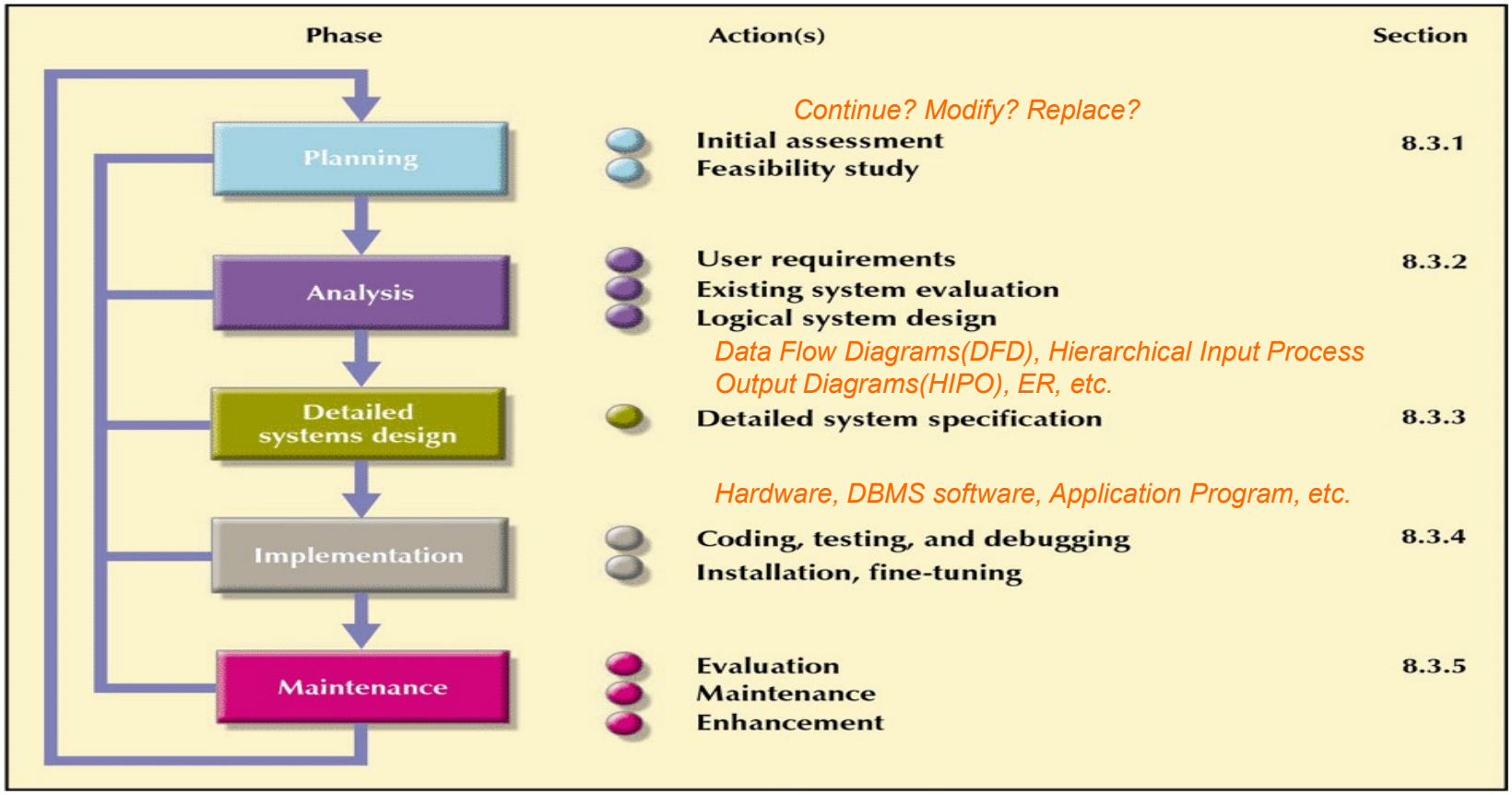
Database Design

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

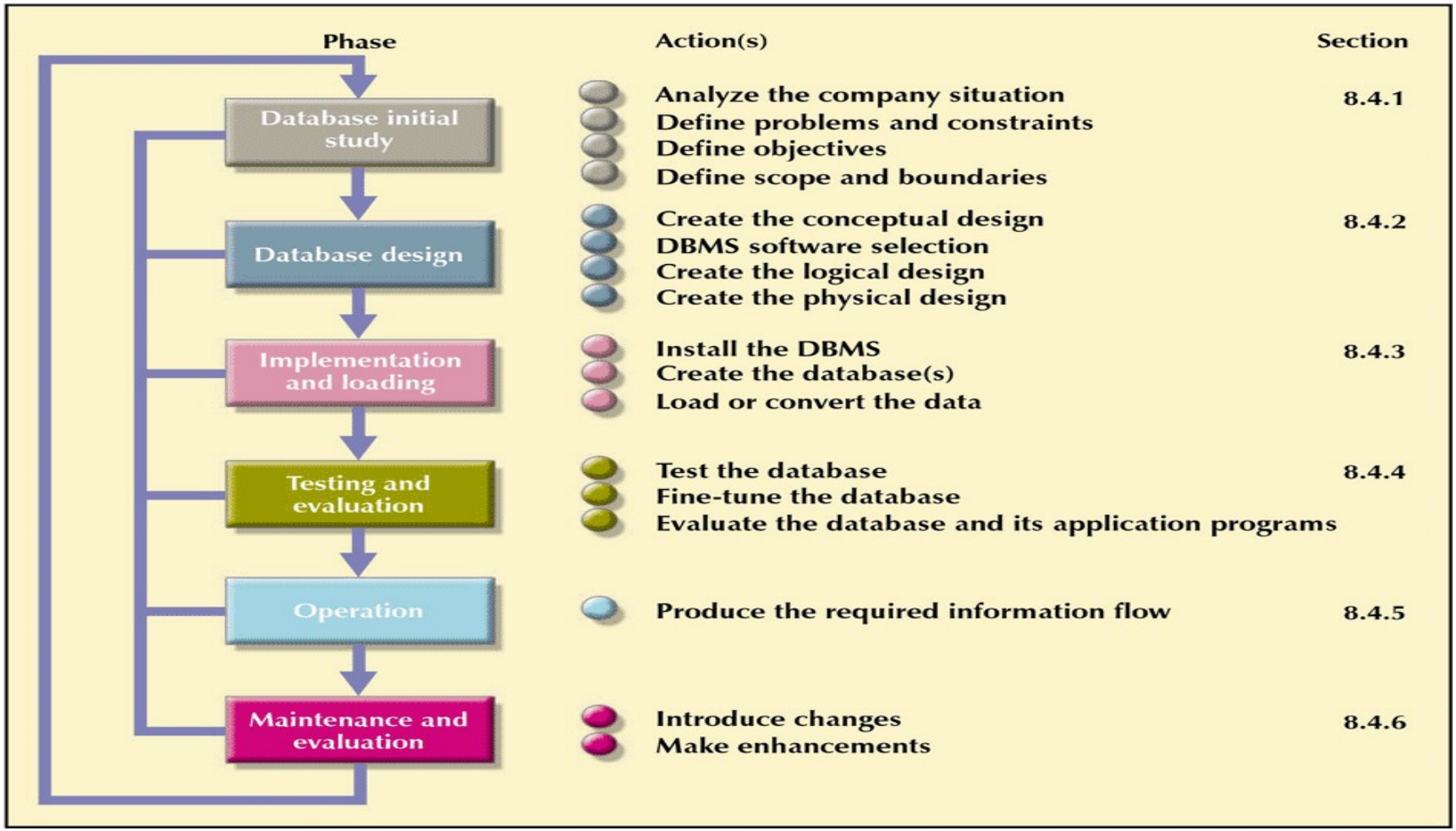
Systems Development Life Cycle (SDLC)



Corrective maintenance: system errors
Adaptive maintenance: business environment changes
Perfective maintenance: system enhancement

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

The Database Life Cycle (DBLC)



From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 1: The Database Initial Study

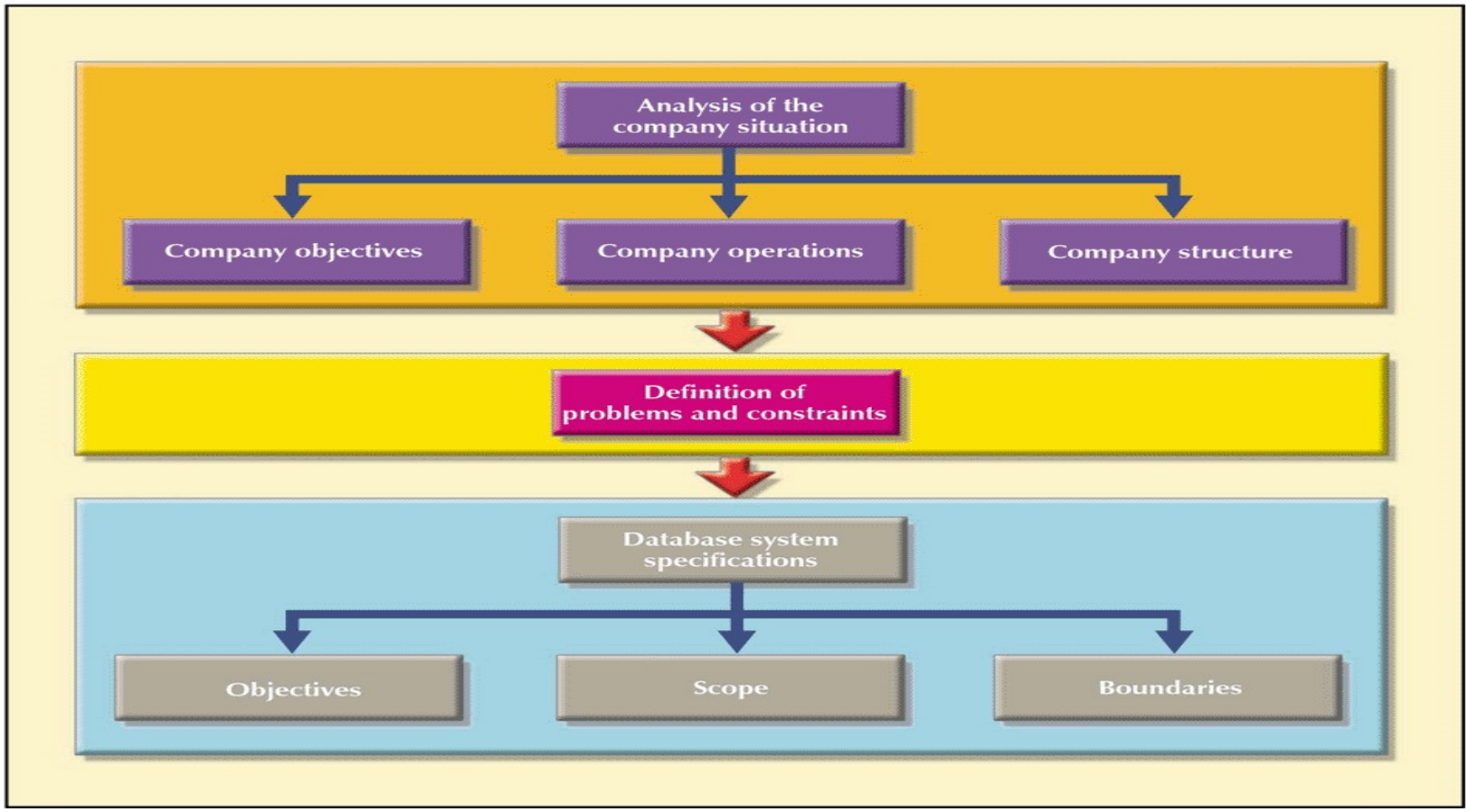
People-oriented: talk with end users

- Overall purpose:
  - Analyze the company situation *Operating environment, mission, structure, etc.*
  - Define problems and constraints
  - Define objectives *Initialization, Interface? Data sharing?*
  - Define scope and boundaries

    Hardware, software, budget, etc.
- Interactive and iterative processes required to complete the first phase of the DBLC successfully

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Summary of Activities in the Database Initial Study



