

# Introduction to Data Mining

## Lecture1 Introduction

Jun Huang

Anhui University of Technology

Spring 2018

huangjun_cs@163.com

# Welcome



- Jun Huang
  - Technology of Computer Application, Ph.D
  - University of Chinese Academy of Sciences
  - Research Interests: Machine Learning, Data Mining, Multimedia Content Analysis, etc.
- Office: Room 227, Yifu Building
- Phone: 17353766628
- Wechat

# Syllabus(Tentative)



- Introduction
- Data pre-processing
- Classification
- Clustering
- Association rules
- Advanced Topics in Data Mining and Machine Learning
  - Multi-Label Classification
  - Multi-View/Modal Learning
  - Multiple Clustering
  - Paralleled and Distributed Computing

# Introduction



- Textbook and references
- Prerequisites
- Grading scheme
- What is data mining?
- Tasks of data mining
- Procedure of data mining

# Textbook and References



- Textbook
  - The Top 10 Algorithms in Data Mining. Xindong Wu, Kumar Vipin. Chinese Edition (translated by Wenbo Li and Suyan Wu), Tsinghua University Press.
  - Data Mining, Concepts and Techniques. Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2011(Third Edition).
- References
  - Research paper. To be announced in class.

# Prerequisites



- Data structure
- Algorithm
- Database
- Programming skills:Python, Java, Matlab,.etc.
- A mini survey
  - How many people were major in computer science?
  - How many people took machine learning courses before?
  - How many people took database courses before?

# Grading Scheme



- Assignments (20%)
- Course Project (30%)
  - One Project
  - Develop an algorithm and hand in a project report
  - To be evaluated in technical innovation, performance, thoroughness of the work, clarity of presentation
- Final Exame (50%)

# About the Project



- Choose a topic from a list of selected topics
- Read through some related researh papers and fully understand them
- Implement and experimentally evaluate the major method
- Identify advantages and disadvantages
- Improve the method in effectiveness or efficiency, implement and experimentally evaluate your improvement(plus)
- Write a technical report and give presentation

# How to do a good project



- Start early
  - It takes time to understand and think
- Discuss with me
  - Maybe I can give some suggestions or ideas
- Implement concretely
  - Understand the advantages and disadvantages
- Think creatively

# Why take this course?



- Data Mining is so hot
- Solve many interesting problems in real applications, e.g. business management,WWW,science exploration
- Turn raw data into knowledge
- Promising in research of many disciplines
- Data miners'job market: many well-paid positions

# Objectives of this course



- Introduce the motivation of data mining
- Outline principles, major algorithms
- Introduce applications
- Introduce advanced topics
- Enhance independent research capability

# Policies



- Students are expected to attend all classes
- No late homework will be accepted
- All work must be efforts of your own
- No Plagiarism!

# What motivated data mining?



- The explosive growth of data
- Data collection and data availability
  - Computer hardware & software develop dramaticlly
  - The amount of data collected and stored doubles/triples per year vs. CPU speed increases 15% per year(till 2003)
- Many types of databases
  - Oject-oriented, spatial, temporal, time-series, text, multimedia, Web

# What motivated data mining?
## Business World



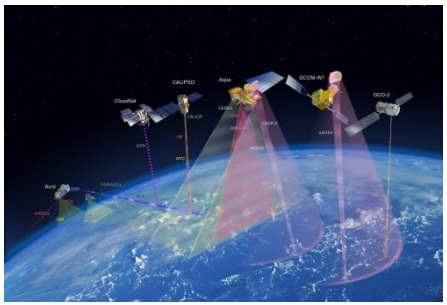
- Tremendous of data being collected and stored
  - E-commerce
  - Transactions
  - Stocks
  - Credit card transactions
- Strong competitive pressure to extract and use knowledge hidden in the data to provide customized CRM

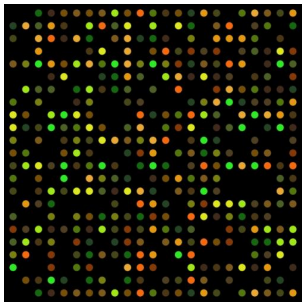


# What motivated data mining?
## Scientific World



- Tremendous of data being collected and stored
  - Remote sensing
  - Bioinformatics(Microarrays)
  - Scientific simulation
- Scientists need strong data analysis to assist research, such as classification, segmetation, etc.





# What motivated data mining?



- We are drowing in data, but starving for knowledge
  - Data rich, knowledge poor
  - Decision makers, domain experts have biases or errors
- Automated analysis of massive data sets

# What is data mining?

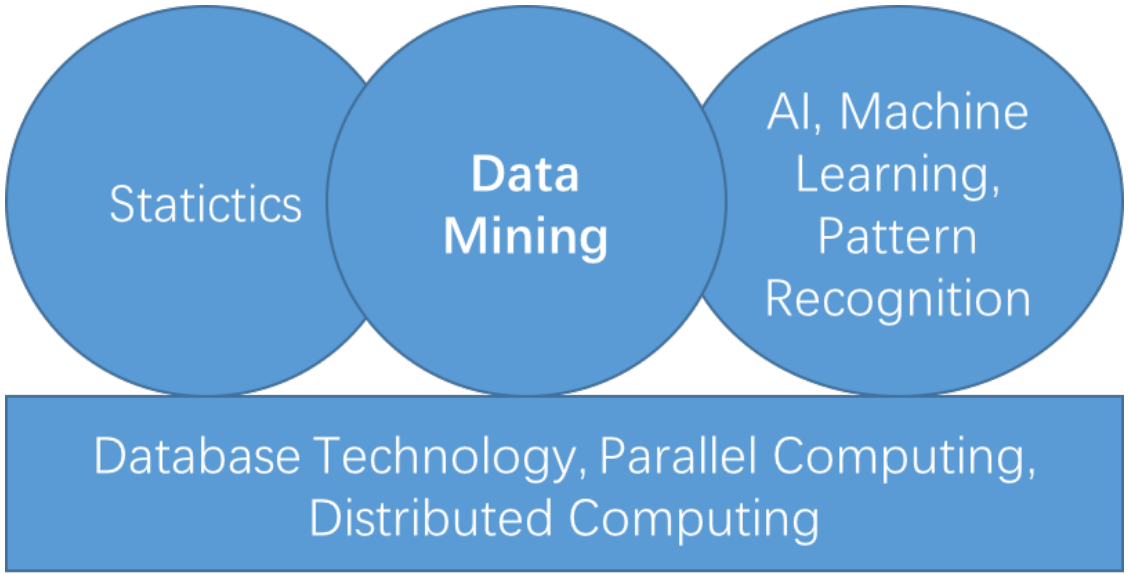


- **Data Mining**: Discover valid, novel, usefull, and understandable patterns in massive database

