SDS 3120: Data Mining

Dr.Mutua Kilai

May 2024

Department of Pure and Applied Sciences

Purpose of the Course

ⓘ Important

The purpose of the course is to equip learners with necessary skills for developing, manipulating, recovering and maintaining databases.

Learning Outcomes

- By the end of the Course the learner should be able to:
  - Explain concepts of data mining
  - Apply mining on different kinds of data
  - Apply mining on different kind of knowledge
  - Discuss principles of pattern analysis and machine learning
  - Demonstrate representation discovery

Course Content

- In this course we will consider the following problems:
  - Classification, cluster and outlier analysis
  - Mining time series and sequence data
  - Text mining, web mining and pattern analysis

Course Content Cont’d

- We cover fundamental data mining ideas:
  - Clustering, SVM
  - Semi-supervised learning
  - Information retrieval
  - Most important algorithms in data mining

Logistics of the Course

- **Course Website:** https://sam-mutua.github.io/datamin23/
- All Lecture Notes will be uploaded per week.
- All Assignments will also be posted in the course website.
- The website also has useful resources on data mining.

Programming Languages

- We will use R and Python programming languages which are **Free and Open Source.**
- We will also use SQL for databases lectures.
- Details on Installing the softwares can be found here

Data Mining

- **Data Mining** is the process of automatically discovering useful information in large data repositories.
- Data mining techniques are used in order to find useful patterns.
- Data mining is a key part of **knowledge discovery in databases(KDD)**.

KDD Process

Data Mining: Searching for patterns in data

- The knowledge discovery process is an iterative sequence with several steps.

Steps of KDD

1. **Data Cleaning** - The removal of noise and inconsistent data.
2. **Data Integration** - Combining of multiple data sources.
3. **Data Selection** - Data relevant to analysis is selected
4. **Data Transformation**- Data are transformed into various forms.
5. **Data Mining** - Extracting data patterns
6. **Pattern Evaluation** - Identify interesting patterns in data

Key Points

ⓘ Important

Data Mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data.

💡 Tip

The data sources can be:

- Databases
- Data warehouse
- Web
- Other information repositories

Motivation

- Specific challenges that motivated data mining:
  - *Scalability*: Data mining algorithms to handle massive data
  - *High dimensionality*: Encounter data sets with thousands of attributes.
  - *Heterogenous and Complex Data*: Need to handle complex attributes
  - *Data Ownership and Distribution*: Data that is geographically distributed not in one place.

Origin of Data Mining

- Data mining draws upon ideas such as:
  - Sampling, estimation, hypothesis testing
  - Search algorithms
  - Modeling techniques
  - Pattern recognition
- Data mining has also adopted ideas from other areas such as:
  - Optimization
  - Evolutionary computing
  - Information theory
  - Signal processing

What kind of data can be mined?

- Data mining can be applied to any kind of data as long as the data are meaningful.
- The most basic forms for data mining applications are:
  - Database data
  - Data warehouse data
  - Transactional data

Cont’d

- Data mining can also be applied to other forms of data such as:
  - Data streams
  - Ordered/sequence data
  - data graph
  - spatial data
  - text data
  - Multimedia data