From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 2: Database Design

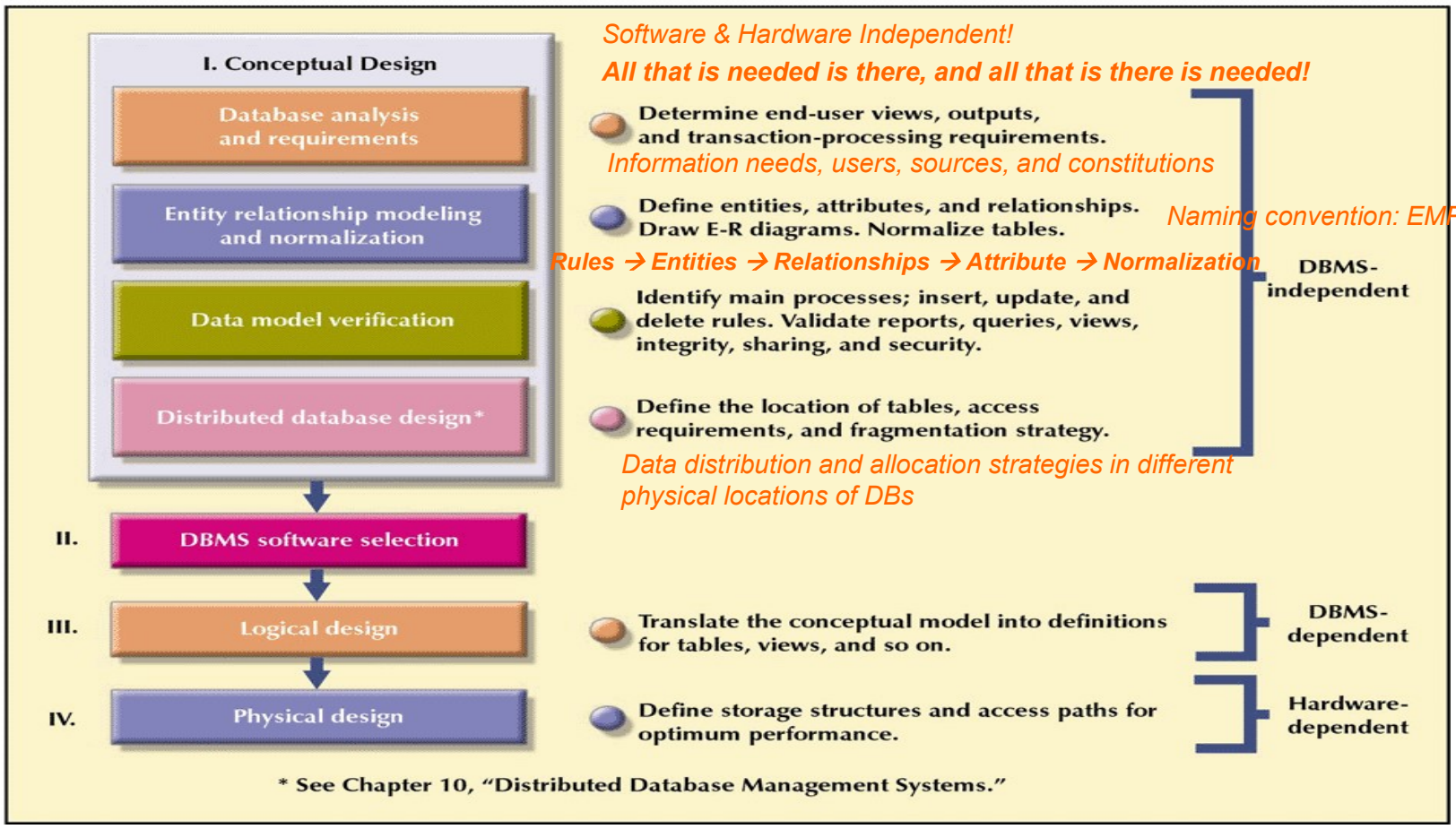
- Necessary to concentrate on the data
- Characteristics required to build database model
- Two views of data within system:
  - Business(Manager's) view of data as information source

    WHAT?
  - Designer's view of data structure, its access, and the activities required to transform the data into information

    HOW?

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Procedure Flow in the Database Design



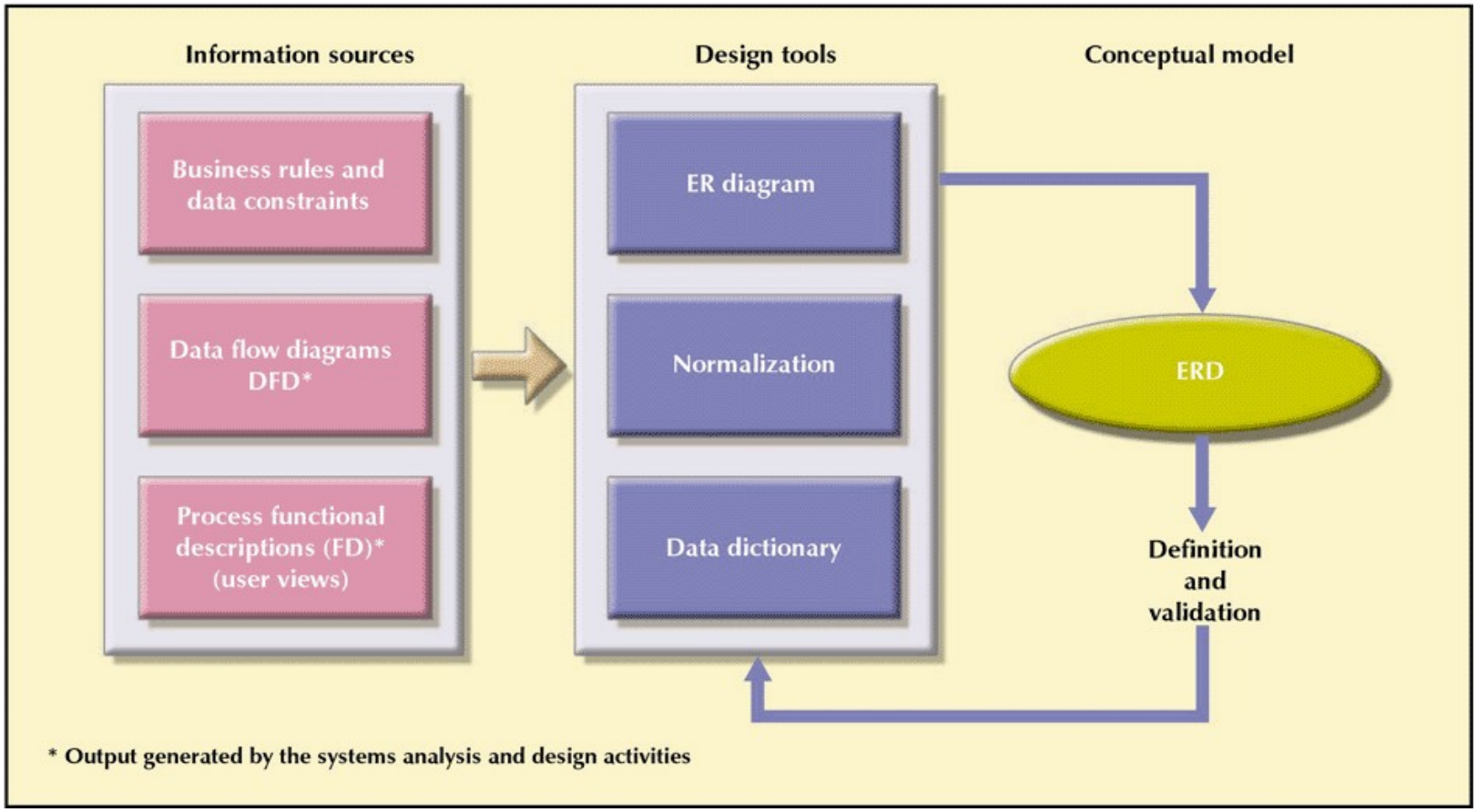
Verification VS Validation?

Verification: Is that a "right" model? → Design, Concept, Job

Validation: Is that a "correct" model? → Output

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Conceptual Design Tools and Information Sources



From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Data Model Verification

End-user data views and their required transactions, Access paths and security, Business-imposed data requirements/constraints

- Model must be verified against proposed system processes to corroborate that intended processes can be supported by database model
- Revision of original design starts with a careful reevaluation of entities, followed by a detailed examination of attributes that describe these entities
- Define design's major components as *modules:*
  - An information system component that handles a specific function

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

DBMS Software Selection

- Critical to the information system's smooth operation

- Advantages and disadvantages should be carefully studied

Cost: purchase, maintenance, operational, license, installation, training, etc.
Features and tools
Underlying model: hierarchical, network, relational, object-oriented, etc.
Portability/Interoperability
Hardware requirements

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Logical Design

- Used to translate conceptual design into internal model for a selected database management system *MySQL, Oracle, Access, MongoDB, etc.*
- Logical design is software-dependent
- Requires that all objects in the model be mapped to specific constructs used by selected database software

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

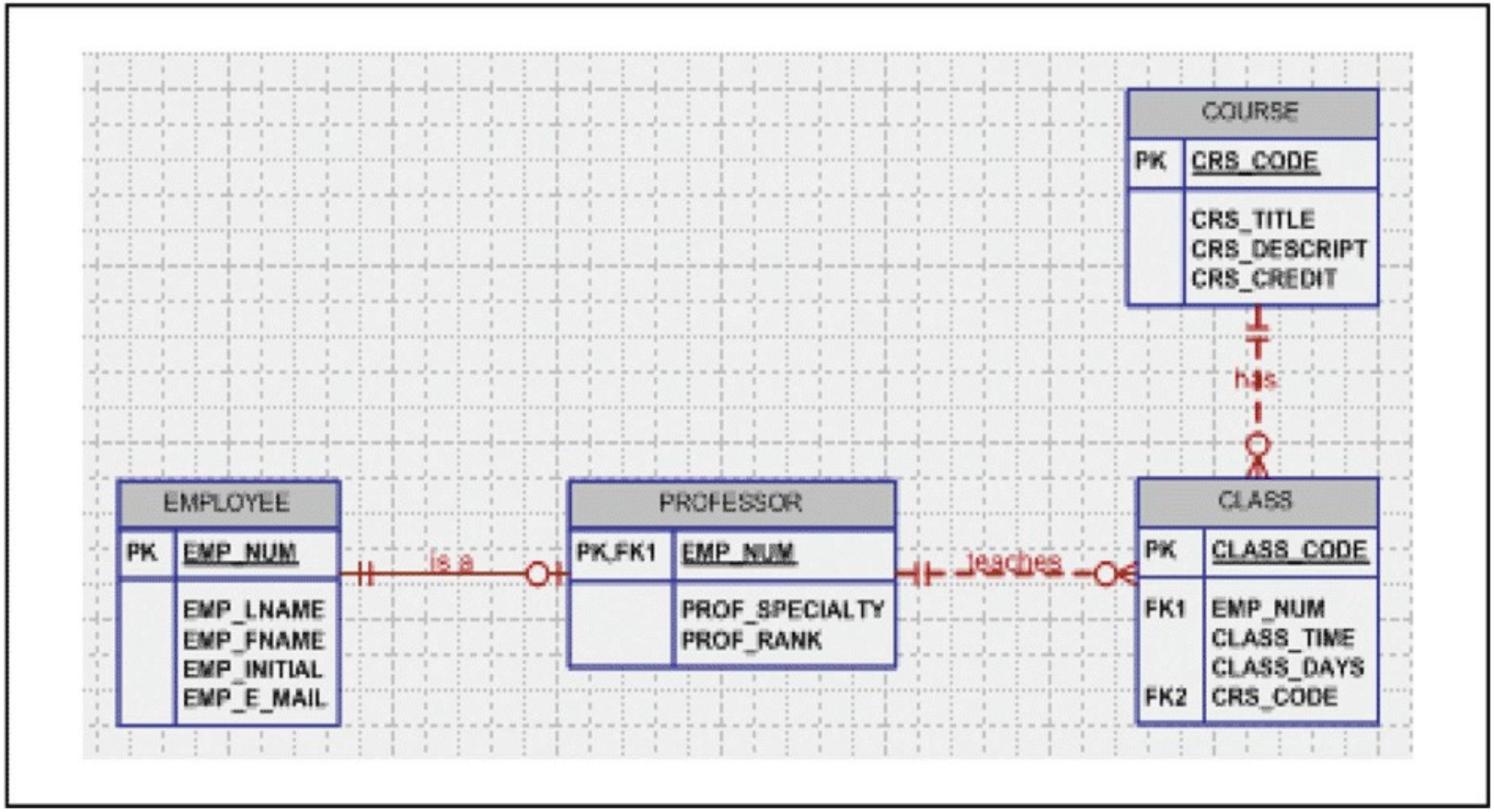
A Simple Conceptual Model

[Logical Design]

CLASS_CODE → Type: numeric, Range: low value=1000, high value=9999, Display format: 9999, Length: 4

CLASS_DAYS → Type: character, Display format: XXX, Valid entries: MWF, TTh, M, T, W, Th, F, S, Length: 3

CLASS_TIME → Type: character, Display format: 99:99(24-hour clock), Range: 06:00 to 22:00, Length: 5



From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Physical Design

- Process of selecting data storage and data access characteristics of the database
- Storage characteristics are a function of device types supported by the hardware, type of data access methods supported by system, and DBMS
- Particularly important in the older hierarchical and network models
- Becomes more complex when data are distributed at different locations

LOCATION + PERFORMANCE!

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 3: Implementation and Data Loading

- New database implementation requires the creation of special storage-related constructs to house the end-user tables and meet specified technical requirements

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Performance

- One of the most important factors in certain database implementations
- Not all DBMSs have performance-monitoring and fine-tuning tools embedded in their software
- There is no standard measurement for database performance

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Security

- Data must be protected from access by unauthorized users
- Must provide for the following:
  - Physical security *Personnel physical access*
  - Password security
  - Access rights
  - Audit trails *Check access violations. Although after-the-fact, its mere existence can discourage unauthorized use.*
  - Data encryption *Data 암호화*
  - Diskless workstations *No information download allowed*

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Backup and Recovery

- Database can be subject to data loss through unintended data deletion and power outages
- Data backup and recovery procedures
  - Create a safety valve
  - Allow database administrator to ensure availability of consistent data

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Integrity

- Enforced through proper use of primary and foreign key rules

Concurrency Control

- Feature that allows simultaneous access to a database while preserving data integrity
- Failure to maintain can quickly destroy a database's effectiveness

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 4: Testing and Evaluation

- Occurs in parallel with applications programming
- Database tools used to prototype applications
- If implementation fails to meet some of the system's evaluation criteria
  - Fine-tune specific system and DBMS configuration parameters
  - Modify the physical design
  - Modify the logical design
  - Upgrade or change the DBMS software and/or the hardware platform

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 5: Operation

- Once the database has passed the evaluation stage, it is considered operational
- Beginning of the operational phase starts the process of system evolution

Problems that could not have been foreseen during the testing phase begin to surface!

