15CSE401 MLDM

Data Mining: Tasks and Challenges

Lecture 2
18 Jul 2020

Today's plan

- Review data mining definitions
- Data mining as Knowledge discovery process
- Data mining tasks
- Challenges

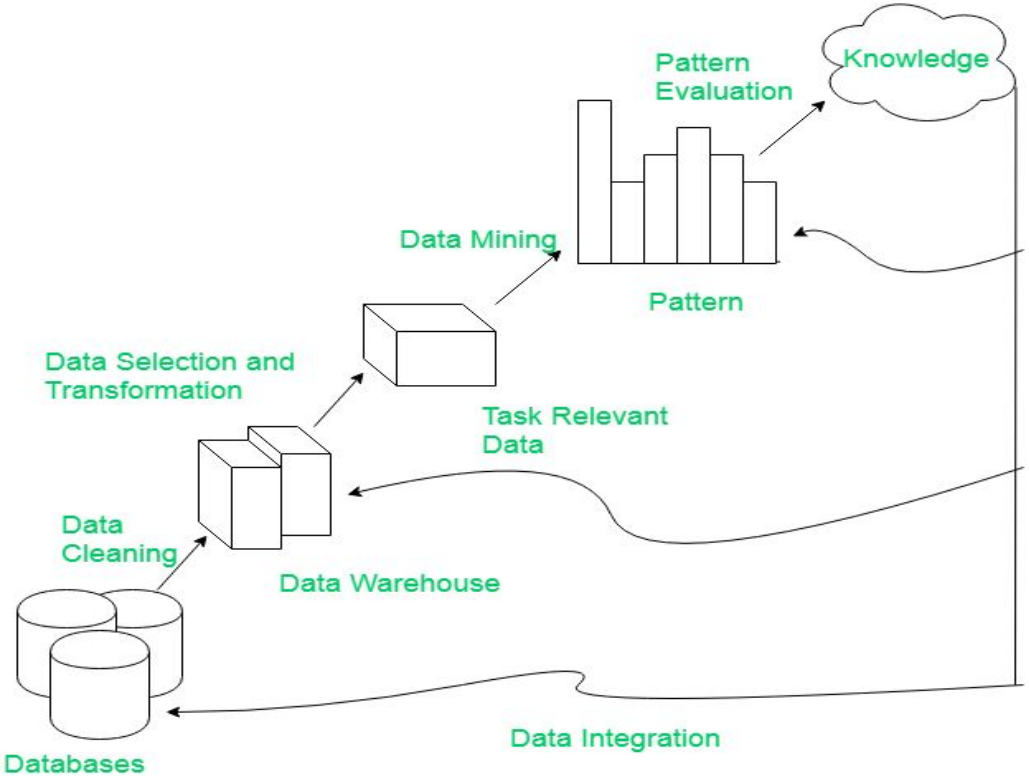
- Ref: Ch-1, Han et al, 'Data Mining: concepts and techniques'

- Quiz
- Discussion

What is data mining?

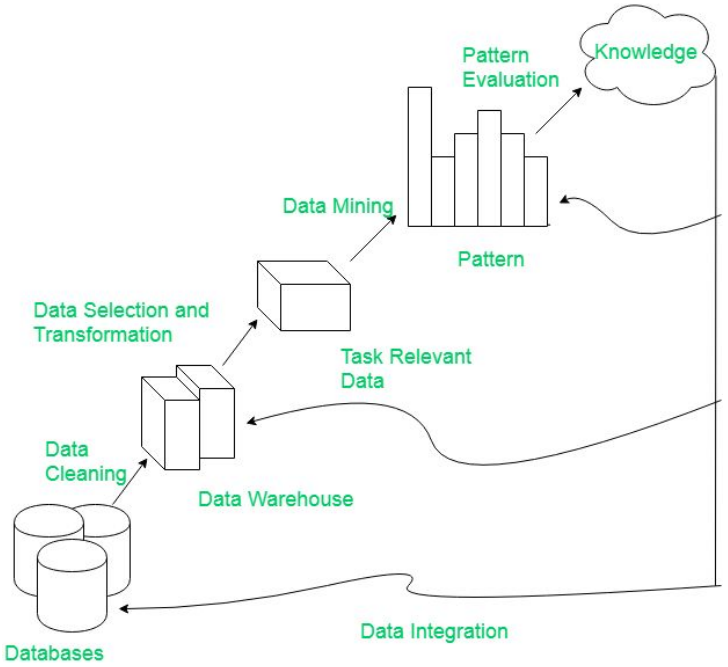
- Automatic extraction of interesting patterns from massive dataset
- These patterns are non-trivial, implicit and previously unknown
- Data mining turns data into knowledge
  - So a Knowledge Discovery from Database (KDD) process
- It could be used as part of
  - Automatic query processing
  - Expert systems

