

Some data analytics elements

Genoveva Vargas Solar

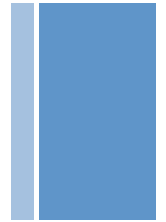
French Council of scientific research, LIG-LAFMIA, France

Genoveva.Vargas@imag.fr

<http://www.vargas-solar.com/teaching>

<http://www.vargas-solar.com>

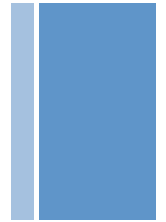
What to do with all these data?



- Aggregation and Statistics
 - Data warehouse and OLAP
- Indexing, Searching, and Querying
 - Keyword based search
 - Pattern matching (XML/RDF)
- Knowledge discovery
 - Data Mining
 - Statistical Modeling

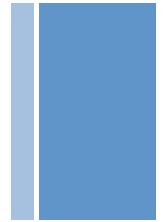
OLAP & data warehouses

Analysis of huge volumes of data



DOMAINE	APPLICATIONS
Super markets	Consumers behavior analysis Cross sales
Banks	Research on fraud contexts Credit preconditions identification
Insurance companies	Selection and pricing models Accident cause analysis
Airlines Automobile	Control quality Order prevision
Telecommunication	Pricing simulation

Data Warehouse



A huge amount of data

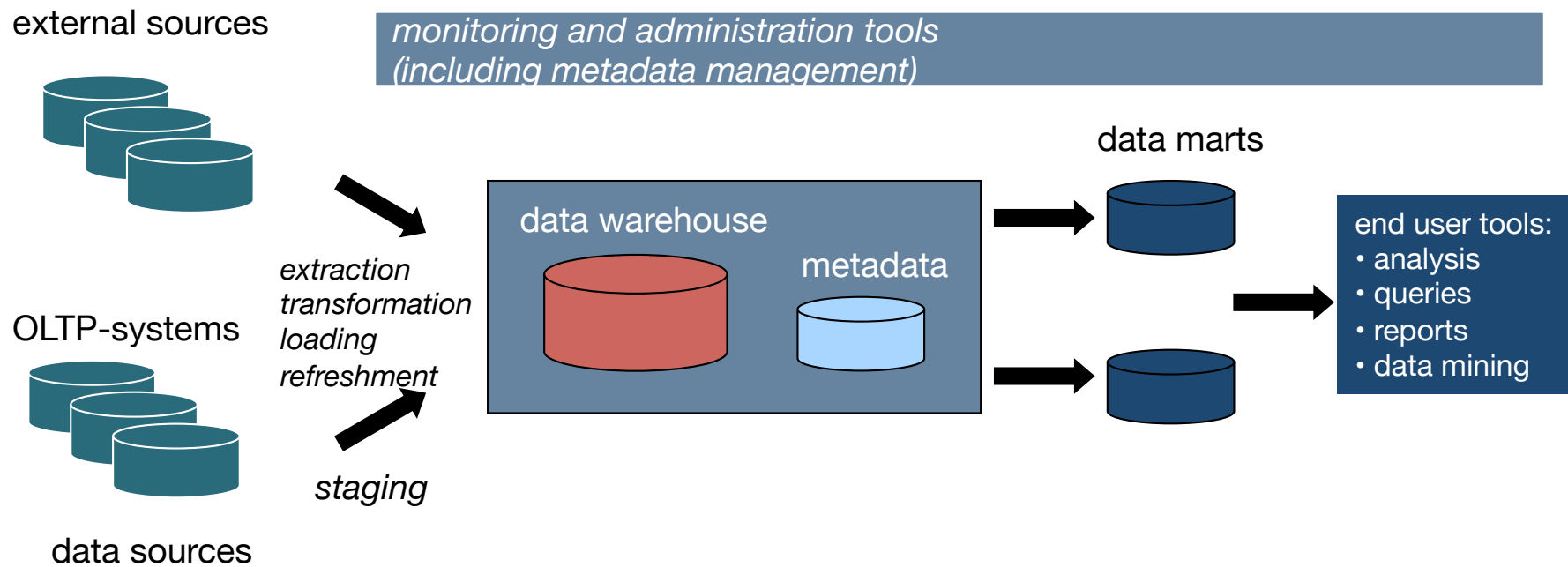
- selected
- integrated
- organized
- “historized”

for supporting decisional applications

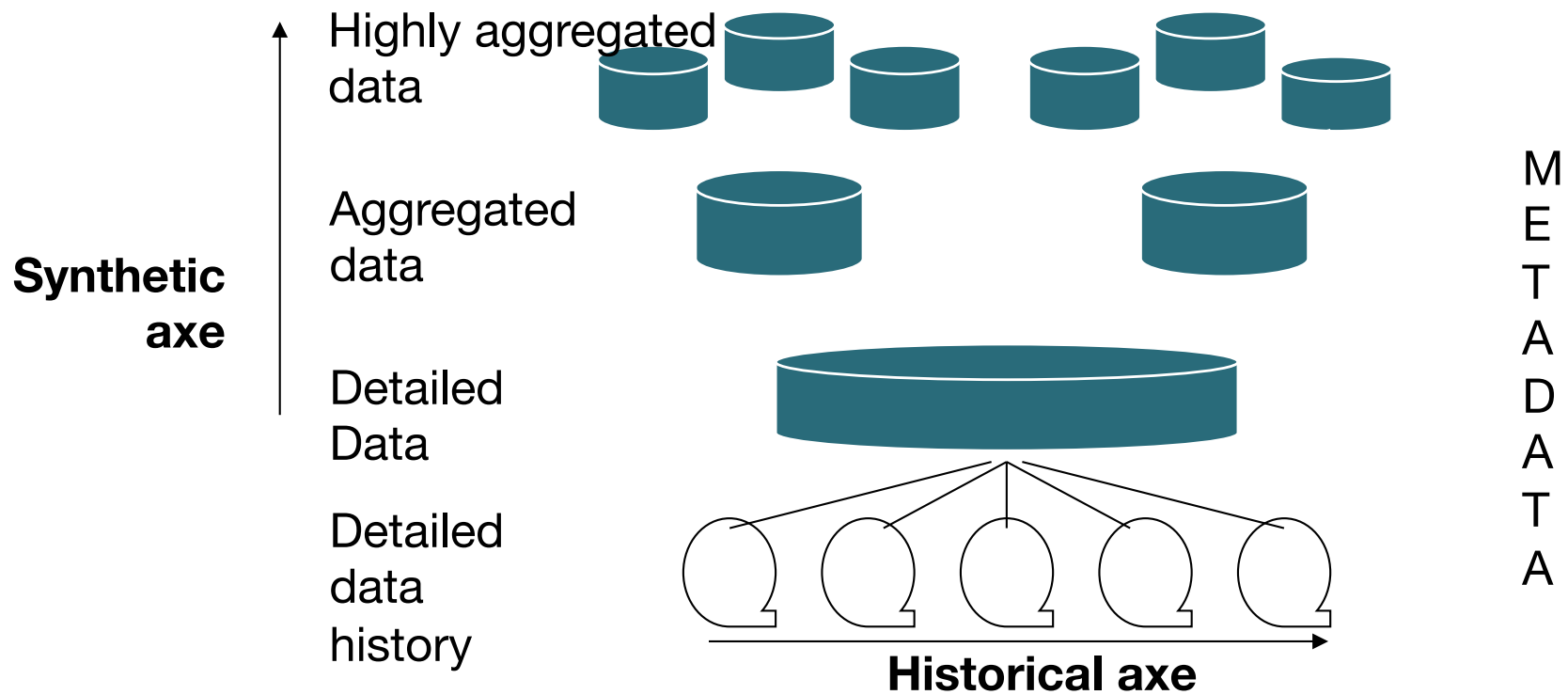
Differences with operational databases (*OLTP* vs. *OLAP*)

<i>Characteristics</i>	<i>Operational Databases</i>	<i>Data warehouses</i>
Data	Current	Historical
Use	Operational support for an enterprise	Support of analysis in an enterprise
Execution unit	Transaction	Query
Amount of managed data	Decenies	Millions
Access model	Read/Write	Read (mainly)
User type	Employee	Decision maker
Number of users	Mille	Cent

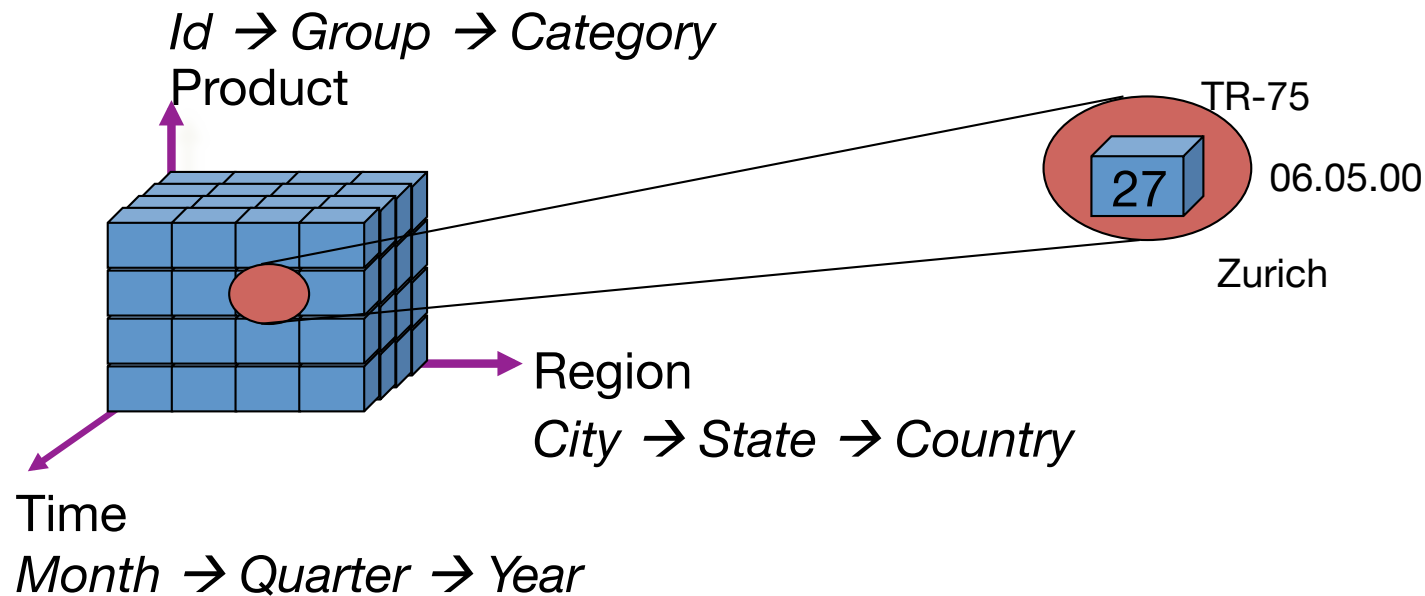
Architecture



Internal structure



Multidimensional data models



- Description of orthogonal dimensions
- Specification of measures

Dimensions

Product

Top

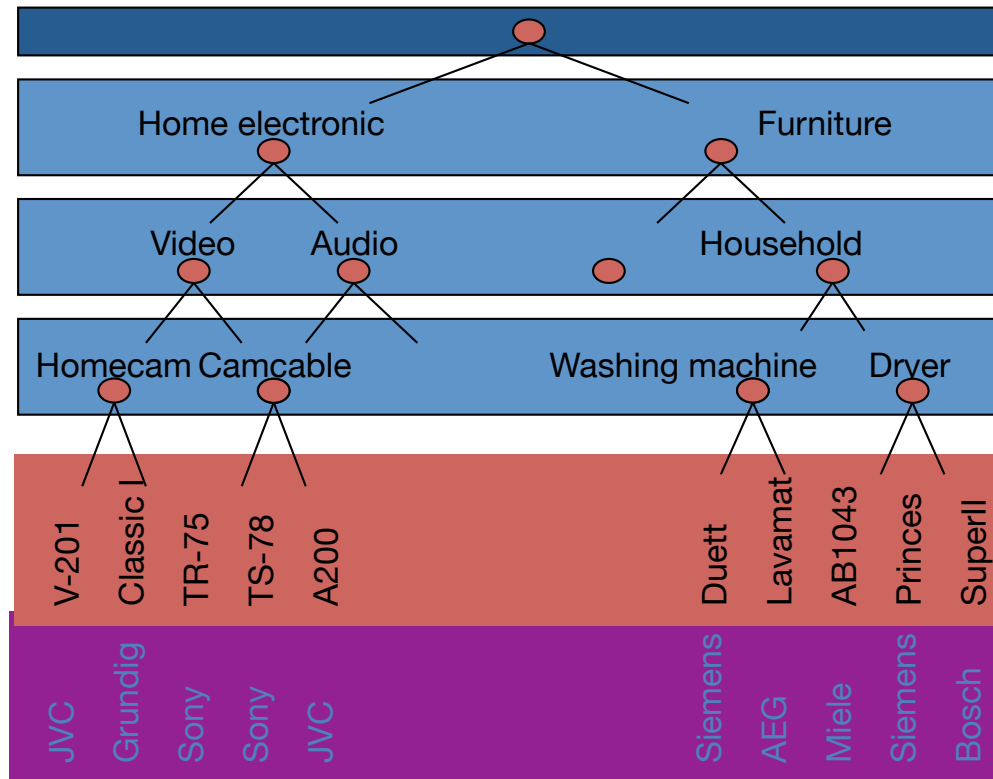
Branch

Group

Family

Article

Brand



Location

Top

Region

State

City

Shop

Time

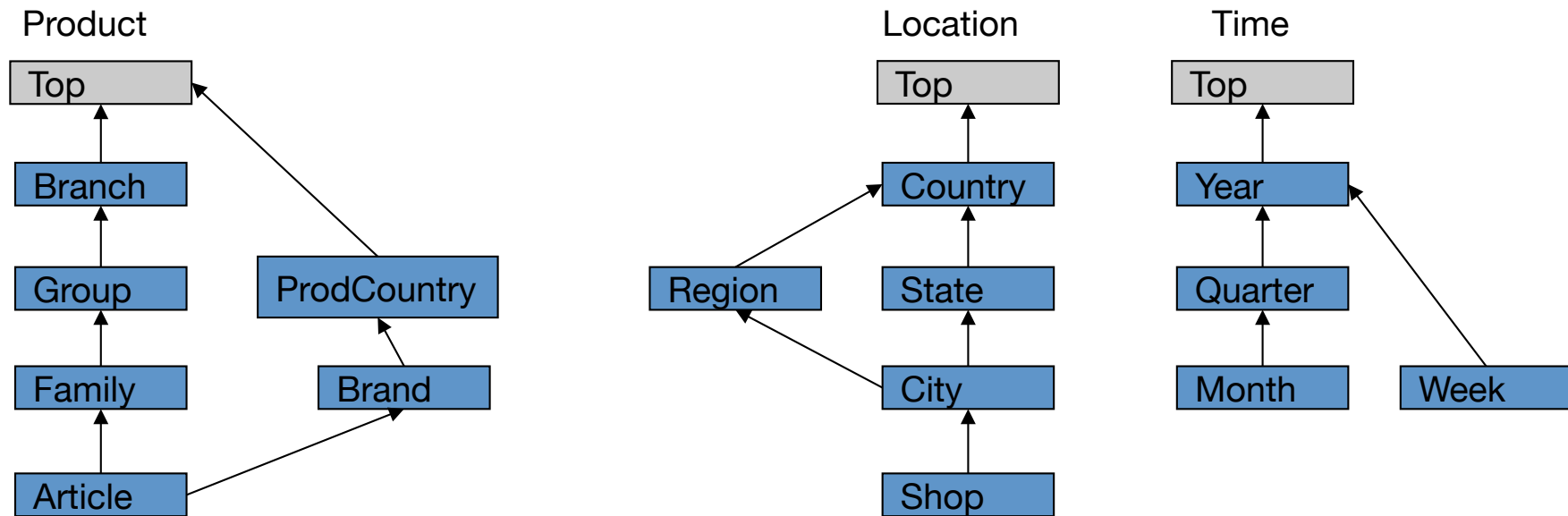
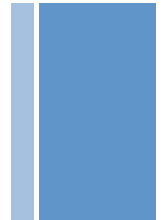
Top