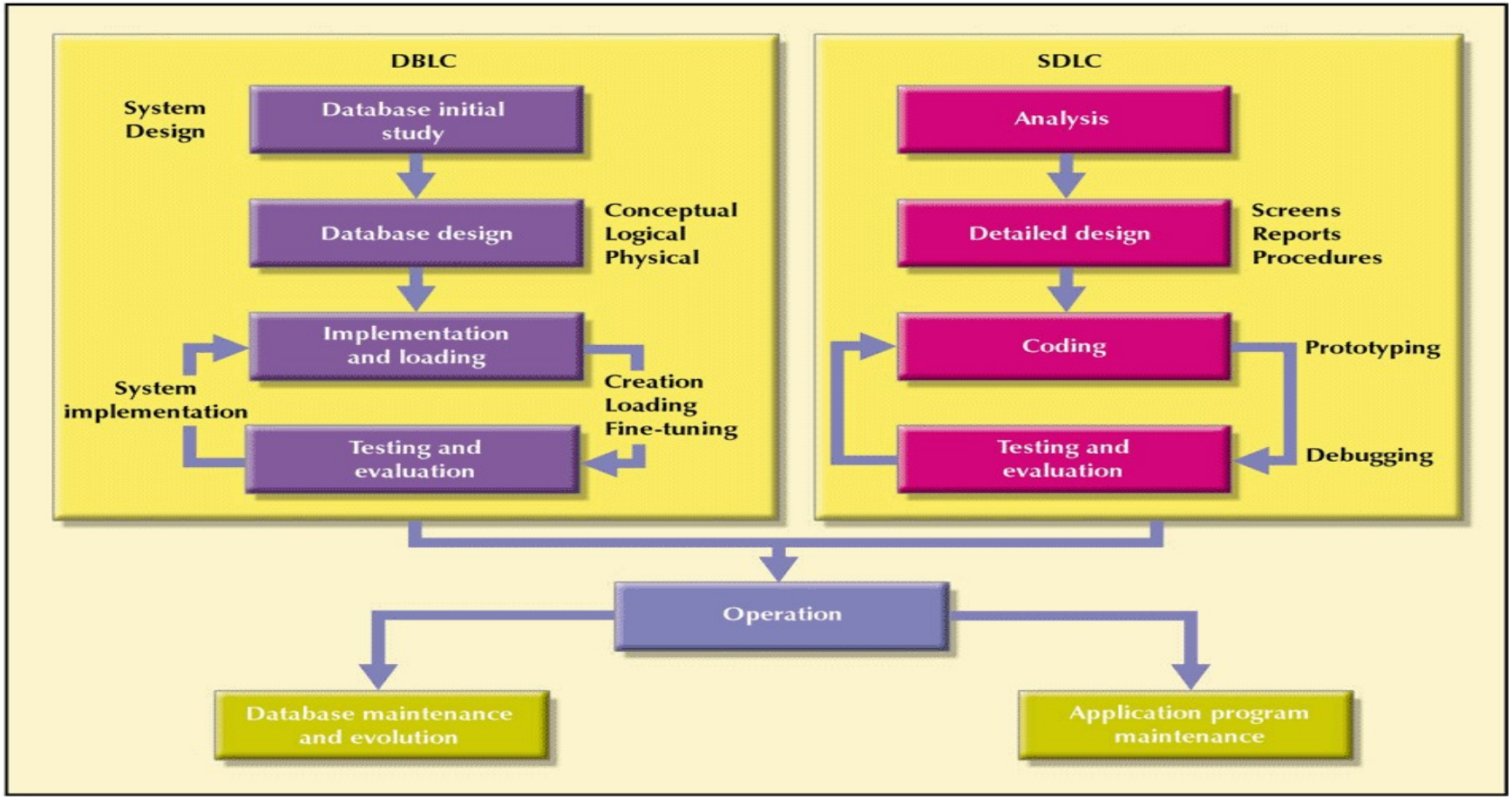
From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Phase 6: Maintenance and Evolution

- Required periodic maintenance:
  - Preventive maintenance *Action before problem - backup*
  - Corrective maintenance *Action after problem - recovery*
  - Adaptive maintenance *Enhance performance, Add new features, etc.*
- Assignment of access permissions and their maintenance for new and old users
- Generation of database access statistics
- Periodic security audits
- Periodic system-usage summaries

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Parallel Activities in the DBLC and the SDLC



From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Database Design Strategies

- Two classical approaches to database design:
  - Top-down design (entity → attribute)
    - Identifies data sets
    - Defines data elements for each of those sets
  - Bottom-up design (attribute → entity)
    - Identifies data elements (items)
    - Groups them together in data sets

Complementary! ER: maybe top-down but Normalization: bottom-up
Bottom-up: small DB with few entities, attribute, … Top-down: the number, variety, and complexity of entities, relations, …

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Centralized vs. Decentralized Design

- Database design may be based on two very different design philosophies:
  - Centralized design
    - Productive when the data component is composed of a relatively small number of objects and procedures
  - Decentralized design
    - Used when the data component of system has considerable number of entities and complex relations on which very complex operations are performed

Synonyms and homonyms (같은 철자의 다른 뜻)
Entity and entity subtypes
Conflicting object definitions

From Rob and Coronel (2004), Database Systems: Design, Implementation, and Management

Data Management: Warehousing, Analyzing, Mining, and Visualization

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Management

- Managing data is difficult for various reasons:
  - The amount of data increases exponentially with time.
  - Data are scattered throughout organizations.
  - Data are collected by many individuals using several methods.
  - External data needs to be considered in making organizational decisions.
  - Data security, quality, and integrity are critical.

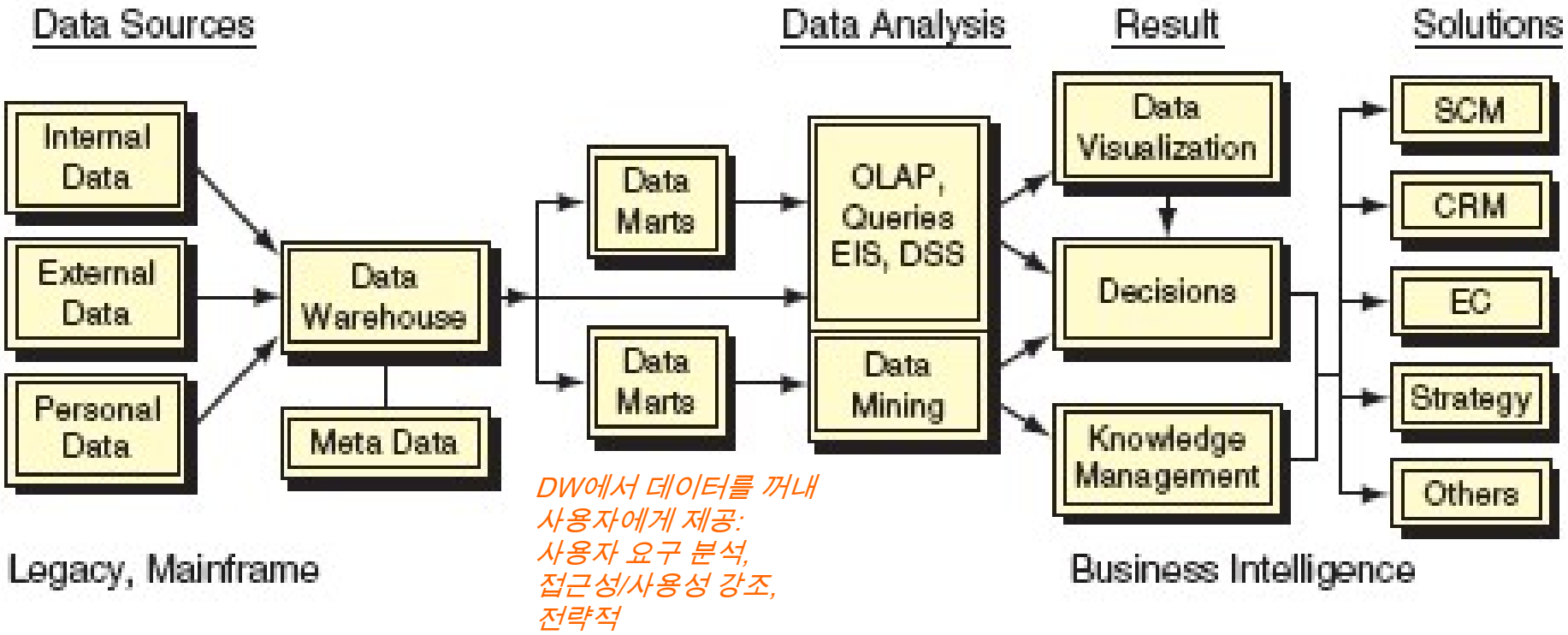
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Life Cycle Process

*ERP: 기업의 기간시스템 중 하나. 경영계획수립, 예산관리, 생산관리, 재고관리, 구매관리 등 기업의 핵심적인 기능들을 통합적으로 구현한 시스템. 예) 생산계획 수립 -> 생산에 필요한 물품들에 대한 구매계획 수립 -> 구매/생산에 따른 재고 Update 등 일관되게 관리 (*고객관리: CRM, 공급망관리: SCM, 지식경영: KMS 등)*

Legacy: 어떤 기업이나 조직이 갖고 있는 기존의 시스템, 데이터, 데이터베이스 등

OLAP: Online Analytical Processing
EIS: Executive IS
DSS: Decision Support System
SCM: Supply Chain Mgmt.
CRM: Customer Relationship Mgmt.
EC: e-Commerce



DW에서 데이터를 꺼내 사용자에게 제공:
사용자 요구 분석,
접근성/사용성 강조,
전략적

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Sources

- Internal Data Sources are usually stored in the corporate databases and are typically about people, products, services, and processes.
- Personal Data is documentation on the expertise of corporate employees usually maintained by the employee.
- External Data Sources range from commercial databases to government reports.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Methods for Collecting Raw Data

- Data collection can take place:
  - in the field
  - from individuals
  - via manual methods
    - time studies with timekeeping device
    - surveys
    - observations
    - contributions from experts
  - using instruments and sensors
  - transaction processing systems
  - via electronic transfer
  - from a web site

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Quality

Data quality (DQ) is an extremely important issue since quality determines the data's usefulness as well as the quality of the decisions based on the data.

- Intrinsic DQ: Accuracy, objectivity, believability, and reputation.
- Accessibility DQ: Accessibility and access security.
- Contextual DQ: Relevancy, value added, timeliness, completeness, amount of data.
- Representation DQ: Interpretability, ease of understanding, concise representation, consistent representation.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

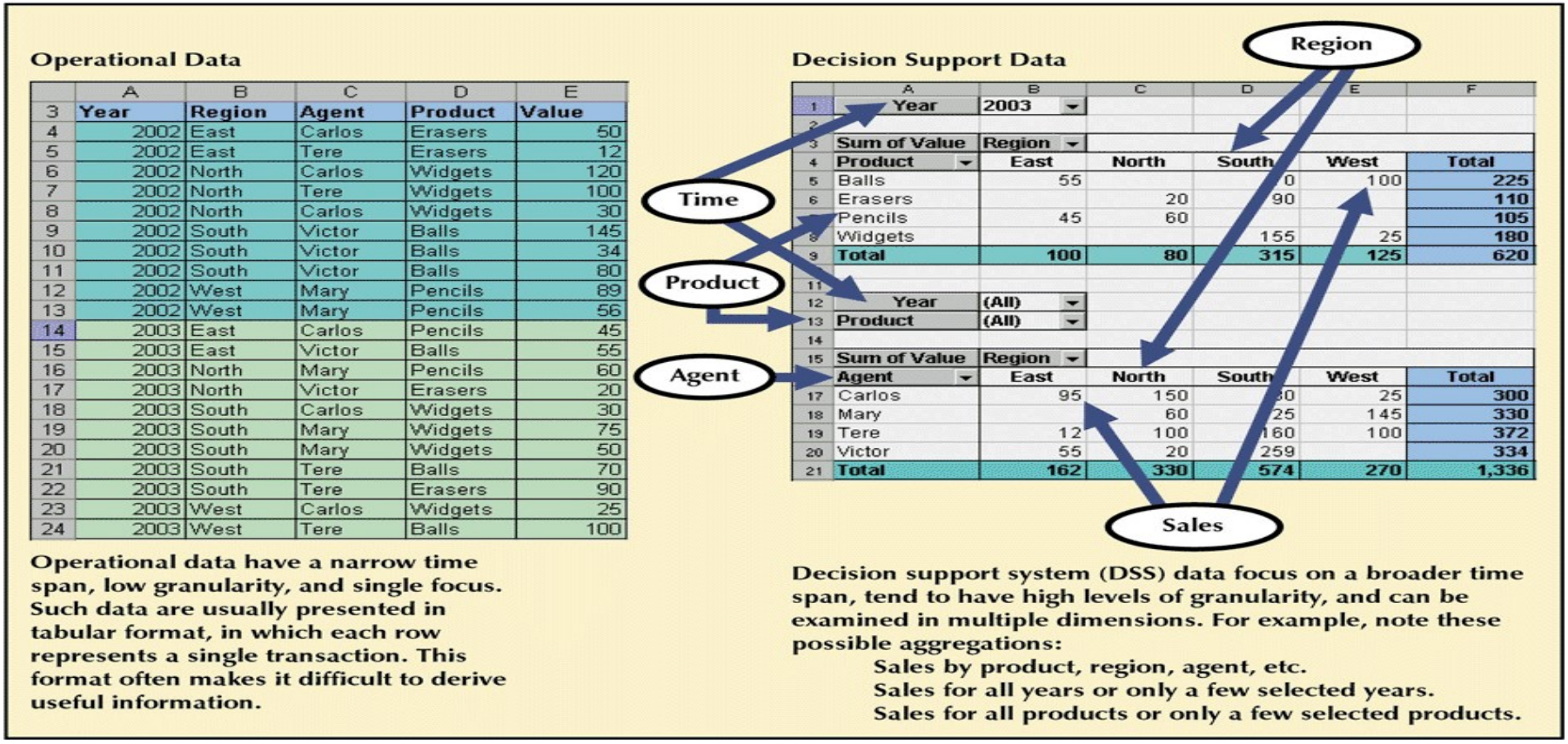
Transactional (Operational) vs. Analytical (Decision Support) Data Processing

Transactional processing takes place in operational systems that provide the organization with the capability to perform business transactions and produce transaction reports.

