



Data Mining: Concepts and Techniques

(3rd ed.)

— Chapter 1 —

Farid Feyzi

Grading Policy

- Mid-Exam: 25%
- Final Exam: 40%
- Research Work (with Presentation): 15(up to 25)%
- Project: 20%

Chapter 1. Introduction

- Why Data Mining? 
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Why Data Mining?



- The Explosive Growth of Data: from terabytes to petabytes
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Why do we need data mining?

- Really, really huge amounts of raw data!!
 - In the digital age, TB of data is generated by the second
 - Mobile devices, digital photographs, web documents.
 - Facebook updates, Tweets, Blogs, User-generated content
 - Transactions, sensor data, surveillance data
 - Queries, clicks, browsing
 - Cheap storage has made possible to maintain this data
- Need to analyze the raw data to extract knowledge

Why do we need data mining?

- “The data is the computer”
 - Large amounts of data can be more powerful than complex algorithms and models
 - Google has solved many Natural Language Processing problems, simply by looking at the data
 - Example: misspellings, synonyms
 - Data is power!
 - Today, the collected data is one of the biggest assets of an online company
 - Query logs of Google
 - The friendship and updates of Facebook
 - Tweets and follows of Twitter
 - Amazon transactions

Data Mining as the Evolution of Information Technology

- 1960s:

- Data

- 1970s:

- Rel

- 1980s:

- RD

- Appl

- 1990s:

- Data

dat

- 2000s

- Stream c

- Data mir

- Web tec

Advanced Data Analysis

(late-1980s to present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
- Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
- Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
- Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.

ductive, etc.)

etc.)

Web

mation systems

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining? 
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

What Is Data Mining?



- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems



Knowledge Discovery (KDD) Process

The knowledge discovery process is an iterative sequence of the following steps:

- 1. Data cleaning (to remove noise and inconsistent data)
- 2. Data integration (where multiple data sources may be combined)
- 3. Data selection (where data relevant to the analysis task are retrieved from the database)
- 4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

Knowledge Discovery (KDD) Process



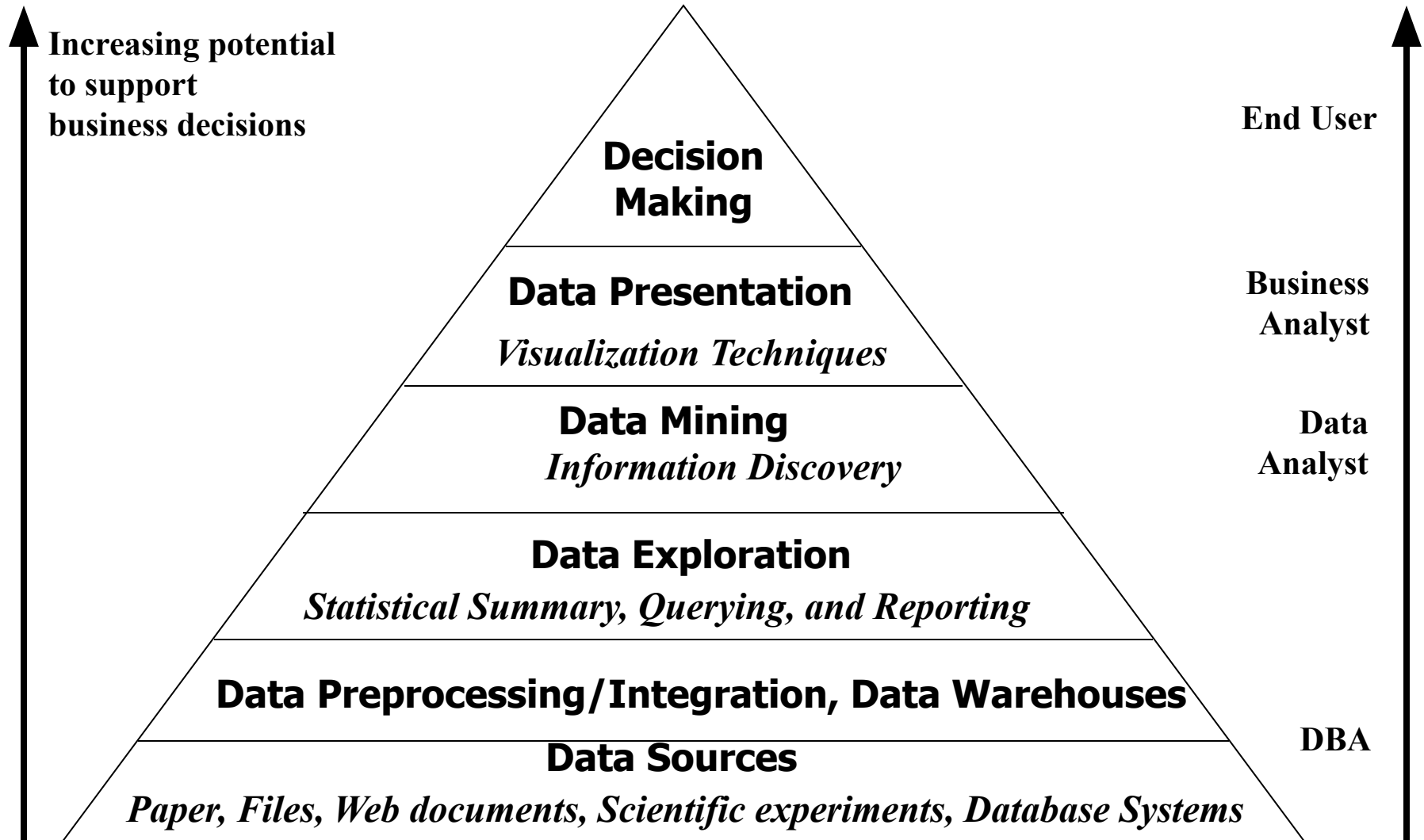
The knowledge discovery process is an iterative sequence of the following steps:

- 5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
- 6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- 7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

