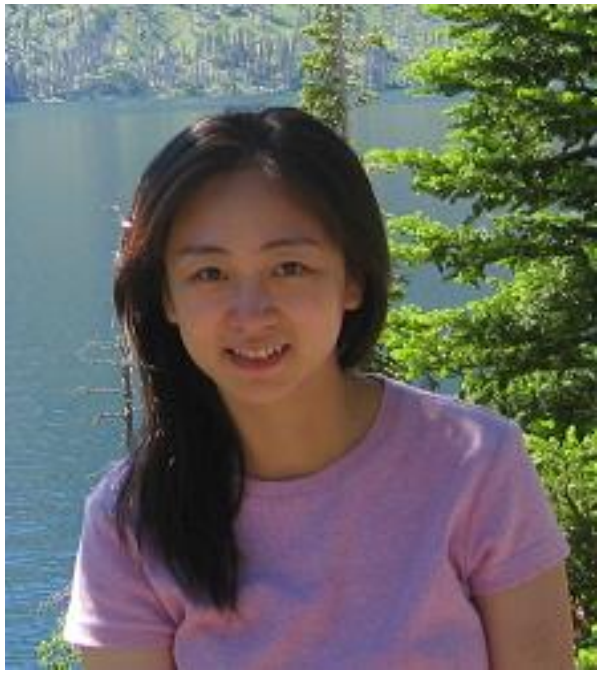
Data Mining

Ying Liu, Prof., Ph.D

School of Computer Science and Technology
University of Chinese Academy of Sciences
Data Mining and High Performance Computing Lab

Welcome

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

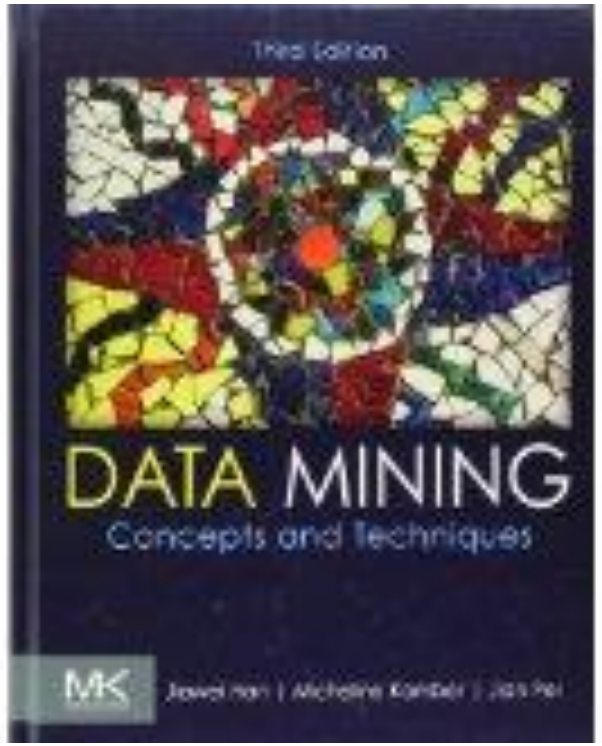


Useful Information

- Teaching Assistants
  - Wei, Qiancheng
  - Yuan, Yi
- Class: Monday & Wednesday 8:30 - 10:10, 教 1-207
- 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%)
  - 3 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

- Choose a topic from the following topics
- Read through some related research papers and fully understand them
- Develop and implement the method
- To be evaluated by the ranking

Project

- Option 1: 天池大赛（学习赛）
  - **题目——银行客户认购产品预测**

    (https://tianchi.aliyun.com/competition/entrance/531993/introduction)
  - 下载数据
  - 数据预处理
  - 建模
  - 模型验证
  - 提交结果到天池

银行客户认购产品预测

赛题简介

本次教学赛是陈博士发起的数据分析系列赛事第1场 —— 银行客户认购产品预测

赛题以银行产品认购预测为背景，想让你来预测下客户是否会购买银行的产品。在和客户沟通的过程中，我们记录了和客户联系的次数，上一次联系的时长，上一次联系的时间间隔，同时在银行系统中我们保存了客户的基本信息，包括：年龄、职业、婚姻、之前是否有违约、是否有房贷等信息，此外我们还统计了当前市场的情况：就业、消费信息、银行同业拆解率等。

用户购买预测是数字化营销领域中的重要应用场景，通过这道赛题，鼓励学习者利用营销活动信息，为企业提供销售策略，也为消费者提供更适合的商品推荐。

金融数据分析比赛的目的是为了更好地带动数据科学初学者一起玩起来，因此我们鼓励所有选手，基于赛题发表notebook分享，内容包含但不限于**对赛题的理解、数据分析及可视化、算法模型的分析以及数据分析思路**等内容。

【配套学习资源】：天池AI学习

【教学赛】是阿里云天池面向高校开放的、以AI教学为目的的公益教学赛事，欢迎申报。

申请链接>>

银行客户认购产品预测

赛题背景

赛题以银行产品认购预测为背景，想让你来预测下客户是否会购买银行的产品。在和客户沟通的过程中，我们记录了和客户联系的次数，上一次联系的时长，上一次联系的时间间隔，同时在银行系统中我们保存了客户的基本信息，包括：年龄、职业、婚姻、之前是否有违约、是否有房贷等信息，此外我们还统计了当前市场的情况：就业、消费信息、银行同业拆解率等。

赛题任务

To DO：预测用户是否进行购买产品

字段	说明
age	年龄
job	职业：admin, unknown, unemployed, management…
marital	婚姻：married, divorced, single
default	信用卡是否有违约: yes or no
housing	是否有房贷: yes or no
contact	联系方式：unknown, telephone, cellular
month	上一次联系的月份：jan, feb, mar, …
day_of_week	上一次联系的星期几：mon, tue, wed, thu, fri
duration	上一次联系的时长（秒）

Project

- Option 2: 天池大赛（教学赛）
  - **题目—— 心跳信号分类预测**

    (https://tianchi.aliyun.com/competition/entrance/531883/introduction)
  - 下载数据
  - 数据预处理
  - 建模
  - 模型验证
  - 提交结果到天池

心跳信号分类预测

赛题背景

赛题以医疗数据挖掘为背景，要求选手使用提供的心跳信号传感器数据训练模型并完成不同心跳信号的分类的任务。为了更好的引导大家入门，还特别为本赛题定制了学习方案，其中包括数据科学库、通用流程和baseline方案学习三部分。

通过对本方案的完整学习，可以帮助掌握数据竞赛基本技能。同时我们也将提供专属的视频直播学习通道。

一、赛题数据

赛题以预测心电图心跳信号类别为任务，数据集报名后可见并可下载，该数据来自某平台心电图数据记录，总数据量超过20万，主要为1列心跳信号序列数据，其中每个样本的信号序列采样频次一致，长度相等。为了保证比赛的公平性，将会从中抽取10万条作为训练集，2万条作为测试集A，2万条作为测试集B，同时会对心跳信号类别（label）信息进行脱敏。

