Chapter1_applicationDM.ppt

Chapter2_0.ppt

Chapter3_0.ppt

Chapter3_1.ppt

Chapter3_2.ppt

Chapter3_3.ppt

Chapter4_0.ppt

Chapter4_1.ppt

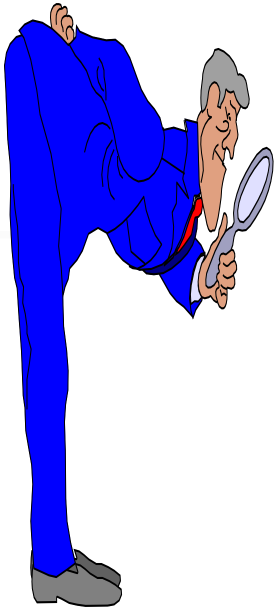
Chapter4_2.ppt

By Babu Ram Dawadi

Data Mining:
Trends and Applications

Babu Ram Dawadi

Data Mining??



- Data Mining:

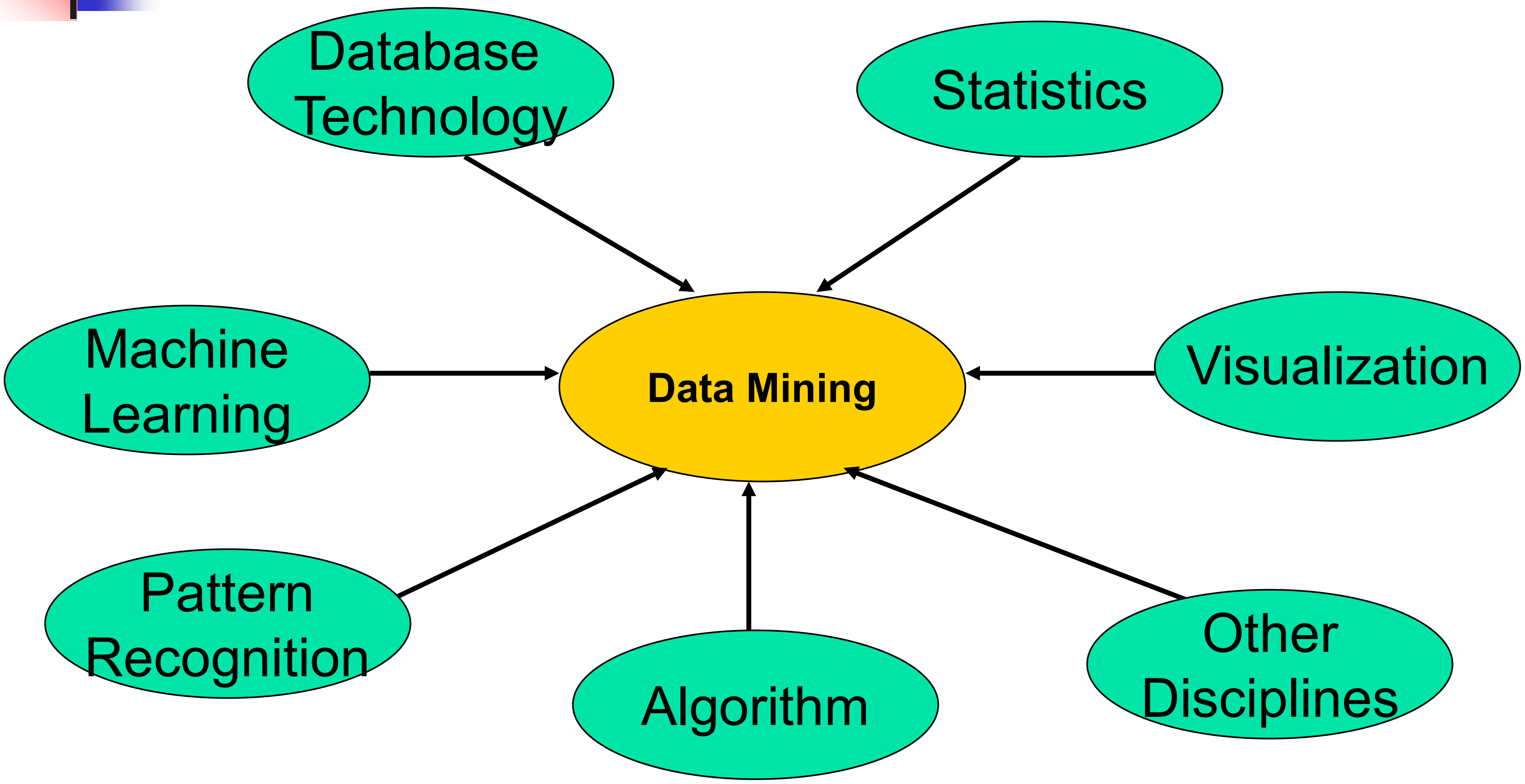
  ❖The process of Discovering meaningful patterns & trends often previously unknown, by shifting large amount of data, using pattern recognition, statistical and Mathematical techniques.

  ❖A group of techniques that find relationship that have not previously been discovered

What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- Alternative names and their “inside stories”:
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?
  - (Deductive) query processing.
  - Expert systems

Data Mining: Confluence of Multiple Disciplines



Data Mining: On What Kinds of Data?

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

Data Mining Applications

- Data mining is a young discipline with wide and diverse applications
  - There is still a nontrivial gap between general principles of data mining and domain-specific, effective data mining tools for particular applications
- Some application domains
  - Biomedical and DNA data analysis
  - Financial data analysis
  - Retail industry
  - Telecommunication industry

Biomedical Data Mining and DNA Analysis

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Humans have around 100,000 genes
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
  - Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
  - Data cleaning and data integration methods developed in data mining will help

DNA Analysis: Examples

- Similarity search and comparison among DNA sequences
  - Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
  - Identify gene sequence patterns that play roles in various diseases
- **Association analysis**: identification of co-occurring gene sequences
  - Most diseases are not triggered by a single gene but by a combination of genes acting together
  - Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples
- **Path analysis**: linking genes to different disease development stages
  - Different genes may become active at different stages of the disease
  - Develop pharmaceutical interventions that target the different stages separately

Data Mining for Financial Data Analysis

- Loan payment prediction/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating
- Classification and clustering of customers for targeted marketing
  - multidimensional segmentation by nearest-neighbor, classification, decision
- Detection of money laundering and other financial crimes
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry

- Retail industry: huge amounts of data on sales, customer shopping history, etc.
- Applications of retail data mining
  - Identify customer buying behaviors
  - Discover customer shopping patterns and trends
  - Improve the quality of customer service
  - Achieve better customer retention and satisfaction
  - Enhance goods consumption ratios
  - Design more effective goods transportation and distribution policies

Data Mining for Telecomm. Industry (1)

- A rapidly expanding and highly competitive industry and a great demand for data mining
  - Understand the business involved
  - Identify telecommunication patterns
  - Catch fraudulent activities
  - Make better use of resources
  - Improve the quality of service
- Multidimensional analysis of telecommunication data
  - Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

Data Mining for Telecomm. Industry (2)

- Fraudulent pattern analysis and the identification of unusual patterns
  - Identify potentially fraudulent users and their atypical usage patterns
  - Detect attempts to gain fraudulent entry to customer accounts
  - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
  - Find usage patterns for a set of communication services by customer group, by month, etc.
  - Promote the sales of specific services
  - Improve the availability of particular services in a region

Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Examples of Data Mining Systems (1)

- IBM Intelligent Miner
  - A wide range of data mining algorithms
  - Scalable mining algorithms
  - Toolkits: neural network algorithms, statistical methods, data preparation, and data visualization tools
  - Tight integration with IBM's DB2 relational database system
- SAS Enterprise Miner
  - A variety of statistical analysis tools
  - Data warehouse tools and multiple data mining algorithms
- Mirosoft SQLServer 2000
  - Integrate DB and OLAP with mining
  - Support OLEDB for DM standard

Examples of Data Mining Systems (2)

- SGI MineSet
  - Multiple data mining algorithms and advanced statistics
  - Advanced visualization tools
- Clementine (SPSS)
  - An integrated data mining development environment for end-users and developers
  - Multiple data mining algorithms and visualization tools
- DBMiner (DBMiner Technology Inc.)
  - Multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
  - Efficient, association and sequential-pattern mining functions, and visual classification tool
  - Mining both relational databases and data warehouses

Data Mining and Intelligent Query Answering

- Query answering
  - Direct query answering: returns exactly what is being asked
  - Intelligent (or cooperative) query answering: analyzes the intent of the query and provides generalized, neighborhood or associated information relevant to the query
- Some users may not have a clear idea of exactly what to mine or what is contained in the database
- Intelligent query answering analyzes the user's intent and answers queries in an intelligent way

Data Mining and Intelligent Query Answering (2)

- A general framework for the integration of data mining and intelligent query answering
  - Data query: finds concrete data stored in a database
  - Knowledge query: finds rules, patterns, and other kinds of knowledge in a database

- Ex. Three ways to improve on-line shopping service
  - Informative query answering by providing summary information
  - Suggestion of additional items based on association analysis
  - Product promotion by sequential pattern mining

Data Mining: Merely Managers' Business or Everyone's?

- Data mining will surely be an important tool for managers' decision making
  - Bill Gates: "Business @ the speed of thought"
- The amount of the available data is increasing, and data mining systems will be more affordable
- Multiple personal uses
  - Mine your family's medical history to identify genetically-related medical conditions
  - Mine the records of the companies you deal with
  - Mine data on stocks and company performance, etc.
- Invisible data mining
  - Build data mining functions into many intelligent tools

Trends in Data Mining (1)

- Application exploration
  - development of application-specific data mining system
  - Invisible data mining (mining as built-in function)

- Scalable data mining methods
  - Constraint-based mining: use of constraints to guide data mining systems in their search for interesting patterns

- Integration of data mining with database systems, data warehouse systems, and Web database systems

Summary

- Domain-specific applications include biomedicine (DNA), finance, retail and telecommunication data mining
- There exist some data mining systems and it is important to know their power and limitations
- Intelligent query answering can be integrated with mining

Data Warehousing

- Data
  - Raw piece of information that is capable of being moved and store.
- Database
  - An organized collection of such data in which data are managed in tabular form with relationship.
- Data Warehouse
  - System that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lots of reports by the use of mining tools.

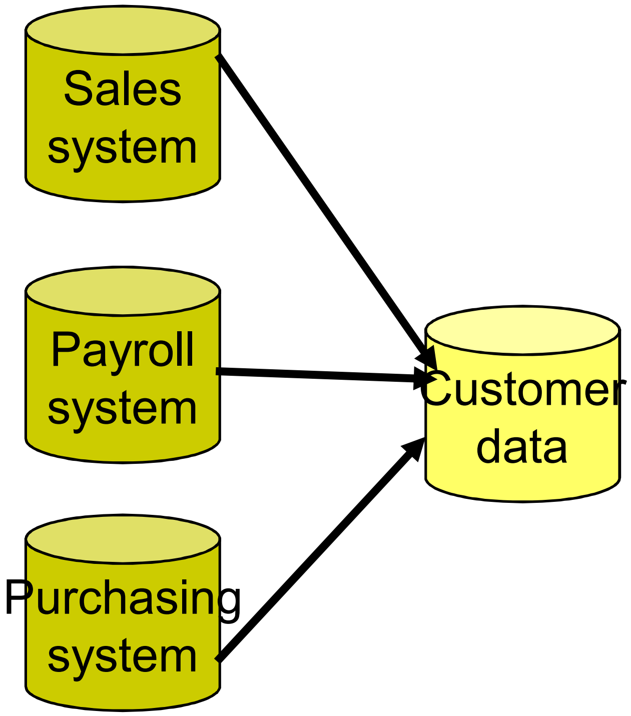
Data Warehouse…

  - “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”

- Data warehousing:
  - The process of constructing and using data warehouses.
  - Is the process of extracting & transferring operational data into informational data & loading it into a central data store (warehouse)

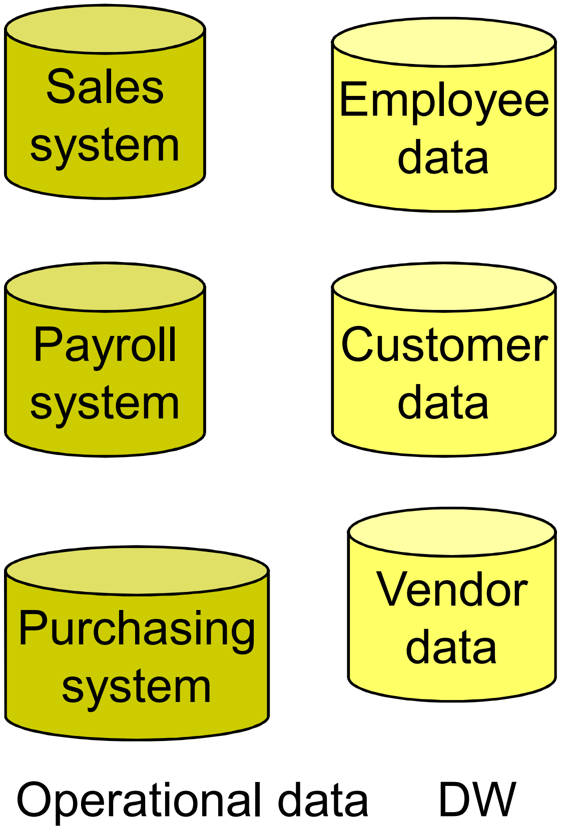
Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.



Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

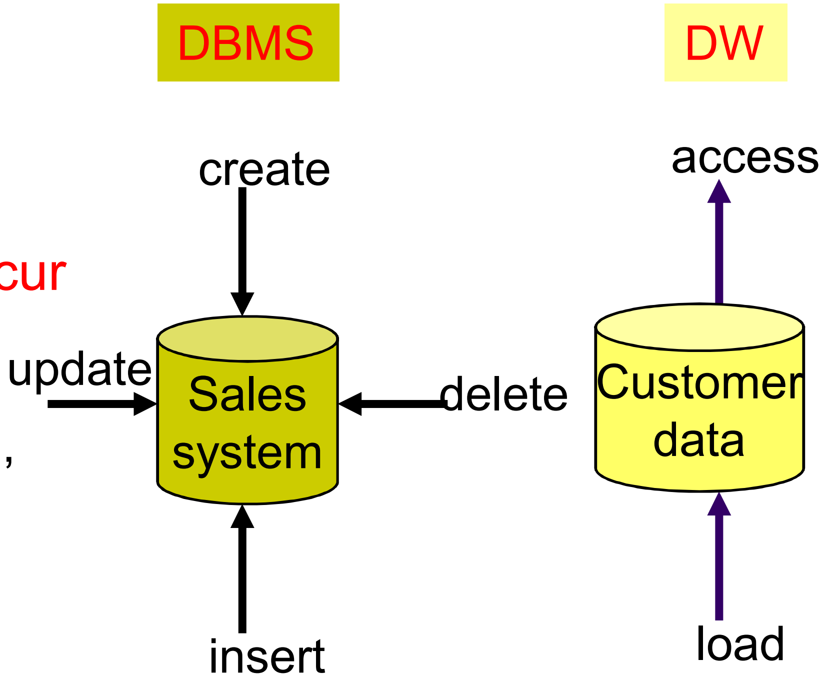


Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

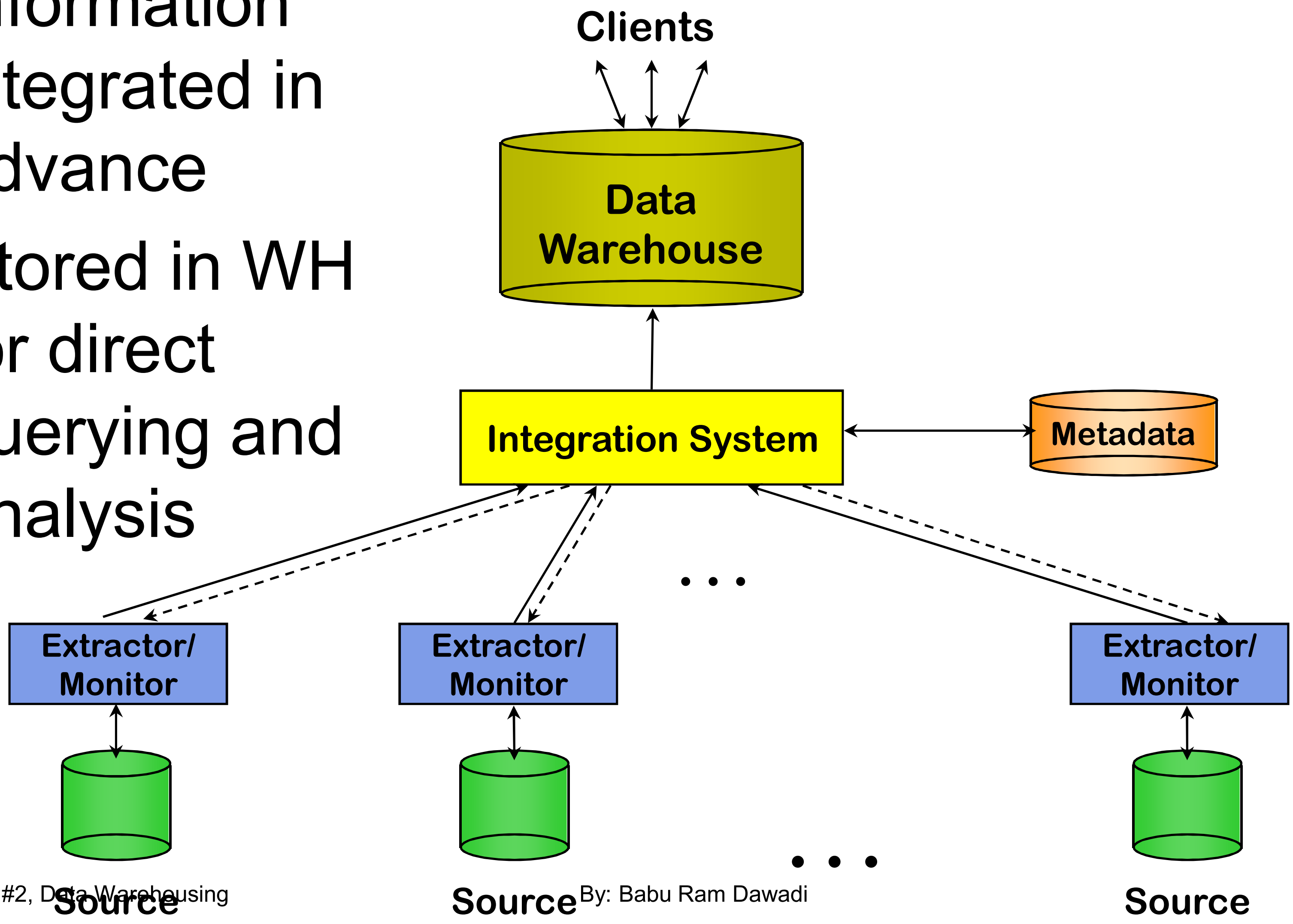


Data Warehouse Usage

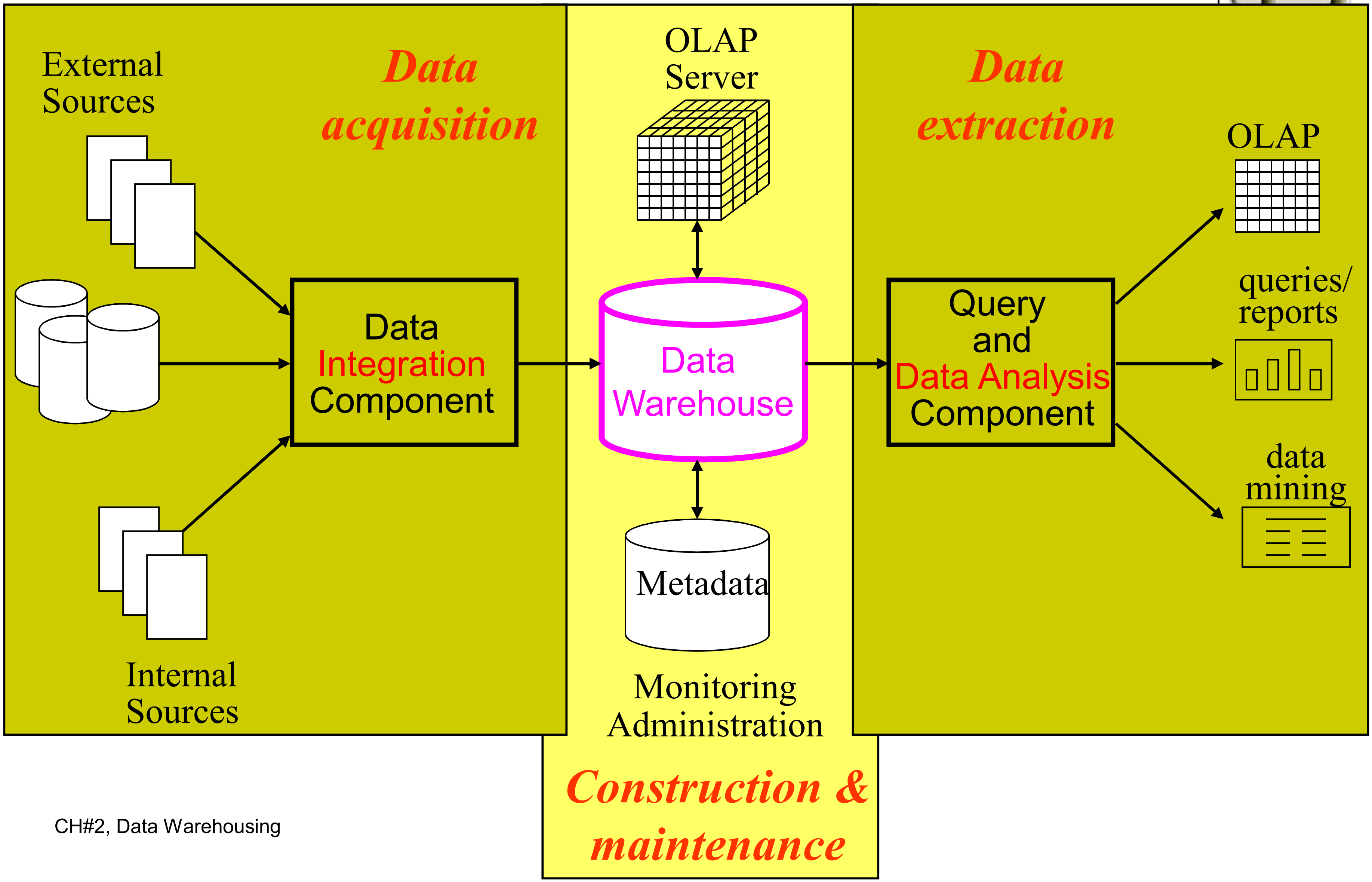
- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

The Warehousing Approach

- Information integrated in advance
- Stored in WH for direct querying and analysis



General Architecture



3 main phases

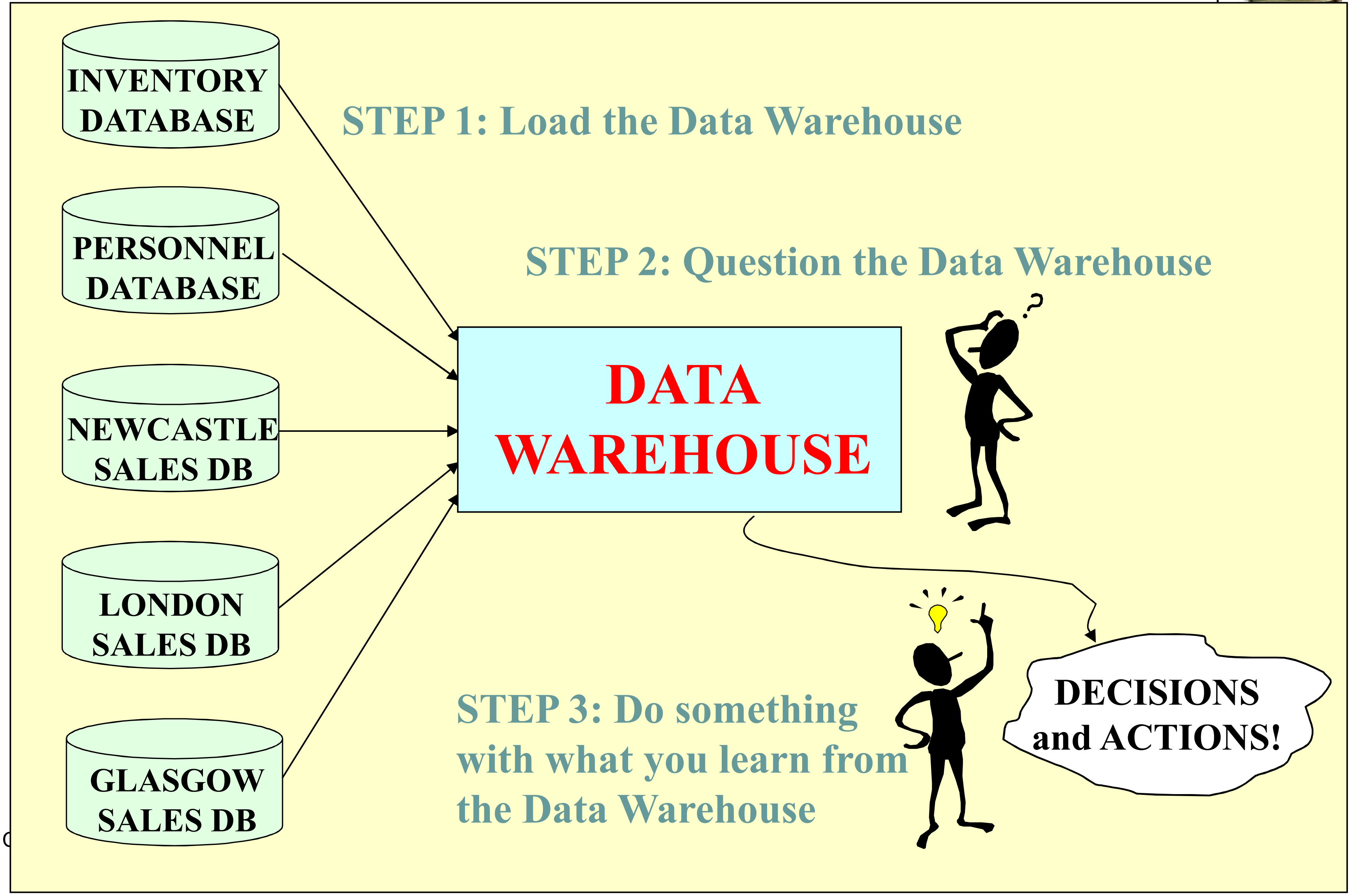
- Data acquisition
  - relevant data collection
  - Recovering: transformation into the data warehouse model from existing models
  - Loading: cleaning and loading in the DWH
- Storage
- Data extraction
  - Tool examples: Query report, SQL, multidimensional analysis (OLAP tools), datamining
- + evolution and maintenance

DW Monitoring

- Identify growth factors and rate
- Identify what data is being used
- Identify who is using the data, and when
  - → Avoid constant growth
  - → Plan for evolution (trends)
- Control response time (latency)

DATA WAREHOUSING

THE USE OF A DATA WAREHOUSE



INVENTORY DATABASE

PERSONNEL DATABASE

NEWCASTLE SALES DB

LONDON SALES DB

GLASGOW SALES DB

STEP 1: Load the Data Warehouse

DATA WAREHOUSE

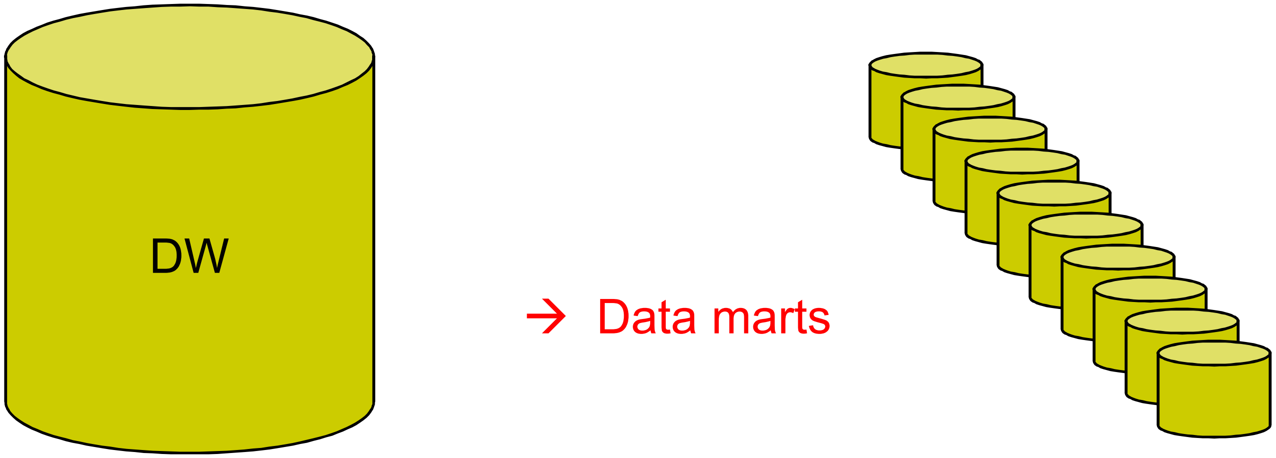
STEP 2: Question the Data Warehouse

STEP 3: Do something with what you learn from the Data Warehouse

DECISIONS and ACTIONS!

Partitioning

- To improve performances & flexibility without giving up on the details



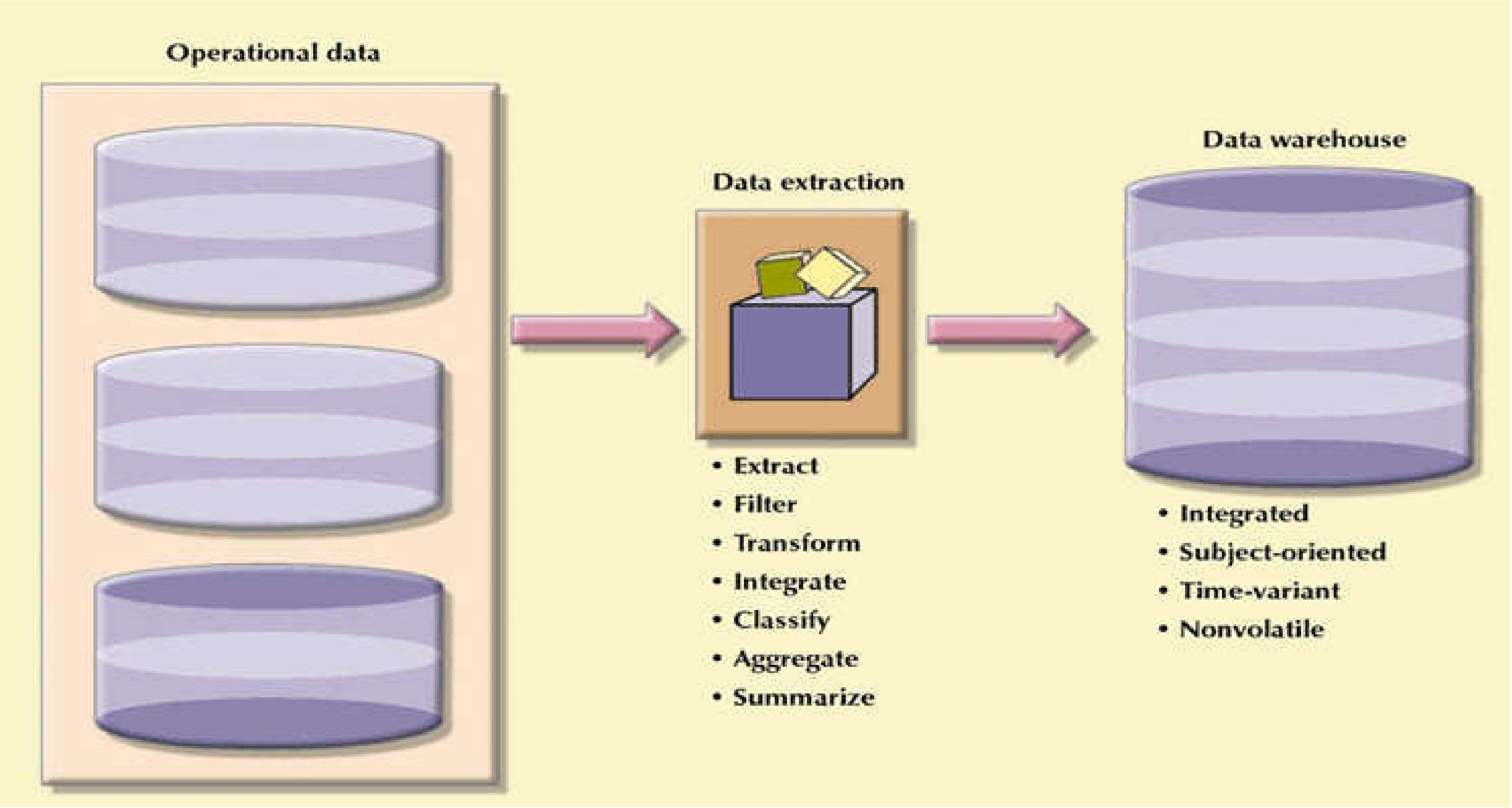
→ Data marts

- By date, business type, geography, …

The Need for Data Analysis

- Managers must be able to track daily transactions to evaluate how the business is performing
- By tapping into the operational database, management can develop strategies to meet organizational goals
- Data analysis can provide information about short-term tactical evaluations and strategies

Creating a Data Warehouse



Factors Common to Data Warehousing

- Dynamic framework for decision support that is always a work in progress
- Must satisfy:
  - Data integration and loading criteria
  - Data analysis capabilities with acceptable query performance
  - End-user data analysis needs
- Apply database design procedures

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation(aggregation).
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: Decision Support requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats

Decision Support Systems

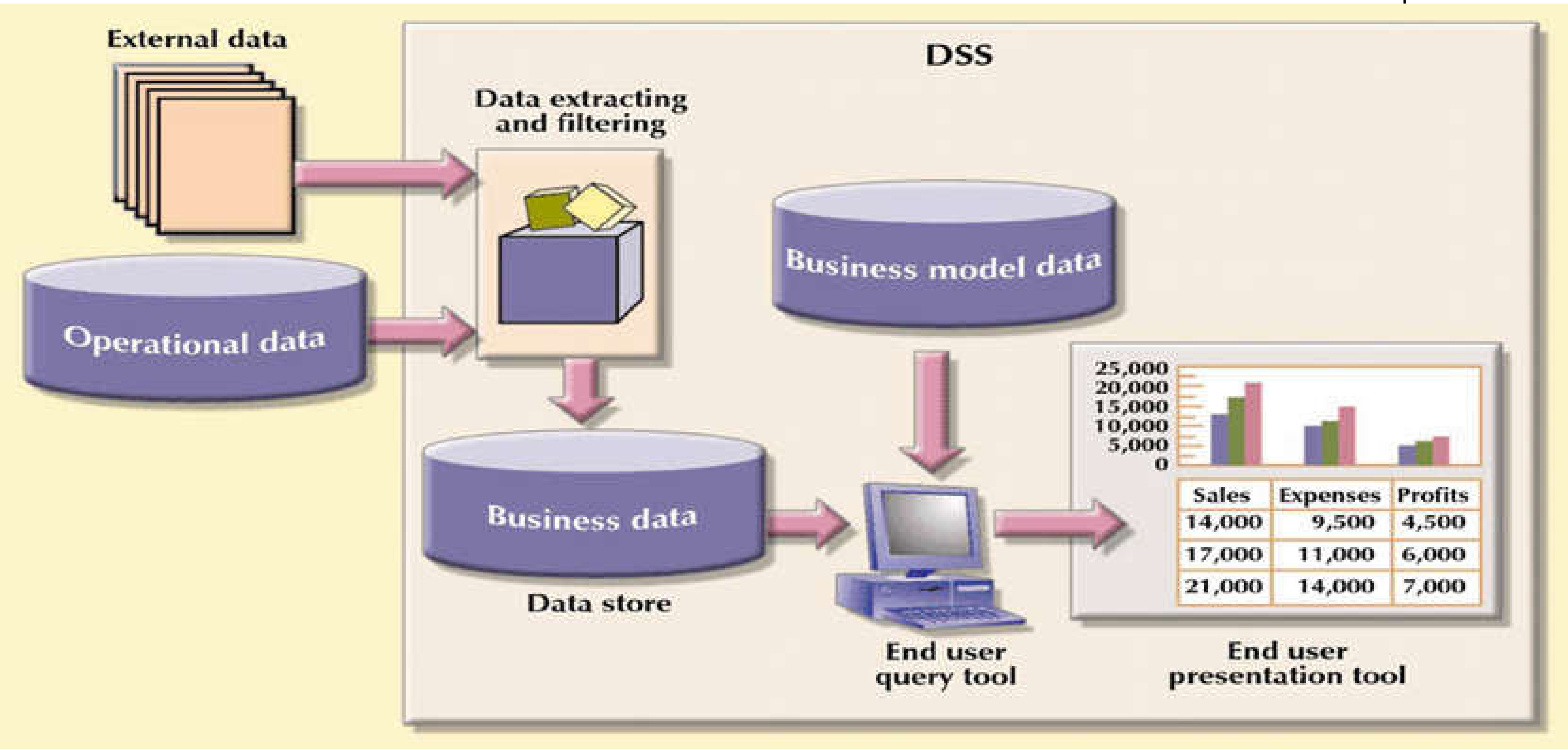
- Methodology (or series of methodologies) designed to extract information from data and to use such information as a basis for decision making

- Decision support system (DSS):
  - Arrangement of computerized tools used to assist managerial decision making within a business
  - Usually requires extensive data "massaging" to produce information
  - Used at all levels within an organization
  - Often tailored to focus on specific business areas
  - Provides ad hoc query tools to retrieve data and to display data in different formats

Decision Support Systems (continued)

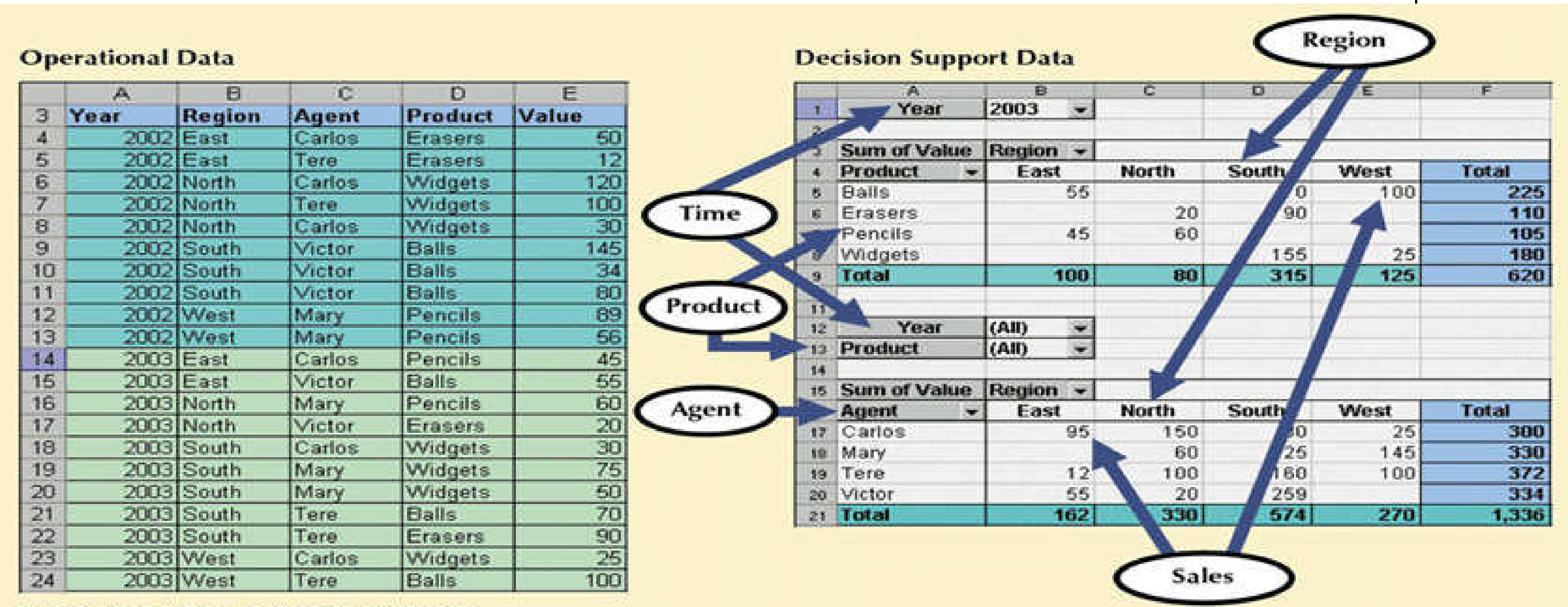
- Composed of four main components:
  - Data store component
    - Basically a DSS database
  - Data extraction and filtering component
    - Used to extract and validate data taken from operational database and external data sources
  - End-user query tool
    - Used to create queries that access database
  - End-user presentation tool
    - Used to organize and present data

Main Components of a Decision Support System (DSS)



Sales	Expenses	Profits
14,000	9,500	4,500
17,000	11,000	6,000
21,000	14,000	7,000

Transforming Operational Data Into Decision Support Data



Operational Data

	A	B	C	D	E
3	Year	Region	Agent	Product	Value
4	2002	East	Carlos	Erasers	50
5	2002	East	Tere	Erasers	12
6	2002	North	Carlos	Widgets	120
7	2002	North	Tere	Widgets	100
8	2002	North	Carlos	Widgets	30
9	2002	South	Victor	Balls	145
10	2002	South	Victor	Balls	34
11	2002	South	Victor	Balls	80
12	2002	West	Mary	Pencils	89
13	2002	West	Mary	Pencils	56
14	2003	East	Carlos	Pencils	45
15	2003	East	Victor	Balls	55
16	2003	North	Mary	Pencils	60
17	2003	North	Victor	Erasers	20
18	2003	South	Carlos	Widgets	30
19	2003	South	Mary	Widgets	75
20	2003	South	Mary	Widgets	50
21	2003	South	Tere	Balls	70
22	2003	South	Tere	Erasers	90
23	2003	West	Carlos	Widgets	25
24	2003	West	Tere	Balls	100

Operational data have a narrow time span, low granularity, and single focus. Such data are usually presented in tabular format, in which each row represents a single transaction. This format often makes it difficult to derive useful information.

Decision Support Data

	A	B	C	D	E	F
1	Year	2003				
2						
3	Sum of Value	Region				
4	Product	East	North	South	West	Total
5	Balls	55		0	100	225
6	Erasers		20	90		110
7	Pencils	45	60			105
8	Widgets			155	25	180
9	Total	100	80	315	125	620
10						
11						
12	Year	(All)				
13	Product	(All)				
14						
15	Sum of Value	Region				
16	Agent	East	North	South	West	Total
17	Carlos	95	150	0	25	300
18	Mary		60	25	145	330
19	Tere	12	100	160	100	372
20	Victor	55	20	259		334
21	Total	162	330	574	270	1,336

Decision support system (DSS) data focus on a broader time span, tend to have high levels of granularity, and can be examined in multiple dimensions. For example, note these possible aggregations:

Sales by product, region, agent, etc.
Sales for all years or only a few selected years.
Sales for all products or only a few selected products.

Designing DSS

- DSS is the more general term referring to all kinds of analysis of existing data in order to make better decisions, like: data mining, OLAP, Simulation etc…

- DSS design differs considerably from that of an online transaction processing (OLTP). In contrast to OLTP, DSS are used only for queries.

Designing DSS

- Designing a DSS seeks particular importance on:
  - Requirement of the end user
  - Software requirement
  - Hardware requirement

- End user requirement
  - Discuss with the end user
  - People who need to use DSS produce a huge variety of queries
  - Some are interested only on a particular part of the information so that they may prefer to optimize the application completely in order to speed up the query process.

DSS …

- Software Requirement
  - Type of software depends very much on the requirement of the end user.
  - Working on a client/server environment allows flexibility in choosing the appropriate software for end users.
  - For data mining, software can be split into two parts:
    - The first works with the algorithms on the database server
    - The second work on the local workstation.

DSS…

- Hardware Requirement
  - A large DW can contain hundreds of thousands of giga bytes.
    - So DW is designed by Engineer with knowledge of both hardware and software
  - For data mining, it is not always necessary to have a very large database and a large database server.

ON-LINE ANALYTICAL PROCESSING (OLAP)

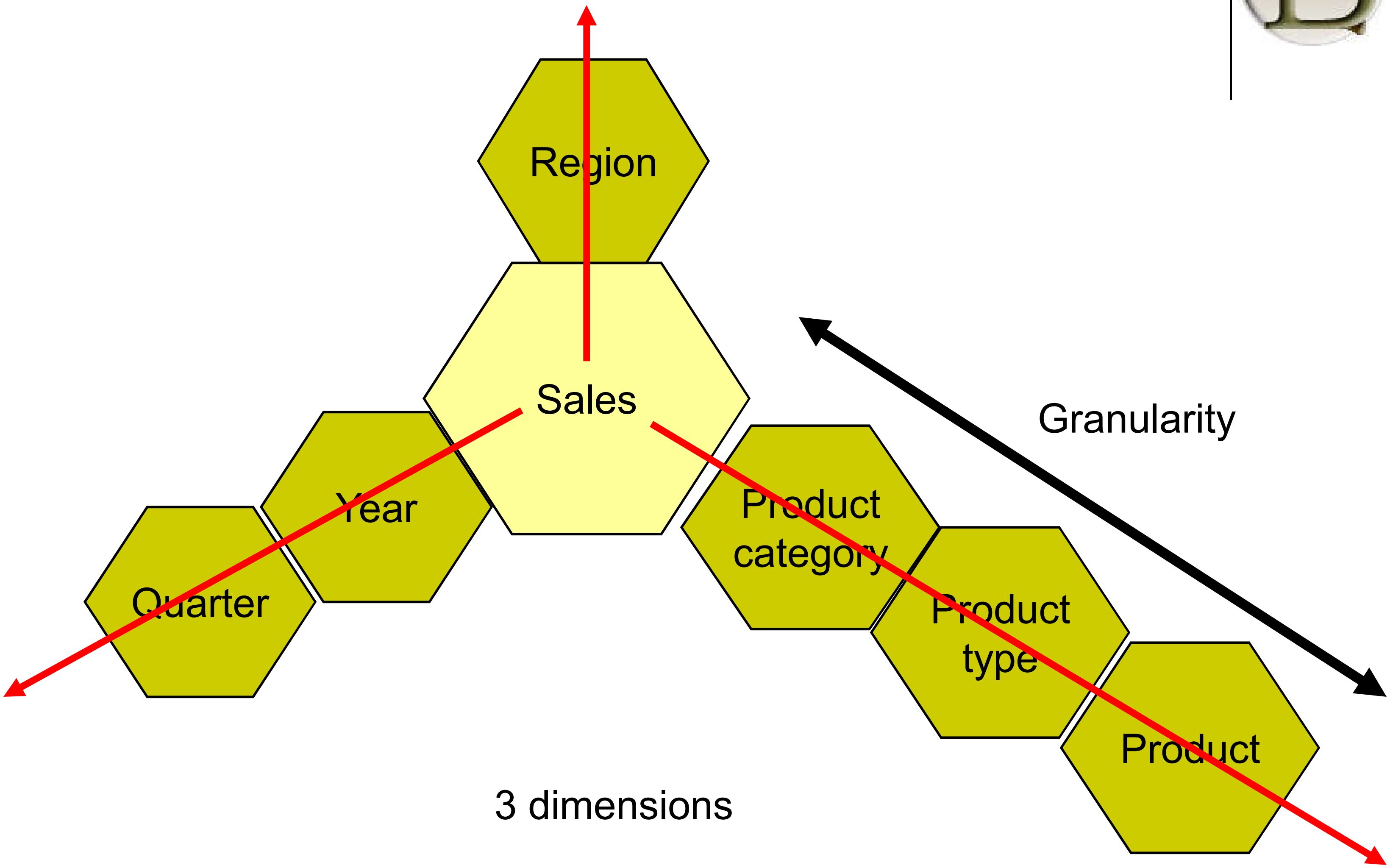
OLAP

WHAT IS OLAP?

DEFINITION :

‘OLAP applications and tools are those that are designed to ask ad hoc, complex queries of large multidimensional collections of data. It is for this reason that OLAP is often mentioned in the context of Data Warehouses’.

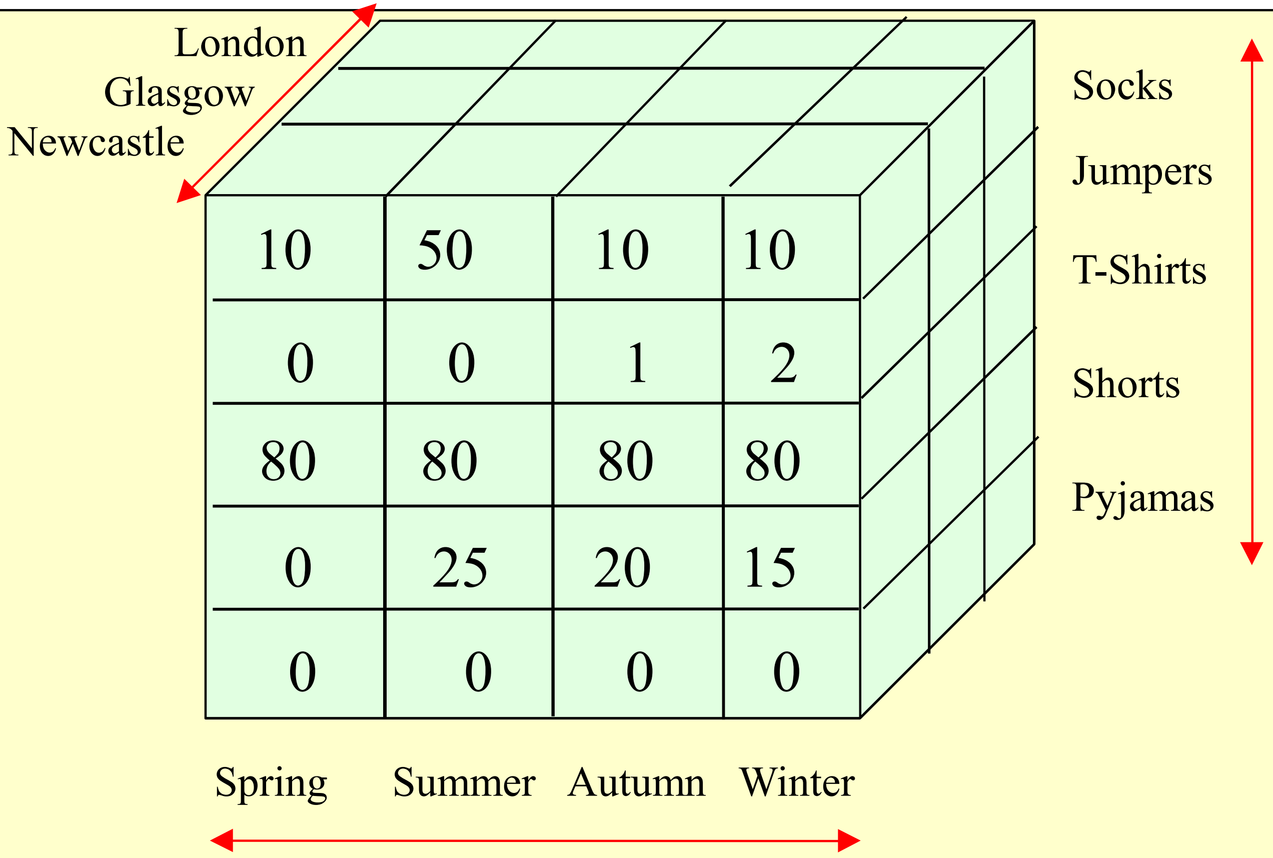
The Multidimensional Idea



3 dimensions

OLAP

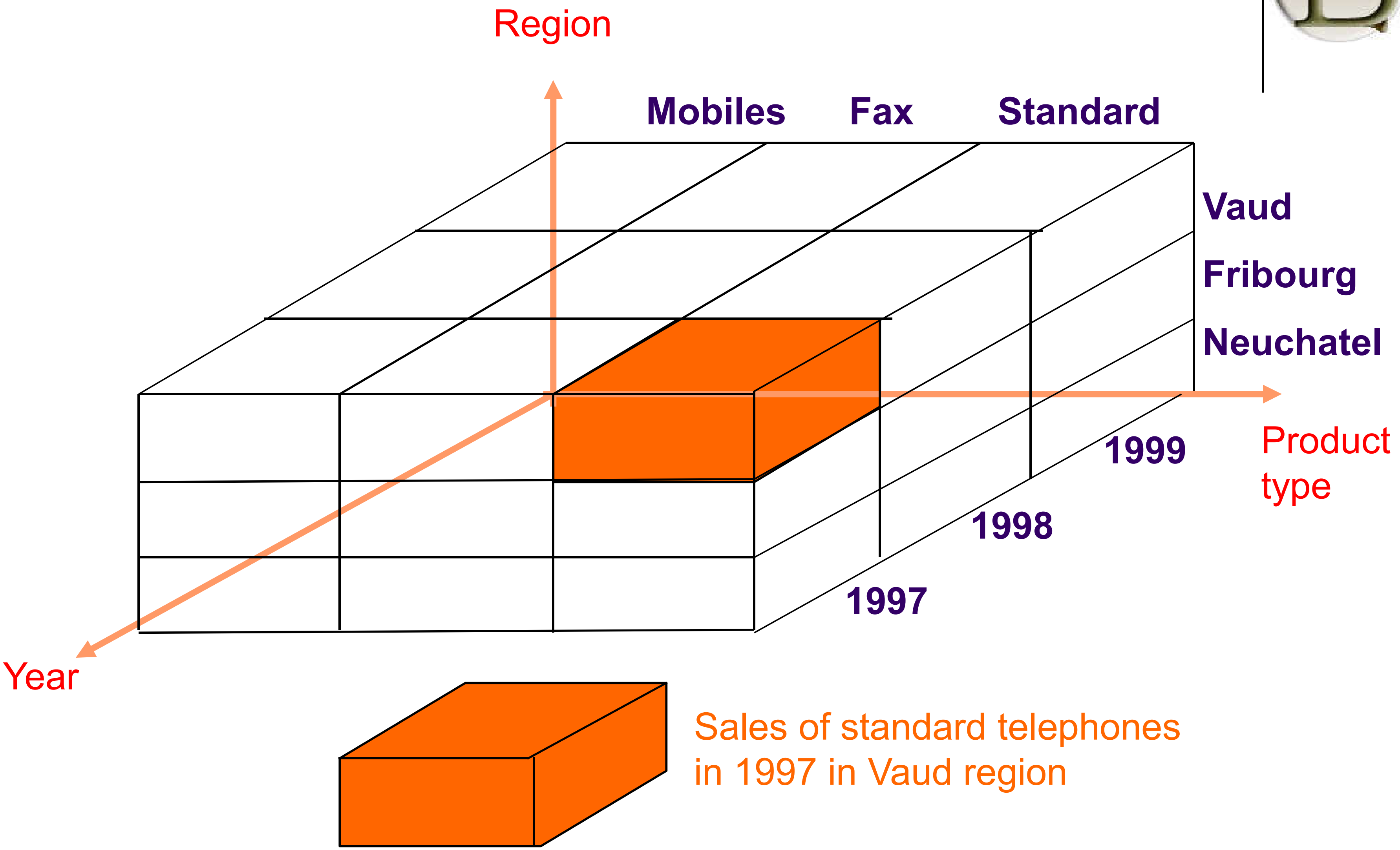
MULTDIMENSIONAL DATA MODEL



	Spring	Summer	Autumn	Winter
Socks	10	50	10	10
Jumpers	0	0	1	2
T-Shirts	80	80	80	80
Shorts	0	25	20	15
Pyjamas	0	0	0	0

Example: Three dimensions – Product, Sales, Area, and Season

Storage: The Cube



Sales of standard telephones in 1997 in Vaud region

OLAP Terminology

- A **data cube** supports viewing/modelling of a variable (a set of variables) of interest. **Measures** are used to report the values of the particular variable with respect to a given set of dimensions.

- A **fact table** stores measures as well as keys representing relationships to various dimensions.

- **Dimensions** are perspectives with respect to which an organization wants to keep record.

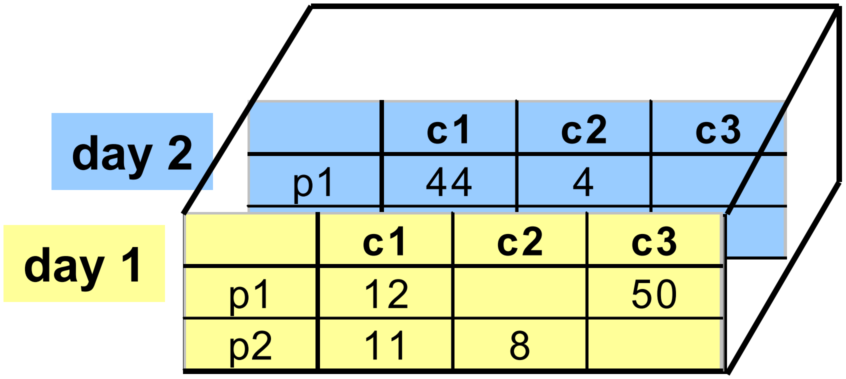
- A **star schema** defines a fact table and its associated dimensions.

3-D Cube

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Multi-dimensional cube:



day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

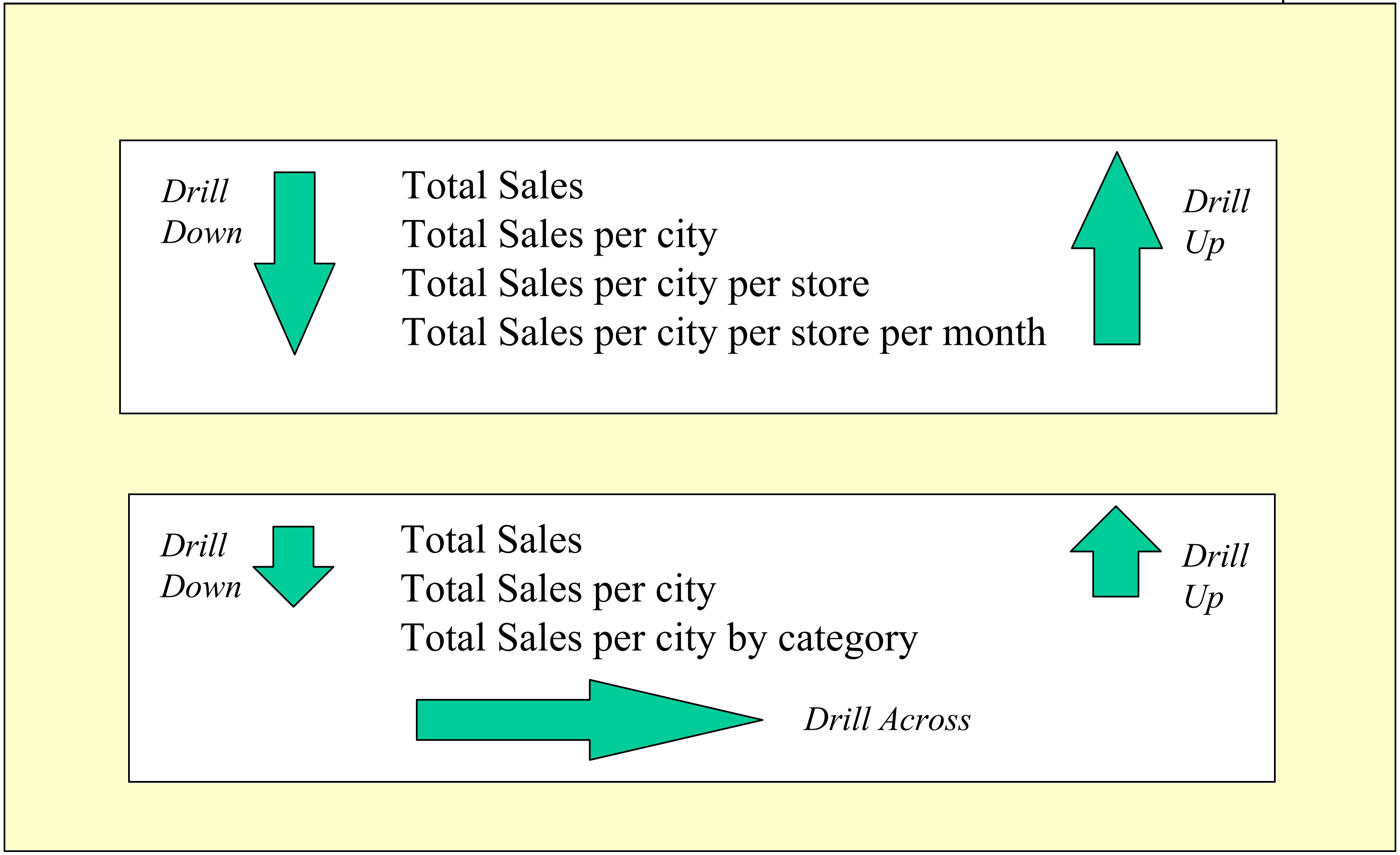
dimensions = 3

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice:
  - *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes.*
- Other operations
  - *drill across: involving (across) more than one fact table*

OLAP

TYPICAL OLAP OPERATIONS



Drill Down

Total Sales
Total Sales per city
Total Sales per city per store
Total Sales per city per store per month

Drill Up

Drill Down

Total Sales
Total Sales per city
Total Sales per city by category

Drill Up

Drill Across

By a drill up opperation examine sales By country rather than city level

	location by city			
	Istanbul	Ankara	Berlin	Münih
PC	20	30	50	40
Printer	15	5	10	20

roll up →

	location y country	
	Türkiyy	Almanya
PC	50	90
Printer	20	30

	measure is sales	
	Time 2002	
PC	50	
Printer	23	

Drill down →

	2002			
	Q1	Q2	Q3	Q4
PC	10	15	20	5
Printer	5	10	5	3

- when performed by dimension reduction
  - one or more dimensions are removed from the cube
- Ex a sales cube with location and time
  - roll-up may remove the time dimension
  - aggregation of total sales by location
    - rather than by location and by time

	location by country	
	Türkiye	Almanya
PC	50	90
Printer	20	30

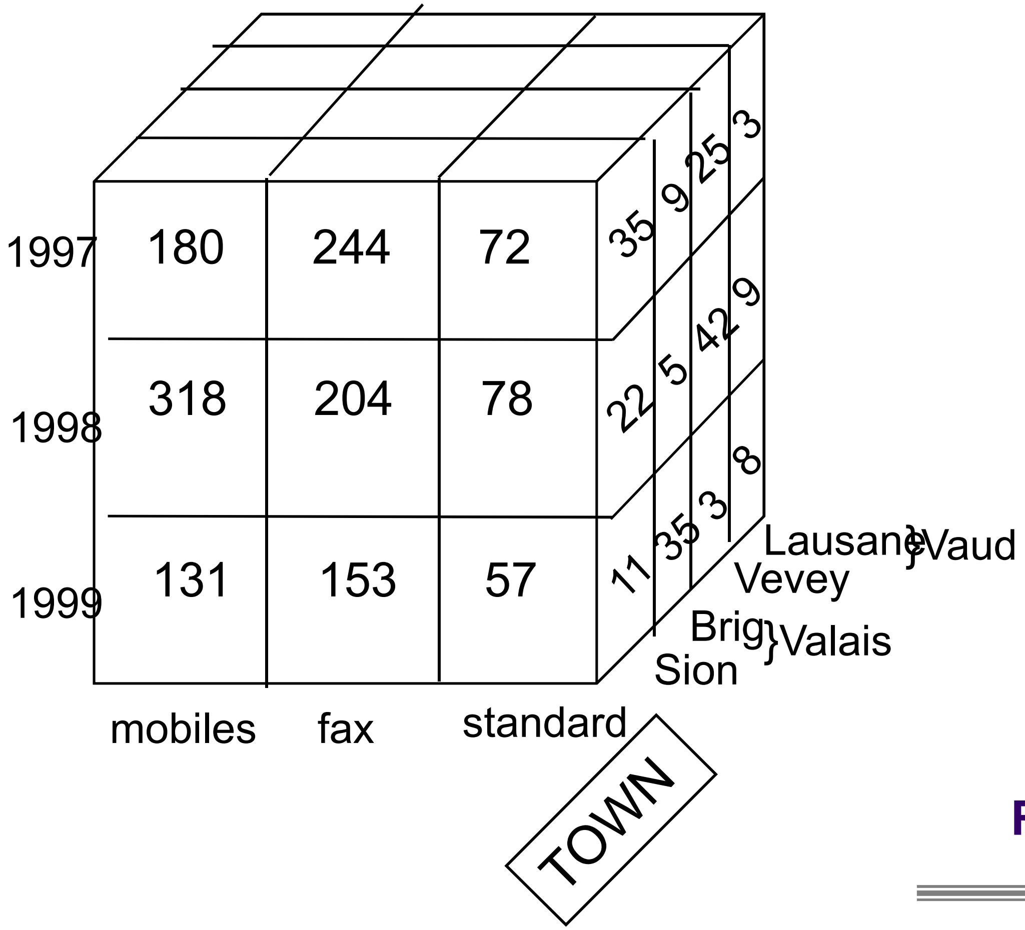
Two dimensional cuboid

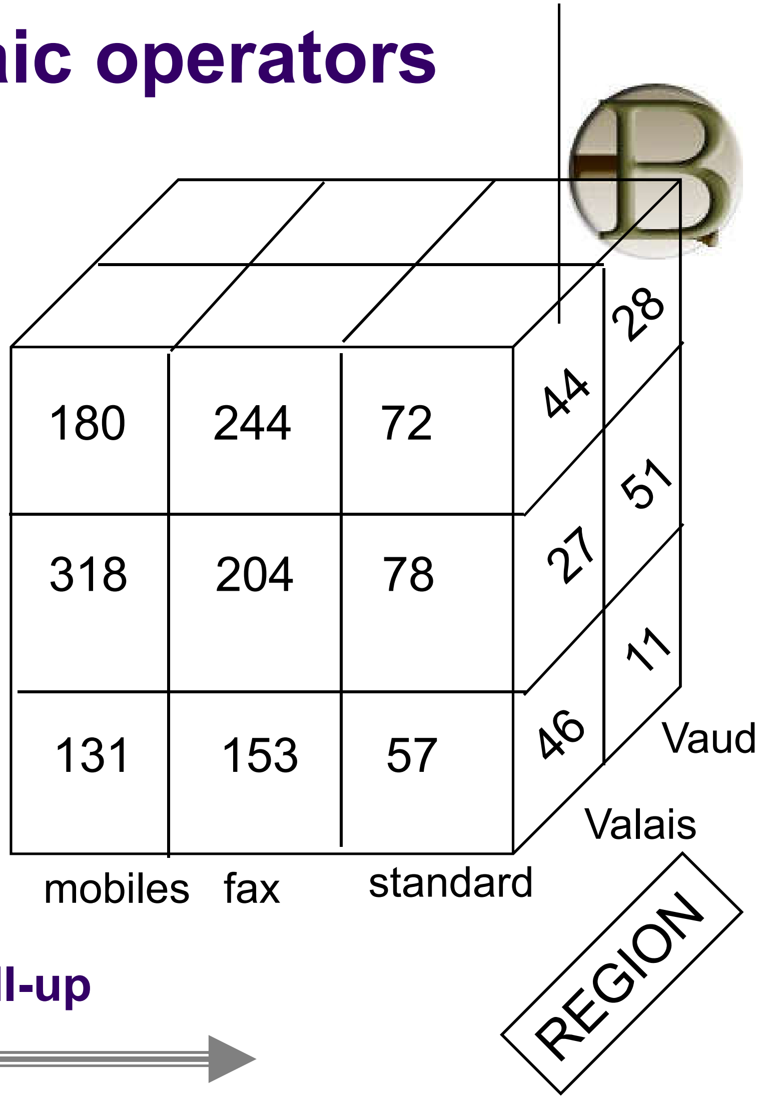
→

	locat All
PC	140
Printer	50

One dim. cuboid

Roll-up and Drill-down algebraic operators





Roll-up

Less detailed: go up in the granularity hierarchy

Drill-down

More detailed: go down in the granularity hierarchy

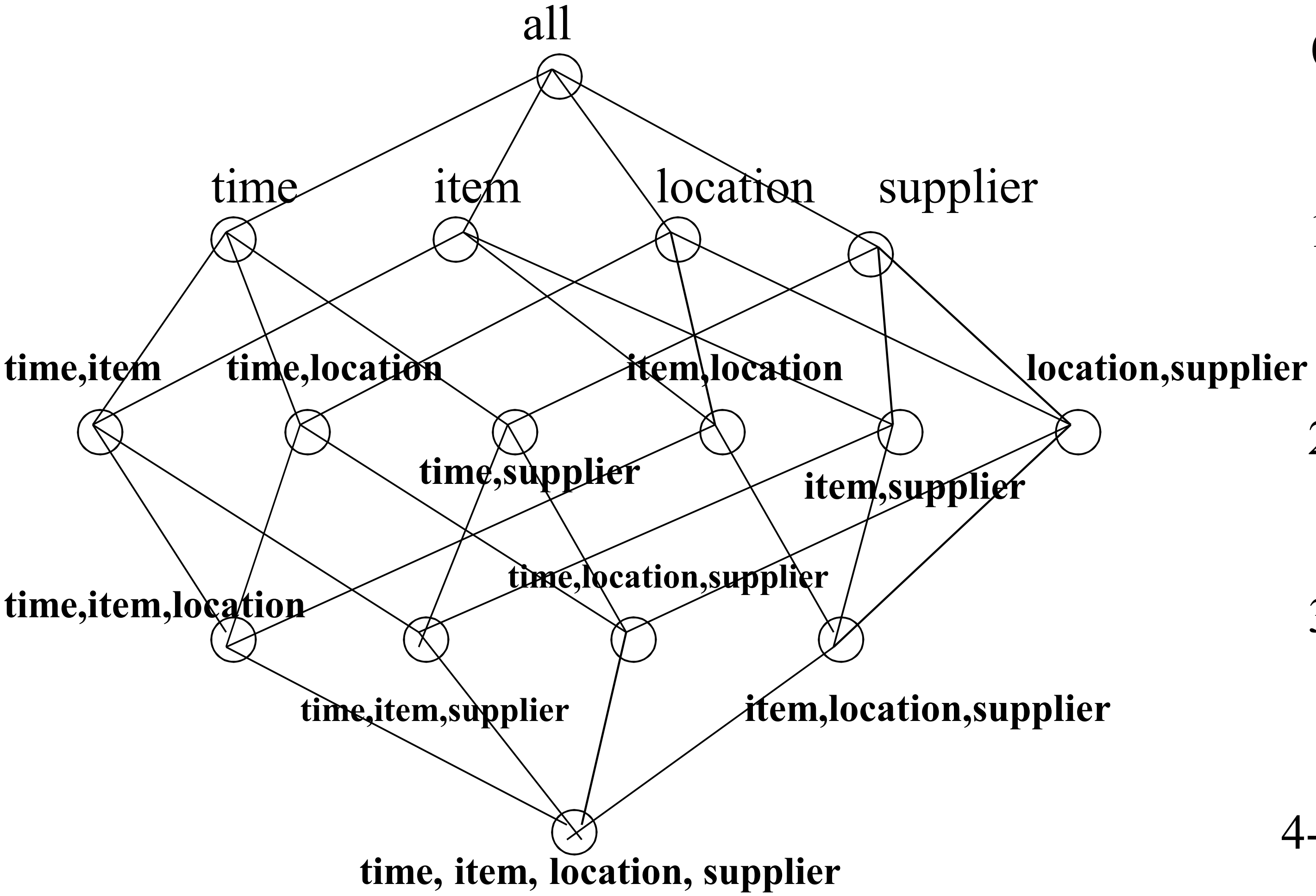
Slice and dice

- Slice: a selection on one dimension of the cube resulting in subcube
- Ex: sales data are selected for dimension time using time =spring
- dice: defines a subcube by performing a selection on two or more dimensions
- Ex: a dice opp. Based on
  - location="london" or "glasgow" and
  - time =spring or summer and
  - item = "T-shirts" or "Pyjamas"

