DATA MINING

Presented and Created By: Shawni Dutta.

BACKGROUND

- Rapid advances in data collection and storage technology has enabled organizations to accumulate large amount of data.
- Extracting useful information has proven extremely challenging.
- Traditional data analysis may not be fruitful due to the massive data size.
- Again, even if the dataset size of small, due to non-traditional nature of data, traditional data analysis may not be applied.

WHAT IS DATA MINING?

- Data mining is a technology that blends traditional data analysis methods with sophistical algorithms for processing large volumes of data.
- Data mining is the process of automatically discovering useful information in large data repositories.
- Data mining techniques are applied to large databases for finding novel and useful patterns that might remain unknown otherwise.
- They also provide information about whether a newly arrived customer will spend more than 100 USD at a department store.

INFORMATION RETRIEVAL SYSTEM AND DATA MINING

- Not all information retrieval tasks are data mining.
- For example, looking up individual records using database management system or finding particular web pages via a query to an Internet search engine are example of Information retrieval task.
- Although, these tasks are important.
- But, Data mining techniques have been used to enhance information retrieval systems.

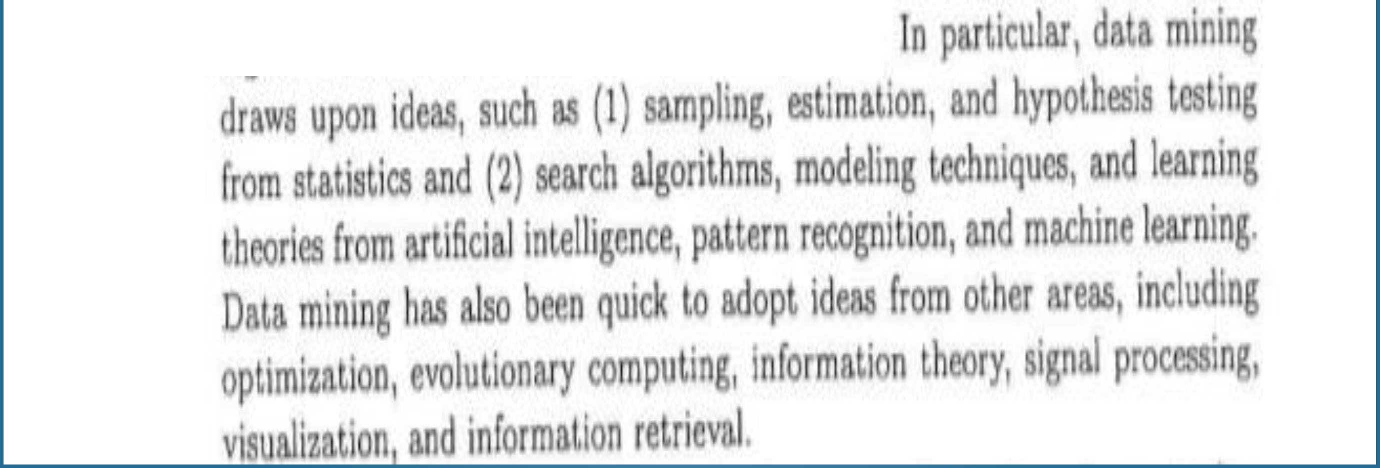
DATA MINING TURNS A LARGE COLLECTION OF DATA INTO KNOWLEDGE

- A search engine (e.g., Google) receives hundreds of millions of queries every day.
- Each query can be viewed as a transaction where the user describes her or his information need. What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time?
- Interestingly, some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone.
- For example, Google's Flu Trends uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms.
- A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, Flu Trends can estimate flu activity up to two weeks faster than traditional systems can

