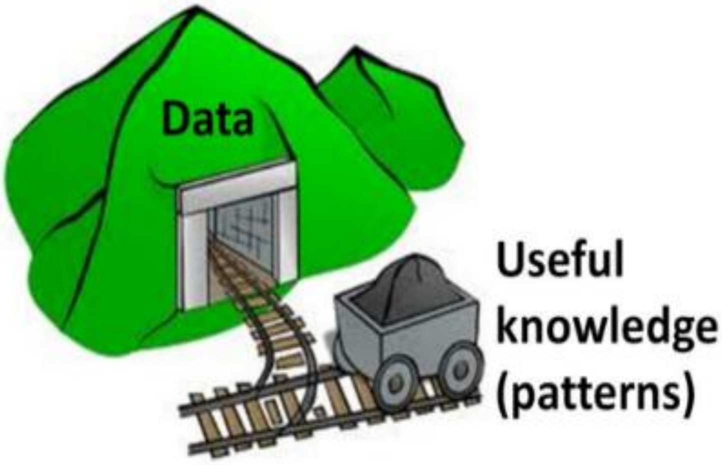
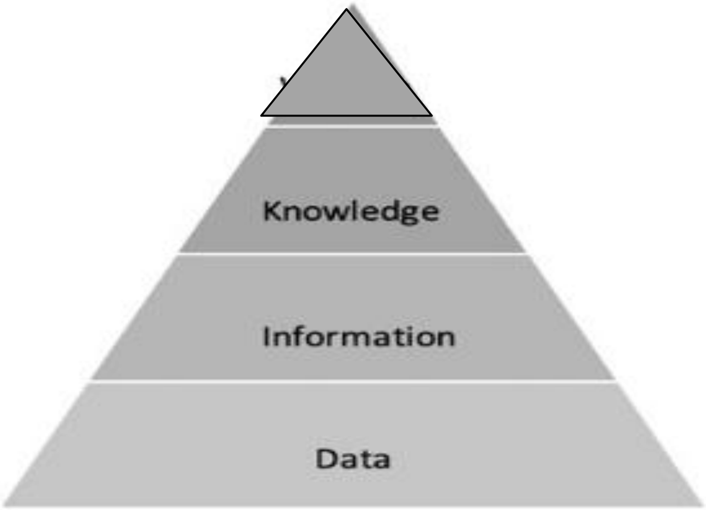
Definition of 'Data Mining'