Year

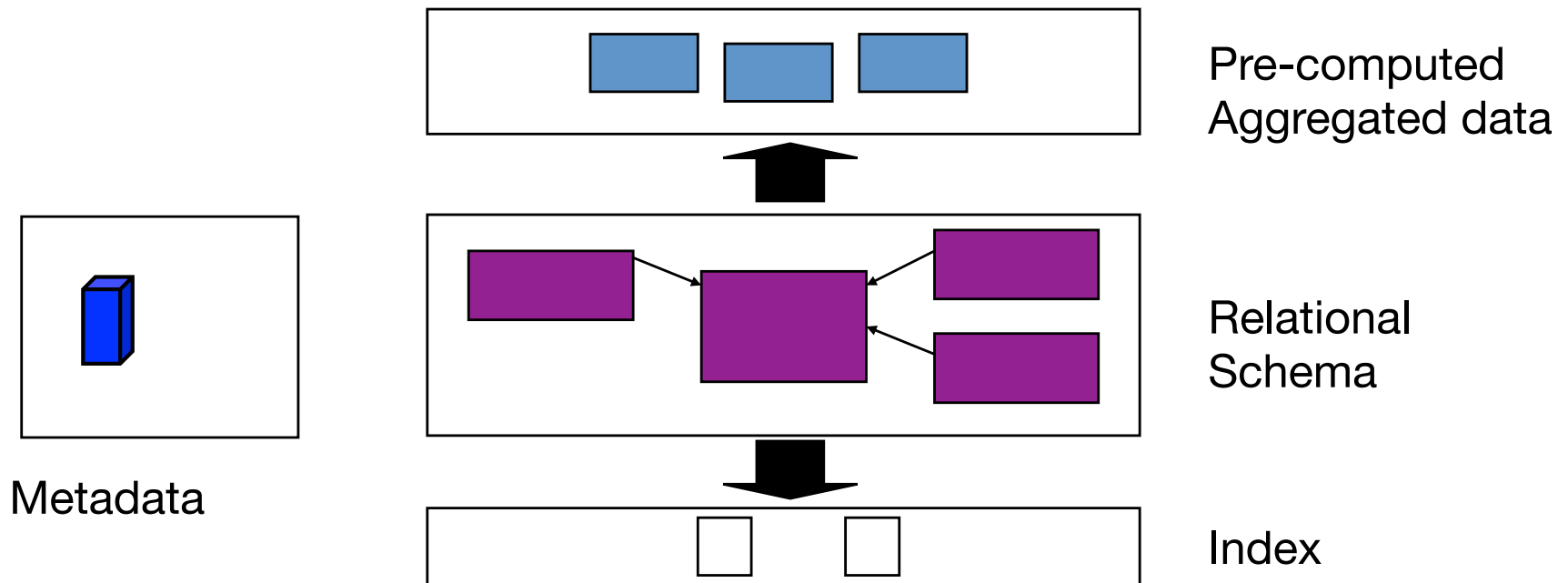
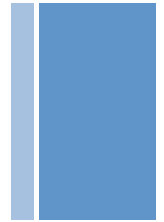
Quarter

Month

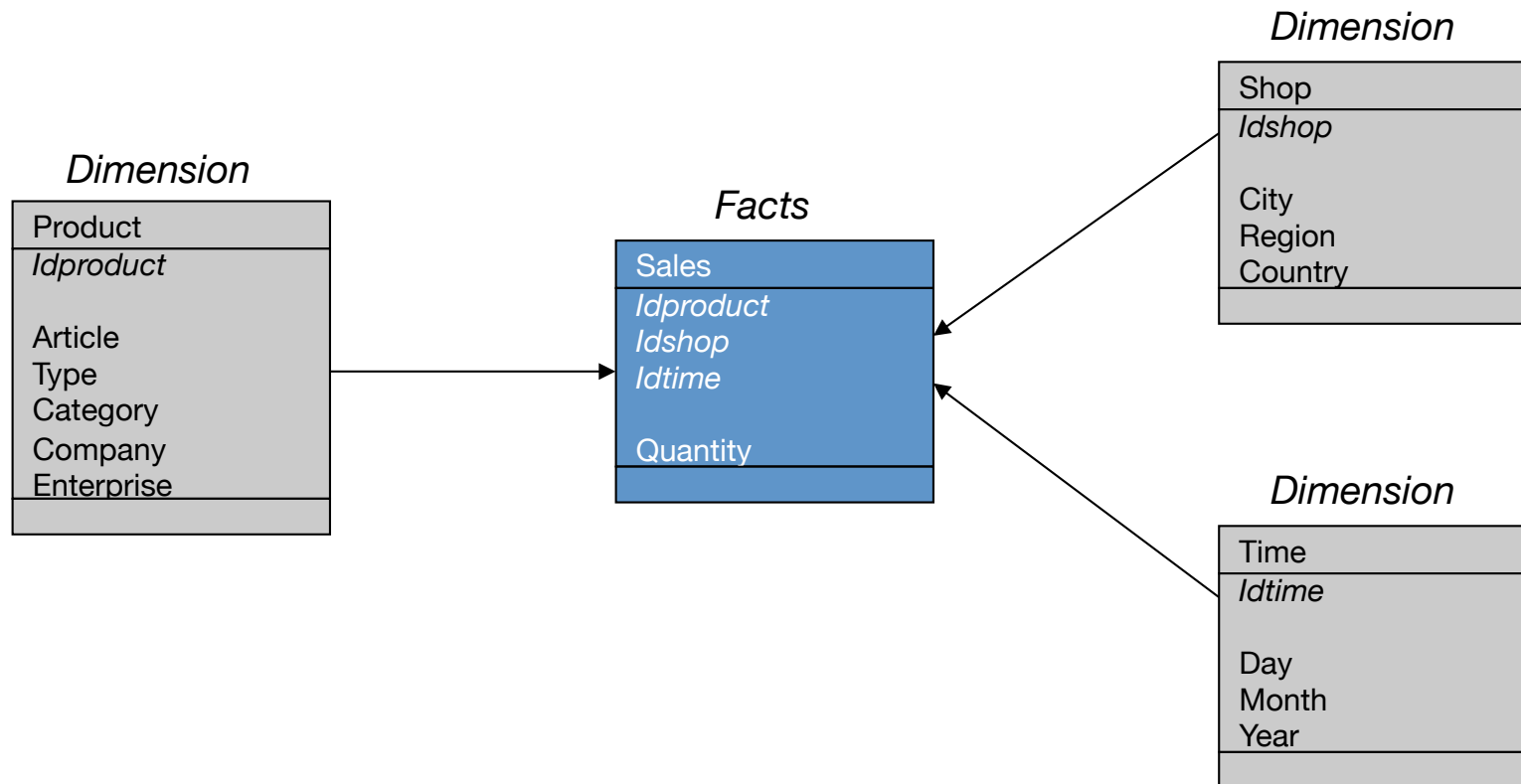
Dimension schemata



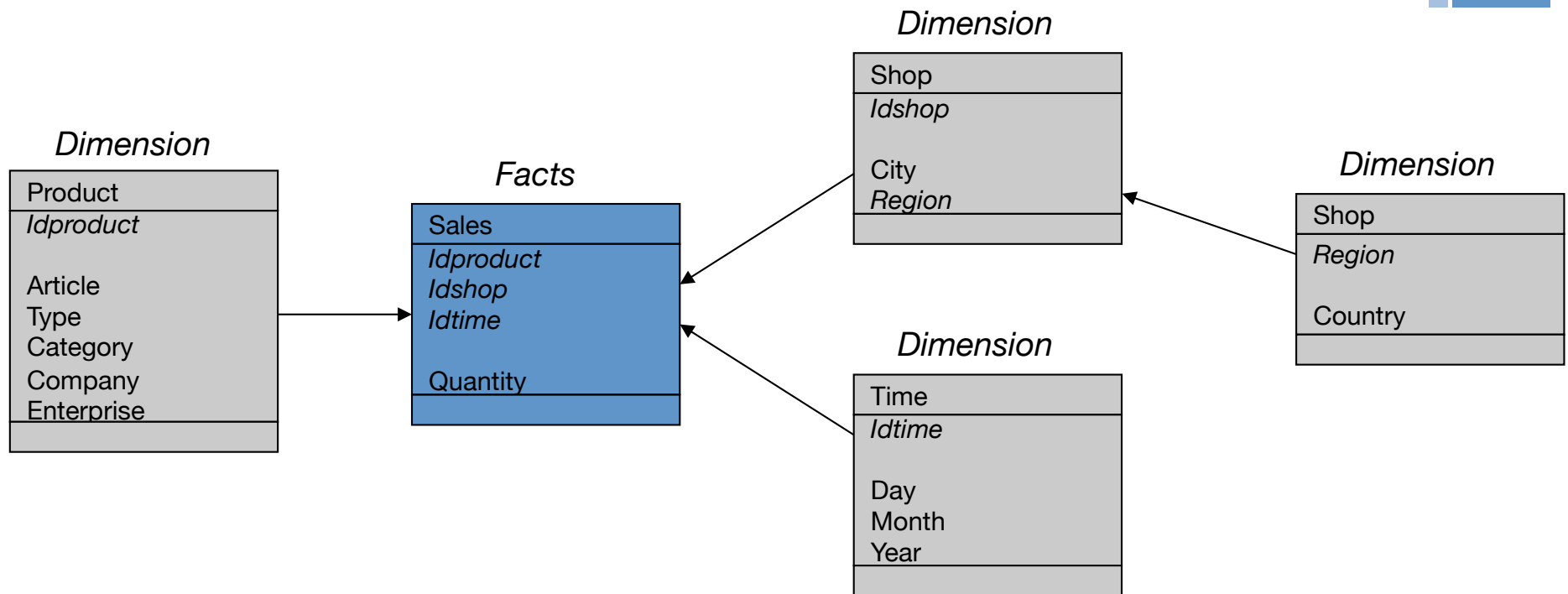
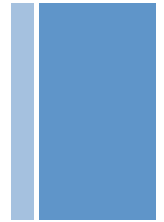
Relational data warehouse



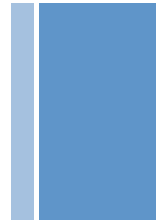
Star schema



Star schema



Pre-computed aggregated data

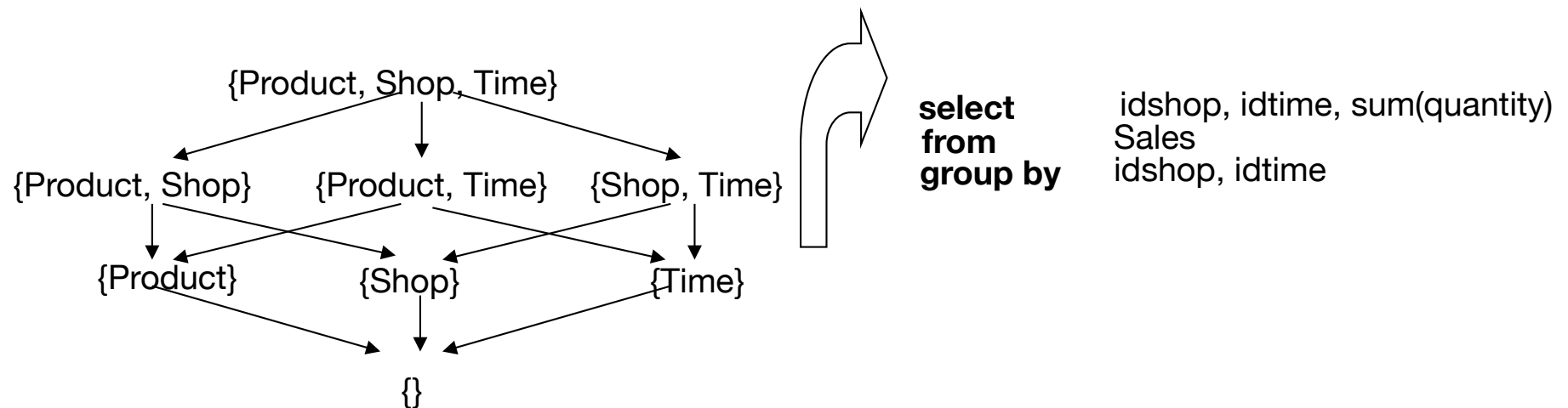


- Computed aggregated and stored in advance
- Potentially re-usable for computing other queries
- Represented by materialized views

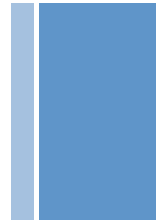
```
Create materialized view VPP(idproduit,total)
as
select      idproduit, sum(quantity) as total
from        Sales
group by    idproduit
```

Synthesis graph

- For a facts relation representing a k dimensional cube, there are 2^k views
- Heuristic selection of aggregated data to be materialized



Binary index (*bitmap*)

