Concept and Applications of Data Mining

Week 1

Topics

- Introduction
- Syllabus
- Data Mining Concepts
- Team Organization

Introduction Session

- Your name and major
- The definition of data mining
- Your expectation from this course

Course Syllabus

- Syllabus

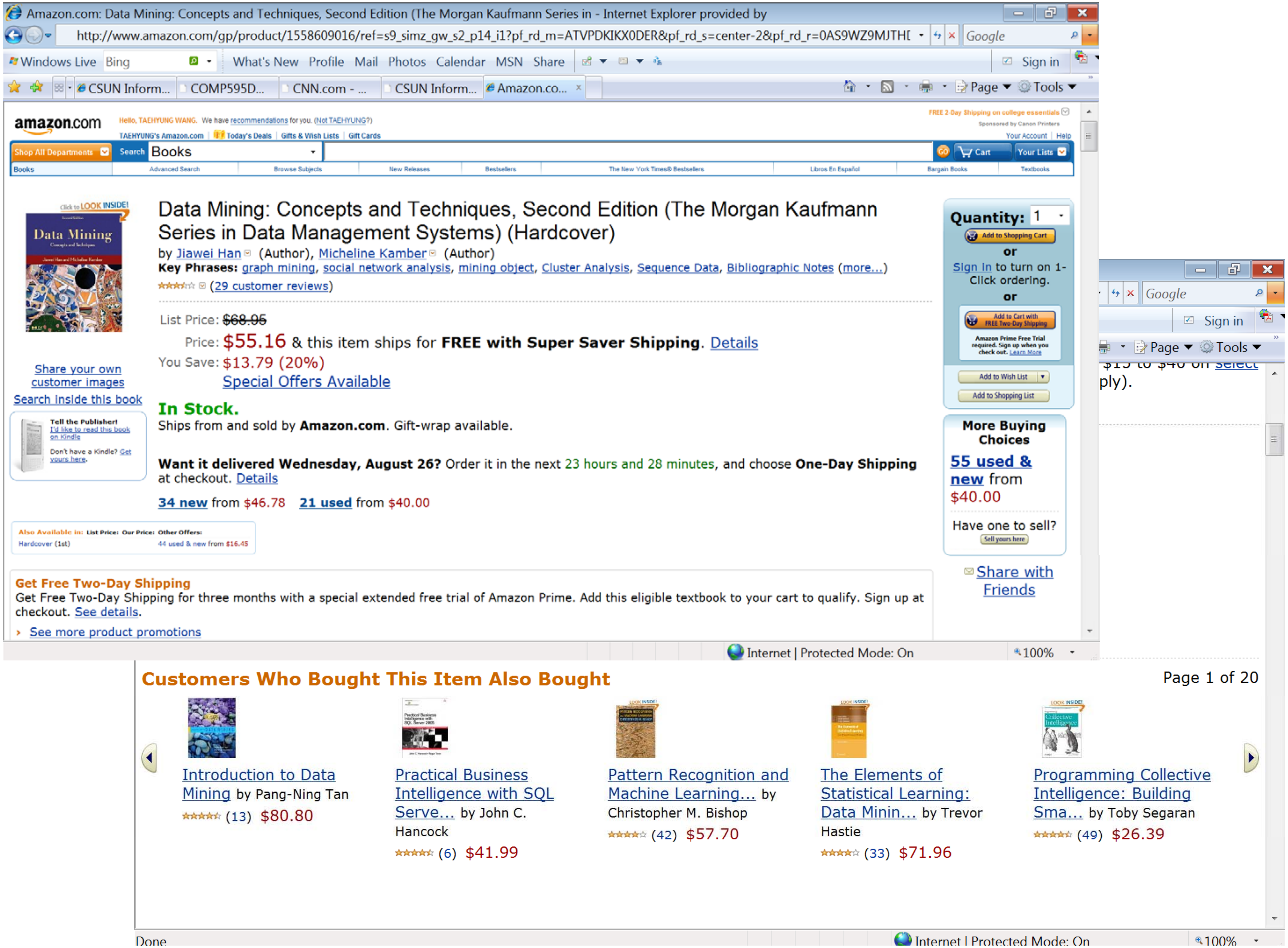
Data Mining Applications

WAL★MART
ALWAYS LOW PRICES.
Always
Walmart.com
citi
CHASE
STATE FARM
Auto
Life
Fire
INSURANCE
ups
at&t
verizonwireless
Authorized Agent
ebY
amazon.com
NCBI
AMGEN
Google
YAHOO!
msn
NATIONAL SECURITY AGENCY
NSA
UNITED STATES OF AMERICA
DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION
U.S. DEPARTMENT OF HOMELAND SECURITY
CENTRAL INTELLIGENCE AGENCY
UNITED STATES OF AMERICA
NASA

Classes of Data-Mining Applications in 2003

Data-Mining Applications	Percentage
Banking	13
Bioinformatics/biotech	10
Direct marketing/fundraising	10
Fraud detection	9
Scientific data	9
Insurance	8
Telecommunication	8
Medical/pharmaceuticals	6
Retail	6
e-Commerce/Web	5
Other	4
Investment/stocks	3
Manufacturing	2
Security	2
Supply chain analysis	2
Travel	2
Entertainment	1

Source: www.kdnuggets.com



Hello, TAEHYUNG WANG. We have recommendations for you. (Not TAEHYUNG?)

TAEHYUNG's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

FREE 2-Day Shipping on college essentials

Sponsored by Canon Printers

Your Account | Help

Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) (Hardcover)

by Jiawei Han (Author), Micheline Kamber (Author)

Key Phrases: graph mining, social network analysis, mining object, Cluster Analysis, Sequence Data, Bibliographic Notes (more...)

(29 customer reviews)

List Price: ~~$68.95~~

Price: $55.16 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $13.79 (20%)

Special Offers Available

Share your own customer images

Search inside this book

Tell the Publisher!
I'd like to read this book on Kindle

Don't have a Kindle? Get yours here.

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Wednesday, August 26? Order it in the next 23 hours and 28 minutes, and choose **One-Day Shipping** at checkout. Details

34 new from $46.78 **21 used** from $40.00

Also Available in:	List Price:	Our Price:	Other Offers:
Hardcover (1st)			44 used & new from $16.45

Quantity: 1

or

Sign in to turn on 1-Click ordering.

or

Amazon Prime Free Trial required. Sign up when you check out. Learn More

More Buying Choices

55 used & new from $40.00

Have one to sell?

Share with Friends

Get Free Two-Day Shipping

Get Free Two-Day Shipping for three months with a special extended free trial of Amazon Prime. Add this eligible textbook to your cart to qualify. Sign up at checkout. See details.

› See more product promotions

$15 to $40 on select

ply).

Customers Who Bought This Item Also Bought

Page 1 of 20

Introduction to Data Mining by Pang-Ning Tan (13) $80.80

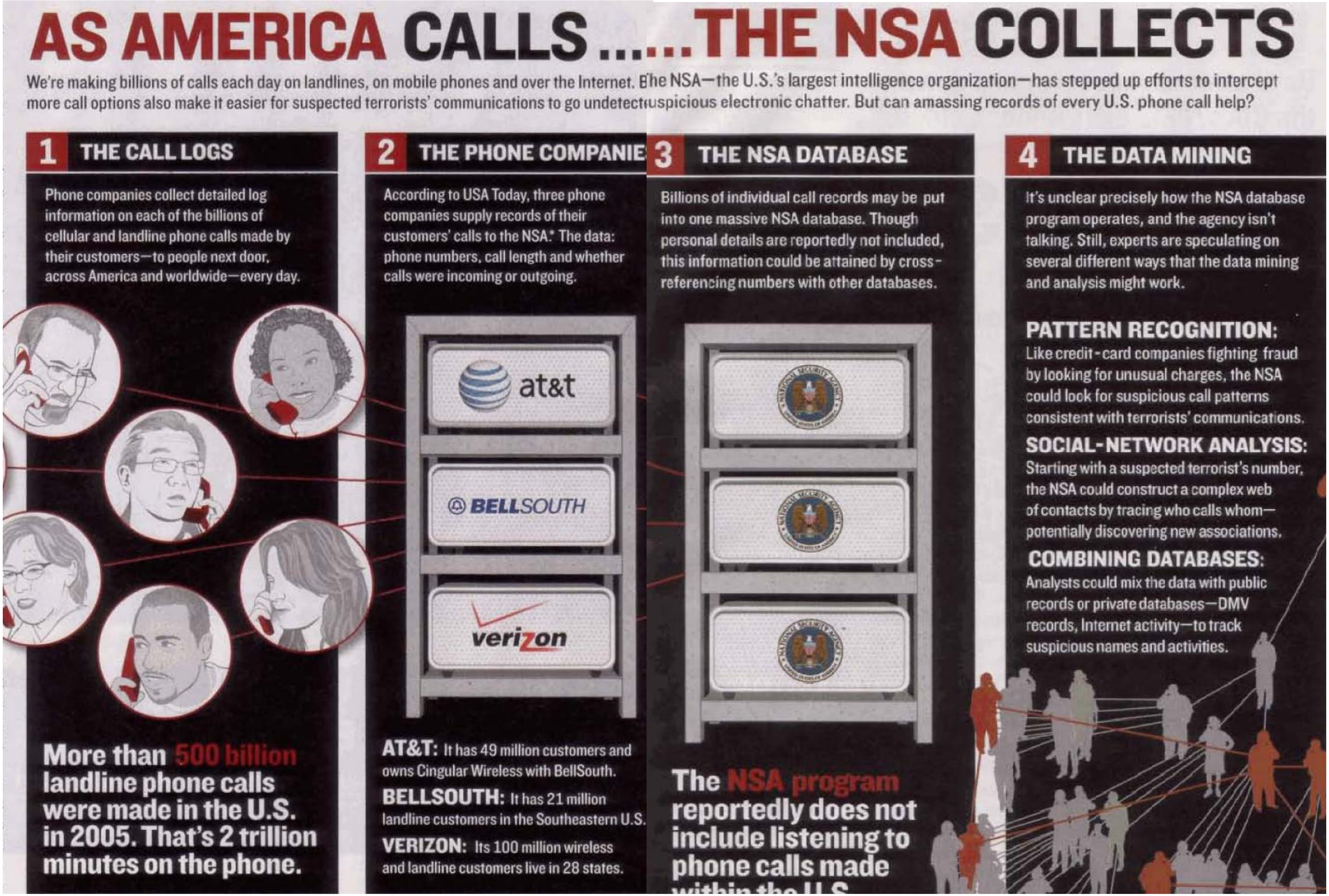
Practical Business Intelligence with SQL Serve... by John C. Hancock (6) $41.99

