



**ST402**  
**Statistical Data Mining**  
Lecture 01: Introduction  
Prof. Roshan Dharshana Yapa  
Department of Statistics and Computer Science  
University of Peradeniya



### Course Overview


- Introduction to Data Mining
  - Why Data Mining? What is Data Mining? On What Kind of Data?
- Basic Data Mining Tasks
- Database/OLTP Systems, Data warehousing, Online Analytical Processing (OLAP)
- Data Mining Methods
  - Fuzzy Sets and Fuzzy Logic, Information Retrieval, Decision Trees
- Decision support systems, Web search engines
- Statistics, Machine learning, Pattern Matching
- Similarity measures, Neural Networks, Genetic Algorithms
- Classification
  - Statistical Based Algorithms, Distance Based Algorithms, Decision Tree Based Algorithms



10/22/2020 ST402 2

### Course Evaluation

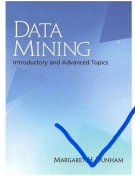
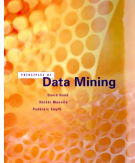
- Assignment – 15%
- Quiz – 15%
- End Semester – 70%

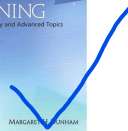


10/22/2020 ST402 3

### Text Books

- Principles of Data Mining –
  - David Hand, Heikki Mannila, Padhraic Smyth
- Data Mining- Introductory and Advanced Topics
  - Margaret H. Dunham, S. Sridhar

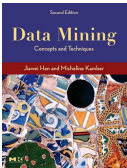




10/22/2020 ST402 4

### Text Books

- Data Mining: Concepts & Techniques
- Jiawei Han & Micheline Kanmber

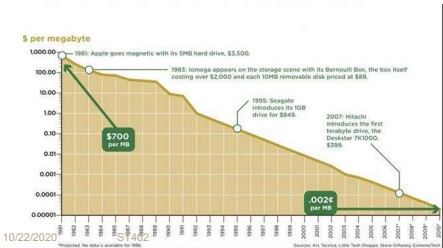


### Introduction

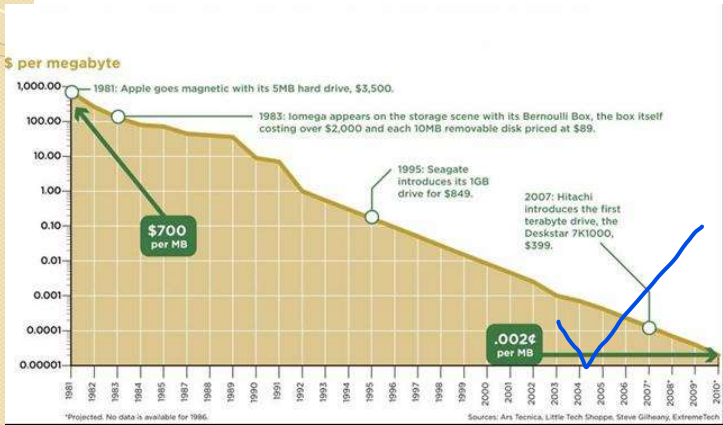


### Why Data Mining?

- Explosive growth in generation and storage of information
- More and more operations of enterprises are computerized due to technology advances in data acquisition
- At the end of 2004, HD storage was estimated to cost only about US\$1/GB



### Why Data Mining?



### Wal-Mart – A CASE Study

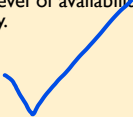
- A large chain of departmental store in USA
- More than 4000 stores world wide in 2004
- Generate more than 20 million point of sales transactions every day
- Managed a database of more than 460 terabytes (in 2004)

1 terabyte =  $10^{12}$  bytes  
Roughly equivalent to storing 12 million books



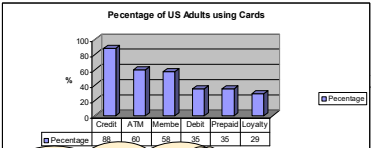
### Why Data Mining?

- Growth in Online Transaction Processing Data (*OLTP*)
  - The first database systems were implemented in the 1960's and 1970's
  - Many enterprises have more than 40-50 years of experience in using database systems
  - Therefore, they have accumulated large amount of data.
- NB: What is *OLTP*?
  - Conventional database systems are often design for day-to-day running of an organization and are called *OLTP* systems.
  - These systems are designed to capture business transactions online and are optimized for high throughput, a high level of availability and in some cases high level of information security.



