DICE
ANALYTICS

DATA SCIENCE & MACHINE LEARNING COURSE

https://www.facebook.com/diceanalytics/
https://pk.linkedin.com/company/diceanalytics

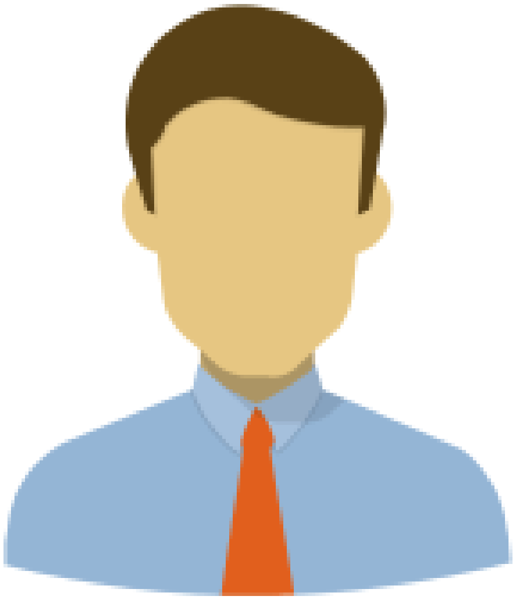
Association Rule Mining

- An Association Rule is a pattern that states when Event A occurs, another Event B occurs with certain probability.

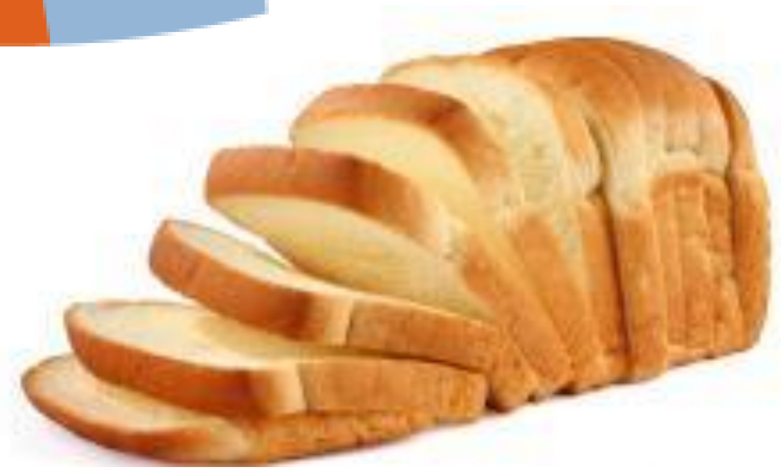
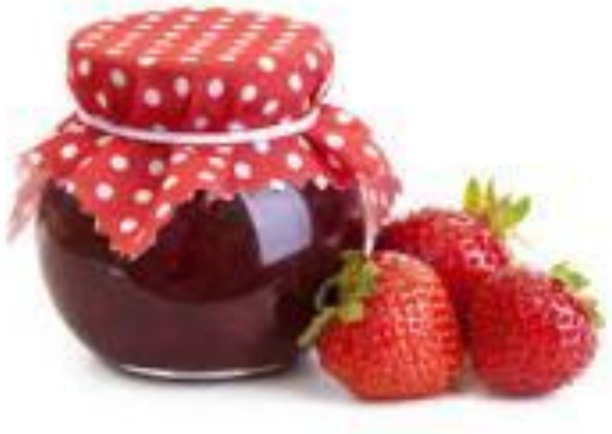
- These are *if/then* statements that help discover relationships between unrelated data in a data repository.

