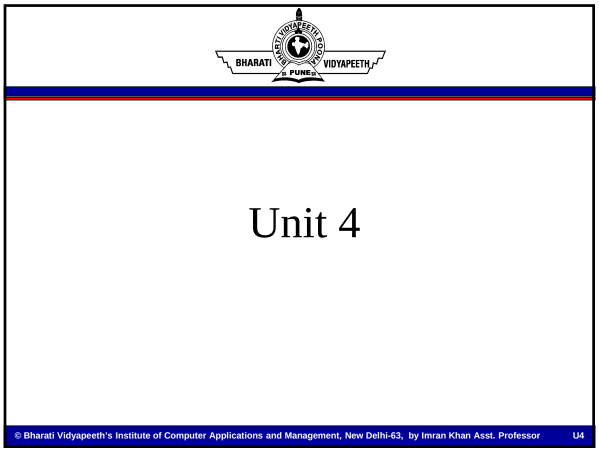
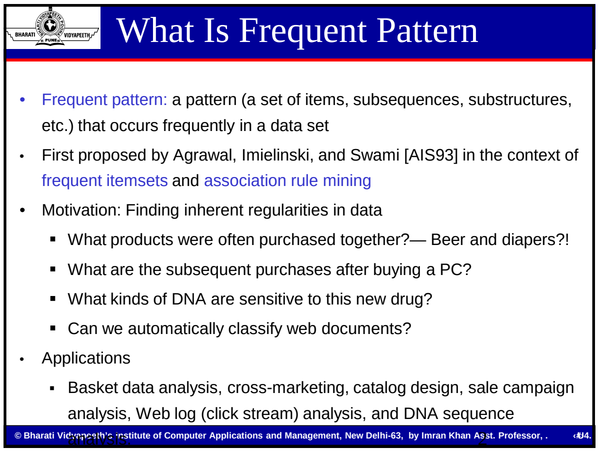


Unit 4

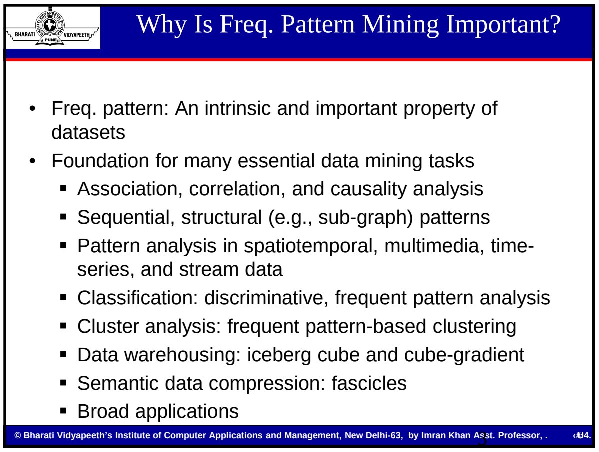
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor U4



What Is Frequent Pattern

- 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?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

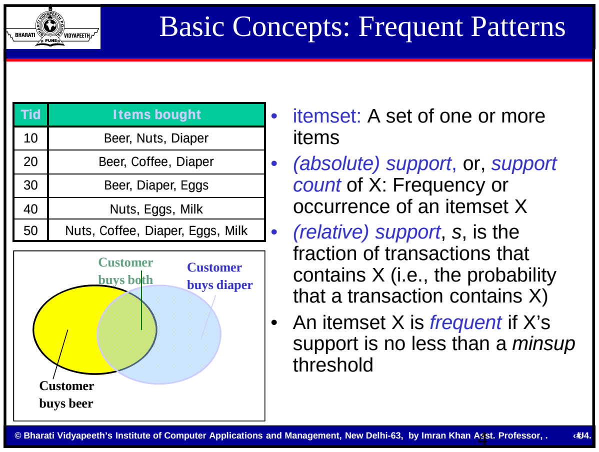
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . U4.



Why Is Freq. Pattern Mining Important?

- Freq. pattern: An intrinsic and important property of datasets
- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: discriminative, frequent pattern analysis
  - Cluster analysis: frequent pattern-based clustering
  - Data warehousing: iceberg cube and cube-gradient
  - Semantic data compression: fascicles
  - Broad applications

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . U4.

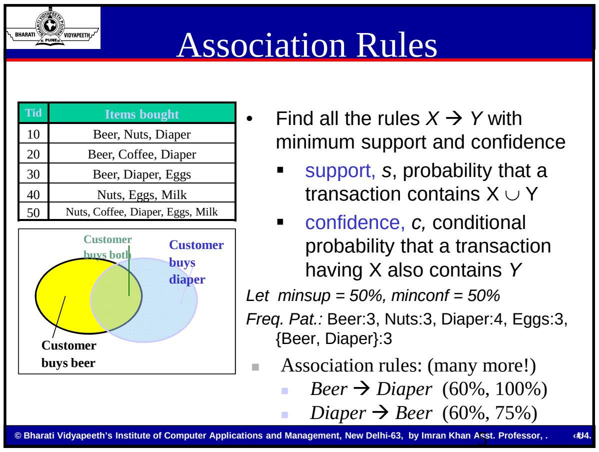


Basic Concepts: Frequent Patterns

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Customer buys both
Customer buys diaper
Customer buys beer

- itemset: A set of one or more items
- *(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



Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

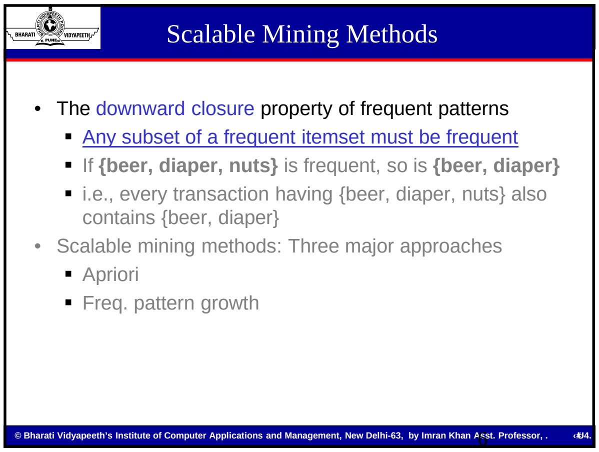
Customer buys both
Customer buys diaper
Customer buys beer

- Find all the rules *X* → *Y* with minimum support and confidence
  - support, *s*, probability that a transaction contains X ∪ Y
  - confidence, *c,* conditional probability that a transaction having X also contains *Y*

Let minsup = 50%, minconf = 50%

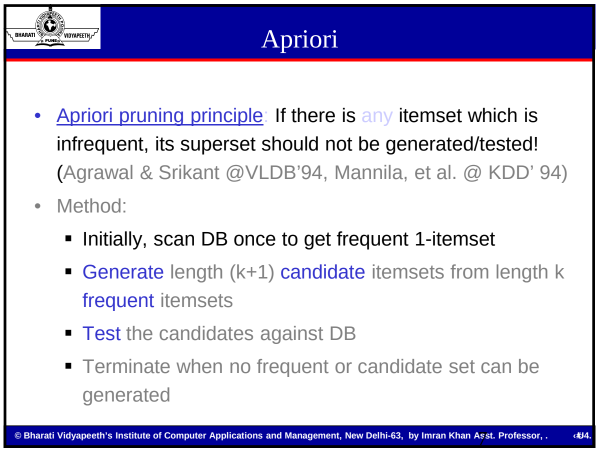
Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)



