Data Warehousing
資料倉儲

Data Warehousing, Data Mining, and Business Intelligence

1001DW02
MI4
Tue. 6,7 (13:10-15:00) B427

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.im.tku.edu.tw/~myday/

2011-09-13

Syllabus

週次	日期	內容（Subject/Topics）
1	100/09/06	Introduction to Data Warehousing
2	100/09/13	Data Warehousing, Data Mining, and Business Intelligence
3	100/09/20	Data Preprocessing: Integration and the ETL process
4	100/09/27	Data Warehouse and OLAP Technology
5	100/10/04	Data Warehouse and OLAP Technology
6	100/10/11	Data Cube Computation and Data Generation
7	100/10/18	Data Cube Computation and Data Generation
8	100/10/25	Project Proposal
9	100/11/01	期中考試週

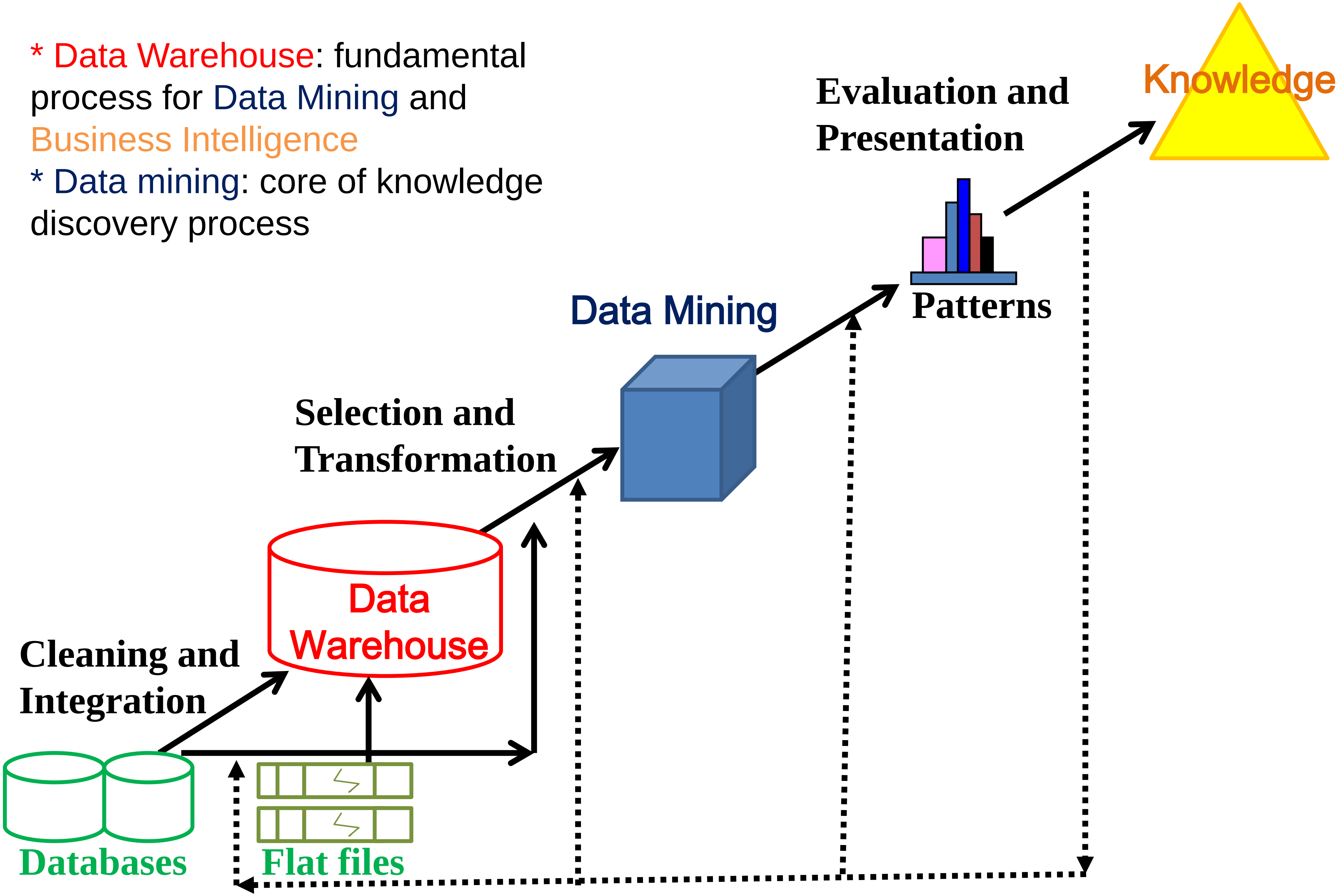
Syllabus

週次	日期	內容（Subject/Topics）
10	100/11/08	Association Analysis
11	100/11/15	Classification and Prediction
12	100/11/22	Cluster Analysis
13	100/11/29	Sequence Data Mining
14	100/12/06	Social Network Analysis
15	100/12/13	Link Mining
16	100/12/20	Text Mining and Web Mining
17	100/12/27	Project Presentation
18	101/01/03	期末考試週

