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

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

Cosmology with photometric classification

Algorithms for Photometric Classification

- PSNID (Sako et al 2011)
- PELICAN (Pasquet et al 2019)
- SuperNNova (Moller et al 2020)
- snmachine (Lochner et al 2016)
- Villar et al (2019)
- Karpenka et al (2013)
- Newling et al (2011)
- Bloom et al (2012) - transients & variables
- Godines et al (2019) - microlensing

At least dozens more for other transients and variable!
Comparing different classification methods

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

AUC: Area Under ROC Curve -- Closer to 1 is better
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 et al 2019)
- SuperNNova (Moller et al 2020)
- snmachine (Lochner et al 2016)
- Villar et al (2019)
- Karpenka et al (2013)
- Newling et al (2011)
- Bloom et al (2012) - transients & variables
- Godines et al (2019) - microlensing

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)
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?

### Data and compute sizes:
- Final volume of raw image data = 60 PB
- Final catalog size (DR11) = 15 PB
- Peak compute power in LSST data centers = about 2 PFLOPS

### Network bandwidths:
- Summit (Cerro Pachón) - Base (La Serena) = 600 Gbps
- Base (La Serena) to Archive (NCSA) = 2 x 100 Gbps

### Alert Production:
- Real-time alert latency = 60 seconds
- Estimated number of alerts per night = up to about 10 million

### Data Releases:
- Number of Data Releases = 11
- Images collected = 5.5 million 3.2 Gigapixel images

### Estimated counts for DR1
- (produced from first 6 months of observing)
  - Objects = 18 billion; Sources = 350 billion
  - (single epoch); Forced Sources = 0.75 trillion

### Estimated counts for DR11
- Objects = 37 billion; Sources = 7 trillion
  - (single epoch); Forced Sources = 30 trillion
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:


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

→ 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
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?
The Metric

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

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

Malz et al (2019, including K.Ponder)
AJ, 158, 5, 171
Collecting models to simulate data
### Summary of Models used in PLAsTiCC

<table>
<thead>
<tr>
<th>Model Class</th>
<th>Model Description</th>
<th>Contributor(s)</th>
<th>Nevent Gen</th>
<th>Nevent train</th>
<th>Nevent test</th>
<th>redshift range</th>
</tr>
</thead>
<tbody>
<tr>
<td>90: SNIa</td>
<td>WD detonation, Type Ia SN</td>
<td>RK</td>
<td>16,453,270</td>
<td>2,213</td>
<td>1,659,831</td>
<td>&lt; 1.6</td>
</tr>
<tr>
<td>67: SNIa-91bg</td>
<td>Peculiar type Ia: 91bg</td>
<td>SG, LG</td>
<td>1,329,510</td>
<td>208</td>
<td>40,193</td>
<td>&lt; 0.9</td>
</tr>
<tr>
<td>52: SNIax</td>
<td>Peculiar SNax</td>
<td>SJ, MD</td>
<td>8,660,920</td>
<td>183</td>
<td>63,064</td>
<td>&lt; 1.3</td>
</tr>
<tr>
<td>42: SNII</td>
<td>Core Collapse, Type II SN</td>
<td>SG, LG:RK, JRP:VAV</td>
<td>59,198,660</td>
<td>1,193</td>
<td>1,000,150</td>
<td>&lt; 2.0</td>
</tr>
<tr>
<td>62: SNIIbc</td>
<td>Core Collapse, Type IIbc SN</td>
<td>VAV:RK, JRP</td>
<td>22,599,840</td>
<td>484</td>
<td>175,094</td>
<td>&lt; 1.3</td>
</tr>
<tr>
<td>95: SLSN-I</td>
<td>Super-Lum. SN (magnetar)</td>
<td>VAV</td>
<td>90,640</td>
<td>175</td>
<td>35,782</td>
<td>&lt; 3.4</td>
</tr>
<tr>
<td>15: TDE</td>
<td>Tidal Disruption Event</td>
<td>VAV</td>
<td>58,550</td>
<td>405</td>
<td>13,555</td>
<td>&lt; 2.6</td>
</tr>
<tr>
<td>64: KSN</td>
<td>Kilonova (NS-NS merger)</td>
<td>DK, GN</td>
<td>43,150</td>
<td>100</td>
<td>131</td>
<td>&lt; 0.3</td>
</tr>
<tr>
<td>88: AGN</td>
<td>Active Galactic Nuclei</td>
<td>SD</td>
<td>175,500</td>
<td>390</td>
<td>1,01,424</td>
<td>&lt; 3.4</td>
</tr>
<tr>
<td>92: RRL</td>
<td>RR lyrae</td>
<td>SD</td>
<td>200,200</td>
<td>239</td>
<td>197,155</td>
<td>0</td>
</tr>
<tr>
<td>65: M-dwarf</td>
<td>M-dwarf stellar flare</td>
<td>SD</td>
<td>800,800</td>
<td>981</td>
<td>93,494</td>
<td>0</td>
</tr>
<tr>
<td>16: EB</td>
<td>Eclipsing Binary stars</td>
<td>AP</td>
<td>220,200</td>
<td>924</td>
<td>96,572</td>
<td>0</td>
</tr>
<tr>
<td>53: Mira</td>
<td>Pulsatizing variable stars</td>
<td>RH</td>
<td>1,490</td>
<td>30</td>
<td>1,453</td>
<td>0</td>
</tr>
<tr>
<td>6: uLens-Single</td>
<td>u-lens from single lens</td>
<td>RD, AA: EB, GN</td>
<td>2,820</td>
<td>151</td>
<td>1,303</td>
<td>0</td>
</tr>
<tr>
<td>791: uLens-Binary</td>
<td>u-lens from binary lens</td>
<td>RD, AA</td>
<td>1,010</td>
<td>0</td>
<td>533</td>
<td>0</td>
</tr>
<tr>
<td>992: ILOT</td>
<td>Intermed. Lumi. Optical Trans.</td>
<td>VAV</td>
<td>4,521,970</td>
<td>0</td>
<td>1,702</td>
<td>&lt; 0.4</td>
</tr>
<tr>
<td>993: CaRT</td>
<td>Calcium Rich Transient</td>
<td>VAV</td>
<td>2,834,500</td>
<td>0</td>
<td>9,680</td>
<td>&lt; 0.9</td>
</tr>
<tr>
<td>994: PISN</td>
<td>Pair Instability SN</td>
<td>VAV</td>
<td>5,650</td>
<td>0</td>
<td>1,172</td>
<td>&lt; 1.9</td>
</tr>
<tr>
<td>995: uLens-String</td>
<td>u-lens from cosmic strings</td>
<td>DC</td>
<td>30,020</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>Sum of all models</td>
<td></td>
<td>117,128,700</td>
<td>7,846</td>
<td>3,492,888</td>
<td>—</td>
</tr>
</tbody>
</table>

