# ASSOCIATION RULE MINING

## *Market Basket Analysis*

**Analysis of customer buying habits by finding associations and correlations between the different items that customers place in their "shopping basket"**

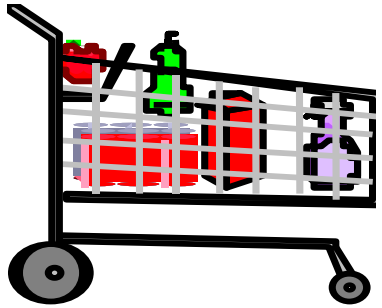
**Milk, eggs, sugar, bread**

**Customer1**

**Milk, eggs, cereal, bread**

**Customer2**

**Eggs, sugar**



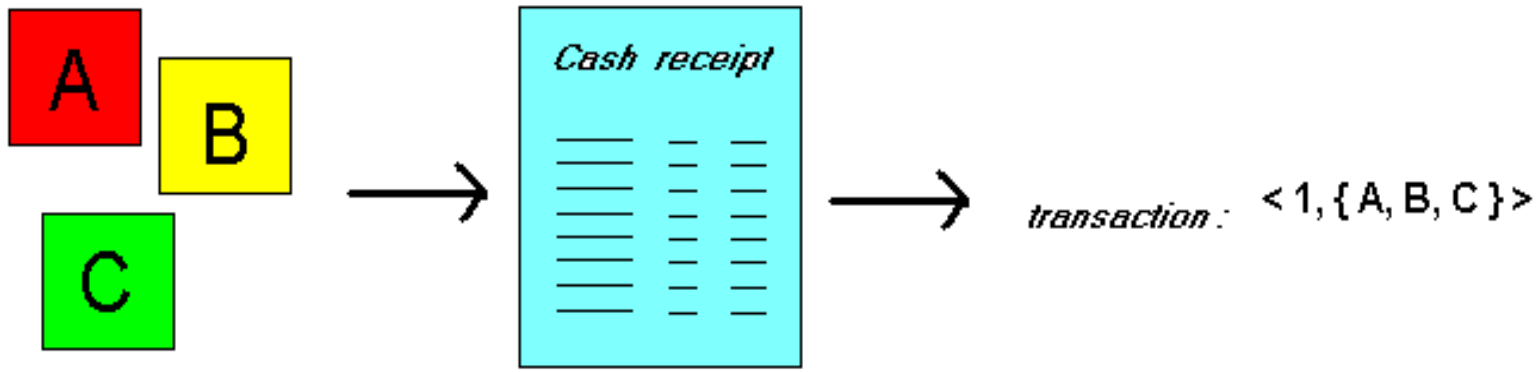
**Customer3**

# ASSOCIATION RULE MINING

## *Market Basket Analysis*

**Given: A database of customer transactions (e.g., shopping baskets), where each transaction is a set of items (e.g., products)**

**Find: Groups of items which are frequently purchased together**



# ASSOCIATION RULE MINING

## *Basic Concepts*

**Given database of transactions, each transaction is a list of items (purchased by a customer in a visit)**

**Find all rules that correlate the presence of one set of items with that of another set of items**

**Example:** ***98% of people who purchase tires and auto accessories also get automotive services done***

# ASSOCIATION RULE MINING

**Extract information on purchasing behavior**

**"IF buys coke and sausage, THEN also buy mustard with high probability“**

**Actionable information: can suggest...**

**New store layouts and product assortments**

**Which products to put on promotion**

# ASSOCIATION RULE MINING

**Useful:**

**"On Thursdays, super store consumers often purchase rice and meat together.“**

**Trivial:**

**"Customers who purchase maintenance agreements are very likely to purchase large appliances."**

**unexpected:**

**"When a new hardaware store opens, one of the most sold items is toilet rings."**

# ASSOCIATION RULE MINING

## *Association Rules: Basics*

- **Support:** denotes the frequency of the rule within transactions.

$$\text{support}(A \Rightarrow B\ [s, c]) = p(A \cup B) = \text{support}(\{A,B\})$$

- **Confidence:** denotes the percentage of transactions containing A which contain also B.

$$\text{confidence}(A \Rightarrow B\ [s, c]) = p(B|A) = p(A \cup B) / p(A) = \text{support}(\{A,B\}) / \text{support}(\{A\})$$

# ASSOCIATION RULE MINING

## *Association Rules: Basics*

- **Minimum support $\sigma$ :**
  - **High** $\Rightarrow$ **few** frequent itemsets
    $\Rightarrow$ **few** valid rules which occur **very often**
  - **Low** $\Rightarrow$ **many** valid rules which occur **rarely**

