# BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

## Programme: M.Sc. Statistics

## Course Title: Introduction to Big Data Analytics

## Course Code: 23ST02VAC

# Unit-II

# BIG DATA ANALYTICS

**Dr. T. Jai Sankar**
**Associate Professor and Head**
**Department of Statistics**

**Ms. S. Soundarya**
**Guest Faculty**
**Department of Statistics**

# INTRODUCTION TO BIG DATA ANALYTICS

## UNIT - II

## DATA ANALYTICS

Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

### Types of Data Analysis

The four types of data analysis are:

- Descriptive Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

Below, we will introduce each type and give examples of how they are utilized in business.

### Descriptive Analysis

The first type of data analysis is descriptive analysis. It is at the foundation of all data insight. It is the simplest and most common use of data in business today. Descriptive analysis answers the “what happened” by summarizing past data, usually in the form of dashboards.

Business applications of descriptive analysis include:

- KPI dashboards
- Monthly revenue reports
- Sales leads overview

### Predictive Analysis

Predictive analysis attempts to answer the question “what is likely to happen”. This type of analytics utilizes previous data to make predictions about future outcomes.

This type of analysis is another step up from the descriptive and diagnostic analyses. Predictive analysis uses the data we have summarized to make logical predictions of the outcomes of events. This analysis relies on statistical modelling, which requires added technology and manpower to forecast. It is also important to understand that forecasting is only an estimate; the accuracy of predictions relies on quality and detailed data.

Business applications of predictive analysis include:

- Risk Assessment
- Sales Forecasting
- Using customer segmentation to determine which leads have the best chance of converting
- Predictive analytics in customer success teams

**Prescriptive Analysis**

The final type of data analysis is the most sought after, but few organizations are truly equipped to perform it. Prescriptive analysis is the frontier of data analysis, combining the insight from all previous analyses to determine the course of action to take in a current problem or decision.

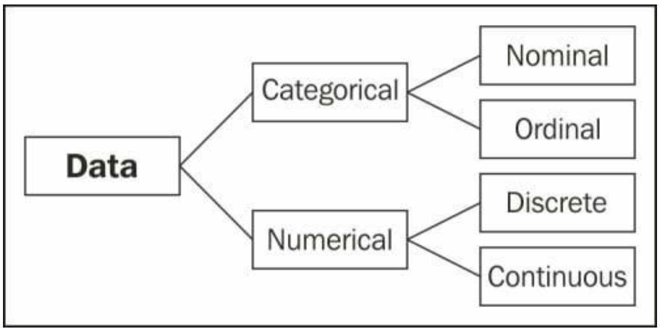
Prescriptive analysis utilizes state of the art technology and data practices. It is a huge organizational commitment and companies must be sure that they are ready and willing to put forth the effort and resources.

**Nature of data in data analytics**

Nature of Data: the nature of data under descriptive statistics is sets. A set is simply a collection of numbers that behaves in predictable ways. Data reflects real life, and there are patterns everywhere to be found. Descriptive analysis describes those patterns.

**NATURE OF DATA**

Data is the plural of datum, so it is always treated as plural. We can find data in all the situation of the world around us, in all the structured or unstructured, in continuous or discrete conditions, in weather records, stock market logs, in photo albums, music playlist, or in our Twitter accounts. In fact, data can be seen as the essential raw material of any kind of human activity. Data are known facts or things used as basis for inference or reckoning. As shown in the following figure, we can see Data in two distinct ways: Categorical and Numerical:



- Categorical data are values or observations that can be sorted into groups or categories. There are two types of categorical values, nominal and ordinal.
- A nominal variable has no intrinsic ordering to its categories. For example, housing is a categorical variable having two categories (own and rent).
- An ordinal variable has an established ordering. For example, age as a variable with three orderly categories (young, adult, and elder).
- Numerical data are values or observations that can be measured. There are two kinds of numerical values, discrete and continuous .
- Discrete data are values or observations that can be counted and are distinct and separate. For example, number of lines in a code.
- Continuous data are values or observations that may take on any value within a finite or infinite interval. For example, an economic time series such as historic gold prices.

