

# **Introduction to Data Mining**

**CI603 - Data Mining**

# Module Schedule

1. Introduction to Data Mining

2. Data

3. Association Analysis

4. Cluster Analysis

5. Classification

6. Naïve Bayes Classifier

7. Singular Value  
Decomposition

7. Principal Component Analysis

8. Support Vector Machines

9. Anomaly Detection

10. Information Retrieval

11. Exploratory Data Analysis

12. Data Mining Methodology

13. Different Approaches

# Assessment and Reading List

- **Assessment**

- ▶ 100% Coursework

- **Reading List**

- ▶ **Introduction to Data Mining** - Pang-Ning Tan, Michael Steinbach, Anju Karpatne, Vipin Kumar
- ▶ **Introduction to Information Retrieval** - Christopher D Manning, Prabhakar Raghavan, Hinric Schutze

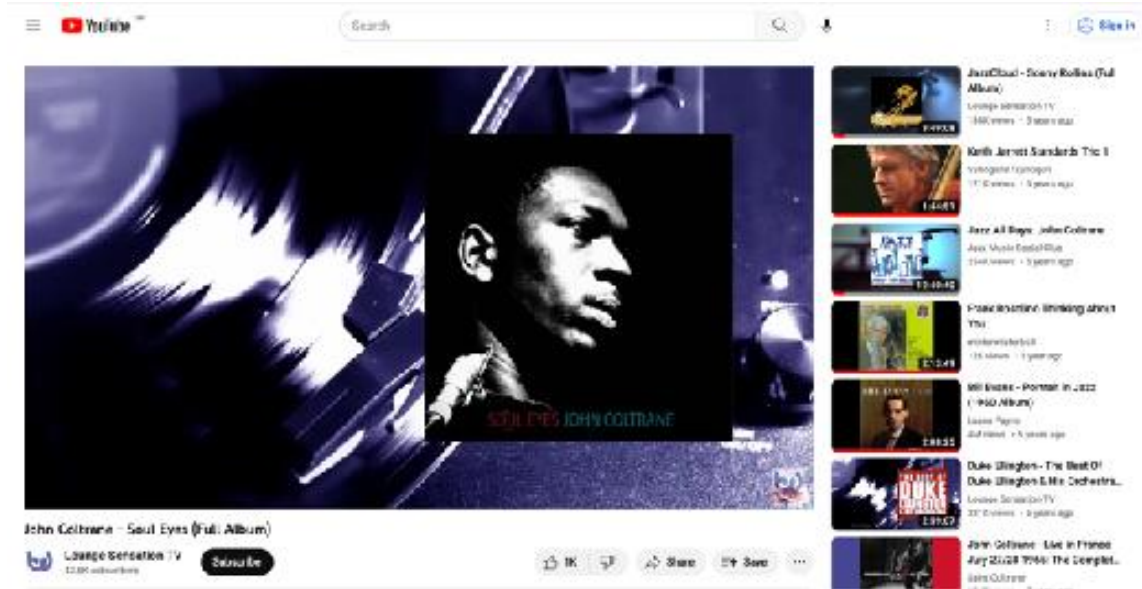
# The Data Explosion

- The term refers to the rapid growth of digital data **generated** and **stored** globally.
- There is an exponential growth of data resulting from the increasing use of technologies such as:
  - smartphones,
  - social media,
  - the Internet of things (IoT),
  - cloud computing.





# Data Explosion Contributors Examples



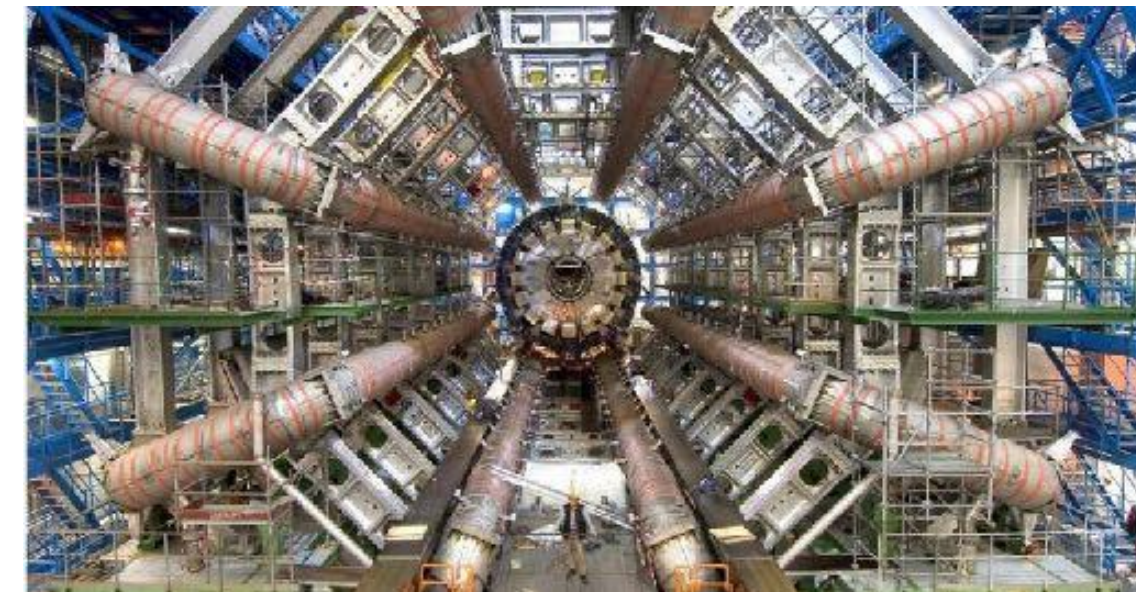
Over 400 hours of video content is uploaded to YouTube every minute



