Data Mining

Introduction

Mahesh Kumar
[maheshkumar@andc.du.ac.in]

Course Web Page
[www.mkbhandari.com/mkwiki]

Outline

1. What Is Data Mining?
2. Challenges
3. Data Mining Origins
4. Data Mining Tasks

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 and Industry: Web, e-commerce, transactions, stocks,
    - → Science and Engineering: Remote sensing, bioinformatics, scientific simulation
    - → Society and everyone: News, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
  - The amount of data (volume), its complexity (variety), and the rate at which it is being collected and processed(velocity), have simply become too challenging for humans to analyze unaided.
- Thus, there is a great need for automated tools for extracting useful information from big data despite the challenges posed by its enormity and diversity.

Data Mining: Confluence of Multiple Disciplines

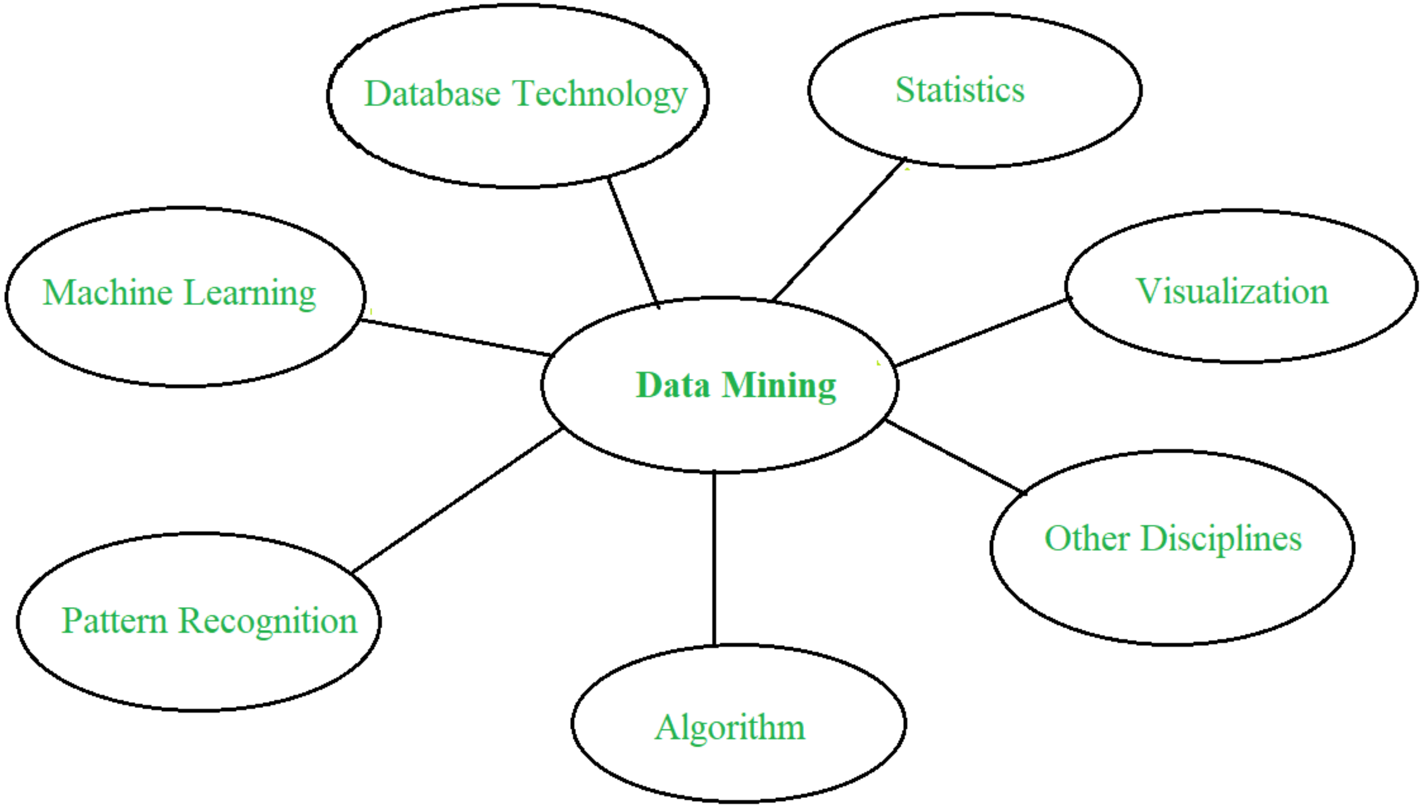


Figure 1: Data Mining : Confluence of Multiple Disciplines [3]

