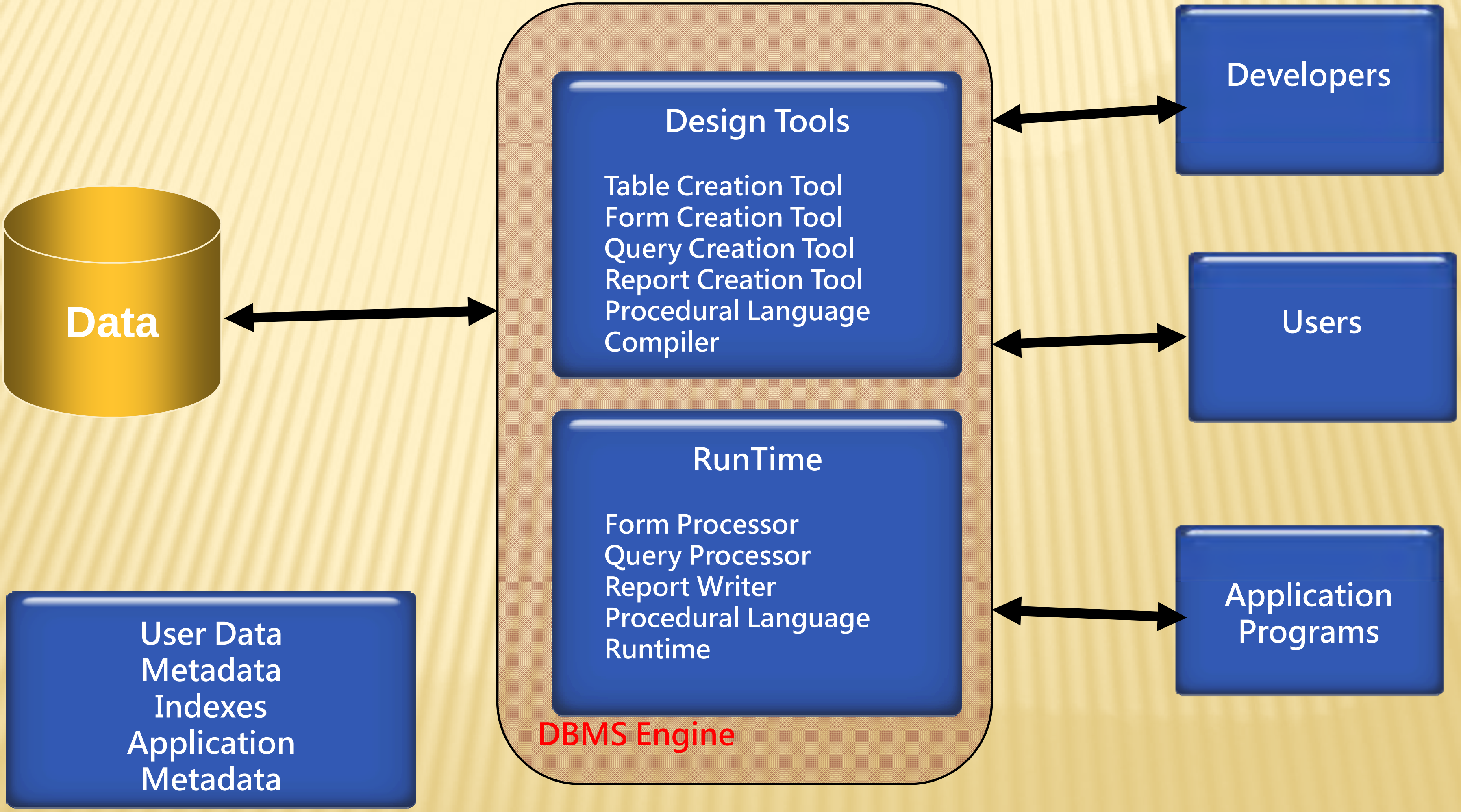
DATABASE DEVELOPMENTS

Tools, Applications and Trends – July 2011

AGENDA

- CASE Tools
- Data Mining Methods and Applications
- Recent Database Technology Development

A TYPICAL DATABASE SYSTEM



Data

Design Tools

Table Creation Tool
Form Creation Tool
Query Creation Tool
Report Creation Tool
Procedural Language
Compiler

RunTime

Form Processor
Query Processor
Report Writer
Procedural Language
Runtime

DBMS Engine

Developers

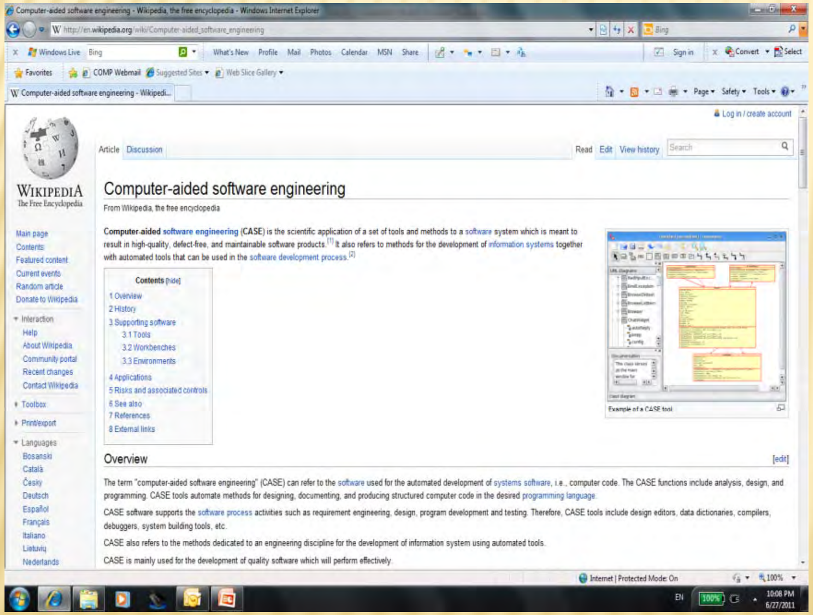
Users

Application Programs

User Data
Metadata
Indexes
Application
Metadata

WHAT IS CASE

- Computer-aided software engineering (CASE) is the scientific application of a set of tools and methods to a software system which is meant to result in high-quality, defect-free, and maintainable software products.



Computer-aided software engineering

From Wikipedia, the free encyclopedia

Computer-aided software engineering (**CASE**) is the scientific application of a set of tools and methods to a software system which is meant to result in high-quality, defect-free, and maintainable software products.[1] It also refers to methods for the development of information systems together with automated tools that can be used in the software development process.[2]

Contents [hide]
1 Overview
2 History
3 Supporting software
3.1 Tools
3.2 Workbenches
3.3 Environments
4 Applications
5 Risks and associated controls
6 See also
7 References
8 External links

Example of a CASE tool

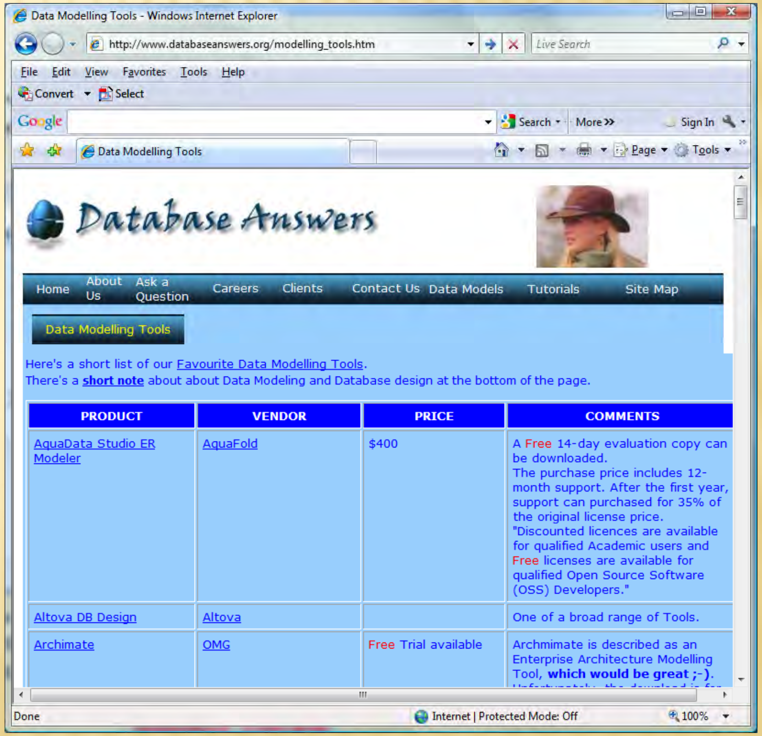
Overview [edit]

The term "computer-aided software engineering" (CASE) can refer to the software used for the automated development of systems software, i.e., computer code. The CASE functions include analysis, design, and programming. CASE tools automate methods for designing, documenting, and producing structured computer code in the desired programming language.

CASE software supports the software process activities such as requirement engineering, design, program development and testing. Therefore, CASE tools include design editors, data dictionaries, compilers, debuggers, system building tools, etc.

CASE also refers to the methods dedicated to an engineering discipline for the development of information system using automated tools.

CASE is mainly used for the development of quality software which will perform effectively.



Data Modelling Tools

Here's a short list of our Favourite Data Modelling Tools.
There's a **short note** about about Data Modeling and Database design at the bottom of the page.

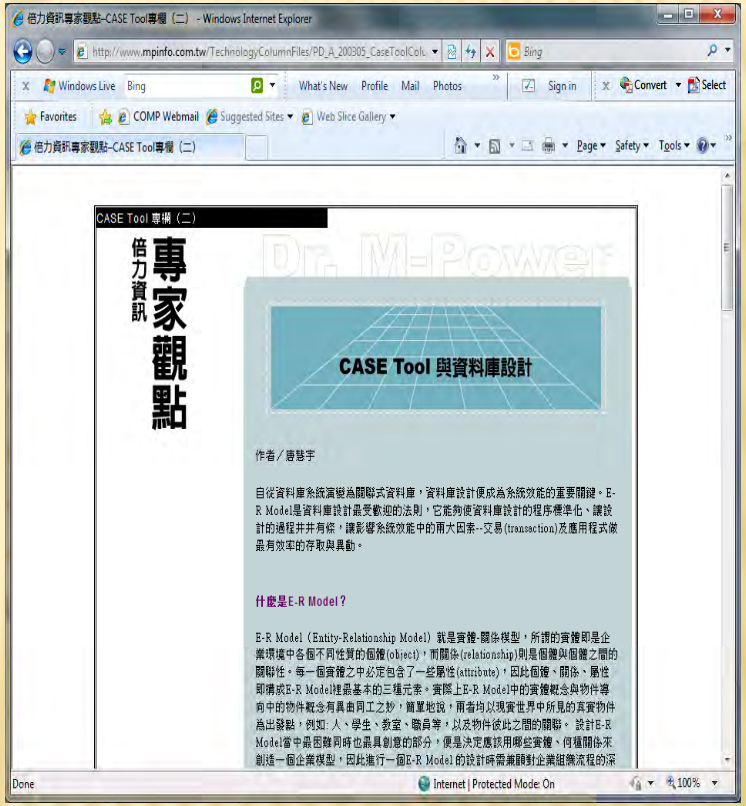
PRODUCT	VENDOR	PRICE	COMMENTS
AquaData Studio ER Modeler	AquaFold	$400	A Free 14-day evaluation copy can be downloaded. The purchase price includes 12-month support. After the first year, support can purchased for 35% of the original license price. "Discounted licences are available for qualified Academic users and Free licenses are available for qualified Open Source Software (OSS) Developers."
Altova DB Design	Altova		One of a broad range of Tools.
Archimate	OMG	Free Trial available	Archmimate is described as an Enterprise Architecture Modelling Tool, **which would be great ;-)**.

