



Data Mining

Introduction to data mining



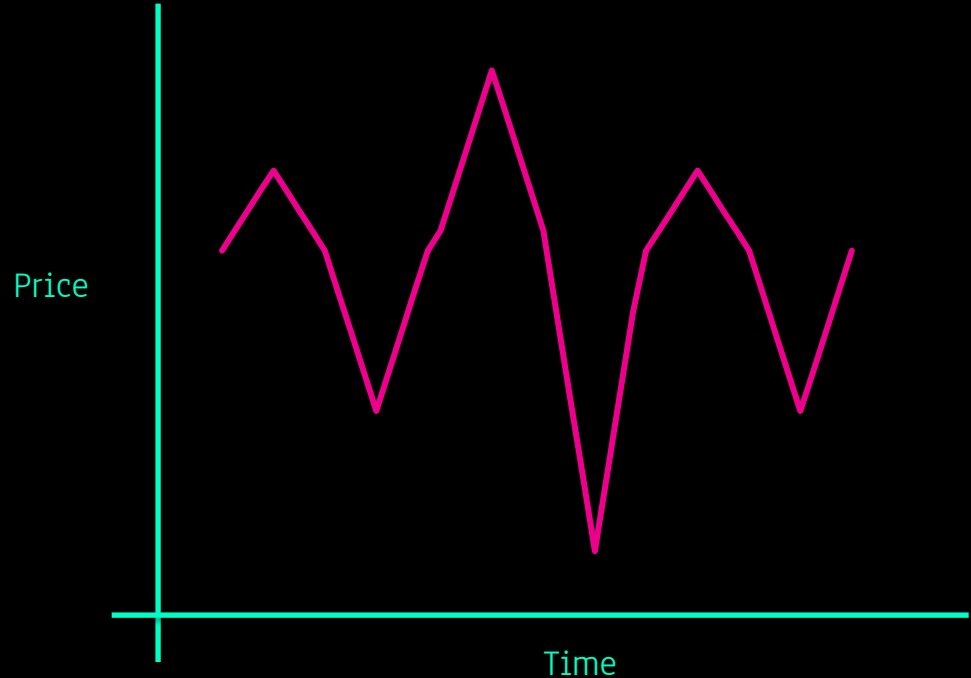
01

What is Data
Mining?

Data Mining

The following graph shows the **stock price** of a company.

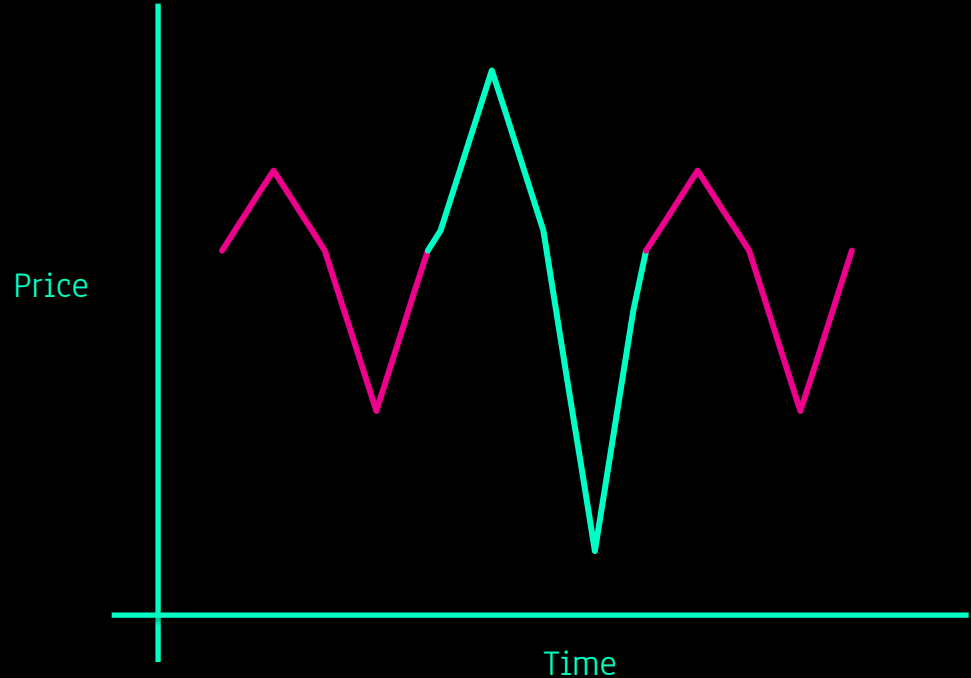
Do you find any pattern in it?



