Data Mining

Concepts & Algorithms

DSA 8102: Data mining, data storage and retrieval

Strathmore
UNIVERSITY

D. Owuor (Ph.D.)

Objectives

- Determine why data mining is in high demand.
- Describe the concept of data mining and the technologies associated with it.
- Identify which kinds of data may be mined,
- Describe the data mining algorithms and areas of applications.
- Discuss significant issues in data mining.

Strathmore
UNIVERSITY

Data Mining

Strathmore
UNIVERSITY

Why data mining?

The digital age

- Recent advances in data collection and storage technology have led to accumulation of all sorts of data.
- Unfortunately, these accumulation of data in disparate structures became overwhelming.
- The initial chaos led to the creation of unstructured and structured databases (i.e., RDBMS).
- The rapid growth of different forms of data led to the advent of **Big Data**.
- Today, we have far more information that we can handle from business transactions, scientific data, satellite pictures, text reports, military intelligence, etc.

Sources of data

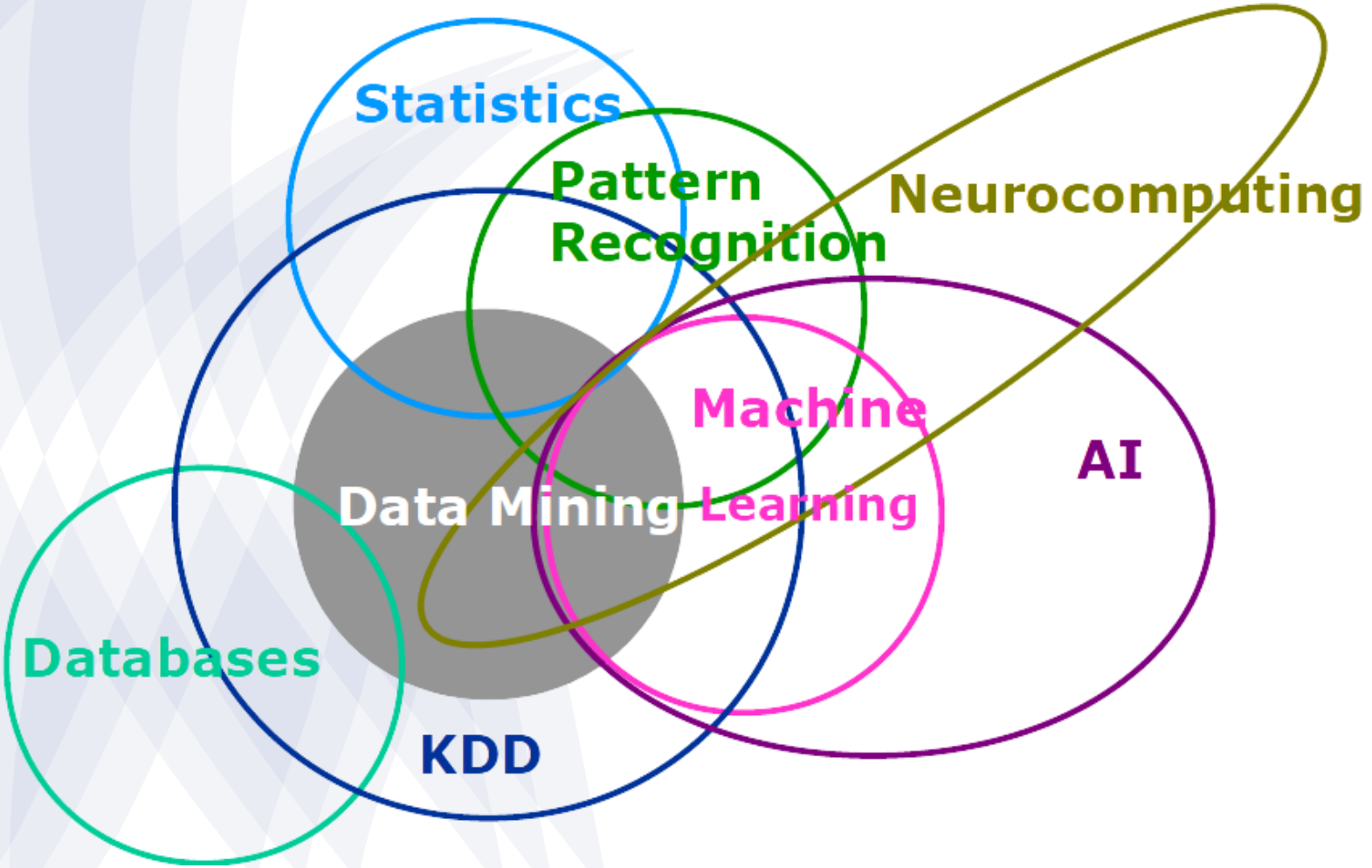
- Here is a non-exclusive list of a variety of information:
  - Business transactions
  - Scientific data
  - Medical and personal data
  - Surveillance video and pictures
  - Satellite sensing
  - Games
  - CAD and Software Engineering data
  - Text reports and memos (email messages)
  - World Wide Web repositories