N-DİMENSİONAL CUBE

- A data cube is referred to as a **cuboid**
- The lattice of cuboids forms a data cube.
- The cuboid holding the lowest level of summarization is called a base cuboid.
  - the 4-D cuboid is the base cuboid for the given four dimensions
- The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid.
  - typically denoted by all

Cube: A Lattice of Cuboids



0-D(apex) cuboid

1-D cuboids

2-D cuboids

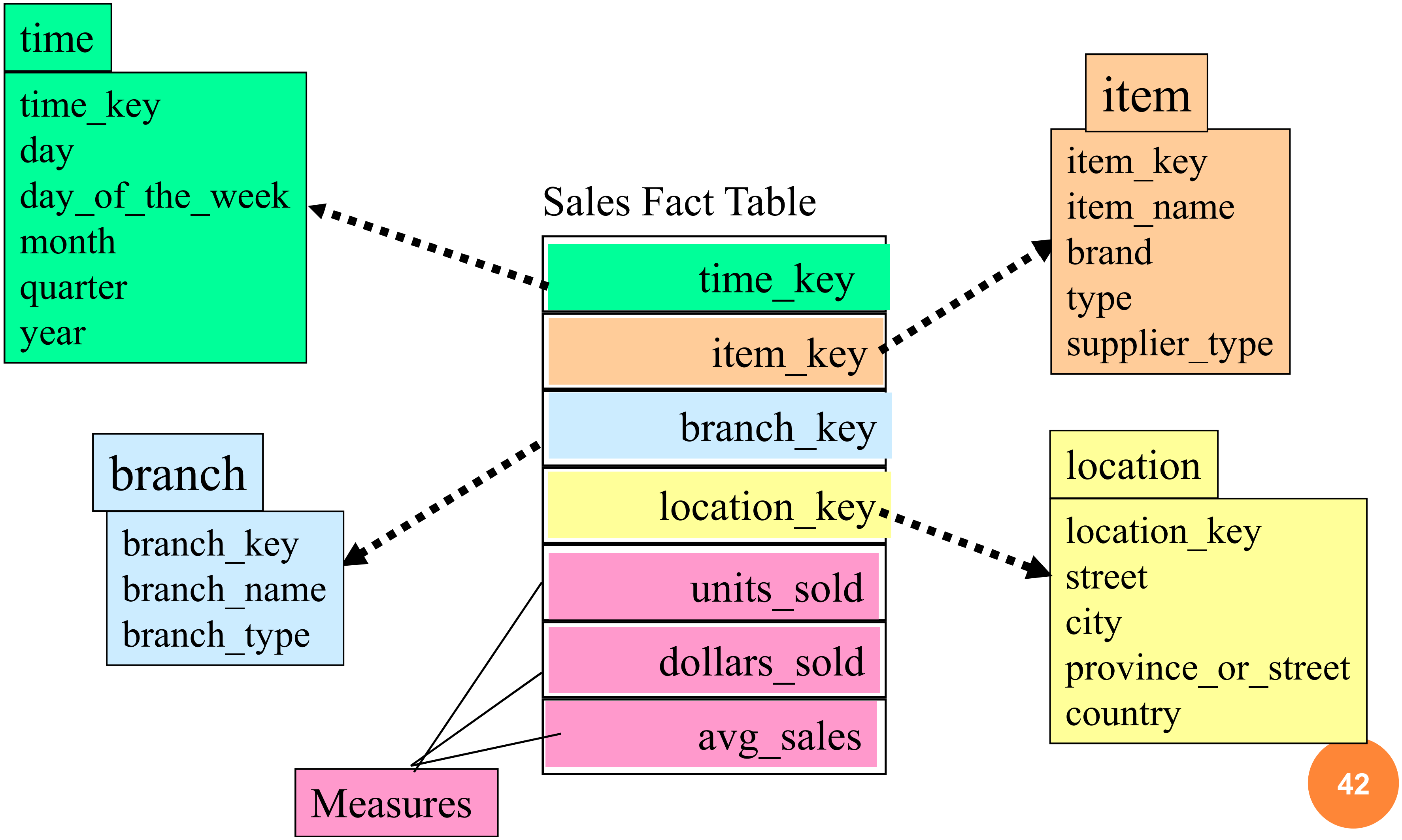
3-D cuboids

4-D(base) cuboid

CONCEPTUAL MODELING OF DATA WAREHOUSES

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellation: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

EXAMPLE OF STAR SCHEMA



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

location

location_key
street
city
province_or_street
country

Measures

DEFINING A STAR SCHEMA IN DMQL

- Cube Definition (Fact Table)

  define cube <cube_name> [<dimension_list>]:   <measure_list>

- Dimension Definition (Dimension Table)

  define dimension <dimension_name> as
    (<attribute_or_subdimension_list>)

define cube sales_star [time, item, branch, location]:
  dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars), units_sold = count(*)

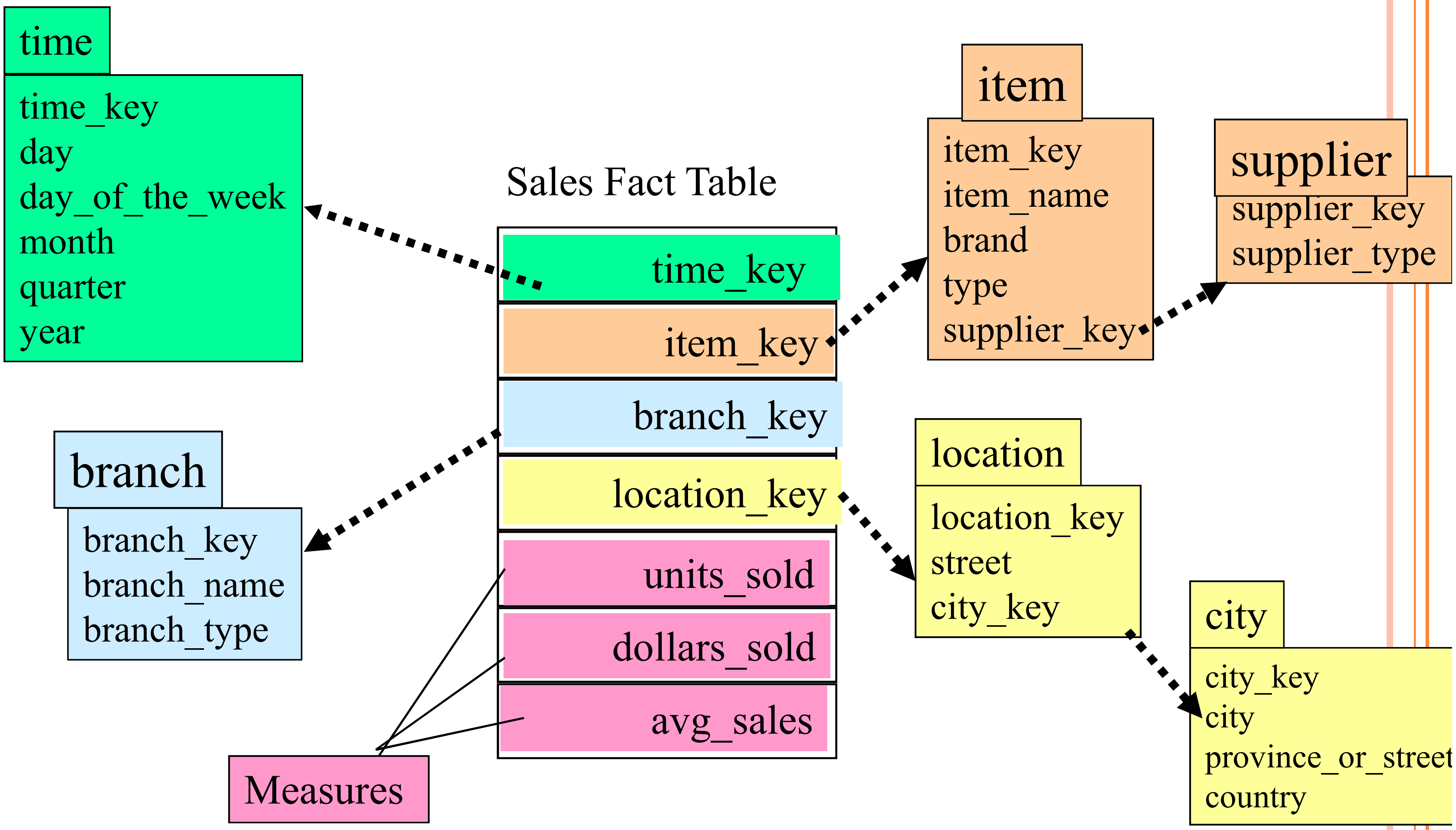
define dimension time as (time_key, day, day_of_week, month, quarter, year)

define dimension item as (item_key, item_name, brand, type, supplier_type)

define dimension branch as (branch_key, branch_name, branch_type)

define dimension location as (location_key, street, city, province_or_state, country)

Example of Snowflake Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_key

supplier

supplier_key
supplier_type

branch

branch_key
branch_name
branch_type

location

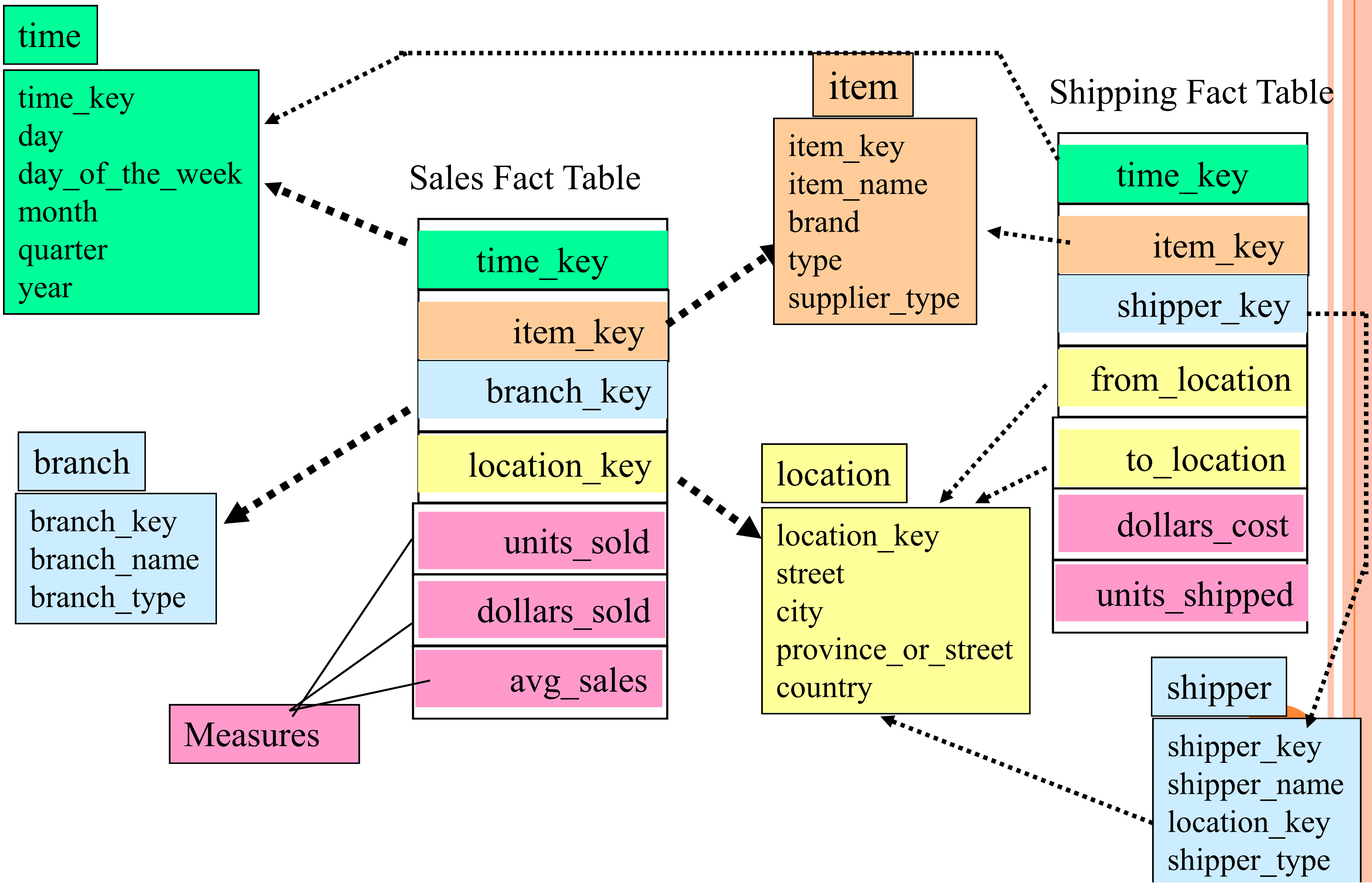
location_key
street
city_key

city

city_key
city
province_or_street
country

Measures

Example of Fact Constellation



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

branch

branch_key
branch_name
branch_type

item

item_key
item_name
brand
type
supplier_type

location

location_key
street
city
province_or_street
country

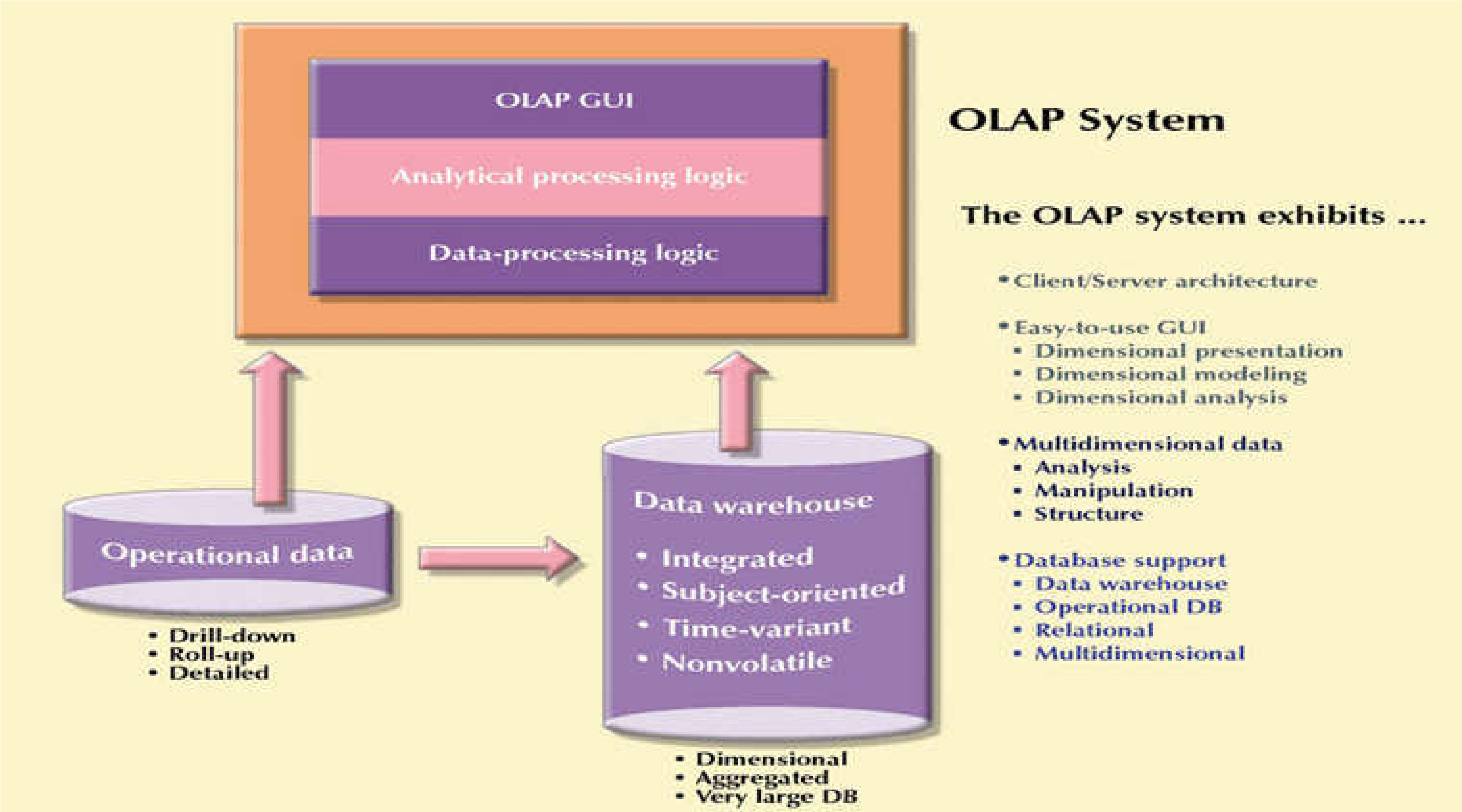
Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

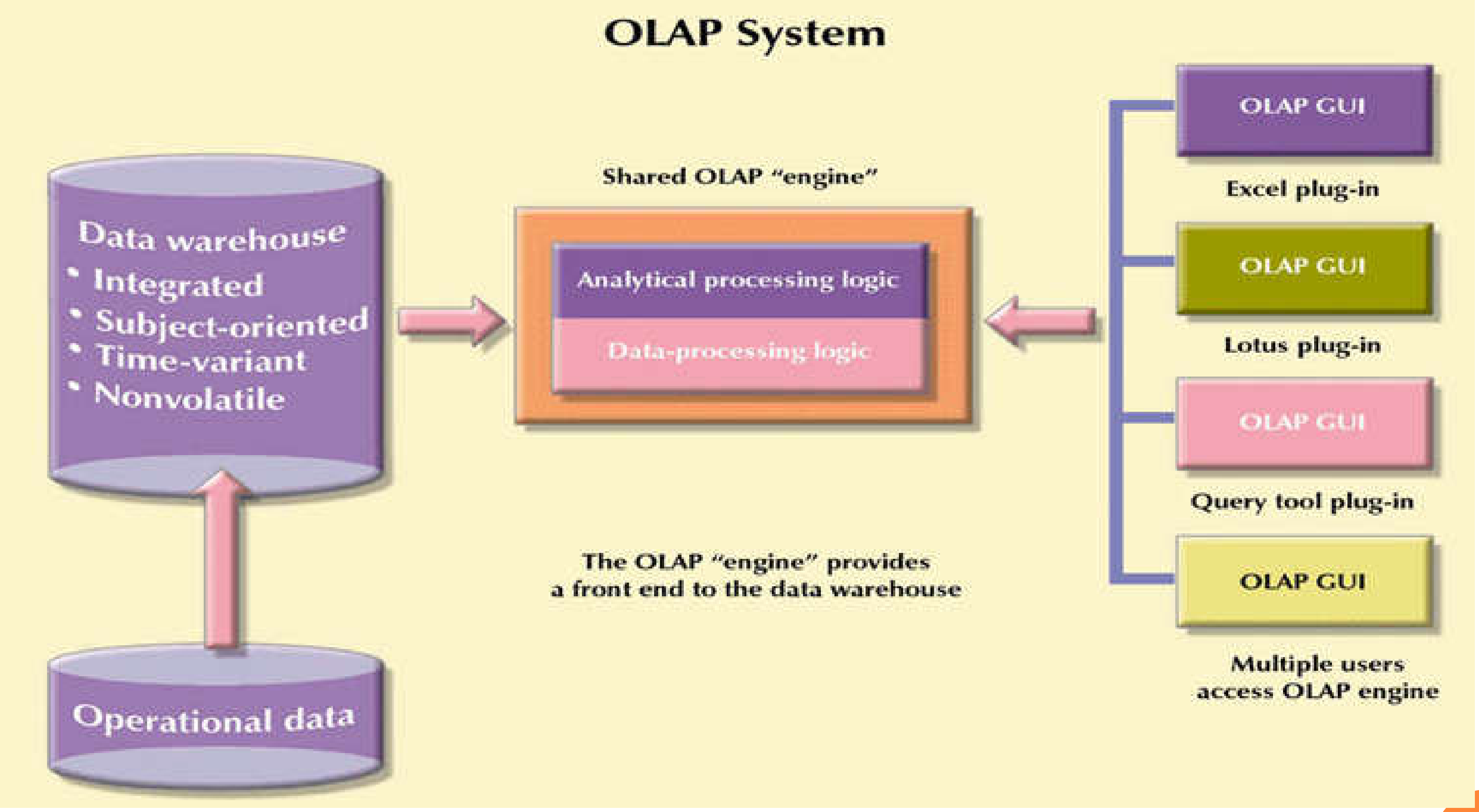
shipper

shipper_key
shipper_name
location_key
shipper_type

OLAP Client/Server Architecture



OLAP Server Arrangement



OLTP VS. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

OLTP –VS- OLAP

- On Line Transaction Processing -- OLTP
  - Maintain a database that is an accurate model of some real-world enterprise
    - Short simple transactions
    - Relatively frequent updates
    - Transactions access only a small fraction of the database
- On Line Analytic Processing -- OLAP
  - Use information in database to guide strategic decisions
    - Complex queries
    - Infrequent updates (Load)
    - Transactions access a large fraction of the database

Business Information:

“How you gather, manage,
and use information
will determine whether
you win or lose.”

– Bill Gates

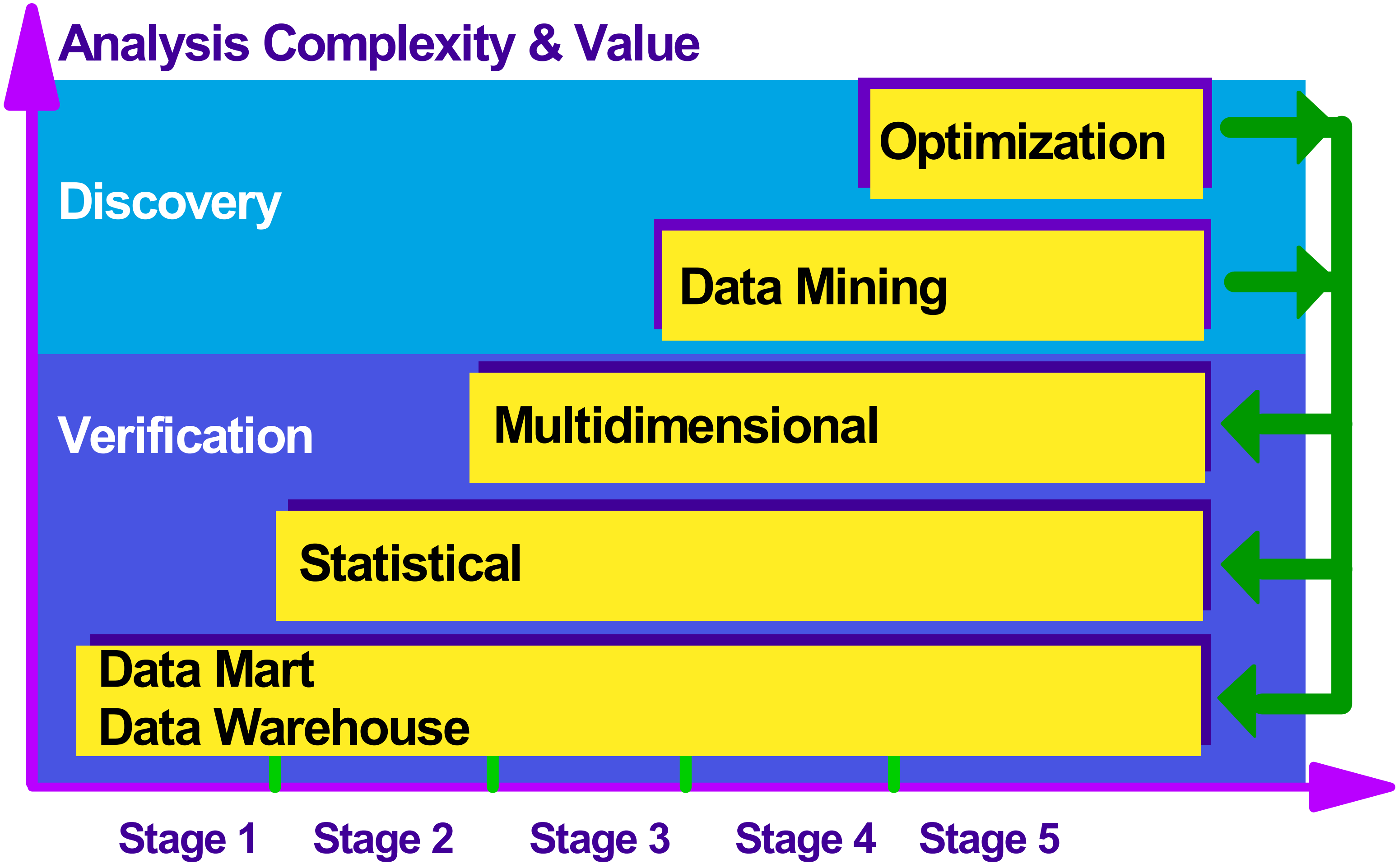
What is BI?

- Business Intelligence means using your data assets to make better business decisions.

- Business intelligence involves the gathering, management, and analysis of data for the purpose of turning that data into useful information which is then used to improve decision making.

- Organizations can then make more strategic decisions about how to administer clients and programs. These practices can also reduce operating costs through more effective financial analysis, risk management, and fraud management.

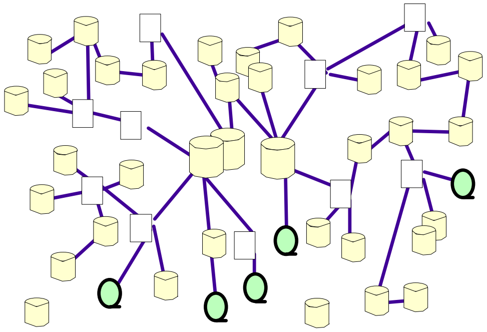
Business Intelligence solutions start with data warehouses and data marts



Data… Information….Decisions

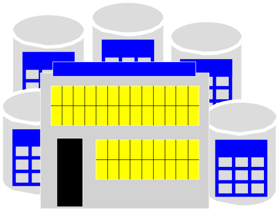
Data to Information to Decisions

Data



- On-line Updates
- Batch Feeds
- Operational Data Store

Information Management



- Data Warehouse
- Data Mart
- Data Transformation
- Data Synchronization

Access

- Query & Reporting
- Data Mining
- On-line Analytical Processing
- Summary and detail
- Drill capability

Knowledge discovery in databases

- KDD is the process of identifying valid,potentially useful and understandable patterns & relationships in data

⌦ Knowledge = patterns & relationships

knowledge discovery =

data preparation + data mining + evaluation/interpretation of discovered patterns/relationships

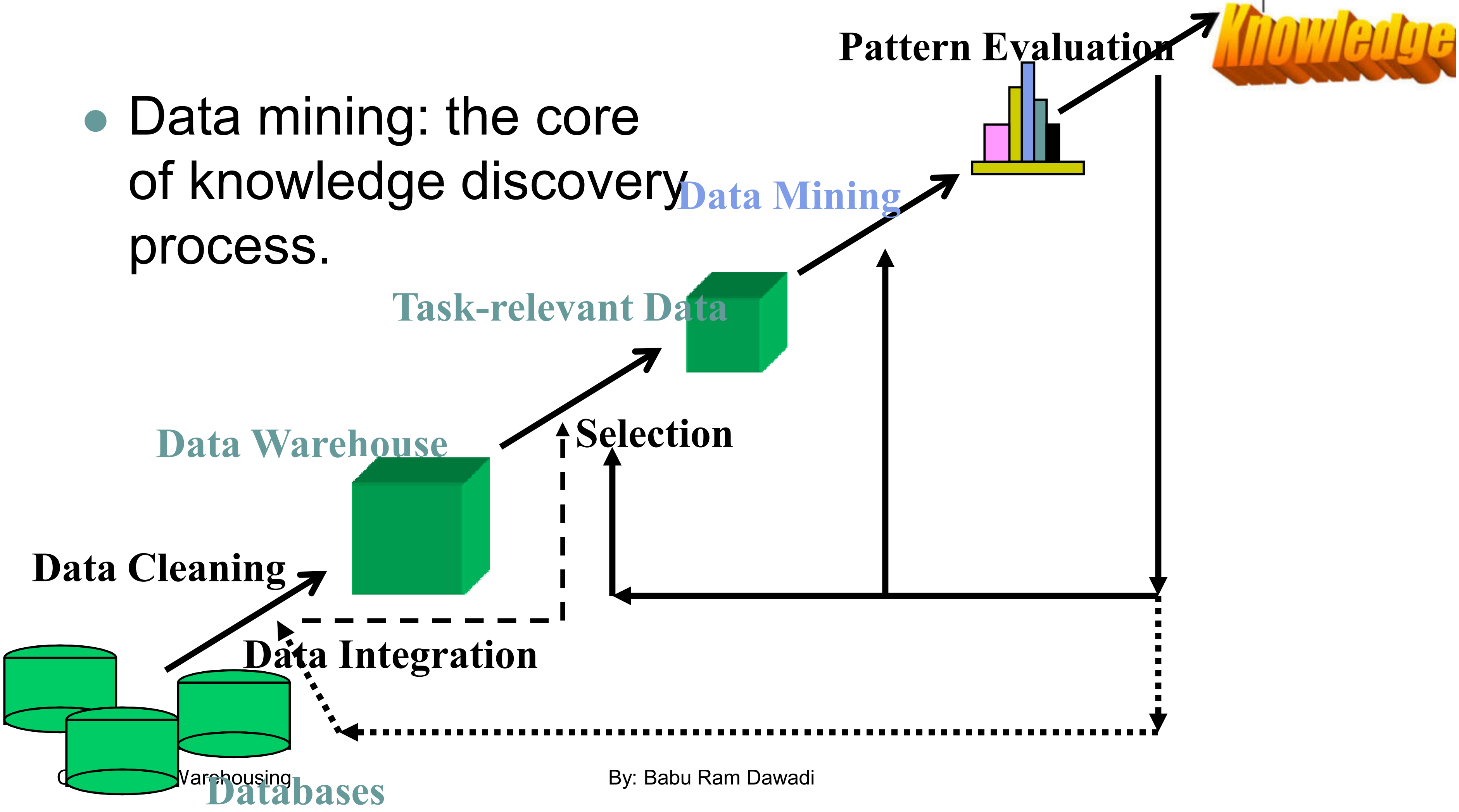
- Nowadays, ⌦ KDD = data mining

Knowledge Discovery in Database Environment (Stages)

- There are six stages of KDD which are:
  - Data selection
  - Cleaning
  - Enrichment
  - Coding
  - Data mining
  - reporting

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



KDD : Data selection

- Data Selection
  - It is the first stage of KDD process in which we collect and select the data set or database required to work with
  - Data sets are obtained from operational databases
  - Obtaining information from centralized databases can be difficult, reasons may be:
    - Data set may need conversion from one format to another
      - Eg: Excel files to access files

KDD: Data selection

- Different quality of data in different parts are available
- Making choice on right data is important
- Investigations should be made on any data warehouses available in an organization.
- A well maintained DW helps to make data selection job convenient by providing right data set necessary for analysis.

- Data Cleaning
  - This is the second stage of KDD.
  - Data set obtained is never perfectly cleaned.
  - We may not be aware of to what extent it is polluted.

KDD: Data selection

- Data in real world is dirty:
  - Incomplete: lack attribute values
  - Noisy: contains errors
    - Human errors
    - Not available when collected
    - Not entered due to misunderstanding
    - Malfunction of hardware/software
    - Mistake data entry
  - Inconsistent: contains discrepancies codes
- The cleaning phenomena should try to eliminate all the above mentioned defects by the stage of de-duplication, domain consistency, disambiguation

KDD …

- Enrichment
  - Enrichment is the process of adding additional information to the databases or accessing additional databases to obtain extra information.
  - Eg: an airline company might cooperate with telephone company to enhance its marketing policy. A telephone company maintains large databases comprising the call behavior of customers & create telephone profiles of the basis of these data.

KDD …

  - These telephone profiles could be used by airlines to identify interesting new groups of target customers
  - So data miners can collect all the necessary information from additional bought – in databases.
  - Obtaining information from other organizations may involve some tedious procedures.
- Coding:
  - Coding is one of the most important stage where further cleaning and transformation of data is done.

KDD …

- Coding…
  - It can range from simple SQL Queries to using sophisticated high level languages depending upon requirement.
  - Some polluted records can be easily filtered out by using SQL queries. (Eg: Records with most of the field empty can be easily traced and removed)
  - Coding is the creative activity which involves creative transformation of data.
  - It can be used to obtained more simpler form of the complete, detailed database.

KDD: Coding

- Coding…
  - Example: to the table of the magazine publisher, we can apply following coding steps:
    - Convert address to region (area codes)
    - Birth date to age
    - Divide income by 1000
    - Divide credit by 1000
    - Convert owners yes/no to 1/0
    - Convert purchase date to month starting from 1990
    - Perform filtering

KDD: Coding

- Coding: applying steps 1 to 6

Client No	Age	income	Credit	Car owner	H. Woner	Region	Month of purchase	Mag. purchased
203	20	18.5	17.8	0	0	1	52	Car
203	20	18.5	17.8	0	0	1	42	Music
209	25	36.0	26.6	1	0	1	Null	Comic
203	20	18.5	17.8	0	0	1	48	house

KDD: Coding

- Coding: applying step 7

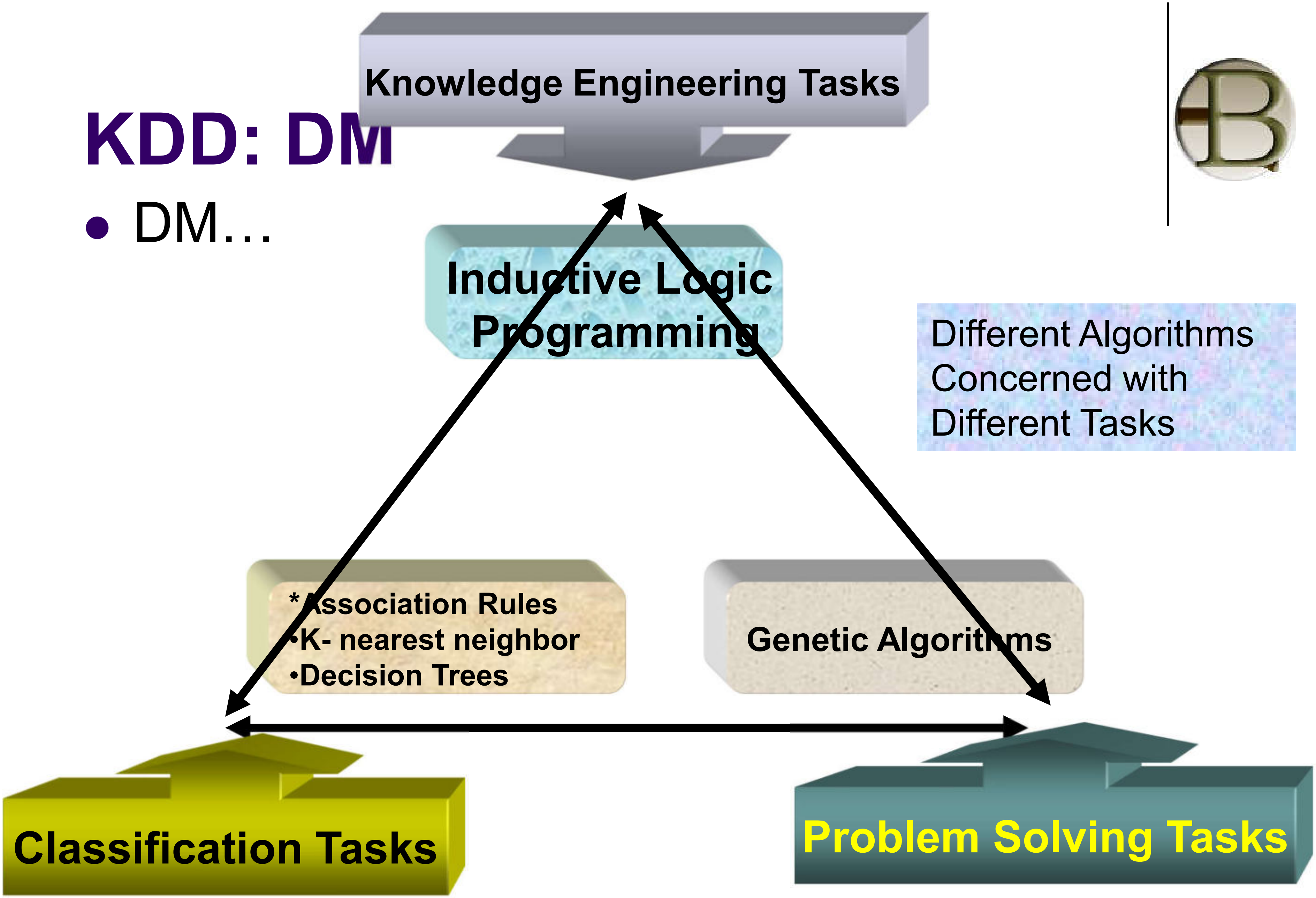
Client NO	Age	Income	Credit	Car owner	House owner	Region	Car. Mag.	House Mag.	Sport Mag.	Music Mag.	Comic Mag.
203	20	18.5	17.8	0	0	1	1	1	0	1	0
209	25	36.0	26.6	1	0	1	0	0	0	0	1

KDD: DataMining

- Data Mining:
  - All the cleanings, transformations and enrichment are performed on data, so that we can extract the most useful information from it, and this is performed in data mining stage of KDD.
  - It consists of different rules, techniques, and algorithms used for mining purpose.
  - These are involved in performing following three tasks:
    - Knowledge Engineering Tasks
    - Classification tasks
    - Problem solving tasks

KDD: DM

- DM…



Different Algorithms Concerned with Different Tasks

KDD: Data mining

- Knowledge engineering:
  - is the process of finding right formal representation of certain body of knowledge in order to represent it in a knowledge based system
    - Eg: Expert Systems (medical diagnostic system)
- Classification tasks:
  - Classification is the process of dividing data into no. of classes. Eg: class of customers
- Problem Solving Tasks:
  - It involves finding solutions of remedies to the problems that arise. Eg: why are people not going to cinema hall?

KDD…

- For finding useful patterns in databases, it is necessary to choose right algorithms and right tools.
- For choosing right data mining algorithms following three points should be considered:
  - Quality of input [No. of records, attributes, numeric]
  - Quality of output [yes/no results, statistics]
  - Performance [CPU load]

KDD: Reporting

- This stage involves documenting the results obtained from learning algorithms.
- Any report writer or graphical tools can be used
- It basically combines two functions:
  - Analysis of results obtained from mining.
  - Application of results to new data
- Different data visualization tools like scatter diagrams available for showing different patterns or clusters of data can be used.

Chapter 3: Data Preprocessing

- Preprocess Steps
  - Data cleaning
  - Data integration and transformation
  - Data reduction

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

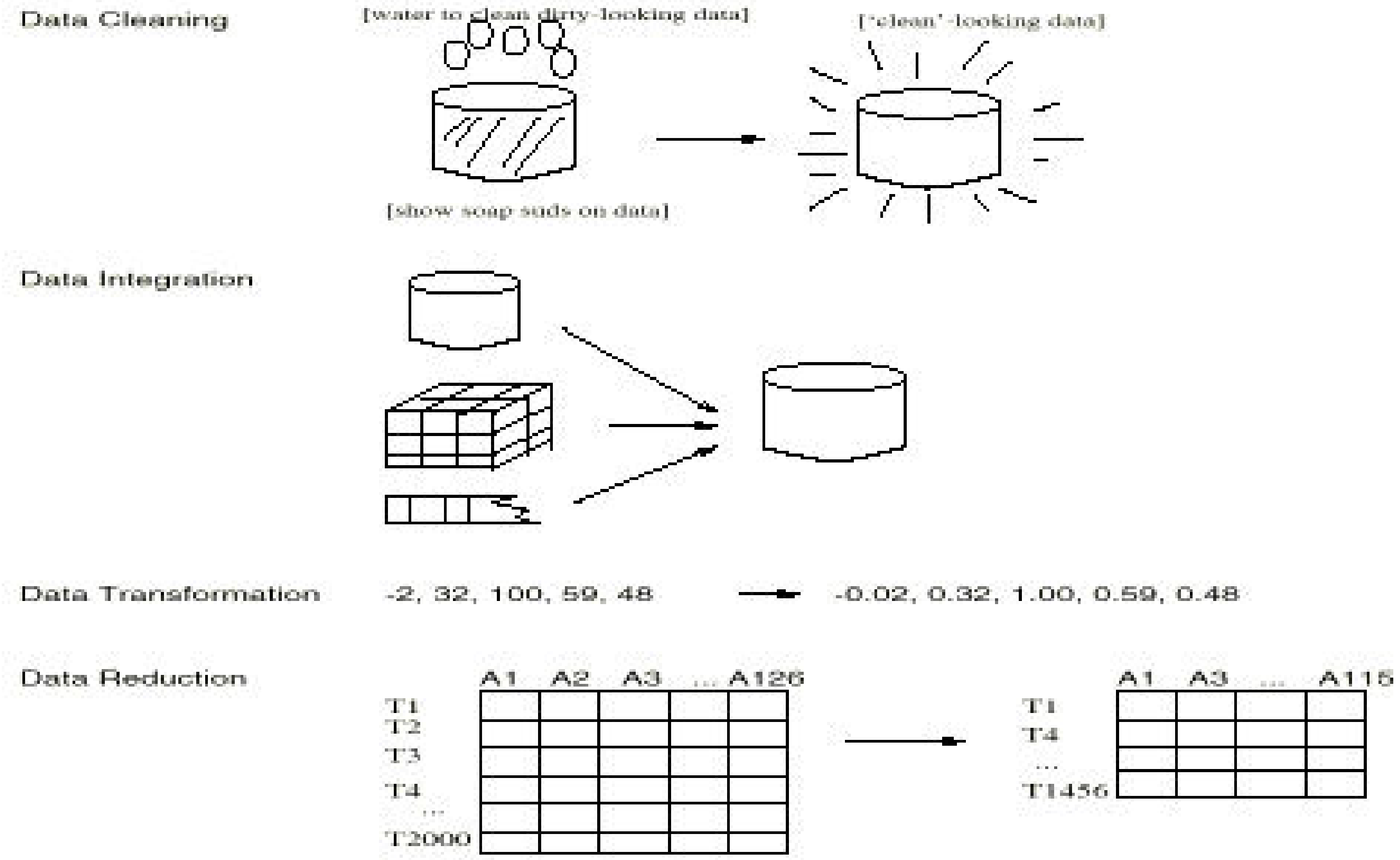
Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

Forms of data preprocessing



Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data

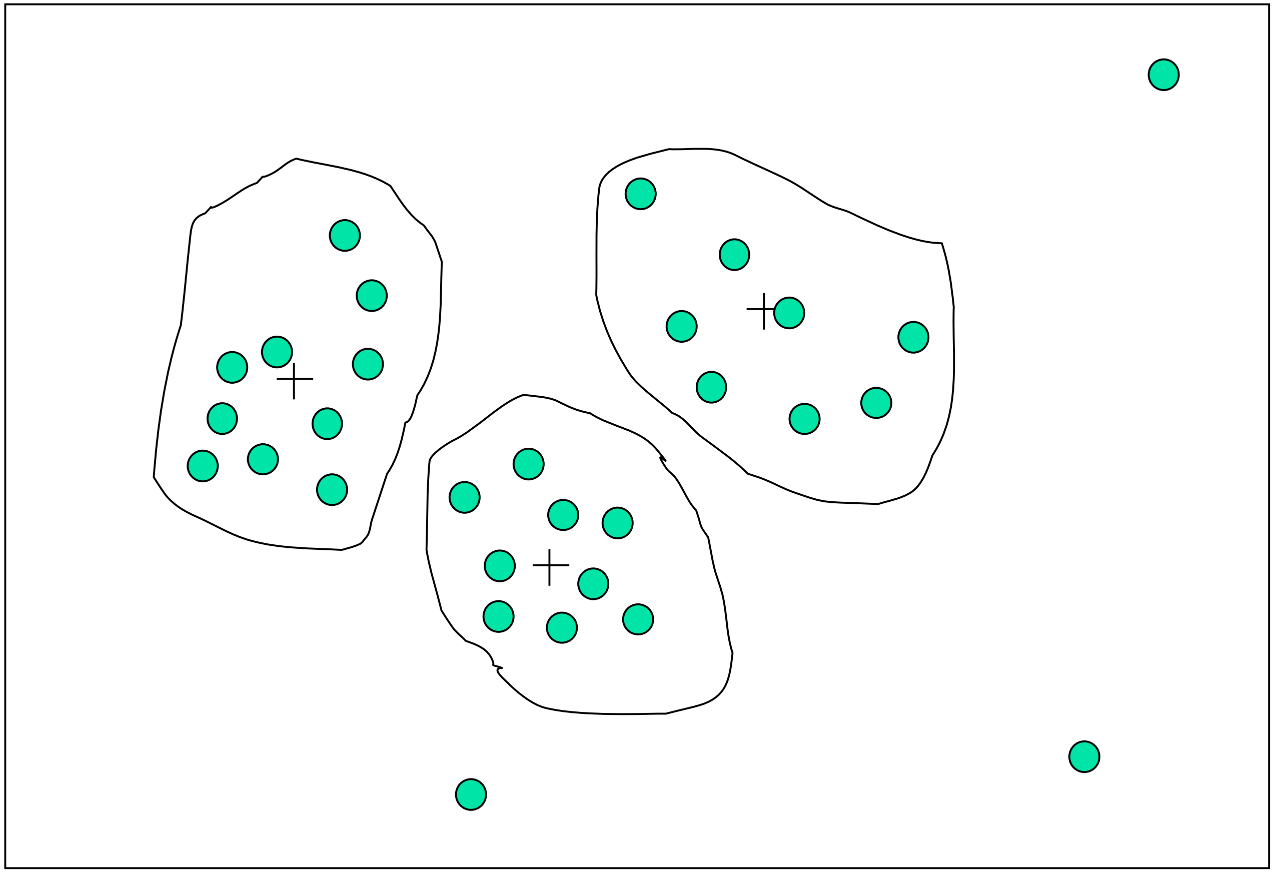
- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median
  - Equal-width (distance) partitioning:
    - It divides the range into *N* intervals of equal size: uniform grid
    - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
- Equal-depth (frequency) partitioning:
    - It divides the range into *N* intervals, each containing approximately same number of samples
    - Managing categorical attributes can be tricky.
- Combined computer and human inspection
  - detect suspicious values and check by human

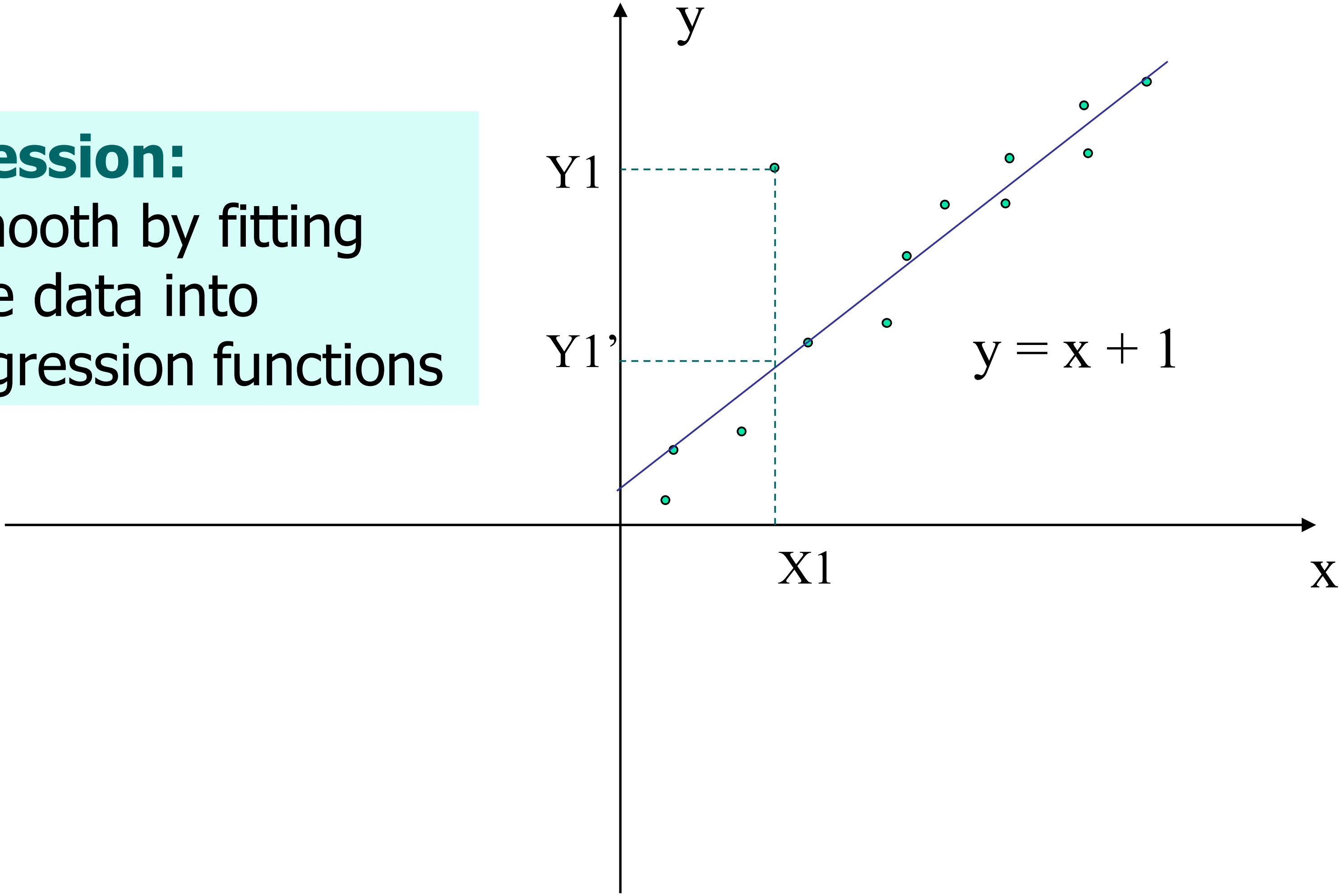
Cluster Analysis

Clustering: detect and remove outliers



Regression

Regression:
smooth by fitting the data into regression functions



Data Integration

- Data integration:
  - combines data from multiple sources.
  - Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id $\equiv$ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundant Data in Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table.
  - Redundant data may be able to be detected by correlational analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- **min-max normalization**
  - Min-max normalization performs a linear transformation on the original data.
  - Suppose that mina and maxa are the minimum and the maximum values for attribute A. Min-max normalization maps a value v of A to v' in the range [new-mina, new-maxa] by computing:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range $12,000 to $98,000 normalized to [0.0, 1.0]. Then $73600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

Data Transformation: Normalization

- **Z-score Normalization:**
  - In z-score normalization, attribute A are normalized based on the mean and standard deviation of A. a value v of A is normalized to v’ by computing:

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - where μ: mean, σ: standard deviation
  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
  - This method of normalization is useful when the actual minimum and maximum of attribute A are unknown.

Data Transformation: Normalization

Normalization by Decimal Scaling

- Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A.
- The number of decimal points moved depends on the maximum absolute value of A.
- a value v of A is normalized to v' by computing: v' = (v / 10j). Where j is the smallest integer such that Max(|v'|)<1.

Data Reduction Strategies

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- Data reduction
  - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction
  - Data Compression

Data Cube Aggregation

- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with

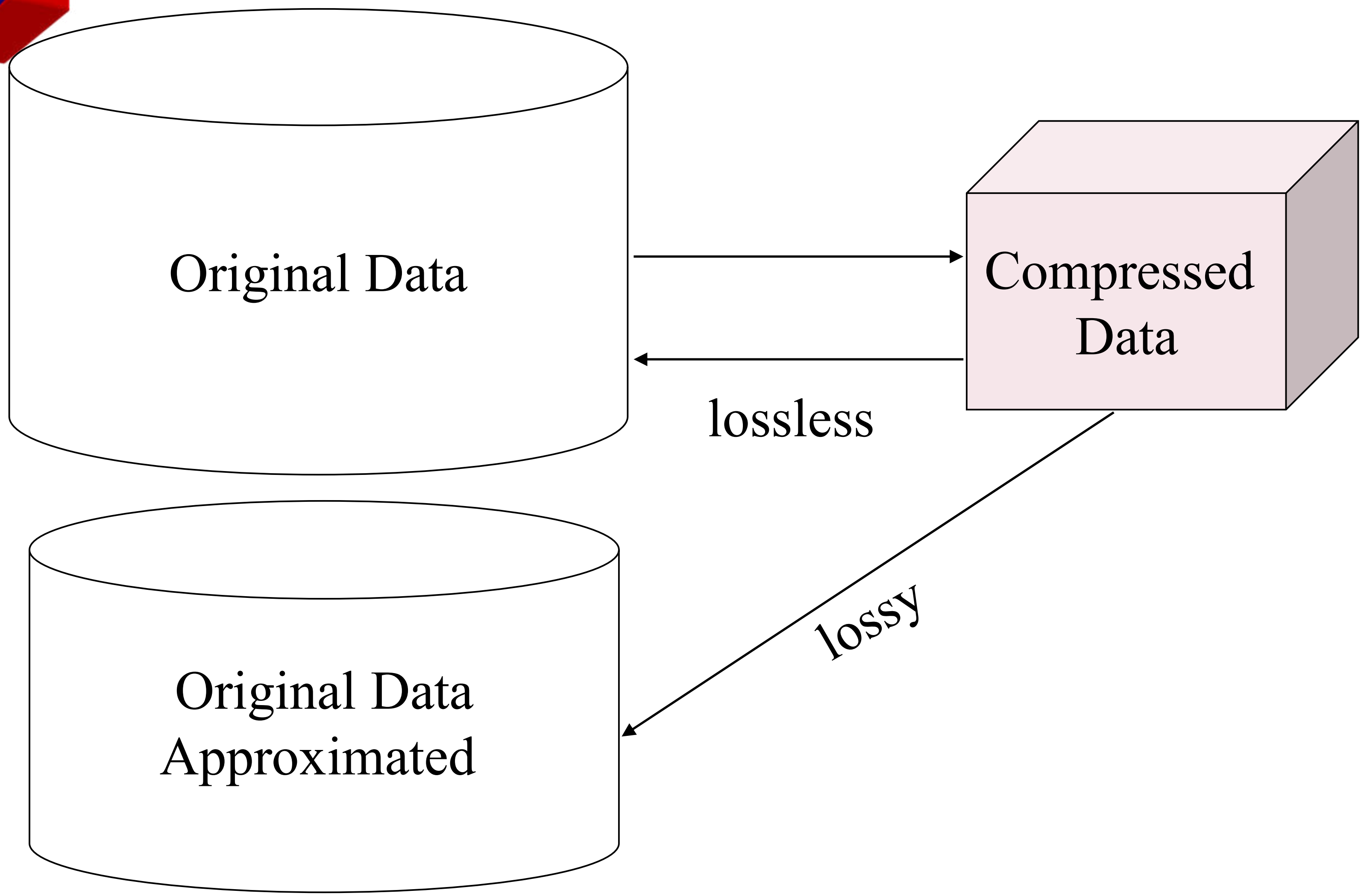
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task

- Queries regarding aggregated information should be answered using data cube, when possible

Data Compression

- String compression
  - Typically lossless
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Data Compression



Two Styles of Data Mining

- **Descriptive data mining**
  - characterize the general properties of the data in the database
  - finds patterns in data and the user determines which ones are important
- **Predictive data mining**
  - perform inference on the current data to make predictions
  - we know what to predict
- **Not mutually exclusive**
  - used together
  - Descriptive → predictive
- Eg. Customer segmentation – descriptive by clustering
- Followed by a risk assignment model – predictive by ANN

Descriptive Data Mining (1)

- Discovering new patterns inside the data
- Used during the data exploration steps
- Typical questions answered by descriptive data mining
  - what is in the data
  - what does it look like
  - are there any unusual patterns
  - what dose the data suggest for customer segmentation
- users may have no idea
  - which kind of patterns may be interesting

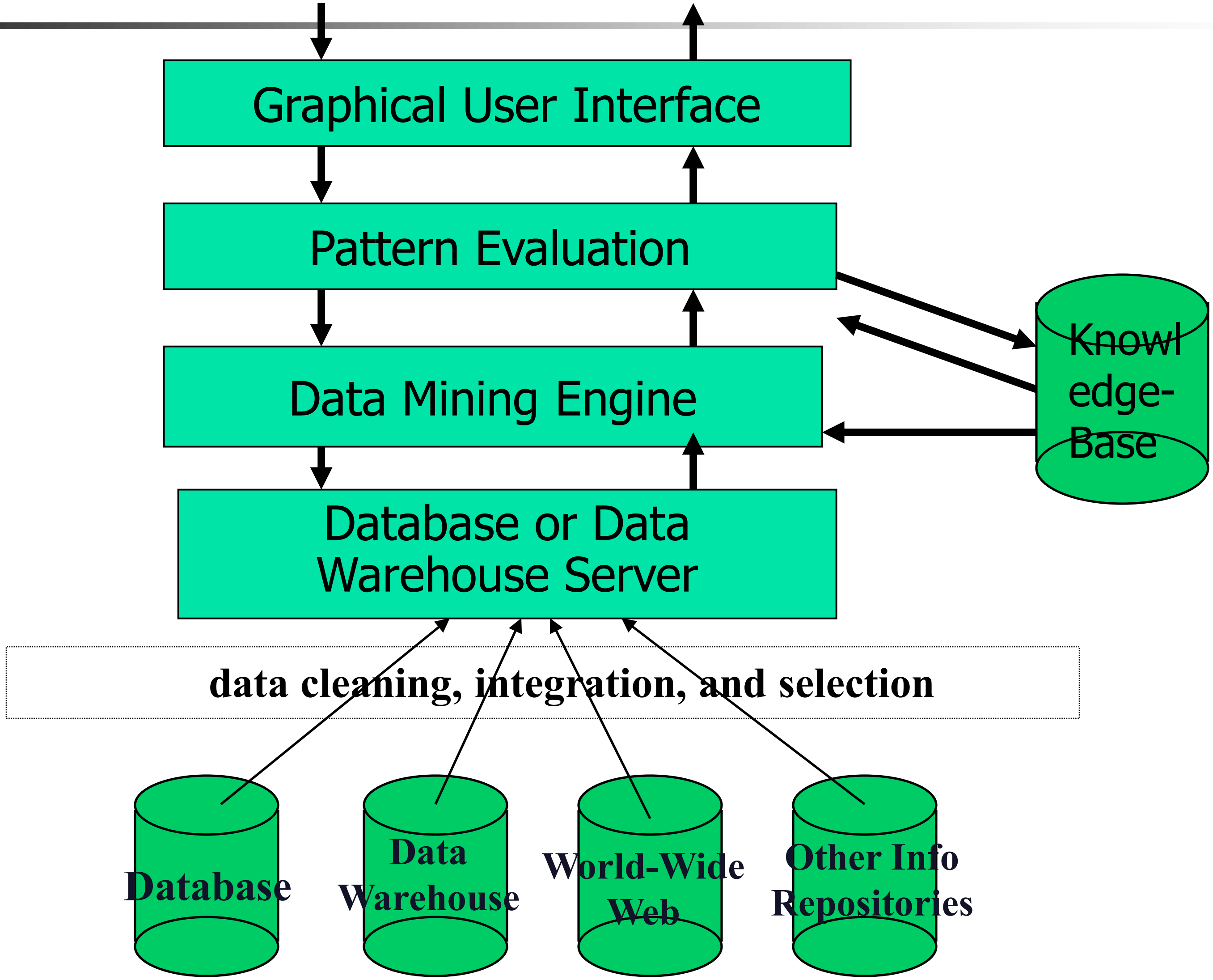
Descriptive Data Mining (2)

- patterns at verious granularities
  - geograph
    - country - city - region - street
  - student
    - university - faculty - department - minor
- Fuctionalities of descriptive data mining
  - Clustering
    - Ex: customer segmentation
  - summarization
  - visualization
  - Association
    - Ex: market basket analysis

Predictive Data Mining

- Using known examples the model is trained
  - the unknown function is *learned from data*
- the more data with known outcomes is available
  - the better the predictive power of the model
- Used to predict outcomes whose inputs are known but the output values are not realized yet
- Never 100% accurate
- Its performance on unknown data is much more important

Architecture: Typical Data Mining System



Architecture: DM system

- A good data mining architecture will help to make best use of software environment, perform DM tasks in efficient and timely manner, interoperate and exchange information with other information systems, be adaptable to varying user requirement.

- Data mining system architecture includes the consideration of coupling a data mining system with a database or data warehouse system

Coupling

- There are several possible designs such as no coupling, loose coupling, semi tight coupling and tight coupling

- No coupling
  - means that a data mining system will not utilize any function of a database or data warehouse system

  - It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file

Loose Coupling

- Loose coupling means that a data mining system will use some facilities of a database or data warehouse system
- Fetching data from a data repository managed by database or data warehouse system, and then storing the mining results either in a file or in a designated place in a database or data warehouse

Semi tight coupling

- besides linking a data mining system to database or data warehouse system, efficient implementations of a few essential data mining primitives can be provided in the database or data warehouse system

- These primitives can include sorting, indexing, aggregation, histogram analysis, multi-way join, and some pre-computation of some essential statistical measures, such as sum, count max, min, and so on.

Tight Coupling

  - Tight coupling means that a data mining systems smoothly integrated into the database or data warehouse system
  - The data mining subsystem is treated as one functional component of an information system
  - This approach is highly desirable since it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment

- A well designed data mining system should offer tight or semi tight coupling with a database or data warehouse system.

Data Mining Techniques

- DM is not so much a single technique, as the idea that there is more knowledge hidden in the data than shows itself on the surface.

- Any data that helps extract more out of data is useful. So DM techniques form quite a heterogeneous group

DM Techniques..

- Query tools
- Statistical Techniques
- Visualization
- OLAP
- Case-Based Learning (K- Nearest Neighbor)
- Decision Trees
- Association rules
- Neural Network
- Genetic Algorithms & many more…

Query Tools

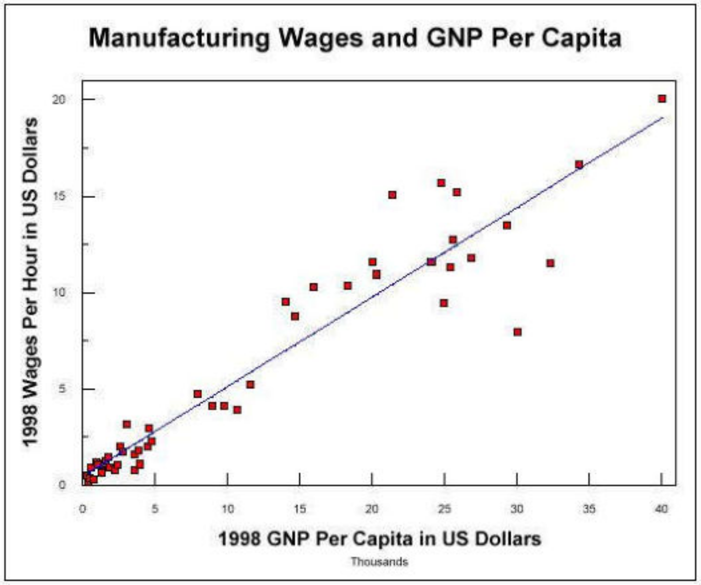
- The first step in data mining should always be a rough analysis of the data set using traditional query tools.
- Applying simple SQL can achieve wealth of information
- Almost 80% of the interesting information can be abstracted from a database using SQL,
- The remaining 20% requires more advanced techniques where 20% hidden Information might have vital importance.

