

**DS 5230**

**Unsupervised Machine Learning and Data Mining  
Association Analysis**

**Steve Morin, Ph.D.**

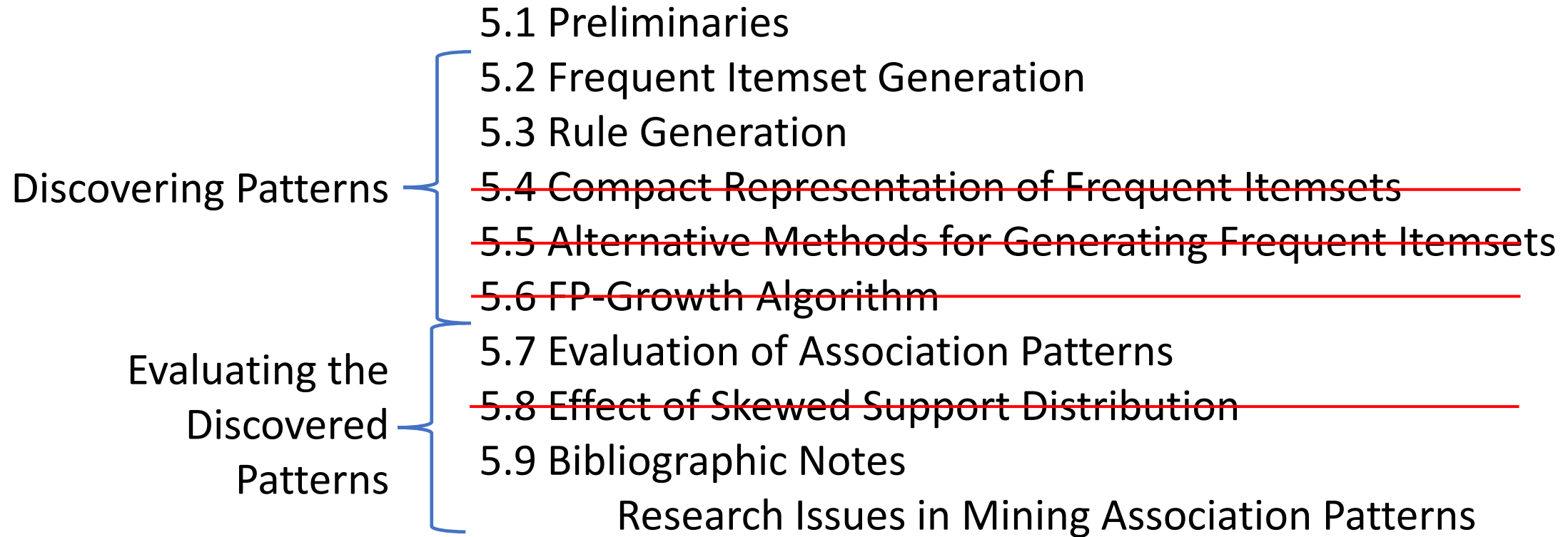
**s.morin@northeastern.edu**



Recommended  
Reading

# Introduction to Data Mining - Chapter 5

## Association Analysis: Basic Concepts and Algorithms



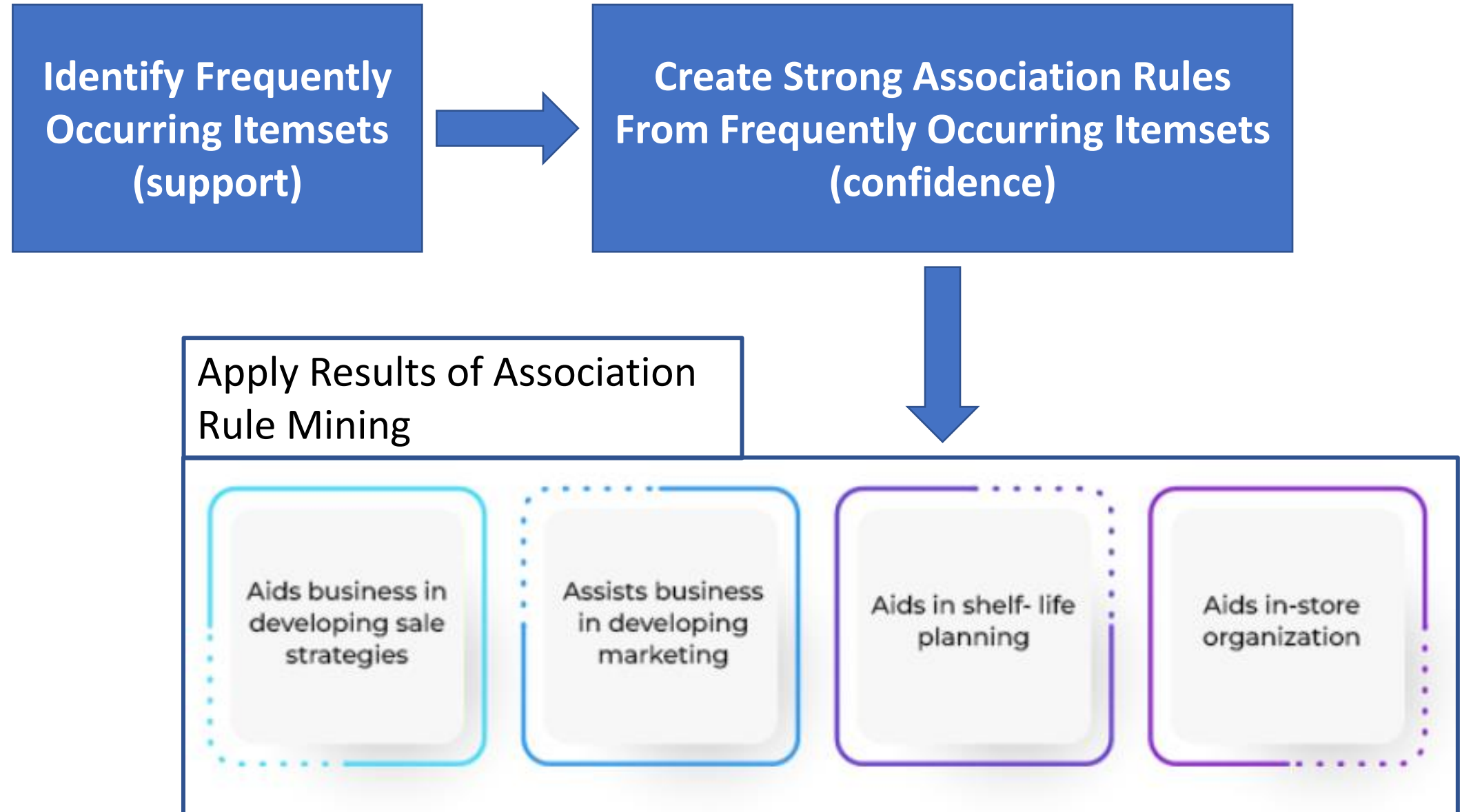
### Assumptions in this chapter:

- Input data set consists of binary attributes called items.
- The presence of an item is assumed more important than its absence therefore an item is an asymmetric binary attribute.
- Only frequent patterns are considered important.



# Introduction

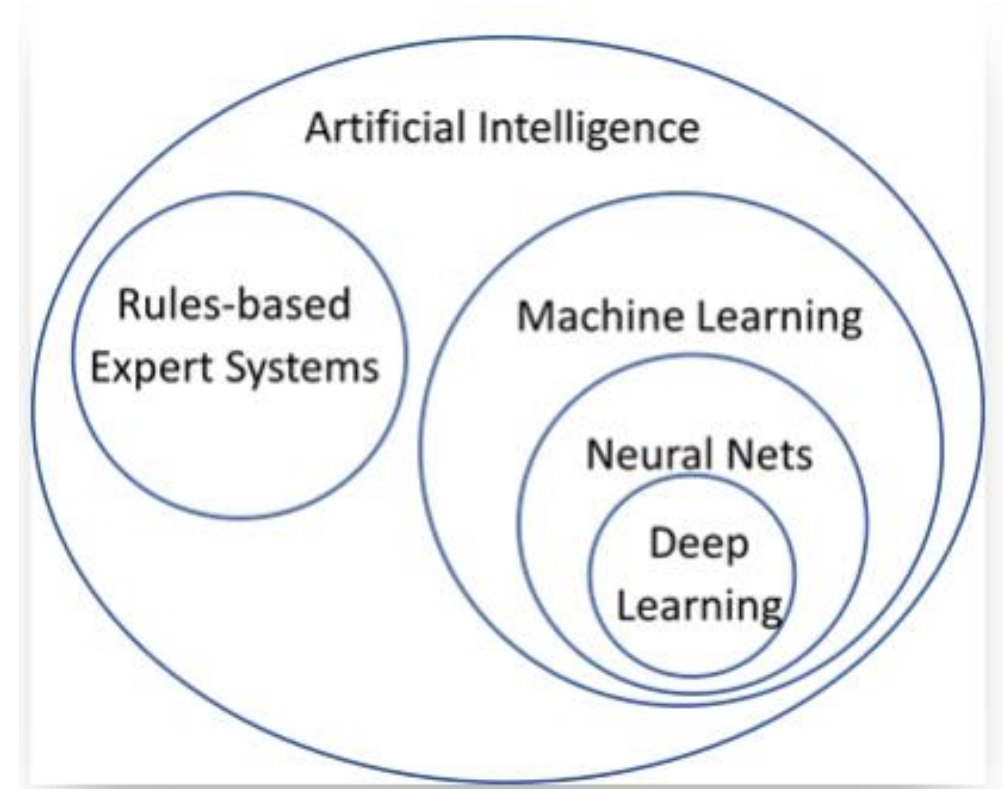
# Association Rule Mining – An Overview



# Association Rule Learning - Introduction

What is association rule learning?

Association rule learning is a ***rule-based machine learning*** method for discovering interesting relations between variables in large databases.



[https://en.wikipedia.org/wiki/Association\\_rule\\_learning](https://en.wikipedia.org/wiki/Association_rule_learning)

[https://en.wikipedia.org/wiki/Rule-based\\_machine\\_learning](https://en.wikipedia.org/wiki/Rule-based_machine_learning)

# Association Rule Learning - Introduction

## What is rule-based machine learning?

Rule-based machine learning encompasses any machine learning method that applies some form of learning algorithm that identifies, learns, or evolves rules to better understand data.

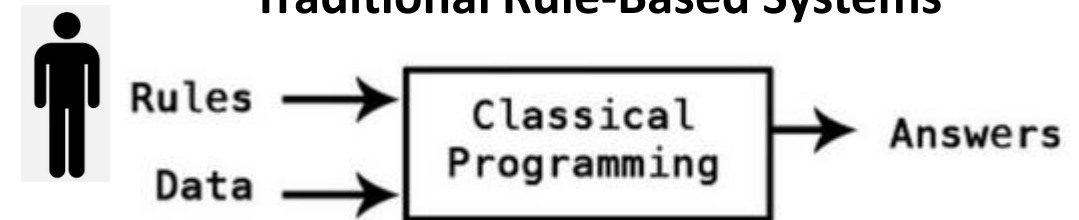
The defining characteristic of a rule-based machine learner is the identification and utilization of a set of rules that collectively represent the knowledge captured by the system.

Rule-based machine learning is distinct from traditional rule-based systems which are often hand-crafted by a human needing to apply prior domain knowledge to manually construct rules and curate a rule set.

## Rule-Based Machine Learning



## Traditional Rule-Based Systems



[https://en.wikipedia.org/wiki/Association\\_rule\\_learning](https://en.wikipedia.org/wiki/Association_rule_learning)

[https://en.wikipedia.org/wiki/Rule-based\\_machine\\_learning](https://en.wikipedia.org/wiki/Rule-based_machine_learning)

# Association Rule Learning – Introduction (continued)

Association rule learning is intended to identify strong rules discovered in databases using some measures of interestingness.

In any given transaction with a variety of items, association rules are meant to discover the rules that determine how or why certain items are connected.

Based on the concept of strong rules, association rules were introduced for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

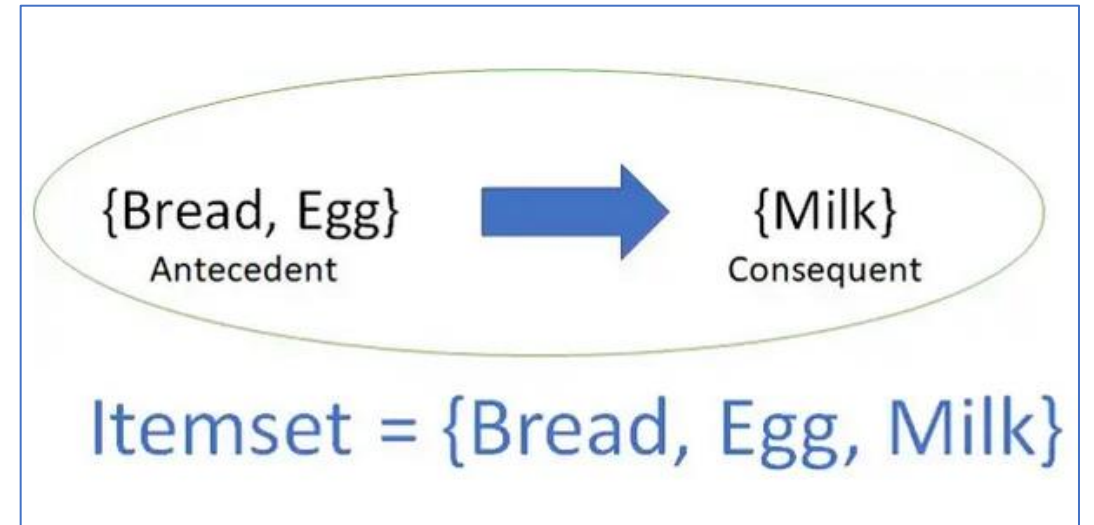




# Association Rule Learning – Introduction (continued)

For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.

Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.



# Association Rule Learning – Introduction (continued)

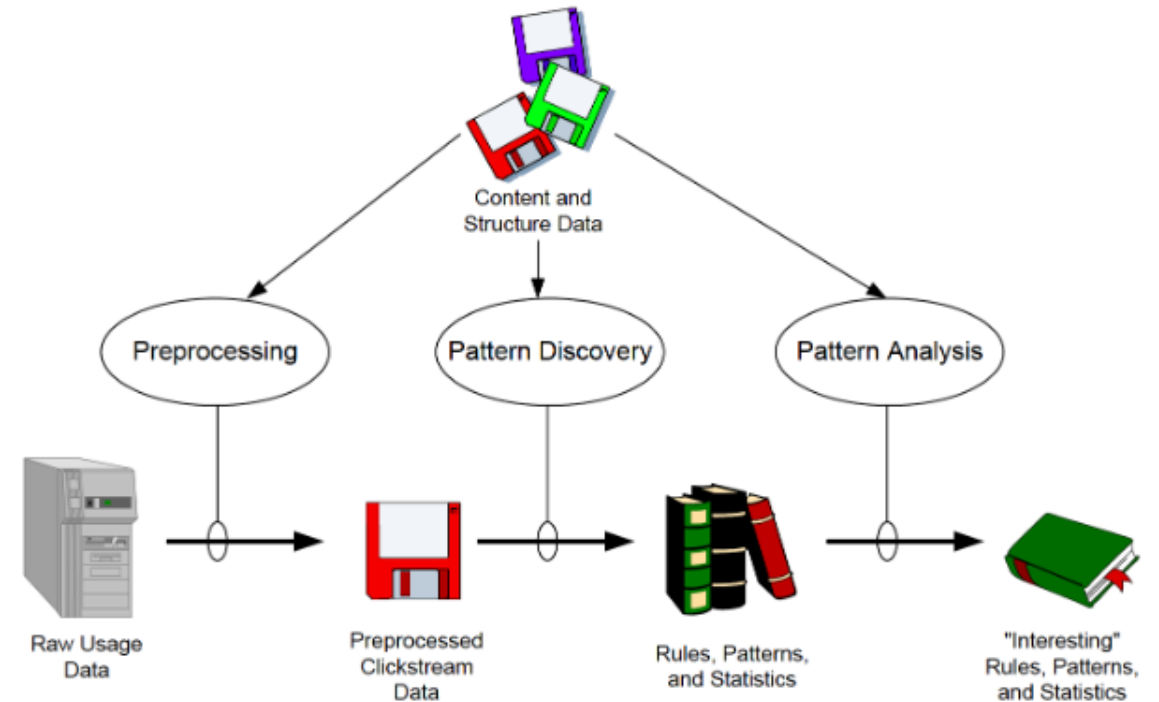
In addition to the above example from market basket analysis, association rules are employed today in many application areas including:

- Web usage mining,
- intrusion detection,
- continuous production, and
- bioinformatics.