Pattern Recognition and Machine Learning... by Christopher M. Bishop (42) $57.70

The Elements of Statistical Learning: Data Minin... by Trevor Hastie (33) $71.96

Programming Collective Intelligence: Building Sma... by Toby Segaran (49) $26.39

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA



AS AMERICA CALLS......THE NSA COLLECTS

We're making billions of calls each day on landlines, on mobile phones and over the Internet. Ehe NSA—the U.S.'s largest intelligence organization—has stepped up efforts to intercept more call options also make it easier for suspected terrorists' communications to go undetectuspicious electronic chatter. But can amassing records of every U.S. phone call help?

1 THE CALL LOGS

Phone companies collect detailed log information on each of the billions of cellular and landline phone calls made by their customers—to people next door, across America and worldwide—every day.

More than 500 billion landline phone calls were made in the U.S. in 2005. That's 2 trillion minutes on the phone.

2 THE PHONE COMPANIE

According to USA Today, three phone companies supply records of their customers' calls to the NSA.* The data: phone numbers, call length and whether calls were incoming or outgoing.

AT&T: It has 49 million customers and owns Cingular Wireless with BellSouth.

BELLSOUTH: It has 21 million landline customers in the Southeastern U.S.

VERIZON: Its 100 million wireless and landline customers live in 28 states.

3 THE NSA DATABASE

Billions of individual call records may be put into one massive NSA database. Though personal details are reportedly not included, this information could be attained by cross-referencing numbers with other databases.

The NSA program reportedly does not include listening to phone calls made within the U.S.

4 THE DATA MINING

It's unclear precisely how the NSA database program operates, and the agency isn't talking. Still, experts are speculating on several different ways that the data mining and analysis might work.

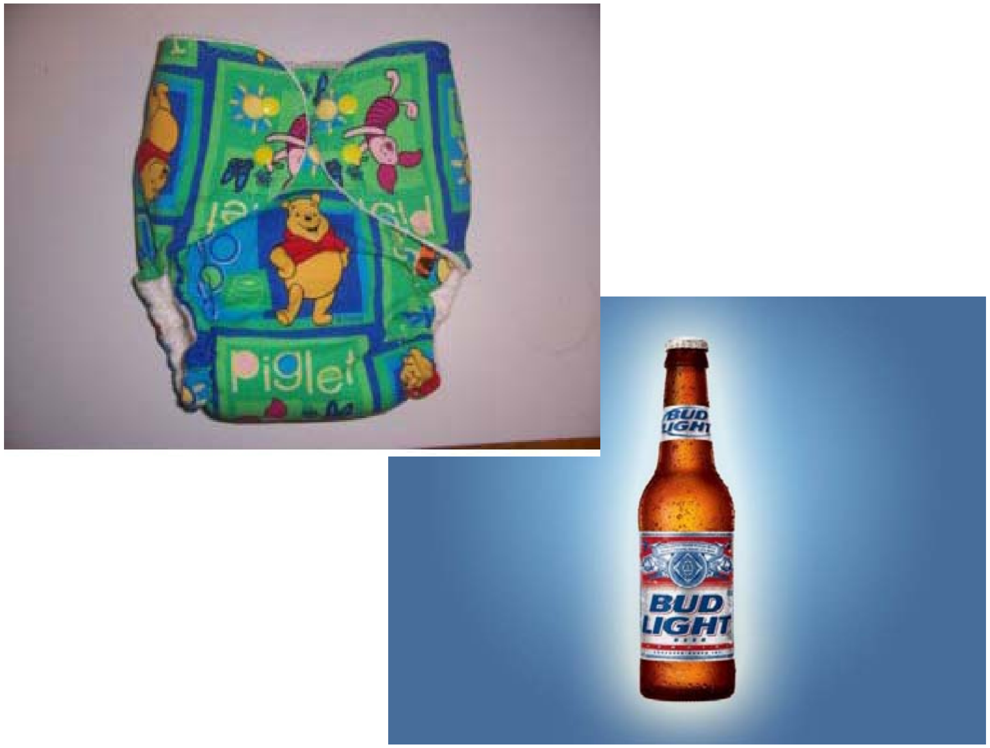
PATTERN RECOGNITION: Like credit-card companies fighting fraud by looking for unusual charges, the NSA could look for suspicious call patterns consistent with terrorists' communications.

SOCIAL-NETWORK ANALYSIS: Starting with a suspected terrorist's number, the NSA could construct a complex web of contacts by tracing who calls whom—potentially discovering new associations.

COMBINING DATABASES: Analysts could mix the data with public records or private databases—DMV records, Internet activity—to track suspicious names and activities.

