数据挖掘

刘莹，博士，教授

中国科学院大学计算机科学与技术学院
中国科学院大学数据挖掘与高性能计算实验室

Welcome

- 首席/主讲
  - 刘莹，教授
  - Computer Engineering, Ph.D, Northwestern University, USA, 2005
  - Research interests
    - Data Mining, Artificial Intelligence, High Performance Computing, etc.
  - Email: yingliu@ucas.ac.cn

Welcome

- 助教
  - 姜小平
  - Email: jiangxiaoping17@mails.ucas.ac.cn

Useful Information

- Class: Monday 3:20 - 5:00, 教404
- 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) (英文版)

  - 韩家炜等，数据挖掘：概念与技术（第3版） (中文版)，2012，机械工业出版社

Prerequisites

- 数据结构
- 算法
- 概率统计
- 编程语言：C/C++ (preferred), Python, Java, etc.

Grading Scheme

- Assignments (30%)
  - 4 homework assignments
- Course Project (30%)
  - Group project (2-3 students/group)
  - Solve a real-world problem
    - Develop an algorithm
    - Hand in a project report
  - In class presentation
  - To be evaluated in performance, technical innovation, thoroughness of the work, clarity of presentation
- Final Exam (40%)

About the Assignments

- 4 written assignments
  - written parts
  - in-class lab assignments
    - Implement some algorithms in class
    - Under the supervision of the instructors

About the Project

- Solve a real-world problem

It is going to be a competition !

- Choose a topic from the provided topics
- Read through some related research papers and fully understand them
- Develop and implement the method
- Write a technical report
- To be evaluated in performance, technical innovation, clarity of presentation

About the Project

- Option 1: 垃圾邮件分类
  - https://challenge.datacastle.cn/v3/cmptDetail.html?id=352
  - 赛题任务：给定邮件文本信息，建立分类模型，判断哪些邮件属于垃圾邮件。

About the Project

About the Project

- Option 2:内存故障预测
  - https://tianchi.aliyun.com/competition/entrance/532055
  - 赛题任务：根据一段时间内的内存系统日志、内存故障数据，通过科学的方式来预测某块内存在未来一段时间是否会出现故障，输出预测未来7天会发生内存故障的机器集合，且附带预测时间间隔。

About the Project

https://tianchi.aliyun.com/competition/entrance/532055/information

数据描述

数据中提供的24个日志模版可以理解为对系统日志长文本的进行了关键字提取，模版为这些关键字的组合，其中hwerr表示hardware error模块，sel表示sel模块，但我们对模版其他的关键字进行了脱敏。

列名	实例	说明
collect_time	2019/1/14 18:19	日志发生时间
hwerr	1	脱敏后的hardware error模块
sel	1	脱敏后的sel模块
serial_number	server_31576	电脑编号

提交说明

选手需要用自己训练好的模型在测试集上预测结果（未来7天是否出现故障）并将预测为会出现故障的机器和预测时间间隔(pti：时间间隔，以分钟为单位)保存为csv格式提交。

形式如下：

```
server_1,2019-08-15 00:00:00,14

server_123,2019-08-16 02:12:00,1200
```

How to Do a Good Project?

- Start early
  - It takes time to understand, learn and think
- Discuss with me
  - Maybe I can give some suggestions or ideas
- Implement concretely
  - Understand the pros and cons
- 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
- Classification/prediction
- Association rules
- Clustering
- Applications
  - credit scoring, target marketing, oil exploration, radar target detection & recognition
- Advanced topics

Objectives of This Course

- Introduce the motivation of data mining
- Outline principles, major algorithms
- Introduce applications
- Introduce advanced topics

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

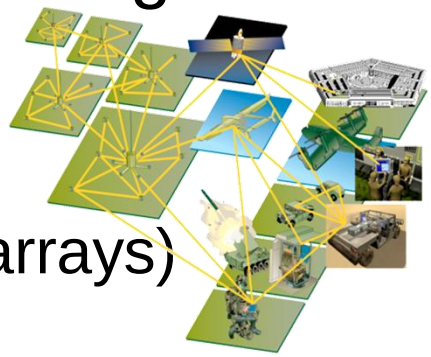
What Motivated Data Mining – Business World

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

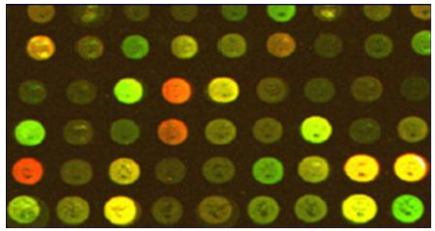
- Strong competitive pressure to extract and use the knowledge hidden in the data to provide customized CRM

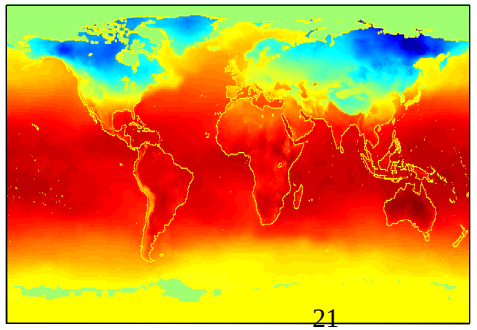
What Motivated Data Mining – Scientific World

- Tremendous of data being collected and stored
  - Remote sensing
  - Bioinformatics (Microarrays)
  - Scientific simulation



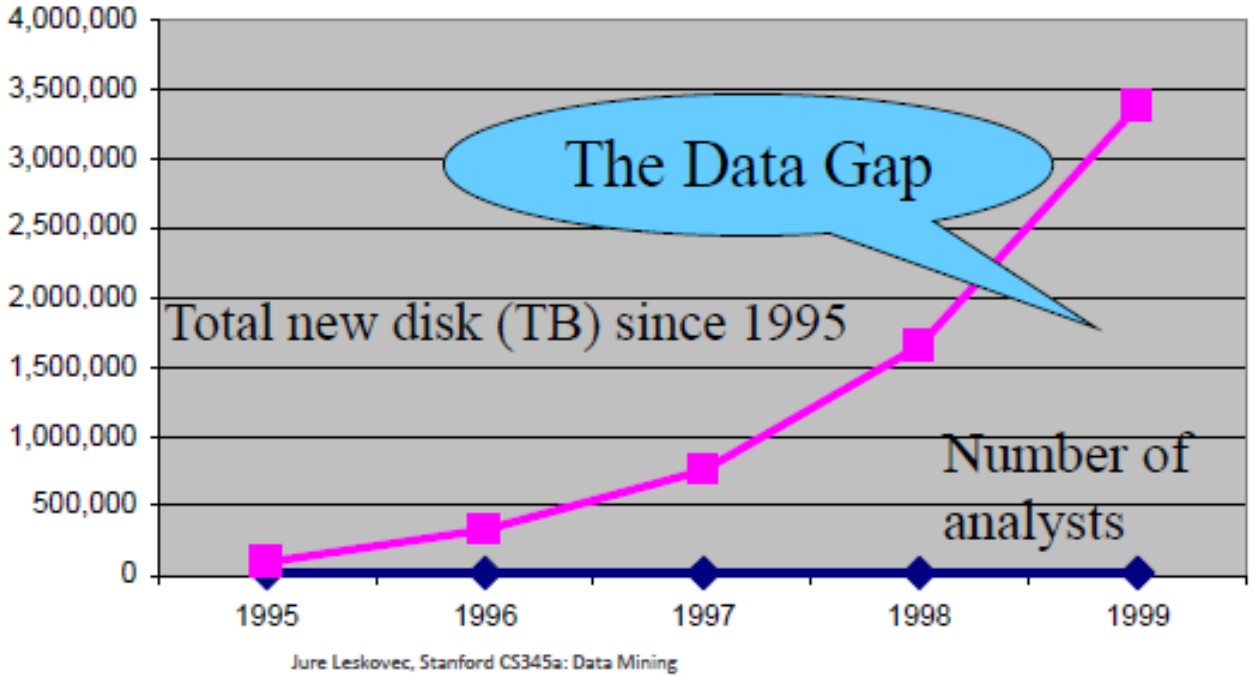
- Scientists need strong data analysis to assist research, such as classification, segmentation, etc.





