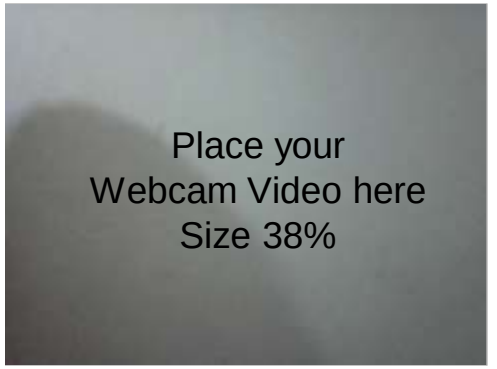
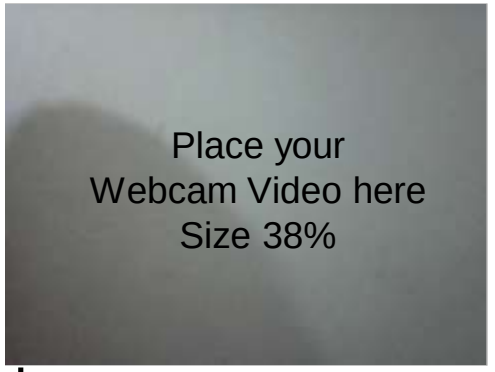
Data Mining

Aiswarya V
Department of Computer Science
Amrita Vishwa Vidyapeetham

Why Data mining?



- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …

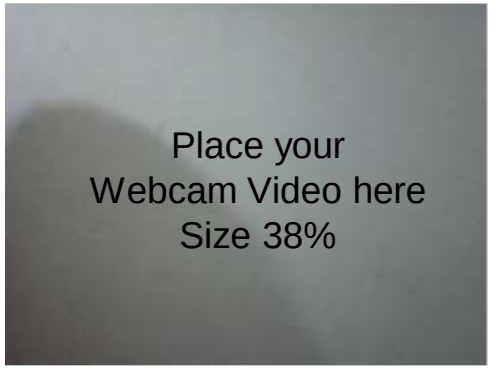


- Society and everyone: news, digital cameras, YouTube

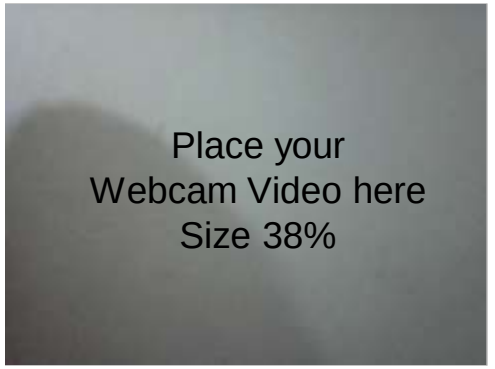
- We are drowning in data, but starving for knowledge!

- "Necessity is the mother of invention"–Data mining–Automated analysis of massive data sets

Evolution of Data science

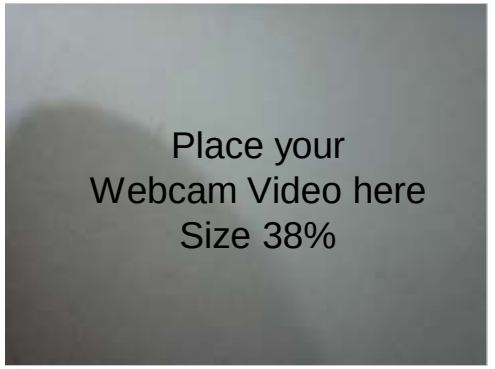


- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.

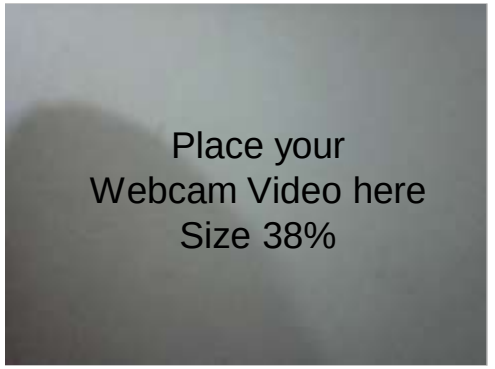


- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

Evolution of Database Technology

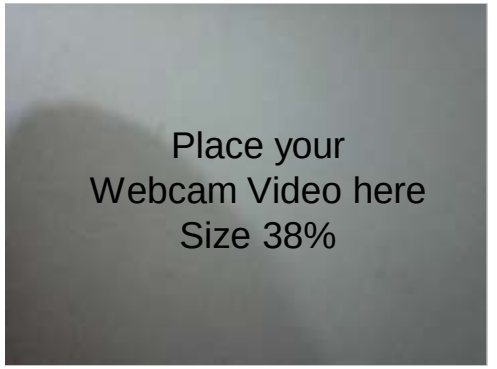


- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models



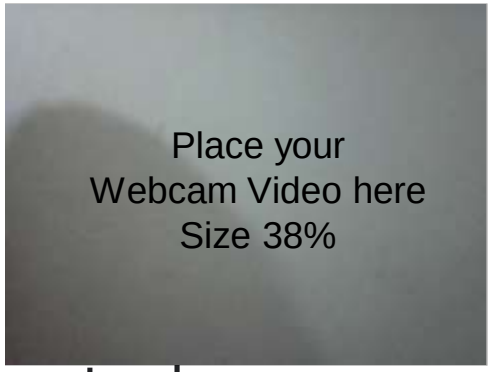
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

What is Data mining?



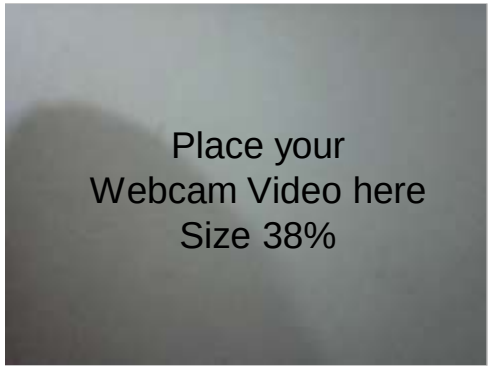
- Data mining (knowledge discovery from data)
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Characteristics of a data mining system



- Large quantities of data. The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.
- Noisy, incomplete data. ...
- Complex data structure. ...
- Heterogeneous data stored in legacy systems.

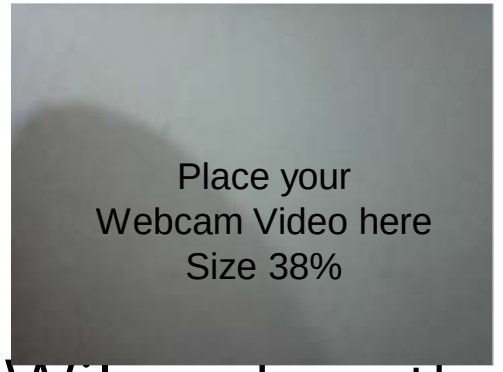
Benefits



Data mining benefits include:

- It helps companies gather reliable information.
- It's an efficient, cost-effective solution compared to other data applications.
- It helps businesses make profitable production and operational adjustments.
- Data mining uses both new and legacy systems.
- It helps businesses make informed decisions

References



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.