*a* sum > 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.

---


**Simulation Source code:** [http://snana.uchicago.edu](http://snana.uchicago.edu)

---

**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 Hoyle (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**

---

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
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
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

- 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

- 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
- Comparing model to DDF/WFD
- Comparing real data to model

Classification codes to search for unphysical correlations

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 Milky Way dust?

What is a Hubble diagram?

What is cadence?

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

September 28, 2018 - December 17, 2018

PLAsTiCC Astronomical Classification
Can you help make sense of the Universe?

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

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

- **Astronomers**: 43.4%
- **STEM**: 36.2%
- **Data Science**: 7.8%
- **Other**: 9.1%
- **Unknown**: 3.8%

Great Square of Pegasus
What components did the top 20 solutions have?

- Boosted Decision Trees
- Ensemble
- Data Augmentation
- Neural Nets
- SNe LC Fitters
- Bazin
- SALT2
- Redshift Models
- Gaussian Process

Many solutions forked a Kernel with LightGBM Boosted Decision Trees.

Developed by Astronomers

Schematic credit: Mi Dai
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

- 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!

Stay tuned!
Other ways to convince people to solve your problems...
Cosmostatistics Initiative

Aim: To create an interdisciplinary community around data 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.

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

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)
Our approach:

N-body simulation → Our mock catalog → Real Galaxy catalog

- **Known cosmology**
  - Dark Matter particle 3-positions → particle correlation function
  - 3-velocities
  - Redshift

- **Known cosmology**
  - Galaxy angular positions → galaxy correlation function → "environment"
  - Redshift
  - Photometry

- **Unknown cosmology**
  - Galaxy angular positions → galaxy correlation function → "environment"
  - Redshift
  - Spectra (Photometry)

Slide Credit: Alex Malz
Start with real data from Galaxy and Mass Assembly (GAMA) catalog.

Define environment

Every environment curve has associated photometry

Malz & Ponder et al, in prep

Create any photometric bands from spectra

~140,000 galaxies
~7 redshift bins
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/
Apply model to mock catalog particles

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

Malz & Ponder et al, in prep

Red: GAMA
Blue: Mock

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
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.

RESSPECT

Recommendation System for Spectroscopic Follow-Up

Goal: Build a sample for photometric supernova cosmology.

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?
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