LECTURE 1:
A BRIEF INTRODUCTION TO DATA MINING

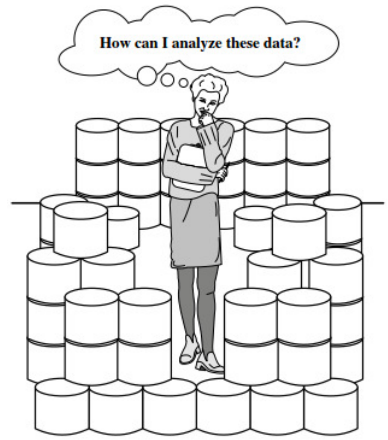
By: Jiawei Han (Additions and modifications: Siamak Sarmady)

Outline

- →Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

Why Data Mining?

- **Explosive Growth of Data:** our capability of generating, collecting, storing, and managing data has grown tremendously in the last 50 years.
  - Major sources of abundant data
    - **Business:** e-commerce, transactions, stocks, …
    - **Science:** Remote sensing, bioinformatics, scientific simulation, …
    - **Computerized Society:** web, news, digital media, social networks, …
    - **Government:** gathers information about people, economy, healthcare,…
  - We are drowning in data but starving for knowledge! (knowledge is deeply buried in data)
- **Main Reason for Data Mining:** Gathering Knowledge



Why Data Mining?

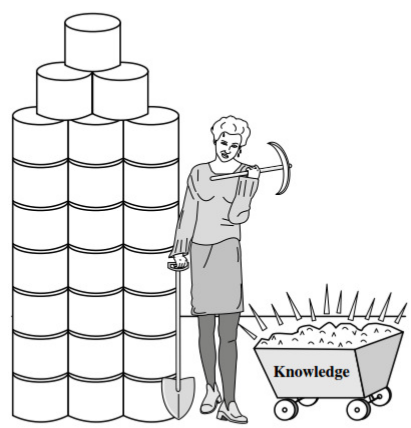
- Decisions are often made based on not the information obtained from the data repositories. They are typically made based on the decision maker's intuition.
  - That's because the decision makers (top management) do not have the tools to extract the valuable knowledge embedded in the vast amounts of data.
- The knowledge obtained from experts and research might be biased or wrong. We can use data mining to verify the knowledge.
- **Data mining:** automated and scalable analysis of massive data sets

Outline

- Why Data Mining?
- →What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

What Is Data Mining?

- **Data mining (knowledge discovery from data):**
  - **Aim:** extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - **Data mining, a misnomer?** we don't mine data, we mine something from data. Should rather be "Knowledge mining from data"!
  - **Examples:** Google's Flu Trends is an example. It uses the trend of searching specific terms as an indication of flu activity in an area. It estimates flu activity and trend up to two weeks faster than traditional systems. The method could be used to predict a contagion.

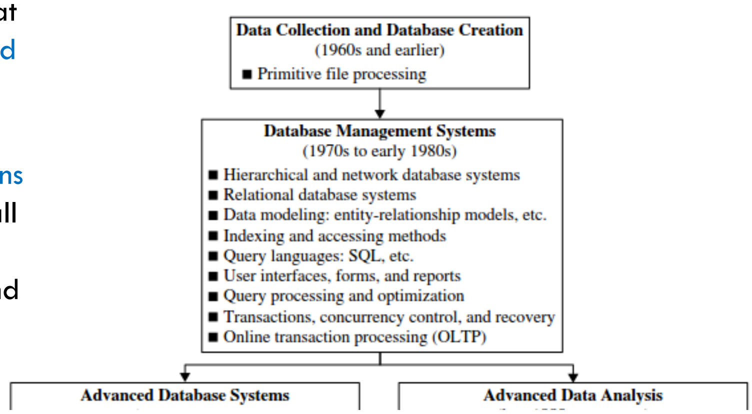


What Is Data Mining?

- **Alternative names:**
  - Knowledge discovery (mining) in databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Data archeology
  - Data dredging
  - Information harvesting
  - Business intelligence
- **Relationship with other disciplines:**
  - Machine learning, pattern recognition, statistics, databases, business intelligence, big data, ……

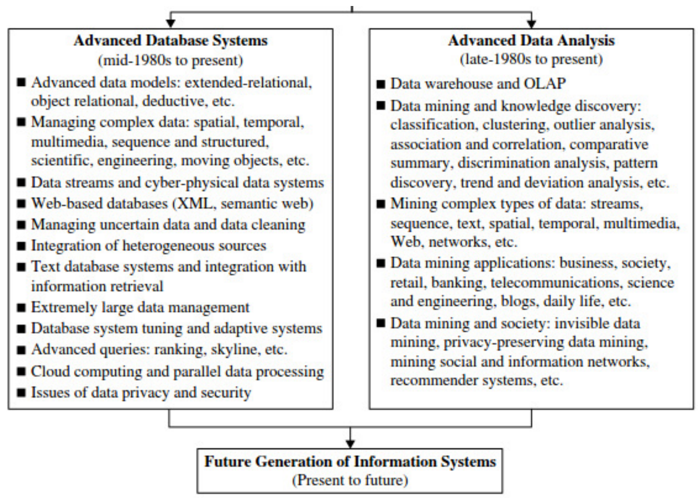
Evolution of Information Science

- **OLTP:** Online Transaction Processing (OLTP) is a class of applications that are suitable for transaction-oriented systems such as banking systems.
- **Transaction:** An operation comprising of several sub-operations is performed in an atomic way. If all sub-operations are successful, the transaction is deemed successful and closed, otherwise the effects of all operations are reversed (i.e. modifications on objects and databases).

