BITS Pilani

BITS Pilani
Hyderabad Campus

Prof.Aruna Malapati
Professor
Department of CSIS

Data Mining - Introduction

Today's Learning objective

innovate achieve lead

- Course logistics
- Data Mining Definition
- DM vs. Machine learning (ML)
- Data Mining Process
- Data Mining Tasks
- Challenges in Data Mining

Course Team

1. Instructor:

   Prof.Aruna Malapati

2. Ph.D TAs

   Ms. Priya Baju

   Mr. Suchit Bhai Patel

   Mr. Danny Muzata

Project

- Max of 3 members in a group.
- Implement and test Data Mining algorithms taught.
- The team will give you only ideas/datasets however you can pick up your problems.

Learning Objectives

innovate achieve lead

- Identify Data Mining problems.
- Apply appropriate algorithms to solve Data Mining problems.
- Implement validation approaches.
- Combine multiple approaches to get better results

Course Modules

- Introduction to Data Mining
- Data Preprocessing
- Dimension Reduction
- Association Rule Mining
- Clustering
- Outlier Analysis
- Classification Techniques

Basic mathematics and statistics

innovate achieve lead

- Reading mathematical notations
- Basic concepts of probability theory (distributions, conditional probability, independence)
- Summary statistics like mean, median, mode, standard deviation, covariance
- Matrices and their basic operations

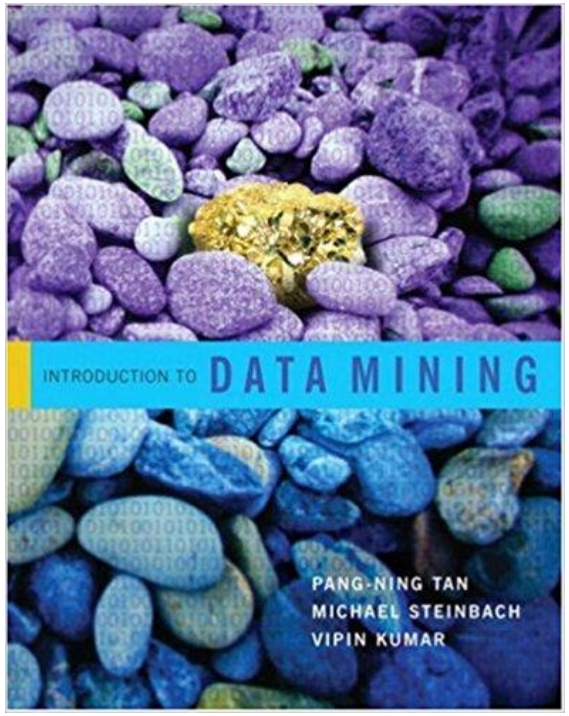
Evaluation

Component	Duration	Weightage (%)	Date & Time	Nature of Component
Mid Term Exam	90 Mins.	30	15/03/2024 11.00 - 12.30PM	Closed Book
Quizzes (Best 2 out 3)	30 Mins	10	Q1- 10/02/2024 Q2- 23/03/2024 Q3- 20/04/2024	Closed Book
Project (Phase-1 evaluation before Pre-Mid)	--	20	TBA	Open Book
Comprehensive	3 Hours	40	16/05 AN	Closed Book

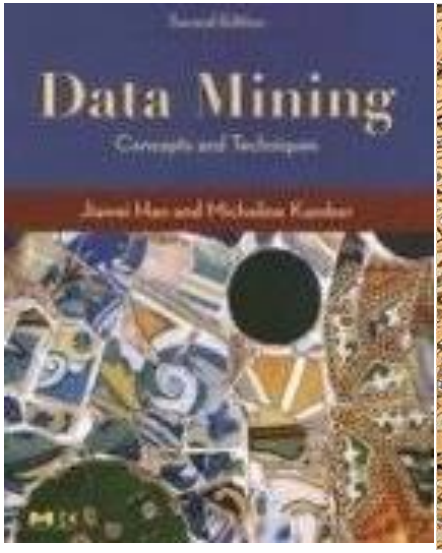
Chamber Consultation Hour: Sat 11.30-12.30 @ H 132
Notices: CMS only
Make-up Policy: Prior Permission is a must, and Make-up shall be granted only in genuine cases based on the individual's needs and circumstances. The recommendation from the chief warden is necessary to request a make-up.