Data mining: a step in KDD process



Data mining: a step in KDD process

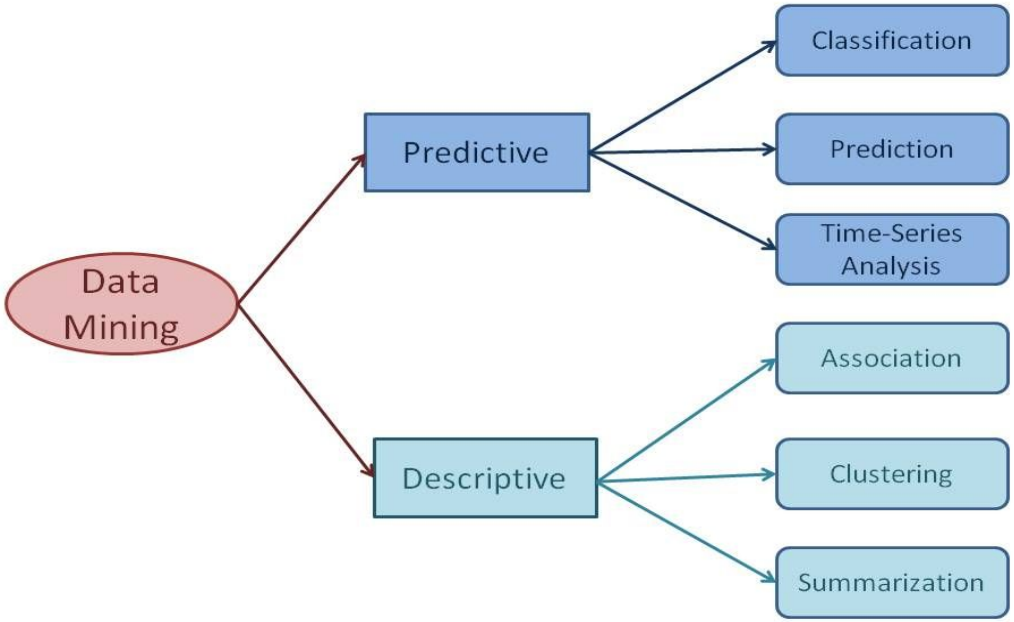
- **Data cleaning** - missing values, noise
- **Data integration**- warehouses and other sources
- **Data selection**- relevance to analysis/ task in hand
- **Data transformation**- mapping and coding
- **Data mining** - patterns identification
- **Pattern evaluation** - measure of interest
- **Knowledge representation** - visualization



What kinds of data can be mined?

- **Databases**
  - Relational databases (tables)
  - Entity Relation model represents the data semantics
  - NoSQL databases - Graph databases
- **Data warehouses**
  - Repository collected from multiple sources
  - Eg. company have branches at different locations
  - Cleaning → integration → transformation → loading → periodical refreshment
- **Transactional Data**
  - Customer transactions (purchase/booking/cancellation..)
- **Other sources**
  - Genome, Images, Text, Social media, Web.....