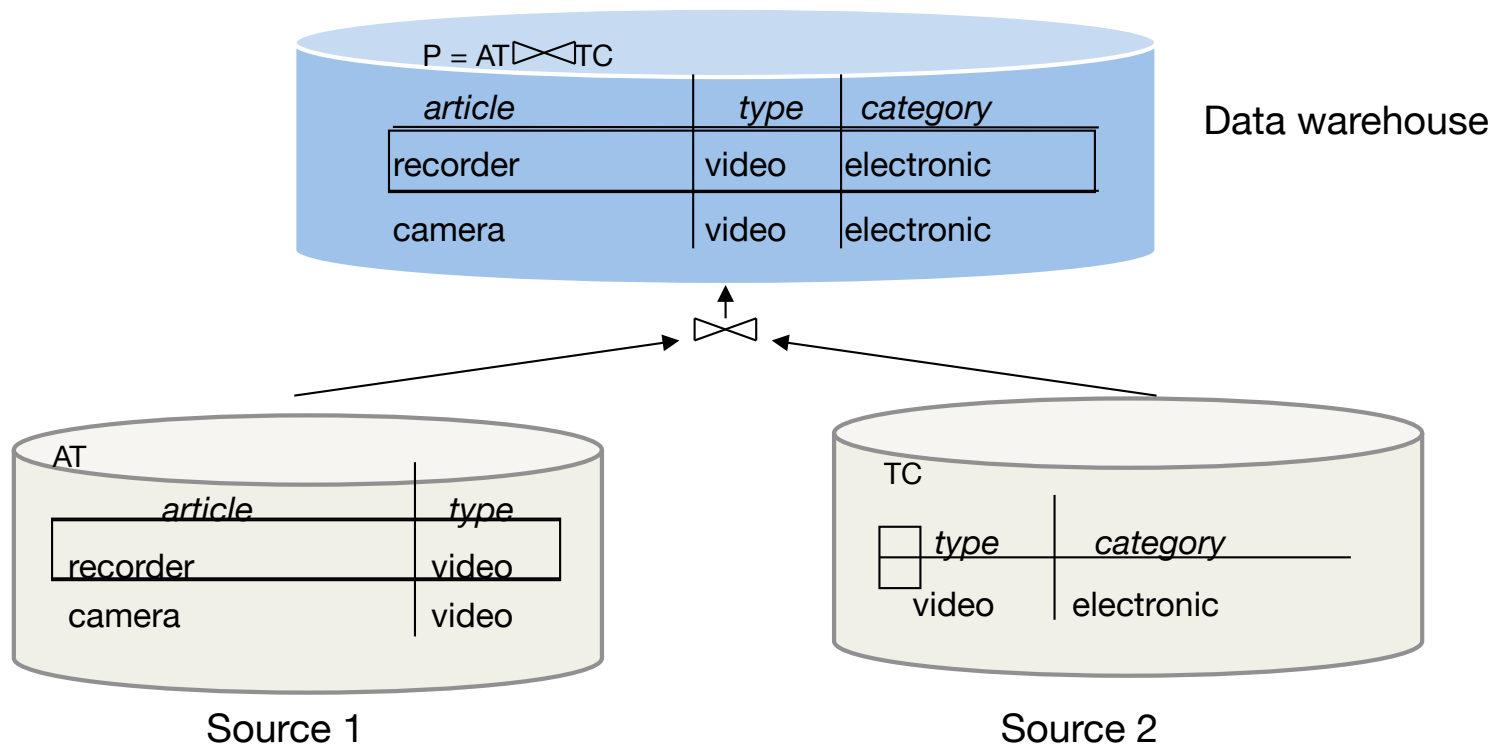


Sequence of bits where each bit specifies whether the associated register has a property

Category = electronics

Category	Bitmap
electronics	1
household	0
electronic	0
furniture	1
electronic	0
furniture	1
clothes	0

Construction example

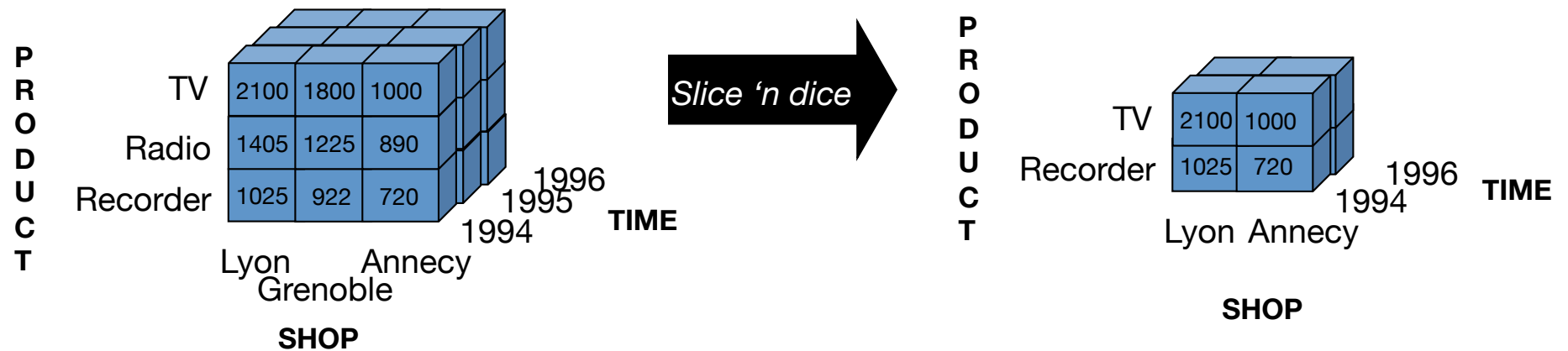
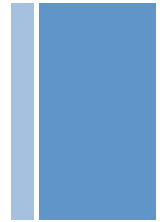


Construction issues

19

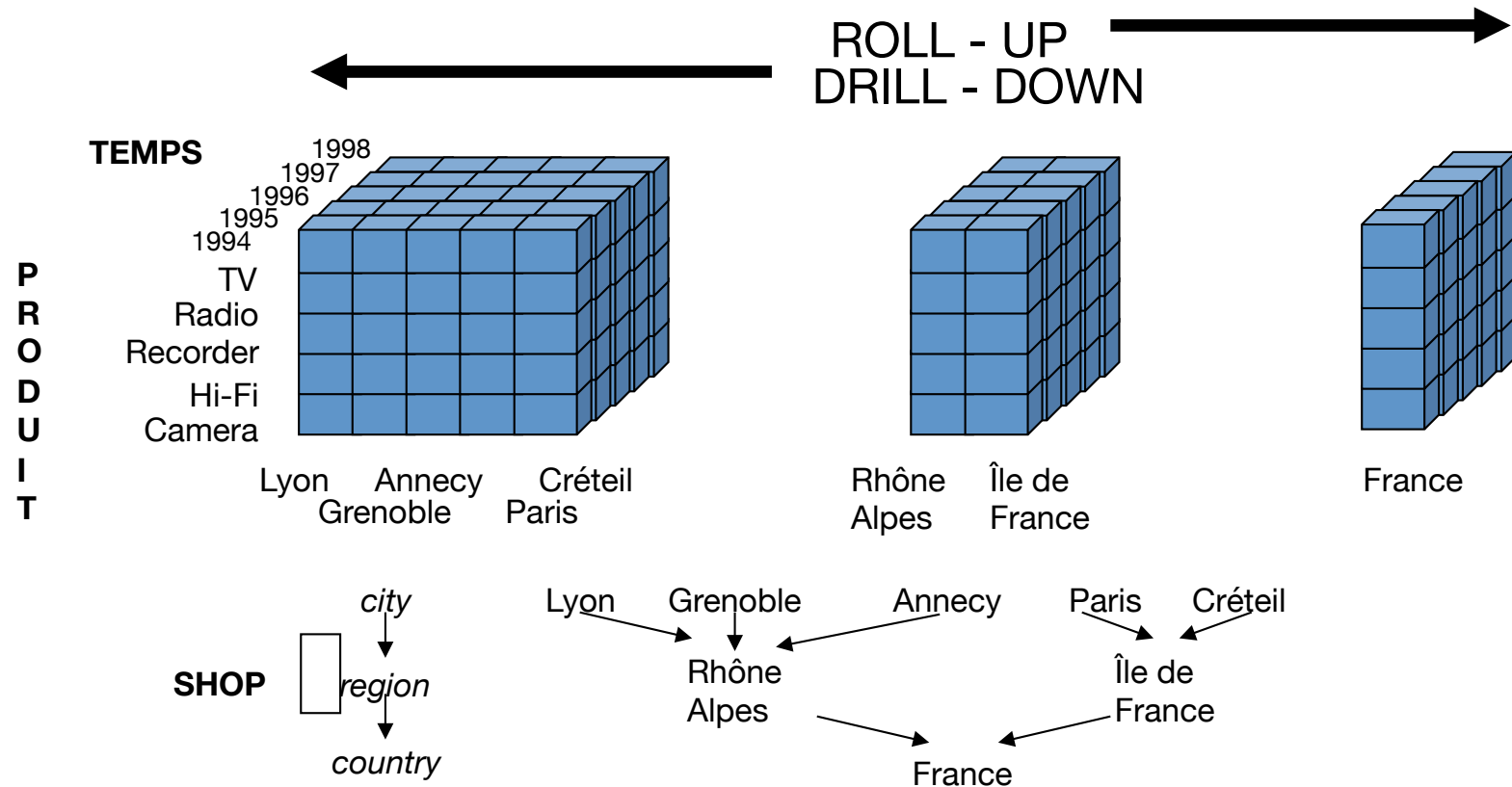
- Massif charge of data
 - Solutions: Massif charge of data
 - Solutions: Partitioned and sorted slots
- Parallel execution of input/output and computations
 - Partitioned and sorted slots
 - Parallel execution of input/output and computations

SLICE 'N DICE



Slice-dice product != radio, (SALES)
 shop!=Grenoble,
 time!=1995

ROLL UP, DRILL DOWN

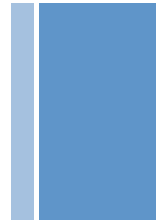


Final remarks



- Storage, indexation, interrogation
- Few updates
- Large volumes of data
- Analysis/synthesis queries without real time constraints

Technology



- Specialized DBMS: RedBrick (Informix), Oracle 8i (Oracle), DB2 (IBM), SQL Server (Microsoft)
- Schema specification: Warehouse Builder (Oracle), Visual Warehouse (IBM)
- Warehouse construction: Data Joiner (DB2), Extract (ETI), Data Transformation Services (Microsoft)
- Analysis tools: Express (Oracle), OLAP Services (Microsoft), MetaCube (Informix), Intelligent Miner (IBM)

What is Data Mining?

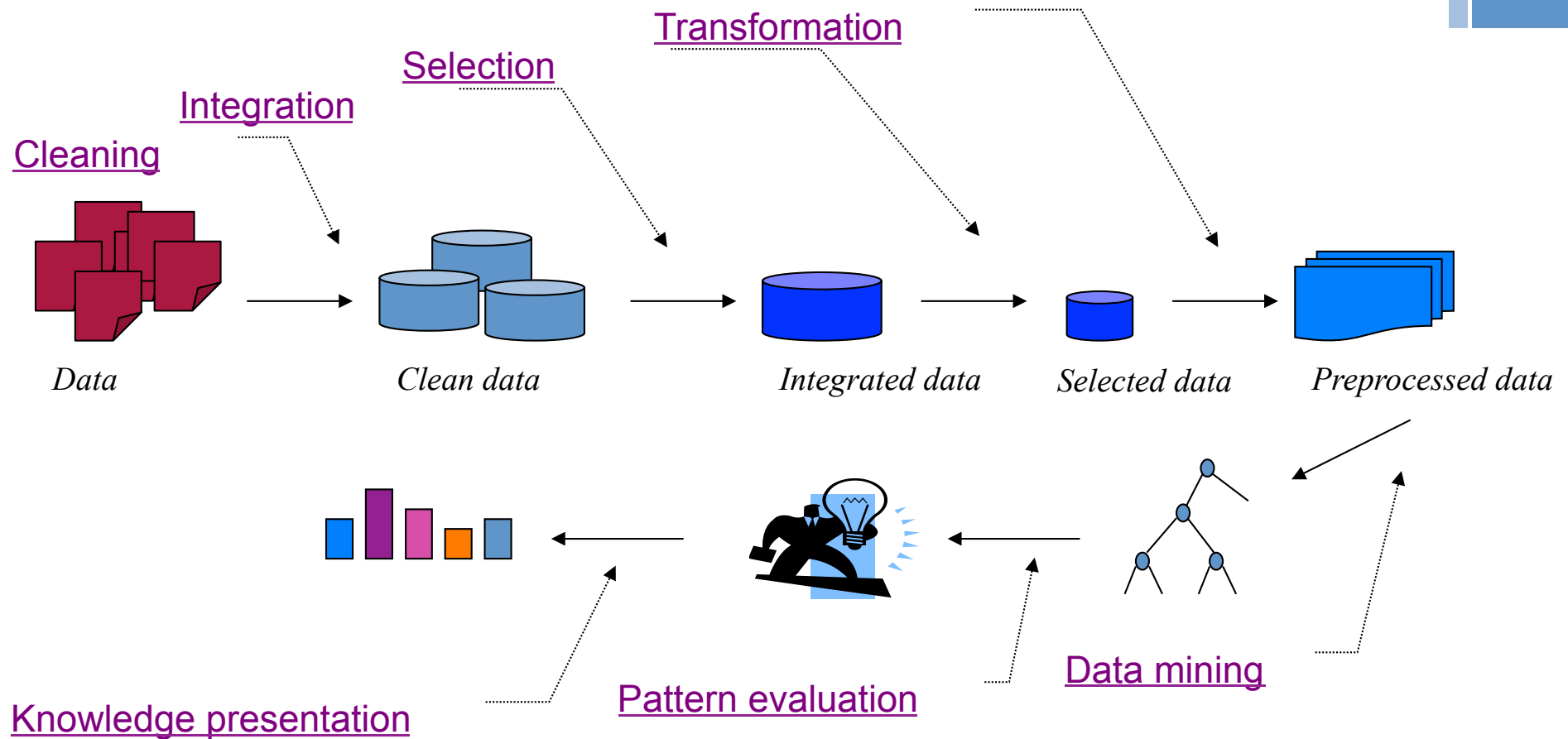
- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Data mining

Preliminary definitions

- Data mining why?
 - Huge amounts of data in electronic forms
 - Turn data into useful information and knowledge for broad applications
 - Market analysis, business management, decision support
- Data mining the process of discovering interesting knowledge from huge amounts of data
 - Patterns, associations, changes, anomalies and significant structures
 - Databases, data warehouses, information repositories