Data Mining

Data mining is a process of discovering useful **patterns or trends** within large datasets.

Data mining can help companies to develop more effective marketing strategies, increase sales, and decrease costs.





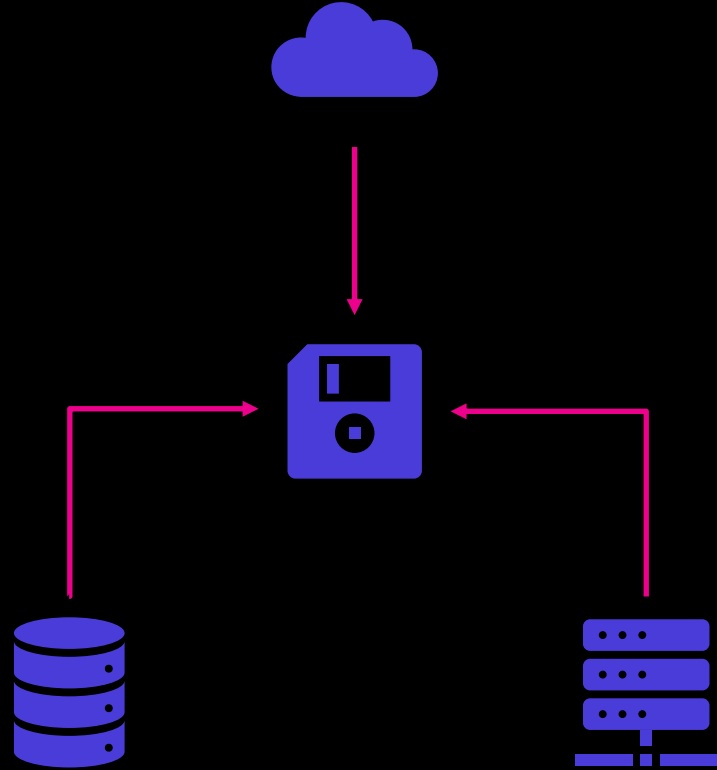
02

Steps Involved In Data Mining

Data Collection

Gather and collect the relevant data that you intend to analyze.

This data can come from several sources such as **Databases**, **Data Warehouses**, ...



Data Cleaning and Preprocessing

This involves identifying and correcting inconsistencies in the data.

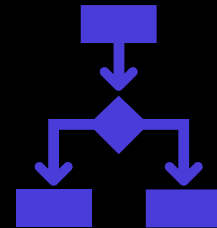
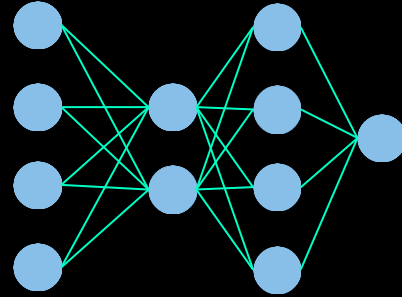
Example : A table which consists of missing data.

Employees	Exceeded KPIs	Leadership Capability	Aged < 30	Outcome
6	6	2		Promoted
4		2	4	Not Promoted

Data Modeling

Data Modeling involves **building models** to represent the data and to **predict future outcomes**.

We have a variety of data mining models available, such as **Decision trees**, **Neural networks**, **Support Vector Machines**, ...