A supplementary activity to transaction processing is called **analytical processing,** which involves the analysis of accumulated data. These analyses place strategic information in the hands of decision makers to enhance productivity and make better decisions, leading to greater competitive advantage.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Transforming Operational Data Into Decision Support 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

	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
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

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Warehouse

DATA COLLECTION FOR SUPPORTING DECISION MAKINGS

- Defined in many different ways
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

  Nonvolatile: 데이터가 변하지 않는
- Data warehousing:
  - The process of constructing and using data warehouses

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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 usually converted.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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".

Explicit data: 의도적으로 정보가 제공되는 데이터
(예. 설문조사, 회원등록)
Implicit data: 간접적으로 축적되는 데이터
(예. Clickstreams, SNS)

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Warehouse—Time Variant

- Instagram page

"Insoo is going to lunch early with his best friend, Gyueun at Hanam Pork House near school"

Explicit data:

Insoo and Gyueun are going to eat barbeque for lunch and where they are going to eat.

Implicit data:

Insoo has a typical lunch time, Insoo and Gyueun are not vegetarians, Insoo and Gyueun went to school today, etc.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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*.

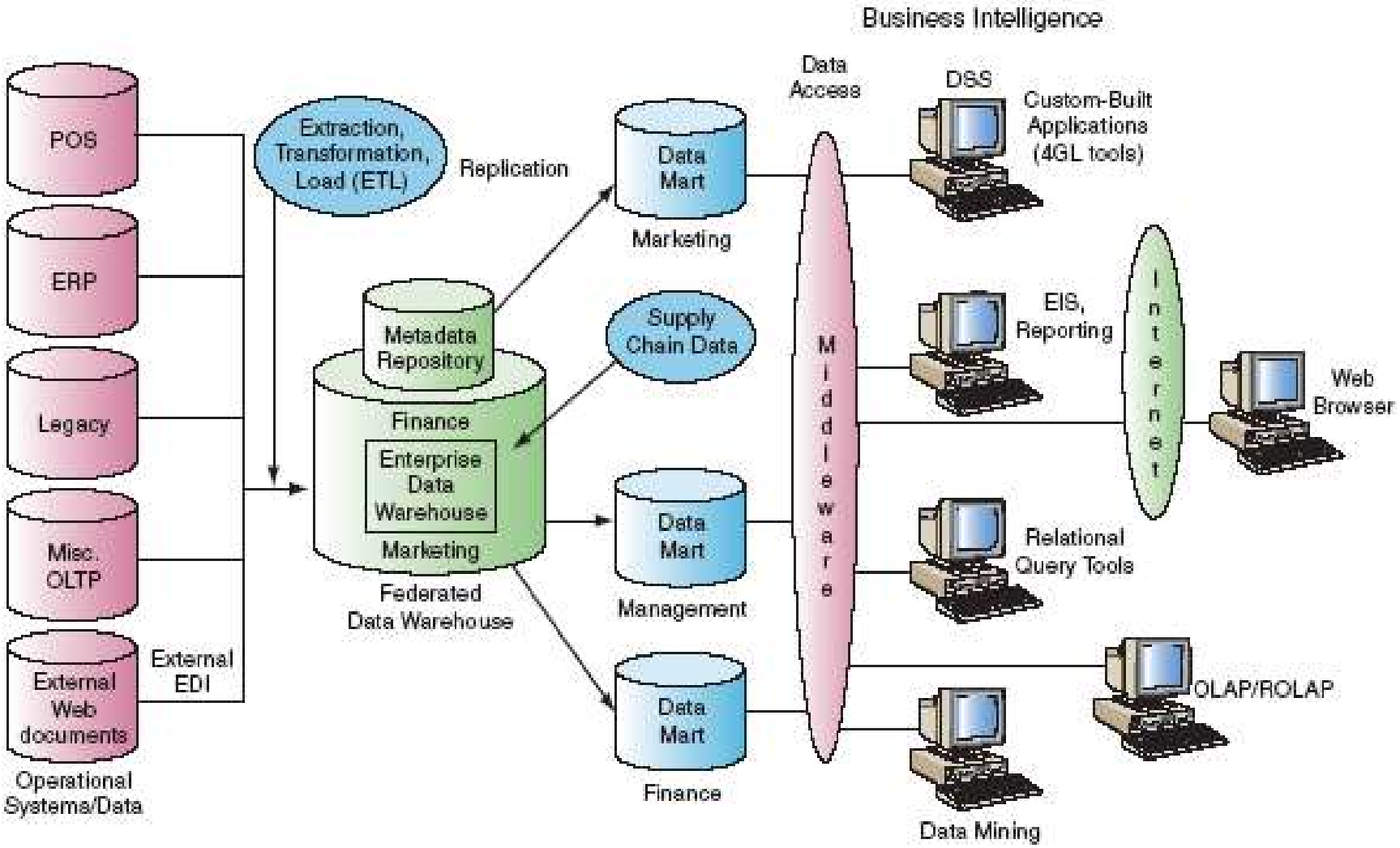
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

CHARACTERISTIC	OPERATIONAL DATABASE DATA	DATA WAREHOUSE DATA
Integrated	Similar data can have different representations or meanings. For example, Social Security numbers may be stored as ###-##-#### or as #########, and a given condition may be labeled as T/F or 0/1 or Y/N. A sales value may be shown in thousands or in millions.	Provide a unified view of all data elements with a common definition and representation for all business units.
Subject-oriented	Data are stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, and credit amounts.	Data are stored with a subject orientation that facilitates multiple views of the data and facilitates decision making. For example, sales may be recorded by product, by division, by manager, or by region.
Time-variant	Data are recorded as current transactions. For example, the sales data may be the sale of a product on a given date, such as $342.78 on 12-MAY-2004.	Data are recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.
Nonvolatile	Data updates are frequent and common. For example, an inventory amount changes with each sale. Therefore, the data environment is fluid.	Data cannot be changed. Data are added only periodically from historical systems. Once the data are properly stored, no changes are allowed. Therefore, the data environment is relatively static.

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

POS: Point of Sale
OLTP: Online Transaction Processing
ROLAP: Relational OLAP

Data Warehouse



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- OLAP Distinct features
  - Use multidimensional data analysis techniques: advanced data presentation, aggregation/consolidation/classification, computation, data modeling(what-if scenarios, impact analysis, etc.)
  - Provide advanced database support: multiple linking and queries
  - Provide easy-to-use end-user interfaces
  - Support client/server architecture

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Operational vs. Multidimensional View of Sales

Database name: Ch12_Text

Table name: DW_INVOICE

	INV_NUM	INV_DATE	CUS_NAME	INV_TOTAL
▶	2034	15-May-04	Dartonik	$1,400.00
	2035	15-May-04	Summer Lake	$1,200.00
	2036	16-May-04	Dartonik	$1,350.00
	2037	16-May-04	Summer lake	$3,100.00
	2038	16-May-04	Trydon	$400.00

Table name: DW_LINE

	INV_NUM	LINE_NUM	PROD_DESCRIPTION	LINE_PRICE	LINE_QUANTITY	LINE_AMOUNT
▶	2034	1	Optical Mouse	$45.00	20	$900.00
	2034	2	Wireless RF remote and laser pointer	$50.00	10	$500.00
	2035	1	Everlast Hard Drive, 60 GB	$200.00	6	$1,200.00
	2036	1	Optical Mouse	$45.00	30	$1,350.00
	2037	1	Optical Mouse	$45.00	10	$450.00
	2037	2	Roadster 56KB Ext. Modem	$120.00	5	$600.00
	2037	3	Everlast Hard Drive, 60 GB	$205.00	10	$2,050.00
	2038	1	NoTech Speaker Set	$50.00	8	$400.00

Multidimensional View of Sales

	Time Dimension		
Customer Dimension	**15-May-04**	**16-May-04**	**Totals**
Dartonik	**$1,400.00**	**$1,350.00**	**$2,750.00**
Summer Lake	**$1,800.00**	**$3,100.00**	**$4,900.00**
Trydon		**$400.00**	**$400.00**
Totals	**$3,200.00**	**$4,850.00**	**$8,050.00**

Sales are located in the intersection of a customer row and time column

Aggregations are provided for both dimensions

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Online Analytical Processing

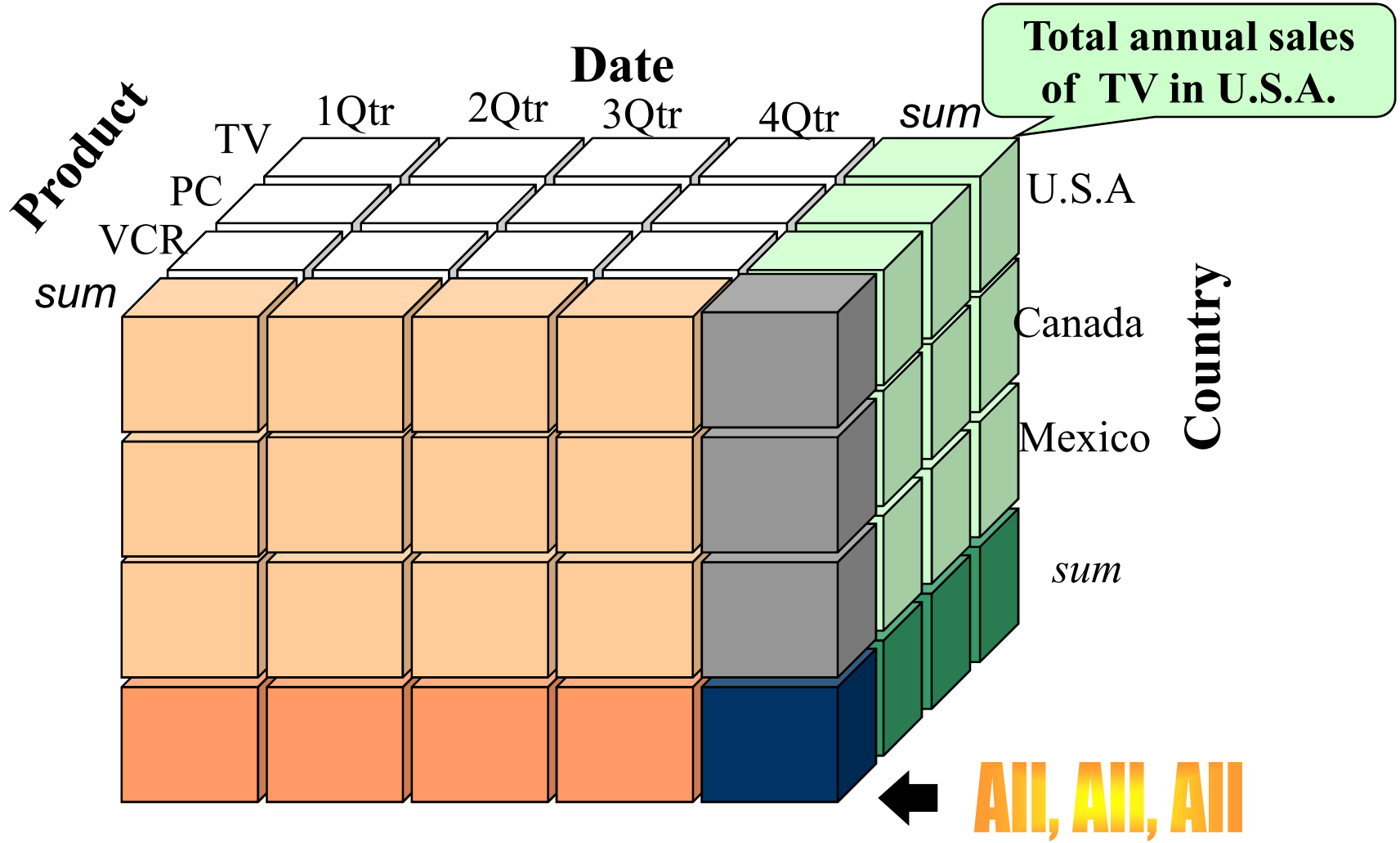
Online analytical processing (OLAP) is a set of tools that analyze and aggregate data to reflect business needs of the company. These business structures (multidimensional views of data) allow users to quickly answer business questions. OLAP is performed on Data Warehouses and Marts.