# What is data mining?

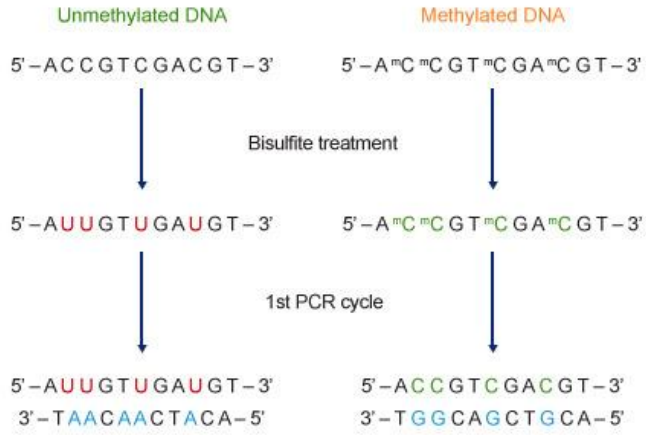


- Cross disciplines
  - Databases
  - Machine Learning: decision tree, Bayesian Classifier, etc.
  - Statistics: regression, etc.
  - Neural networks



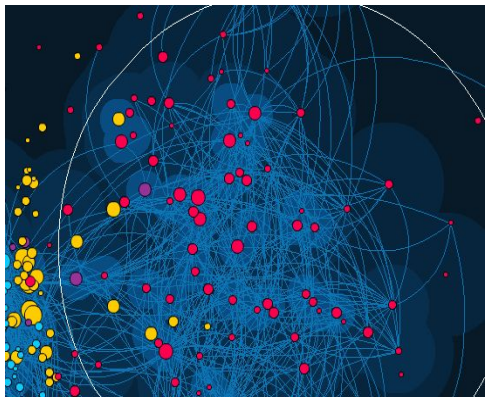
# Why not Traditional Data Analysis?

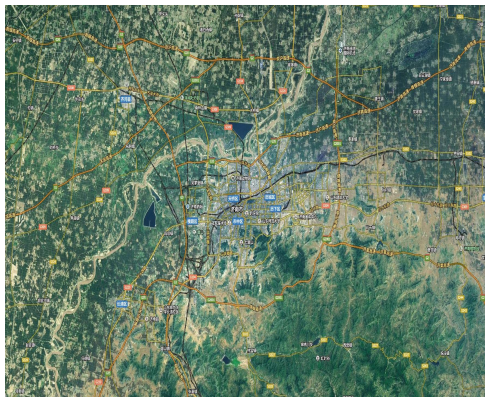
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions



# Why not traditional data analysis?

- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Graphs, social networks
  - Spatial, temporal, multimedia, text, and web data
- New and sophisticated applications





# Why not traditional data analysis?



| **Database** | **Data Mining** |
|---|---|
| Storage-oriented | Discover knowledge from data in databases |
| Provide simple queries | |
| **Data warehouse** | **Data Mining** |
| Subject-oriented | Advanced data analysis tools |
| A multidimensional view of data | |
| Operations to access summarized data | |
| **Statistical algorithms** | **Data Mining** |
| Based on many hypothesis | Less hypothesis;Find patterns in large number of samples;Abnomal patterns |
| Find patterns in small number of samples | |

# Characteristics of Data Mining



- Massive dataset
- Automatically searching for interesting patterns from history data
- Fast
- Scalable
- Update easily
- Practical
- Decision support

# Exercises



1. Could you present an application of data mining in business domain?
2. Could you present an application of data mining in scientific domain?

# What kinds of Patterns?

## Association rules

- **Association rules**: Detect sets of attributes or items that frequently co-occur in many database records and rules among them

On Thursdays, during 4-11pm customers often purchase diapers and beers together!

# Ex.1 Market basket analysis and management

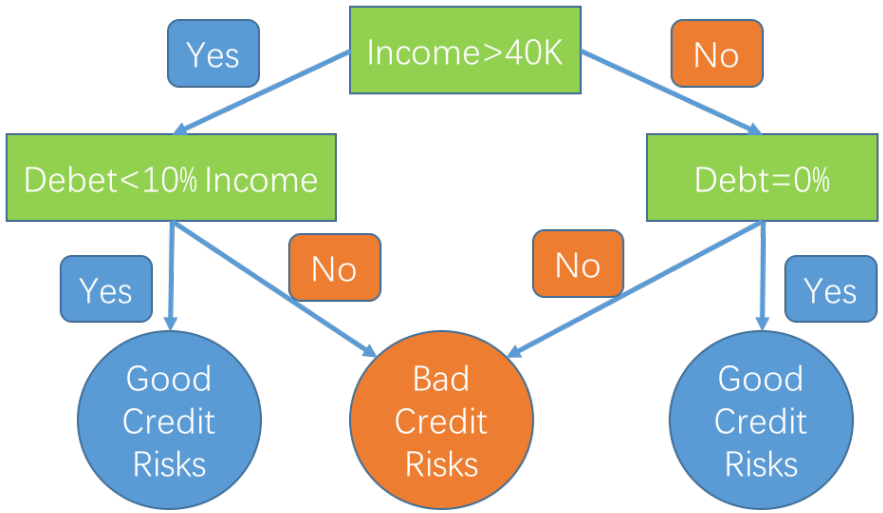


- Where does the data come from?
  - supermarket transactions, membership cards, discount coupons, customer complaint calls
- Cross-marketing analysis
  - What products were often purchased together? Purchase recommendation, cross selling
  - What are the subsequent purchases after buying a give product?
- Target-marketing
  - What types of customers buy what products
- Catalog design


# What kinds of Patterns?

## Classification

- **Classification**: Build a model of classes on training dataset, and then, assign a new record to one of several predefined classes




- **Decision Tree**
- **Rule1: if Income<=40K and Debt=0 then good**
- **Rule2: if Income>40K and Debt < 10% of Income then good**