Knowledge Discovery (KDD) Process

* Data Warehouse: fundamental process for Data Mining and Business Intelligence

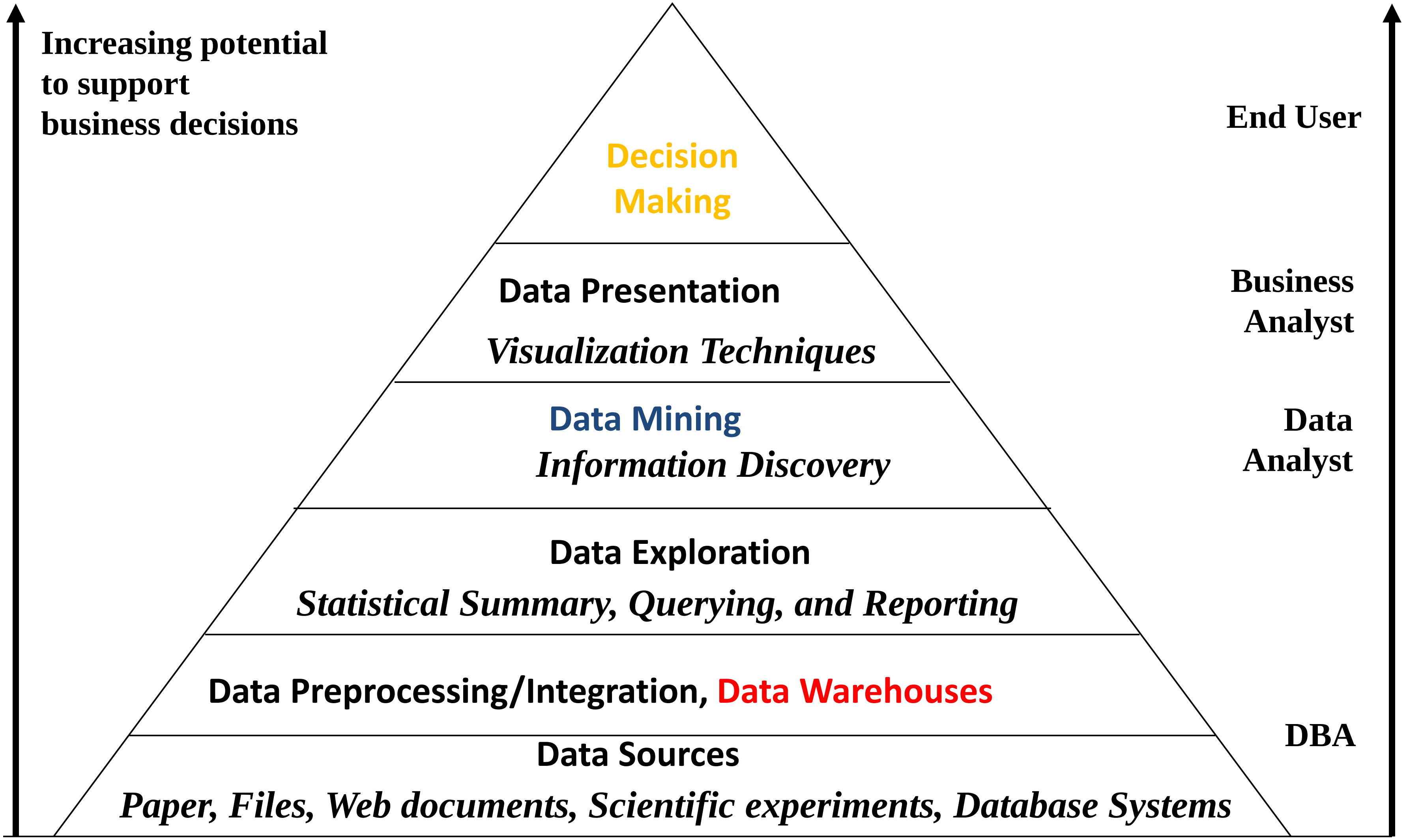
* Data mining: core of knowledge discovery process



Source: Han & Kamber (2006)

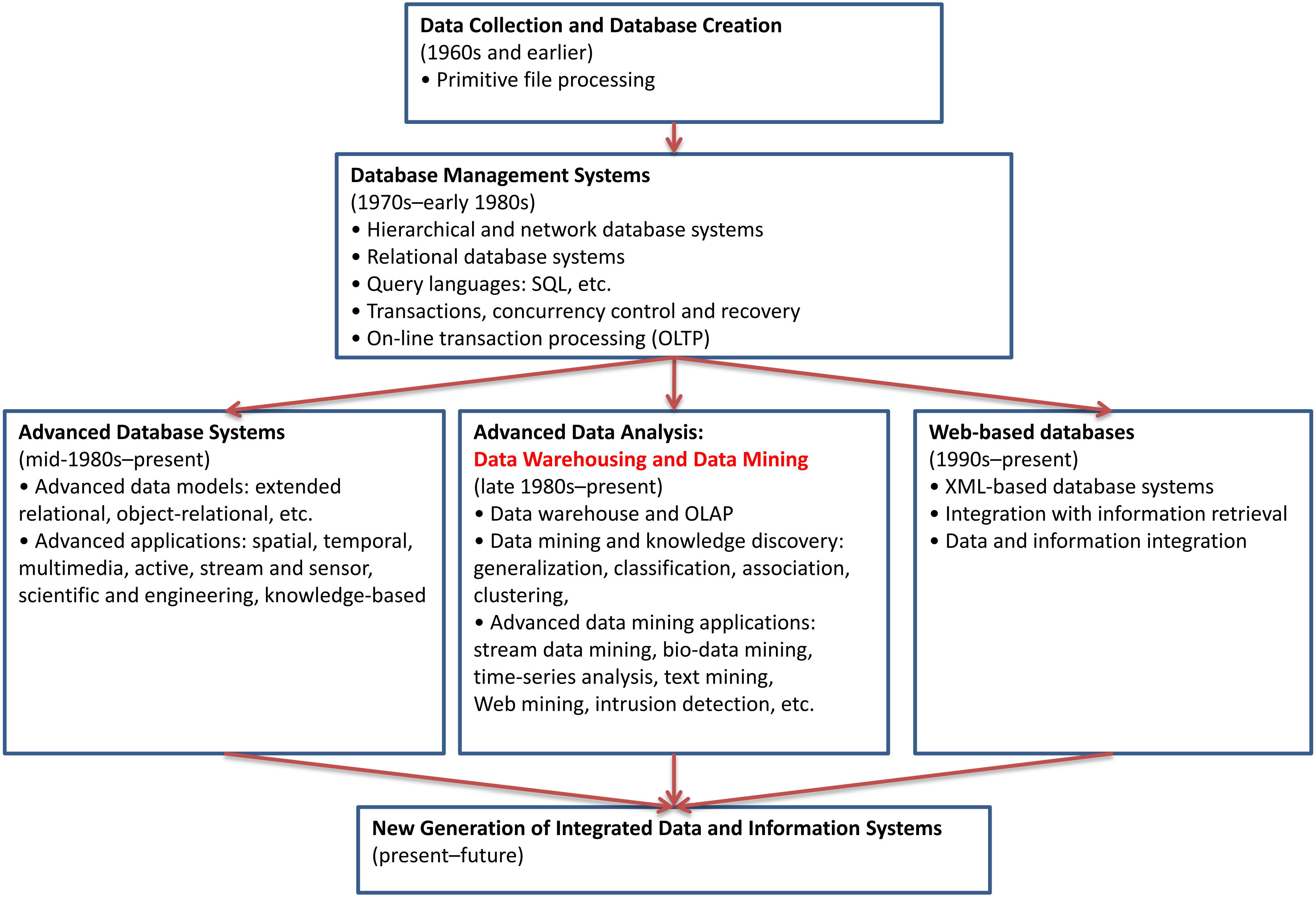
Data Warehouse

Data Mining and Business Intelligence


Increasing potential
to support
business decisions
Decision
Making
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems
End User
Business
Analyst
Data
Analyst
DBA

Source: Han & Kamber (2006)

Evolution of Database Technology



Data Collection and Database Creation
(1960s and earlier)

- Primitive file processing

Database Management Systems
(1970s–early 1980s)

- Hierarchical and network database systems
- Relational database systems
- Query languages: SQL, etc.
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)

- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis:
Data Warehousing and Data Mining
(late 1980s–present)

- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering,
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.

