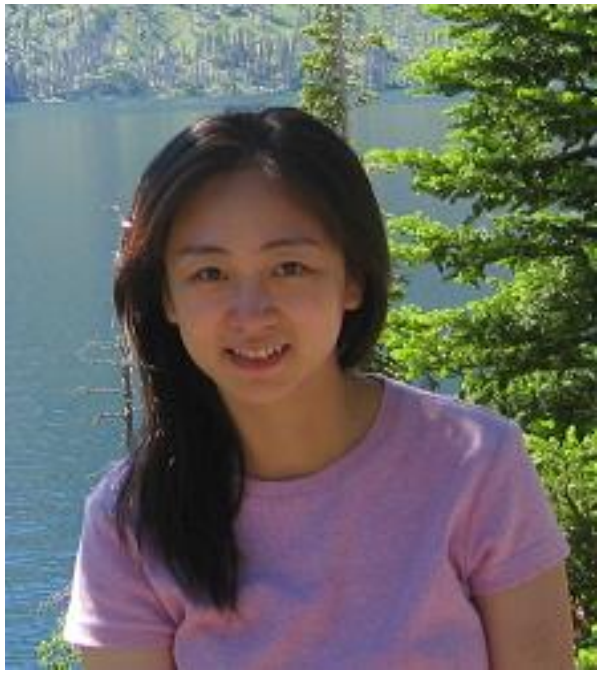
Data Mining

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

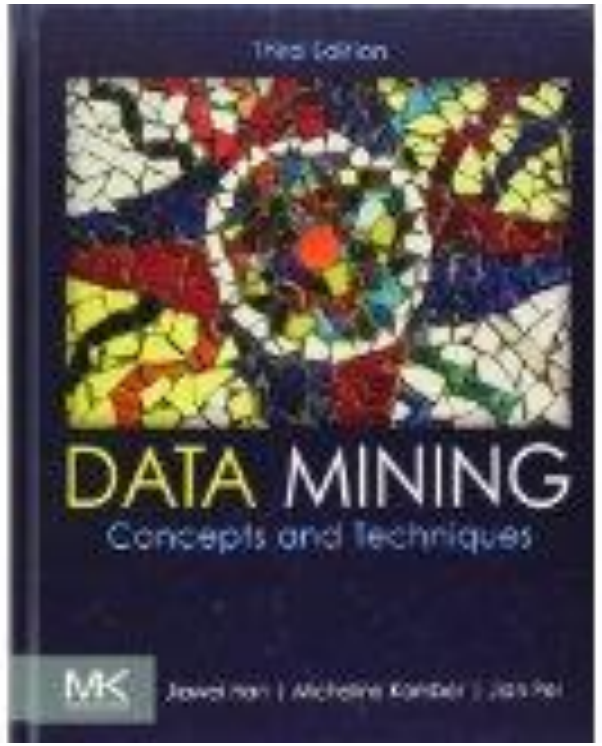


Useful Information

- Teaching Assistants
  - Jiang, Wen
  - Cui, Zhenyu
  - Wang, Wei
- 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%)
  - 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

- Option 1: 2020 CCF大数据与计算智能大赛
  - **题目——风电机组异常数据识别与清洗**

  (https://www.datafountain.cn/competitions/451)
  - 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

datafountain.cn/competitions/451

拖拽上传

赛制规则 数据与评测 排行榜 参赛队伍 交流讨论 常见问题 报名参赛

- 大赛介绍
- 主题方向
- 赛题名称
- 赛题背景
- 赛题任务
- 组织架构
- 赛程信息
- 参赛奖励
- 激励机制
- 大赛规则
- 参赛交流

赛题名称

风电机组异常数据识别与清洗

赛题背景

风能是一种环境友好且经济实用的可再生能源。中国是世界排名第一的风力发电国家、新装风力发电设备装机容量最大的国家，并且保持快速增长。由于风力发电正处于飞速发展阶段，风电场数量和规模不断扩大，然而受地理条件和环境因素限制，风电场多位于偏僻遥远的平原、山区或海上，因此为风电公司引入SCADA系统（数据采集与监视控制系统）对风电场群的日常运行进行集中监控、调度和管理，但风电机组受设备、环境、运行状态等因素影响，SCADA系统实时采集的风机运行数据会存在有大量异常值和缺失值，这些"脏数据"的存在严重影响后续的风电机组状态分析、故障诊断等功能。因此，识别并排除风电机组的异常数据具有重要的探究意义。

赛题任务

依据提供的8台风力电机1年的10min间隔SCADA运行数据，包括时间戳信息、风速信息和功率信息等，利用机器学习相关技术，建立鲁棒的风电机组异常数据检测模型，用于识别并剔除潜在的异常数据，提高数据质量。

此任务未给出异常数据标签，视为聚类任务，为引导选手向赛题需求对接，现简单阐述异常数据定义。异常数据是由风机运行过程与设计运行工况出现较大偏离时产生，如风速仪测风异常导致采集的功率散点明显偏离设计风功率。


datafountain.cn/competitions/451
赛制规则
数据与评测
排行榜
参赛队伍
交流讨论
常见问题
报名参赛
大赛介绍
主题方向
赛题名称
赛题背景
赛题任务
组织架构
赛程信息
参赛奖励
激励机制
大赛规则
参赛交流
赛程信息
2020/08/28 大赛启动仪式
2020/08/28-2020/10/16 A榜期间，参赛报名并网上提交资料,提交优化
2020/10/16 截止报名及组队
2020/10/18-2020/10/18 B榜期间，根据新的B榜分数优化，以最后一次提交成绩为准
2020/10/19-2020/11/13 评审、公示、上线平台
2020/11/24 总决赛颁奖、供需对接
参赛奖励
2个题目分别设奖：
奖项
数量
奖金
新能源数据治理类一等奖
1支团队
每支团队奖金30000元及证书
新能源数据治理类二等奖
2支团队
每支团队奖金20000元及证书

About the Project

- Option 2: in-class competition
  - **题目：天体光谱智能识别分类**
  - 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。
  - 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。
  - Read through some related research papers and fully understand them
  - Develop and implement the method
  - To be evaluated by the ranking in class

About the Project

- Option 3: in-class competition
  - 题目：**Foreign Object Detection on Roadways**
  - Given an input image, an algorithm is expected to produce a set of tight boxes around objects with classification labels.
  - We provide 766 images, taken on the road at the Zhongguancun Campus. The foreign objects in the images include yellow/black clippers, screws and small wrenches for screwing.
  - Read through some related research papers and fully understand them
  - Develop and implement the method
  - To be evaluated by the ranking in class


