

Anomaly Detection in Astronomy

Michelle Lochner

Senior Lecturer

University of the Western Cape/
South African Radio Astronomy
Observatory



UNIVERSITY *of the*
WESTERN CAPE

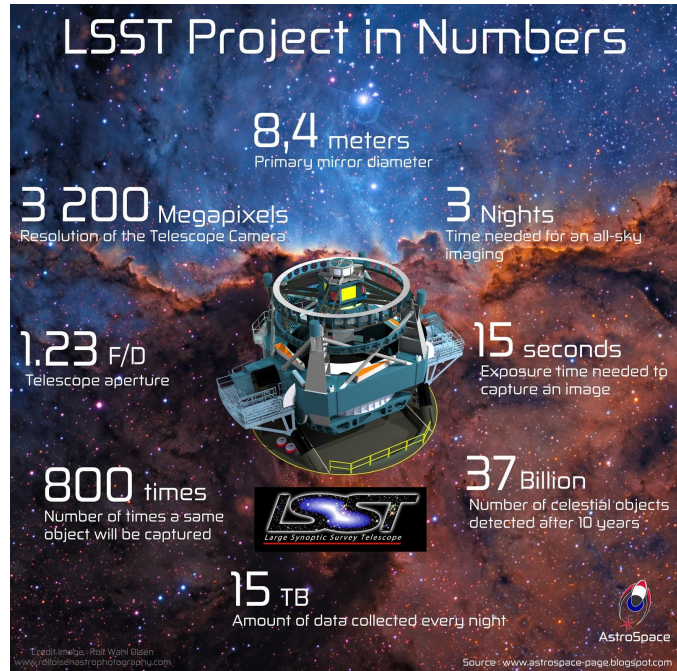


SARAO
South African Radio
Astronomy Observatory



The Vera C. Rubin Observatory

Legacy Survey of Space and Time (LSST)



Rubin Obs/NSF/AURA & Bruno C. Quint

Day 154



Vera C. Rubin Observatory

The Square Kilometre Array



SKA1-mid

the SKA's mid-frequency instrument



Location:
South Africa

Frequency range:
350 MHz
to
15.3 GHz
with a goal of 24 GHz



197 dishes
(including 64 MeerKAT dishes)



Maximum baseline:
150km

SKA1-low

the SKA's low-frequency instrument



Location: Australia

Frequency range:
50 MHz
to
350 MHz



~131,000
antennas spread between
512 stations



Maximum baseline:
~65km

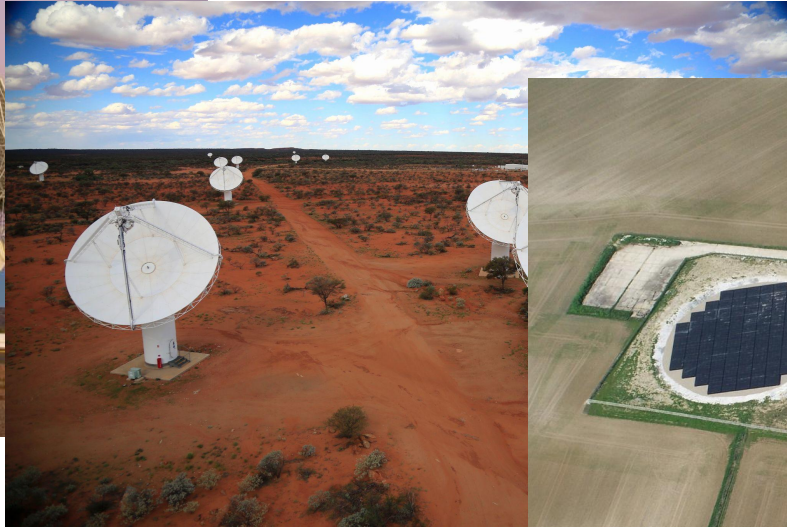
Square Kilometre Array Observatory

MeerKAT



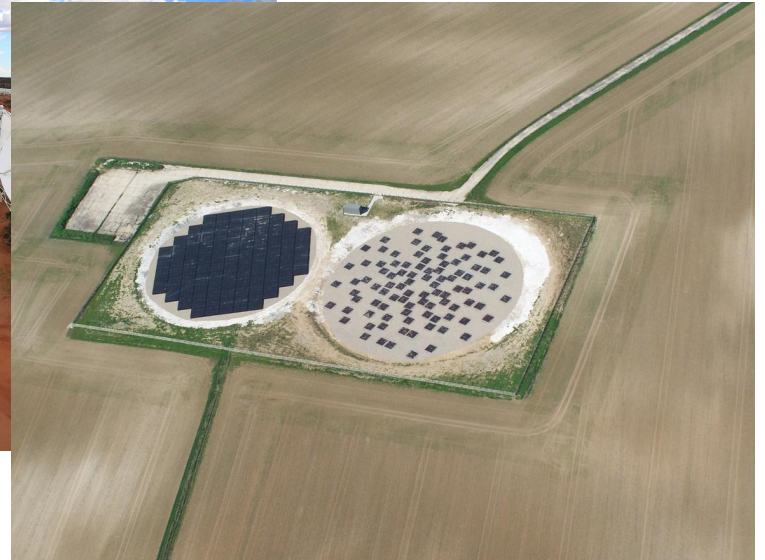
SARAO

ASKAP



CSIRO

LOFAR



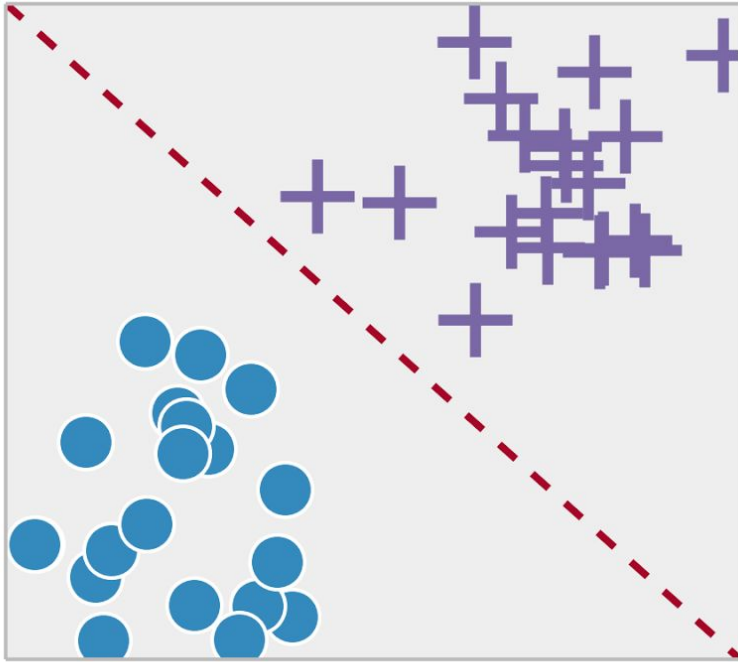
LOFAR

We're facing a data explosion



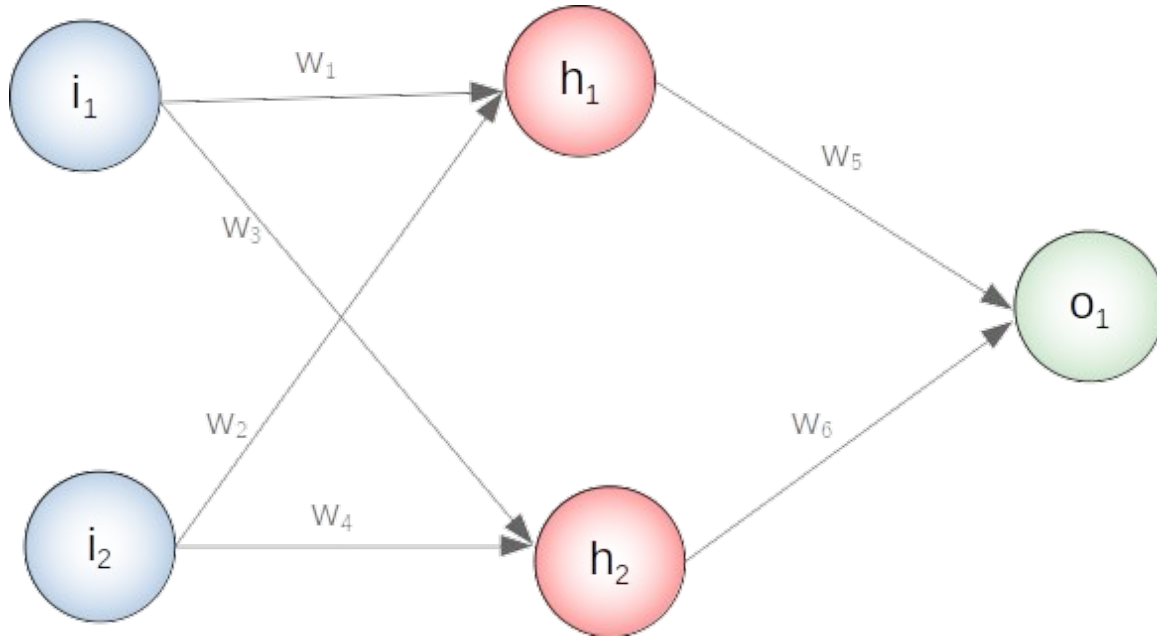
Machine Learning

Supervised Machine Learning



Automatically learns a **model** to map inputs to outputs, using a **training set**.

Machine Learning



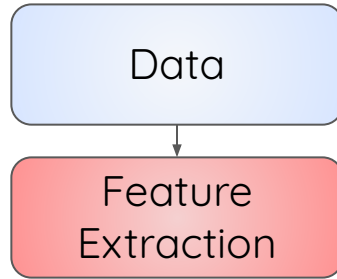
$$\tanh[w_5 \tanh(w_1 i_1 + w_2 i_2) + w_6 \tanh(w_3 i_1 + w_4 i_2)]$$

Some Definitions

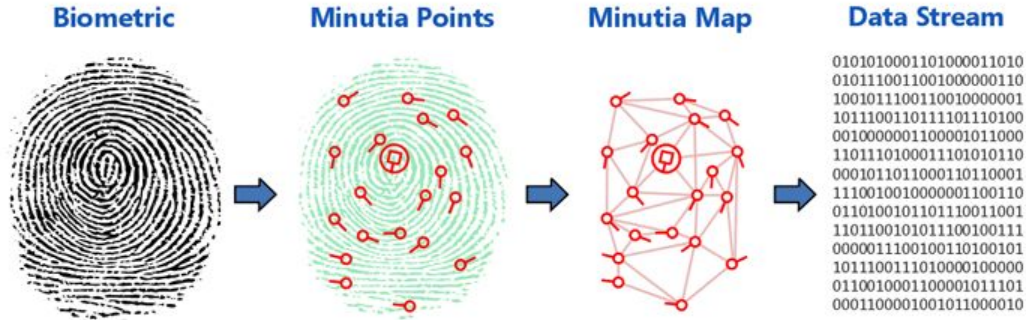
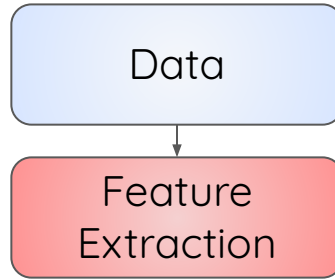
Some Definitions

