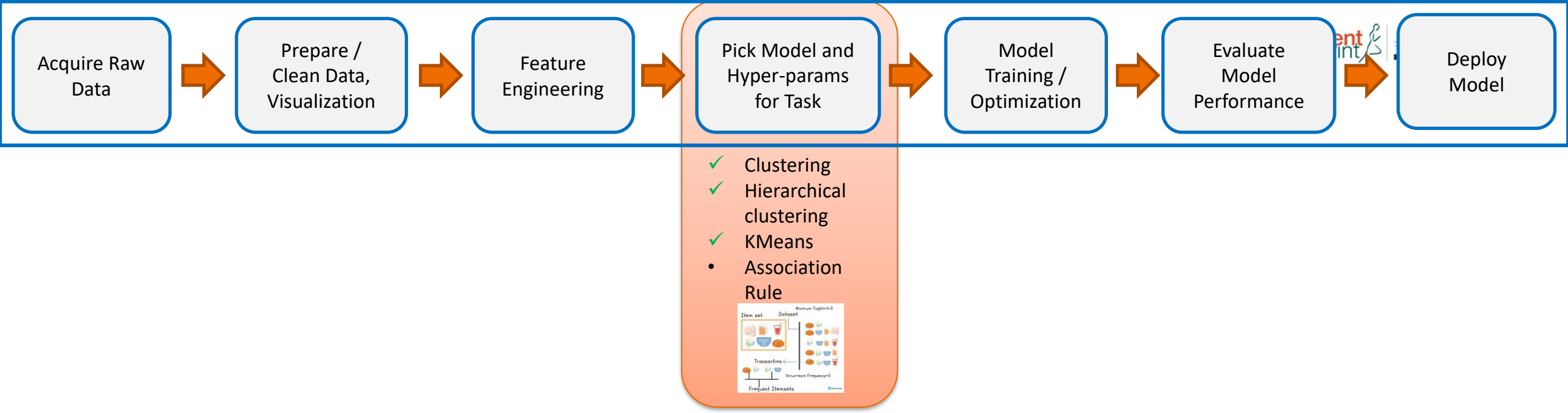


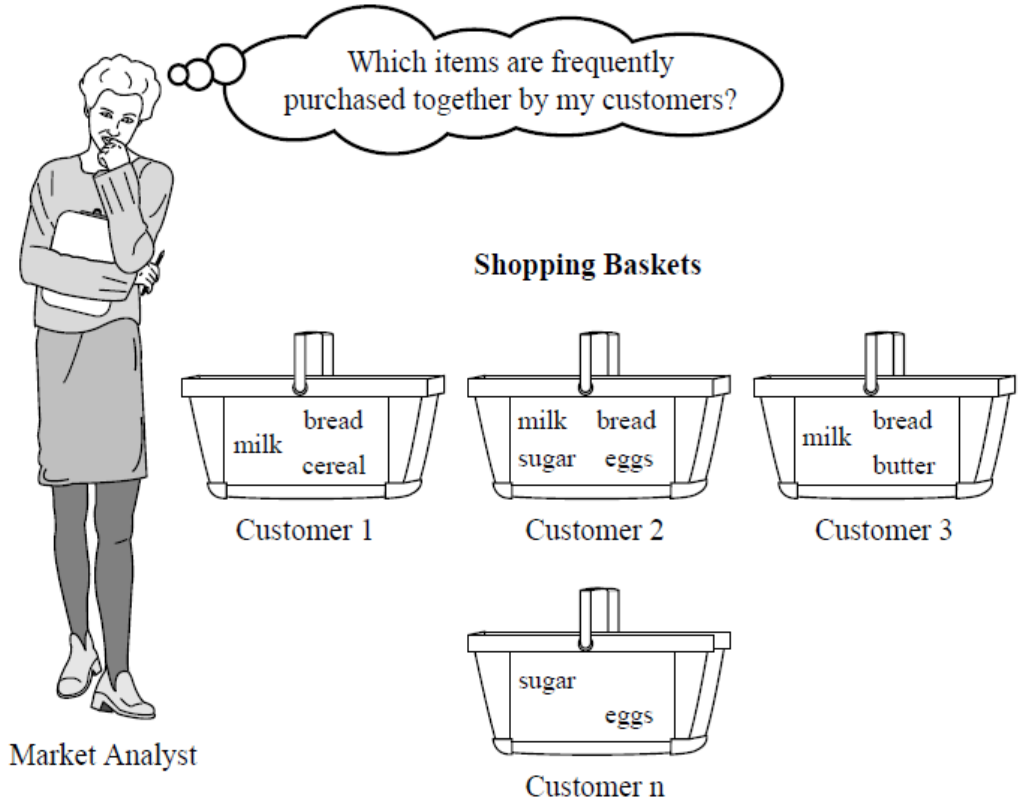
Association Rules

Identifying Co-occurring Patterns

What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— bread and butters?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Market Basket Analysis