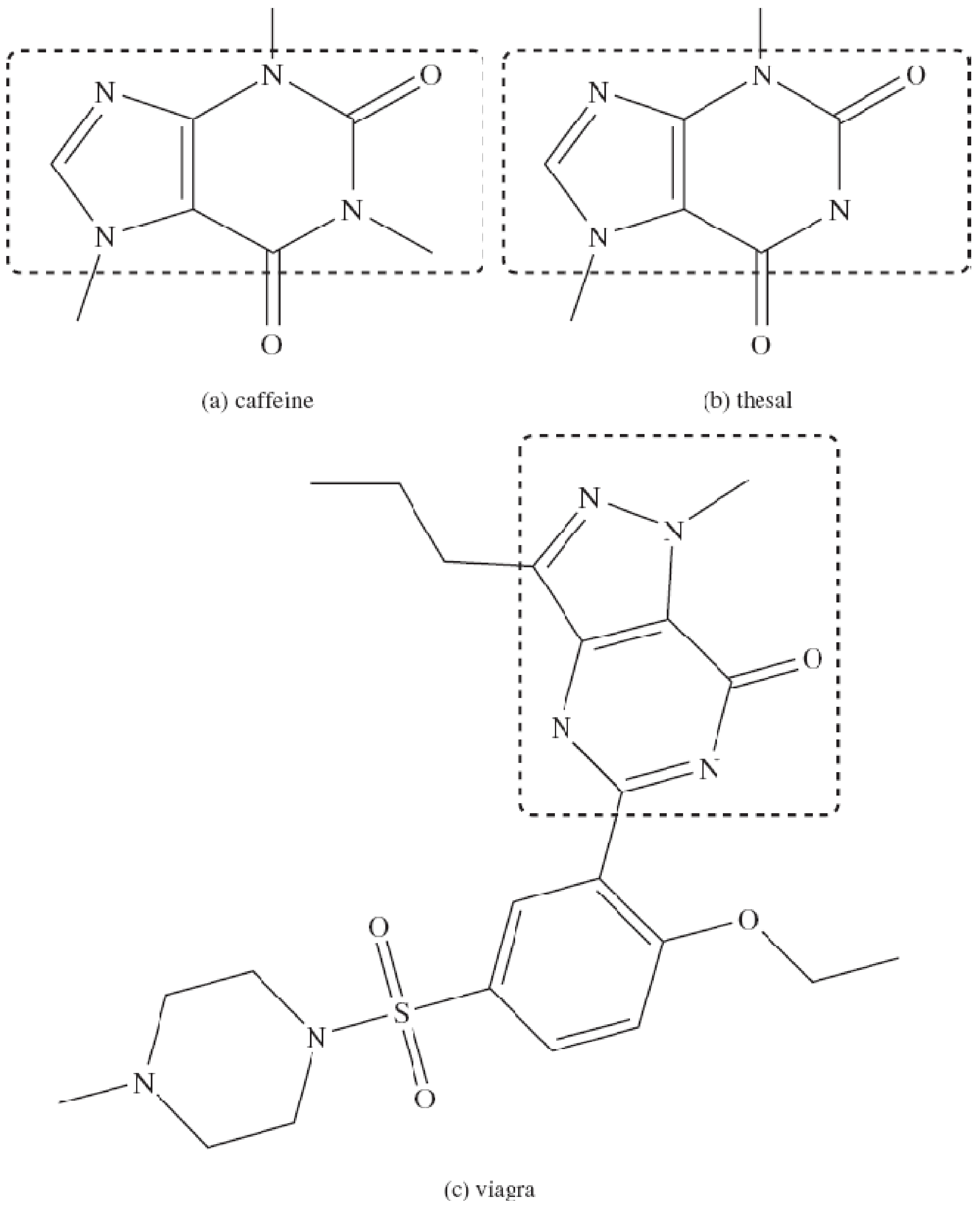
Newsweek, May 22, 2006

Market Basket Analysis



Chemistry Informatics

Figure 9.14 A Chemical database.

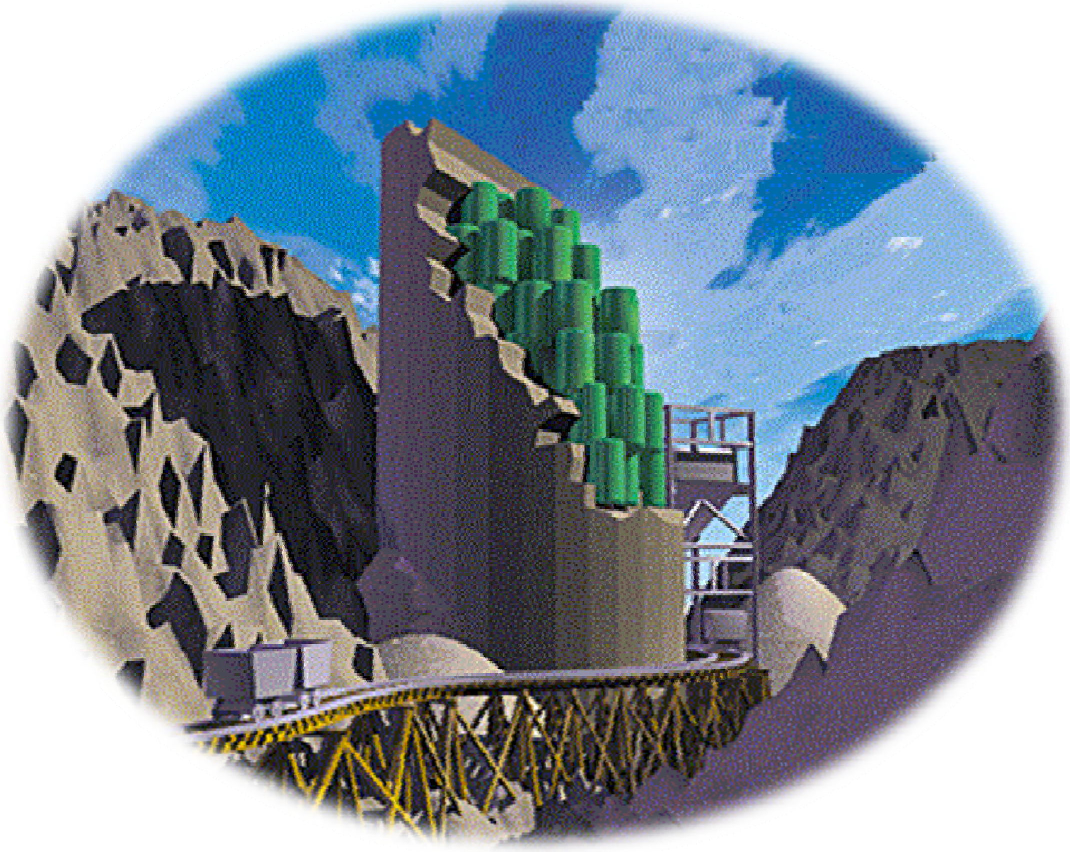


(a) caffeine

(b) thesal

(c) viagra

What is Data Mining?



Source: Cover page of *Advanced in Knowledge Discovery and Data Mining ,* edited by U. Fayyad, G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, MIT Press

