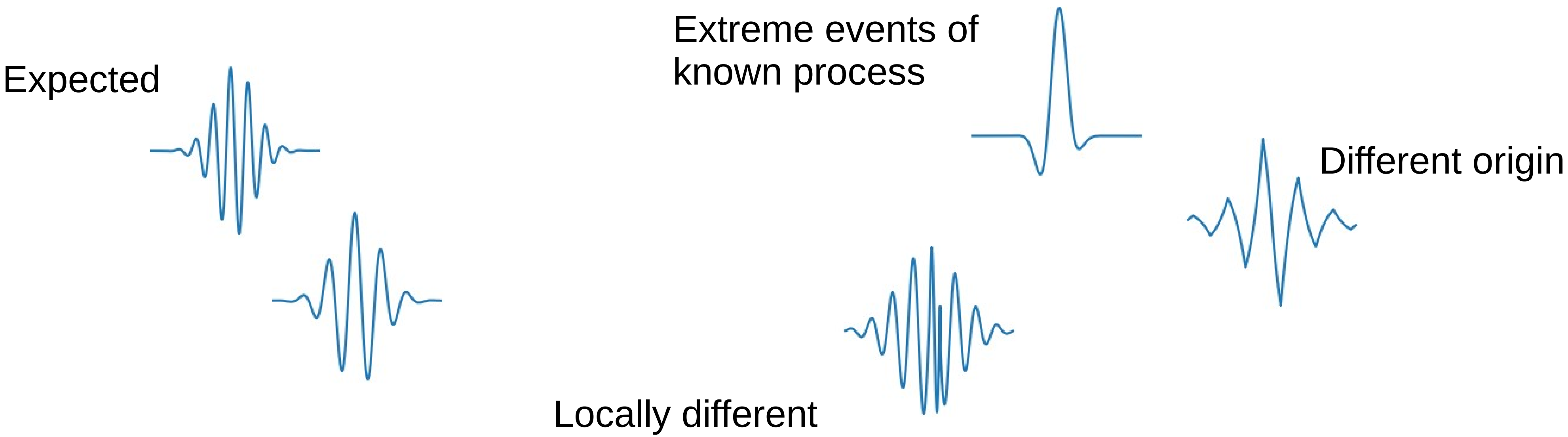
Signatures to help interpretability of anomalies

Emmanuel Gangler

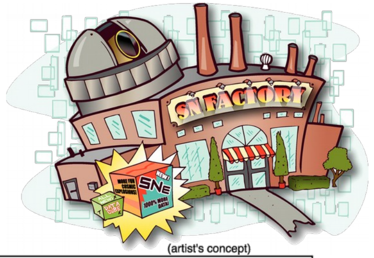
Interpreting Anomalies

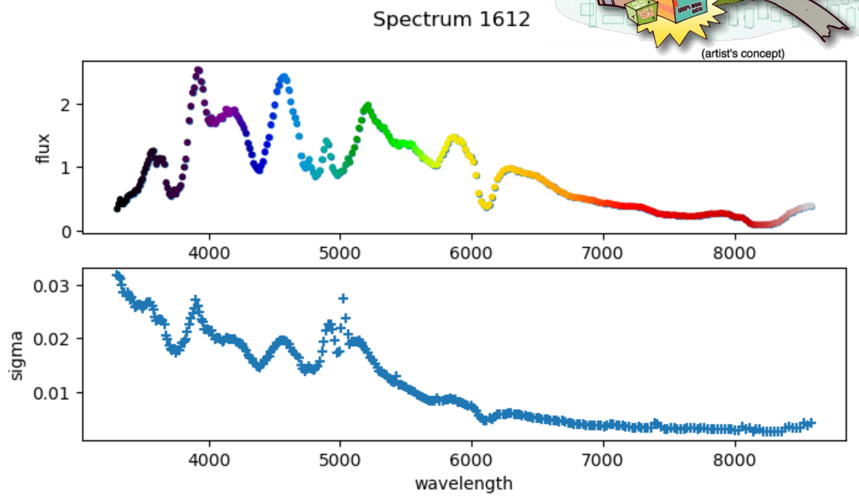
Anomalies can come in different flavors



→ How to make the difference ?

SNFactory dataset

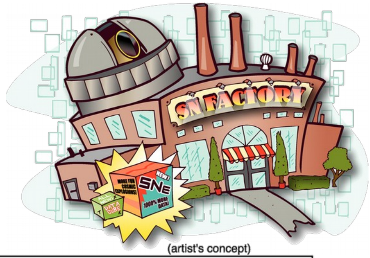




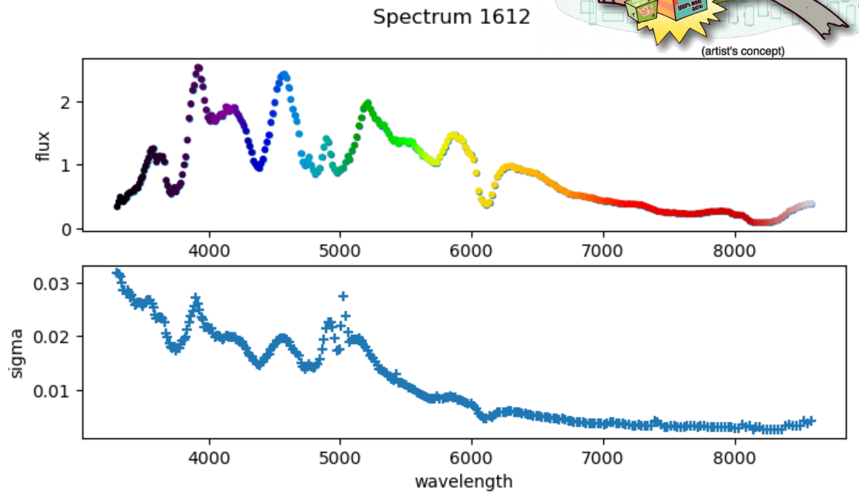
- Public astronomical dataset
  arXiv:2005.03462
  - **2323 spectra** of Type Ia supernovae
    - 288 spectral bins
    - With noise estimate

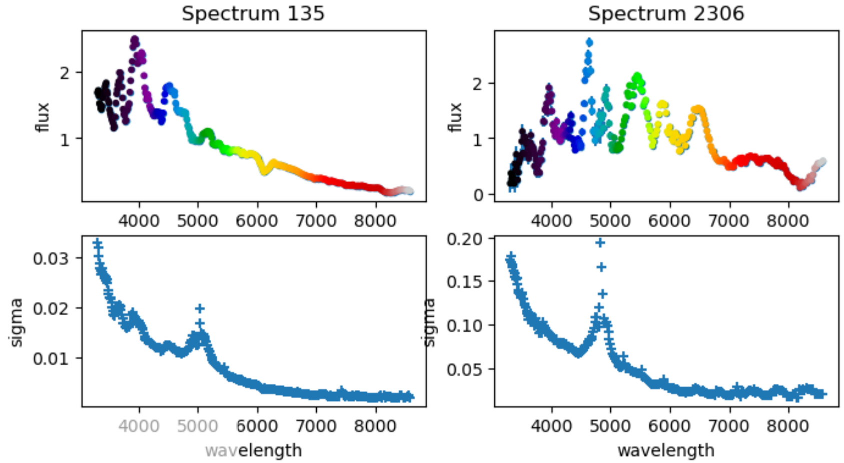
      → **576 features**

SNFactory dataset



- Public astronomical dataset
  arXiv:2005.03462
  - **2323 spectra** of Type Ia supernovae
    - 288 spectral bins
    - With noise estimate
      → **576 features**
- Interest of this dataset for anomalies
  - High internal variability
  - Expert tagging of anomalies
  - Noisy data making the task difficult
  - Local data artifacts

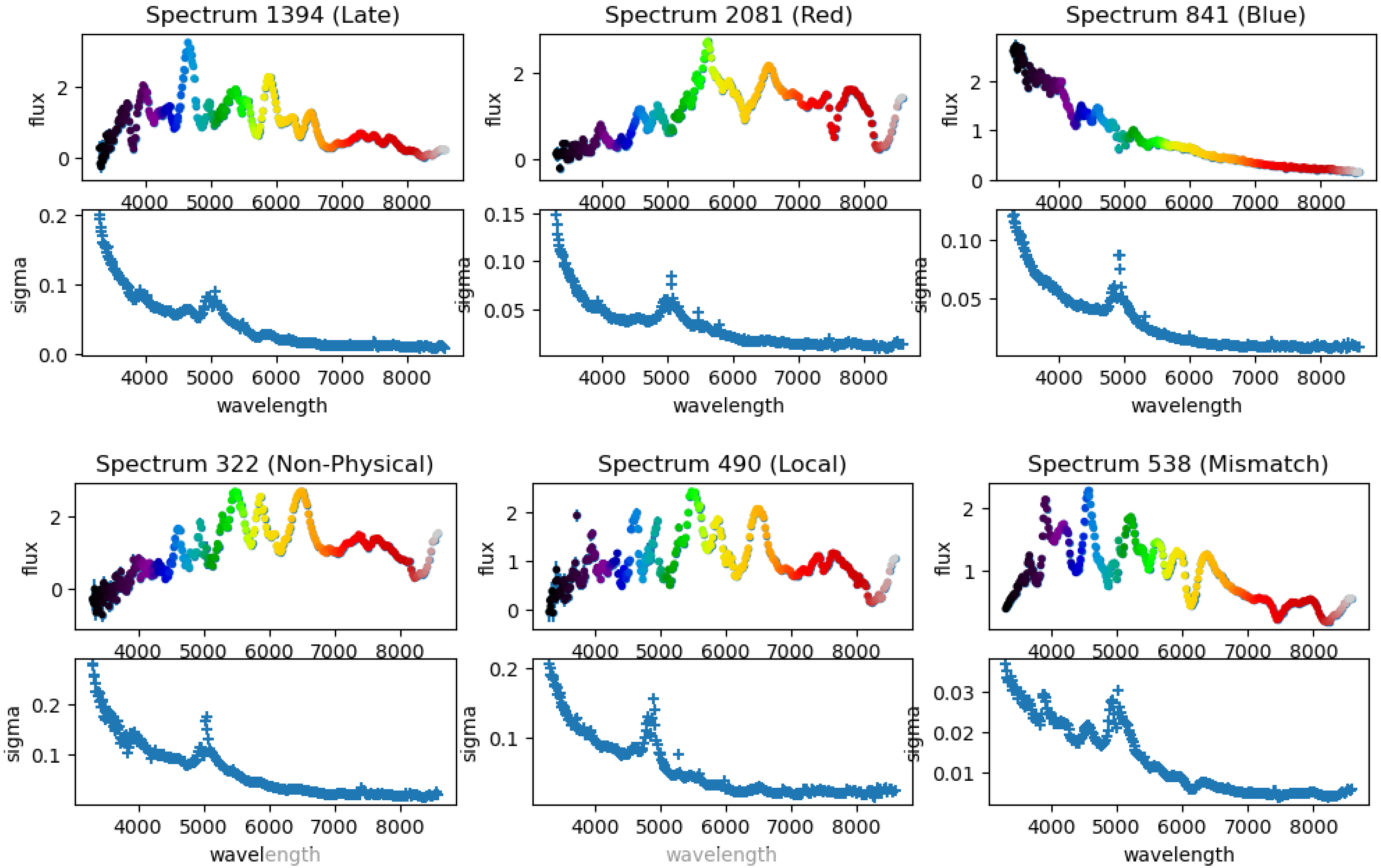




Some expected "anomalies" :

