Distance Metrics for Machine Learning in Time-Domain Astronomy

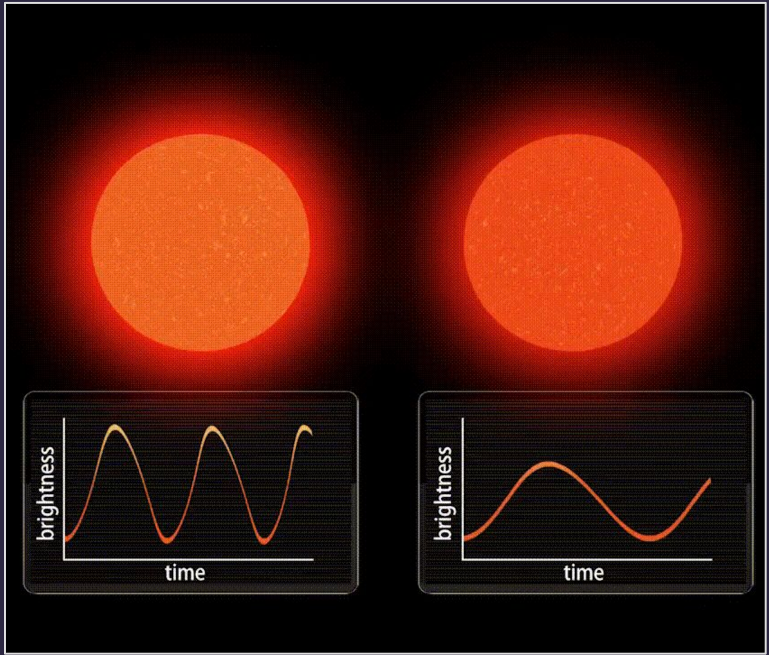
Siddharth Chaini (17275)
Department of Physics, IISER Bhopal

Advisors:
Prof. Ashish Mahabal (Caltech)
Prof. Ajit Kembhavi (IUCAA)
Prof. Sukanta Panda (IISER Bhopal)

Time-Domain Astronomy

Subfield of astronomy related to how cosmic objects change with time

- Typically 3 main categories of objects:
  - Moving Objects (e.g. Asteroids)
  - Transients - Fade over time (e.g. Supernovae)
  - Variable Objects (e.g. Pulsating Stars)
- Changes in brightness can be studied with light curves
- Important right now because of improved telescopes like the Zwicky Transient Facility (ZTF) and the Rubin Observatory



Credits: CAASTRO

Big Data and Machine Learning

- ZTF is already expected to produce over 3.2 petabytes of data by the end of its lifecycle.
- Rubin Observatory will take this to new levels - 15 petabytes expected by 2033
- Machine Learning:
  Computer algorithms that learn tasks and patterns from data.
  - Supervised Learning
  - Unsupervised Learning
- Data in Time-Domain Astronomy: Light curves (in 2 filters for ZTF - g/r)
  but these are irregular and sparse.
- Instead, we extract features from light curves (variability/periodicity/shape/etc.)
- But 108 total features (2 filters) - HIGHLY DIMENSIONAL!!!
  Sanchez-Saez et al. (2021)

A new approach: Distance Metrics

- A distance tells us about the degree of closeness of two physical objects or ideas.

  Even distances between light curves (and their features) !

Calculate distance between light curve feature in four ways:

1. Euclidean Distance: $d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$

2. Cityblock Distance: $d(x,y) = \sum_{i=1}^{n} |y_i - x_i|$

3. Canberra Distance: $d(x,y) = \sum_{i=1}^{n} \frac{|y_i - x_i|}{|x_i| + |y_i|}$

4. Braycurtis Distance: $d(x,y) = \frac{\sum_{i=1}^{n} |y_i - x_i|}{\sum_{i=1}^{n} |x_i + y_i|}$

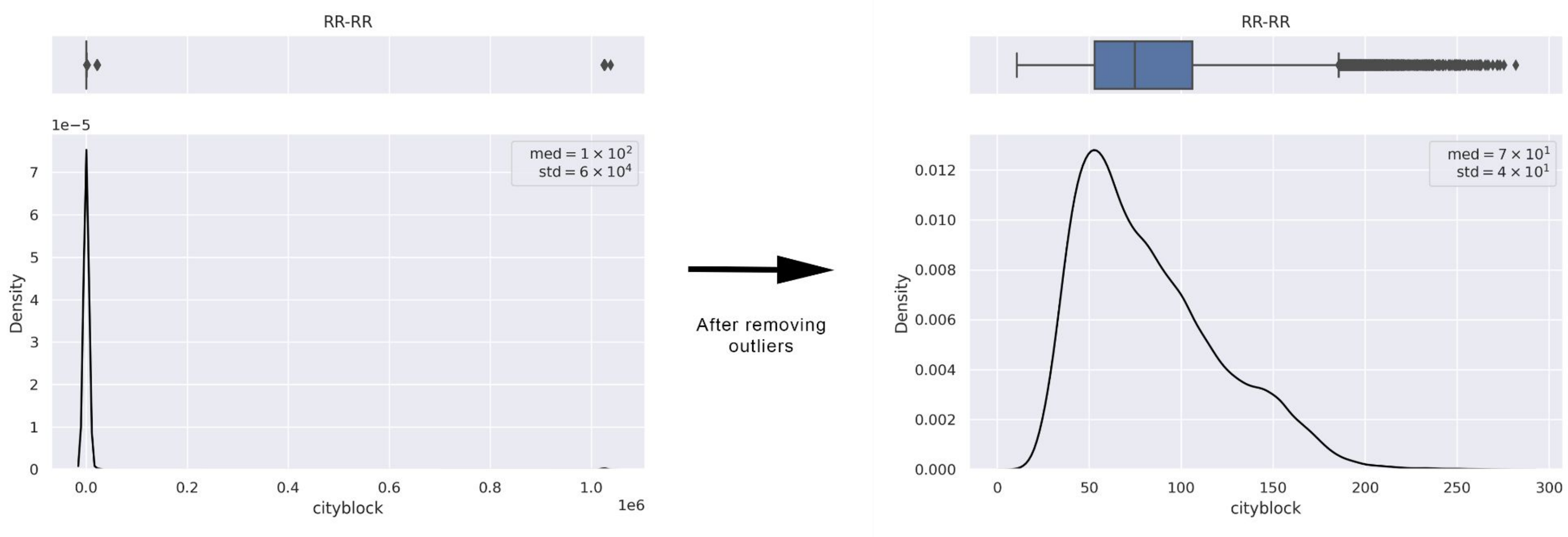
Take 600 objects from each of these 4 classes:

1. BY Draconis variables (BYDra)
2. RR Lyrae variables (RR)
3. Mira variables (Mira)
4. Eclipsing Algol variables (EA)

And calculate all distances pairwise.

Results (I) - Outlier Removal

- For known objects: Distances between objects from the same class



Results (II) - Classification

For unknown objects:

1. Drop outliers
2. Shuffle data, split data into training set and test set
3. Create a “canonical” feature set for each class using training set (median)
4. Calculate distance between test object and the canonical feature set for all 4 classes
5. Set label based on class with minimum distance

Distance Metric	Accuracy
Euclidean	73.61%
Cityblock	82.29%
Canberra	94.10%
Braycurtis	84.03%

Future work

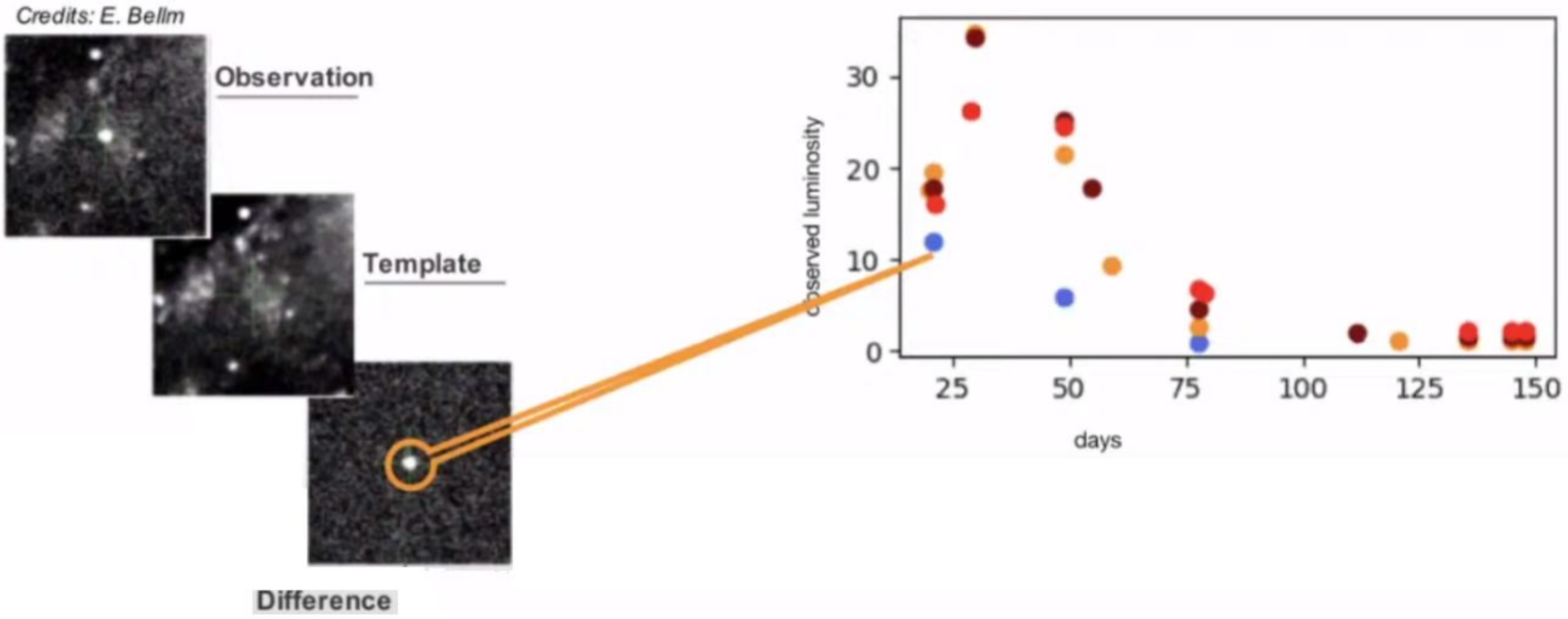
1. Why does Canberra perform better?
2. Combine multiple distances
3. Increase dataset size - number of objects as well as classes
4. Eliminate redundant features - decrease dimensionality
5. Find distance between light curves directly, without feature extraction

