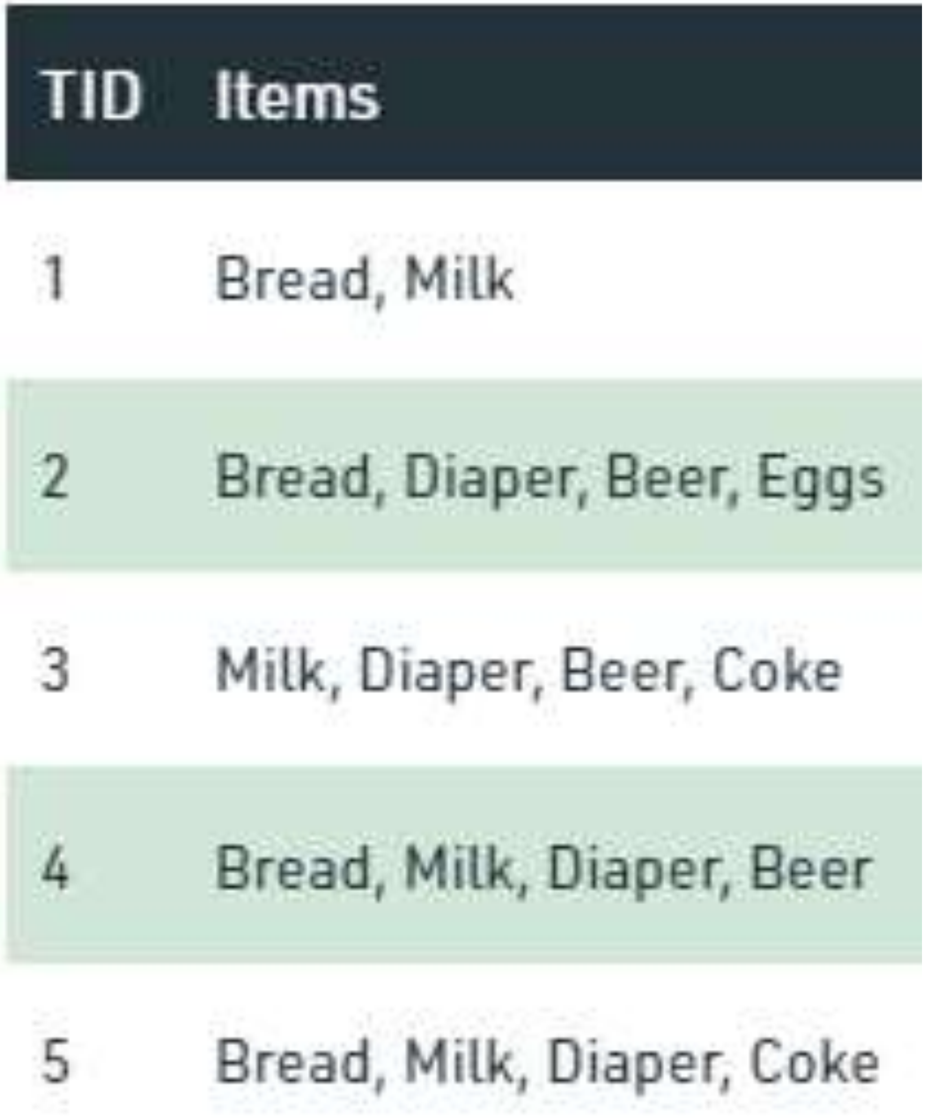
Association Rule Mining

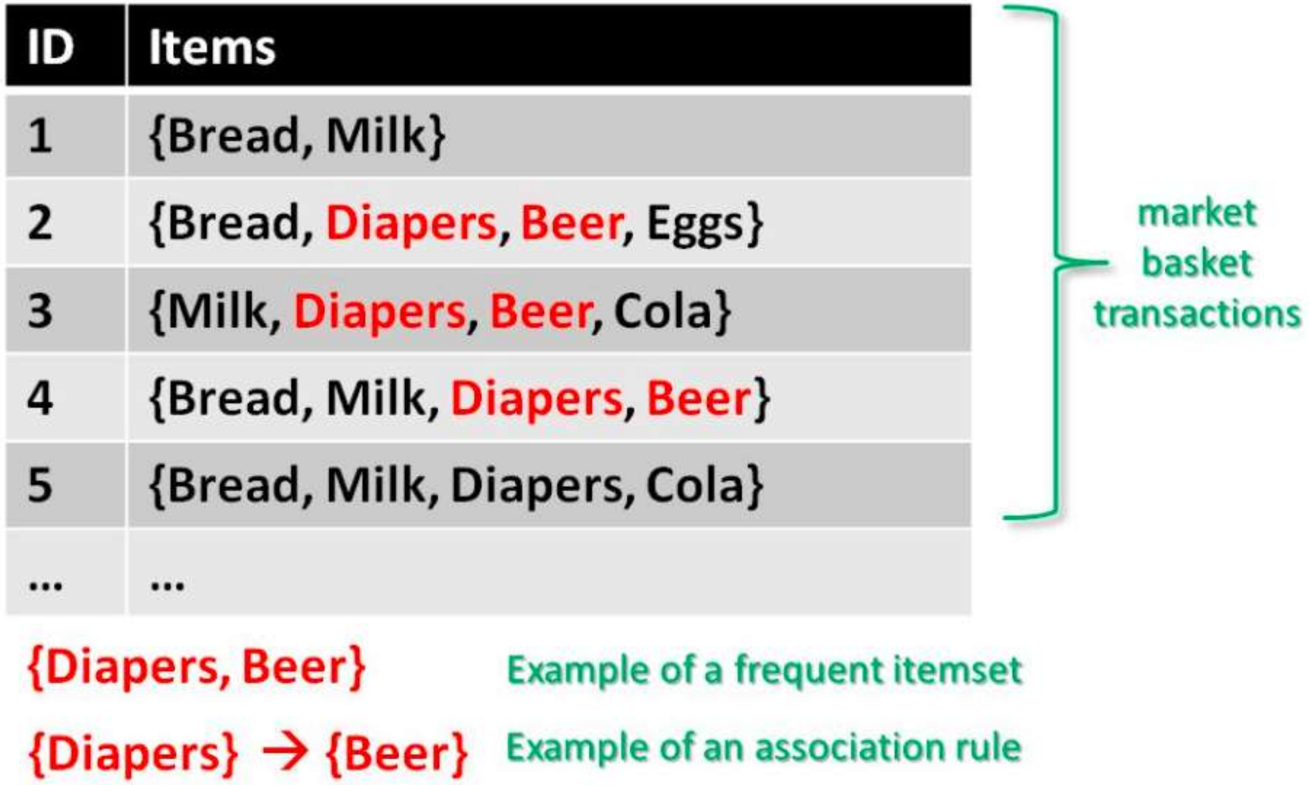
- When we go grocery shopping, we often have a standard list of things to buy. Each shopper has a distinctive list, depending on one’s needs and preferences. A housewife might buy healthy ingredients for a family dinner, while a bachelor might buy beer and chips. Understanding these buying patterns can help to increase sales in several ways. If there is a pair of items, X and Y, that are frequently bought together.
    - Both X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other.
    - Promotional discounts could be applied to just one out of the two items.
    - Advertisements on X could be targeted at buyers who purchase Y.
    - X and Y could be combined into a new product, such as having Y in flavors of X.
- While we may know that certain items are frequently bought together, the question is, how do we uncover these associations?
- Besides increasing sales profits, association rules can also be used in other fields. In medical diagnosis for instance, understanding which symptoms tend to co-morbid can help to improve patient care and medicine prescription

- Association rules analysis is a technique to uncover how items are associated to each other.
- Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.
- Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.
- Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.



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

- Given is a set of transaction data. You can see transactions numbered 1 to 5.
- Each transaction shows items bought in that transaction.
- You can see that *Diaper is bought with Beer* in three transactions.
- Similarly, *Bread is bought with milk* in three transactions making them both frequent item sets.



ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

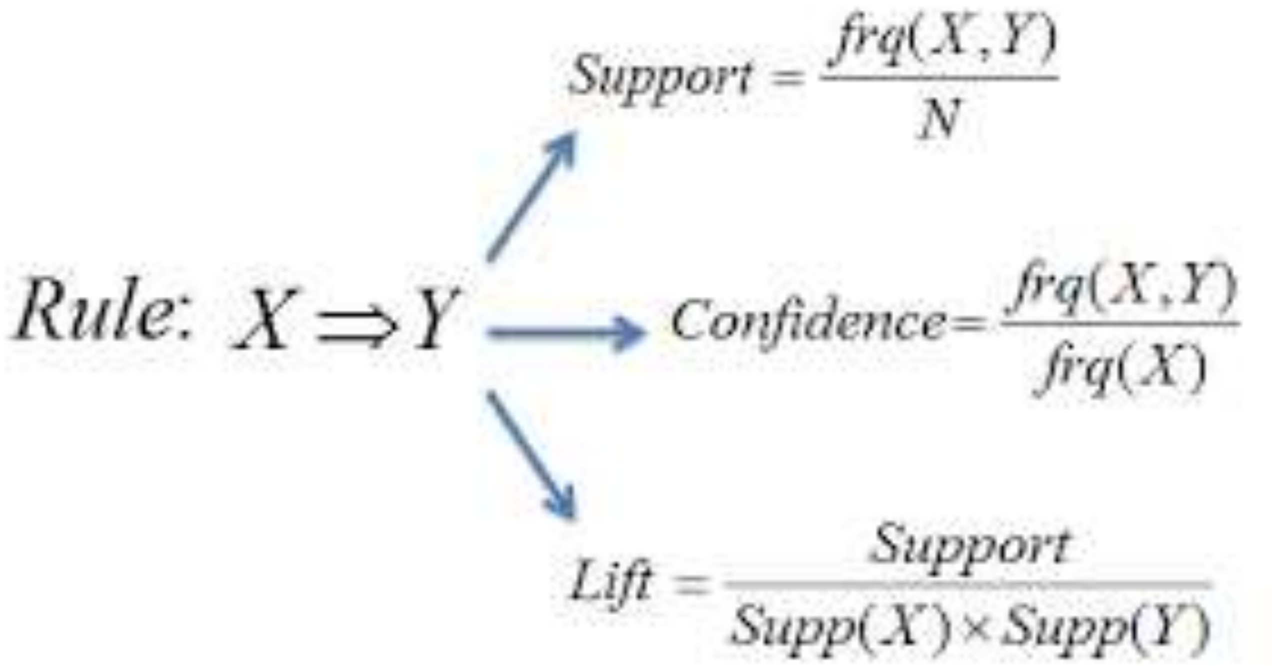
{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

- **Association Rule –** An implication expression of the form X -> Y, where X and Y are any 2 itemsets.
- Example: {Milk, Diaper}->{Beer}
- An association rule has two parts: an antecedent (if) and a consequent (then).
- An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent.
- *“If a customer buys bread, he’s 70% likely of buying milk.”, {Bread}->{Milk}*
- In the above association rule, bread is the antecedent and milk is the consequent.

- Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following three parameters, the important relationships are observed:

1. **Support**: Support indicates how frequently the if/then relationship appears in the database. Or fraction of transactions that contain both X and Y. It measures frequency of association.
2. **Confidence**: Confidence tells about the number of times these relationships have been found to be true. Or It measures how often items in Y appear in transactions that contain X. It measures the strength of association.
3. **Lift:** Lift gives the correlation between A and B in the rule A=>B. Correlation shows how one item-set A effects the item-set B.