09-11-2018 星期二 18:00:06
1
2
3
4

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
- Classification
- Clustering
- Association rules
- Applications
- 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

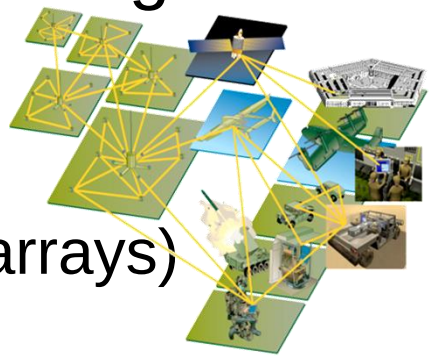
What Motivated Data Mining – Business World

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

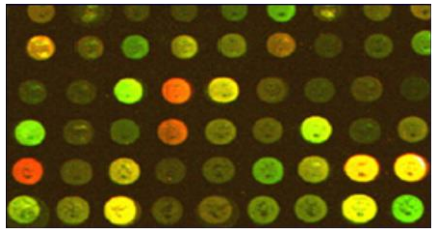
- Strong competitive pressure to extract and use the knowledge hidden in the data to provide customized CRM

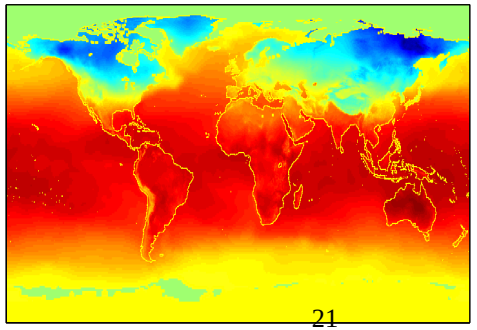
What Motivated Data Mining – Scientific World

- 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?

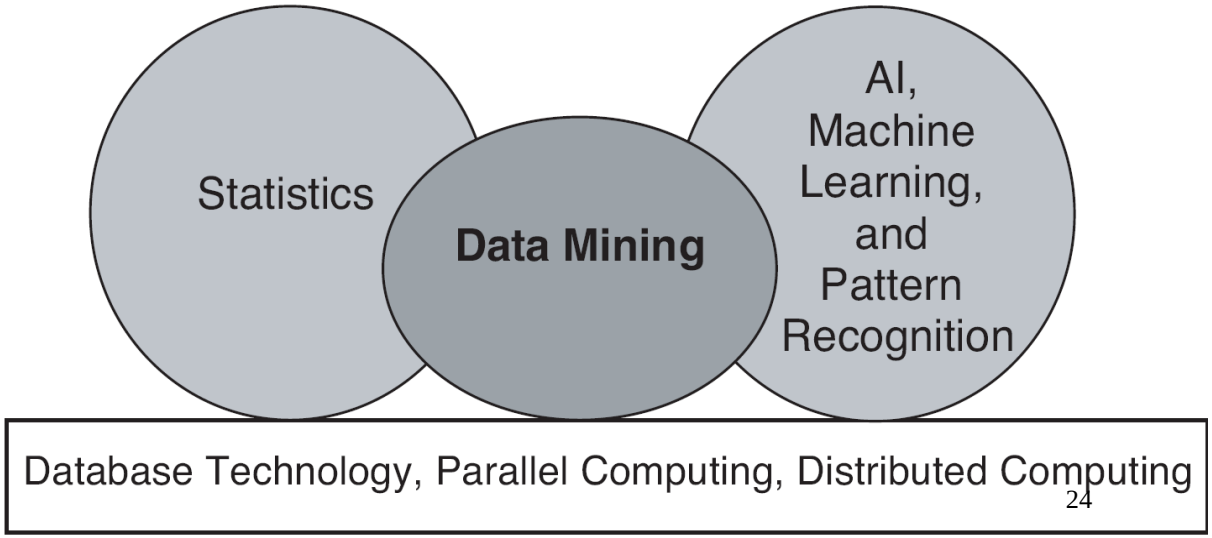
- We are drowning in data, but starving for knowledge!
  - Data rich, knowledge poor
  - Decision makers, domain experts have biases or errors
- Automated analysis of massive data sets

What is Data Mining?

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

What is Data Mining?

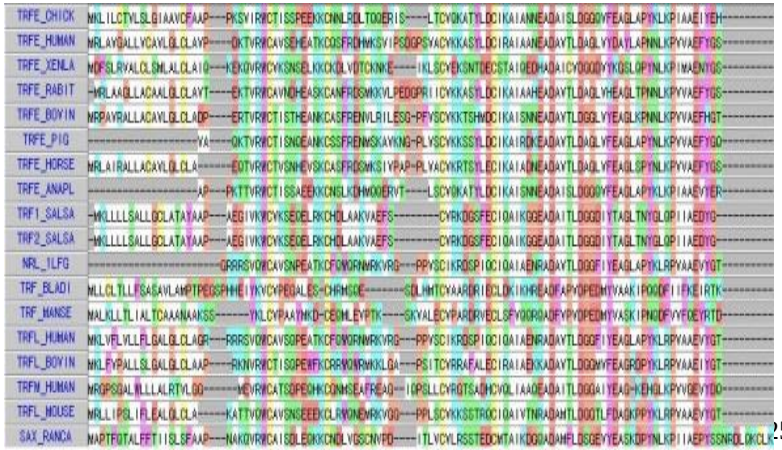
- Cross Disciplines
  - Databases
  - Machine learning: decision tree, Bayesian classifier, etc.
  - Statistics: regression, etc.
  - Neural networks
  - Parallel/Distributed computing



Why Not Traditional Data Analysis?

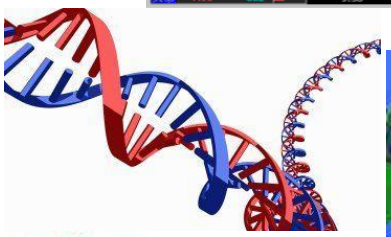
- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data

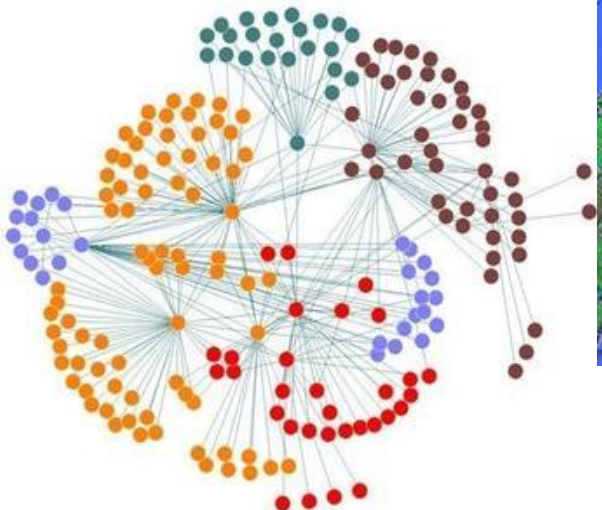
- High-dimensionality of data
  - DNA sequences may have tens of thousands of dimensions