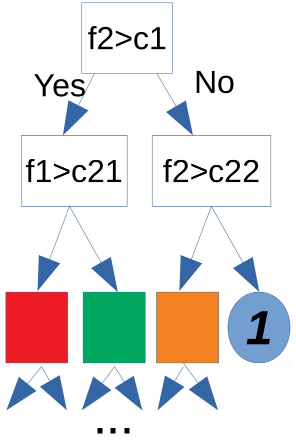
Physical

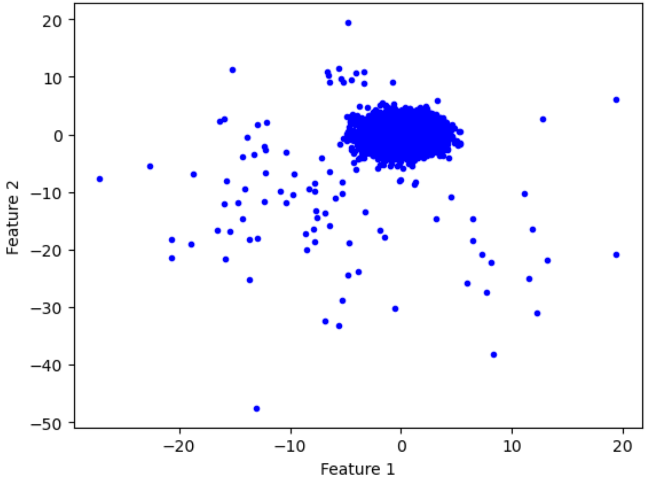
Artifacts



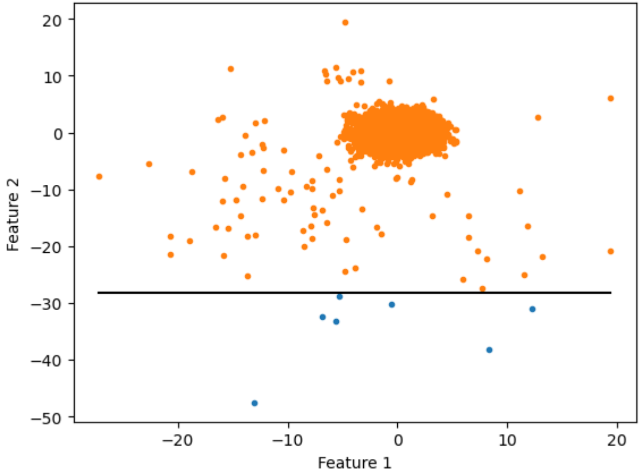
Isolation Forest Density Estimation

- **Random Tree:**
  - Select a feature randomly
  - Select a random threshold within the range spanned by the feature for the (sub) sample
  - Repeat for each subsample
  - Stop when only 1 point in the sub-partition
- **Random Forest:**
  - Average depth on T trees (proxy for local density)

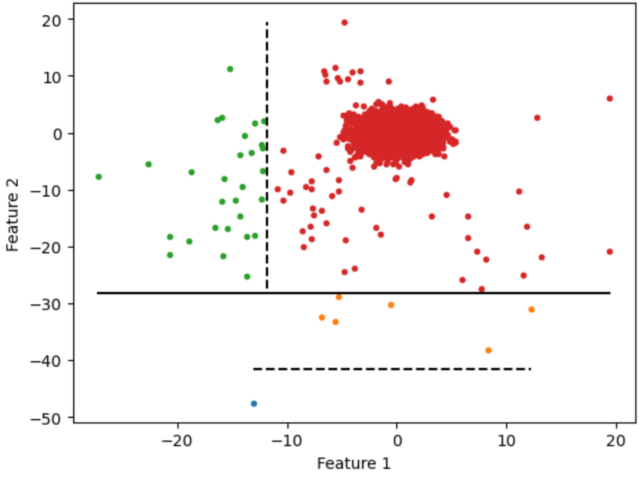




Depth 0

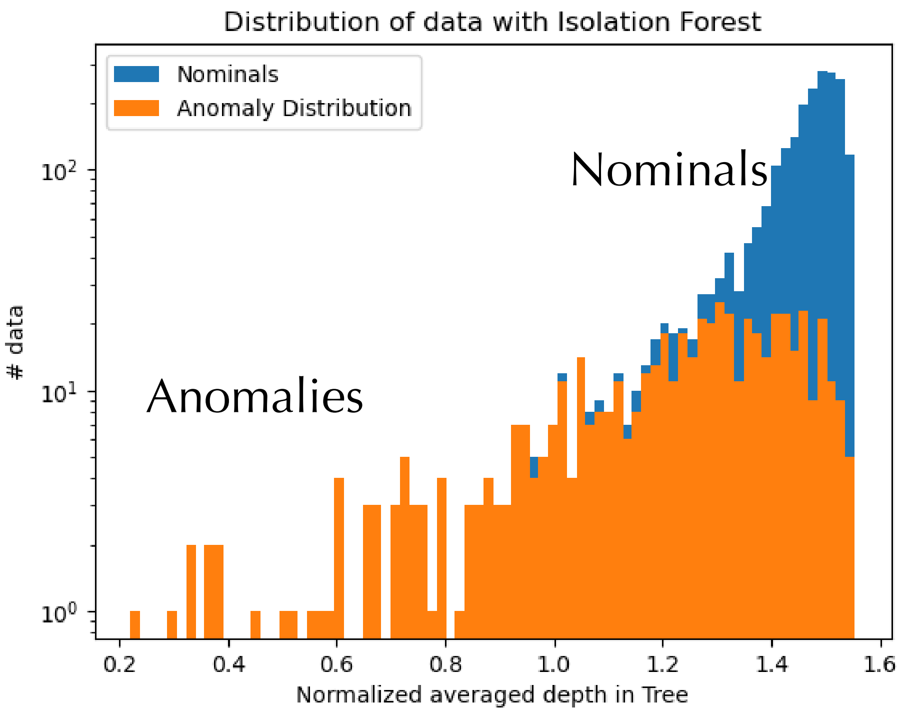


Depth 1

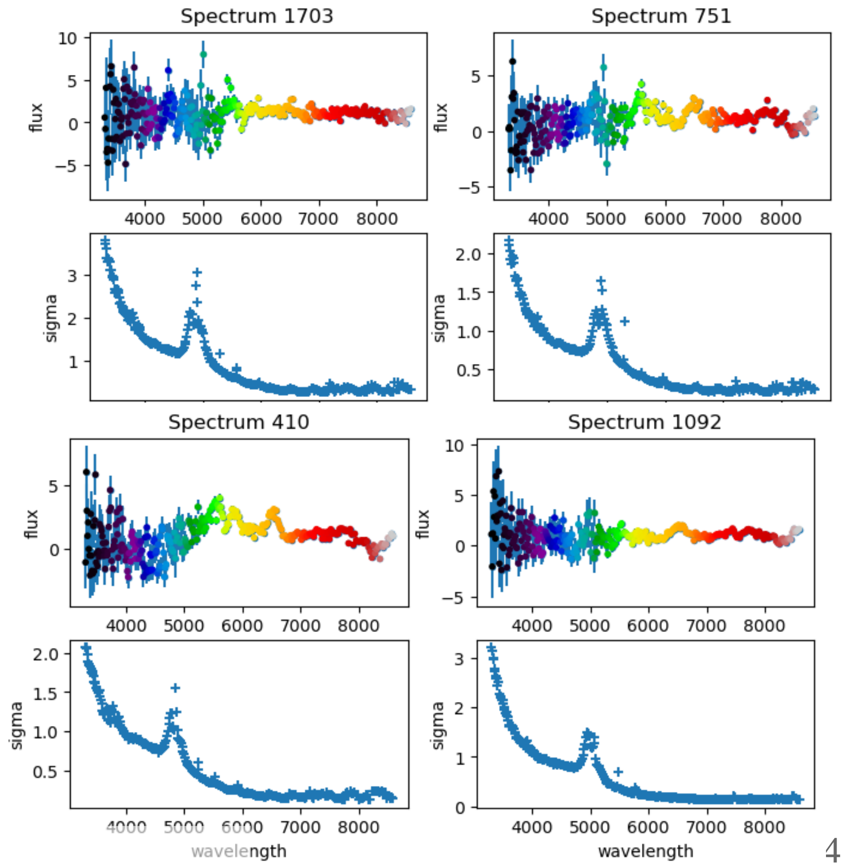


Depth 2

Isolation Forest for SNFactory dataset



Top 4 outliers

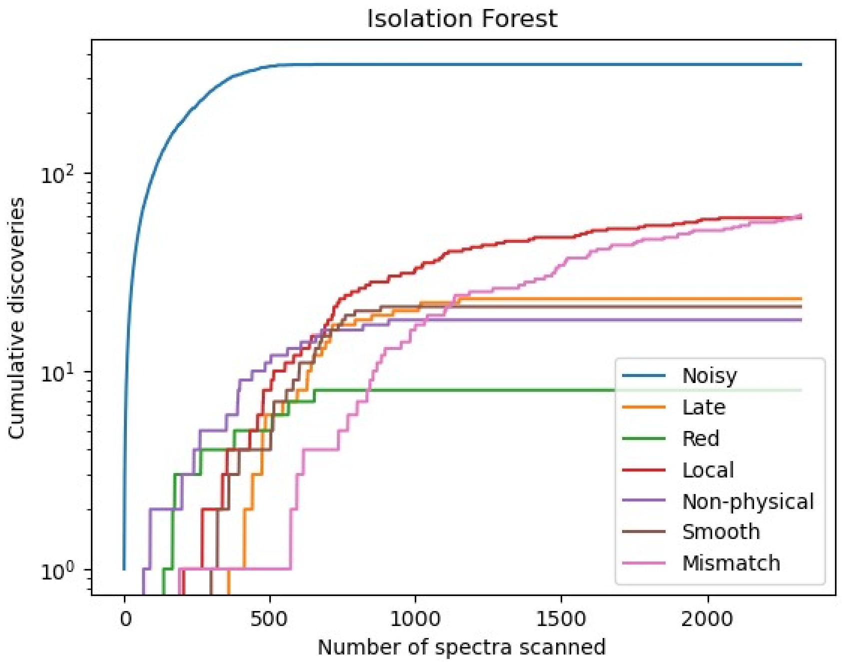


Outliers are anomalies of the "noisy" type

Discovery efficiency

- IF very efficient for **dominant anomaly**
  → Noisy data dominate (AUC=0.985)

- Less efficient for **other classes**
  - **Rank** of last anomaly type dicovered: 360 (expected 326)
  - For non-noise anomalies : AUC=0.61



Some common Questions:

- **Why** are some data taged as anomalies ?
- Are there **different classes** of anomalies ?
- Can I find **more anomalies** of a given kind ?
- Can I imporove discovery of **new anomalies** ?