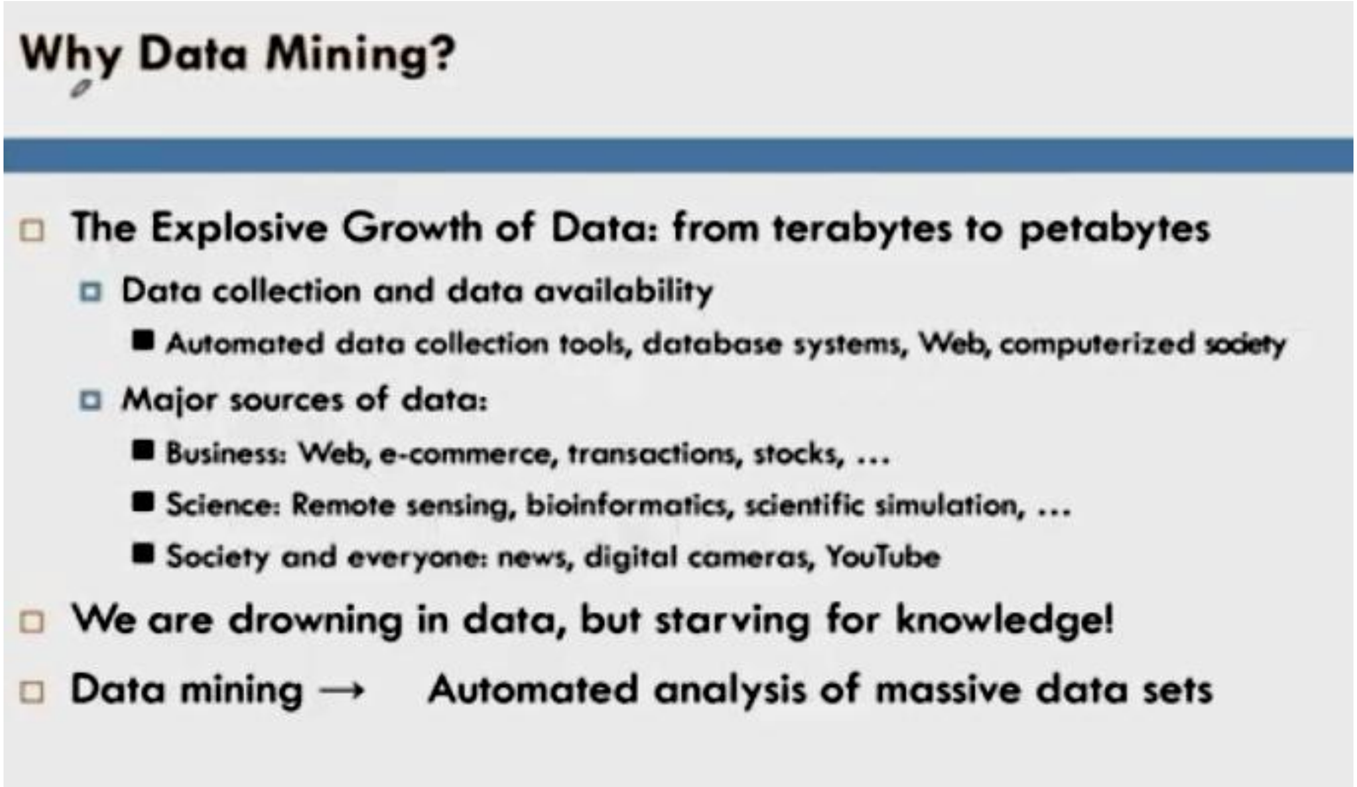
Data Pyramid



Data: refers to raw symbols or facts that represent properties of objects, events, and their environments.

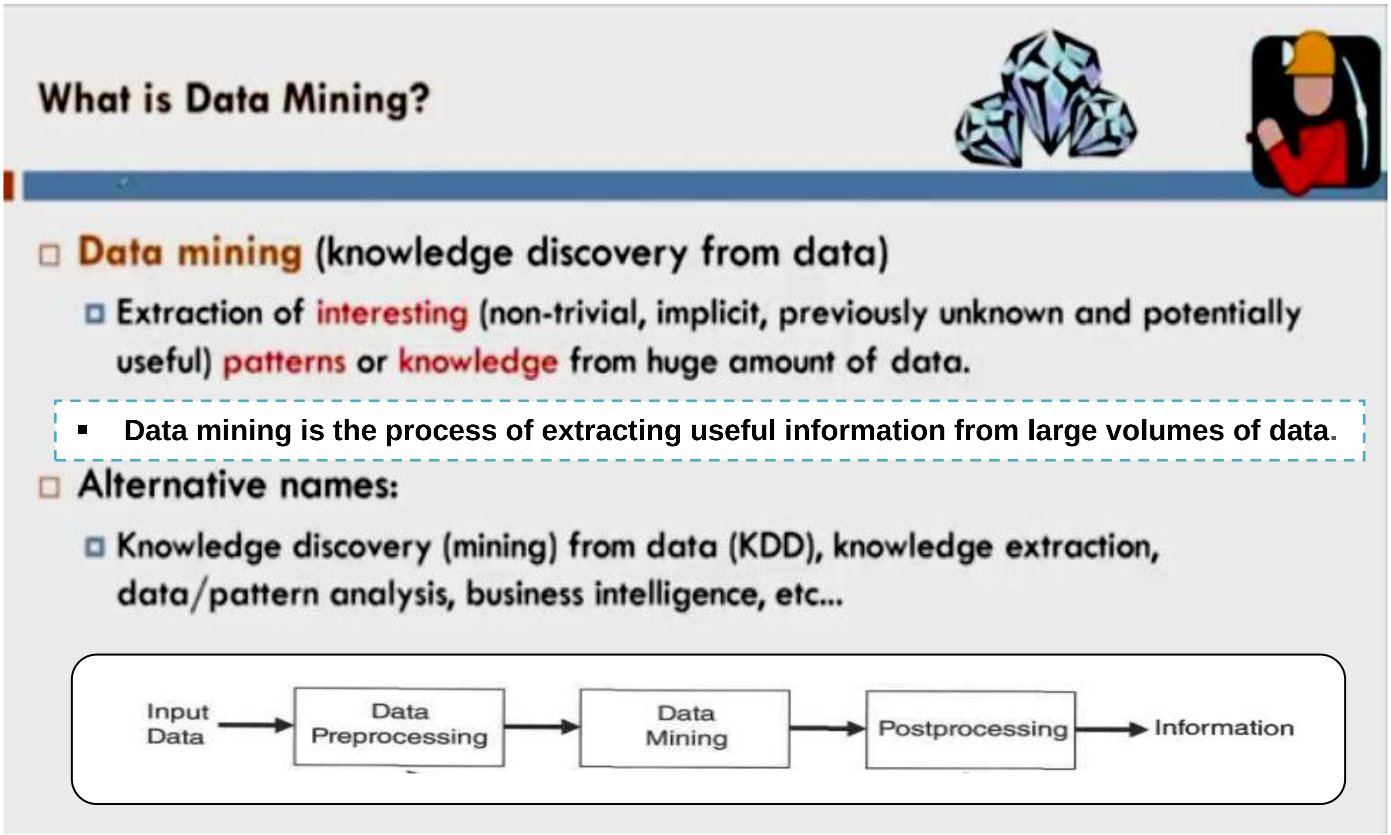
Information: Data that is processed, organized, or presented in a context to make it meaningful.

