Speeding-up the search for anomalies in ZTF times series with interpretable machine learning

Emmanuel Gangler
and the SNAD team

Particules pLasmas Univers applicationS
Laboratoire de Physique de Clermont

cnrs

UNIVERSITÉ Clermont Auvergne

Why do we care about anomalies ?

Anomalies are at the source of many serendipitous discoveries
Larger dataset → new discoveries

Some challenges :

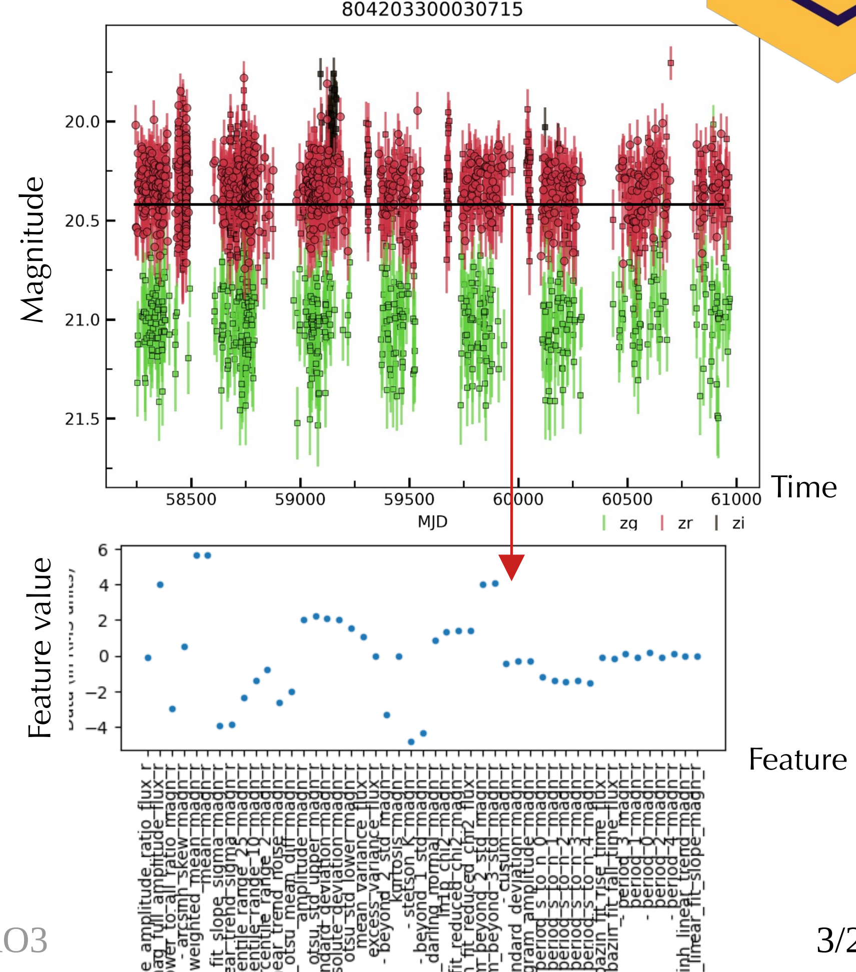
- Normal behavior may be not known
  - Usually in large data sets → **Unsupervised learning, fast algorithms.**
- Interpreting the underlying behavior/properties of anomalous data
  - Field expert knowledge needed (here : mostly astronomy)

How to help the expert ?

- **Find** anomalies
- **Understand** Machine decision
- Find **more interesting** anomalies

ZTF DR23 dataset

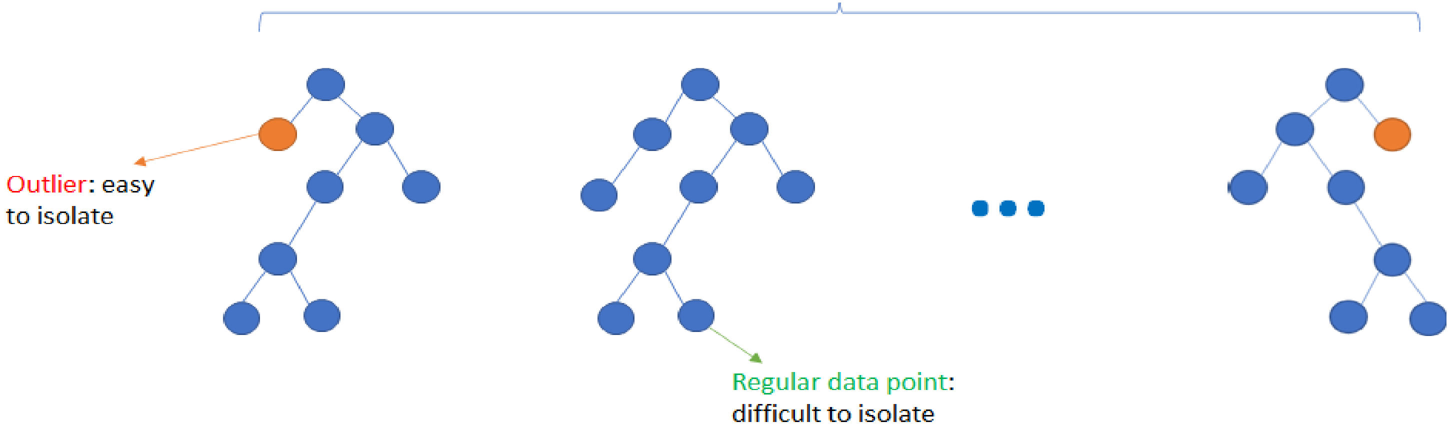
- Selection of **720 Million Light-Curves**
  - ~ **7 years** of observations
  - Northern hemisphere
  - n_obs >100 in r with good quality
  - Feature extraction in r band (Malanchev 2021)
    → 47 **features**
  - **130 Gbytes**
- Use of the SNAD viewer to scrutinize individual objects (Malanchev 2023)



Isolation Forest

Explainable by desgin

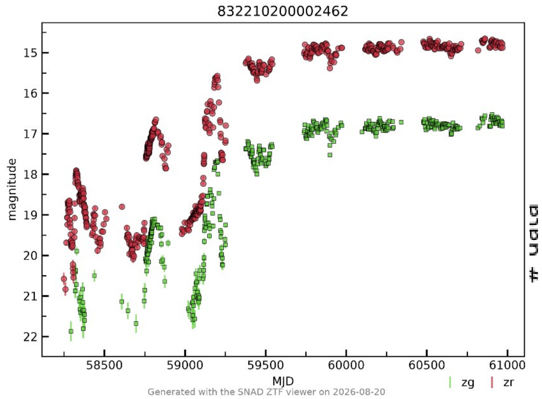
- **Random Tree:**
  - Select a feature + a threshold randomly
  - Partition the sample and repeat until only 1 sample left in the partition
  - Shallow in the tree means outlier
- **Random Forest:** **Average** on T trees (proxy for local density)