# Ex.2 Credit Scoring

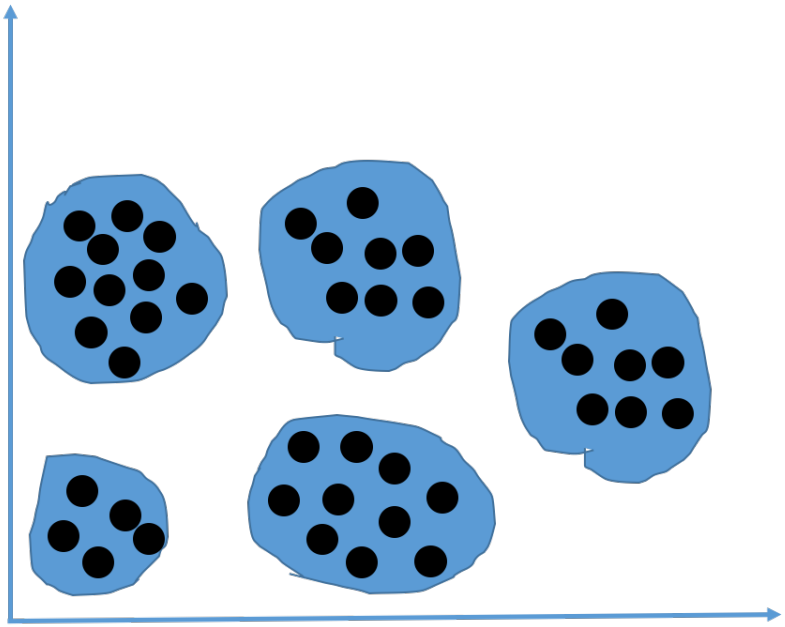


- Where does the data come from?
  - credit card transactions, credit card payments, loan payments, demographic data
- Predict the probability to bankrupt or charge-off
- Reduce the credit risk to the banks
- Increase the profitability of the banks

# What kinds of Patterns?
## Clustering



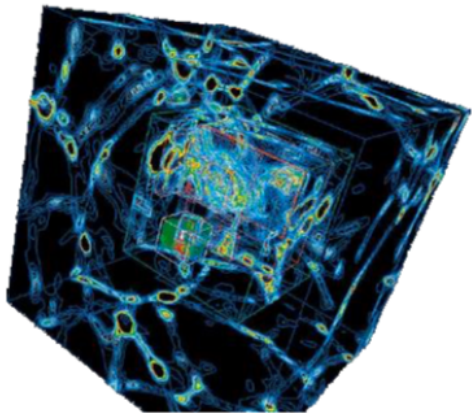
- **Clustering**: Partition the dataset into groups such that elements in a group hae lower inter-group similarity and higher intra-group similarity



# Ex.3 Scientific simulation



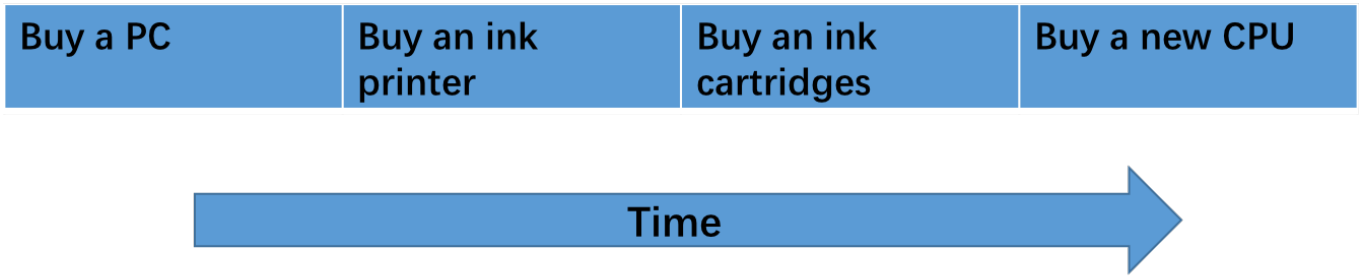
- Cosmological simulation
  - Simulate the formation of the galaxy
  - Enormous particles at each evoluation stage, beyond the capability of human being to analyze



# What kinds of Patterns?
## Sequence mining



- **Sequence mining**: Give a set of sequences, find the complete set of frequent subsequences

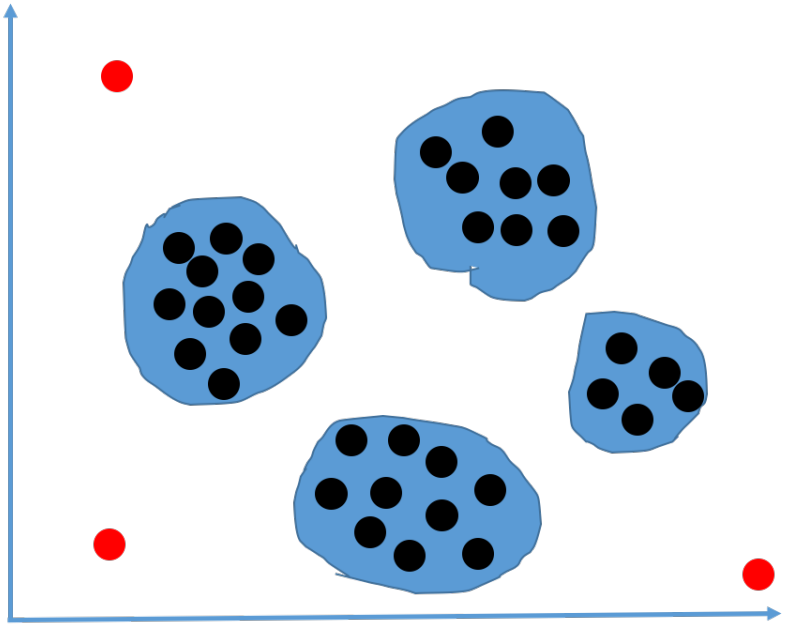



**Marking strategy: recommend a new CPU for the customer 9 months after his first purchase**

# What kinds of Patterns?
## Anomaly detection



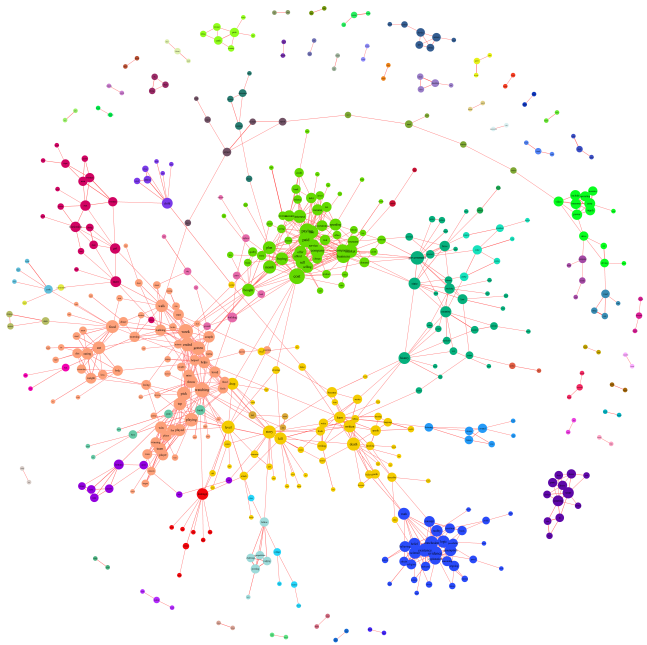
- **Anomaly detection**: Give a set of $n$ objects, and $k$, the number of expected anomolies, find the top $k$ objects that are considerably dissimilar or inconsistent with the remaining data