Anomaly signature

- Anomaly score for 1 tree

Depth 0

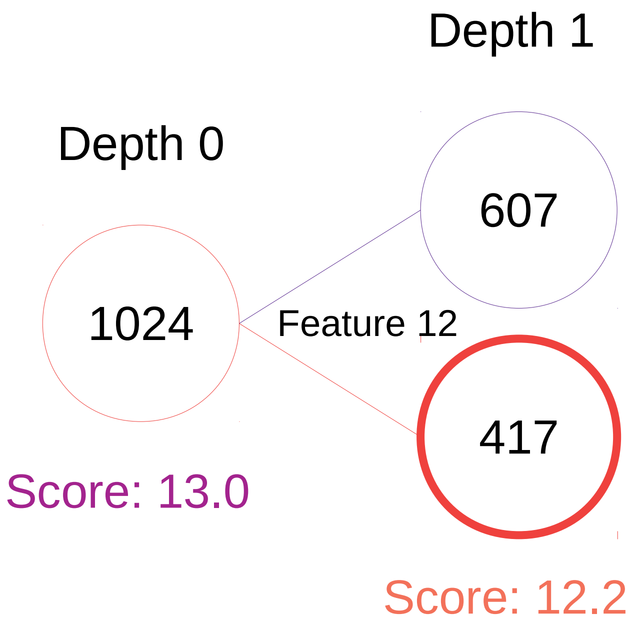
Number
of samples
reaching the node

Score: 13.0

Before any decision :
Expected score for anomaly = Average tree depth

Anomaly signature

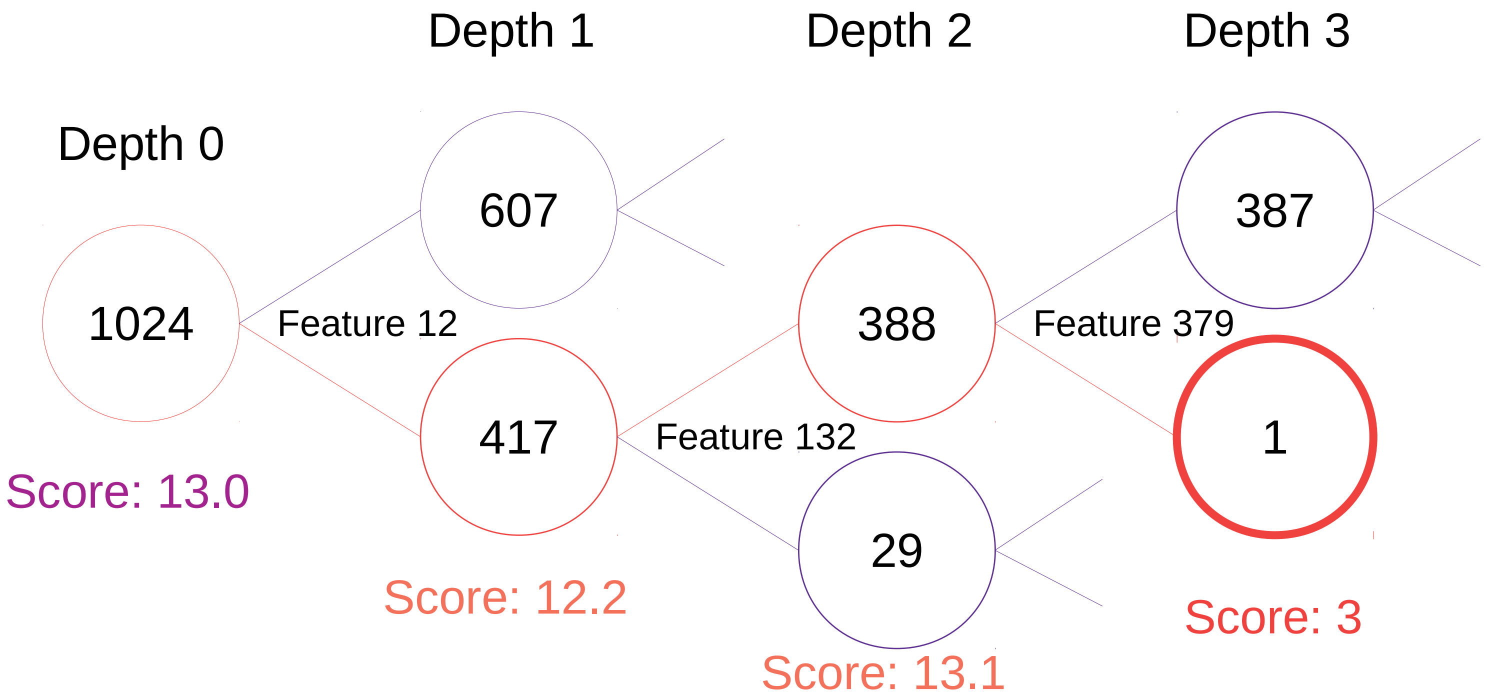
- Anomaly score for 1 tree



After cut 1 : score for anomaly = Average tree depth for 417 elements +1

Anomaly signature

- Anomaly score for 1 tree



For this outlier : Feature 12 = - 0.8 Feature 132 = + 0.9 **Feature 379 = - 10.1**

Anomaly signature

- Anomaly score (= average depth) :

$$S = S_0 + \frac{1}{T} \sum_{t,f_i} \delta S_{t,f_i}$$

→ Feature importance (the lower, the more anomalous)

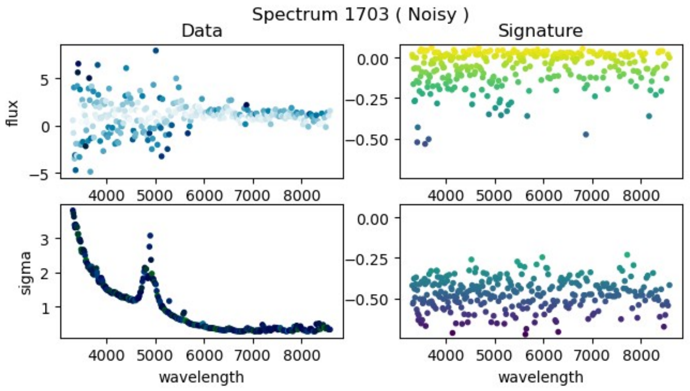
$$S_{f_i} = \frac{1}{n_i} \sum_{t} \delta S_{t,f_i}$$

Can be computed for a single data element, or a subsample

Signature & interpretability

Top 1 Anomaly

Data Signature

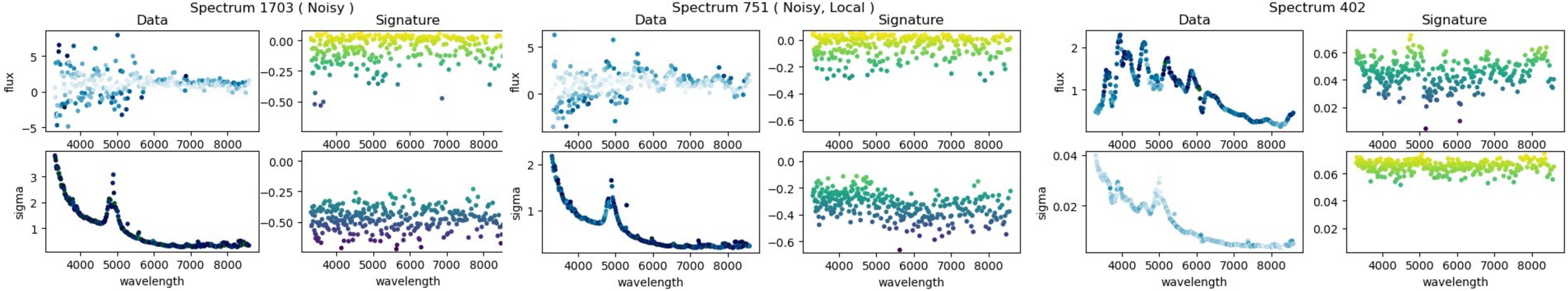


- Signature highlights where the data is anomalous
  - Negative score = anomalous
  - Interpretation : decision based on sigma

Signature & interpretability

Top 1 Anomaly | Top 2 Anomaly | Top 1 Nominal

Data | Signature | Data | Signature | Data | Signature

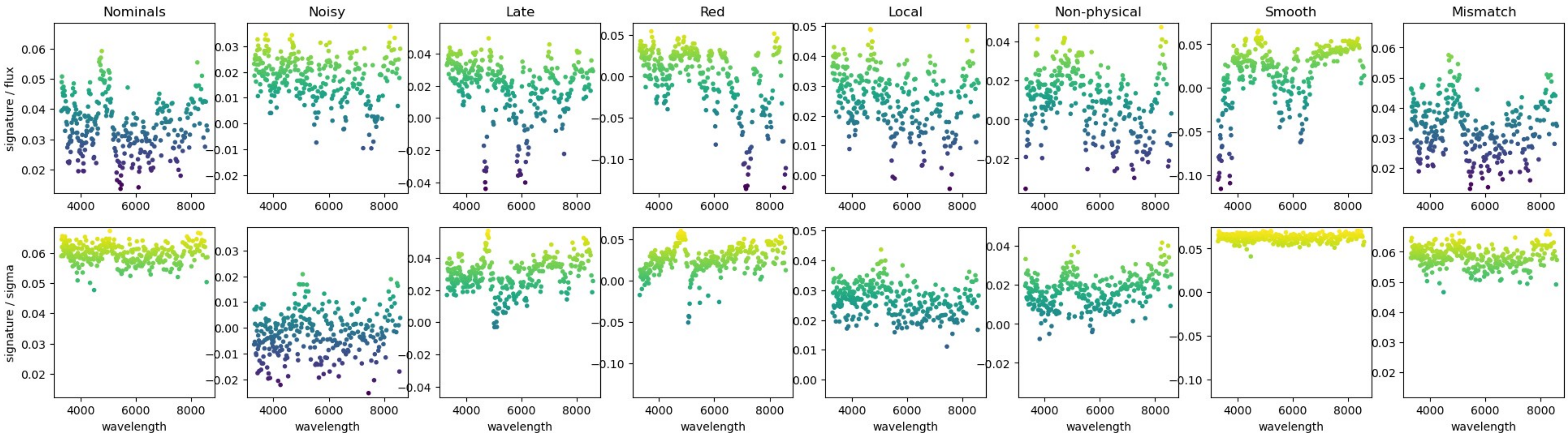


- Signature highlights where the data is anomalous
  - Negative score = anomalous
  - Interpretation : decision based on sigma
  - Positive score = nominal

Signatures as anomaly tags

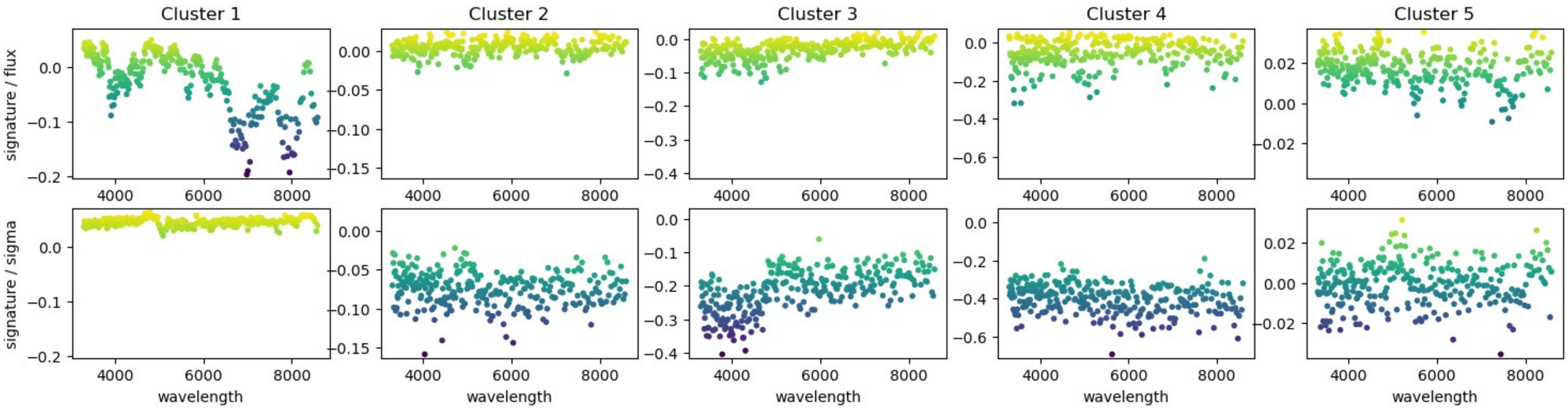
- Different kind of anomalies have distinctive signatures
- On one glance, expert knows where to search in the data

