

# Learning Outcome

---

**Goal:** Provide an overview of data mining.

- Business Intelligence Basics
- Data Mining Basics
  - Define data mining
  - Data mining vs. databases
  - Basic data mining tasks
  - Data mining development
  - Data mining issues

# Business Intelligence

---

- In today's business environment, three things are certain
  - Competition is more intense than ever
  - The quantity of information is increasing proportionally
  - Markets and products evolve faster than ever
- Businesses need
  - To have the appropriate information
  - To make informed decisions
  - To take timely action

# Business Intelligence: **Process**

---

- Business Plan
- Architecture
- Project Planning
- Data Acquisition
- Implementation of the business intelligence solution
  - Data Warehouse
  - Intelligence tools e.g., data mining
- Evaluate the use of business intelligence

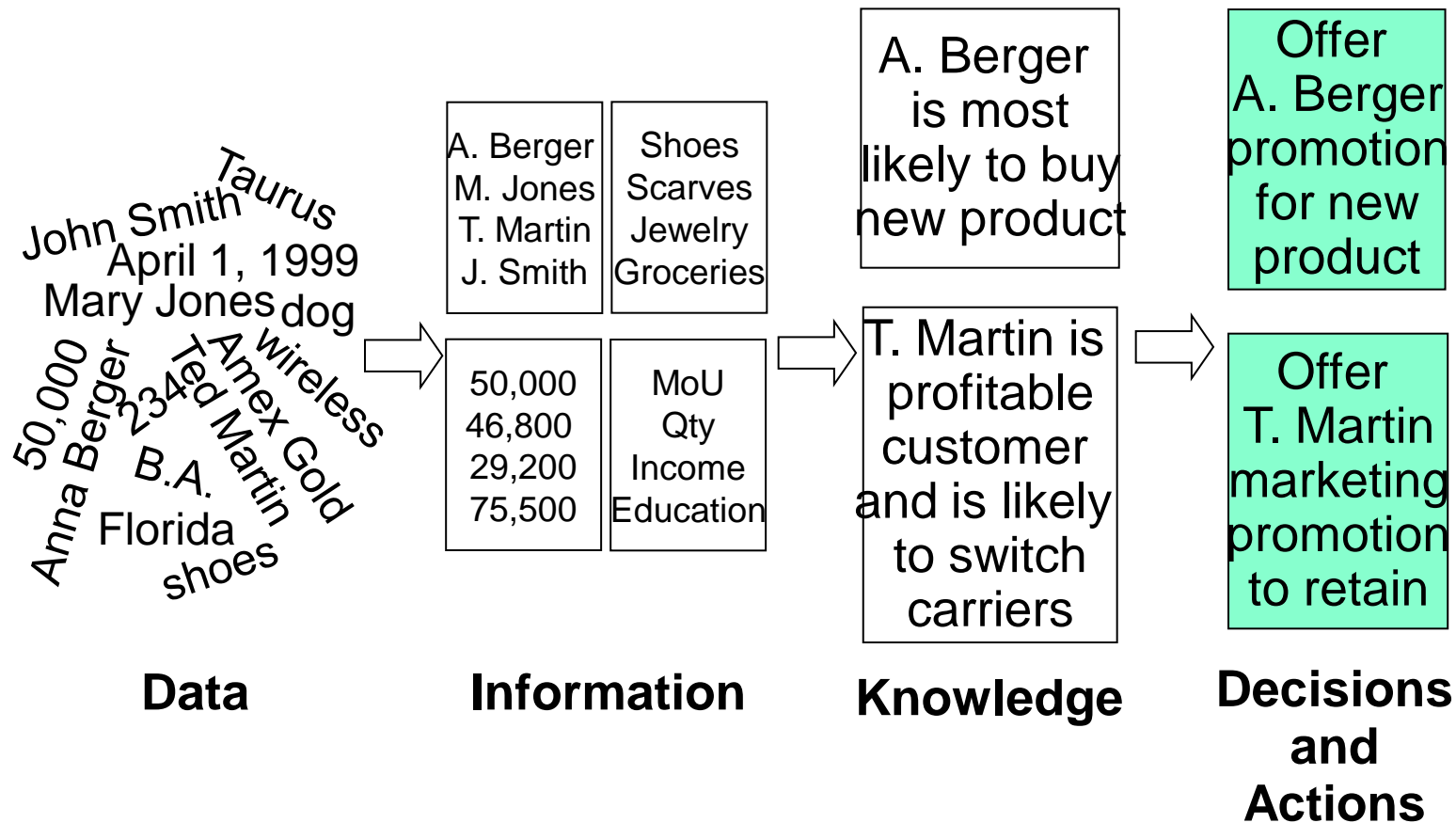
# Business Intelligence

---

- According to the International Data Corporation (IDC), in 65 organisations, the average return on business intelligence investments was over 400% over a period of 2.3 years.
- Implies services and products in the areas of
  - Data Mining
  - Data Warehousing

