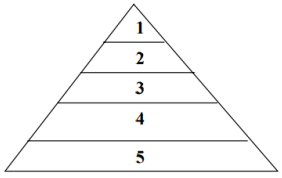
Aspect	Descriptive Data Mining	Predictive Data Mining
Goal	Understand and summarize	Make predictions
Objective	Describe existing data	Forecast future outcomes
Focus	Past and current patterns	Future trends and patterns
Data Used	Historical data	Historical and current data
Purpose	Gain insights	Inform decision-making
Techniques	Clustering, summarization,	Regression, classification,
	association rules,	time series analysis,
	visualization	machine learning algorithms
Examples	Market basket analysis,	Sales forecasting,
	customer segmentation	churn prediction
Output	Patterns, trends, summaries	Predictions, probabilities
Evaluation	Accuracy is not a primary	Accuracy, precision,
	concern	recall, F1-score, etc.
Usage	Exploratory data analysis,	Decision support, risk
	report generation,	management, recommendation
	data understanding	systems, marketing

Data Generalization & Summarization

Data generalization (DWDM)

» A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.



Conceptual levels

Approaches:

- Data cube approach(OLAP approach)
- Attribute-oriented induction approach

1. Data cube approach:

- It is also known as OLAP approach.
- It is an efficient approach as it is helpful to make the past selling graph.
- In this approach, computation and results are stored in the Data cube.
- It uses Roll-up and Drill-down operations on a data cube.
- These operations typically involve aggregate functions, such as count(), sum(), average(), and max().
- These materialized views can then be used for decision support, knowledge discovery, and many other applications.

2.Attribute-Oriented Induction:

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
  - Collect the task-relevant data(*initial relation*) using a relational database query
  - Perform generalization by attribute removal or attribute generalization.
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
  - Interactive presentation with users.

Basic Principles of Attribute-Oriented Induction

- **Data focusing**: task-relevant data, including dimensions, and the result is the *initial relation*.
- **Attribute-removal**: remove attribute *A* if there is a large set of distinct values for *A* but (1) there is no generalization operator on *A*, or (2) *A*'s higher level concepts are expressed in terms of other attributes.
- **Attribute-generalization**: If there is a large set of distinct values for *A*, and there exists a set of generalization operators on *A*, then select an operator and generalize *A*.
- **Attribute-threshold control**: typical 2-8, specified/default.
- **Generalized relation threshold control**: control the final relation/rule size.

Analytical Characterization

- Analytical characterization is used to help and identifying the weakly relevant, or irrelevant attributes.
- We can exclude these unwanted irrelevant attributes when we preparing our data for the mining.

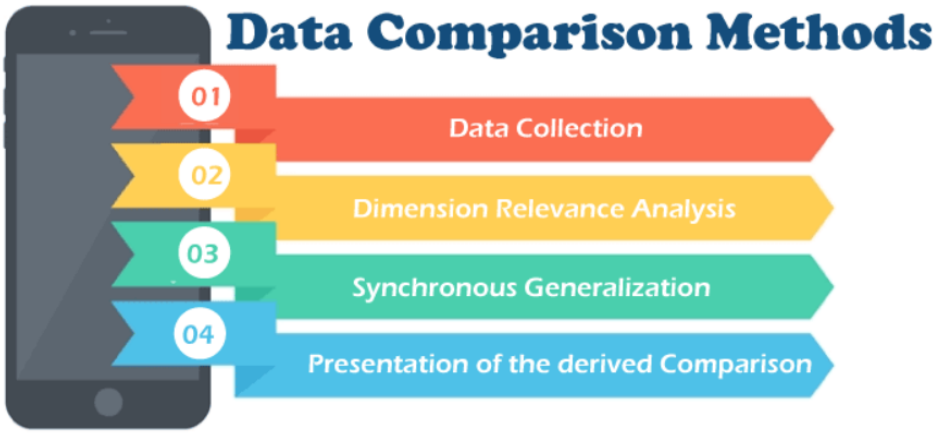
EXAMPLE: We want to characterize the class or in other words, we can say that suppose we want to compare the classes. Now the confusing question is that What if we are not sure which attribute we should include for the class characterization or class comparison? If we specify too many attributes, then these attributes can be a solid reason to slow down the overall process of data mining.
We can solve this problem with the help of analytical characterization.