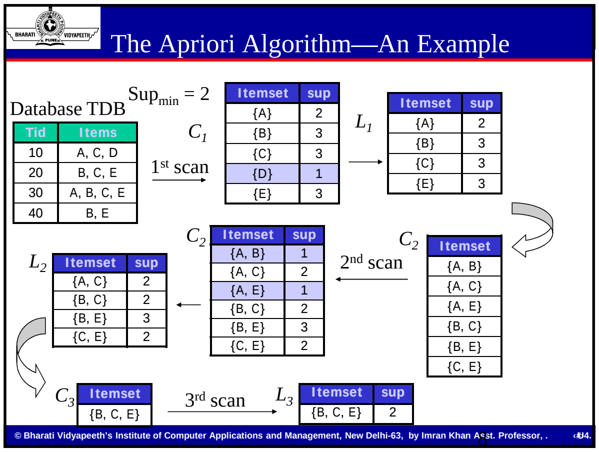
Scalable Mining Methods

- The downward closure property of frequent patterns
  - Any subset of a frequent itemset must be frequent
  - If **{beer, diaper, nuts}** is frequent, so is **{beer, diaper}**
  - i.e., every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Scalable mining methods: Three major approaches
  - Apriori
  - Freq. pattern growth



Apriori

- 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 (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent or candidate set can be generated



The Apriori Algorithm—An Example

$Sup_{min} = 2$

Database TDB

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

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

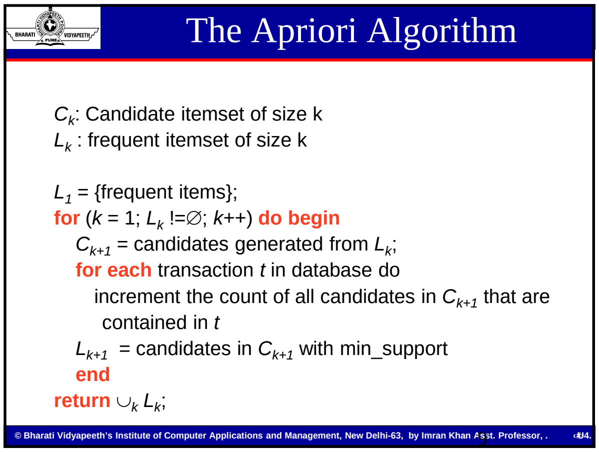
C_3

Itemset
{B, C, E}

3rd scan

L_3

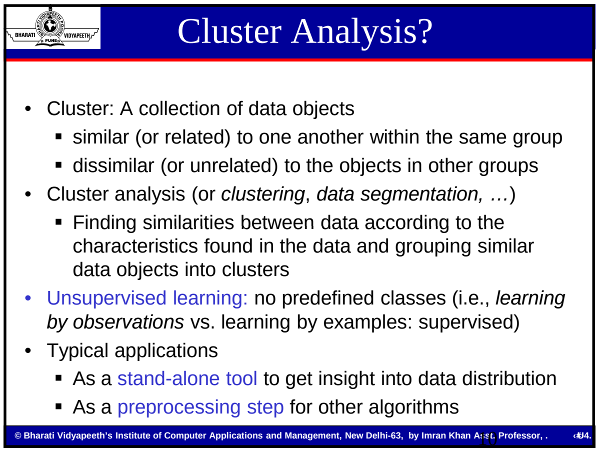
Itemset	sup
{B, C, E}	2



The Apriori Algorithm

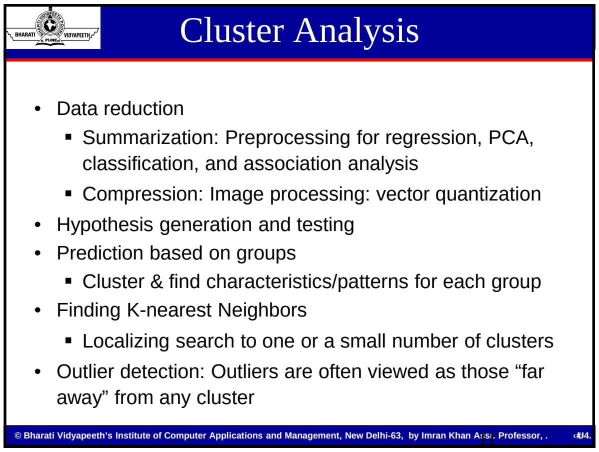
C_k: Candidate itemset of size k
L_k : frequent itemset of size k

```
$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do
        increment the count of all candidates in $C_{k+1}$ that are
          contained in $t$
    $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
    end
return $\cup_k L_k$;
```



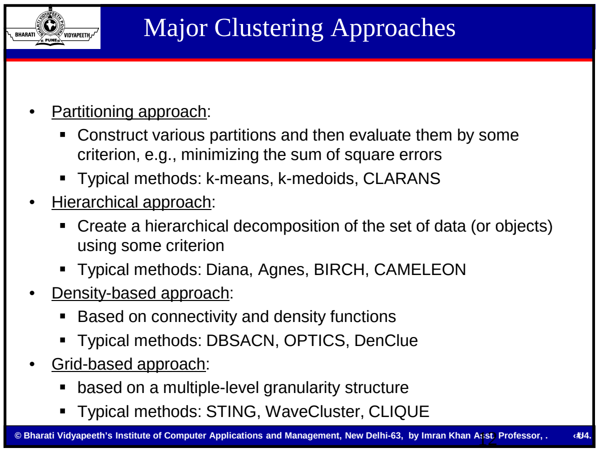
Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



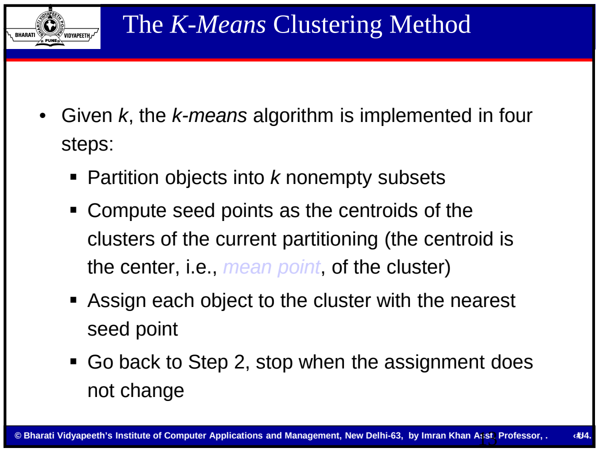
Cluster Analysis

- Data reduction
  - Summarization: Preprocessing for regression, PCA, classification, and association analysis
  - Compression: Image processing: vector quantization
- Hypothesis generation and testing
- Prediction based on groups
  - Cluster & find characteristics/patterns for each group
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection: Outliers are often viewed as those "far away" from any cluster



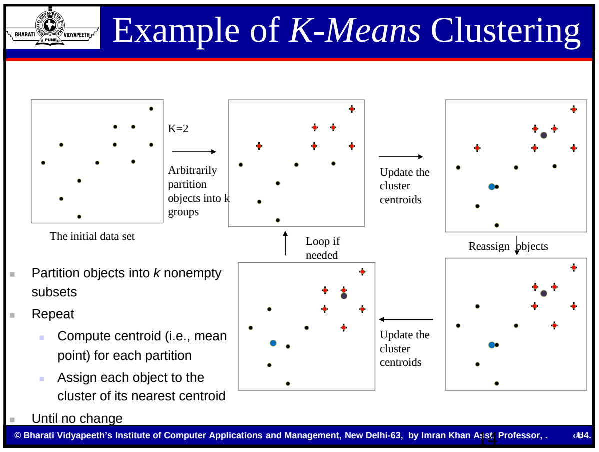
Major Clustering Approaches

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue
- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE



The *K-Means* Clustering Method

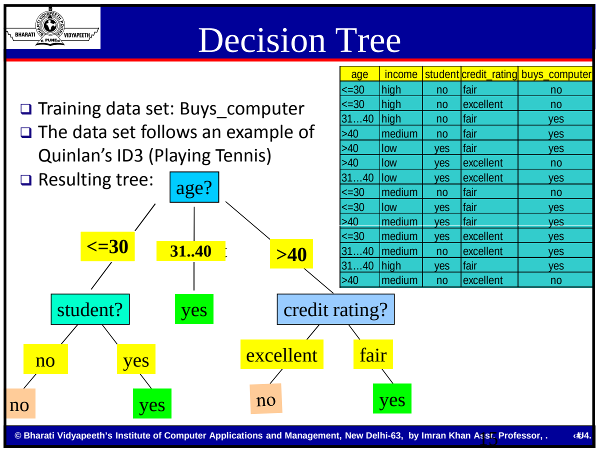
- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when the assignment does not change



Example of *K-Means* Clustering

K=2
Arbitrarily partition objects into k groups
The initial data set
Update the cluster centroids
Reassign objects
Loop if needed
Update the cluster centroids

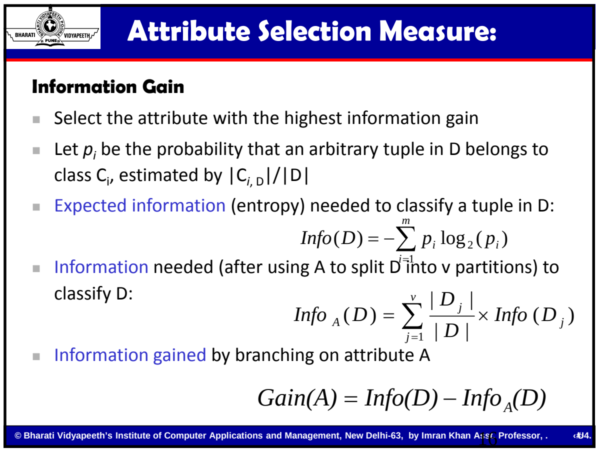
- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change



Decision Tree

- ❑ Training data set: Buys_computer
- ❑ The data set follows an example of Quinlan's ID3 (Playing Tennis)
- ❑ Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…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



Attribute Selection Measure:

Information Gain

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

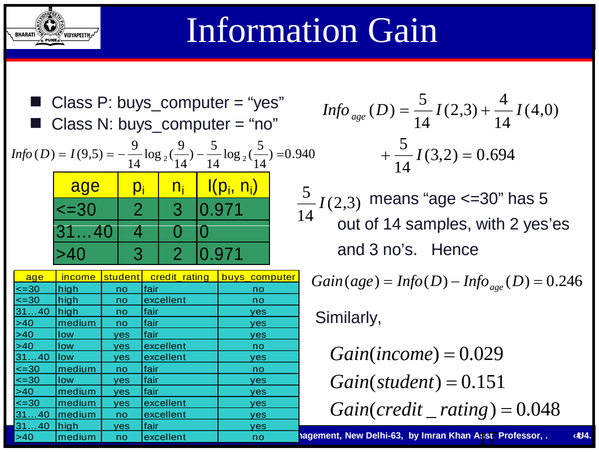
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$



Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…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

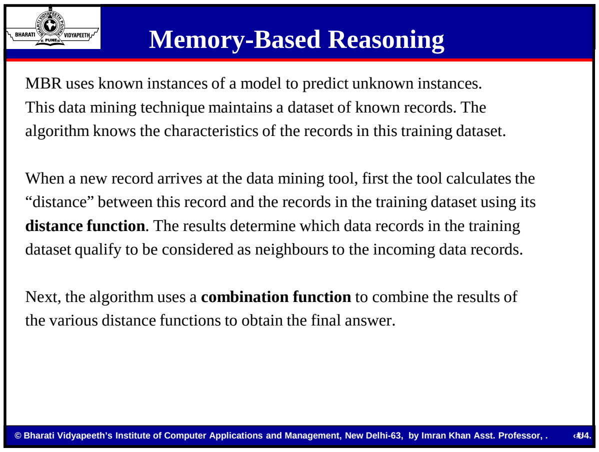
$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$



Memory-Based Reasoning

MBR uses known instances of a model to predict unknown instances. This data mining technique maintains a dataset of known records. The algorithm knows the characteristics of the records in this training dataset.

When a new record arrives at the data mining tool, first the tool calculates the "distance" between this record and the records in the training dataset using its **distance function**. The results determine which data records in the training dataset qualify to be considered as neighbours to the incoming data records.

Next, the algorithm uses a **combination function** to combine the results of the various distance functions to obtain the final answer.

For solving a data mining problem using MBR, we are concerned with three critical issues:

- Selecting the most suitable historical records to form the training dataset.
- Establishing the best way to compose the historical record.
- Determining the two essential functions, namely, the distance function and the combination function.

Link Analysis

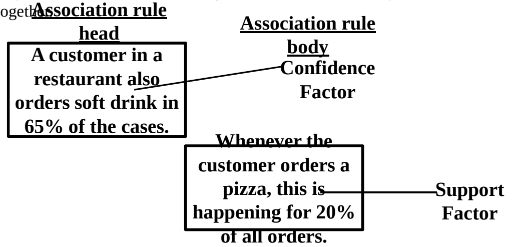
This algorithm is extremely useful for finding patterns from relationships. The link analysis technique mines relationships and discovers knowledge.

For eg. If the Fast Food Restaurant owner in the case study has to apply link analysis technique to mine data from the data warehouse, he might find out that in more than 80% of the cases, customers order a soft drink if they order a pizza. The restaurant owner will try to analyse the link between the two products and promote them together.

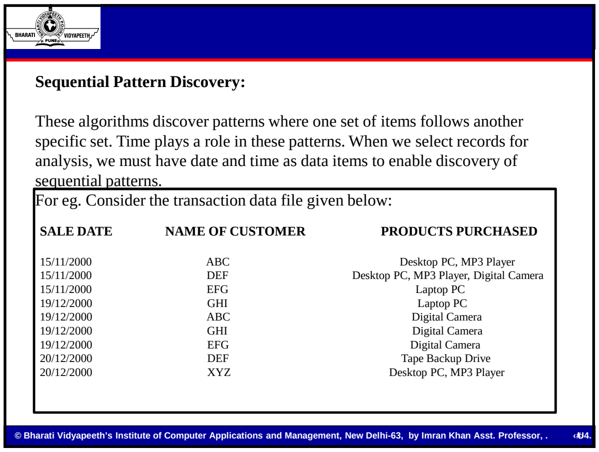
Depending upon the types of knowledge discovery, link analysis techniques have three types of applications: **associations discovery**, **sequential pattern discovery** and **similar time discovery**.

Associations Discovery:

These algorithms find combinations where the presence of one item suggests the presence of another.
When we apply these algorithms to the daily sales of the fast food restaurant, they will uncover affinities among menu items that are likely to be ordered together.



Association rule head
A customer in a restaurant also orders soft drink in 65% of the cases.
Association rule body
Confidence Factor
Whenever the customer orders a pizza, this is happening for 20% of all orders.
Support Factor



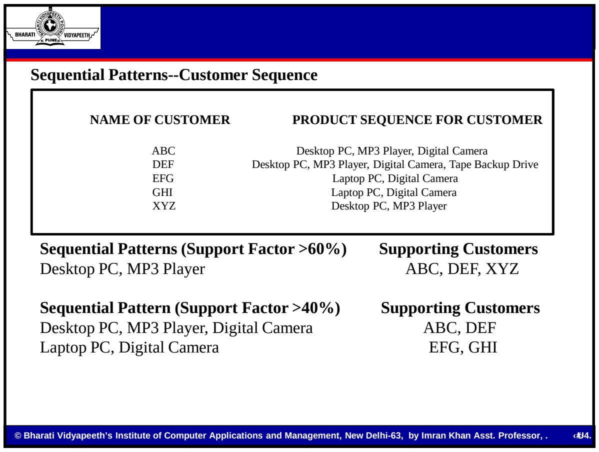
Sequential Pattern Discovery:

These algorithms discover patterns where one set of items follows another specific set. Time plays a role in these patterns. When we select records for analysis, we must have date and time as data items to enable discovery of sequential patterns.