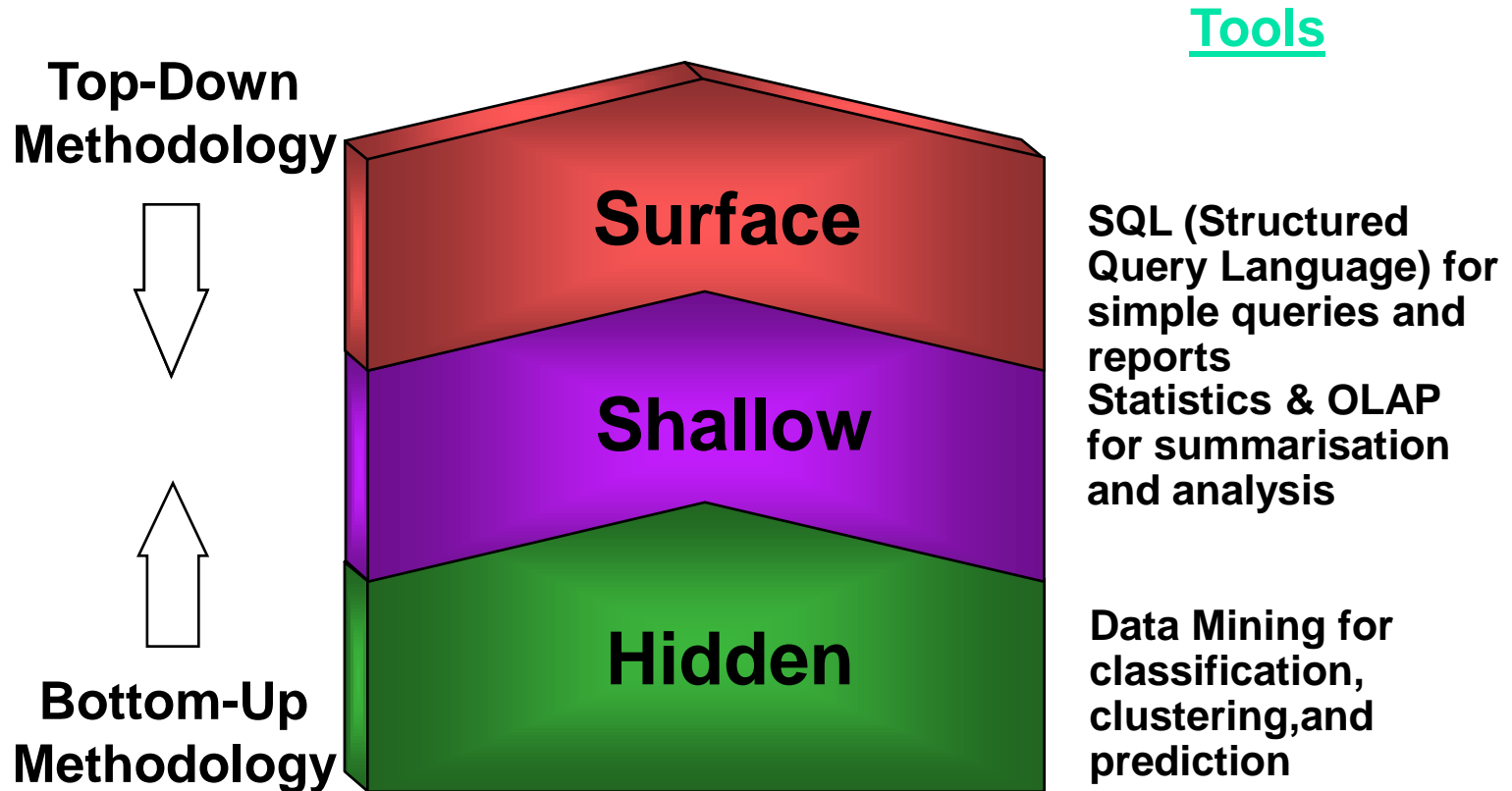
# Business Intelligence

## ***Data-Information-Knowledge-Decision***



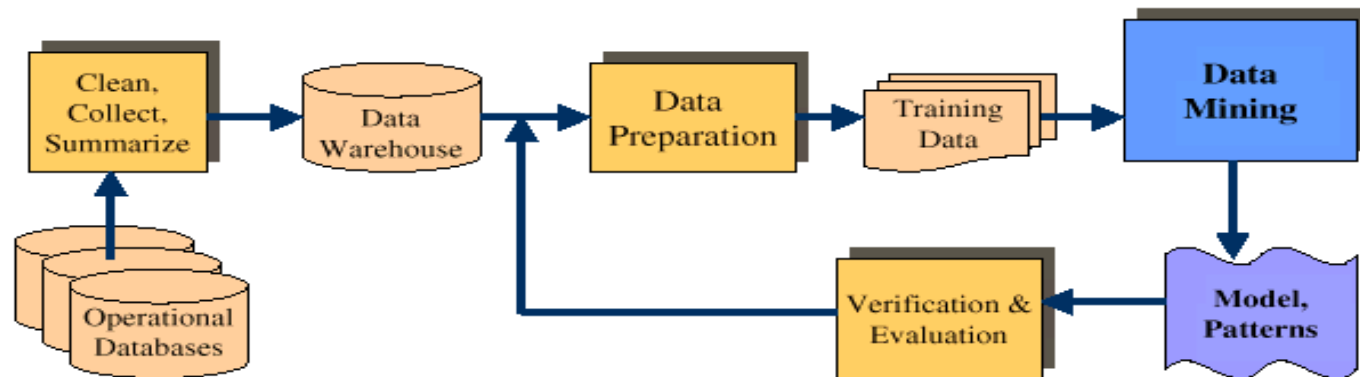
# Business Intelligence

## *Business Knowledge*



# What is Data Mining?

- Non-trivial extraction of **implicit, previously unknown** and **potentially useful** information from data.
- Knowledge Discovery in Databases (KDD) Or Data Mining is a non-trivial process of identifying **valid, novel, potentially useful**, and **ultimately understandable** patterns in large set of data



# How much Data We Have

---

- EMC-sponsored research from IDC that for the first time measures and forecasts the amounts and types of digital information created and copied in the world - and whether it is generated from individuals or businesses
- The primary drivers include rich media, user-generated content and 1.6 billion Internet users.
- The 2006 digital universe was 161 billion gigabytes (161 exabytes) in size.
- This digital universe equals approximately three million times the information in all the books ever written - or the equivalent of 12 stacks of books, each extending more than 93 million miles from the earth to the sun.



# How much we will have by the year 2010

---

- **DIGITAL CAMERAS:** The number of images captured on exceeded 150 billion worldwide. The number of images captured on cell phones hit almost 100 billion. IDC is forecasting the capture of more than 500 billion images by 2010.
- **EMAIL:** The number of email mailboxes has grown from 253 million in 1998 to nearly 1.6 billion in 2006. i.e., excluding spam - accounted for 6 exabytes.
- **INTERNET:** In 1996 there were only 48 million people using the Internet. The Worldwide Web was just two years old. By 2006, there were 1.1 billion users on the Internet. By 2010, IDC expects another 500 million users to come online.
- A groundbreaking new study forecasts that a staggering 988 billion Gigabytes of digital information will have been created by 2010.