## RELATIONAL DATA BASE TO BIG DATA

Big Data is a Database that is different and advanced from the standard database. The Standard Relational databases are efficient for storing and processing structured data. It uses the table to store the data and structured query language (SQL) to access and retrieve the data.

Big Data is the type of data that includes unstructured and semi-structured data. There are specific types of database known as NoSQL databases, There are several types of NoSQL Databases and tools available to store and process the Big Data. NoSQL Databases are optimized for data analytics using the Big Data such as text, images, logos, and other data formats such as XML, JSON. The big data is helpful for developing data-driven intelligent applications.

### Difference Between Big Data and Database

Given below is the difference between Big Data and Database:

- Big Data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases. A traditional database is not able to capture, manage, and process the high volume of data with low-latency While Database is a collection of information that is organized so that it can be easily captured, accessed, managed and updated.

- Big Data refers to technologies and initiatives that involve data that is too diverse i.e. varieties, rapid-changing or massive for skills, conventional technologies, and infrastructure to address efficiently While Database management system (DBMS) extracts information from the database in response to queries but it in restricted conditions.
- There can be any varieties of data while DB can be defined through some schema.
- It is difficult to store and process while Databases like SQL, data can be easily stored and process.

Here's the road map for this introductory post:

- Overview of database engines
- Data modeling
- Entity-relationship modeling
- Relational model

**1. Overview of Database Engines:** So why should we use a database? Well, the first reason is that a database gives a lot of useful abstractions. Secondly, it also has these properties known as **ACID** (Atomicity, Consistency, Isolation, Durability).

- **Atomicity**: Operations executed by the database will be atomic / "all or nothing." For example, if there are 2 operations, the database ensures that either both of them happen or none of them happens.
- **Consistency**: Anyone accessing the database should see consistent results.
- **Isolation**: If there are multiple clients trying to access the database, there will be multiple transactions happening simultaneously. The database needs to be able to isolate these transactions.
- **Durability**: When writing a result into the database, we should be guaranteed that it won't go away.

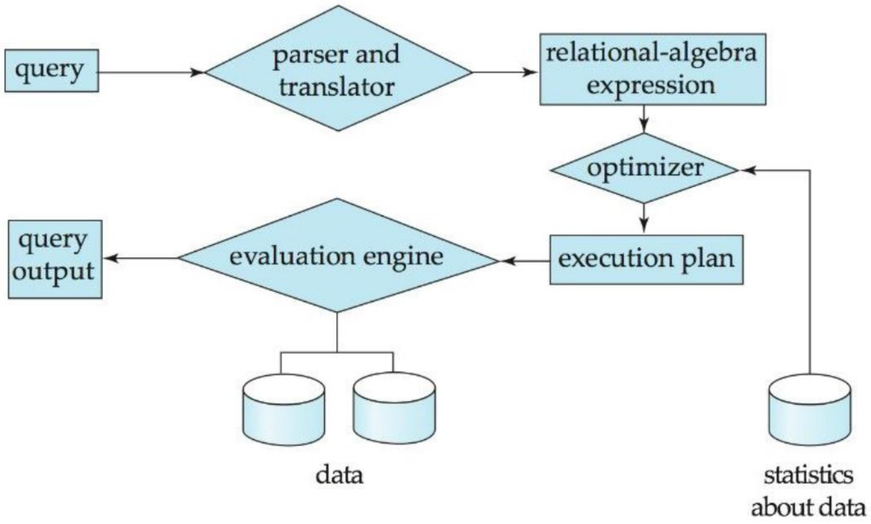
In a database engine, there are 2 main components: the storage manager and the query processor**.** The storage manager is the interface between the database and the operating system. It is responsible for authorization, interaction with the OS file system (accessing storage and organizing files), and efficient data storage/modification (indexing, hashing, buffer management).

One very important piece of the storage manager is the transaction manager**.** It ensures the database is consistent (if a failure occurs) and atomic. It also does concurrency control to make sure multiple operations result in a consistent database.