A smart city might store data from millions of sensors in order to optimise traffic flow and reduce energy consumption.



The growth of wearable devices and telemedicine is leading to an explosion in the amount of health-related data being generated.



The Large Hadron Collider generates over 30 petabytes of data per year.



# Some Internet Live Stats



5,191,330,128

Internet Users in the world

5.35 billion Jan2024  
66.2% of World population



1,924,670,896

Total number of Websites

Over 2 billion websites



191,925,915,515

Emails sent today

361 billion emails sent  
worldwide each day



3,051,886,685

Facebook active users



1,071,675,539

Google+ active users



383,926,141

Twitter active users



5,885,273,871

Google searches today

8.5 Billion google searches  
every day



5,710,054

Blog posts written today



584,387,781

Tweets sent today



422,735,272

Pinterest active users



385,915,680

Skype calls today



156,415

Websites hacked today



5,566,872,637

Videos viewed today  
on YouTube



67,267,485

Photos uploaded today  
on Instagram



120,563,499

Tumblr posts today



457,450

Computers sold today



3,042,215

Smartphones sold today



261,790

Tablets sold today

# Some Figures

- 80% of data will be unstructured by 2025.
- 90% of all data in existence today was created in the past two years.
- Worldwide data is expected to hit **175 zettabytes** by 2025.
- 90 ZB of this data will be from IoT devices.



# Units of data storage have dramatically changed

1 megabyte (Mb)	1,000 kilobytes
1 gigabyte (Gb)	1,000 megabytes
1 terabyte (Tb)	1,000 gigabytes
1 petabyte (Pb)	1,000 terabytes
1 exabyte (Eb)	1,000 petabytes
1 zettabyte (Zb)	1,000 exabytes
1 yottabyte (Yb)	1,000 zettabyte



<https://www.old-computers.com/>





# The 5 Vs of Big Data

The five characteristics of big data are: Volume, Velocity, Veracity, Variety, Value

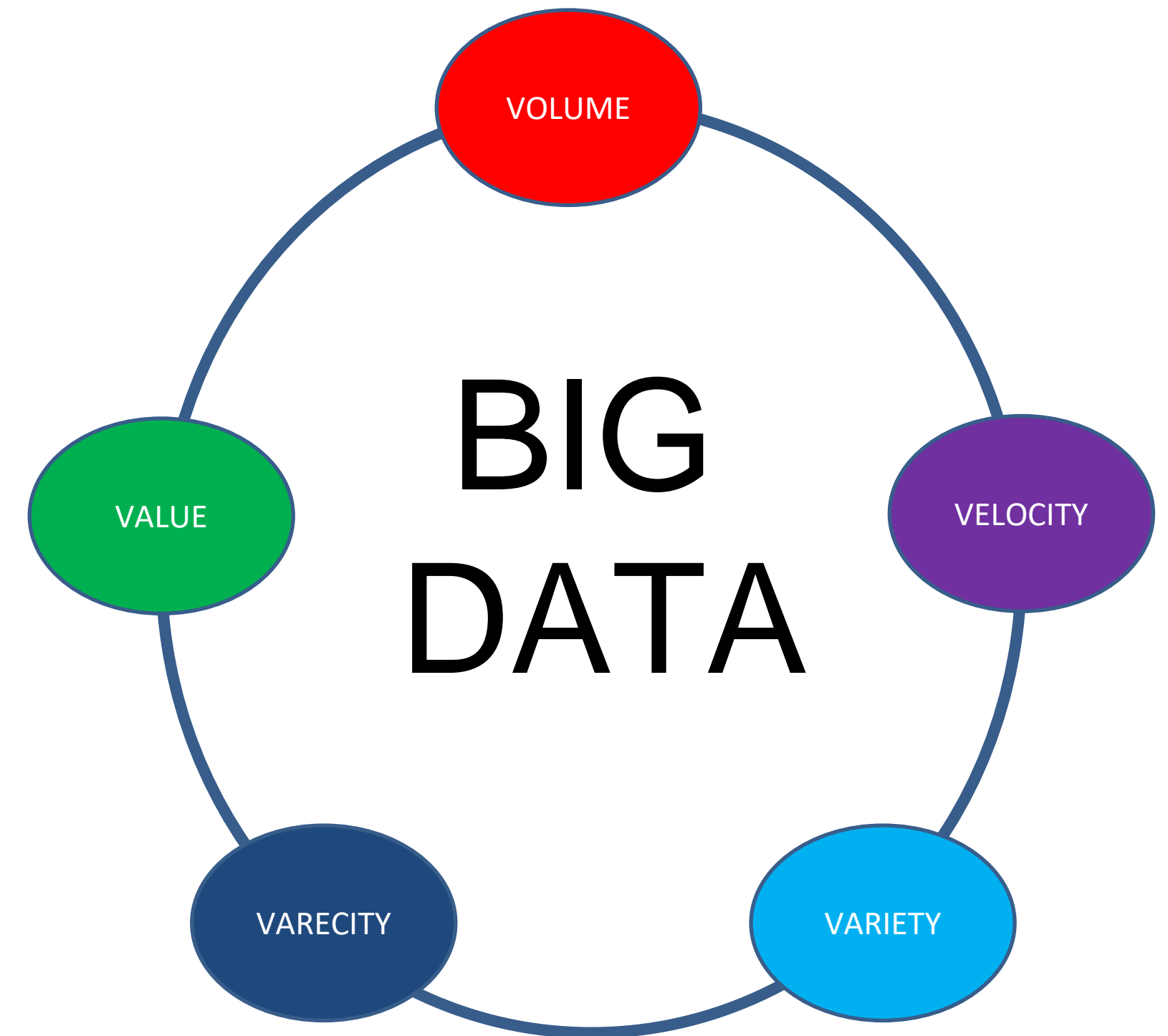
**Volume** – This refers to the size of the Big Data. The data can be referred to as Big Data or not.

