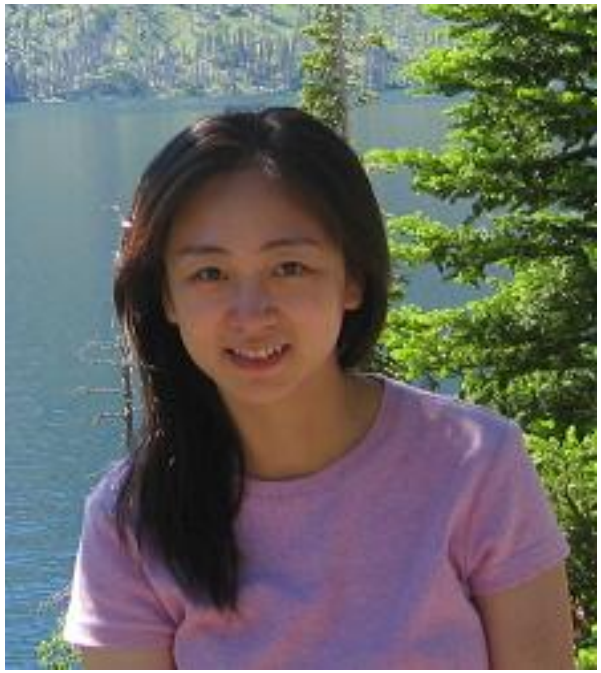
Data Mining

Ying Liu, Prof., Ph.D

School of Computer and Control
University of Chinese Academy of Sciences
The Key Lab of Big Data Mining and Knowledge Management

Welcome

- Ying Liu
  - Computer Engineering, Ph.D, Northwestern University, USA
  - Research interests
    - Data Mining
    - High Performance Computing
    - Big Data
  - Email: yingliu@ucas.ac.cn

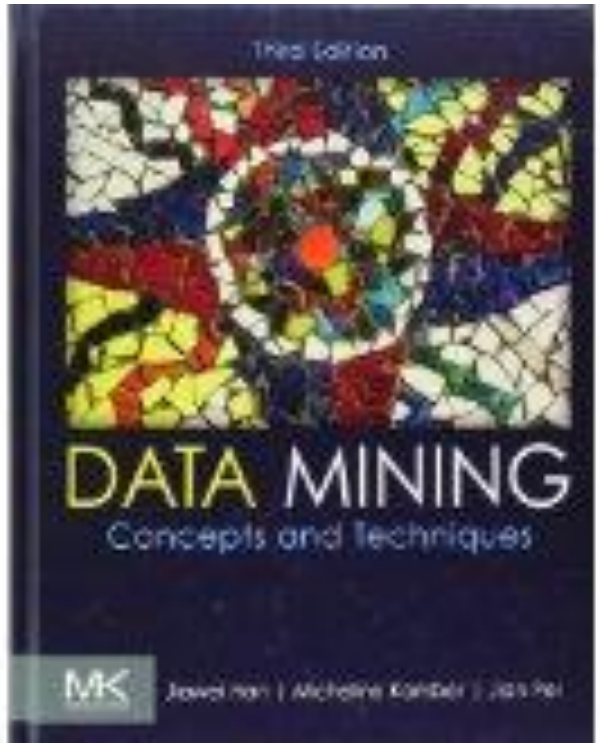


Useful Information

- Teaching Assistants
  - Leng, Jiaxu (442675812@qq.com)
  - Liu, Jinyi (807272279@qq.com)
  - Zhang, Tianlin (zhangtianlin668@163.com)
  - Quan, Pei (931995765@qq.com)
- Class: Monday & Wednesday 8:30 - 10:10, 教1-101
- Website: http://sep.ucas.ac.cn

Textbook and References

- Textbook
  - Data Mining, Concepts and Techniques. Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2011 (Third Edition)



- References
  - Research papers. To be announced in class.

Prerequisites

- Data Structure
- Algorithm
- Database
- Programming: C/C++ (preferred), Python, 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 homework assignments
- Course Project (30%)
  - Group project (4 students/group)
  - Solve a real problem: propose an algorithm/approach and implement it
- Final Exam (40%)
  - In class, closed book

About the Project

- Option 1:
  - CCF Big Data & Computational Intelligence Contest (https://www.datafountain.cn/projects/2018CCF/index.html)
  - Choose a topic from 基金间的相关性预测/面向电信行业存量用户的智能套餐个性化匹配模型/2018中国气象“神气”大数据算法与应用大赛-算法赛
  - Read through some related research papers and fully understand them
  - Develop and Implement the method
  - To be evaluated by the ranking or feedback from the contest


DF,CCF指定专业大数据
安全 | https://www.datafountain.cn
DataFountain
首页
全部赛事
专家库
企业办赛
个人主页
帮 助
登录
注册
2018 CCF BDCI 6th
大数据与计算智能大赛
Big Data & Computational Intelligence Contest
热门赛题
2018 CCF BDCI 6th
大数据与计算智能大赛
Alibaba Group
阿里巴巴集团
2018 CCF 大数据与计算智能大赛（ BDCI 2018)
https://www.datafountain.cn/competitions/314/details
22:03
2018/9/5

首页-中国科学院大学 | 基金间的相关性预测 - DF

安全 | https://www.datafountain.cn/competitions/312/details

拖拽上传

DataFountain 首页 全部赛事 专家库 企业办赛 个人主页 帮助 登录 注册

基金间的相关性预测

主办方：中国计算机学会 & 宜信大数据

机器学习 相关性预测

2018/09/06 13:23:38 littlebai参加了基金间的相关性预测

289 支队伍 / 315 参赛选手

奖金

¥1,000,000

97%

报名参赛

A 榜 B 榜 A 榜 B 榜

开赛 初赛 08.29 ~ 10.21 复赛 10.22 ~ 11.11 结束

赛制规则

数据下载与评测

相关问题

初赛排行榜

复赛排行榜

赛制规则

大赛介绍

2018 CCF大数据与计算智能大赛（BDCI 2018）由教育部高等学校计算机类专业教学指导委员会、沈阳市人民政府指导共同指导，中国计算机学会主办，以前沿技术与应用为导向，汇聚政产学研用智慧。

BDCI大赛通过连续五年举办，累计吸引来自25个国家、1000余所科高校、1200余家企事业单位、80余所科研机构的30000余人参与，已成为最具影响力的大数据赛事品牌之一。BDCI 2018立足国际化、产业化、规模化、普及化，面向金融、电信、汽车等方向发布6道赛题，面向全球开放报名。数据驱动，智见未来！