Why Analytical Characterization?

Analytical Characterization is a very important activity in data mining due to the following reasons;

- Due to the limitation of the OLAP tool about handling the complex objects.
- Due to the lack of an automated generalization,
- we must explicitly tell the system which attributes are irrelevant and must be removed, and similarly,
- we must explicitly tell the system which attributes are relevant and must be included in the class characterization.

Attribute Relevance Analysis



1. Data Collection:

- It is collecting the data for both the target class and the contrasting class by query processing.

2. **Dimension relevance analysis:**:

- This step identifies a set of dimensions and attributes on which the selected relevance measure is to be applied.
- The relation obtained by such an application of Attribute Oriented Induction is called the candidate relation of the mining task.

3. S**ynchronous Generalization:**:

- We evaluate each attribute in the candidate relation using the selected relevance analysis measure.
- This step results in an initial target class working relation and initial contrasting class working relation.

4. **Presentation of the derived comparison:**:

- We need to perform the Attribute Oriented Induction process using a less conservative set of attribute generalization thresholds.

If descriptive mining is

- Class characterization, only ITCWR is included.
- Class Comparison both ITCWR and ICCWR are included.

Relevance Measures

Some of the methods of quantitative relevance measure are:

- Information Gain (ID3)
- Gain Ratio (C4.5)
- Gini Index
- Chi^2 contingency table statistics
- Uncertainty Coefficient

What are statistical measures in large databases?

There are several descriptive statistical measures to mine in large databases in data mining i.e used for knowledge discovery in large databases.

These measures are listed down below.

- **Measuring Central Tendency.**
- **Measuring the Dispersion of Data.**
- **Boxplot Analysis.**
- **Visualization of Boxplot Dispersion.**
- **Histogram Analysis.**
- **Quantile Plot.**
- **Quantile-Quantile Plot.**
- **Scatter Plot.**
- **Loess Curve.**

Measuring The Central Tendency

Mean:

- It is the Arithmetic average of the given data.
- For Weighted mean, we use this formula, x=(sum(wi*xi)/sum(wi).

Median:

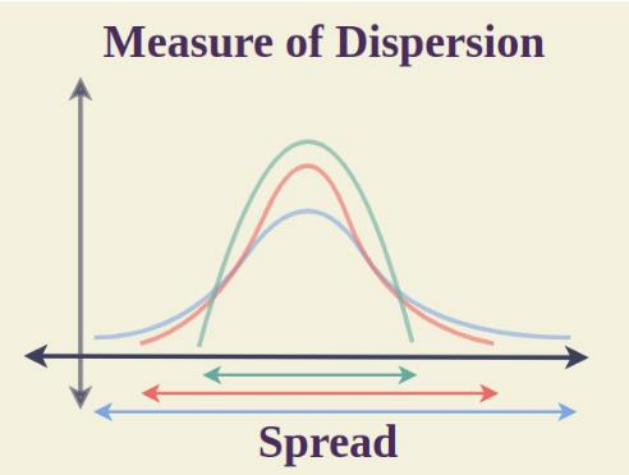
- It is a holistic measure of data.
- Given in order, It is nothing but the middlemost value of the dispersed data.
- If there are odd no values then the middle value will be the median.
- If there are even no values then the median is the average of two middle values.
- It can also be estimated by using,

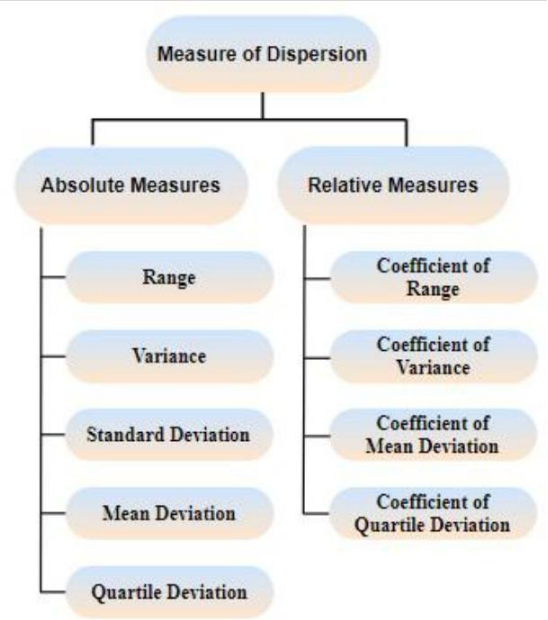
Mode:

- It is nothing but the value that occurs most frequently in the data.
- If there is only one mode in the data then it is a unimodal data.
- If there are two modes in the data then it is bimodal data.
- If there are three modes in the data then it is trimodal data.
- The empirical formula of mode is, **mode=3median-2mean**

Measuring The Dispersion Of Data

Quartiles, Outliers, and Boxplots:





The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
2. **Variance:** average of the squared differences between each data point and the mean. Variance $(\sigma^2) = \Sigma(X-\mu)^2/N$
3. **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\surd\sigma$.
4. **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.

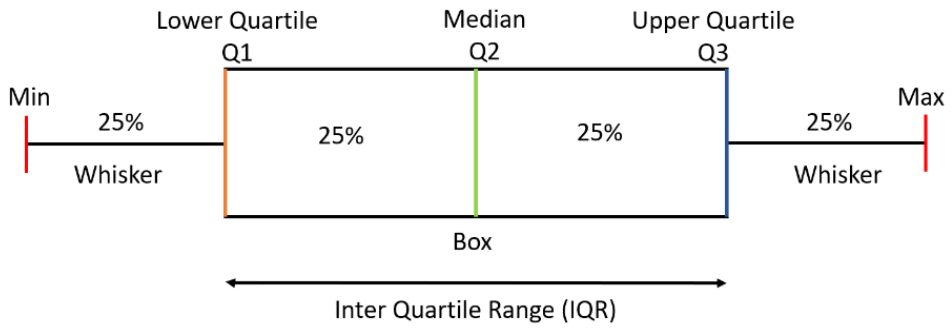
5. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

- **Quartiles**: Those are nothing but the 1/4th of the data, Q1 (25th percentile), Q3 (75th percentile).
- **Inter-quartile range**: It is the differences between the 75th and 25th quartile (IQR = Q3 – Q1).
- **Five number summary**: It describes five values -> **min, Q1, M, Q3, max**.
- **Boxplot**: Ends of the box are the quartiles, the median is marked, whiskers(two lines outside the box extend to Minimum and Maximum) and plot outlier individually.

- **Outlier**: It is usually, a value higher/lower than 1.5 x IQR.

A box plot gives a five-number summary of a set of data which is-

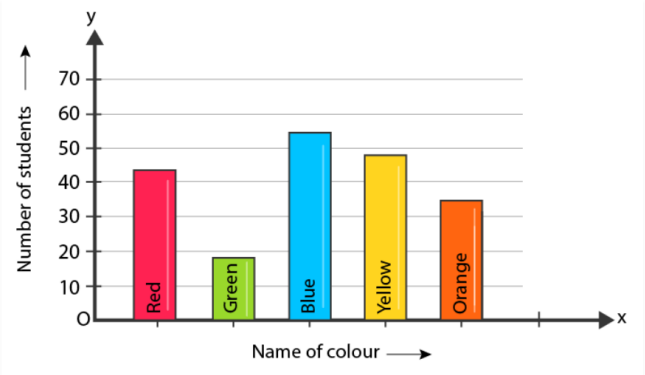
- **Minimum** – It is the minimum value in the dataset excluding the outliers
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.



Statistical Graphs (bar graph, pie graph, line graph, etc.)

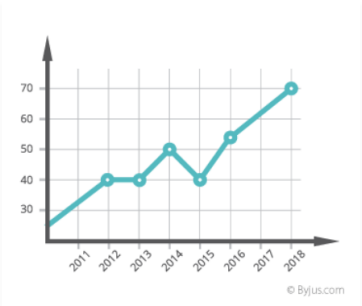
Bar Graph

Bar graphs are the pictorial representation of grouped data in vertical or horizontal rectangular bars, where the length of bars is proportional to the measure of data. The chart's horizontal axis represents categorical data, whereas the chart's vertical axis defines discrete data.



