ECLT 5810/SEEM5750
Mining Association Rules

Association Rule: Basic Concepts

- Given: (1) a large database of transactions, (2) 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
  - E.g., *98% of people who purchase tires and auto accessories also get automotive services done*

Association Rule: Basic Concepts

- Applications
  - $* \Rightarrow$ *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales)
  - Home Electronics $\Rightarrow *$ (What other products should the store stocks up?)
  - Attached mailing in direct marketing
  - Insurance claim analysis

Association Rule: Support and Confidence

- "cake" and "bread" $\Rightarrow$ "milk" [0.5%, 60%]
  - Support and confidence are 2 measures of rule interestingness
  - 0.5% of all the transactions show that cake and bread and milk are purchased together.
  - 60% of the customers (transactions) who purchased (contained) cake and bread also purchased (contained) milk.
  - Association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold**.
  - The thresholds can be set by users or domain experts.

Association Rule: Support and Confidence

- Find all the rules $X_1, \ldots X_m \Rightarrow Y_1, \ldots Y_k$ with minimum confidence and support
  - support, s, probability that a transaction contains $\{X_1,\ldots X_m, Y_1,\ldots Y_k\}$
  - confidence, c, conditional probability that a transaction having $\{X_1,\ldots X_m\}$ also contains $\{Y_1,\ldots Y_k\}$

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

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

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

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

Association Rule: Basic Concepts

- *Itemset*: A set of items
- *k-itemset*: A set of k items
- *Support count* of a itemset = number of transactions containing the itemset
- An *itemset satisfies minimum support* (also called the *frequent itemset*) if the support count of the itemset >= min_sup x total number of transactions

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

Let minimum support 50%, we have

- itemset *{ A ,C } satisfying the minimum support*
- itemset { *A, C* } is a frequent itemset

Mining Association Rule: An Example

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

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

- For rule A $\Rightarrow$ C:
  - support = support({A, C}) = 50%
  - confidence = support({A, C})/support({A}) = 66.6%
- **Apriori principle:**
  - *Any subset of a frequent itemset must be frequent*

Mining Frequent Itemsets: Two Key Steps

- Step 1: Find the *frequent itemsets*: the sets of items that have minimum support
  - 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
  - Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Step 2: Use the frequent itemsets to generate association rules.

