

# Data Mining: Association Analysis

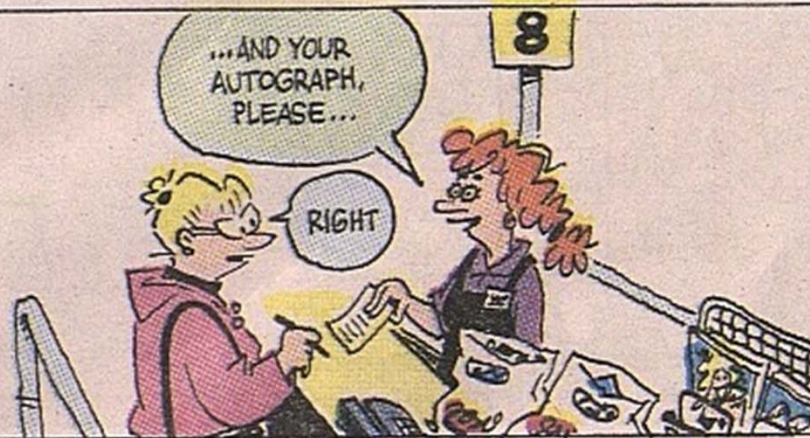
Tom Heskes





# Betty

By Delaney & Rasmussen



© 2001 by NEA, Inc. www.cortina.com 11-18





## Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

### Market-Basket transactions

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count ( $\sigma$ )**
  - Frequency of occurrence of an itemset
  - E.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- **Frequent itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Definition: Association Rule

- **Association rule**
  - An implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets
  - Example:  
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$
- **Rule evaluation metrics**
  - **Support (s):**  
Fraction of transactions that contain both  $X$  and  $Y$
  - **Confidence (c):**  
Measures how often items in  $Y$  appear in transactions that contain  $X$

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$$
$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$
$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

## Association Rule Mining Task

- Given a set of transactions  $T$ , the goal of association rule mining is to find all rules having
  - support  $\geq$  *minsup* threshold
  - confidence  $\geq$  *minconf* threshold
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds $\Rightarrow$  **Computationally prohibitive!**

## Association Rule Mining Task

### Example of Rules:

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$  ( $s=0.4$ ,  $c=0.67$ )  
 $\{\text{Milk, Beer}\} \rightarrow \{\text{Diaper}\}$  ( $s=0.4$ ,  $c=1.0$ )  
 $\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}$  ( $s=0.4$ ,  $c=0.67$ )  
 $\{\text{Beer}\} \rightarrow \{\text{Milk, Diaper}\}$  ( $s=0.4$ ,  $c=0.67$ )  
 $\{\text{Diaper}\} \rightarrow \{\text{Milk, Beer}\}$  ( $s=0.4$ ,  $c=0.5$ )  
 $\{\text{Milk}\} \rightarrow \{\text{Diaper, Beer}\}$  ( $s=0.4$ ,  $c=0.5$ )

- All the above rules are binary partitions of the same itemset:  
 $\{\text{Milk, Diaper, Beer}\}$
- Rules originating from the same itemset have **identical support** but can have **different confidence**
- Thus, we may decouple the support and confidence requirements

# Mining Association Rules

- Two-step approach:

1. Frequent itemset generation

Generate all itemsets whose support  $\geq$  minsup

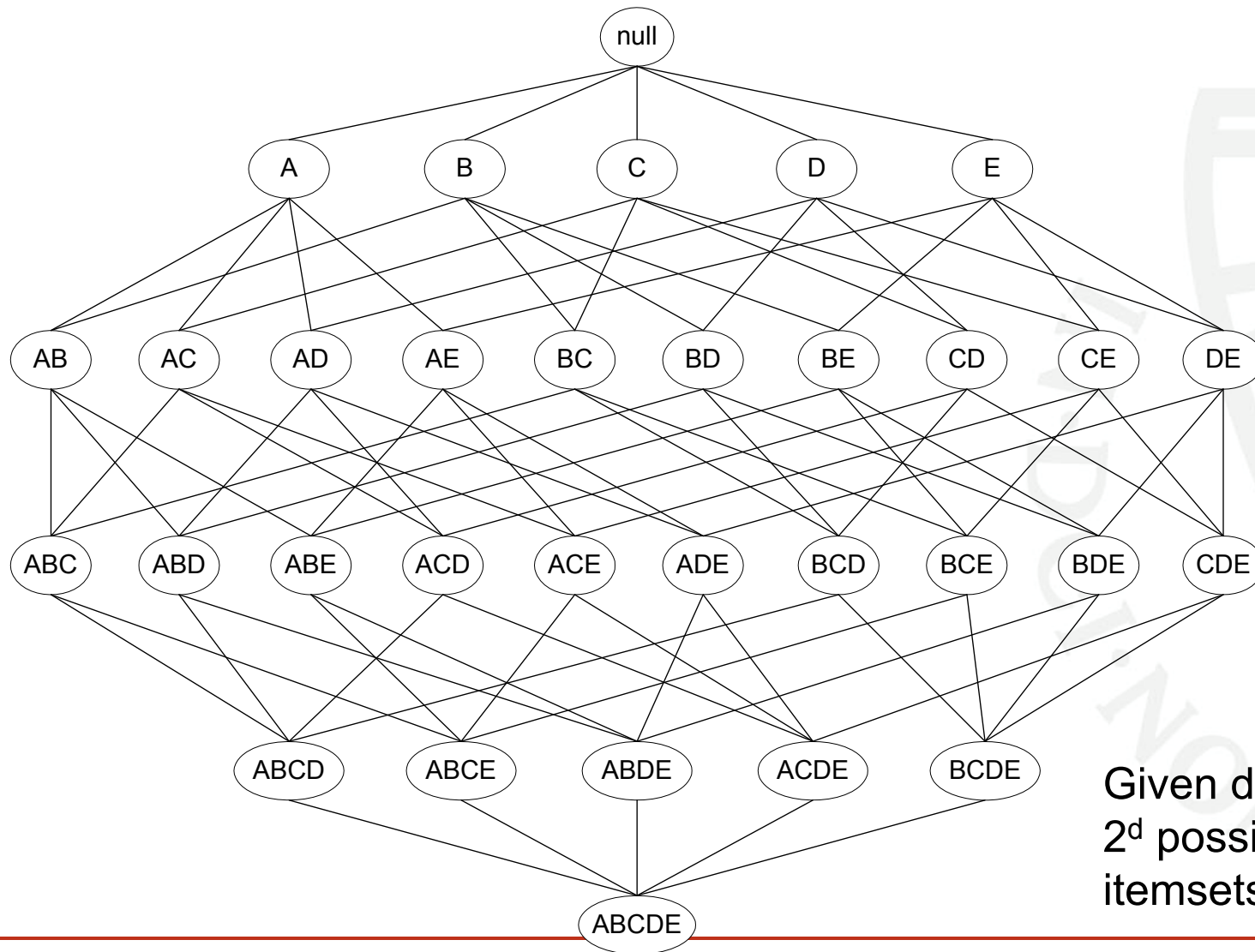
2. Rule generation

Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive



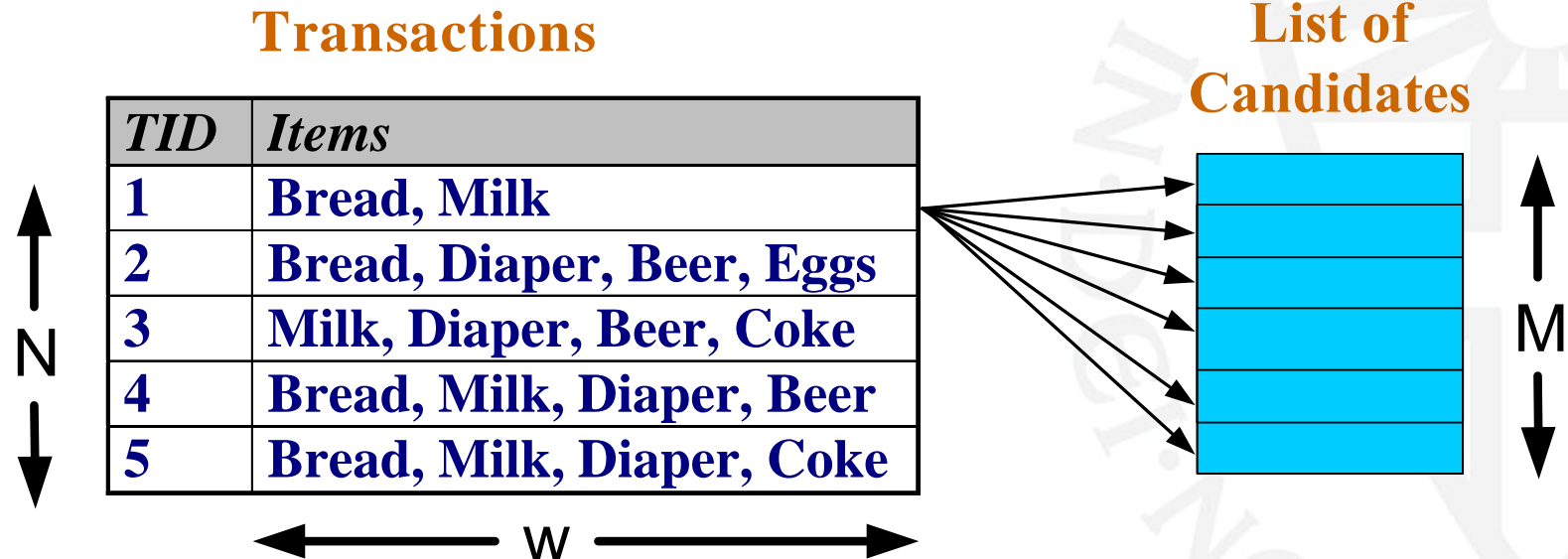
# Frequent Itemset Generation



Given  $d$  items, there are  $2^d$  possible candidate itemsets

## Frequent Itemset Generation

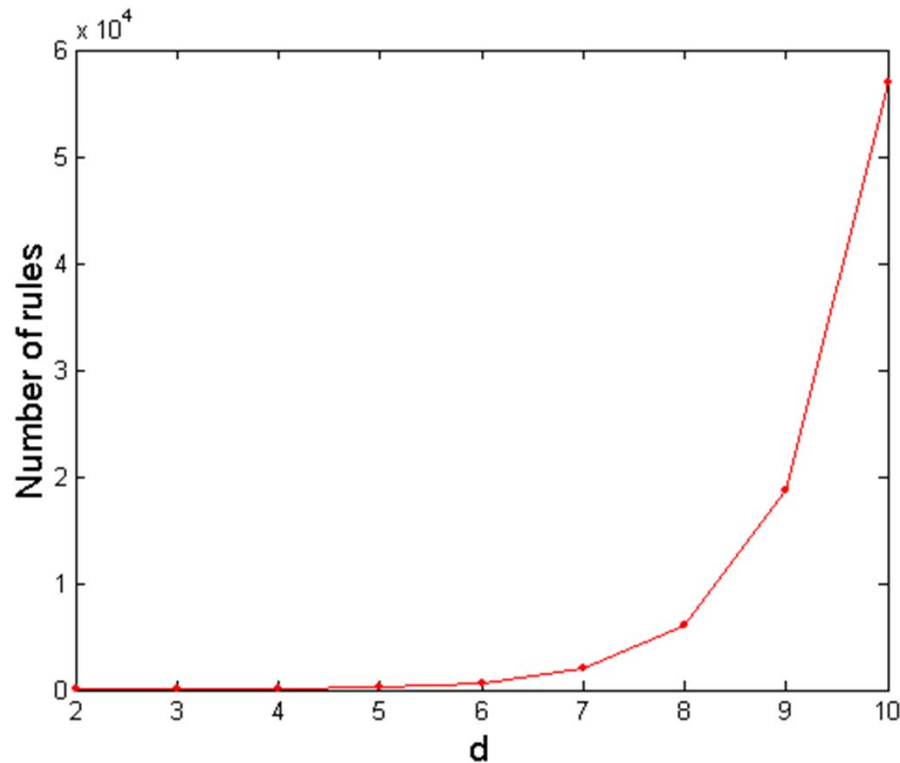
- Brute-force approach:
  - Each itemset in the lattice is a **candidate** frequent itemset
  - Count the support of each candidate by scanning the database



- Match each transaction against every candidate
- Complexity  $\sim O(N M w) \Rightarrow$  **Expensive since  $M = 2^d$  !!!**

# Computational Complexity

- Given  $d$  unique items:
  - Total number of itemsets =  $2^d$
  - Total number of possible association rules:



$$R = \sum_{k=1}^d \left[ \binom{d}{k} \times \sum_{j=1}^{k-1} \binom{k}{j} \right]$$
$$= 3^d - 2^{d+1} + 1$$

If  $d=6$ ,  $R = 602$  rules

Note:  $\{\dots\} \rightarrow \emptyset$  and  $\emptyset \rightarrow \{\dots\}$  not allowed

## Frequent Itemset Generation Strategies

- Reduce the **number of candidates** ( $M$ )
  - Complete search:  $M=2^d$
  - Use pruning techniques to reduce  $M$
- Reduce the **number of transactions** ( $N$ )
  - Reduce size of  $N$  as the size of itemset increases
  - Used by Direct Hash & Pruning and vertical-based mining algorithms
- Reduce the **number of comparisons** ( $N M$ )
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction



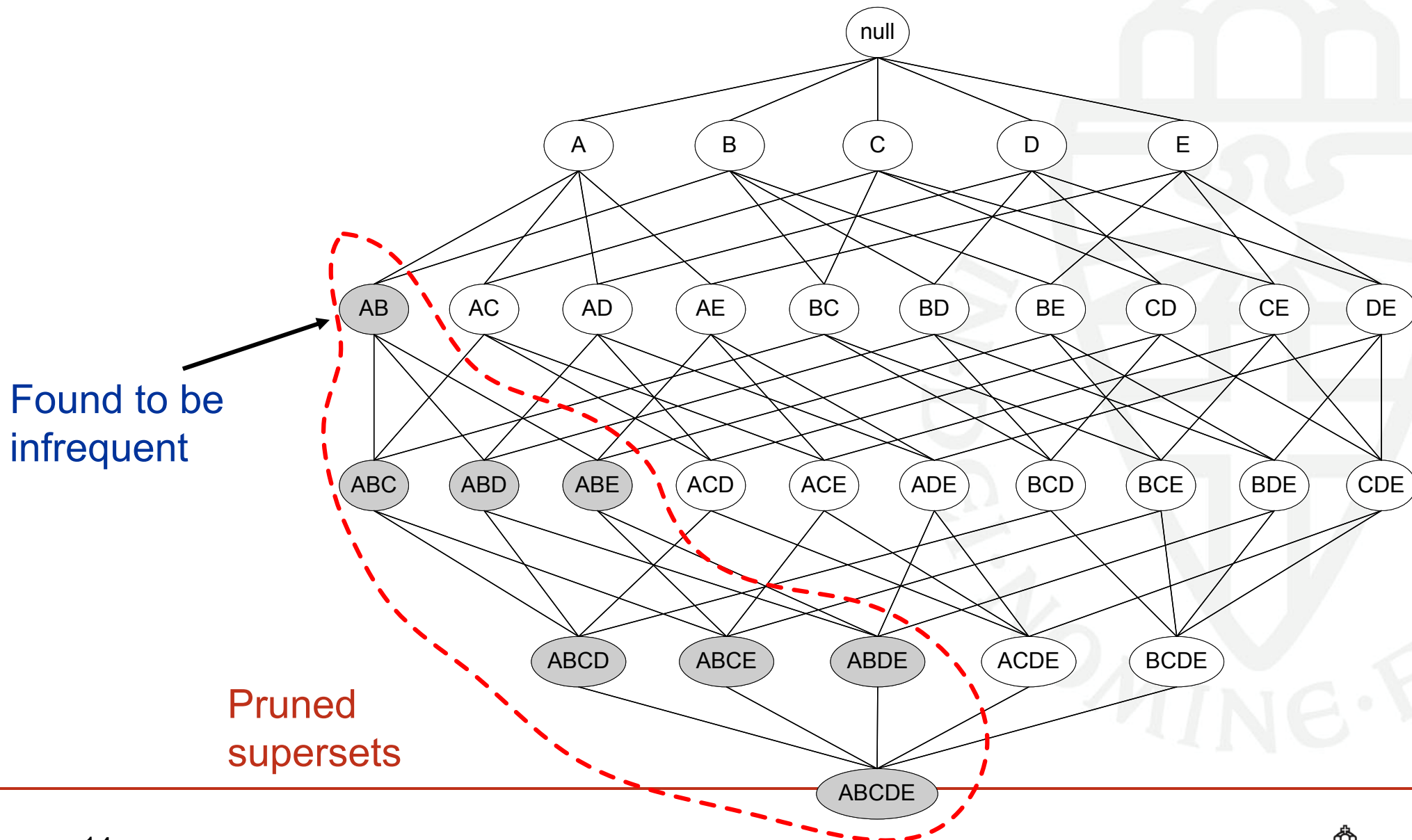
## Reducing Number of Candidates

- **Apriori principle:**
  - If an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

- Support of an itemset never exceeds the support of its subsets
- This is known as the **monotone** property of support

# Illustrating Apriori Principle



## Illustrating Apriori Principle

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1



Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)



Item set	Count
{Bread,Milk,Diaper}	3

Minimum Support = 3

If every subset is considered,

$${}^6C_1 + {}^6C_2 + {}^6C_3 = 41$$

With support-based pruning,

$$6 + 6 + 1 = 13$$

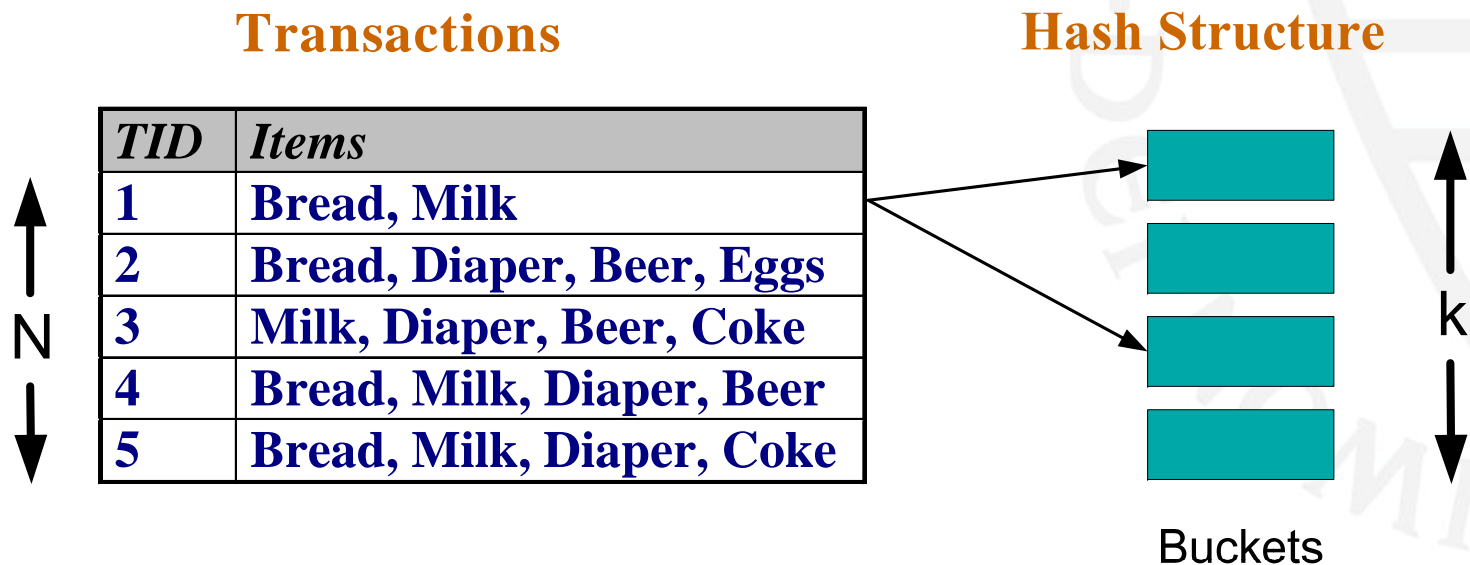
# Apriori Algorithm

- Let  $k=1$
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
  - Generate length  $(k+1)$  candidate itemsets from length  $k$  frequent itemsets
  - Prune candidate itemsets containing subsets of length  $k$  that are infrequent
  - Count the support of each candidate by scanning the DB
  - Eliminate candidates that are infrequent, leaving only those that are frequent



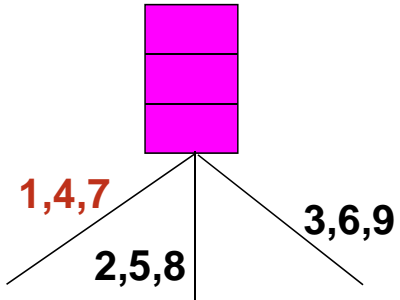
## Reducing Number of Comparisons

- Candidate counting:
  - Scan the database of transactions to determine the support of each candidate itemset
  - To reduce the number of comparisons, store the candidates in a hash structure
  - Instead of matching each transaction against every candidate, match it against candidates contained in the hashed buckets