For eg. Consider the transaction data file given below:

SALE DATE	NAME OF CUSTOMER	PRODUCTS PURCHASED
15/11/2000	ABC	Desktop PC, MP3 Player
15/11/2000	DEF	Desktop PC, MP3 Player, Digital Camera
15/11/2000	EFG	Laptop PC
19/12/2000	GHI	Laptop PC
19/12/2000	ABC	Digital Camera
19/12/2000	GHI	Digital Camera
19/12/2000	EFG	Digital Camera
20/12/2000	DEF	Tape Backup Drive
20/12/2000	XYZ	Desktop PC, MP3 Player

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



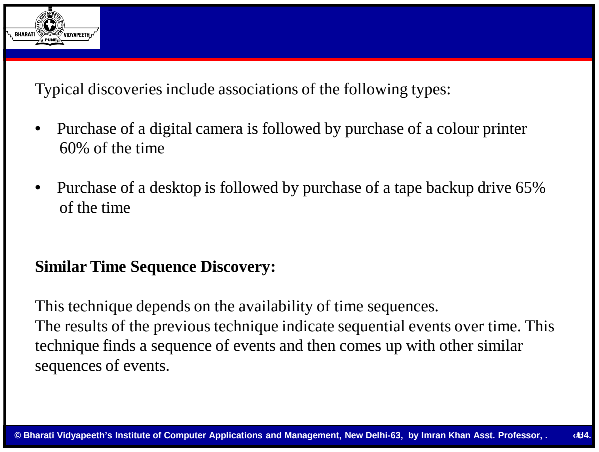
Sequential Patterns--Customer Sequence

NAME OF CUSTOMER	PRODUCT SEQUENCE FOR CUSTOMER
ABC	Desktop PC, MP3 Player, Digital Camera
DEF	Desktop PC, MP3 Player, Digital Camera, Tape Backup Drive
EFG	Laptop PC, Digital Camera
GHI	Laptop PC, Digital Camera
XYZ	Desktop PC, MP3 Player

Sequential Patterns (Support Factor >60%)	**Supporting Customers**
Desktop PC, MP3 Player	ABC, DEF, XYZ

Sequential Pattern (Support Factor >40%)	**Supporting Customers**
Desktop PC, MP3 Player, Digital Camera	ABC, DEF
Laptop PC, Digital Camera	EFG, GHI

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



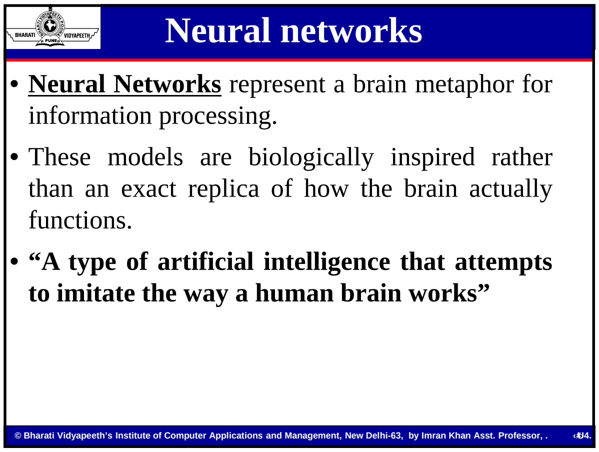
Typical discoveries include associations of the following types:

- Purchase of a digital camera is followed by purchase of a colour printer 60% of the time
- Purchase of a desktop is followed by purchase of a tape backup drive 65% of the time

Similar Time Sequence Discovery:

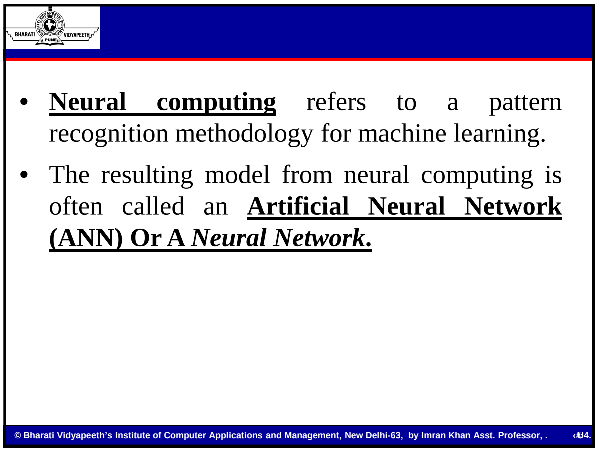
This technique depends on the availability of time sequences.
The results of the previous technique indicate sequential events over time. This technique finds a sequence of events and then comes up with other similar sequences of events.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.

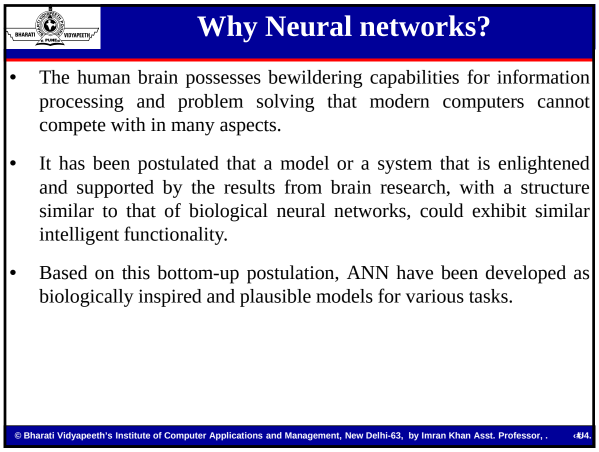


Neural networks

- **Neural Networks** represent a brain metaphor for information processing.
- These models are biologically inspired rather than an exact replica of how the brain actually functions.
- **"A type of artificial intelligence that attempts to imitate the way a human brain works"**

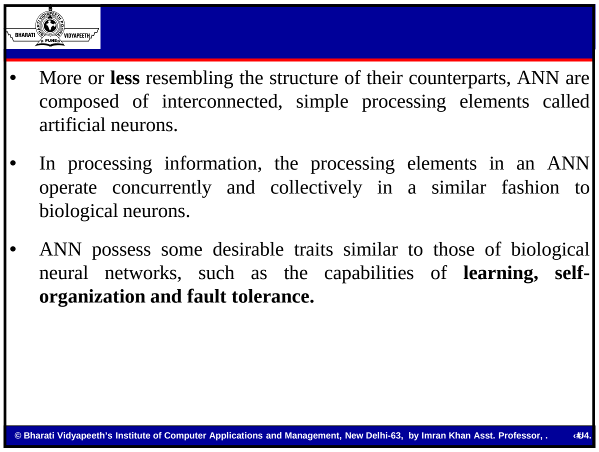


- **Neural computing** refers to a pattern recognition methodology for machine learning.
- The resulting model from neural computing is often called an **Artificial Neural Network (ANN) Or A *Neural Network*.**

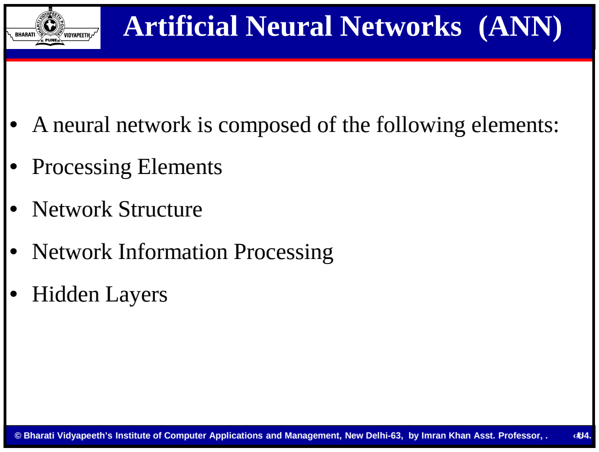


Why Neural networks?