Strathmore
UNIVERSITY

Data Mining Concept

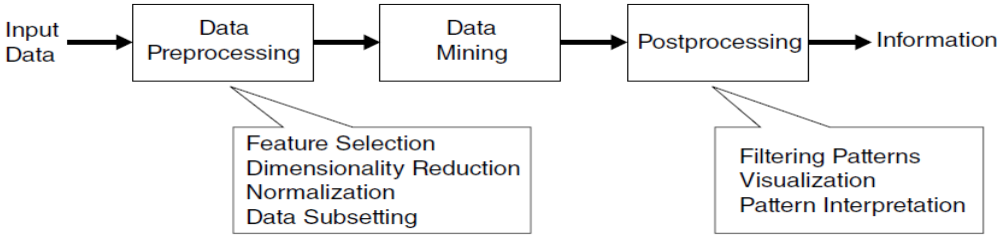
Multidisciplinary

Strathmore UNIVERSITY

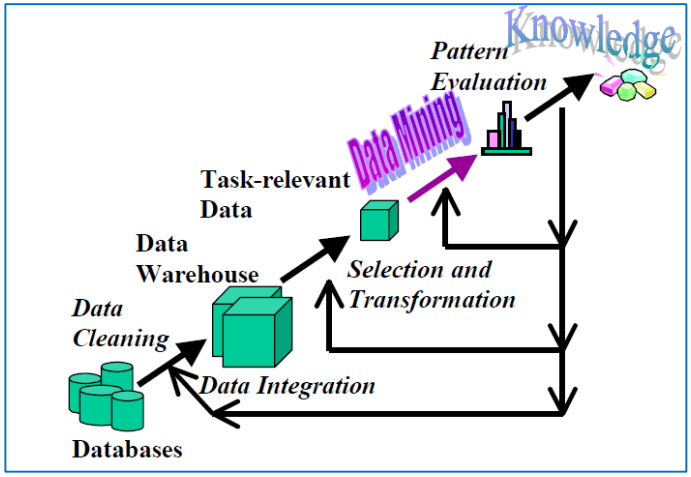


Data Mining

- Data mining is a crucial part of (KDD) knowledge discovery in databases.
- It is the process of converting raw data into useful information.
  - **Input data:** Flat files, spreadsheets or relational tables)
  - **Data preprocessing:** Cleaning data to remove noise and duplicate observations
  - **Data mining:** Techniques are applied to extract patterns potentially useful
  - **Data Analysis:** Either through visualization or pattern interpretation.
  - **Information:** Output of the knowledge discovery

