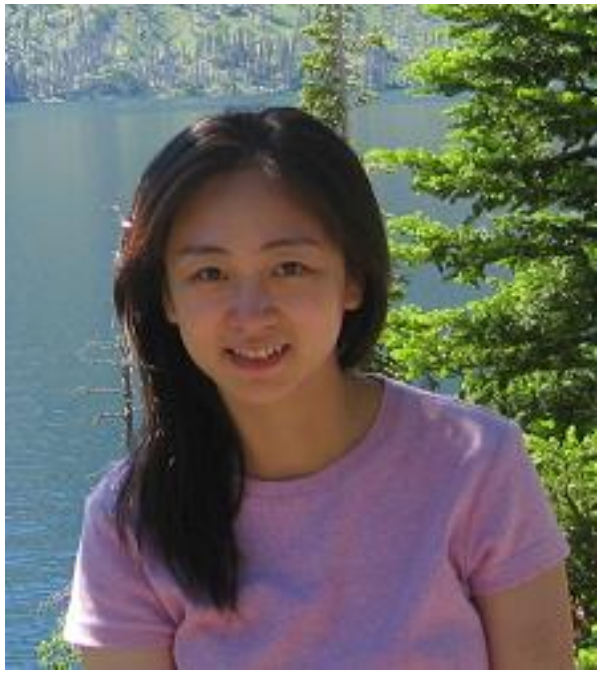
Data Mining

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

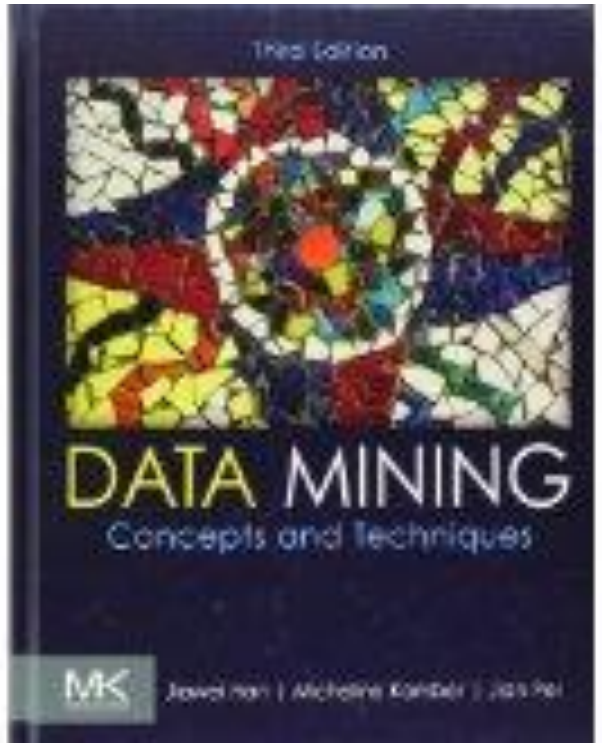


Useful Information

- Teaching Assistants
  - Wei, Qiancheng
  - Jiang, Wen
- 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: 天池大数据竞赛（正式赛）
  - **题目——糖尿病性黄斑水肿 (DME) 患者的Anti-VEGF抗血管内皮生长因子(简称Anti-VEGF)治疗转归预测**

  (https://tianchi.aliyun.com/competition/entrance/531929/introduction)

  初赛时间：2021/09/01 - 2021/10/31

  复赛时间：2021/11/02 – 2021/11/19

糖尿病性黄斑水肿 (DME) 患者的Anti-VEGF抗血管内皮生长因子(简称Anti-VEGF)治疗转归预测

任务和主题

抗血管内皮生长因子（Anti-VEGF）治疗指一组可减少新生血管生长或黄斑水肿的药物。Anti-VEGF药物可用于治疗多种导致黄斑（位于眼球后极部视网膜）下新生血管生长或黄斑水肿的疾病。但是，有相当一部分患者对这种疗法无反应或反应不充分，据不同研究报告，这个数字在10% 到 50% 之间，尽管他们每月都规律接受Anti-VEGF注射治疗。

在这次同步竞赛中，您需要建立机器学习模型来预测糖尿病性黄斑水肿 (DME) 患者对治疗的反应。您将使用医院里收集的数千张影像，预测经过负荷治疗后，6个月时对Anti-VEGF治疗的反应。如果成功，眼科医生就能借此定制治疗计划，并确保为患者提供及时有效的治疗。

比赛数据说明

CSV 文件:

patient ID	gender	age	diagnosis	preVA	anti-VEGF	preCST	preIRF	preSRF	prePED	preHRF	VA	continue injection	CST	IRF	SRF	PED	HRF
			1=wet														
	1=male		AMD		1=bevacizumab		1=present	1=present	1=present	1=present		1=yes		1=present	1=present	1=present	1=present
	2=female		2=PCV		2=ranibizumab		0=absent	0=absent	0=absent	0=absent		0=no		0=absent	0=absent	0=absent	0=absent
			3=DME		3=aflibercept												
			4=RVO		4=conbercept												
			5=CME		0=not receiving anti-VEGF												
			6=fellow eye(not receive anti-VEGF)														
			9=other diagnosis														

1.**Pretreatment Data** 每只眼睛的治疗前数据包含以下信息

- 性别值，1代表男性，2代表女性
- 患者年龄。每个年龄值均为整数
- 每只眼睛的诊断结果，为含义如下的类值：
  - 1=wet AMD 湿性AMD
  - 2=PCV 息肉样脉络膜血管病变

糖尿病性黄斑水肿 (DME) 患者的Anti-VEGF抗血管内皮生长因子(简称Anti-VEGF)治疗转归预测

结果提交要求（例如格式等）

包含患者 ID 和预测治疗后信息的 csv 文件。

patient ID	preCST	VA	continue	CST	IRF	SRF	HRF
0000-1315R							

数据格式

一个 csv 文件，其中每一列需为每个患者 ID 存储以下的预测治疗后信息。

- 患者 ID
- preCST 治疗前中心视网膜厚度是整数值，单位为微米。
- VA 是视力值
- 持续注射是整数，值为 0 和 1，其中 0 代表“否”，1 代表“是”
- CST 中心视网膜厚度是整数值，单位为微米
- IRF 视网膜层间积液是整数值，值为 0 和 1，其中 0 代表“无”，1 代表“有”
- SRF 神经上皮下积液是整数值，值为 0 和 1，其中 0 代表“无”，1 代表“有”
- HRF 视网膜高反射是整数值，值为 0 和 1，其中 0 代表“无”，1 代表“有”

检测指标及评判标准

对于每一行预测结果，对于preCST、VA和CST，预测误差在正负5%以内，该条目可得1分。对于其他4个0/1预测条目，预测正确可得1分，其他情况不得分。以最终总得分数排名。

虽然preIRF、preSRF、preHED、preHRF和HED的预测值并不直接参与得分评估，但选手仍然需要对其进行预测，以根据治疗前后症状情况等进行继续治疗判断。