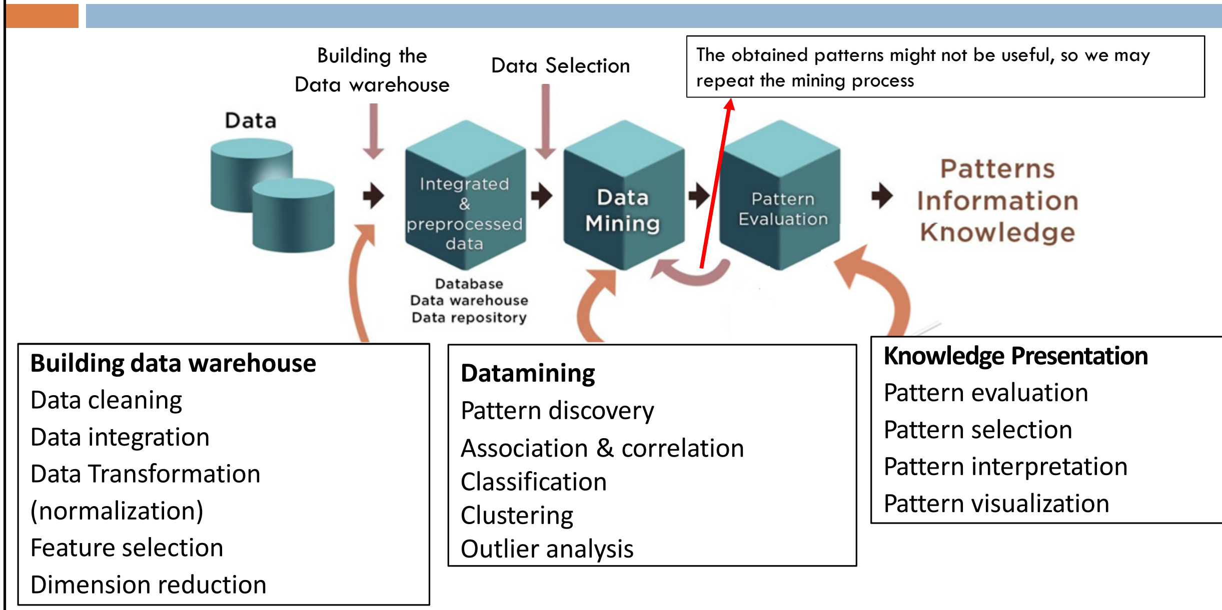


Evolution of Information Science

- **Data warehouse:** a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. Requires data cleaning, data integration, OLAP.
- **OLAP** (Online Analytical Processing): analysis techniques with functionalities like: summarization, consolidation, aggregation, and viewing data from different angles and aspects.
- **Data mining tools:** frequent pattern extraction, Classification, Clustering, Outlier/Anomaly detection, Characterization of changes over time.



Data Mining: A Knowledge Discovery (KDD) Process

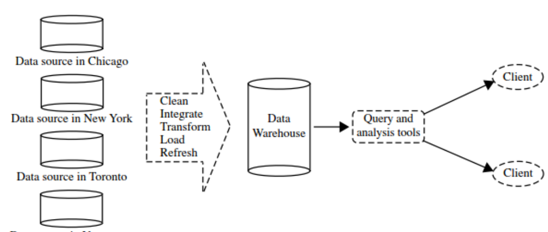


Knowledge Discovery (KDD) Process (an example framework)

- Data Mining usually involves
  - **Building data warehouse:** refers to the above four operations
  - **Data Selection:** to select relevant data for an analysis or mining task from the database
  - **Data mining:** to use intelligent methods to extract data patterns
  - **Pattern evaluation:** to identify truly interesting patterns that represent knowledge
  - **Gathering knowledge:** conclude interesting patterns found from data mining into knowledge form
  - **Knowledge presentation:** to present mined knowledge in visual forms

