Welcome to:

Data Mining

Unit objectives

After completing this unit, you should be able to:

- Understand what data mining and data warehouse is
- Learn the various steps in data mining process
- Learn the importance of data mining
- Gain knowledge on the use case of data mining
- Gain knowledge on the challenges of data mining

Introduction to data mining

- Data
  - Data can be described by any facts, numbers, or text.

- Data can be classified into below three categories:
  - Structured data.
  - Unstructured data.
  - Semi-structured data.

- Information:
  - The patterns, associations, or relationships among all this data can provide information.

- Knowledge:
  - Information is transformed to knowledge on historical and future patterns.

Why data mining?

- Abundant growth of data:
  - Availability of huge amounts of data in the repository.
  - The imminent need for transforming such data into useful information and knowledge.
  - Use the knowledge for analyzing applications related to market analysis, production control, explorations in science etc.
  - Number of records too large (millions or billions).
  - High dimensional (attributes/features/fields) data (thousands).

- Major sources of abundant data:
  - Sales in retail, policy and claim data in insurance, medical history data in health care.
  - Financial data in banking and securities.
  - Web navigation, on-line collections.

What is data mining?

- Data mining
  - Extracting or “mining” knowledge from large amounts of data.
  - The process of mining actionable information from large sets of data.
  - Term is a misnomer.

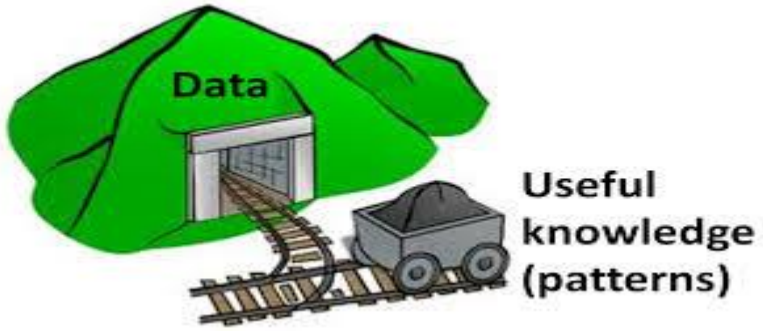


Figure: Knowledge pattern of data

- Other terms for data mining
  - Knowledge extraction, knowledge mining from data, data/pattern analysis, data dredging and data archeology, Knowledge Discovery from Data (KDD).

Need for data mining tools

- Human analysis breaks down with volume and dimensionality
  - How quickly can one digest 1 million records, with 100 attributes.
  - High rate of growth, changing sources.

- What is done by non-statisticians?
  - Select a few fields and fit simple models or attempt to visualize.

Evolution of data mining?

- 1960s
  - Database and information technology lead to sophisticated and powerful database systems.
- 1970s
  - Database system progressed from hierarchical, network database system to relational database systems.
- 1980s
  - Advanced data analysis.
- 1990s
  - Data warehouse, data mining, Worldwide Web and web based databases.

Process involve in data mining? (1 of 2)