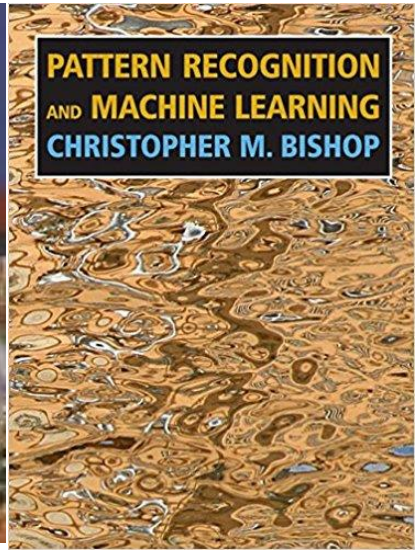
Books

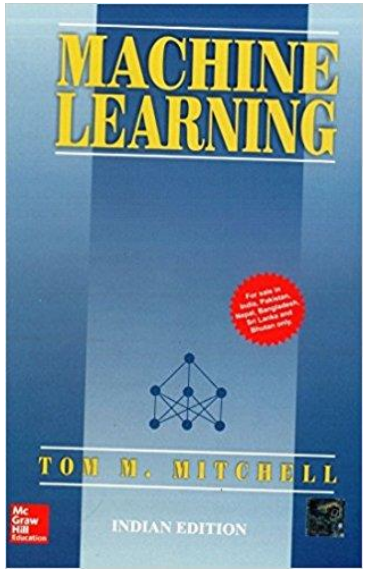
TEXT BOOK



REFERENCE BOOKS



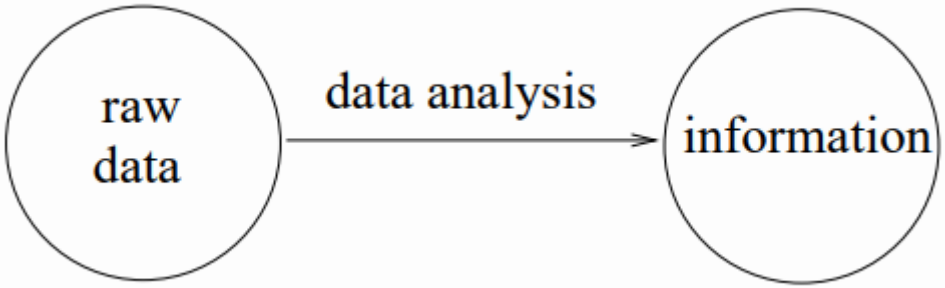




What is Data mining (DM)?

- 'A non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." (Fayyad et al. 1996)
- "Automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies." (wikipedia)
- "Analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner." (Hand et al. 2001)

Data vs. Information?