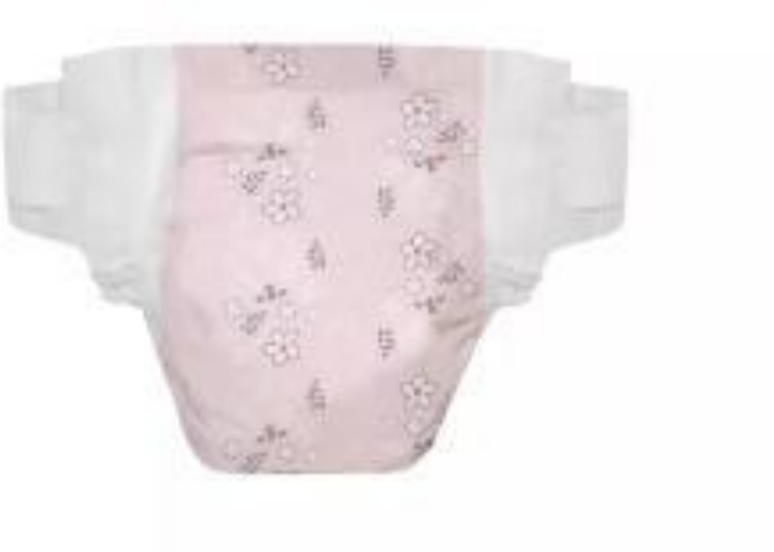
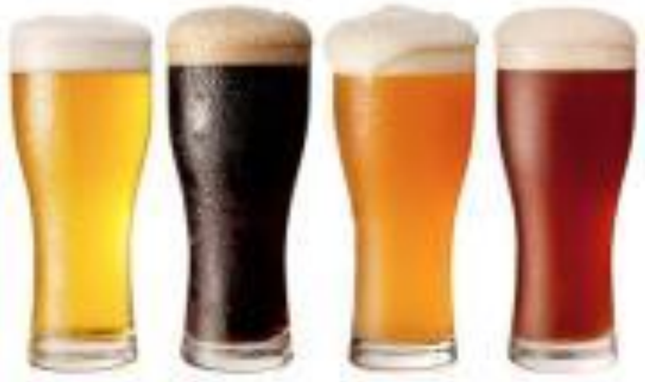
- Algorithms : Apriori, etc.
- Example : Market Basket Analysis

Association Rule Mining (Cases)



Walmart

1)  + 

2)  + 

Association Rule Mining (Terms)

Rule (A $\Longrightarrow$ B)

- A is called L.H.S (Left Hand Side)
- B is called R.H.S (Right Hand Side)

- Used to show Association among two items
- If A is diaper and B is beer, it means when a customer buys diaper, he would buy beer too.

Association Rule Mining (Terms)

$$\text{Support}\ (A \Longrightarrow B) = \frac{\text{Freq}\ (A \text{ and } B)}{N}$$

$$= P\ (A\ \&\ B)$$

- Support means the probability of the customer buying Item A and Item B together among all sales transactions.
- Range 0 to 1

Association Rule Mining (Terms)

$$\text{Confidence } (A \Longrightarrow B) = \frac{P\,(A \text{ and } B)}{P\,(A)}$$

$$= P\,(B \mid A)$$

- Confidence means that if a customer picks up Item A, how he is likely to buy Item B?.
- The maximum value of confidence has to be 1.

Is Confidence Enough?

	Basketball	No basketball	Total
Cereal	2000	1750	3750
No Cereal	1000	250	1250
Total	3000	2000	5000

Sup(B→C) = 40%

P(B) = 60%

Conf(B→C) = 66.67%

P(C) = 75%

Is Confidence Enough?

	Basketball	No basketball	Total
Cereal	2000	1750	3750
No Cereal	1000	250	1250
Total	3000	2000	5000

Sup(B→nC) = 20%

P(B) = 60%

Conf(B→nC) = 33.33%

P(nC) = 25%

Association Rule Mining (Terms)

$$\textbf{Lift}\ (A \Longrightarrow B) = \frac{P\ (A \text{ and } B)}{P\ (A) \times P\ (B)}$$

$$= \frac{\text{Confidence}\ (A \Longrightarrow B)}{P\ (B)}$$

- Lift is a true comparison between naive model and our model.
- It means how more likely a customer buy both, compared to buy separately.
- Range can be from 0 to +inf
- If 1 then independent

Sensitivity: Internal

DICE ANALYTICS

Is Confidence Enough?

	Basketball	No basketball	Total
Cereal	2000	1750	3750
No Cereal	1000	250	1250
Total	3000	2000	5000

Sup(B→C) = 40%

Conf(B→C) = 66.67%

Lift(B→C) = 0.89

P(B) = 60%

P(C) = 75%

Is Confidence Enough?

	Basketball	No basketball	Total
Cereal	2000	1750	3750
No Cereal	1000	250	1250
Total	3000	2000	5000

Sup(B→nC) = 20%

Conf(B→nC) = 33.33%

Lift(B→nC) = 1.33

P(B) = 60%

P(nC) = 25%

DICE
ANALYTICS

Is Lift Enough?