Data

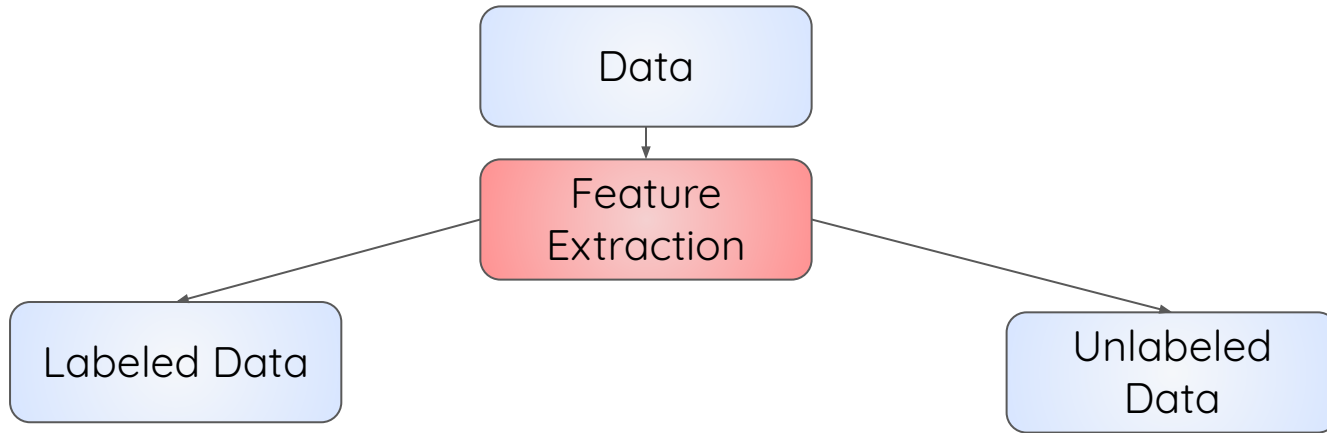
Some Definitions



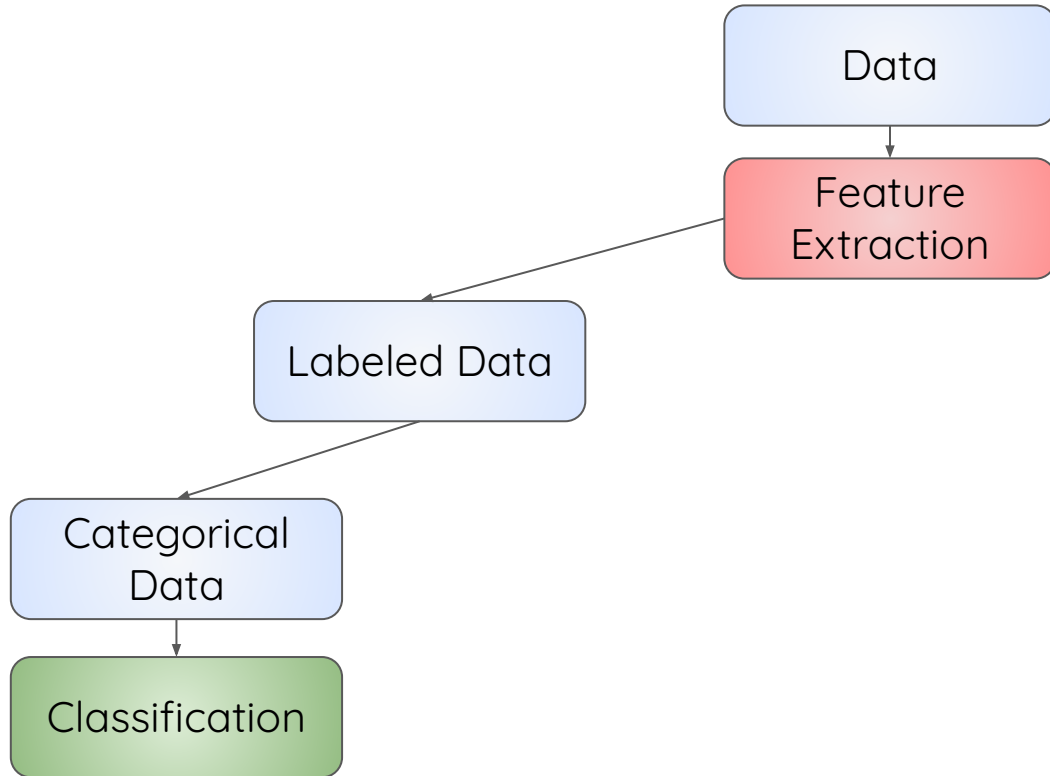
Some Definitions



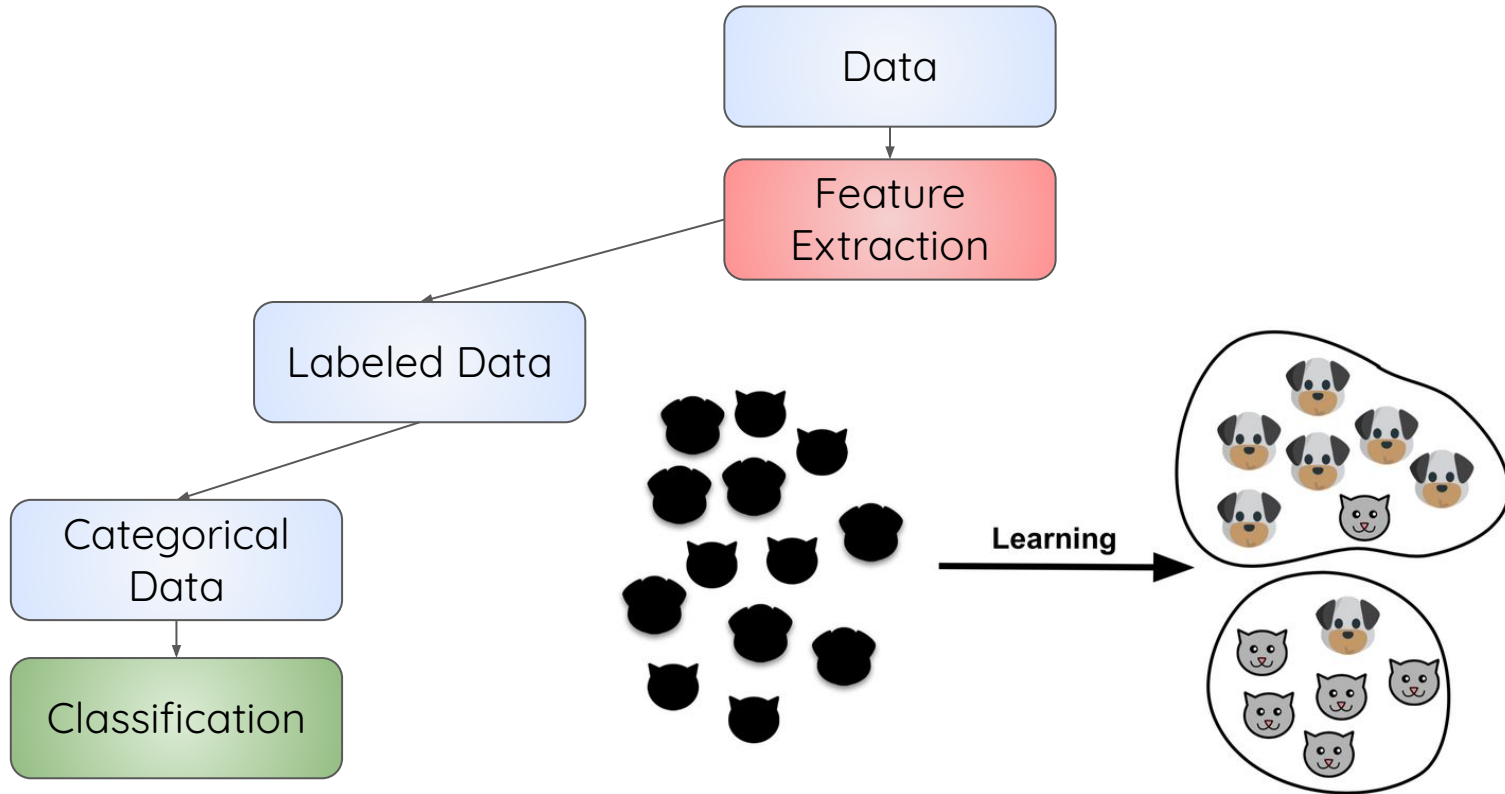
Some Definitions



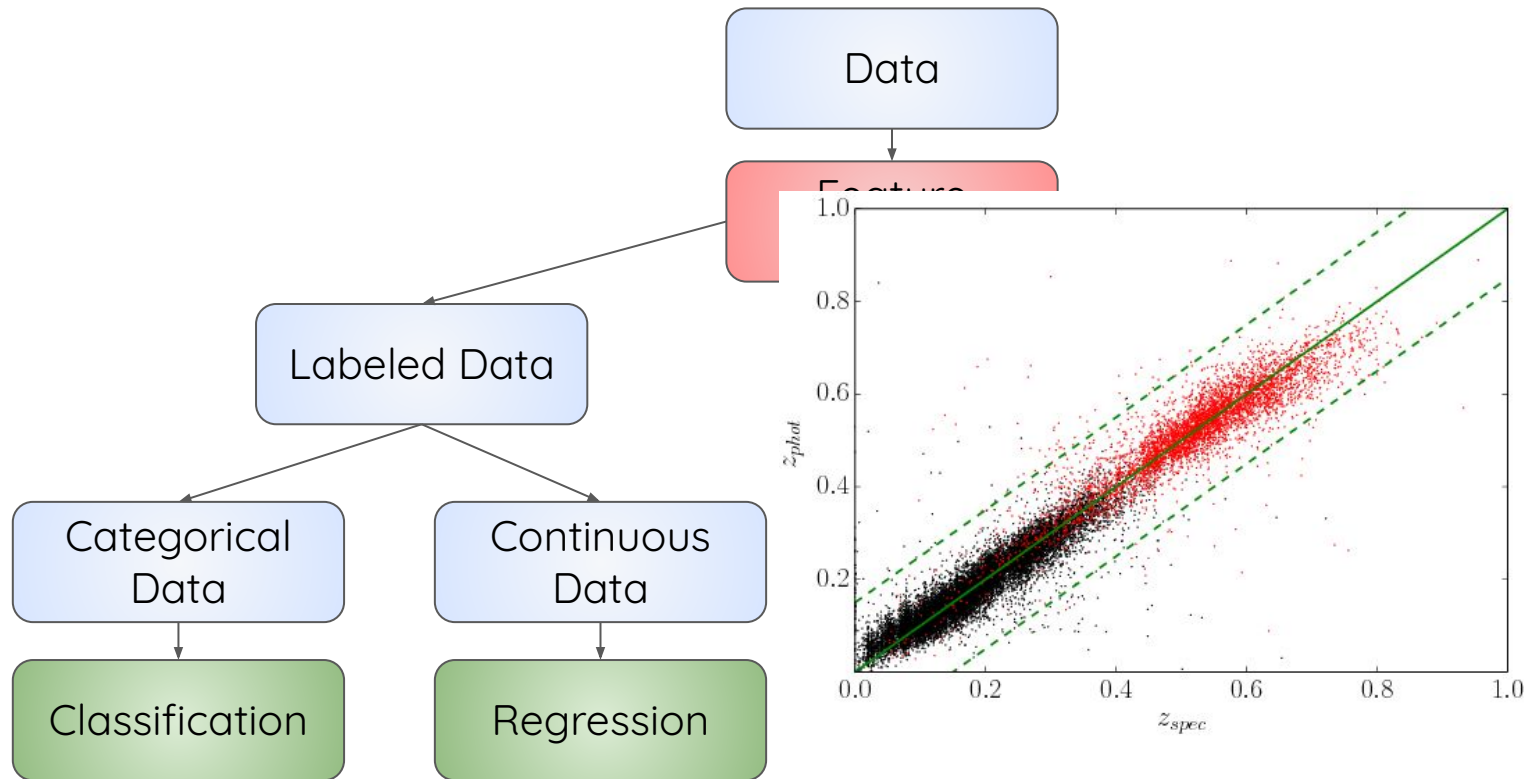
Some Definitions



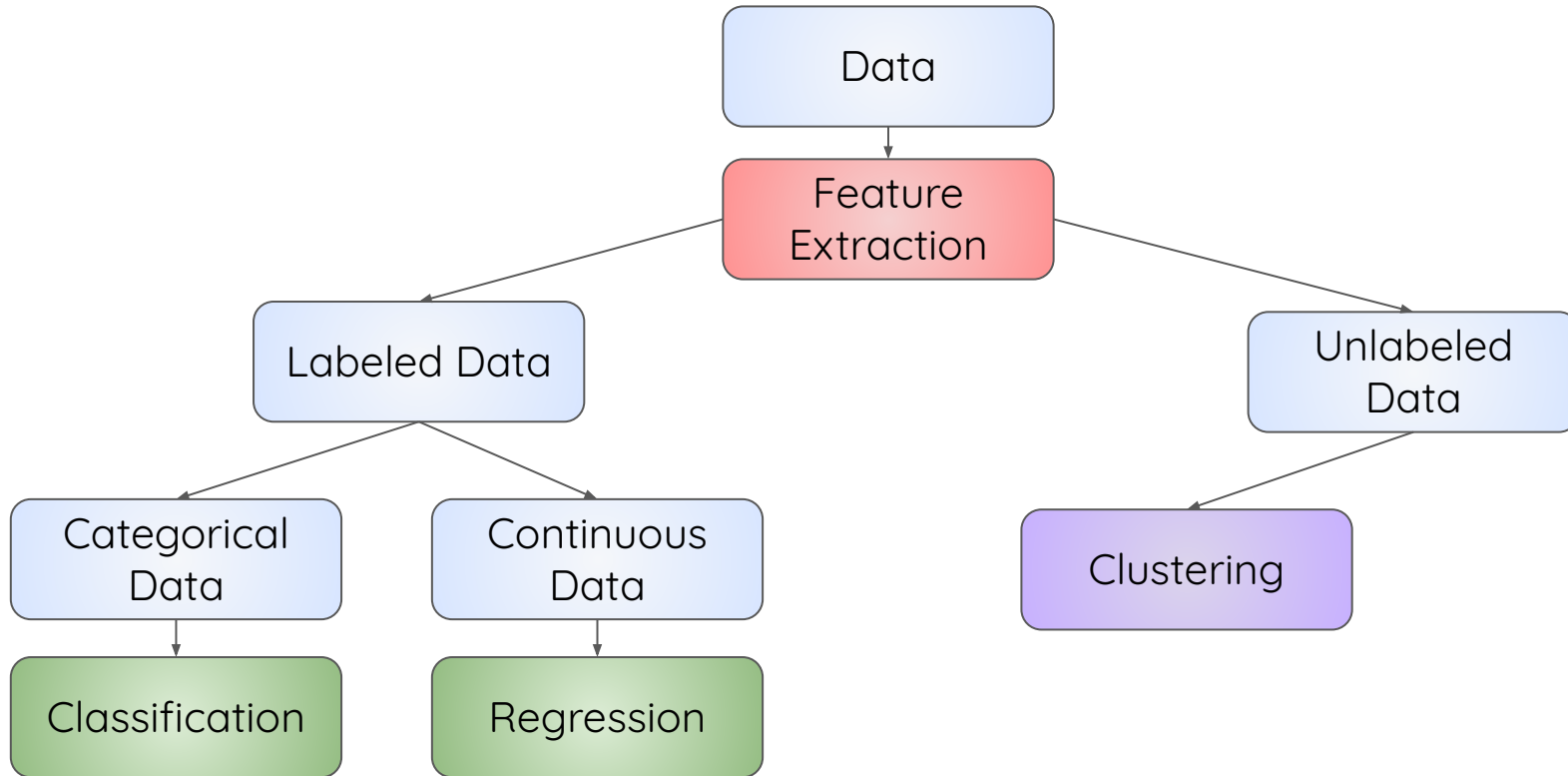
Some Definitions



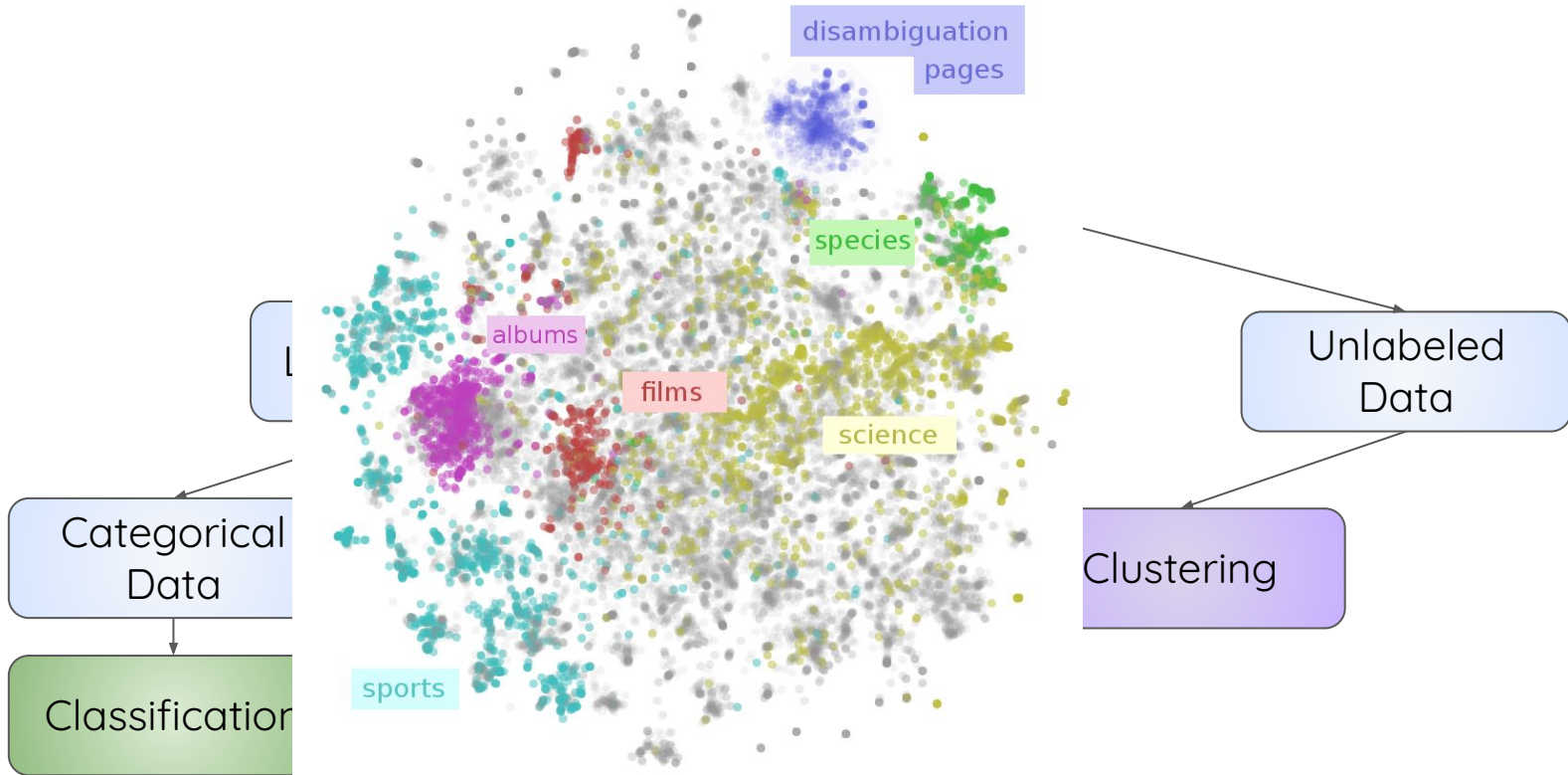
Some Definitions



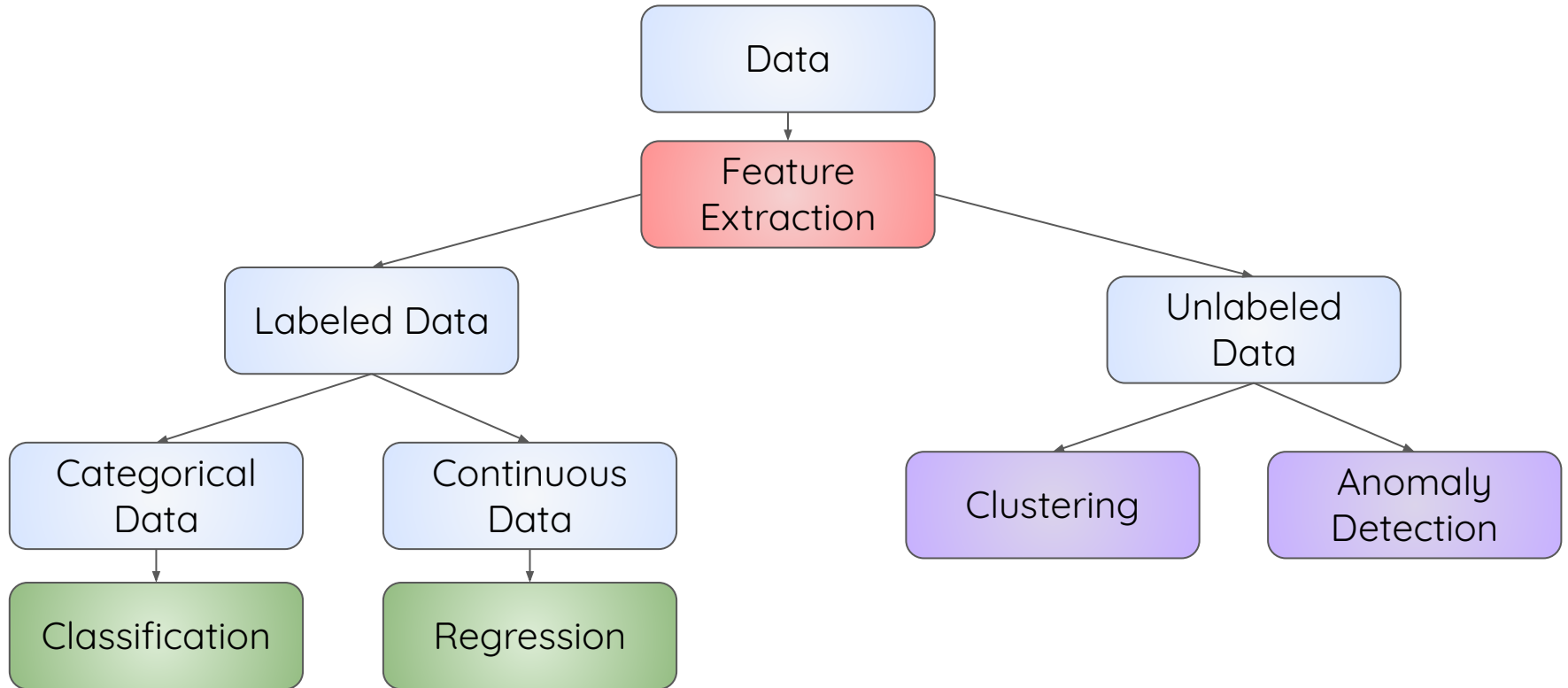
Some Definitions



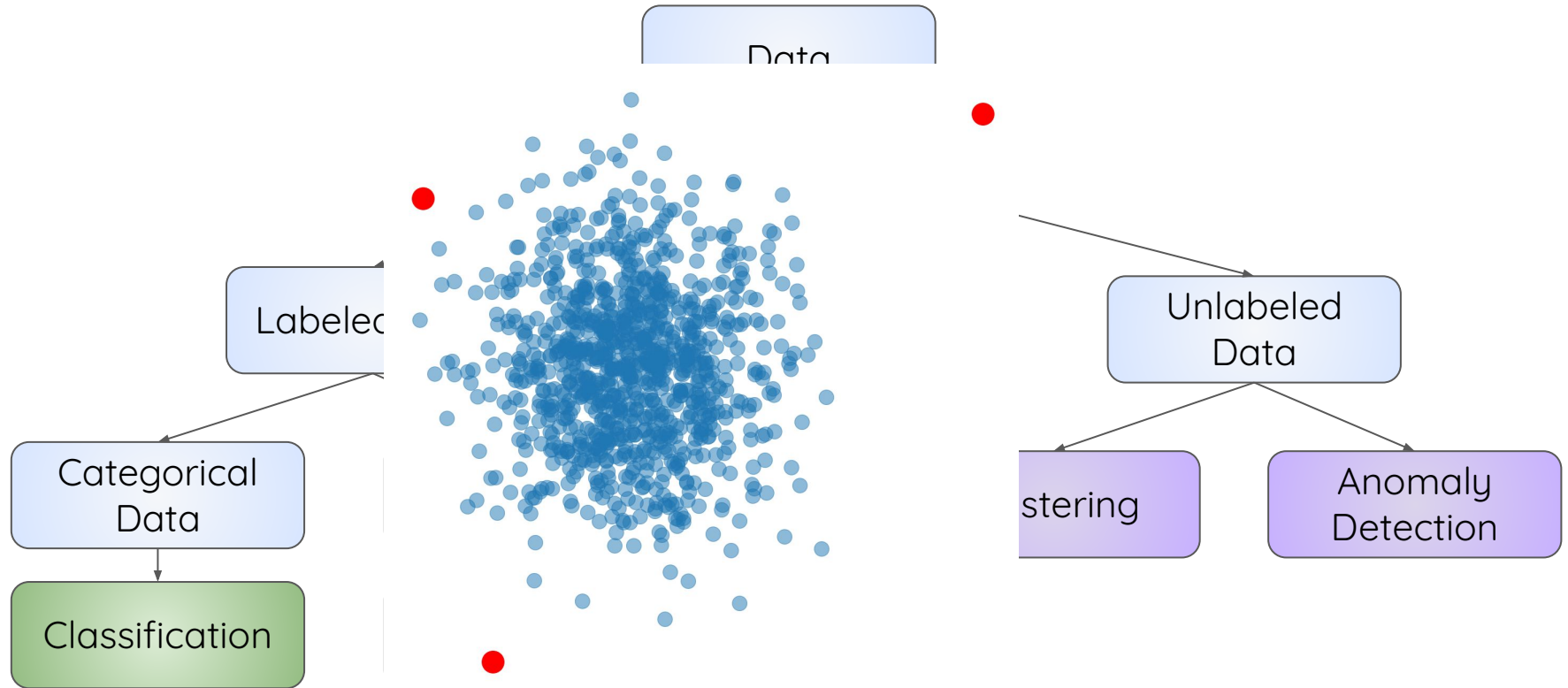
Some Definitions



Some Definitions



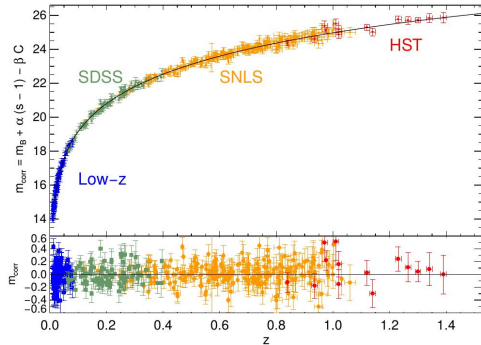
Some Definitions