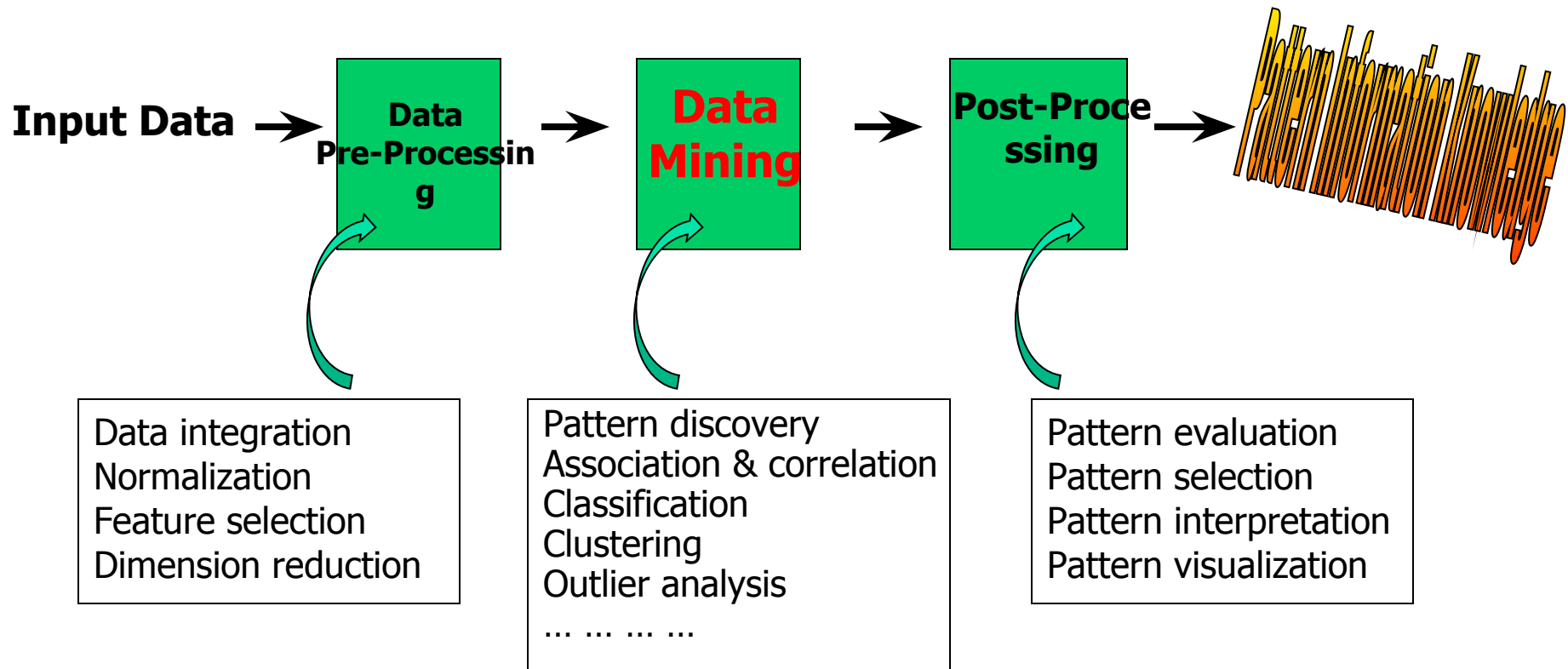
Example: A Web Mining Framework

- Web mining usually involves
 - Data cleaning
 - Data integration from multiple sources
 - Warehousing the data
 - Data cube construction
 - Data selection for data mining
 - Data mining
 - Presentation of the mining results
 - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



KDD Process: A Typical View from ML and Statistics




- This is a view from typical machine learning and statistics communities

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining 
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Multi-Dimensional View of Data Mining

■ **Data to be mined**

- Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

■ **Knowledge to be mined (or: Data mining functions)**

- Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
- Descriptive vs. predictive data mining
- Multiple/integrated functions and mining at multiple levels


■ **Techniques utilized**

- Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

■ **Applications adapted**

- Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined? 
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
 - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data (incl. bio-sequences)
 - Structure data, graphs, social networks and multi-linked data
 - Object-relational databases
 - Heterogeneous databases and legacy databases
 - Spatial data and spatiotemporal data
 - Multimedia database
 - Text databases
 - The World-Wide Web

The data is also very complex

- Multiple types of data: tables, time series, images, graphs, etc
- Spatial and temporal aspects
- Interconnected data of different types:
 - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images through cameras, queries to search engines

Example: transaction data

- Billions of real-life customers:
 - WALMART: 20M transactions per day
 - AT&T 300 M calls per day
 - Credit card companies: billions of transactions per day.
- The point cards allow companies to collect information about specific users

Example: document data

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: 4 million articles (and counting)
- Online news portals: steady stream of 100's of new articles every day
- Twitter: ~300 million tweets every day

Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 500 million users
- Twitter: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

Example: genomic sequences

- <http://www.1000genomes.org/page.php>
- Full sequence of 1000 individuals
- 3×10^9 nucleotides per person \square 3×10^{12} nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Example: environmental data

- Climate data (just an example)

<http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php>

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”
- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”
- **Spatiotemporal** data

Behavioral data

- Mobile phones today record a large amount of information about the user behavior
 - GPS records position
 - Camera produces images
 - Communication via phone and SMS
 - Text via facebook updates
 - Association with entities via check-ins
- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.
- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.
- Data collected for millions of users on a daily basis

So, what is Data?

- Collection of data **objects** and their **attributes**
- An attribute is a property or characteristic of an object
 - Examples: eye color of a person, temperature, etc.
 - Attribute is also known as **variable**, **field**, **characteristic**, or **feature**
- A collection of attributes describe an object
 - Object is also known as **record**, **point**, **case**, **sample**, **entity**, or **instance**

Objects

Attributes

Attributes				
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

Size: Number of objects

Dimensionality: Number of attributes

Sparsity: Number of populated object-attribute pairs

Types of Attributes

- There are different types of attributes
 - **Categorical**
 - Examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
 - **Nominal** (no order or comparison) vs **Ordinal** (order but not comparable)
 - **Numeric**
 - Examples: dates, temperature, time, length, value, count.
 - **Discrete** (counts) vs **Continuous** (temperature)
 - Special case: **Binary** attributes (yes/no, exists/not exists)

Numeric Record Data

- If data objects have the same **fixed set** of **numeric attributes**, then the data objects can be thought of as **points** in a multi-dimensional space, where each **dimension** represents a distinct attribute
- Such data set can be represented by an **n-by-d data matrix**, where there are **n** rows, one for each object, and **d** columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Categorical Data

- Data that consists of a collection of records, each of which consists of a **fixed set** of **categorical** attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

Document Data

- Each document becomes a 'term' vector,
 - each term is a component (attribute) of the vector,
 - the value of each component is the number of times the corresponding term occurs in the document.
 - **Bag-of-words** representation – no ordering

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

- Each record (transaction) is a **set of items**.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

- A set of items can also be represented as a **binary vector**, where each attribute is an item.
- A document can also be represented as a **set of words** (no counts)