	Basketball	No basketball	Total
Cereal	100	1000	1100
No Cereal	1000	100000	101000
Total	1100	101000	102100

Sup(B→C) = 0.10%

Conf(B→C) = 9.09%

P(B) = 1%

P(C) = 1%

Lift(B→C) = 8.44

Association Rule Mining (Terms)

$$\textbf{Leverage}\ (A \Longrightarrow B) =$$

$$P\ (A\ \text{and}\ B) \;-\; P\ (A) \times P\ (B)$$

- Lift may find very strong associations for less frequent items, while Leverage tends to prioritize items with higher frequencies/support in the dataset.
- Range from -1 to 1
- If near to 0 then independent

Association Rule Mining (Terms)

$$\textbf{Conviction}\ (A \Longrightarrow B)$$

$$= \frac{1 - \text{Support}(B)}{1 - \text{Confidence}(A \Longrightarrow B)}$$

- Conviction tells us the %age about Rule (A => B) being incorrect if association between A and B was an accidental chance.
- Range is from 0 to inf
- If near to 1 then independent

Is Lift Enough?

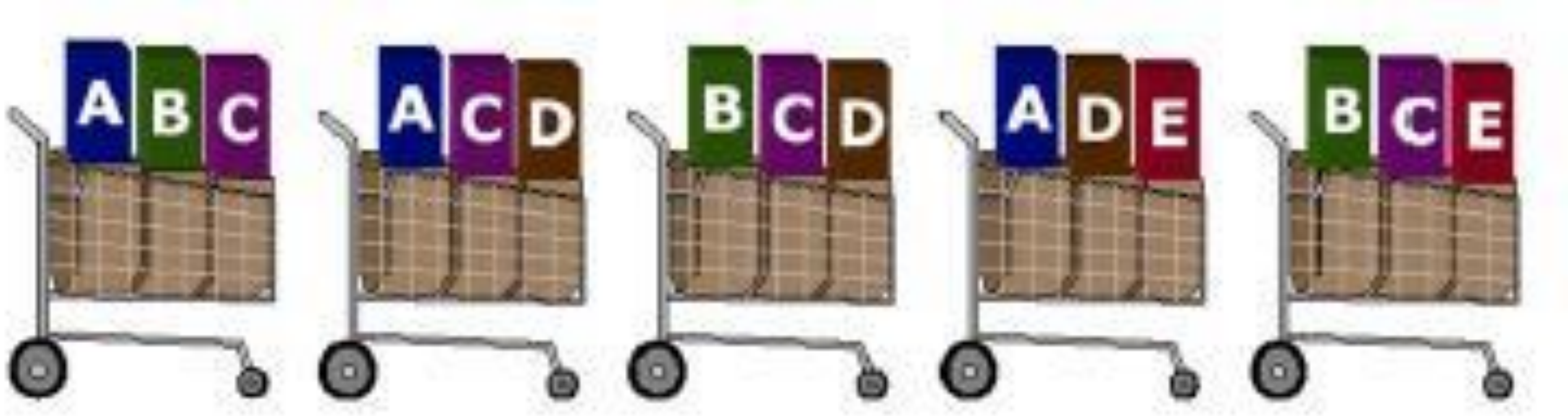
	Basketball	No basketball	Total
Cereal	100	1000	1100
No Cereal	1000	100000	101000
Total	1100	101000	102100

