Advanced Analytical Theory and Methods: Association Rules

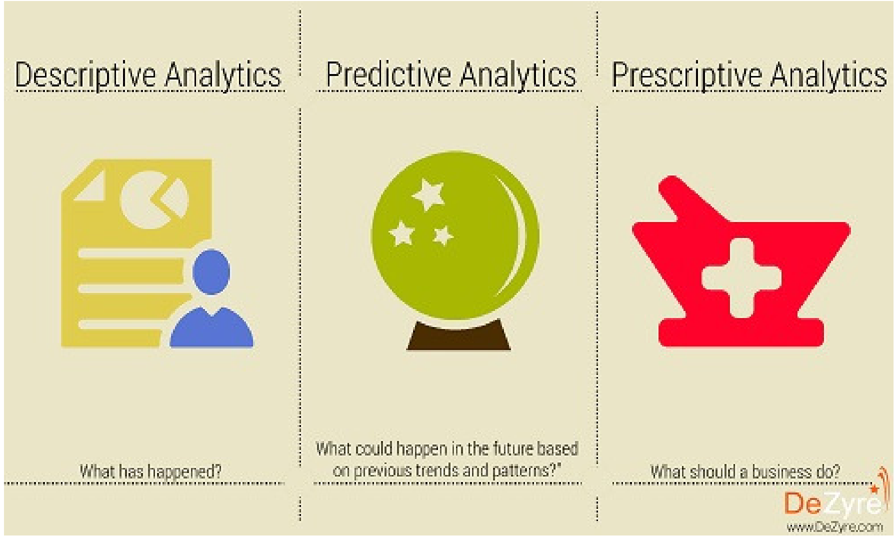
Prof. Dr. Shamim Akhter

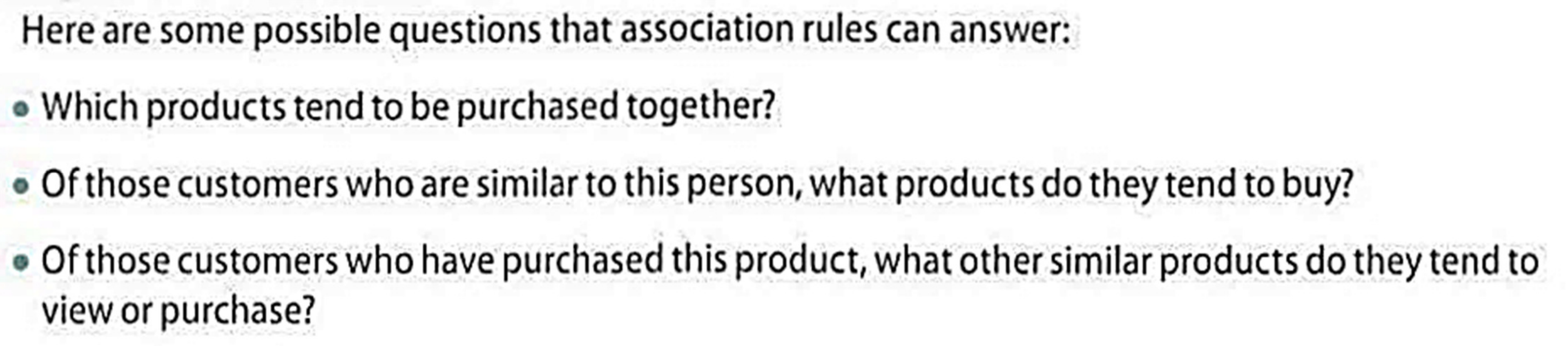
Professor, Dept. of CSE

Ahsanullah University of Science and Technology

Association Rules

- Association rules represent interesting associations and relationships hidden in a large dataset.
- These are unsupervised but descriptive, not predictive learning methods.