Project

- Option 2: 天池大数据竞赛（长期赛）
  - **题目——天猫复购预测**

(https://tianchi.aliyun.com/competition/entrance/231576/introduction)

天猫复购预测

赛题背景

商家有时会在特定日期，例如Boxing-day，黑色星期五或是双十一（11月11日）开展大型促销活动或者发放优惠券以吸引消费者，然而很多被吸引来的买家都是一次性消费者，这些促销活动可能对销售业绩的增长并没有长远帮助，因此为解决这个问题，商家需要识别出哪类消费者可以转化为重复购买者。通过对这些潜在的忠诚客户进行定位，商家可以大大降低促销成本，提高投资回报率（Return on Investment, ROI）。众所周知的是，在线投放广告时精准定位客户是件比较难的事情，尤其是针对新消费者的定位。不过，利用天猫长期积累的用户行为日志，我们或许可以解决这个问题。

我们提供了一些商家信息，以及在“双十一”期间购买了对应产品的新消费者信息。你的任务是预测给定的商家中，哪些新消费者在未来会成为忠实客户，即需要预测这些新消费者在6个月内再次购买的概率。

数据描述

数据集包含了匿名用户在 "双十一 "前6个月和"双十一 "当天的购物记录，标签为是否是重复购买者。出于隐私保护，数据采样存在部分偏差，该数据集的统计结果会与天猫的实际情况有一定的偏差，但不影响解决方案的适用性。训练集和测试集数据见文件data_format2.zip，数据详情见下表。

字段名称	描述
user_id	购物者的唯一ID编码
age_range	用户年龄范围。<18岁为1；[18,24]为2； [25,29]为3； [30,34]为4；[35,39]为

Project

- Option 3: in-class competition
  - **题目：天体光谱智能识别分类**
  - **背景**
    - 光谱天文望远镜每个观测夜晚都能采集万余条光谱，使得传统的人工或半人工的利用模板匹配的方式不能很好应对，需要高效而准确的天体光谱智能识别分类算法。
  - **任务**
    - 利用14万个天体的光谱数据进行模型训练，把测试集中的未知天体分成行星（star），星系（galaxy）和类星体（qso）三类。

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

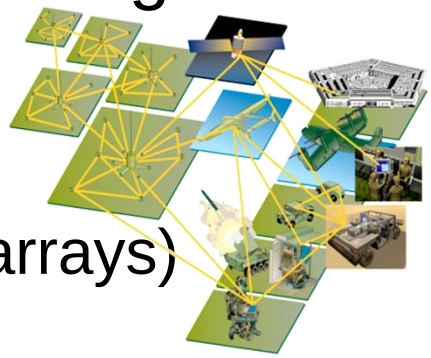
What Motivated Data Mining – Business World

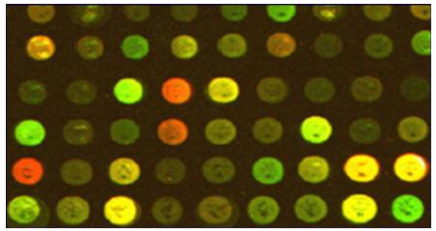
- Tremendous of data being collected and stored
  - E-commerce
  - Transactions
  - Stocks
  - Credit card transactions

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

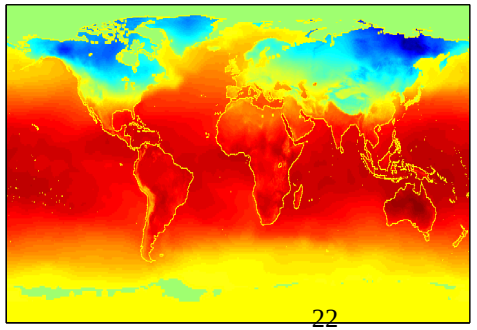
What Motivated Data Mining – Scientific World

- 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?

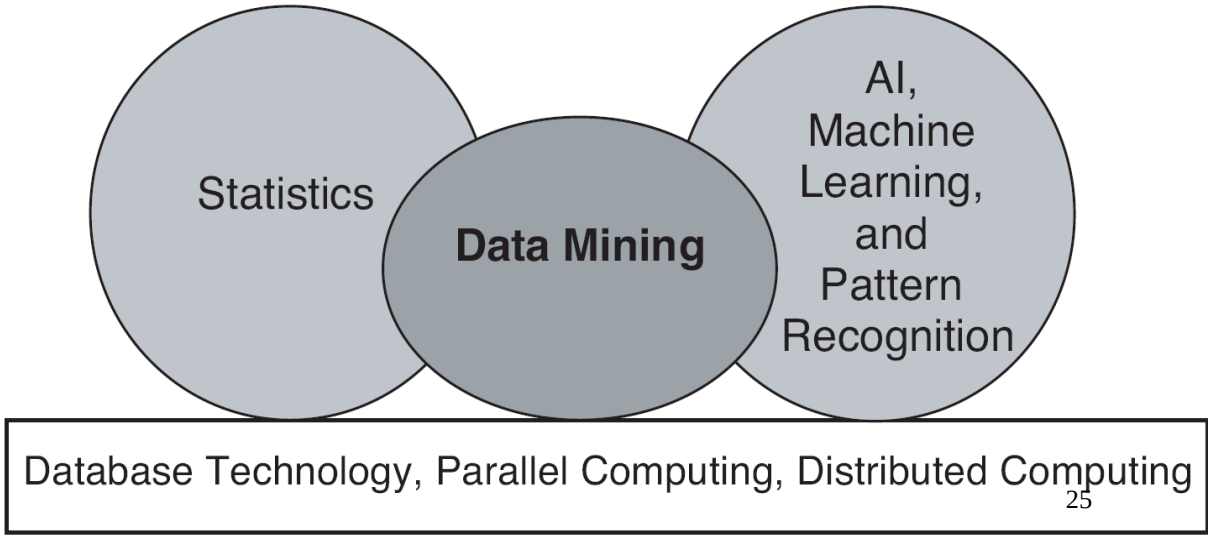
- We are drowning in data, but starving for knowledge!
  - Data rich, knowledge poor
  - Decision makers, domain experts have biases or errors
- Automated analysis of massive data sets

What is Data Mining?

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

What is Data Mining?

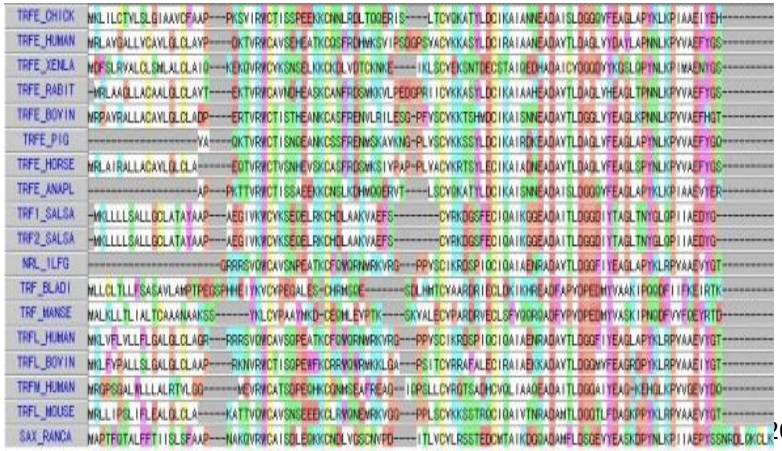
- Cross Disciplines
  - Databases
  - Machine learning: decision tree, Bayesian classifier, etc.
  - Statistics: regression, etc.
  - Neural networks
  - Parallel/Distributed computing



Why Not Traditional Data Analysis?