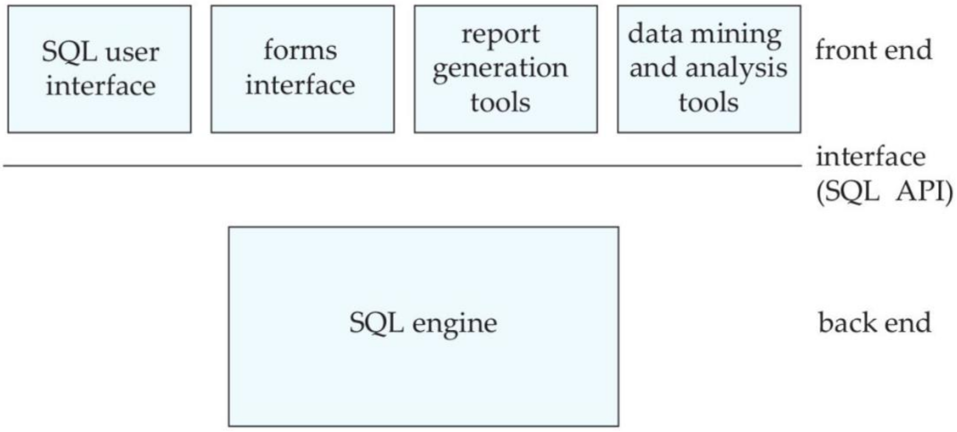
For those who are not familiar, transactions are collections of operations for a single task. Examples include:

- Assume a constraint balance > 0
- Deduct 50 from A
- Add 50 to the balance of B
- Store the new balance

On the other hand, the query processor is responsible for 3 major jobs: parsing and translation, optimization, and evaluation. The diagram below gives an overview of the query processor:

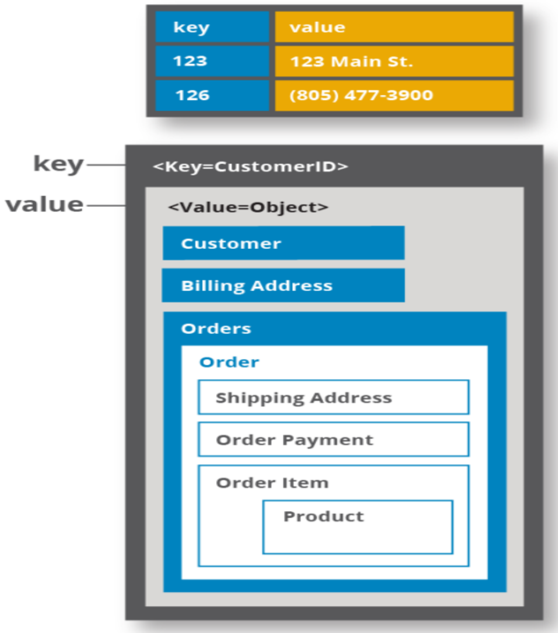


For most of the time, we can think of our database as a black box, as seen in the diagram below (the SQL engine). We ask queries of our database (via SQL API), and the database gives us the answer. The front end that we see includes SQL user interface, forms interface, report generation tools, data mining/analysis tools…



## KEY-VALUE STORE

A key-value store**,** or key-value database, is a type of data storage software program that stores data as a set of unique identifiers, each of which have an associated value. This data pairing is known as a "key-value pair." The unique identifier is the "key" for an item of data, and a value is either the data being identified or the location of that data.



| key | value |
|---|---|
| 123 | 123 Main St. |
| 126 | (805) 477-3900 |

### An example of a key-value store

The key could be anything, depending on restrictions imposed by the database software, but it needs to be unique in the database so there is no ambiguity when searching for the key and its value. The value could be anything, including a list or another key-value pair. Some database software allows you to specify a data type for the value.

### Key-value pair mean

A key-value pair is two pieces of data associated with each other. The key is a unique identifier that points to its associated value, and a value is either the data being identified or a pointer to that data.

