Understanding and Implementing Novelty and Outlier Detection

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Understanding outliers and novelties

Novelty and outlier detection uses

Algorithms for outlier and novelty detection

Local Outlier Factor

Elliptic Envelope

Isolation Forest

Outlier and Novelty

Outlier

A data point that differs significantly from other data points in the same data set.

Outlier

A data point that differs significantly from other data points in the same data set.

Novelty

A data point encountered in prediction that differs significantly from any data points encountered during training.

Novelty

A data point **encountered in prediction** that differs significantly from any data points encountered during training.

Novelty

A data point encountered in prediction that differs significantly from any data points encountered during training.

Outliers and Novelties

Outliers	Novelties
Anomalous points in training dataset	Anomalous points in test dataset
Unsupervised	Semi-supervised
Outliers, by definition, will never form a dense cluster	Novelties could possibly form a dense cluster

Outlier Detection

Fit regions in the dataset where data points are the most concentrated, deviant observations are outliers.

Novelty Detection

Training data not polluted by outliers, try to detect whether new observations are deviant.

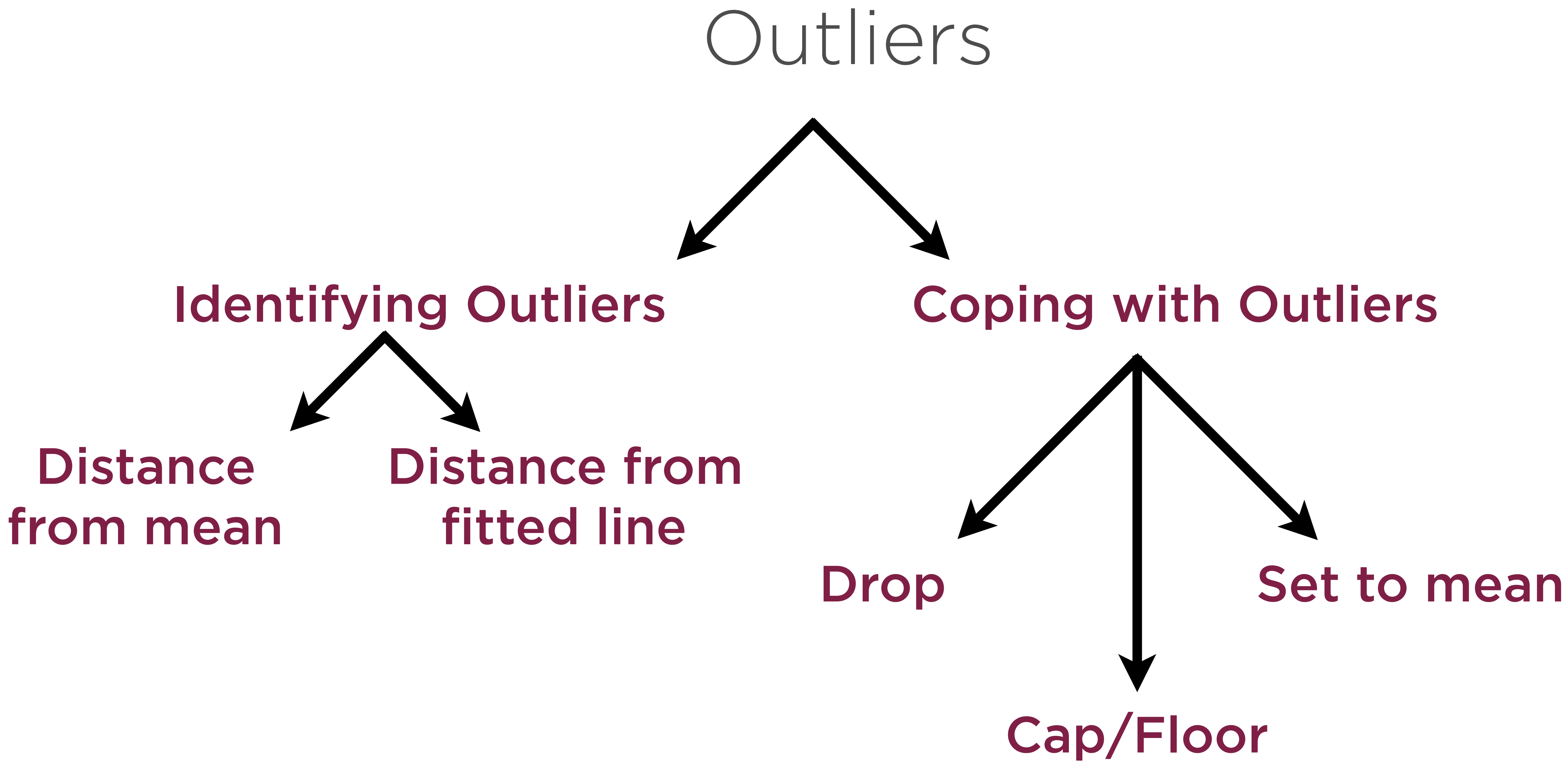
Uses of Outlier and Novelty Detection

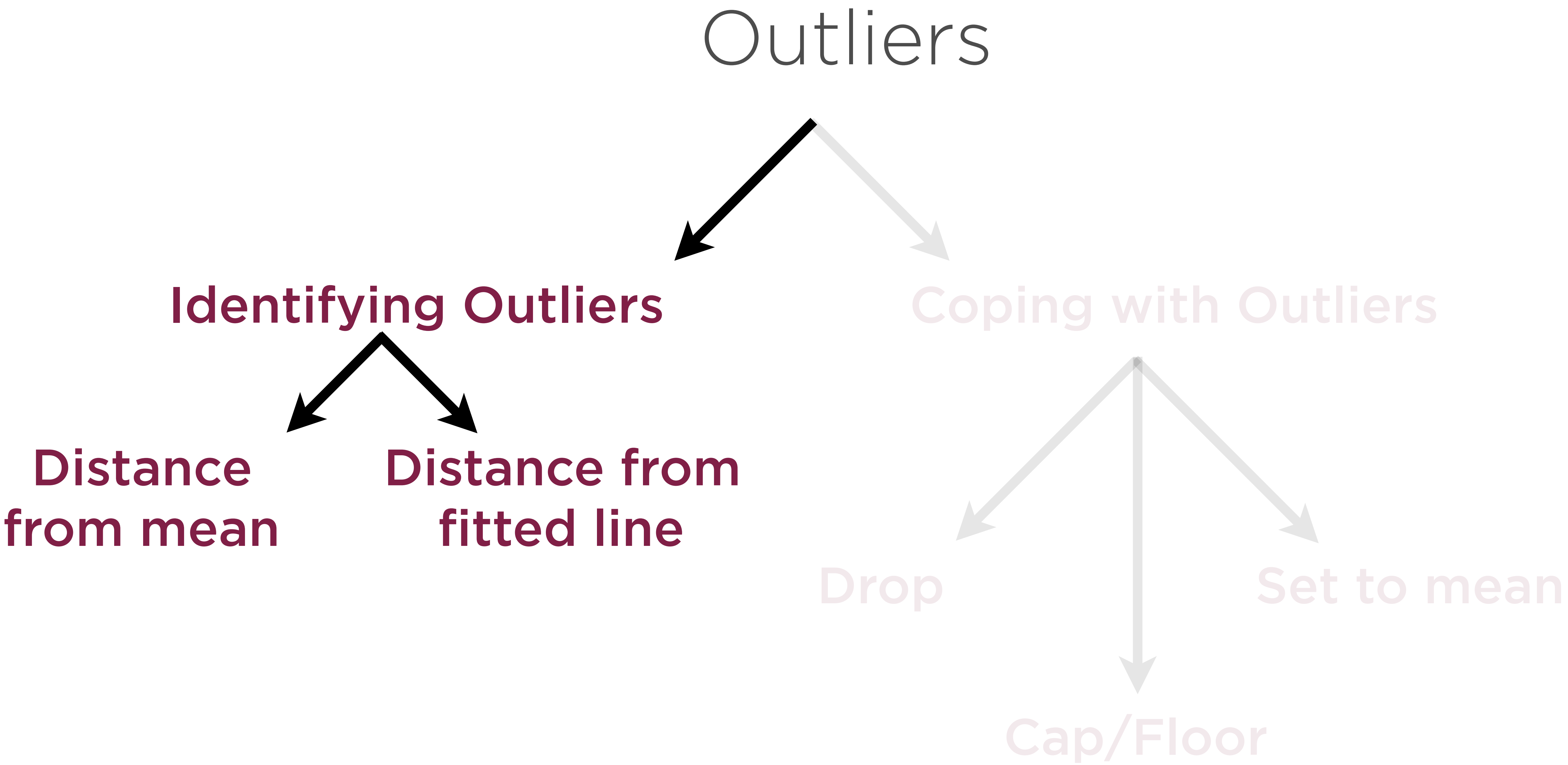
Detecting anomalous data i.e. fraudulent credit card transactions

Detecting errors in data collection or processing

Cleaning and preparing data for ML models

Outlier and Novelty Detection


Outliers
Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Cap/Floor
Set to mean


Outliers
Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Set to mean
Cap/Floor

Identifying Outliers

Distance from mean

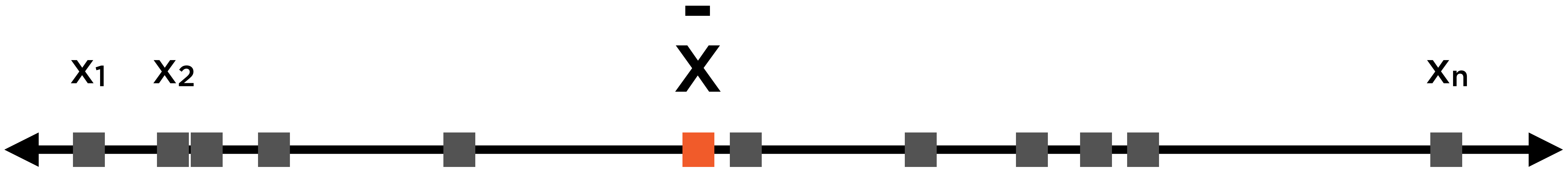
Distance from fitted line

Identifying Outliers

Distance from mean

Distance from fitted line

Mean and Variance



Mean and variance succinctly summarize a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Variance and Standard Deviation