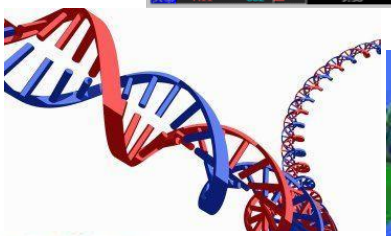
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data

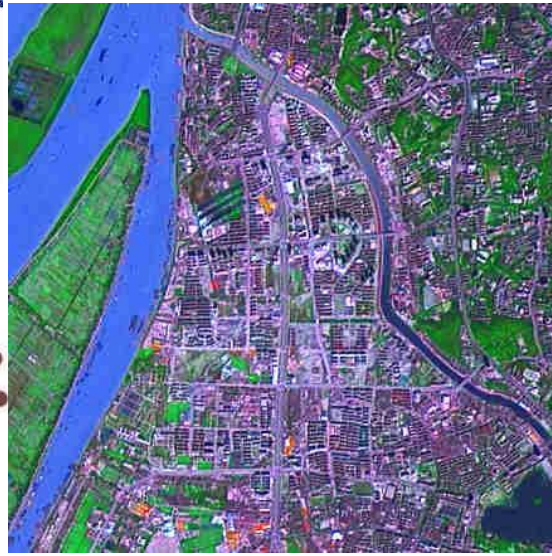
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions



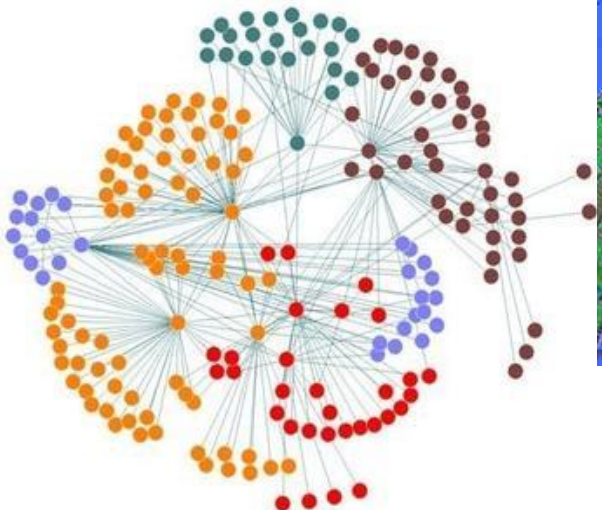
Why Not Traditional Data Analysis?

- 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

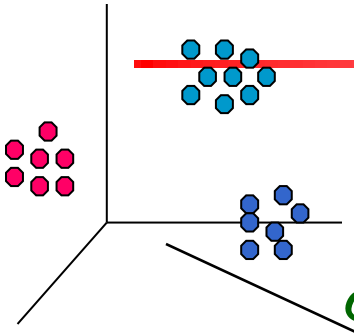
- Massive dataset
- Automatically searching for interesting patterns from historical data
- Fast
- Scalable
- Update easily
- Practical
- Decision support

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

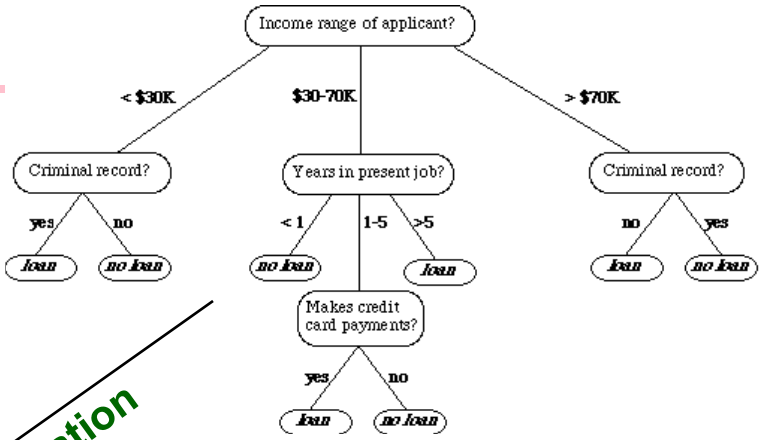


Clustering

Data

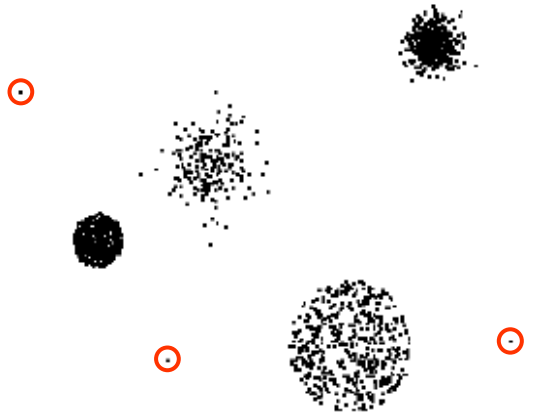
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Classification



Association Rules

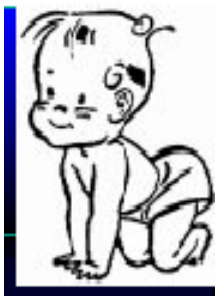
Anomaly Detection



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!



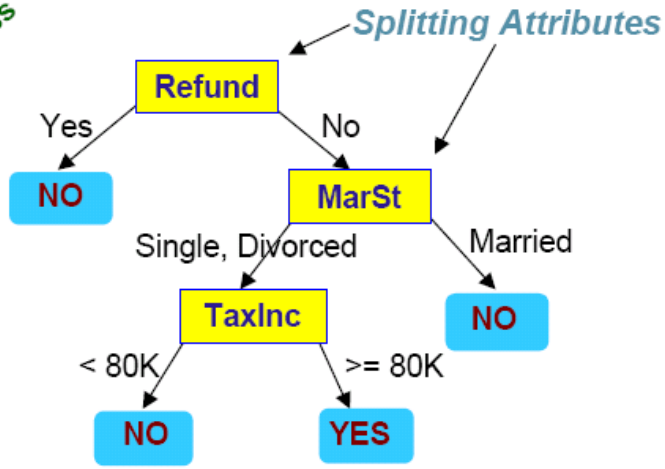
Ex. 1: Production Recommendation

- Where does the data come from?
  - supermarket transactions, membership cards, e-commerce orders, 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