http://en.wikipedia.org/wiki/Computer-aided_software_engineering

http://www.databaseanswers.org/modelling_tools.htm

DATABASE CASE TOOLS

- Types
  - Data modeling
  - Form creation
  - Query builder
  - Report creation
- Availability
  - Free or Almost Free
  - Commercial


倍力資訊專家觀點-CASE Tool專欄（二） - Windows Internet Explorer
http://www.mpinfo.com.tw/TechnologyColumnFiles/PD_A_200305_CaseToolColu
Bing
Windows Live
What's New
Profile
Mail
Photos
Sign in
Convert
Select
Favorites
COMP Webmail
Suggested Sites
Web Slice Gallery
倍力資訊專家觀點-CASE Tool專欄（二）
Page
Safety
Tools
Done
Internet | Protected Mode: On
100%

CASE Tool 專欄（二）

倍力資訊 專家觀點

Dr. M-Power

CASE Tool 與資料庫設計

作者／唐慧宇

自從資料庫系統演變為關聯式資料庫，資料庫設計便成為系統效能的重要關鍵。E-R Model是資料庫設計最受歡迎的法則，它能夠使資料庫設計的程序標準化、讓設計的過程井井有條，讓影響系統效能中的兩大因素--交易(transaction)及應用程式做最有效率的存取與異動。

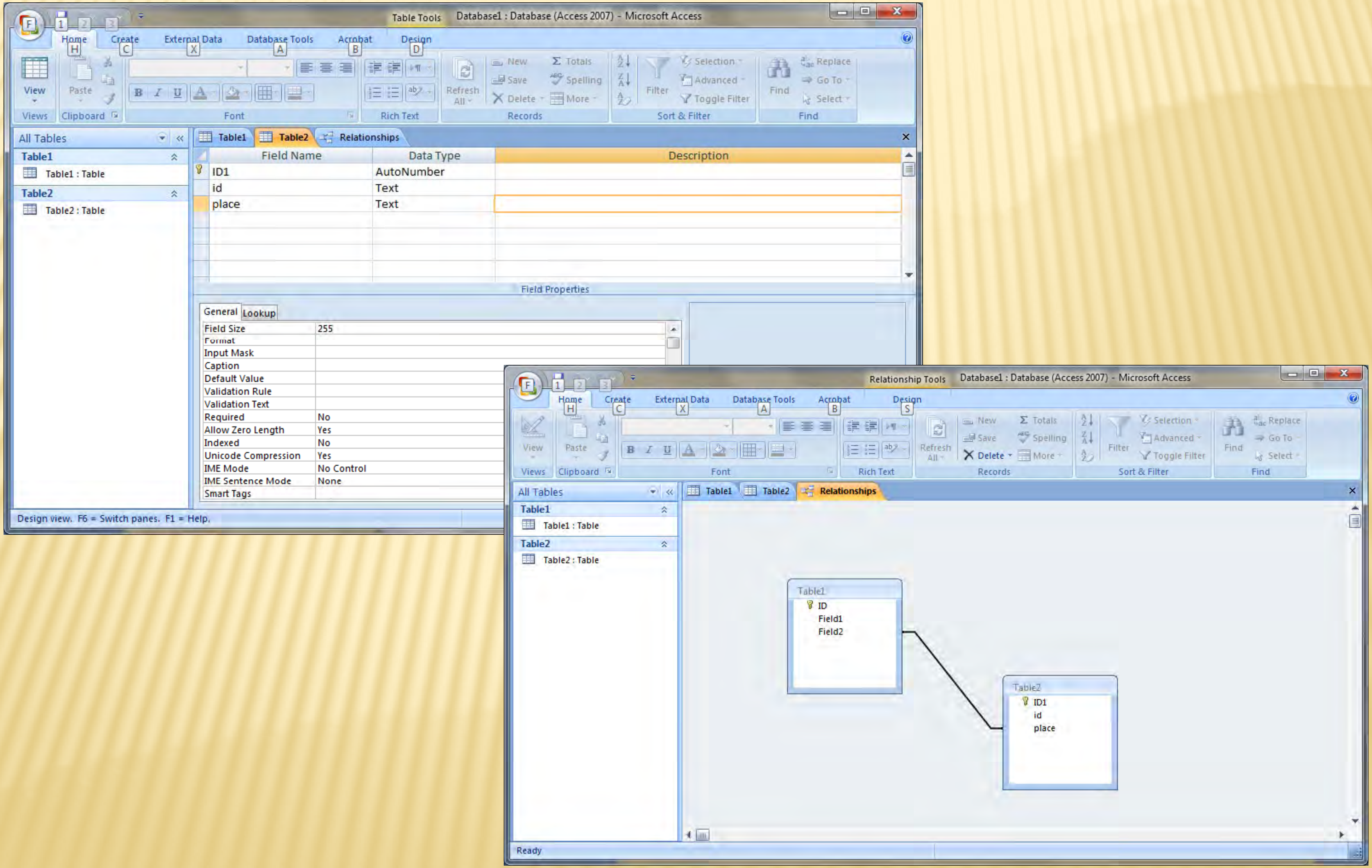
什麼是E-R Model？

E-R Model (Entity-Relationship Model) 就是實體-關係模型，所謂的實體即是企業環境中各個不同性質的個體(object)，而關係(relationship)則是個體與個體之間的關聯性。每一個實體之中必定包含了一些屬性(attribute)，因此個體、關係、屬性即構成E-R Model裡最基本的三種元素。實際上E-R Model中的實體概念與物件導向中的物件概念有異曲同工之妙，簡單地說，兩者均以現實世界中所見的真實物件為出發點，例如: 人、學生、教室、職員等，以及物件彼此之間的關聯。 設計E-R Model當中最困難同時也最具創意的部分，便是決定應該用哪些實體、何種關係來創造一個企業模型，因此進行一個E-R Model 的設計時需兼顧對企業組織流程的深

www.mpinfo.com.tw/TechnologyColumnFiles/PD_A_200305_CaseToolColumn.htm

www.mpinfo.com.tw/TechnologyColumnFiles/PD_A_200307_CaseToolColumn.htm

MS ACCESS TOOLS

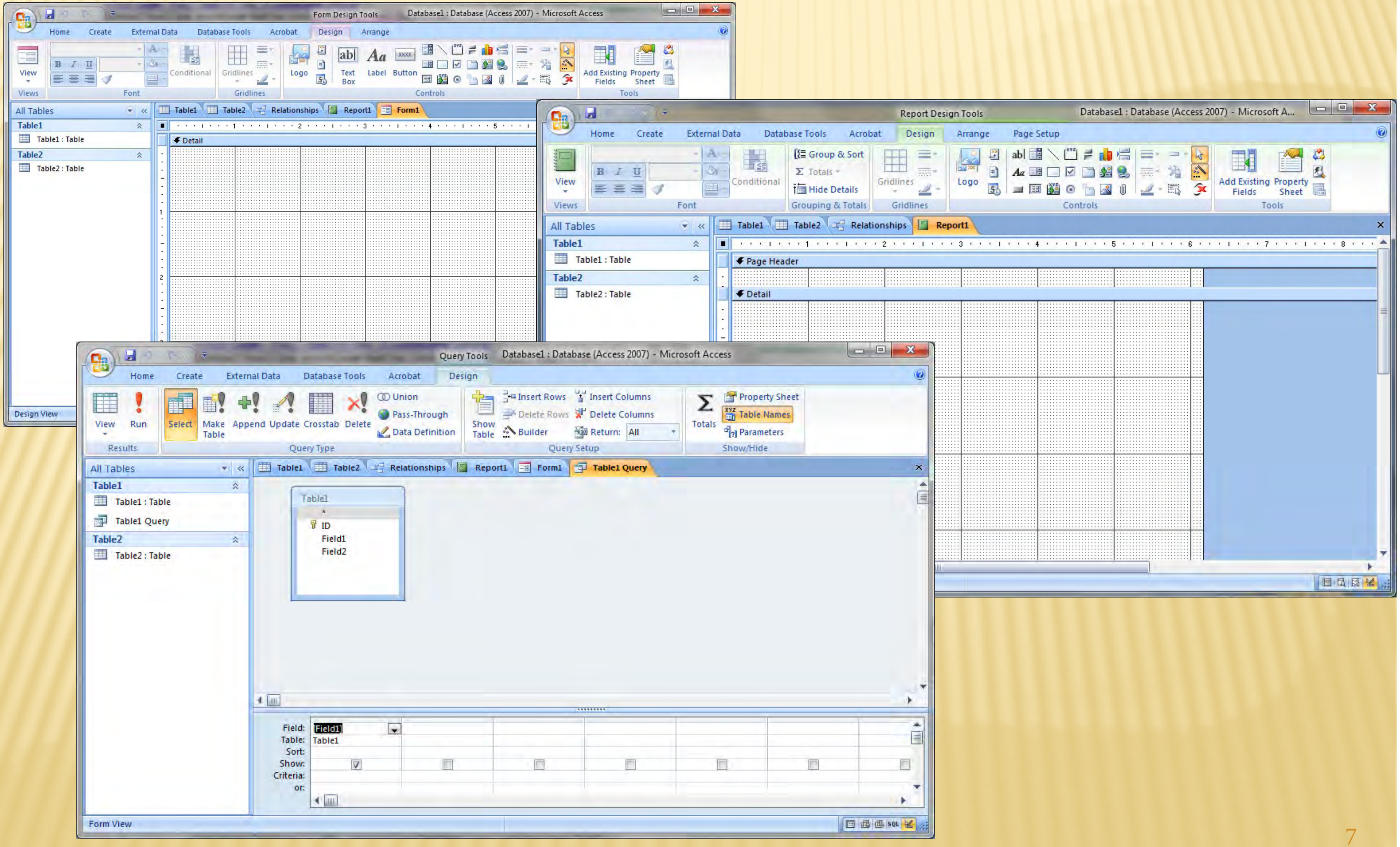


Field Name	Data Type	Description
ID1	AutoNumber	
id	Text	
place	Text	

Field Size	255
Format	
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	No
Allow Zero Length	Yes
Indexed	No
Unicode Compression	Yes
IME Mode	No Control
IME Sentence Mode	None
Smart Tags	

Design view. F6 = Switch panes. F1 = Help.