Average signature/expert-defined tag



Signatures & Clustering

- K-means on signatures for top 10 % anomalous data (232 spectra)
  - Very unbalanced : 90 % of those are tagged “Noisy”
  - Contains 7 % of nominals



- All anomalies belong to Cluster 1 !
- Only 39 elements : easier to analyse
- Still 2 classes of anomalies not found… + Choice of K is empirical

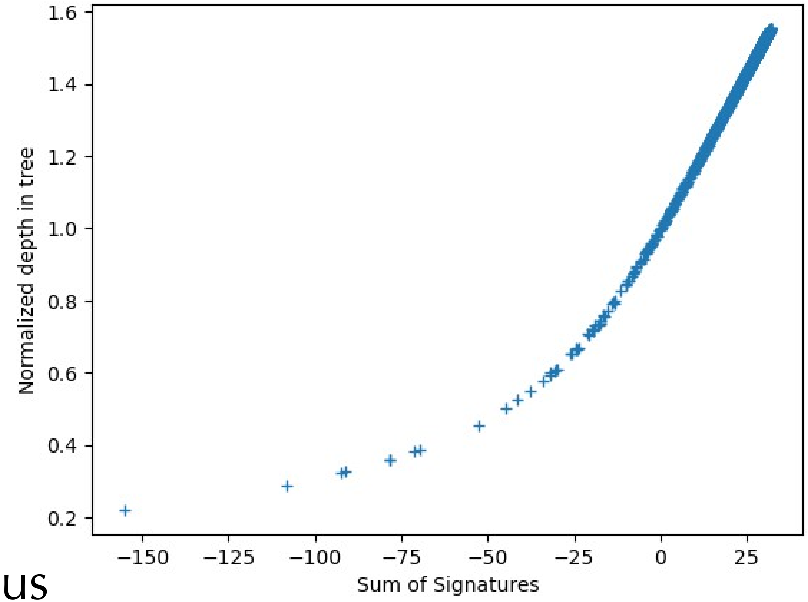
Recursive approach to novelty discovery with signatures

- Signature allow to derive a weighted anomaly score

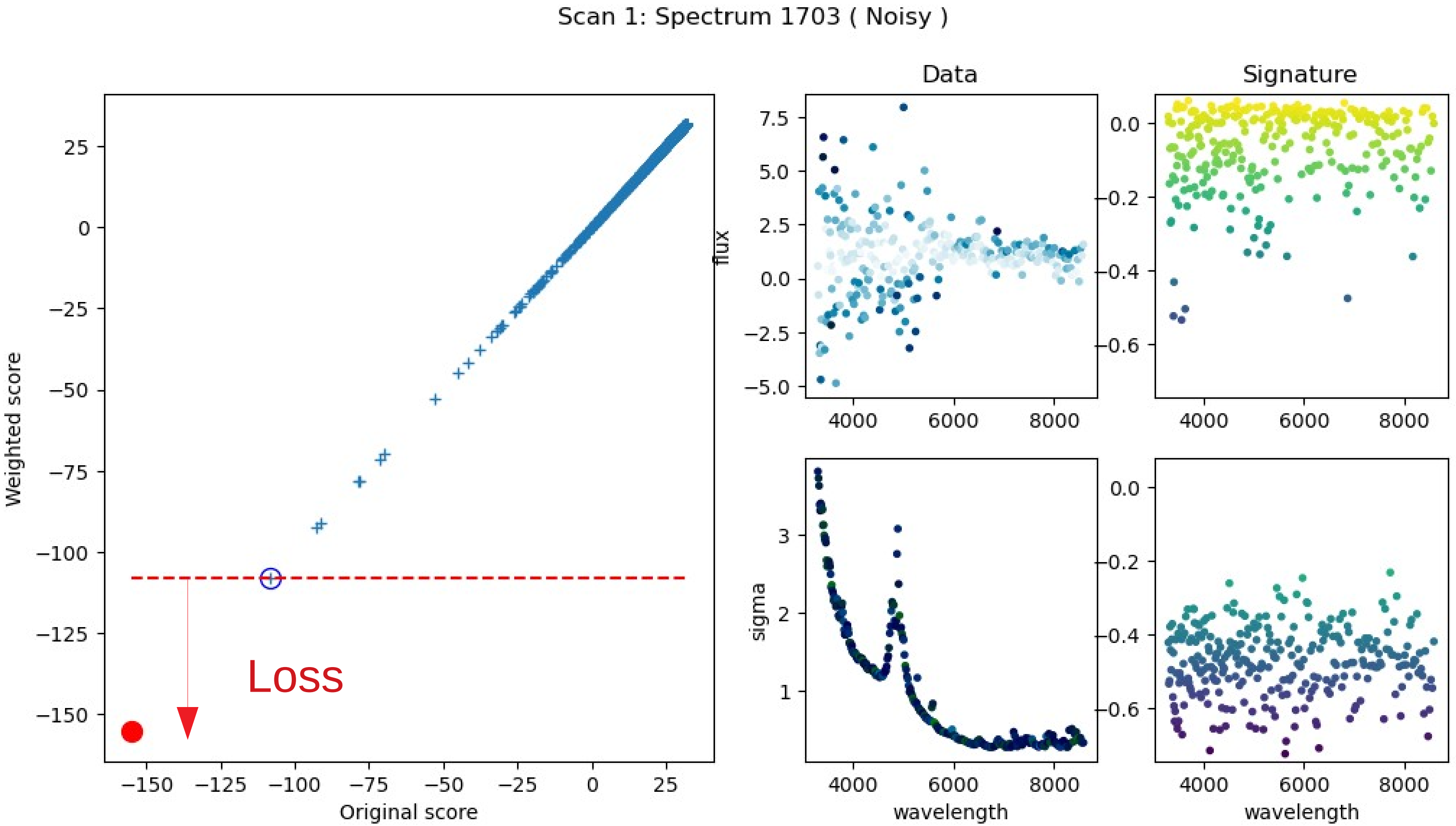
$$S^j = \sum_i \alpha_i S^j_{f_i}$$

  - Initialization of weights to 1

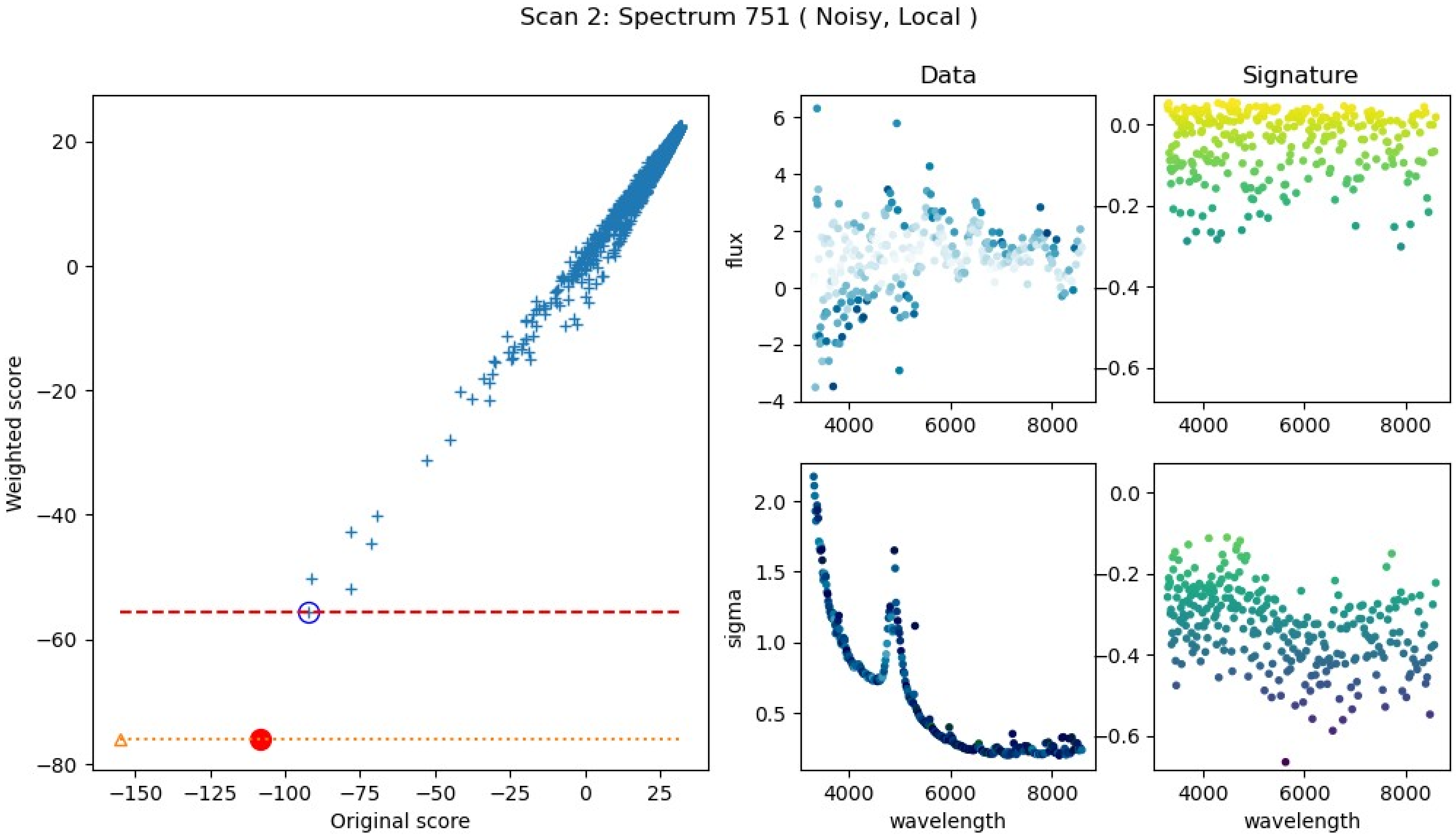
- For each data examined by the expert :
  - Tag as either wanted or unwanted
  - Update the weights / Hinge loss function
  - Propose to the expert the next mostly anomalous