Web-based databases
(1990s–present)

- XML-based database systems
- Integration with information retrieval
- Data and information integration

New Generation of Integrated Data and Information Systems
(present–future)

Source: Han & Kamber (2006)

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

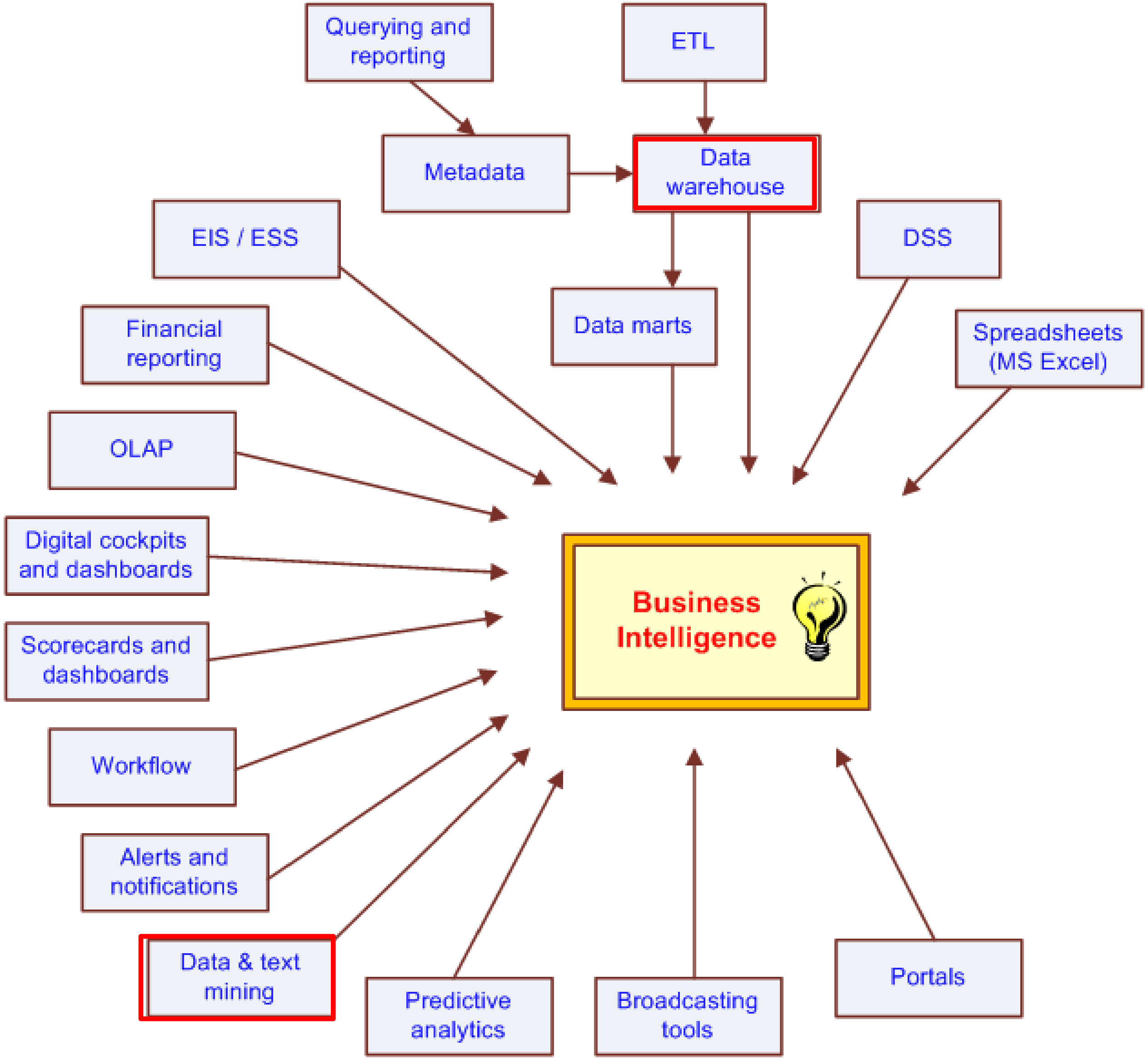
Source: Han & Kamber (2006)

A Brief History of Business Intelligence (BI)

- The term BI was coined by the Gartner Group in the mid-1990s
- Concept of BI
  - 1970s - MIS reporting - static/periodic reports
  - 1980s - Executive Information Systems (EIS)
  - 1990s - OLAP, dynamic, multidimensional, ad-hoc reporting -> coining of the term "BI"
  - 2005+ Inclusion of AI and Data/Text Mining capabilities; Web-based Portals/Dashboards
  - 2010s - yet to be seen

