David Eccles School of Business

THE UNIVERSITY OF UTAH

Association Rule Mining

Data Mining for Business Intelligence

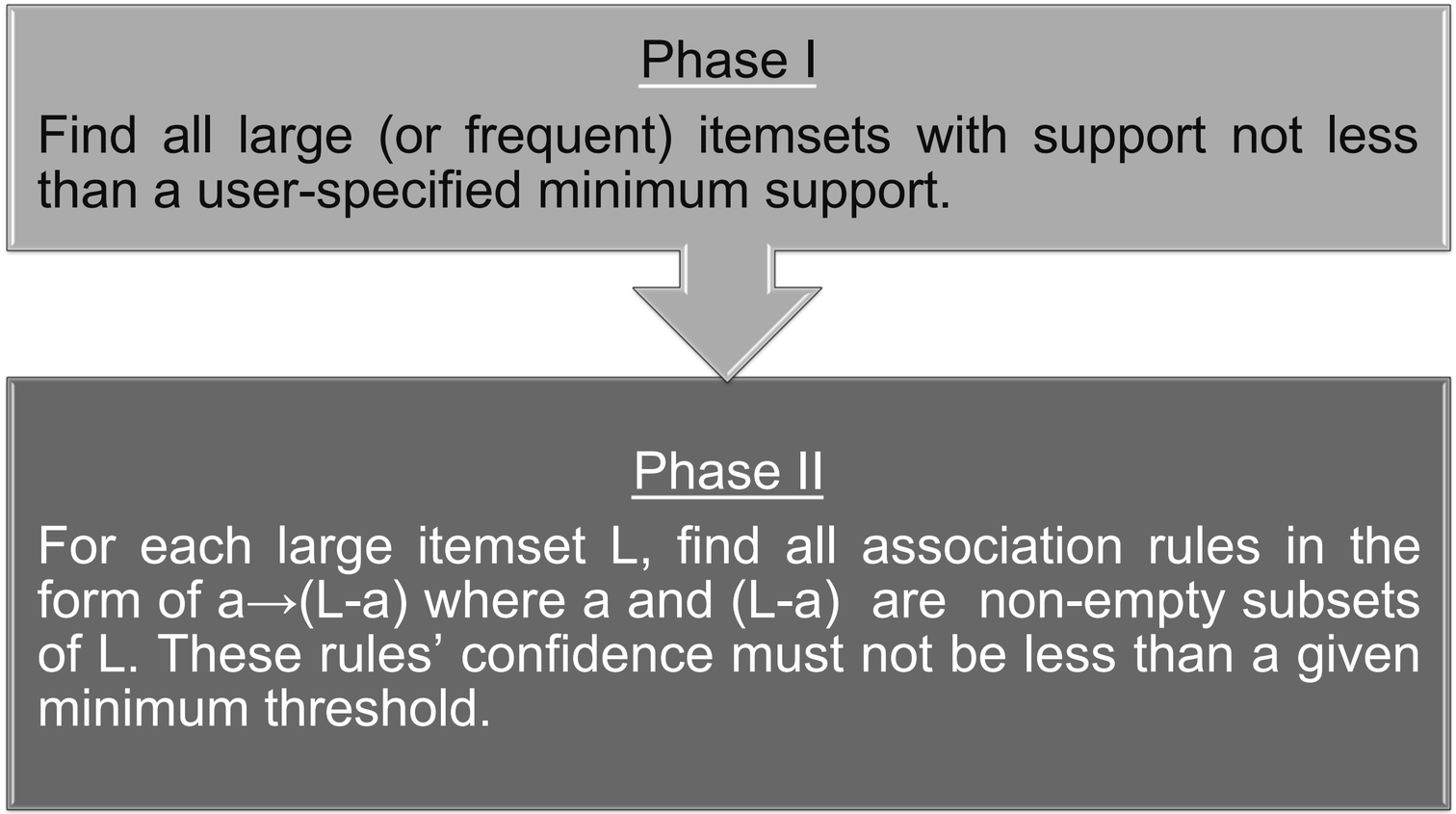
AGENDA

- Motivating examples and main approach
- Set basics and important metrics
  - Support, confidence and lift
- Apriori property and Apriori algorithm
- Data preparation decisions
- Extension of association rules

ASSOCIATION RULE MINING

- Association rule mining: find all association rules with support and confidence not less than the user-specified minimum support and confidence levels in the DB.
- For small problems, the process of mining association rules is not that complex.
- How about a transaction database with 1 billion transactions and 1 million different items?
- An efficient algorithm is needed!