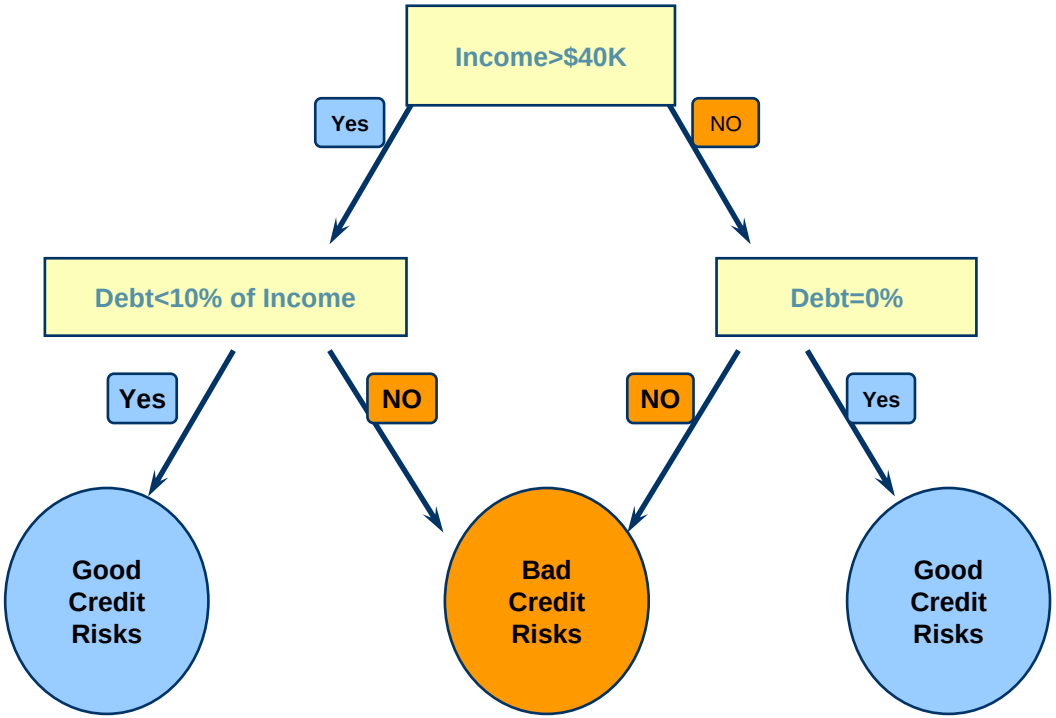


The splitting attribute at a node is determined based on the Gini index.

- Decision Tree

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

Ex.2 Credit Scoring



- Decision Tree

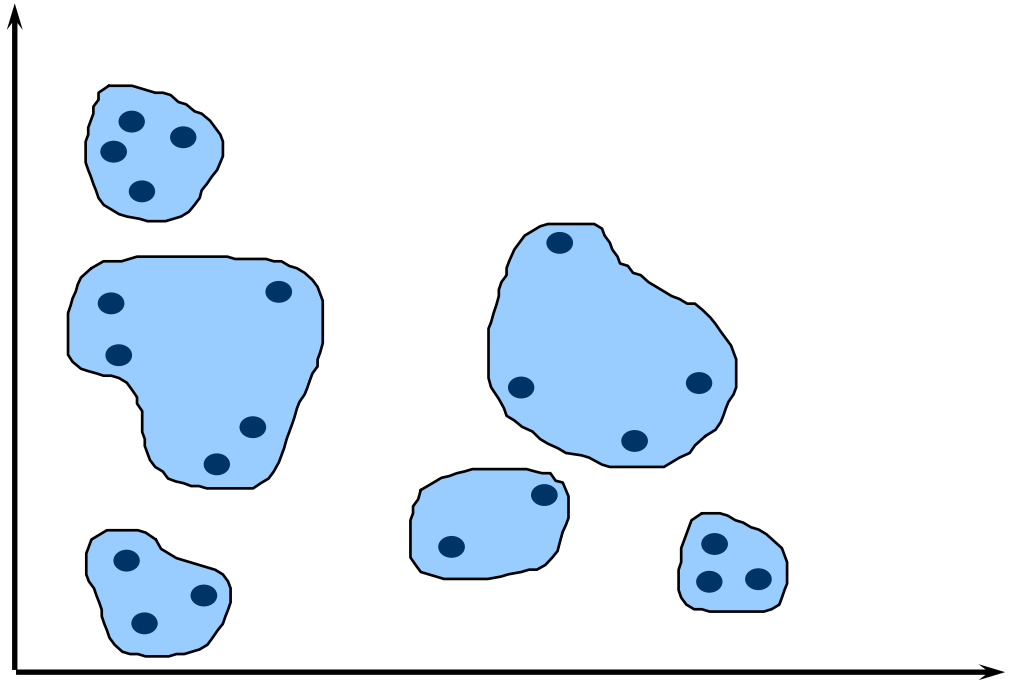
rule 1：if (Income<=$40k) and (Debt=0) then “good”
rule 2: if (Income>$40K) and (Debt<10% of Income) then “good”