How Much Information in 2003

- http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

What is Data Mining?

- Misnomer??
- Gold Mining vs. Sand (Rock) Mining
- Knowledge Discovery from Data (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archaeology
- Data dredging

Data Mining is an Interdisciplinary and Multidisciplinary Field

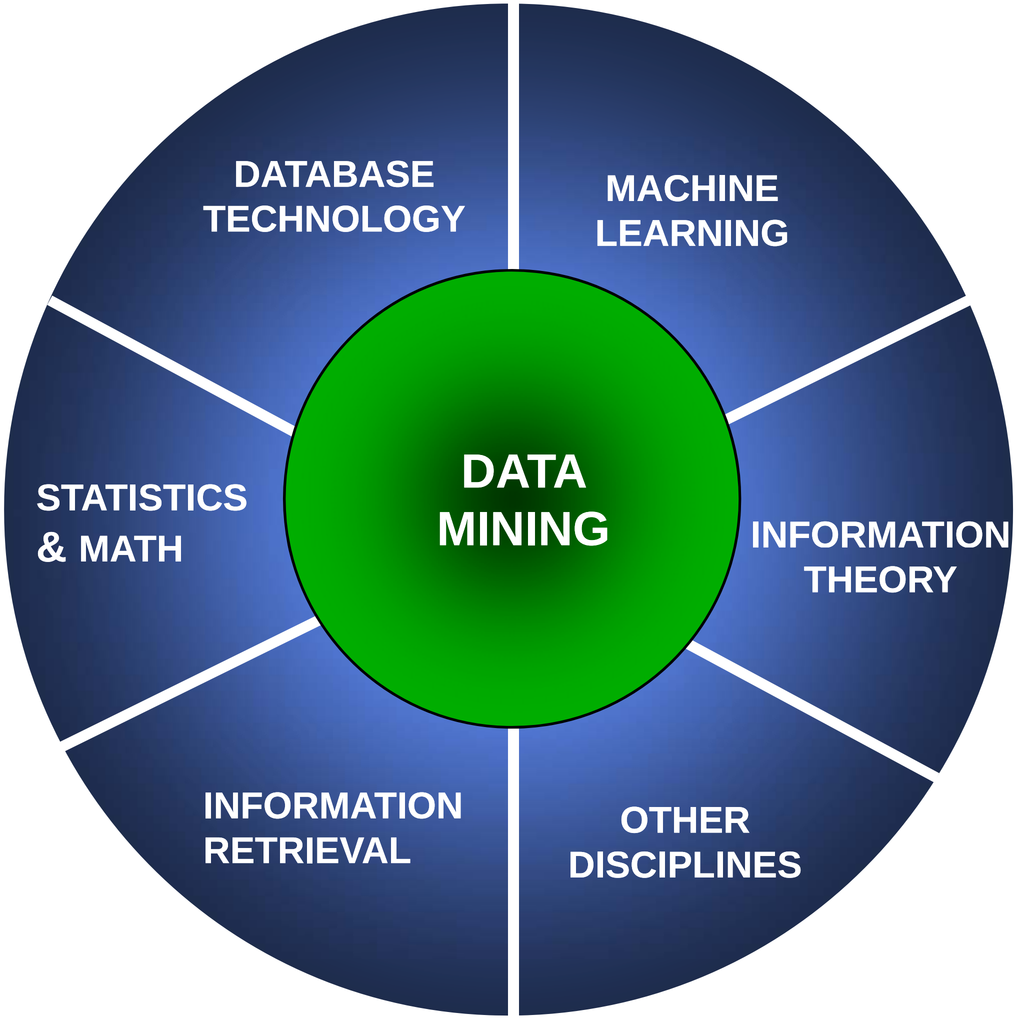
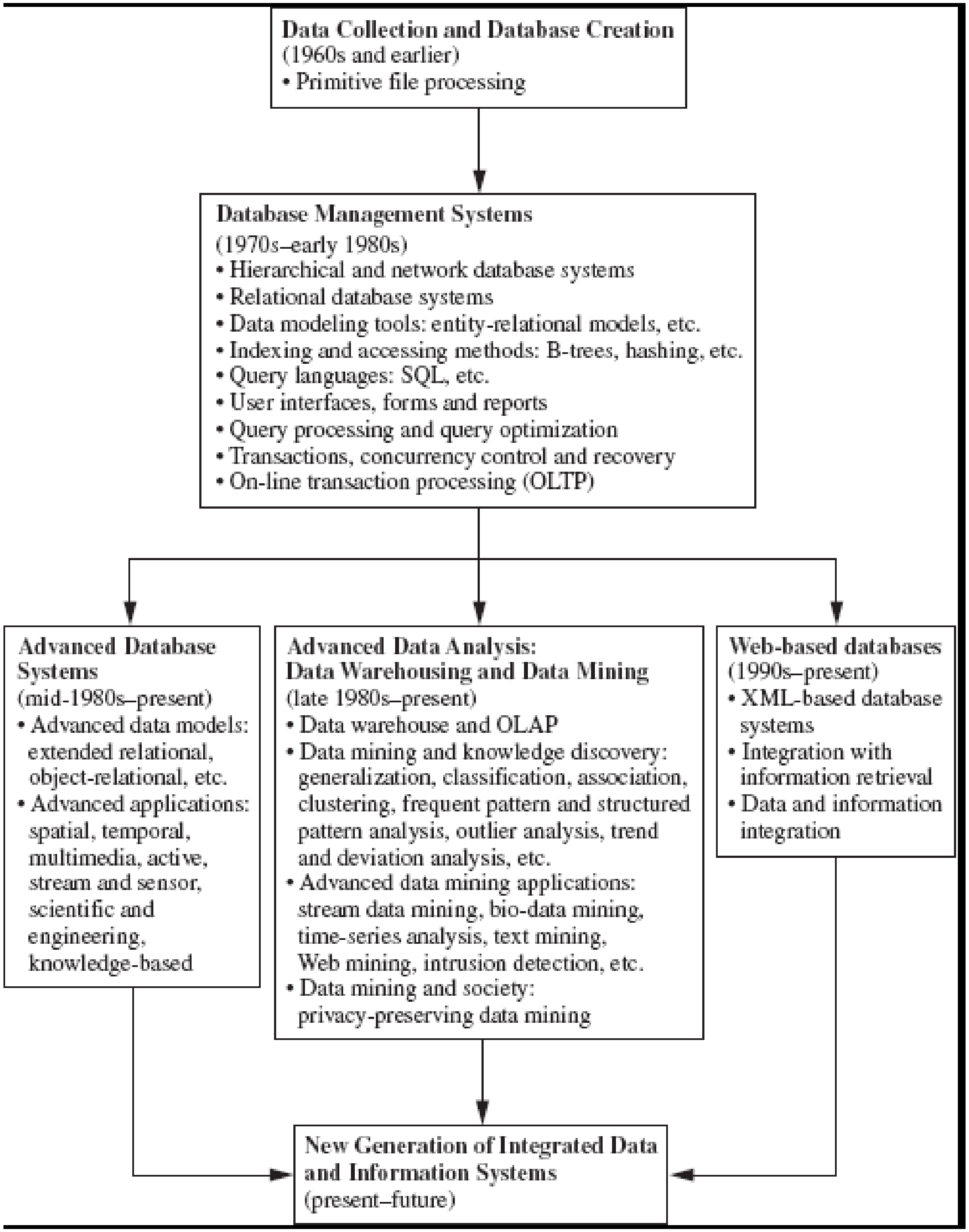