字段表

Field	Description
id	为心跳信号分配的唯一标识
heartbeat_signals	心跳信号序列
label	心跳信号类别（0、1、2、3）

Project

- Option 3: 天池大赛（算法大赛）
  - **题目——支撑分布式储能系统优化部署的新能源产量预测**

    (https://tianchi.aliyun.com/competition/entrance/532022/introduction)
  - 比赛时间
    - 报名阶段：2022年8月15日至2022年9月15日
    - 初赛阶段：2022年9月1日至2022年9月30日
    - 复赛阶段：2022年10月5日至2022年11月5日
    - 决赛阶段：时间待定

支撑分布式储能系统优化部署的新能源产量预测

背景介绍

本算法大赛由国际联合实验室UNiLAB– Big Data Analytics for Smart Energy Systems创办，旨在推动引入先进人工智能和数据分析技术突破智慧能源领域关键问题。首届比赛设立三个主题赛道，依托阿里云天池平台开展，详细题目设置、数据和评价标准附后。

赛程安排

本次大赛分为初赛、复赛和决赛三个阶段，整体时间安排如下：

报名阶段：2022年8月15日至2022年9月15日
初赛阶段：2022年9月1日至2022年9月30日
复赛阶段：2022年10月5日至2022年11月5日
决赛阶段：时间待定

支撑分布式储能系统优化部署的新能源产量预测

赛题背景

在电力系统中，需要根据分布在不同地理位置的用户实时需求，向用户调度所需能源存储在用户本地。为优化分布式储能系统中的能源部署，准确预测用户用电需求极为重要。随着新能源技术的普及，用户能够自主安装新能源设备提供自身电力需求，从而减少对传统电能的需求和花销。我们提出支撑分布式储能系统优化部署的新能源产量预测问题。需要参与者利用人工智能相关技术根据历史数据预测用户未来新能源产量。

赛题描述

图1 用户用电来源示意图

如图1所示，用户日常用电由两部分组成：太阳能发电和从电力市场购买。太阳能发电量受环境太阳能量辐射强度有关，表示为$P(W)=\eta EA$，其中E为太阳光辐射强度，单位为W/m^2,A=2为太阳能电池板面积，单位为m^2,η=0.5为太阳能电池板转换效率。当太阳能供给不足时，用户从电力市场购买额外电能。

问题：给定用户平均每日总用电需求，以及与太阳能生产量相关的历史环境信息，对用户未来太阳能生产量进行预测。

数据说明

初赛数据

同一地点310天中每天每小时的归一化环境温度（temp.csv）和归一化风向和风速数据(wind.csv)，300天中每天6：00-20：00每小时归一化太阳辐射强度(sunshine.csv)数据。

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
  - Widely used 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
- Clustering
- Association rules
- Applications
  - credit scoring, target marketing, oil exploration, radar target detection & recognition
- Big 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

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

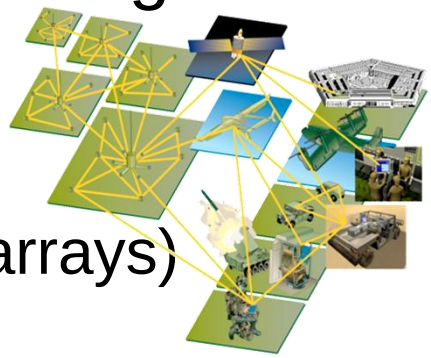
What Motivated Data Mining – Business World

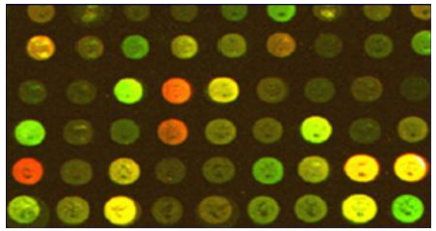
- Tremendous of data being collected and stored
  - E-commerce
  - Transactions
  - Stocks
  - Credit card transactions

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

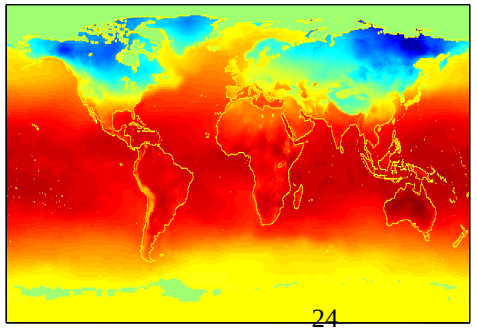
What Motivated Data Mining – Scientific World

- 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?

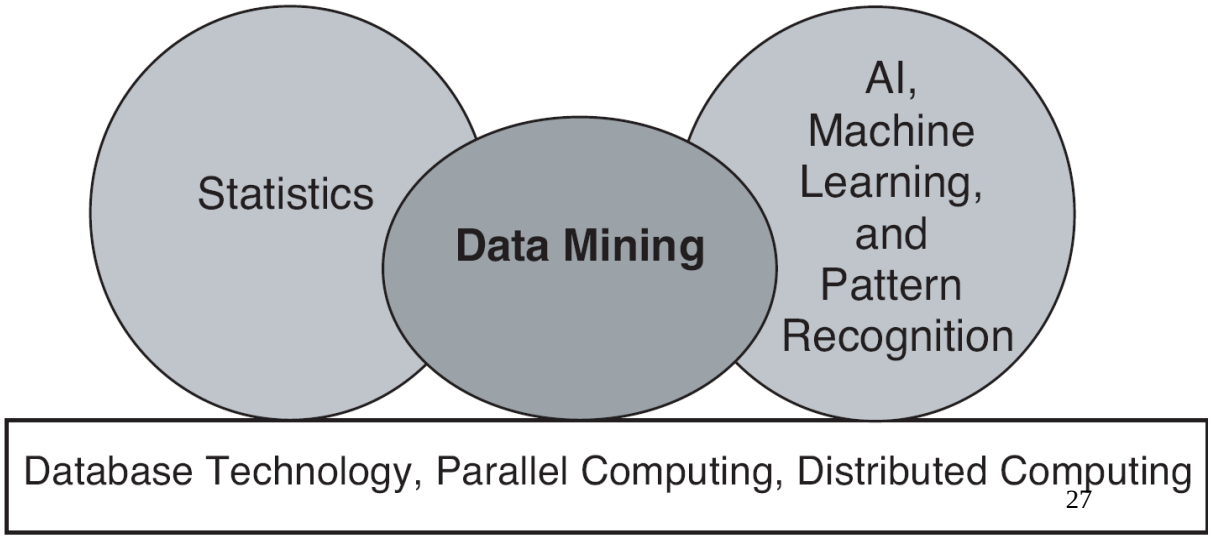
- We are drowning in data, but starving for knowledge!
  - Data rich, knowledge poor
  - Decision makers, domain experts have biases or errors
- Automated analysis of massive data sets

What is Data Mining?

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

What is Data Mining?