# What kinds of Patterns?
## Community Analysis



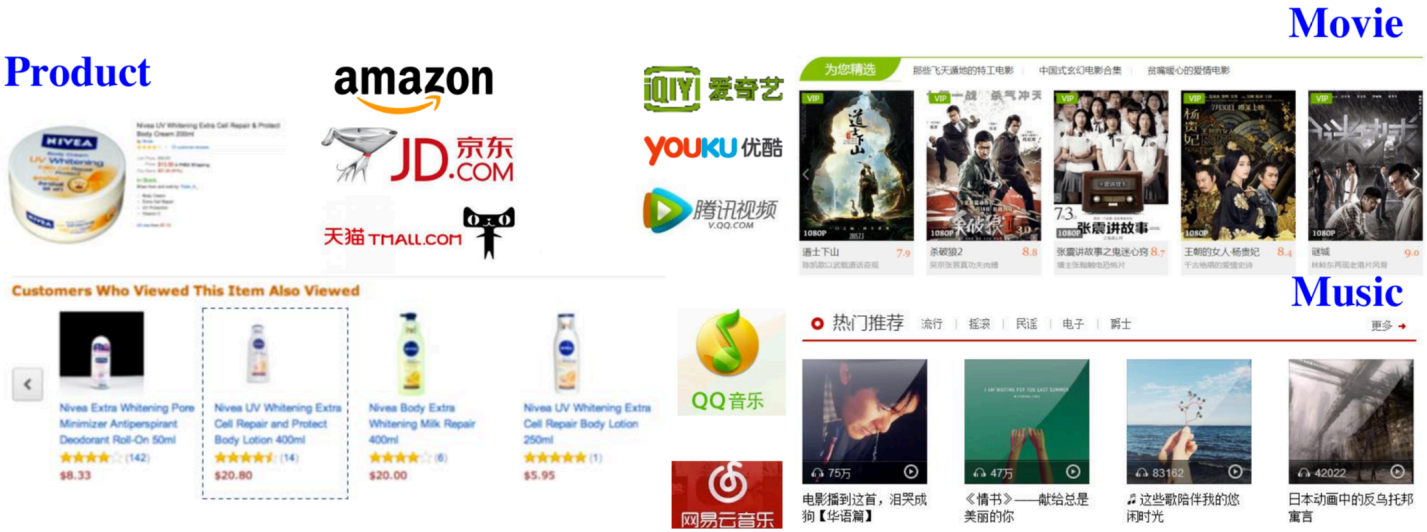
- **Community Analysis**: In social media mining, analyzing communities is essential.
  - How to detect communities?
  - How do communities evolve and how to study evolving communities?
  - How to evaluate deteted communities?



# What kinds of Patterns?

## Recommender systems

- **Recommender systems**:
  - Recommend products that would be interesting to individuals
  - Build a function, $f: U \times I \rightarrow \mathbb{R}$, for user set $U$ and item set $I$



# Exercises



1. Please present an example where data mining is crucial to the success of the business. What data mining techniques are used? (What kinds of patterns are mined?)
2. Can you describe other possible kind of knowledge that needs to be discovered by data mining methods but not been mentioned in class yet?

# On What Kinds of Data?



1. Database-oriented data sets and applications
   - Relational database, data warehouse, transactional database
2. Advanced database applications
   - Data streams
   - Spatial data
   - Text database
   - Multimedia data
   - Time-series
   - Bio-medical data
   - Network traffic data

# Relational Databases



- Structured data
  - Table-records-attributes
  - Accessed by queries, SQL
- Online transactional processing (OLTP)
  - Insert a student "Jun Huang" into class "Introduction to Data Mining", Spring 2018

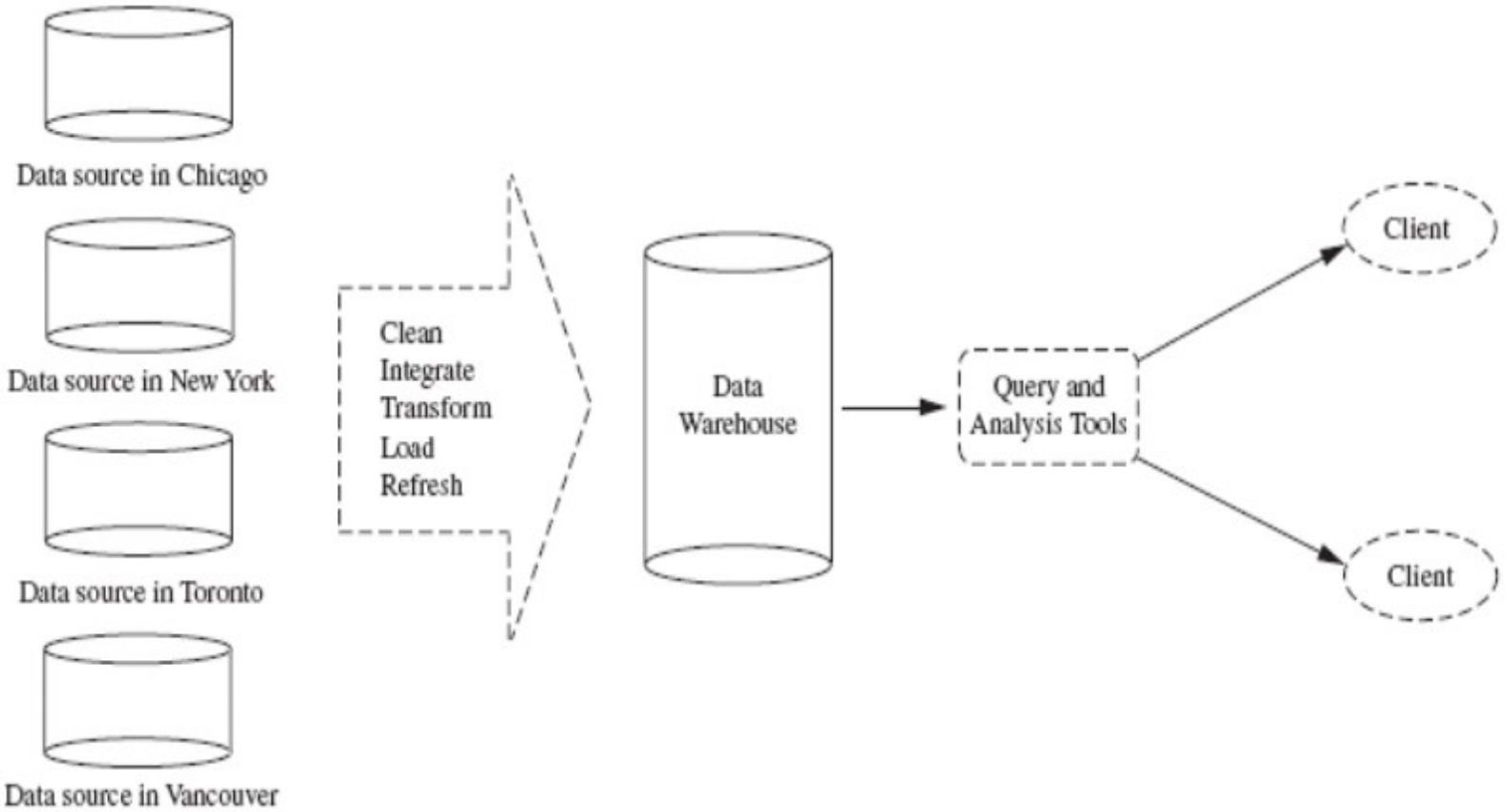
| **Name** | **Time** | **Course** | **Score** | **Room** |
|---|---|---|---|---|
| Jun Huang | Spring 2018 | Introduction to Data Mining | 90 | 002 |
| Lucy | Spring 2017 | Introduction to Data Mining | 70 | 002 |
| Merlisa | Spring 2017 | Math | 80 | 001 |
| Jack | Fall 2016 | Machine Learning | 80 | 001 |