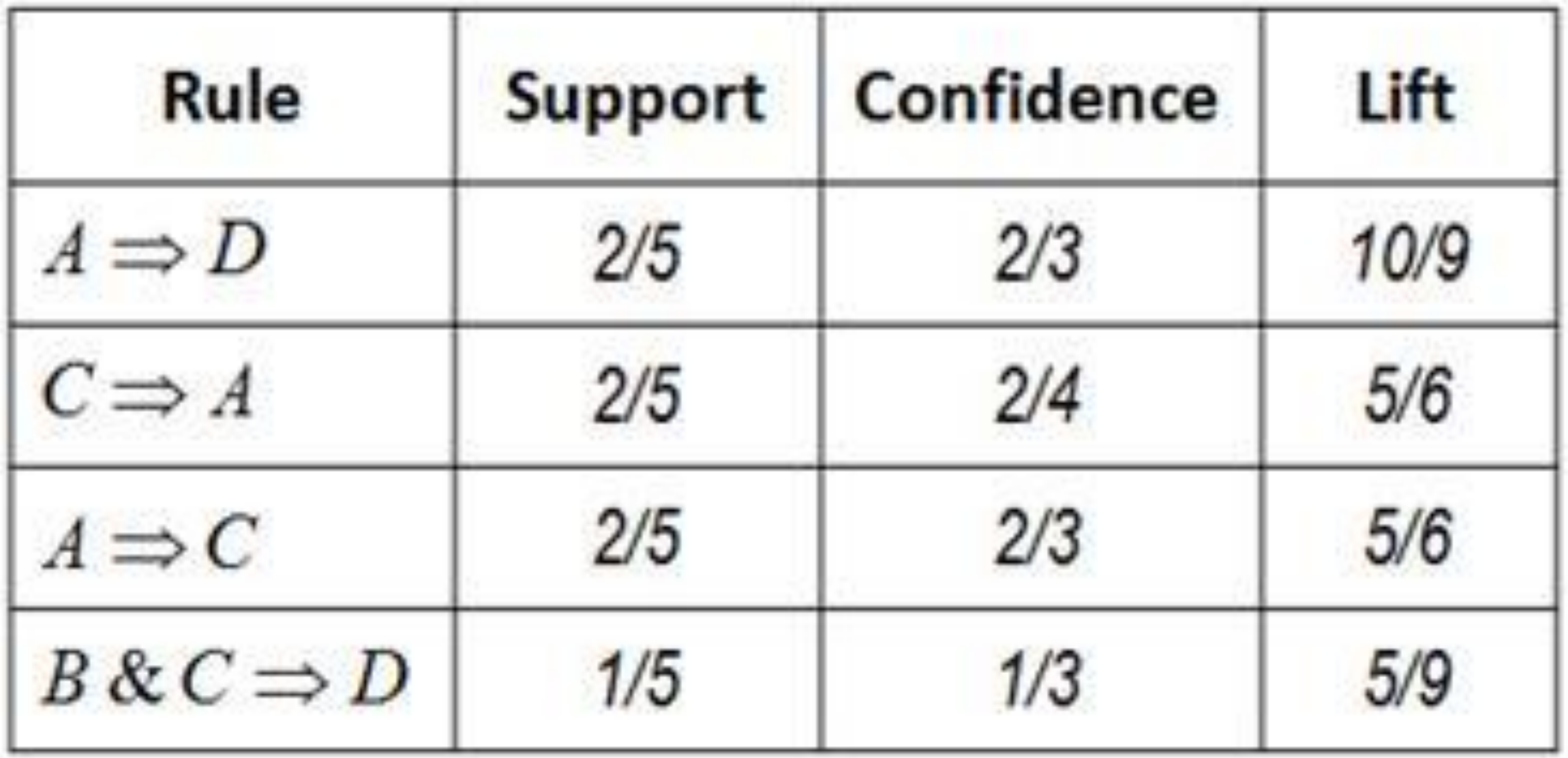
Sup(B→C) = 0.10% **Lev(B→C) = 0.09%** **P(B) = 1%**

Conf(B→C) = 9.09% **Cov(B→C) = 1.08** **P(C) = 1%**

Lift(B→C) = 8.44

Association Rule Mining (Example)