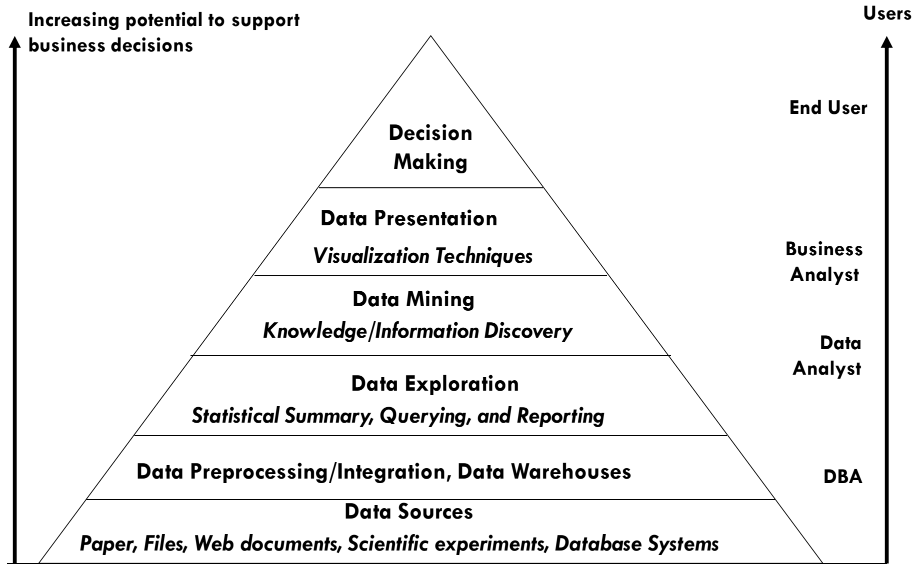
Building a Data warehouse

- **Data warehouse:** is a repository of information collected from multiple sources (e.g. all branches of a bank or chain store) , stored under unified schema, and usually residing at a single site. It is created via a process of:
  - **Data cleaning:** to remove noise and inconsistent data
  - **Data integration from multiple sources:** to combine multiple data sources into a single schema
  - **Data Transformation:** to transform data and consolidate it into appropriate form i.e. normalization, formatting values into a single format (e.g. different date formats)
  - **Data loading (cube construction):** to perform aggregation and summary operation and to build multidimensional representations of a desired attribute
  - **Periodical Data Refresh**



Data Mining in Business Intelligence

Who does what?



Outline

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - →What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

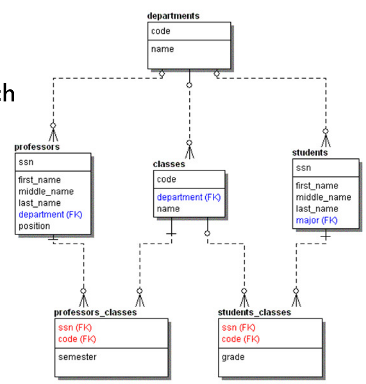
Data View: On What Kinds of Data?

- **Structured and semi-structured data**
  - Database data: relational/object-relational
  - Data warehouse data
  - Transactional data
- **Unstructured data**
  - Text data and Web data
  - Spatial and spatiotemporal data
  - Multimedia data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Graphs, social networks and information networks

Database Data (1)

- **Relational Database:** a collection of tables (mostly represent entities), with a set of columns (attributes). Each row in a table is called record or a tuple (in NoSQL).
  - **Tuple:** Each tuple stores an entry or object that is identified with a unique key.
  - **Entity-relationship (ER) model:** A semantic data model that describes the database i.e. the entities and relationships
- **Example:** the database of "All Electronics" company:
  - The company(data) is described using some tables, some of tables represent entities. Tables include: customer, item, employee and branch

customer	*(cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .)*
item	*(item_ID, brand, category, type, price, place_made, supplier, cost, . . .)*
employee	*(empl_ID, name, category, group, salary, commission, . . .)*
branch	*(branch_ID, name, address, . . .)*
purchases	*(trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)*
items_sold	*(trans_ID, item_ID, qty)*
works_at	*(empl_ID, branch_ID)*



Database Data (2)

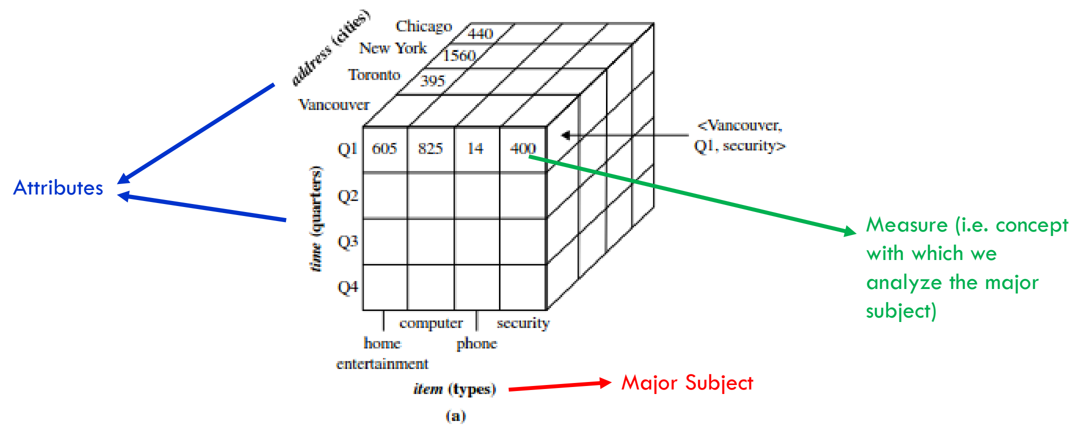
- **Database Queries:** a query written in SQL language.
  - A **complex query** is transformed into a set of relational operations, such as join, selection and projections.
  - It is then optimized for efficient processing (by changing the queries and usage of cache).
  - The result is either a subset of database data or an operation that is performed on the database.
- **Mining Relational Databases:**
  - The structure of typical database is not very suitable for OLAP and data mining applications. That's because these databases are normally designed with OLTP (i.e. transactional) applications.
  - However it is still possible to discover trends and simple data patterns in such database.
    - It is possible to predict credit risk of new customers based on their income, age and previous credit information.
    - It is possible to discover a significant increase in price of a product.