What is Data Mining?

- Data Mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial</u>, <u>implicit</u>, <u>previously unknown</u> and <u>potentially useful</u>) patterns or knowledge from huge amount of data.

- *<u>Alternative names</u>*
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Not all information discovery tasks are considered to be data mining
  - Looking up individual records using a database management system (query processing)
  - Finding particular Web pages via a query to an Internet search engine (web browsing)
  - (Deductive) expert systems

Data Mining: On What Kinds of Data?

- Database-oriented data sets and application
  - Relational database, data warehouse, transactional database.

- Advanced data sets and advanced application
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (including bio-sequence)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimeda database
  - Text databases
  - The world-wide web

Data Mining and Knowledge Discovery

- Data Mining is an integral part of **knowledge discovery in databases (KDD)**, which is the overall process of converting raw data into useful information

- Consists of a series of transformation steps, from data preprocessing to postprocessing of data mining results.

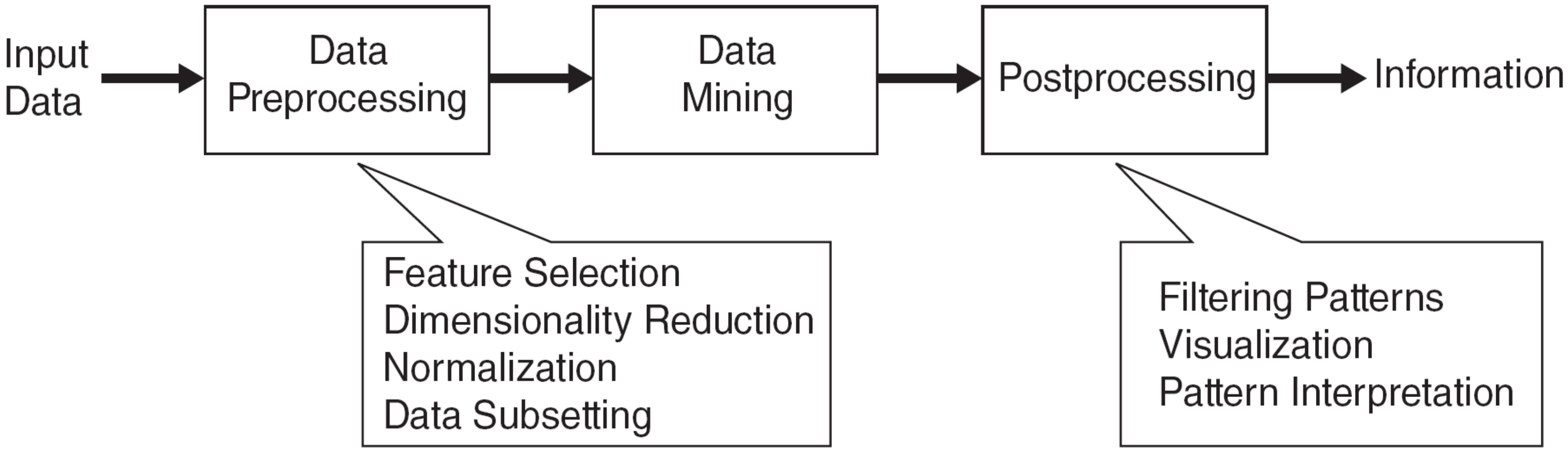


Figure 1. The process of knowledge discovery in databases (KDD). **[1]**

Data Mining and Knowledge Discovery

- *The input data*
  - stored in a variety of formats (flat files, spreadsheets, or relational tables)
  - may reside in a centralized data repository or be distributed across multiple sites.