# Data Warehouse



- A subject-oriented, integrated, cleaned collection of data in support of management' s decision making process
- Data from multiple databases
- Consistency checking in data warehouses
- Data warehouses can answer OLAP queries efficiently
  - Online analytical processing (OLAP)
  - Find the average class score of "Jun Huang"in the last 3 years, grouped by semesters
- Many patterns are summarization of data
  - Roll-up, drill-down

# Data Warehouse



# Tansactional Databases



- $I = \{x_1, x_2, ..., x_n\}$ is the set of items
- An itemset is a subset of $I$
- A transaction is a tuple $(tid, X)$
  - $tid$: Transaction Id
  - $X$: Itemset
- A transactional database is a set of transactions

| **Tid** | **Itemset** |
|---|---|
| T100 | milk, bread, beer, diaper |
| T200 | beer, cook, fish, potato, orange, apple |
| ... | ... |

# Spatial Data



- Spatial Information
  - Geographical database
  - VLSI chip design databases
  - Satellite/remote sensing image database
  - Medical image database
- Spatial Patterns
  - Find characteristics of homes near a give location
  - Change in trend of metropolitan poverty rates based on distances from major highways

| 编号 | 中心 | 正右方 | 右上方 | 面积 |
|---|---|---|---|---|
| 1 | 居民地 | 绿地 | 水体 | 120 |
| 2 | 绿地 | 水体 | 水体 | 40 |
| 3 | 水体 | 居民地 | 居民地 | 610 |
| ... | ... | ... | ... | ... |

# Time Series

- A sequence of values that change over time
  - Sequences of stock price at every 5 minutes
  - Daily temperature
  - Power supply
  - Electrocardiogram
- Typical Operations
  - Similarity search
  - Trend analysis
  - Periodic patter discovery


# Text Databases & Multimedia Databases



- HTML web documents
- XML documents
- Digital libraries
- Annotated multimedia databases
  - Image, audio and video data
- Typical Operations
  - Similarity-based pattern maching
  - Image classification

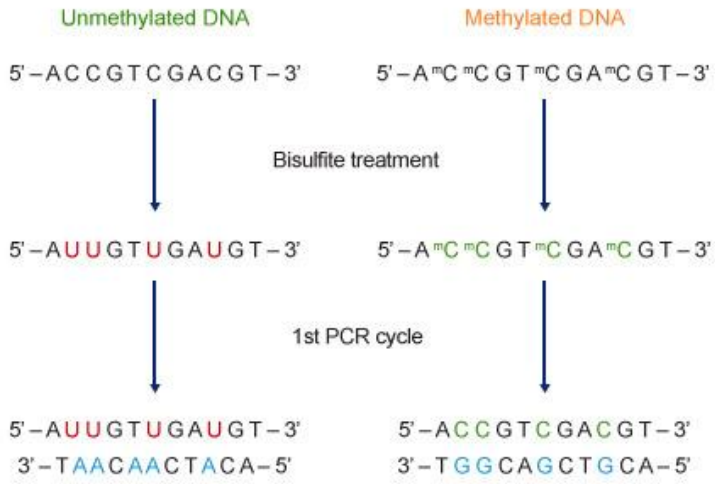

# Data Streams



- Data in the form of continuous arrival in multiple, rapid, time-varying, possibly unpredictable and unbounded streams
  - Dynamically changing patterns, high volume, infinite, quick response, no re-scan
- Many applications
  - Stock exchange, network monitoring, telecommunicatioins data management, web application, sensor network, etc.

# Biomedical Data

- Bio-sequences
  - DNA; very long sequences of nucleotideos
  - Similarity search
  - Identify sequential patterns that play roles in various diseases
  - Association analysis: co-occurring gene



# World-Wide Web



- The WWW is huge, widely distributed, global infoormation service center for
  - Information service: news, advertisements, consumer information, financial management, education, goverment, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for data mining
- Challenges
  - Too huge for effectie data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