13:24 2018/9/6

面向电信行业存量用户的 ×

安全 | https://www.datafountain.cn/competitions/311/details

DataFountain 首页 全部赛事 专家库 企业办赛 个人主页 帮 助 登录 注册

面向电信行业存量用户的智能套餐个性化匹配模型

主办方：中国计算机学会 & 中国联通研究院

数据挖掘 分类预测

2018/09/05 21:50:40 ShawnyXiao参加了面向电信行业存量用户的智能套餐个性化匹配模型

502 支队伍 / **573** 参赛选手

奖金

¥1,000,000

报名参赛

A 榜 B 榜 A 榜 B 榜

开赛 初赛 08.29 ~ 10.21 复赛 10.22 ~ 11.11 结束

赛制规则

数据下载与评测

相关问题

初赛排行榜

复赛排行榜

赛制规则

大赛介绍

2018 CCF大数据与计算智能大赛（BDCI 2018）由教育部高等学校计算机类专业教学指导委员会、沈阳市人民政府指导共同指导，中国计算机学会主办，以前沿技术与应用为导向，汇聚政产学研用智慧。

BDCI大赛通过连续五年举办，累计吸引来自25个国家、1000余所科高校、1200余家企事业单位、80余所科研机构的30000余人参与，已成为最具影响力的大数据赛事品牌之一。BDCI 2018立足国际化、产业化、规模化、普及化，面向金融、电信、汽车等方向发布6道赛题，面向全球开放报名。数据驱动，智见未来！

CH 英 22:05 2018/9/5

首页-中国科学院大学 | 2018 中国气象“神气”

安全 | https://www.datafountain.cn/competitions/315/details

拖拽上传

DataFountain 首页 全部赛事 专家库 企业办赛 个人主页 帮助 登录 注册

2018 中国气象“神气”大数据算法与应用大赛-算法赛

主办方：华风集团创新研究院 & 北京天译科技有限公司 & 北京晓数科技有限公司

机器学习 图像识别

2018/09/06 12:18:03 土豆磊参加了2018 中国气象“神气”大数据算法与应用大赛-算法赛

18 支队伍 / **18** 参赛选手

奖金

¥100,000

49%

参赛

A 榜 B 榜

开赛 初赛 09.05 ~ 10.19 结束

赛制规则

数据下载与评测

相关问题

初赛排行榜

赛制规则

大赛介绍

2018中国气象“神气”大数据算法与应用大赛(2018 Meteorology Big Data Algorithm & Application Contest，简称“MBDAA”)是气象大数据的算法、应用大型挑战赛事。旨在激发社会各行业、高校、科研院所等对气象数据应用的想象力和创新力；提高社会各行业对气象数据应用的能力空间。

本届赛事分为算法题、应用题、创意题三大类，本赛题为算法题。通过行业数据与气象数据结合计算，进行创新探索，利用新一代大数据技术促进产业发展，帮助传统行业发展找到更多竞争制高点。

CH 13:33 2018/9/6

About the Project

- Option 2:
  - Contest in class
  - Assigned a topic (to be announced)
  - Read through some related research papers and fully understand them
  - Develop and Implement the method
  - To be evaluated by the ranking in class

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
- Think creatively

Why Take This Course ?

- Data mining is 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

➢*Data Mining is very useful*!

Syllabus (Tentative)

- Introduction
- Data warehouse
- Data pre-processing
- Association rules
- Classification
- Clustering
- Applications
- Advanced topics

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 (individual assignment) or of your approved team (group assignment)

No Plagiarism!

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

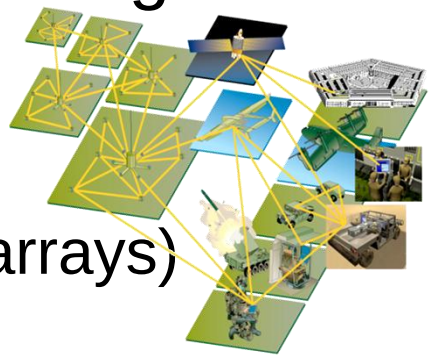
What Motivated Data Mining – Business World

- Tremendous of data being collected and stored
  - E-commerce
  - Transactions
  - Stocks
  - Credit card transactions

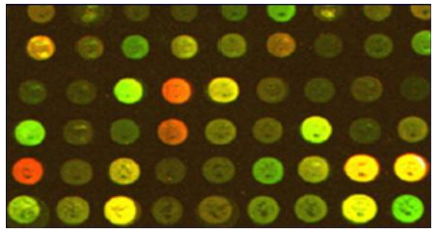
- Strong competitive pressure to extract and use the knowledge hidden in the data to provide customized CRM

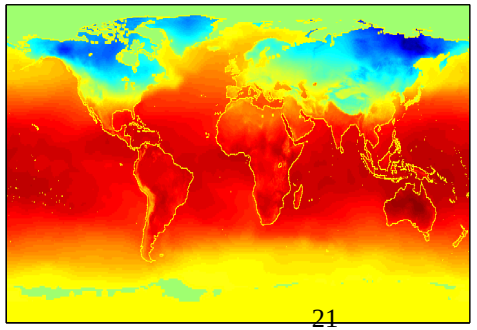
What Motivated Data Mining – Scientific World

- 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?