In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

## Web usage mining (simplified view)

---



# Analyzing Large Data Sets for Frequent Itemsets

Huge amounts of transaction data is generated daily.

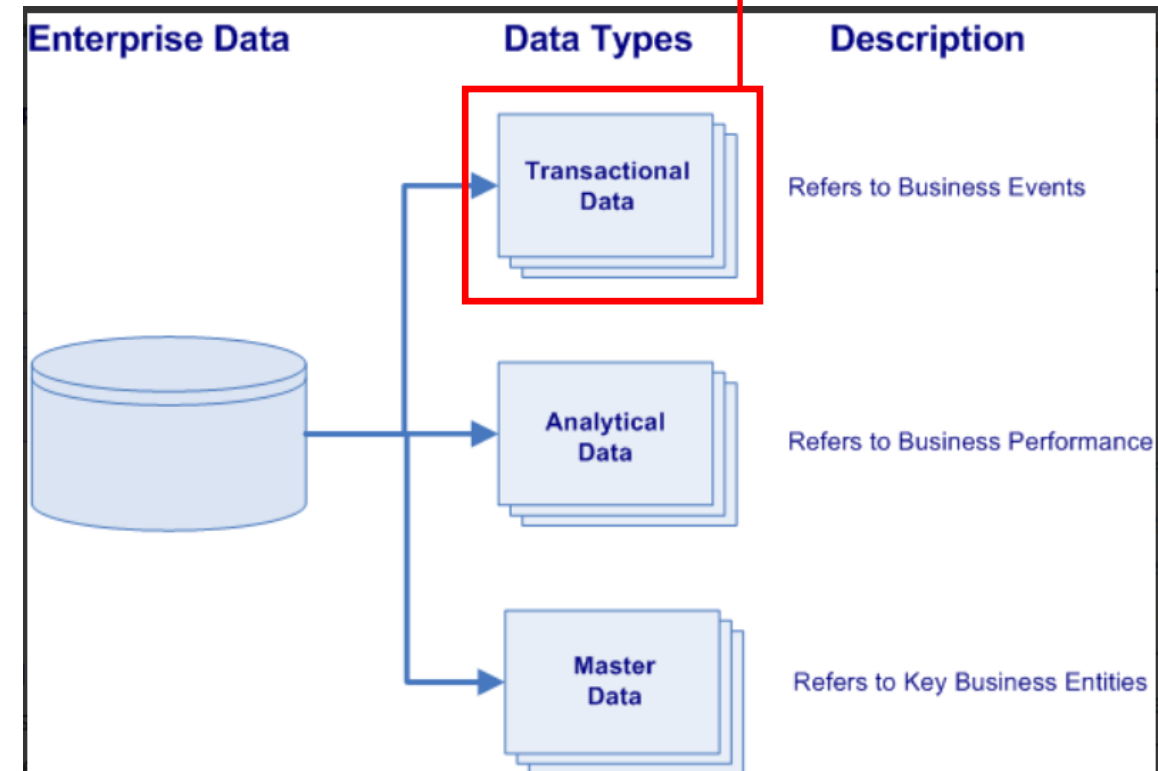
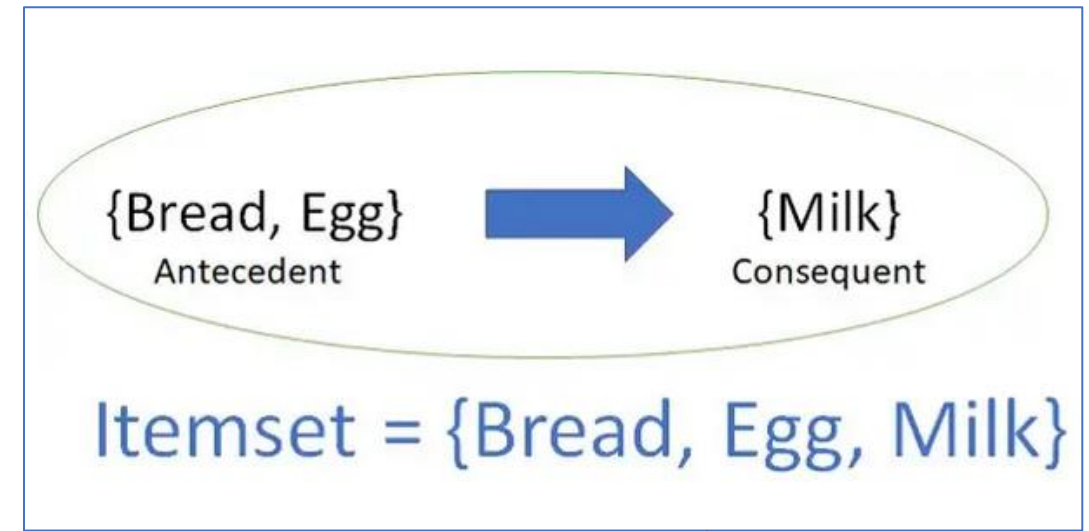
Retailers are interested in analyzing transaction data to learn about purchasing behaviors.

The results of these analyses can be used in:

- marketing promotions
- inventory management
- customer relationship management

Association analysis is used to discover interesting relationships in large data sets.

The uncovered relationships can be represented in the form of sets of items present in many transaction. These sets of items are often called **frequent itemsets** or **association rules**.



# Two Key Issues to Consider When Applying Association Analysis to Market Basket Data

- Computationally Expensive for Large Data Sets
- Evaluating Discovered Patterns
  - Prevent the Generation of Spurious Patterns (patterns that happen by chance)
  - Rank Patterns in Terms of an Interestingness Measure





# Association Analysis Preliminaries

# Transaction / Market Basket Data – An Introduction

Record data is a collection of data objects, i.e., records each of which has a fixed set of attributes (data fields).

Record data is usually stored in a flat file or can be extracted from a relational database.

Transaction data is a special type of record data where each record, i.e., transaction involves a set of items.

For example, the set of items bought by a customer during one shopping trip constitutes a transaction and, along with a unique Transaction Identification (TID) can be considered a record or data object.

This type of data is often called market basket data because the set of items in each record are the products in a person's market basket.



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



# Transaction / Market Basket Data – Binary Representation

Transaction data is a collection of data objects. Each data object has a TID, and a set of items purchased in the transaction.

Each row represents the purchases of a customer at a particular time.

Transaction data objects are often transformed into a set of data objects with binary asymmetric attributes.

Most often the transformed data attributes are binary indicating whether an item was purchased (1) or not purchased (0).

## Transaction Data

TID	Items
0	{Nutmeg, Eggs, Onion, Yogurt, Milk, Kidney Beans}
1	{Nutmeg, Eggs, Onion, Yogurt, Dill, Kidney Beans}
2	{Milk, Eggs, Kidney Beans, Apple}
3	{Unicorn, Kidney Beans, Yogurt, Corn, Milk}
4	{Eggs, Ice cream, Onion, Corn, Kidney Beans}

## Transformed Transaction Data

TID	Apple	Corn	Dill	Eggs	Ice cream	Kidney Beans	Milk	Nutmeg	Onion	Unicorn	Yogurt
0	0	0	0	1	0	1	1	1	1	0	1
1	0	0	1	1	0	1	0	1	1	0	1
2	1	0	0	1	0	1	1	0	0	0	0
3	0	1	0	0	0	1	1	0	0	1	1
4	0	1	0	1	1	1	0	0	1	0	0

# Symmetric and Asymmetric Binary Attributes

## Symmetric Binary Attribute

A binary attribute is symmetric if both of its states are equally valuable and carry the same weight.

For example, an attribute that records gender is symmetric in a data set where knowing a person's gender is important.

## Asymmetric Binary Attribute

A binary variable is asymmetric when one of the two states (e.g., state 1) is interpreted as more informative than the other state (e.g., state 0) .

For example, an attribute that records whether an item was purchased (1) or not (0) is asymmetric because in most cases we are interested in the relatively small number of items that were purchased and not the much larger set of items that were not purchased.

If we met a friend in the grocery store, would we ever say the following?

*“I see our purchases are very similar since we didn’t buy most of the same things.”*



# Definitions: Itemset, Support Count and Support

## Itemset

A collection of one or more items

Example: {Milk, Bread, Diaper}

k-itemset

An itemset that contains k items

Consider the 3-itemset in the transaction data on the right: {Milk, Bread, Diaper}

## Support count ( $\sigma$ )

Frequency of occurrence of the itemset

e.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

## Support

Fraction of transactions that contain the itemset

e.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5 = 0.40$

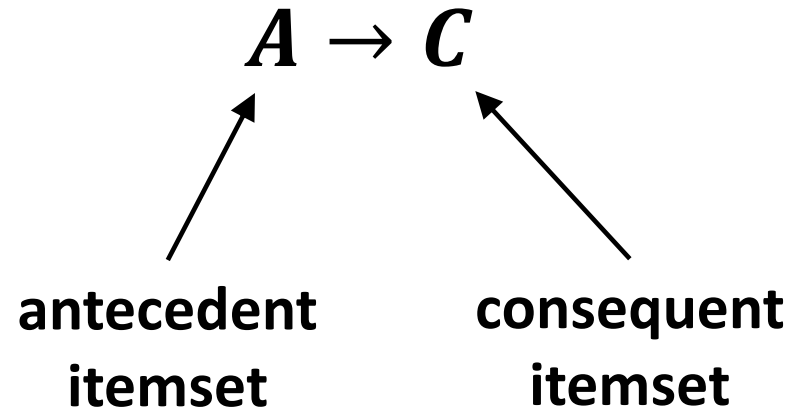
## Frequent Itemset

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

# Expressing Association Rules

Given a rule



$A$  and  $C$  are disjoint itemsets

The rule above is an implication expression.

The strength of an association rule can be measured in terms of its support and its confidence.

# Association Rules and Their Metrics

## Association Rule

An association rule is an implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets, e.g.,  $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

## Association Rule Evaluation Metrics

The support  $s$  is the fraction of transactions that contain both  $X$  and  $Y$

$$s(X \rightarrow Y) = \frac{\text{count of transactions that contain both } X \text{ and } Y}{\text{total number of transactions}}$$
$$= \frac{\sigma(X \cup Y)}{|T|}$$

The confidence  $c$  measures how often items in  $Y$  appear in transactions that contain  $X$

$$c(X \rightarrow Y) = \frac{\text{count of transactions that contain both } X \text{ and } Y}{\text{count of transactions that contain } X}$$
$$= \frac{\sigma(X \cup Y)}{\sigma(X)}$$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$X \rightarrow Y$$

$$\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$$

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

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

# Why Use Support and Confidence?

## Support

- Support enables us to discern between association rules that occur by chance versus those that occur due to a real pattern in behavior.
- Support enables efficient discovery of association rules.

## Confidence

- Measures the reliability of the association rule.
- For a given rule  $X \rightarrow Y$ , the higher the confidence, the more likely it is for  $Y$  to be present in transactions that contain  $X$ .

# Caution!

An association rule does not necessarily imply causality.

Instead, it can sometimes suggest a strong co-occurrence relationship between items in the antecedent and the consequent.

# Association Rule Mining Task

Given a set of transactions, 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

⇒ **Computationally prohibitive!**

# Association Rule Mining Task – Computational Complexity

Given  $d$  unique items in an association rule mining task:

Total number of possible frequent itemsets =  $2^d$

Total number of possible association rules =  $3^d - 2^{d+1} + 1$

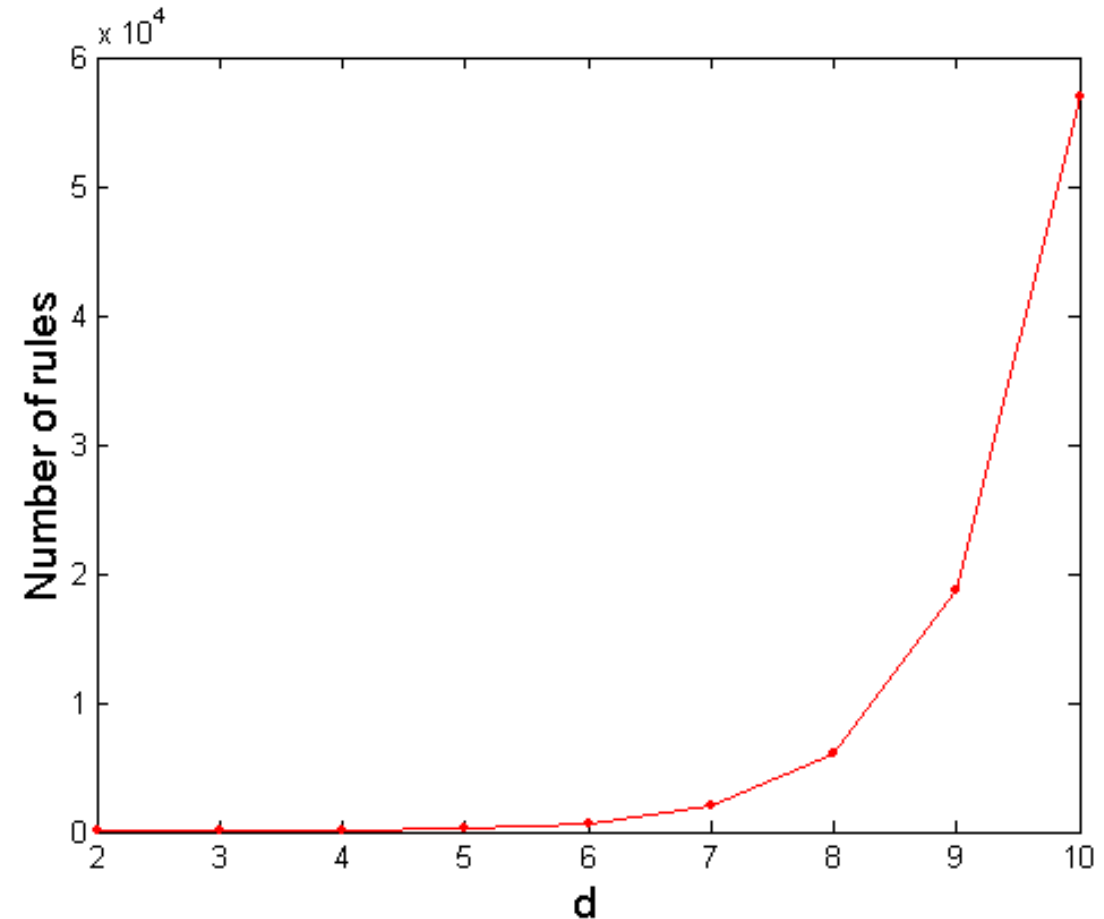
<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Unique Items: {Bread, Milk, Diaper, Eggs, Coke, Beer}

Number of Unique Items: 6

Number of Possible Frequent Itemsets = 64

Number of Possible Association Rules = 602



# Association Rule Mining Task – Opportunities to Reduce Computational Complexity

<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 of Rules:

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

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

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

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

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

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

All the above rules are binary partitions of the same itemset: {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

name	formula
support	$s(A \rightarrow C) = s(A \cup C)$
confidence	$confidence(A \rightarrow C) = \frac{s(A \rightarrow C)}{s(A)}$



# Association Rule Mining Task – Two-Step Approach

Two-step approach:

1. Frequent Itemset Generation
  - Generate all itemsets whose support  $\geq \text{minsup}$
2. Rule Generation
  - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
  - Keep all rules for which confidence  $\geq \text{minconf}$
  - These are called strong rules

Frequent itemset generation is still computationally expensive.

