PLAsTiCC: Convincing other people to solve your problems

Kara Ponder
Berkeley Center for Cosmological Physics
Computational Data Science Fellow
kponder@berkeley.edu

AI Seminar @ SLAC 02/20/2020

Outline

- Why do we need someone else to solve our problems?
- PLAsTiCC: Photometric LSST Astronomical Time-Series Classification Challenge
- Other ways to convince people to solve your problems: COIN
  - Data driven mock galaxy catalogs
  - Recommendation System for Spectroscopic follow-up

Type Ia Supernova

- Thermonuclear explosion of a white dwarf star
- At peak brightness, 5 billion times brighter than the sun!
- Explode with approximately the same mass so we assume they are standard candles
- Need **spectroscopy** to distinguish them from other supernovae

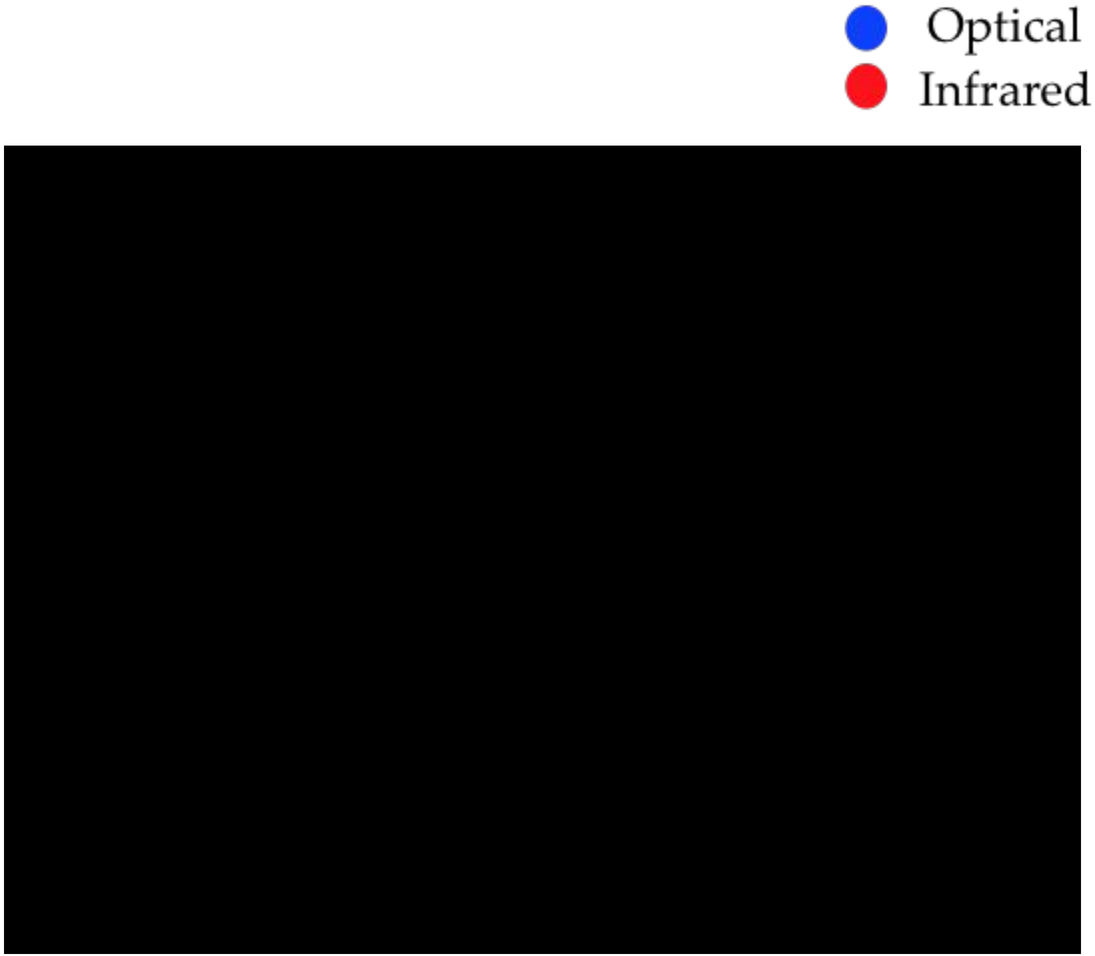


Image credit: Brian Hayden

Spectroscopy versus Photometry

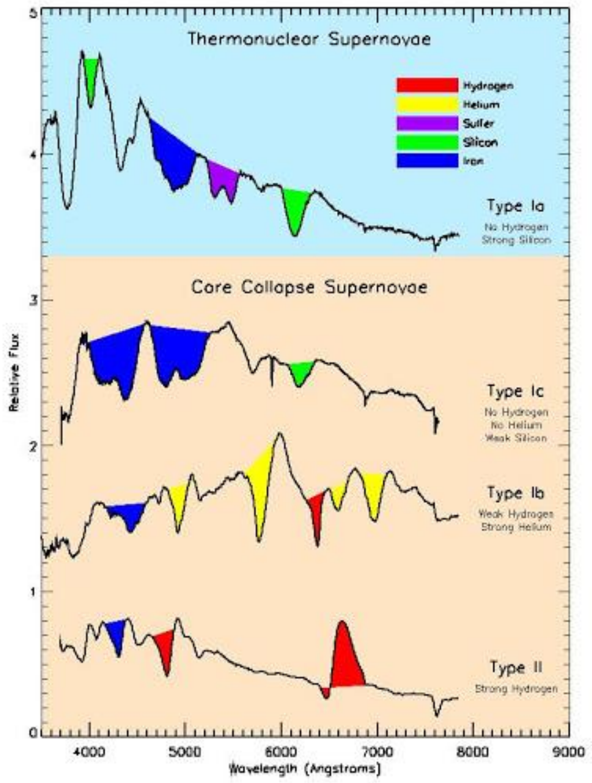


Image credit: Dan Kasen

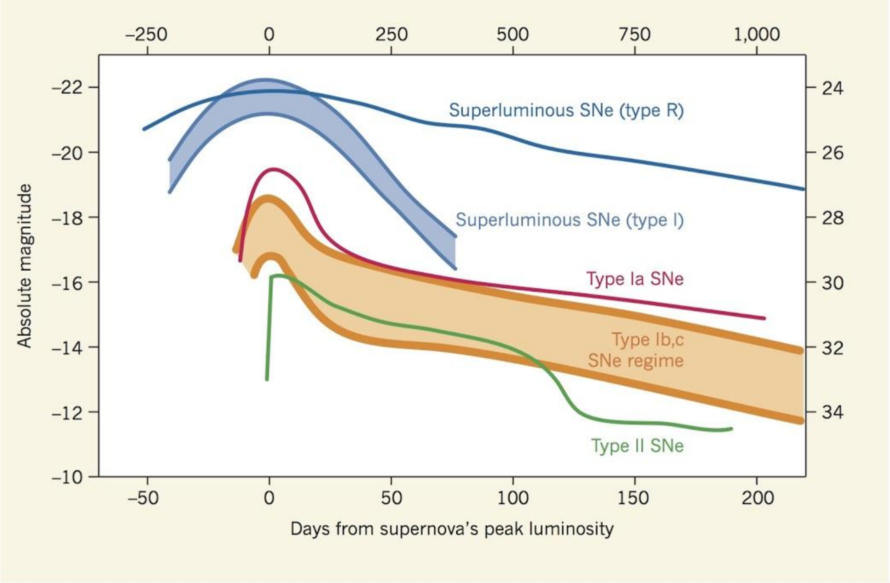
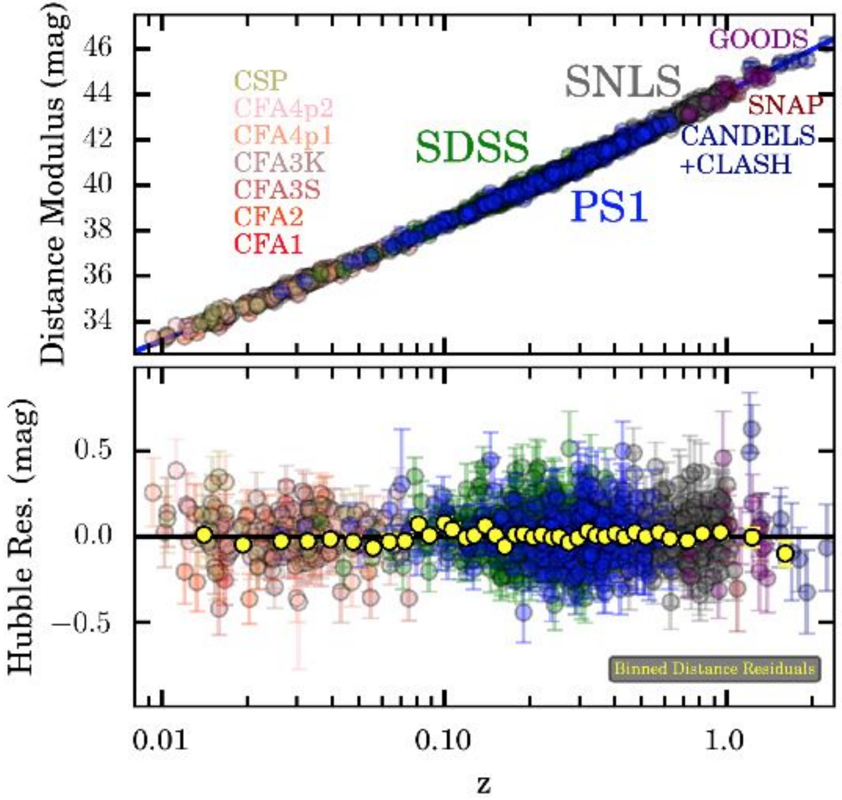


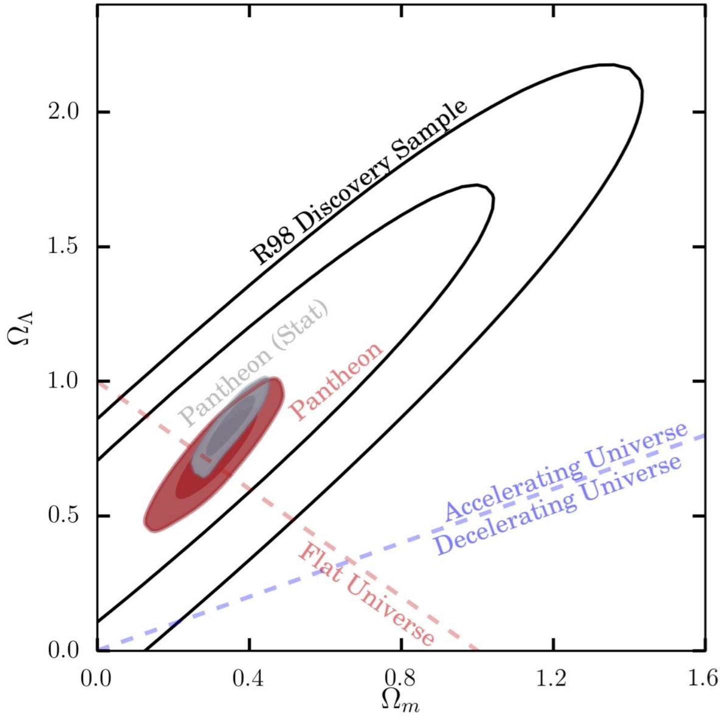
Image credit: Stephen Smartt, Nature (2012)

What's the appeal of photometry?

It's cheap!