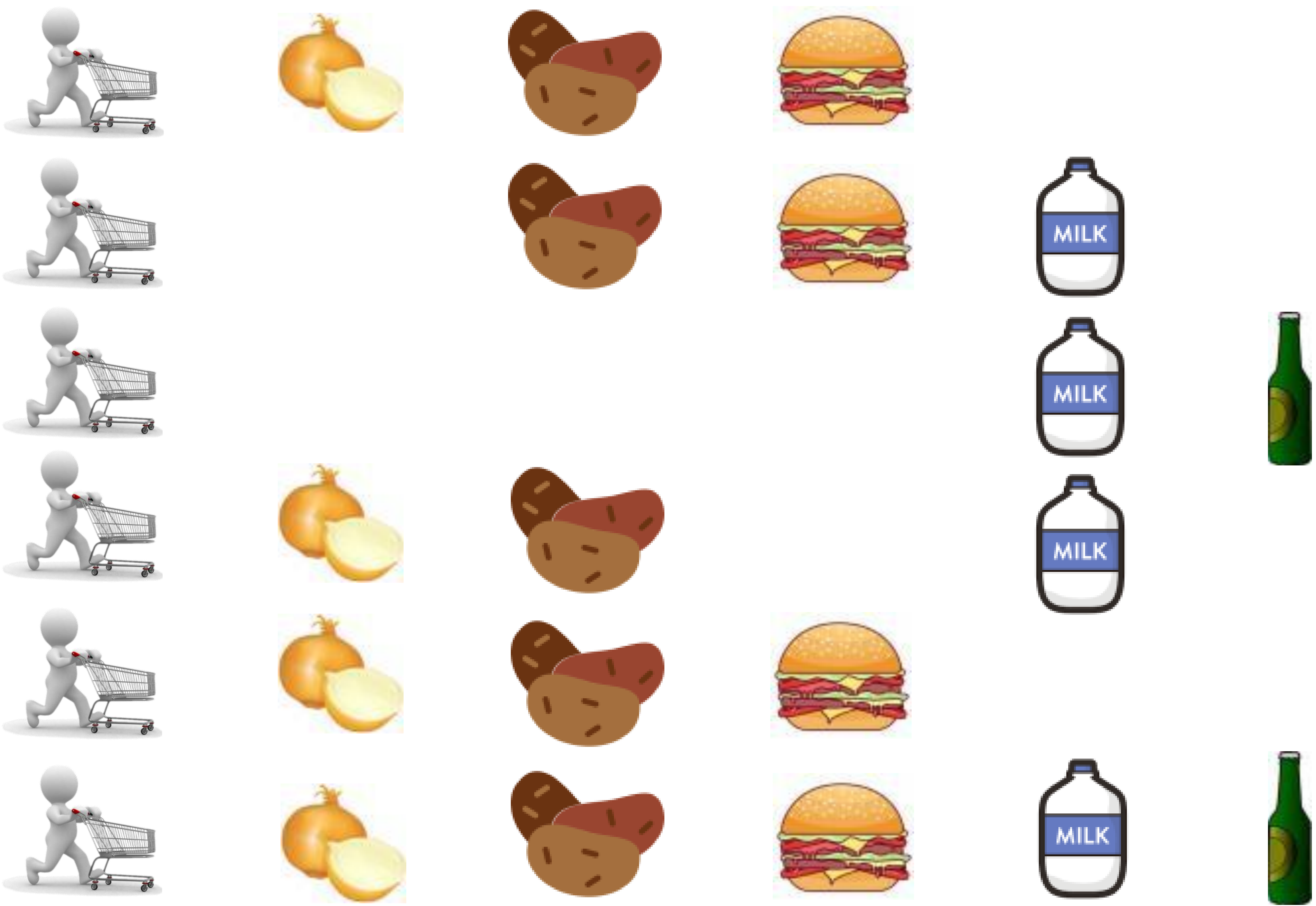
Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \,\&\, C \Rightarrow D$	1/5	1/3	5/9

Apriori Algorithm

- **Find the frequent itemsets: the sets of items that have minimum support:**
  - **A subset of a frequent itemset must also be a frequent itemset „**
    - **Generate length (k+1) candidate itemsets from length k frequent itemsets, and „**
    - **Test the candidates against DB to determine which are in fact frequent**

