

---

# Outlier detection

Dalya Baron  
Carnegie Observatories

---

*Vatican Observatory Summer School on Big Data and  
Machine Learning 2023 (VOSS-2023)*

# What is an outlier?

---

- ❖ **Bad object:** reduction problems, cosmic rays, legos floating around the earth...
- ❖ **Misclassified object:** objects that were incorrectly selected into our dataset. For example: a star in a sample of quasars, variable star accidentally classified as a transient, etc.
- ❖ **Tails of distributions:** objects of the same class that show extreme values in one of their properties.
- ❖ **Unknown unknowns:** objects we did not know we should be looking for, and might represent something new and exciting.

# What is an outlier?

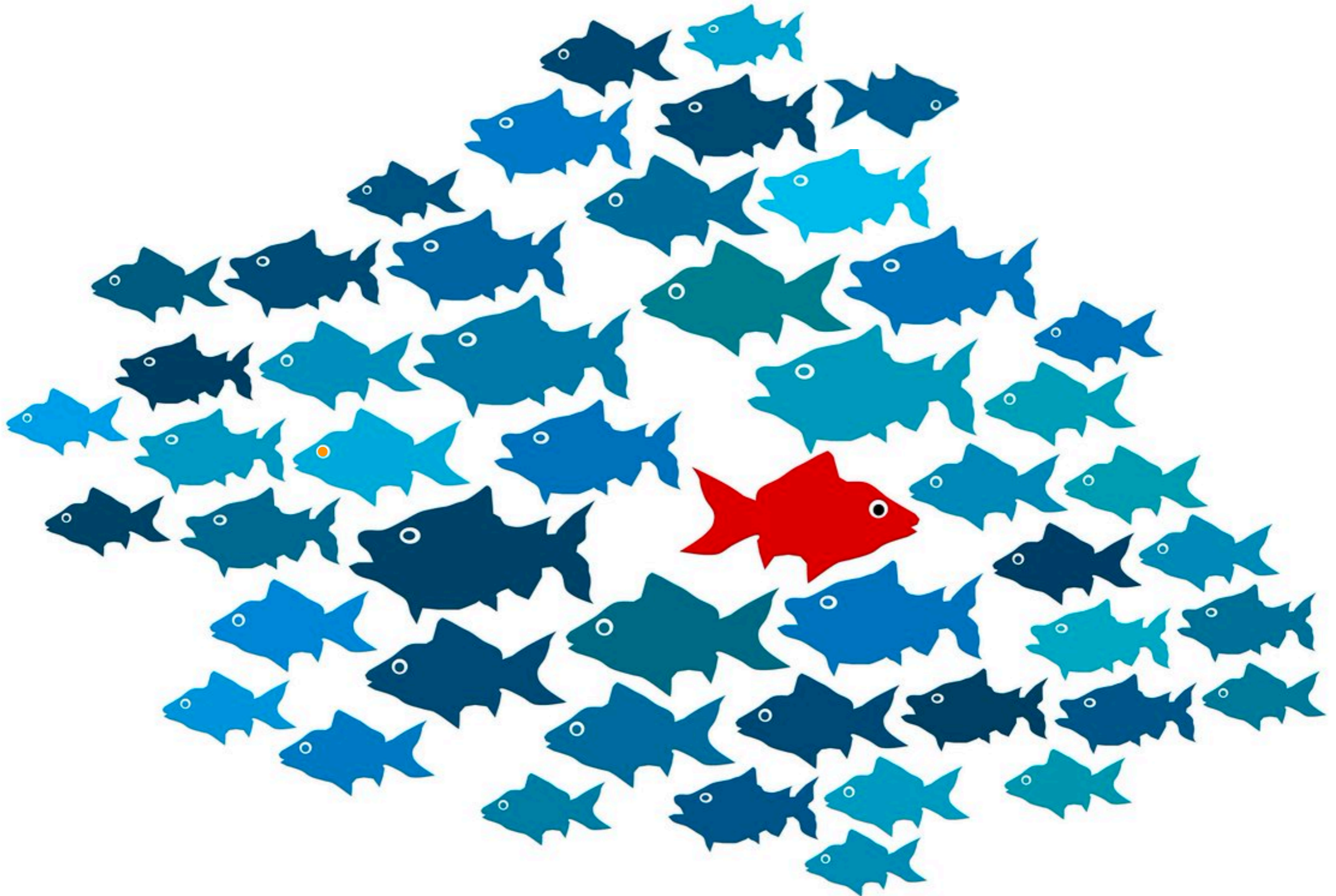
---

- ❖ **Bad object:** reduction problems, cosmic rays, legos floating around the earth...
- ❖ **Misclassified object:** objects that were incorrectly selected into our dataset. For example: a star in a sample of quasars, variable star accidentally classified as a transient, etc.
- ❖ **Tails of distributions:** objects of the same class that show extreme values in one of their properties.
- ❖ **Unknown unknowns:** objects we did not know we should be looking for, and might represent something new and exciting.

In astronomy, processes that take place on a shorter timescale will appear rare in our datasets.

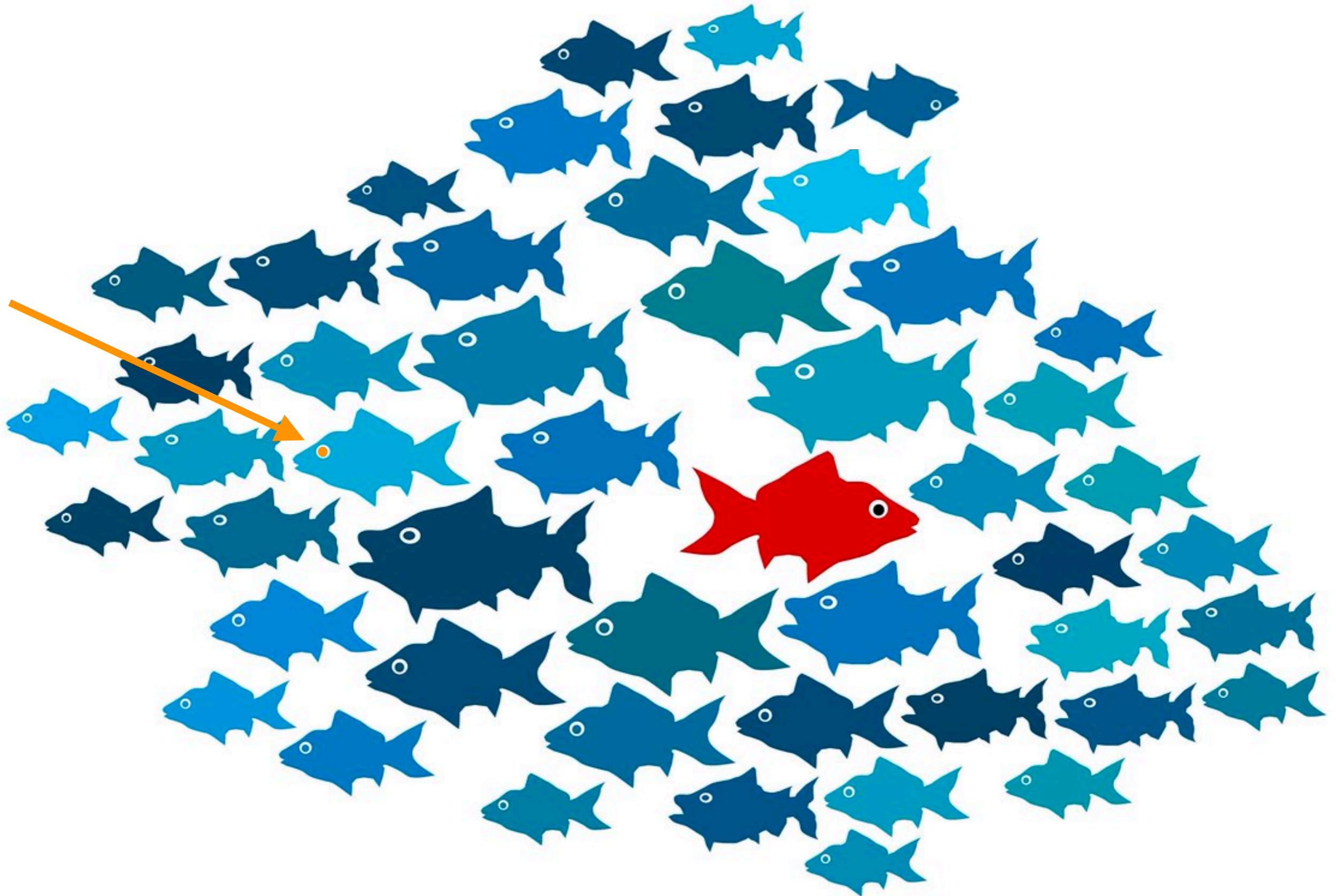
# Where is the outlier?

---



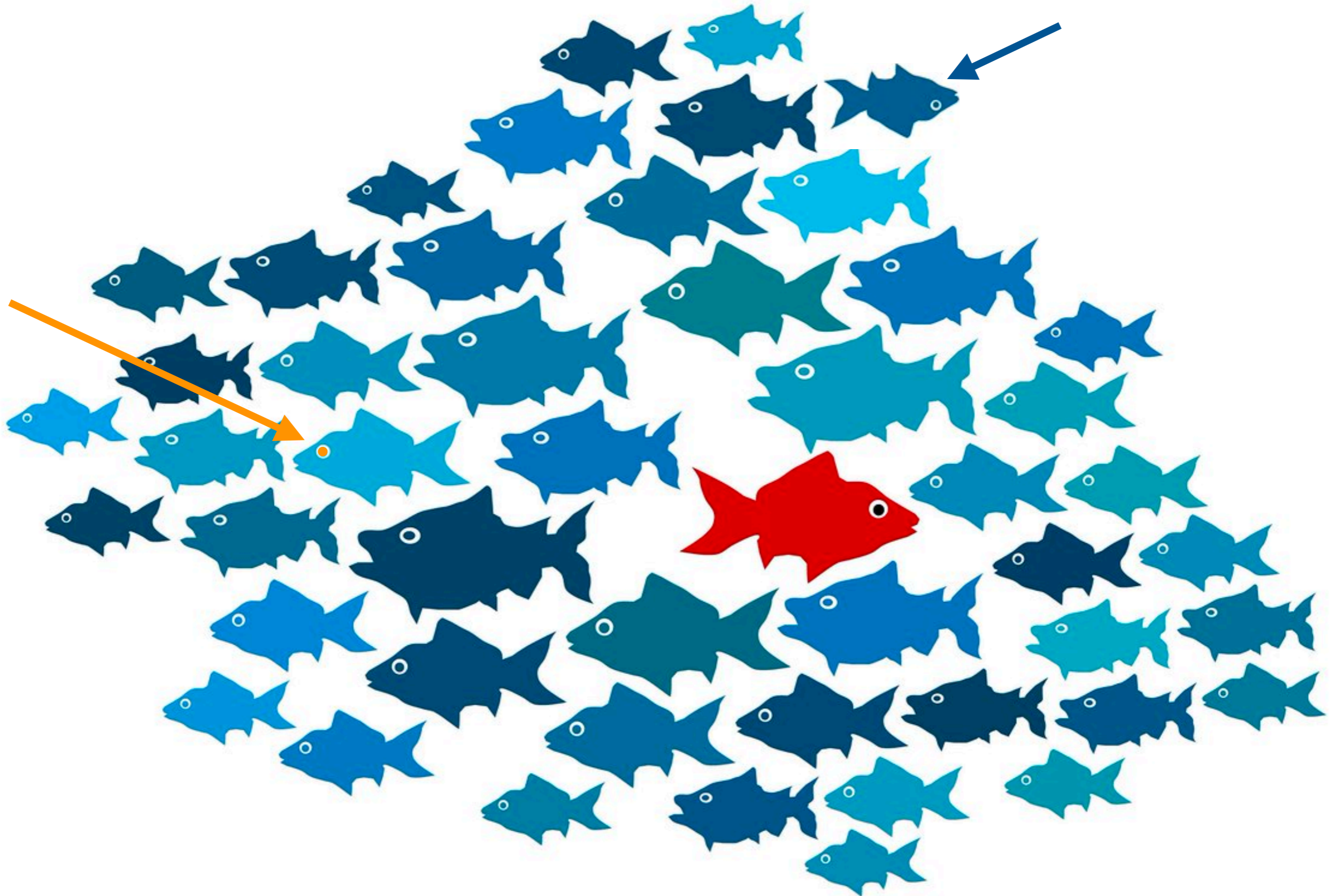
# Where is the outlier?