Statistical Methods (Statistical DM)

- There are many well-established statistical techniques for data analysis, particularly for numeric data
  - applied extensively to data from scientific experiments and data from economics and the social sciences

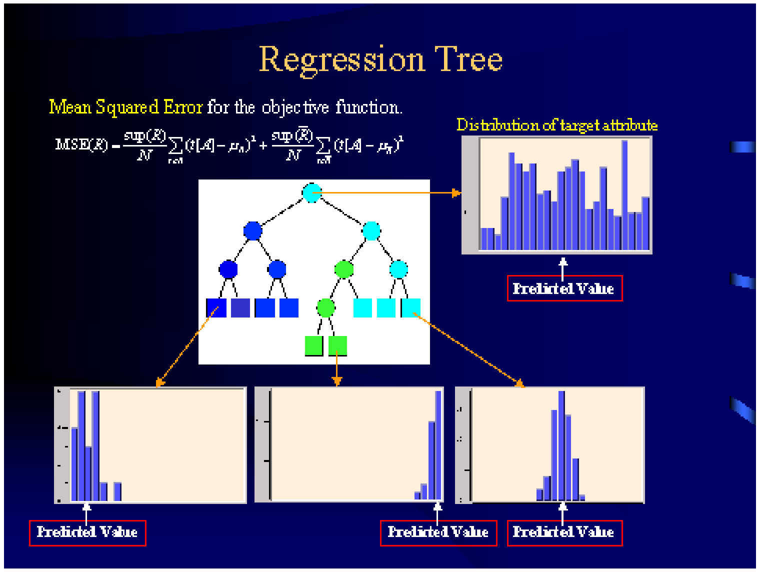
Regression

- predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric



Statistical Methods (Statistical DM)

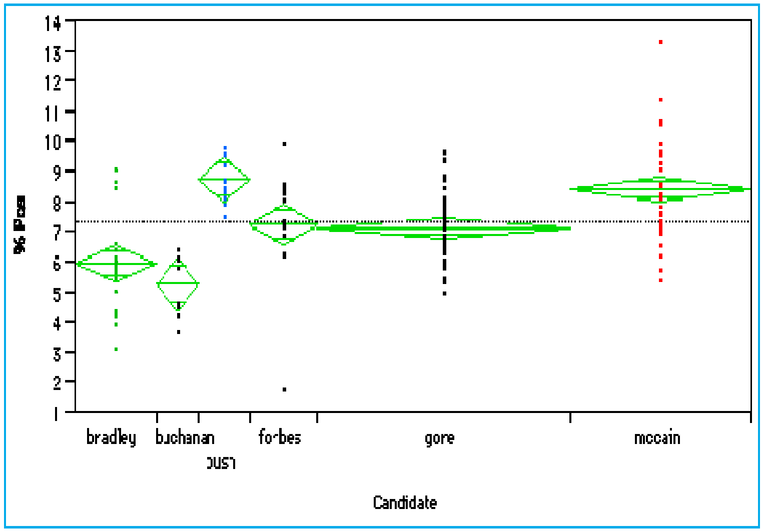
- **Regression trees**
  - Binary trees are used for classification and prediction
  - Similar to decision trees:Tests are performed at the internal nodes
  - In a regression tree the mean of the objective attribute is computed and used as the predicted value
- **Analysis of variance**
  - Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)



Regression Tree

Mean Squared Error for the objective function.

$$MSE(R) = \frac{sup(R)}{N}\sum_{t\in R}(t[A]-\mu_R)^2 + \frac{sup(\bar{R})}{N}\sum_{t\in \bar{R}}(t[A]-\mu_{\bar{R}})^2$$



Visualization Techniques (Visual DM)

- Visualization: use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data

- Visual Data Mining: the process of discovering implicit but useful knowledge from large data sets using visualization techniques

Computer Graphics

Multimedia Systems

Human Computer Interfaces

High Performance Computing

Pattern Recognition

Visualization

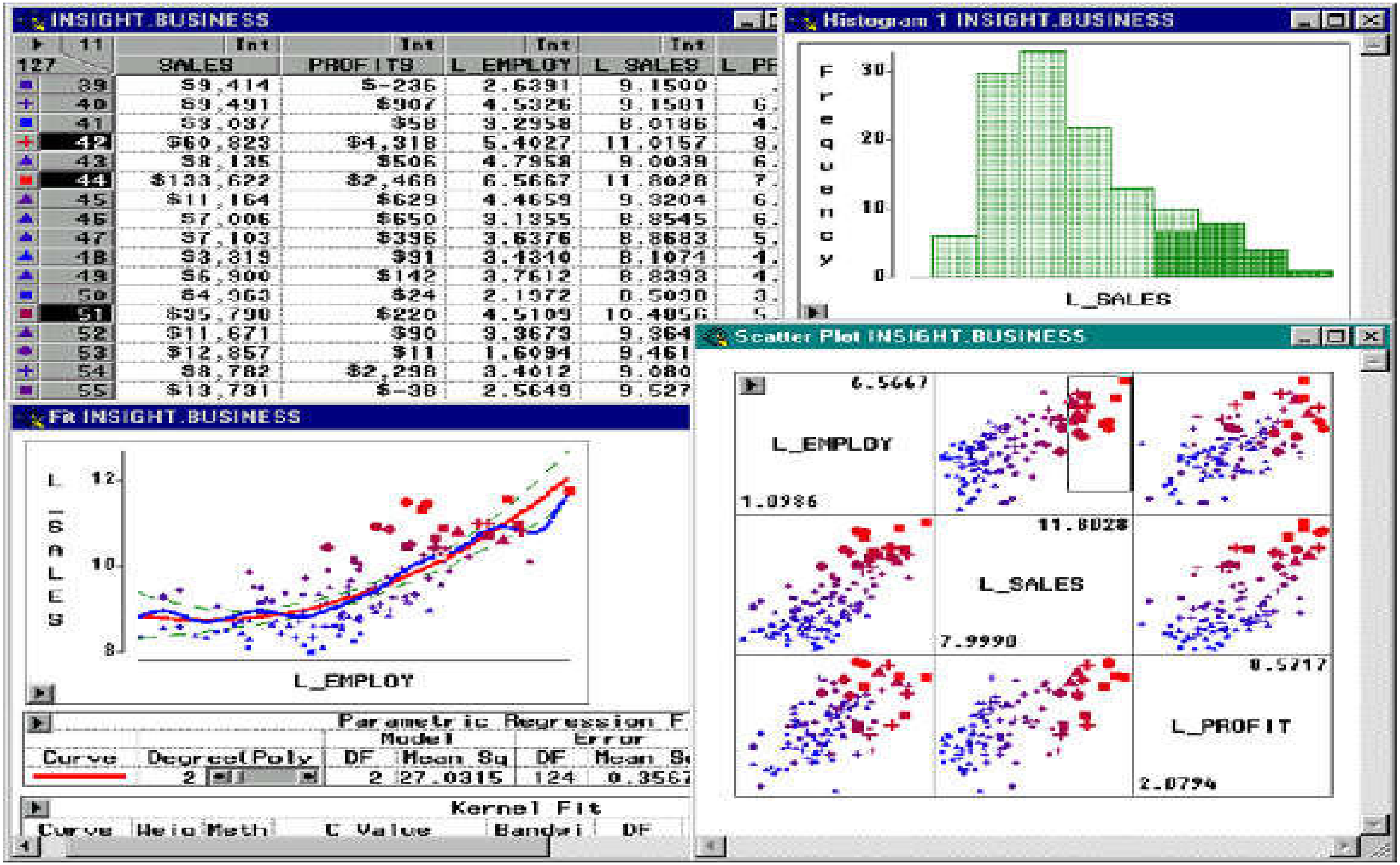
- Purpose of Visualization
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data.
  - Help to find interesting regions and suitable parameters for further quantitative analysis.
  - Provide a visual proof of computer representations derived

Data Mining Result Visualization

- Presentation of the results or knowledge obtained from data mining in visual forms

- Examples
  - Scatter plots
  - Decision trees
  - Association rules
  - Clusters
  - Outliers

Visualization of Data Mining Results: Scatter Plots



INSIGHT.BUSINESS

11 / 127	SALES (Int)	PROFITS (Int)	L_EMPLOY (Int)	L_SALES (Int)	L_PF
39	$9,414	$-236	2.6391	9.1500	.
40	$9,491	$907	4.5326	9.1581	6.
41	$3,037	$58	3.2958	8.0186	4.
42	$60,823	$4,318	5.4027	11.0157	8.
43	$8,135	$506	4.7958	9.0039	6.
44	$133,622	$2,468	6.5667	11.8028	7.
45	$11,164	$629	4.4659	9.3204	6.
46	$7,006	$650	3.1355	8.8545	6.
47	$7,103	$396	3.6376	8.8683	5.
48	$3,319	$91	3.4340	8.1074	4.
49	$6,900	$142	3.7612	8.8393	4.
50	$4,963	$24	2.1972	8.5030	3.
51	$35,790	$220	4.5109	10.4856	5.
52	$11,671	$90	3.3673	9.364	
53	$12,857	$11	1.6094	9.461	
54	$8,782	$2,298	3.4012	9.080	
55	$13,731	$-38	2.5649	9.527	

Histogram 1 INSIGHT.BUSINESS

Scatter Plot INSIGHT.BUSINESS

Fit INSIGHT.BUSINESS

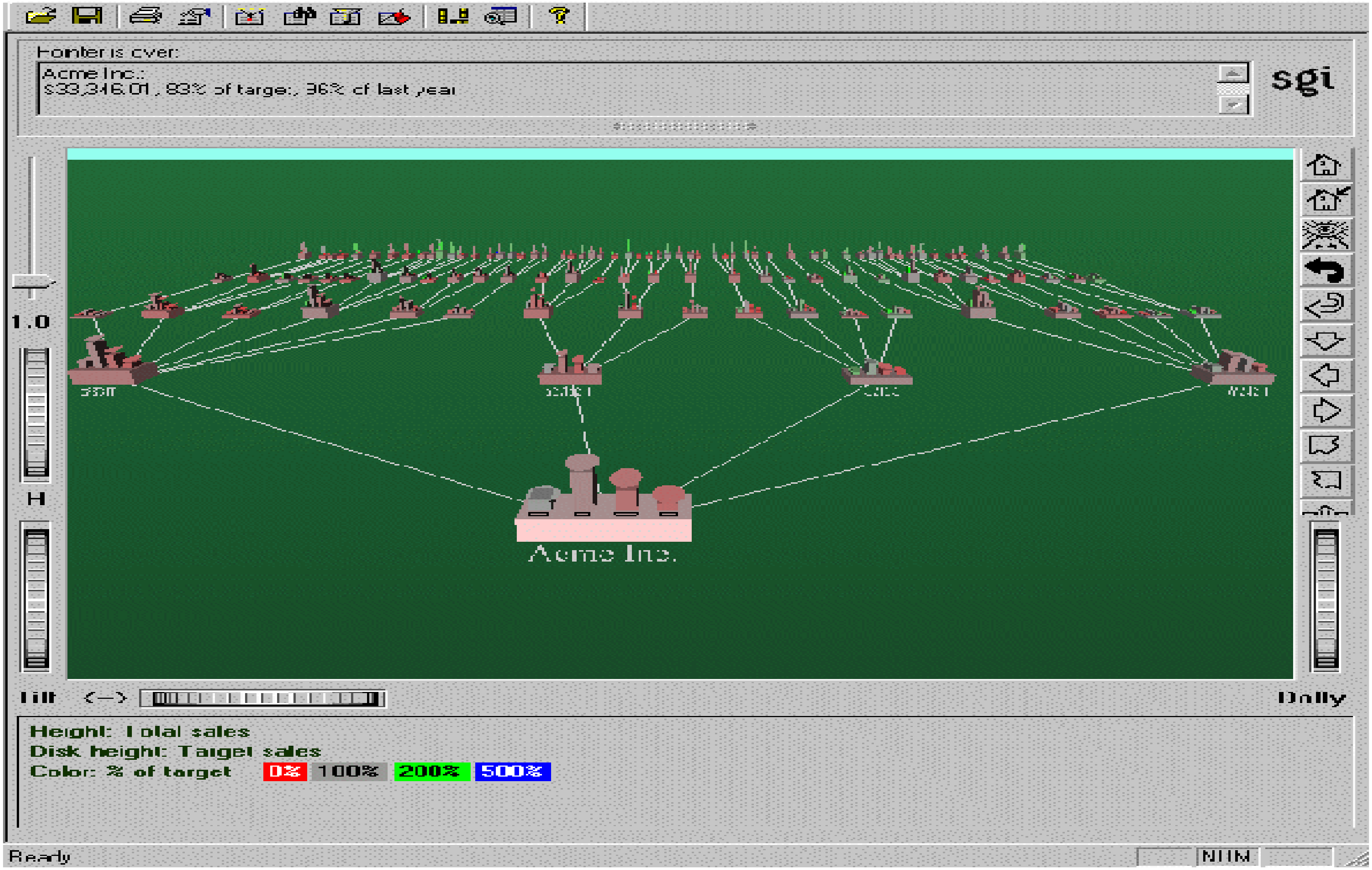
Parametric Regression F

Curve	Degree(Poly)	Model DF	Model Mean Sq	Error DF	Error Mean S
🟥	2	2	27.0315	124	0.3567

Kernel Fit

Curve	Weig Meth	C Value	Bandwi	DF

Visualization of a Decision Tree : MineSet 3.0

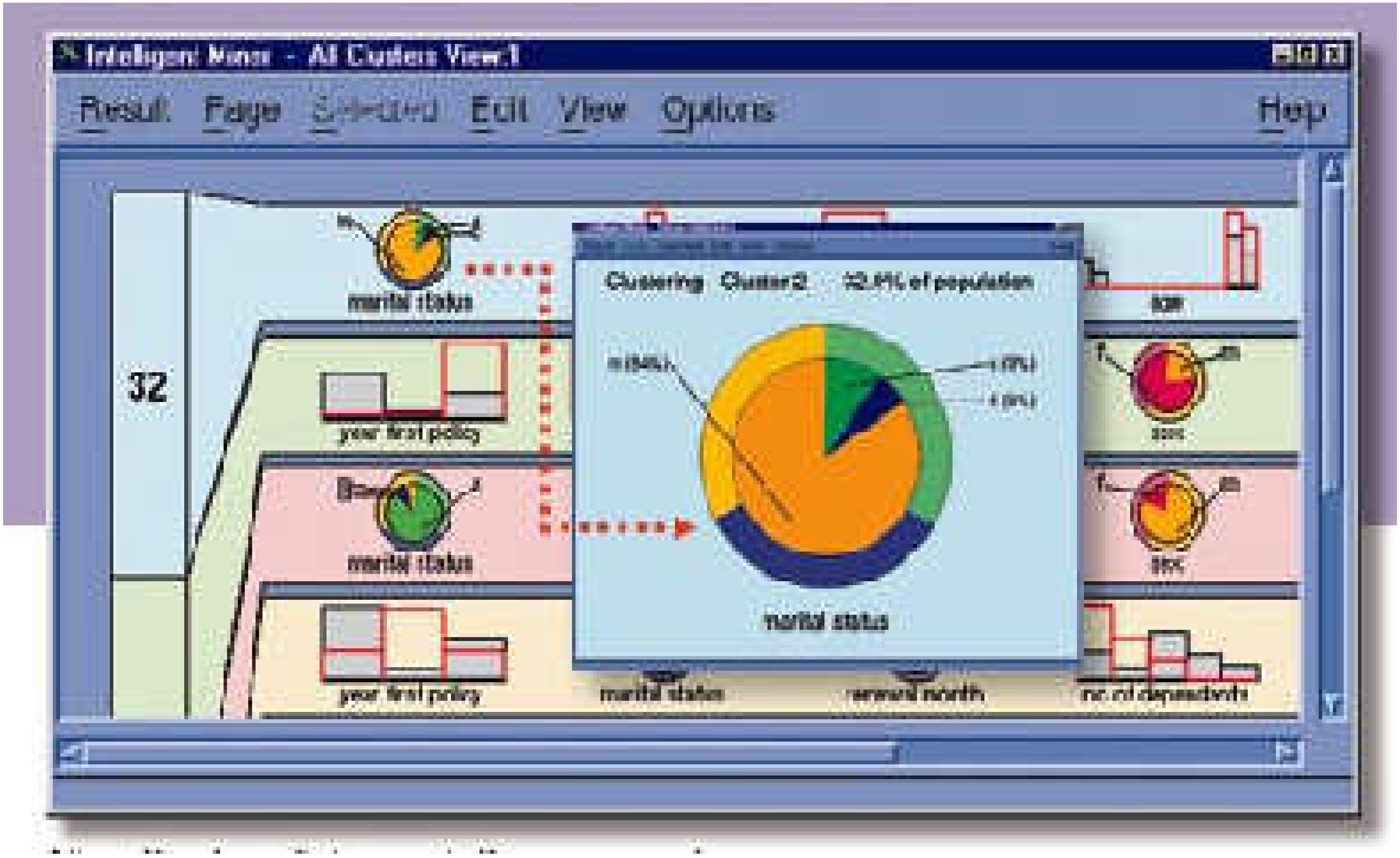


Pointer is over:

Acme Inc.:
$33,346.01 , 83% of target, 36% of last year

Height: Total sales
Disk height: Target sales
Color: % of target 🟥 0% ▨ 100% 🟩 200% 🟦 500%

Visualization of Cluster Grouping in IBM Intelligent Miner

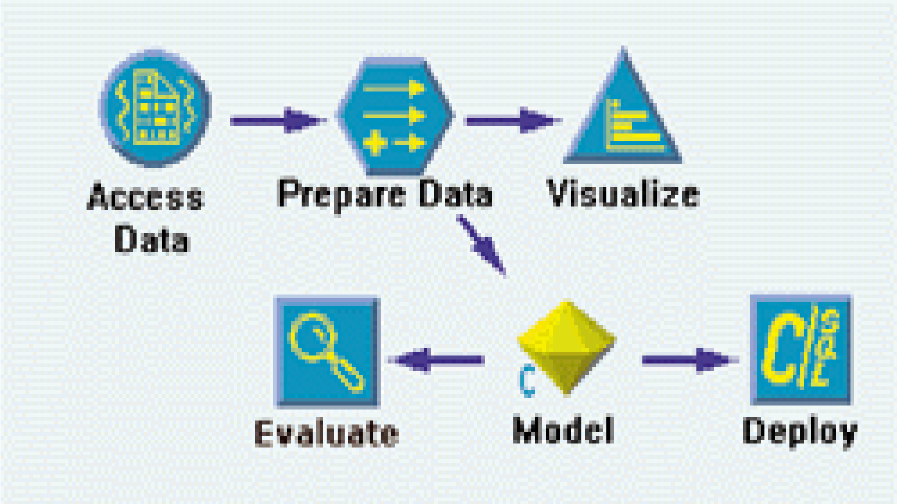


Data Mining Process Visualization

- Presentation of the various processes of data mining in visual forms so that users can see
  - Data extraction process
  - Where the data is extracted
  - How the data is cleaned, integrated, preprocessed, and mined
  - Method selected for data mining
  - Where the results are stored
  - How they may be viewed

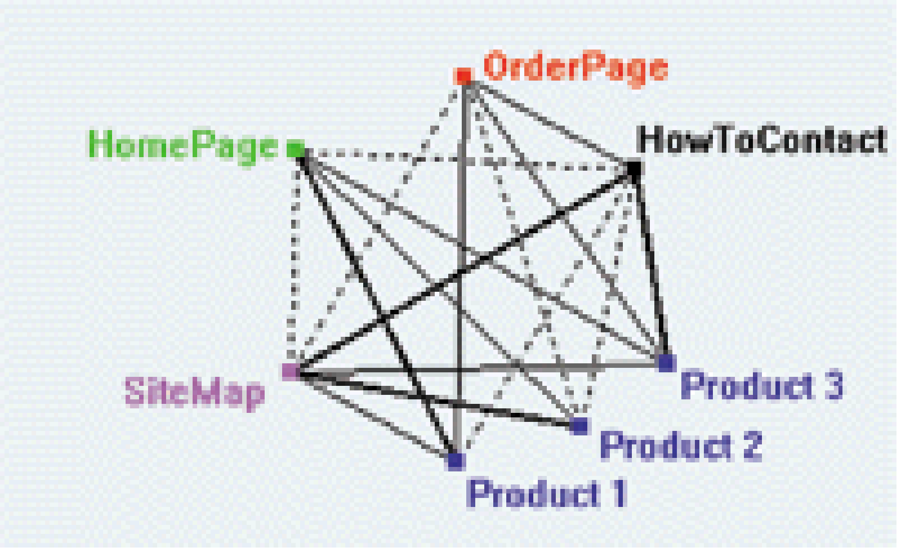
Visualization of Data Mining Processes by Clementine

E.g.: Knowledge Flow in WEKA



See your solution discovery process clearly

Understand variations with visualized data



Likelihood & Distance

- Records that are close to each other are very alike & records that are vary far removed from each other represents individuals that have little in common.

- Examples: 3 field values (age, income & Credit): these three attributes form a three dimensional data space and we can analyze the distances between records in this space.

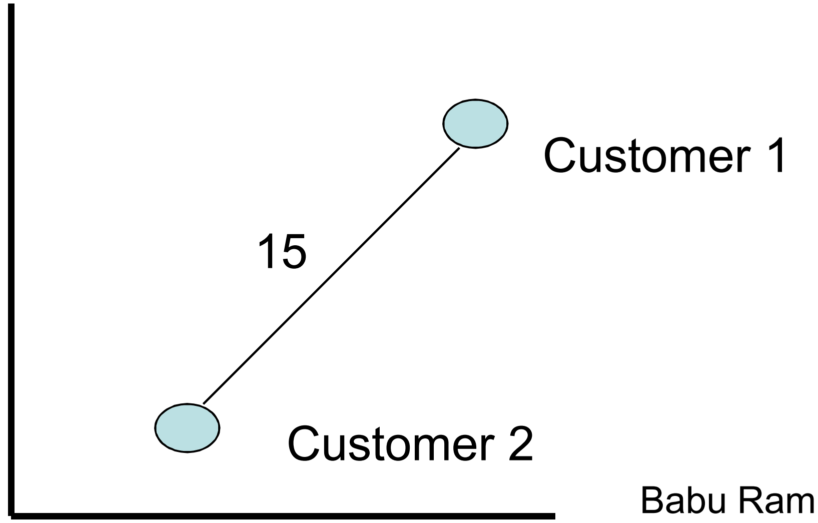
- Assume: Age range: 1 to 100 yrs, Income range: 0 to 100,000$ & Credit range: 0 to 50,000$

- If it is used without correction, then income/credit is more distinctive than age. So for better analysis, we need **coding.**

- **Coding: divide income & age separation by 1000.**

Likelihood & Distance

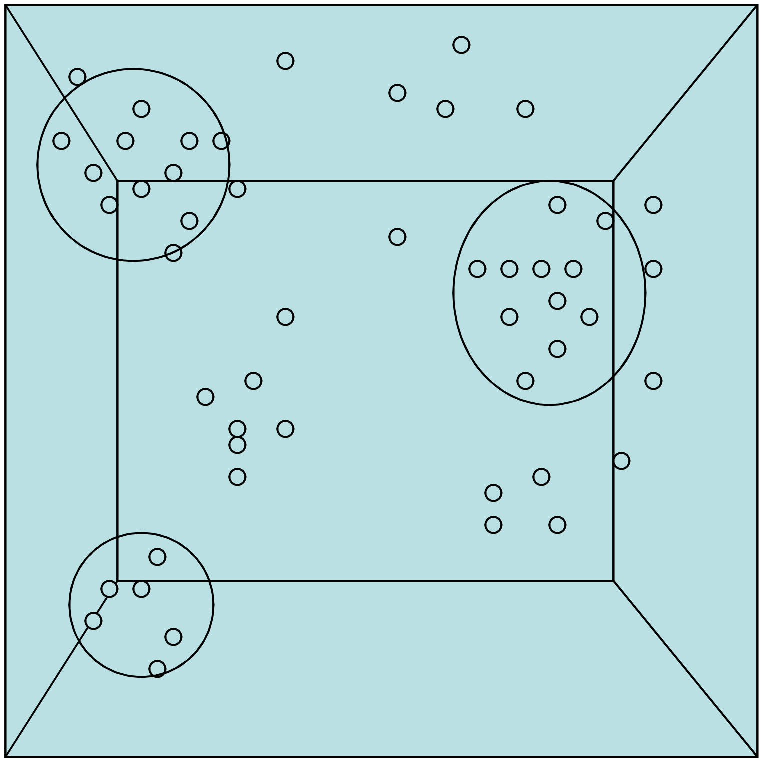
	Age	Income	Credit
Customer 1	32	40,000	10,000
Customer 2	24	30,000	2,000
Differences	8	10,000	8,000
Coding	8	10	8
distance	SquareRoot(8*8+10*10+8*8)=15		



Records Become points in Multidimensional data space

Likelihood & Distance

Sometimes it is possible to identify a visual cluster of potential Customers that are very likely to buy certain product



OLAP Tools

- Idea of dimensionality (data are managed into multidimensional cube)
  - Eg: how much is sold?? (Zero Dimensional)
  - What type of magazines are sold in a designated area per month and to what age group? (a 4-dimensional question: product, area, purchase date and age)
- OLAP Operations (Slicing,Dicing,Rolling,Drilling)

K-Nearest Neighbor

- When we interpret records as points in a data space, we can define the concepts of neighborhood.
  - "Records that are close to each other live in each others neighborhood."
  - "Records of the same type will be close to each other in the data space."
  - i.e. customers of same type will show same behavior.

K-Nearest Neighbor

- Basic principle of K-nearest neighbor is
  - "Do as your neighbor Do"

- If we want to predict the behavior of certain individual, we first to look at the behavior of, for example: ten individuals that are close to him/her in the data space.

- We calculate the average of 10 neighbors and this average value will be the prediction for our individual.

- K=> No. of Neighbors. 5-Nearest neighbor => 5 neighbors to be taken for calculation

K-Nearest Neighbor

- Draw Back
  - If we want to make a prediction for each element in the data set containing n-records, then we have to compare each record with every other records which leads to complexity.
  - For million records, we may need billion comparisons.
  - So not desirable for large data sets.
  - Doesn't work for high attributes or high dimensional data.

Decision Tree: Outline

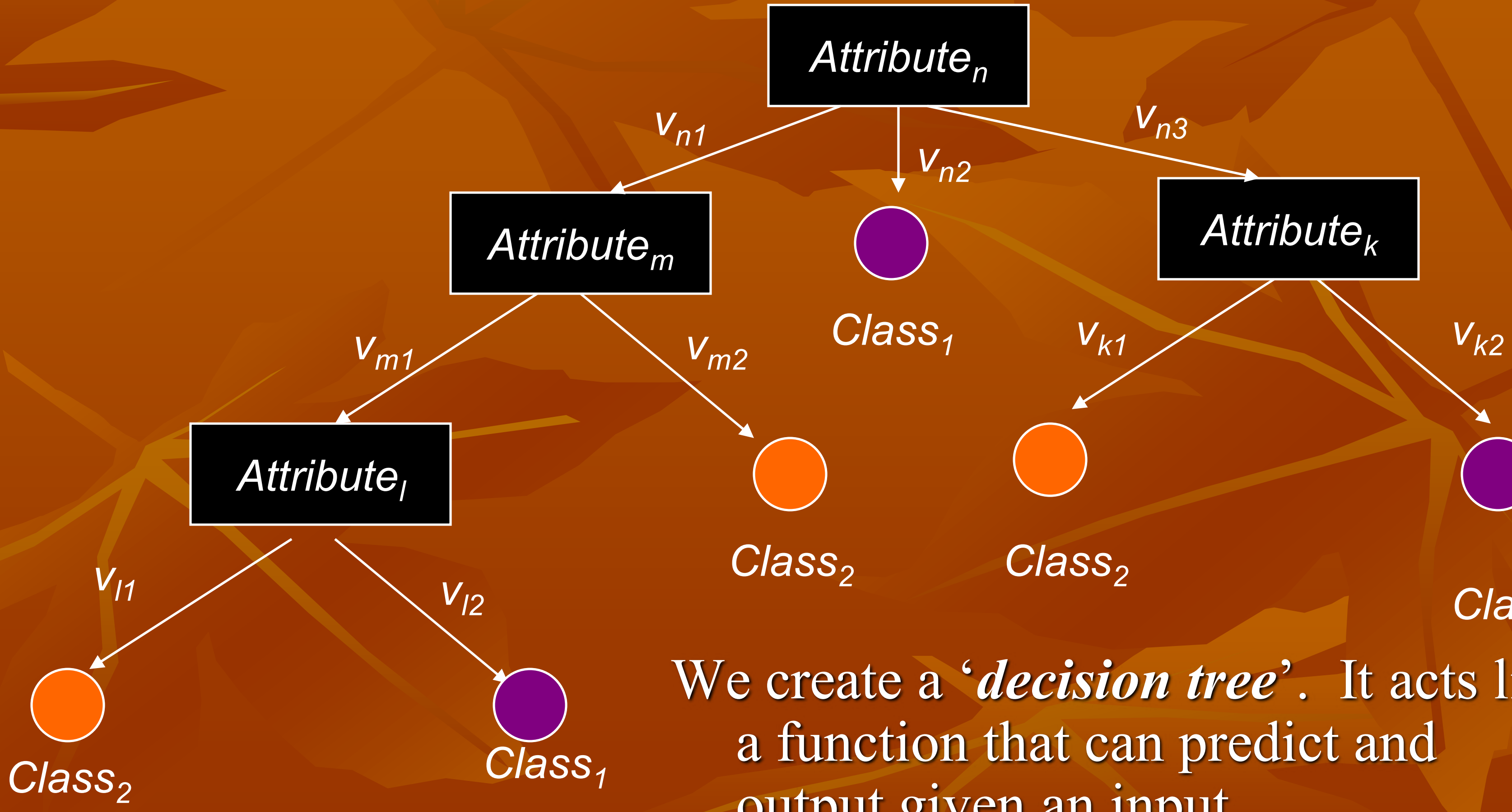
- Decision tree representation
- ID3 learning algorithm
- Entropy, information gain
- Overfitting

Defining the Task

- Imagine we’ve got a set of data containing several types, or ***classes***.
  - E.g. information about customers, and class=whether or not they buy anything.

- Can we predict, i.e ***classify***, whether a previously unseen customer will buy something?

An Example Decision Tree



We create a ‘***decision tree***’. It acts like a function that can predict and output given an input

Decision Trees

- The idea is to ***ask a series of questions***, starting at the root, that will lead to a leaf node.
- The ***leaf node provides the classification***.

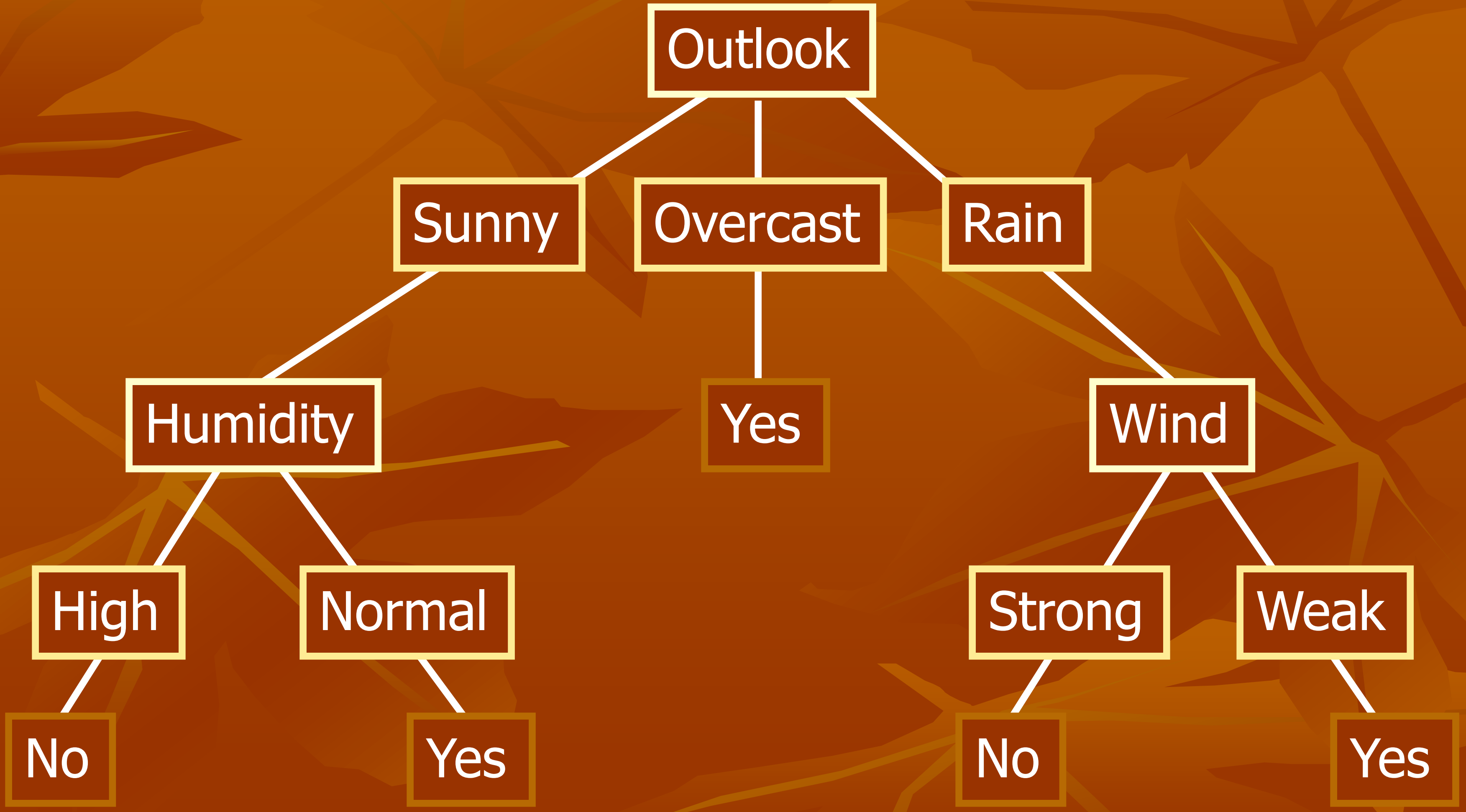
Algorithm for Decision Tree Induction

- Basic algorithm
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

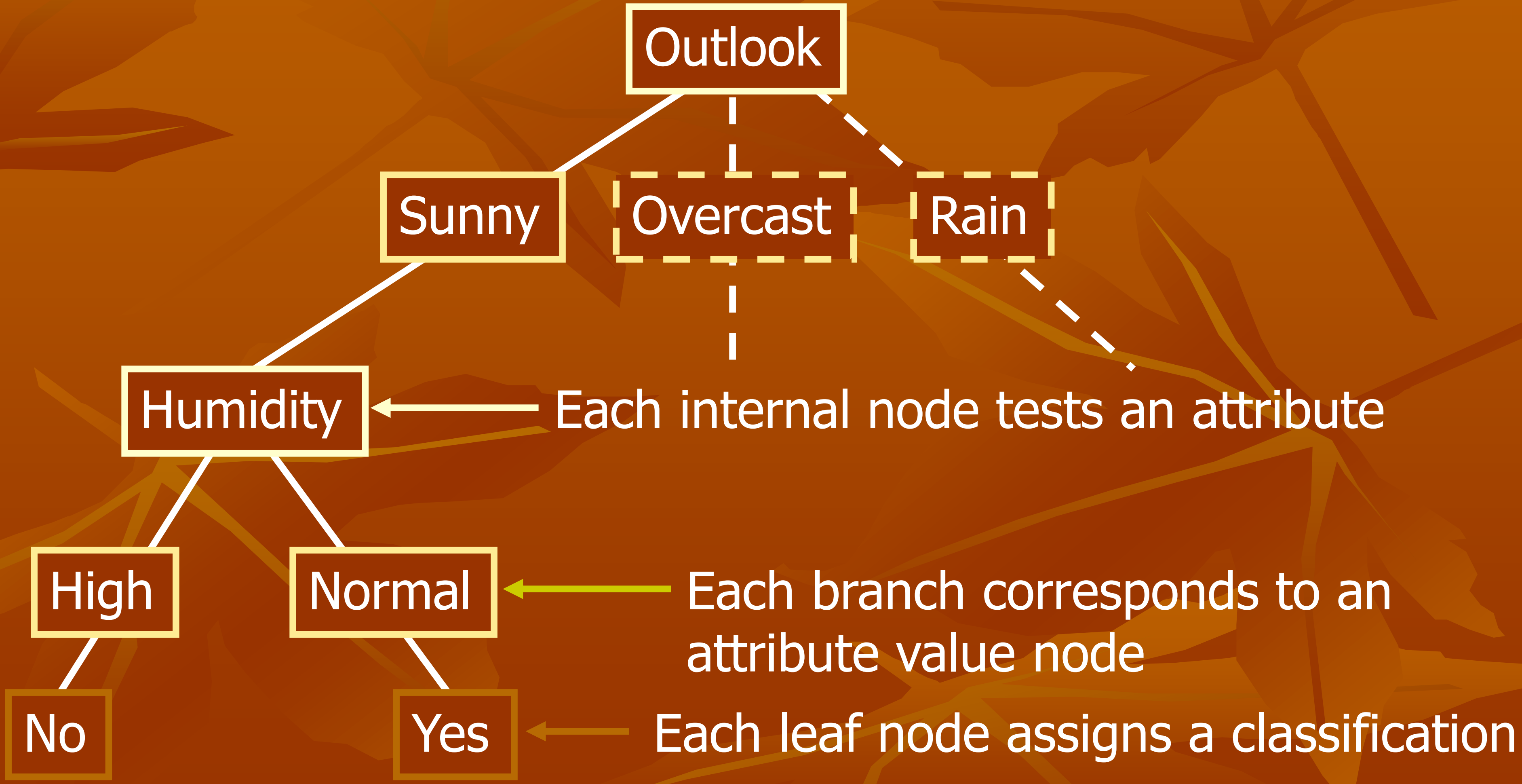
Classification by Decision Tree Induction

- Decision tree
  - A flow-chart-like tree structure
  - Internal node denotes a test on an attribute
  - Branch represents an outcome of the test
  - Leaf nodes represent class labels or class distribution
- Decision tree generation consists of two phases
  - Tree construction
    - At start, all the training examples are at the root
    - Partition examples recursively based on selected attributes
  - Tree pruning
    - Identify and remove branches that reflect noise or outliers
- Once the tree is build
  - Use of decision tree: Classifying an unknown sample

Decision Tree for PlayTennis



Decision Tree for PlayTennis



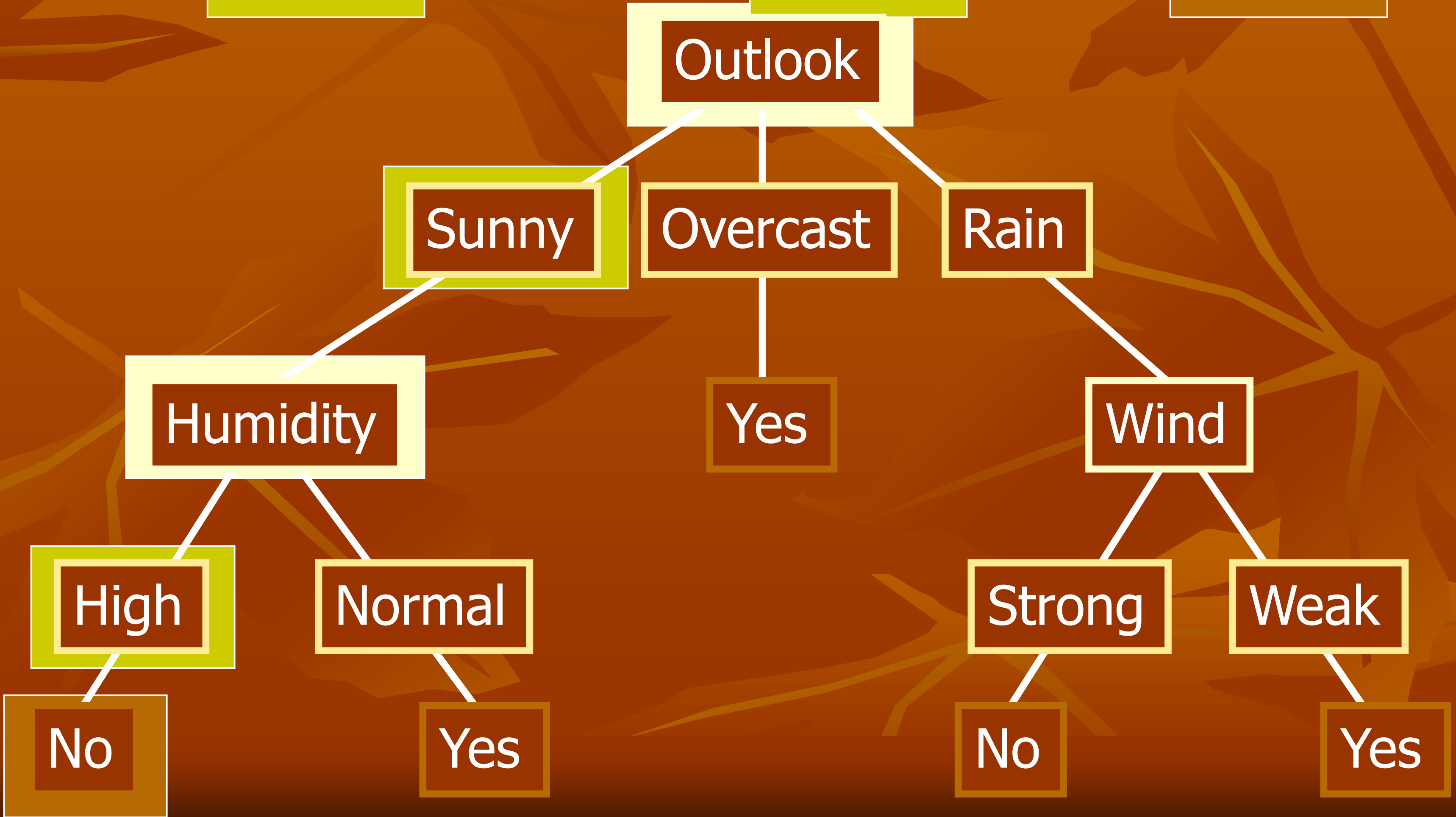
Each internal node tests an attribute

Each branch corresponds to an attribute value node

Each leaf node assigns a classification

Decision Tree for PlayTennis

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	?No



Outlook

Sunny
Overcast
Rain

Humidity
Yes
Wind

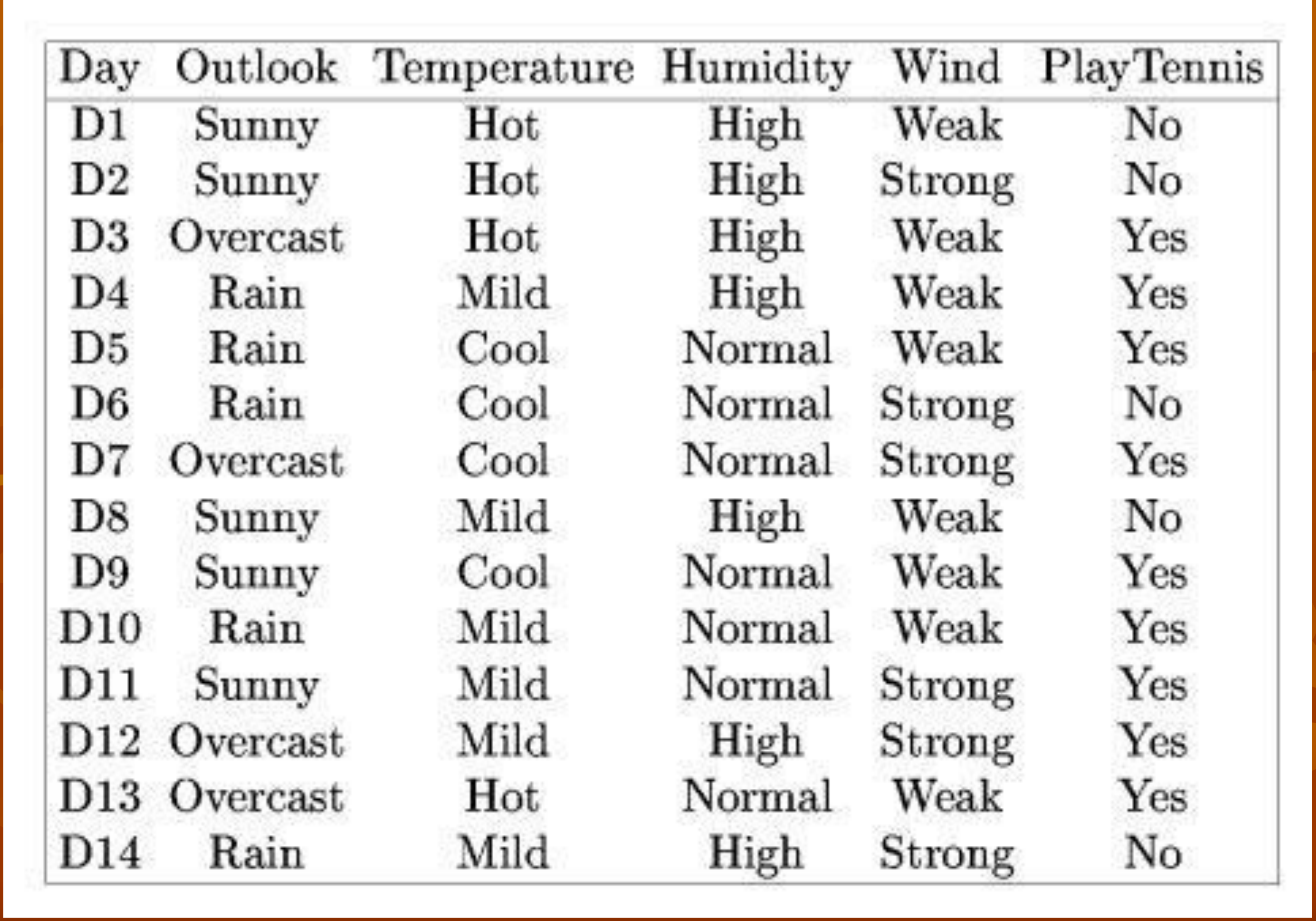
High
Normal
Strong
Weak

No
Yes
No
Yes

Decision Trees

Consider these data:

A number of examples of weather, for several days, with a classification ‘PlayTennis.’



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

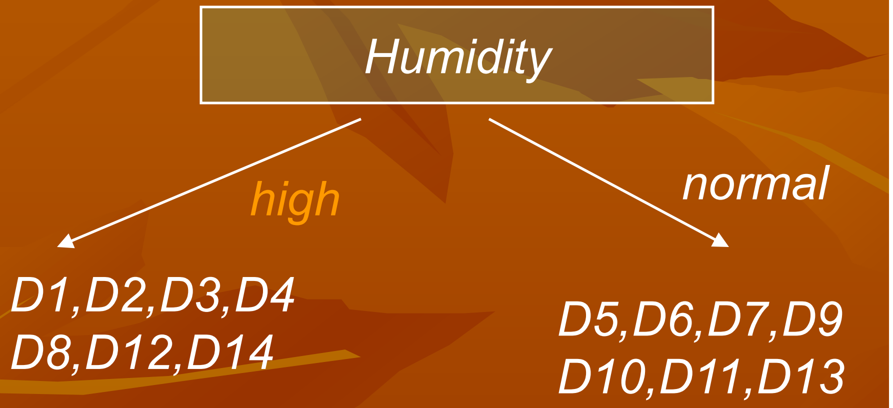
Decision Tree Algorithm

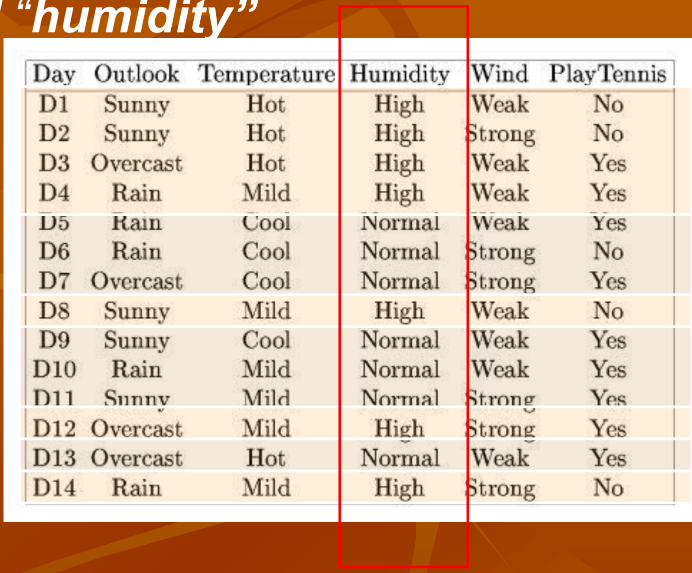
Building a decision tree

1. Select an attribute
2. Create the subsets of the example data for each value of the attribute
3. For each subset
   - *if not all the elements of the subset belongs to same class repeat the steps 1-3 for the subset*

Building Decision Trees

*Let’s start building the tree from scratch. We first need to decide which attribute to make a decision. Let’s say we selected “**humidity**”*

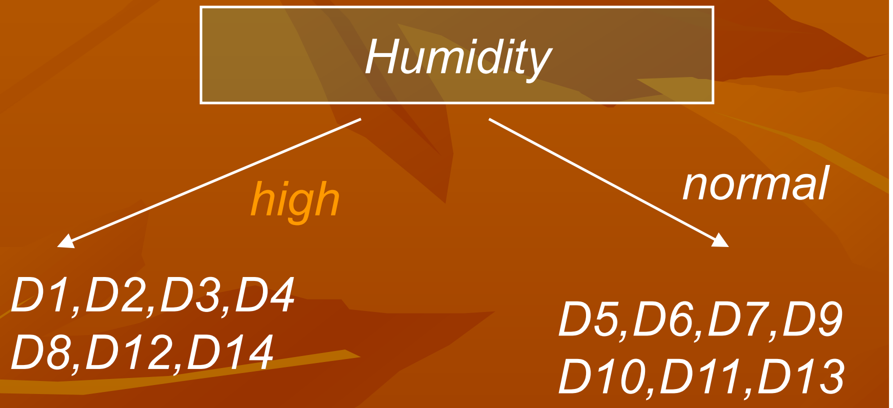




Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*Now lets classify the first subset D1,D2,D3,D4,D8,D12,D14 using attribute “**wind**”*



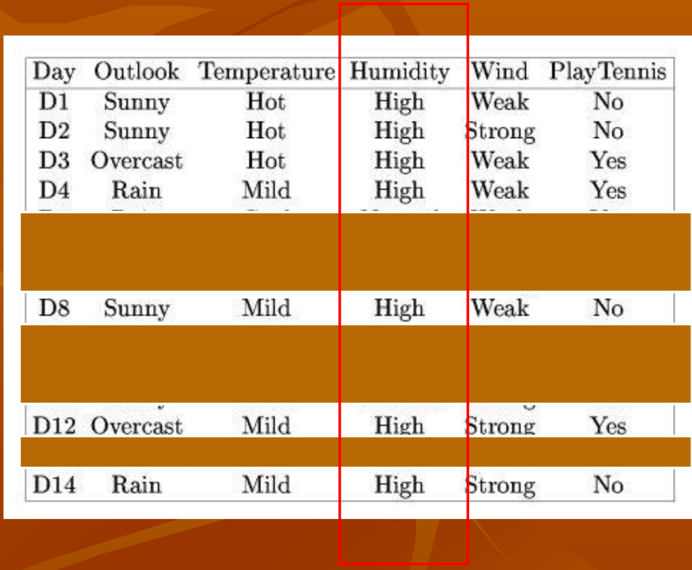
Humidity

high

normal

D1,D2,D3,D4
D8,D12,D14

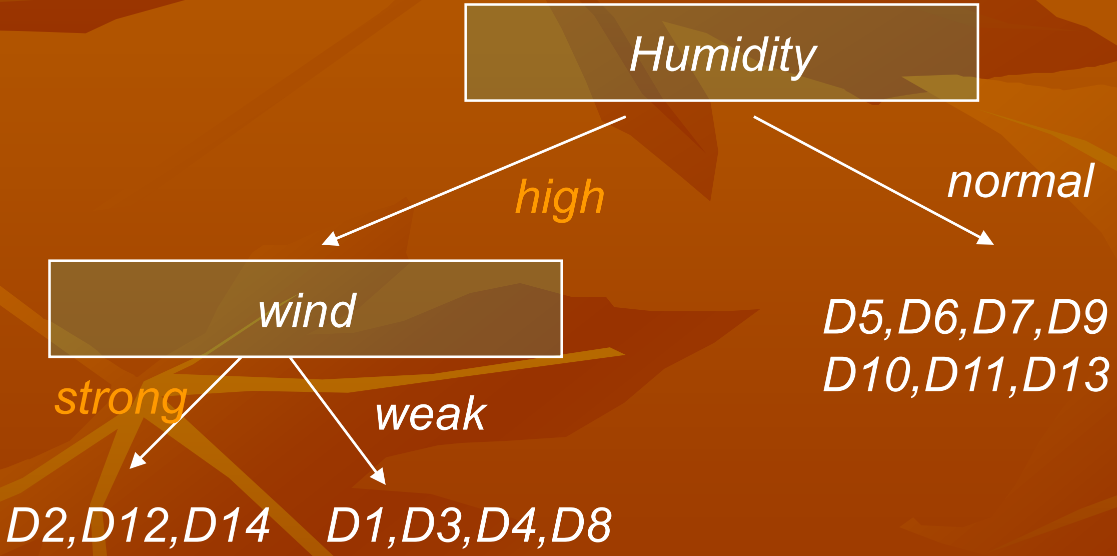
D5,D6,D7,D9
D10,D11,D13



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*Subset D1,D2,D3,D4,D8,D12,D14 classified by attribute "**wind**"*



Humidity

high

normal

wind

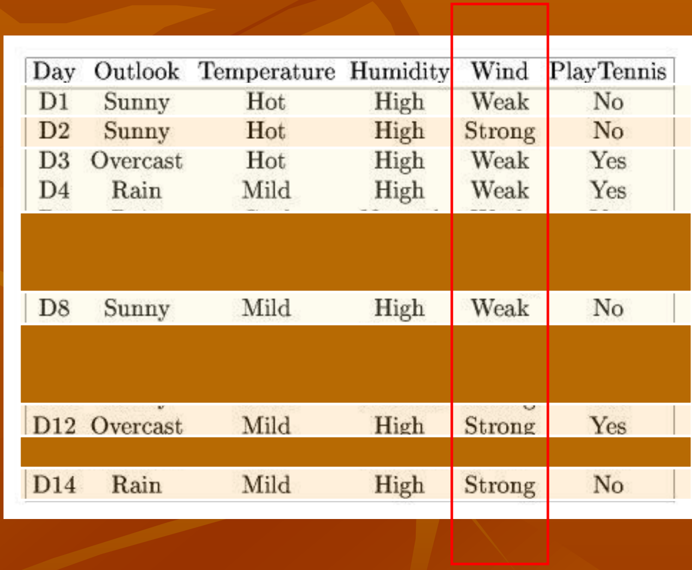
strong

weak

D2,D12,D14

D1,D3,D4,D8

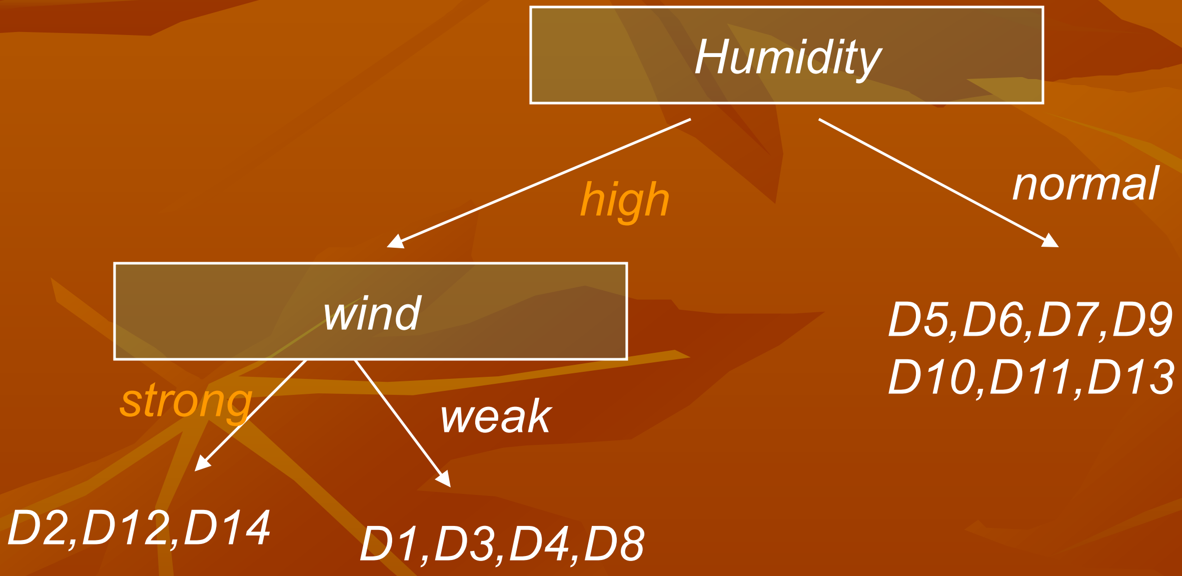
D5,D6,D7,D9
D10,D11,D13



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*Now lets classify the subset D2,D12,D14 using attribute "**outlook**"*



Humidity

high

normal

wind

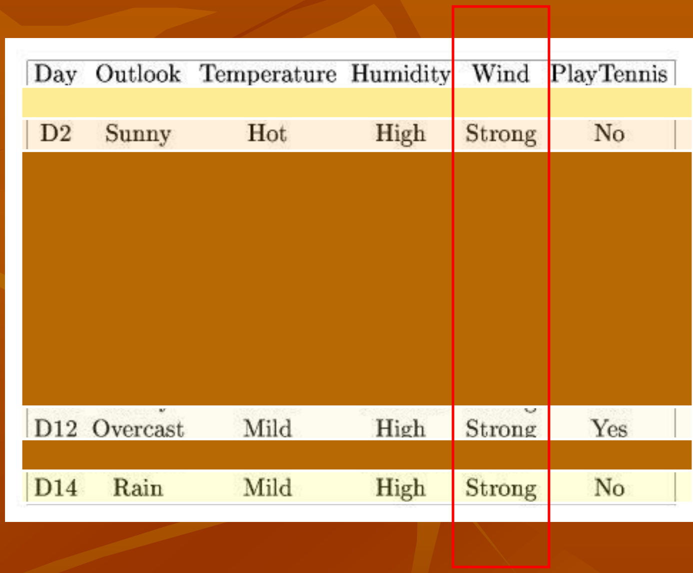
strong

weak

D2,D12,D14

D1,D3,D4,D8

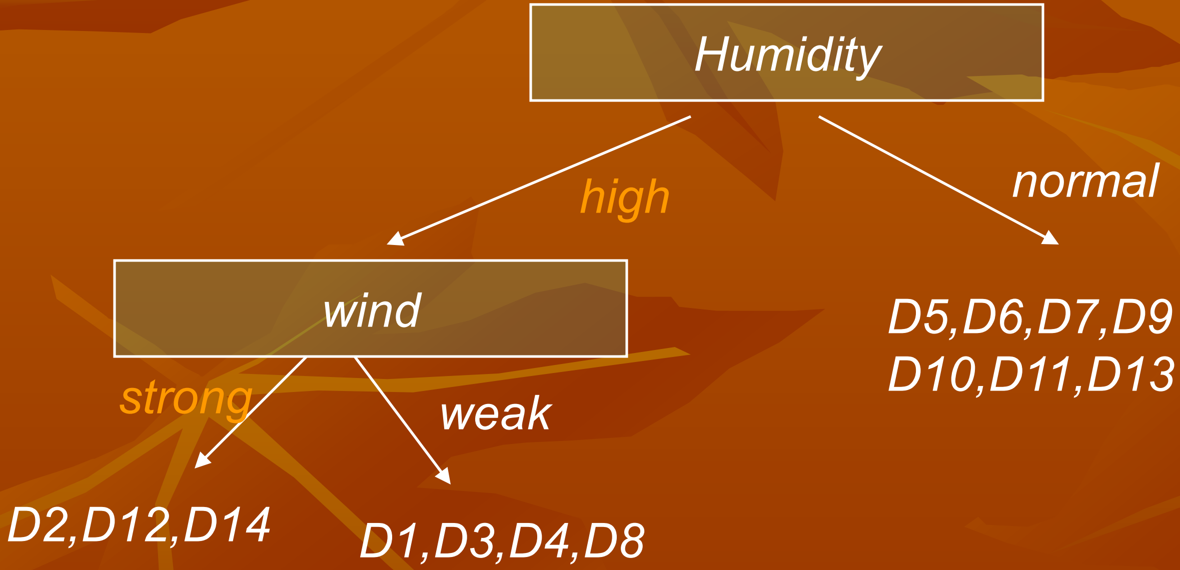
D5,D6,D7,D9
D10,D11,D13

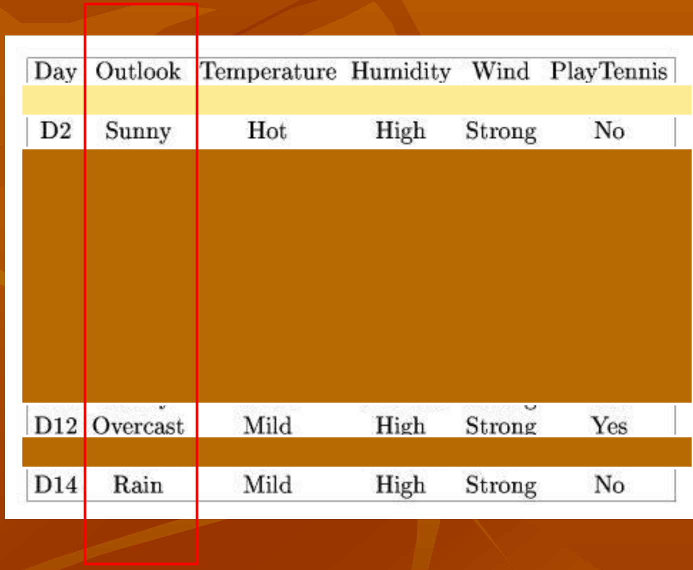


Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D2	Sunny	Hot	High	Strong	No
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*Subset D2,D12,D14 classified by “**outlook**”*

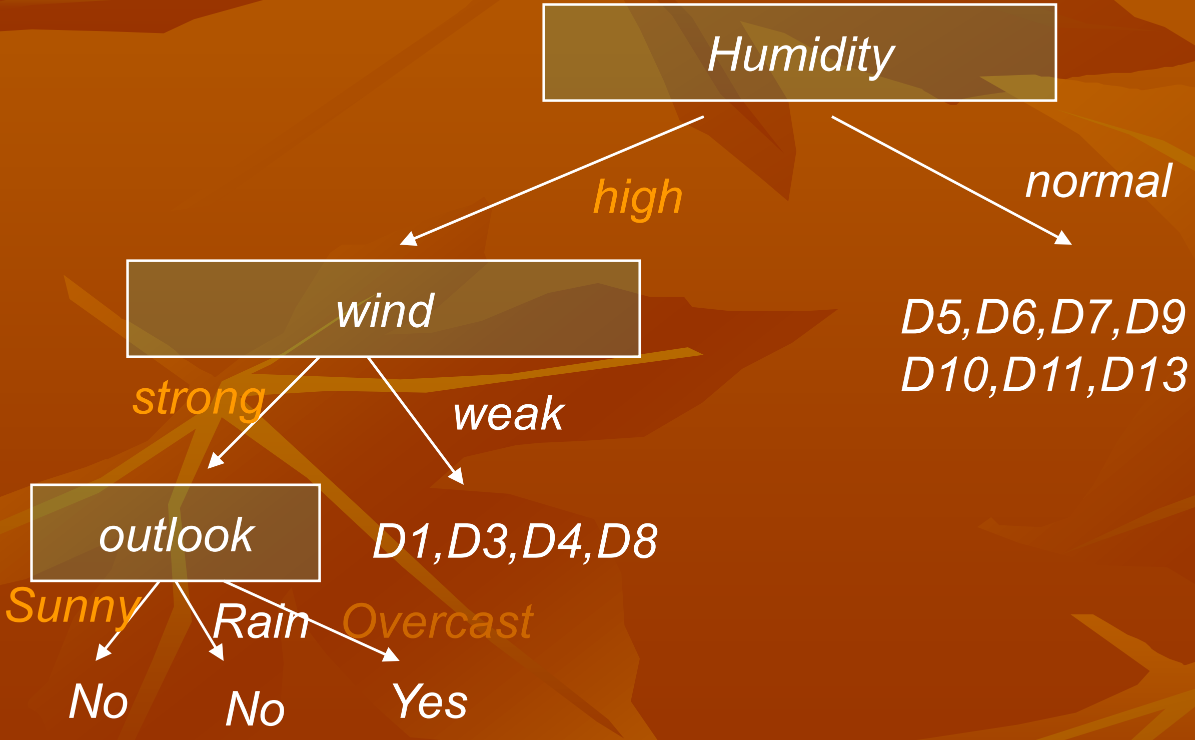




Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D2	Sunny	Hot	High	Strong	No
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*subset D2,D12,D14 classified using attribute “**outlook**”*



Humidity

high

normal

wind

D5,D6,D7,D9
D10,D11,D13

strong

weak

outlook

D1,D3,D4,D8

Sunny

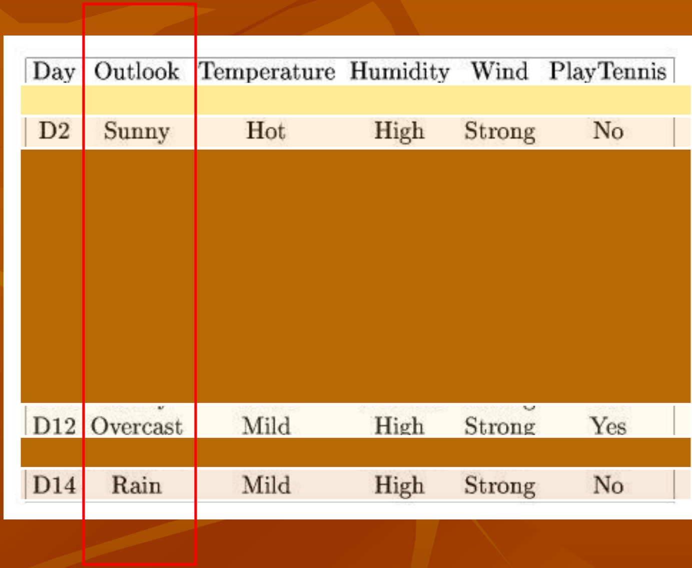
Rain

Overcast

No

No

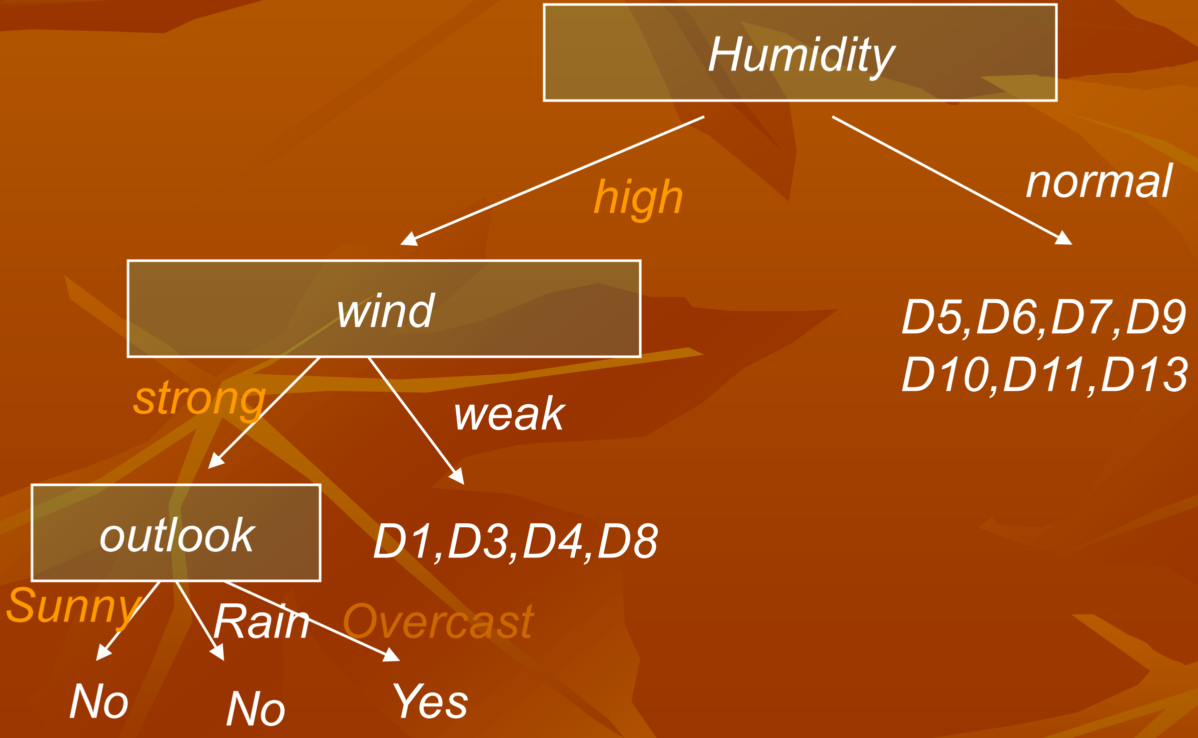
Yes



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D2	Sunny	Hot	High	Strong	No
D12	Overcast	Mild	High	Strong	Yes
D14	Rain	Mild	High	Strong	No

