DOUGLAS COLLEGE

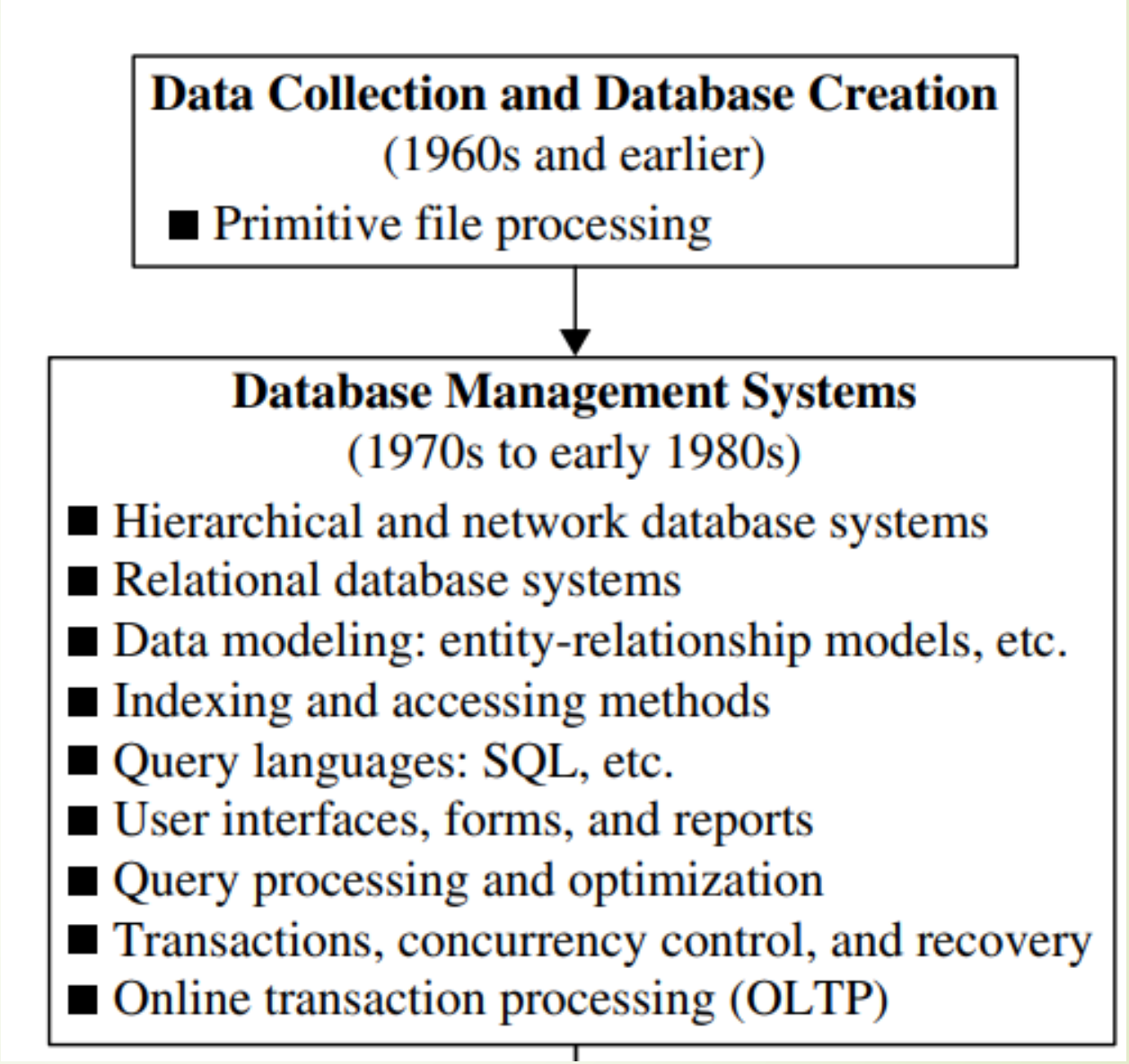
# FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

## CSIS 3290

## WHY DATA MINING

FATEMEH AHMADI (PH.D.)

# Data Mining as the Evolution of Information Technology



# Data Mining as the Evolution of Information Technology

**Advanced Database Systems**
(mid-1980s to present)

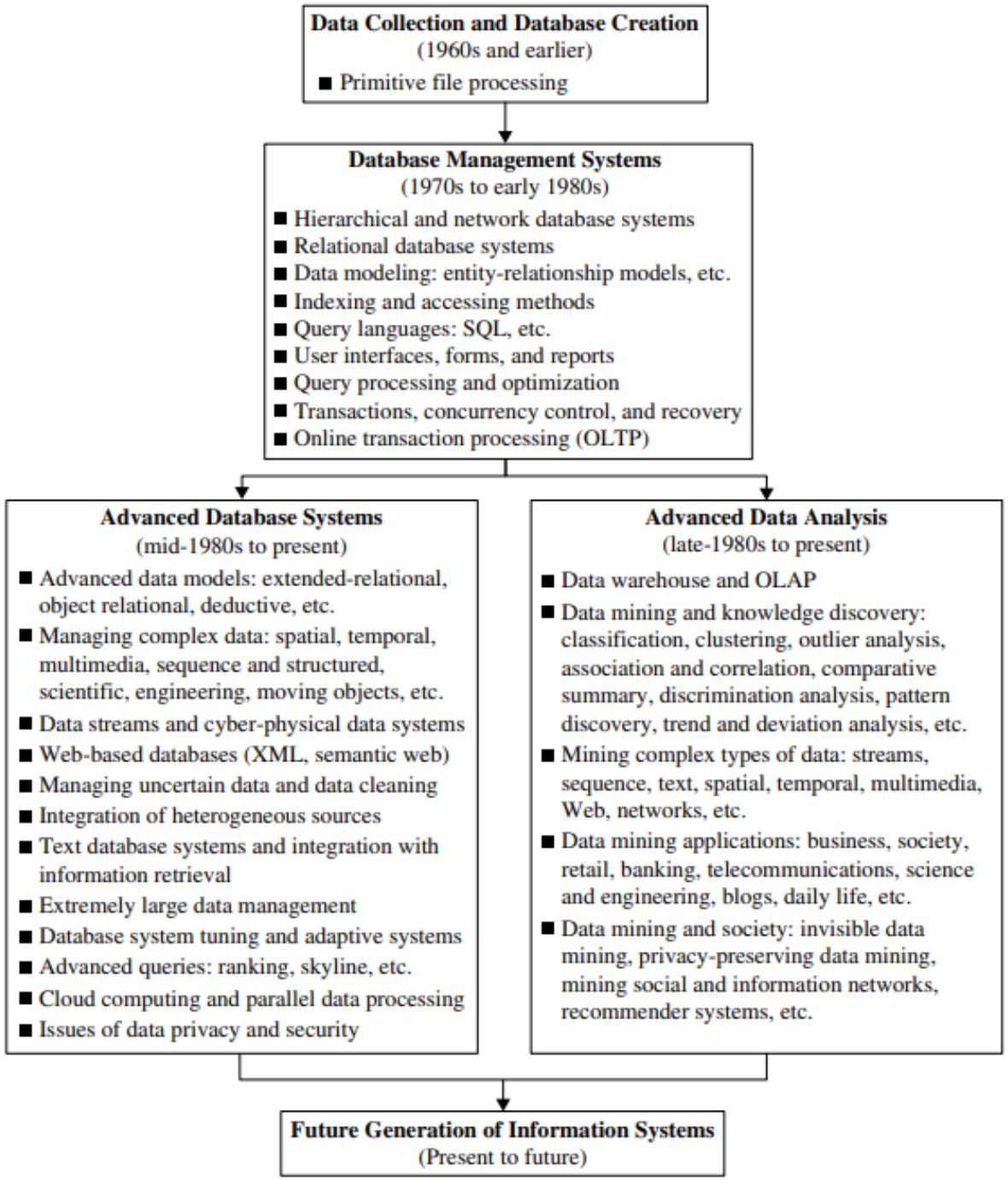
- Advanced data models: extended-relational, object relational, deductive, etc.
- Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
- Data streams and cyber-physical data systems
- Web-based databases (XML, semantic web)
- Managing uncertain data and data cleaning
- Integration of heterogeneous sources
- Text database systems and integration with information retrieval
- Extremely large data management
- Database system tuning and adaptive systems
- Advanced queries: ranking, skyline, etc.
- Cloud computing and parallel data processing
- Issues of data privacy and security

