# Looking for anomalous variability: wandering unsupervised

Alex Razim, Ruder Bošković Institut, Zagreb, Croatia
NOIRLab Rare Gems conference, Tucson, USA, 22.05.2024

# Intro slides that got chopped and recycled

| **2000: ROTSE, ASAS** | **2025: ZTF, LSST** |
|---|---|
| ~$10^3$ - $10^8$ of LCs | ~$10^{10}$ of LCs |
| 100-500 obs' per object | up to 1000 obs' per object |
| single-bands, $m_V \lesssim 15$ | (u)griz(y) bands, $m_r \lesssim 24$ |

**Supervised vs. Unsupervised vs. Semi-supervised ML**

- supervised: train the algorithm to map observed features to pre-determined labels
- unsupervised: determine internal topology of the dataset in a feature space, no labels used
- semi-supervised: use partially labelled dataset to study the topology of the dataset better

**Common applications for ML in astronomy**

- star-galaxy classification (Odewahn et al. 1992, Bertin & Arnouts 1996)
- morphological galaxy classification (Storrie-Lombardi et al. 1992, Dieleman et al. 2015)
- photo-z (Firth et al. 2003)
- spectra classification (von Hippel et al. 1994, Folkes et al. 1996)
- solar activity prediction (Lundstedt & Wintoft 1994)

# Intro slides that got chopped and recycled

| **2000: ROTSE, ASAS** | **2025: ZTF, LSST** |
|---|---|
| ~$10^3$ - $10^8$ of LCs | ~$10^{10}$ of LCs |
| 100-500 obs' per object | up to 1000 obs' per object |
| single-bands, $m_V \lesssim 15$ | (u)griz(y) bands, $m_r \lesssim 24$ |

**Supervised vs. Unsupervised vs. Semi-supervised ML**

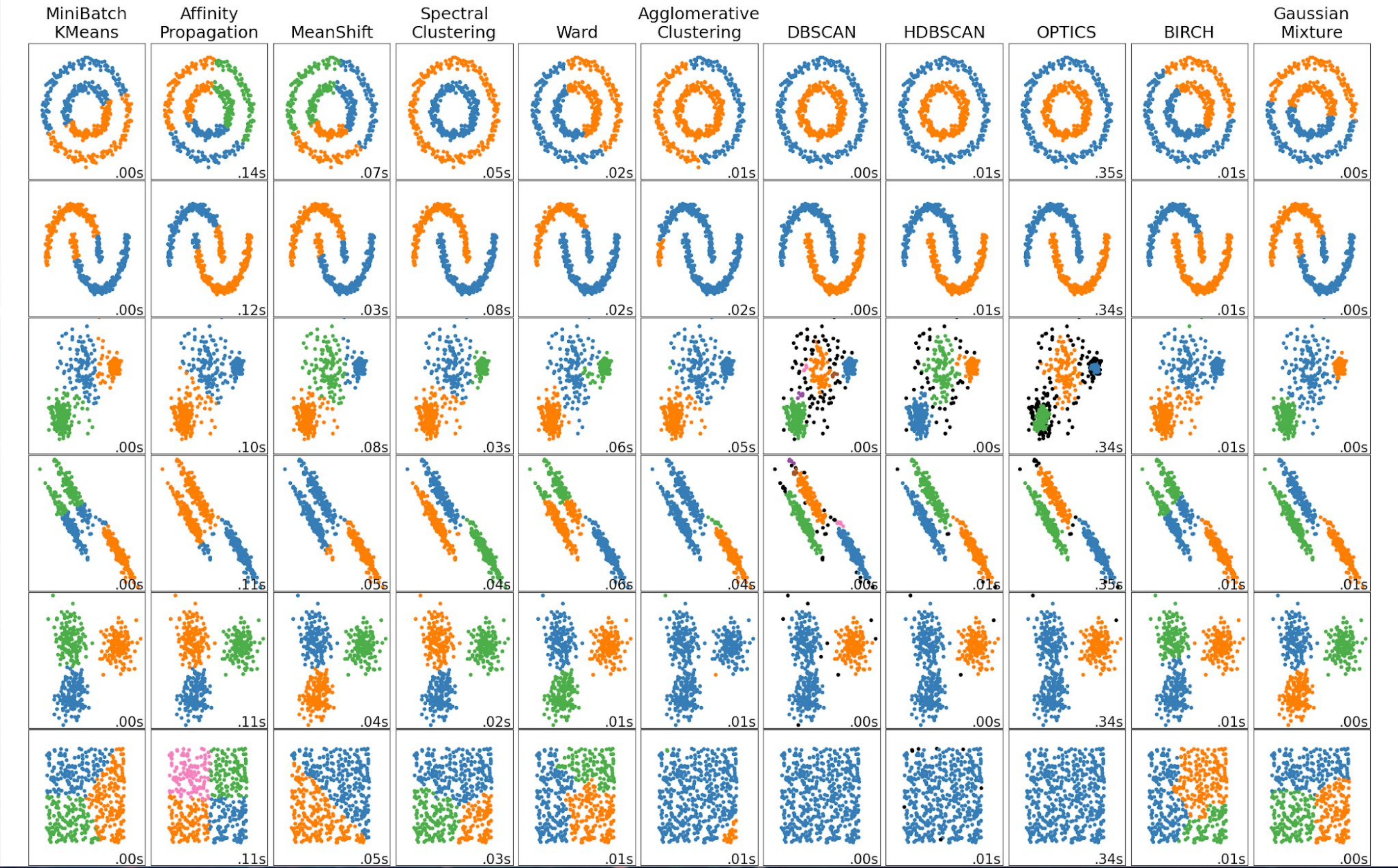
- supervised: train the algorithm to map observed features to pre-determined labels
- unsupervised: determine internal topology of the dataset in a feature space, no labels used
- semi-supervised: use partially labelled dataset to study the topology of the dataset better