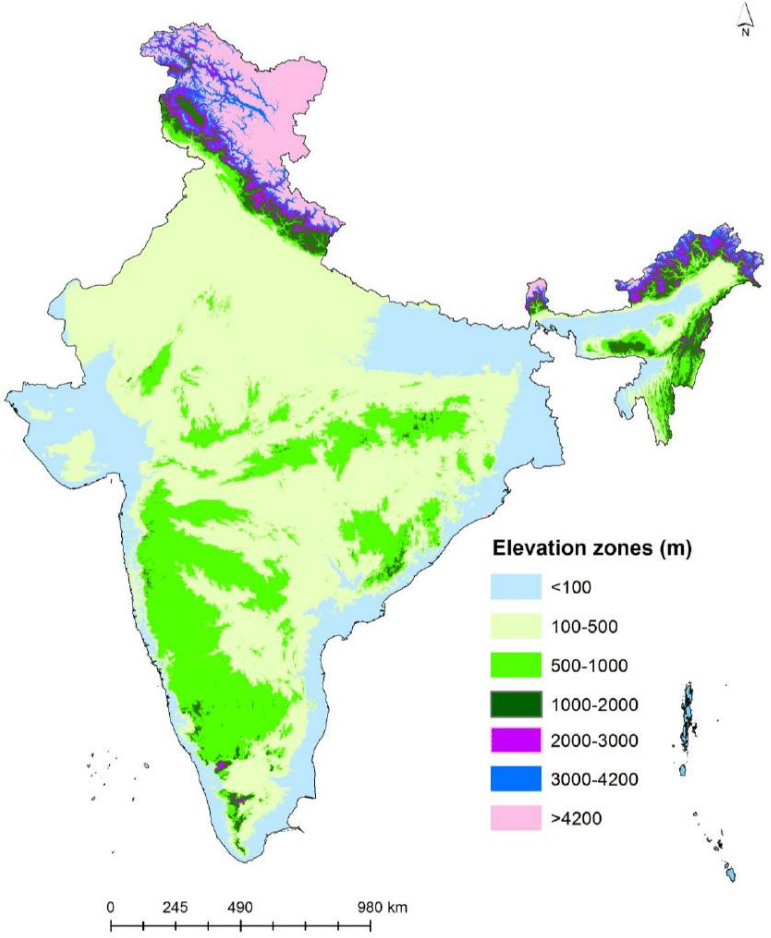


- raw data = unprocessed, uninterpreted facts (e.g, measurements)
- information = knowledge that has meaning, “interpreted data”
- relative terms: the resulting information from one process may be source data for another process

Example: Data vs. Information

- data: elevation measurements across India over many years ⇒ pieces of information:

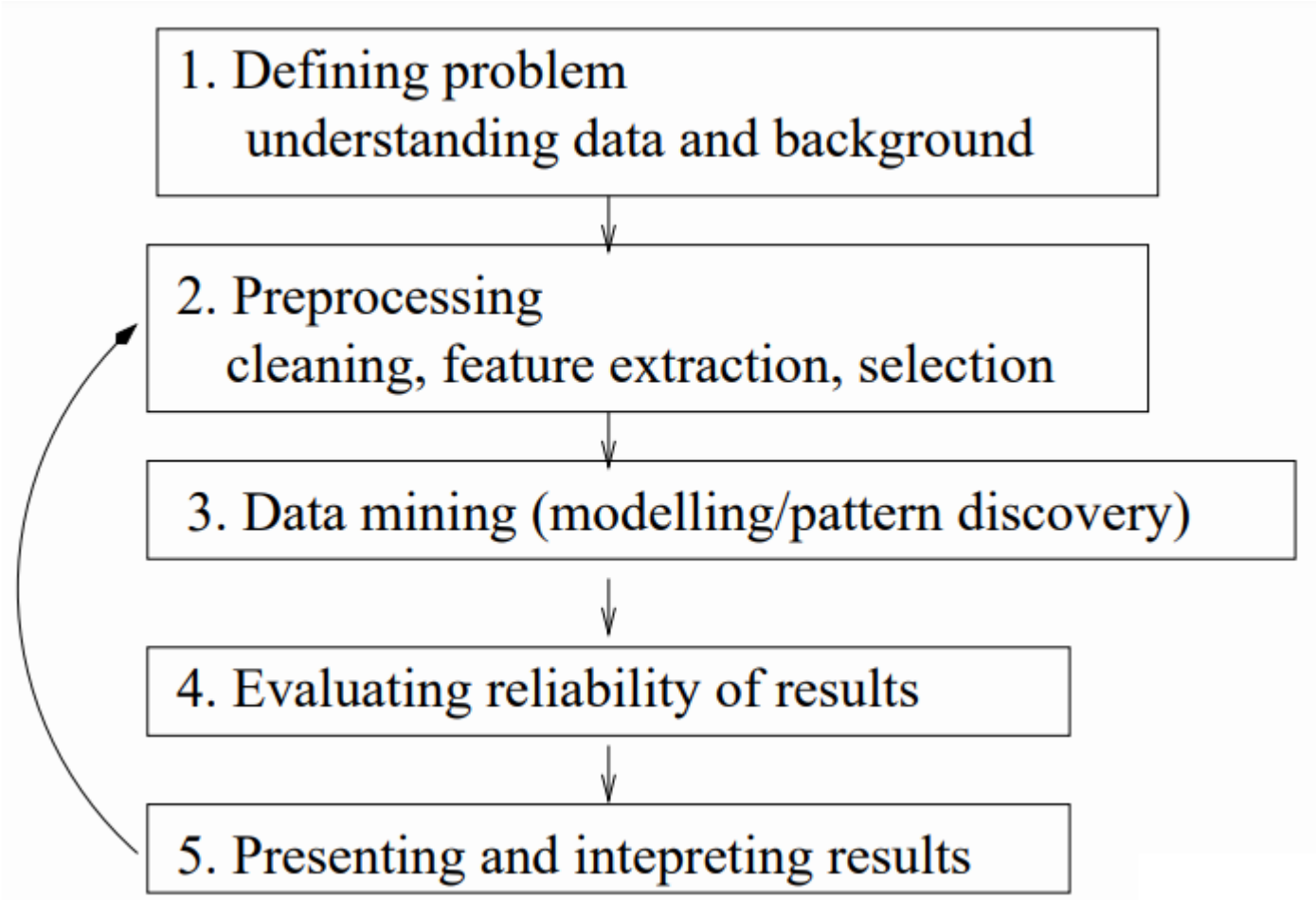
- The highest point is in the Himalayas.
- Most of India is lowland (≤ 200m from the sea level)
- Mean elevation is about 154m
- The landmass raises about 9mm per year in Tamilnadu area.