# Data Mining as the Evolution of Information Technology

**Advanced Data Analysis**
(late-1980s to present)

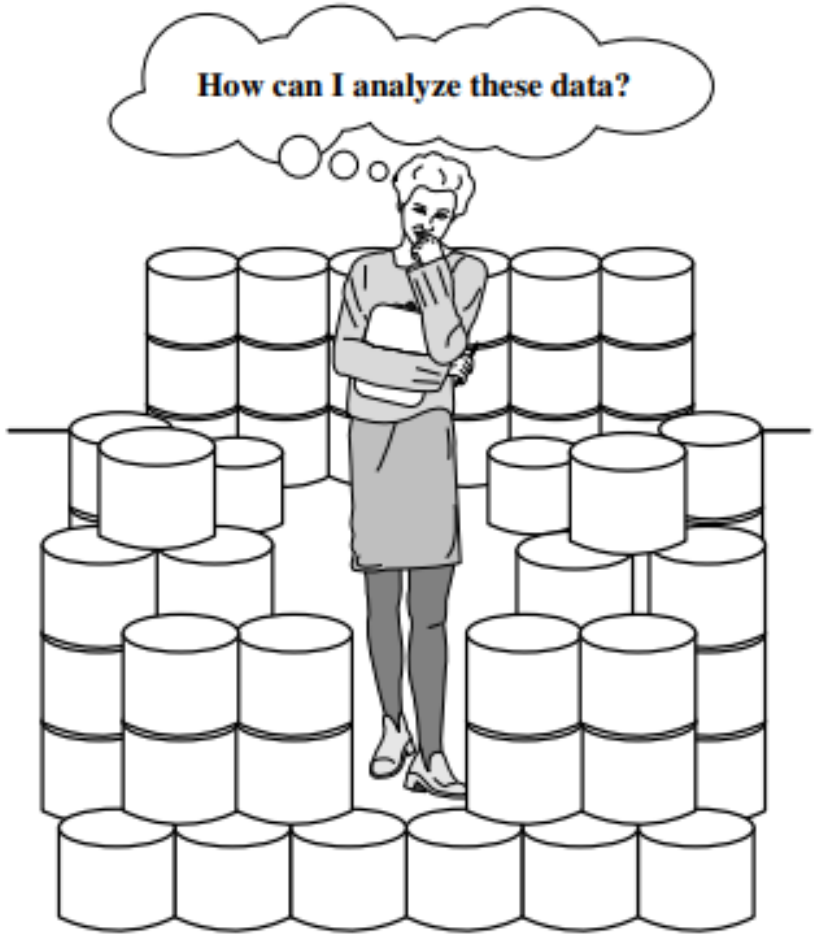
- Data warehouse and OLAP
- Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
- Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
- Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
- Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.

# Data Mining as the Evolution of Information Technology



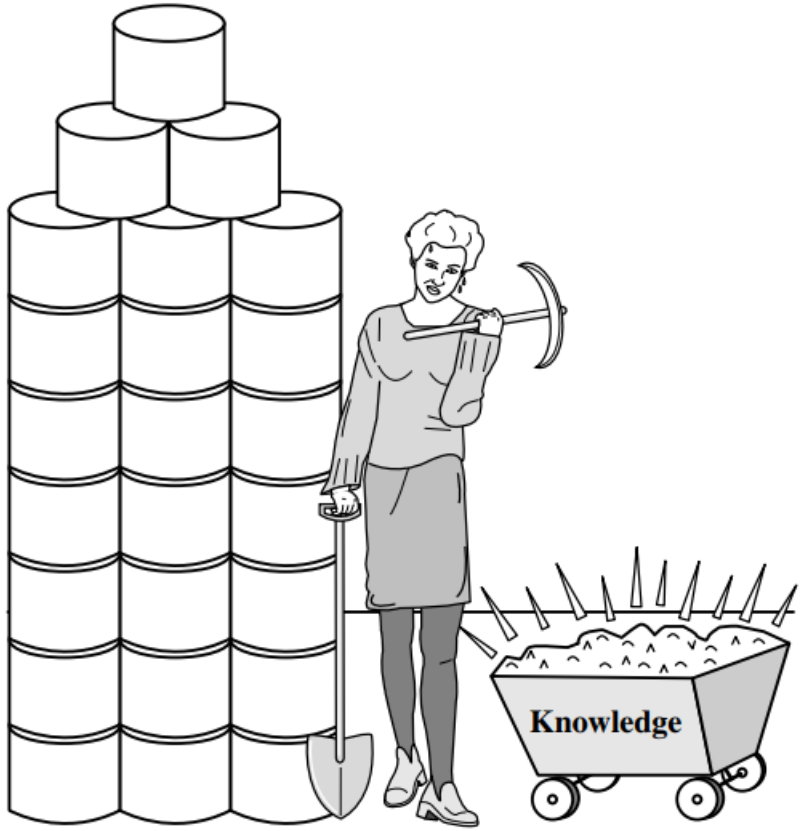
# Data Mining as the Evolution of Information Technology

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a **data rich but information poor situation**. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools.



# What Is Data Mining?

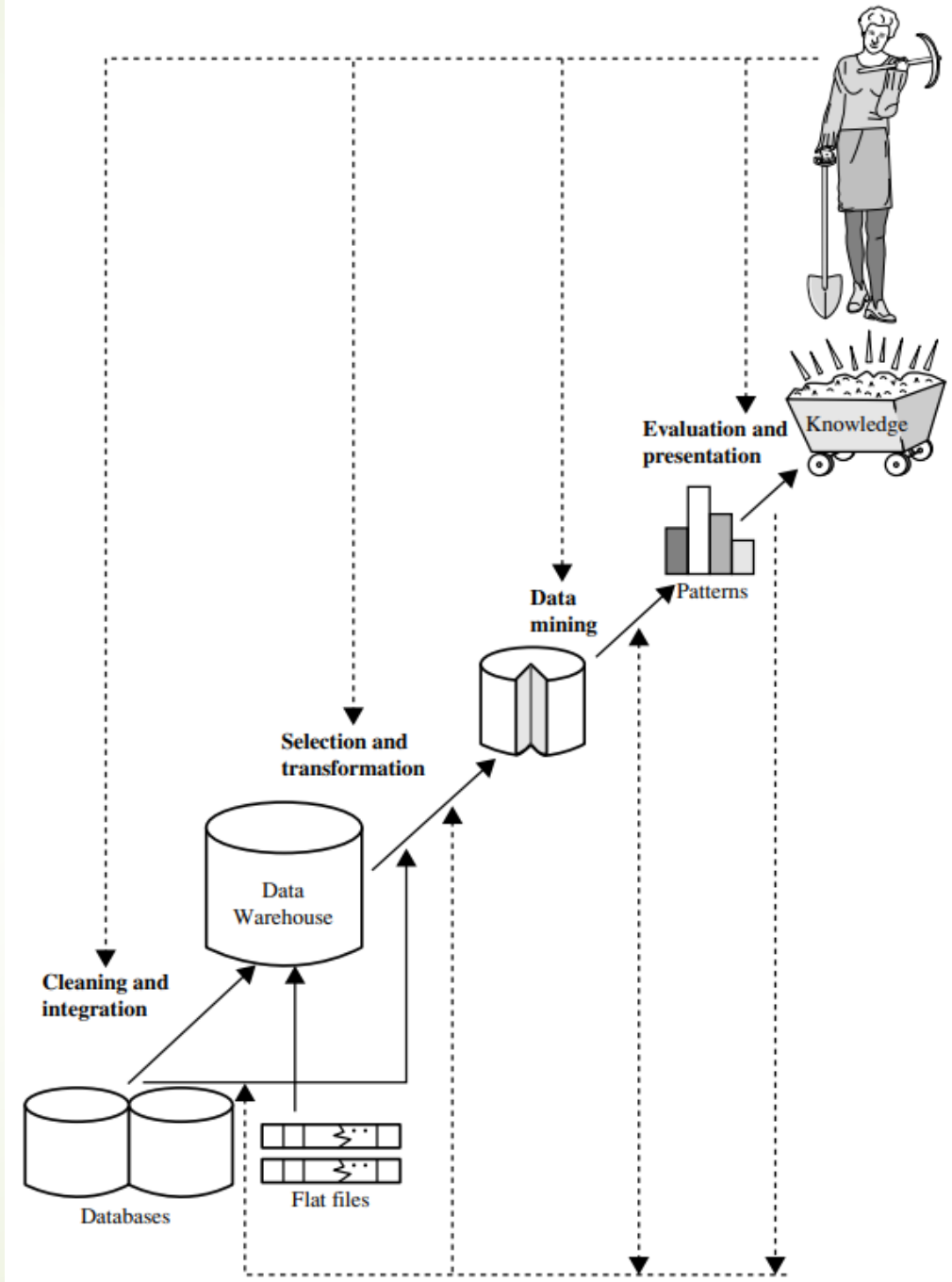
- Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.
- In addition, many other terms have a similar meaning to data mining—for example, **knowledge mining from data**, **knowledge extraction, data/pattern analysis, data archaeology**, and **data dredging**.
- Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**, while others view data mining as merely an essential step in the process of knowledge discovery



