

Data Mining

Ying Liu, Associate Prof., Ph.D

University of Chinese Academy of Sciences

Welcome

■ Instructor: Ying Liu

- Computer Engineering, Ph.D
- Northwestern University, USA
- Research interests
 - data mining, high performance computing, etc.
- Email: yingliu@ucas.ac.cn

Information

- Teaching Assistants
 - Ma, Pengshan, sdxmps@126.com
 - Wang, Zhongya, enriquewang@126.com
 - Yang, Jiajun, 398966925@qq.com
- Class: Monday & Wednesday 8:30 - 10:10, 教1-002
- Website: <http://www2ucas.ac.cn>

Textbook and References

■ Textbook

- Data Mining, Concepts and Techniques. Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2006.

■ References

- Research papers. To be announced in class.

Prerequisites

- Data Structure
- Algorithm
- Database
- Programming: C/C++ (preferred), Java

A Mini Survey

- How many people were major in computer science?
- How many people took machine learning courses before?
- How many people took statistics courses before?
- How many people took database courses before?

Grading Scheme

- Assignments (30%)
 - 2-3 homework assignments
- Course Project (30%)
 - One project. Group project (3-4 students/group)
 - Develop an algorithm and hand in a project report
 - To be evaluated in technical innovation, performance, thoroughness of the work, clarity of presentation
- Final Exam (40%)
 - In class, open book, open notes

About the Project

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

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 pros and cons
- Think creatively

Policies

- Students are expected to attend all classes
- No late homework will be accepted
- All work must be efforts of your own (individual assignment) or of your approved team (group assignment)

No Plagiarism!

Why Take This Course ?

- Data mining is hot

- Solve many interesting problems in real applications, e.g. business management, market analysis, science exploration
- Turn raw data into knowledge
- Promising in research of many disciplines
- Data miners' job market: many well-paid positions (median \$70k in U.S.)

➤ *Data Mining is very useful!*

Syllabus (Tentative)

- Introduction
- Data warehouse
- Data pre-processing
- Association Rules
- Classification
- Clustering
- Applications
- New trends in data mining

Objectives of This Course

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

What Motivated Data Mining?

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

What Motivated Data Mining – Business World

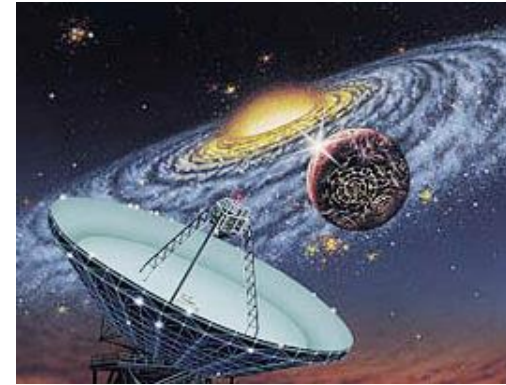
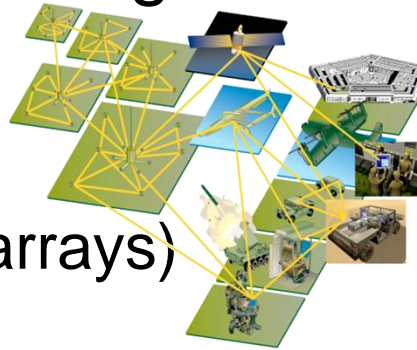
- Tremendous of data being collected and stored
 - E-commerce
 - Transactions
 - Stocks
 - Credit card transactions
- Strong competitive pressure to extract and use the knowledge hidden in the data to provide customized CRM



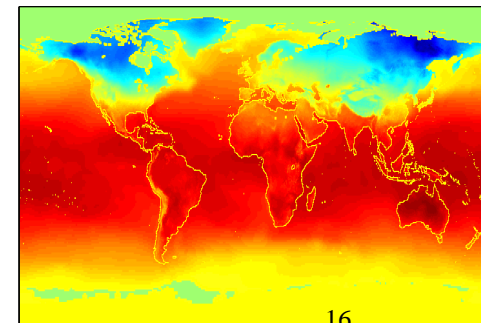
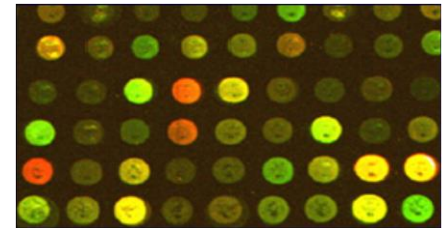
What Motivated Data Mining – Scientific World

- Tremendous of data being collected and stored

- Remote sensing
- Bioinformatics (Microarrays)
- Scientific simulation



- Scientists need strong data analysis to assist research, such as classification, segmentation, etc.



What Motivated Data Mining?

- We are drowning 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?