Figure 1.1 The evolution of database system technology



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
- Data modeling tools: entity-relational models, etc.
- Indexing and accessing methods: B-trees, hashing, etc.
- Query languages: SQL, etc.
- User interfaces, forms and reports
- Query processing and query optimization
- 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, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining and society: privacy-preserving data mining

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)

Data Mining is a Process of knowledge discovery

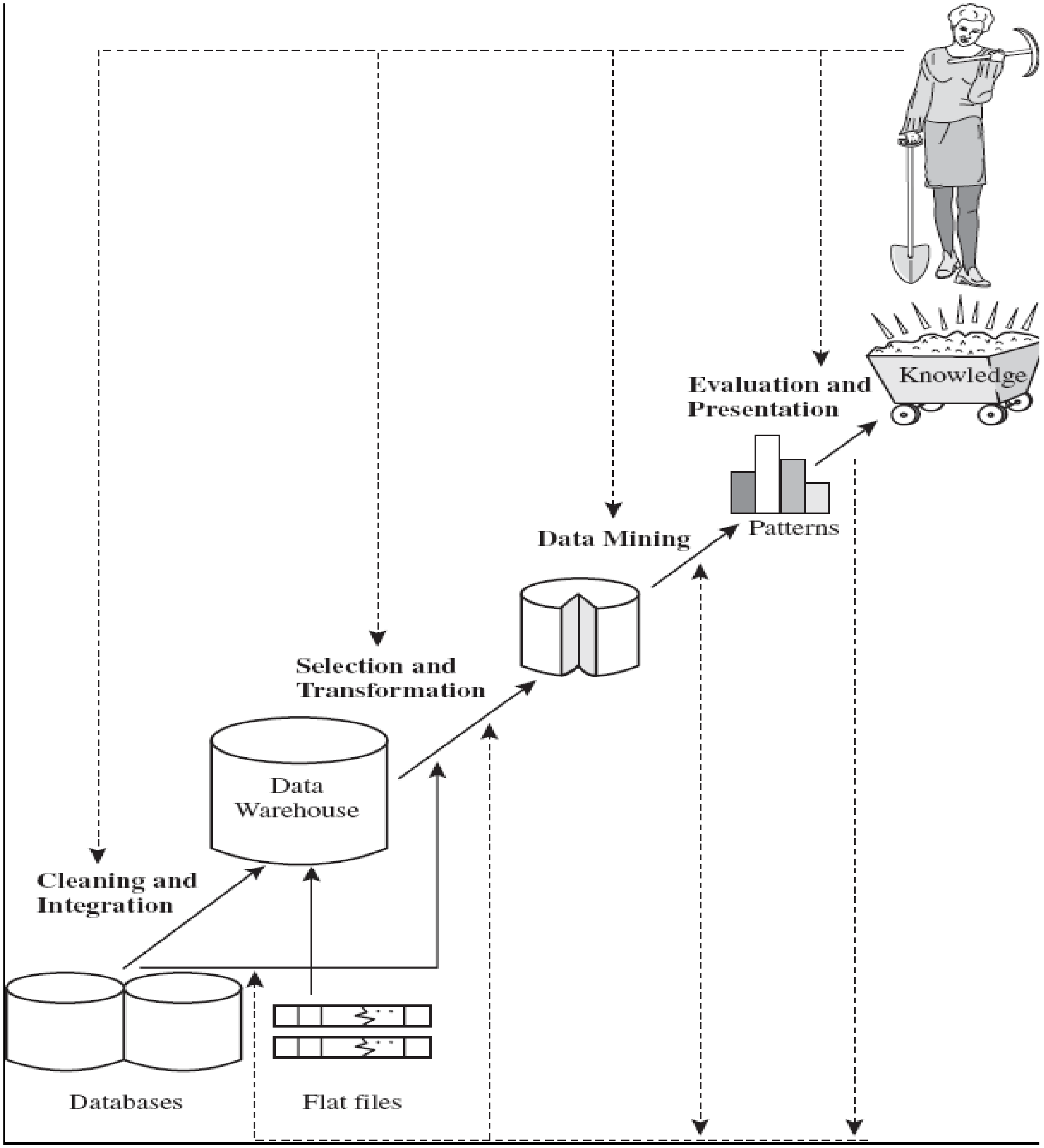


Figure 1.4 Data mining as a step in the process of knowledge discovery

Architecture of a Data Mining System

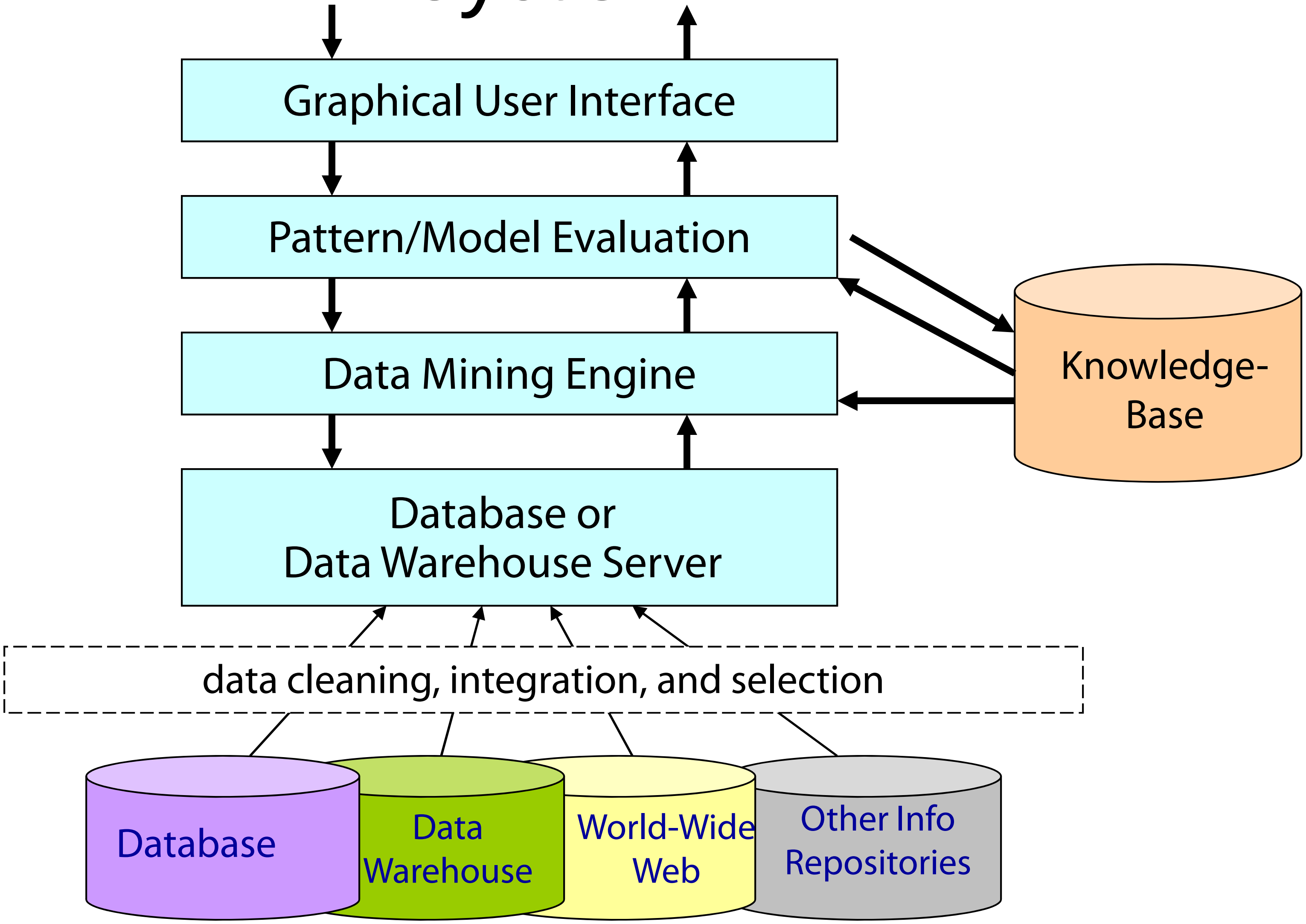
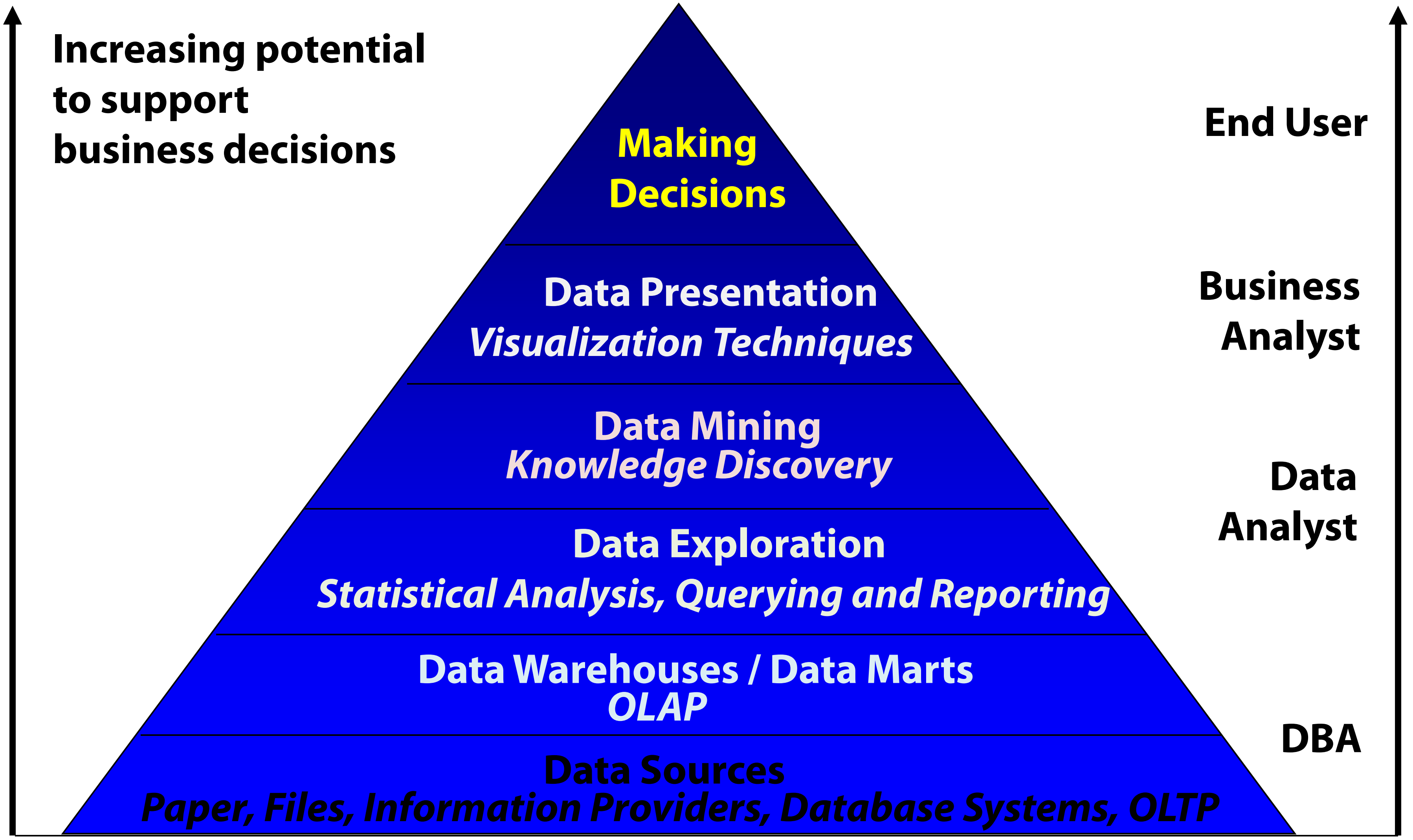


Figure 1.5 Architecture of a typical data mining system

Data Mining and Stakeholders



Increasing potential to support business decisions

Making Decisions

Data Presentation
Visualization Techniques

Data Mining
Knowledge Discovery

Data Exploration
Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts
OLAP

Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

End User

Business Analyst

Data Analyst

DBA

Data Types - Perspective on Structure

- Structured
- Semi-structured
- Unstructured

Structured Data (1)

- Data is organized in semantic entities
- Similar entities are grouped together (relations or classes)
- Entities in the same group have the same descriptions (attributes, features)

Structured Data (2)

- Descriptions for all entities in a group (schema)
- Attributes
  - Have same defined formats
  - Have predefined lengths
  - Follow same orders

Semi-structured Data (1)

- Semi-structured data are organized in semantic entities
- Similar entities are grouped together
- Entities in same group may not have same attributes

Semi-structured Data (2)

- Attributes
  - Order of attributes not necessarily important
  - Not all attributes may be required
  - Size of same attributes in a group may differ
  - Type of same attributes in a group may differ

XML

```
<bank-1>
    <customer>
              <customer_name> Hayes </customer_name>
              <customer_street> Main </customer_street>
              <customer_city> Harrison </customer_city>
              <account>
                          <account_number> A-102 </account_number>
                          <branch_name>  Perryridge </branch_name>
                          <balance> 400 </balance>
              </account>
              <account>
                 …
              </account>
    </customer>
       .
       .
</bank-1>
```

Unstructured Data (1)

- Masses of computerized data
  - which do not have a data structure
  - which is easily readable by a machine

Unstructured Data (2)

"Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data – commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations and Web pages."-- DM Review Magazine, February 2003 Issue

Data Types – Perspective on Representation

- Numeric and categorical
- Quantitative and qualitative
- Nominal and ordinal
- Static and dynamic (temporal)

Numeric and Categorical Data (1)

- Numeric data
  - – Real number data, integer number data
  - – Properties
    - Order relations (2 < 5)
    - Distance relation (d(2.3, 4.2) = 1.9)
    - Equality relation (2 = 2)

Numeric and Categorical Data (2)

- Categorical (symbolic) values
  - Equality relation
    - Blue = Blue or Rea <> Blue
  - Categorical values can be converted to a numeric values
    - Gender (male, female) → (0, 1)

Quantitative and Qualitative Data

- Quantitative data
  - Numeric values are quantitative values
  - Height, weight, salary
- Qualitative data
  - Nominal
  - Ordinal