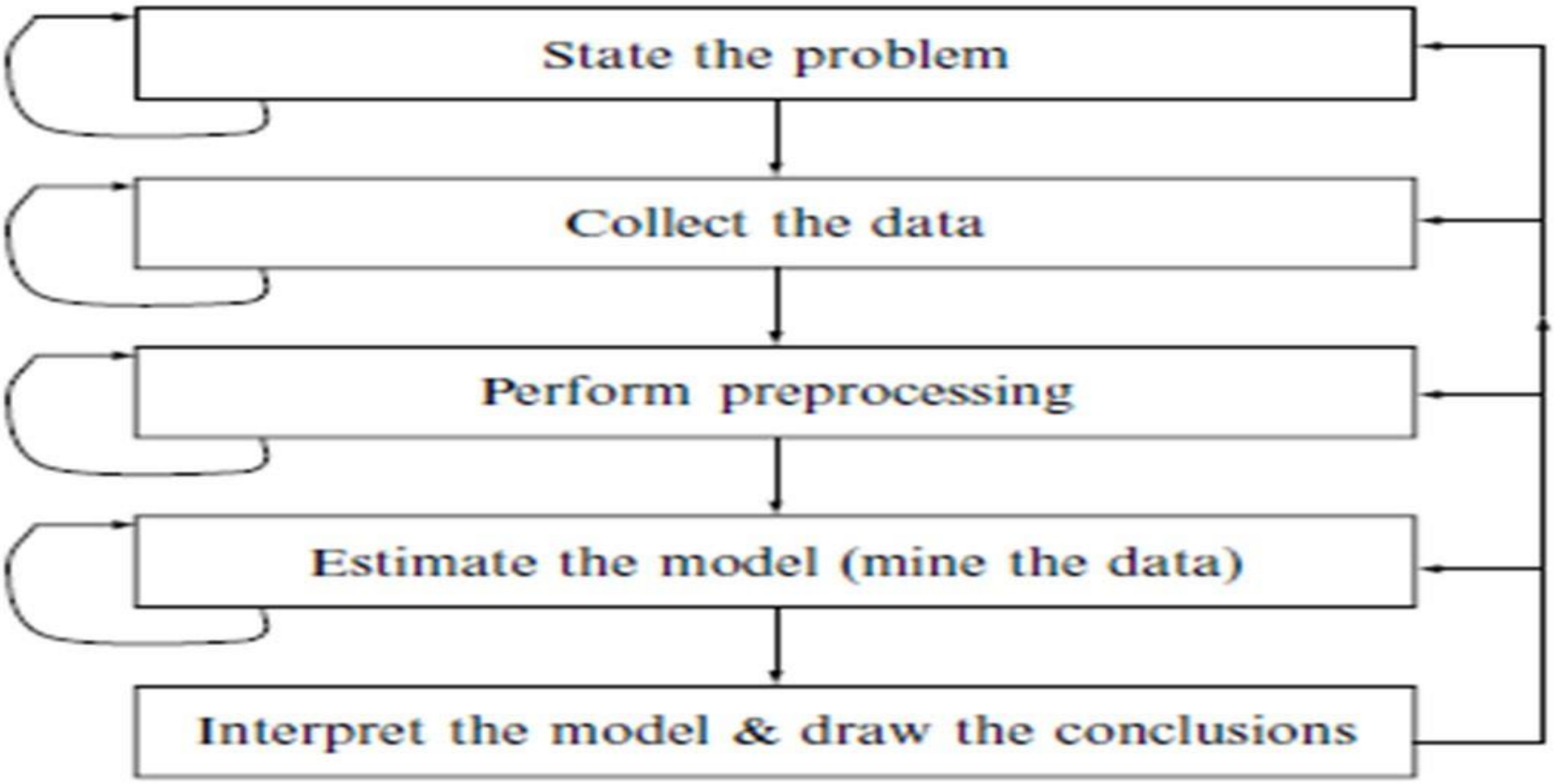


Figure: Process involve in data mining

Process involve in data mining? (2 of 2)

- Building a data mining model is part of a larger process which includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment.

- This process is defined by the six basic steps:
  - Defining the problem.
  - Preparing data.
  - Exploring data.
  - Building models.
  - Exploring and validating models.
  - Deploying and updating models.

Data mining process?