A key-value pair is the fundamental data structure of a key-value store or key-value database, but key-value pairs have existed outside of software for much longer. A telephone directory is a good example, where the key is the person or business name, and the value is the phone number. Stock trading data is another example of a key-value pair. In this case, you may have a key associated with values for the stock ticker, whether the trade was a buy or sell, the number of shares, or the price of the trade.

### Key-value store advantages

There are a few advantages that a key-value store provides over traditional row-column-based databases. Thanks to the simple data format that gives it its name, a key-value store can be very fast for read and write operations. And key-value stores are very flexible, a valued asset in modern programming as we generate more data without traditional structures.

Also, key-value stores do not require placeholders such as “null” for optional values, so they may have smaller storage requirements, and they often scale almost linearly with the number of nodes.

### Key-value database use cases

The advantages listed above naturally lend themselves to several popular use cases for key-value databases.

- Web applications may store user session details and preference in a key-value store. All the information is accessible via user key, and key-value stores lend themselves to fast reads and writes.
- Real-time recommendations and advertising are often powered by key-value stores because the stores can quickly access and present new recommendations or ads as a web visitor moves throughout a site.
- On the technical side, key-value stores are commonly used for in-memory data caching to speed up applications by minimizing reads and writes to slower disk-based systems. Hazelcast is an example of a technology that provides an in-memory key-value store for fast data retrieval.

### Distributed key-value store

A distributed key-value store builds on the advantages and use cases described above by providing them at scale. A distributed key-value store is built to run on multiple computers working together, and thus allows you to work with larger data sets because more servers with more memory now hold the data.

## DATA SOURCES IN BUSINESS

A data source is **a** place where information is obtained. The source can be a database, a flat file, an XML file, or any other format that a system can read. The input is recorded as a collection of records that contain information used in the business process.

- **Define data needs :** Determine what data is needed, why it's needed, and how it will be used. This helps to avoid collecting irrelevant data and ensure that data collection is aligned with business goals.
- **Ensure data quality:** Effective data sourcing can significantly impact the quality of analysis and insights.
- **Establish a governance framework:** Robust data governance can help to ensure data quality, compliance, and availability.
- **Prioritize data security:** Data security is an important consideration when collecting data.
- **Use analytics:** Analytics can provide insights into business performance and help with decision- making.
- **Use data integration platforms :** Data integration platforms can help to streamline the data sourcing process and improve efficiency.
- **Align with business objectives :** Data initiatives should address specific business needs to generate real value.

## OBSERVER THE DATA RANGES

The range of data is the difference between the highest and lowest values in a set of data. It's a measure of the spread of the data, but it doesn't indicate how the data is distributed.

To calculate the range of data, subtract the lowest value from the highest value. For example, if the data set is {2, 5, 8, 10, 3}, the range is 10 - 2 = 8.

Big data is a large amount of data that can be analyzed and understood to provide actionable insights. Big data can come from a variety of sources, such as ocean observatories, which collect data to help answer research questions.

Here are some ways to analyze big data:

- **Artificial intelligence (AI) :** AI can help visualize and analyze big data to provide insights in real time.
- **Machine learning :** Machine learning can help analyze big data to provide insights in real time.
- **Modern database technologies :** Modern database technologies can help analyze big data to provide insights in real time.

## DEFINITION OF OUTLIER DETECTION

Outliers are generally defined as samples that are exceptionally far from the mainstream of data. There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise.

An outlier may also be explained as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution.

Therefore, Outlier Detection may be defined as the process of detecting and subsequently excluding outliers from a given set of data. There are no standardized Outlier identification methods as these are largely dependent upon the data set. Outlier Detection as a branch of data mining has many applications in data stream analysis.

This paper focuses on the problems of detecting outlier over data stream and the specific techniques used for detecting outlier over streaming data in data mining. We would also focus on outlier detection methods and recent researches on outlier analysis.

Our discussion will also cover areas of standard applications of Outlier Detection, such as Fraud detection, public health, and sports and touch upon the various approaches like Proximity-based approaches and Angle-based approaches.