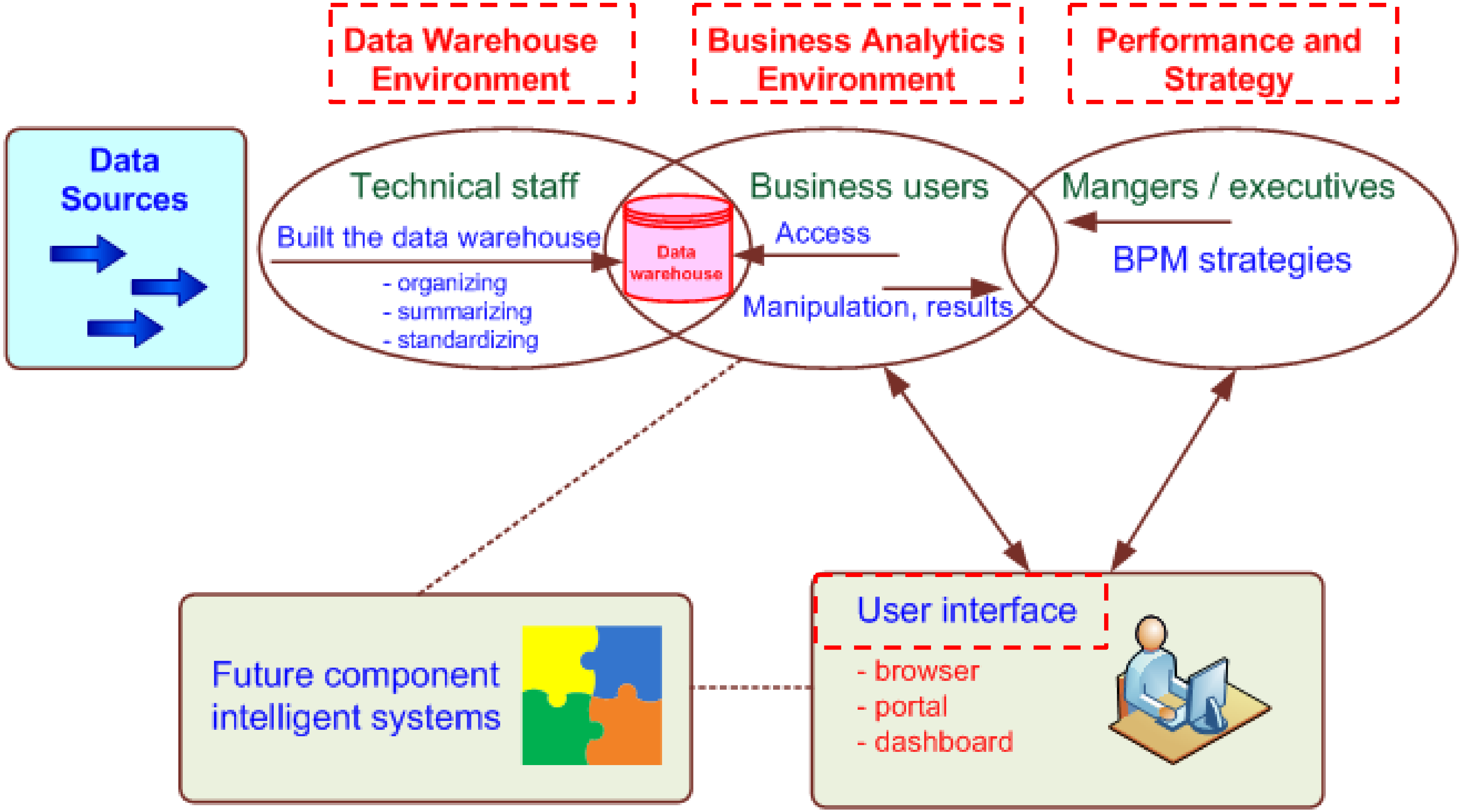
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Evolution of Business Intelligence (BI)



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

A High-Level Architecture of BI



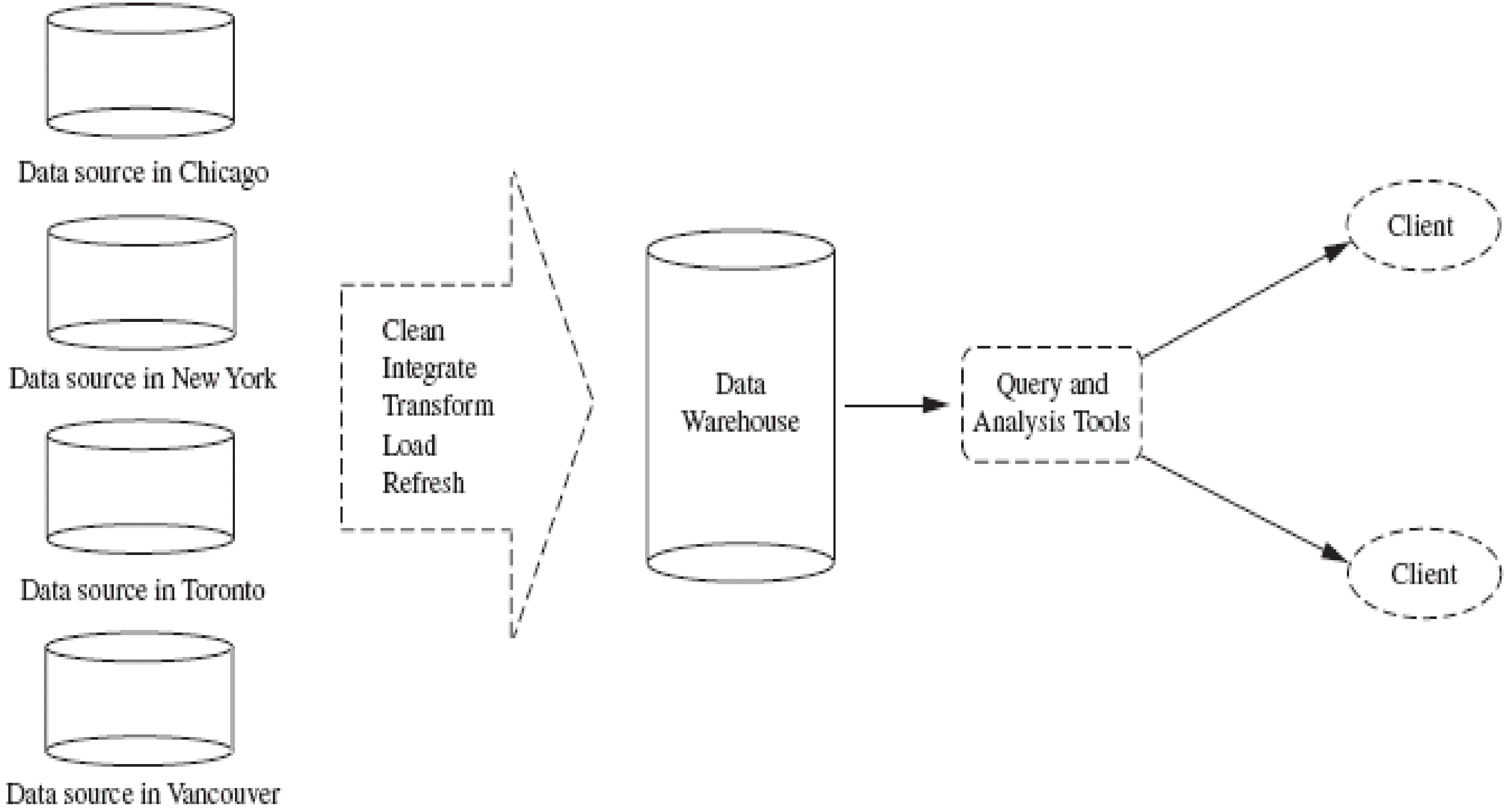
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

The Architecture of BI

- A BI system has four major components
  - a data warehouse, with its source data
  - business analytics, a collection of tools for manipulating, mining, and analyzing the data in the data warehouse;
  - business performance management (BPM) for monitoring and analyzing performance
  - a user interface (e.g., dashboard)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Typical framework of a data warehouse