Nominal Data

- Utility customer type (residential, commercial, industrial, governmental)
- Use different symbols, characters, and numbers
- These values can be coded alphabetically as A, B, and C, or numerically as 1, 2, and 3
- Order-less

Ordinal Data

- The rank of the student in a class
- Ordinal variables is a categorical variable for which an order relation is defined but not a distance relation
- The ordered scale need not be necessarily linear; difference between 4th and 5th students are different to that of 14th and 15th students

Static and Dynamic Data

- Static data
  - Attribute values do not change with time
- Dynamic data
  - Attribute values change with time

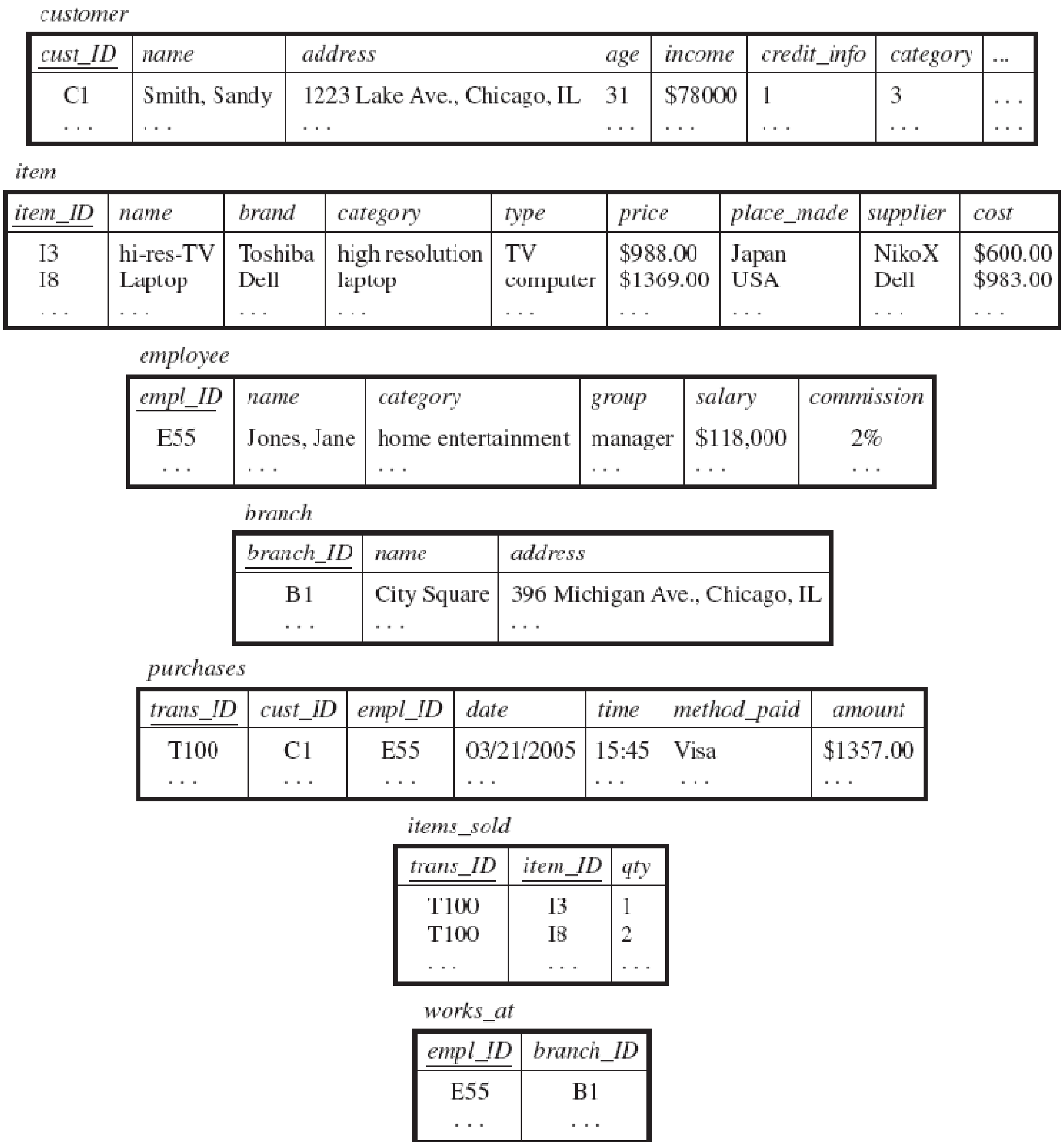
Data Repositories

- Transactional database
- Relational database
- Data warehouse
- Advanced database
- Data stream
- The World Wide Web

Transactional Database

TID	***List of item_IDs***
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Table 5.1 Transactional data for an *AllElectronics* branch



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
. . .	. . .

Figure 1.6. Fragments of Relations

From a **Relational Database** for *AllElectronics*

Data Warehouse (Mart)

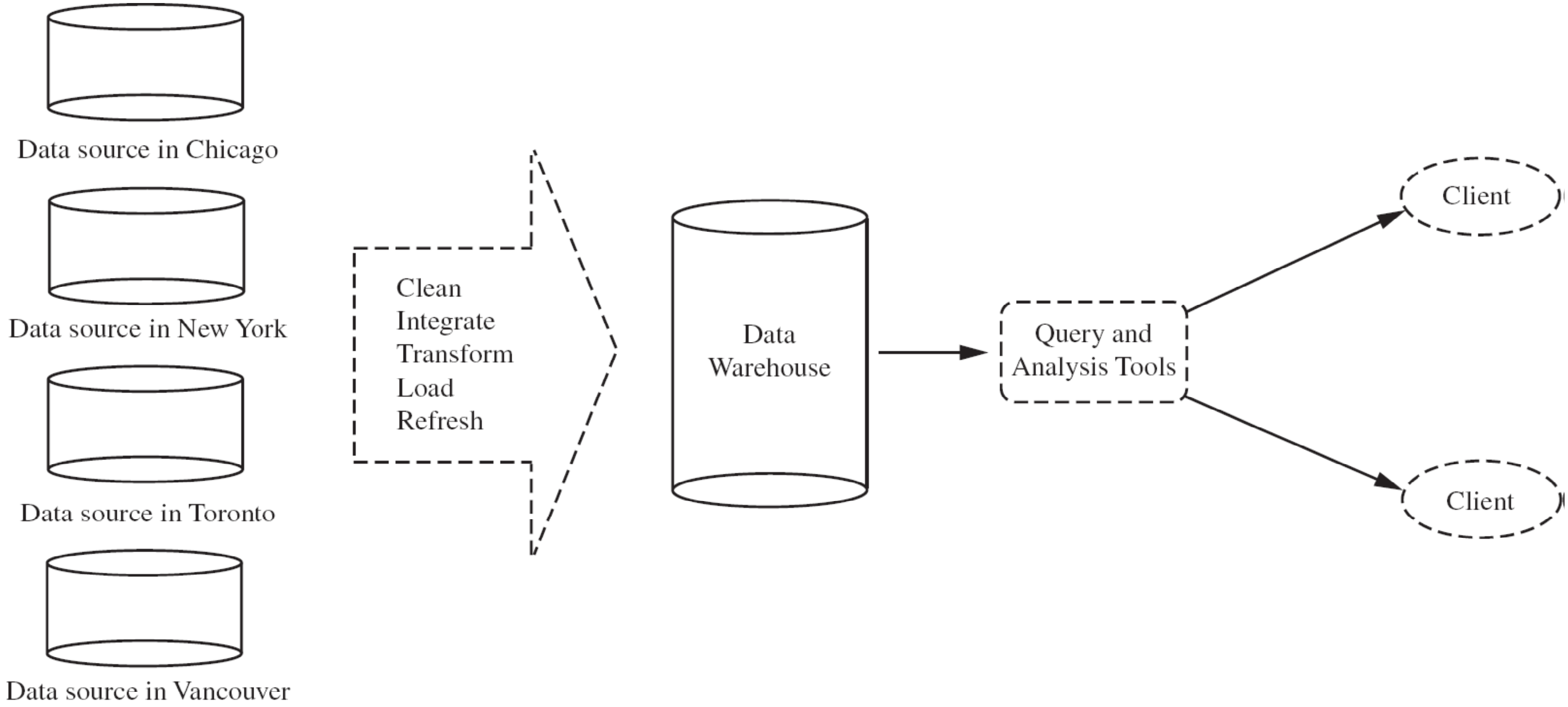
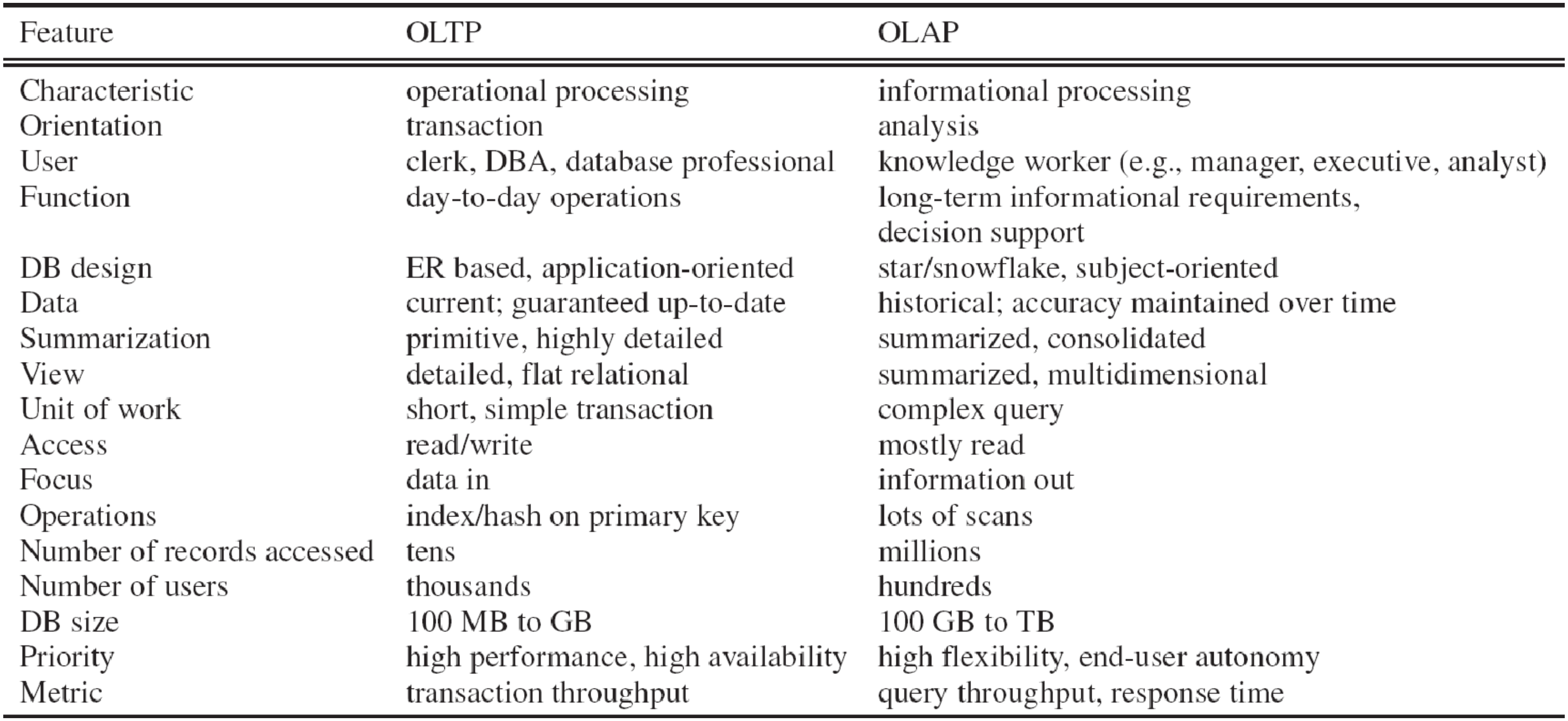