- **Minimum confidence $\gamma$ :**
  - **High** $\Rightarrow$ **few** rules, but all "**almost logically true**"
  - **Low** $\Rightarrow$ **many** rules, many of them very "**uncertain**"

- **Typical values:** $\sigma$ = 2 -10 %, $\gamma$ = 70 - 90 %

# ASSOCIATION RULE MINING

## *Rule Measures: Support & Confidence*

***Let minimum support 50%, and minimum confidence 50%, we have***

$A \Rightarrow C$ **(50%, 66.6%)**

$C \Rightarrow A$ **(50%, 100%)**

| Transaction ID | Items Bought |
|---|---|
| 2000 | A,B,C |
| 1000 | A,C |
| 4000 | A,D |
| 5000 | B,E,F |

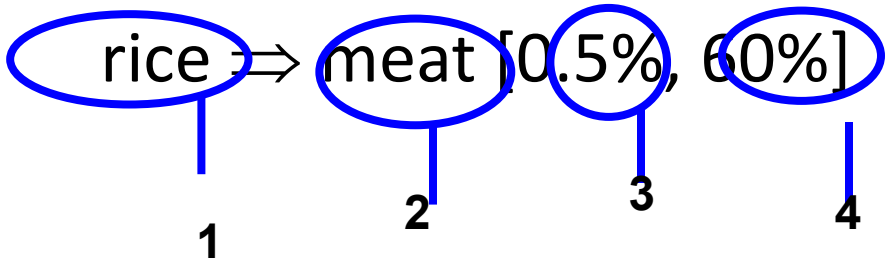
# ASSOCIATION RULE MINING

## *Association Rules: Basics*

- **Typical representation formats for association rules:**
  - rice $\Rightarrow$ meat [0.5%, 60%]
  - buys:rice $\Rightarrow$ buys:meat [0.5%, 60%]
  - "**IF** buys rice, **THEN** buys meat in 60% of the cases. rice and meat are bought together in 0.5% of the rows in the database."
- **Other representations (used in Han's book):**
  - buys(x, "rice") $\Rightarrow$ buys(x, "meat") [0.5%, 60%]
  - major(x, "CS") ^ takes(x, "DB") $\Rightarrow$ grade(x, "A") [1%, 75%]

# ASSOCIATION RULE MINING

## *Association Rules: Basics*




"**IF** buys rice,
**THEN** buys meat
in 60% of the cases
in 0.5% of the rows"

1. **Antecedent**, left-hand side (LHS), body
2. **Consequent**, right-hand side (RHS), head
3. **Support**, frequency ("in how big part of the data the things in left- and right-hand sides occur together")
4. **Confidence**, strength ("if the left-hand side occurs, how likely the right-hand side occurs")

# ASSOCIATION RULE MINING

## *Apriori Algorithm*

**Apriori algorithm finds frequent itemsets (itemsets with minimum support)**

**Association rules can then be generated from frequent itemsets**

**It consists of two steps:**

- **Generation of candidate itemsets**
- **Pruning of itemsets which are infrequent**