Data warehouse (1)

- **Major Subjects:** Data warehouses (cubes) are typically organized around major subjects (e.g., customer, item, supplier, and activity) around which we want to get some insight.
  - Data typically provides historical perspective e.g. the past 6 to 12 months in summarized form.
    - Instead of storing details of each sales transaction, a summary of the transactions per item type for each store (or each sales region) is stored.
- **Data Cube:** A multidimensional data structure in which each dimension corresponds to an attribute in the schema (e.g. item, time, region). It provides a multidimensional view of data and precomputed summarizations of it.
  - One of these attributes, is normally the major subject of the cube (e.g. item)
  - Since the aggregations and summarizations are precomputed, access to them is very fast.
- **Measure:** Each cell of a data cube stores the value of an aggregate measure such sum(sales amount). This measure is the main concept of each cube.
  - We first decide what are the measures that we want to analyze (e.g. sales)
  - We may build several cubes for different measures of interest like "Sales value", "Sales count".

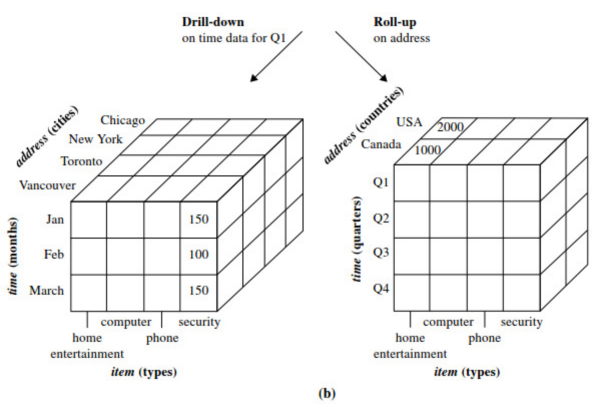
Data warehouse (2)

- **Example:** This shows one of the data cubes used in "All Electronics" company.
  - **Attributes:** Time, Item*, Address.
  - **Measure:** Sales amount (in $1000)which



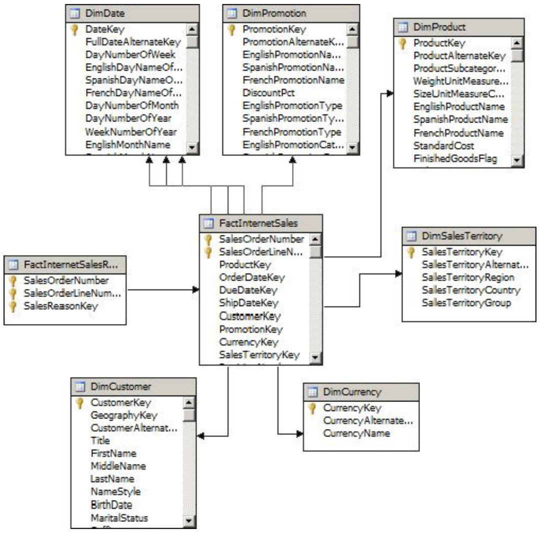
Data warehouse (3)

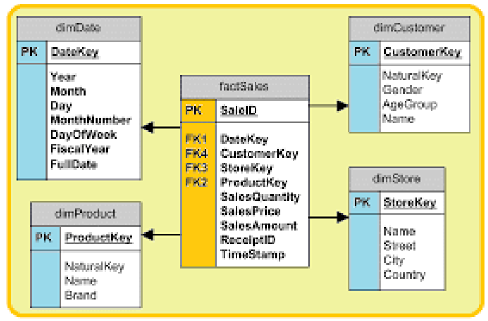
- Data is stored at different levels (i.e. granularity) for each attribute. It is possible to view the cube at different levels of each attribute.
  - **Roll-up:** view the cube at higher levels of an attribute (e.g. countries or states instead of cities)
  - **Drill-down:** view the cube at lower levels of an attribute (e.g. months instead of seasons or years)



Data warehouse (4)

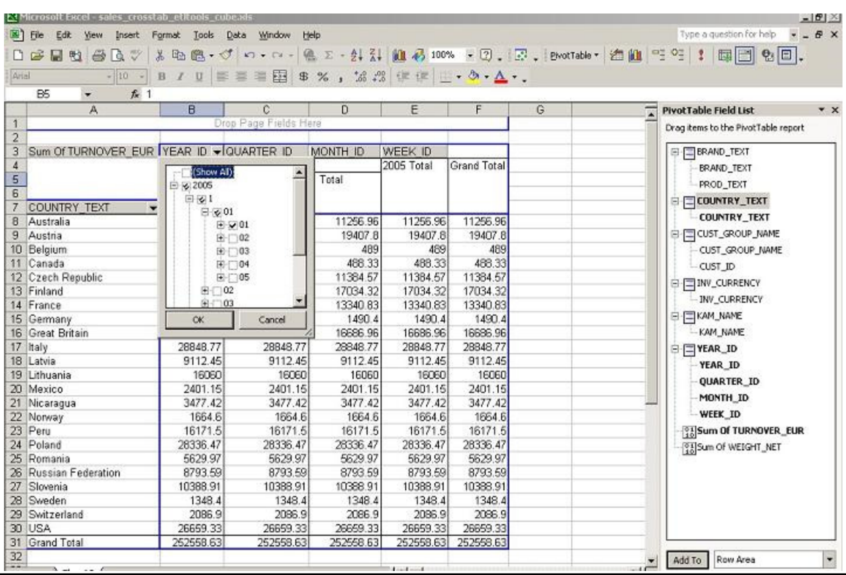
- In order to build a cube, data is integrated from several relational tables…