---



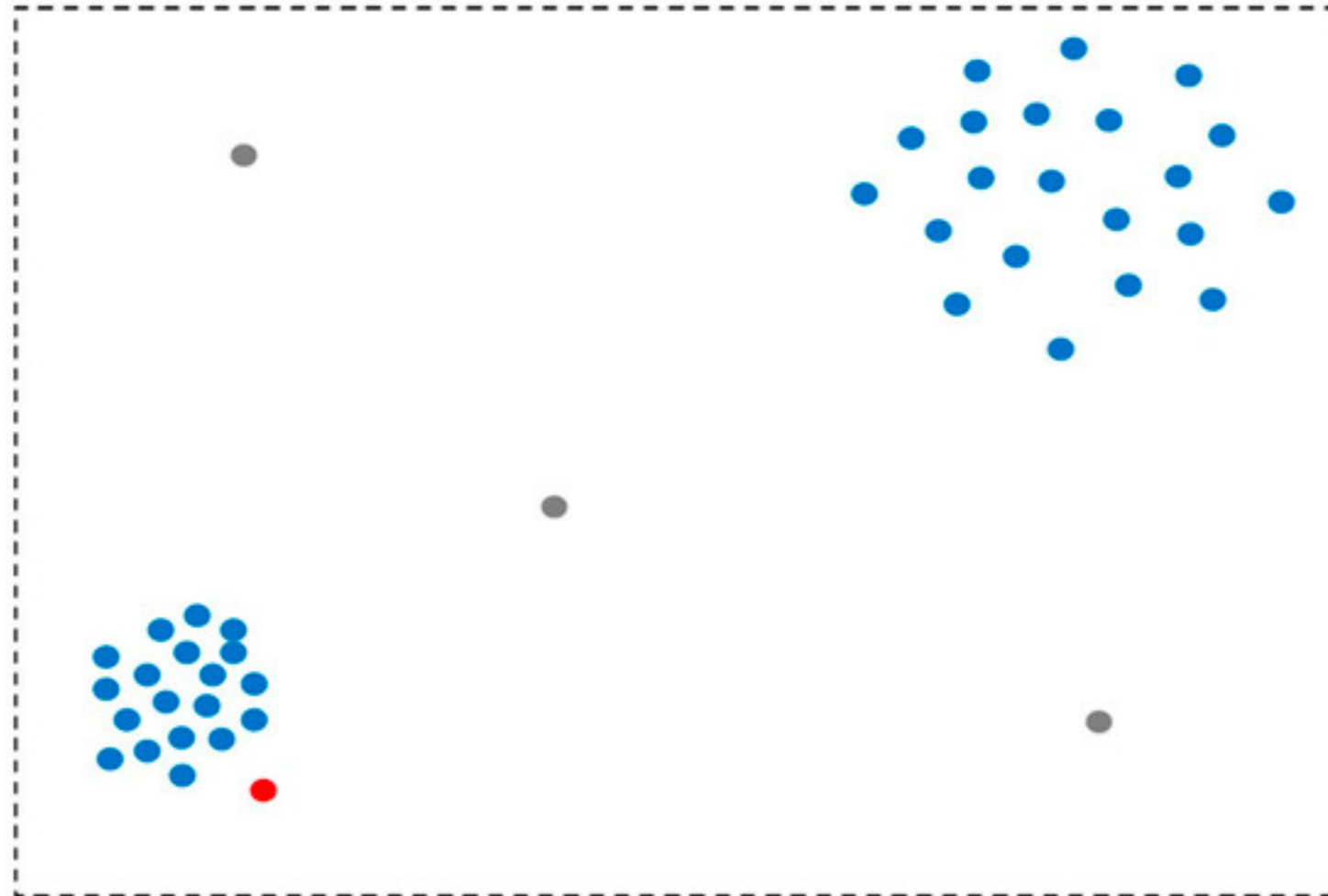
# Where is the outlier?

---



# Different types of outliers in feature space

---



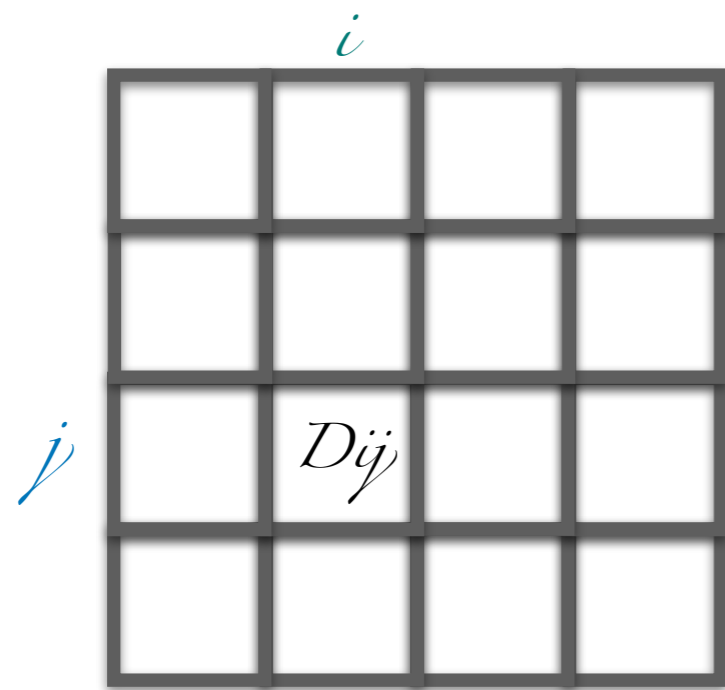
# How can we detect outliers?

---

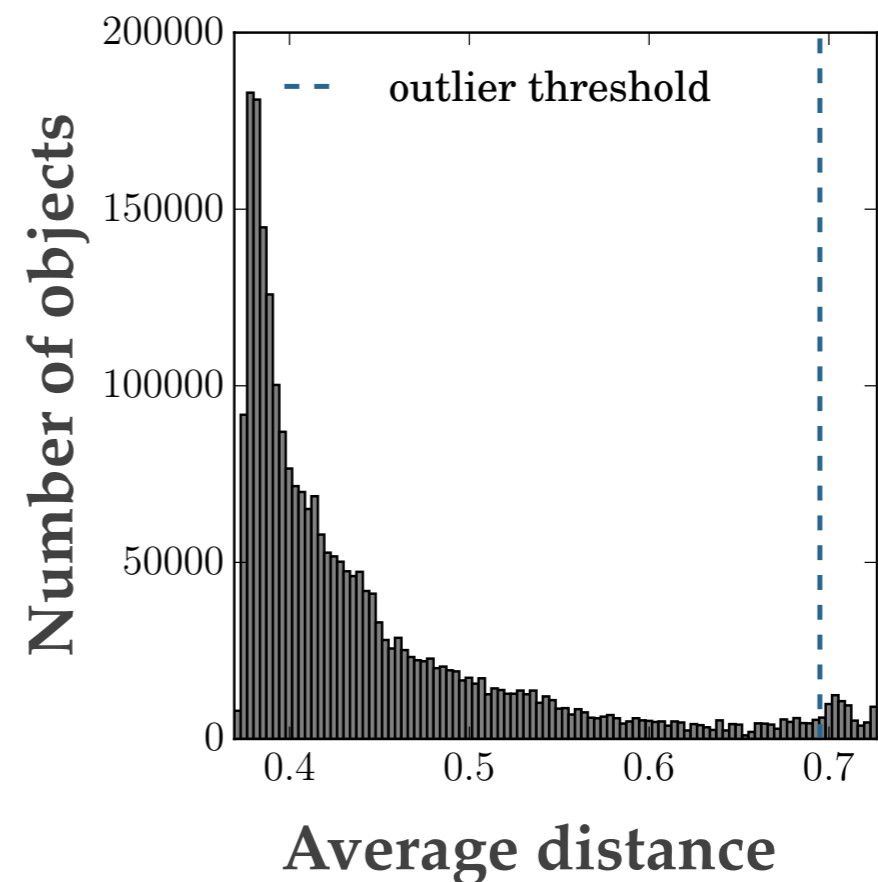
0. Have a domain expert manually-inspect all the objects in the dataset.

# How can we detect outliers?

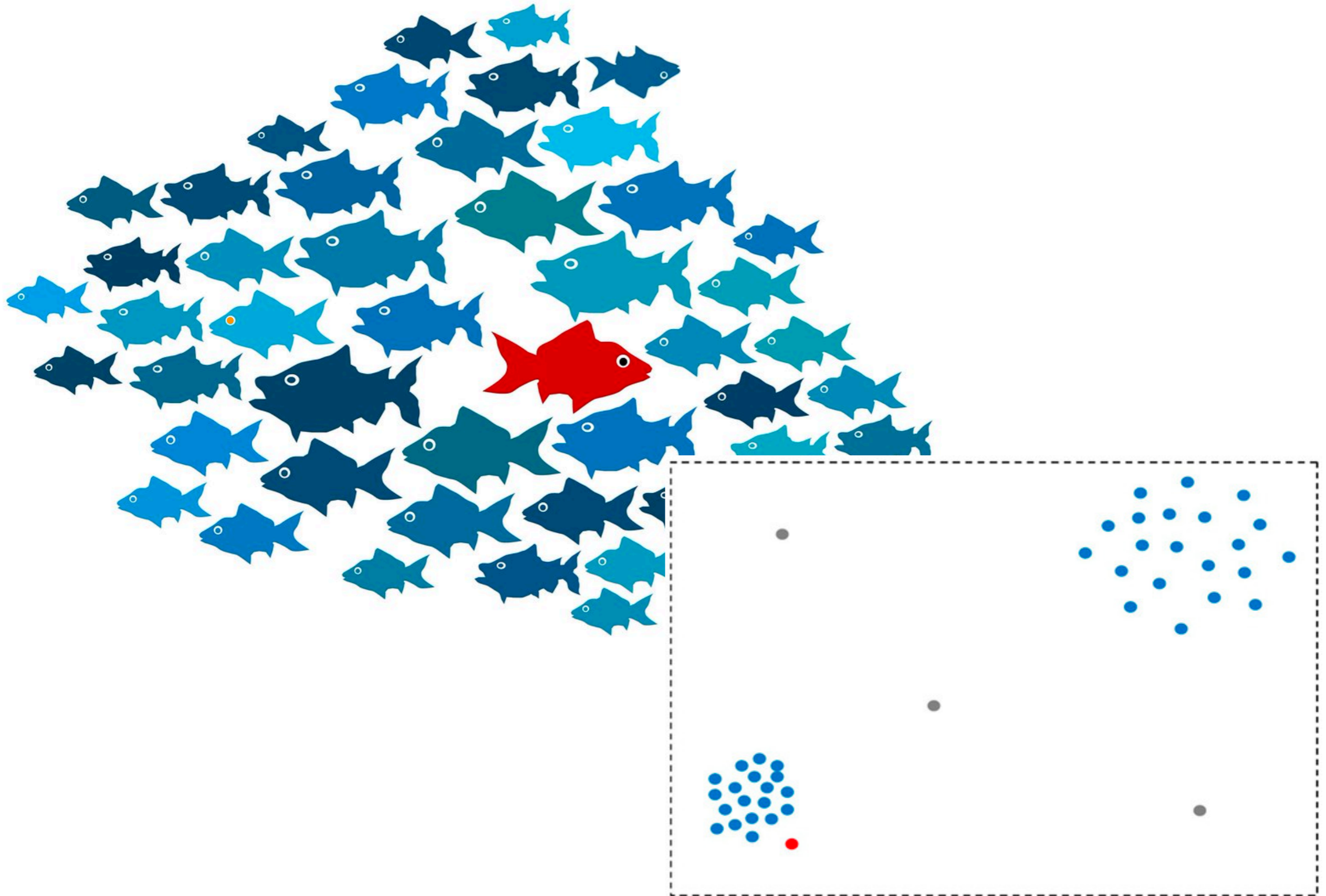
1. Measure pair-wise distances between all the objects in the sample and identify objects with large distances.



Distance matrix



# What type of outliers will we find?



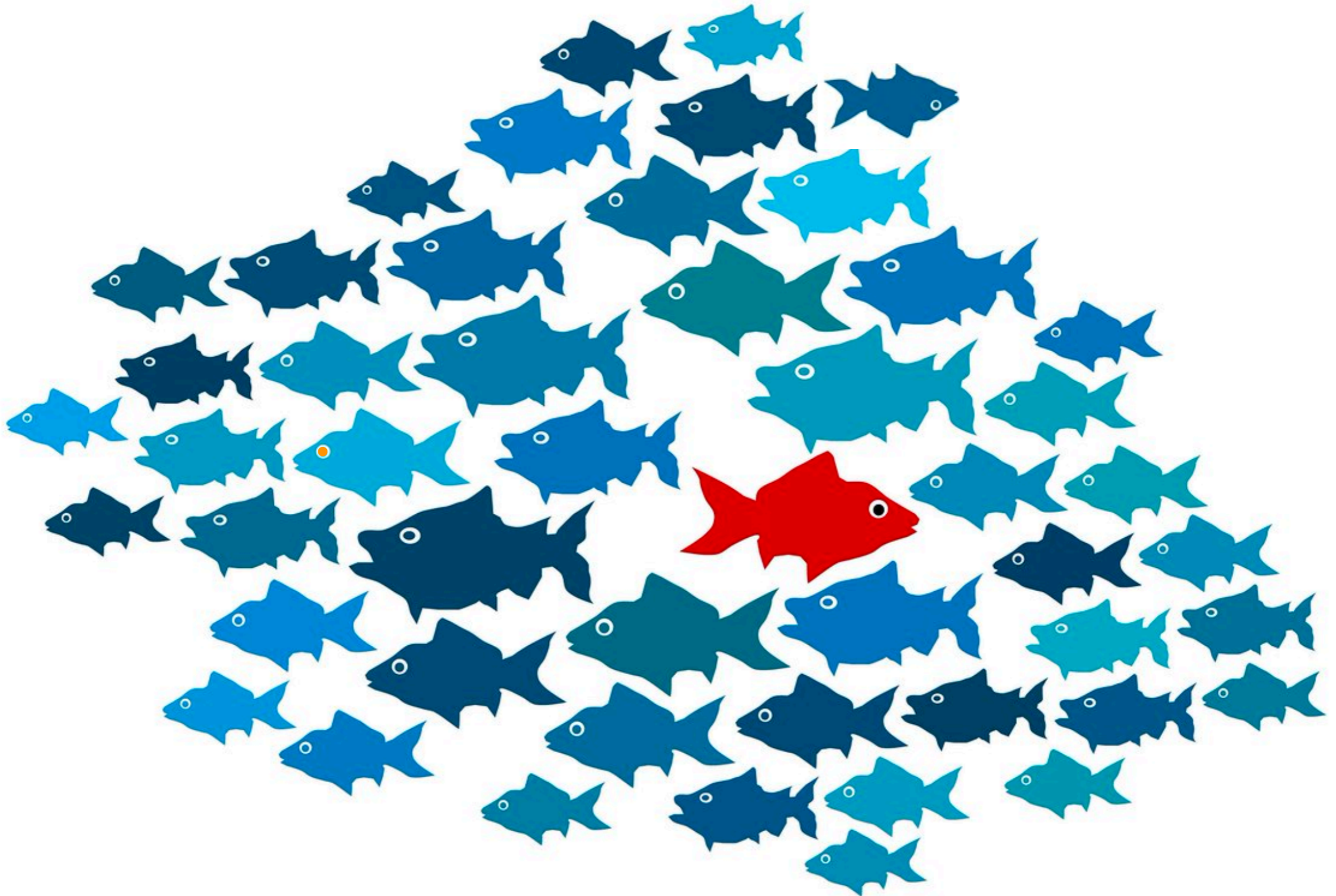
# How can we detect outliers?

---

2. Using Supervised Learning algorithms:
  - In the framework of a classification task, objects that have a relatively-low probability to belong to a class will be considered outliers (e.g., Random Forest).

# Which of the outliers will we find?

---



# How can we detect outliers?

---

## 2. Using Supervised Learning algorithms:

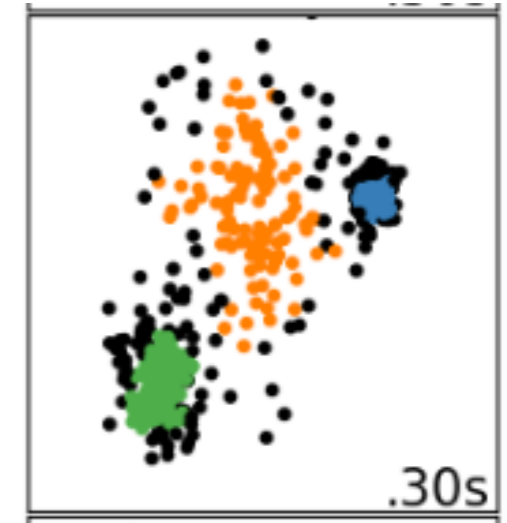
- In the framework of a classification task, objects that have a relatively-low probability to belong to a class will be considered outliers (e.g., Random Forest).
- Aspects to consider:
  - The type of outliers we will find will strongly depend on the classification problem.
  - Features that are less relevant for the classification task can be ignored and objects that show outlying properties in these features might not be flagged.
  - Objects with a known class but with outlying properties might not get flagged.

# How can we detect outliers?

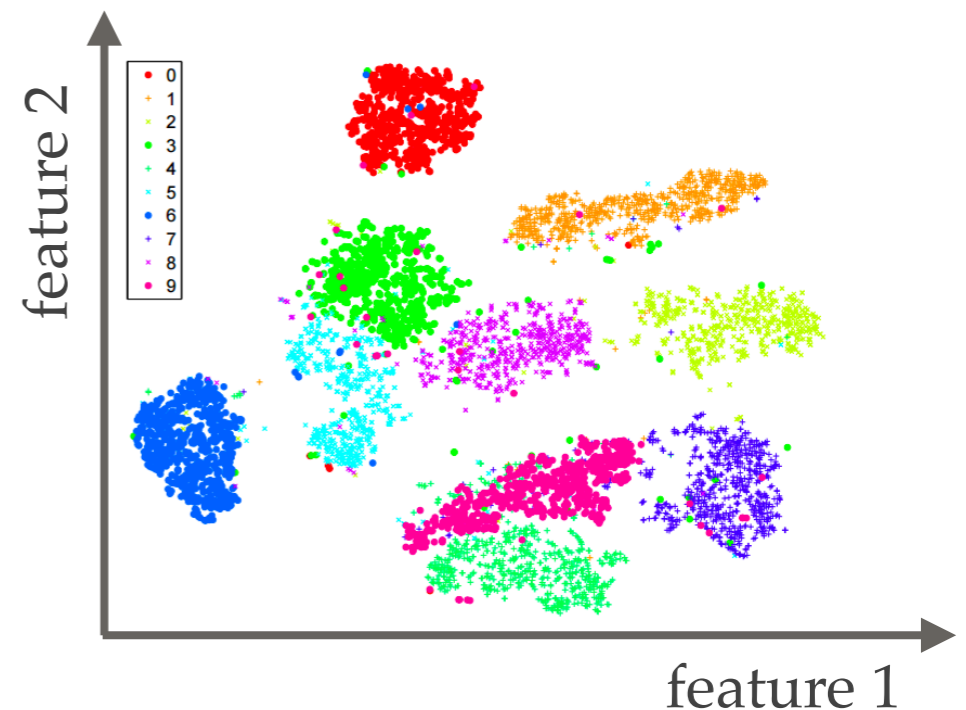
## 3. Using Unsupervised Learning algorithms:

- Some clustering algorithms (e.g., Hierarchical clustering, DBSCAN, OPTICS, GMMs) flag outliers.

**OPTICS**; sklearn gallery:



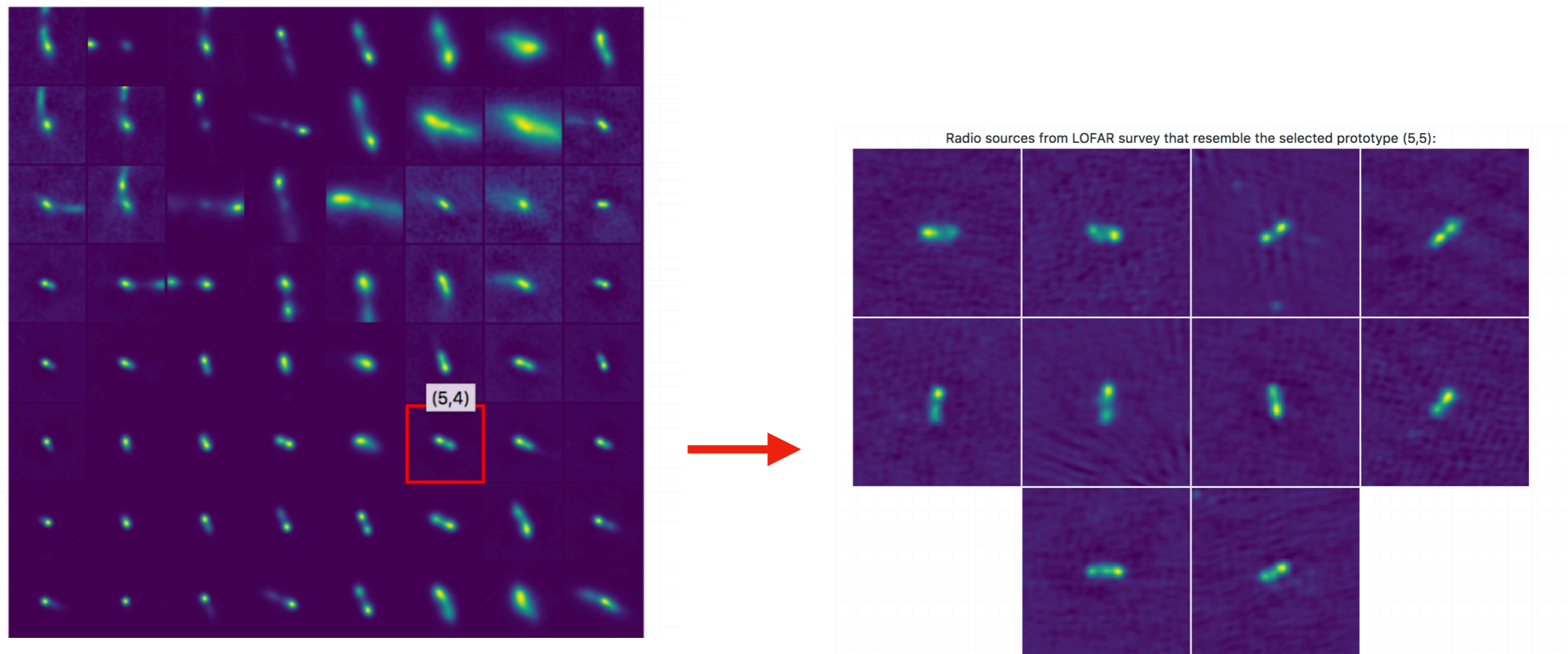
- Apply a dimensionality reduction algorithm and identify outliers in the low-dimensional representation.



# How can we detect outliers?

## 3. Using Unsupervised Learning algorithms:

- Using self-organizing maps: select outliers as objects that show a large distance to all prototypes, compared to the other objects.



Taken from J.  
Harwood's  
presentation

# How can we detect outliers?

---

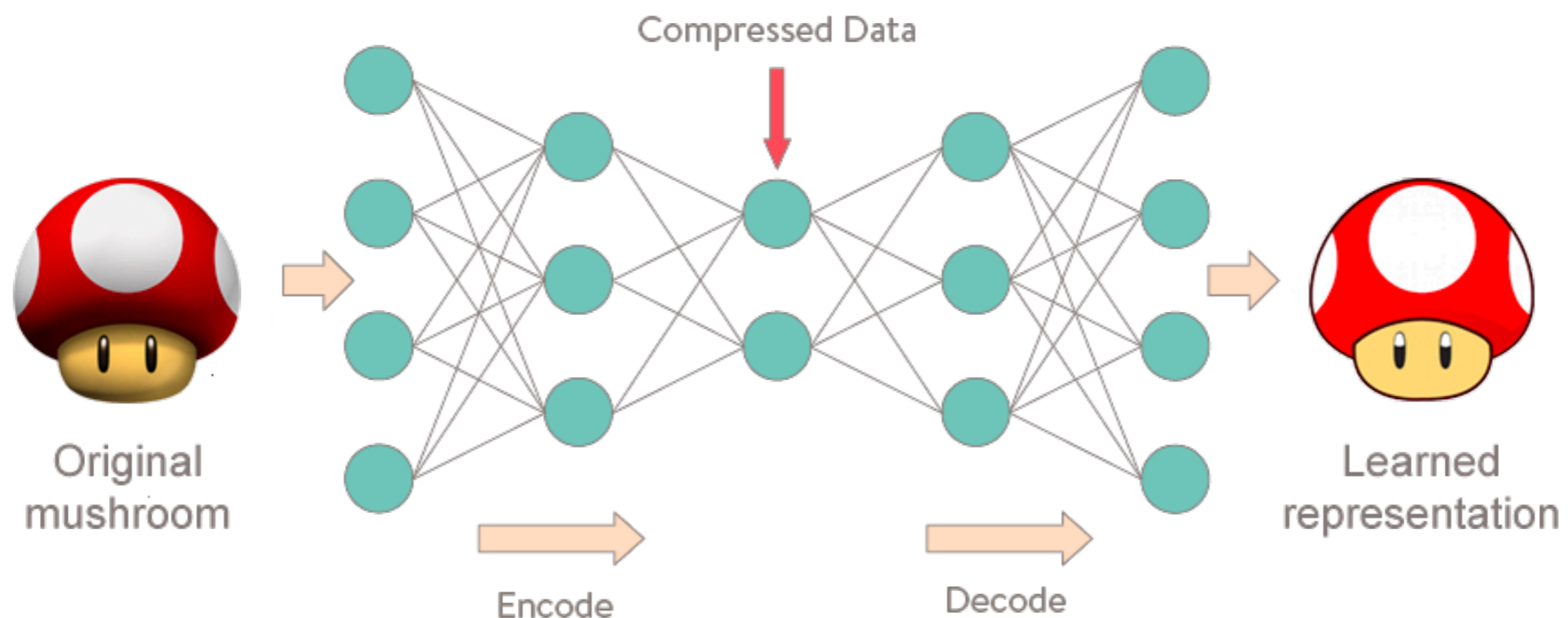
## 3. Using Unsupervised Learning algorithms:

- For methods that can reconstruct the object after the dimensionality reduction, compute the reconstruction error and select objects with a large reconstruction error (e.g., PCA, NMF, auto-encoders).

# How can we detect outliers?

## 3. Using Unsupervised Learning algorithms:

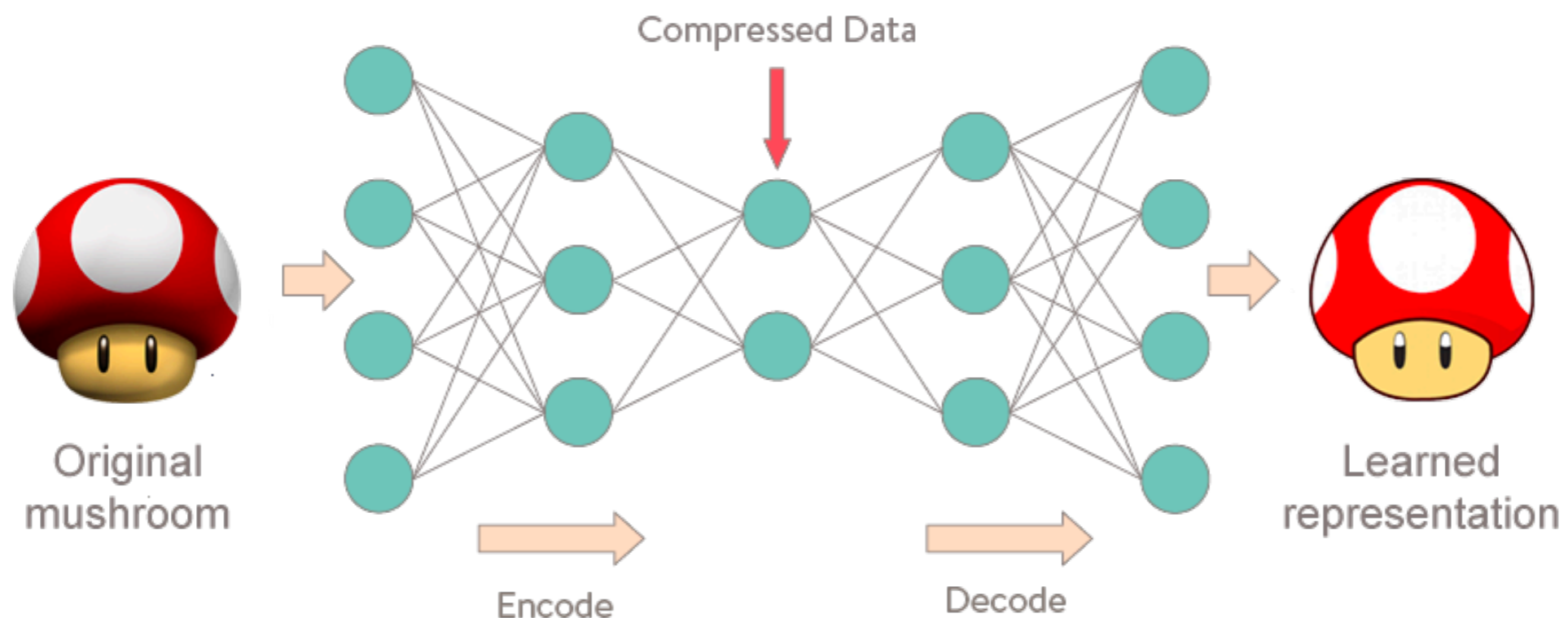
- For methods that can reconstruct the object after the dimensionality reduction, compute the reconstruction error and select objects with a large reconstruction error (e.g., PCA, NMF, auto-encoders).



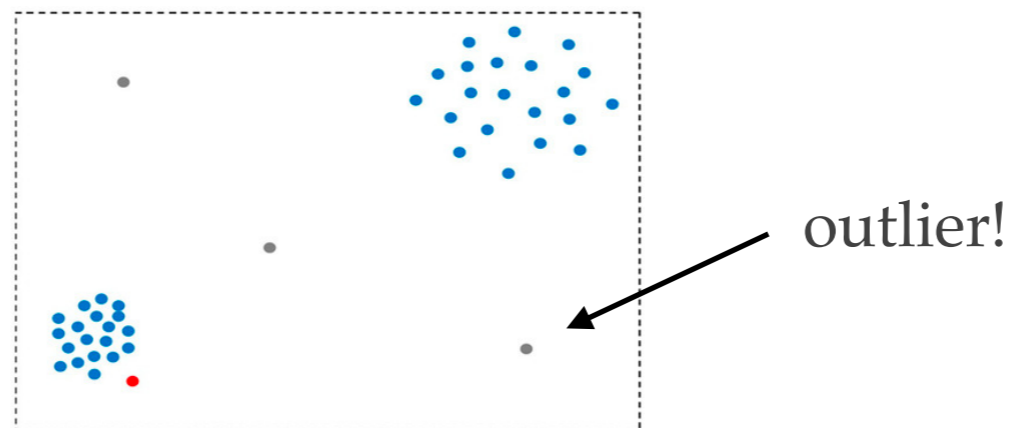
# How can we detect outliers?

## 3. Using Unsupervised Learning algorithms:

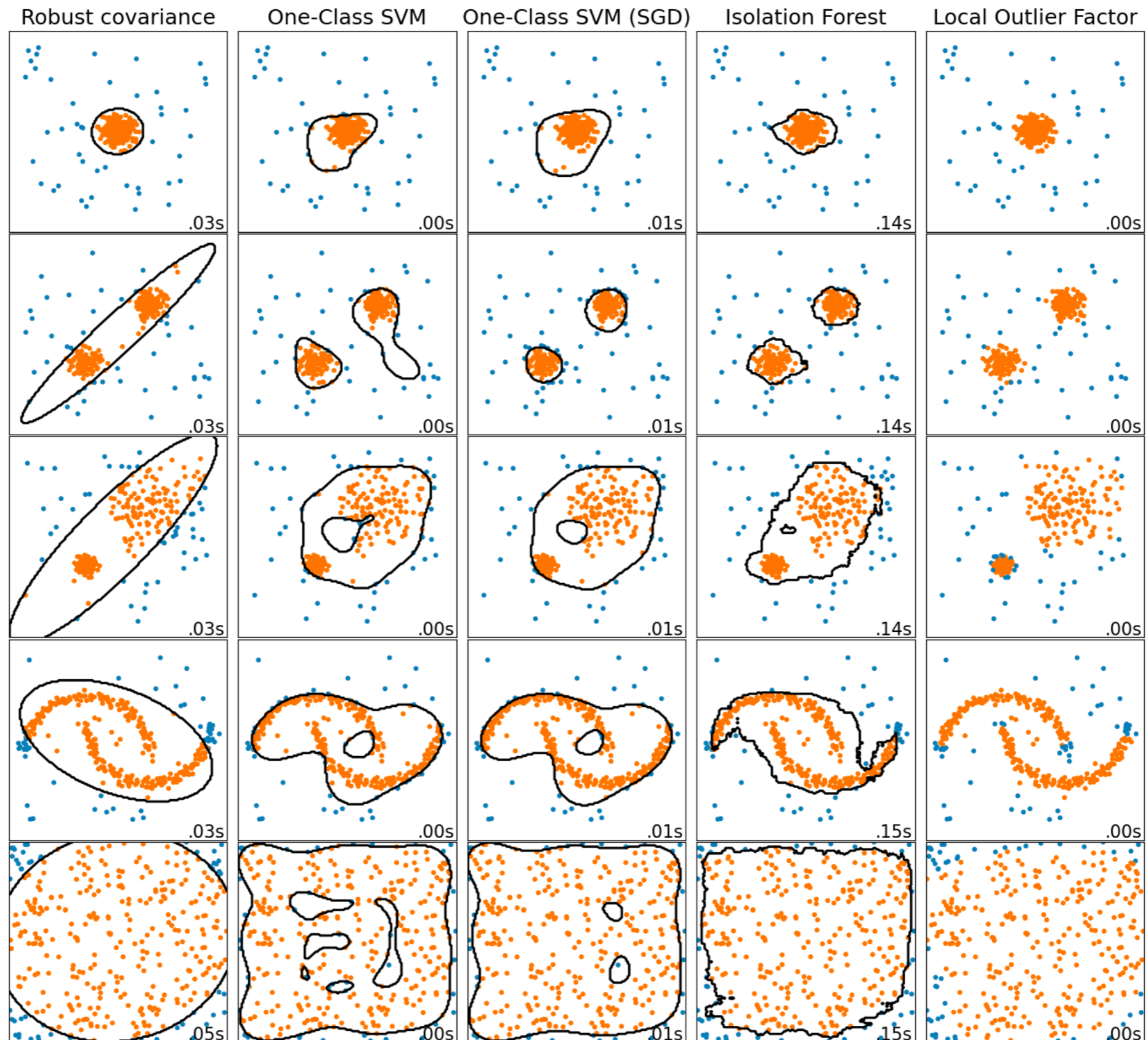
- For methods that can reconstruct the object after the dimensionality reduction, compute the reconstruction error and select objects with a large reconstruction error (e.g., PCA, NMF, auto-encoders).



Latent space / low dimensional embedding:



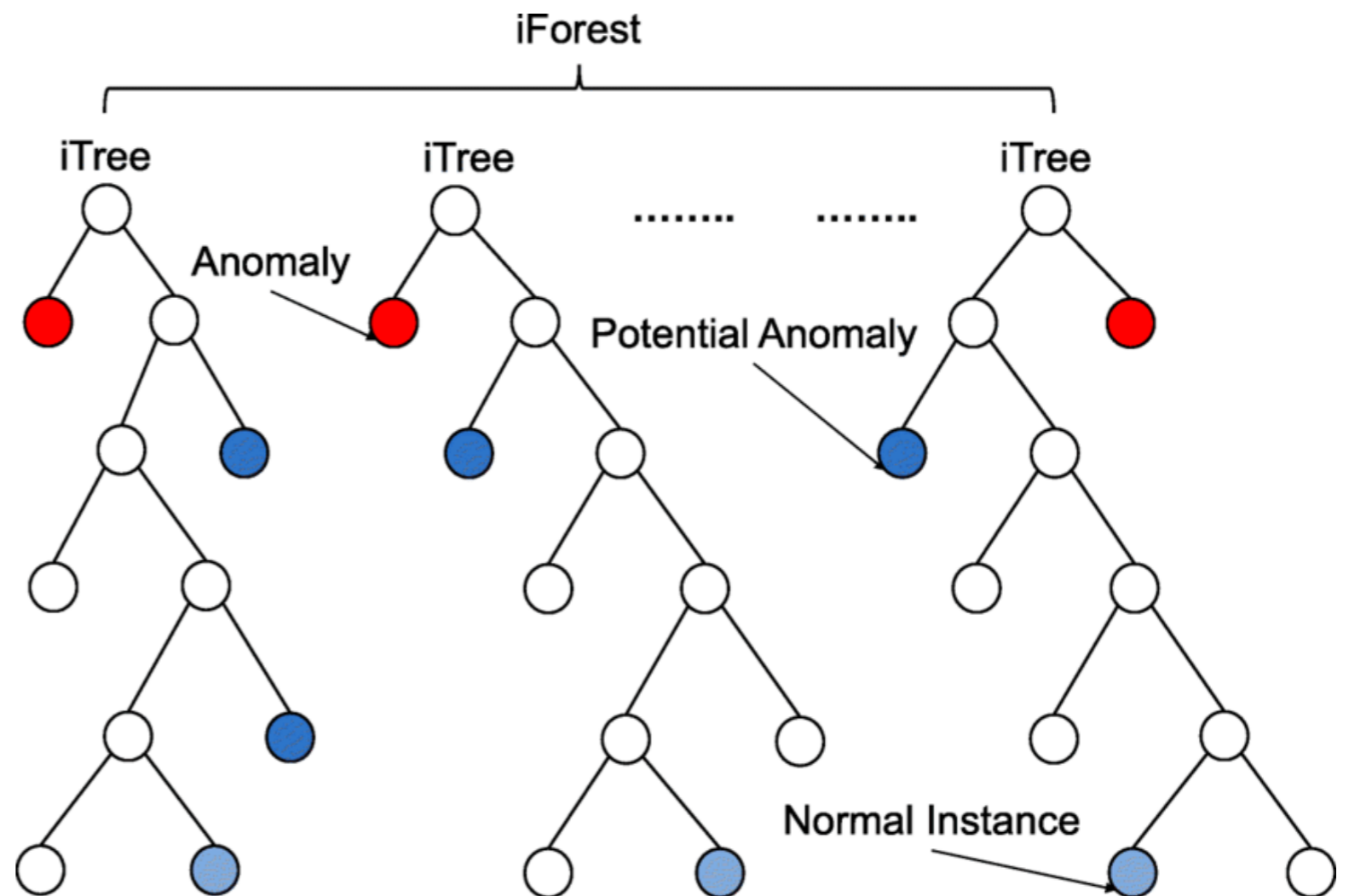
# Outlier Detection Algorithms



Taken from [here](#).

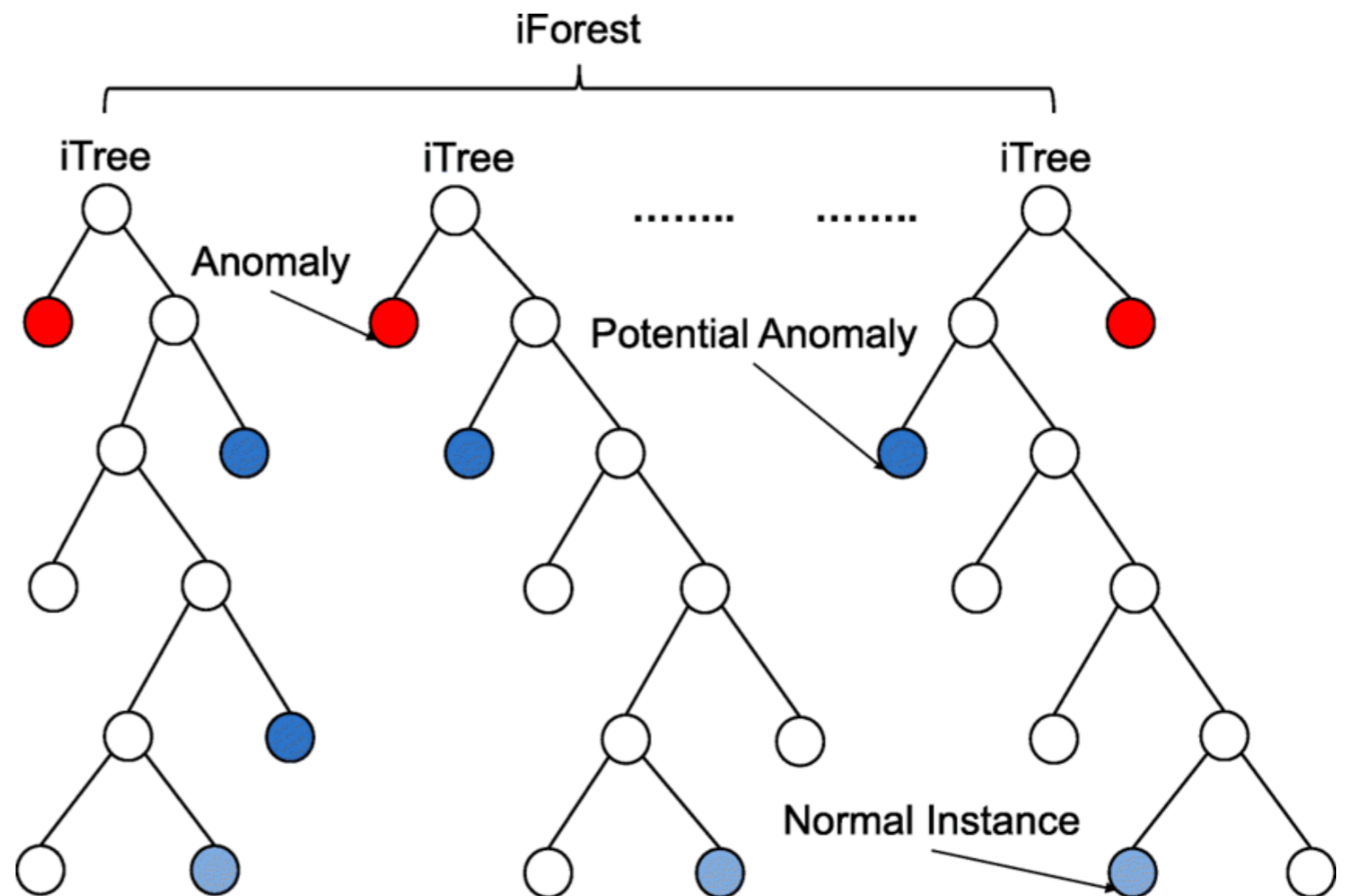
# Isolation Forests

- ❖ We build random trees by randomly-selecting a feature and then randomly-selecting a split value between the minimum and maximum of the feature value.
- ❖ Outliers will tend to separate from the other objects earlier in the process.



# Isolation Forests

- ❖ We build random trees by randomly-selecting a feature and then randomly-selecting a split value between the minimum and maximum of the feature value.
- ❖ Outliers will tend to separate from the other objects earlier in the process.



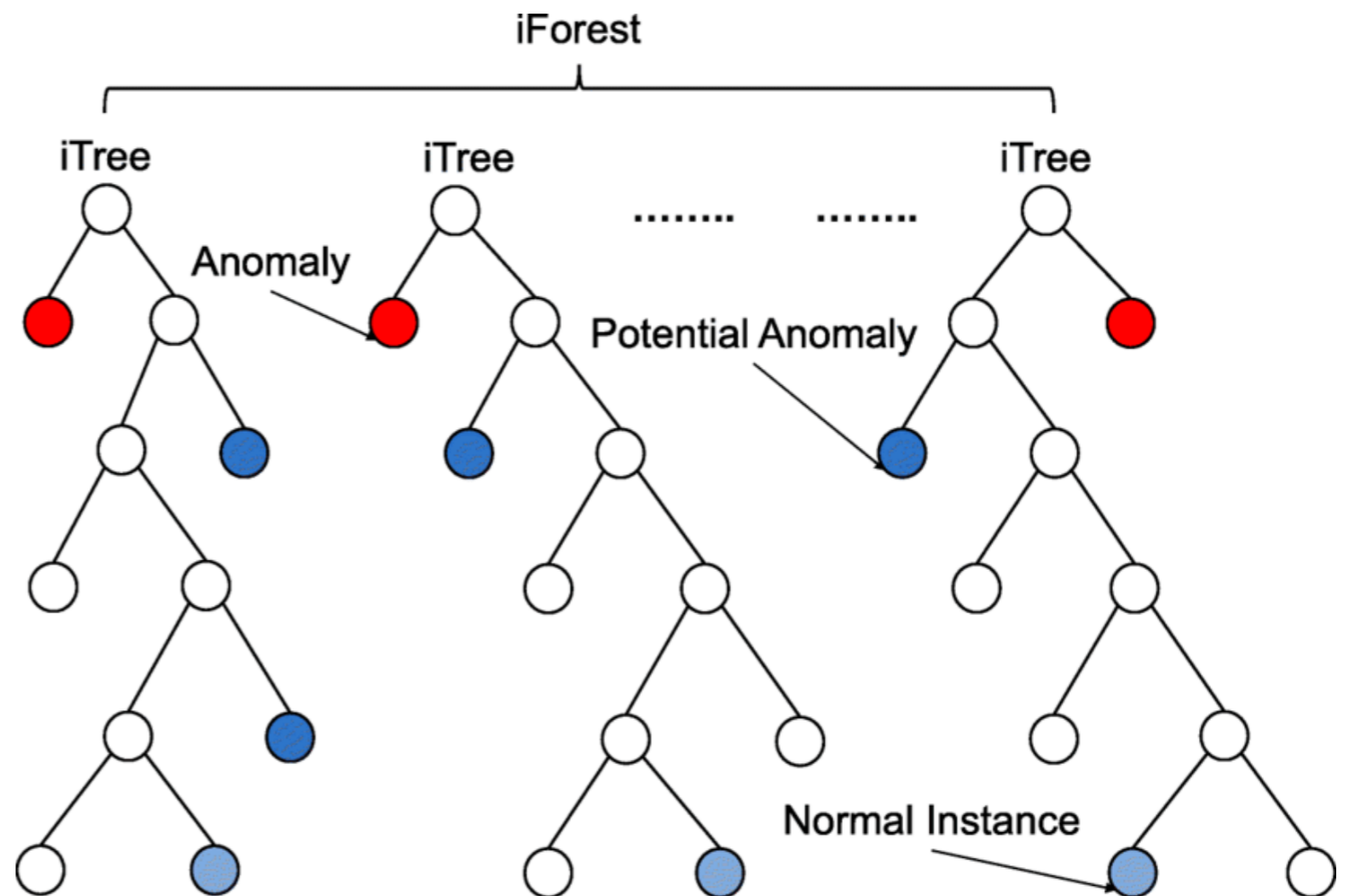
- ❖ For a set of random trees, the outlier score is proportional to the average path length within the different trees.
- ❖ The user provides the number of trees in the forest and the fraction of objects to be marked as outliers.

# Isolation Forests

- ❖ We build random trees by randomly-selecting a feature and then randomly-selecting a split value between the minimum and maximum of the feature value.
- ❖ Outliers will tend to separate from the other objects earlier in the process.

[Sklearn implementation](#)

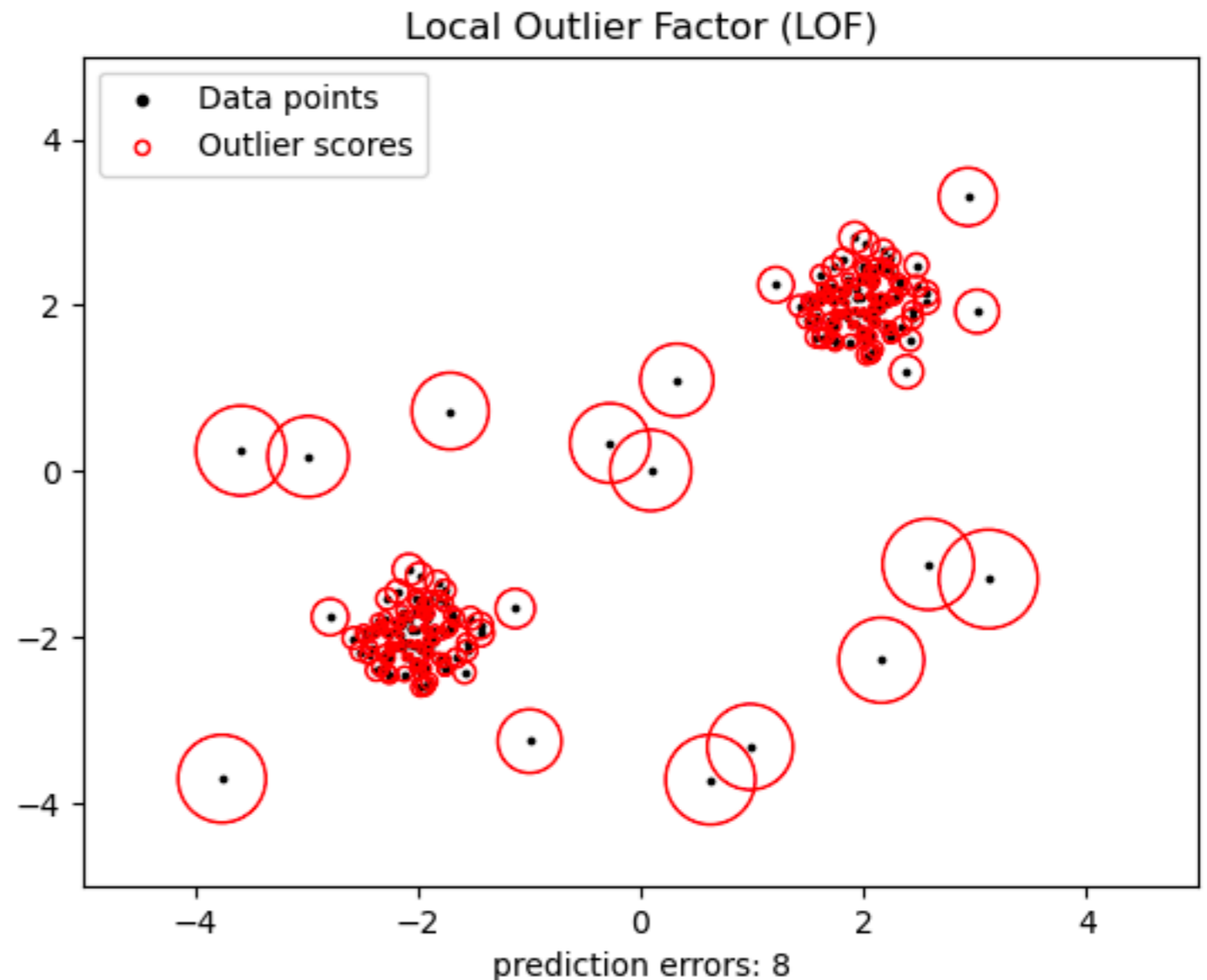
- ❖ For a set of random trees, the outlier score is proportional to the average path length within the different trees.
- ❖ The user provides the number of trees in the forest and the fraction of objects to be marked as outliers.



# Local Outlier Factor

- ❖ The algorithm computes the *local density deviation* of a given point with respect to its *neighbors*.
- ❖ The local density around an object is estimated using its distance to its  $k$  nearest neighbors.

Outliers are objects that have substantially lower local density than their neighbors.

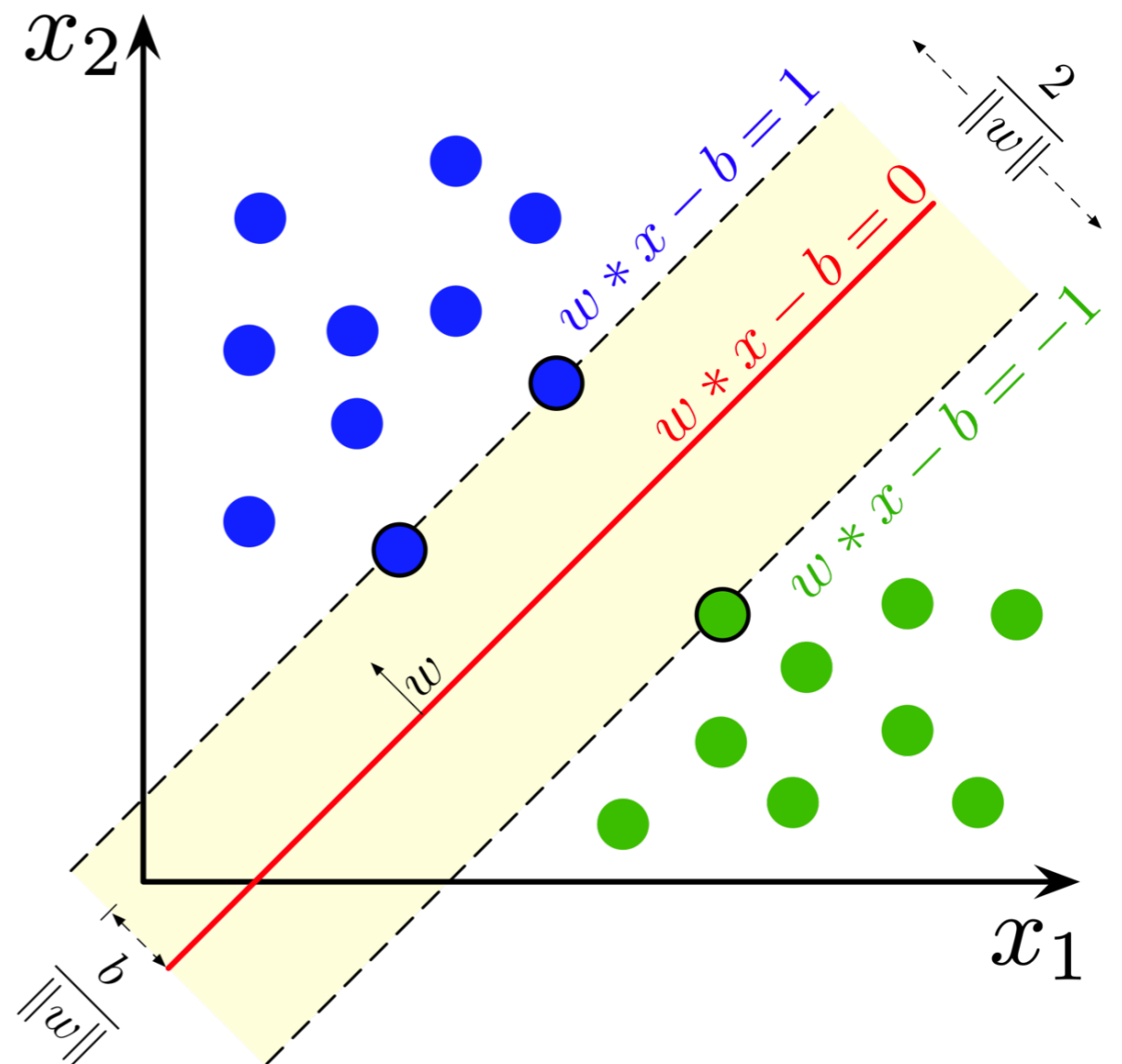
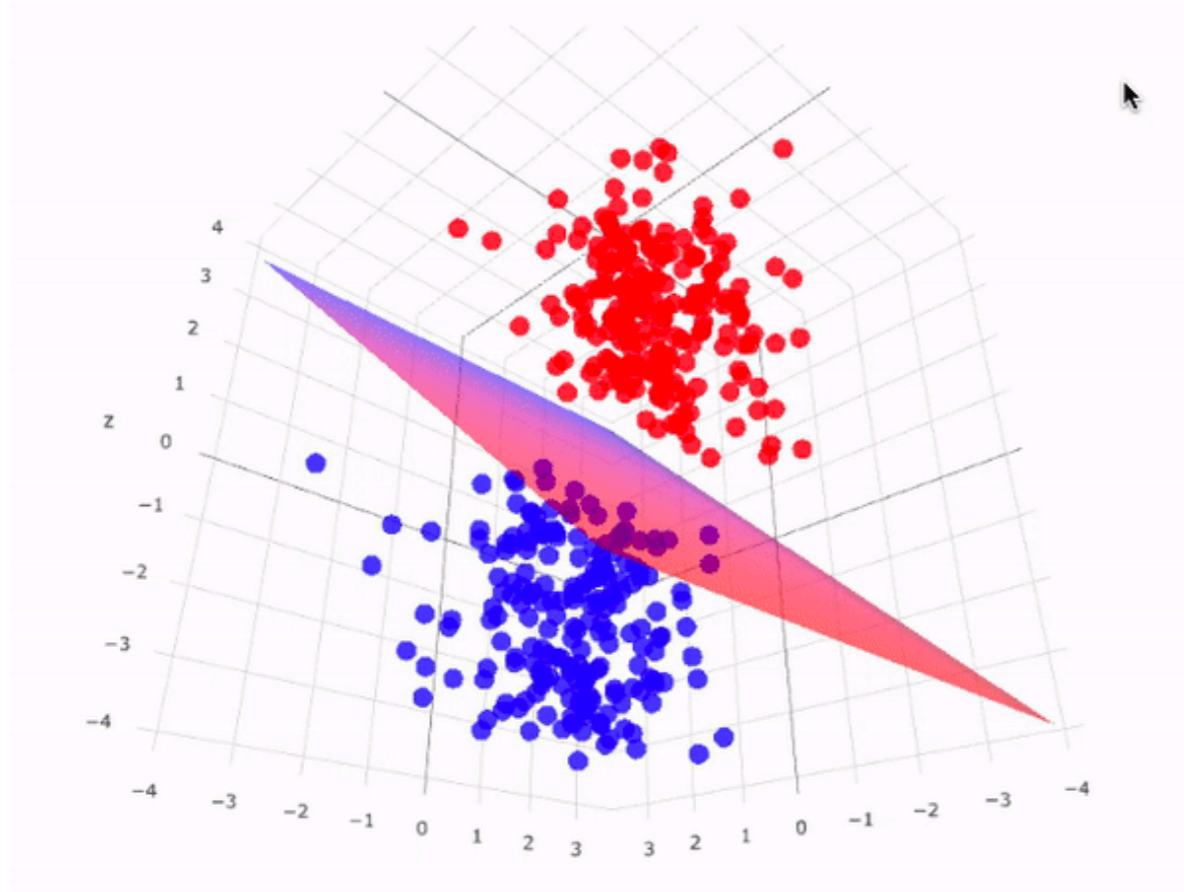


# One-Class SVM

SVM: Supervised Machine Learning algorithm used for **classification**. The constructed model is a **multidimensional hyper-plane** that best separates the two classes.

For example, our dataset has two features and two classes. SVM will find the one-dimensional hyper-plane that provides the best separation of the two classes.

Training stage: given a labeled dataset, find the best multidimensional hyperplane.

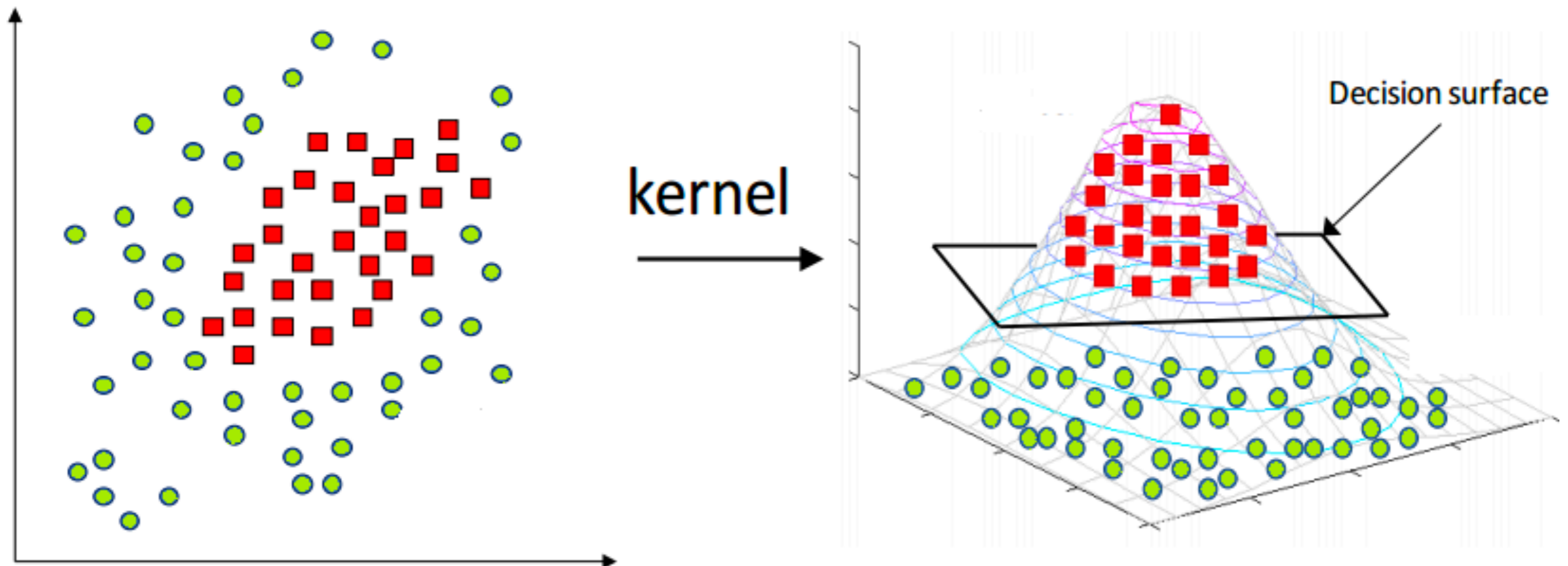


# One-Class SVM

For datasets which are not-linearly separable, we apply the **Kernel Trick**.

We map the input dataset onto a higher dimension, using a predefined kernel, where the data is linearly-separable.