### Why Data Mining?

- Growth in data due to cards
  - Growing use of credit cards and loyalty cards is an important area of data growth
  - In many developed & developing countries there has been a tremendous growth in the use of loyalty cards



If 50 transactions per CC per year, CC companies would be collecting 100 billion trans per year!

|               | Cards(millions) | Population(millions) | Cards per capita |
|---------------|-----------------|----------------------|------------------|
|               | 755             | 293                  | 2.58             |
|               | 177             | 1294                 | 0.14             |
|               | 148             | 184                  | 0.80             |
|               | 126             | 60                   | 2.10             |
|               | 121             | 127                  | 0.95             |
|               | 109             | 45                   | 1.31             |
|               | 95              | 47                   | 2.02             |
|               | 60              | 22                   | 2.73             |
|               | 56              | 39                   | 1.44             |
|               | 51              | 31                   | 1.65             |
| Total Top Ten | 1700            | 2180                 | 0.78             |
| Total Global  | 2362            | 6443                 | 0.37             |

Top ten credit card holding countries (VISA & MASTER) in 2003

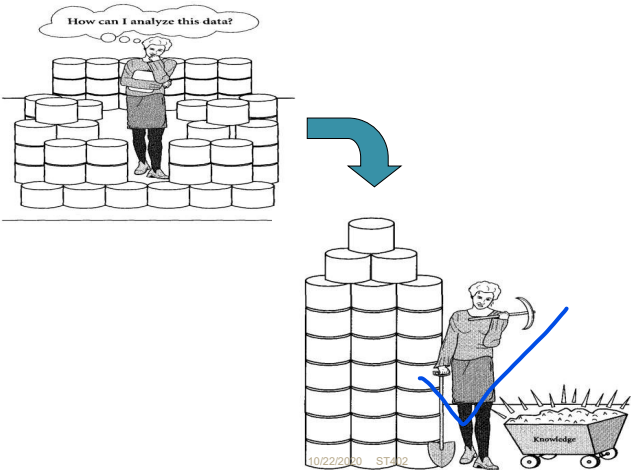
Source: International Card Manufacturers Association

### Why Data Mining?

- Growth in Data Storage Capacity
  - Annual disk storage sales in 2003 were 16,000 petabytes ( $10^{15}$  bytes) of storage compared to only around 4000 petabytes in 2000
- Growth in data due to the Web
- Decline the cost of Processing
- Growth of data due to the other sources
  - Telephone transactions
  - Medical transactions
  - Immigration and customs transactions
  - Banking transactions
  - Shopping transactions, etc.



## Why Data Mining



## Data Mining Applications

- Data mining can be classified in to six groups
  - Prediction and Description
    - "Would this customer buy a product?", "is this customer likely to leave"
    - Sales forecasting and analysis
    - **Techniques:** Selecting some or all the objects available in the database to predict other variables of interest.
  - Customer Relationship Marketing (CRM)
    - Customers have a lifetime value. (Not just a value of a single sale)
    - Analysing customer profiles, discovering sales triggers, critical issues that determine client loyalty and helps in improving customer relations
    - **Techniques:** Cluster analysis to identify customers suitable for cross-selling other products.

## Data Mining Applications

- Customer Profiling
  - Process of using the relevant and available information to describe the characteristics of a group of customers and to identify their discriminators from other customers or ordinary customers and drivers for their purchasing decisions
  - Profiling can help an enterprise identify its most valuable customers and differentiate their needs and values.
  - Profiling may also facilitate loan and credit card approval or approval of insurance applications for their valued customers.
- Outlier Identification and Fraud Detection
  - Many uses of data mining in identifying outliers, fraud or unusual cases
  - Ex. Identify unusual expense claims by staff, identify anomalies in expenditure between similar units of an enterprise, fraud involve in credit or telephone cards

## Data Mining Applications

- Customer Segmentation
  - Identify individuals in the market based on their status and needs
  - This helps for promoting the cross-selling of services and increasing customer relations
  - Used for evaluating of performance of various banking channels such as phone banking and internet banking
  - Understand and predict customer behaviour and profitability, to develop new products and services, and to market new offerings
- Web Site Design and Promotion
  - Discover how users navigate a web site and results can help in improving the site design and making it more visible on the web
  - Improve cross-selling by suggesting to a web customer that he/she may be interested in, through correlating properties about the customer, or the items the person has ordered.