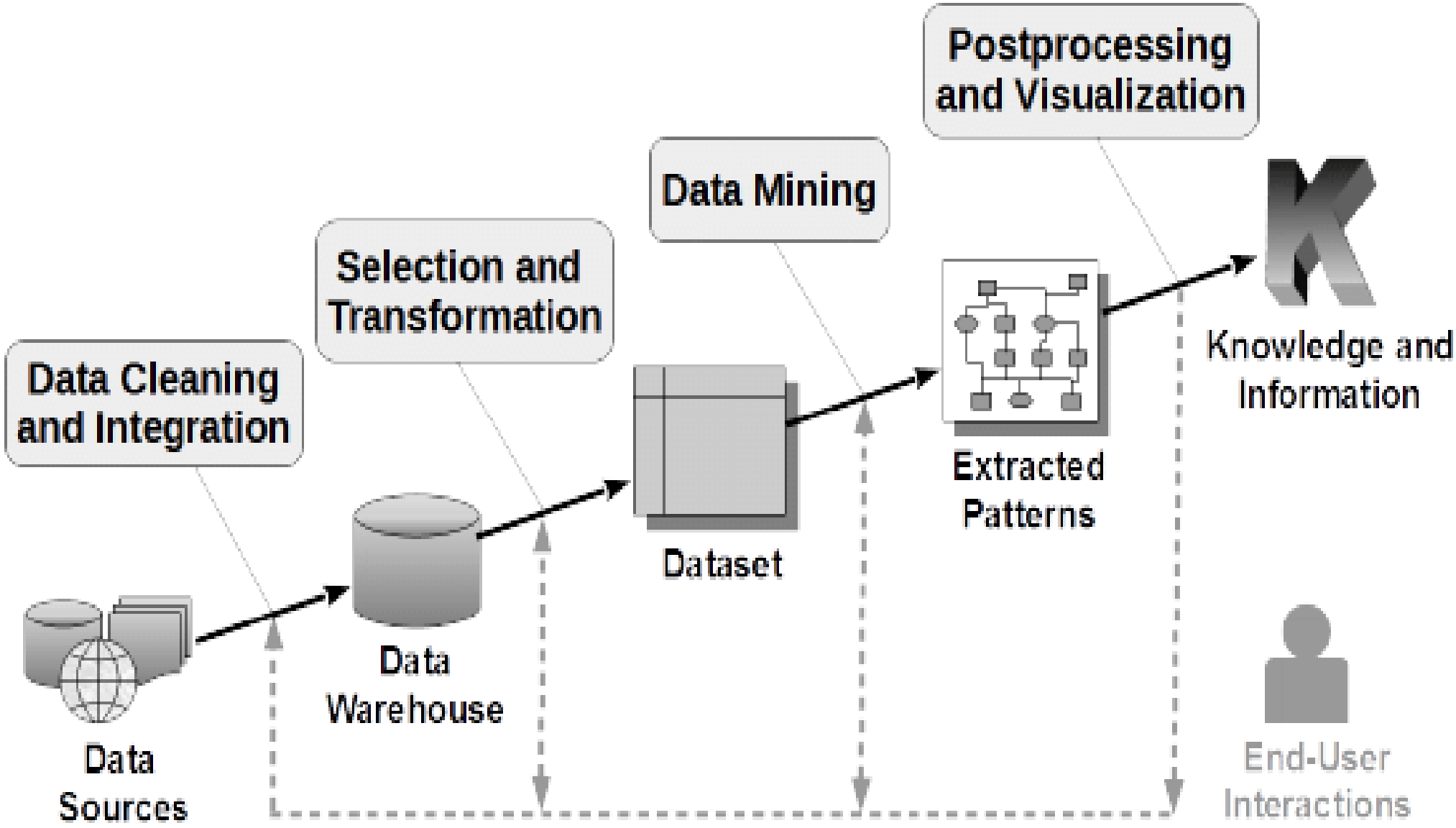


Figure: Data mining process

KDD process model (1 of 2)

- Knowledge Discovery from Data (KDD), refers to the process of exploring relevant information, which is hidden in data. It also emphasizes on method which helps with "high-level" application.

- This problem is of interest to person who does research in statistics, machine learning, databases, artificial intelligence, pattern recognition, acquisition of knowledge for the smart system & data/information visualization.

- The primary objective of the KDD processes to extract data knowledge, that depends on the contexts of big database.