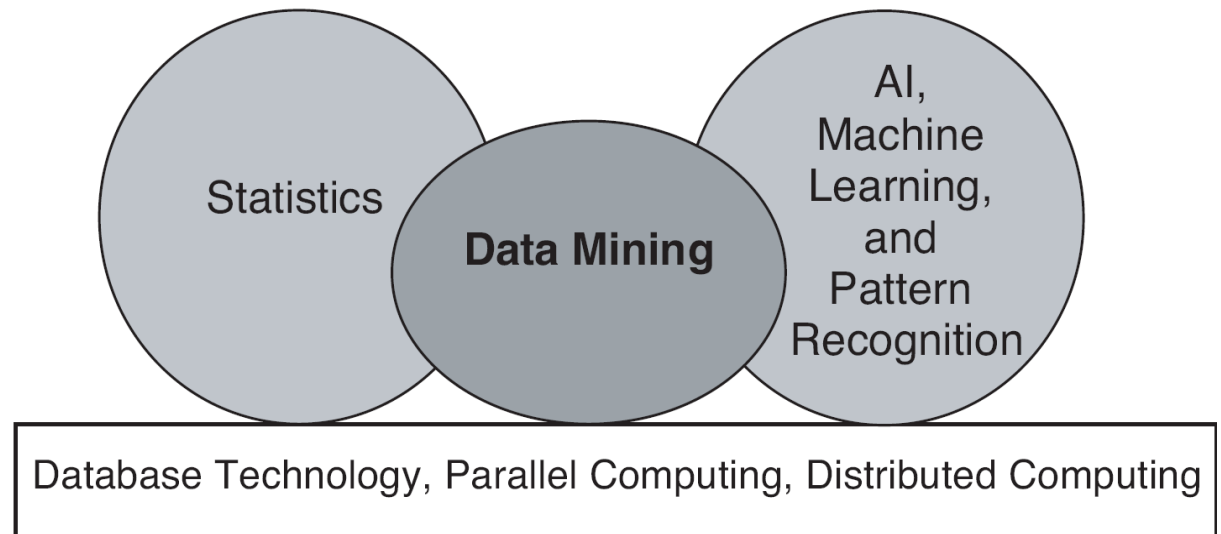
- Data mining — Discover valid, novel, useful, and understandable patterns in massive datasets



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

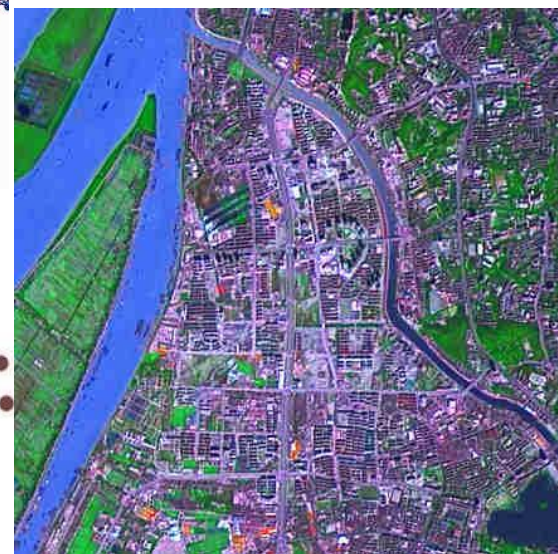
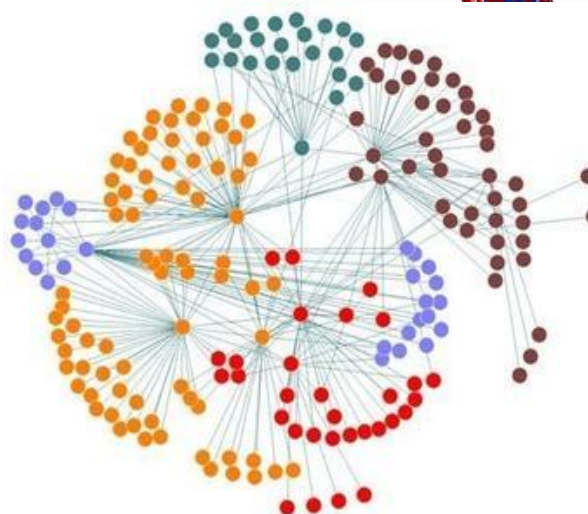


- High-dimensionality of data
 - DNA sequences may have tens of thousands of dimensions

TRFE_CHICK	WHLICLTNLSLBIANVCFAP--PKSVIRICTISSPEEXKCNLRLTODERIS--LTCYQKATLDCIKAIANNEADATSLGGQVFEADLAPINLKPVAEITYEH
TRFE_HUMAN	MRLAYGALLYGAYLQLCLAYP--OKTVRICAIVSEHEATKQDFRHHKSVIPDGGPSYACVKKASTLDCIRAIANNEADAVTLDAQLYIDAFLAPNLKPVAEITYGS
TRFE_XENLA	WFLSLRYVALQLHMLALCLATG--KEXQVRKCVKSNELKXCKLYOTCNKE--IKLSCEYKSNTECESTATGDAICYYQGYKQSLQFYNLKPVAEITYGS
TRFE_RABIT	MRLAQLLACAAQLCLAYT--EXTVRICAVNHKASKCANFRDSMKVLPEDQPIICVKKASTLDCIKAIANNEADAVTLDAQLYHEADLTPNLKPVAEITYGS
TRFE_BOVIN	MSPAYRALLACAYLQLCLADP--ERTVRICITISTHEANICASFRENILRILESQ-PFYSCNKTSTHWDIKAIANNEADAVTLGGQVFEADLAPINLKPVAEITYGS
TRFE_PIG	--YA--OKTVRICTISNDEANICSSPFRENKAYKING-PLYSCKYKSSLDCIKAIANNEADAVTLDAQLYFEADLAPINLKPVAEITYGS
TRFE_HORSE	MRLAIRALLACAYLQLCLA--EDTVRICTVSNHNSKASPFQDWKSIYRAP-PLVACYKRTSTLDCIKAIANNEADAVTLDAQLYFEADLAPINLKPVAEITYGS
TRFE_ANPL	--AP--PKTTVRICTISSAEDKXCNLKHMOQERYT--LSCYQKATLDCIKAIANNEADATSLGGQVFEADLAPINLKPVAEITYGS
TRF1_SALSA	WLLLLSALLDGLATAYAP--AEGIVKVKYKSEDELKCHILANVAEFS--CYRKQGSFEDQAIKGGADATLGGQVITAGLTNYLQLOPIIAEDITYGS
TRF2_SALSA	WLLLLSALLDGLATAYAP--AEGIVKVKYKSEDELKCHILANVAEFS--CYRKQGSFEDQAIKGGADATLGGQVITAGLTNYLQLOPIIAEDITYGS
NRL_ILFG	--GRRSVQKAVSNPEATKCFQWQRMKVRG--PPYSCIKRQSPIDCQIAIENNEADAVTLGGQVITAGLAPINLKPVAEITYGS
TRF_BLAD1	WLLQLTLISABAVLHMTPEQSPHLEIKVQXPEALES-CHNGGE--SOLHNTCYAARDRILODKIKHNEADAPYQEDMIVAAKIPQDPIIIEVIRTK
TRF_HANSE	WALLLLTILALTDAAANAKSS--YKLCYRAATNKD-CEHLEYPK--SKYALECYPARDRVDELSPYQGRADAPYQEDMIVAAKIPQDPIIIEVIRTK
TRF1_HUMAN	WHLVLLVLLGLALQLCLAGR--RRSVQKAVSQPEATKCFQWQRMKVRG--PPYSCIKRQSPIDCQIAIENNEADAVTLGGQVITAGLAPINLKPVAEITYGS
TRF1_BOVIN	WHLVYRALLSGLALQLCLAP--RINVIRICTISQPEVFKCRQWQRMKVLDA--PSITCYRAFALEDICRAIENNEADAVTLGGQVITAGLAPINLKPVAEITYGS
TRF1_HUMAN	WROPSGALWLLALRTVLDG--VEYRVKATSPQEHKCNSEATFEAD--IGPOLLCHRTSAINHVLIAADADATLGGQVITAGLAPINLKPVAEITYGS
TRF1_HOUSE	WHLIPSLIFLEALQLCLA--KATTYQKAVSNSEEDCLRWQNMKVRG--PPLSCYKSSSTROCIQAIYTNNEADATLGGQVITAGLAPINLKPVAEITYGS
SAX_RANCA	NAPTFTALFFTIILSBAAP--NAKTVRICAISLEBKXCNLYSSCNFD--ITLVCYLSSTEDCMTAKDQADHFLSGEYVYKAPINLKPVAEITYGS