**Common applications for ML in astronomy**

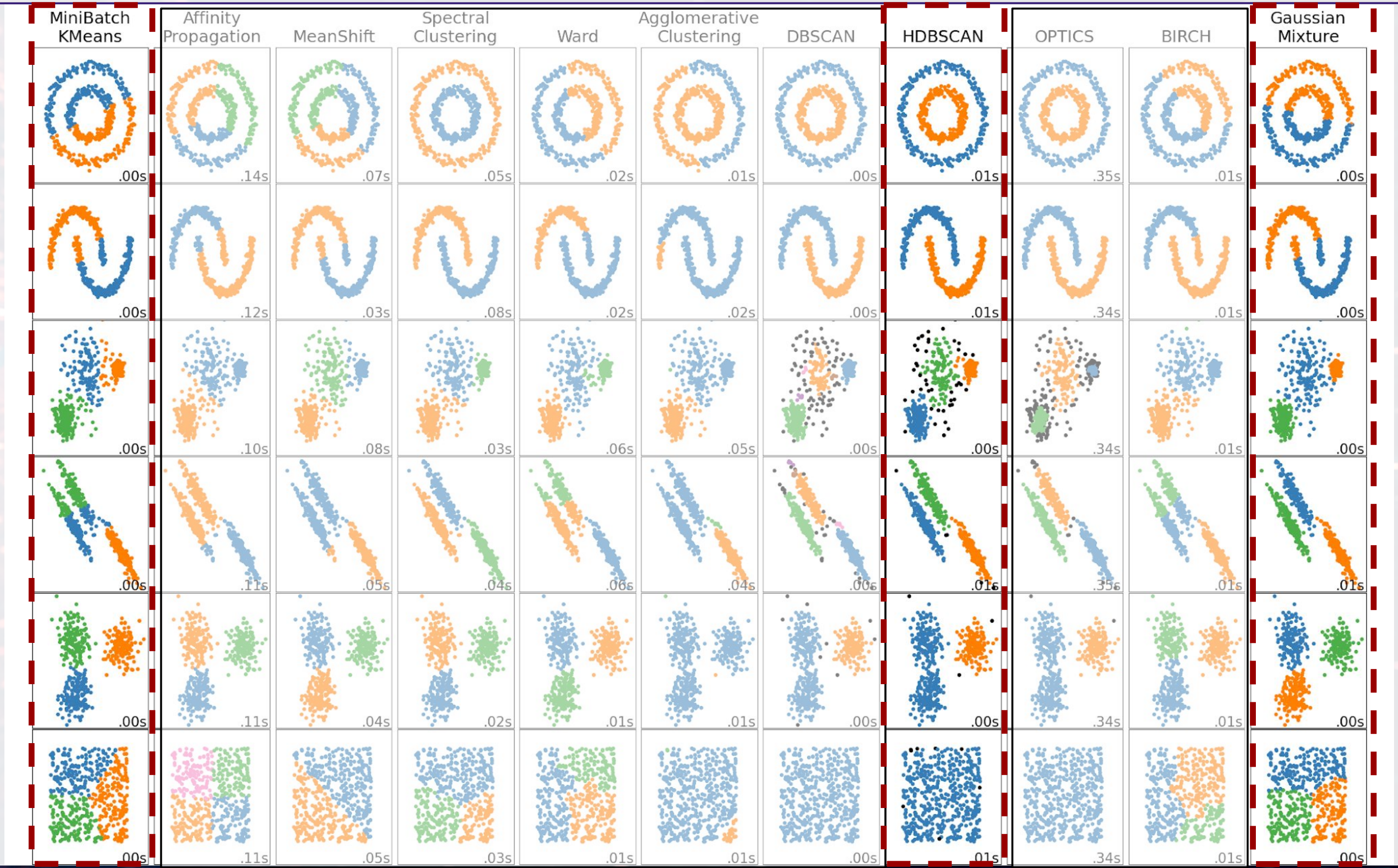
- star-galaxy classification (Odewahn et al. 1992, Bertin & Arnouts 1996)
- morphological galaxy classification (Storrie-Lombardi et al. 1992, Dieleman et al. 2015)
- photo-z (Firth et al. 2003)
- spectra classification (von Hippel et al. 1994, Folkes et al. 1996)
- solar activity prediction (Lundstedt & Wintoft 1994)