PHASES IN ASSOCIATION RULE MINING



Focus is on I

II is straightforward

E.g. Find all association rules in the example with 60% or more support and 80% or more confidence.
Large itemsets: Itemsets that have a support not less than the specified threshold.

APRIORI ALGORITHM FOR PHASE I

- An efficient algorithm to discover all large itemsets from a huge database with large number of items.
- Developed by two researchers from IBM Almaden Research Lab.
- Based on the ***Apriori property***

APRIORI PROPERTY

Apriori property: "Any subset of a large itemset must be large"

TID	Items
100	1,3,4,6
200	2,3,5,7
300	1,2,3,5,8
400	2,5,9,10
500	1,4

- Minimum support =40%

{2,3,5} is large →

{2}, {3}, {5}, {2,3}, {2,5} and {3,5} must be large

How can we use this property to mine large itemsets?

APRIORI ALGORITHM – Phase 1

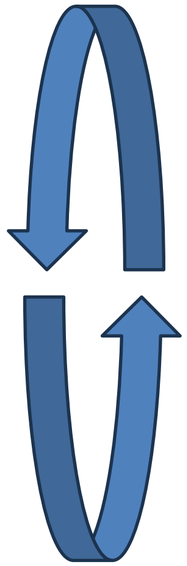
Initiation

Step 1: Set k = 1. Scan DB one time to find all large 1- itemsets.

Step 2: Increment k. Generate candidate K-itemsets from large (k-1)-itemsets.

Step 3: Filter out non-large k-itemsets from candidate k-itemsets based on their support levels (another DB scan)

Iteration



Go back to step 2 if candidate k-itemsets exist.

Termination

Stop when no more candidate itemsets can be generated.

PHASE 1: STEP 2 OF APRIORI ALGORITHM

- Candidate k-itemsets are k-itemsets that could be large.
- Why generate candidate k-itemsets only from large (k-1) itemsets?

PHASE 1: STEP 2,3 OF APRIORI ALGORITHM

Itemsets: L1,L2

L1(n): the n-th element in itemset L1.
L2(n): the n-th element in itemset L2

e.g if L1 = {eggs,milk,bread} then L1(3) = bread
e.g if L2 = {eggs,milk,cheese} then L1(3) = cheese

L1(1) = L2(1) and L1(2) = L2(2) and L1(3) ≠ L2(3)

- Step 2: Join:
  - If k=2, simply merge every two unique 1-itemsets into a 2-itemset. Else ->
  - If k>2
    - Sort all items in the large (k-1)-itemsets
    - Find and merge any two joinable (k-1)-itemsets, L1 and L2
    - Joinable? Two large (k-1)-itemsets, L1 and L2, that are joinable must satisfy the following conditions:
      - L1(1)=L2(1) and L1(2)=L2(2) and …. L1(k-2)=L2(k-2) L1(k-1) ≠ L2(k-1)
- Step 3: Prune: prune non-large itemsets generated in step 2

EXAMPLE: APRIORI ALGORITHM

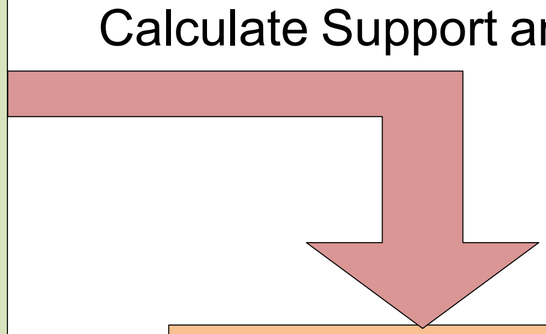
TID	Items
100	1,3,4,6
200	2,3,5,7
300	1,2,3,5,8
400	2,5,9,10
500	1,4

- Minimum support =40%
- Minimum confidence =70%

EXAMPLE: APRIORI ALGORITHM

Tid	Items
100	1, 3, 4, 6
200	2, 3, 5, 7
300	1, 2, 3, 5, 8
400	2, 5, 9, 10
500	1, 4

Minimum Support: 40%



Large 1-itemset:

{1} support=3/5=60%

{2} support=3/5=60%

{3} support=3/5=60%

{4} support=2/5=40%

{5} support=3/5=60%

EXAMPLE: APRIORI ALGORITHM

Large 1-itemset:

{1} support=3/5=60%

{2} support=3/5=60%

{3} support=3/5=60%

{4} support=2/5=40%

{5} support=3/5=60%

Generate Itemset

Candidate 2-itemset:

{1, 2} {1, 3} {1, 4} {1, 5}

{2, 3} {2, 4} {2, 5}

{3, 4} {3, 5}

{4, 5}

EXAMPLE: APRIORI ALGORITHM

Minimum Support: 40%

Candidate 2-itemset:

{1, 2} {1, 3} {1, 4} {1, 5}
{2, 3} {2, 4} {2, 5}
{3, 4} {3, 5}
{4, 5}

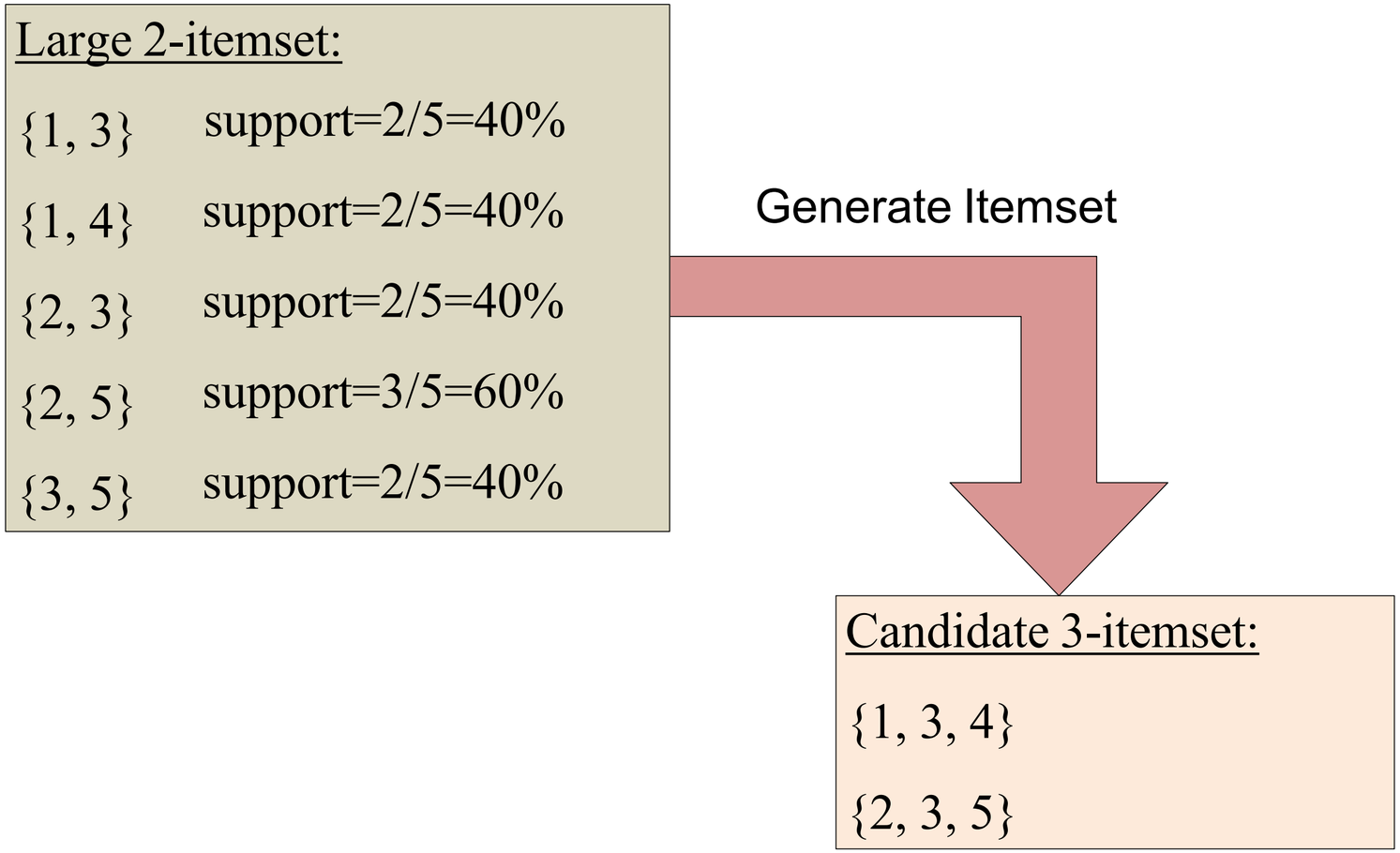
Calculate Support and Prune

Tid	Items
100	1, 3, 4, 6
200	2, 3, 5, 7
300	1, 2, 3, 5, 8
400	2, 5, 9, 10
500	1, 4