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

- Storage-oriented
- Provide simple queries

Data mining

Discover knowledge from data in databases

■ Data warehouse

- Subject-oriented
- A multidimensional view of data
- Operations to access summarized data

Advanced data analysis tools

■ Statistical algorithms

- Based on many hypothesis
- Find patterns in small number of samples

Less hypothesis

Find patterns in large number of samples

Abnormal patterns

Characteristics of Data Mining

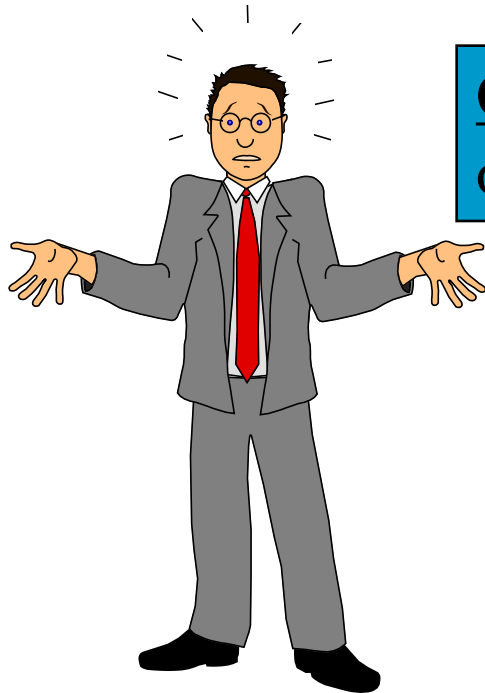
- Massive dataset
- Automatically searching for interesting patterns from historical data
- Fast
- Scalable
- Update easily
- Practical
- Decision support

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
 - 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

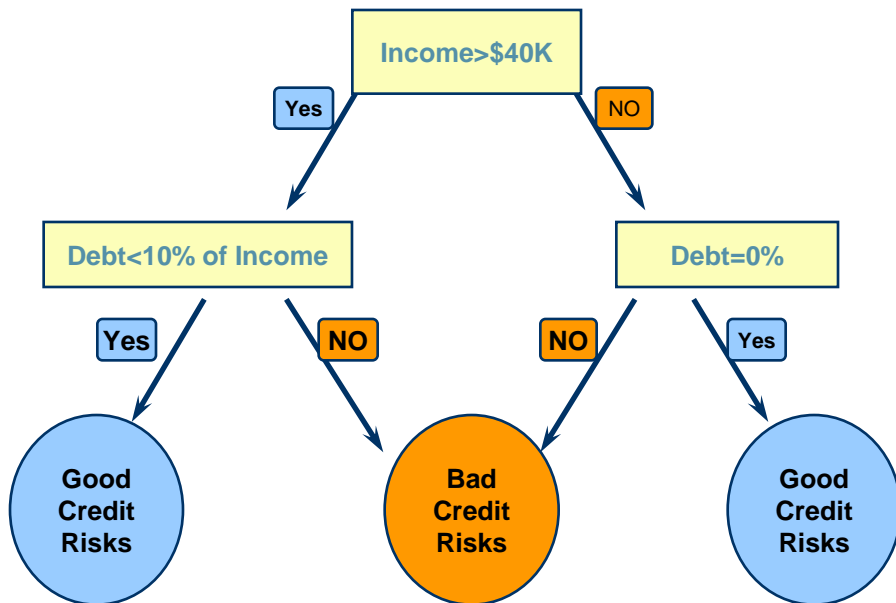
- Where does the data come from?
 - supermarket transactions, membership cards, discount coupons, customer complaint calls
- Cross-marketing analysis
 - What products were often purchased together?
Purchase recommendation, cross selling
 - What are the subsequent purchases after buying a given product?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



What Kinds of Patterns?

