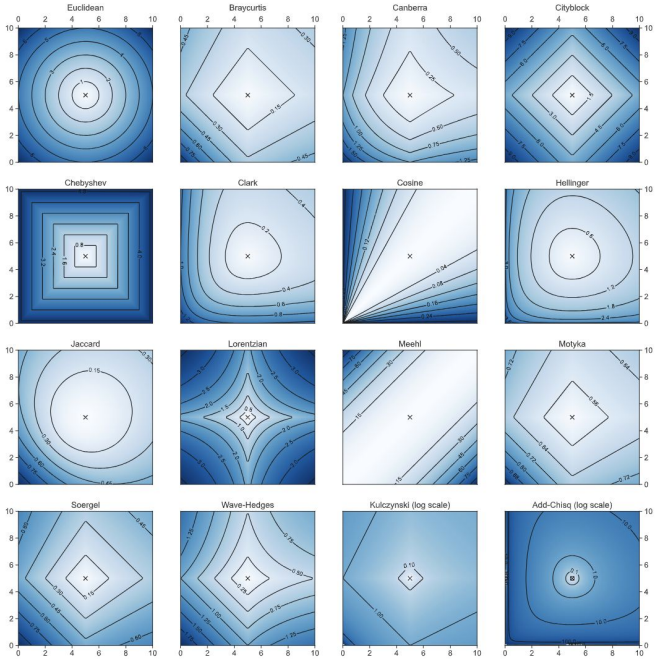
Classifying the Sky with ZTF



Caltech

Ashish Mahabal
Deputy Director, Center for Data Driven Discovery, Caltech

The Promises and Perils of Long Time: Recent Advances in Astronomical Time Series

JSM, Portland, 2024-08-07

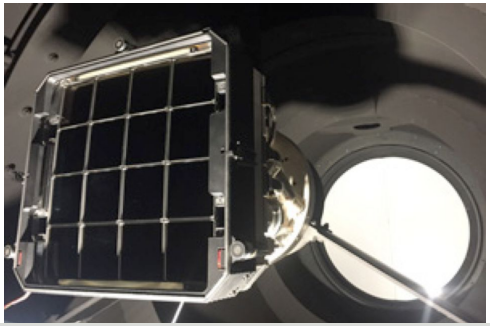
Outline

- About ZTF
- Its transients, variables etc
- Time series - irregular, gappy, heteroskedastic
- Standad technique: get features and classify; use CNNs
- Using XGB and DNN
- Carrying confidances, ambiguities
- Firgate
- Using distance mterics
- RWE mode

- Exploring Foundation Models
- Combining diverse sources (QS)
- Using for future surveys:
  - Roman STRIDE
  - Rubin anomalies
- Anomaly detection
- ZARTH
- Summary
  - Lot being done
  - Archives
  - FMs

Zwicky Transient Facility

Systematic Exploration of the Dynamic Sky

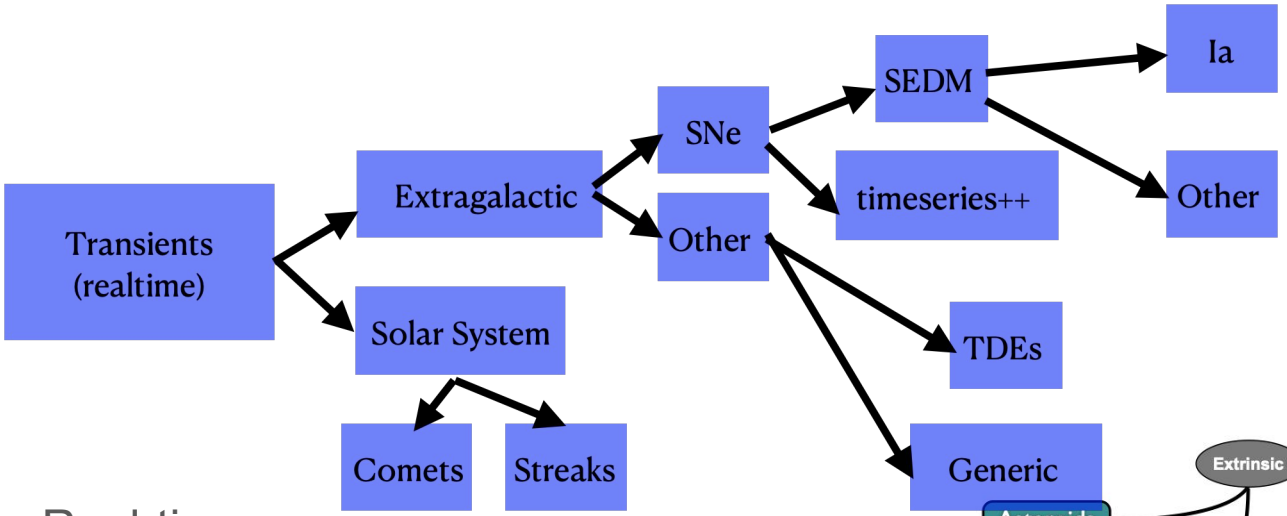


1.2m telescope; 47 sq. deg FOV; cadence: 2-3 nights

Filter(s)	#PSFcat-*sci* sources	#Aperturecat-*sci* sources	#PSFcat-*ref* sources	#Aperturecat-*ref* sources
g	211,215,662,616	134,418,348,720	2,592,206,617	813,670,209
r	575,149,635,926	359,183,821,995	3,442,961,605	1,168,563,737
i	73,590,761,419	42,716,536,579	1,493,676,149	484,736,547