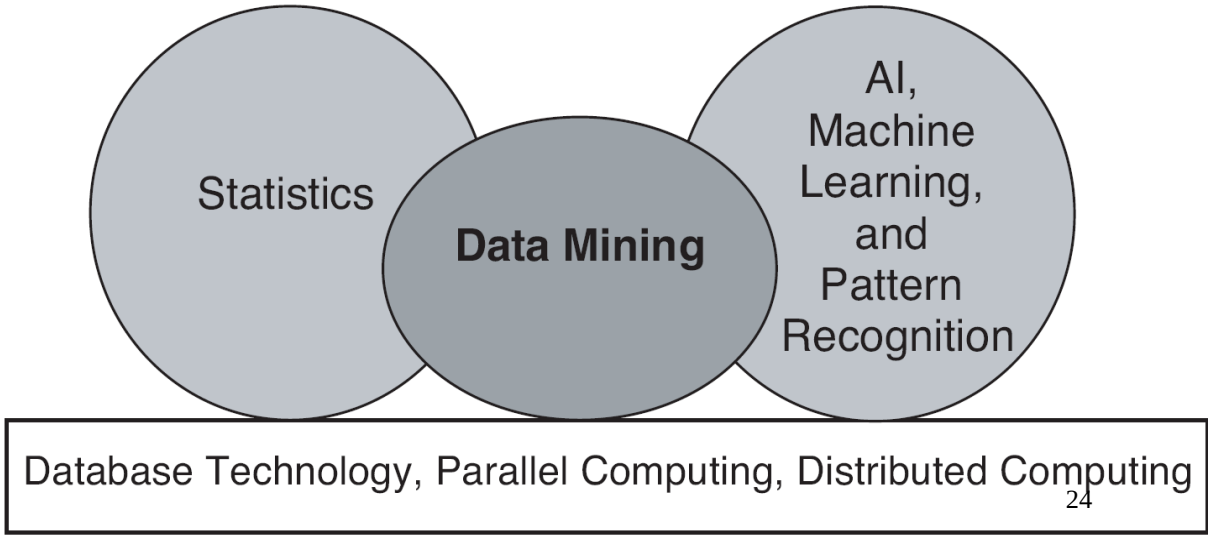
- We are drowning in data, but starving for knowledge!
  - Data rich, knowledge poor
  - Decision makers, domain experts have biases or errors
- Automated analysis of massive data sets

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
  - Parallel/Distributed computing



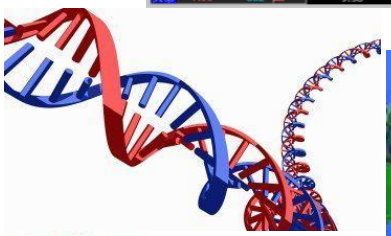
Why Not Traditional Data Analysis?

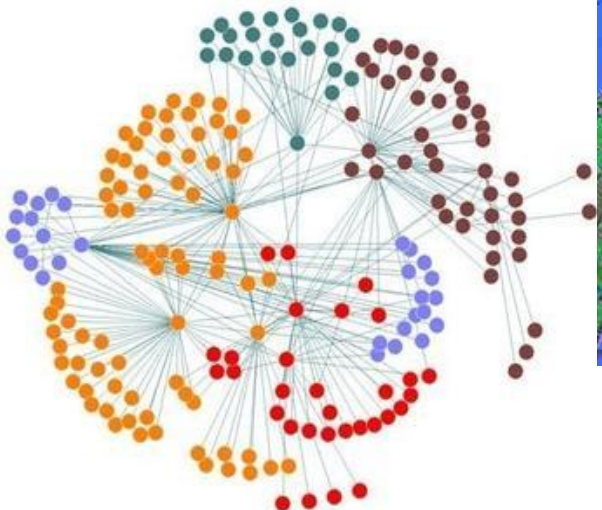
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data

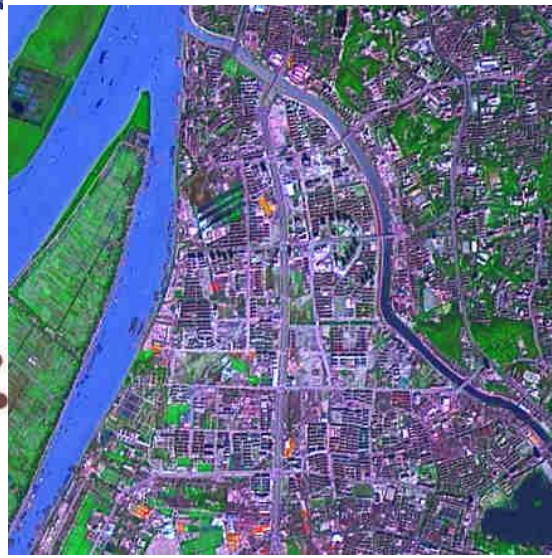
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions

Why Not Traditional Data Analysis?

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







Why Not Traditional Data Analysis?

- Database
  - Storage-oriented
  - Provide simple queries
- Data warehouse
  - Subject-oriented
  - A multidimensional view of data
  - Operations to access summarized data
- Statistical algorithms
  - Based on many hypothesis
  - Find patterns in small number of samples

Data mining

Discover knowledge from data in databases

Advanced data analysis tools

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

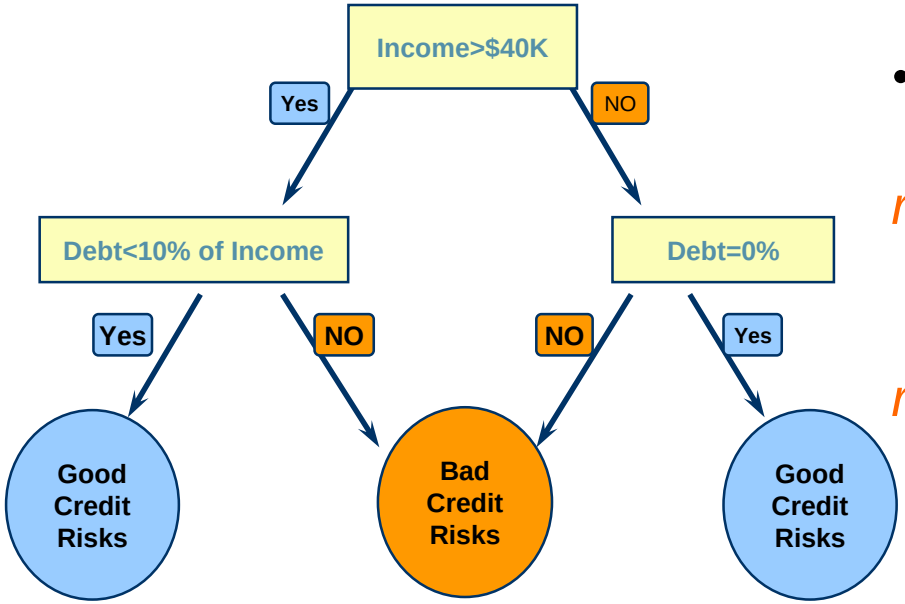
On Thursdays, during 4-11pm customers often purchase diapers and beers together!