What Motivated Data Mining?

- There is often information “hidden” in the data that is not readily evident
- Human analysts take weeks to discover useful information
- Much of the data is never analyzed at all



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?

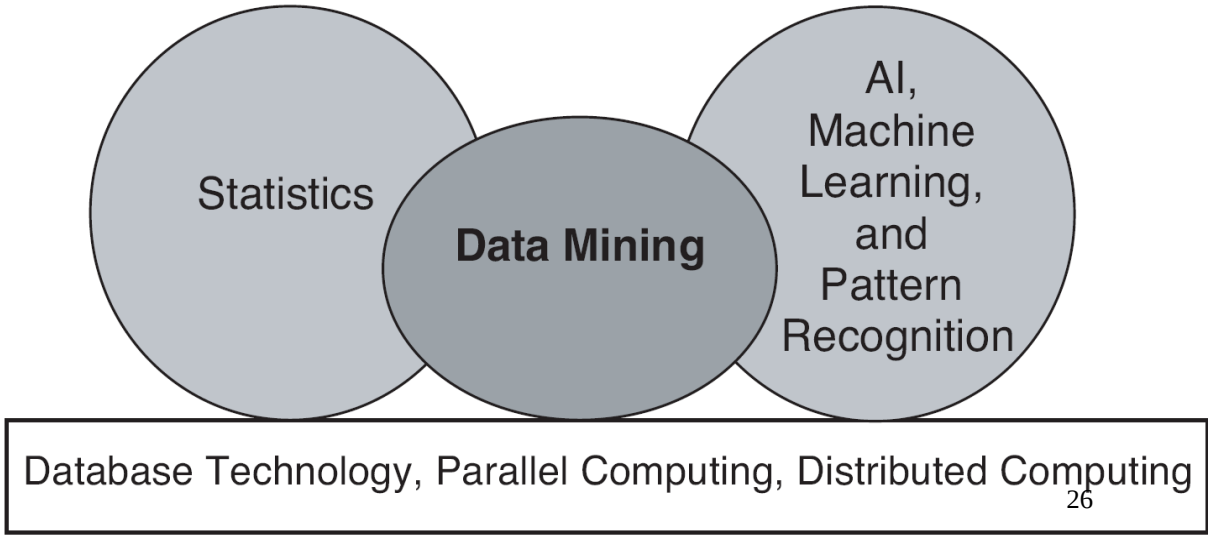
- Data mining — Discover valid, novel, useful, and understandable patterns in massive datasets

What is Data Mining?

- Automatically analyze large databases to find patterns that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern

What is Data Mining?

- Cross Disciplines
  - Databases
  - Machine learning: decision tree, Bayesian classifier, etc.
  - Statistics: regression, etc.
  - Neural networks
  - Parallel/Distributed computing



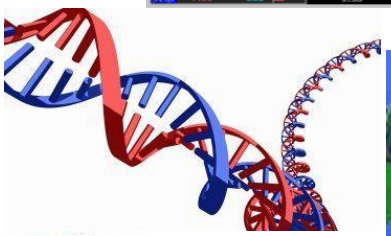
Why Not Traditional Data Analysis?

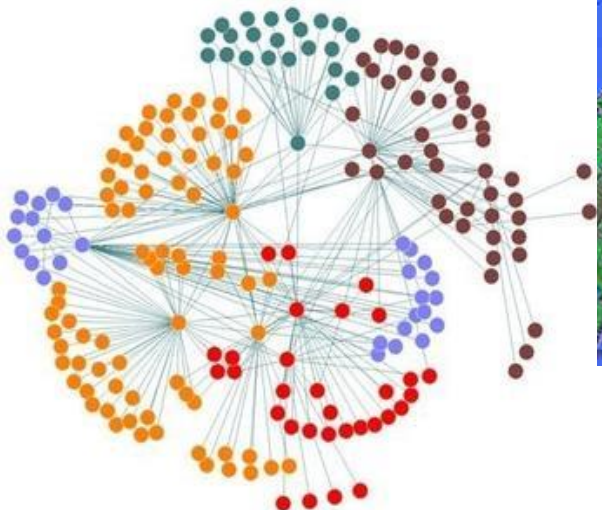
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data

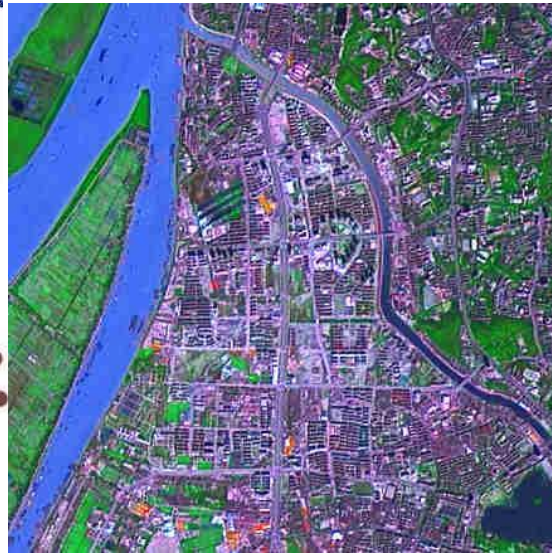
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions

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

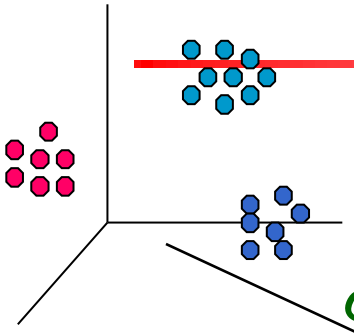
Find patterns in large number of samples