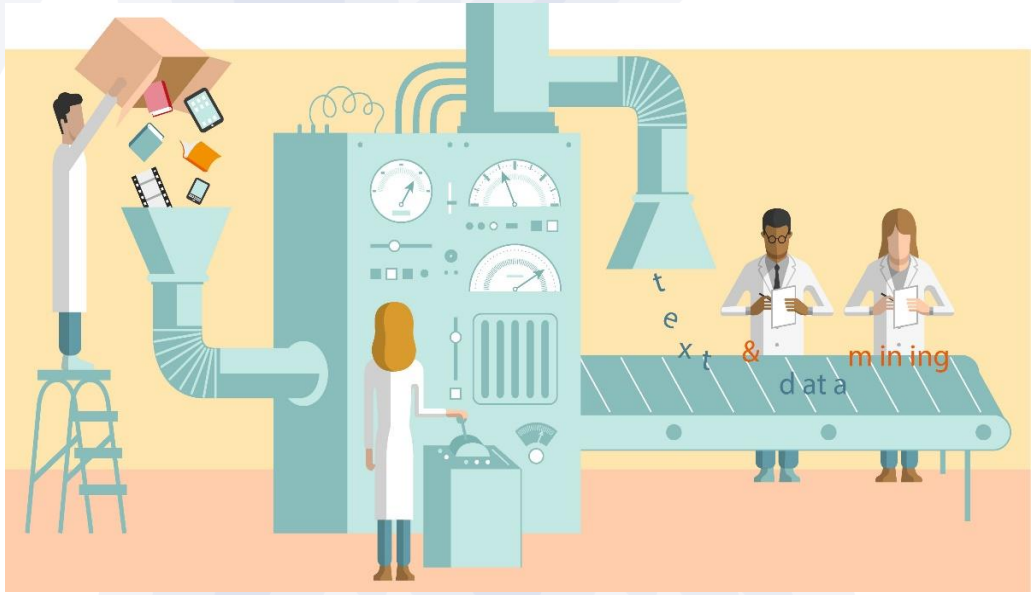


The process of knowledge discovery in databases (Tan et al. 2013)



(Zaïane, 1999)

Data for mining



- In principle data mining is not specific to one type of data.
- Examples of data sources that can be mined:
  - Flat files
  - Relational databases
  - Data warehouse
  - Transaction databases
  - Multimedia databases
  - Spatial databases
  - Time-series databases
  - World Wide Web

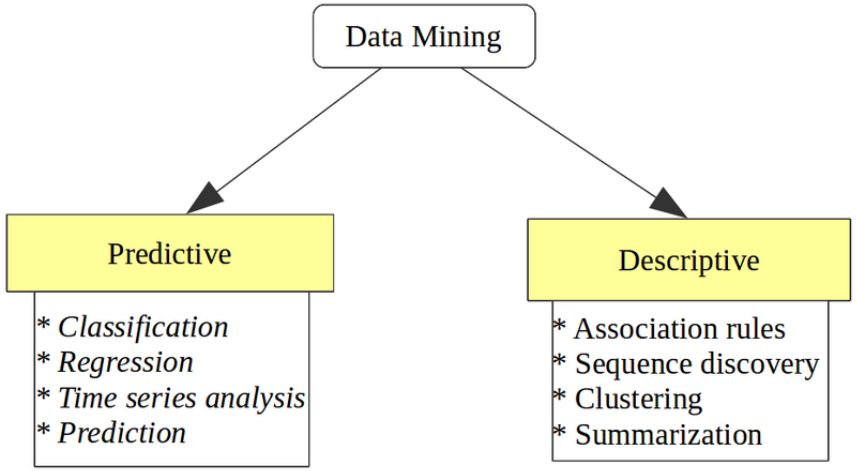
Strathmore
UNIVERSITY

Data Mining Tasks & Algorithms

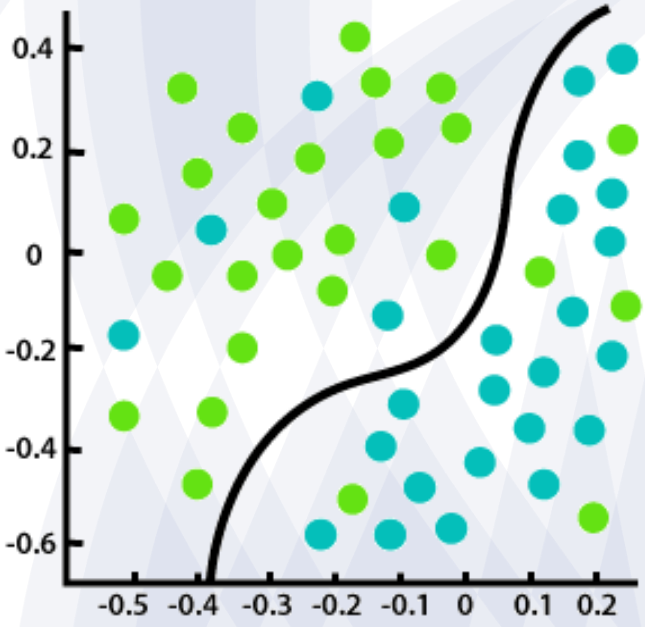
Common mining tasks

Strathmore UNIVERSITY

- **Predictive data mining:** is used to predict future outcomes.
- **Descriptive data mining:** focuses on what has happened.
- Most common data mining tasks.
  - Classification
  - Regression
  - Time series analysis
  - Association rules
  - Clustering
  - Sequence discovery
  - Summarization