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"

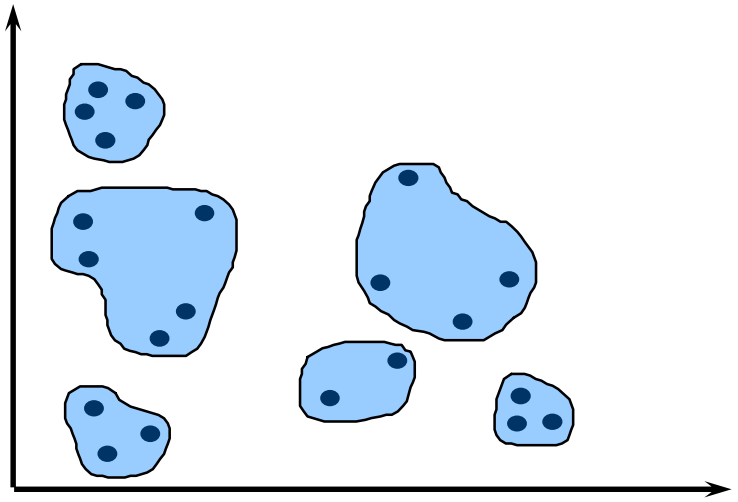
rule 2: if (Income>$40K) and (Debt<10% of Income) then "good"

Ex.2 Credit Scoring

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

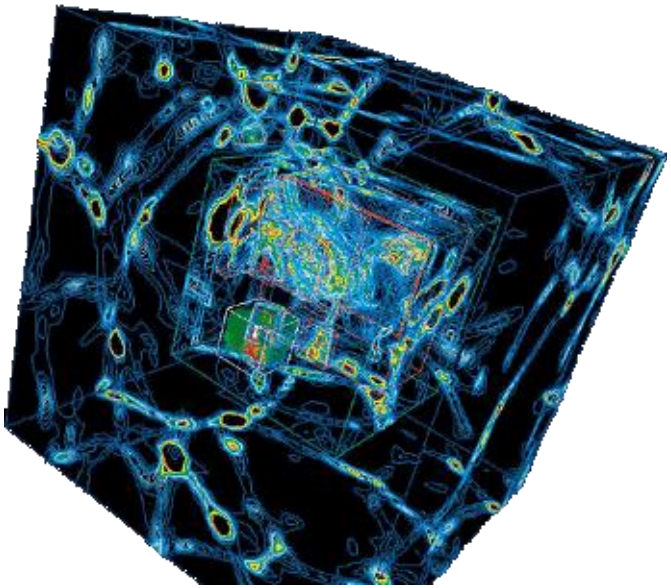
What Kinds of Patterns?

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



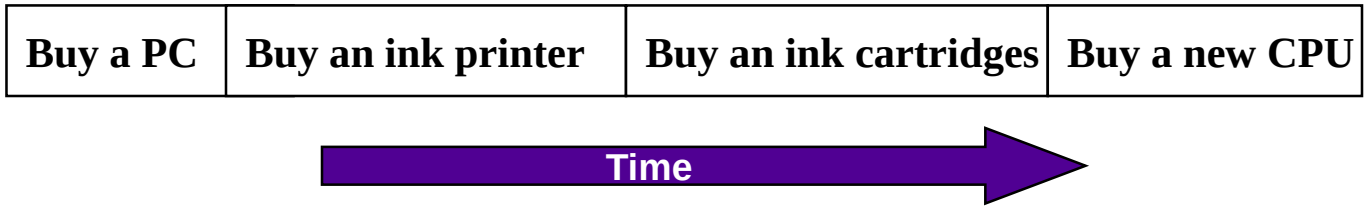
Ex.3 Scientific Simulation

- 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?

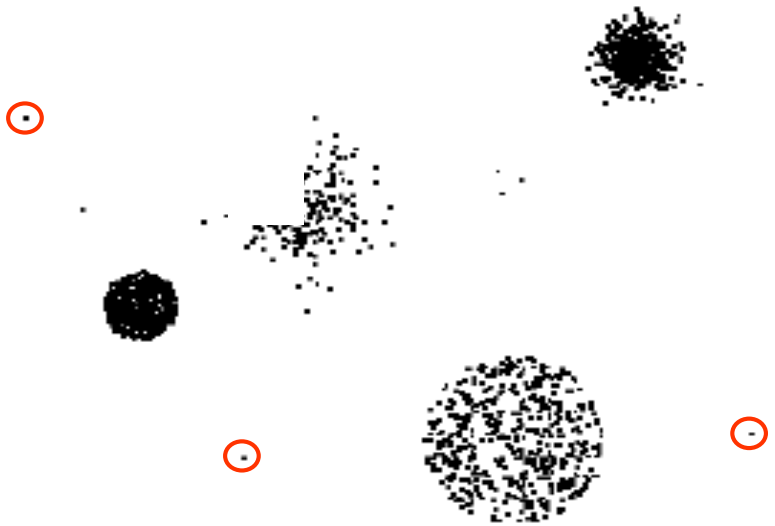
- Sequence mining
  - Given a set of sequences, find the complete set of frequent subsequences



Marketing stragegy: 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



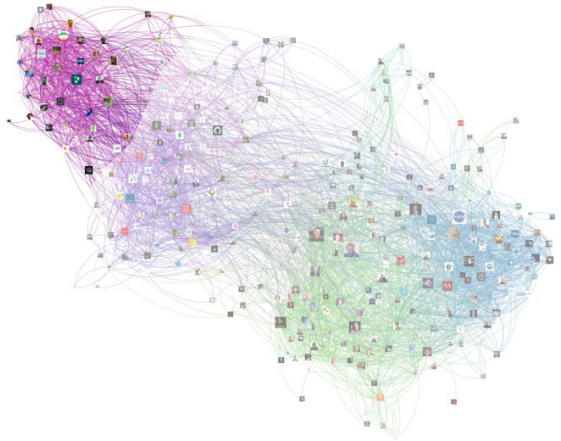
Anomalies may be valuable!

What Kinds of Patterns?

- 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 detected communities?



What Kinds of Patterns?