Abnormal patterns

Characteristics of Data Mining

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

Data Mining Tasks

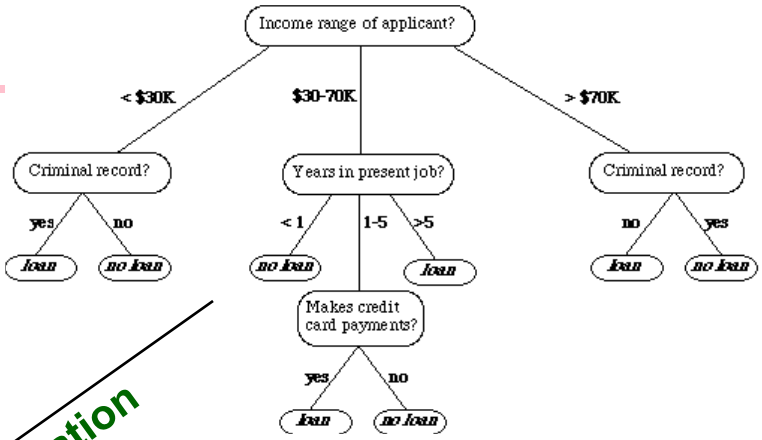


Clustering

Data

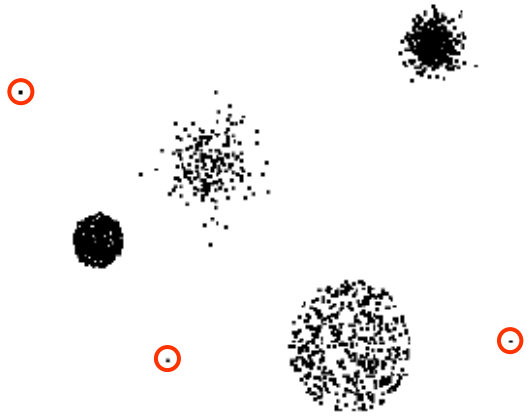
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Classification



Association Rules

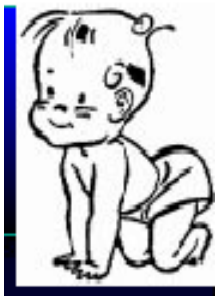
Anomaly Detection



Association Rules Mining

- 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!



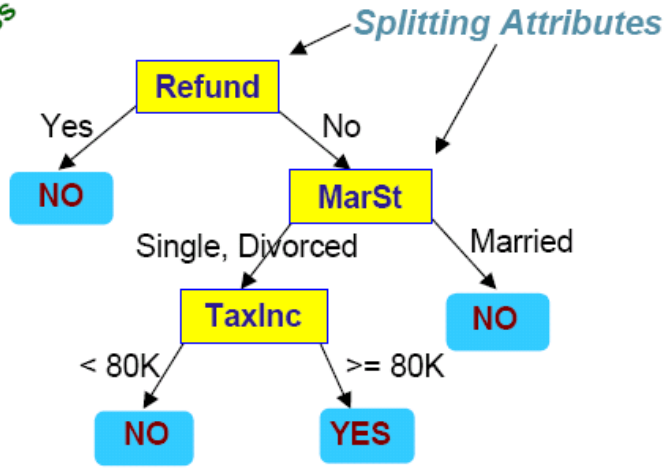
Ex. 1: Production Recommendation

- Where does the data come from?
  - supermarket transactions, membership cards, shopping carts, discount coupons
- Discover individual products, or groups of products that tend to occur together in transactions
- Determine recommendations and cross-sell and up-sell opportunities
- Improve the efficiency of a promotional campaign

Classification

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

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



The splitting attribute at a node is determined based on the Gini index.

- Decision Tree

rule 1： if (Refund='no') and (MarSt = 'Single, Divorced') and (TaxInc >= 80K) then "Cheat"

Ex.2 Credit Scoring

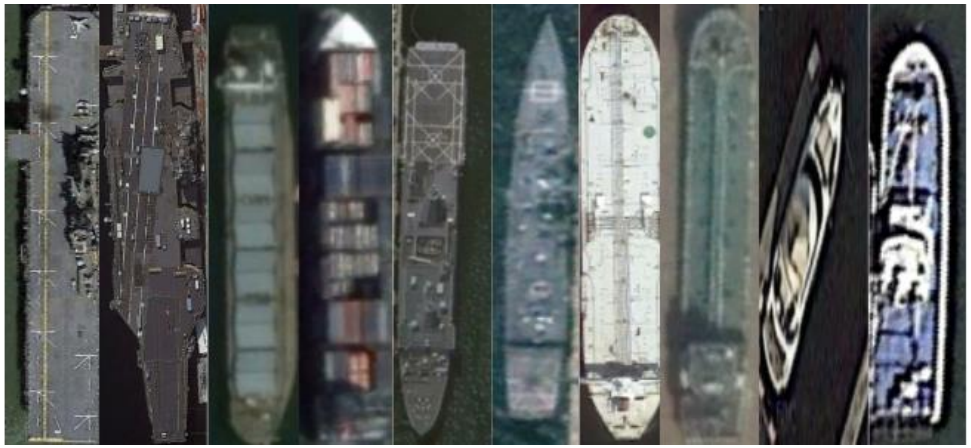
- Where does the data come from?
  - Credit card transactions, credit card payments, loan payments, demographic data
- Predict the probability to bankrupt or charge-off
- Reduce the credit risk to the banks
- Increase the profitability of the banks

Ex.2 Credit Scoring

- Decision Tree

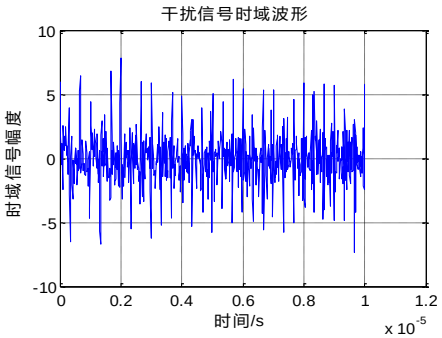
rule 1： if (Income<=$40k) and (Debt=0) then "good"
rule 2: if (Income>$40K) and (Debt<10% of Income) then "good"

Ex.3 目标识别

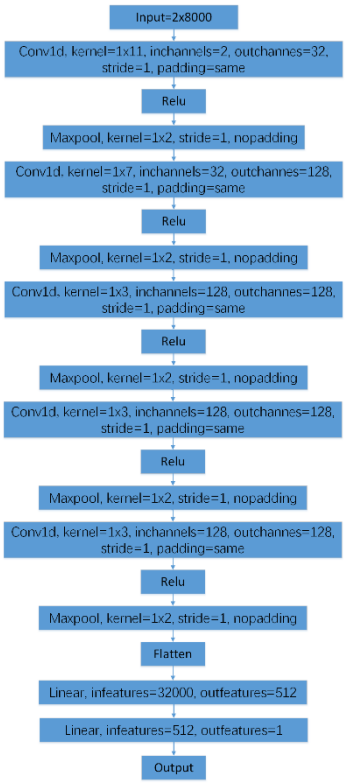


	登陆舰	航母	货船	集装箱	军舰_1	军舰_2	大型油轮	小型油轮	游艇	渔船
误分率	13%	6.5%	3.3%	16%	10%	6.5%	3.3%	0%	3.3%	3.3%

Ex.4 雷达信号干扰识别



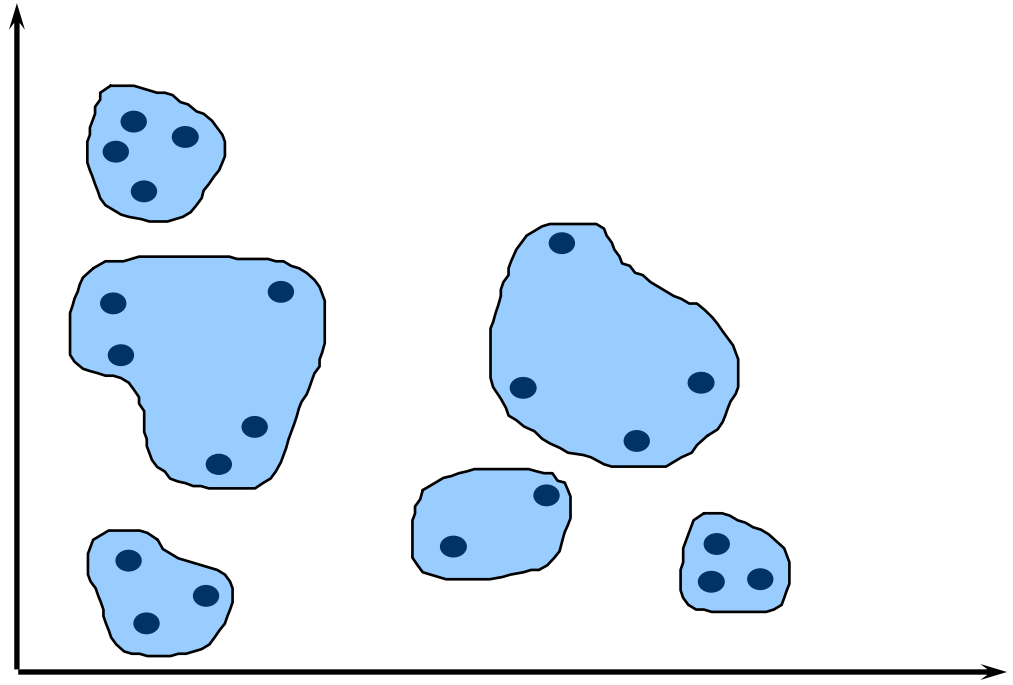
原始干扰信号



True \ Predicted	线性调频	Barker 码	Frank 码	噪声调幅	噪声调频	灵巧噪声	梳状谱	其它
线性调频	100%	0%	0%	0%	0%	0%	0%	0%
Barker 码	0%	61%	38%	0%	0%	0%	0%	1%
Frank 码	0%	27%	72%	0%	0%	0%	0%	1%
噪声调幅	0%	0%	0%	100%	0%	0%	0%	0%
噪声调频	0%	0%	0%	100%	0%	0%	0%	0%
灵巧噪声	0%	0%	0%	0%	0%	100%	0%	0%
梳状谱	0%	0%	0%	0%	0%	20%	80%	0%

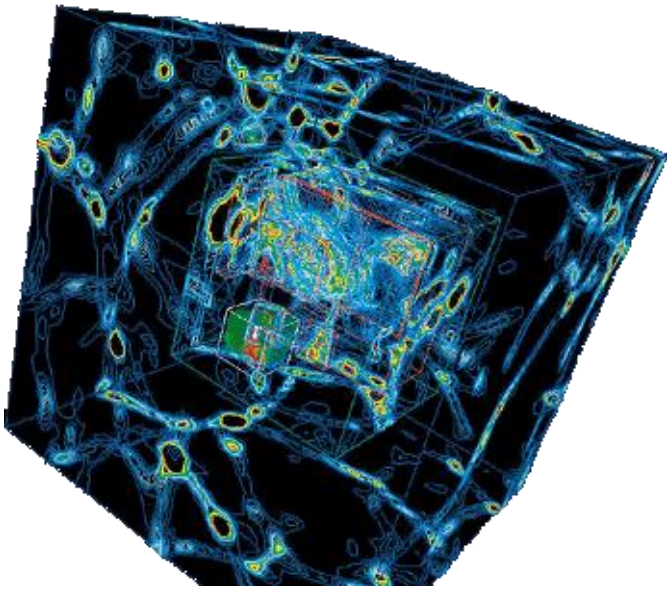
Clustering

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



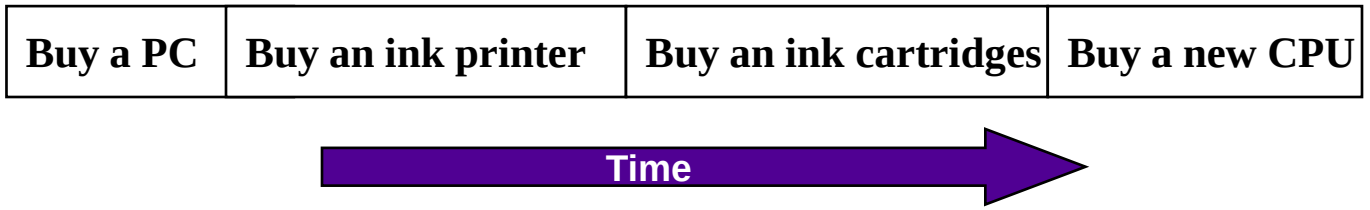
Ex.3 Scientific Simulation

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



Sequence Mining

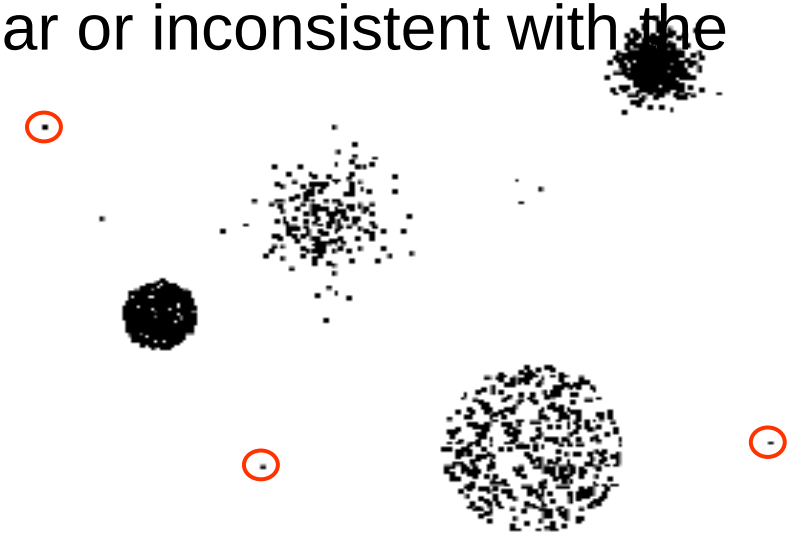
- Given a set of sequences, find the complete set of frequent subsequences



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

Anomaly Detection

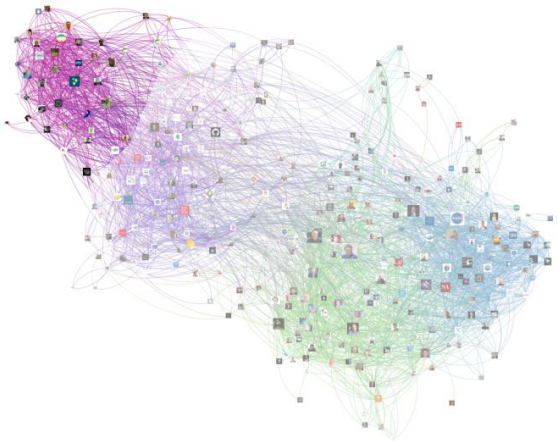
- What are anomalies?
  - The set of objects are considerably dissimilar from the remaining of the data
- Given a set of *n* objects, and *k,* the number of expected anomalies, find the top *k* objects that are considerably dissimilar or inconsistent with the remaining data



Anomalies may be valuable!

Social Analysis

- In social media mining
  - Detect communities
  - Communities evolution



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 iQIY

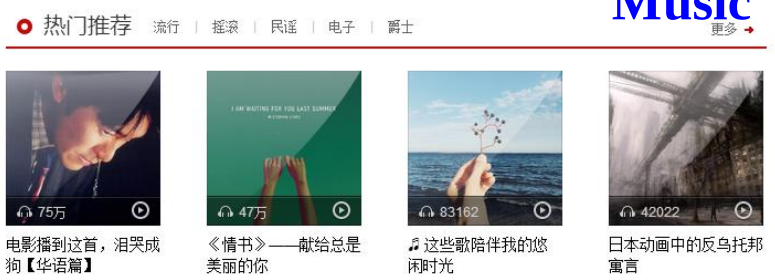
优酷 YOUKU

腾讯视频 V.QQ.COM

Music

QQ音乐

网易云音乐



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

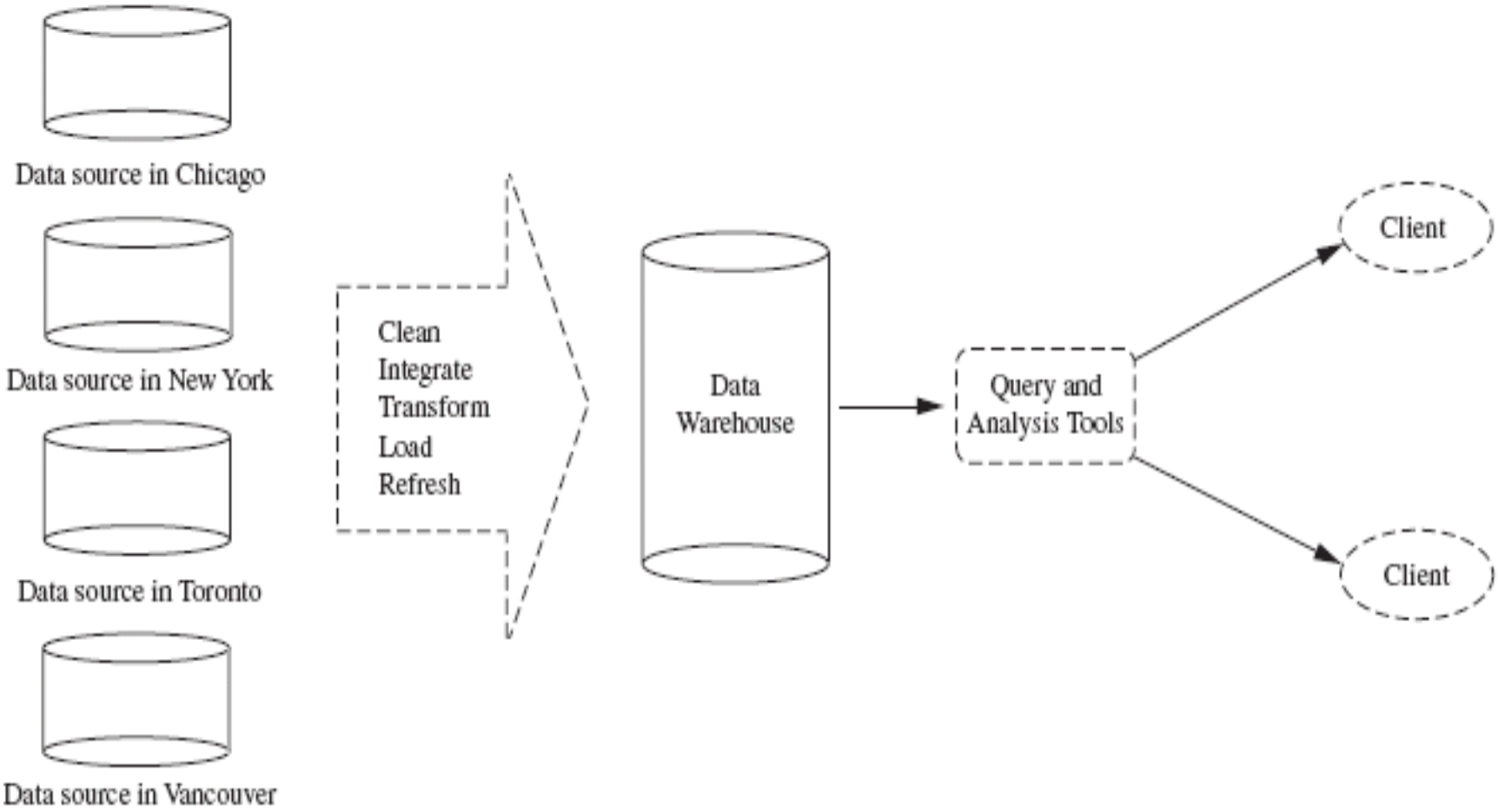
- Structured data
  - Table – records – attributes
  - Accessed by queries, SQL

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



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
…	…

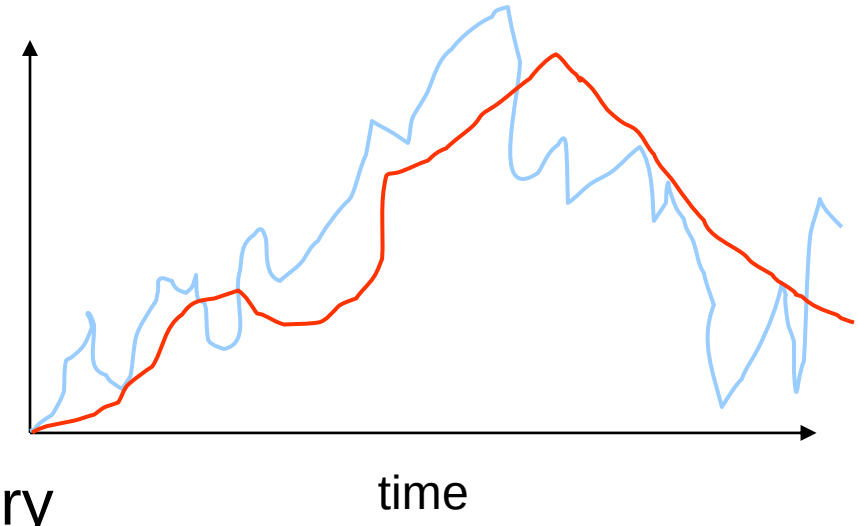
Spatial Data

- 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

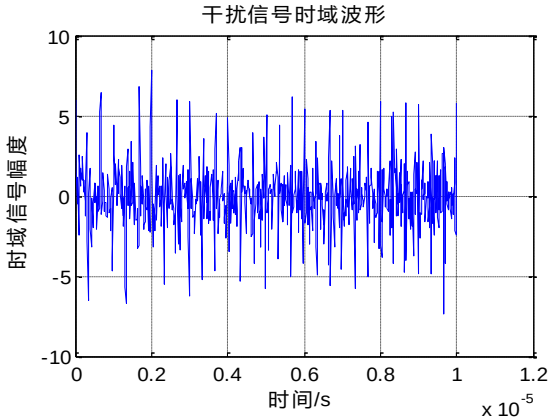


Text Databases

- HTML web documents
- XML documents
- Digital libraries

Multimedia Databases

- Multimedia databases
  - Image, audio and video data
  - Typical operations
    - Similarity-based pattern matching
    - 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, telecommunications data management, web application, sensor networks, etc.

Biomedical Data

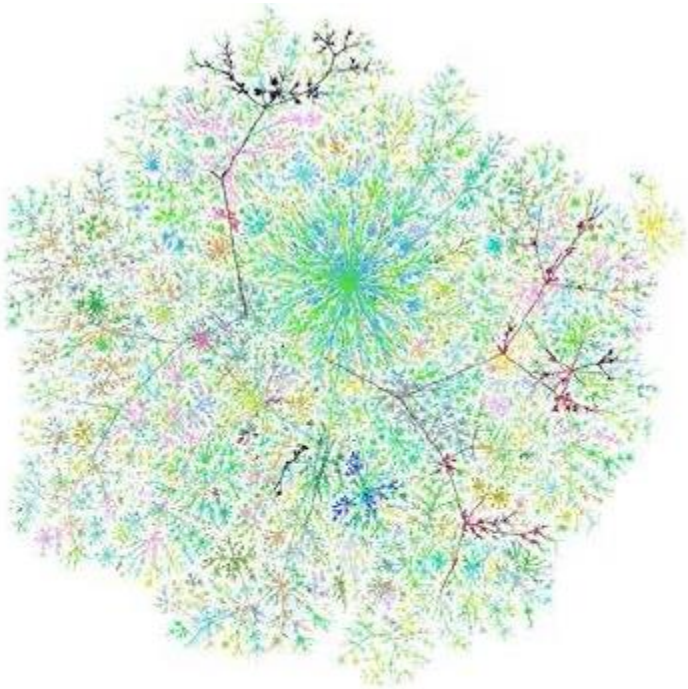
- Bio-sequences
  - DNA: very long sequences of nucleotides
  - Similarity search
  - Identify sequential patterns that play roles in various diseases
  - Association analysis: co-occurring gene sequences

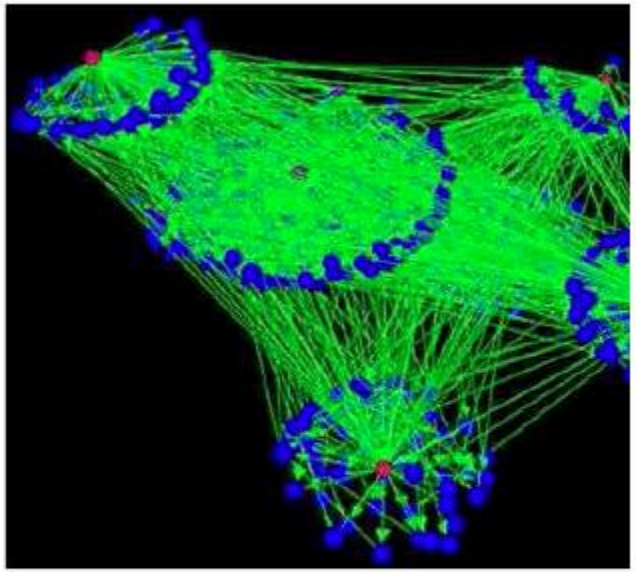
World-Wide Web

- The WWW is huge, widely distributed, global information service centre
  - 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

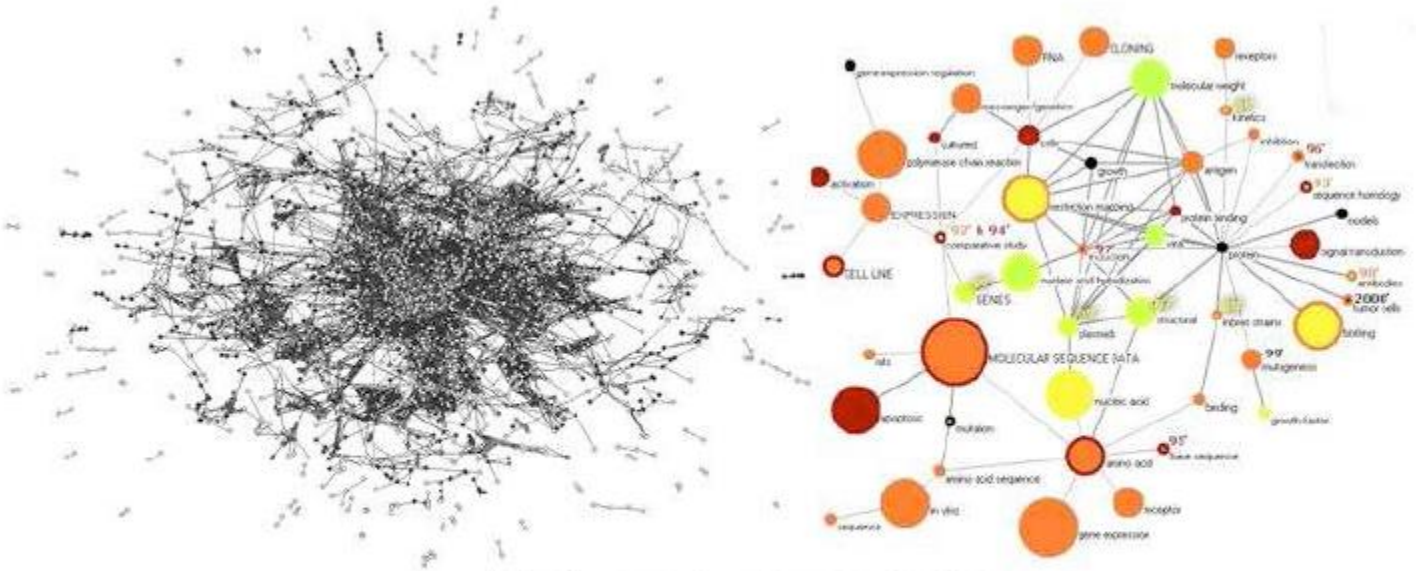




The images are downloaded from
http://www.maths.bris.ac.uk/~maarw/graphs/graph.html
and http://www.netdimes.org/new/?q=node/17

Graph

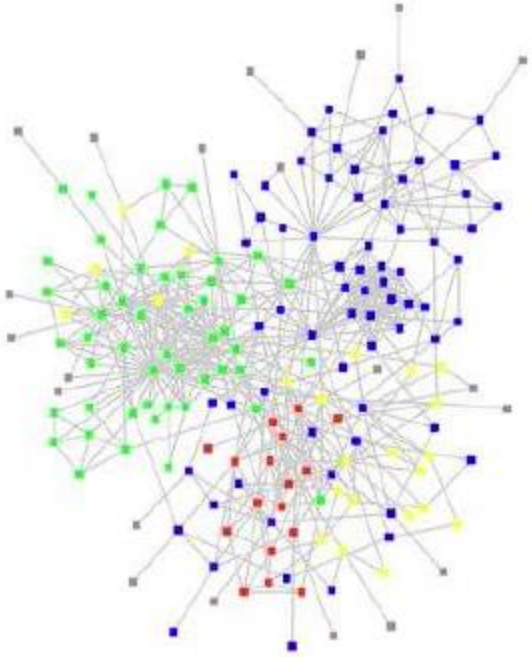
- Citation graph



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

Graph

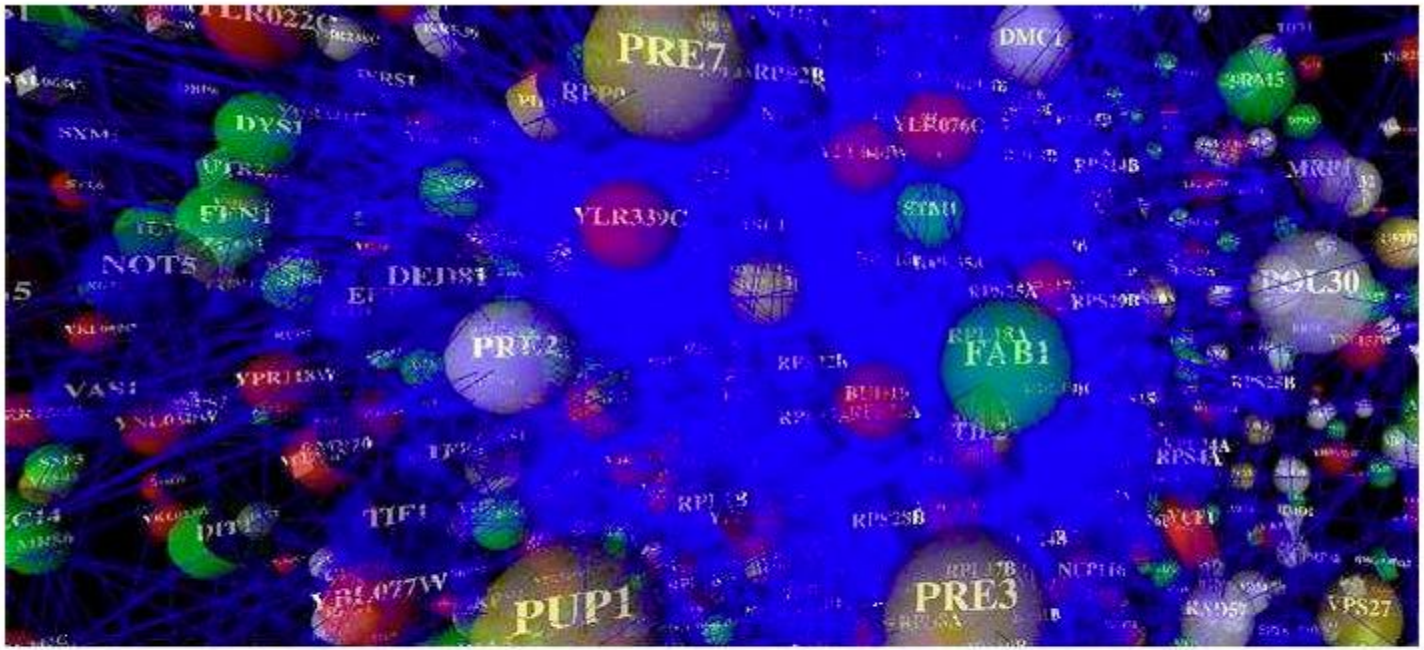
- Friendship graph



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/

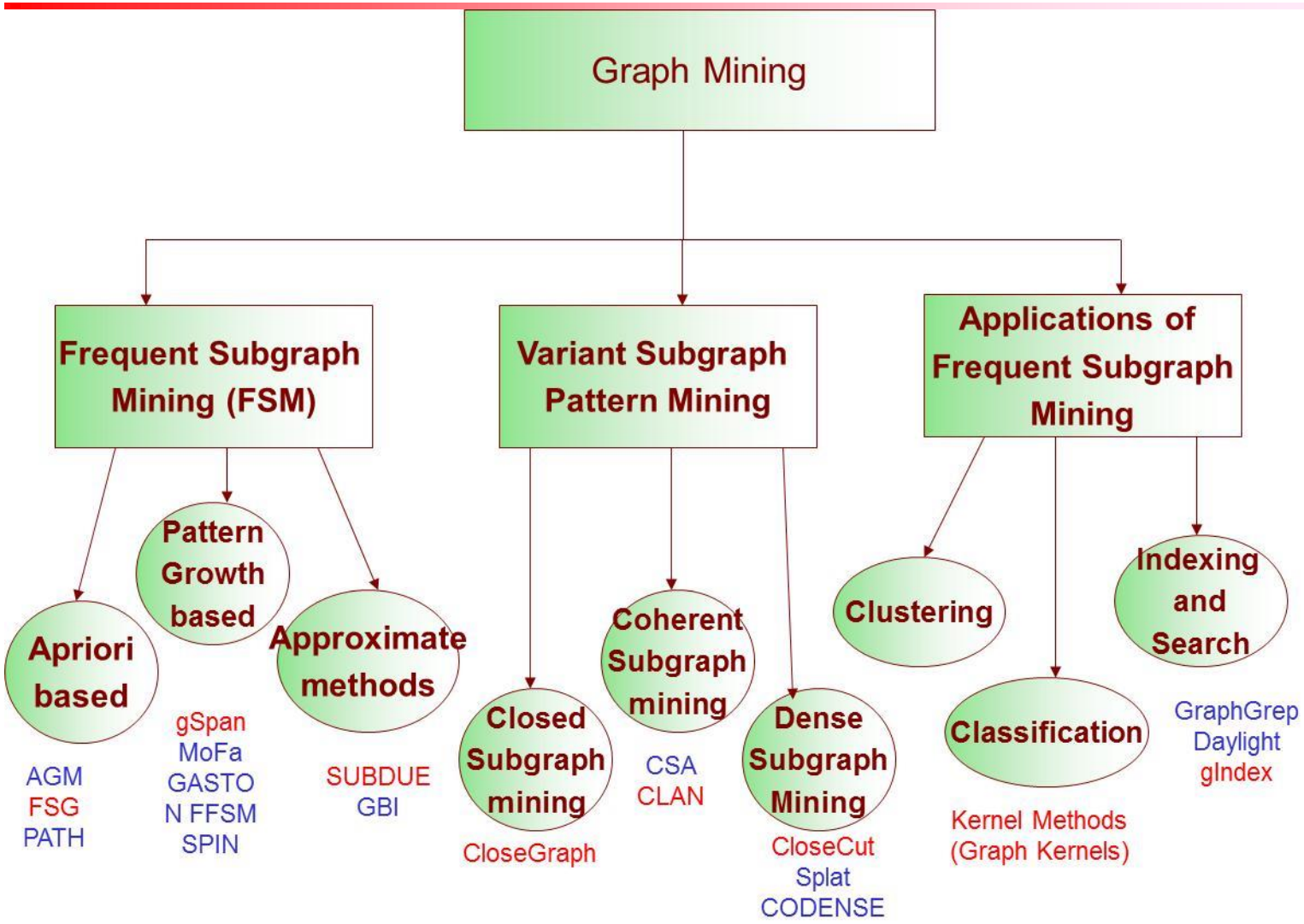
Graph

- 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

Applications

- Banking: loan/credit card approval
  - Predict good customers based on old customers
- Retail, telecommunication: customer relationship management
  - Identify those who are likely to leave for a competitor
- Retail: targeted marketing
  - Identify likely responders to promotions
- Telecommunications, finance: fraud detection
  - from an online stream of event identify fraudulent events

Applications (Continued)

- Medicine: disease outcome, effectiveness of treatments
  - Analyze patient disease history: find relationship between diseases
- Science: scientific data analysis
  - Identify new galaxies by searching for clusters
- WWW: website/store design and promotion
  - Find affinity of visitor to pages and modify layout

Success Cases

- Credit scoring
  - 根据中国人民银行的人口数据、信用卡、贷款、准贷记卡数据，挖掘信贷行为与信用表现的关系
  - 利用预测模型，首次建立了中国人民的信用局评分模型
  - K_S值达到0.51，获北京市科学技术二等奖
- Reservoir prediction
  - Predict the reservoir levels for Kipper 1 oil well
  - Build prediction models on BHP-Billiton's oil well log data
  - Achieved 75+% accuracy, almost as good as a domain expert