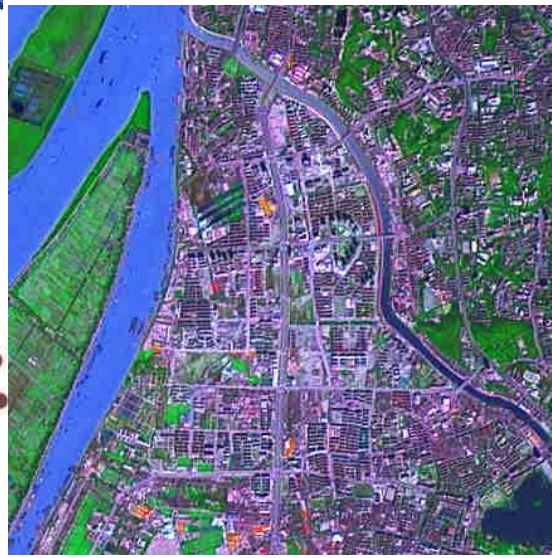


Why Not Traditional Data Analysis?

- 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

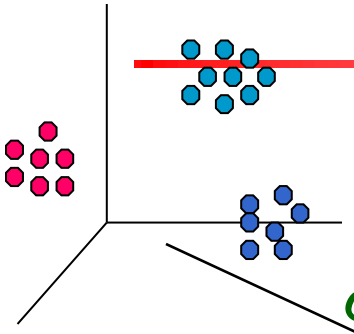
- Massive dataset
- Automatically searching for interesting patterns from historical data
- Fast
- Scalable
- Update easily
- Practical
- Decision support

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

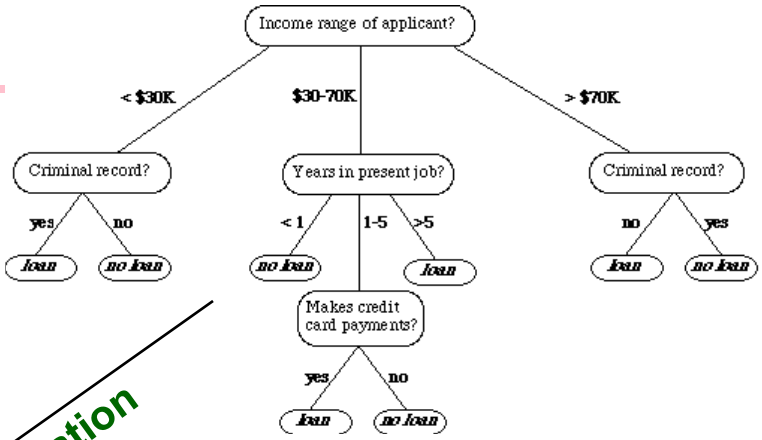


Clustering

Data

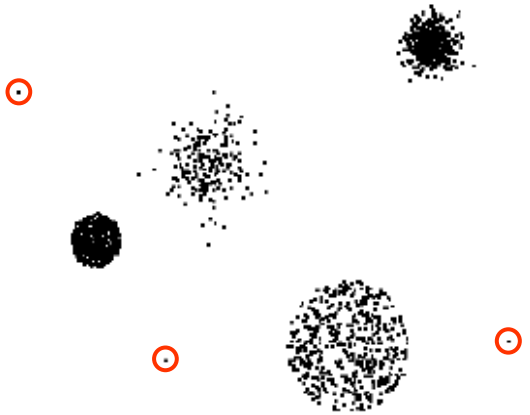
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Classification



Association Rules

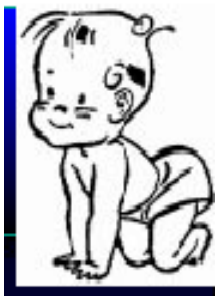
Anomaly Detection



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!