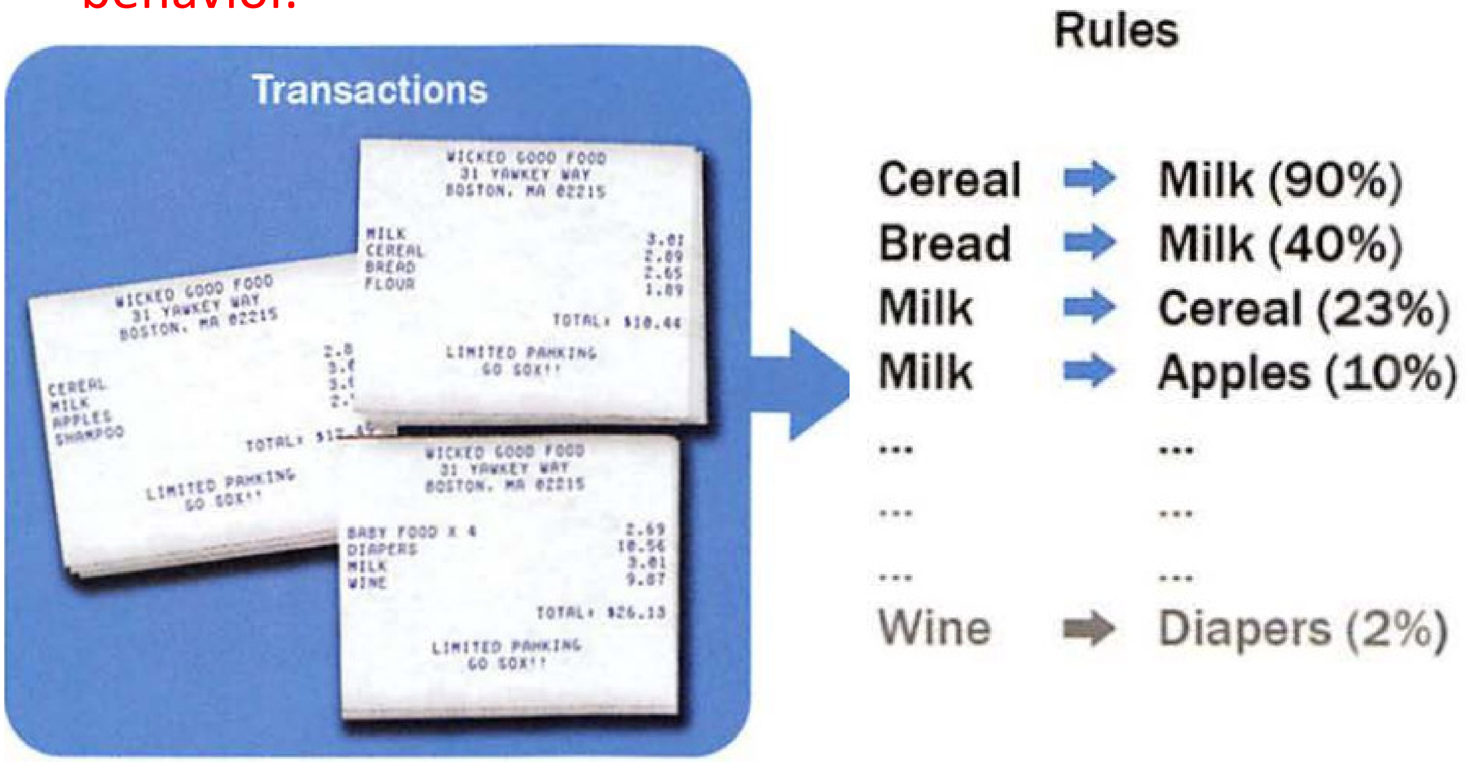



Here are some possible questions that association rules can answer:

- Which products tend to be purchased together?
- Of those customers who are similar to this person, what products do they tend to buy?
- Of those customers who have purchased this product, what other similar products do they tend to view or purchase?

General logic behind association rules

- Given a large collection of transactions, each transaction consists of one or more items.
- Association rules review the items being purchased to see what items are frequently bought together and to discover a list of rules describing the purchasing behavior.

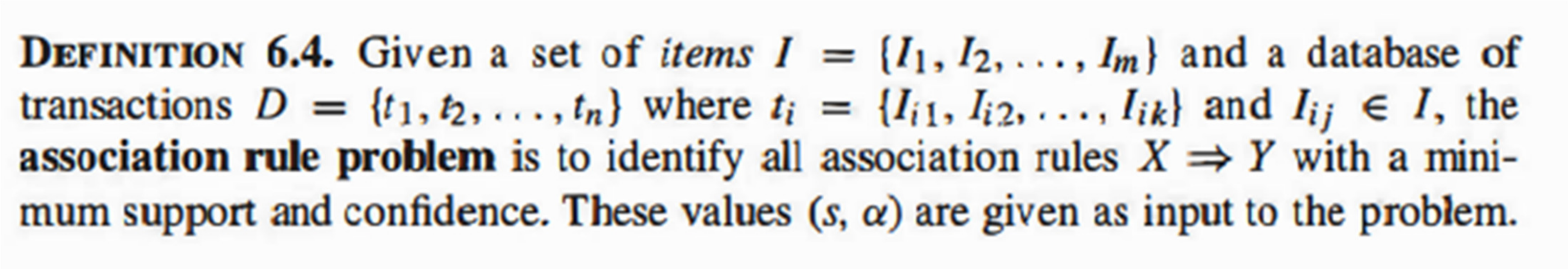


Rules

Cereal	➡	Milk (90%)
Bread	➡	Milk (40%)
Milk	➡	Cereal (23%)
Milk	➡	Apples (10%)
...		...
...		...
...		...
Wine	➡	Diapers (2%)

The first three rules suggest that when cereal is purchased, milk is purchased 90% of the time. When bread is purchased, 40% of the time milk is purchased. When milk is purchased, 23% of the time cereal is purchased.

Each uncovered rule is in the form X->Y, meaning that when item X is observed, item Y is also observed. In this case, the left-hand side (LHS) of the rule is X, and the right-hand side (RHS) of the rule is Y.



Definition 6.4. Given a set of *items* $I = \{I_1, I_2, \ldots, I_m\}$ and a database of transactions $D = \{t_1, t_2, \ldots, t_n\}$ where $t_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the **association rule problem** is to identify all association rules $X \Rightarrow Y$ with a minimum support and confidence. These values (s, α) are given as input to the problem.

