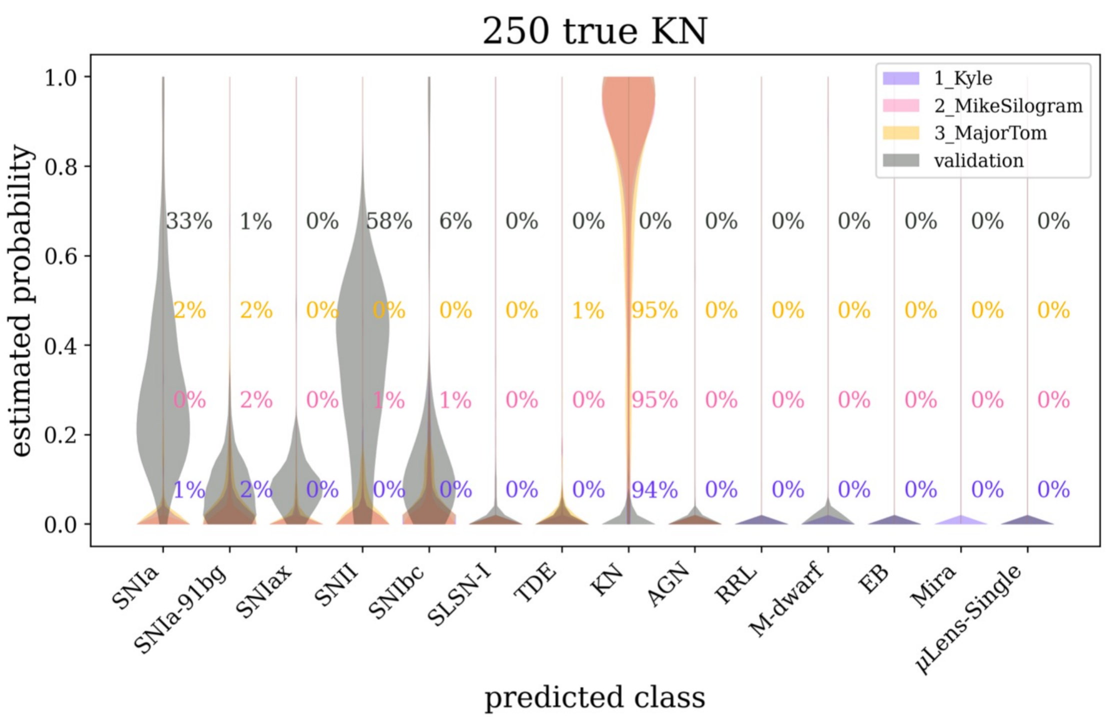
- Class imbalance (?!)
- Non-representativity
  - Data degradation
  - Model spec-z
- 'Other' class
  - probing the leader board:
    - Tuned combination: a power law combination of SNe



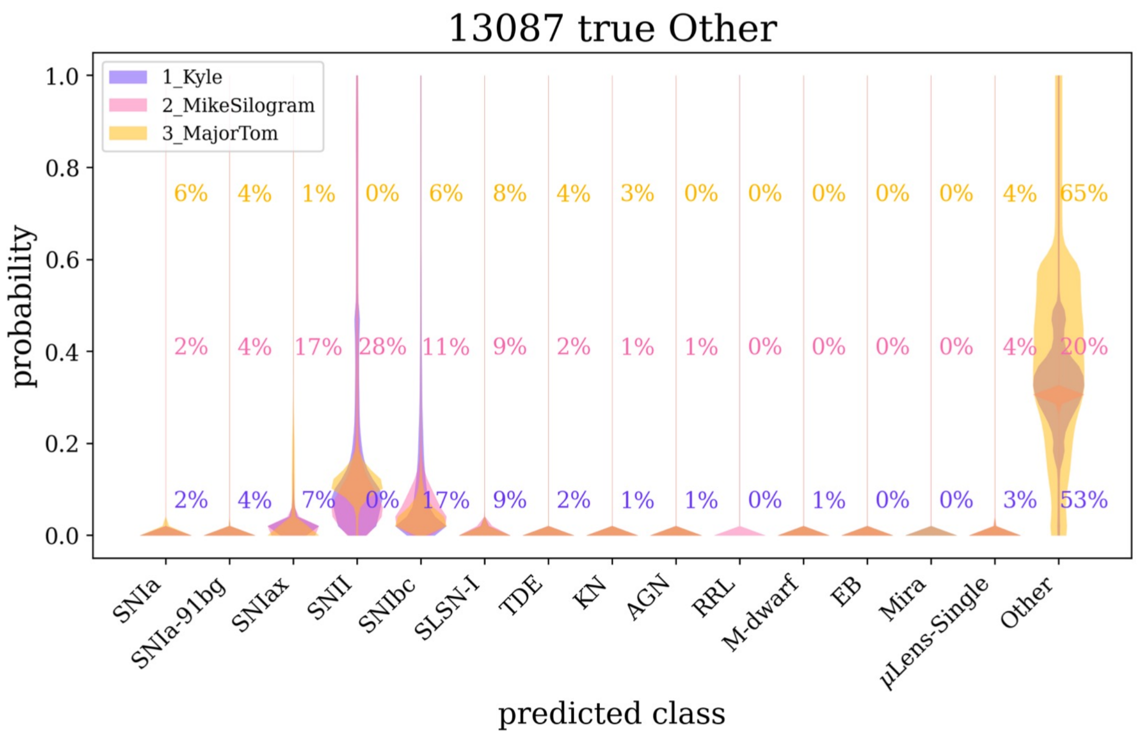
Pseudo-confusion matrices

How good is the classification?

Well-classified



Less well-classified



Conclusion

- **data augmentation of sparse non-representative training data** are key to accurate performance on the full test set
- **a metric weighted by the number of objects in a class** ensures that classification performance on the most populous class does not dominate overall performance
- **data fitting** and **feature selection** are the most computationally time-consuming parts of any classification challenge
- one successful approach is **pulling out a larger set of features over which to train**, and then **determining the feature importance** before classifying the objects on the **smaller** subset of features
- **no strong preference** for any one machine learning architecture was found; successful classification procedures often combine a range of methods from neural networks to tree-based approaches