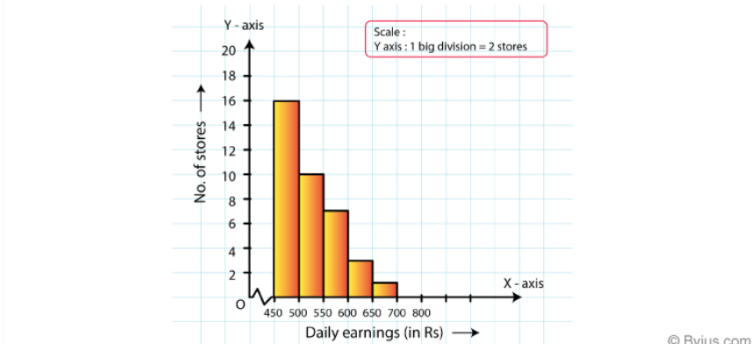
Line Graph

A graph that utilizes points and lines to represent change over time is defined as a line graph. In other words, it is a chart that shows a line joining several points or a line that shows the relation between the points. The diagram depicts quantitative data between two changing variables with a straight line or curve that joins a series of successive data points. Linear charts compare these two variables on a vertical and horizontal axis.



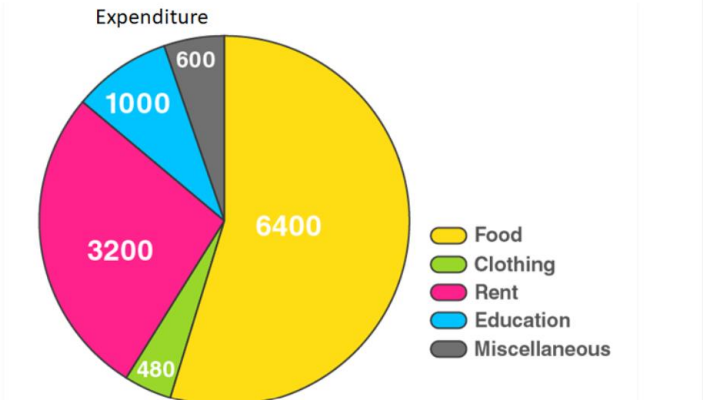
Histogram

A histogram chart displays the frequency of discrete and continuous data in a dataset using connected rectangular bars. Here, the number of observations that fall into a predefined class interval represented by a rectangular bar.



Pie Chart

A pie chart used to represent the numerical proportions of a dataset. This graph involves dividing a circle into various sectors, where each sector represents the proportion of a particular element as a whole. This is also called a circle chart or circle graph.



what is association rule?

- It is simple **If/Then** statements that help discover relationships between seemingly **independent relational** databases or other data repositories.
- It is suitable for **non-numeric**, **categorical data** and requires just a little bit more than simple counting.
- It is a procedure which aims to observe frequently **occurring patterns**, **correlations**, or **associations** from datasets found in various kinds of databases such as **relational databases**, **transactional databases**, and other forms of repositories.
- Association rule learning is a type of **unsupervised learning** technique that checks for the **dependency of one data item on another data item**

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Examples of association rules

{Diaper} → {Beer},
{Milk, Bread} → {Diaper,Coke},
{Beer, Bread} → {Milk},

An association rule has 2 parts:

- **an antecedent (if) and**
- **a consequent (then)**

> *"If a customer buys bread, he's 70% likely of buying milk."*

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:

There are several metrics.

- **Support**
- **Confidence**
- **Lift**

Support

- Support measures the frequency or occurrence of a particular itemset (a combination of items) in the dataset.
- It is calculated as the number of transactions containing the itemset divided by the total number of transactions.
- High support indicates that the itemset is frequent in the dataset.

$$\text{Supp(X)}=\frac{Freq(X)}{T}$$

Confidence

- Confidence measures the strength of the association between two items (antecedent and consequent) in a rule.
- It is calculated as the support of the itemset containing both the antecedent and consequent divided by the support of the antecedent.
- High confidence indicates that when the antecedent is present, there is a strong likelihood of the consequent being present as well.

$$\text{Confidence}=\frac{Freq(X,Y)}{Freq(X)}$$

Lift

- Lift measures how much more likely the consequent is to be bought when the antecedent is bought compared to when it is bought independently.
- It is calculated as the confidence of the rule divided by the support of the consequent.