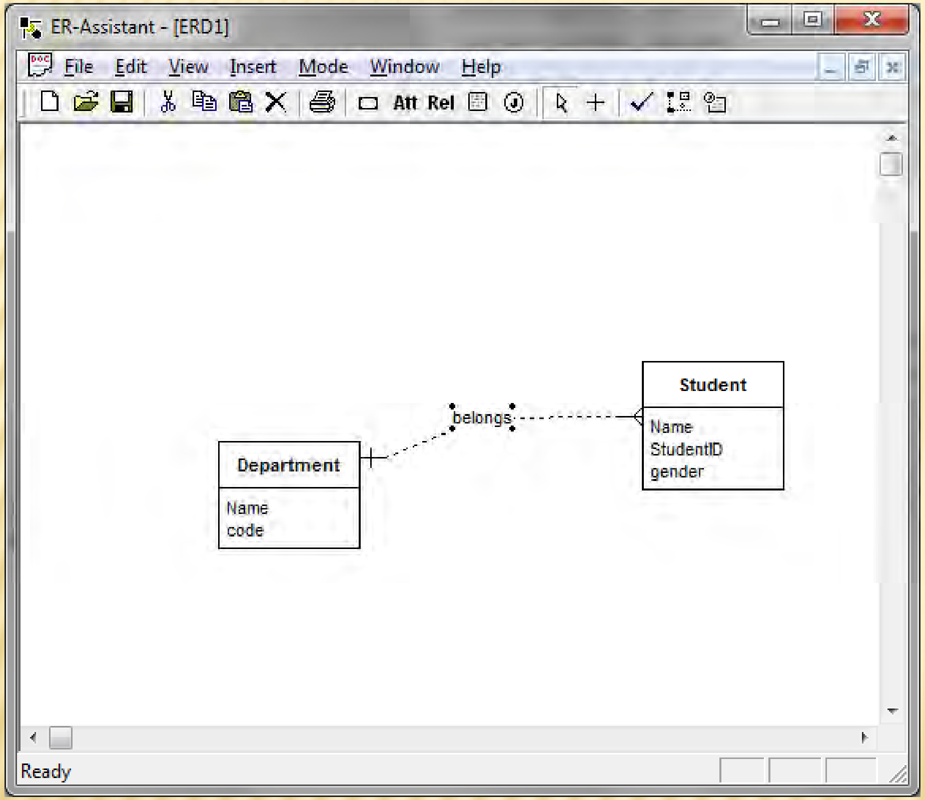
MS ACCESS TOOLS



Field: Field1
Table: Table1
Sort:
Show: ☑ □ □ □ □ □ □
Criteria:
or:

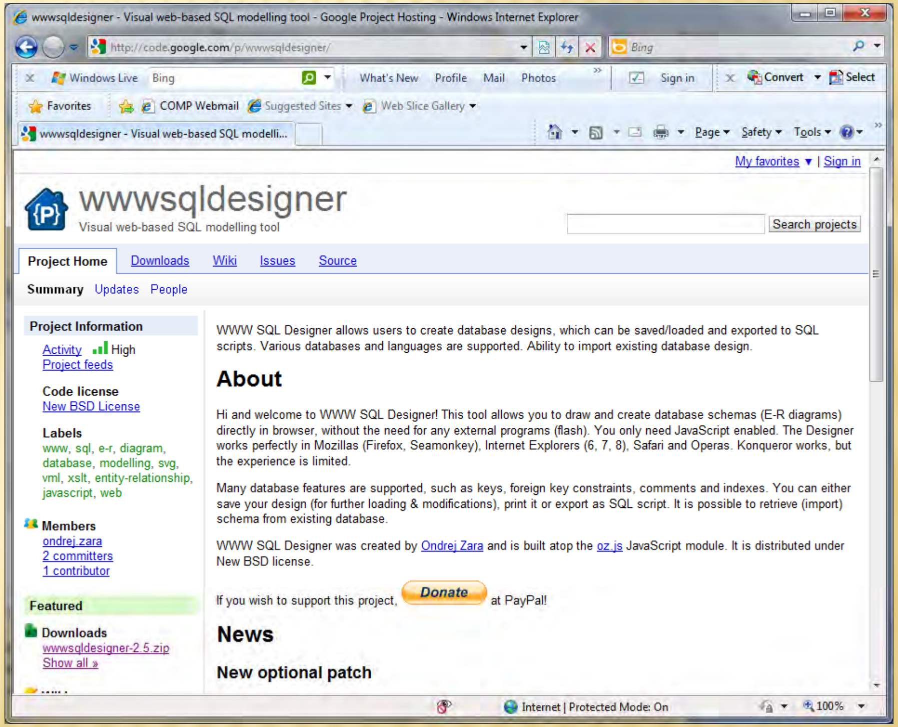
Return: All

DATA MODELING TOOL –ER ASSISTANT



http://highered.mcgraw-hill.com/sites/0072942207/student_view0/e_r_assistant.html

http://code.google.com/p/wwwsqldesigner/


wwwsqldesigner - Visual web-based SQL modelling tool - Google Project Hosting - Windows Internet Explorer
http://code.google.com/p/wwwsqldesigner/
Bing
Windows Live
What's New Profile Mail Photos
Sign in
Convert
Select
Favorites
COMP Webmail
Suggested Sites
Web Slice Gallery
wwwsqldesigner - Visual web-based SQL modelli...
Page Safety Tools
My favorites | Sign in
Search projects
Project Home Downloads Wiki Issues Source
Summary Updates People
Internet | Protected Mode: On
100%

wwwsqldesigner

Visual web-based SQL modelling tool

Project Information

Activity High

Project feeds

Code license

New BSD License

Labels

www, sql, e-r, diagram, database, modelling, svg, vml, xslt, entity-relationship, javascript, web

Members

ondrej.zara

2 committers

1 contributor

Featured

Downloads

wwwsqldesigner-2.5.zip

Show all »

WWW SQL Designer allows users to create database designs, which can be saved/loaded and exported to SQL scripts. Various databases and languages are supported. Ability to import existing database design.

About

Hi and welcome to WWW SQL Designer! This tool allows you to draw and create database schemas (E-R diagrams) directly in browser, without the need for any external programs (flash). You only need JavaScript enabled. The Designer works perfectly in Mozillas (Firefox, Seamonkey), Internet Explorers (6, 7, 8), Safari and Operas. Konqueror works, but the experience is limited.

Many database features are supported, such as keys, foreign key constraints, comments and indexes. You can either save your design (for further loading & modifications), print it or export as SQL script. It is possible to retrieve (import) schema from existing database.

WWW SQL Designer was created by Ondrej Zara and is built atop the oz.js JavaScript module. It is distributed under New BSD license.

If you wish to support this project, Donate at PayPal!

News

New optional patch

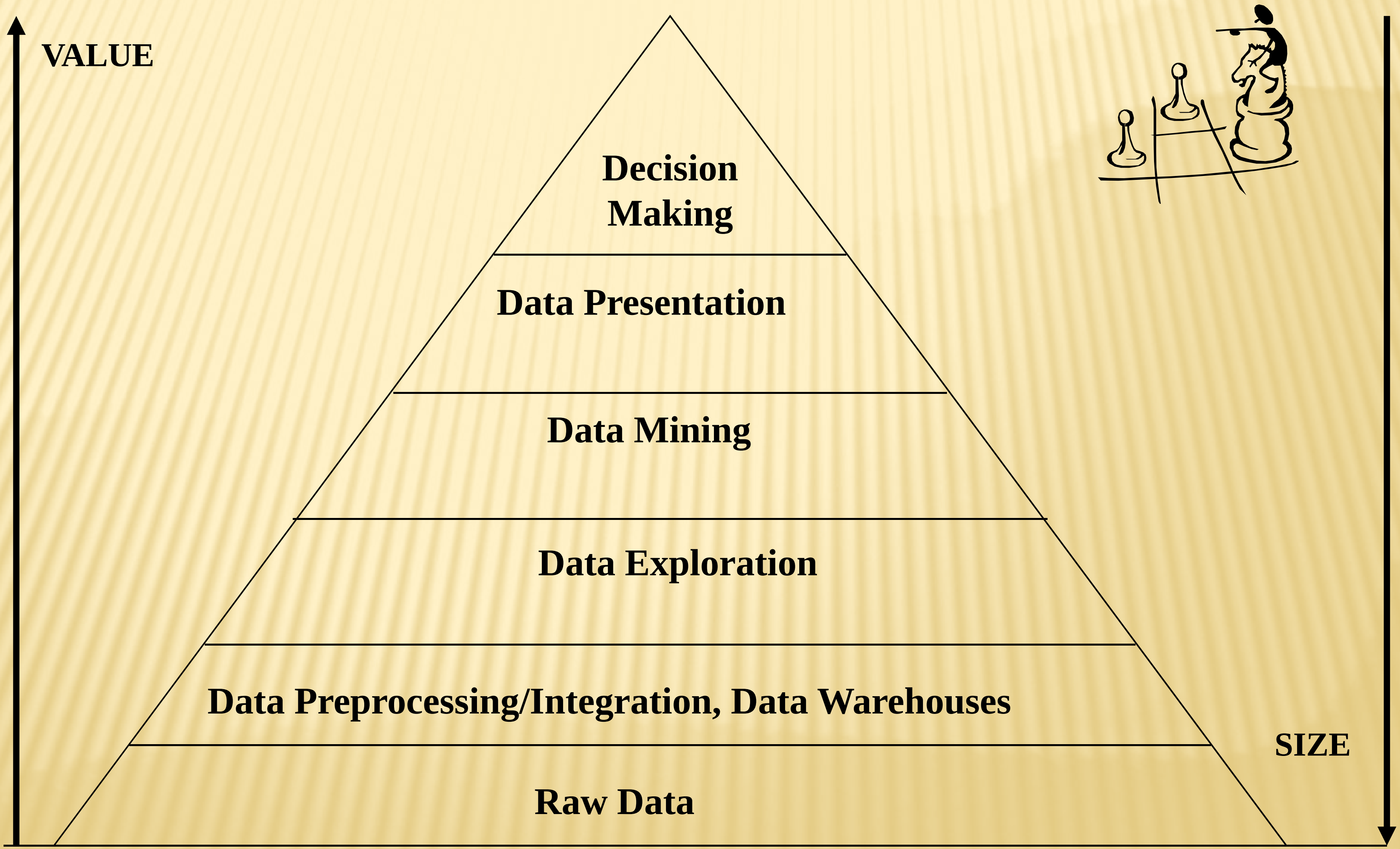
WHY DATA MINING?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated sensors, database systems, Web services
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Social media: FB, Twitter, forums, YouTube
- How to deal with different types of data?
  - Numbers, text, images, audios, videos
- http://www.youtube.com/watch?v=IPboKPWpOVo

WHY NOT TRADITIONAL DATA ANALYSIS?