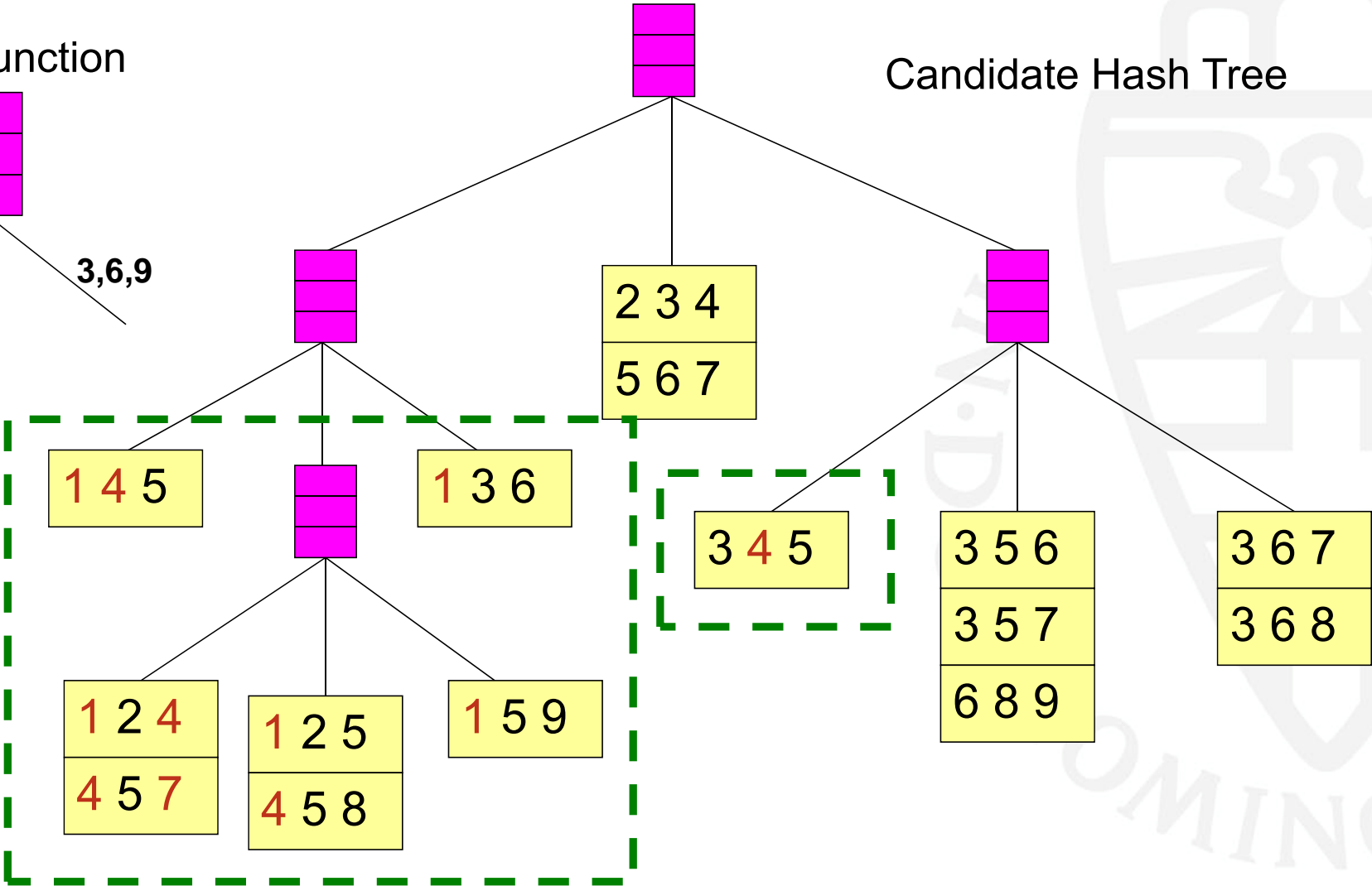
# Generate Hash Tree

Hash Function



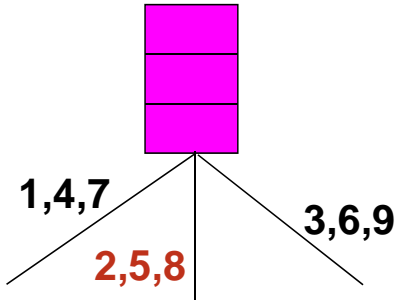
Candidate Hash Tree

Hash on  
1, 4 or 7



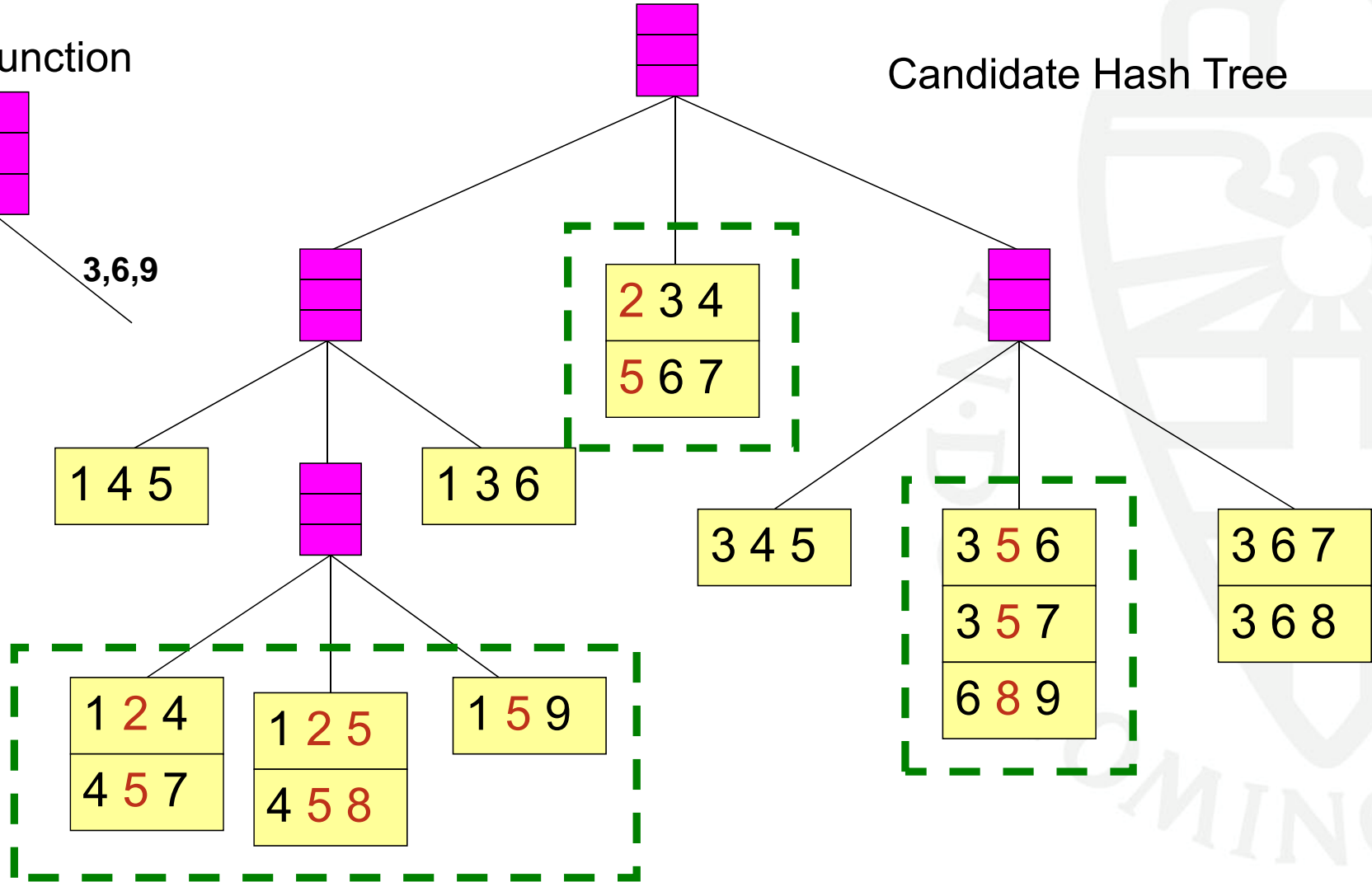
# Generate Hash Tree

Hash Function



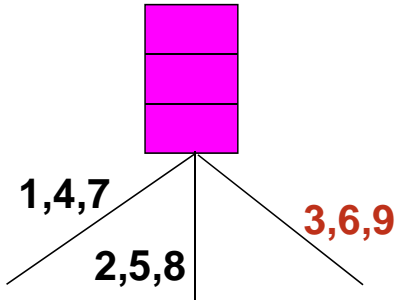
Candidate Hash Tree

Hash on  
2, 5 or 8



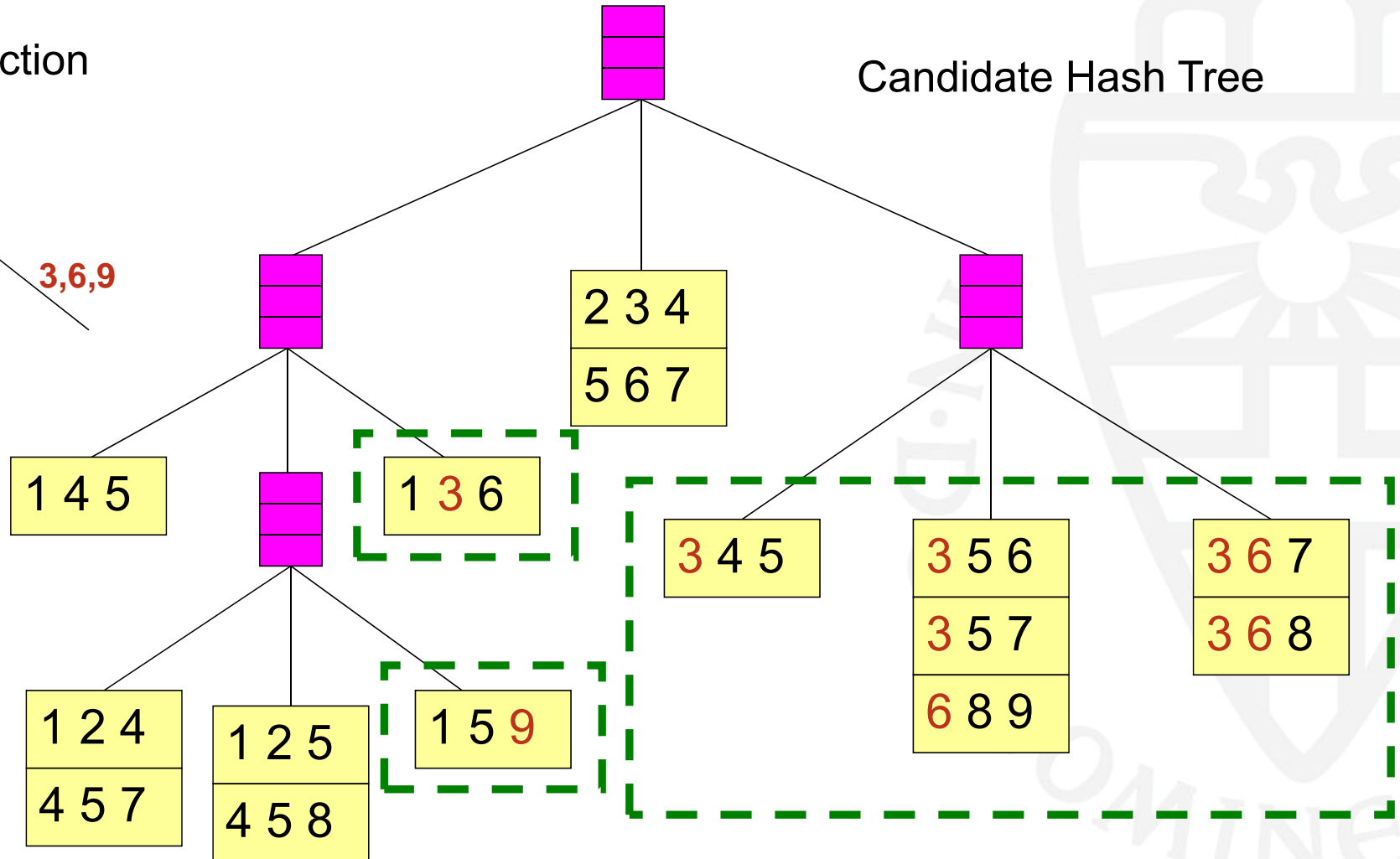
## Generate Hash Tree

Hash Function



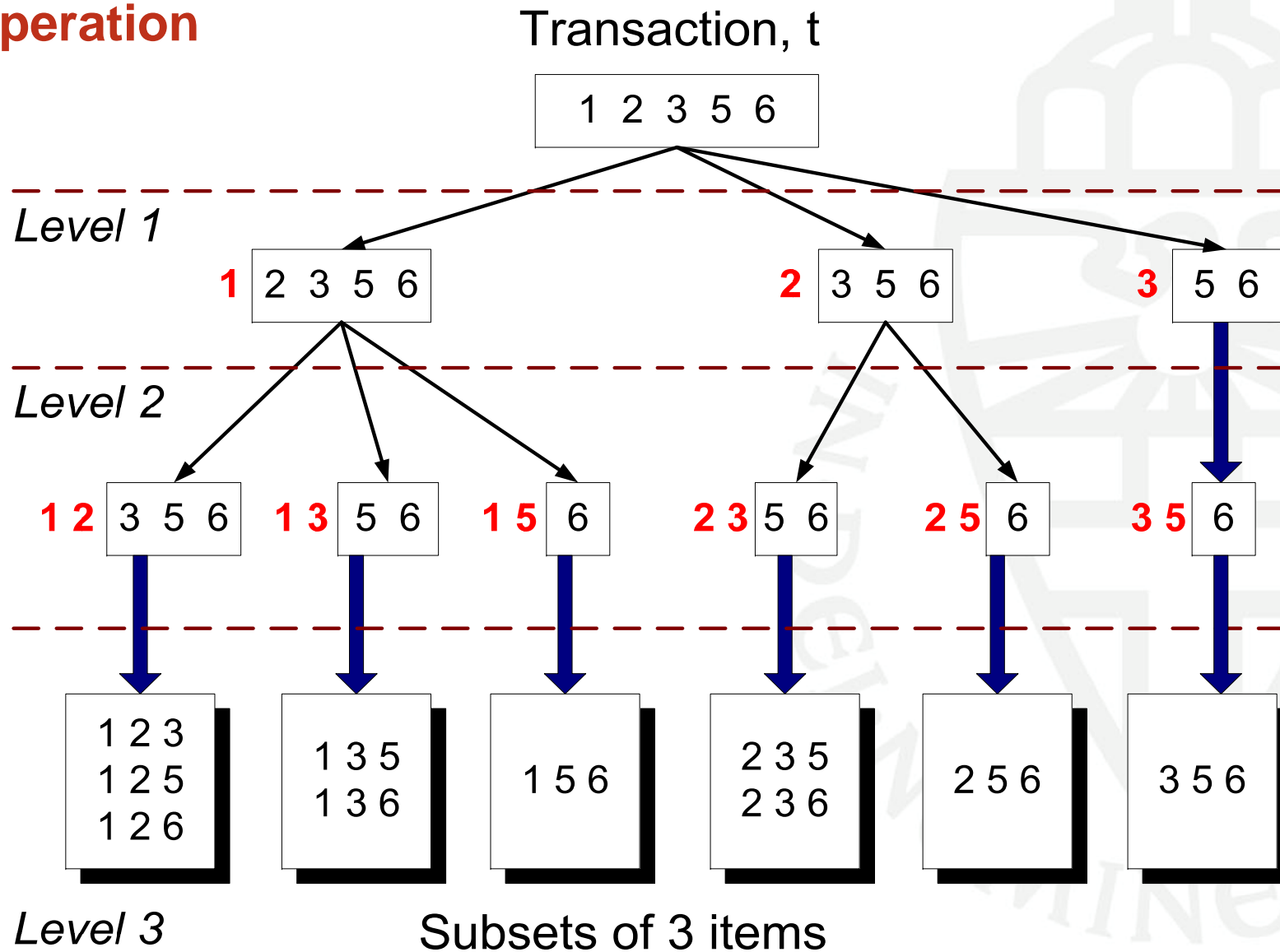
Candidate Hash Tree

Hash on  
3, 6 or 9

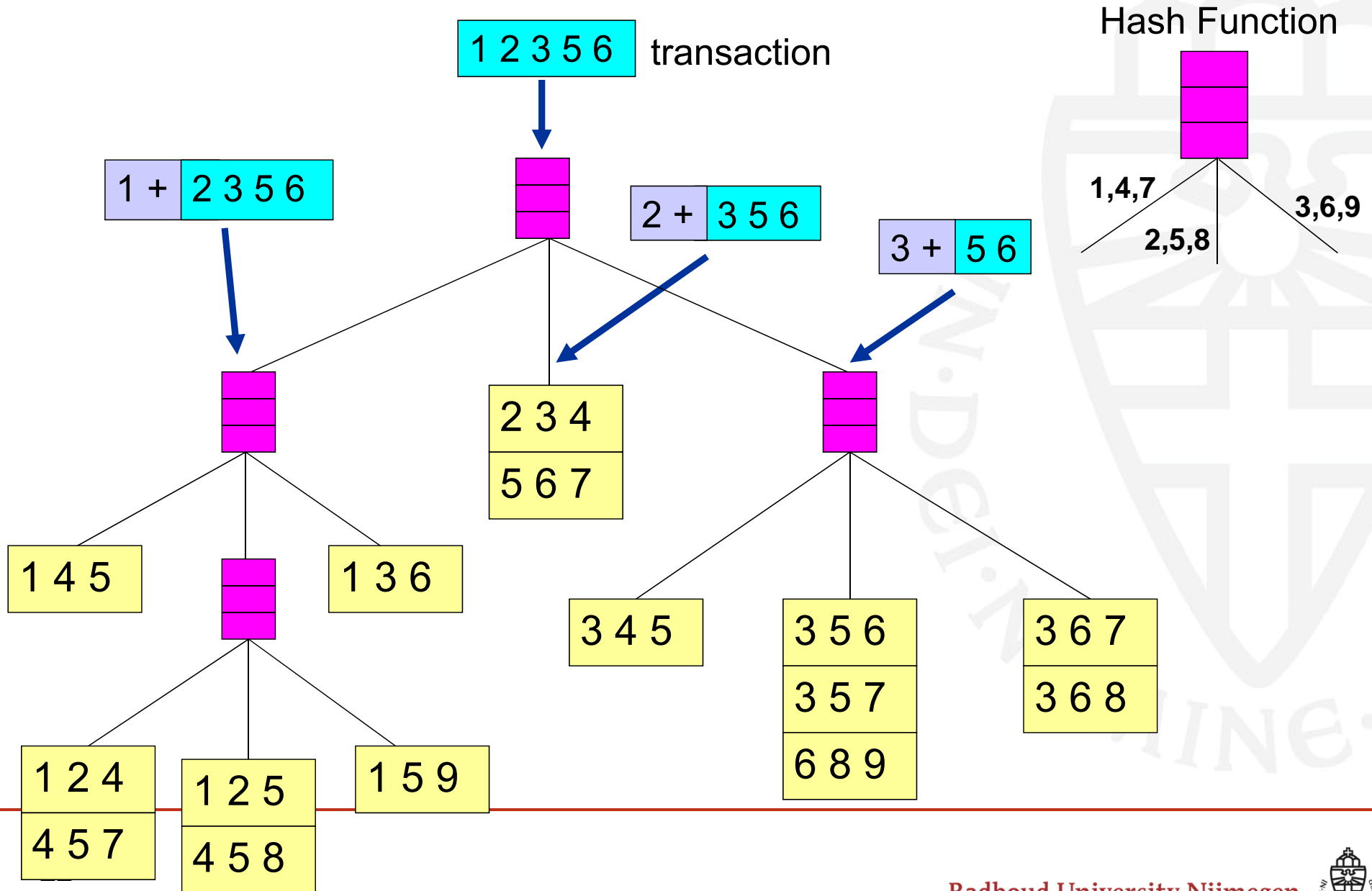




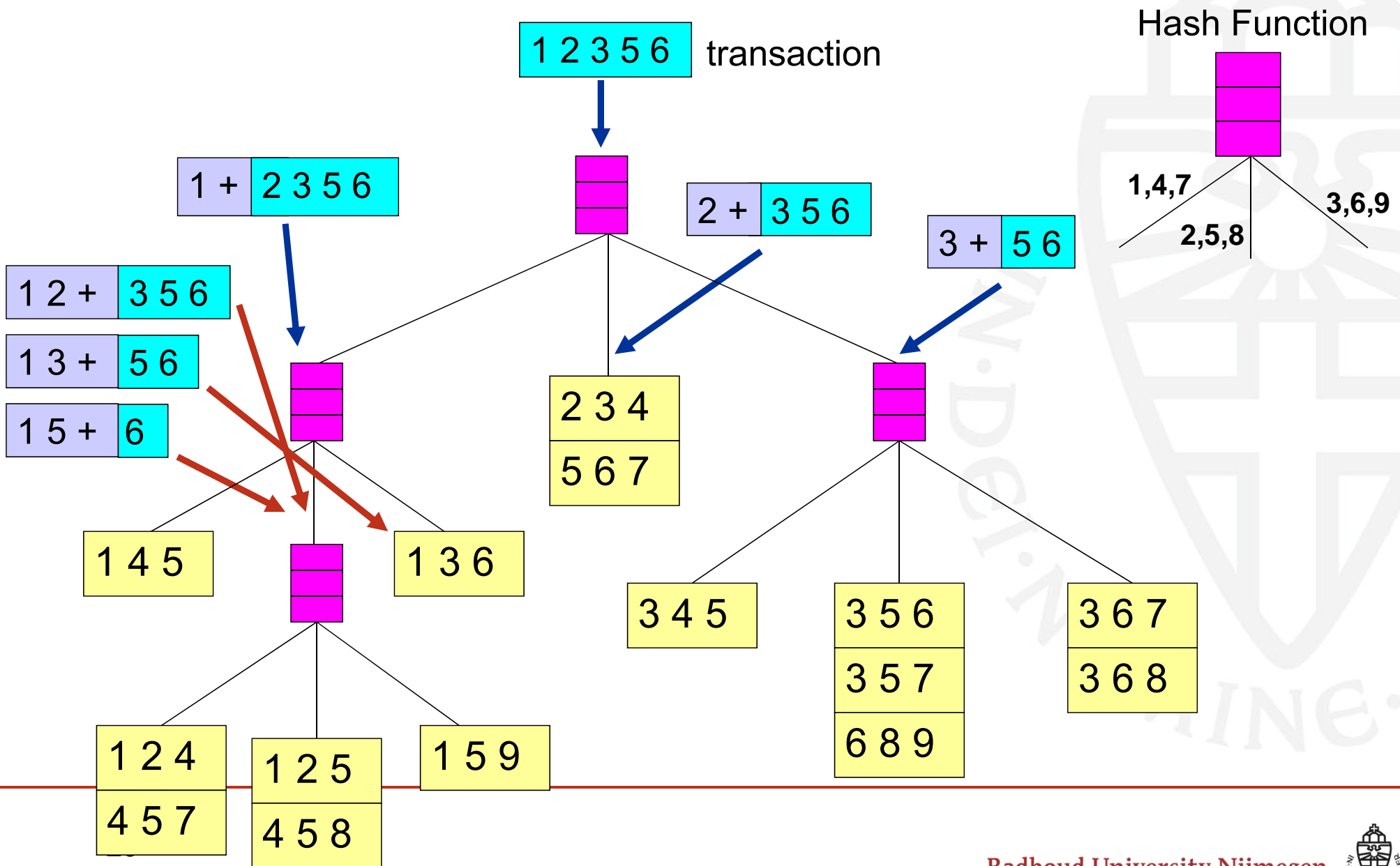
## Subset Operation



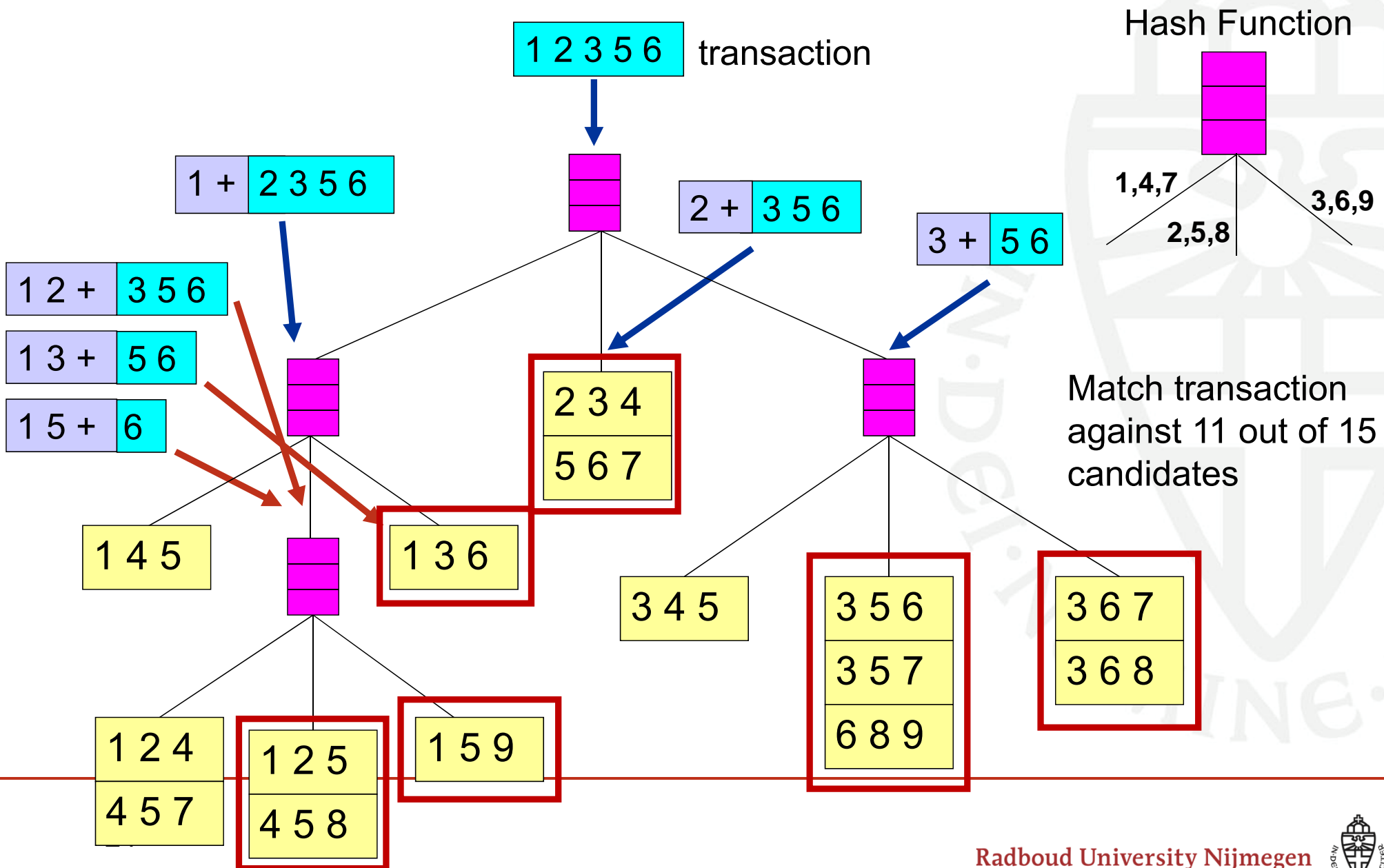
## Subset Operation Using Hash Tree



## Subset Operation Using Hash Tree



# Subset Operation Using Hash Tree





## Factors Affecting Complexity

- Choice of minimum **support threshold**
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (**number of items**) of the data set
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase
- **Size of database**
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions
- **Average transaction width**
  - transaction width increases with denser data sets
  - this may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)

## Compact Representation of Frequent Itemsets

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

- Some itemsets are redundant because they have identical support as their supersets
- Number of frequent itemsets  $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$
- Need a compact representation

## Factors Affecting Complexity

- An itemset is **closed** if none of its immediate supersets has the same support as the itemset
- Compact representation of itemsets without loss of support info

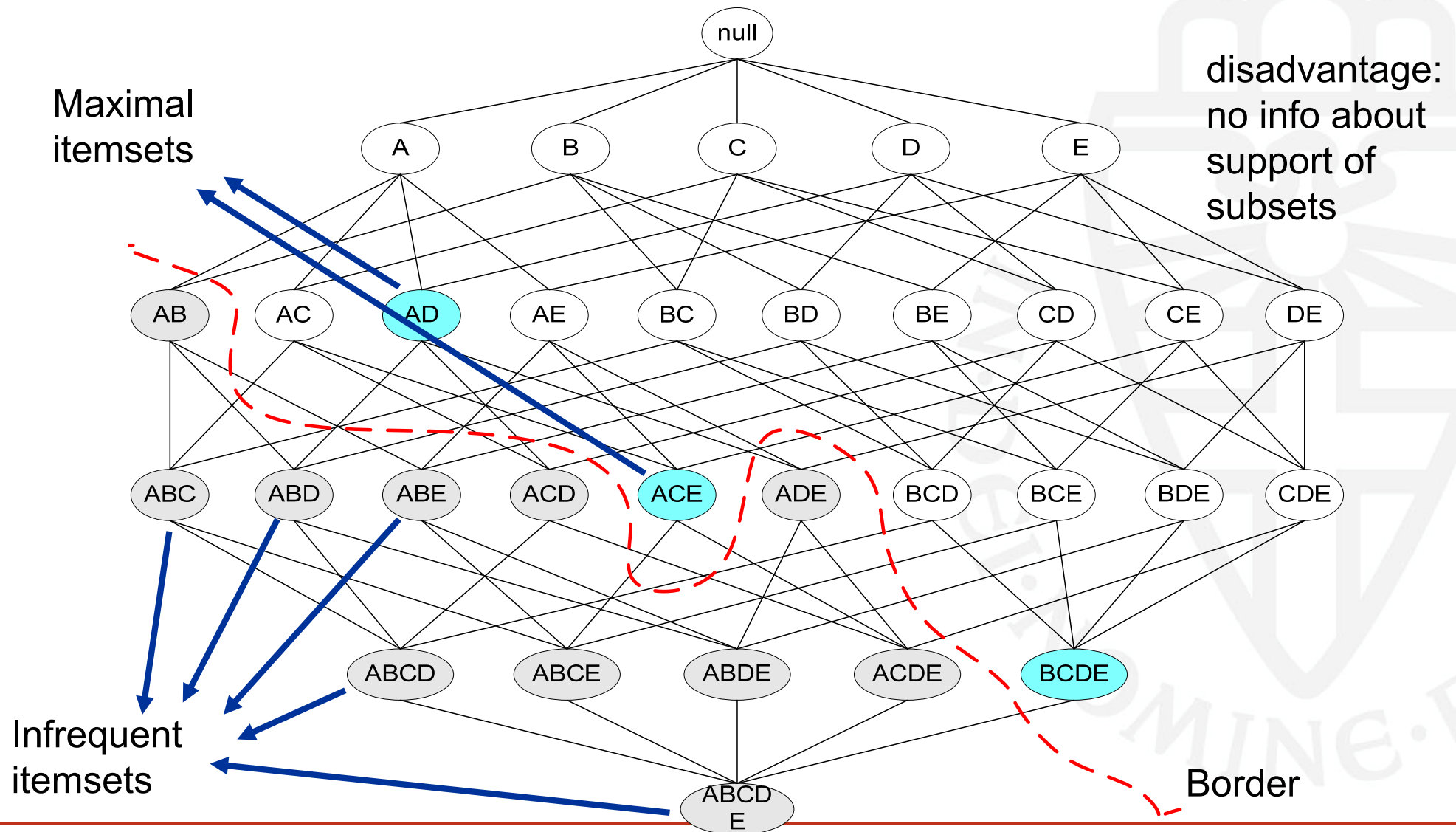
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

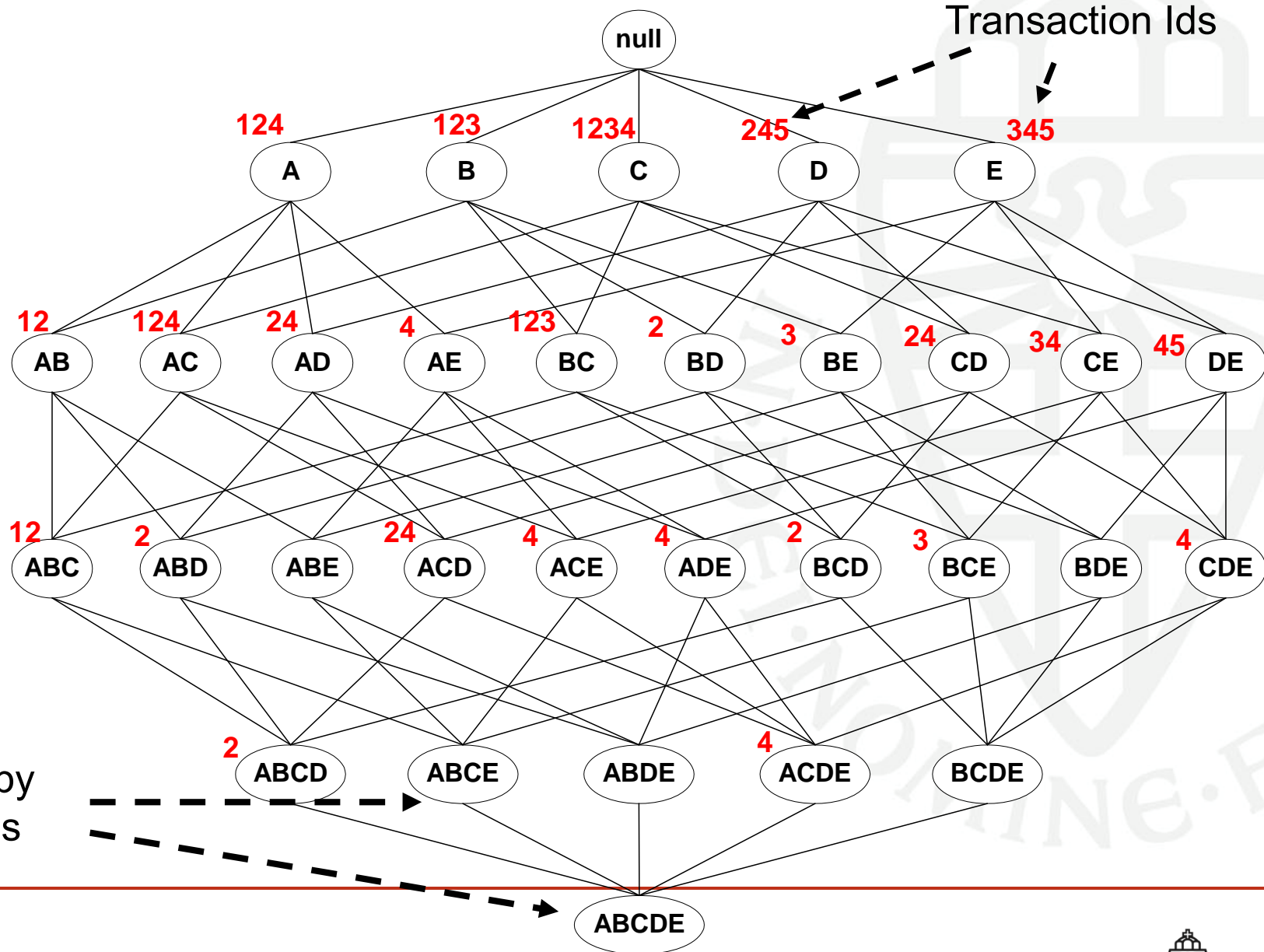
## Maximal Frequent Itemset

An itemset is **maximal frequent** if none of its immediate supersets is frequent