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D →

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D →

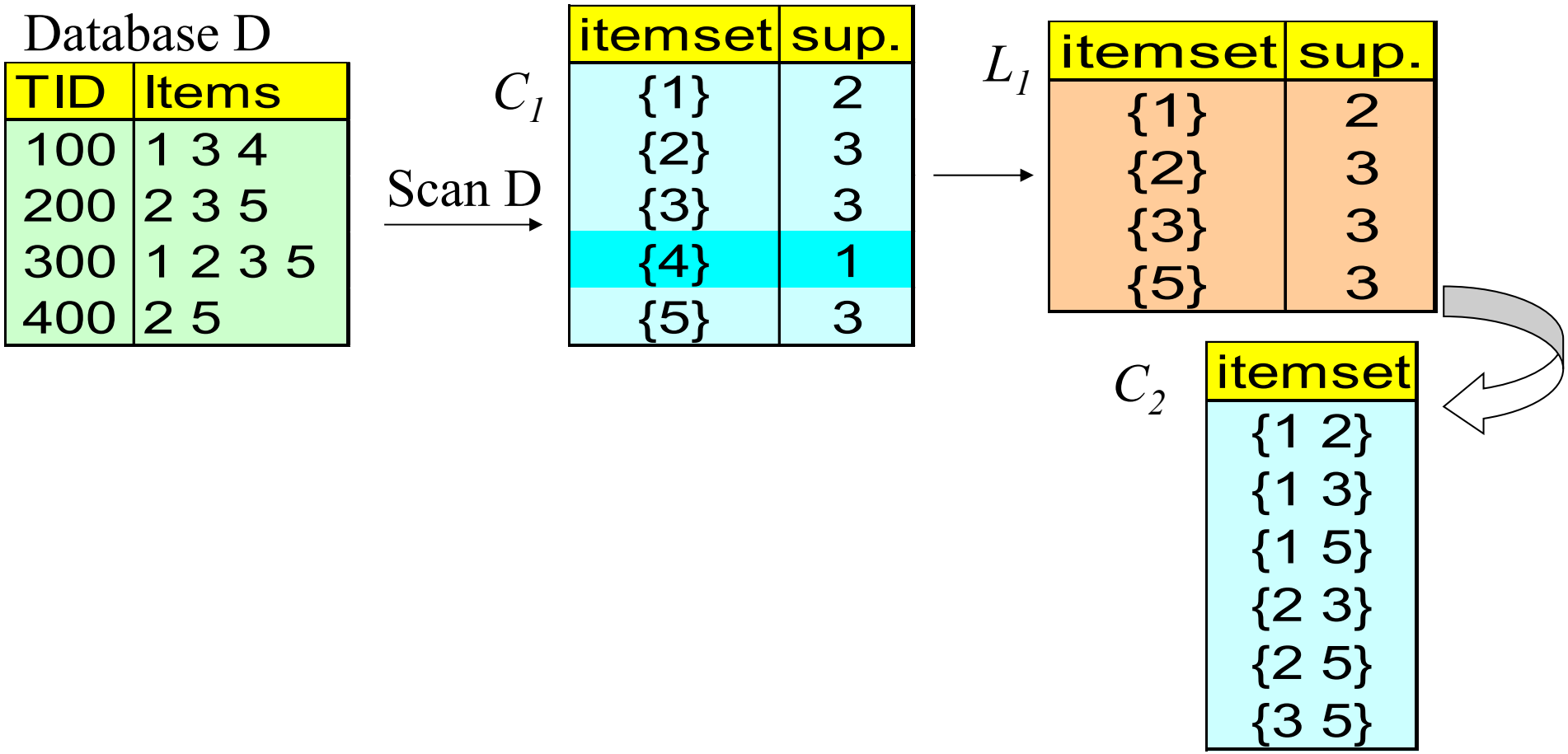
C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