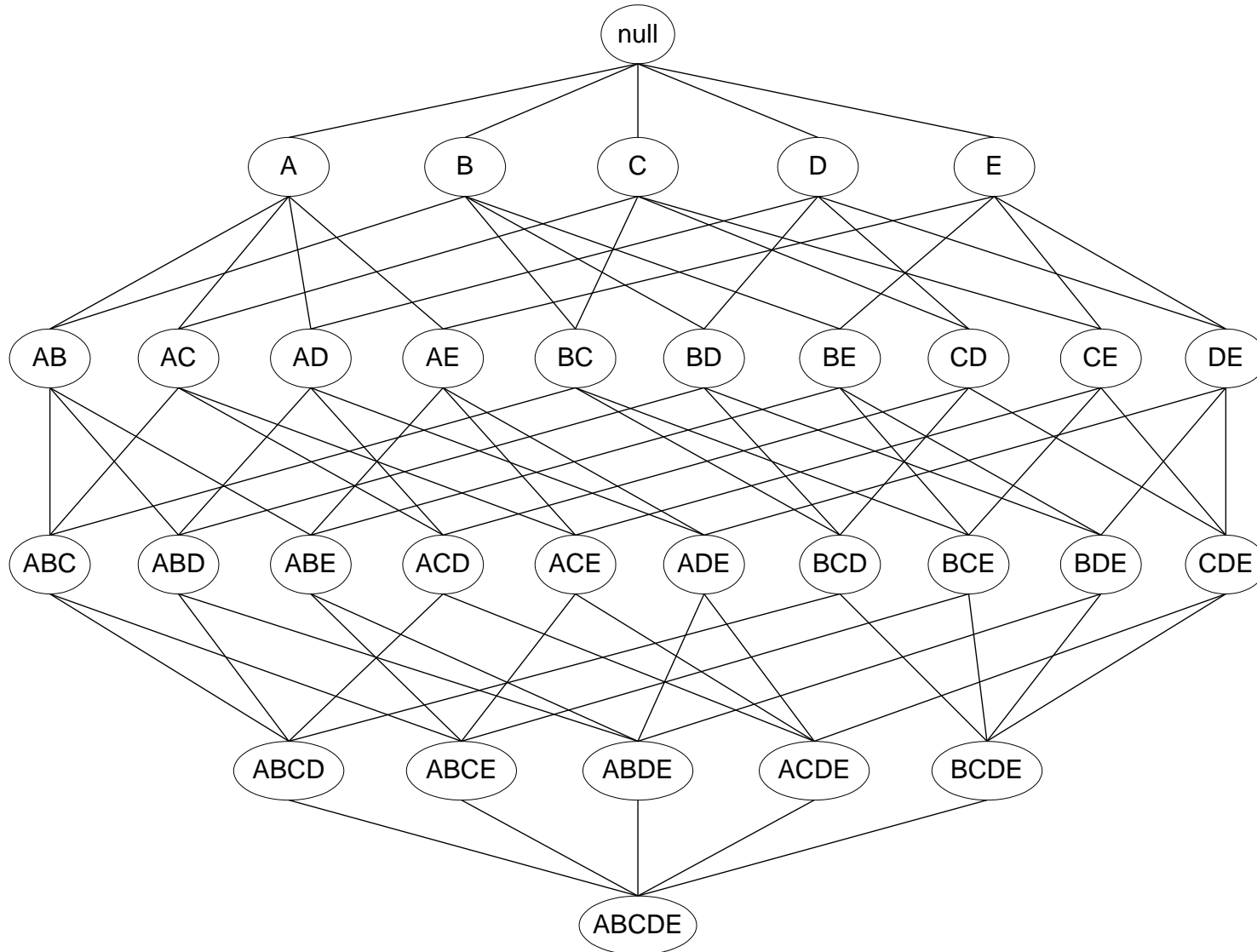
**Stop Here – Go To Evaluation of Association  
Patterns**



# Frequent Itemset Generation

# Frequent Itemset Generation

An Itemset Lattice for  $I = \{a, b, c, d, e\}$

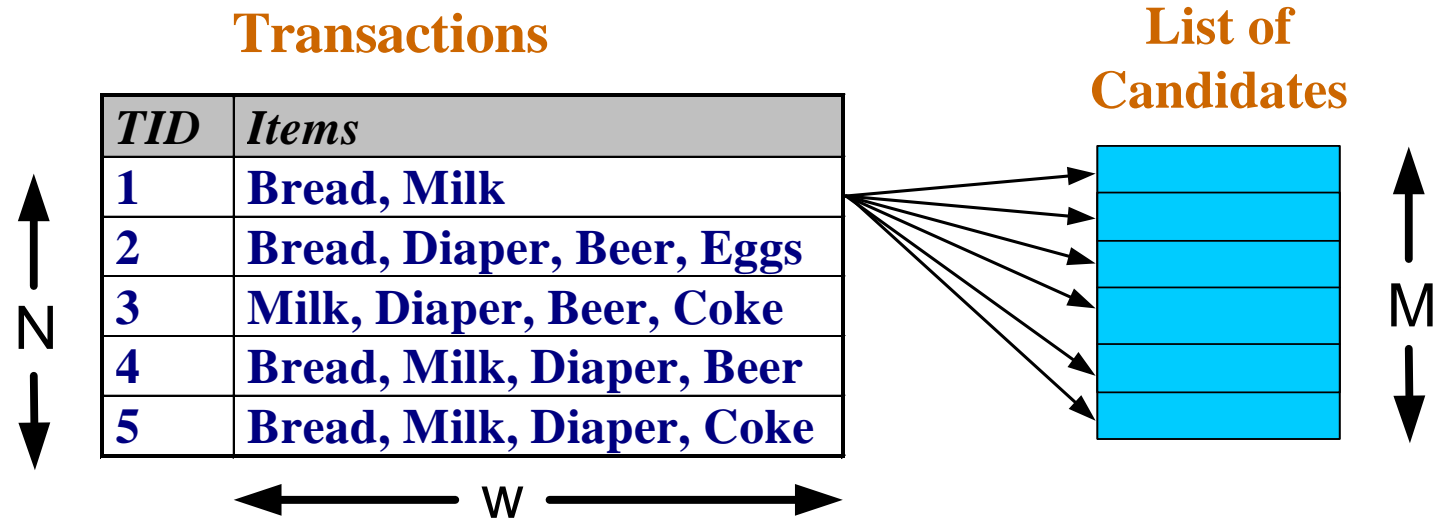


- Given  $k$  items, there are  $2^k$  possible candidate itemsets (including the null set).
- Because  $k$  can be very large in many practical applications the search space for itemsets that needs to be explored is exponentially large.

# Frequent Itemset Generation – A Brute Force Approach

## Brute-force approach:

- Each itemset in the lattice is a candidate frequent itemset
- Count the support of each candidate by scanning the list of transactions



- Match each transaction against every candidate
- Complexity  $\sim O(NMw) \Rightarrow$  Expensive since  $M = 2^k - 1!!!$

In the scanning of the list of transactions the support for the candidate itemset {Bread, Milk} is incremented three times because the itemset is contained in transactions 1, 4 and 5.

# Frequent Itemset Generation – Three Approaches to Reducing Computational Complexity

Reduce the **number of candidates** (M)

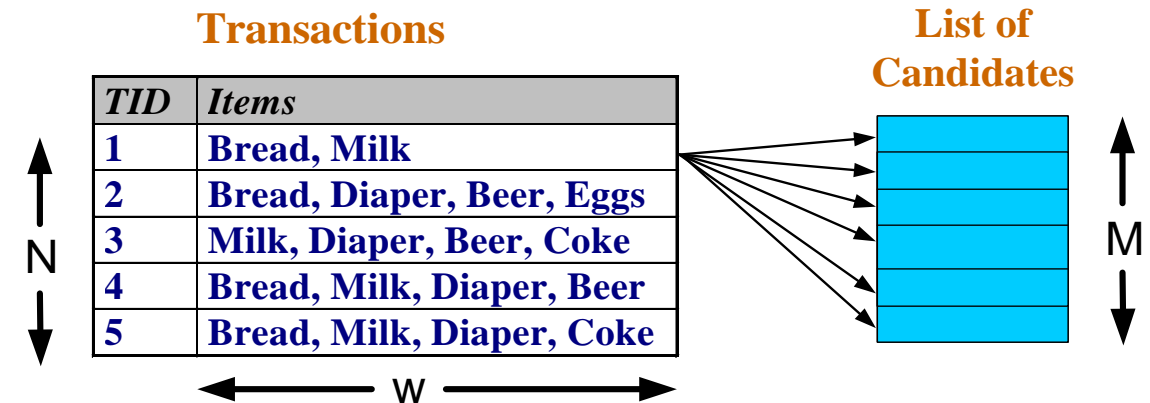
- Complete search:  $M=2^k$
- Use pruning techniques to reduce M

Reduce the **number of transactions** (N)

- Reduce size of N as the size of itemset increases
- Used by DHP and vertical-based mining algorithms

Reduce the **number of comparisons** (NM)

- Use efficient data structures to store the candidates or transactions
- No need to match every candidate against every transaction



# Frequent Itemset Generation – Reducing the Number of Candidates

## Apriori principle:

If an itemset is frequent, then all its subsets must also be frequent

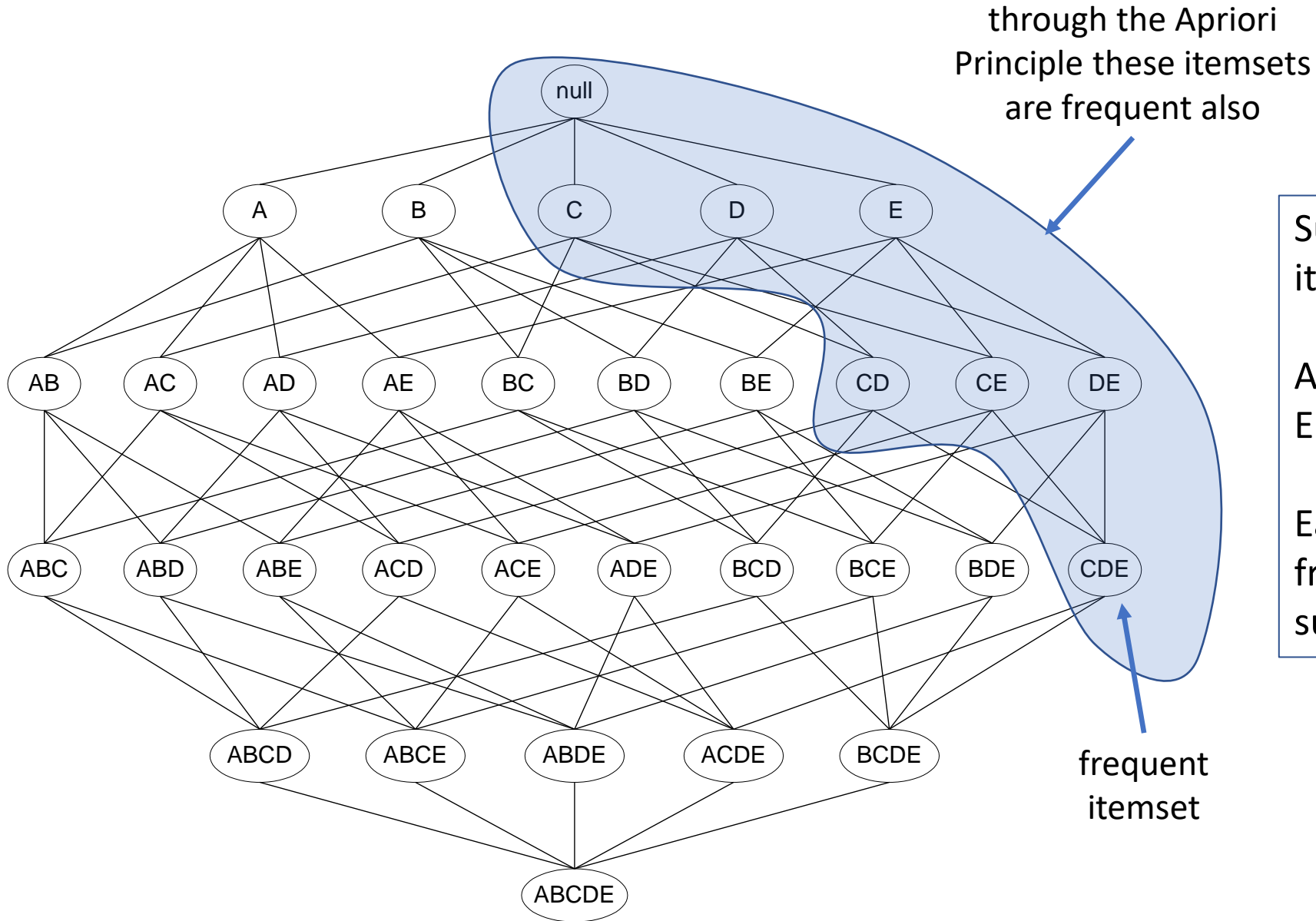
Apriori principle holds due to the following property of the support measure:

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

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

# Apriori Principle – Frequent Itemset Identification

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



Suppose  $\{C, D, E\}$  is a frequent itemset.

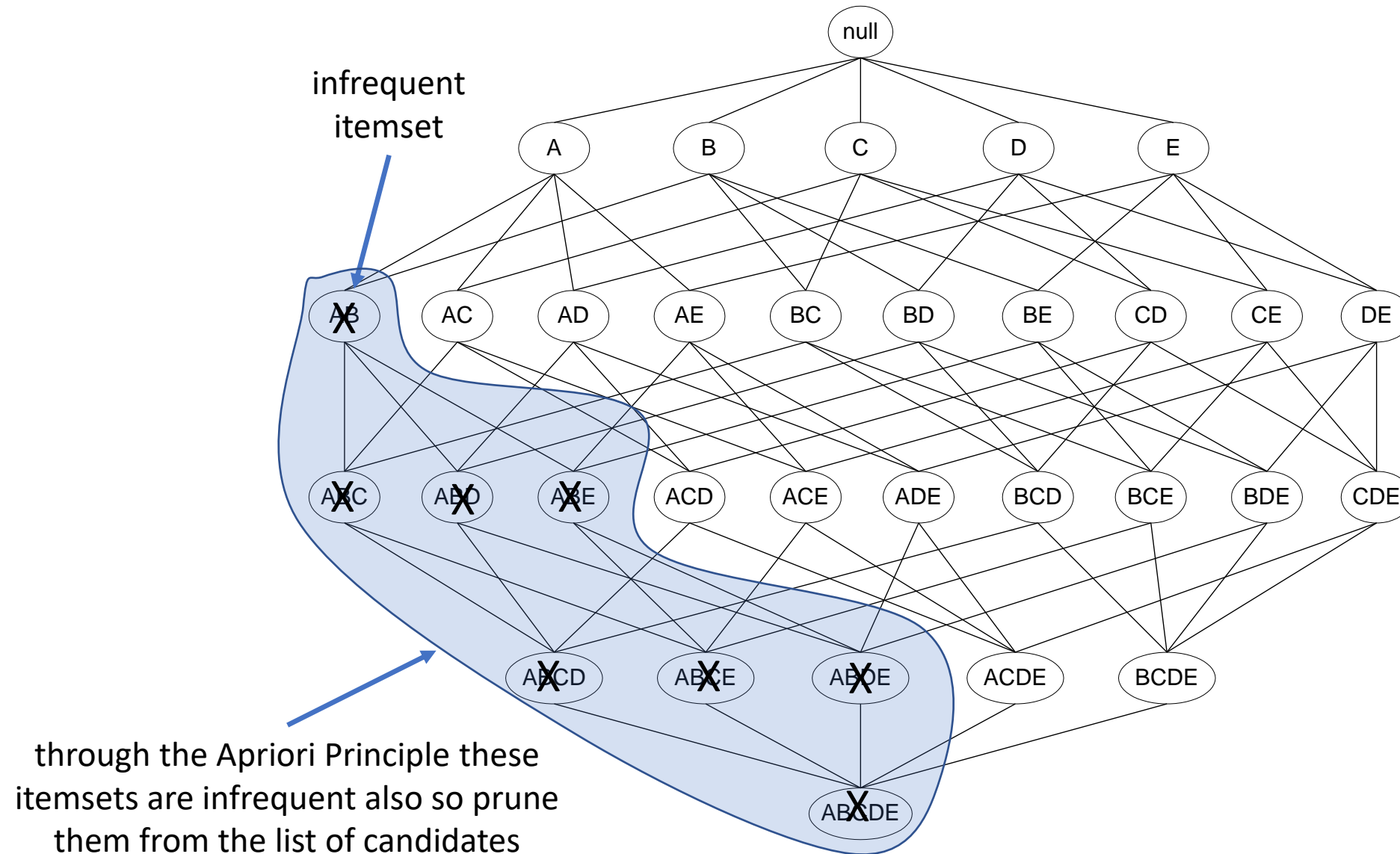
Any transaction that contains  $\{C, D, E\}$  must also contain all its subsets.

