The Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC)

Mi Dai

Rutgers University

on behalf of the PLAsTiCC Team:

Tarek Allam Jr., Anita Bahmanyar, Rahul Biswas, Alexandre Boucaud, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, Richard Kessler, Michelle Lochner, Ashish A. Mahabal, Alex I. Malz, Kaisey S. Mandel, Juan Rafael Martínez-Galarza, Jason D. McEwen, Daniel Muthukrishna, Gautham Narayan, Hiranya Peiris, Christina M. Peters, Kara Ponder, Christian N. Setzer, The LSST Dark Energy Science Collaboration, The LSST Transients, Variable Stars Science Collaboration

The Photometric LSST Astronomical Time Series Classification Challenge

The Large Synoptic Telescope (LSST)

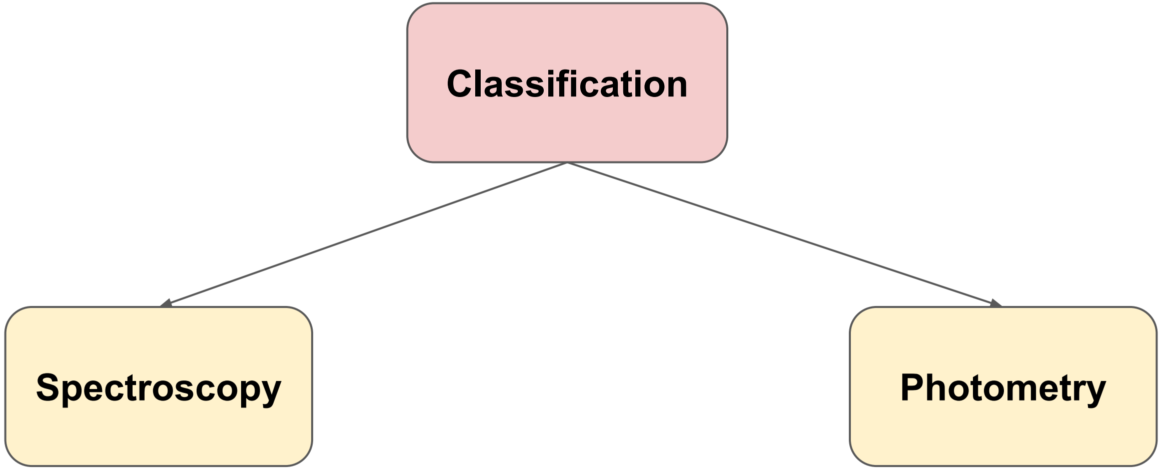
Credit: LSST Project/NSF/AURA

This telescope will produce the deepest, widest, image of the Universe:

- 27-ft (8.4-m) mirror, the width of a singles tennis court
- 3200 megapixel camera
- Each image the size of 40 full moons
- 37 billion stars and galaxies
- 10 year survey of the sky
- 10 million alerts, 1000 pairs of exposures, 15 Terabytes of data .. every night!

https://www.lsst.org/

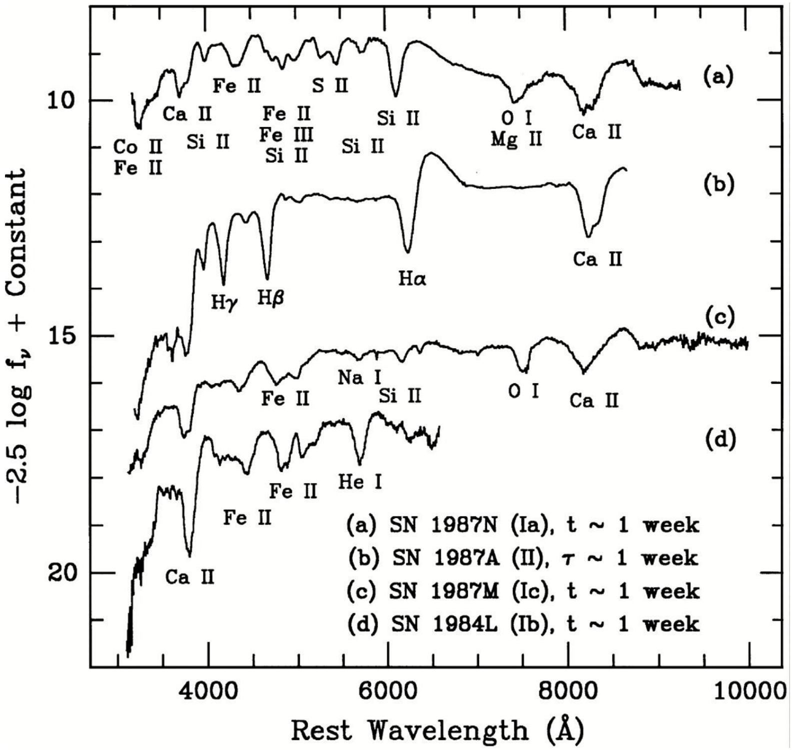
The Photometric LSST Astronomical Time Series Classification Challenge