### Outlier Detection Techniques

For outlier identification in a dataset, it is very important to keep in mind the context and finding answer the very basic and pertinent question: “Why do I want to detect outliers?” The context will explain the meaning of your findings.

Remember two important questions about your dataset in times of outlier identification:

- Which and how many features am I considering for outlier detection? (univariate / multivariate)
- Can I assume a distribution(s) of values for my selected features? (parametric / non-parametric)

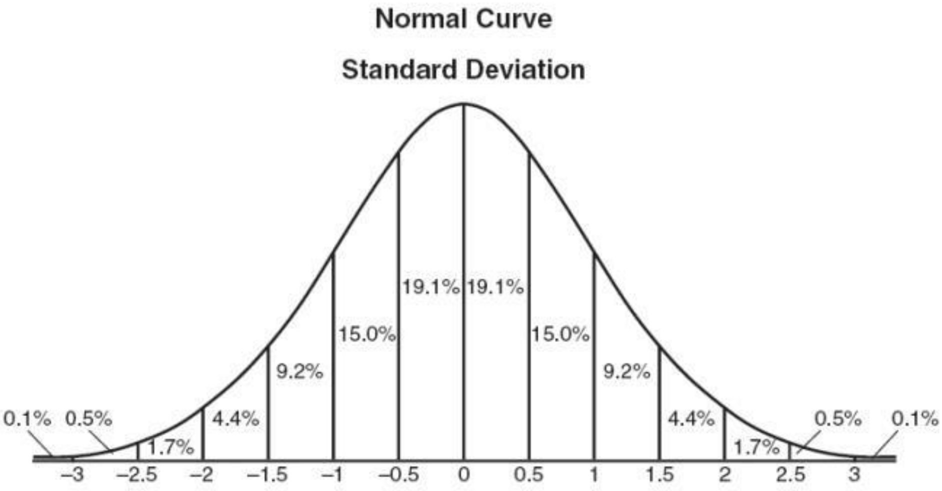
### Four Outlier Detection Techniques

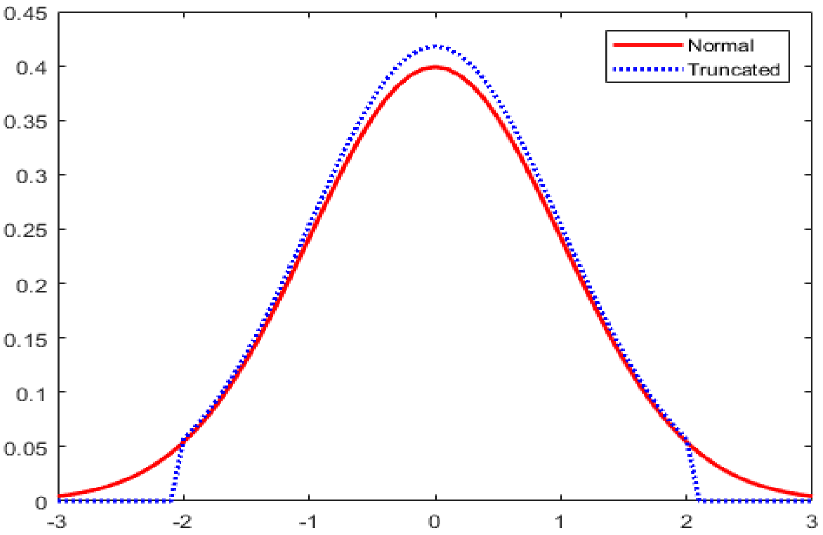
**1. Numeric Outlier :** Numeric Outlier is the simplest, nonparametric outlier detection technique in a one-dimensional feature space. The outliers are calculated by means of the IQR (InterQuartile Range). For example, the first and the third quartile (Q1, Q3) are calculated. An outlier is then a data point xi that lies outside the inter quartile range. Using the inter quartile multiplier value k=1.5, the range limits are the typical upper and lower whiskers of a box plot. This technique can easily be implemented in KNIME Analytics Platform using the Numeric Outliers node.