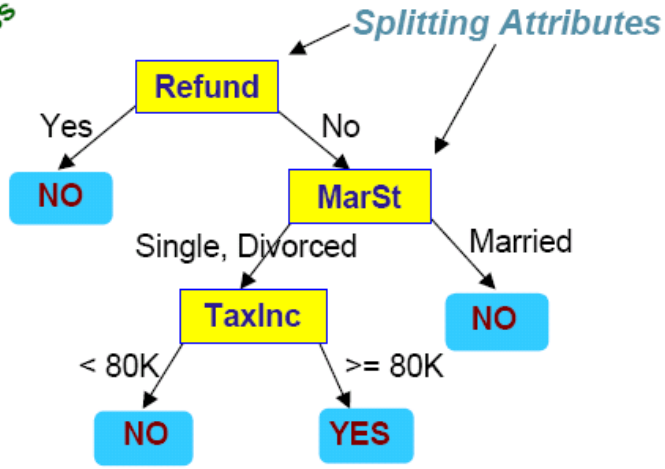
Ex. 1: Production Recommendation

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

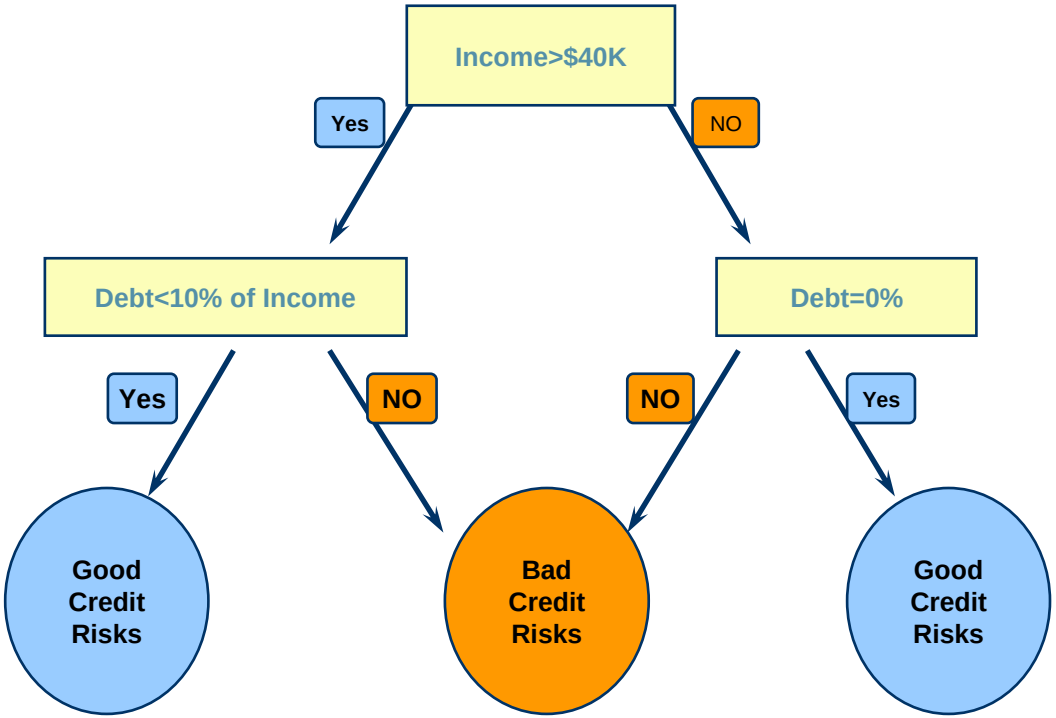


The splitting attribute at a node is determined based on the Gini index.

- Decision Tree

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

Ex.2 Credit Scoring



- Decision Tree

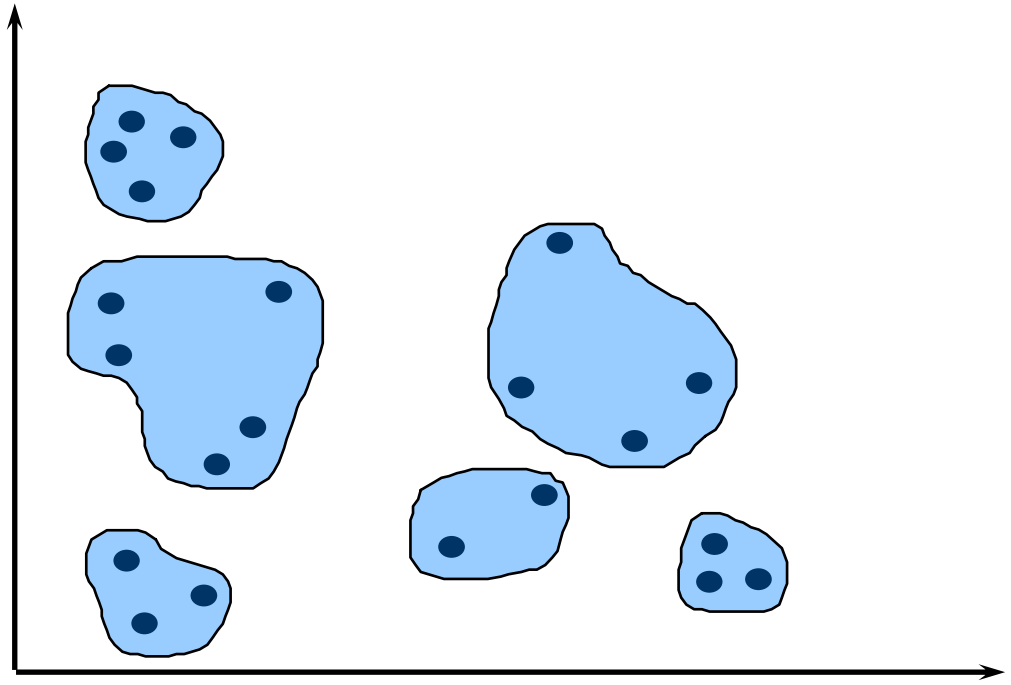
rule 1：if (Income<=$40k) and (Debt=0) then "good"
rule 2: if (Income>$40K) and (Debt<10% of Income) then "good"

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

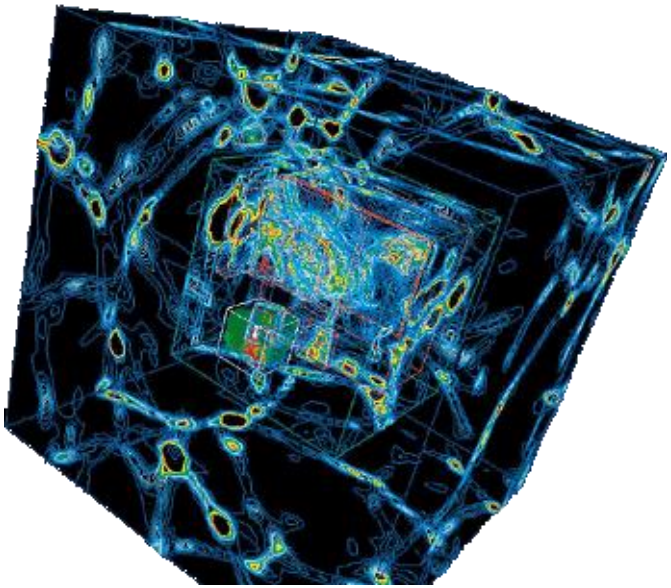
Clustering

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



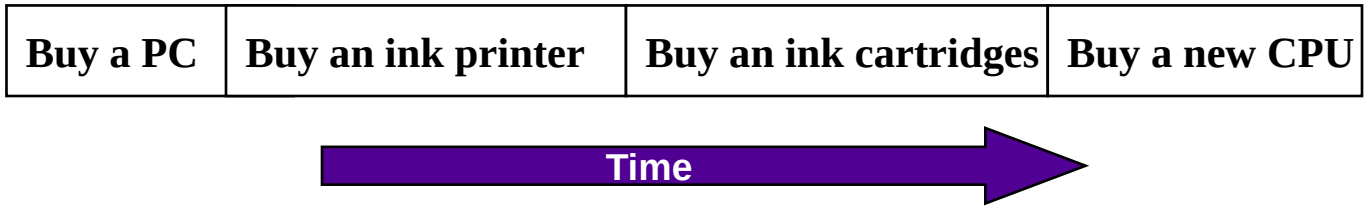
Ex.3 Scientific Simulation

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



Sequence Mining

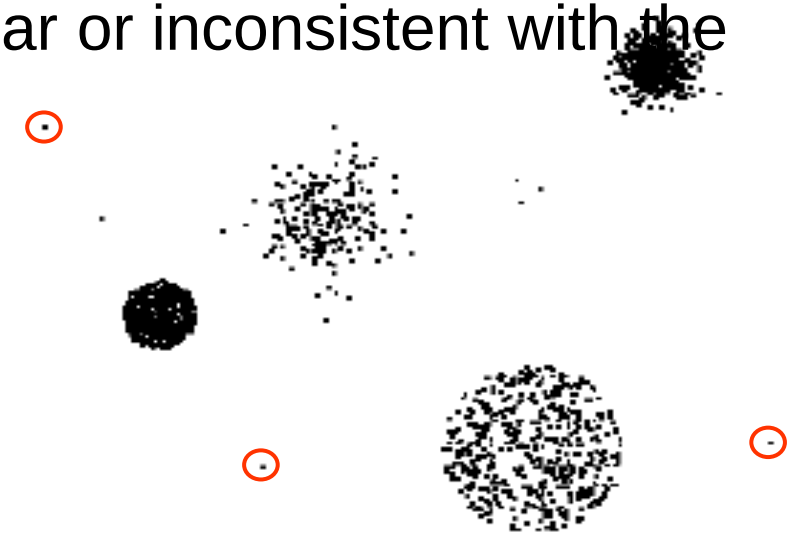
- Given a set of sequences, find the complete set of frequent subsequences