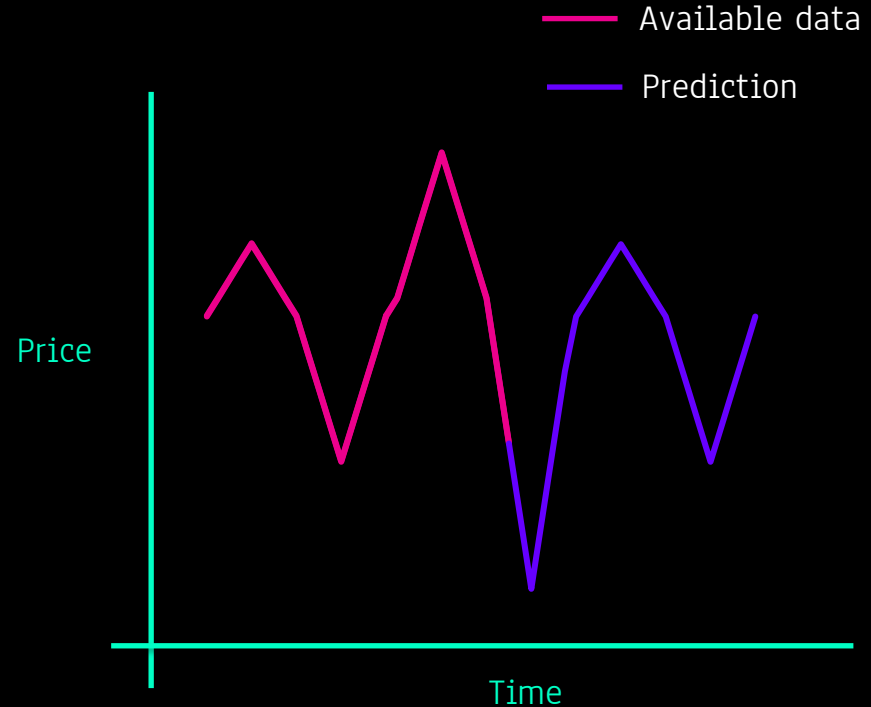
03

Why is Data Mining
Important?

Predictive Modeling

Data mining can be used to build predictive models that can forecast **future events or outcomes**.

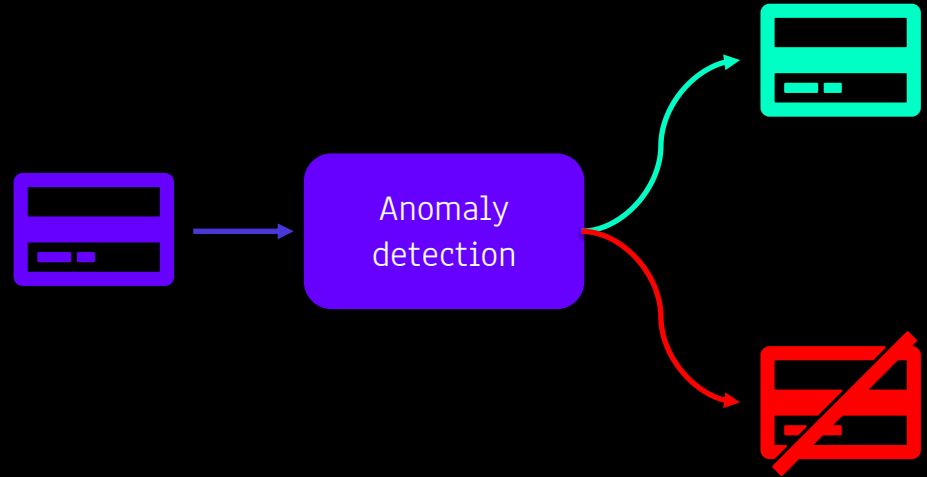
For example, it can be used to predict sales trend, market demand, ...



Fraud Detection

Data mining is a powerful tool for detecting **fraudulent activities**. **Unusual patterns** and **anomalies** in data can trigger alerts and investigations.

Example : Detecting fraudulent credit card transactions.





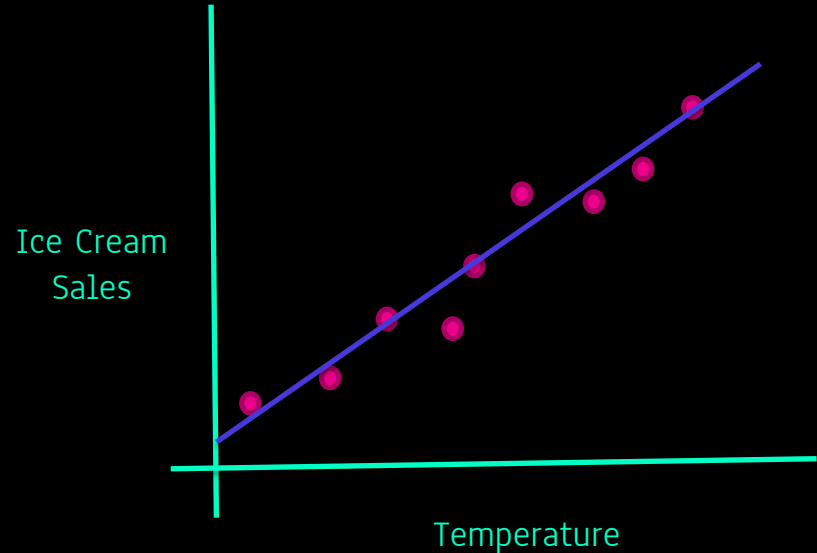
04

Some Prediction Mechanisms

Regression

Regression is a technique which is used to identify the **relationship** between different attributes and **predict** the value of one attribute with the help of other.