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

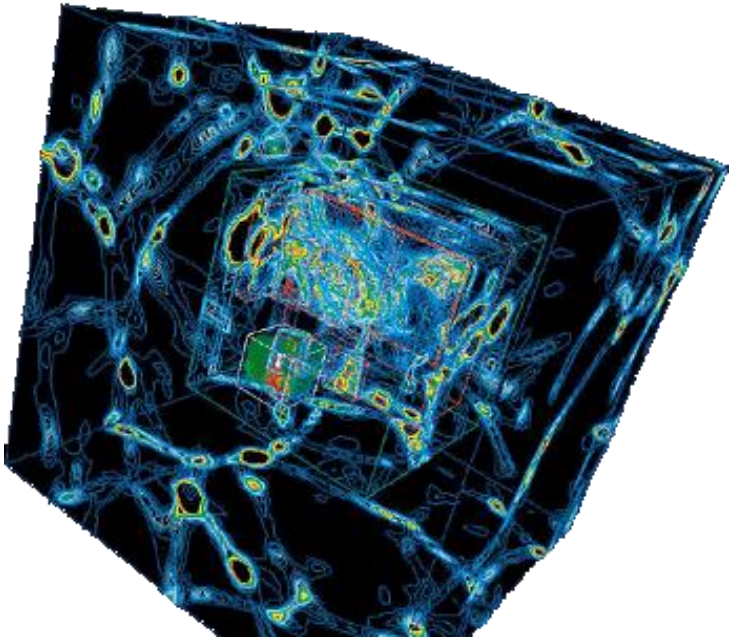
Clustering

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



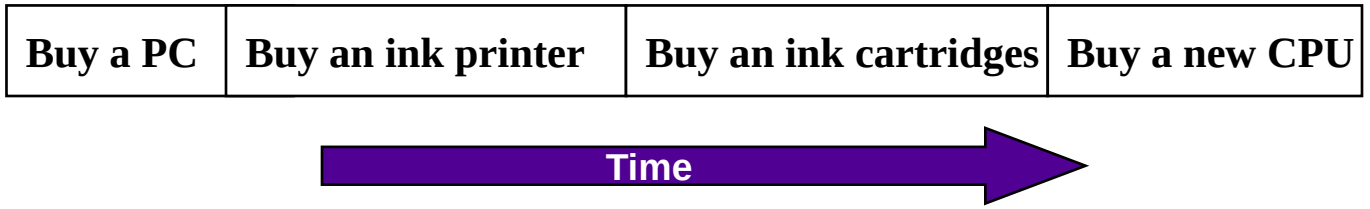
Ex.3 Scientific Simulation

- 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

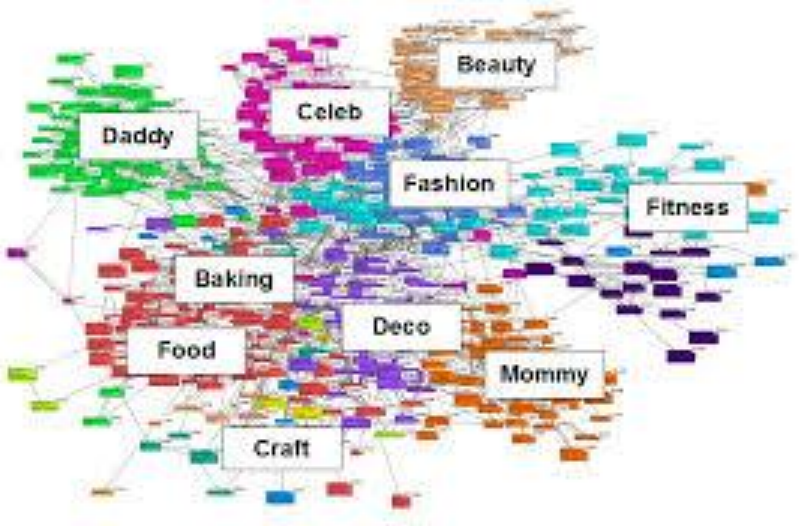
Anomaly Detection

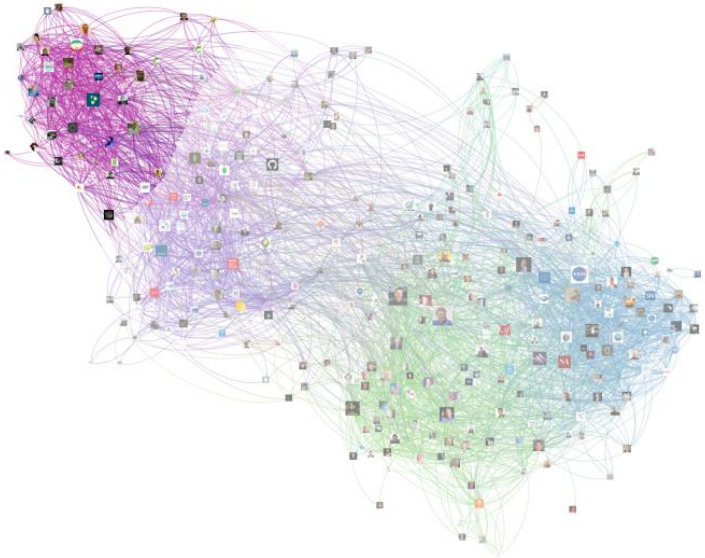
- What are anomalies?
  - The set of objects are considerably dissimilar from the remaining of the data
- Given a set of *n* objects, and *k,* the number of expected anomalies, find the top *k* objects that are considerably dissimilar or inconsistent with the remaining data

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.COM 京东

天猫 TMALL.COM

Movie

iQIYI 爱奇艺

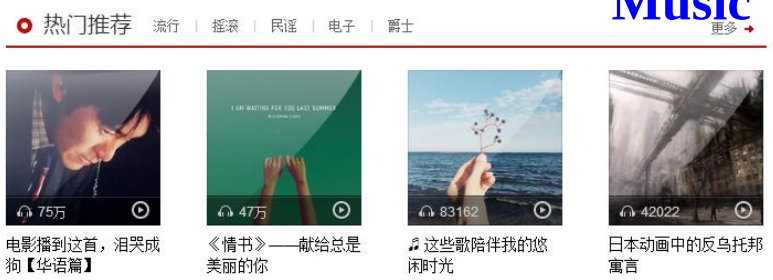
YOUKU 优酷

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

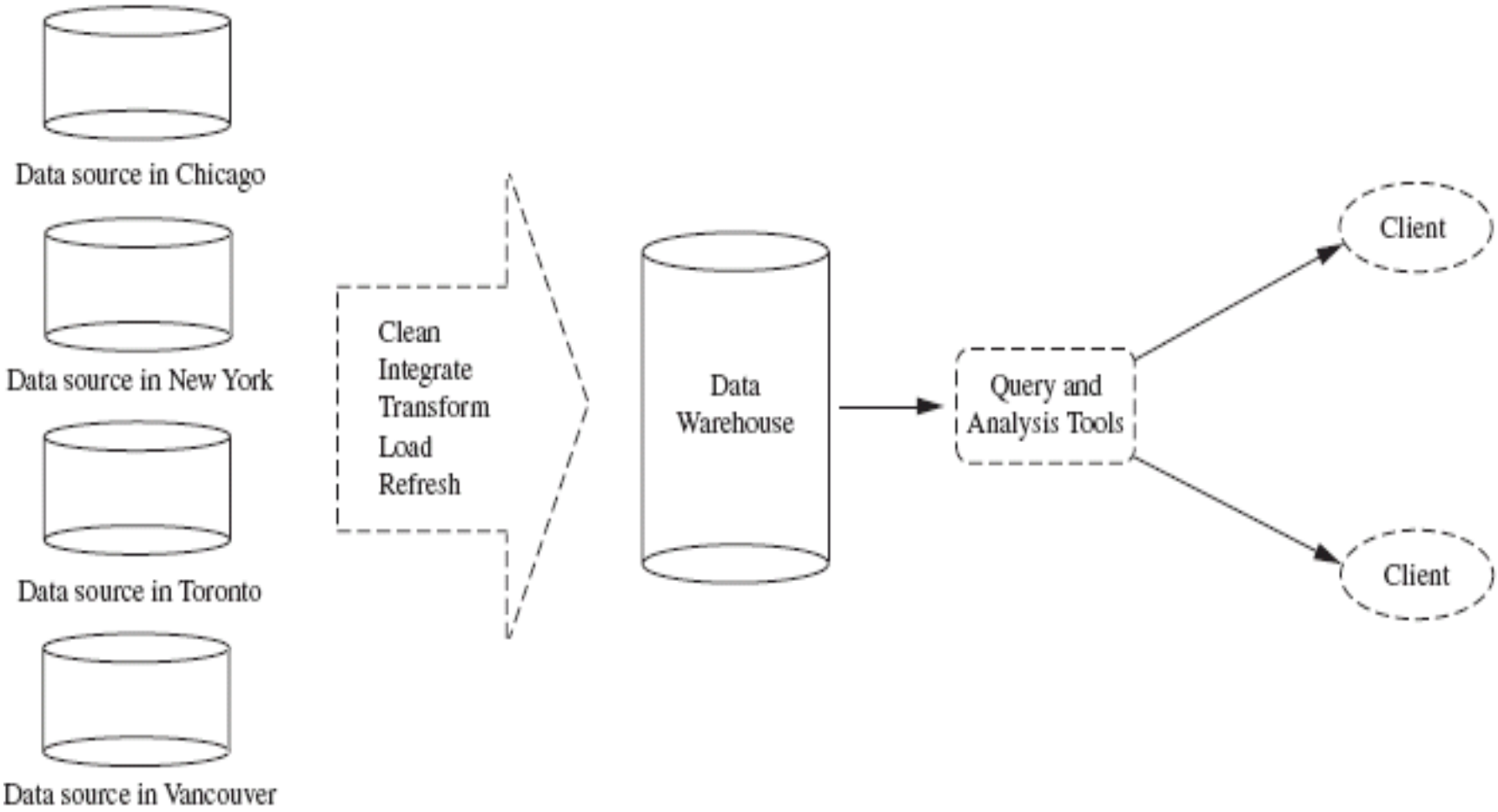
- Structured data
  - Table – records – attributes
  - Accessed by queries, SQL
- Online transactional processing (OLTP)
  - Insert a student “Ying Liu” into class “Introduction to Data Mining”, fall 2014