Discovery knowledge process



Simplified KDD discovery process

28

- Database/ data warehouse construction
 - Cleaning, integration, selection and transformation
- Iterative process: data mining
 - Data mining, pattern evaluation, knowledge representation

Major tasks in data mining

■ Description

- Describes the data set in a concise and summarized manner
- Presents interesting general properties of data

■ Prediction

- Constructs one or a set of models
- Performs inference on the data set
- Attempts to predict the behavior of new data sets

Data mining system tasks

- Class description
 - Concise and succinct summarization of a collection of data → characterization
 - Distinguishes it from others → comparison or discrimination
 - Aggregation: count, sum, avg
 - Data dispersion: variance, quartiles, etc.
 - Example: compare European and Asian sales of a company, identify important factors which discriminate the two classes
- Association
- Classification
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
 - Discovery of correlations or association relationships among a set of items
 - Expressed in the form a rule: attribute value conditions that occur frequently together in a given set of data
 - $X \rightarrow Y$: database tuples that satisfy X are likely to satisfy Y
 - Transaction data analysis for directed marketing, catalog design, etc.
- Classification
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
 - Analyze a set of training data (i.e., a set of objects whose class label is known)
 - Construct a model for each class based on the data features
 - A decision tree or classification rules are generated
 - Better understanding of each class
 - Classification of future data
 - Diseases classification to help to predict the kind of diseases based on the symptoms of patients
 - Classification methods proposed in machine learning, statistics, database, neural networks, rough sets.
 - Customer segmentation, business modelling and credit analysis
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
 - Predict possible values of some missing data or the value distribution of certain attributes in a set of objects
 - Find a set of relevant attributes to the attribute of interest (e.g., by some statistical analysis)
 - Predict the value distribution based on the set of data similar to the selected objects
 - An employee's potential salary can be predicted based on the salary distribution of similar employees in a company
 - Regression analysis, generalized linear models, correlation analysis, decision trees used in quality prediction
 - Genetic algorithms and neural network models also popular
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
- Clustering
 - Identify clusters embedded in the data
 - Cluster is a collection of data objects similar to one another
 - Similarity expressed by distance functions specified by experts
 - Good cluster method produces high quality clusters to ensure that
 - inter cluster similarity is low
 - intra cluster similarity is high
 - Cluster the houses of Cholula according to their house category, floor area and geographical locations
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
- Clustering
- Time series analysis
 - Analyze large set of time-series data to find regularities and interesting characteristics
 - Search for similar sequences, sequential patterns, periodicities, trends and derivations
 - Predict the trend of the stock values for a company based on its stock history, business situation, competitors' performance and current market

Data mining challenges

- Handling of different types of data
 - Knowledge discovery system should perform efficient and effective data mining on different kinds of data
 - Relational data, complex data types (e.g. structured data, complex data objects, hypertext, multimedia, spatial and temporal, transaction, legacy data)
 - Unrealistic for one single system
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
 - Running times predictable and acceptable in large databases
 - Algorithms with exponential or medium order polynomial complexity are not practical
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

38

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
 - Discovered knowledge must
 - Portray the contents of a database accurately:
 - Useful for certain applications
 - Uncertainty measures (approximate or quantitative rules)
 - Noise and exceptional data: statistical, analytical and simulative models and tools
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
 - Different kinds of knowledge can be discovered
 - Examine from different views and present in different forms
 - Express data mining requests and discovered knowledge in high level languages or graphical interfaces
 - Knowledge representation techniques
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
 - Difficult to predict what can be discovered
 - High level data mining query should be treated as a probe disclosing interesting traces to be further explored
 - Interactive discovery: refine queries, dynamically change data focusing, progressively deepen a data mining process, flexibly view data and data mining results at multiple abstraction levels and different angles
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
 - Mine distributed and heterogeneous (structure, format, semantic)
 - Disclose high level data regularities in heterogeneous databases hardly discovered by query systems
 - Huge size, wide distribution and computational complexity of data mining methods → parallel and distributed algorithms
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security
 - Data viewed from different angles and abstraction levels → threaten security and privacy
 - When is it invasive and how to solve it?
 - Conflicting goals
 - Data security protection vs. Interactive data mining of multiple level knowledge from different angles

Roadmap

43

- ✓ Preliminary definitions
- Data mining approaches
 - Algorithms “pleyade”
 - Data mining from a database perspective
- Future research directions

Data mining approaches

- Needs the integration of approaches from multiple disciplines
 - Database systems & data warehousing
 - Statistics, machine learning, data visualization, information retrieval, high performance computing
 - Neural networks, pattern recognition, spatial data analysis, image databases, spatial processing, probabilistic graph theory and inductive logic programming
- Large set of data mining methods
 - Machine learning: classification and induction problems
 - Neural networks: classification, prediction, clustering analysis tasks
 - Scalability and efficiency
- Data structures, indexing, data accessing techniques

Data analysis vs. data mining

- Data analysis
 - Assumption driven
 - Hypothesis is formed and validated against data
- Data mining
 - Discovery-driven
 - Patterns are automatically extracted from data
 - Substantial search efforts
 - High performance computing
 - Parallel, distributed and incremental data mining methods
 - Parallel computer architectures

Classifying data mining techniques

46

- What kinds of databases to work on
 - DMS classified according to the kinds of database on which data mining is performed
 - Relational, transactional, OO, deductive, spatial, temporal, multimedia, heterogeneous, active, legacy, Internet-information base
- What kind of knowledge to be mined
 - Kind of knowledge
 - Association, characteristic, classification, discriminant rules, clustering evolution, deviation analysis
 - Abstraction level of the discovered knowledge
 - Generalized knowledge, primitive-level knowledge, multiple-level knowledge
- What kind of techniques to be utilized
 - Driven methods
 - Autonomous, data-driven, query-driven, interactive
 - Data mining approach
 - Generalization-based, pattern-based, statistics and mathematical theories, integrated approaches

Data mining algorithms

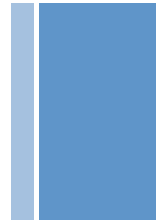
- Prediction
 - Forecast the value of a variable based on the previous knowledge of that variable
 - Algorithms
 - Classification
 - Prediction
 - Trend detection
 - Disasters prediction like floods, earthquake, volcanoes eruptions
- Description
 - Discover knowledge contained in a data collection → decision making
 - Algorithms
 - Clustering
 - Association rules
 - Business and scientific areas

Roadmap

48

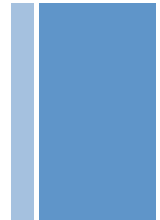
- ✓ Preliminary definitions
- Data mining approaches
 - ✓ Algorithms “pleyade”
 - Data mining from a database perspective
- Future research directions

Knowledge Discovery in Multi-Databases



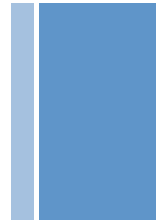
- A major challenge in multi-DBs: semantic heterogeneity.
 - Multi-databases: low level heterogeneity but high-level regularity (e.g., school grading systems)
 - The role of generalization-based knowledge discovery: raise concept levels and ease semantic heterogeneity
- Knowledge discovery and query transformation
 - Not only an exchangeable “export schema” but also a common high level “language” (“vocabulary”)
 - Each local database system provides two-way transformation between low and high level data
 - The transformation contributes to high level knowledge exchange (for KDD) and query/result interpretation (for interoperability)

Data Mining in WWW



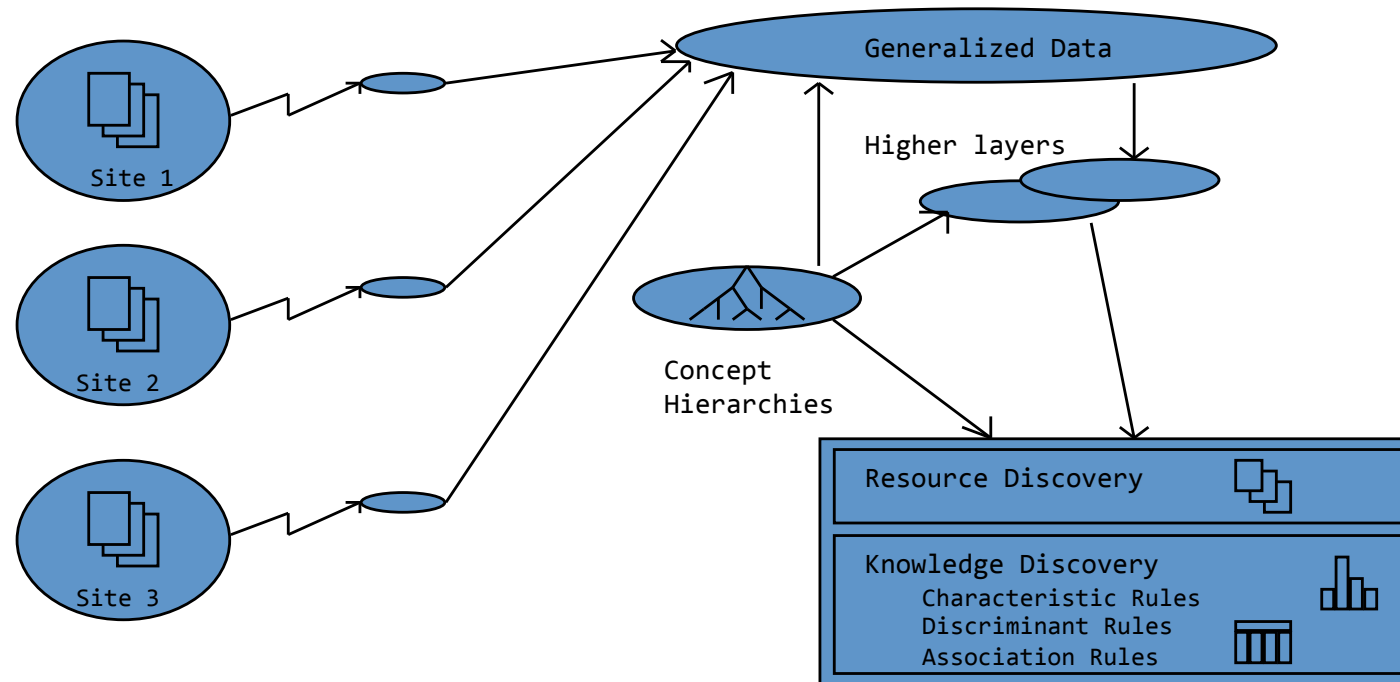
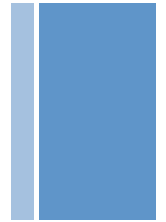
- Internet (WWW): A world-wide distributed information system based on hypertext (semi-structured, HTML, Java)
- Structured Web hypothesis (Etzioni' 96) and three subtasks in Web mining:
 - **Resource discovery:** locating documents and services on the Web.
 - WebCrawler, Alta Vista --- scan millions of Web documents and create index of words (too many irrelevant, outdated responses).
 - MetaCrawler --- mines robot-created indices.
 - Future: automatic text categorization and construction of Web directories.
 - **Information extraction:** Automatic information extraction from newly discovered Web sources
 - Harvest: uses a model of semi-structured documents.
 - Internet Learning Agent and Shopbot : learn about Web services.
 - **Generalization:** Uncover general patterns at individual sites and across multiple sites
 - Relying on feedbacks from multi-users to solve the labeling problem.

A Multi-Layered Database Model for Web Mining



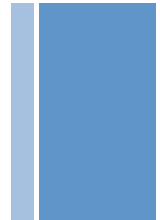
- A multiple layered database model based on semi-structured data hypothesis
- Layer-0: An unstructured, massive, primitive, diverse global information-base
- Layer-1: A relatively structured, descriptor-like, massive, distributed database by data analysis, transformation and generalization techniques
 - Tools to be developed for descriptor extraction
- Higher-layers: Further generalization to form progressively smaller, better structured, and less remote databases for efficient browsing, retrieval, and information discovery

Internet Information Mining: A WebMiner Proposal



Jiawei Han, Simon Fraser University, Canada

Data Mining Applications



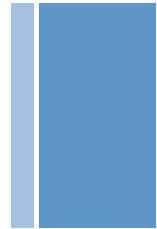
- Numerous data mining applications.
 - Querying database knowledge
 - Multi-level data browsing
 - Performance prediction
 - Market analysis
 - Database design and query optimization
 - Intelligent query answering.
- Intelligent Query Answering
 - Extended data model: Schemas, hierarchies, multi-layered databases, generalized relations/cubes, data mining tools.
 - Intelligent answering: Multi-level summaries & statistics, neighborhood info, ‘roll-up’ & ‘drill-down’ facilities.
 - “What kind of houses can be bought with \$500K in L.A.?” summary, width extension, height extension, etc.

Systems for Data Warehousing and Data Mining



- Systems for Data Warehousing
 - Arbor Software: Essbase
 - Oracle (IRI): Express
 - Cognos: PowerPlay
 - Redbrick Systems: Redbrick Warehouse
 - Microstrategy: DSS/Server
- Systems or Research Prototypes for Data Mining
 - IBM: QUEST (Intelligent Miner)
 - Information Discovery Inc.: Data Mining Suite
 - Silicon Graphics: MineSet
 - Integral Solutions Ltd.: Clementine
 - SFU (DBTech): DBMiner
 - Rutgers: DataMine, George Mason: INLEN, GMD: Explora

Major Knowledge Discovery Methods in KDD Systems



- Database-oriented approach: Quest, DBMiner, etc.
- OLAP-based (data warehousing) approach: Informix-MetaCube, Redbricks, Essbase, etc.
- Machine learning: AQ15, ID3/C4.5, Cobweb, etc.
- Knowledge representation & reasoning: e.g., IMACS.
- Rough sets, fuzzy sets: Datalogic/R, 49er, etc.
- Statistical approaches, e.g., KnowledgeSeeker.
- Neural network approach, e.g., NeuroRule (Lu et al. '95).
- Inductive logic programming: Muggleton & Raedt '94, etc.
- Deductive DB integration: KnowlegeMiner (Shen et al. '96).
- Visualization approach: VisDB(Keim, Kriegel, et al. '94).
- Multi-strategy mining: INLEN, KDW+, Explora, etc.

Roadmap

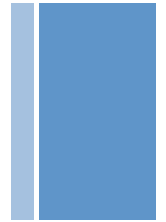
56

- ✓ Preliminary definitions
- ✓ Data mining approaches
- Future research directions

- Data warehousing: integrating data from multiple sources into large warehouses and support on-line analytical processing and business decision making
 - Data mining (knowledge discovery in databases): Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
- Necessity: **Data explosion problem**--- automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases

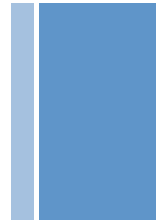
We are drowning in data, but starving for knowledge!