**Velocity** – This refers to the speed at which data is being collected.

**Variety** - This refers to whether data is structured, semi structured or unstructured.

**Veracity**- This refers to the quality of the data, the sources where the data came from, need to check data for accuracy before using it for decision making.

**Value** – This refers to how useful the data for decision making.






# Key Drivers of Data Growth

- Main drivers:
  - ▶ Increase in **storage capacities** and **lower cost**.
  - ▶ Increase in **data processing capabilities**.
  - ▶ Increase in **data generated** and **made available**.
- **Data management** has become **extremely difficult**.
- Moreover, it has raised concerns about **data privacy** and **security**.
- But it also has led to new opportunities.



# Big Data Opportunities

- Improve decision-making.
  - Enhanced customer experiences.
  - Predictive analytics.
  - Improved operational efficiency.
  - New business opportunities.
- 



# Knowledge as a Competitive Advantage

**In God we trust;  
all others must bring data.**  
- William Edwards Deming -

The goal is to turn data into information, and  
information into insight.

—Carly Fiorina

If you torture the data long  
enough, it will confess.

Ronald Coase

The analysis of data will not  
by itself produce new ideas...

Edward De Bono

**Knowledge is of  
more value than gold.**

Solomon

**Data is the new oil.  
It's only useful when  
it's refined!**

Jess Greenwood,

**DATA IS THE NEW GOLD**

**WITHOUT DATA**

YOU'RE JUST ANOTHER PERSON  
WITH AN OPINION

W. EDWARDS DEMING

**DATA IS GREAT, BUT  
STRATEGY IS BETTER**

STEVEN SINOFKY  
PICTUREQUOTES.COM



# Data Rich but Knowledge Poor

- **Simply collecting sheer volumes of data will not deliver the insights they are craving.**
- **Traditional data analysis techniques can't cope with big data:**
  - ▶ Scalability.
  - ▶ High Dimensionality (Curse of dimensionality).
  - ▶ Heterogeneous and Complex data.
  - ▶ Ownership and Distribution.
  - ▶





# What is Data Mining?

- Data mining is the **process** of automatically **discovering useful information** in **large data repositories**.
- **Data mining techniques** are deployed to **interrogate large data sets** in order to:
  - **find novel and useful patterns** that might otherwise remain unknown.
  - **predict the outcomes** of future observations.



**Not all information discovery tasks** are considered to be data mining

# Information Discovery Examples:

- Entering a query into a search engine and getting a list of results (**keyword search**).
- Generating a concise summary of a text document or set of documents (**text summarisation**).
- Identifying named entities in a text, such as people, organisations, and locations (***named entity recognition***).
- Determining the sentiment expressed in a piece of text, such as positive, negative, or neutral (***sentiment analysis***).



# Information Discovery Examples:

- ~~Entering a query into a search engine and getting a list of results (**keyword search**).~~
- ~~Generating a concise summary of a text document or set of documents (**text summarisation**).~~
- ~~Identifying named entities in a text, such as people, organisations, and locations (***named entity recognition***).~~
- ~~Determining the sentiment expressed in a piece of text, such as positive, negative, or neutral (***sentiment analysis***).~~

**None of them** can be considered **data mining** because it involves more advanced techniques to **uncover patterns and relationships** in **large datasets**.