Supernova Classification

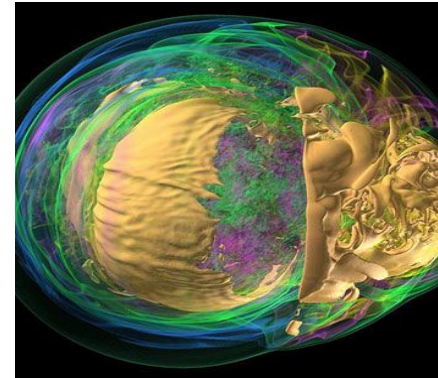
Supernova Classification

Type Ia

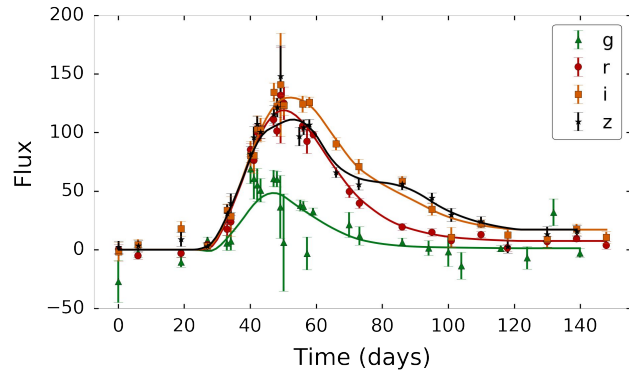


Conley et al. (2011)

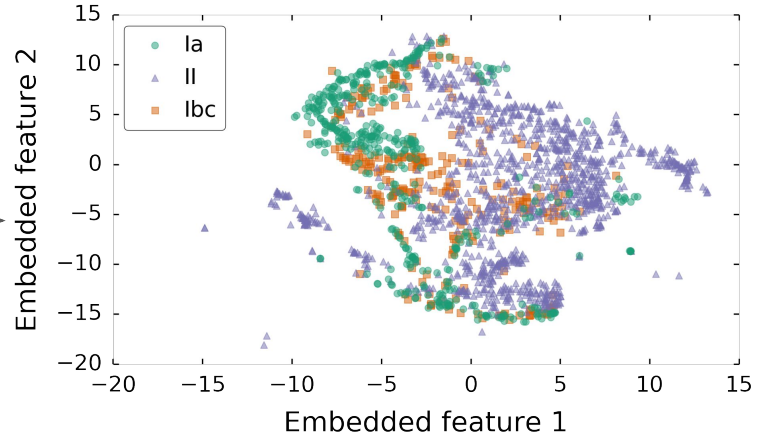
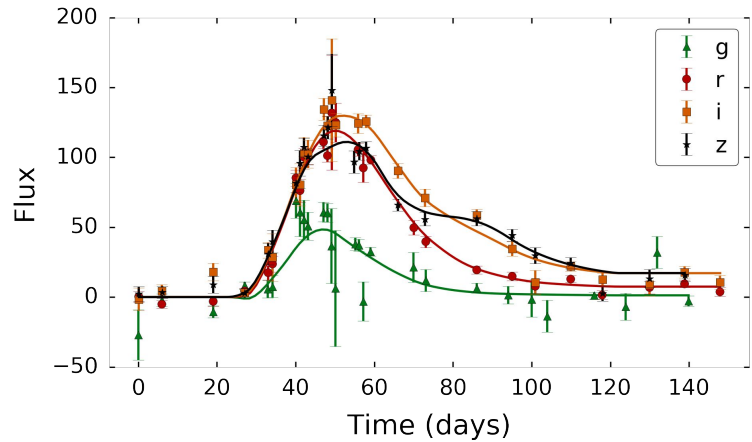
Core Collapse



Supernova Classification



Supernova Classification

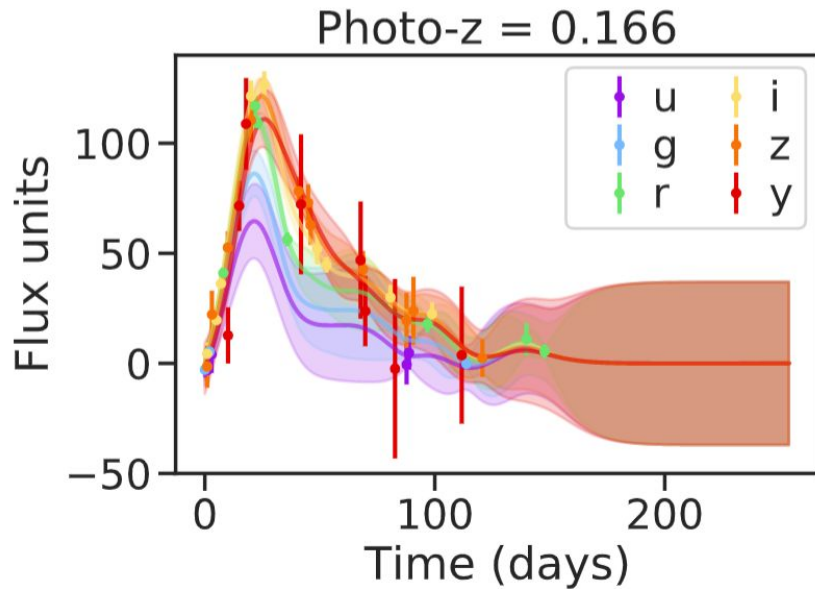


Lochner et al. (2016) - 1603.00882

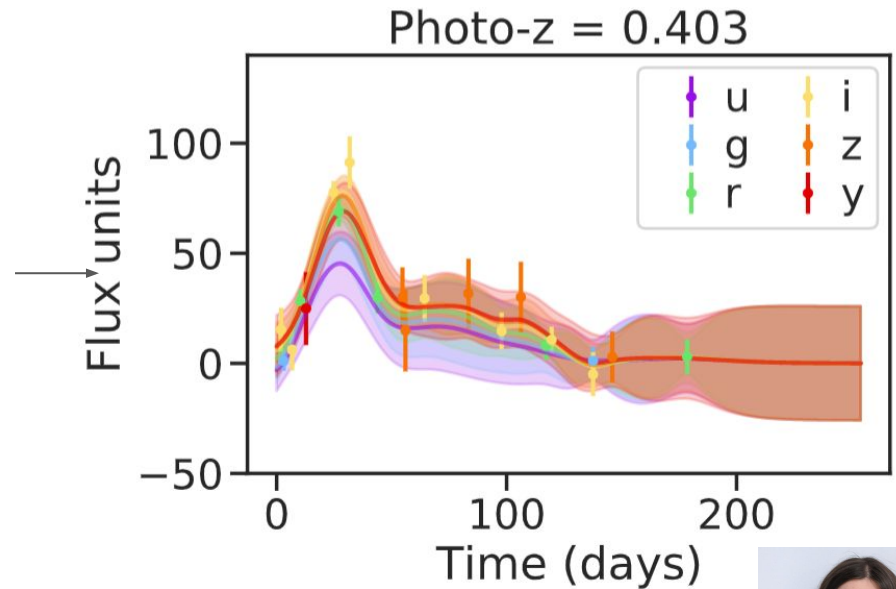
Sooknunan et al. (2018) - 1811.08446

PLAsTiCC

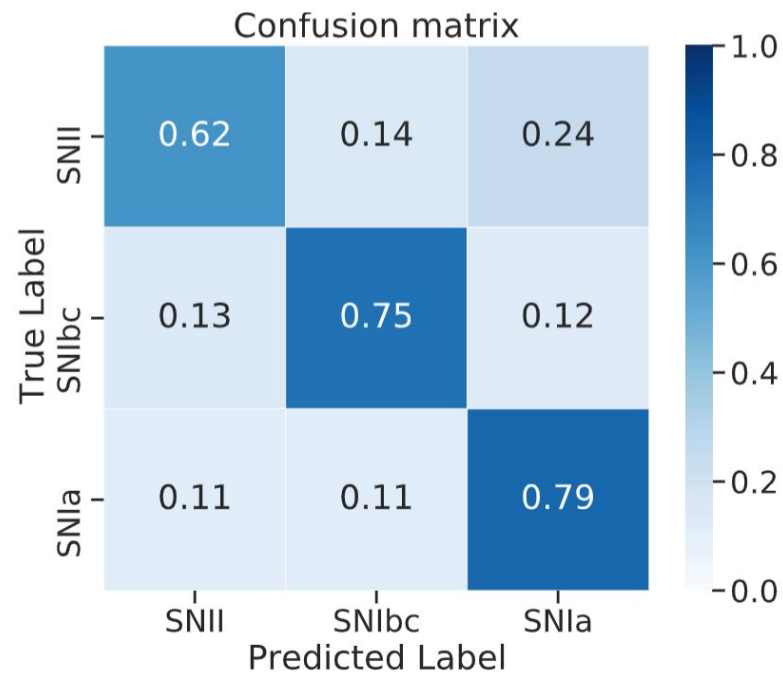
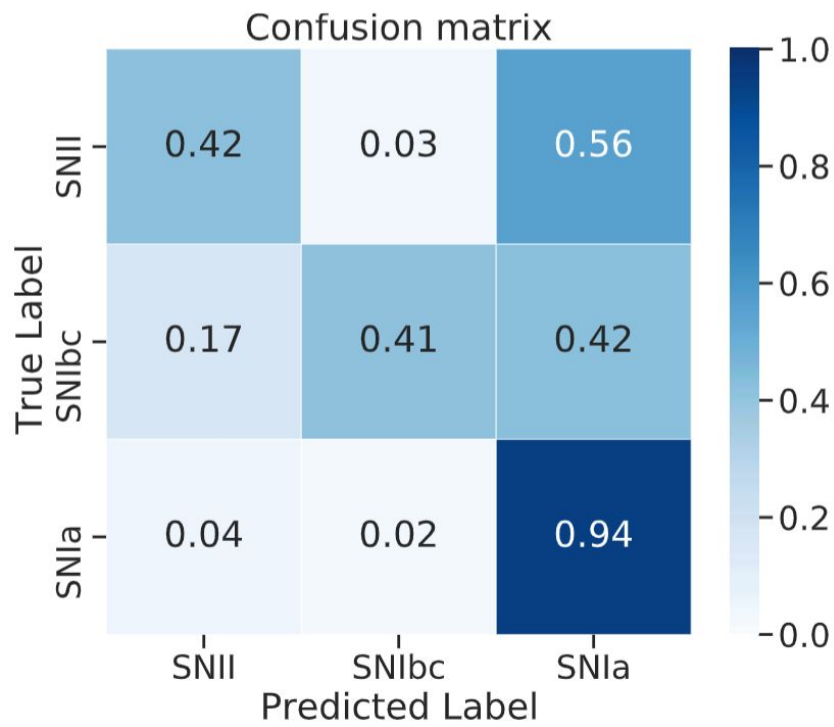
Original



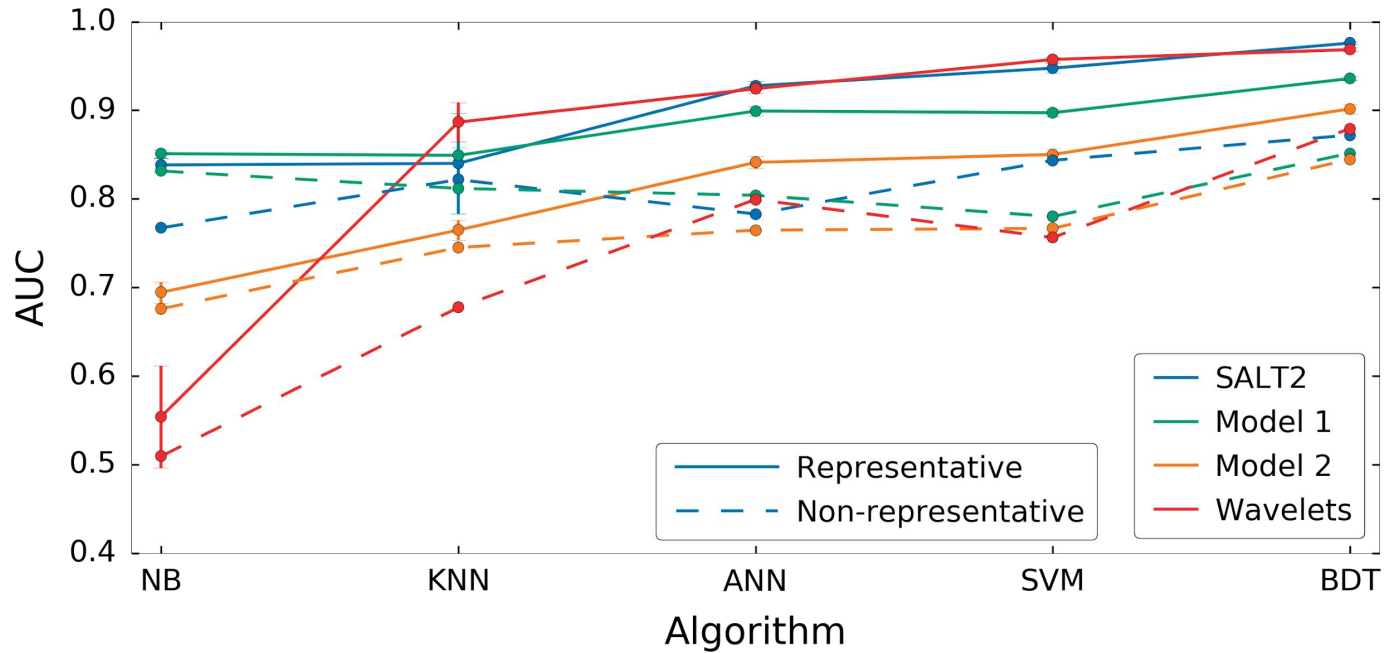
Synthetic



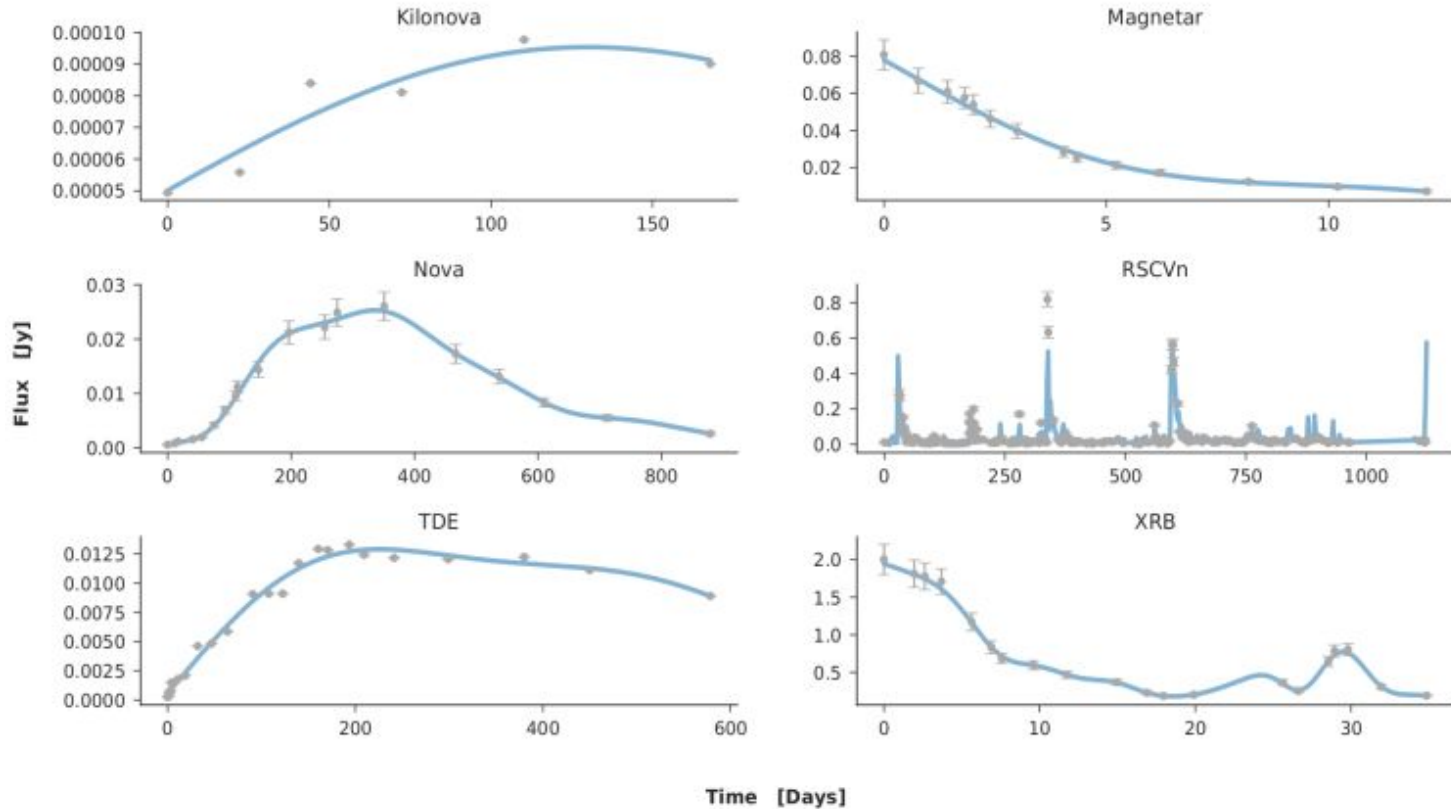
PLAsTiCC



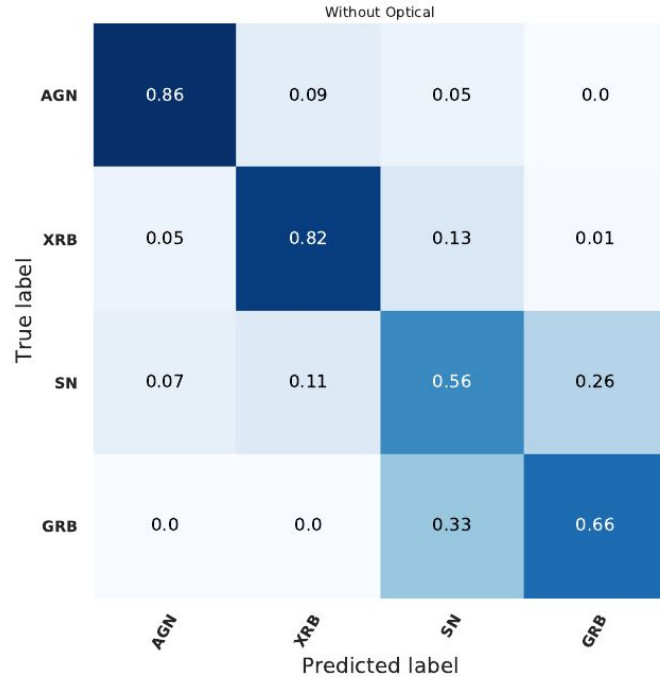
Supernova Classification



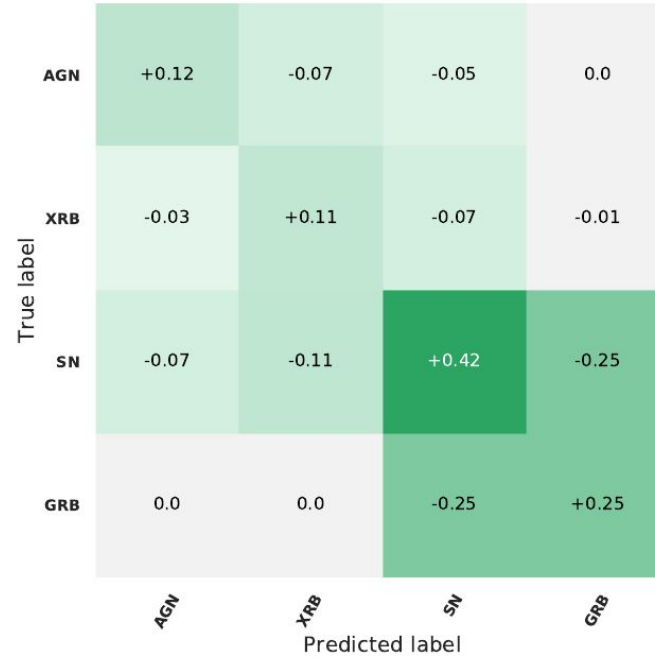
Multiwavelength Transient Classification



Multiwavelength Transient Classification

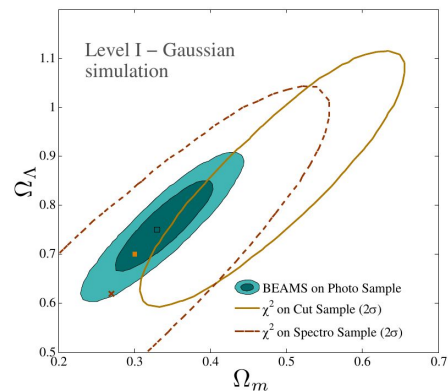
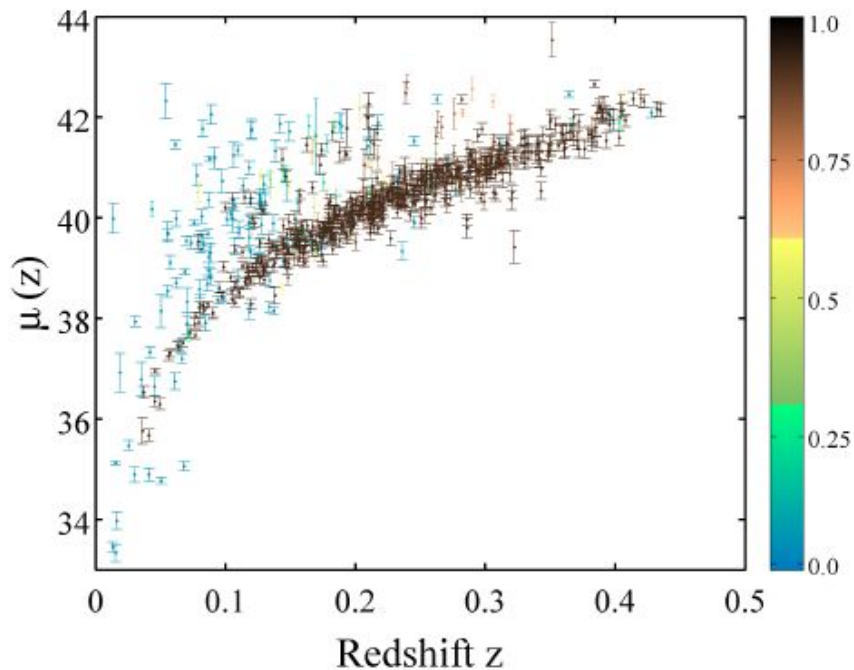


(a) Confusion matrix without optical feature



(b) Confusion matrix showing the difference when optical feature is added

Science with Imperfect Classification



Bayesian Estimation Applied to Multiple
Species (BEAMS)

Kunz, Bassett & Hlozek - 0611004

Newling et al. - 1110.6178

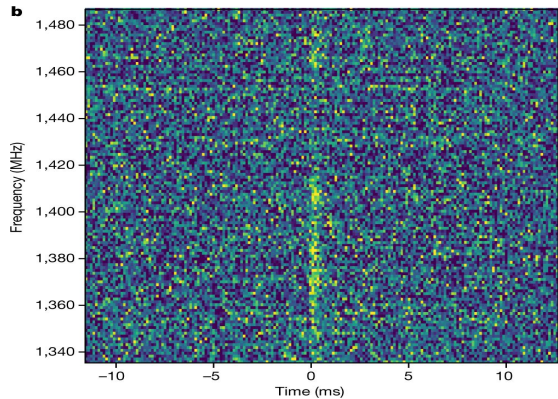
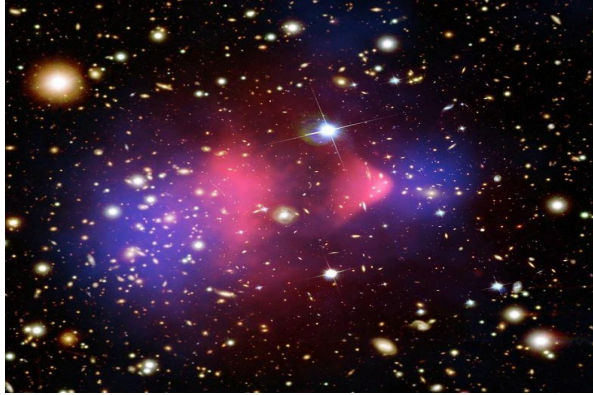
Hlozek et al. - 1111.5328

Lochner et al. - 1205.3493

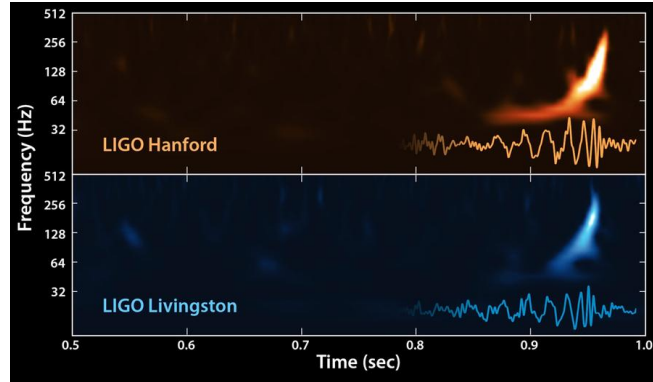
Roberts et al. - 1704.07830

Anomaly Detection

Known Unknowns - rare events

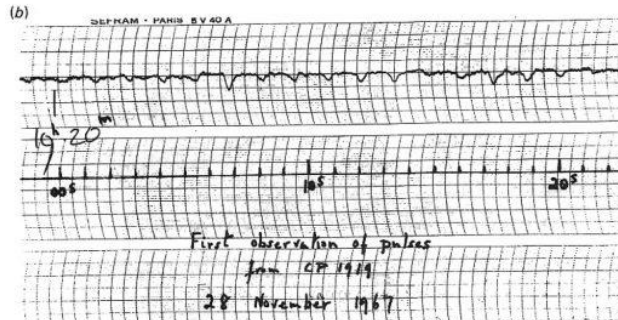
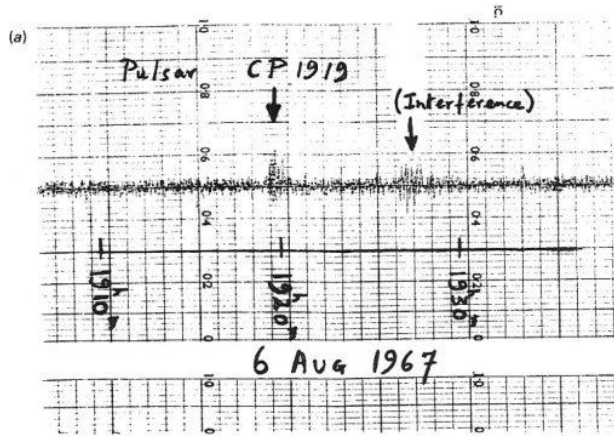


Ravi et al. (2019)



Hubble/ NASA/ ESA

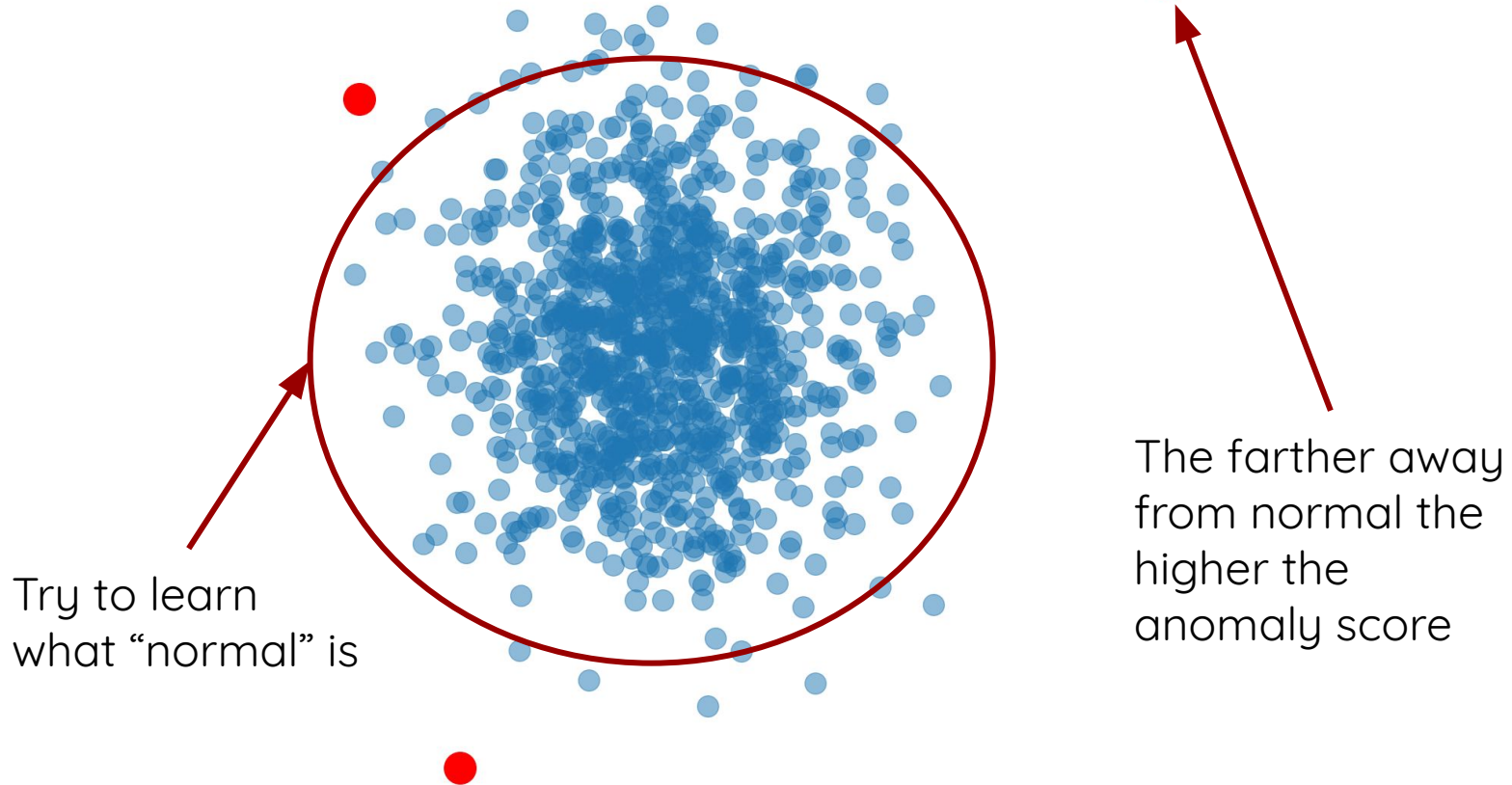
Unknown Unknowns - new anomalies



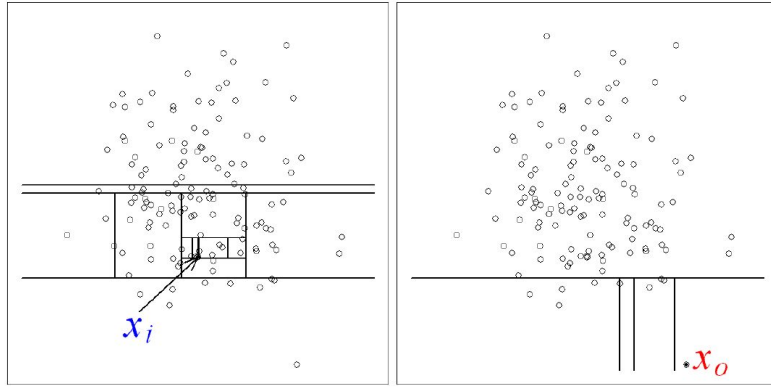
How do we discover new phenomena...

...among 10 million possibilities?

Anomaly Detection

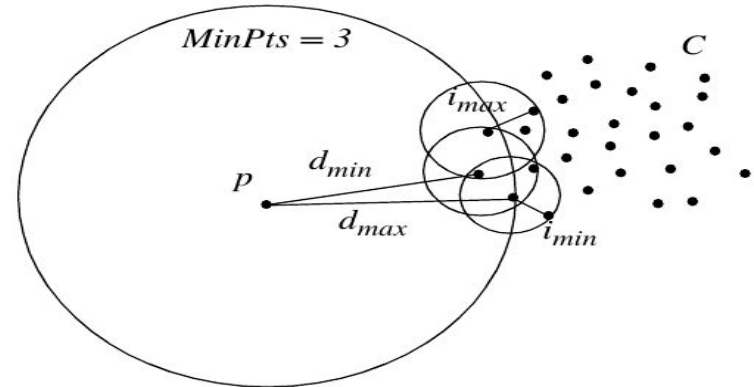


Anomaly Detection Algorithms

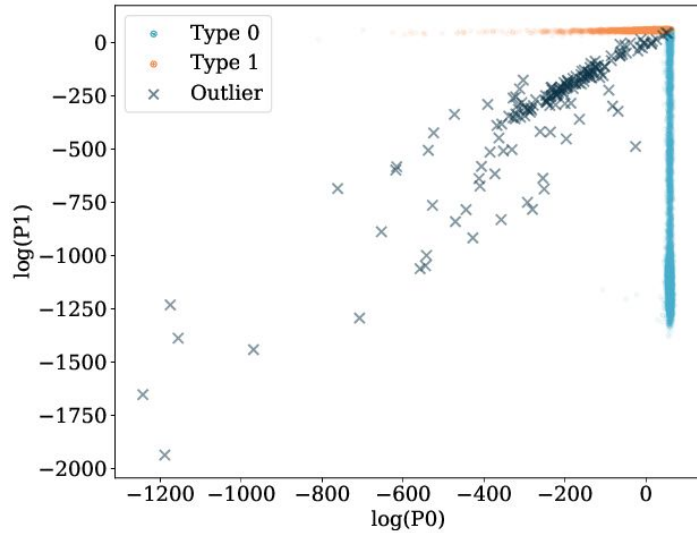


Isolation Forest (Liu,
Ting & Zhou; 2008)

Local Outlier Factor
(Breunig et al; 2008)

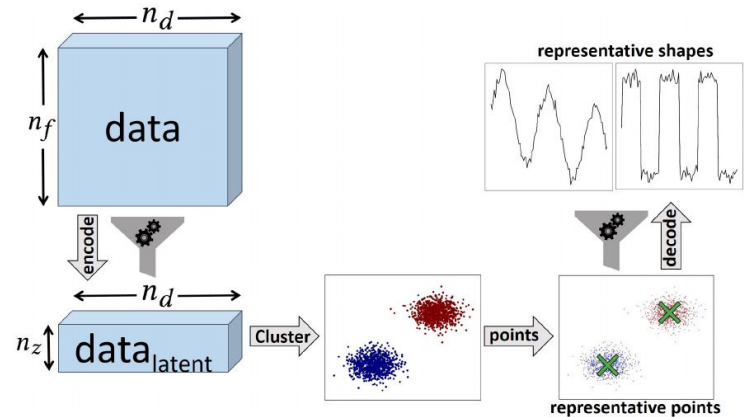


Anomaly Detection Algorithms



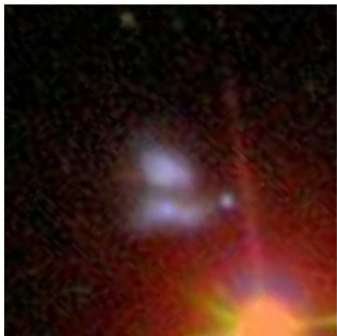
BADAC (Roberts, Bassett & Lochner - 1902.08627)

DRAMA (Vafaei Sadr, Bassett & Kunz - 1909.04060)

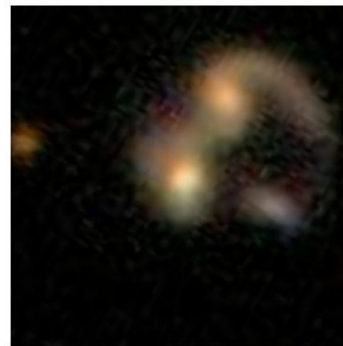


Anomaly Detection Isn't Enough

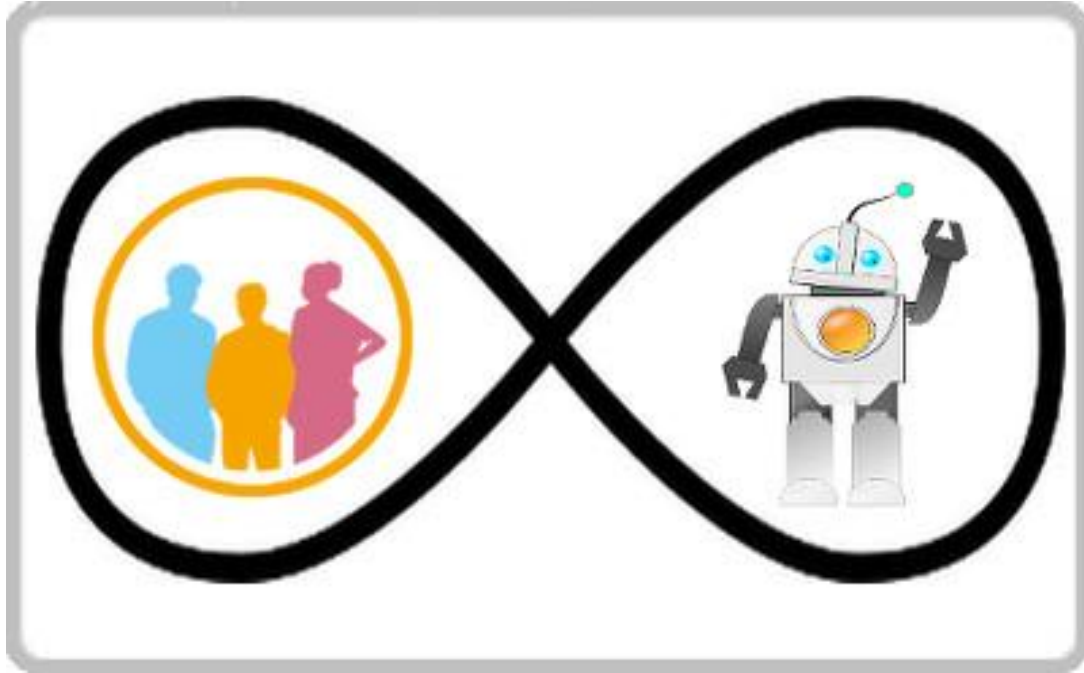
Artefacts



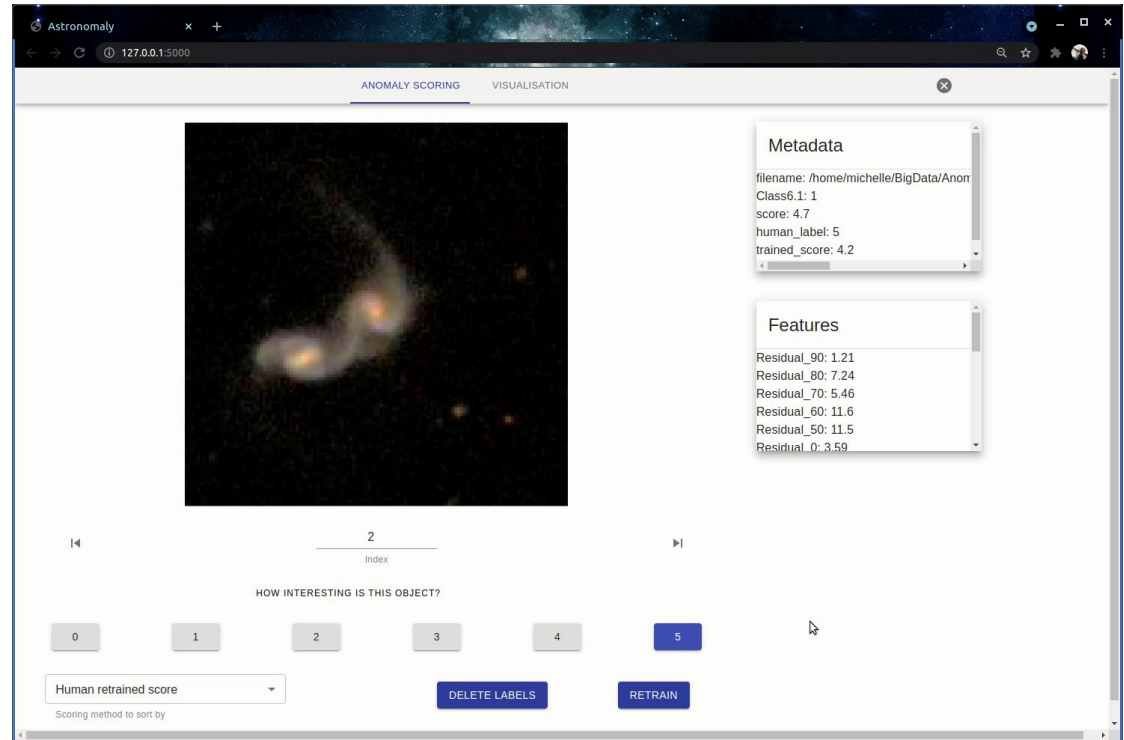
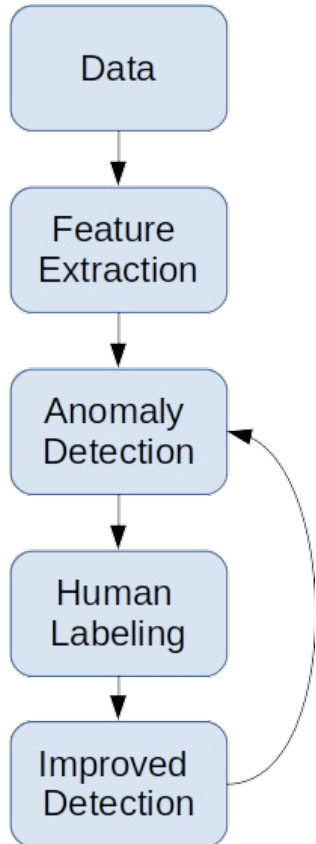
Real sources



Active Learning



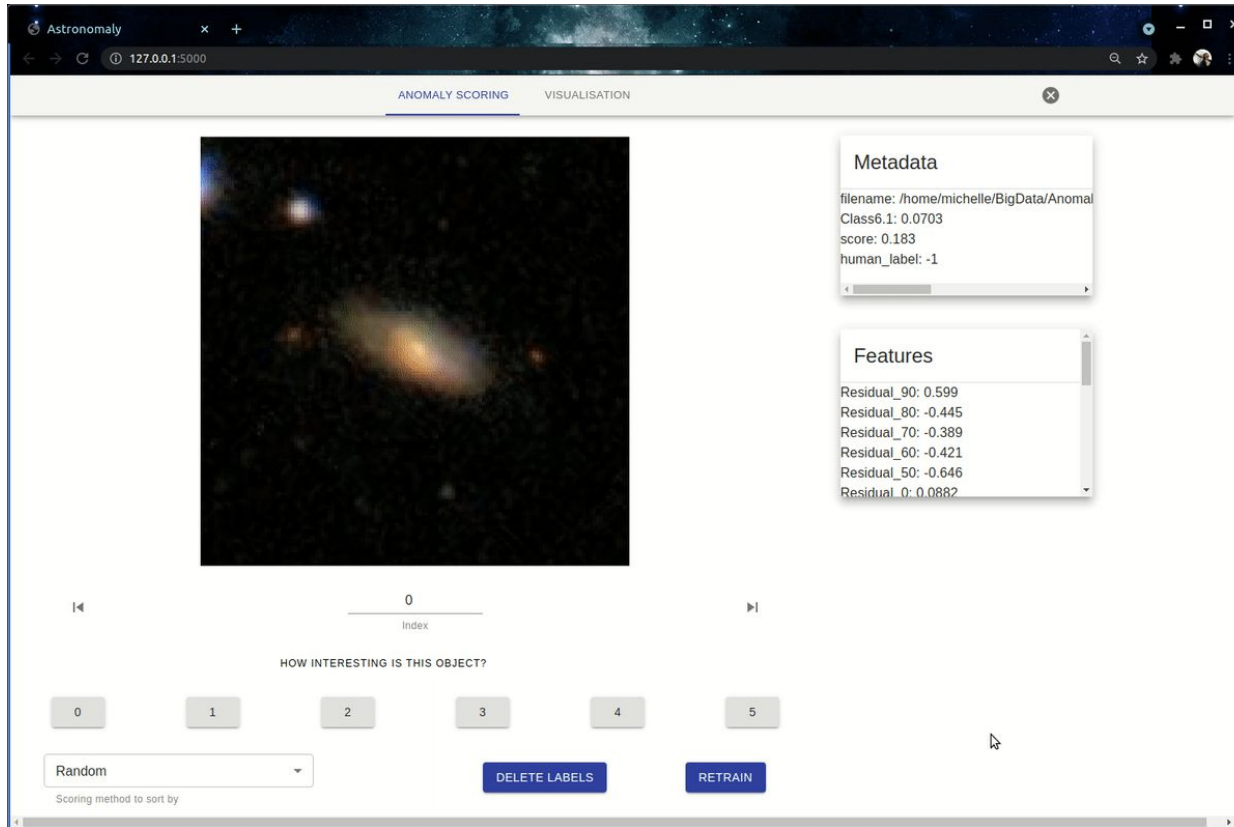
Astronomy



Lochner and Bassett (2020) - [2010.11202](https://arxiv.org/abs/2010.11202)

<https://github.com/MichelleLochner/astronomy>

Galaxy Zoo - Random Examples



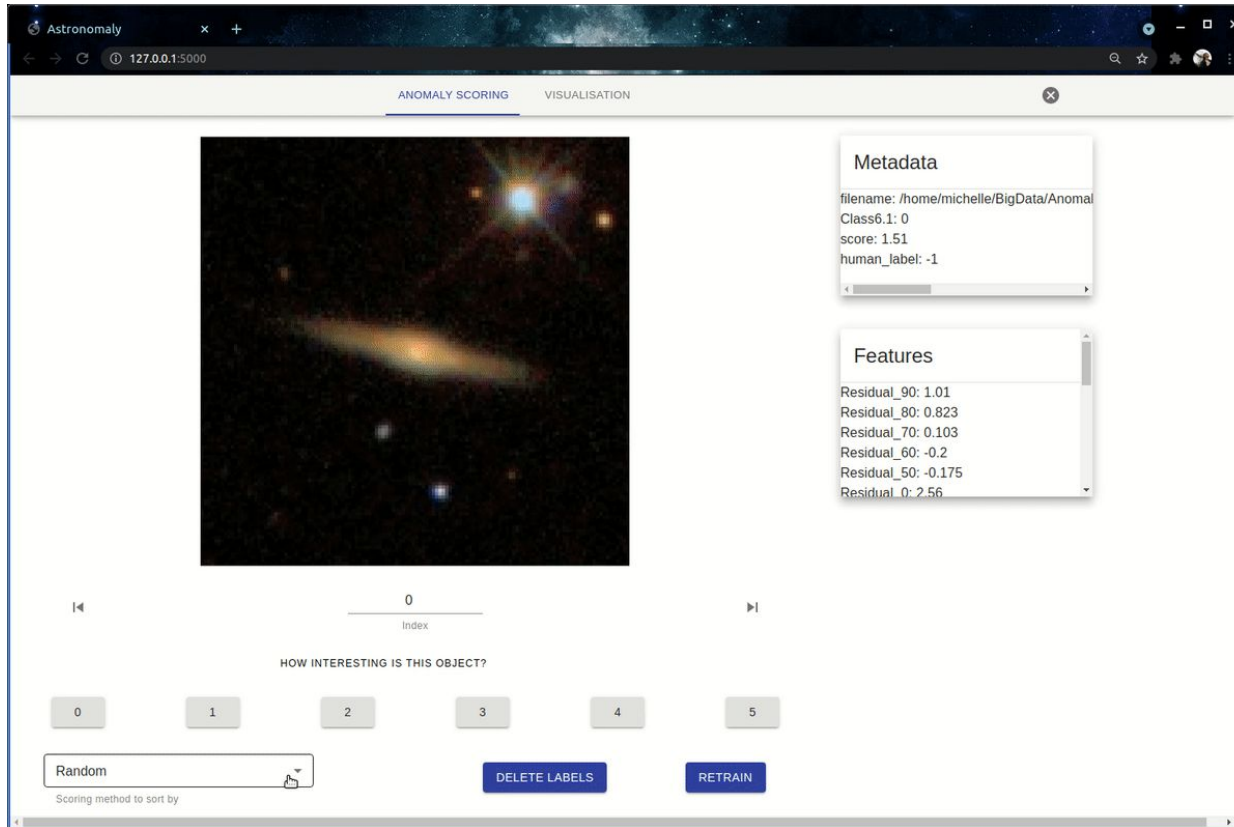
The screenshot displays the Galaxy Zoo web interface. The browser address bar shows the URL 127.0.0.1:5000. The interface is divided into two tabs: "ANOMALY SCORING" (active) and "VISUALISATION".

The main content area features a large image of a galaxy. To the right of the image are two panels:

- Metadata:**
 - filename: /home/michelle/BigData/Anomal
 - Class6.1: 0.0703
 - score: 0.183
 - human_label: -1
- Features:**
 - Residual_90: 0.599
 - Residual_80: -0.445
 - Residual_70: -0.389
 - Residual_60: -0.421
 - Residual_50: -0.646
 - Residual_0: 0.0882

Below the image is a navigation bar with a "0" index and "Index" label. Below that is a question: "HOW INTERESTING IS THIS OBJECT?". Below the question are five buttons labeled "0", "1", "2", "3", "4", and "5". At the bottom left is a dropdown menu set to "Random" with the text "Scoring method to sort by" below it. At the bottom center are two buttons: "DELETE LABELS" and "RETRAIN".

Galaxy Zoo - Machine Learning



The screenshot displays the 'ANOMALY SCORING' interface within a browser window titled 'Astronomy'. The URL is 127.0.0.1:5000. The interface is split into two tabs: 'ANOMALY SCORING' (active) and 'VISUALISATION'. The main area shows a central image of a galaxy. To the right, there are two panels: 'Metadata' and 'Features'. Below the image, there is a slider for 'Index' (set to 0) and a question 'HOW INTERESTING IS THIS OBJECT?' with buttons for ratings 0 through 5. At the bottom, there is a 'Random' button, a 'Scoring method to sort by' dropdown, and 'DELETE LABELS' and 'RETRAIN' buttons.

Metadata

- filename: /home/michelle/BigData/Anomal
- Class6.1: 0
- score: 1.51
- human_label: -1

Features

- Residual_90: 1.01
- Residual_80: 0.823
- Residual_70: 0.103
- Residual_60: -0.2
- Residual_50: -0.175
- Residual_0: 2.56

0
Index

HOW INTERESTING IS THIS OBJECT?

0 1 2 3 4 5

Random

Scoring method to sort by

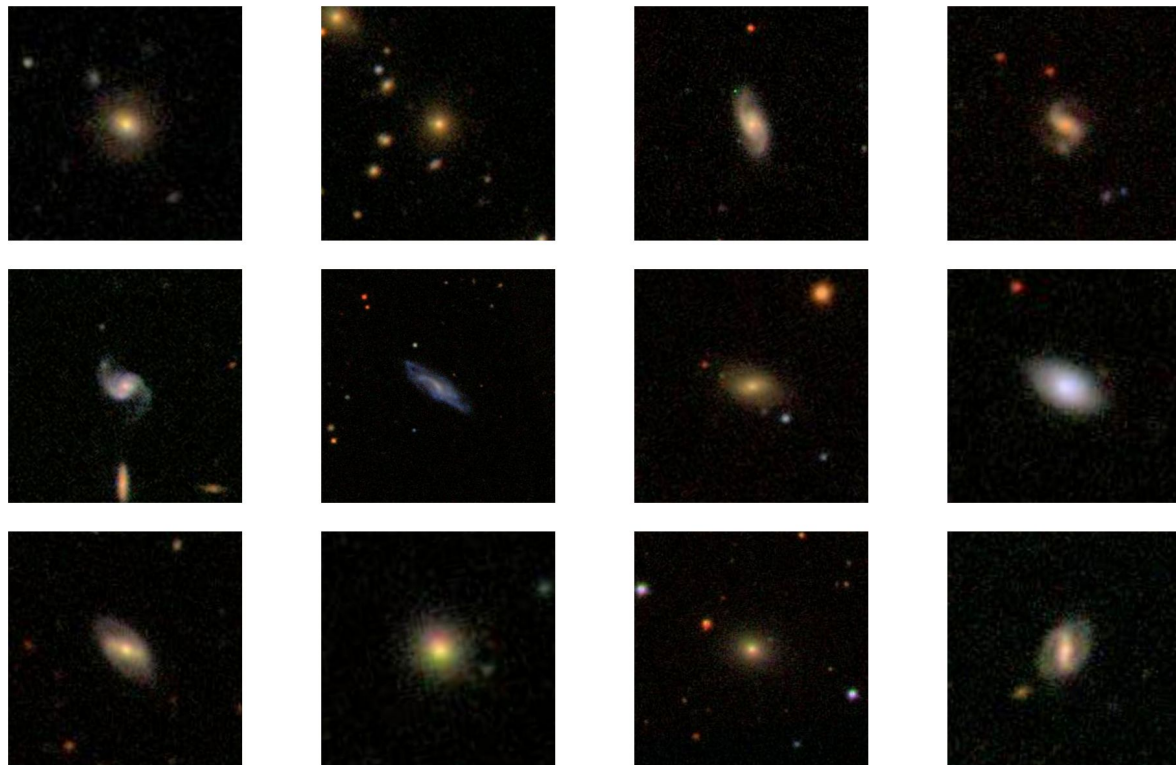
DELETE LABELS RETRAIN

Galaxy Zoo - Active Learning

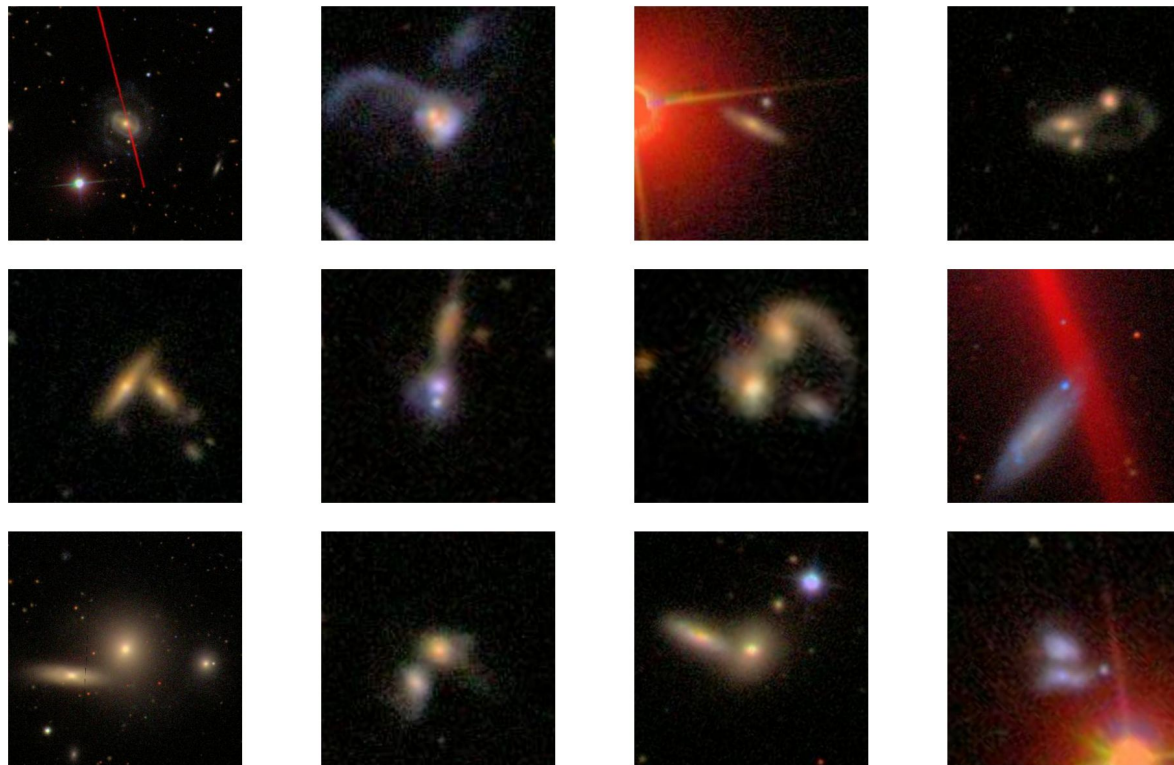
The screenshot displays the 'Astronomy' web application interface, which is used for active learning in galaxy classification. The interface is divided into several sections:

- Navigation:** The browser address bar shows '127.0.0.1:5000'. The application has two tabs: 'ANOMALY SCORING' (active) and 'VISUALISATION'.
- Image:** A central image of a galaxy is displayed.
- Metadata:** A panel on the right shows the following information:
 - filename: /home/michelle/BigData/Anorr
 - Class6.1: 0.916
 - score: 5
 - human_label: 5
 - trained_score: 4.53
- Features:** A panel on the right shows the following Residual values:
 - Residual_90: 2.58
 - Residual_80: 23.2
 - Residual_70: 12.9
 - Residual_60: 8.1
 - Residual_50: 5.38
 - Residual_0: 3.19
- Index:** A slider below the image shows the current 'Index' is 0.
- Interest Rating:** A row of buttons labeled '0' through '5' is used to rate the object's interest. The '5' button is currently selected.
- Actions:** A dropdown menu for 'Human retrained score' is set to '0'. Below it are buttons for 'DELETE LABELS' and 'RETRAIN'.

Galaxy Zoo - Random Examples



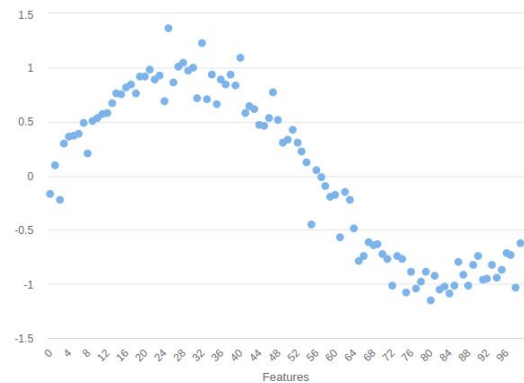
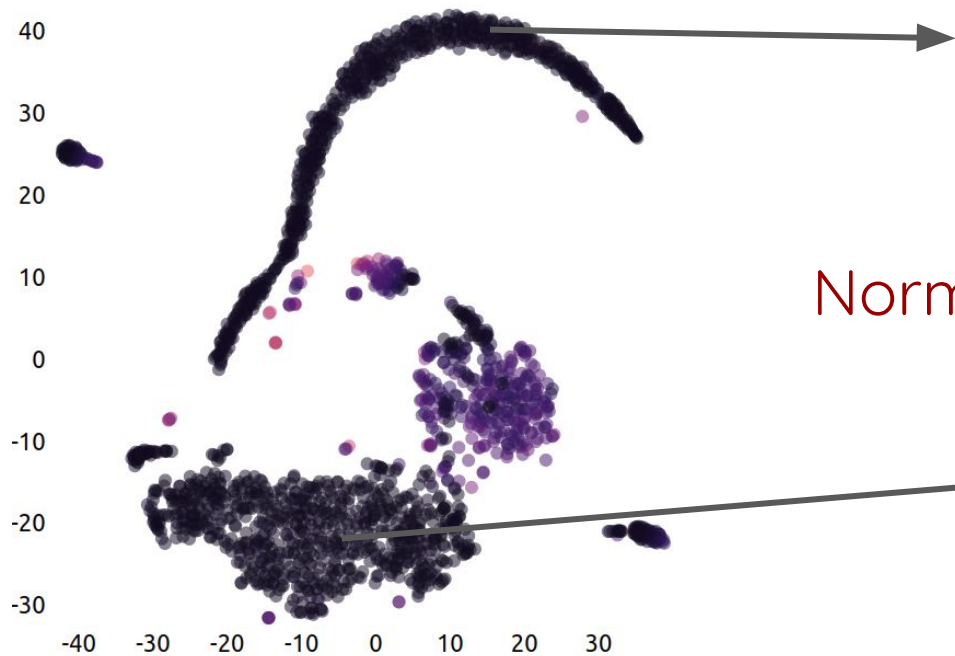
Galaxy Zoo - No active Learning