# ASSOCIATION RULE MINING

## *Apriori Algorithm: Finding frequent itemsets*

**It has an iterative approach known as a level-wise search**

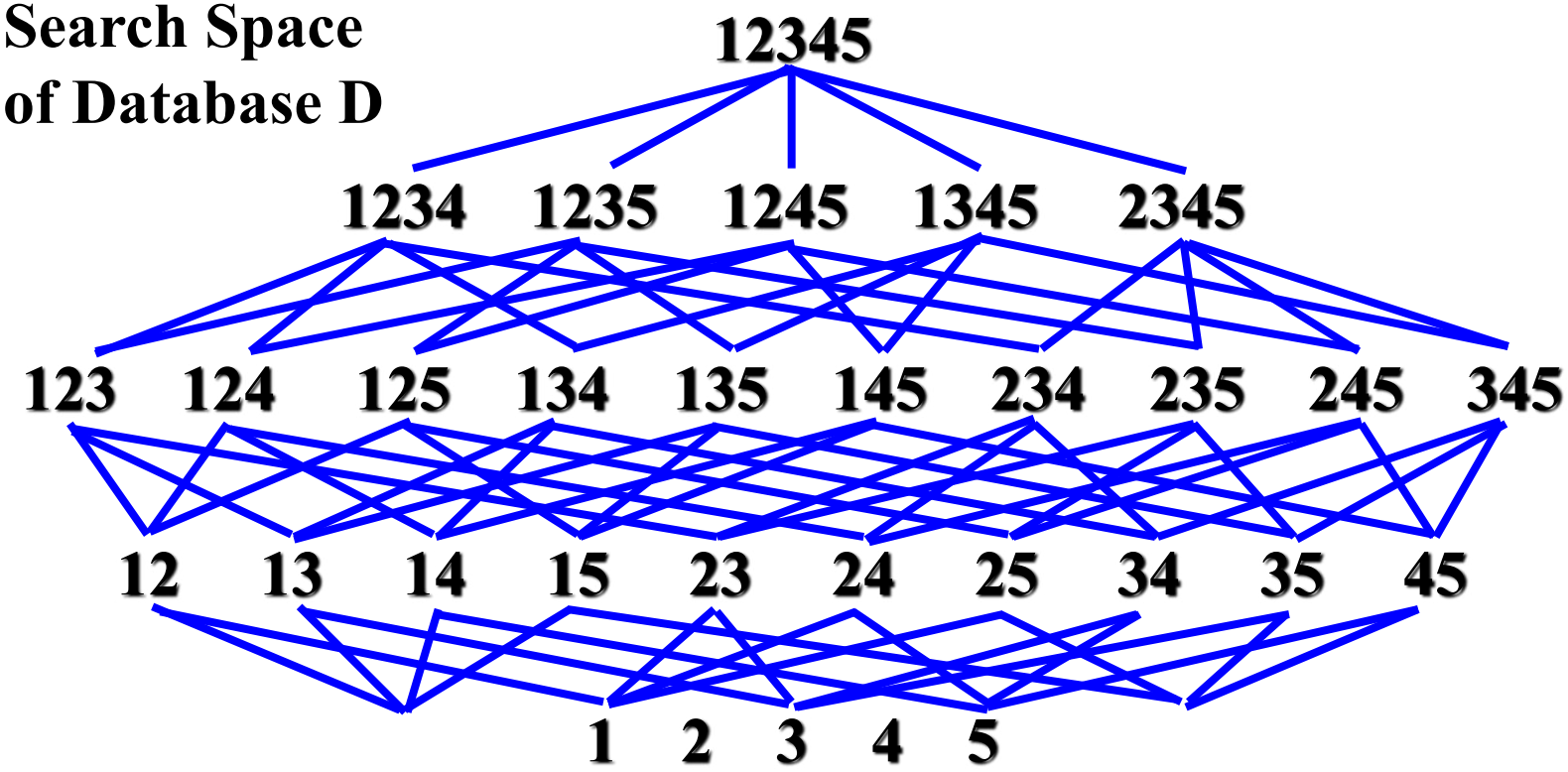
**First, the set of frequent 1-itemsets is found (called $L_1$)**
**$L_1$ is used to find frequent 2-itemsets ($L_2$)**
**$L_2$ is used to find $L_3$, and so on…**

**The finding of each L requires one full scan of the database**

# ASSOCIATION RULE MINING

## *Apriori Algorithm: Example*

**Search Space of Database D**



# ASSOCIATION RULE MINING

## *Apriori Algorithm: Finding frequent itemsets*

**The efficiency of the level-wise generation of frequent itemsets is improved by an important property (called the Apriori property)**