Each of these subsets is in turn a frequent itemset with at least the support of the itemset  $\{C, D, E\}$ .



# Apriori Principle – Support-Based Pruning of the Search Space

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



# An Illustration of the Apriori Principle

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

Minimum Support Count = 3

If every subset is considered,

$${}^6C_1 + {}^6C_2 + {}^6C_3 \\ 6 + 15 + 20 = 41$$

With support-based pruning,  
6

$${}^nC_r = \frac{n!}{(n-r)!r!}$$

Items  
(1-itemsets)

Item
Bread
Coke
Milk
Beer
Diaper
Eggs

# An Illustration of the Apriori Principle (continued)

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

Minimum Support Count = 3

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3$   
 $6 + 15 + 20 = 41$

With support-based pruning,  
6

Items (1-itemsets)

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

Coke and Eggs do not meet the minimum support count of 3, i.e., they are infrequent itemsets.

The Apriori Principle ensures that all supersets involving Coke and Eggs will be infrequent.

# An Illustration of the Apriori Principle (continued)

Items (1-itemsets)

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

Prune Coke and Eggs  
because all supersets will  
be infrequent.

Minimum Support Count = 3

Pairs (2-itemsets)

Itemset
{Bread,Milk}
{Bread, Beer }
{Bread,Diaper}
{Beer, Milk}
{Diaper, Milk}
{Beer,Diaper}

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3$   
 $6 + 15 + 20 = 41$   
  
With support-based pruning,  
 $6 + 6$


TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

# An Illustration of the Apriori Principle (continued)

Items (1-itemsets)

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

Prune Coke and Eggs because all  
supersets will be infrequent.



Minimum Support Count = 3

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3$   
 $6 + 15 + 20 = 41$

With support-based pruning,  
 $6 + 6$

Pairs (2-itemsets)

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

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

{Beer, Bread} and {Beer, Milk}  
do not meet the minimum  
support count of 3, i.e., they  
are infrequent itemsets.

The Apriori Principle ensures  
that all supersets involving  
{Beer, Bread} and {Beer, Milk}  
will be infrequent.

# An Illustration of the Apriori Principle (continued)

Items (1-itemsets)

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

Prune Coke and Eggs because all supersets with these will be infrequent.

Minimum Support Count = 3

If every subset is considered,  
 ${}^6C_1 + {}^6C_2 + {}^6C_3$   
 $6 + 15 + 20 = 41$

With support-based pruning,  
 $6 + 6 + 1 = 13$

Pairs (2-itemsets)

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

Prune {Beer, Bread} and {Beer, Milk} because all supersets with these will be infrequent.

Triplets (3-itemsets)

Itemset
{Bread,Diaper,Milk}

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

# An Illustration of the Apriori Principle (continued)

Items (1-itemsets)

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

Prune Coke and Eggs because all supersets will be infrequent.

Minimum Support Count = 3

Pairs (2-itemsets)

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

Prune {Beer, Bread} and {Beer, Milk} because all supersets with these will be infrequent.

If every subset is considered,

$${}^6C_1 + {}^6C_2 + {}^6C_3 \\ 6 + 15 + 20 = 41$$

With support-based pruning,

$$6 + 6 + 1 = 13$$

68% Reduction in  
itemset candidates!

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

Triplets (3-itemsets)

Itemset	Count
{Bread, Diaper, Milk}	2

# Apriori Algorithm for Frequent Itemset Generation

```
1:  $k = 1$ 
2:  $F_1 = \{i | i \in I \wedge \sigma(\{i\}) \geq N * \text{minsup}\}$            {Find all frequent 1-itemsets}
3: repeat
4:    $k = k + 1$ 
5:    $C_k = \text{candidate\_gen}(F_{k-1})$            {Generate candidate itemsets}
6:    $C_k = \text{candidate\_prune}(C_k, F_{k-1})$        {Prune candidate itemsets}
7:   for each transaction  $t \in T$  do
8:      $C_t = \text{subset}(C_k, t)$            {Identify all candidates that belong to t}
9:     for each candidate itemset  $c \in C_t$  do
10:       $\sigma(c) = \sigma(c) + 1$            {Increment support count}
11:   end for
12: end for
13:  $F_k = \{i | i \in I \wedge \sigma(\{i\}) \geq N * \text{minsup}\}$        {Find all frequent k-itemsets}
14: until  $F_k = \emptyset$ 
15:  $\text{Result} = \cup F_k$ 
```



# Apriori Algorithm - Candidate Generation

$$C_k = \text{candidate\_gen}(F_{k-1})$$

This operation generates new candidate k-itemsets based on the frequent (k-1)-itemsets found in the previous iteration.

# Apriori Algorithm - Candidate Pruning

$$C_k = \text{candidate\_prune}(C_k, F_{k-1})$$

This operation eliminates some of the candidate k-itemsets using support-based pruning, i.e., by removing k-itemsets whose subsets are known to be infrequent in previous iterations. Pruning is done without computing the actual support of these k-itemsets.

# Apriori Algorithm - Support Counting

**for** each transaction  $t \in T$  **do**

$C_t = \text{subset}(C_k, t)$

{Identify all candidates that belong to t}

**for** each candidate itemset  $c \in C_t$  **do**

$\sigma(c) = \sigma(c) + 1$

{Increment support count}

**end for**

**end for**

This operation is the process of determining the frequency of occurrence for every candidate itemset that survives the candidate pruning step.

# Apriori Algorithm - Candidate Elimination

$$F_k = \{i | i \in I \wedge \sigma(\{i\}) \geq N * \text{minsup}\}$$

This operation eliminates all itemsets whose support counts are less than  $N * \text{minsup}$  where  $N$  is the number of transactions.

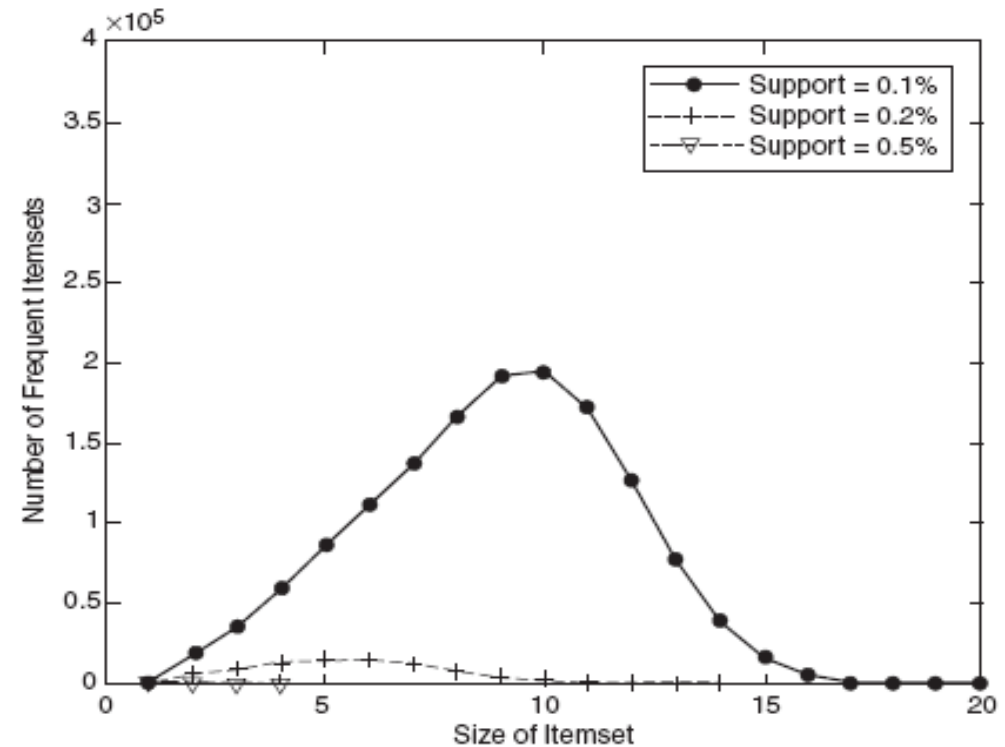
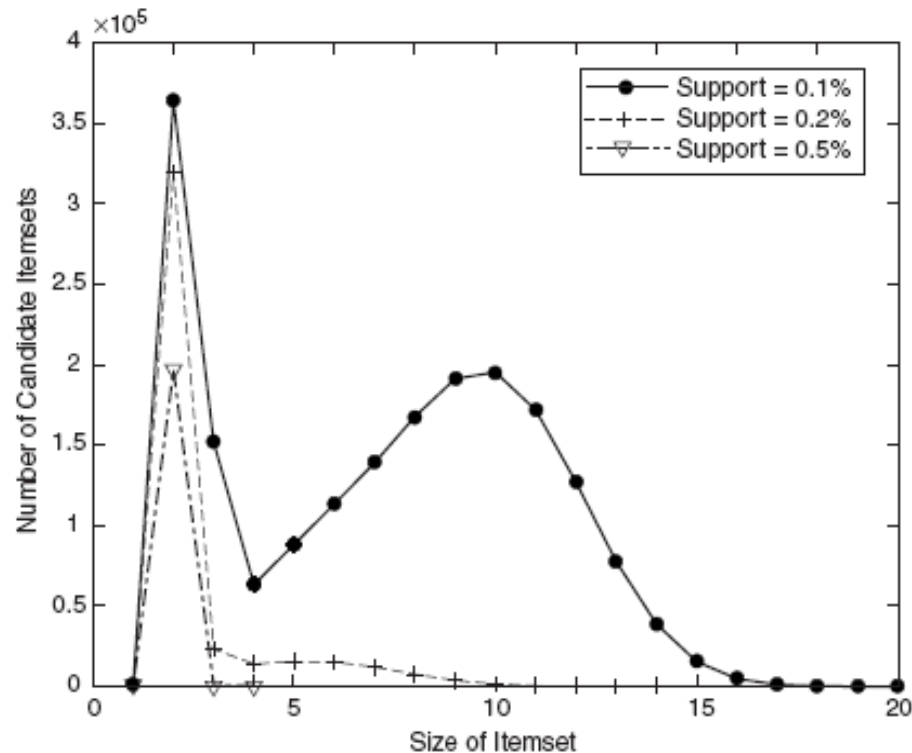
# Apriori Algorithm Computational Complexity - Runtime and Storage

## Support Threshold

Lowering the support threshold often results in:

- more itemsets being declared as frequent
- tends to increase maximum size of frequent itemsets

Both increase computational cost.



# **Apriori Algorithm Computational Complexity - Runtime and Storage**

## **Number of Items (Dimensionality)**

As the number of items in the data set increases more space will be required to store the support counts of items.

If the number of frequent itemsets also grows with the dimensionality of the data, the runtime and storage requirements will increase because of the larger number of candidate itemsets generated by the algorithm.

# Apriori Algorithm Computational Complexity - Runtime and Storage

## Number of Transactions

Because the Apriori algorithm makes repeated passes over the transaction data set, its runtime increases with a larger number of transactions.

```
7: for each transaction  $t \in T$  do  
8:    $C_t = \text{subset}(C_k, t)$                                 {Identify all candidates that belong to t}  
9:   for each candidate itemset  $c \in C_t$  do  
10:     $\sigma(c) = \sigma(c) + 1$                             {Increment support count}  
11:  end for
```

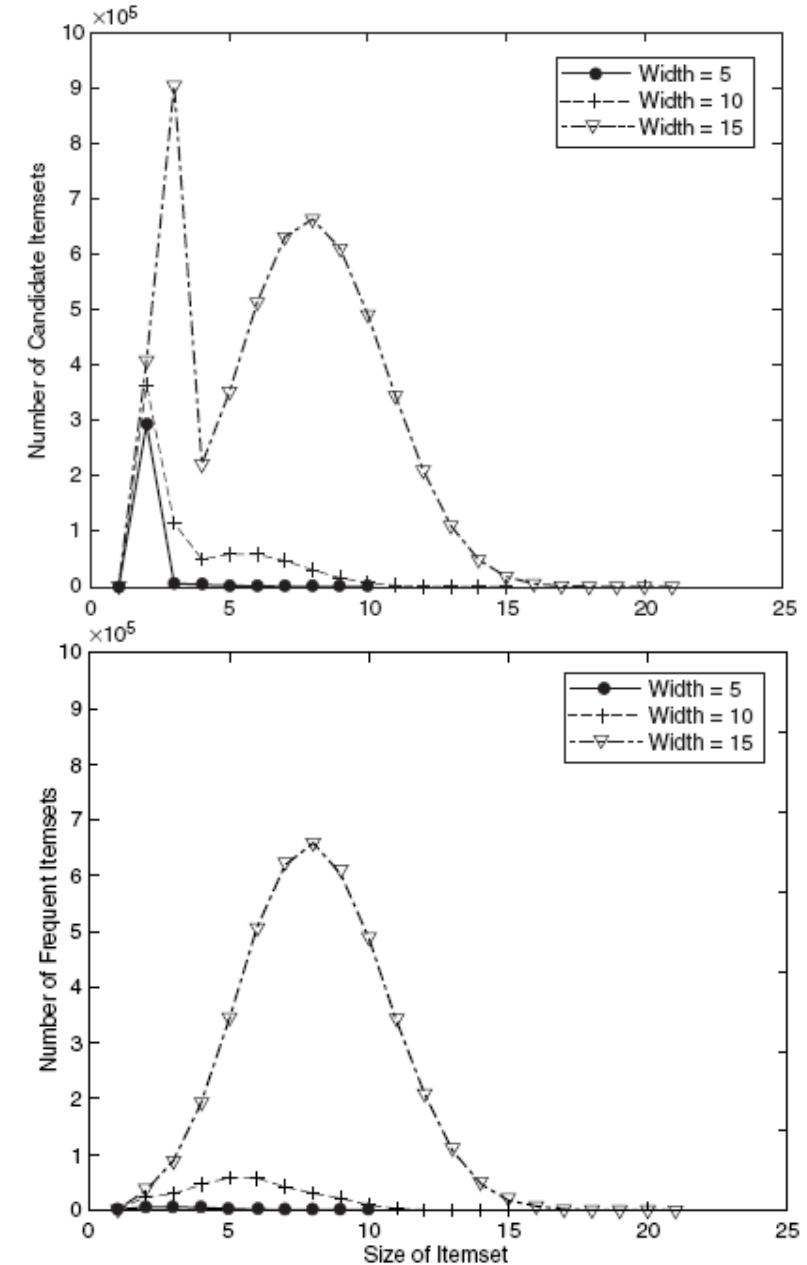
# Apriori Algorithm Computational Complexity - Runtime and Storage

## Average Transaction Width

For dense data sets the average transaction width can be very large.

This affects the complexity of the Apriori algorithm in two ways:

1. The maximum size of frequent itemsets tends to increase as the average transaction width increases. As a result, more candidate itemsets must be examined during candidate generation and support counting.
2. As the transaction width increases more itemsets are contained in a transaction. This will increase the time required for support counting.







# Rule Generation

# Rule Generation

## How to Extract Association Rules Efficiently from a Given Frequent Itemset

Each frequent  $k$ -itemset,  $Y$ , can produce up to  $2^k - 2$  association rules, ignoring rules that have empty antecedents or consequents.

An association rule can be extracted by partitioning the itemset  $Y$  into two nonempty subsets  $X$  and  $Y-X$  such that  $X \rightarrow Y - X$  satisfies the confidence threshold.

Note that all such rules meet the confidence threshold because they are generated from a frequent itemset.

# Rule Generation - Example

## How to Extract Association Rules Efficiently from a Given Frequent Itemset

Let  $Z = \{a, b, c\}$  be a frequent itemset with support  $s = \sigma(\{a, b, c\})/N$  where  $N$  is the total number of transactions.