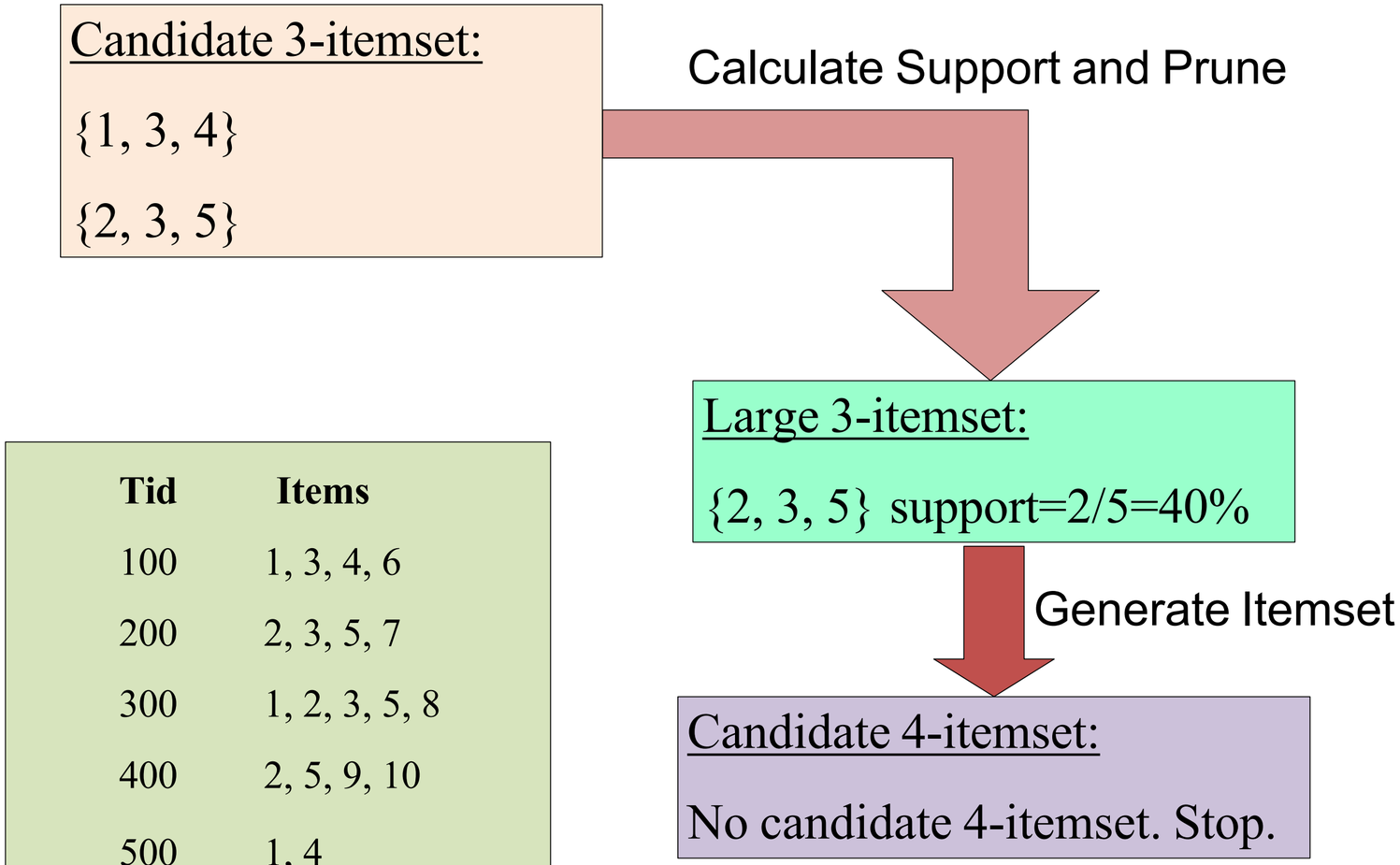
Large 2-itemset:

{1, 3} support=2/5=40%
{1, 4} support=2/5=40%
{2, 3} support=2/5=40%
{2, 5} support=3/5=60%
{3, 5} support=2/5=40%