Data warehouse (5)

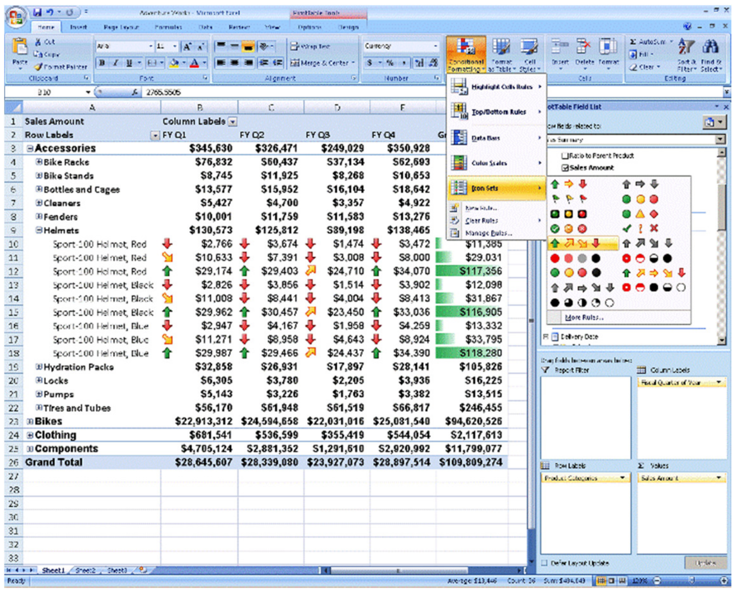
- Note that the cube is not visualized as a multidimensional graphical object! (A cube connected to excel)



Sum Of TURNOVER_EUR	YEAR_ID	QUARTER_ID	MONTH_ID	WEEK_ID	
			Total	2005 Total	Grand Total
COUNTRY_TEXT					
Australia			11256.96	11256.96	11256.96
Austria			19407.8	19407.8	19407.8
Belgium			489	489	489
Canada			488.33	488.33	488.33
Czech Republic			11384.57	11384.57	11384.57
Finland			17034.32	17034.32	17034.32
France			13340.83	13340.83	13340.83
Germany			1490.4	1490.4	1490.4
Great Britain			16686.96	16686.96	16686.96
Italy	28848.77	28848.77	28848.77	28848.77	28848.77
Latvia	9112.45	9112.45	9112.45	9112.45	9112.45
Lithuania	16060	16060	16060	16060	16060
Mexico	2401.15	2401.15	2401.15	2401.15	2401.15
Nicaragua	3477.42	3477.42	3477.42	3477.42	3477.42
Norway	1664.6	1664.6	1664.6	1664.6	1664.6
Peru	16171.5	16171.5	16171.5	16171.5	16171.5
Poland	28336.47	28336.47	28336.47	28336.47	28336.47
Romania	5629.97	5629.97	5629.97	5629.97	5629.97
Russian Federation	8793.59	8793.59	8793.59	8793.59	8793.59
Slovenia	10388.91	10388.91	10388.91	10388.91	10388.91
Sweden	1348.4	1348.4	1348.4	1348.4	1348.4
Switzerland	2086.9	2086.9	2086.9	2086.9	2086.9
USA	26659.33	26659.33	26659.33	26659.33	26659.33
Grand Total	252558.63	252558.63	252558.63	252558.63	252558.63

Data warehouse (6)

- Earlier versions



Transactional Data and Databases

- **Transactional Database:** each record (tuple) in such a database captures a transaction (e.g. customer purchase, inventory operation, order, quote, …). A transaction normally includes a unique ID and a list of items making up the transaction.
  - Notice that the transaction details might be stored in more than one table (e.g. a table for transactions and another one for its items).

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
…	…

- **Mining example:** An analyst might use the above database to gain useful knowledge (e.g. PC and Monitor).
  - "Which items sold well together". Answering that question may allow a marketer to increase the sales by bundling specific products together with some discount.

Mining other kinds of Data (Unstructured?)

- We did not discuss some other kinds of data that can be mined (such as data streams, spatial and temporal data, time-series, ordered/sequence data, text and web, multi-media, graphs and networked data).
- These other data bring new challenges on how to:
  - **Structure:** handle data carrying special structures (e.g. sequences, trees, graphs and networks)
  - **Semantic:** handle specific semantics (e.g. ordering, html, image, audio, video, connectivity, …)
  - **Patterns in Rich Structures:** mine patterns that occur in rich structures and semantics (i.e. within a network or an Html)
    - Finding patterns (e.g. frequent ones) could be difficult or need special algorithms…

Mining other kinds of Data (Unstructured?) – Example Applications

- **Temporal data:**
  - analyzing the timing of bank branch transactions might help with scheduling of tellers based on the volume of customer traffic.
  - Mining stock exchange data can uncover trends and help in devising better investment strategies.
- **Streams:**
  - By mining data streams on a network we can identify intrusions based on anomalies in message flows.
- **Spatial data:**
  - We may look for changes in poverty rates based on the distance of cities to highways and availability of specific facilities.