Standard deviation is the square root of variance

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Outliers

Points that lie more than 3 standard deviations from the mean are often considered outliers

Outliers

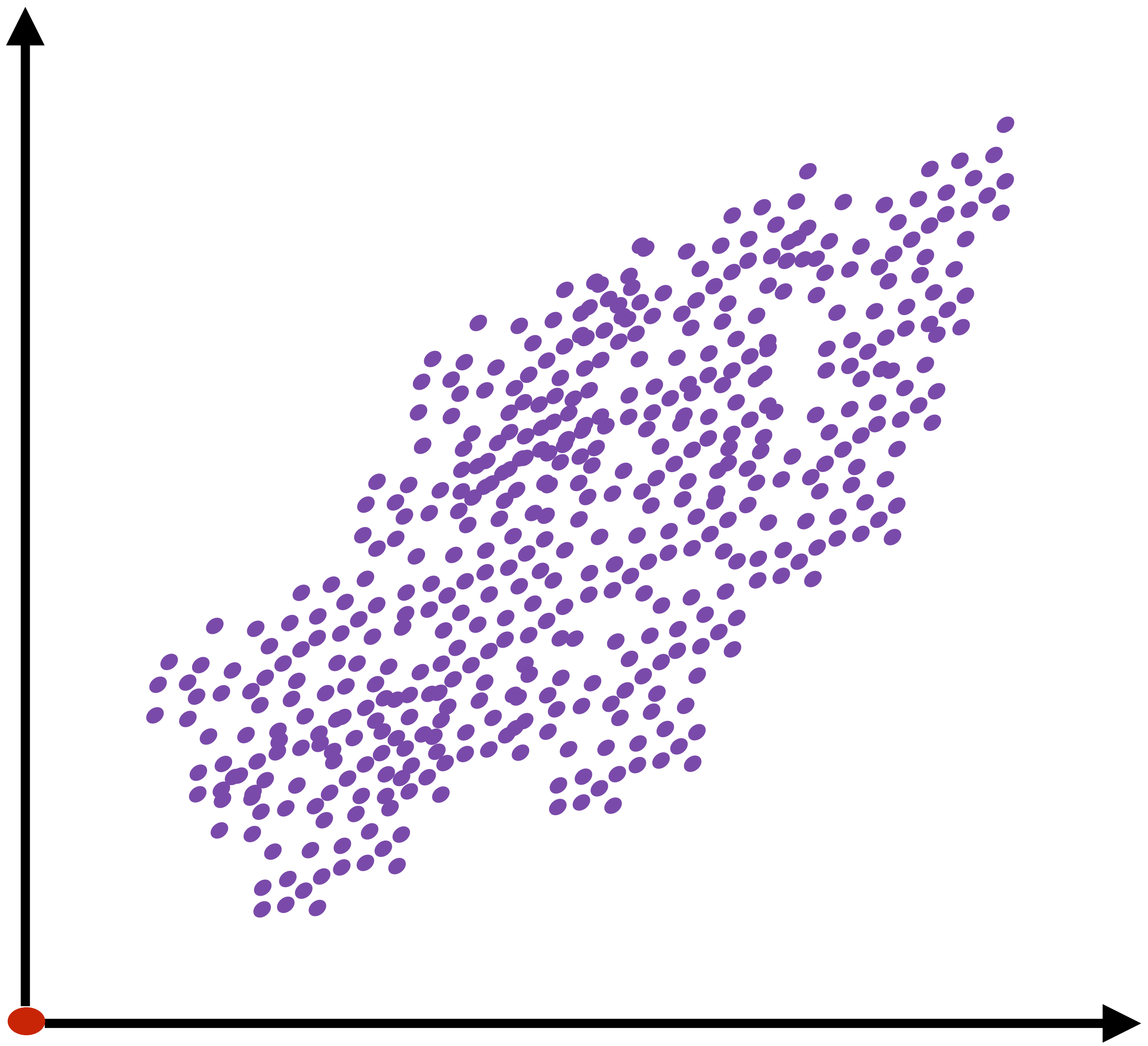
Points that lie more than 3 standard deviations from the mean are often considered outliers

Identifying Outliers

Distance from mean

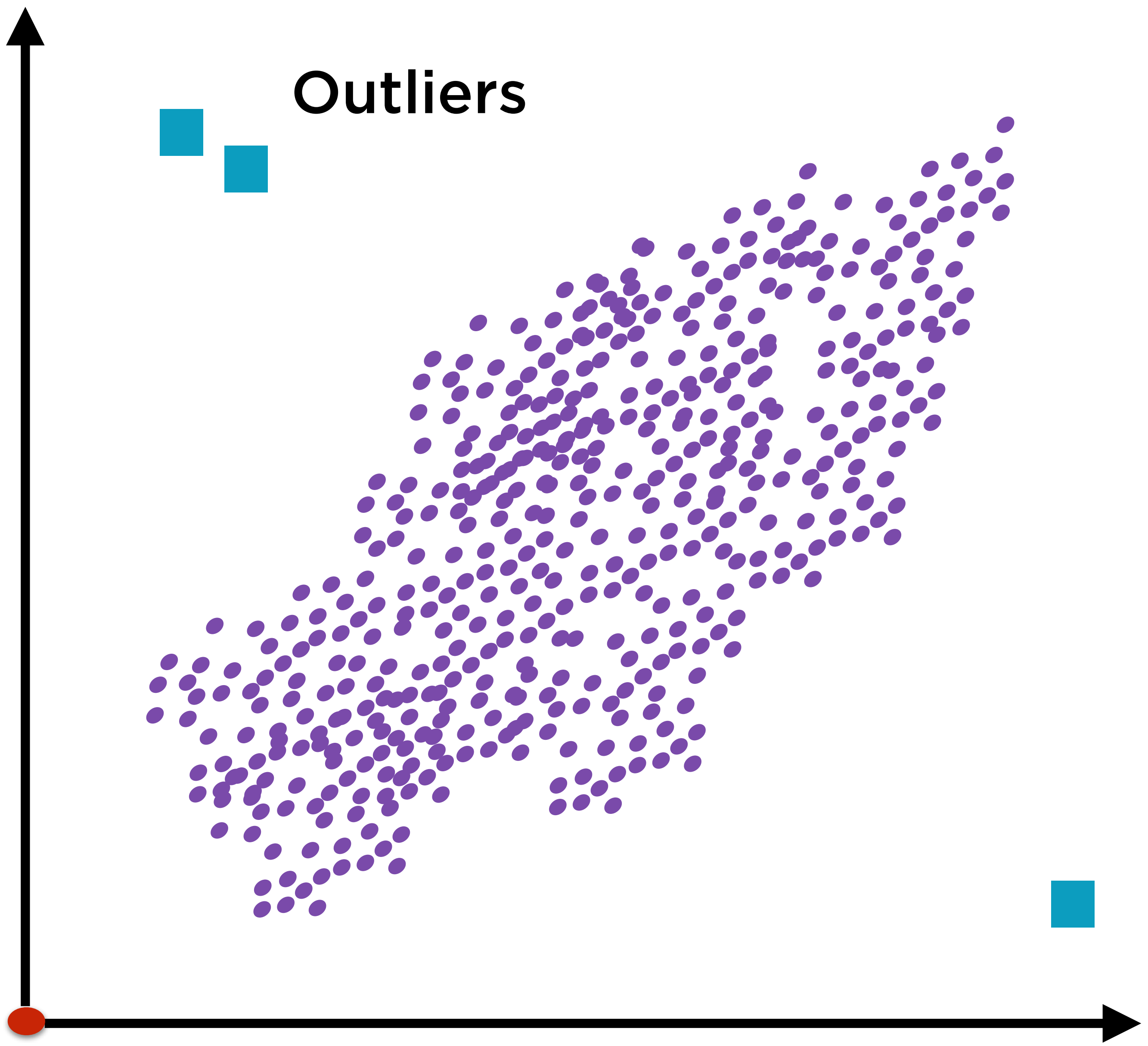
Distance from fitted line

Outliers



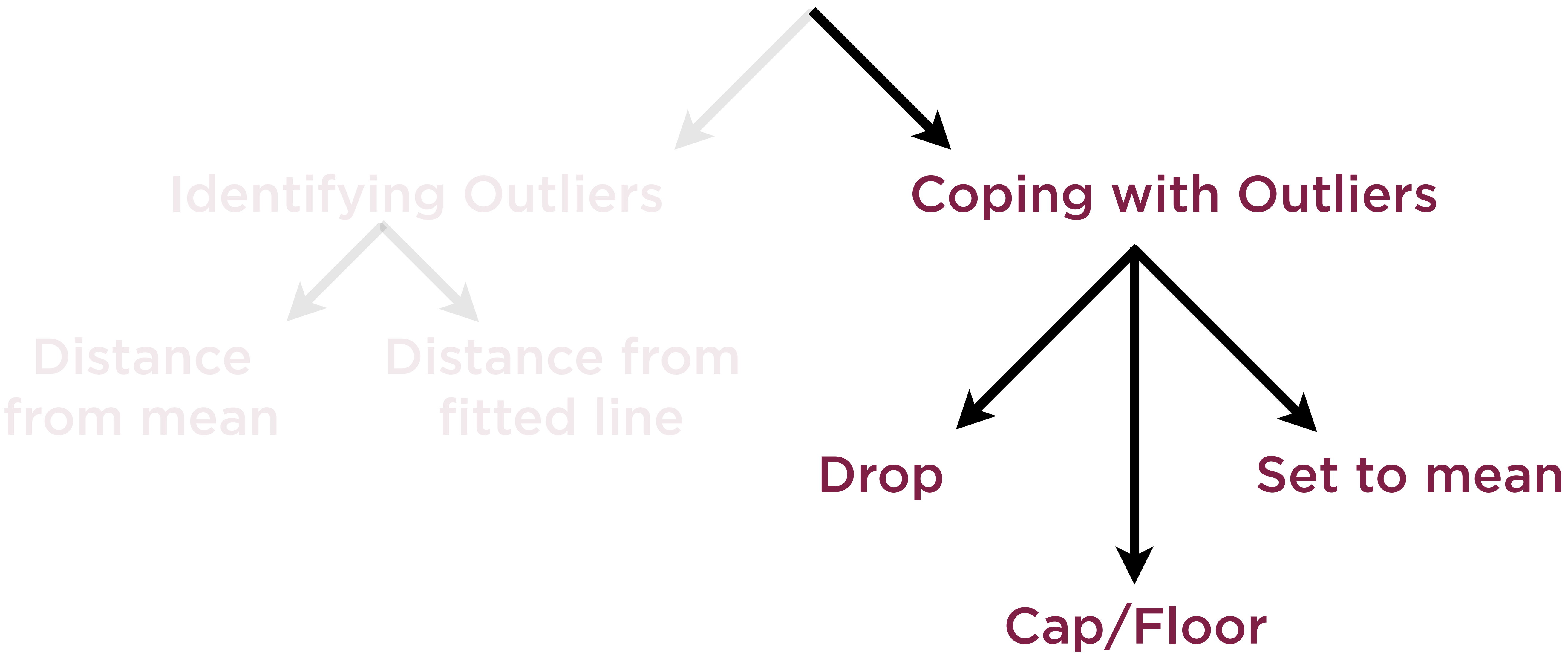
Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers



Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers


Identifying Outliers
Distance from mean
Distance from fitted line
Coping with Outliers
Drop
Cap/Floor
Set to mean

Coping with Outliers

Always start by scrutinizing outliers

If erroneous observation

- Drop if all attributes of that point are erroneous
- Set to mean if only one attribute is erroneous

Coping with Outliers

If genuine, legitimate outlier

- Leave as-is if model not distorted
- Cap/Floor if model is distorted
    - Need to first standardize data
    - Cap positive outliers to +3
    - Floor negative outliers to -3

scikit-learn algorithms can be used for both outlier as well as novelty detection

Outlier and Novelty Detection Algorithms in scikit-learn

Local Outlier Factor	**Elliptic Envelope**	**Isolation Forest**

Local Outlier Factor

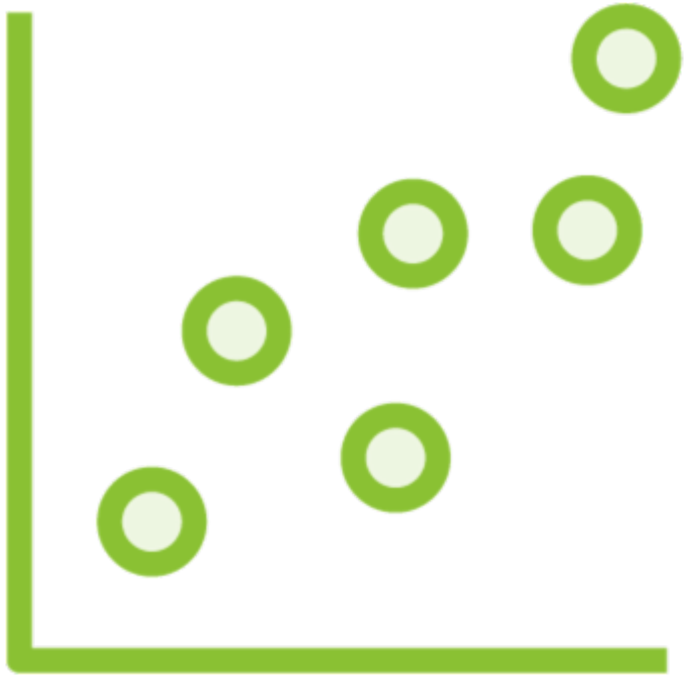
Outlier and Novelty Detection Algorithms in scikit-learn

Local Outlier Factor

Elliptic Envelope

Isolation Forest

Local Outlier Factor

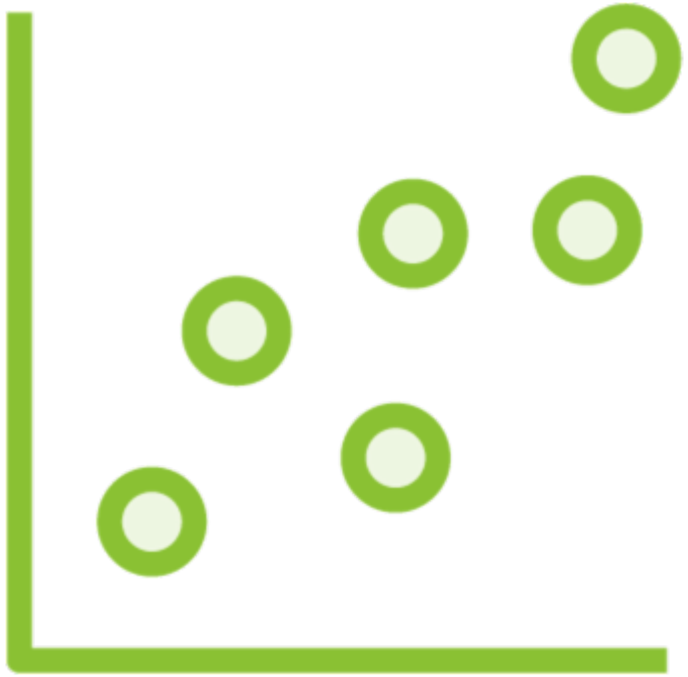


For each point, compute a score called the Local Outlier Factor (LOF) score

Flag as outlier if

- Point is far from its nearest neighbors
- Those neighbors are close to each other

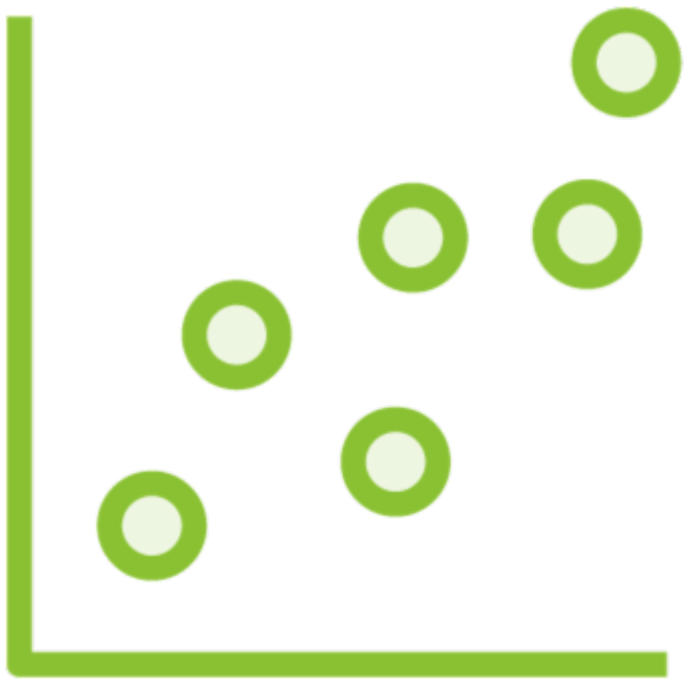
Local Outlier Factor



Use K-nearest neighbors algorithm to find neighbors

Number of neighbors to be considered is a parameter

Local Outlier Factor



Calculate the average density of neighbors

- How close the neighbors are to each other, on average

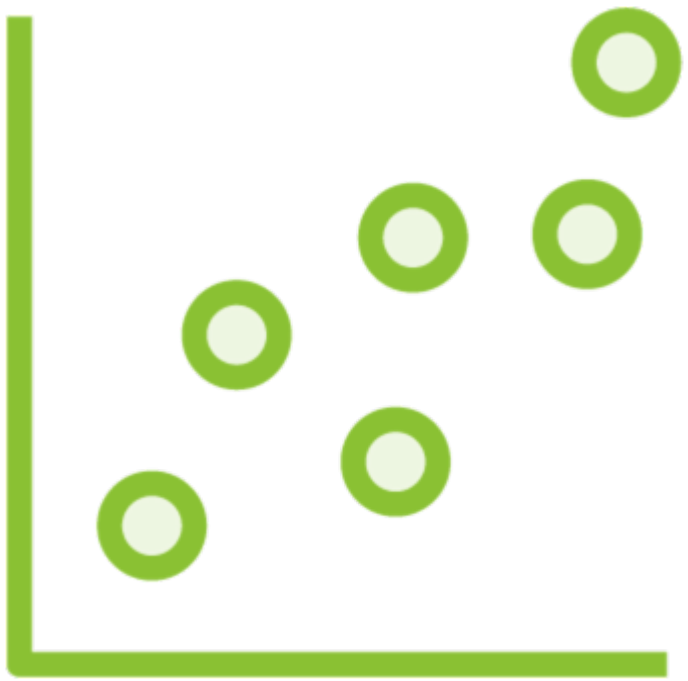
Calculate the average density of candidate point

- How close the point is to neighbors

Compare the two

Determines **how isolated** a particular sample is with respect to its **surrounding neighborhood**