Supernova Cosmology

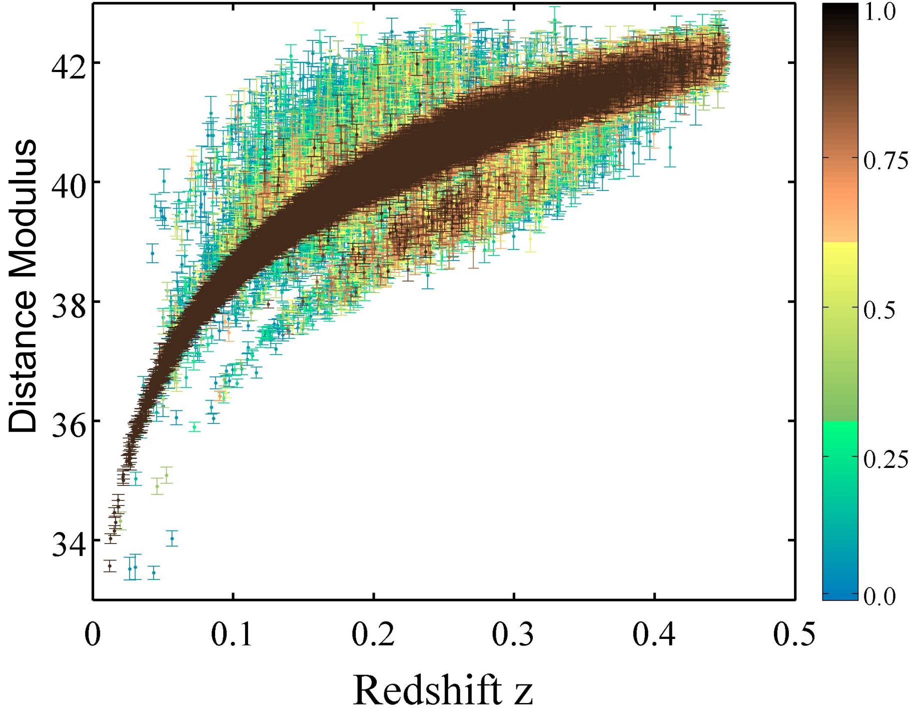




State-of-the-art: 1048 SNe Ia from Pantheon.
25 years to gather ~1000 SNe Ia

Scolnic et al (2018), ApJ, 859, 101S

Cosmology with photometric classification



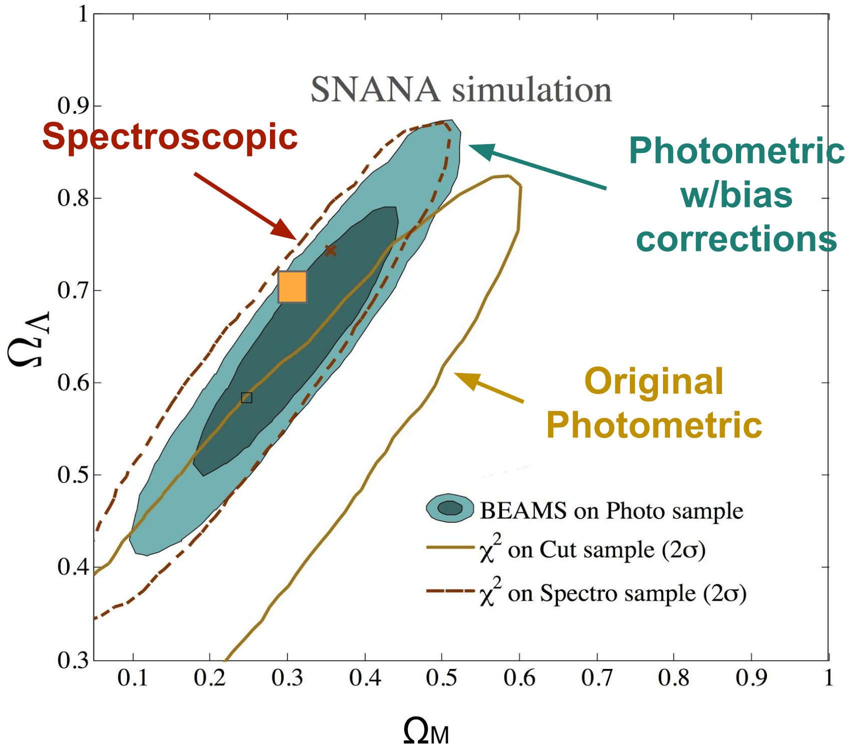


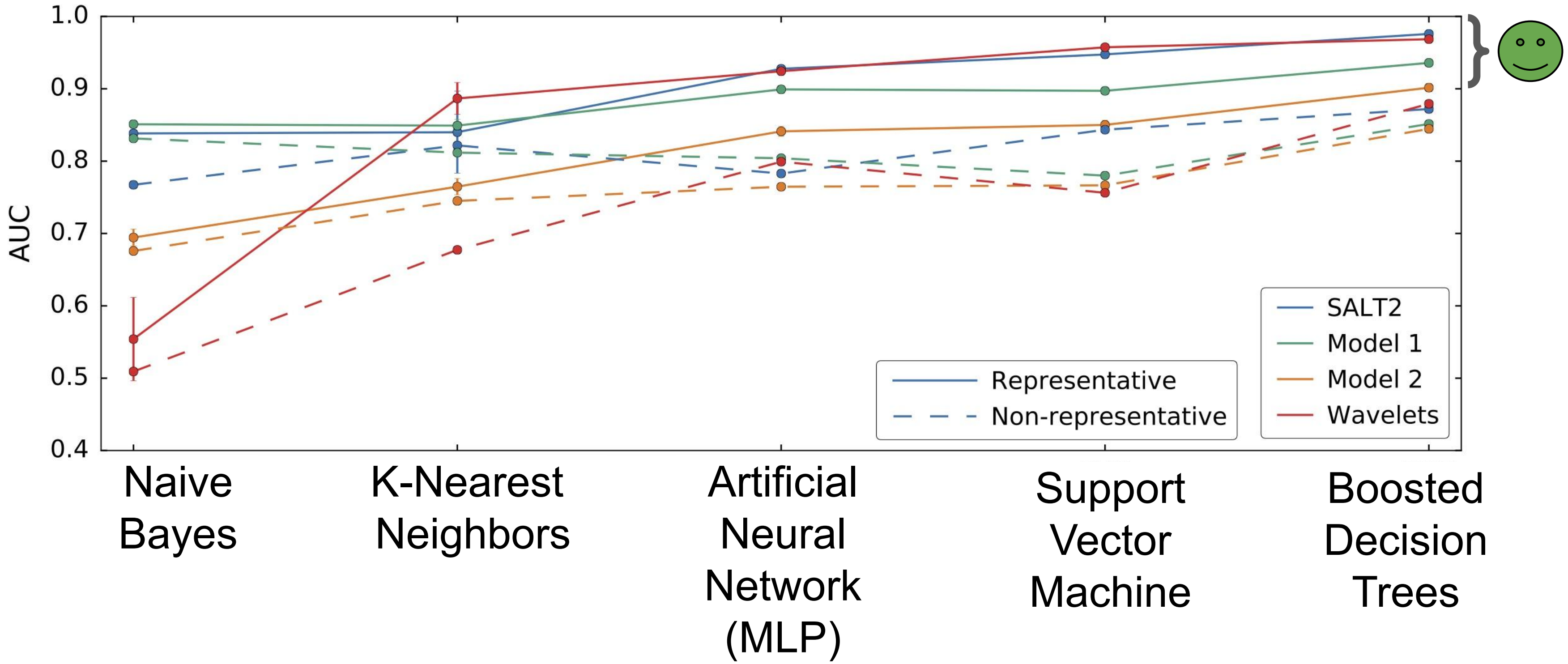
Image credits: Hlozek et al (2012) ApJ, 752, 2, 79

Algorithms for Photometric Classification

- PSNID (Sako et al 2011)
- PELICAN (Pasquet el at 2019)
- SuperNNova (Moller et al 2020)
- snmachine (Lochner et al 2016)
- Dai et al (2018)
- Ishida et al (2019) & Ishida, de Souza (2013)
- Villar et al (2019)
- Karpenka et al (2013)
- Newling et al (2011)
- Bloom et al (2012) - transients & variables
- Godines et al (2019) - microlensing
- Eyer et al (2004), Kim et al (2015) - variable stars

At least dozens more for other transients and variable!

Comparing different classification methods

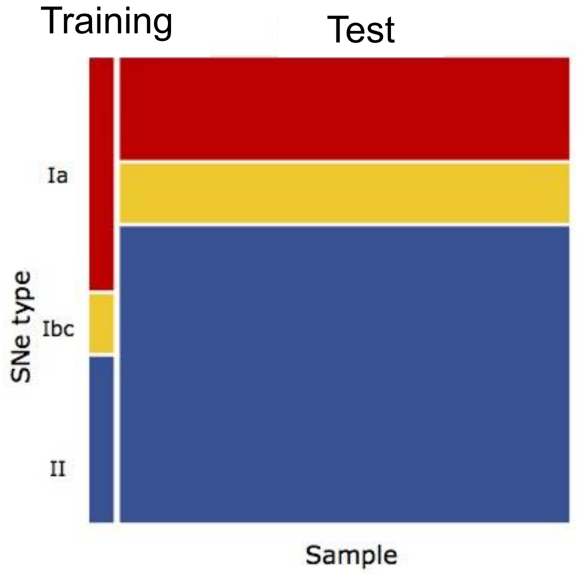


AUC: Area Under ROC Curve --
Closer to 1 is better

Figure credit: Lochner et al (ApJS, 2016), arXiv: 1611.07042

Major issues for photometric classification

Machine learning techniques are dependent on large, homogenous and representative training sets...



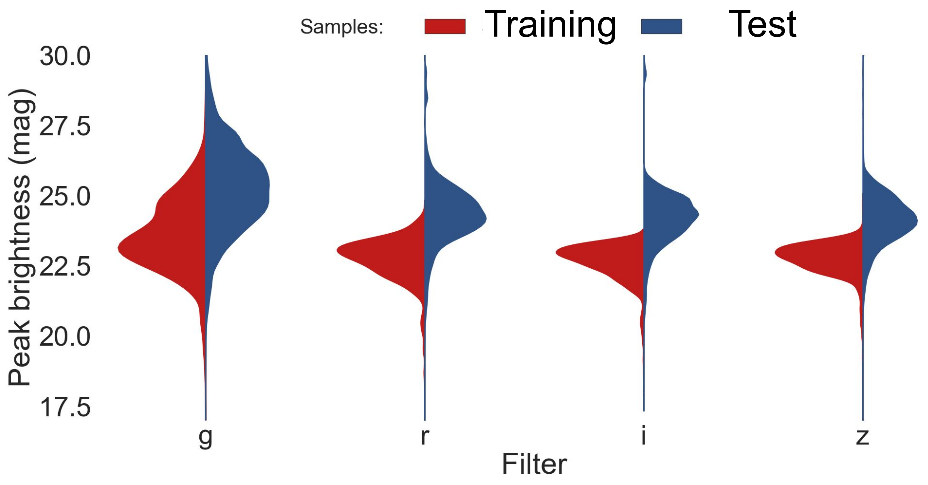


Figure credits: **Ishida** et al., 2019 MNRAS - arXiv:1804.03765

Algorithms for Photometric Classification

- PSNID (Sako et al 2011)
- PELICAN (Pasquet el at 2019)
- SuperNNova (Moller et al 2020)
- snmachine (Lochner et al 2016)
- Dai et al (2018)
- Ishida et al (2019) & Ishida, de Souza (2013)
- Villar et al (2019)
- Karpenka et al (2013)
- Newling et al (2011)
- Bloom et al (2012) - transients & variables
- Godines et al (2019) - microlensing
- Eyer et al (2004), Kim et al (2015) - variable stars

At least dozens more for other transients and variable!