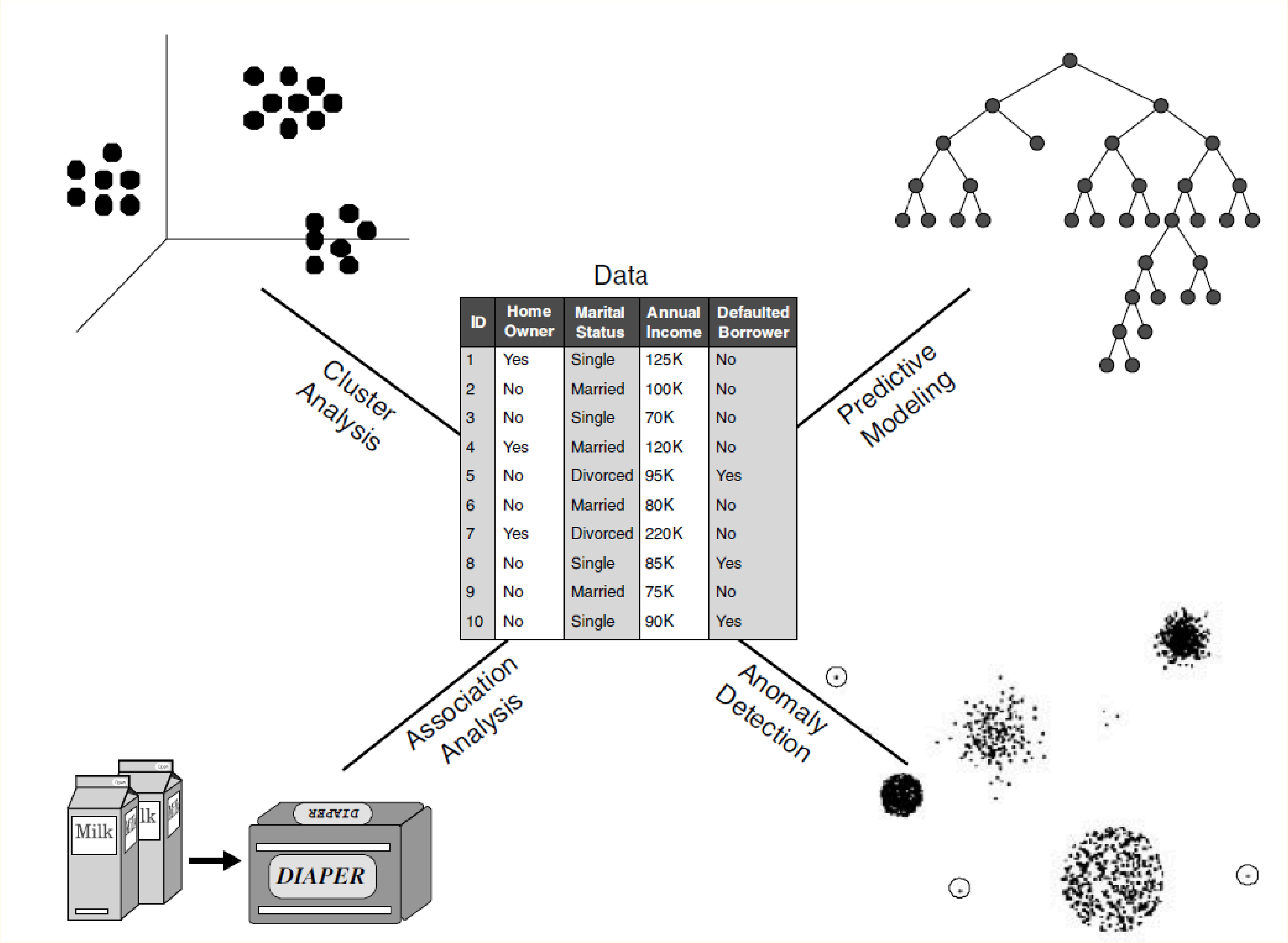
Data Mining Tasks

- Data mining are divided into two major categories.

Predictive Tasks: The objective is to predict the value of a particular attribute based on the value of other attributes

Descriptive Tasks: Here the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize underlying relationships in data.