- *Preprocessing*
  - Transform the raw input data into an appropriate format for subsequent analysis.
  - The steps involved includes
    - → Integrating data from multiple sources
    - → cleaning data to remove noise and duplicate observations
    - → selecting records and features that are relevant to the data mining task at hand
  - Since, in many ways data can be collected and stored, data preprocessing is perhaps the most laborious and time-consuming step in the overall knowledge discovery process.

Data Mining and Knowledge Discovery

- *Postprocessing*
  - *“Closing the loop”*
    - → the process of integrating data mining results into decision support systems.
    - → ***For example***, in business applications, the insights offered by data mining results can be integrated with campaign management tools so that effective marketing promotions can be conducted and tested.
  - Ensures that only valid and useful results are incorporated into the decision support system.
  - Example, visualization, which allows analysts to explore the data and the data mining results from a variety of viewpoints.
  - Statistical measures or hypothesis testing methods can also be applied during postprocessing to eliminate spurious data mining results.

Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

The Origins of Data Mining

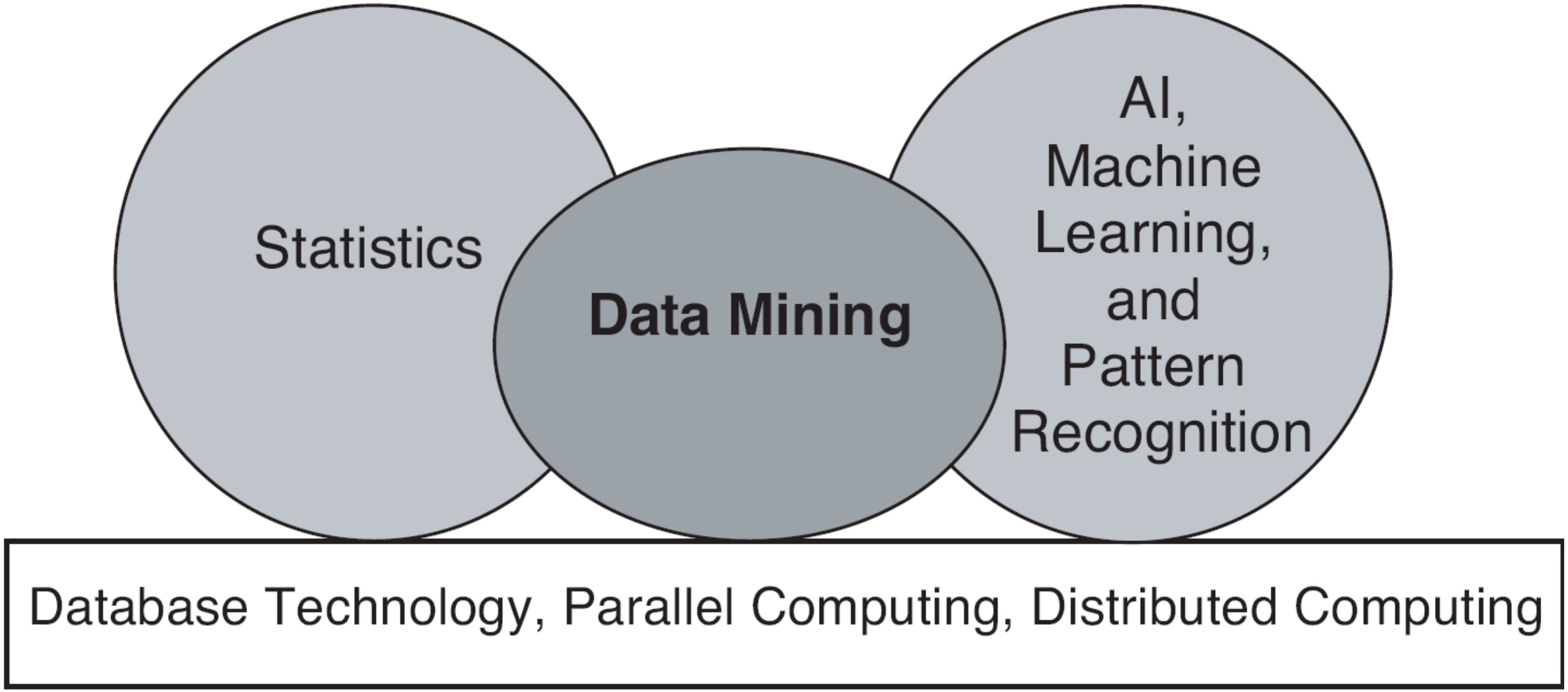

Statistics
Data Mining