Knowledge: Formulate a set of facts in rules to use them to make decision.



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 data:
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- Data mining → Automated analysis of massive data sets



What is Data Mining?

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

- **Data mining is the process of extracting useful information from large volumes of data.**

- Alternative names:
  - Knowledge discovery (mining) from data (KDD), knowledge extraction, data/pattern analysis, business intelligence, etc...

Pattern

Pattern is an arrangement of repeated parts. It is a summarization of relationships in the data, perhaps holding for only a few records or a few variables or both.

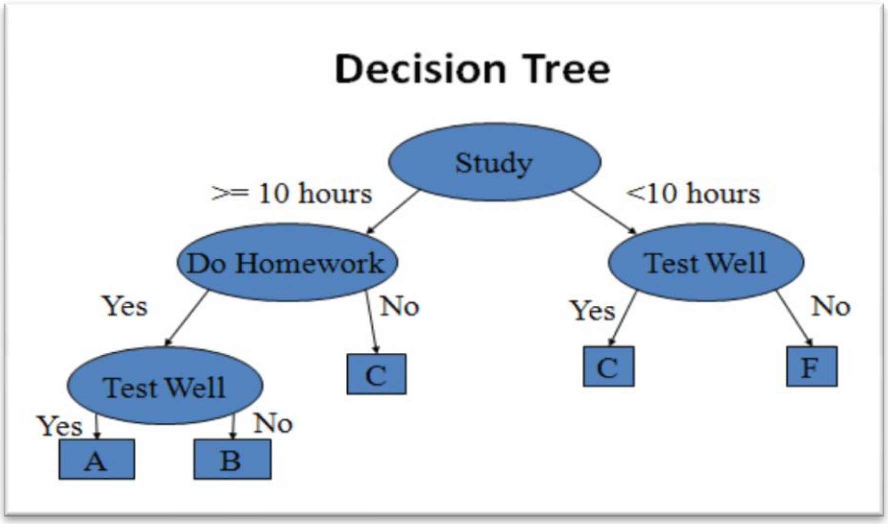
Example:

Shape	**Color**	**Weight**
Box	**Red**	**100**
Box	**Red**	**200**
Box	**Red**	**300**
Box	**Blue**	**400**
Cone	**Blue**	**400**