# What is (not) Data Mining?

---

- **What is not Data Mining?**

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

- **What is Data Mining?**

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

# Database Processing vs. Data Mining Processing

---

- Query
    - Well defined
    - SQL
  - Data
    - Operational data
  - Output
    - Precise
    - Subset of database
- Query
    - Poorly defined
    - No precise query language
  - Data
    - Not operational data
  - Output
    - Fuzzy
    - Not a subset of database

# Query Examples

---

## ■ Database

- Find all credit applicants with last name of Smith.
- Identify customers who have purchased more than £10,000 in the last month.
- Find all customers who have purchased milk

## ■ Data Mining

- Find all credit applicants who are poor credit risks. (classification)
- Identify customers with similar buying habits. (Clustering)
- Find all items which are frequently purchased with milk. (association rules)

# Supervised Learning /Unsupervised Learning

---

- Supervised Learning
  - Build a learner model using data instances of known origin.
  - Use the model to determine the outcome new instances of unknown origin
- Unsupervised Learning
  - A data mining method that builds models from data without predefined classes.

# Basic Data Mining Tasks

---

## ■ **Classification**

- maps data into predefined groups or classes
- Supervised learning

## ■ **Clustering**

- groups similar data together into clusters.
- Unsupervised learning

## ■ **Association Rules**

- uncovers relationships among data.
- Unsupervised learning

# Other Applications

---

- IBM Advanced Scout analyzes NBA game statistics
  - A large number of data generated for each game, player, teams, etc. (see <http://www.nba.com>)
  - Points per game, rebounds per game, field goal percentage, 3 point field goal percentage, free throw percentage, assists per game, assists per turnover, steals per game, blocks per game, turnovers per game, fouls per game, ...
- Used by NBA coaching staffs to discover interesting patterns in basketball game data.



# Classification: Definition

---

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Task: Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