# Data is very large – Large-scale Data is Everywhere

There has been enormous data growth in both **commercial** and **scientific** databases due to advances in data generation and collection technologies

## *New mantra*

Gather whatever data you can whenever and wherever possible.

## *Expectations*

Gathered data will have value either for the purpose collected or for a purpose not envisioned.

*Cyber Security*

*Transaction Data*

*Social Networking: Twitter*

*Sensor Networks*

fMRI Data from Brain

Sky Survey Data

Gene Expression Data

*Computational Simulations*

Surface Temperature of Earth

Example Domains



# Data is very **complex**

- Multiple **types** of data: tables, time series, images, graphs, etc
- **Spatial** and **temporal** aspects
- **Interconnected** data of different types:  
Data collected from mobile phones - location of the user, contact information, attending venues, opinions through twitter, images through cameras, queries to search engines

# Example: transaction data

Billions of real-life customers:

Supermarkets: 20M transactions per day

Credit card companies: billions of transactions per day.

The supermarket loyalty cards allow companies to collect information about specific users

# Example: document data

Web as a document repository: estimated 50 billion of web pages

Wikipedia: over 4 million articles

Online news portals: steady stream of 100's of new articles every day

Twitter: ~300 million tweets every day



# Behavioural data

Mobile phones today record a large amount of information about the user behavior

- GPS records position

- Connections to contacts

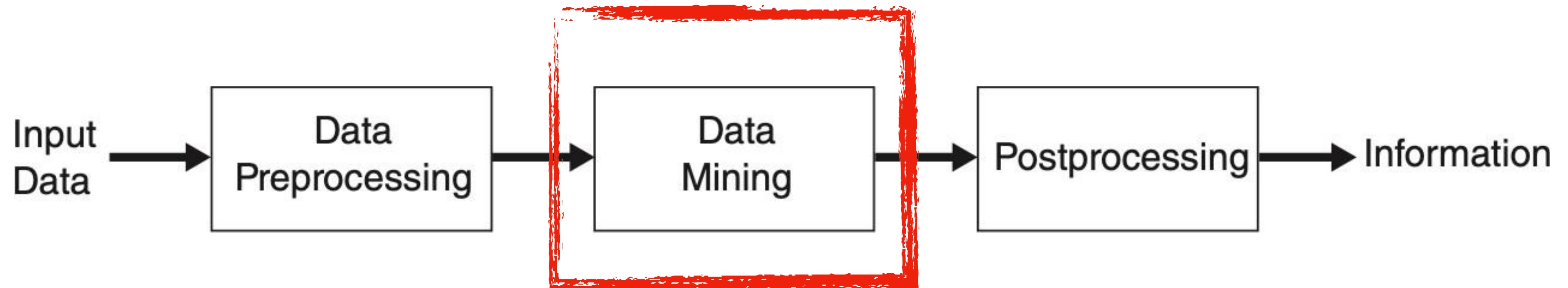
- Chosen venues to attend

Amazon collects customer browsing and purchase activity along with reviews and ratings assigned to items purchased.

Search engines can record browsing activity, queries entered, pages returned and user clicks.

# Knowledge Discovery in Databases

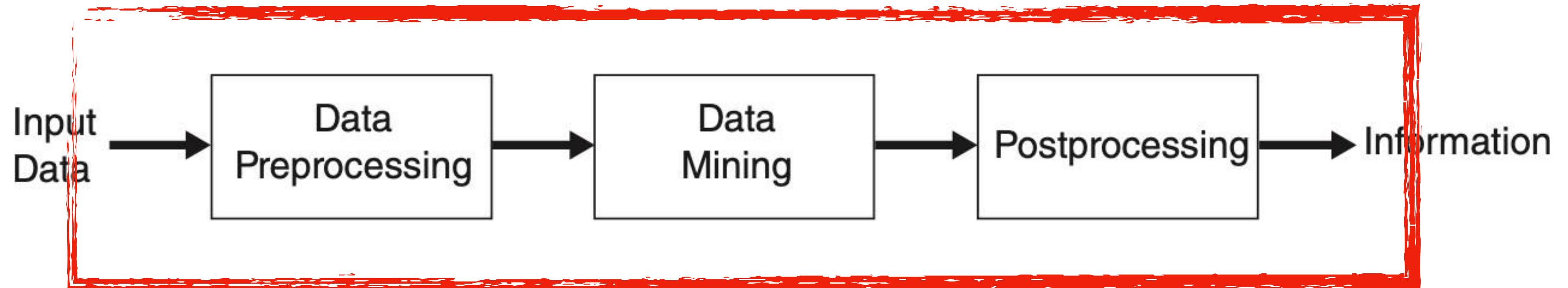
- Knowledge Discovery has been defined as the '**non-trivial extraction of implicit, previously unknown and potentially useful information from data**'
- Data mining has traditionally been viewed as an **intermediate process** within the *KDD framework*.