Mining other kinds of Data (Unstructured?) – Example Applications

- **Text Data:**
  - By mining historical text data, we can identify the evolution of hot topics.
  - By mining user comments about products we can assess customer sentiments about the product and market
- **Multimedia data:**
  - We can identify objects, classify them by assigning semantic labels or tags.
  - By mining videos we can identify the sequence of events (e.g. in a football game).
- **Web:**
  - We can classify web pages and uncover dynamics, association and relationships between them
  - We can specify their relevance to a topic or search query
- **Mining different types at the same time:**
  - We may find much more interesting patterns and knowledge by mining several related data with different types.

Outline

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - →What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

Data Mining Functions

Patterns we discover, typically depend on the functions or methods we use to discover them.

- **Methods of data mining:**
  - Class/Concept Description: Characterization and Discrimination
  - Mining Frequent Patterns, Associations and Correlations
  - Predictive analysis - Classification and Regression
  - Cluster Analysis
  - Outlier Analysis

Data Mining Function: (1) Class Description: Characterization and Discrimination

- Data entries can be categorized or associated with classes or concepts (by obvious properties of items). These classes and concepts should be described with concise and precise terms.
  - In AllElectronics company classes of items for sales include: computers, printers, …
  - Classes of customers include: bigSpenders and budgetSpenders
- We may then find patterns in relation to classes/concepts … (e.g. increase of job opportunities increases the sales of computer equipment)
- The classes can be derived in two ways:
  - **Data characterization:** summarizes the data based on the characteristics or features of the target class of data
    - Classifying sales items as computers, printers, … i.e. based on their properties
    - Classifying customers by their age, income range, employment, credit rating …
  - **Data discrimination:** describes the classes by comparison of the target class with one or a set of contrasting classes
    - Compare customers and put 20% of them in the bigSpenders and 80% in budgetSpenders class
  - **By both:** uses a mix of both characterization and discrimination methods to specify classes (male/female customers that are bigSpenders or budgetSpenders)

Data Mining Function: (2) Mining Frequent Patterns, Associations and Correlations

- Frequent patterns (or frequent itemsets)
  - Frequent itemsets – *with each other*
    - What items are frequently purchased together in your Walmart?
  - Frequent subsequences (sequential patterns) – *after each other*
    - like the customer tend to buy a computer first, then later a printer, then tonners…
    - Periodicity, trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Frequent substructures – *with similar structure*
    - Refer to different structural forms (trees, graphs, lattices) that might be combined with itemsets or subsequences
- **A current research question:** How to mine such patterns and rules efficiently in large datasets?

Data Mining Function: (2) Mining Frequent Patterns, Associations and Correlations

- **Frequent itemsets:** an example of such a mined rule
    - Buys(X,"computer") → Buys(X, "software") [support = 1%, confidence = 50%]
    - Or written simpler: computer → software [1%, 50%]
  - **Association:** Involves a predicate like "Buys", and an association rule (i.e. the part after →)
    - If rule section has just one predicate it is called "single-dimensional association rule".
    - A multi-dimensional rule:
      Age(X, "20..29") ∧ income(X,"40k..49k") → Buys(X, "laptop") [support = 1%, confidence = 50%]
  - **Support:** analysis shows 1% of the analyzed transactions show these purchases together
  - **Confidence (certainty):** means if a customer buys a computer, there is 50% chance that he buys a software as well
  - **Interesting rules:** typically, only rules that exceed a "minimum support threshold" and a "minimum confidence threshold" are kept.
- **Correlation:** just like association but expresses the changes of two things being correlated

Data Mining Function: (3) Predictive analysis - Classification and Regression

- **Classification and label prediction:** is a supervised method since it is trained with properly classified (labeled) examples
  - **Training data (examples):** some items (that have attributes) with classes (label) specified for them
  - **Model:** a function constructed based on the training data (examples) that maps attributes to classes
  - **Prediction:** the model (function) can now be used to predict the class (label) of unseen data.
    - Predict whether a patient has cancer using his data, Predict whether the risk level of a loan, predict whether a transaction is fraud or not.

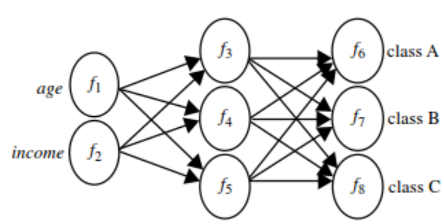
- **Regression:** a supervised method, very similar to classification, but continuous.
  - Models continuous functions. Finds a function describing the relation of parameters and a concept
  - Statistical method, predicts missing or unknown numerical data values rather than discrete classes

Data Mining Function: (3) Predictive analysis - Classification and Regression

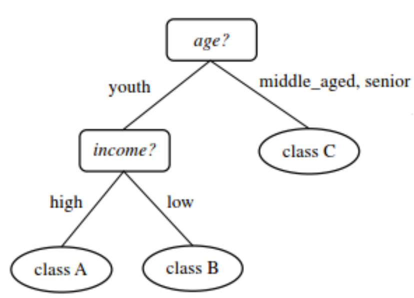
- **Classification methods:** Decision trees, naïve Bayes classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …

age(X, "youth") AND income(X, "high") → *class(X, "A")*
age(X, "youth") AND income(X, "low") → *class(X, "B")*
age(X, "middle_aged") → *class(X, "C")*
age(X, "senior") → *class(X, "C")*