There are six candidate association rules  $X \rightarrow Y$  that can be generated from  $Z$ :

Association Rule	Support	Confidence
$\{a, b\} \rightarrow \{c\}$	$\sigma(\{a, b\} \cup \{c\})/N$	$\sigma(\{a, b\} \cup \{c\})/\sigma(\{a, b\})$
$\{a, c\} \rightarrow \{b\}$	$\sigma(\{a, c\} \cup \{b\})/N$	$\sigma(\{a, c\} \cup \{b\})/\sigma(\{a, c\})$
$\{b, c\} \rightarrow \{a\}$	$\sigma(\{b, c\} \cup \{a\})/N$	$\sigma(\{b, c\} \cup \{a\})/\sigma(\{b, c\})$
$\{a\} \rightarrow \{b, c\}$	$\sigma(\{a\} \cup \{b, c\})/N$	$\sigma(\{a\} \cup \{b, c\})/\sigma(\{a\})$
$\{b\} \rightarrow \{a, c\}$	$\sigma(\{b\} \cup \{a, c\})/N$	$\sigma(\{b\} \cup \{a, c\})/\sigma(\{b\})$
$\{c\} \rightarrow \{a, b\}$	$\sigma(\{c\} \cup \{a, b\})/N$	$\sigma(\{c\} \cup \{a, b\})/\sigma(\{c\})$

As each of their support is identical to the support for  $Z$ , all the rules satisfy the support threshold.

Note that since  $Z$  is a frequent itemset all subsets of  $Z$  are frequent. The support counts for all frequent itemsets were found during itemset generation therefore there is no need to read the entire transaction data set again and the confidence for each rule can be calculated from previous work.

# Rule Generation – Confidence-Based Pruning

Theorem 5.2.

Let  $Y$  be an itemset and  $X$  is a subset of  $Y$ .

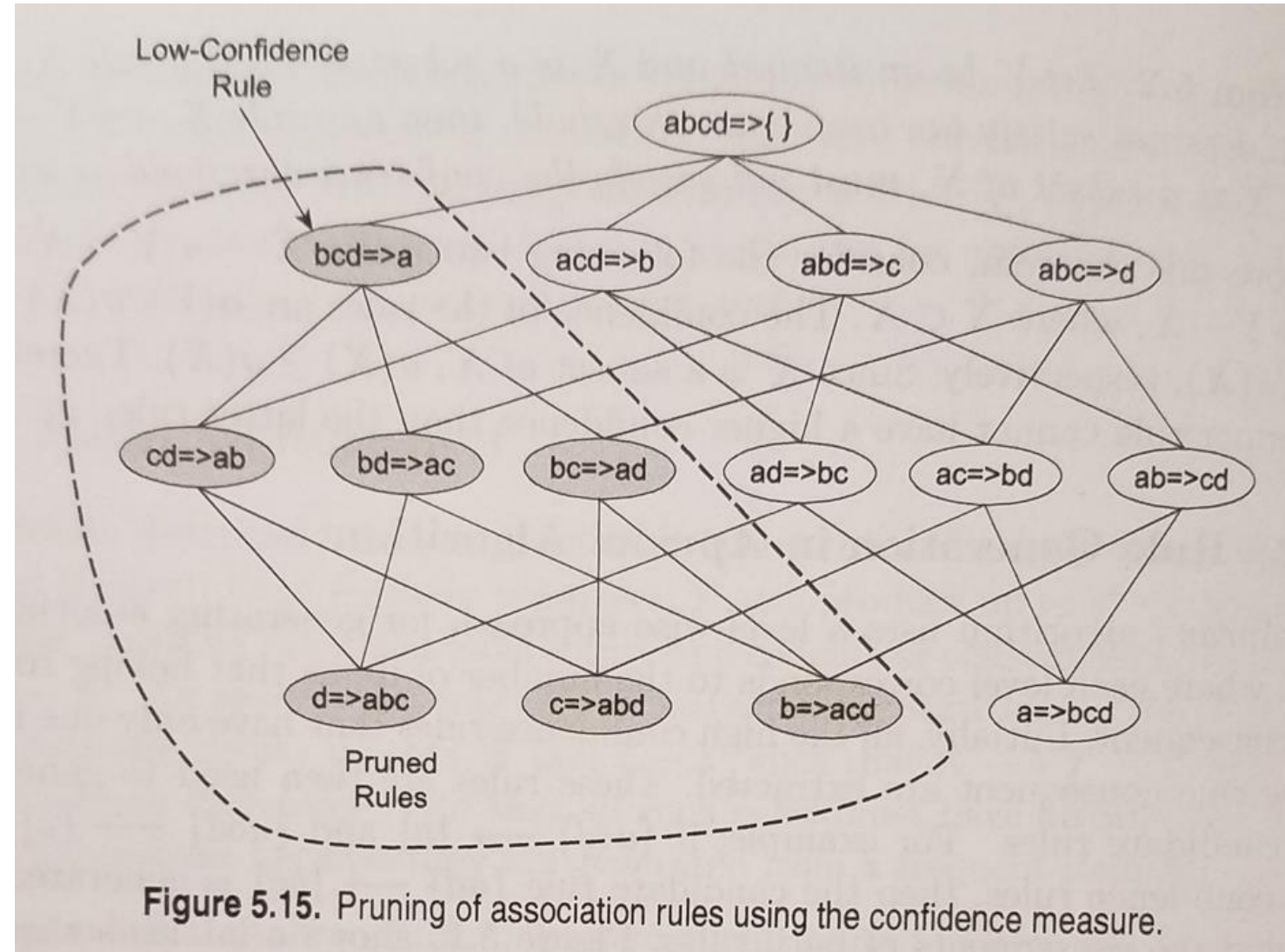
If a rule

$$X \rightarrow Y - X$$

does not satisfy the confidence threshold, then any rule

$$\tilde{X} \rightarrow Y - \tilde{X},$$

where  $\tilde{X}$  is a subset of  $X$ , must not satisfy the confidence threshold as well.

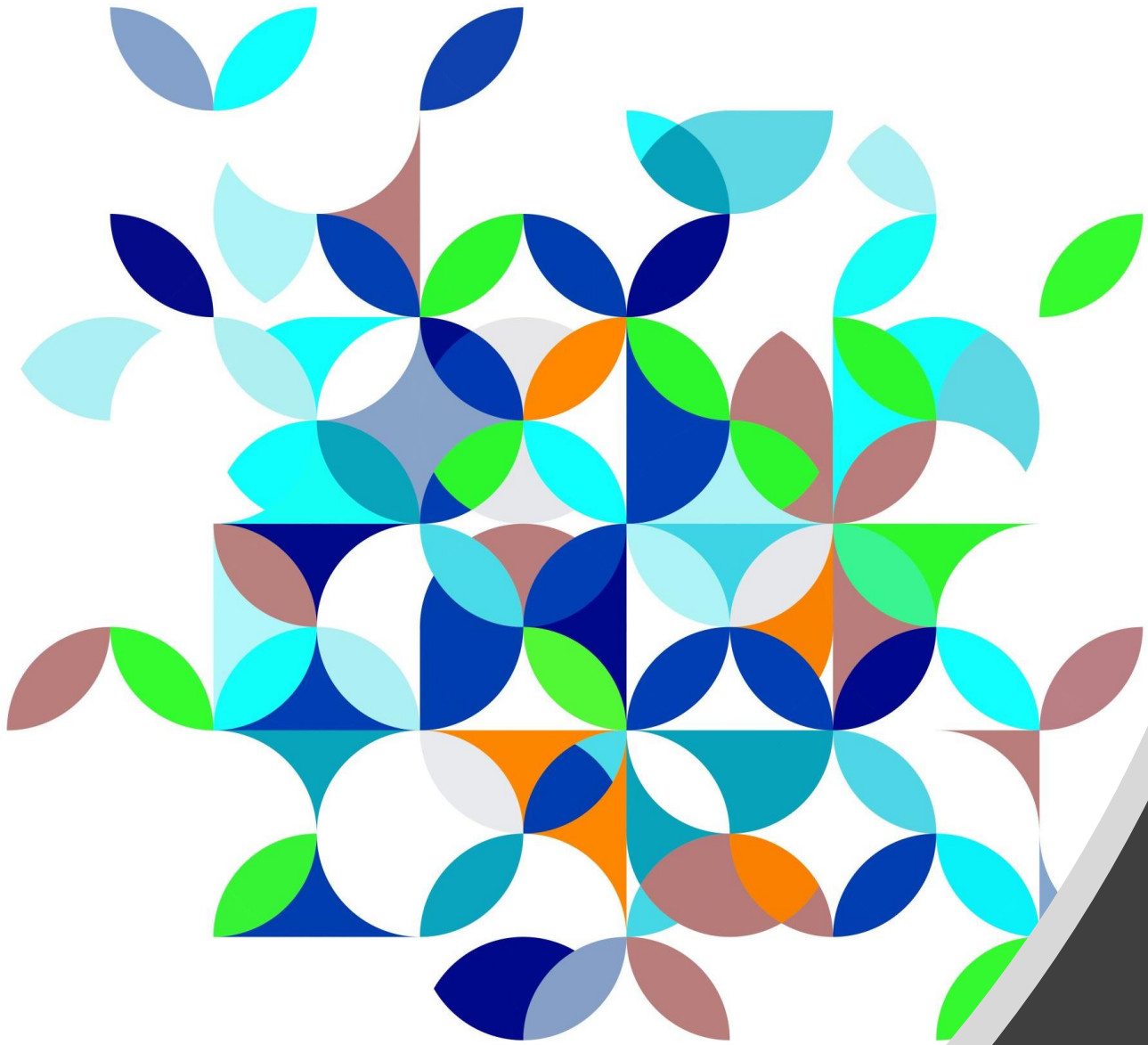


# Rule Generation in Apriori Algorithm

```
1: for each frequent k-itemset  $f_k, k \geq 2$  do  
2:    $H_1 = \{i | i \in f_k\}$                                 {1-item consequents of the rule}  
3:   call ap_genrules( $f_k, H_1$ )  
4: end for
```

Procedure ap\_genrules( $f_k, H_m$ )

```
1:  $k = |f_k|$   
2:  $m = |H_m|$   
3: if  $k > m$  then  
4:    $H_{m+1} = \text{candidate\_gen}(H_m)$   
5:    $H_{m+1} = \text{candidate\_prune}(H_{m+1}, H_m)$   
6:   for each  $h_{m+1} \in H_{m+1}$  do  
7:      $\text{conf} = \sigma(f_k) / \sigma(f_k - h_{m+1})$   
8:     if  $\text{conf} \geq \text{minconf}$  then  
9:       output the rule  $(f_k - h_{m+1}) \rightarrow h_{m+1}$   
10:    else  
11:      delete  $h_{m+1}$  from  $H_{m+1}$   
12:    endif  
13: end for  
14: call ap_genrules( $f_k, H_{m+1}$ )  
15: end if
```



# Evaluation of Association Patterns

# Evaluation of Association Patterns

The size and dimensionality of commercial databases can be very large, and we can easily end up with thousands or even millions of patterns many of which may not be very interesting.

Identifying the most interesting patterns is not a trivial task.

It is important to establish a set of well-accepted criteria for evaluating the quality of association patterns.

<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

Unique Items:

{Bread, Milk, Diaper, Eggs, Coke, Beer}

Number of Unique Items: 6

Number of Possible  
Frequent Itemsets = 64

Number of Possible  
Association Rules = 602

# Evaluation of Association Patterns

## Objective Interestingness Measures

Objective interestingness measures:

- can be used to rank itemsets or rules
- provide a straightforward way of dealing with enormous number of patterns
- can provide statistical information
- include:
  - support
  - confidence
  - others



# Evaluation of Association Patterns

## Subjective Interestingness Measures

A pattern is considered subjectively uninteresting unless it reveals unexpected information about the data or provides useful knowledge that can lead to profitable action.

For example, the rule  $\{Butter\} \rightarrow \{Bread\}$  may not be interesting despite having high support and confidence because the relationship represented by the rule might seem rather obvious.



# Evaluation of Association Patterns

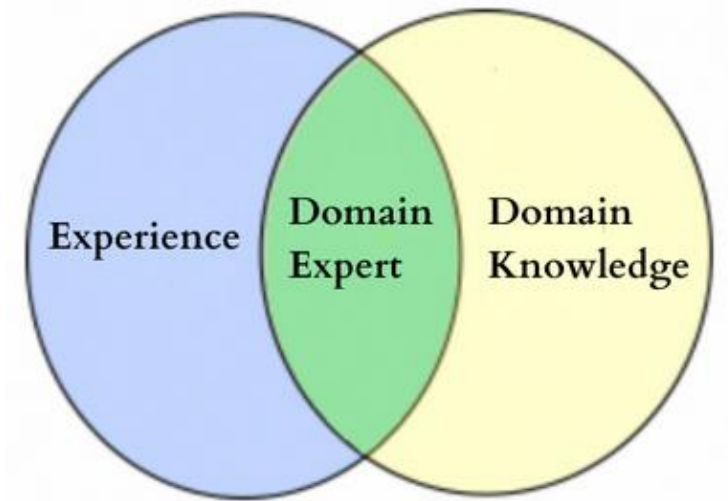
## Subjective Interestingness Measures (continued)

On the other hand, the rule  $\{Diapers\} \rightarrow \{Beer\}$  is interesting because the relationship is quite unexpected and may suggest a new cross-selling opportunity.

Incorporating subjective knowledge into pattern evaluation is a difficult task because it requires prior information from domain experts.

In this example we might need a retail domain expert to better understand the interestingness of the rule  $\{Diapers\} \rightarrow \{Beer\}$ .

In what follows we will only discuss objective interestingness measures.



# Objective Interestingness Measures

Objective interestingness measures:

- are data-driven
- domain-independent
- require only thresholds for filtering low-quality patterns
- usually computed based on frequency counts tabulated in a contingency table

# Understanding Confidence from a Statistical Perspective

Given two frequent itemsets  $X$  and  $Y$  with the association rule  $X \rightarrow Y$  the confidence of this association rule can be expressed as

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{\text{count of transactions with both } X \text{ and } Y}{\text{count of transactions with } X} = \frac{P(X \cap Y)}{P(X)} = P(Y|X)$$

association analysis measure
conditional probability

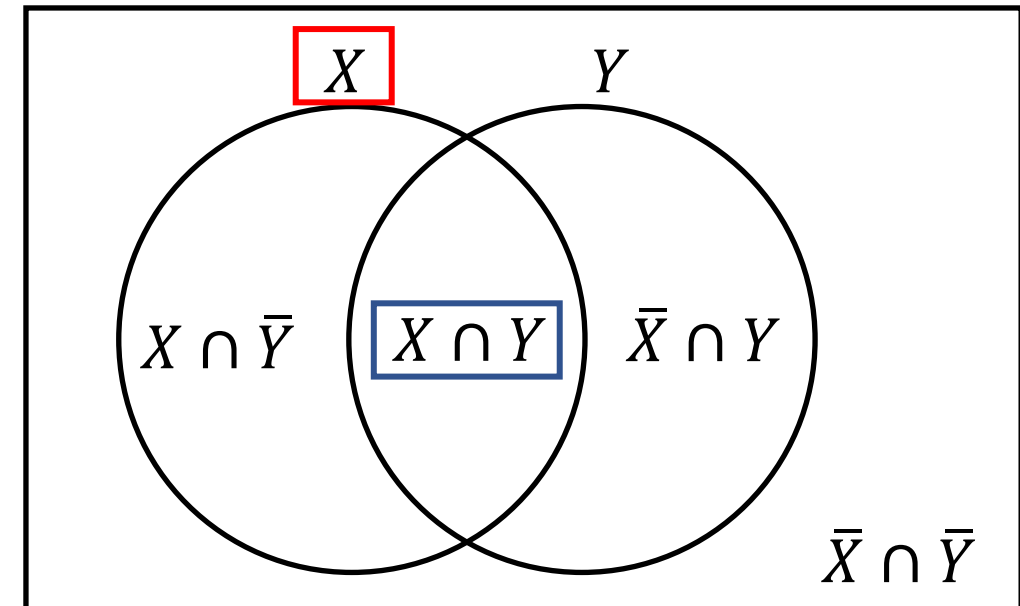
# Transactions Count Contingency

## Table for the rule $X \rightarrow Y$

$$N = |X| + |\bar{X}| = |Y| + |\bar{Y}|$$

	$Y$	$\bar{Y}$	
$X$	$ X \cap Y $	$ X \cap \bar{Y} $	$ X $
$\bar{X}$	$ \bar{X} \cap Y $	$ \bar{X} \cap \bar{Y} $	$ \bar{X} $
	$ Y $	$ \bar{Y} $	$N$

## Transactions Venn Diagram



**Stop Here – Go To Objective Measures Brief**

**objective\_measures\_brief.pptx**

# Computing Interestingness Measures

- Given  $X \rightarrow Y$  or  $\{X, Y\}$ , information needed to compute interestingness can be obtained from a contingency table.
- $\bar{X}$  ( $\bar{Y}$ ) indicate the absence of  $X$  ( $Y$ ) in a transaction.
- $f_{ij}$  denotes a frequency count:
  - $f_{11}$  is the number of times  $X$  and  $Y$  appear together in a transaction.
  - $f_{10}$  is the number of times  $X$  but not  $Y$  appear together in a transaction.