## The Data Mining Process

- Decision makers of today's business need an information system that is "customer-centric"
- OLTP and legacy systems usually not suitable for data mining.
- Conventional database systems are often designed for day-to-day running of an organization (Online Transaction Processing).
- These systems are designed to capture business transactions online and are optimized for high throughput, high level of availability and information security.
- Hence a separate database is needed to store information needed for data mining.

## A Typical Data Mining Process

- Typical Data mining process is likely to include the following steps,
  - Requirement Analysis
    - The enterprise decision makers need to formulate goals that the data mining process is expected to achieve.
    - The business problem must be clearly defined.
    - The techniques to be used and the data that is required are likely to be different for different goals.
    - If the goals are clearly defined is easier to evaluate the results.
  - Data Selection and Collection
    - This step may include finding the best source database for the data that is required.

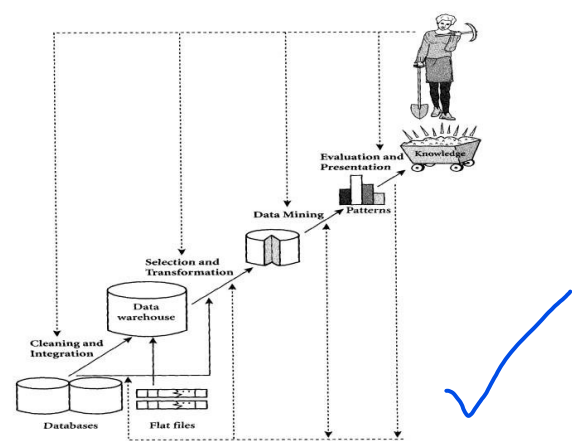
## A Typical Data Mining Process

- Cleaning and Preparing Data
  - Essentially, a data store that integrates data from a number of databases may need to be created.
  - When integrating, one often encounters problems like identifying data, dealing with missing data, data conflicts and ambiguity etc.
  - Some times this needs more that 50% of effort in data mining project.
- Data Mining Exploration and Validations
  - Assuming that the user has access to one or more data mining tools, a data mining model may be constructed based on enterprise's needs.
  - It may be possible to take a sample of data and apply a number of techniques. For each technique the results should be evaluated and their significance interpreted.
  - This is an iterative process which should lead to selection of one or more techniques that are suitable for further exploration, testing and validation.

## A Typical Data Mining Process

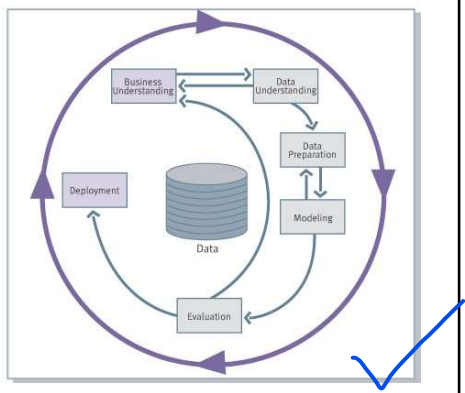
- Implementing, Evaluating and Monitoring
  - Once a model has been selected and validated, the model can be implemented for use by decision makers
  - This may involve software development for generating reports, for results visualization and explanation
  - If more than one technique is available for given data mining task, it is important to select the results and choose the best technique.
  - Every enterprise evolves with time and therefore, monitoring is likely to lead from time-to-time to refinement of tools and techniques that have been implemented
- Results Visualization
  - Explaining the results of data mining to the decision makers is an important step.
  - Sophisticated data visualisation tools are being developed to display results that deal with more than two dimensions.

### A Typical Data Mining Process



### CRISP-DM (CRoss Industry Standard Process for Data Mining)

- Another data mining model proposed by consortium of vendors and users (Including SPSS)



### CRISP-DM

- Business Understanding
  - Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
- Data Understanding
  - Starts with initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.



### CRISP-DM

- Data Preparation
  - Tasks include table, record and attribute selection as well as transformation and cleaning of data for modelling tools.
- Modelling
  - In this phase, various modelling techniques are selected and applied and their parameters are calibrated to optimal values.
  - Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
  - Therefore, stepping back to the data preparation phase is often necessary.