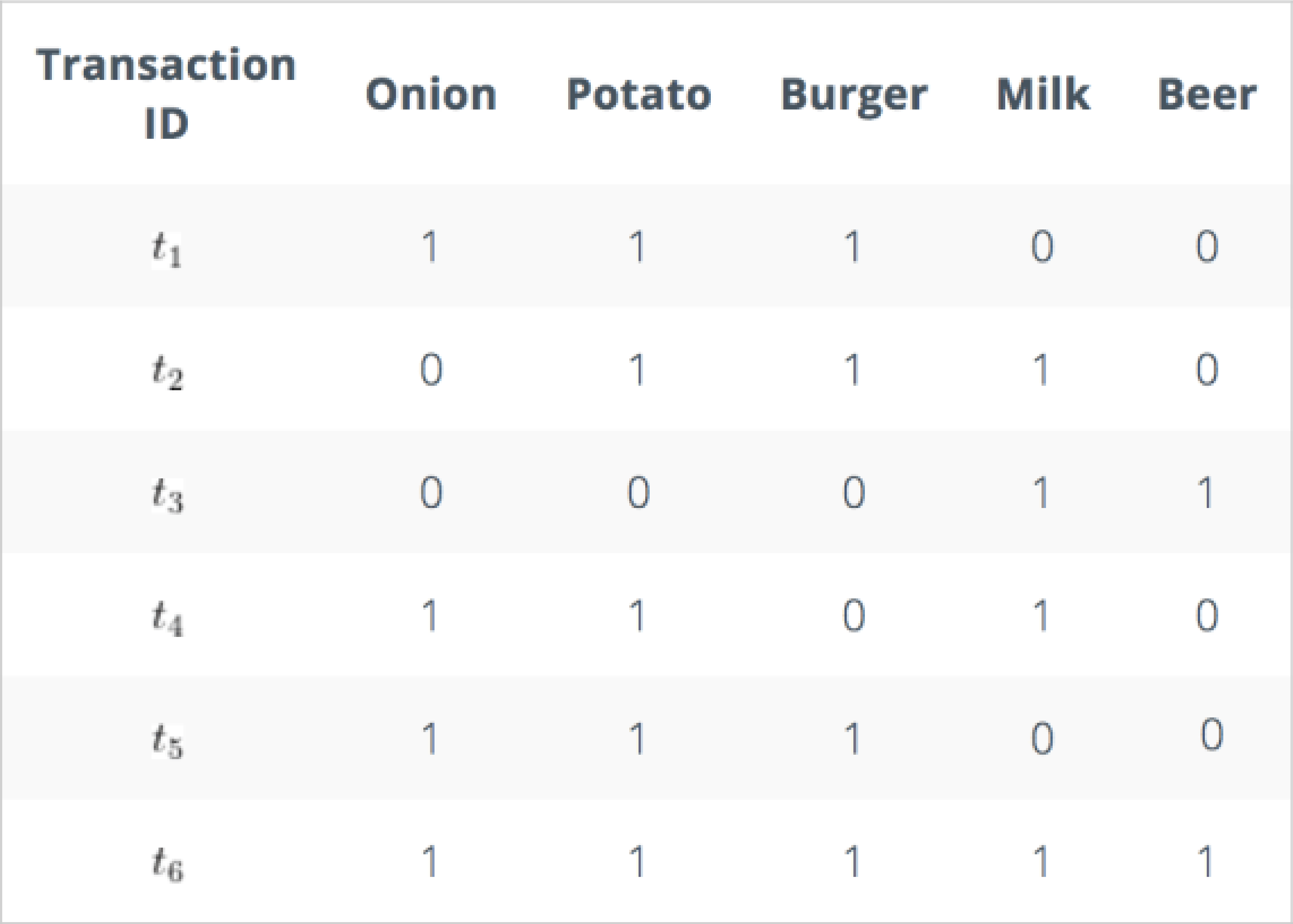
- **Use the frequent itemsets to generate association rules.**

Apriori Algorithm (Steps)