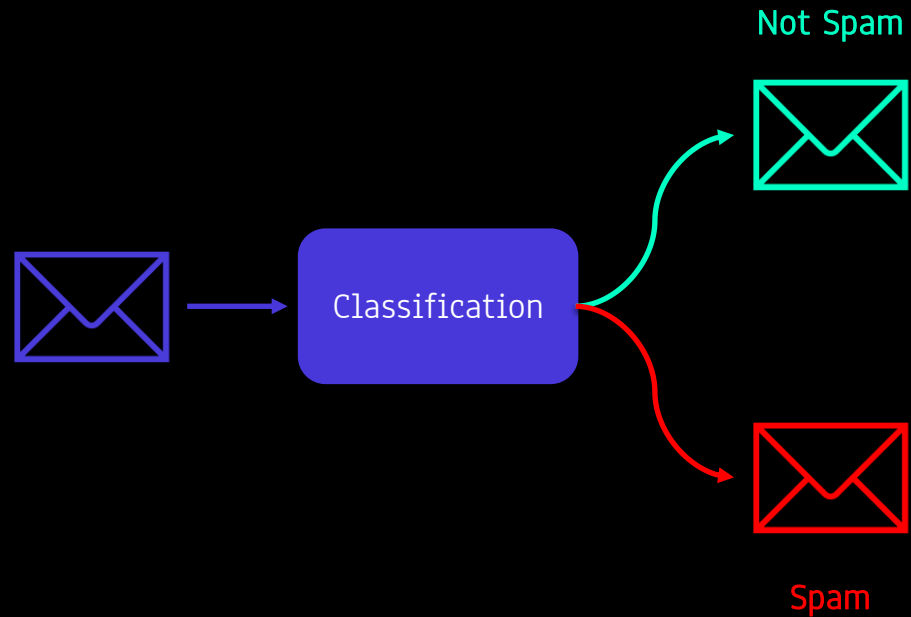
Example : Relationship between **Temperature** and **Ice Cream Sales**.



Classification

Classification refers to the process of **categorizing data** into classes based on their properties.

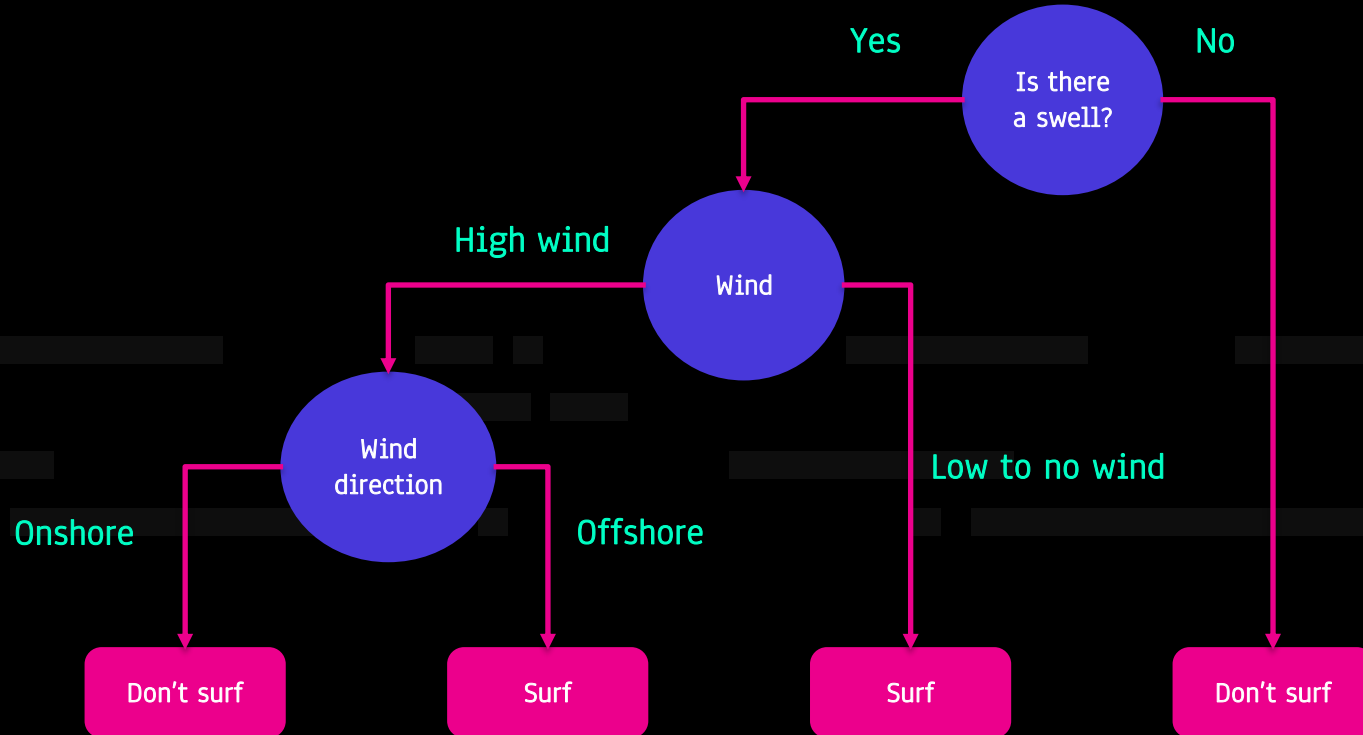
Example : Classification of an email as **spam** or **not**.