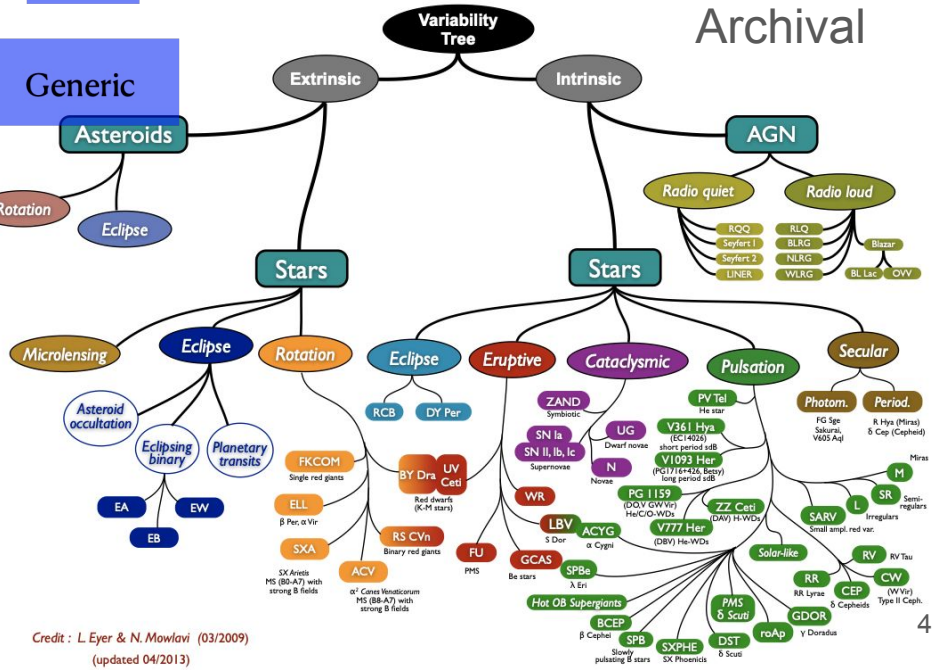
DR21: number of sources (May 2024). Hundreds/thousands points per source.