Sparsity: average number of products bought by a customer

Ordered Data

- Genomic **sequence** data

```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

- Data is a long **ordered** string

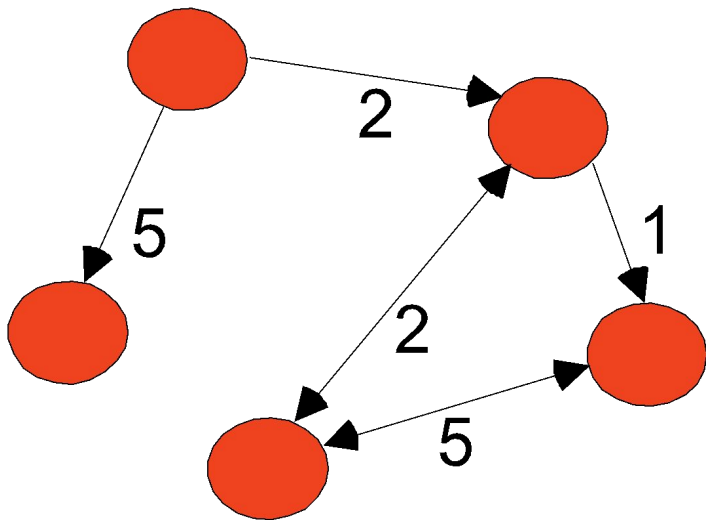
Ordered Data

- Time series
 - Sequence of ordered (over “time”) numeric values.



Graph Data

- Examples: Web graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>
```

```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>
```


```
<li>
```

```
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>
```

```
<li>
```

```
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined? 
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Data Mining Function: (1) Generalization

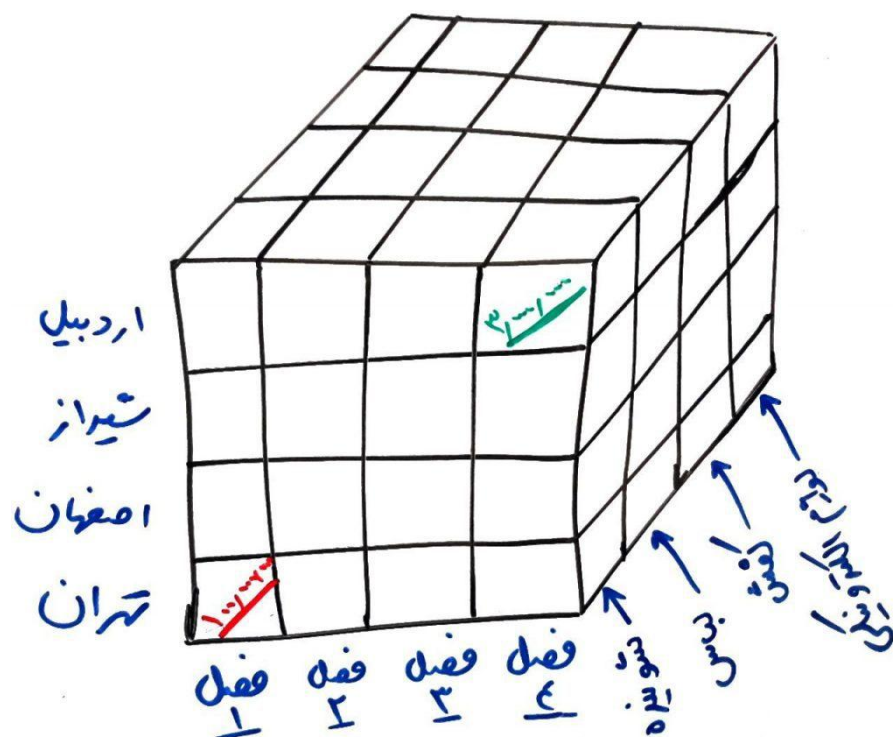
- Information integration and data warehouse construction
 - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology (**See the Next Slide**)
 - Scalable methods for computing (i.e., materializing) multidimensional aggregates
 - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data cube technology

داده ها در دو بعد

ذخیره شده اند

Student #1	19	18.5	185	مرد
Student #2	21	17.0	162	زن
Student #3	22	15.5	177	زن
Student #4	20	18	156	مرد
⋮				



در مکعب داده، داده ها به صورت چند بُعدی نمایش داده می شوند و هر بُعد یک ویژگی از انبار داده ما را نمایش می دهد (زمان فروش، مکان فروش، نوع اجناس فروخته شده)

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
 - A typical association rule
 - Diaper \square Beer [0.5%, 75%] (support, confidence)
 - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: (3) Classification

- Classification and label prediction
 - Construct models (functions) based on some training examples
 - Describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
 - Predict some unknown class labels
- Typical methods
 - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
 - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

Data Mining Function: (3) Classification

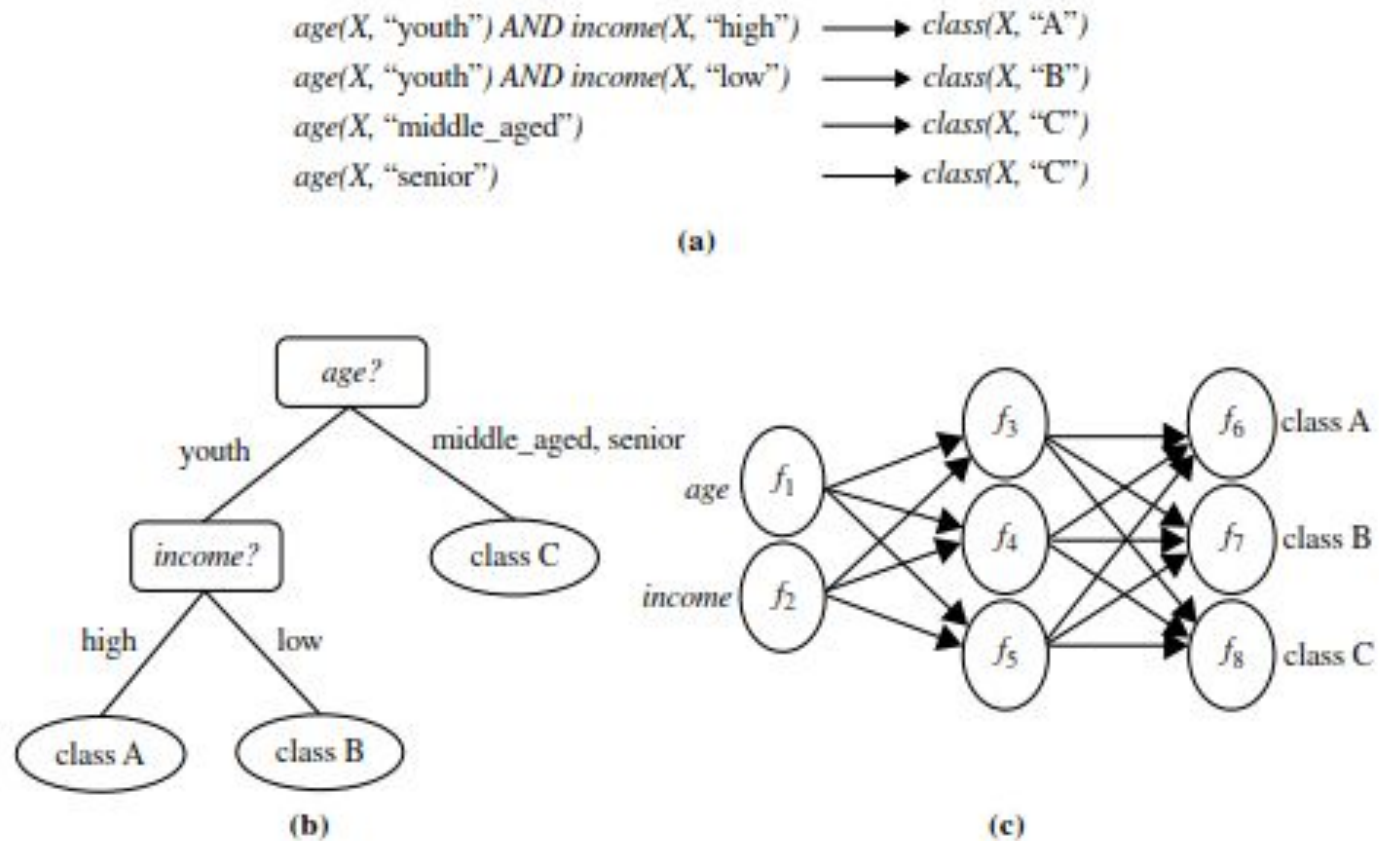


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: (4) Cluster Analysis

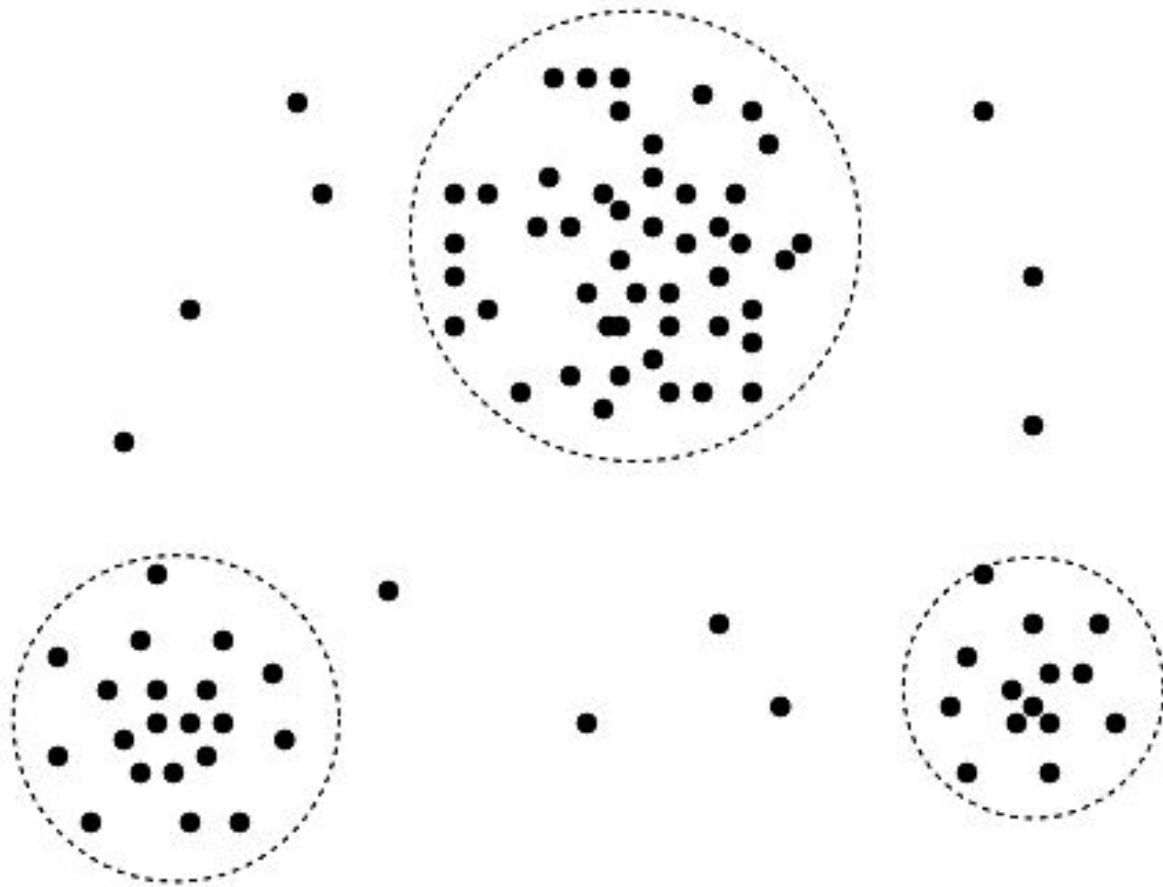


Figure 1.10 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
 - Outlier: A data object that does not comply with the general behavior of the data