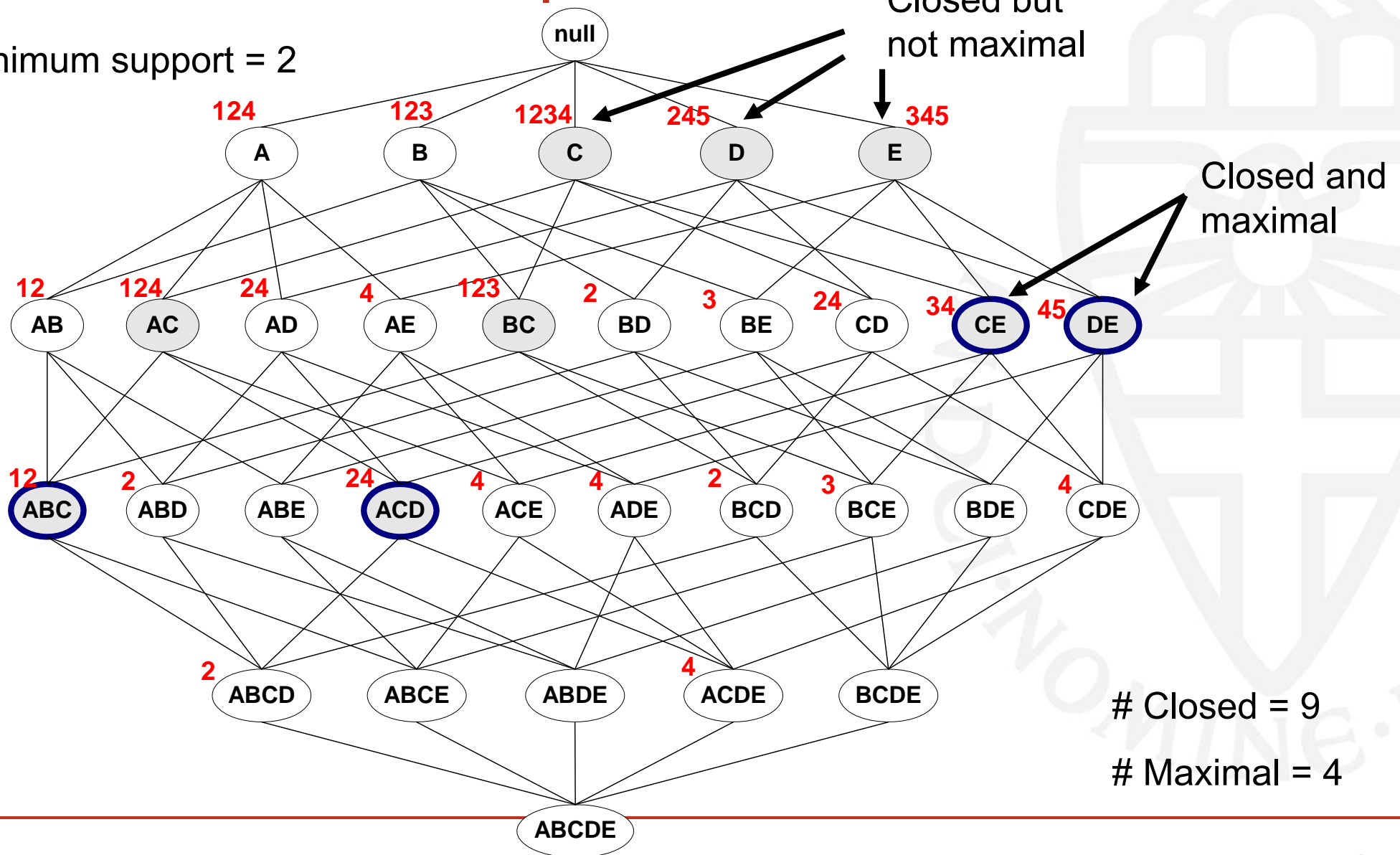
# Maximal vs Closed Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



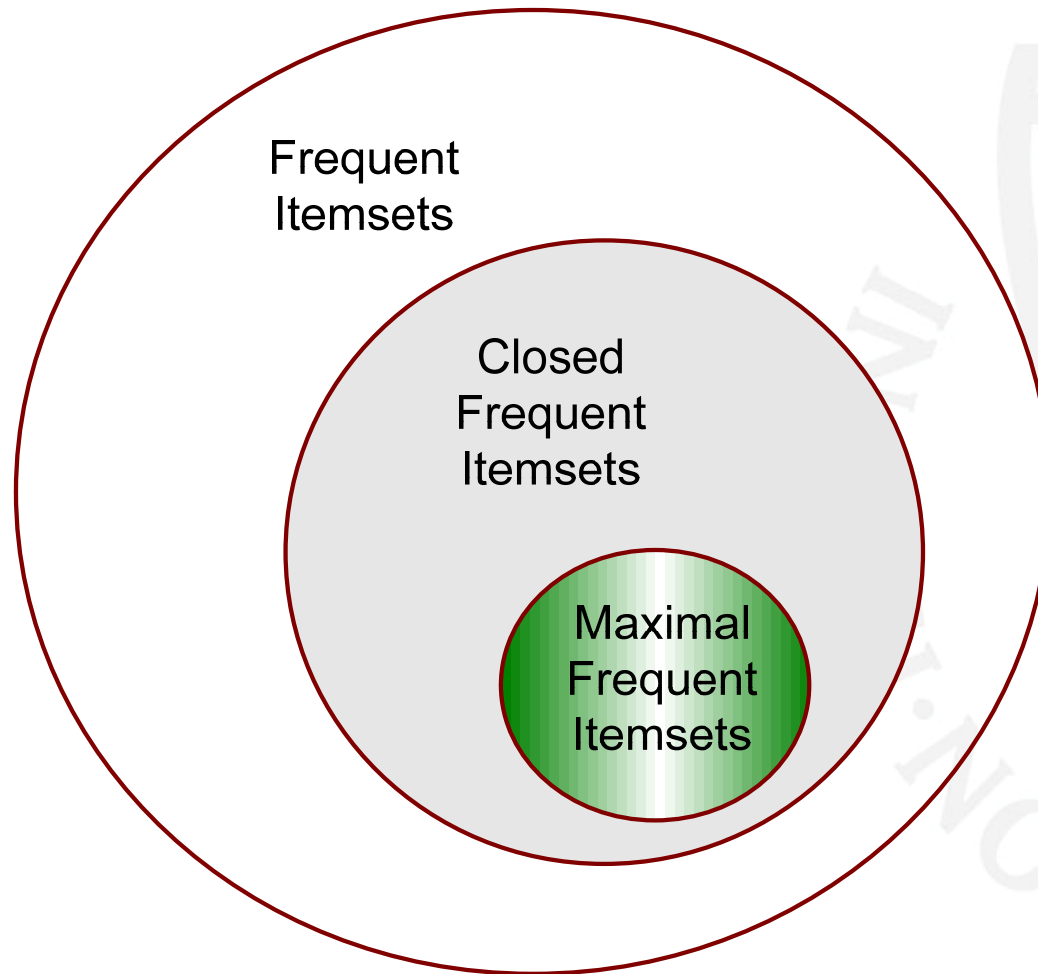
# Maximal vs Closed Frequent Itemsets

Minimum support = 2



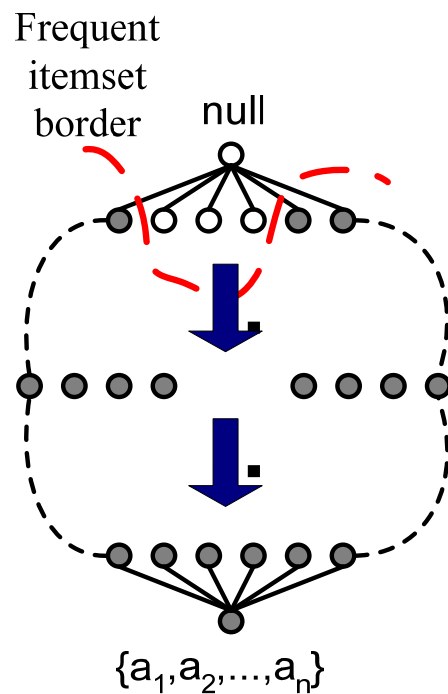


## Maximal vs Closed Itemsets

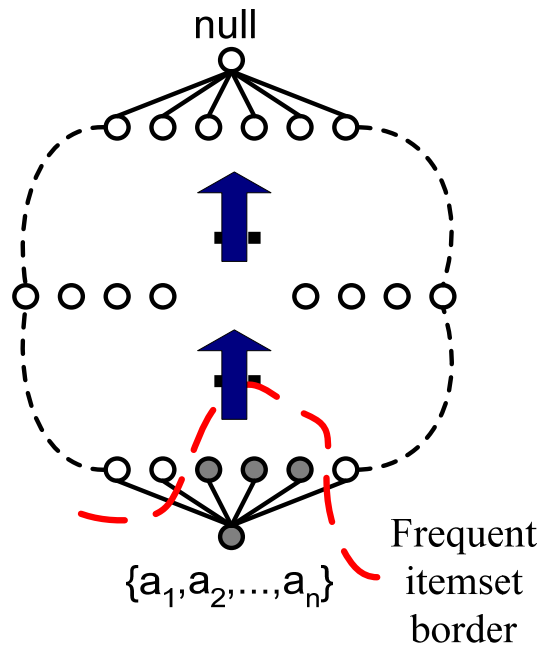


## Alternative Methods for Frequent Itemset Generation

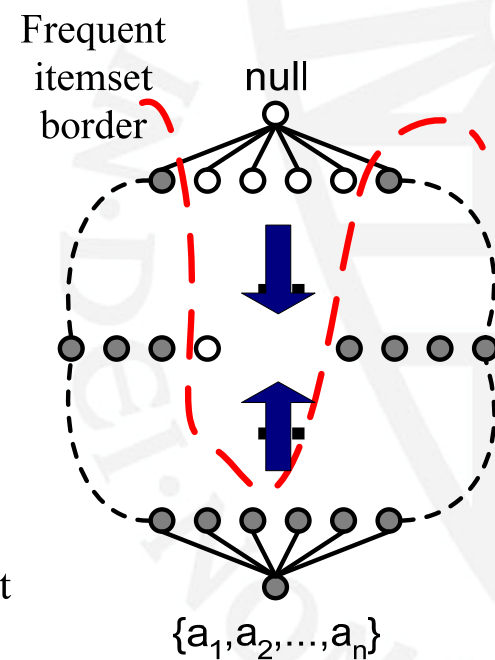
- General-to-specific (as Apriori) vs Specific-to-general