https://irsa.ipac.caltech.edu/data/ZTF/docs/releases/ztf_release_notes_latest

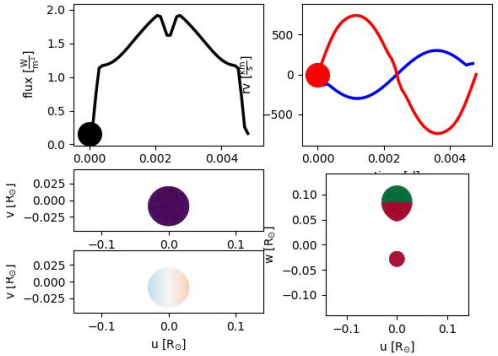


Real-time

Archival

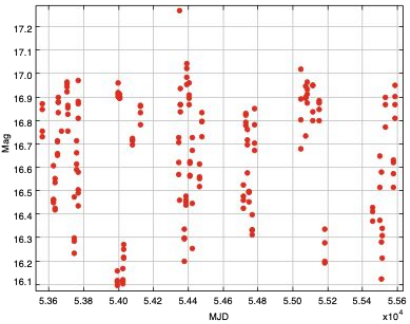


Solar System, Galactic, Extra-galactic science

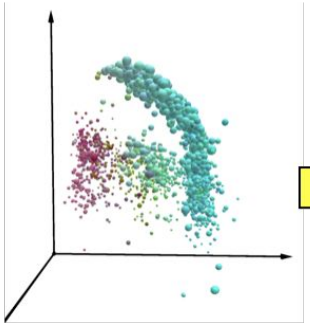


Light curves (time series) are the primary currency

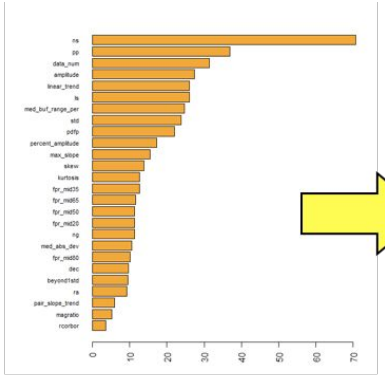
Feature-based



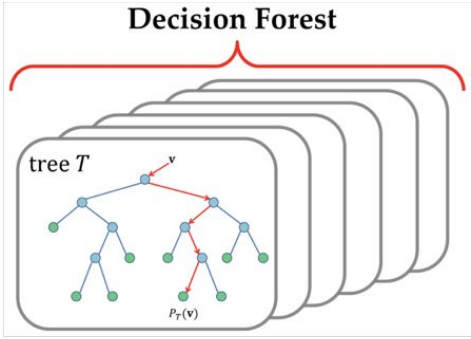
Light curves



Feature vectors



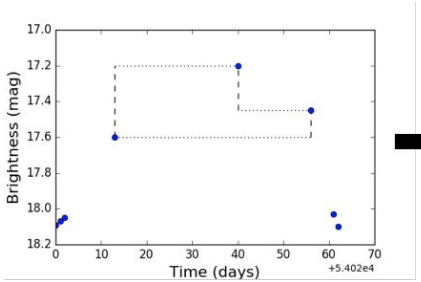
Dimensionality Reduction



Classification

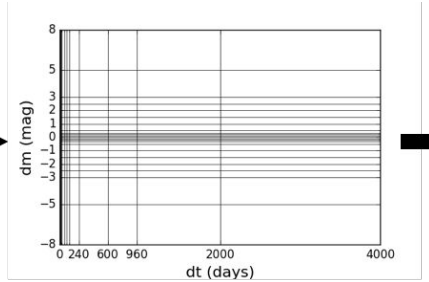
Survey Differences: area, bands, cadence, depth, exposure, ...

Variable points -> normalizing dimensionality



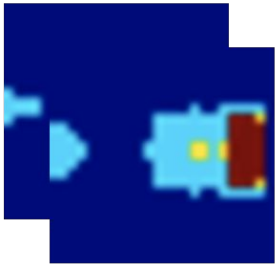
Light curves

n points

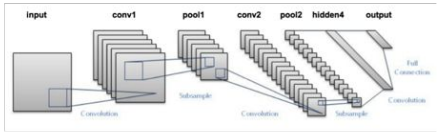


Density representation

n * (n-1)/2 points



Equi-area images



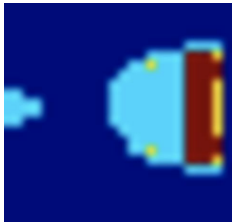
Convolutional Neural Network

Use the CNN hammer

dmdt images

RR Lyrae

RS CVn



LPV

Mahabal, Sheth et al., 1709.06257

Biswas, McIver, Mahabal, arXiv:1910.12143

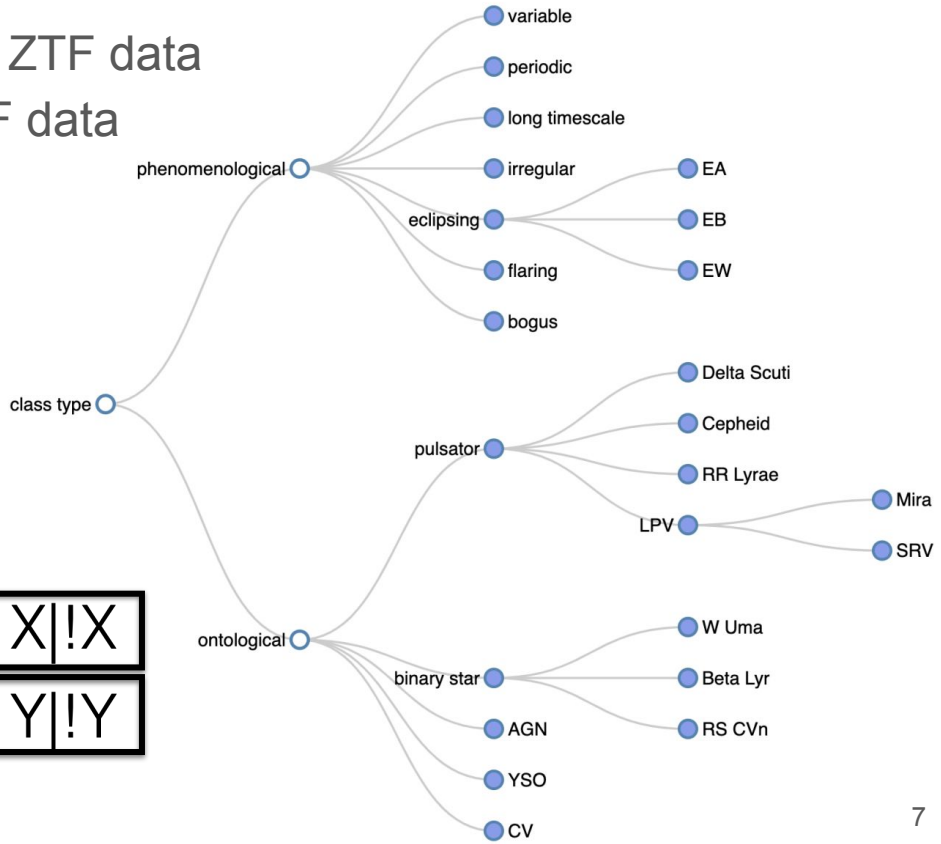
Hierarchical/stackable Classification Through Independent Binary Classifiers

