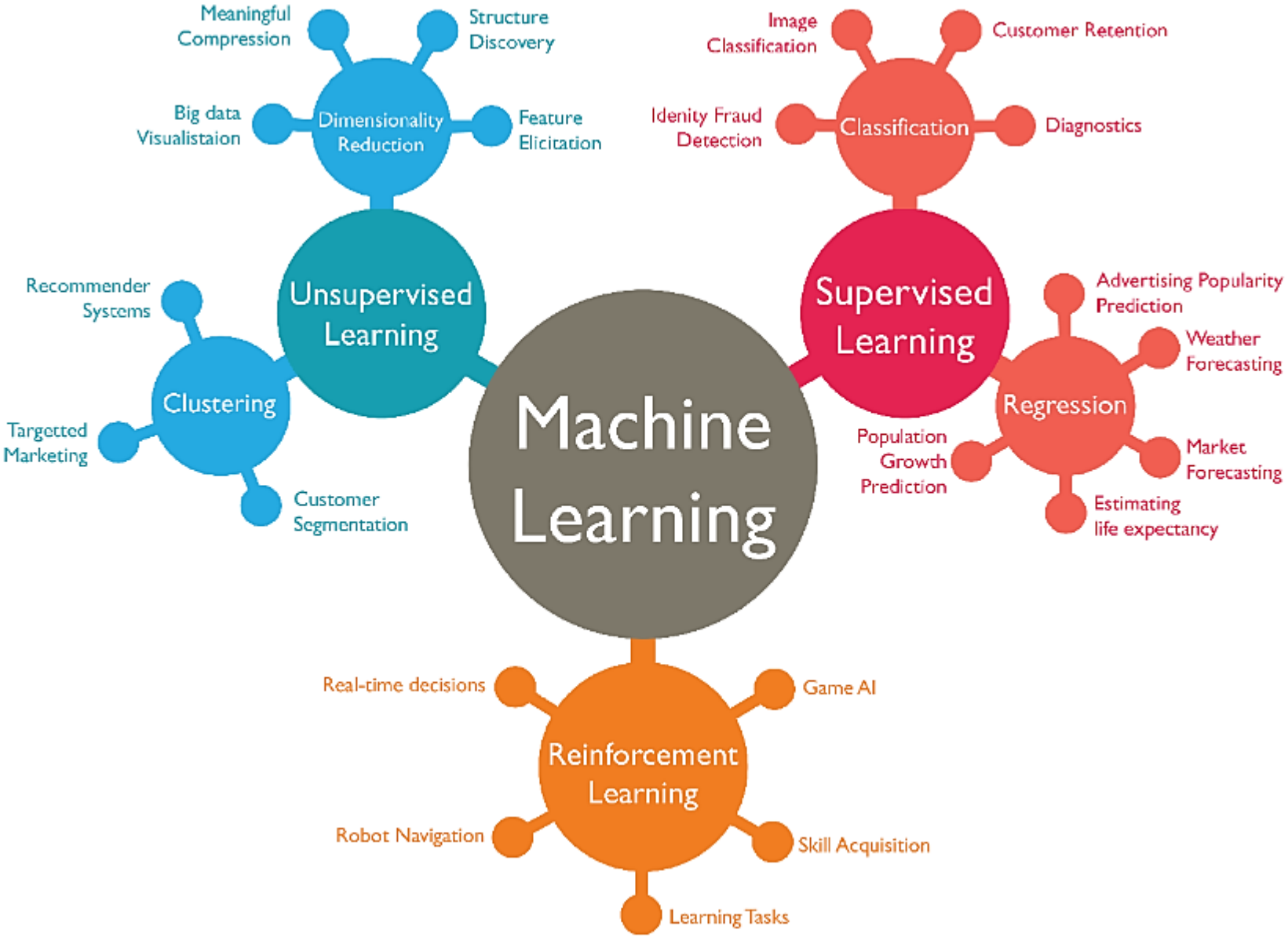
Advanced Clustering Techniques

Dimensionality reduction and Outlier detection

Machine learning applications



Introduction to Dimensionality Reduction

- **What is Dimensionality Reduction?**
  - Reducing the number of features (dimensions) while maintaining as much information as possible.
  - Benefits: Speed up model training, reduce storage, and eliminate noisy or redundant data.
- **Key Techniques**:
  - **Principal Component Analysis (PCA)**:
    - Linear transformation to project data into a lower-dimensional space.
    - Captures maximum variance with fewer components.
  - **t-Distributed Stochastic Neighbor Embedding (t-SNE)**:
    - Non-linear dimensionality reduction technique for visualizing high-dimensional data.
    - Great for clustering visualization.
  - **Linear Discriminant Analysis (LDA)**:
    - Supervised method focusing on maximizing class separability.
  - **Feature Selection**: Selecting a subset of relevant features for model training.

Dimensionality Reduction Examples

- **Text Categorization:** With vast amounts of online data, dimensionality reduction helps classify text documents into predefined categories by reducing the feature space (like word or phrase features) while maintaining accuracy.

- **Image Retrieval**: As image data grows, indexing based on visual content (colour, texture, shape) rather than just text descriptions has become essential. This allows for better retrieval of images from large databases.

- **Gene Expression Analysis**: Dimensionality reduction accelerates gene expression analysis, helping classify samples (e.g., leukaemia) by identifying key features, improving both speed and accuracy.

- **Intrusion Detection**: In cybersecurity, dimensionality reduction helps analyse user activity patterns to detect suspicious behaviours and intrusions by identifying optimal features for network monitoring.

Advantages of Dimensionality Reduction