Rule Based



Neural Network



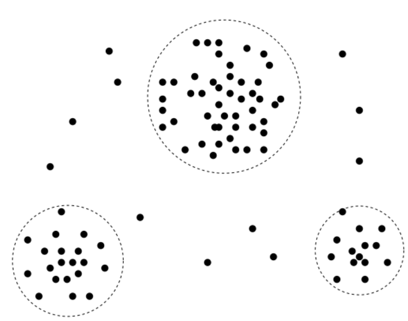
Decision Tree

Data Mining Function: (3) Predictive analysis - Classification and Regression

- **Typical applications:**
  - Credit card fraud detection, spam detection, direct marketing, classifying stars, diseases, web-pages, …
- **Example:**
  - Sales prediction (class):
    - Existing sales information provides details about the items (properties, price, sales period etc., total sales, profit). Based on this data, it is possible to build a model that predicts a new item will fall into which of these classes: good_response, midl_response, no_response
    - Some classifiers could reveal which attribute has the most significant effect on the class selection. Using a decision tree method might possibly reveal that the most important factor in determining the level of sales is price (and then brand, …)
  - Sales prediction (amount):
    - In previous example assume, instead of sales class (level of sales) we are interested in an approximate of the sales amount each item will have. This time we are predicting the value of a continuous function.

Data Mining Function: (4) Cluster Analysis

- **Clustering:** to put dispersed data items into groups or clusters based on their attributes.
- **Unsupervised learning:** no training using classified/labeled examples is done
- **Purpose:** to finds distribution patterns and distinguish similar items
- **Principle:** maximizing intra-class (items inside a class) similarity & minimizing interclass (items in different classes) similarity. Meaning that it tries to put items very similar to each other in the same groups while putting those different from each other into different groups.
- **Applications:** There are many applications
  - Clustering customers and target each group with different campaigns or offers, grouping news items
- **Methods:** K-Means, Hierarchical Clustering, Correlation clustering, Mean Shift Clustering…



Data Mining Function: (5) Outlier Analysis

- **Outlier:** A data object that does not comply with the general behavior of the data (or it somehow is very different with the majority of data)
  - **Noise:** In some occasions we will remove these data assuming that they are affected by noise. This might improve the performance of data mining functions
  - **Anomaly (rare events):** In some other occasions, the outlier data is exactly what we are looking for. Abnormal network activity for example, might point to intrusions. In a factory, such data might point to failures and problems. In bank transactions, such anomalies may point to fraud (based on transaction attributes such as amount, location, purchase frequency).
- **Methods:** by product of clustering or regression analysis, …
  - **Using statistical tests:** assuming a specific distribution, those too different from the distribution are assumed as outliers.
  - **Distance measures and Clustering:** Data that is dispersed far outside the clusters might be outlier
  - **Density based methods:** can identify outliers in a local region (areas inside the cluster that are sparse but have a few items) even if they seem normal from distribution point of view

Other Patterns - Structure and Network Analysis

Some other patterns can be discovered using custom and innovative methods:

- **Graph mining**
  - Finding frequent subgraphs (e.g., chemical compounds), 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 in multiple information networks (social, …): friends, family, classmates, …
- **Web mining**
  - Search
    - PageRank to Google (i.e. discover importance from links)
  - Analysis of Web information networks
    - opinion mining, usage mining, Web community discovery …

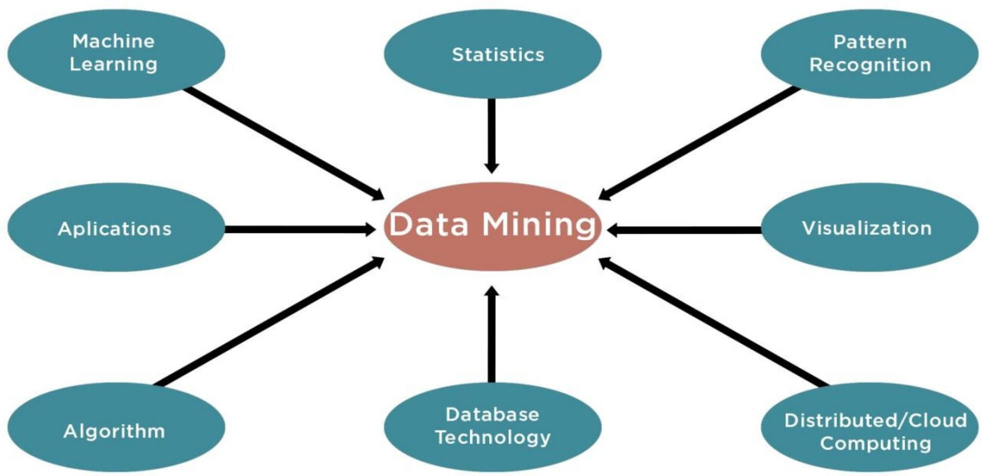
Are all Patterns Interesting?

- Mining process can potentially find thousands or millions of patterns or rules.
  - A mining algorithm is complete, if it can find all the interesting (and uninteresting) patterns.
- As mentioned earlier, an interesting pattern should satisfy a minimum support and confidence threshold.
- In another language, a pattern is interesting if
  1. **Easily understood by humans:** some discovered patterns might be too complex or not easily understandable and therefore usable to humans
  2. **Valid on new or test data with some degree of certainty:** apply to data other than the test data
  3. **Potentially useful:** provide some type of benefit, otherwise no one might be interested in them
  4. **Novel:** people will find a pattern or knowledge interesting if it is not already known
  - A pattern is also interesting if it validates a hypothesis that the user sought to confirm.
- An interesting pattern represents knowledge. So the output of Mining process is typically knowledge.