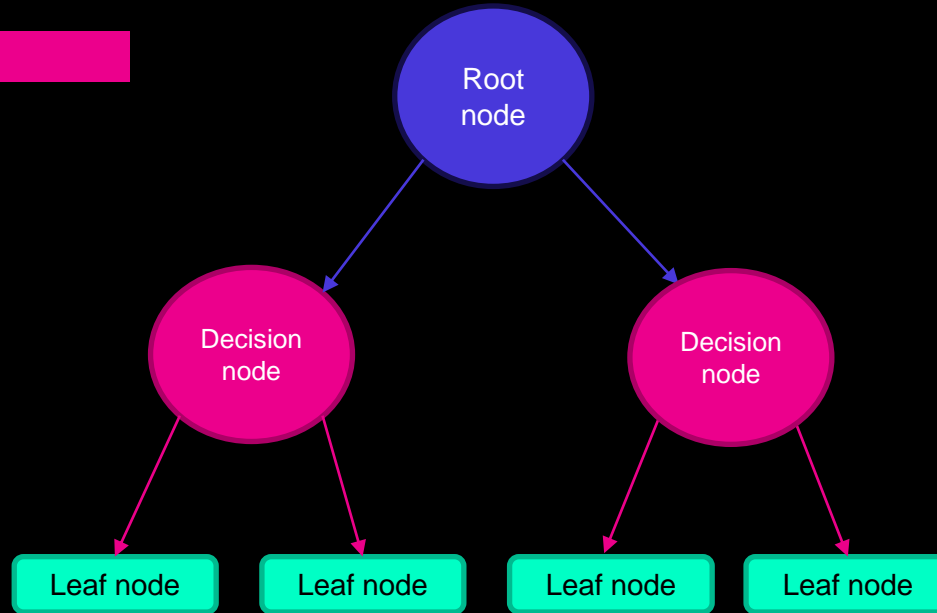
Let's imagine that you are trying to assess
whether or not you should go surf



You may use the following decision rules to make a choice



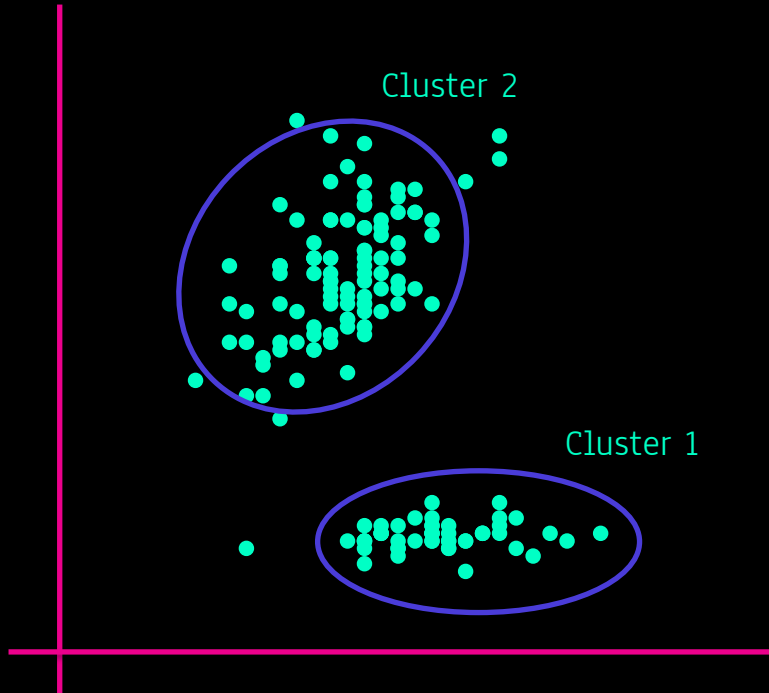
Structure of a decision tree



Clustering

Clustering is a data mining technique used to **group similar data points together** based on their similarities.