- 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

Product

amazon

JD.COM 京东

天猫 TMALL.COM

Movie

爱奇艺 iQIYI

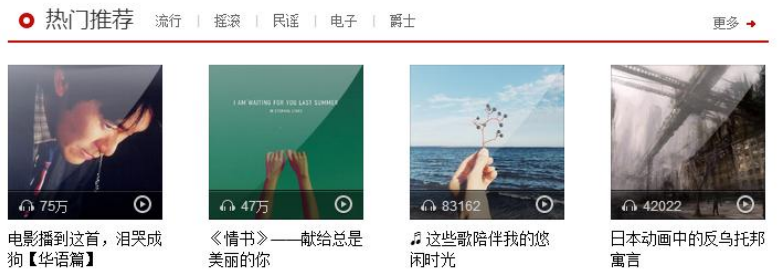
优酷 YOUKU

腾讯视频 V.QQ.COM

Music

QQ音乐

网易云音乐



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

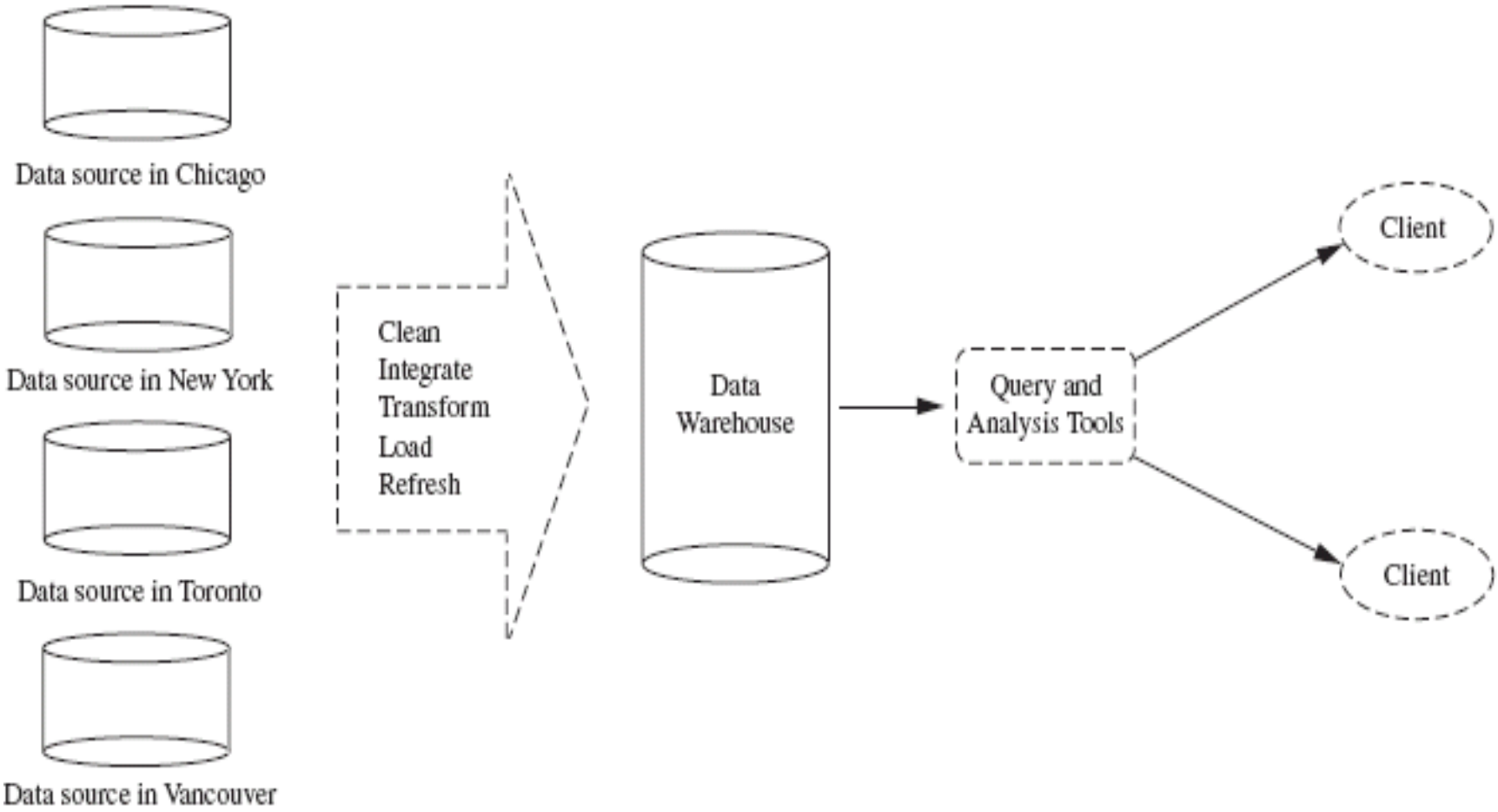
- Structured data
  - Table – records – attributes
  - Accessed by queries, SQL
- Online transactional processing (OLTP)
  - Insert a student “Ying Liu” into class “Introduction to Data Mining”, fall 2014

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

- $I=\{x_1, \ldots, x_n\}$ 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

与关系数据库不同的是，空间数据库中的数据很难计算

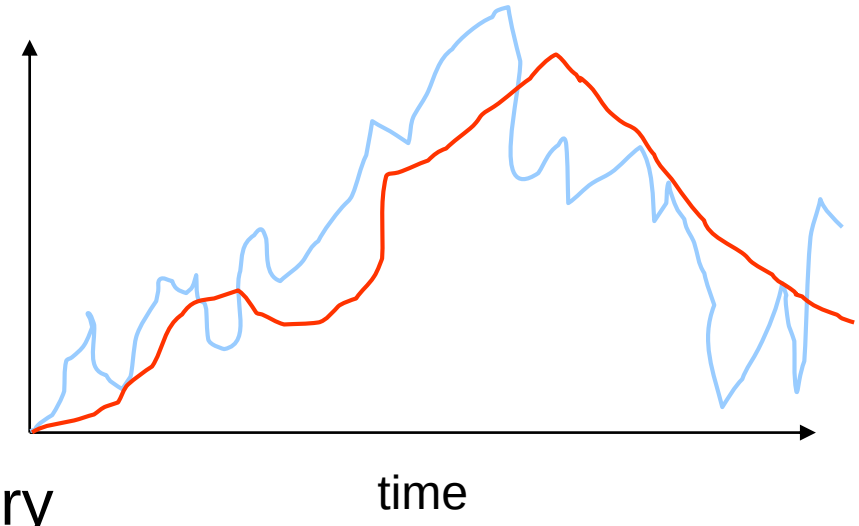
- Spatial information
  - Geographic databases (map)
  - VLSI chip design databases
  - Satellite/remote sensing image databases
  - Medical image database
- Spatial patterns
  - Find characteristics of homes near a given location
  - Change in trend of metropolitan poverty rates based on distances from major highways