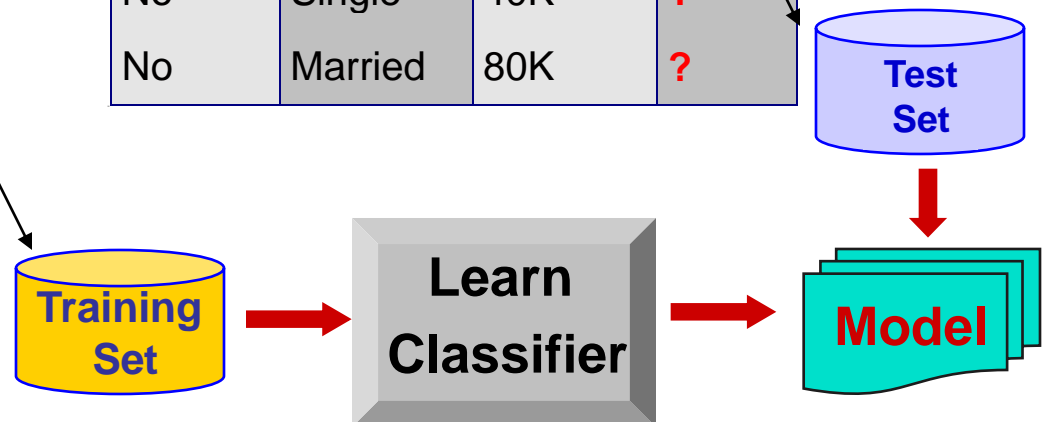


# Classification Example

categorical  
categorical  
continuous  
class

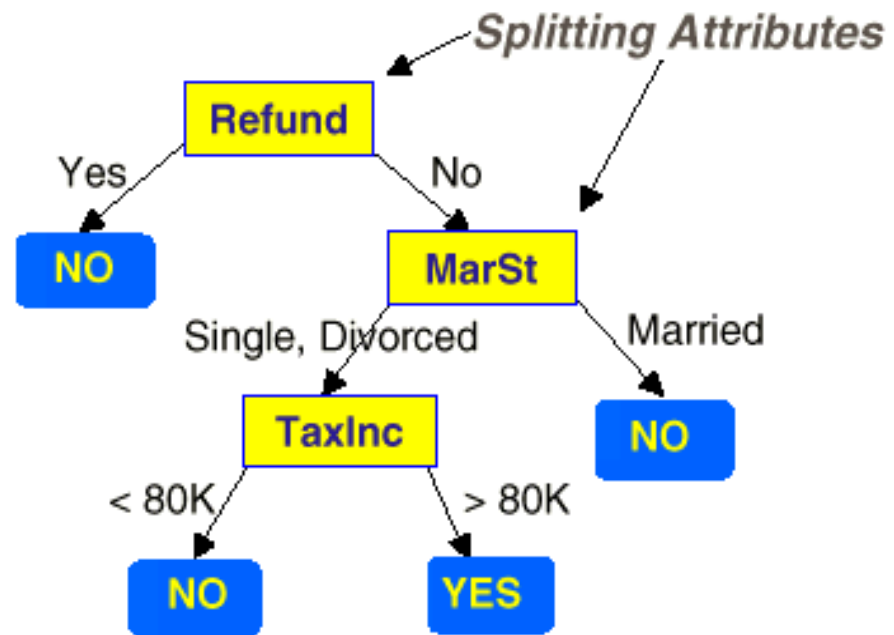
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



# Example of Decision Trees

<i>Tid</i>	<i>Refund</i>	<i>Marital Status</i>	<i>Taxable Income</i>	<i>Cheat</i>