**With the help of this property, the search space is reduced**

***Apriori Property:*** **All non-empty subsets of a frequent itemsets must also be frequent**

# ASSOCIATION RULE MINING

## *Apriori Algorithm: Finding frequent itemsets*

**A subset of a frequent itemset must also be a frequent itemset**

**i.e., if {*AB*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset**

**The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties**

# ASSOCIATION RULE MINING

## *Apriori Algorithm: Example of Generating Candidates*

***L3={abc, abd, acd, ace, bcd}***

**Self-joining: *L3*L3***

***abcd*** **from *abc*** **and *abd***

***acde*** **from *acd*** **and *ace***

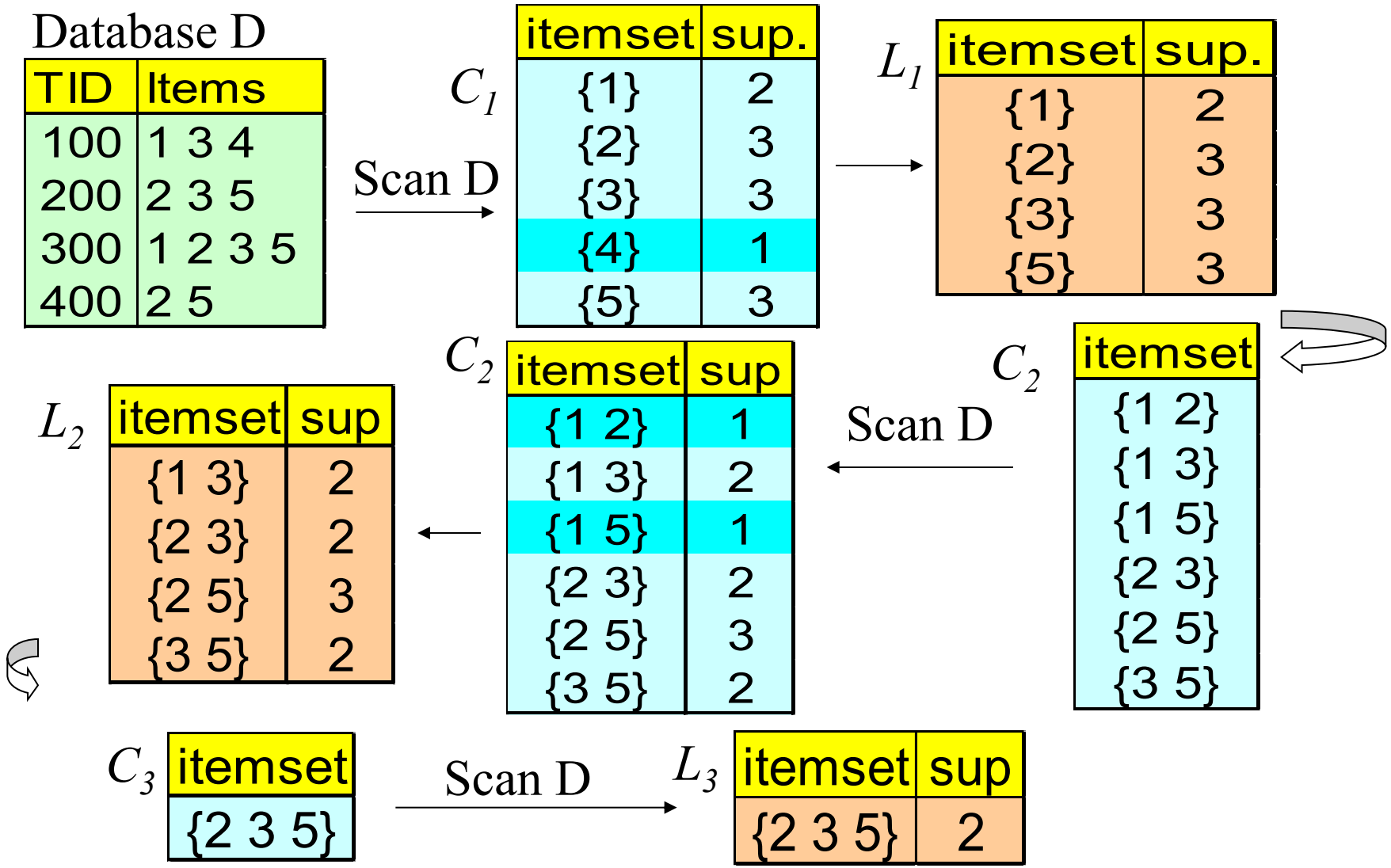
**Pruning:**

***acde*** **is removed because *ade*** **is not in *L3***

***C4={abcd}***

# ASSOCIATION RULE MINING

*Apriori Algorithm (min support =2 means 50%)*



# ASSOCIATION RULE MINING

## *Strong Rules from Frequent Itemsets*

**Once frequent itemsets have been found, we can convert them into association rules**

*AllElectronics* database

| TID | List of item_ID's |
|---|---|
| T100 | I1, I2, I5 |
| T200 | I2, I4 |
| T300 | I2, I3 |
| T400 | I1, I2, I4 |
| T500 | I1, I3 |
| T600 | I2, I3 |
| T700 | I1, I3 |
| T800 | I1, I2, I3, I5 |
| T900 | I1, I2, I3 |

**Suppose the data contains the frequent itemset $\{I_1, I_2, I_5\}$**

$I1 \wedge I2 \Rightarrow I5$, $confidence = 2/4 = 50\%$

$I1 \wedge I5 \Rightarrow I2$, $confidence = 2/2 = 100\%$

$I2 \wedge I5 \Rightarrow I1$, $confidence = 2/2 = 100\%$

$I1 \Rightarrow I2 \wedge I5$, $confidence = 2/6 = 33\%$

$I2 \Rightarrow I1 \wedge I5$, $confidence = 2/7 = 29\%$

$I5 \Rightarrow I1 \wedge I2$, $confidence = 2/2 = 100\%$