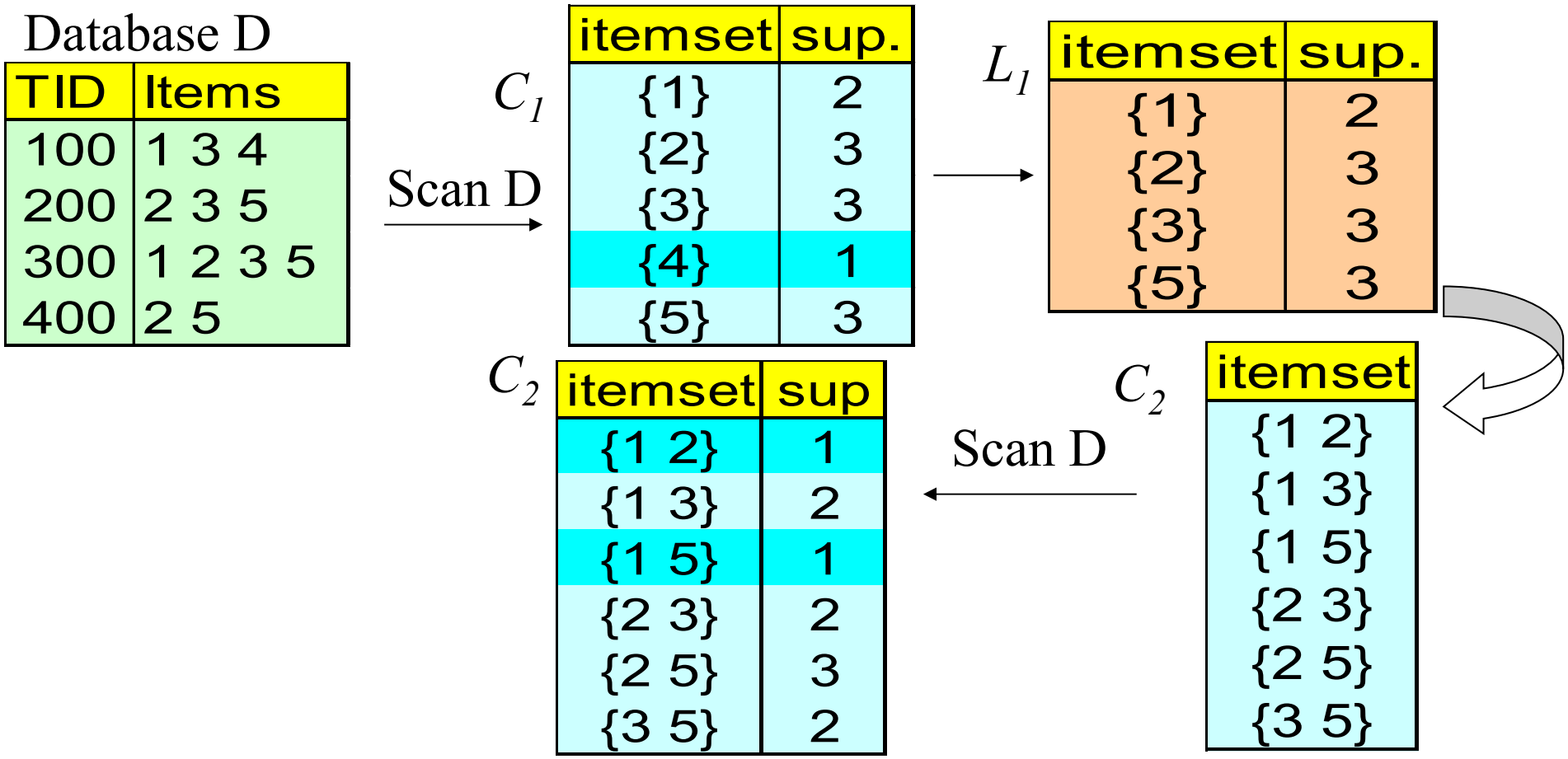
→ L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