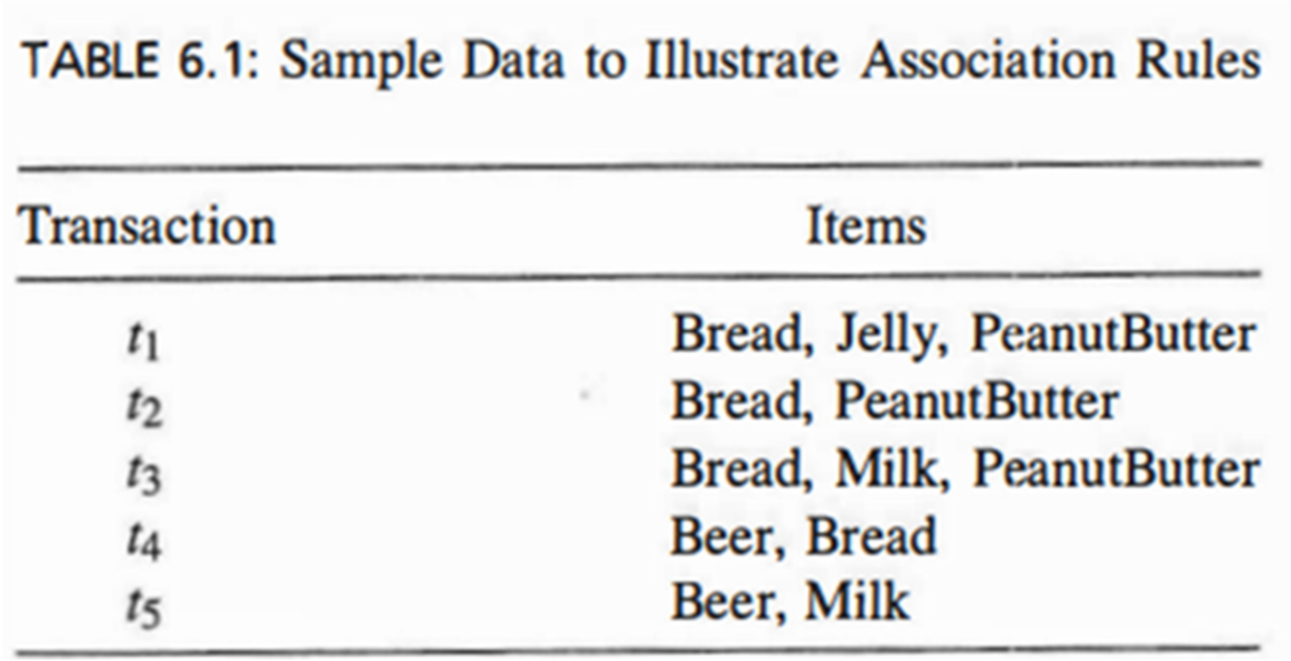
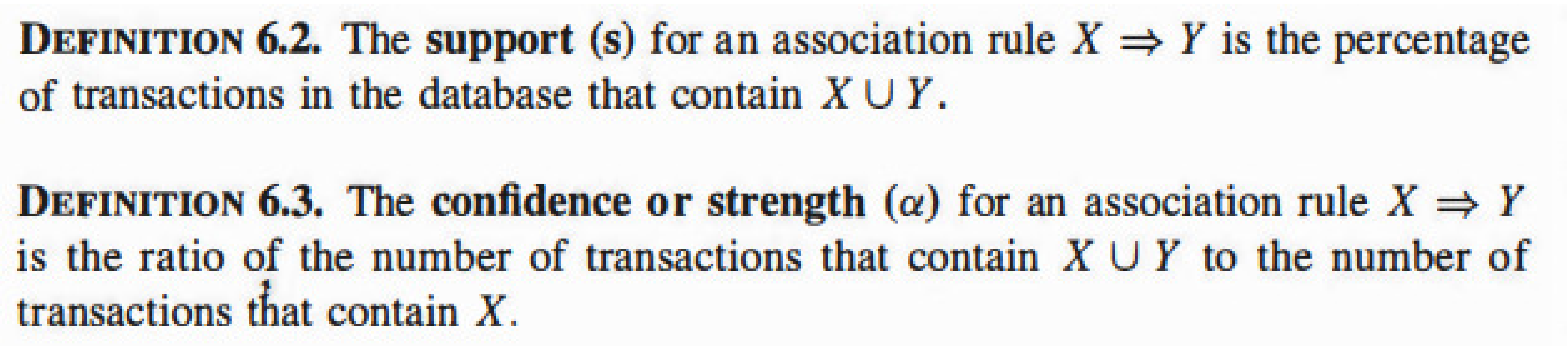