**2. Z-Score :** Z-score technique assumes a Gaussian distribution of the data. The outliers are the data points that are in the tails of the distribution and therefore far from the mean.

$$z = \frac{x - \mu}{\sigma}$$

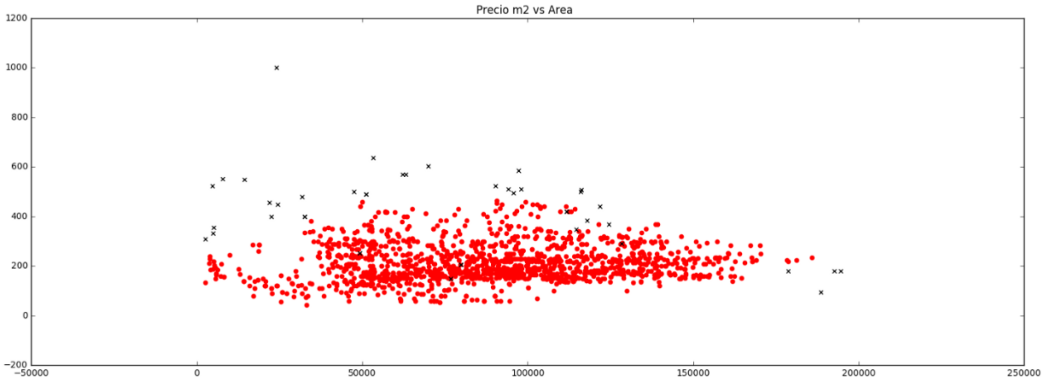
After making the appropriate transformations to the selected feature space of the dataset, the z-score of any data point can be calculated with the following expression:



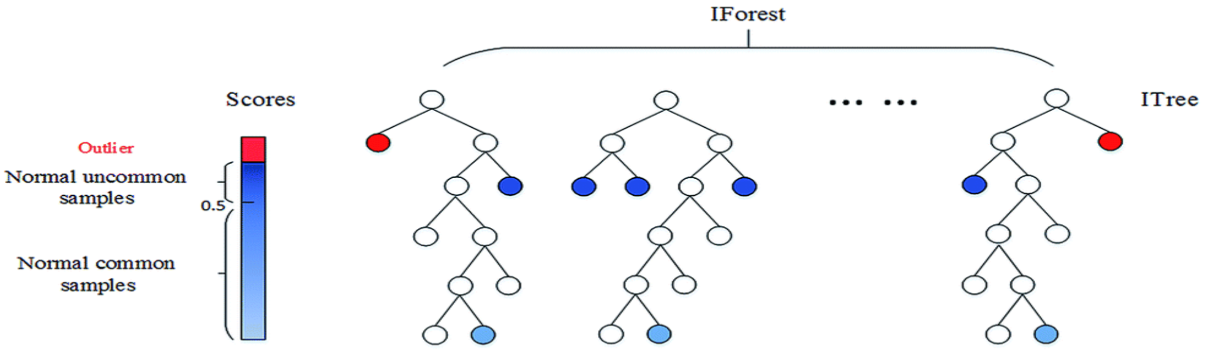


When computing the z-score for each sample on the data set a threshold must be specified. Some good ‘thumb-rule’ thresholds can be 2.5, 3, 3.5 or more standard deviations.

**3. DBSCAN:** This Outlier Detection technique is based on the DBSCAN clustering method. DBSCAN is a nonparametric, density-based outlier detection method in a one or multi-dimensional feature space. Here, all data points are defined either as Core Points, Border Points or Noise Points.  To put it in simpler words, Core Points are data points that have at least MinPts neighbouring data points within a distance ε.



**4. Isolation Forest :** This nonparametric method is ideal for large datasets in a one or multi-dimensional feature space. The isolation number is of paramount importance in this Outlier Detection technique. The isolation number is the number of splits needed to isolate a data point. This number of splits is ascertained by following these steps:



It requires fewer splits to isolate an outlier than it does to isolate a non-outlier, i.e. an outlier has a lower isolation number in comparison to a non-outlier point. A data point is therefore defined as an outlier if its isolation number is lower than the threshold.