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

DATA MINING AND BUSINESS INTELLIGENCE



VALUE

Decision Making

Data Presentation

Data Mining

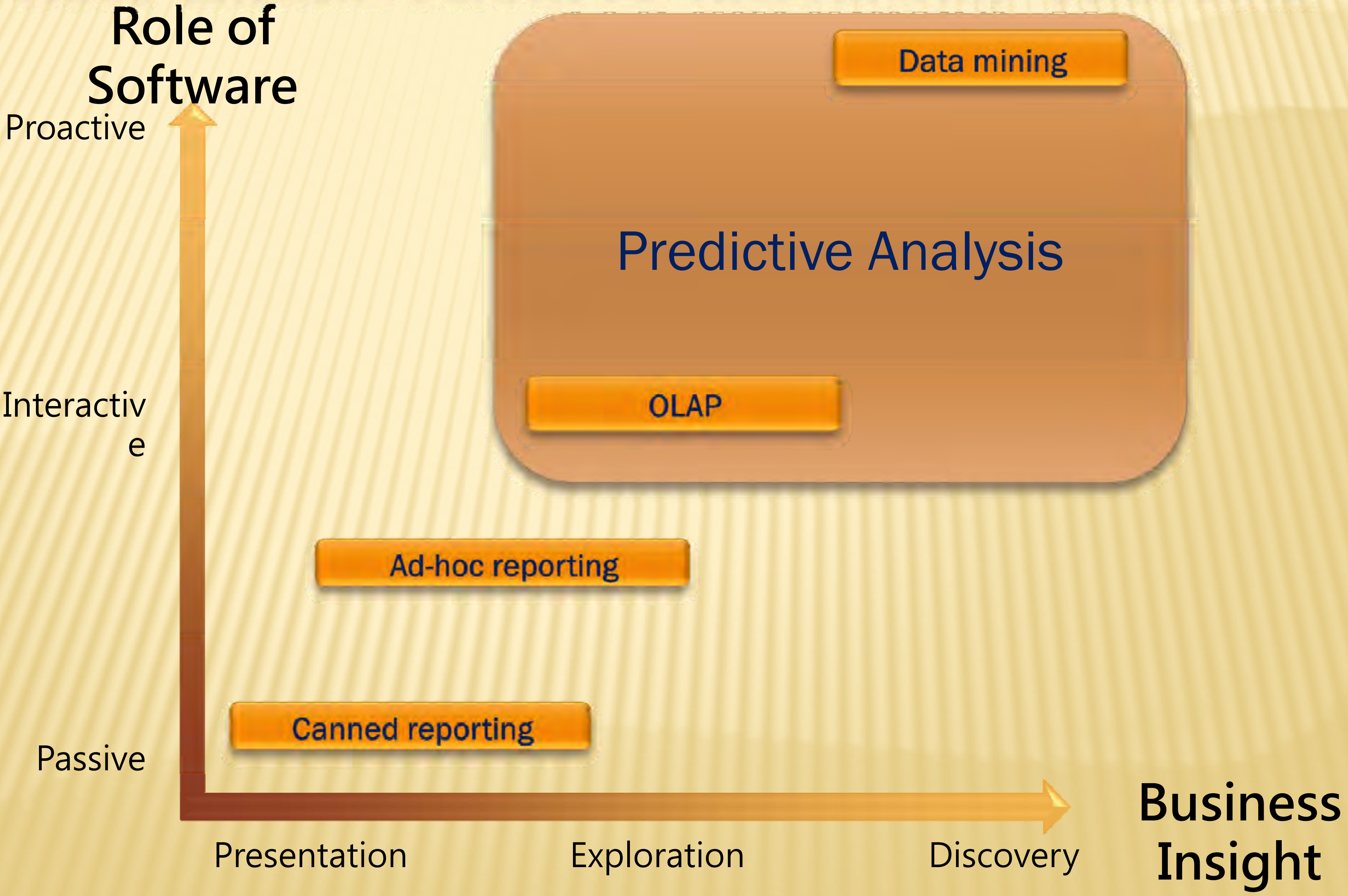
Data Exploration

Data Preprocessing/Integration, Data Warehouses

Raw Data

SIZE

DATA MINING ENABLES PREDICTIVE ANALYSIS



Role of Software

Proactive

Interactiv
e

Passive

Data mining

Predictive Analysis

OLAP

Ad-hoc reporting

Canned reporting

Presentation

Exploration

Discovery

Business Insight

DATA MINING ANALYSIS

- Frequent patterns, association, correlation vs. causality
  - E.g. Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
  - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values
- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis

DATA MINING IN PRACTICES

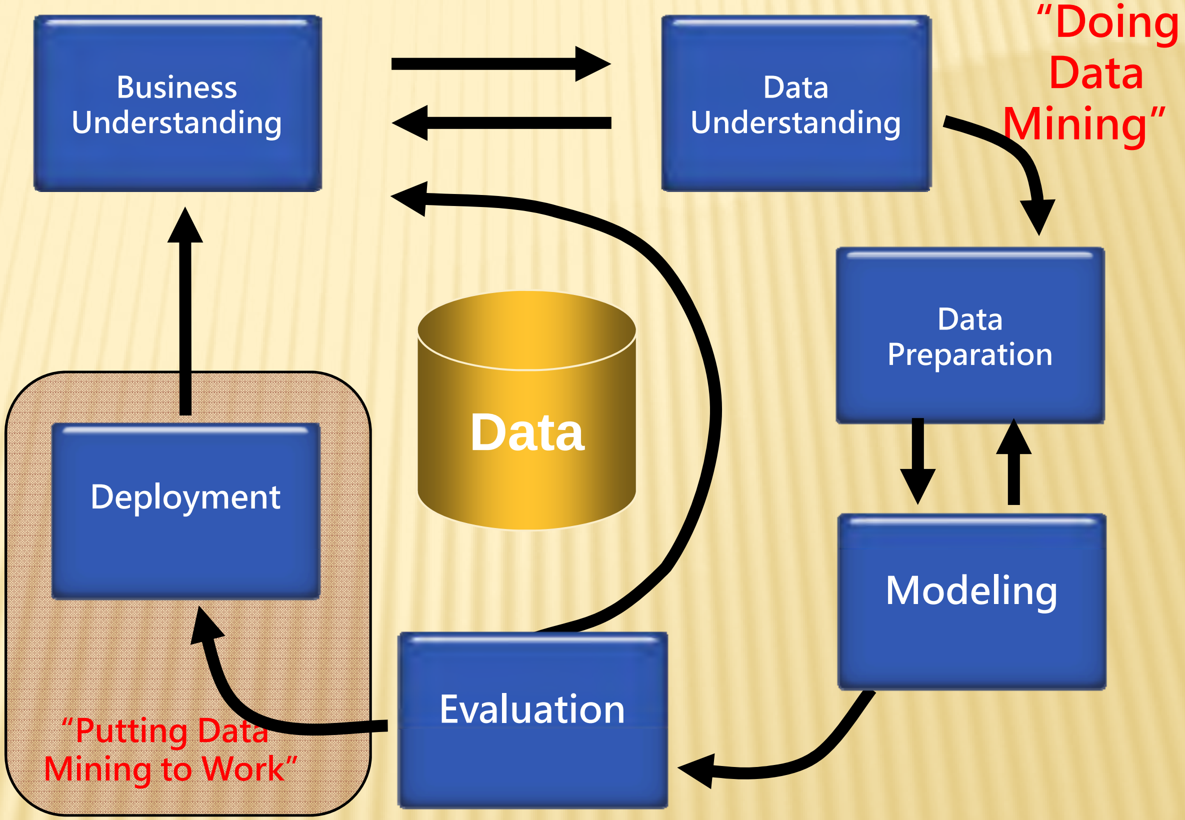
- CRM – buying behaviors
- Banking – personal loan
- Telecommunication – mobile packages
- Maintenance Services – machinery part pre-ordering
- Education – online study patterns
- Social media – sentimental analysis in FB postings

DATA MINING TECHNIQUES

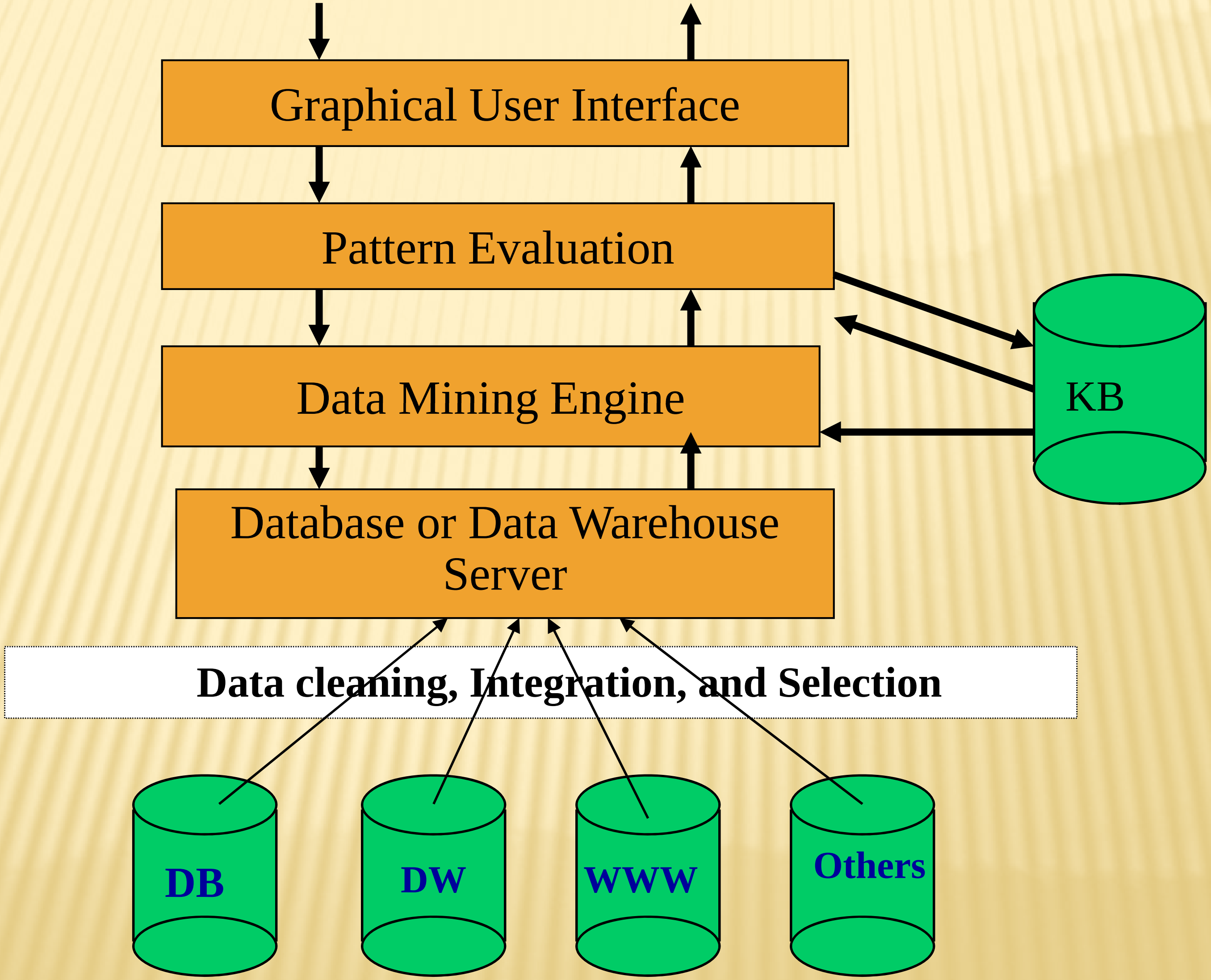
- Neural Networks
- Generalized EM And K-means Cluster Analysis
- General CART Models
- General CHAID Models
- Interactive Trees (C&RT and CHAID)
- Boosted Tree Classifiers and Regression
- Association Rules
- MARSPlines
- Machine Learning(Bayesian, Support Vectors and Nearest neighbors)
- Random Forests for Regression and Classification
- Generalized Additive Models (GAM)
- Feature Selection and Variable Screening