Building Decision Trees

*Now lets classify the subset D1,D3,D4,D8 using attribute “**outlook**”*



Humidity

high

normal

wind

D5,D6,D7,D9
D10,D11,D13

strong

weak

outlook

D1,D3,D4,D8

Sunny

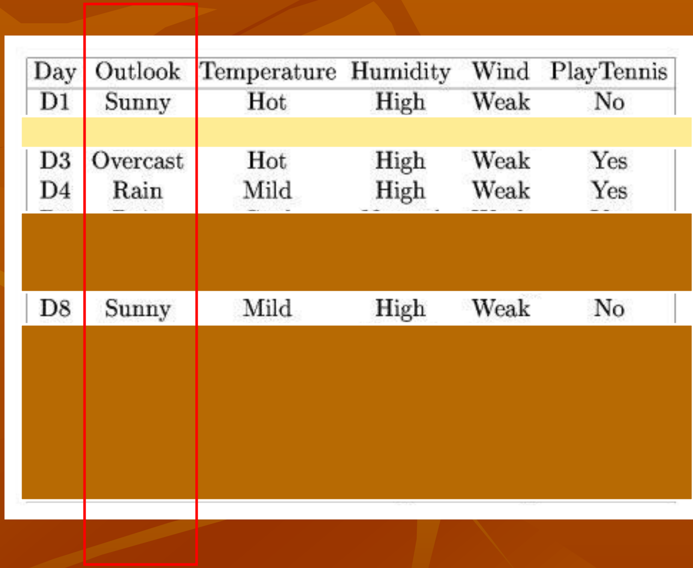
Rain

Overcast

No

No

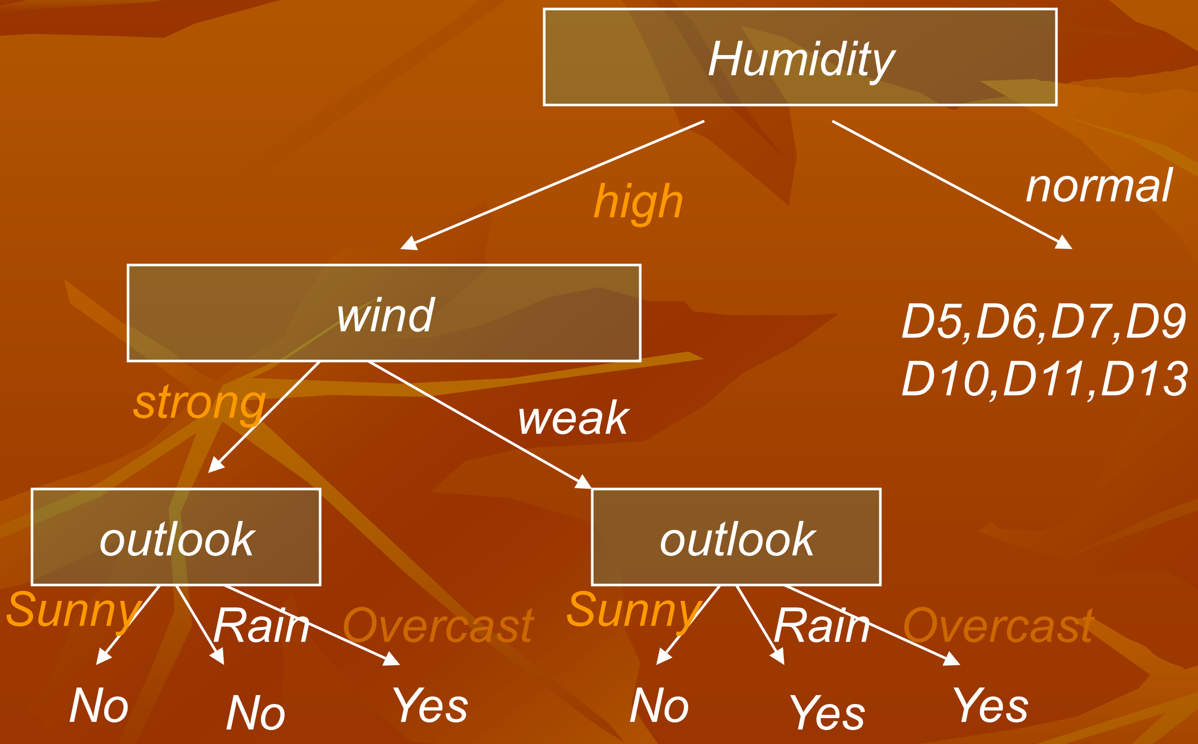
Yes



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D8	Sunny	Mild	High	Weak	No

Building Decision Trees

*subset D1,D3,D4,D8 classified by “**outlook**”*



Humidity

high

normal

wind

D5,D6,D7,D9
D10,D11,D13

strong

weak

outlook

outlook

Sunny

Rain

Overcast

Sunny

Rain

Overcast

No

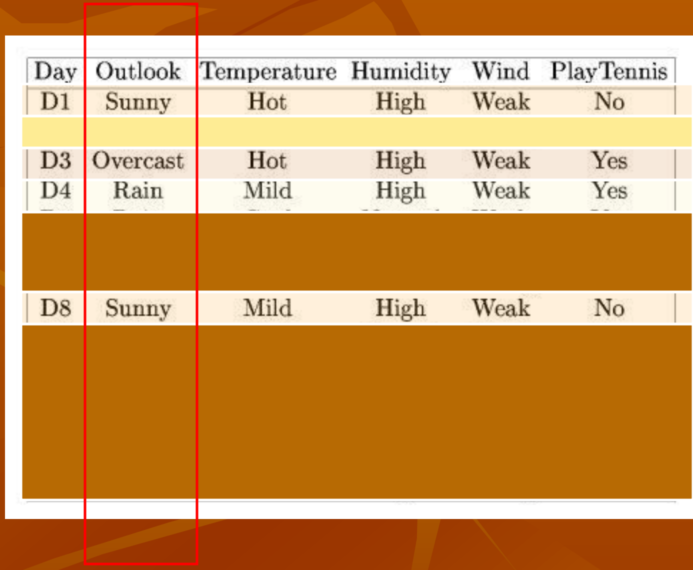
No

Yes

No

Yes

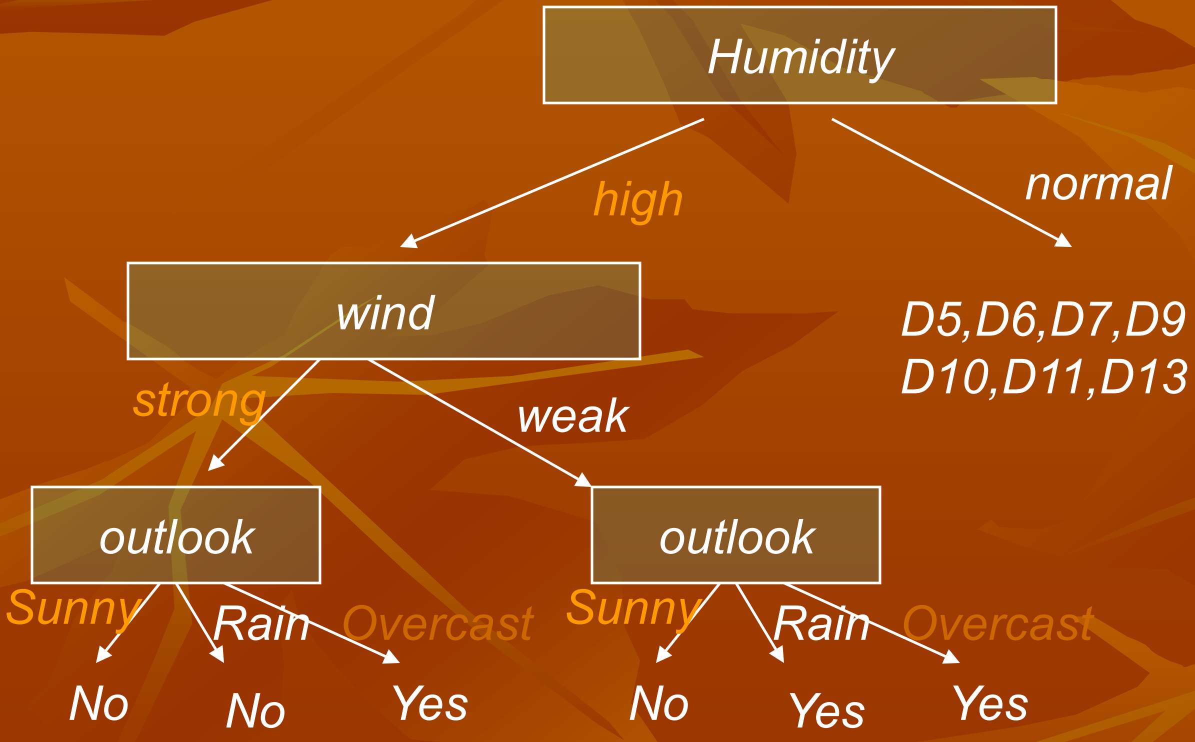
Yes

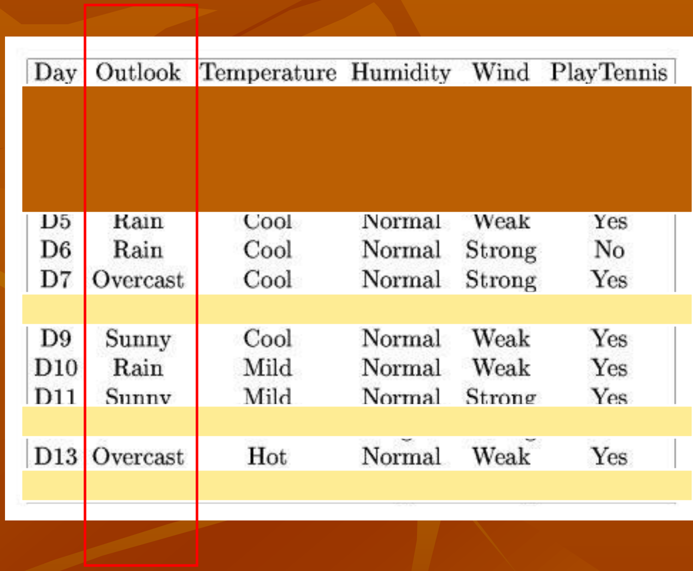


Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D8	Sunny	Mild	High	Weak	No

Building Decision Trees

*Now classify the subset D5,D6,D7,D9,D10,D11,D13 using attribute “**outlook**”*

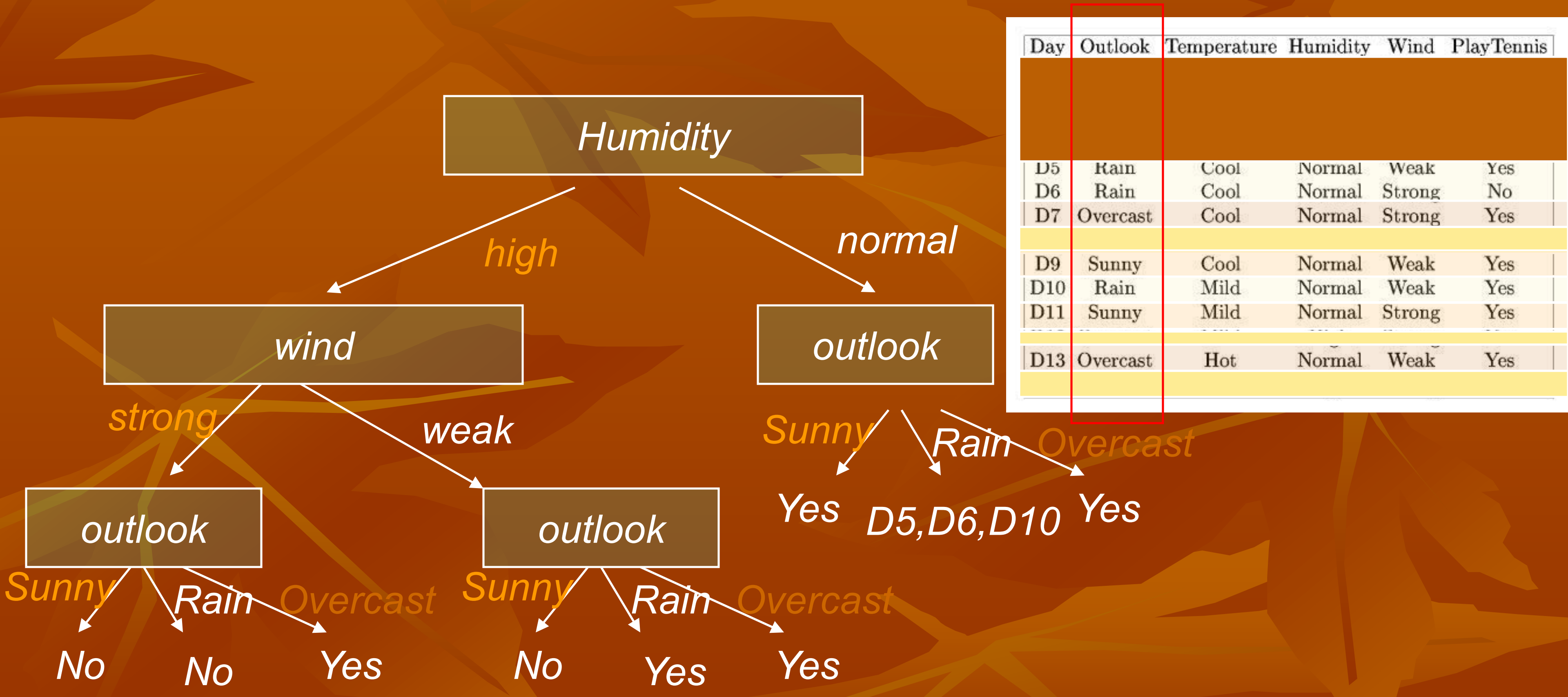




Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Building Decision Trees

*subset D5,D6,D7,D9,D10,D11,D13 classified by “**outlook**”*



Humidity

high

normal

wind

outlook

strong

weak

Sunny

Rain

Overcast

Yes

D5,D6,D10

Yes

outlook

outlook

Sunny

Rain

Overcast

Sunny

Rain

Overcast

No

No

Yes

No

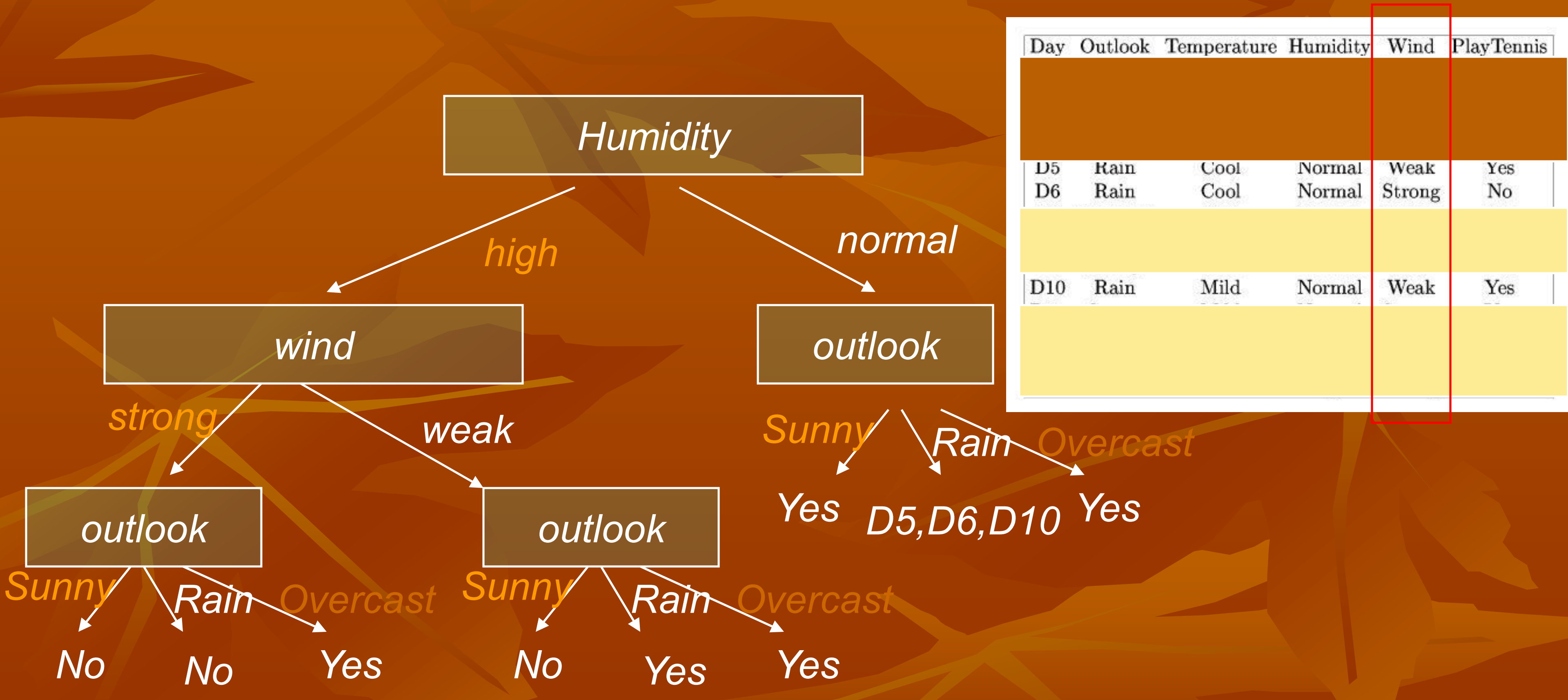
Yes

Yes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Building Decision Trees

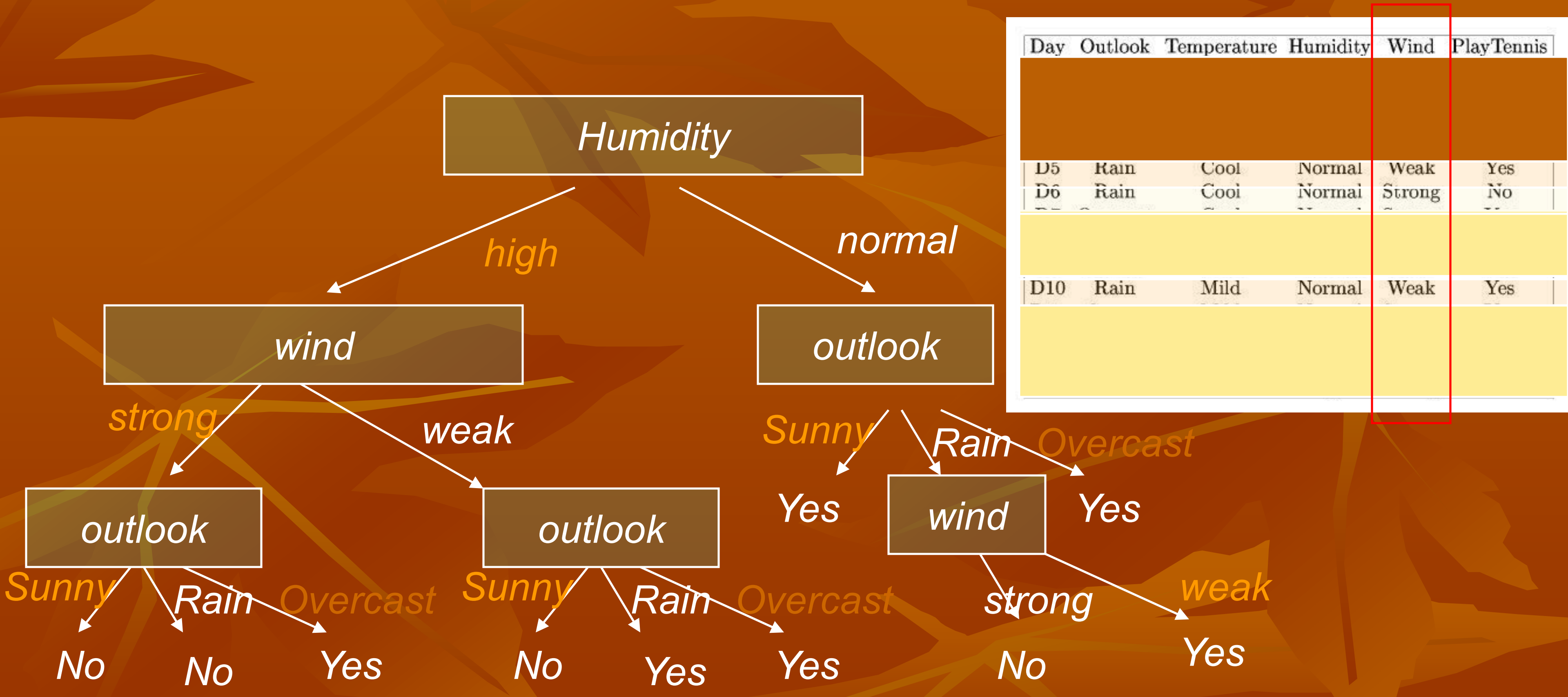
*Finally classify subset D5,D6,D10by “**wind**”*



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes

Building Decision Trees

*subset D5,D6,D10 classified by “**wind**”*

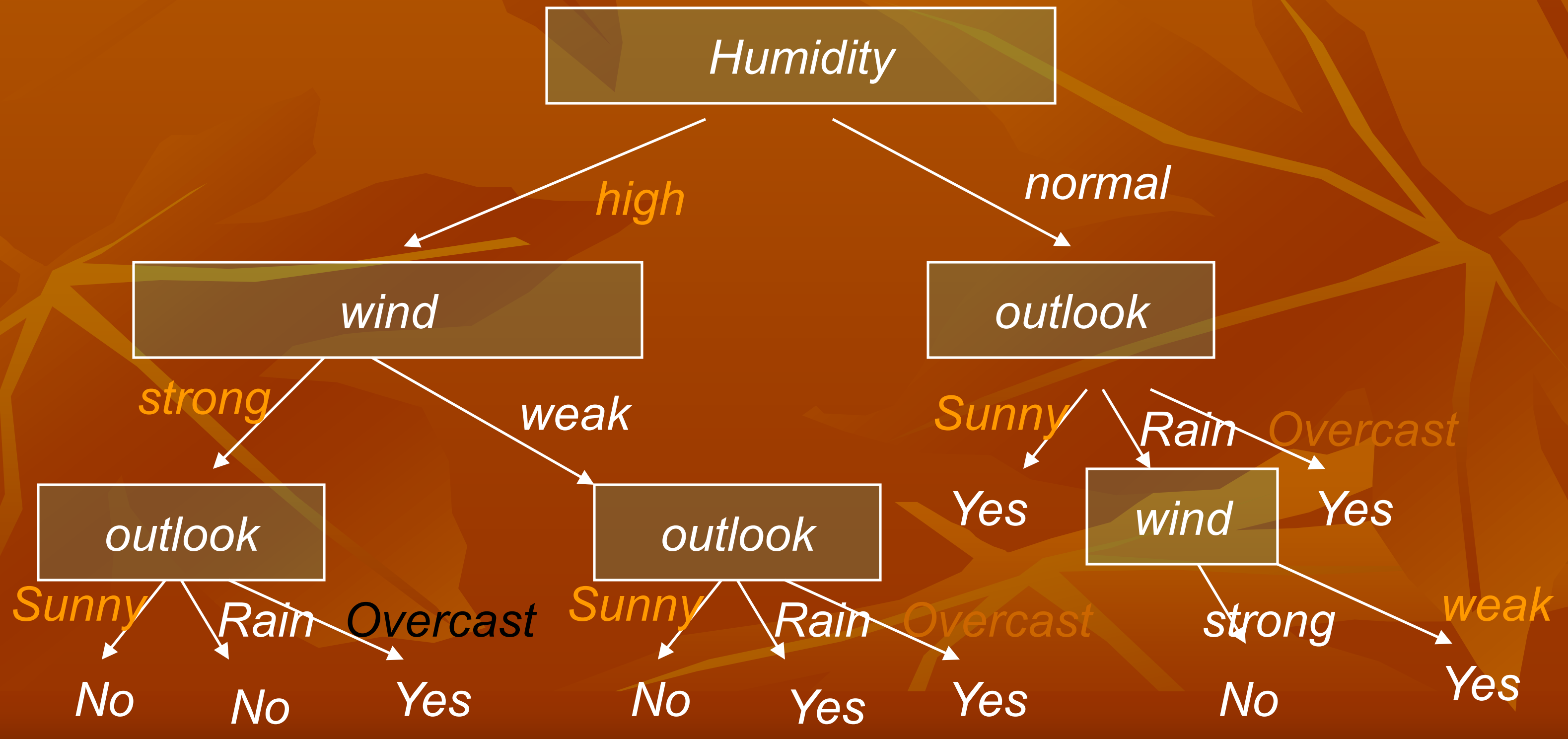


Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes

Decision Trees and Logic

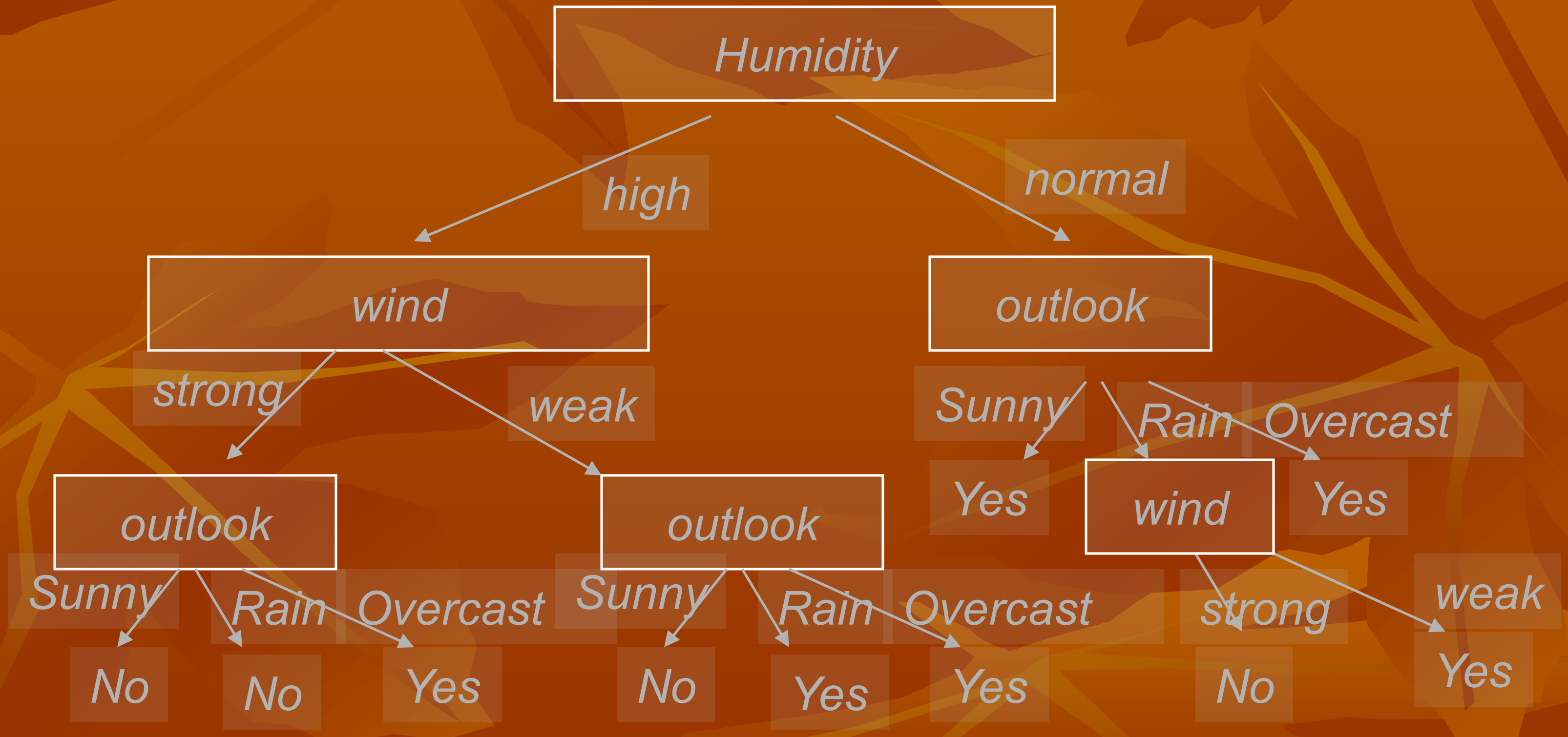
The decision tree can be expressed as an expression or if-then-else sentences:

(humidity=high ∧ wind=strong ∧ outlook=overcast) ∨
(humidity=high ∧ wind=weak ∧ outlook=overcast) ∨
(humidity=normal ∧ outlook=sunny) ∨
(humidity=normal ∧ outlook=overcast) ∨
(humidity=normal ∧ outlook=rain ∧ wind=weak) ⇒ 'Yes'



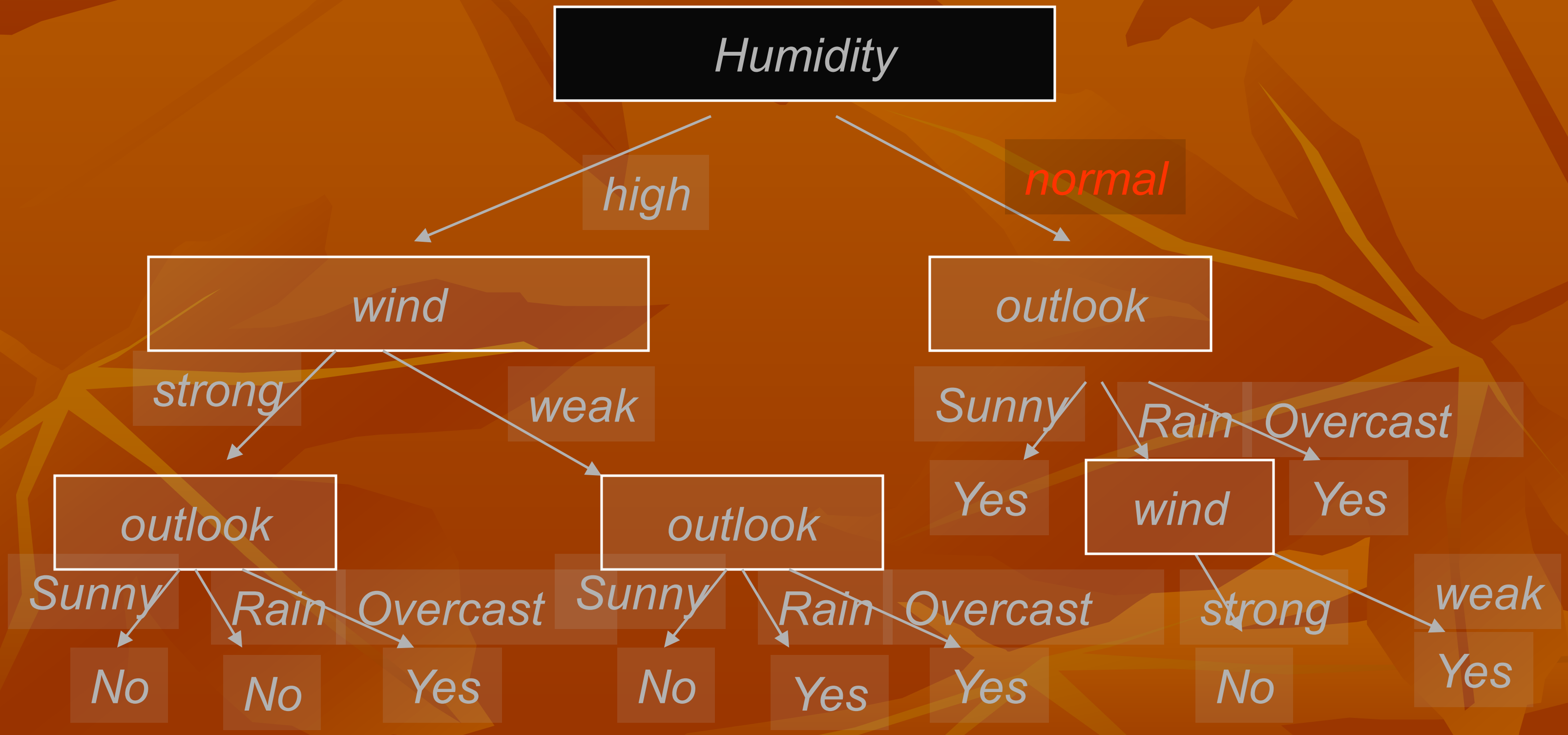
Using Decision Trees

Now let's classify an unseen example: <sunny,hot,normal,weak>=?



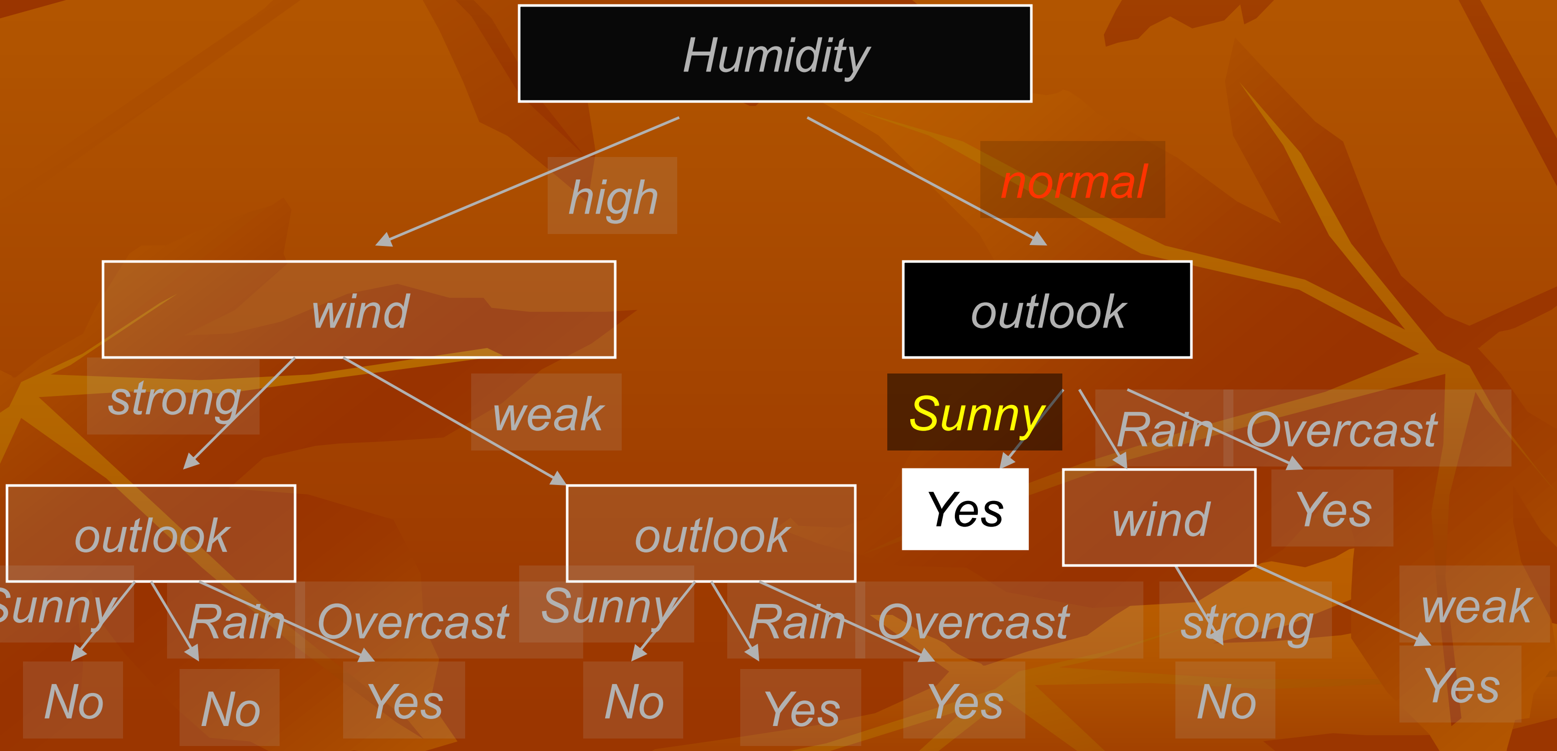
Using Decision Trees

Classifying: <sunny,hot,normal,weak>=?



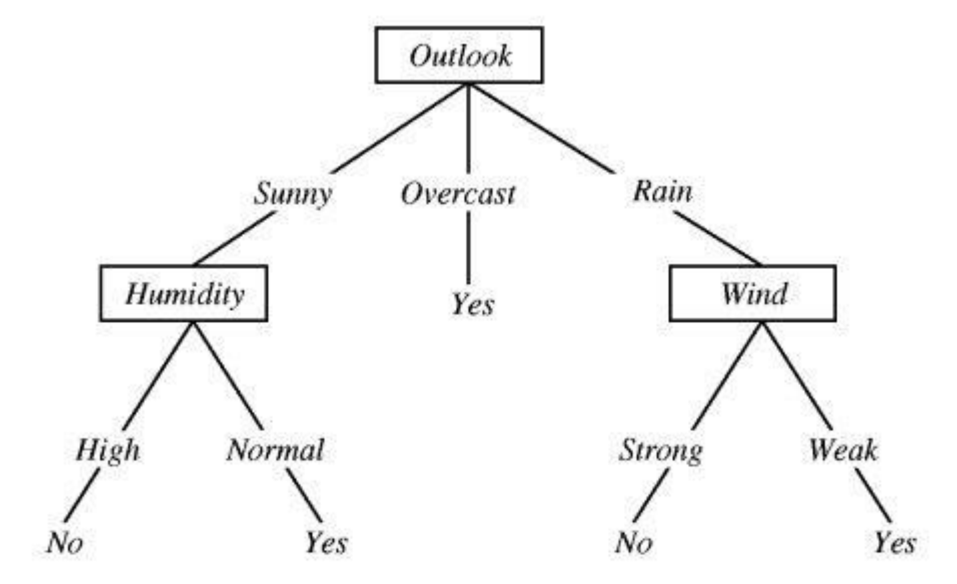
Using Decision Trees

Classification for: <sunny,hot,normal,weak>=Yes



A Big Problem…

Here's another tree from the ***same training data*** that has a different attribute order:



Which attribute should we choose for each branch?

Choosing Attributes

- We need a way of ***choosing the best attribute*** each time we add a node to the tree.
- Most commonly we use a measure called ***entropy***.
- Entropy measure the degree of ***disorder*** in a set of objects.

Entropy

- In our system we have
  - 9 positive examples
  - 5 negative examples

- The ***entropy, E(S),*** of a set of examples is:
  - $E(S) = \sum_{i=1}^{c} -p_i \log p_i$
  - Where c = no of classes and p_i = ratio of the number of examples of this value over the total number of examples.

- $P+ = 9/14$
- $P- = 5/14$
- $E = - 9/14 \log_2 9/14 - 5/14 \log_2 5/14$
- $E = \underline{0.940}$

- In a ***homogenous*** (totally ordered) system, the entropy is 0.

- In a ***totally heterogeneous*** system (totally disordered), all classes have equal numbers of instances; the entropy is 1

Entropy

- We can evaluate ***each attribute*** for their entropy.
  - E.g. evaluate the attribute "*Temperature*"
  - Three values: 'Hot', 'Mild', 'Cool.'
- So we have three subsets, one for each value of 'Temperature'.

$S_{hot}=\{D1,D2,D3,D13\}$

$S_{mild}=\{D4,D8,D10,D11,D12,D14\}$

$S_{cool}=\{D5,D6,D7,D9\}$

We will now find:

$E(S_{hot})$
$E(S_{mild})$
$E(S_{cool})$

Entropy

S_{hot}= {D1,D2,D3,D13}

Examples:
2 positive
2 negative

Totally heterogeneous + disordered therefore:
p_{+}= 0.5
p_{-}= 0.5

Entropy(S_{hot}),=
-0.5$\log_2$0.5
-0.5$\log_2$0.5 = **1.0**

S_{mild}= {D4,D8,D10, D11,D12,D14}

Examples:
4 positive
2 negative

Proportions of each class in this subset:
p_{+}= 0.666
p_{-}= 0.333

Entropy(S_{mild}),=
-0.666$\log_2$0.666
-0.333$\log_2$0.333 = **0.918**

S_{cool}={D5,D6,D7,D9}

Examples:
3 positive
1 negative

Proportions of each class in this subset:
p_{+}= 0.75
p_{-}= 0.25

Entropy(S_{cool}),=
-0.25$\log_2$0.25
-0.75$\log_2$0.75 = **0.811**

Gain

Now we can compare the entropy of the system ***before*** we divided it into subsets using "Temperature", with the entropy of the system ***afterwards***. This will tell us how good "Temperature" is as an attribute.

The entropy of the system after we use attribute "Temperature" is:

$$(|S_{hot}|/|S|)*E(S_{hot}) + (|S_{mild}|/|S|)*E(S_{mild}) + (|S_{cool}|/|S|)*E(S_{cool})$$

$$(4/14)*1.0 \quad + \quad (6/14)*0.918 \quad + \quad (4/14)*0.811 = \underline{\mathbf{0.9108}}$$

This difference between the entropy of the system before and after the split into subsets is called the ***gain***:

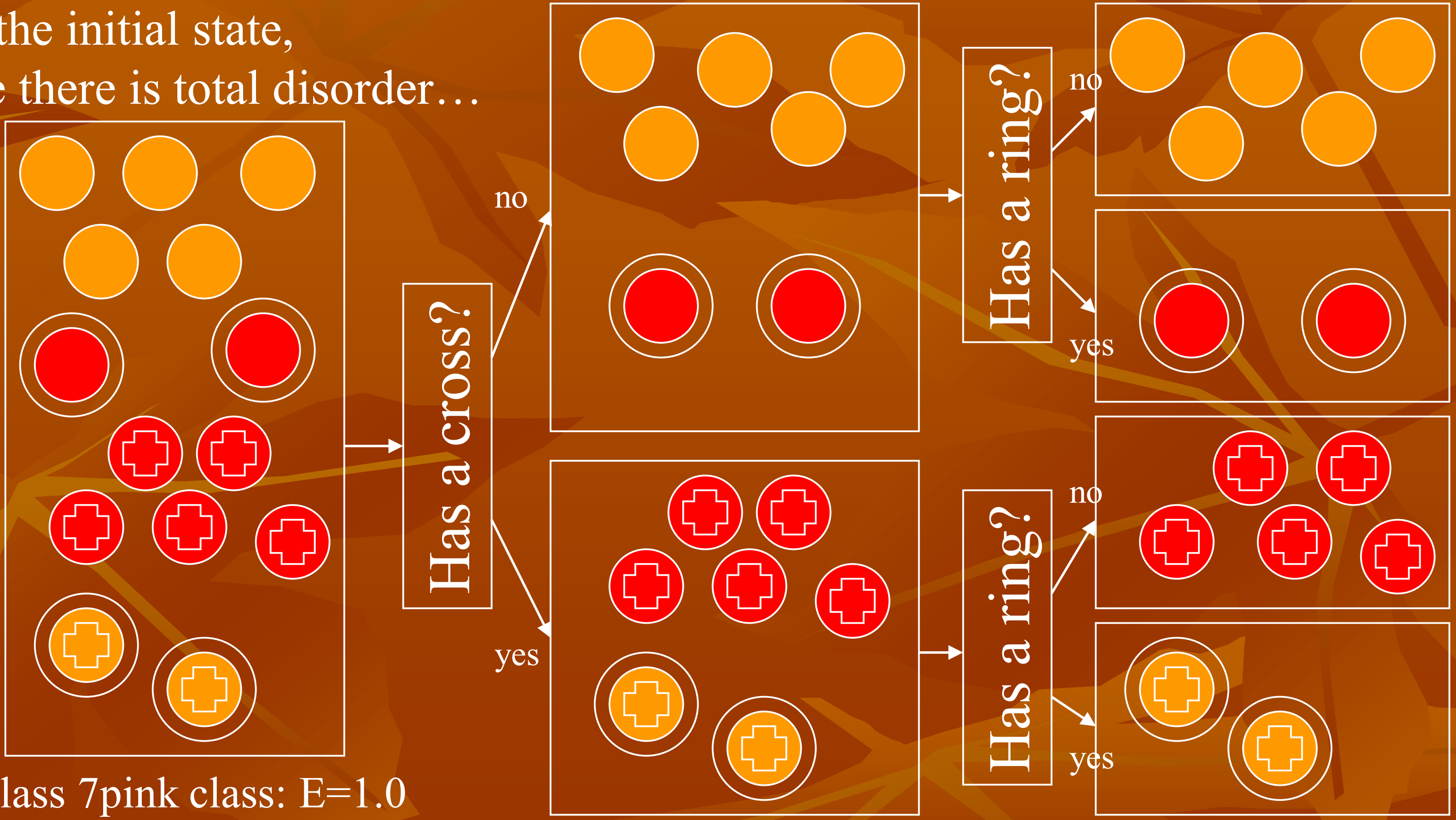
$$\text{Gain}(S, A) = \overbrace{\text{Entropy}(S)}^{\text{E(before)}} - \overbrace{\sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)}^{\text{E(afterwards)}}$$

$$\text{Gain(S,Temperature)} = \quad 0.940 \quad - \quad 0.9108 \quad = \underline{\mathbf{0.029}}$$

Decreasing Entropy

From the initial state,
Where there is total disorder…

…to the final state where all
subsets contain a single class



7red class 7pink class: E=1.0

Both subsets
E=-2/7log2/7 –5/7log5/7

All subset: E=0.0

Tabulating the Possibilities

Attribute=value	|+|	|-|	E	E after dividing by attribute A	Gain
Outlook=sunny	2	3	-2/5 log 2/5 – 3/5 log 3/5 = 0.9709	0.6935	0.2465
Outlook=o'cast	4	0	-4/4 log 4/4 – 0/4 log 0/4 = 0.0		
Outlook=rain	3	2	-3/5 log 3/5 – 2/5 log 2/5 = 0.9709		
Temp'=hot	2	2	-2/2 log 2/2 – 2/2 log 2/2 = 1.0	0.9108	0.0292
Temp'=mild	4	2	-4/6 log 4/6 – 2/6 log 2/6 = 0.9183		
Temp'=cool	3	1	-3/4 log 3/4 – 1/4 log 1/4 = 0.8112		
Etc…					

… etc

This shows the entropy calculations…

Table continued…

E for each subset of A	Weight by proportion of total	E after A is the sum of the weighted values	Gain = (E before dividing by A) – (E after A)
-2/5 log 2/5 – 3/5 log 3/5 = **0.9709**	0.9709 x 5/14 = **0.34675**	0.6935	0.2465
-4/4 log 4/4 – 0/4 log 0/4 = **0.0**	0.0 x 4/14 = **0.0**		
-3/5 log 3/5 – 2/5 log 2/5 = **0.9709**	0.9709 x 5/14 = **0.34675**		
-2/2 log 2/2 – 2/2 log 2/2 = **1.0**	1.0 x 4/14 = **0.2857**	0.9109	0.0292
-4/6 log 4/6 – 2/6 log 2/6 = **0.9183**	0.9183 x 6/14 = **0.3935**		
-3/4 log 3/4 – 1/4 log 1/4 = **0.8112**	0.8112 x 4/14 = **0.2317**		

…and this shows the gain calculations

Gain

- We calculate the gain for all the attributes.
- Then we see which of them will bring more ‘order’ to the set of examples.

- Gain(S,Outlook) = **0.246**
- Gain(S,Humidity) = 0.151
- Gain(S,Wind) = 0.048
- Gain(S, Temp’) = 0.029

- The first node in the tree should be the one with the highest value, i.e. ‘*Outlook*’.

ID3 (Decision Tree Algorithm: (Quinlan 1979))

- ID3 was the first proper decision tree algorithm to use this mechanism:

Building a decision tree with ID3 algorithm

1. *Select the attribute with the most gain*
2. *Create the subsets for each value of the attribute*
3. *For each subset*
    1. *if not all the elements of the subset belongs to same class repeat the steps 1-3 for the subset*

Main Hypothesis of ID3: The simplest tree that classifies training examples will work best on future examples (Occam's Razor)

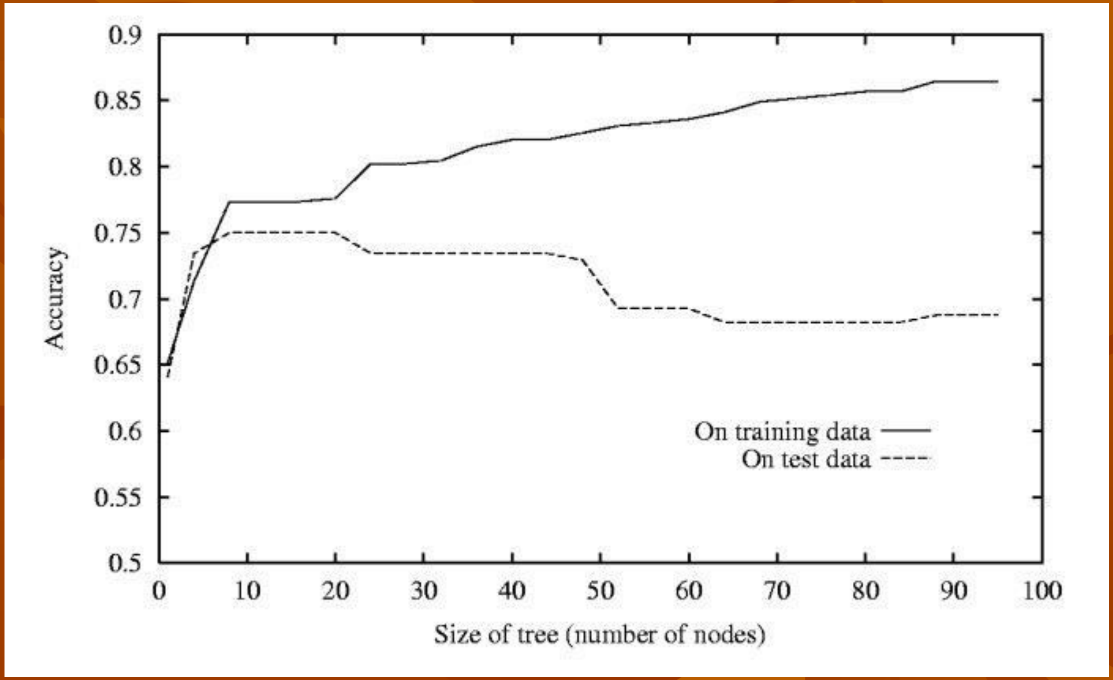
ID3 (Decision Tree Algorithm)

```
•Function DecisionTtreeLearner(Examples, TargetClass, Attributes)

•create a Root node for the tree
•if all Examples are positive, return the single-node tree Root, with label = Yes
•if all Examples are negative, return the single-node tree Root, with label = No
•if Attributes list is empty,
    • return the single-node tree Root, with label = most common value of TargetClass in Examples
•else
    •A  =  the attribute from Attributes with the highest information gain with respect to Examples
    •Make A the decision attribute for Root
    •for each possible value v of A:
        •add a new tree branch below Root, corresponding to the test A = v
        •let Examplesv be the subset of Examples that have value v for attribute A
        •if Examplesv is empty then
            •add a leaf node below this new branch with label = most common value of TargetClass in
            Examples
        •else
            •add the subtree DTL(Examplesv, TargetClass, Attributes - { A })
        •end if
•end
•return Root
```

The Problem of Overfitting

- Trees may grow to include ***irrelevant attributes***
- ***Noise*** may add ***spurious nodes*** to the tree.
- This can cause ***overfitting*** of the ***training data*** relative to ***test data.***



Hypothesis *H* **overfits** the data if there exists *H'* with greater error than *H*, over training examples, but less error than *H* over entire distribution of instances.

Fixing Over-fitting

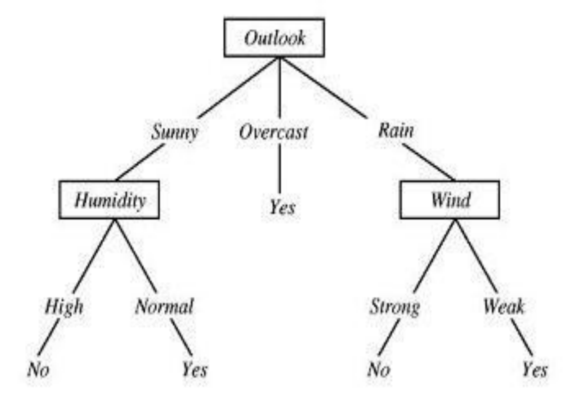
Two approaches to pruning

Prepruning: Stop growing tree during the training when it is determined that there is not enough data to make reliable choices.

Postpruning: Grow whole tree but then remove the branches that do not contribute good overall performance.

Rule Post-Pruning

Rule post-pruning



•prune (generalize) each rule by removing any preconditions (*i.e., attribute tests) that result in improving its accuracy over the validation set*

•sort pruned rules by accuracy, and consider them in this order when classifying subsequent instances

•**IF (Outlook = Sunny) ^ (Humidity = High) THEN PlayTennis = No**

•Try removing **(Outlook = Sunny)** condition or **(Humidity = High)** condition from the rule and select whichever pruning step leads to the biggest improvement in accuracy on the validation set (or else neither if no improvement results).

•converting to rules improves readability

Advantage and Disadvantages of Decision Trees

- Advantages:
  - Easy to understand and map nicely to a production rules
  - Suitable for categorical as well as numerical inputs
  - No statistical assumptions about distribution of attributes
  - Generation and application to classify unknown outputs is very fast

- Disadvantages:
  - Output attributes must be categorical
  - Unstable: slight variations in the training data may result in different attribute selections and hence different trees
  - Numerical input attributes leads to complex trees as attribute splits are usually binary

Assignment

Given the training data set, to identify whether a customer buys computer or not, Develop a Decision Tree using ID3 technique.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Association Rules

- Example1: a female shopper buys a handbag is likely to buy shoes
- Example2: when a male customer buys beer, he is likely to buy salted peanuts

- It is not very difficult to develop algorithms that will find this associations in a large database
  - The problem is that such an algorithm will also uncover many other associations that are of very little value.

Association Rules

- It is necessary to introduce some measures to distinguish interesting associations from non-interesting ones

- Look for associations that have a lots of examples in the database: support of an association rule

- May be that a considerable group of people who read all three magazines but there is a much larger group that buys A & B, but not C; association is very weak here although support might be very high.

Associations….

- Percentage of records for which C holds, within the group of records for which A & B hold: confidence

- Association rules are only useful in data mining if we already have a rough idea of what is we are looking for.

- We will represent an association rule in the following way:
  - MUSIC_MAG, HOUSE_MAG=>CAR_MAG
  - *Somebody that reads both a music and a house magazine is also very likely to read a car magazine*

Associations…

- Example: shopping Basket analysis

Transactions	Chips	Rasbari	Samosa	Coke	Tea
T1		X	X		
T2	X	X			
T3		X	X		X

Example…

- 1. find all frequent Itemsets:
- (a) 1-itemsets
  - K= [{Chips}C=1,{Rasbari}C=3,{Samosa}C=2, {Tea}C=1]
- (b) extend to 2-itemsets:
  - L=[{Chips, Rasbari}C=1, {Rasbari,Samosa}C=2,{Rasbari,Tea}C=1,{Samosa,Tea}C=1]
- (c) Extend to 3-Itemsets:
  - M=[{Rasbari, Samosa,Tea}C=1

Examples..

- Match with the requirements:
  - Min. Support is 2 (66%)
  - (a) >> K1={{Rasbari}, {Samosa}}
  - (b) >>L1={Rasbari,Samosa}
  - (c) >>M1={}
- Build All possible rules:
  - (a) no rule
  - (b) >> possible rules:
    - Rasbari=>Samosa
    - Samosa=>Rasbari
  - (c) No rule
- ***Support:*** *given the association rule X1,X2…Xn=>Y, the support is the Percentage of records for which X1,X2…Xn and Y both hold true.*

Example..

- Calculate Confidence for b:
  - Confidence of [Rasbari=>Samosa]
    - {Rasbari,Samosa}C=2/{Rasbari}C=3
    - =2/3
    - 66%
  - Confidence of Samosa=> Rasbari
    - {Rasbari,Samosa}C=2/{Samosa}C=2
    - =2/2
    - 100%
- ***Confidence**: Given the association rule X1,X2….Xn=>Y, the confidence is the percentage of records for which Y holds within the group of records for which X1,X2…Xn holds true.*

The A-Priori Algorithm

- Set the threshold for *support* rather high – to focus on a small number of best candidates,

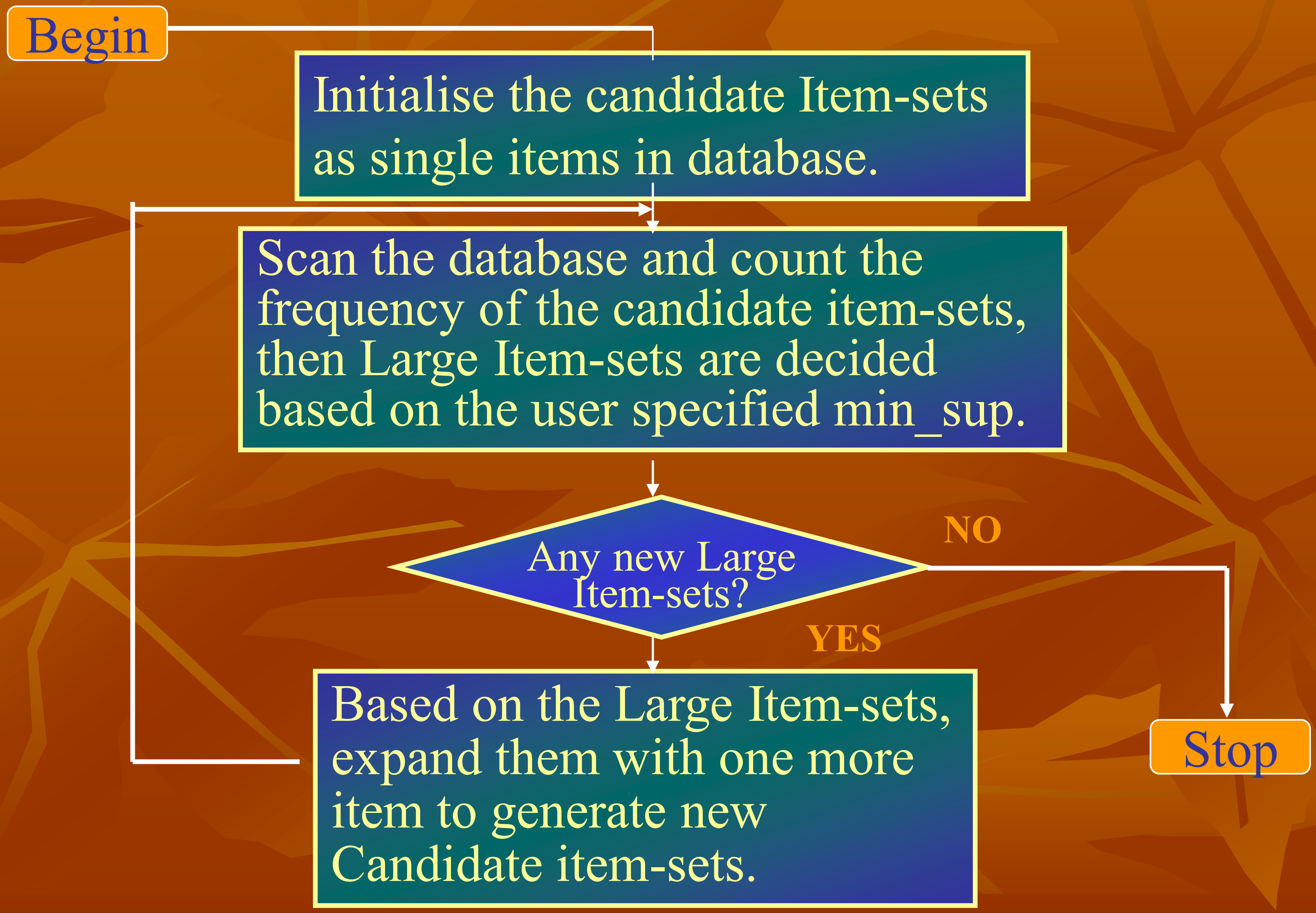
- Observation: *If a set of items X has support s, then each subset of X must also have support at least s.*

(if a pair {i,j} appears in say, 1000 baskets, then we know there are at least 1000 baskets with item i and at least 1000 baskets with item j)

Algorithm:

1) Find the set of candidate items – those that appear in a sufficient number of baskets by themselves
2) Run the query on only the candidate items

Apriori Algorithm



Begin

Initialise the candidate Item-sets as single items in database.

Scan the database and count the frequency of the candidate item-sets, then Large Item-sets are decided based on the user specified min_sup.

Any new Large Item-sets?

NO

YES

Based on the Large Item-sets, expand them with one more item to generate new Candidate item-sets.

Stop

Apriori: A Candidate Generation-and-test Approach

- Any subset of a frequent itemset must be frequent
  - if **{beer, diaper, nuts}** is frequent, so is **{beer, diaper}**
  - Every transaction having {beer, diaper, nuts} also contains {beer, diaper}
- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!
- The performance studies show its efficiency and scalability

The Apriori Algorithm — An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

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

L_1

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

C_2

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

2nd scan

C_2

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

L_2

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

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Problems with A-priori Algorithms

- It is costly to handle a huge number of candidate sets. For example if there are 10^4 large 1-itemsets, the Apriori algorithm will need to generate more than 10^7 candidate 2-itemsets. Moreover for 100-itemsets, it must generate more than $2^{100} \approx 10^{30}$ candidates in total.

- The candidate generation is the inherent cost of the Apriori Algorithms, no matter what implementation technique is applied.
- To mine a large data sets for long patterns – this algorithm is NOT a good idea.
- When Database is scanned to check Ck for creating Lk, a large number of transactions will be scanned even they do not contain any k-itemset.

Artificial Neural Network: Outline

- Perceptrons
- Multi-layer networks
- Backpropagation

  - Neuron switching time : > 10^{-3} secs
  - Number of neurons in the human brain: ~10^{11}
  - Connections (synapses) per neuron : ~10^{4}–10^{5}
  - Face recognition : 0.1 secs
  - High degree of parallel computation
  - Distributed representations

Human Brain

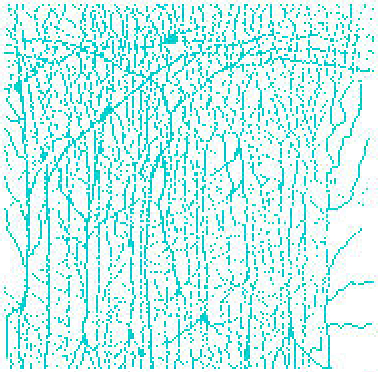
- Computers and the Brain: A Contrast
  - Arithmetic: 1 brain = 1/10 pocket calculator
  - Vision: 1 brain = 1000 super computers
  - Memory of arbitrary details: computer wins
  - Memory of real-world facts: brain wins
  - A computer must be programmed explicitly
  - The brain can learn by experiencing the world

Definition

- ". . . Neural nets are basically mathematical models of information processing . . ."
- ". . . (neural nets) refer to machines that have a structure that, at some level, reflects what is known of the structure of the brain . . ."
- "A neural network is a massively parallel distributed processor . . . "

Properties of the Brain

- Architectural
  - **80,000 neurons per square mm**
  - **10^{11} neurons - 10^{15} connections**
  - **Most axons extend less than 1 mm (local connections)**
- Operational
  - **Highly complex, nonlinear, parallel computer**
  - **Operates at millisecond speeds**



A typical network of neurons

Interconnectedness

- Each neuron may have over a thousand synapses
- Some cells in cerebral cortex may have 200,000 connections
- Total number of connections in the brain "network" is astronomical—greater than the number of particles in known universe

Brain and Nervous System

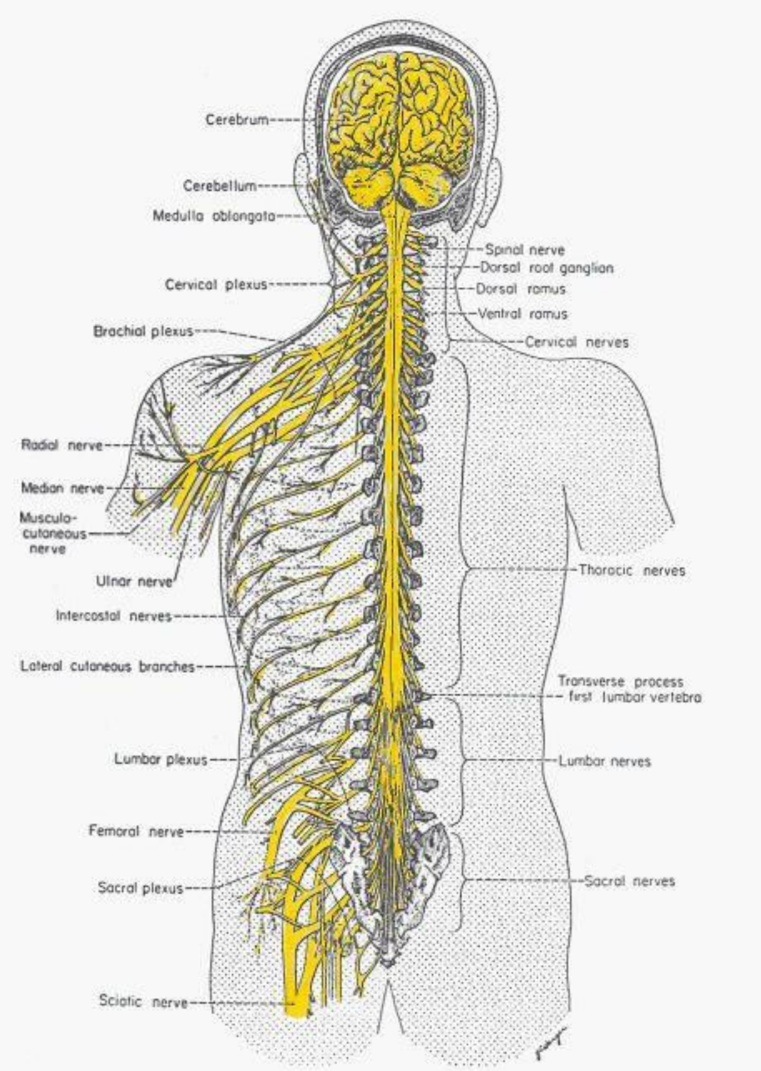
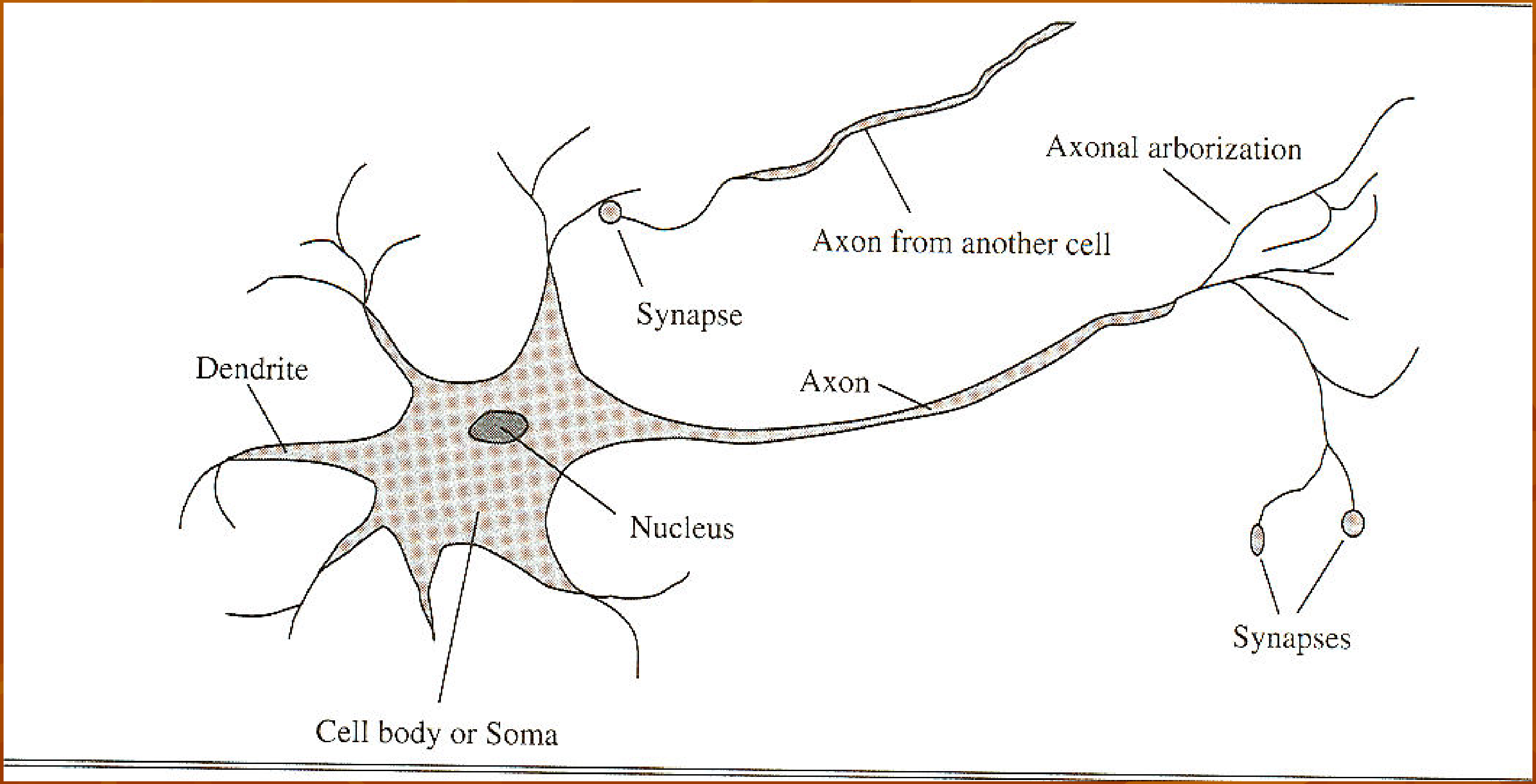


Fig. 2. The human central nervous system, exposed by dissection from the dorsal aspect. Shows the brain, spinal cord and the proximal parts of the spinal nerves. Compare this with the generalized vertebrate plan shown in Figure 1.

- Around 100 billion neurons in the human brain.
- Each of these is connected to many other neurons (typically 10000 connections)
- Regions of the brain are (somewhat) specialised.
- Some neurons connect to senses (input) and muscles (action).

Detail of a Neuron



The Question

Humans find these tasks relatively simple

- We learn by example
- The brain is responsible for our ‘computing’ power

If a machine were constructed using the fundamental building blocks found in the *brain* could it *learn* to do ‘difficult’ tasks ???