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
…	…

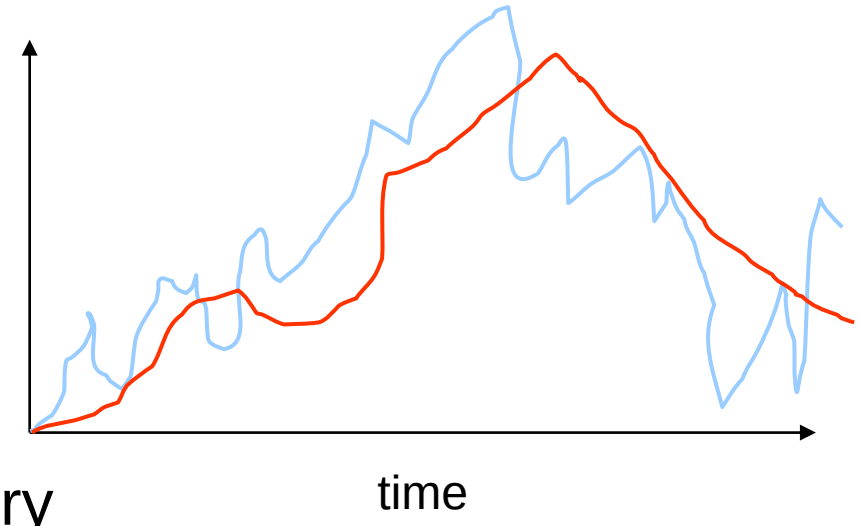
Spatial Data

- 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

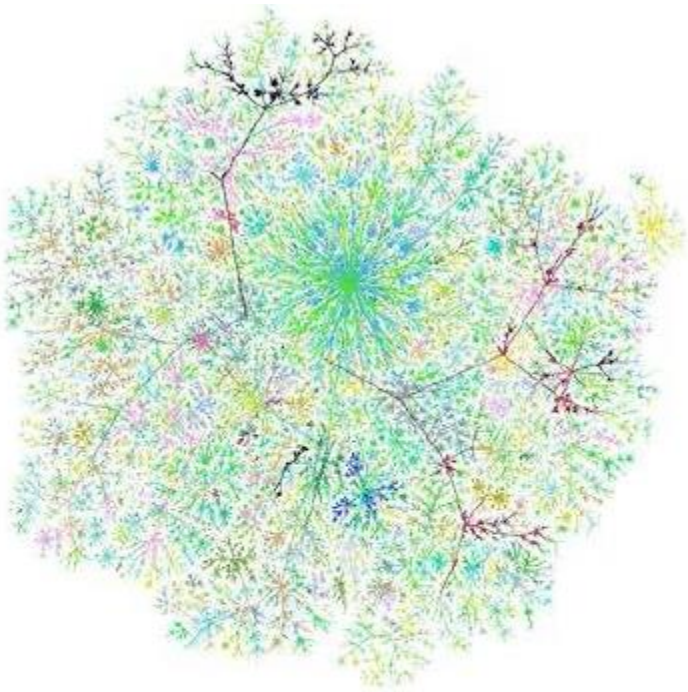
- The WWW is huge, widely distributed, global information service center
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for data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