# What Is Data Mining?

The knowledge discovery process:

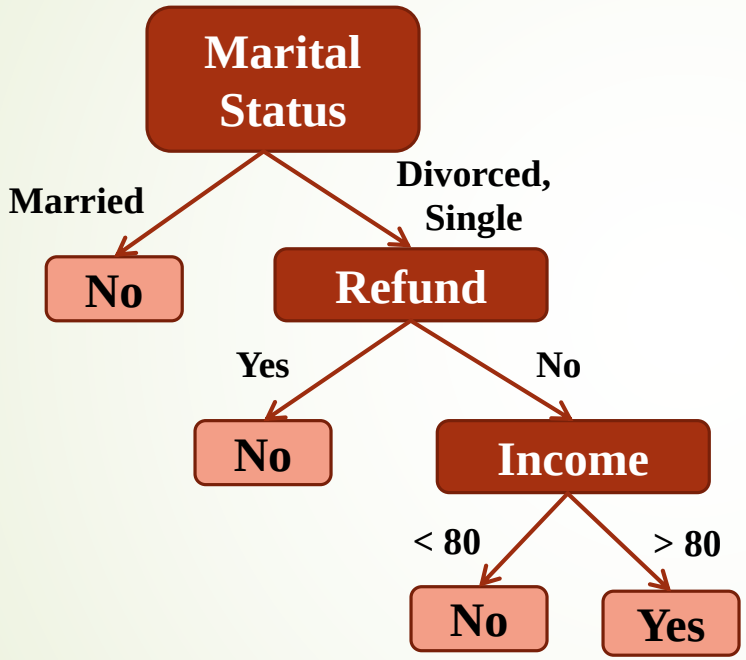
1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures—see Section 1.4.6)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users

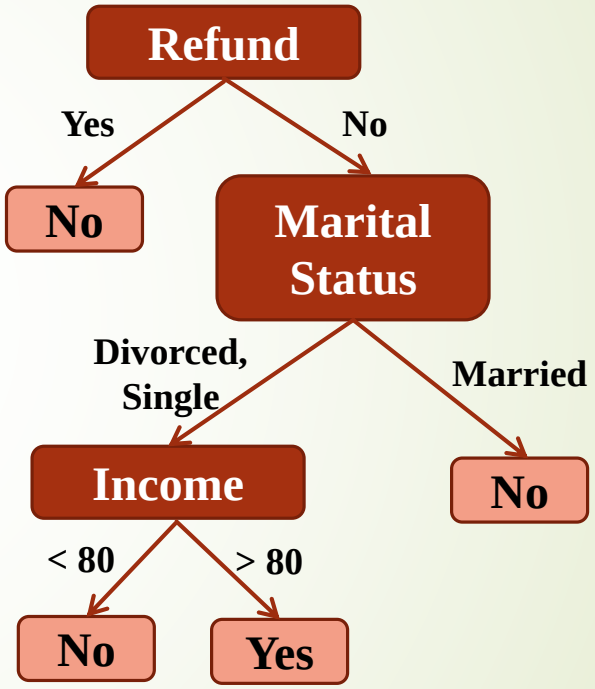


# Example: Decision Trees

| T_ID | Refund | Marital Status | Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125 | No |
| 2 | No | Married | 100 | No |
| 3 | No | Single | 70 | No |
| 4 | Yes | Married | 120 | No |
| 5 | No | Divorced | 95 | Yes |
| 6 | No | Married | 60 | No |
| 7 | Yes | Divorced | 220 | No |
| 8 | No | Single | 85 | Yes |
| 9 | No | Married | 75 | No |
| 10 | No | Single | 90 | No |

# Example: Decision Trees




Decision Tree No.1

Decision Tree No.2

There could be more than one decision tree per a table

# Example: Market Basket Analysis

**{X→Y}: {Butter, Bread}→ Milk**

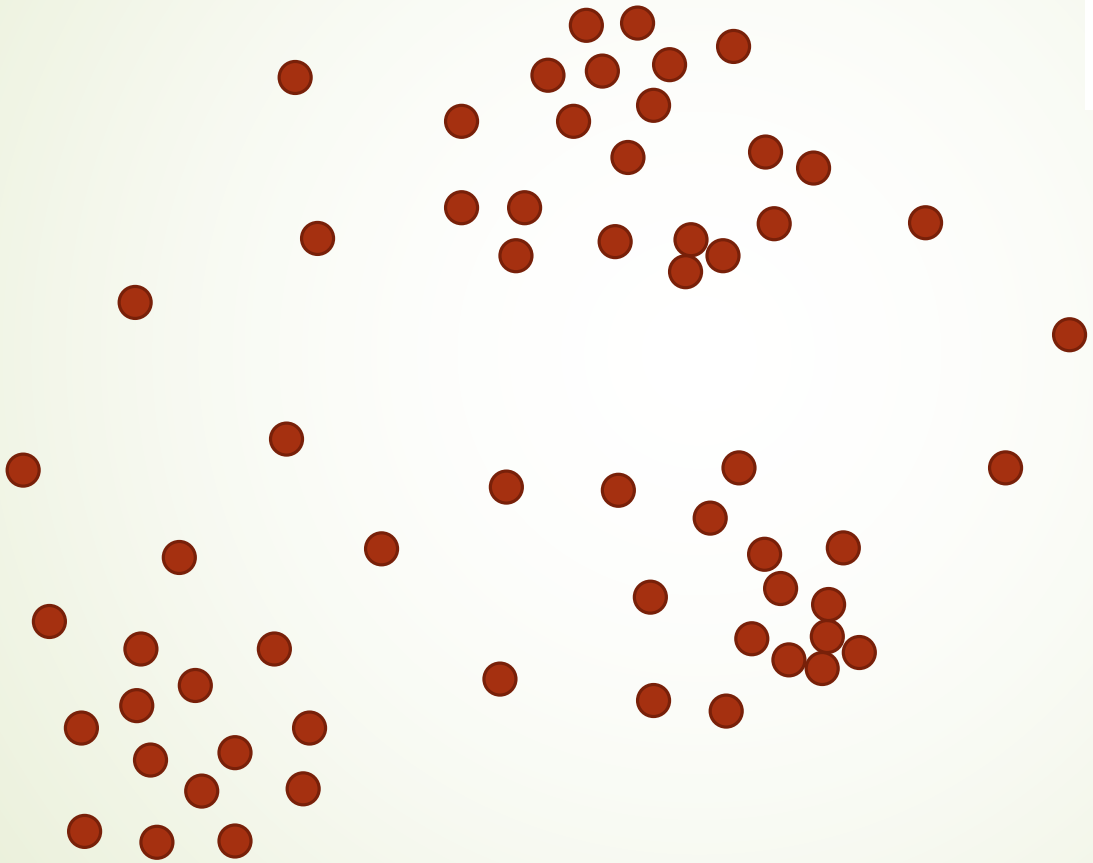
| Transaction | Milk | Bread | Butter |
|---|---|---|---|
| **1** | 1 | 1 | 0 |
| **2** | 0 | 0 | 1 |
| **3** | 0 | 0 | 0 |
| **4** | 1 | 1 | 1 |
| **5** | 0 | 1 | 0 |