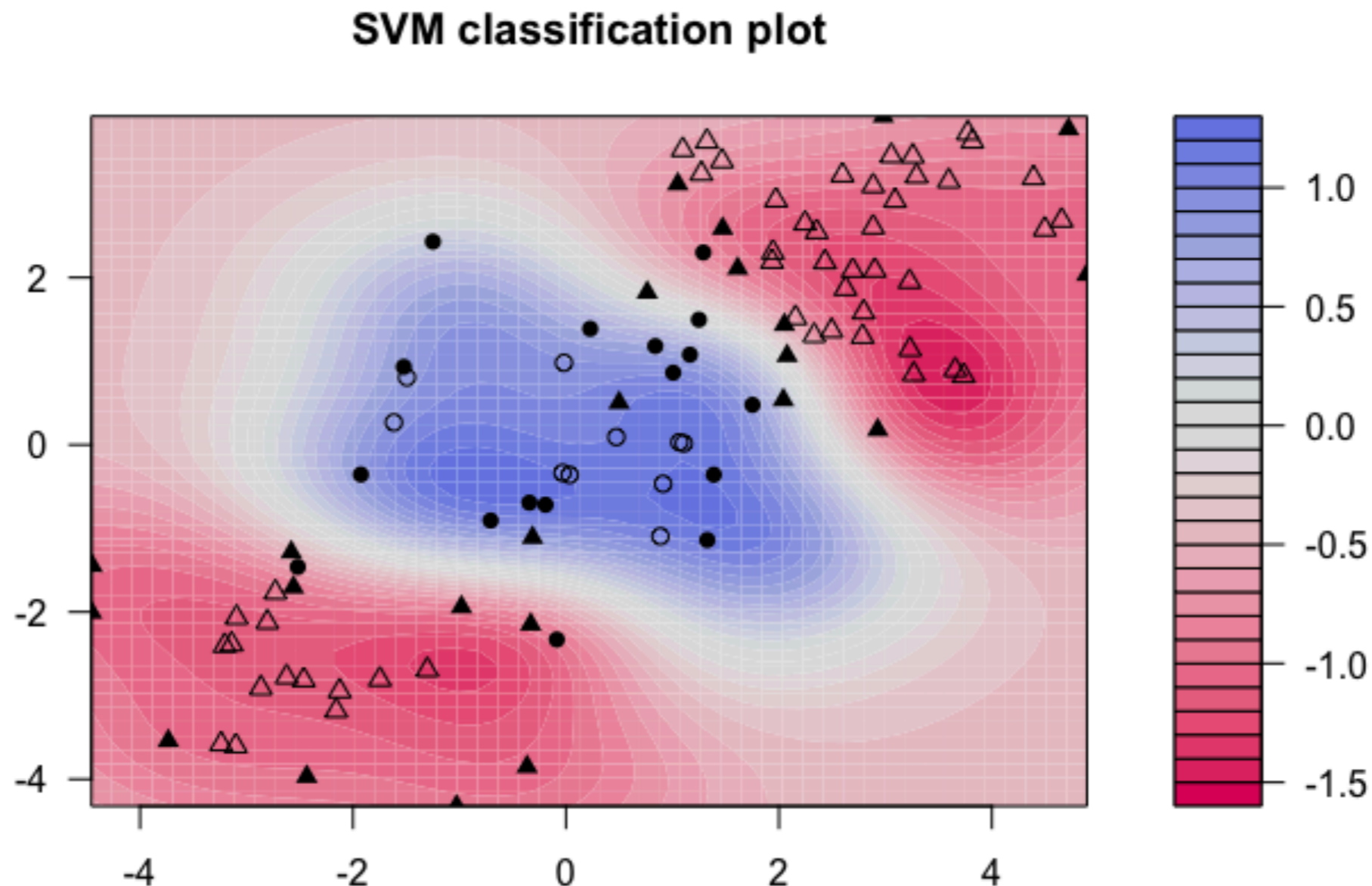
The kernel shape is a **hyper-parameter** of the algorithm, along with parameters that control its shape!



# One-Class SVM

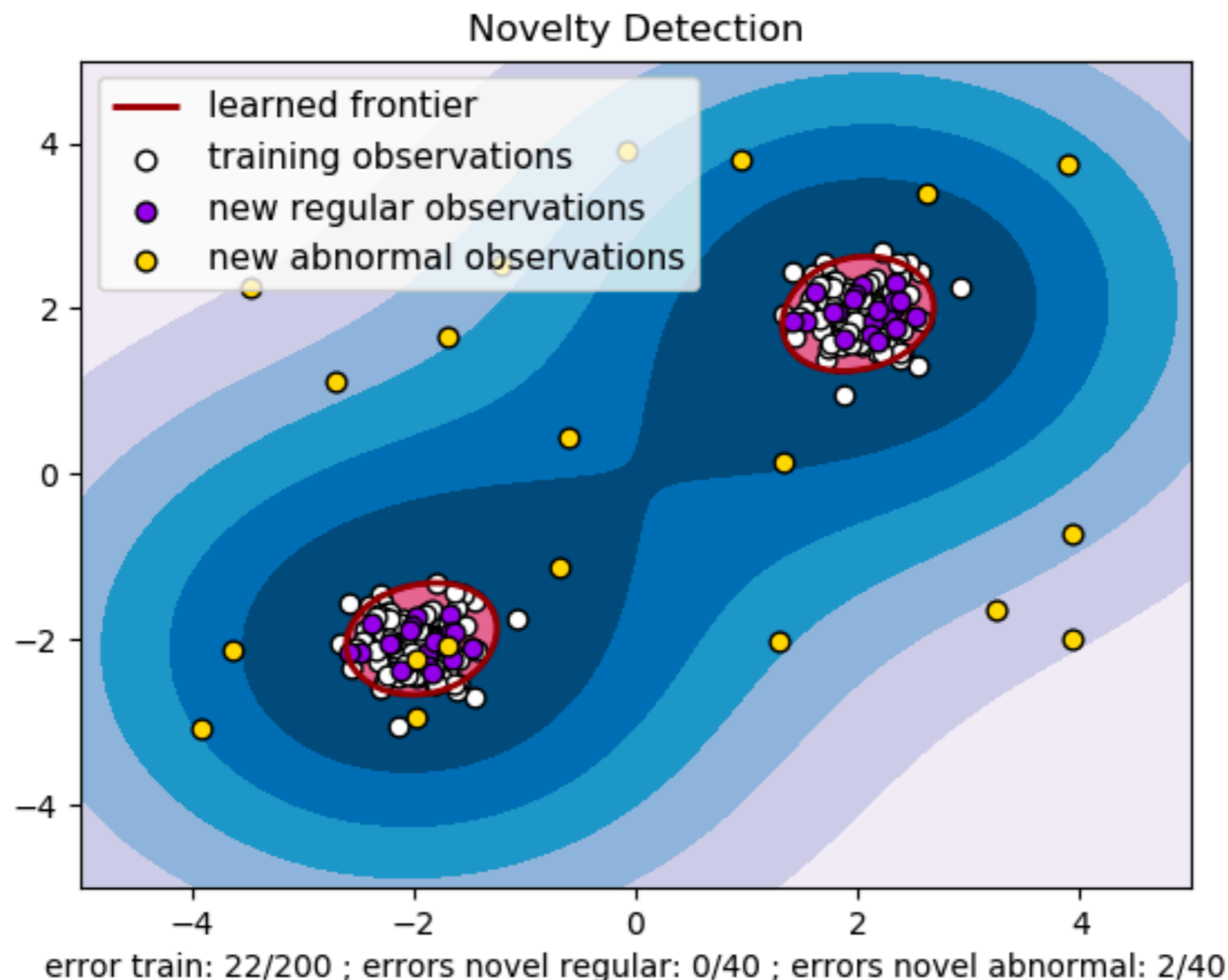
Prediction on previously unseen datasets: use the decision boundary to classify the objects according to their features.

The distance of an object from a decision boundary can be used as a classification uncertainty.



# One-Class SVM

In One-class SVM, instead of using a hyperplane to separate between two classes, we use a hypersphere to encompass all the objects. The outliers will be defined as objects that are outside of this hypersphere.



# Outlier Detection: aspects to consider

---

The best practice is application-specific:

- ❖ **Goal:** identify and remove outliers from our dataset
  - Use feature cuts.
  - Dimensionality reduction with the same metric as the one assumed later in your model.

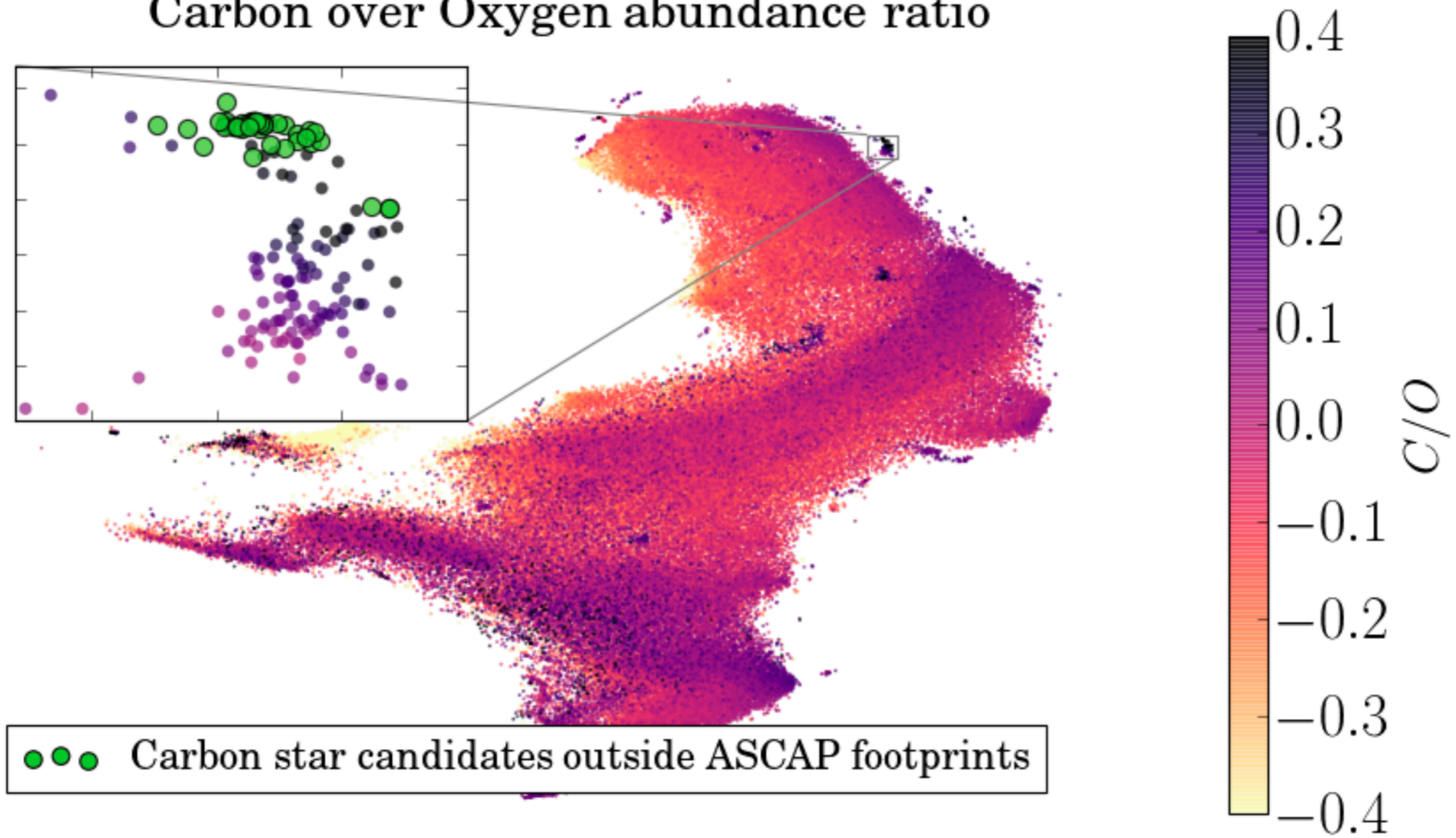
# Outlier Detection: aspects to consider

---

The best practice is application-specific:

- ❖ **Goal:** identify and remove outliers from our dataset
  - Use feature cuts.
  - Dimensionality reduction with the same metric as the one assumed later in your model.
- ❖ **Goal:** identify new and interesting objects in the dataset
  - Stay as close as possible to the raw dataset.
  - Euclidean distance / variance-based methods may limit the type of outliers we will find.
  - Use several algorithms and several metrics.

## Carbon over Oxygen abundance ratio



tSNE map of APOGEE stars, Reis et al. (2018)