- **Faster Computation**: With fewer features, machine learning algorithms can process data more quickly. This results in faster model training and testing, which is particularly useful when working with large datasets.

- **Better Visualization**: As we saw in the earlier figure, reducing dimensions makes it easier to visualize data, revealing hidden patterns.

- **Prevent Overfitting**: With fewer features, models are less likely to memorize the training data and overfit. This helps the model generalize better to new, unseen data, improving its ability to make accurate predictions.

Disadvantages of Dimensionality Reduction

- **Data Loss & Reduced Accuracy** – Some important information may be lost during dimensionality reduction, potentially affecting model performance.

- **Choosing the Right Components** – Deciding how many dimensions to keep is difficult, as keeping too few may lose valuable information, while keeping too many can lead to overfitting.

What Is Outlier Discovery (analysis)?

- **What are outliers?**
  - The set of objects are considerably dissimilar from the remainder of the data
- **Problem**: Define and find outliers in large data sets

- **Applications:**
  - Credit card fraud detection
  - Customer segmentation
  - Medical analysis
  - Telecommunication fraud detection
  - Network intrusion detection
  - Fault detection

Anomaly/Outlier Detection

- **What is Outlier Detection?**
  - Identifying data points that differ significantly from the rest of the dataset.
  - Applications: Fraud detection, anomaly detection, quality control.

- **Outlier Detection Methods**:
  - **Z-Score**: Identifies outliers based on how many standard deviations a point is from the mean.
  - **Isolation Forest**: Tree-based method for anomaly detection.
  - **One-Class SVM**: Suitable for outlier detection in high-dimensional datasets.

Methods in anomaly detection

Supervised Anomaly Detection
Labels available for both normal data and anomalies Cannot detect unknown and emerging anomalies