KDD process model (2 of 2)

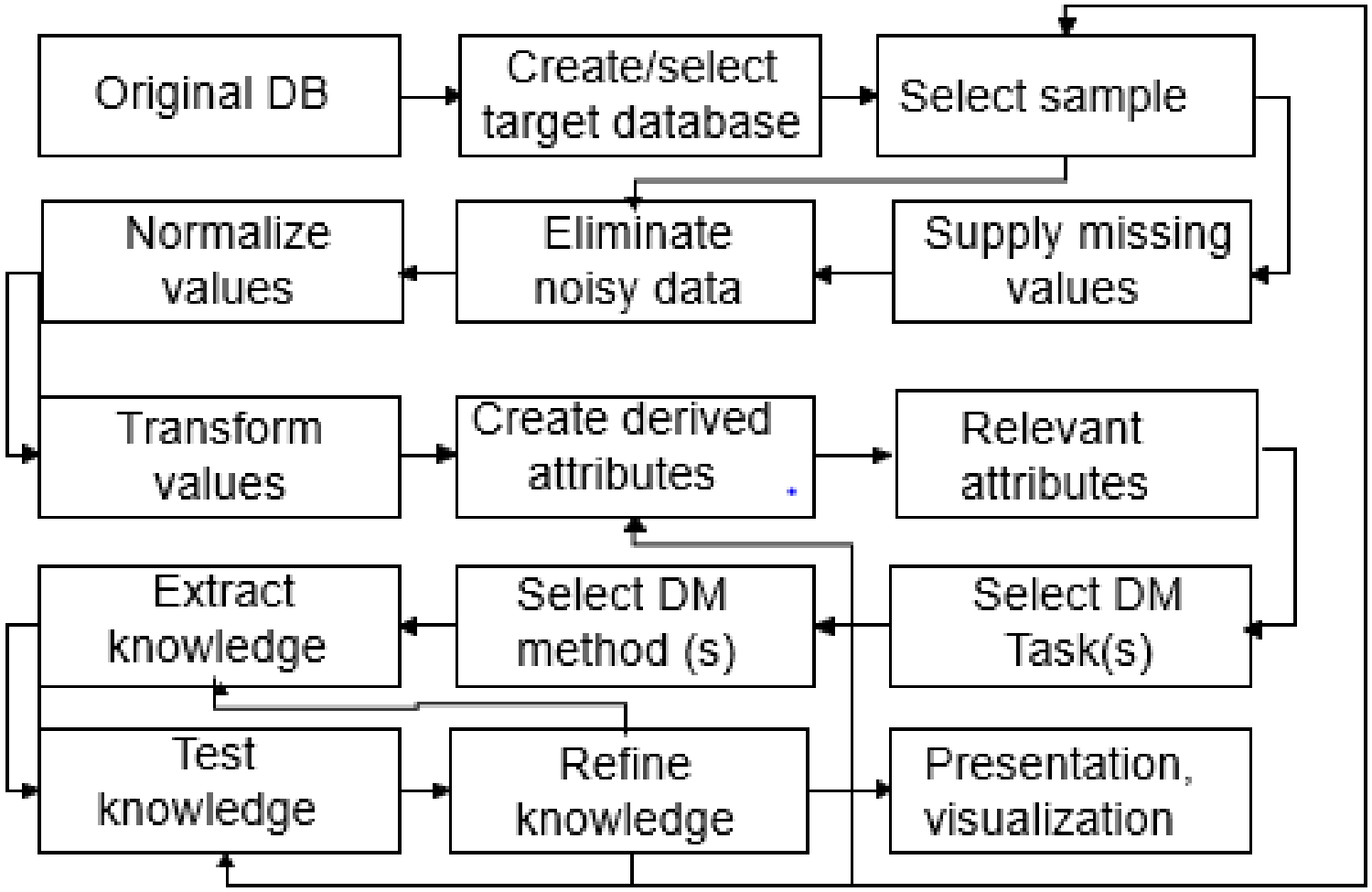