Success Cases (Continued)

- 面向天体模拟的高性能聚类算法
  - 利用聚类算法HOP，挖掘大规模天体模拟数据中的星系
  - 提出并实现了并行计算方法，获得高加速比
  - 被美国圣地亚哥超级计算中心(SDSC)使用
- High utility itemsets mining
  - 提出High utility itemsets mining算法
  - 在美国某连锁超市交易数据中，挖掘出高利润的商品集合

Success Cases (Continued)

- 光学遥感图像海面舰船识别

每类舰船的误分率

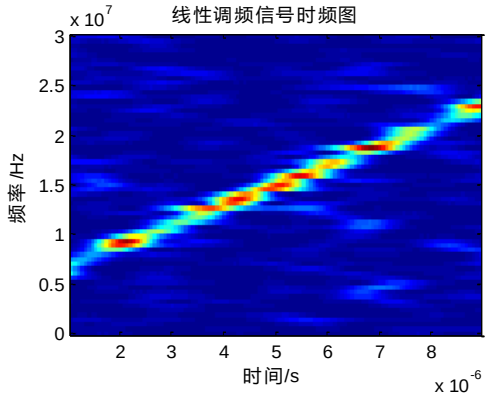
	登陆舰	航母	货船	集装箱	军舰_1	军舰_2	大型油轮	小型油轮	游艇	渔船
误分率	13%	6.5%	3.3%	16%	10%	6.5%	3.3%	0%	3.3%	3.3%

Success Cases (Continued)

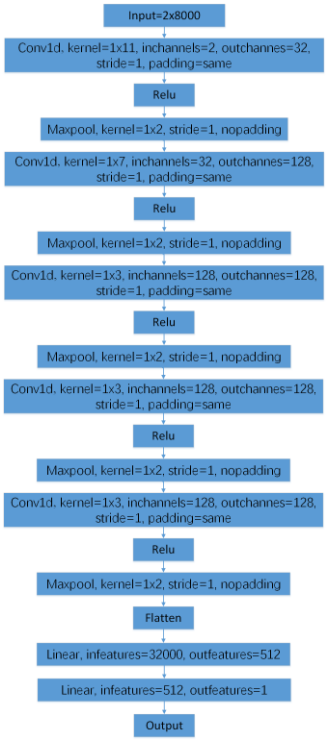
- 路面异物识别

Success Cases (Continued)

- 雷达信号干扰识别



时频图



True \ Predicted	线性调频	Barker 码	Frank 码	噪声调幅	噪声调频	灵巧噪声	梳状谱
线性调频	100%	0%	0%	0%	0%	0%	0%
Barker 码	0%	92%	8%	0%	0%	0%	0%
Frank 码	0%	0%	100%	0%	0%	0%	0%
噪声调幅	0%	0%	0%	92%	0%	8%	0%
噪声调频	0%	0%	0%	0%	100%	0%	0%
灵巧噪声	0%	0%	0%	0%	0%	100%	0%
梳状谱	0%	0%	0%	0%	0%	0%	100%

Exercises

Google, Baidu, Facebook, etc. are important Internet companies.

1. What kinds of data do they process?
2. What data mining techniques are they using to discover valuable knowledge?

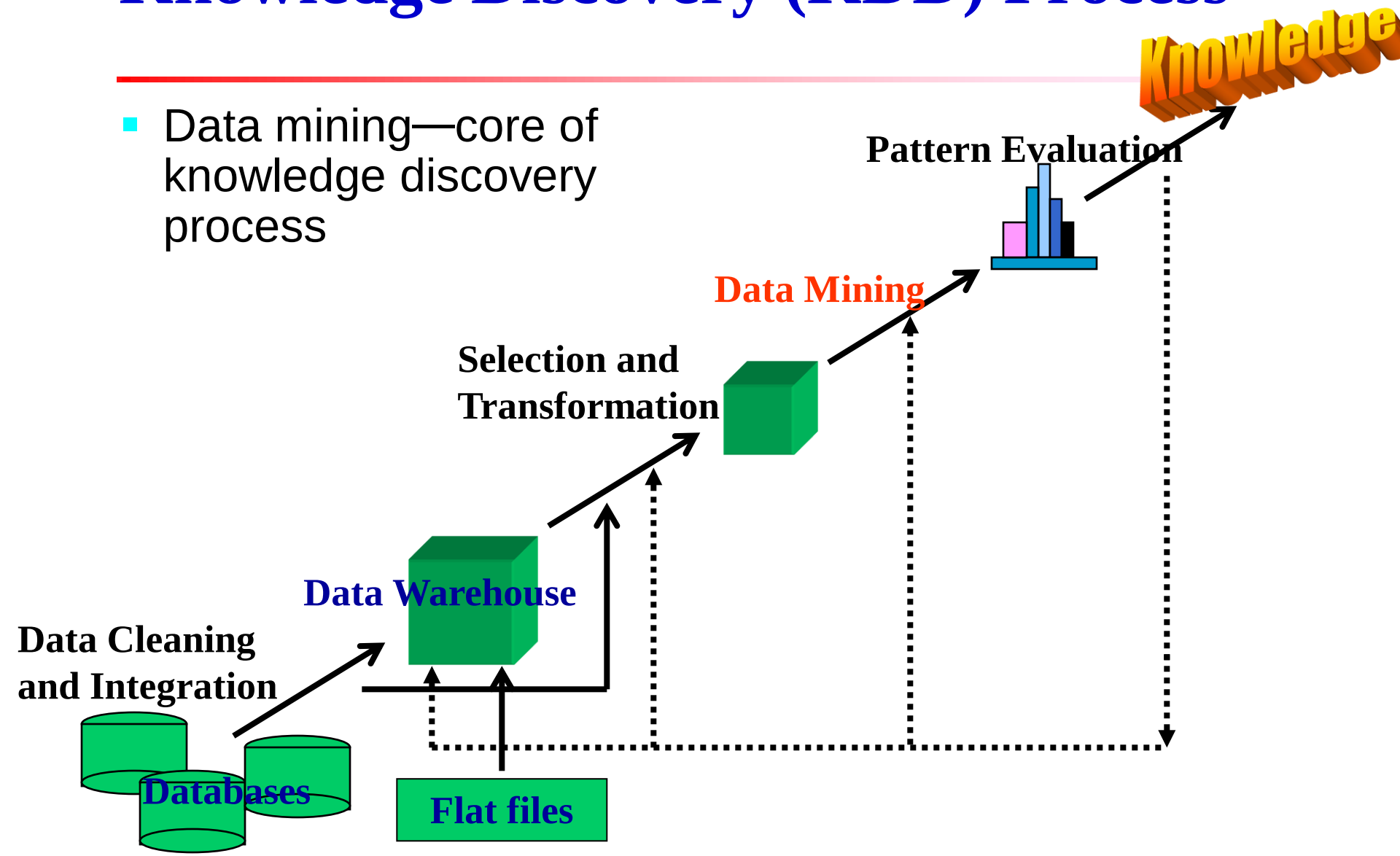
Exercises

Data mining is one of the important data analysis methods in scientific applications.

1. What kinds of data do they process?
2. What data mining techniques are they using to discover valuable knowledge?

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