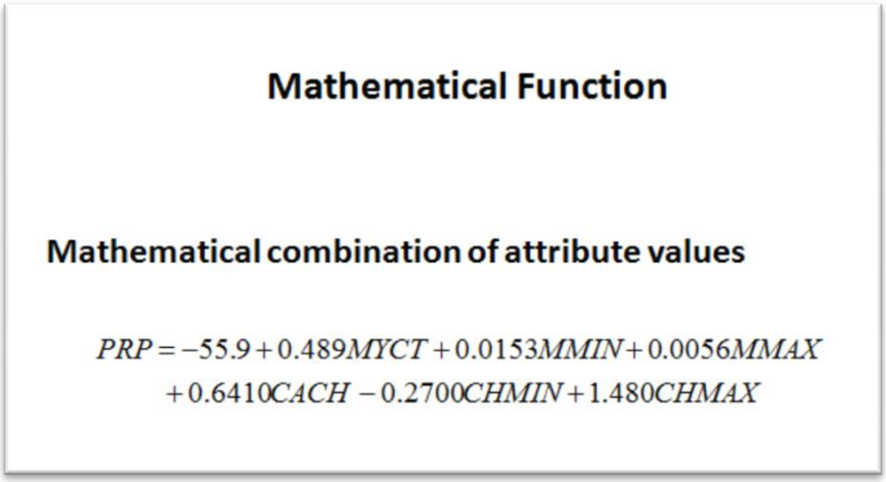
1. **If Shape= Box then Color=Red**
2. **If Shape= Box then Weight <=300**
3. **If Color= Blue then Weight = 400**

Model

Model is a global summary of data sets that can describe or summarize the unknown relations between the data items (or subset of patterns). It is represented as mathematical function, set of rules, decision tree, neural network.



Decision Tree



Mathematical Function

Mathematical combination of attribute values

$$\begin{aligned} PRP = & -55.9 + 0.489MYCT + 0.0153MMIN + 0.0056MMAX \\ & + 0.6410CACH - 0.2700CHMIN + 1.480CHMAX \end{aligned}$$

Data Mining Tasks

Data mining tasks are generally divided into two major categories:

Predictive tasks

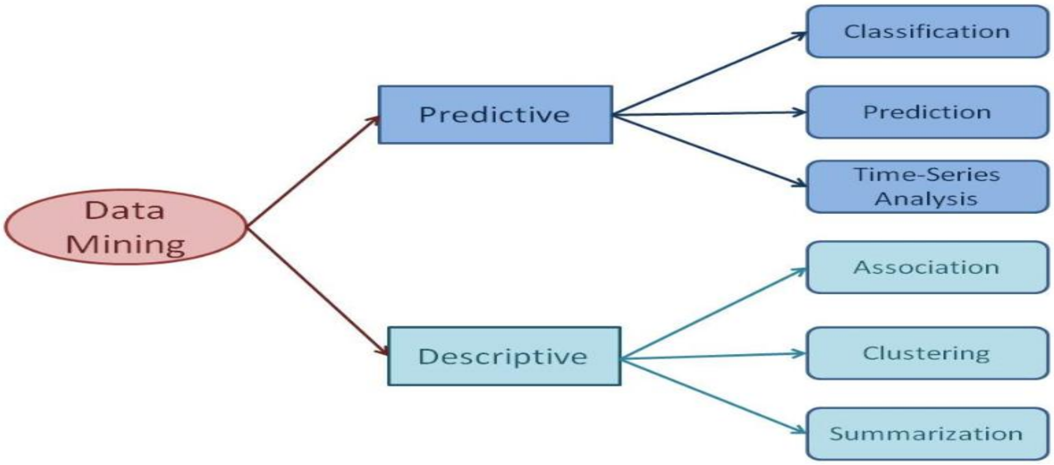
The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Descriptive tasks

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

Different Data Mining Tasks

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining.