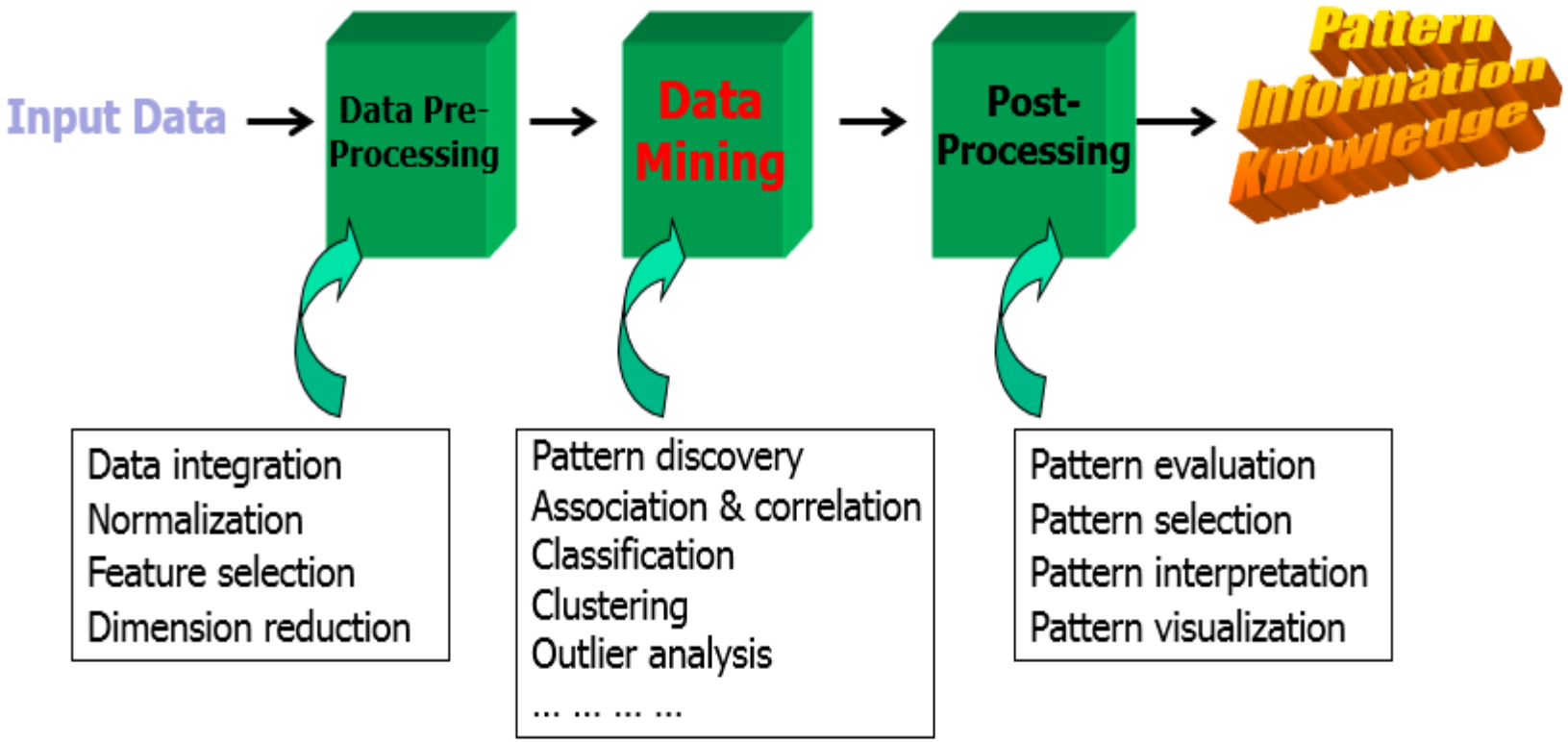