AI,
Machine
Learning,
and
Pattern
Recognition
Database Technology, Parallel Computing, Distributed Computing

Figure 2: Data mining as a confluence of many disciplines. **[1]**

The Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High Dimensional
  - Heterogeneous
  - Complex
  - Distributed

- A key component of the emerging field of data science and data-driven discovery

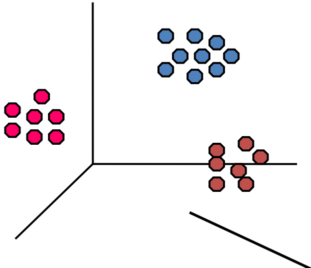
Data Mining Tasks

- Data mining tasks are generally divided into **two** major categories:

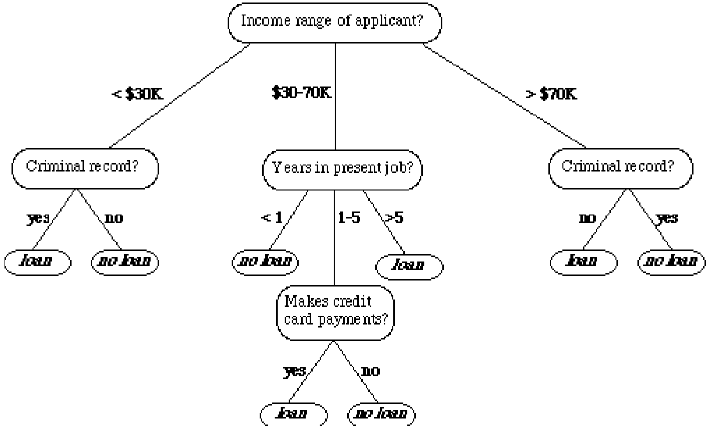
1. **Predictive Task**
   - → To predict the value of a particular attribute based on the values of other attributes.
   - → Attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

2. **Descriptive Task**
   - → To derive human-interpretable patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data.
   - → Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Data Mining Tasks



Clustering



Predictive Modeling

Tid	Refund	Marital Status	Taxable Income	Cheat
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	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules

Anomaly Detection

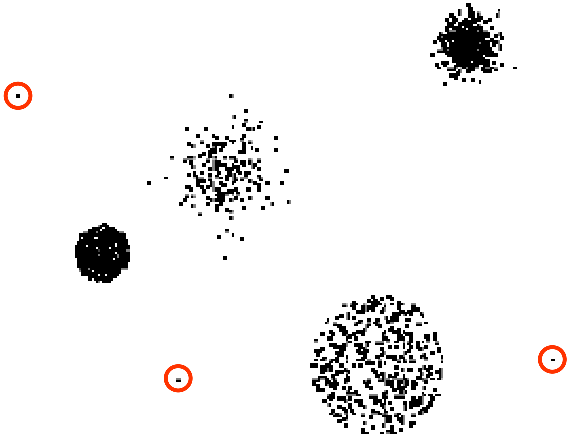


Figure 3: Four of the core data mining tasks.**[1]**

Predictive Modeling

- **Predictive Modeling** refers to the task of building a model for the target variable as a function of the explanatory variables.