Basic Ideas in Machine Learning

- Machine learning is focused on *inductive* learning of hypotheses from examples.
- Three main forms of learning:
  - *Supervised learning*: Examples are tagged with some “expert” information.
  - *Unsupervised learning*: Examples are placed into categories without guidance; instead, generic properties such as “similarity” are used.
  - *Reinforcement learning*: Examples are tested, and the results of those tests used to drive learning.

Neural Network: Characteristics

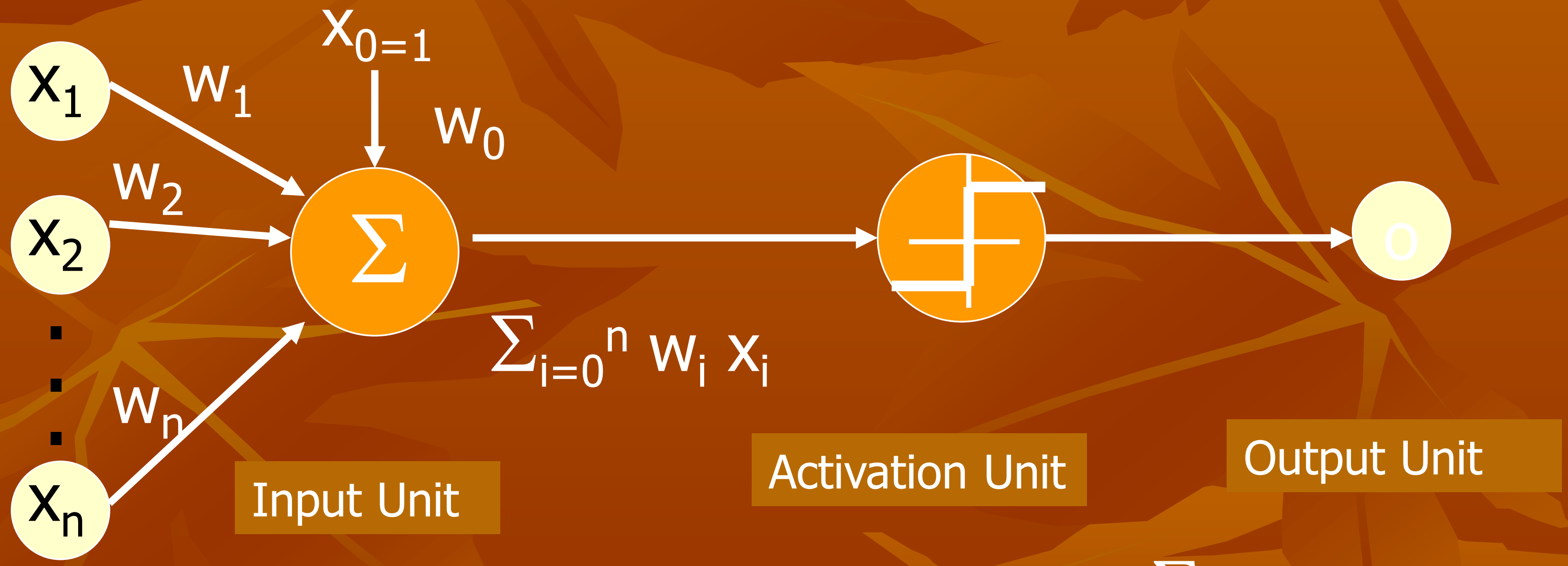
- Highly parallel structure; hence a capability for fast computing
- Ability to learn and adapt to changing system parameters
- High degree of tolerance to damage in the connections
- Ability to learn through parallel and distributed processing

Neural Networks

- A neural Network is composed of a number of nodes, or units, connected by links. Each link has a numeric weight associated with it.

- Each unit has a set of input links from other units, a set of output links to other units, a current activation level, and a means of computing the activation level at the next step in time.

- Linear treshold unit (LTU)

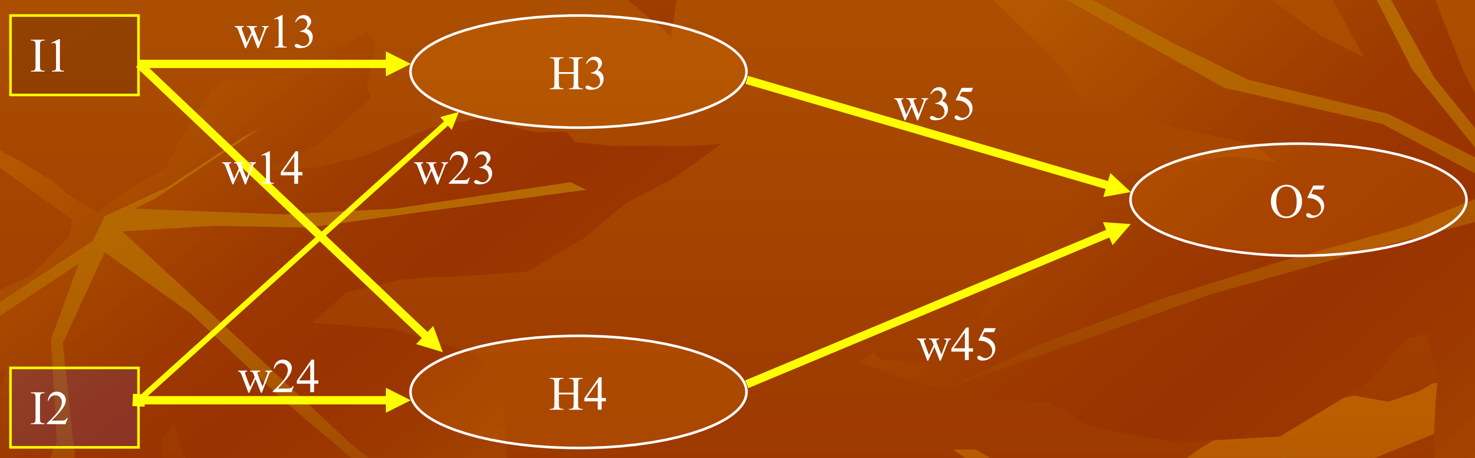


$\Sigma_{i=0}^{n} w_i x_i$

$$o(x_i) = \begin{cases} 1 \text{ if } \Sigma_{i=0}^{n} w_i x_i > 0 \\ -1 \text{ otherwise} \end{cases}$$

Layered network

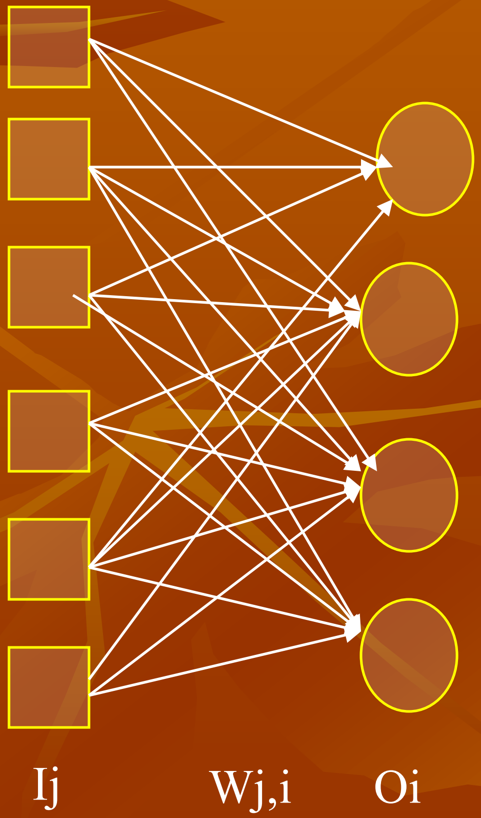
- Single layered
- Multi layered

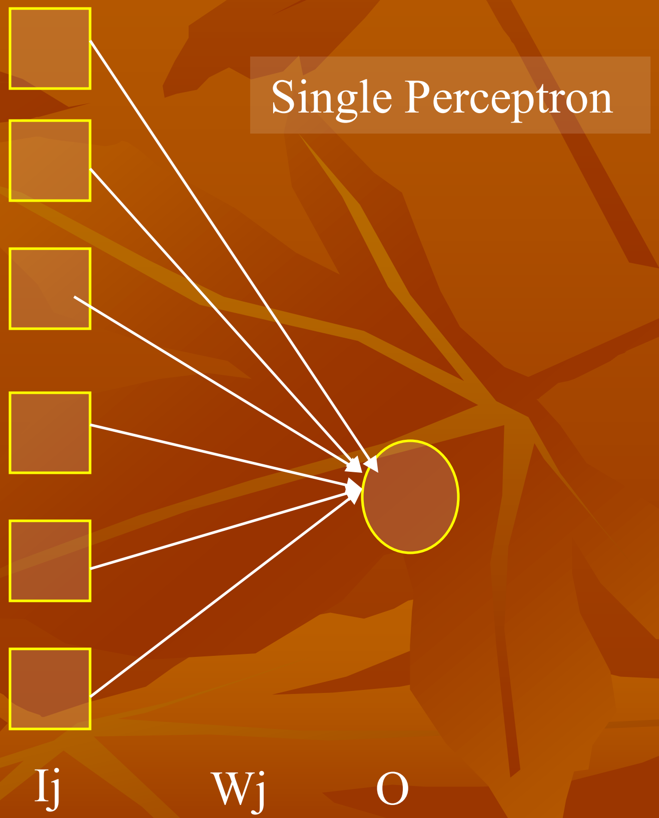


Two layer, feed forward network with two inputs, two hidden nodes and one output node.

Perceptrons

- A single-layered, feed-forward network can be taken as a perceptron.





Perceptron Learning Rule

$w_i = w_i + \Delta w_i$

$\Delta w_i = \eta\ (t - o)\ x_i$

t=c(x) is the target value

o is the perceptron output

η Is a small constant (e.g. 0.1) called *learning rate*

- If the output is correct (t=o) the weights w_i are not changed
- If the output is incorrect ($t \neq o$) the weights w_i are changed such that the output of the perceptron for the new weights is *closer* to t.

Genetic Algorithm

- Derived inspiration from biology
- The most fertile area for exchange of views between biology and computer science is ‘evolutionary computing’
- This area evolved from three stages or less independent development:
    - Genetic algorithms
    - Evolutionary programming
    - Evolution strategies

GA..

- The investigators began to see a strong relationship between these areas, and at present, genetic algorithms are consideered to be among the most successful machine-learning techniques.

- In the “origin of species”, Darwin described the theory of evolution, with the ‘natural selection’ as the central notion.
  - Each species has an overproduction of individuals and in a tough struggle for life, only those individuals that are best adapted to the environment survive.

- The long DNA molecules, consisting of only four building blocks, suggest that all the heriditary information of a human individual, or of any living creature, has been laid down in a language of only four letters (C,G,A & T in language of genetics)

How large is the decision space?

- If we were to look at every alternative, what would we have to do? Of course, it depends.....

- Think: enzymes
  - Catalyze all reactions in the cell
  - Biological enzymes are composed of amino acids
  - There are 20 naturally-occurring amino acids
  - Easily, enzymes are 1000 amino acids long
  - 20^1000 = (2^1000)(10^1000) ≈ 10^1300

- A reference number, a benchmark:

10^80 ≈ number of atomic particles in universe

How large is the decision space?

- Problem: Design an icon in black & white
  How many options?
  - Icon is 32 x 32 = 1024 pixels
  - Each pixel can be on or off, so 2^1024 options
  - 2^1024 ≈ (2^20)^50 ≈ (10^6)^50 = 10^300
- Police faces
  - 10 types of eyes
  - 10 types of noses
  - 10 types of eyebrows
  - 10 types of head
  - 10 types of head shape
  - 10 types of mouth
  - 10 types of ears
  - but already we have 10^7 faces

GA..

- The collection of genetic instruction for human is about 3 billion letters long
  - Each individual inherits some characteristics of the father and some of the mother.
  - Individual differences between people, such as hair color and eye color, and also pre-disposition for diseases, are caused by differences in genetic coding
    - Even the twins are different in numerous aspects.

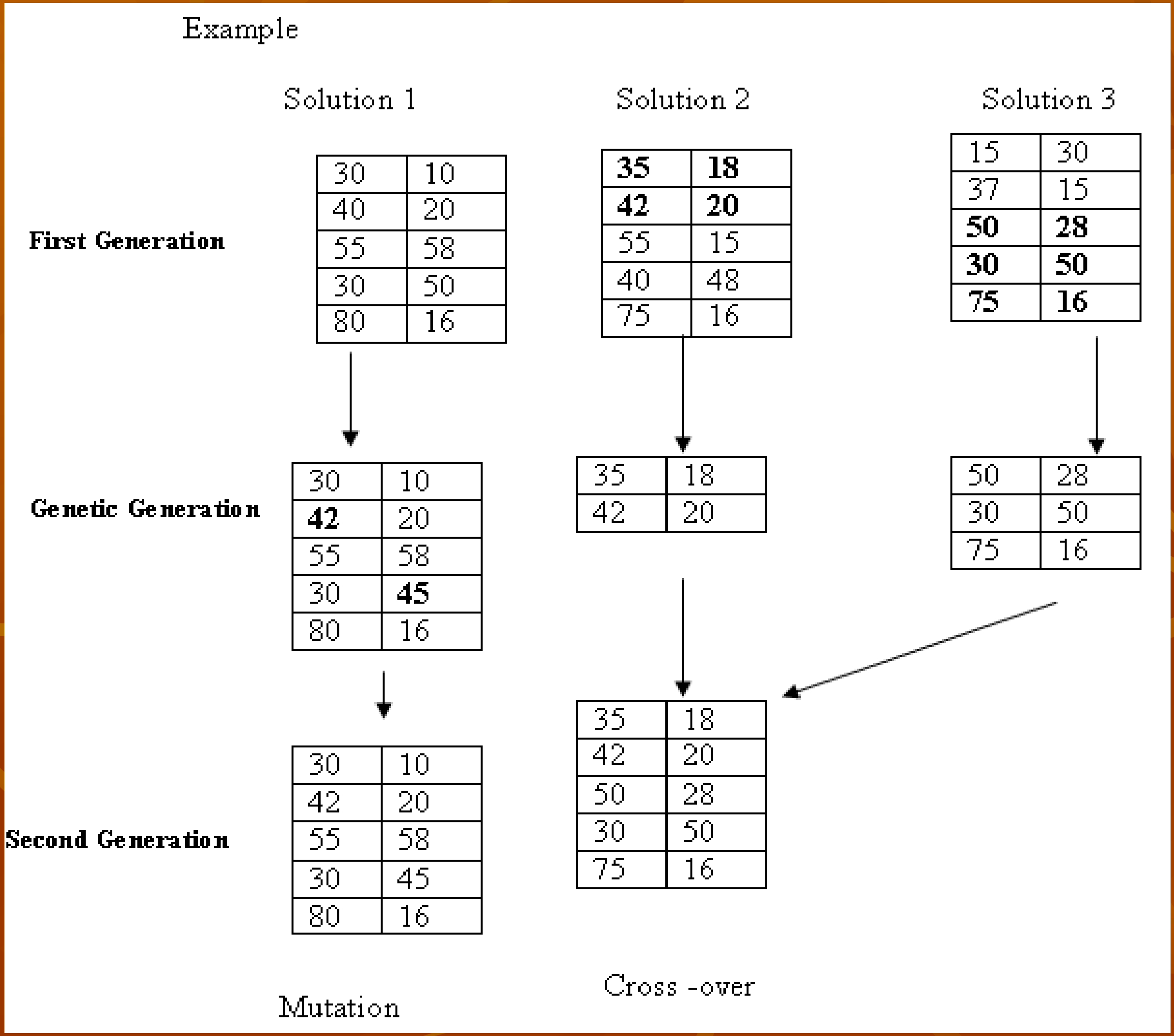
Genetic Algorithm Components

- Selection
  - determines how many and which individuals breed
  - premature convergence sacrifices solution quality for speed
- Crossover
  - select a random crossover point
  - successfully exchange substructures
  - 00000 x 11111 at point 2 yields 00111 and 11000
- Mutation
  - random changes in the genetic material (bit pattern)
  - for problems with billions of local optima, mutations help find the global optimum solution
- Evaluator function
  - rank fitness of each individual in the population
  - simple function (maximum) or complex function

GA..

- Following are the formula for constructing a genetic algorithm for the solution of problem
  - Write a good coding in terms of strings of limited alphabets
  - Invent an artificial environment in the computer where solution can join each other
  - Develop ways in which possible solutions can be combined. Like father's and mother's strings are simply cut and after changing, stuck together again called cross- over
  - Provide an initial population or solution set and make the computer play evolution by removing bad solutions from each generation and replacing them with mutations of good solutions
  - Stop when a family of successful solutions has been produced

Example



Example

First Generation

Solution 1

30	10
40	20
55	58
30	50
80	16

Solution 2

35	**18**
42	**20**
55	15
40	48
75	16

Solution 3

15	30
37	15
50	**28**
30	**50**
75	**16**

Genetic Generation

30	10
42	20
55	58
30	**45**
80	16

35	18
42	20

50	28
30	50
75	16

Second Generation

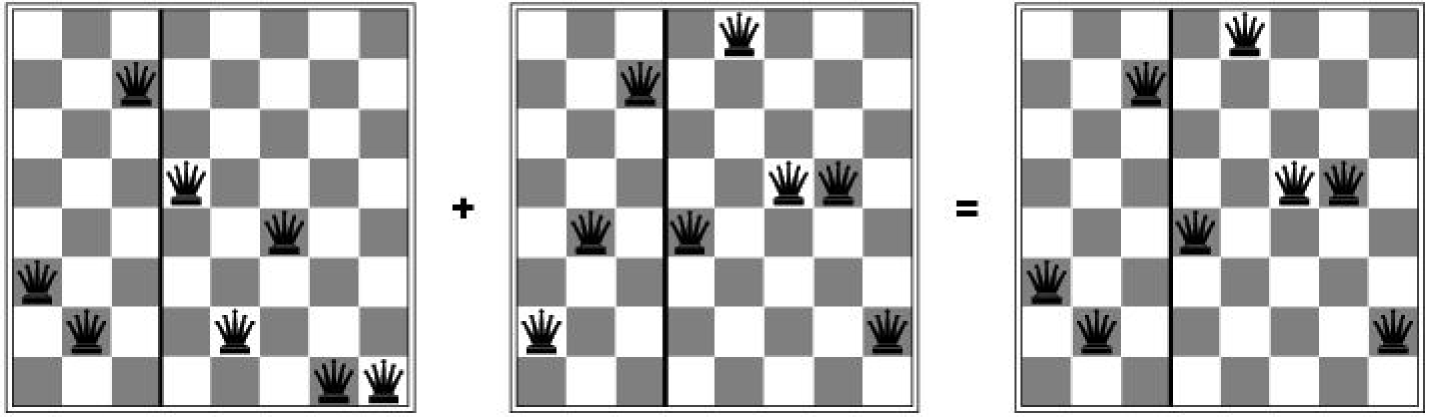
30	10
42	20
55	58
30	45
80	16

Mutation

35	18
42	20
50	28
30	50
75	16

Cross -over

Genetic algorithms



Clustering

By : Babu Ram Dawadi

Clustering

- cluster is a collection of data objects, in which the objects similar to one another within the same cluster and dissimilar to the objects in other clusters

- Cluster analysis is the process of finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

- Clustering: Given a database D = {t1, t2, .., tn}, a distance measure dis(ti, tj) defined between any two objects ti and tj, and an integer value k, the clustering problem is to define a mapping f: D -> {1, …, k} where each ti is assigned to one cluster Kj, 1<=j<=k. here 'k' is the number of clusters.

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Categories of Clustering

main categories (classes) of clustering methods

- Partition-based
- Hierarchical
- Density-based
- Grid-based
- Model-based

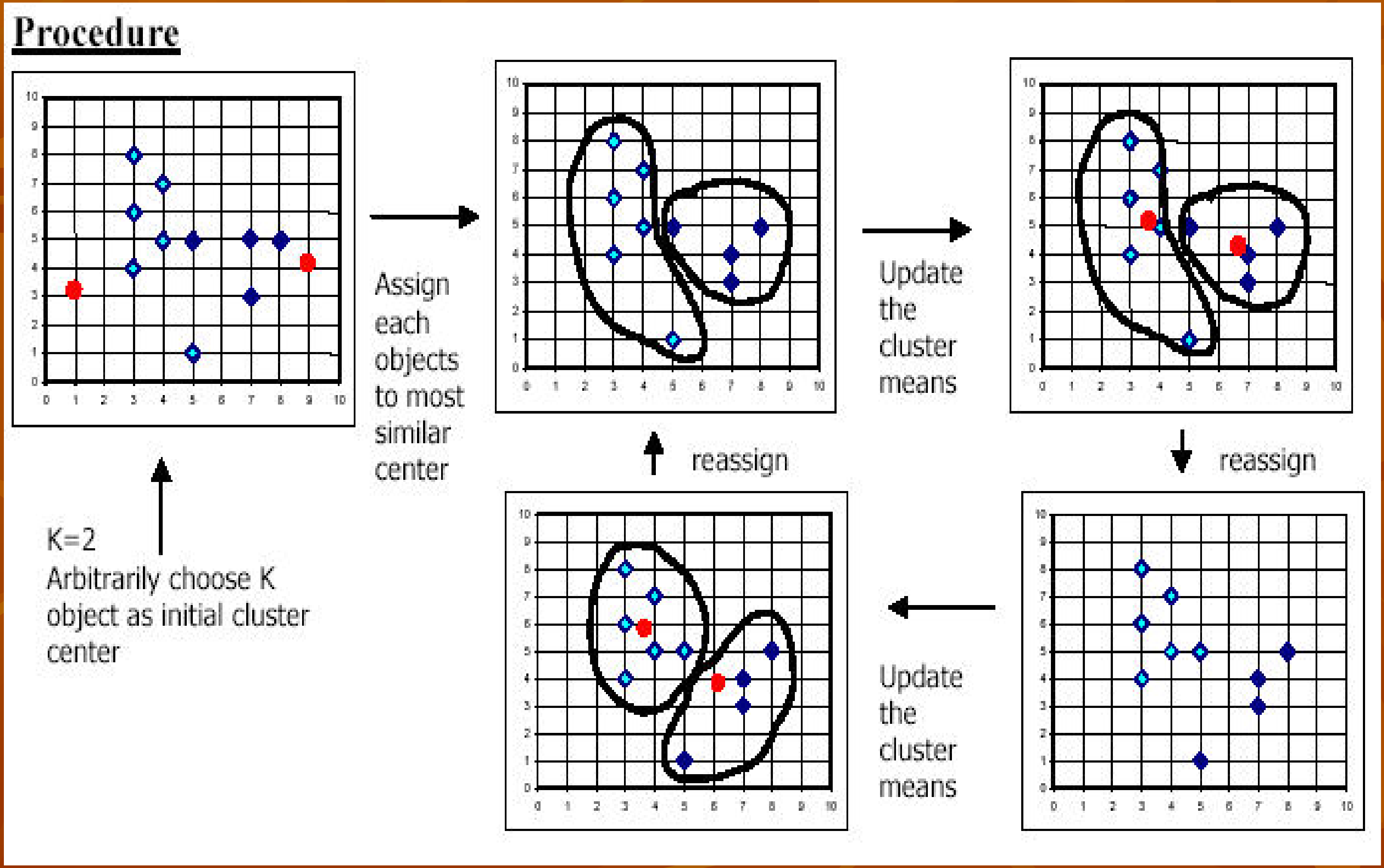
Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids): Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

1. Chose k number of clusters to be determined
2. Chose k objects randomly as the initial cluster centers
3. Repeat
    1. Assign each object to their closest cluster center, using Euclidean distance
    2. Compute new cluster centers, calculate mean point
4. Until
    1. No change in cluster centers or
    2. No object change its clusters

The *K-Means* Clustering Method



Procedure

K=2
Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

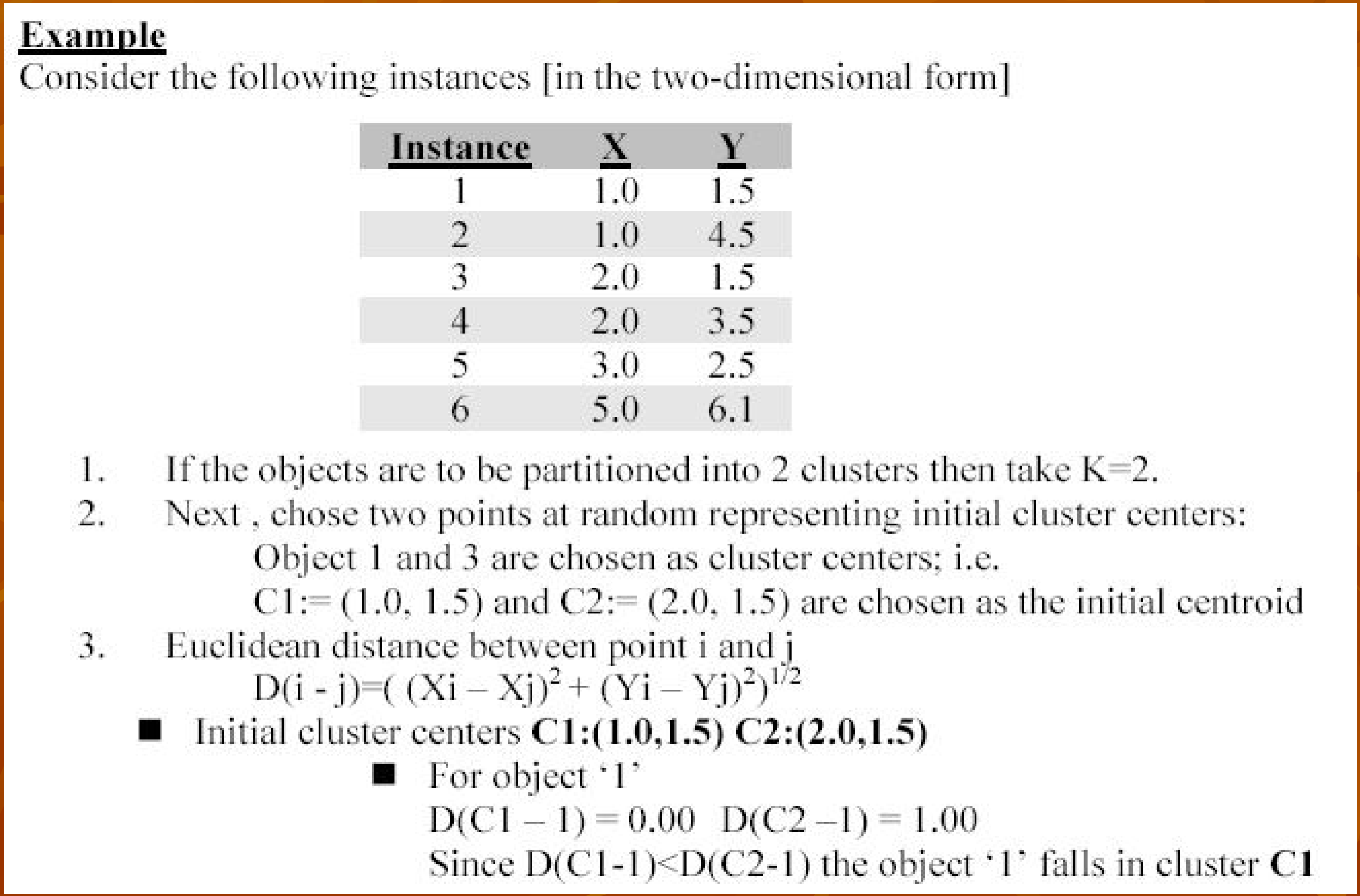
Update the cluster means

reassign

Update the cluster means

reassign

K-Means Clustering



Example

Consider the following instances [in the two-dimensional form]

Instance	X	Y
1	1.0	1.5
2	1.0	4.5
3	2.0	1.5
4	2.0	3.5
5	3.0	2.5
6	5.0	6.1

1. If the objects are to be partitioned into 2 clusters then take K=2.
2. Next , chose two points at random representing initial cluster centers:

   Object 1 and 3 are chosen as cluster centers; i.e.

   C1:= (1.0, 1.5) and C2:= (2.0, 1.5) are chosen as the initial centroid
3. Euclidean distance between point i and j

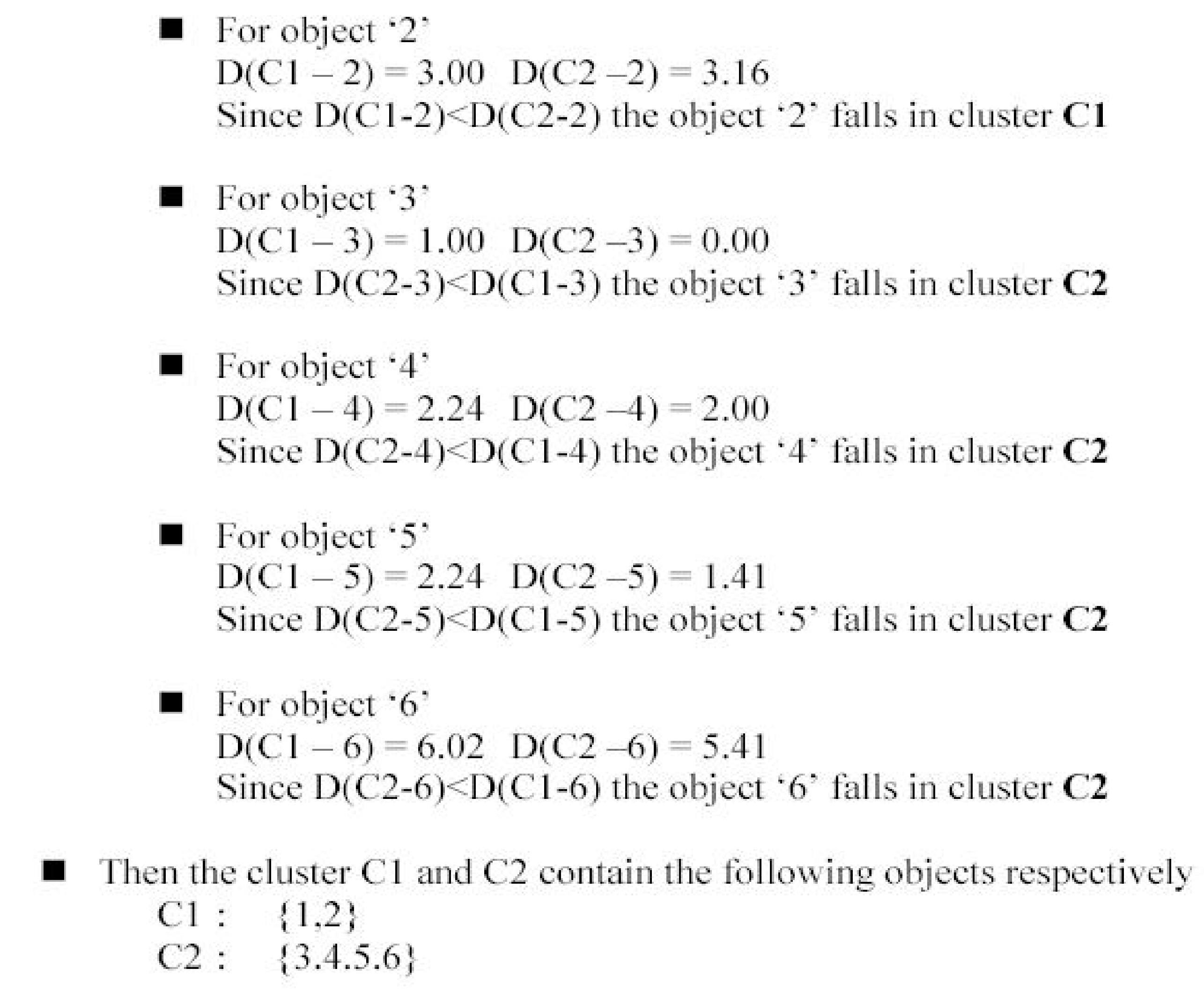
   $D(i - j)=((Xi - Xj)^2 + (Yi - Yj)^2)^{1/2}$

- Initial cluster centers **C1:(1.0,1.5) C2:(2.0,1.5)**
  - For object '1'

    D(C1 – 1) = 0.00   D(C2 –1) = 1.00

    Since D(C1-1)<D(C2-1) the object '1' falls in cluster **C1**

K-Means Clustering



- For object '2'
  D(C1 – 2) = 3.00   D(C2 –2) = 3.16
  Since D(C1-2)<D(C2-2) the object '2' falls in cluster **C1**

- For object '3'
  D(C1 – 3) = 1.00   D(C2 –3) = 0.00
  Since D(C2-3)<D(C1-3) the object '3' falls in cluster **C2**

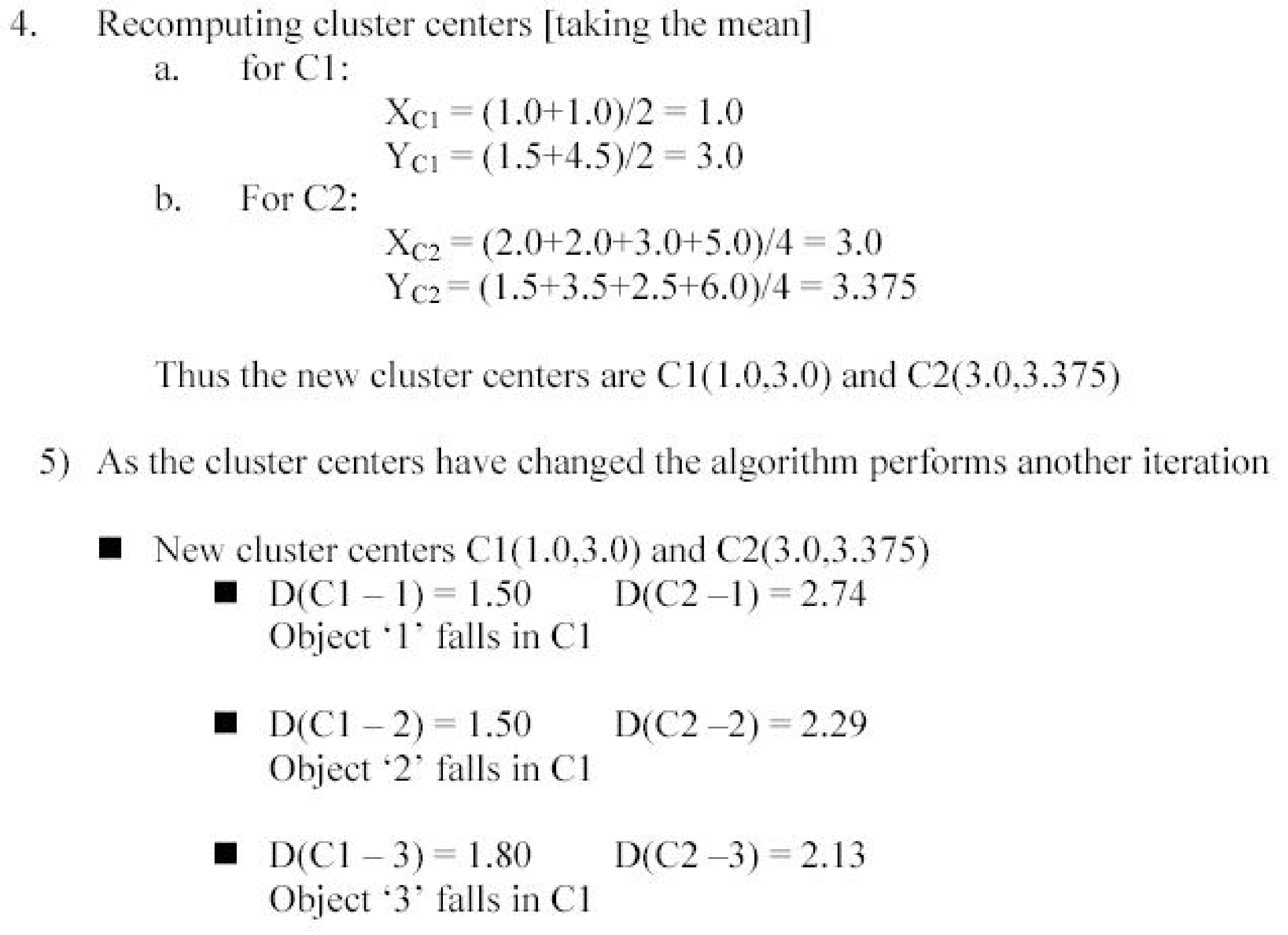
- For object '4'
  D(C1 – 4) = 2.24   D(C2 –4) = 2.00
  Since D(C2-4)<D(C1-4) the object '4' falls in cluster **C2**

- For object '5'
  D(C1 – 5) = 2.24   D(C2 –5) = 1.41
  Since D(C2-5)<D(C1-5) the object '5' falls in cluster **C2**

- For object '6'
  D(C1 – 6) = 6.02   D(C2 –6) = 5.41
  Since D(C2-6)<D(C1-6) the object '6' falls in cluster **C2**

- Then the cluster C1 and C2 contain the following objects respectively
  C1 :   {1.2}
  C2 :   {3.4.5.6}

K-Means Clustering



4. Recomputing cluster centers [taking the mean]

   a. for C1:

      $X_{C1} = (1.0+1.0)/2 = 1.0$

      $Y_{C1} = (1.5+4.5)/2 = 3.0$

   b. For C2:

      $X_{C2} = (2.0+2.0+3.0+5.0)/4 = 3.0$

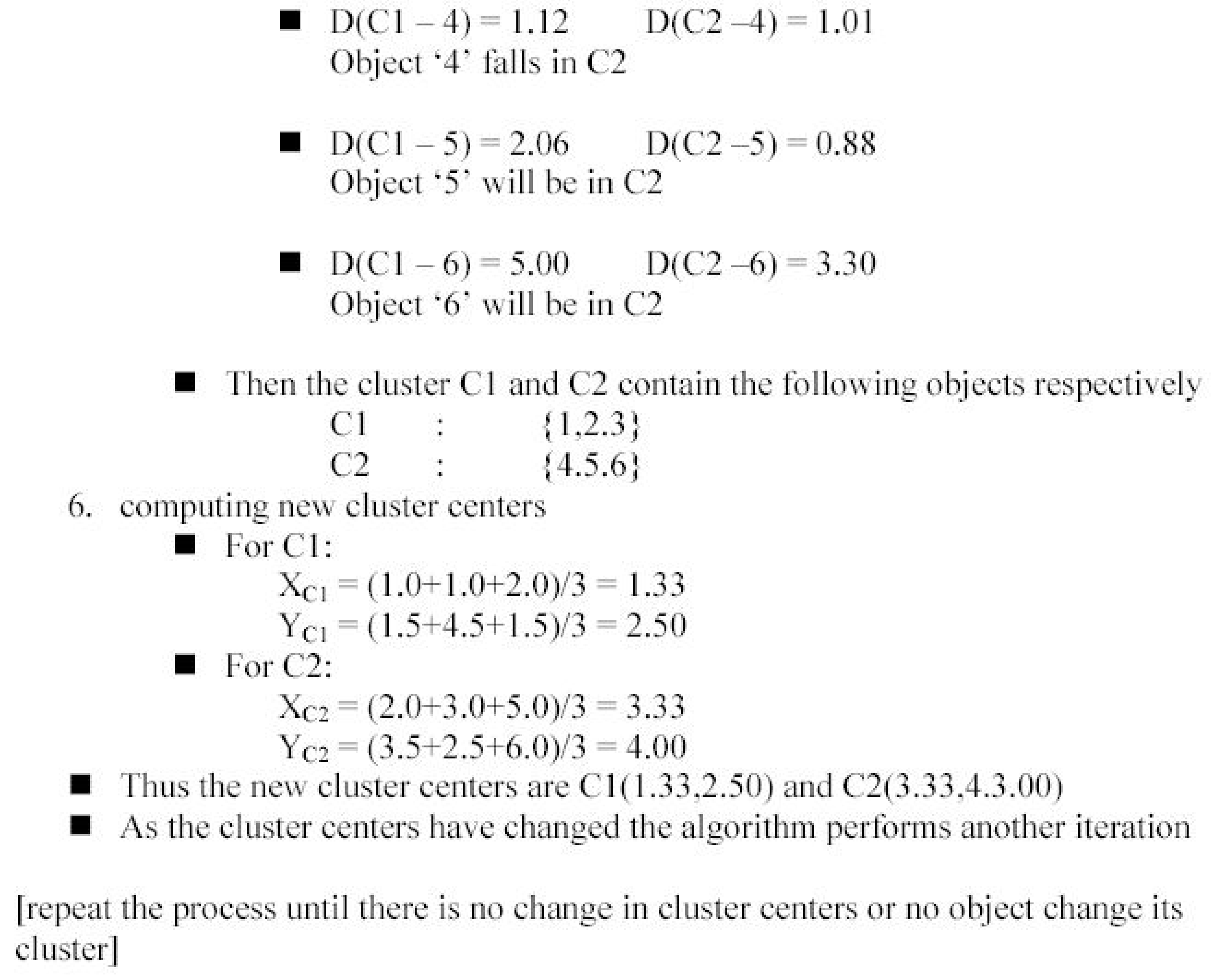
      $Y_{C2} = (1.5+3.5+2.5+6.0)/4 = 3.375$

   Thus the new cluster centers are C1(1.0,3.0) and C2(3.0,3.375)

5) As the cluster centers have changed the algorithm performs another iteration

- New cluster centers C1(1.0,3.0) and C2(3.0,3.375)
  - D(C1 – 1) = 1.50 D(C2 –1) = 2.74
Object ‘1’ falls in C1
  - D(C1 – 2) = 1.50 D(C2 –2) = 2.29
Object ‘2’ falls in C1
  - D(C1 – 3) = 1.80 D(C2 –3) = 2.13
Object ‘3’ falls in C1

K-Means Clustering



- D(C1 – 4) = 1.12 D(C2 –4) = 1.01
  Object '4' falls in C2

- D(C1 – 5) = 2.06 D(C2 –5) = 0.88
  Object '5' will be in C2

- D(C1 – 6) = 5.00 D(C2 –6) = 3.30
  Object '6' will be in C2

- Then the cluster C1 and C2 contain the following objects respectively
  C1 : {1,2.3}
  C2 : {4.5.6}

6. computing new cluster centers
   - For C1:
     $X_{C1} = (1.0+1.0+2.0)/3 = 1.33$
     $Y_{C1} = (1.5+4.5+1.5)/3 = 2.50$
   - For C2:
     $X_{C2} = (2.0+3.0+5.0)/3 = 3.33$
     $Y_{C2} = (3.5+2.5+6.0)/3 = 4.00$

- Thus the new cluster centers are C1(1.33,2.50) and C2(3.33,4.3.00)
- As the cluster centers have changed the algorithm performs another iteration

[repeat the process until there is no change in cluster centers or no object change its cluster]

Weakness of K-means

- Applicable only when *mean* is defined, then what about categorical data?
- Need to specify *K,* the *number* of clusters, in advance
  - run the algorithm with different K values
- Unable to handle noisy data and *outliers*
- Works best when clusters are of approximately of equal size

Hierarchical Clustering

Clustering comes in a form of a tree – *dendrogram* visualizing how data contribute to individual clusters

Clustering is realized in a successive manner through:

- successive splits, or
- successive aggregations

Hierarchical Clustering

Provides graphical illustration of relationships between the data in the form of *dendrogram*

Dendrogram is a binary tree

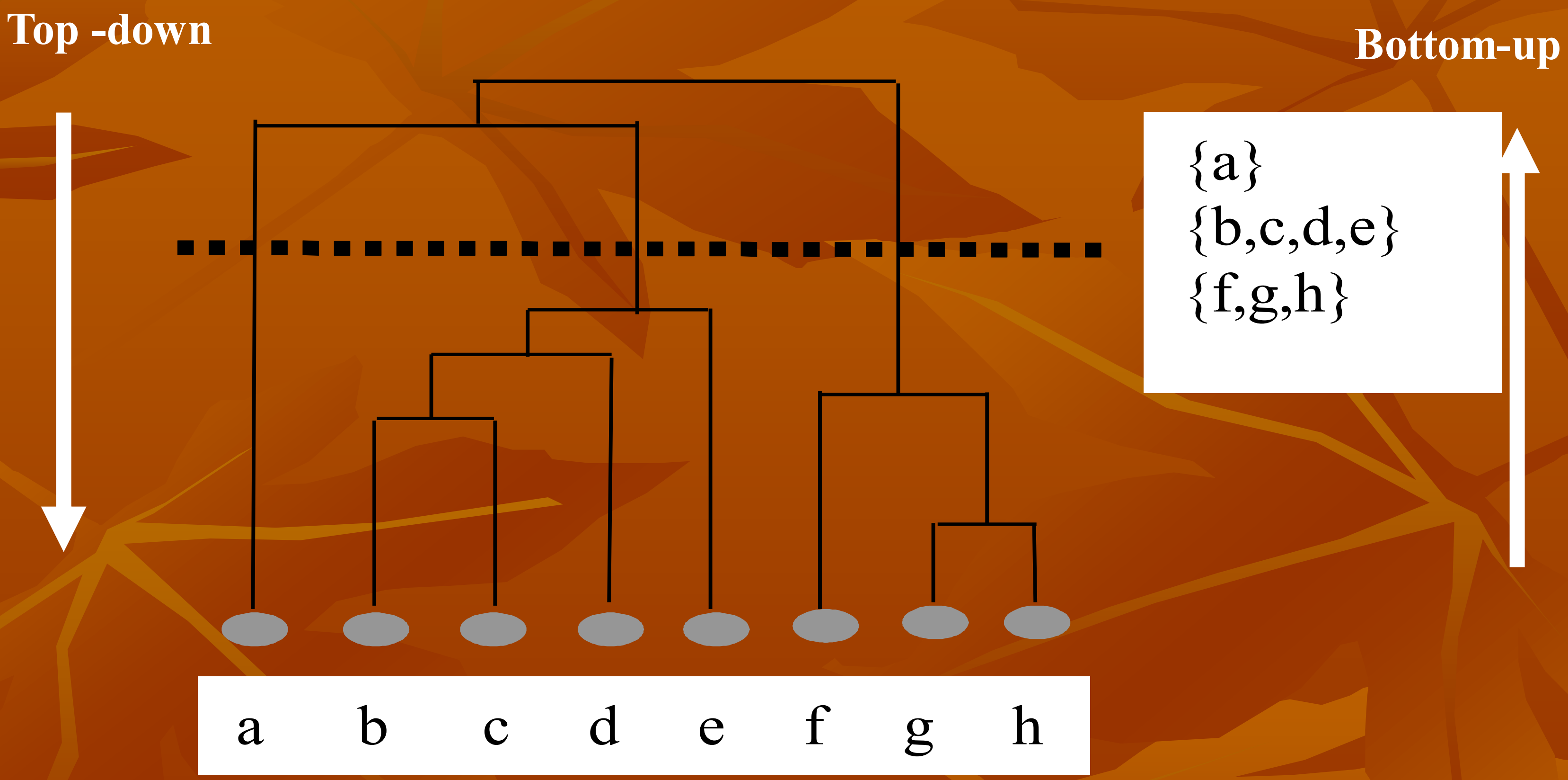
Two fundamental approaches:

- Bottom – up (agglomerative approach)
- Top-down (divisive approach)

Hierarchical Clustering: Types

- Agglomerative(**Bottom-up or agglomerative)**:
  - starts with as many clusters as there are records, with each cluster having only one record. Then pairs of clusters are successively merged until the number of clusters reduces to k.
  - at each stage, the pair of clusters are merged which are nearest to each other. If the merging is continued, it terminates in the hierarchy of clusters which is built with just a single cluster containing all the records.
- *Divisive algorithm (*__Top-down or divisive):__ takes the opposite approach from the agglomerative techniques. These starts with all the records in one cluster, and then try to split that clusters into smaller pieces.

Hierarchical Clustering



Top -down

Bottom-up

{a}
{b,c,d,e}
{f,g,h}

a b c d e f g h

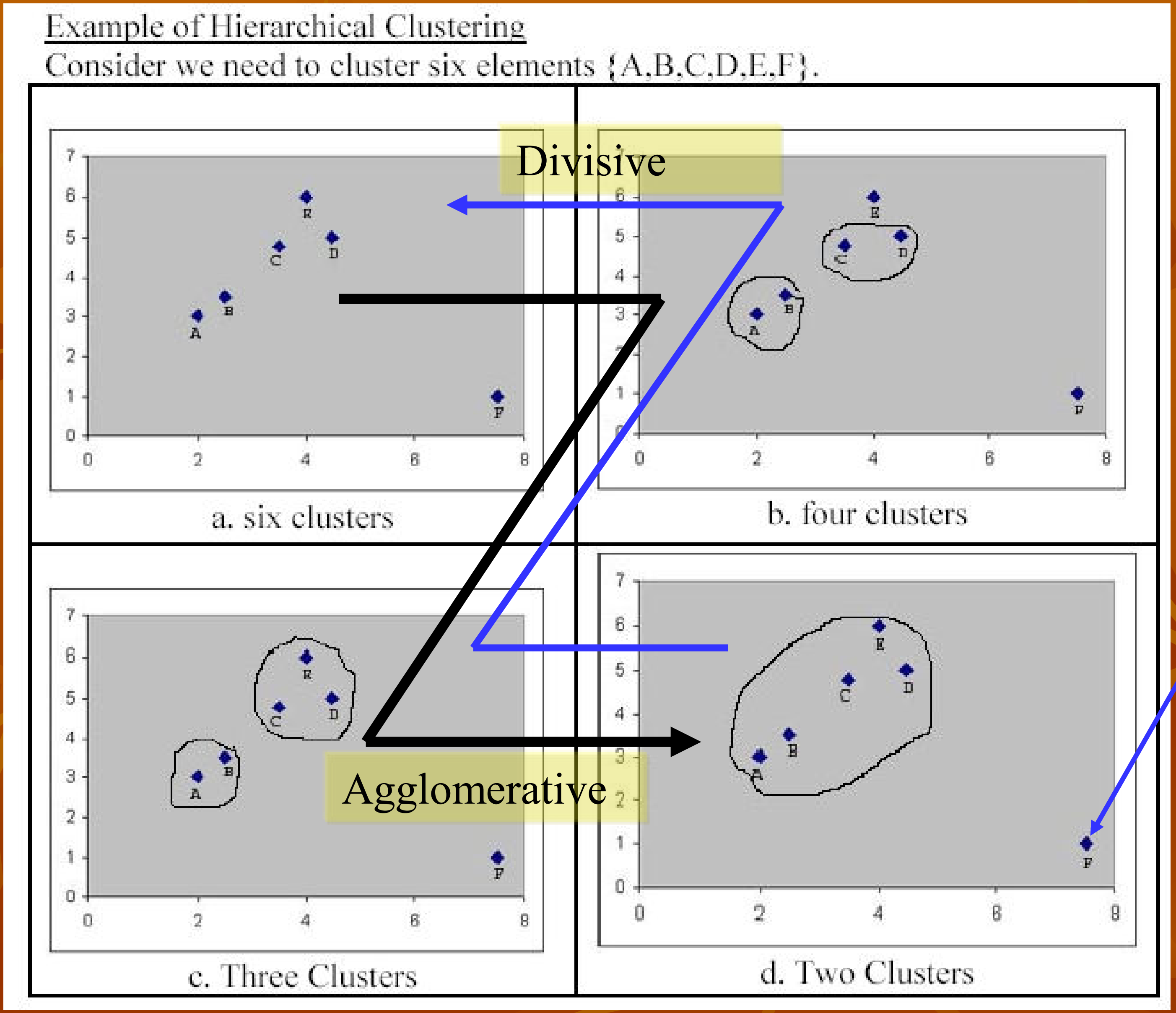
Hierarchical methods

- **Agglomerative methods start with each object in the data forming its own cluster, and then successively merge the clusters until one large cluster is formed (that encompasses the entire dataset)**

- **Divisive methods start by considering the entire data as one cluster and then split up the cluster(s) until each object forms one cluster**

Example of Hierarchical Clustering

Consider we need to cluster six elements {A,B,C,D,E,F}.



a. six clusters

b. four clusters

c. Three Clusters

d. Two Clusters

Remove Outlier

Density-Based Clustering Methods (DENCLUE)

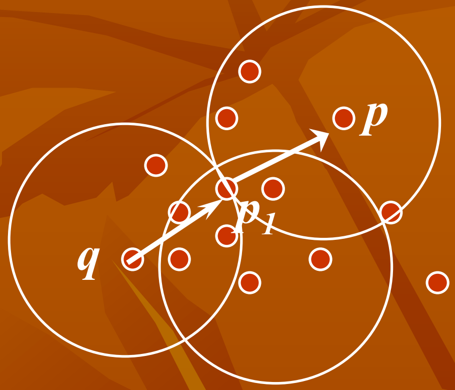
- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Density Based Spatial Clustering of Application Noise
  - OPTICS: Ordering Points to Identify the Clustering Structures
  - CLIQUE : Clustering in Clique

Density-Based Clustering: Background

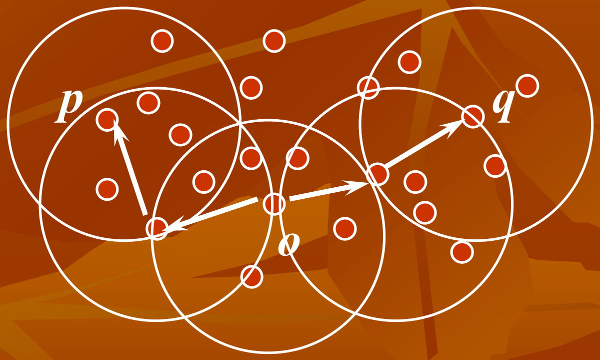
- The basic terms
  - The neighbourhood of an object that enclosed in a circle with radius Eps is called Eps - neighbourhood of that object
  - Eps neighbourhood with minimum object points is called core object.
  - An object A from a dataset is directly density reachable from object B where A is the member of Eps-neighbourhood of B and B is a core object.

Density-Based Clustering:

- Density-reachable:
  - A point p is density-reachable from a point q wrt. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

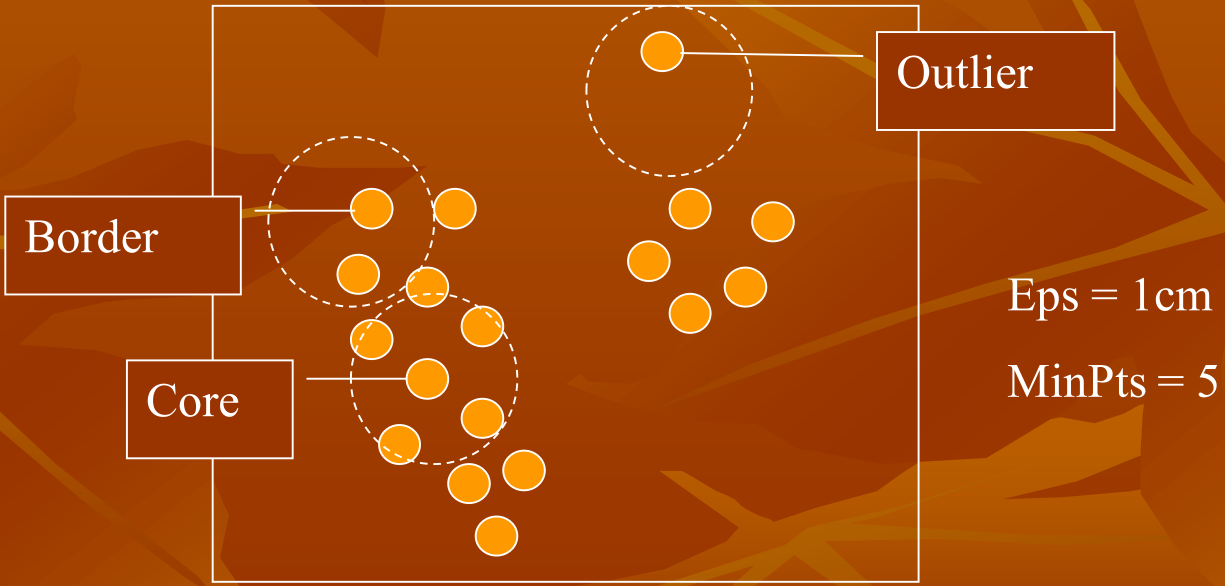


- Density-connected
  - A point p is density-connected to a point q wrt. *Eps*, *MinPts* if there is a point o such that both, p and q are density-reachable from o wrt. *Eps* and *MinPts*.



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



DBSCAN: The Algorithm

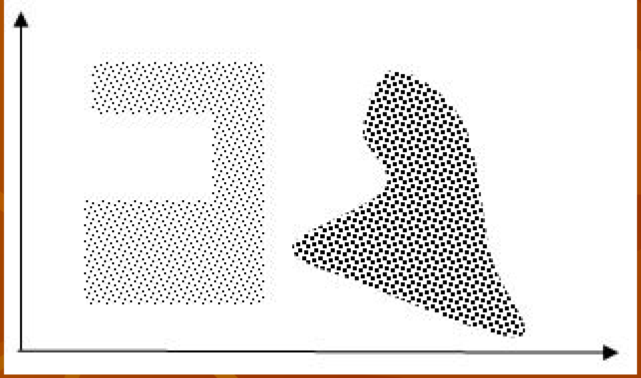
- Arbitrarilyy select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Grid-Based Clustering Method

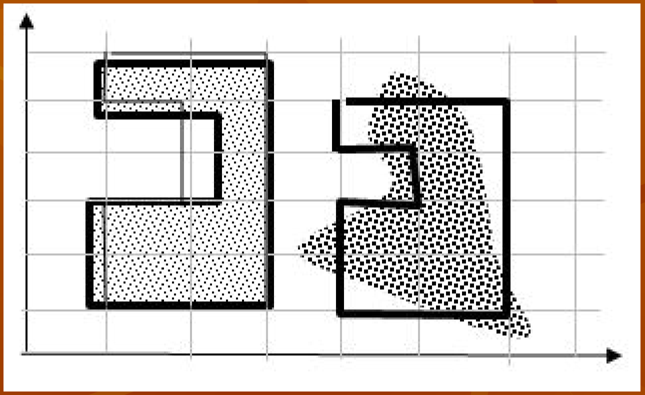
- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

Grid-Based Clustering

Describe structure in data in the language of generic geometric constructs – *hyperboxes* and their combinations



Collection of clusters of different geometry



Formation of clusters by merging adjacent hyperboxes of the grid

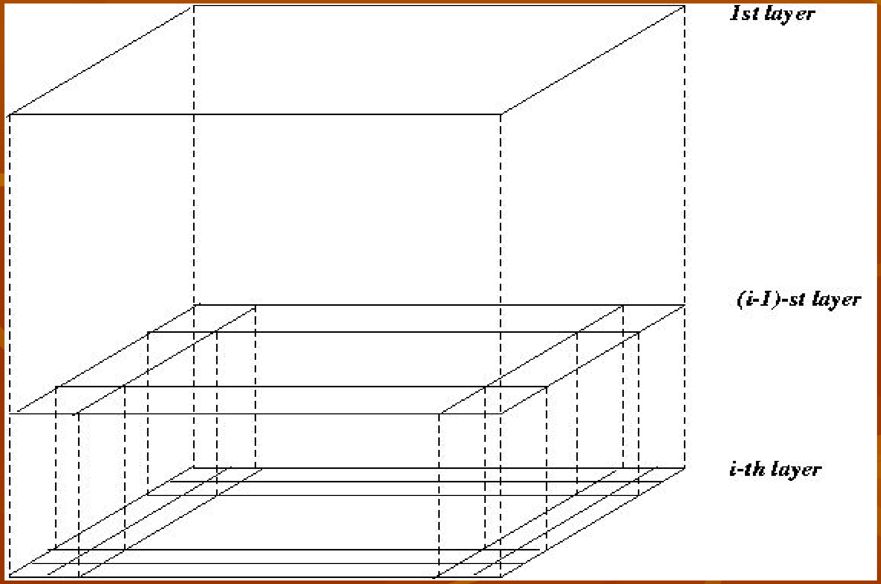
Grid-Based Clustering Steps

- Formation of the grid structure
- Insertion of data into the grid structure
- Computation of the density index of each hyperbox of the grid structure
- Sorting the hyperboxes with respect to the values of their density index
- Identification of cluster centers (viz. the hyperboxes of the highest density)
- Traversal of neighboring hyperboxes and merging process
- Choice of the grid:
  - too rough grid may not help capture the details of the structure in the data.
  - too detailed grid produces a significant computational overhead.

STING: A Statistical Information Grid

- The spatial area is divided into rectangular cells

- There are several levels of cells corresponding to different levels of resolution



STING: A Statistical Information Grid

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
  - *count*, *mean*, *s*, *min*, *max*
  - type of distribution—normal, *uniform*, etc.
- For each cell in the current level compute the confidence interval

STING: A Statistical Information Grid

- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached

- Advantages:
  - Query-independent, easy to parallelize, incremental update
  - *O(K),* where *K* is the number of grid cells at the lowest level

- Disadvantages:
  - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

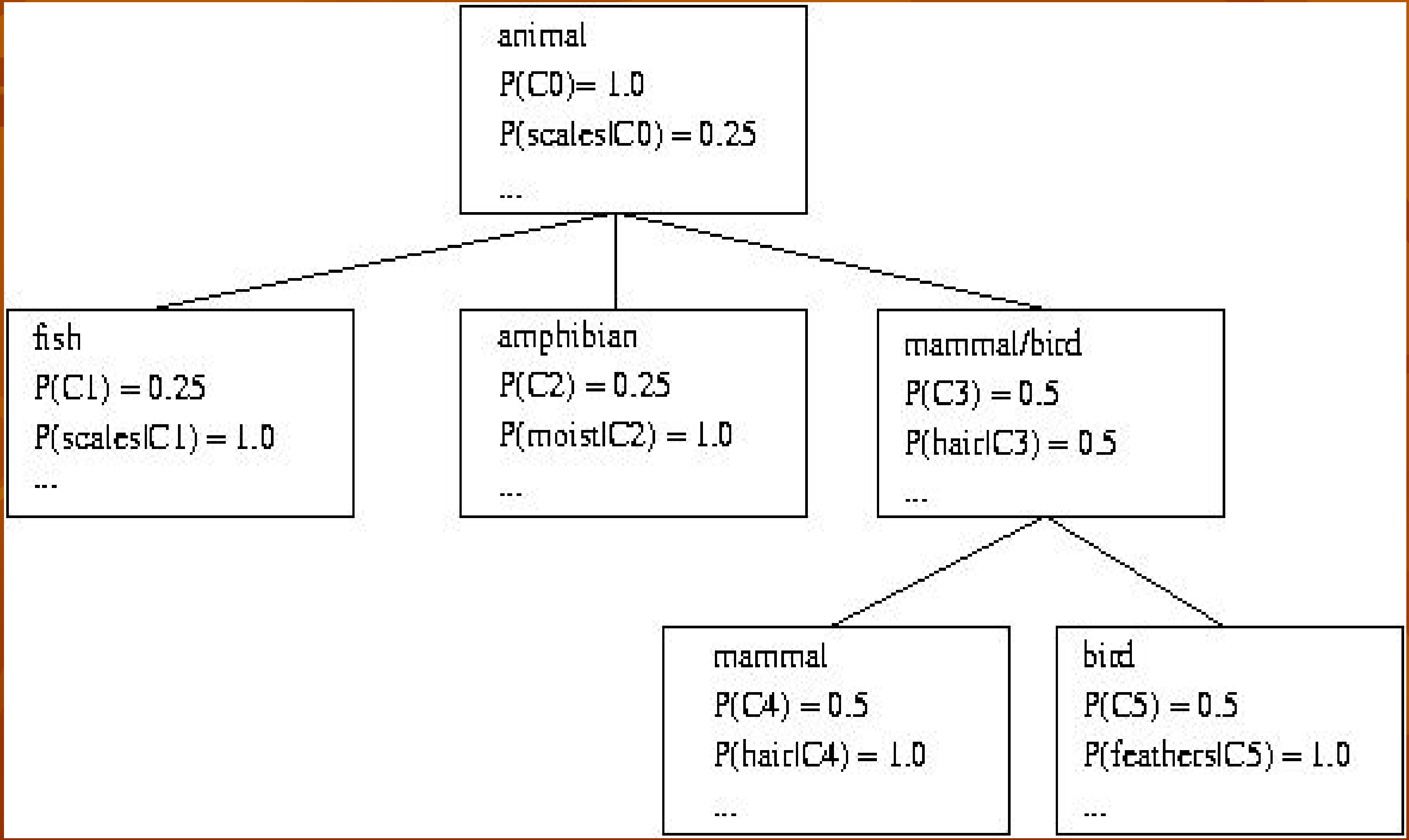
Model-Based Clustering Methods

- Attempt to optimize the fit between the data and some mathematical model

- Statistical and AI approach
  - COBWEB (Fisher'87)
    - A popular a simple method of incremental conceptual learning
    - Creates a hierarchical clustering in the form of a classification tree
    - Each node refers to a concept and contains a probabilistic description of that concept

COBWEB Clustering Method

A classification tree



Summary

- Cluster analysis groups objects based on their similarity and has wide applications

- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods

- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches

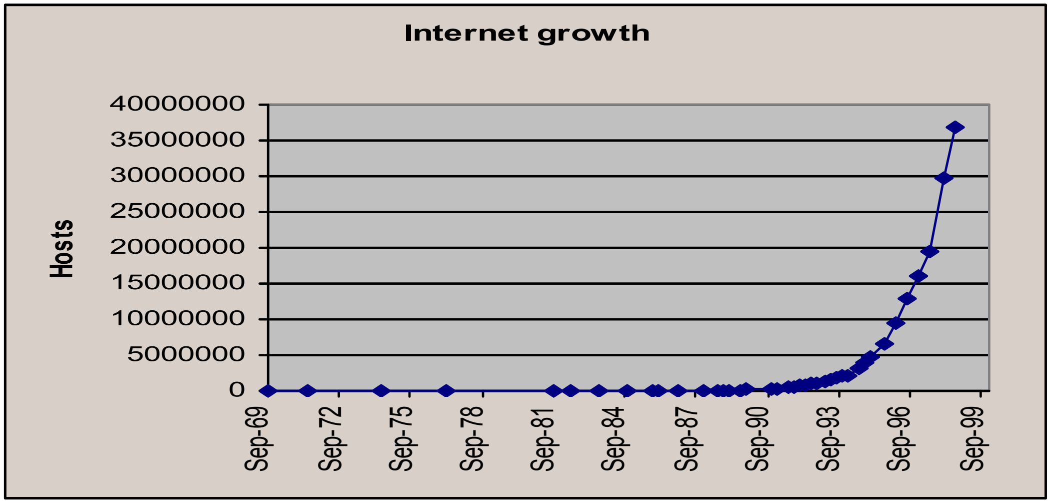
Mining the World-Wide Web

- The WWW is huge, widely distributed, global information service centre for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information

- WWW provides rich sources for data mining
- Challenges
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

Mining the World-Wide Web

- Growing and changing very rapidly



- Broad diversity of user communities
- Only a small portion of the information on the Web is truly relevant or useful
  - 99% of the Web information is useless to 99% of Web users
  - How can we find high-quality Web pages on a specified topic?