**Support: Supp(X) = 1/5= 0.2 ~ 20%**

**(**Since it occurs in 20% of all transactions**)**

**Confidence : Conf(X)= Supp(X U Y)/ Supp(X) = 0.2/0.2=1 ~ 100% (**100% of the times a customer buys butter and bread, milk is bought as well**)**

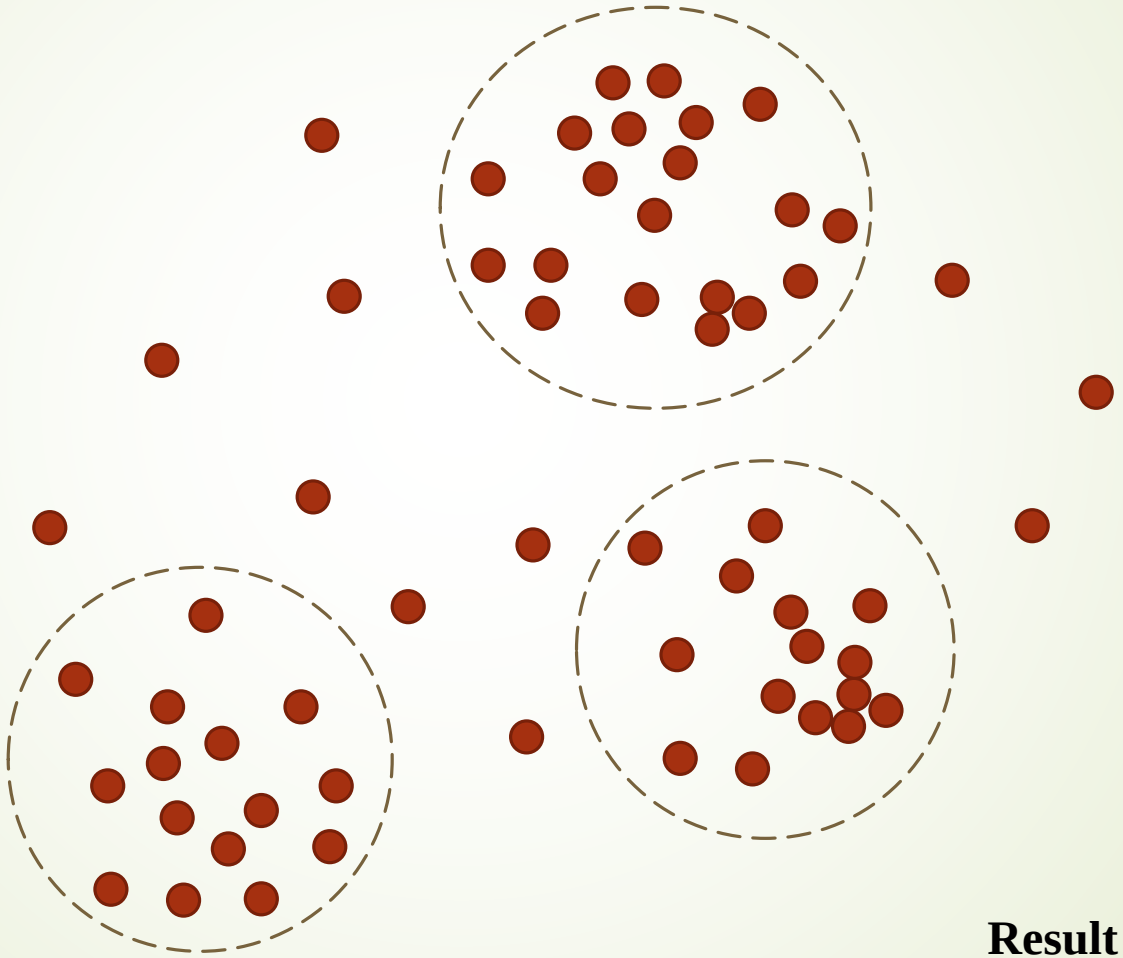
# Example: Clustering



**Famous Clustering Algorithms**

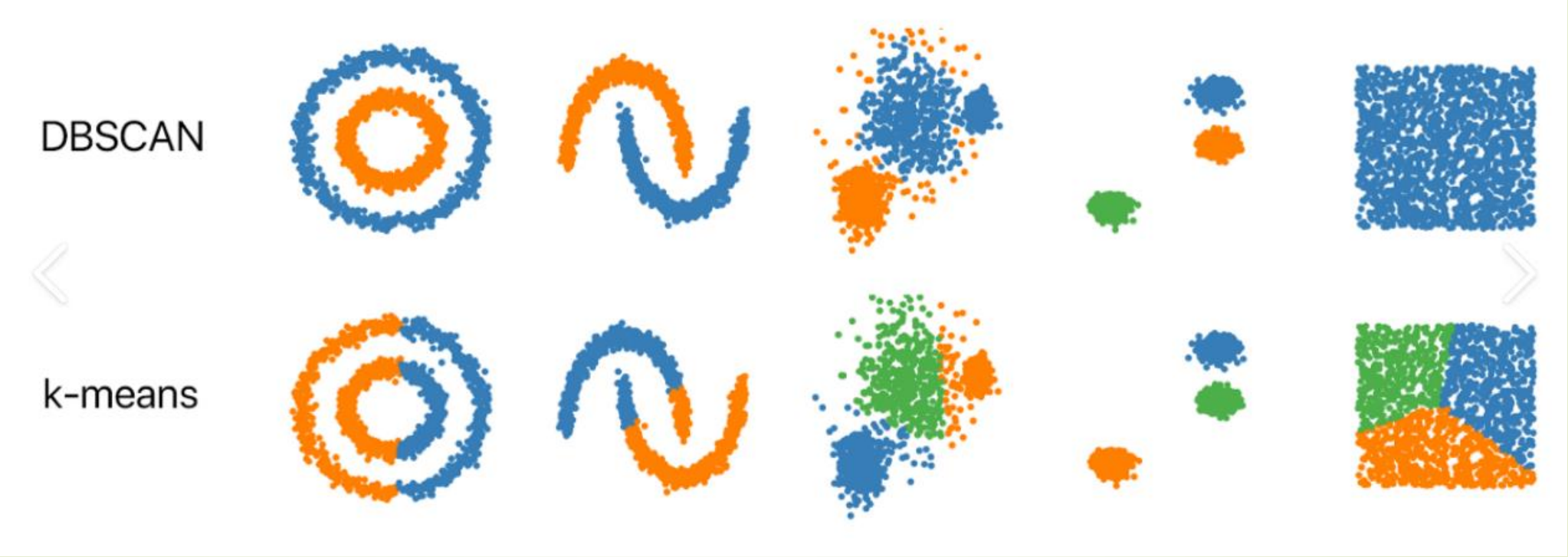
- ✓ **K-Means**
- ✓ **DBSCAN**
- ✓ **Hierarchical Clustering**
- ✓ **…**