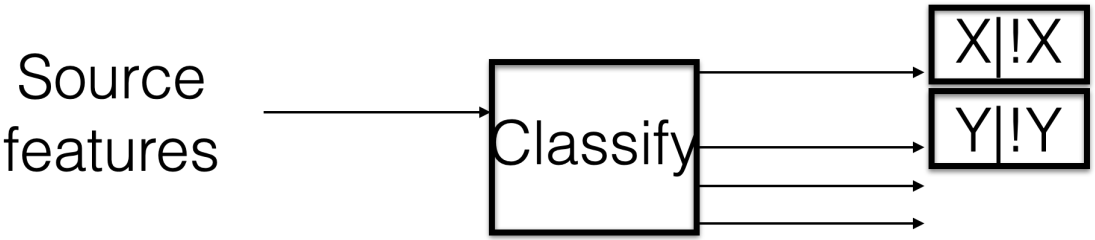
- ***Phenomenological***: based on *just* the ZTF data
- ***Ontological***: based on *not just* the ZTF data

Since the classifiers are independent

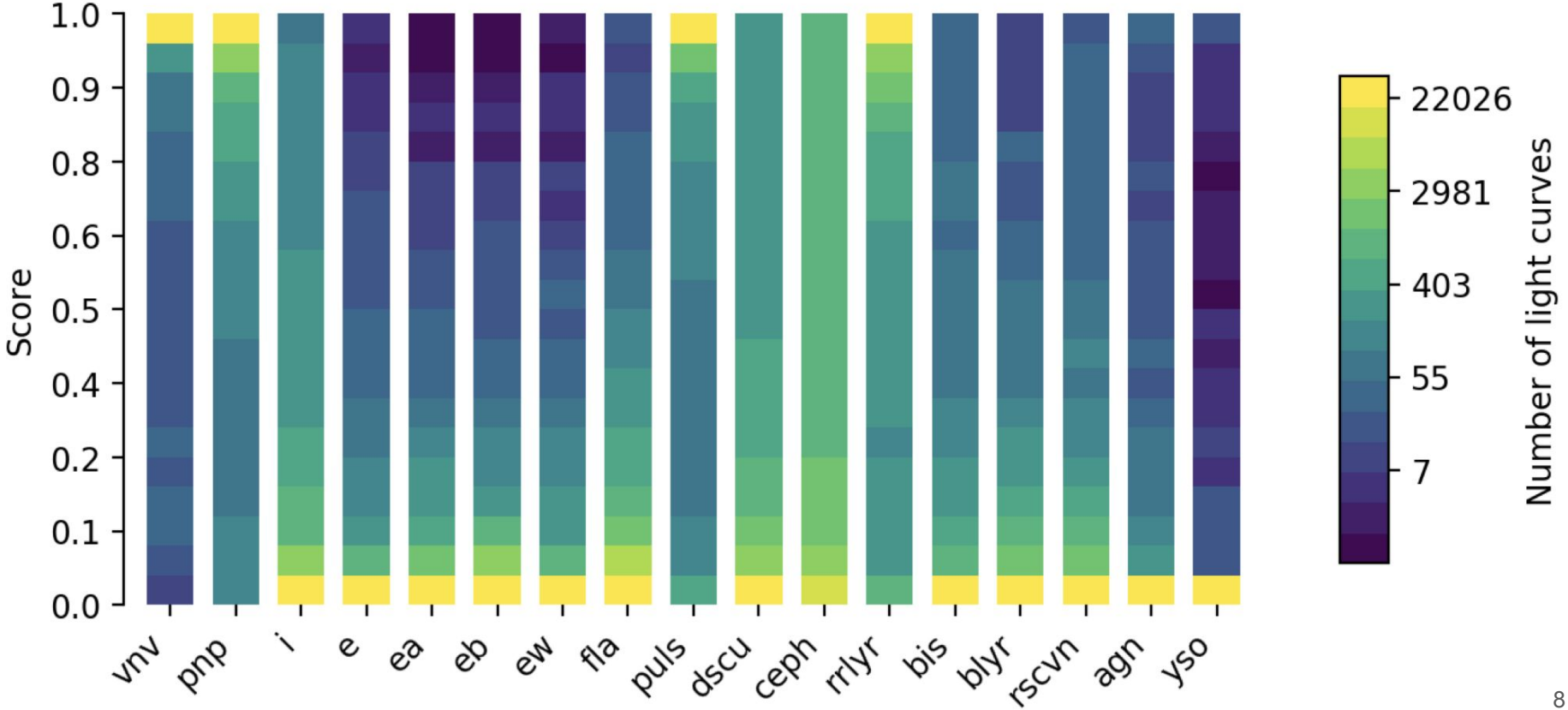
no restrictions on p(sum)