Mostly supervised

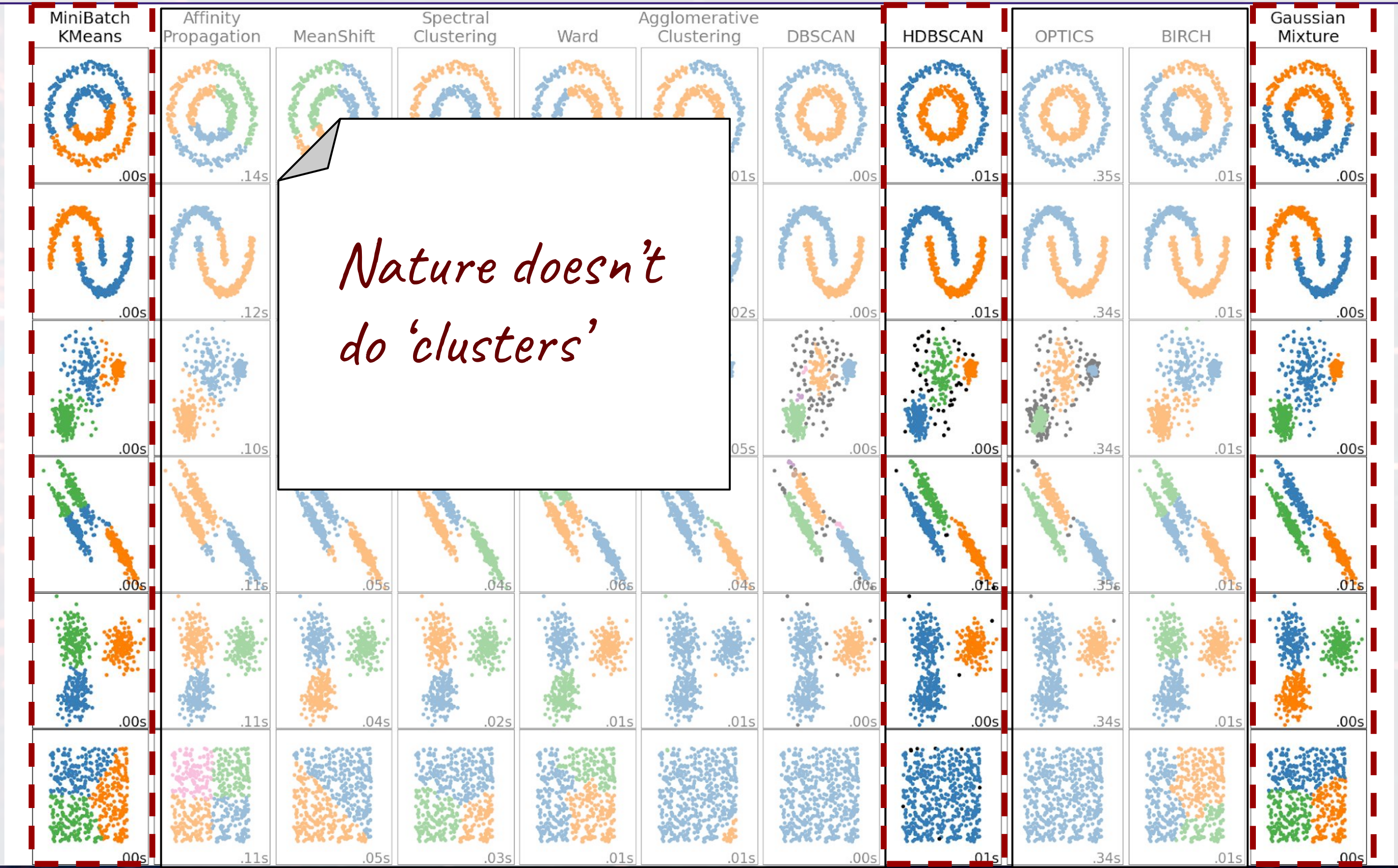
# UML algorithms: clustering



# UML algorithms: clustering



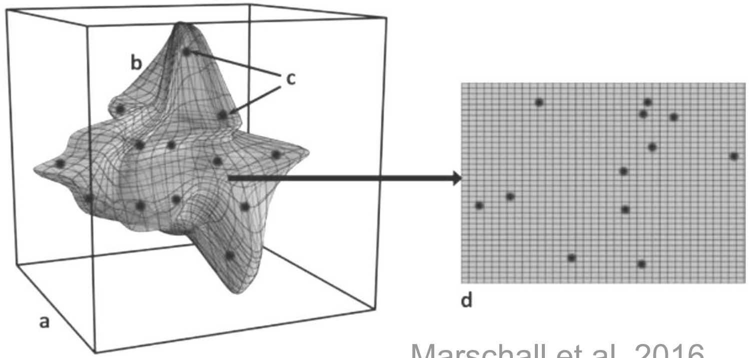
# UML algorithms: clustering



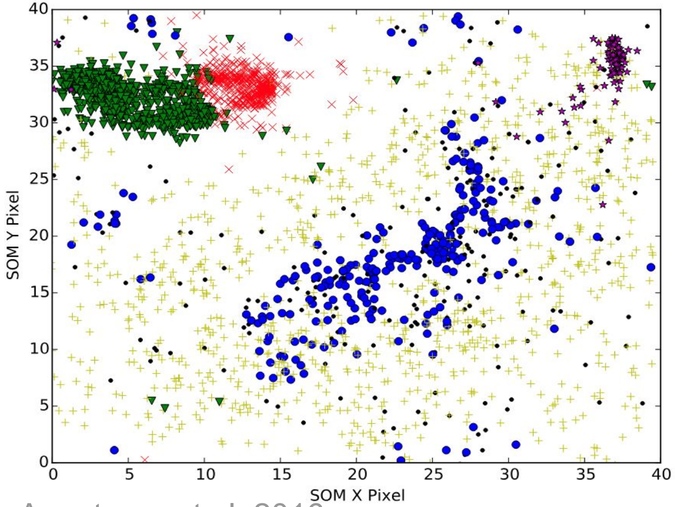
# UML algorithms: dimensionality reduction