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

Anomaly Detection

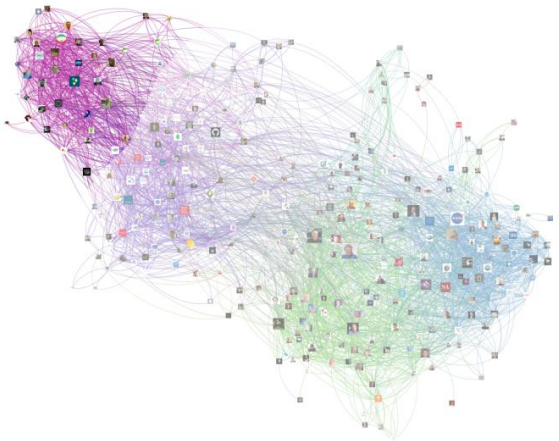
- What are anomalies?
  - The set of objects are considerably dissimilar from the remaining of the data
- Given a set of *n* objects, and *k,* the number of expected anomalies, find the top *k* objects that are considerably dissimilar or inconsistent with the remaining data



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

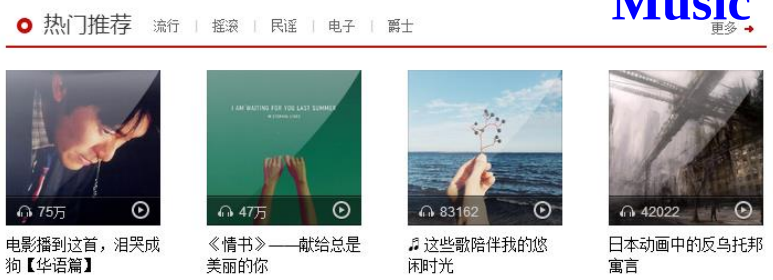
YOUKU 优酷

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

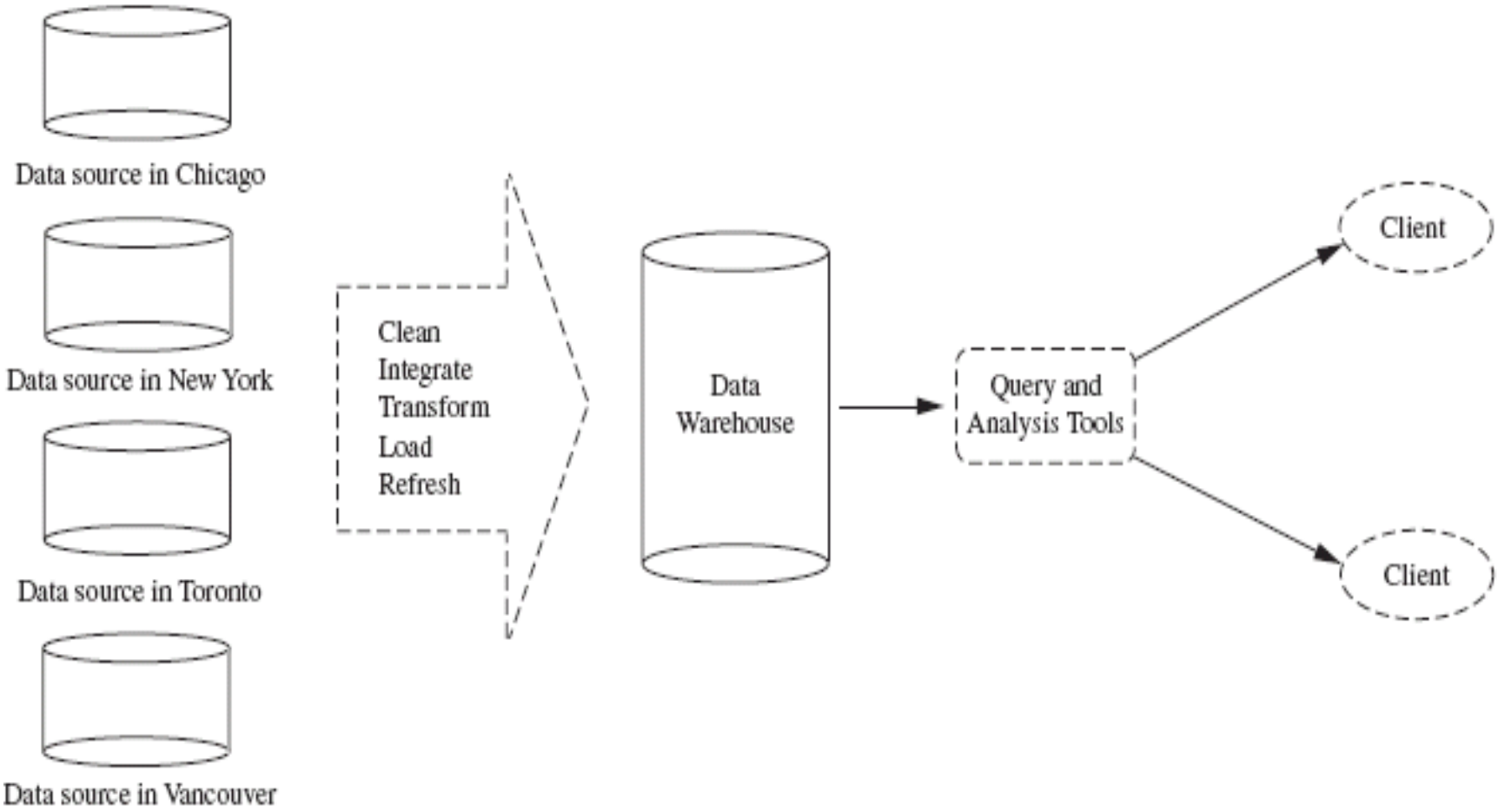
- Structured data
  - Table – records – attributes
  - Accessed by queries, SQL
- Online transactional processing (OLTP)
  - Insert a student “Ying Liu” into class “Introduction to Data Mining”, fall 2014