Figure: KDD process model architecture

Research challenges for KDD



- This is a view from typical machine learning and statistics communities

Figure: Research challenges of KDD

Data mining: On what kinds of data?

- Database data
- Data warehouse: A repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.
- Transactional data: Each record in a transactional database captures a transaction.
- Other kinds of data: Say, time-related or sequence data, data streams, spatial data, engineering design data, hypertext and multimedia data, graph and network data, the web.

Scenario: Need for database

- Consider all electronics, a successful international company having branches around the world.
- Number of transactions in each branch is vast, so each branch has its own set of databases.
- The president of all electronics needs a report providing an analysis of the company's sales per item type per branch for the third quarter.
- Since the relevant data is very huge and also spread out over several databases located at numerous remote locations it is a difficult task. How such report can be produced.

Mining on different kinds of data

- Datasets generated from transactional database applications
  - Data from relational database, transactional database and data warehouse.

- Advanced data sets from various sources
  - Streams of data and data generated from sensor.
  - Data from time-series, temporal data, data related to bio-informatics.
  - Structured data, graph data, social network data and multi-linked data.
  - Data from object-relational DB.
  - Data from heterogeneous DB & legacy DB.
  - Data from space and spatiotemporal.
  - Data from multimedia.
  - Textual data.
  - Data from internet.

Types of data mining tasks

- General descriptive knowledge
  - Summarizations.
  - Symbolic descriptions of subsets.

- Discriminative knowledge
  - Distinguish between K classes.
  - Accurate classification (also black box).
  - Separate spaces.

CRISP-DM (1 of 2)

- CRISP-DM
- Stands for Cross Industry Standard Process for Data Mining.
- Industry-proven way to guide your data mining efforts.
- In CRISP-DM, the life cycle of a data mining project is divided into six phases. The sequence of the phases is not strict, can be moved back and forth between different phases as per the need of the business" -ibm.com.
  - Understanding business
  - Understanding data
  - Preparation of data
  - Modeling
  - Evaluation
  - Deployment

CRISP-DM: Elaborate view (2 of 2)

