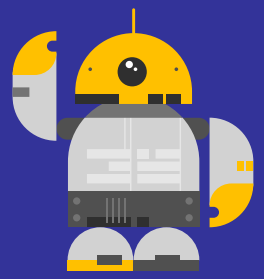


Database Systems

Lecture 3

Dr. T. Akilan

takilan@lakeheadu.ca

This Lecture

- Introduction to data mining
- Data
- Database
- Data warehouse
- Recap - data mining challenges
- Applications
- Summary
- Pop quiz
- Then we move onto data and its attributes

Data Mining Challenges

- **Heterogeneous data**
  - Different sources (databases) with different fields
  - Legacy databases with outdated information
  - Noise and missing information
  - User-submitted information of questionable quality

- **Efficiency** and **scalability**
  - Your algorithm can extract knowledge from an encyclopaedia article in 15 seconds!
  - There are more than 6M articles in English Wikipedia (18 Jan. 2021)
  - It will take nearly 3 years to finish ...
  - **So, is it an efficient and scalable algorithm?**

Data Mining Challenges Cont.

- **Outliers**
  - A piece of data that is **very unlike everything else around** it
  - Including it causes **large differences** in the **average statistics**, but excluding it requires **special exception rules**
    - ✓ E.g., how much cash do you have in your pockets?
    - ✓ 10-people random sample: \$10, \$10, \$10, \$10, \$10, \$10, \$10, \$10, \$10, \$**1,000,000**
    - ✓ Average amount people have = \$100,009?
  - **Outlier analysis:**
    - ✓ It may uncover fraudulent usage of credit cards by **detecting purchases of unusually large amounts** for a given account number in comparison to regular charges incurred by the same account.
    - ✓ Usage of the card's geographical location
    - ✓ Any other ideas?

Data Mining Challenges Cont.

- **High dimensionality**
  - Most complete data warehouse have a lot of information (dimensions) about each item
    - E.g., A shopping database can have very detailed data about purchases
    - But "the customer bought 2% milk on special"
    - ✓ Can cause us to discover patterns that are too specific to be useful, or to miss more general patterns
  - **Idea → Dimensionality reduction**
    - ✓ abstracting away some details
    - 2% milk on special = 2% milk = milk = dairy products = groceries?
    - i.e., from multidimensional space to lower dimensional space

Data Mining Challenges Cont.

- **Handling uncertainty, noise, or incompleteness of data**
  - Data often contain noise, errors, exceptions, or uncertainty, or are incomplete.
  - Errors and noise may **confuse the data mining process**, leading to the derivation of **erroneous patterns.**
  - Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.
- **Background knowledge**
  - Some patterns are obvious from the data for us, because of our background knowledge
    - ✓E.g., we know a text document mentioning "wall street" is probably about finance
    - ✓Because we know the New York Stock Exchange (NYSE) - a major financial institution is on Wall Street
  - **How to include such knowledge into a database system?**

Data Mining Challenges Cont.

- **Evaluating** the knowledge
  - One can mine tremendous amount of patterns
  - How to know which ones are good or bad?
  - Different evaluation metrics for different patterns and applications
    - ✓ Predictive, coverage, statistical measures (precision, recall, accuracy, f-measure), computational complexity, etc.

Applications

- **Basket data analysis** and **targeted marketing**
  - Given a database of customers with demographic information, location, and past purchase behaviour
  - Determine the profile of the most profitable customers
  - Tailor advertisement campaigns to attract and retain these customers

- **Fraud detection**
  - Automobile Insurance Bureau of Massachusetts had a database of insurance claims, including over **60 attributes** such as claimant, type of accident, type of injury/treatment, and expert opinion of **real vs. fraud**
  - **Dimension reduction** methods used to obtain weighted variables, then identified subsets of **characteristics strongly correlated with fraud**

Applications

- **Web page analysis:**
  - Page ranking, for example, Google search engine results (e.g., using BFS)
  - Recommender systems (Amazon)
  - Clicks-to-Customers
    - ✓ 50% of Dell's customers order their computer through the web, but 0.5% of visitors of Dell's web page become customers
    - ✓ Dell has navigation history of visitors through their site
    - ✓ Cluster customers through their **click sequences**, and design web pages **to maximize the number of customers**