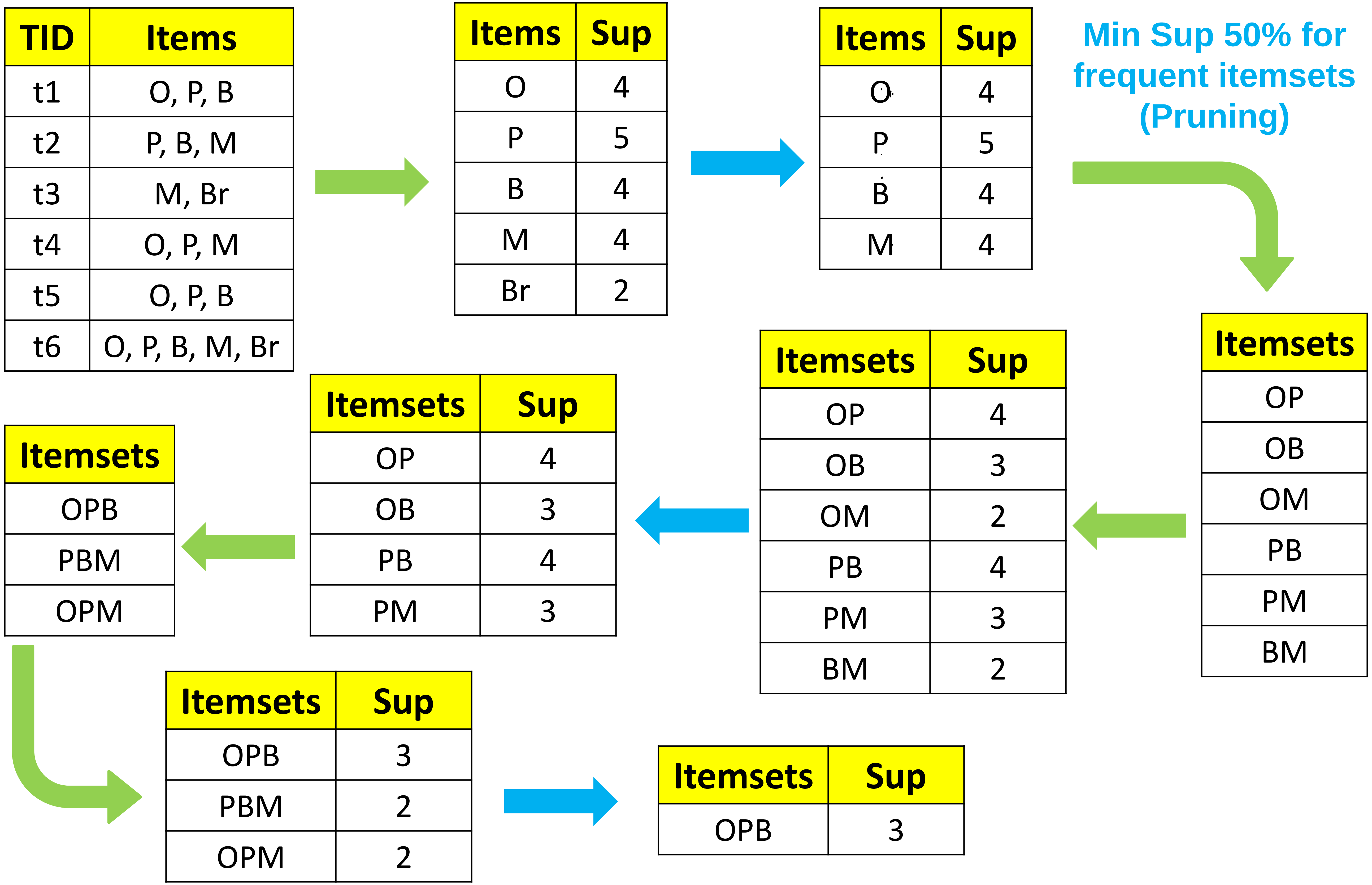
Apriori Algorithm (Steps)

Convert DB to One-Hot Encoding



Transaction ID	Onion	Potato	Burger	Milk	Beer
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	0
t_6	1	1	1	1	1

Apriori Algorithm (Steps)



TID	Items
t1	O, P, B
t2	P, B, M
t3	M, Br
t4	O, P, M
t5	O, P, B
t6	O, P, B, M, Br

Items	Sup
O	4
P	5
B	4
M	4
Br	2

Items	Sup
O	4
P	5
B	4
M	4

Min Sup 50% for frequent itemsets (Pruning)

Itemsets
OP
OB
OM
PB
PM
BM

Itemsets	Sup
OP	4
OB	3
OM	2
PB	4
PM	3
BM	2

Itemsets	Sup
OP	4
OB	3
PB	4
PM	3

Itemsets
OPB
PBM
OPM

Itemsets	Sup
OPB	3
PBM	2
OPM	2

Itemsets	Sup
OPB	3

Apriori Algorithm (Steps)

Final Frequent Items sets using algorithm

Itemsets	Support
O	4
P	5
B	4
M	4
OP	4
OB	3
PB	4
PM	3
OPB	3

Association Rules will be made

Apriori Algorithm

➢ Challenges:

- Multiple scans of transaction database
- Huge number of candidates
- Tedious workload of support calculation for each candidate

➢ Improving of Apriori:

- Reduce number of transaction database scan
- Shrink number of candidates
- Facilitate support counting of candidates