- Two types of Predictive Modeling tasks:

  1. **Classification,** used for discrete target variables

     → *For example*, predicting whether a Web user will make a purchase at an online bookstore is a classification task because the target variable is binary-valued.

  2. **Regression,** used for continuous target variables

     → *For example*, forecasting the future price of a stock is a regression task because price is a continuous-valued attribute.

- The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.

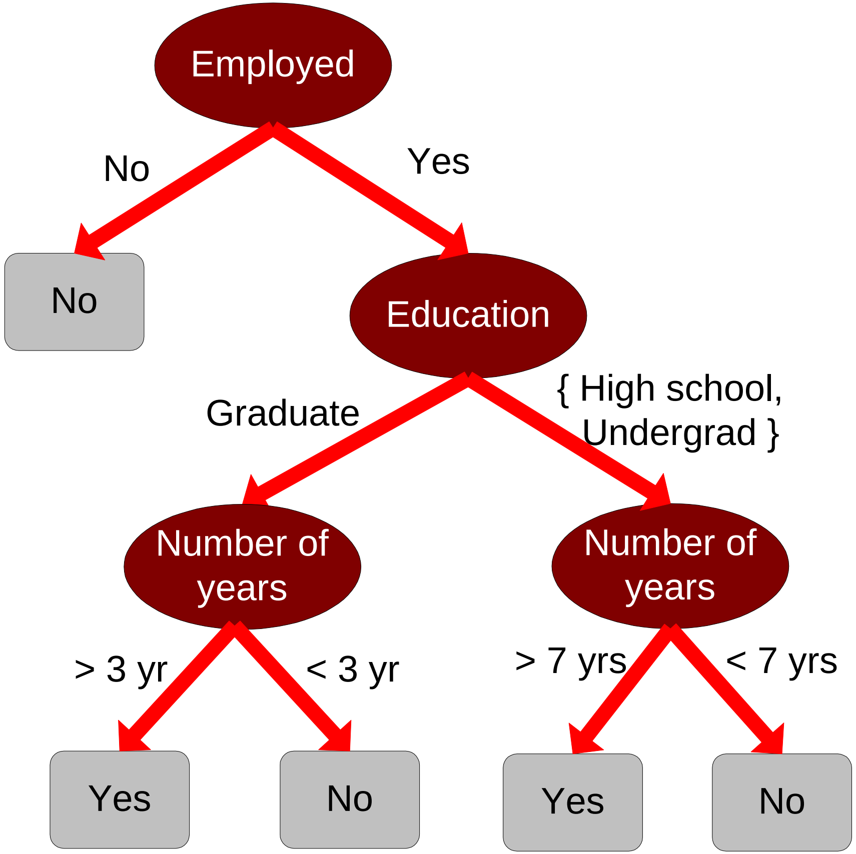
Predictive Modeling - Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness



Example of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc

- Identifying intruders in the cyberspace

Association Analysis

- Given a set of records each of which contain some number of items from a given collection
    - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

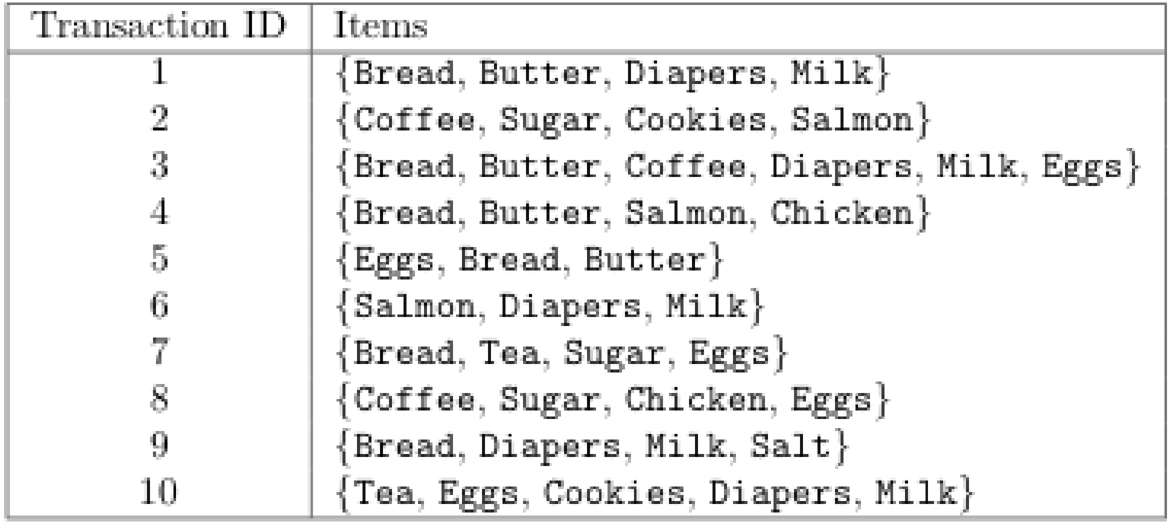
Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Analysis