编号	中心	正右方	右上方	面积
1	居民地	绿地	水体	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
…	…	…	…	…

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

分割方法



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
    - Deep learning

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

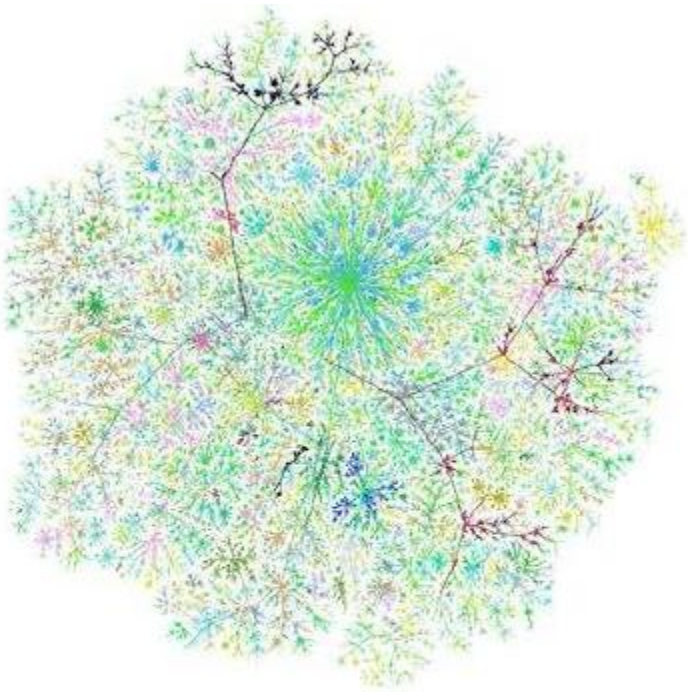
- The WWW is huge, widely distributed, global information service center for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for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

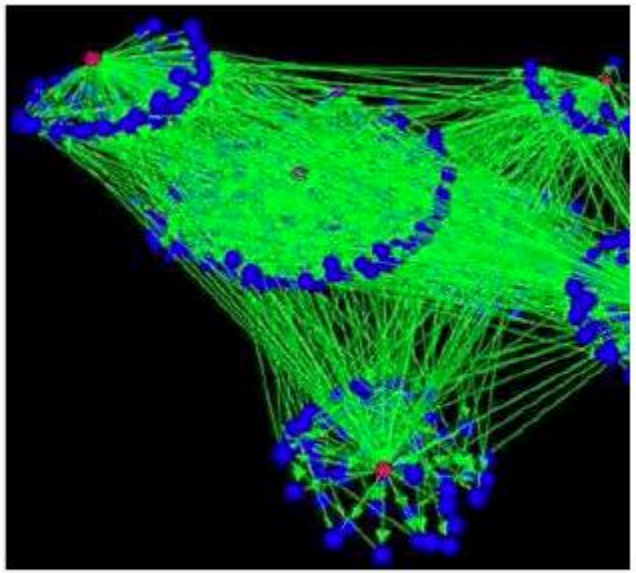
World-Wide Web

- 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

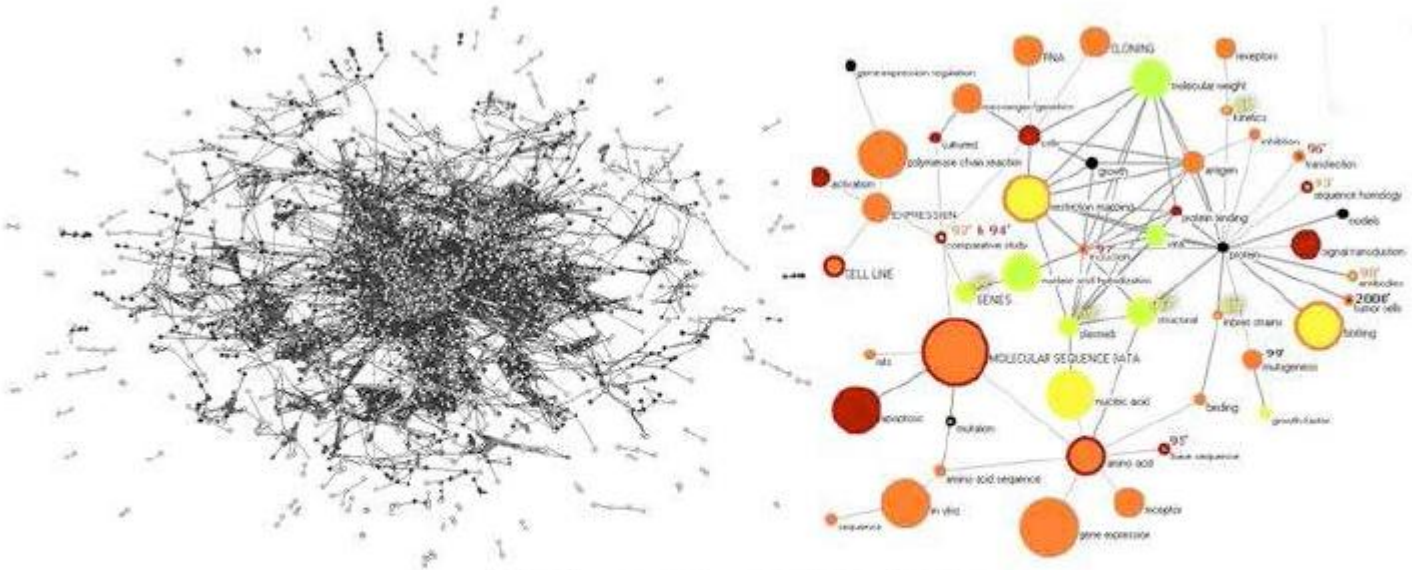




The images are downloaded from
http://www.maths.bris.ac.uk/~maarw/graphs/graph.html
and http://www.netdimes.org/new/?q=node/17