- Lift greater than 1 indicates a positive association, meaning that the presence of the antecedent increases the likelihood of the consequent.

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

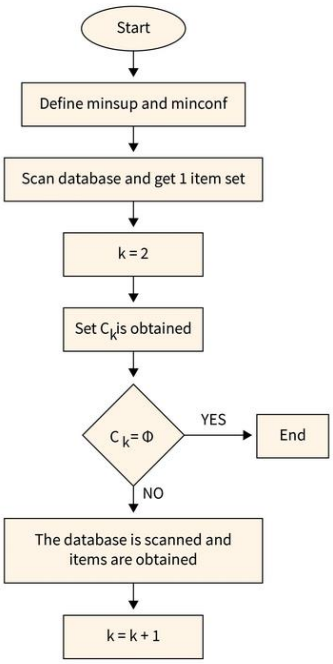
$$\text{Lift}=\frac{Supp(X,Y)}{Supp(X)\times Supp(Y)}$$

Association rule learning can be divided into three types of algorithms.

1. **Apriori**
2. **Eclat**
3. **F-P Growth Algorithm**

Apriori Algorithm?

- Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.
- Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.
- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.



Advantages of Apriori Algorithm

- It is used to calculate large itemsets.
- Simple to understand and apply.

Disadvantages of Apriori Algorithms

- Apriori algorithm is an expensive method to find support since the calculation has to pass through the whole database.
- Sometimes, you need a huge number of candidate rules, so it becomes computationally more expensive

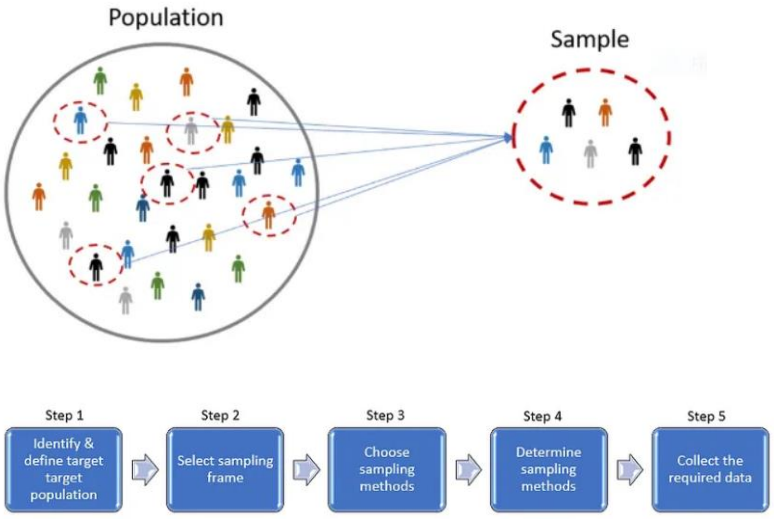
Types Of Association Rules In Data Mining

There are typically four different types of association rules in data mining. They are

- Multi-relational association rules
- Generalized Association rule
- Interval Information Association Rules
- Quantitative Association Rules

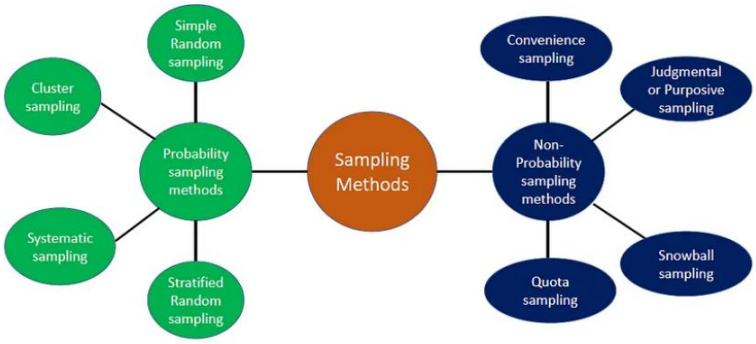
Sampling Techniques— Statistical approach in Machine learning.

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.