Data Mining: Classification Schemes



- Different views, different classifications:
 - Kinds of knowledge to be discovered,
 - Kinds of databases to be mined, and
 - Kinds of techniques adopted.
- Knowledge to be mined:
 - Summarization (characterization), comparison, association, classification, clustering, trend, deviation and pattern analysis, etc.
 - Mining knowledge at different abstraction levels:
 - primitive level, high level, multiple-level, etc.
- Databases to be mined:
 - Relational, transactional, object-oriented, object-relational, active, spatial, time-related, textual, multi-media, heterogeneous, legacy, etc.
- Techniques adopted:
 - Database-oriented, machine learning, neural network, statistics, visualization, etc.

Mining Different Kinds of Knowledge in Large Databases



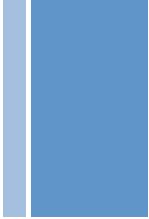
- **Characterization:** Generalize, summarize, and possibly contrast data characteristics, e.g., grads/undergrads in CS.
- **Association:** Rules like "buys(x, milk) \rightarrow buys(x, bread)".
- **Classification:** Classify data based on the values in a classifying attribute, e.g., classify cars based on gas mileage.
- **Clustering:** data to form new classes, e.g., cluster houses to find distribution patterns.
- **Trend and deviation analysis:** Find and characterize evolution trend, sequential patterns, similar sequences, and deviation data, e.g., stock analysis.
- **Pattern-directed analysis:** Find and characterize user-specified patterns in large databases.

Conclusions

- **Data mining:** A rich, promising, young field with broad applications and many challenging research issues.
- **Recent progress:** Database-oriented, efficient data mining methods in relational and transaction DBs.
- **Tasks:** Characterization, association, classification, clustering, sequence and pattern analysis, prediction, and many other tasks.
- **Domains:** Data mining in extended-relational, transaction, object-oriented, spatial, temporal, document, multimedia, heterogeneous, and legacy databases, and WWW.
- **Technology integration:**
 - Database, data mining, & data warehousing technologies.
 - Other fields: machine learning, statistics, neural network, information theory, knowledge representation, etc.

Statistics

Random Sample and Statistics



5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Table 1.2: Iris Dataset: sepal length

- *Population*: is used to refer to the set or universe of all entities under study
- However, looking at the entire population may not be feasible, or may be too expensive
- Instead, we draw a random sample from the population, and compute appropriate *statistics* from the sample, that give estimates of the corresponding population parameters of interest

Statistic

- Let S_i denote the random variable corresponding to data point x_i , then a *statistic* $\hat{\theta}$ is a function $\hat{\theta} : (S_1, S_2, \dots, S_n) \rightarrow \mathbb{R}$
- If we use the value of a statistic to estimate a population parameter, this value is called a *point estimate of the parameter*, and the statistic is called as an *estimator of the parameter*

Empirical Cumulative Distribution Function

$$\hat{F}(x) = \frac{\sum_{i=1}^n I(S_i \leq x)}{n}$$

Where

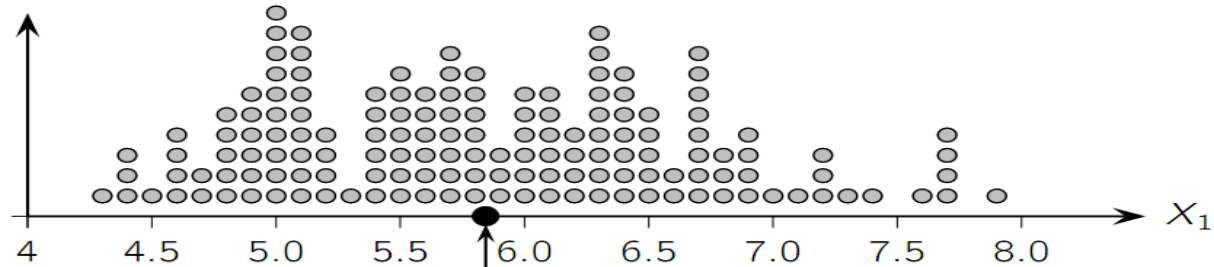
$$I(S_i \leq x) = \begin{cases} 1 & \text{if } S_i \leq x \\ 0 & \text{if } S_i > x \end{cases}$$

Inverse Cumulative Distribution Function

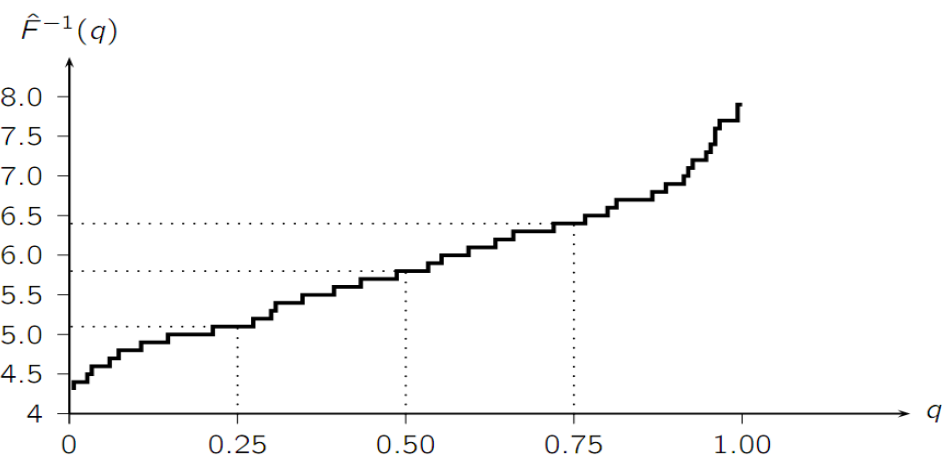
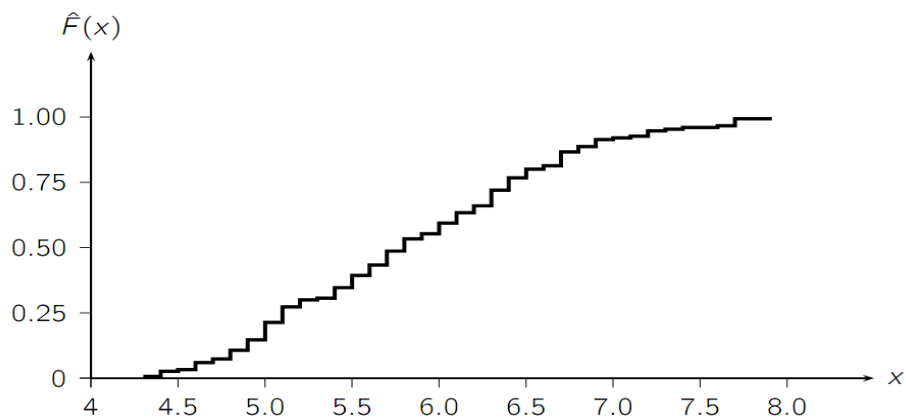
$$F^{-1}(q) = \min\{x : F(x) > q\} \quad \text{for } q \in [0, 1]$$

Exan

Frequency



$$\hat{\mu} = 5.843$$



Measures of Central Tendency (Mean)

Population Mean:

$$\mu = E[X] = \sum_x x f(x)$$

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Sample Mean (Unbiased not robust):

$$\hat{\mu} = \sum_x x \hat{f}(x) = \sum_x x \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n S_i}{n}$$

$$E[\hat{\mu}] = E \left[\frac{\sum_{i=1}^n S_i}{n} \right] = \frac{1}{n} \sum_{i=1}^n E[S_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

Measures of Central Tendency (Median)

Population Median:

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

or

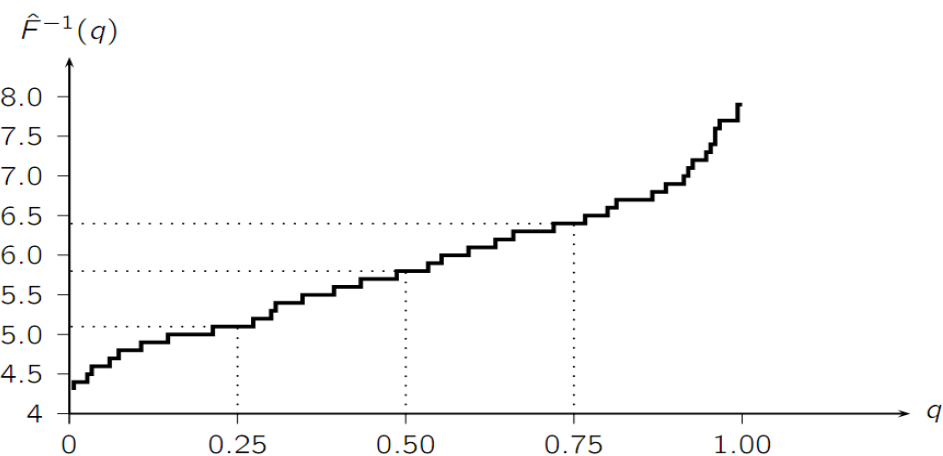
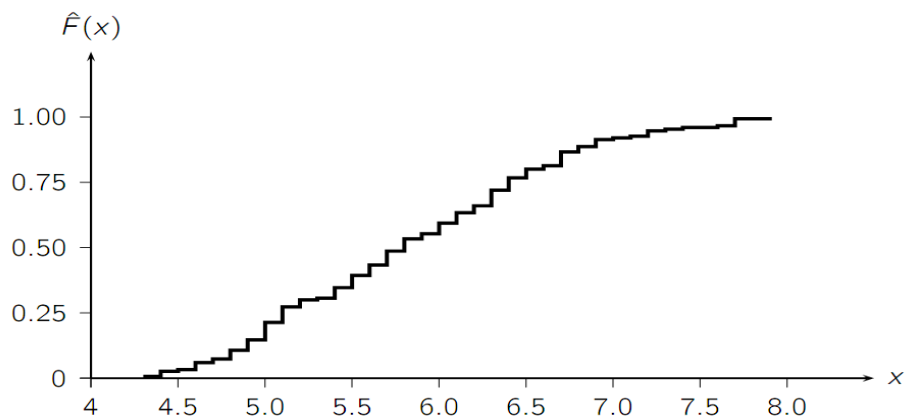
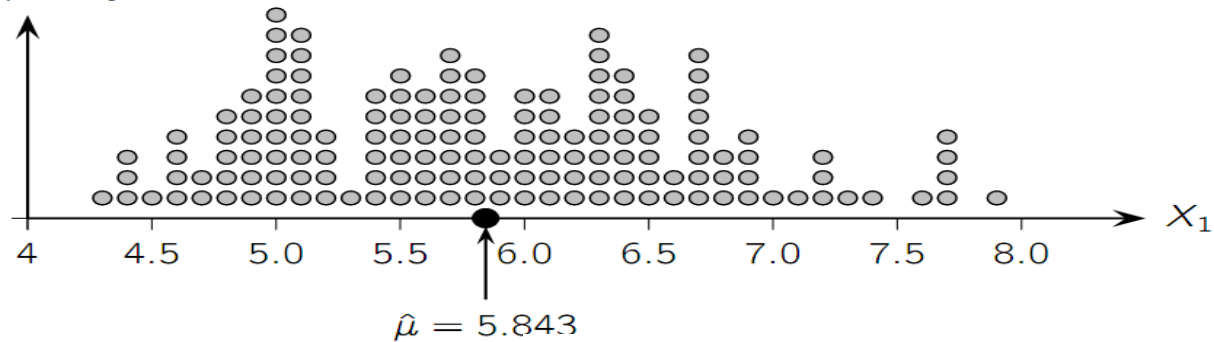
$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

Sample Median:

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

Exan

Frequency



Measures of Dispersion (Range)

Range:

$$r = \max_x \{x\} - \min_x \{x\}$$

Sample Range:

$$\hat{r} = \max_i \{S_i\} - \min_i \{S_i\} = \max_i \{x_i\} - \min_i \{x_i\}$$

- ❑ Not robust, sensitive to extreme values

Measures of Dispersion (Inter-Quartile Range)

Inter-Quartile Range (IQR):

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

Sample IQR:

$$\widehat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

- ▣ More robust

Measures of Dispersion (Variance and Standard Deviation)

Variance:

$$\text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

$$\begin{aligned} \sigma^2 = \text{var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Measures of Dispersion

(Variance and Standard Deviation)

Variance:

$$\text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

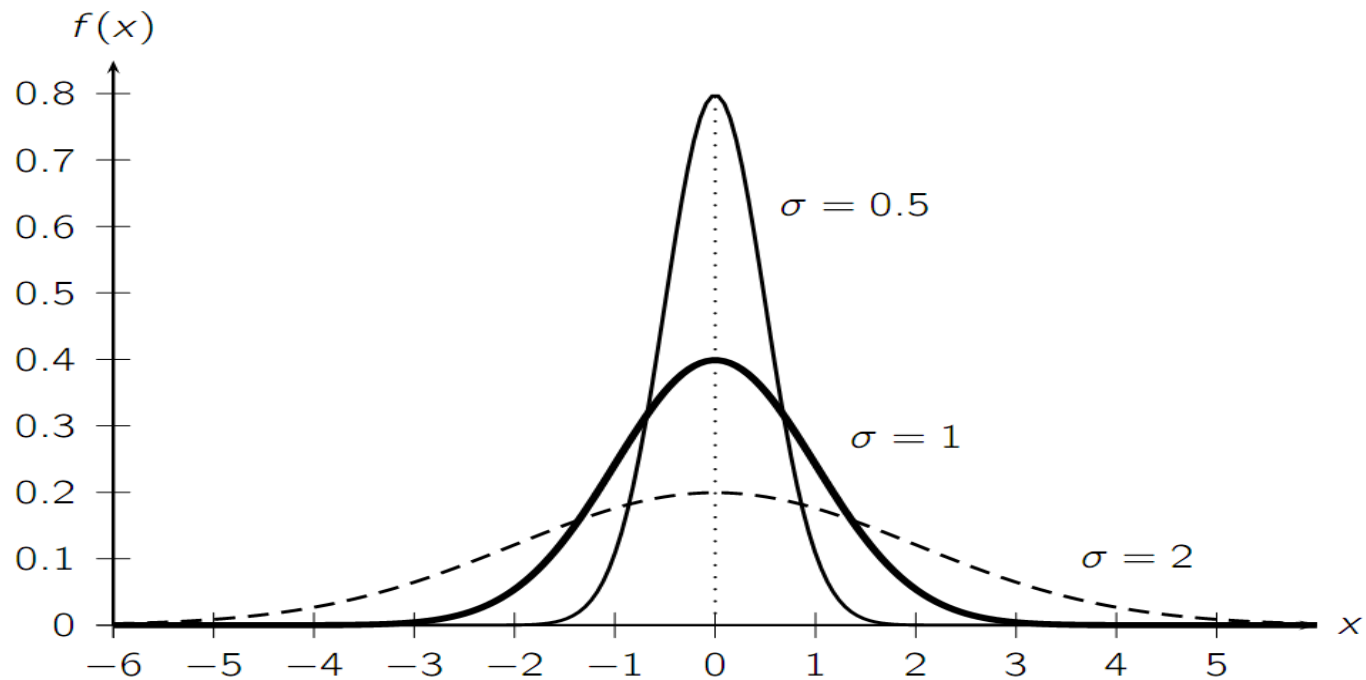
$$\begin{aligned} \sigma^2 = \text{var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Sample Variance & Standard Deviation:

$$\hat{\sigma}^2 = \sum_x (x - \hat{\mu})^2 \hat{f}(x) = \sum_x (x - \hat{\mu})^2 \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n (S_i - \hat{\mu})^2}{n}$$