1. Classification

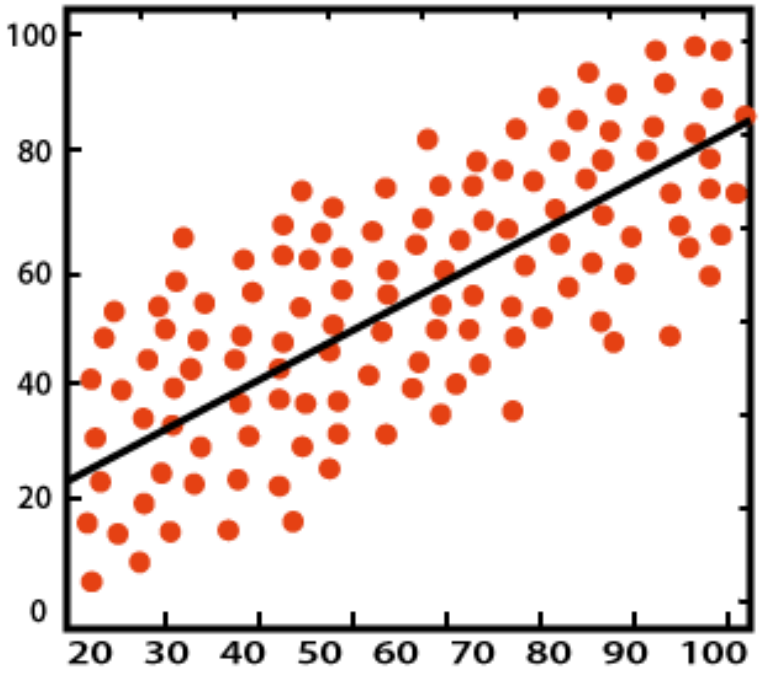


- Predict if a data point belongs to one of the predefined classes. The predictions are usually from a known data set.
- Common examples:
  - Determining whether a particular credit card transaction is fraudulent or not;
  - Assessing whether a mortgage application is a good or bad credit risk;
  - Diagnosing whether a particular disease is present or not;
- Popular algorithms:
  - Decision trees, neural networks, Bayesian models, induction rules etc.

2. Regression

Strathmore UNIVERSITY

- One of the oldest statistical technique used to predict a numeric or continuous value.
- Common examples:
  - Predicting unemployment rate for the following year.
  - Estimating insurance premium.
- Popular algorithms:
  - Linear regression and logistic regression.



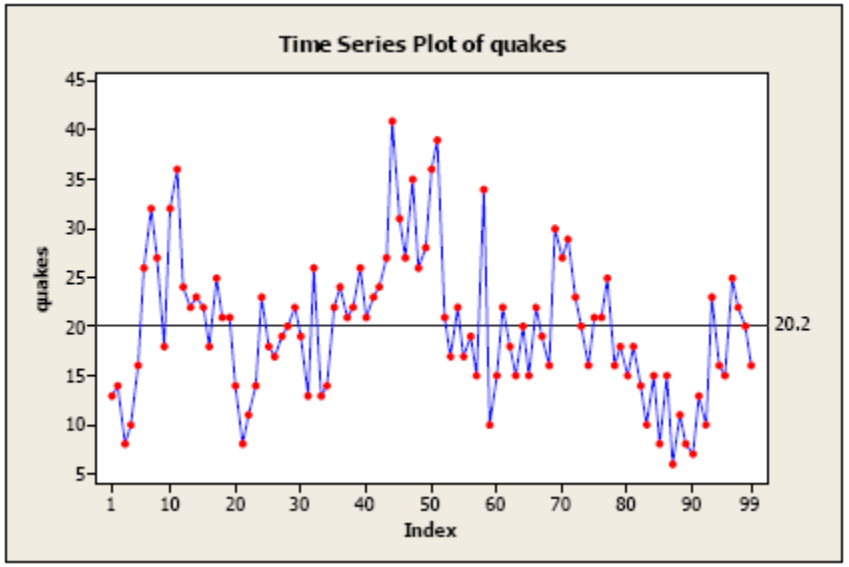
3. Clustering

- Data is partitioned into several meaningful groups. Clustering differs from classification in that there is no target variable.
- Common examples:
  - Finding customer segments in a company based on transaction, web and customer call data.
  - For gene expression clustering, where very large quantities of genes may exhibit similar behaviour.
- Popular algorithms:
  - K-means clustering, DBSCAN