Galaxy Zoo - Active Learning

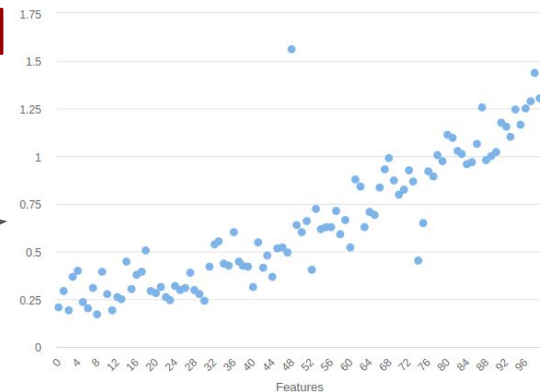


Synthetic data



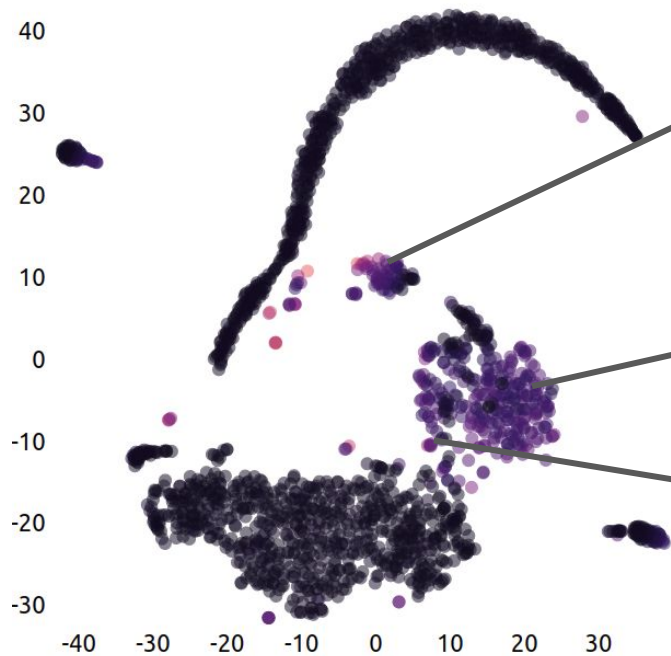
Anomaly Score: 0.270122

Normal

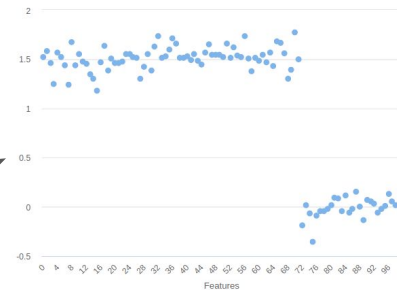


Anomaly Score: 0.241527

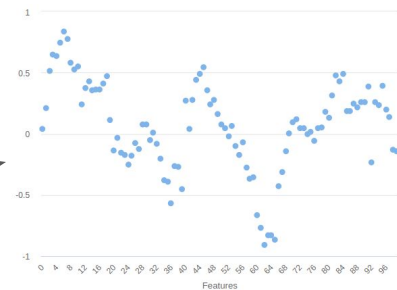
Synthetic data



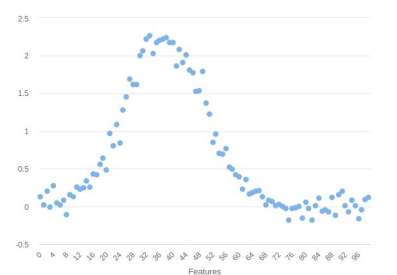
Anomalous



Anomaly Score: 3.112012

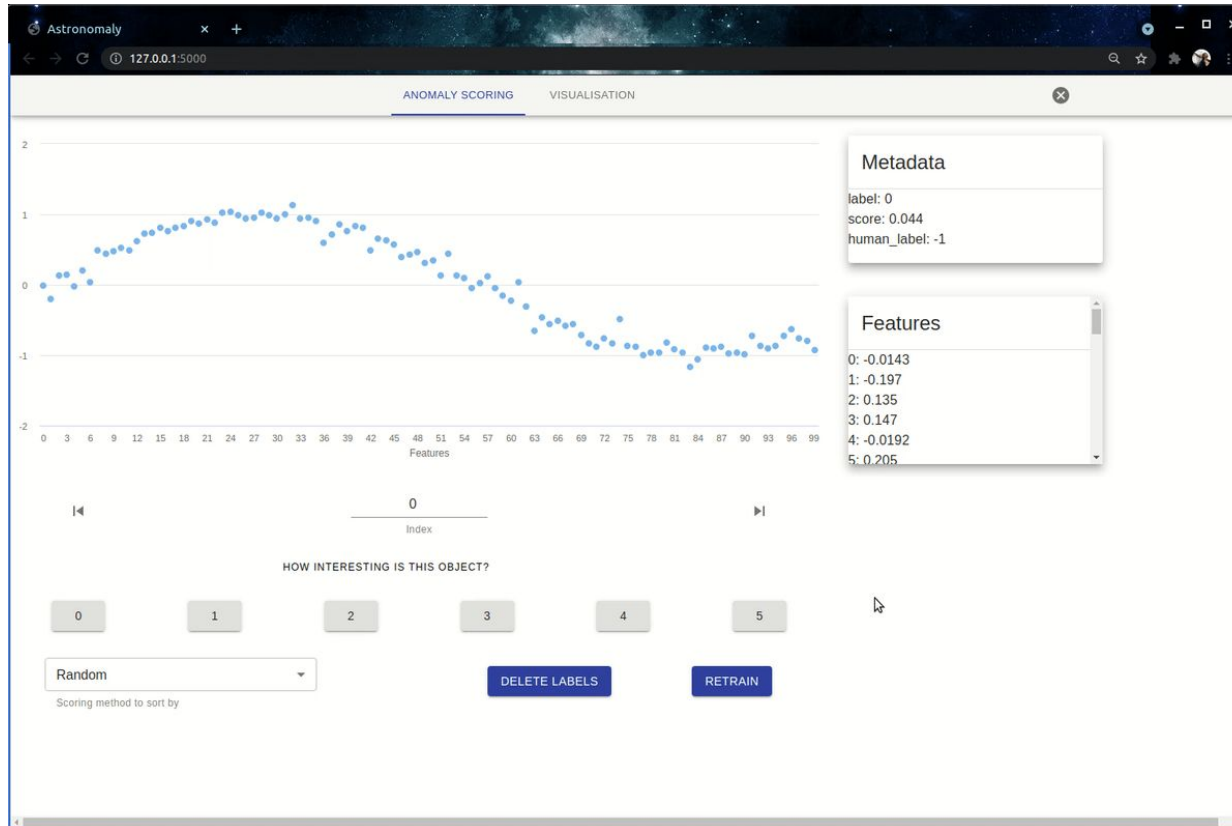


Anomaly Score: 1.303772

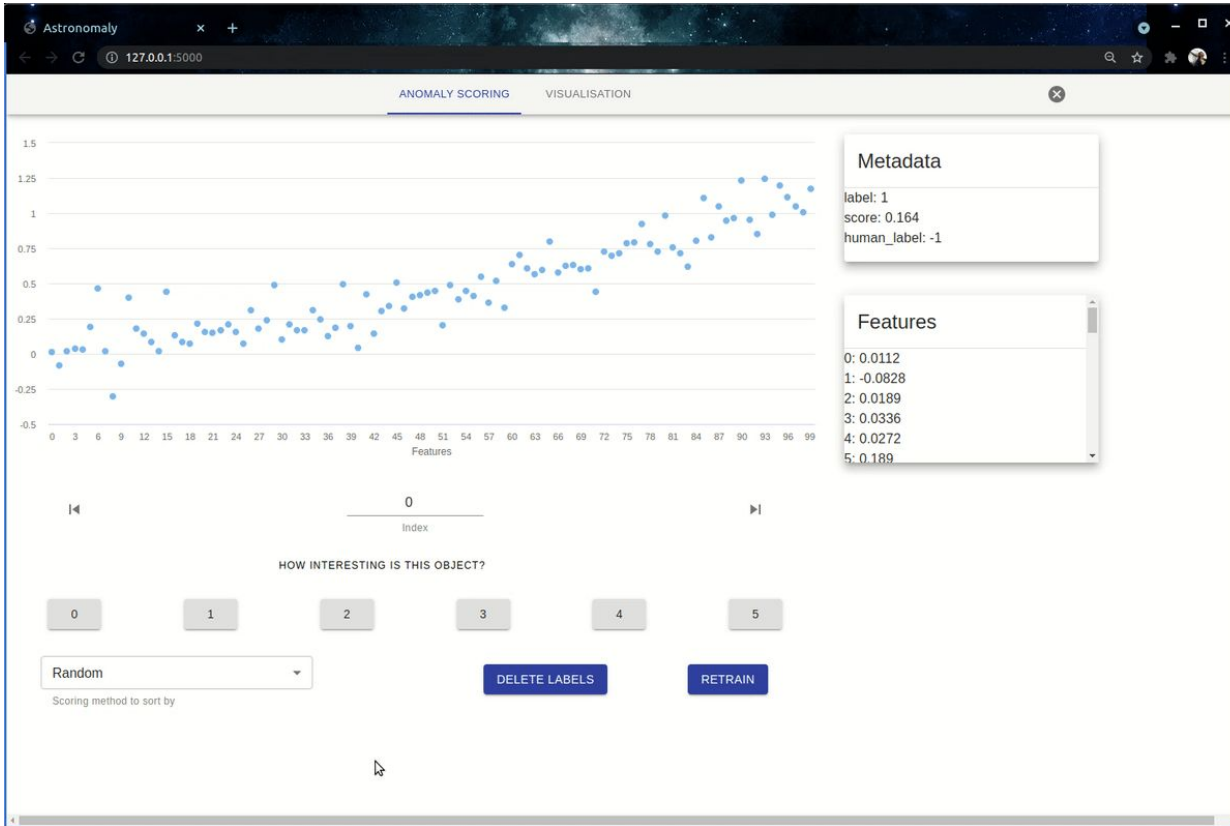


Anomaly Score: 3.92992

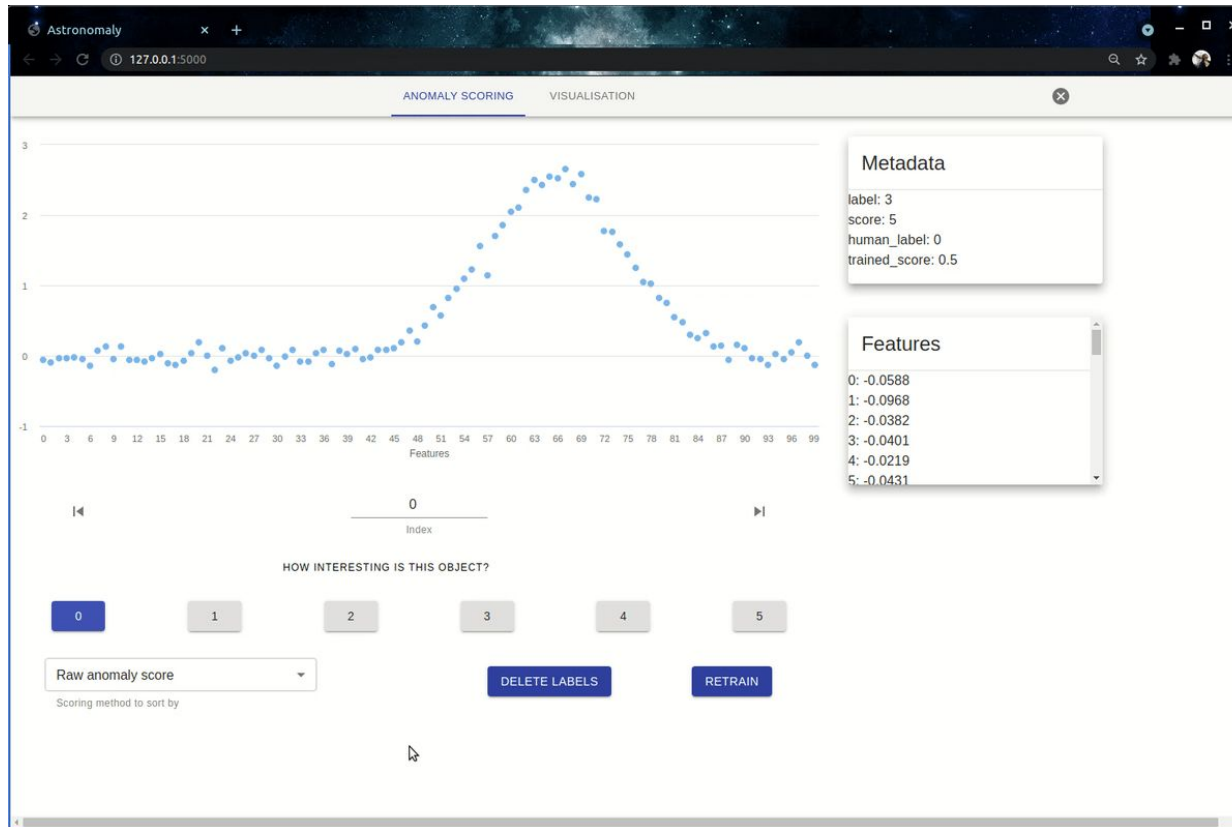
Synthetic data - Random



Synthetic data - No active Learning

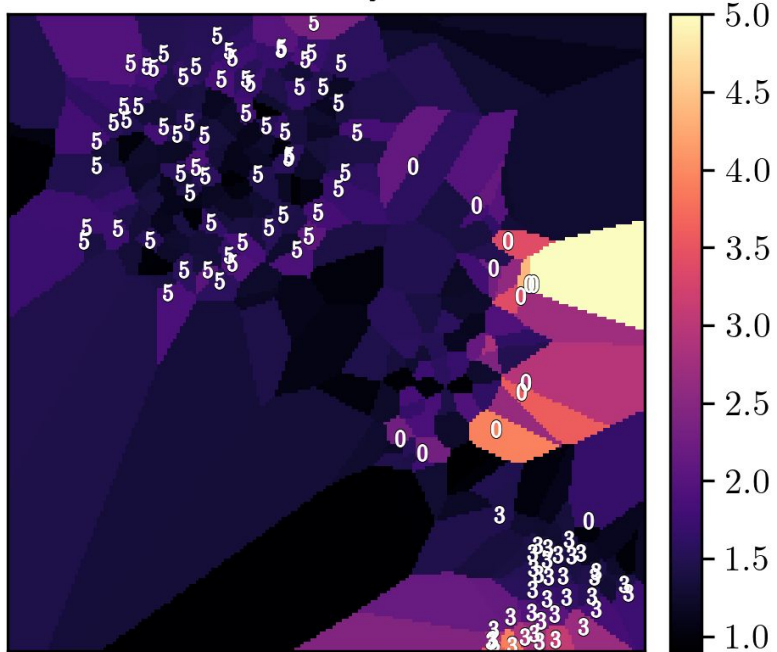


Synthetic data - Active Learning

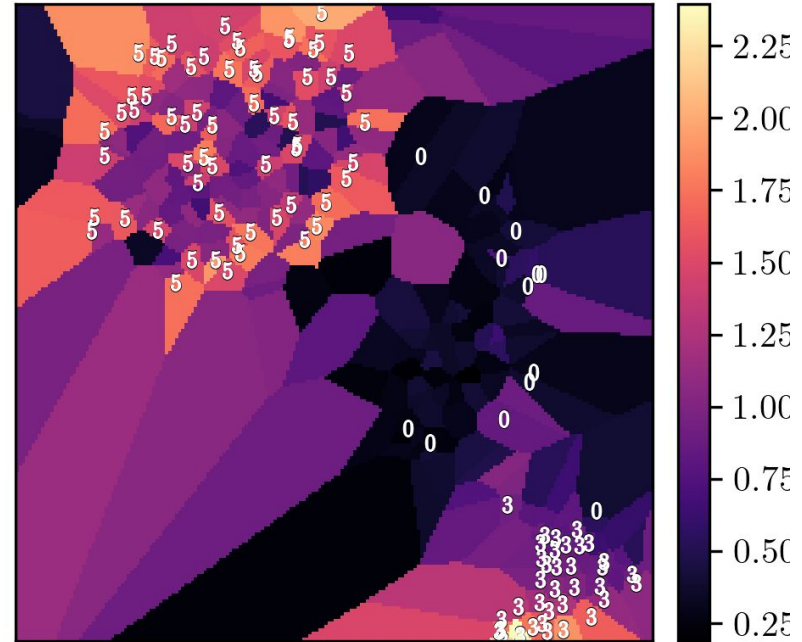


Visualisation with Synthetic Data

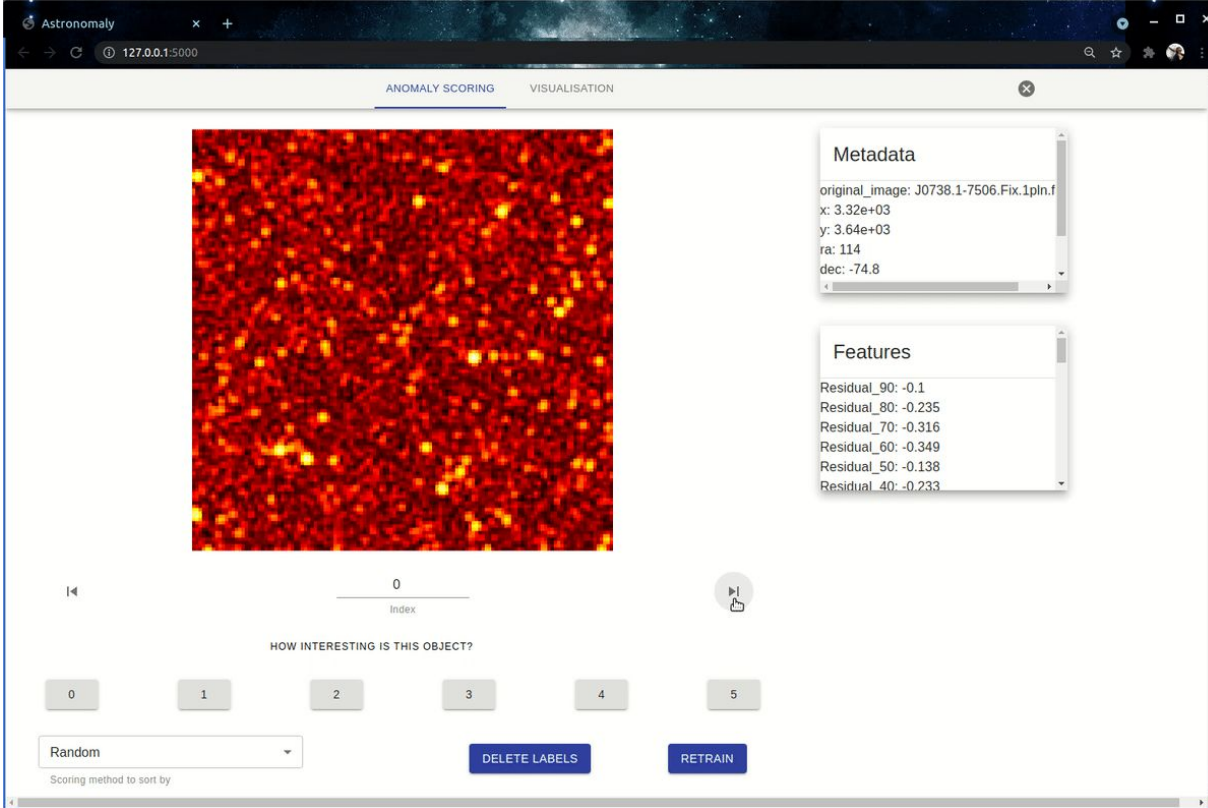
Raw anomaly score



Trained anomaly score



MeerKAT Data - Random Examples



The screenshot displays the 'ANOMALY SCORING' interface of the Astronomy application. The main visualization is a square field of stars, with a central region highlighted in a darker red, indicating an anomaly. The interface includes a 'Metadata' panel on the right with the following information:

```
original_image: J0738.1-7506.Fix.1pln.f
x: 3.32e+03
y: 3.64e+03
ra: 114
dec: -74.8
```

Below the metadata is a 'Features' panel with the following data:

```
Residual_90: -0.1
Residual_80: -0.235
Residual_70: -0.316
Residual_60: -0.349
Residual_50: -0.138
Residual_40: -0.233
```

At the bottom of the interface, there is a slider labeled 'HOW INTERESTING IS THIS OBJECT?' with a value of 0. Below the slider are five buttons labeled 0, 1, 2, 3, 4, and 5. A dropdown menu is set to 'Random', and there are two buttons: 'DELETE LABELS' and 'RETRAIN'. The text 'Scoring method to sort by' is visible below the dropdown menu.

MeerKAT Data - Anomalies

The screenshot shows a web browser window with the URL 127.0.0.1:5000. The page has two tabs: "ANOMALY SCORING" (active) and "VISUALISATION". The main content area displays a large, square, red-toned astronomical image filled with numerous small, bright yellow and orange spots, representing a field of stars or galaxies. Below the image, there is a navigation bar with a left arrow, a slider set to "10" with "Index" below it, and a right arrow with a share icon. Below the navigation bar, the text "HOW INTERESTING IS THIS OBJECT?" is followed by five buttons labeled "0", "1", "2", "3", "4", and "5". At the bottom left, there is a dropdown menu currently set to "Random" with the text "Scoring method to sort by" below it. At the bottom center, there are two blue buttons: "DELETE LABELS" and "RETRAIN". On the right side of the interface, there are two panels: "Metadata" and "Features".

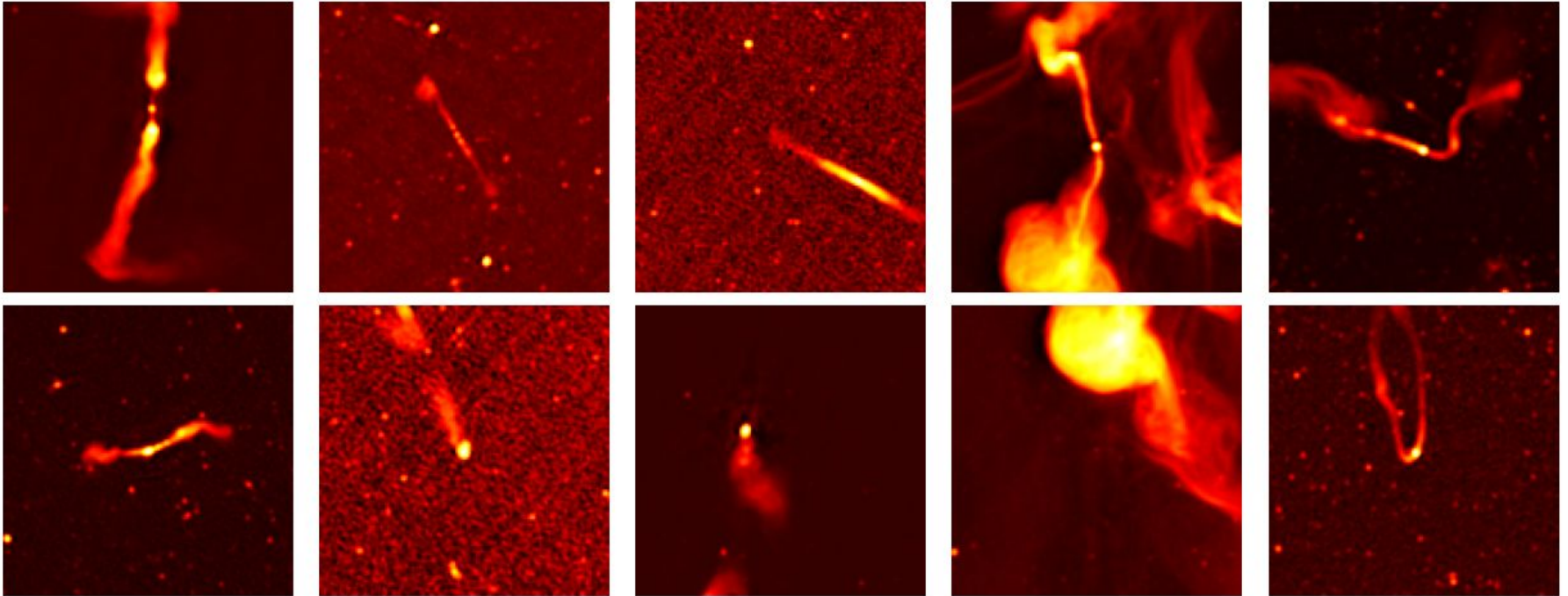
Metadata

- original_image: J0314.3-4525.Fix.1pln.f
- x: 1.46e+03
- y: 2.6e+03
- ra: 49.3
- dec: -45.5

Features

- Residual_90: -0.257
- Residual_80: -0.249
- Residual_70: 0.0513
- Residual_60: -0.0836
- Residual_50: -0.189
- Residual_40: -0.239

MeerKAT Galaxy Cluster Legacy Survey



Anomalies in DECAALS

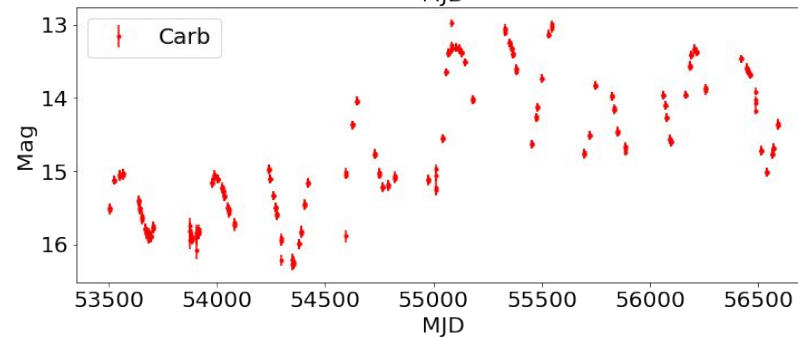
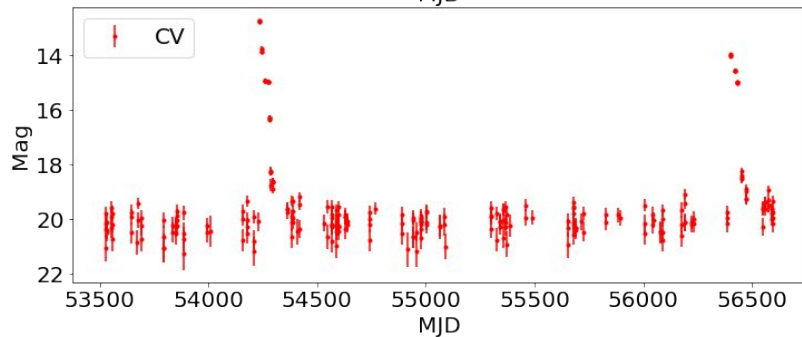
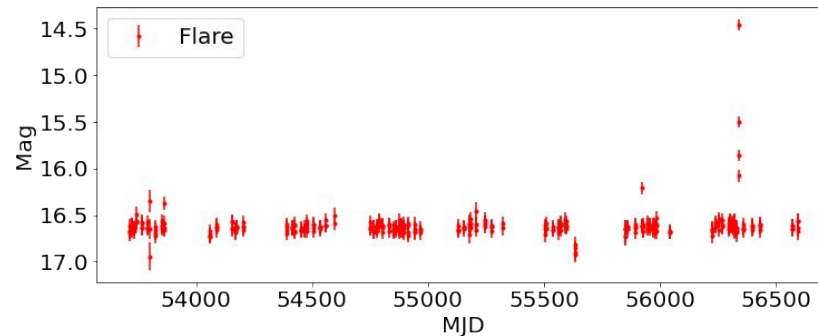
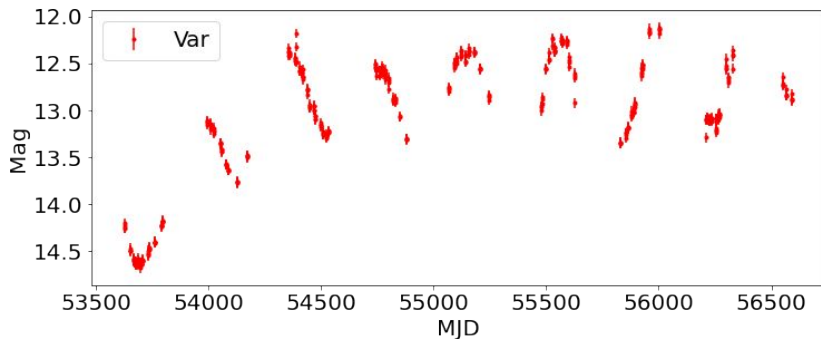
Verlon Etsebeth (MSc student)