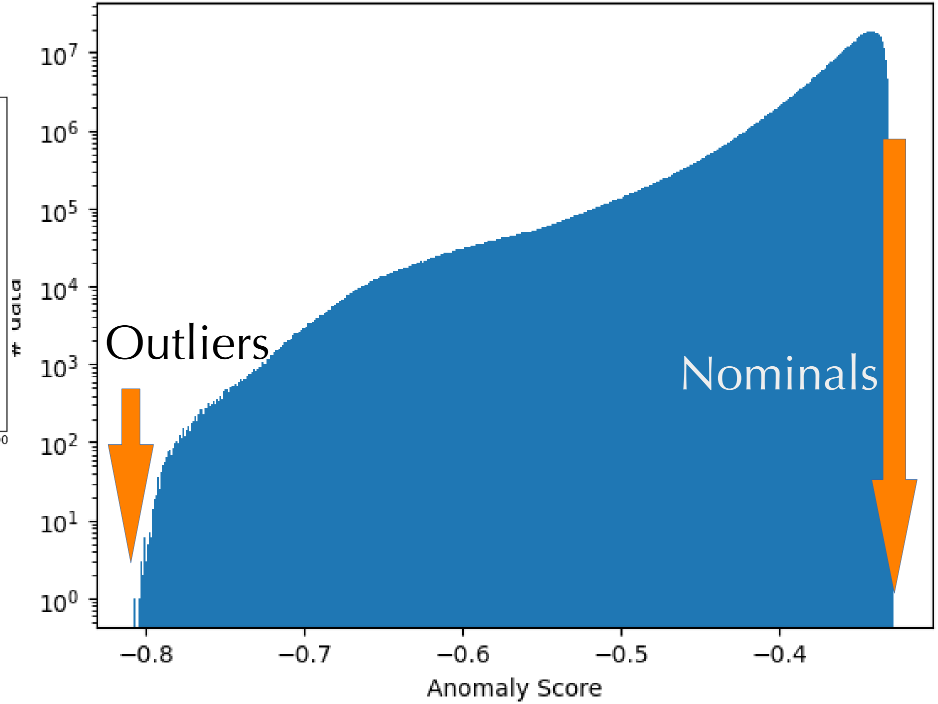


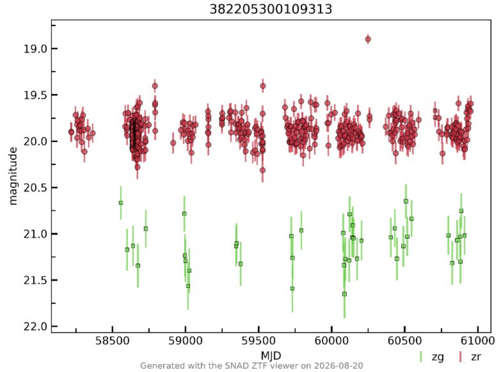
Isolation Forest for ZTF DR23

- Isolation Forest run through Coniferest Package (Kornilov 2025)

Training time : 10 min
Scoring time : 10 min
64 CPU, 256 Gb RAM







→ **YSO**

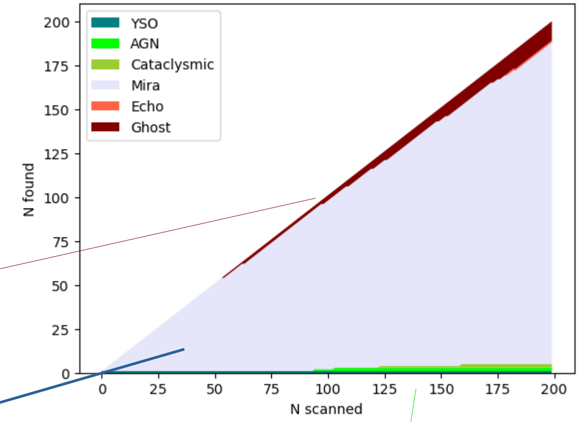
~8000 candidates in ZTF (Zhang2023)

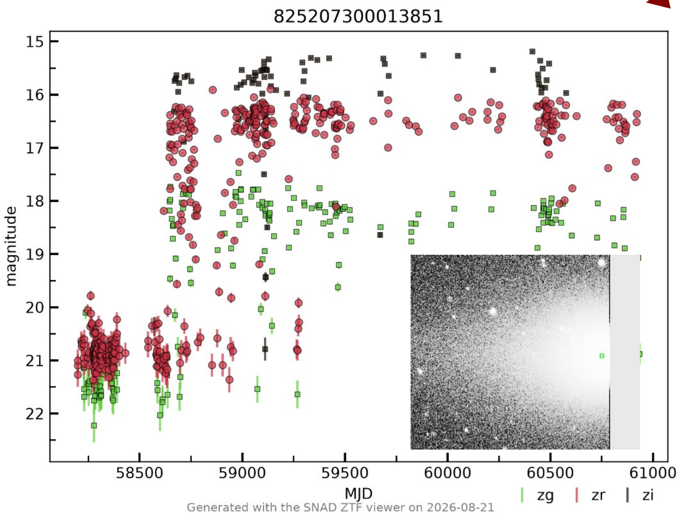
→ A random star

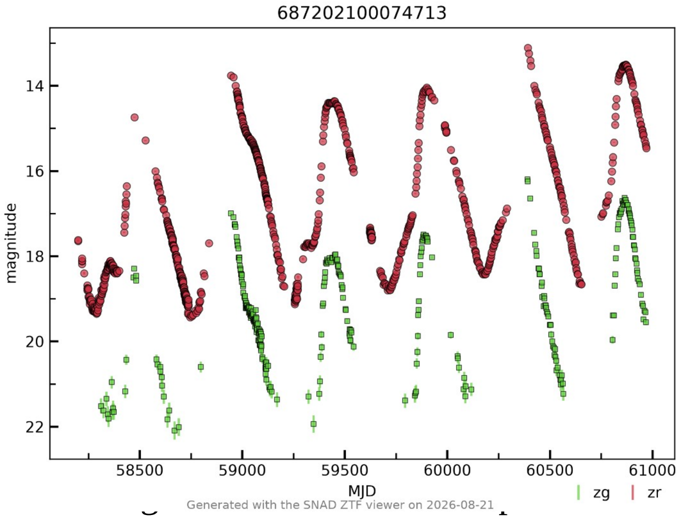
Census of outliers…

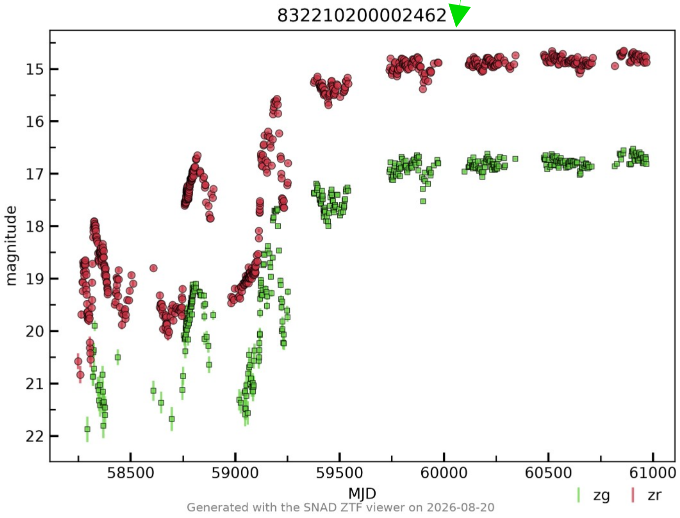
- **Scanning takes human (expert) time !**
  - Budget of 200 objects
  - Most of them are Miras

    → **Lack of diversity !**





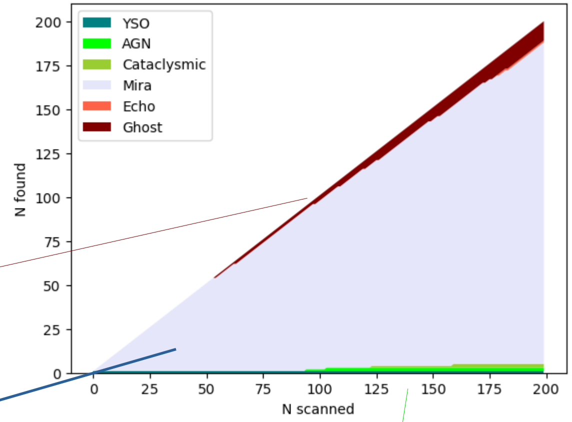


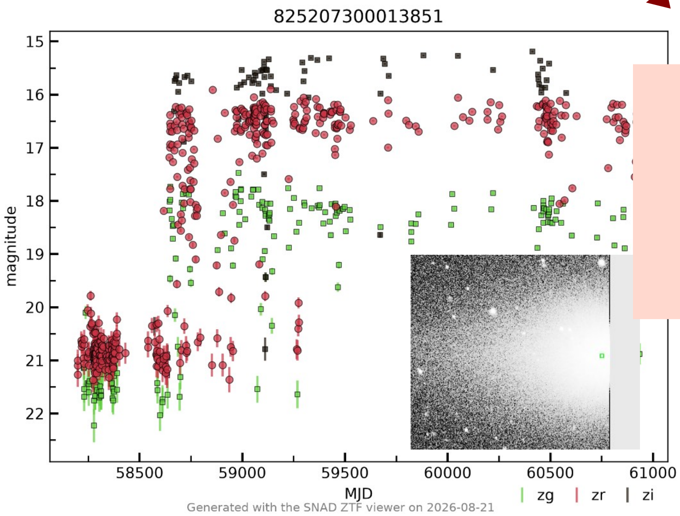


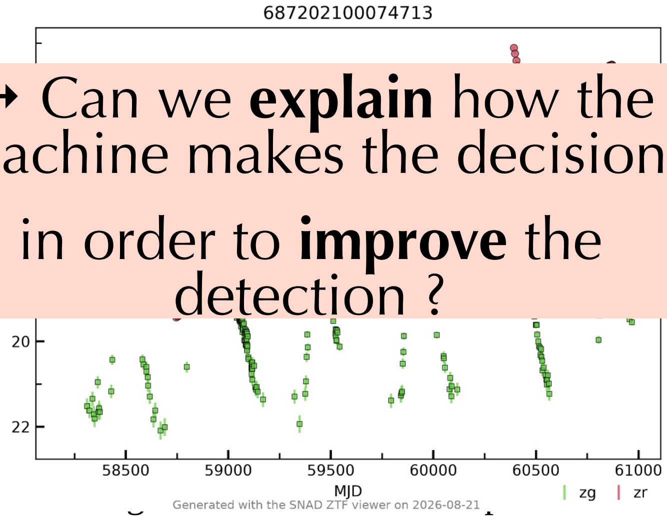
Census of outliers...

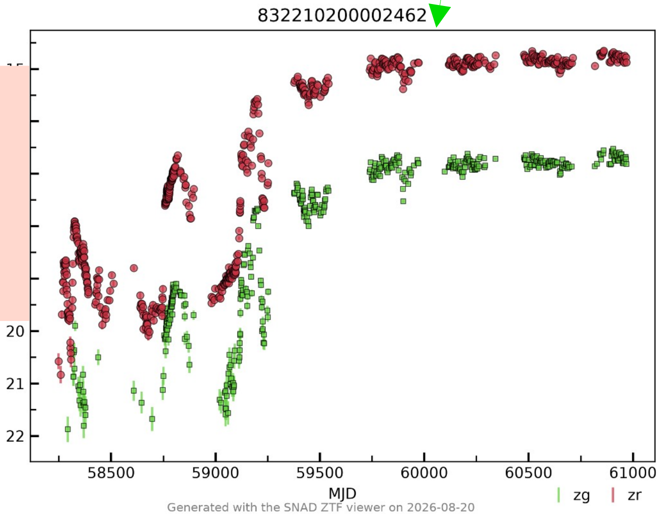
- **Scanning takes human (expert) time !**
  - Budget of 200 objects
  - Most of them are Miras

    → **Lack of diversity !**









→ Can we **explain** how the machine makes the decision

in order to **improve** the detection ?

Explainable machine learning

A system that produces **details** or **reasons**
To makes its functioning **clear** or **easy to understand**
for humans (Barredo Arrieta 2020)

In practice:
feature importance

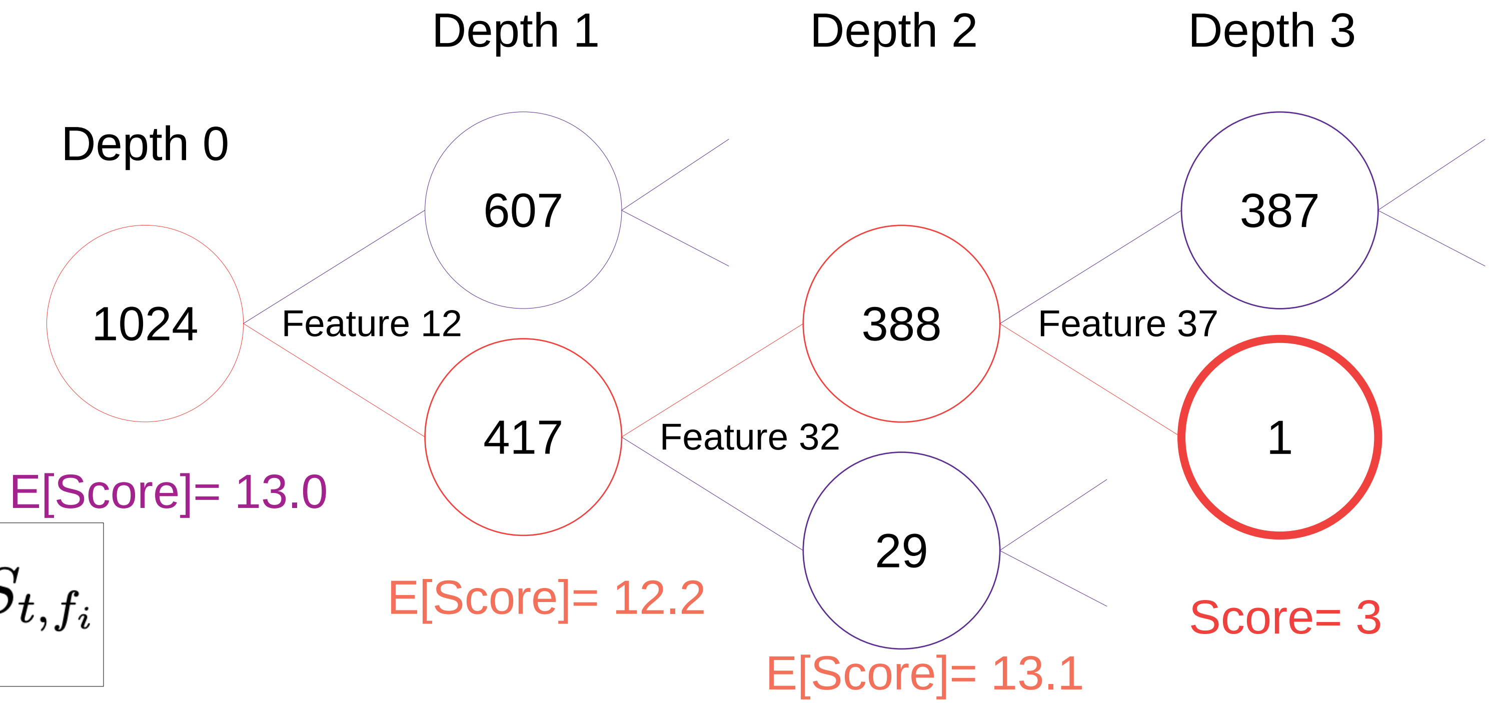
- Popular choices:
  - **Model-agnostic** methods
    - Probe the model from outside
      - *Computationally costly*
      - Prone to instability
    - **SHAP**
      - *Global measure*
      - *Scales poorly with features*
    - **LIME**
      - Local measure
      - Surrogate model
      - *Depends on hyperparameters*

- What we would like:
  - *Simple and fast*
    - Fast and accurate
    - No hyperparameters
    - Sub-Linear in features
  - Local measure
    - ***Explaning individual anomalies***
  - Surrogate model
    - Approximates well Isolation Forest scoring for ALL data

  → **Model-specific method**

The Signature method E.G. 2024

- What does a given feature brings to the score ?



$$S_{f_i} = \frac{1}{n_i} \sum_t \delta S_{t,f_i}$$

→ Average on all trees

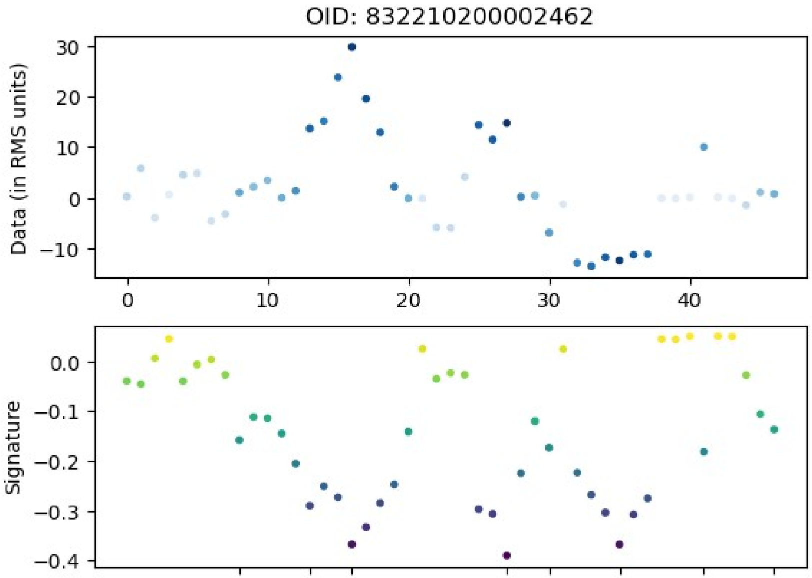
Feature 12 = - 0.8 Feature 32 = + 0.9 **Feature 37 = - 10.1**

Signature, Visualization and Interpretation for ZTF

Training time : 10 min
Scoring time : 10 min
Signature time : 30 min

Data and Signature

For Top-1 outlier (YSO)



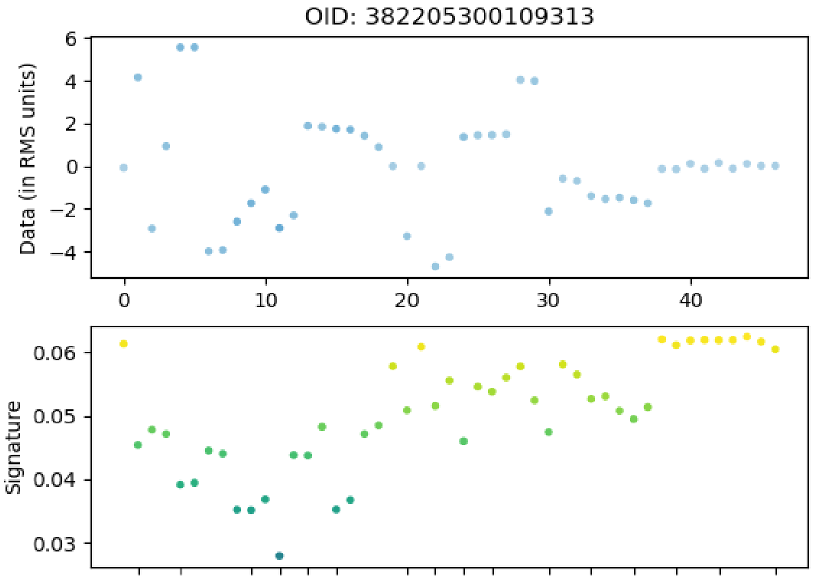
Negative values
Outlier ←

Distribution params.

χ2 Params.

Period parameters

For top-1 Nominal (star)



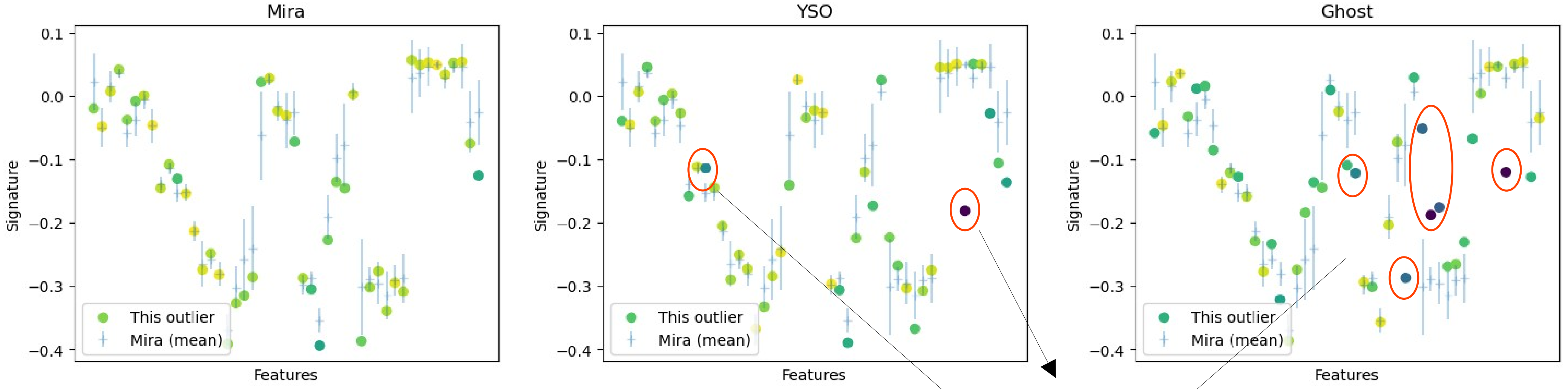
Positive values
→ Nominals

Features

Signatures are specific to outlier classes

Signatures compared to Mira average signature

Top-1 mira | YSO | Top-1 Artefact



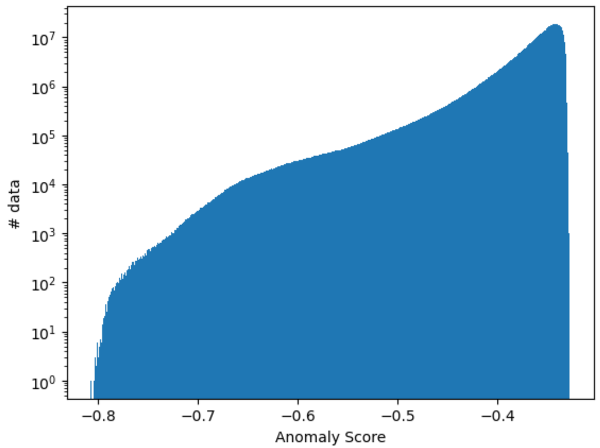
Object follows average pattern in signature space

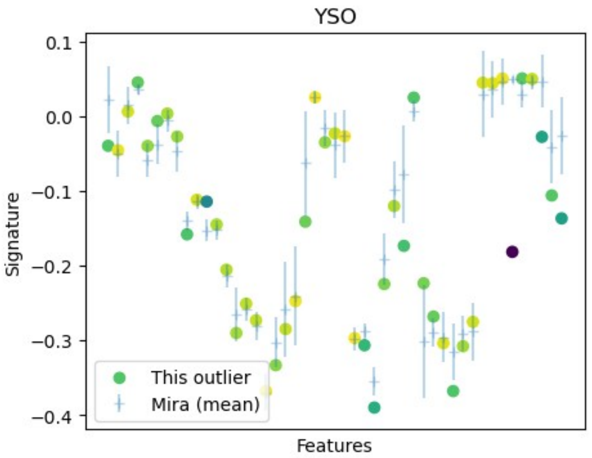
Distinctive pattern in signature space

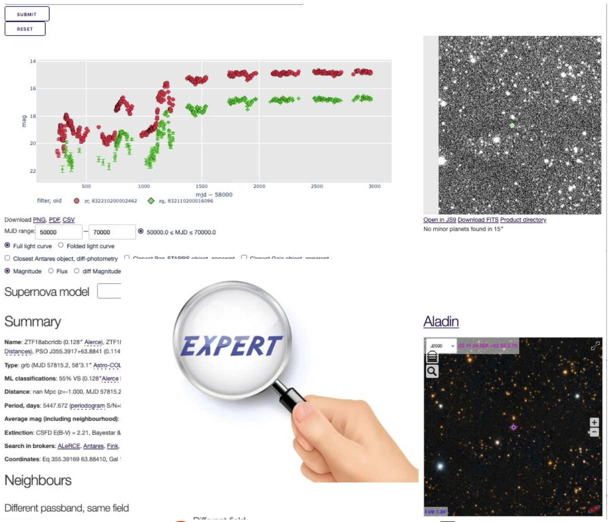
Can we leverage signatures to find more interesting anomalies ?

→ Active Learning

Machine would learn better if a few labels were known
→ iterative knowledge improvement
With **Human-in-the-loop**







Step 1:
Train model

Step 2:
Select best outlier

Step 3:
Ask Expert

□ interesting
□ not interesting

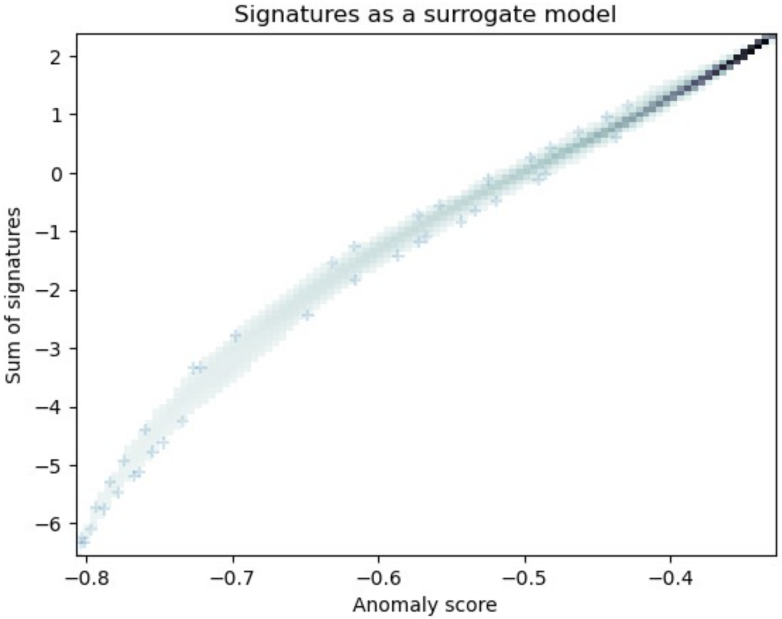
Step 4:
Update labels

Feature selection

Example task:
What is the best feature set to find
Non-Miras, non-artefacts ?

Procedure:
- Let redefine the anomaly score as $Sc_i = \sum_f \epsilon_f \, s_{fi}$
  $\epsilon_f \in \{0, 1\}$

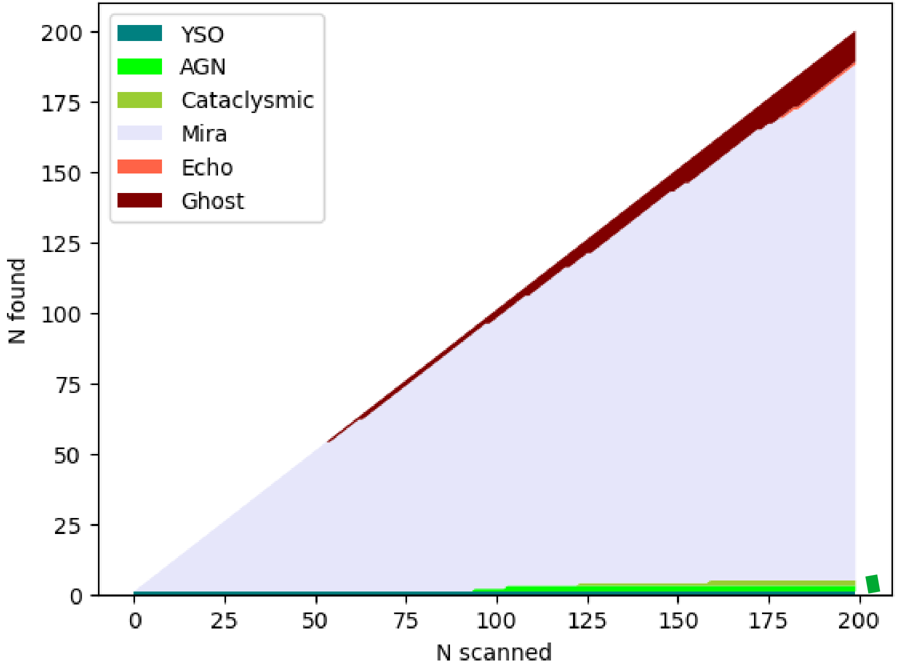
- Turn on/off features to minimize loss function



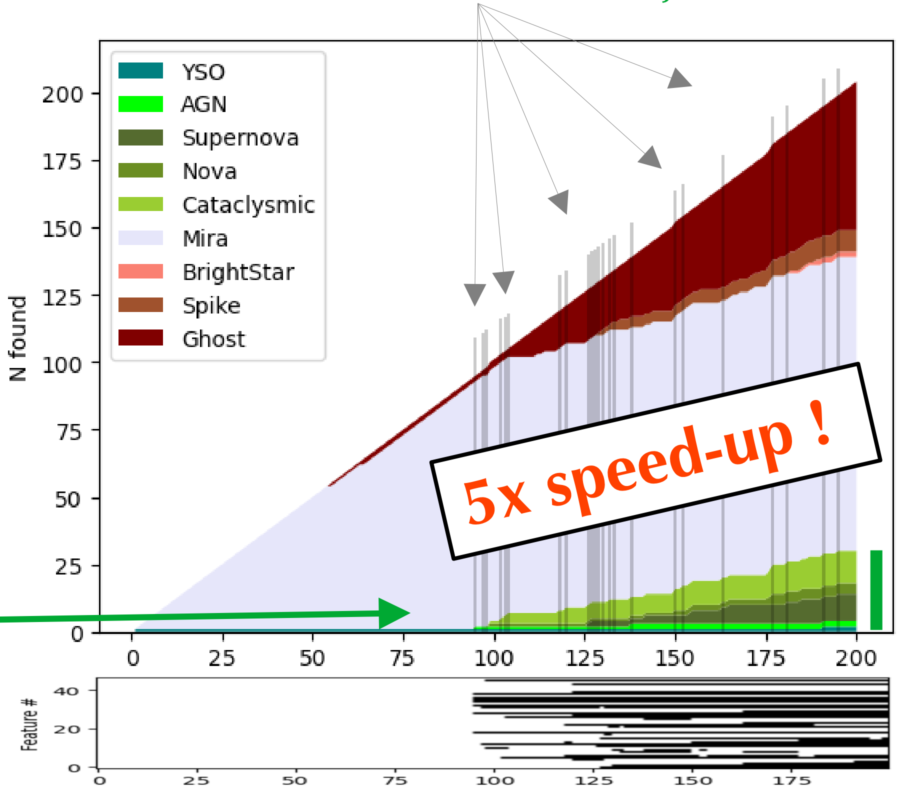
Feature Selection

Training time : 10 min
Scoring time : 10 min
Signature time : 30 min
Retraining time : 1 min

Passive: 5 objects

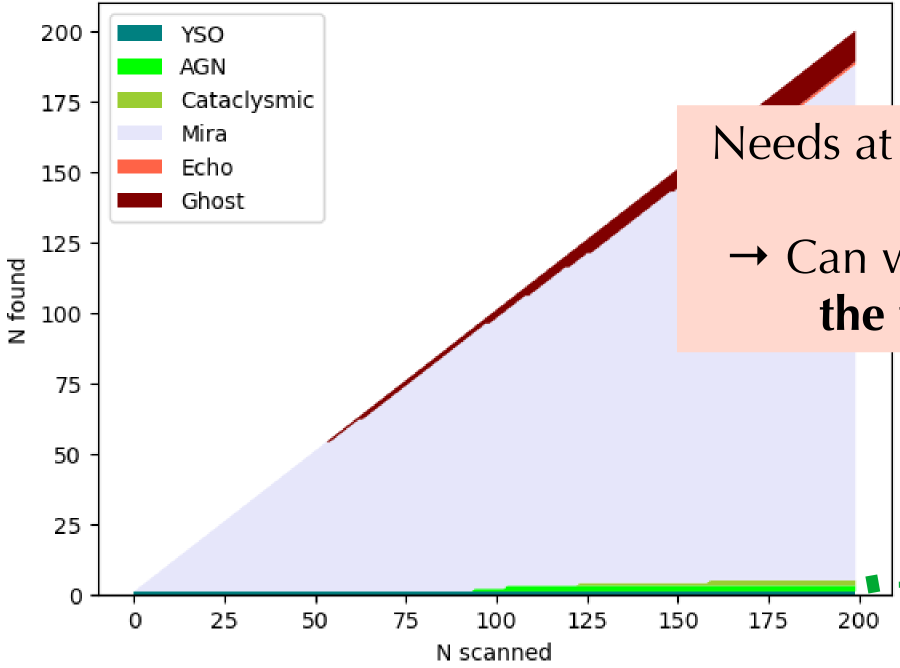


Active : 26 objects

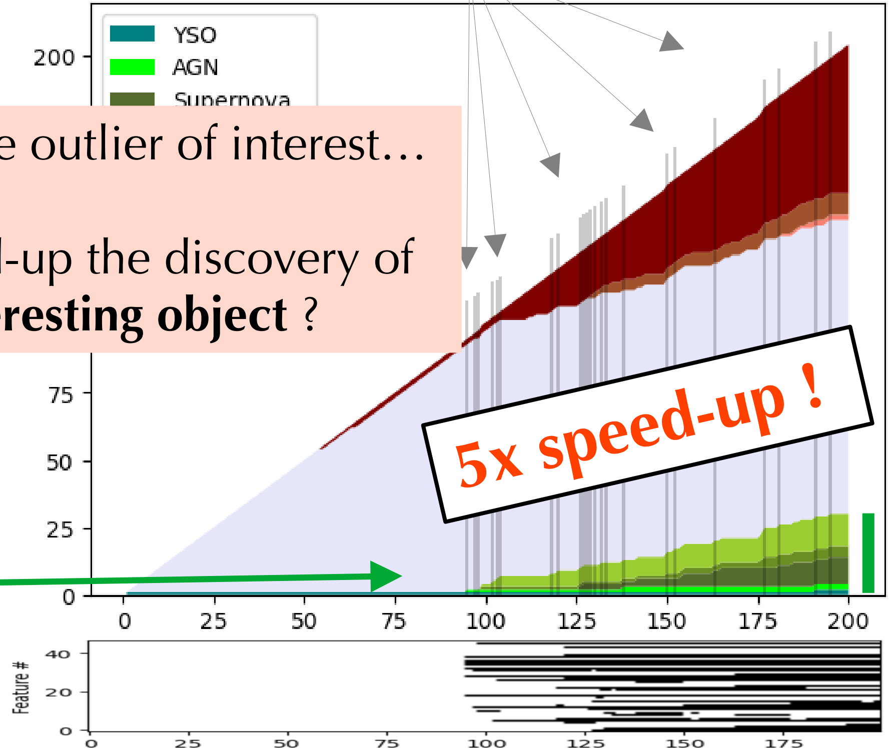


Feature Selection

Passive: 5 objects



Active : 26 objects



Needs at least one outlier of interest…

→ Can we speed-up the discovery of **the first interesting object** ?

5x speed-up !

Diversity search

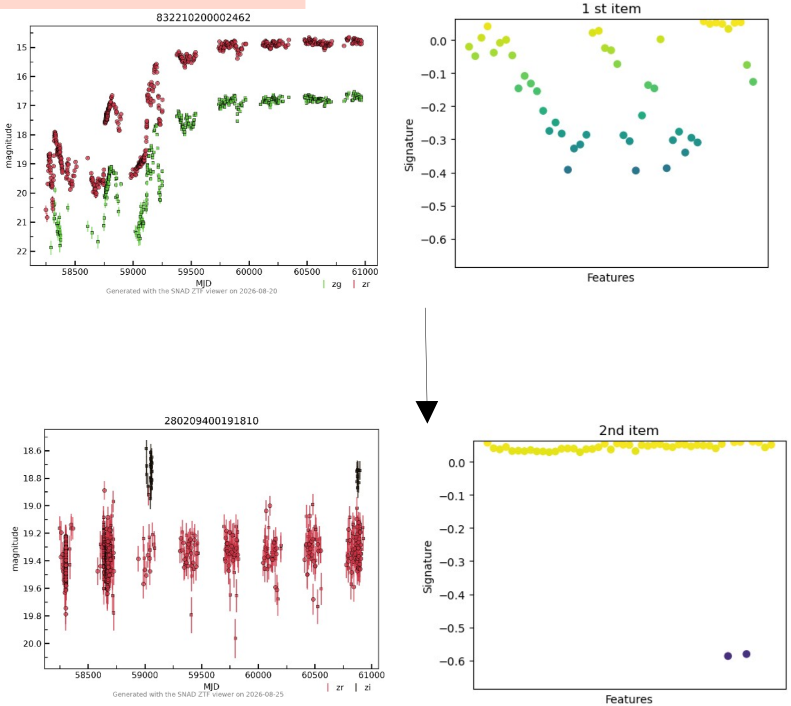
Searching for **unknown unknowns** !

Training time : 10 min
Scoring time : 10 min
Signature time : 30 min
Iteration time : 1 min

- Leverage distances in signature space

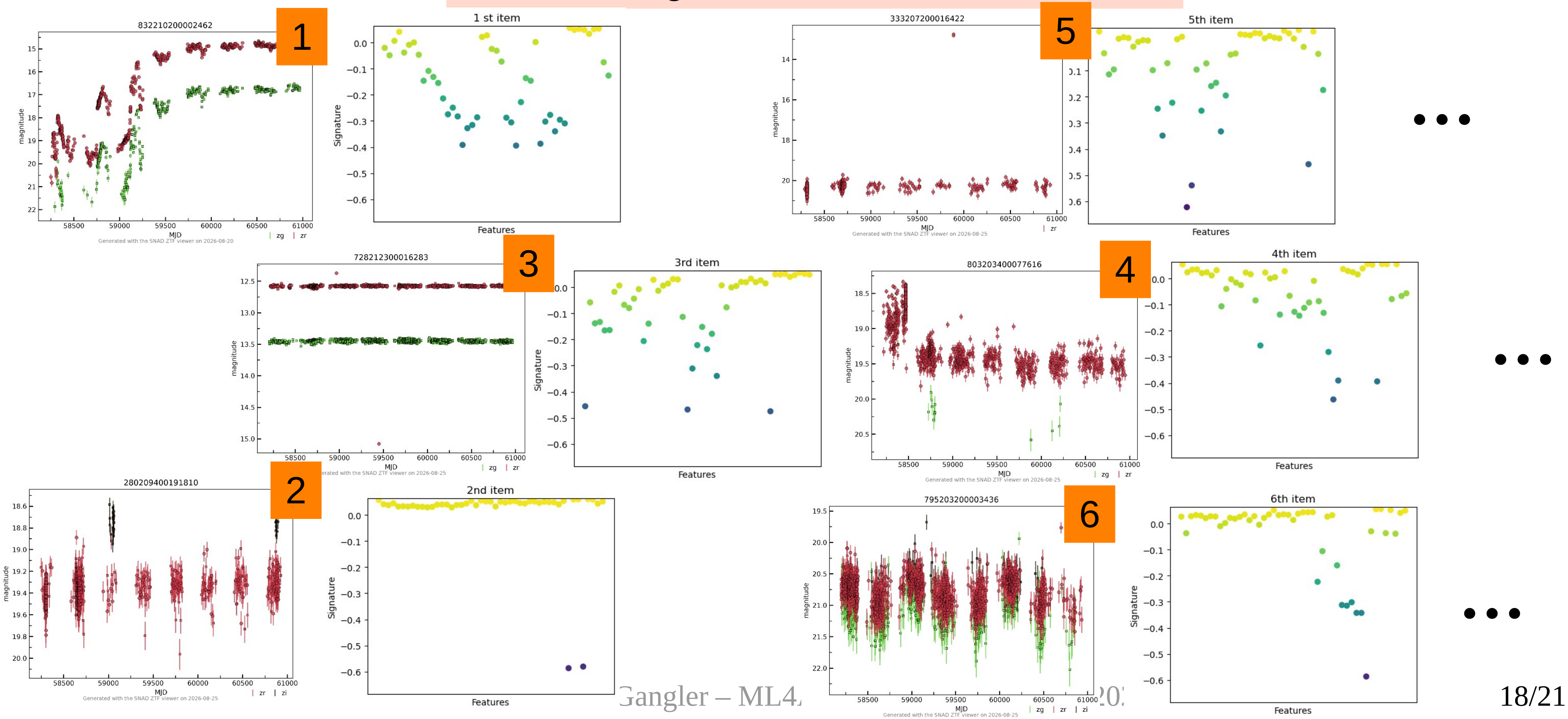
Procedure:

- Compute distances to labeled objects
- Select the object farthest from labelled objects
- add to labeled list and iterate



Diversity search

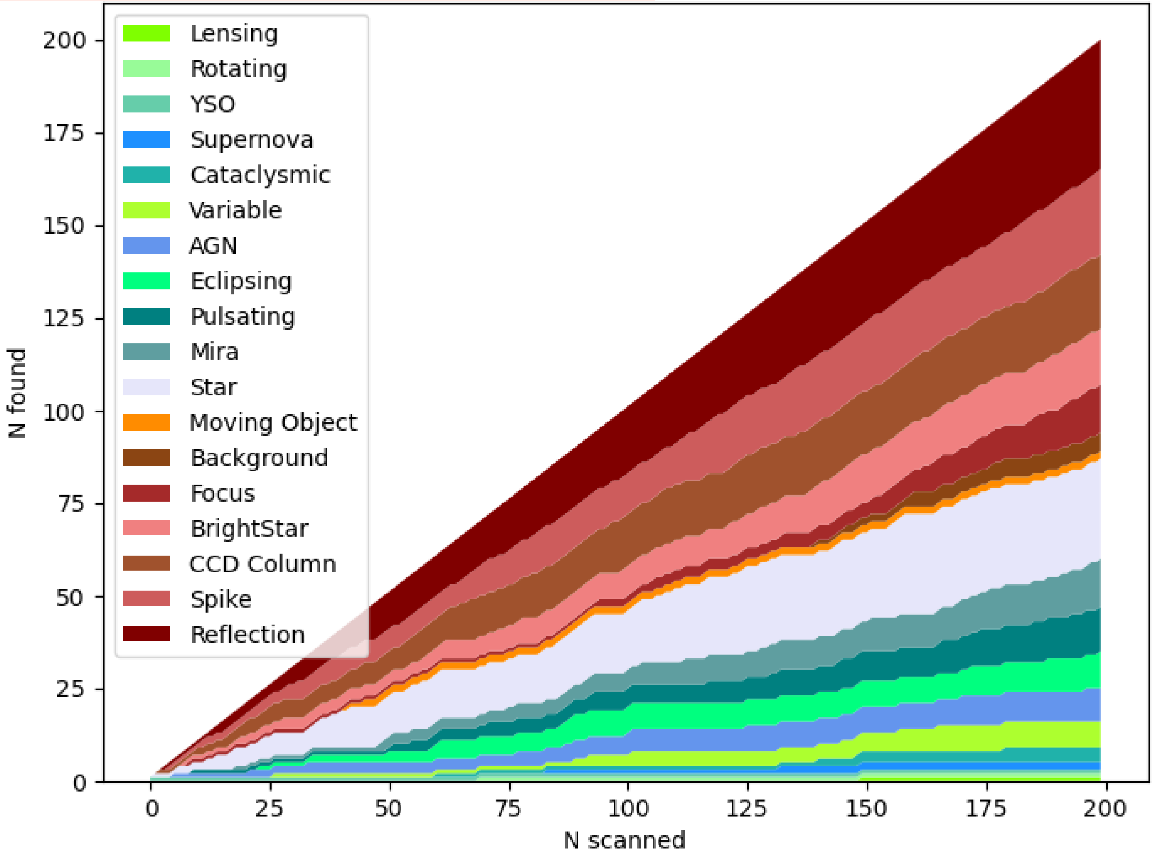
Searching for unknown unknowns !



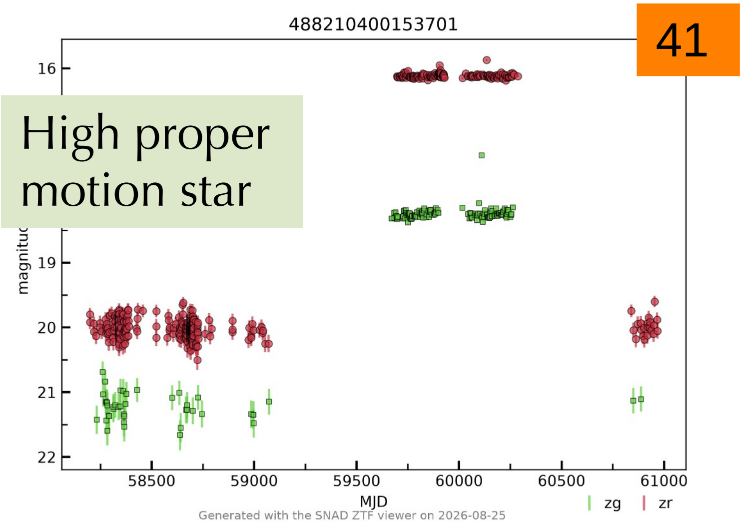
Diversity search

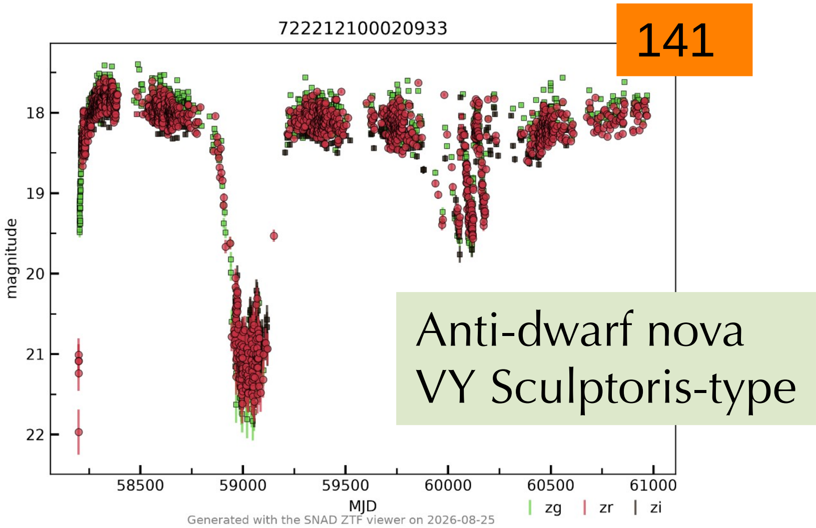
Searching for unknown unknowns !

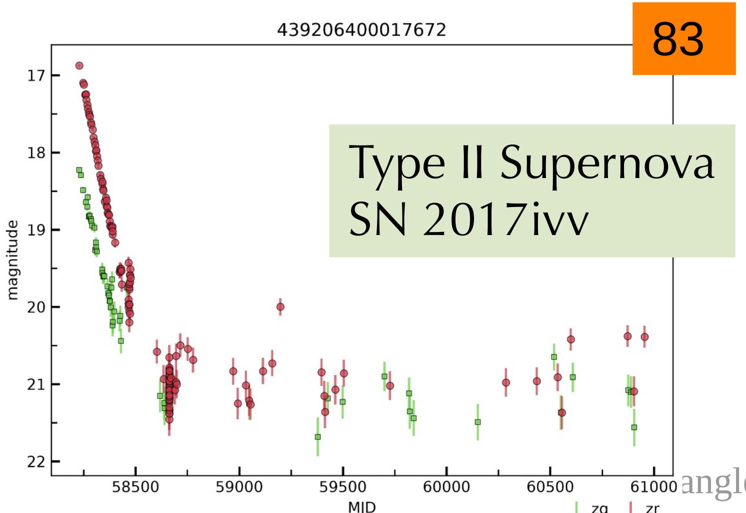
- Results : (after 200 scans)
  - Increased diversity
  - 55 % of volume is occupied by artefacts!
  - Rare classes present

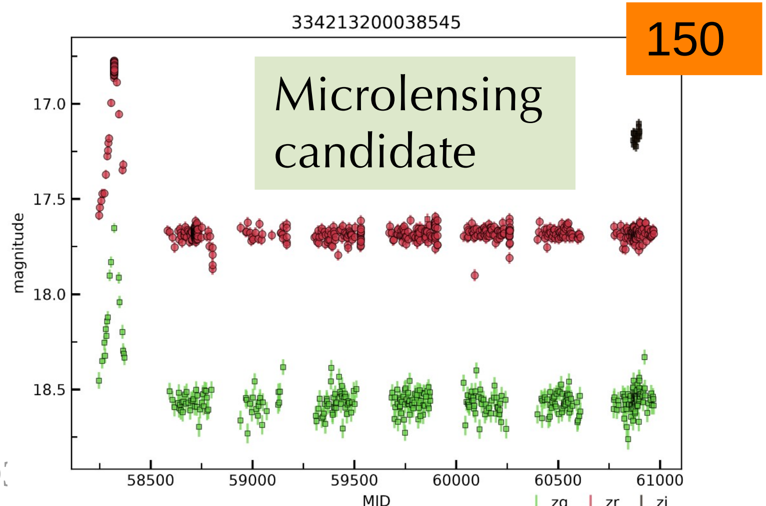


Some rare events :









Take-aways

Anomalies are not unique…
They depend on expert ineterst ← Humans are involved

- **Finding outliers** is not the bottleneck
- **Lack of labels** → ***how to minimize expert time in labeling data ?***
- **Interpretability** is key to bridge AI and human expertise

- **Anomaly signature** is a metric for **feature importance** tuned for anomalies
  - ***Diversity of use cases :***
    - → Visualization
    - → Feature selection
    - → **Fast** Data mining
  - **Model-specific, Domain agnostic**
    - Works for ANY numerical tabular data

      → **and YOU ?**

Stay tuned on

SNAD

https://snad.space

https://github.com/snad-space/coniferest