DATA MINING IN BRIEF

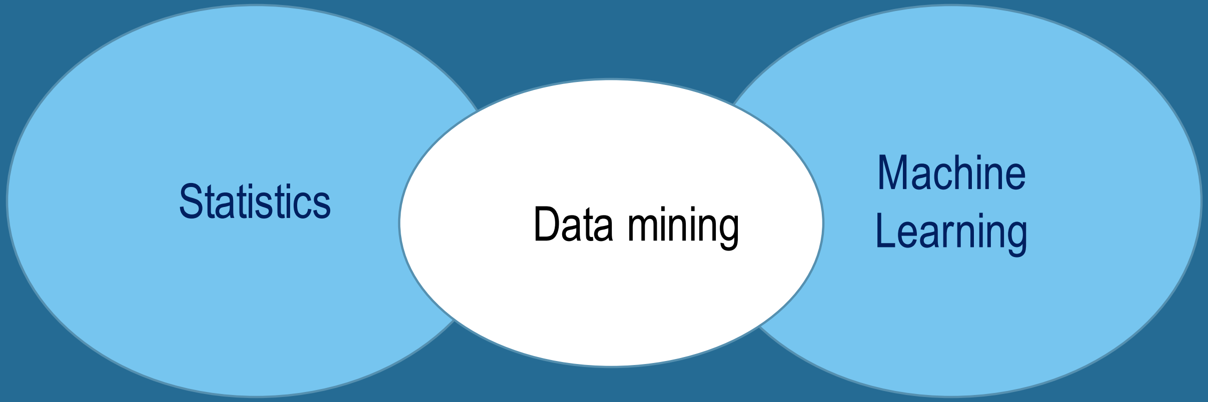
- Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.
- The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

ORIGINS OF DATA MINING



In particular, data mining draws upon ideas, such as (1) sampling, estimation, and hypothesis testing from statistics and (2) search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning. Data mining has also been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval.

DATA MINING AS A CONFLUENCE OF MANY DISCIPLINES



DATA MINING IN BUSINESS

- Data Mining techniques can be used to support a wide range of business intelligence applications such as customer profiling, targeted marketing, work-flow management, fraud detection etc.
- It can also help retailers answer important questions such as-
    - Who are the most profitable customers?
    - What is the revenue outlook of the company for the next year?

DATA MINING IN SCIENCE

- For understanding the Earth's climate system, NASA has developed a series of Earth's-orbiting satellites that continuously generate global observations of the land surface, oceans, and atmosphere.
- Traditional methods may not be feasible to analyze these large datasets.
- It can also help Earth scientists answer important questions such as-
  - What is the relationship between the frequency and intensity of ecosystem disturbances such as droughts and hurricanes to global warming?
  - How well can we predict the beginning and end of growing season for a region?

DATA MINING IN TELECOMMUNICATION

- In this, data mining gains a competitive advantage and reduce customer churn by understanding demographic characteristics and predicting customer behavior.
- Increases customer loyalty and improve profitability by providing customized services.
- As it supports customer strategy by developing appropriate marketing campaigns and pricing strategies.

DATA MINING IN EDUCATION

- There is a newly emerging field, called Educational Data Mining. As it concerns with developing methods. That discover knowledge from data originating from educational Environments.
- The goals of EDM are identified as predicting students' future learning behavior, studying. We use data mining by an institution to take accurate decisions. And also to predict the results of the student.
- With the results, the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured. And used to develop techniques to teach them.

Q: DISCUSS WHETHER OR NOT EACH OF THE FOLLOWING ACTIVITIES IS A DATA MINING TASK.

- *Dividing the customers of a company according to their gender.*

A: No. This is a simple database query.

- *Dividing the customers of a company according to their profitability.*

A: No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

Q: DISCUSS WHETHER OR NOT EACH OF THE FOLLOWING ACTIVITIES IS A DATA MINING TASK.

- *Computing the total sales of a company.*

A: No. Again, this is simple accounting.

- *Sorting a student database based on student identification numbers.*

A: No. Again, this is a simple database query.

- *Extracting the frequencies of a sound wave.*

A: No. This is signal processing.

Q: DISCUSS WHETHER OR NOT EACH OF THE FOLLOWING ACTIVITIES IS A DATA MINING TASK.

- *Predicting the future stock price of a company using historical records.*

A: Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modeling.

- *Monitoring the heart rate of a patient for abnormalities*

A: Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection.

DATA MINING ISSUES

- Scalability
- High Dimensionality

KNOWLEDGE DISCOVERY PROCESS (KDD)

- In addition, many other terms have a similar meaning to data mining—for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.
- Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data, or KDD.**
- However, others view data mining as merely an essential step in the process of knowledge discovery.