Contingency table

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

$f_{11}$ : support count of  $X$  and  $Y$

$f_{10}$ : support count of  $X$  and  $\bar{Y}$

$f_{01}$ : support count of  $\bar{X}$  and  $Y$

$f_{00}$ : support count of  $\bar{X}$  and  $\bar{Y}$



Used to define various interestingness measures such as support, confidence, Gini, entropy, etc.

# Contingency Tables in Association Analysis

## Market Basket Data

TID	Items
0	{Eggs, Milk, Onion, Yogurt, Kidney Beans, Nutmeg}
1	{Eggs, Onion, Yogurt, Dill, Kidney Beans, Nutmeg}
2	{Apple, Eggs, Kidney Beans, Milk}
3	{Milk, Yogurt, Corn, Unicorn, Kidney Beans}
4	{Eggs, Onion, Corn, Ice cream, Kidney Beans}



## Transformed Market Basket Data

TID	Apple	Corn	Dill	Eggs	Ice cream	Kidney Beans	Milk	Nutmeg	Onion	Unicorn	Yogurt
0	0	0	0	1	0	1	1	1	1	0	1
1	0	0	1	1	0	1	0	1	1	0	1
2	1	0	0	1	0	1	1	0	0	0	0
3	0	1	0	0	0	1	1	0	0	1	1
4	0	1	0	1	1	1	0	0	1	0	0



	antecedents	consequents	antecedent support	consequent support	support	confidence
0	(Eggs)	(Kidney Beans)	0.8	1.0	0.8	1.0
1	(Kidney Beans)	(Eggs)	1.0	0.8	0.8	0.8

- 5 transactions
- 11 items
- $2^{11} - 1 = 2047 =$  possible itemsets
- $3^{11} - 2^{12} + 1 = 173052$  possible association rules

Generate Frequent Itemsets (min support = 0.70):

{Eggs}

{Kidney Beans}

{Eggs, Kidney Beans}

Generate Strong Association Rules (min confidence = 0.80:

{Eggs} → {Kidney Beans}

{Kidney Beans} → {Eggs}

# Contingency Tables in Association Analysis (continued)

Transformed Market Basket Data

	TID	Apple	Corn	Dill	Eggs	Ice cream	Kidney Beans	Milk	Nutmeg	Onion	Unicorn	Yogurt
0	0	0	0	0	1	0	1	1	1	1	0	1
1	1	0	0	1	1	0	1	0	1	1	0	1
2	2	1	0	0	1	0	1	1	0	0	0	0
3	3	0	1	0	0	0	1	1	0	0	1	1
4	4	0	1	0	1	1	1	0	0	1	0	0



Transformed Market Basket Data – Only Frequent Itemsets from Strong Association Rules

	Eggs	Kidney Beans
0	1	1
1	1	1
2	1	1
3	0	1
4	1	1



Contingency Table for {Eggs} → {Kidney Beans}

	<i>Kidney Beans</i>	<i><math>\overline{Kidney Beans}</math></i>	
<i>Eggs</i>	4	0	4
<i><math>\overline{Eggs}</math></i>	1	0	1
	5	0	5

$\text{support}(\text{Eggs} \rightarrow \text{Kidney Beans}) = P(\text{Eggs} \cap \text{Kidney Beans}) = 4/5 = 0.80$   
 $\text{confidence}(\text{Eggs} \rightarrow \text{Kidney Beans}) = P(\text{Eggs} \cap \text{Kidney Beans}) / P(\text{Eggs}) = 4/4 = 1.00$

	antecedents	consequents	antecedent support	consequent support	support	confidence
0	(Eggs)	(Kidney Beans)	0.8	1.0	0.8	1.0



# Limitations of the Support-Confidence Framework

## Support

The drawback of support is that many potentially interesting patterns involving low support items might be eliminated by the support threshold. The subject of mining interesting infrequent patterns will not be covered in this course.

## Confidence

The drawback of confidence is illustrated with an example in the next three slides.

# Drawback of Confidence - Example 1: Coffee vs Tea

Suppose we are interested in analyzing the relationship between people who drink coffee and people who drink tea.

We poll a group of people about beverage preference, summarize their responses and conduct an association analysis.

# Drawback of Confidence - Example 1: Coffee vs Tea (continued)

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000



Data Set from Polling

Person	Tea	Coffee	...
1	0	1	...
2	1	0	...
3	1	1	...
4	1	0	...
...			

Association Rule: Tea  $\rightarrow$  Coffee

Support =  $150/1000 = 0.15$

Confidence =  $P(\text{Coffee} \mid \text{Tea}) = 150/200 = 0.75$

Confidence  $> 50\%$ , meaning people who drink tea are more likely to drink coffee than not drink coffee. So, the rule seems reasonable.

# Drawback of Confidence - Example 1: Coffee vs Tea (continued)

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000



Data Set from Polling

Person	Tea	Coffee	...
1	0	1	...
2	1	0	...
3	1	1	...
4	1	0	...
...			

Association Rule: Tea  $\rightarrow$  Coffee

Support =  $150/1000 = 0.15$

Confidence =  $P(\text{Coffee} \mid \text{Tea}) = 150/200 = 0.75$

But  $P(\text{Coffee}) = 0.80$ , which means knowing that a person drinks tea reduces the probability that the person drinks coffee!

Note that  $P(\text{Coffee} \mid \overline{\text{Tea}}) = 650/800 = 0.81$

**The association rule  
Coffee  $\rightarrow$  Tea is  
misleading despite its  
high confidence  
value.**

# Drawback of Confidence - Example 2: Tea vs Honey

Suppose we are interested in analyzing the relationship between people who drink tea and people who use honey.

We poll a group of people, summarize their responses and conduct an association analysis.

# Drawback of Confidence - Example 2: Tea vs Honey (continued)

Contingency Table

	<i>Honey</i>	$\overline{Honey}$	
<i>Tea</i>	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000



Data Set from Polling

Person	Tea	Honey	...
1	0	1	...
2	1	0	...
3	1	1	...
4	1	0	...
...			

Association Rule:  $Tea \rightarrow Honey$

Support =  $100/1000 = 0.10$

Confidence =  $P(Honey \mid Tea) = 100/200 = 0.50$

Confidence = 50%, which may mean that drinking tea has little influence whether honey is used or not. So, rule seems uninteresting.

But  $P(Honey) = 120/1000 = 0.12$  hence tea drinkers are far more likely to have honey. So, rule seems more interesting.

Note that  $P(Honey \mid \overline{Tea}) = 20/800 = 0.025!!$

**Using confidence  
might lead to the  
false conclusion that  
the association rule  
 $Tea \rightarrow Honey$  is  
uninteresting.**

# Understanding the Limitations of Confidence from a Statistical Perspective

The support  $s(A, B)$  of a pair of variables (events)  $A$  and  $B$  measures the probability of the two variables occurring together. Hence, the joint probability  $P(A, B)$  can be written

$$P(A, B) = s(A, B) = \frac{\sigma(A \cup B)}{N} = \frac{f_{11}}{N}$$

If we assume that  $A$  and  $B$  are statistically independent, i.e., there is no relationship between occurrences of  $A$  and  $B$ , then  $P(A, B) = P(A) \times P(B)$ .

---

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$

$f_{11}$ : support count of  $A$  and  $B$

$f_{10}$ : support count of  $A$  and  $\bar{B}$

$f_{01}$ : support count of  $\bar{A}$  and  $B$

$f_{00}$ : support count of  $\bar{A}$  and  $\bar{B}$

$$N = f_{+1} + f_{+0} = f_{1+} + f_{0+}$$

# Understanding the Limitations of Confidence from a Statistical Perspective (continued)

Hence, under the assumption of statistical independence between  $A$  and  $B$ , the support  $s_{indep}(A, B)$  of  $A$  and  $B$  can be written as

$$s_{indep}(A, B) = s(A) \times s(B) \quad \text{or equivalently} \quad s_{indep}(A, B) = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N}$$

---

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$

$f_{11}$ : support count of  $A$  and  $B$

$f_{10}$ : support count of  $A$  and  $\bar{B}$

$f_{01}$ : support count of  $\bar{A}$  and  $B$

$f_{00}$ : support count of  $\bar{A}$  and  $\bar{B}$

$$N = f_{+1} + f_{+0} = f_{1+} + f_{0+}$$



# Understanding the Limitations of Confidence from a Statistical Perspective (continued)

If the support between two variables,  $s(A, B)$  is equal to  $s_{indep}(A, B)$ , then  $A$  and  $B$  can be considered to be unrelated to each other.

However, if  $s(A, B)$  is widely different from  $s_{indep}(A, B)$ , then  $A$  and  $B$  are most likely dependent.

Hence, any deviation of  $s(A, B)$  from  $s(A) \times s(B)$  can be seen as an indication of a statistical relationship between  $A$  and  $B$ .

Since the confidence measure only considers the deviance of  $s(A, B)$  from  $s(A)$  and not from  $s(A) \times s(B)$ , it fails to account for the support of the consequent, namely  $s(B)$ .

This results in the detection of spurious patterns (e.g.,  $\{\text{Tea}\} \rightarrow \{\text{Coffee}\}$ ) and the rejection of truly interesting patterns (e.g.,  $\{\text{Tea}\} \rightarrow \{\text{Honey}\}$ ), as illustrated earlier.

# Understanding the Relationship Between the Confidence of an Association Rule and the Support of its Consequent

Consider the association rule  $X \rightarrow Y$  and the criterion that

$$c(X \rightarrow Y) = s(Y)$$

What does this mean from a statistical perspective?

We can derive a statistical relationship as follows:

$$c(X \rightarrow Y) = s(Y)$$

$$\frac{\sigma(X \cup Y)}{\sigma(X)} = \sigma(Y)$$

$$\frac{P(X \cap Y)}{P(X)} = P(Y)$$

$$P(X \cap Y) \equiv P(X, Y) = P(X) \times P(Y)$$

**Multiplication  
rule for  
independent  
events.**

As you can see this criterion means that  $X$  and  $Y$  are statistically independent.

# Understanding the Relationship Between the Confidence of an Association Rule and the Support of its Consequent (continued)

Consider the case where  $X$  and  $Y$  are not statistically independent.

If the criterion is

$$c(X \rightarrow Y) > s(Y)$$

then

$$P(X, Y) > P(X) \times P(Y)$$

and we find the  $X$  and  $Y$  are positively correlated indicating that as the occurrence of  $X$  increases so does the occurrence of  $Y$ .  $X \rightarrow Y$  is a very interesting association rule!

If the criterion is

$$c(X \rightarrow Y) < s(Y)$$

then

$$P(X, Y) < P(X) \times P(Y)$$

and we find that  $X$  and  $Y$  are negatively correlated indicating that as the occurrence of  $X$  increases the occurrence of  $Y$  will decrease.  $X \rightarrow Y$  *is not a very* interesting association rule.

# Drawback of Confidence - Example 1: Coffee vs Tea from a Statistical Perspective

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Association Rule:  $Tea \rightarrow Coffee$

quantity	value
$P(Coffee)$	$800/1000 = 0.80$
$P(Tea)$	$200/1000 = 0.20$
$P(Coffee) \times P(Tea)$	$0.80 \times 0.20 = 0.16$
$P(Coffee, Tea)$	$150/1000 = 0.15$
$c(Tea \rightarrow Coffee)$	$150/200 = 0.75$
$s(Tea \rightarrow Coffee)$	$150/1000 = 0.15$
$s(Coffee)$	$800/1000 = 0.80$

- Since  $P(Coffee) \times P(Tea) = 0.16 \neq P(Coffee, Tea) = 0.15$  we conclude that Tea and Coffee are not statistically independent.
- Since  $P(Coffee, Tea) = 0.15 < P(Coffee) \times P(Tea) = 0.16$  we conclude that Coffee and Tea are negatively correlated.
- We can see additional evidence of the negative correlation because  $c(Tea \rightarrow Coffee) = 0.75 < s(Coffee) = 0.80$ .
- If this association rule came from market basket data, the negative correlation indicates that as sales for Tea increases sales for coffee would decrease making  $Tea \rightarrow Coffee$  not a very interesting association rule even though the support and confidence for the rule are relatively high.

# Drawback of Confidence - Example 2: Tea vs Honey from a Statistical Perspective

Contingency Table

	<i>Honey</i>	$\overline{Honey}$	
<i>Tea</i>	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000

Association Rule:  $Tea \rightarrow Honey$

quantity	value
$P(Honey)$	$120/1000 = 0.12$
$P(Tea)$	$200/1000 = 0.20$
$P(Honey) \times P(Tea)$	$0.12 \times 0.20 = 0.024$
$P(Honey, Tea)$	$100/1000 = 0.10$
$c(Tea \rightarrow Honey)$	$100/200 = 0.50$
$s(Tea \rightarrow Honey)$	$120/1000 = 0.10$
$s(Honey)$	$120/1000 = 0.12$

- Since  $P(Honey) \times P(Tea) = 0.024 \neq P(Honey, Tea) = 0.10$  we conclude that Honey and Tea are not statistically independent.
- Since  $P(Honey, Tea) = 0.10 > P(Honey) \times P(Tea) = 0.024$  we conclude that Honey and Tea are positively correlated.
- We can see additional evidence of the positive correlation because  $c(Tea \rightarrow Honey) = 0.50 > s(Honey) = 0.12$ .
- If this association rule came from market basket data, the relatively low values for rule support and confidence might lead us to conclude that this was an uninteresting rule, yet we find that the probability of sales of Honey jumps from 12% to 50% when someone is buying Tea. Furthermore, due to the positive correlation, more Tea sales means more Honey sales.

# Measures for Association Rules

## What kind of rules do we really want?

Confidence( $X \rightarrow Y$ ) should be sufficiently high:

To ensure that people who buy  $X$  will more likely buy  $Y$  than not buy  $Y$ .

Confidence( $X \rightarrow Y$ )  $>$  support( $Y$ ):

Otherwise, rule will be misleading because having item  $X$  reduces the chance of having item  $Y$  in the same transaction.

Is there any measure that capture this constraint?

- Answer: Yes. There are many of them.
- They capture the deviance of  $s(A, B)$  from  $s_{indep}(A, B)$  and therefore are not susceptible to the limitations of the confidence measure.
- They account for statistical dependence.

**Stop Here – Go To Objective Measures Brief  
Slide 15**

**objective\_measures\_brief.pptx**

# Measures that Account for Statistical Dependence

Interest Factor (also known as Lift)

Piatetsky-Shapiro (PS) Measure

Correlation Analysis

IS Measure



# Measures that Account for Statistical Dependence

## Interest Factor (also known as Lift)

The interest factor  $I(A, B)$  is defined as

$$I(A, B) = \frac{s(A, B)}{s(A) \times s(B)} = \frac{N f_{11}}{f_{1+} f_{+1}}$$

$$I(A, B) \begin{cases} = 1, & \text{if } A \text{ and } B \text{ are independent} \\ > 1, & \text{if } A \text{ and } B \text{ are positively related} \\ < 1, & \text{if } A \text{ and } B \text{ are negatively related} \end{cases}$$

---

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$

$f_{11}$ : support count of  $A$  and  $B$   
 $f_{10}$ : support count of  $A$  and  $\bar{B}$   
 $f_{01}$ : support count of  $\bar{A}$  and  $B$   
 $f_{00}$ : support count of  $\bar{A}$  and  $\bar{B}$   
 $N = f_{+1} + f_{+0} = f_{1+} + f_{0+}$

# Measures that Account for Statistical Dependence

## Interest Factor (also known as Lift) – An Example

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Association Rule:  $Tea \rightarrow Coffee$

Quantity	Value
$s(Tea \rightarrow Coffee)$	0.15
$c(Tea \rightarrow Coffee) = P(Coffee \mid Tea)$	$150/200 = 0.75$
$s(Coffee) = P(Coffee)$	$800/1000 = 0.80$
$I(Tea, Coffee)$	$(150/1000)/((200/1000)*(800/1000)) = 0.9375$

$$I(A, B) = \frac{s(A, B)}{s(A) \times s(B)} = \frac{Nf_{11}}{f_{1+}f_{+1}}$$

- Confidence =  $P(Coffee \mid Tea) = 0.75$  but  $P(Coffee) = 0.80!!$
- Note that  $I = 0.15 / (0.2 \times 0.8) = 0.9375$  therefore Coffee and Tea are not statistically independent and since  $I < 1$  they are negatively associated.
- While the support and confidence are relatively high the negative association makes this rule uninteresting.
- Is it enough to use confidence and interest for pruning?