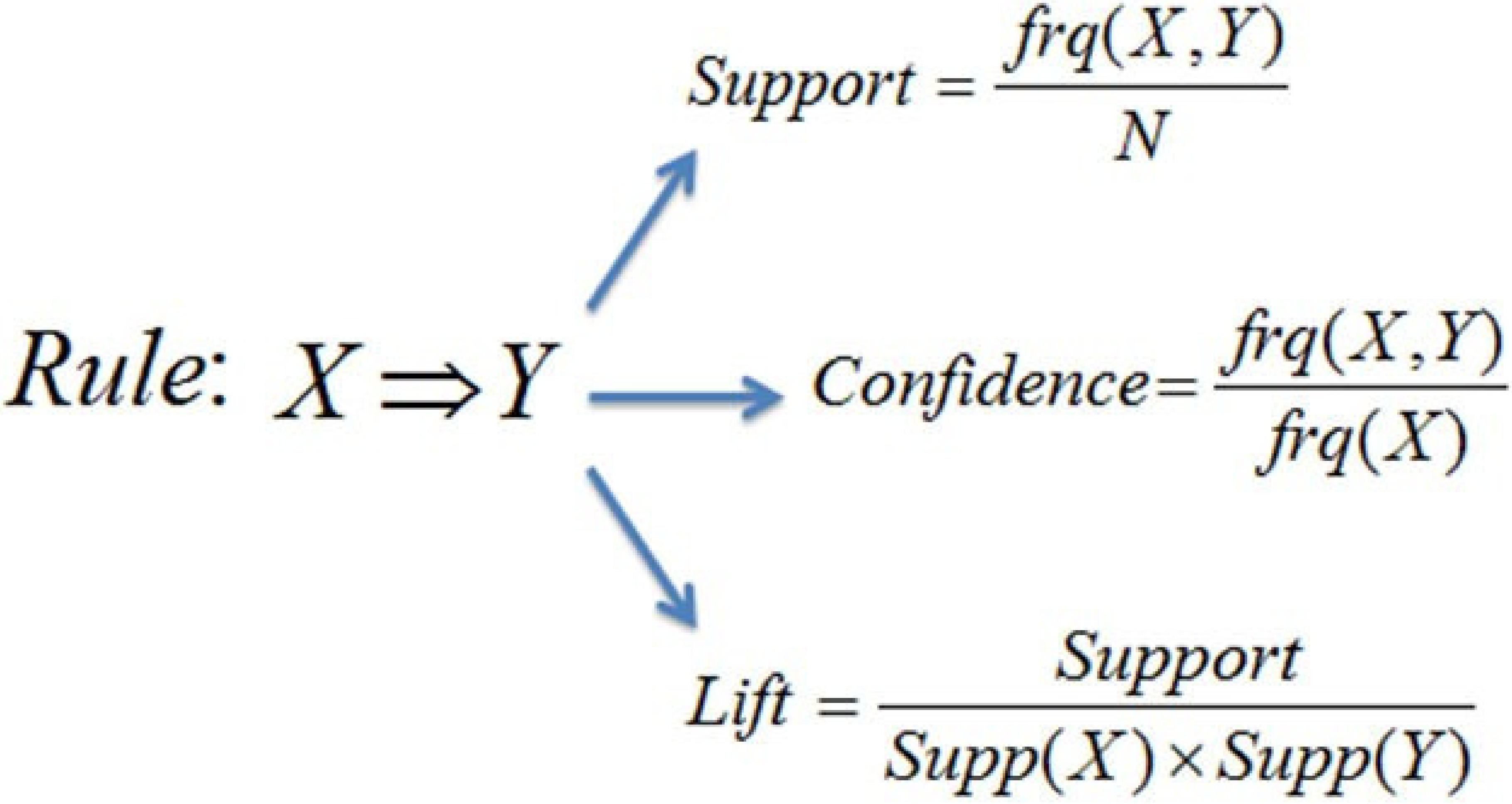
TABLE 6.1: Sample Data to Illustrate Association Rules

Transaction	Items
t_1	Bread, Jelly, PeanutButter
t_2	Bread, PeanutButter
t_3	Bread, Milk, PeanutButter
t_4	Beer, Bread
t_5	Beer, Milk



Definition 6.2. The **support** (**s**) for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$.

Definition 6.3. The **confidence or strength** (α) for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X.