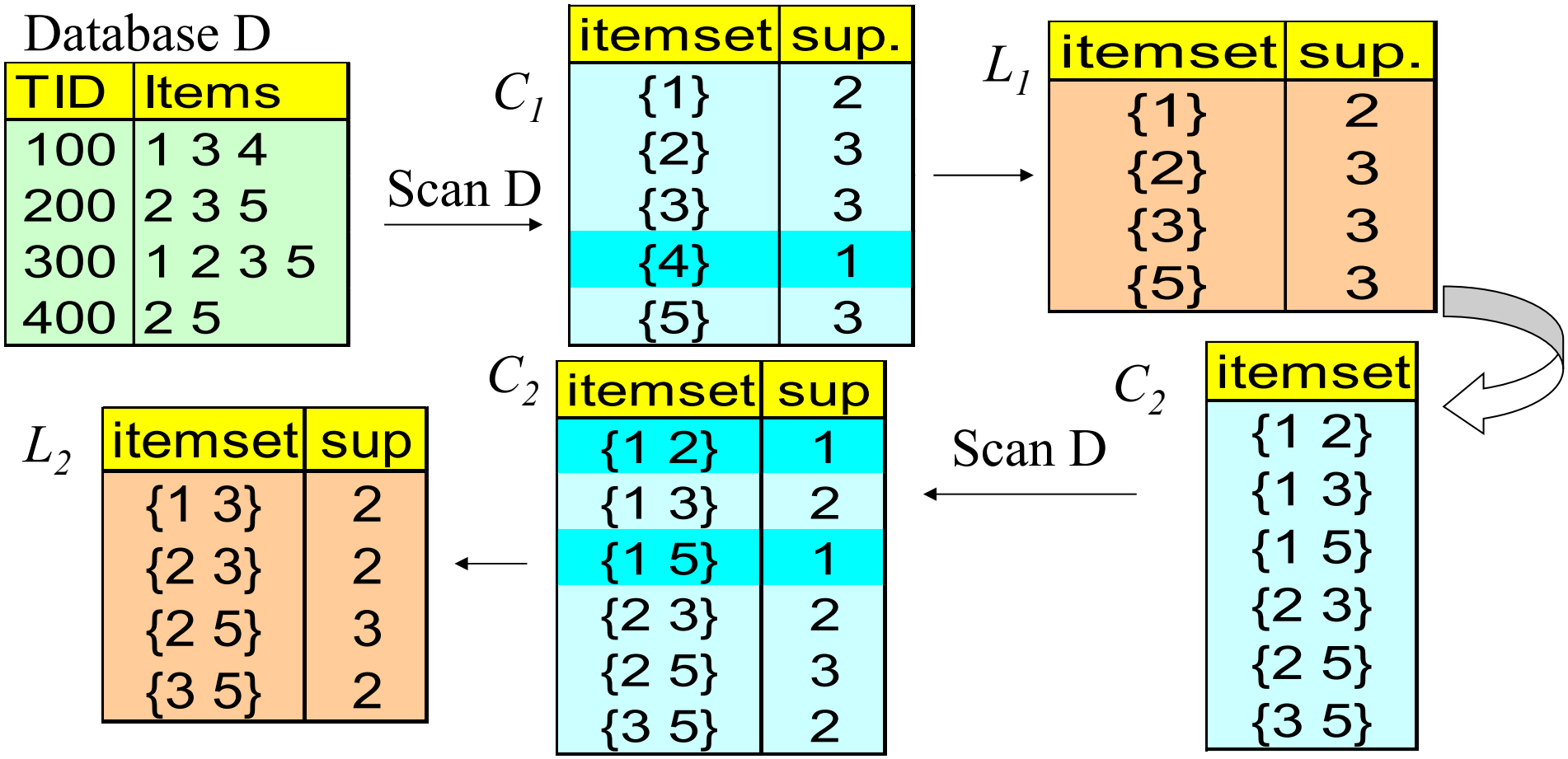
The Apriori Algorithm — Example



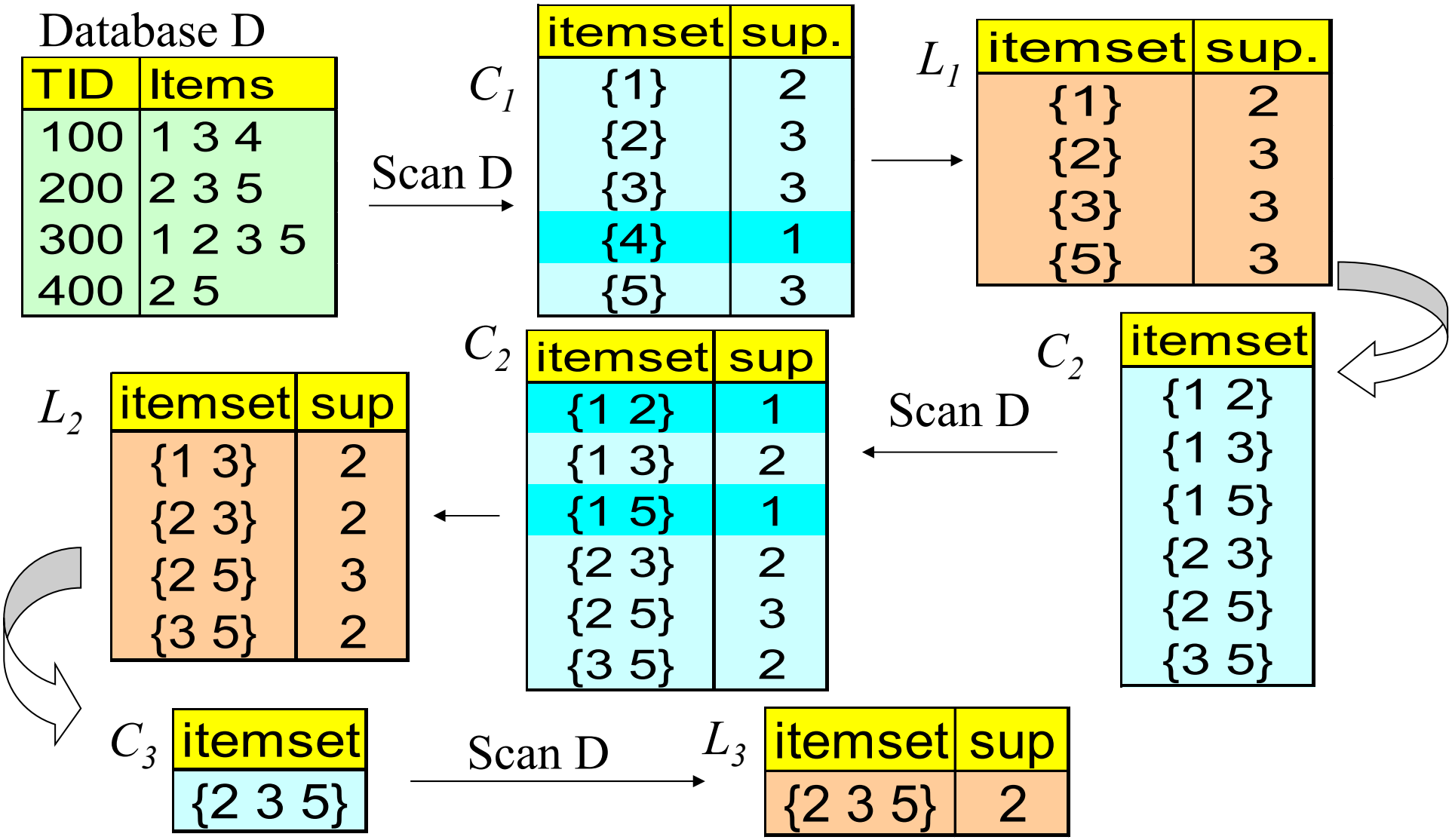
The Apriori Algorithm — Example



The Apriori Algorithm — Example



The Apriori Algorithm — Example



Step 1: The Apriori Algorithm (Find frequent itemsets)

- Notations:

  C_k: Candidate itemsets of size k (We keep a support count for each candidate)

  L_k : Frequent itemsets of size k
- Join Step: C_k is generated by joining $L_{k-1}$with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- **Pseudo-code:**

L_1 = {frequent items};
for ($k = 1; L_k \neq \emptyset; k++$) do begin
  C_{k+1} = candidates generated from L_k; // join and prune steps
  for each transaction *t* in database do
    increment the count of all candidates in C_{k+1} that are contained in *t*
  L_{k+1} = candidates in C_{k+1} satisfying *min_support*
  end
return $\cup_k L_k$;

Example of Generating Candidates

- L_3={*abc, abd, acd, ace, bcd*}
- Self-joining: L_3*L_3
  - *abcd* from *abc* and *abd*
  - *acde* from *acd* and *ace*
- Pruning:
  - *acde* is removed because *ade* is not in L_3
- C_4={*abcd*}

How to Generate Candidates?