# Measures that Account for Statistical Dependence

## Piatesky-Shapiro (PS) Measure

The Piatesky-Shapiro (PS) measure is defined as

$$PS = s(A, B) - s(A) \times s(B) = \frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$$

$$PS(A, B) \begin{cases} = 0, \text{ if } A \text{ and } B \text{ are independent} \\ > 0, \text{ if } A \text{ and } B \text{ have a positive relationship} \\ < 0, \text{ if } A \text{ and } B \text{ have a negative relationship} \end{cases}$$

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$

$f_{11}$ : support count of  $A$  and  $B$   
 $f_{10}$ : support count of  $A$  and  $\bar{B}$   
 $f_{01}$ : support count of  $\bar{A}$  and  $B$   
 $f_{00}$ : support count of  $\bar{A}$  and  $\bar{B}$   
 $N = f_{+1} + f_{+0} = f_{1+} + f_{0+}$

# Measures that Account for Statistical Dependence

## Correlation Analysis

The correlation analysis between two binary variables can be measured using the  $\varphi$  – *coefficient* which is defined as

$$\varphi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}} = \frac{s(A, B) - s(A) \times s(B)}{\sqrt{s(A) \times (1 - s(A)) \times s(B) \times (1 - s(B))}}$$

The  $\varphi$  – *coefficient* is a normalized version of the PS measure and ranges from -1 to +1.

- A value of -1 is a perfect negative relationship.
- A value of 0 means no relationship.
- A value of +1 is a perfect positive relationship.

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$

$f_{11}$ : support count of  $A$  and  $B$   
 $f_{10}$ : support count of  $A$  and  $\bar{B}$   
 $f_{01}$ : support count of  $\bar{A}$  and  $B$   
 $f_{00}$ : support count of  $\bar{A}$  and  $\bar{B}$   
 $N = f_{+1} + f_{+0} = f_{1+} + f_{0+}$

# Measures that Account for Statistical Dependence: Correlation Analysis – An Example

$$\varphi = \frac{s(A, B) - s(A) \times s(B)}{\sqrt{s(A) \times (1 - s(A)) \times s(B) \times (1 - s(B))}}$$

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

$$\varphi = -0.0625$$

Association Rule: Tea → Coffee

quantity	value
P(Coffee)	800/1000 = 0.80
P(Tea)	200/1000 = 0.20
P(Coffee) x P(Tea)	0.80 x 0.20 = 0.16
P(Coffee, Tea)	150/1000 = 0.15
c(Tea → Coffee)	150/200 = 0.75
s(Tea → Coffee)	150/1000 = 0.15
s(Coffee)	800/1000 = 0.80

Contingency Table

	<i>Honey</i>	$\overline{Honey}$	
<i>Tea</i>	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000

$$\varphi = 0.5847$$

Association Rule: Tea → Honey

quantity	value
P(Honey)	120/1000 = 0.12
P(Tea)	200/1000 = 0.20
P(Honey) x P(Tea)	0.12 x 0.20 = 0.024
P(Honey, Tea)	100/1000 = 0.10
c(Tea → Honey)	100/200 = 0.50
s(Tea → Honey)	120/1000 = 0.10
s(Honey)	120/1000 = 0.12

# Measures that Account for Statistical Dependence

## IS Measure

The IS measure is defined as

$$IS = \sqrt{I(A, B) \times s(A, B)} = \frac{s(A, B)}{\sqrt{s(A) \times s(B)}} = \frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}$$

Since IS is the geometric mean between the interest factor and the support of a pattern, IS is large when both the interest factor and the support are large. Hence if the interest factor of two patterns are identical, the IS has a preference of selecting the pattern with the higher support.

The value of IS varies from 0 to 1 where:

- IS = 0 corresponds to no co-occurrence of the two variables
- IS = 1 corresponds to perfect co-occurrence of the two variables

# Measures that Account for Statistical Dependence: IS Measure – An Example

$$IS = \frac{s(A, B)}{\sqrt{s(A) \times s(B)}}$$

Contingency Table

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Association Rule: Tea → Coffee

$$IS = 0.375$$

quantity	value
P(Coffee)	800/1000 = 0.80
P(Tea)	200/1000 = 0.20
P(Coffee) x P(Tea)	0.80 x 0.20 = 0.16
P(Coffee, Tea)	150/1000 = 0.15
c(Tea → Coffee)	150/200 = 0.75
s(Tea → Coffee)	150/1000 = 0.15
s(Coffee)	800/1000 = 0.80

Contingency Table

	<i>Honey</i>	$\overline{Honey}$	
<i>Tea</i>	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000

Association Rule: Tea → Honey

$$IS = 0.6455$$

quantity	value
P(Honey)	120/1000 = 0.12
P(Tea)	200/1000 = 0.20
P(Honey) x P(Tea)	0.12 x 0.20 = 0.024
P(Honey, Tea)	100/1000 = 0.10
c(Tea → Honey)	100/200 = 0.50
s(Tea → Honey)	120/1000 = 0.10
s(Honey)	120/1000 = 0.12

# A Survey of Measures that Capture the Deviance Between $s(A, B)$ and $s_{indep}(A, B)$

The multitude of measures with varying ranges will have different behaviors when applied to the evaluation of association rules.

The measures we have defined above are not exhaustive and there exist many alternative measures for capturing different properties of relationships between pairs of binary variables.

The table on the next slide provides some of the definitions for these measures in terms of the frequency counts of a 2 x 2 contingency table.



# A Survey of Measures that Capture the Deviance Between $s(A, B)$ and $s_{indep}(A, B)$ (continued)

Other measures  
in the literature.

Measure (Symbol)	Definition
Correlation ( $\phi$ )	$\frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$
Odds ratio ( $\alpha$ )	$(f_{11} f_{00}) / (f_{10} f_{01})$
Kappa ( $\kappa$ )	$\frac{N f_{11} + N f_{00} - f_{1+} f_{+1} - f_{0+} f_{+0}}{N^2 - f_{1+} f_{+1} - f_{0+} f_{+0}}$
Interest ( $I$ )	$(N f_{11}) / (f_{1+} f_{+1})$
Cosine ( $IS$ )	$(f_{11}) / (\sqrt{f_{1+} f_{+1}})$
Piatetsky-Shapiro ( $PS$ )	$\frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2}$
Collective strength ( $S$ )	$\frac{f_{11} + f_{00}}{f_{1+} f_{+1} + f_{0+} f_{+0}} \times \frac{N - f_{1+} f_{+1} - f_{0+} f_{+0}}{N - f_{11} - f_{00}}$
Jaccard ( $\zeta$ )	$f_{11} / (f_{1+} + f_{+1} - f_{11})$
All-confidence ( $h$ )	$\min \left[ \frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}} \right]$

# Consistency Among Objective Interestingness Measures

Can the measures available produce similar ordering results when applied to a set of association patterns?

If the measures are consistent, then we can choose any one of them as our evaluation metric.

Otherwise, it is important to understand what their differences are in order to determine which measure is more suitable for analyzing certain types of patterns.

# Consistency Among Objective Interestingness Measures (continued)

Consider the 10 contingency tables below.  
They have been selected to illustrate the differences among the measures.

The rankings of the 10 contingency tables by each of the interestingness measures are also given.

Contingency Tables

Exam ple	f <sub>11</sub>	f <sub>10</sub>	f <sub>01</sub>	f <sub>00</sub>
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 by Measure Type  
(1 is most interesting and 10 is least interesting)

	$\phi$	$\alpha$	$\kappa$	$I$	$IS$	$PS$	$S$	$\zeta$	$h$
$E_1$	1	3	1	6	2	2	1	2	2
$E_2$	2	1	2	7	3	5	2	3	3
$E_3$	3	2	4	4	5	1	3	6	8
$E_4$	4	8	3	3	7	3	4	7	5
$E_5$	5	7	6	2	9	6	6	9	9
$E_6$	6	9	5	5	6	4	5	5	7
$E_7$	7	6	7	9	1	8	7	1	1
$E_8$	8	10	8	8	8	7	8	8	7
$E_9$	9	4	9	10	4	9	9	4	4
$E_{10}$	10	5	10	1	10	10	10	10	10

# Consistency Among Objective Interestingness Measures (continued)

The results shown on the last slide suggests that the measures greatly differ from each other and can provide conflicting information about the interestingness of a pattern.

No measure is universally best for all applications.

Next, we study some of the properties of the measures that play an important role in determining if they are suited for a certain application.

# Properties of Objective Interestingness Measures

## Inversion Property

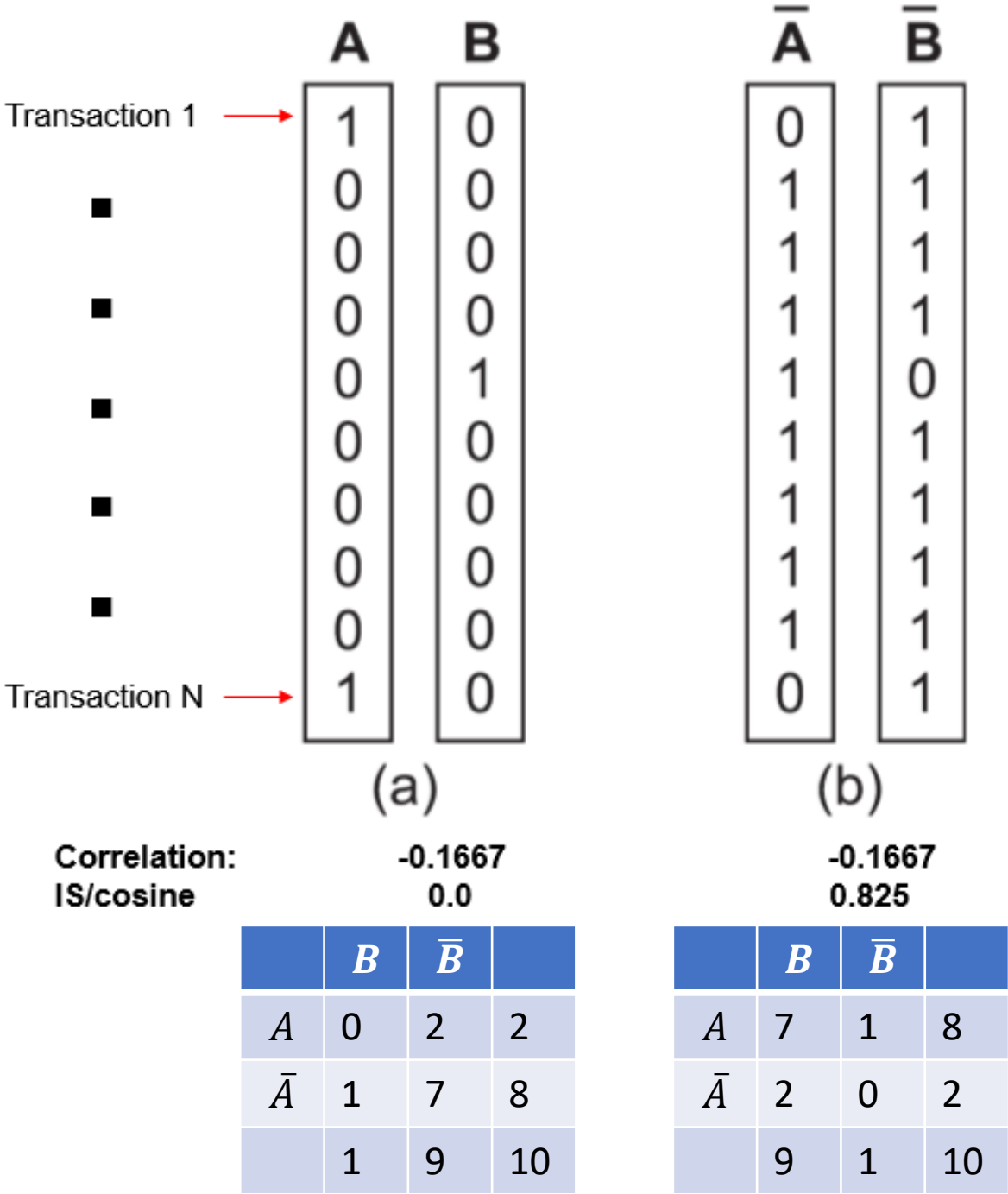
Consider the binary vectors  $A$  and  $B$ .

The 0/1 value in each column vector indicates whether a transaction (row) contains a particular itemset (column).

The vectors  $\bar{A}$  and  $\bar{B}$  are the inverted versions of  $A$  and  $B$ , i.e., the inversion transformation has been applied to the binary vectors.

If a measure is invariant under the inversion operation, then its value for the vector pair  $\{\bar{A}, \bar{B}\}$  should be identical to its value for  $\{A, B\}$ .

An objective measure  $M$  is invariant under the inversion operation if its value remains unchanged when exchanging the frequency counts  $f_{11}$  with  $f_{00}$  and  $f_{10}$  with  $f_{01}$ .



# Properties of Objective Interestingness Measures

## Inversion Property (continued)

Measures that are invariant under the inversion operation include:

- Correlation ( $\phi$  – coefficient )
- Odds Ratio
- Kappa
- Collective Strength

These measures are useful when the presence (1's) of a variable is as important as its absence (0's), i.e., symmetric binary variables.

Measures that are not invariant under the inversion operation include:

- Interest Factor
- IS

These measures are useful for asymmetric binary variables where we are looking to capture relationships based on the presence of a variable.



invariant under the inversion



not invariant under the inversion

Measure (Symbol)	Definition
Correlation ( $\phi$ )	$\frac{N f_{11} - f_{1+} f_{+1}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$
Odds ratio ( $\alpha$ )	$(f_{11} f_{00}) / (f_{10} f_{01})$
Kappa ( $\kappa$ )	$\frac{N f_{11} + N f_{00} - f_{1+} f_{+1} - f_{0+} f_{+0}}{N^2 - f_{1+} f_{+1} - f_{0+} f_{+0}}$
Interest ( $I$ )	$(N f_{11}) / (f_{1+} f_{+1})$
Cosine ( $IS$ )	$(f_{11}) / (\sqrt{f_{1+} f_{+1}})$
Piatetsky-Shapiro ( $PS$ )	$\frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2}$
Collective strength ( $S$ )	$\frac{f_{11} + f_{00}}{f_{1+} f_{+1} + f_{0+} f_{+0}} \times \frac{N - f_{1+} f_{+1} - f_{0+} f_{+0}}{N - f_{11} - f_{00}}$
Jaccard ( $\zeta$ )	$f_{11} / (f_{1+} + f_{+1} - f_{11})$
All-confidence ( $h$ )	$\min \left[ \frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}} \right]$

# Properties of Objective Interestingness Measures: Scaling Property

Consider the two contingency tables for mask use and susceptibility to Covid.

These tables are used to study the relationship between mask wearing and susceptibility to Covid.

The second contingency table has data from the same population but has two times as many Mask wearers and 3 times as many Covid-Free people.

The underlying association should be independent of the relative number of Mask wearing and Covid-Free subjects.

Let  $T$  be a contingency table with frequency counts  $[f_{11}; f_{10}; f_{01}; f_{00}]$ . Let  $T'$  be the transformed contingency table with scaled frequency counts  $[k_1k_3f_{11}; k_2k_3f_{10}; k_1k_4f_{01}; k_2k_4f_{00}]$ , where  $k_1, k_2, k_3, k_4$  are positive constants used to scale the two rows and the two columns of  $T$ . An objective measure  $M$  is invariant under the row/column scaling operation if  $M(T) = M(T')$ .

$$Odds - Ratio = \frac{f_{11}f_{00}}{f_{01}f_{10}}$$

	Covid-Positive	Covid-Free	
Mask	20	30	50
No-Mask	40	10	50
	60	40	100

Odds-Ratio = 200/1200 = 0.1667

	Covid-Positive	Covid-Free	
Mask	40	180	340
No-Mask	40	30	180
	120	400	520

2x →

↓ 3x

Odds-Ratio = 1200/7200 = 0.1667

# Properties of Objective Interestingness Measures

## Scaling Property (continued)

The odds-ratio is the only objective measure that is invariant under contingency table row/column scaling.