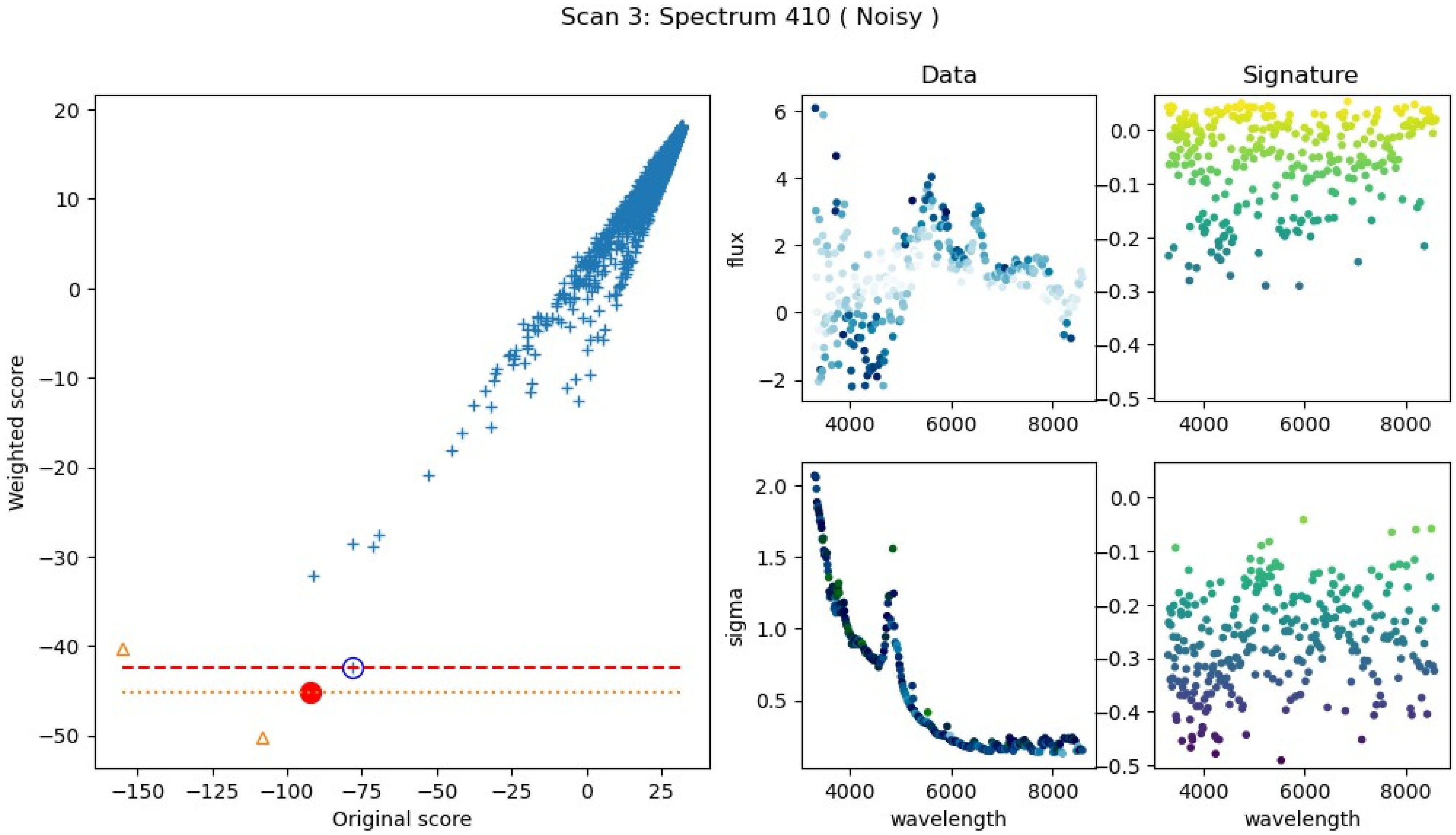
Discovery-oriented iterative process
→ tag everything as unwanted



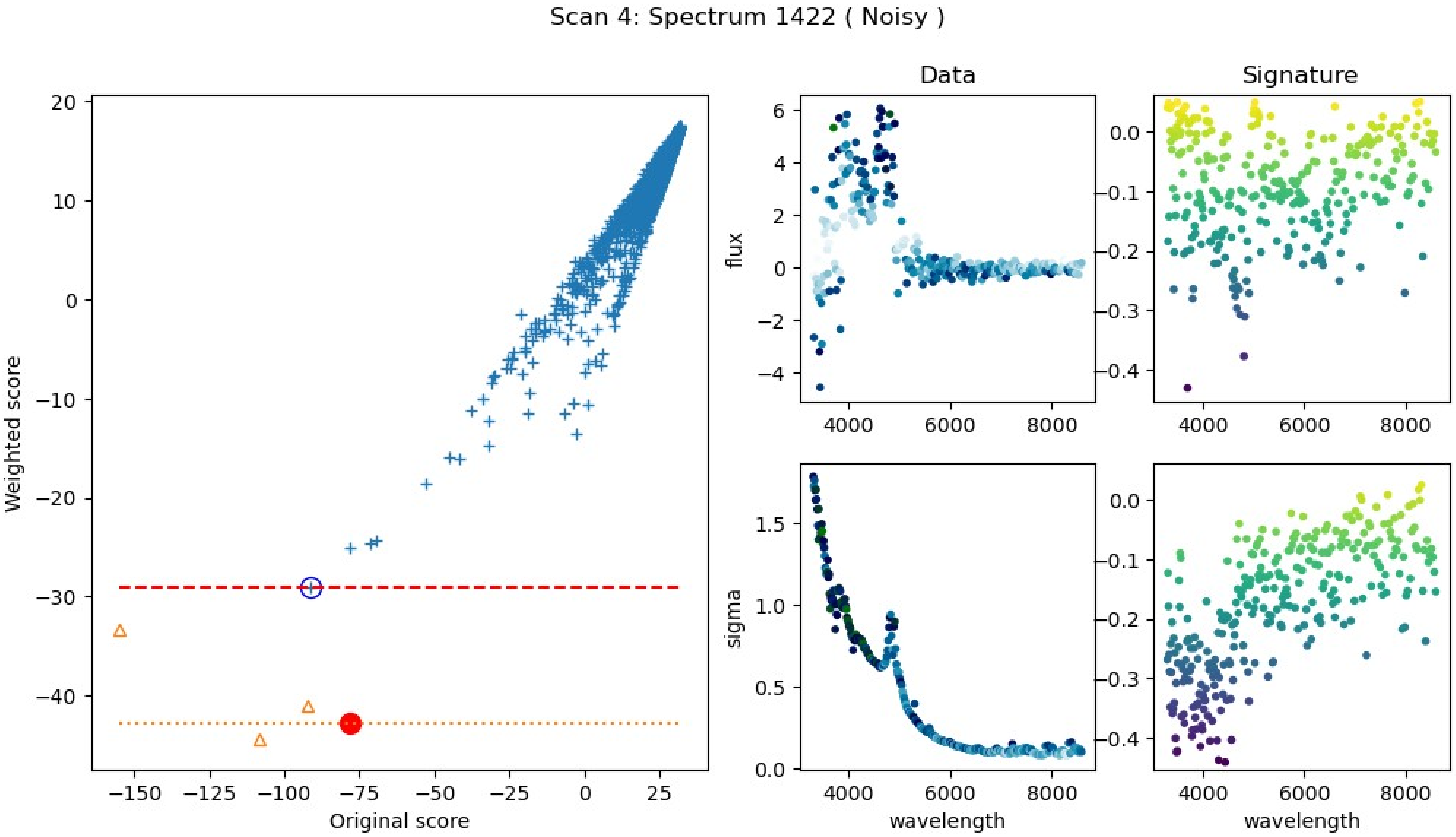
Discovery-oriented iterative process
→ tag everything as unwanted



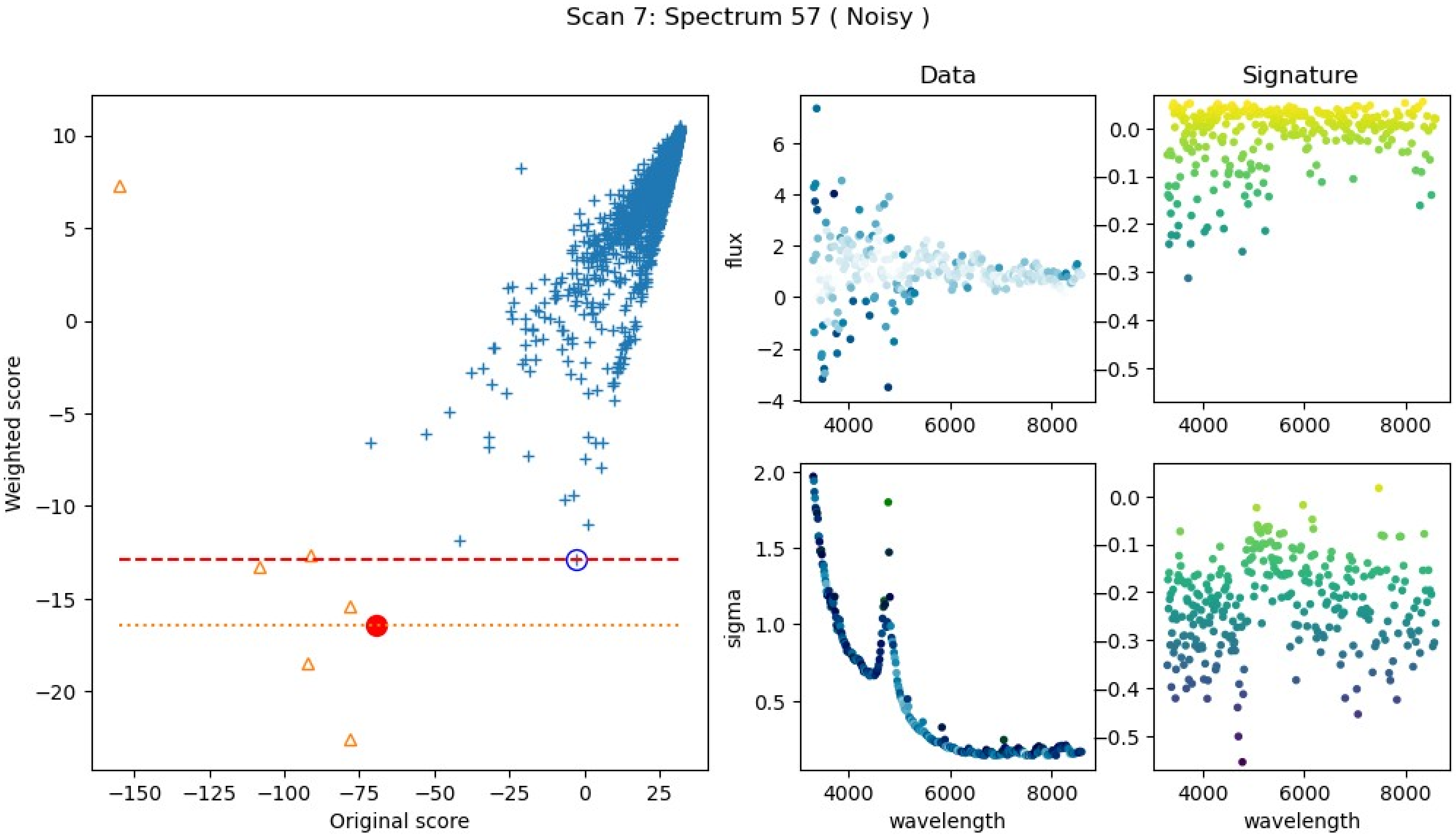
Discovery-oriented iterative process
→ tag everything as unwanted



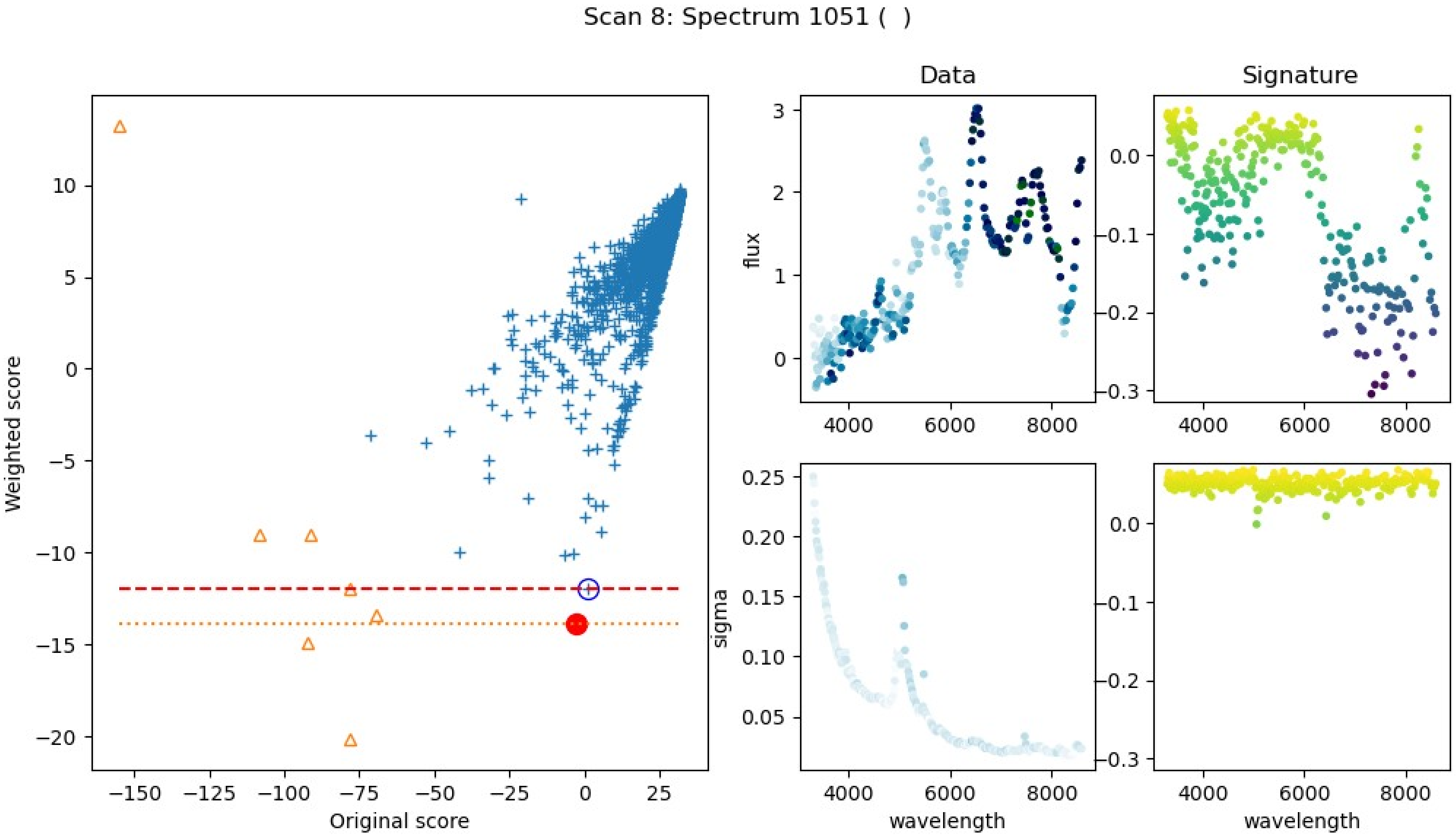
Discovery-oriented iterative process
→ tag everything as unwanted



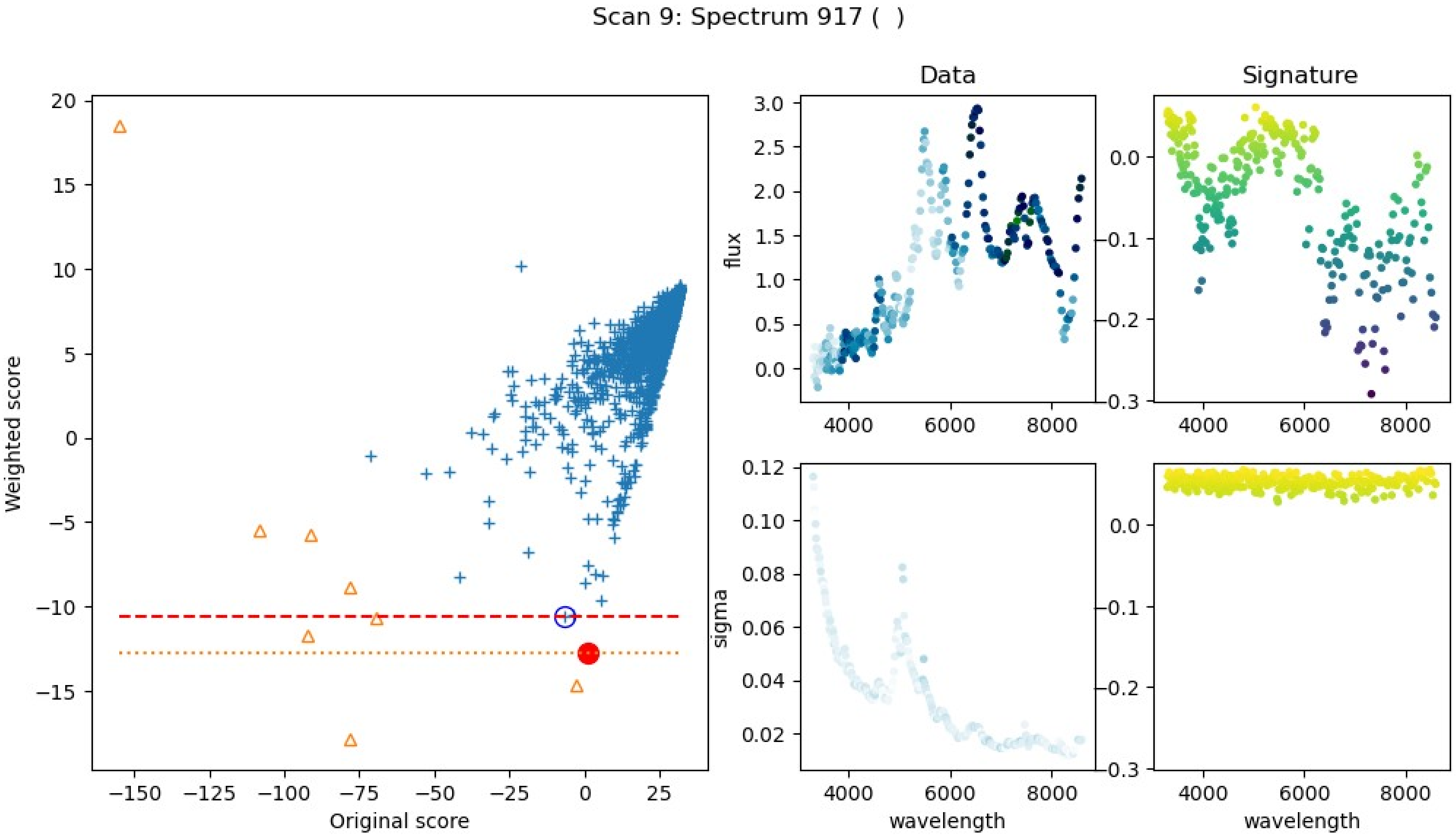
Discovery-oriented iterative process
→ tag everything as unwanted



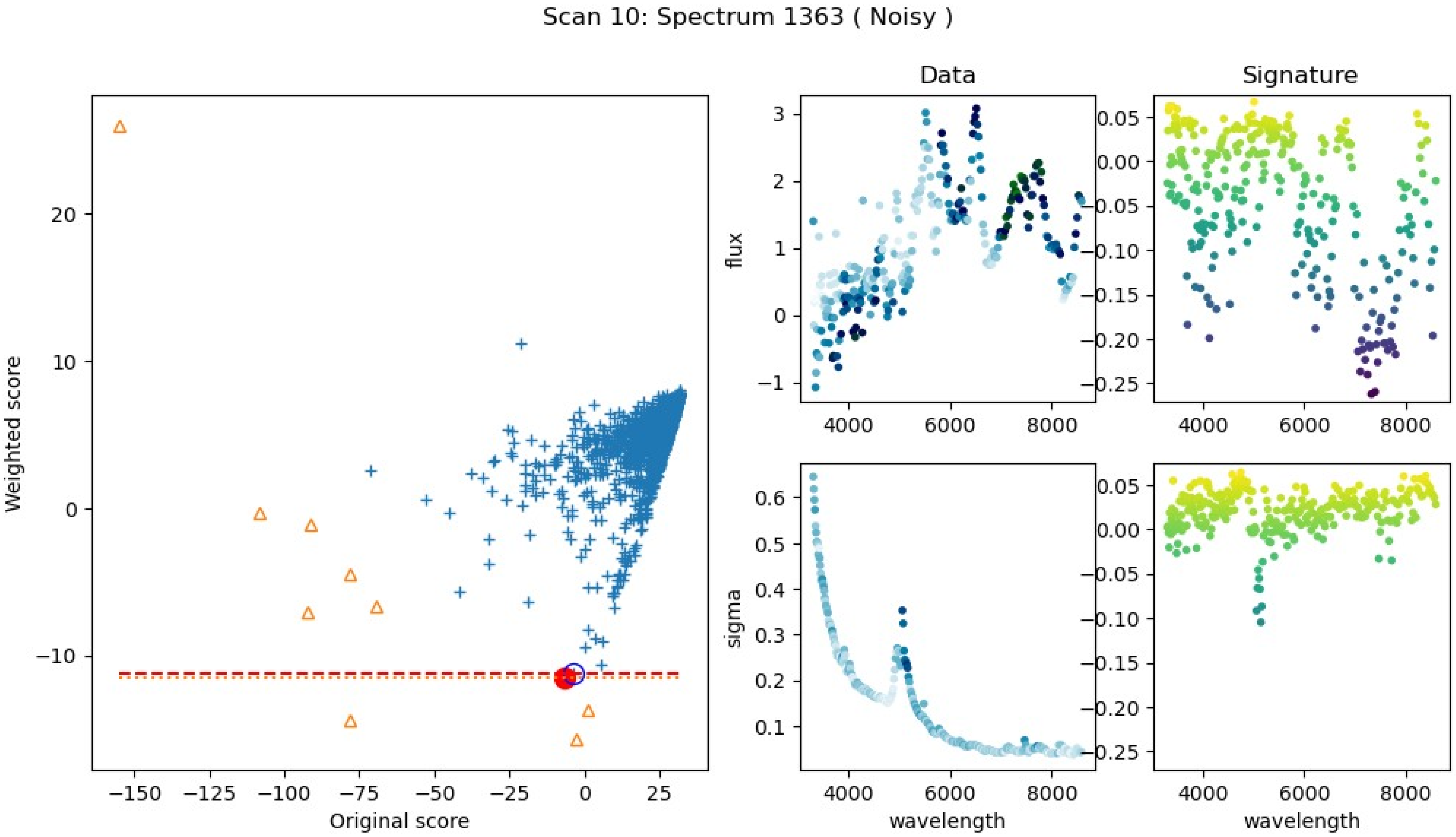
Discovery-oriented iterative process → tag everything as unwanted