■ Classification

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



• Decision Tree

rule 1: if (Income ≤ \$40k) and (Debt = 0) then “good”

rule 2: if (Income > \$40K) and (Debt < 10% of Income) then “good”

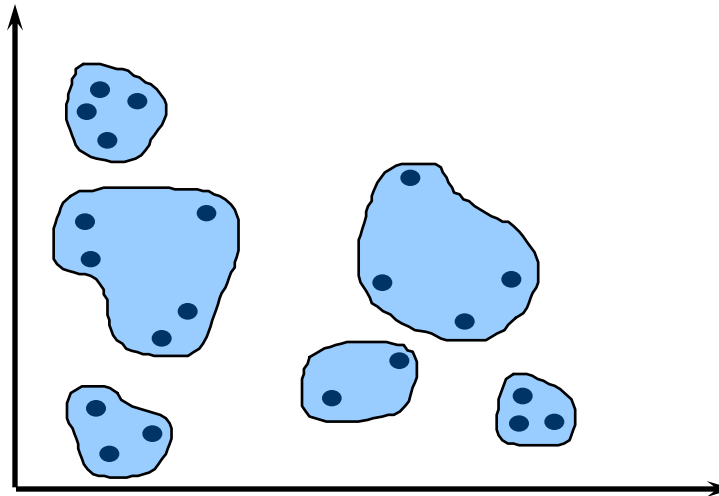
Ex.2 Credit Scoring

- Where does the data come from?
 - credit card transactions, credit card payments, loan payments
- Predict the probability to bankrupt or charge-off
- Reduce the credit risk to the banks
- Increase the profitability of the banks
- Other examples
 - Insurance fraud detection
 - Network intrusion detection

What Kinds of Patterns?

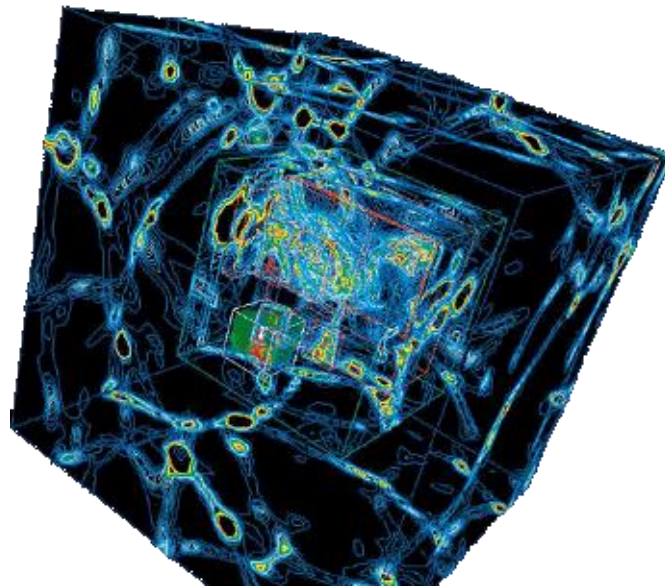
■ Clustering

- Partition the dataset into groups such that elements in a group have 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 evolution stage, beyond the capability of human being to analyze



What Kinds of Patterns?

■ Sequence mining

- Given a set of sequences, find the complete set of frequent subsequences

Buy a PC	Buy an ink printer	Buy an ink cartridges	Buy a new CPU
----------	--------------------	-----------------------	---------------

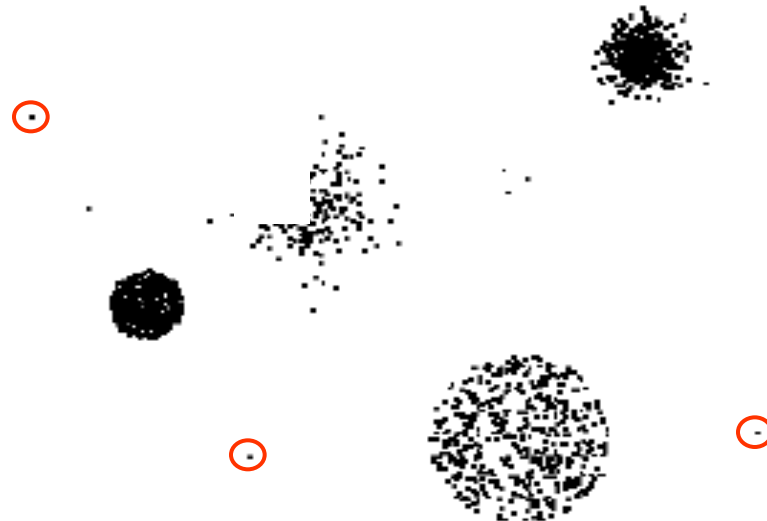


Marketing strategy: recommend a new CPU for the customer 9 months after his first purchase

What Kinds of Patterns?

■ Anomaly detection

- Given a set of n objects, and k , the number of expected anomalies, find the top k objects that are considerably dissimilar or inconsistent with the remaining data



Anomalies may be valuable!

Exercises

1. Please present an example where data mining is crucial to the success of the business. What data mining techniques are the business 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?