In particular it may not be 1





Using multiple classifiers together



Classification related ZTF work (SCoPe++) …

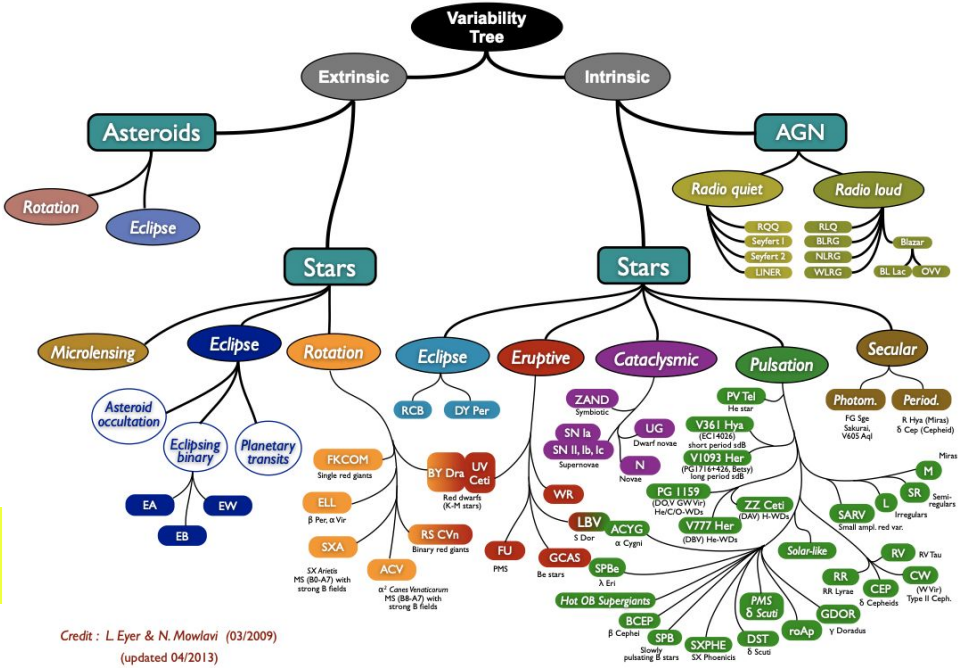
Mahabal et al. 2017 http://arxiv.org/abs/1709.06257v1
Duev et al 2019 http://arxiv.org/abs/1904.05920v2
Duev et al. 2021 https://arxiv.org/abs/2102.13352
Coughlin et al. 2020 https://arxiv.org/abs/2009.14071
Van Roestel et al. 2021 https://arxiv.org/abs/2102.11304
Fremling 2021 https://arxiv.org/abs/2104.12980

https://arxiv.org/abs/2312.00143

Healy et al.