(a) General-to-specific



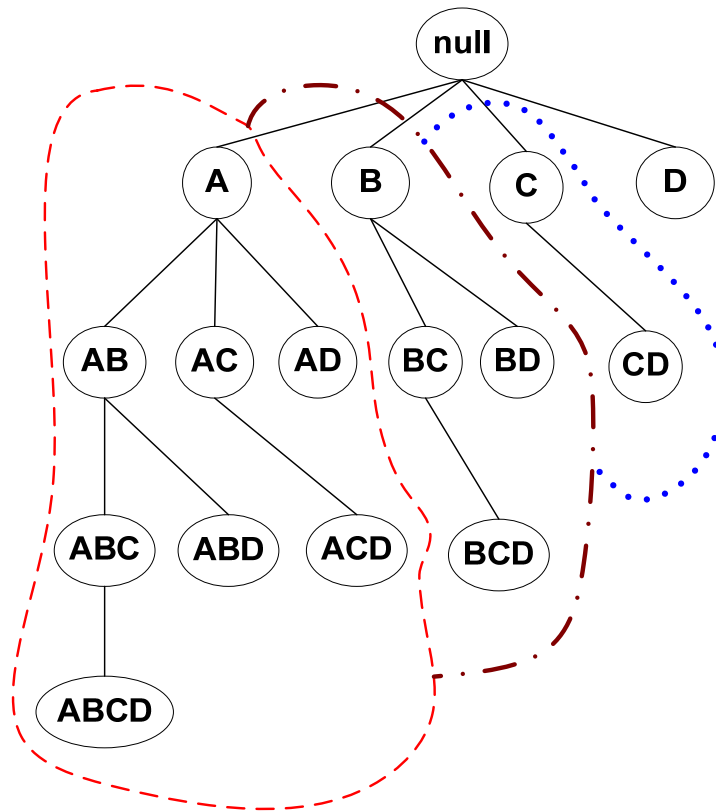
(b) Specific-to-general



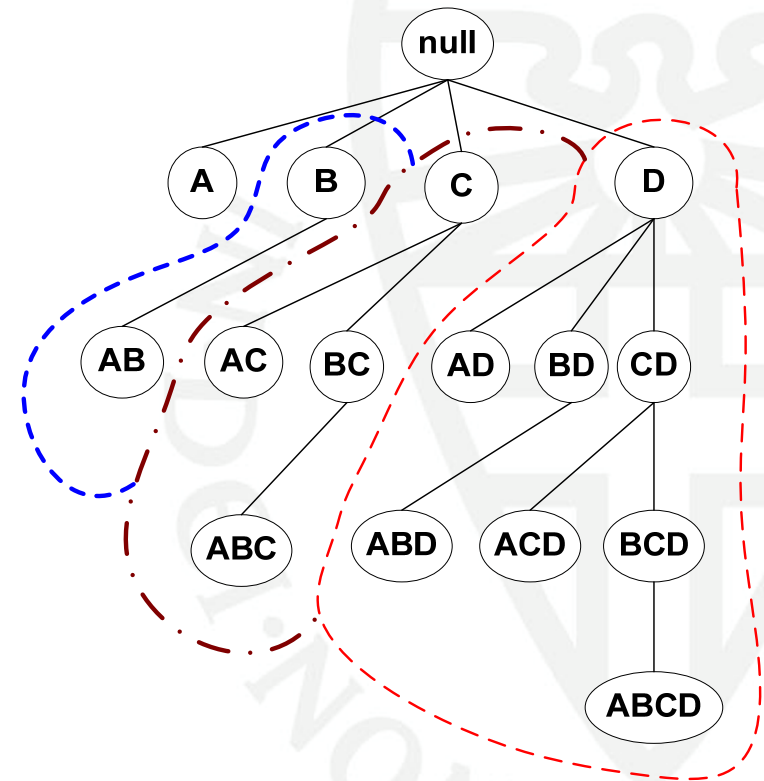
(c) Bidirectional

## Alternative Methods for Frequent Itemset Generation

- Equivalent Classes



(a) Prefix tree

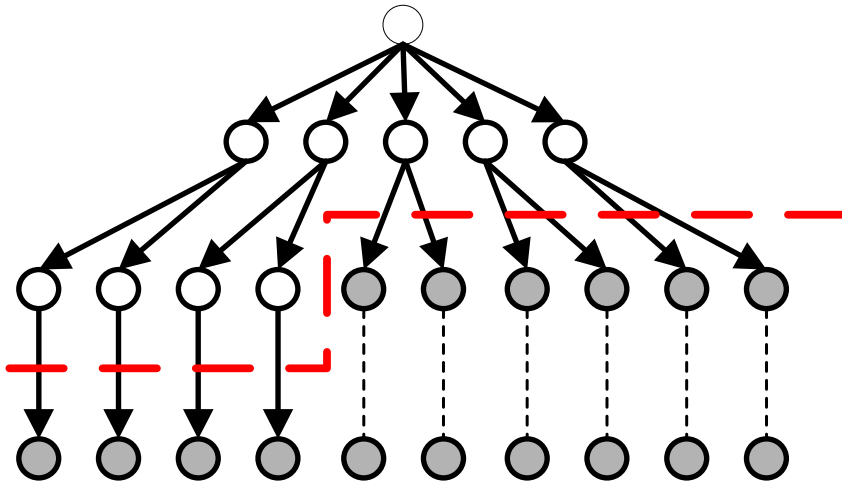


(b) Suffix tree

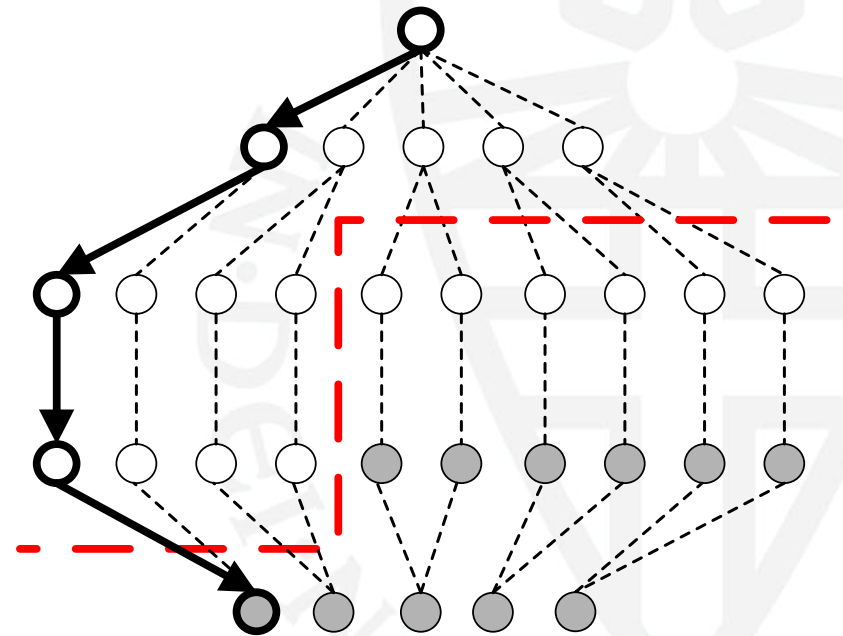
ation

Depth first

- Breadth-first vs depth-first



(a) Breadth first



(b) Depth first

## Rule Generation

- Given a frequent itemset  $L$ , find all non-empty subsets  $f \subset L$  such that  $f \rightarrow L - f$  satisfies the minimum confidence requirement
- If  $\{A,B,C,D\}$  is a frequent itemset, candidate rules:  
ABC  $\rightarrow$  D,      ABD  $\rightarrow$  C, ACD  $\rightarrow$  B, BCD  $\rightarrow$  A,  
A  $\rightarrow$  BCD,      B  $\rightarrow$  ACD, C  $\rightarrow$  ABD, D  $\rightarrow$  ABC,  
AB  $\rightarrow$  CD,      AC  $\rightarrow$  BD, AD  $\rightarrow$  BC, BC  $\rightarrow$  AD,  
BD  $\rightarrow$  AC,      CD  $\rightarrow$  AB
- If  $|L| = k$ , then there are  $2^k - 2$  candidate association rules (ignoring  $L \rightarrow \emptyset$  and  $\emptyset \rightarrow L$ )

## Rule Generation

- How to efficiently generate rules from frequent itemsets?
- In general, confidence does not have a monotone property:

$c(ABC \rightarrow D)$  can be larger or smaller than  $c(AB \rightarrow D)$

- But confidence of rules generated from the **same itemset** does have a monotone property
- For example,  $L = \{A, B, C, D\}$ :

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

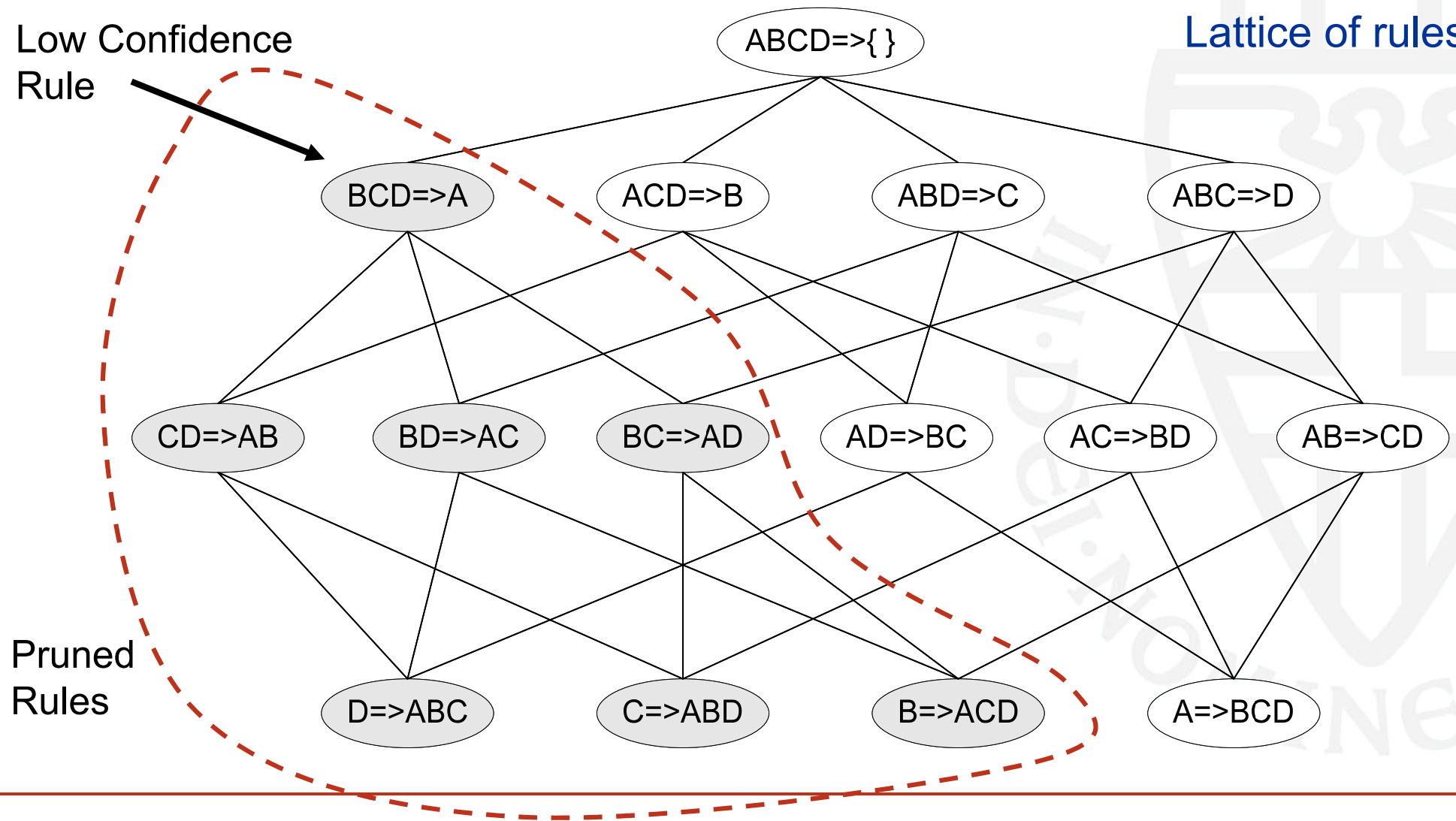
- Confidence is a decreasing function of the number of items on the RHS of the rule



## Rule Generation for Apriori Algorithm

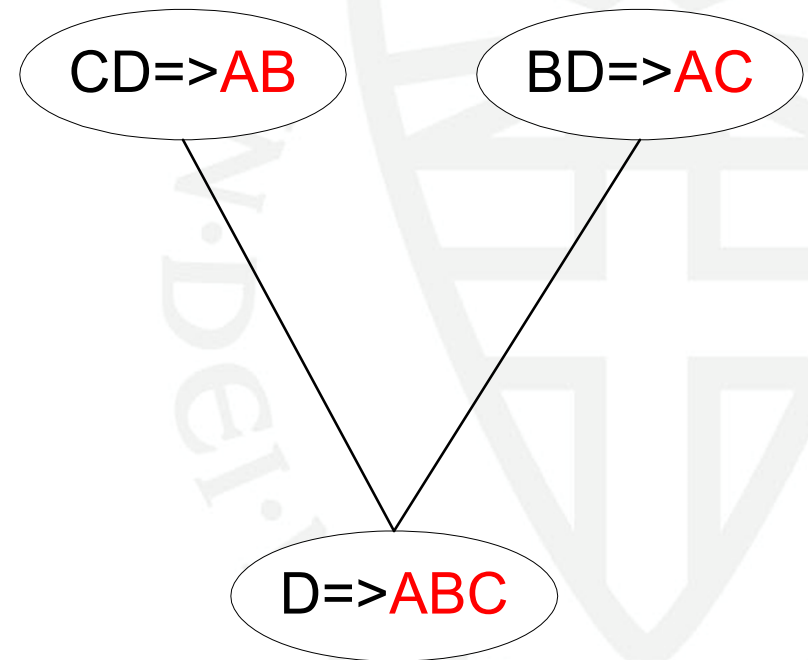
Low Confidence  
Rule

Lattice of rules



## Rule Generation for Apriori Algorithm

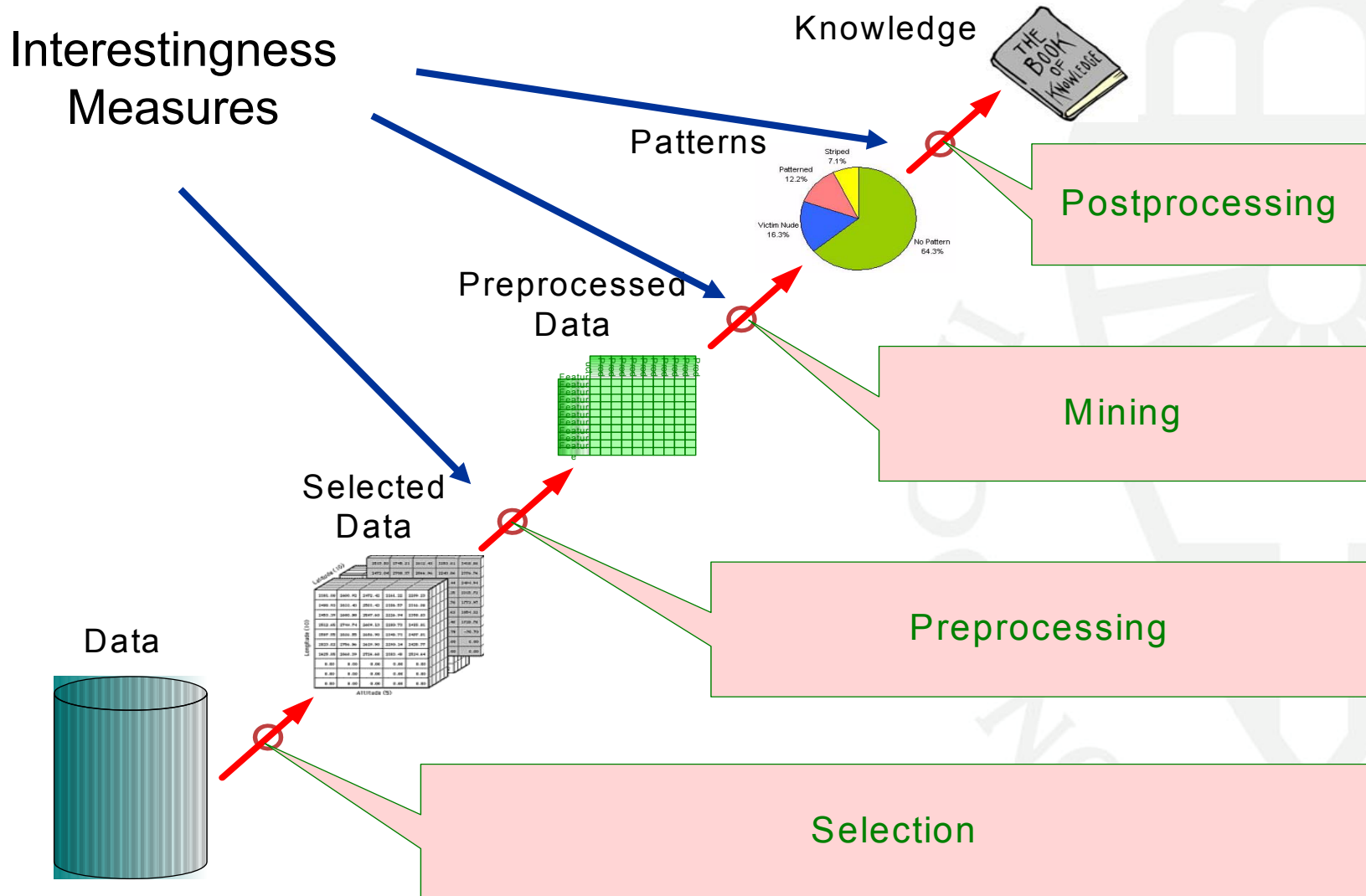
- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent
- $\text{join}(\text{CD} \Rightarrow \text{AB}, \text{BD} \Rightarrow \text{AC})$  would produce the candidate rule  $\text{D} \Rightarrow \text{ABC}$
- Prune rule  $\text{D} \Rightarrow \text{ABC}$  if its subset  $\text{AD} \Rightarrow \text{BC}$  does not have high confidence



## Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - redundant if  $\{A,B,C\} \rightarrow \{D\}$  and  $\{A,B\} \rightarrow \{D\}$  have the same support and confidence
- **Interestingness measures** can be used to prune/rank the derived patterns
- In the original formulation of association rules, support and confidence are the only measures used

# Application of Interestingness Measure



## Computing Interestingness Measure

- Given a rule  $X \rightarrow Y$ , information needed to compute rule interestingness can be obtained from a **contingency table**

	Y	$\bar{Y}$	
X	$f_{11}$	$f_{10}$	$f_{1+}$
$\bar{X}$	$f_{01}$	$f_{00}$	$f_{0+}$
	$f_{+1}$	$f_{+0}$	$ T $

$f_{11}$ : support of X and Y  
 $f_{10}$ : support of X and  $\bar{Y}$   
 $f_{01}$ : support of  $\bar{X}$  and Y  
 $f_{00}$ : support of  $\bar{X}$  and  $\bar{Y}$

Used to define various measures:  
support, confidence, lift, Gini, J-measure, etc.