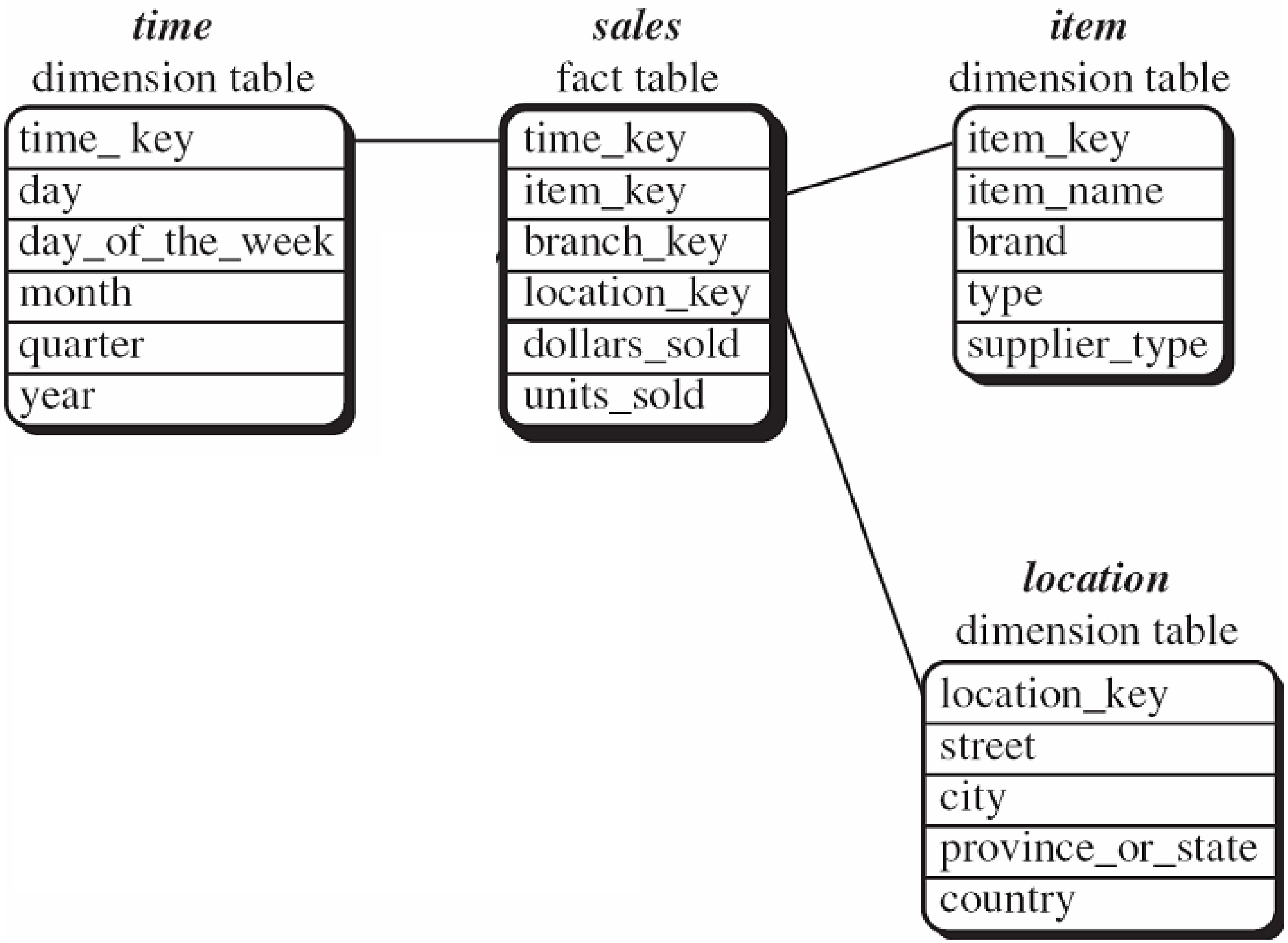
Figure 1.7 Typical framework of a data warehouse for *AllElectronics*



Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Table 3.1 Comparison between OLTP and OLAP systems

Star Schema of a Data Warehouse for Sales



time
dimension table

time_ key
day
day_of_the_week
month
quarter
year

sales
fact table

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

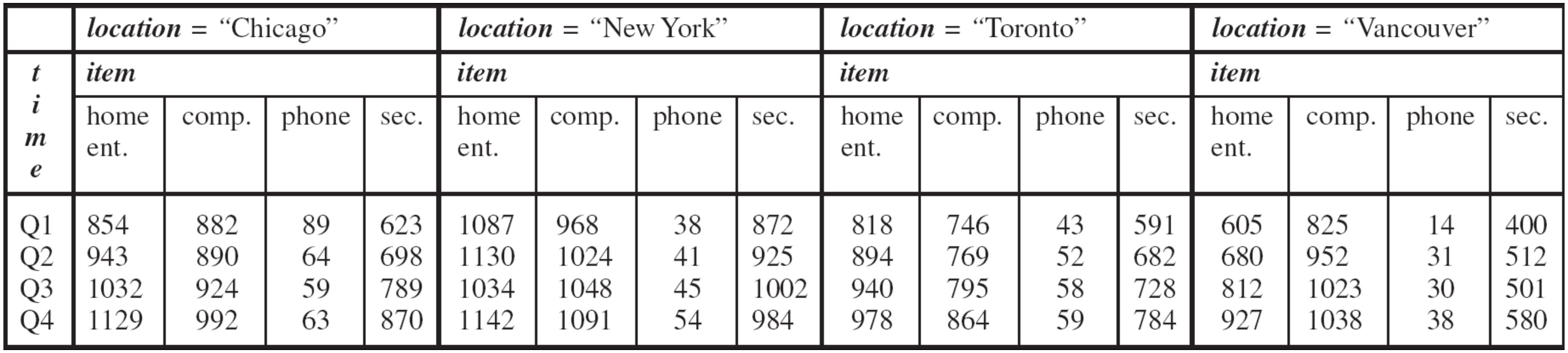
item
dimension table

item_key
item_name
brand
type
supplier_type

location
dimension table

location_key
street
city
province_or_state
country

Figure 3.4 Star schema of a data warehouse for sales



	location = "Chicago"				*location* = "New York"				*location* = "Toronto"				*location* = "Vancouver"			
time	***item***				***item***				***item***				***item***			
	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Table 3.3 A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollar_sold* (in thousands).