# Clustering



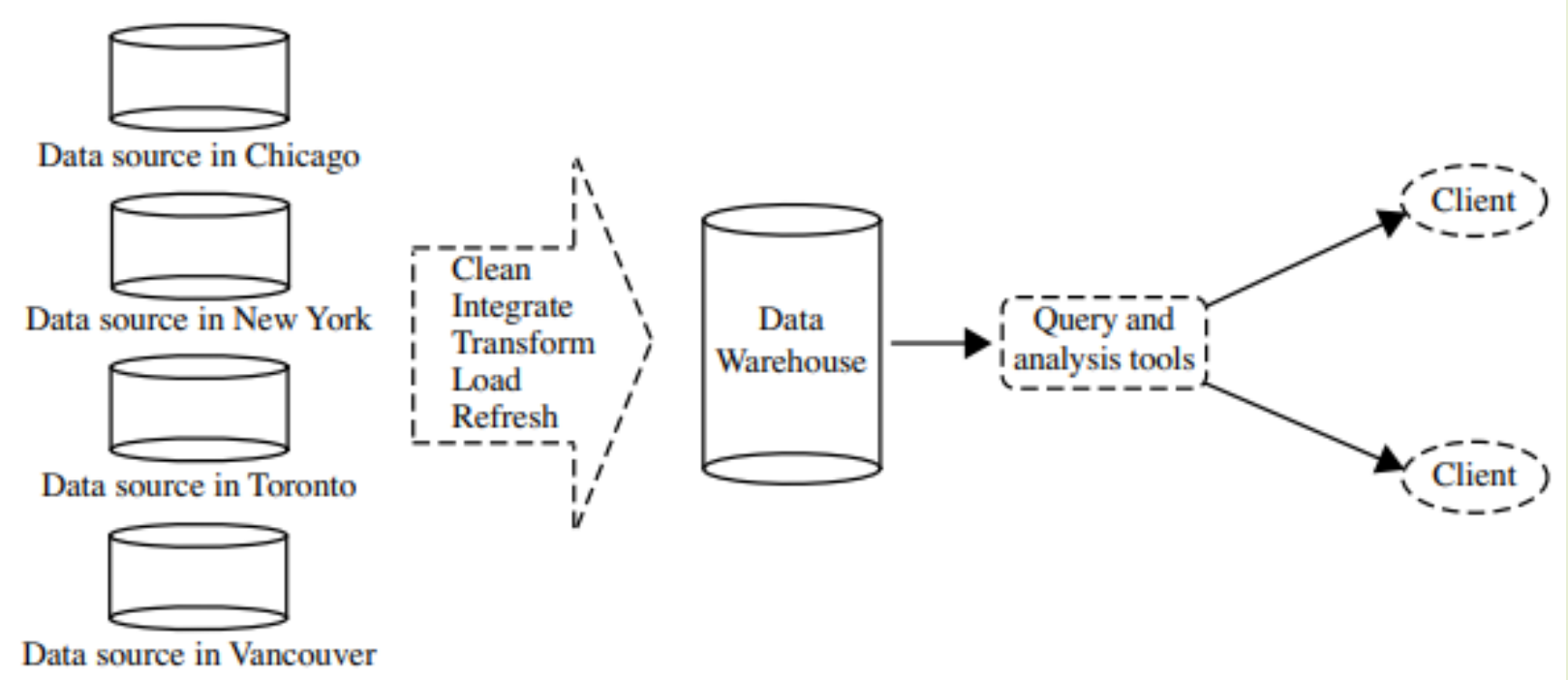
Result of K-Means Clustering

# Clustering




https://www.bing.com/images/search?view=detailV2&ccid=3LdJGuOj&id=8CC2488C0F8B17E22DD76E9D59F2ACDFDA158EEE&thid=OIP.3LdJGuOjoIHFG7m8W2NmXQHaCw&mediaurl=https%3a%2f%2fmiro.medium.com%2fmax%2f1200%2f1*KqWII7sFp1JL0EXwJGpqFw.png&cdnurl=https%3a%2f%2fth.bing.com%2fth%2fid%2fR.dcb7491ae3a3a251c51bb9bc5b63665d%3frik%3d7o4V2f%252bs8lmdbg%26pid%3dImgRaw%26r%3d0&exph=448&expw=1200&q=dbscan&simid=607992534376195191&FORM=IRPRST&ck=D05FFCD01B189C12EA922B280124E7E8&selectedIndex=3&ajaxhist=0&ajaxserp=0

# Data Warehouses



# Data Cube

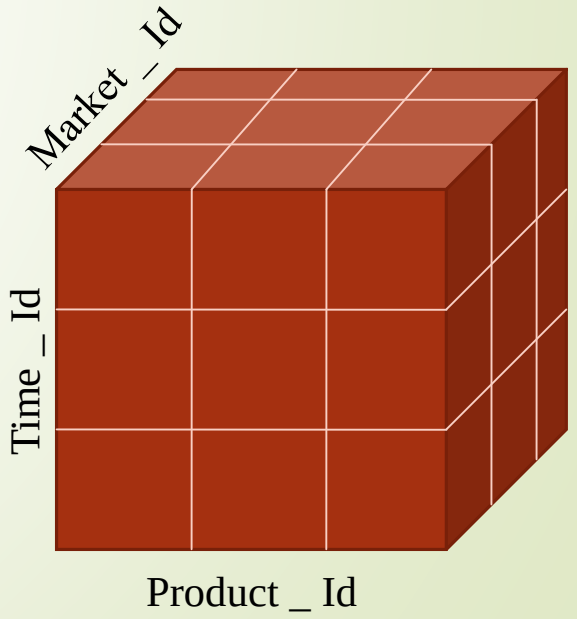
| Market _ Id | Product _ Id | Time _ Id | Sales _ Amt |
|---|---|---|---|
| M1 | P1 | Q1 | 1000 |
| M1 | P2 | Q1 | 2000 |
| M1 | P3 | Q4 | 1500 |
| M2 | P1 | Q3 | 500 |
| M2 | P2 | Q1 | 800 |
| M2 | P3 | Q4 | 0 |
| M3 | P1 | Q1 | 5000 |
| M3 | P2 | Q2 | 8000 |
| M3 | P3 | Q2 | 10 |