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



The splitting attribute at a node is determined based on the Gini index.

# Classification: Application 1

---

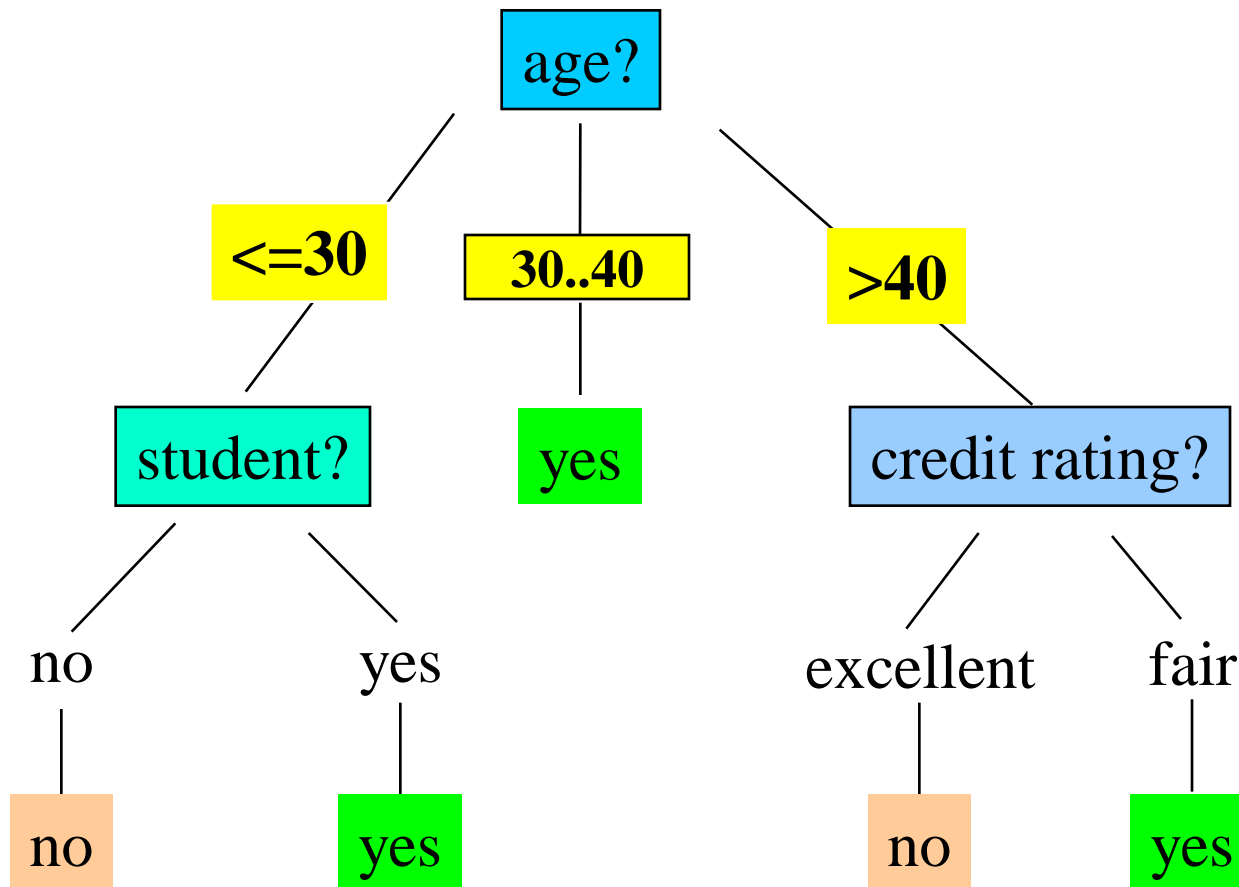
- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