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

Relational Databases

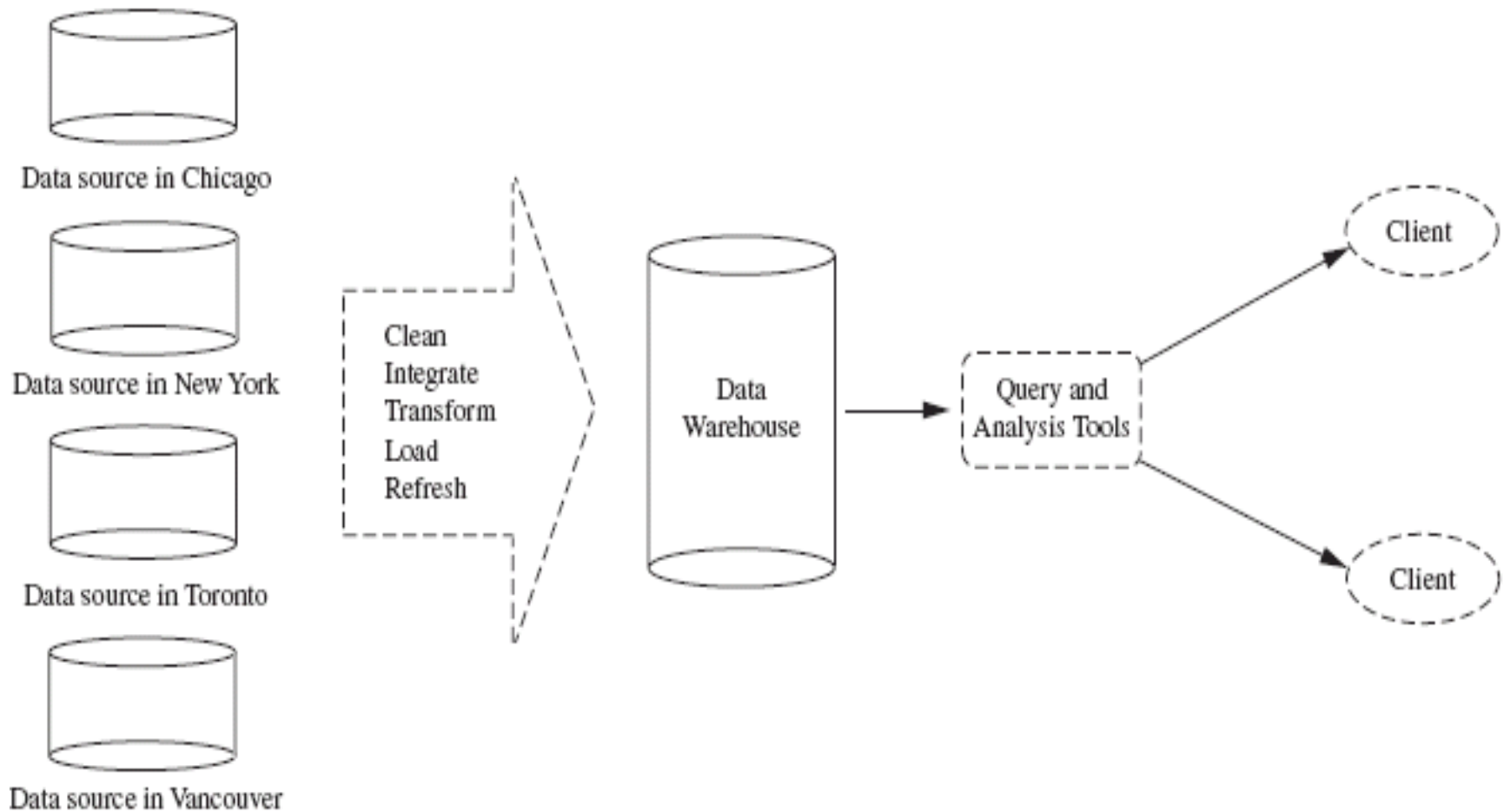
- Structured data
 - Table – records – attributes
 - Accessed by queries, SQL
- Online transactional processing (OLTP)
 - Insert a student “Ying Liu” into class “Introduction to Data Mining”, fall 2014

Name	Time	Course	score	Room
Ying Liu	Fall 2014	Introduction to Data Mining	90	002
Tom	Fall 2014	Math	85	001
Merlisa	Spring 2014	Compiler	70	001
George	Fall 2014	Graphics	92	001

Data Warehouses

- 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 “Ying Liu” in the last 3 years, grouped by semesters
- Many patterns are summarization of data
 - Roll-up, drill-down

Data Warehouses



Transactional Databases

- $I = \{x_1, \dots, x_n\}$ is the set of **items**
- An **itemset** is a subset of I
- A **transaction** is a tuple (tid, X)
 - Transaction ID tid
 - Itemset X
- 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

- Geographic databases (map)
- VLSI chip design databases
- Satellite/remote sensing image databases
- Medical image database

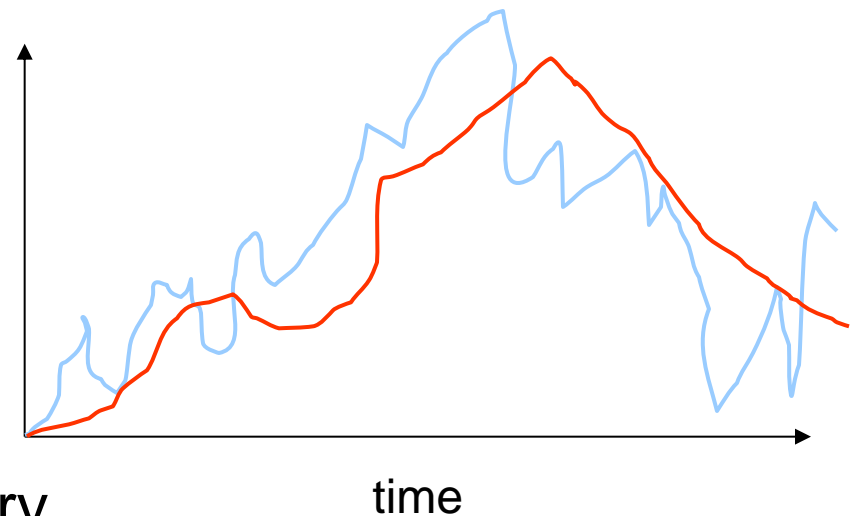
编号	中心	正右方	右上方	面积
1	居民地	绿地	水体	100
2	绿地	水体	水体	50
3	水体	居民地	居民地	600
4	水体	绿地	绿地	54

■ Spatial patterns

- Find characteristics of homes near a given location
- Change in trend of metropolitan poverty rates based on distances from major highways

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 pattern 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 matching



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, telecommunications data management, web application, sensor networks, etc.