1. Classification

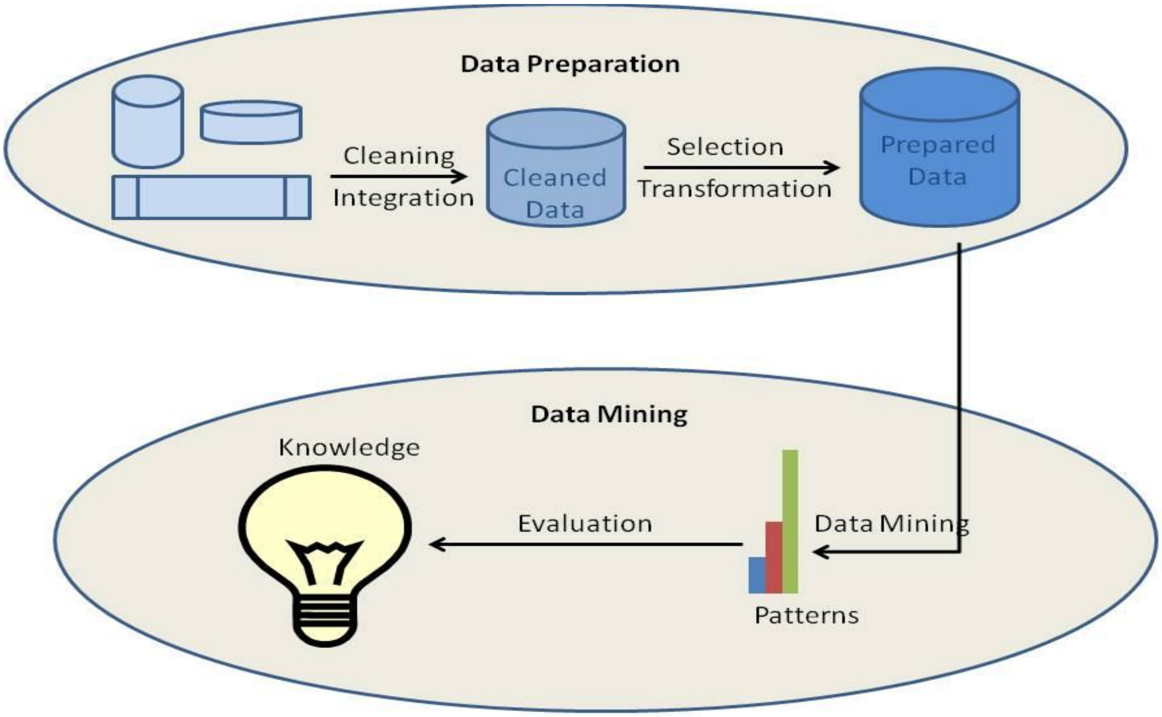
Classification derives a model from classified data to determine the class of a new object based on its attributes.

2. Clustering

Clustering finds groups of closely related data objects in data set so that objects that belong to the same cluster (group) are more similar to each other and more dissimilar with objects in other cluster.

KDD process

KDD (knowledge discover from data) is a process of detecting interesting patterns in the data. It includes: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation are to be completed in the given order.



1. **Data Cleaning**

Data cleaning is the process where the data gets cleaned (remove noise and inconsistent data, and handle missing values).

2. **Data Integration**

Data integration is the process where data from different data sources are integrated into one.

3. Data Selection

Data selection is the process where the data relevant to the analysis is retrieved from the database (data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations).

4. Data Transformation

Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining.

5. Data Mining

Data mining is the core process where intelligent methods are applied to extract data patterns,

6. Pattern Evaluation

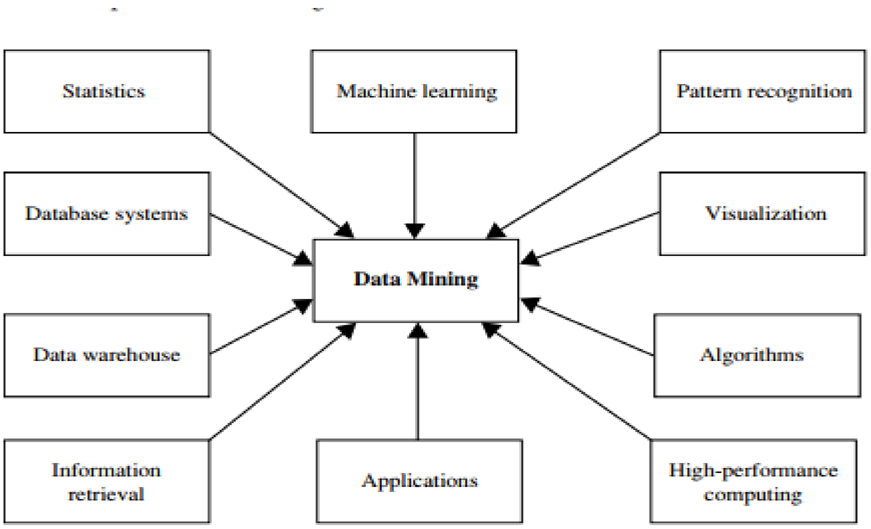
The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures.