**A Fact Table for a Supermarket Application**

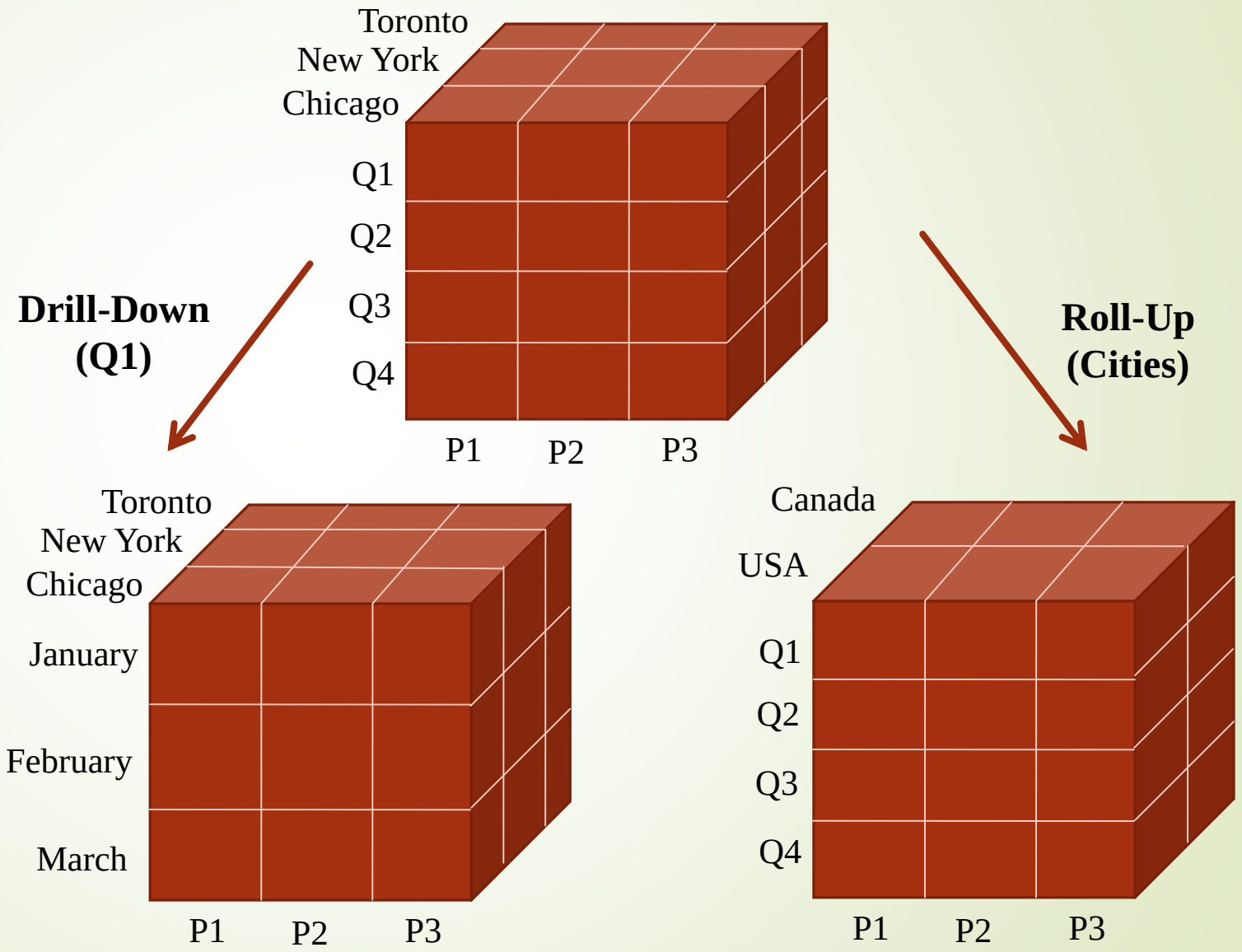
**Table Name: *Sales***

**Operations in Data Cubes:**

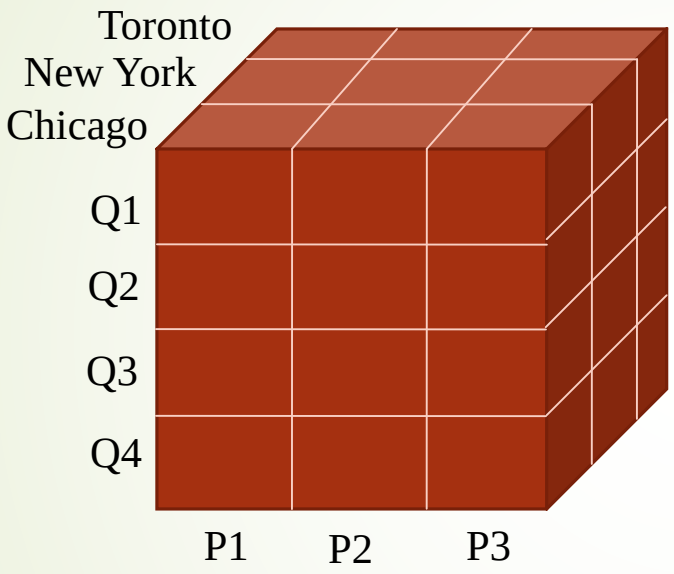
- ✓ **Drill down**
- ✓ **Roll Up**
- ✓ **Slice**
- ✓ **Dice**
- ✓ **…**



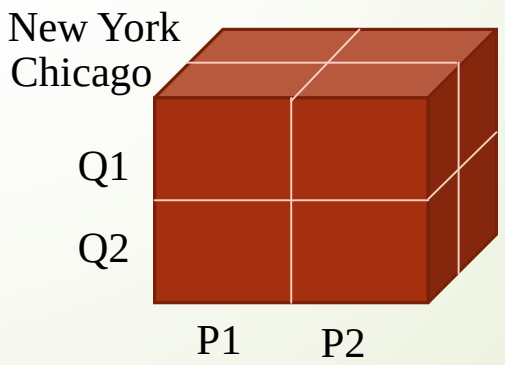
# Data Cube


Toronto
New York
Chicago
Q1
Q2
Q3
Q4
P1
P2
P3
Drill-Down
(Q1)
Roll-Up
(Cities)
Toronto
New York
Chicago
January
February
March
P1
P2
P3
Canada
USA
Q1
Q2
Q3
Q4
P1
P2
P3

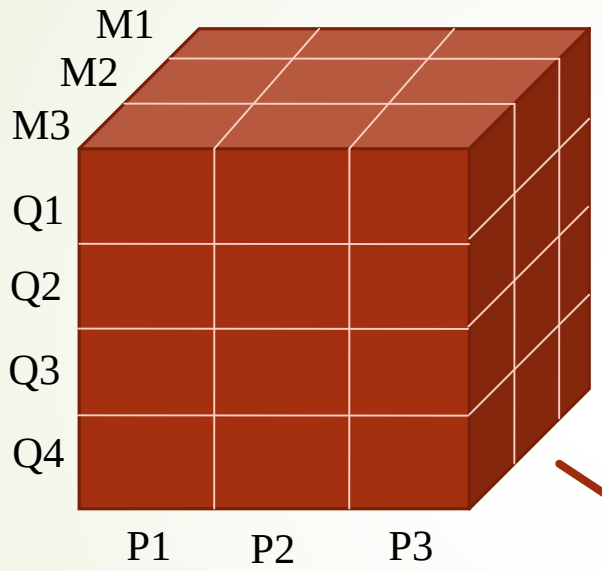
# OLAP Operations: Dice




**Dice**
**(Time = Q1 or Q2**
**Product= P1 or P2**
**Market = Chicago or New York)**



# OLAP Operations: Slice




**Slice**
**(Time Dimension = Q1)**

| | P1 | P2 | P3 |
|---|---|---|---|
| M1 | | | |
| M2 | | | |
| M3 | | | |

**Dimensional Tables for the Supermarket Application:** To describe additional information on each dimension

| Market _ Id | City | State | Region |
|---|---|---|---|
| M1 | New York | New York | East |
| M2 | Toronto | New Jersey | North |
| M3 | Oakland | California | West |

**Table Name: *Market***

**Table Name: *Product***

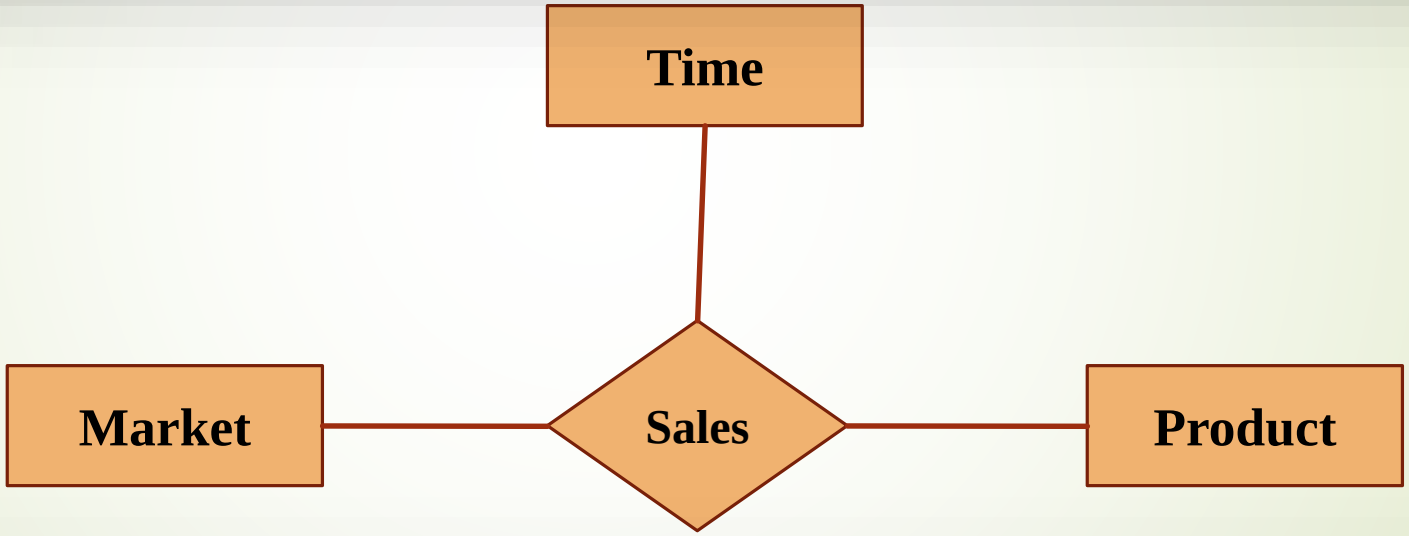
| Product _ Id | Name | Category | Price |
|---|---|---|---|
| P1 | Tissue | Soft Goods | 1.98 |
| P2 | Soft Drink | Drink | 2.96 |
| P3 | Cold cuts | Meat | 1.78 |