## Drawback of Confidence

- Association rule: **Tea**  $\rightarrow$  **Coffee**
- Confidence =  $P(\text{Coffee}|\text{Tea}) = 0.75$
- But  $P(\text{Coffee}) = 0.9$
- Although confidence is high, rule is uninteresting
- $P(\text{Coffee}|\overline{\text{Tea}}) = 0.9375$

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100



## Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)
  - $P(S \cap B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \cap B) = P(S) \times P(B) \Rightarrow$  Statistical independence
  - $P(S \cap B) > P(S) \times P(B) \Rightarrow$  Positively correlated
  - $P(S \cap B) < P(S) \times P(B) \Rightarrow$  Negatively correlated

## Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{P(Y | X)}{P(Y)}$$

$$Interest = \frac{P(X, Y)}{P(X)P(Y)} \quad (= Lift \quad \text{for binary variables})$$

$$PS = P(X, Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X, Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}$$

## Example: Lift/Interest

- Association rule: **Tea**  $\rightarrow$  **Coffee**
- Confidence =  $P(\text{Coffee}|\text{Tea}) = 0.75$
- But  $P(\text{Coffee}) = 0.9$
- **Lift** =  $0.75/0.9 = 0.8333$  ( $< 1$ , therefore is negatively associated)

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

## Drawback of Lift and Interest

	Y	$\bar{Y}$	
X	10	0	10
$\bar{X}$	0	90	90
	10	90	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

	Y	$\bar{Y}$	
X	90	0	90
$\bar{X}$	0	10	10
	90	10	100

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

- Not invariant under **inversion operation** ( $0 \rightarrow 1$  and  $1 \rightarrow 0$ )

## Interestingness Measures (1)

#	Measure	Formula
1	$\phi$ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's ( $\lambda$ )	$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio ( $\alpha$ )	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\bar{A}\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A}\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha - 1}{\alpha + 1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\bar{A}\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A}\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$
6	Kappa ( $\kappa$ )	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
7	Mutual Information ( $M$ )	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$
8	J-Measure ( $J$ )	$\max \left( P(A,B) \log \left( \frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left( \frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), \right. \\ \left. P(A,\bar{B}) \log \left( \frac{P(A \bar{B})}{P(A)} \right) + P(\bar{A}B) \log \left( \frac{P(\bar{A} B)}{P(\bar{A})} \right) \right)$
9	Gini index ( $G$ )	$\max \left( P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] \right. \\ \left. - P(B)^2 - P(\bar{B})^2, \right. \\ \left. P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] \right. \\ \left. - P(A)^2 - P(\bar{A})^2 \right)$

## Interestingness Measures (2)

10	Support ( $s$ )	$P(A, B)$
11	Confidence ( $c$ )	$\max(P(B A), P(A B))$
12	Laplace ( $L$ )	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction ( $V$ )	$\max\left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{A}B)}\right)$
14	Interest ( $I$ )	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine ( $IS$ )	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's ( $PS$ )	$P(A, B) - P(A)P(B)$
17	Certainty factor ( $F$ )	$\max\left(\frac{P(B A)-P(B)}{1-P(B)}, \frac{P(A B)-P(A)}{1-P(A)}\right)$
18	Added Value ( $AV$ )	$\max(P(B A) - P(B), P(A B) - P(A))$
19	Collective strength ( $S$ )	$\frac{P(A,B)+P(\bar{A}\bar{B})}{P(A)P(B)+P(\bar{A})P(\bar{B})} \times \frac{1-P(A)P(B)-P(\bar{A})P(\bar{B})}{1-P(A,B)-P(\bar{A}\bar{B})}$
20	Jaccard ( $\zeta$ )	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen ( $K$ )	$\sqrt{P(A, B)} \max(P(B A) - P(B), P(A B) - P(A))$

## Interestingness Measures

- There are lots of measures of interestingness proposed in the literature
- Some measures are good for certain applications, but not for others
- What criteria should we use to determine whether a measure is good or bad?
- What about Apriori-style support based pruning? How does this affect these measures?



## Properties of A Good Measure

- **Piatetsky-Shapiro**: 3 properties a good measure  $M$  must satisfy:
  - $M(A,B) = 0$  if  $A$  and  $B$  are statistically independent
  - $M(A,B)$  increase monotonically with  $P(A,B)$  when  $P(A)$  and  $P(B)$  remain unchanged
  - $M(A,B)$  decreases monotonically with  $P(A)$  [or  $P(B)$ ] when  $P(A,B)$  and  $P(B)$  [or  $P(A)$ ] remain unchanged

## Comparing Different Measures


Example	$f_{11}$	$f_{10}$	$f_{01}$	$f_{00}$
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Rankings of contingency tables using various measures:

#	$\phi$	$\lambda$	$\alpha$	$Q$	$Y$	$\kappa$	$M$	$J$	$G$	$s$	$c$	$L$	$V$	$I$	$IS$	$PS$	$F$	$AV$	$S$	$\zeta$	$K$
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

## Property under Variable Permutation

	<b>B</b>	<b><math>\bar{B}</math></b>
<b>A</b>	p	q
<b><math>\bar{A}</math></b>	r	s



	<b>A</b>	<b><math>\bar{A}</math></b>
<b>B</b>	p	r
<b><math>\bar{B}</math></b>	q	s

- Does  $M(A,B) = M(B,A)$ ?
- Symmetric measures:  
support, lift, collective strength, cosine, Jaccard, etc
- Asymmetric measures:  
confidence, conviction, Laplace, J-measure, etc

## Property under Row/Column Scaling

- Grade-Gender Example (Mosteller, 1968):

	Male	Female	
High	1	4	5
Low	2	3	5
	3	7	10

	Male	Female	
High	2	40	34
Low	4	30	42
	6	70	76



2x



10x

- Mosteller: Underlying association should be independent of the relative number of male and female students in the samples

## Property under Inversion Operation

		A	B		C	D		E	F
Transaction 1	→	1	0		0	1		0	0
	■	0	0		1	1		1	0
		0	0		1	1		1	0
	■	0	0		1	1		1	0
		0	1		1	0		1	1
	■	0	0		1	1		1	0
		0	0		1	1		1	0
	■	0	0		1	1		1	0
		0	0		1	1		1	0
Transaction N	→	1	0		0	1		0	0

(a) (b) (c)

## Example: $\phi$ -Coefficient

- $\phi$ -coefficient is analogous to correlation coefficient for continuous variables
- invariant under **inversion operation**

	Y	$\bar{Y}$	
X	60	10	70
$\bar{X}$	10	20	30
	70	30	100

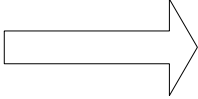
$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

	Y	$\bar{Y}$	
X	20	10	30
$\bar{X}$	10	60	70
	30	70	100

$$\phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

## Property under Null Addition

	<b>B</b>	<b><math>\bar{B}</math></b>
<b>A</b>	p	q
<b><math>\bar{A}</math></b>	r	s



	<b>B</b>	<b><math>\bar{B}</math></b>
<b>A</b>	p	q
<b><math>\bar{A}</math></b>	r	s + k

- **Invariant** measures:  
support, cosine, Jaccard, ...
- **Non-invariant** measures:  
correlation, Gini, mutual information, odds ratio...

## Different Measures have Different Properties

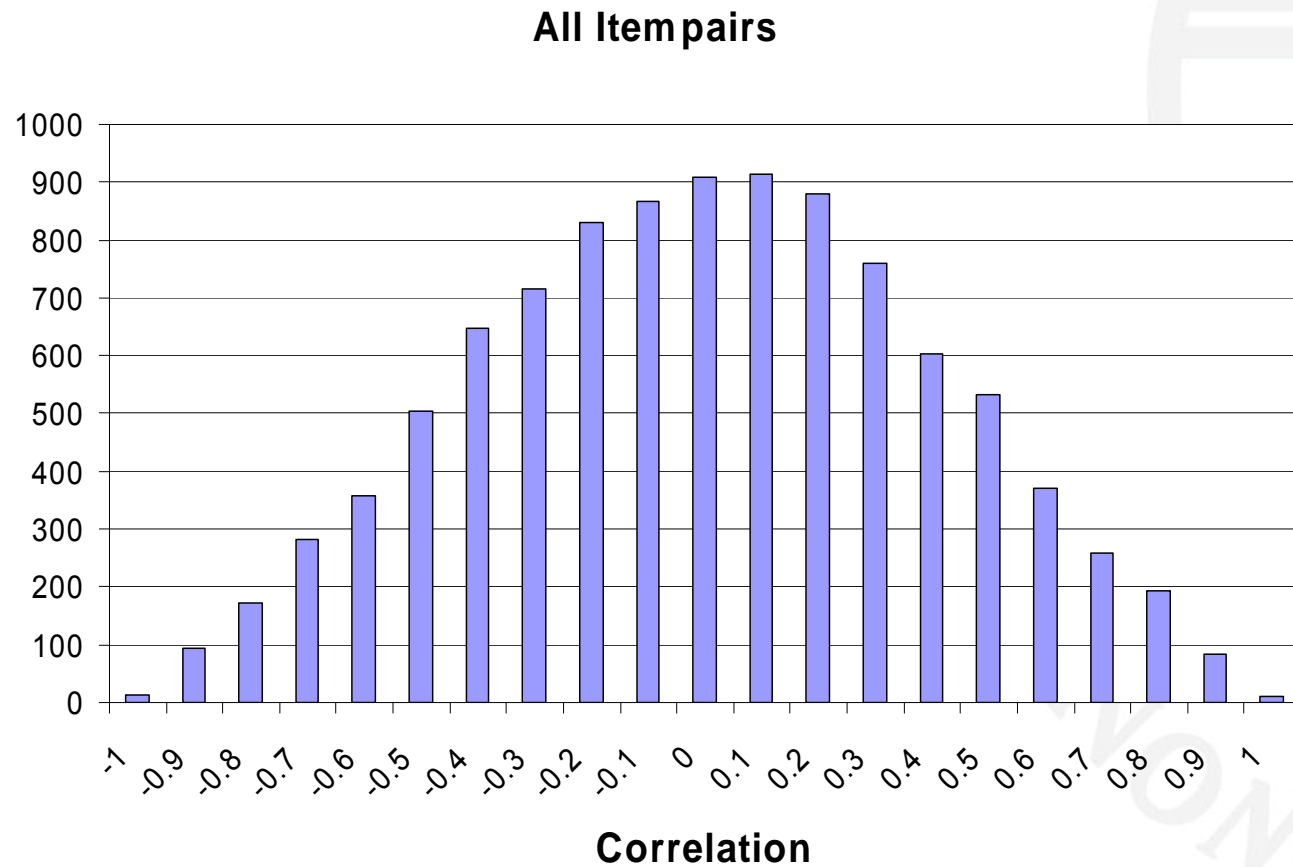
Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
$\Phi$	Correlation	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
$\lambda$	Lambda	0 ... 1	Yes	No	No	Yes	No	No*	Yes	No
$\alpha$	Odds ratio	0 ... 1 ... $\infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
$\kappa$	Cohen's	-1 ... 0 ... 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 ... 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 ... 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 ... 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 ... 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 ... 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 ... 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 ... 1 ... $\infty$	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 ... 1 ... $\infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 ... 0 ... 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 ... 0 ... 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 ... 1 ... 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 ... 1 ... $\infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
$\zeta$	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}}-1\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right) \dots 0 \dots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No



## Support-based Pruning

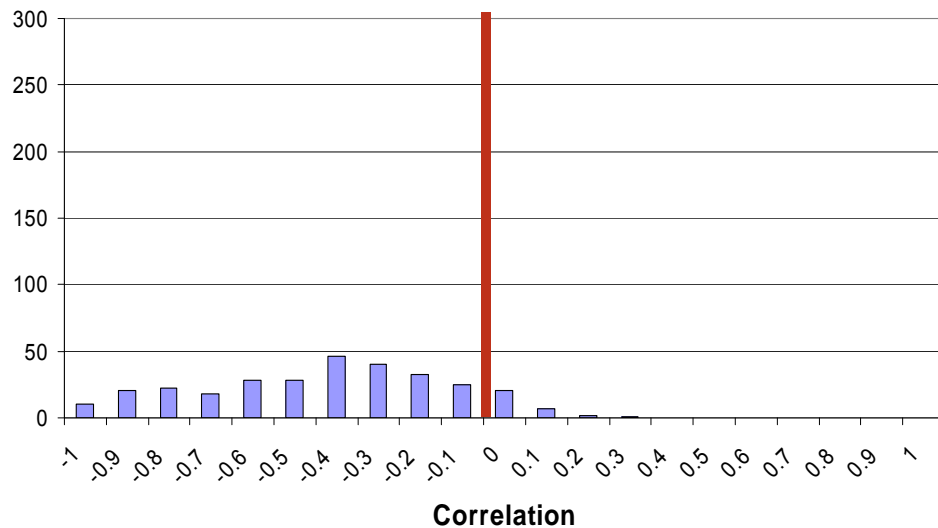
- Most of the association rule mining algorithms use support measure to prune rules and itemsets
- Study effect of support pruning on correlation of itemsets
  - Generate 10000 random contingency tables
  - Compute support and pairwise correlation for each table
  - Apply support-based pruning and examine the tables that are removed