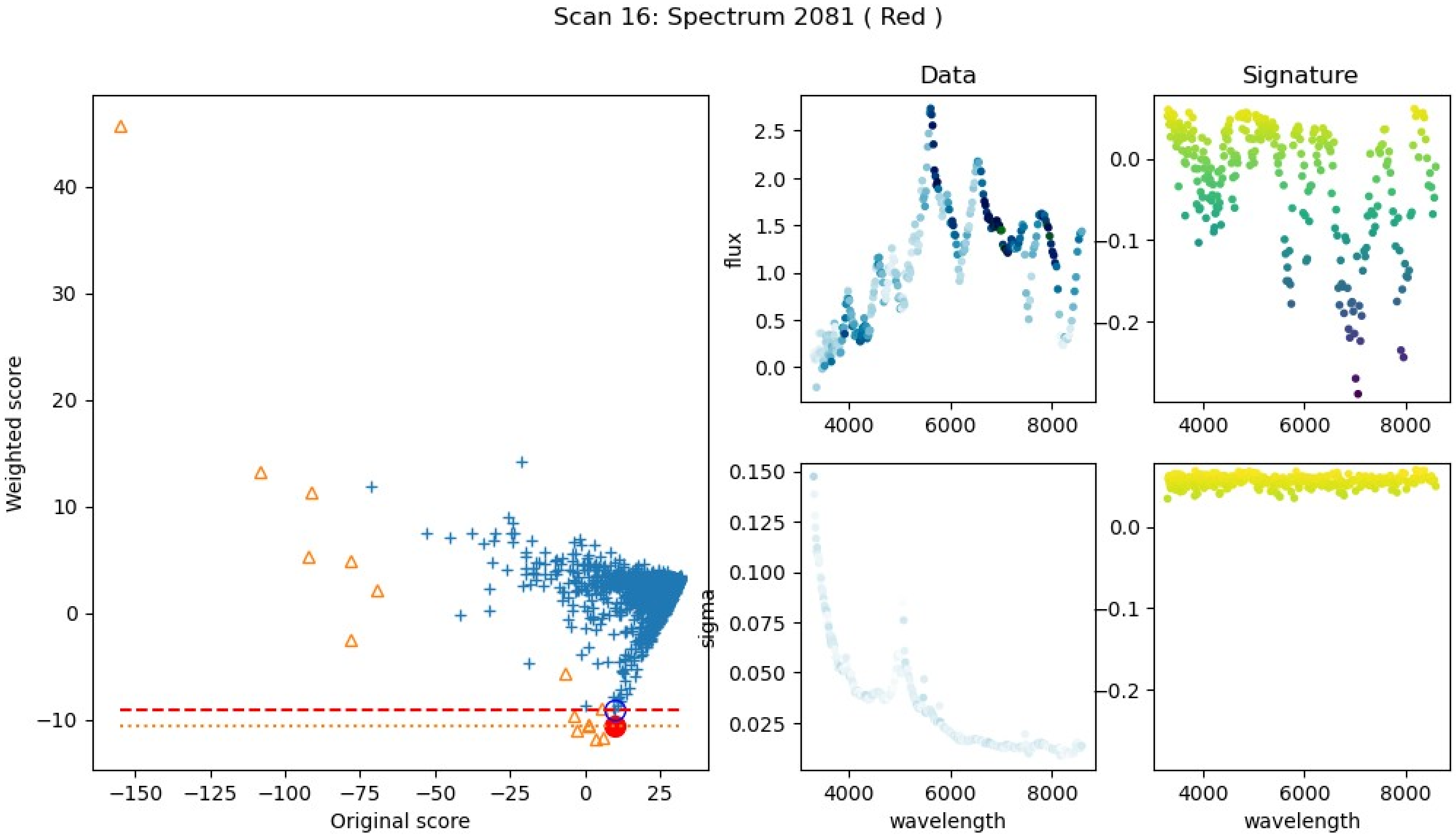
Discovery-oriented iterative process
→ tag everything as unwanted



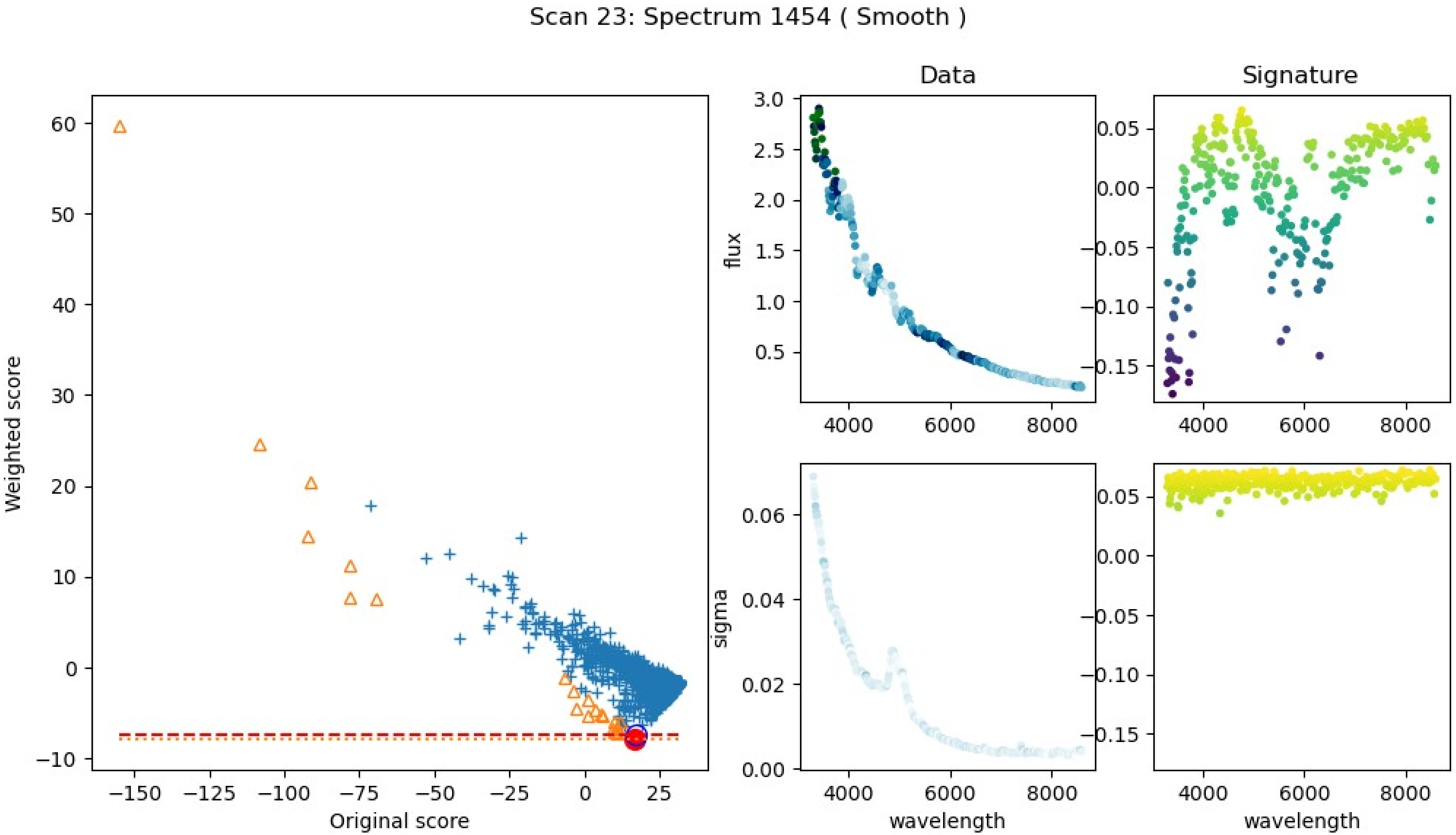
Discovery-oriented iterative process
→ tag everything as unwanted



Discovery-oriented iterative process
→ tag everything as unwanted

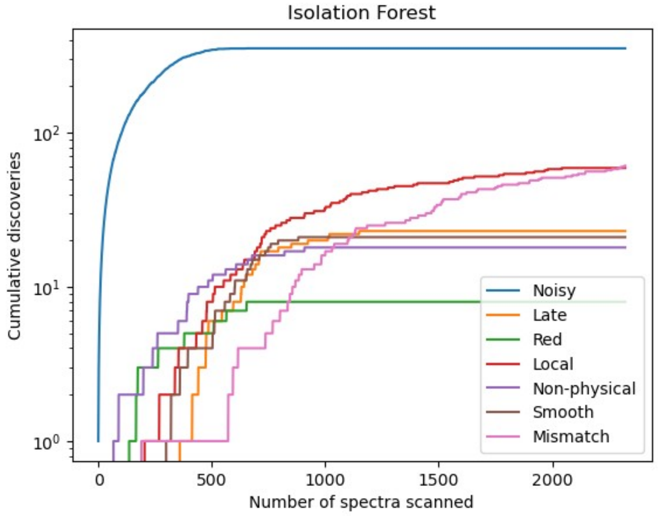


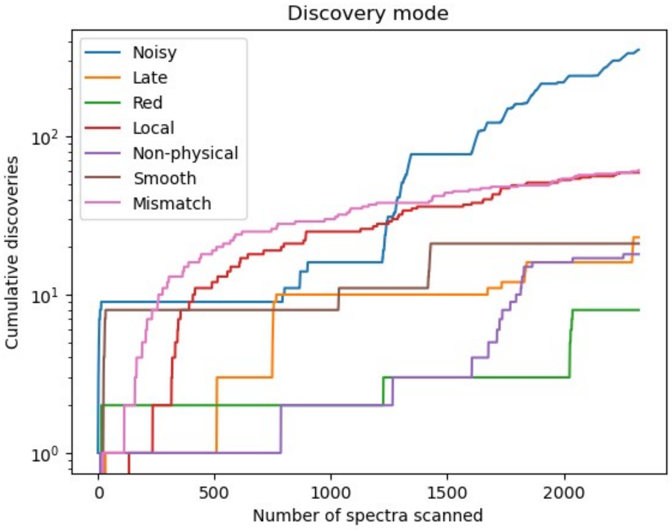
Discovery-oriented iterative process
→ tag everything as unwanted



Different tasks :

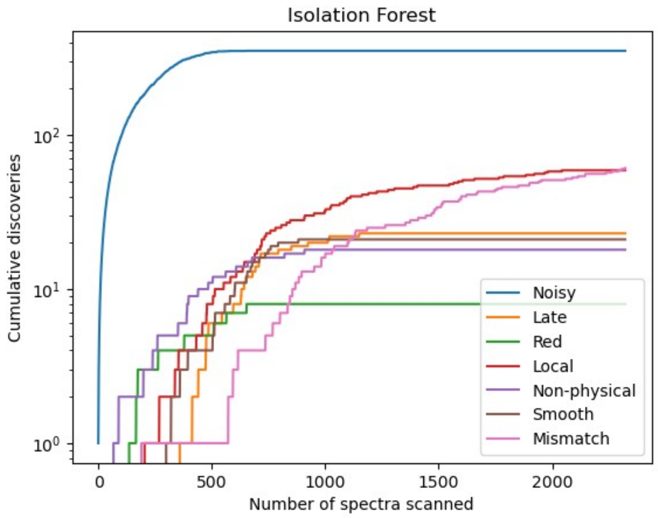
- Isolation Forest
  - Default
  - ***AUC(Noisy)***=0.98
  - AUC(Others)=0.60
  - Rank of last class=360
- Discovery mode
  - All unwanted
  - AUC(Noisy)=0.25
  - AUC(Others)=0.51
  - ***Rank*** of last class=133

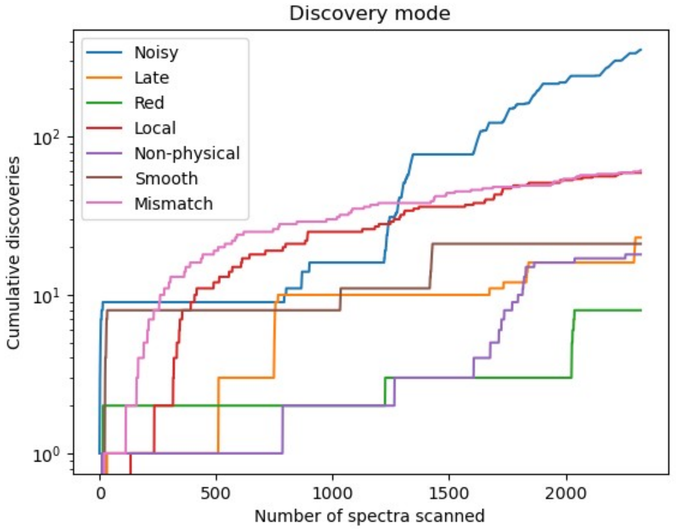


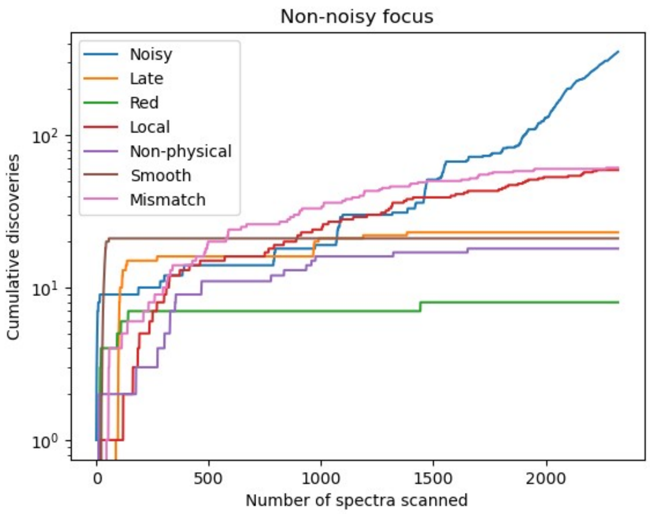


Different tasks :