The goal is to **discover natural groupings** in a dataset **without any prior knowledge** about the groupings.





05

Types of Clustering

Partition Based Clustering

Partitioning objects into **k number of clusters** where each partition represents one cluster.

Examples:

- K-Means Clustering
- CLARA (**Clustering LARge Applications**)
- CLARANS (**Clustering Large Applications based upon RANdomized Search**)

Hierarchical Clustering

Depending upon the **hierarchy**, these clustering methods create a cluster having a **tree-type structure** where each newly formed clusters are made using priorly formed clusters.

Examples:

- CURE (**Clustering Using Representatives**)
- BIRCH (**Balanced Iterative Reducing Clustering and using Hierarchies**)
- Linkage clustering (**Single/complete/average linkage**)

Density Based Clustering

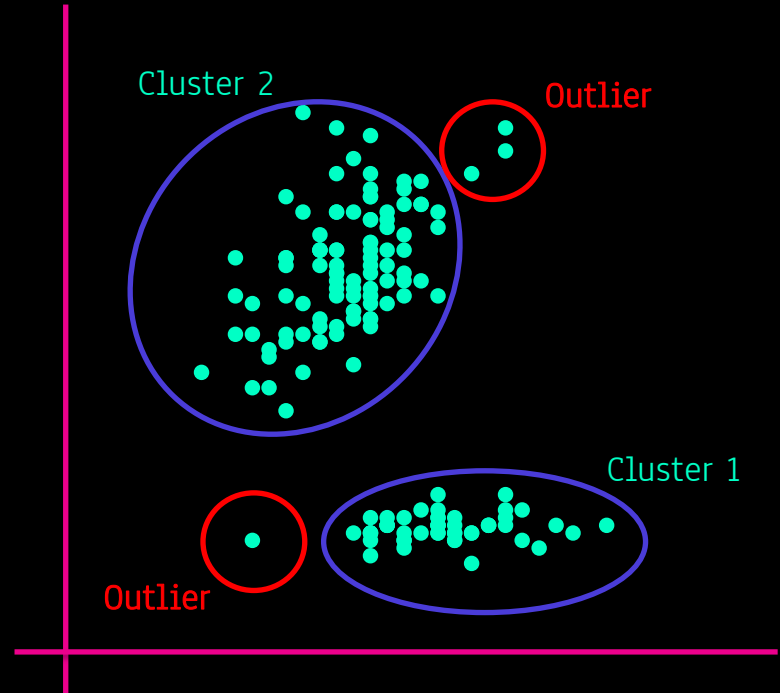
These methods of clustering recognize clusters of **dense regions** that possess some similarity and are **distinct from low dense regions** of the space.

Examples:





- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)

Outlier Analysis

Outlier analysis is the process of identifying and examining data points that **significantly differ** from the rest of the dataset.



Applications of Outlier Analysis

-  **Healthcare:** Identify patients with unusual medical conditions.
-  **Marketing:** Identify customers with high or low purchasing habits.
-  **Cybersecurity:** Detect abnormal network behaviour or suspicious activity.
-  **Banking:** Prevent credit card frauds by detecting unusual usage.



06

Real World
Applications

Spotify

Spotify uses clustering algorithms to recommend music by **grouping users into clusters** with similar listening preferences.

It analyzes the music listening history of users and **identifies patterns** and **commonalities** among them.



Visa



Visa uses anomaly detection algorithms to identify fraudulent transactions by establishing a **baseline of normal user behaviour**.

Any **deviation** from this baseline, such as unusual transaction amounts, locations, or patterns, **raises suspicion and triggers fraud alerts**.



Amazon



Amazon uses predictive modeling to **forecast sales** by analyzing historical sales data, market trends, and various factors like seasonality, promotions, and customer behaviour.

These models help **optimize inventory, plan logistics,** and **make informed decisions** to meet customer demand efficiently.





Thank you!