$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

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

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

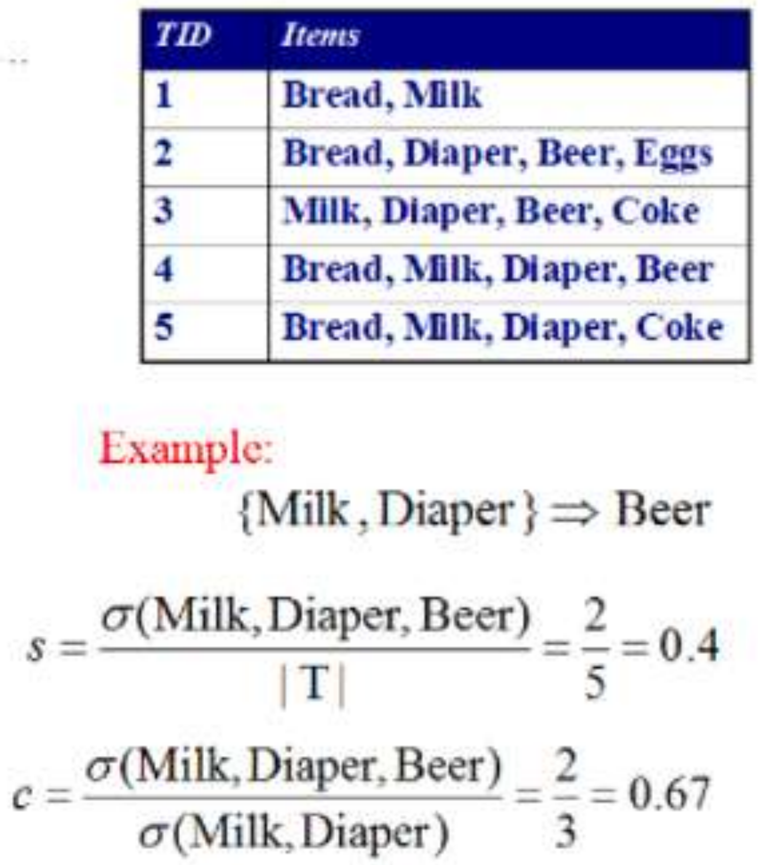
Computer=>Anti−virusSoftware

[Support=20%,confidence=60%]

Above rule says:

1.) 20% transaction show Anti-virus software is bought with purchase of a Computer

2.) 60% of customers who purchase Anti-virus software is bought with purchase of a Computer



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

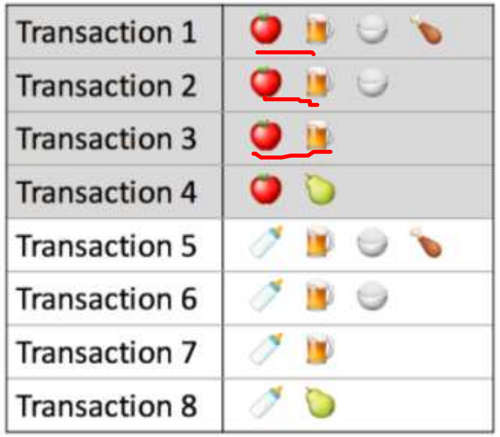
Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

- **Measure 1: Support**. This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In Table 1 below, the support of {apple} is 4 out of 8, or 50%. Itemsets can also contain multiple items. For instance, the support of {apple, beer, rice} is 2 out of 8, or 25%.
- **Measure 2: Confidence**. This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears. In Table 1, the confidence of {apple -> beer} is 3 out of 4, or 75%.



Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

$$\text{Support}\,\{🍎\} = \frac{4}{8}$$

$$\text{Confidence}\,\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\,\{🍎, 🍺\}}{\text{Support}\,\{🍎\}}$$

$$= \frac{3}{4}$$

Applications

1.Market Basket Analysis:

- This is the most typical example of association mining. Data is collected using barcode scanners in most supermarkets. This database, known as the “market basket” database, consists of a large number of records on past transactions. A single record lists all the items bought by a customer in one sale. Knowing which groups are inclined towards which set of items gives these shops the freedom to adjust the store layout and the store catalog to place the optimally concerning one another.

2.Medical Diagnosis:

- Association rules in medical diagnosis can be useful for assisting physicians for curing patients. Diagnosis is not an easy process and has a scope of errors which may result in unreliable end-results. Using relational association rule mining, we can identify the probability of the occurrence of illness concerning various factors and symptoms. Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.

3.Census Data:

- Every government has tons of census data. This data can be used to plan efficient public services(education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products). This application of association rule mining and data mining has immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.

4.Protein Sequence:

- Proteins are sequences made up of twenty types of amino acids. Each protein bears a unique 3D structure which depends on the sequence of these amino acids. A slight change in the sequence can cause a change in structure which might change the functioning of the protein. This dependency of the protein functioning on its amino acid sequence has been a subject of great research.

5. Entertainment

- Services like Netflix and Spotify can use association rules to fuel their content recommendation engines. Machine learning models analyze past user behavior data for frequent patterns, develop association rules and use those rules to recommend content that a user is likely to engage with, or organize content in a way that is likely to put the most interesting content for a given user first.