Local Outlier Factor

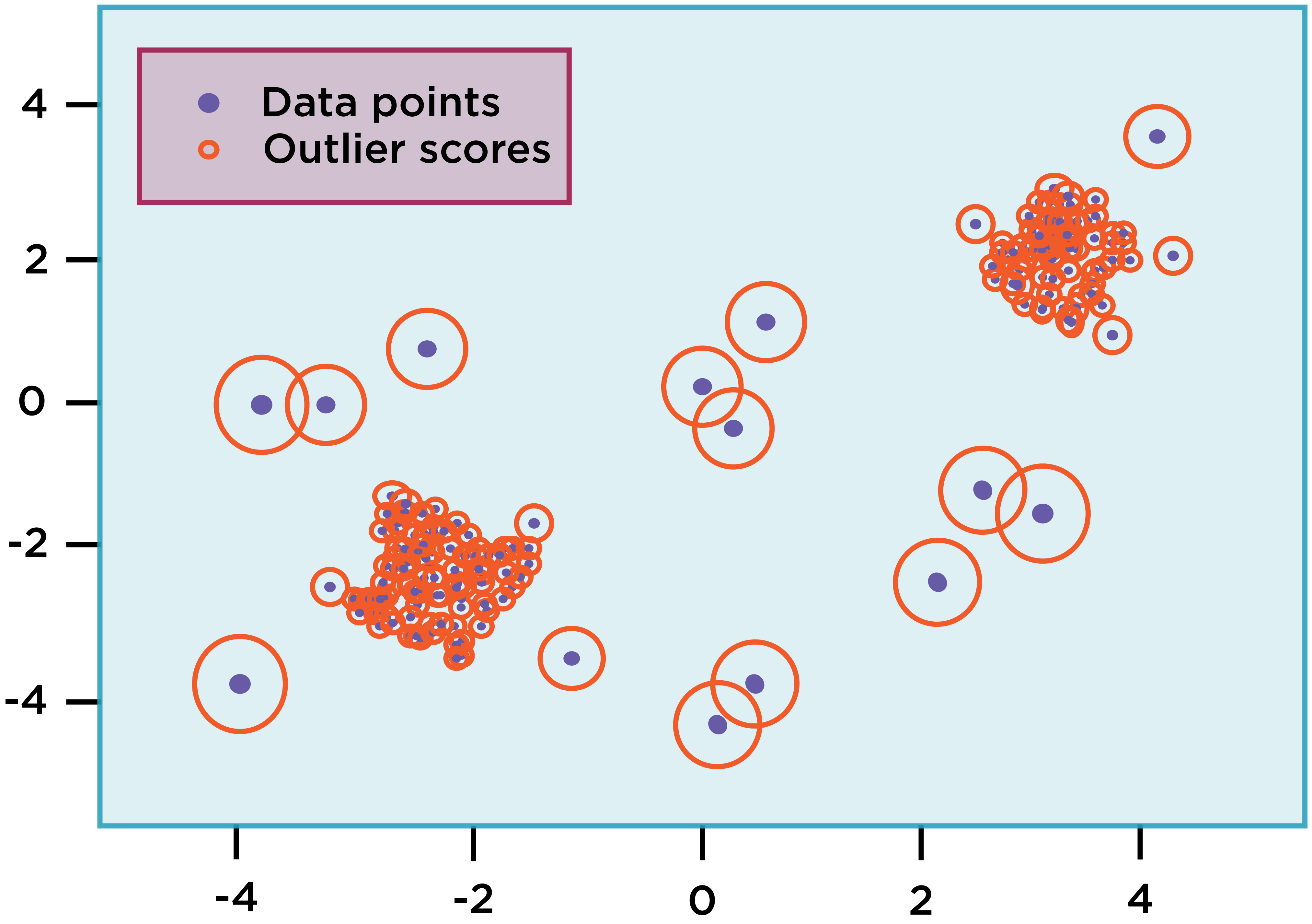


Works well with moderately high dimensionality data

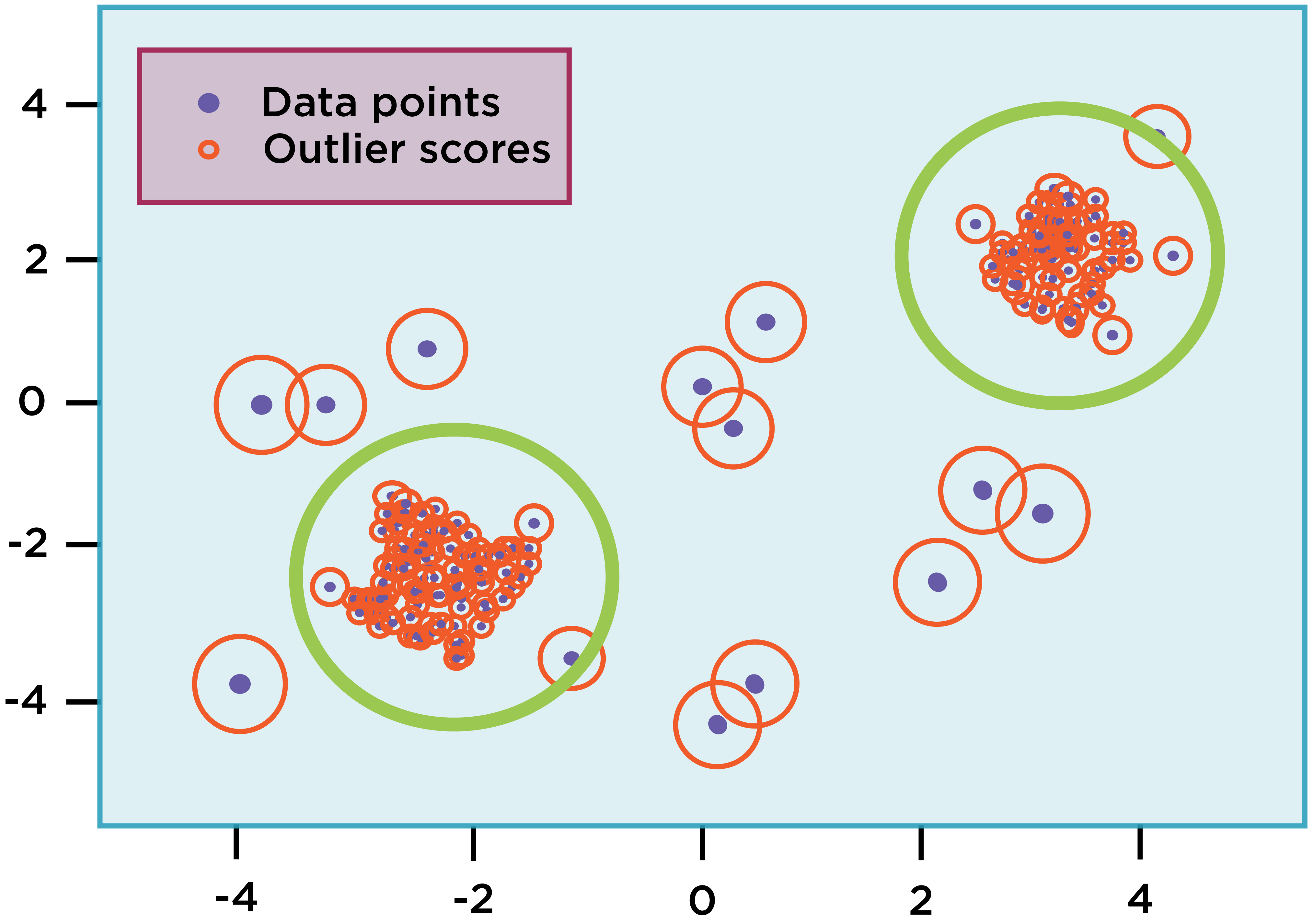
Considers both local and global properties

- Due to use of K-nearest-neighbors

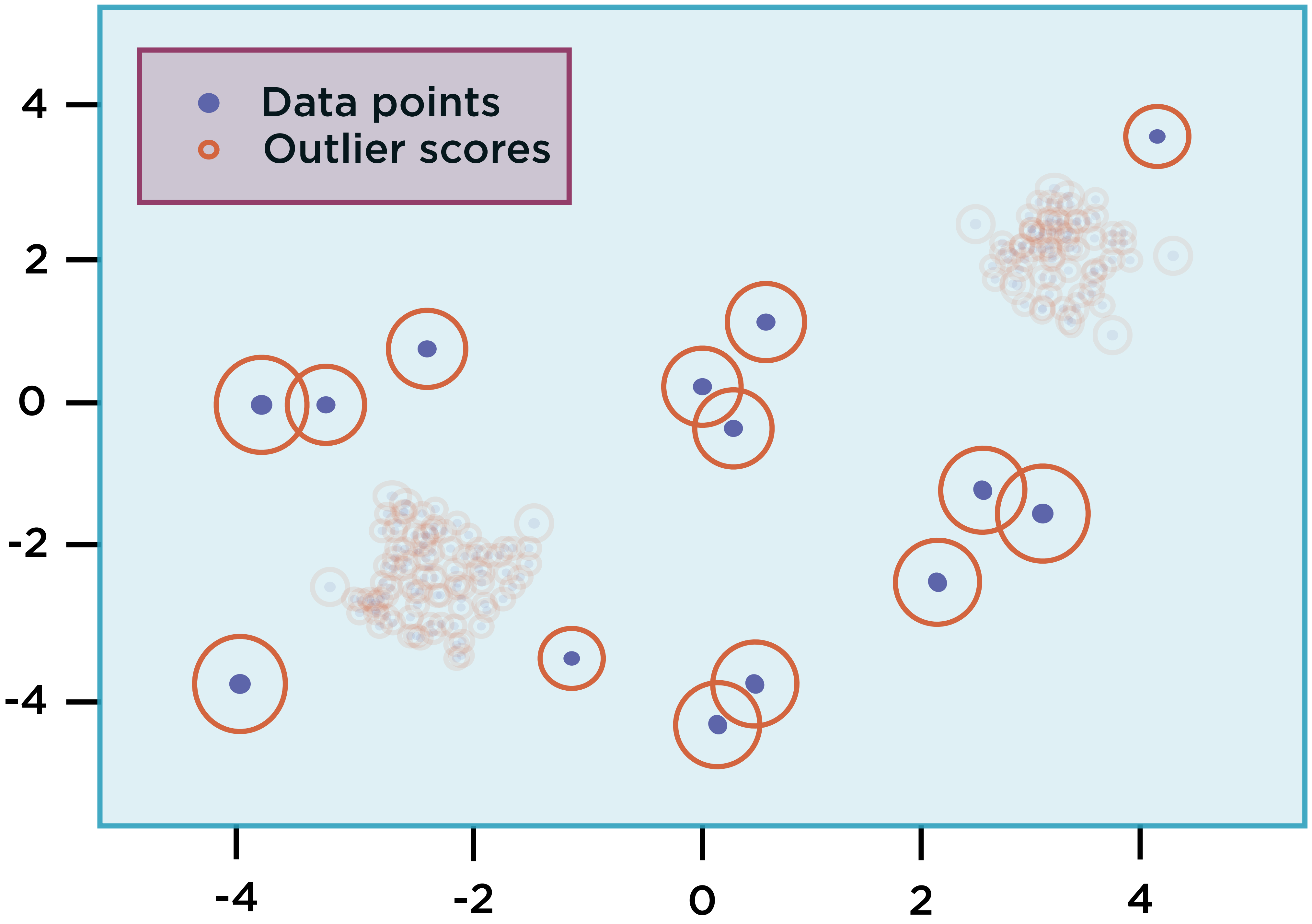
Outlier Detection with Local Outlier Factor