https://zenodo.org/records/11127912

Classifications on 80+ fields

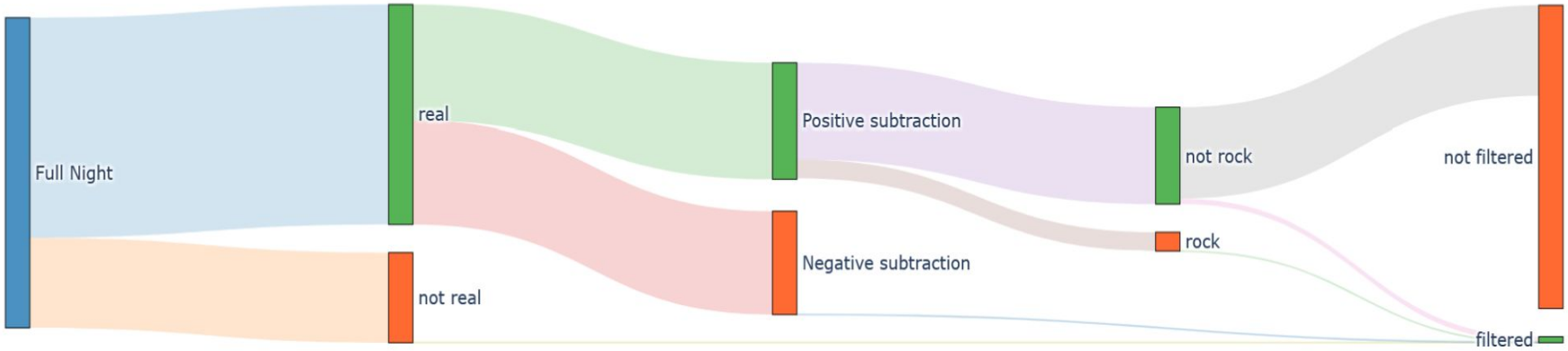


Ashish Mahabal

FRIGATE: Fritz Gap Analysis
What are different groups interested in

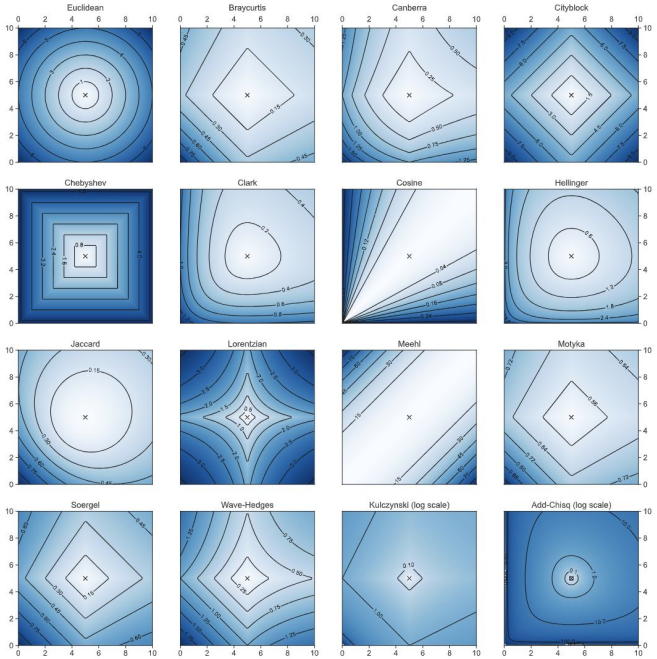
Kira, Theo, Ashish

Filtering of ZTF Data

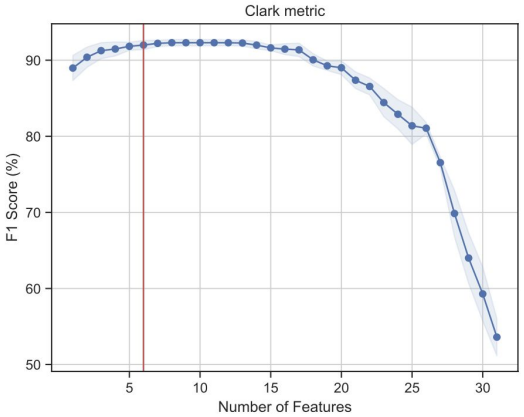


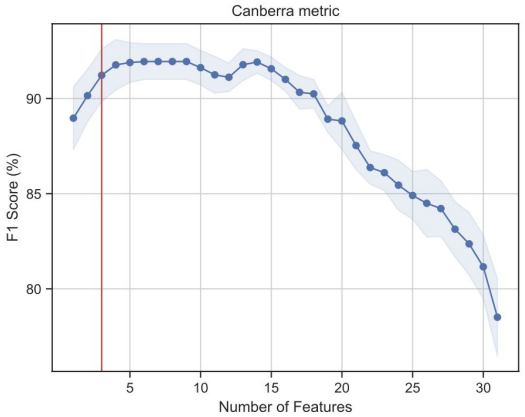
Distance metric based classifier

Chaini, Mahabal, Kembhavi, Bianco



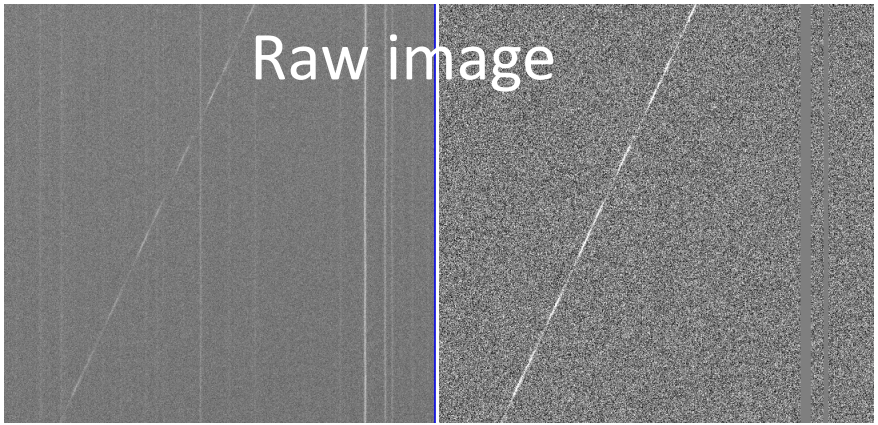
Fewer features required



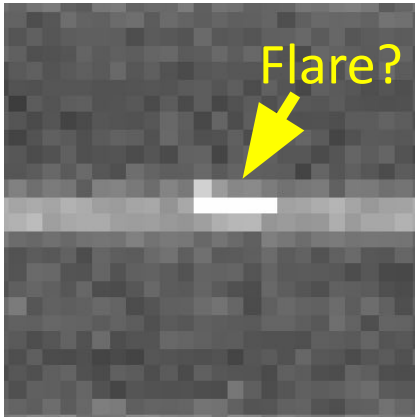


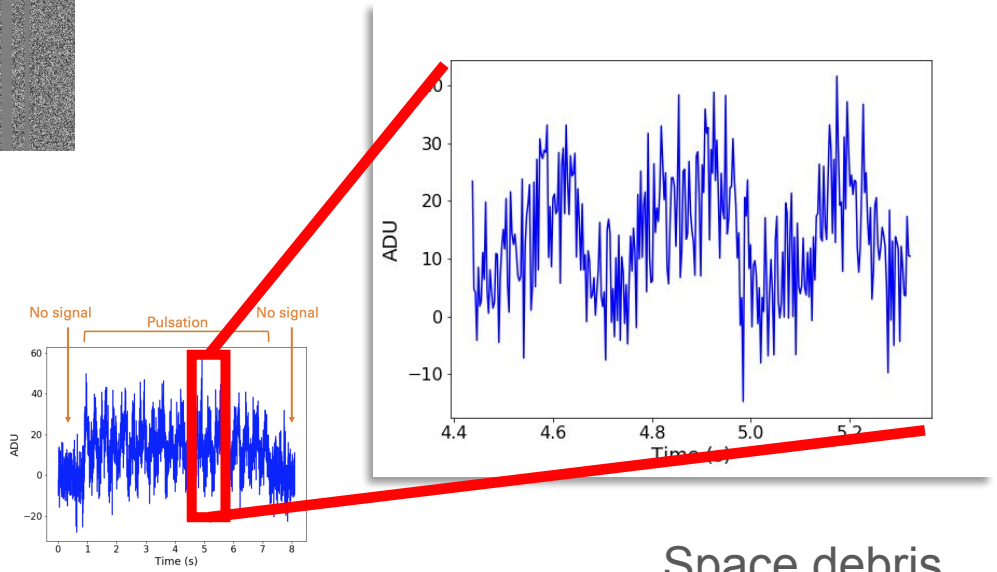
https://arxiv.org/abs/2403.12120

RWE Data Mining: Image Subtraction



Andreoni, Smith,
Mahabal, Graham,
Daniels, Bianco, …





Space debris

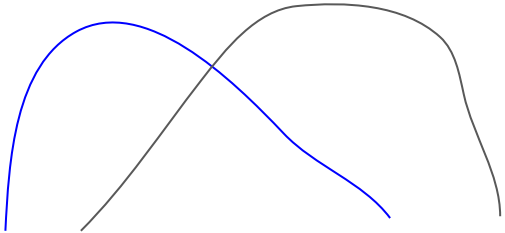
Foundation models - one line primer

It generalizes well and hence has Zero-shot applicability

Examples:

- LLMs (e.g. ChatGPT)
- DALL-E, midjourney, SORA, …
- SAM aka Segment Anything Model

Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).



Foundation models for time series

- Totem from Caltech: https://github.com/SaberaTalukder/TOTEM (uses tokenization)
- Moment: https://github.com/moment-timeseries-foundation-model/moment
- NBeats, DeepAR, Informer, TimesNet (all forecasting)
- TimesFM (Google): https://github.com/google-research/timesfm (forecasting - decoder only)
- Chronos (Amazon): Language model trained on tokens
- TimeGEN-1 (Microsoft Azure): Released two weeks back

The power of archives

DPOSS, Paloma-Quest, PTF, iPTF, ZTF done from the same 1.2m telescope

Provide variability information over several decades

First project I had done was with DPOSS 25 years ago

We now have better capability to explore archives and we should do that

That will help with discoveries from future surveys

Roman/Rubin

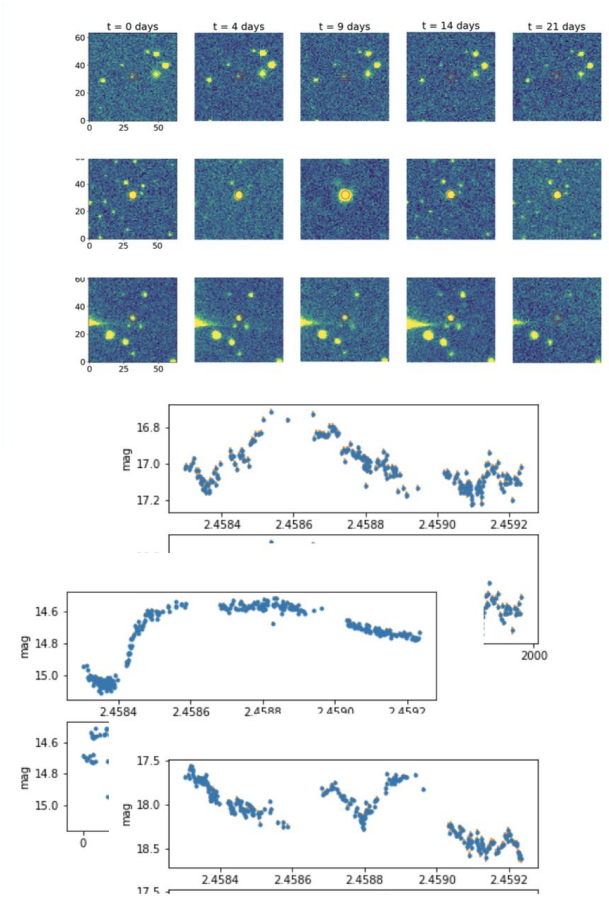
Roman STRIDE: Time domain umbrella group - meets monthly