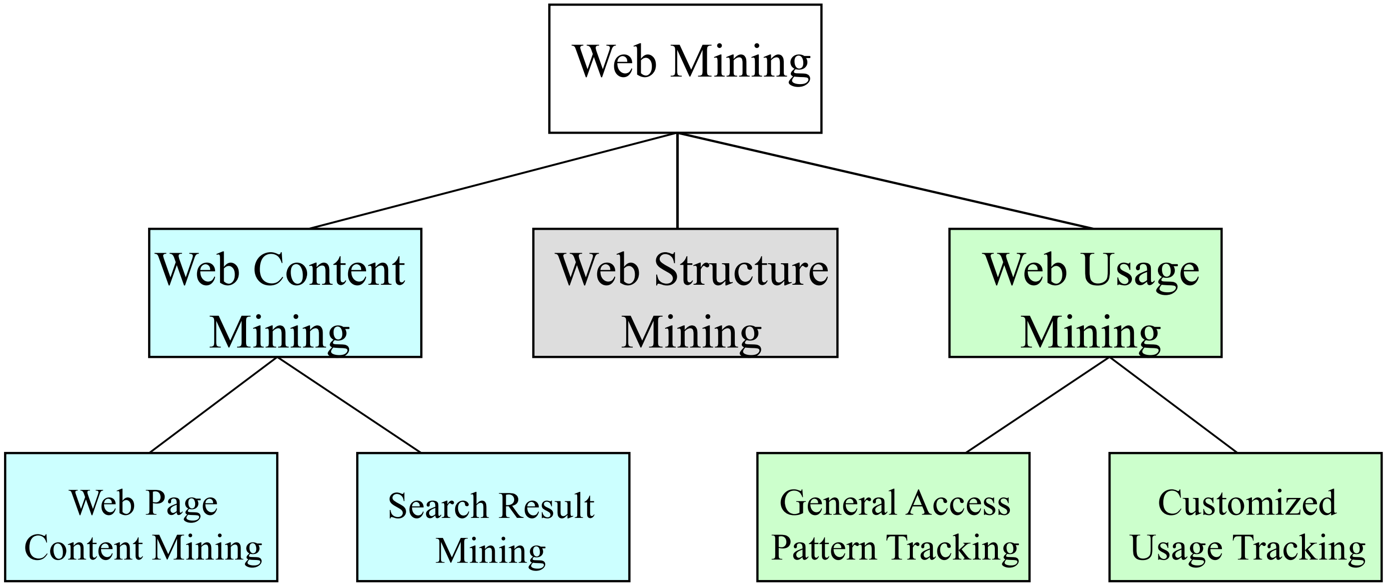
Web search engines

- Index-based: search the Web, index Web pages, and build and store huge keyword-based indices

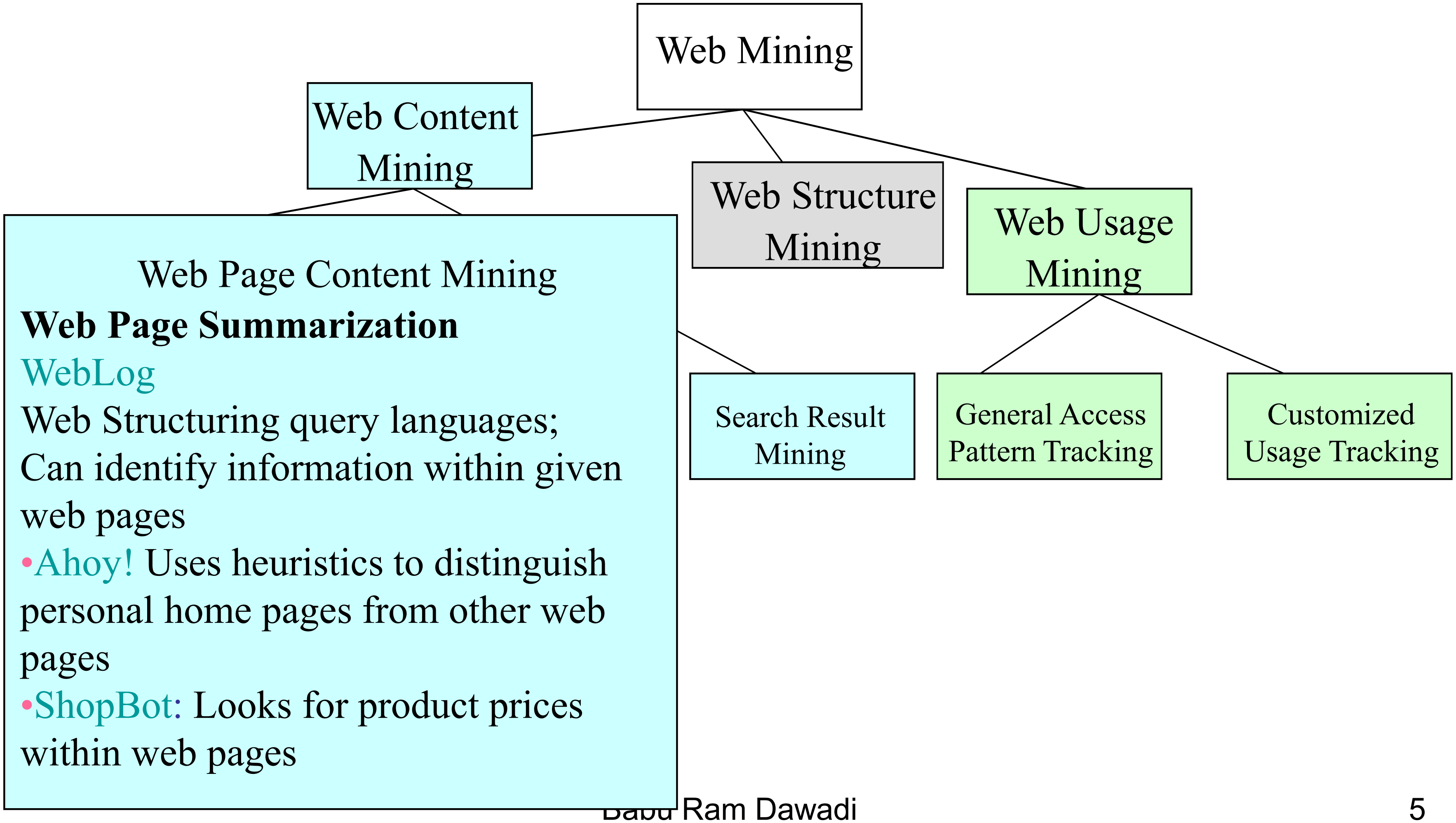
- Help locate sets of Web pages containing certain keywords

- Deficiencies
  - A topic of any breadth may easily contain hundreds of thousands of documents

Web Mining Taxonomy



Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Web Page Summarization

WebLog

Web Structuring query languages;

Can identify information within given web pages

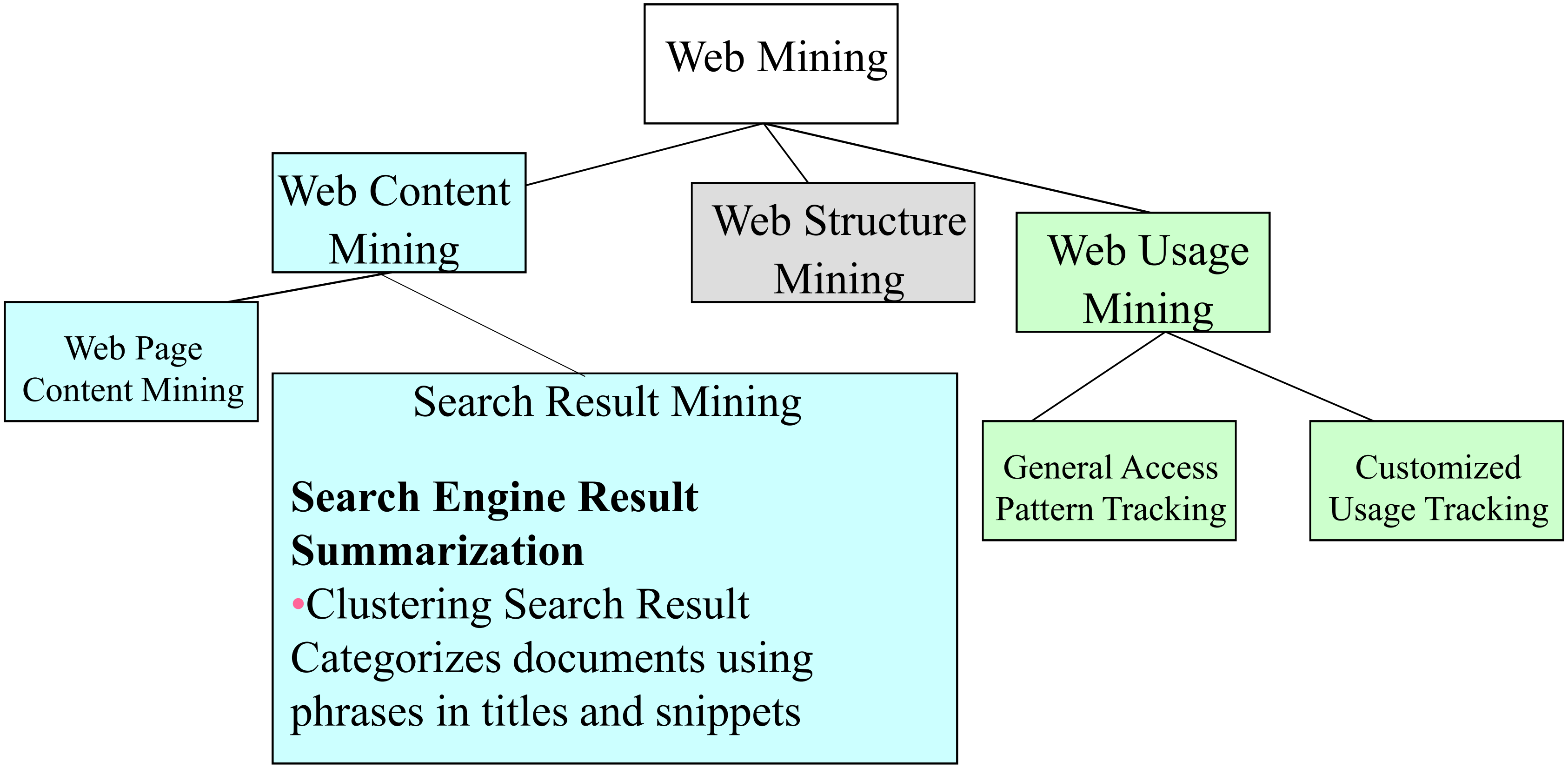
- Ahoy! Uses heuristics to distinguish personal home pages from other web pages
- ShopBot: Looks for product prices within web pages

Search Result Mining

General Access Pattern Tracking

Customized Usage Tracking

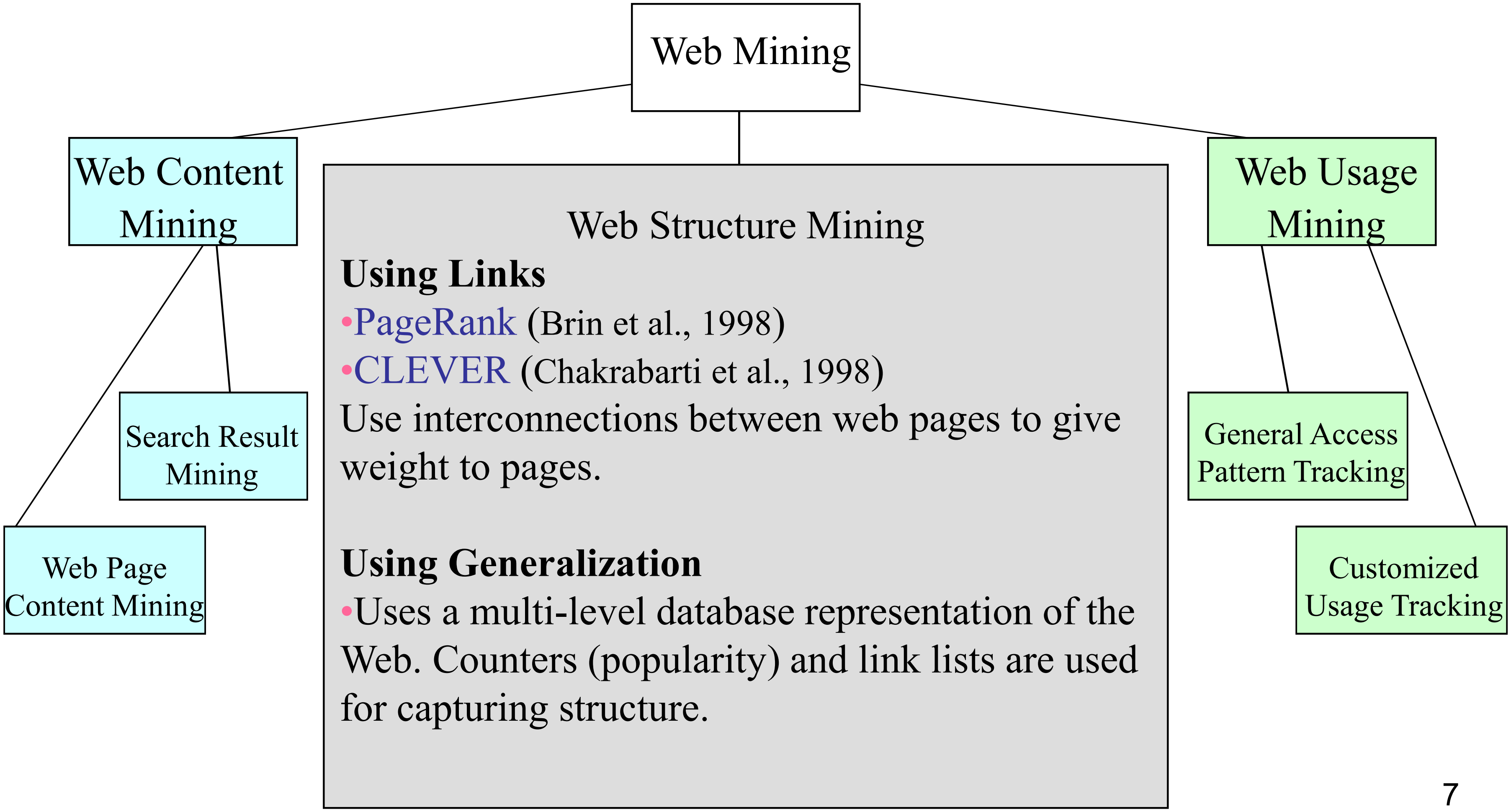
Mining the World-Wide Web



Search Engine Result Summarization

- Clustering Search Result Categorizes documents using phrases in titles and snippets

Mining the World-Wide Web



Web Mining

Web Content Mining

Search Result Mining

Web Page Content Mining

Web Structure Mining

Using Links

- PageRank (Brin et al., 1998)
- CLEVER (Chakrabarti et al., 1998)

Use interconnections between web pages to give weight to pages.

Using Generalization

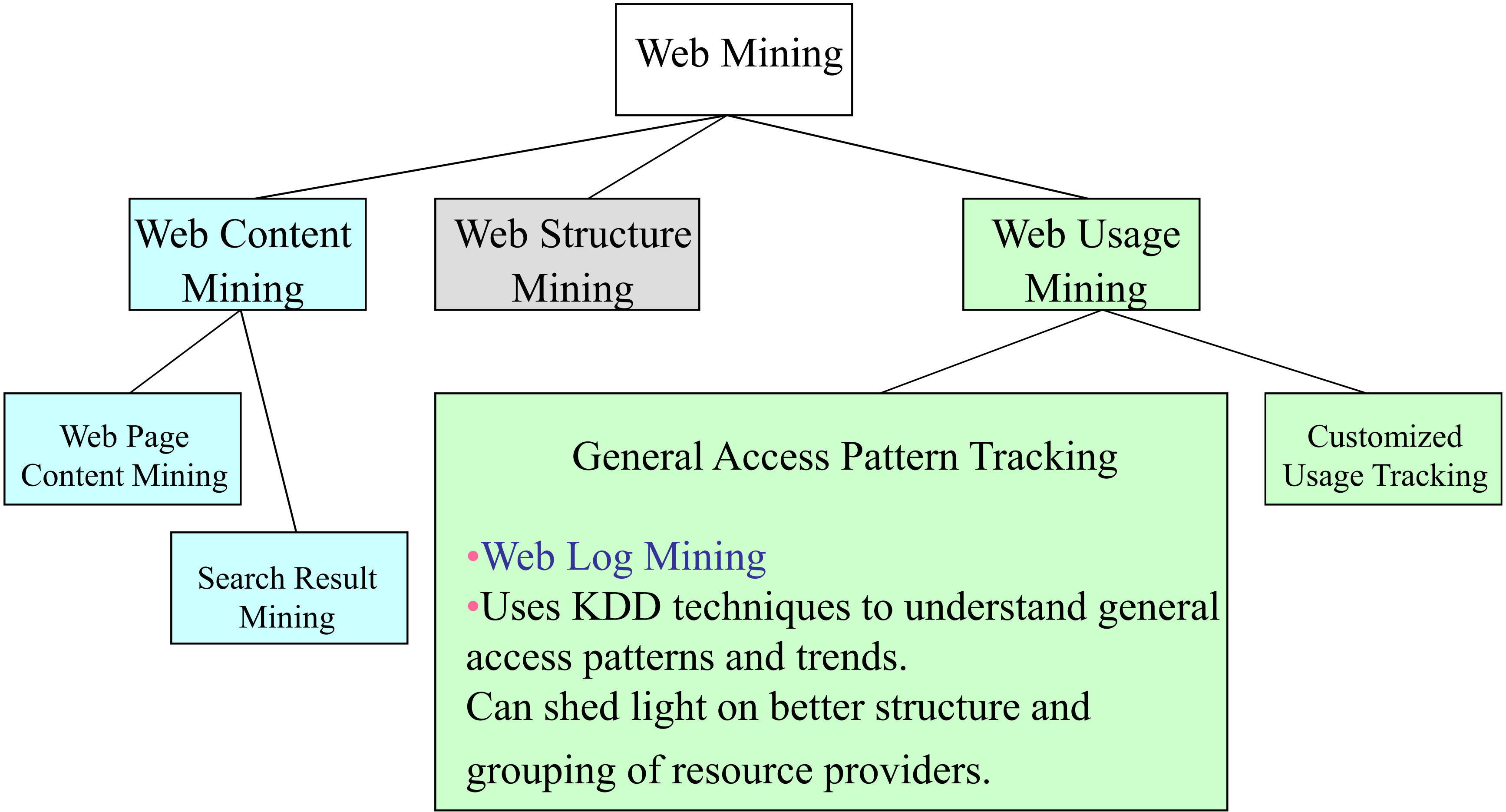
- Uses a multi-level database representation of the Web. Counters (popularity) and link lists are used for capturing structure.

Web Usage Mining

General Access Pattern Tracking

Customized Usage Tracking

Mining the World-Wide Web



Web Mining

Web Content Mining

Web Structure Mining

Web Usage Mining

Web Page Content Mining

Search Result Mining

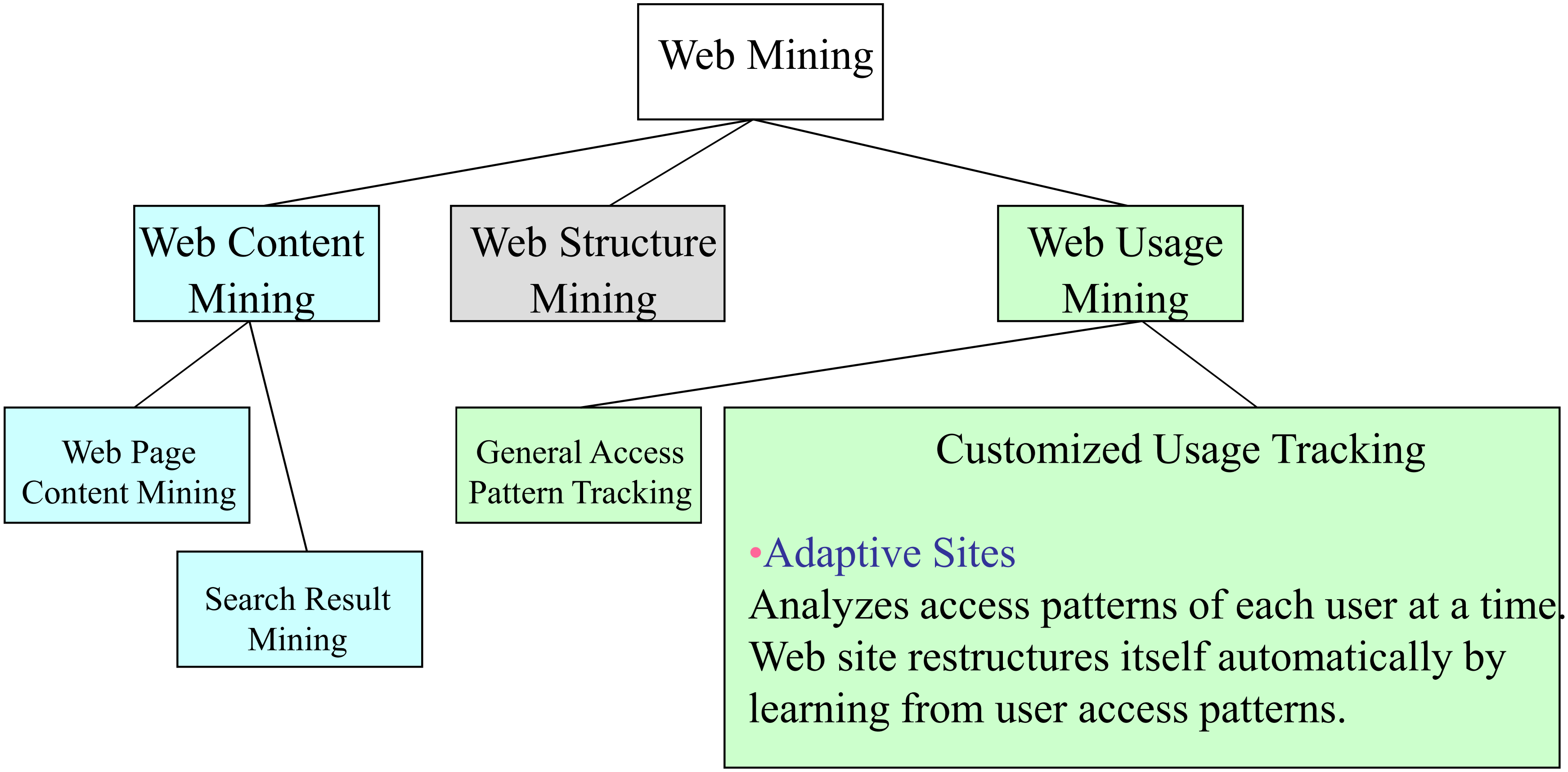
General Access Pattern Tracking

- Web Log Mining
- Uses KDD techniques to understand general access patterns and trends.

Can shed light on better structure and grouping of resource providers.

Customized Usage Tracking

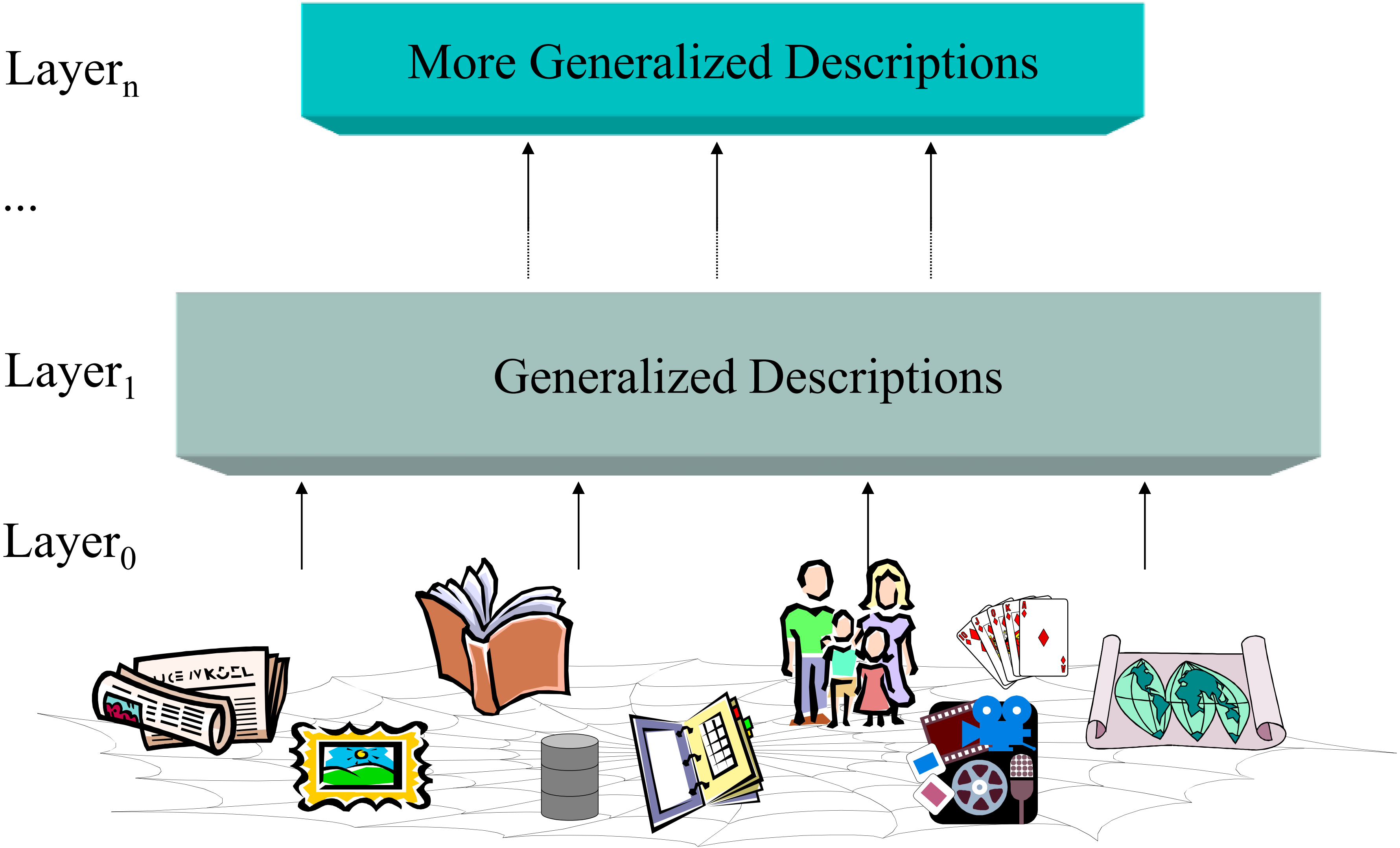
Mining the World-Wide Web



- Adaptive Sites

Analyzes access patterns of each user at a time.
Web site restructures itself automatically by learning from user access patterns.

Multiple Layered Web Architecture



Layer$_n$

More Generalized Descriptions

…

Layer$_1$

Generalized Descriptions

Layer$_0$

Mining the World-Wide Web

Layer-0: Primitive data

Layer-1: dozen database relations representing types of objects (metadata)

document, organization, person, software, game, map, image,…

- **document**(file_addr, authors, title, publication, publication_date, abstract, language, table_of_contents, category_description, keywords, index, multimedia_attached, num_pages, format, first_paragraphs, size_doc, timestamp, access_frequency, links_out,...)

- **person**(last_name, first_name, home_page_addr, position, picture_attached, phone, e-mail, office_address, education, research_interests, publications, size_of_home_page, timestamp, access_frequency, ...)

- **image**(image_addr, author, title, publication_date, category_description, keywords, size, width, height, duration, format, parent_pages, colour_histogram, Colour_layout, Texture_layout, Movement_vector, localisation_vector, timestamp, access_frequency, ...)

Mining the World-Wide Web

Layer-2: simplification of layer-1

- **doc_brief**(file_addr, authors, title, publication, publication_date, abstract, language, category_description, key_words, major_index, num_pages, format, size_doc, access_frequency, links_out)

- **person_brief** (last_name, first_name, publications,affiliation, e-mail, research_interests, size_home_page, access_frequency)

Layer-3: generalization of layer-2

- **cs_doc**(file_addr, authors, title, publication, publication_date, abstract, language, category_description, keywords, num_pages, form, size_doc, links_out)

- **doc_summary**(affiliation, field, publication_year, count, first_author_list, file_addr_list)

- **doc_author_brief**(file_addr, authors, affiliation, title, publication, pub_date, category_description, keywords, num_pages, format, size_doc, links_out)

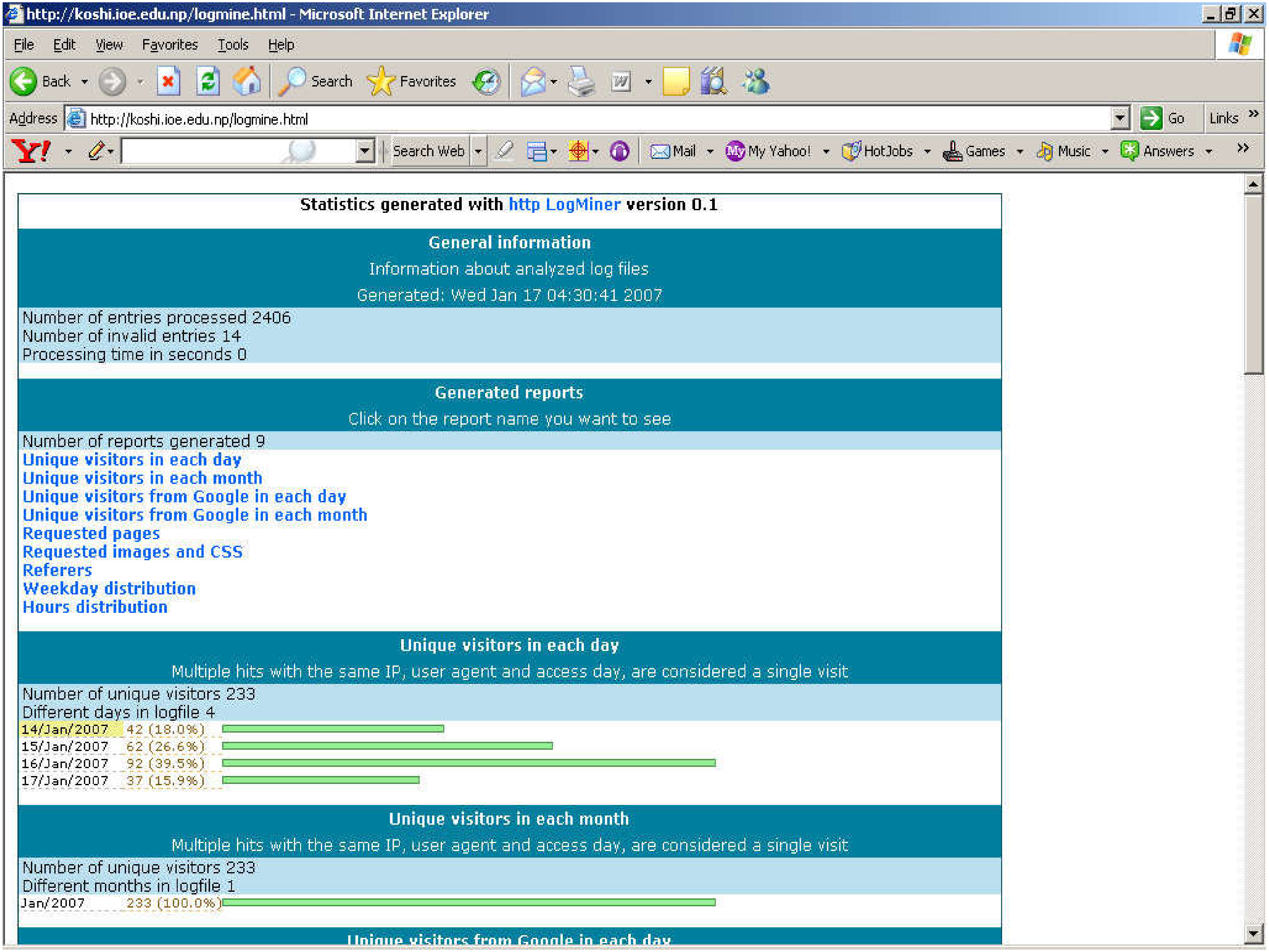
- **person_summary**(affiliation, research_interest, year, num_publications, count)

Web Usage Mining

- Mining Web log records to discover user access patterns of Web pages
- Applications
  - Target potential customers for electronic commerce
  - Enhance the quality and delivery of Internet information services to the end user
  - Improve Web server system performance
  - Identify potential prime advertisement locations
- Web logs provide rich information about Web dynamics
  - Typical Web log entry includes the URL requested, the IP address from which the request originated, and a timestamp

Techniques for Web usage mining

- Construct multidimensional view on the Weblog database
  - Perform multidimensional OLAP analysis to find the top *N* users, top *N* accessed Web pages, most frequently accessed time periods, etc.
- Perform data mining on Weblog records
  - Find association patterns, sequential patterns, and trends of Web accessing
  - May need additional information,e.g., user browsing sequences of the Web pages in the Web server buffer
- Conduct studies to
  - Analyze system performance, improve system design by Web caching, Web page prefetching, and Web page swapping



Statistics generated with http LogMiner version 0.1

General information

Information about analyzed log files

Generated: Wed Jan 17 04:30:41 2007

Number of entries processed 2406
Number of invalid entries 14
Processing time in seconds 0

Generated reports

Click on the report name you want to see

Number of reports generated 9

- Unique visitors in each day
- Unique visitors in each month
- Unique visitors from Google in each day
- Unique visitors from Google in each month
- Requested pages
- Requested images and CSS
- Referers
- Weekday distribution
- Hours distribution

Unique visitors in each day

Multiple hits with the same IP, user agent and access day, are considered a single visit

Number of unique visitors 233
Different days in logfile 4

```
14/Jan/2007   42 (18.0%)
15/Jan/2007   62 (26.6%)
16/Jan/2007   92 (39.5%)
17/Jan/2007   37 (15.9%)
```

Unique visitors in each month

Multiple hits with the same IP, user agent and access day, are considered a single visit

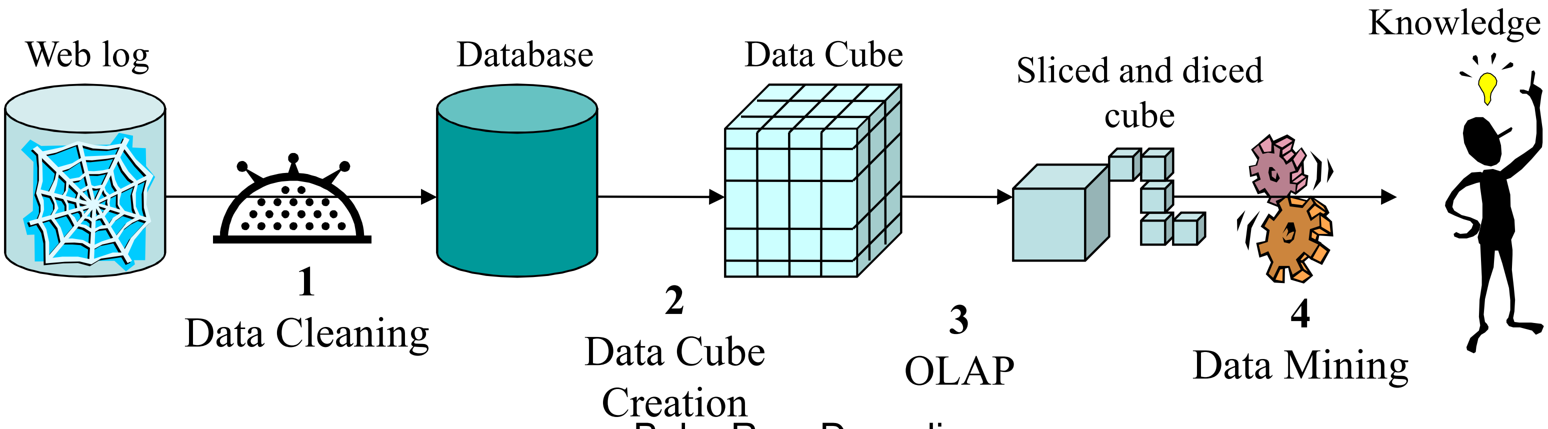
Number of unique visitors 233
Different months in logfile 1

```
Jan/2007      233 (100.0%)
```

Unique visitors from Google in each day

Mining the World-Wide Web

- Design of a Web Log Miner
  - Web log is filtered to generate a relational database
  - A data cube is generated form database
  - OLAP is used to drill-down and roll-up in the cube
  - OLAM is used for mining interesting knowledge



What is Customer Relationship Management?

- MORE than a technology
  - A philosophy .. a strategy … a discipline for handling information and designing business processes
- For CRM to be truly effective,
  - an organization must first understand who its customers are and what their value is over a lifetime.
  - The company must then determine what the needs of its customers are and how best to meet those needs.
  - Next, the organization must look into all of the different ways information about customers comes into a business, where and how this data is stored and how it is currently used.”
- CRM “brings together lots of pieces of information about customers, sales, marketing effectiveness, responsiveness and market trends” (“What is CRM”)

Relationship Marketing

- Relationship Marketing is a Process
  - communicating with your customers
  - listening to their responses
- Companies take actions
  - marketing campaigns
  - new products
  - new channels
  - new packaging

The Move Towards Relationship Management

- E-commerce companies want to customize the user experience
- Credit card companies want to recommend good restaurants and hotels in new cities
- Phone companies want to know your friends and family
- Bottom line: Companies want to be in the business of serving customers rather than merely selling products

CRM is Revolutionary

- Banks have been in the business of managing the spread between money borrowed and money lent
- Insurance companies have been in the business of managing loss ratios
- Telecoms have been in the business of completing telephone calls
- Key point: More companies are beginning to view customers as their primary asset

The Electronic Trail

- A customer places a catalog order over the telephone
- At the local telephone company
  - time of call, number dialed, long distance company used, …
- At the long distance company (for the toll-free number)
  - duration of call, route through switching system, …
- At the catalog
  - items ordered, call center, promotion response, credit card used, inventory update, shipping method requested, …

The Electronic Trail-- continued

- At the credit card clearing house
  - transaction date, amount charged, approval code, vendor number, …
- At the bank
  - billing record, interest rate, available credit update, …
- At the package carrier
  - zip code, time stamp at truck, time stamp, …
- Bottom line: Companies do keep track of data

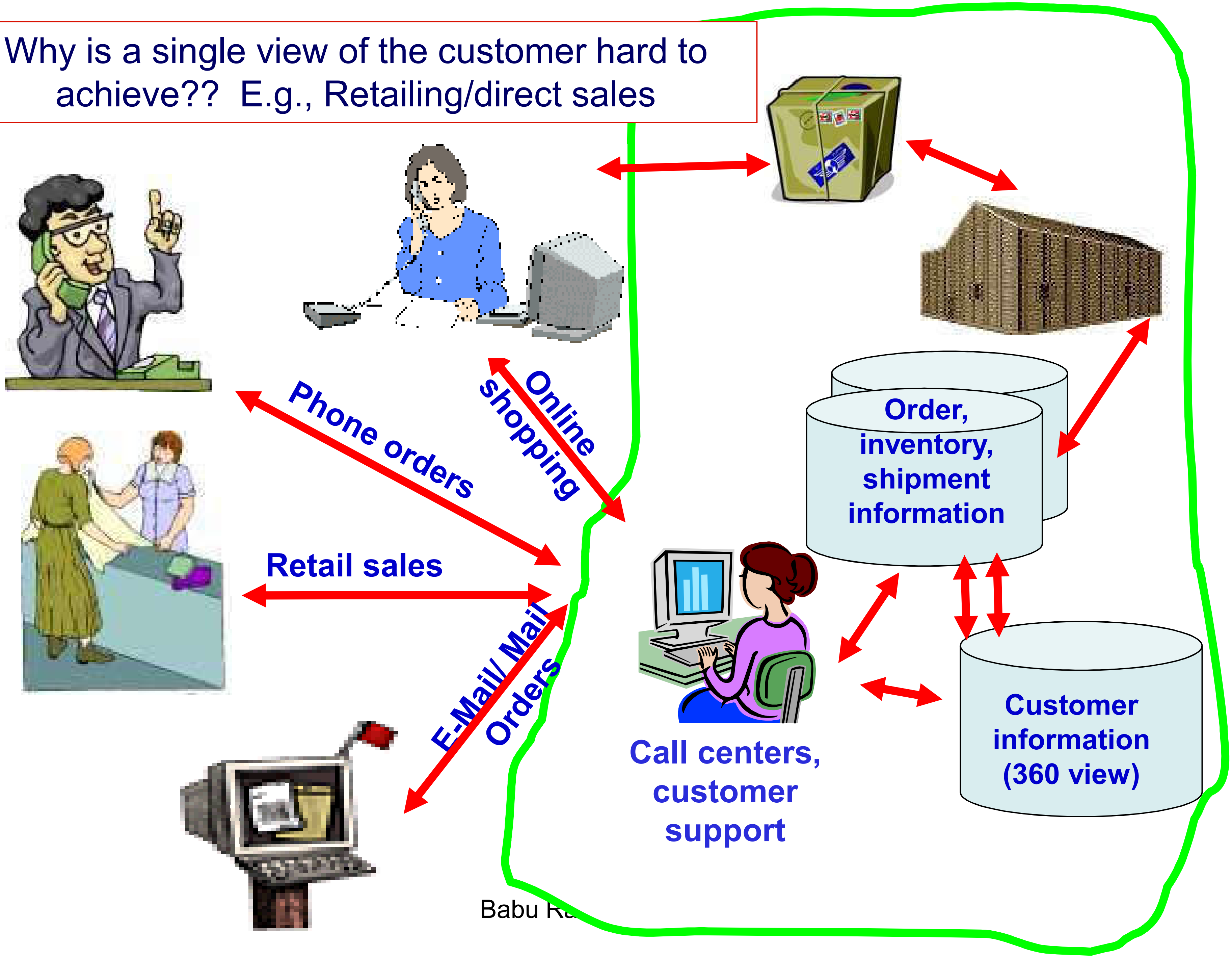
CRM Requires Learning and More

- Form a learning relationship with your customers
  - Notice their needs
    - On-line Transaction Processing Systems
  - Remember their preferences
    - Decision Support Data Warehouse
  - Learn how to serve them better
    - Data Mining
  - Act to make customers more profitable

The Importance of Channels

- Channels are the way a company interfaces with its customers
- Examples
  - Direct mail
  - Email
  - Banner ads
  - Telemarketing
  - Billing inserts
  - Customer service centers
  - Messages on receipts
- Key data about customers come from channels

Why is a single view of the customer hard to achieve?? E.g., Retailing/direct sales


Phone orders
Online shopping
Retail sales
E-Mail/ Mail Orders
Order, inventory, shipment information
Call centers, customer support
Customer information (360 view)

Goals of CRM

- Typical CRM goals:
  - Acquire customers (direct marketing, identifying)
  - Retain customers (customer service contacts)
  - Enhance revenue from customers (profiling, customization)
- Track/analyze customer needs; profiling
  - Data warehousing, analytic software
- Increasing revenues/profits by marketing -- ‘everyone is in sales’
- Monitor sales, better planning

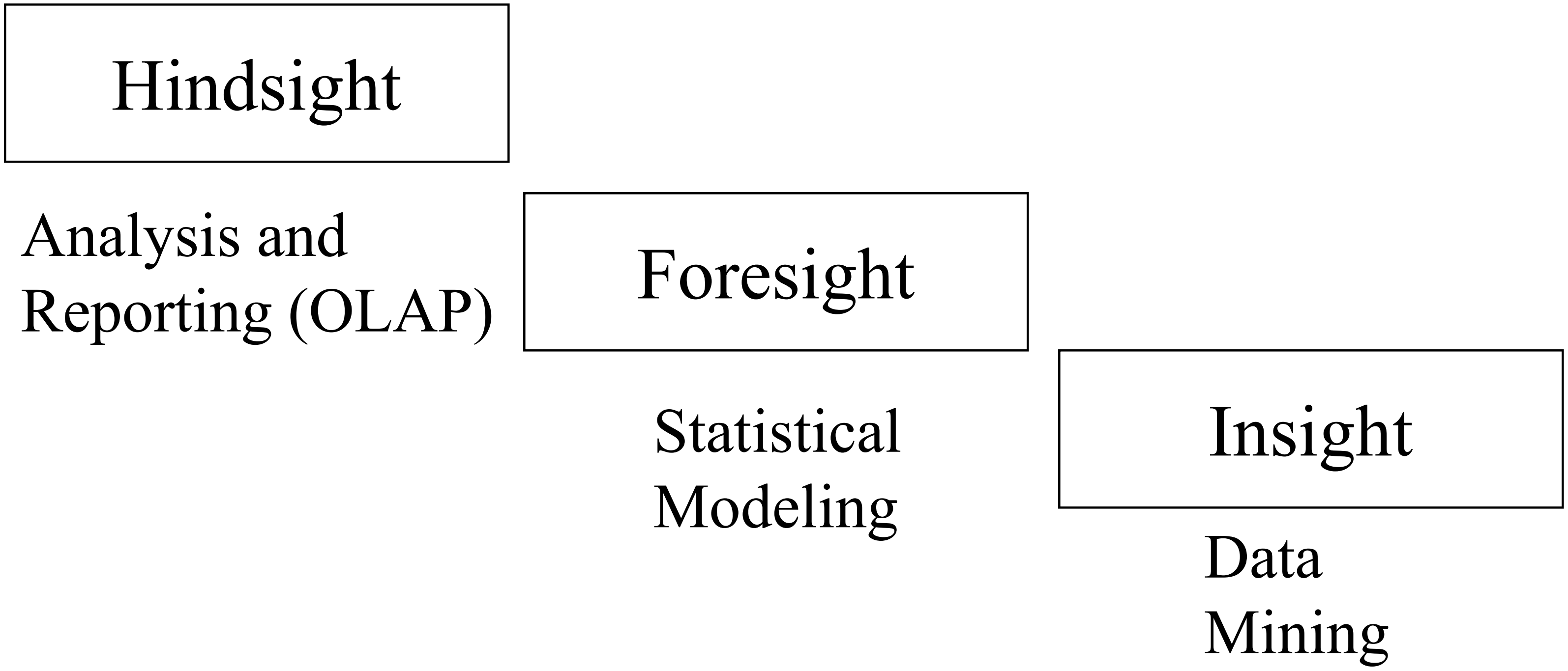
Technologies that underlie CRM applications

- Many technologies can contribute to CRM applications, depending on the area of focus
  - Most are basic ICTs that need to be integrated, e.g.,
- **Data collection**
- **Data storage/maintenance** (e.g., customer data bases, data warehouses or marts)
- **Data analysis/interpretation** (e.g., data mining and analytic software)
- **Data presentation** (e.g., web-based portals, mobile devices)
- **Workflow/process** automation software packages
- Newer technologies relate to web-based and mobile communication channels

Channels

- Channels are the source of data
- Channels are the interface to customers
- Channels enable a company to get a particular message to a particular customer
- Channel management is a challenge in organizations
- CRM is about serving customers through all channels

Where Does Data Mining Fit In?



Hindsight

Analysis and Reporting (OLAP)

Foresight

Statistical Modeling

Insight

Data Mining

What is a Customer

- A transaction?
- An account?
- An individual?
- A household?
- The customer as a transaction
  - purchases made with cash are anonymous
  - most Web surfing is anonymous
  - we, therefore, know little about the consumer

A Customer is an Account

- More often, a customer is an account
- Retail banking
  - checking account, mortgage, auto loan, …
- Telecommunications
  - long distance, local, ISP, mobile, …
- Insurance
  - auto policy, homeowners, life insurance, …

Customers Play Different Roles

- Parents buy back-to-school clothes for teenage children
  - children decide what to purchase
  - parents pay for the clothes
  - parents “own” the transaction
- Parents give college-age children cellular phones or credit cards
  - parents may make the purchase decision
  - children use the product
- It is not always easy to identify the customer

The Customer’s Lifecycle

- Childhood
  - birth, school, graduation, …
- Young Adulthood
  - choose career, move away from parents, …
- Family Life
  - marriage, buy house, children, divorce, …
- Retirement
  - sell home, travel, hobbies, …
- Much marketing effort is directed at each stage of life

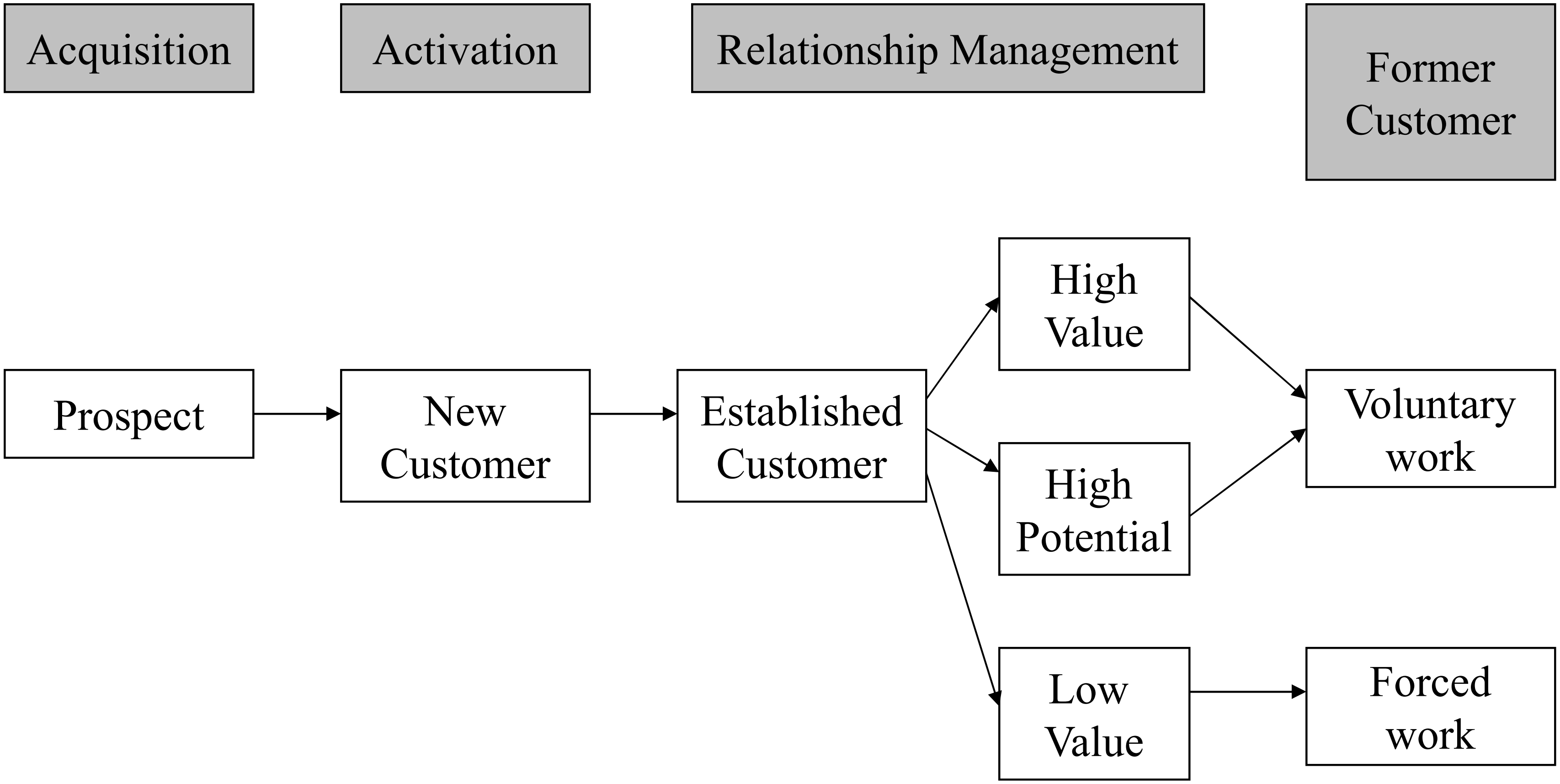
The Customer’s Lifecycle is Unpredictable

- It is difficult to identify the appropriate events
  - graduation, retirement may be easy
  - marriage, parenthood are not so easy
  - many events are “one-time”
- Companies miss or lose track of valuable information
  - a woman gets married, changes her last name, and merges her accounts with spouse
- It is hard to track your customers so closely, but, to the extent that you can, many marketing opportunities arise

Customers Evolve Over Time

- Customers begin as prospects
- Prospects indicate interest
  - fill out credit card applications
  - apply for insurance
  - visit your website
- They become new customers
- After repeated purchases or usage, they become established customers
- Eventually, they become former customers
  - either voluntarily or involuntarily

Business Processes Organize Around the Customer Lifecycle



Different Events Occur Throughout the Lifecycle

- Prospects receive marketing messages
- When they respond, they become new customers
- They make initial purchases
- They become established customers and are targeted by cross-sell and up-sell campaigns
- Some customers are forced to leave (cancel)
- Some leave (cancel) voluntarily
- Others simply stop using the product

Privacy is a Serious Matter

- Data mining and CRM raise some privacy concerns
- These concerns relate to the collection of data, more than the analysis of data
- The next few slides illustrate marketing mistakes that can result from the abundance and availability of data

The Story of Former Senior Chief Petty Officer Timothy McVeigh

- Several years ago, Timothy used an AOL account, with an anonymous alias
- Under marital status, he listed “gay”
- A colleague discovered the account and called AOL to verify that the owner was Timothy
- AOL gave out the information over the phone
- Timothy was discharged (three years short of his pension)

Serious Privacy Violations

- AOL breached its own policy by giving out confidential user information
  - AOL paid an undisclosed sum to settle with Timothy and suffered bad press as well
  - Timothy received an honorable discharge with full retirement pension

Friends, Family, and Others

- In the 1990s, MCI promoted the “Friends and Family” program
- They asked existing customers for names of people they talked with often
- If these friends and family signed up with MCI, then calls to them would be discounted
- Early in 1999, BT (formerly British Telecom) took the idea one step beyond
- BT invented a new marketing program
  - discounts to the most frequently called numbers

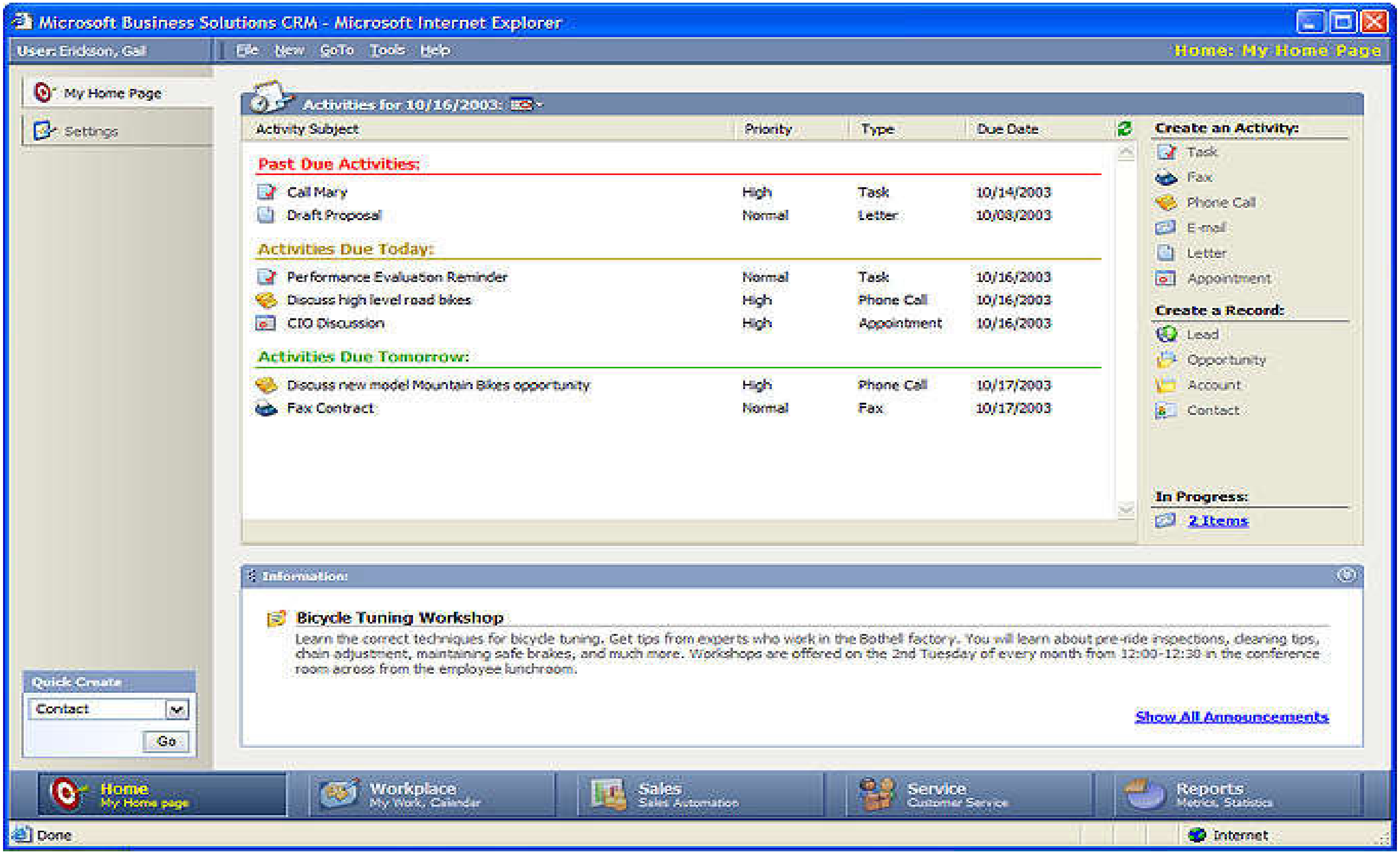
BT Marketing Program

- BT notified prospective customers of this program by sending them their most frequently called numbers
- One woman received the letter
  - uncovered her husband’s cheating
  - threw him out of the house
  - sued for divorce
- The husband threatened to sue BT for violating his privacy
- BT suffered negative publicity

No Substitute for Human Intelligence

- Data mining is a tool to achieve goals
- The goal is better service to customers
- Only people know what to predict
- Only people can make sense of rules
- Only people can make sense of visualizations
- Only people know what is reasonable, legal, tasteful
- Human decision makers are critical to the data mining process

Microsoft CRM Dynamics http://www.microsoft.com/BusinessSolutions/content /demos/MSCRMdemos/full_demo.htm



Activities for 10/16/2003:

Activity Subject	Priority	Type	Due Date
Past Due Activities:			
Call Mary	High	Task	10/14/2003
Draft Proposal	Normal	Letter	10/08/2003
Activities Due Today:			
Performance Evaluation Reminder	Normal	Task	10/16/2003
Discuss high level road bikes	High	Phone Call	10/16/2003
CIO Discussion	High	Appointment	10/16/2003
Activities Due Tomorrow:			
Discuss new model Mountain Bikes opportunity	High	Phone Call	10/17/2003
Fax Contract	Normal	Fax	10/17/2003

Information:

Bicycle Tuning Workshop

Learn the correct techniques for bicycle tuning. Get tips from experts who work in the Bothell factory. You will learn about pre-ride inspections, cleaning tips, chain adjustment, maintaining safe brakes, and much more. Workshops are offered on the 2nd Tuesday of every month from 12:00-12:30 in the conference room across from the employee lunchroom.

Show All Announcements

Privacy & Security Aspects

- Privacy-Preserving Data Mining

  How do we mine data when we can't even look at it?

- Individual Privacy
  - Nobody should know more about any entity after the data mining than they did before
- Organization Privacy
  - Protect knowledge about a collection of entities
    - Individual entity values may be known to all parties
    - Which entities are at which site may be secret

Privacy constraints don’t prevent data mining

- Goal of data mining is summary results
  - Association rules
  - Classifiers
  - Clusters
- The results alone need not violate privacy
  - Contain no individually identifiable values
  - Reflect overall results, not individual organizations

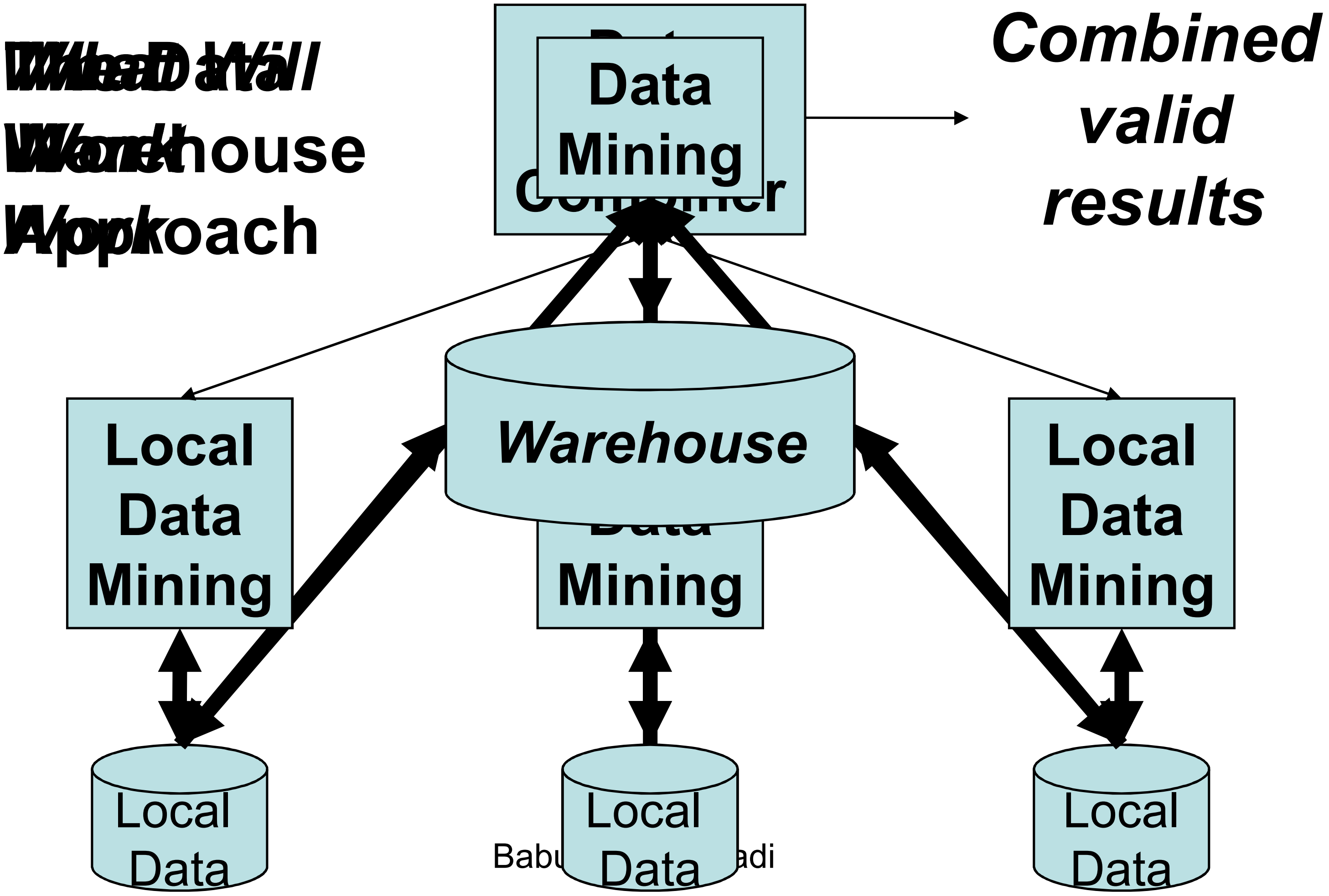
The problem is computing the results without access to the data!

Example: Privacy of Distributed Data

Data Warehouse Approach

What Won't Work

What Will Work



Privacy-Preserving Data Mining: Who?

- Government / public agencies. Example:
  - The Centers for Disease Control want to identify disease outbreaks
  - Insurance companies have data on disease incidents, seriousness, patient background, etc.
  - But can/should they release this information?
- Industry Collaborations / Trade Groups. Example:
  - An industry trade group may want to identify best practices to help members
  - But some practices are trade secrets
  - How do we provide “commodity” results to all (Manufacturing using chemical supplies from supplier X have high failure rates), while still preserving secrets (manufacturing process Y gives low failure rates)?

Privacy-Preserving Data Mining: Who?

- Multinational Corporations
  - A company would like to mine its data for globally valid results
  - But national laws may prevent transborder data sharing
- Public use of private data
  - Data mining enables research studies of large populations
  - But these populations are reluctant to release personal information

Data Mining as a Threat to Security

- Data mining gives us “facts” that are not obvious to human analysts of the data
- Enables inspection and analysis of huge amounts of data
- Possible threats:
  - Predict information about classified work from correlation with unclassified work (e.g. budgets, staffing)
  - Detect “hidden” information based on “conspicuous” lack of information
  - Mining “Open Source” data to determine predictive events (e.g., Pizza deliveries to the Pentagon)
- It isn’t the data we want to protect, but correlations among data items

Multimedia Data Mining

Multimedia Data Mining

- Multimedia data types
  - any type of information medium that can be represented, processed, stored and transmitted over network in digital form
  - Multi-lingual text, numeric, images, video, audio, graphical, temporal, relational, and categorical data.