Different steps of sampling

Different Types of Sampling Techniques



Different types of sampling methods

1. **Reducing Computational Burden**:
    - Large databases can contain millions or even billions of records. Analyzing the entire dataset may not be feasible due to computational limitations.
    - Sampling allows you to work with a much smaller subset of the data, reducing the computational burden and speeding up analysis.
2. **Estimating Population Statistics**:
    - When you take a random or stratified sample from a large database, you can use the statistics computed from the sample to estimate population parameters.
    - For example, you can estimate the mean, variance, or proportions within the population based on the sample statistics.
3. **Hypothesis Testing and Confidence Intervals**:
    - Sampling is essential for conducting hypothesis tests and calculating confidence intervals.
    - You can test hypotheses and make inferences about the population using statistical tests and confidence intervals based on sample data.
4. **Exploratory Data Analysis (EDA)**:
    - Sampling is useful for initial data exploration and visualization. You can generate summary statistics, histograms, and other plots from the sample to gain insights and identify patterns.
5. **Resource Allocation**:
    - Sampling allows you to allocate resources more efficiently. Instead of processing the entire dataset, you can allocate computational resources, storage, and memory based on the size of the sample.
6. **Real-Time and Streaming Data**:
    - In real-time or streaming data scenarios, where new data is continuously arriving, it may not be feasible to store and analyze all data.
    - Continuous sampling methods can be used to select representative samples from the incoming data stream for real-time analysis.

Common sampling techniques include:

1. **Simple Random Sampling**:
    - In simple random sampling, each record in the database has an equal probability of being selected.
    - It is often implemented using random number generators or hashing functions.
2. **Stratified Sampling**:
    - Stratified sampling divides the dataset into subgroups (strata) based on certain characteristics (e.g., age, region, category).
    - A random sample is then taken from each stratum.

- This ensures that each subgroup is represented in the sample, making it useful for ensuring diversity.

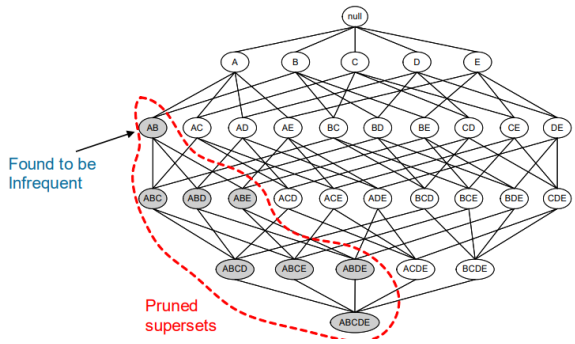
3. **Systematic Sampling**:
   - Systematic sampling selects every nth record from the dataset after a random start.
   - It can be efficient and provide a representative sample if there is no specific order or pattern in the data.
4. **Cluster Sampling**:
   - Cluster sampling divides the dataset into clusters or groups.
   - A random sample of clusters is selected, and then all records within the selected clusters are included in the sample.
   - It is useful when the dataset naturally has a clustered structure.
5. **Reservoir Sampling**:
   - Reservoir sampling is used for sampling from a continuous data stream or large database where the total number of records is unknown in advance.
   - It maintains a fixed-size reservoir and replaces records in the reservoir with a certain probability as new data arrives.
6. **Sequential Sampling**:
   - Sequential sampling is used for dynamic data where new data arrives over time.
   - It continuously updates the sample as new records are added and old records are removed to maintain a representative sample.

Aspect	Data Generalization	Data Summarization
Purpose	Reduce detail and abstraction	Provide a concise overview
Focus	Higher-level, abstract categories	Key data characteristics
Retained Information	Reduced detail	Retains more detail
Privacy Preservation	Strong privacy protection	Limited privacy protection
Techniques	Clustering, binning, ontology	Statistical summaries, visualization, summary statistics
Example	Age groups from specific ages	Mean and standard deviation of numerical data, distribution of categorical data
Use Cases	Data anonymization, hierarchical data, aggregation	Exploratory data analysis, reporting, visualization, dashboard creation

Mining single-dimensional Boolean association rules from transactional transactional databases.

