Outlier Detection

Dalya Baron (Tel Aviv University)

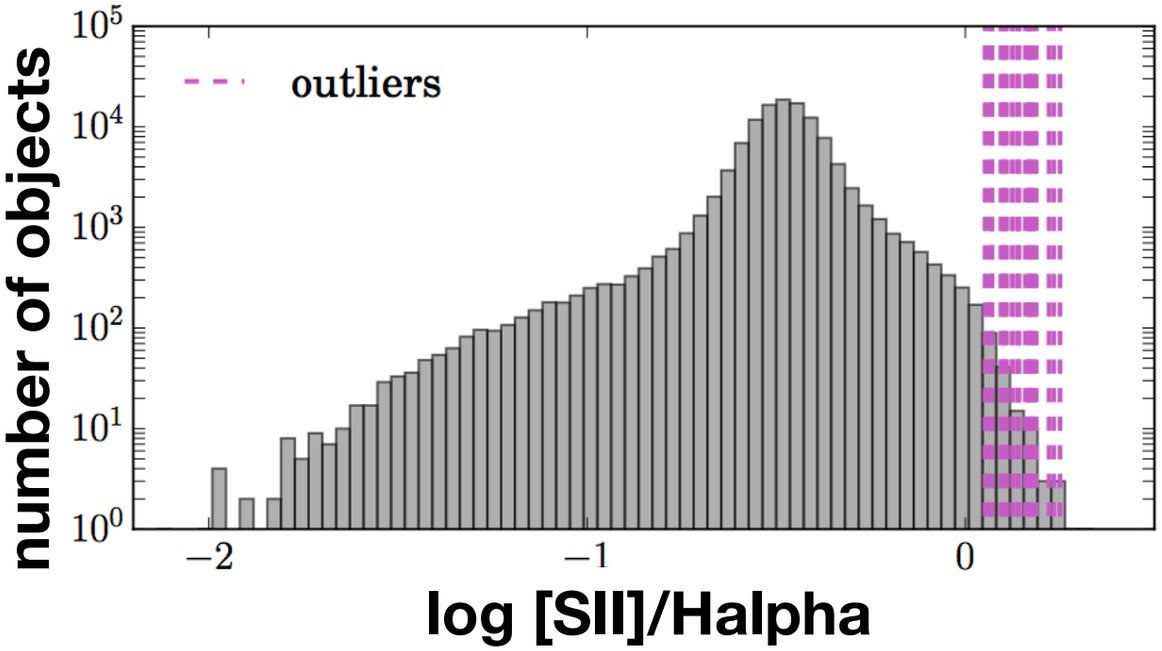
XXX Winter School, November 2018

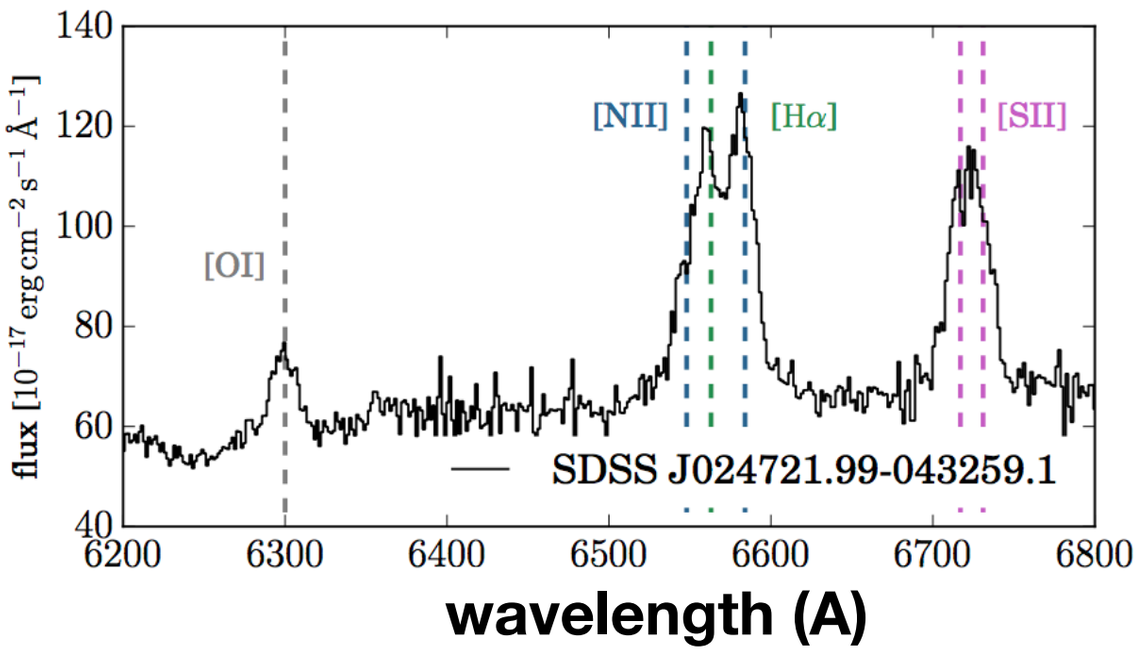
What is an outlier?

- **“Bad” object:** artifacts, cosmic rays, bad reduction.
- **Misclassified object:** star classified as QSO, variable star classified as SN.
- **Tail of a distribution:** most luminous SN, fastest accreting BH.
- **Unknown unknowns:** completely new objects we did not know we should be looking for.
- **In astronomy:** processes which happen on shorter time scales.

How do we find outliers?

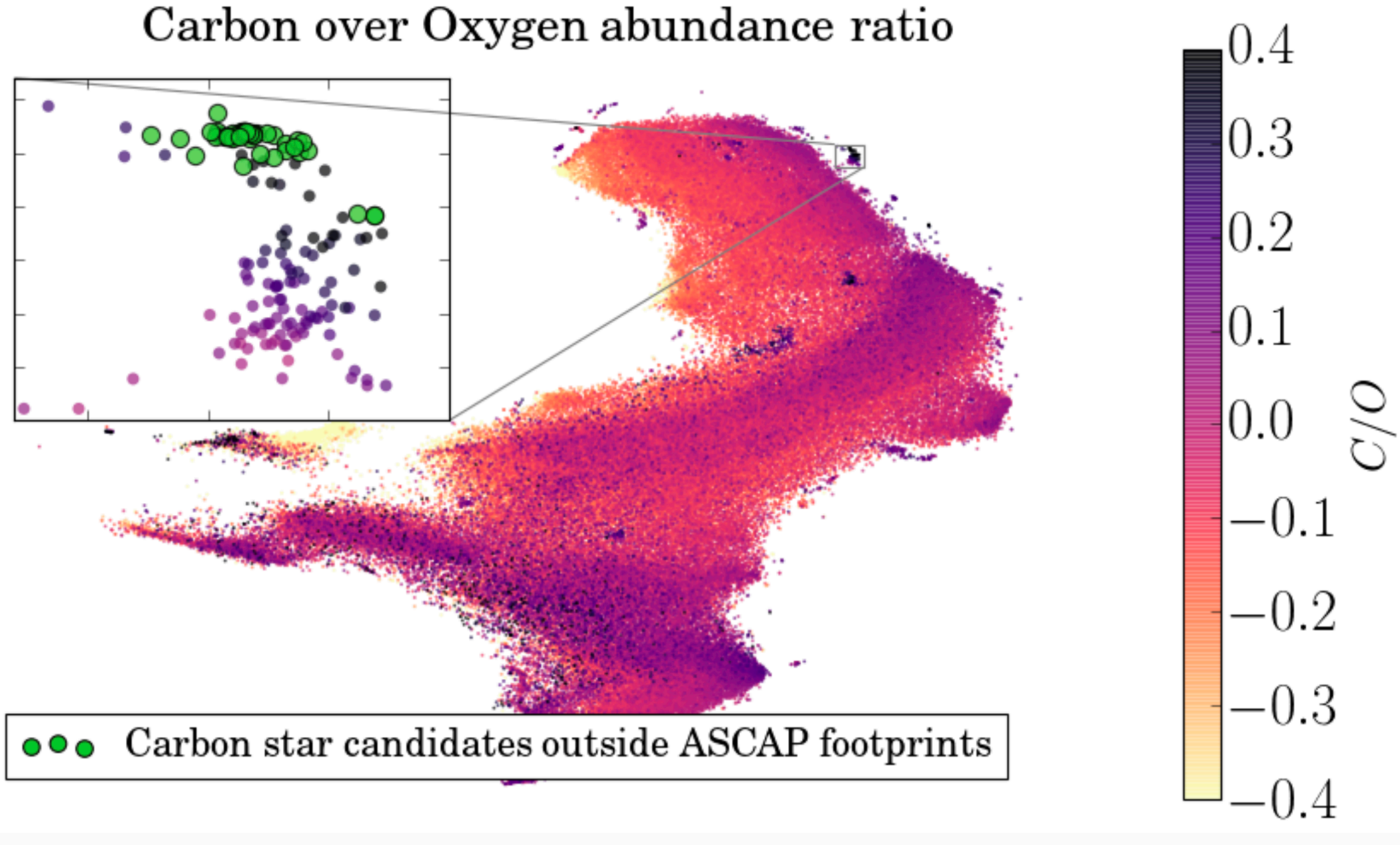
1. Stay as close as possible to the original dataset. Measured features will usually not help (see Baron & Poznanski 2017).





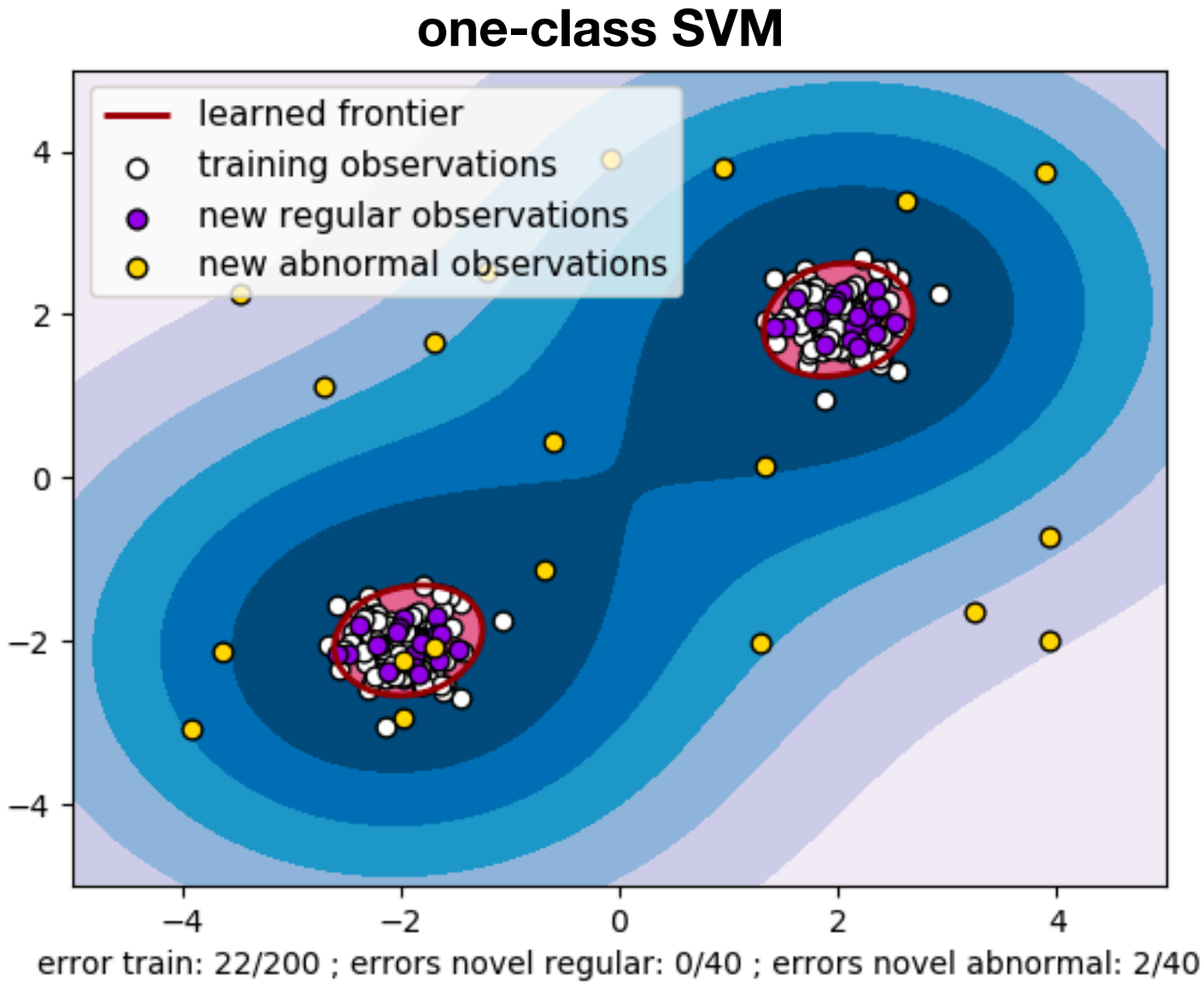
How do we find outliers?

1. Stay as close as possible to the original dataset. Measured features will usually not help (see Baron & Poznanski 2017).



How do we find outliers?

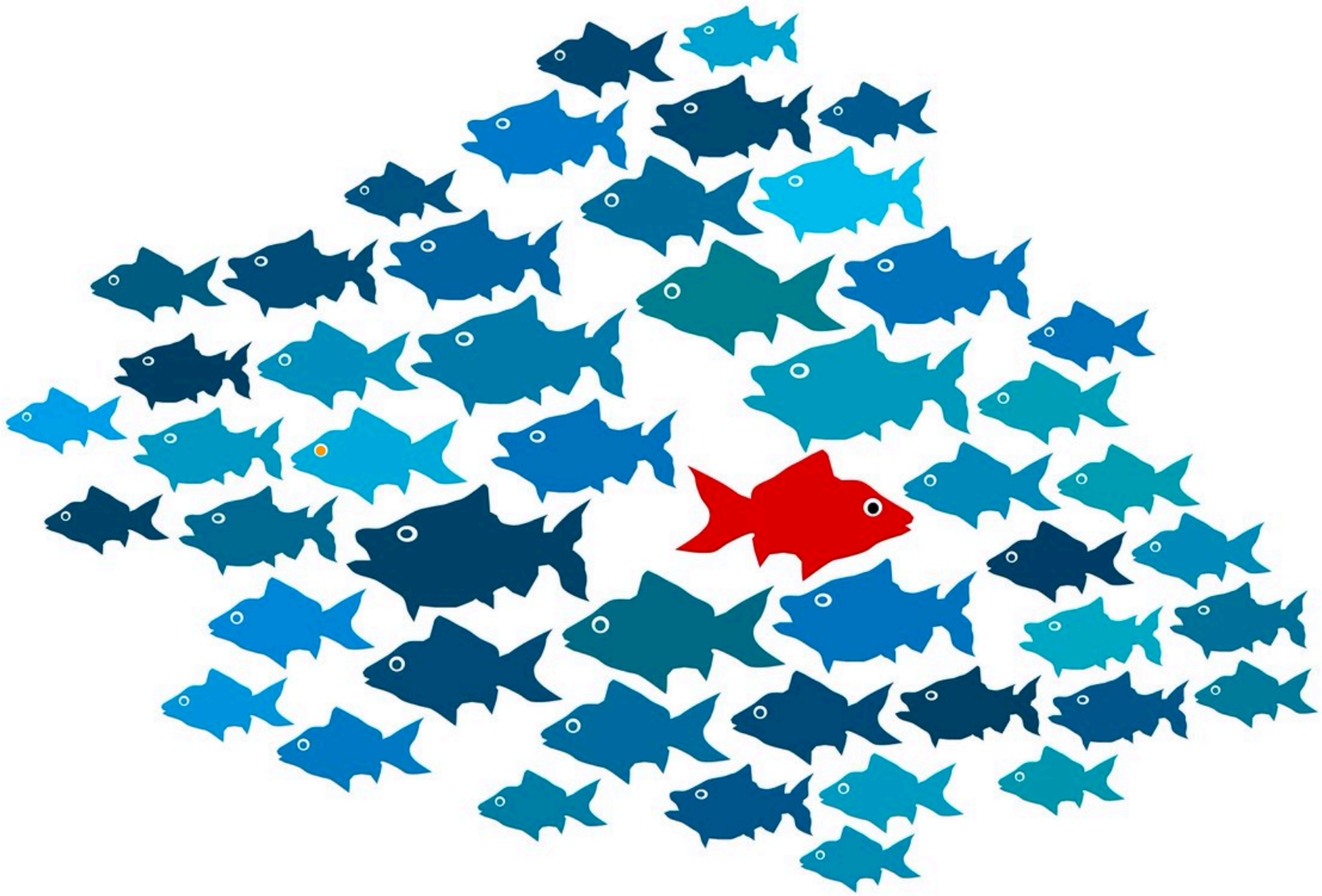
2. Supervised learning-based outlier detection will uncover the outliers that “shout the strongest”.



Additional supervised algorithms that can be used: **RF, ANN, etc..**

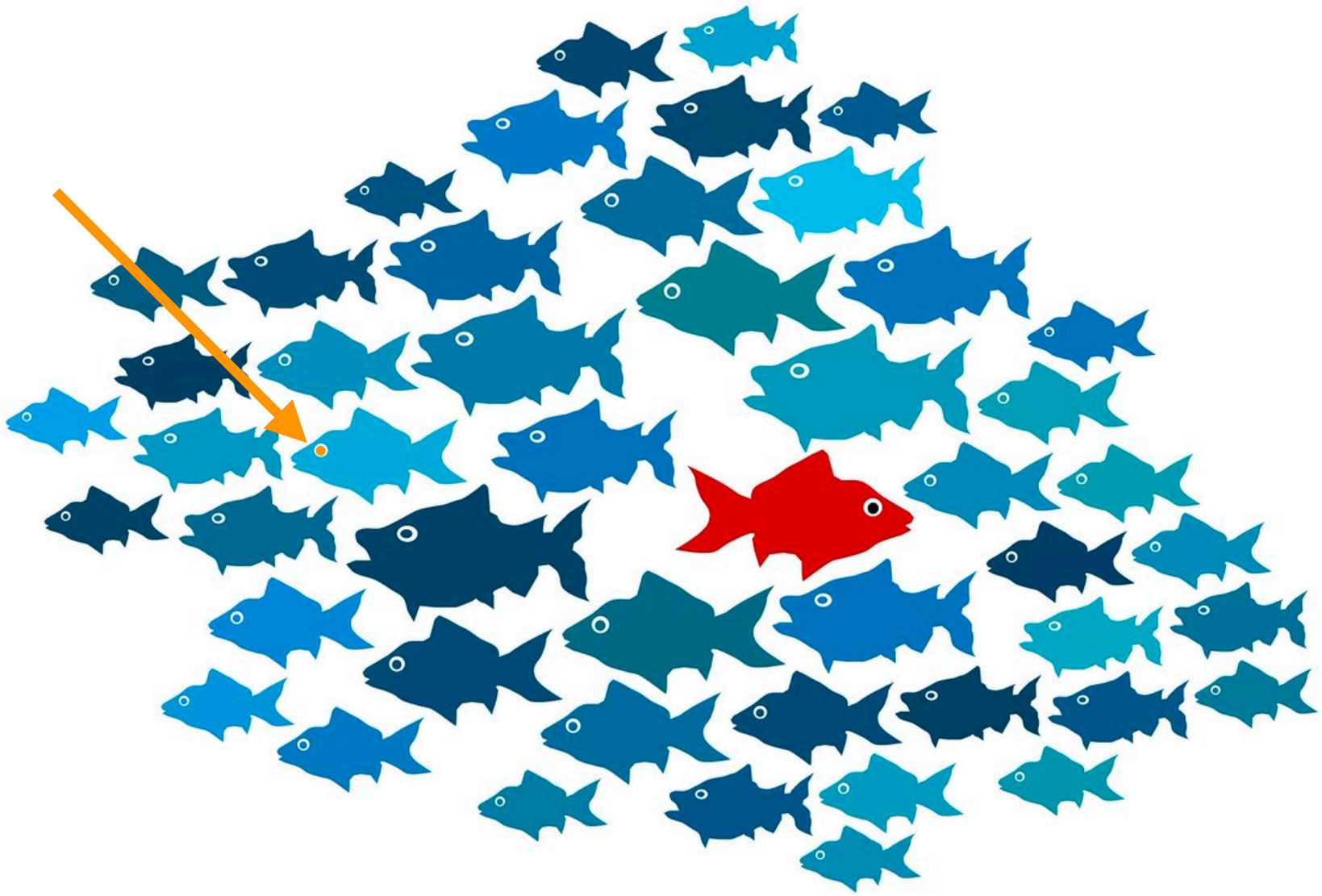
How do we find outliers?

2. Supervised learning-based outlier detection will uncover the outliers that “shout the strongest”.



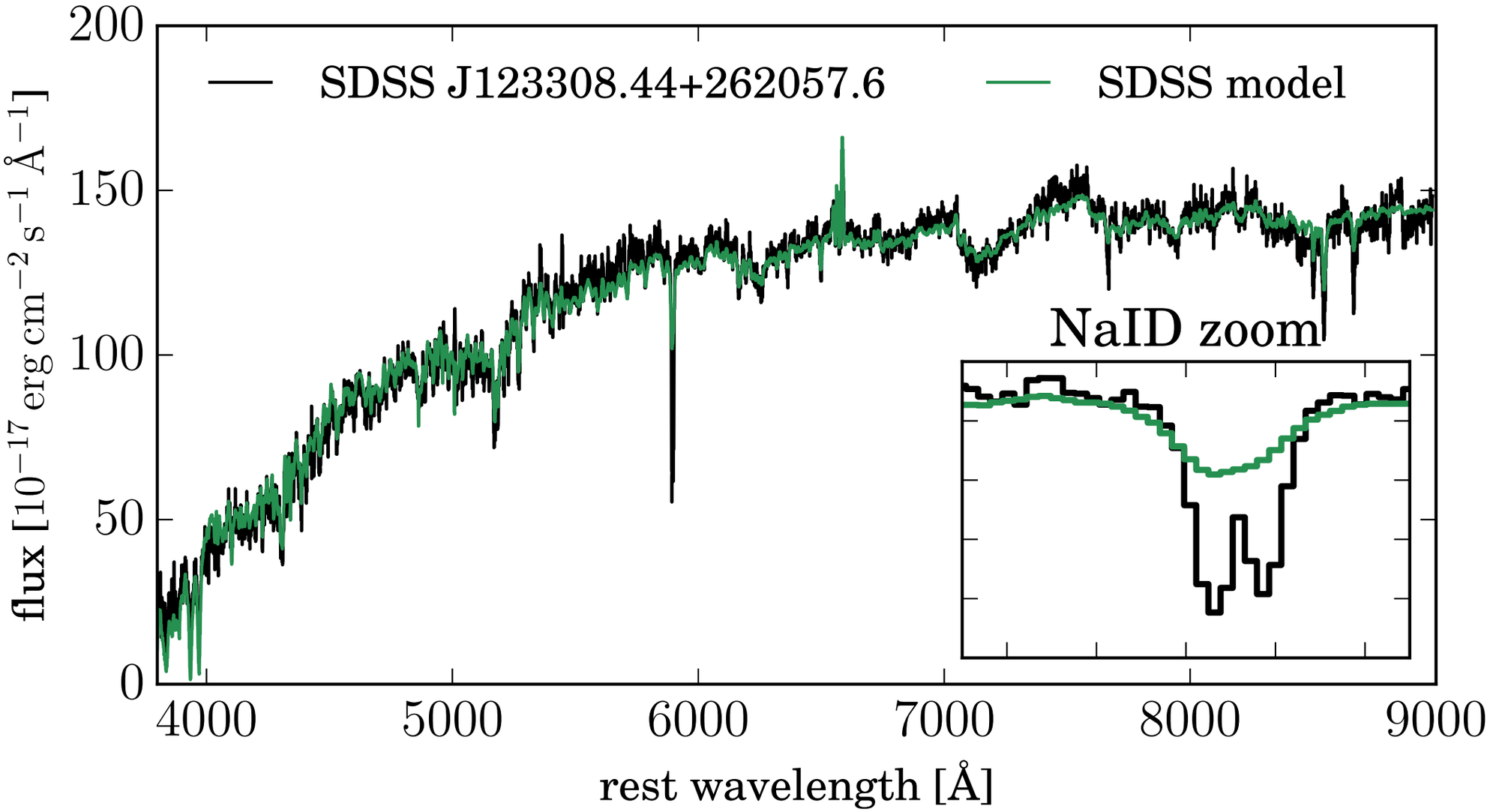
How do we find outliers?

2. Supervised learning-based outlier detection will uncover the outliers that “shout the strongest”.



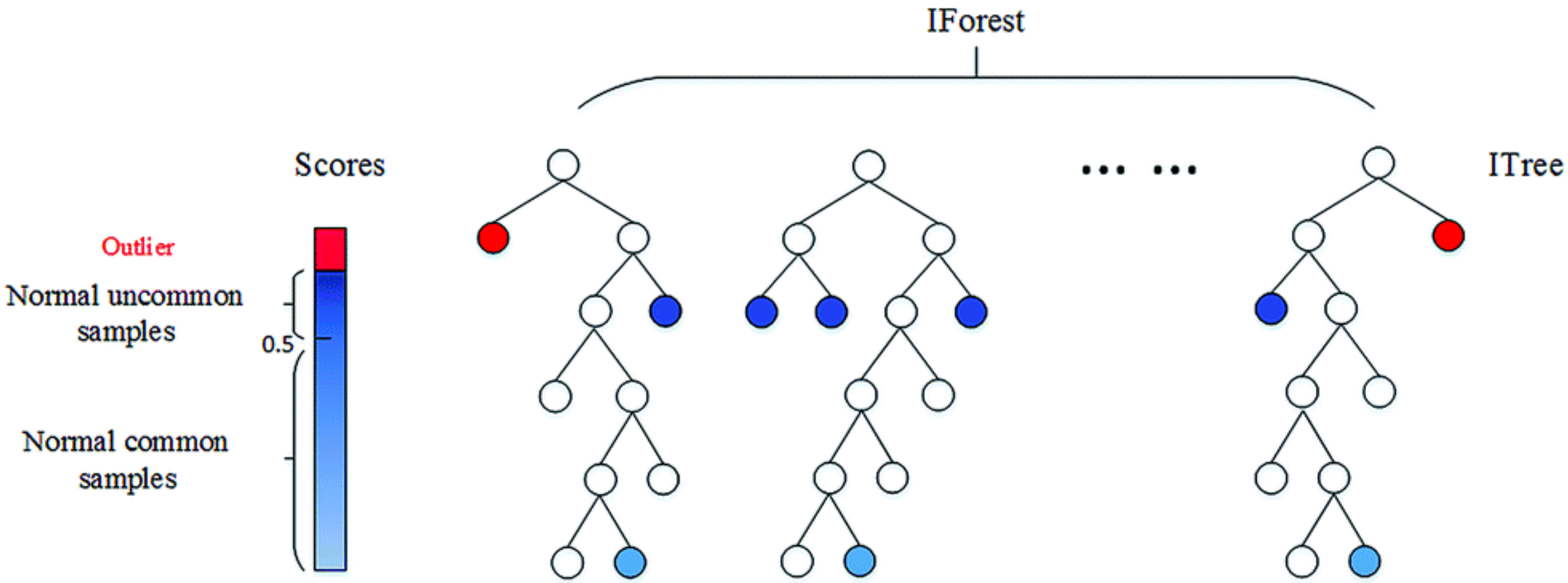
How do we find outliers?

2. Supervised learning-based outlier detection will uncover the outliers that “shout the strongest” (see Baron & Poznanski 2017).



Isolation Forests

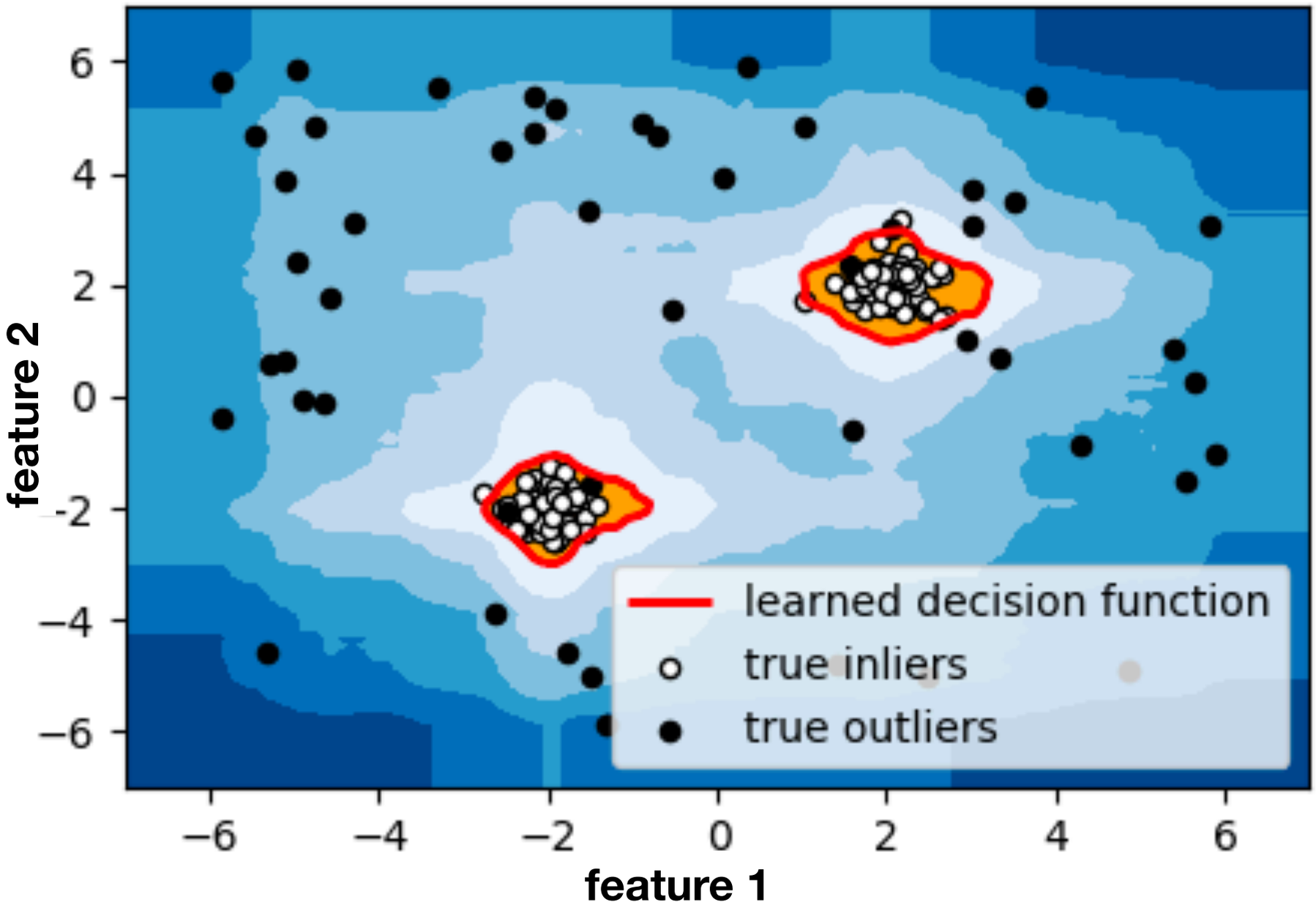
A fast algorithm that does not require distance measurements.
Outliers are objects that are separated from the rest of the dataset higher in the tree.



http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

Isolation Forests

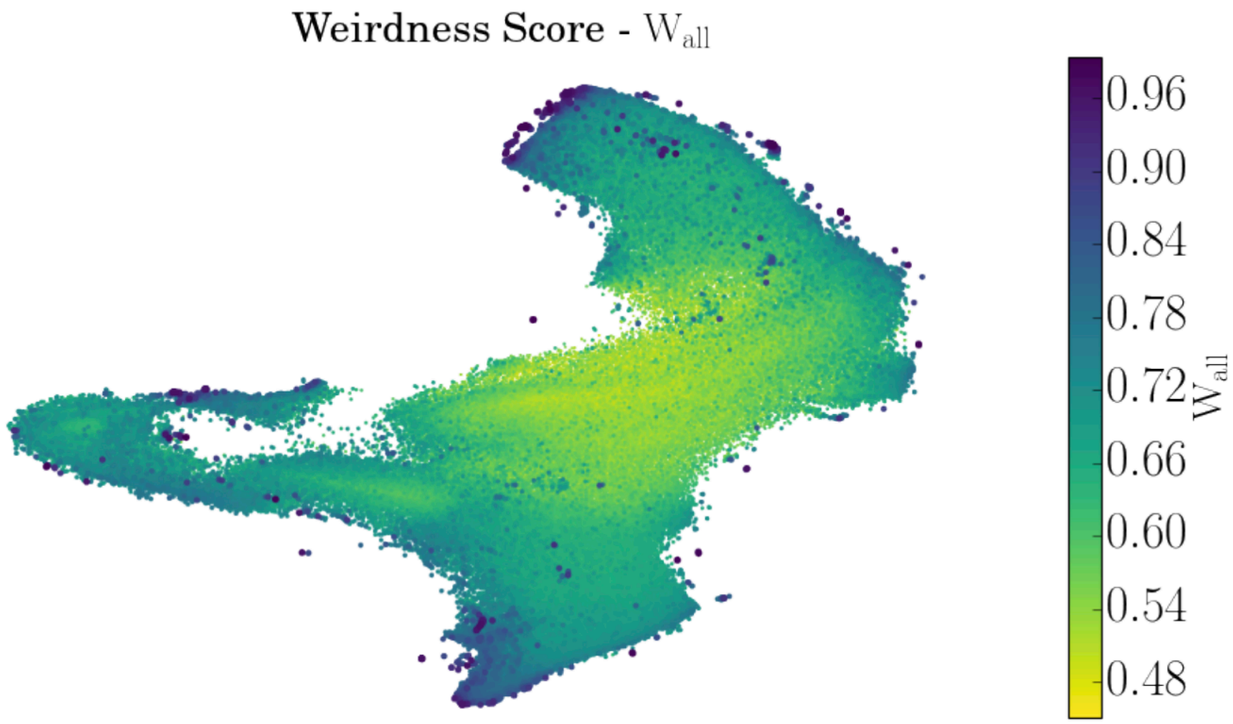
Isolation Forests will also find the outliers that shout the most.



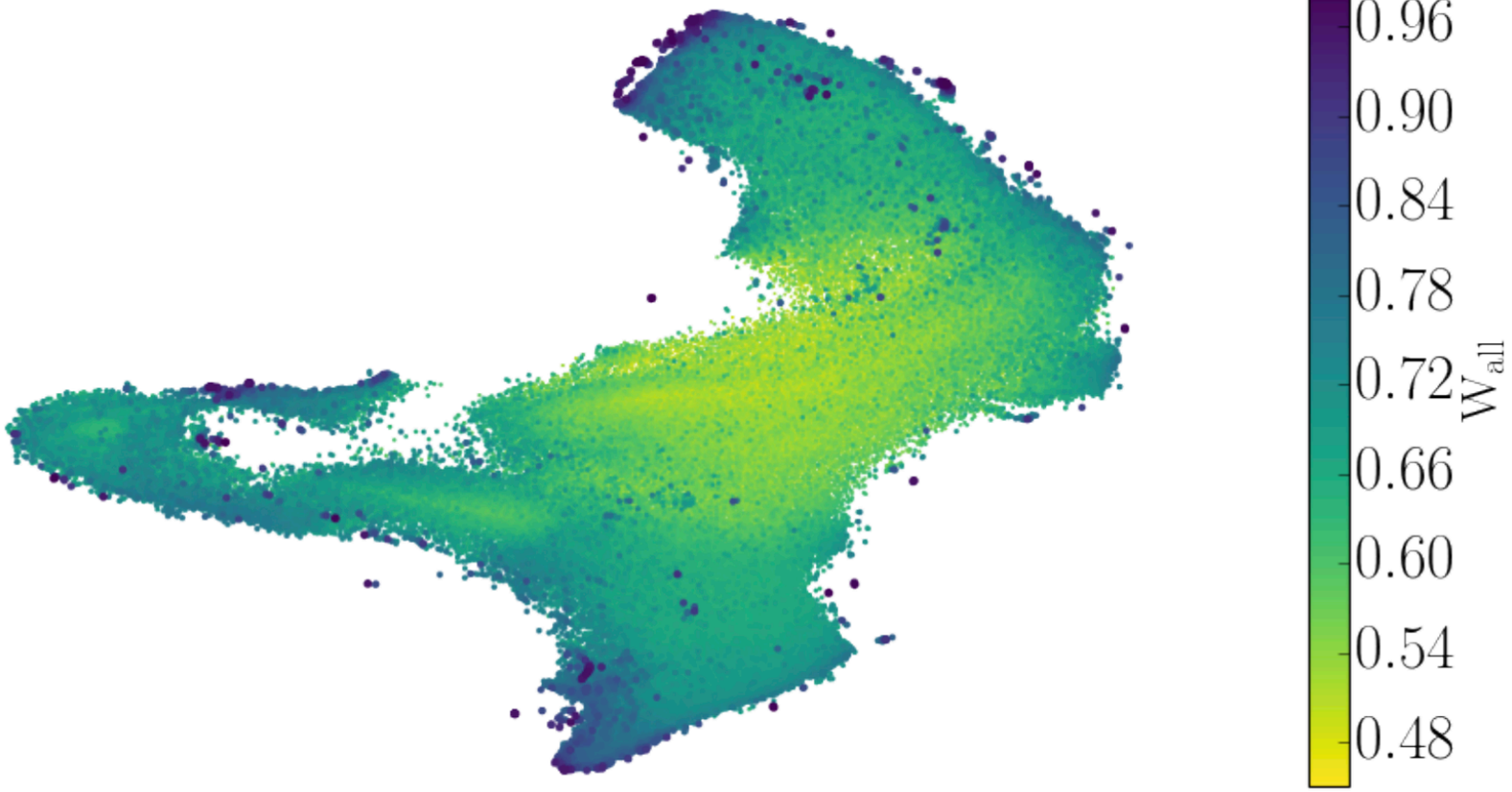
Outlier detection on spectra using unsupervised RF

See: Baron & Poznanski 2017, Reis+18

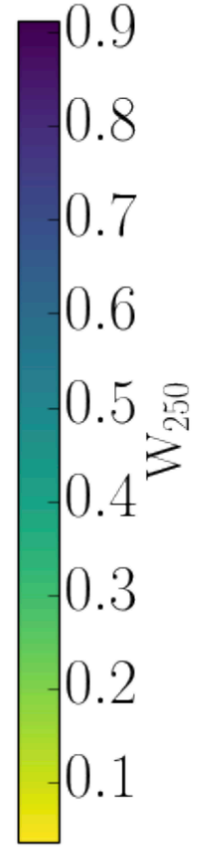
Use unsupervised Random Forest to define distances between the objects in the sample. Define outliers are objects with a large average distance from all the rest, or from their ~250 nearest neighbors. Use **tSNE** to explore outlier clusters.



Weirdness Score - W_{all}

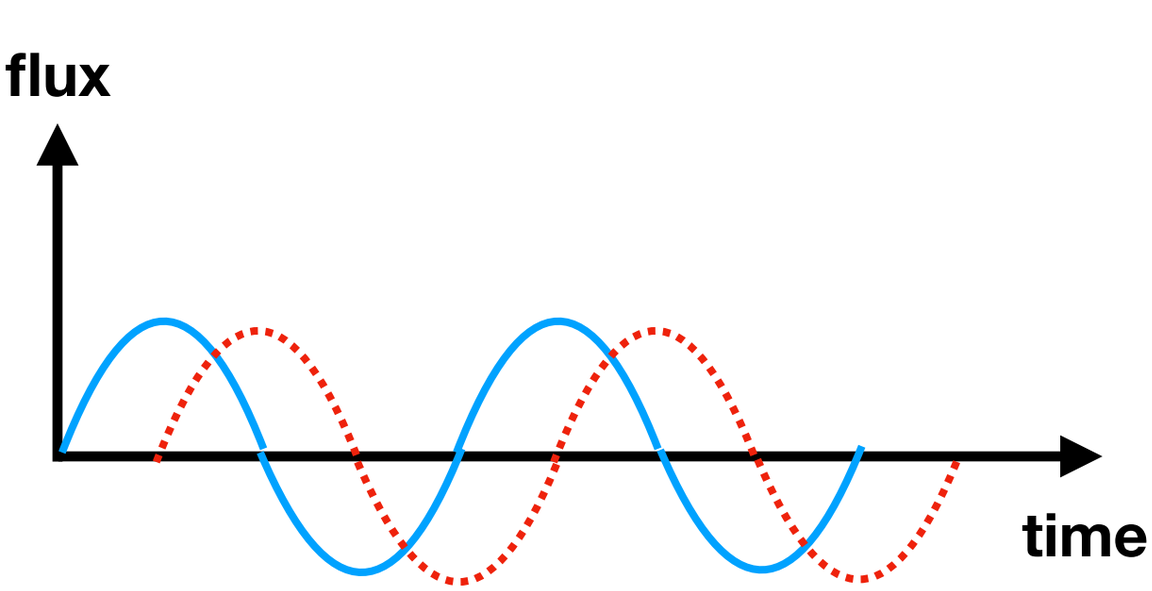


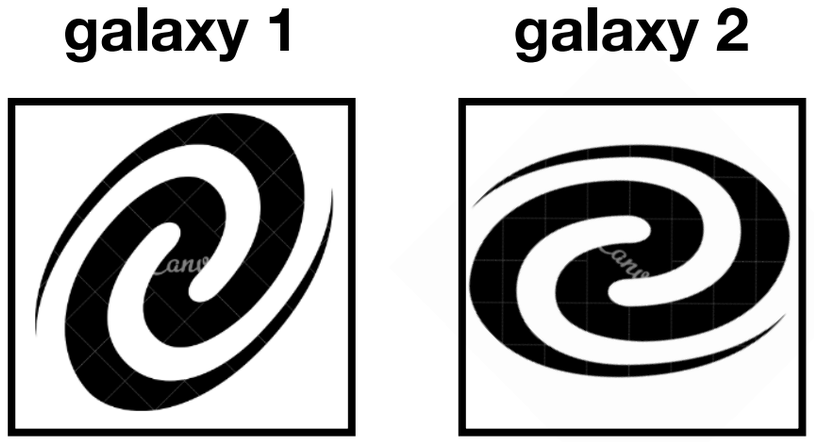
Weirdness Score - W_{250}



Outlier detection on other datasets using unsupervised RF?

The unsupervised Random Forest assumes a regular grid, and thus will work for spectra or extracted features.
It will not work for images or time series, because it does not have translational and rotational symmetry!

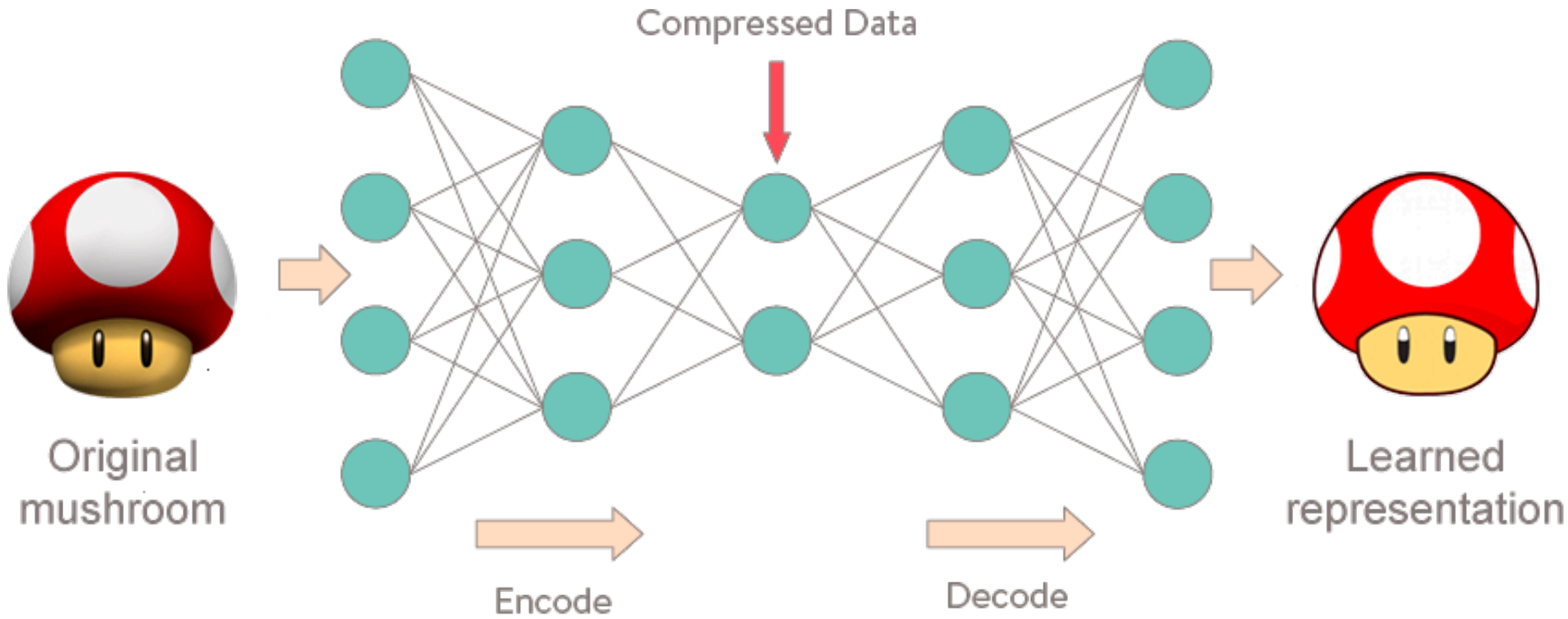




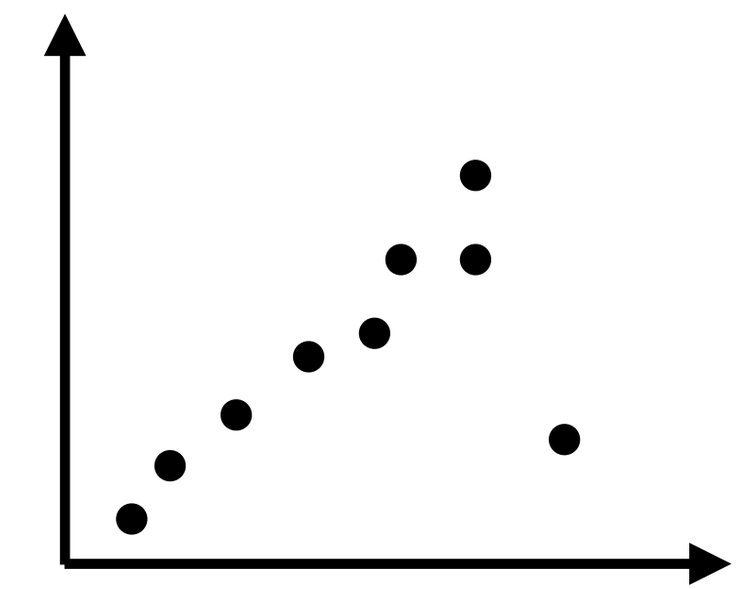
possible solution: find a representation of the signal on a regular grid (e.g., FFT of time series).

Outlier detection using Autoencoders

Autoencoders can represent images (CNN) and time-series (RNN) well.
Use the latent space to look for, and examine outliers.

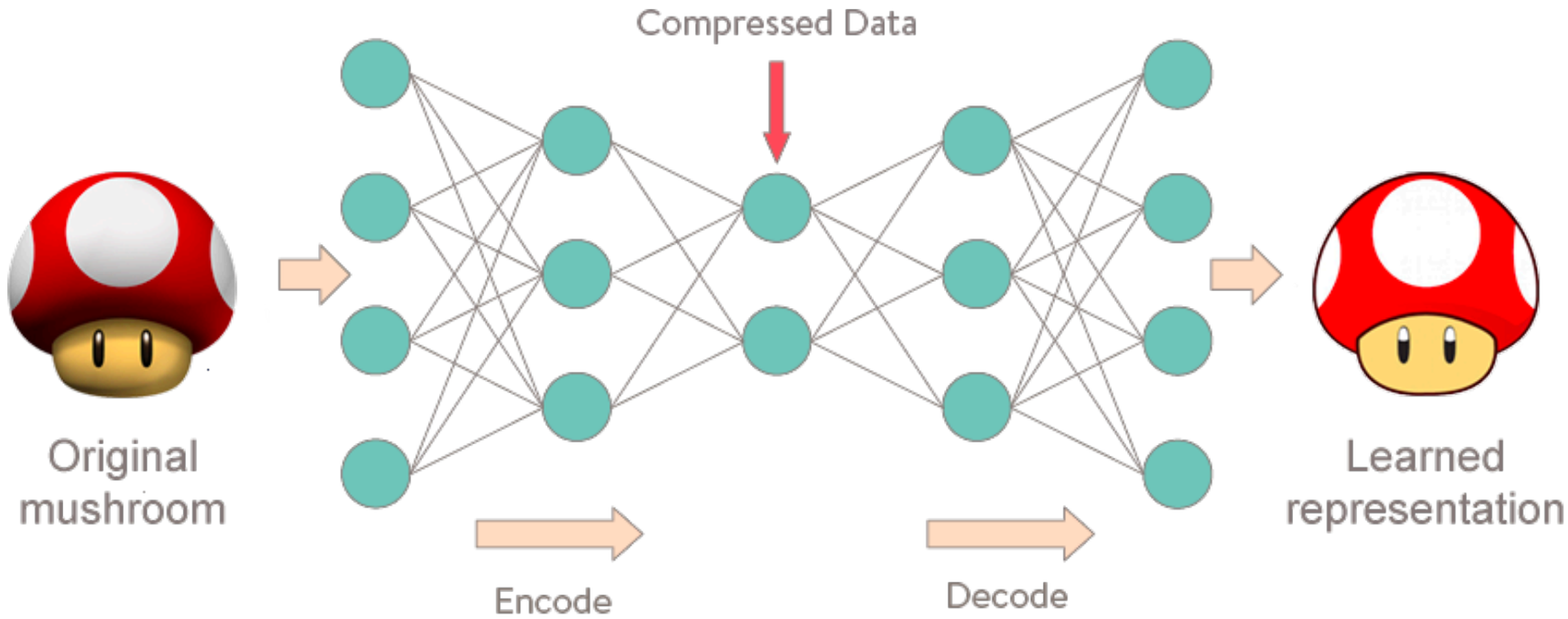


Latent (compressed) space:

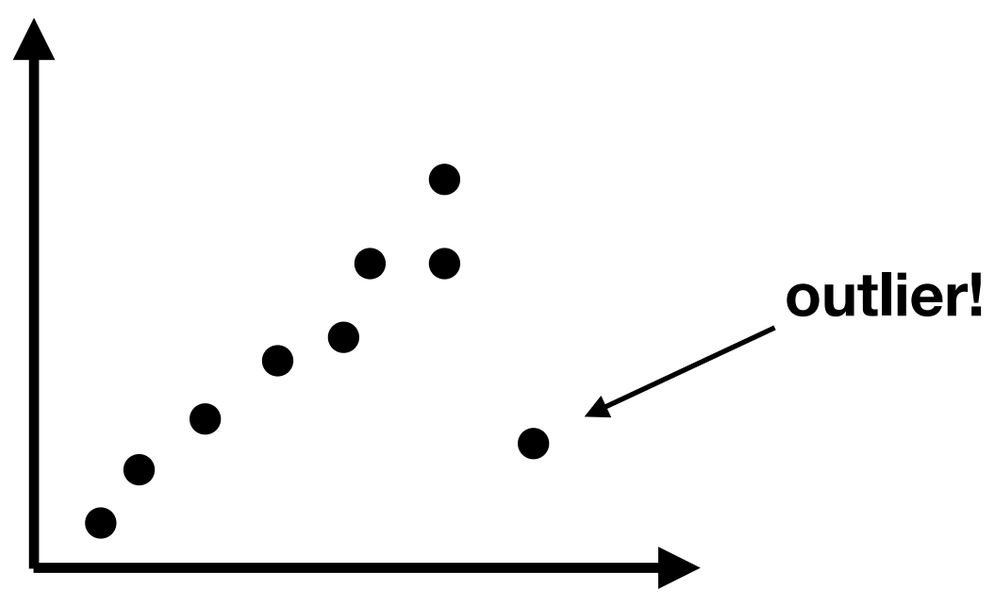


Outlier detection using Autoencoders

Autoencoders can represent images (CNN) and time-series (RNN) well.
Use the latent space to look for, and examine outliers.

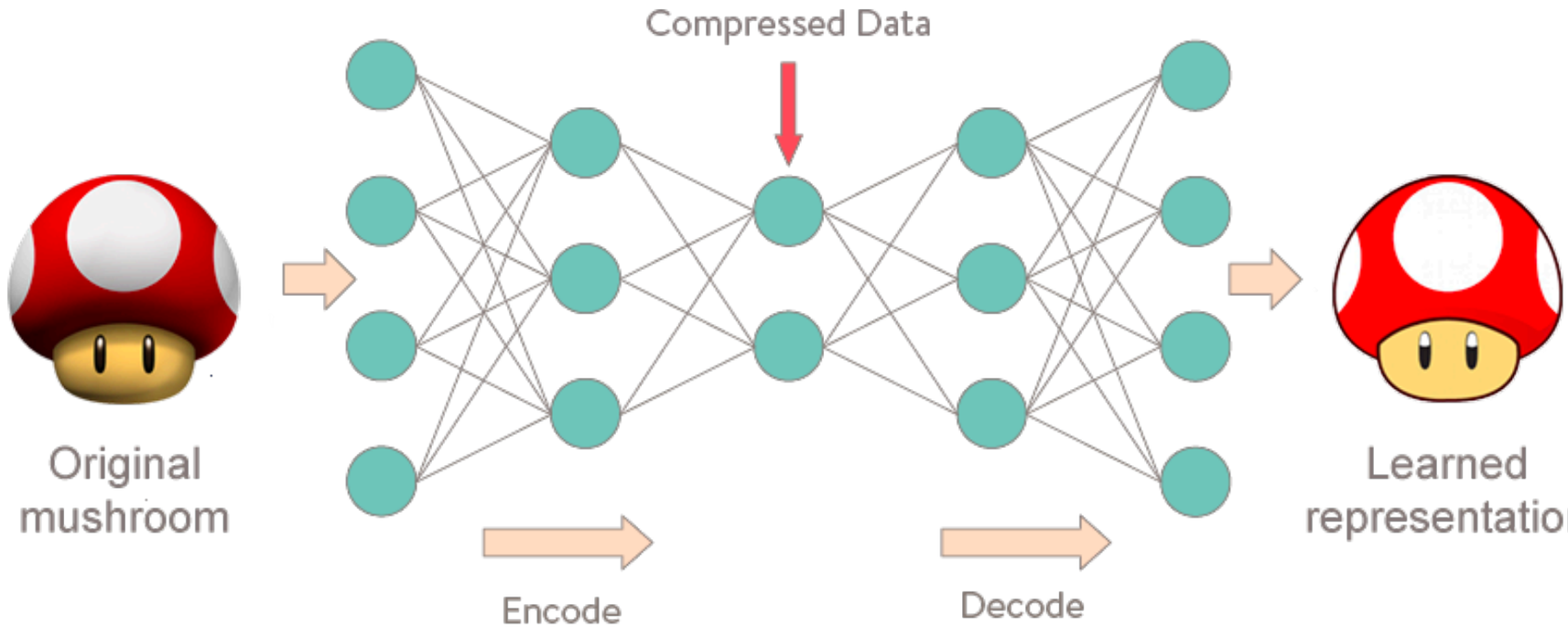


Latent (compressed) space:

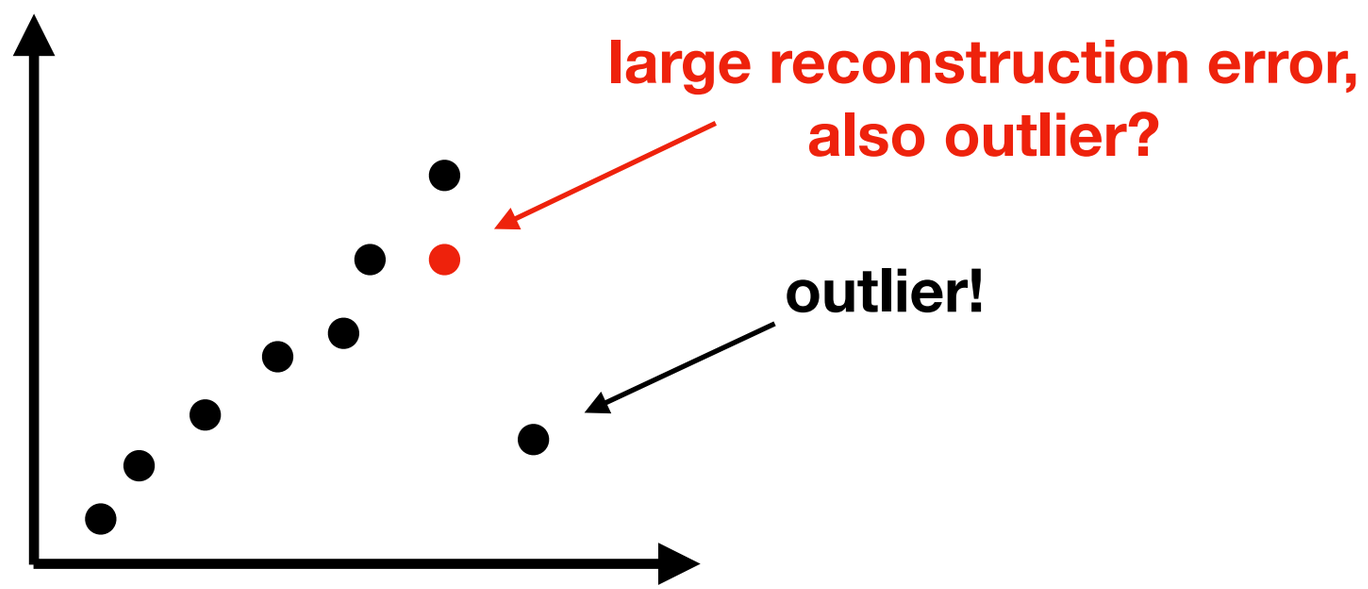


Outlier detection using Autoencoders

Autoencoders can represent images (CNN) and time-series (RNN) well.
Use the latent space to look for, and examine outliers.



Latent (compressed) space:



Questions?

Tutorials

Summary

- Machine learning algorithms are just a tool. Not every problem is appropriate for machine learning.
- Unsupervised (and most supervised) algorithms are not black boxes!!
- Don’t trust me.
- Make friends with computer scientists and engineers.
- Don’t be afraid to change off-the-shelf tools.
- **Open questions:**
  - Uncertainties in supervised and unsupervised algorithms.
  - Unsupervised: we need better cost functions.

Tell me more about your science!

Thanks! :)

Tutorials

- Clustering algorithms, dimensionality reduction, supervised learning and outlier detection: https://github.com/dalya/XXX_winter_school

- Unsupervised Random Forest and outlier detection: https://github.com/dalya/WeirdestGalaxies

- Outliers and tSNE in APOGEE: https://github.com/ireis/APOGEE_tSNE_nb

- Probabilistic Random Forest: https://github.com/ireis/PRF