Produce dependency rules based on following table which will predict occurrence of an item based on occurrences of other items.



Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

Table 1: Market basket data. **[1]**

Association Analysis - Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management

- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period

- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

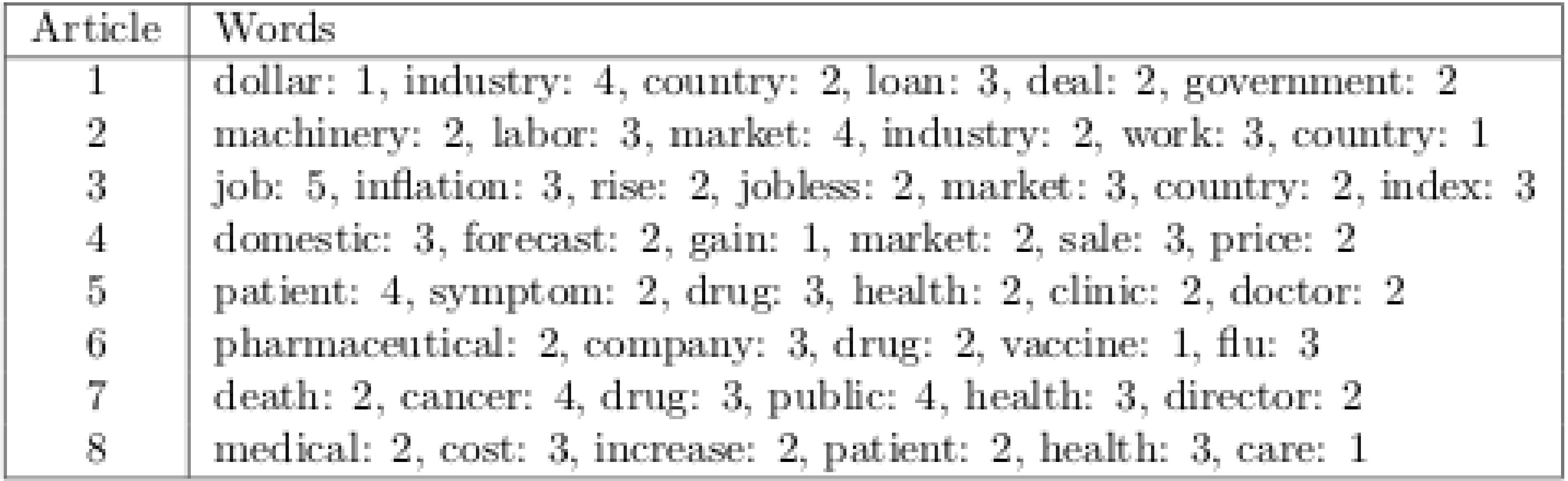
Cluster Analysis

- **Cluster Analysis** seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters.

  - Clustering can be used to group sets of related customers, find areas of the ocean that have a significant impact on the Earth’s climate, and compress data.

  - Example (*Document Clustering*)**,** The collection of news articles can be grouped based on their respective topics. Each article is represented as a set of word-frequency pairs (w, c), where w is a word and c is the number of times the word appears in the article.