5. Apriori Algorithm: Mining Single-Dimension Boolean AR

- It is used to mine Boolean, single-level, and single-dimension ARs.
- Apriori Principle



- Apriori algorithm:
  - Uses prior knowledge of frequent itemset properties.
  - It is an iterative algorithm known as level-wise search.
  - The search proceeds level-by-level as follows:
    - First determine the set of frequent 1-itemset; L1
    - Second determine the set of frequent 2-itemset using L1: L2
    - Etc.

5.3. Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan → C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan → C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan → L_3

Itemset	sup
{B, C, E}	2

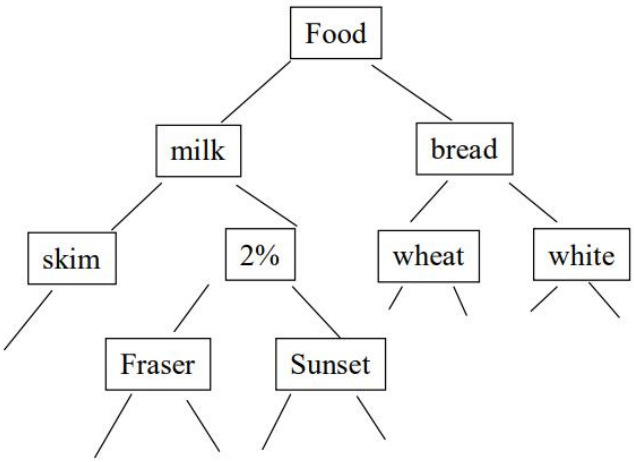
5.5. Challenges

- Multiple scans of transaction database
- Huge number of candidates
- Tedious workload of support counting for candidates
- Improving Apriori:
    - general ideas
    - Reduce passes of transaction database scans
    - Shrink number of candidates
    - Facilitate support counting of candidates
    - Easily parallelized

Mining multilevel association rules from transactional databases

7. Multiple-Level Association Rules

- Items often form hierarchy.
- Items at the lower level are expected to have lower support.
- Rules regarding itemsets at appropriate levels could be quite useful.
- Transaction database can be encoded based on dimensions and levels
- We can explore shared multi-level mining



Multi-level Association

- Uniform Support- the same minimum support for all levels
  - + One minimum support threshold. No need to examine itemsets containing any item whose ancestors do not have minimum support.
  - – Lower level items do not occur as frequently. If support threshold
    - too high ⇒ miss low level associations
    - too low ⇒ generate too many high level associations

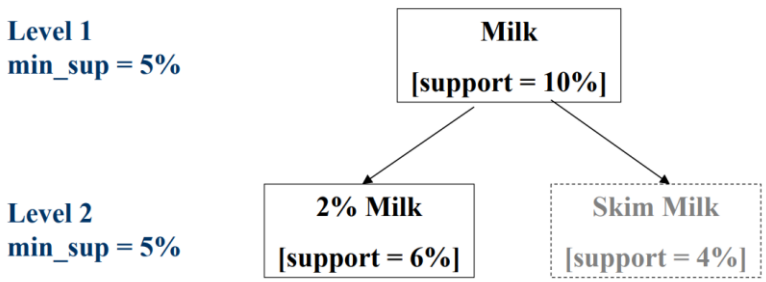
- Two multiple-level mining associations strategies:
  - Uniform Support
  - Reduced support

- Uniform Support: the same minimum support for all levels
  - One minimum support threshold.
  - No need to examine itemsets containing any item whose ancestors do not have minimum support.
  - Drawback:
    - Lower level items do not occur as frequently. If support threshold

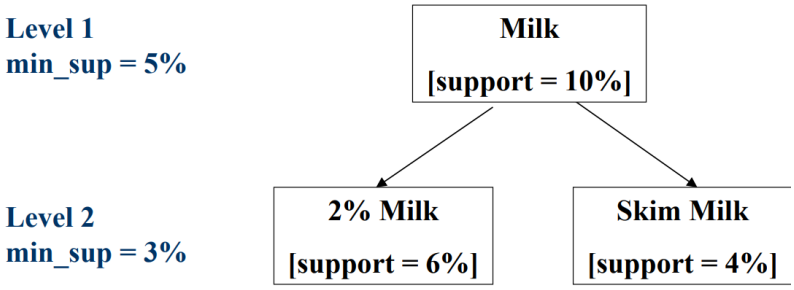
    too high ➔ miss low level associations
    too low ➔ generate too many high level assoc.