Thank you!

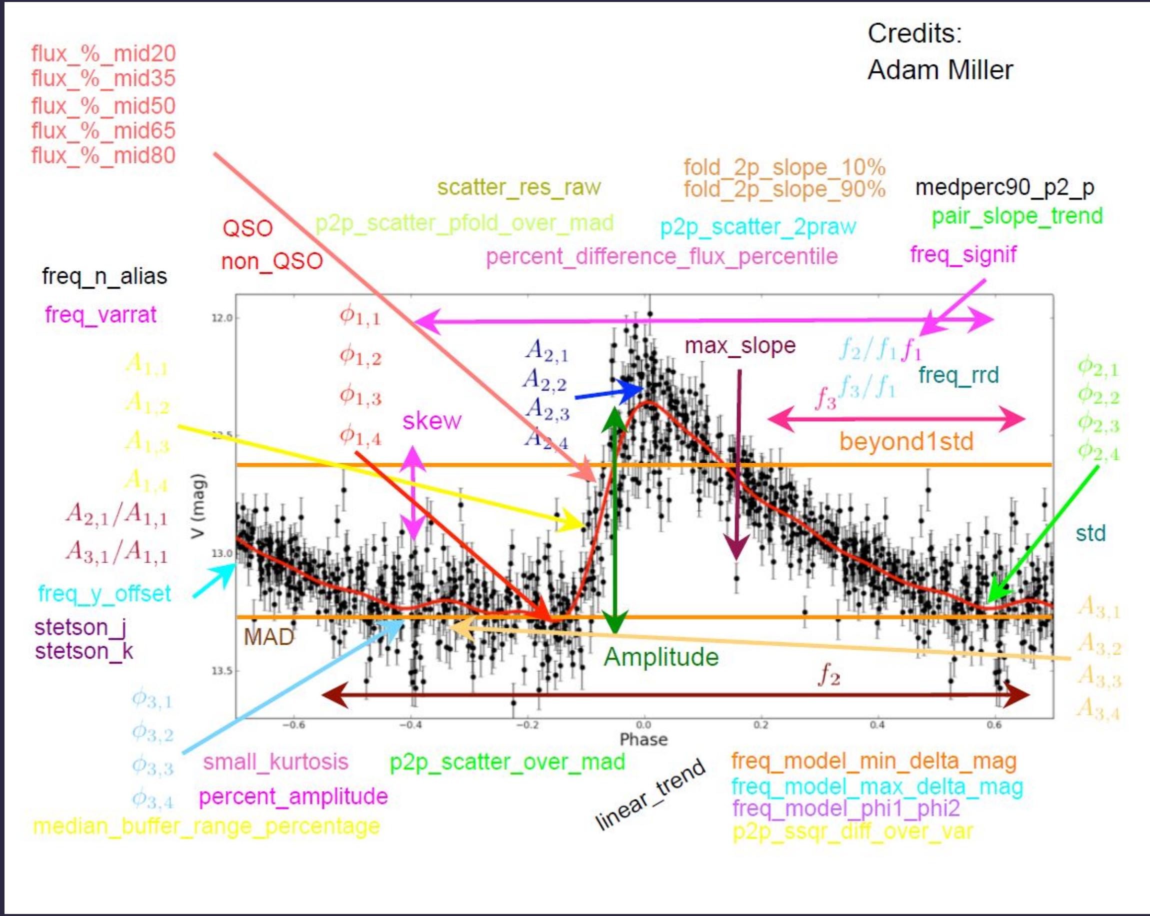
Appendix

App. 01

Light Curve Construction



Feature Extraction



Definition of Distance

Definition 1. *The distance* d *between two points, in a set* X*, is a function* $d : X \times X \to [0, \infty)$ *that gives a distance between each pair of points in that set such that, for all* $x, y, z \in X$*, the following properties hold:*

1. $d(x, y) = 0 \iff x = y$ *(identity of indiscernibles)*
2. $d(x, y) = d(y, x)$ *(symmetry)*
3. $d(x, y) \leq d(x, z) + d(z, y)$ *(triangle inequality)*

The above three axioms also imply the following condition:

$$d(x, y) \geq 0, \text{for all } x, y \in X$$

App. 04

Notebooks

1. distance analysis.ipynb
2. LCdistance_classifier.ipynb