- ROLAP (Relational OLAP) is an OLAP database implemented on top of an existing multiple relational database: multidimensional data schema support within the RDBMS by normalization & queries
- MOLAP (Multidimensional OLAP) is a specialized multidimensional data store such as a Data Cube. The multidimensional view is physically stored in specialized data files.

Data Cube를 만들고 접근: 3D plotting of data, static, not subject to change, cannot be created by ad-hoc queries, faster

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

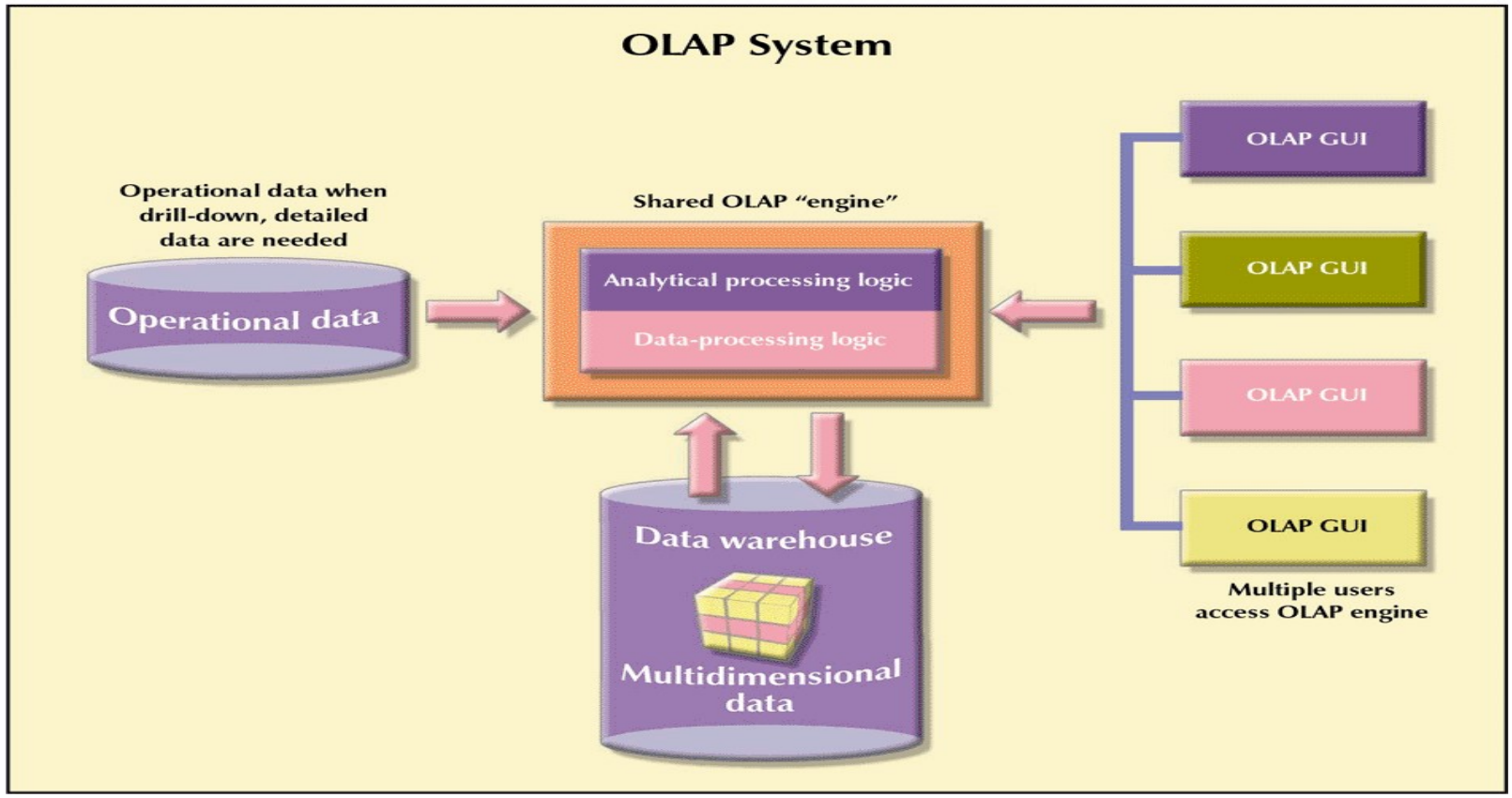
A Sample Data Cube



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

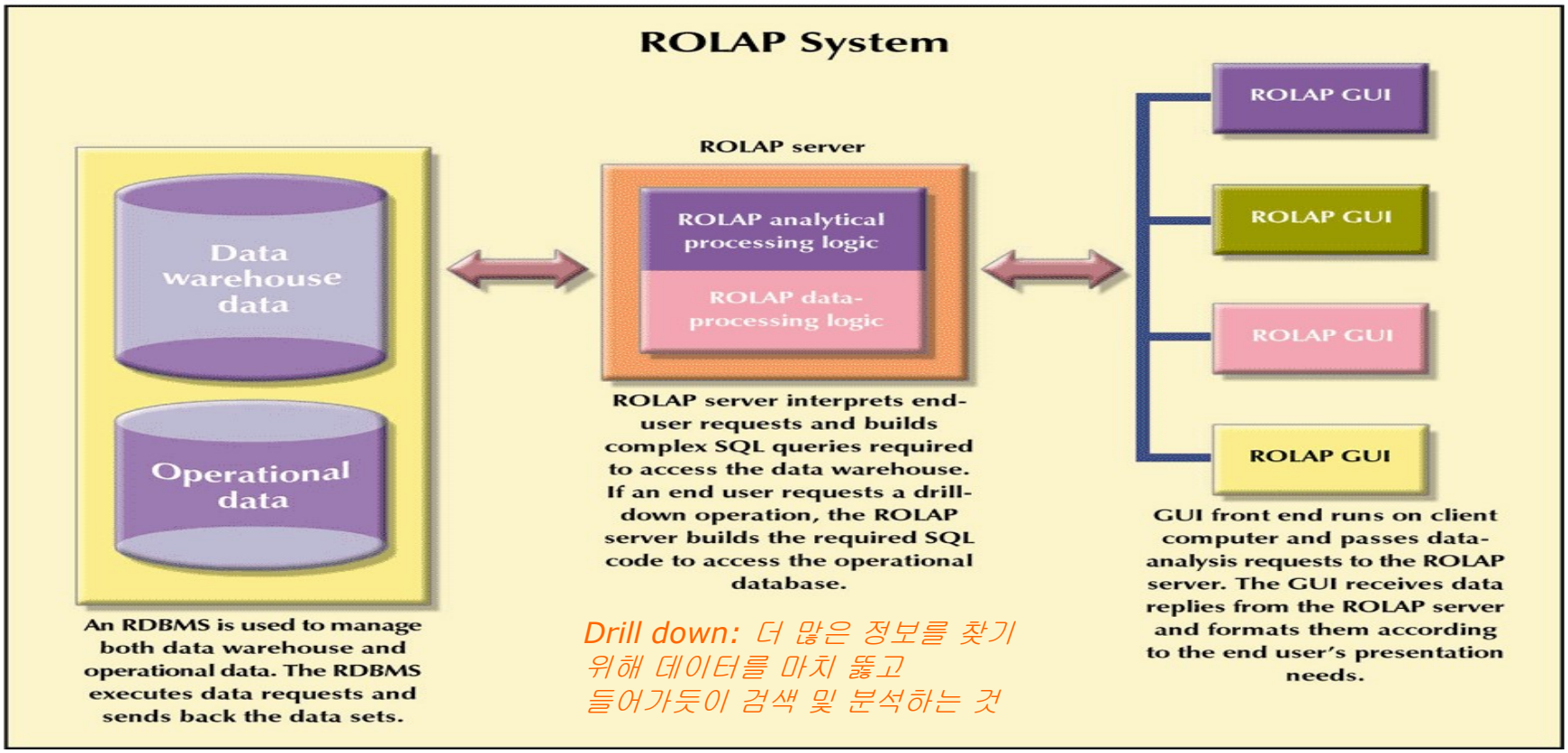
OLAP System

Front end through which end users access and analyzed
Can be directly or indirectly linked to Operational data → Possible to extracts data from an operational database and then stores it in a multidimensional structure for further data analysis (similarly acts as Data Mart)



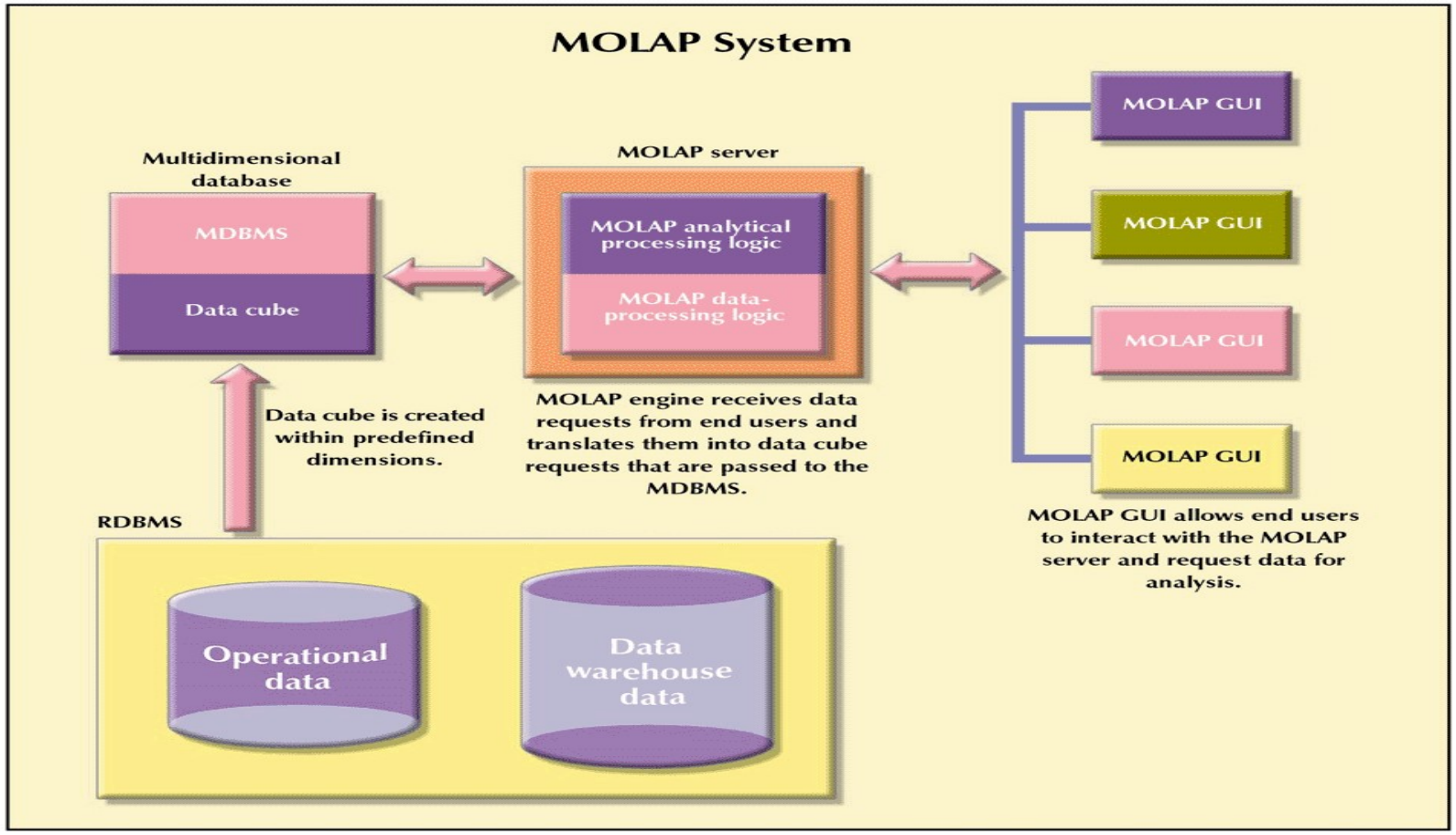
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Typical ROLAP Architecture



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Typical MOLAP Architecture

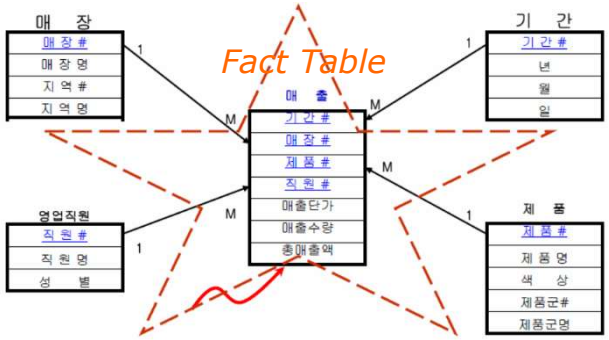


From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Relational vs. Multidimensional OLAP

Dimensional Table

Fact Table



INTEGRATED!!