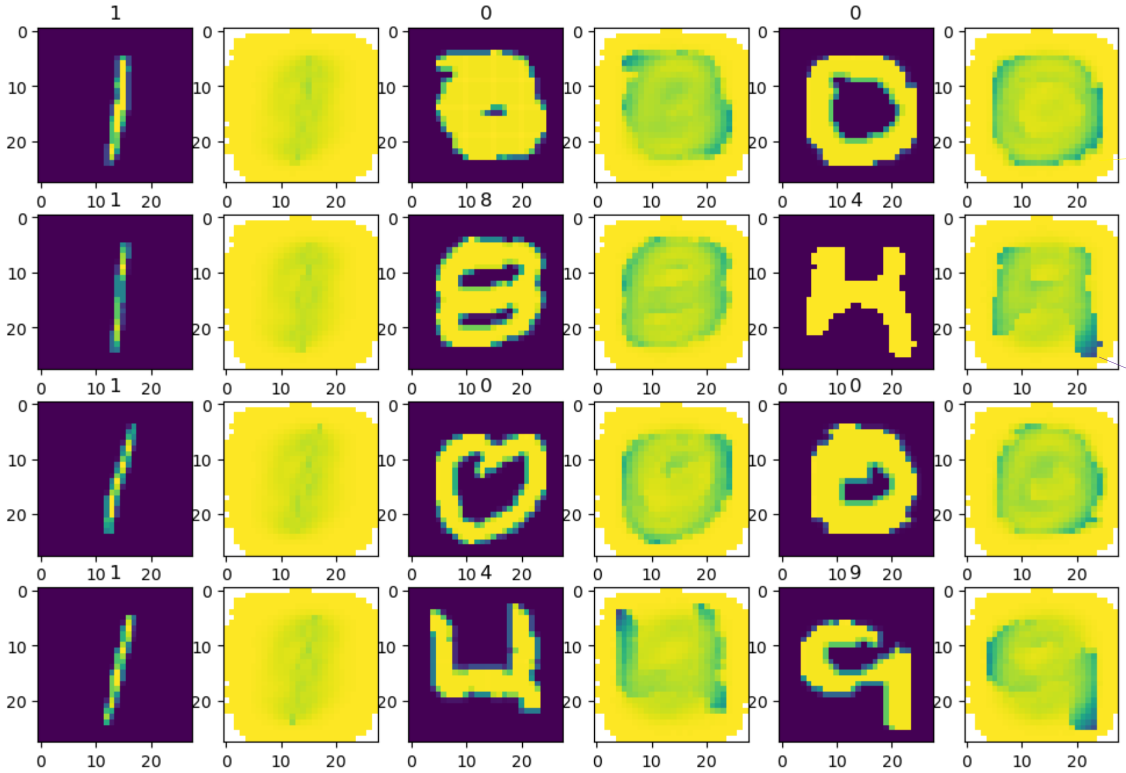
- Isolation Forest
  - Default
  - ***AUC(Noisy)***=0.98
  - AUC(Others)=0.60
  - Rank of last class=360
- Discovery mode
  - All unwanted
  - AUC(Noisy)=0.25
  - AUC(Others)=0.51
  - ***Rank*** of last class=133
- Active learning
  - Only non-noisy wanted
  - AUC(Noisy)=0.19
  - ***AUC(Others)***=0.70
  - Rank of last class=87







Signatures work also on MNIST:



Unimportant

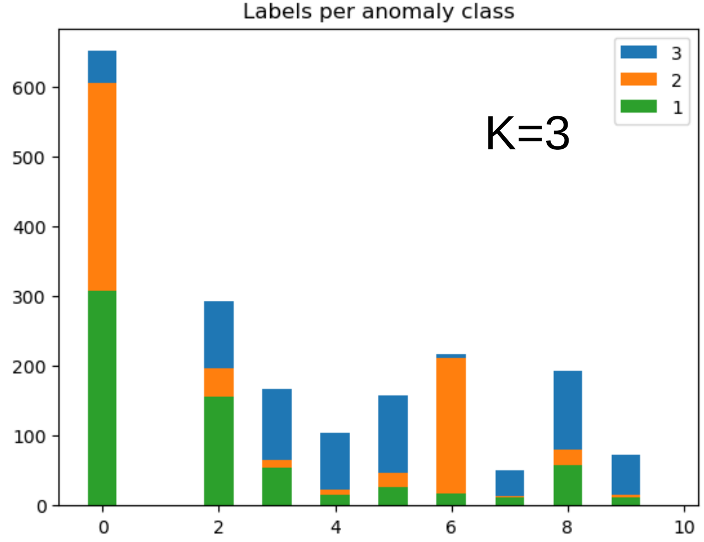
Strong signature

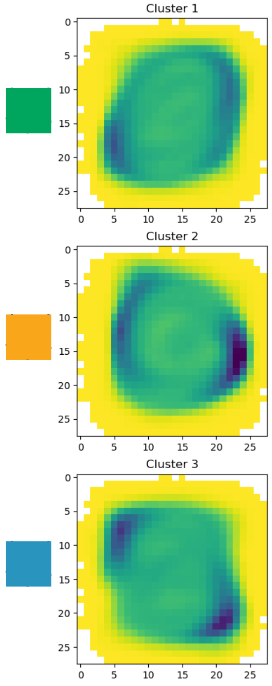
Nominals :
Small signature

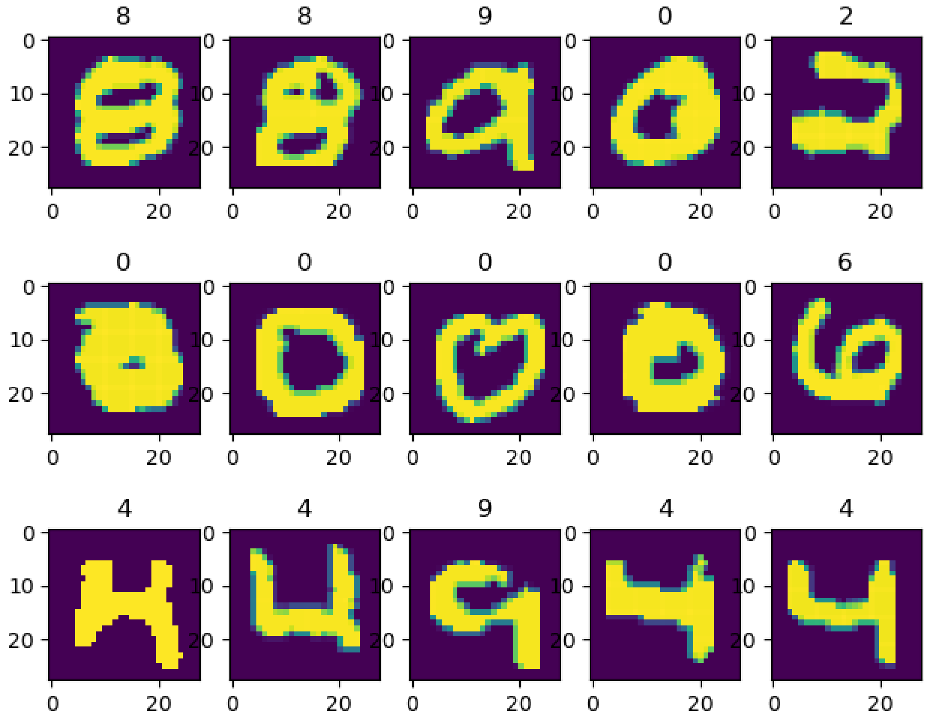
Outliers :
Signature indicates pixel contribution
→ different outliers have different signatures !

Using signatures to classify outliers

- Kmeans on signatures from 1950 outliers (depth<average depth)



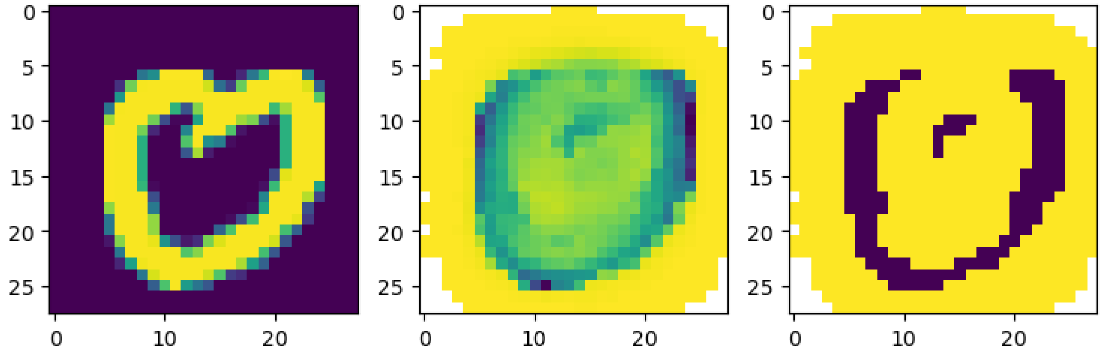




Cluster average signature

Examples / cluster

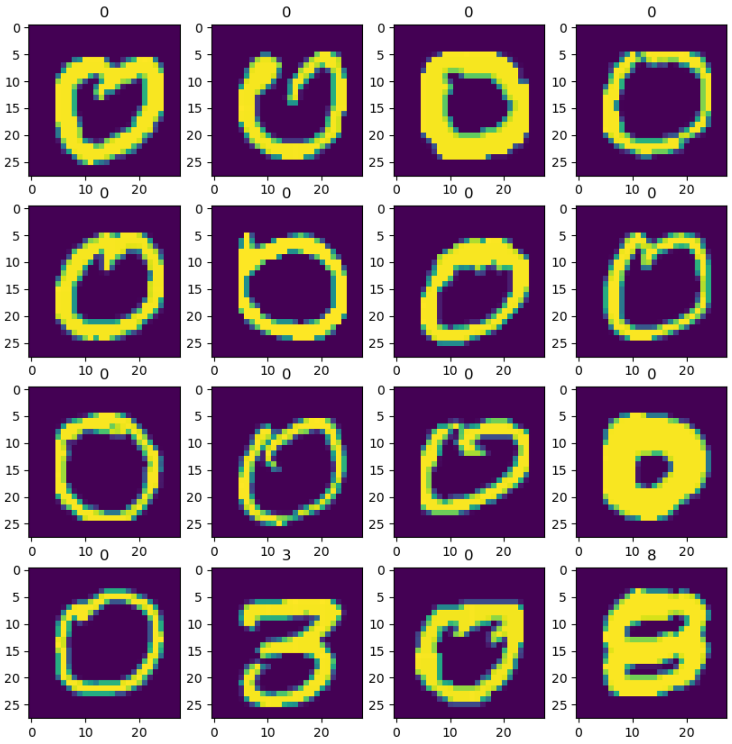
Using signature to select more of the same



Anomaly → Signature → Feature Mask

↓

Retraining with selected features



→ Similar Anomalies

Conclusions

- **Anomaly Signature** is a metric for feature importance
- **Domain agnostic / method aware** (Isolation Forest)
  - Works with any tabluar data
- ***Many use cases:***
  - Interpretability of the decisions
  - Visualisation of outliers
  - Feature selection
  - Categorization of outliers
  - Active learning of anomalies

This is only the beginning !

Stay tuned on

SNAD