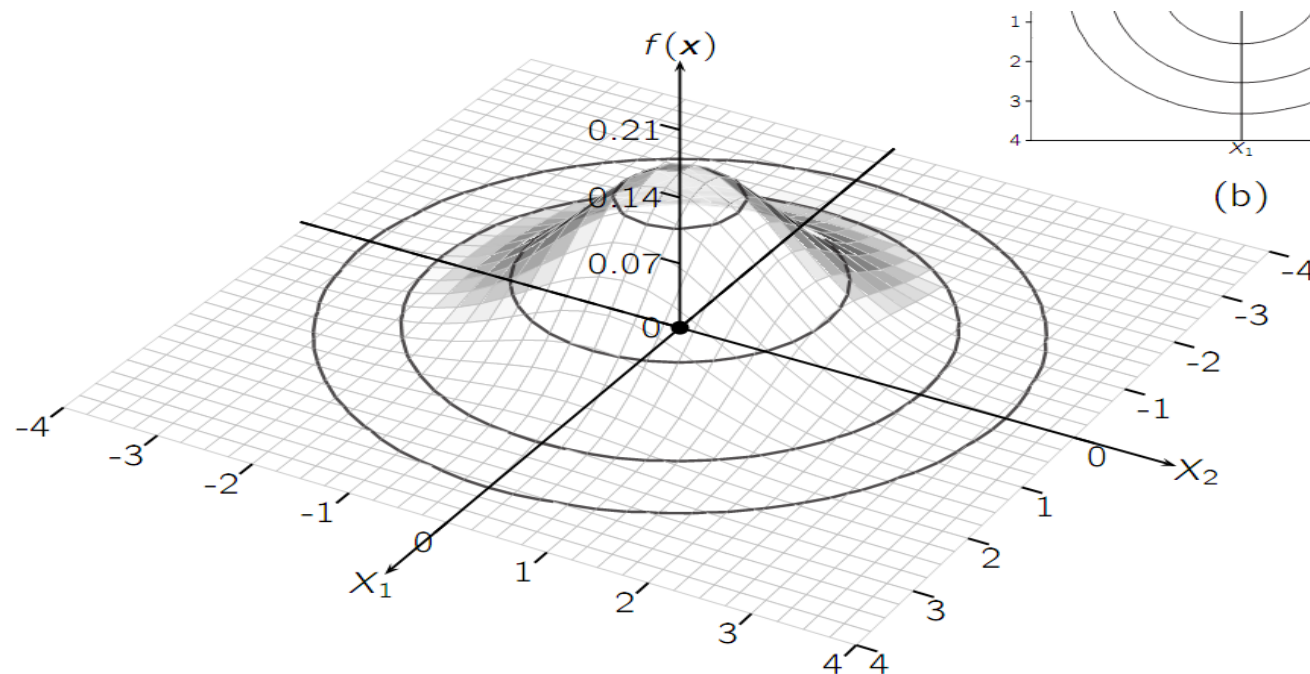
Univariate Normal Distribution

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$



Multivariate Normal Distribution

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp \left\{ -\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2} \right\}$$





Thanks **Merci**

Gracias



Contact: Genoveva Vargas-Solar, CNRS, LIG-LAFMIA

Genoveva.Vargas@imag.fr

<http://www.vargas-solar.com/teaching>

Knowledge discovery phases

- Data cleaning
 - handle noisy, erroneous, missing or irrelevant data (e.g., AJAX)
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation



Knowledge discovery phases

79

- Data cleaning
- Data integration
 - multiple, heterogeneous data sources may be integrated into one
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
- Data selection
 - relevant data for the analysis task retrieved from the database
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation



Knowledge discovery phases

81

- Data cleaning
- Data integration
- Data selection
- Data transformation
 - data transformed or consolidated into forms appropriate for mining (i.e., aggregation)
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

82

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining:
 - intelligent methods are applied in order to extract data patterns
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

83

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation:
 - identify the truly interesting patterns representing knowledge ← interestingness measures
- Knowledge presentation

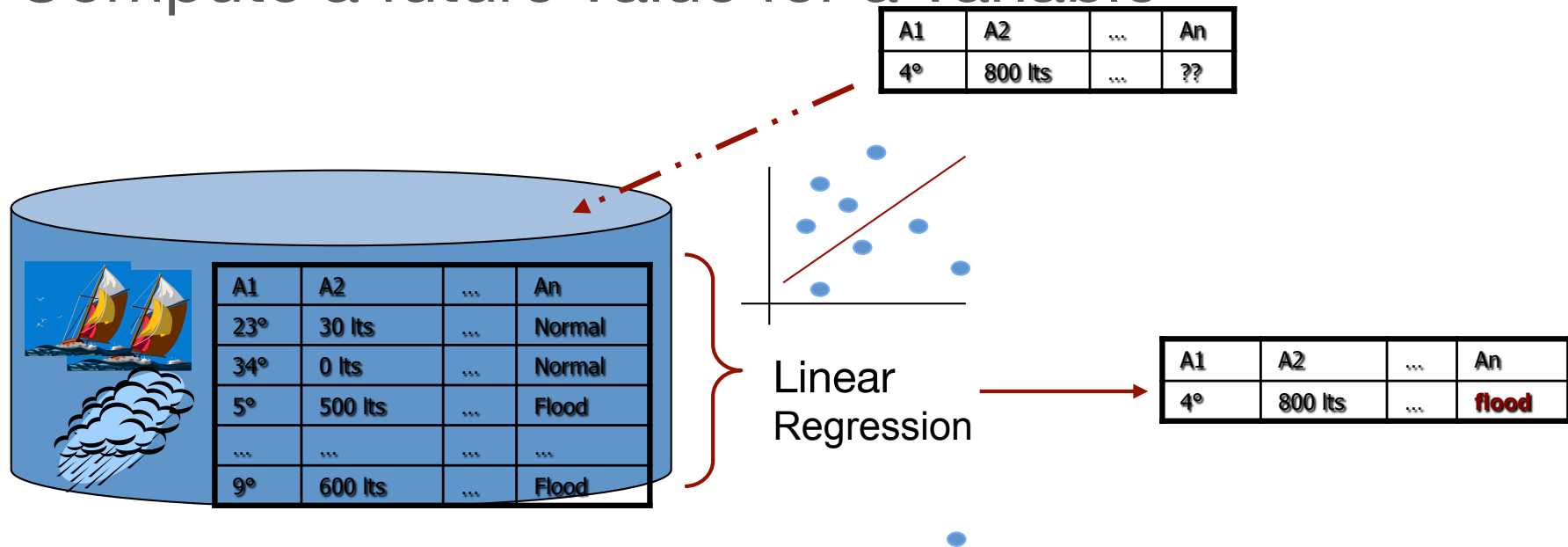
Knowledge discovery phases

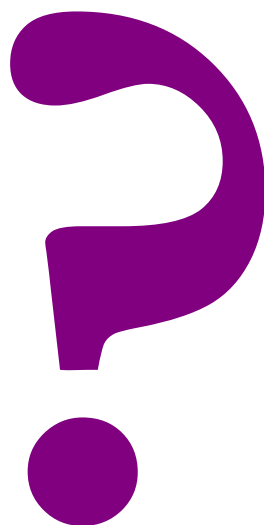
- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation
 - visualization and knowledge representation techniques
 - present knowledge to the “decision makerPattern evaluation”



Prediction

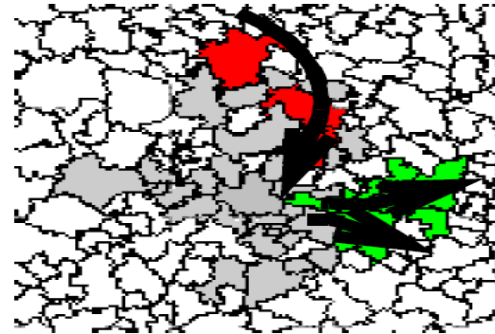
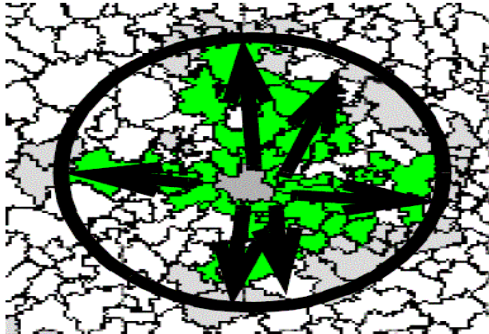
- Compute a future value for a variable

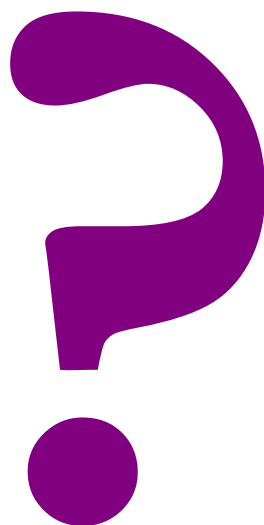




Trend detection

- Discover information, given an object and its neighbors





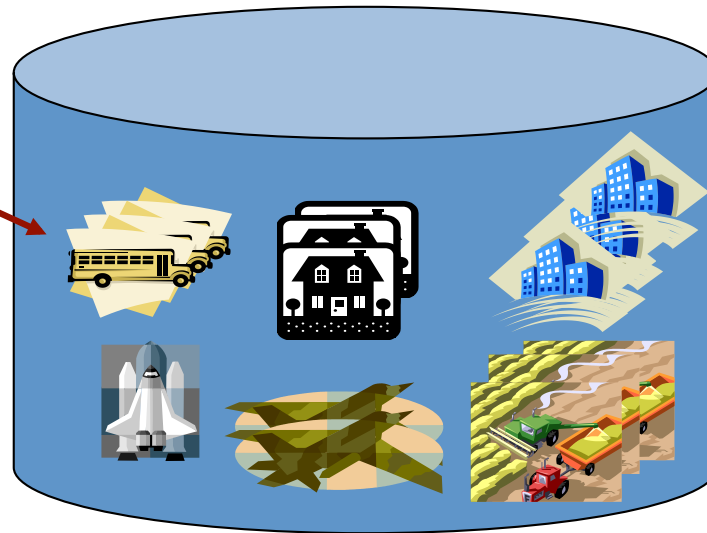
Classification

90

- General principle and definitions
- Classification based on decision trees
- Methods for performance improvement

General principle

- Given a set of classes identify whether a new object belongs to one of them



Definitions

- Process which
 - Finds the common properties among a set of objects in a database
 - Classifies them into different classes according to a classification model
 - Classification model
 - Sample database E treated as a training set where each tuple
 - Consists of the same set of multiple attributes (or features)
 - Has a known class identity associated with it
 - Objective
 - First analyze the training data and develop an accurate description of model for each class using the features
 - Class descriptions used to classify future test data or develop a better description (classification rules)
-
- U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996
 - S.M. Weiss, C.A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*, Morgan Kaufman, 1991

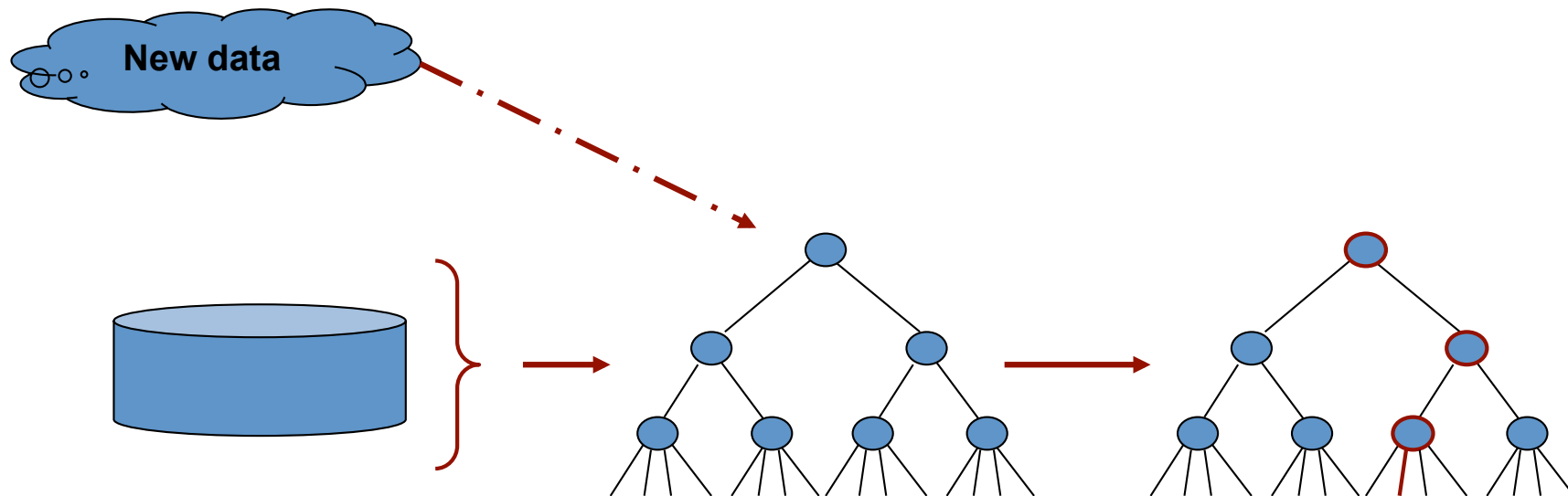
Classification

93

- ✓ General principle and definitions
- Classification based on decision trees
- Methods for performance improvement

Decision trees

- Organized data with respect to variable class
- Algorithms: ID3, C4.5, C5, CART, SLIQ, SPRINT