Data Cube for Sales

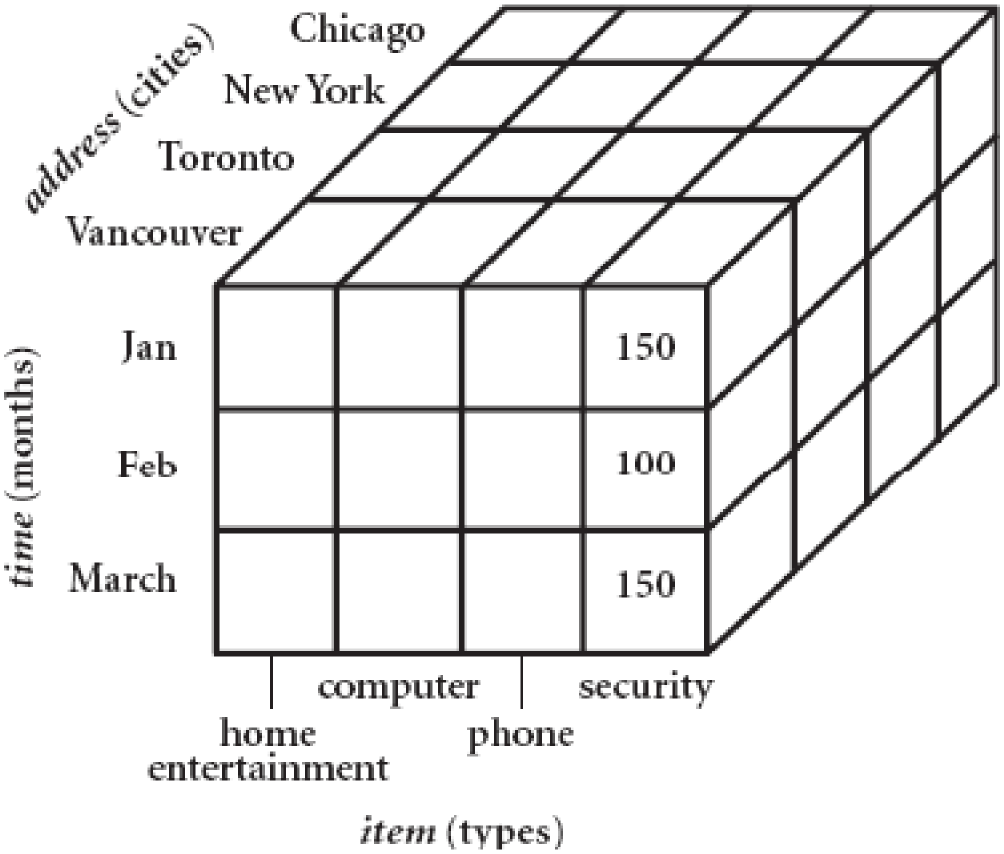
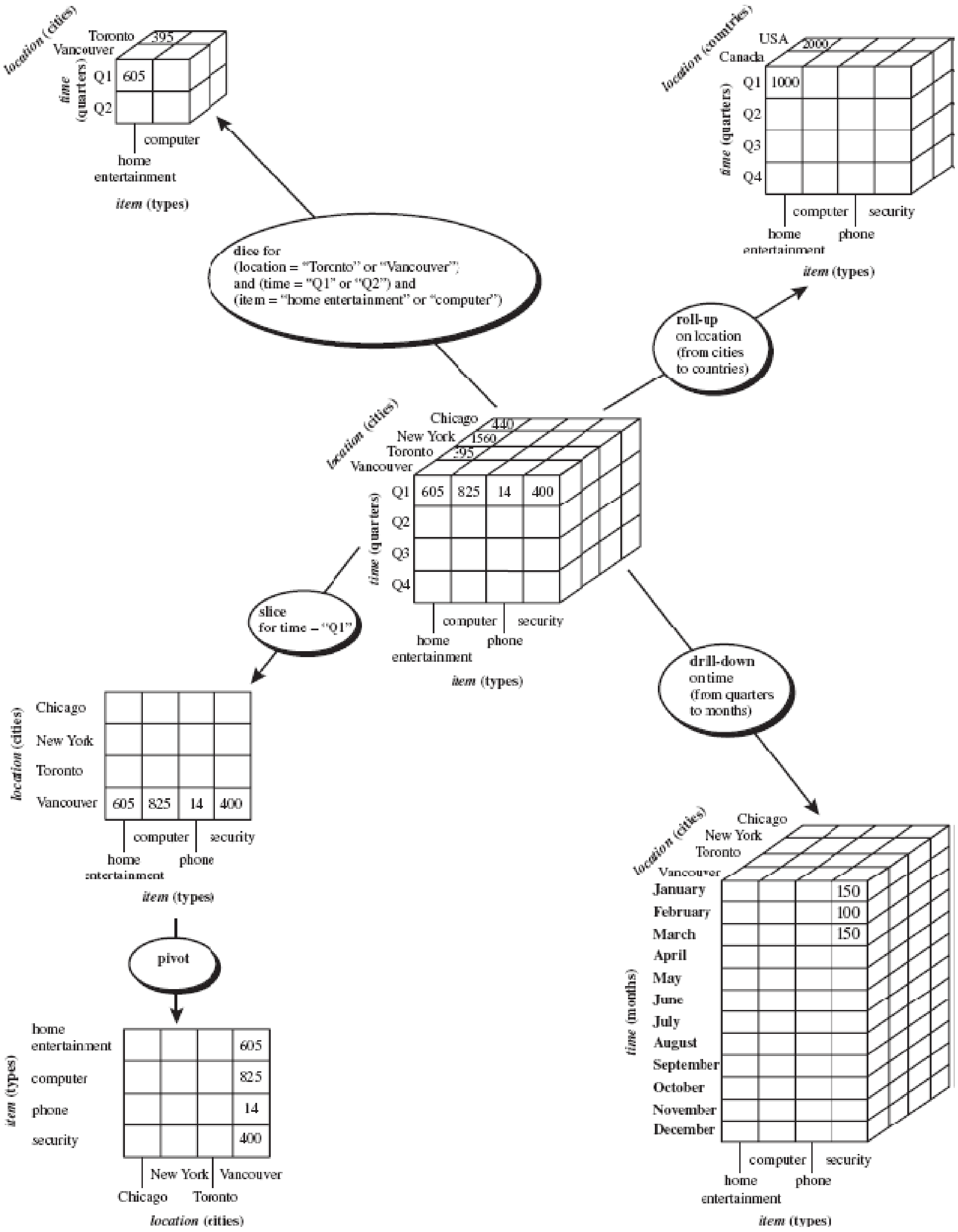


Figure 3.1 A 3-D data cube representation of the data in Table 3.3, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollar_sold* (in thousands).

Figure 3.10. Examples of Typical OLAP operations on multidimensional data cube, commonly used for data warehousing



Advanced Databases

- Object-relational databases
- Temporal databases
- Sequence databases
- Time-series databases
- Spatial databases
- Saptio-temporal databases
- Text databases
- Heterogeneous databases

Data Streams

- The features of data stream: huge or possibly infinite volume, dynamically changing, flowing in and out in a fixed order, allowing only one or a small number of scans, and demanding fast (often real-time) response time

The World Wide Web (1)

- The WWW serves a huge, distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services

The WWW (2)

- The challenges for KD
  - Size
  - Complexity
  - Dynamic
  - Diversity
  - Relevance

Lab Activities

- Introduction to R
- Organize your team
  - Each team consist of three (four) students
  - Email your team information (names and email addresses) to the instructor by the end of today’s lab session
- Read the chapter 2 of the lecture text book and do team homework assignment #1
- Read the chapters 1, 2 and 3 of the lab text book
- Brainstorm on the topic of you group project

(Team) Homework Assignment #1

- Do Example 2.1, 2.6, 2.7, and Exercise 2.18. Note that you need to use R for 2.18 (b).
- Prepare for the results of the homework assignment
- Due date
  – beginning of the lecture on Friday February 4th.

Next Week Topics

- Data types and data repositories (Section 1.3)
- Data preprocessing (Ch. 2)