Cluster Analysis

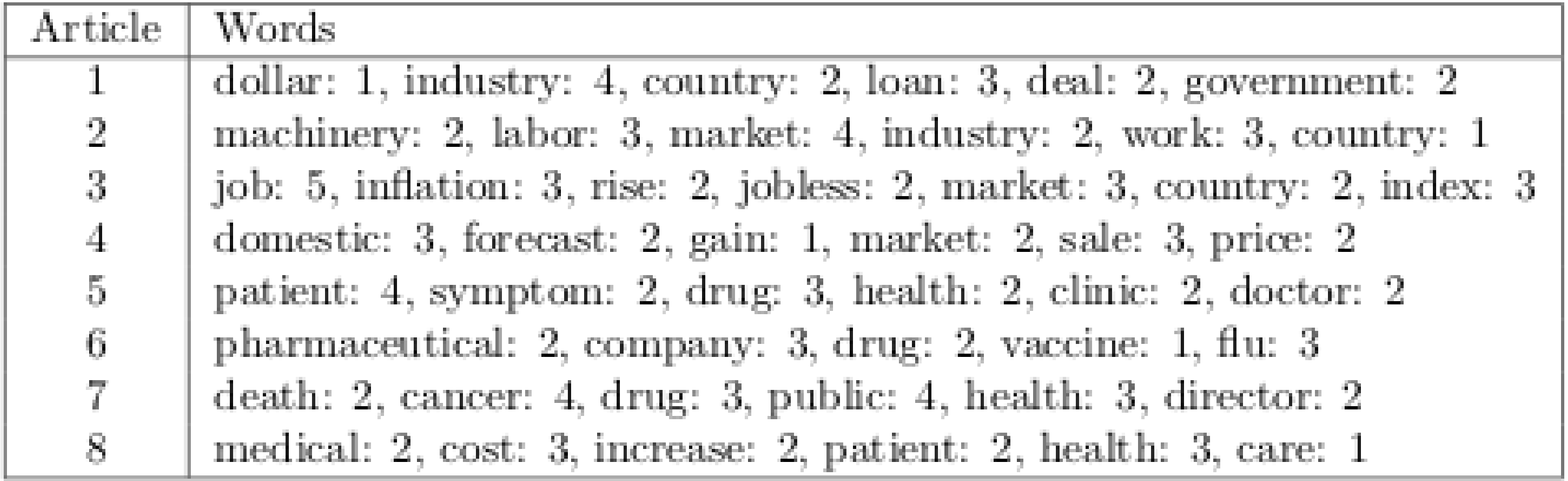


Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

Table 2: Collection of news articles **[1]**

- There are two clusters in the dataset

Cluster Analysis



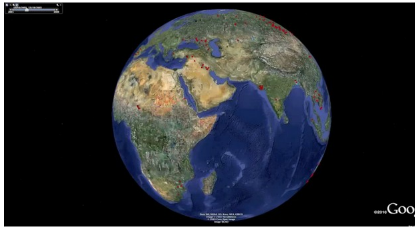
Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

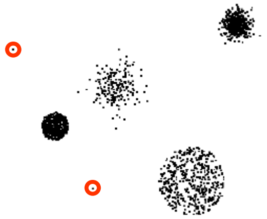
Table 2: Collection of news articles **[1]**

- There are two clusters in the dataset
  - → *News about Economy*, first four articles.
  - → *News about Helth Care*, last four articles.

Deviation/Anomaly/Change Detection

- Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of the data.
  - Such observations are known as anomalies or outliers.
  - The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous.
  - ***Applications***





    → Credit Card Fraud Detection

    → Network Intrusion Detection

    → Identify anomalous behavior from sensor networks for monitoring and surveillance

    → Detecting changes in the global forest cover.

References

1. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education.

2. Data Mining, IIT Kharagpur, Prof. Pabitra Mitra, NPTEL.

3. https://www.geeksforgeeks.org/data-mining-process/