DM vs. Machine learning (ML)?

- assumptions on the origin of data
  - ML: there is an underlying model that has generated data ⇒ find the best explaining model
  - DM: data is primary, there isn't necessarily any model or may be only local patterns ⇒ find interesting patterns
- different emphasis:
  - ML: more focus on supervised methods (given sample output, learn a model: input → output)
  - DM: more focus on pattern discovery and unsupervised methods (learn inherent structure in data)

DM process


1. Defining problem
understanding data and background
2. Preprocessing
cleaning, feature extraction, selection
3. Data mining (modelling/pattern discovery)
4. Evaluating reliability of results
5. Presenting and intepreting results

1. Defining the problem

- Understanding data: what variables measure/describe?
- What are data types? How much there is data?
- What kinds of patterns would be interesting or useful to find?
- What is already known?
- It is worth studying some background theory!
- Difficulty: How people from different fields find the same language?

2. Preprocessing

- Combining data from different sources (may require transformations)
- Preliminary analysis: means, standard deviations and distributions of variables, correlations, ... (e.g., with statistical tools)
- Data cleaning: handling missing values, detecting and correcting errors
- Deriving new potentially good features (e.g., combining variables, discretization, normalization)
- Possibly dimension reduction (combines feature extraction and selection)

2. Preprocessing (cont'd)

- Choosing good features for the modelling purpose and pruning out irrelevant or redundant features Feature selection can also be done during modelling
- If possible, don't prune anything, before you know which variables are useless
- If computationally feasible, try many alternative versions of the data (with alternative features)
- Feature selection methods never guarantee optimal results! (NP-hard problem, can't try all $2^k - 1$ possible feature sets!)

3. Data mining

- Always good to begin from dependency analysis! → choosing features and modeling methods
- Usually, descriptive modeling helps in building a predictive model
  - e.g., gene–habit–disease data
  - Descriptive: Find the 100 most significant association rules related to variable Diabetes
  - Predictive: Learn (from selected data) the best model that predicts diabetes
  - Typical building blocks of the DM step dependency analysis, classification, clustering, outlier detection It is always good to begin with It is always good to begin with