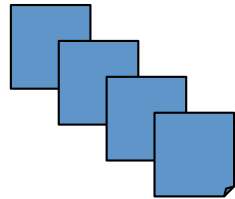


Decision trees

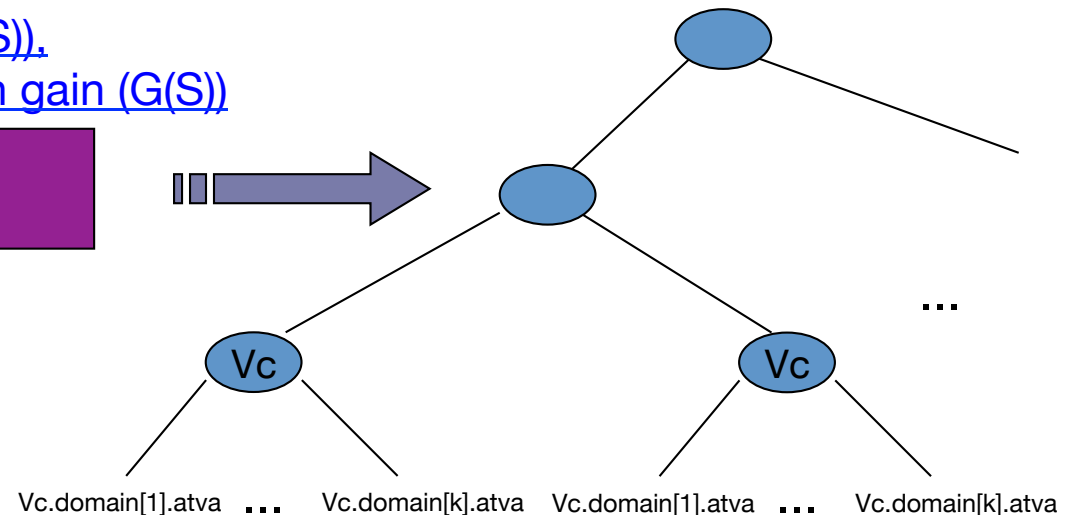
- A decision tree based classification method is a supervised learning method
 - Constructs a decision tree from a set of examples
 - The quality function of a tree depends on the classification accuracy and the size of the tree
- Choose a subset of training examples (a window) to form the decision tree
 - If the tree does not give the correct answer for all the objects
 - a selection of exceptions is added to the window
 - The process continues until the right decision tree is found
 - A tree in which
 - Each leaf carries a class name
 - Interior node specifies an attribute with a branch corresponding to each possible value of that attribute

Prediction algorithm: Interactive Dichotomizer (ID3)

A_1	...	A_i	V_c



- Top down
- Greedy
- [Entropy \(\$I\(S\)\$ \)](#).
- [Information gain \(\$G\(S\)\$ \)](#)



Data collection:

- Set of attributes

• $A_i \rightarrow \{ \langle \text{value}_1, \text{occurrence number} \rangle, \dots, \langle \text{value}_j, \text{occurrence number} \rangle \}$

- Class variable denotes values

characterizing the represented model

$V_c \rightarrow \{ \langle \text{value}_1, \text{occurrence number} \rangle, \dots, \langle \text{value}_j, \text{occurrence number} \rangle \}$

```
dat{
    Tuple[] domain;
}
```

```
Tuple{
    String atVa;
    Number nOc;
}
```


Information gain of an attribute

$$G(A_i) = I - I(A_i)$$

- $G(A_i)$ = Information gain for attribute A_i
- I = Entropy of the class variable
- $I(A_i)$ = Entropy of attribute A_i

Attribute entropy

$$I(A_i) = \sum_{j=1}^{nv(A_i)} \frac{n_{ij}}{n} I_{ij}$$

- $nv(A_i)$ = The different values number that the attribute A_i can take.
- n_{ij}/n = The probability that the attribute A_i appears in the collection
- n = The number of the rows in the data collection
- I_{ij} = Entropy of the attribute A_i with value j

Entropy of the values of an attribute A_i

- Given the value j of attribute A_i :

$$I_{ij} = - \sum_{k=1}^{nc} \frac{n_{ijk}}{n_{ij}} \log_2 \frac{n_{ijk}}{n_{ij}}$$

- nc = class variable domain cardinality
- n_{ijk}/n_{ij}
 - Given a value j of attribute A_i and a value k of the class variable
 - Probability of the occurrence of tuples in the collection containing j and k
- $\log_2 n_{ijk}/n_{ij}$ = number of digits need for representing the probability n_{ijk}/n_{ij} in binary system

Gain table

- For each attribute of the data collection
 - Compute information gain
- Order the table

Attribute	Information gain
A_1	0.6
A_2	0.5
,	
,	
,	
A_i	0.1

Decision tree construction: first step

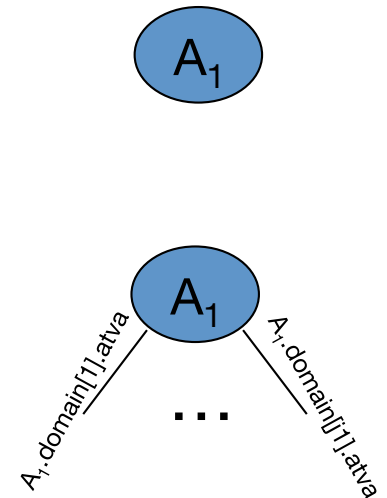
10
1

- Identify the class variable
- Compute variable base
- Compute gain table
- Root: the attribute with the highest gain
- Edges
 - Number: Vc domain cardinality
 - Label: vc in Vc.domain

A_1	...	A	Vc

Attribute	Information gain
A_1	0.6
A_2	0.5
...	
A_k	0.1

$A_1 \rightarrow \{ \langle \text{value}_1, \text{occurrence number} \rangle, \dots, \langle \text{value}_{j_1}, \text{occurrence number} \rangle \}$



Decision tree construction:

step2..n

- Class variable: root
- Select one of the root's edges
- Compute the gain table
- Node_i:
 - Attribute with the highest gain in the new table
- Edges
 - Number: Vc domain cardinality
 - Label: vc in Vc.domain

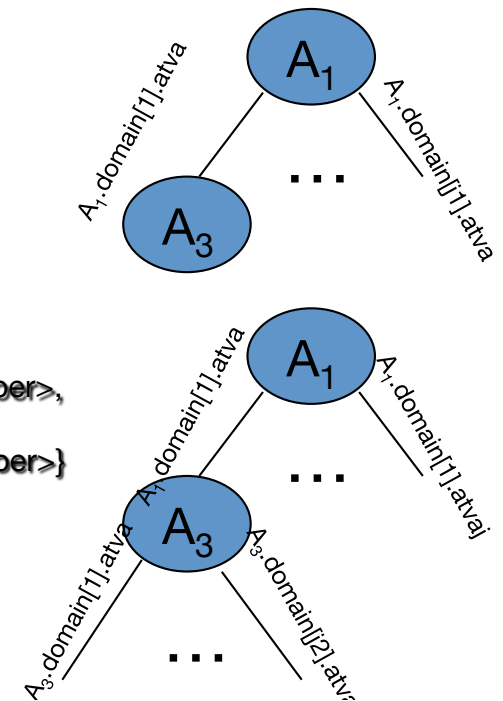
A_1	...	A_i	Vc

Attribute	Information gain
A_2	0.5
A_3	0.7
...	
A_i	0.3

$A_3 \Rightarrow \{ \langle \text{value}_1, \text{occurrence number} \rangle, \dots, \langle \text{value}_j, \text{occurrence number} \rangle \}$

→ Recursively compute nodes n_{i+1} until each root's edges have

• The original class variable is the last node and



Classification

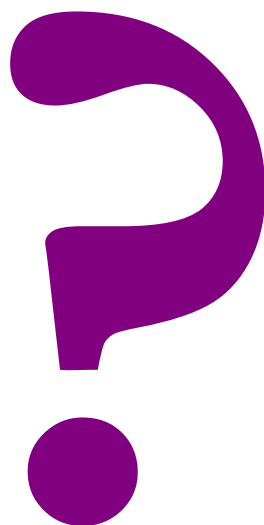
10
3

- ✓ General principle and definitions
- ✓ Classification based on decision trees
- Methods for performance improvement

Performance improvement

10
4

- Scaling up problems
 - Relatively well performance in small databases
 - Poor performance or accuracy reduction with large training sets
- Databases indices to improve on data retrieval but not in classification efficiency
 - R. Agrawal, S. Ghosh, T. Imielinsky, B. Iyer, A. Swami, An interval classifier for database mining applications, Proceedings of the 18th International Conference on Very Large Databases, August, 1992
- DBMiner improve classification accuracy: multi-level classification technique
 - Classification accuracy in large databases with attribute oriented induction and classification methods
 - J. Han, Y. Fu, Exploration of the power of attribute-oriented induction in data mining, In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996
 - J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, O.R. Zaiane, DBMiner: A system for mining knowledge in large relational databases, In Proceedings of the International Conference on Data Mining and Knowledge Discovery, August, 1996
- SLIQ (Supervised Learning in QUEST)
 - Mining classification rules in large databases
 - Decision tree classifier for numerical and categorical attributes
 - Pre-sorting technique, tree pruning
 - P.K. Chan, S.J. Stolfo, Learning arbiter and combiner trees from partitioned data for scaling machine learning, Proceedings of the 1st International Conference On Knowledge discovery and Data mining, August, 1995



Clustering

10
6

- General principle and definitions
- Randomized search for clustering large applications
- Focusing methods
- Clustering feature and CF trees

Clustering

- Discover a set of classes given a data collection



Definitions

- Process of grouping physical or abstract objects into classes of similar objects
→ Clustering or unsupervised classification
- Helps to construct meaningful partitioning of a large set of objects
 - Divide and conquer methodology
 - Decompose a large scale system into smaller components to simplify design and implementation
- Identifies clusters or densely populated regions
 - According to some distance measurement
 - In a large multidimensional data set
 - Given a set of multidimensional data points
 - The data space is usually not uniformly occupied
 - Data clustering identifies the sparse and the crowded places
 - Discovers the overall distribution patterns of the data set

Approaches

- As a branch of statistics, clustering analysis extensively studied focused on distance-based clustering analysis
 - AutoClass with Bayesian networks
 - P. Cheeseman, J. Stutz, Bayesian classification (AutoClass): Theory and results, In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds. Advances in Knowledge Discovery and Data mining, AAAI/MIT Press, 1996
 - Assume that all data point are given in advance and can be scanned frequently
- In machine learning clustering analysis
- Clustering analysis

Approaches

- As a branch of statistics
- In machine learning clustering analysis
 - Refers to unsupervised learning
 - Classes to which an object belongs to are not pre-specified
 - Conceptual clustering
 - Distance measurement may not be based on geometric distance but on that a group of objects represents a certain conceptual class
 - One needs to define a similarity between the objects and then apply it to determine the classes
 - Classes are collections of objects low interclass similarity and high intra class similarity
- Clustering analysis

Approaches

11
1

- As a branch of statistics
- In machine learning clustering analysis
- Clustering analysis
 - Probability analysis
 - Assumption that probability distributions on separate attributes are statistically independent one another (not always true)
 - The probability distribution representation of clusters → expensive clusters' updates and storage
 - Probability-based tree built to identify clusters is not height balanced
 - Increase of time and space complexity
- D. Fisher, Improving inference through conceptual clustering, *In Proceedings of the AAAI Conference*, July, 1987
- D. Fisher, Optimization and simplification of hierarchical clusterings, *In Proceedings of the 1st Conference on Knowledge Discovery and Data mining*, August, 1985

Clustering

11
2

- ✓ General principle and definitions
- Randomized search for clustering large applications
- Focusing methods
- Clustering feature and CF trees

Clustering Large applications based upon randomized Search[62]

11
3

- PAM (Partitioning Around Medoids)
 - Finds k clusters in n objects
 - First finding a representation object for each cluster
 - The most centrally located point in a cluster: medoid
 - After selecting k medoids,
 - Tries to make a better choice
 - Analyzing all possible pairs of objects such that one object is a medoid and the other is not
 - The measure of clustering quality is calculated for each such combination
 - Cost of a single iteration $O(k(n-k)^2)$ → inefficient if k is big
- CLARA (Clustering Large Applications)

Clustering Large applications based upon randomized Search

11
4

- PAM (Partitioning Around Medoids)
- CLARA (Clustering Large Applications)
 - Uses sampling techniques
 - A small portion of the real data is chosen as a representative of the data
 - Medoids are chosen from this sample using PAM
 - If the sample is selected in a fairly random manner
 - Correctly represents the whole data set
 - The representative objects (medoids) will be similar to those chosen for the whole data set
- CLARANS integrate PAM and CLARA
 - Searching only the subset of the data set not confining it to any sample at any given time
 - Draw a sample randomly in each step
 - Clustering process as searching a graph where every node is a potential solution

Clustering

11
5

- ✓ General principle and definitions
- ✓ Randomized search for clustering large applications
- Focusing methods
- Clustering feature and CF trees

Focusing methods*

- CLARANS assumes that the objects are all stored in main memory
 - Not valid for large databases →
 - Disk based methods required
 - R*-trees[11] tackle the most expensive step (i.e., calculating the distances between two clusters)
- Reduce the number of considered objects: *focusing on representative objects*
 - A centroid query returns the most central object of a leaf node of the R*-tree where neighboring points are stored
 - Only these objects used to compute medoids of the clusters
 - ☺ The number of objects is reduced
 - ☹ Objects that could have been better medoids are not considered
- Restrict the access to certain objects that do not actually contribute to the computation: computation performed only on pairs of objects that can improve the clustering quality
 - Focus on relevant clusters
 - Focus on a cluster

*M. Ester, H.P. Kriegel and X. Xu, Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification, In *Proceedings of the 4th Symposium on Large Spatial Databases*, August, 1995

Clustering

11
7

- ✓ General principle and definitions
- ✓ Randomized search for clustering large applications
- ✓ Focusing methods
- Clustering feature and CF trees

Clustering feature and CF trees

- R-trees not always available and time consuming construction
- BIRCH (Balancing Iterative Reducing and Clustering)
 - Clustering large sets of points
 - Incremental method
 - Adjustment of memory requirements according to available size

Clustering feature and CF trees: concepts

11
9

■ Clustering feature

- CF is the triplet summarizing information about subclusters of points. Given n-dimensional points in a subcluster $\{X_i\}$