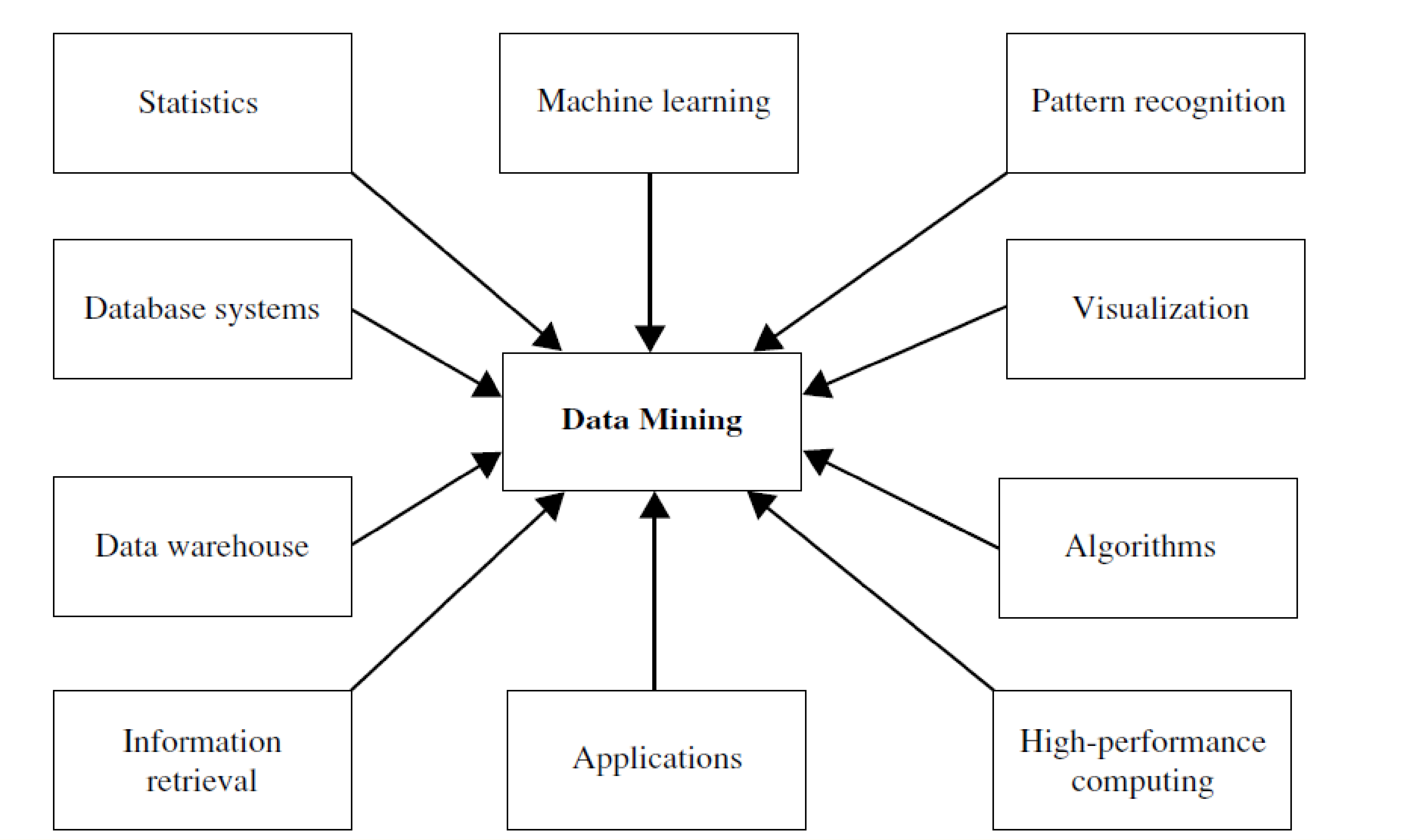
Cont’d



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Four core data mining tasks

Technologies Used



Technologies used in data mining

Application of Data Mining

Business Intelligence (BI) technologies provide historical, current and predictive views of business operations. Examples include: reporting, predictive analytics, benchmarking etc

Web Search Engines is a specialized computer server that searches for information on the web. The search results of a user query are often returned as a list.

Take Home Exercise

Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

References

1. S. Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, San Francisco, CA, 2003.
2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining. Pearson Education Limited 2014. ISBN 10: 1-292-02615-4
3. Jiawei Han, Micheline Kamber, Jian Pei: Data Mining Concepts and Techniques 2015, Morgan Kaufmann Publishers