- The human brain possesses bewildering capabilities for information processing and problem solving that modern computers cannot compete with in many aspects.
- It has been postulated that a model or a system that is enlightened and supported by the results from brain research, with a structure similar to that of biological neural networks, could exhibit similar intelligent functionality.
- Based on this bottom-up postulation, ANN have been developed as biologically inspired and plausible models for various tasks.

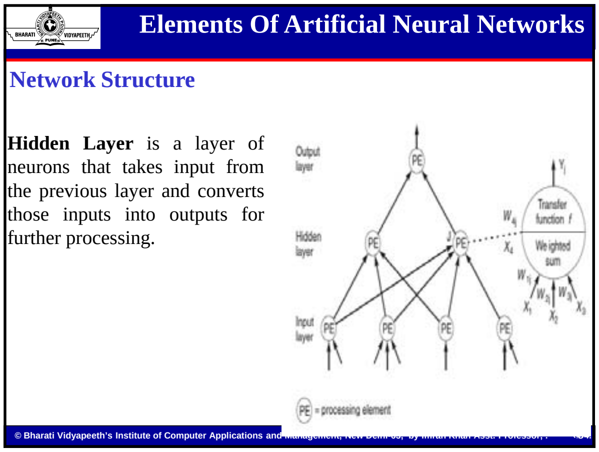


- More or **less** resembling the structure of their counterparts, ANN are composed of interconnected, simple processing elements called artificial neurons.
- In processing information, the processing elements in an ANN operate concurrently and collectively in a similar fashion to biological neurons.
- ANN possess some desirable traits similar to those of biological neural networks, such as the capabilities of **learning, self-organization and fault tolerance.**



Artificial Neural Networks (ANN)

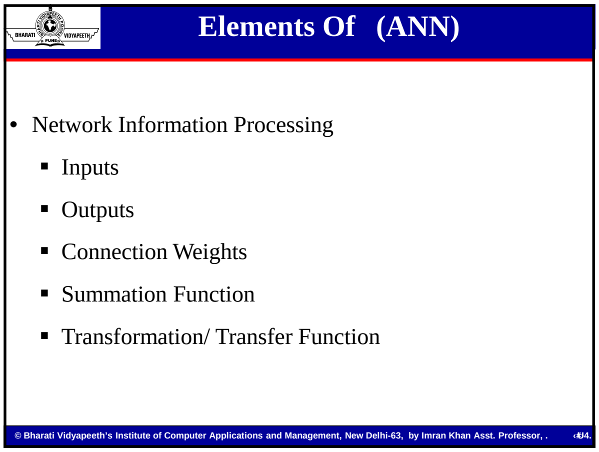
- A neural network is composed of the following elements:
- Processing Elements
- Network Structure
- Network Information Processing
- Hidden Layers



Elements Of Artificial Neural Networks

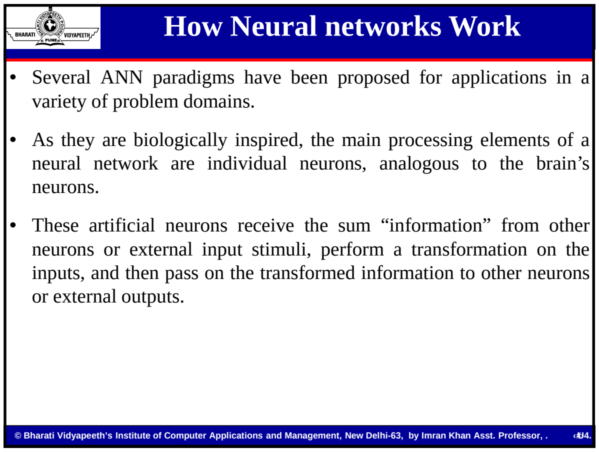
Network Structure

Hidden Layer is a layer of neurons that takes input from the previous layer and converts those inputs into outputs for further processing.



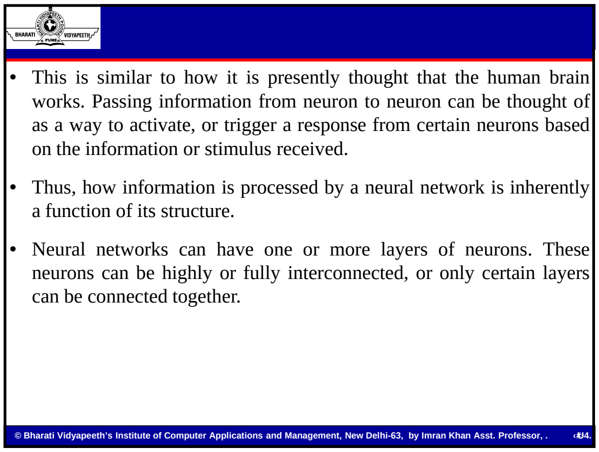
Elements Of (ANN)

- Network Information Processing
  - Inputs
  - Outputs
  - Connection Weights
  - Summation Function
  - Transformation/ Transfer Function

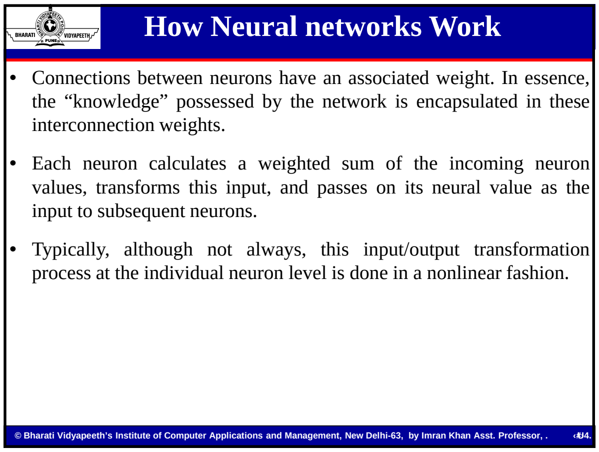


How Neural networks Work

- Several ANN paradigms have been proposed for applications in a variety of problem domains.
- As they are biologically inspired, the main processing elements of a neural network are individual neurons, analogous to the brain's neurons.
- These artificial neurons receive the sum "information" from other neurons or external input stimuli, perform a transformation on the inputs, and then pass on the transformed information to other neurons or external outputs.



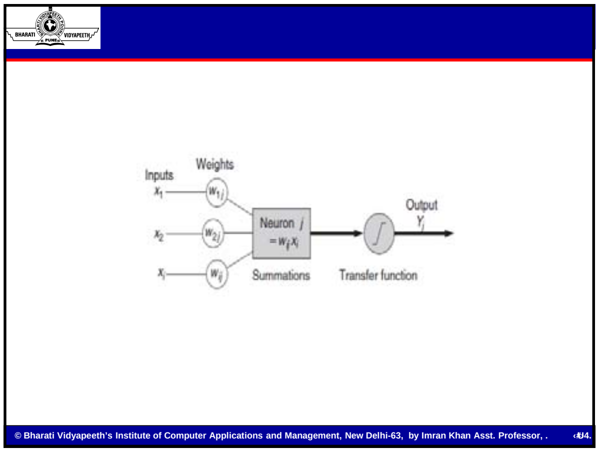
- This is similar to how it is presently thought that the human brain works. Passing information from neuron to neuron can be thought of as a way to activate, or trigger a response from certain neurons based on the information or stimulus received.
- Thus, how information is processed by a neural network is inherently a function of its structure.
- Neural networks can have one or more layers of neurons. These neurons can be highly or fully interconnected, or only certain layers can be connected together.