- Reduced Support: reduced minimum support at lower levels
  - There are 4 search strategies:
    - Level-by-level independent
    - Level-cross filtering by k-itemset
    - Level-cross filtering by single item
    - Controlled level-cross filtering by single item

Multi-level mining with uniform support



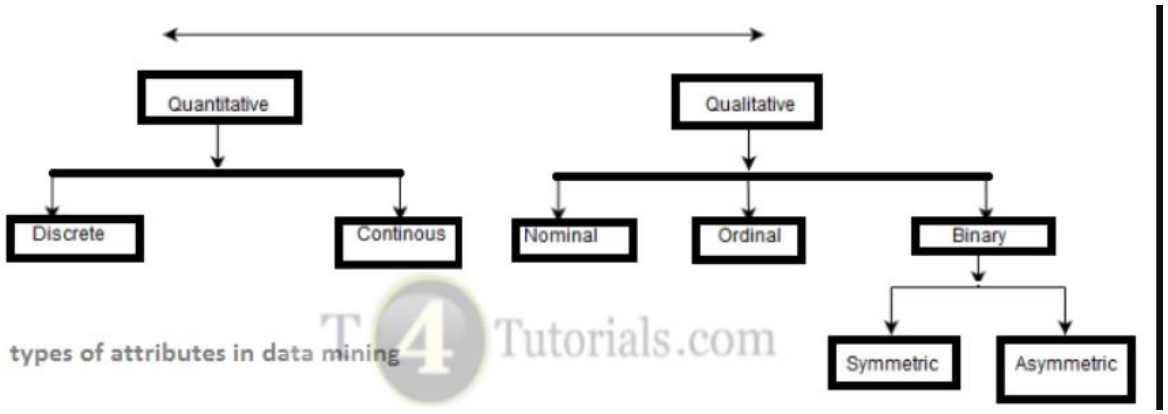
Multi-level mining with reduced support



Multi-level Association

- Reduced Support- reduced minimum support at lower levels
  - There are 4 search strategies:
    - Level-by-level independent
    - Level-cross filtering by k-itemset
    - Level-cross filtering by single item
    - Controlled level-cross filtering by single item

attributes types in data mining



Nominal Attributes

Nominal data is in alphabetical form and not in an integer. Nominal Attributes are Qualitative Attributes.

example, sates and colors are the attribute and New, Pending, Working, Complete, Finish and Black, Brown, White, and Red are the values

Binary Attributes

Binary data have only two values/states. For example, here HIV detected can be only Yes or No.
Binary Attributes are Qualitative Attributes.

Examples of Binary Attributes

Attribute	Value
HIV detected	Yes, No
Result	Pass, Fail

Ordinal Attributes

All Values have a meaningful order.

For example, Grade-A means highest marks,

B means marks are less than A,

C means marks are less than grades A and B, and so on.

Ordinal Attributes are Quantitative Attributes.

Examples of Ordinal Attributes

Attribute	Value
Grade	A, B, C, D, F
BPS- Basic pay scale	16, 17, 18

Discrete Attributes

Certainly, here's a very concise comparison between itemsets and association rules in tabular form:

Aspect	Itemset	Association Rule
Definition	Item collections	Relationship statement
Types	Frequent or infrequent	Derived from frequent itemsets
Purpose	Discover associations	Describe associations
Measures	Support, count	Support, confidence, lift

Aspect	Data Suppression	Data Masking
Purpose	Remove sensitive data	Protect sensitive data
Goal	Total removal	Obfuscation/replacement
Use Cases	De-identification	Testing, sharing, security
Data Removal	Complete removal	Data replaced with fake
Irreversible Transformation	Yes (data loss)	No (reversible)
Example	Removing SSNs	Replacing names with pseudonyms
Data Integrity	Potential impact	Preserves data integrity
Compliance	GDPR, HIPAA, CCPA	Data privacy regulations
Security vs. Privacy	Privacy-focused	Balances security and privacy
Application	Raw data, analysis	Testing, development, sharing
Use in Testing	Rarely used (impact)	Commonly used
Use in Production	Rarely used (data loss)	Commonly used (secure data)