## Effect of Support-based Pruning

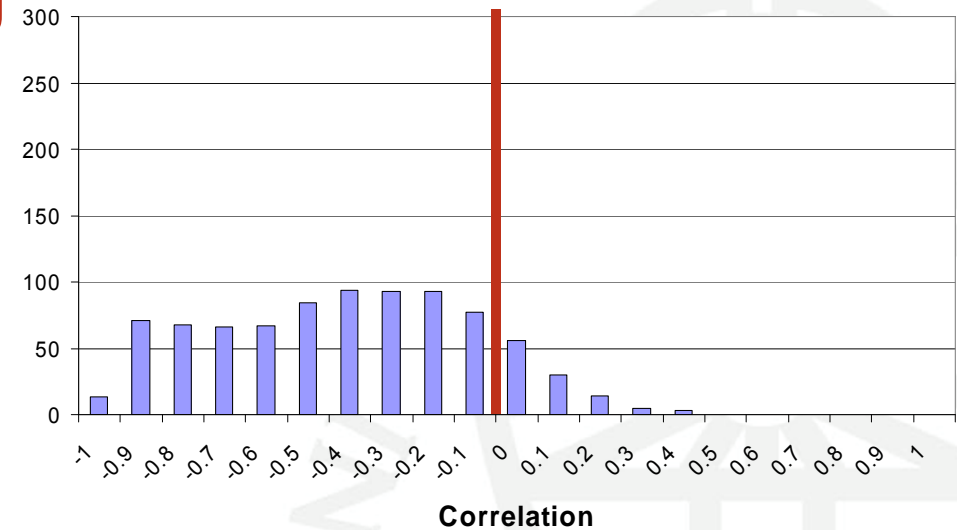


# Effect of Support-based Pruning

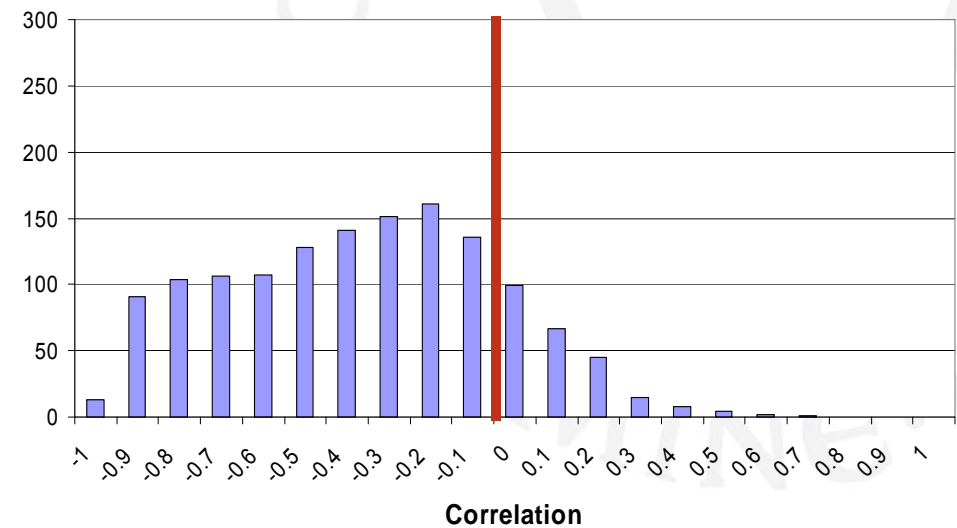
Support < 0.01



Support < 0.03



Support < 0.05



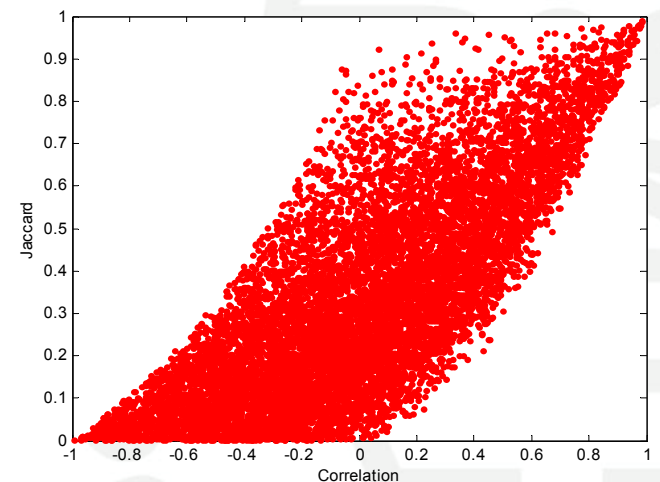
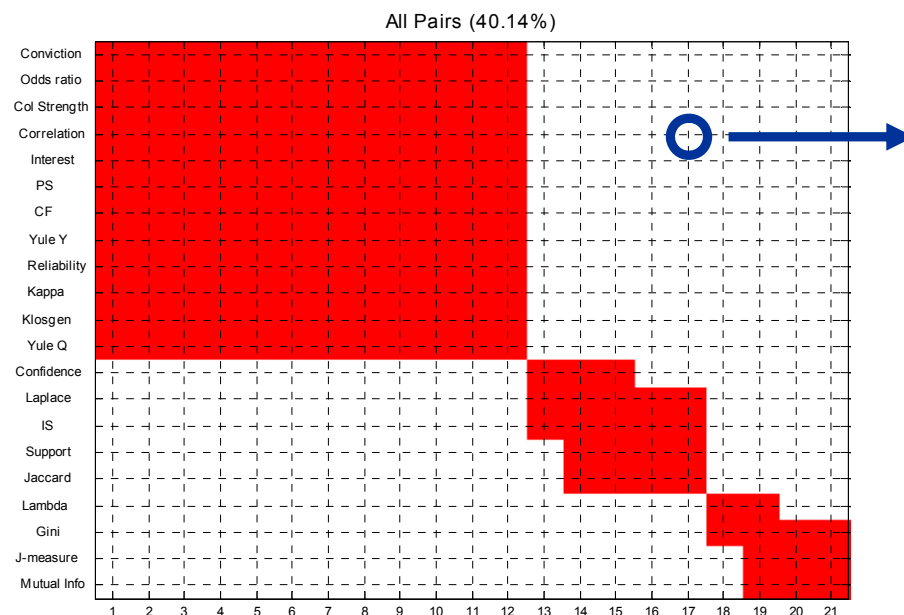
Support-based pruning eliminates mostly negatively correlated itemsets

## Effect of Support-based Pruning

- Investigate how support-based pruning affects other measures
- Steps:
  - Generate 10000 contingency tables
  - Rank each table according to the different measures
  - Compute the pair-wise correlation between the measures

## Effect of Support-based Pruning

- Without support pruning (all pairs)

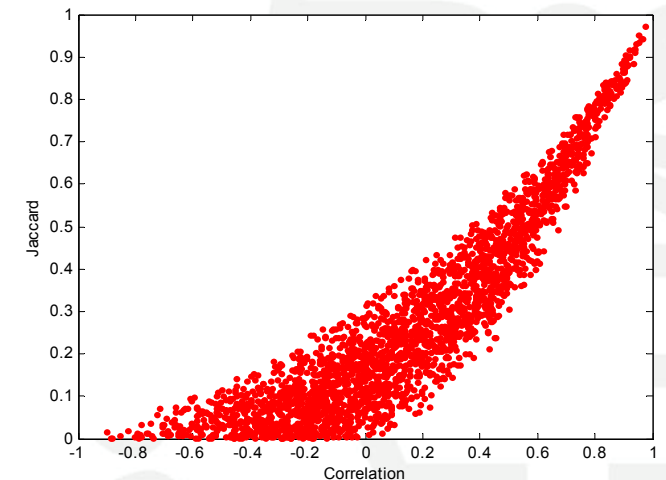
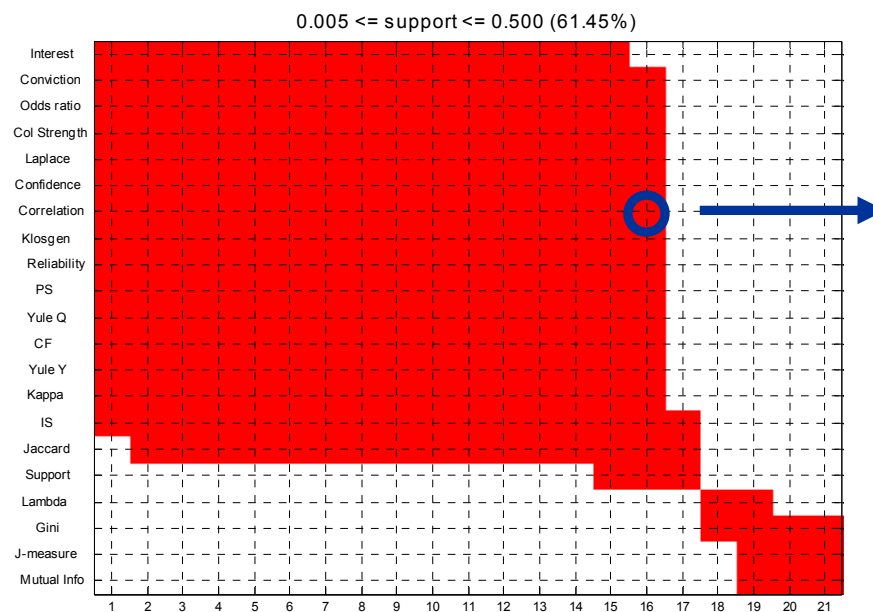


Scatter Plot between Correlation & Jaccard Measure

- Red cells indicate correlation between the pair of measures  $> 0.85$
- 40.14% pairs have correlation  $> 0.85$

## Effect of Support-based Pruning

- $0.5\% \leq \text{support} \leq 50\%$

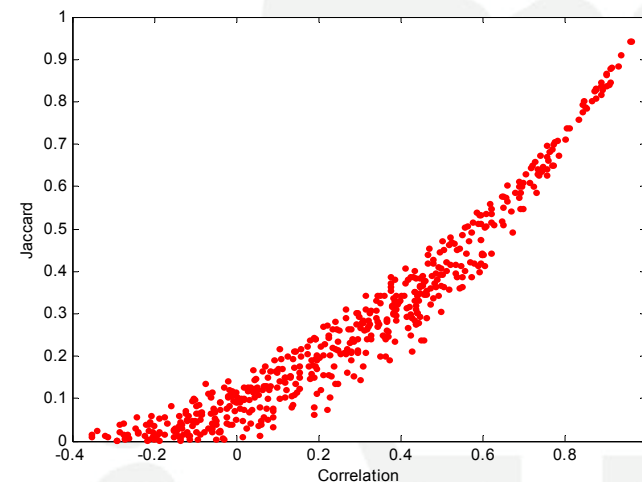
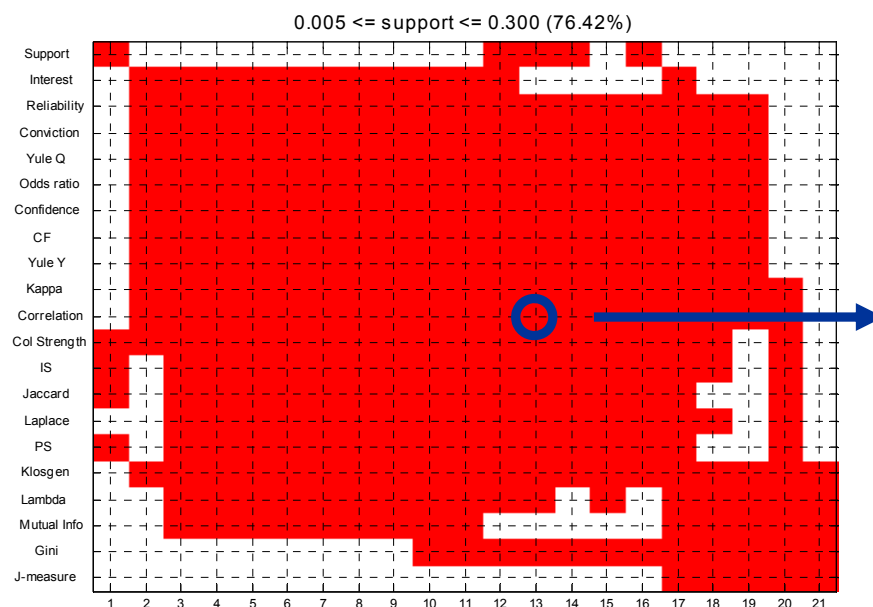


Scatter Plot between Correlation & Jaccard Measure

- 61.45% pairs have correlation  $> 0.85$

## Effect of Support-based Pruning

- $0.5\% \leq \text{support} \leq 30\%$



Scatter Plot between Correlation & Jaccard Measure

- 76.42% pairs have correlation  $> 0.85$

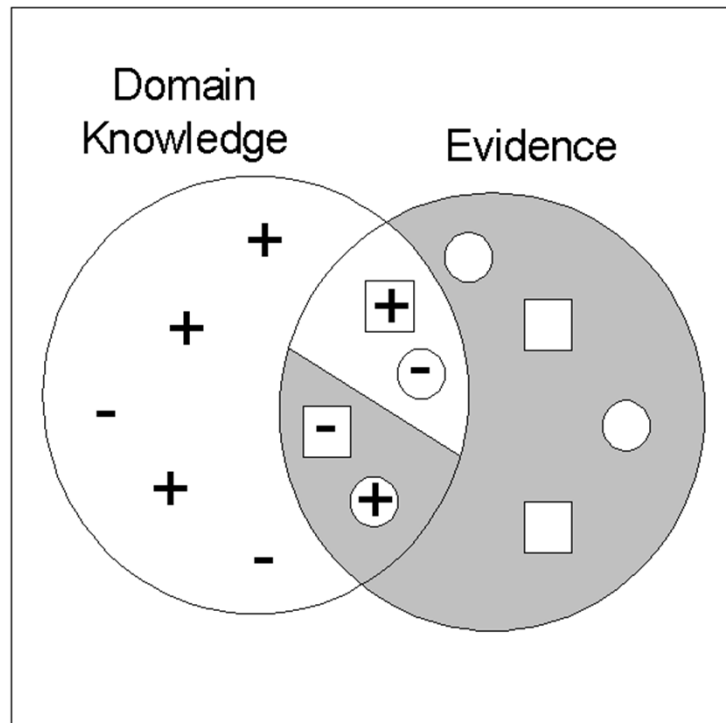
## Subjective Interestingness Measure

- Objective measure:
  - Rank patterns based on statistics computed from data
  - e.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).
- Subjective measure (Silberschatz & Tuzhilin):
  - Rank patterns according to user's interpretation
  - A pattern is subjectively interesting if it contradicts the expectation of a user
  - A pattern is subjectively interesting if it is actionable



## Interestingness via Unexpectedness

- Need to model expectation of users (domain knowledge)



- + Pattern expected to be frequent
- Pattern expected to be infrequent
- Pattern found to be frequent
- Pattern found to be infrequent
- + - Expected Patterns
- + Unexpected Patterns

- Need to combine expectation of users with evidence from data (i.e., extracted patterns)