- **Biological** and **medical data analysis:**
  - classification, cluster analysis, biological sequence analysis, biological network analysis

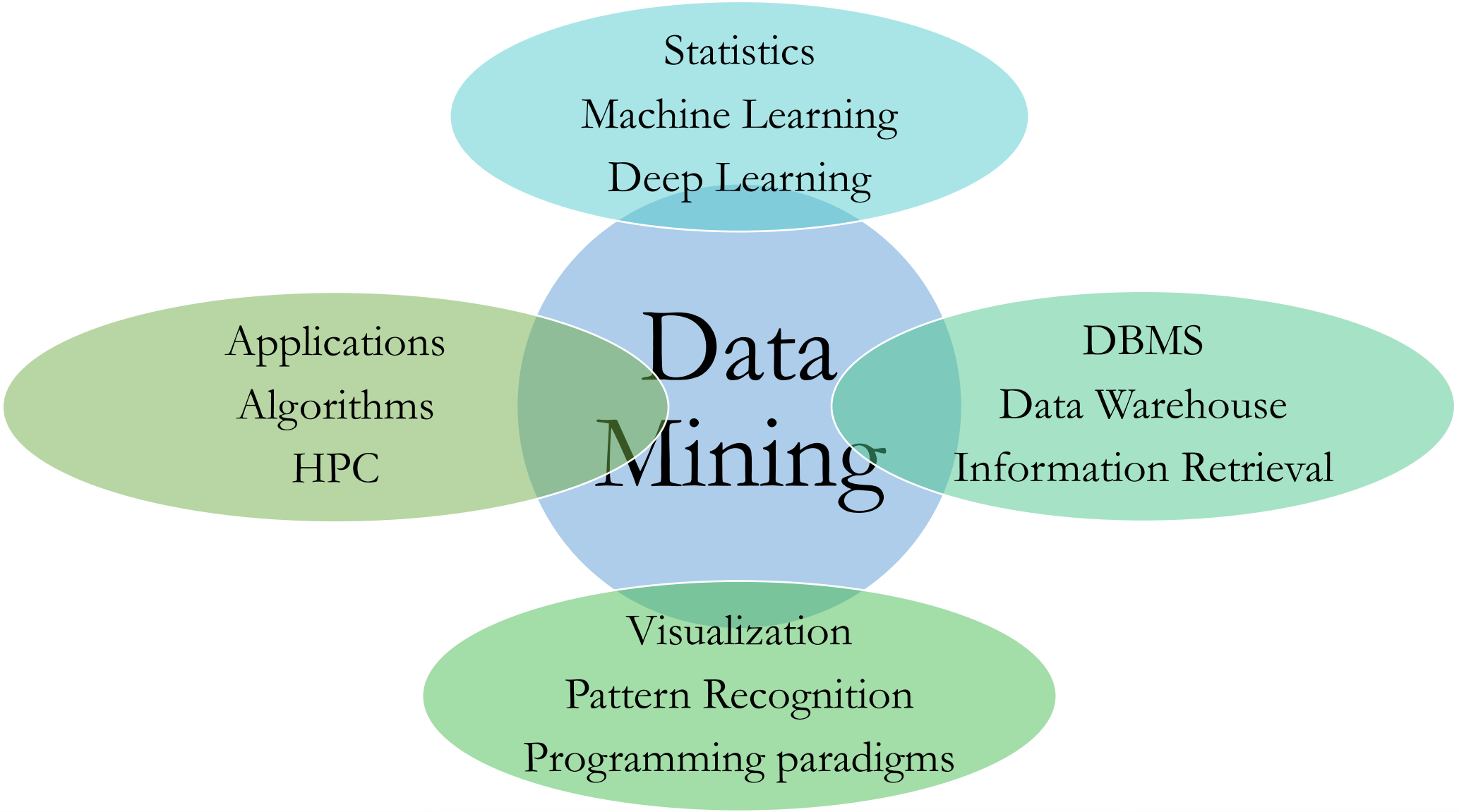
- Engineering research and development (Watson)

Summary

- Data (collected and generated) and individual databases are being consolidated into massive data warehouses
  - Getting knowledge from this massive amount of data is a challenge

- Data mining: extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns from huge amount of data
  - Different types of data
  - Different patterns of interest
  - Different applications
  - Different challenges

Summary Cont.