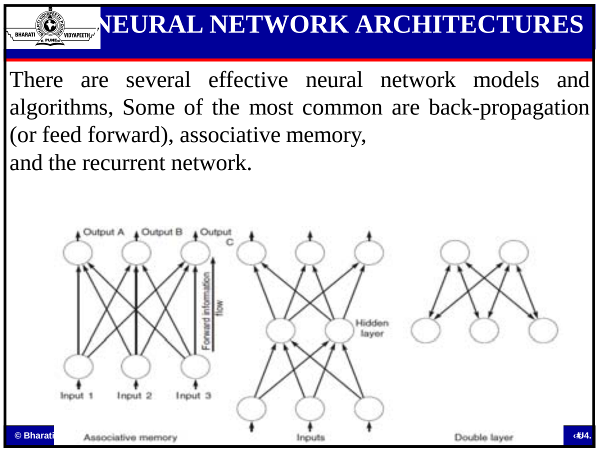
How Neural networks Work

- Connections between neurons have an associated weight. In essence, the “knowledge” possessed by the network is encapsulated in these interconnection weights.
- Each neuron calculates a weighted sum of the incoming neuron values, transforms this input, and passes on its neural value as the input to subsequent neurons.
- Typically, although not always, this input/output transformation process at the individual neuron level is done in a nonlinear fashion.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



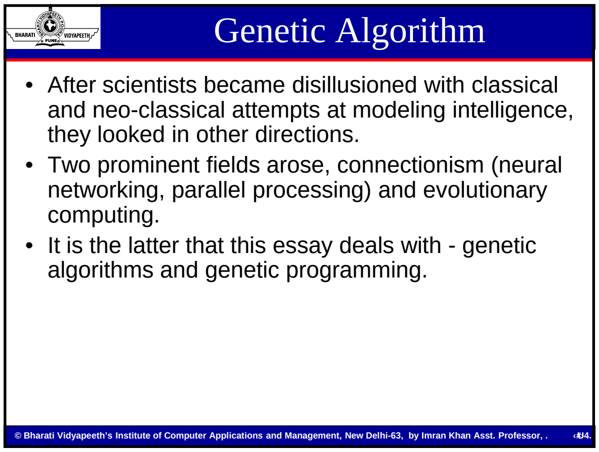
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



NEURAL NETWORK ARCHITECTURES

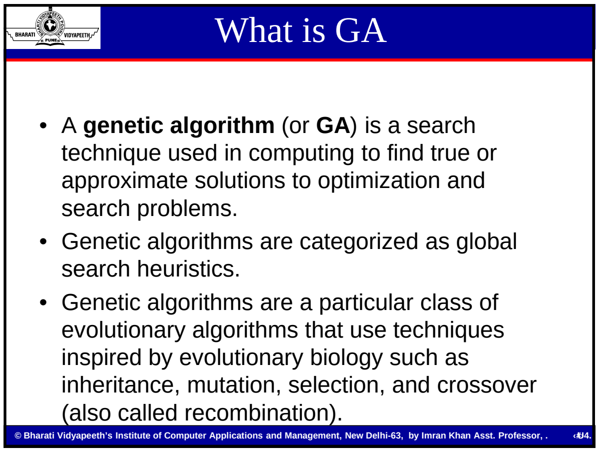
There are several effective neural network models and algorithms, Some of the most common are back-propagation (or feed forward), associative memory,
and the recurrent network.

© Bharati ‹#›U4.



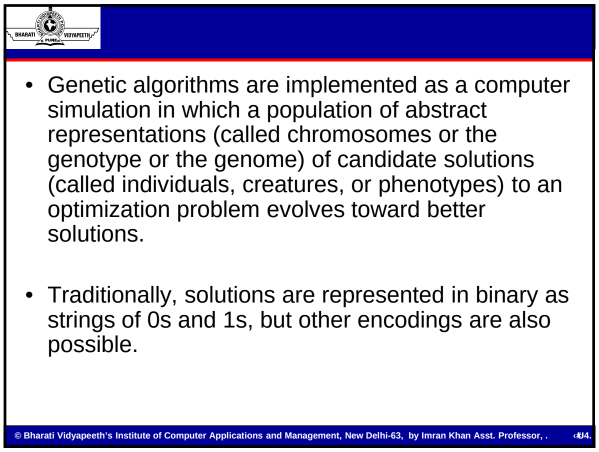
Genetic Algorithm

- After scientists became disillusioned with classical and neo-classical attempts at modeling intelligence, they looked in other directions.
- Two prominent fields arose, connectionism (neural networking, parallel processing) and evolutionary computing.
- It is the latter that this essay deals with - genetic algorithms and genetic programming.

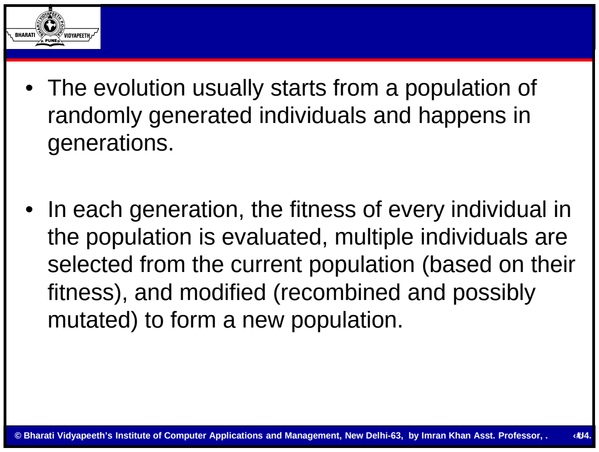


What is GA

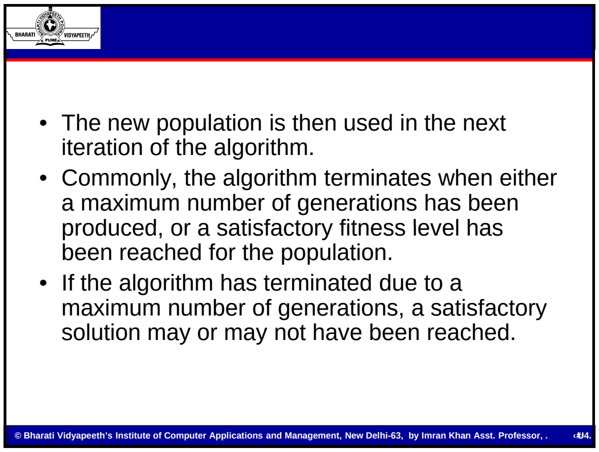
- A **genetic algorithm** (or **GA**) is a search technique used in computing to find true or approximate solutions to optimization and search problems.
- Genetic algorithms are categorized as global search heuristics.
- Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination).



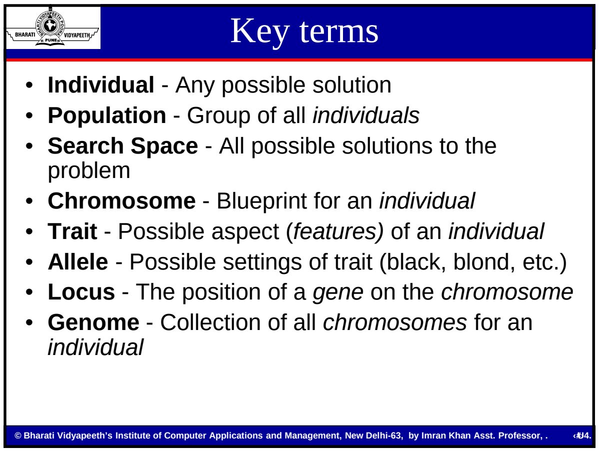
- Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions.
- Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.