$$Rule{:}\ X \Rightarrow Y$$

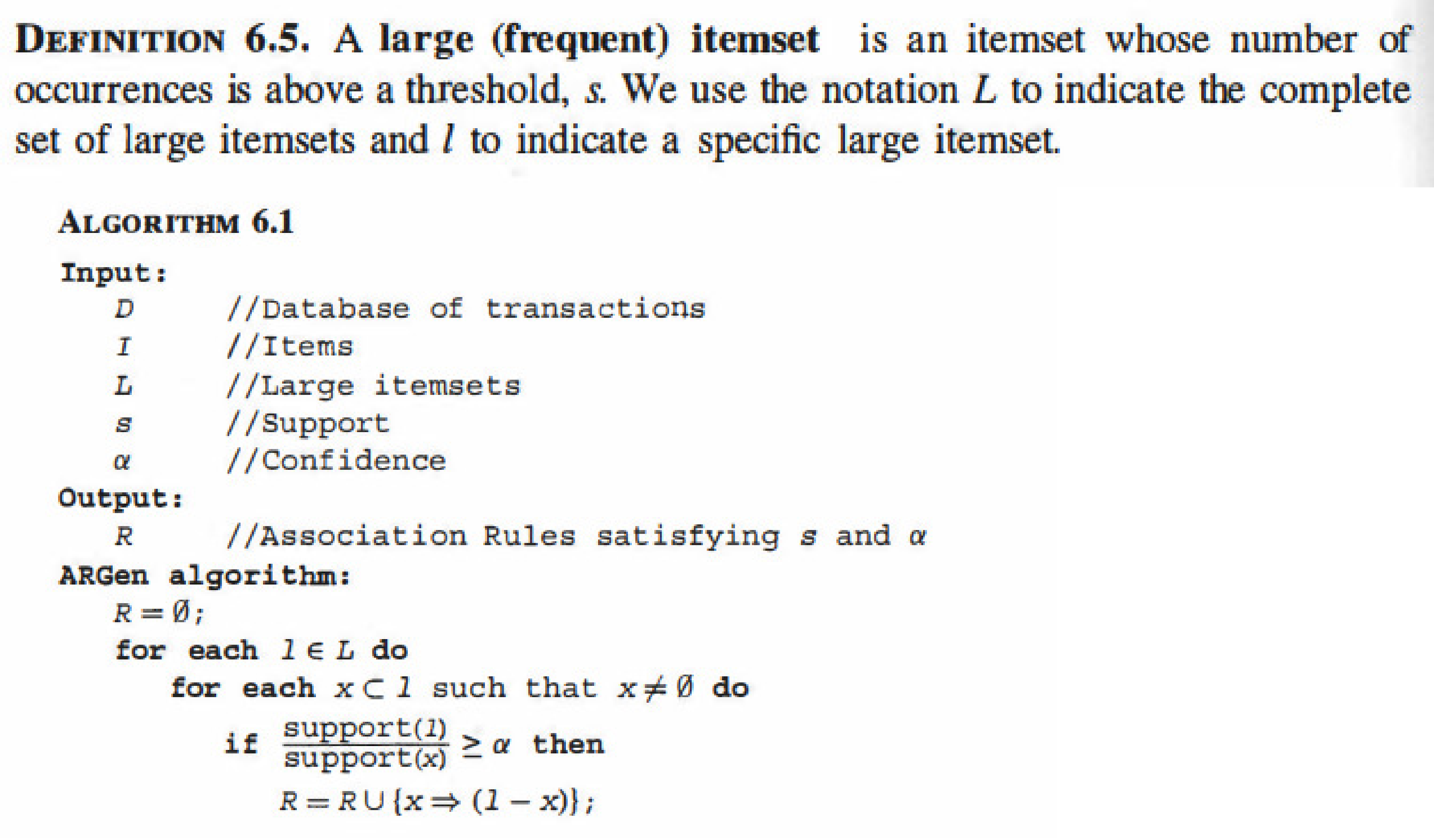
$$Support = \frac{frq(X,Y)}{N}$$

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

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

Apriori Algorithm

- The most well-known association rule algorithm is used in most commercial products.
- It uses the following property, which we call the large itemset property: any subset of a large itemset must be large.



Definition 6.5. A large (frequent) itemset is an itemset whose number of occurrences is above a threshold, *s*. We use the notation *L* to indicate the complete set of large itemsets and *l* to indicate a specific large itemset.

Algorithm 6.1

```
Input:
   D      //Database of transactions
   I      //Items
   L      //Large itemsets
   s      //Support
   α      //Confidence
Output:
   R      //Association Rules satisfying s and α
ARGen algorithm:
   $R = \emptyset$;
   for each $l \in L$ do
      for each $x \subset l$ such that $x \neq \emptyset$ do
         if $\frac{support(l)}{support(x)} \geq \alpha$ then
            $R = R \cup \{x \Rightarrow (l - x)\}$;
```

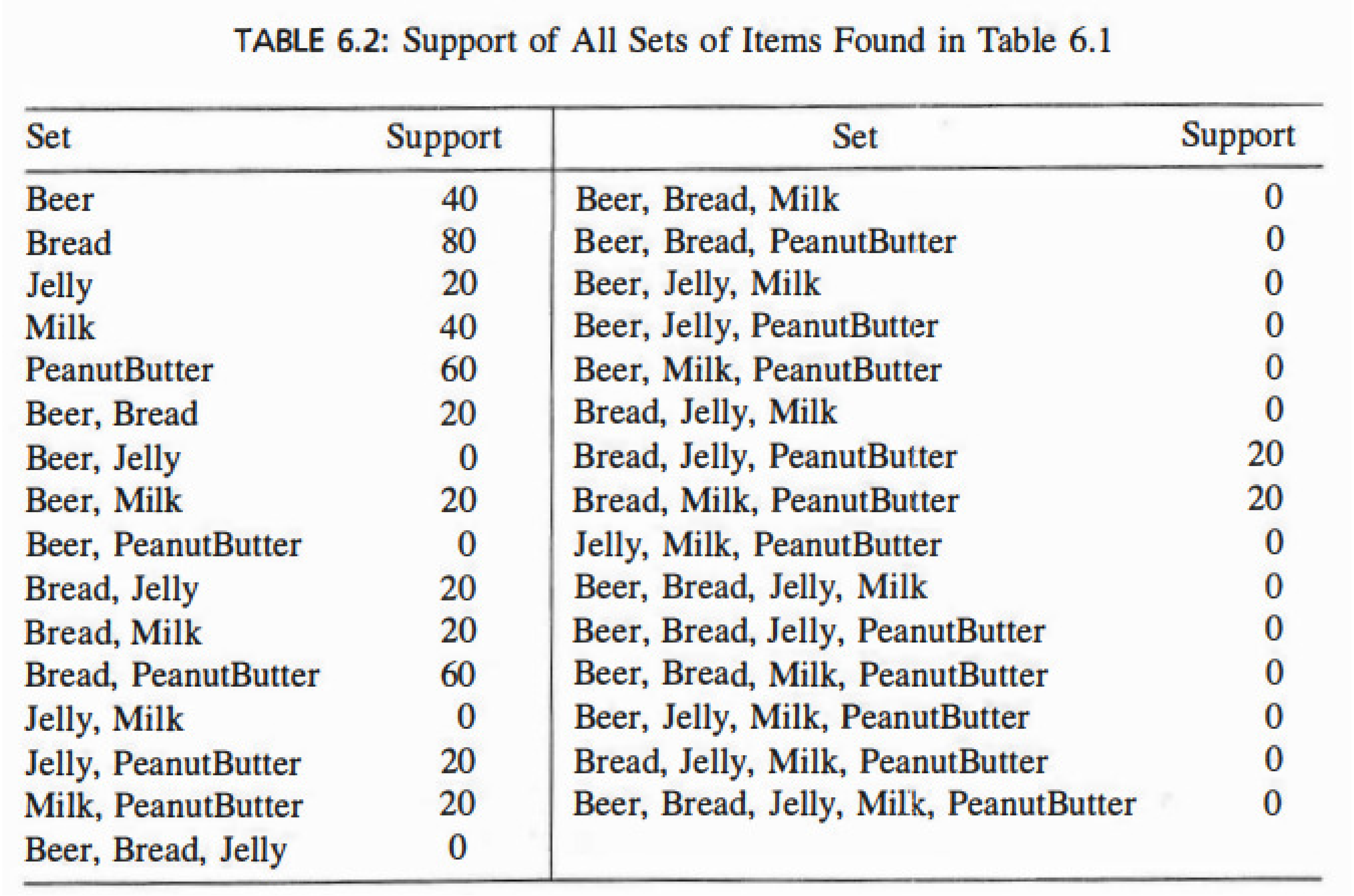
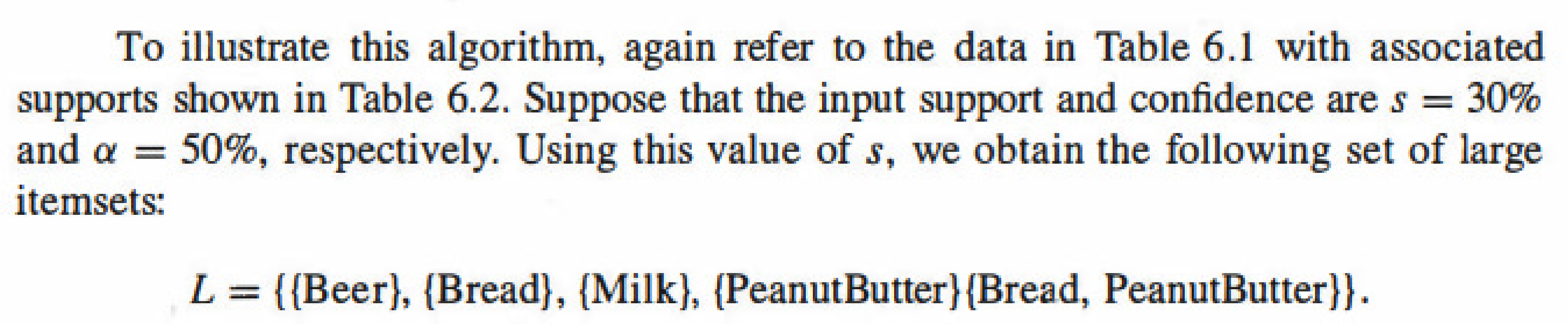