Biomedical Data

■ Bio-sequences

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



World-Wide Web

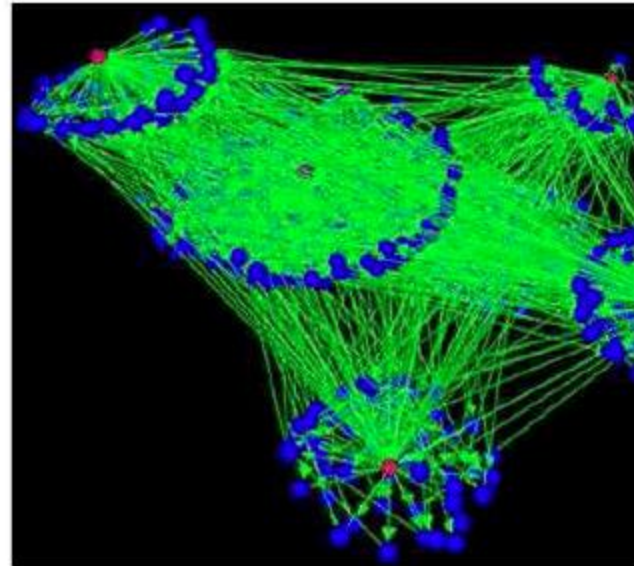
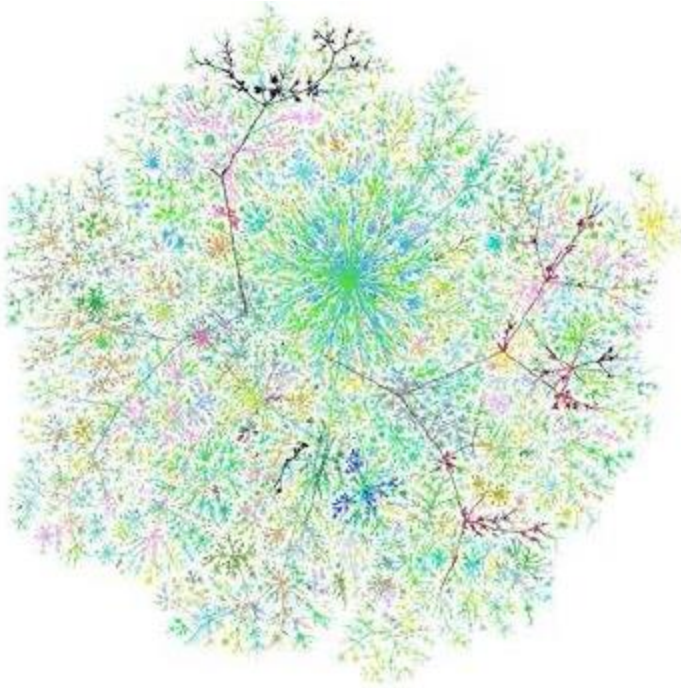
- The WWW is huge, widely distributed, global information service center for
 - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
 - Hyper-link information
 - Access and usage information
- WWW provides rich sources for data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

World-Wide Web

- Web Usage: Logs and IP package header streams
 - Mine Weblog records to discover user accessing patterns of Web pages
- Web Content
 - Extract knowledge from a Web documents, automatic categorization
- Web Structure
 - Identifying interesting graph patterns among different Web pages

Graph

■ Internet graph



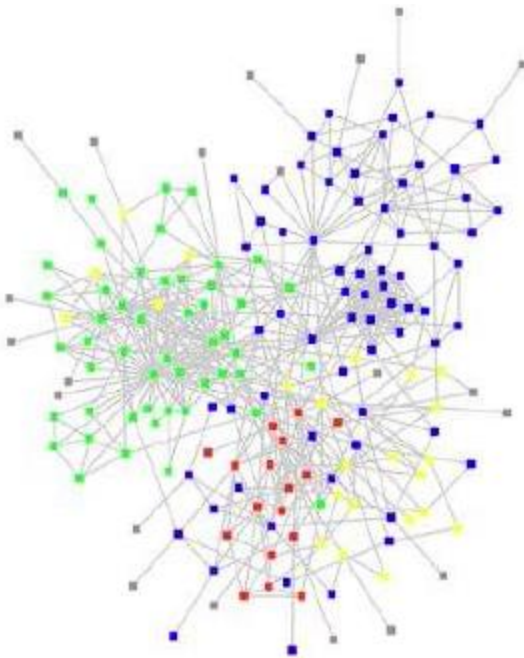
The images are downloaded from
<http://www.maths.bris.ac.uk/~maarw/graphs/graph.html>
and <http://www.netdimes.org/new/?q=node/17>