## CRISP-DM

- Evaluation
  - At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective.
  - Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.
- Deployment
  - Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.



## Data Mining Tasks



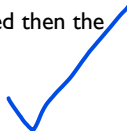
## Data Mining Tasks

- It is convenient to categorize Data Mining methods into tasks
    - According to the different types of objectives
    - Should not be unique, further division is allowed if possible
1. Exploratory Data Analysis (EDA)
  2. Predictive Modeling
  3. Descriptive Modeling
  4. Discovering Patterns and Rules
  5. Retrieval by Content
  6. Sequence discovery / Sequence Analysis
  7. Web Mining / Search Engines

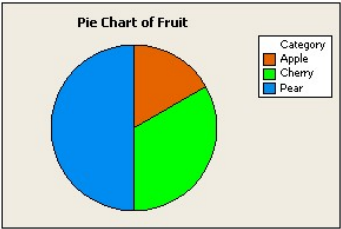
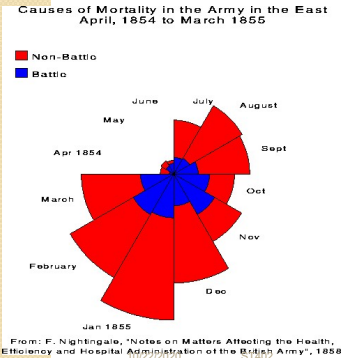


## I. Exploratory Data Analysis (EDA)

- Interactive and Visual
- Simply explore the data without clear idea
- When the number of attributes increased then the complexity of the analysis also increased
  - E.g.: Pie Chart, Coxcomb plot, Time series plot, Scatter plot
  - Consider a Scatter plot, when number of variables increased then the dimensionality is also increased

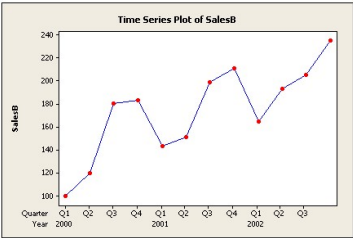


# I. Exploratory Data Analysis (EDA) : Examples

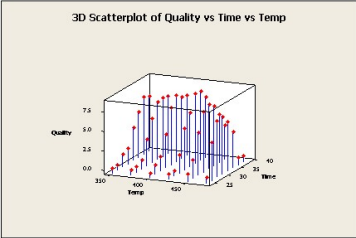


Use to display the proportion of each data category relative to the whole data set.

# I. Exploratory Data Analysis (EDA) : Examples



Company's quarterly sales for 2001 to 2003



Need to determine the optimal time and temperature for reheating a new frozen food

## 2. Predictive Modeling - Classification

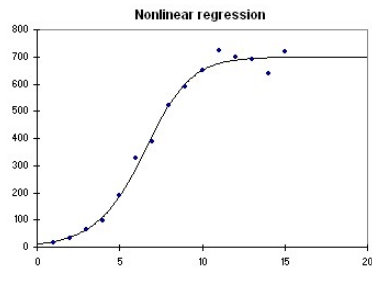
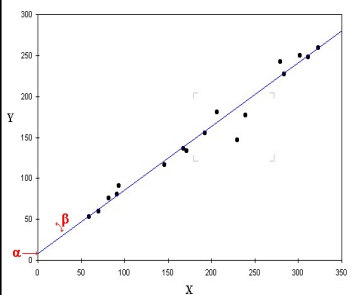
- Maps data into pre-defined classes/groups
- These classes are determined before examined the data
  - E.g.:An input pattern is classified into one of several classes based on its similarity to the predefined classes
    - Character recognition
  - Identifying good risk for granting a long or insurance
  - Whether an is likely to respond for a direct mail solicitation
  - Decision Tree is widely used technique

## 2. Predictive Modeling - Regression

- Maps data into a real valued prediction variable
- Assumes that the target data fit into some known type of function
  - Linear
  - Non-linear
  - etc
- Determines the best model that describes the data
  - E.g.: Person wished to reach a certain level of savings before his retirement
    - Use a linear regression model which based on current values and some past values