# Training Dataset

This follows an example from Quinlan's ID3

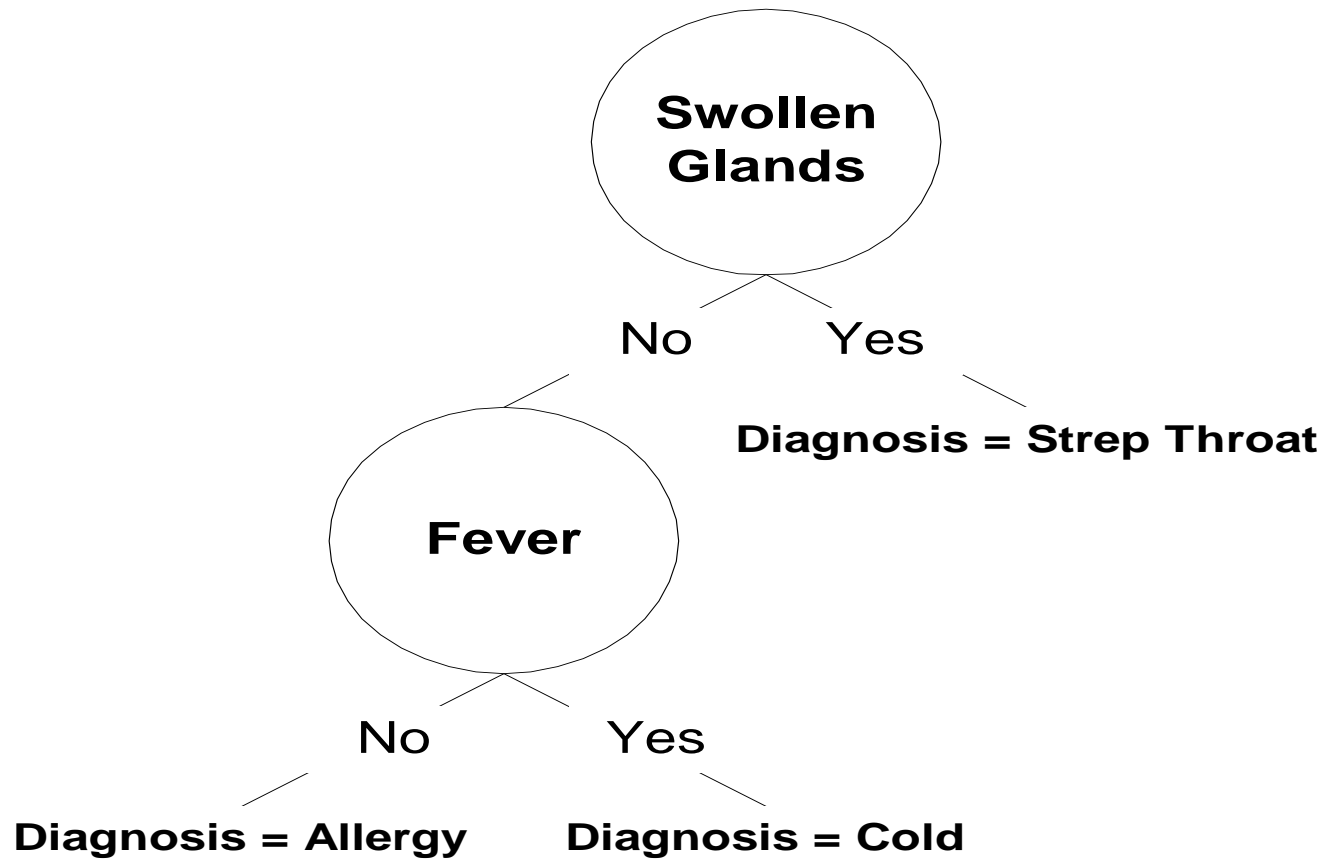
age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