- PCA (+non-linear);
- Self-Organizing Maps (SOM);
- Uniform manifold approximation and projection (UMAP);
- t-distributed stochastic neighbor embedding (tSNE);
- Neural gas;
- Autoencoders;

Two guiding principles: proximity (not always Euclidian!) and continuity

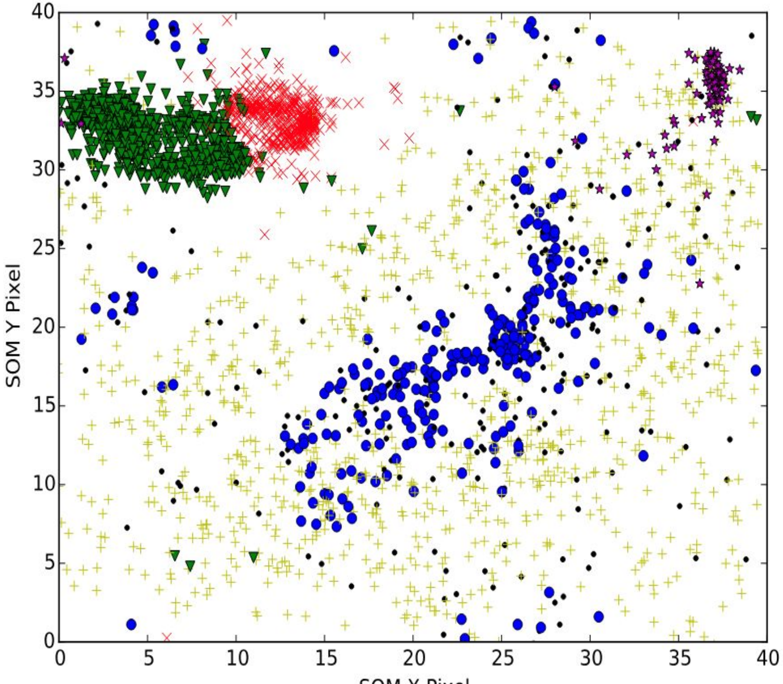


Marschall et al. 2016

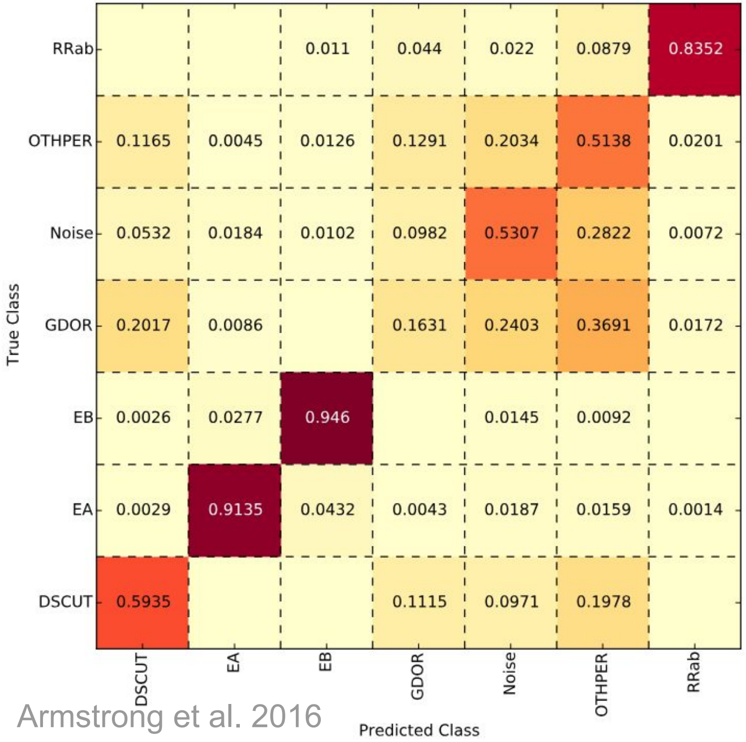


Armstrong et al. 2016

# Dimensionality reduction: N-D -> 1D



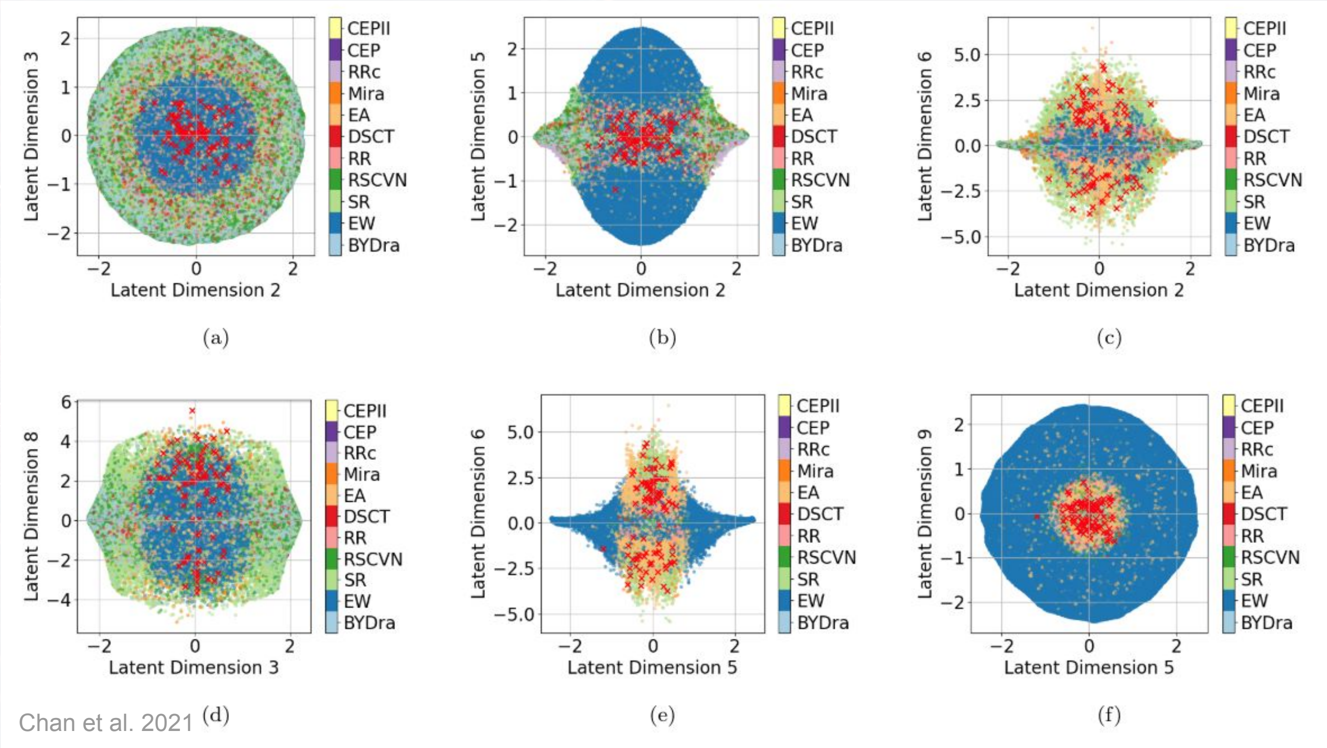

Armstrong et al. 2016



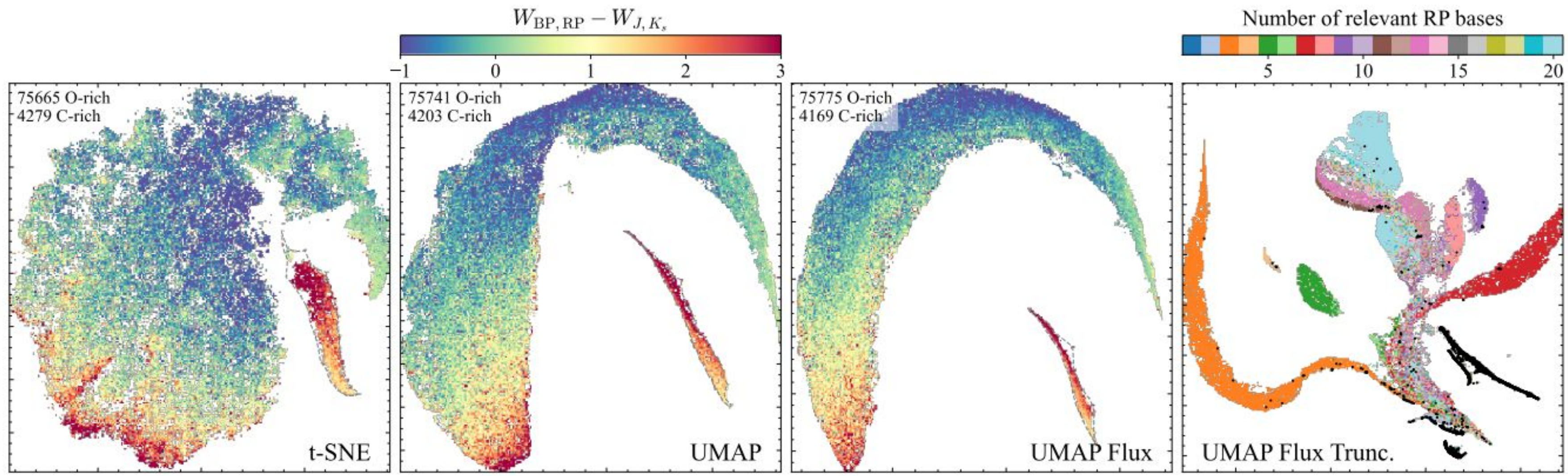

Armstrong et al. 2016

# UML algorithms: anomaly detection

All of the above