TVS/ISSC: many initiatives

Anomaly detection

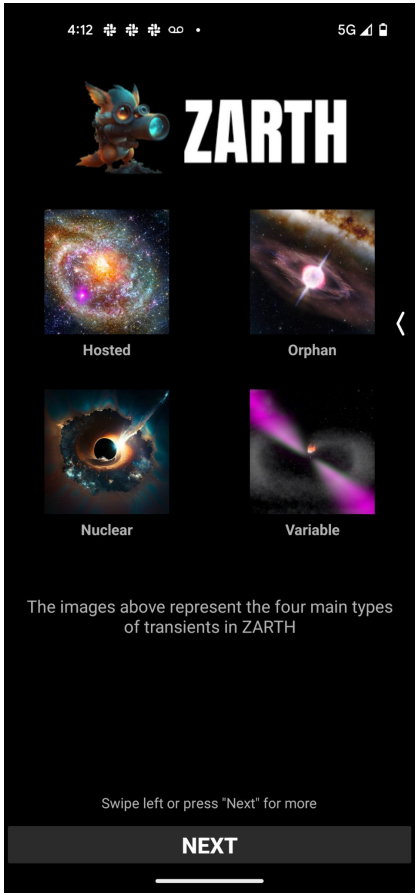
Other surveys like Argus, FAST will make this more fun

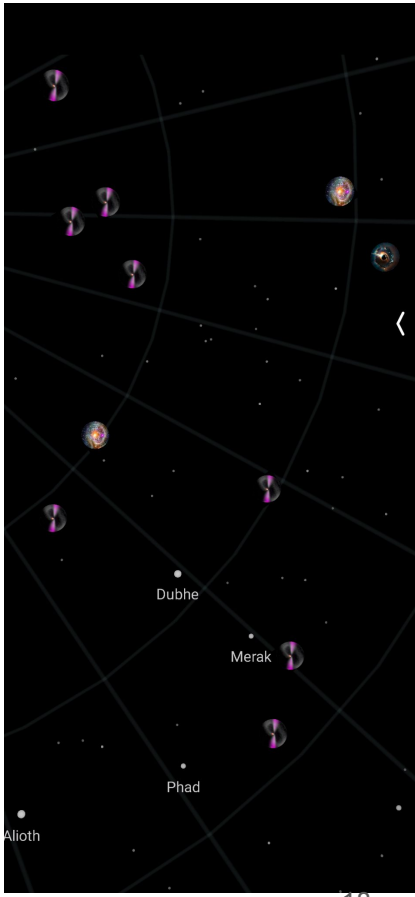
No Anomaly Left Behind



With PPurohit, SParikh, YHassan, T Jegou Du Laz, …

ZARTH - Pokemon GO for ZTF transients





Get from android playstore
Coming to iPhones soon

Ashish Mahabal

With **D Thummar, D Pindawala, A Arora**, T Jegou Du Laz, A Bhavsar, I Kostadinova, Naman Dharmani…

Four primary classes plus two ambiguous classes

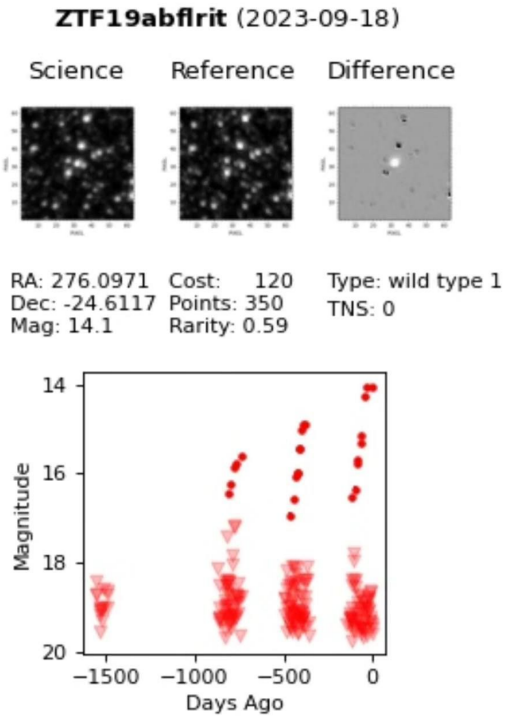
Hosted

Nuclear

Orphan

Variable

20-200 fresh ZTF transients every good night
Many gamification elements
Points for catching
Leaderboards
Streaks
Badges coming soon



Game currently for outreach only for general public
Astronomy students can learn a lot
And also provide feedback

Wild Type 1
Wild Type 2

Ideal to introduce in the classroom

Summary

- Statistical and traditional ML still used extensively in astronomical time series
- Archives underexplored
- Modern ML methods - foundation models based on attention - mostly need regular data, and typically trained for forecasting
- Explainability and interpretibility crucial

ashish at caltech.edu