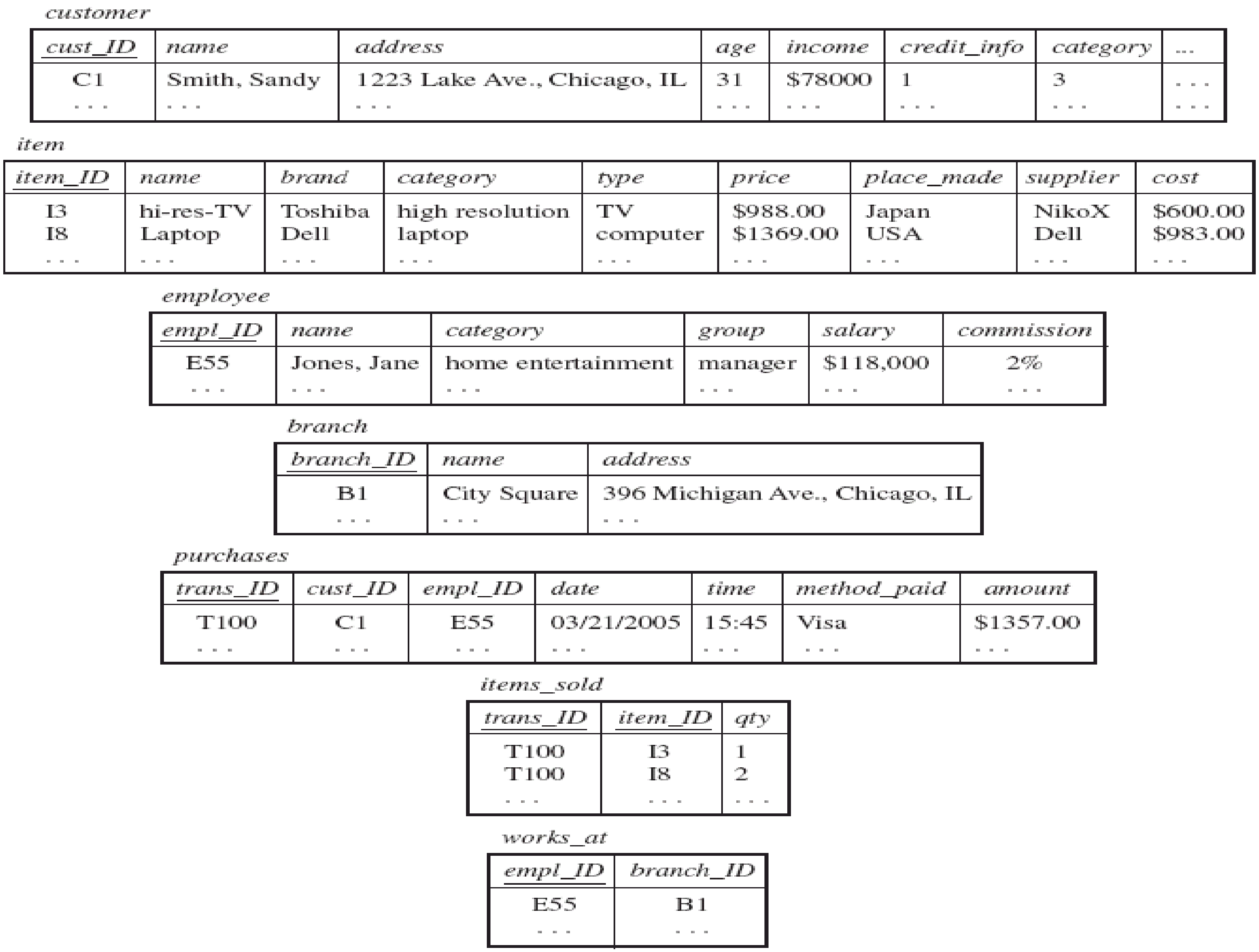


Source: Han & Kamber (2006)

ETL

- Extraction
- Transformation
- Loading

Relational Database



customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

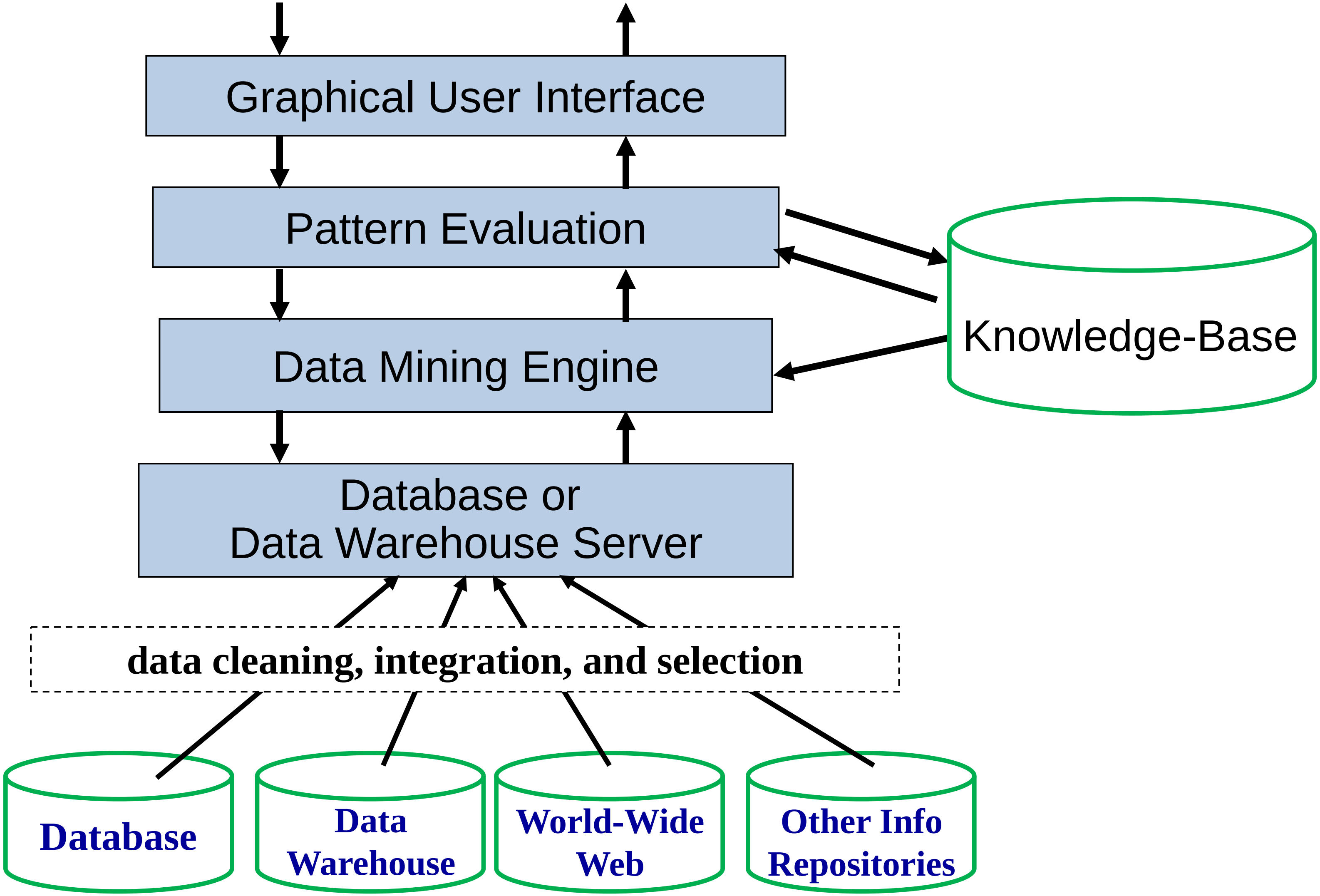
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