- $CF = \left(N, \overrightarrow{LS}, SS \right)$
 - N is the number of points in the subcluster
 - LS is the linear sum on N points
 - SS is the squares sum of data points

$$\sum_{i=1}^N \overrightarrow{X_i}$$
$$\sum_{i=1}^N X_i^2$$

■ Clustering features

- Are sufficient for computing clusters
- Summarize information about subclusters of points instead of storing all points
 - Constitute an efficient storage information method since they

Clustering feature and CF trees: concepts

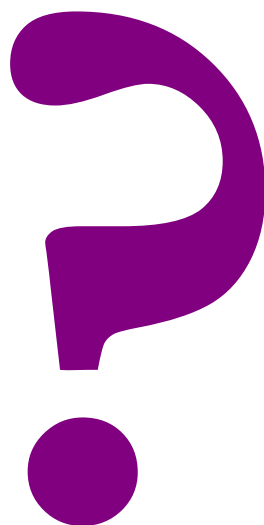
12
0

- Clustering feature tree
 - Branching factor B specifies the maximum number of children
 - Threshold T specifies the maximum diameter of subclusters stored at the leaf nodes
 - Changing the T we can change the size of the tree
 - Non leaf nodes are storing sums of their children CF's → summarize information about their children
 - Incremental method: built dynamically as data points are inserted
 - A point is inserted in the closes leaf entry
 - If the diameter of the cluster stored in the leaf node after insertion is larger than T
 - Split it and eventually other nodes
 - After insertion the information about the new point is transmitted to the root

Clustering feature and CF trees: concepts

12
1

- Clustering feature tree
 - The size of CF tree can be changed by changing T
 - If the size of the memory needed for storing the CF tree is larger than the size of the main memory
 - Then a larger T is specified and the tree is rebuilt
 - Rebuild process is done by building a new tree from the leaf nodes of the old tree
 - Reading all the points is not necessary
- CPU and I/O costs of BIRCH $O(N)$
 - Linear scalability of the algorithm with respect to the number of points
 - Insensitivity of the input order
 - Good quality of clustering of the data
- T. Zhang, R. Ramakrishnan, M. Livy, BIRCH: an efficient data clustering method for very large databases, In Proceedings of the ACM SIGMOD International Conference on Management of Data, June 1996



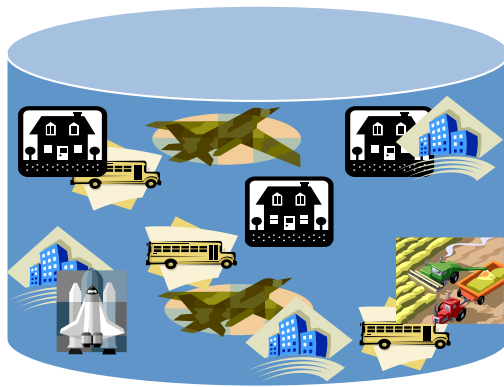
Association rules

12
3




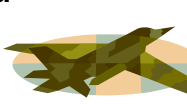
- General principle
- A priori algorithm: an example
- Mining generalized rules
- Improving efficiency of mining association rules

Association rules

- Given a data collection determine possible relationships among the variables describing such data
- Relationships are expressed as association rules $X \rightarrow Y$



Near( , ) \rightarrow Cheap house

Near( , )
and
Near( , ) \rightarrow Expensive house

Association rules

- ✓ General principle
- A priori algorithm: an example
- Mining generalized rules
- Other issues on mining association rules
 - Interestingness of discovered association rules
 - Improving efficiency of mining association rules

Mathematical model

- Let $I = \{i_1, \dots, i_n\}$ be a set of literals called items
- D a set of transaction where each t in T is a set of items such that
$$T \subseteq I$$
- Each transaction has in TID
- Let X be a set of items, T is said to contain X iff X in T
- An association rule $X \rightarrow Y$ where X in I , Y in I and X does not intersect Y
 - Holds in the transaction set D with confidence c if $c\%$ of the transactions in D that contain X also contain Y
 - Has support s in the transaction set D if $s\%$ of transactions in D contain the intersection of X and Y

Mathematical model

- Confidence denotes the strength of implication
- Support indicates the frequencies of the occurring patterns in the rule
 - Reasonable to pay attention to rules with reasonably large support: strong rules
 - Discover strong rules in large data bases
 - Discover large item sets
 - the sets of itemsets that have transaction support above a predetermined minimum support s
 - Use large itemsets to generate association rules for the database

Algorithm a priori*

Database D

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

Scan D
→

C₁

Itemset	s
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L₁

Itemset	s
{A}	2
{B}	3
{C}	3
{E}	3

- In each iteration
 - Construct a candidate set of large itemsets
 - Count the number of occurrences in of each candidate itemset
 - Determine large itemsets based on a pre-determined minimum support
- In the first iteration
 - Scan all transactions to count the number of occurrences for each item

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

Algorithm a priori*

C_2

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

Scan D
→

Itemset	s
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

L_1

Itemset	s
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

■ Second iteration

- Discover 2-itemsets

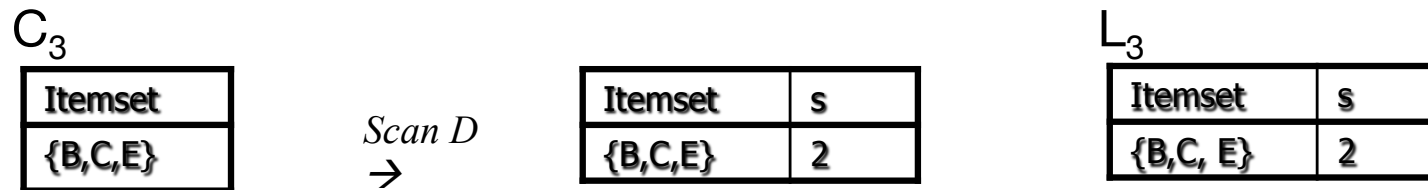
- Candidate set C_2 : $L_1 * L_1$

- C_2 consists of $\left(\frac{|L_1|}{2}\right)$ 2-itemsets

$$L_k * L_k = \{X \cup Y \mid X, Y \in L_k, |X \cap Y| = k - 1\}$$

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

Algorithm a priori*



- From L_2
 - two large 2-itemsets are identified with the same first item: {B,C} and {B,E}
 - {C,E} is a two large 2-itemset? YES!
- No candidate 4-itemset \rightarrow END
- **HOMEWORK: Analyze DHP in**

J.-S. Park, P.S. Yu, An effective hash based algorithm for mining association rules, Proceedings of the ACM SIGMOD, May, 1995

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

Association rules

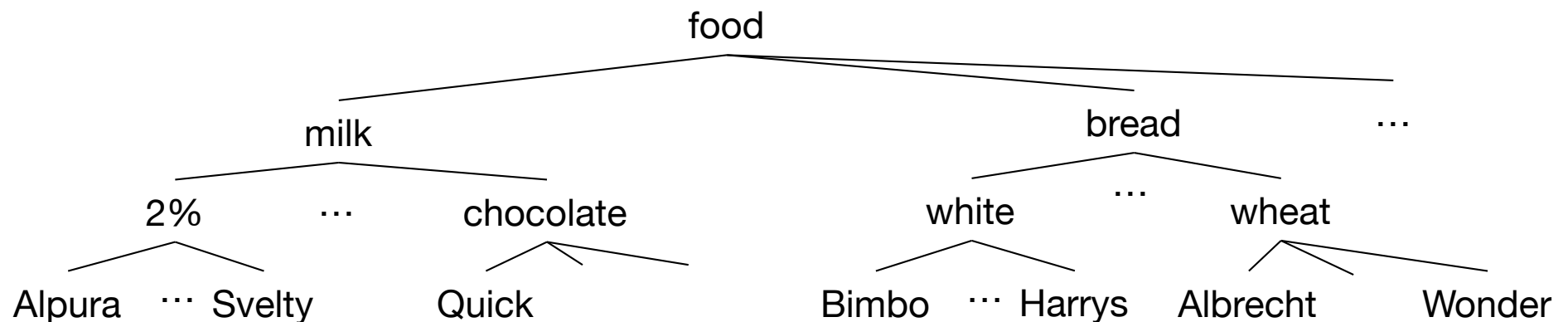
- ✓ General principle
- ✓ A priori algorithm: an example
- Mining generalized rules
- Other issues on mining association rules
 - Interestingness of discovered association rules
 - Improving efficiency of mining association rules

Mining generalized and multiple-level association rules

13
2

- Interesting associations among data items occur at a relatively high concept level
 - Purchase patterns in a transaction database many not show substantial regularities at a primitive data level (e.g., bar code level)
 - Interesting regularities at some high concept level such as milk and bread
- Study association rules at a generalized abstraction level or at multiple levels

Mining generalized and multiple-level association rules*



- The bar codes of 1 gallon of Alpura 2% milk and 1lb of Wonder wheat bread: what for?
- 80% of the customers that purchase milk also purchase bread
- 70% of people buy wheat bread if they buy 2% milk

* J. Han, Y. Fu, Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21th International Conference of Very Large Databases*, September 1995

Mining generalized and multiple-level association rules*

13
4

- Low level associations may be examined only when
 - High level parents are large at their corresponding levels
 - Different levels may adopt different minimum support thresholds
- Four algorithms developed for efficient mining of association rules
 - Based on different ways of sharing multiple level mining processes and reduction of encoded transaction tables
- Mining of quantitative association rules
 - R. Srikant, R. Agrawal, Mining Generalized Association Rules, *Proceedings of the 21st International Conference on Very Large Databases*, September, 1995
- Meta rule guided mining of association rules in relational databases
 - Y. Fu, J. Han, Meta rule guided mining of association rules in relational databases, *Proceedings of the 1st International Workshop on Integration of Knowledge with Deductive and Object Oriented Databases (KDOOD)*, Singapore, December, 1995
 - W. Shen, K. Ong, B. Mitbander, C. Zaniolo, Metaqueries for data mining, In U.M. Fayard, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data mining*, AAAI/MIT Press, 1996

* J. Han, Y. Fu, Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21th International Conference of Very Large Databases*, September 1995

Association rules

- ✓ General principle
- ✓ A priori algorithm: an example
- ✓ Mining generalized rules
- Other issues on mining association rules
 - Interestingness of discovered association rules
 - Improving efficiency of mining association rules

Interestingness of discovered association rules

13
6

- Not all discovered association rules are strong (i.e., passing the minimum support and minimum confidence thresholds)
- Consider a survey done in a university of 5000 students: students and activities they engage in the morning
 - 60% of the students play basket ball, 75% eat cereal, 40% both play basket ball and eat cereal
 - Suppose that a mining program runs
 - Minimal student support $s = 2000$
 - Minimal confidence is 60%
 - Play basket ball \rightarrow eat cereal
 - $2000/3000 = 0,66$
 - Pb!!! The overall percentage of students eating cereal is 75% > 66%
 - Playing basket ball and eating cereal are negatively associated: being involved in one decreases the likelihood of being involved in the other

Interestingness of discovered association rules

13
7

- Filter out misleading associations
 - $A \rightarrow B$ is interesting if its confidence exceeds a certain measure
 - Test of statistical independence

$$\frac{P(A \cap B)}{P(A)} - P(B) > d$$

$$P(A \cap B) - P(A) * P(B) > k$$

- Interestingness studies

- G. Piatetsky-Shapiro, Discovery analysis and presentation of strong rules, In G. Piatetsky-Shapiro and W.J. Frawley, eds. *Knowledge Discovery in Databases*, AAAI/MIT press, 1991
- A. Silberschatz, M. Stonebraker, J.D. Ullman, Database research: Achievements and opportunities into the 21st century, In *Report of an NSF Workshop on the Future of Database Systems Research*, May, 1995
- R. Srikant, R. Agrawal, Mining generalized association rules, *Proceedings of the 21st International Conference on Very Large Databases*, September, 1995

Association rules

- ✓ General principle
- ✓ A priori algorithm: an example
- ✓ Mining generalized rules
- Other issues on mining association rules
 - ✓ Interestingness of discovered association rules
 - Improving efficiency of mining association rules

Improving the efficiency of mining association rules

13
9

- Database scan reduction:
 - Profit from database scans C_i in order to compute in advance L_i and L_{i+1}
 - M.S. Chen, J.S. Park, P.S. Yu, Data mining for path traversal patterns in a Web Environment, *Proceedings of the 16th International Conference on Distributed Computing Systems*, May, 1996
- Sampling: mining with the adjustable accuracy
- Incremental updating of discovered association rules
- Parallel data mining

Improving the efficiency of mining association rules

14
0

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
 - Frequent basis for mining transaction data to capture behavior
 - Efficiency more important than accuracy
 - Attractive due to the increasing size of databases
- H. Mannila, H. Toivonen, A. Inkeri Verkamo, Efficient algorithms for discovering association rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, July, 1994
- J.-S. Park, M.S. Chen, P.S. Yu, Mining association rules with adjustable accuracy, IBM research report, 1995
- R. Srikant, R. Agrawal, 1995, *ibidem*.
- Incremental updating of discovered association rules
- Parallel data mining

Improving the efficiency of mining association rules

14
1

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
 - Frequent basis for mining transaction data to capture behavior
 - Efficiency more important than accuracy
 - Attractive due to the increasing size of databases
- H. Mannila, H. Toivonen, A. Inkeri Verkamo, Efficient algorithms for discovering association rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, July, 1994
- J.-S. Park, M.S. Chen, P.S. Yu, Mining association rules with adjustable accuracy, IBM research report, 1995
- R. Srikant, R. Agrawal, 1995, *ibidem*.
- Incremental updating of discovered association rules
- Parallel data mining

Improving the efficiency of mining association rules

14
2

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
- Incremental updating of discovered association rules
 - On data base updates →
 - Maintenance of discovered association rules required
 - Avoid redoing data mining on the whole updated database
 - Rules can be invalidated and weak rules become strong
 - Reuse information of the large itemsets and integrate the support information of new ones
 - Reduce the pool of candidate sets to be examined
 - D.W. Cheung, J. Han, V. Ng, C.Y. Wong, Maintenance of discovered association rules in large databases: an incremental updating technique, *In Proceedings of the International Conference on Data Engineering*, February, 1996
- Parallel data mining

Improving the efficiency of mining association rules

14
3

- Database scan reduction:
 - Sampling: mining with the adjustable accuracy
 - Incremental updating of discovered association rules
 - Parallel data mining
 - Progressive knowledge collection and revision based on huge transaction databases
 - DB partitioned → inter-node data transmission for making decisions can be prohibitively large
-
- IBM *Scalable POWERparallel Systems*, Technical report GA23-2475-02, February, 1995
 - J.S. Park, M.S. Chen, P.S. Yu, Efficient parallel data mining for association rules, *Proceedings of the 4th International Conference on Information and Knowledge Management*, November, 1995