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
…	…

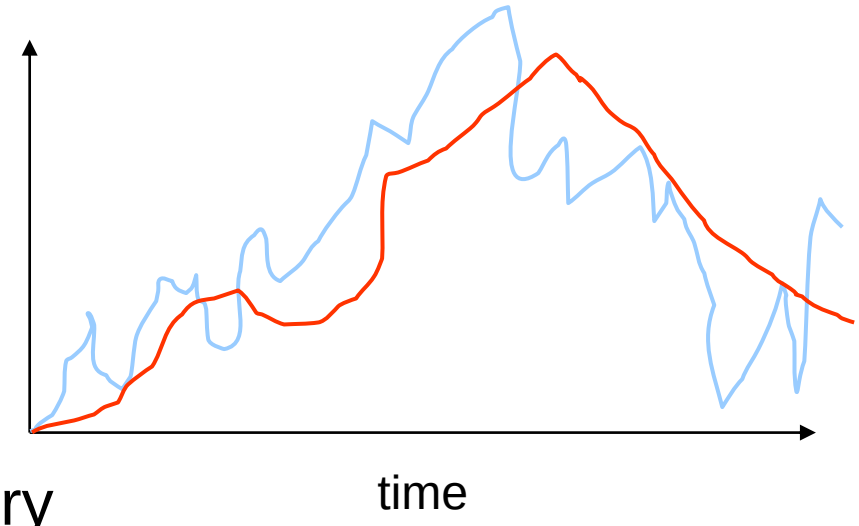
Spatial Data

- 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


Earth-Kind
Sustainable Landscaping Principles & Practices
Landscape Design:

- 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

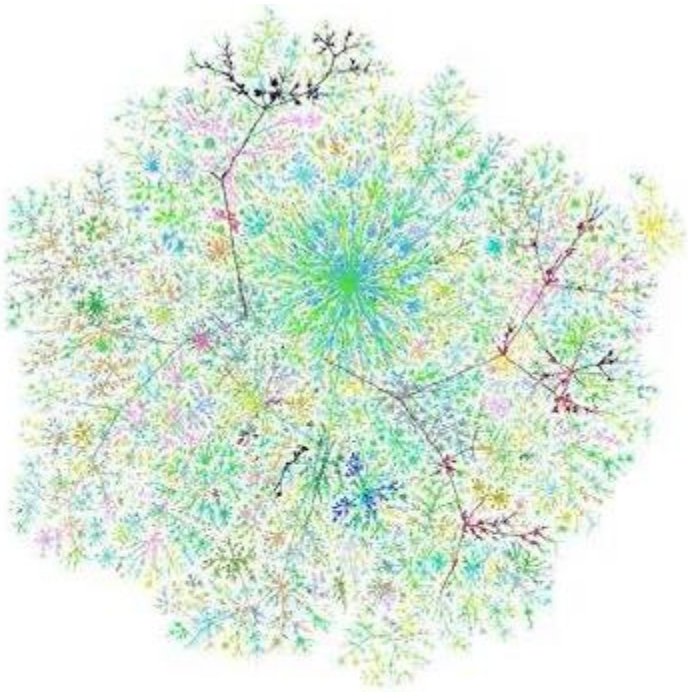
- The WWW is huge, widely distributed, global information service center for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for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