 - Noise or exception? — One person's garbage could be another person's treasure
 - Methods: by product of clustering or regression analysis, ...
 - Useful in fraud detection, rare events analysis

What can you do with the data?

- Suppose that you are the owner of a supermarket and you have collected billions of **market basket** data. What information would you extract from it and how would you use it?

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Product placement

Catalog creation

Recommendations

- What if this was an online store?

What can you do with the data?

- Suppose you are a search engine and you have a **toolbar log** consisting of
 - pages browsed,
 - queries,
 - pages clicked,
 - ads clicked

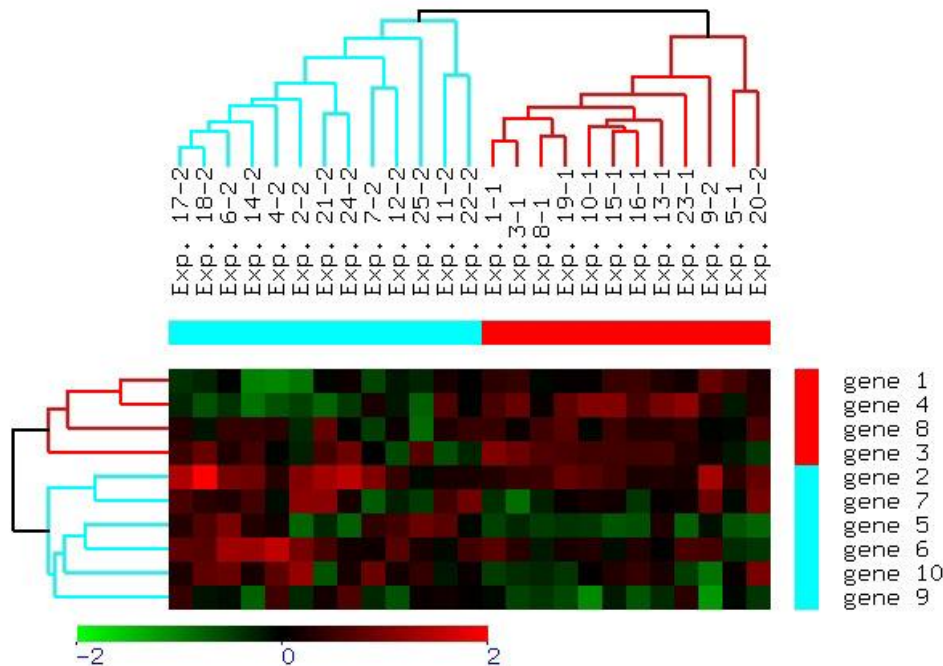
Ad click prediction

Query reformulations

each with a **user id** and a **timestamp**. What information would you like to get out of the data?

What can you do with the data?

- Suppose you are biologist who has **microarray expression data**: thousands of genes, and their expression values over thousands of different settings (e.g. tissues). What information would you like to get out of your data?



Groups of genes and tissues

What can you do with the data?

- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get out of your data?



What can you do with the data?

- You are the owner of a social network, and you have full access to the social graph, what kind of information do you want to get out of your graph?
 - Who is the most important node in the graph?
 - What is the shortest path between two nodes?
 - How many friends two nodes have in common?
 - How does information spread on the network?

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
 - Trend, time-series, and deviation analysis: e.g., regression and value prediction
 - Sequential pattern mining
 - e.g., first buy digital camera, then buy large SD memory cards
 - Periodicity analysis (in time-series)
 - Motifs and biological sequence analysis
 - Approximate and consecutive motifs
 - Similarity-based analysis
- Mining data streams
 - Ordered, time-varying, potentially infinite, data streams

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequential pattern mining:
 - an important data mining task with a wide range of applications from **text analysis to market basket analysis**