TOP-10 ALGORITHM SELECTED AT ICDM'06

- #1: C4.5
- #2: K-Means
- #3: SVM
- #4: Apriori
- #5: EM
- #6: PageRank
- #7: AdaBoost
- #7: kNN
- #7: Naive Bayes
- #10: CART



A TYPICAL DATA MINING SYSTEM

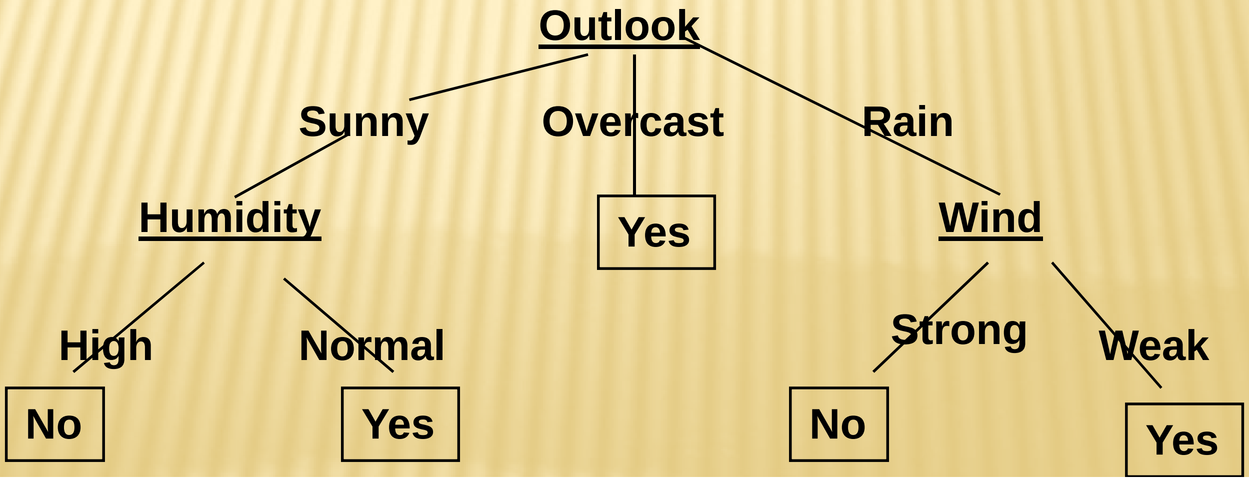


DECISION TREES

- Many inductive knowledge acquisition algorithms generate („induce“) classifiers in form of decision trees.
- A **decision tree** is a simple recursive structure for expressing a sequential classification process.
  - Leaf nodes denote classes
  - Intermediate nodes represent tests
- Decision trees classify instances by sorting them down the tree from the root to some leaf node which provides the classification of the instance.

DECISION TREE: EXAMPLE

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



ID3

- Learns trees by constructing them top down.
- Initial question: “Which attribute should be tested at the root of the tree?” ->each attribute is evaluated using a statistical test to see how well it classifies.
- A descendant of the root node is created for each possible value of this attribute.
- Entire process is repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree.

THE BASIC ID3 ALGORITHM

ID3 (samples, Tattr, Attrs)

```
Create a Root node of the tree
If all samples are positive, return the single-node tree Root, with label +
If all samples are negative, return the single-node tree root, with label –
If attrs is empty, return the single-node tree Root, with label = most common value of
        Tattr in samples
Otherwise begin
    A = the attribute from Attrs that best classifues samples
    The decision attribute for Root = A
    For each positive attribute v_i of A
        Add a new tree branch below Root, corresponding to the
            test A= v_i
        Let sample_v_i be the subset of samples that have
            value v_i for A
        If samples_v_i is empty
            Add a leaf node below this new branch with
            label = most common value of Tattr in samples
        Else below this new branch add the subtree
            ID3(samples_v_i, Tattr, samples – {A})
    End
Return Root
```

THE APRIORI ALGORITHM

- The Apriori Algorithm is for frequent item associations
  - buys(x, “diapers”) → buys(x, “beers”) [0.5%, 60%]
- Technical terms
  - Frequent Itemsets: The sets of item which has minimum support (denoted by L_i for i^{th}-Itemset).
  - Join Operation: To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself.

THE APRIORI ALGORITHM

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

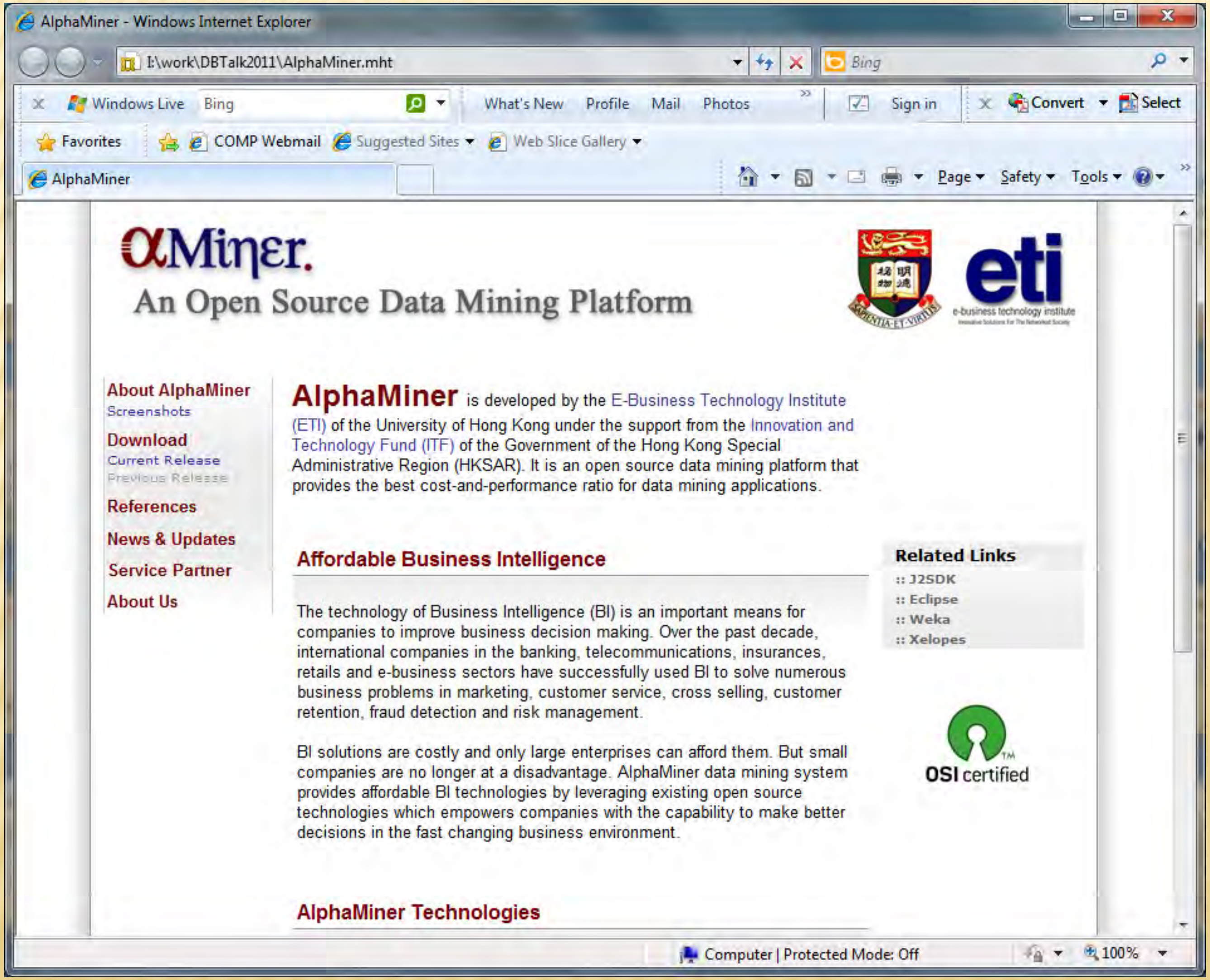
```
$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
     $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do
            increment the count of all candidates in $C_{k+1}$ that are contained in $t$
    $L_{k+1}$  = candidates in $C_{k+1}$ with min_support
    end
return $\cup_k L_k$;
```

ASSOCIATION RULES: EXAMPLE

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

- Minimum support count required is 2
- After mining, we find
  - L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.

ALPHAMINER



An Open Source Data Mining Platform

AlphaMiner is developed by the E-Business Technology Institute (ETI) of the University of Hong Kong under the support from the Innovation and Technology Fund (ITF) of the Government of the Hong Kong Special Administrative Region (HKSAR). It is an open source data mining platform that provides the best cost-and-performance ratio for data mining applications.

Affordable Business Intelligence

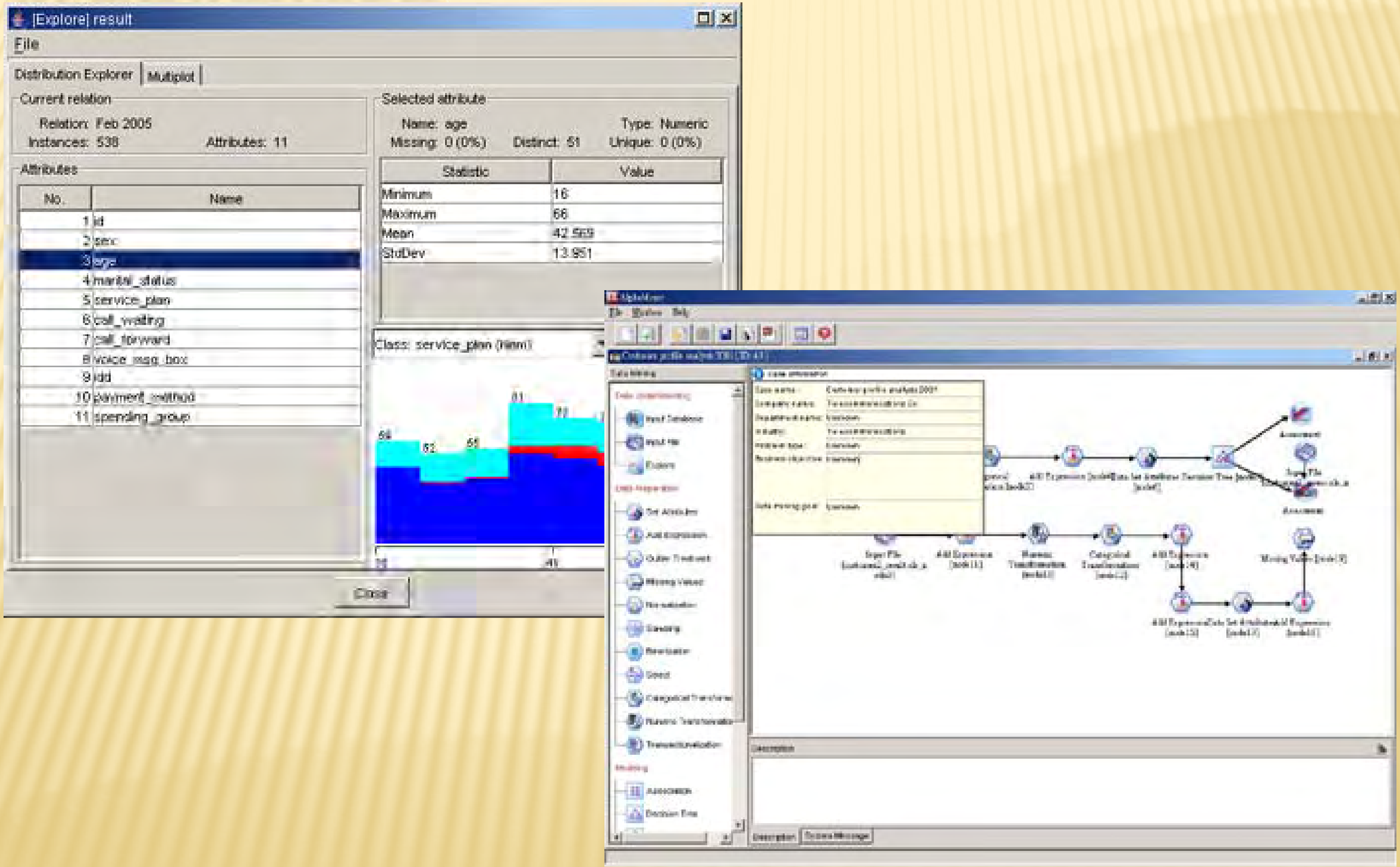
The technology of Business Intelligence (BI) is an important means for companies to improve business decision making. Over the past decade, international companies in the banking, telecommunications, insurances, retails and e-business sectors have successfully used BI to solve numerous business problems in marketing, customer service, cross selling, customer retention, fraud detection and risk management.