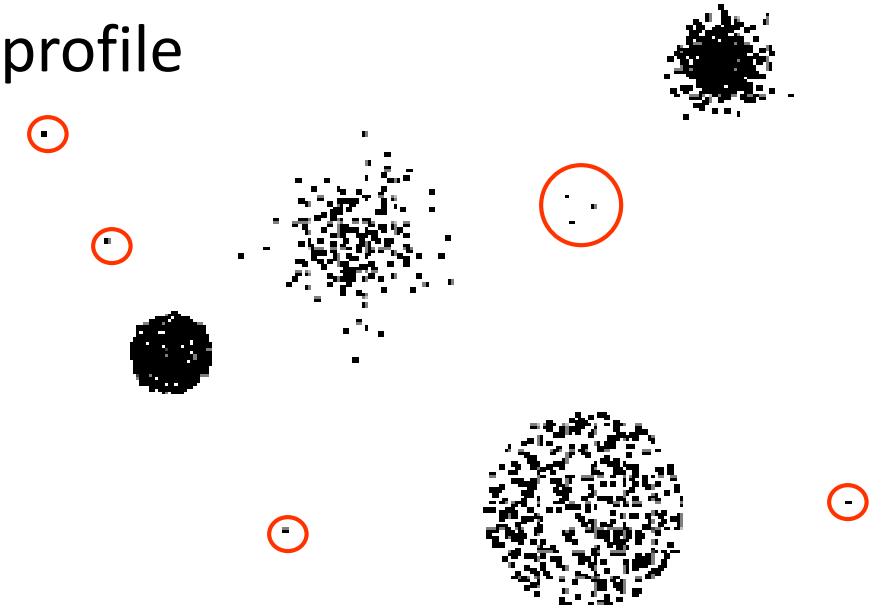
Semi-supervised Anomaly Detection
Labels available only for normal data Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies.

Unsupervised Anomaly Detection
No labels assumed Based on the assumption that anomalies are very rare compared to normal data

Anomaly Detection Schemes

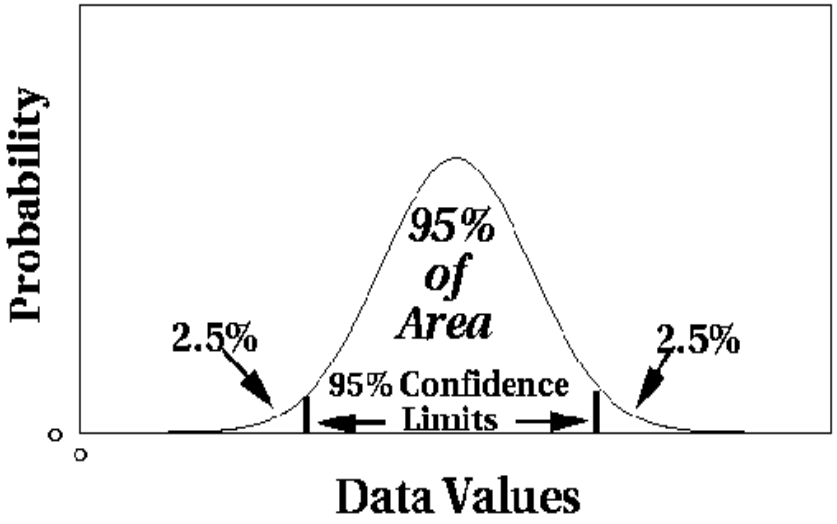
- General Steps
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based



Outlier Discovery: Statistical Approaches

- Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Apply a statistical test that depends on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers
- Drawbacks
  - most tests are for single attribute
  - In many cases, data distribution may not be known



Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
  - We need multi-dimensional analysis without knowing data distribution

- Distance-based outlier: A DB(p, D)-outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O

- Algorithms for mining distance-based outliers
  - Index-based algorithm
  - Nested-loop algorithm
  - Cell-based algorithm

Distance-based Approaches

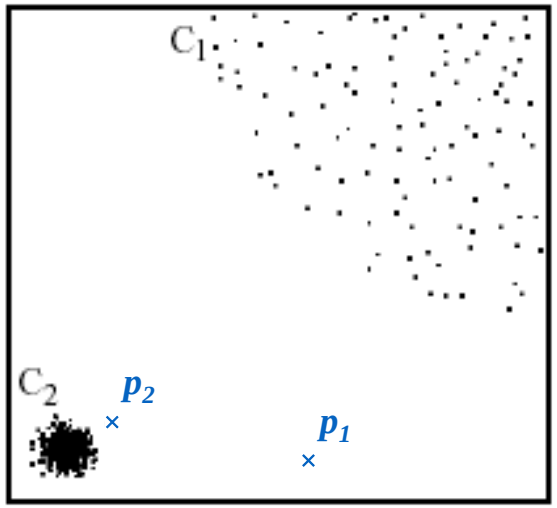
- Data is represented as a vector of features
- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *D*
    - The top n data points whose distance to the kth nearest neighbor is greatest
    - The top n data points whose average distance to the k nearest neighbors is greatest

Density-based: LOF approach

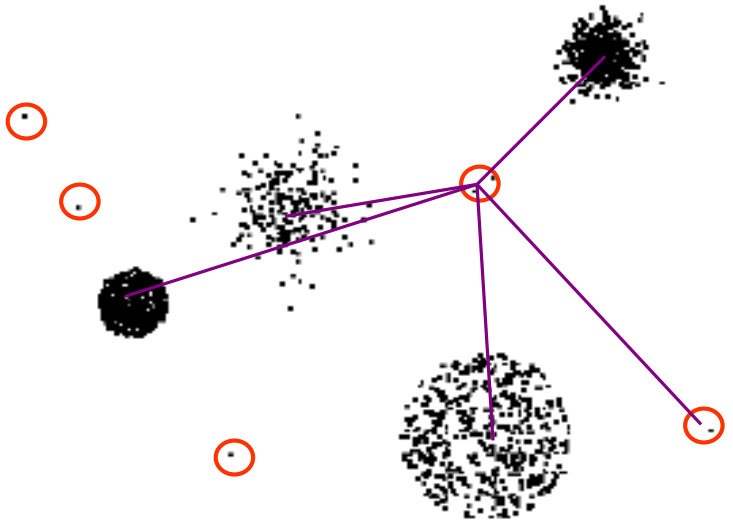
- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers



Anomaly Detection

- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack

- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data

Anomaly detection uses cases

Monitor Traffic Conditions

Network and Cyber Security Detect Intrusion

Monitor Biomarkers

Predictive maintenance, Production safety Monitoring

Agriculture Pest, Water Control

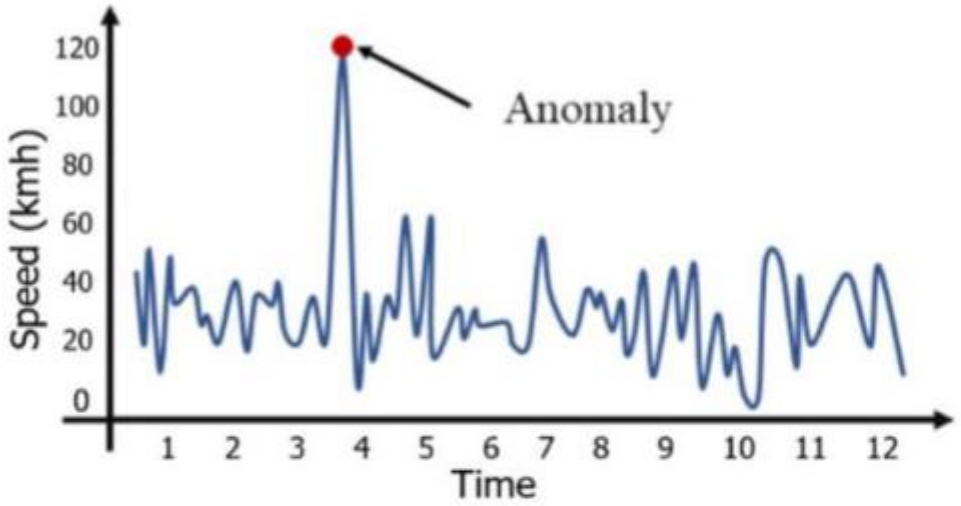
Monitor Energy Consumption

https ://www.slideserve.com/ablanco/detecting-outliers-in-iot-data-streams-powerpoint-ppt-presentation