4. Time Series Analysis

Strathmore UNIVERSITY

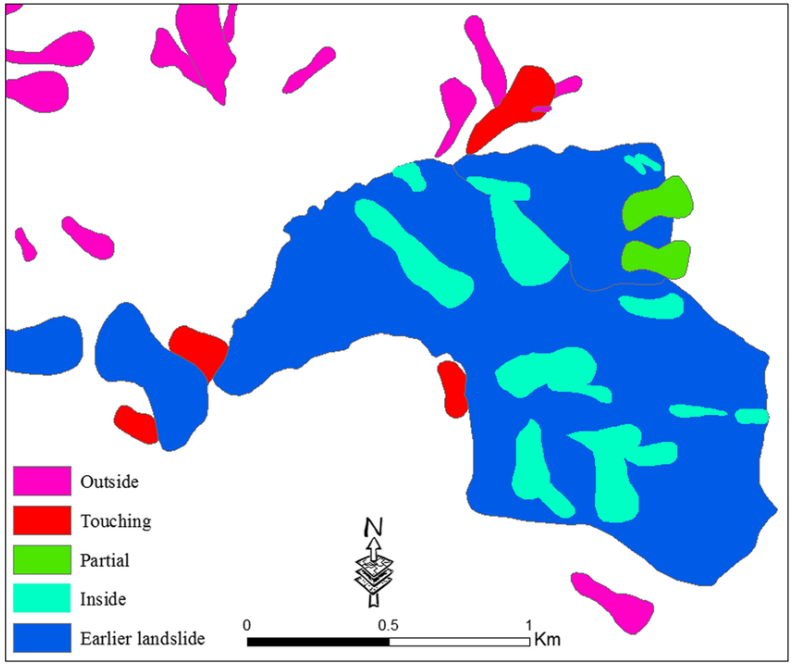
- Predict the value of the target variable for a future time frame based on historical value.
- Common examples:
  - Forecasting of sales events,
  - Tracking daily, hourly, or weekly weather data.
  - Tracking changes in application performance.
  - Tracking network logs.
- Popular algorithms:
  - Exponential smoothing, autoregressive integrated moving average (ARIMA), regression.



5. Summarization

Strathmore UNIVERSITY

- Involves techniques for finding a compact description of a dataset. Presents useful information about the data, e.g. mean, charts, graphs, etc.
- Common examples:
  - Monitor the activity of a network
  - Compare different entities.
- Popular algorithms:
  - Multivariate visualization



6. Association Rules

Strathmore UNIVERSITY

- Describes relationship between items of a data set.
- Common examples:
  - Finding out which items in a supermarket are purchased together, and which items are never purchased together,
  - Determining the proportion of cases in which a new drug will exhibit dangerous side effects.
- Popular algorithms:
  - Apriori algorithm, FP-Growth.