Graph

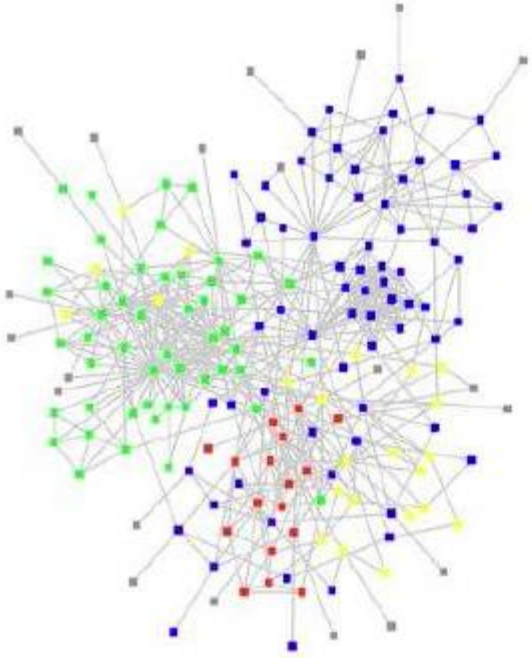
- Citation graph



The images are downloaded from
http://www.emeraldinsight.com/fig/2780600403005.png
and www.bordalierinstitute.com/target1.html