- Citation graph



Graph

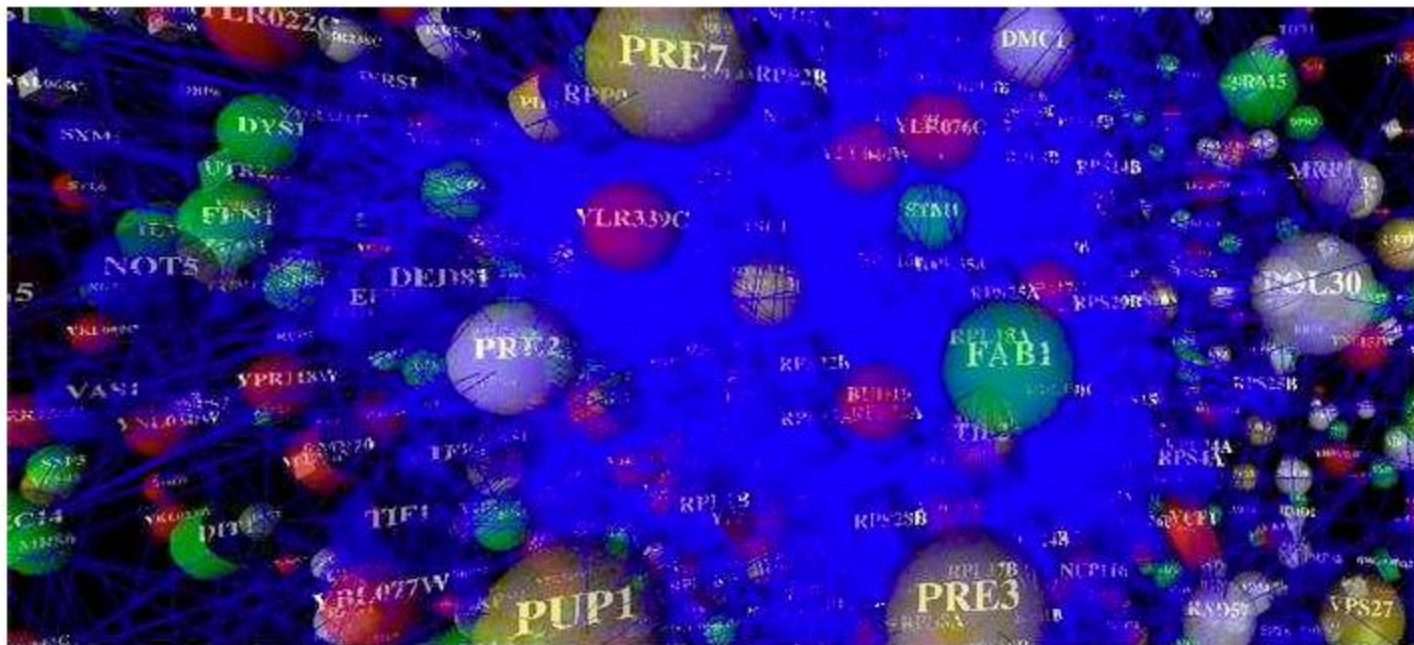
■ Friendship graph



The images are downloaded from
<http://www.thenetworkthinker.com/>
and [http://myweb20list.com/blog/2008/03/23/
new-amazing-facebook-photo-mapper/my-facebook-friend-graph/](http://myweb20list.com/blog/2008/03/23/new-amazing-facebook-photo-mapper/my-facebook-friend-graph/)

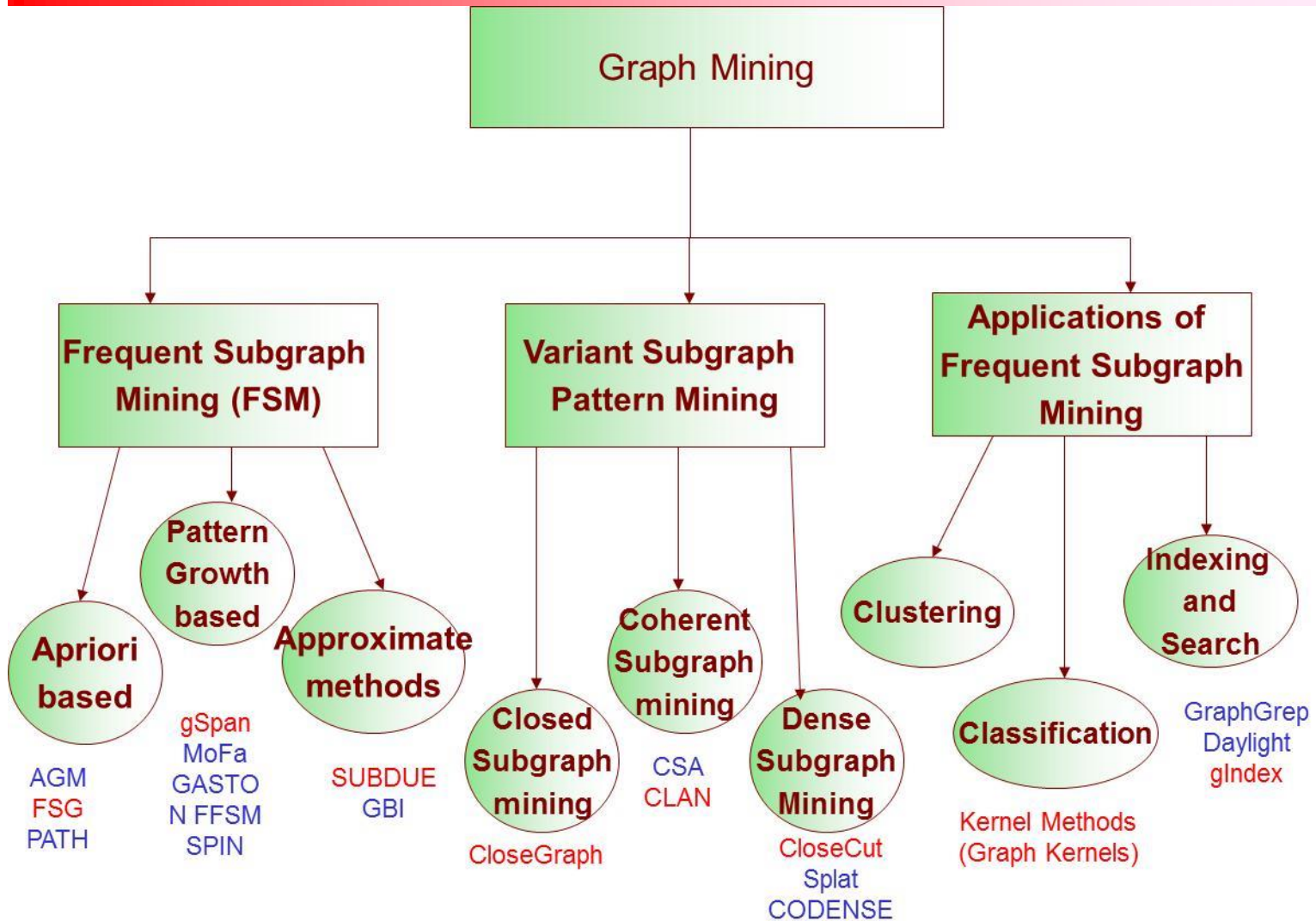
Graph

■ Protein interaction graph



The images are downloaded from
<http://bioinformatics.icmb.utexas.edu/lgl/Images/rsomZoom.jpg>