Proprietary: 상표/특허

CHARACTERISTIC	ROLAP	MOLAP
Schema	Uses star schema Additional dimensions can be added dynamically	Uses data cubes Additional dimensions require re-creation of the data cube
Database size	Medium to large	Small to medium
Architecture	Client/server Standards-based Open	Client/server Proprietary
Access	Supports ad hoc requests Unlimited dimensions	Limited to predefined dimensions
Resources	High	Very high
Flexibility	High	Low
Scalability	High	Low
Speed	Good with small data sets; average for medium to large data sets	Faster for small to medium data sets; average for large data sets

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

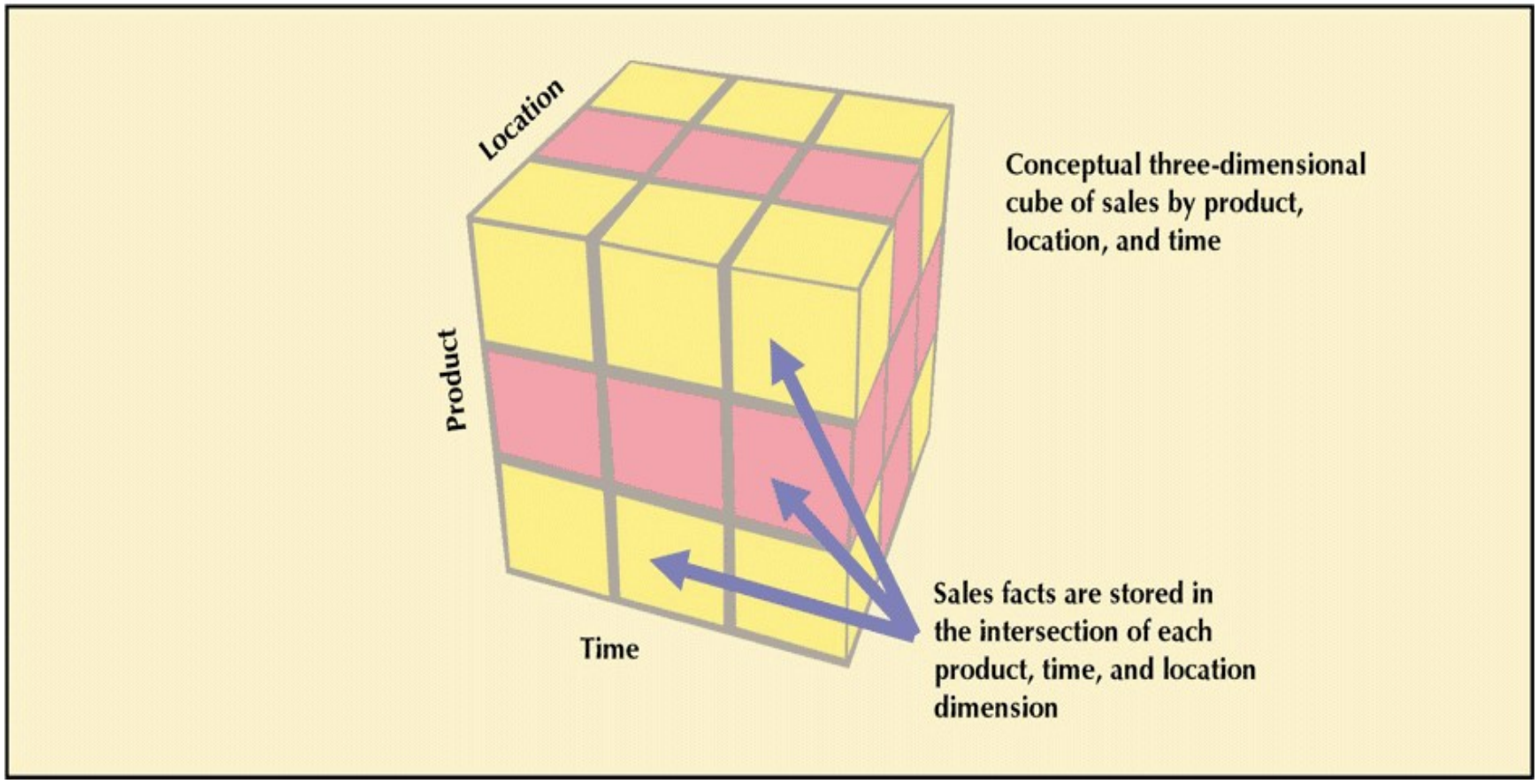
The Data Cube

Multidimensional databases are specialized data stores that organize facts by dimensions, such as geographical region, product line, salesperson, time. The data in these databases are usually preprocessed and stored in *data cubes.*

- One intersection might be the quantities of a product sold by specific retail locations during certain time periods.
- Another matrix might be sales volume by department, by day, by month, by year for a specific region
- Cubes provide faster:
  - Queries
  - Slices and Dices of the information
  - Rollups (하위에서 상위로 쌓아감)
  - Drill Downs (상위에서 하위로 파감)

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Three-Dimensional View of Sales



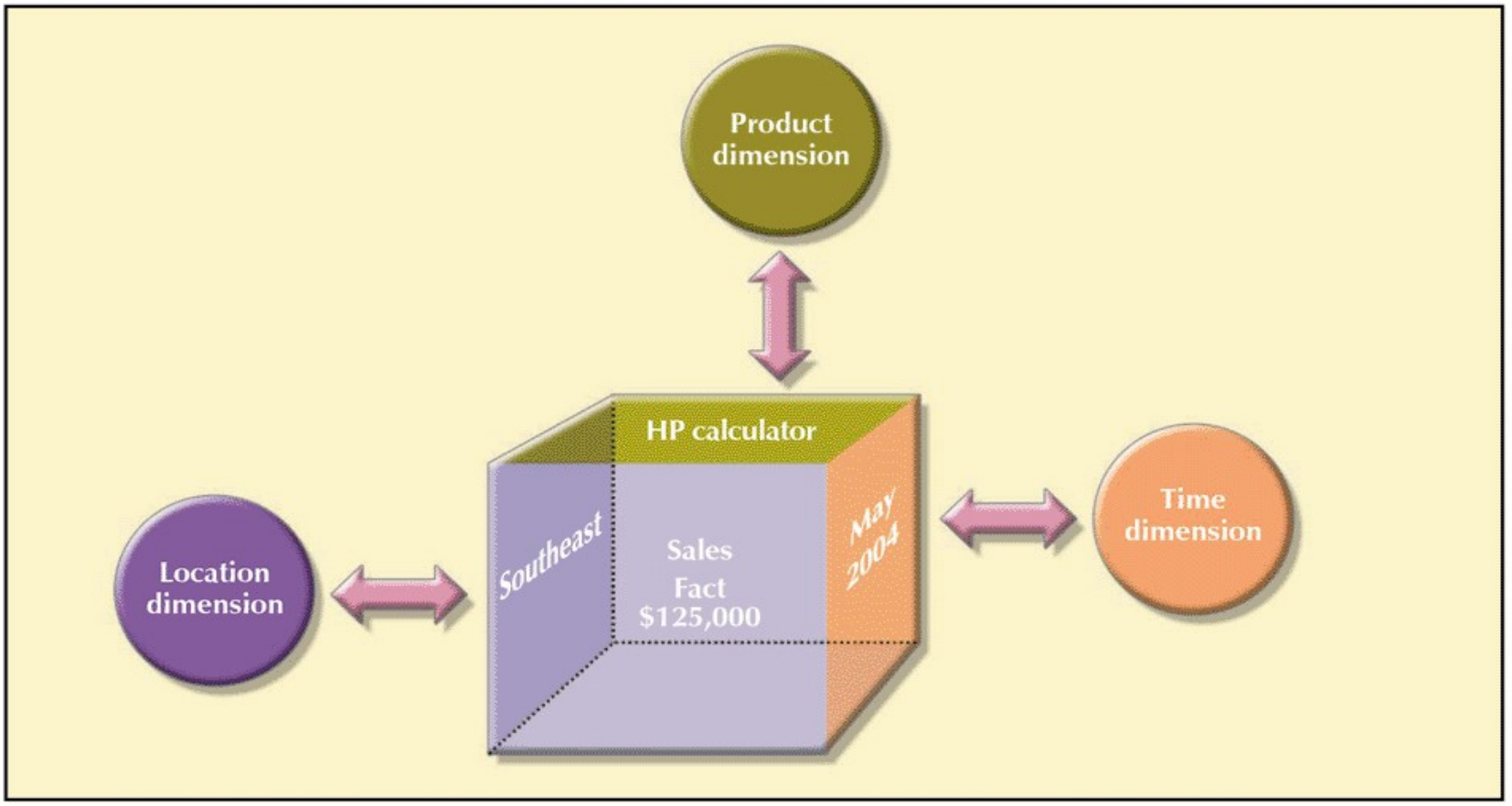
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Star Schema

- Data modeling technique used to map multidimensional decision support data into a relational database
- Creates the near equivalent of a multidimensional database schema from the existing relational database
- Yield an easily implemented model for multidimensional data analysis, while still preserving the relational structures on which the operational database is built
- Has four components: facts, dimensions, attributes, and attribute hierarchies

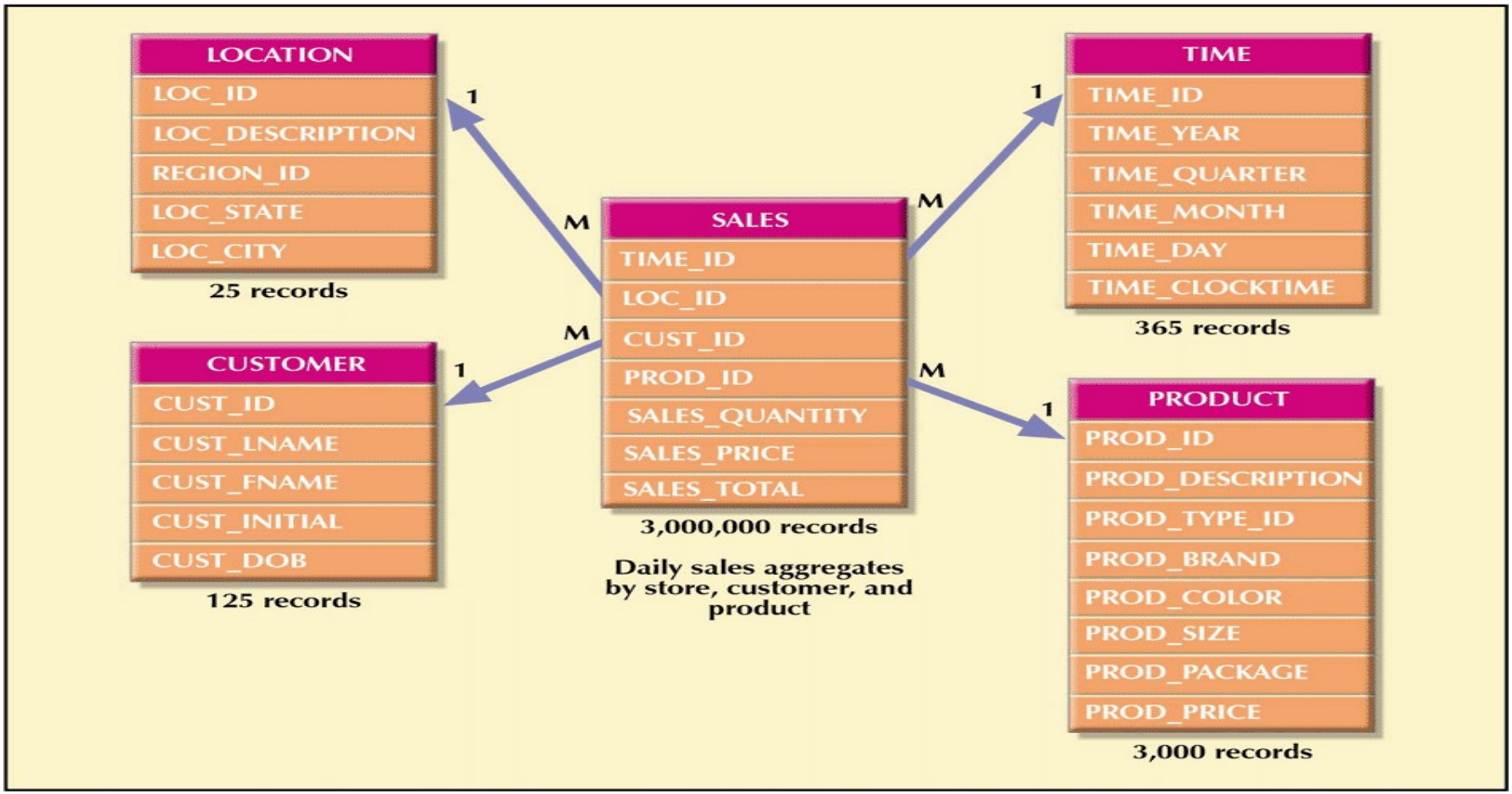
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Simple Star Schema