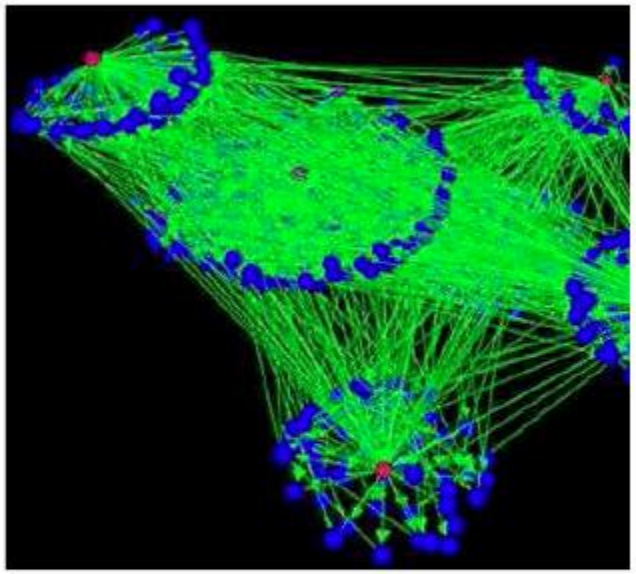
World-Wide Web

- 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

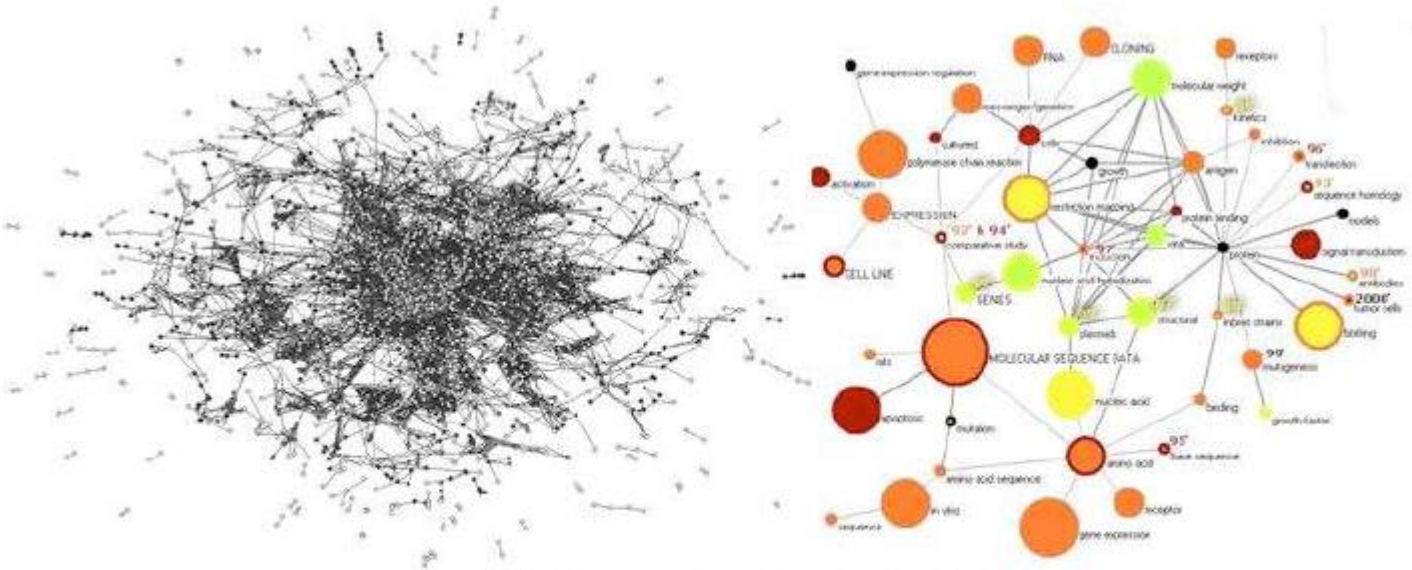




The images are downloaded from
http://www.maths.bris.ac.uk/~maarw/graphs/graph.html
and http://www.netdimes.org/new/?q=node/17