 - This database contains four sequences(ordered list of itemsets). Each **sequence** represents the items purchased by a customer at different times.

SID	Sequence
1	$\langle \{a, b\}, \{c\}, \{f, g\}, \{g\}, \{e\} \rangle$
2	$\langle \{a, d\}, \{c\}, \{b\}, \{a, b, e, f\} \rangle$
3	$\langle \{a\}, \{b\}, \{f, g\}, \{e\} \rangle$
4	$\langle \{b\}, \{f, g\} \rangle$

- Find the sequences of items frequently bought by customers

Structure and Network Analysis

- Graph mining
 - Finding frequent subgraphs (e.g., chemical compounds-malware analysis), trees (XML), substructures (web fragments)
- Information network analysis
 - Social networks: actors (objects, nodes) and relationships (edges)
 - e.g., author networks in CS, terrorist networks
 - Multiple heterogeneous networks
 - A person could be multiple information networks: friends, family, classmates, ...
 - Links carry a lot of semantic information: Link mining
- Web mining
 - Web is a big information network: from PageRank to Google
 - Analysis of Web information networks
 - Web community discovery, opinion mining, usage mining, ...

Evaluation of Knowledge

- Are all mined knowledge interesting?
 - One can mine tremendous amount of “patterns” and knowledge
 - Some may fit only certain dimension space (time, location, ...)
 - Some may not be representative, may be transient, ...
- Evaluation of mined knowledge → directly mine only interesting knowledge?
 - Descriptive vs. predictive
 - Coverage(for classification-Similar to support)
 - Typicality vs. novelty
 - Accuracy(for classification)
 - Timeliness
 - ...

What can we do with data mining?

- Some examples:
 - Frequent itemsets and Association Rules extraction
 - Coverage
 - Clustering
 - Classification
 - Ranking
 - Exploratory analysis

Frequent Itemsets and Association Rules

- Given a set of records each of which contain some number of items from a given collection;
 - Identify sets of items (**itemsets**) occurring frequently together
 - Produce **dependency rules** which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Itemsets Discovered:

{Milk,Coke}
{Diaper, Milk}

Rules Discovered:

{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Frequent Itemsets: Applications