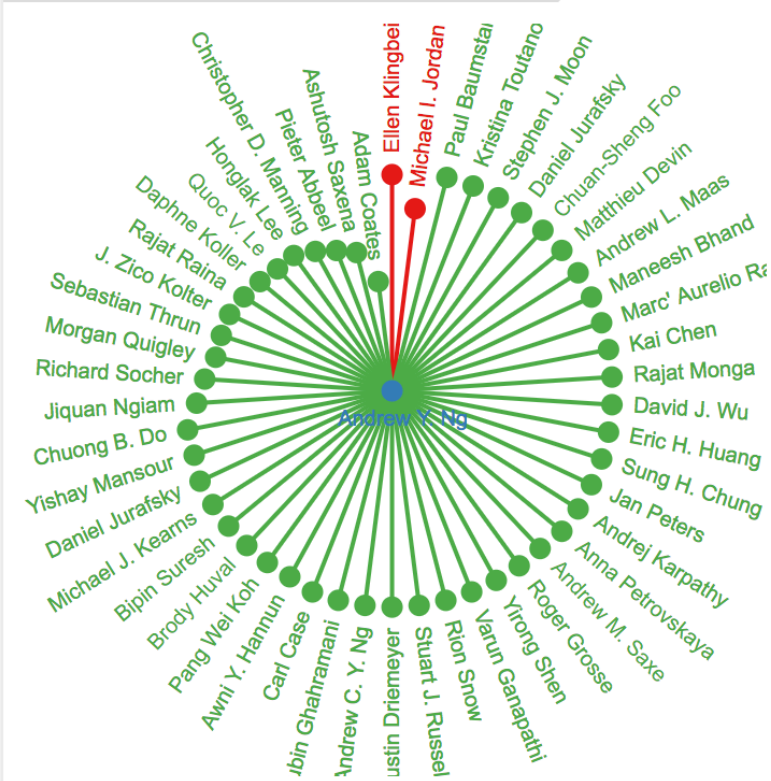
# World-Wide Web

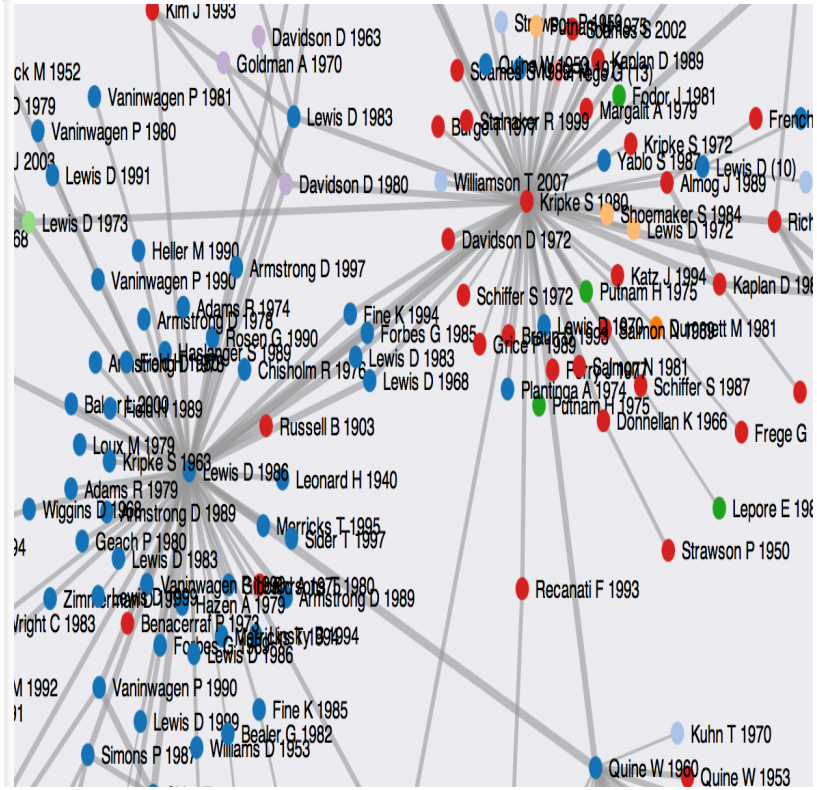


- Web-Usage: Logs and IP package head streams
  - Mining weblog records to discover users accessing patterns of Web pages
- Web Content
  - Extract knowledge from Web documenta, automatic categorization
- Web Structure
  - Identifying interesting graph patterns among different Web pages