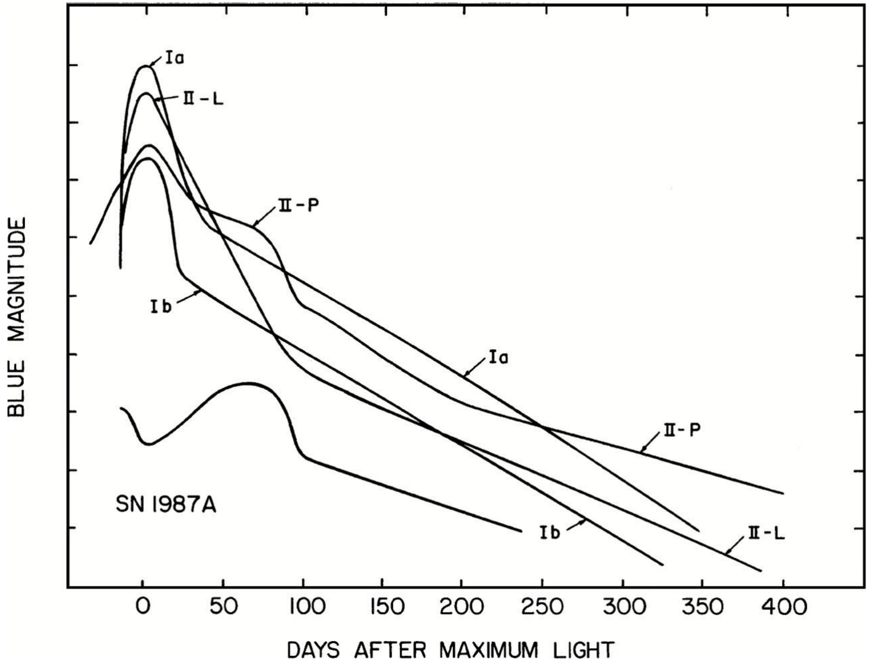


The Photometric LSST Astronomical Time Series Classification Challenge

Spectroscopy

Photometry



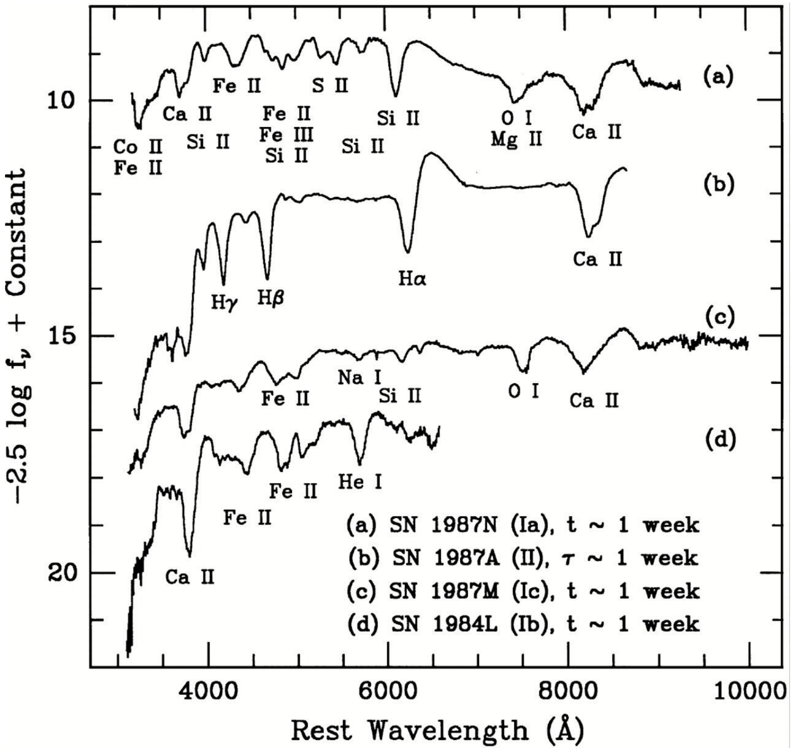


Filippenko 1997

The Photometric LSST Astronomical Time Series Classification Challenge

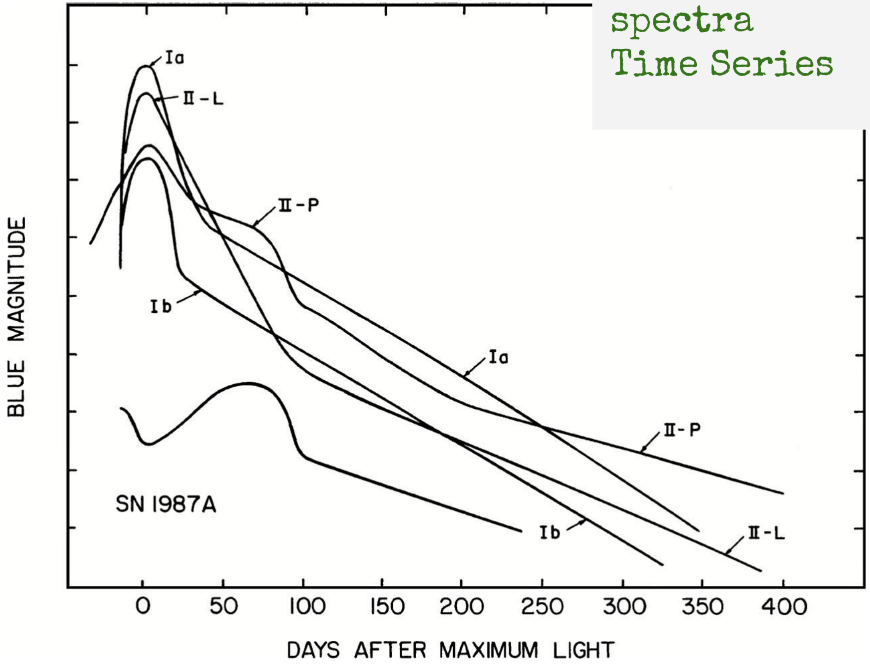
Spectroscopy

Expensive



Photometry

Cheaper
Multi-bands
Low resolution spectra
Time Series

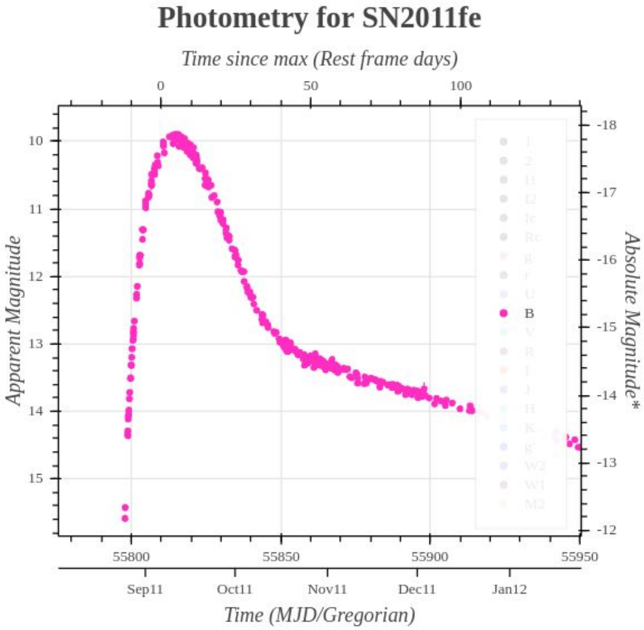


Filippenko 1997

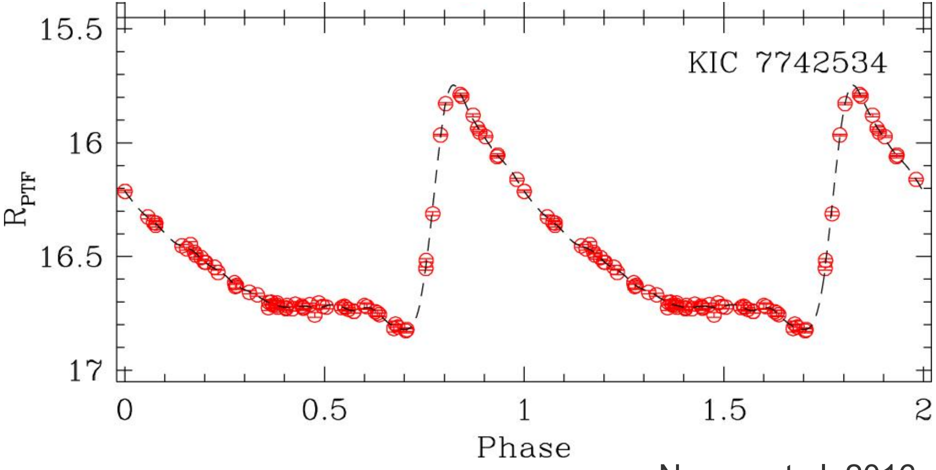
The Photometric LSST **Astronomical Time Series** Classification Challenge

⇩

Light curves of
Transients and Variable Stars



Plotted using the Open Supernova Catalog



Ngeow et al. 2016

Pre-PLAsTiCC: SNPhotCC

SUPERNOVA PHOTOMETRIC CLASSIFICATION CHALLENGE

RICHARD KESSLER,[1,2] ALEX CONLEY,[3] SAURABH JHA,[4] STEPHEN KUHLMANN[5]

Results from the Supernova Photometric Classification Challenge

RICHARD KESSLER,[1,2] BRUCE BASSETT,[3,4,5] PAVEL BELOV,[6] VASUDHA BHATNAGAR,[7] HEATHER CAMPBELL,[8]
ALEX CONLEY,[9] JOSHUA A. FRIEMAN,[1,2,10] ALEXANDRE GLAZOV,[6] SANTIAGO GONZÁLEZ-GAITÁN,[11]
RENÉE HLOZEK,[12] SAURABH JHA,[13] STEPHEN KUHLMANN,[14] MARTIN KUNZ,[15] HUBERT LAMPEITL,[8]
ASHISH MAHABAL,[16] JAMES NEWLING,[3] ROBERT C. NICHOL,[8] DAVID PARKINSON,[17]
NINAN SAJEETH PHILIP,[18] DOVI POZNANSKI,[19,20] JOSEPH W. RICHARDS,[20,21]
STEVEN A. RODNEY,[22] MASAO SAKO,[23] DONALD P. SCHNEIDER,[24]
MATHEW SMITH,[25] MAXIMILIAN STRITZINGER,[26,27,28]
AND MELVIN VARUGHESE[29]

- Held in 2010
- Classification on SN Ia and Core-collapse
- Within the astronomy community
- Sample size ~ several thousands
- The post-challenge data has been used for developing methods for photometric classification of supernovae

Why citizen science?

- Citizen science is vital for astronomy
- Industry drives rapid advances in machine learning (ML)
- LSST data rate demands ML for identifying time-domain events
- Citizen scientists now include thousands of ML experts
- Kaggle provides a platform for ML experts to tackle interesting supervised-learning questions

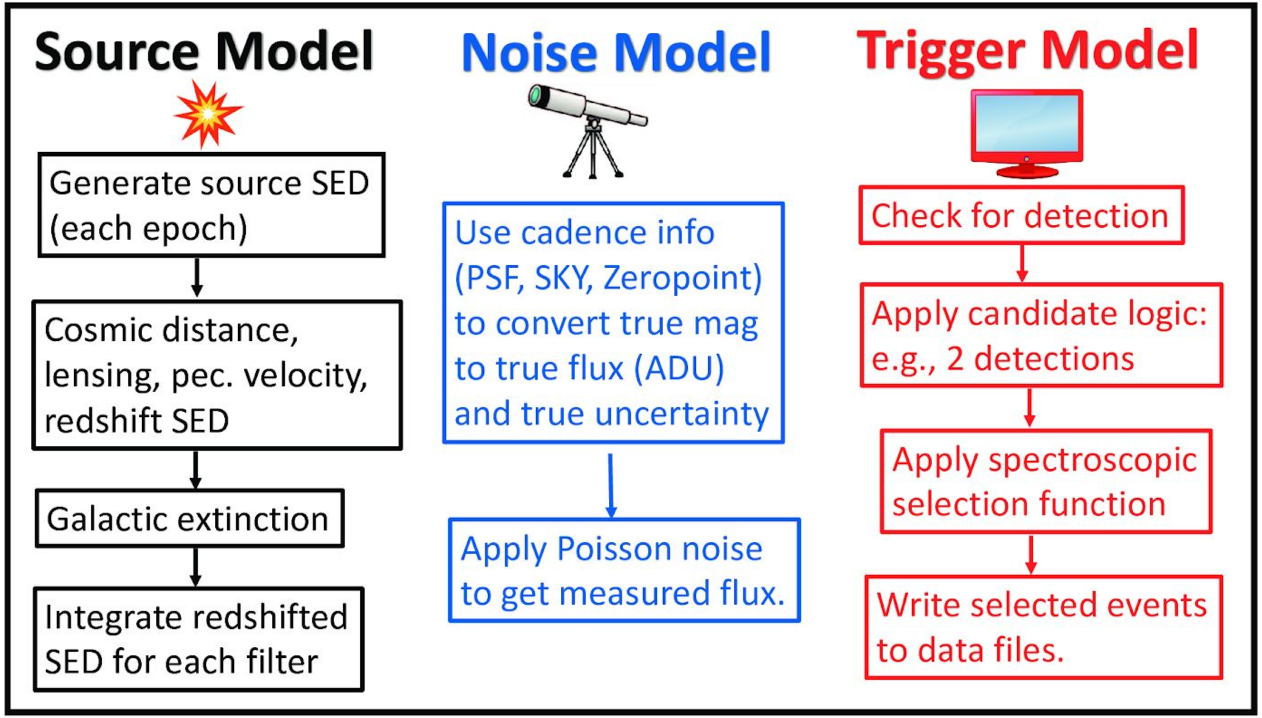
Credit: Kaggle

Slide credit: Gautham Narayan

The “Challenge”

- Types are unbalanced
- Small number in the training set
- The training set is not representative of the test data
- Season gaps
- Non-uniform cadence
- Unknown Class 99

Simulation



Kessler et al. 2019

Models

Summary of Models used in PLAsTiCC

model class num[a]: name	model description	contributor(s)[b]	Nevent Gen[c]	Nevent train[d]	Nevent test[e]	redshift range[f]
90: SNIa	WD detonation, Type Ia SN	RK	16,353,270	2,313	1,659,831	< 1.6
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	< 0.9
52: SNIax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	< 1.3
42: SNII	Core Collapse, Type II SN	SG,LG:RK,JRP:VAV	59,198,660	1,193	1,000,150	< 2.0
62: SNIbc	Core Collapse, Type Ibc SN	VAV:RK,JRP	22,599,840	484	175,094	< 1.3
95: SLSN-I	Super-Lum. SN (magnetar)	VAV	90,640	175	35,782	< 3.4
15: TDE	Tidal Disruption Event	VAV	58,550	495	13,555	< 2.6
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	< 0.3
88: AGN	Active Galactic Nuclei	SD	175,500	370	101,424	< 3.4
92: RRL	RR lyrae	SD	200,200	239	197,155	0
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	0
16: EB	Eclipsing Binary stars	AP	220,200	924	96,572	0
53: Mira	Pulsating variable stars	RH	1,490	30	1,453	0
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2,820	151	1,303	0
991: μLens-Binary	μ-lens from binary lens	RD,AA	1,010	0	533	0
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1,702	< 0.4
993: CaRT	Calcium Rich Transient	VAV	2,834,500	0	9,680	< 0.9
994: PISN	Pair Instability SN	VAV	5,650	0	1,172	< 1.9
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	0
TOTAL	Sum of all models		117,128,700	7,846	3,492,888	—

[a]num>990 were all in unknown class 99 during the competition. An extra digit is added here to distinguish each model.
[b]Co-author initials. Colon separates independent methods.
[c]Number of generated events, corresponding to the true population without observational selection bias.
[d]Labeled subset from spectroscopic classification. 0 → predicted from theory, not convincingly observed, or very few observations.
[e]Unlabeled sample. `PLAsTiCC` goal is to label this sample.
[f]Redshift> 0 for extragalactic models; Redshift= 0 for Galactic models.

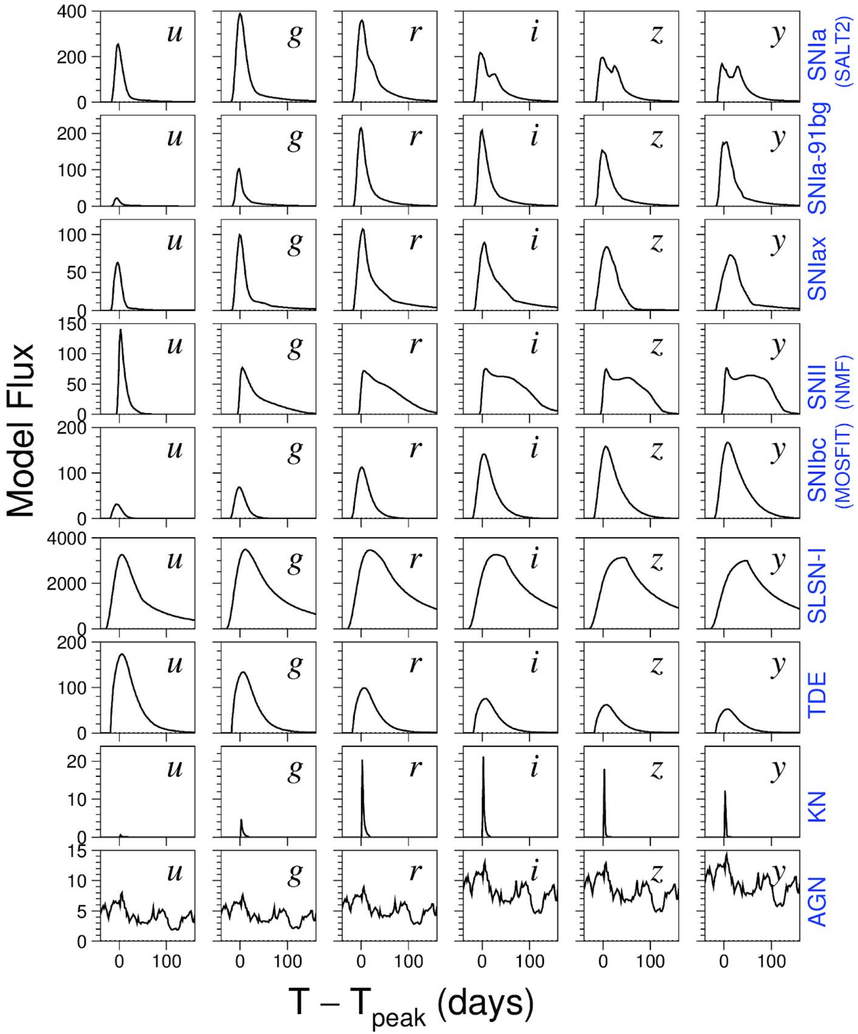
Model Contributors:
AA: Arturo Avelino (Harvard U.)
EB: Etienne Bachelet (LCO)
DC: David Chernoff (Cornell U.)
MD: Mi Dai (Rutgers U.)
SD: Scott Daniel (U.Washington)
RD: Rosanne Di Stefano (Harvard U.)
LG: Lluís Galbany (U.Pitt)
SG: Santiago González-Gaitán (U.Lisbon)
RH: Renée Hlozek (U.Toronto)
SJ: Saurabh Jha (Rutgers U.)
DK: Dan Kasen (U.C. Berkeley)
RK: Rick Kessler (U.Chicago)
GN: Gautham Narayan (STScI)
JRP: Justin Pierel (U. South Carolina)
AP: Andrej Prsa (Villanova U.)
VAV: Ashley Villar (Harvard U.)

Unblinded Data Files: http://doi.org/10.5281/zenodo.2539456

Simulation Source code: http://snana.uchicago.edu

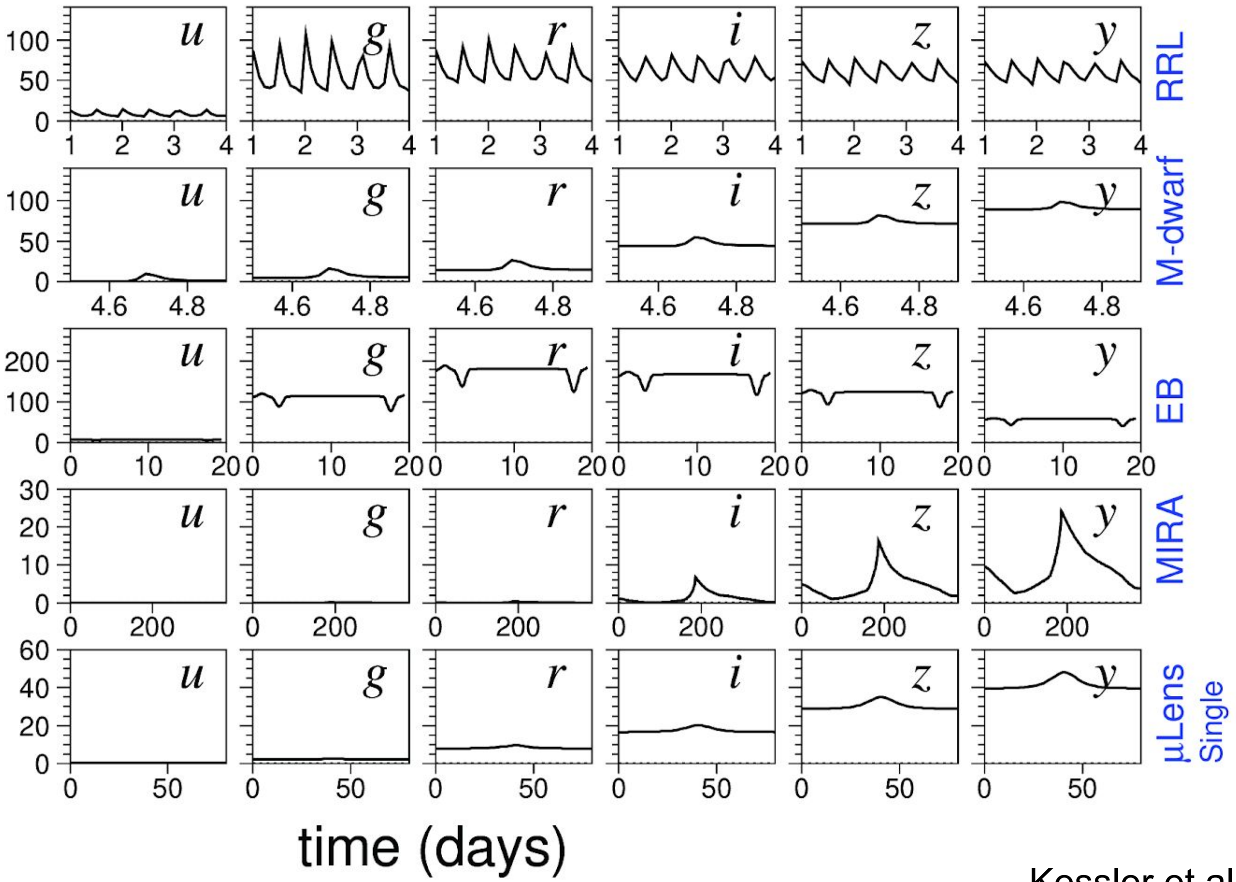
Kessler et al. 2019
Slide credit: Rick Kessler

Models



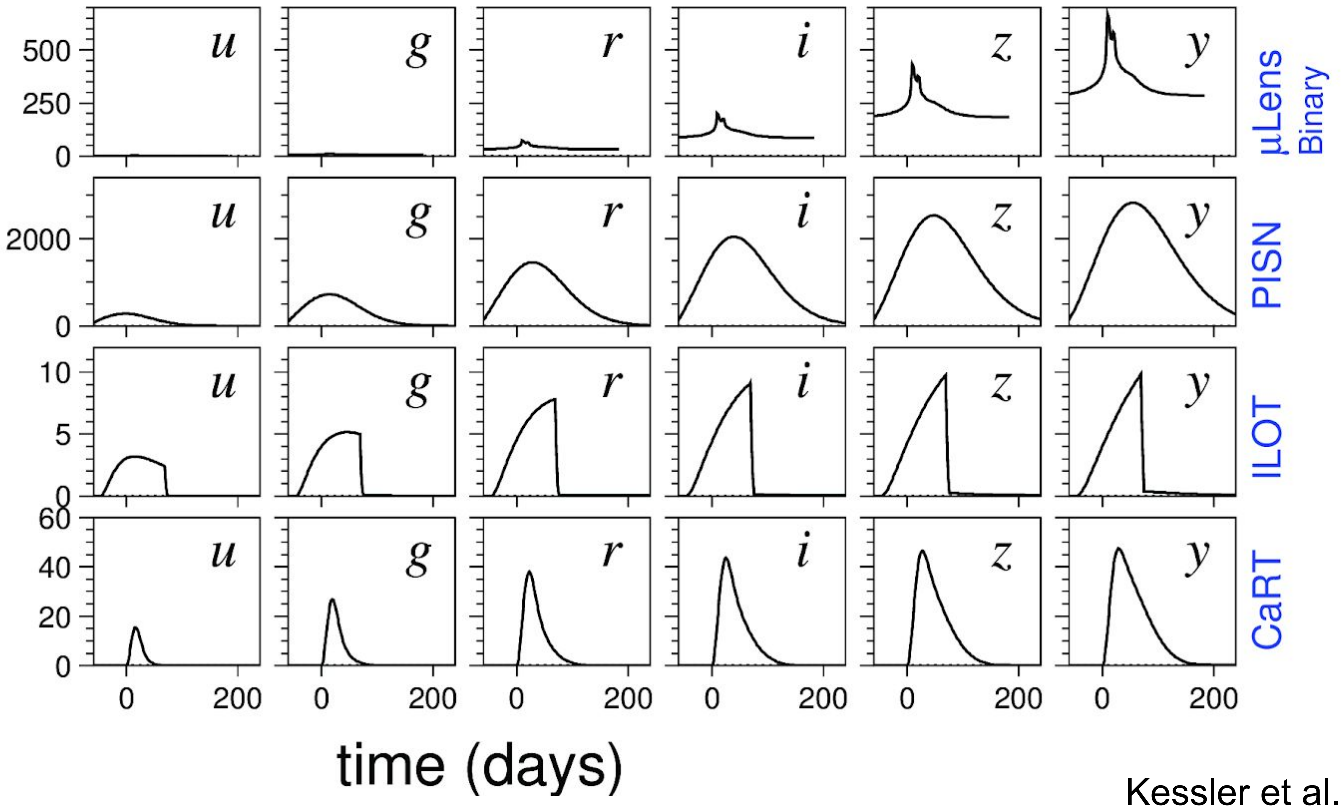
Kessler et al. 2019

Models



Kessler et al. 2019

Models

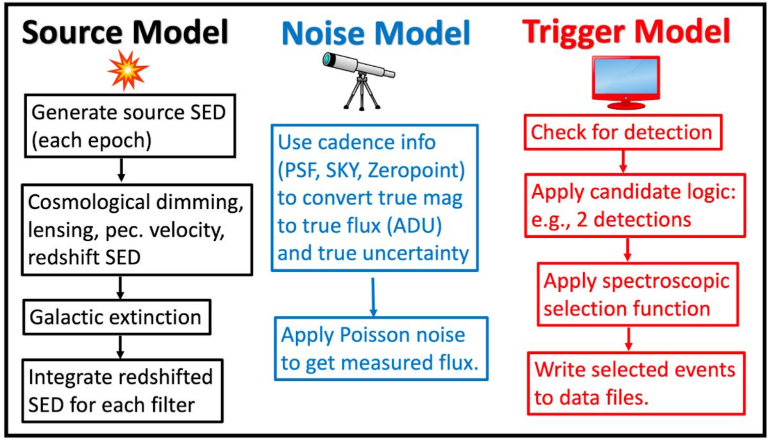


Kessler et al. 2019

Validation Efforts

Because we have SIMULATED data, there are several areas where we may introduce biases or non-physical correlations:

- Every box is a potential source for errors
- The source code was used to generate this data set, but it has never been used for galactic transients
- 3 million new SEDs added as inputs to the simulations!
- New codes and SEDs are an excellent source of bugs !



Kessler et al. 2019

Slide credit: Kara Ponder

Validation Efforts

Method to the Madness:
How to validate PLAsTiCC

Distribution tests

- Maximum Flux
- Minimum Flux
- Redshift
- Rates

Light curves

Visual inspection : limited to ~few hundred objects

- DDF
- WFD
- Model

Comparing model to DDF/WFD

Comparing real data to model

Classification codes to search for unphysical correlations

Meta data

- Ra
- Dec
- l
- b
- Milky Way Dust
- spec/photo-z
- distance modulus

Specialized tests per model.
Such as period-luminosity relations

- We created a private GitHub repository with a skeleton jupyter notebook and python3 environment.
- Each model had at least two validators each time the full set of simulations were regenerated
- A data scientist from Kaggle also reviewed our data

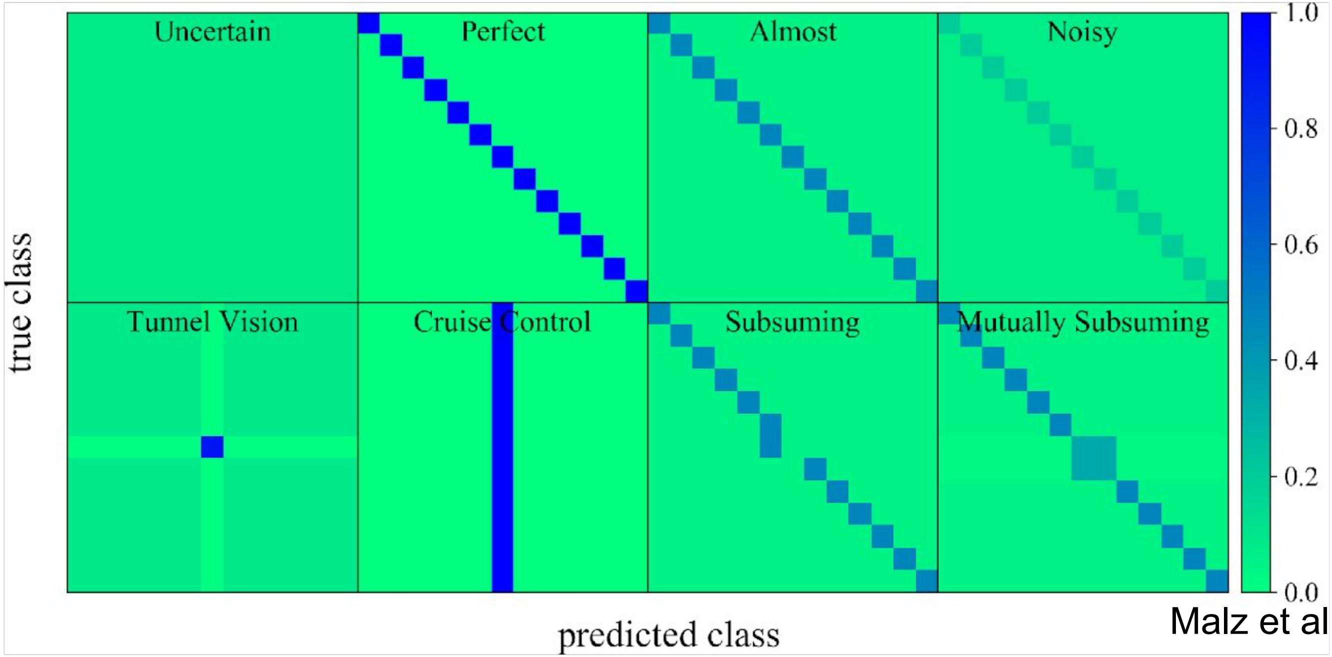
Slide credit: Kara Ponder

The Metric

- The metric needs to be probabilistic
- The metric depends on the science goal
- We need to select a metric that balances a variety of goals

The Metric

- The metric needs to be probabilistic
- The metric depends on the science goal
- We need to select a metric that balances a variety of goals

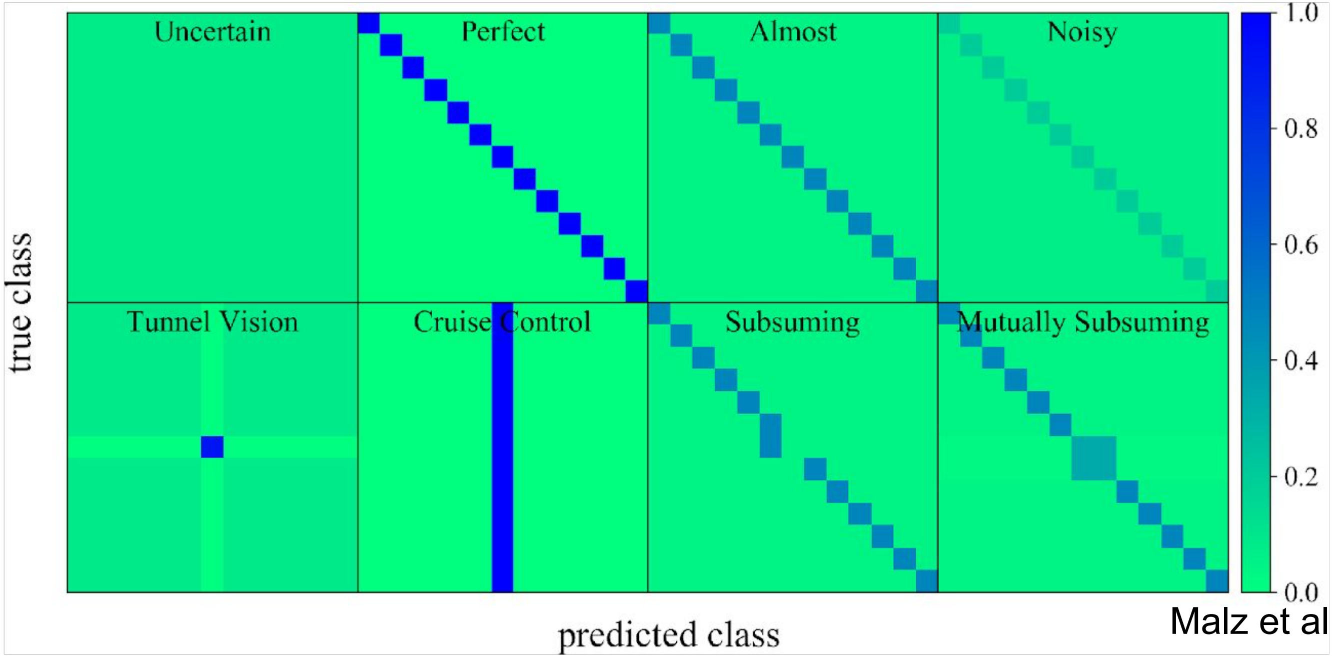


Malz et al. 2018

The Metric

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

- The metric needs to be probabilistic
- The metric depends on the science goal
- We need to select a metric that balances a variety of goals



Malz et al. 2018

Kaggle Competition

Featured Prediction Competition

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

$25,000
Prize Money

LSST Project · 1,094 teams · 7 months ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Overview

- Description
- Evaluation
- Prizes
- Timeline
- PLAsTiCC's Team

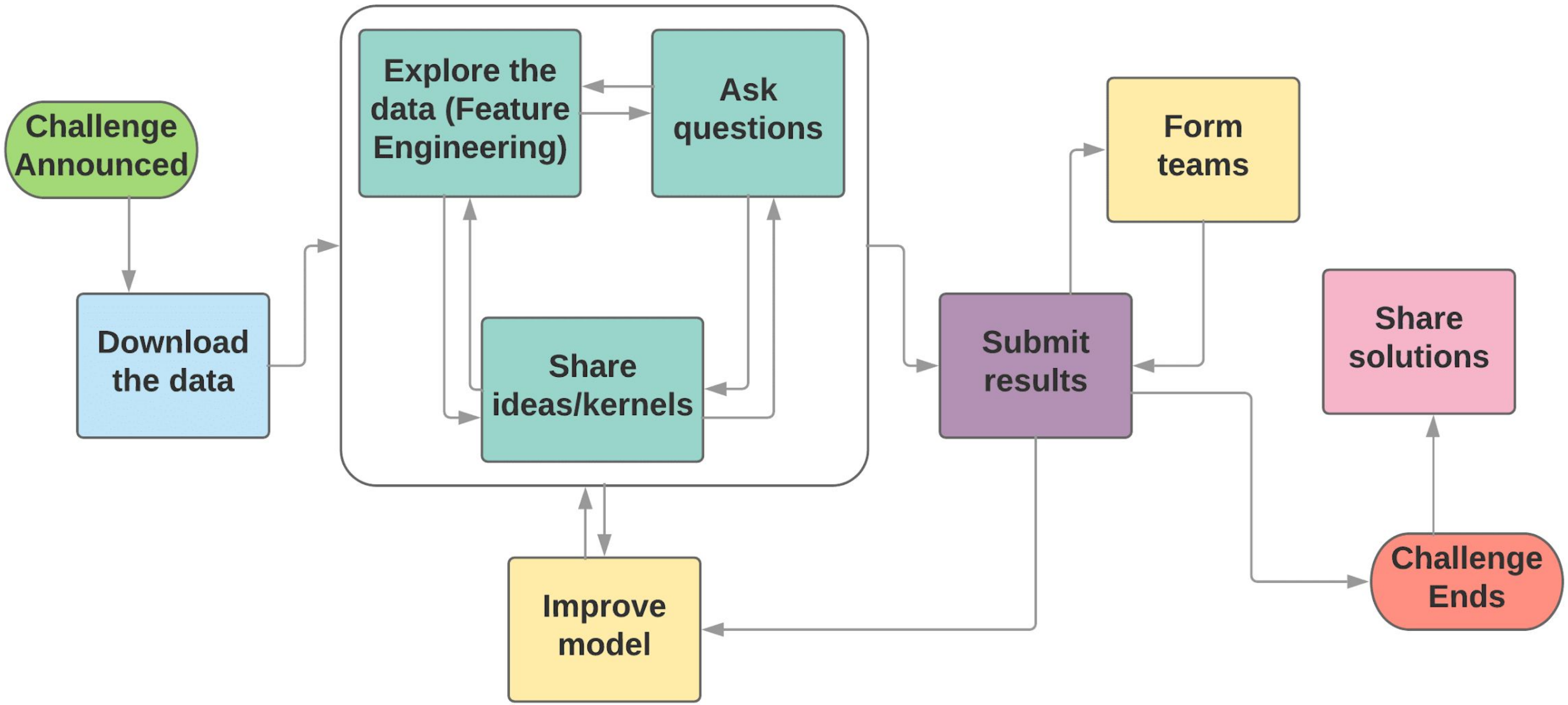
Help some of the world's leading astronomers grasp the deepest properties of the universe.

The human eye has been the arbiter for the classification of astronomical sources in the night sky for hundreds of years. But a new facility -- the Large Synoptic Survey Telescope (LSST) -- is about to revolutionize the field, discovering 10 to 100 times more astronomical sources that vary in the night sky than we've ever known. Some of these sources will be completely unprecedented!

The Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) asks Kagglers to help prepare to classify the data from this new survey. Competitors will classify astronomical sources that vary with time into different classes, scaling from a small training set to a very large test set of the type the LSST will discover.

More background information is available here.

A Kaggler’s journey to PLAsTiCC solutions



Leaderboard

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

LSST Project · 1,094 teams · 7 months ago

$25,000
Prize Money

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | Late Submission

Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 67% of the test data.

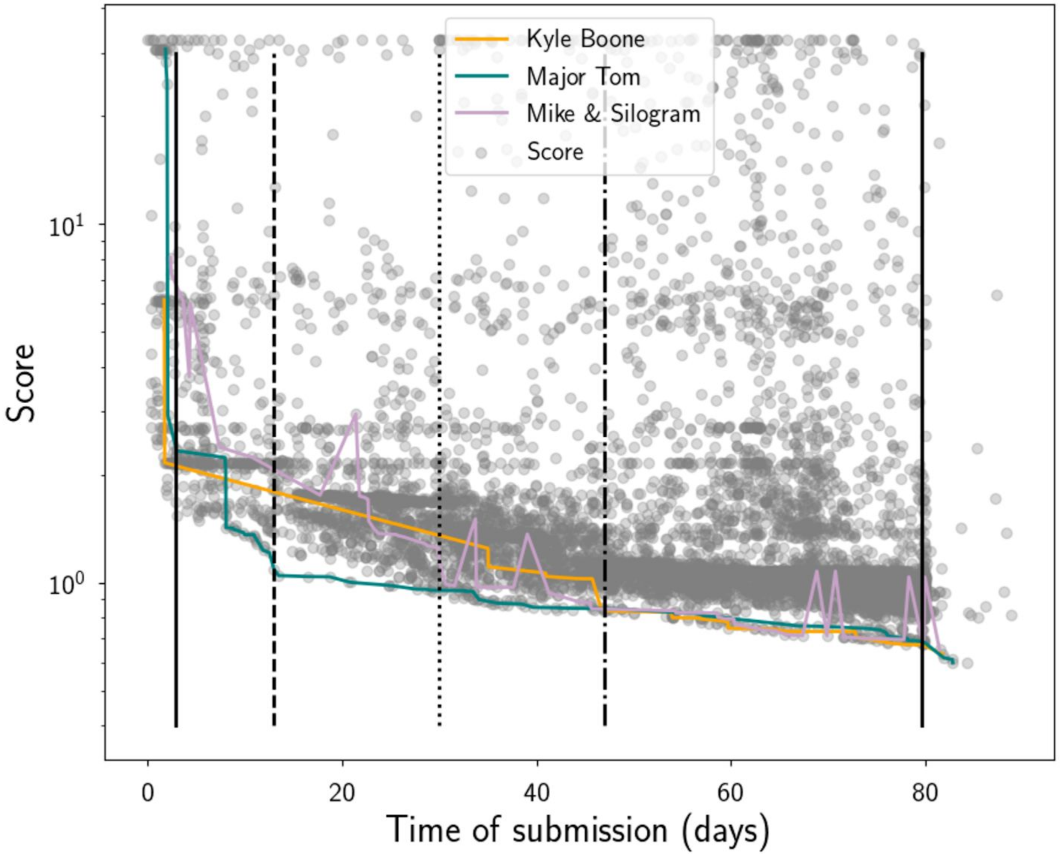
This competition has completed. This leaderboard reflects the final standings.

Refresh

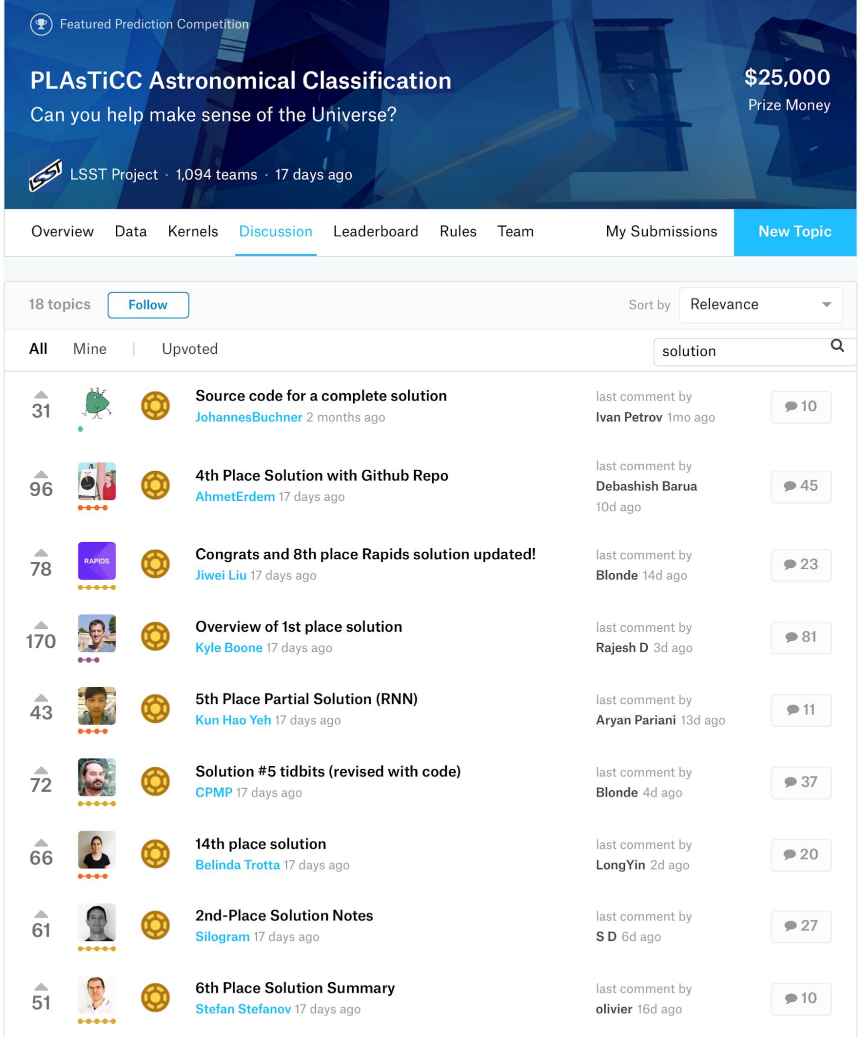
In the money | Gold | Silver | Bronze

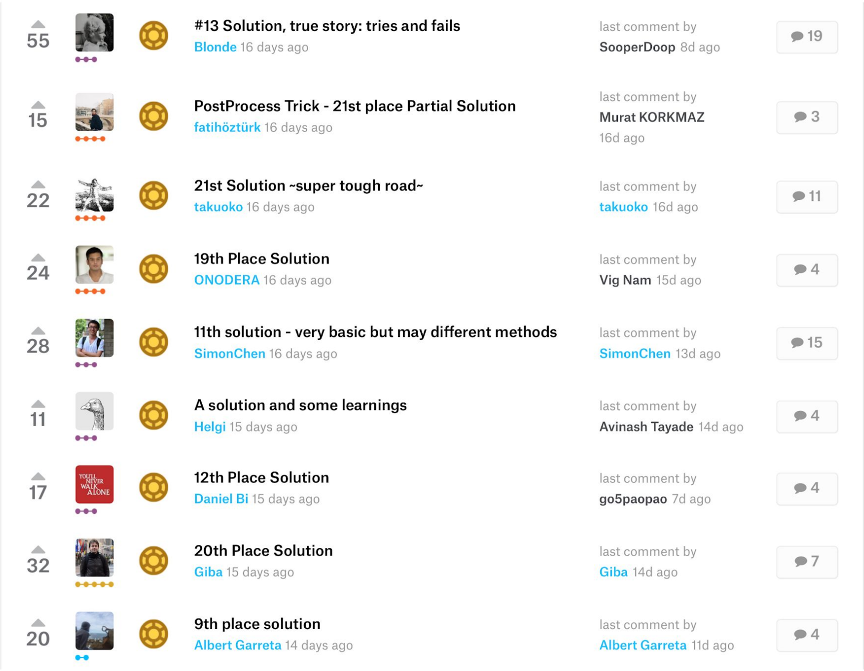
#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Kyle Boone			0.68503	104	7mo
2	▲2	Mike & Silogram			0.69933	176	7mo
3	▼1	Major Tom			0.70016	366	7mo
4	▼1	AhmetErdem			0.70423	233	7mo
5	—	SKZ Lost in Translation			0.75229	337	7mo
6	▲2	Stefan Stefanov			0.80173	28	7mo
7	▲3	hklee			0.80836	63	7mo
8	▼1	rapids.ai			0.80905	133	7mo
9	▼3	Three Musketeers			0.81312	313	7mo

Team scores over time

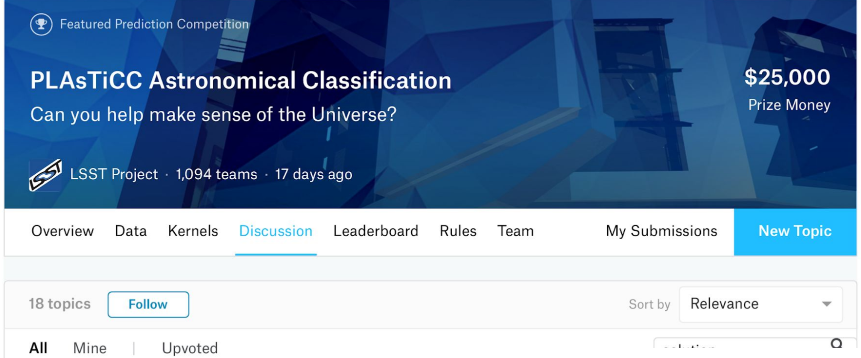


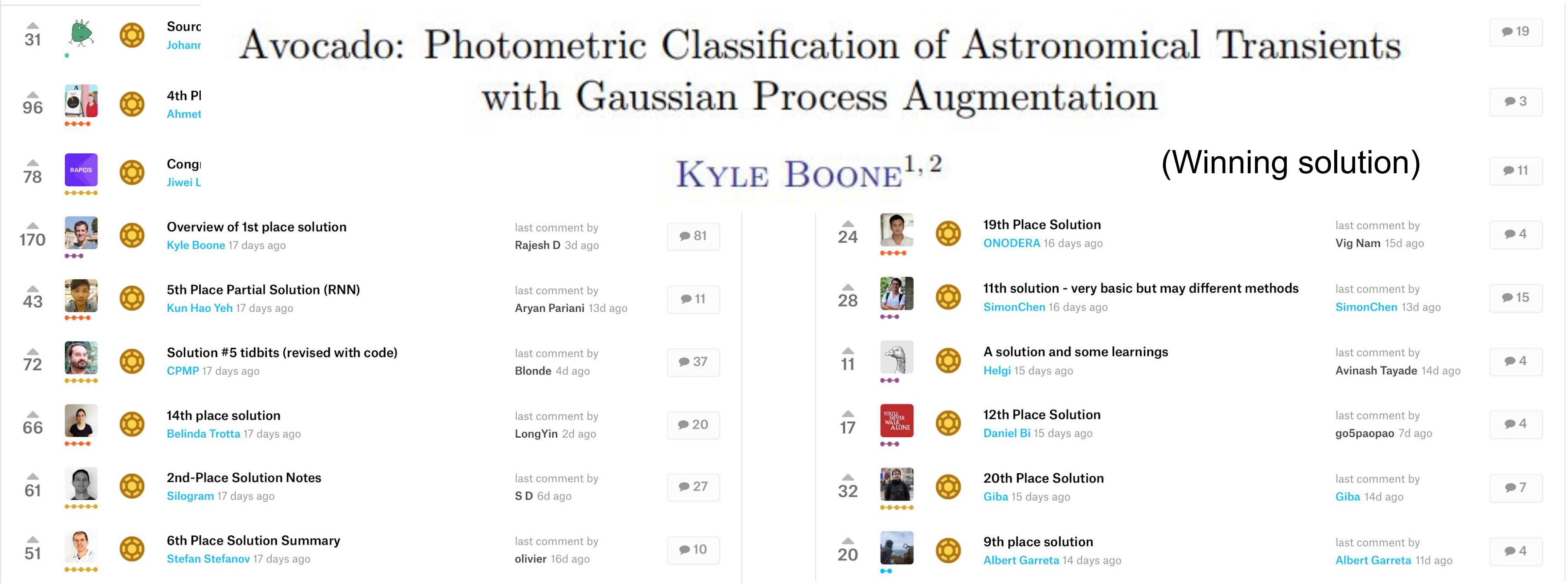
Solutions posted on Kaggle





Solutions posted on Kaggle





Solutions posted on Kaggle

Featured Prediction Competition

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

$25,000 Prize Money

LSST Project · 1,094 teams · 17 days ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | New Topic

18 topics Follow Sort by Relevance

All Mine | Upvoted

Votes	Topic	Author	Last comment by	Comments
31	Sourc	Johanr		19
96	4th Pl	Ahmet		3
78	Congi	Jiwei L		11
170	Overview of 1st place solution	Kyle Boone 17 days ago	Rajesh D 3d ago	81
43	5th Place Partial Solution (R	Kun Hao Yeh 17 days ago		
72	Solution #5 tidbits (revised v	CPMP 17 days ago		
66	14th place solution	Belinda Trotta 17 days ago	LongYin 2d ago	20
61	2nd-Place Solution Notes	Silogram 17 days ago	S D 6d ago	27
51	6th Place Solution Summary	Stefan Stefanov 17 days ago	olivier 16d ago	10
24	19th Place Solution	ONODERA 16 days ago	Vig Nam 15d ago	4
			imonChen 13d ago	15
			vinash Tayade 14d ago	4
17	12th Place Solution	Daniel Bi 15 days ago	go5paopao 7d ago	4
32	20th Place Solution	Giba 15 days ago	Giba 14d ago	7
20	9th place solution	Albert Garreta 14 days ago	Albert Garreta 11d ago	4

Avocado: Photometric Classification of Astronomical Transients with Gaussian Process Augmentation

Kyle Boone[1,2]

(Winning solution)

PLAsTiCC results paper by Hložek et al, in prep

Useful Features

- Light curve fitting -- Bazin, GP, template fitting (SALT2, SN templates)
- Flux ratio (color)
- Flux difference
- Host galaxy photo-z
- flux * distance ** 2

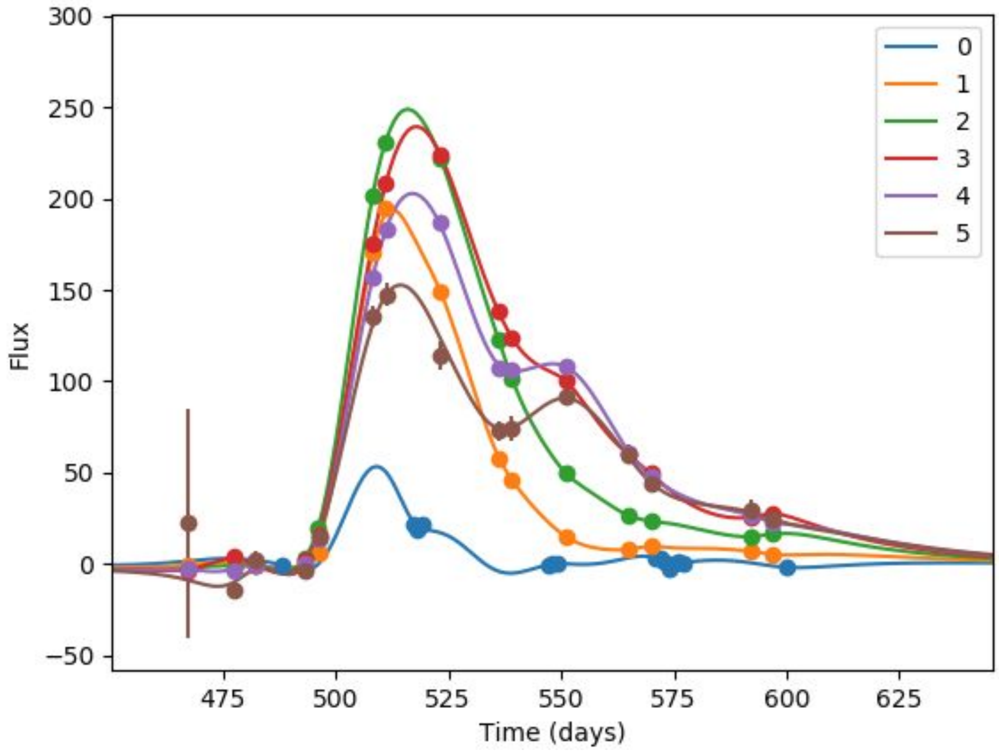


Figure credit: Kyle Boone

Popular Models among Kagglers

Gradient Boosting

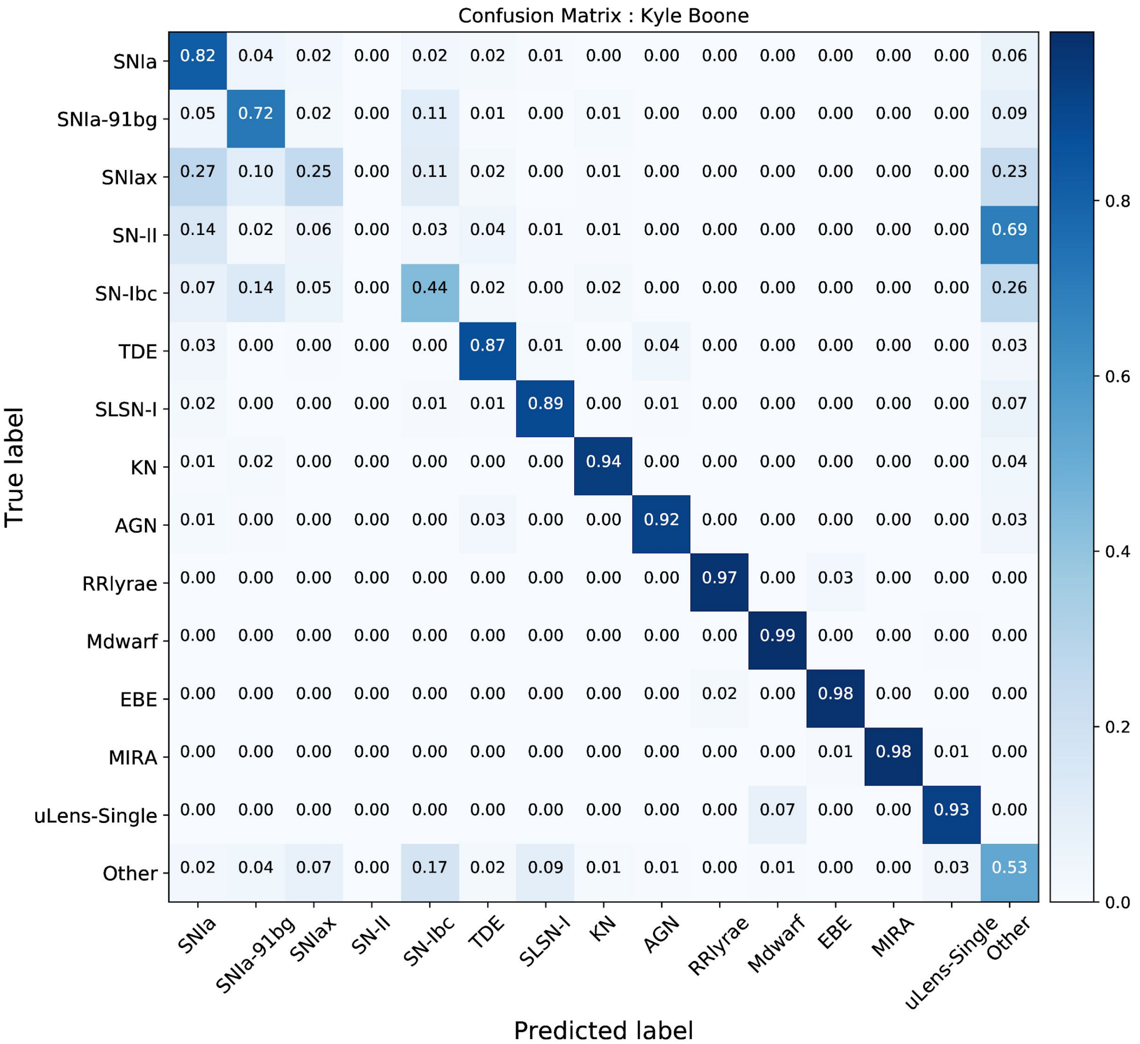
- LightGBM
- XGBoost
- CatBoost

Neural Net

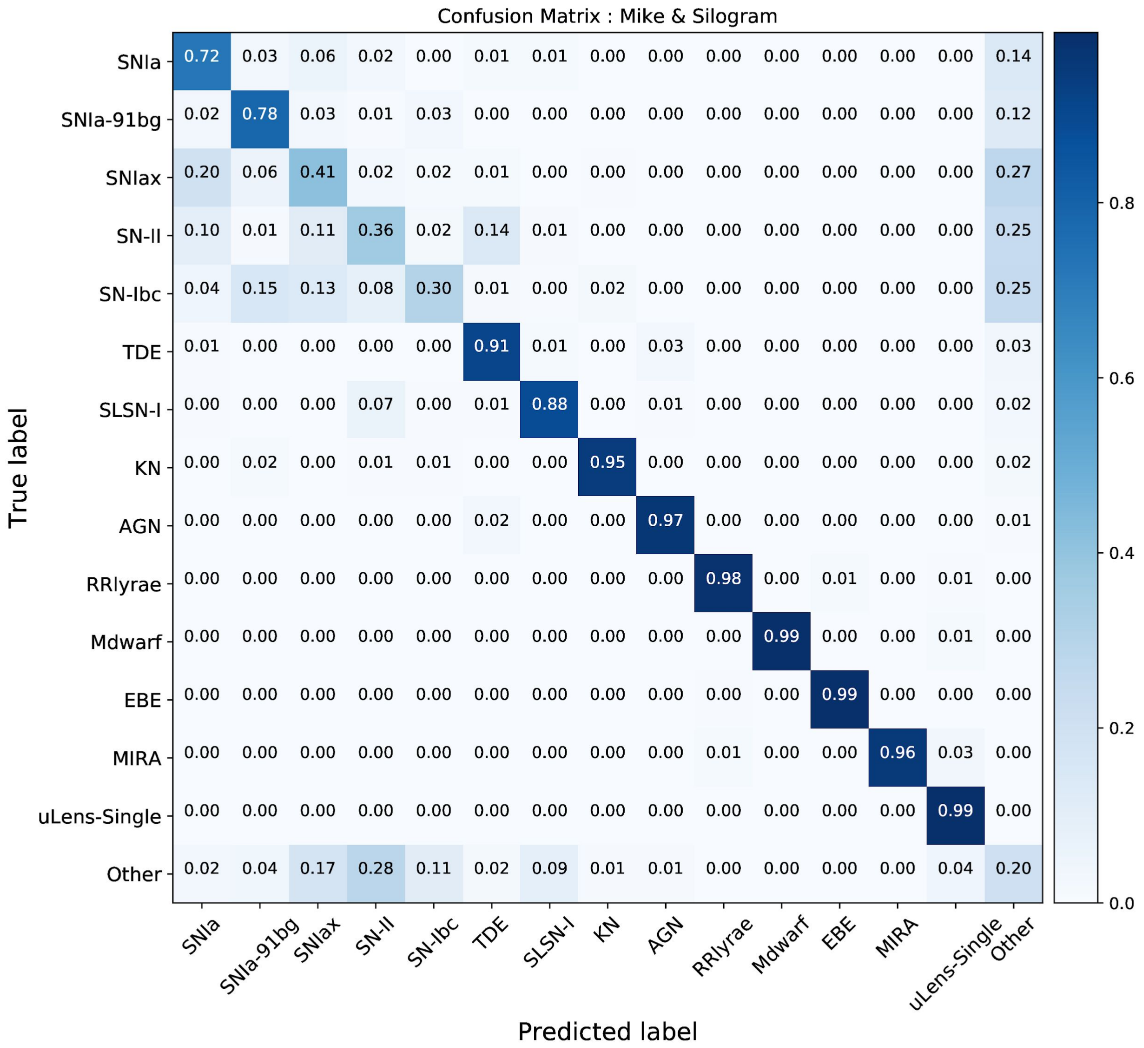
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Multi Layer Perceptron (MLP)
- Autoencoders