- Text mining: finding associated phrases in text
 - There are lots of documents that contain the phrases “association rules”, “data mining” and “efficient algorithm”
- Recommendations:
 - Users who buy this item often buy this item as well
 - Users who watched James Bond movies, also watched Jason Bourne movies.
- Recommendations make use of item and user similarity

Association Rule Discovery: Application

- Supermarket **shelf management**.
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
 - A classic rule --
 - If a customer buys diaper and milk, then he is very likely to buy beer.
 - So, don't be surprised if you find six-packs stacked next to diapers!

Clustering Definition

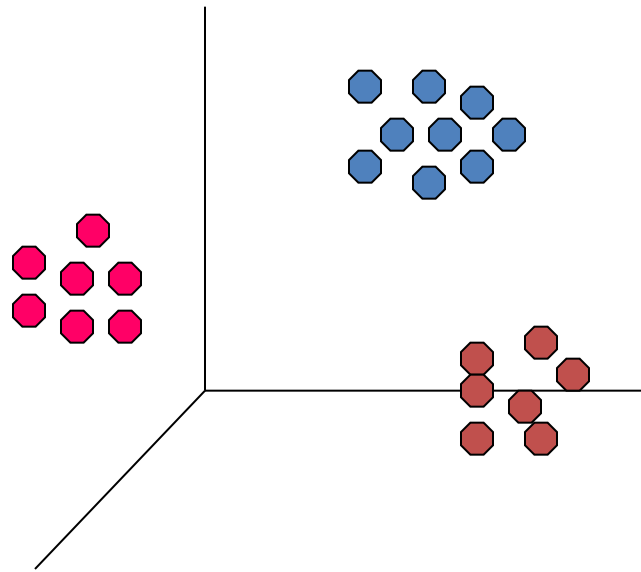
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
 - Data points in one cluster are more similar to one another.
 - Data points in separate clusters are less similar to one another.
- Similarity Measures?
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

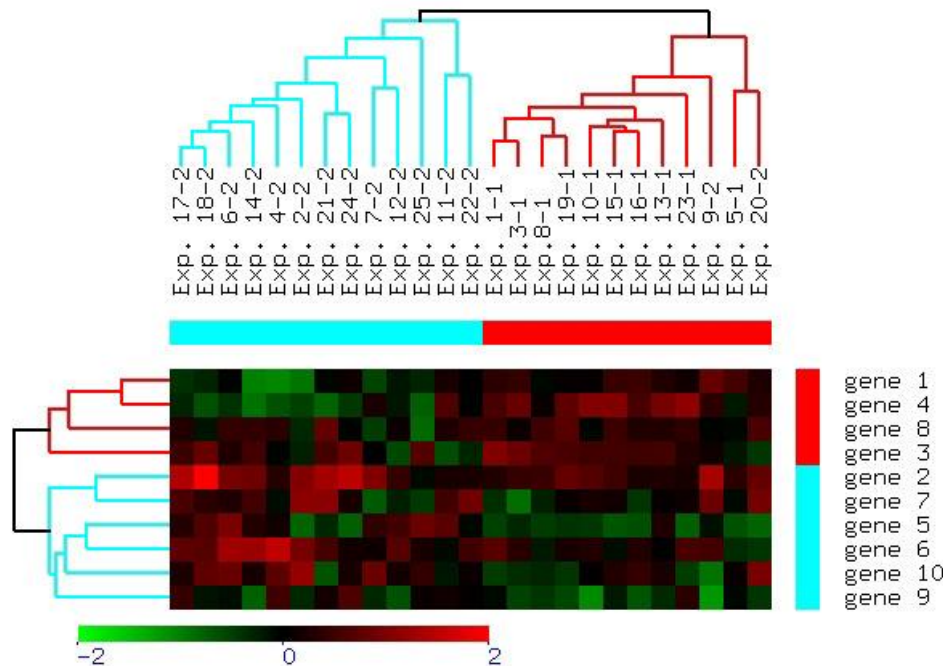
Intracuster distances
are minimized

Intercluster distances
are maximized



Clustering: Application 1

- Bioinformatics applications:
 - Goal: Group genes and tissues together such that genes are coexpressed on the same tissues



Clustering: Application 2

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
 - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Clustering of S&P 500 Stock Data

- Observe Stock Movements every day.
- Cluster stocks if they change similarly over time.

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN, Bay-Net-work-Down, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-Down, Tellabs-Inc-DOWN, Natl-Semiconduct-DOWN, Orac1-DOWN, SGI-DOWN, Sun-DOW N	Technology1-DOWN
2	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP

Coverage

- Given a set of customers and items and the transaction relationship between the two, select a small set of items that “covers” all users.
 - For each user there is at least one item in the set that the user has bought.
- Application:
 - Create a catalog to send out that has at least one item of interest for every customer.

Classification: Definition

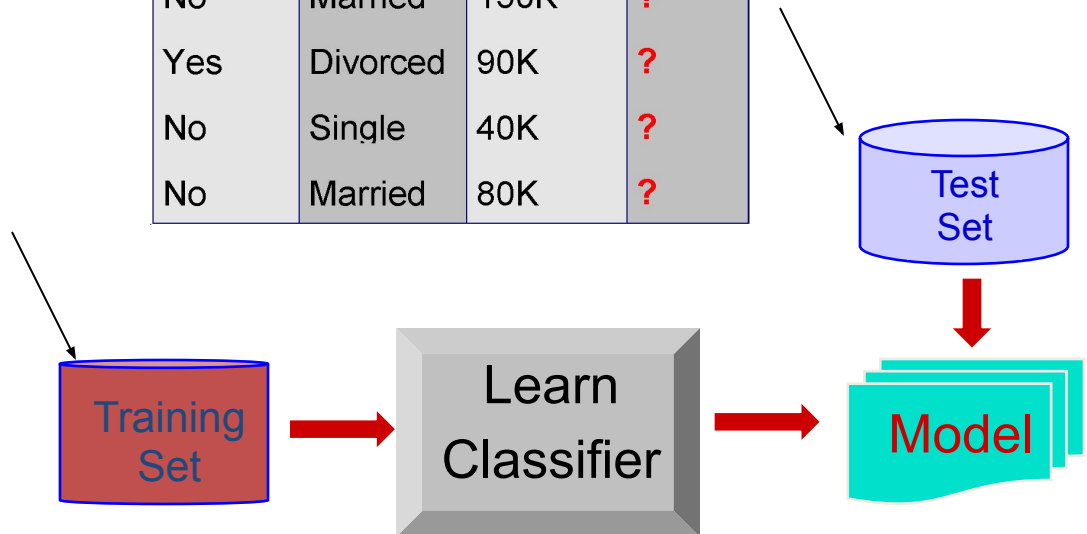
- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Ad Click Prediction