Graph



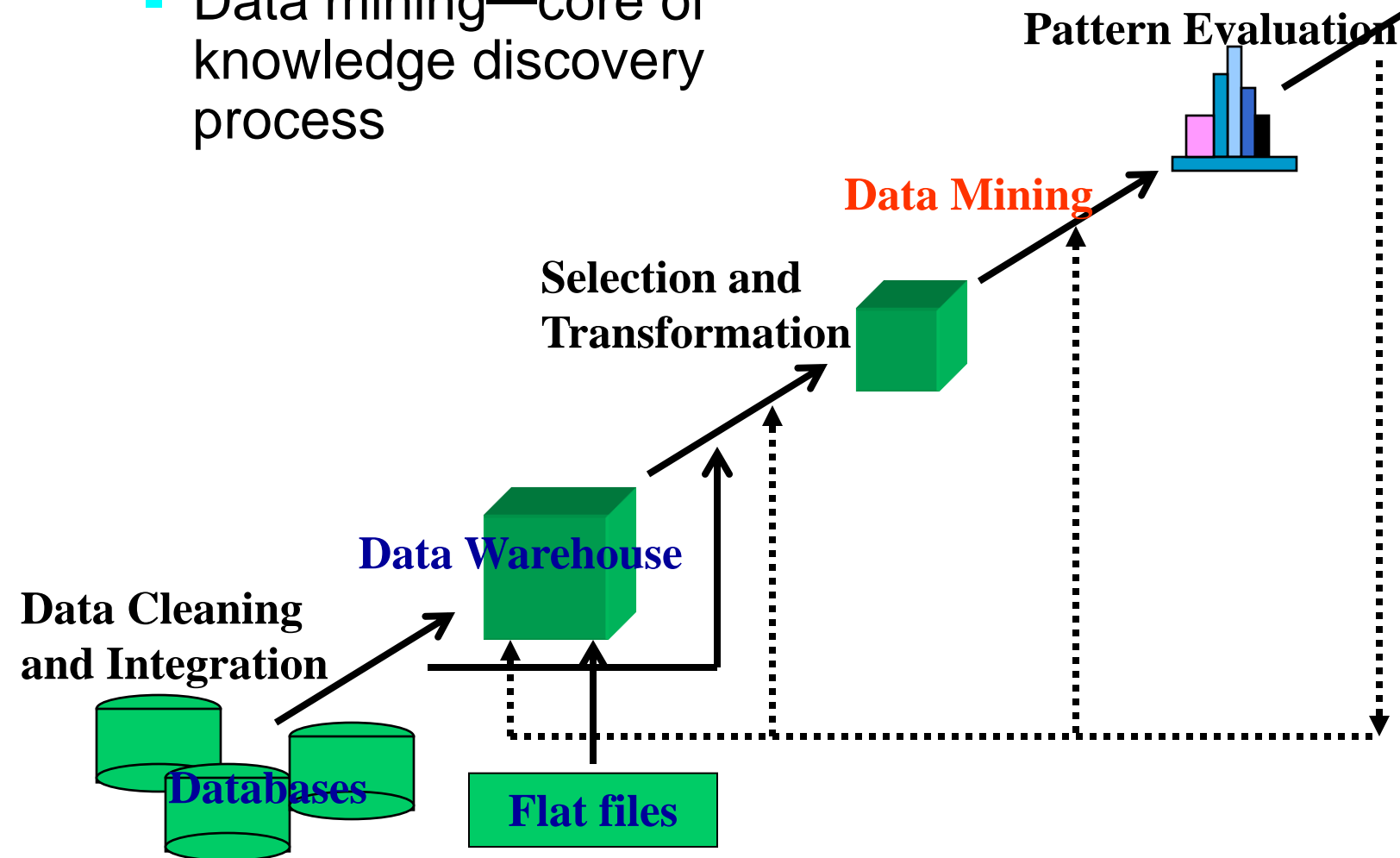
Questions

1. Assume you are a software engineer in an Internet company, e.g. Google, Baidu. What kinds of data you may process and what data mining techniques you will use to discover valuable knowledge?

Knowledge Discovery (KDD) Process

Knowledge

- 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
- Objective 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 search
- Search for only interesting patterns: An optimization problem — Challenging
 - Can a data mining system find only the interesting patterns?
 - Approaches
 - First generate all the patterns and then filter out the uninteresting 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., bio, stream, Web, graph
- Performance: efficiency, effectiveness, and scalability
- Parallel, distributed and incremental mining methods
- Pattern evaluation: the interestingness problem
- Handling noise and incomplete data
- Incorporation of background knowledge

Research Issues in Data Mining

- User interaction
 - Data mining query languages
 - Expression and visualization of data mining results
- Applications and social impacts
 - Domain-specific data mining
 - Protection of data security, integrity, and privacy

Important Resources

- Data mining conferences
 - ACM SIGKDD, IEEE ICDM, SIAM DM, PKDD, PAKDD
- Database conferences
 - ACM SIGMOD, VLDB, ACM PODS, IEEE ICDE, EDBT, ICDT
- Important journals
 - ACM Data Mining and Knowledge Discovery
 - IEEE Transactions on Knowledge and Data Engineering
 - Knowledge and Information Systems