7. Knowledge Representation

The information mined from the data needs to be presented to the user in an appealing way where visualization and knowledge representation techniques are used to present mined knowledge to users.

Data Mining Techniques

Data mining integrates approaches and techniques from various disciplines such as machine learning, statistics, artificial intelligence, neural networks, database management, data warehousing, data visualization, spatial data analysis, probability graph theory etc. In short, data mining is a multi-disciplinary field.



Examples of Kind of Data can be mined

1. **Flat Files:** simple data files in text or binary format with structure known by the data mining algorithms.
2. **Relational database.**
3. **Multimedia Database**: images, audio, and text media, high dimension data.
4. **Spatial Database:** geographical information like maps.
5. **Time Series Database:** like stoke market data, or logged activities.

Challenges in Data Mining

Though data mining is very powerful, it faces many challenges during its implementation. The challenges could be related to performance, data, methods and techniques used etc. The data mining process becomes successful when the challenges or issues are identified correctly and sorted out properly.

Noisy and Incomplete Data

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors.

Complex Data

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

Performance

The performance of the data mining system mainly depends on the efficiency of algorithms and techniques used. If the algorithms and techniques designed are not up to the mark, then it will affect the performance of the data mining process adversely.

Distributed Data

Real world data is usually stored on different platforms in distributed computing environments. It could be in databases, individual systems, or even on the Internet. It is practically very difficult to bring all the data to a centralized data repository mainly due to organizational and technical reasons.

Data Mining Applications

Data mining applications are vast and varied, with applications across industries and disciplines. Here are some common areas where data mining techniques are applied:

1. **Business and Marketing:** Data mining in business and marketing is used for shopping cart analysis to understand customer purchasing behavior and perform customer segmentation for targeted marketing campaigns. Predictive modeling for sales forecasting and customer churn prediction. Sentiment analysis of social media data provides a recommendation system to understand customer opinions and feedback and recommend personalized products.

2. **Finance:** Data mining techniques are most commonly used for detecting fraud in banking transactions, risk assessment and credit scoring for loan approval, stock market analysis and forecasting, and predicting customer lifetime value for marketing strategies.

3. **Healthcare:** Healthcare data mining is the discovery of patterns, correlations, and insights from large data sets generated in the healthcare industry. The most common tasks of data mining in healthcare include disease prediction and diagnosis, Drug discovery and development, Patient monitoring and personalized treatment recommendations, and Health outcome prediction for patient care management.

4. **Telecommunications:** Data mining techniques are most commonly used for detecting fraud in banking transactions, risk assessment and credit scoring for loan approval, stock market analysis and forecasting, and predicting customer lifetime value for marketing strategies.

5. **Manufacturing and Supply Chain:** Predictive maintenance of machinery and systems, supply chain optimization, demand forecasting, quality control, and error detection in manufacturing processes.

6. **Education:** Adaptive learning systems for personalized education and dropout prediction and prevention strategies, student performance prediction and early intervention, and adaptive learning systems.

7. **Government and Public Sector:** To extract useful information and patterns from large amounts of data collected by government agencies and organizations, data mining uses advanced analytical techniques. Fraud detection in public welfare programs, Crime pattern analysis for law enforcement, and Traffic flow prediction and optimization.

8. **E-commerce and Retail:** Data mining plays a crucial role in the E-commerce and retail industries, offering insights into customer behavior, market trends, product performance, and more. Product recommendation systems, Price optimization and dynamic pricing, and Inventory management and demand forecasting.

9. **Energy and Utilities:** Data mining within the energy and utilities sector includes extricating important insights and patterns from large datasets produced by different operations within these businesses. Energy consumption prediction and optimization, equipment failure prediction for planning, and renewable energy forecasting.

10. **Media and Entertainment:** Data mining is the process of collecting valuable information and patterns from a large amount of data on various aspects of media consumption, audience behavior, content preferences, or anything else that might be relevant to this industry. Content recommendation systems, segmentation of audiences for targeted advertising, and Box Office revenue estimates.