## OUTLIER DETECTION METHODS

### Models for Outlier Detection Analysis

There are several approaches to detecting Outliers. Outlier detection models may be classified into the following groups:

**1. Extreme Value Analysis :** Extreme Value Analysis is the most basic form of outlier detection and great for 1-dimension data. In this Outlier analysis approach, it is assumed that values which are too large or too small are outliers. Z-test and Student's t-test are classic examples. These are good heuristics for initial analysis of data but they do not have much value in multivariate settings. Extreme Value Analysis is largely used as final step for interpreting outputs of other outlier detection methods.

**2. Linear Models**: In this approach, the data is modelled into a lower-dimensional sub-space with the use of linear correlations. Then the distance of each data point to a plane that fits the sub-space is being calculated. This distance is used to find outliers. PCA (Principal Component Analysis) is an example of linear models for anomaly detection.

**3. Probabilistic and Statistical Models:** In this approach, Probabilistic and Statistical Models assume specific distributions for data. They make use of the expectation-maximization (EM) methods to estimate the parameters of the model. Finally, they calculate the probability of membership of each data point to calculated distribution. The points with a low probability of membership are marked as outliers.

**4. Proximity-based Models**: In this method, outliers are modelled as points isolated from the rest of the observations. Cluster analysis, density-based analysis, and nearest neighbourhood are the principal approaches of this kind.

**5. Information-Theoretic Models:** In this method, the outliers increase the minimum code length to describe a data set.

### Uses of Outlier Detection Methods

### 1. High Dimensional Outlier Detection :

Real-world data sets are mostly very high dimensional. In many applications, data sets may contain thousands of features. The traditional outlier detection approaches such as PCA and LOF will not be effective. High Contrast Subspaces for Density-Based Outlier Ranking (HiCS) method explained in this paper as an effective method to find outliers in high dimensional data sets.

### 2. Proximity Method

Once you have explored the simpler extreme value methods, consider moving onto proximity-based methods.

- Use clustering methods to identify the natural clusters in the data (such as the k-means algorithm).
- Identify and mark the cluster centroids.
- Identify data instances that are a fixed distance or percentage distance from cluster centroids.
- Filter out the outliers candidate from training dataset and assess the model’s performance.

### 3. Projection Method

Projection methods are relatively simple to apply and quickly highlight extraneous values.

- Use projection methods to summarize your data to two dimensions (such as PCA, SOM or Sammon's mapping).
- Visualize the mapping and identify outliers by hand.
- Use proximity measures from projected values or codebook vectors to identify outliers.
- Filter out the outliers candidate from training dataset and assess the model's performance.

### Outlier Detection Applications

- ❖ Outlier Detection has been mostly studied in the context of multiple application domains. Many algorithms have been proposed for outlier detection in high-dimensional data, uncertain data, stream data, and time-series data.
- ❖ By its inherent nature, network data provides very different challenges that need to be addressed in a special way. Network data humongous in volume, contains nodes of different types, rich nodes with associated attribute data, noisy attribute data, noisy link data, and is dynamically evolving in multiple ways.
- ❖ The concept of Outlier Detection from a networks perspective opens up a whole new dimension of outlier detection research. The detected outliers, which cannot be found by traditional outlier detection techniques, provide new insights into the application area.
- ❖ The algorithms can be applied to several areas, including social network analysis, cyber-security, distributed systems, health care, and bio-informatics. Since both the amount of data as well as the linkage increase in a variety of domains, such network-based techniques will find more applications and more opportunities for research for various settings.

### Outlier Detection and Data Mining

Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from the data mining community.

Data mining involves algorithms of data mining, machine learning, statistics, and natural language processing, attempts to extract high quality, useful information from unstructured formats. The recent years have seen a tremendous increase in the adoption of text mining for business applications.