| Time _ Id | Week | Month | Quarter |
|---|---|---|---|
| T1 | Wk-1 | January | First |
| T2 | Wk-13 | April | Second |
| T3 | Wk-25 | July | Third |
| T3 | Wk-37 | October | Fourth |

**Table Name: *Time***

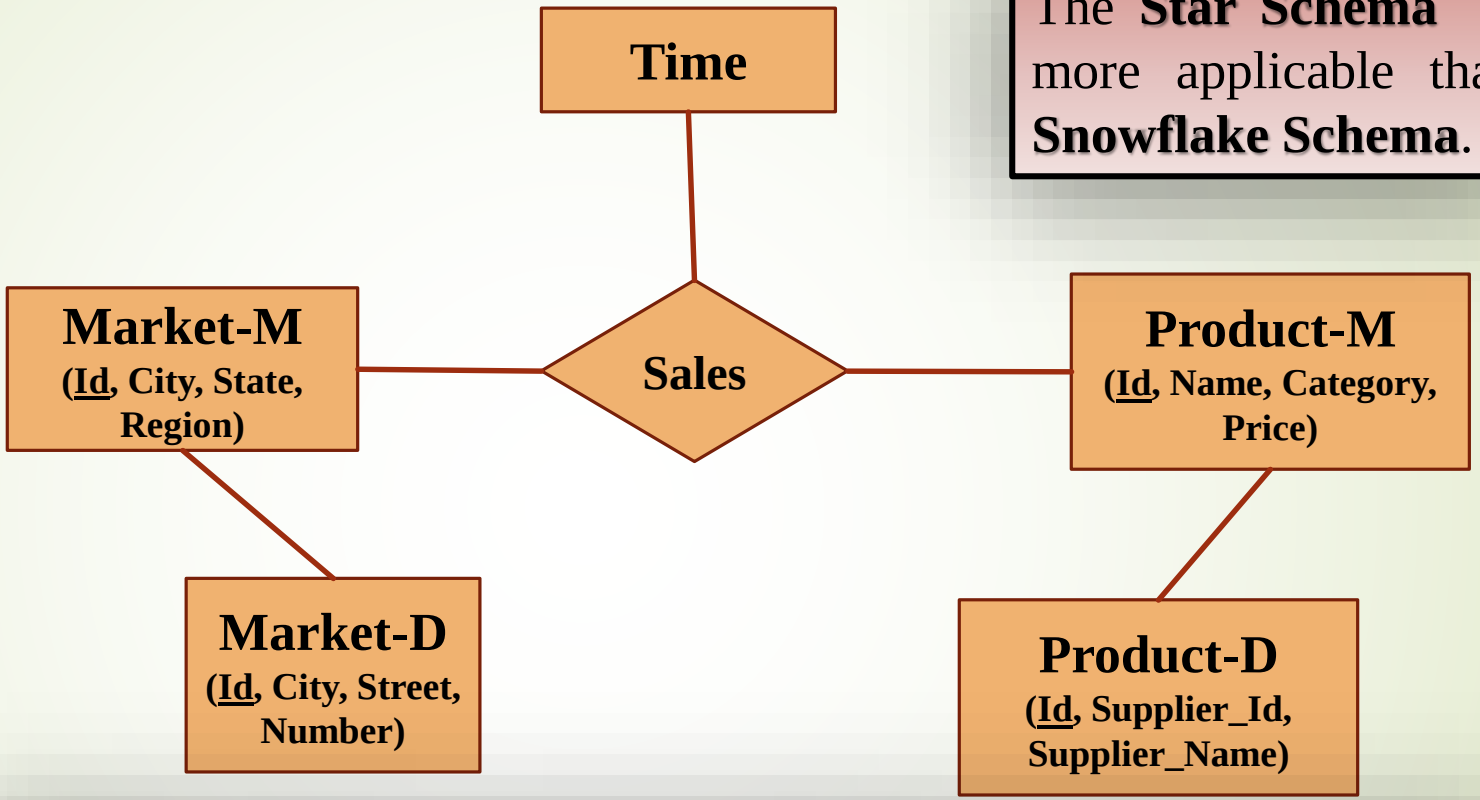
# Star Schema

In **Star Schema**, the fact table is at the center and the dimension tables radiate it.




If dimension tables are normalized and each becomes several tables, the star schema will be more complex called **Snowflake Schema**.

# Snowflake Schema

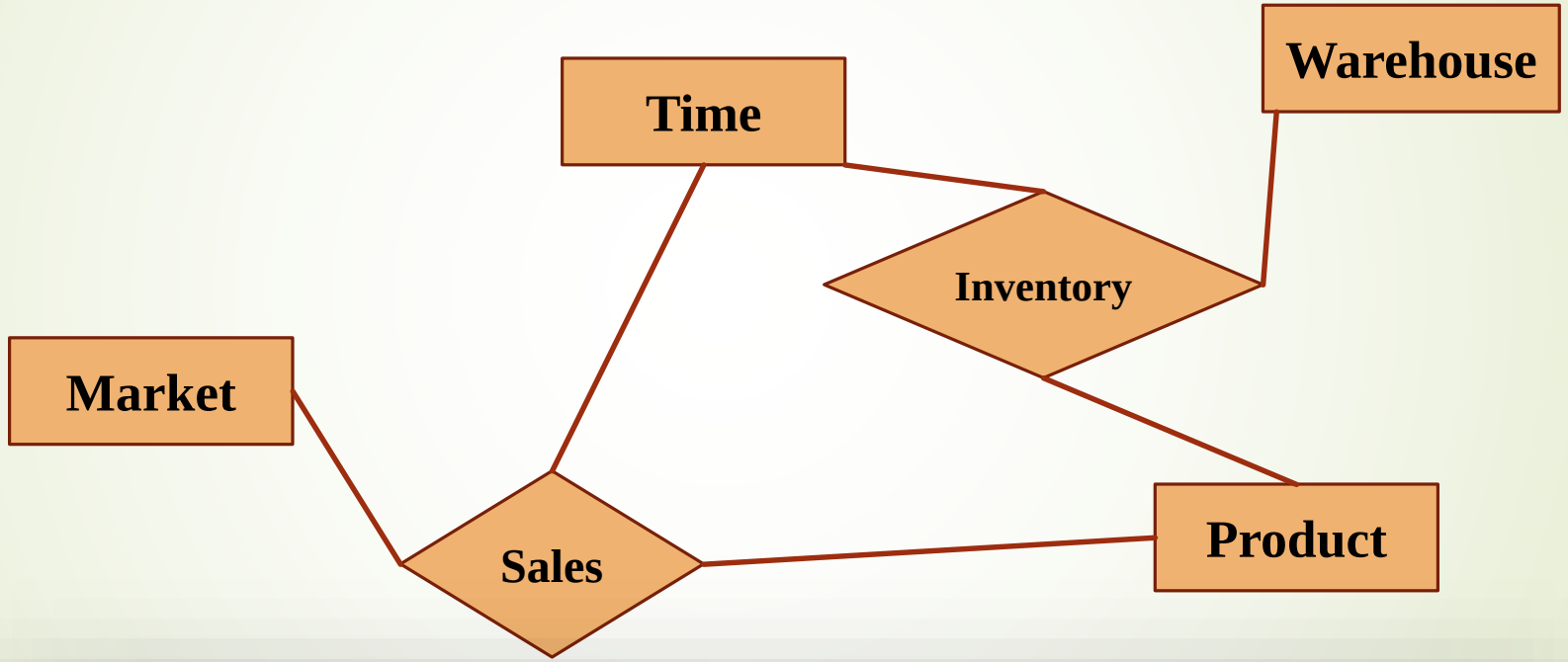



The **Star Schema** is more applicable than **Snowflake Schema**.