Data Mining Functionalities

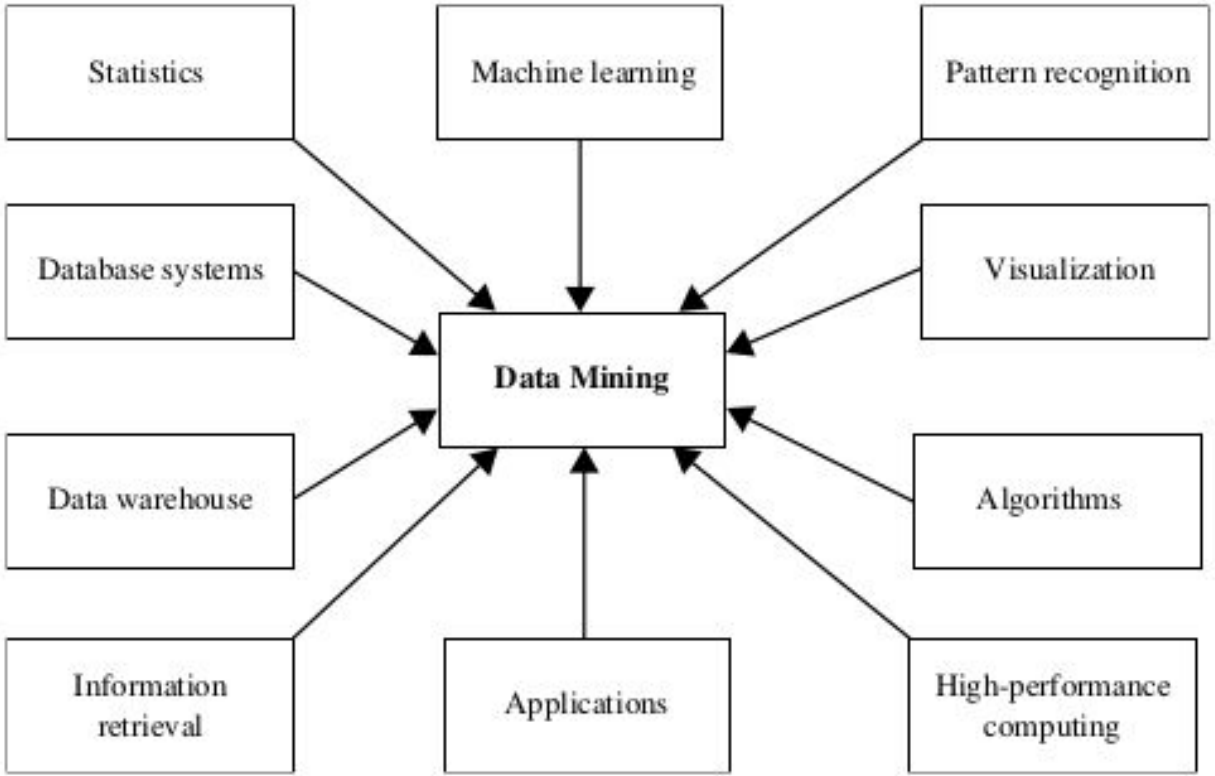


- Data **characterization**
  - Summarizing data of the target class
- Data **discrimination**
  - Comparison of target class with one or more set of comparative classes

Are all patterns are interesting?

- A pattern is **interesting** if
  - **Easy** to understand
  - **Valid** on new data (previously unseen)
  - Potentially **useful**
  - **Novel**
- **Interestingness** can be measured by:
  - Support - how frequently the pattern occur
  - Confidence- how meaningful the pattern
  - Belief (expectedness) - patterns correlation with user's belief
  - Completeness: can cover all interesting patterns?

Data Mining: Confluence of Multiple Disciplines



Data Mining: Challenges

- **Mining Technology**
  - **Diversity** of data (sources)
  - **Performance:**
    - efficiency, effectiveness, scalability
  - **Subjectivity** (of interestingness)
    - Evaluation of patterns
  - Incorporation of **background knowledge**
  - **Noise** and **incomplete data**
  - **Scalable** with **Parallel and distributed** techniques
    - Algorithms should support parallel/distributed technologies
  - **Knowledge fusion**
    - Integrating knowledges

Data Mining: Challenges

- **User interaction**
  - Need flexible interface and exploratory environment
  - Should be dynamic
  - Flexible query languages
  - Presentation and visualization of results
- **Data mining and society**
  - How can we use for benefit of society?
  - How can we guard its misuse?
  - Privacy-preserving data mining

Interesting discussion on “Does data mining affect Our society and polity” (optional)
https://www.youtube.com/watch?v=vaZHIyxonOE

Summary

- Data mining - discovering interesting patterns from massive data
- It is a knowledge discovery process
- Data sources can be databases, warehouses, transaction history, web, genome, etc
- Functionality- Predictive and descriptive
- Evaluation of patterns
- Challenges - Technological, user interface, social and ethical

- Quiz

Next

- Review of Machine learning concepts
  - Some basic concepts of machine learning