*Knowledge Discovery in Databases Framework*

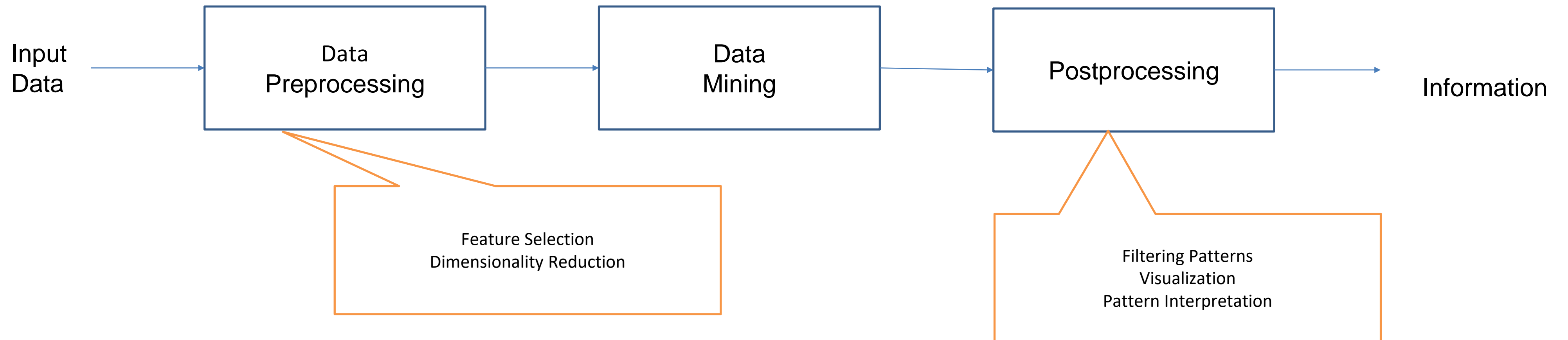
# Origin of Data Mining

- It has emerged over the years as an academic field within computer science, focusing on all aspects of KDD, including **data preprocessing**, **mining**, and **postprocessing**.