# Graph

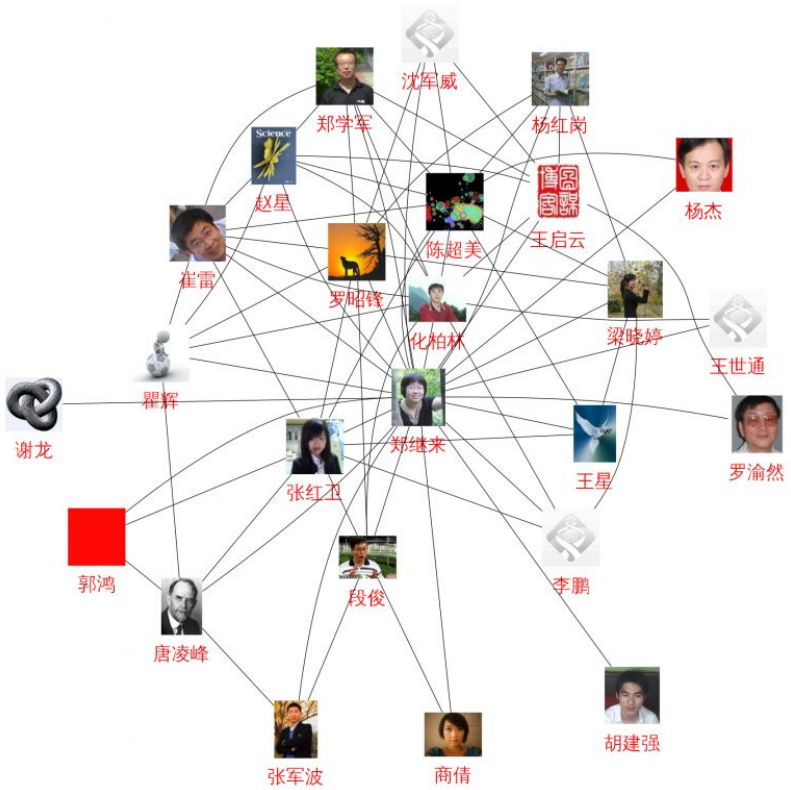
- Citation Graph



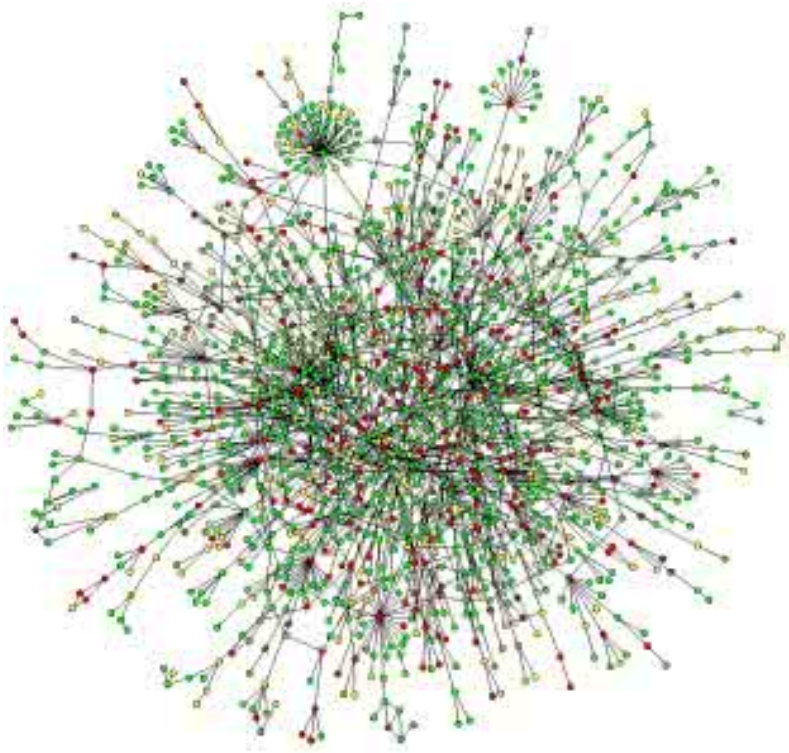
# Graph

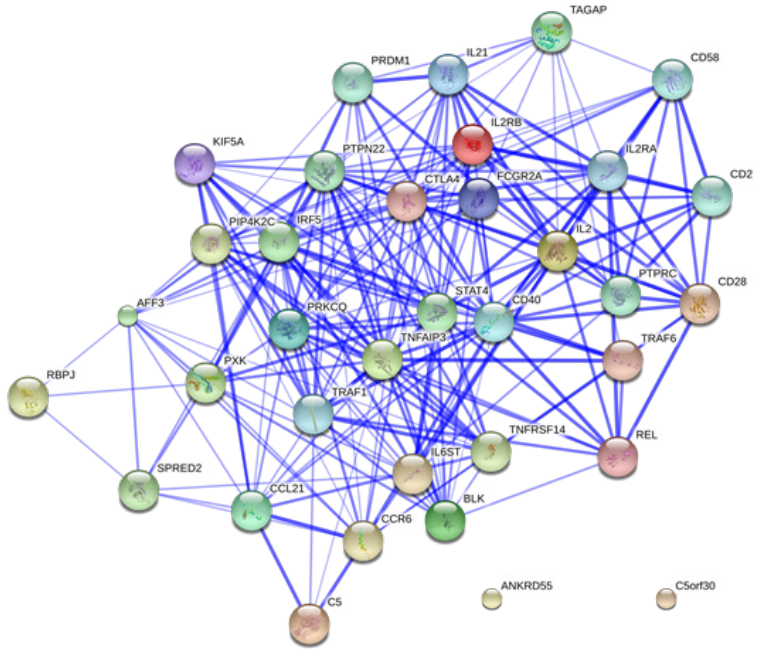
- Friendship Graph

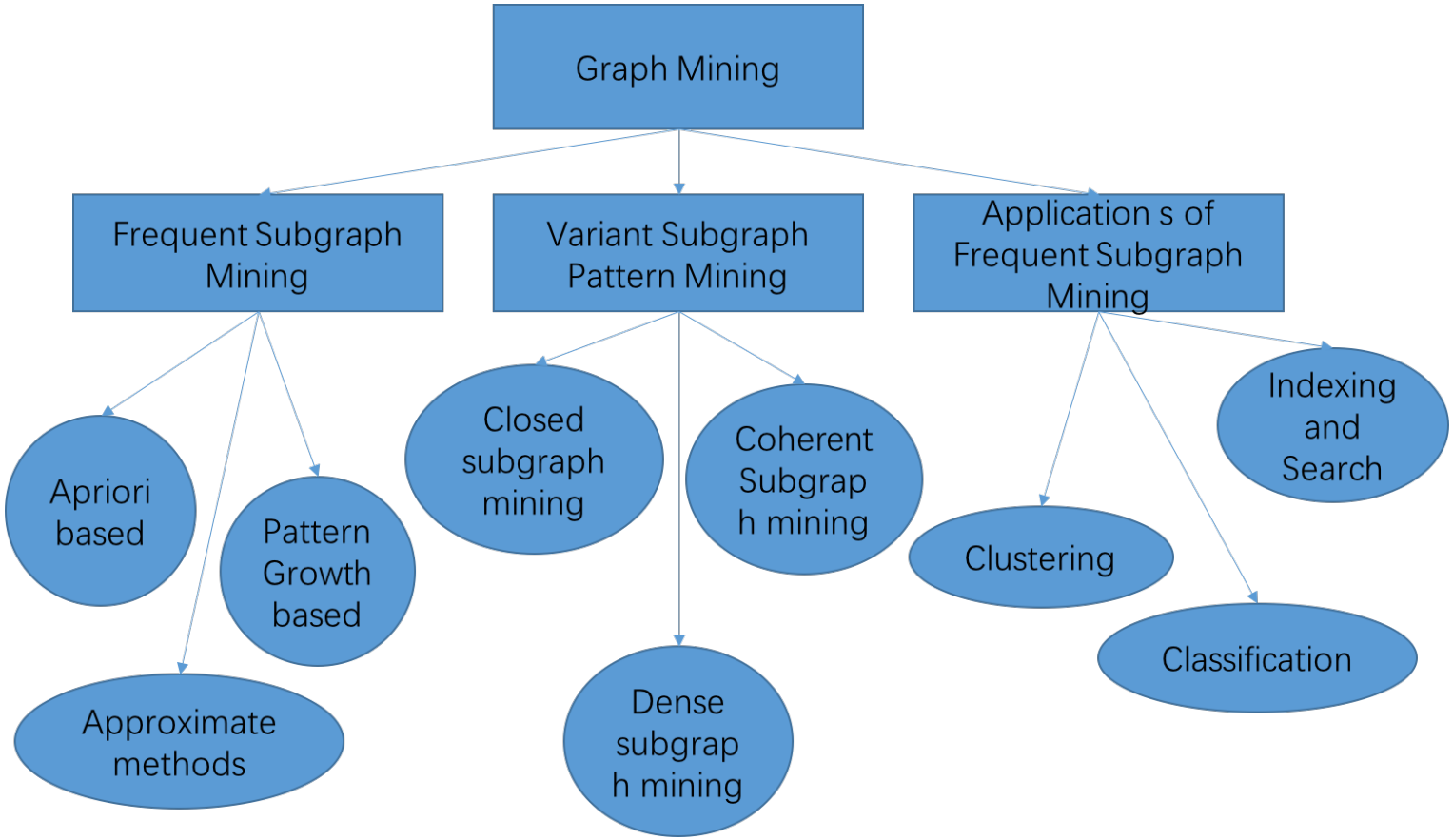


# Graph

- Protein Interaction Graph

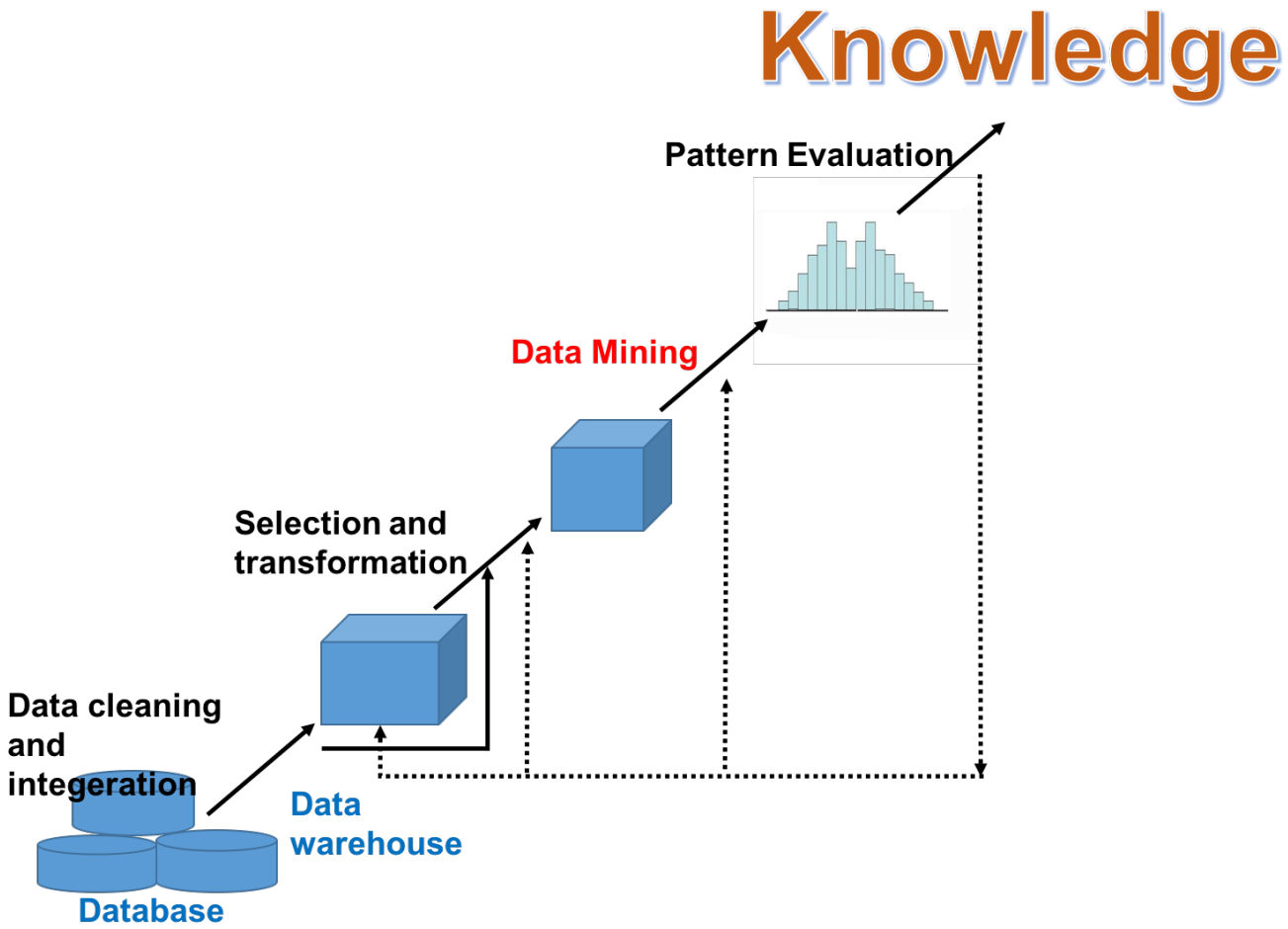






# Knowledge Discovery(KDD) Process

- Data Mining-Core of knowledge discovery process



# Key Steps in KDD Process



- Learning the application domain
  - Relevant prior knowledge and goals of application
- Creating a target data resource
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing the mining algorithm(s) to search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

# Are All the "Discovered" Patterns Interesting?



- Data mining may generate thousands of patterns: Not all of them are interesting
- Interestingness measures
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- Objectie vs. subjective interestingness measures
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

# Find All and Only Interesting Patterns?



- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive searching
- Search for only interesting patterns: An optimization problem - Challenging
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninterestings ones
    - Guide and constrain the discovery process

# Research Issues in Data Mining



- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., Web, graph, bioinformatics, stream
  - Performance: efficiency, effectiveness, and scalability