In general, the frequency of items in a contingency table closely depends on the sample of transactions used to generate the table.

Any change in the sampling procedure may affect a row and/or column scaling transformation.

A measure that is expected to be invariant to differences in the sampling procedure must not change with row and/or column scaling.



# Properties of Objective Interestingness Measures: Null Addition Property

Suppose we are interested in analyzing the relationship between a pair of words, such as data and mining, in a set of documents.

If a collection of articles about ice fishing is added to the data set, should the association between data and mining be affected?

This process of adding unrelated data (in this case, documents) to a given data set is known as the null addition operation.

An objective measure  $M$  is invariant under the null addition operation if it is not affected by increasing  $f_{00}$ , while all other frequencies in the contingency table stay the same.

$$\text{Jaccard} = \zeta = \frac{f_{11}}{f_{1+} + f_{+1} - f_{11}}$$

	$B$	$\overline{B}$	
$A$	700	100	800
$\overline{A}$	100	100	200
	800	200	1000

$$\text{Jaccard} = 0.78$$

	$B$	$\overline{B}$	
$A$	700	100	800
$\overline{A}$	100	1100	1200
	800	1200	2000

$$\text{Jaccard} = 0.78$$

# Properties of Objective Interestingness Measures

## Null Addition Property (continued)

For applications such as document analysis or market basket analysis we would like objective measures that remains invariant under the null addition operation.

Otherwise, the relationship between words can be made to change simply by adding enough documents that do not contain both words.

Measures that satisfy the null addition property include:

- IS
- Jaccard

Measures that violate the null addition property include:

- Interest Factor
- PS
- Odds-Ratio
- Correlation ( $\varphi$  – *coefficient*)

# Properties of Objective Interestingness Measures

## A Summary

Symbol	Measure	Inversion	Null Addition	Scaling
$\phi$	$\phi$ -coefficient	Yes	No	No
$\alpha$	odds ratio	Yes	No	Yes
$\kappa$	Kappa	Yes	No	No
$I$	Interest	No	No	No
$IS$	Cosine	No	Yes	No
$PS$	Piatetsky-Shapiro's	Yes	No	No
$S$	Collective strength	Yes	No	No
$\zeta$	Jaccard	No	Yes	No
$h$	All-confidence	No	Yes	No
$s$	Support	No	No	No

# Properties of Objective Interestingness Measures

Based on these properties which objective measure is best for Market Basket Analysis?

The green boxes indicate useful properties for each measure. It looks like for rules evaluation cosine ( $IS$ ) and Jaccard ( $\zeta$ ) are best. For itemset evaluation it looks like All-Confidence ( $h$ ) would be best.

Symbol	Measure	Inversion	Null Addition	Scaling
$\phi$	$\phi$ -coefficient	Yes	No	No
$\alpha$	odds ratio	Yes	No	Yes
$\kappa$	Kappa	Yes	No	No
$I$	Interest	No	No	No
$IS$	Cosine	No	Yes	No
$PS$	Piatetsky-Shapiro's	Yes	No	No
$S$	Collective strength	Yes	No	No
$\zeta$	Jaccard	No	Yes	No
$h$	All-confidence	No	Yes	No
$s$	Support	No	No	No

# Simpson's Paradox

An observed association between variables may be influenced by the presence of other confounding factors, i.e., variables, that are not included in the analysis.

The hidden variables may cause the observed association to disappear or reverse its direction, a phenomenon known as Simpson's Paradox.

# Simpson's Paradox – An Example

---

Contingency table of the association between buying an HDTV and buying an exercise machine.

The association rule and resulting counts were extracted from a transaction data set.

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

---

Evaluation of the Association Pattern Using Confidence:

$\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}: c = 99/180 = 0.55$

$\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}: c = 54/120 = 0.45$

Customers ***who do buy HDTVs*** are more likely to buy exercise machines.

---

# Simpson's Paradox – An Example (continued)

The analytics team returned to the transaction data and stratified it by age group: college students vs working adults.

A three-way contingency table was generated using the new variable Customer Group.

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

Evaluation of the Association Patterns With Stratification by Age Group Using Confidence:

## College Students

{HDTV = Yes}  $\rightarrow$  {Exercise Machine = Yes}:  $c = 1/10 = 0.100$

{HDTV = No}  $\rightarrow$  {Exercise Machine = Yes}:  $c = 4/34 = 0.118$

## Working Adults

{HDTV = Yes}  $\rightarrow$  {Exercise Machine = Yes}:  $c = 98/170 = 0.577$

{HDTV = No}  $\rightarrow$  {Exercise Machine = Yes}:  $c = 50/86 = 0.581$

**This reversal in the direction of association is known as Simpson's Paradox.**

Customers ***who do not buy HDTVs*** are more likely to buy exercise machines.

# Simpson's Paradox – An Example (continued)

## Understanding this Reversal of Association Direction

$$\{\text{HDTV}=\text{Yes}\} \rightarrow \{\text{Working Adult}\}: c = (98+72)/(1+9+98+72) = 170/180 = 0.944$$

Most customers who buy HDTVs are working adults.

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

$$\{\text{Exercise Machine} = \text{Yes}\} \rightarrow \{\text{Working Adult}\}: c = (98+50)/(1+4+98+50) = 148/153 = 0.967$$

Most customers who buy exercise machines are working adults.

Since working adults form the largest fraction of customers for both HDTVs and exercise machines, they both look related and the rule  $\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}$  turns out to be stronger in the combined data than what it would have been if the data is stratified.

Customer group acts as a hidden variable that affects both the fraction of customers who buy HDTVs and those who buy exercise machines.

If we factor out the effect of the hidden variable by stratifying the data, we see that the relationship between buying HDTVs and exercise machines is not direct but show up as an indirect consequence of the effect of the hidden variable.



# Simpson's Paradox – The Lesson

Proper stratification is needed to avoid generating spurious patterns resulting from Simpson's Paradox.

## Example 1:

Market basket data from a major supermarket chain should be stratified according to store location.

## Example 2:

Medical records from various patients should be stratified according to confounding factors such as age and gender.

# Simpson's Paradox – The Lesson

Proper stratification is needed to avoid generating spurious patterns resulting from Simpson's Paradox.

## Example 1:

Market basket data from a major supermarket chain should be stratified according to store location.

## Example 2:

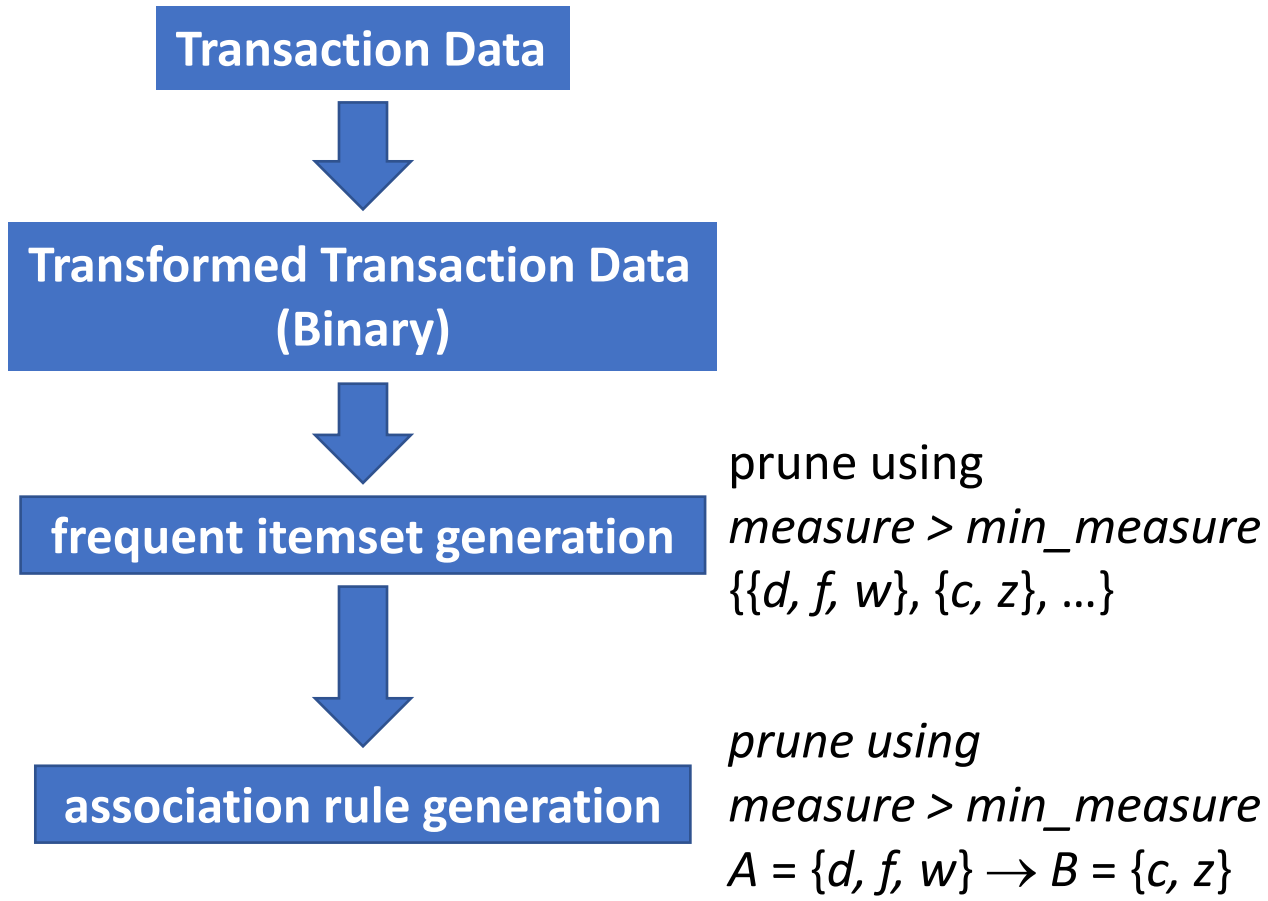
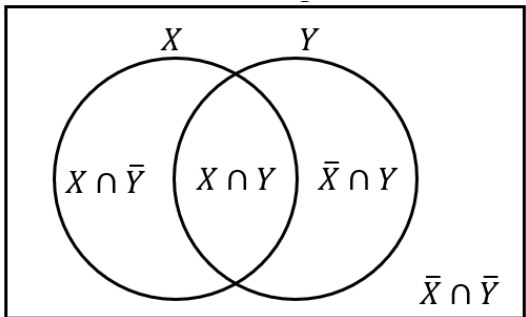
Medical records from various patients should be stratified according to confounding factors such as age and gender.

# A Review of Association Analysis

support count of  $A$  and  $B$  =  
count of transactions that contain both  $A$  and  $B$

- $f_{11}$  = support count of  $A$  and  $B$
- $f_{10}$  = support count of  $A$  and  $\bar{B}$
- $f_{01}$  = support count of  $\bar{A}$  and  $B$
- $f_{00}$  = support count of  $\bar{A}$  and  $\bar{B}$
- $f_{+1}$  = support count of  $B$
- $f_{+0}$  = support count of  $\bar{B}$
- $f_{1+}$  = support count of  $A$
- $f_{0+}$  = support count of  $\bar{A}$
- $N$  = total transaction count =  $f_{+1} + f_{+0} = f_{1+} + f_{0+}$

	$B$	$\bar{B}$	
$A$	$ A \cap B  = f_{11}$	$ A \cap \bar{B}  = f_{10}$	$ A  = f_{1+}$
$\bar{A}$	$ \bar{A} \cap B  = f_{01}$	$ \bar{A} \cap \bar{B}  = f_{00}$	$ \bar{A}  = f_{0+}$
	$ B  = f_{+1}$	$ \bar{B}  = f_{+0}$	$N$



An example of a measure:  
IS Measure:  $IS(A \rightarrow B) = IS(A, B)$

$$IS = \frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}$$



# Research Issues in Mining Association Patterns

# Introduction to Data Mining

2<sup>nd</sup> Edition

Copyright 2019

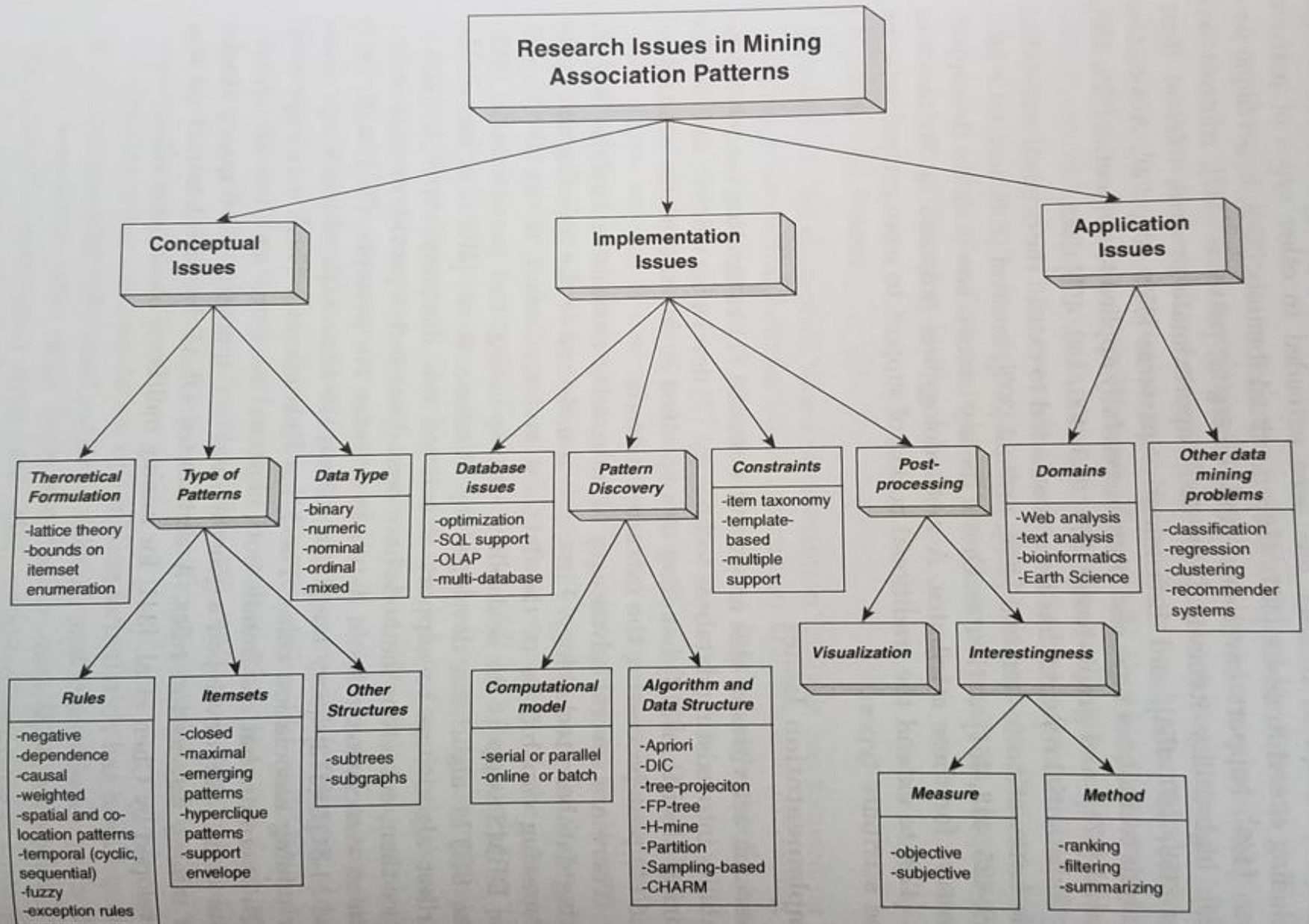


Figure 5.31. An overview of the various research directions in association analysis.

# Introduction to Data Mining - Chapter 6

## Association Analysis: Advanced Concepts

- 6.1 Handling Categorical Attributes
- 6.2 Handling Continuous Attributes
- 6.3 Handling a Concept Hierarchy
- 6.4 Sequential Patterns
- 6.5 Subgraph Patterns
- 6.6 Infrequent Patterns

### **The advanced concepts:**

- Extends to data sets with symmetric binary, categorical and continuous attributes.
- Extends to more complex entities such as sequences and graphs.



# Appendix