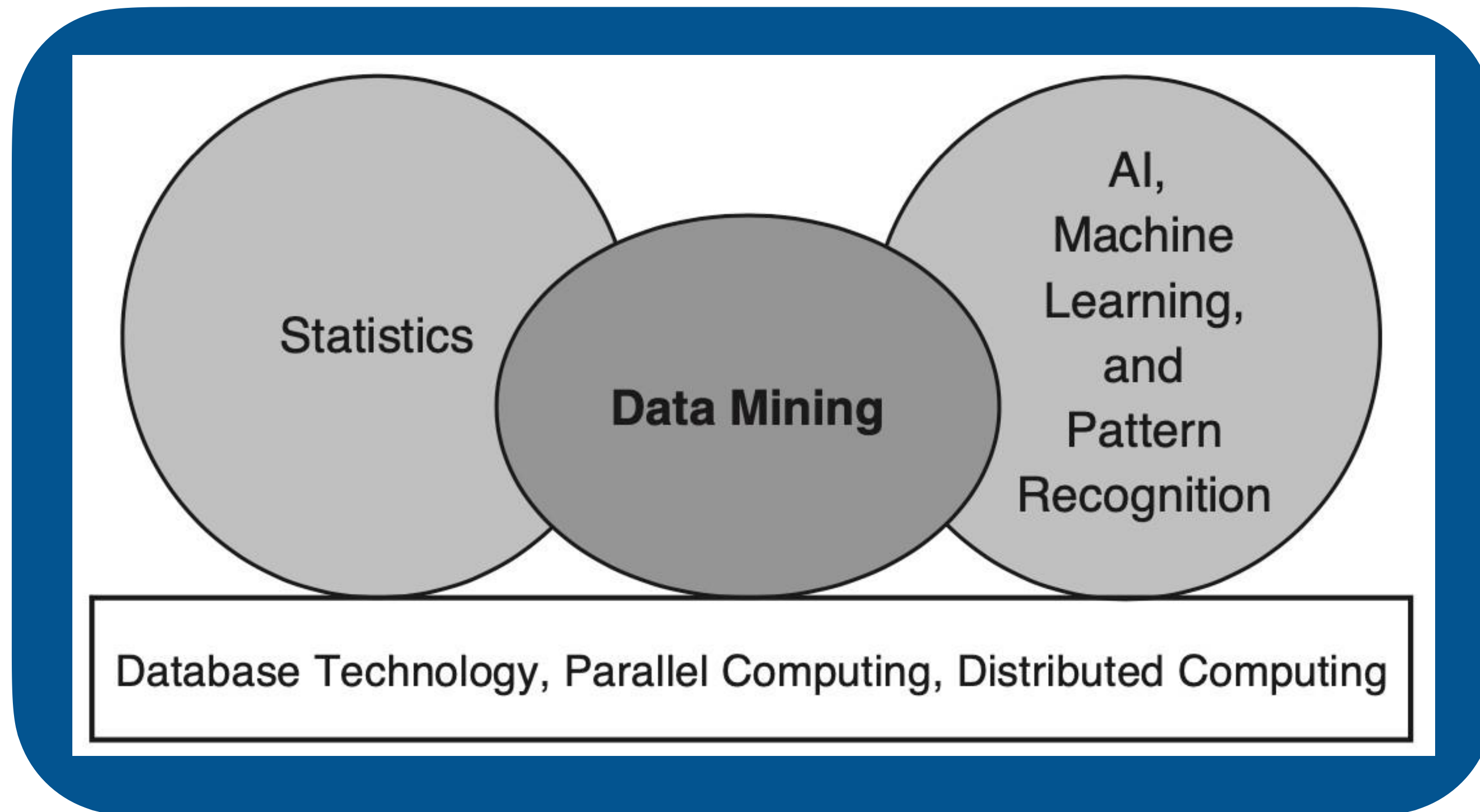


# How does data mining work?



# Origin of Data Mining

- Data Mining is the **combination of several disciplines.**



# Machine Learning

- “**Machine Learning** is the field of study that gives computers the ability to learn without being **explicitly** programmed.” (Arthur Samuel, 1959)
- Machine learning, is a **subfield of AI** that focuses on the development of **algorithms** and **statistical models** that can **learn from and make predictions** or **decisions** based on **data**.
- It includes various techniques such as: **supervised learning**, **unsupervised learning**, **reinforcement learning**, and others that can be used for various applications such as **image classification**, **speech recognition**, and **natural language processing**.

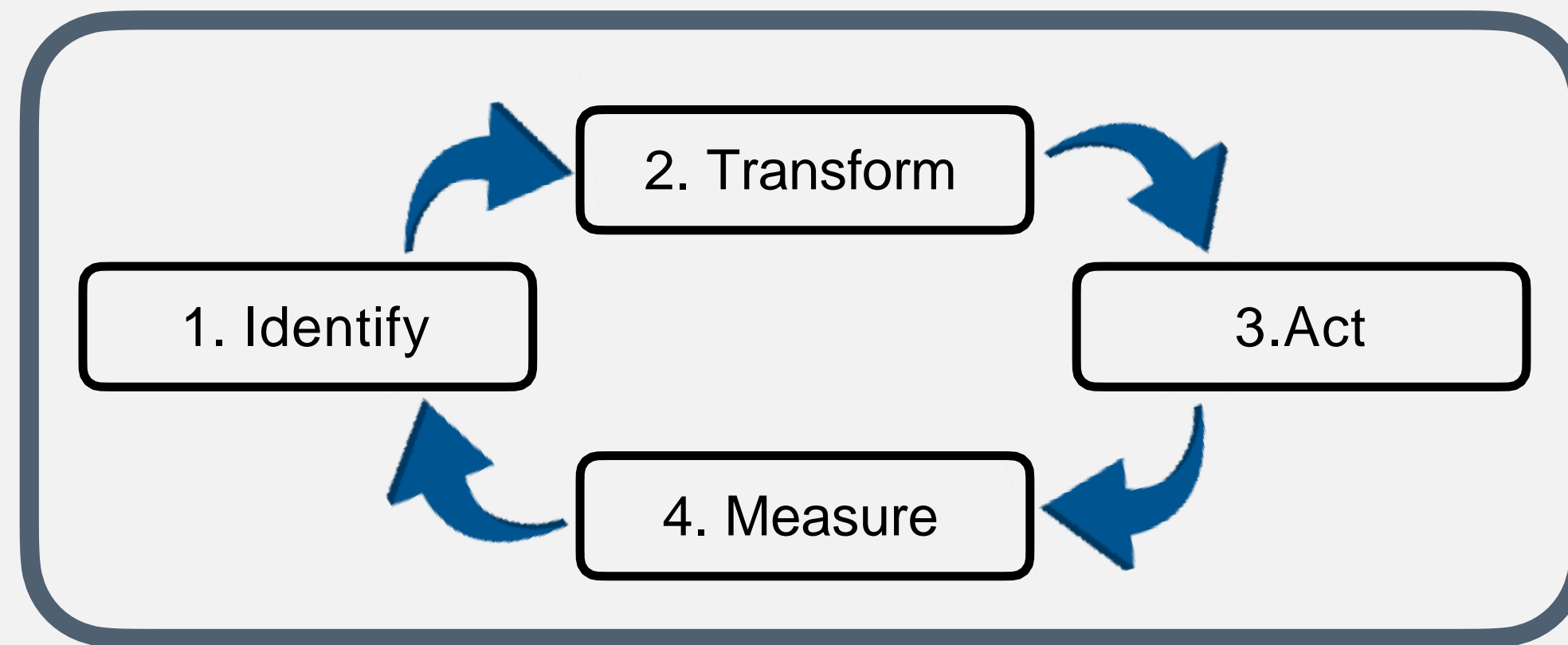


# Data Mining vs Machine Learning

- Both concepts are related but distinct fields of study.
  - **Data mining** provides the input data for machine learning algorithms to learn from.
  - **Machine learning algorithms** provide the methods for analysing and making predictions based on the data mined from large datasets.
- In practice, many data mining techniques (i.e. *clustering* and *decision tree induction*), are also used in machine learning, and many machine learning models (i.e. *neural networks* and *support vector machines*), can also be used for data mining.
- So while there are some differences between the two fields, the lines between them are often blurred.