BI solutions are costly and only large enterprises can afford them. But small companies are no longer at a disadvantage. AlphaMiner data mining system provides affordable BI technologies by leveraging existing open source technologies which empowers companies with the capability to make better decisions in the fast changing business environment.

AlphaMiner Technologies

www.eti.hku.hk/alphaminer/cur_release.html

ALPHAMINER



[Explore] result

File

Distribution Explorer | Multiplot

Current relation

Relation: Feb 2005
Instances: 538 Attributes: 11

Attributes

No.	Name
1	id
2	sex
3	age
4	marital_status
5	service_plan
6	call_waiting
7	call_forward
8	voice_msg_box
9	idd
10	payment_method
11	spending_group

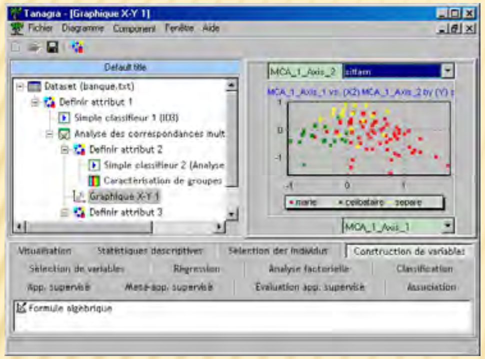
Selected attribute

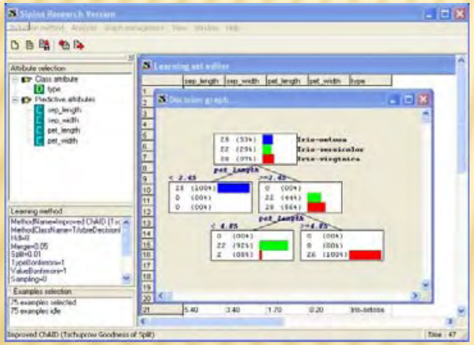
Name: age Type: Numeric
Missing: 0 (0%) Distinct: 51 Unique: 0 (0%)

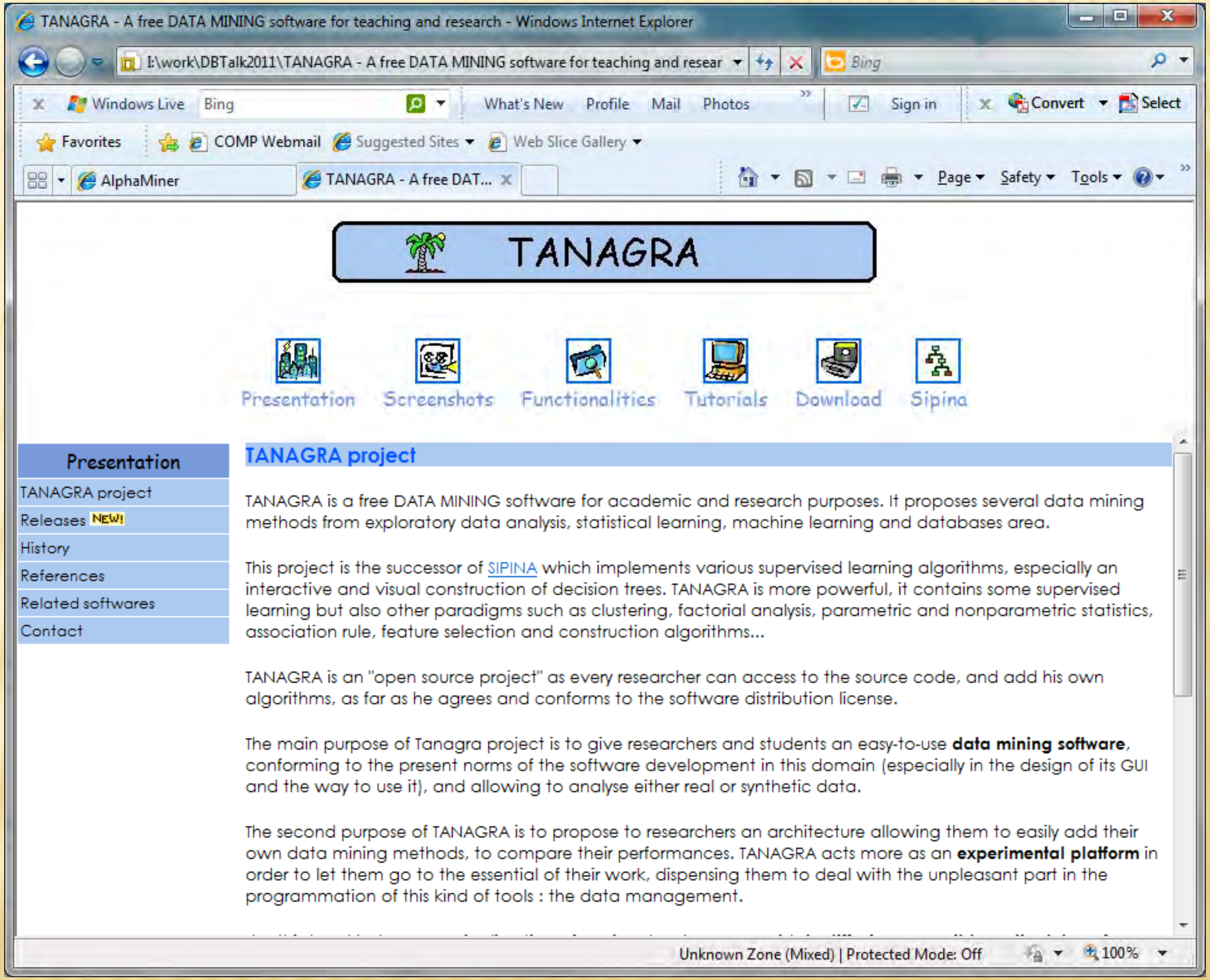
Statistic	Value
Minimum	16
Maximum	66
Mean	42.569
StdDev	13.951

Class: service_plan (Nom)

TANAGRA







TANAGRA project

TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.

This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms...

TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.

The main purpose of Tanagra project is to give researchers and students an easy-to-use **data mining software**, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyse either real or synthetic data.

The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an **experimental platform** in order to let them go to the essential of their work, dispensing them to deal with the unpleasant part in the programmation of this kind of tools : the data management.

eric.univ-lyon2.fr/~ricco/tanagra/index.html

DMINER

DATA MINING ON SOFTWARE PROGRAMS

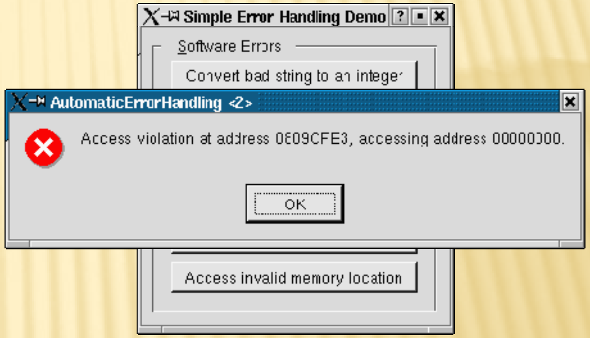
- Software is “full of bugs”
  - Windows 2000, 35 million lines of code
    - 63,000 known bugs at the time of release, 2 per 1000 lines

Courtesy to CNN.com

- Software failure costs
  - Ariane 5 explosion due to “errors in the software of the inertial reference system” (Ariaen-5 flight 501 inquiry board report http://ravel.esrin.esa.it/docs/esa-x-1819eng.pdf)
  - A study by the National Institute of Standards and Technology found that software errors cost the U.S. economy about $59.5 billion annually http://www.nist.gov/director/prog-ofc/report02-3.pdf
- Testing and debugging are laborious and expensive
  - “50% of my company employees are testers, and the rest spends 50% of their time testing!” —Bill Gates, in 1995

WHAT BUGS

- Crashing bugs
  - Symptoms: segmentation faults
  - Reasons: memory access violations
  - Tools: Valgrind, CCured


Simple Error Handling Demo
Software Errors
Convert bad string to an integer
AutomaticErrorHandling <2>
OK
Access invalid memory location

Access violation at address 0E09CFE3, accessing address 00000000.

- Noncrashing bugs
  - Symptoms: unexpected outputs
  - Reasons: logic or semantic errors
    - if ((m >= 0)) or if ((m >= 0) and (m <= ubound))
    - j = i or j= i+1
  - Tools: No sound tools

STATIC PROGRAM ANALYSIS

- Methodology
  - Examine source code directly
  - Enumerate all the possible execution paths without running the program
  - Check user-specified properties
- Strengths
  - Check all possible execution paths
- Problems
  - Shallow semantics
  - Properties can be directly mapped to source code structure

DATA MINING APPROACH

- A graph classification problem
  - Every execution gives one behavior graph
  - Two sets of instances: correct and incorrect
- Values of classification
  - Classification itself does not readily work for bug localization
    - Classifier only labels each run as either correct or incorrect as a whole
    - It does not tell when abnormality happens
  - When abnormality happens?
    - Incremental classification?

REVIEW

- CASE Tools – definition, samples
- Data Mining and its Methods
  - Why, analysis types and methods
- Database Technology
  - Traditional DBMS, DDBMS. MDBMS
  - Recent developments

DATABASE MANAGEMENT SYSTEMS

- Relational DBMS
  - Tables
- Object-Oriented DBMS
  - Objects
- Distributed DBMS
- Multimedia DBMS
- …

DISTRIBUTED DBMS

*DDBMS to Avoid **"islands of information"** problem…*

A "Distributed Database" is a logically interrelated collection of shared data (and a description of this data), physically distributed over a computer network.

A "Distributed DBMS" (DDBMS) is a Software system that permits the management of the distributed database and makes the distribution transparent to users.