Graph

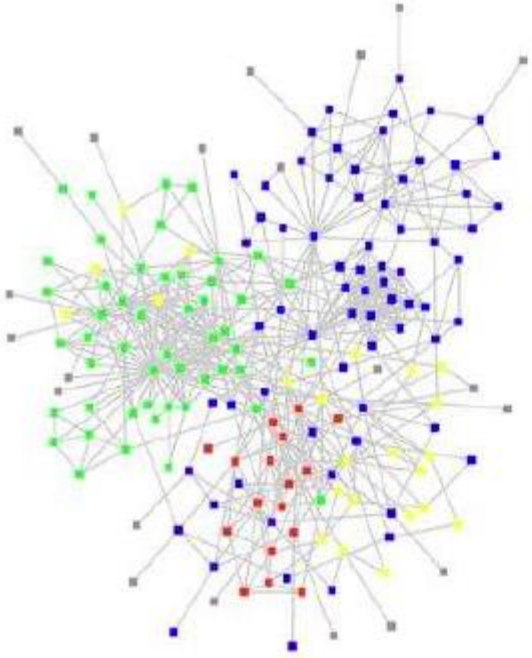
- Citation graph



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

Graph

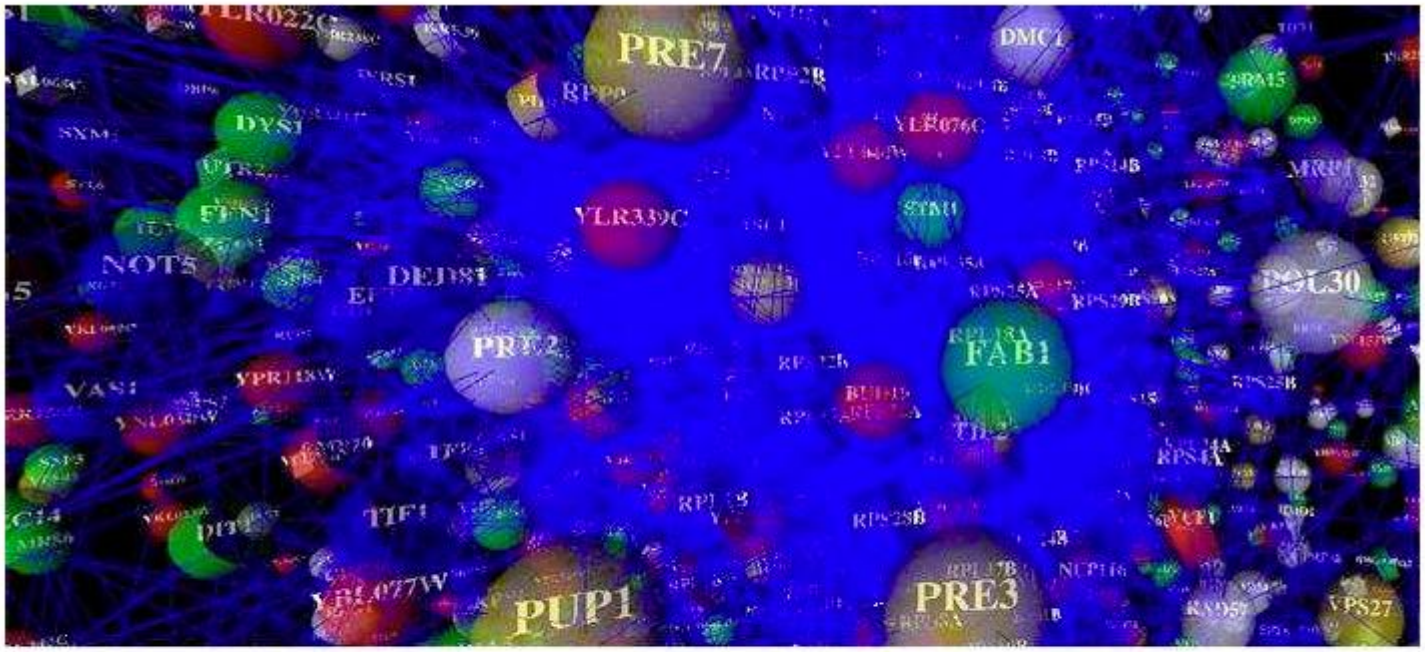
- Friendship graph



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/

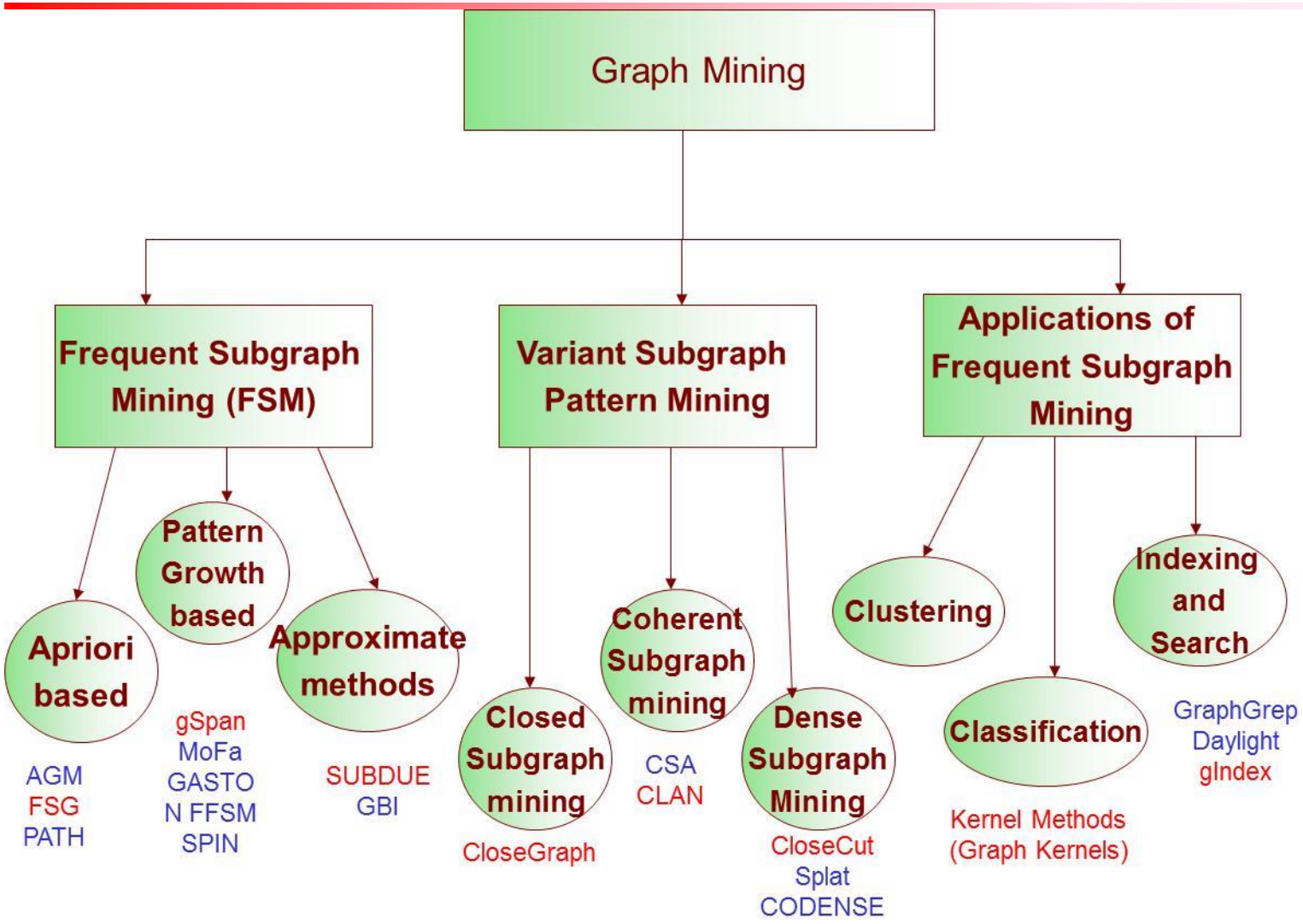
Graph

- Protein interaction graph



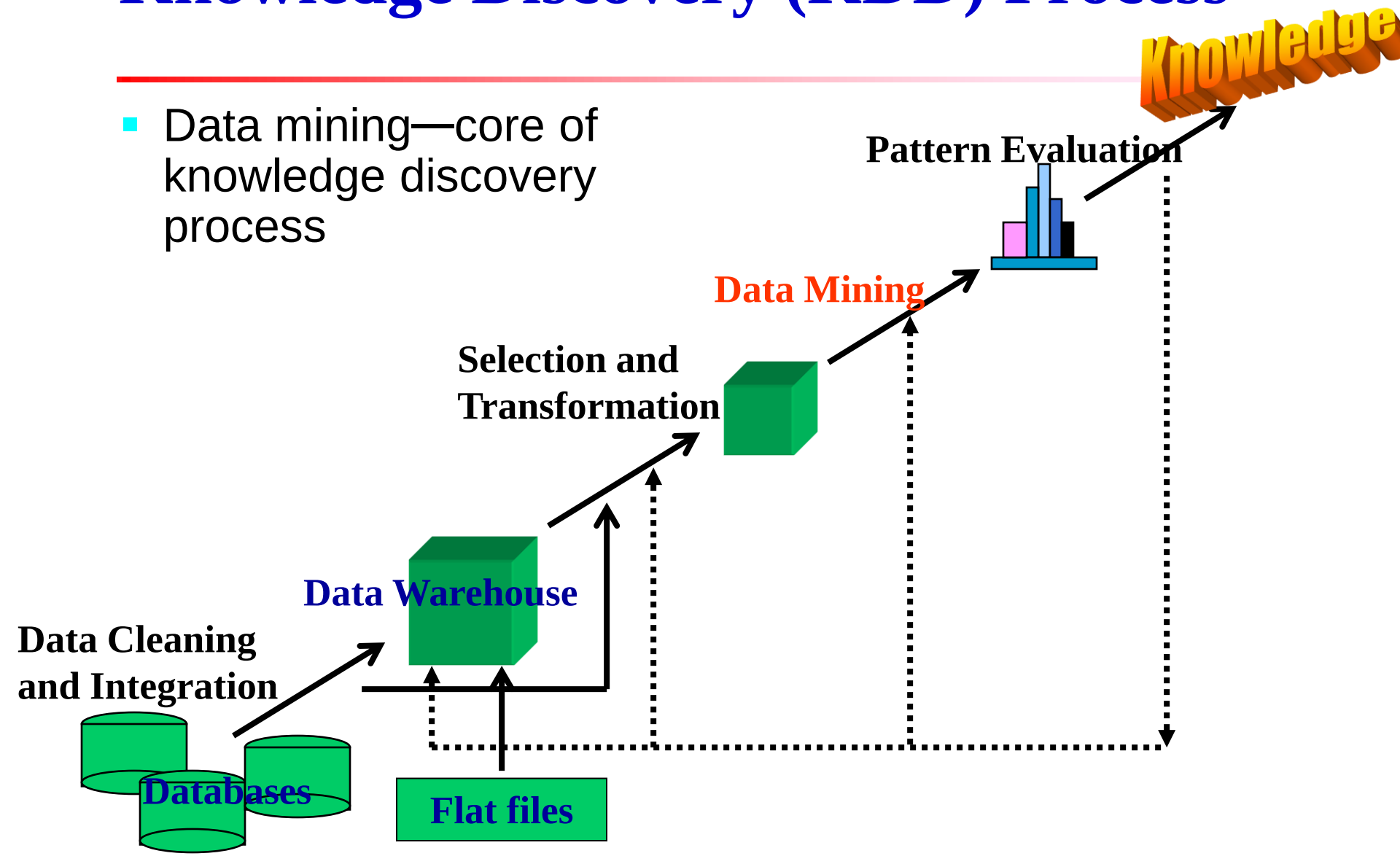
The images are downloaded from
http://bioinformatics.icmb.utexas.edu/lgl/Images/rsomZoom.jpg

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 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