# Output: A Decision Tree for “*buys\_computer*”



**Table 1.1 • Hypothetical Training Data for Disease Diagnosis**

<b>Patient ID#</b>	<b>Sore Throat</b>	<b>Fever</b>	<b>Swollen Glands</b>	<b>Congestion</b>	<b>Headache</b>	<b>Diagnosis</b>
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold



**Table 1.2 • Data Instances with an Unknown Classification**

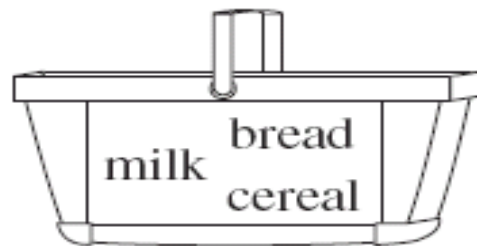
<b>Patient ID#</b>	<b>Sore Throat</b>	<b>Fever</b>	<b>Swollen Glands</b>	<b>Congestion</b>	<b>Headache</b>	<b>Diagnosis</b>
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?



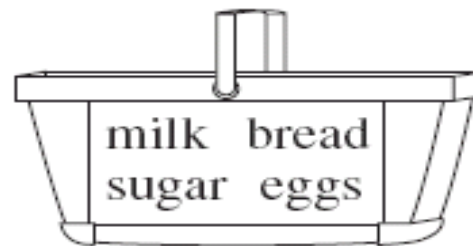
# Market Basket Analysis (cont...)

Which items are frequently purchased together by my customers?