+ modified supervised
algorithms



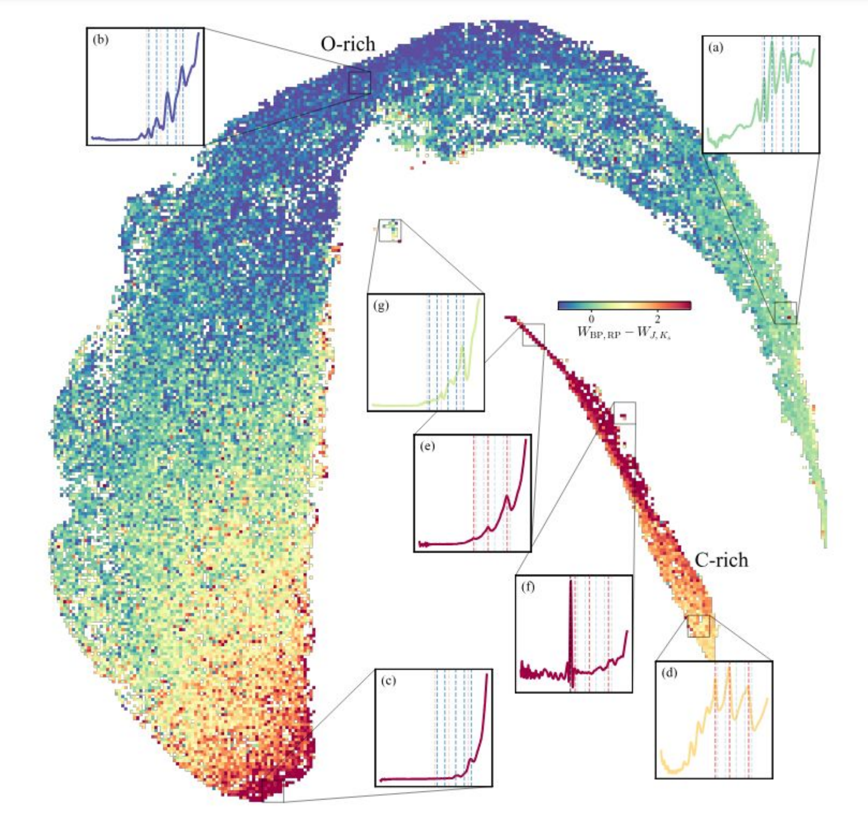
# UML algorithms as a form of dark magic




Sanders & Matsunaga, 2023

# UML algorithms as a form of dark magic

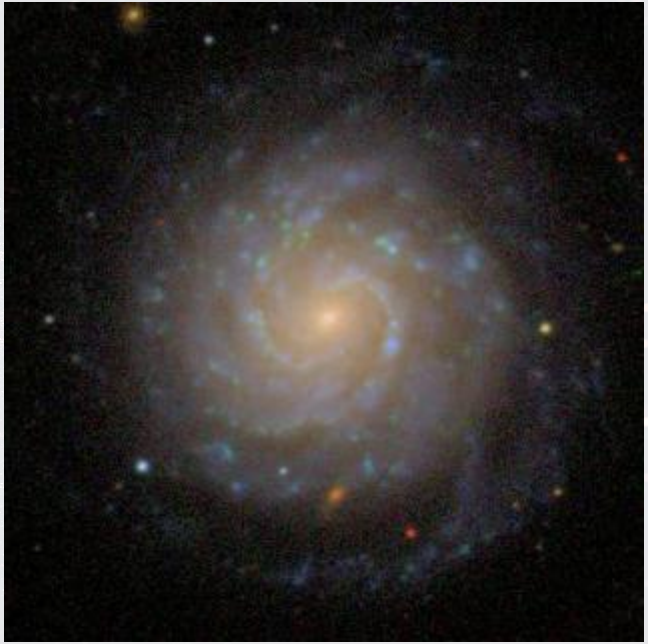
- not as popular as supervised (and self-supervised in case of LLM) in industry → not so many well-developed tools
- interpretability is worse than for supervised ML (read: horrible)
- not many tools are adapted for data with uncertainties




Sanders & Matsunaga, 2023

## Astronomical LCs as a source of unyielding pain

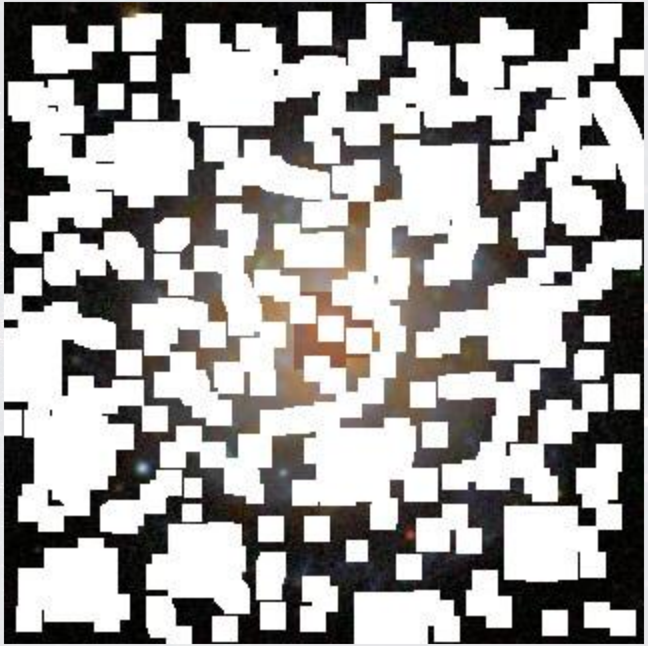
- Variable sizes of data vectors
- Uneven sampling
- Phase gaps
- No universal period finding algorithm (+aliasing)
- We need to take into account uncertainties
- Outliers can ruin period finding (and, consequently, everything)
- Good interpolations are computationally expensive (Gaussian Processes!)



# Astronomical LCs as a source of unyielding pain

- Variable sizes of data vectors
- Uneven sampling
- Phase gaps
- No universal period finding algorithm (+aliasing)
- We need to take into account uncertainties
- Outliers can ruin period finding (and, consequently, everything)
- Good interpolations are computationally expensive (Gaussian Processes!)

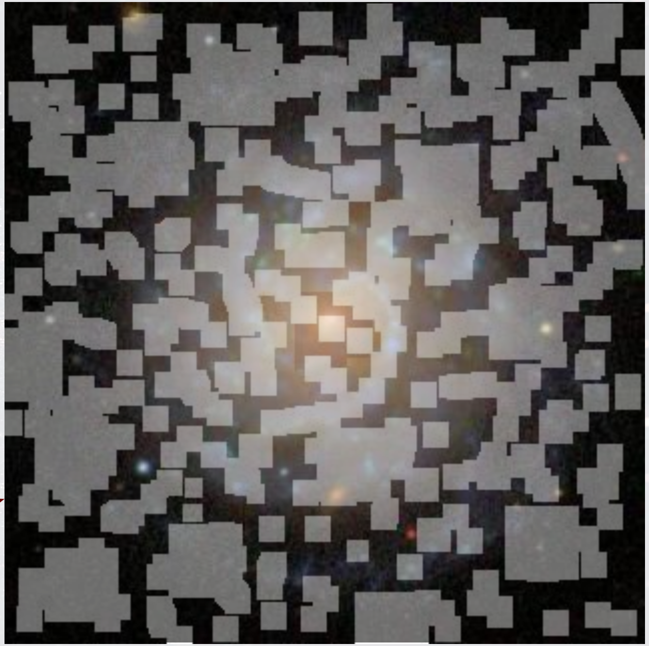
*If galaxy images were like LCs*



# Astronomical LCs as a source of unyielding pain

- Variable sizes of data vectors
- Uneven sampling
- Phase gaps
- No universal period finding algorithm (+aliasing)
- We need to take into account uncertainties
- Outliers can ruin period finding (and, consequently, everything)
- Good interpolations are computationally expensive (Gaussian Processes!)

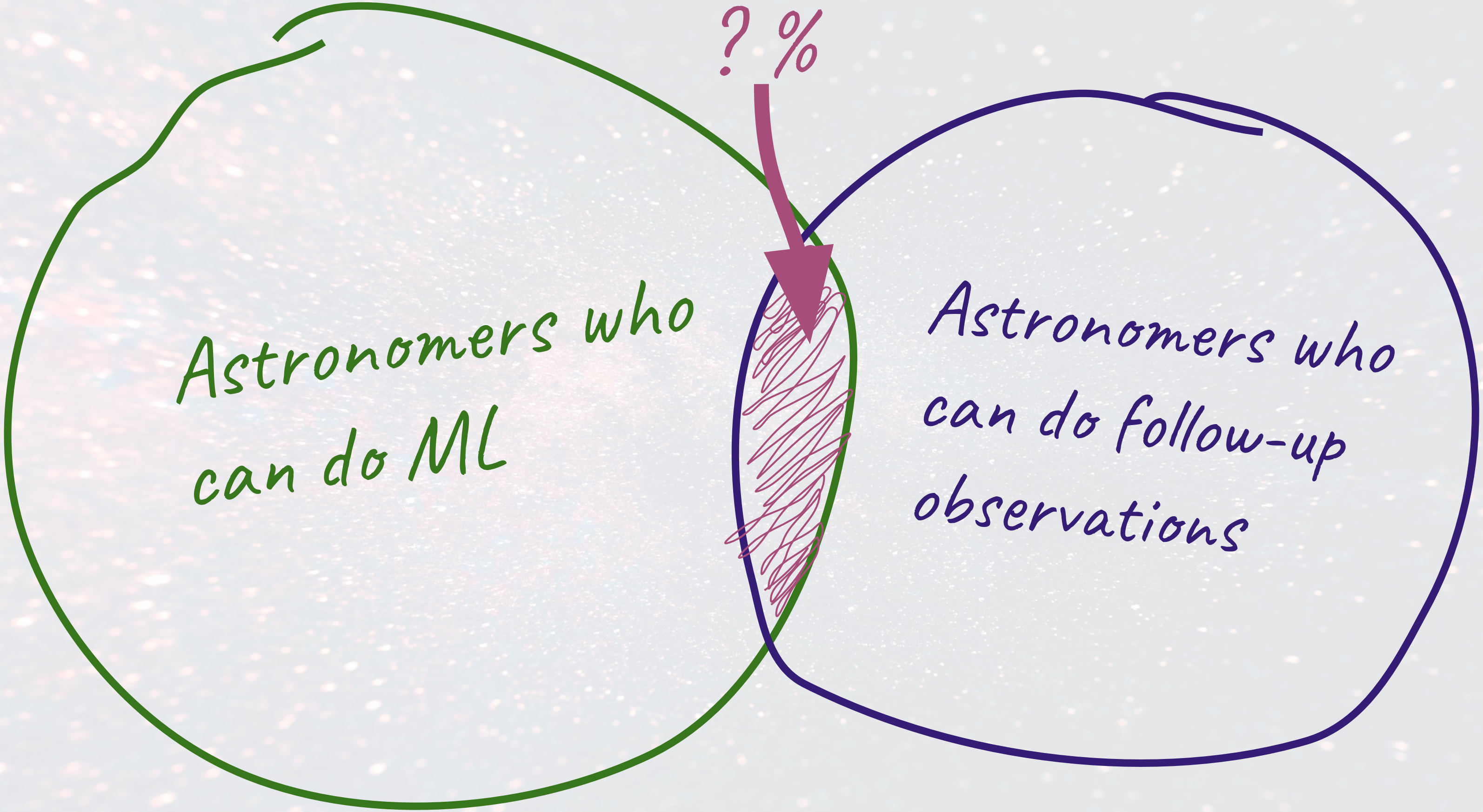
*We need simulations and interpolations*