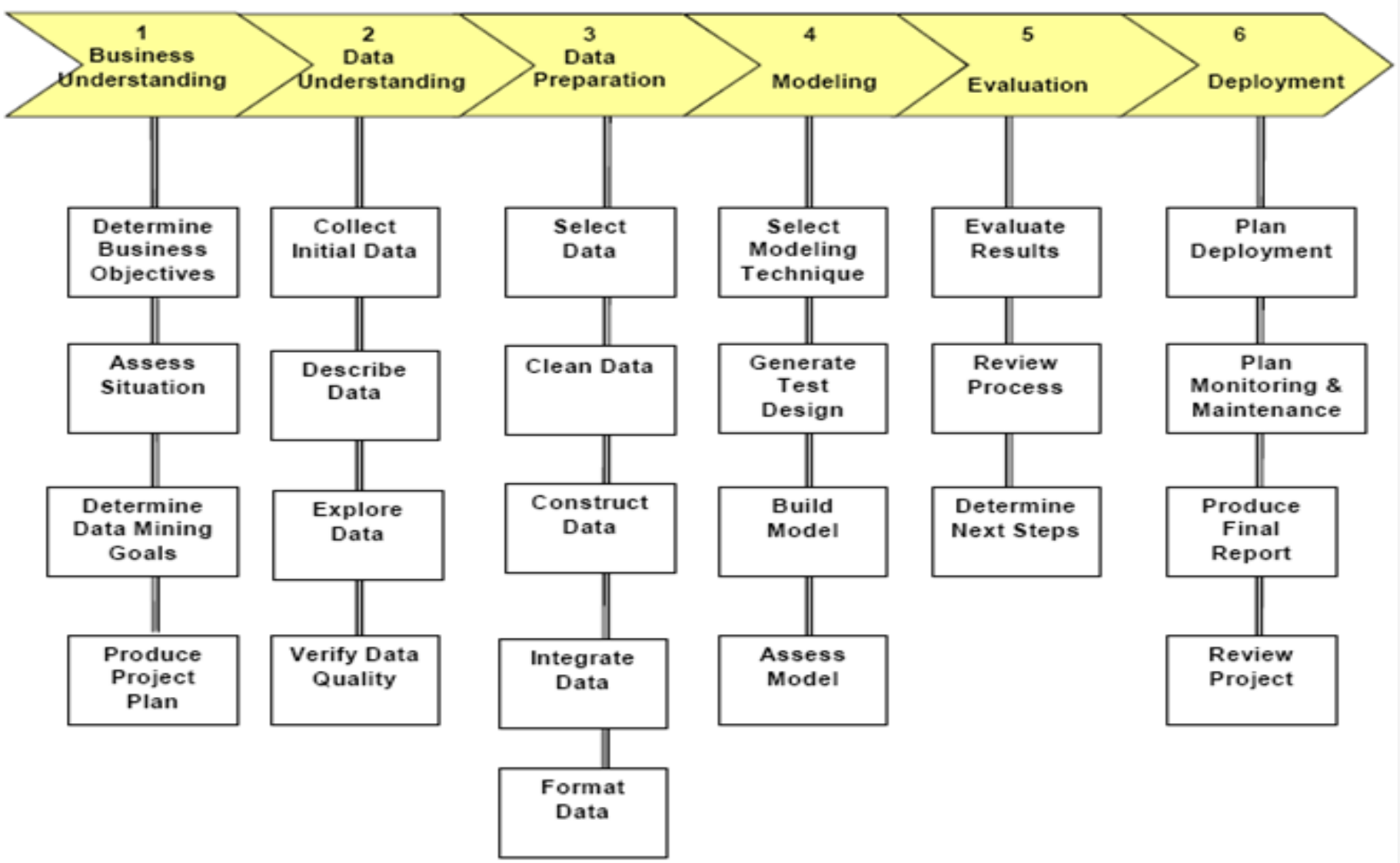


Figure: CRISP-DM: Elaborate view

Components of DM methods

- Representation: Language for patterns/models, expressive power.
- Evaluation: Scoring methods for deciding what is a good fit of model to data.
- Search: Method for enumerating patterns/models.

Data mining operations

- Verification driven:
  - Validating hypothesis.
  - Querying and reporting (spread sheets, pivot tables).
  - Multidimensional analysis (dimensional summaries); On-Line Analytical Processing.
  - Statistical analysis.

- Discovery driven:
  - Exploratory data analysis.
  - Predictive modelling.
  - Database segmentation.
  - Link analysis.
  - Deviation detection.

Data mining techniques

- Association rules: detect sets of attributes that frequently co-occur, and rules among them.
- e.g., 90% of the people who buy cookies, also buy milk (60% of all grocery shoppers buy both).
- Sequence mining (categorical): Discover sequences of events that commonly occur together.
- .e.g., In a set of DNA sequences ACGTC is followed by GTCA after a gap of 9, with 30% probability.
- CBR or Similarity search: Given a database of objects, and a “query” object, find the object(s) that are within a user-defined distance of the queried object, or find all pairs within some distance of each other.

Applications of data mining

- Sales/Marketing:
  - Identify buying patterns from customers.
  - Find the association among customer demographic characteristics.
  - Predict response to mailing campaigns.
  - Market basket analysis.

- Banking:
  - Credit card fraudulent detection.
  - Identify ‘loyal’ customers.
  - Predict customers who are likely to change their credit card affiliation.
  - Determine credit card spending rate by customer groups.
  - Find the hidden correlation’s between different financial indicators.
  - Identify rules related to stock trading from historical market data.

Predictive analytics

- Information is the oil of the 21st century, and analytics is the combustion engine.”-Peter Sondergaard, senior VP and global head of research for Gartner.

- “Predictive analytics helps connects data to effective action by drawing reliable conclusions about current conditions and future events”-Gareth Herschel, Research Director, Gartner Group.

- There is no need to “learn” to calculate payroll.

- Learning is used when:
  - Human expertise does not exist (navigating on mars).
  - Humans are unable to explain their expertise (speech recognition).
  - Solution changes in time (routing on a computer network).
  - Solution needs to be adapted to particular cases (user biometrics).

Where predictive analytics is used

- Predictive customer analytics:
  - Acquiring customers.
  - Growing with customer.
  - Retaining the best customer.

- Predictive operational analytics:
  - Plan
  - Manage
  - Maximize

- Predictive threat and fraud analytics:
  - Monitoring process
  - Detection process
  - Control

Issues and challenges in predictive analytics or data mining

- Methods where mining is perfumed.
- Interaction with user.
- Efficiency and scalable nature of data.
- Diversification of data.
- About societal data.