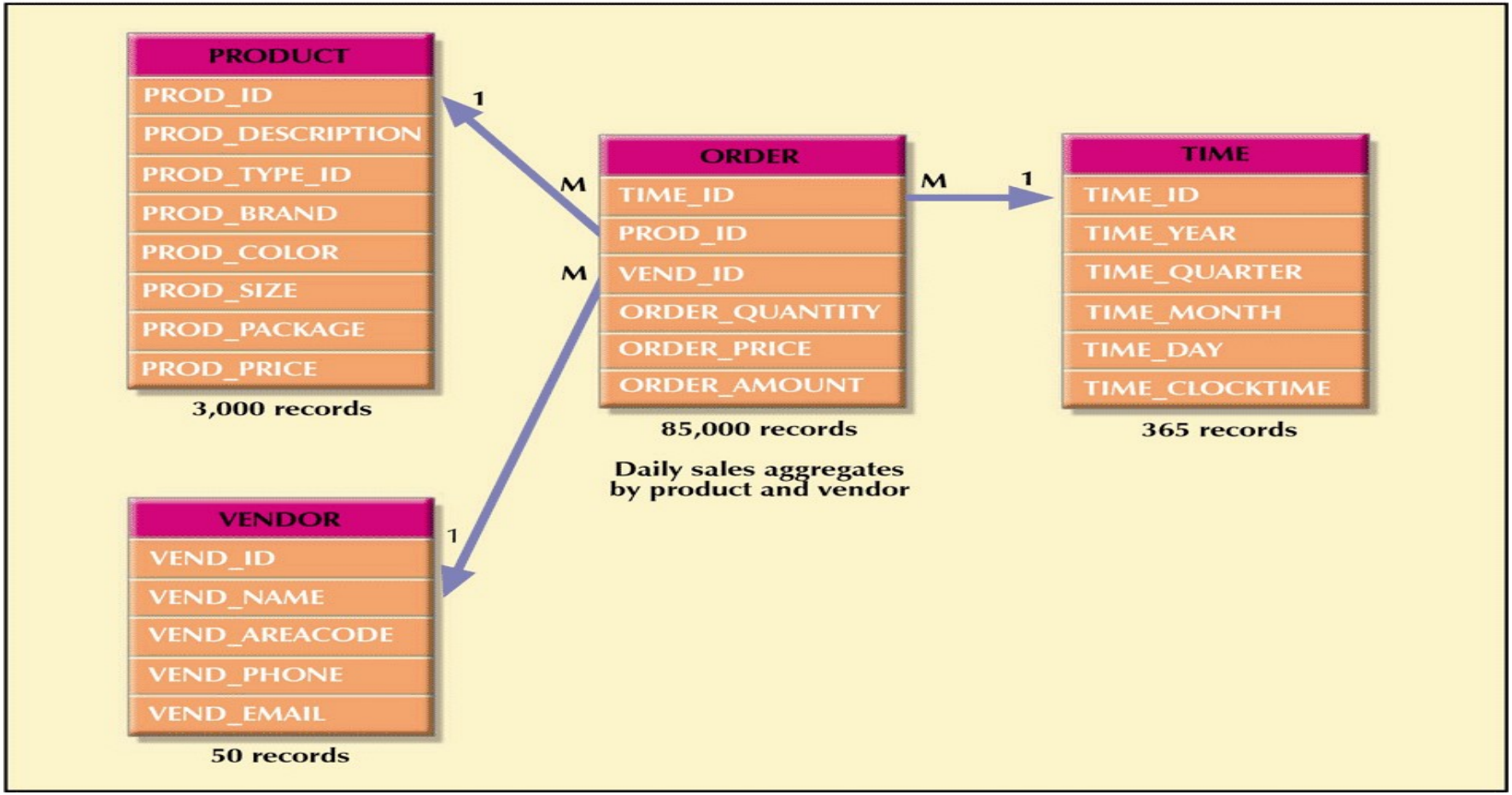
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Star Schema for Sales



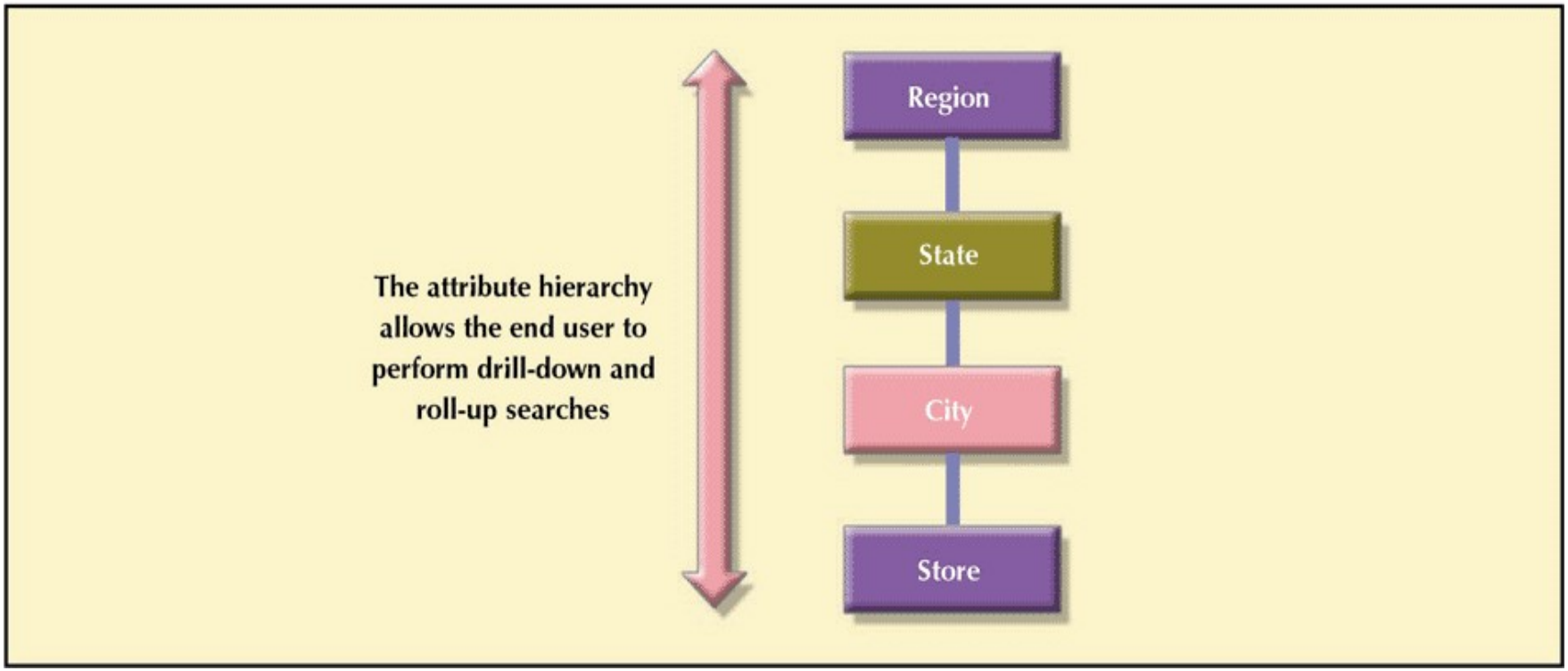
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Star Schema for Orders



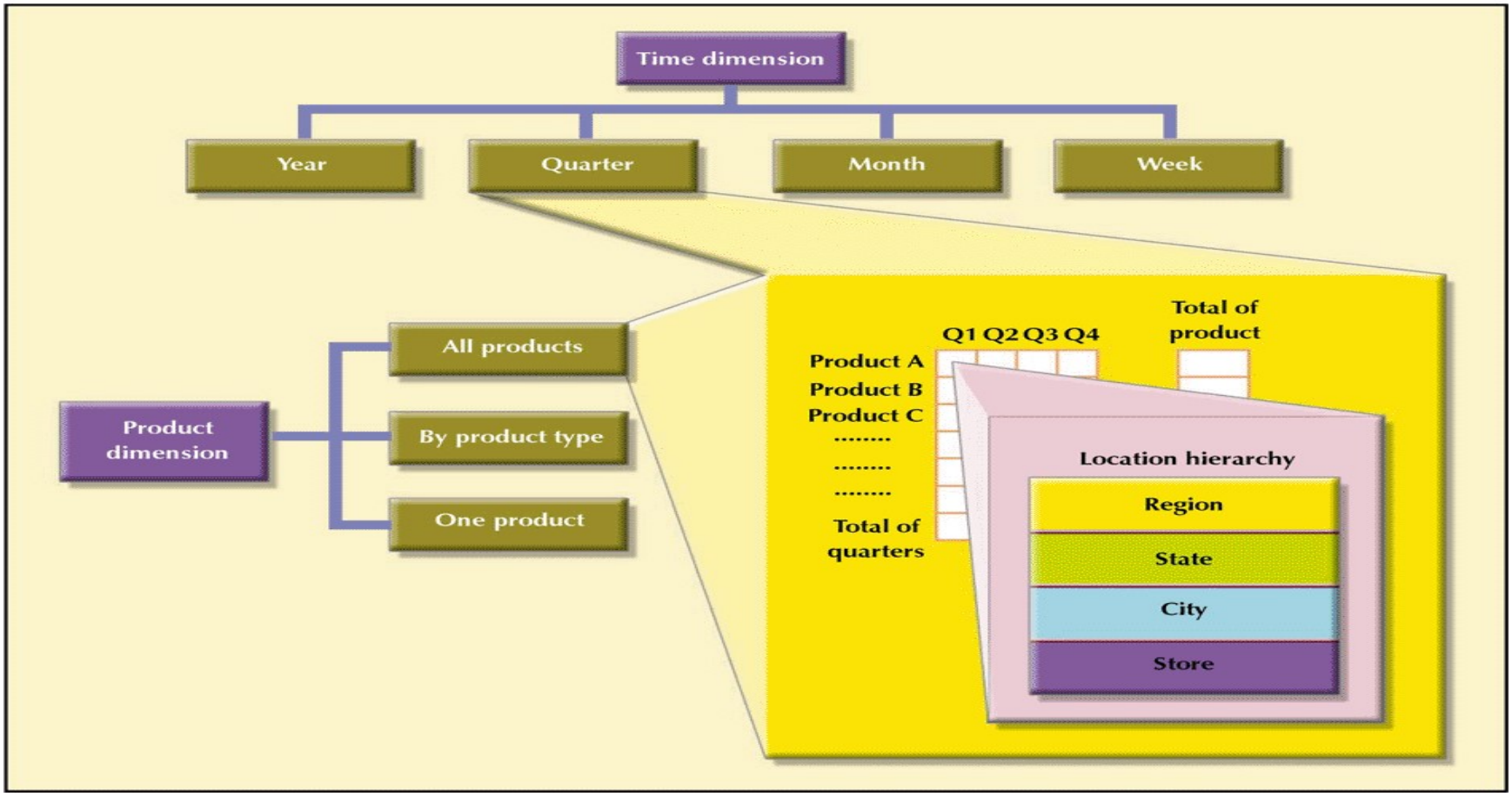
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Location Attribute Hierarchy



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Attribute Hierarchies In Multidimensional Analysis



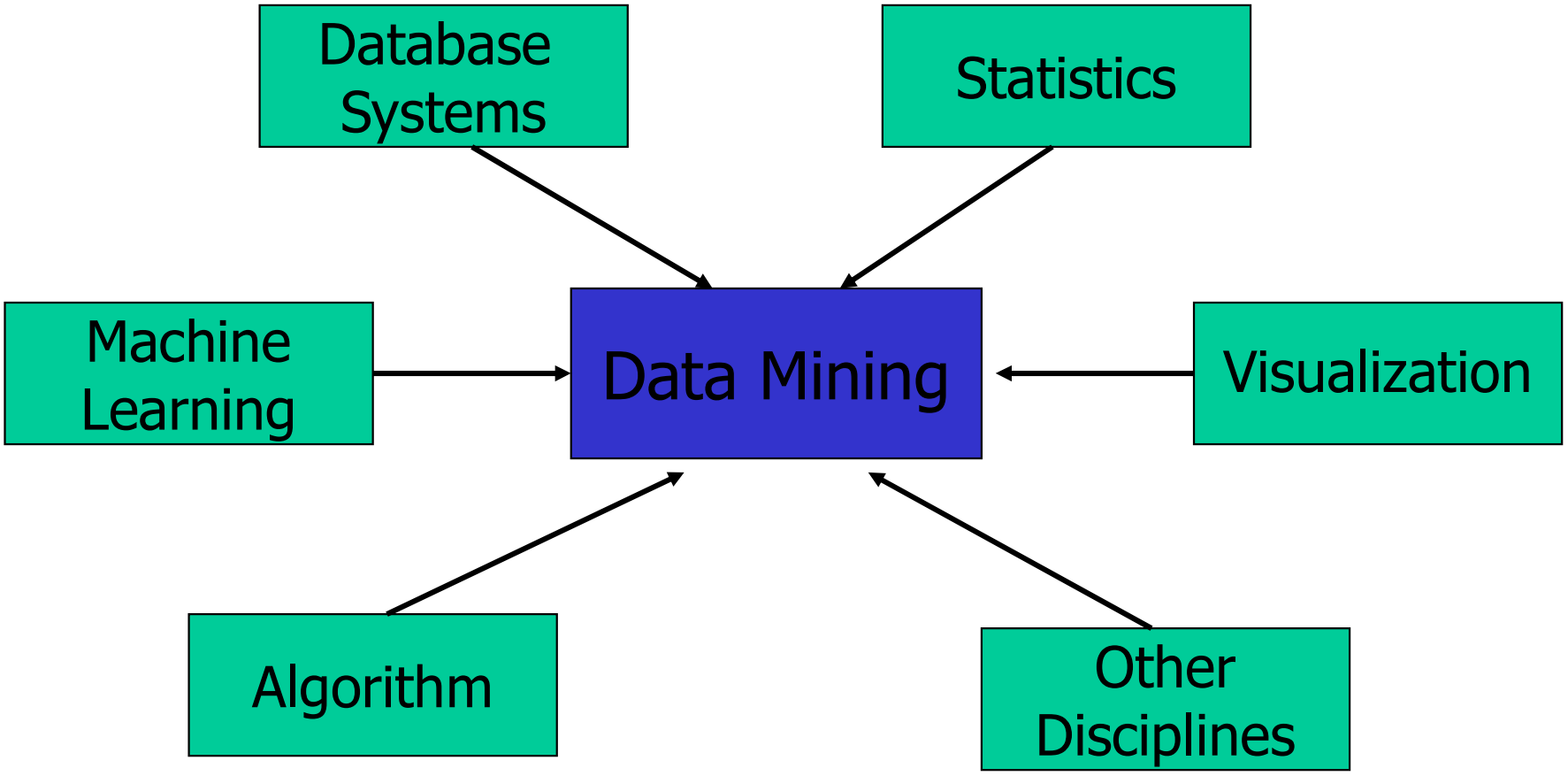
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Mining