Types of Anomalies: Case of traffic analysis

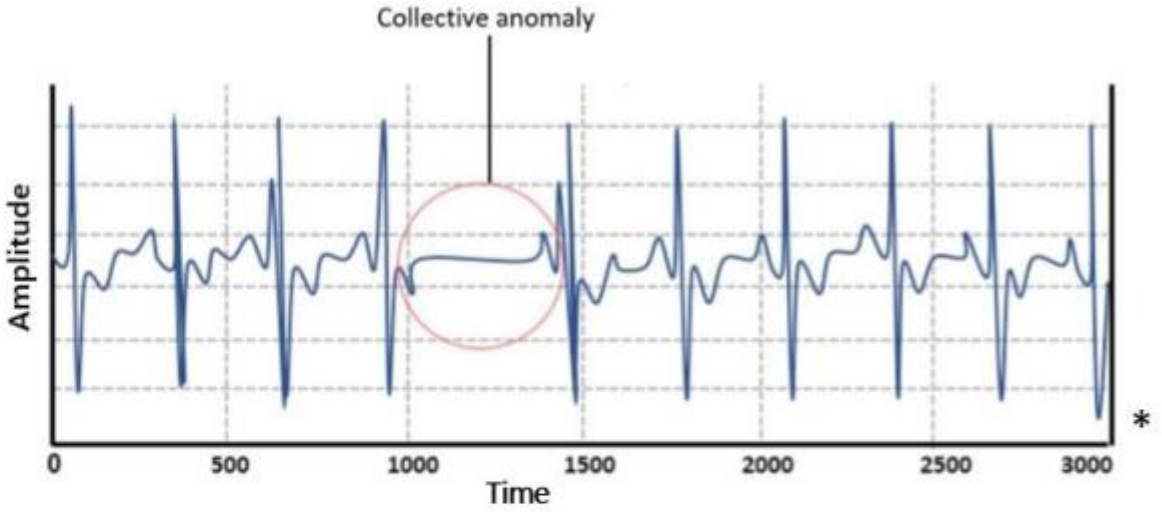
1. Point anomaly

A data point that is inconsistent from the rest of the data



2. Collective anomaly

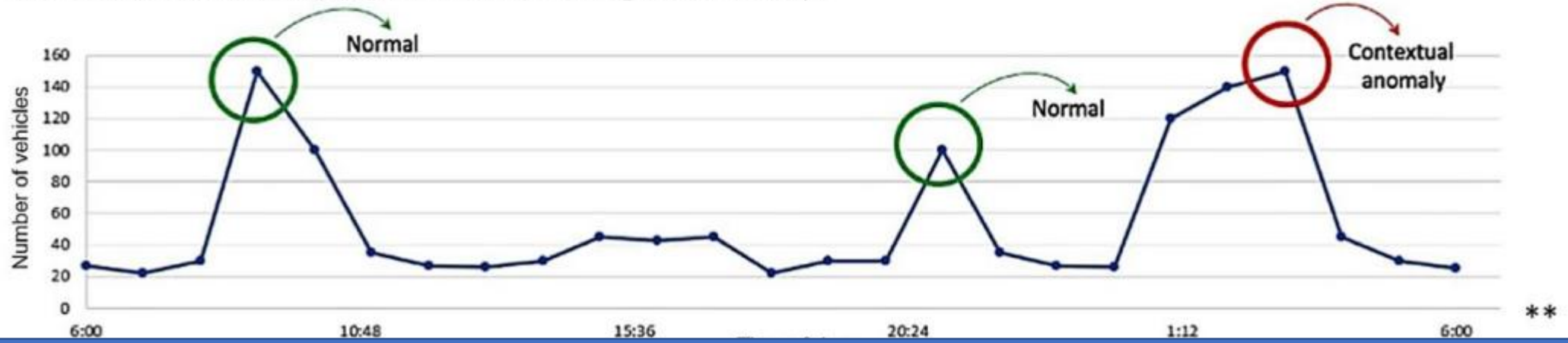
Sequence of linked data point that are inconsistent with the entire dataset.



*

3. Contextual anomaly

A conditional anomaly based on context e.g. time of day.