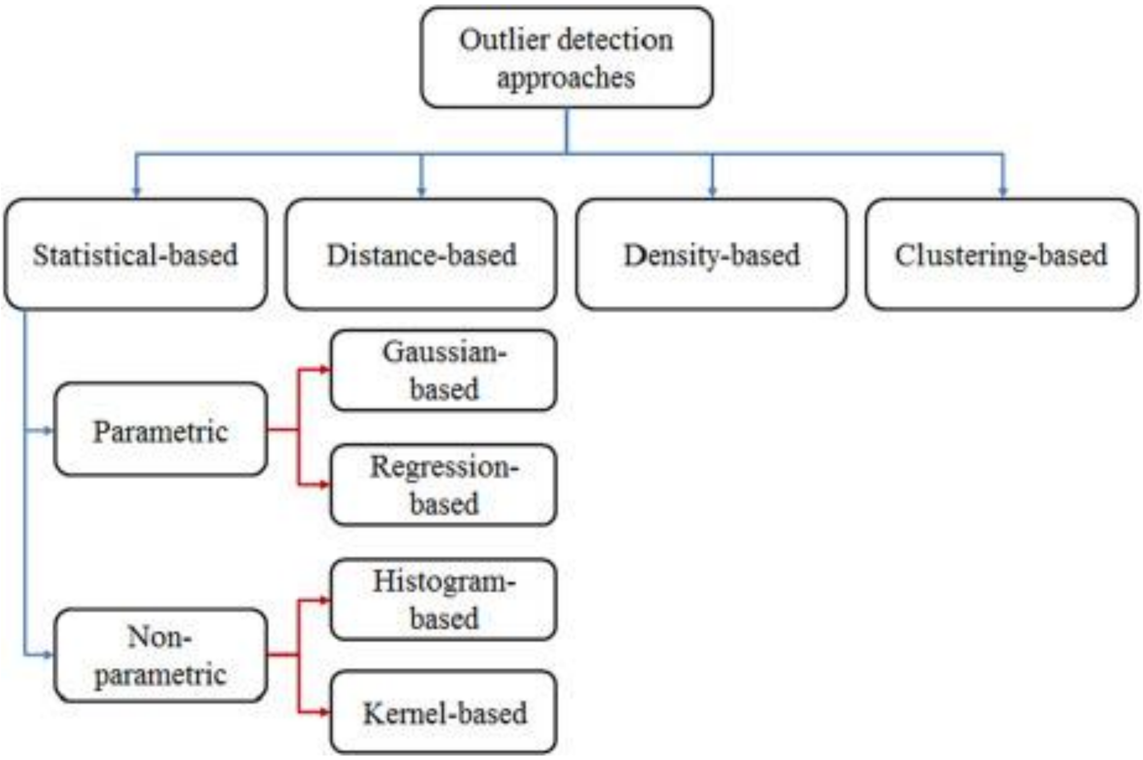
**Wrapping Up**

You may also go for a combined course in Text Mining and Data Analytics, to learn about the major techniques for mining and analyzing text data to discover interesting patterns, extract useful knowledge, and support decision making, with an emphasis on statistical approaches.

You will also need to learn detailed analysis of text data. Prior knowledge of statistical approaches helps in robust analysis of text data for pattern finding and knowledge discovery.

You will love to experiment with explorative data analysis for Hierarchical Clustering, Corpus Viewer, Image Viewer, and Geo Map. You can also learn to interactively explore the dendrogram, read the documents from selected clusters, observe the corresponding images, and locate them on a map.

## OUTLIER ELIMINATION IN BIG DATA ANALYTICS



Detecting and eliminating outliers is the act of removing individual data vectors from a larger set of data, which is critical in nearly any quantitative field such as ML. The quality of the data is just as crucial as the quality of the classification model in machine learning and any quantitative discipline.

Outliers are nothing but data points that differ significantly from other observations. They are the points that lie outside the overall distribution of the dataset. Outliers, if not treated, can cause serious problems in statistical analyses.

When you decide to remove outliers, document the excluded data points and explain your reasoning. You must be able to attribute a specific cause for removing outliers. Another approach is to perform the analysis with and without these observations and discuss the differences.