**Dimension tables are rarely normalized**:

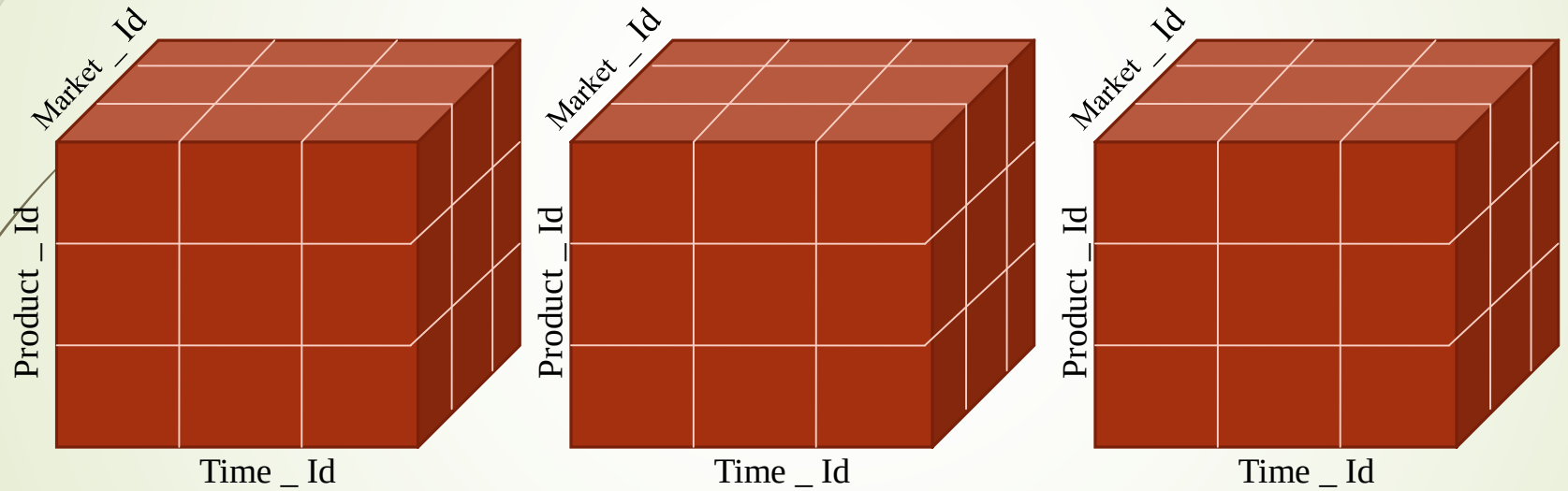
- They are updated so infrequently that update anomalies are not an issue.
- They are so small compared with the fact table that saved space due to the elimination of redundancy is negligible.

# Constellation Schema




In **Constellation Schema**, there are more than one fact table.

# The Forth Dimension

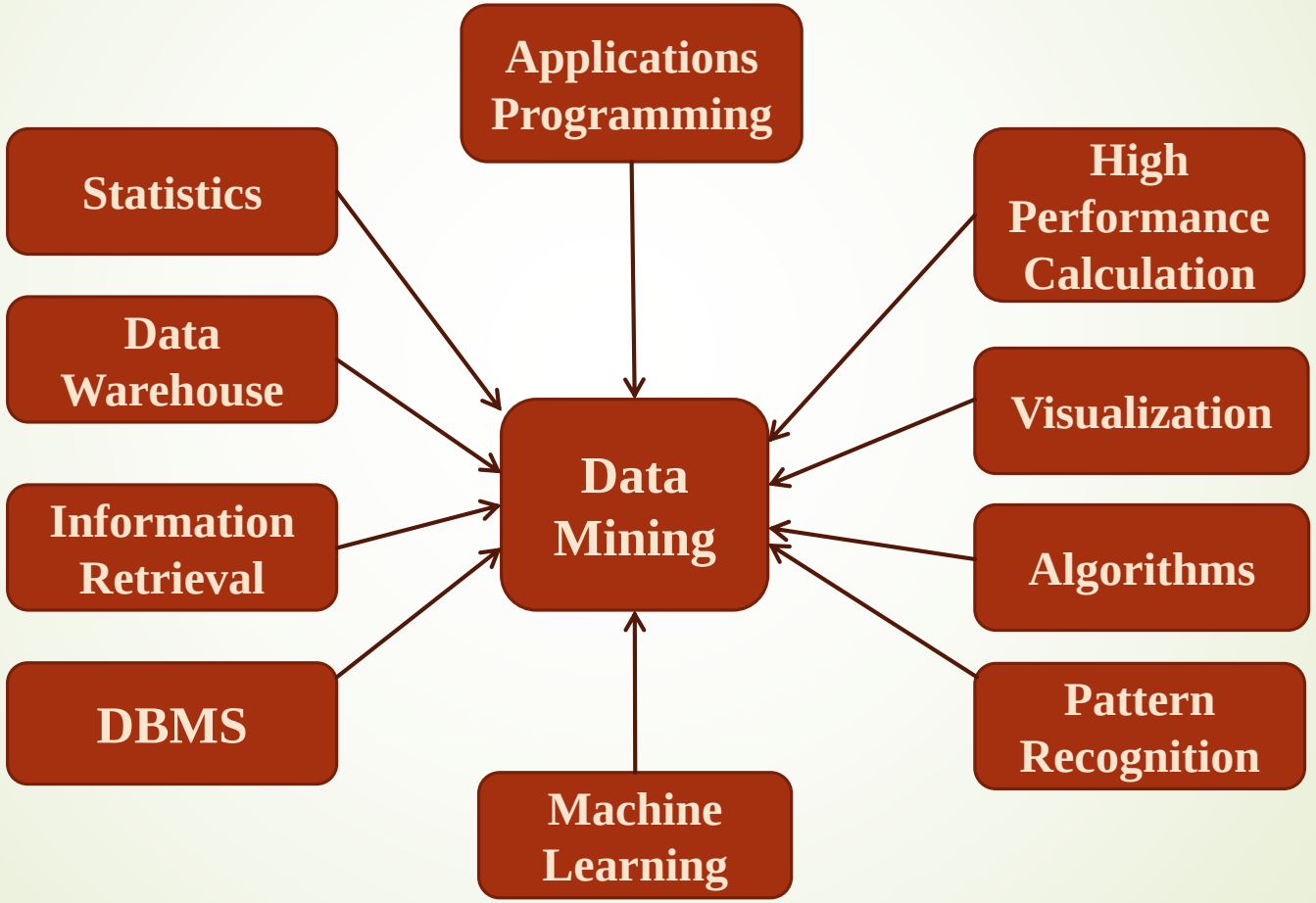



Supplier: S1

Supplier: S2

Supplier: S3

# Technologies Used in Data Mining



# Reference

Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 1.