  - Parallel, distributed and incremental mining methods
  - Pattern evaluation: the interesting problem
  - Handling noise and incomplete data
  - Incorporation of background knowledge
- User Interaction
  - Data mining query languages
  - Expression and visualization of data mining results
- Application and socail impacts
  - Domain-specific data mining
  - Protection of data security, integrity, and privacy

# Top 10 Challenging Problems in Data Mining Research



1. Developing a Unifying Theory of Data Mining
2. Scaling Up for High Dimensional Data and High Speed Data Streams
3. Mining Sequence Data and Time Series Data
4. Mining Complex Knowledge from Complex Data
5. Data Mining in a Network Setting
6. Distributed Data Mining and Mining Multi-agent Data
7. Data Mining for Biological and Environmental Problems
8. Data-Mining-Process Related Problems
9. Security, Privacy and Data Integrity
10. Dealing with Non-static, Unbalanced and Cost-sensitive Data

# Important Resources



- Data mining Conferences
  - KDD, ICDM, WWW, WSDM, SDM, ECML-PKDD, PAKDD
- Database Conferences
  - SIGMOD, VLDB, PODS, ICDE, EDBT, ICDT
- Important Journals
  - IEEE Transactions on Knowledge and Data Engineering (TKDE)
  - Knowledge and Information Systems (KAIS)
  - Data Mining and Knowledge Discovery: An International Journal (DMKD)
  - ACM Data Mining and Knowledge Discovery (TKDD)
- Useful Website
  - https://www.kdnuggets.com/
  - https://www.kaggle.com/