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.

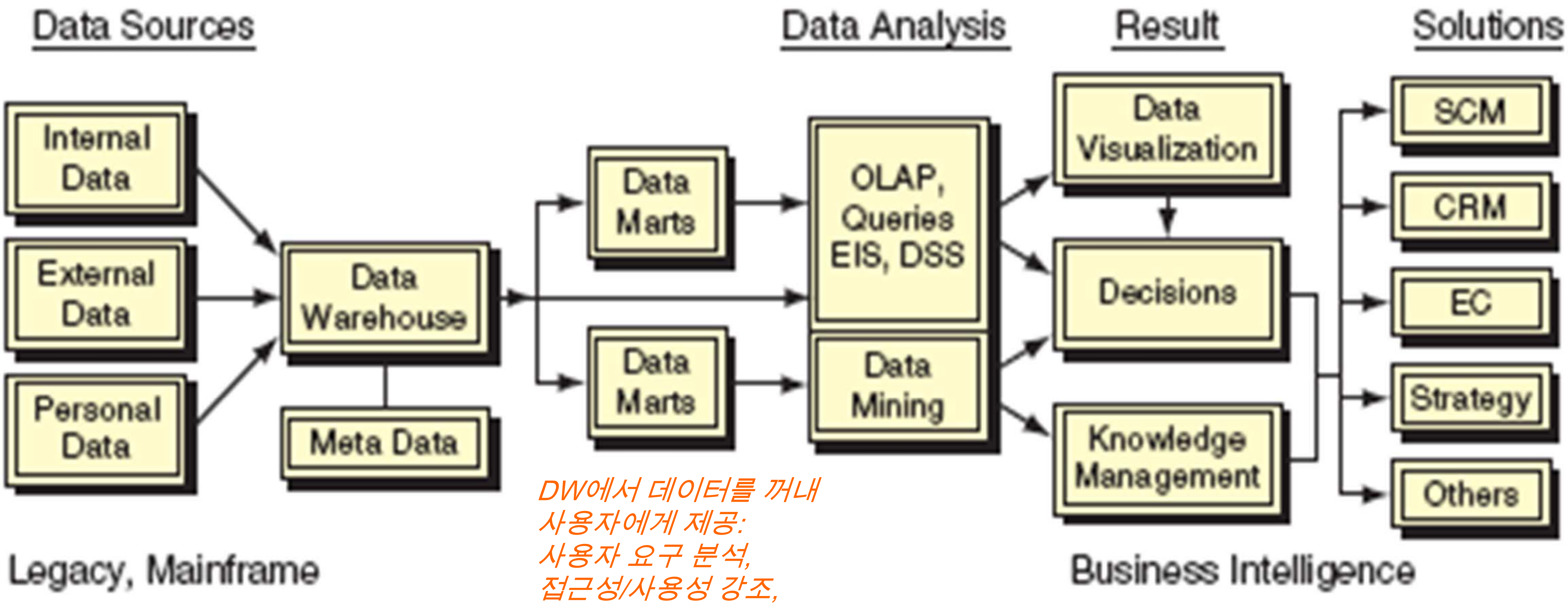
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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

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



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

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

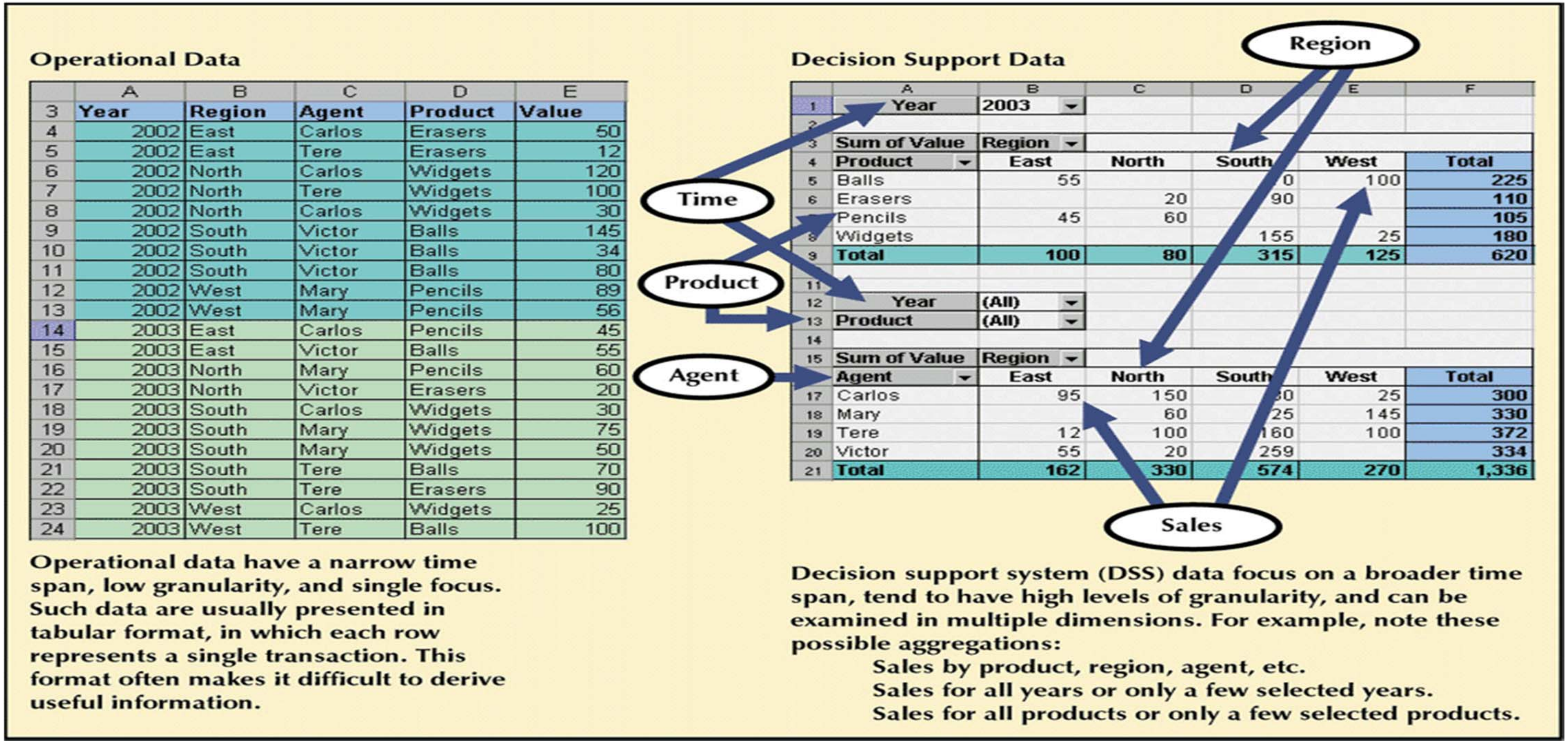
Transactional (Operational) vs. Analytical (Decision Support) Data Processing

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



Operational Data

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

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

Decision Support Data

	A	B	C	D	E	F
1	**Year**	2003				
2						
3	**Sum of Value**	**Region**				
4	**Product**	**East**	**North**	**South**	**West**	**Total**
5	Balls	55		0	100	**225**
6	Erasers		20	90		**110**
7	Pencils	45	60			**105**
8	Widgets			155	25	**180**
9	**Total**	**100**	**80**	**315**	**125**	**620**
11						
12	**Year**	**(All)**				
13	**Product**	**(All)**				
14						
15	**Sum of Value**	**Region**				
16	**Agent**	**East**	**North**	**South**	**West**	**Total**
17	Carlos	95	150	30	25	**300**
18	Mary		60	25	145	**330**
19	Tere	12	100	160	100	**372**
20	Victor	55	20	259		**334**
21	**Total**	**162**	**330**	**574**	**270**	**1,336**

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

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

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, but not rigorously.
  - 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: 전원이 끊겨도 데이터가 소멸되지 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

- Facebook page

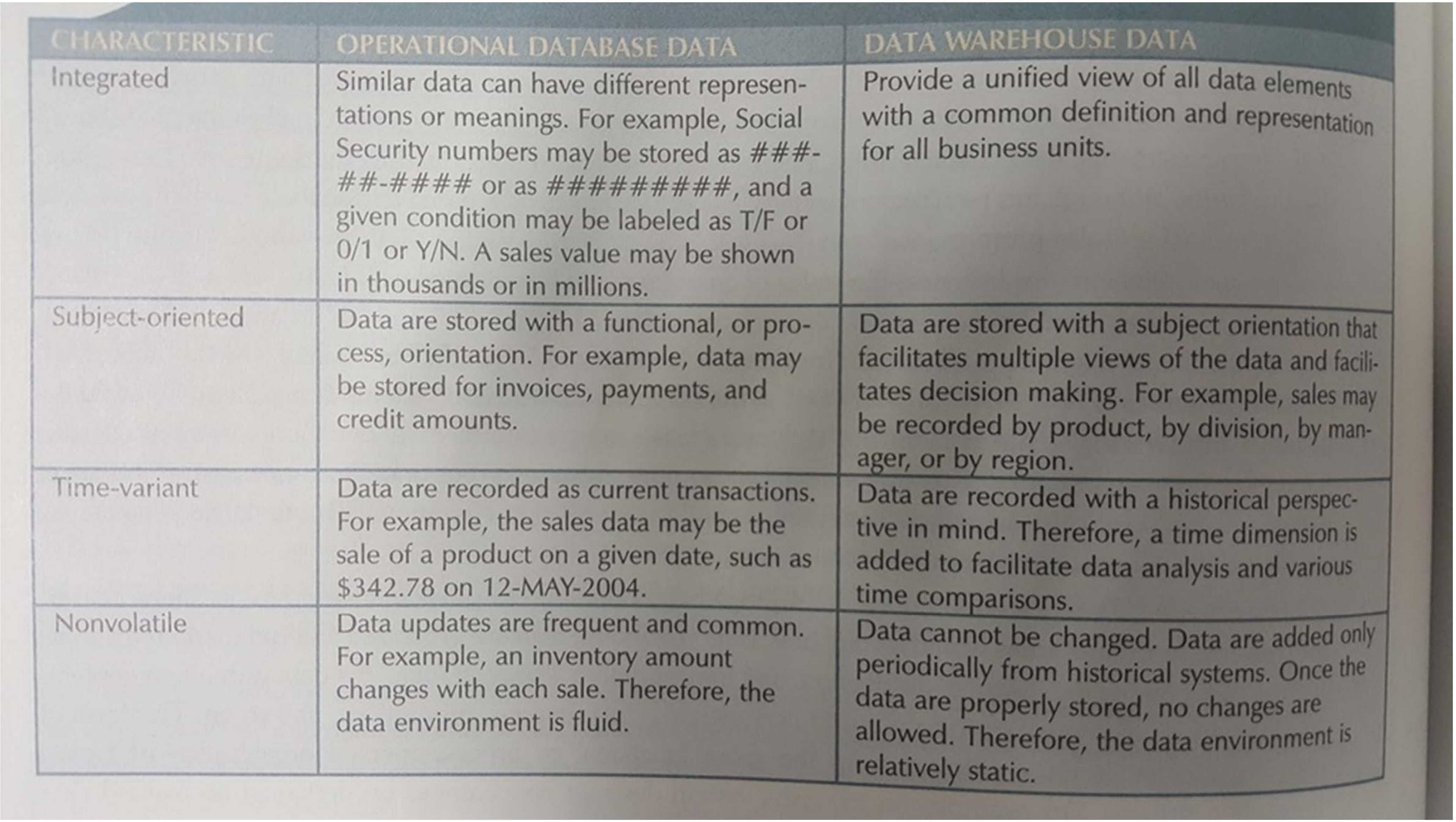
“Jinwoo is going to lunch early with his best friend, Sunghyun at Iron Pit BBQ near school”

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*.

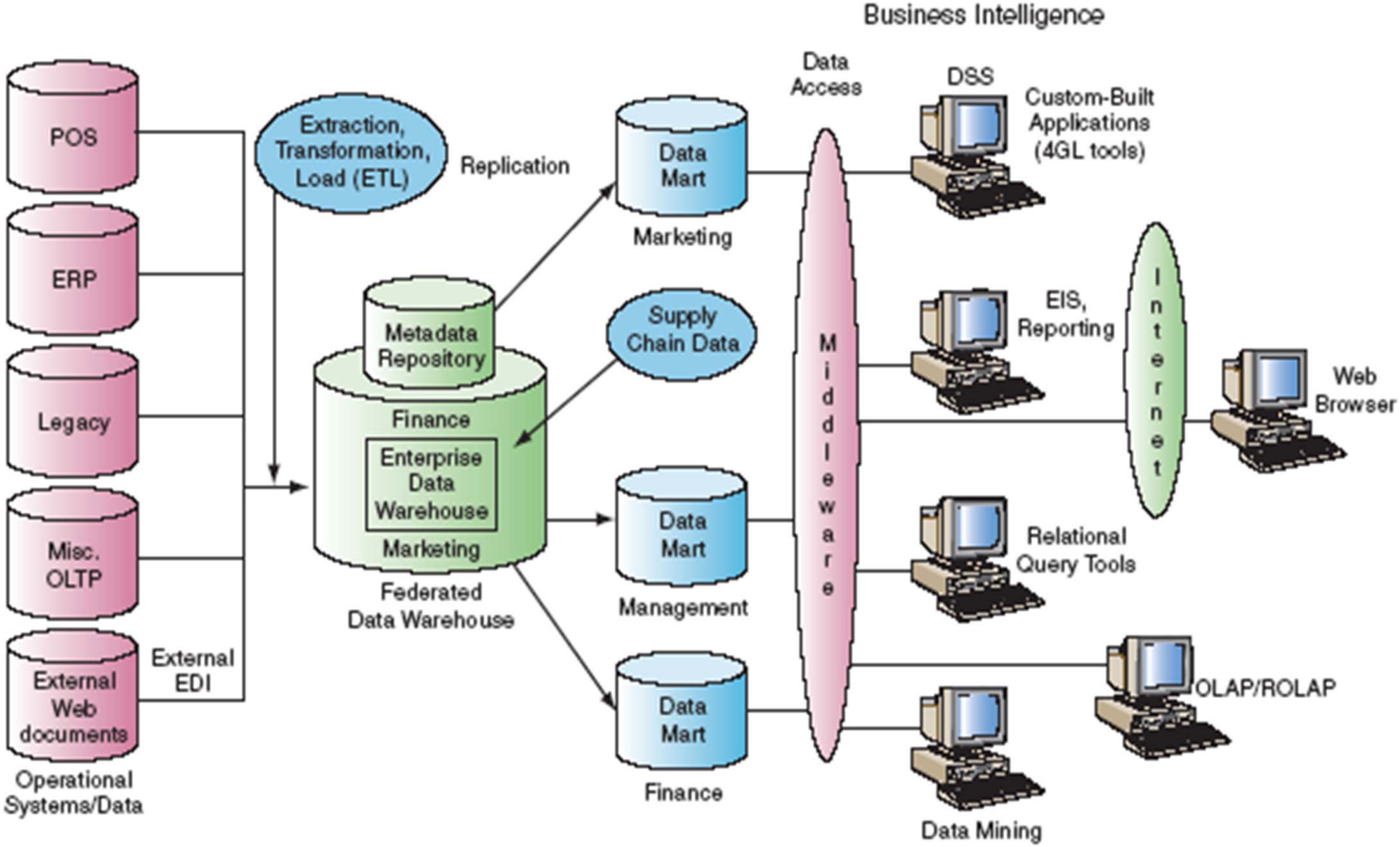
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques



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

The 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

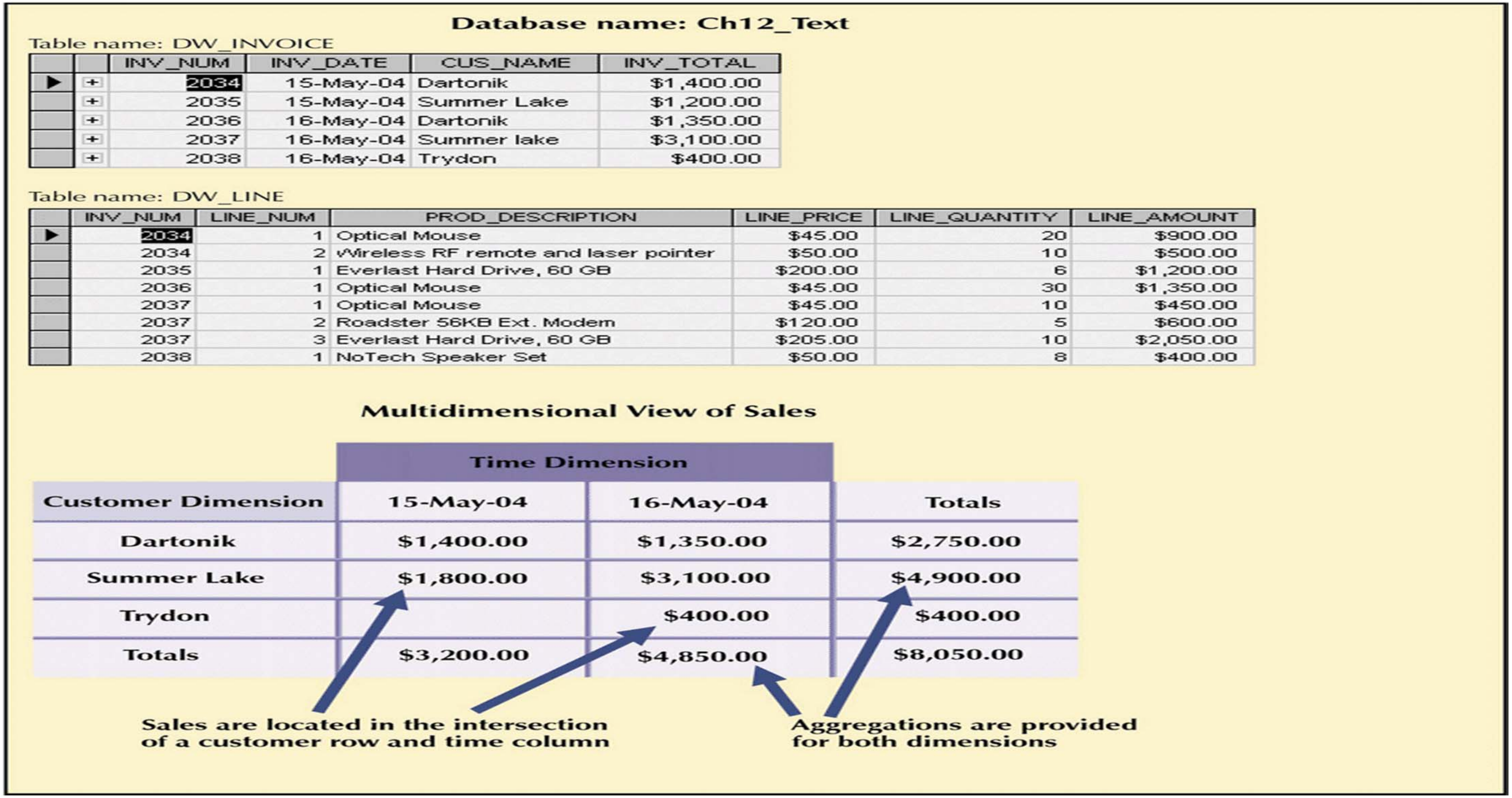
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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.

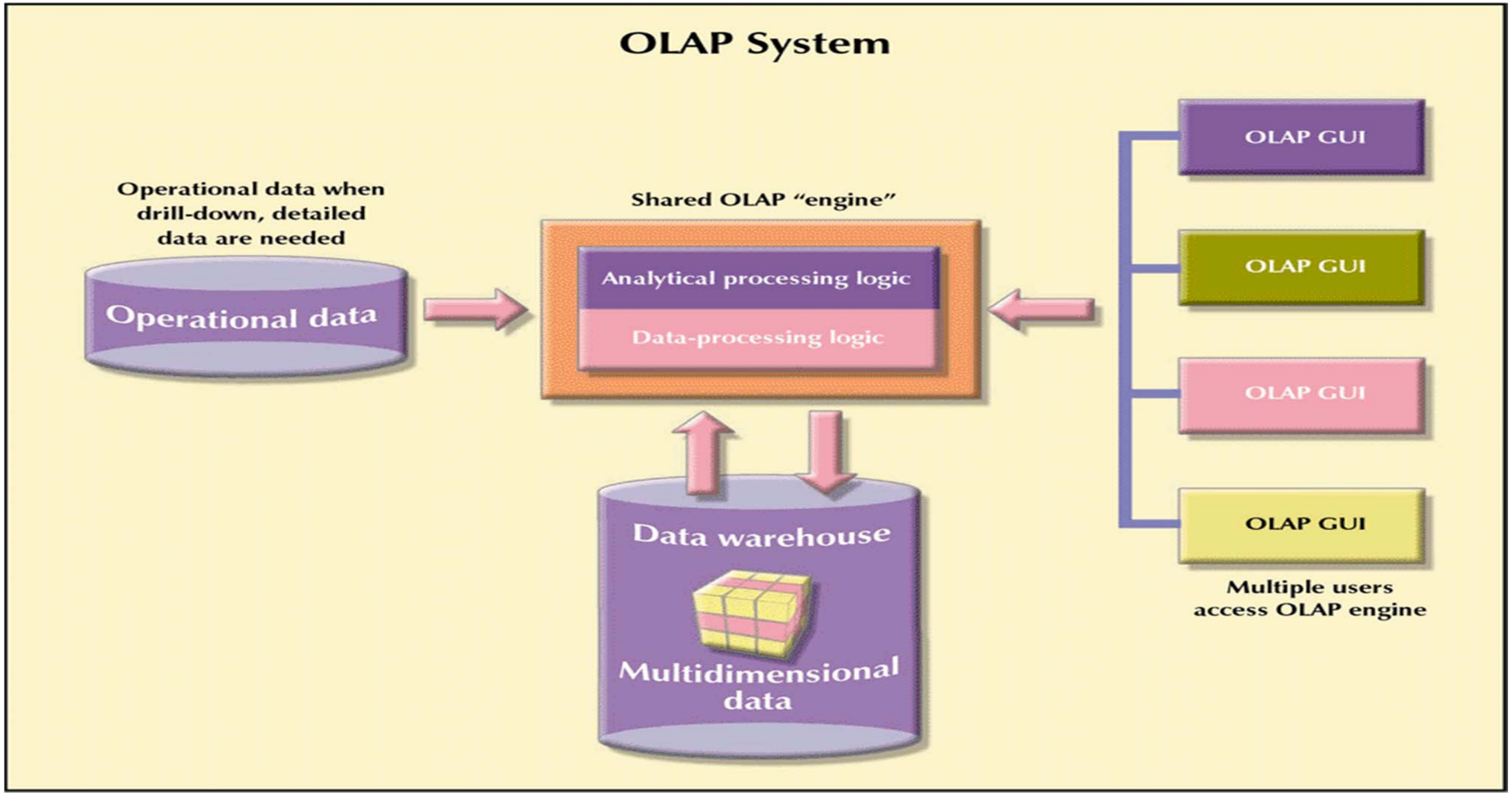
- ROLAP (Relational OLAP) is an OLAP database implemented on top of an existing multiple relational database: multidimensional data schema support within the RDBMS by normalization & queries
- MOLAP (Multidimensional OLAP) is a specialized multidimensional data store such as a Data Cube. The multidimensional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

Data Cube: 3D plotting of data, static, not subject to change,
cannot be created by ad-hoc queries, CBR, 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

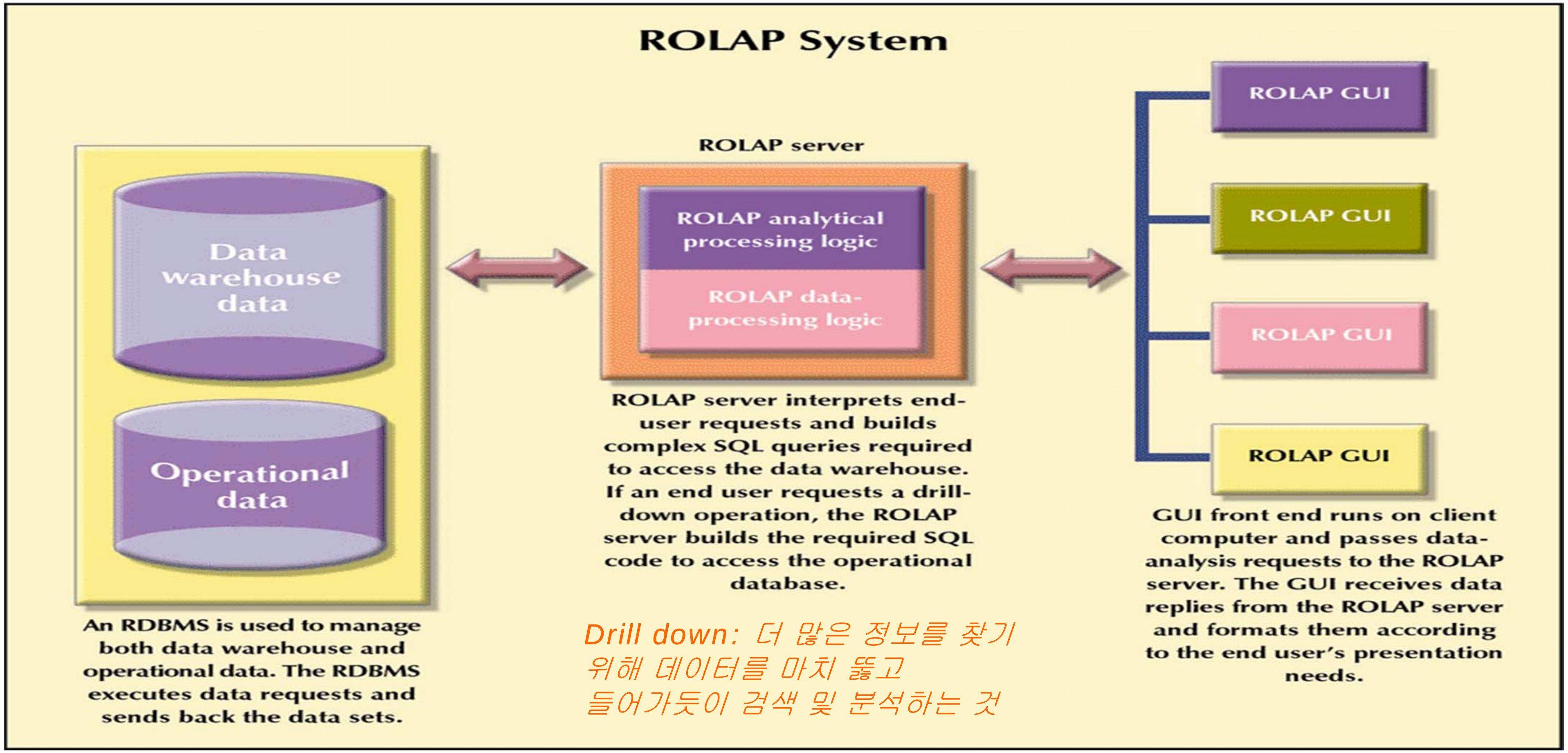
OLAP System

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



An RDBMS is used to manage both data warehouse and operational data. The RDBMS executes data requests and sends back the data sets.

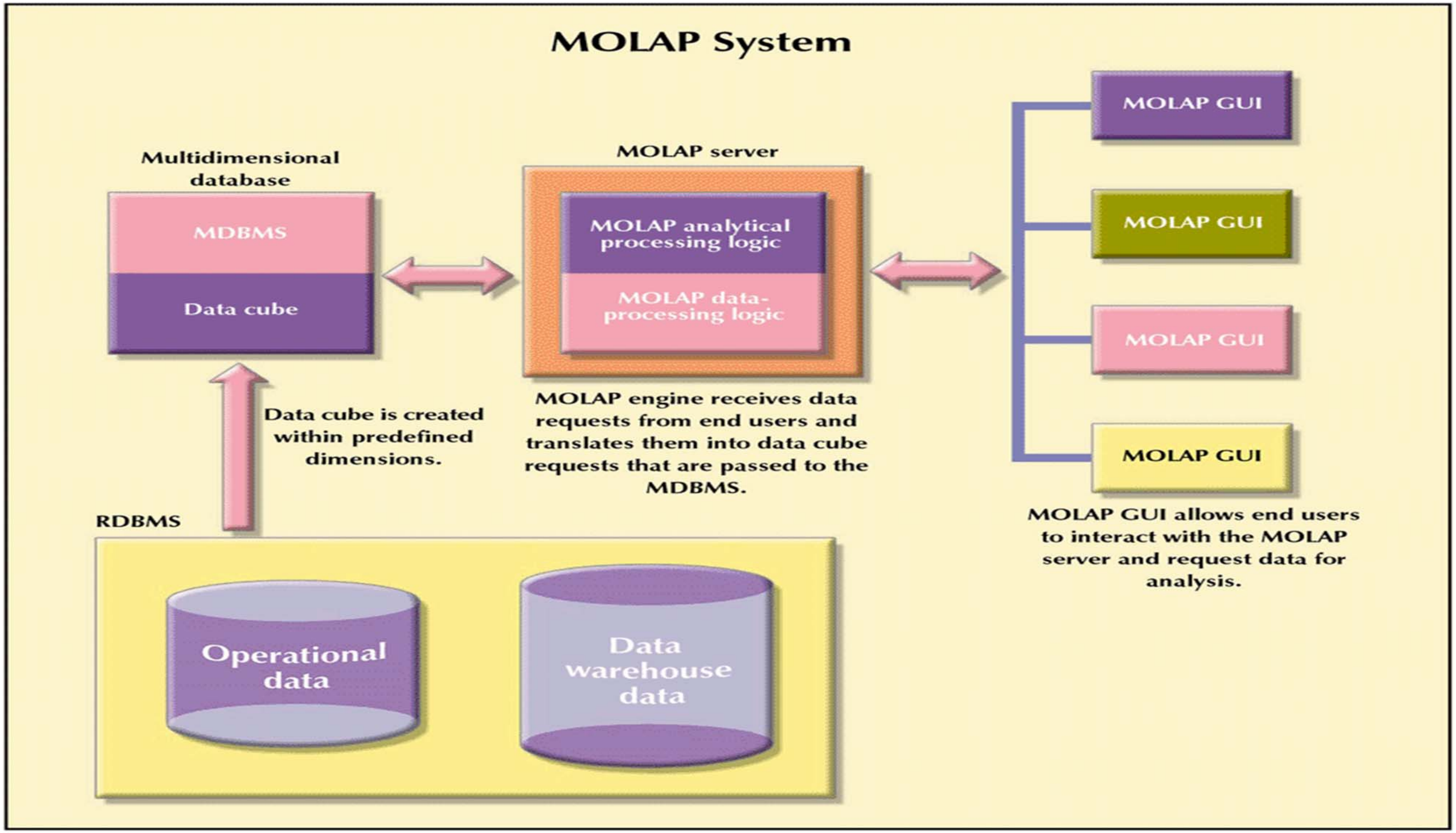
ROLAP server interprets end-user requests and builds complex SQL queries required to access the data warehouse. If an end user requests a drill-down operation, the ROLAP server builds the required SQL code to access the operational database.

GUI front end runs on client computer and passes data-analysis requests to the ROLAP server. The GUI receives data replies from the ROLAP server and formats them according to the end user's presentation needs.

Drill down: 더 많은 정보를 찾기 위해 데이터를 마치 뚫고 들어가듯이 검색 및 분석하는 것

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

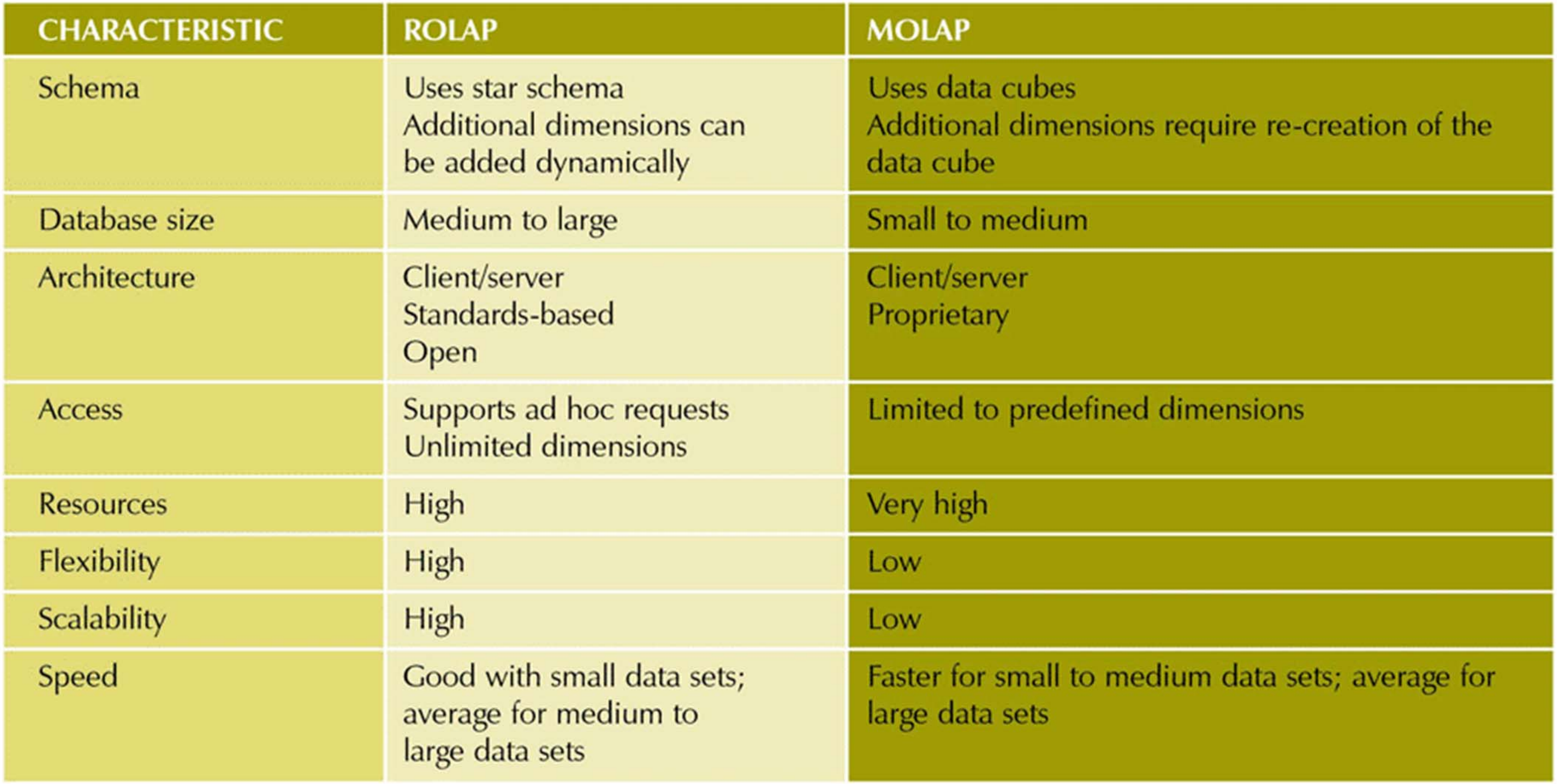

MOLAP System
Multidimensional database
MDBMS
Data cube
MOLAP server
MOLAP analytical processing logic
MOLAP data-processing logic
MOLAP GUI
MOLAP GUI
MOLAP GUI
MOLAP GUI
Data cube is created within predefined dimensions.
MOLAP engine receives data requests from end users and translates them into data cube requests that are passed to the MDBMS.
MOLAP GUI allows end users to interact with the MOLAP server and request data for analysis.
RDBMS
Operational data
Data warehouse 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

Relational vs. Multidimensional OLAP

Star Schema: Fact tables + Requested Dimensions
Proprietary: 상표/특허

INTEGRATED!!



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

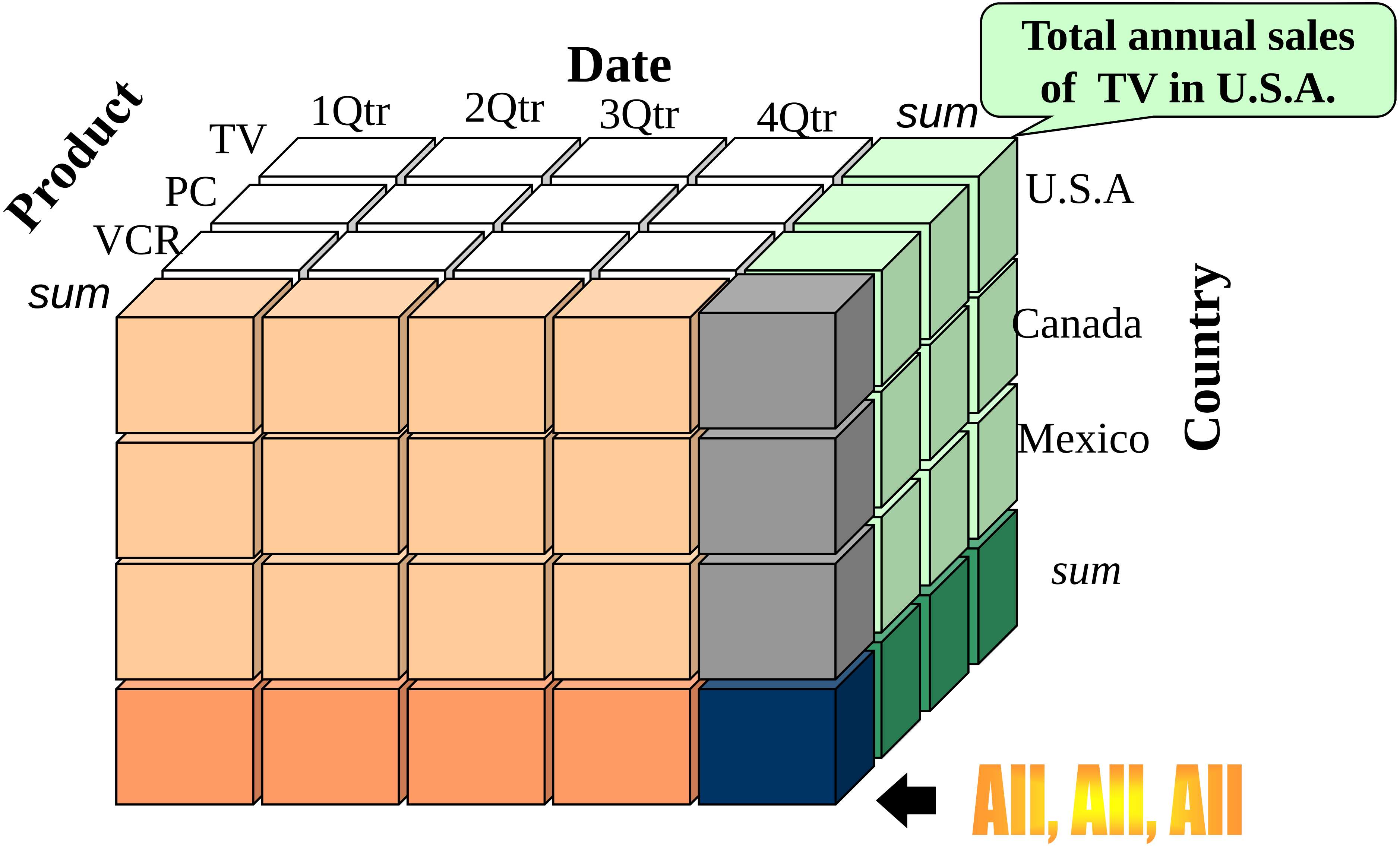
The Data Cube

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
  - Drill Downs

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

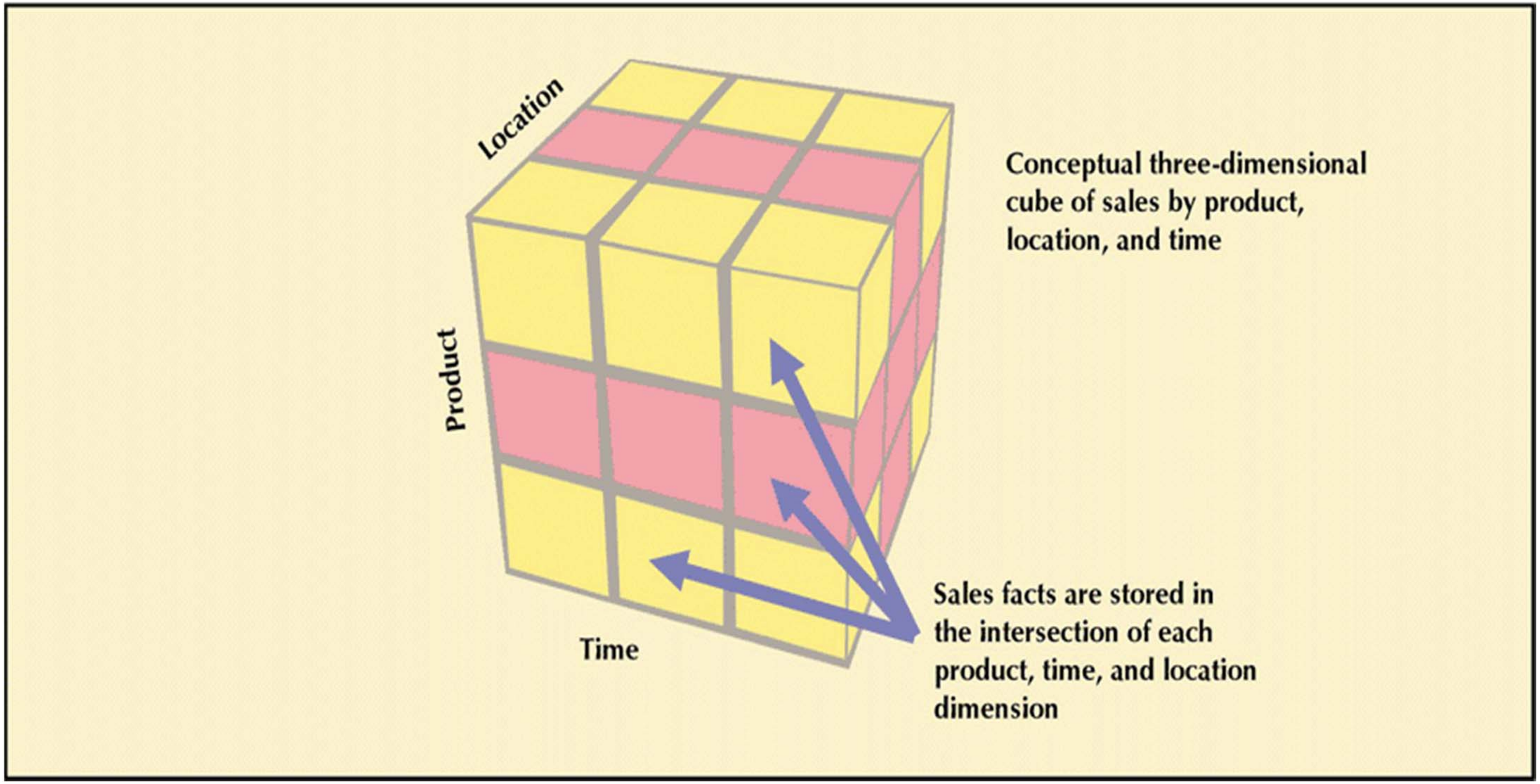


From Turban et al. (2004), Information Technology for Management

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

Han, Kamber (2001) Data Mining: Concepts and Techniques

Three-Dimensional View of Sales



Conceptual three-dimensional cube of sales by product, location, and time

Sales facts are stored in the intersection of each product, time, and location dimension

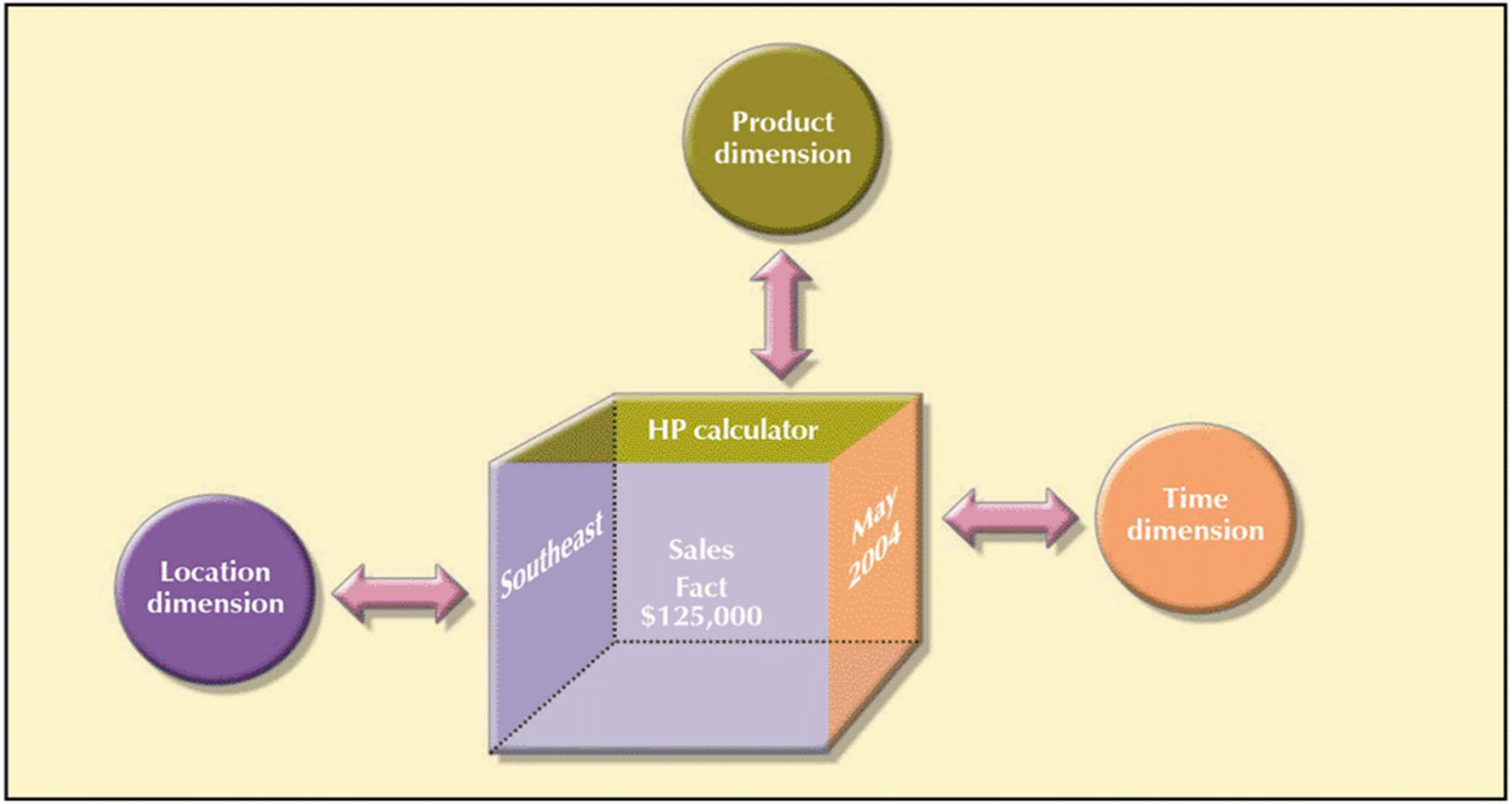
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Star Schemas

- 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

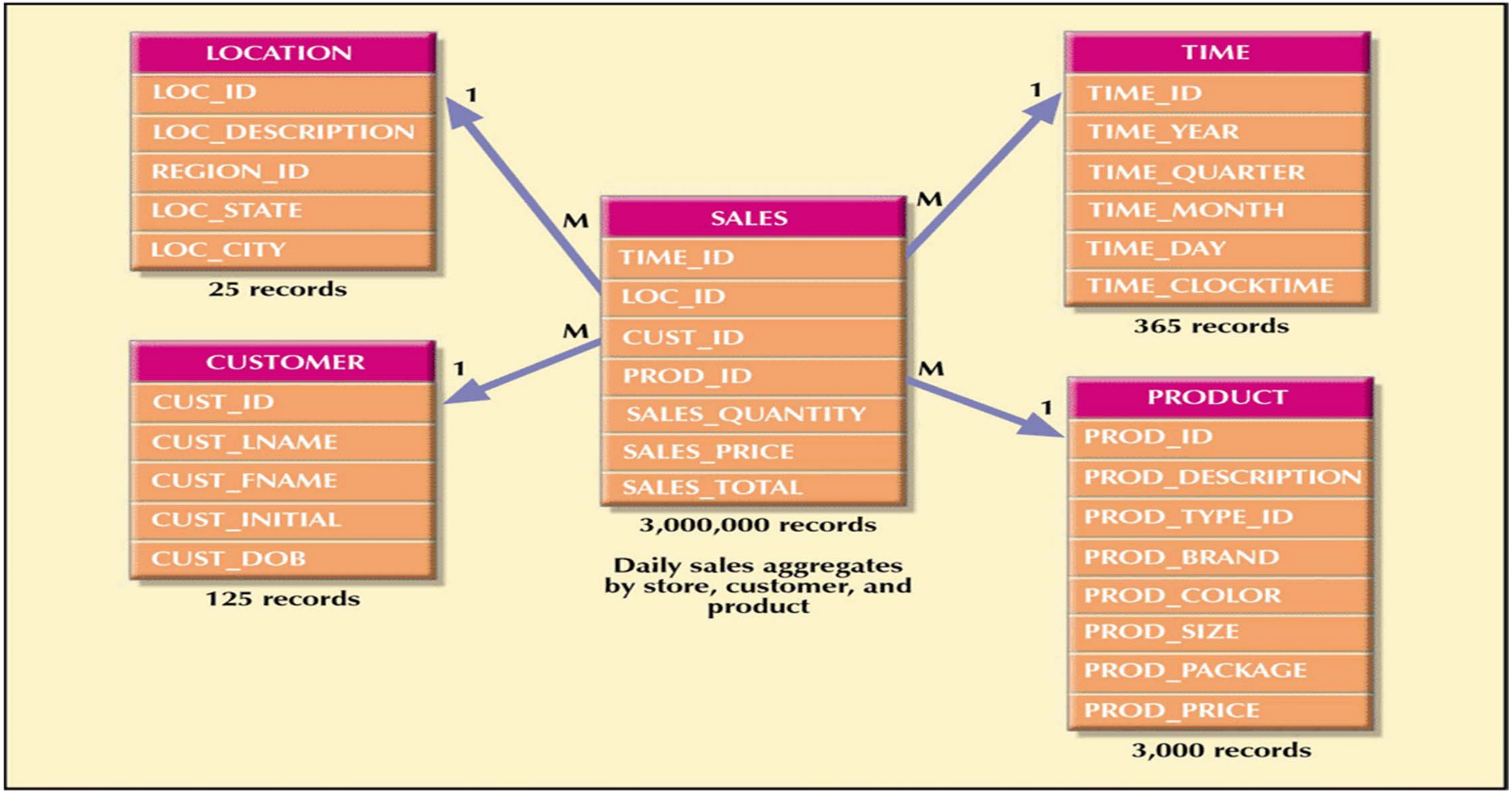


From Turban et al. (2004), Information Technology for Management

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

Han, Kamber (2001) Data Mining: Concepts and Techniques

Star Schema for Sales



LOCATION
LOC_ID
LOC_DESCRIPTION
REGION_ID
LOC_STATE
LOC_CITY

25 records

CUSTOMER
CUST_ID
CUST_LNAME
CUST_FNAME
CUST_INITIAL
CUST_DOB

125 records

SALES
TIME_ID
LOC_ID
CUST_ID
PROD_ID
SALES_QUANTITY
SALES_PRICE
SALES_TOTAL

3,000,000 records

Daily sales aggregates by store, customer, and product

TIME
TIME_ID
TIME_YEAR
TIME_QUARTER
TIME_MONTH
TIME_DAY
TIME_CLOCKTIME

365 records

PRODUCT
PROD_ID
PROD_DESCRIPTION
PROD_TYPE_ID
PROD_BRAND
PROD_COLOR
PROD_SIZE
PROD_PACKAGE
PROD_PRICE

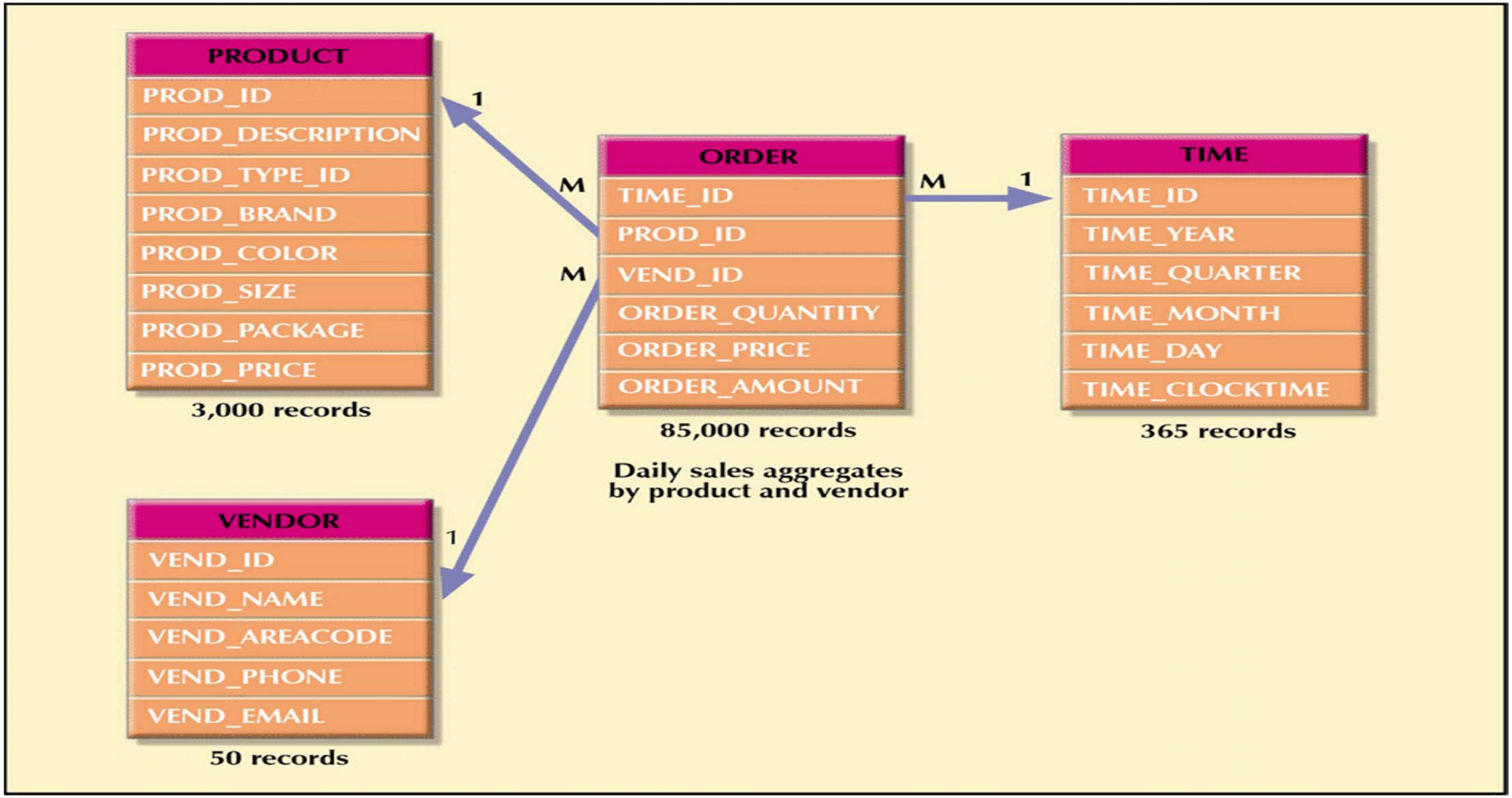
3,000 records

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



PRODUCT
PROD_ID
PROD_DESCRIPTION
PROD_TYPE_ID
PROD_BRAND
PROD_COLOR
PROD_SIZE
PROD_PACKAGE
PROD_PRICE

3,000 records

ORDER
TIME_ID
PROD_ID
VEND_ID
ORDER_QUANTITY
ORDER_PRICE
ORDER_AMOUNT

85,000 records

Daily sales aggregates
by product and vendor

TIME
TIME_ID
TIME_YEAR
TIME_QUARTER
TIME_MONTH
TIME_DAY
TIME_CLOCKTIME

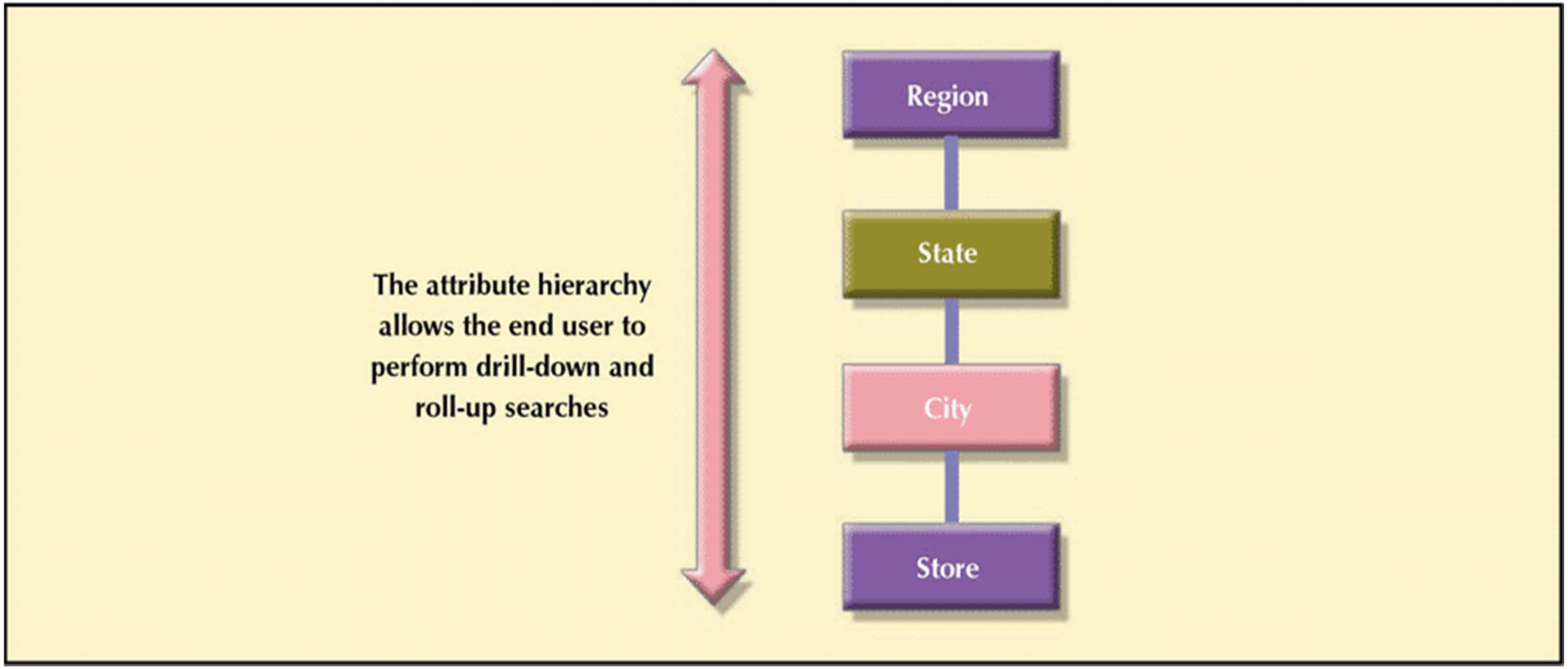
365 records

VENDOR
VEND_ID
VEND_NAME
VEND_AREACODE
VEND_PHONE
VEND_EMAIL

50 records

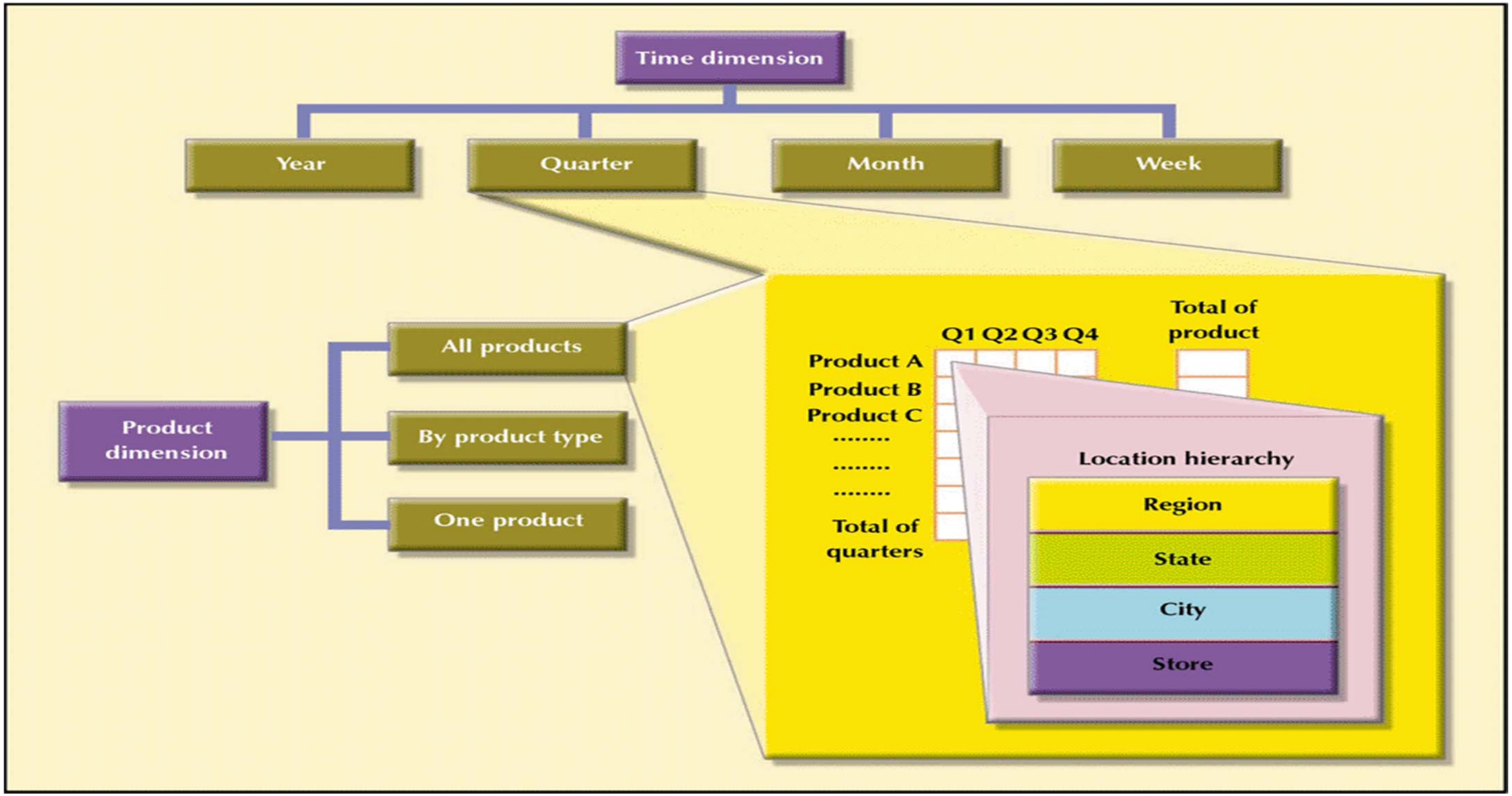
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Location Attribute Hierarchy


The attribute hierarchy allows the end user to perform drill-down and roll-up searches
Region
State
City
Store

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



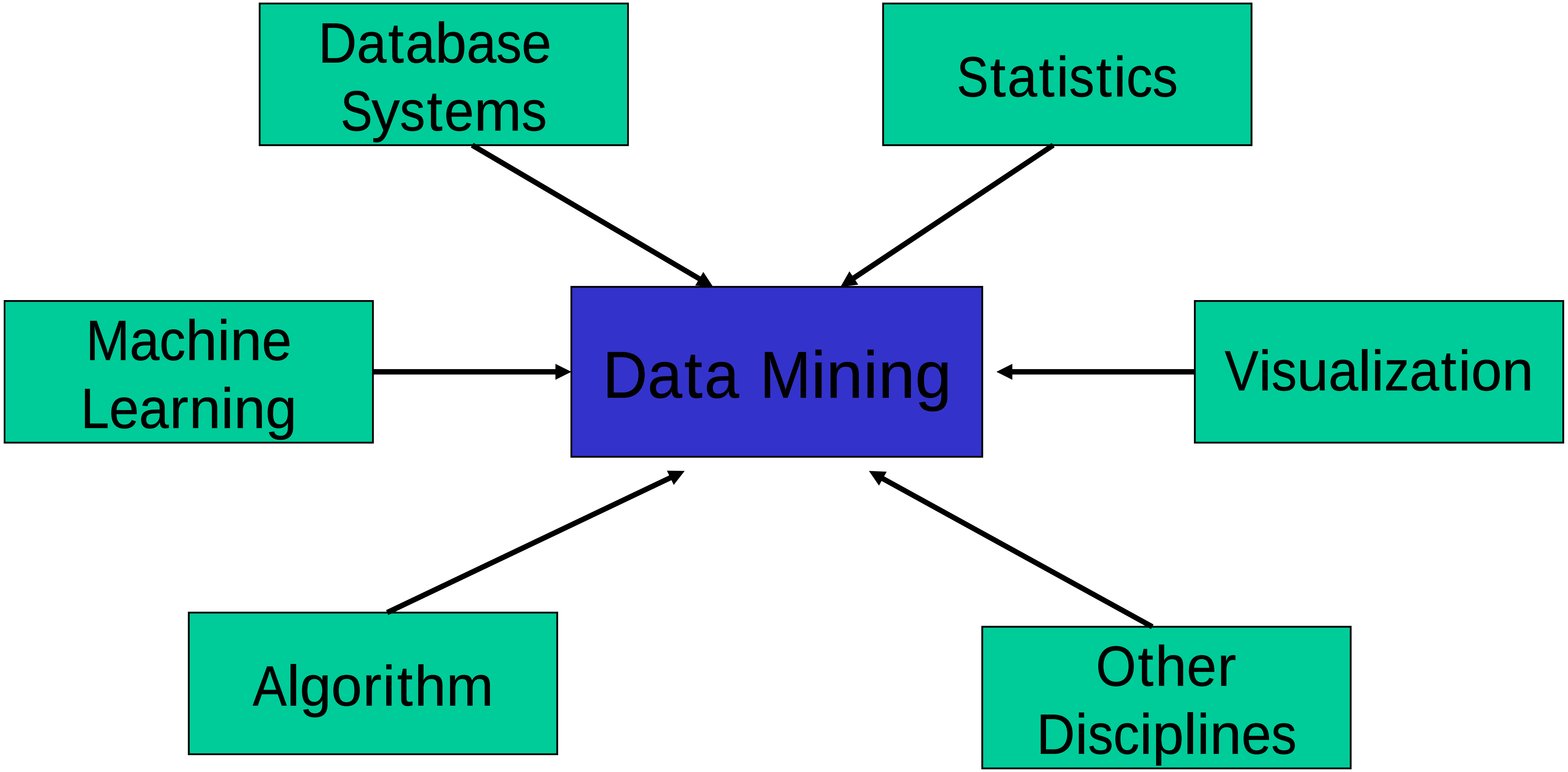
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

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 small 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, 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

- 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
  - Retail industry
  - 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

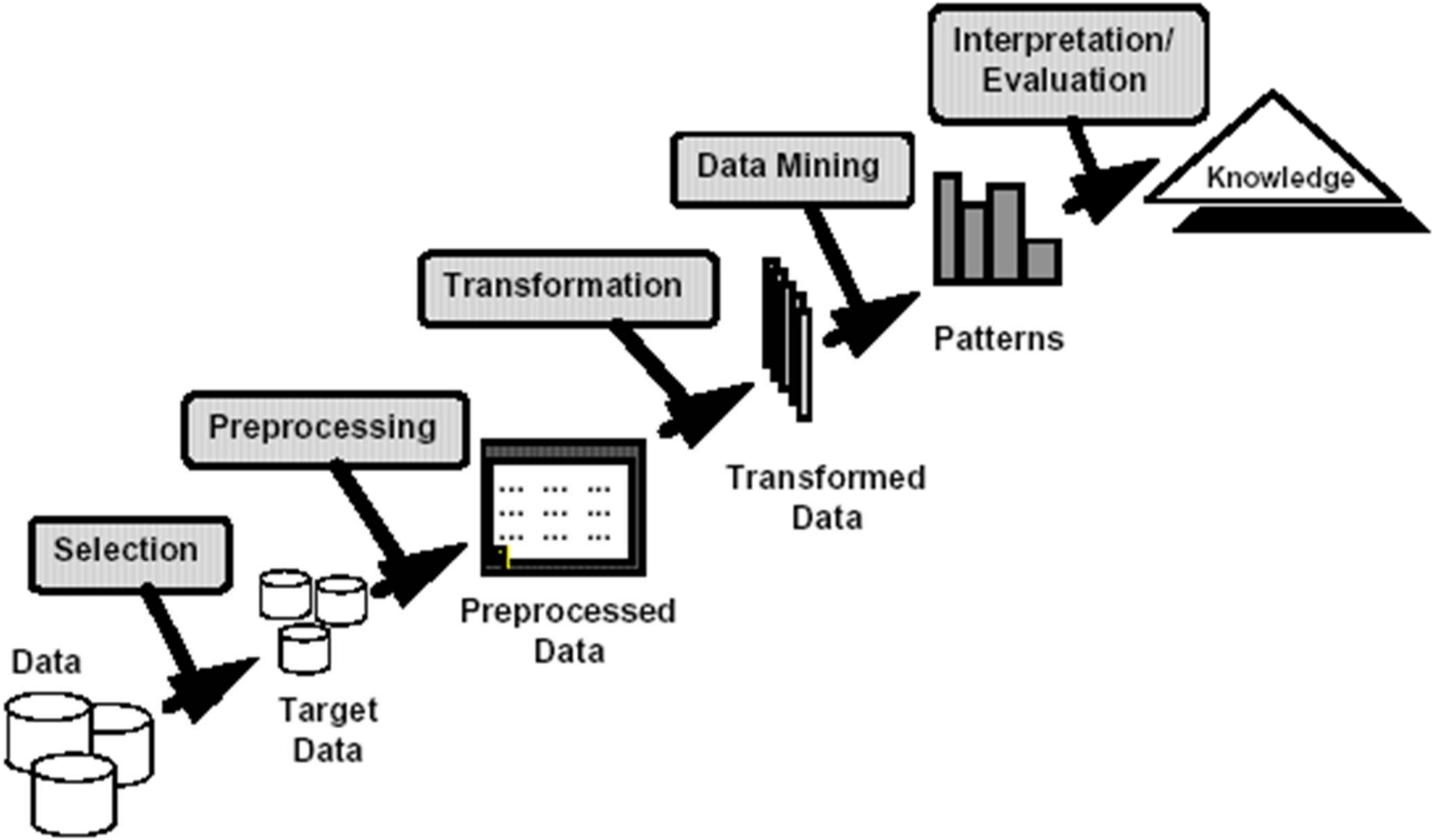
“Other” Mining Environments

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

Data Mining Functionalities

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown or missing numerical values

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

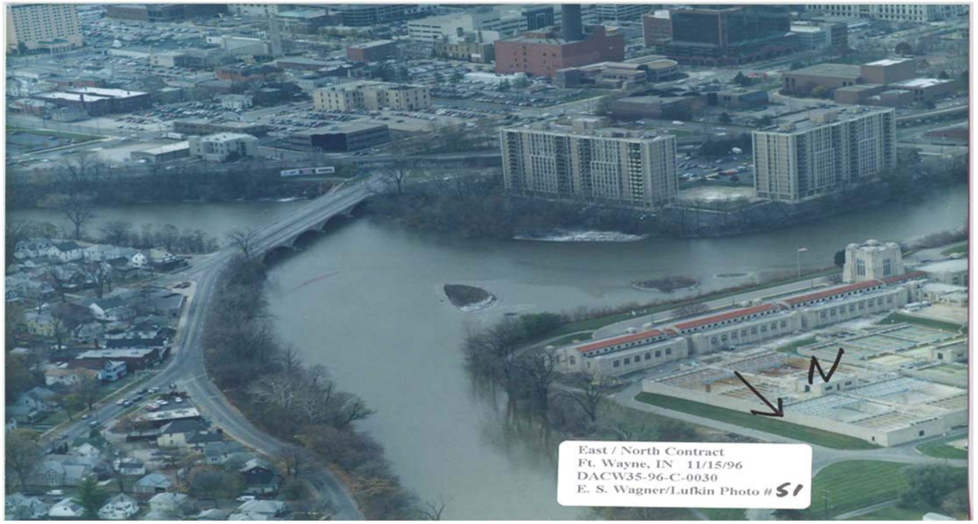
Data Mining Functionalities (2)

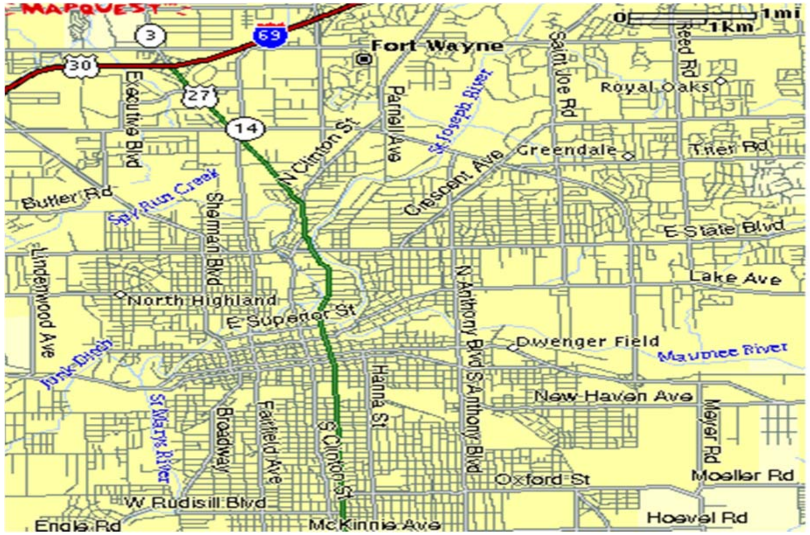
- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

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 Case Study

- Fort Wayne – IN: Flood Control Project
  - Phase I: CTRL-EAST, $4,488,450.21, 11/1/95-10/23/98
  - Phase II: East-North, $12,107,880.46, 1/6/97-11/5/98
  - Phase III: CTRL, $ 6,018,981.54, 9/14/98-8/6/99
  - Phase IV: West, 5/28/99-





From Turban et al. (2004), Information Technology for Management

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

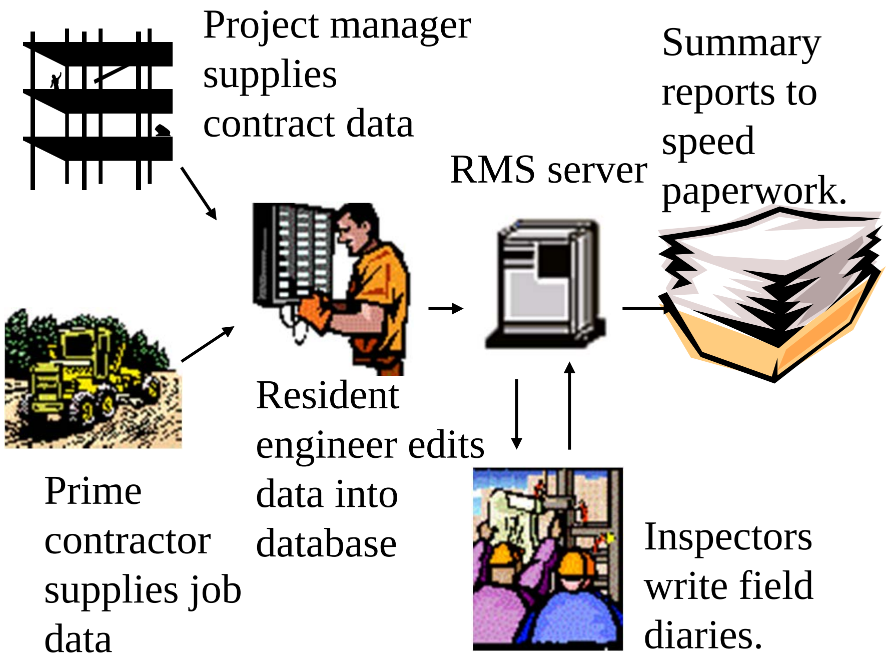
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Mining Case Study

Data Collection and Extraction:

Resident Management System - RMS

- ❑ Manages Civil Works projects.
- ❑ Was developed by US Corps of Engineers (1996)
- ❑ Consists of about 80 database tables, each of which has about more than 20 attributes.
- ❑ Contains data on construction project planning, contract administration, quality assurance, payments, correspondence, submittal management, safety and accident administration, modification processing, and management reporting.

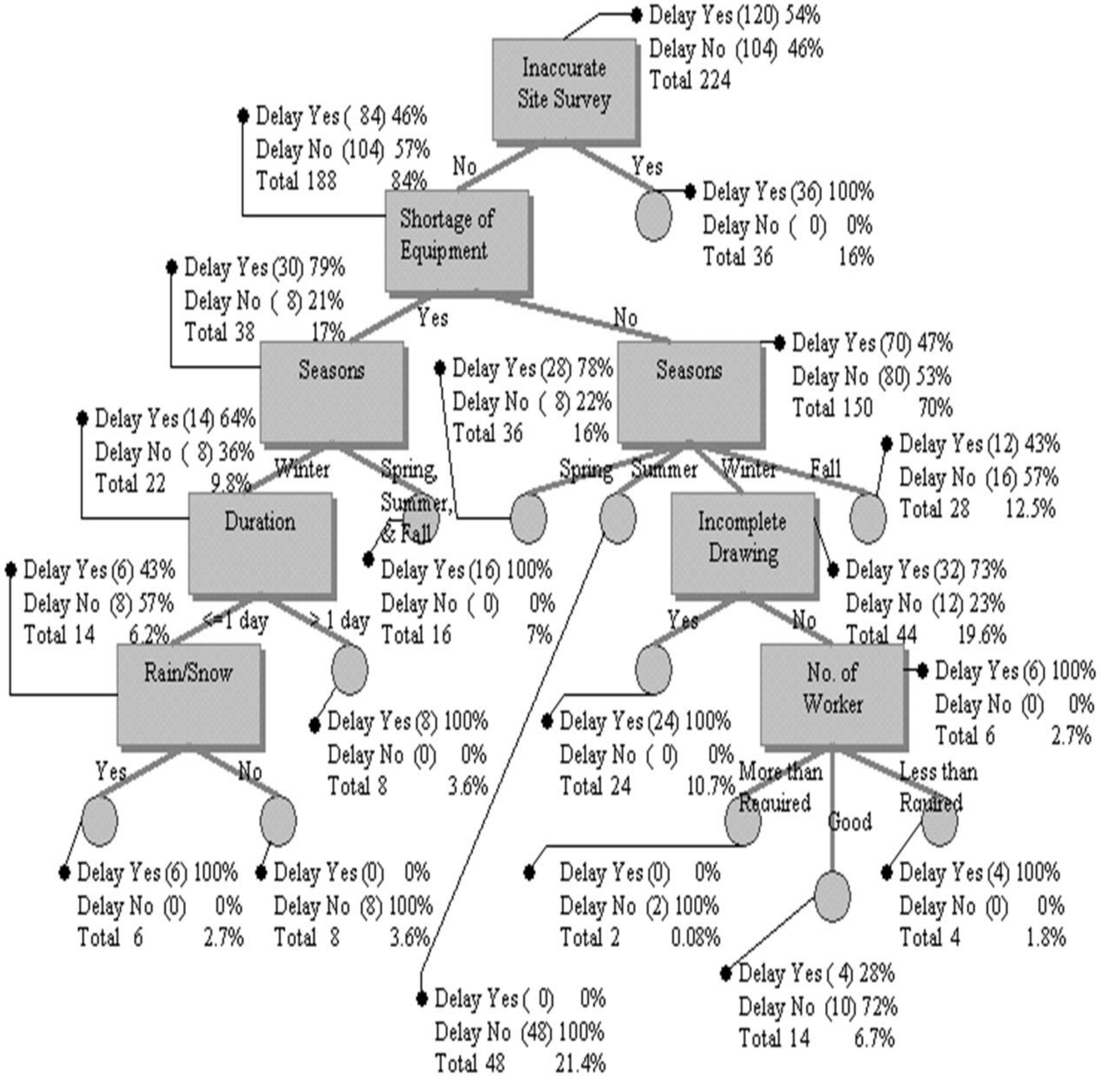

Project manager supplies contract data
Summary reports to speed paperwork.
RMS server
Resident engineer edits data into database
Prime contractor supplies job data
Inspectors write field diaries.

RMS in a construction project

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 Case Study

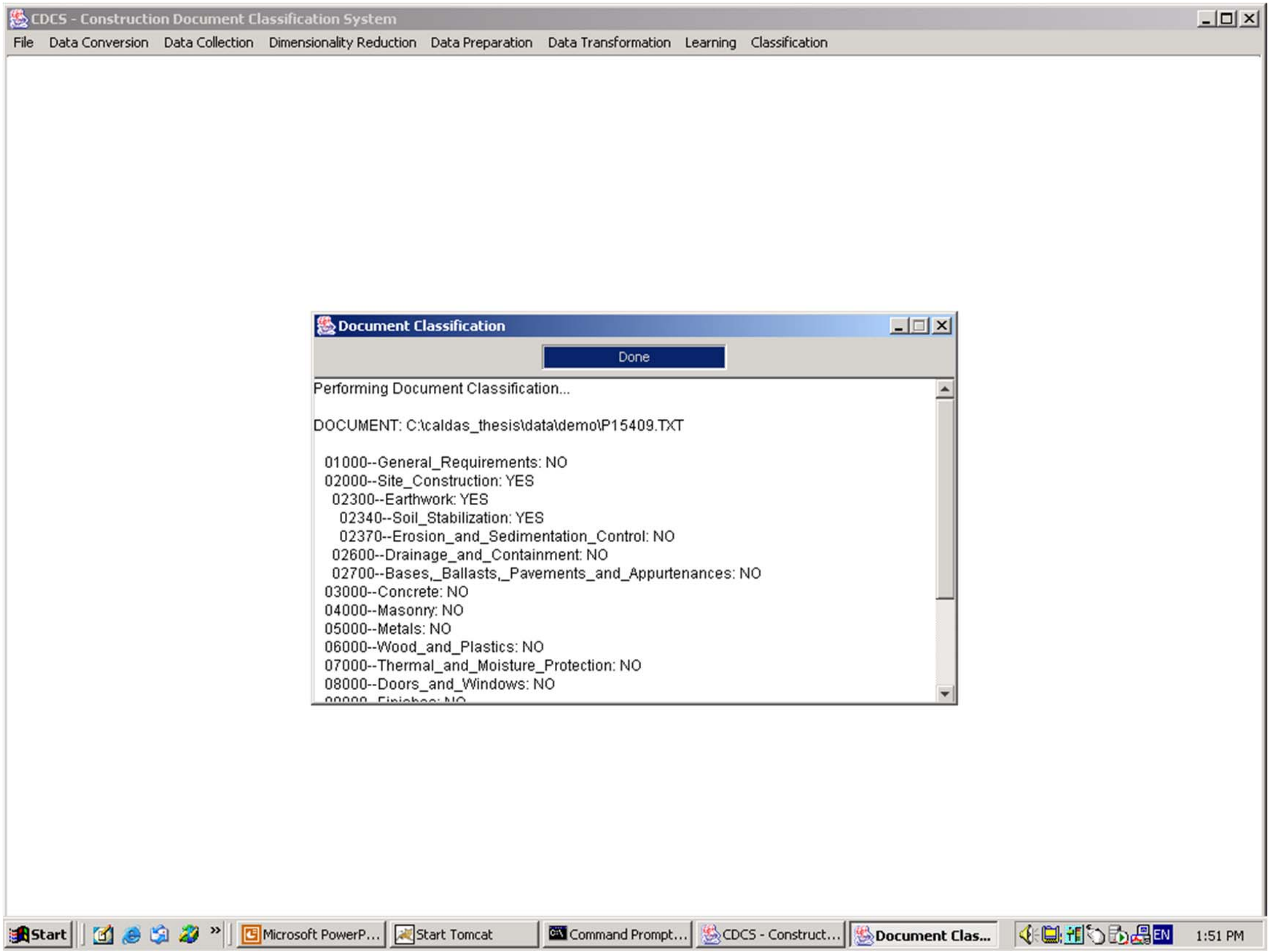
- **Results from C4.5 Decision Trees**



- Weather considered responsible for delays by site managers, appear not to be the most important cause in determining delays.
- Activities with “Inaccurate Site Surveys” are always delayed in the schedule.
- Shortage of Equipment, Seasons, and Incomplete Drawing are also very significant factors compared to other factors.

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

Document Classification

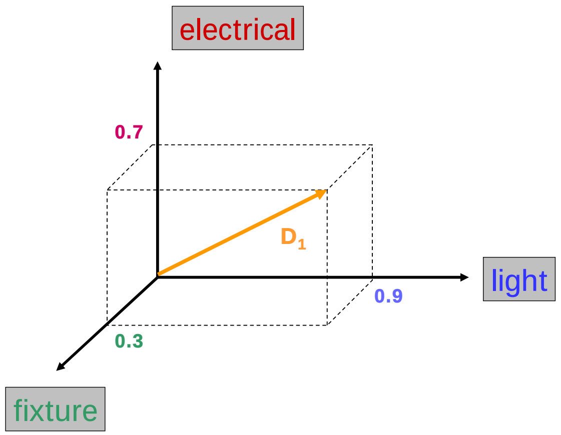


```
Performing Document Classification...

DOCUMENT: C:\caldas_thesis\data\demo\P15409.TXT

  01000--General_Requirements: NO
  02000--Site_Construction: YES
    02300--Earthwork: YES
      02340--Soil_Stabilization: YES
      02370--Erosion_and_Sedimentation_Control: NO
    02600--Drainage_and_Containment: NO
    02700--Bases,_Ballasts,_Pavements_and_Appurtenances: NO
  03000--Concrete: NO
  04000--Masonry: NO
  05000--Metals: NO
  06000--Wood_and_Plastics: NO
  07000--Thermal_and_Moisture_Protection: NO
  08000--Doors_and_Windows: NO
  09000--Finishes: NO
```

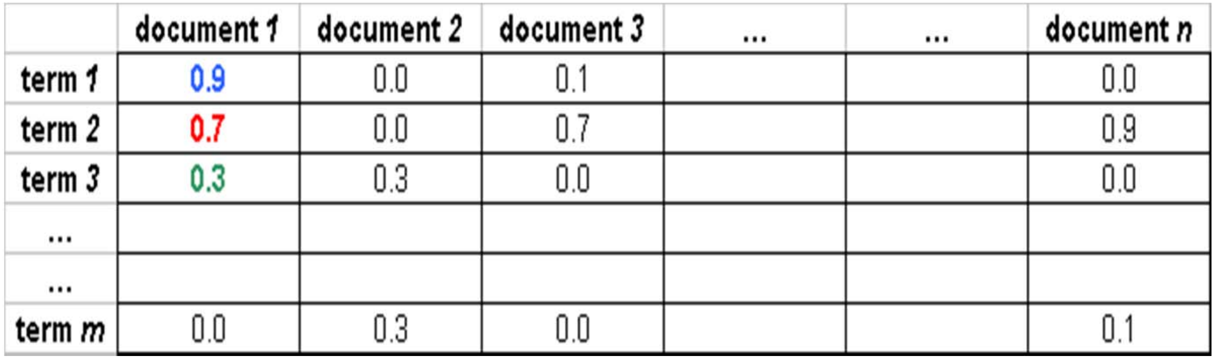
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Document Representation

- Project documents are represented as vectors in a multi-dimensional space.
- Vector coordinate values are defined by the index terms weights.
- Project document collection can be represented as a *m x n* matrix.
- Project document collection is parsed and indexed.



$$\mathbf{\vec{d}_1 = (0.9, 0.7, 0.3, \dots, 0.0)}$$

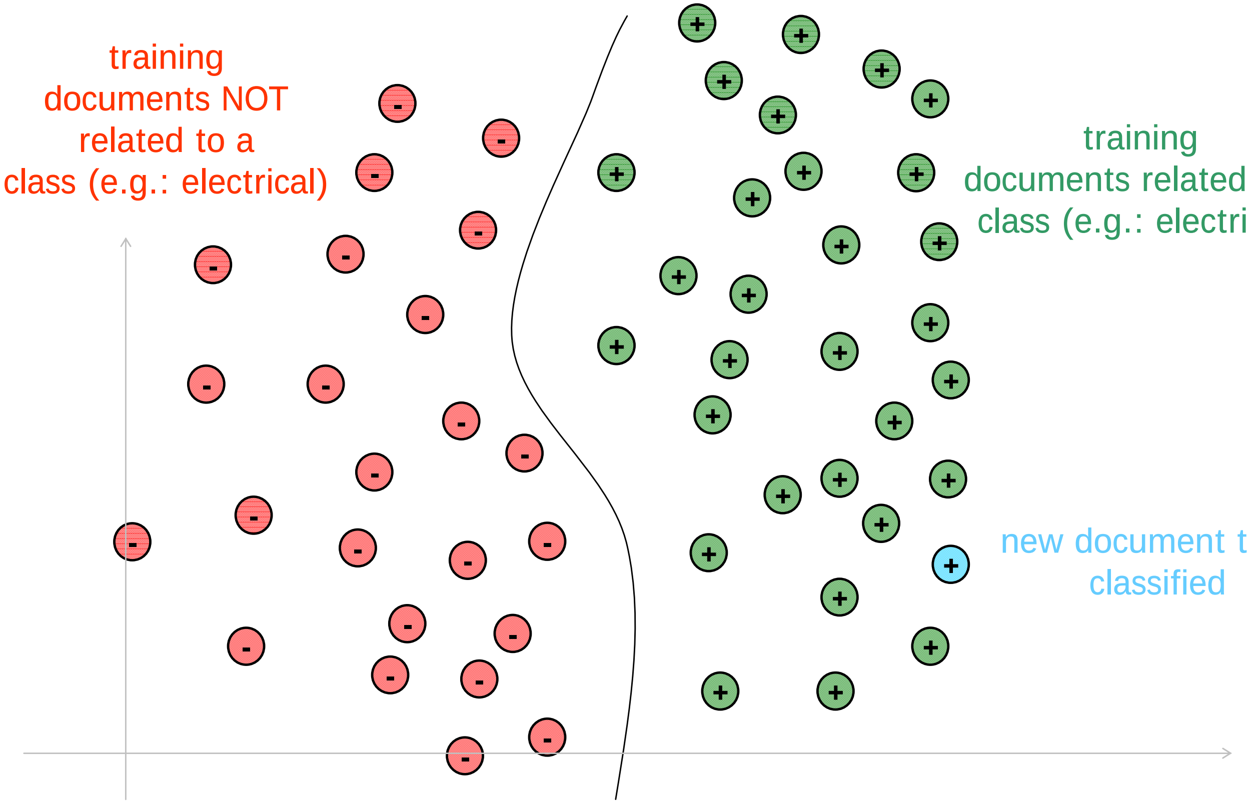


	document *1*	document *2*	document *3*	...	...	document *n*
term *1*	0.9	0.0	0.1			0.0
term *2*	0.7	0.0	0.7			0.9
term *3*	0.3	0.3	0.0			0.0
...						
...						
term *m*	0.0	0.3	0.0			0.1

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

Classification

- Previously classified documents are used to create classification models.
- Classification models are used to classify new documents.



training documents NOT related to a class (e.g.: electrical)

training documents related to a class (e.g.: electrical)

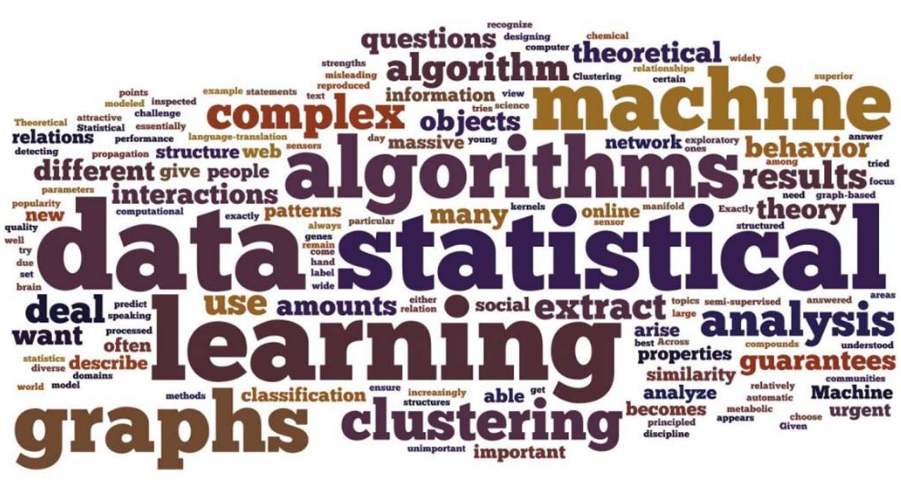
new document to be classified

Decision Boundary (classification model)

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 Visualization

- Data visualization refers to presentation of data by technologies such as digital images, geographical information systems, graphical user interfaces, multidimensional tables and graphs, virtual reality, three-dimensional presentations, videos and animation.



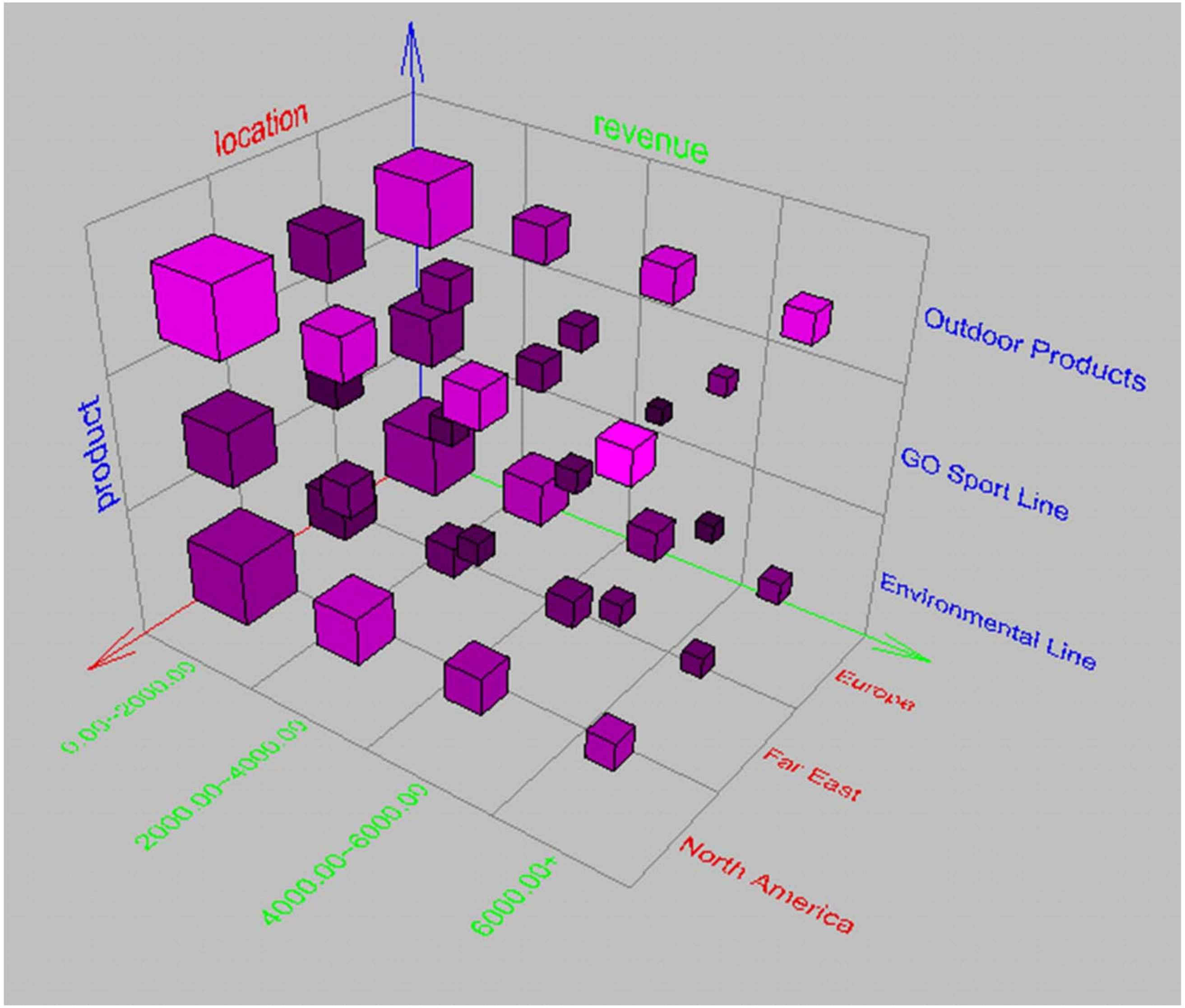
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

Data Visualization

- Geographical information system (GIS) is a computer-based system for capturing, storing, checking, integrating, manipulating, and displaying data using digitized maps. Every record or digital object has an identified geographical location. It employs spatially oriented databases.
- Visual interactive modeling (VIM) uses computer graphic displays to represent the impact of different management or operational decisions on objectives such as profit or market share.
- Virtual reality (VR) is interactive, computer-generated, three-dimensional graphics delivered to the user. These artificial sensory cues cause the user to “believe” that what they are doing is real.

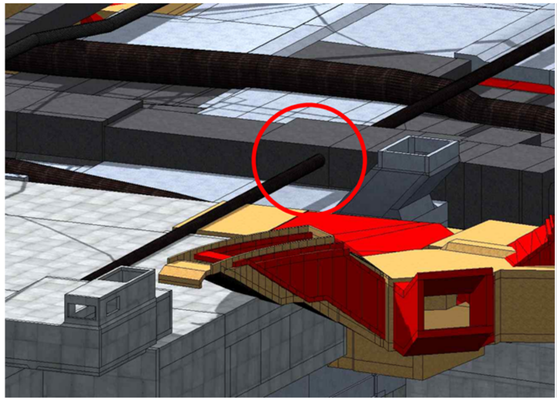
From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

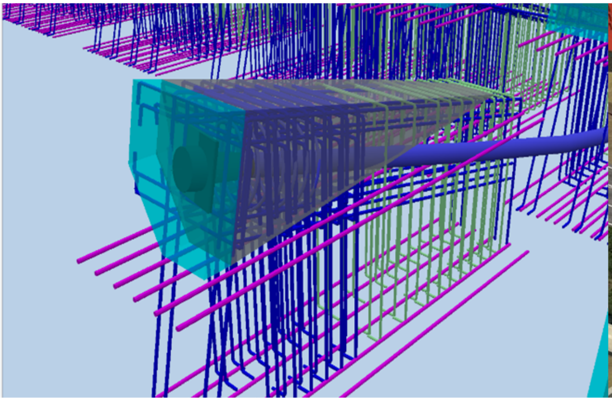
Visualization Example

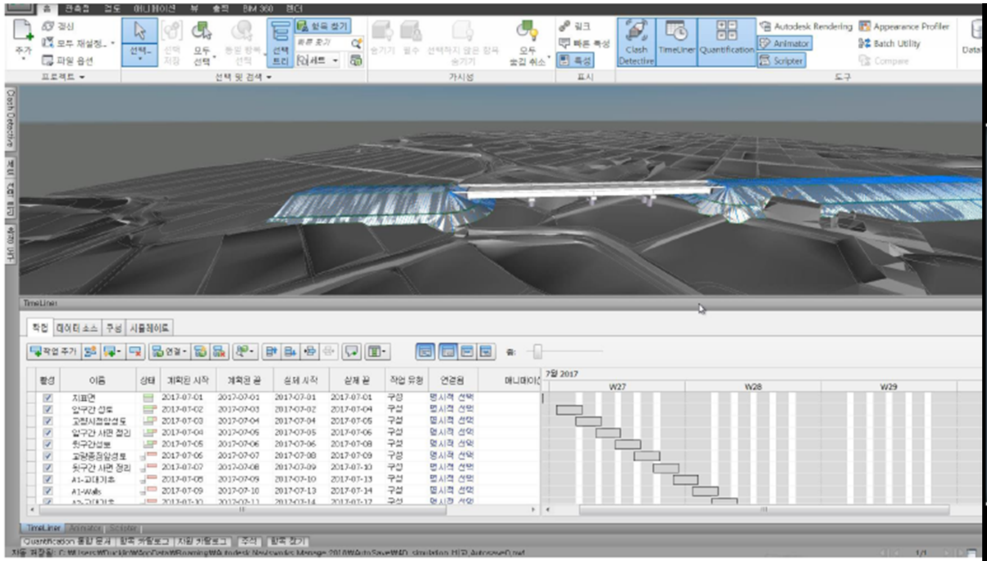


From Turban et al. (2004), Information Technology for Management
Rob and Coronel (2004), Database Systems: Design, Implementation, and Management
Han, Kamber (2001) Data Mining: Concepts and Techniques

BIM





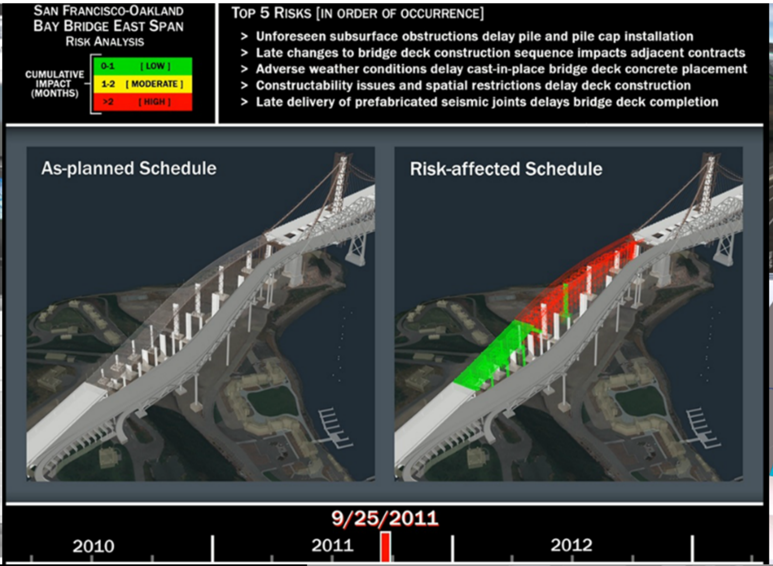


작업 | 데이터 소스 | 구성 | 시뮬레이션

활성	이름	상태	계획된 시작	계획된 끝	실제 시작	실제 끝	작업 유형	연결됨
☑	지표면		2017-07-01	2017-07-01	2017-07-01	2017-07-01	구성	명시적 선택
☑	앞구간 성토		2017-07-02	2017-07-03	2017-07-02	2017-07-04	구성	명시적 선택
☑	교량시점앞성토		2017-07-03	2017-07-04	2017-07-04	2017-07-05	구성	명시적 선택
☑	앞구간 사면 정리		2017-07-04	2017-07-05	2017-07-05	2017-07-06	구성	명시적 선택
☑	뒷구간성토		2017-07-05	2017-07-06	2017-07-06	2017-07-08	구성	명시적 선택
☑	교량종점앞성토		2017-07-06	2017-07-07	2017-07-08	2017-07-09	구성	명시적 선택
☑	뒷구간 사면 정리		2017-07-07	2017-07-08	2017-07-09	2017-07-10	구성	명시적 선택
☑	A1-교대기초		2017-07-08	2017-07-09	2017-07-10	2017-07-13	구성	명시적 선택
☑	A1-Walls		2017-07-09	2017-07-10	2017-07-13	2017-07-14	구성	명시적 선택
☑	A2-교대기초		2017-07-10	2017-07-11	2017-07-14	2017-07-17	구성	명시적 선택

7월 2017 W27 W28 W29

Quantification 통합 문서 | 항목 카탈로그 | 자원 카탈로그 | 주석 | 항목 찾기



SAN FRANCISCO-OAKLAND BAY BRIDGE EAST SPAN RISK ANALYSIS

CUMULATIVE IMPACT (MONTHS)

0-1	[LOW]
1-2	[MODERATE]
>2	[HIGH]

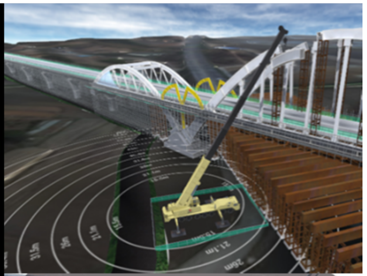
TOP 5 RISKS [IN ORDER OF OCCURRENCE]

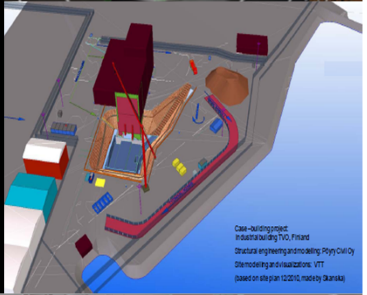
- > Unforeseen subsurface obstructions delay pile and pile cap installation
- > Late changes to bridge deck construction sequence impacts adjacent contracts
- > Adverse weather conditions delay cast-in-place bridge deck concrete placement
- > Constructability issues and spatial restrictions delay deck construction
- > Late delivery of prefabricated seismic joints delays bridge deck completion

As-planned Schedule

Risk-affected Schedule

9/25/2011



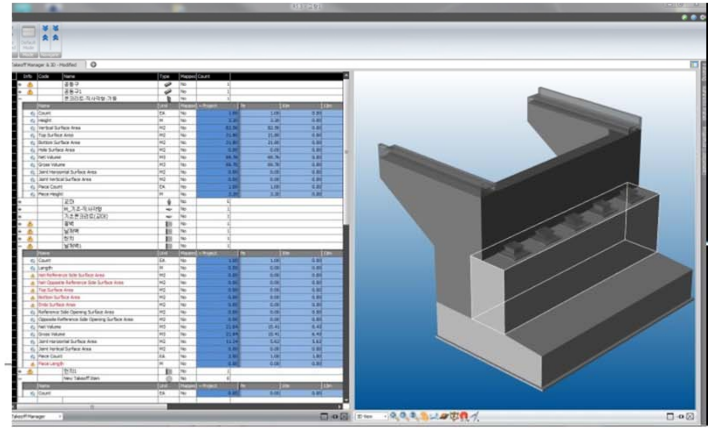


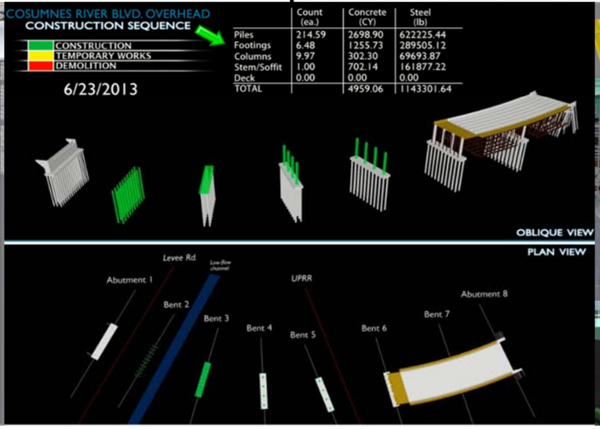
Case–building project:
Industrial building TVO, Finland

Structural engineering and modelling: Pöyry Civil Oy

Site modelling and visualizations: VTT

(based on site plan 12/2010, made by Skanska)



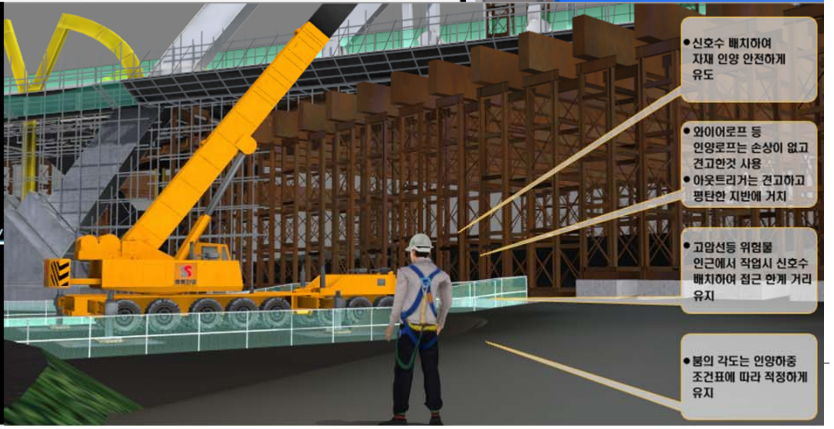


COSUMNES RIVER BLVD. OVERHEAD
CONSTRUCTION SEQUENCE

- 🟩 CONSTRUCTION
- 🟨 TEMPORARY WORKS
- 🟥 DEMOLITION

6/23/2013

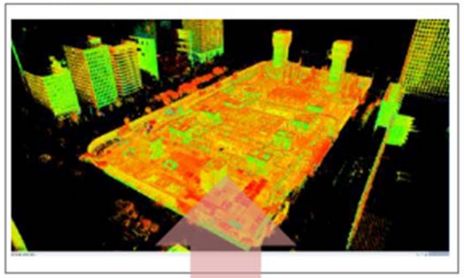
	Count (ea.)	Concrete (CY)	Steel (lb)
Piles	214.59	2698.90	622225.44
Footings	6.48	1255.73	289505.12
Columns	9.97	302.30	69693.87
Stem/Soffit	1.00	702.14	161877.22
Deck	0.00	0.00	0.00
TOTAL		4959.06	1143301.64



- 신호수 배치하여 자재 인양 안전하게 유도
- 와이어로프 등 인양로프는 손상이 없고 견고한것 사용
- 아웃트리거는 견고하고 평탄한 지반에 거치
- 고압선등 위험물 인근에서 작업시 신호수 배치하여 접근 한계 거리 유지
- 붐의 각도는 인양하중 조건표에 따라 적정하게 유지

BIM

현황 : SCAN DATA



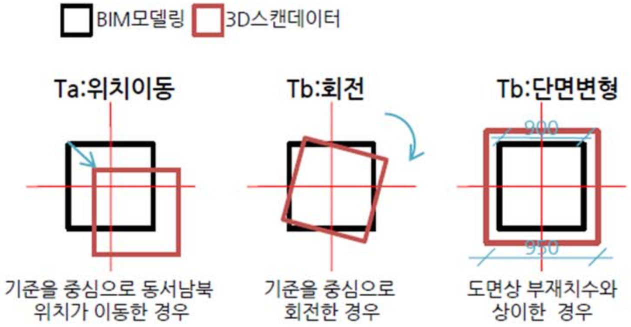
3D SCAN 데이터와 BIM모델링
DATA와 중첩을 통해 정합성 검토

도면 : BIM Model

보고서 작성 범례

1. 카테고리 구분

C.1	시공오차	Ta	위치이동
		Tb	회전
		Tc	단면변형
C.2	시공누락		
C.3	시공도서상이		
C.4	정보부족		



□ BIM모델링 □ 3D스캔데이터

Ta:위치이동

기준을 중심으로 동서남북 위치가 이동한 경우

Tb:회전

기준을 중심으로 회전한 경우

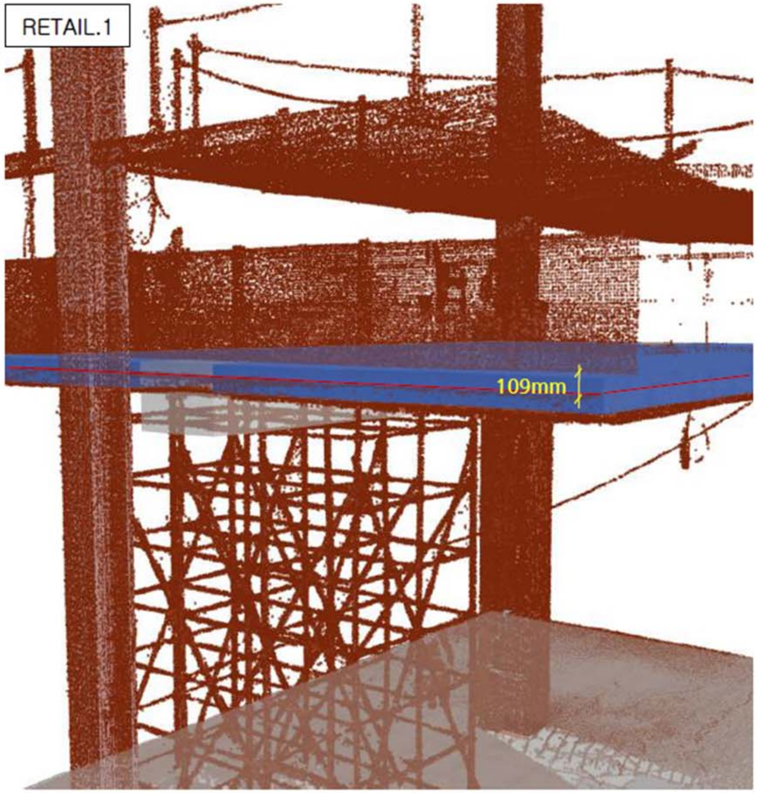
Tb:단면변형

도면상 부재치수와 상이한 경우

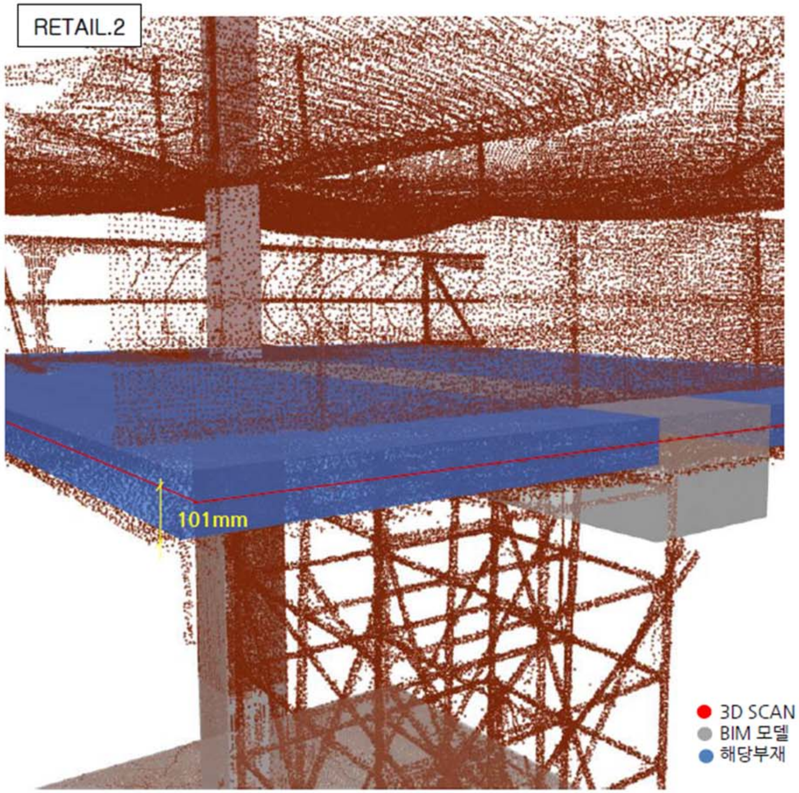
2. 오차범위 별 색깔 구분

구분	기준		
	위치이동	회전변형	단면변형
🟥 A	51mm이상	2.1°이상	두께의7% 초과
🟦 B	36~50mm	1.6° ~ 2.0°	두께의 5% 초과 두께의 7% 이하
🟨 C	21~35mm	1.0° ~ 1.5°	두께의 3% 초과 두께의 5% 이하

시공오차	내용	슬라브가 하부 방향으로 처짐 : A Level	위치 정보	RETAIL.B-B1F-009, RETAIL.D-B1F-010

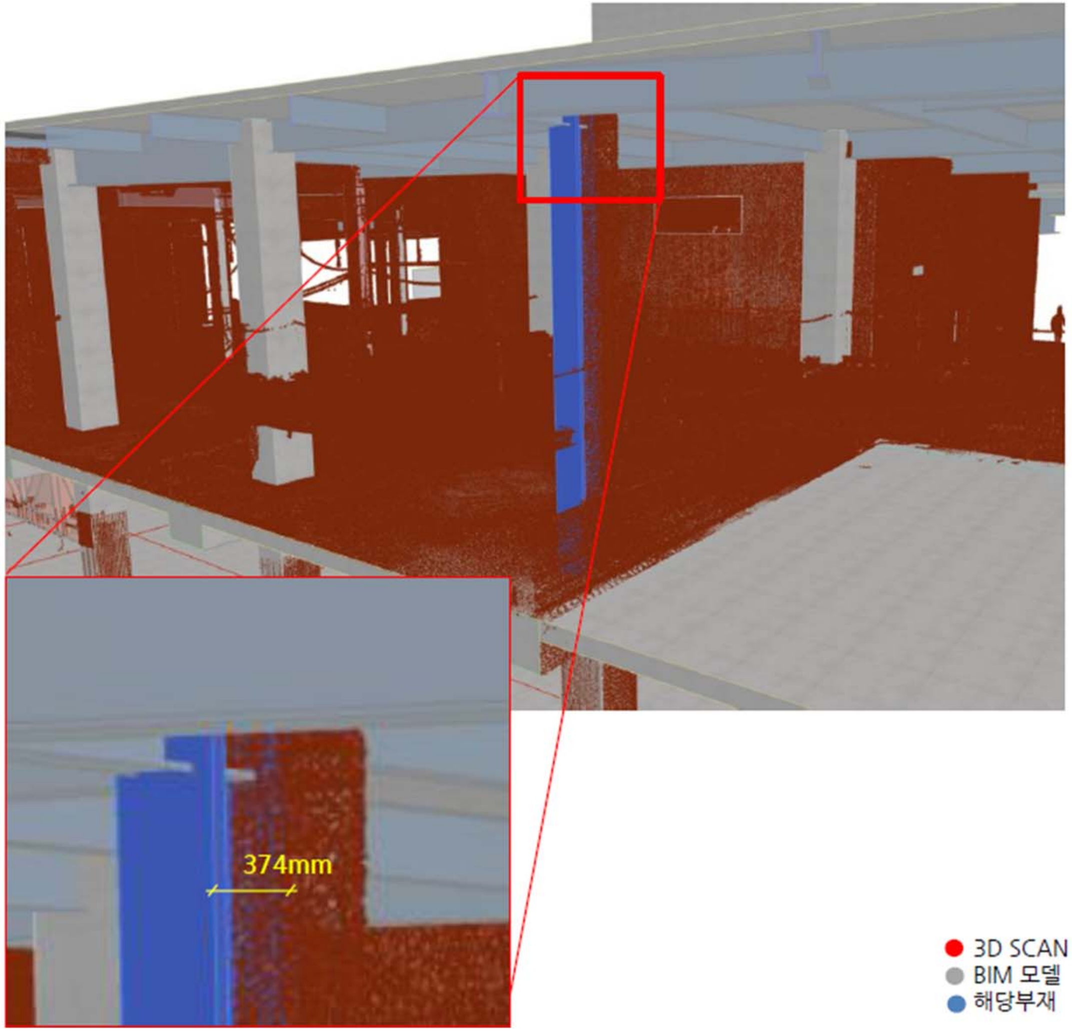


RETAIL.1



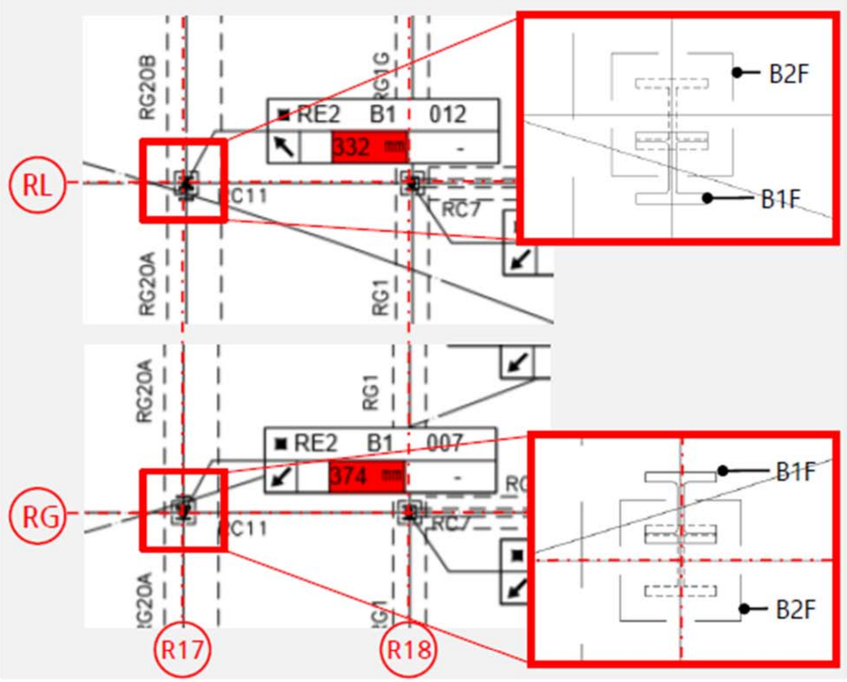
RETAIL.2

시공오차/위치이동	내용 RC11기둥 계획도 위치와 시공 위치 상이 : A Level	위치 정보	RETAIL2-B1F-007, 012

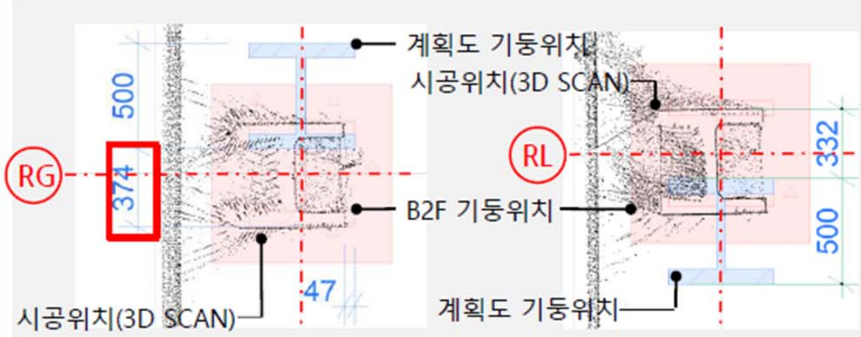


● 3D SCAN
● BIM 모델
● 해당부재

KEY MAP



평면 이미지



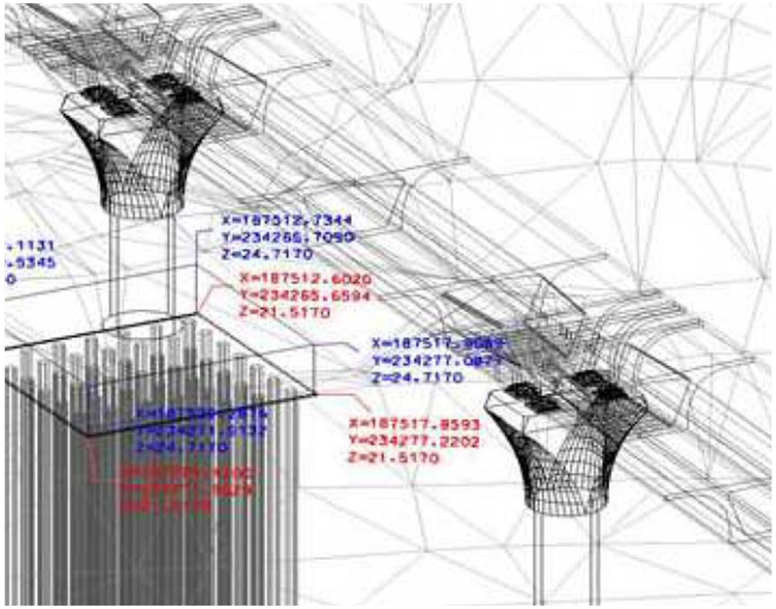
BIM

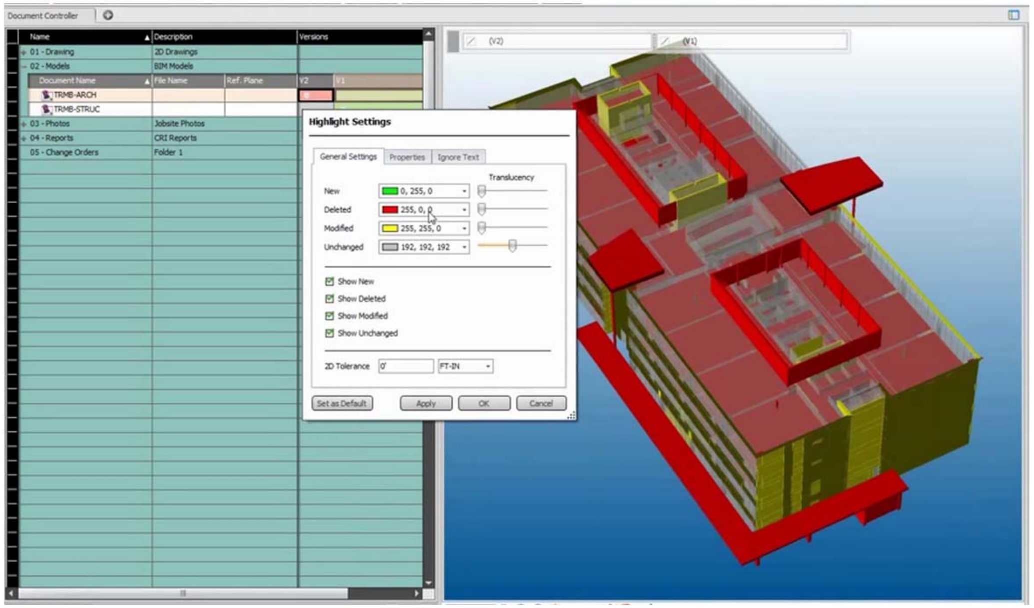
좌표기준점으로의 활용

- BIM 모델이 소수 측량 기준점을 기준으로 정확하게 배치되었을 경우, 모델의 임의의 위치에서의 좌표를 즉각적으로 얻을 수 있기 때문에 추가적인 측량 작업을 대신할 수 있음

설계변경 이력관리

- BIM 모델과 설계변경 문서 연계
- 계약 문서와 3D 모델 비교





Name	Description	Versions
01 - Drawing	2D Drawings	
02 - Models	BIM Models	

Document Name	File Name	Ref. Plane	V2	V1
TRMB-ARCH				
TRMB-STRUC				

Name	Description	Versions
03 - Photos	Jobsite Photos	
04 - Reports	CRI Reports	
05 - Change Orders	Folder 1	

Highlight Settings

General Settings | Properties | Ignore Text

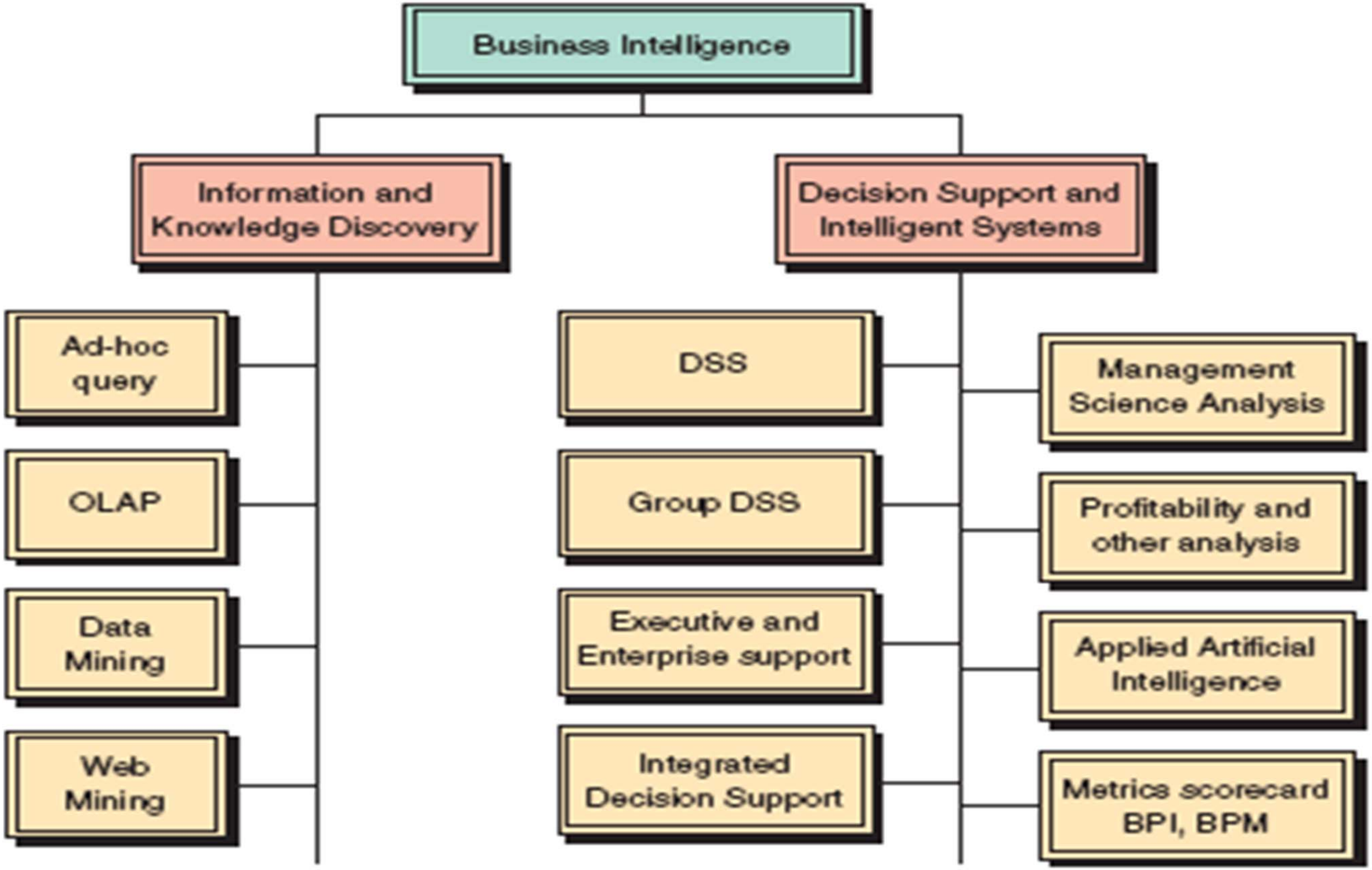
		Translucency
New	🟩 0, 255, 0	
Deleted	🟥 255, 0, 0	
Modified	🟨 255, 255, 0	
Unchanged	⬜ 192, 192, 192	

☑ Show New
☑ Show Deleted
☑ Show Modified
☑ Show Unchanged

2D Tolerance 0' FT-IN

Set as Default | Apply | OK | Cancel

Business Intelligence


Business Intelligence
Information and Knowledge Discovery
Decision Support and Intelligent Systems
Ad-hoc query
OLAP
Data Mining
Web Mining
DSS
Group DSS
Executive and Enterprise support
Integrated Decision Support
Management Science Analysis
Profitability and other analysis
Applied Artificial Intelligence
Metrics scorecard BPI, BPM

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