Definitions

- Subfield of data mining that deals with an extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia databases

  - Influence on related interdisciplinary fields

  - Databases – extension of the KDD (rule patterns)

  - Information systems – multimedia information analysis and retrieval – content-based image and video search and efficient storage organization

Information model

- Data segmentation
  - Multimedia data are divided into logically interconnected segments (objects)
  - Pattern extraction
  - Mining and analysis procedures should reveal some relations between objects on the different level
  - Knowledge representation
  - Incorporated linked patterns

Generalizing Spatial and Multimedia Data

- Spatial data:
  - Generalize detailed geographic points into clustered regions, such as business, residential, industrial, or agricultural areas, according to land usage
  - Require the merge of a set of geographic areas by spatial operations
- Image data:
  - Extracted by aggregation and/or approximation
  - Size, color, shape, texture, orientation, and relative positions and structures of the contained objects or regions in the image
- Music data:
  - Summarize its melody: based on the approximate patterns that repeatedly occur in the segment
  - Summarized its style: based on its tone, tempo, or the major musical instruments played

Similarity Search in Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
  - Labor-intensive if performed manually
  - Results are typically of poor quality if automated

- Content-based retrieval systems
  - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

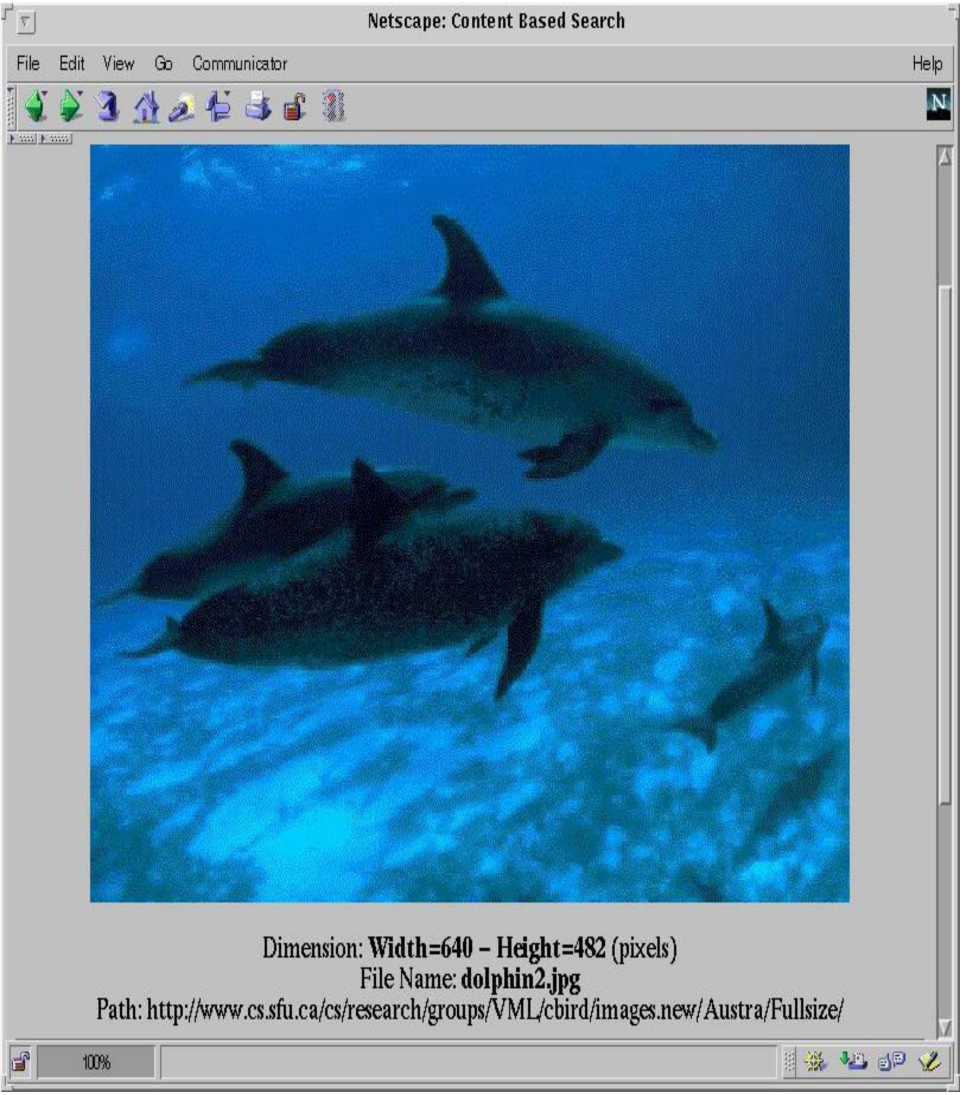
Multidimensional Analysis of Multimedia Data

- Multimedia data cube
  - Design and construct similar to that of traditional data cubes from relational data
  - Contain additional dimensions and measures for multimedia information, such as color, texture, and shape

- The database does not store images but their descriptors
  - Feature descriptor: a set of vectors for each visual characteristic
    - Color vector: contains the color histogram
    - MFC (Most Frequent Color) vector: five color centroids
    - MFO (Most Frequent Orientation) vector: five edge orientation centroids

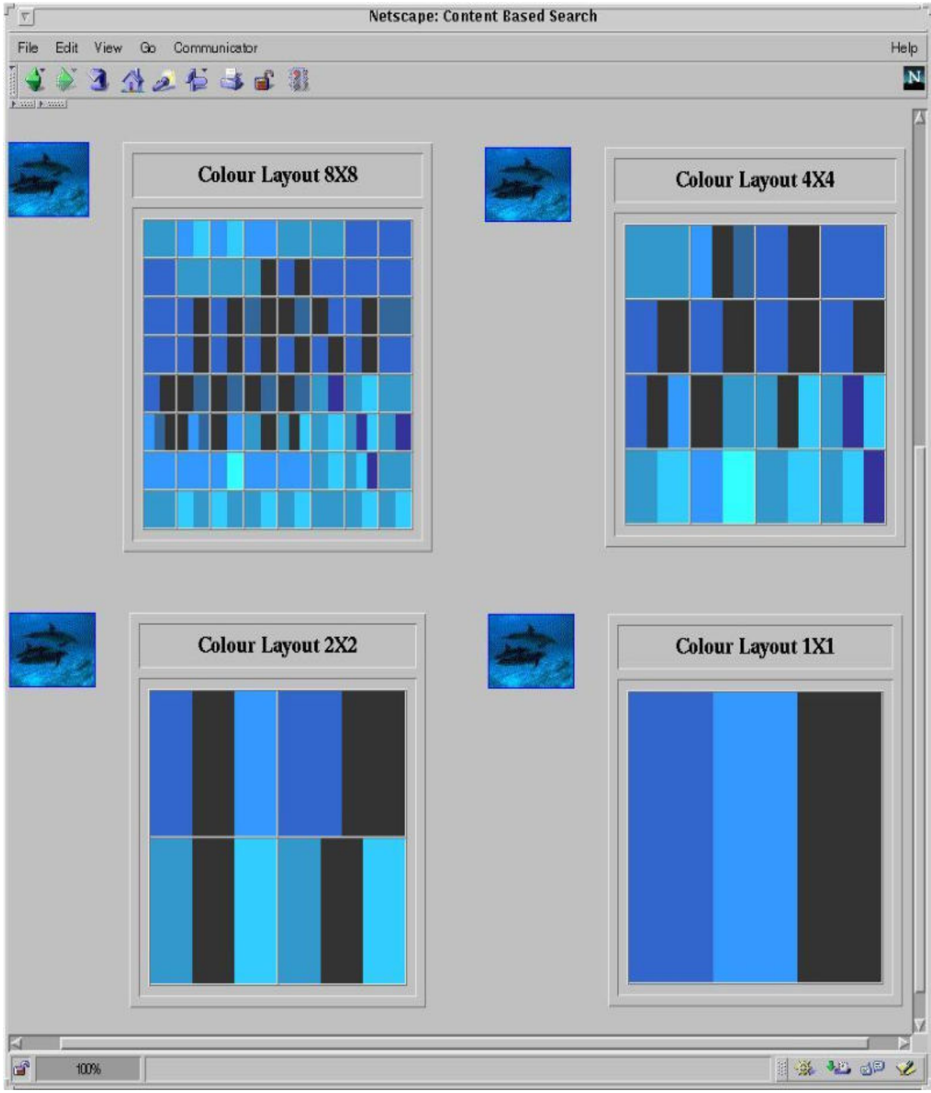
  - Layout descriptor: contains a color layout vector and an edge layout vector

Multi-Dimensional Search in Multimedia Databases



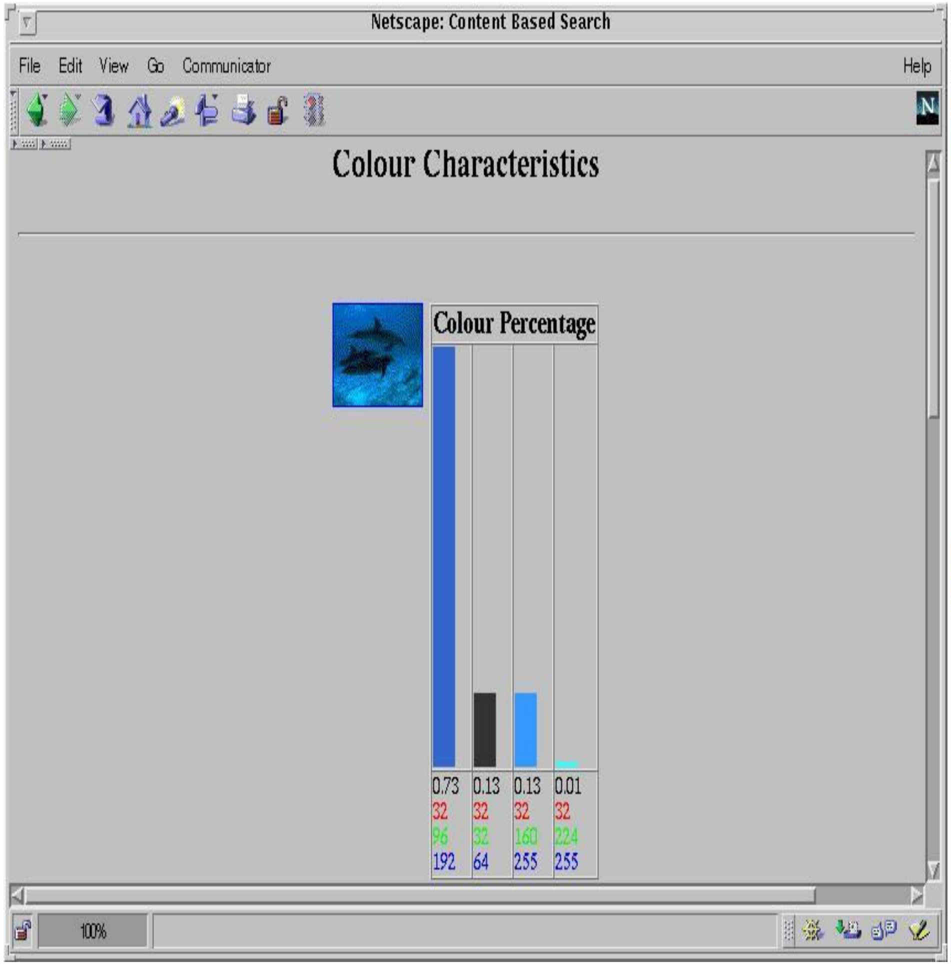
Netscape: Content Based Search
File Edit View Go Communicator Help
100%

Dimension: **Width=640 – Height=482** (pixels)
File Name: **dolphin2.jpg**
Path: http://www.cs.sfu.ca/cs/research/groups/VML/cbird/images.new/Austra/Fullsize/



Multi-Dimensional Analysis in Multimedia Databases

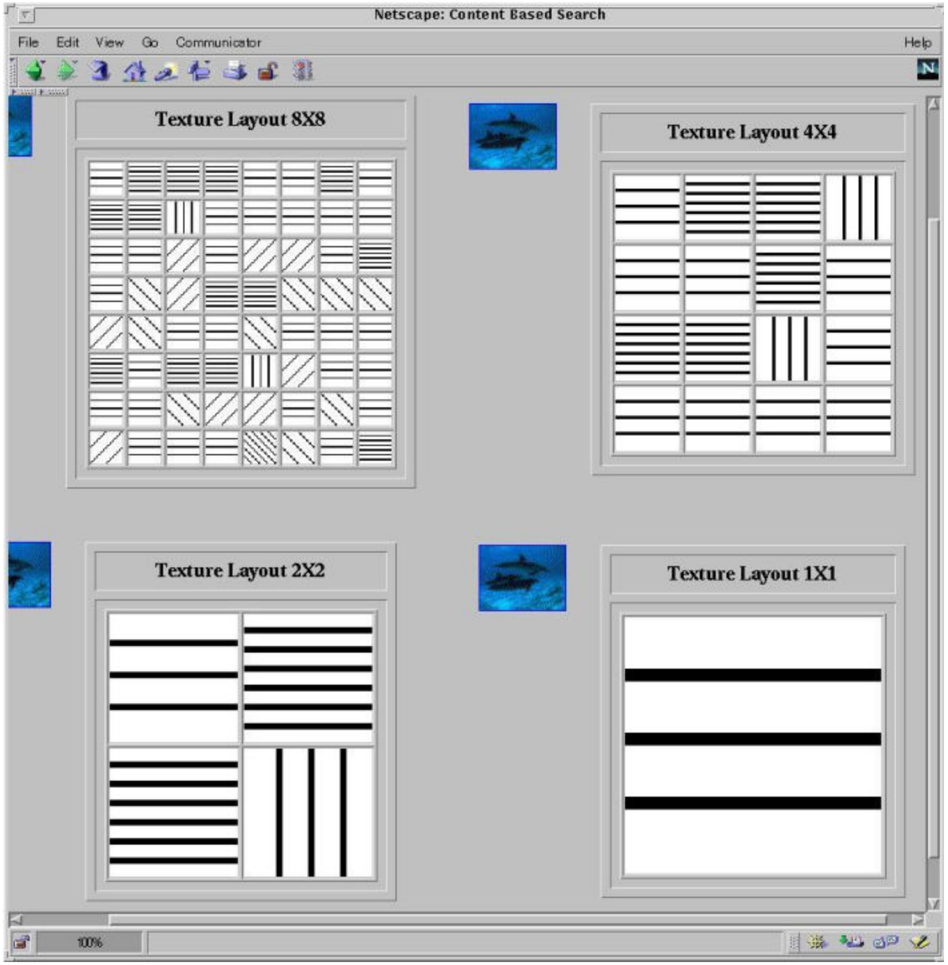
Color histogram


Netscape: Content Based Search
File Edit View Go Communicator Help
100%

Colour Characteristics

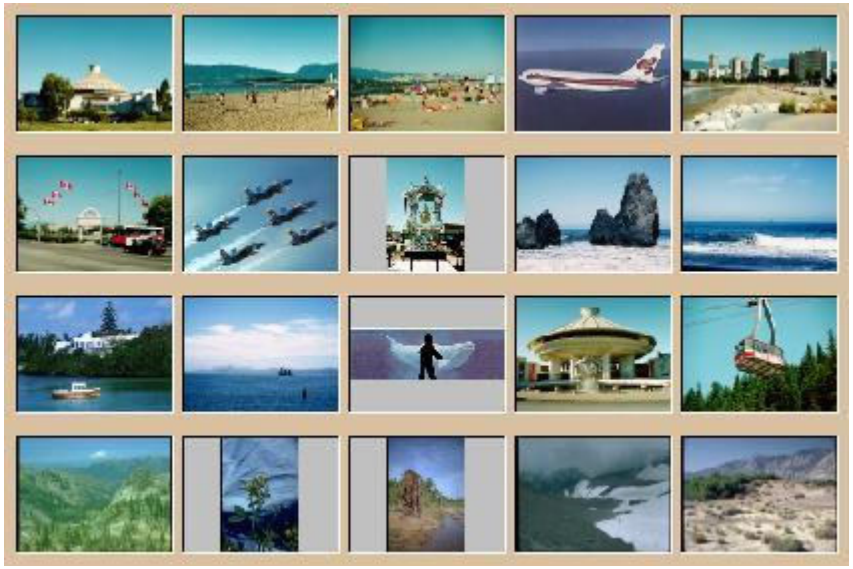
Colour Percentage			
0.73	0.13	0.13	0.01
32	32	32	32
96	32	160	224
192	64	255	255

Texture layout



Mining Multimedia Databases

Refining or combining searches



Search for "blue sky"

(top layout grid is blue)

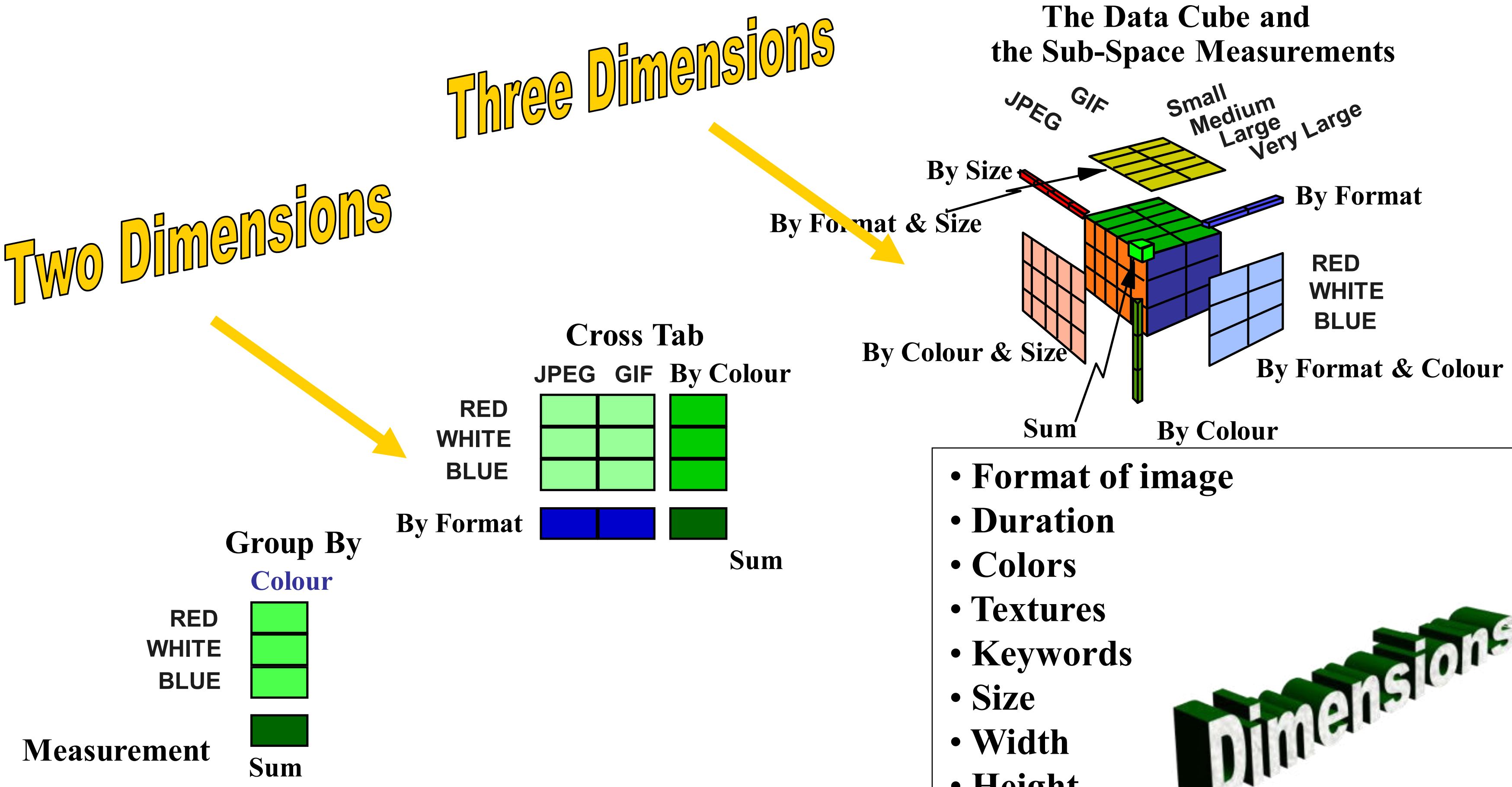
Search for "airplane in blue sky"

(top layout grid is blue and keyword = "airplane")

Search for "blue sky and green meadows"

(top layout grid is blue and bottom is green)

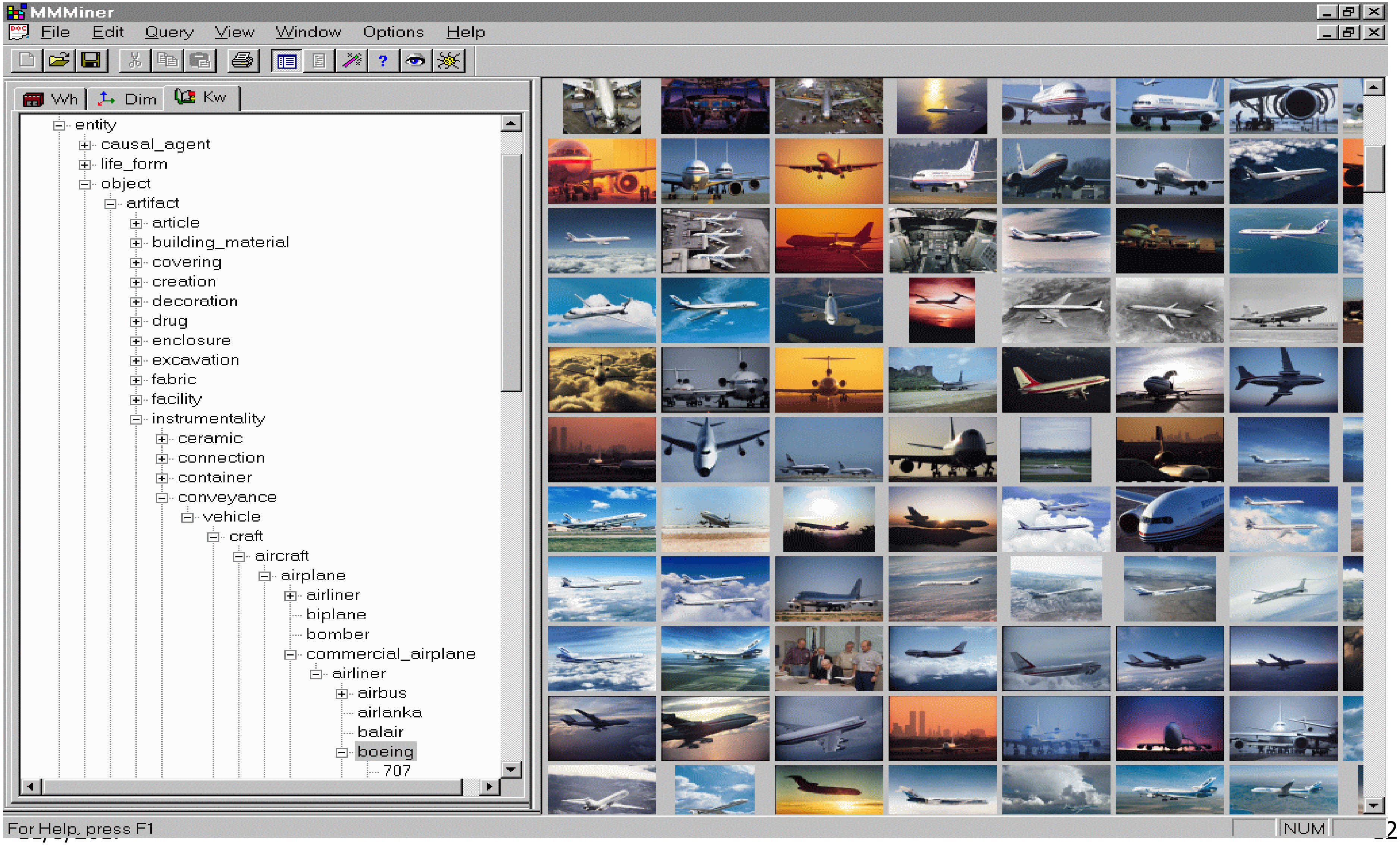
Mining Multimedia Databases



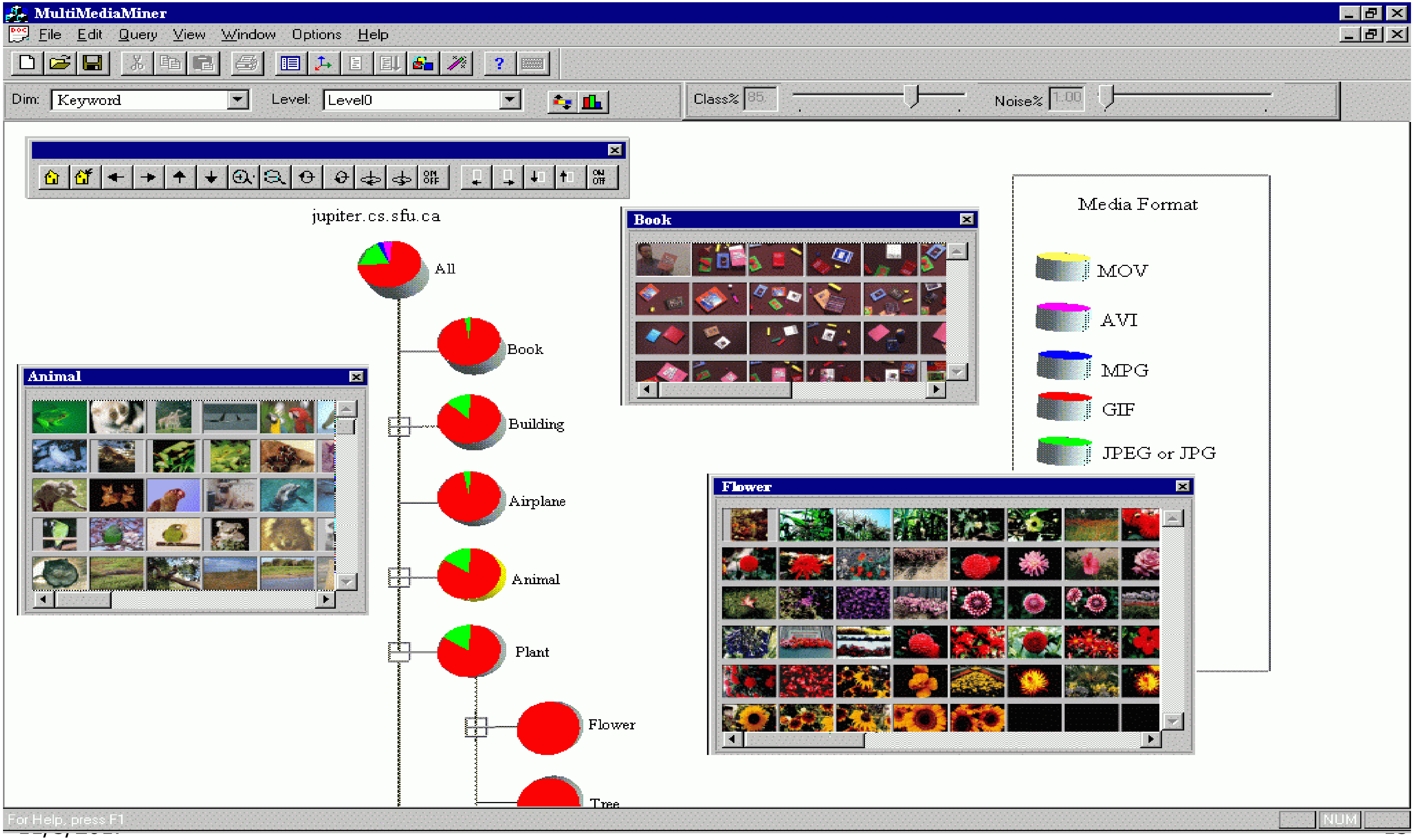
- Format of image
- Duration
- Colors
- Textures
- Keywords
- Size
- Width
- Height
- Internet domain of image
- Internet domain of parent pages
- Image popularity

Dimensions

Mining Multimedia Databases in MultiMediaMiner



Classification in MultiMediaMiner



Dim: Keyword

Level: Level0

Class% 85

Noise% 1.00

For Help, press F1

Mining Associations in Multimedia Data

- Associations between image content and non-image content features
  - "If at least 50% of the upper part of the picture is blue, then it is likely to represent sky."
- Associations among image contents that are not related to spatial relationships
  - "If a picture contains two blue squares, then it is likely to contain one red circle as well."
- Associations among image contents related to spatial relationships
  - "If a red triangle is between two yellow squares, then it is likely a big oval-shaped object is underneath."

Mining Associations in Multimedia Data

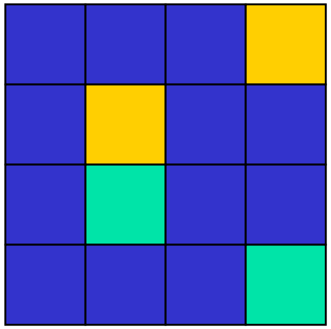
- Special features:
  - Need occurrences besides Boolean existence, e.g.,
    - "Two red square and one blue circle" implies theme "air-show"
  - Need spatial relationships
    - Blue on top of white squared object is associated with brown bottom
  - Need multi-resolution and progressive refinement mining
    - It is expensive to explore detailed associations among objects at high resolution
    - It is crucial to ensure the completeness of search at multi-resolution space

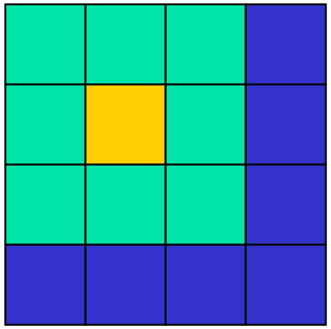
Mining Multimedia Databases

Spatial Relationships from Layout

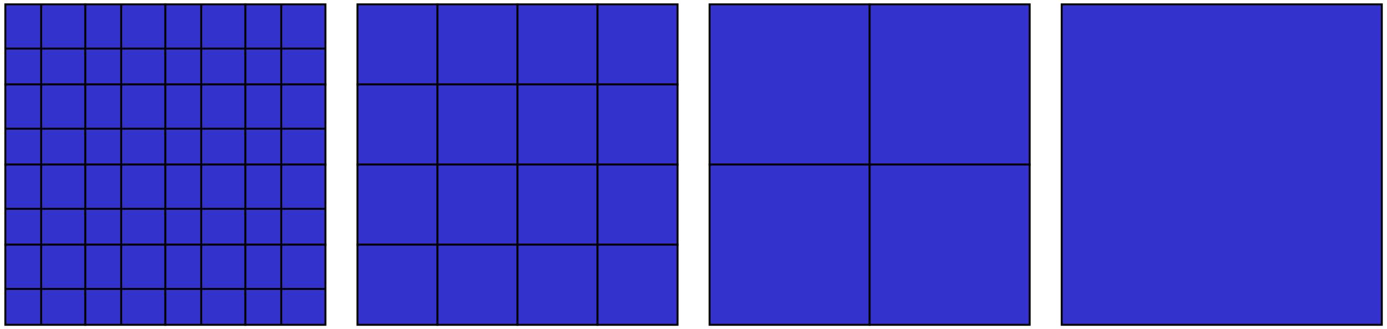
property **P1** ***on-top-of*** property **P2**

property **P1** ***next-to*** property **P2**



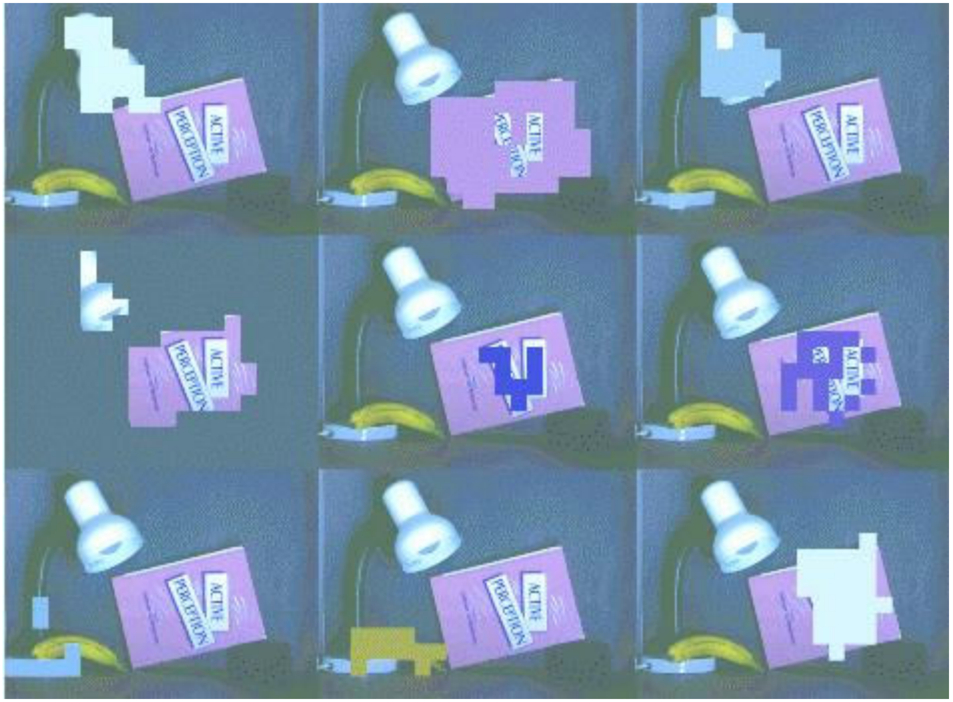


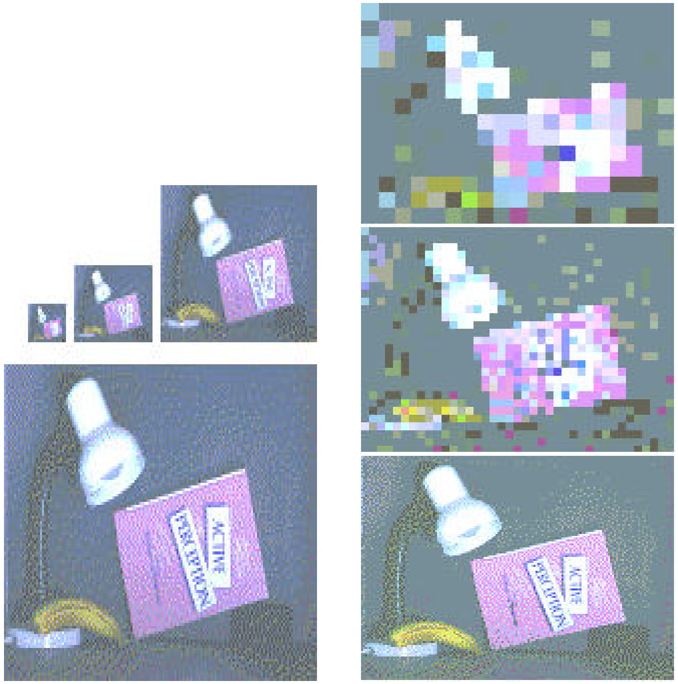
Different Resolution Hierarchy



Mining Multimedia Databases

From Coarse to Fine Resolution Mining





Data Mining:

— Mining Text and Web Data —

Han & Kambr

Babu Ram Dawadi

Mining Text and Web Data

- Text mining, natural language processing and information extraction: An Introduction
- Text categorization methods
- Mining Web linkage structures
- Summary

Mining Text Data: An Introduction

Data Mining / Knowledge Discovery

Structured Data

HomeLoan (
Loanee: *Frank Rizzo*
Lender: *MWF*
Agency: *Lake View*
Amount: *$200,000*
Term: *15 years*
)

Multimedia

Loans($200K,[map],...)

Free Text

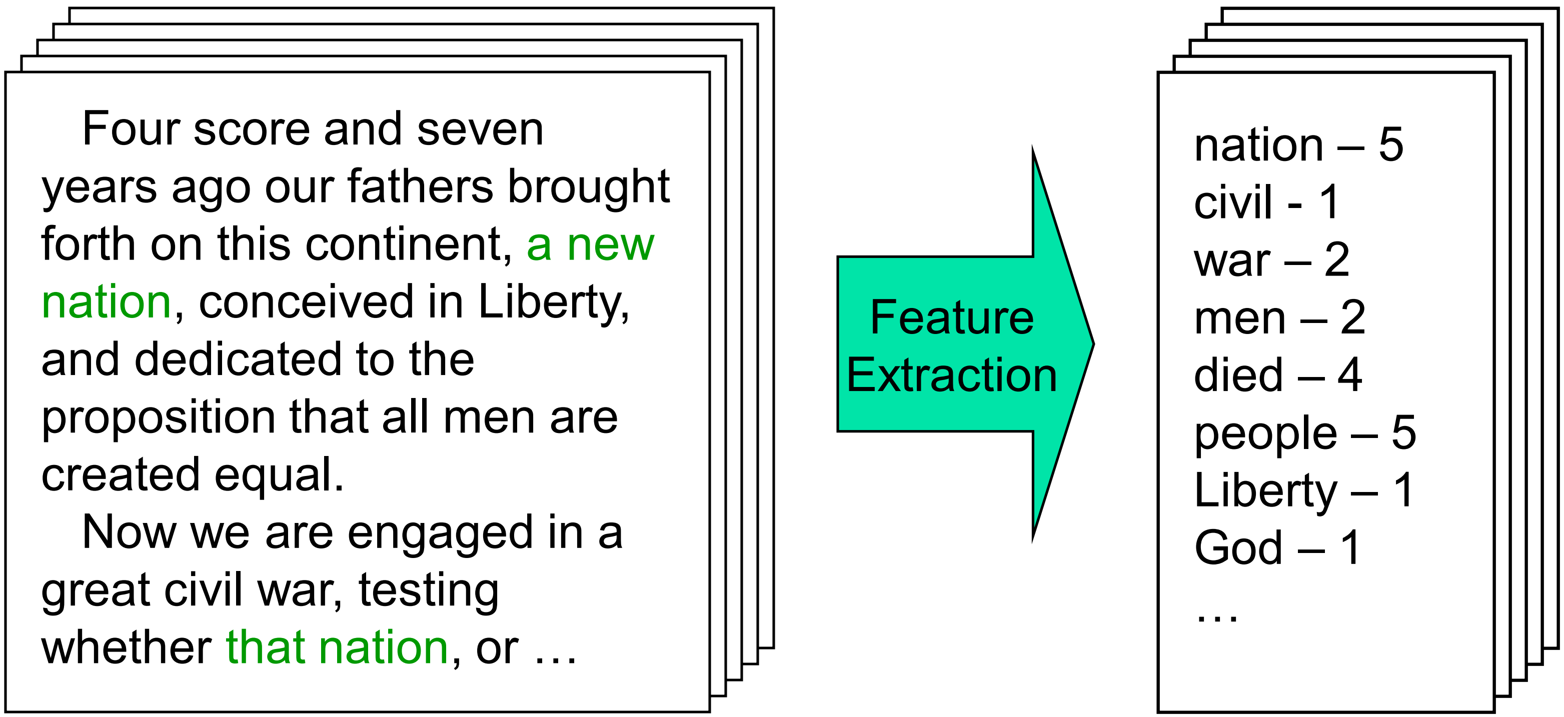
Frank Rizzo bought his home from Lake View Real Estate in 1992.

He paid $200,000 under a15-year loan from MW Financial.

Hypertext

<a href>*Frank Rizzo* </a> *Bought*
<a hef>*this home*</a>
from <a href>*Lake View Real Estate*</a>
In <b>*1992*</b>.
<p>...

Bag-of-Tokens Approaches



Documents

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or …

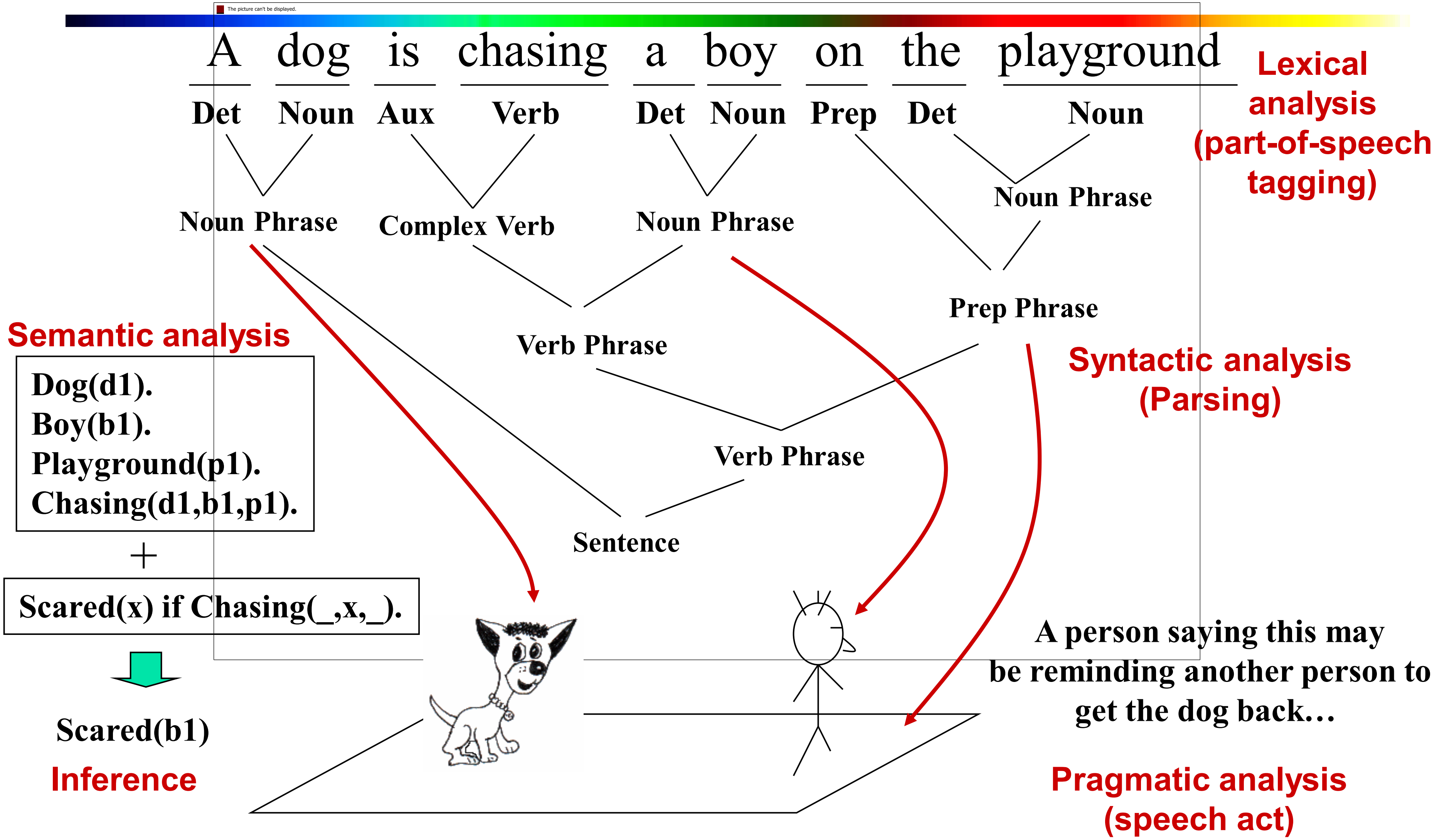
Feature Extraction

Token Sets

nation	5
civil	1
war	2
men	2
died	4
people	5
Liberty	1
God	1
…	

Loses all order-specific information!
Severely limits context!

Natural Language Processing



The picture can't be displayed.

A dog is chasing a boy on the playground

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase Complex Verb Noun Phrase Noun Phrase

Prep Phrase

Verb Phrase

Verb Phrase

Sentence

Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

A person saying this may be reminding another person to get the dog back…

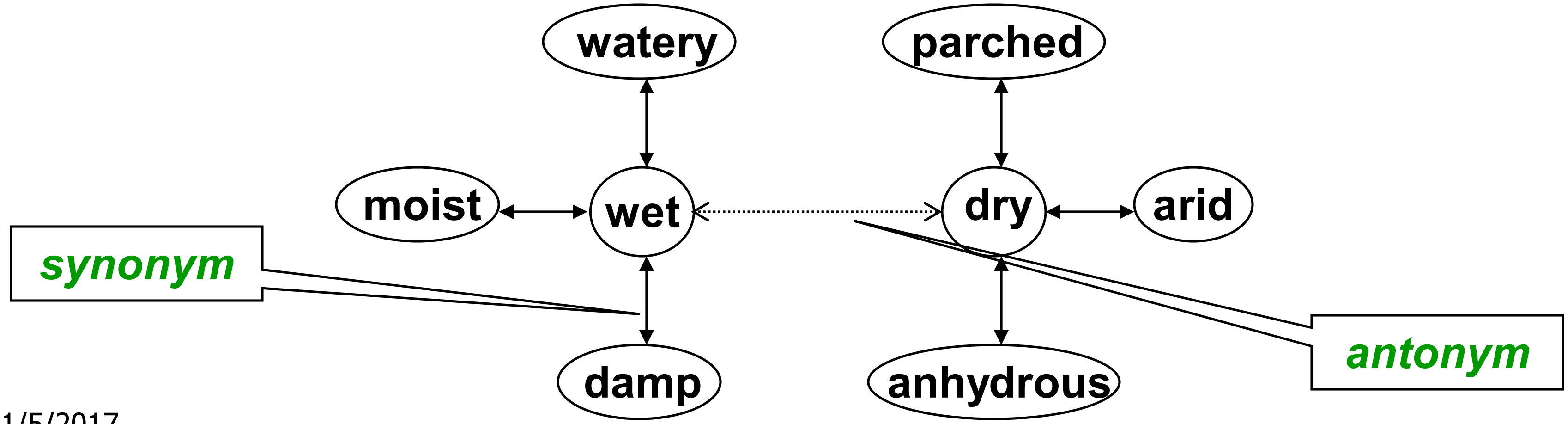
Pragmatic analysis (speech act)

(Taken from ChengXiang Zhai, CS 397cxz – Fall 2003)

WordNet

An extensive lexical network for the English language

- Contains over 138,838 words.
- Several graphs, one for each part-of-speech.
- *Synsets* (synonym sets), each defining a semantic sense.
- Relationship information (antonym, hyponym, meronym …)
- Downloadable for free (UNIX, Windows)
- Expanding to other languages (Global WordNet Association)



Text Databases and Information Retrieval (IR)

- Text databases (document databases)
  - Large collections of documents from various sources: news articles, research papers, books, digital libraries, e-mail messages, and Web pages, library database, etc.
  - Data stored is usually *semi-structured*
  - Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data

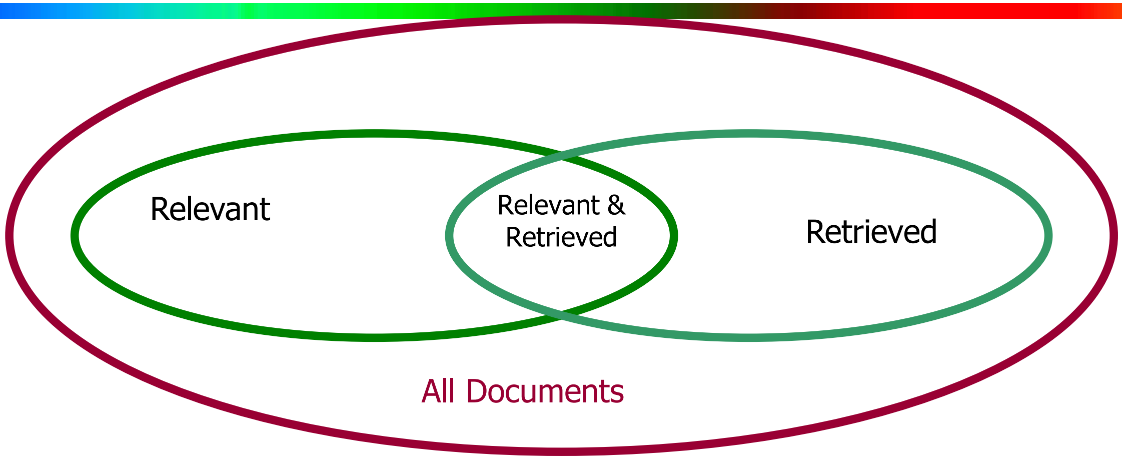
- Information retrieval
  - A field developed in parallel with database systems
  - Information is organized into (a large number of) documents
  - Information retrieval problem: locating relevant documents based on user input, such as keywords or example documents

Information Retrieval

- Typical IR systems
  - Online library catalogs
  - Online document management systems

- Information retrieval vs. database systems
  - Some DB problems are not present in IR, e.g., update, transaction management, complex objects
  - Some IR problems are not addressed well in DBMS, e.g., unstructured documents, approximate search using keywords and relevance

Basic Measures for Text Retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Information Retrieval Techniques

- Basic Concepts
  - A document can be described by a set of representative keywords called index terms.
  - Different index terms have varying relevance when used to describe document contents.
  - This effect is captured through the assignment of numerical weights to each index term of a document. (e.g.: frequency,)
- DBMS Analogy
  - Index Terms → Attributes
  - Weights → Attribute Values

Keyword-Based Retrieval

- A document is represented by a string, which can be identified by a set of keywords

- Queries may use expressions of keywords
  - E.g., car *and* repair shop, tea *or* coffee, DBMS *but not* Oracle
  - Queries and retrieval should consider synonyms, e.g., repair and maintenance

- Major difficulties of the model
  - Synonymy: A keyword *T* does not appear anywhere in the document, even though the document is closely related to *T*, e.g., data mining
  - Polysemy: The same keyword may mean different things in different contexts, e.g., mining

Types of Text Data Mining

- Keyword-based association analysis
- Automatic document classification
- Similarity detection
  - Cluster documents by a common author
  - Cluster documents containing information from a common source
- Link analysis: unusual correlation between entities
- Sequence analysis: predicting a recurring event
- Anomaly detection: find information that violates usual patterns
- Hypertext analysis
  - Patterns in anchors/links
    - Anchor text correlations with linked objects

Keyword-Based Association Analysis

- Motivation
  - Collect sets of keywords or terms that occur frequently together and then find the association or correlation relationships among them
- Association Analysis Process
  - Preprocess the text data by parsing, stemming, removing stop words, etc.
  - Evoke association mining algorithms
    - Consider each document as a transaction
    - View a set of keywords in the document as a set of items in the transaction
  - Term level association mining
    - No need for human effort in tagging documents
    - The number of meaningless results and the execution time is greatly reduced

Text Classification

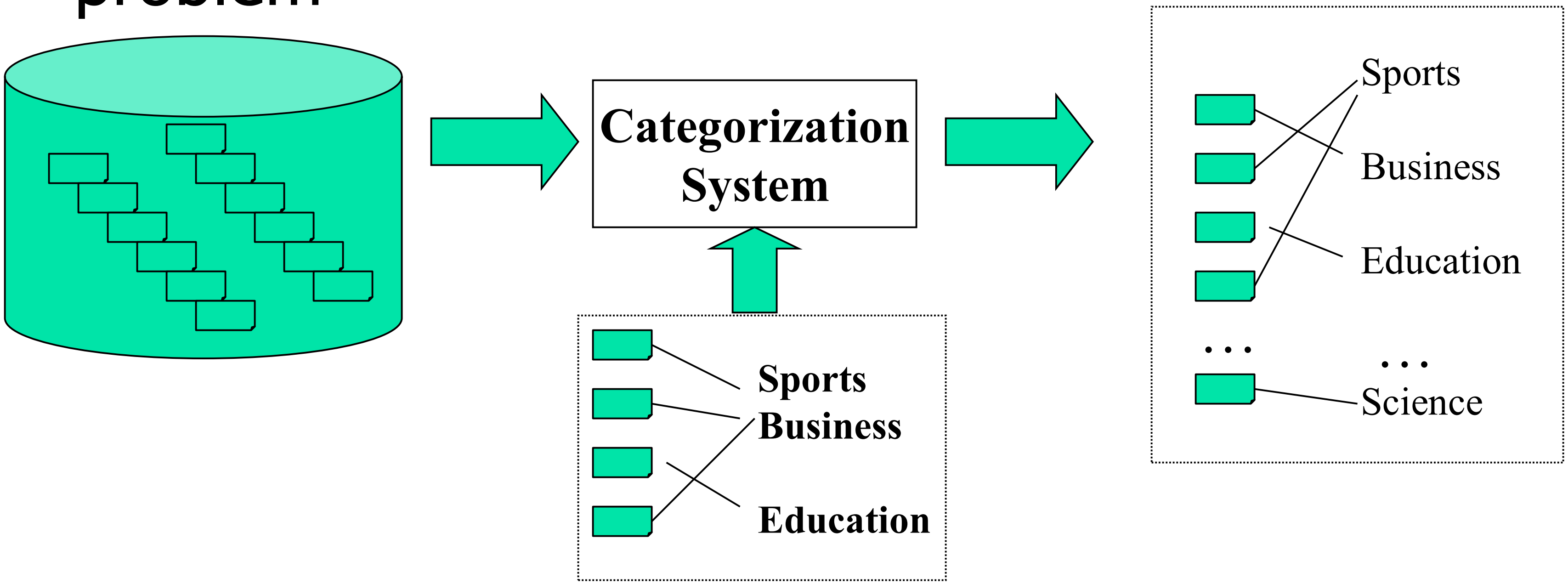
- Motivation
  - Automatic classification for the large number of on-line text documents (Web pages, e-mails, corporate intranets, etc.)
- Classification Process
  - Data preprocessing
  - Definition of training set and test sets
  - Creation of the classification model using the selected classification algorithm
  - Classification model validation
  - Classification of new/unknown text documents
- Text document classification differs from the classification of relational data
  - Document databases are not structured according to attribute-value pairs

Document Clustering

- Motivation
  - Automatically group related documents based on their contents
  - No predetermined training sets or taxonomies
  - Generate a taxonomy at runtime
- Clustering Process
  - Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
  - Hierarchical clustering: compute similarities applying clustering algorithms.
  - Model-Based clustering (Neural Network Approach): clusters are represented by “exemplars”. (e.g.: SOM)

Text Categorization

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard classification (supervised learning) problem



Applications

- News article classification
- Automatic email filtering
- Webpage classification
- Word sense disambiguation
- … …

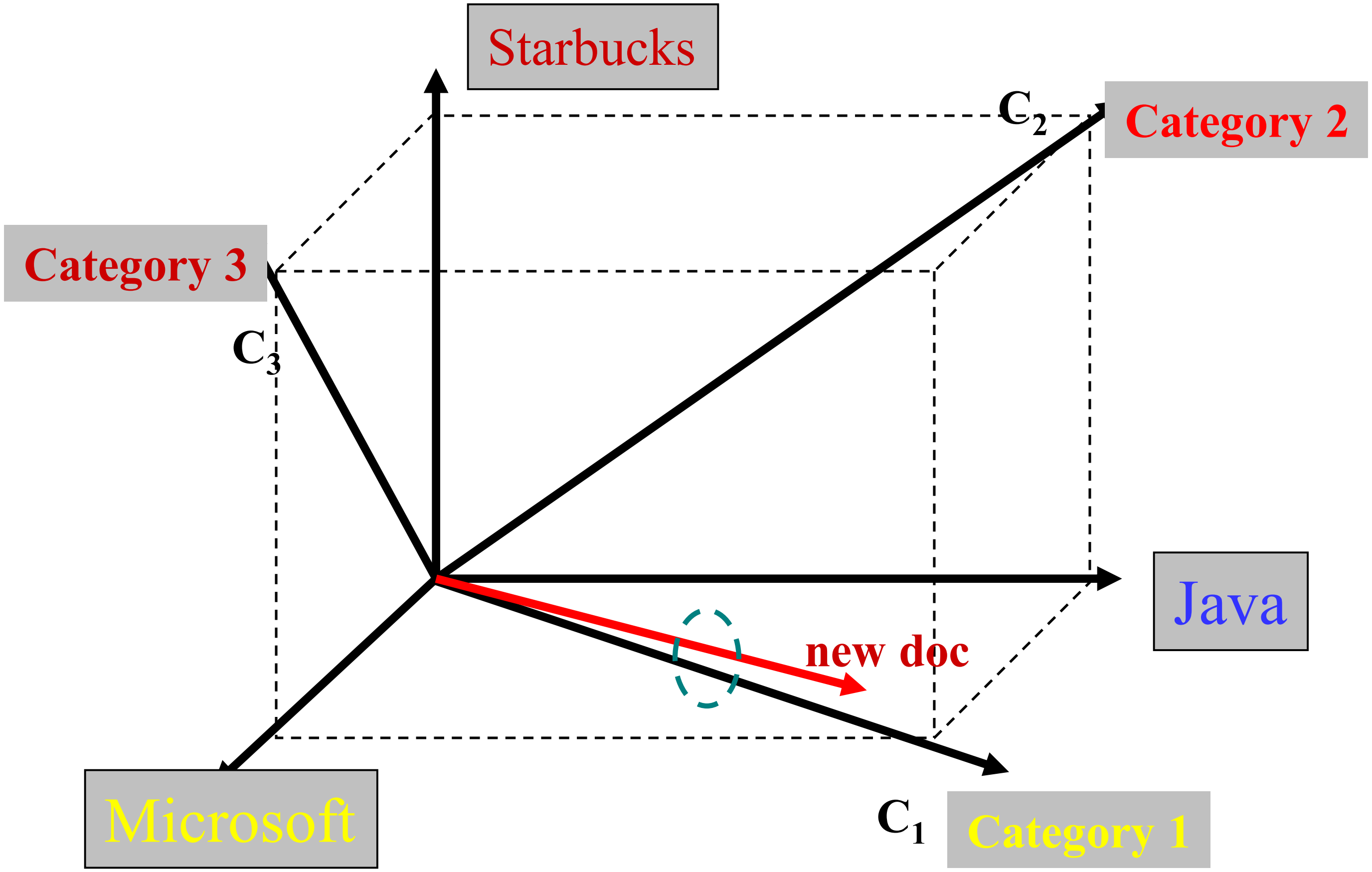
Categorization Methods

- Manual: Typically rule-based
  - Does not scale up (labor-intensive, rule inconsistency)
  - May be appropriate for special data on a particular domain
- Automatic: Typically exploiting machine learning techniques
  - Vector space model based
    - K-nearest neighbor (KNN)
    - Decision-tree (learn rules)
    - Neural Networks (learn non-linear classifier)
    - Support Vector Machines (SVM)
  - Probabilistic or generative model based
    - Naïve Bayes classifier

Vector Space Model

- Represent a doc by a term vector
  - Term: basic concept, e.g., word or phrase
  - Each term defines one dimension
  - N terms define a N-dimensional space
  - Element of vector corresponds to term weight
  - E.g., d = (x_1,…,x_N), x_i is "importance" of term i
- New document is assigned to the most likely category based on vector similarity.

VS Model: Illustration



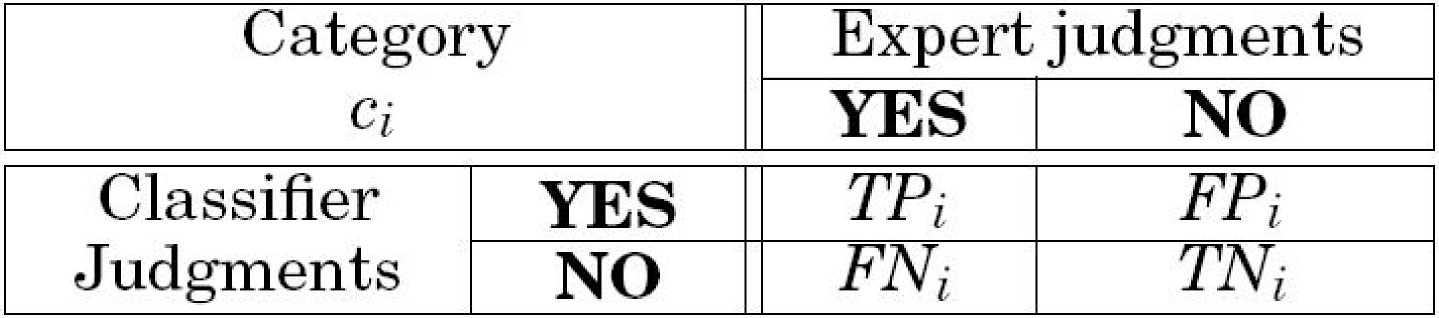
Categorization Methods

- Vector space model
  - K-NN
  - Decision tree
  - Neural network
  - Support vector machine
- Probabilistic model
  - Naïve Bayes classifier
- Many, many others and variants exist [F.S. 02]
  - e.g. Bim, Nb, Ind, Swap-1, LLSF, Widrow-Hoff, Rocchio, Gis-W, … …

Evaluations

- Effectiveness measure
  - Classic: Precision & Recall

Table II. The Contingency Table for Category c_i



Category c_i		Expert judgments	
		YES	**NO**
Classifier Judgments	**YES**	TP_i	FP_i
	NO	FN_i	TN_i

- Precision $$\hat{\pi}_i = \frac{TP_i}{TP_i + FP_i}$$
- Recall $$\hat{\rho}_i = \frac{TP_i}{TP_i + FN_i}.$$

Evaluation (con’t)

- Benchmarks
  - Classic: Reuters collection
    - A set of newswire stories classified under categories related to economics.
- Effectiveness
  - Difficulties of strict comparison
    - different parameter setting
    - different “split” (or selection) between training and testing
    - various optimizations … …
  - However widely recognizable
    - Best: Boosting-based committee classifier & SVM
    - Worst: Naïve Bayes classifier
  - Need to consider other factors, especially efficiency

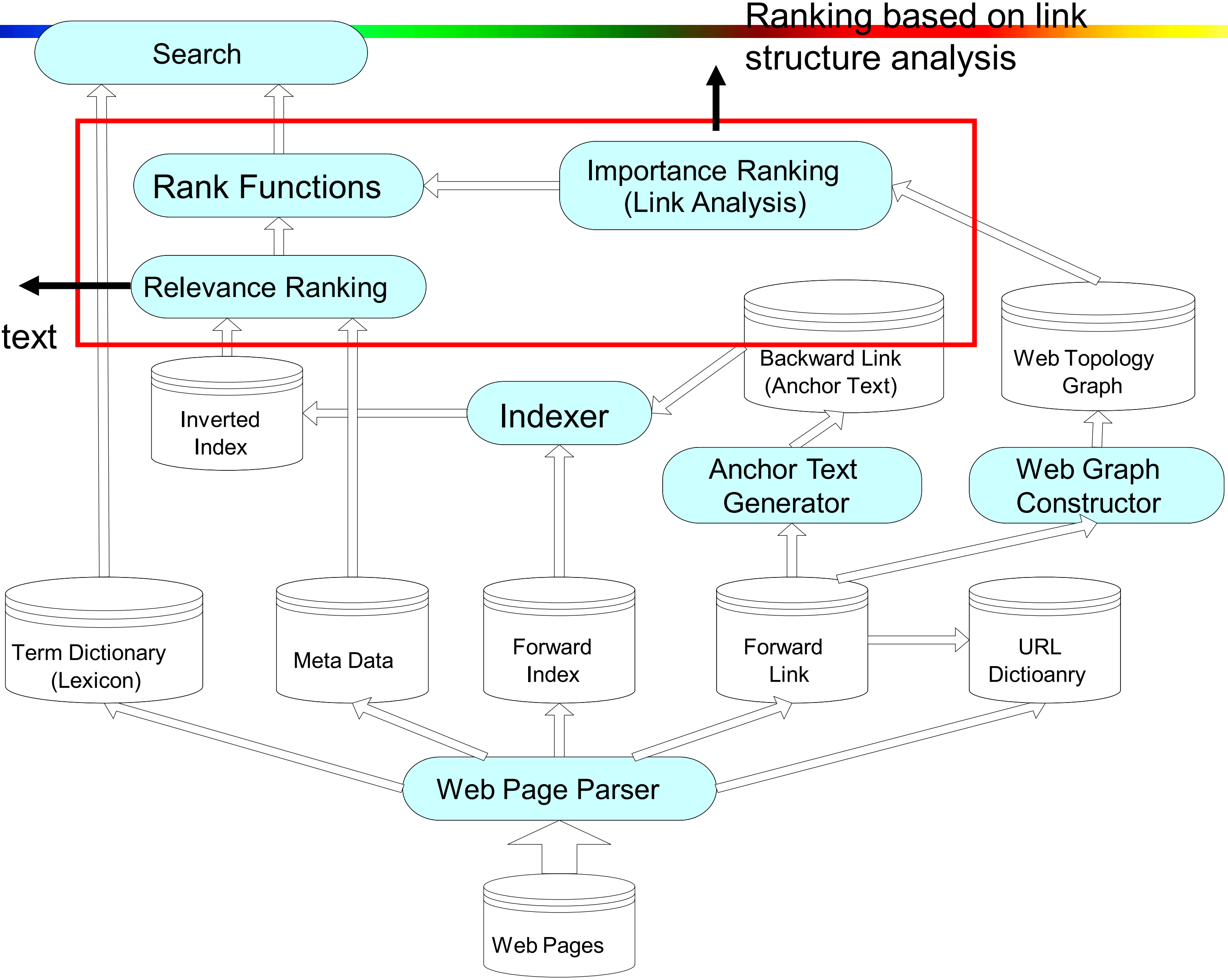
Mining Text and Web Data

- Text mining, natural language processing and information extraction: An Introduction
- Text categorization methods
- Mining Web linkage structures
  - Based on the slides by Deng Cai
- Summary

Outline

- Background on Web Search
- VIPS (VIsion-based Page Segmentation)
- Block-based Web Search
- Block-based Link Analysis
- Web Image Search & Clustering

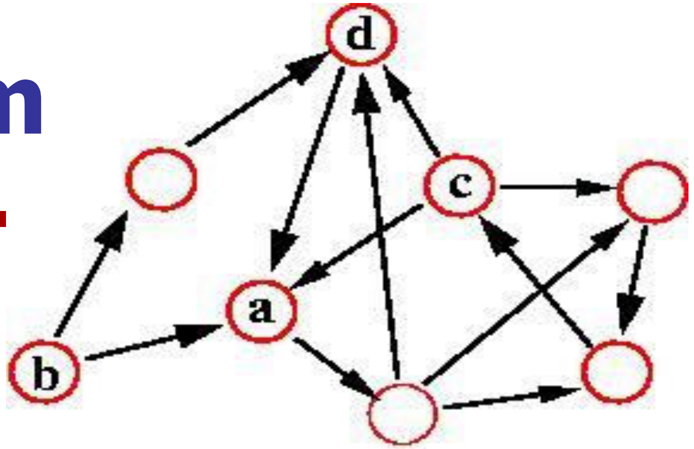
Search Engine – Two Rank Functions



Ranking based on link structure analysis

Similarity based on content or text

The PageRank Algorithm



s(a) ~ s(b) + s(c) + s(d) ?

- Basic idea
  - ***significance of a page is determined by the significance of the pages linking to it***

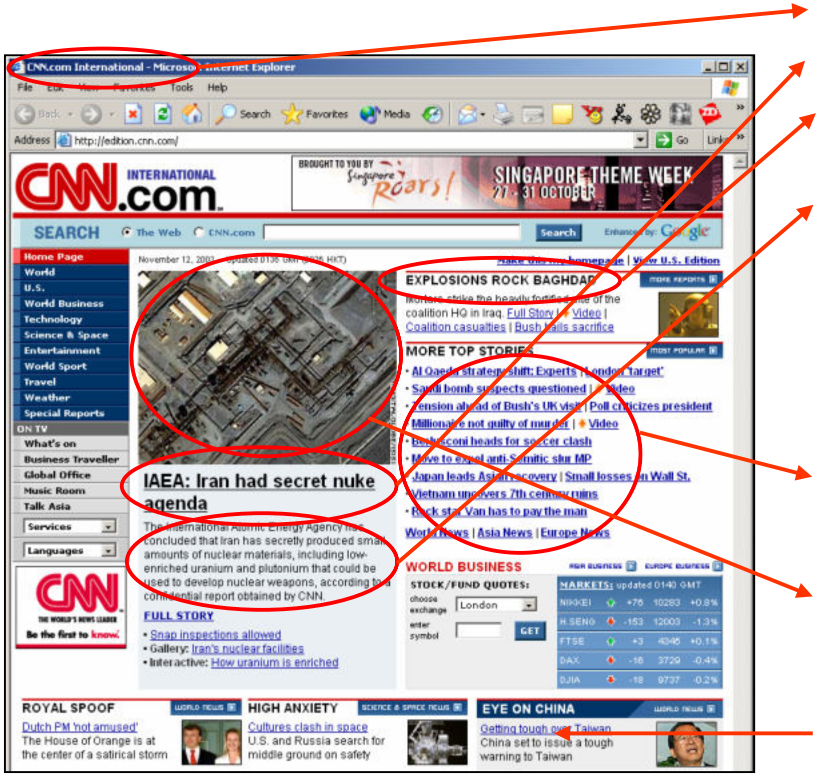
$$A_{ij} = \begin{cases} 1 & if\ \ page\ i\ links\ to\ page\ j \\ 0 & otherwise \end{cases}$$

- More precisely:
  - Link graph: adjacency matrix *A*,
  - Constructs a probability transition matrix *M* by renormalizing each row of *A* to sum to 1 $\varepsilon U + (1-\varepsilon)M$ $U_{ij} = 1/n\ for\ all\ i, j$
  - Treat the web graph as a markov chain (random surfer)
  - The vector of PageRank scores *p* is then defined to be the stationary distribution of this Markov chain. Equivalently, p is the principal right eigenvector of the transition matrix $(\varepsilon U + (1-\varepsilon)M)^T$

$$(\varepsilon U + (1-\varepsilon)M)^T p = p$$

Layout Structure

- Compared to plain text, a web page is a 2D presentation
  - Rich visual effects created by different term types, formats, separators, blank areas, colors, pictures, etc
  - Different parts of a page are not equally important



CNN.com International - Microsoft Internet Explorer

CNN.com INTERNATIONAL

BROUGHT TO YOU BY Singapore Roars! SINGAPORE THEME WEEK 27 - 31 OCTOBER

CNN THE WORLD'S NEWS LEADER Be the first to know.

November 12, 2003 -- Updated 0136 GMT (0936 HKT)

Make this my homepage | View U.S. Edition

IAEA: Iran had secret nuke agenda

The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials, including low-enriched uranium and plutonium that could be used to develop nuclear weapons, according to a confidential report obtained by CNN.

FULL STORY

- Snap inspections allowed
- Gallery: Iran's nuclear facilities
- Interactive: How uranium is enriched

EXPLOSIONS ROCK BAGHDAD MORE REPORTS

Mortars strike the heavily fortified site of the coalition HQ in Iraq. Full Story | Video | Coalition casualties | Bush hails sacrifice

MORE TOP STORIES MOST POPULAR

- Al Qaeda strategy shift: Experts | London 'target'
- Saudi bomb suspects questioned | Video
- Tension ahead of Bush's UK visit | Poll criticizes president
- Millionaire not guilty of murder | Video
- Berlusconi heads for soccer clash
- Move to expel anti-Semitic slur MP
- Japan leads Asian recovery | Small losses on Wall St.
- Vietnam uncovers 7th century ruins
- Rock star Van has to pay the man

World News | Asia News | Europe News

WORLD BUSINESS ASIA BUSINESS EUROPE BUSINESS

STOCK/FUND QUOTES: choose exchange London; enter symbol GET

MARKETS: updated 0140 GMT

NIKKEI	▲	+76	10283	+0.8%
H.SENG	▼	-153	12003	-1.3%
FTSE	▲	+3	4345	+0.1%
DAX	▼	-18	3729	-0.4%
DJIA	▼	-18	9737	-0.2%

ROYAL SPOOF WORLD NEWS

Dutch PM 'not amused' The House of Orange is at the center of a satirical storm

HIGH ANXIETY SCIENCE & SPACE NEWS

Cultures clash in space U.S. and Russia search for middle ground on safety

EYE ON CHINA WORLD NEWS

Getting tough over Taiwan China set to issue a tough warning to Taiwan

Title: CNN.com International

H1: IAEA: Iran had secret nuke agenda

H3: EXPLOSIONS ROCK BAGHDAD

…

TEXT BODY (with position and font type): The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials including low enriched uranium and plutonium that could be used to develop nuclear weapons according to a confidential report obtained by CNN…

Hyperlink:

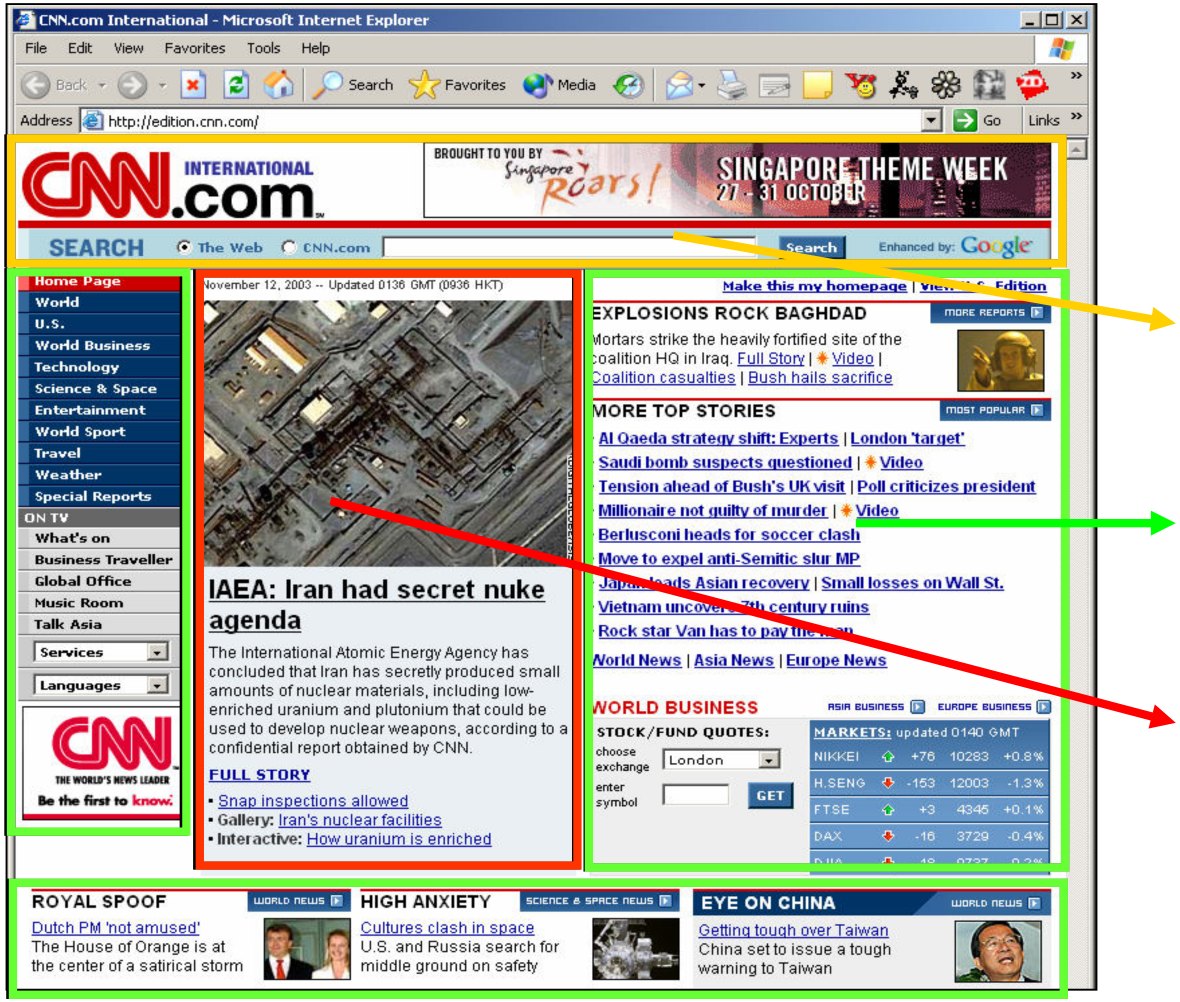
- URL: http://www.cnn.com/...
- Anchor Text: Al oaeda…

Image:

- URL: http://www.cnn.com/image/...
- Alt & Caption: Iran nuclear …

Anchor Text: CNN Homepage News …

Web Page Block—Better Information Unit


CNN.com International - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://edition.cnn.com/
Go Links
CNN.com INTERNATIONAL
BROUGHT TO YOU BY Singapore Roars!
SINGAPORE THEME WEEK 27 - 31 OCTOBER
SEARCH ● The Web ○ CNN.com Search Enhanced by: Google
Home Page
World
U.S.
World Business
Technology
Science & Space
Entertainment
World Sport
Travel
Weather
Special Reports
ON TV
What's on
Business Traveller
Global Office
Music Room
Talk Asia
Services
Languages
CNN THE WORLD'S NEWS LEADER
Be the first to know.