- The evolution usually starts from a population of randomly generated individuals and happens in generations.
- In each generation, the fitness of every individual in the population is evaluated, multiple individuals are selected from the current population (based on their fitness), and modified (recombined and possibly mutated) to form a new population.

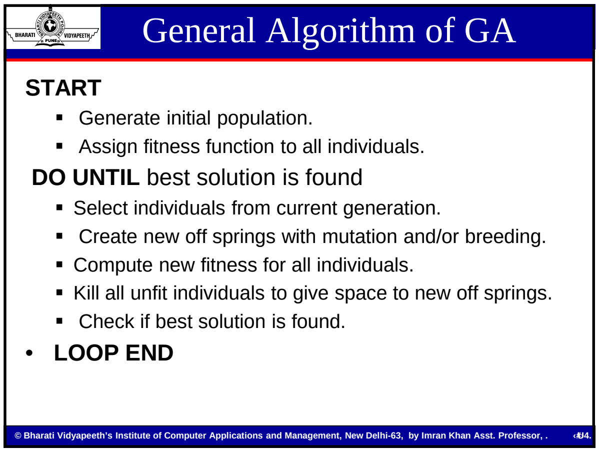


- The new population is then used in the next iteration of the algorithm.
- Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.
- If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.



Key terms

- **Individual** - Any possible solution
- **Population** - Group of all *individuals*
- **Search Space** - All possible solutions to the problem
- **Chromosome** - Blueprint for an *individual*
- **Trait** - Possible aspect (*features)* of an *individual*
- **Allele** - Possible settings of trait (black, blond, etc.)
- **Locus** - The position of a *gene* on the *chromosome*
- **Genome** - Collection of all *chromosomes* for an *individual*



General Algorithm of GA

START

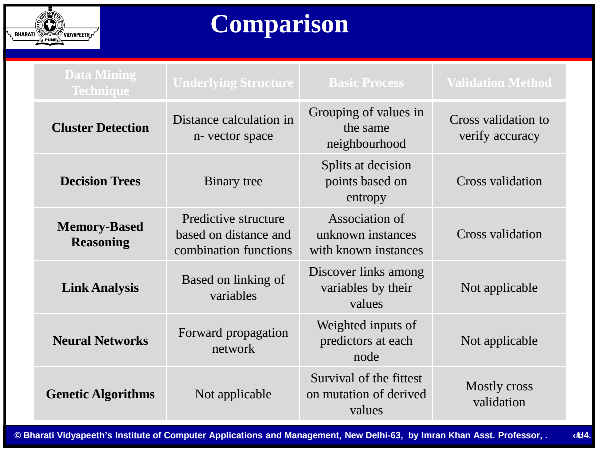
- Generate initial population.
- Assign fitness function to all individuals.

DO UNTIL best solution is found

- Select individuals from current generation.
- Create new off springs with mutation and/or breeding.
- Compute new fitness for all individuals.
- Kill all unfit individuals to give space to new off springs.
- Check if best solution is found.

- **LOOP END**

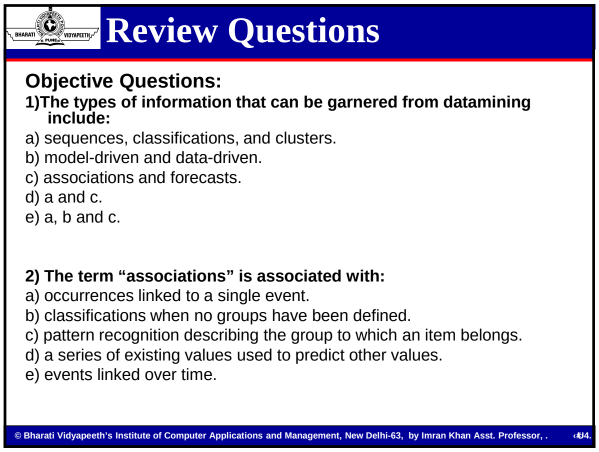
© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



Comparison

Data Mining Technique	Underlying Structure	Basic Process	Validation Method
Cluster Detection	Distance calculation in n- vector space	Grouping of values in the same neighbourhood	Cross validation to verify accuracy
Decision Trees	Binary tree	Splits at decision points based on entropy	Cross validation
Memory-Based Reasoning	Predictive structure based on distance and combination functions	Association of unknown instances with known instances	Cross validation
Link Analysis	Based on linking of variables	Discover links among variables by their values	Not applicable
Neural Networks	Forward propagation network	Weighted inputs of predictors at each node	Not applicable
Genetic Algorithms	Not applicable	Survival of the fittest on mutation of derived values	Mostly cross validation

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



Review Questions

Objective Questions:

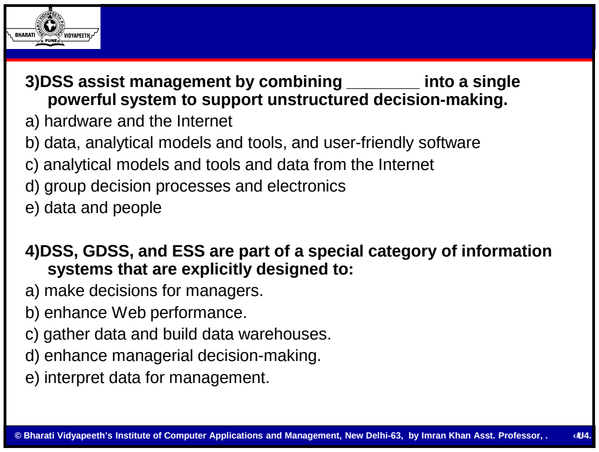
1)The types of information that can be garnered from datamining include:

a) sequences, classifications, and clusters.
b) model-driven and data-driven.
c) associations and forecasts.
d) a and c.
e) a, b and c.

2) The term "associations" is associated with:

a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . ‹#›U4.



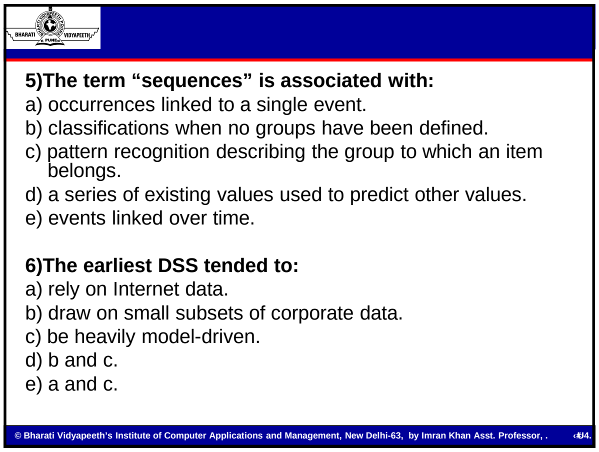
3)DSS assist management by combining ________ into a single powerful system to support unstructured decision-making.

a) hardware and the Internet
b) data, analytical models and tools, and user-friendly software
c) analytical models and tools and data from the Internet
d) group decision processes and electronics
e) data and people

4)DSS, GDSS, and ESS are part of a special category of information systems that are explicitly designed to:

a) make decisions for managers.
b) enhance Web performance.
c) gather data and build data warehouses.
d) enhance managerial decision-making.
e) interpret data for management.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . U4.