Source: Han & Kamber (2006)

Architecture of a typical data mining system



Graphical User Interface

Pattern Evaluation

Data Mining Engine

Knowledge-Base

Database or
Data Warehouse Server

data cleaning, integration, and selection

Database

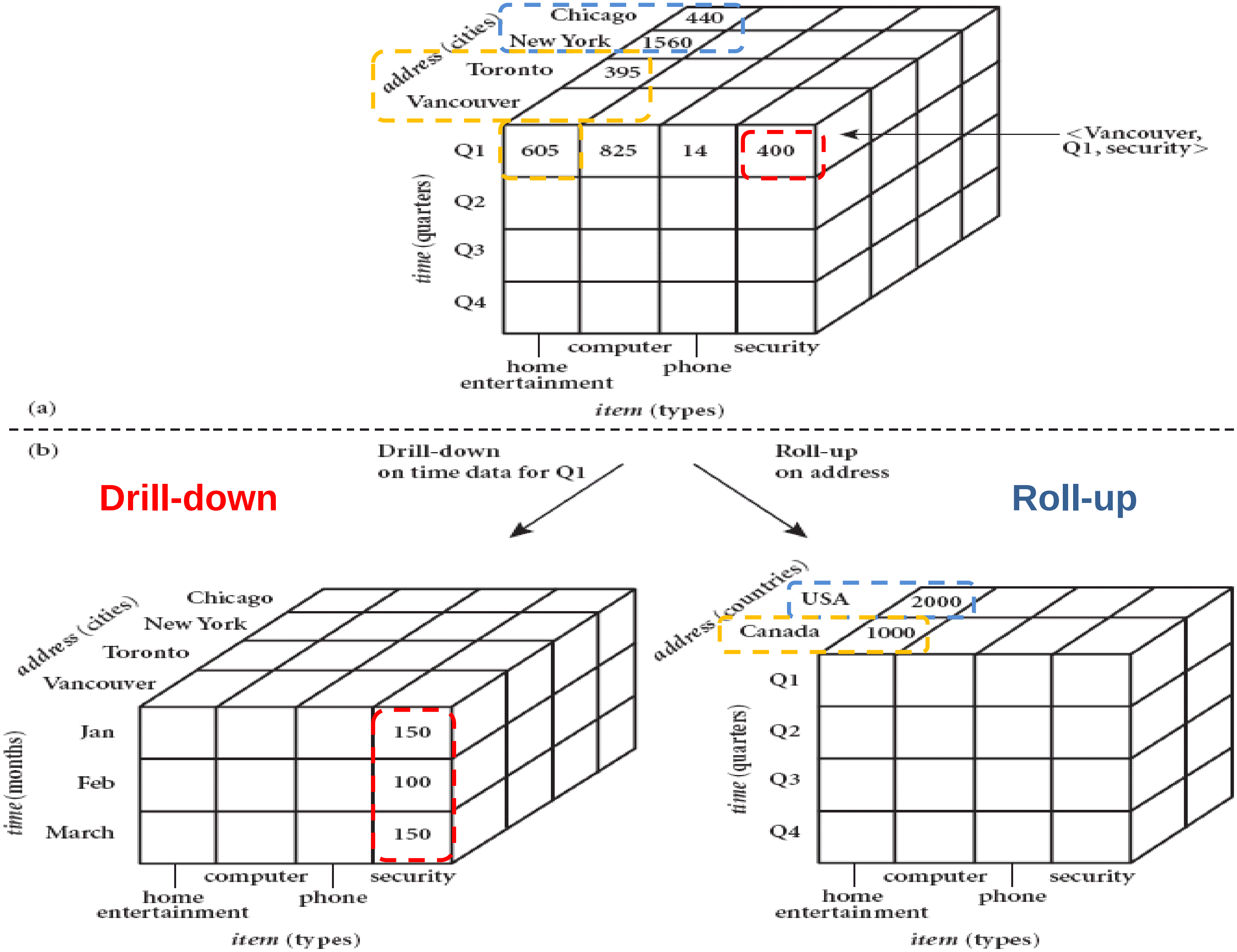
Data Warehouse

World-Wide Web

Other Info Repositories

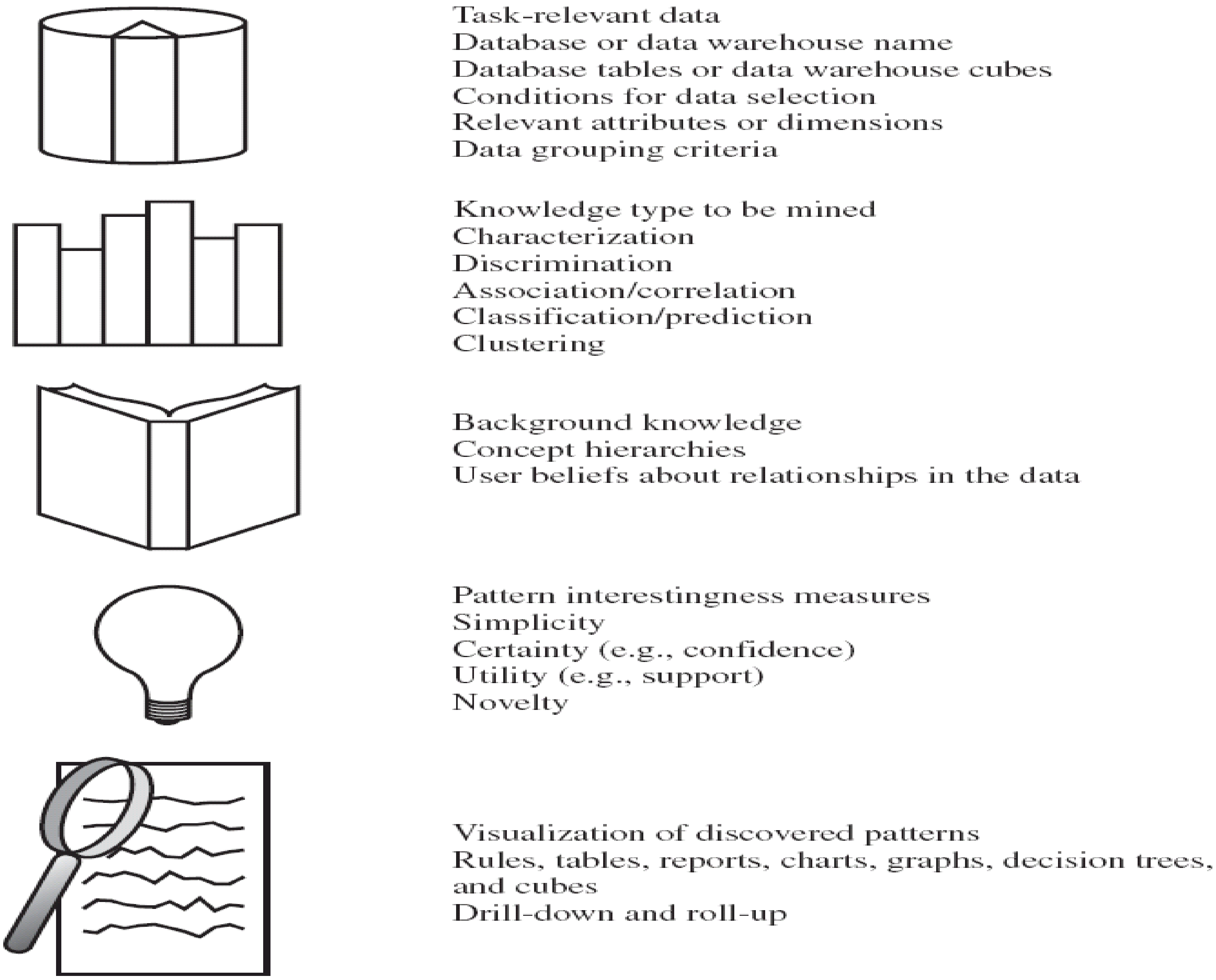
Source: Han & Kamber (2006)

Multidimensional data cube for data warehousing



Source: Han & Kamber (2006)

Primitives for specifying a data mining task



Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria

Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering

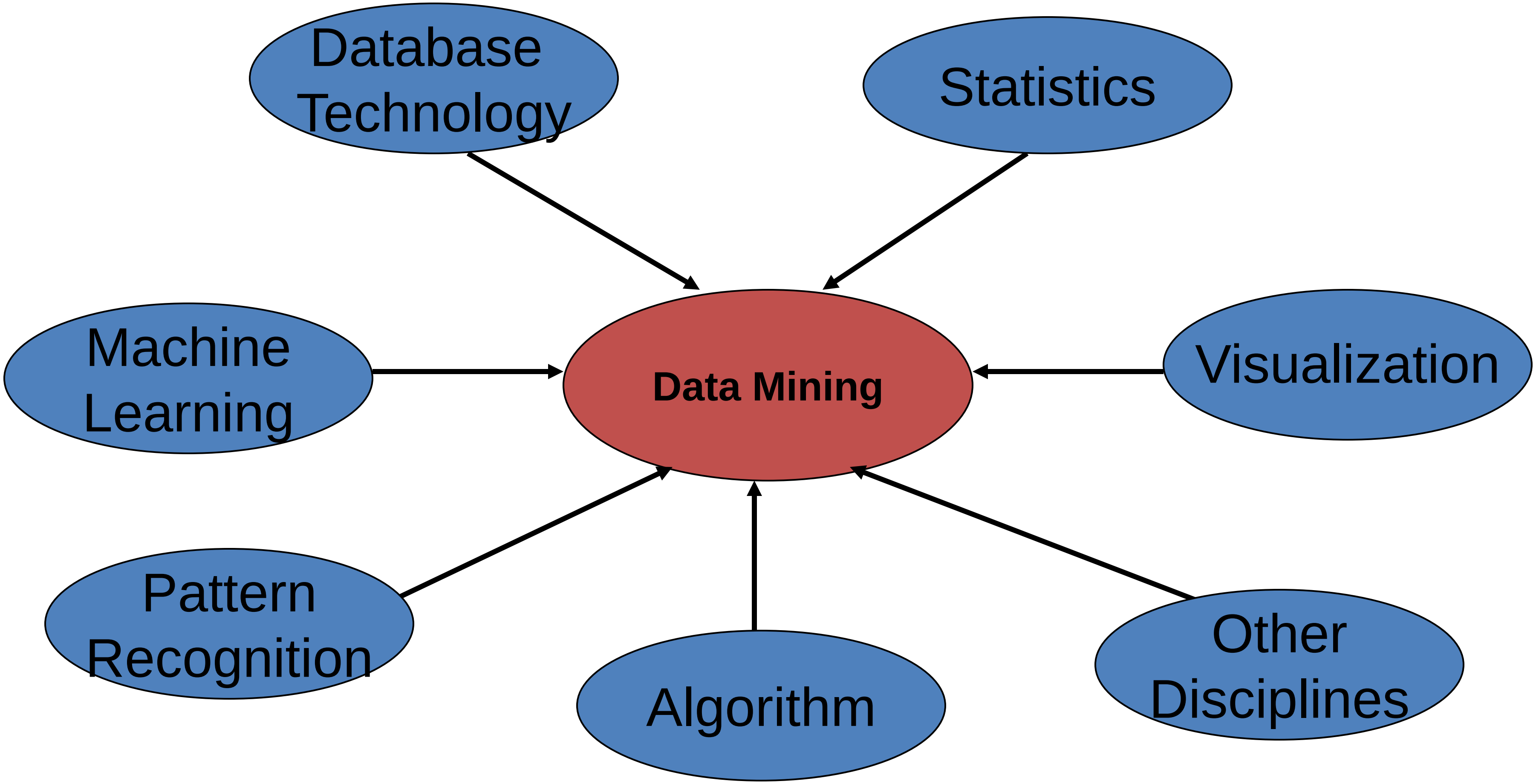
Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

Source: Han & Kamber (2006)

Data Mining: Confluence of Multiple Disciplines



Source: Han & Kamber (2006)

Differences between a data warehouse and a database

- Data warehouse:
  - A data warehouse is a repository of information collected from multiple sources over a history of time stored under a unified schema and used for data analysis and decision support
  - There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another.
- Database:
  - A database is a collection of interrelated data that represents the current status of the stored data.
  - A database system supports ad-hoc query and on-line transaction processing.