# The Virtuous Cycle of Data Mining

- Data mining itself is the process of finding useful patterns and rules in large volumes of data.
- To be successful, data mining must **become an integral part of a larger business process**, the *virtuous cycle of data mining*.



*The virtual cycle of data mining*



# Identify Business Opportunities

- Data mining starts by **identifying the right business opportunities**.
  - ▶ **Interviewing business experts.**
  - ▶ **Willingness to act** on the results.
  - ▶ The impact of whatever actions are taken **must be measurable**.



# Transform Data into Information

- **Success** is about **making business sense of the data**, not using particular algorithms or tools.
- Identifying the **right data sources** and **bringing them together** are critical success factors.
- What can be done with available data?

# Act on the Information

- The results of data mining **must feed into business processes.**
- Data mining **makes business decisions more informed.**
- It may be applied to a **particular activity** that would have been done anyway—but with more (or less) confidence that the action will work.
- But it also could be incorporated into another systems.





# Measure the Results

- Modelling efforts should be measured.
- How can results be measured?
- Every data mining effort has lessons that can be applied to future efforts. The question is:
  - ▶ **what to measure** and
  - ▶ **how to approach the measurement so it provides the best input for future use.**





# A Learning Process, but...

- Data mining is a way of **learning from the past** in order to **make better decisions in the future**.
  - ▶ Learning things that **aren't true**.
    - Patterns may not represent any underlying rule.
    - The data set may not reflect the relevant population
    - Data may be at the wrong level of detail
  - ▶ Learning things that are **true**, but **not useful**.
    - Learning things that are already known (or should be known)
    - Learning things that can't be used.

# Undirected data mining

Unlabelled data

Unsupervised learning

# Directed data mining

Labelled data

Supervised learning



# Directed data mining

- Focuses **on one or more variables** that are **targets**, and the historical data contains examples of all the target values.
- Directed data mining **does not look for just any pattern in the data**, but for **patterns that explain the target values**.



*Task driven*





# Undirected data mining

- In undirected data mining, the **goal is to find overall patterns.**
- After patterns have been detected, it is the **responsibility of a person to interpret them and decide whether they are useful.**



*Data driven*





# Directed vs Undirected

- e.g. *Fraud Detection*
  - ▶ A **directed approach** would search for new records that are similar to cases known to be fraudulent.
  - ▶ An **undirected approach** would look for new records that are unusual.





# Data Mining Tasks

- Data mining tasks are **technical activities** that can be described independently of any particular business goal.
- If a business goal is well-suited to data mining, it can usually be phrased in terms of the following tasks:
  - ▶ **Preparing data for mining**
  - ▶ **Exploratory data analysis**
  - ▶ **Binary response modelling (also called binary classification)**
  - ▶ **Classification of discrete values and predictions**
  - ▶ **Estimation of numeric values**
  - ▶ **Finding clusters and associations**
  - ▶ **Applying a model to new data**



# Preparing Data for Mining

- **Data preparation** is usually the **most time-consuming part** of a data mining project.
- The amount of effort required **depends on the nature of the data sources** and the **requirements of particular data mining techniques**.
- Some data preparation is required to **fix problems with the source data**, but much of it is to **enhance the information content** of the data. **Better data means better models**.



# Exploratory Data Analysis

- **Exploratory data analysis** is a **statistical approach** that aims at **discovering** and **summarising** a dataset.
- An exploratory data analysis may be a **report** or a **collection of graphs** that describe something of interest.
- Data exploration typically explore the **structure of the dataset**, the **variables** and **their relationships**.
- It is useful for **profiling**.

## Central Tendencies

- Mean
- Median
- mode

## Distribution

- Variance
- Frequency
- Quantiles

## Relationship between attributes

- Correlation



# Classification

- It is one of the **most common data mining tasks**, seems to be a human imperative.
- Classification consists of **assigning a newly presented object to one of a set of predefined classes**.
- The classification task is characterised by a **well-defined definition of the classes**, and a **model set consisting of preclassified examples**.
- Most common techniques for classification: **decision trees, logistic regression and neural networks**.
- Other alternatives: **similarity models, memory-based reasoning, and naïve Bayesian models**.