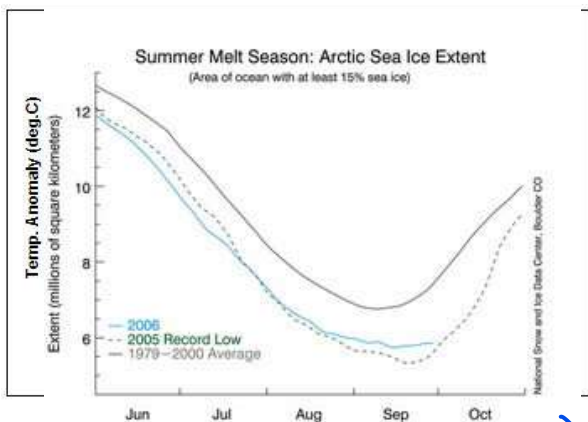
## 2. Predictive Modeling - Regression



## 2. Predictive Modeling – Time Series Analysis

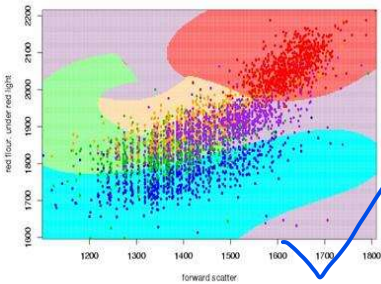
- Examine whether an attribute varies over **time**
- Values are obtained as evenly spaced time
  - Weekly, Daily, Yearly etc.
- To examine similarities among different time series
- Examine it's behavior according to the structure
- Used to predict future values
  - E.g.:
    - Yearly rainfall in a country
    - Global warming

## 2. Predictive Modeling – Time Series Analysis



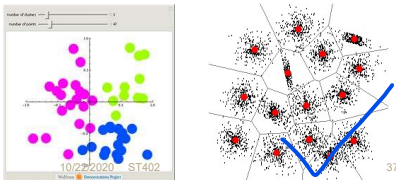
## 2. Predictive Modeling – Prediction

- This is a composition of
  - Classification
  - Regression
  - Time Series Analysis



### 3. Descriptive Modeling - Clustering

- Similar to Classification
- Classes are not pre-defined
  - Known as Unsupervised learning or clustering
- Experts can interpret the clusters at the end of the exercise
  - E.g.: Identifying purchasing patterns in a Supermarket
    - According to the personal details
    - Or food choice
  - Methods: K-Means Algorithm



### 3. Descriptive Modeling - Summarization

- Maps data into subsets associated with simple descriptions
- Characterizes the content of the database
  - E.g.: The average GPA can be considered to compare the students on their performance during an academic year



### 4. Discovering Patterns and Rules: (Association Rule Mining/ Market Basket Analysis)

- Is a technique that analysis a set of transaction captured at a supermarket checkout.
- Each transaction being a list of products or items purchased by one customer.
- The aim of association rules mining is to determine which items are purchased together frequently, so that they may be grouped together on store shelves or the information may be used for cross-selling.

Example 1: Analysis of purchases in a supermarket

|           |         |        |        |            |         |
|-----------|---------|--------|--------|------------|---------|
| customer1 | pizza   | beer   | cheese | bread      | chips   |
| customer2 | milk    | bread  | ham    | cigarettes |         |
| customer3 | yoghurt | sugar  | flour  | cornflakes | napkins |
| customer4 | shampoo | beer   | chips  | newspaper  | pizza   |
| xustomer5 | chips   | coffee | beer   | pizza      | cream   |
| customer6 | jam     | rolls  | butter | beer       |         |

If pizza and beer are in one purchase, it is likely that chips are also in that purchase.



### 5. Retrieval by Content

- Before the process is taken place user has an exact pattern
- Target is to find a matching pattern
- Commonly used for text and image data
  - E.g.:
    - Searching Google (Text mining)
    - Cancer detection (Image analysis/processing)



## 6. Sequence discovery / Sequential Analysis

- Determines patterns based on a time sequence of actions
- E.g.:
  - Most people who buy a CD player may be found to purchase CDs within one week
  - If 70% of the users of web page **A** follow one of the following pattern behavior
    - Pattern 1 - [A, B, C]
    - Pattern 2 - [A, D, B, C]
    - Pattern 3 - [A, E, B, C]

10/22/2020 ST402

41

## 7. Web Mining / Search Engines

- Web Data Mining
  - Web mining may divided into several categories as, web content mining, web structure mining, web usage mining etc.
- Search Engines
  - They are huge databases of Web pages as well as software packages for indexing and retrieving the pages that enable users to find information of interest to them.

10/22/2020 ST402

42