Outlier Detection with Local Outlier Factor



Outlier Detection with Local Outlier Factor



`sklearn.neighbors.LocalOutlierFactor`

estimator cannot be used directly for novelty detection, need to set novelty=True

Elliptic Envelope

Outlier and Novelty Detection Algorithms in scikit-learn

Local Outlier Factor

Elliptic Envelope

Isolation Forest

Elliptic Envelope

Assumes data is drawn from a normal i.e. Gaussian distribution

Draw an elliptical envelope through the central data points

All points outside the ellipse are considered outliers

Elliptic Envelope

Elliptic envelope is drawn using a Robust Covariance estimate

Assumes data is drawn from a known distribution e.g. Gaussian Normal

Robust Covariance

Covariance matrix simply summarizes pair-wise covariance of vectors

If greater values of one variable correspond to greater values in another

Or vice versa

Covariance is positive

Robust Covariance

Trivial to compute, but can be fragile

- Time-series data with illiquid stocks

Usually calculated using maximum likelihood estimate

Very sensitive to outliers in the dataset

Robust Covariance

Use complex but robust procedure called Minimum Covariance Determinant

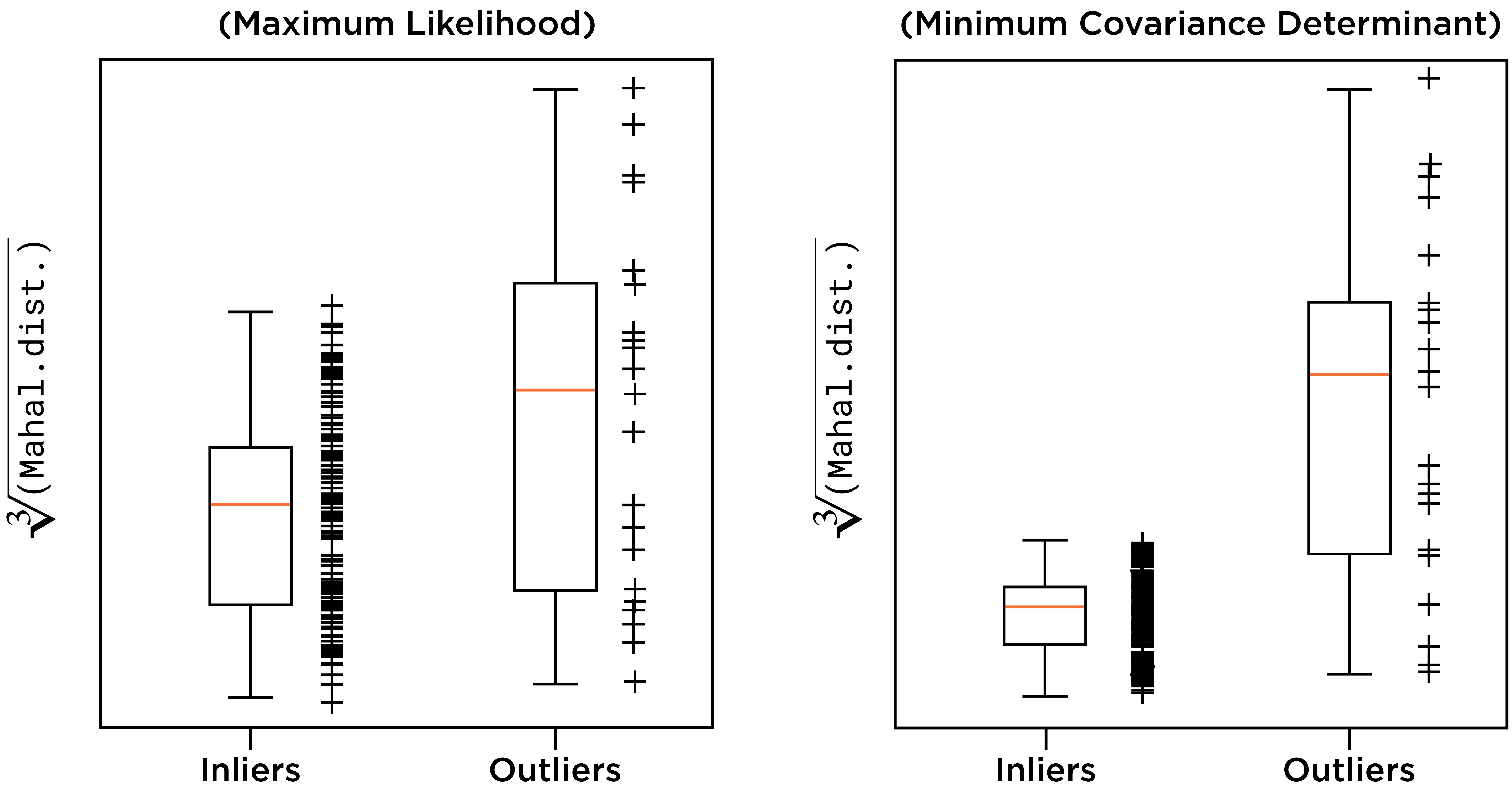
- Uses Mahalanobis Distance

Estimation is robust to outliers

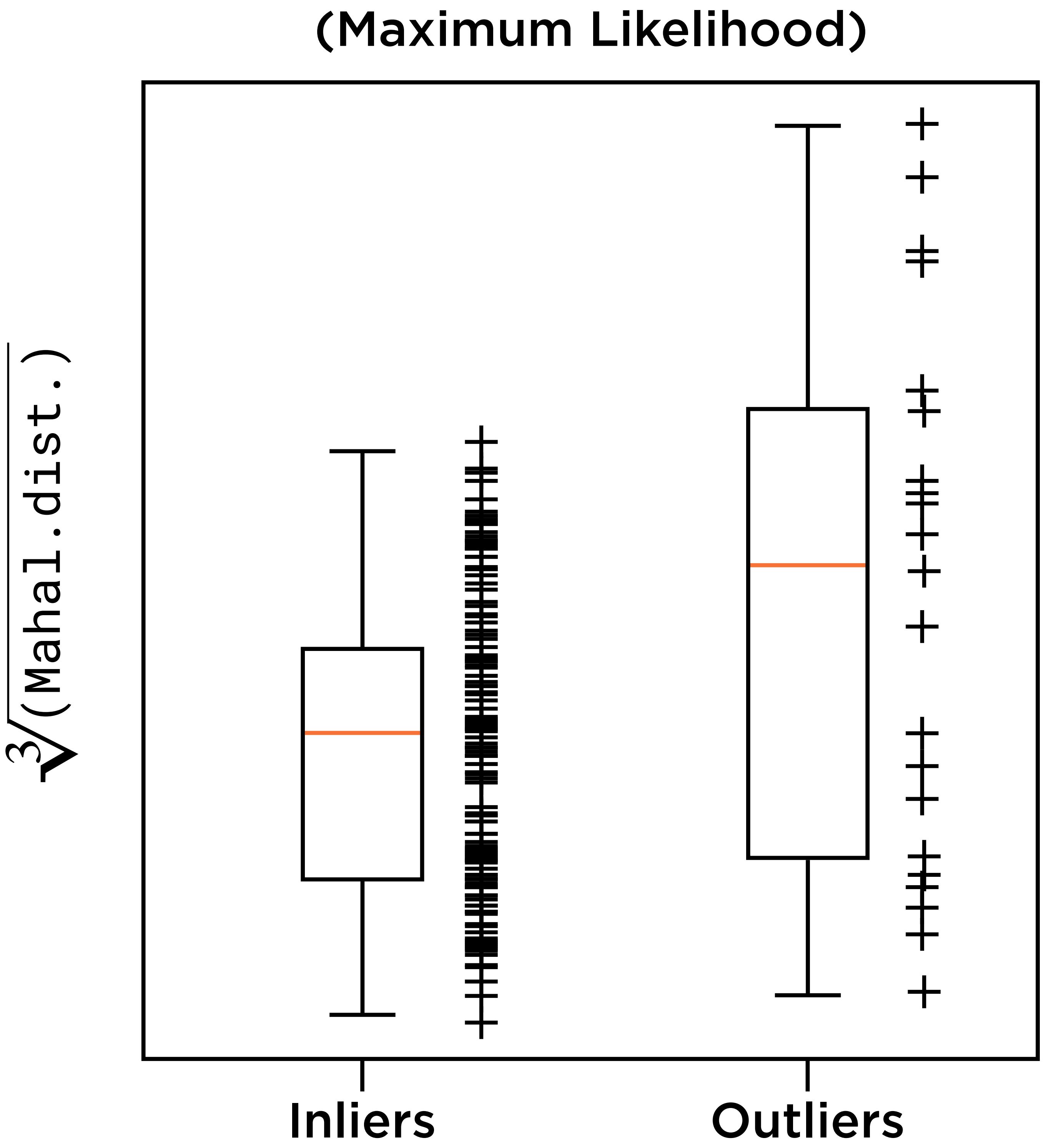
Mahalanobis Distance

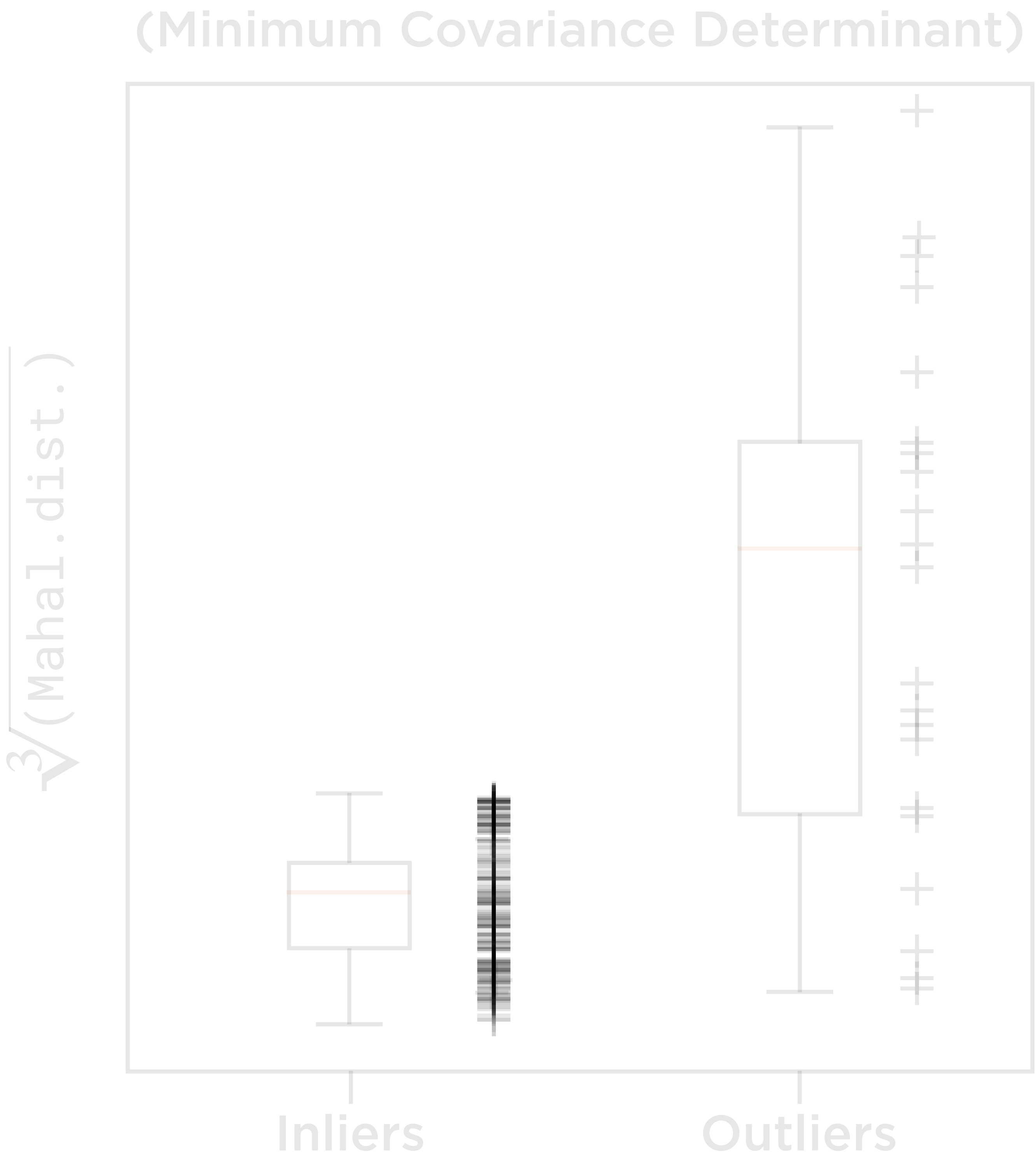
Measure of distance between two points; similar to Euclidean (L2) distance, but with the difference that each dimension is normalized to have equal variance.

Mahalanobis Distances of a Contaminated Dataset

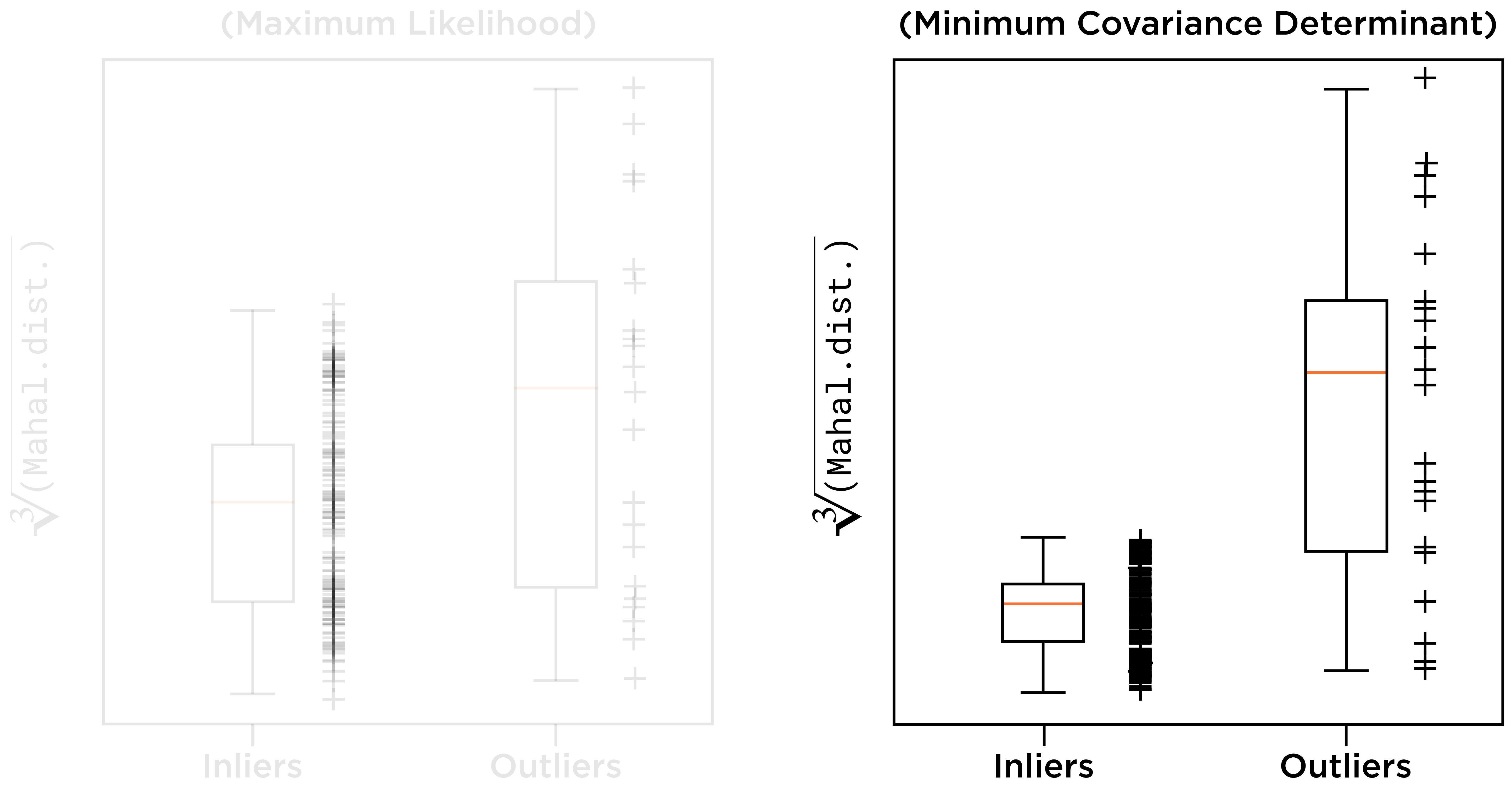


Mahalanobis Distances of a Contaminated Dataset





Mahalanobis Distances of a Contaminated Dataset



Isolation Forest

Outlier and Novelty Detection Algorithms in scikit-learn

Local Outlier Factor

Elliptic Envelope

Isolation Forest

Isolation Forest

Use Random Forests (common ML technique) to identify outliers

Forests of Decision Trees

Works particularly well for data of moderately high dimensionality

Jockey or Basketball Player?

