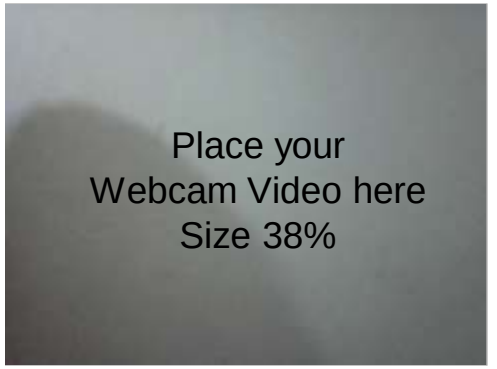
Data mining

Aiswarya V
Department of Computer Science
Amrita Vishwa Vidyapeetham

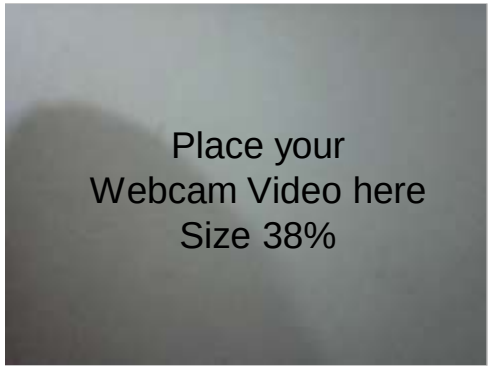
Types of data mined

- Data mining can be applied to any kind of data as long as the data are meaningful for a target application.
- The most basic forms of data for mining applications are database data , data warehouse data and transactional data.

Data Mining: On What Kinds of Data?

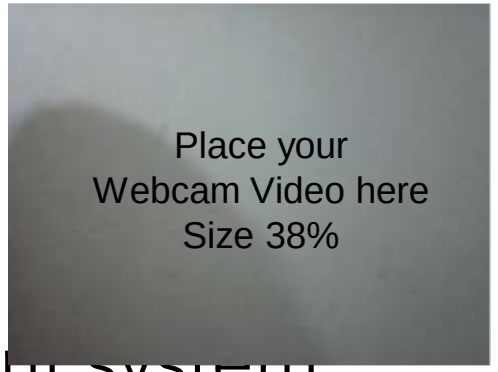


- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Other data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data



- Object-relational databases
- Heterogeneous databases and legacy databases
- Spatial data and spatiotemporal data
- Multimedia database
- Text databases
- The World-Wide Web

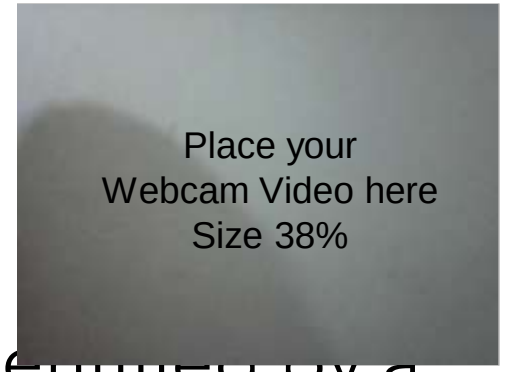
Relational database



- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.

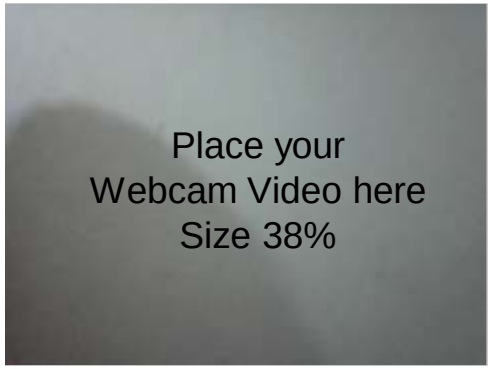
- A relational database is a collection of tables, each of which is assigned a unique name.
- Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).

<Content Slide>



- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases.
- An ER data model represents the database as a set of entities and their relationships

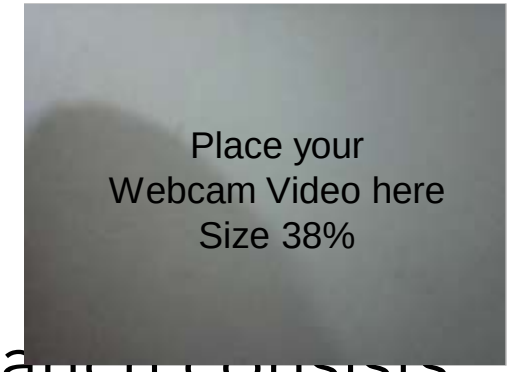
Example



A relational database for an Electronics store.

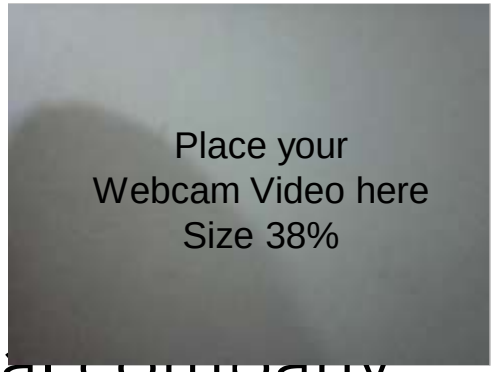
- The fictitious AllElectronics store is used to illustrate concepts.
- The company is described by the following relation tables: customer, item, employee, and branch. The relation customer consists of a set of attributes describing the customer information, including a unique customer identity number (cust ID), customer name, address, age, occupation, annual income, credit information, and category.

<Content Slide>



- Similarly, each of the relations item, employee, and branch consists of a set of attributes describing the properties of these entities.
- Tables can also be used to represent the relationships between or among multiple entities.
- In our example, these include purchases (customer purchases items, creating a sales transaction handled by an employee), items sold (lists items sold in a given transaction), and works at (employee works at a branch of AllElectronics).

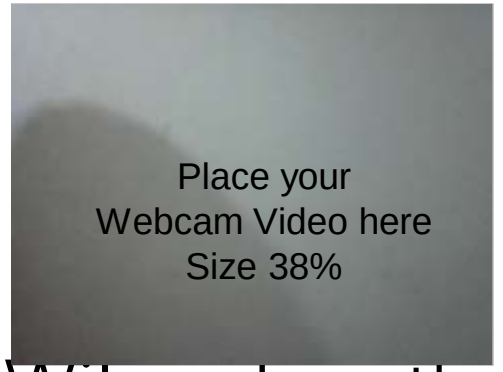
Data Warehouses



- Suppose that AllElectronics is a successful international company with branches around the world.
- Each branch has its own set of databases.
- The president of AllElectronics has asked you to provide an analysis of the company's sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites.
- If AllElectronics had a data warehouse, this task would be easy. Here comes the application of Datawarehouse

- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes.
- A data cube for summarized sales data of AllElectronics .
- The cube has three dimensions: address (with city values Chicago, New York, Toronto, Vancouver), time (with quarter values Q1, Q2, Q3, Q4), and item(with item type values home entertainment, computer, phone, security).
- The aggregate value stored in each cell of the cube is sales amount (in thousands)

Reference



1. William H Inmon "Building the Data Warehouse", Wiley, Fourth Edition 2005.

2. K.P.Soman, Shyam Diwakar and V.Ajay, "Insight into data mining Theory and Practice", Prentice Hall of India, 2006.

3. Jiawei Han, Micheline Kamber and Jian Pei, "Data mining concepts and Techniques", Third Edition, Elsevier Publisher, 2006.