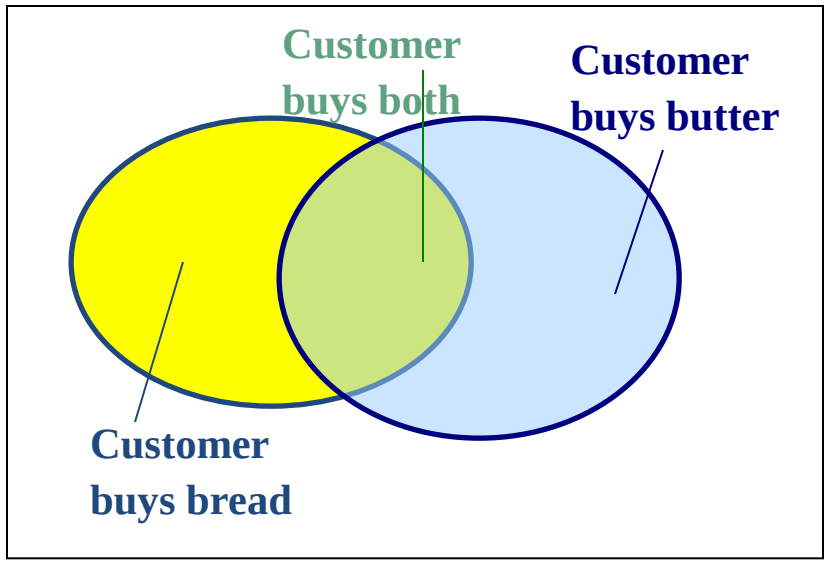


computer ⇒ *antivirus software* [*support* = 2%, *confidence* = 60%]

Source: Han & Kamber (2006)

Basic Concepts: Frequent Patterns

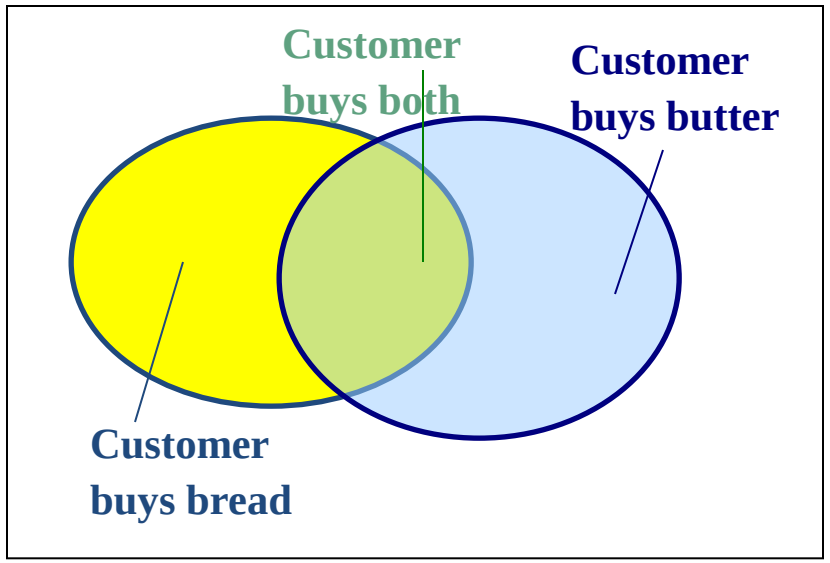
Tid	Items bought
10	bread, Nuts, butter
20	bread, Coffee, butter
30	bread, butter, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, butter, Eggs, Milk



- itemset: A set of one or more items
- k-itemset X = {x1, …, xk}
- (absolute) support, or, support count of X: Frequency or occurrence of an itemset X
- (relative) support, s, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is frequent if X's support is no less than a minsup threshold

Basic Concepts: Association Rules

Tid	Items bought
10	bread, Nuts, butter
20	bread, Coffee, butter
30	bread, butter, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, butter, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, s, probability that a transaction contains $X \cup Y$
  - confidence, c, conditional probability that a transaction having X also contains Y. $P(Y|X)$