4. Evaluating reliability of results

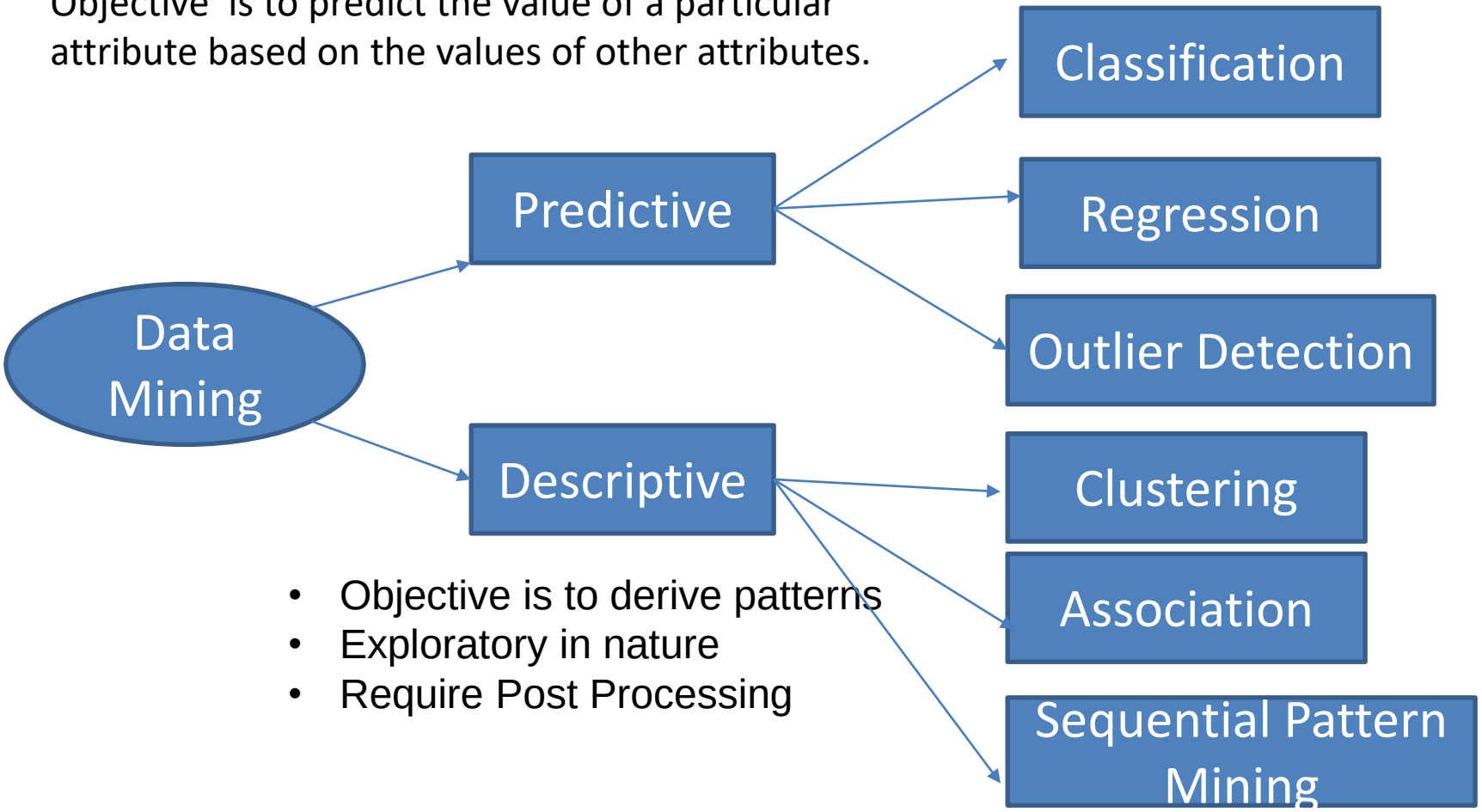
- Are discovered patterns or models sensible?
- Validating predictive models easy:
- determine goodness measure (or error measure)
- learn a model from a training set and evaluate the error in the test set
- training and test sets have to representative!
- alternative (if too little data) cross-validation
- tools: statistical significance testing, use of validation data

5. Presenting and interpreting results

- present results illustratively so that essential things are emphasized
- domain experts have an important role!
- Did you find something new? Could you formulate a hypothesis based on results? What should be studied further?
- leads often to a new DM round; try new variables and possibly other methods

Data Mining Tasks

Objective is to predict the value of a particular attribute based on the values of other attributes.



- Objective is to derive patterns
- Exploratory in nature
- Require Post Processing

Challenges in Data Mining

innovate achieve lead

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
  - High complexity of data
  - Noisy and unreliable
  - Dynamically evolving
  - High dimensionality
  - Multiple heterogeneous sources
- New and sophisticated applications

Take home message

- Data Mining refers to non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data
- Data Mining covers topics including warehousing, association analysis, clustering, classification, anomaly detection, etc. (based on the type of mined knowledge), as well as transaction data mining, stream data mining, sequence data mining, graph data mining etc. (based on the type of data)
- Data Mining has wide applications in many fields in business, science, engineering, education, and many more.