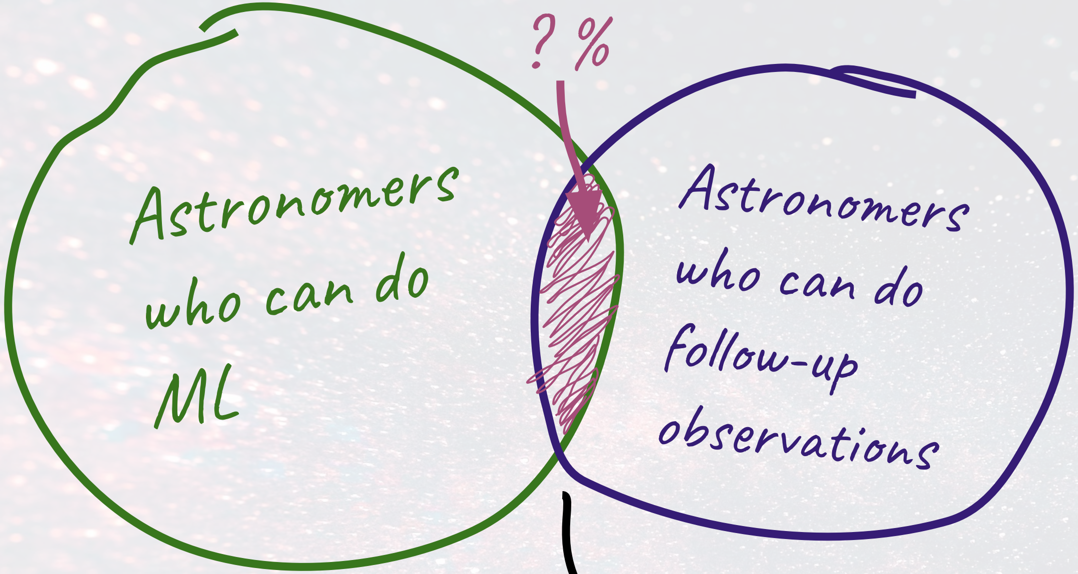


# What are the ‘rare gems’ for the variable sources?

- **Unknown unknowns**
  - Boyajian star & Co (Boyajian et al. 2016);
  - Blue large-amplitude pulsators (BLAPs, Pietrukowicz et al. 2017)
- **Known unknowns (intermediate types, rapidly passing evolutionary stages)**
  - Fast Yellow Pulsating Supergiants (FYPS, Dorn-Wallenstein et al. 2020)
  - changing-state AGNs (Sanchez-Saez et al. 2021)

- **Anomalous objects of common categories**
  - Multi-mode anomalous Cepheid (Soszynski et al. 2020)
  - Beat type II Cepheid (Smolec et al. 2018)
  - Magnetic chemically peculiar stars (Bernhard et al. 2021)
- **Rare, but already known subtypes**
  - High-Amplitude Delta Scuti (Lee et al. 2008)
  - Anomalous Cepheids (Soszynski et al. 2015, 2017)
- **‘Ordinary’ objects in non-ordinary environments**
- **‘Ordinary’ objects with high value for the ‘hot topics’**


? %
Astronomers who can do ML
Astronomers who can do follow-up observations




Don't look for new algorithms
Don't look for fancy features
Look for the right co-authors

# MapLC project (SMASH, Slovenia, starting Feb 2025)

**Objectives:**

- Develop a UML-based software package for astronomical variability data analysis;
- Develop a tool for managing feature sets, coming from different sources;
- Compare different algorithm/feature set combination for several science cases.

**Deliverables:**

- Work datasets with several feature sets;
- UML-visualization and interpretation Python package;
- An comparison of algrorithm/feature set performance for the test science cases;
- Detection/classification catalogues for test science cases

**Feature sets:**

- pre-developed by the LSST TVS SC community;
- pre-developed by the alert brokers;
- home-brewed ML-based

**UML algrorithms:**

- Self-Oranising Maps;
- UMAP;
- HDBSCAN

**Test science cases:**

- Blue Large-Amplitude Pulsators (BLAPs);
- Yellow Pulsating Supergiants (YPSs);
- Tidal Disruption events (TDEs);
- Supernovae (SNe);
- Young Stellar Objects (YSOs)

# Challenges, infrastructural requirements and solutions

| Challenges | Solutions |
| --- | --- |
| • Improve datasets ‘discoverability’<br>• Improve UML interpretability<br>• Deal with uneven sampling and phase gaps<br>• Adapt UML algorithms to data with uncertainties<br>• Look for laws and relations in higher dimensionalities<br>• Adopt ‘anomaly-oriented’ mindset | • Data archives APIs; Schema browsers; tutorials for crossmatching/forced photometry, tutorials for quality cuts. (Software development trainings - invest in making code reusable!)<br>• performance comparison papers for feature sets and period finding algorithms<br>• fast LC simulations and interpolation algorithms<br>• computer scientists’ help needed for improving interpretability<br>• better visualization<br>• data imputation (including UML), semi-supervised ML<br>• ‘anomalies-oriented’ follow-up calls, papers? (proactive approach)<br>• ‘anomaly-oriented’ projects. Not ‘we can reproduce the already existing classification’, but ‘let’s take the objects from the ‘Other’ category and figure out what they are’. |