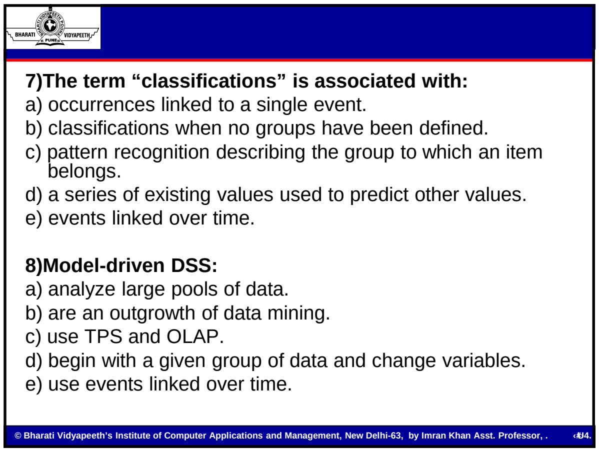
5)The term “sequences” is associated with:

a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

6)The earliest DSS tended to:

a) rely on Internet data.
b) draw on small subsets of corporate data.
c) be heavily model-driven.
d) b and c.
e) a and c.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . U4.



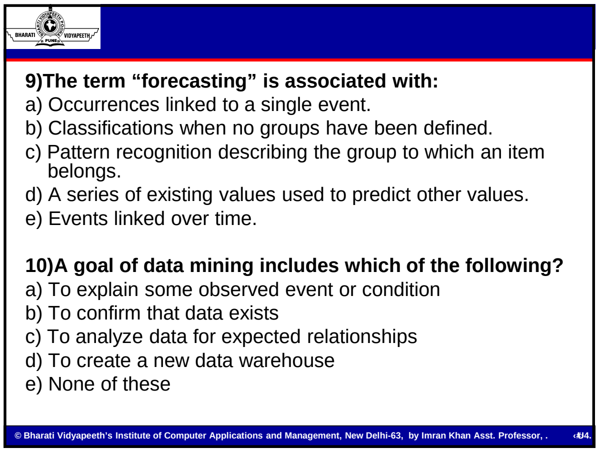
7)The term “classifications” is associated with:

a) occurrences linked to a single event.
b) classifications when no groups have been defined.
c) pattern recognition describing the group to which an item belongs.
d) a series of existing values used to predict other values.
e) events linked over time.

8)Model-driven DSS:

a) analyze large pools of data.
b) are an outgrowth of data mining.
c) use TPS and OLAP.
d) begin with a given group of data and change variables.
e) use events linked over time.

© Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi-63, by Imran Khan Asst. Professor, . U4.

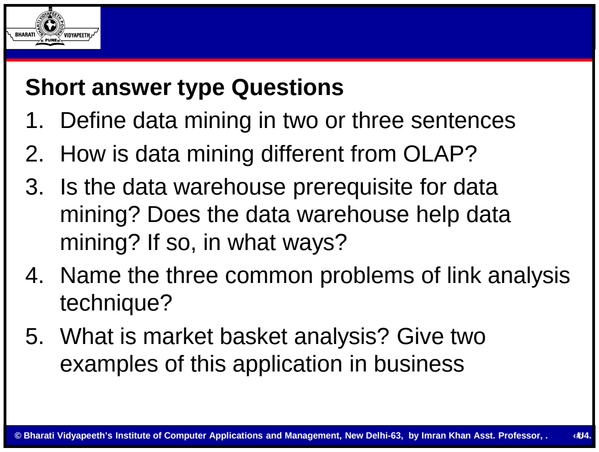


9)The term “forecasting” is associated with:

a) Occurrences linked to a single event.
b) Classifications when no groups have been defined.
c) Pattern recognition describing the group to which an item belongs.
d) A series of existing values used to predict other values.
e) Events linked over time.

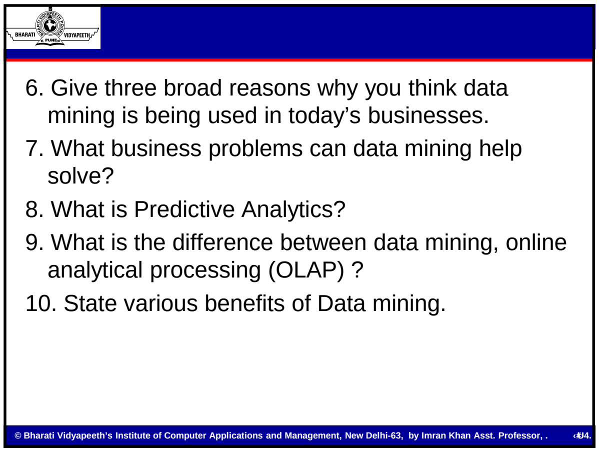
10)A goal of data mining includes which of the following?

a) To explain some observed event or condition
b) To confirm that data exists
c) To analyze data for expected relationships
d) To create a new data warehouse
e) None of these

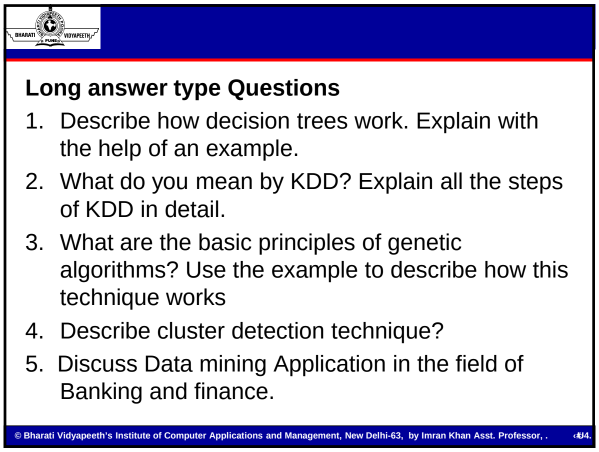


Short answer type Questions

1. Define data mining in two or three sentences
2. How is data mining different from OLAP?
3. Is the data warehouse prerequisite for data mining? Does the data warehouse help data mining? If so, in what ways?
4. Name the three common problems of link analysis technique?
5. What is market basket analysis? Give two examples of this application in business

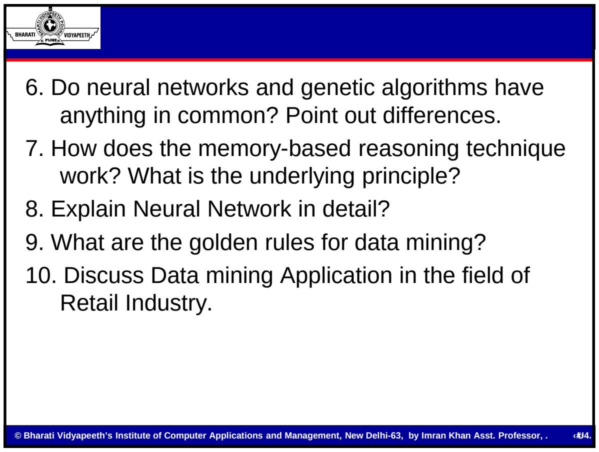


6. Give three broad reasons why you think data mining is being used in today’s businesses.
7. What business problems can data mining help solve?
8. What is Predictive Analytics?
9. What is the difference between data mining, online analytical processing (OLAP) ?
10. State various benefits of Data mining.



Long answer type Questions

1. Describe how decision trees work. Explain with the help of an example.
2. What do you mean by KDD? Explain all the steps of KDD in detail.
3. What are the basic principles of genetic algorithms? Use the example to describe how this technique works
4. Describe cluster detection technique?
5. Discuss Data mining Application in the field of Banking and finance.



6. Do neural networks and genetic algorithms have anything in common? Point out differences.
7. How does the memory-based reasoning technique work? What is the underlying principle?
8. Explain Neural Network in detail?
9. What are the golden rules for data mining?
10. Discuss Data mining Application in the field of Retail Industry.