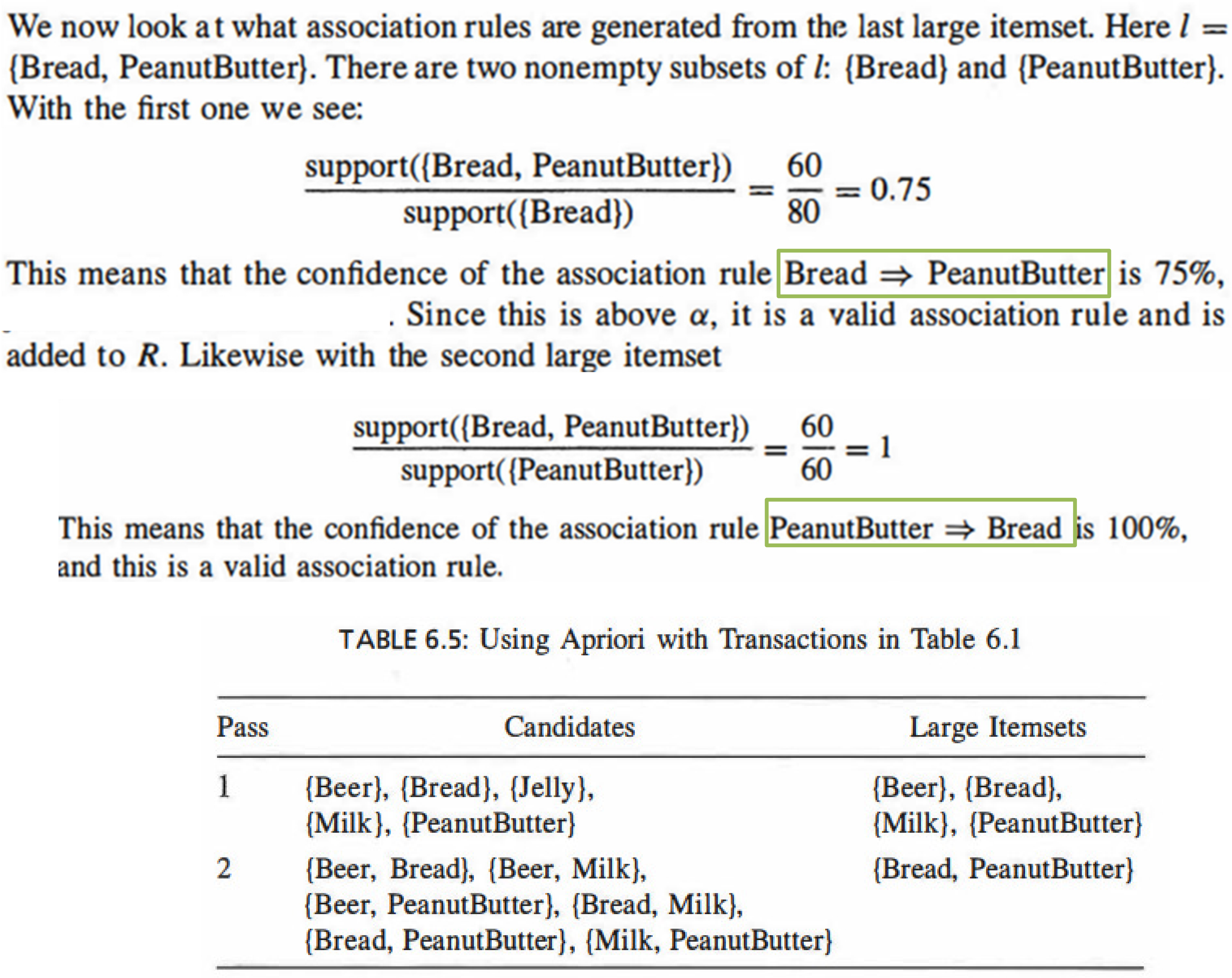
TABLE 6.2: Support of All Sets of Items Found in Table 6.1