- **All are limited by representativeness.**
- **No single algorithm can classify all transients and variables.**
- **All lack early classification abilities.**

In the future, will we need photometric classification?

Vera Rubin Observatory
Legacy Survey of Space and Time (LSST)

- Previously: Large Synoptic Survey Telescope
- 8.4 meter mirror
- 10 year survey beginning 2023
- 6 **photometric** filters: UV to NIR
- 3.2 Gpixel camera the size of a small car being built here at SLAC!
- Two observing patterns:
    - Wide Fast Deep (WFD)
    - Deep Drilling Field (DDF)

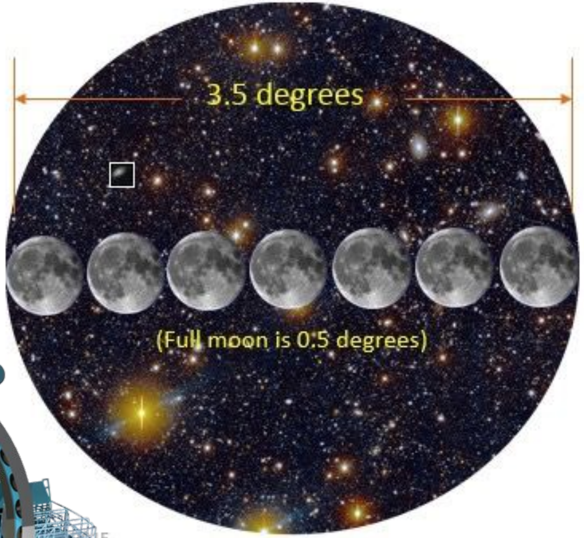


Image Credit: LSST

Image credit: LSSTC

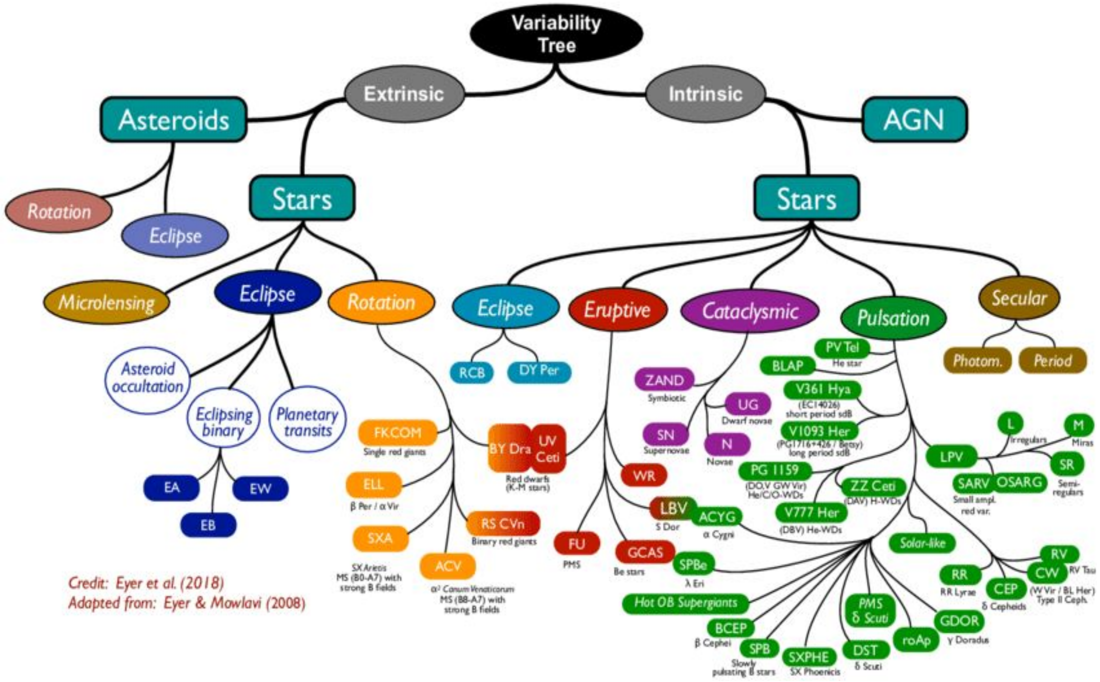
Era of Big Data

- With up to 10 million transient alerts per night, the community will be drowning in possible transients!
- LSST will give 10,000 **spectroscopically** confirmed Type Ia SNe over 10 years!
- But what about the other possible 100,000s supernovae observed?

How can we bring the groups together?

We’ll have a stream of up to 10 million transients and variables per night

With such large, distinct communities, we decided to use a data challenge to unite them.



Photometric LSST Astronomical Time-Series Classification Challenge: PLAsTiCC

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

Other photometric classification challenges: SNPhotCC

Asked 2 questions:

How well can you classify a Type Ia SN?

Can we classify early? -- No one participated in this part

SUPERNOVA PHOTOMETRIC CLASSIFICATION CHALLENGE

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

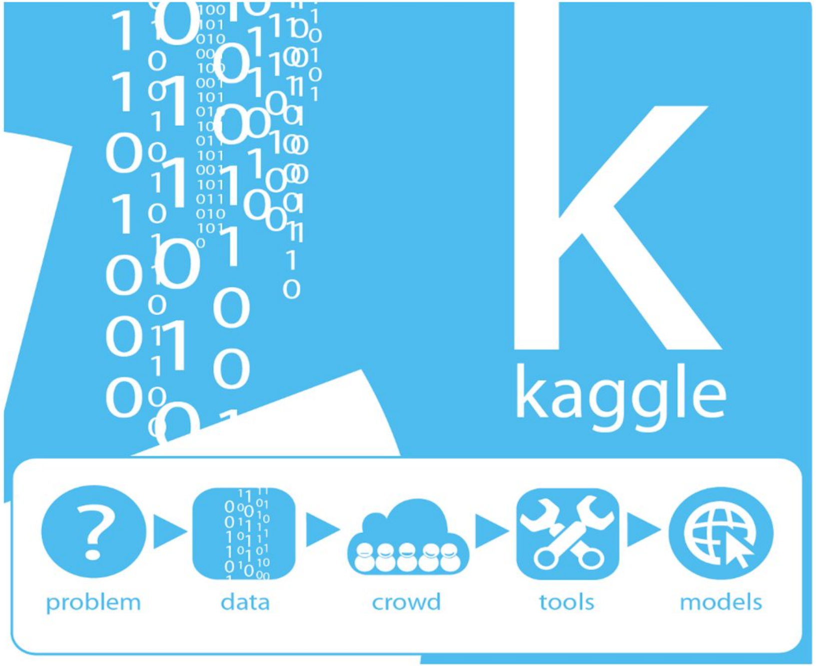
Results from the Supernova Photometric Classification Challenge

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 to prepare for the Dark Energy Survey (DES)
→ Only had Supernova Ia, Ibc, II templates
→ Thousands of objects
→ Only available to Astronomy community

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

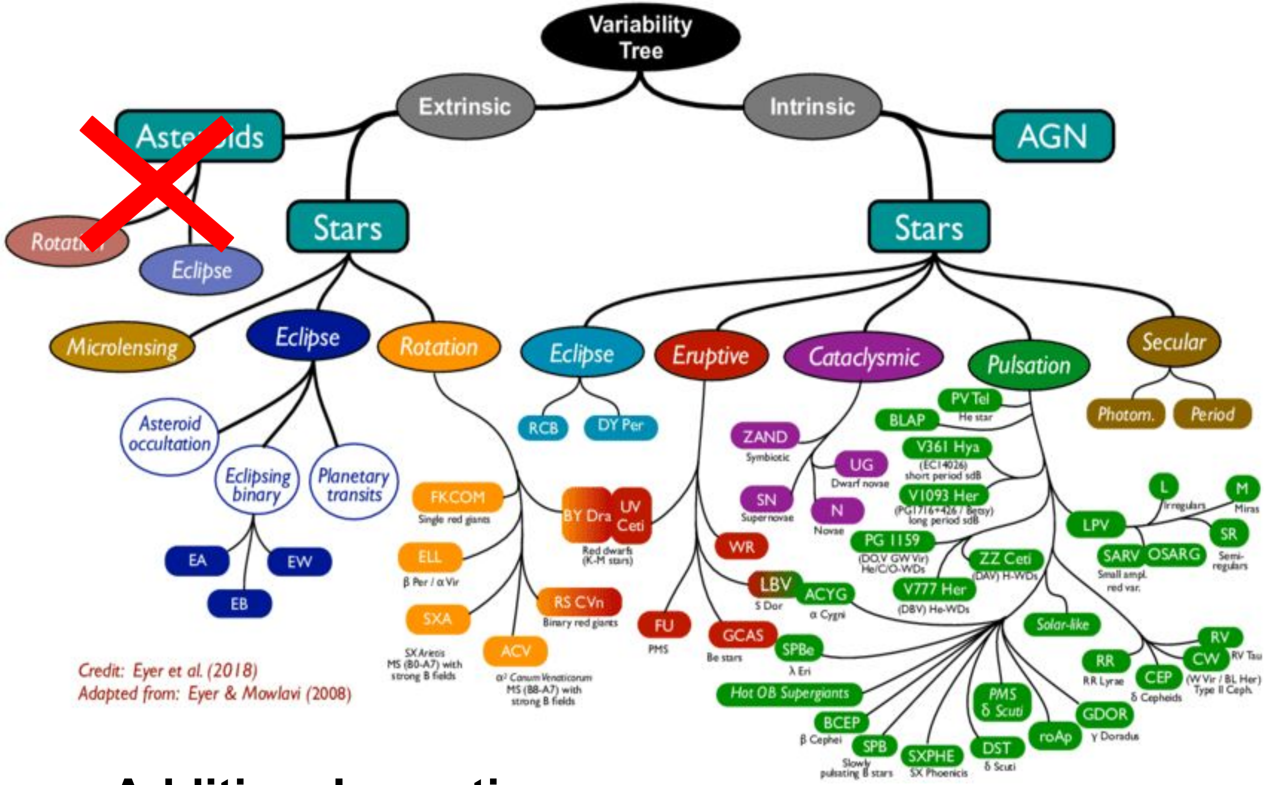


Credit: Kaggle

Slide credit: Gautham Narayan

The Question

What question do you want participants to answer?



How well can you classify **ALL** transients and variables?