Summary Cont.

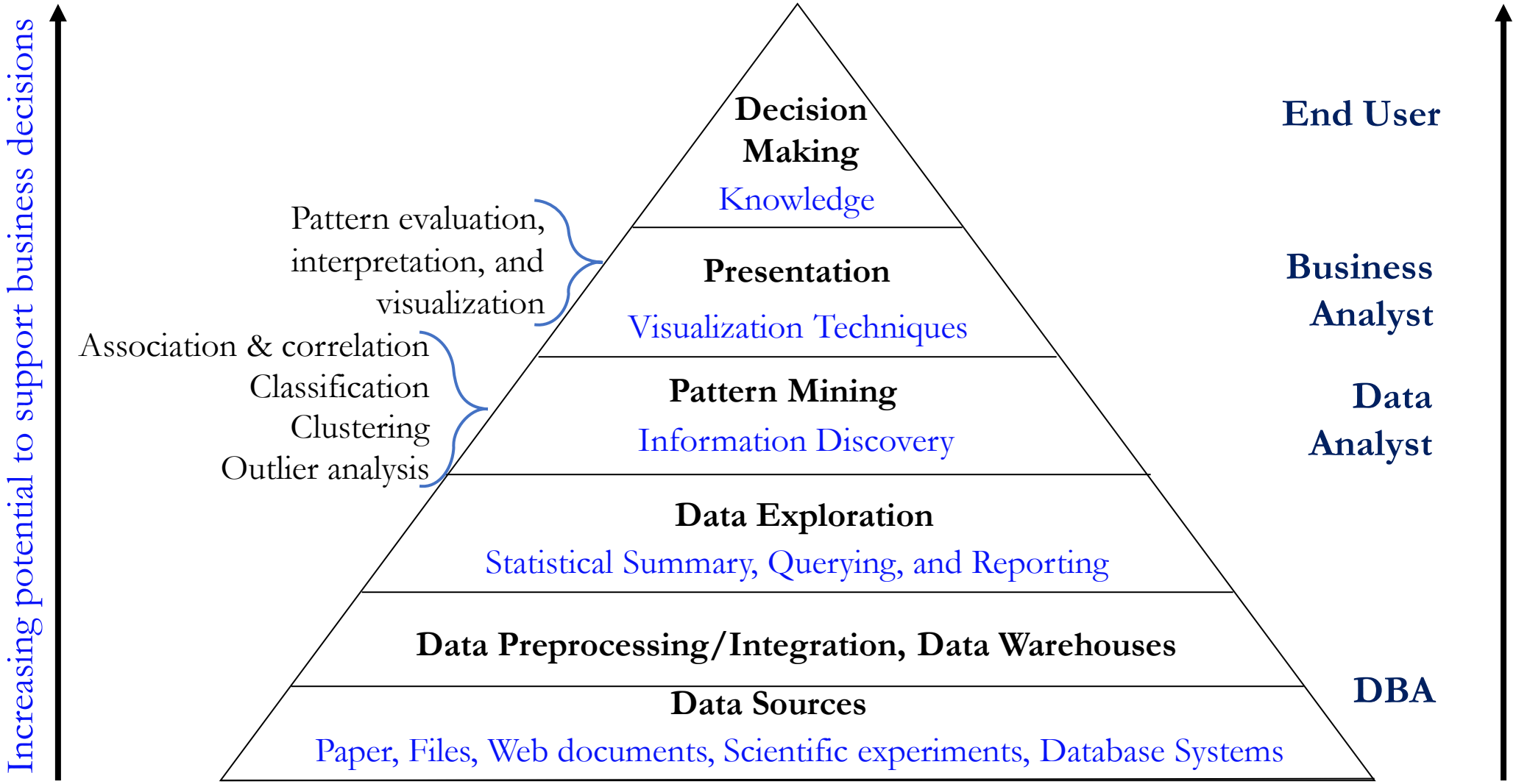


Image: Data Mining: Concepts and Techniques

Exercises

- How is a data warehouse different from a database? How are they similar?

- Describe three challenges to data mining.