Set	Support	Set	Support
Beer	40	Beer, Bread, Milk	0
Bread	80	Beer, Bread, PeanutButter	0
Jelly	20	Beer, Jelly, Milk	0
Milk	40	Beer, Jelly, PeanutButter	0
PeanutButter	60	Beer, Milk, PeanutButter	0
Beer, Bread	20	Bread, Jelly, Milk	0
Beer, Jelly	0	Bread, Jelly, PeanutButter	20
Beer, Milk	20	Bread, Milk, PeanutButter	20
Beer, PeanutButter	0	Jelly, Milk, PeanutButter	0
Bread, Jelly	20	Beer, Bread, Jelly, Milk	0
Bread, Milk	20	Beer, Bread, Jelly, PeanutButter	0
Bread, PeanutButter	60	Beer, Bread, Milk, PeanutButter	0
Jelly, Milk	0	Beer, Jelly, Milk, PeanutButter	0
Jelly, PeanutButter	20	Bread, Jelly, Milk, PeanutButter	0
Milk, PeanutButter	20	Beer, Bread, Jelly, Milk, PeanutButter	0
Beer, Bread, Jelly	0		



To illustrate this algorithm, again refer to the data in Table 6.1 with associated supports shown in Table 6.2. Suppose that the input support and confidence are $s = 30\%$ and $\alpha = 50\%$, respectively. Using this value of s, we obtain the following set of large itemsets:

$$L = \{\{\text{Beer}\}, \{\text{Bread}\}, \{\text{Milk}\}, \{\text{PeanutButter}\}\{\text{Bread, PeanutButter}\}\}.$$



We now look at what association rules are generated from the last large itemset. Here $l = \{\text{Bread, PeanutButter}\}$. There are two nonempty subsets of l: {Bread} and {PeanutButter}. With the first one we see:

$$\frac{\text{support}(\{\text{Bread, PeanutButter}\})}{\text{support}(\{\text{Bread}\})} = \frac{60}{80} = 0.75$$

This means that the confidence of the association rule Bread $\Rightarrow$ PeanutButter is 75%, . Since this is above α, it is a valid association rule and is added to R. Likewise with the second large itemset

$$\frac{\text{support}(\{\text{Bread, PeanutButter}\})}{\text{support}(\{\text{PeanutButter}\})} = \frac{60}{60} = 1$$

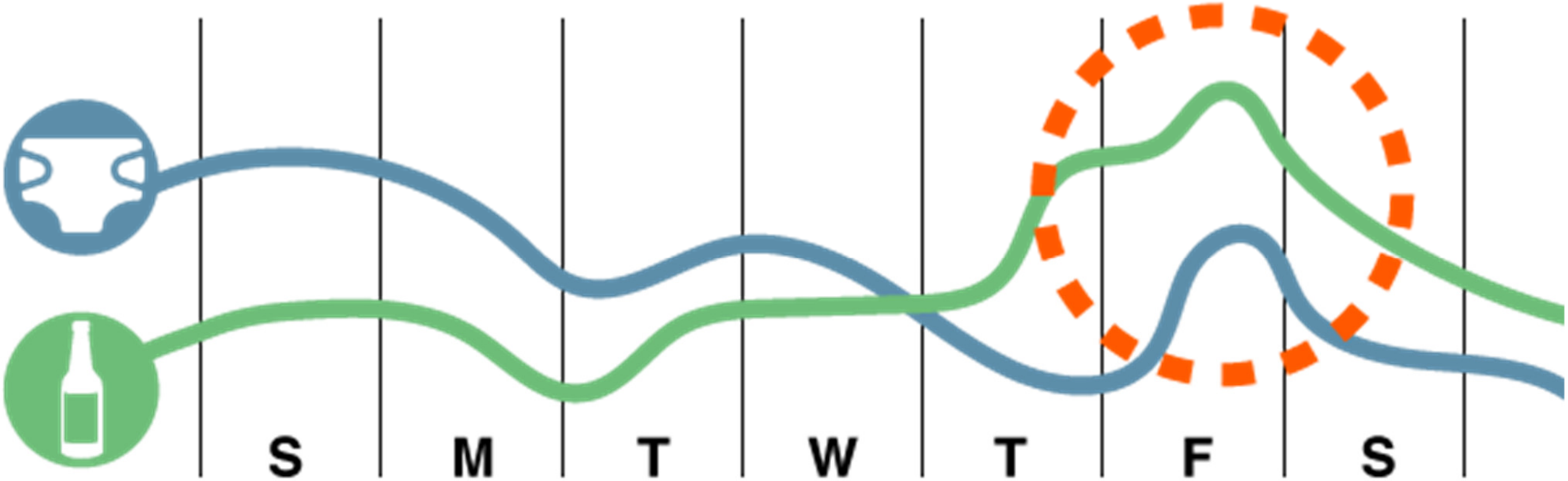
This means that the confidence of the association rule PeanutButter $\Rightarrow$ Bread is 100%, and this is a valid association rule.

TABLE 6.5: Using Apriori with Transactions in Table 6.1

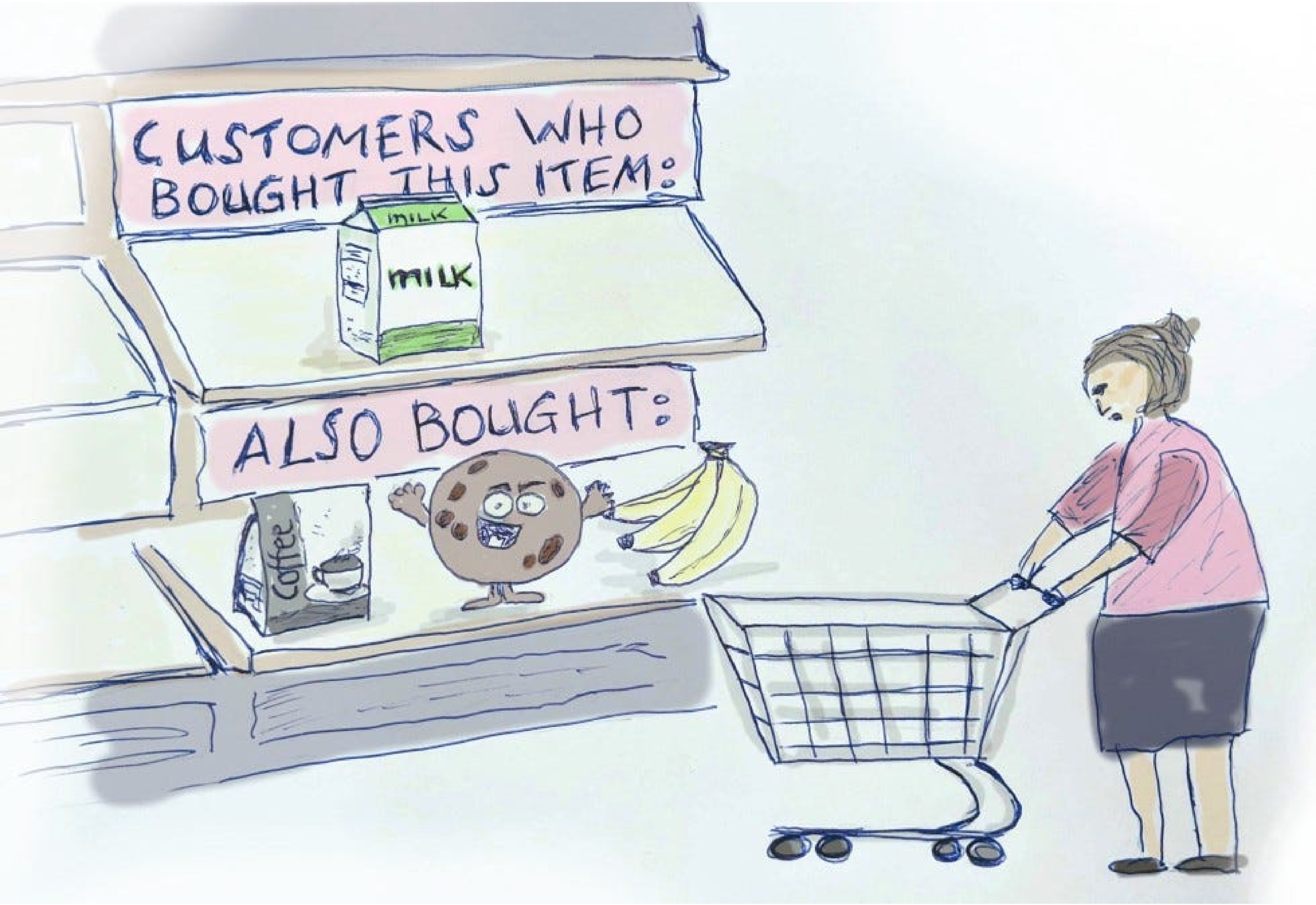
Pass	Candidates	Large Itemsets
1	{Beer}, {Bread}, {Jelly}, {Milk}, {PeanutButter}	{Beer}, {Bread}, {Milk}, {PeanutButter}
2	{Beer, Bread}, {Beer, Milk}, {Beer, PeanutButter}, {Bread, Milk}, {Bread, PeanutButter}, {Milk, PeanutButter}	{Bread, PeanutButter}

A classic example of association rules in data mining.

- On Friday afternoons, young American males who buy diapers also have a predisposition to buy beer.



- A supermarket has 200,000 customer transactions. About 4,000 transactions, or about 2% of the total number of transactions, include the purchase of diapers.
- About 5,500 transactions (2.75%) include the purchase of beer. Of those, about 3,500 transactions, 1.75%, include both the purchase of diapers and beer.
- Based on the percentages, that large number should be much lower. However, the fact that about 87.5% of diaper purchases include the purchase of beer indicates a link between diapers and beer.


CUSTOMERS WHO BOUGHT THIS ITEM:
MILK
MILK
ALSO BOUGHT:
coffee

Example 2

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

{ Milk, Diaper} => {Beer}

Support	**{Milk,Diaper U Beer}/N =2/5=0.4**
Confidence	**2/3=0.67**
Lift	**Confidence/Support(y)=Support/{Support(x).Support(y)} = 0.67/{3/5}=0.67/0.6=1.11** **Lift value near 1 indicates X& Y appear almost together.** **Lift > 1 means they appear together more than expected** **Lift < 1 means they appear together less than expected** **Greater Lift value indicates a stronger association**