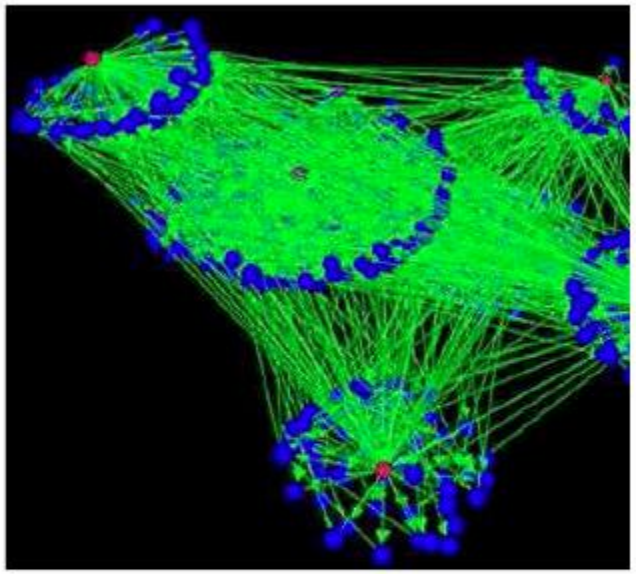
World-Wide Web

- 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

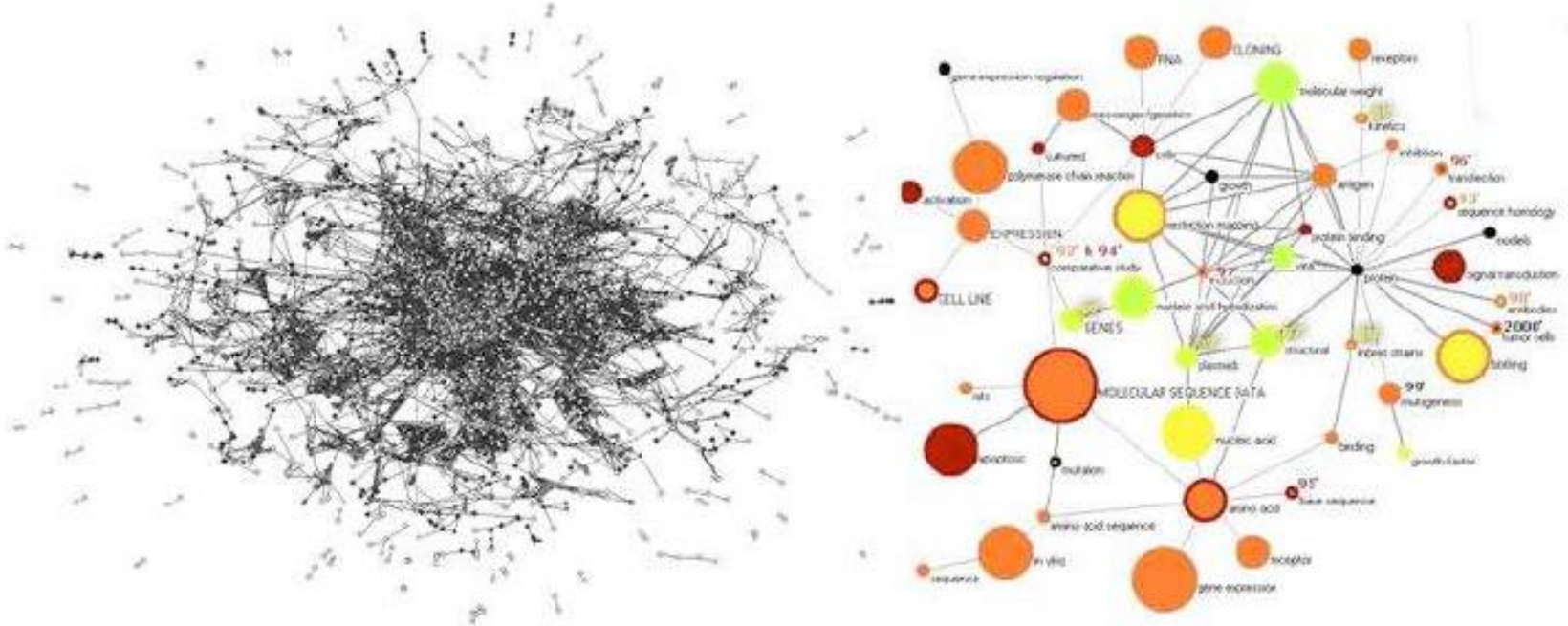




The images are downloaded from
http://www.maths.bris.ac.uk/~maarw/graphs/graph.html
and http://www.netdimes.org/new/?q=node/17