KNOWLEDGE DISCOVERY PROCESS (KDD)

- The knowledge discovery process is shown in as an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

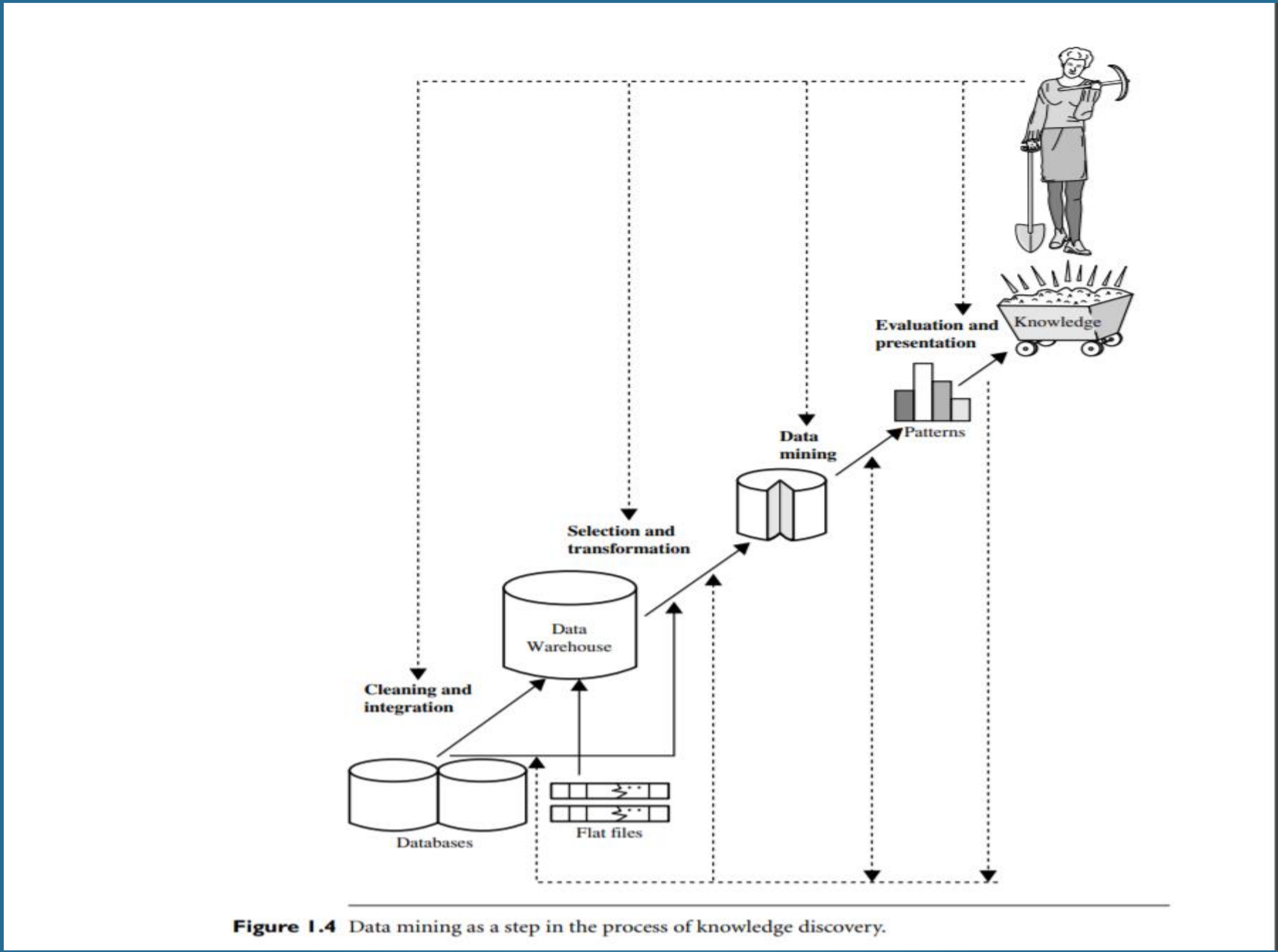


Figure 1.4 Data mining as a step in the process of knowledge discovery.