 - Goal: Predict if a user that visits a web page will click on a displayed ad. Use it to target users with high click probability.
 - Approach:
 - Collect data for users over a period of time and record who clicks and who does not. The {click, no click} information forms the class attribute.
 - Use the history of the user (web pages browsed, queries issued) as the features.
 - Learn a classifier model and test on new users.

Classification: Application 2

- Fraud Detection
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - **Label** past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

Link Analysis Ranking

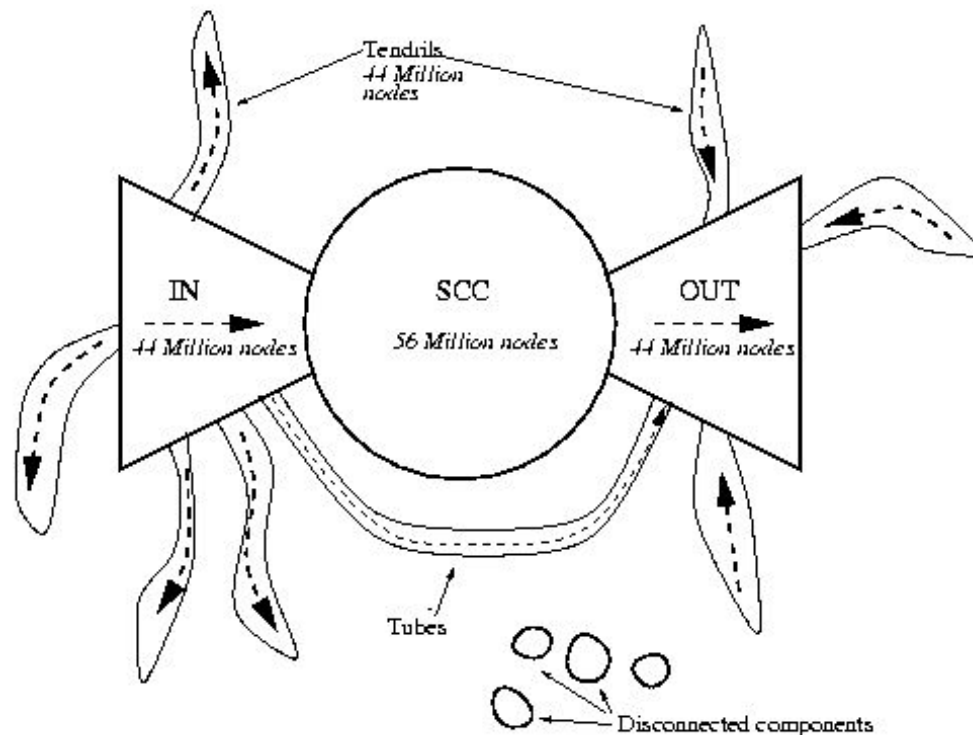
- Given a collection of web pages that are linked to each other, rank the pages according to importance (**authoritativeness**) in the graph
 - Intuition: A page gains authority if it is linked to by another page.
- Application: When retrieving pages, the authoritativeness is factored in the ranking.

Exploratory Analysis

- Trying to understand the data as a **physical phenomenon**, and describe them with simple metrics
 - What does the web graph look like?
 - How often do people repeat the same query?
 - Are friends in facebook also friends in twitter?
- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- It helps our understanding of the world, and can lead to **models** of the phenomena we observe.

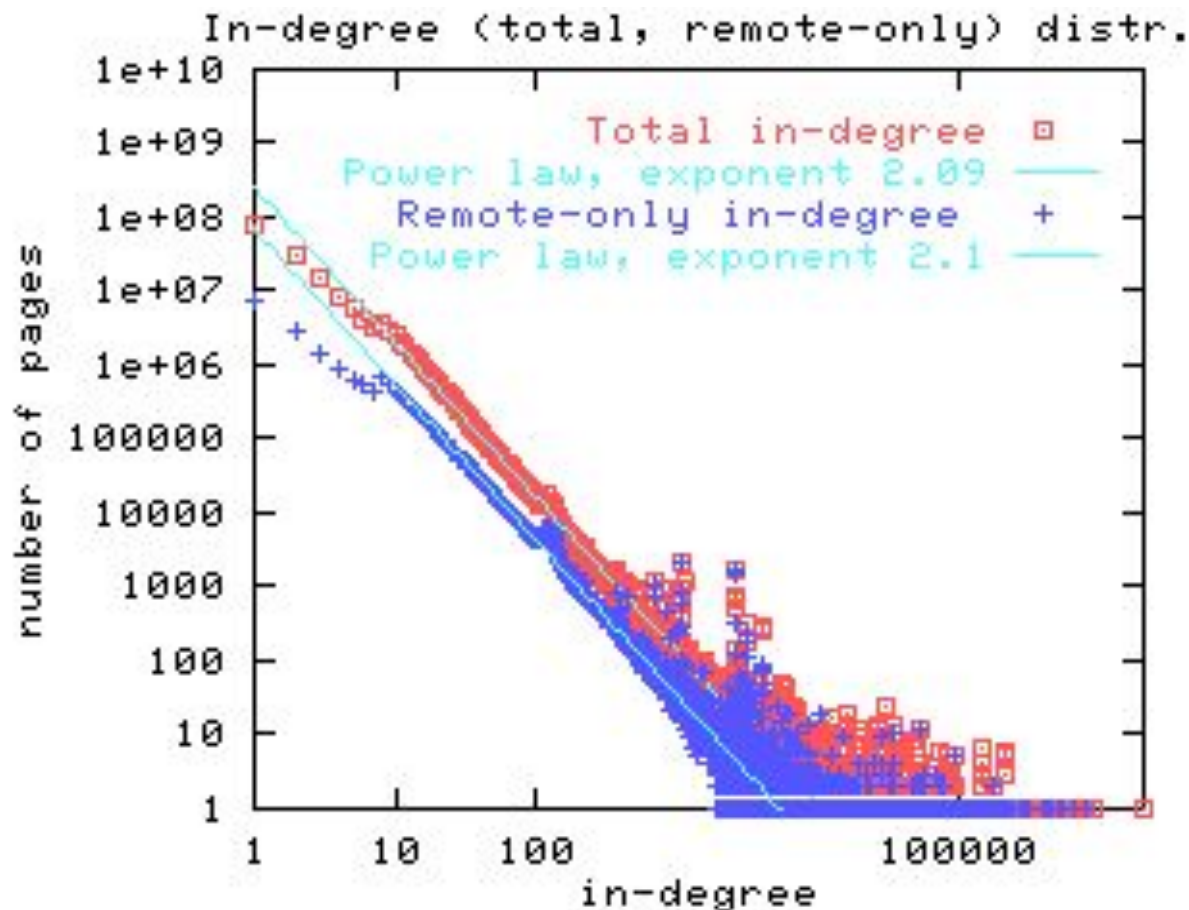
Exploratory Analysis: The Web

- What is the structure and the properties of the web?
- The Bow-Tie Structure of the Web



Exploratory Analysis: The Web

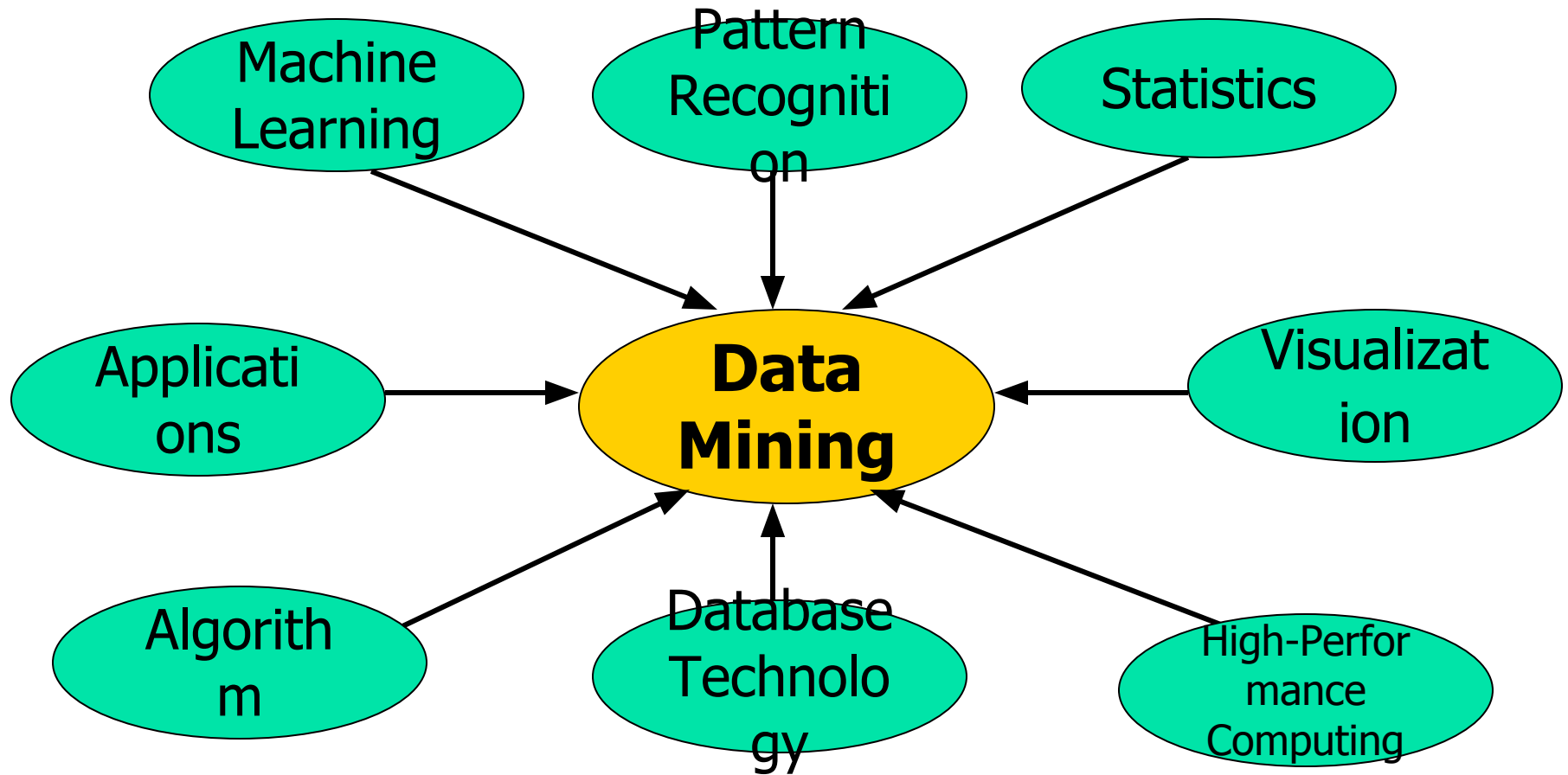
- What is the distribution of the incoming links?



Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used? 
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary


Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
 - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
 - Micro-array may have tens of thousands of dimensions
- High complexity of data
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data
 - Structure data, graphs, social networks and multi-linked data
 - Heterogeneous databases and legacy databases
 - Spatial, spatiotemporal, multimedia, text and Web data
 - Software programs, scientific simulations
- New and sophisticated applications


Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted? 
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining 
- A Brief History of Data Mining and Data Mining Society
- Summary

Major Issues in Data Mining (1)

- Mining Methodology
 - Mining various and new kinds of knowledge
 - Mining knowledge in multi-dimensional space
 - Data mining: An interdisciplinary effort
 - Boosting the power of discovery in a networked environment
 - Handling noise, uncertainty, and incompleteness of data
 - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
 - Interactive mining
 - Incorporation of background knowledge
 - Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- Efficiency and Scalability
 - Efficiency and scalability of data mining algorithms
 - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
 - Handling complex types of data
 - Mining dynamic, networked, and global data repositories
- Data mining and society
 - Social impacts of data mining
 - Privacy-preserving data mining
 - Invisible data mining

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
 - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
 - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (**KDD**)
 - SIAM Data Mining Conf. (**SDM**)
 - (IEEE) Int. Conf. on Data Mining (**ICDM**)
 - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (**ECML-PKDD**)
 - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (**PAKDD**)
 - Int. Conf. on Web Search and Data Mining (**WSDM**)
- Other related conferences
 - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
 - Web and IR conferences: WWW, SIGIR, WSDM
 - ML conferences: ICML, NIPS
 - PR conferences: CVPR,
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
 - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
 - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
 - Conferences: SIGIR, WWW, CIKM, etc.
 - Journals: WWW: Internet and Web Information Systems,
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Recommended Reference Books

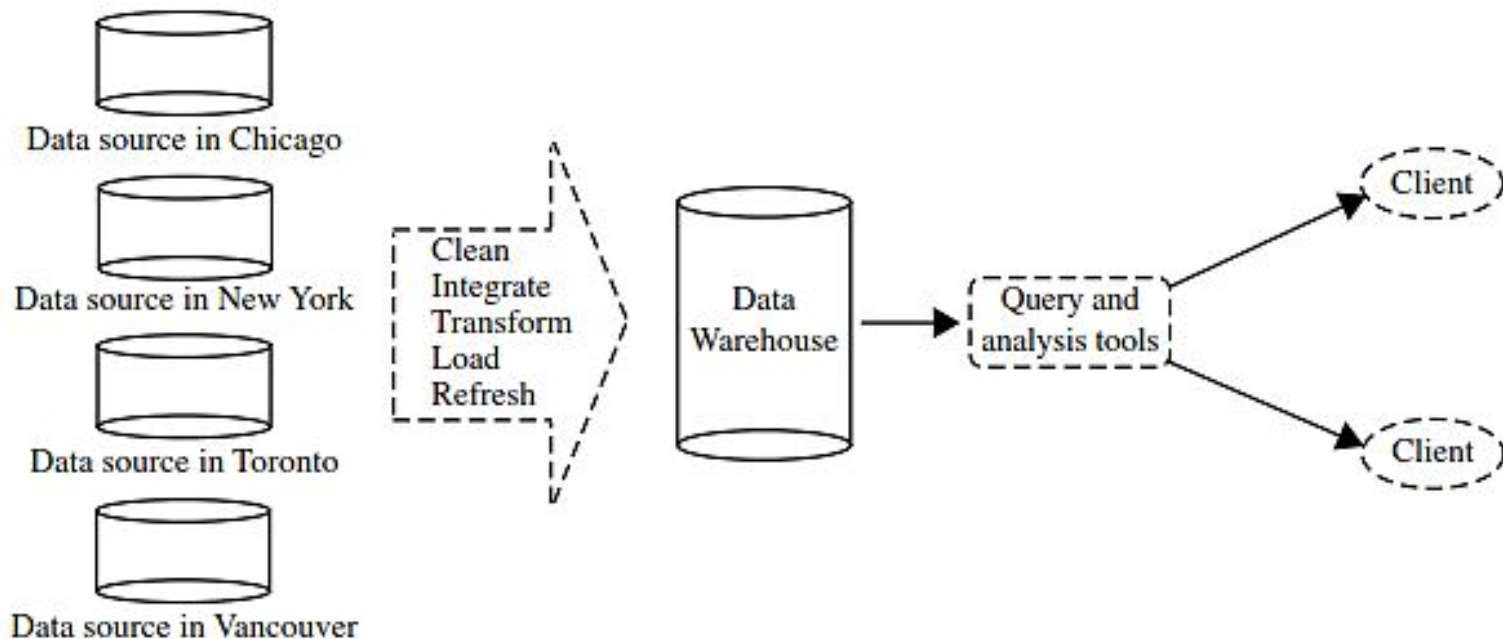
- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkha, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005



Additional Slides

Data Warehouses

- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.



Typical framework of a data warehouse