- Suppose the items in L_{k-1} are listed in an order
- Step 1: self-joining L_{k-1}

  insert into $\boldsymbol{C_k}$

  select $\boldsymbol{p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}}$

  from $\boldsymbol{L_{k-1}\ p, L_{k-1}\ q}$

  where $\boldsymbol{p.item_1=q.item_1, \ldots, p.item_{k-2}=q.item_{k-2}, p.item_{k-1} < q.item_{k-1}}$
- Step 2: pruning

  forall ***itemsets c in*** $\boldsymbol{C_k}$ do

    forall ***(k-1)-subsets s of c*** do

      if *(s is not in* L_{k-1}*)* **then delete** *c* **from** C_k

Step 2: Generating Association Rules from Frequent Itemsets

$$\text{confidence}(A \Rightarrow B)$$
$$= P(B \mid A)$$
$$= \frac{\text{support\_count}(A \cap B)}{\text{support\_count}(A)}$$

1. For each frequent itemset *l*, generate all nonempty subsets of *l*
2. For every nonempty subset *s* of *l*,
   output the rule "*s* $\Rightarrow$ *(l-s)*"
   if support_count(*l*)/support_count(*s*) $\geq$ minimum-confidence

Step 2: Generating Association Rules from Frequent Itemsets

- Consider the frequent itemset {2, 3, 5}
- The candidate association rules are:

2 ^ 3 => 5 confidence = 100%

2 ^ 5 => 3 confidence = 66.7%

3 ^ 5 => 2 confidence = 100%

2 => 3 ^ 5 confidence = 66.7%

3 => 2 ^ 5 confidence = 66.7%

5 => 2 ^ 3 confidence = 66.7%

- If the minimum confidence requirement is 90%, the resulting association rules are:

2 ^ 3 => 5 confidence = 100%

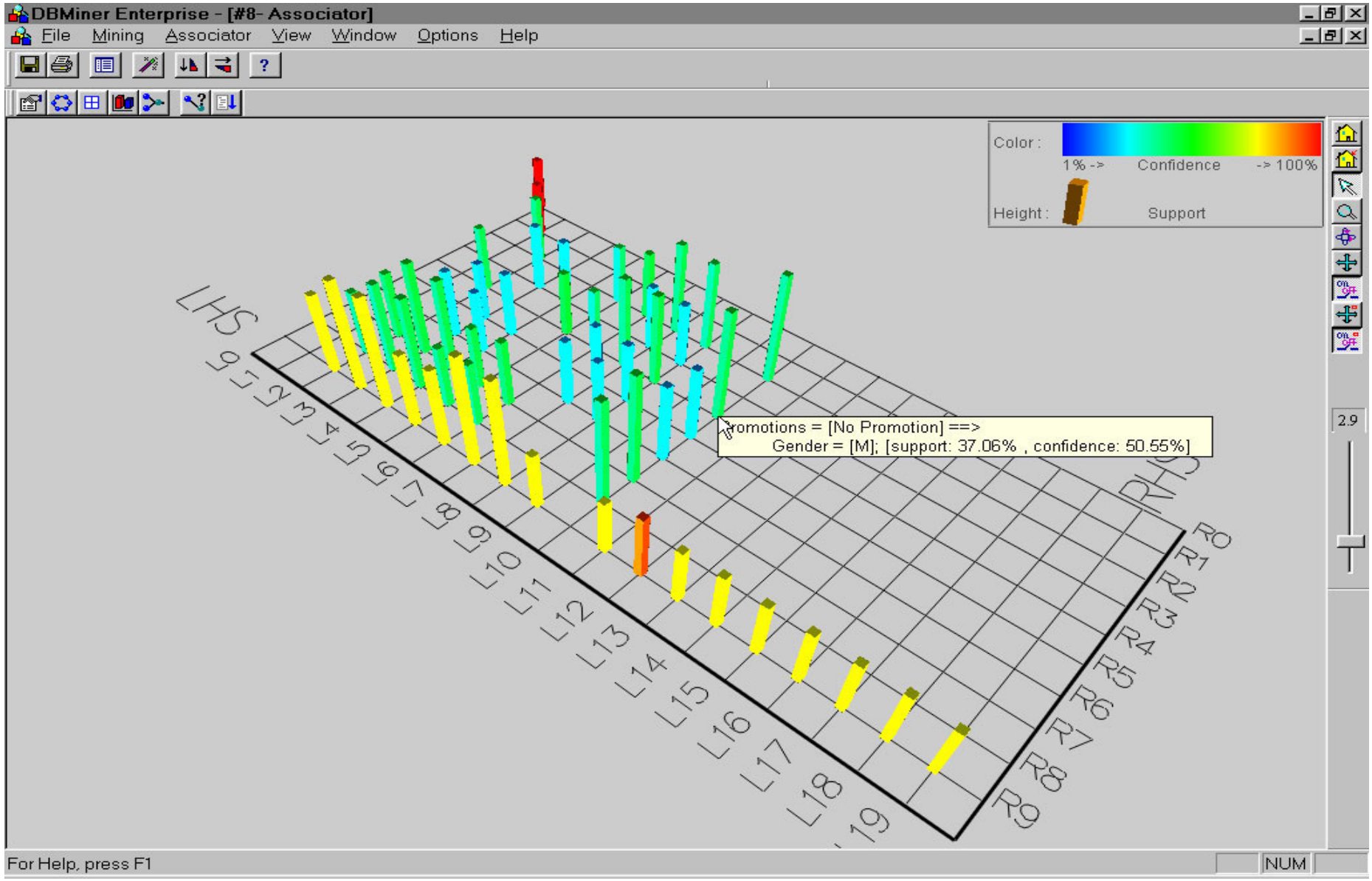
3 ^ 5 => 2 confidence = 100%

Presentation of Association Rules (Table Form)

	Body	Implies	Head	Supp (%)	Conf (%)	F	G	H	I
1	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00'	28.45	40.4				
2	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00'	20.46	29.05				
3	cost(x) = '0.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	59.17	84.04				
4	cost(x) = '0.00~1000.00'	==>	revenue(x) = '1000.00~1500.00'	10.45	14.84				
5	cost(x) = '0.00~1000.00'	==>	region(x) = 'United States'	22.56	32.04				
6	cost(x) = '1000.00~2000.00'	==>	order_qty(x) = '0.00~100.00'	12.91	69.34				
7	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '0.00~500.00'	28.45	34.54				
8	order_qty(x) = '0.00~100.00'	==>	cost(x) = '1000.00~2000.00'	12.91	15.67				
9	order_qty(x) = '0.00~100.00'	==>	region(x) = 'United States'	25.9	31.45				
10	order_qty(x) = '0.00~100.00'	==>	cost(x) = '0.00~1000.00'	59.17	71.86				
11	order_qty(x) = '0.00~100.00'	==>	product_line(x) = 'Tents'	13.52	16.42				
12	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	23.88				
13	product_line(x) = 'Tents'	==>	order_qty(x) = '0.00~100.00'	13.52	98.72				
14	region(x) = 'United States'	==>	order_qty(x) = '0.00~100.00'	25.9	81.94				
15	region(x) = 'United States'	==>	cost(x) = '0.00~1000.00'	22.56	71.39				
16	revenue(x) = '0.00~500.00'	==>	cost(x) = '0.00~1000.00'	28.45	100				
17	revenue(x) = '0.00~500.00'	==>	order_qty(x) = '0.00~100.00'	28.45	100				
18	revenue(x) = '1000.00~1500.00'	==>	cost(x) = '0.00~1000.00'	10.45	96.75				
19	revenue(x) = '500.00~1000.00'	==>	cost(x) = '0.00~1000.00'	20.46	100				
20	revenue(x) = '500.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	19.67	96.14				
21									
22									
23	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
24	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
25	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
26	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
27	cost(x) = '0.00~1000.00' AND order_qty(x) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	33.23				

Sheet1

Visualization of Association Rule Using Plane Graph



Visualization of Association Rule Using Rule Graph