Graph

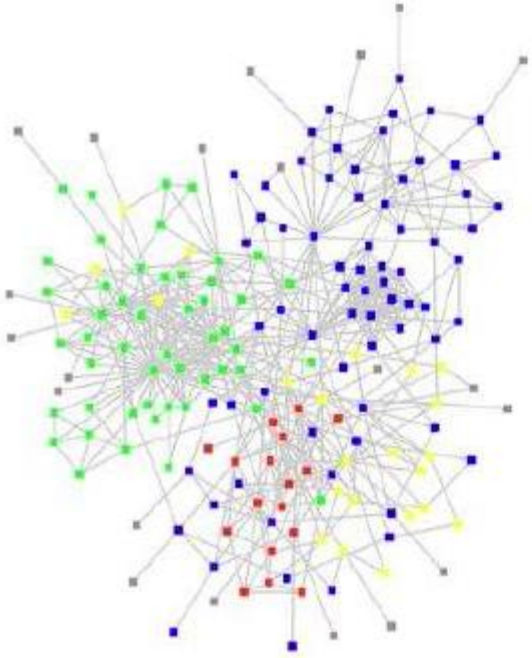
- Citation graph



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

Graph

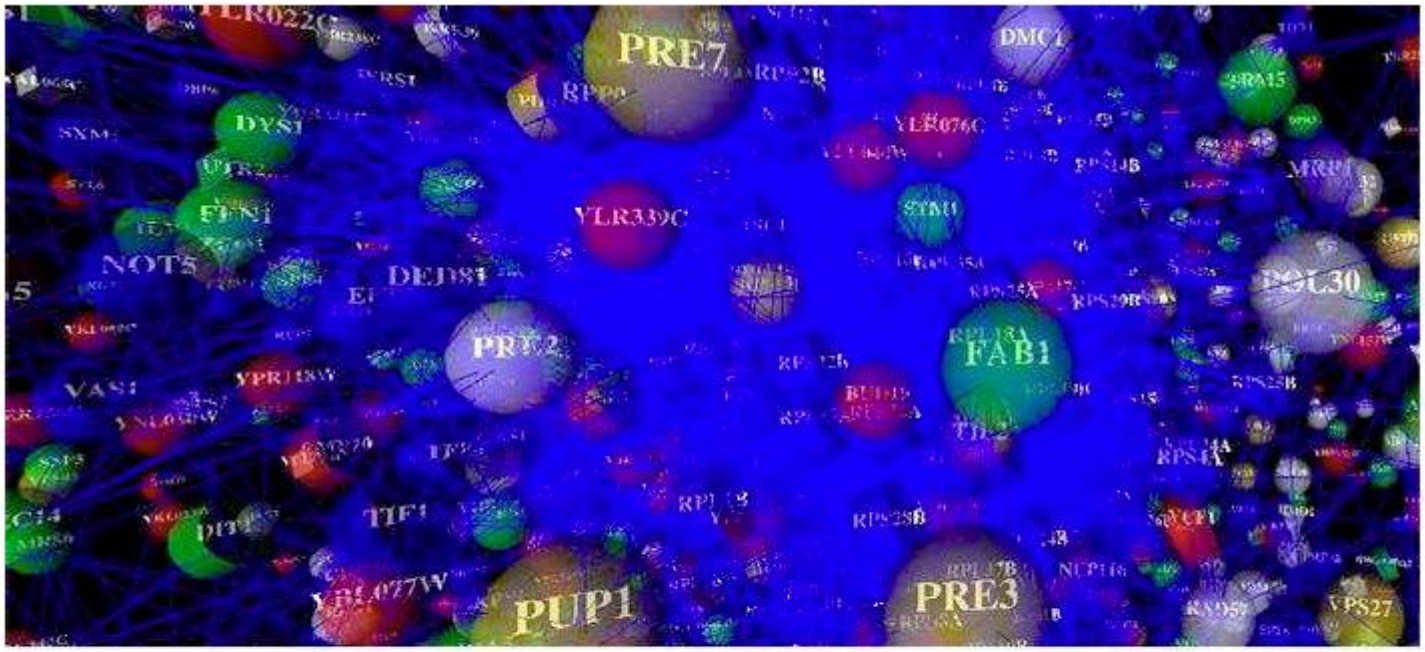
- Friendship graph



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/

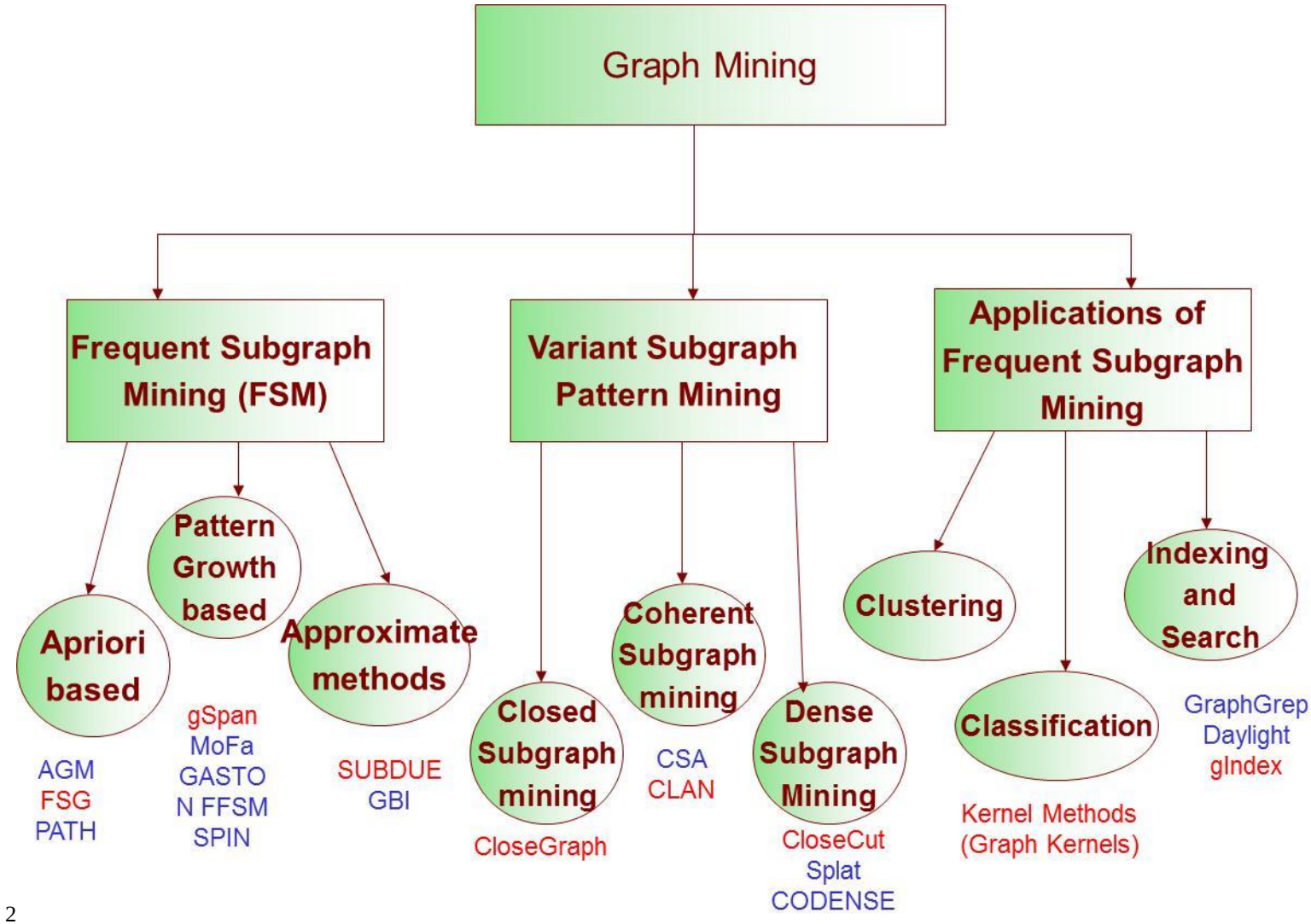
Graph

- Protein interaction graph



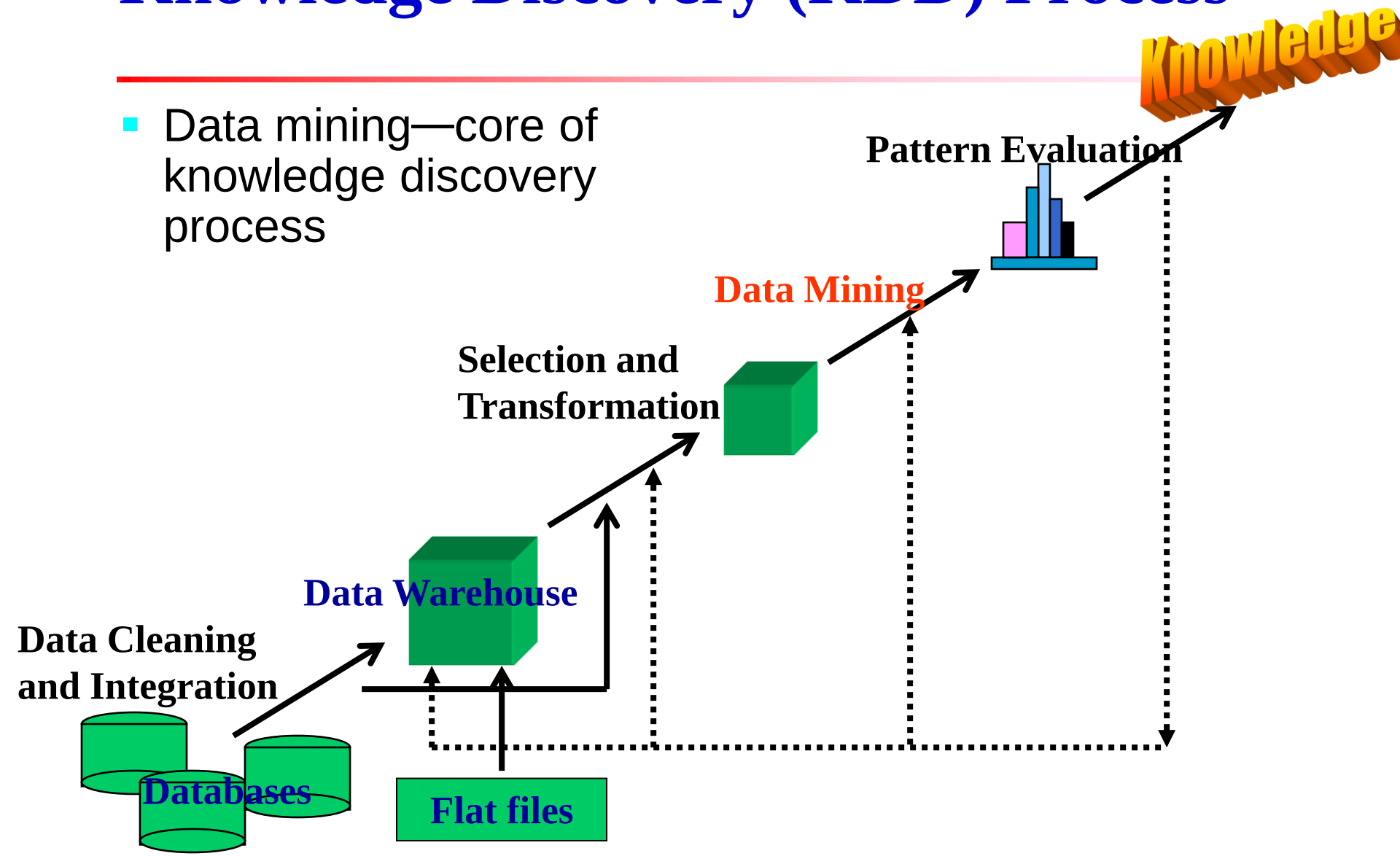
The images are downloaded from
http://bioinformatics.icmb.utexas.edu/lgl/Images/rsomZoom.jpg

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, fused data
  - 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