Outline

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - → What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

Methodology View: Confluence of Multiple Disciplines



Technologies Used - Statistics

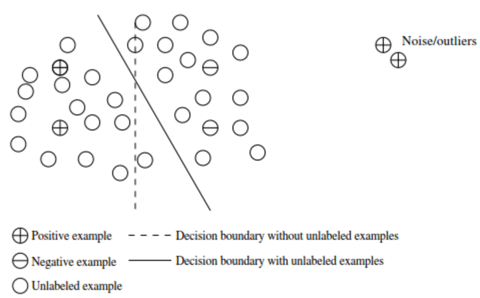
- **Statistical Model:** is a set of mathematical functions that describe the behavior of an item in a target class, in terms of random variables and their probability distributions. The outcome of a data mining task could be models (i.e. a statistical distribution).
  - We can predict the outcome of 10 coin toss with a statistical (probabilistic) model.
  - Using a statistical model, we can predict sales in a specific month and sales of each item.
- **Applications:** prediction (simulation), identifying and handling noise, and missing data, verifying data mining results
- **Statistical hypothesis test (confirmatory data analysis):** a result is called statistically significant if it is unlikely to have occurred by chance.
  - E.g. the relation between specific dosage of a medicine, has significant effect on a disease…

Technologies Used – Machine Learning

- **Machine Learning:** The field that investigates how computers can learn (or improve their performance) using data.
  - **Supervised Learning:** In this method learning happens by providing examples to the algorithm. Examples are data items that are properly classified (i.e. labeled). Algorithm learns how to classify data items and it can classify unseen data items later.
  - **Unsupervised Learning:** These algorithms are unsupervised because no supervision for the learning process through examples is provided. This category of algorithms, put data items into clusters or groups based on their similarities.
  - **Semi-supervised learning:** A mix of the above methods. Small number of labeled examples are used to learn classes. Unlabeled examples are used to refine boundaries and assign the unlabeled data to each cluster.

Technologies Used – Machine Learning

Semi supervised learning



- **Active learning:** The goal is to optimize the model quality by actively acquiring knowledge from human users (there is a limit on how many labels can be asked from users).
- **Focus of research in machine learning:** is on improving and finding models with higher accuracy.
- **Focus of research in data mining:** is mostly on efficiency and scalability for large data sets, as well has the capability of handling complex data types.

Technologies Used – Database Systems and Data Warehouses

- **Data mining capable databases:** recent database systems have built in data analysis capabilities i.e. data warehousing and data mining facilities.
  - A data ware house typically stores data in data cubes. The data cube model facilitates OLAP and allows multidimensional data mining.

- **Database research in data mining field:** focuses on databases and data structures (e.g. data cube) that can handle large data sets. Such databases should be scalable and can handle real-time and fast streaming data.

Technologies Used – Information Retrieval

- **Information retrieval:** is the science of 1) searching for documents or 2) searching for information in documents. Information retrieval assumes:
  - Data under search are unstructured (like data of web pages do not have proper structure).
  - Queries are mainly formed by keywords and do not have complex structures (unlike SQL queries)

- Typical document search methods adopt probabilistic models.
  - **Language model:** is the probability distribution function that determines the frequency of words in documents.
    - The similarity between two documents can be determined by comparing their language models.
  - **Topic model:** A language model that determines the probability distribution function over the documents (vocabulary) of a topic.
  - **Finding documents:** by integrating information retrieval methods and data mining methods, we can find major topics in a collection of documents (and the topic of each document).

Technologies Used – Confluence of Multiple Disciplines?

Data mining field needs techniques from several computer science fields. That's because:

- **Tremendous amount of data:**
  - Algorithms must be scalable to handle big data
- **High-dimensionality of data:**
  - Micro-array may have tens of thousands of dimensions (but sparse data)
- **Highly complex data:**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social and information networks
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- **New and sophisticated applications (in different fields)**

Outline

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - ➢ What Kinds of Applications Are Targeted?

Applications of Data Mining

- **Business intelligence:** technologies that provide historical, current and predictive view of business operations
  - **Capabilities like:** reporting, online analytical processing (OLAP), business performance management, competitive intelligence, benchmarking and predictive analysis. The knowledge gained is used in business decision making.
- **General Business:** Basket data (stock) analysis, targeted marketing, fraud detection
- **Web page search and analysis:** typically includes crawling, indexing and search tasks. The tasks need classification, clustering, PageRank, HITS and other methods and algorithms.
- **Collaborative analysis & recommender systems:** Suggest useful and interesting items based on the history of your interests, or the interests and opinions of the society (content based filtering vs. collaborative filtering).

Applications of Data Mining

- **Biological and medical data analysis:** classification, cluster analysis (microarray data analysis), biological sequence analysis (gene/protein analysis), biological network analysis
- **Social network data:** friend/community discovery, match making…
- **Software engineering**: finding bugs, performance increase, problem root causes
- **Building data mining software:** dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools), invisible data mining systems (places you don't think)