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Examples

{bread} ⇒ {milk}

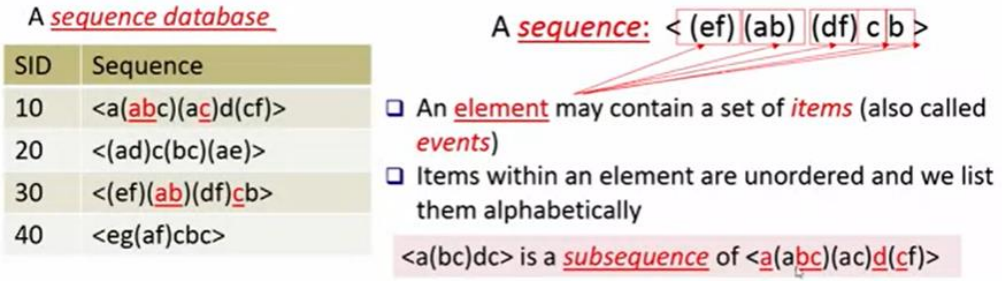
{soda} ⇒ {chips}

{bread} ⇒ {jam}

7. Sequence Discovery

Strathmore UNIVERSITY

- Discovers statistically relevant patterns in sequential data. Event occurrences are usually governed by timing constraints.
- Common examples:
  - Customer shopping sequences: First buy computer, then CD-ROM, then digital camera, within 6 months.
  - Web access patterns
  - Weather prediction.
  - Medical treatments, natural disasters (e.g. earthquakes), stocks and markets.
- Popular algorithms:
  - Apriori, FP-Growth.



Exercise - Group Discussion

Strathmore UNIVERSITY

1. Differentiate between “explainable AI” and “AI” as used in knowledge discovery.
2. Differentiate between clustering and classification.

Strathmore
UNIVERSITY

Categories, Application Areas & Issues

Strathmore
UNIVERSITY

Categories

Categories of data mining tasks

Strathmore UNIVERSITY

- Categorize according to the type of data source
  - spatial data, multimedia data, time-series data, text data, www, etc.
- Categorize according to the data model drawn on
  - relational database, object-oriented database, data warehouse, transactional, etc.
- Categorize according to the kind of knowledge discovered
  - characterization, discrimination, association, classification, clustering, etc.
- Categorize according to mining techniques used
  - machine learning, neural networks, genetic algorithms, statistics, visualization, database-oriented or data warehouse-oriented, etc.

Popular Data Mining Tools

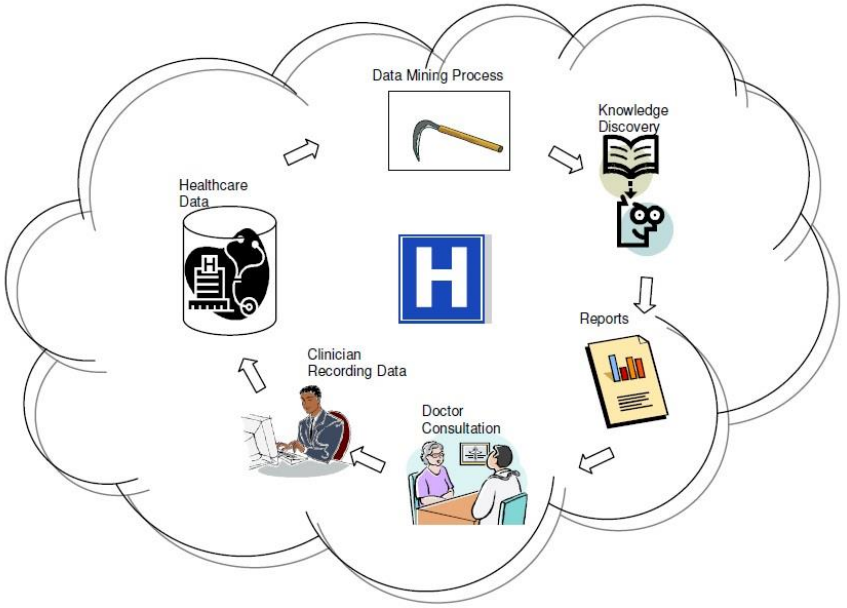
Strathmore
UNIVERSITY

Application Areas

Application Areas

Strathmore UNIVERSITY

- Healthcare:
  - reducing costs and improving patient outcomes
- Education:
  - predicting at-risk students.
- Customer relationship management:
  - know the needs of customers and build loyalty
- Manufacturing engineering:
  - forecasting the overall time for product development
- Finance and banking:
  - correlations and trends in market costs an business information
- Market basket analysis and segmentation:
  - understand the buying habits of customers
- Fraud detection:
  - identify patterns in fraudulent documents
- Etc.



Strathmore
UNIVERSITY

Data Mining Issues

Data Mining Issues

Strathmore UNIVERSITY

- Note that these issues are not exclusive:
  - Security and privacy issues
  - User interface issues
  - Mining methodology issues
  - Performance issues
  - Data sources issues

Exercise

1. Describe the issues that come with data mining.
2. Discuss possible solutions to the issues described above.

References

Strathmore
UNIVERSITY

Thank you!

Any Questions?