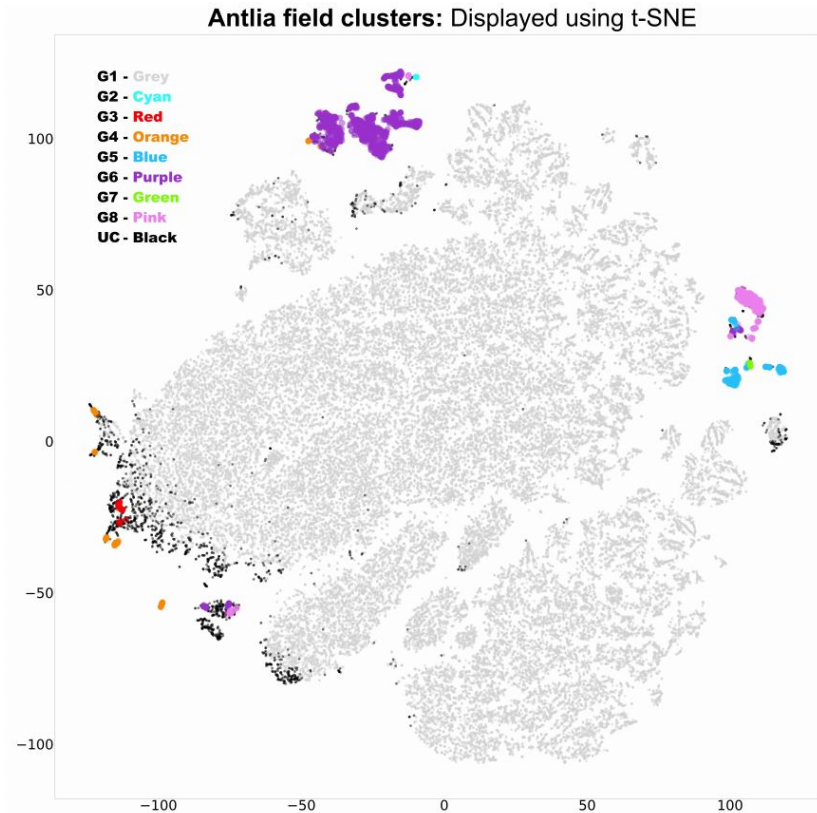
Anomalous Transients



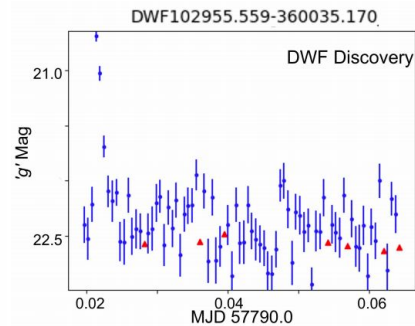
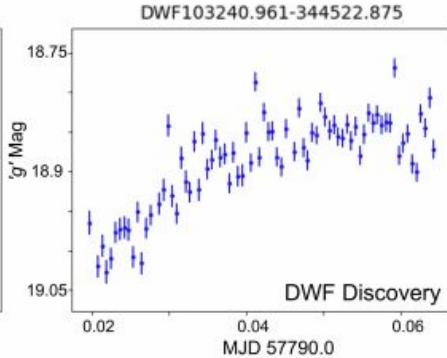
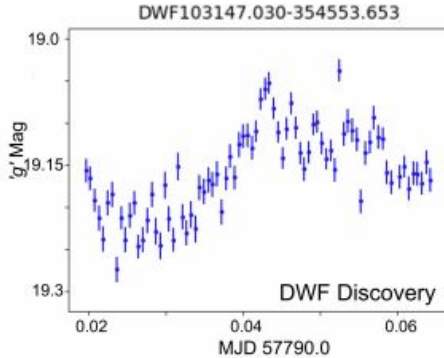
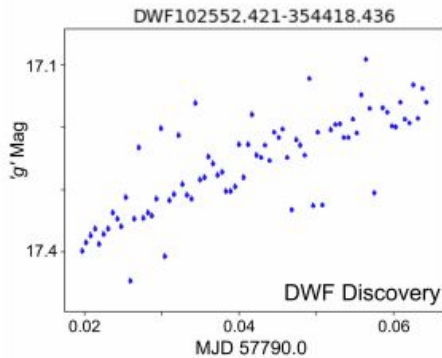
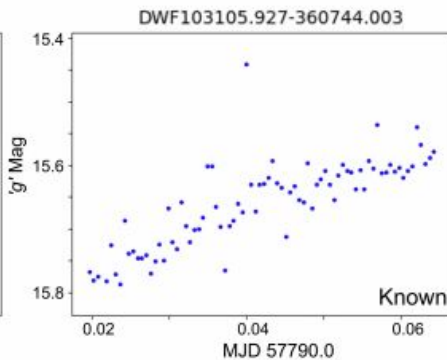
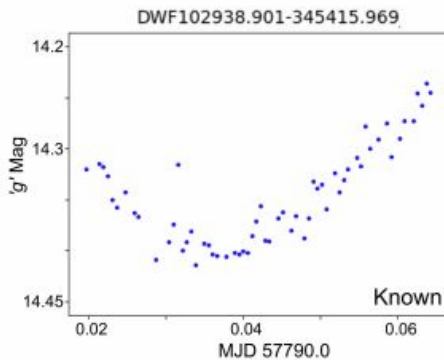
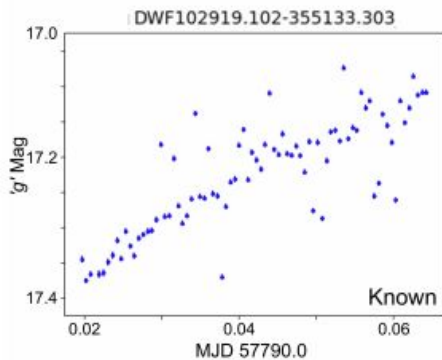
Malema Ramonyai (MSc student)



Astronomy Applied to DWF

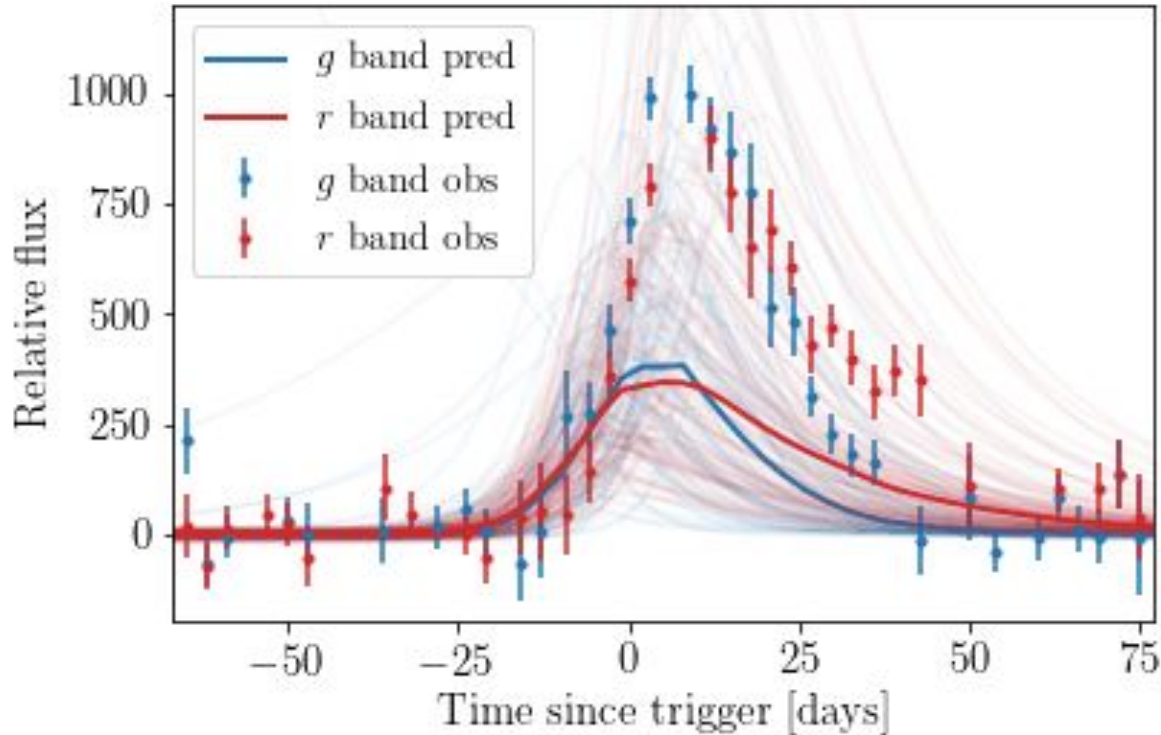


Astrometry Applied to DWF



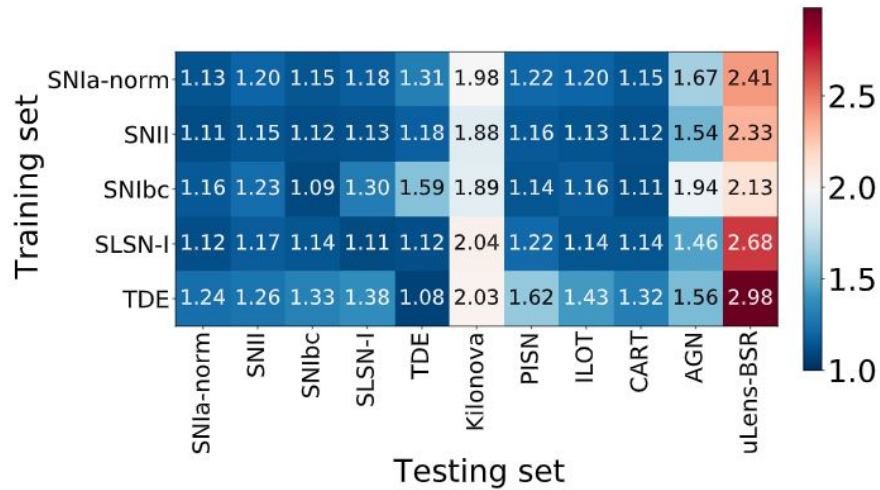
Real-time anomaly detection

Real-time anomaly detection

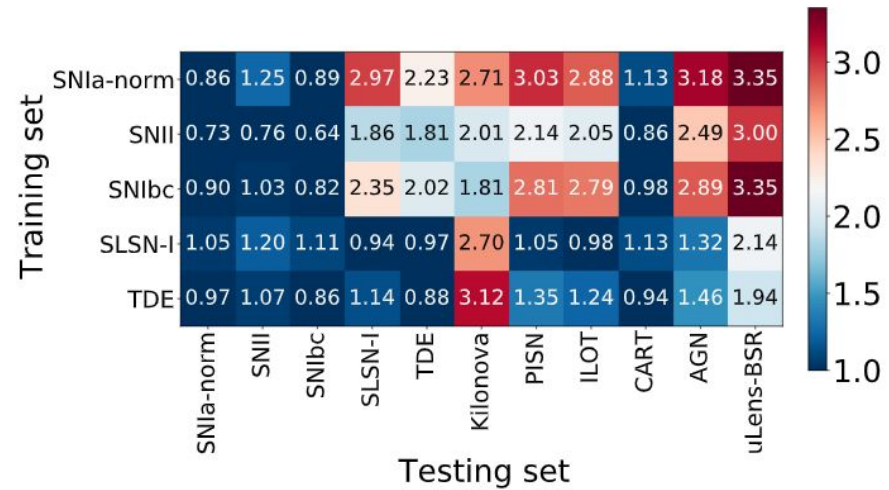


Real-time anomaly detection

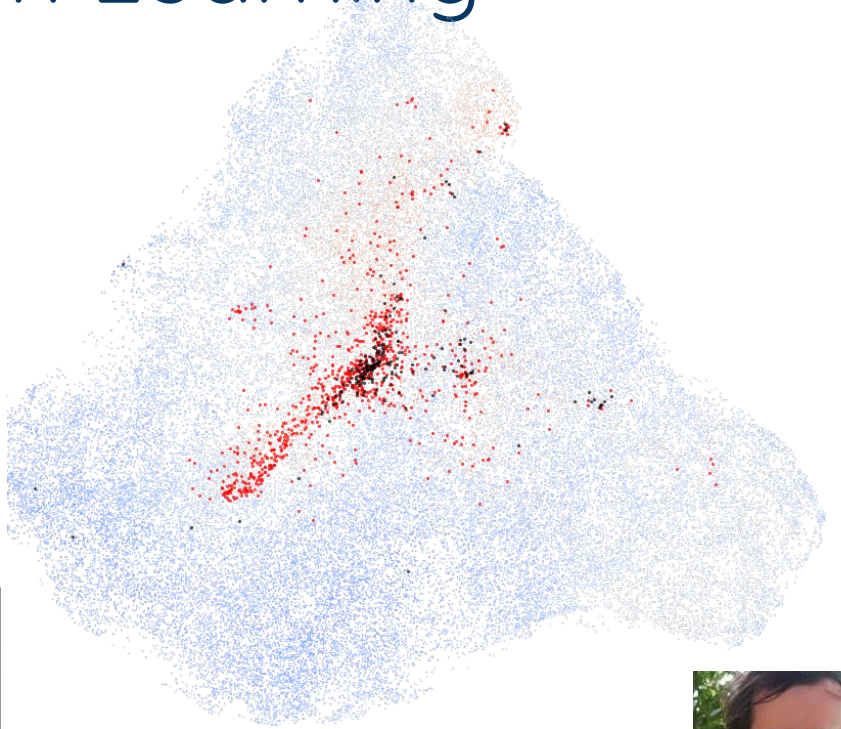
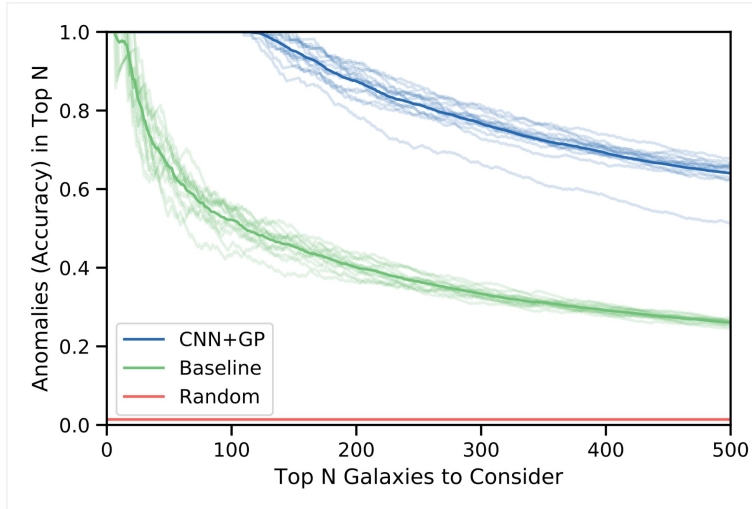
DNN



Bazin



Deep Representation Learning

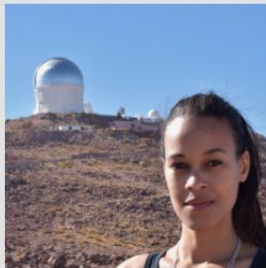


SANGEETA UJJWAL



Country: India
Research Field: Physics
Institute: University of Delhi
Position: Postdoc

SATYA GONTCHO A GONTCHO



Country: United States of America
Research Field: Astrophysics and Cosmology
Institute: Lawrence Berkeley National Laboratory
Position: Project Scientist

SHAZRENE S. MOHAMED



Country: South Africa
Research Field: Astronomy
Institute: South African Astronomical Observatory and University of Cape Town
Position: Faculty

SIPHEPHILE NCUBE



Country: South Africa
Research Field: Physics
Institute: University of the Witwatersrand
Position: Postdoc

SIYI ZHOU



Country: Sweden
Research Field: Theoretical Physics
Institute: Stockholm University
Position: Postdoc

SUDESHNA BORO SAIKIA



Country: Austria
Research Field: Astrophysics and Cosmology
Institute: University of Vienna
Position: Post doctoral researcher

SWARNAMALA SIRSI



Country: India
Research Field: Theoretical Physics
Institute: University of Mysore
Position: Associate Professor (retired)

VALERIA PETTORINO



Country: France
Research Field: Astrophysics and Cosmology
Institute: CEA
Position: STAFF Scientist

Conclusions

- Machine learning is critical in facing the data deluge
- We need automated anomaly detection to ensure scientific discoveries in datasets aren't missed
- Check out Astronomy:
 - <https://arxiv.org/abs/2010.11202>
 - <https://github.com/MichelleLochner/astronomy>
- And the Supernova Foundation:
 - <https://www.supernovafoundation.org/>

A Novel Active Learning Approach

$$\hat{S} = S \tanh(\delta - 1 + \operatorname{arctanh}(\tilde{U}))$$

A Novel Active Learning Approach

ML Anomaly Score

Predicted User Anomaly Score

$$\hat{S} = S \tanh(\delta - 1 + \operatorname{arctanh}(\tilde{U}))$$

Distance penalty term

A Novel Active Learning Approach

The diagram illustrates the components of the equation for the estimated anomaly score \hat{S} . It features four callout boxes with arrows pointing to parts of the equation:

- ML Anomaly Score**: Points to the variable S in the equation.
- Predicted User Anomaly Score**: Points to the variable \tilde{U} in the equation.
- Distance penalty term**: Points to the variable δ in the equation.
- Distance to nearest label**: Points to the variable d in the exponent of the equation.

$$\hat{S} = S \tanh\left(\delta - 1 + \operatorname{arctanh}(\tilde{U})\right)$$
$$\delta = \exp\left(\alpha \frac{d}{d_0}\right)$$

Galaxy Zoo - Active Learning