Fundamental Principle: make distribution transparent to user.

The fact that fragments are stored on different computers is hidden from the users

DISTRIBUTED DBMS

- Characteristics
  - Collection of logically-related shared data
  - Data split into fragments
  - Fragments may be replicated
  - Fragments/replicas allocated to sites
  - Sites linked by a communication network
  - Data at each site is under control of a DBMS
  - DBMSs handle local applications autonomously
  - Each DBMS participates in at least one global application.

DISTRIBUTED DBMS

- Expect Distributed DBMS to have at least the functionality of a typical DBMS

- **Also to have following functionality**
  - Extended communication services
  - Extended Data Dictionary
  - Distributed query processing
  - Extended concurrency control
  - Extended recovery services

DISTRIBUTED DBMS

Advantages

- ✓Reflects organizational structure
- ✓Improved shareability and
- ✓ local autonomy
- ✓Improved availability
- ✓Improved reliability
- ✓Improved performance
- ✓Economics
- ✓Modular growth

Disadvantages

- Complexity
- Cost
- Security
- Integrity control more difficult
- Lack of standards
- Lack of experience
- Database design more complex

DISTRIBUTED DBMS

- Transparency Requirements
  - Distribution transparency
  - Transaction transparency
  - Performance transparency
  - DBMS transparency (only applicable to heterogeneous)

MULTIMEDIA DBMS

- Multimedia Data
  - Different kinds of media—images, video, audio, graphics, hypertext, hypermedia, and other abstract data types.
- Multimedia Object
  - A multimedia document or presentation containing one or more multimedia data.
- Multimedia Database
  - A database containing one or more multimedia object.

MULTIMEDIA DBMS

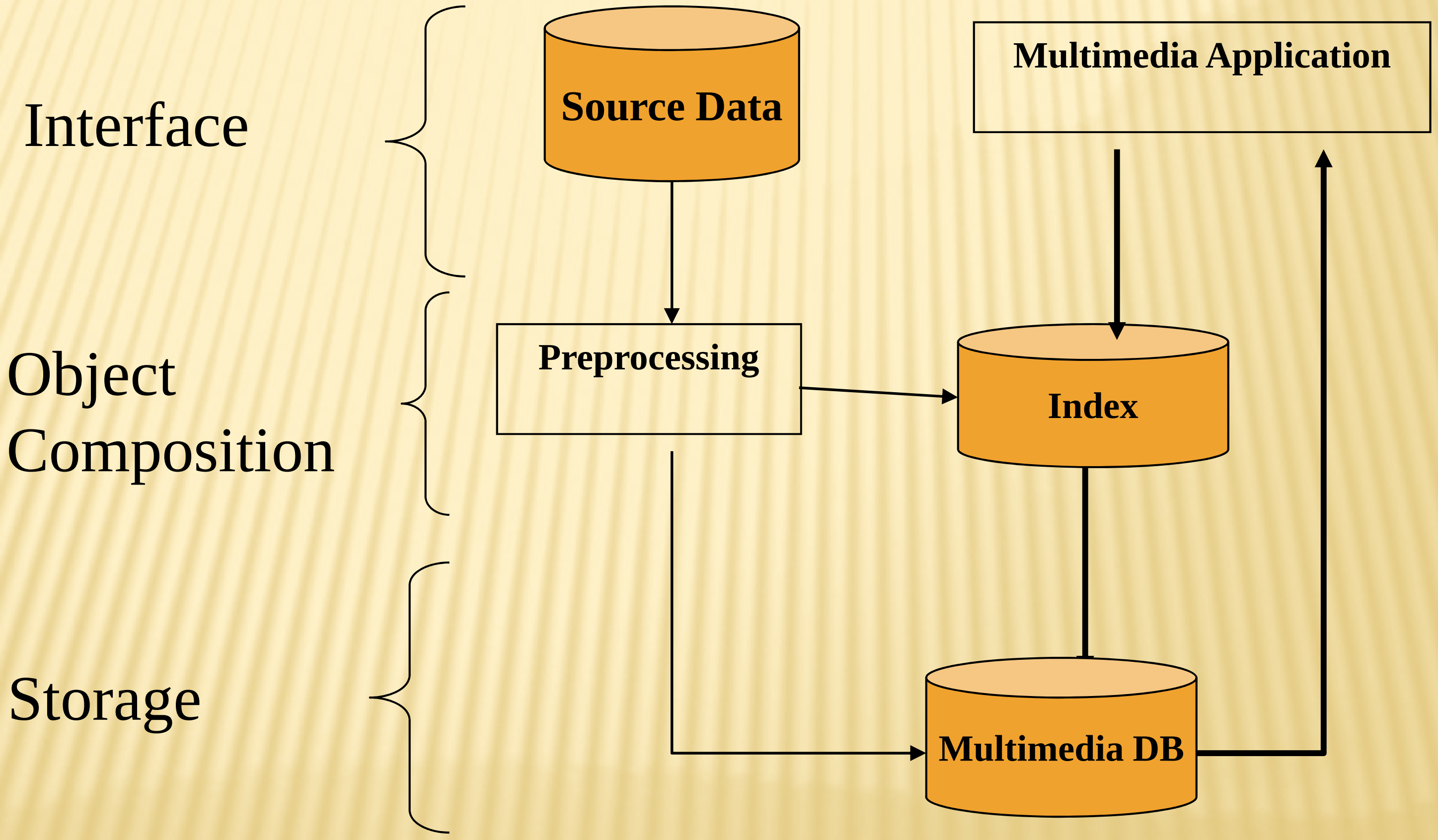
- Characteristics
  - Large object size
  - Synchronous delivery of multimedia objects
  - Multimedia objects may have embedded timing constraints
  - Multimedia object composed of multiple components
  - Queries are not text or numeric based, but content-based
  - Most multimedia transactions are long and requires long processing and retrieval time
  - Multimedia Object presentation is very important

MULTIMEDIA DBMS

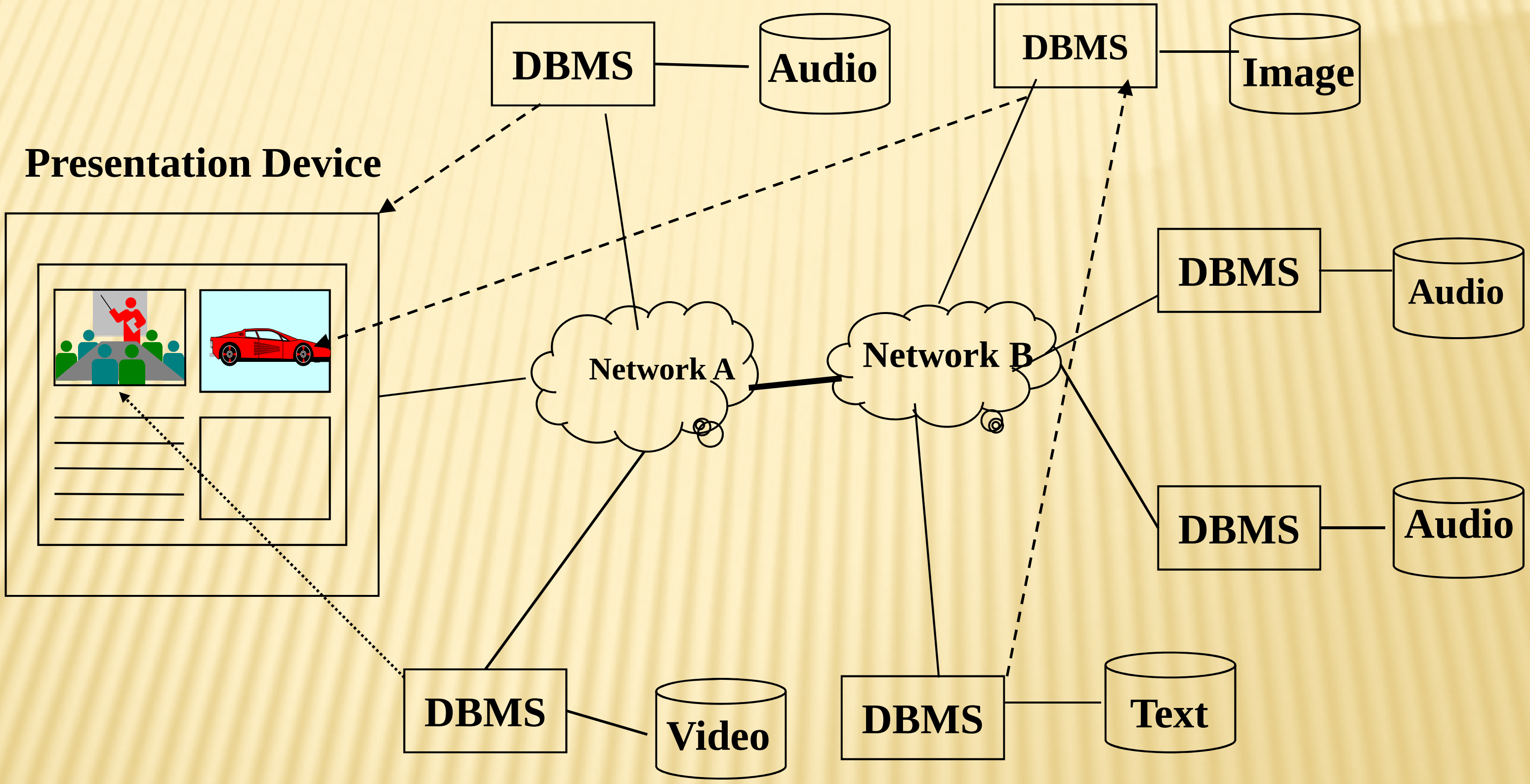
- Expect Multimedia DBMS to have at least the functionality of a typical DBMS

- **Also to have following functionality**
  - Composition and decomposition of Multimedia objects
  - Security and intellectual property protection
  - Concurrency control and locking mechanism
  - Recovery
  - Indexing and clustering

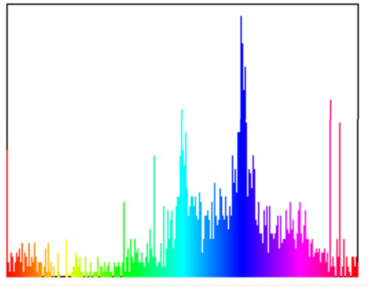
MULTIMEDIA DBMS



DISTRIBUTED MULTIMEDIA DBMS



CONTENT-BASED SEARCHING



- Feature extraction
  - Colors, shapes, textures, and motion
  - Perform aggregation and dimensional reduction
  - Perform similarity and distance calculations among derived features
- Indices store calculated values
- Eliminates need for human annotation

NEW TRENDS IN DATABASE SYSTEMS

- Web Services / XML Databases
- Location-based Databases / Services
- Cloud DBMS
- Data stream management systems
- Flash databases
- Hippocratic databases (data privacy)
- DNA databases
- …

XML IN ACTION - RSS

- RSS (Really Simple Syndication) is an XML application that allows users to “subscribe” to websites.
- Sample uses: Podcasts, Apple iTune Store, news arrival.
- After XHTML, RSS is probably the XML application that web users see most often.