November 12, 2003 -- Updated 0136 GMT (0936 HKT)

IAEA: Iran had secret nuke agenda

The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials, including low-enriched uranium and plutonium that could be used to develop nuclear weapons, according to a confidential report obtained by CNN.

FULL STORY

- Snap inspections allowed
- Gallery: Iran's nuclear facilities
- Interactive: How uranium is enriched

Make this my homepage | View U.S. Edition

EXPLOSIONS ROCK BAGHDAD — MORE REPORTS

Mortars strike the heavily fortified site of the coalition HQ in Iraq. Full Story | Video | Coalition casualties | Bush hails sacrifice

MORE TOP STORIES — MOST POPULAR

Al Qaeda strategy shift: Experts | London 'target'
Saudi bomb suspects questioned | Video
Tension ahead of Bush's UK visit | Poll criticizes president
Millionaire not guilty of murder | Video
Berlusconi heads for soccer clash
Move to expel anti-Semitic slur MP
Japan leads Asian recovery | Small losses on Wall St.
Vietnam uncovers 7th century ruins
Rock star Van has to pay the man

World News | Asia News | Europe News

WORLD BUSINESS — ASIA BUSINESS — EUROPE BUSINESS

STOCK/FUND QUOTES: choose exchange London; enter symbol GET

MARKETS: updated 0140 GMT				
NIKKEI	↑	+76	10283	+0.8%
H.SENG	↓	-153	12003	-1.3%
FTSE	↑	+3	4345	+0.1%
DAX	↓	-16	3729	-0.4%

ROYAL SPOOF — WORLD NEWS

Dutch PM 'not amused'
The House of Orange is at the center of a satirical storm

HIGH ANXIETY — SCIENCE & SPACE NEWS

Cultures clash in space
U.S. and Russia search for middle ground on safety

EYE ON CHINA — WORLD NEWS

Getting tough over Taiwan
China set to issue a tough warning to Taiwan

Web Page Blocks

Importance = Low

Importance = Med

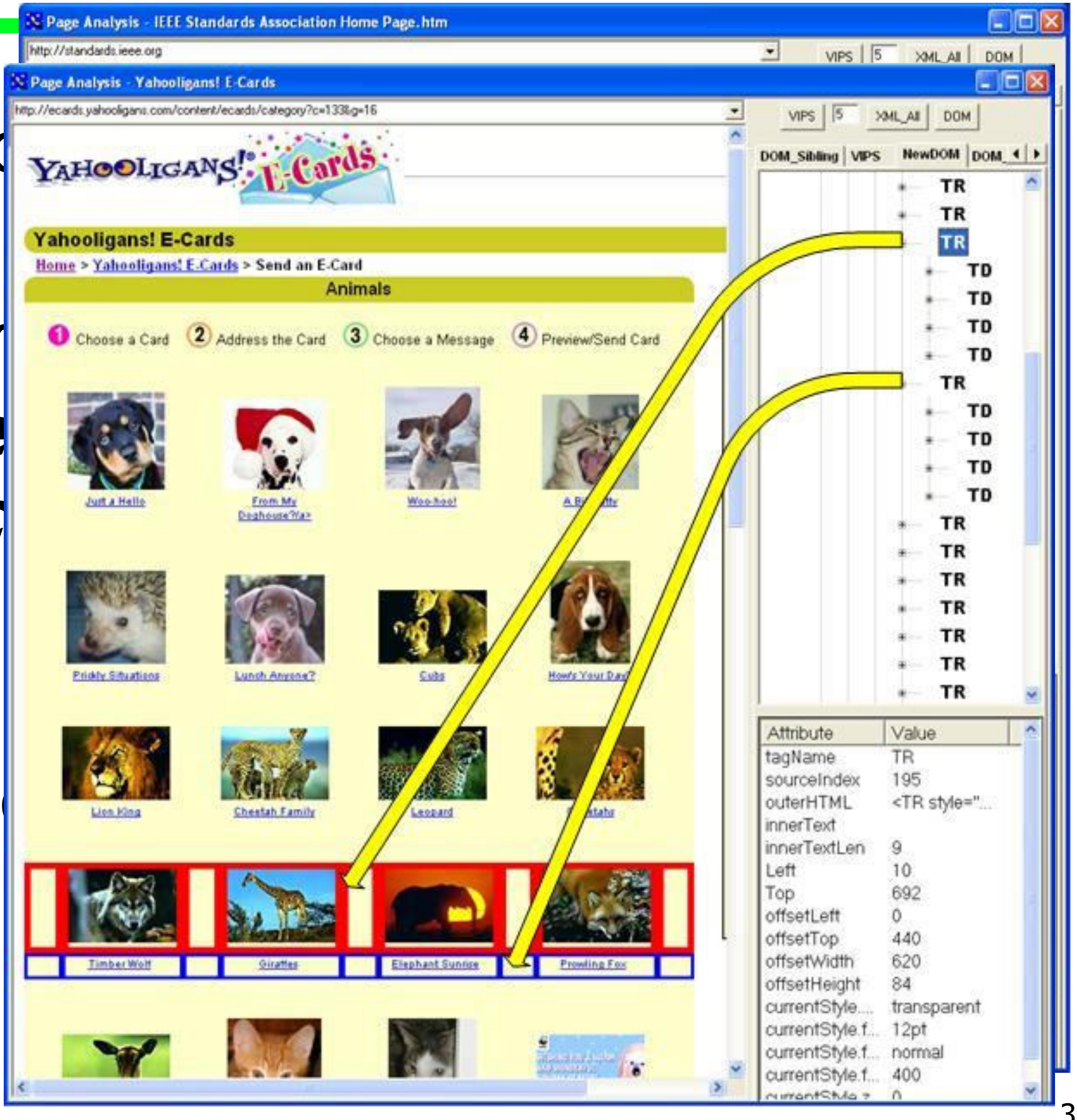
Importance = High

Motivation for VIPS (VIsion-based Page Segmentation)

- Problems of treating a web page as an atomic unit
  - Web page usually contains not only pure content
    - Noise: navigation, decoration, interaction, …
  - Multiple topics
  - Different parts of a page are not equally important
- Web page has internal structure
  - Two-dimension logical structure & Visual layout presentation
  - > Free text document
  - < Structured document
- Layout – the 3rd dimension of Web page
  - 1st dimension: content
  - 2nd dimension: hyperlink

Is DOM a Good Representation of Page Structure?

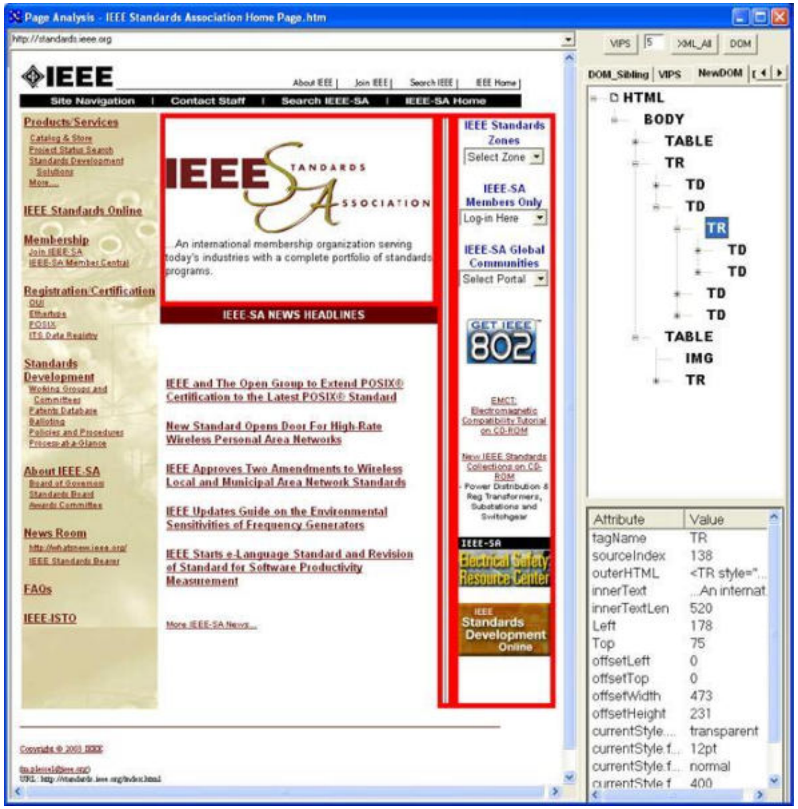
- Page segmentatio
  - Extract structu UL, TITLE, H1-
  - ***DOM is more does not nec structure***
- How about XML?
  - A long way to



Page Analysis - IEEE Standards Association Home Page.htm
http://standards.ieee.org
VIPS
5
XML_All
DOM
Page Analysis - Yahooligans! E-Cards
http://ecards.yahooligans.com/content/ecards/category?c=133&g=16
VIPS
5
XML_All
DOM
DOM_Sibling
VIPS
NewDOM
DOM_
YAHOOLIGANS! E-Cards
Yahooligans! E-Cards
Home > Yahooligans! E-Cards > Send an E-Card
Animals
1 Choose a Card
2 Address the Card
3 Choose a Message
4 Preview/Send Card
TR
TR
TR
TD
TD
TD
TD
TR
TD
TD
TD
TD
TR
TR
TR
TR
TR
TR
TR

Attribute	Value
tagName	TR
sourceIndex	195
outerHTML	<TR style="...
innerText	
innerTextLen	9
Left	10
Top	692
offsetLeft	0
offsetTop	440
offsetWidth	620
offsetHeight	84
currentStyle....	transparent
currentStyle.f...	12pt
currentStyle.f...	normal
currentStyle.f...	400
currentStyle.z...	0

Example of Web Page Segmentation (1)


Page Analysis - IEEE Standards Association Home Page.htm
http://standards.ieee.org
VIPS
5
XML_All
DOM
DOM_Sibling
VIPS
NewDOM
IEEE
About IEEE | Join IEEE | Search IEEE | IEEE Home |
Site Navigation | Contact Staff | Search IEEE-SA | IEEE-SA Home
Products/Services
Catalog & Store
Project Status Search
Standards Development Solutions
More...
IEEE Standards Online
Membership
Join IEEE-SA
IEEE-SA Member Central
Registration/Certification
OUI
Ethertype
POSIX
ITS Data Registry
Standards Development
Working Groups and Committees
Patents Database
Balloting
Policies and Procedures
Process-at-a-Glance
About IEEE-SA
Board of Governors
Standards Board
Awards Committee
News Room
http://whatsnew.ieee.org/
IEEE Standards Bearer
FAQs
IEEE-ISTO
IEEE STANDARDS ASSOCIATION
IEEE-SA NEWS HEADLINES
More IEEE-SA News...
IEEE Standards Zones
Select Zone
IEEE-SA Members Only
Log-in Here
IEEE-SA Global Communities
Select Portal
GET IEEE 802
EMC Electromagnetic Compatibility Tutorial on CD-ROM
New IEEE Standards Collections on CD-ROM
Power Distribution & Reg Transformers, Substations and Switchgear
IEEE-SA Electrical Safety Resource Center
IEEE Standards Development Online
Copyright © 2003 IEEE

...An international membership organization serving today's industries with a complete portfolio of standards programs.

IEEE and The Open Group to Extend POSIX® Certification to the Latest POSIX® Standard

New Standard Opens Door For High-Rate Wireless Personal Area Networks

IEEE Approves Two Amendments to Wireless Local and Municipal Area Network Standards

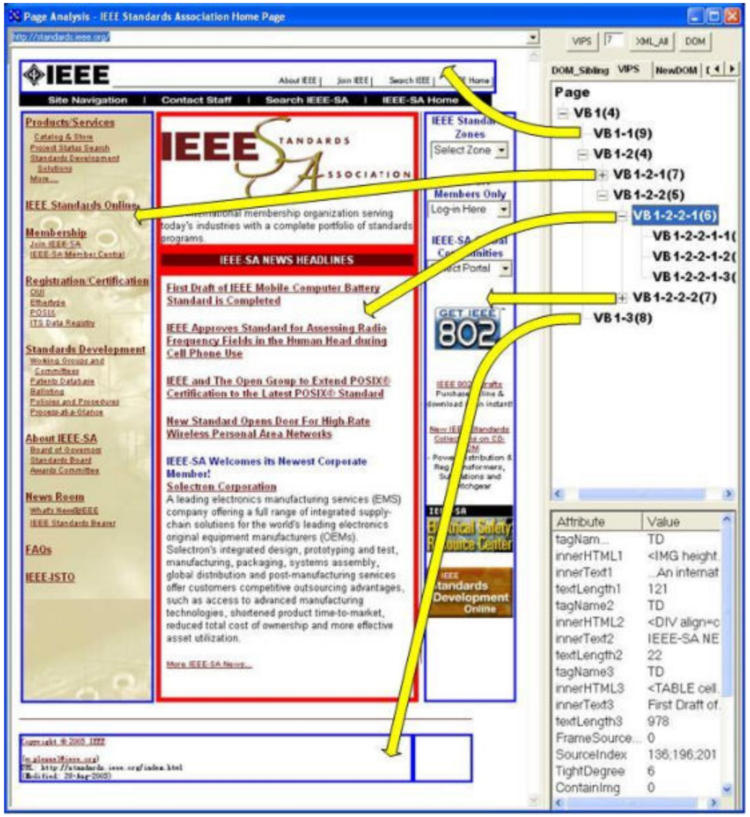
IEEE Updates Guide on the Environmental Sensitivities of Frequency Generators

IEEE Starts e-Language Standard and Revision of Standard for Software Productivity Measurement

```
HTML
  BODY
    TABLE
    TR
      TD
      TD
        TR
          TD
          TD
        TD
        TD
    TABLE
      IMG
      TR
```

Attribute	Value
tagName	TR
sourceIndex	138
outerHTML	<TR style="...
innerText	...An internat
innerTextLen	520
Left	178
Top	75
offsetLeft	0
offsetTop	0
offsetWidth	473
offsetHeight	231
currentStyle....	transparent
currentStyle.f...	12pt
currentStyle.f...	normal
currentStyle.f	400

(DOM Structure)



...An international membership organization serving today's industries with a complete portfolio of standards programs.

First Draft of IEEE Mobile Computer Battery Standard is Completed

IEEE Approves Standard for Assessing Radio Frequency Fields in the Human Head during Cell Phone Use

IEEE and The Open Group to Extend POSIX® Certification to the Latest POSIX® Standard

New Standard Opens Door For High-Rate Wireless Personal Area Networks

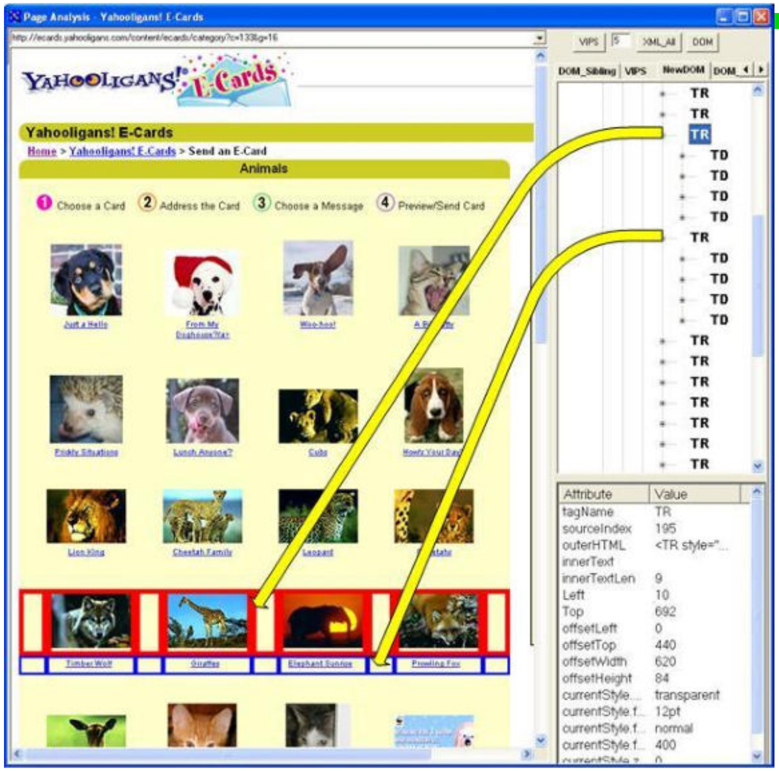
IEEE-SA Welcomes its Newest Corporate Member!
Selectron Corporation
A leading electronics manufacturing services (EMS) company offering a full range of integrated supply-chain solutions for the world's leading electronics original equipment manufacturers (OEMs). Selectron's integrated design, prototyping and test, manufacturing, packaging, systems assembly, global distribution and post-manufacturing services offer customers competitive outsourcing advantages, such as access to advanced manufacturing technologies, shortened product time-to-market, reduced total cost of ownership and more effective asset utilization.

```
Page
  VB 1(4)
    VB 1-1(9)
    VB 1-2(4)
      VB 1-2-1(7)
      VB 1-2-2(5)
        VB 1-2-2-1(6)
          VB 1-2-2-1-1(
          VB 1-2-2-1-2(
          VB 1-2-2-1-3(
        VB 1-2-2-2(7)
    VB 1-3(8)
```

Attribute	Value
tagNam...	TD
innerHTML1	<IMG height.
innerText1	...An internat
textLength1	121
tagName2	TD
innerHTML2	<DIV align=c
innerText2	IEEE-SA NE
textLength2	22
tagName3	TD
innerHTML3	<TABLE cell.
innerText3	First Draft of.
textLength3	978
FrameSource...	0
SourceIndex	136;196;201
TightDegree	6
ContainImg	0

(VIPS Structure)

Example of Web Page Segmentation (2)

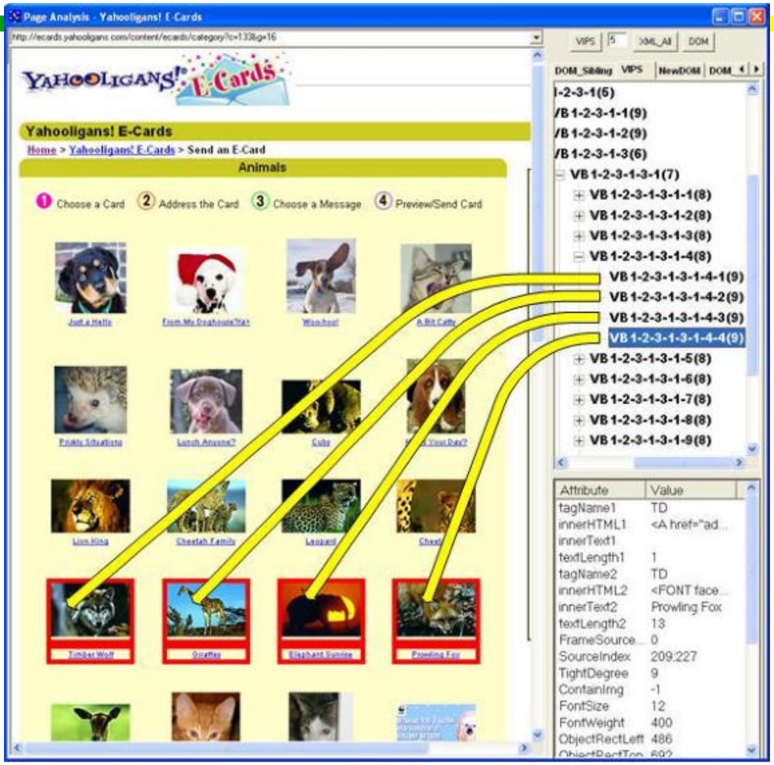


Page Analysis - Yahooligans! E-Cards

```
TR
TR
TR
    TD
    TD
    TD
    TD
TR
    TD
    TD
    TD
    TD
TR
TR
TR
TR
TR
TR
TR
```

Attribute	Value
tagName	TR
sourceIndex	195
outerHTML	<TR style="...
innerText	
innerTextLen	9
Left	10
Top	692
offsetLeft	0
offsetTop	440
offsetWidth	620
offsetHeight	84
currentStyle....	transparent
currentStyle.f...	12pt
currentStyle.f...	normal
currentStyle.f...	400
currentStyle.z...	0

(DOM Structure)



Page Analysis - Yahooligans! E-Cards

```
1-2-3-1(5)
VB 1-2-3-1-1(9)
VB 1-2-3-1-2(9)
VB 1-2-3-1-3(6)
  VB 1-2-3-1-3-1(7)
    VB 1-2-3-1-3-1-1(8)
    VB 1-2-3-1-3-1-2(8)
    VB 1-2-3-1-3-1-3(8)
    VB 1-2-3-1-3-1-4(8)
      VB 1-2-3-1-3-1-4-1(9)
      VB 1-2-3-1-3-1-4-2(9)
      VB 1-2-3-1-3-1-4-3(9)
      VB 1-2-3-1-3-1-4-4(9)
    VB 1-2-3-1-3-1-5(8)
    VB 1-2-3-1-3-1-6(8)
    VB 1-2-3-1-3-1-7(8)
    VB 1-2-3-1-3-1-8(8)
    VB 1-2-3-1-3-1-9(8)
```

Attribute	Value
tagName1	TD
innerHTML1	<A href="ad...
innerText1	
textLength1	1
tagName2	TD
innerHTML2	<FONT face...
innerText2	Prowling Fox
textLength2	13
FrameSource...	0
SourceIndex	209:227
TightDegree	9
ContainImg	-1
FontSize	12
FontWeight	400
ObjectRectLeft	486
ObjectRectTop	692

(VIPS Structure)

- Can be applied on web image retrieval
  - Surrounding text extraction

Web Page Block—Better Information Unit

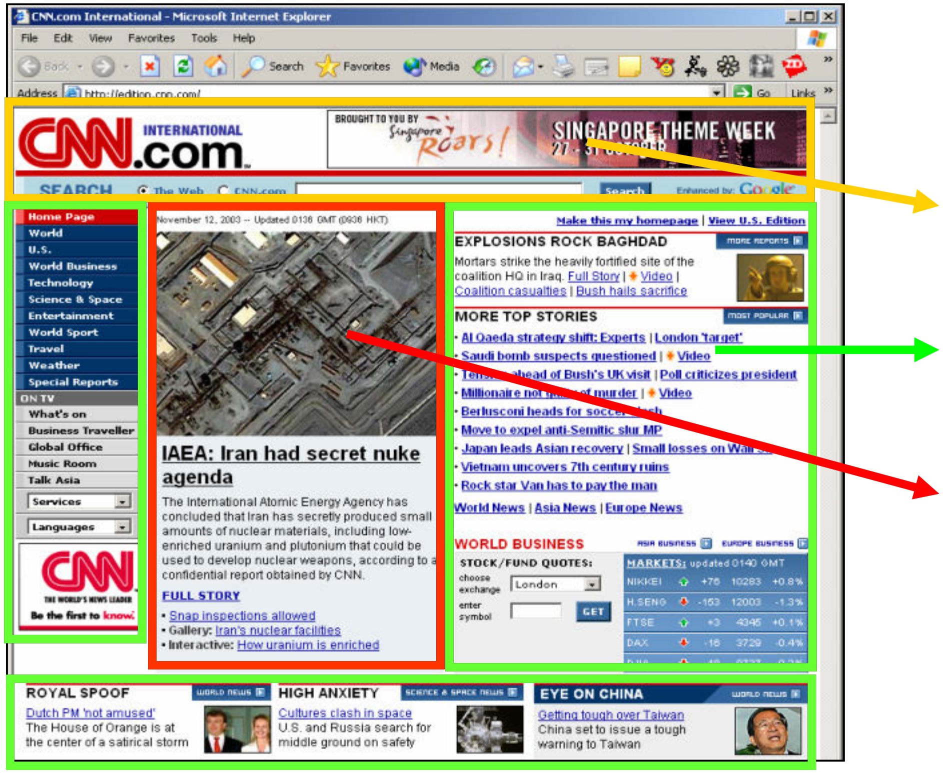
Page Segmentation

- **Vision based approach**

→

Block Importance Modeling

- **Statistical learning**



CNN.com International - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://edition.cnn.com/
Go Links
CNN INTERNATIONAL .com
BROUGHT TO YOU BY Singapore Roars!
SINGAPORE THEME WEEK 27 - 31 OCTOBER
SEARCH ◉ The Web ○ CNN.com Search Enhanced by: Google
Home Page
World
U.S.
World Business
Technology
Science & Space
Entertainment
World Sport
Travel
Weather
Special Reports
ON TV
What's on
Business Traveller
Global Office
Music Room
Talk Asia
Services
Languages
CNN THE WORLD'S NEWS LEADER Be the first to know.
Make this my homepage | View U.S. Edition
MORE REPORTS
MOST POPULAR
ASIA BUSINESS
EUROPE BUSINESS
WORLD NEWS
SCIENCE & SPACE NEWS
WORLD NEWS

November 12, 2003 -- Updated 0138 GMT (0938 HKT)

IAEA: Iran had secret nuke agenda

The International Atomic Energy Agency has concluded that Iran has secretly produced small amounts of nuclear materials, including low-enriched uranium and plutonium that could be used to develop nuclear weapons, according to a confidential report obtained by CNN.

FULL STORY

- Snap inspections allowed
- Gallery: Iran's nuclear facilities
- Interactive: How uranium is enriched

EXPLOSIONS ROCK BAGHDAD

Mortars strike the heavily fortified site of the coalition HQ in Iraq. Full Story | Video | Coalition casualties | Bush hails sacrifice

MORE TOP STORIES

- Al Qaeda strategy shift: Experts | London 'target'
- Saudi bomb suspects questioned | Video
- Tension ahead of Bush's UK visit | Poll criticizes president
- Millionaire not guilty of murder | Video
- Berlusconi heads for soccer clash
- Move to expel anti-Semitic slur MP
- Japan leads Asian recovery | Small losses on Wall St.
- Vietnam uncovers 7th century ruins
- Rock star Van has to pay the man

World News | Asia News | Europe News

WORLD BUSINESS

STOCK/FUND QUOTES:
choose exchange: London
enter symbol: GET

MARKETS: updated 0140 GMT

NIKKEI	↑	+76	10283	+0.8%
H.SENG	↓	-153	12003	-1.3%
FTSE	↑	+3	4345	+0.1%
DAX	↓	-16	3729	-0.4%

ROYAL SPOOF

Dutch PM 'not amused'

The House of Orange is at the center of a satirical storm

HIGH ANXIETY

Cultures clash in space

U.S. and Russia search for middle ground on safety

EYE ON CHINA

Getting tough over Taiwan

China set to issue a tough warning to Taiwan

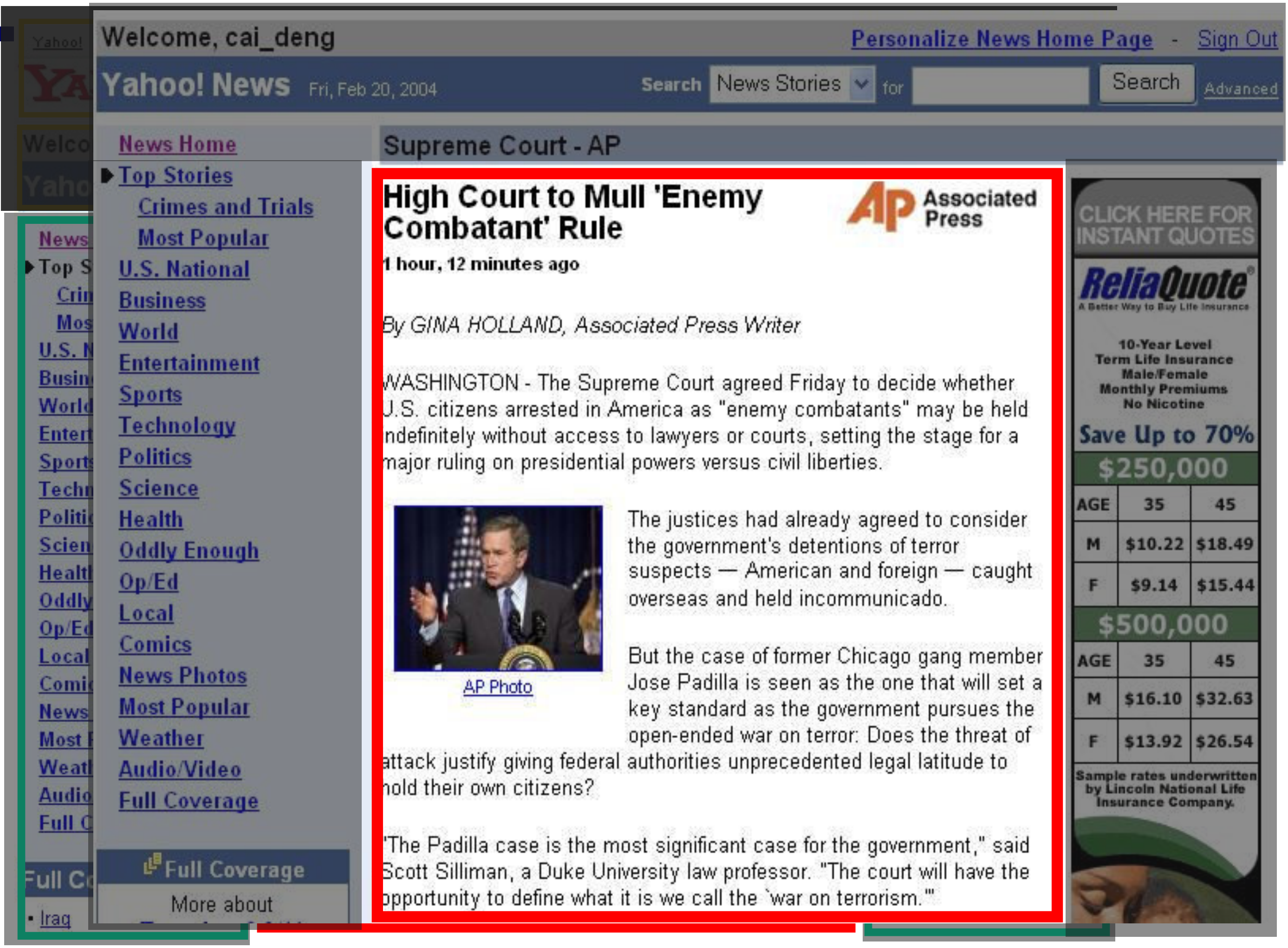
Web Page Blocks

Importance = Low

Importance = Med

Importance = High

A Sample of User Browsing Behavior



Welcome, cai_deng

Personalize News Home Page - Sign Out

Yahoo! News Fri, Feb 20, 2004

News Home
- Top Stories
  - Crimes and Trials
  - Most Popular
- U.S. National
- Business
- World
- Entertainment
- Sports
- Technology
- Politics
- Science
- Health
- Oddly Enough
- Op/Ed
- Local
- Comics
- News Photos
- Most Popular
- Weather
- Audio/Video
- Full Coverage

Full Coverage

More about

Supreme Court - AP

High Court to Mull 'Enemy Combatant' Rule

Associated Press

1 hour, 12 minutes ago

By GINA HOLLAND, Associated Press Writer

WASHINGTON - The Supreme Court agreed Friday to decide whether U.S. citizens arrested in America as "enemy combatants" may be held indefinitely without access to lawyers or courts, setting the stage for a major ruling on presidential powers versus civil liberties.

AP Photo

The justices had already agreed to consider the government's detentions of terror suspects — American and foreign — caught overseas and held incommunicado.

But the case of former Chicago gang member Jose Padilla is seen as the one that will set a key standard as the government pursues the open-ended war on terror: Does the threat of attack justify giving federal authorities unprecedented legal latitude to hold their own citizens?

'The Padilla case is the most significant case for the government," said Scott Silliman, a Duke University law professor. "The court will have the opportunity to define what it is we call the `war on terrorism.'"

Improving PageRank using Layout Structure

- ***Z***: **block-to-page matrix (link structure)**

$$Z_{bp} = \begin{cases} 1/s_b & \text{if there is a link from the } b^{th} \text{ block to the } p^{th} \text{ page} \\ 0 & \text{otherwise} \end{cases}$$

- ***X***: **page-to-block matrix (layout structure)**

$$X_{pb} = \begin{cases} f_p(b) & \text{if the } b^{th} \text{ block is in the } p^{th} \text{ page} \\ 0 & \text{otherwise} \end{cases}$$

$$f \text{ is the block importance function}$$

- **Block-level PageRank**: $W_P = XZ$
  - Compute PageRank on the page-to-page graph
- **BlockRank**: $W_B = ZX$
  - Compute PageRank on the block-to-block graph

Mining Web Images Using Layout & Link Structure (ACMMM'04)

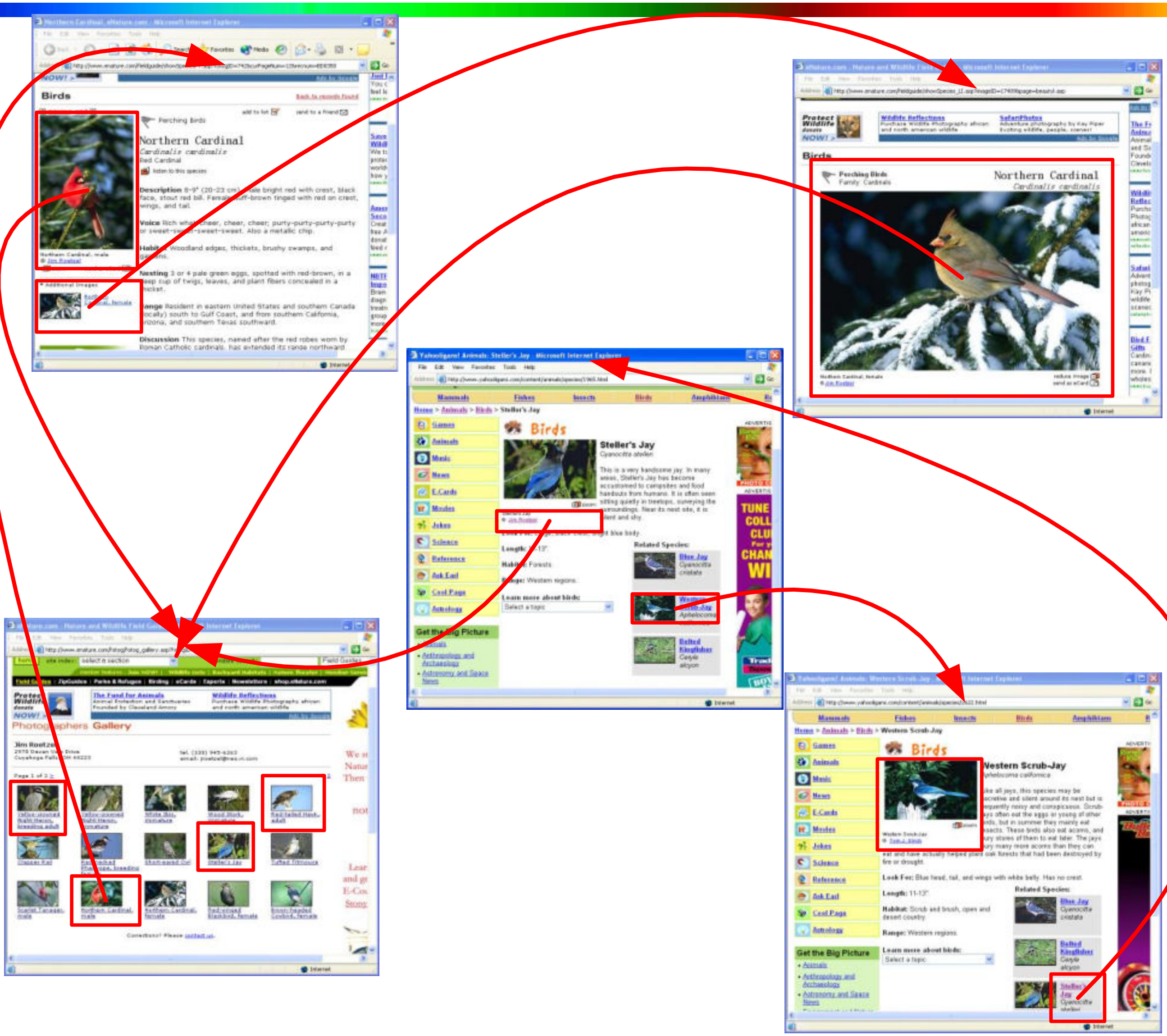


Image Graph Model & Spectral Analysis

- **Block-to-block graph**: $W_B = ZX$
- **Block-to-image matrix (container relation):** Y

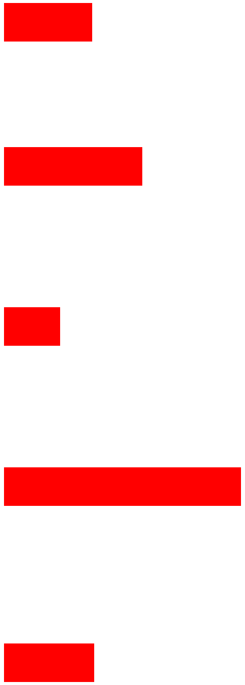
$$Y_{ij} = \begin{cases} 1/s_i & if\, I_j \in b_i \\ 0 & otherwise \end{cases}$$

- **Image-to-image graph:** $W_I = Y^T W_B Y$
- **ImageRank**
  - Compute PageRank on the image graph
- **Image clustering**
  - Graphical partitioning on the image graph

ImageRank

- Relevance Ranking

- Importance Ranking



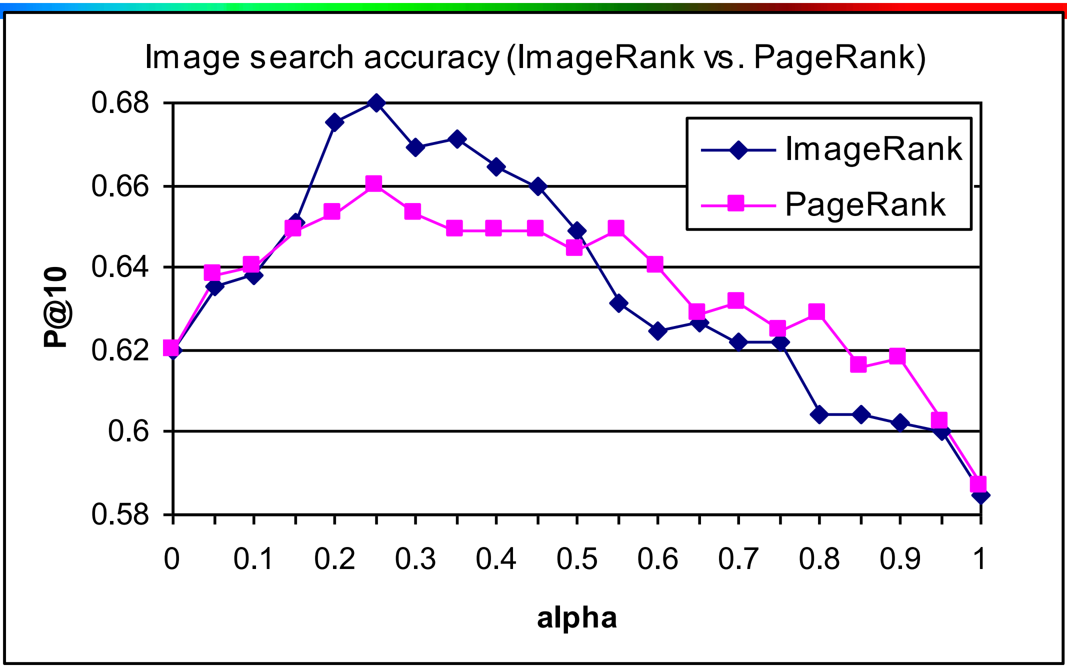
- Combined Ranking

ImageRank vs. PageRank

- Dataset
  - 26.5 millions web pages
  - 11.6 millions images
- Query set
  - 45 hot queries in Google image search statistics
- Ground truth
  - Five volunteers were chosen to evaluate the top 100 results re-turned by the system (iFind)
- Ranking method

$$s(\mathbf{x}) = \alpha \cdot rank_{importance}(\mathbf{x}) + (1-\alpha) \cdot rank_{relevance}(\mathbf{x})$$

ImageRank vs PageRank



- **Image search accuracy using ImageRank and PageRank. Both of them achieved their best results at α=0.25.**

Example on Image Clustering & Embedding

1710 JPG images in 1287 pages are crawled within the website
http://www.yahooligans.com/content/animals/

Six Categories

Mammal

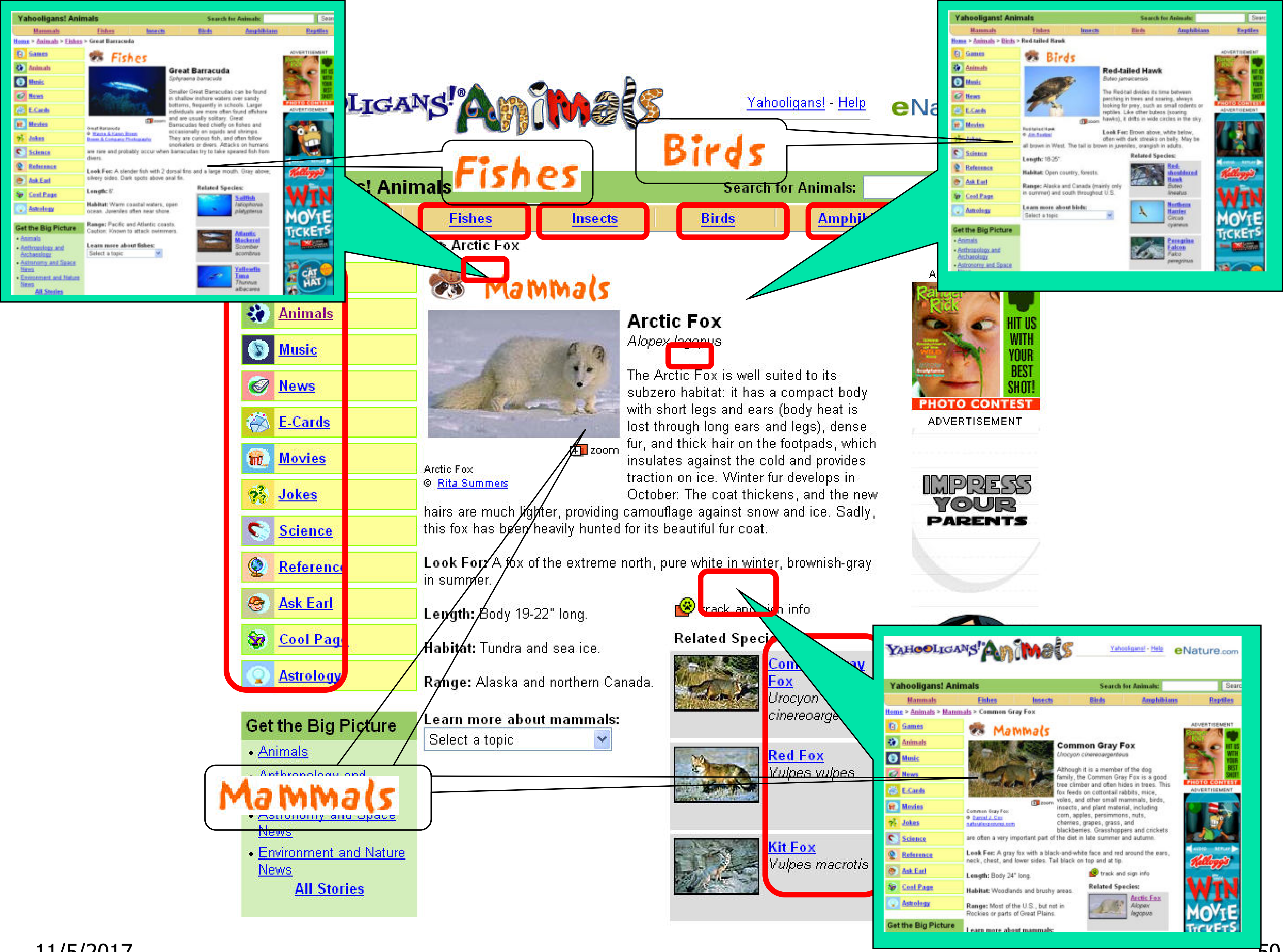
Fish

Reptile

Bird

Amphibian

Insect



LIGANS! Animals

Yahooligans! - Help

! Animals

Search for Animals:

Fishes | Insects | Birds | Amphib

Arctic Fox

Mammals

Animals

Music

News

E-Cards

Movies

Jokes

Science

Referenc

Ask Earl

Cool Pag

Astrology

Get the Big Picture

- Animals
- Anthropology and
- Astronomy and Space News
- Environment and Nature News

All Stories

Arctic Fox

Alopex lagopus

zoom

Arctic Fox
© Rita Summers

The Arctic Fox is well suited to its subzero habitat: it has a compact body with short legs and ears (body heat is lost through long ears and legs), dense fur, and thick hair on the footpads, which insulates against the cold and provides traction on ice. Winter fur develops in October: The coat thickens, and the new hairs are much lighter, providing camouflage against snow and ice. Sadly, this fox has been heavily hunted for its beautiful fur coat.

Look For: A fox of the extreme north, pure white in winter, brownish-gray in summer.

Length: Body 19-22" long.

Habitat: Tundra and sea ice.

Range: Alaska and northern Canada.

Learn more about mammals:

Select a topic

track and sign info

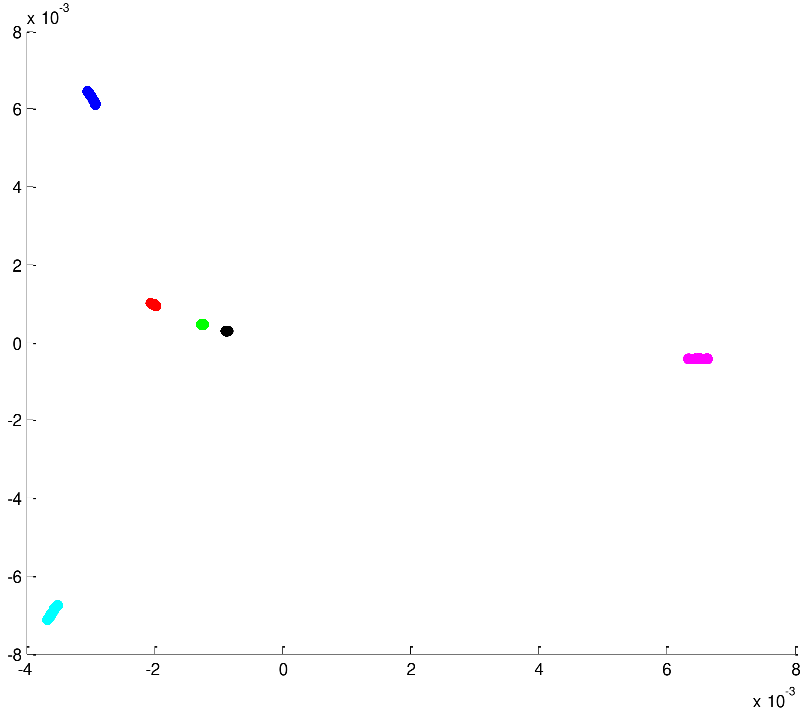
Related Species:

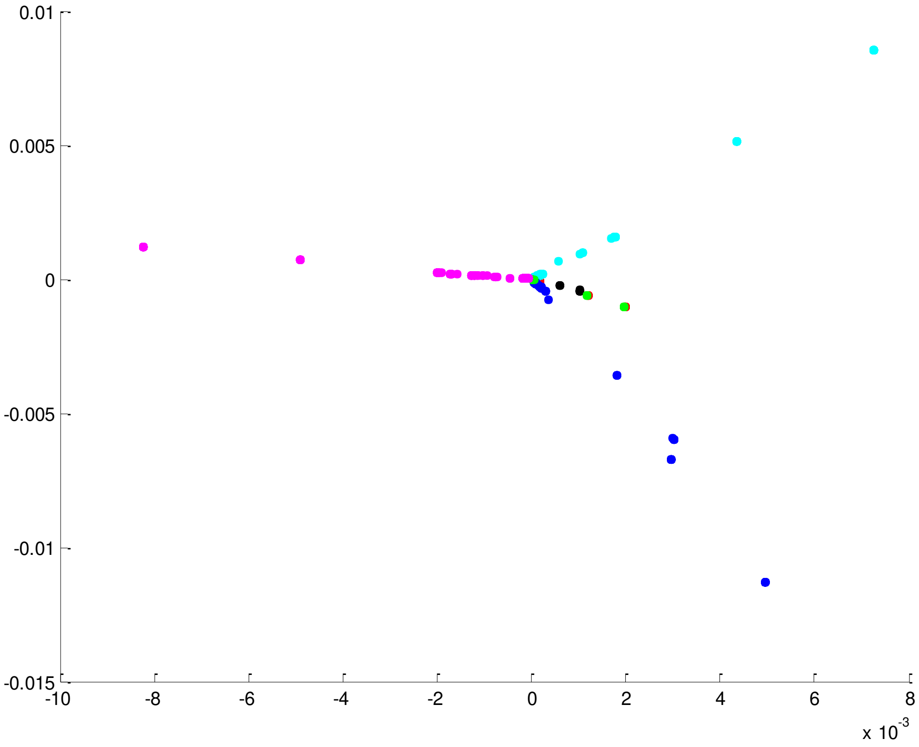
Common Gray Fox
Urocyon cinereoargenteus

Red Fox
Vulpes vulpes

Kit Fox
Vulpes macrotis

2-D embedding of WWW images

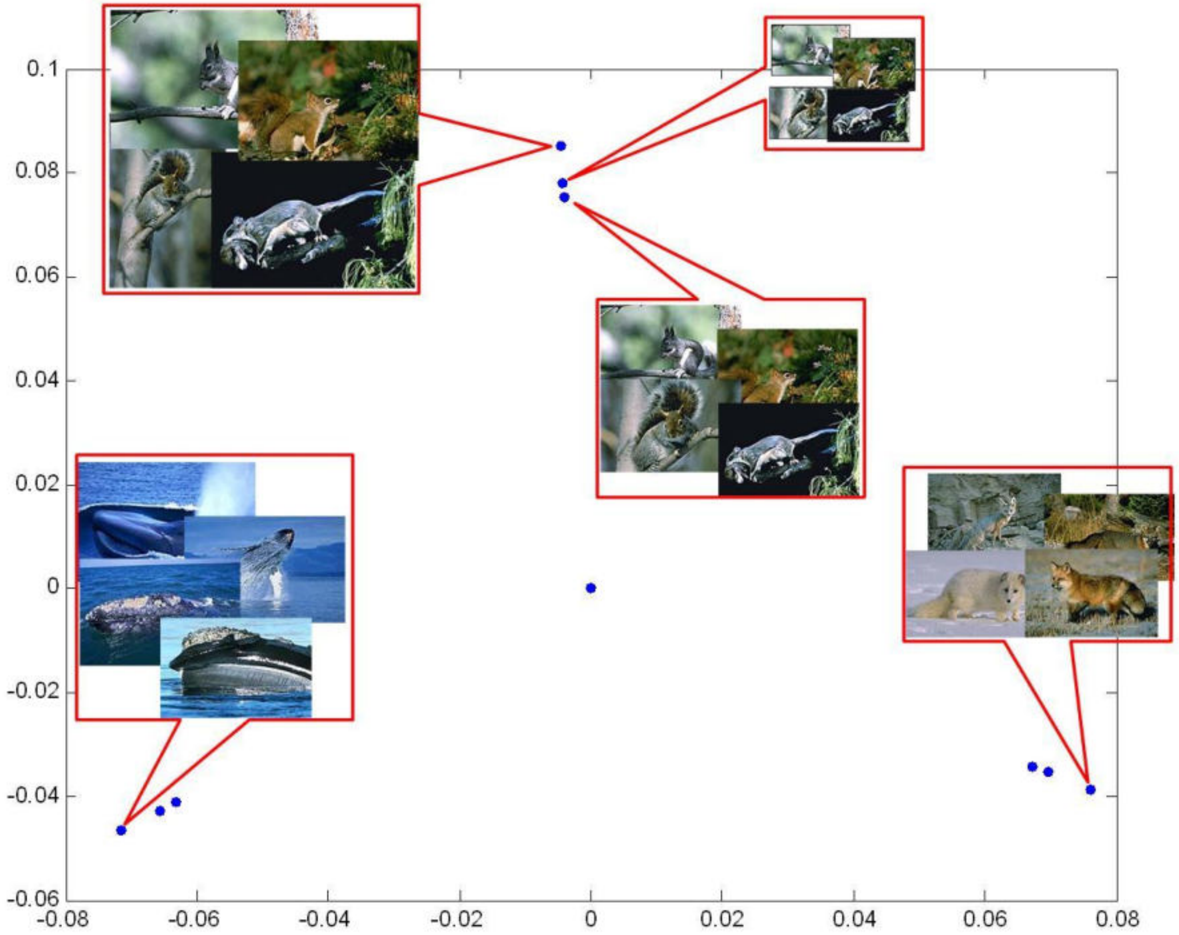




The image graph was constructed from block level link analysis

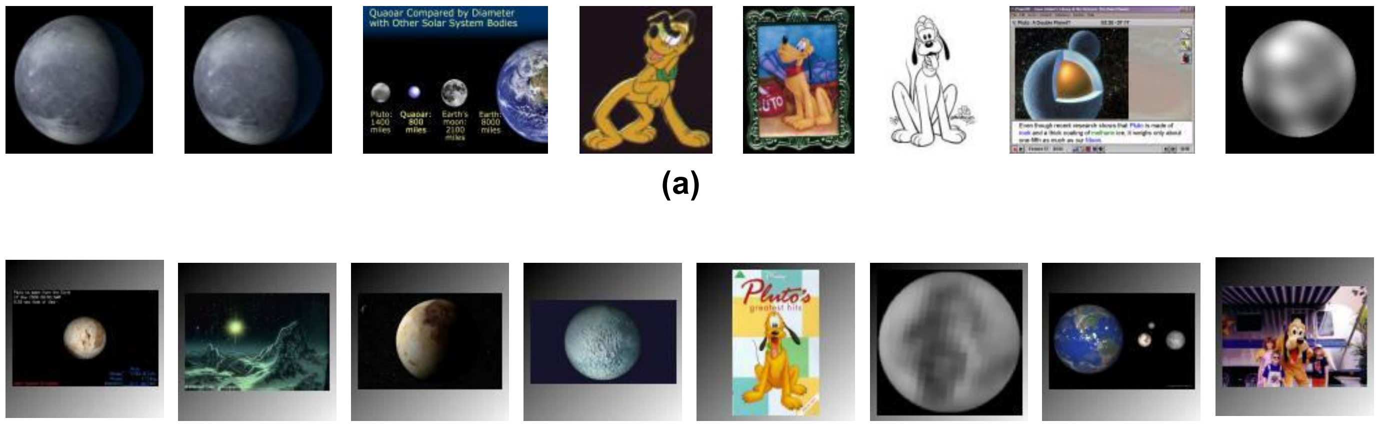
The image graph was constructed from traditional page level link analysis

2-D Embedding of Web Images



- 2-D visualization of the mammal category using the second and third eigenvectors.

Web Image Search Result Presentation



(a)

(b)

Figure 1. Top 8 returns of query "pluto" in Google's image search engine (a) and AltaVista's image search engine (b)

- Two different topics in the search result
- A possible solution:
  - Cluster search results into different semantic groups

Three kinds of WWW image representation

- Visual Feature Based Representation
  - Traditional CBIR
- Textual Feature Based Representation
  - Surrounding text in image block
- Link Graph Based Representation
  - Image graph embedding

Clustering Using Visual Feature

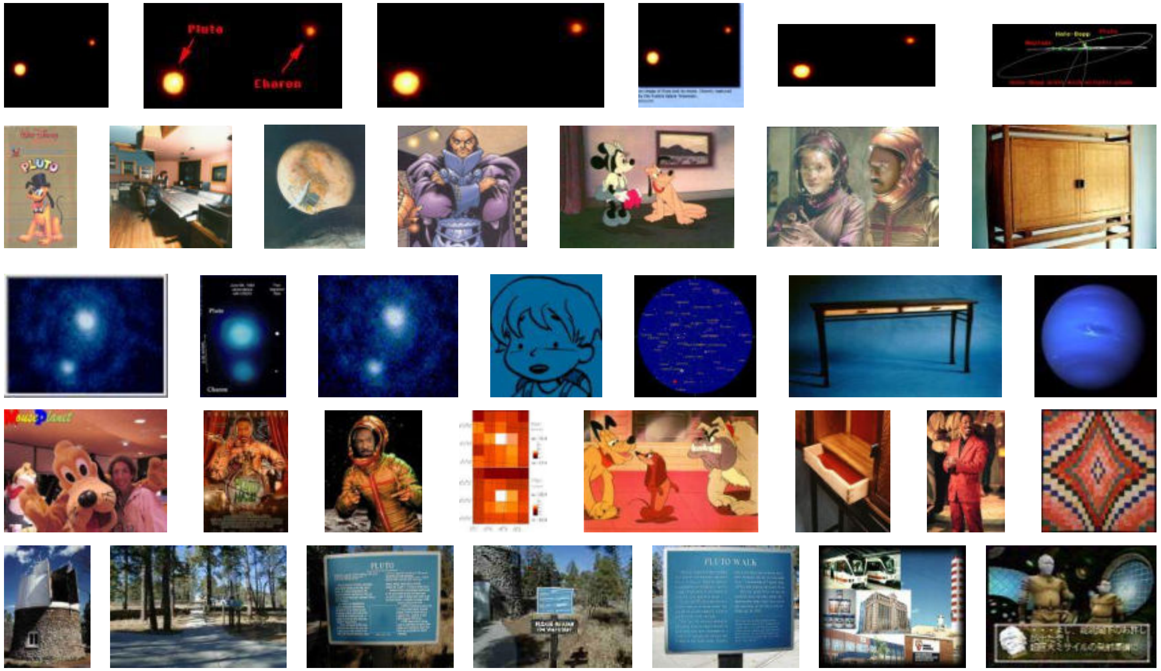


Figure 5. Five clusters of search results of query "pluto" using low level visual feature. Each row is a cluster.

- From the perspectives of color and texture, the clustering results are quite good. Different clusters have different colors and textures. However, from semantic perspective, these clusters make little sense.

Clustering Using Textual Feature

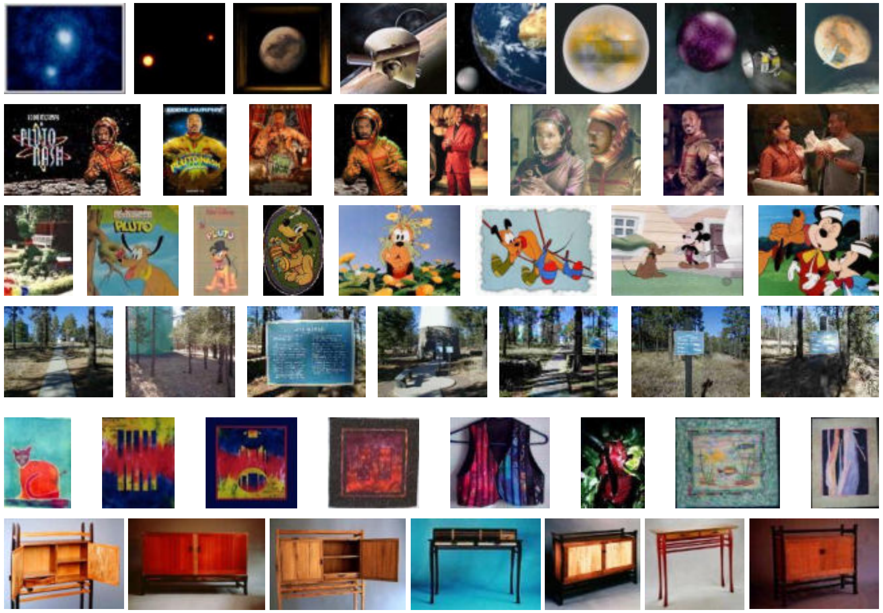


Figure 7. Six clusters of search results of query “pluto” using textual feature. Each row is a cluster

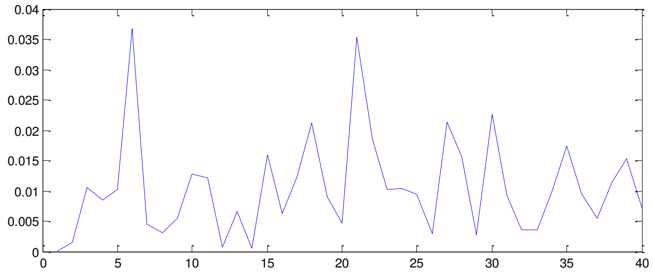


Figure 6. The Eigengap curve with *k* for the “pluto” case using textual representation

- Six semantic categories are correctly identified if we choose *k* = 6.

Clustering Using Graph Based Representation

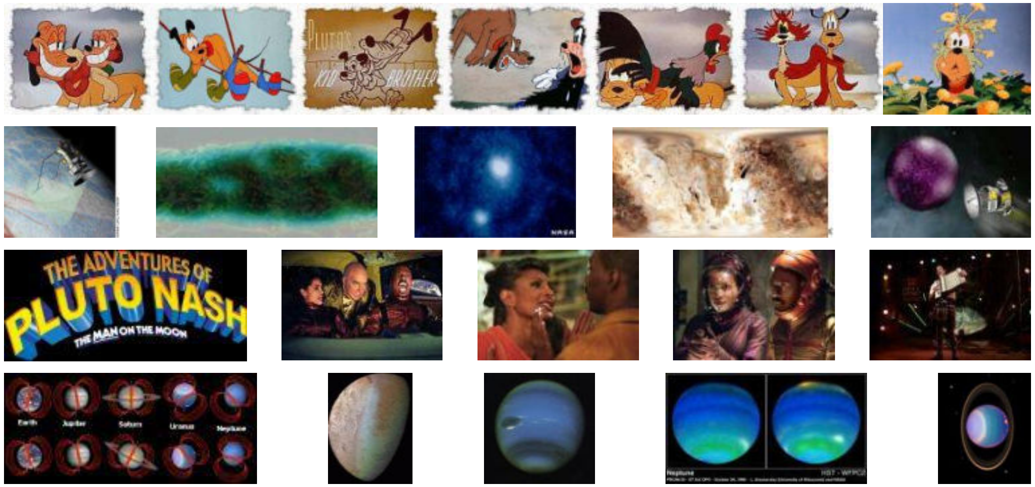


Figure 8. Five clusters of search results of query "pluto" using image link graph. Each row is a cluster

- Each cluster is semantically aggregated.
- Too many clusters.
- In "pluto" case, the top 500 results are clustered into 167 clusters. The max cluster number is 87, and there are 112 clusters with only one image.

Combining Textual Feature and Link Graph

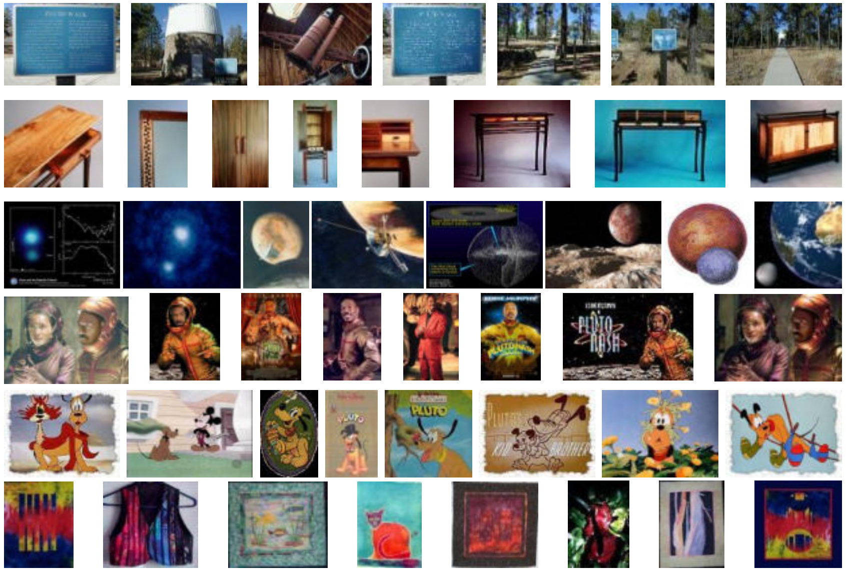


Figure 9. Six clusters of search results of query "pluto" using combination of textual feature and image link graph. Each row is a cluster

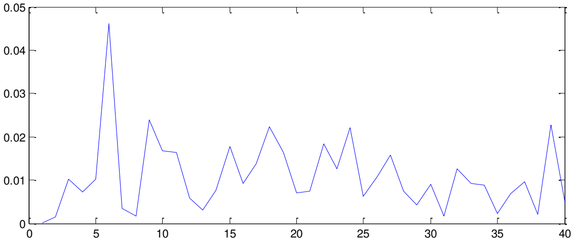
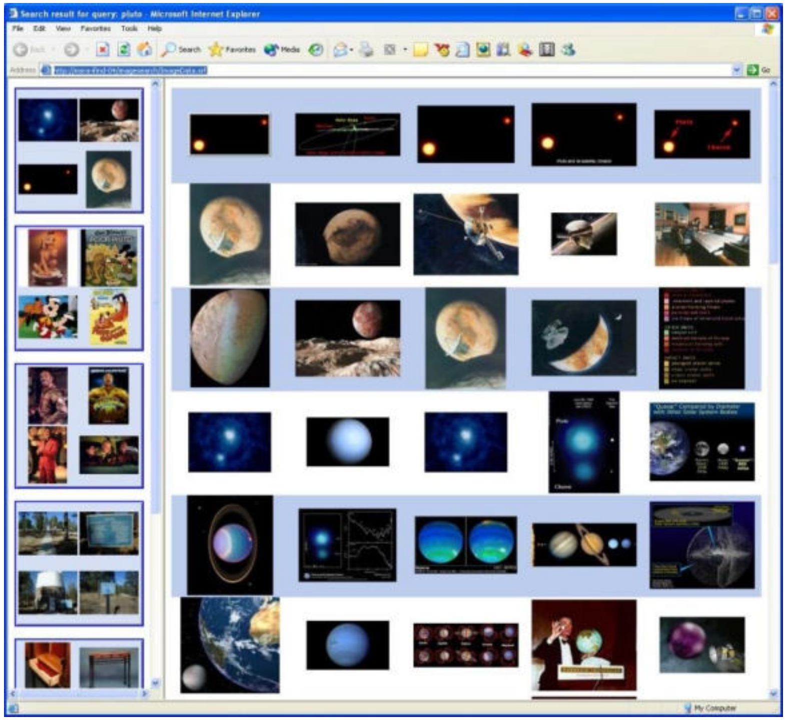


Figure 10. The Eigengap curve with *k* for the "pluto" case using textual and link combination

- Combine two affinity matrix

$$S_{combine}(i,j)=\begin{cases} S_{textual}(i,j) & if\ S_{link}(i,j)=0 \\ 1 & if\ S_{link}(i,j)>0 \end{cases}$$

Final Presentation of Our System


Search result for query: pluto - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
Go
My Computer

- Using textual and link information to get some semantic clusters
- Use low level visual feature to cluster (re-organize) each semantic cluster to facilitate user’s browsing

Summary

- More improvement on web search can be made by mining webpage Layout structure
- Leverage visual cues for web information analysis & information extraction
- Demos:
  - http://www.ews.uiuc.edu/~dengcai2
    - Papers
    - VIPS demo & dll