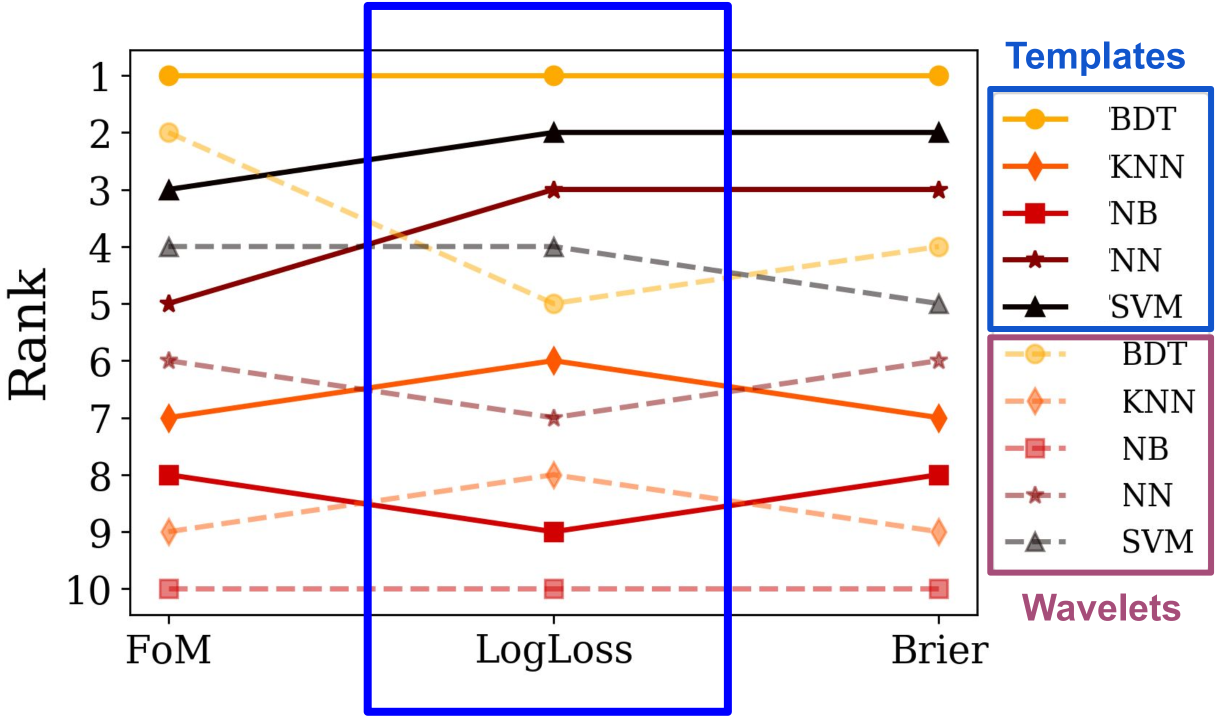
Additional question

Anomaly detection: If we had objects in the test set that were not in the training set, how well are they classified?

Proclam: https://github.com/aimalz/proclam

The Metric

- Needs to cover all different transient and variable classes
- Probabilistic
- Must interface with Kaggle -> need single number
  - Unable to have multiple challenges



Malz et al (2019, including **K.Ponder**) AJ, 158, 5, 171

Collecting models to simulate data

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	—

Galactic

SNe

[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.)

19 Models in total
18 with data
14 in training set

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

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

Kessler et al. 2019, PASP
(including **K.Ponder**)
Slide credit: Rick Kessler

Representativeness

Training set based off spectroscopic sample

- Brighter objects
- Lower redshifts
- More well-sampled light curves
- Different percentages in the training set than in the test set

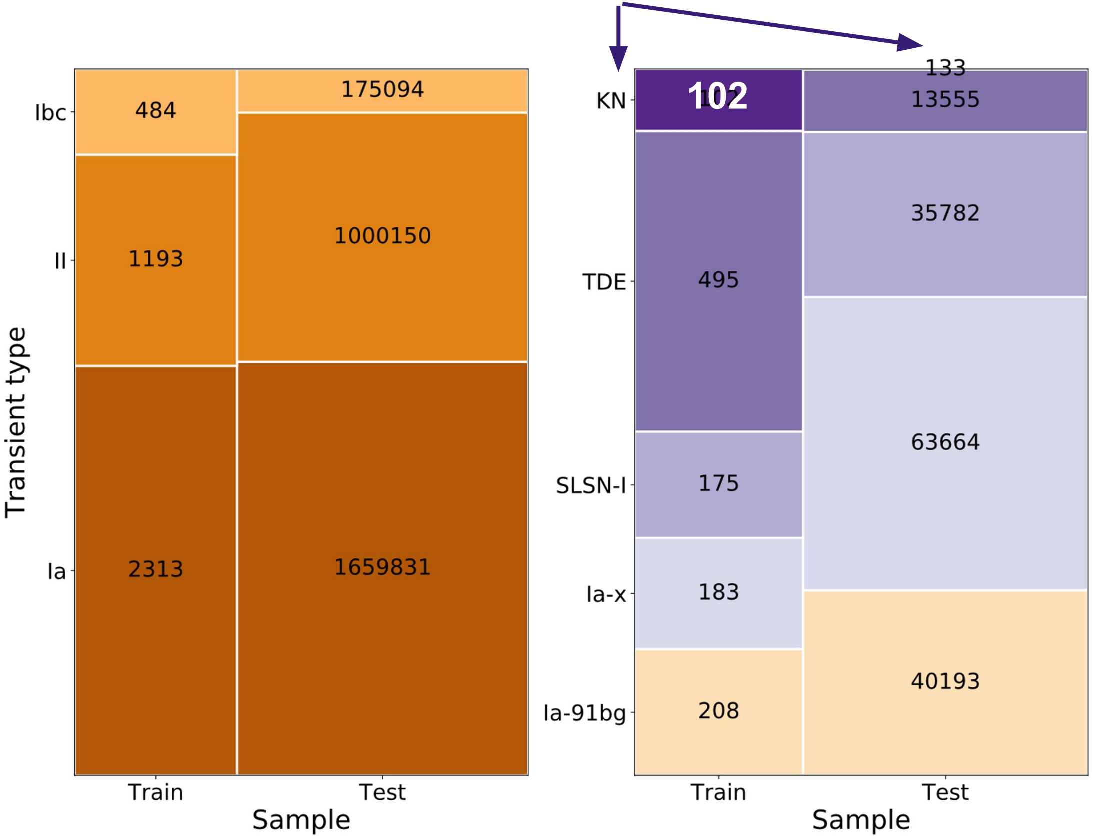


Image credit: Connor Sheere

Validate the simulations

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 (SNANA) had never been used for galactic transients

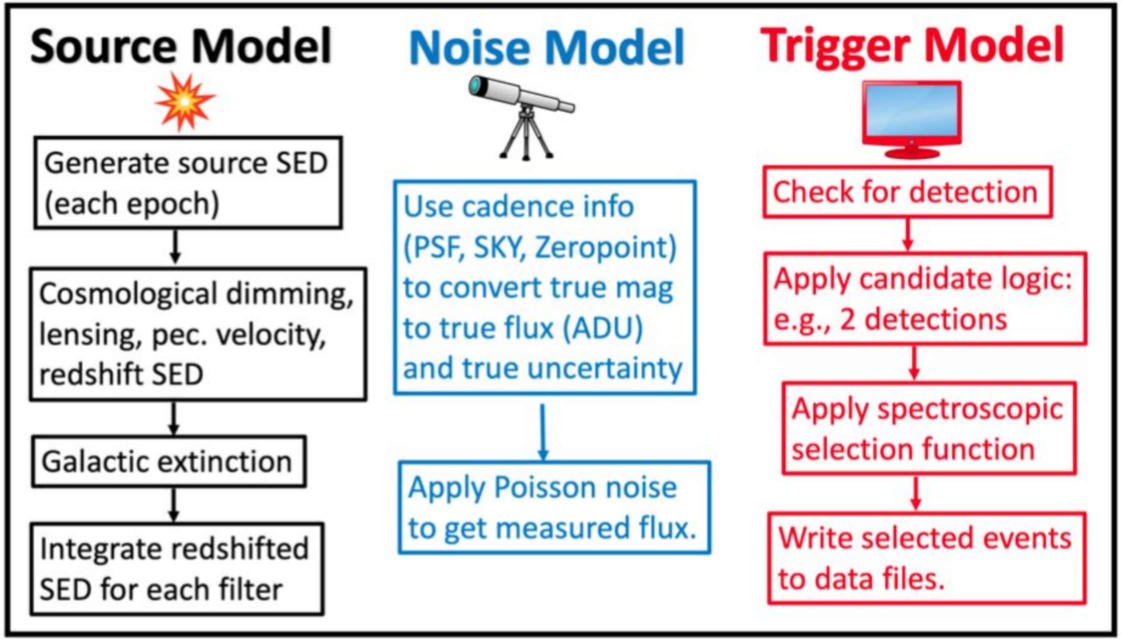


Image credit: Kessler et al (2019)

Method to the madness: How to validate PLAsTiCC

- 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

Method to the madness: How to validate PLAsTiCC

Distribution tests

Maximum Flux

Minimum Flux

Redshift

Rates

- 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

Method to the madness: How to validate PLAsTiCC

Distribution tests

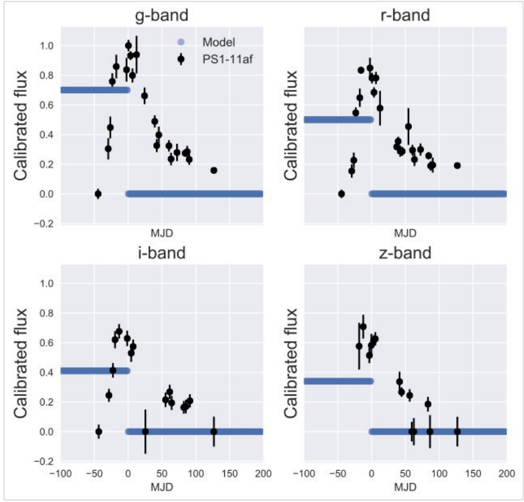
- Maximum Flux
- Minimum Flux
- Redshift
- Rates

Light curves

Visual inspection : limited to ~few hundred objects → DDF, WFD, Model

Comparing model to DDF/WFD

Comparing real data to model



- 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

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

- 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

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

- 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

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

- 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

By the numbers

- More than 1 million new SEDs across several new models in SNANA
- ~3.5 million objects in the test set with <8,000 objects for training
  - 15 classes in the test set, 14 in the training set
- ~450 million observations in over 6 filter bands (18.5 GB)

Even simplified, PLAsTiCC is the largest simulation ever of light curves in the time domain sky in the optical

Slide credit: Gautham Narayan

Run the Challenge!

Preparing to interact with community: Starter Kit

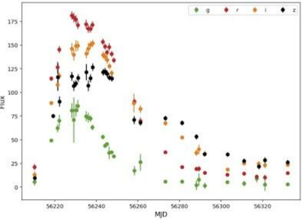
What is LSST?

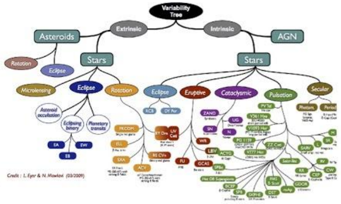
What is a Hubble diagram?

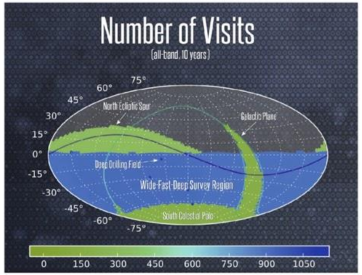
What is cadence?

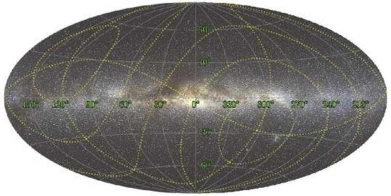
What is Milky Way dust?

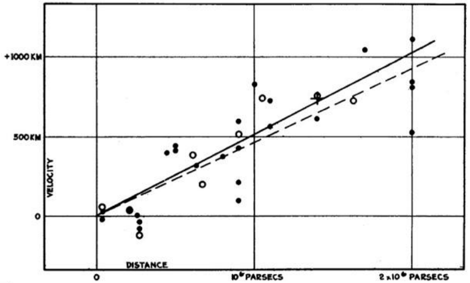
What is redshift?

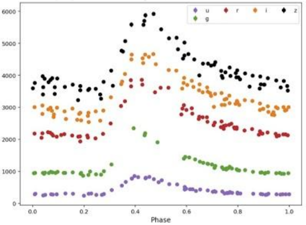


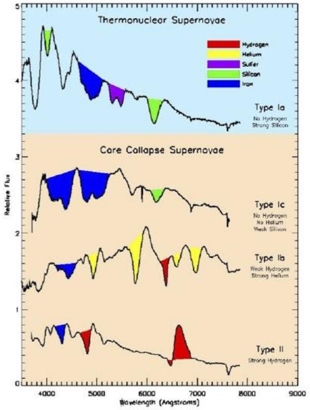


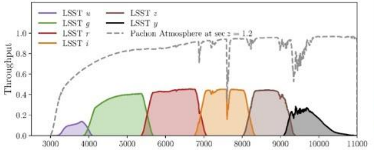


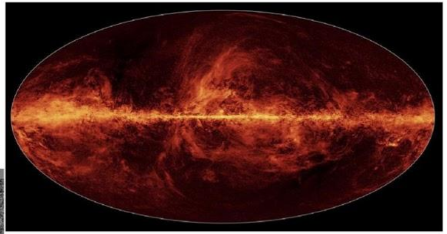


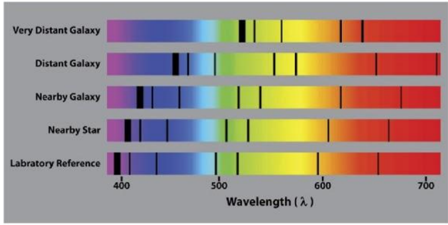


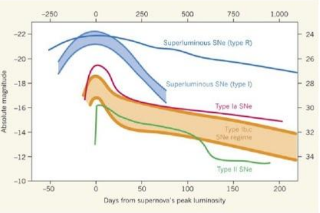


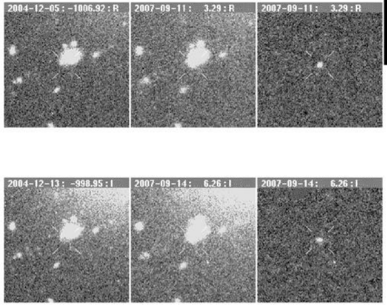












What is photometry? Spectroscopy?

What does the telescope see?

Need to continue to be engaged

Kaggle

Donated by Kaggle

Vital to success of challenge

Featured Prediction Competition

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

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.

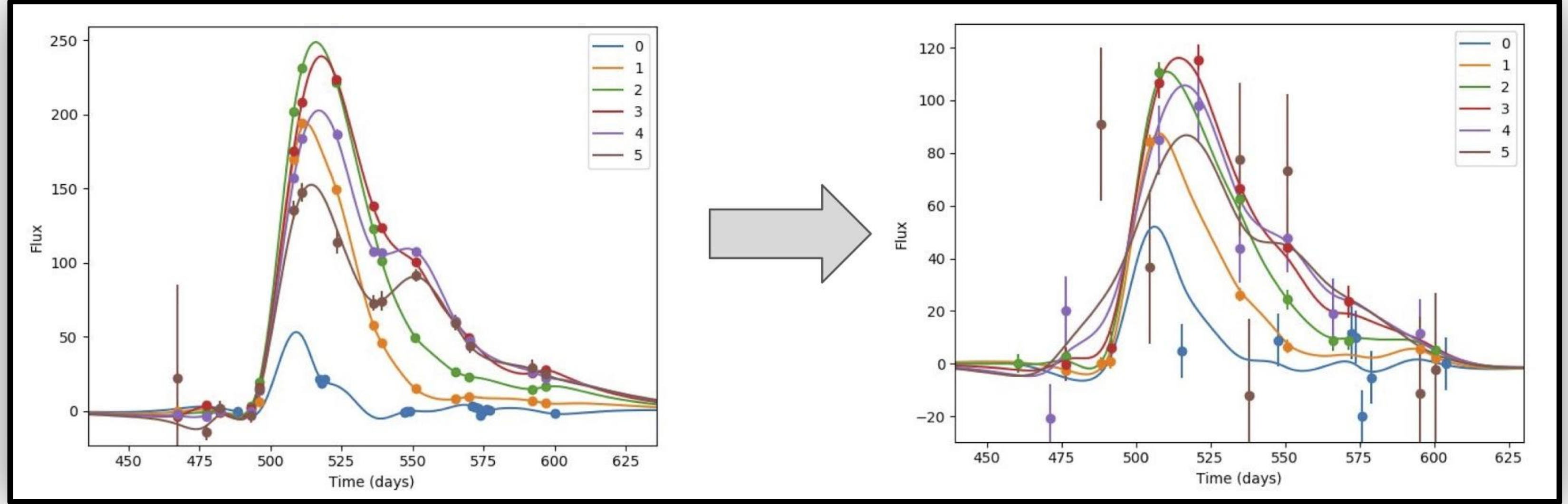
More background information is available here.

1,094	1,325	22,889
Teams	Competitors	Entries

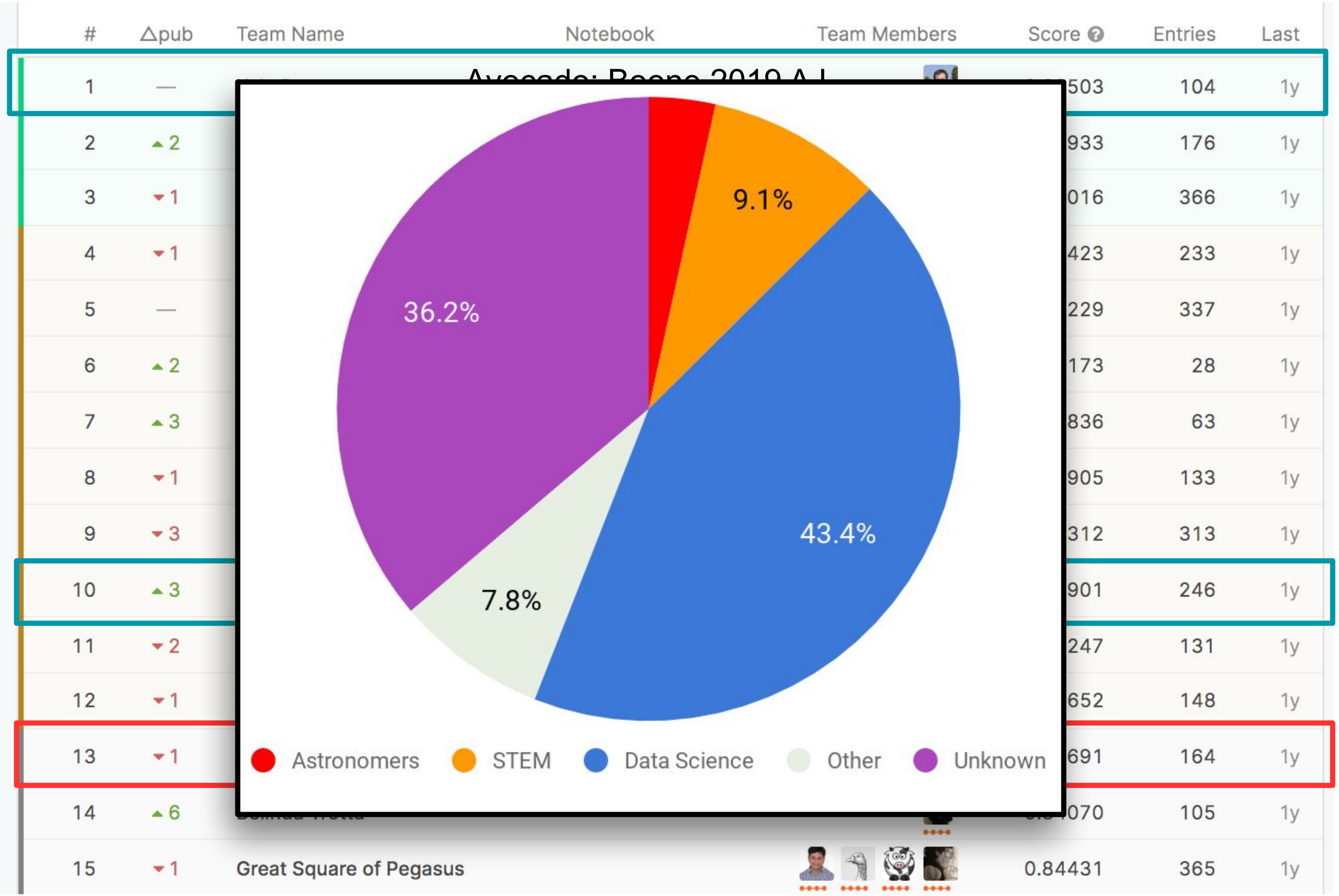
September 28, 2018 - December 17, 2018

#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Kyle Boone	Avocado: Boone 2019 AJ		0.68503	104	1y
2	▲ 2	Mike & Silogram			0.69933	176	1y
3	▼ 1	Major Tom			0.70016	366	1y
4	▼ 1	AhmetErdem			0.70423	233	1y
5	—	SKZ Lost in Translation			0.75229	337	1y
6	▲ 2	Stefan Stefanov			0.80173	28	1y
7	▲ 3	hklee			0.80836	63	1y
8	▼ 1	rapids.ai		RAPIDS	0.80905	133	1y
9	▼ 3	Three Musketeers			0.81312	313	1y
10	▲ 3	J&J	PELICAN: Pasquet et al 2019 A&A		0.81901	246	1y
11	▼ 2	SimonChen			0.82247	131	1y
12	▼ 1	Go Spartans!			0.82652	148	1y
13	▼ 1	Day meets Night	Arxiv: 1909.05032		0.82691	164	1y
14	▲ 6	Belinda Trotta			0.84070	105	1y
15	▼ 1	Great Square of Pegasus			0.84431	365	1y

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Kyle Boone	Avocado: Boone 2019 AJ		0.68503	104	1y
2	▲ 2	Mike & Silogram			0.69933	176	1y
3	▼ 1	Major Tom			0.70016	366	1y

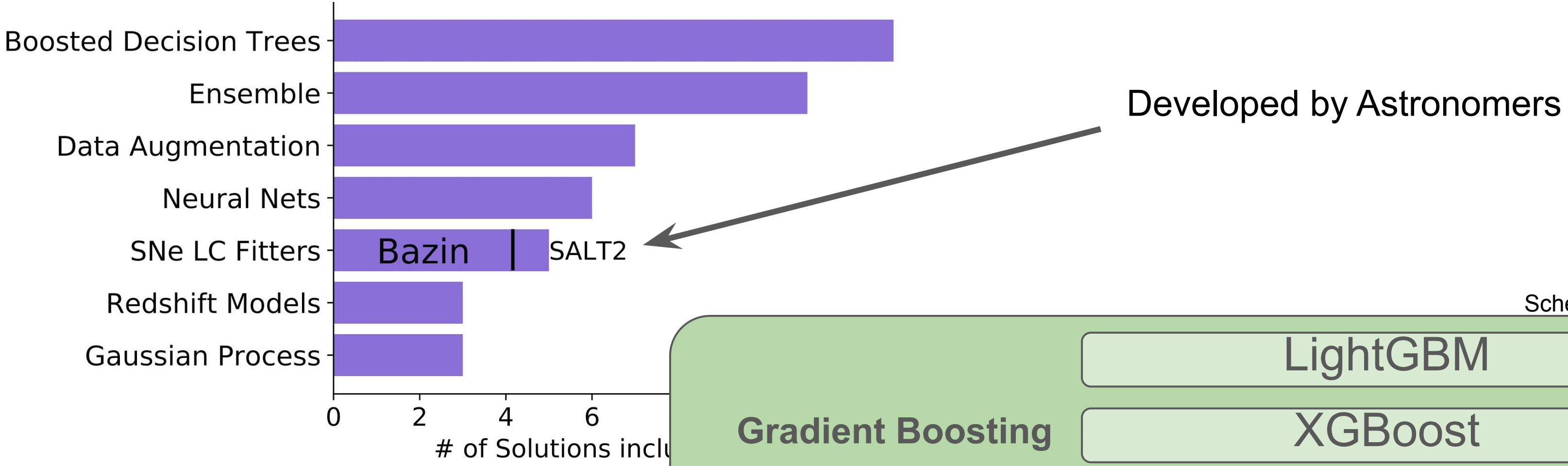


#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
12	▼ 1	Go Spartans!			0.82652	148	1y
13	▼ 1	Day meets Night	Arxiv: 1909.05032		0.82691	164	1y
14	▲ 6	Belinda Trotta			0.84070	105	1y
15	▼ 1	Great Square of Pegasus			0.84431	365	1y

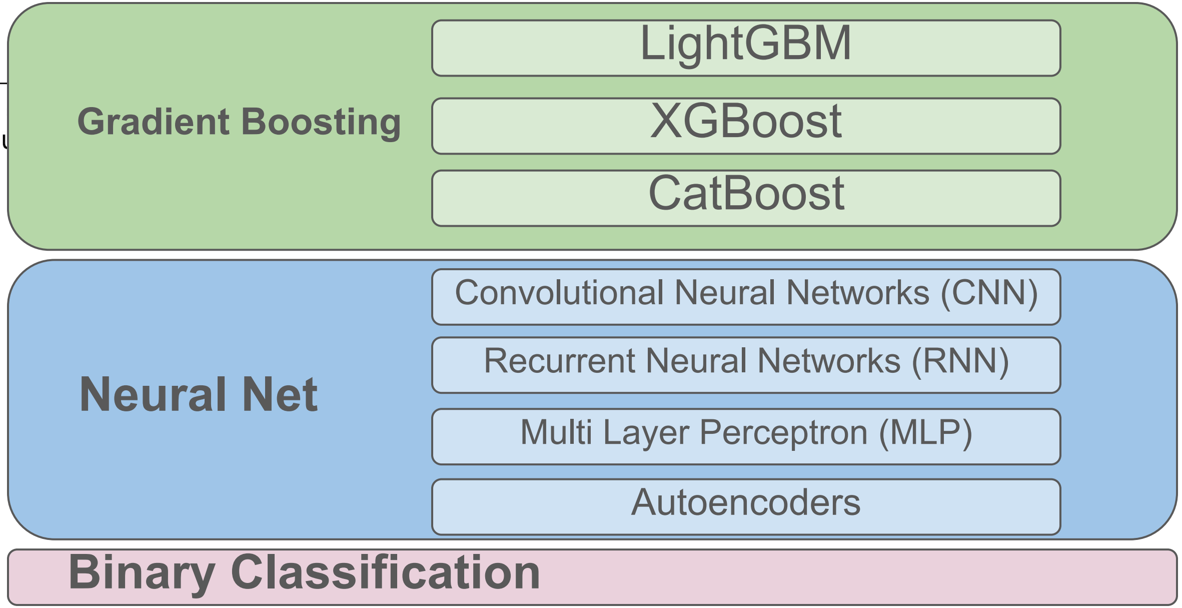


Δpub
Team Name
Notebook
Team Members
Score
Entries
Last
9.1%
36.2%
43.4%
7.8%
Astronomers
STEM
Data Science
Other
Unknown
15
Great Square of Pegasus
0.84431
365
1y

What components did the top 20 solutions have?



Schematic credit: Mi Dai



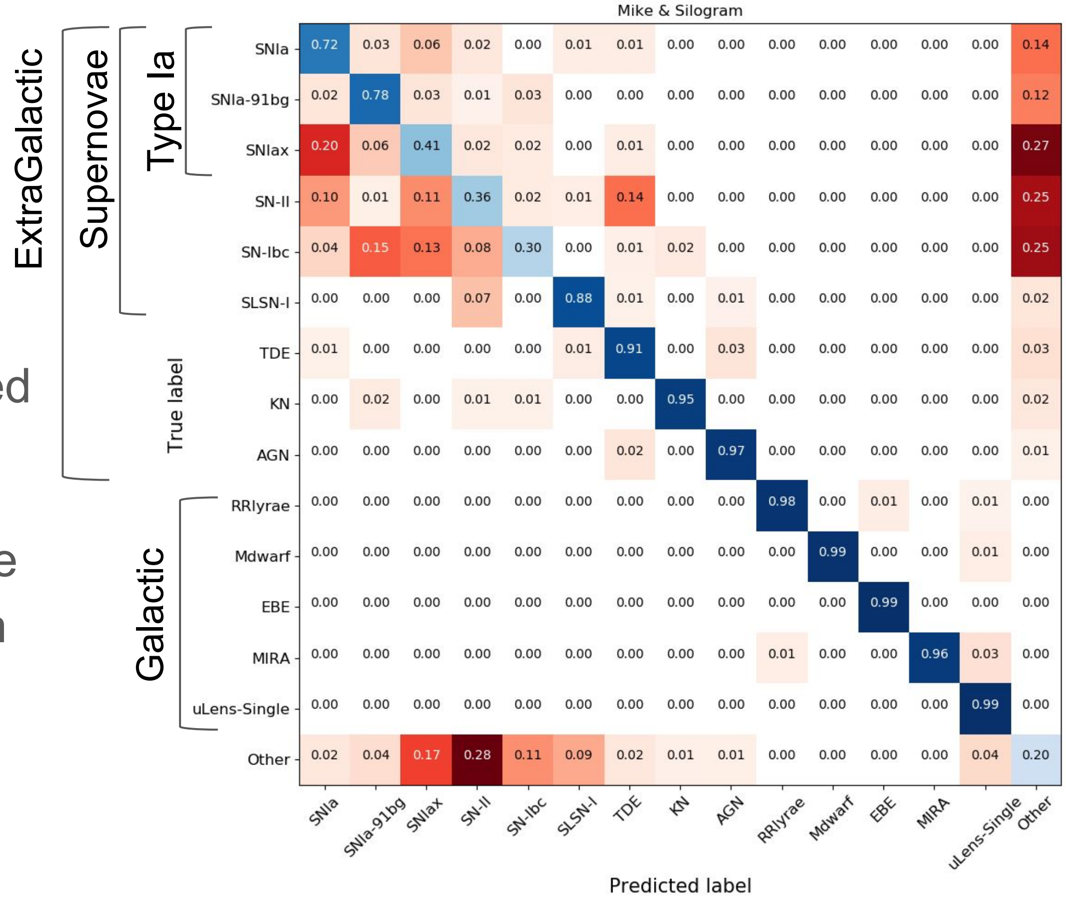
Many solutions forked a Kernel with LightGBM Boosted Decision Trees

Hlozek, **Ponder**, et al, in prep

Things that failed: Anomaly Detection

- No one was able to classify the unknown unknowns
- Most people probed the leaderboard
- Or weighted the classes based on the discussion boards

These will be useful for future studies in anomaly detection even though they failed for PLAsTiCC.



What have we learned?

- It seems to help to be an astronomer
- Data Augmentation may help with non-representativeness
- Utilizing many different methods and combining them may help
- Boosted Decision Trees work well with our data

- People think astronomy is cool, but they like to test their model and move on to the next opportunity
    - Science Competition until January 17, 2019
    - We required documented code on GitHub -- Received 4 entrees
    - All entrees invited to LSST Supernova Science Collaborations meeting this April in Illinois
- Data scientists don’t have ALL the answers, but they help
- We still have a lot of work to do!

PLAsTiCC 2.0

Stay tuned!

- PLAsTiCC was not perfect…
  - Objects were not placed in real galaxies
  - Some of the models were lacking variation
  - We only had 18 models!
  - Catalog only models, could have started from simulated images
- Kaggle dictated we only ask one question
  - We had many! Anomaly detection, Early classification, Type Ia Supernova
- We are still deciding what questions to ask and what community to ask them to!

Other ways to convince people to solve your problems...

Cosmostatistics Initiative

Aim: To create an interdisciplinary community around a
driven problems in astronomy

Method: COIN Residence Programs (CRPs)

- Unstructured meetings
- 12 or less people
- Goal to finish a project in one week
- Researchers in astronomy, statistics, computer science and related fields
- Allows for collaborative research and expertise sharing with concrete goals
- Create a nurturing and intimate environment to build connections

Led by Rafael de Souza, Emille Ishida, Alberto Krone-Martins

https://cosmostatistics-initiative.org/

CRP #6: Chamonix, France

Before the meeting, 14 project ideas were presented.

Chose 3 projects for the week

My project
(co-first author A. Malz):

Use data driven methods to generate mock galaxy catalogs.

August 2019

No connection to my previous work other than its use for cosmology.
Used my transferable skills: **data analysis, statistics, methodology, python, git, ...**

Standard procedure:

N-body simulation → Halo model & Inpainting scheme → Their mock catalog

N-body simulation

Known cosmology

Dark Matter particle 3-positions
- ➔ particle correlation function

3-velocities
Redshift

Halo model & Inpainting scheme

- ➔ Smeared out small scales
- ➔ Dependent on functional approximation

Their mock catalog

Known cosmology

Galaxy angular positions
- ➔ galaxy correlation function

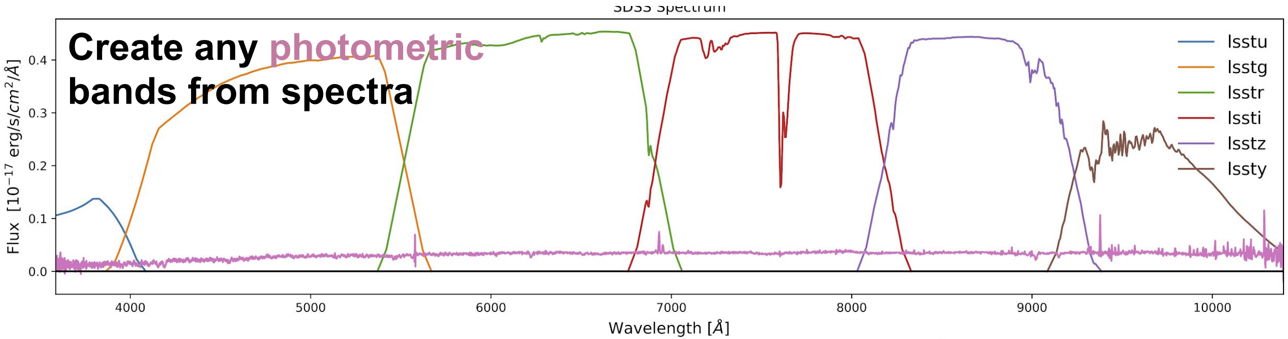
Redshift
Spectra (Photometry)

Slide Credit: Alex Malz

Our approach:

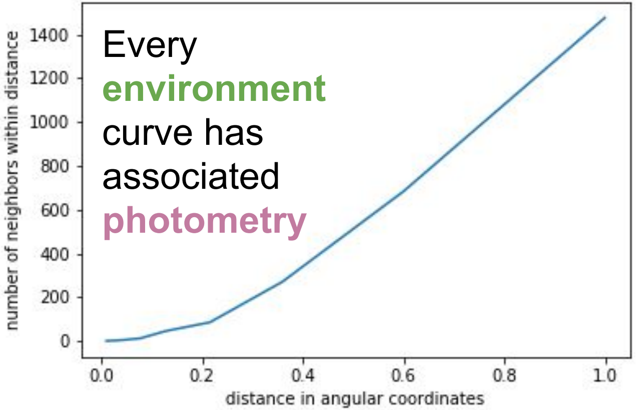
N-body simulation	Our mock catalog	Real Galaxy catalog
Known cosmology	Known cosmology	Unknown cosmology
Dark Matter particle 3-positions ➔ particle correlation function	Galaxy angular positions ➔ galaxy correlation function ➔ **<u>“environment”</u>**	Galaxy angular positions ➔ galaxy correlation function ➔ **“environment”**
3-velocities Redshift	Redshift Photometry	Redshift Spectra (Photometry)

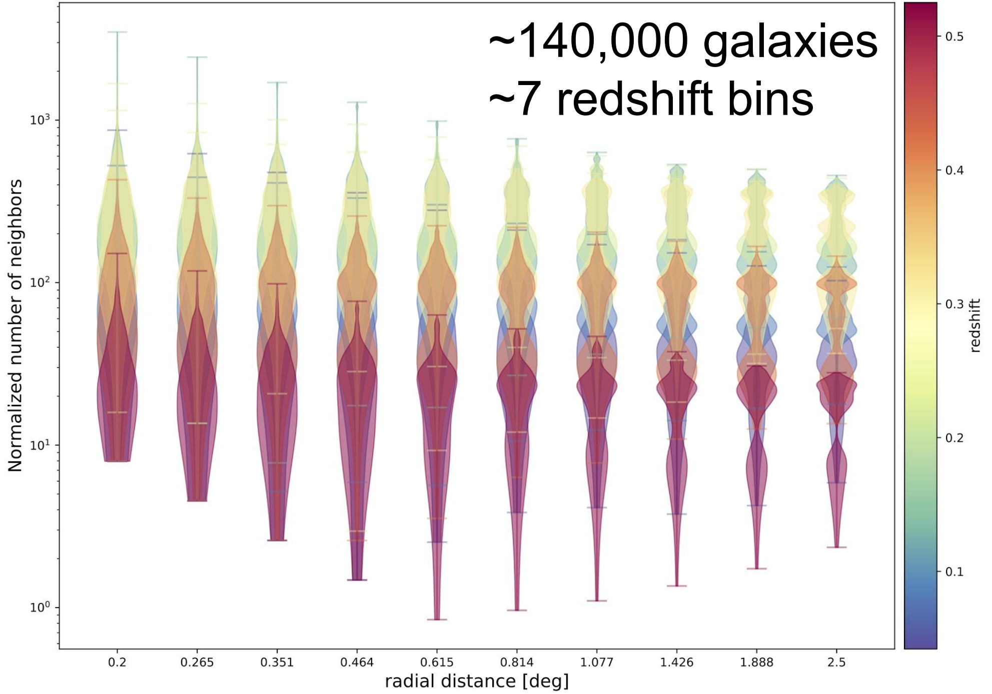
Slide Credit: Alex Malz



Start with real data from Galaxy and Mass Assembly (GAMA) catalog.

Define **environment**

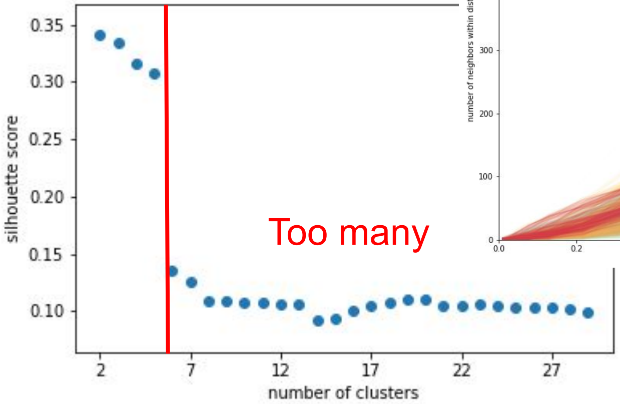


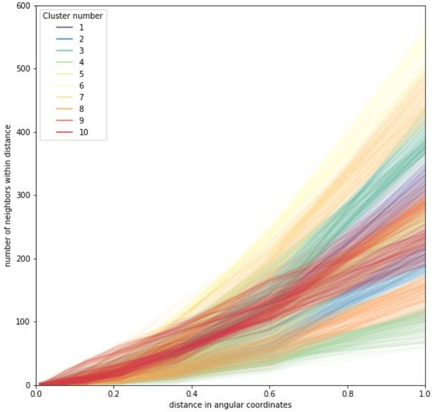


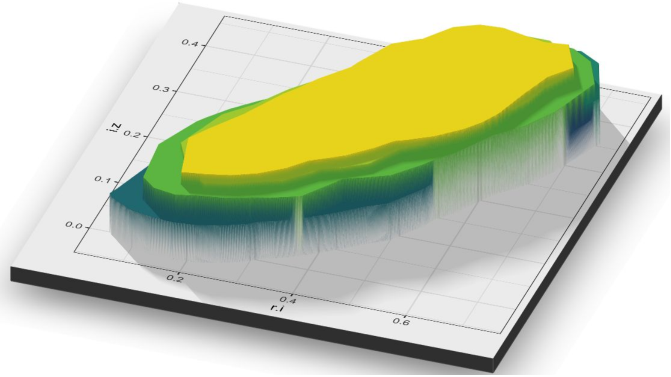
Malz & **Ponder** et al, in prep

Create data driven galaxy property model

- Unsupervised Learning
  - Time Series K-means clustering on **environment** curves
  - Retains vector information
  - How many clusters are supported?
- Model Per redshift bin
- Compare **photometry** per cluster







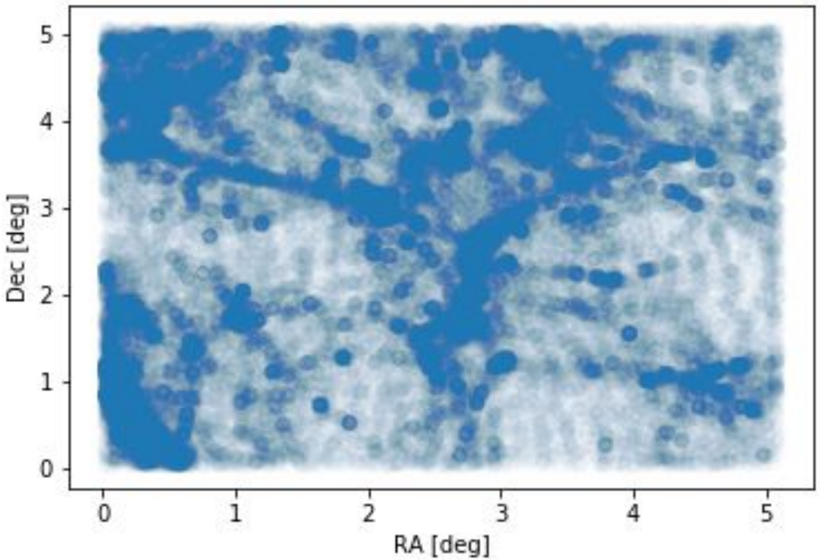
For each **ENVIRONMENT** cluster, fit the synthesized **PHOTOMETRY** with multivariate Gaussian

Malz & **Ponder** et al, in prep

Apply model to mock catalog particles

- Draw sampled particles from N-body simulation (SLICS-HR)

Harnois-Deraps et al (2018)
https://slics.roe.ac.uk/

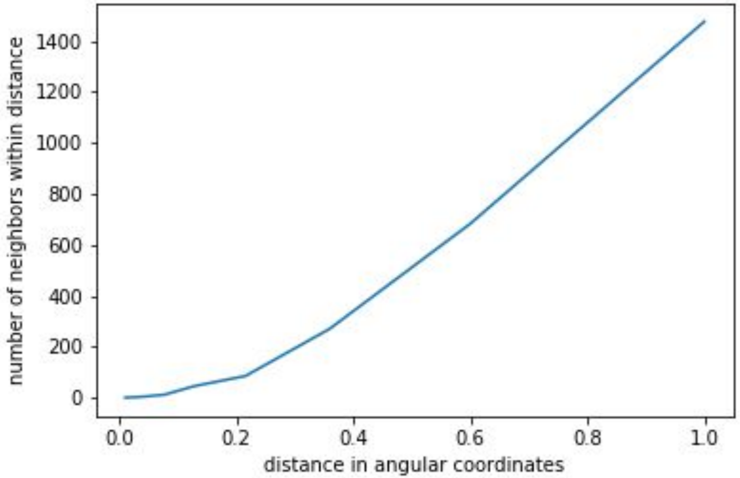


Malz & **Ponder** et al, in prep

All plots on example runs.

Apply model to mock catalog particles

- Draw sampled particles from N-body simulation (SLICS-HR)
- Calculate **environment** of mock catalog object

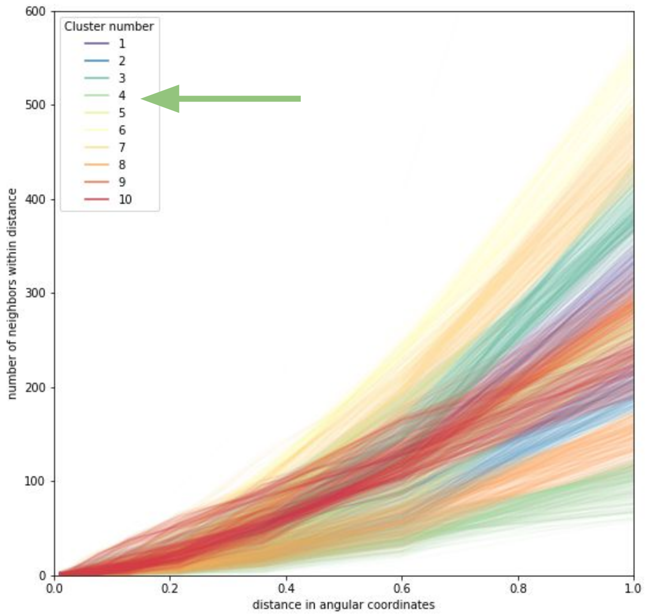


Malz & **Ponder** et al, in prep

All plots on example runs.

Apply model to mock catalog particles

- Draw sampled particles from N-body simulation (SLICS-HR)
- Calculate **environment** of mock catalog object
- Classify **environment** with K-means model

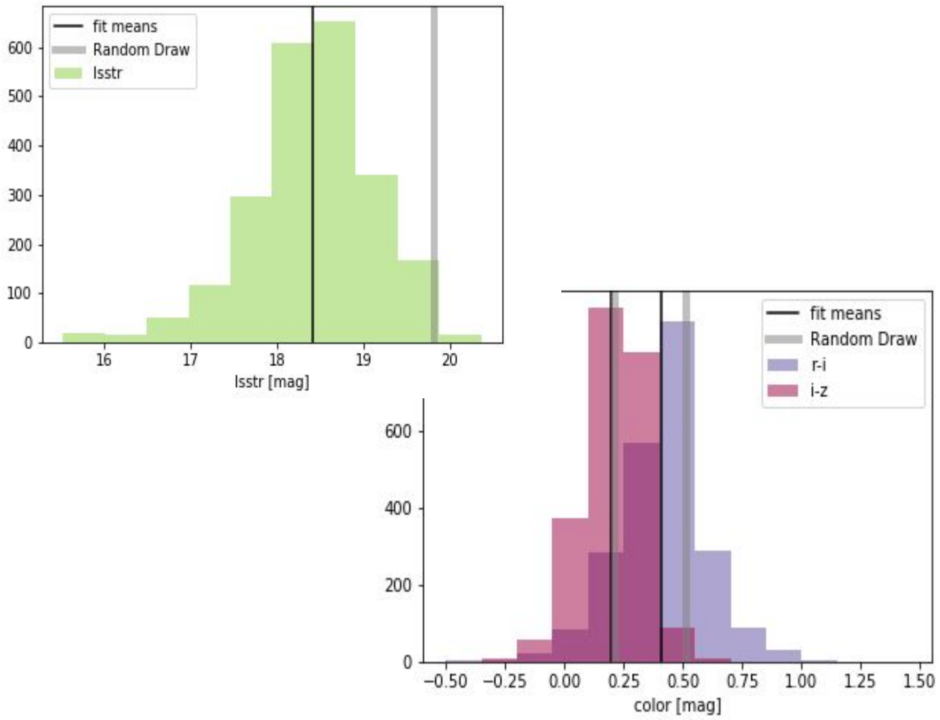


Malz & **Ponder** et al, in prep

All plots on example runs.

Apply model to mock catalog particles

- Draw sampled particles from N-body simulation (SLICS-HR)
- Calculate **environment** of mock catalog object
- Classify **environment** with K-means model
- Randomly draw **photometry** from appropriate **environment** cluster

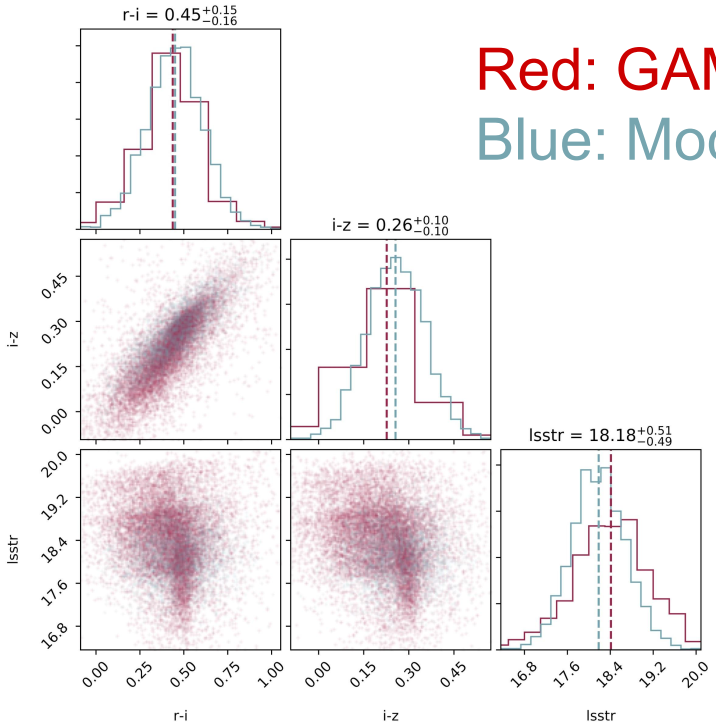


Malz & **Ponder** et al, in prep

All plots on example runs.

Apply model to mock catalog particles

- Draw sampled particles from N-body simulation (SLICS-HR)
- Calculate **environment** of mock catalog object
- Classify **environment** with K-means model
- Randomly draw **photometry** from appropriate **environment** cluster
- Now you have a mock catalog!



Red: GAMA
Blue: Mock

Malz & **Ponder** et al, in prep

All plots on example runs.

Goals:

- Provide proof of concept
- Launch a platform for people to build their own mock catalogs
    - Inputs: Favorite N-body simulation and galaxy survey

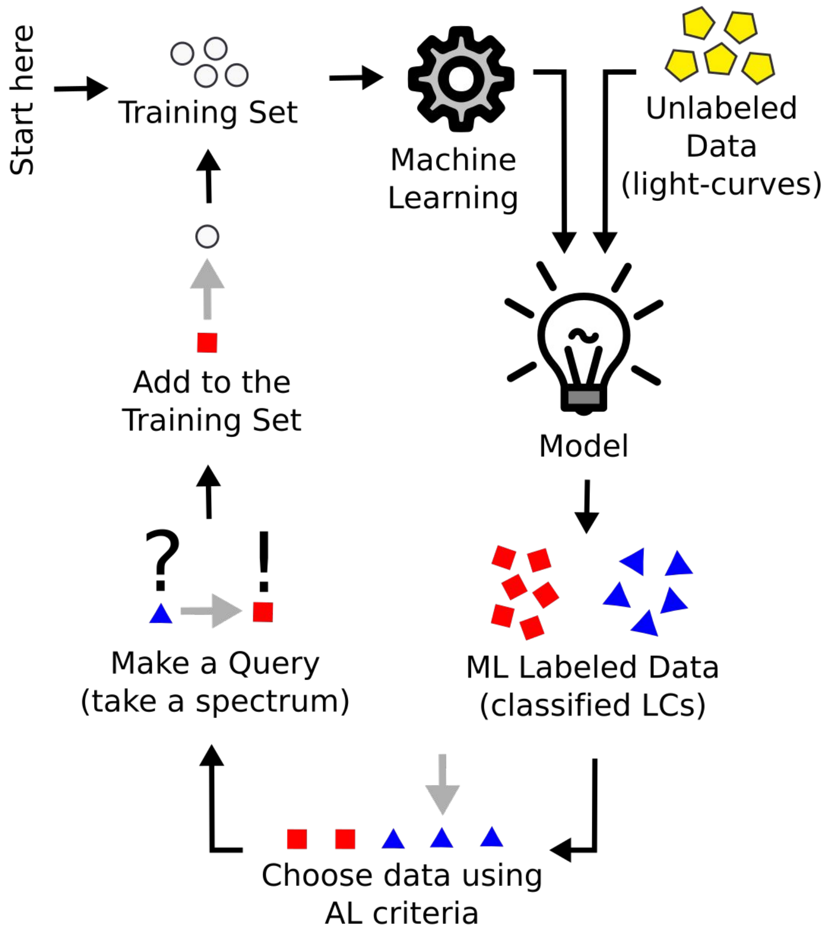
LSST DESC
Dark Energy Science Collaboration

COIN
Cosmostatistics Initiative

LSST DESC & COIN

RESSPECT

Recommendation System for Spectroscopic Follow-up



Active Learning for Photometric Classification of Supernovae

- CRP #4
- Get spectrum for most uncertain object
- Feed back into training set to improve photometric classification

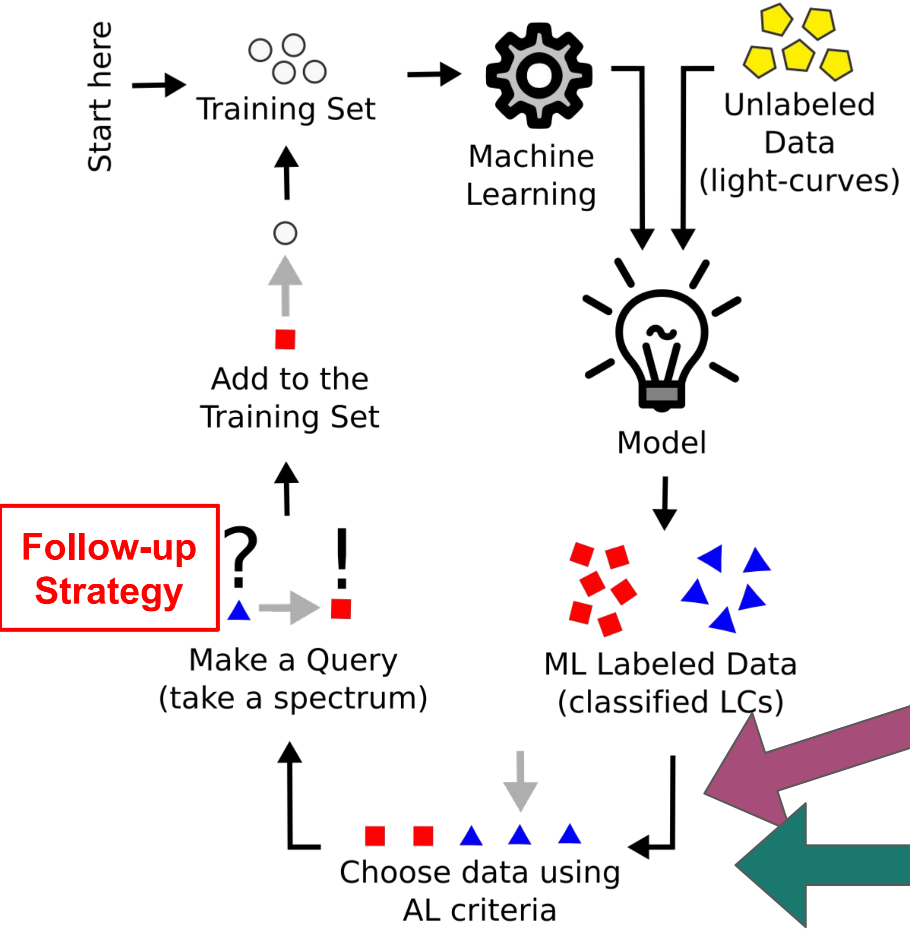
We can use this infrastructure to build **recommendation systems**.

Ishida et al., 2019 - arXiv:astro-ph/1804.03765

RESSPECT

Recommendation System for Spectroscopic Follow-Up

Goal: Build a sample for photometric supernova cosmology.



Telescope Resources

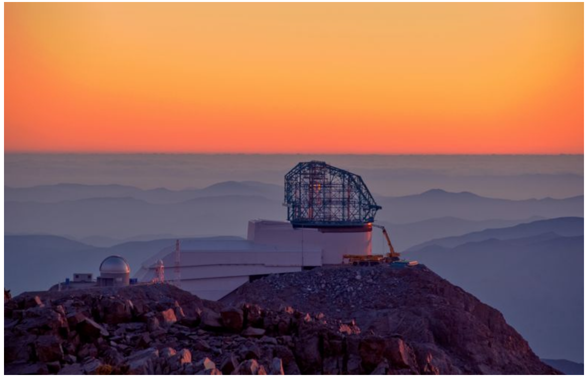
Photometric Cosmology*

Ishida et al., 2019 - arXiv:astro-ph/1804.03765

- There are non-traditional ways of doing research.
- **Decide what is right for your problem**
  - How many different communities are interested? How many people are involved?
  - How much time will take to answer the question?
  - Do you need to collect new data? Generate new data? Is the hardware/software available?
  - Why hasn't this problem been solved?
- **Are Data Challenges are right for you?**
  - **PLAsTiCC**
    - Needed to unify disparate communities
    - New methods appearing quickly in industry data science
    - Needed to motivate a powerhouse to get new simulations
- **Do you need to build a new community?**
  - Do you have a lot of small problems or highly focused problems?
  - **COIN**
    - Goal to build up the Astrostatistics and Astroinformatics community by building individual connections through focused projects
    - Interfacing with existing collaborations to connect experts to non-experts to utilize skills across COIN and the LSST DESC

Will you solve someone else's problem?

VRO, image credit: LSSTC



Thank you!

- Photometric classification is hard
- PLAsTiCC opened opportunities for new approaches
- PLAsTiCC was an overall success! The model paper alone has more than 15 citations
    - We will be able to do more focused SN Ia Cosmology and cadence studies
- COIN: Mock Galaxy Catalog: Outcome will be package or web interface that allows people to create data-driven mock galaxy catalogs
- RESSPECT: Building a recommendation system to maximize cosmology with a photometric sample

Kara Ponder **kponder@berkeley.edu**