Graph

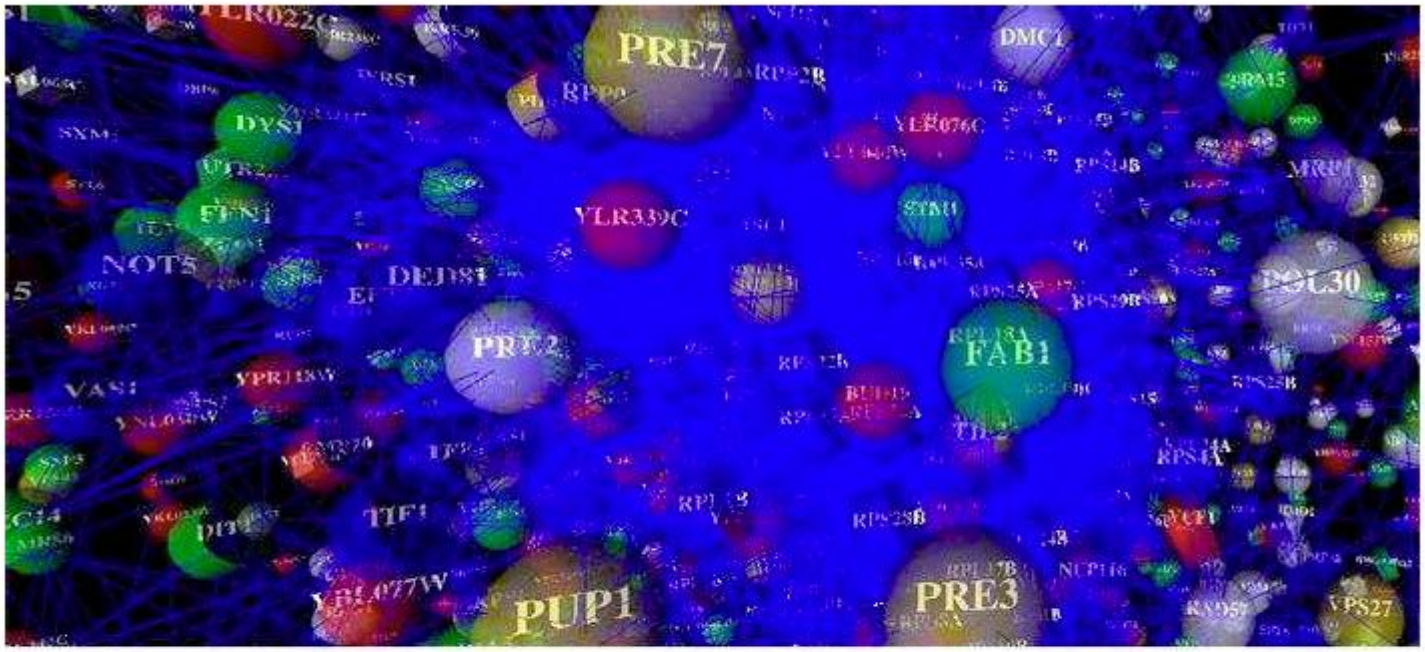
- Friendship graph



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/

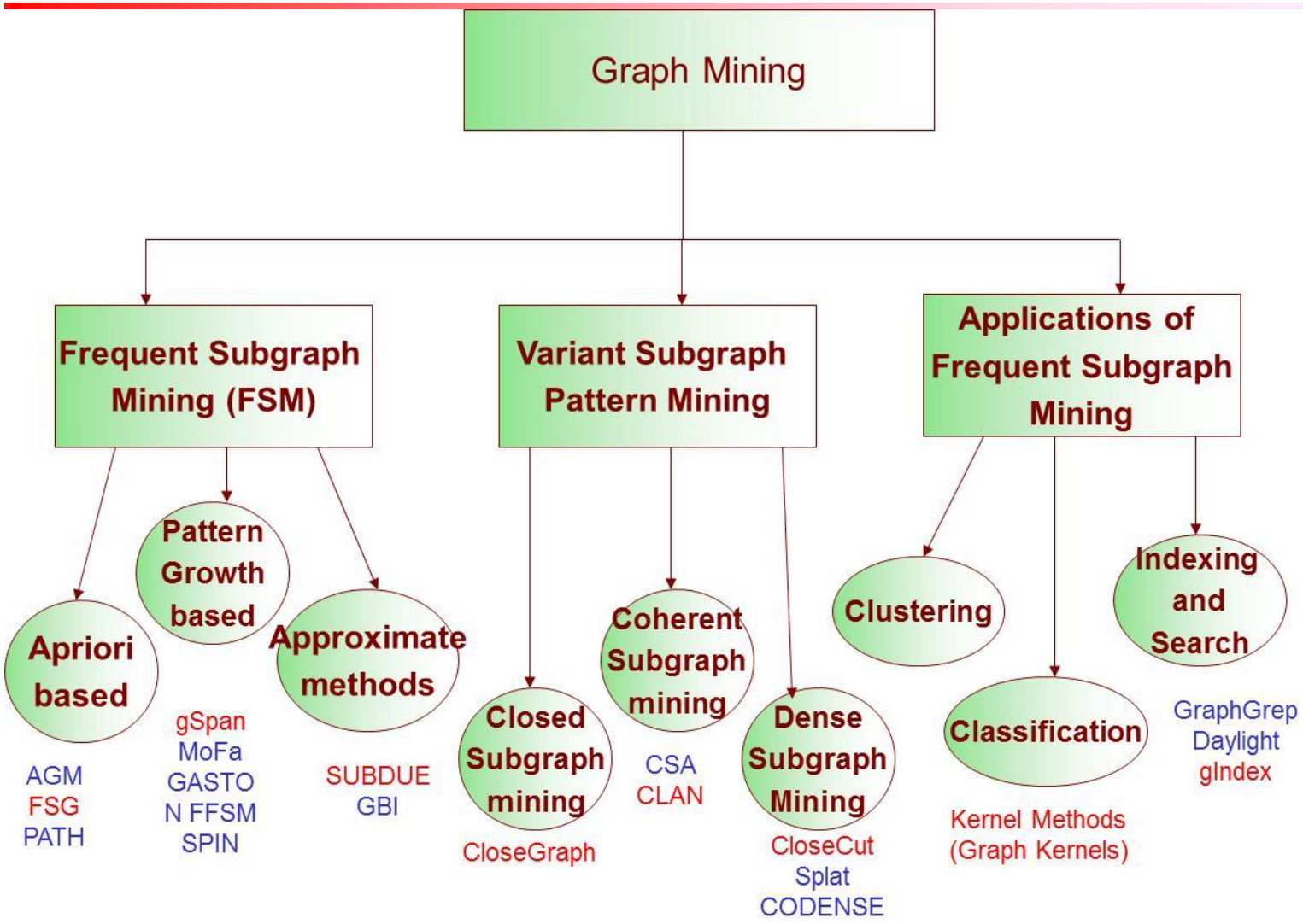
Graph

- Protein interaction graph



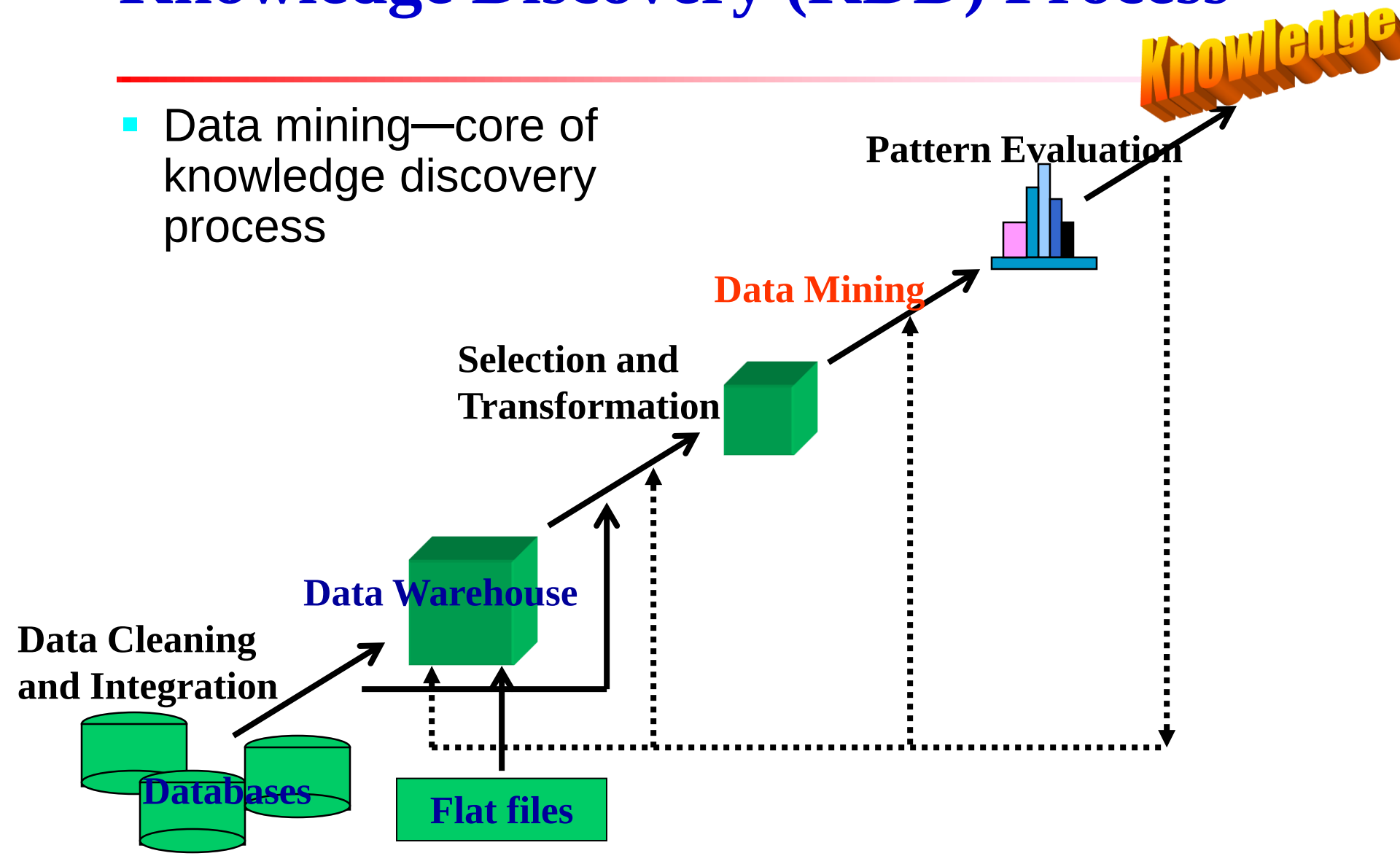
The images are downloaded from
http://bioinformatics.icmb.utexas.edu/lgl/Images/rsomZoom.jpg

Graph


Graph Mining
Frequent Subgraph Mining (FSM)
Variant Subgraph Pattern Mining
Applications of Frequent Subgraph Mining
Apriori based
Pattern Growth based
Approximate methods
Closed Subgraph mining
Coherent Subgraph mining
Dense Subgraph Mining
Clustering
Classification
Indexing and Search
AGM
FSG
PATH
gSpan
MoFa
GASTO
N FFSM
SPIN
SUBDUE
GBI
CloseGraph
CSA
CLAN
CloseCut
Splat
CODENSE
Kernel Methods
(Graph Kernels)
GraphGrep
Daylight
gIndex

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
- 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 遍历 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., Web, graph, bio, stream, image, audio
  - Performance: efficiency, effectiveness, and scalability
  - Parallel, distributed and incremental mining methods
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
  - 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