# Binary Response Modelling (Binary Classification)

- Many business goals boil down to **separating two categories from each other**.
- Techniques such as **logistic regression** are specialised for these sorts of **yes or no models**.
- A response model score can be the **class label itself** or an **estimate of the probability of being in the class of interest**.
- The estimation approach has the great advantage that the **individual records can be rank ordered according to the estimate**.



# Prediction

- **Predictive modelling** is a technique that uses **mathematical and computational methods** to **predict an event or outcome**.
  - ▶ A **mathematical approach** uses an **equation-based model** that describes the phenomenon under consideration. The model is used to **forecast an outcome** at some future state or time based upon changes to the model inputs. The model parameters help **explain how model inputs influence the outcome**. e.g. **Linear regression model**.
  - ▶ The **computational predictive modelling approach** relies on models that are not easy to explain and often require **simulation techniques** to create a prediction (**black box** predictive modelling). e.g. **Neural networks**.
- Other alternatives: **regression, memory-based reasoning, and table lookup models**.



# Finding Clusters, Associations, Affinity Groups, and Anomalies

- Determining what things go together in a shopping cart at the supermarket, and finding groups of shoppers with similar buying habits are both examples of **undirected data mining**.
- Products that tend to sell together are called **affinity groups** and customers with similar behaviours comprise **market segments**.
- **Affinity grouping** is one simple approach to generating rules from data. If two items occur together frequently enough, you can think of how to use this information in marketing campaigns.



# Finding Clusters, Associations, Affinity Groups, and Anomalies

- **Clustering** is the task of **segmenting a heterogeneous population** into a number of more **homogeneous subgroups or clusters**. What distinguishes clustering from classification is that **clustering does not rely on predefined classes**.
- In clustering, there are no predefined classes and no examples. The **records are grouped together on the basis of self-similarity**. It is up to the data miner to **determine what meaning, if any**, to attach to the resulting clusters.
- **Anomaly detection** is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Anomaly detection is widely used in identifying bank fraud, structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

# What Technique for Which Tasks?

- Data mining tasks can be **used in creative ways** for applications outside the ones with which they are most often associated.
  - ▶ Is there a target or targets?
  - ▶ What is the target data like?
  - ▶ What is the input data like?
  - ▶ How important is ease of use?
  - ▶ How important is explicability?





# Is there a Target or Targets?

- **All directed data mining techniques**, including regression, decision trees, and neural networks, require **training with known values** for the **target variables**.
- When the data **does not contain such a target**, one of the undirected techniques such as **clustering** or **exploratory data analysis** is needed.



# What Is the Target Data Like?

- When the target is **numeric** and can take on a **wide range of values** (*continuous target*): **linear regression** models and **neural networks**  $(-\infty, +\infty)$ .
- **Regression** and **table lookup models** can all be used to estimate numeric values also, but they produce a **relatively small number of discrete values**.
- **Memory-based reasoning** is another choice for numeric targets that can produce a wide range of values, but **never outside the range of the original data**.
- When the target is a **binary response** or **categorical variable**: **decision trees**, **logistic regression** and **neural networks**.
- Other alternatives are **similarity models**, **memory-based reasoning**, and **naïve Bayesian models**.

# What Is the Input Data Like?

- **Regression models and neural networks** perform **mathematical operations on the input values** and so **cannot process categorical data or missing values**.
- **Categorical data** can be **recoded or replaced** with numeric fields that capture important features of the categories. It is also possible to input missing values.
- These operations can be **time-consuming** and **inaccurate**, however.
- **Decision trees, table lookup models, and naïve Bayesian models** can easily handle **categorical fields** and **missing values**.



# How Important Is Ease of Use?

- There is often a trade-off between **power**, **accuracy**, and **ease of use**.
- Some techniques require much more data preparation than others.
  - ▶ **Neural networks** require **all inputs to be numeric** and **within a small range of values**. They are also **sensitive to outliers** and **unable to process missing values**.
  - ▶ **Decision trees**, are much more forgiving and **require less data preparation**, but **may not do as good a job**.

# How Important Is Model Explicability?

- For some problems, getting the **right answer fast** is paramount. i.e. *Credit card transaction*.
- At the other extreme, some decisions require a **clear explanation of how the decision was made**. i.e. the granting or denial of a credit may be subject to regulatory review.

# How Important Is Model Explicability?

- Different techniques offer different trade-offs between **accuracy** and **explicability**.
- **Decision trees** arguably offer the **best explanations** because each leaf has a precise description in the form of a rule. The trade-off is that they **may not make use of as much of a variable's inherent information simply comparing it to a splitting value**.
- In a **regression**, when explanatory variables have been standardised, the relative magnitude of the model coefficients show how much each one contributes to the score. In addition, **every small change in the value of an explanatory variable** has an **effect on the score**.
- The regression model makes **more use of the information provided by the explanatory variables** than do **decision trees**.
- **Neural networks** are quite flexible and are capable of modelling quite complex functions very accurately, but are **essentially inexplicable**.