EXAMPLE: APRIORI ALGORITHM



EXAMPLE: APRIORI ALGORITHM



EXAMPLE: GENERATE ASSOCIATION RULES

Large 2- and 3-itemsets:

{1, 3} support=2/5=40%

{1, 4} support=2/5=40%

{2, 3} support=2/5=40%

{2, 5} support=3/5=60%

{3, 5} support=2/5=40%

{2, 3, 5} support =2/5=40%

Candidate rules:

1->3, 3-> 1, 1-> 4, 4->1, 2-> 3, 3-> 2, 2->5, 5->2, 3->5, 5->3

2->3,5 2,3->5 2,5->3 3->2,5 3,5->2 5-> 2,3

Large 1-itemset:

{1} support=3/5=60%

{2} support=3/5=60%

{3} support=3/5=60%

{4} support=2/5=40%

{5} support=3/5=60%

Is Confidence >= 70%?

Association rules:

4->1 (100% confidence)

2->5 (100% confidence)

5->2 (100% confidence)

2,3->5 (100% confidence)

3,5->2 (100%) confidence

confidence(X -> Y) = support(X U Y) /support(X).

confidence(4 ->1) = support({1,4})/support({4}) = .4/.4 = 1.0 or 100%
confidence(3 ->2,5) = support({2,3,5})/support({3}) = .4/.6 ~ .67 or 67%
confidence(5 ->2,3) = support({2,3,5})/support({5}) = .4/.6 ~ .67 or 67%
confidence(2,3 ->5) = support({2,3,5})/support({2,3}) = .4/.4 ~ 1.0 or 100%

EXAMPLE: GENERATE ASSOCIATION RULES (PHASE II)

	Large (Frequent) Item Sets (Support >= 0.2)						
trans id	egg-a milk-b	egg-a bread-b	egg-a soda-b	bread-a milk-b	bread-a soda-b	egg-a milk-b bread-c	egg-a bread-b soda-c
1	1	0	0	0	0	0	0
2	1	1	0	1	0	1	0
3	0	0	0	0	0	0	0
4	0	1	1	0	1	0	1
5	0	0	0	1	0	0	0
Support	0.4	0.4	0.2	0.4	0.2	0.2	0.2
Rules	**Confidence (Min level >= 0.6)**						
a->b or a,b->c	0.667	0.667	0.333	0.500	0.250	0.500	0.500
b>a or b,c->a	0.667	0.500	1.000	0.667	1.000	0.500	1.000
a,c->b	NA					0.500	1.000
a->b,c						0.333	0.333
b->a,c						0.333	0.250
c->a,b						0.625	1.000

David Eccles
School of Business
THE UNIVERSITY OF UTAH

David Eccles School of Business

THE UNIVERSITY OF UTAH

Association Rule Mining

Data Mining for Business Intelligence

AGENDA

- Motivating examples and main approach
- Set basics and important metrics
  - Support, confidence and lift
- Apriori property and Apriori algorithm
- Data preparation decisions
- Extension of association rules

DATA PREPARATION DECISIONS

"Unit of analysis" decision

"Item" decision:
product, brand-product, brand or category

trans id	egg	milk	bread	soda
1	1	1	0	0
2	1	1	1	0
3	0	0	1	0
4	1	0	1	1
5	0	1	1	0

cust id	egg	milk	bread	soda
1	1	1	0	0
2	1	1	1	0
3	1	0	1	1
4	0	1	1	0

date	egg	milk	bread	soda
11/1/2010	1	1	1	0
11/2/2010	1	1	1	1

Decisions depend on applications of rules

EXTENSIONS OF ASSOCIATION RULES

Quantitative association rules	• Consider the quantity of an item in a transaction (e.g. 1 Egg -> 1 Milk, 2 Egg -> 3 Milk)
Hierarchical association rules	• Considers multiple item levels – e.g., product subcategory and category, e.g., • (category) frozen items ➔ dairy items • (subcategory) frozen vegetables ➔ soymilk
Sequential patterns	• Group items by transactions, transactions by customers • E.g., a customer buys Spiderman tends to buy Spiderman 2 at the next visit (transaction)
Inter-transaction association rules	• If the stock price of Microsoft goes down, the stock price of Sun (or IBM) tends to go up next

David Eccles
School of Business
THE UNIVERSITY OF UTAH