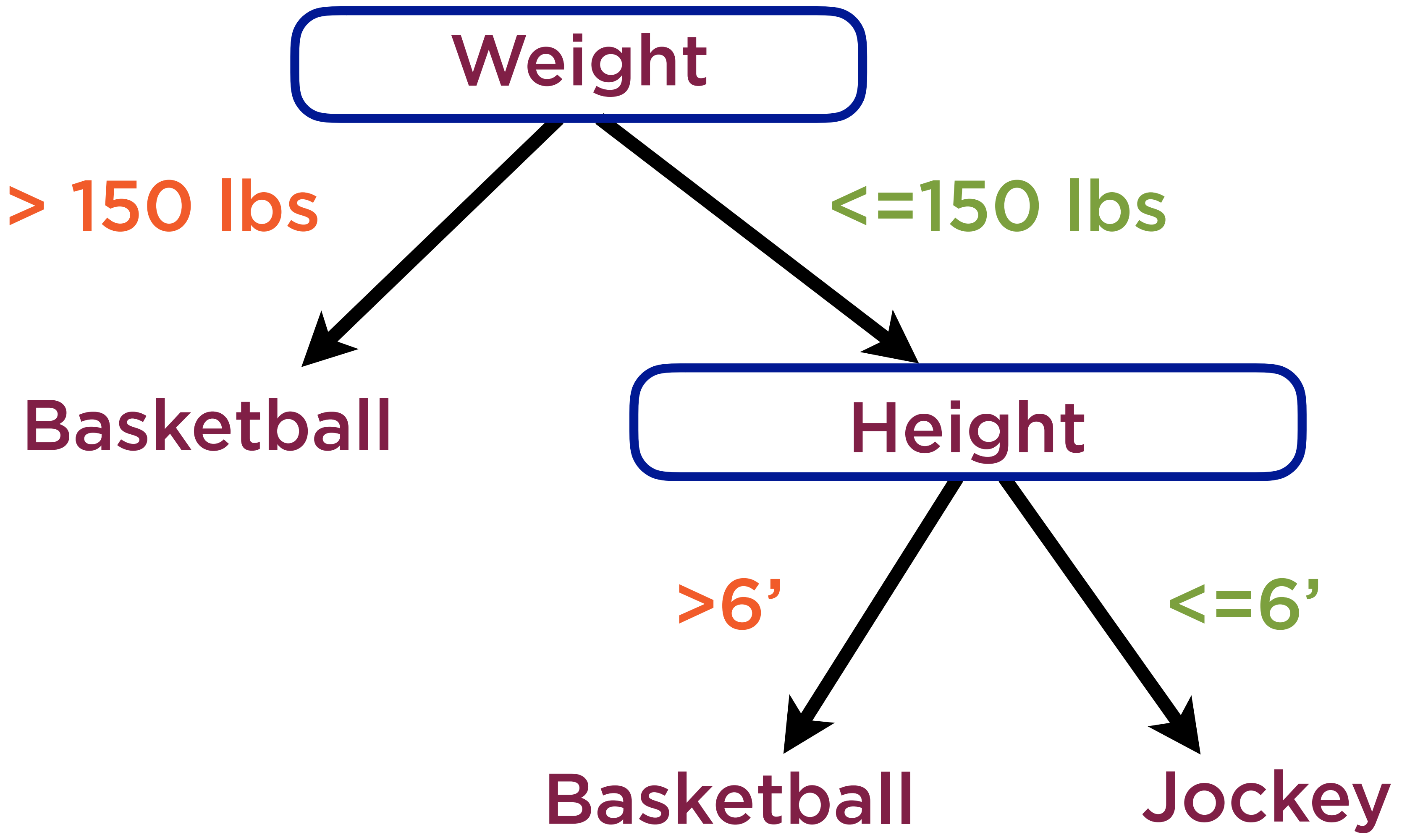
Jockeys

Tend to be light to meet horse carrying limits

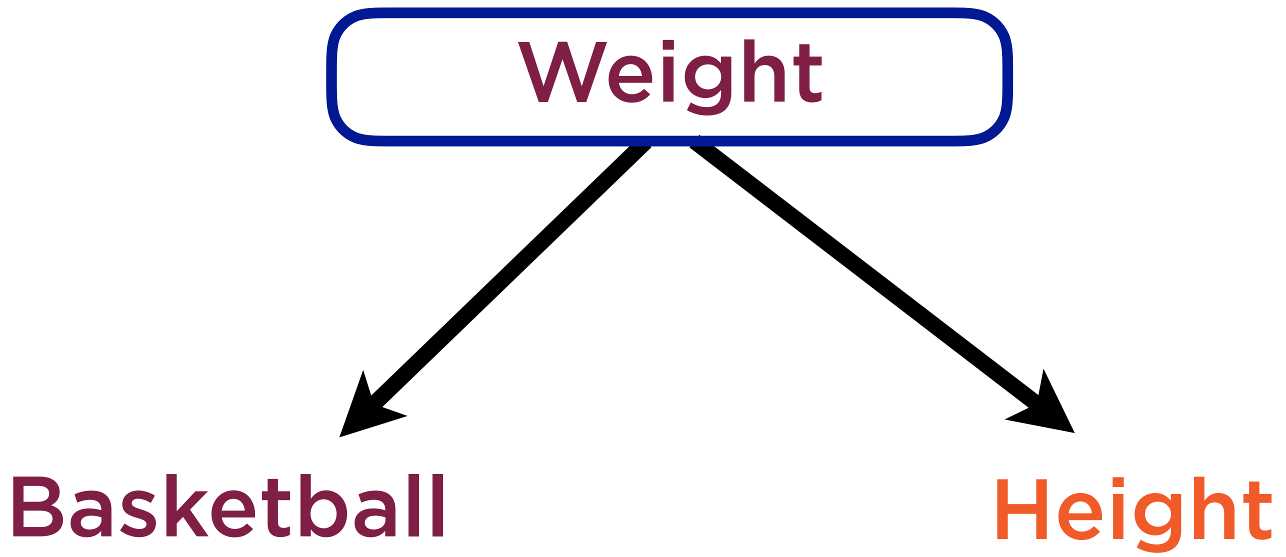
Basketball Players

Tend to be tall, strong and heavy

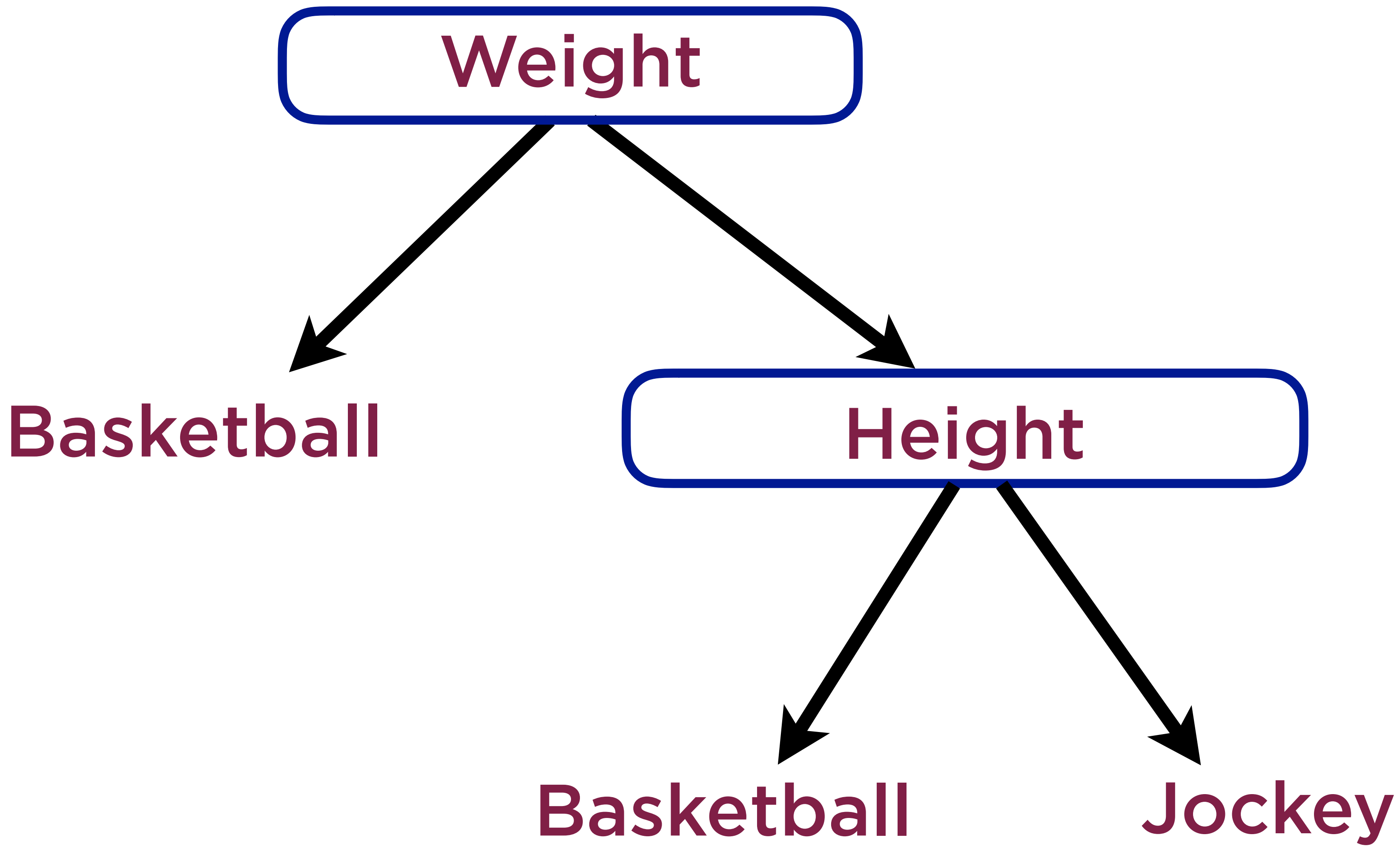
Fit Knowledge Into Rules



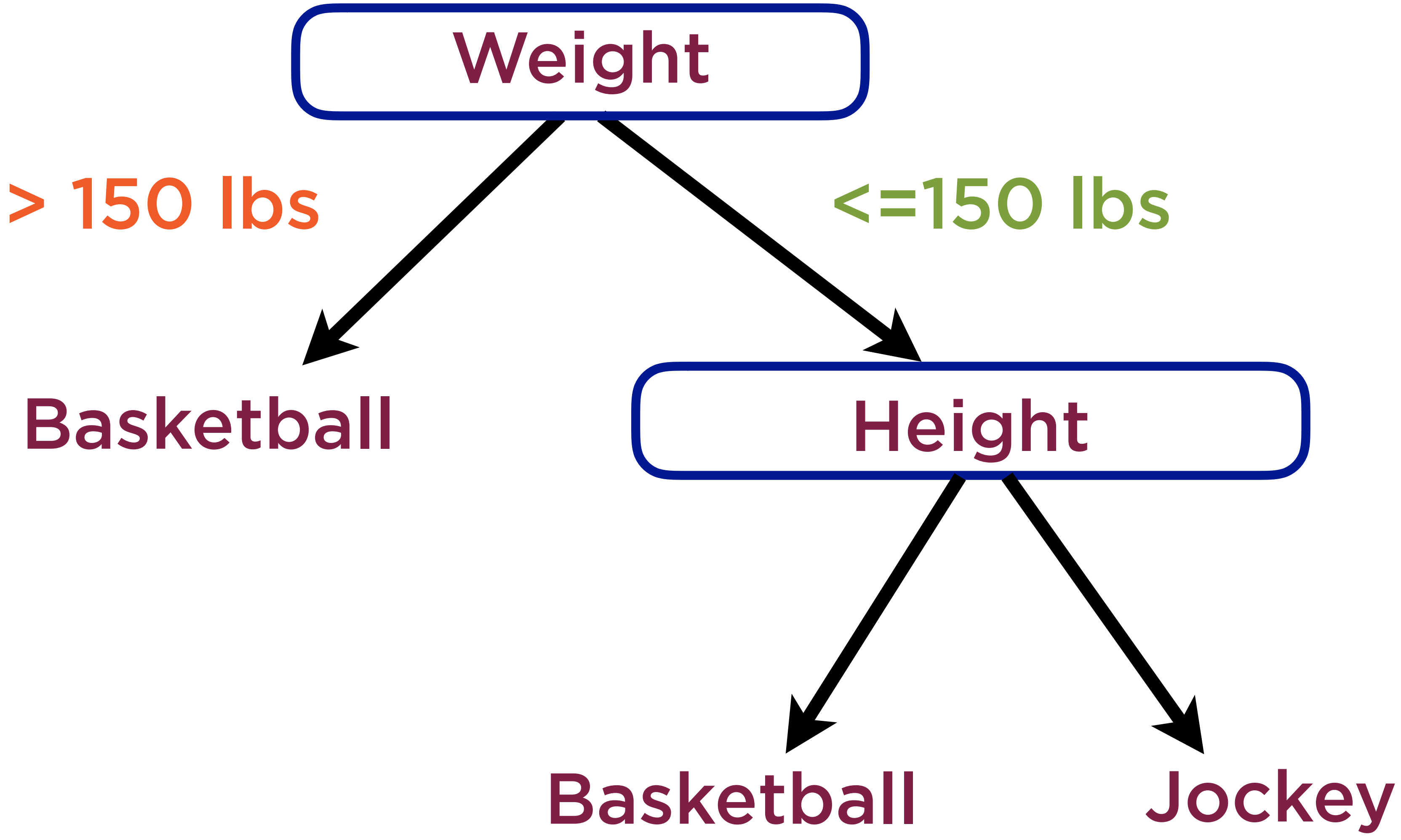
Decision Based on Weight



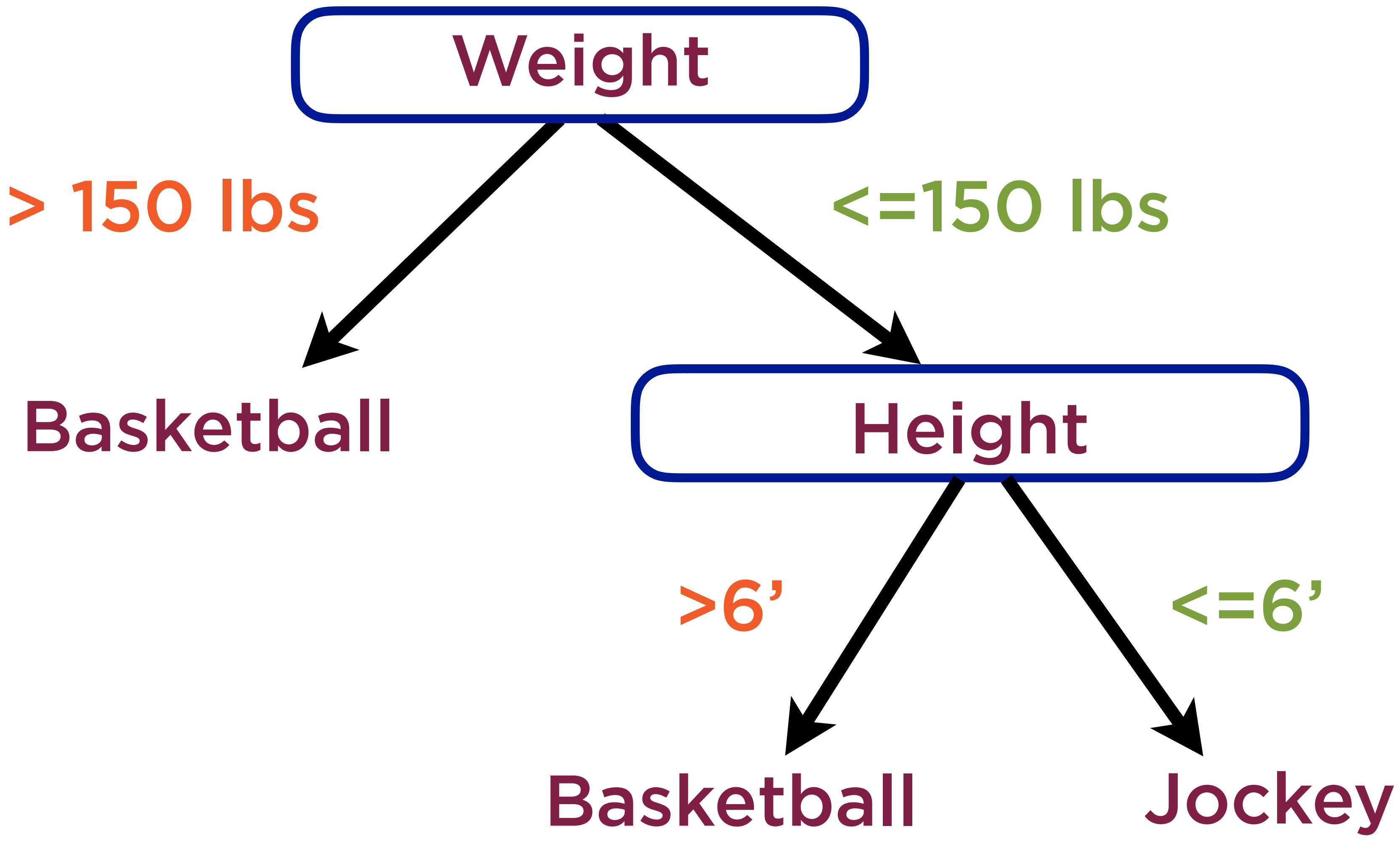
Decision Based on Height



Fit Knowledge Into Rules



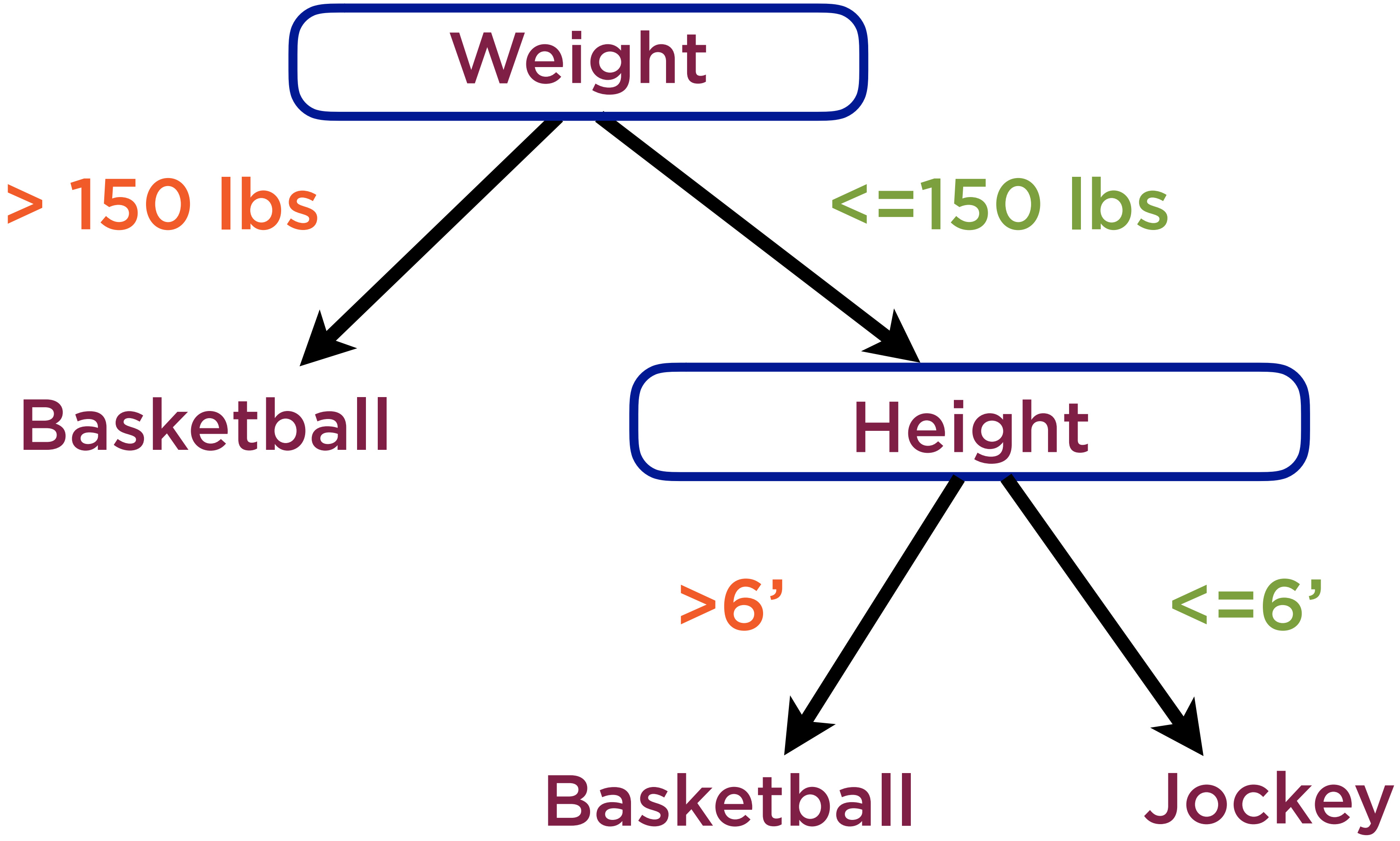
Fit Knowledge Into Rules



Decision Tree

Fit knowledge into rules

Each rule involves a threshold



Isolation Forest

Select a feature of the data point

Split records based on a randomly chosen value of the feature

Continue till a sample is isolated

Isolation Forest

Find how many splits are needed to isolate a point

- Place the point in a category by itself

Smaller the number of splits, the more likely the point is to be an outlier

- Smaller path length from root => greater probability of being outlier

Isolation Forest

Find how many splits are needed to isolate a point

- Place the point in a category by itself

Smaller the number of splits, the more likely the point is to be an outlier

- Smaller path length from root => greater probability of being outlier

Path length averaged over a forest of random trees determines outliers

Demo

Detecting outliers in data using Local Outlier Factor, Isolation Forest and Elliptic Envelope

Demo

Novelty detection using Local Outlier Factor, Isolation Forest and Elliptic Envelope

Demo

Detecting outliers in the head-brain dataset

Summary

Understanding outliers and novelties

Novelty and outlier detection uses

Algorithms for outlier and novelty detection

Local Outlier Factor

Elliptic Envelope

Isolation Forest