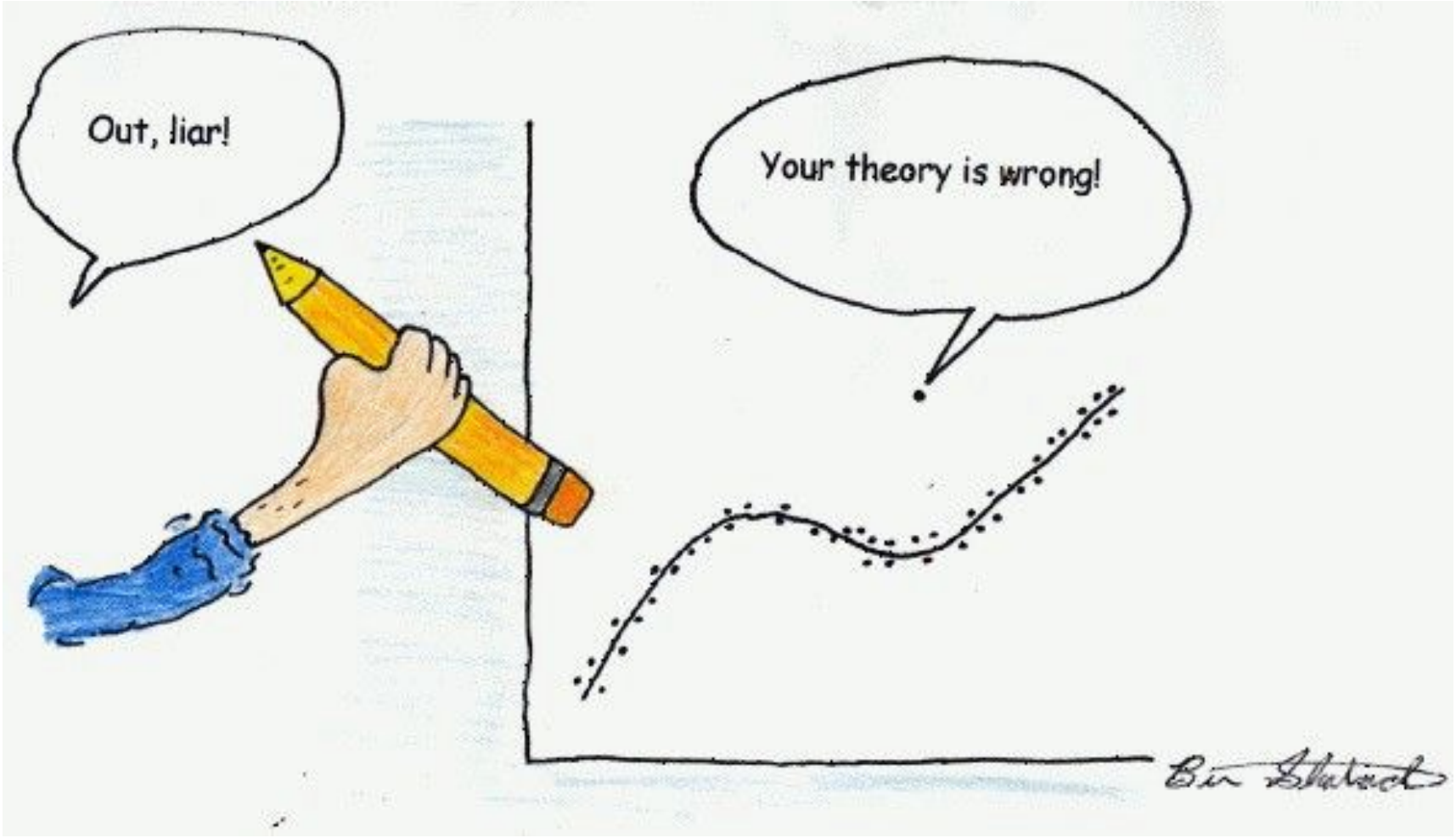
Active Anomaly Detection

TransiXplore, 11 October 2019
LPC - Clermont Ferrand, France

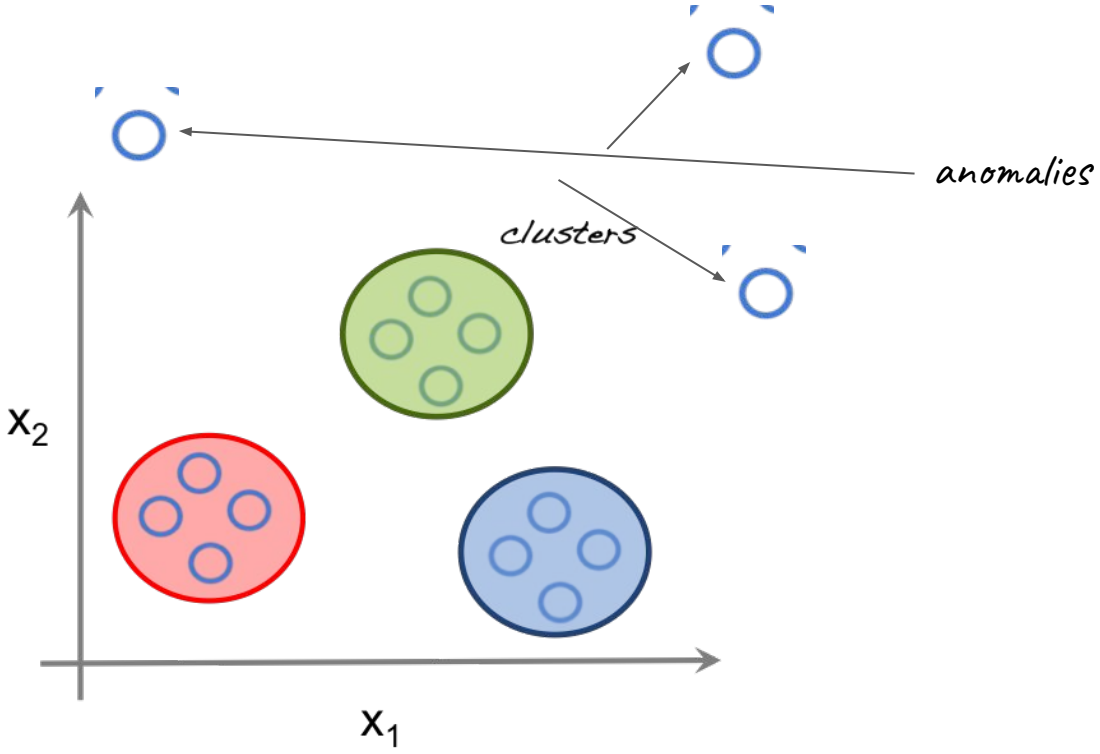
Emille E. O. Ishida
Laboratoire de Physique de Clermont - Université Clermont-Auvergne
Clermont Ferrand, France

SNAD

Stranger things ...



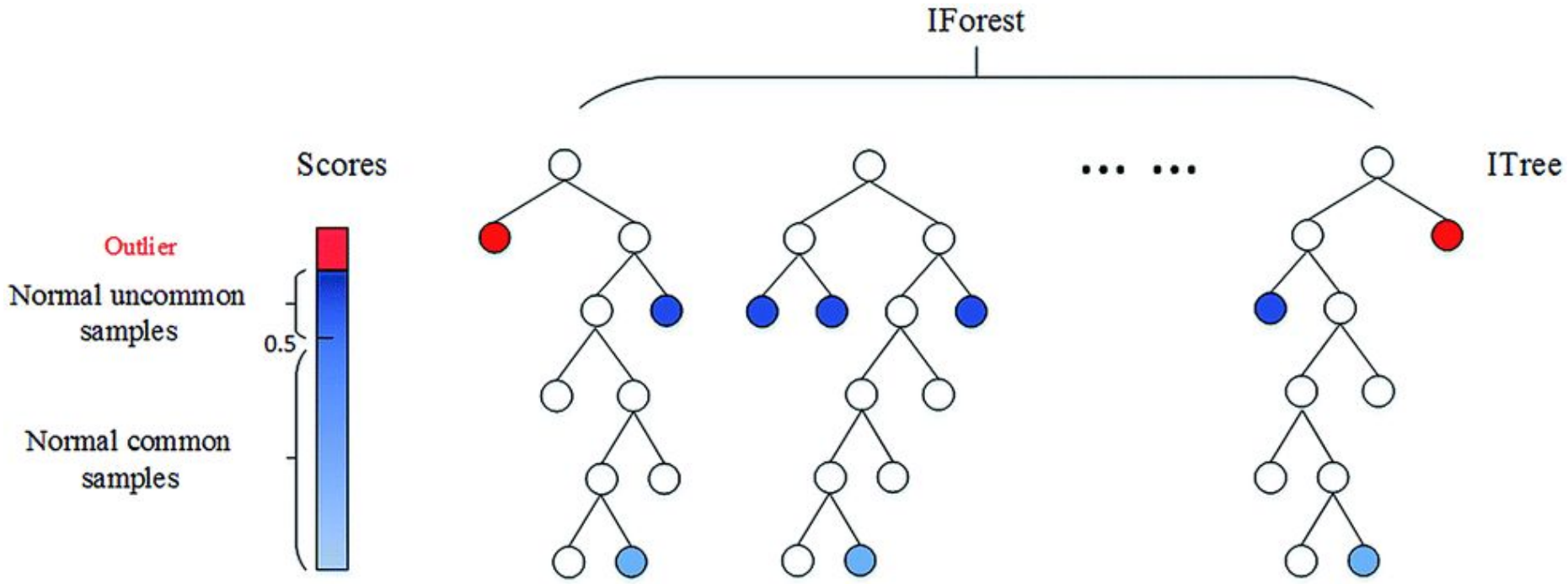
Anomaly Detection



"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Hawkins, 1980

Isolation Forest

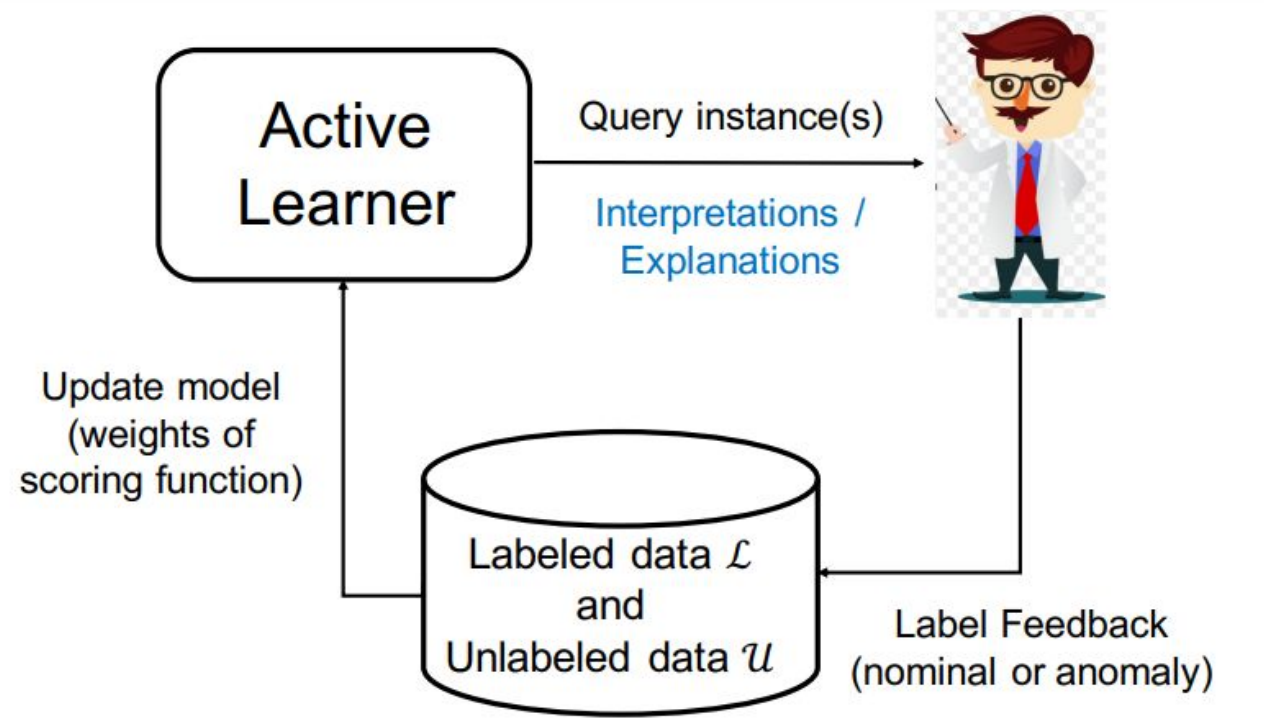


Problem: statistical/numerical anomalies might not be scientific anomalies.

https://donghwa-kim.github.io/iforest.html

Active Learning

or Optimum experimental Design



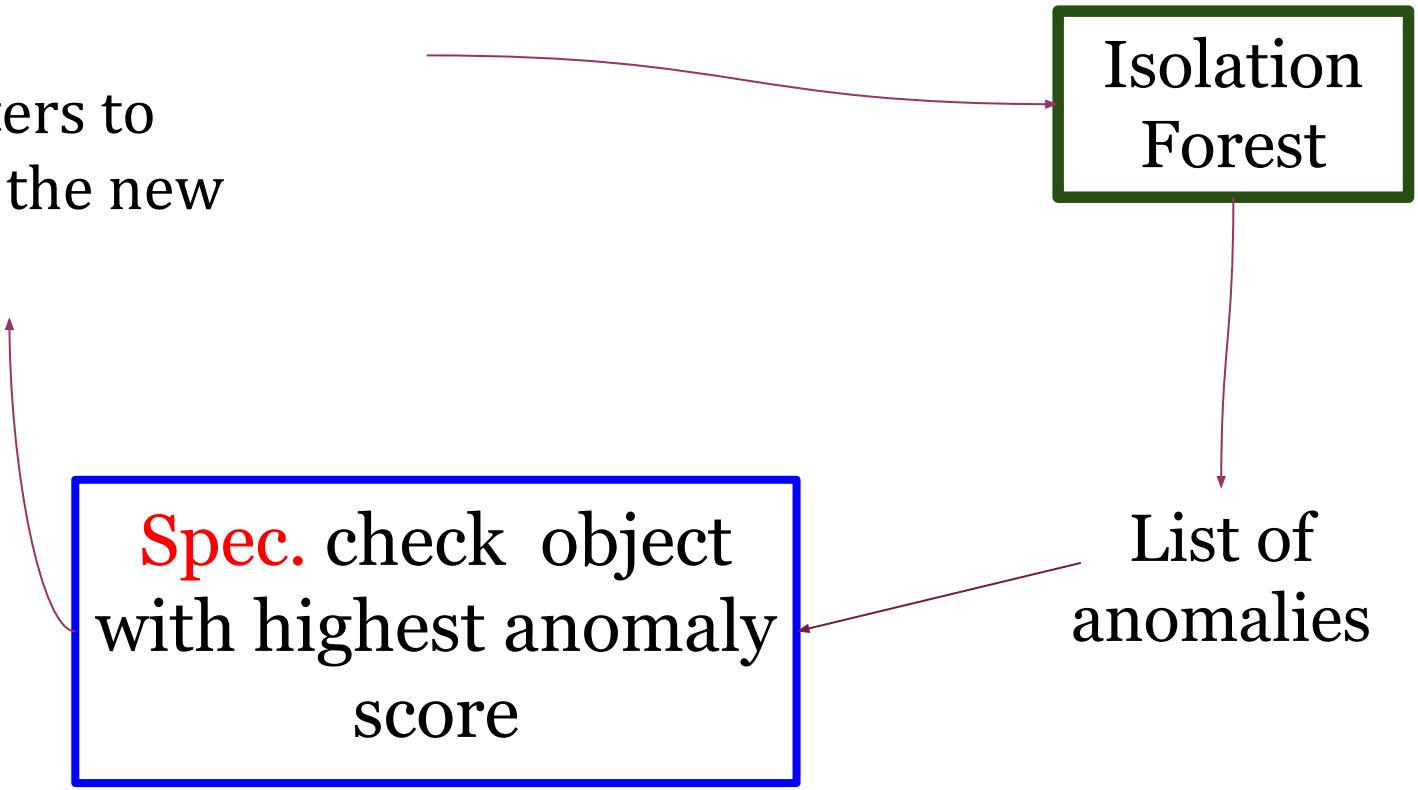
GOAL: Maximize the number of true anomalies presented to the user.

Active Anomaly Detection

A strategy

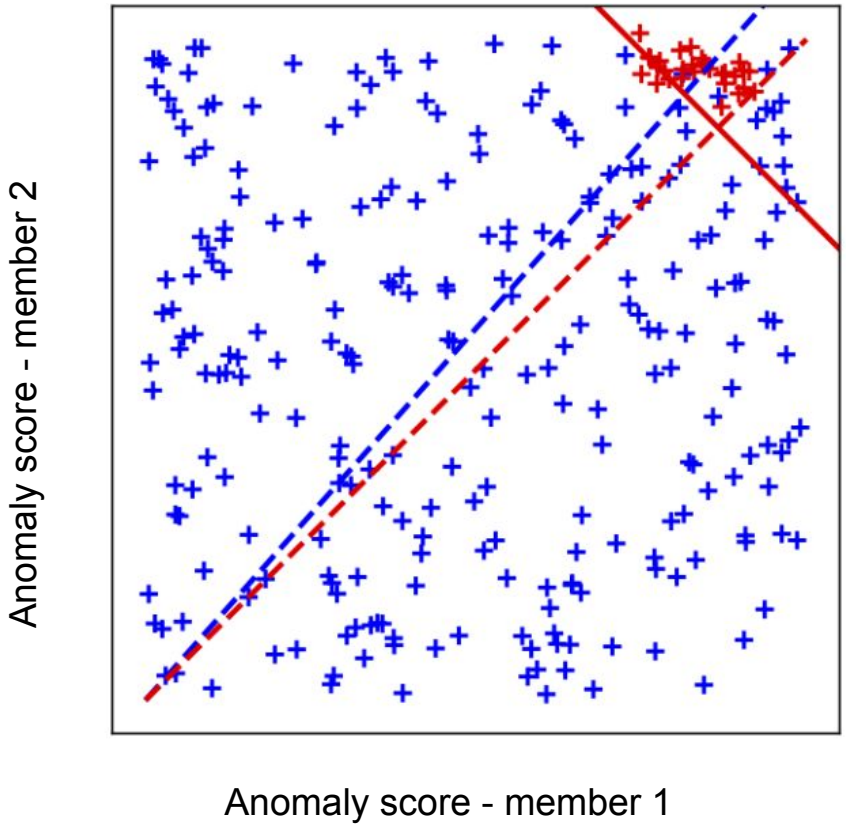
If yes: check next obj in the anomaly score board

If no: update hyperparameters to accommodate the new information



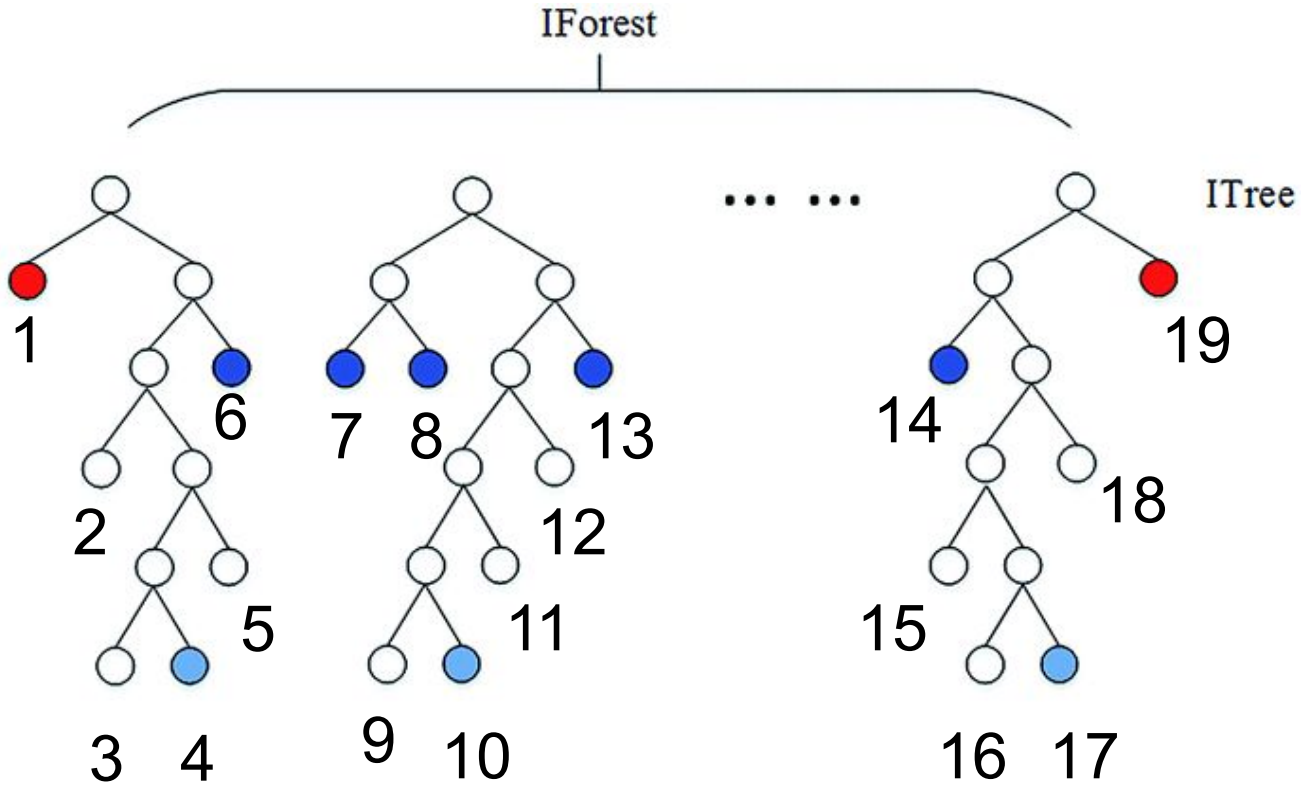
Das, S., Wong, W-K., Dietterich, T., Fern, A. and Emmott, A. (2016). *Incorporating Expert Feedback into Active Anomaly Discovery* in the Proceedings of the IEEE International Conference on Data Mining

Ensemble Anomaly Detection



Ensemble Members ⇛ leaf nodes

wi, i ∈ [1,19]



Active Anomaly Detection

Algorithm 2 Active Anomaly Discovery (AAD)

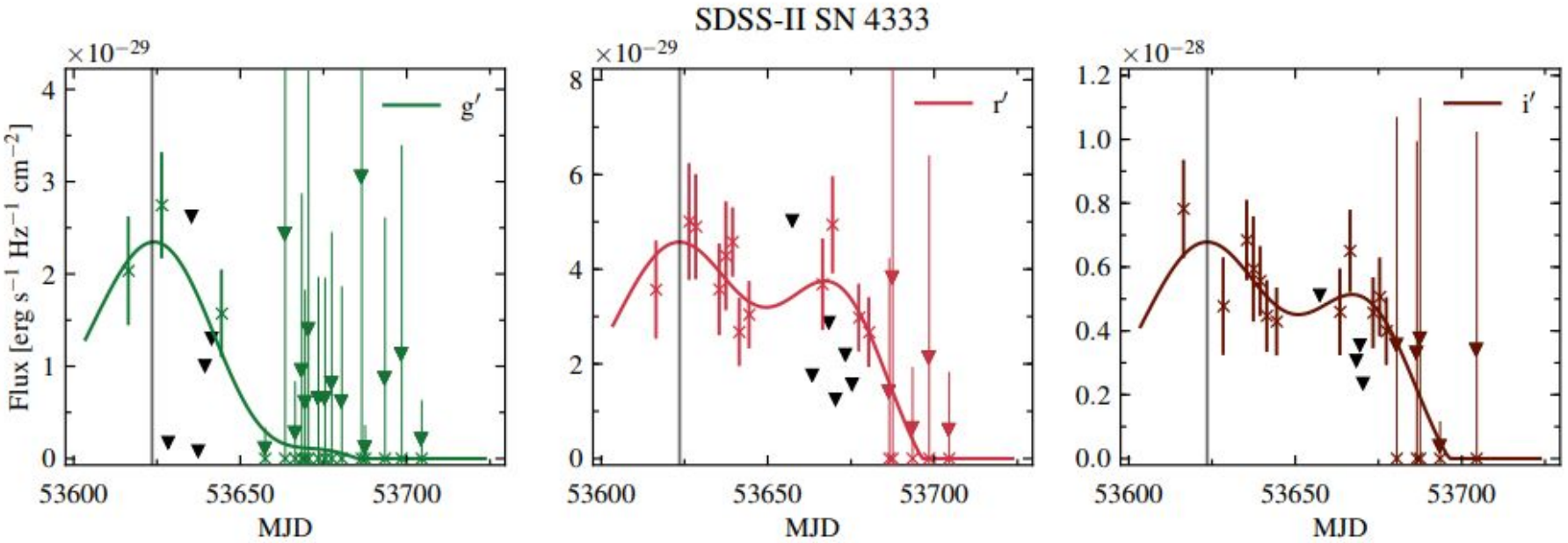
Input: Dataset $\mathbf{H}$, budget B
Initialize the weights $\mathbf{w}^{(0)} = \{\frac{1}{\sqrt{m}}, ..., \frac{1}{\sqrt{m}}\}$
Set $t = 0$
Set $\mathbf{H}_A = \mathbf{H}_N = \emptyset$
while $t \leq B$ **do**
  $t = t + 1$
  Set $\mathbf{a} = \mathbf{H} \cdot \mathbf{w}$ (i.e., $\mathbf{a}$ is the vector of anomaly scores)
  Let $\mathbf{z}_i$ = instance with highest anomaly score (where $i = \arg\max_i(a_i)$)
  Get feedback {*'anomaly'*/*'nominal'*} on $\mathbf{z}_i$
  if $\mathbf{z}_i$ is *anomaly* **then**
    $\mathbf{H}_A = \{\mathbf{z}_i\} \cup \mathbf{H}_A$
  else
    $\mathbf{H}_N = \{\mathbf{z}_i\} \cup \mathbf{H}_N$
  end if
  $\mathbf{w}^{(t)}$ = compute new weights; normalize $\|\mathbf{w}^{(t)}\| = 1$
end while

$$\mathbf{w}^{(t)} = \arg\min_{\mathbf{w},\xi} \frac{C_A}{|\mathbf{H}_A|}\left(\sum_{\mathbf{z}_i \in \mathbf{H}_A} \ell(\hat{q}_\tau(\mathbf{w}^{(t-1)}), \mathbf{w}; (\mathbf{z}_i, y_i))\right) + \frac{1}{|\mathbf{H}_N|}\left(\sum_{\mathbf{z}_i \in \mathbf{H}_N} \ell(\hat{q}_\tau(\mathbf{w}^{(t-1)}), \mathbf{w}; (\mathbf{z}_i, y_i))\right) + \frac{C_\xi}{|\mathbf{H}_A|}\left(\sum_{\mathbf{z}_i \in \mathbf{H}_A} \ell(\mathbf{z}_\tau^{(t-1)} \cdot \mathbf{w}, \mathbf{w}; (\mathbf{z}_i, y_i))\right) + \frac{C_\xi}{|\mathbf{H}_N|}\left(\sum_{\mathbf{z}_i \in \mathbf{H}_N} \ell(\mathbf{z}_\tau^{(t-1)} \cdot \mathbf{w}, \mathbf{w}; (\mathbf{z}_i, y_i))\right) + \|\mathbf{w} - \mathbf{w}_p\|^2 \quad (2)$$

where, $\mathbf{w}_p = \frac{\mathbf{w}_U}{\|\mathbf{w}_U\|} = [\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}]^T$, $\mathbf{z}_\tau^{(t-1)}$ and $\hat{q}_\tau(\mathbf{w}^{(t-1)})$ are

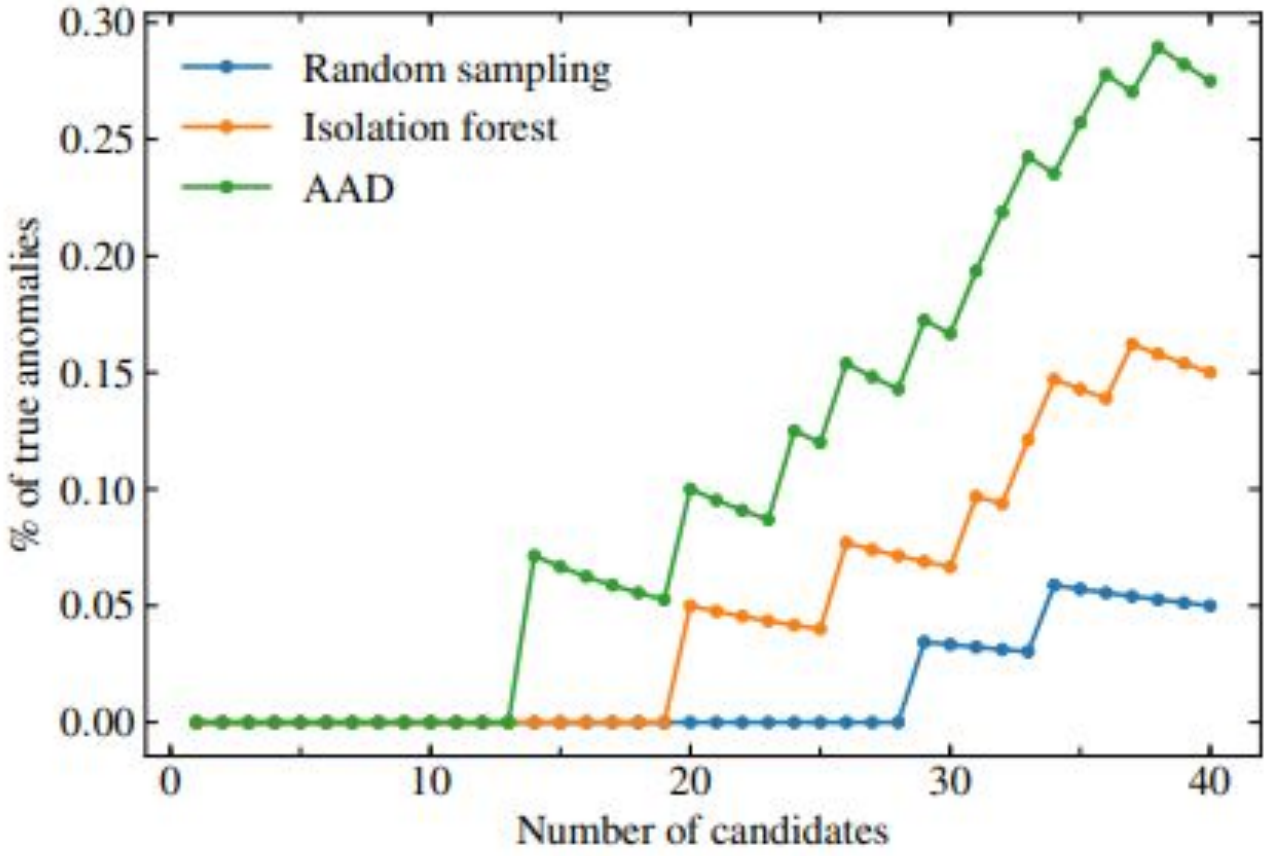
Anomaly Detection in the Open Supernova Catalog

Fit multi-varied GP with upper limits



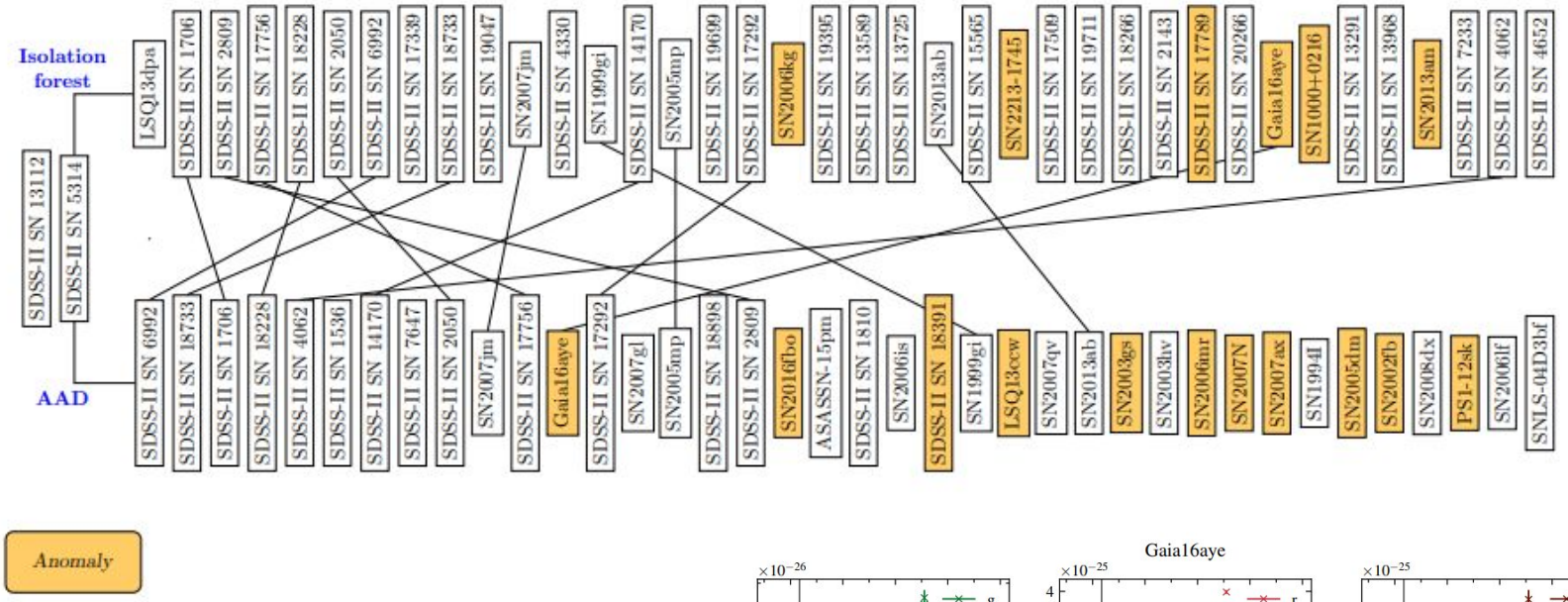
Pruzhinskaya et al., 2019

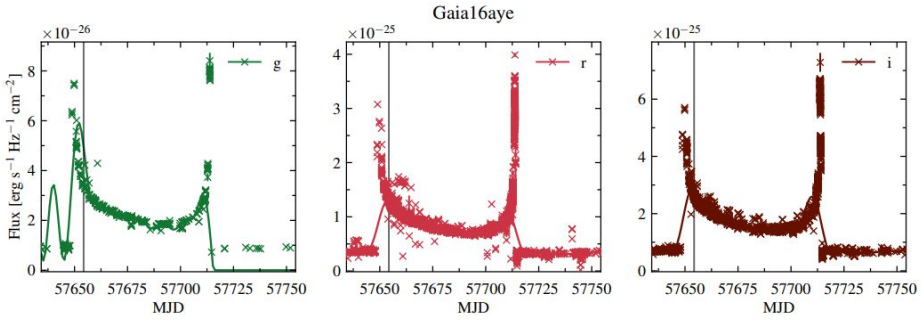
Active Anomaly Detection for time domain discoveries



Ishida et al., 2019 - https://arxiv.org/pdf/1909.13260v1.pdf

Active Anomaly Detection for time domain discoveries





Ishida et al., 2019 - https://arxiv.org/pdf/1909.13260v1.pdf

Extra slides

Pre-processing pipeline

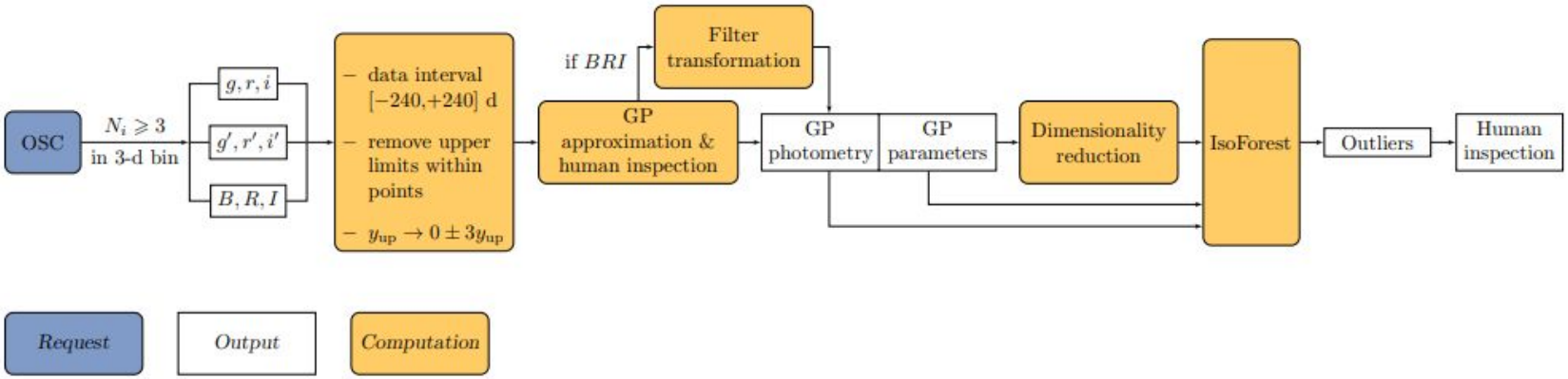


Figure 1. Workflow for the analysis. N_i denotes the number of observations in i'th band. GP photometry includes 364 features: 121×3 normalized fluxes and the LC flux maximum; GP parameters are 9 fitted parameters of the Gaussian process kernel and the log-likelihood of the fit.

Pruzhinskaya et al., 2019