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

<channel>
<title>RSS Example</title>
<description>This is an example of an RSS feed</description>
<link>http://www.domain.com/link.htm</link>
<lastBuildDate>Mon, 28 Aug 2006 11:12:55 -0400 </lastBuildDate>
<pubDate>Tue, 29 Aug 2006 09:00:00 -0400</pubDate>

<item>
<title>Item Example</title>
<description>This is an example of an Item</description>
<link>http://www.domain.com/link.htm</link>
<guid isPermaLink="false"> 1102345</guid>
<pubDate>Tue, 29 Aug 2006 09:00:00 -0400</pubDate>
</item>
</channel>
</rss>
```

XML IN ACTION - GOOGLE

KML (Keyhole Markup Language) is a file format used to display geographic data in an Earth browser such as Google Earth, Google Maps, and Google Maps for mobile.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
                <Placemark>
                <name>Simple placemark</name>
                <description>Attached to the ground. Intelligently places itself
                                at the height of the underlying terrain.</description>
                <Point>
                <coordinates>122.0822035425683,37.42228990140251,0</coordinates>
                </Point>
</Placemark>
</kml>
```

RFID/PML

- Electronic product code (EPC): an unique code for each object
  - RFID Tag
  - RFID reader
- Object Name Service(ONS) : each number corresponds with an address in database
- Product Markup Language (PML)
  - In PML server, PML is used to describe and store information about the item.
- Savant: can work as a router, it get EPC information from the RFID reader, send the information to ONS Server and combine with application program for management of the item.

RFID/PML

```
<pmlcore: Sensor>
        <pmluid:ID>urn:epc:1:4.16.36</pmluid:ID>
        <pmlcore:Observation>
                <pmlcore:DateTime>2002-11-06T13:04:34-
                        06:00</pmlcore:DateTime>
                <pmlcore:Tag>
                        <pmluid:ID>urn:epc:1:2.24.400</pmluid:ID>
                </pmlcore:Tag>
        </pmlcore:Observation>
</pmlcore:Sensor>
```

WHAT IS WDB?

- What are web databases?
  - Two technologies come together
  - Databases
    - Network, Hierarchical, Relational, Object-oriented
    - Systems use for storing, organizing and manipulating data
    - Most businesses have databases for their operations
  - World-Wide Web (WWW)
    - Before the WWW, it was hard to access databases in different networks
    - After mid-1990's, there is almost a web browser accessed by every user
    - People can reach almost sites globally to get products and services
  - Types
    - Using Web as a frontend (Database-to-Web)
    - Using Web as a medium (Database-to-Application-to-Database)

XML DATABASES

- XML is used in many applications now
  - EDI, web services, SOA applications
- Current issues
  - XML Data Model - DTD and XML Schema
  - Query Data Model, Xpath and XQuery
  - Functional Dependencies and Normal Forms
  - XML Storage Techniques: Native and Relational Mappings
  - XML Indexing, Compression, Filtering and Dissemination
  - XML Transaction Management
  - XML Streaming Data
- XML DBMS (en.wikipedia.org/wiki/XML_database)
  - eXist, BaseX, OrientX
  - IBM DB2, MS SQL server, Oracle, Tamino

FLOWR

- FLWOR stands for
  - for, (used to iterate through the result of an XPath expression and to bind a variable intern to each object of the result)
  - let, (used to bind a variable to the whole sequence of objects returned by an XPath expression)
  - where, (used to select those objects from the result returned by an Xpath expression that satisfy given conditions)
  - order by, (used to sort the result of an XPath expression) and
  - return (used to construct the query result)

```
for $c in fn:doc(faculty.xml)/faculty/course
  return
   <course name="{$c/name/text()}" year="{$c/@year}“>    {
   for $s in  fn:doc(students.xml)/students/
                   student[@sid=$c/student/sid]
       return
       <student sid ="{$s/@sid}“>    {
               $s/name,
               $s/surname,
               $c/student[sid=$s/@sid]/grade
               }
      </student> }
    </course>
```

```
for $x in doc("books.xml")/bookstore/book
where $x/price>30
order by $x/title
return $x/title
```

LOCATION-BASED DATABASES

Location-based Databases

FourSquare

GEOLIFE

GeoLife is a GPS-data-driven social networking service where people can share life experiences and connect to each other with their location histories

research.microsoft.com/en-us/projects/geolife/

KEY APPLICATIONS

- Sharing life experiences based on GPS trajectories
- Generic travel recommendations
  - Top interesting locations
  - Travel sequences among locations and
  - Travel experts in a given region
- Collaborative location and activity recommendation
- Personalized friend and location recommendation

COLLABORATIVE LOCATION AND ACTIVITY RECOMMENDATION

Location recommendation given some activity query

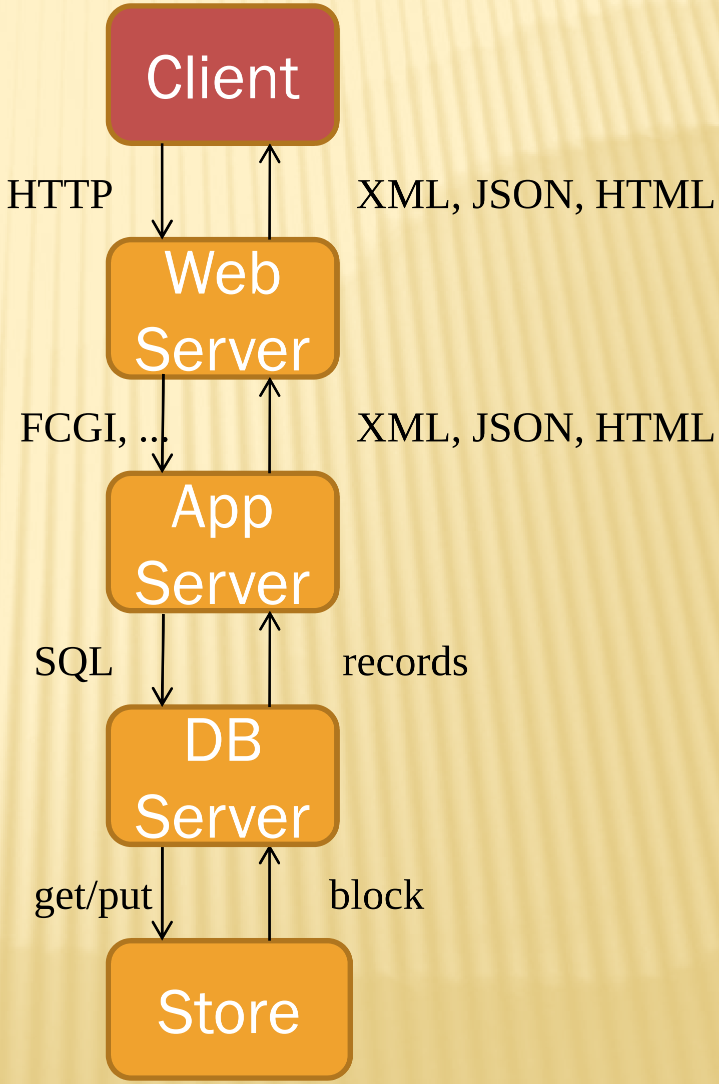
Activity recommendation given some location query

CLOUD DBMS

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and

CLOUD DBMS

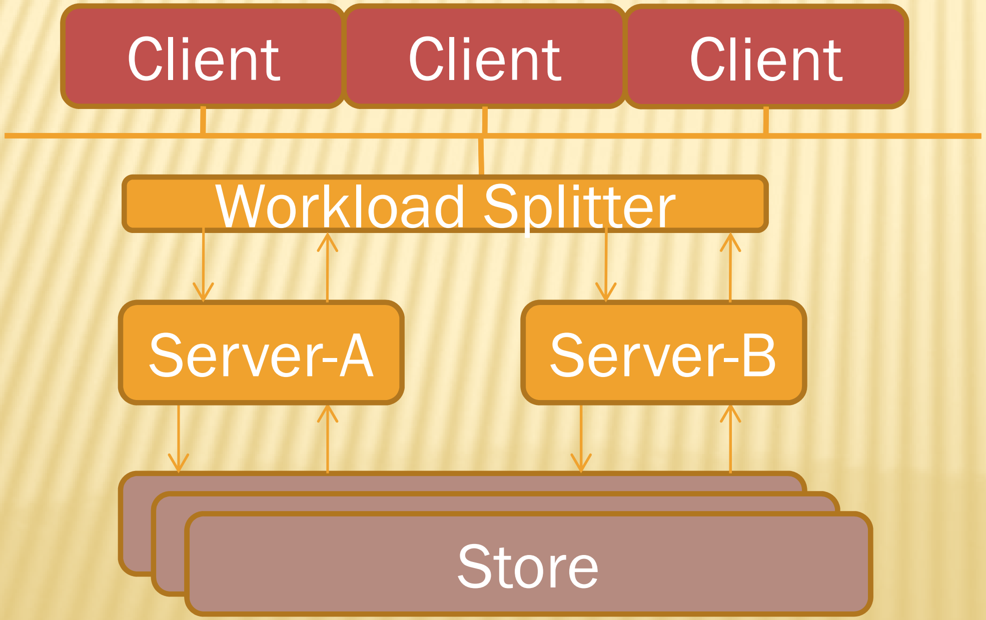
- **Cloud Computing**
- 3 Service Models
  - SaaS, PaaS, IaaS
- 4 Deployment Models
  - Public Cloud, Private Cloud, Community Cloud, Hybrid Cloud
- 5 Essential Characteristics
  - On-demand self-service
  - Resource pooling
  - Rapid elasticity
  - Measured service
  - Broad network access



Source: NIST Definition of Cloud Computing v15

CLOUD DBMS

Application server platform as a service	PaaS
Database platform as a service	PaaS
Identity as a service	PaaS
Storage as a service	IaaS
Software development and test as a service	IaaS
Compute as a service	IaaS



Most popular

App Server as a service

Database as a service

CLOUD DBMS

- Key/value stores: storing all related information about a single item, or object, as a single entity
  - as opposed to having multiple tables in a relational database linked by primary/foreign key relationships.
- Cloud DBMS
  - Xeround
  - Microsoft SQL Azure Database
  - SimpleDB
  - Google AppEngine Data Store
  - Database.com
  - ClearDB
  - CouchOne

CLOUD DBMS

- Two largest types of data management market
  - Transactional Data Management
  - Analytical Data Management
- Which one will benefit from moving to the cloud?

MEMORY DATABASES

- Too expensive before, but flash drives are very cheap now
- However
  - Lack of consistent I/O behavior across Flash Device Models
  - No Reference DBMS design for Flash
  - No Performance Model for Flash
- Define how Flash devices should support DBMS
- Provide DBMS a little more control over I/O behavior

FLASH DEVICES CHARACTERISTICS

- Good

* Great Performance

(40 MB/s Reads, 10 MB/s Writes)

* Low energy consumption

* Potentially safe to power failure

- Bad

* Write Granularity (page level)

* Erase Before Writes (block level)

* Sequential writes within a block

* Limited Lifetime

- Controller (Flash Translation Layer)

* out-of-place updates using log blocks

* Garbage collection

* Mapping between logical address space and physical Flash space

REVIEW

- CASE Tools – definition, samples
- Data Mining and its Methods
  - Why, analysis types and methods
- Database Technology
  - Traditional DBMS, DDBMS. MDBMS
  - Recent developments