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Query processing
  - Expert systems or ML/statistical programs

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Mining: Confluence of Multiple Disciplines



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting(보험), quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Corporate Analysis & Risk Management

Contingency claim: 금융파생상품(미래에 재화나 자산을 교환하는 계약). 파생상품의 기초 자산은 주식, 채권, 금리, 환율, 상품 등 무엇이든지 가능

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent 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

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - 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
  - Anti-terrorism

From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

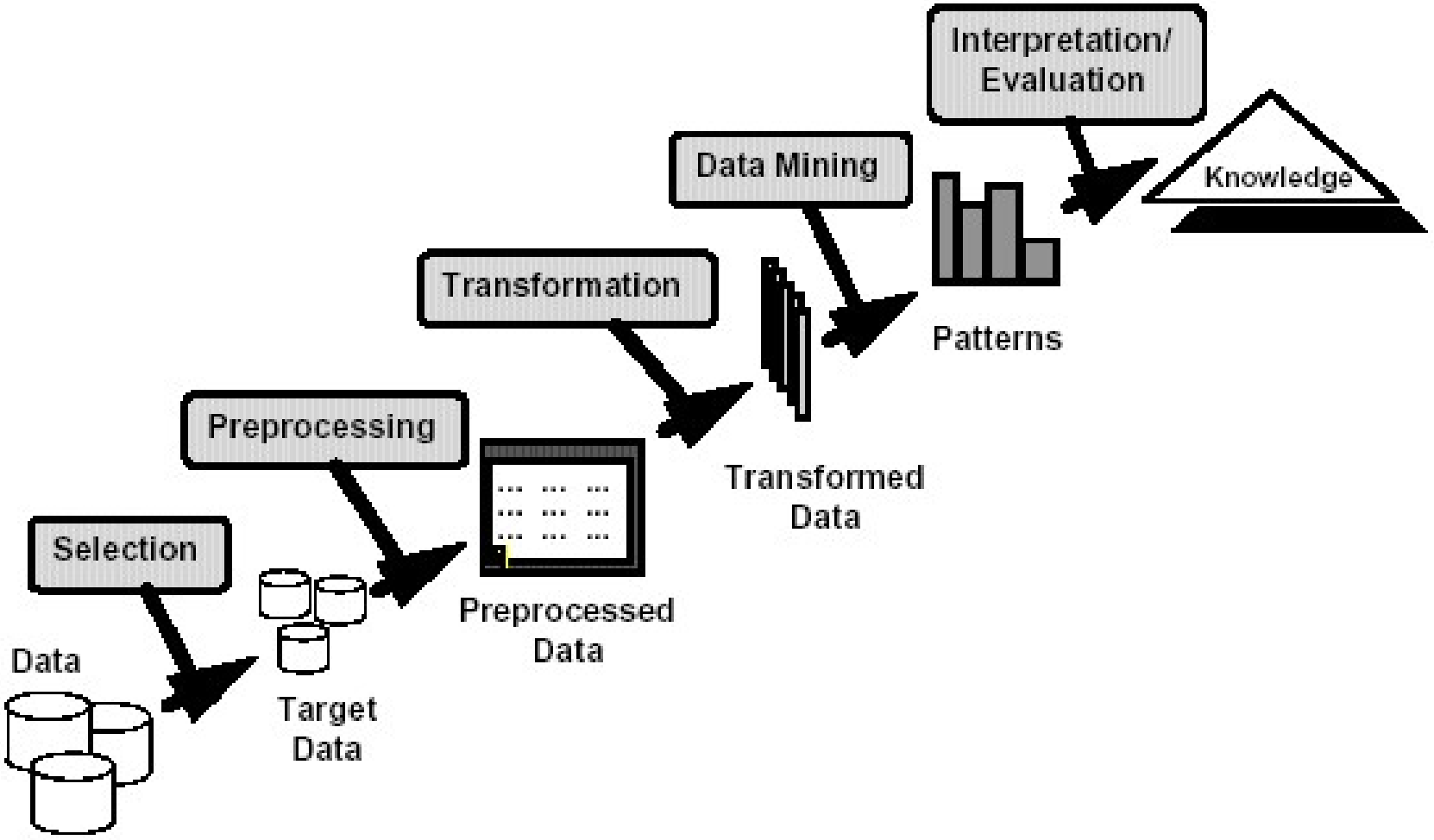
"Other" Mining Environments

In addition to data stored in traditional databases there are other "structures" that can be mined for patterns.

- Text Mining is the application of data mining to non-structured or less-structured text files
- Web Mining is the application of data mining techniques to data related to the World Wide Web. The data may be present in web pages or related to Web activity.
- Spatial Mining is the application of data mining techniques to data that have a location component.
- Temporal Mining is the application of data mining techniques to data that are maintained for multiple points in time.

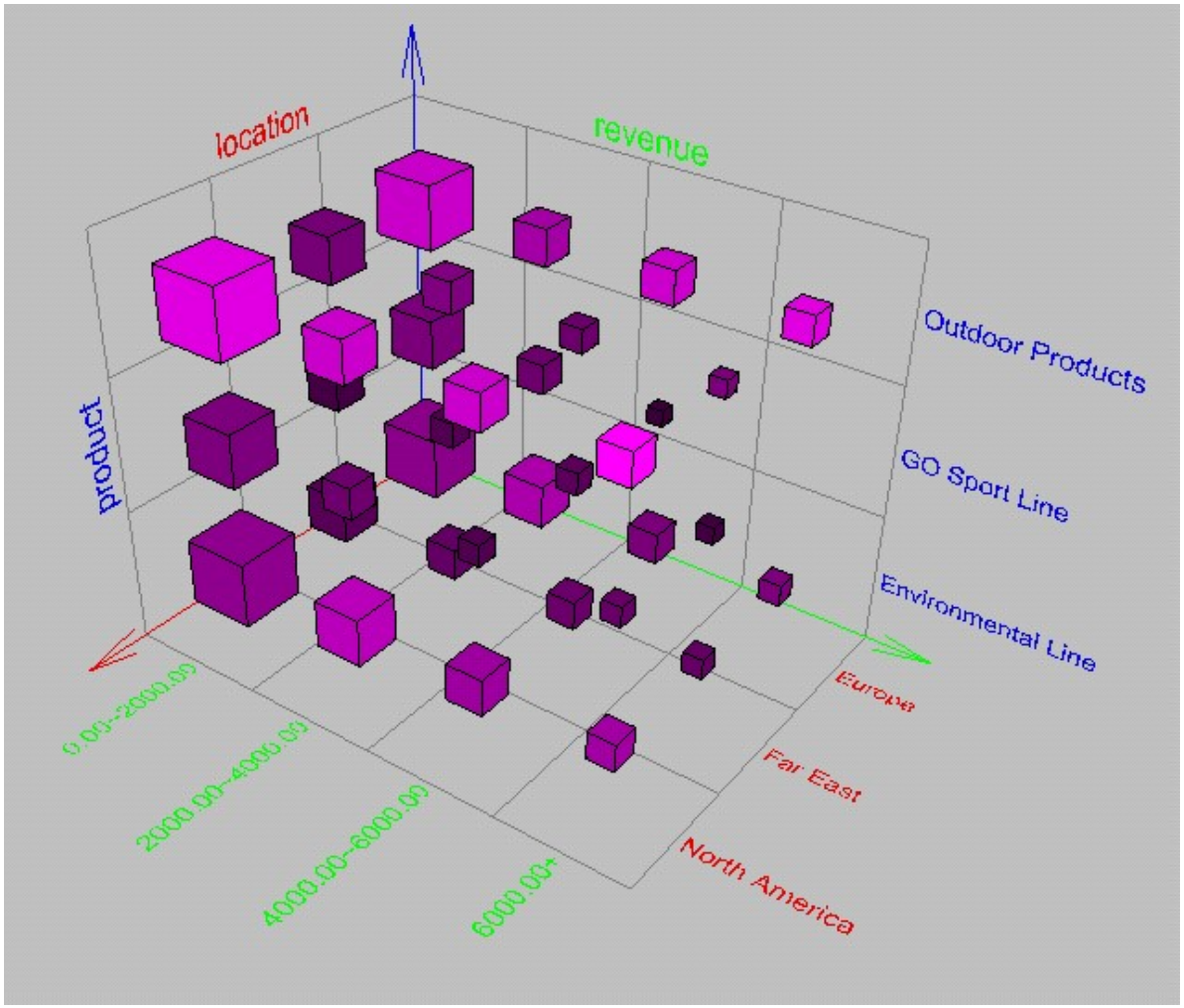
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Steps of a KDD Process



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

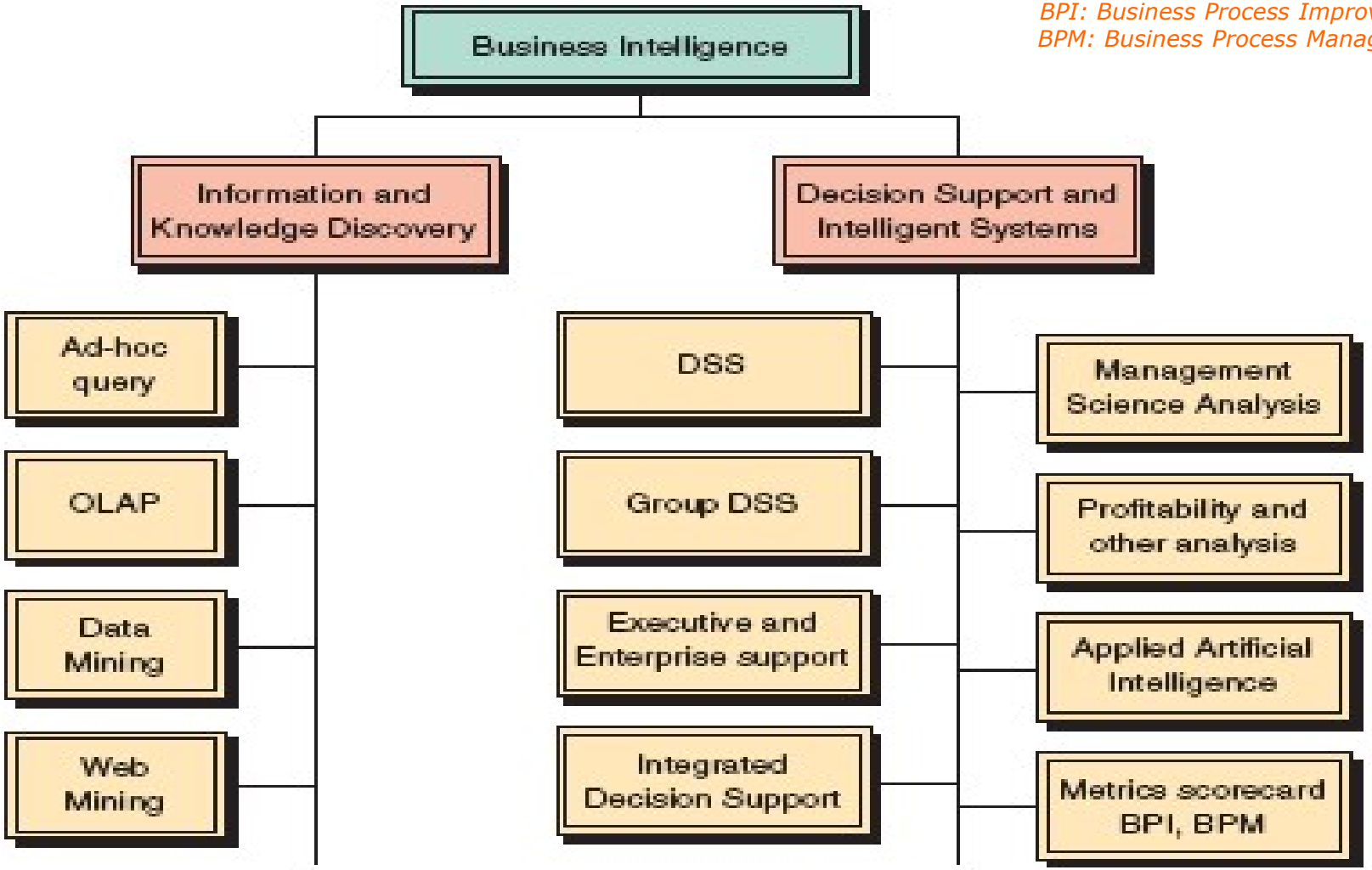
Visualization



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Business Intelligence

BPI: Business Process Improvement
BPM: Business Process Management



From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques