What is Data Mining?

- Discovering **meaningful patterns and trends** using some mathematical algorithm on huge amount of stored data
- Extraction of **interesting, non-trivial, implicit, previously unknown and potentially useful** information or patterns from data
- **Analysis of data and the use of software** techniques for finding patterns and regularities in sets of data

What is Data Mining?

Data mining is the **exploration and analysis** of large quantities of data in order to discover **valid, novel, potentially useful, and ultimately understandable** patterns in data.

Valid: The patterns hold in general.
Novel: We did not know the pattern beforehand.
Useful: We can devise actions from the patterns.
Understandable: We can interpret and comprehend the patterns.

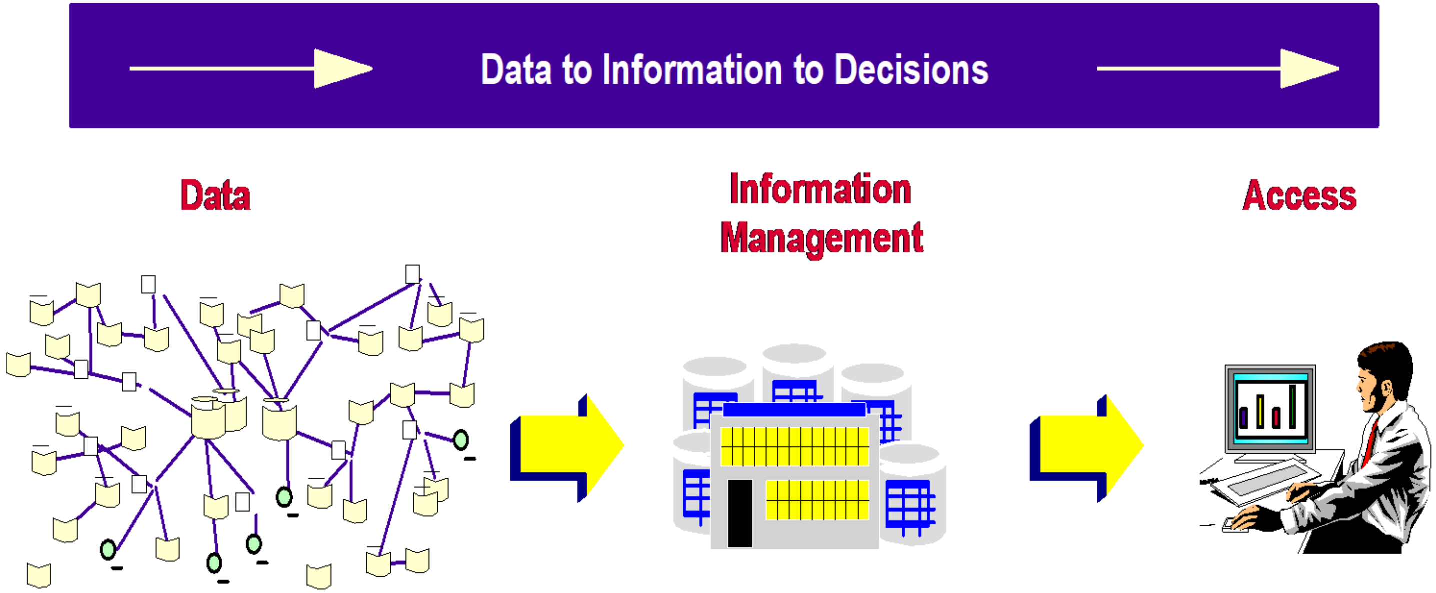
What is Data Mining?

Finding interesting structure in data

- *Structure:* refers to **statistical patterns, predictive models, hidden relationships**

- Examples of tasks addressed by Data Mining
  – Predictive Modeling (classification, regression)
  – Segmentation (Data Clustering)
  – Summarization
  – Visualization

Data… Information….Decisions



Data to Information to Decisions

Data

- On-line Updates
- Batch Feeds
- Operational Data Store

Information Management

- Data Warehouse
- Data Mart
- Data Transformation
- Data Synchronization

Access

- Query & Reporting
- Data Mining
- On-line Analytical Processing
- Summary and detail
- Drill capability

Knowledge discovery in databases

- KDD is the process of identifying valid,potentially useful and understandable patterns & relationships in data

⌦ Knowledge = patterns & relationships

knowledge discovery =

data preparation + data mining + evaluation/interpretation of discovered patterns/relationships

- Nowadays, ⌦ KDD = data mining

Data Mining Step in Detail

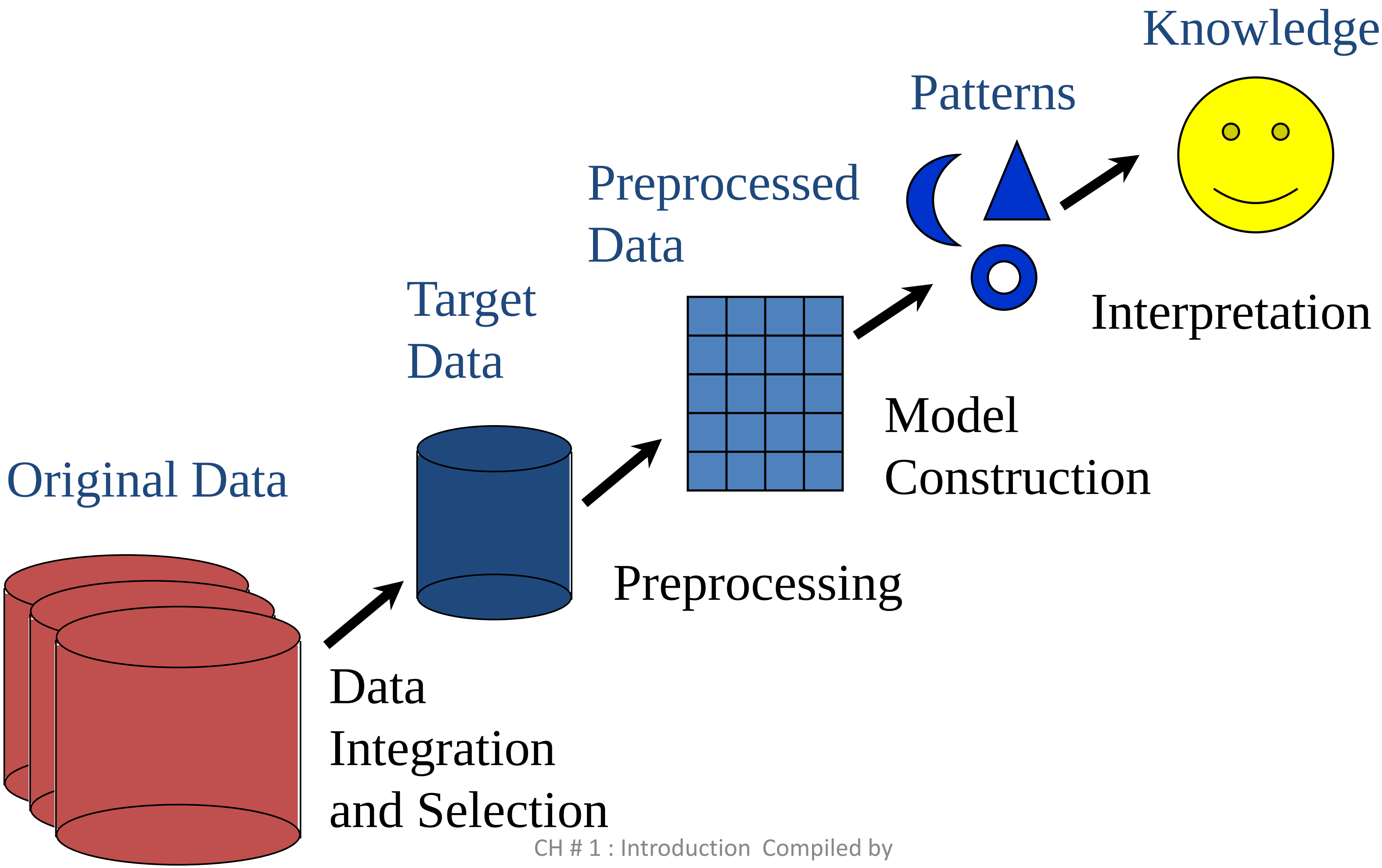
2.1 Data preprocessing

- Data selection: Identify target datasets and relevant fields
- Data cleaning
  - Remove noise and outliers
  - Data transformation
  - Create common units
  - Generate new fields

2.2 Data mining model construction

2.3 Model evaluation

Preprocessing and Mining



Original Data

Data Integration and Selection

Target Data

Preprocessing

Preprocessed Data

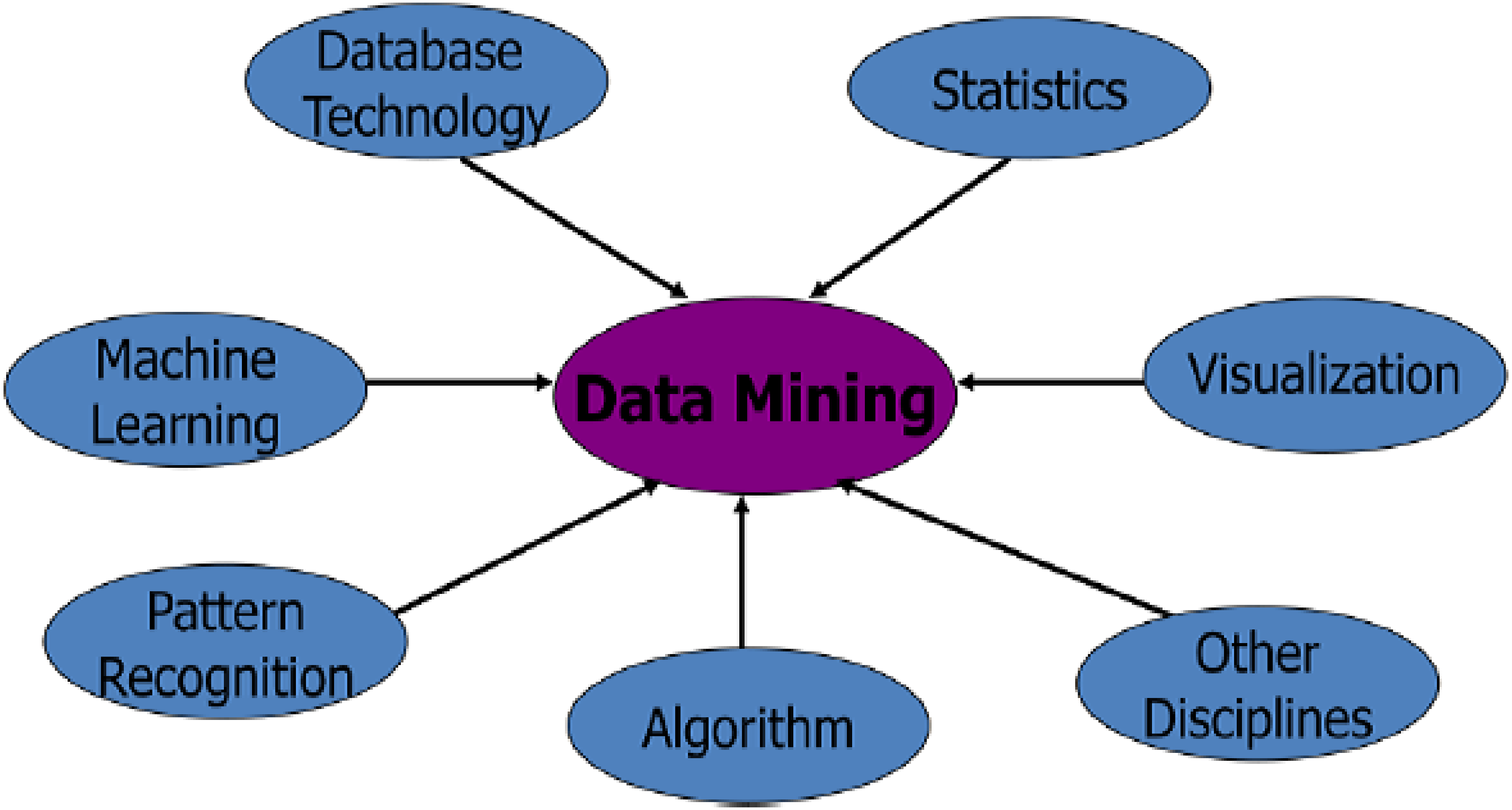
Model Construction

Patterns

Interpretation

Knowledge

Data Mining: Confluence of Multiple Disciplines



Statistics, Machine Learning and Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic (approach to problem solving, learning, or discovery that employs a practical method)
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results

The Need for Data Mining

- The **amount of raw data** stored in corporate data warehouses is **growing rapidly**.
- There is **too much data and complexity** that might be relevant to a specific problem.
- Data mining promises to bridge the analytical gap by giving knowledge workers the **tools to navigate this complex analytical space.**
- The need for information has resulted in the proliferation of data warehouses that integrate information multiple sources **to support decision making.**
- Often include **data from external sources**, such as customer demographics and household information.

Data Mining Motivation

- Changes in the Business Environment
  - Customers becoming more demanding
  - Markets are saturated
- Databases today are huge:
  - More than 1,000,000 entities/records/rows
  - From 10 to 10,000 fields/attributes/variables
  - Gigabytes and terabytes
- Databases a growing at an unprecedented rate
- Decisions must be made rapidly
- Decisions must be made with maximum knowledge

Why Use Data Mining Today?

Human analysis skills are inadequate:

- Volume and dimensionality of the data
- High data growth rate

Availability of:

- Data
- Storage
- Computational power
- Off-the-shelf software
- Expertise

Why Use Data Mining Today?

Competitive pressure!

"The secret of success is to know something that nobody else knows."

Aristotle Onassis

- Competition on service, not only on price (Banks, phone companies, hotel chains, rental car companies)
- Personalization, CRM
- The real-time enterprise
- "Systemic listening"
- Security, homeland defense

An Abundance of Data

- Supermarket scanners, POS data
- Preferred customer cards
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Demographic data
- Sensor networks
- Cameras
- Web server logs
- Customer web site trails

Data Mining: On What Kinds of Data?

- **Database-oriented data sets and applications**

  -Relational database, data warehouse, transactional database

- **Advanced data sets and advanced applications**

  -Data streams and sensor data

  -Time-series data, temporal data, sequence data

  -Structure data, graphs, social networks and multi-linked data

  -Object-relational databases

  -Heterogeneous databases and legacy databases

  -Spatial data and spatiotemporal data

  -Multimedia database, Text databases, The World-Wide Web

Database Processing vs. Data Mining Processing

- Query
  - Well defined
  - SQL

- Data
  - Operational data

- Output
  - Precise
  - Subset of database

- Query
  - Poorly defined
  - No precise query language

- Data
  - Not operational data

- Output
  - Vague
  - Not a subset of database

Data Mining Vs. Database Query

Data Mining

-interesting patterns and facts such as what are the important trends in sells

-more faster than query in trend and pattern analysis

-uses algorithm like machine learning, genetic algorithm

-If we know only vaguely what we are looking for we use data mining.

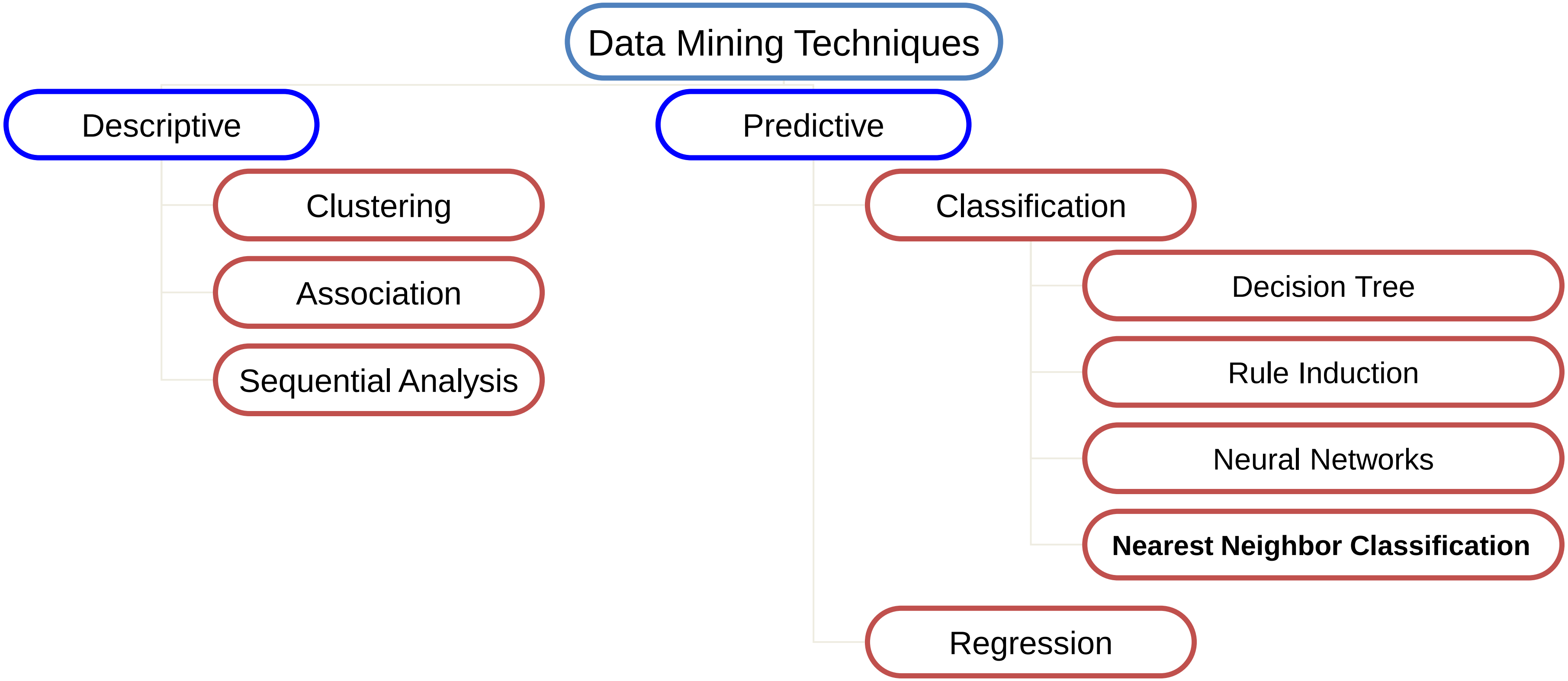
Database

- Normal queries from the database such as what is an average turnover
- If we know exactly what we are looking for, we use query

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Techniques



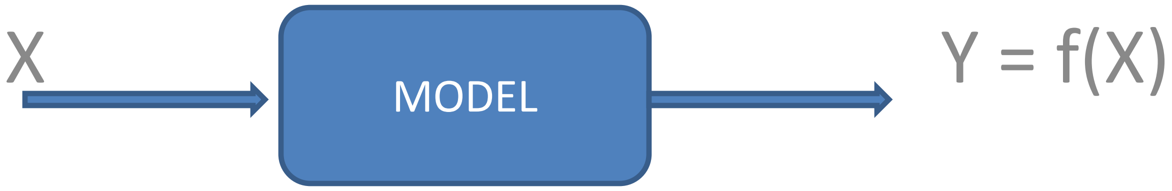
Descriptive Data Mining

- Characterizes the general properties of data in the database
- Finds important patterns or information in data
- Mostly used during data exploration
- It answered following questions:

    What is in the data? What does(n't) it look like? Are there any unusual patterns? What does the data suggest for customer segmentation?

- Functionalities: Clustering, Summarization, Visualization, and Association

Predictive Data Mining



Where X: Vectors of independent variables

Y: Dependent Variable

- It answered following type of questions:

  Who is likely to respond to next product?

  Which customers are likely to leave in the next six months?

- Used to predict outcomes whose inputs are known but the output values are not realized yet.

Data Mining Applications

- **Sales/Marketing**

-What product combinations were purchased together

-Promote their most profitable products and maximize the profit.

-Identify customer's behavior buying patterns.

-Who viewed/bought products in Ecommerce sites

- **Data Mining Applications in Transportation**

-Distribution schedules among warehouses and outlets and analyze loading patterns.

Data Mining Applications

- **Health Care, Medicine , and Insurance**

-Forecasts which customers will potentially purchase new policies.

-Insurance companies to detect risky customers' behavior patterns.

-Biomedical and DNA data analysis

-Identify the patterns of successful medical therapies

for different illnesses.

- **Retail**

-Receive a loyalty, upsell and cross-sell offers, whereas the latter may be offered a win-back deal, for instance.

-Merchandising/display plan

-Customers retention with promotions/sales- record who comes for offers and attract them with cheap products

Data Mining Applications

- **Banking / Finance**

-Adjusting credit scoring for banking institutions

-Credit card fraud detection.

-Identify customers loyalty by analyzing the data of customer's purchasing activities

-Launch different special offers to retain those customers

identify stock trading rules.

- **Websites optimization and searching for "long tail"**
- **Video hosting services to adjust user interface and to improve user experience**
- **Telecommunication industry**

Data Mining Process

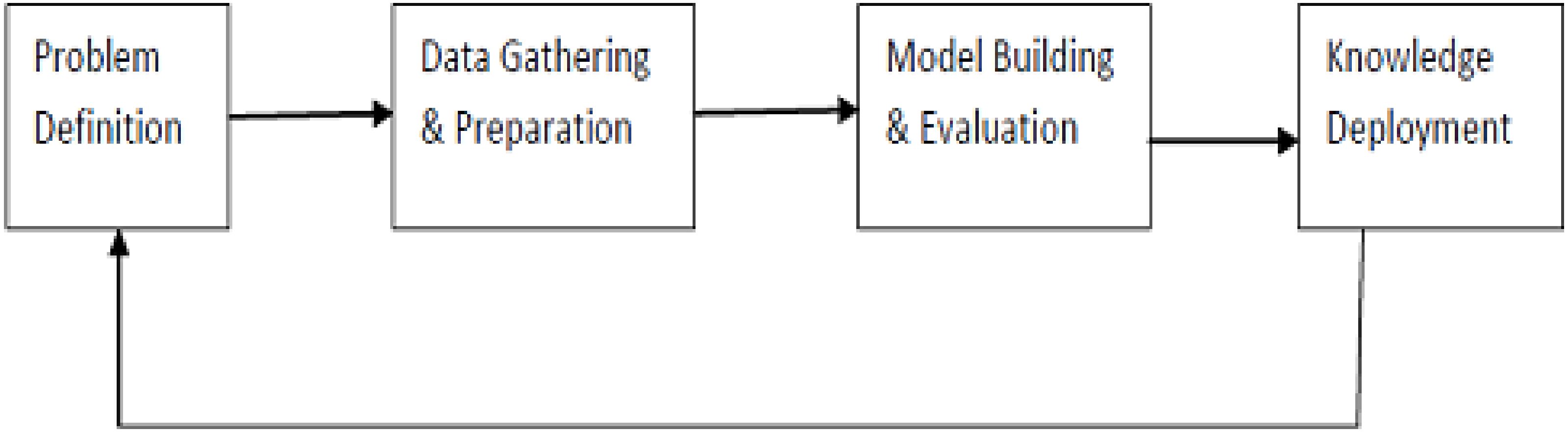


Fig: "Data mining process flow"

Problem Definition:

Focuses on Understanding the project objectives and requirements in terms of business perspective.

Eg: How can I sell more of my product to customer? Which customers are most likely to purchase the product?

Data Mining Process …

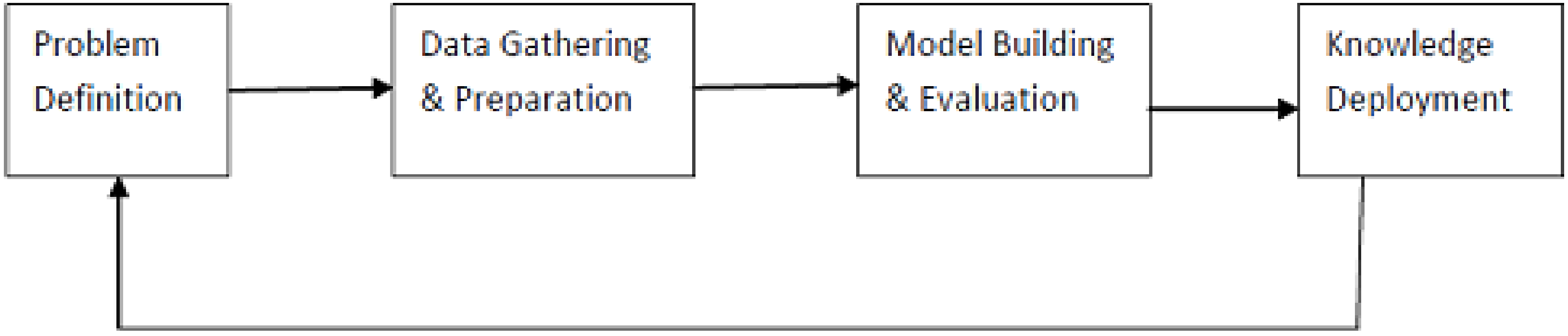


Fig: "Data mining process flow"

Data Gathering and Preparation:

- Data Collection & Exploration.

- Identify data quality, patterns in data.

- Data preparation phase covers all the tasks involved to build the model.

- Data preparation tasks are likely to be performed multiple and not in any prescribed order.

Data Mining Process …

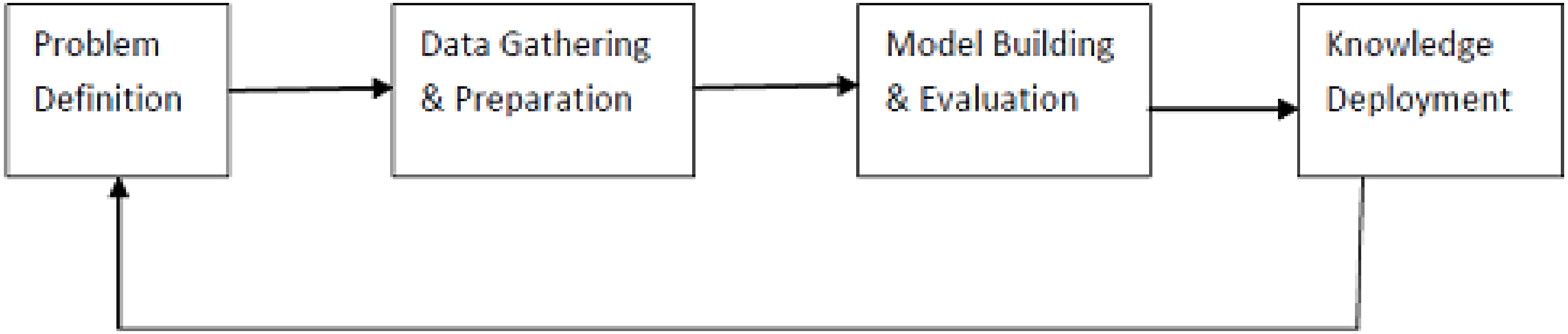


Fig: "Data mining process flow"

Model Building and Evaluation:

-Various modeling techniques are applied and calibrated the parameters to optimal values.

-Evaluate how well the model satisfies the originally stated business goal

Data Mining Process …

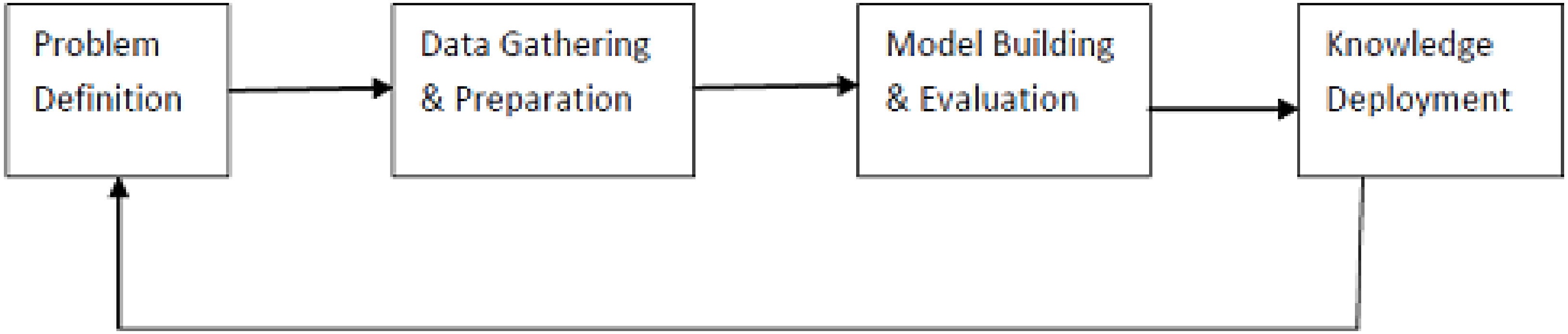


Fig: "Data mining process flow"

Knowledge Deployment:

- Use data mining within a target environment.

- Insight and actionable information can be derived from data

Data, Data, Data everywhere yet ...

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences

⌘ I can't get the data I need

⍓ need an expert to get the data

⌘ I can't understand the data I found

⍓ available data poorly documented

⌘ I can't use the data I found

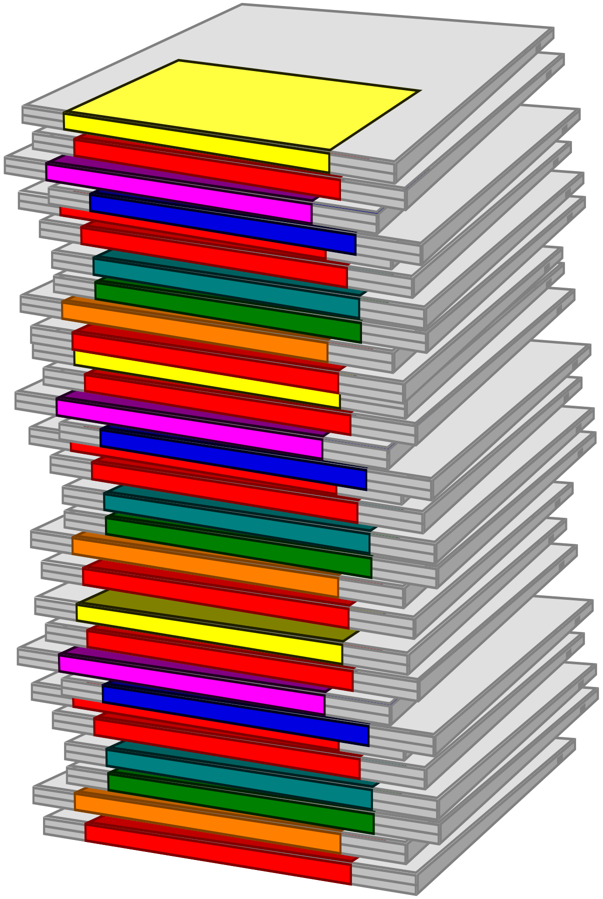
⍓ results are unexpected

⍓ data needs to be transformed from one form to other

What is a Data Warehouse?

A single, complete and consistent store of data obtained from a variety of different sources made available to end users in a what they can understand and use in a business context.

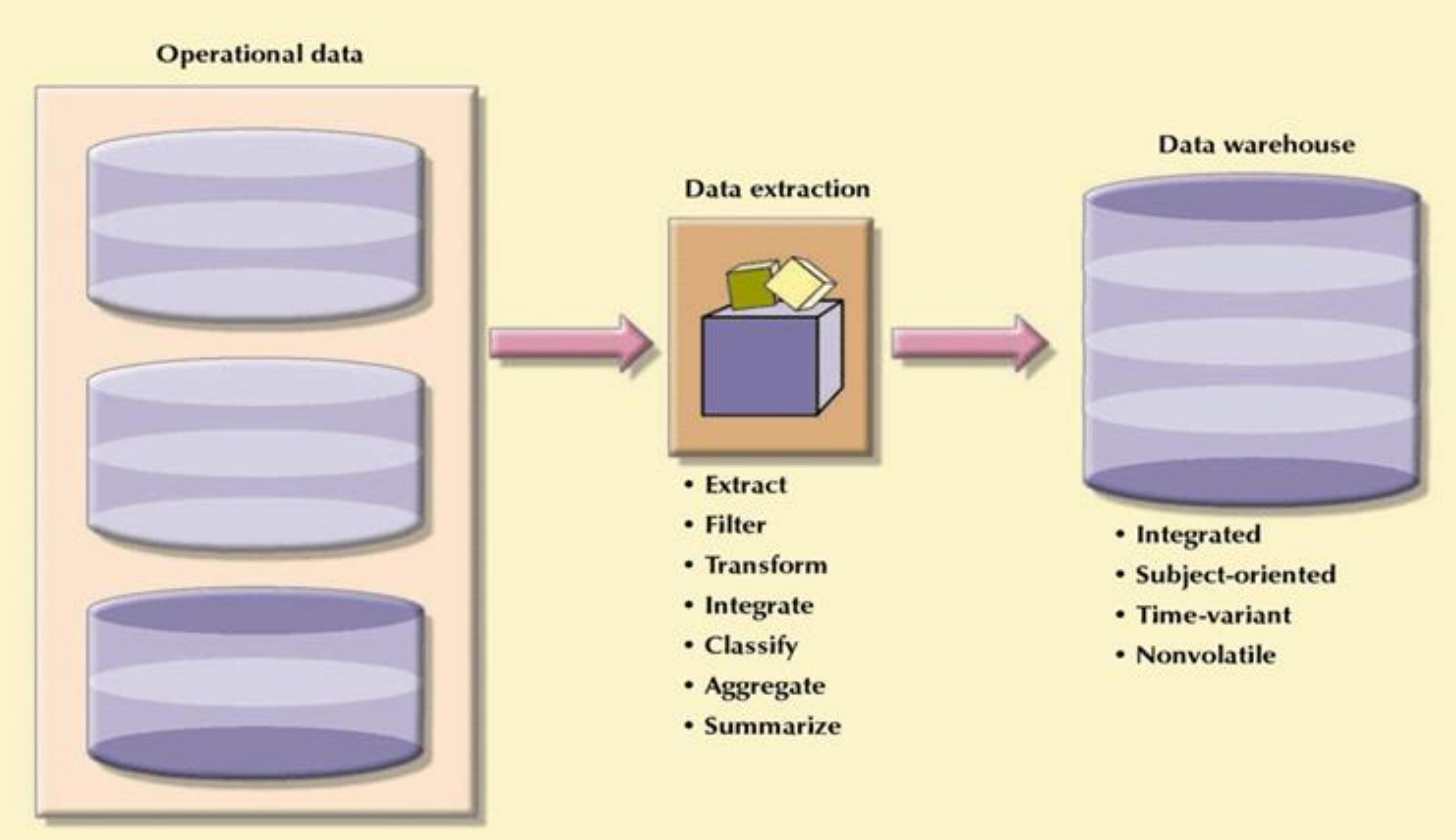
[Barry Devlin]



Data Warehouse

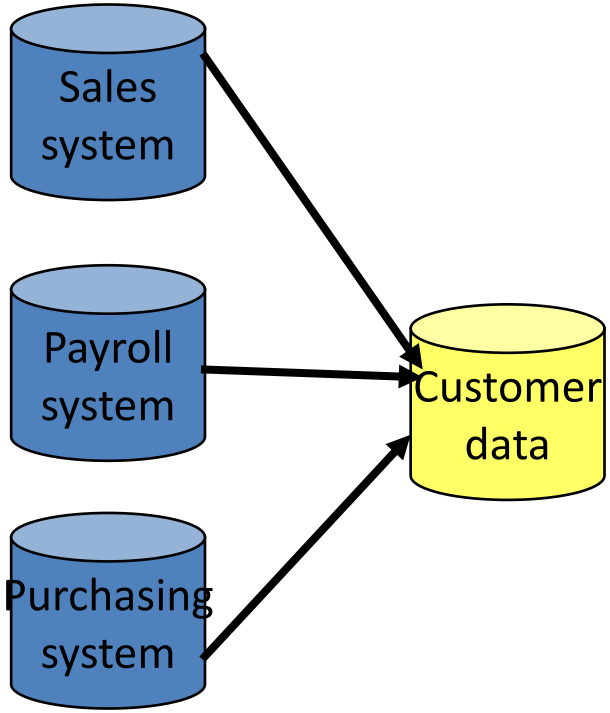
- System that organizes all the data available in an organization, makes it accessible & usable for the all kinds of data analysis and also allows to create a lots of reports by the use of mining tools.
- Large database built from the operational database
- Different from operational database that includes normal daily transactions
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

Creating a Data Warehouse



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 converted.

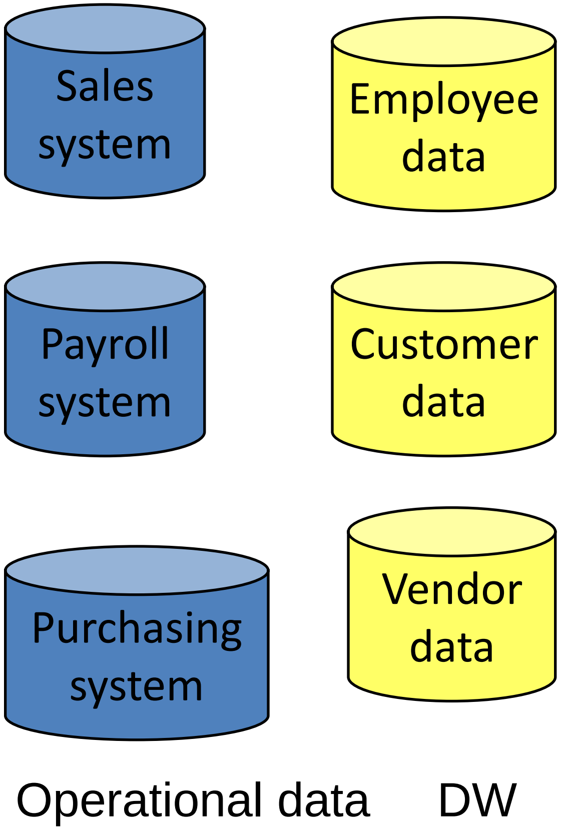


Data Warehouse—Time Variant

- 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".
  - ✓ There must be a connection between the information in the warehouse and the time when it was entered.
  - ✓ Information can be sourced according to period while mining data.

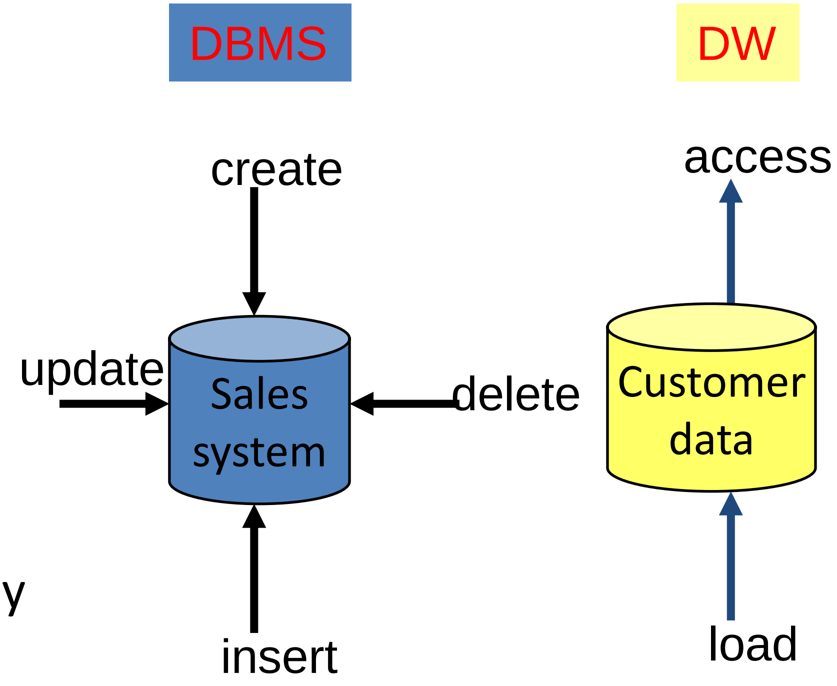
Data Warehouse—Subject-Oriented

- 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.

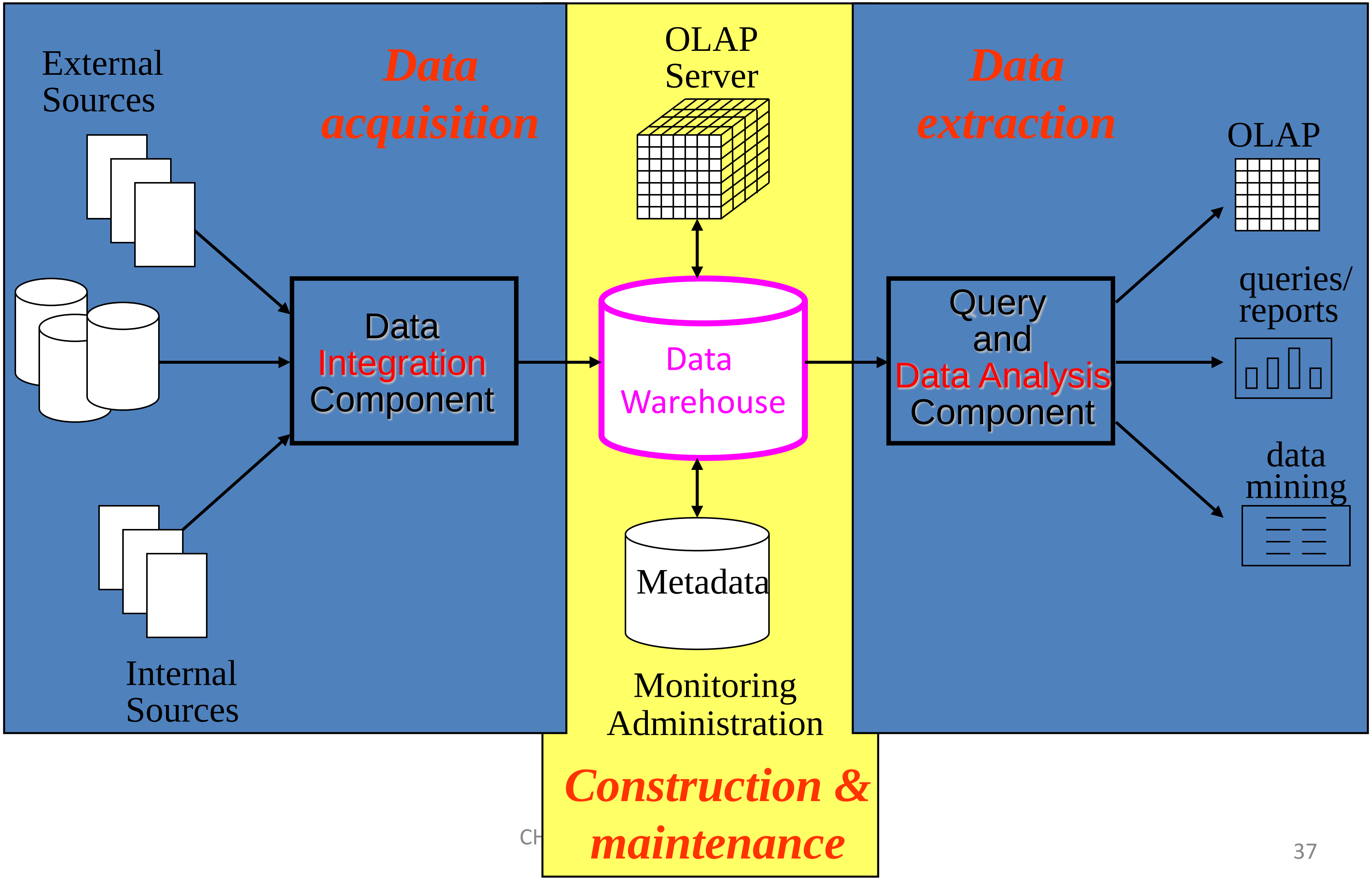


Data Warehouse—Non-Volatile

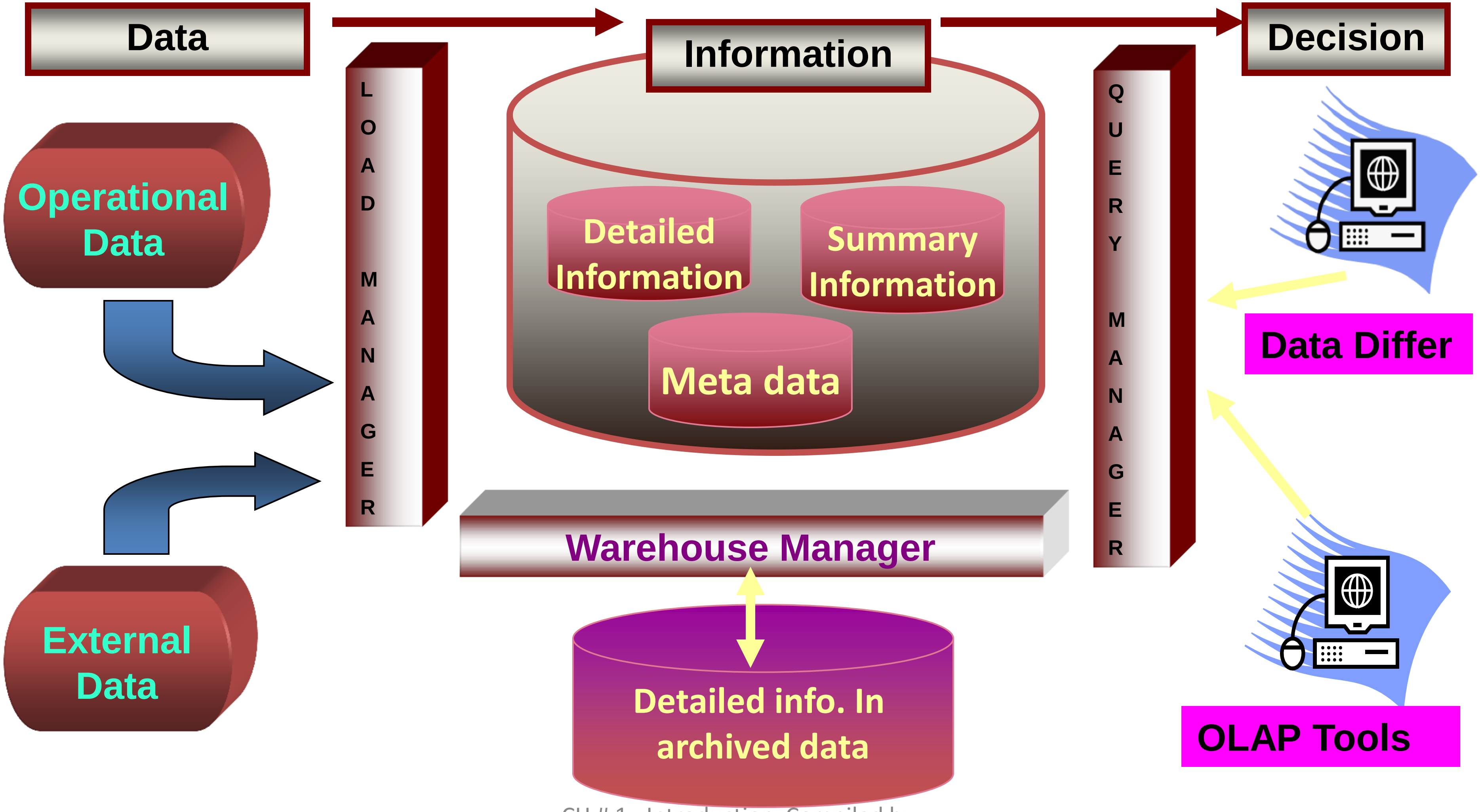
- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data.*



General Architecture



Data Warehouse Architecture



Data Warehouse…

Load Manager:

-perform all the operations necessary to support the **extract and load process**.

-It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse.

Warehouse Manager:

-Performs all the necessary operations to support the warehouse management process.

-It analyzes the data to perform consistency and referential checks.

-Update all existing aggregations and back up data in the data warehouse.

Data Warehouse…

Warehouse Manager …

-It also transforms and merges the source data in the temporary data store into the published data warehouse with creating indexes and business views.

Query Manager:

-Performs all the operations necessary to support the query management process by directing queries to the appropriate tables.

-In some cases it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Data Warehouse …

Detailed Information: -Stores all the detailed information to determine the business requirements to analyze the level at which to retain detailed information in the data warehouse.

***Summary Information*: -**Stores all the predefined aggregations generated by the warehouse manager.

-It is a transient area which will change on an ongoing basis in order to respond to changing query profiles.

-It is essentially a replication to detailed information.

Meta Data: -Meta data is data about data which describes how information is structured within a data warehouse.

-It maps data stores to common view of information with the data warehouse.

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation(aggregation).
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: Decision Support requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats

Comparison of OLTP and Data Warehousing

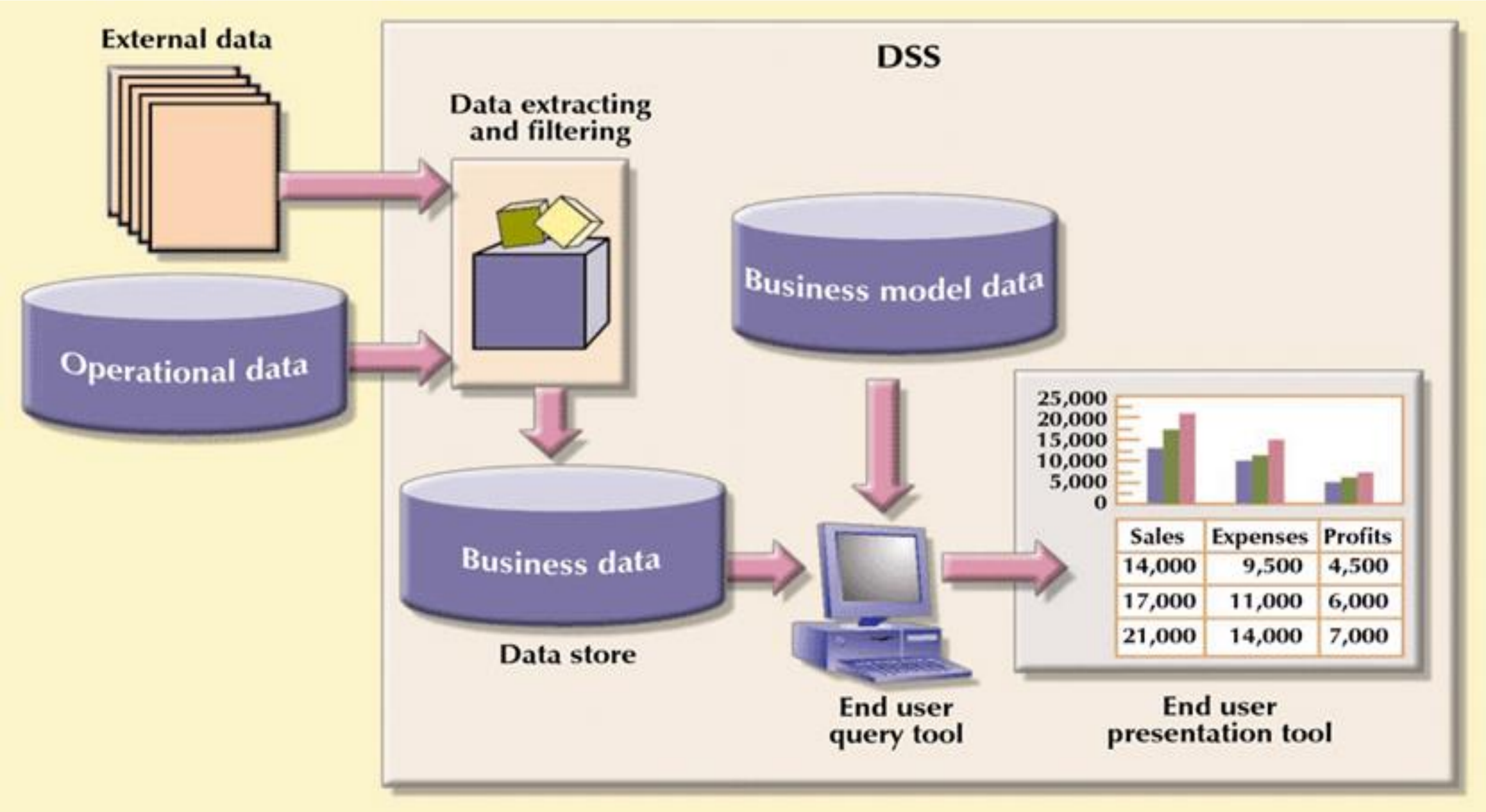
OLTP systems	Data warehousing systems
Holds current data	Holds historic data
Stores detailed data	Stores detailed, lightly, and summarized data
Data is dynamic	Data is largely static
Repetitive processing	Ad hoc, unstructured, and heuristic processing
High level of transaction throughput	Medium to low transaction throughput
Predictable pattern of usage	Unpredictable pattern of usage
Transaction driven	Analysis driven
Application oriented	Subject oriented
Supports day-to-day decisions	Supports strategic decisions
Serves large number of clerical / operational users	Serves relatively lower number of managerial users

Decision Support Systems

- Methodology (or series of methodologies) designed to extract information from data and to use such information as a basis for decision making

- Decision support system (DSS):
  - Arrangement of computerized tools used to assist managerial decision making within a business
  - Usually requires extensive data "massaging" to produce information
  - Used at all levels within an organization
  - Often tailored to focus on specific business areas
  - Provides ad hoc query tools to retrieve data and to display data in different formats

Main Components of a Decision Support System (DSS)



Sales	Expenses	Profits
14,000	9,500	4,500
17,000	11,000	6,000
21,000	14,000	7,000

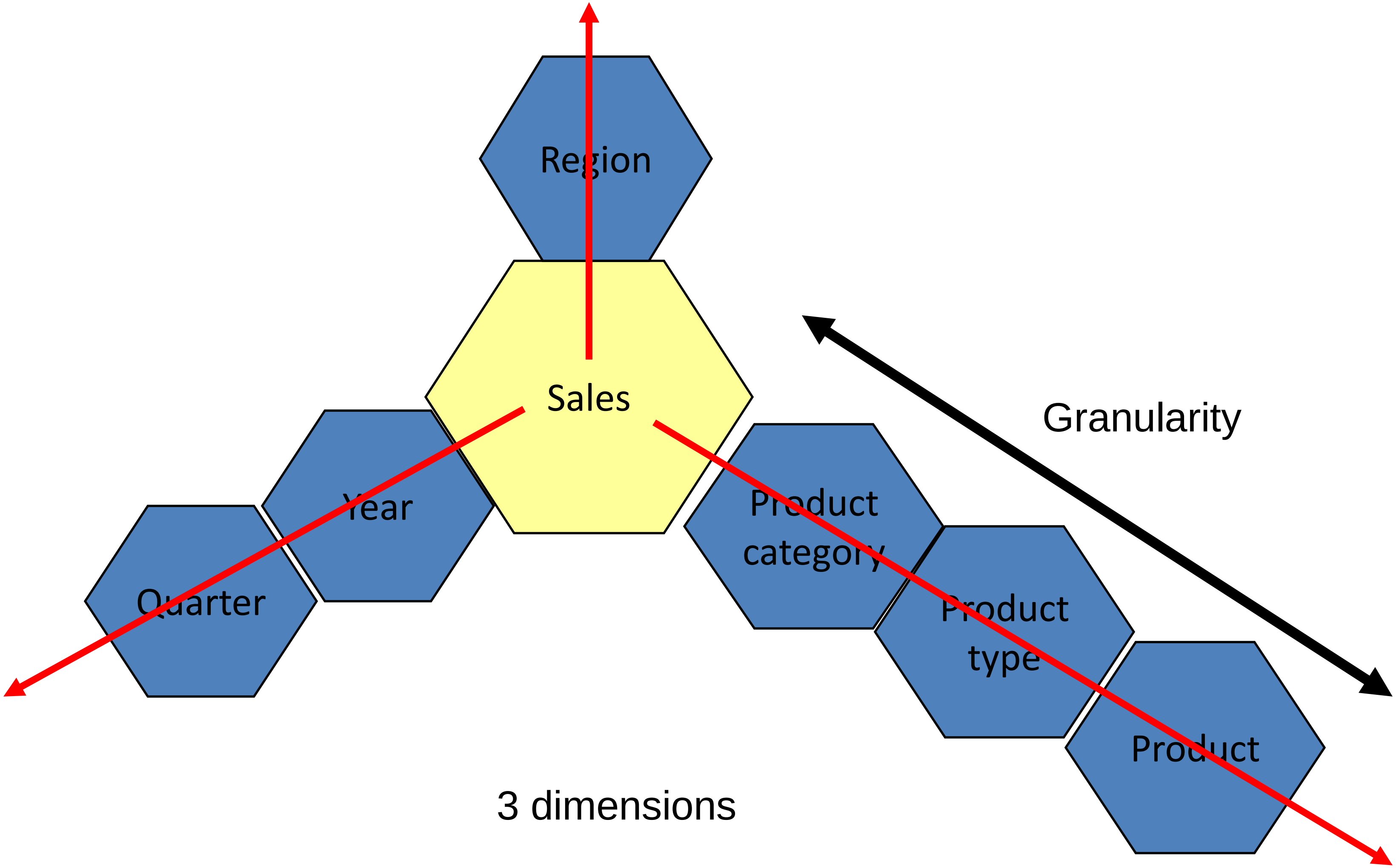
OLAP

WHAT IS OLAP?

DEFINITION :

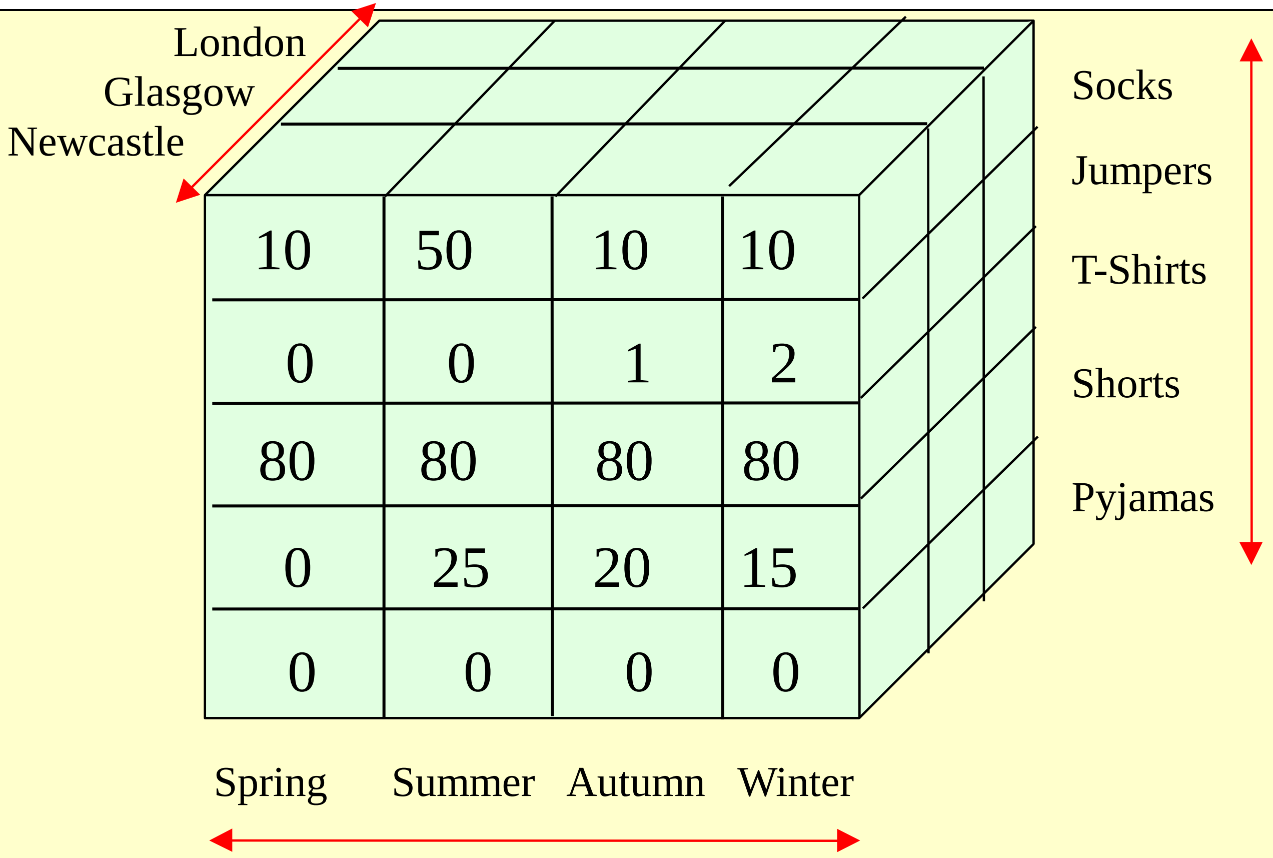
‘OLAP applications and tools are those that are designed to ask ad hoc, complex queries of large multidimensional collections of data. It is for this reason that OLAP is often mentioned in the context of Data Warehouses’.

The Multidimensional Idea



OLAP

MULTDIMENSIONAL DATA MODEL

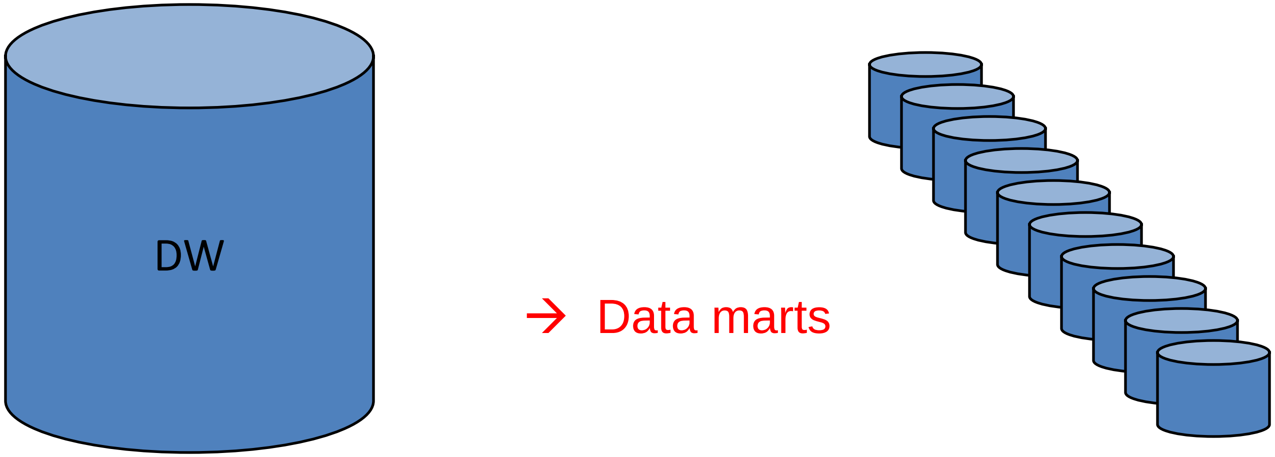


	Spring	Summer	Autumn	Winter
Socks	10	50	10	10
Jumpers	0	0	1	2
T-Shirts	80	80	80	80
Shorts	0	25	20	15
Pyjamas	0	0	0	0

Example: Three dimensions – Product, Sales-Area, and Season

Partitioning

- To improve performances & flexibility without giving up on the details



DW

→ Data marts

- By date, business type, geography, …

Data Mart

- Data Mart is a subset of the information content of a data warehouse that is stored in its own database.

- Data mart may or may not be sourced from an enterprise data warehouse i.e. it could have been directly populated from source data.

- Data mart can improve query performance simply by reducing the volume of data that needs to be scanned to satisfy the query.

- Data marts are created along functional level to reduce the likelihood of queries requiring data outside the mart.

- Data marts may help in multiple queries or tools to access data by creating their own internal database structures.

- Eg: Departmental Store, Banking System.