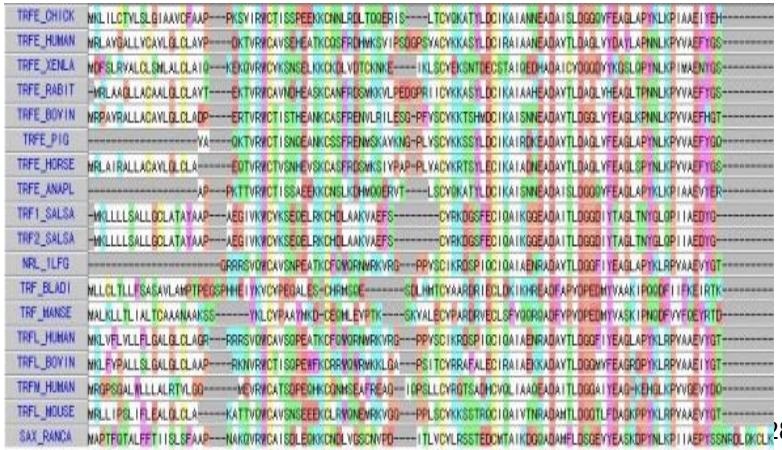
- Cross Disciplines
  - Databases
  - Machine learning: decision tree, Bayesian classifier, etc.
  - Statistics: regression, etc.
  - Neural networks
  - Parallel/Distributed computing



Why Not Traditional Data Analysis?

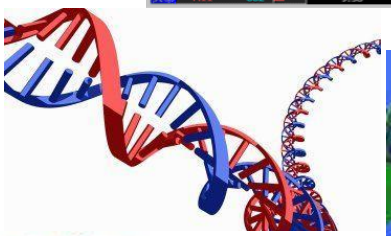
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data

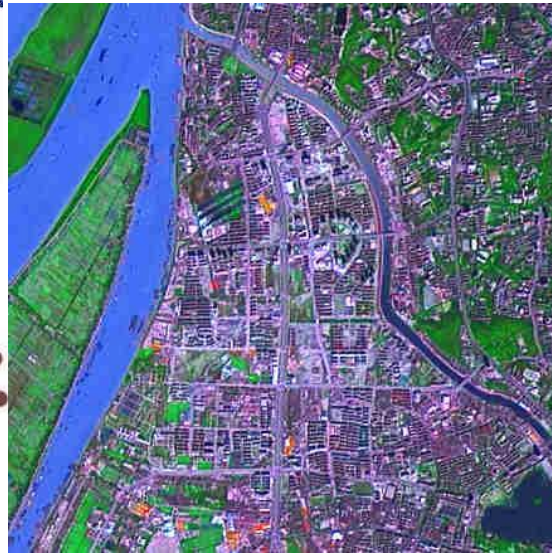
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions



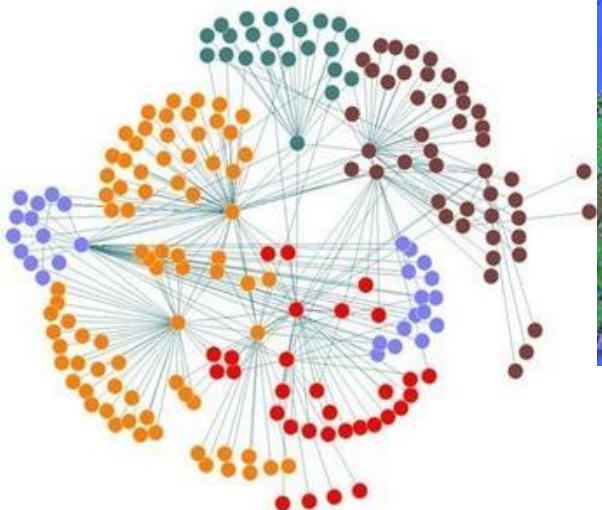
Why Not Traditional Data Analysis?

- 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

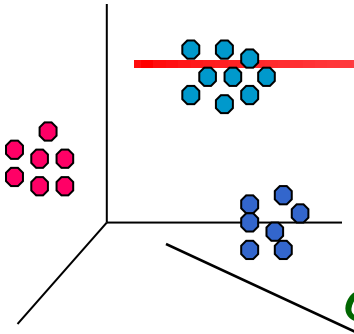
- Massive dataset
- Automatically searching for interesting patterns from historical data
- Fast
- Scalable
- Update easily
- Practical
- Decision support

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 Tasks

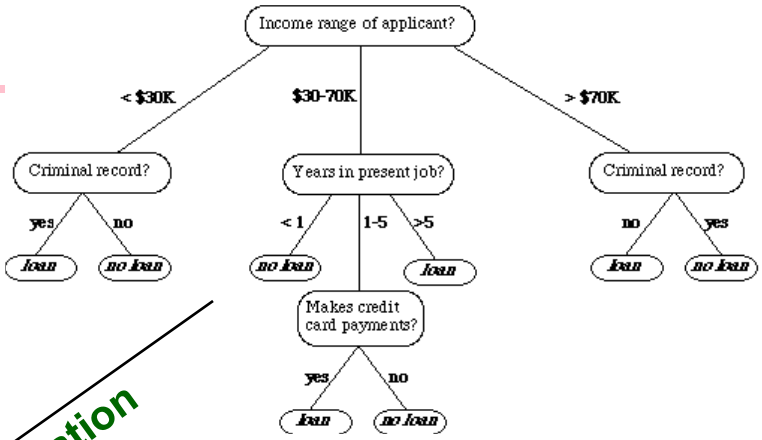


Clustering

Data

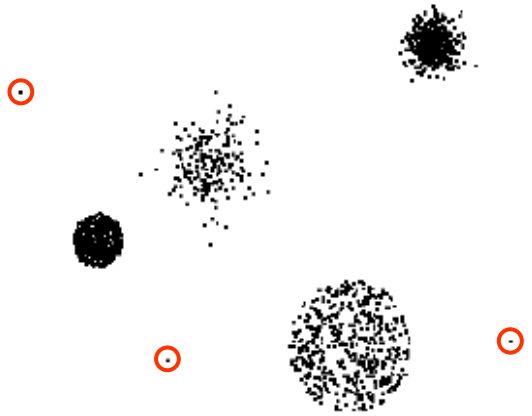
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Classification



Association Rules

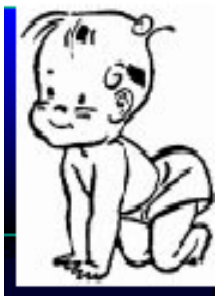
Anomaly Detection



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!



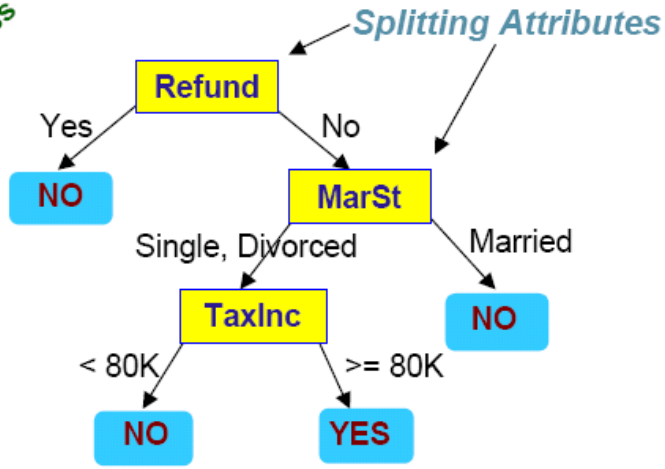
Ex. 1: Production Recommendation

- Where does the data come from?
  - supermarket transactions, membership cards, shopping lists, 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