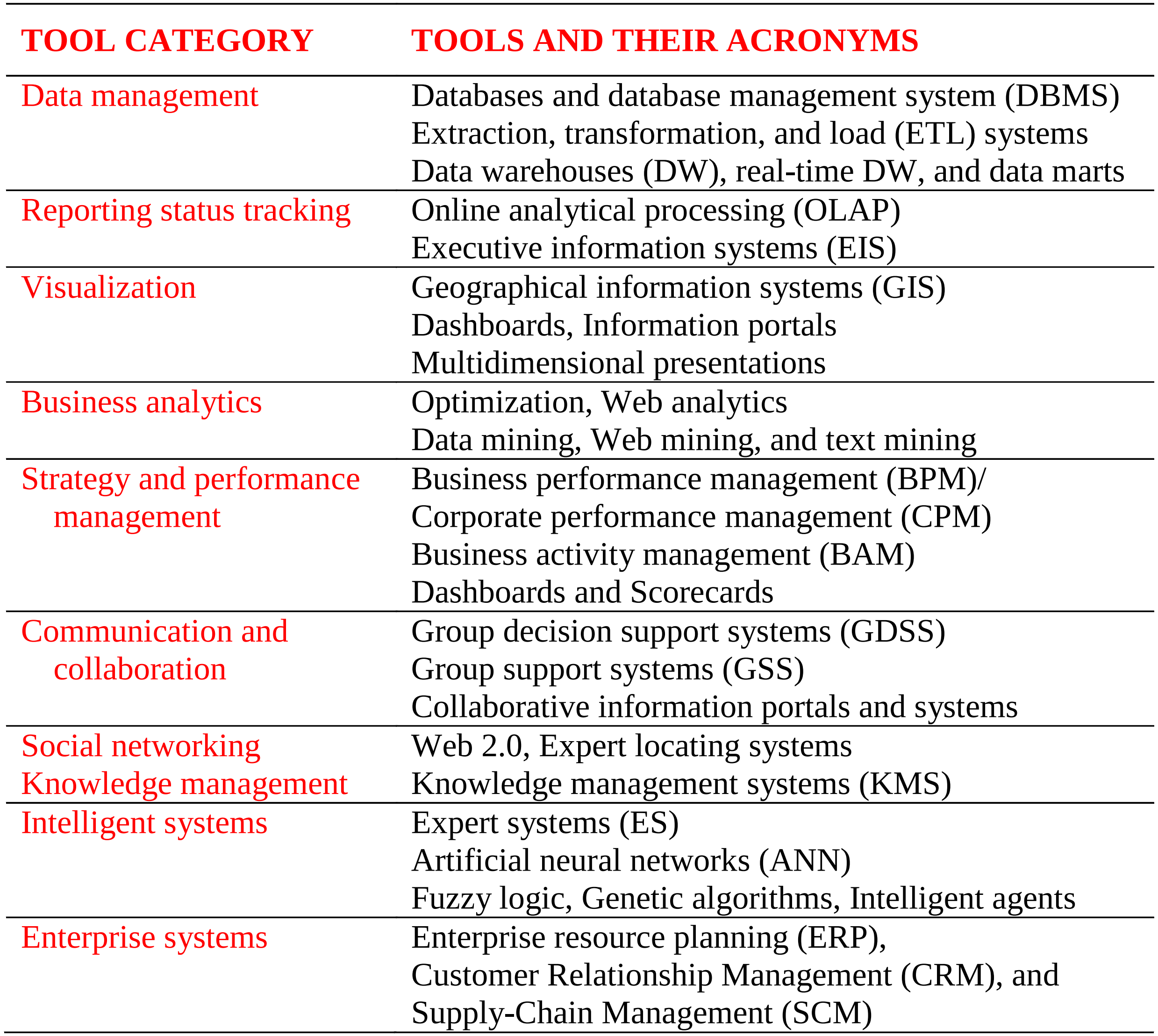
Source: Han & Kamber (2006)

Similarities between a data warehouse and a database

- Both are repositories of information storing huge amounts of persistent data.

Source: Han & Kamber (2006)

Major Tool Categories for Management Support Systems (MSS)



TOOL CATEGORY	TOOLS AND THEIR ACRONYMS
Data management	Databases and database management system (DBMS) Extraction, transformation, and load (ETL) systems Data warehouses (DW), real-time DW, and data marts
Reporting status tracking	Online analytical processing (OLAP) Executive information systems (EIS)
Visualization	Geographical information systems (GIS) Dashboards, Information portals Multidimensional presentations
Business analytics	Optimization, Web analytics Data mining, Web mining, and text mining
Strategy and performance management	Business performance management (BPM)/ Corporate performance management (CPM) Business activity management (BAM) Dashboards and Scorecards
Communication and collaboration	Group decision support systems (GDSS) Group support systems (GSS) Collaborative information portals and systems
Social networking Knowledge management	Web 2.0, Expert locating systems Knowledge management systems (KMS)
Intelligent systems	Expert systems (ES) Artificial neural networks (ANN) Fuzzy logic, Genetic algorithms, Intelligent agents
Enterprise systems	Enterprise resource planning (ERP), Customer Relationship Management (CRM), and Supply-Chain Management (SCM)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

IBM Watson: Smartest Machine On Earth (2011)

- IBM Watson: Final Jeopardy! and the Future of Watson, http://www.youtube.com/watch?v=lI-M7O_bRNg
- Smartest Machine On Earth (2011) 1/4, http://www.youtube.com/watch?v=qIDLd1HUjxY
- Smartest Machine On Earth (2011) 2/4, http://www.youtube.com/watch?v=gg656SKnVQM
- Smartest Machine On Earth (2011) 3/4 , http://www.youtube.com/watch?v=hZ7Hsob-h_Q
- Smartest Machine On Earth (2011) 4/4, http://www.youtube.com/watch?v=ozQG_jIB8SE

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.
- Lucene, http://en.wikipedia.org/wiki/Lucene
- Machine Learning, http://en.wikipedia.org/wiki/Machine_learning