**

Isolation forest (iForest)

- Uses Binary Decision Trees bagging.
- It randomly selects a feature and then selects a split value between the maximum and minimum values.
- The number of splits required to isolate a sample is equivalent to the path length from the root node to the terminating node

Isolation forest

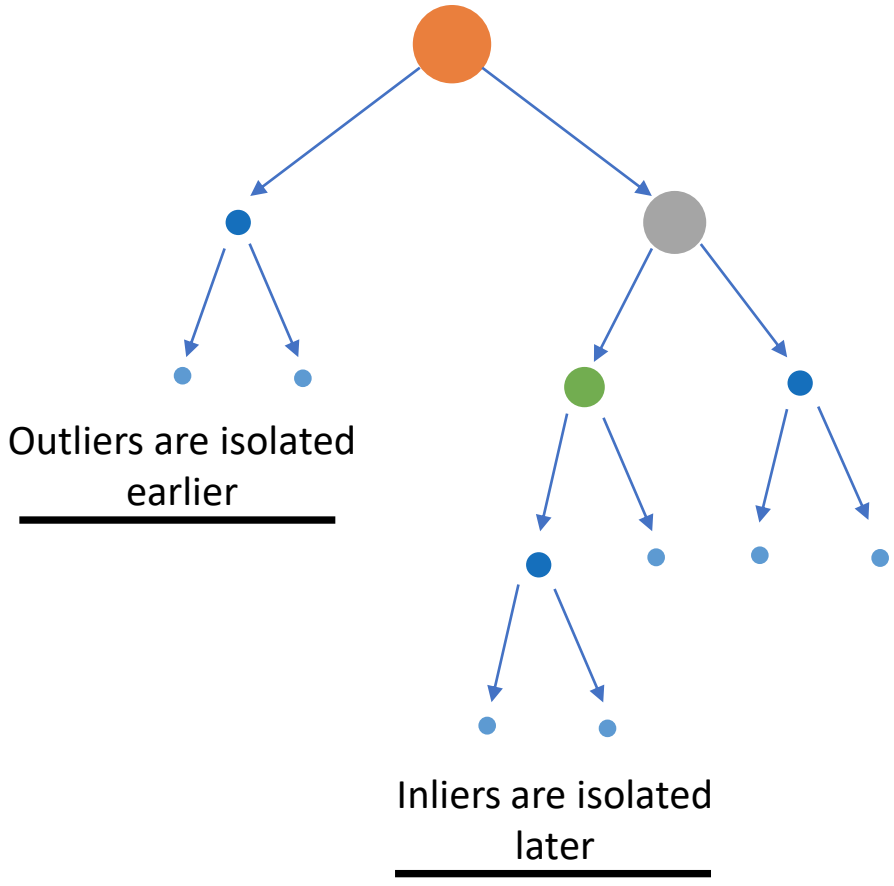
Each observation is given an anomaly score and the following decision can be made on its basis:

- A score close to 1 indicates anomalies
- Score much smaller than 0.5 indicates normal observations
- If all scores are close to 0.5 then the entire sample does not have clearly distinct anomalies

- IForest has a linear time complexity with a low memory requirement which is ideal for high volume data sets.

The Isolation Forest technique

- The Isolation Forest is a technique for the detection of outlier samples.
- Since outliers have features X that differ significantly from most of the samples, they are isolated earlier in the hierarchy of a decision tree.
- Outliers are detected by setting a threshold on the mean length (number of splits) from the top of the tree downwards.
- The Scikit-learn implementation provides a score for each sample that increases from -1 to +1 with the number of splits.



Practical Challenge: Project

- Combine Clustering, Dimensionality Reduction, and Outlier Detection
- Choose a dataset (e.g., Customer data, Financial Transactions, or Medical data).
- Perform the following:
  - Preprocess the data.
  - Apply Dimensionality Reduction (PCA, t-SNE).
  - Apply Clustering (K-Means, DBSCAN).
  - Detect Outliers using Z-Score or Isolation Forest.
  - Visualize results and discuss findings.