What is machine learning?

- Machine learning
  - Study of algorithms that.
  - Improve their performance.
  - At some task.
  - With experience.

- Optimize a performance criterion using example data or past experience.

- Role of statistics: Inference from a sample.

- Role of computer science: Efficient algorithms to:
  - Solve the optimization problem.
  - Representing and evaluating the model for inference.

Growth of machine learning

- Machine learning is preferred approach to:
  - Speech recognition, natural language processing.
  - Computer vision.
  - Medical outcomes analysis.
  - Robot control.
  - Computational biology.

- This trend is accelerating
  - Improved machine learning algorithms.
  - Improved data capture, networking, faster computers.
  - Software too complex to write by hand.
  - New sensors/IO devices.
  - Demand for self-customization to user, environment.
  - It turns out to be difficult to extract knowledge from human experts.
  - Failure of expert systems in the 1980’s.

Applications

- Supervised learning
  - Classification
  - Regression/Prediction
- Unsupervised learning
- Reinforcement learning

Checkpoint (1 of 2)

Multiple choice questions:

1. Why we require data mining?
   a) Help us to reduce data
   b) We cannot handle so much data
   c) Data is in different forms
   d) All of the above

2. What are the different kind of data generating now days?
   a) Structure data
   b) Un-structure data
   c) Semi-Structure data
   d) All of the above

3. The 2 strategies for dealing outliers are:
   a) Detecting and eventually removing outliers as a part of building model phase
   b) Detecting and eventually removing outliers as a part of preprocessing phase
   c) Developing robust modeling methods such that they are sensitive to outliers
   d) Developing robust modeling methods such that they are insensitive to outlier

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. Why we require data mining?
   a) Help us to reduce data
   b) We cannot handle so much data
   c) Data is in different forms
   d) All of the above

2. What are the different kind of data generating now days?
   a) Structure data
   b) Un-structure data
   c) Semi-structure data
   d) All of the above

3. The 2 strategies for dealing outliers are:
   a) Detecting and eventually removing outliers as a part of building model phase
   b) Detecting and eventually removing outliers as a part of preprocessing phase
   c) Developing robust modeling methods such that they are sensitive to outliers
   d) Developing robust modeling methods such that they are insensitive to outlier

Checkpoint (2 of 2)

Fill in the blanks:

1. _____ studies the collection, analysis, interpretation or explanation, and presentation of data.
2. Supervised learning is basically a synonym for _____.
3. Information is transformed to _____ on historical and future patterns.
4. Data mining refers to _____ knowledge from large amounts of data.

True or False:

1. Machine learning is preferred approach to robot control. True/False
2. Data mining is not a kind of transactional data. True/False
3. Unsupervised learning is essentially a synonym for data mining**.** True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Statistics** studies the collection, analysis, interpretation or explanation, and presentation of data.
2. Supervised learning is basically a synonym for **classification**.
3. Information is transformed to **knowledge** on historical and future patterns.
4. Data mining refers to **extracting or mining** knowledge from large amounts of data.

True or False:

1. Machine learning is preferred approach to robot control. **True**
2. Data mining is not a kind of transactional data. **False**
3. Unsupervised learning is essentially a synonym for data mining. **False**

Question bank

Two mark questions:

- What is data mining?
- List the abundant growth of data in data mining.
- List the process involve in data mining.
- What is machine learning?

Four mark questions:

- Describe introduction to data mining.
- Describe the needs of data mining and its goal.
- Describe data mining techniques.
- List the application of data mining.

Eight mark questions:

- Briefly describe machine learning and the growth of machine learning.
- What are the components of data mining methods and its operation?

Unit summary

Having completed this unit, you should be able to:

- Understand what data mining and data warehouse is
- Learn the various steps in data mining process
- Learn the importance of data mining
- Gain knowledge on the use case of data mining
- Gain knowledge on the challenges of data mining