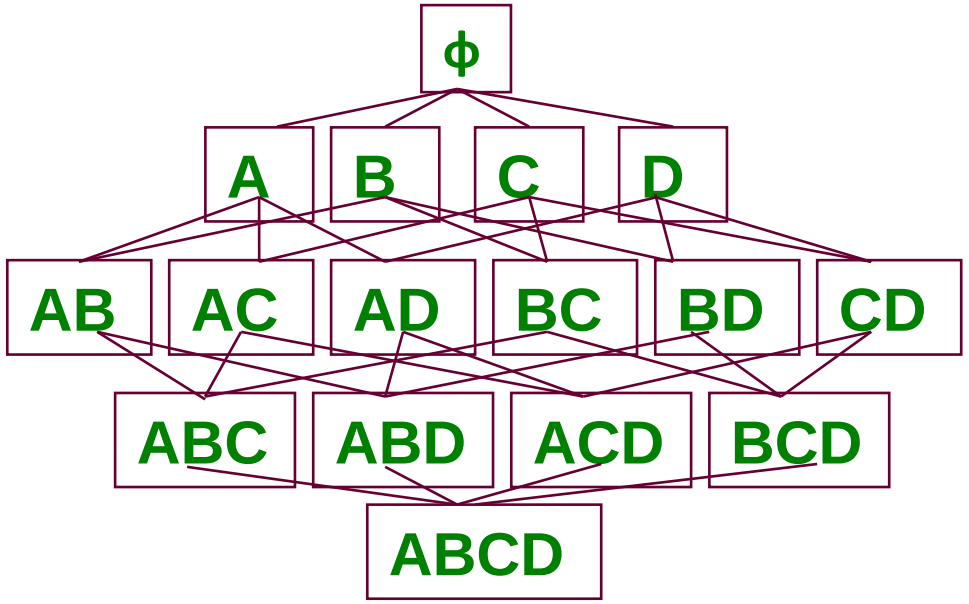
Let minsup = 50%, minconf = 50%

Freq. Pat.: bread: 3, Nuts: 3, butter: 4, Eggs: 3, {bread, butter}: 3

- **Association rules:**
  - bread → butter (60%, 100%)
  - butter → bread (60%, 75%)

Difficulty

- Extremely computationally expensive
- Naïve solution
  - exponential time and memory w.r.t. |I|
  - linear time w.r.t. |D|
- Typically, |I| is in thousands, |D| is in billions…



The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

$Sup_{min} = 2$

1st scan →

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→ L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan ←

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

← L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan → L_3

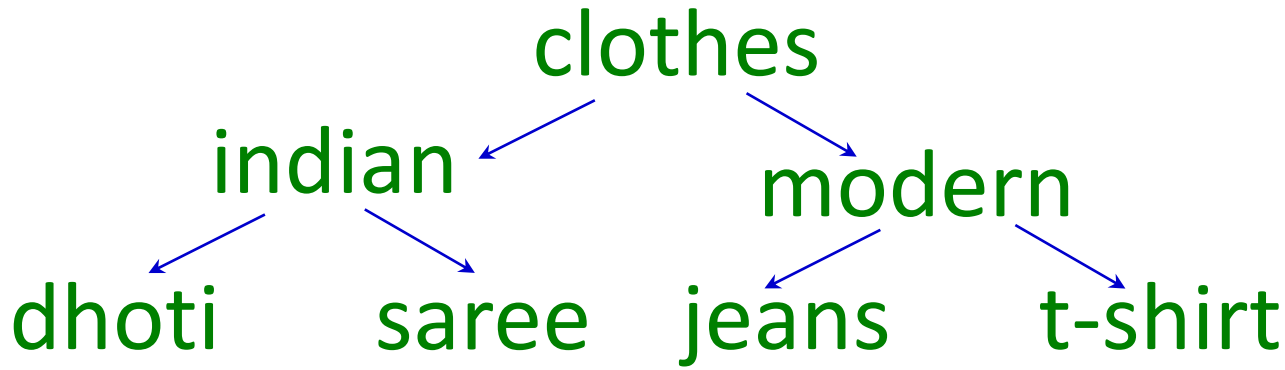
Itemset	sup
{B, C, E}	2

Apriori: A Candidate Generation & Test Approach

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested! (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length candidate (k+1)-itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent candidates can be generated, else iterate

Types of Association Rules

- Boolean association rules
- Hierarchical rules
  - dhoti, saree → t-shirt

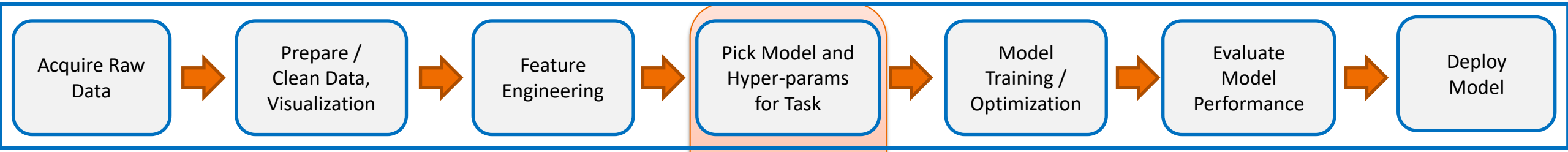


- Quantitative & Categorical rules
  - (Age: 30…39), (Married: Yes) → (NumCars: 2)

More Types

- Cyclic / Periodic rules
  - Sunday → vegetables
  - Christmas → gift items
  - Summer, rich, jobless → ticket to Hawaii
- Constrained rules
  - Show itemsets whose average price > Rs.10,000
  - Show itemsets that have television on RHS
- Sequential rules
  - Star wars, Empire Strikes Back → Return of the Jedi

Summary



- ✓ Association Rule

Thanks!!

Questions?