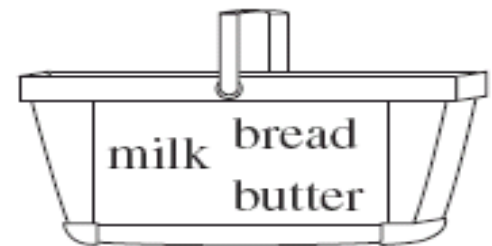
## Shopping Baskets



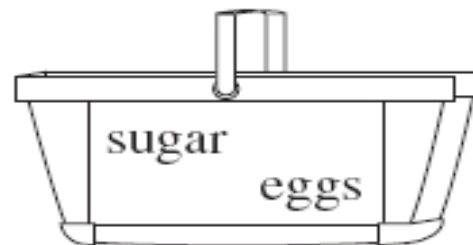
Customer 1



Customer 2



Customer 3



Customer n

Market Analyst

# What Is Association Mining?

---

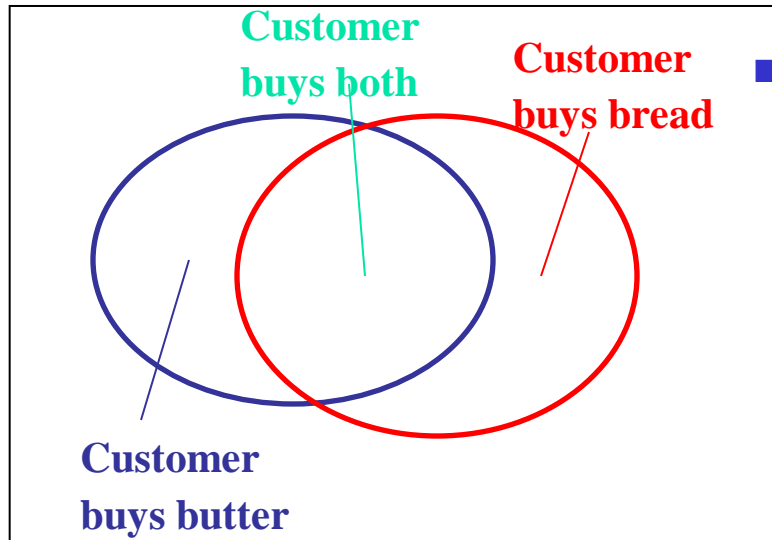
- Association rule mining:
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in large databases.
- Applications:
  - Basket data analysis, cross-marketing, catalog design.

# Association Rule: Basic Concepts

---

- Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)
- Find: all rules that correlate the presence of one set of items with that of another set of items
  - E.g., *98% of people who purchase tires and auto accessories also get automotive services done*
- Applications
  - $* \Rightarrow$  *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales)
  - *Home Electronics*  $\Rightarrow$   $*$  (What other products should the store stocks up?)
  - *Attached mailing* in direct marketing

# Rule Measures: Support and Confidence



Find all the rules  $X \& Y \Rightarrow Z$  with minimum confidence and support

- **support**,  $s$ , **probability** that a transaction contains  $\{X \sqcup Y \sqcup Z\}$
- **confidence**,  $c$ , **conditional probability** that a transaction having  $\{X \sqcup Y\}$  also contains  $Z$

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

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

- $A \Rightarrow C$  (50%, 66.6%)
- $C \Rightarrow A$  (50%, 100%)

# Association Rules Example

---

Transaction	Items
$t_1$	Bread, Jelly, Butter
$t_2$	Bread, Butter
$t_3$	Bread, Milk, Butter
$t_4$	Juice, Bread
$t_5$	Juice, Milk

$I = \{ \text{Juice, Bread, Jelly, Milk, Butter} \}$

Support of {Bread, Butter} is 60%

# Association Rules Ex (cont'd)

---

<b>X → Y</b>	<b>Support</b>	<b>Conf.</b>
Bread → Butter	60%	75%
Butter → Bread	60%	100%
Juice → Bread	20%	50%
Butter → Jelly	20%	33.3%
Jelly → Butter	20%	100%
Jelly → Milk	0%	0%

# Association Rules Ex (cont'd)

---

- Calculate the support and confidence for the following two association rules

$A \rightarrow C$

$C \rightarrow A$

using the market basket data set given below.

<u>TransactionID,</u>	<u>Items Bought</u>
2000	A, B, C
1000	A, C
4000	A, D
5000	B, E, F

# Association Rules Ex (cont'd)

---

Calculate the support and confidence for the following two association rules

Bread  $\rightarrow$  Cheese

Cheese  $\rightarrow$  Bread

using the market basket data set given below.

TransactionID,	Items Bought
2000	Bread , Coffee, Cheese
1000	Bread , Cheese
4000	Bread , Sugar
5000	Coffee, Egg, Tomato



# Clustering Definition

---

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

# Clustering: Application 1

---

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

# Clustering: Application 2

---

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

# Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

<b><i>Category</i></b>	<b><i>Total Articles</i></b>	<b><i>Correctly Placed</i></b>
<b><i>Financial</i></b>	555	364
<b><i>Foreign</i></b>	341	260
<b><i>National</i></b>	273	36
<b><i>Metro</i></b>	943	746
<b><i>Sports</i></b>	738	573
<b><i>Entertainment</i></b>	354	278