Binary Classification

Confusion Matrices



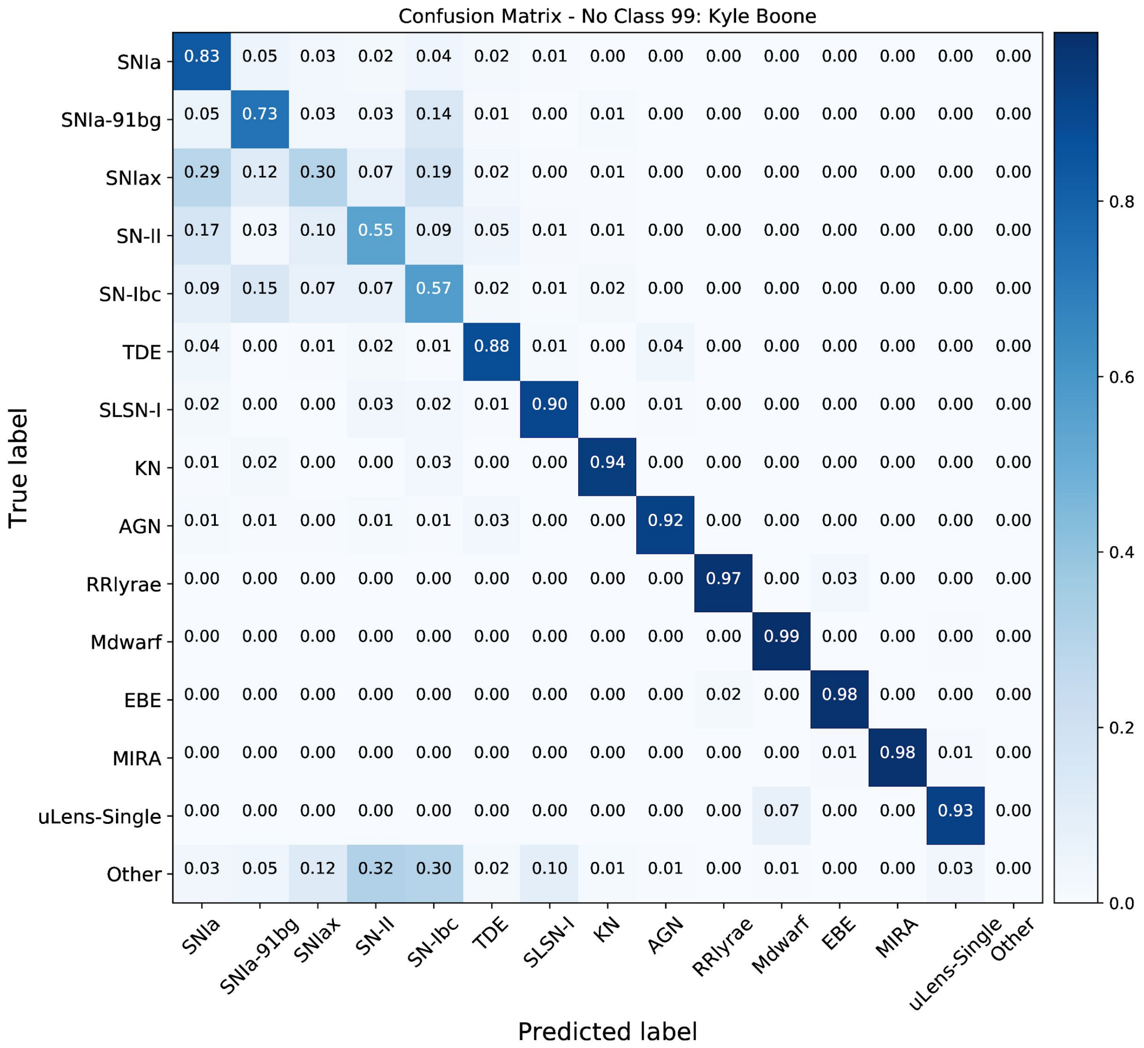
Confusion Matrices



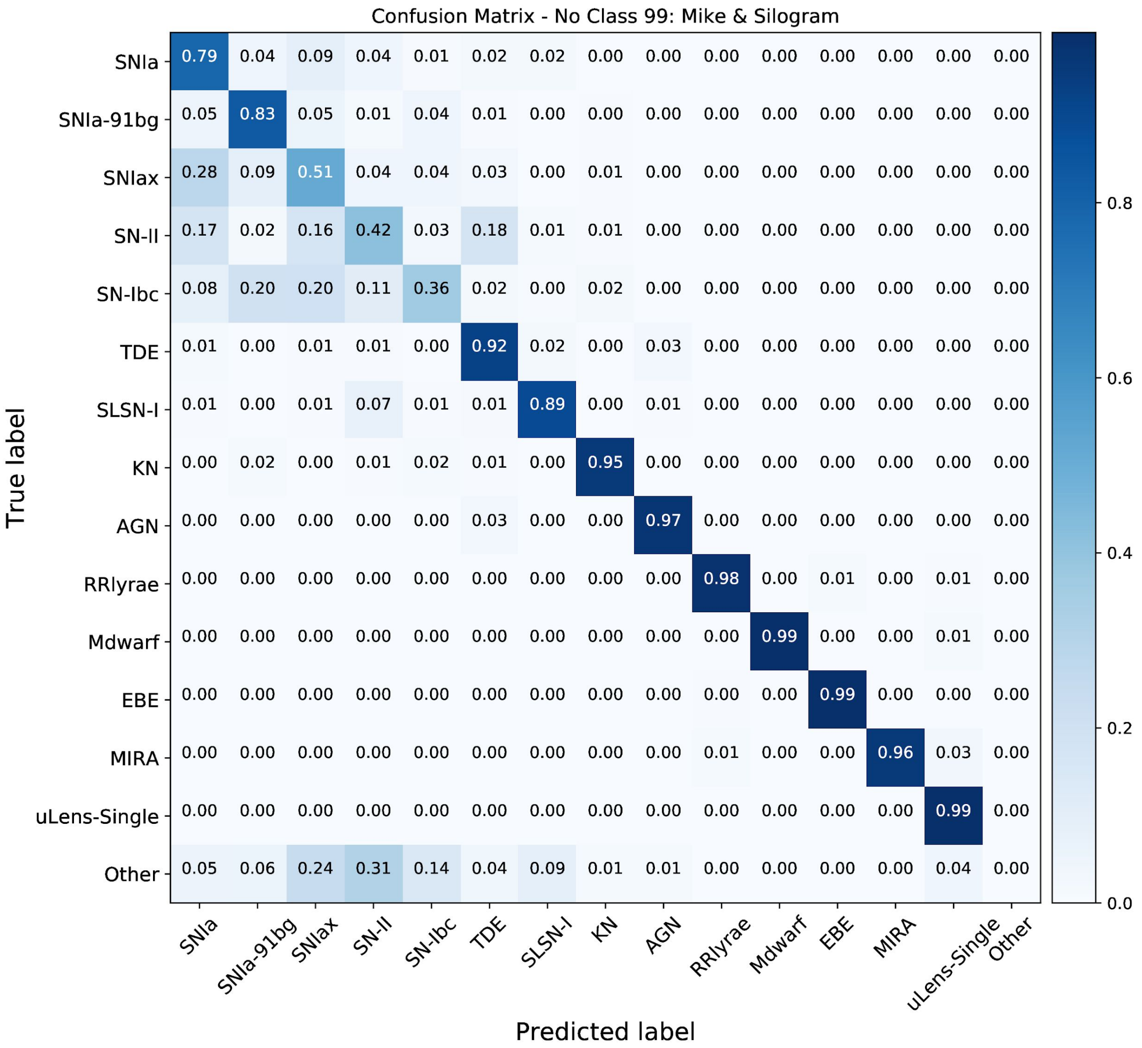
Class 99

- Designed to encourage anomaly detection methods
- Kagglers ended up probing the Leaderboard

Confusion Matrices (excluding class 99)



Confusion Matrices (excluding class 99)



Combining the top solutions

Featured Prediction Competition

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

LSST Project · 1,094 teams · 19 days ago

$25,000
Prize Money

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions New Topic

mamas
3rd place

Top teams' solutions are quite different (finally reached 0.5x !!!)

29

posted in PLAsTiCC Astronomical Classification 18 days ago

I haven't read all of them very carefully, but seems we improved the score by different ways! It's very impressive to me. maybe we can get to 0.5x if we get average of gold medal teams model! (especially, top 5 teams)

UPDATED:

We finally reached 0.5x using 1st, 2nd, 3rd, 4th, 5th subs!
Reallly Great :)

Submission and Description	Private Score	Public Score	Use for Final Score
last_sub.csv.gz 2 minutes ago by mamas 1st * 0.33 + 2nd * 0.33 + (3rd + 4th + 5th)/3 * 0.33	0.61146	0.59828	☐

The PLAsTiCC data

Unblinded Data Release for PLAsTiCC

R. Kessler,[1,2] G. Narayan,[3] A. Avelino,[4] T. Allam Jr.,[5] A. Bahmanyar,[6,7] E. Bachelet,[8] R. Biswas,[9] A. Boucaud,[10,11] P. J. Brown,[12,13] D. F. Chernoff,[14] A. J. Connolly,[15] M. Dai,[16] S. Daniel,[15] R. Di Stefano,[17] M. R. Drout,[6,18] L. Galbany,[19] S. González-Gaitán,[20] M. L. Graham,[15] J. Guillochon,[17] R. Hložek,[6,7] E. E. O. Ishida,[21] S. W. Jha,[16] D. O. Jones,[22] M. Lochner,[23,24] A. A. Mahabal,[25,26] A. I. Malz,[27,28] K. S. Mandel,[29,30] J. R. Martínez-Galarza,[17] J. D. McEwen,[5] D. Muthukrishna,[29] A. O'Grady,[6,7] H. Peiris,[9,31] C. M. Peters,[7] J. R. Pierel,[32] K. Ponder,[33] A. Prša,[34] S. Rodney,[32] C N. Setzer,[9] and V. A. Villar[17]

LSST Dark Energy Science Collaboration and the LSST Transients and Variable Stars Science Collaboration

https://plasticc.org/

https://doi.org/10.5281/zenodo.2535746

By the numbers

- More than 1 million new SEDs across several new models
- 15* classes in training set, one not represented in training
- ~ 3.5 million objects in test set w/ < 8000 objects for training
- ~ 450 million observations (LSST WFD + DDF) in 6 bands ~ 18.5 GB
- Even simplified, PLAsTiCC is the largest simulation of light curves the time-domain sky in the optical ever

Slide credit: Gautham Narayan

Thinking about PLAsTiCC 2.0

- Host-galaxy information
- Realistic photo-z
- Early classification
- Image based challenge

Summary

- 1094 teams have participated on Kaggle
- 18 models were simulated
- Data have already been used by many groups
- More work is needed to digest all the solutions

“KAGGLE IS ADDICTIVE !

ENTER AT YOUR OWN RISK !!!”

"PLAsTiCC IS ADDICTIVE !

ENTER AT YOUR OWN RISK !!!"