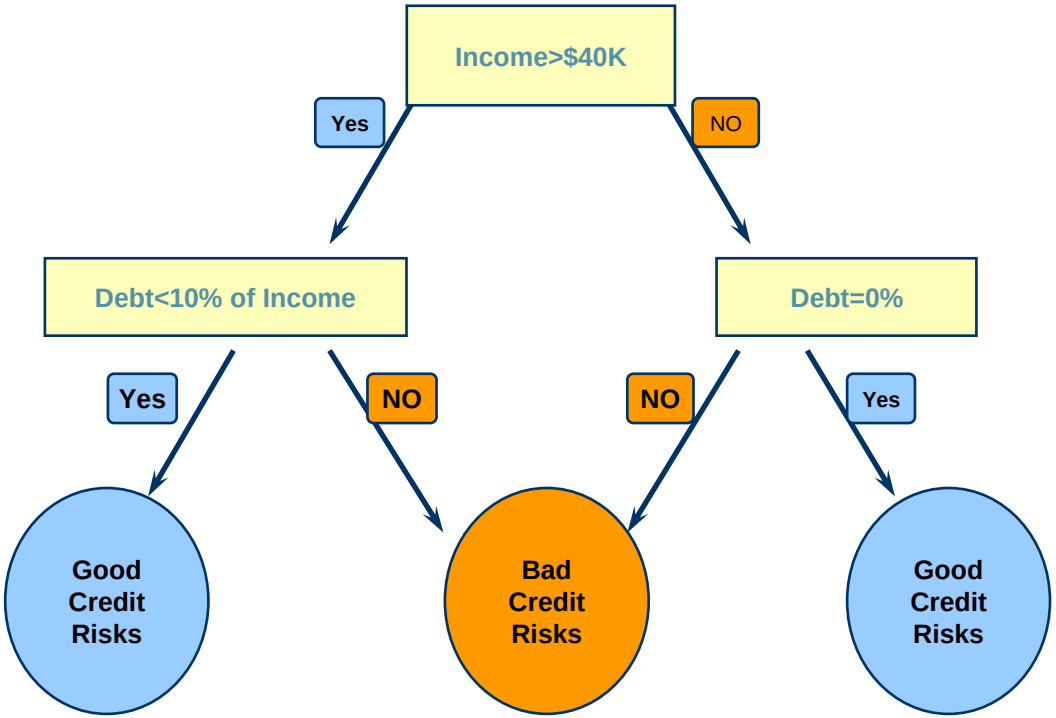


The splitting attribute at a node is determined based on the Gini index.

- Decision Tree

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

Ex.2 Credit Scoring



- Decision Tree

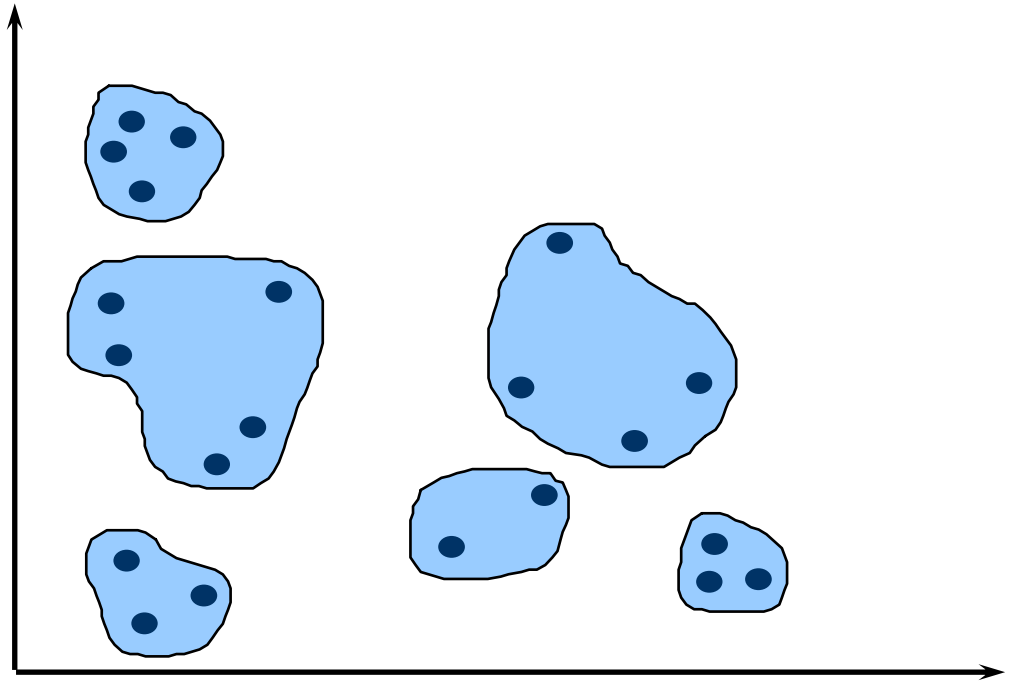
rule 1：if (Income<=$40k) and (Debt=0) then “good”
rule 2: if (Income>$40K) and (Debt<10% of Income) then “good”

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

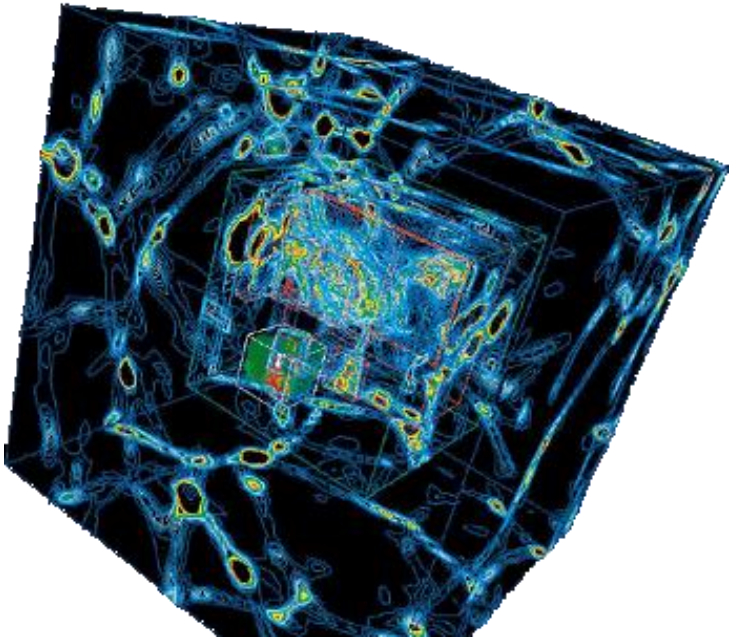
Clustering

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



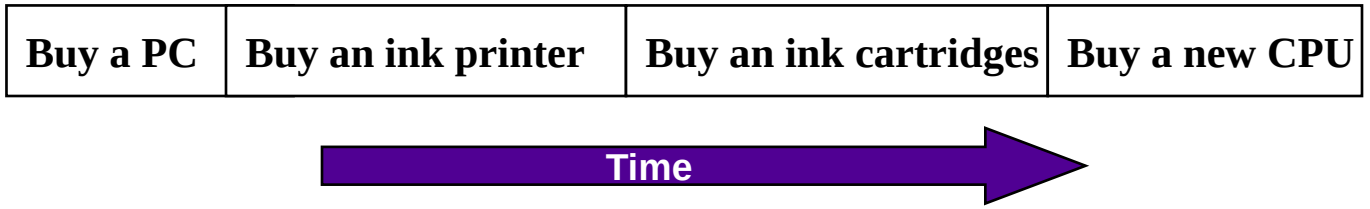
Ex.3 Scientific Simulation

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



Sequence Mining

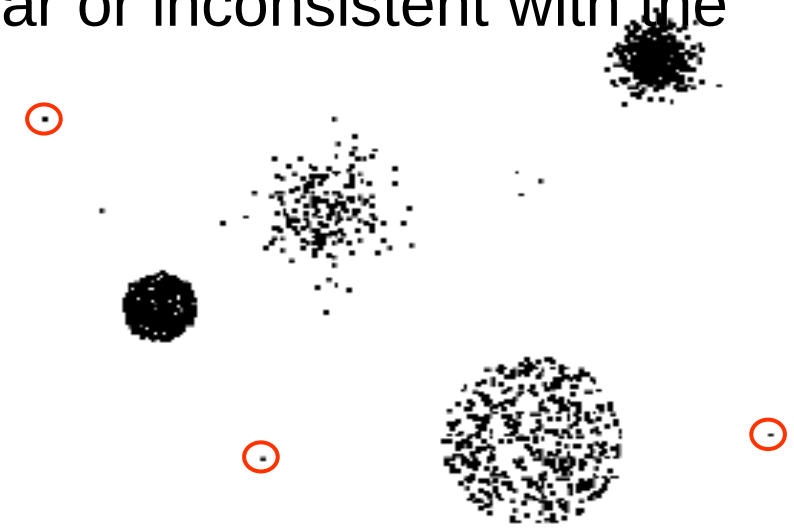
- Given a set of sequences, find the complete set of frequent subsequences



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

Anomaly Detection

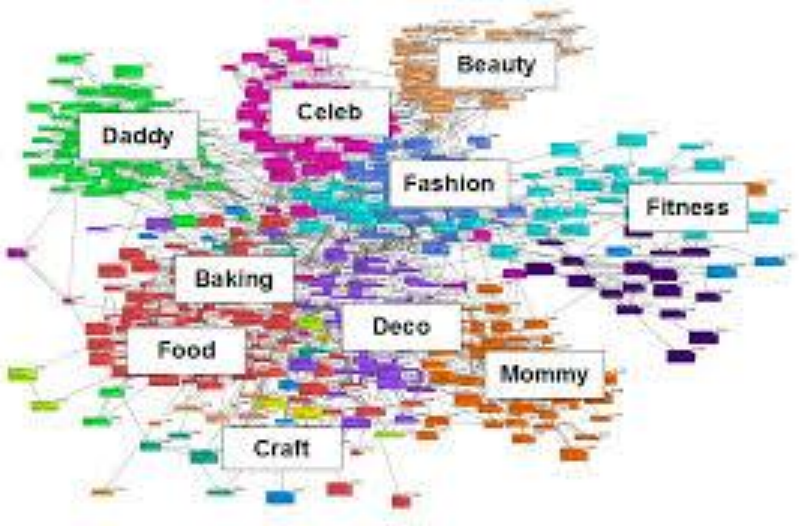
- What are anomalies?
  - The set of objects are considerably dissimilar from the remaining of the data
- Given a set of *n* objects, and *k,* the number of expected anomalies, find the top *k* objects that are considerably dissimilar or inconsistent with the remaining data

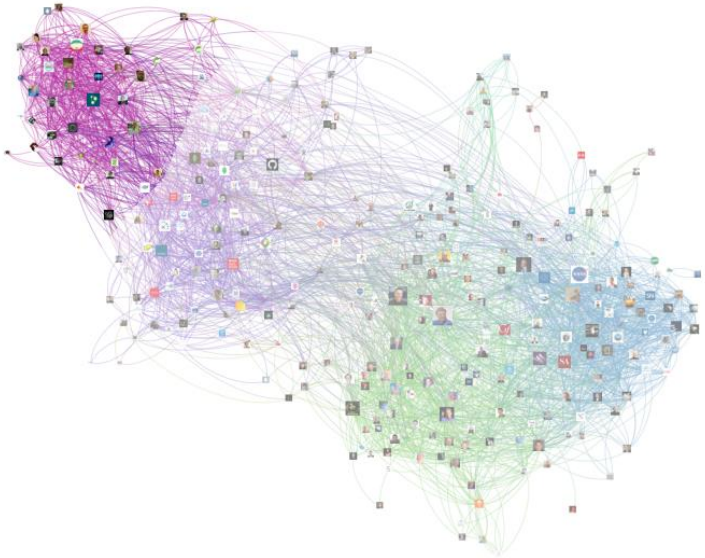


Anomalies may be valuable!

Social Analysis

- 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.京东 .COM

天猫 TMALL.COM

Movie

iQIYI 爱奇艺

YOUKU 优酷

腾讯视频 V.QQ.COM

Music

QQ音乐

网易云音乐

Exercises

1. 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

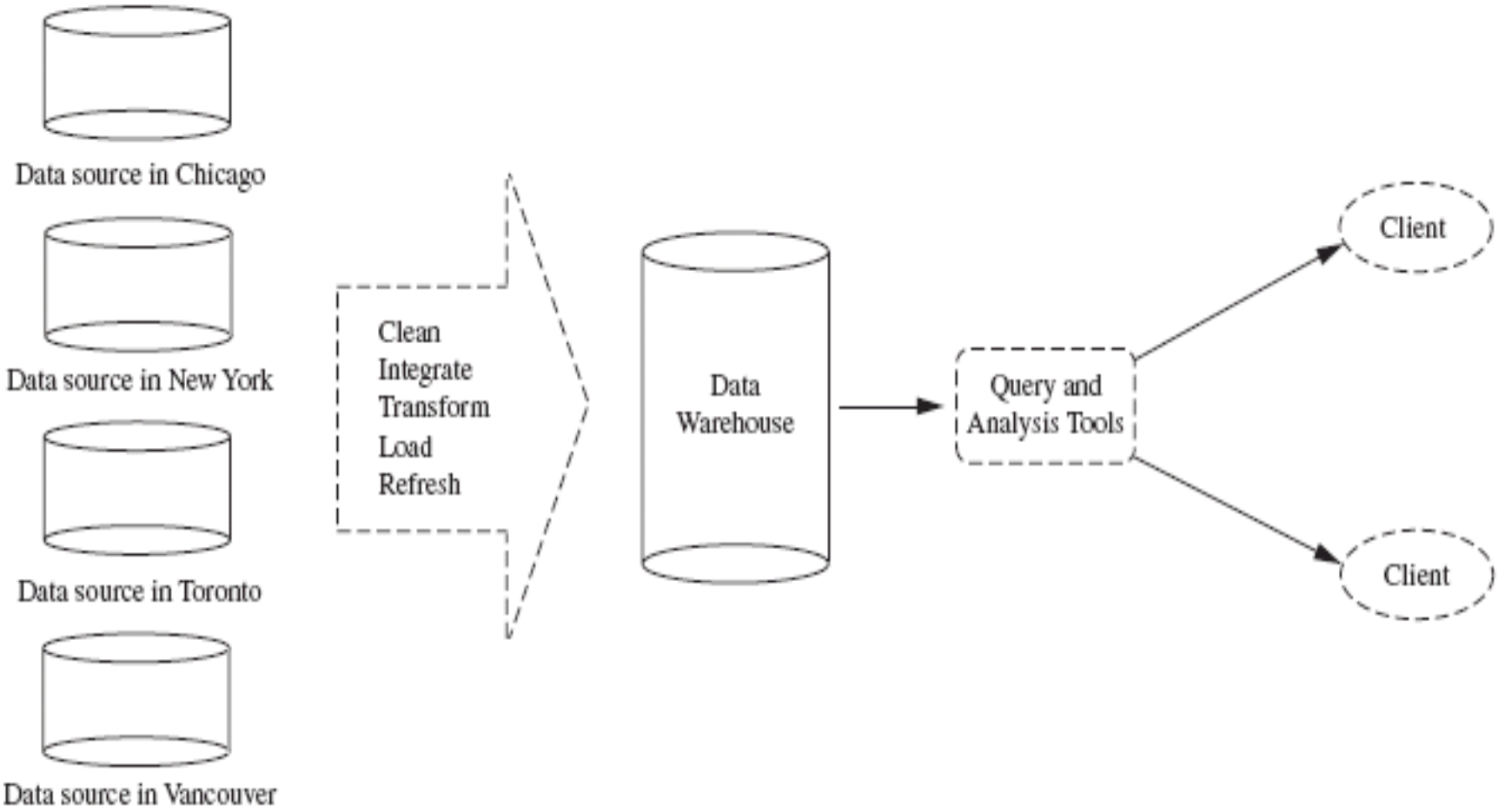
- Structured data
  - Table – records – attributes
  - Accessed by queries, SQL
- Online transactional processing (OLTP)
  - Insert a student “Ying Liu” into class “Introduction to Data Mining”, fall 2014

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
…	…

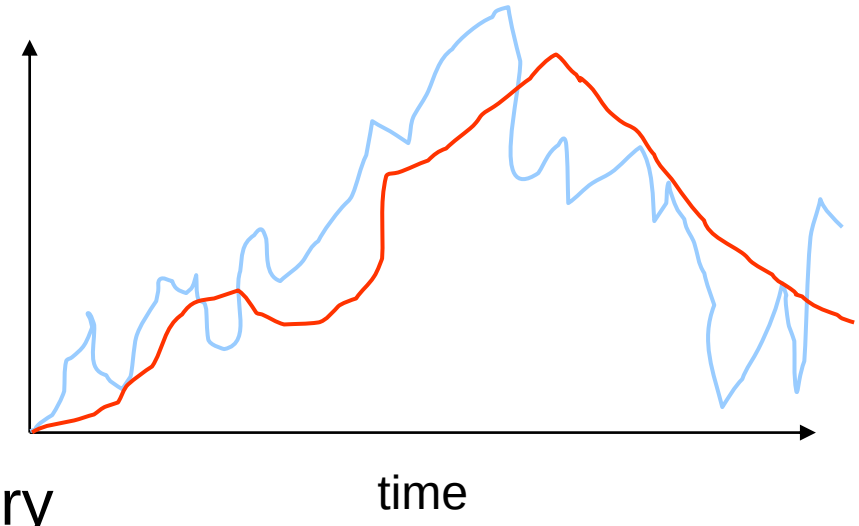
Spatial Data

- 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 & 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

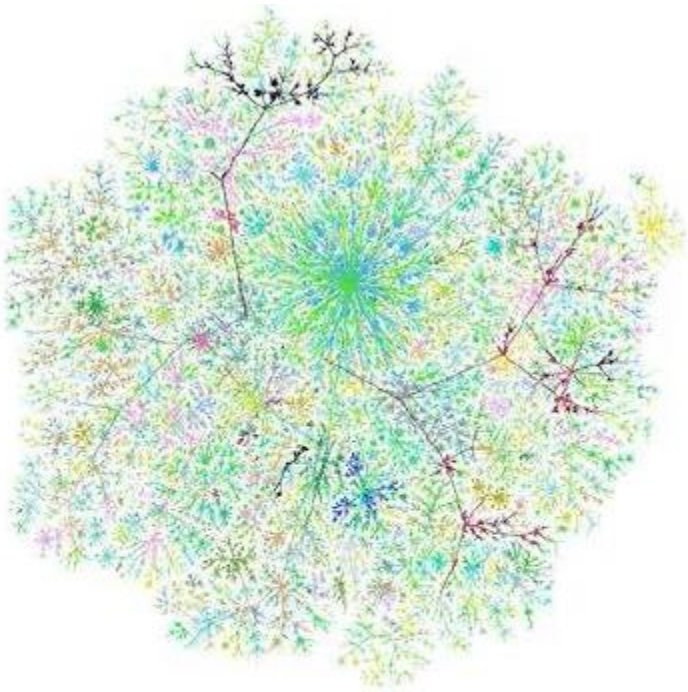
- The WWW is huge, widely distributed, global information service center
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for data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

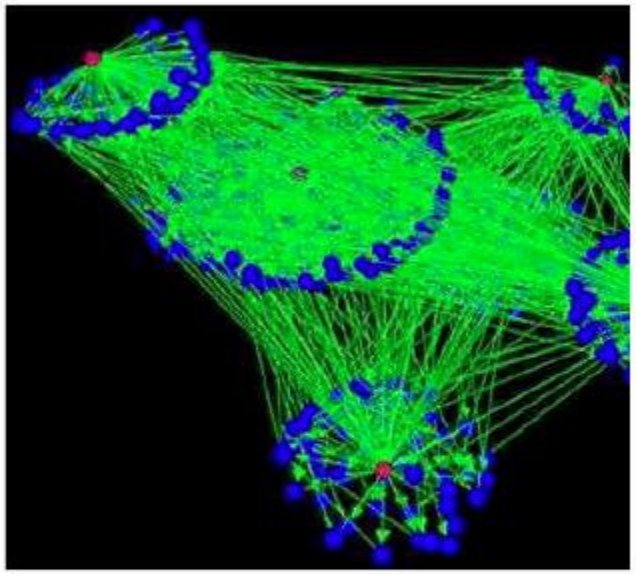
World-Wide Web

- 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

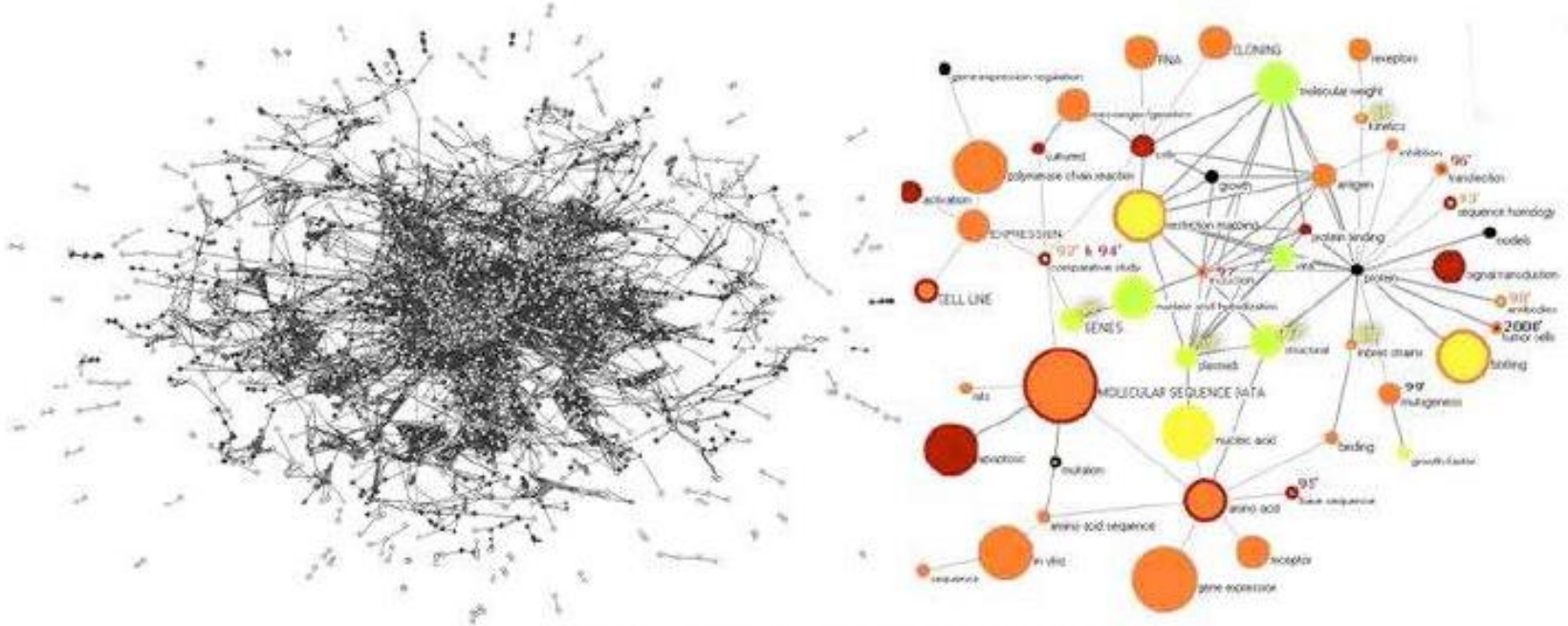




The images are downloaded from
http://www.maths.bris.ac.uk/~maarw/graphs/graph.html
and http://www.netdimes.org/new/?q=node/17

Graph

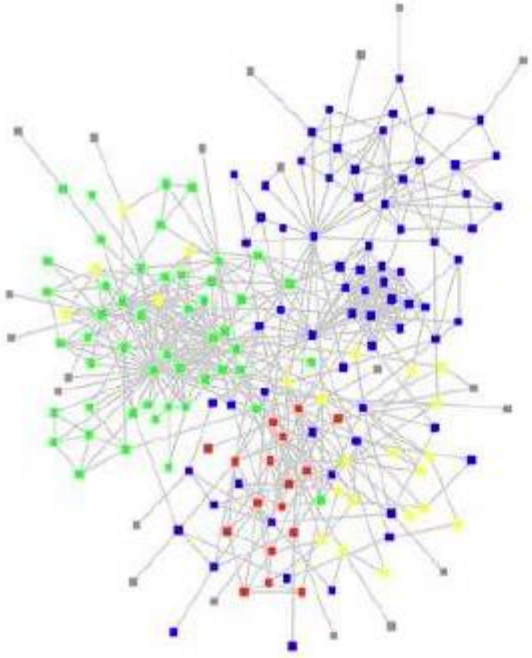
- Citation graph



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

Graph

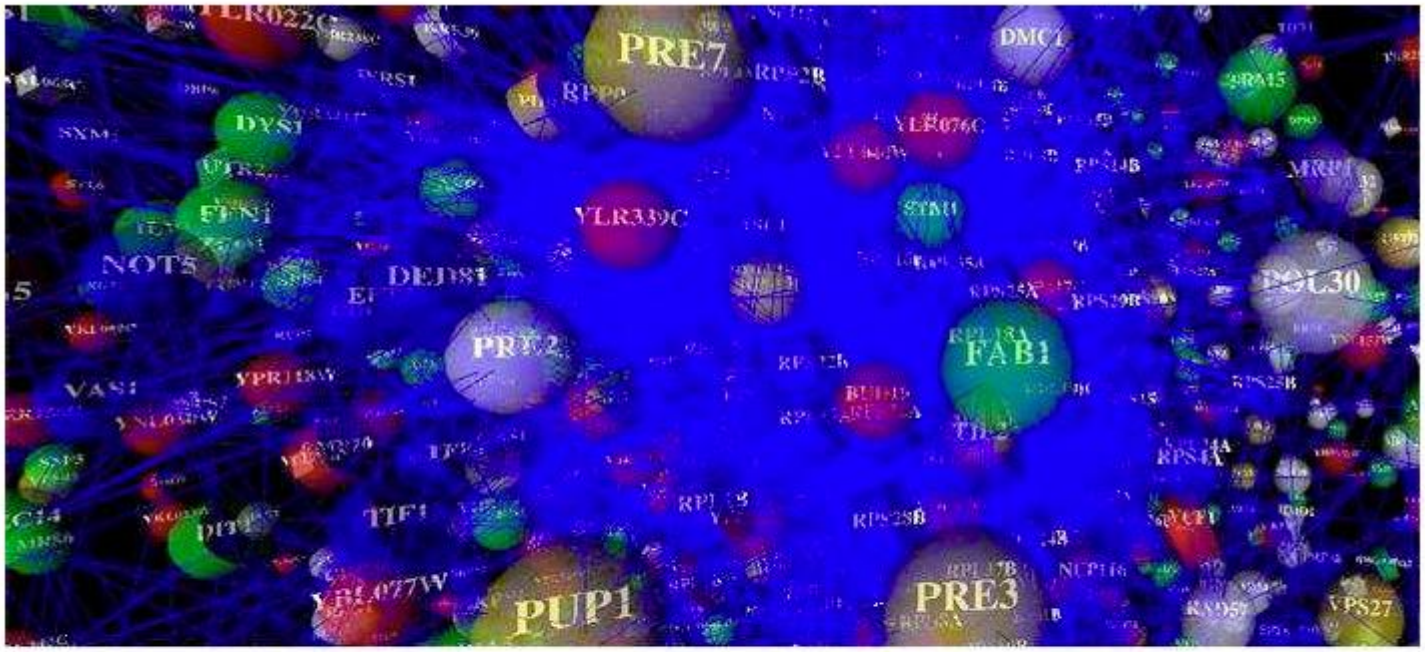
- Friendship graph



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/

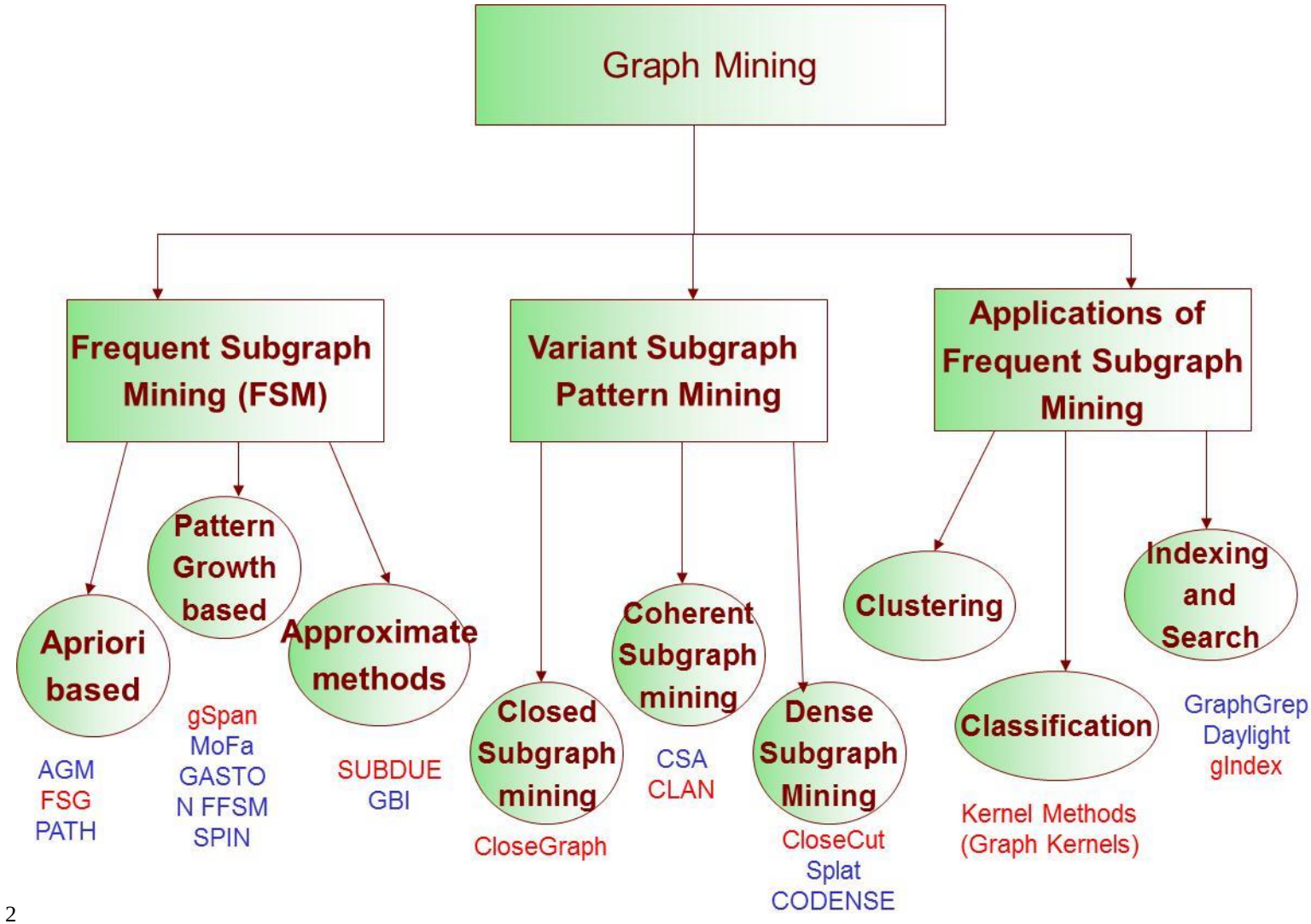
Graph

- Protein interaction graph



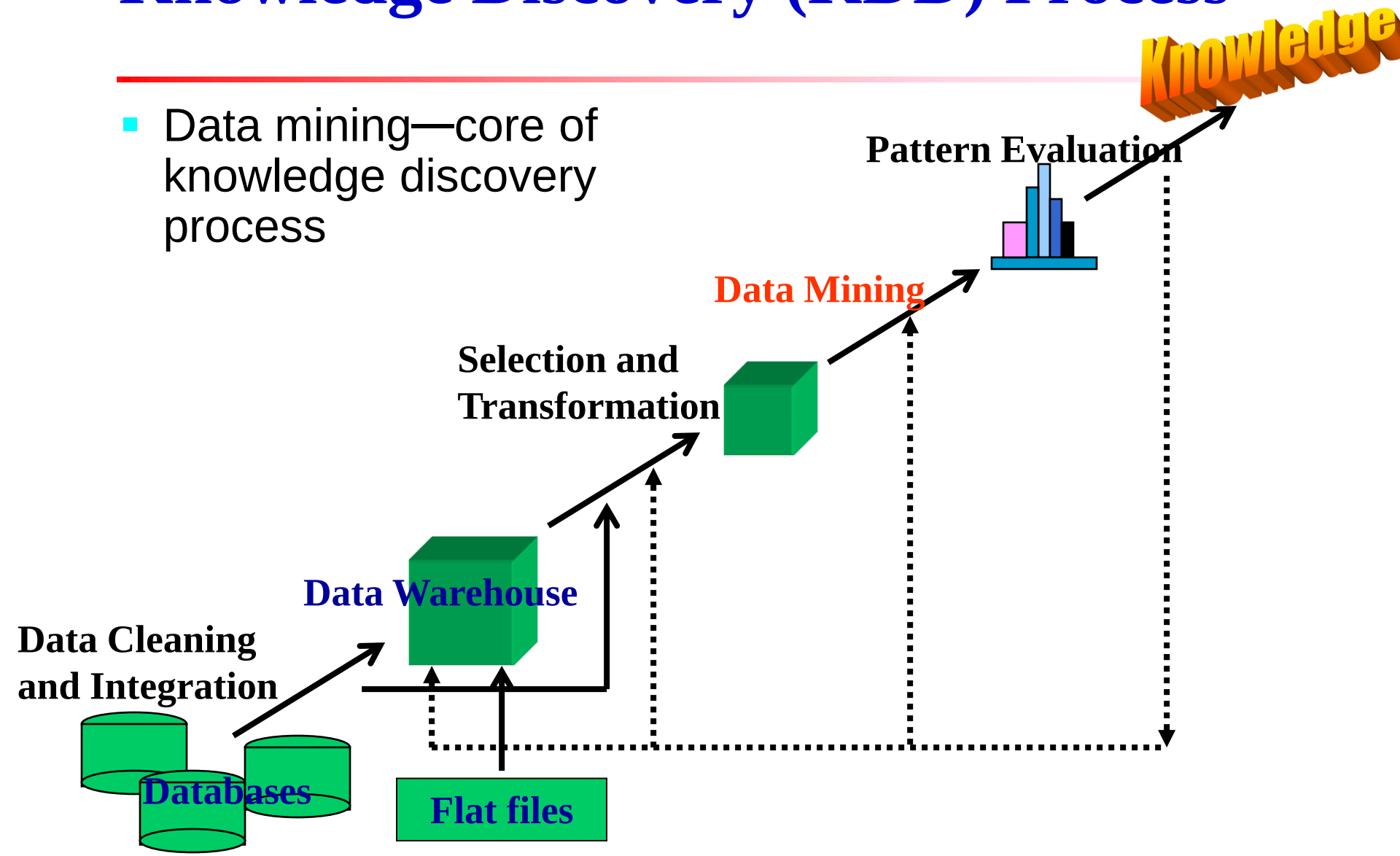
The images are downloaded from
http://bioinformatics.icmb.utexas.edu/lgl/Images/rsomZoom.jpg

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 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., image, audio, text, Web, graph, bio, stream
  - 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