Information Integration

Chapter 20

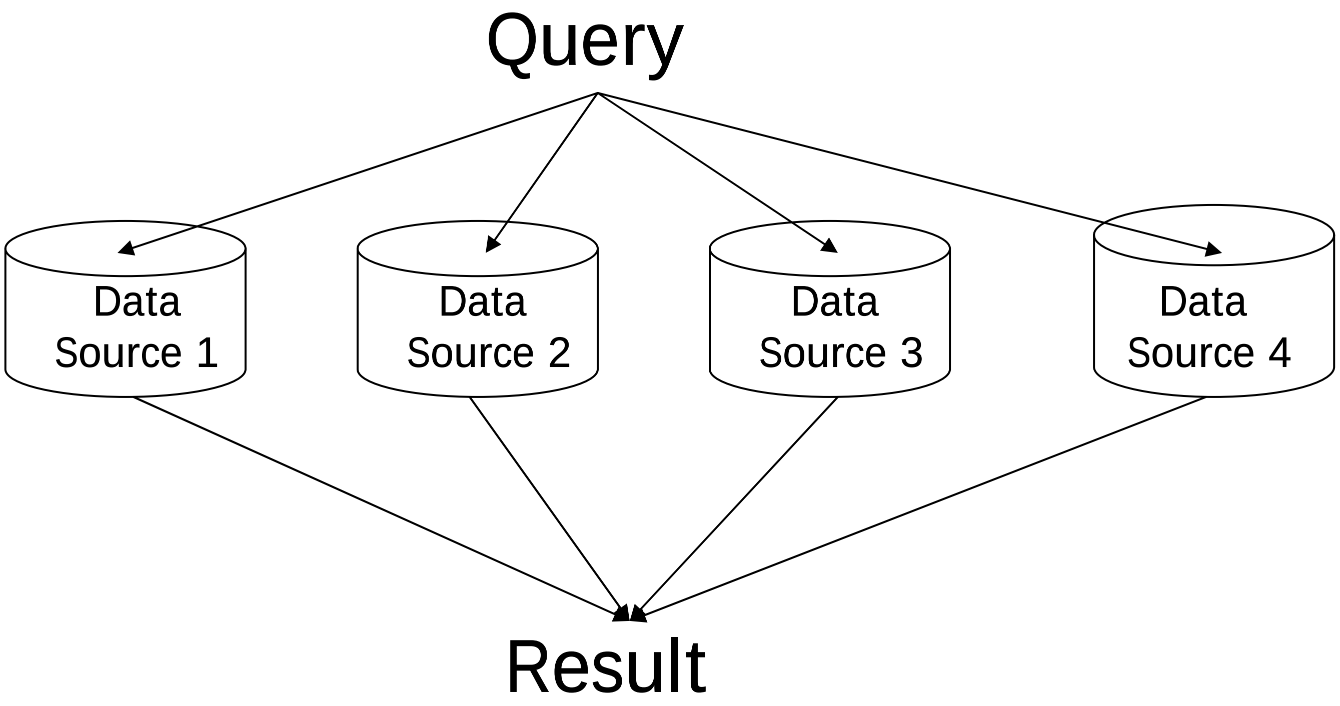
Objectives

- To have a shallow understanding of what a data warehouse and data mining are.

- Lecture outline

- Need for Information Integration (II)
- The Three most common approaches of II
- Problems of II
- OLAP
- Data Mining

- Need for Information Integration

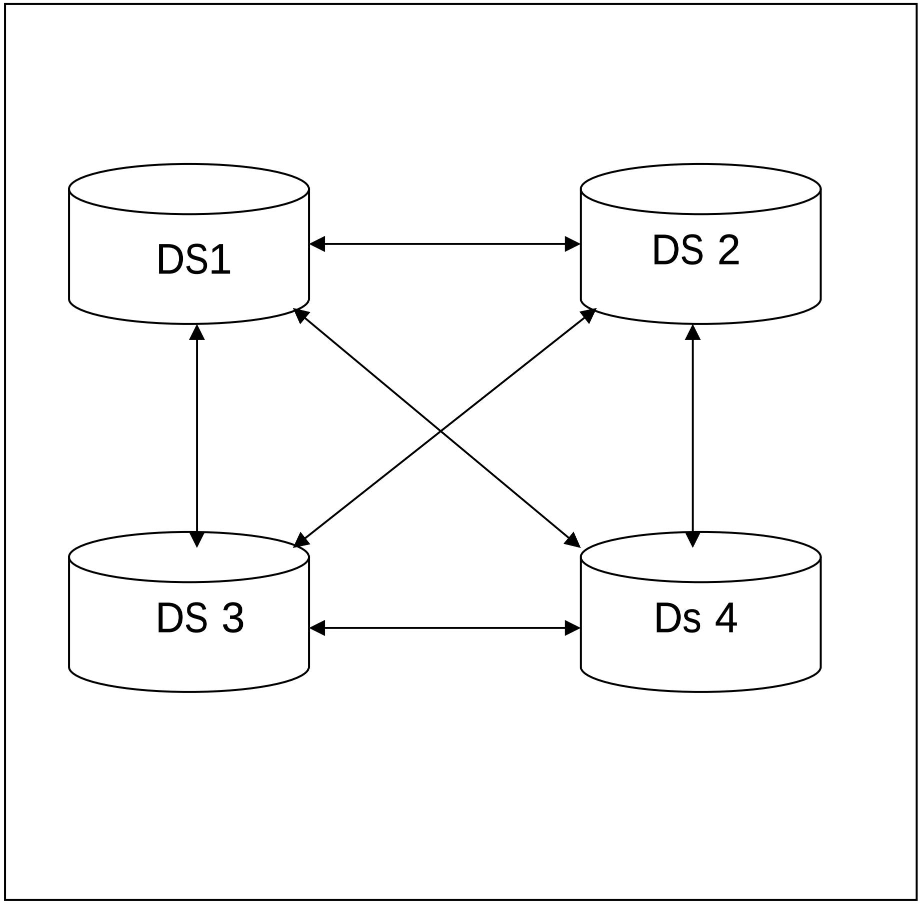


- The Three Most Common Approaches of II

- Federated DBs
- Mediation
- Warehousing

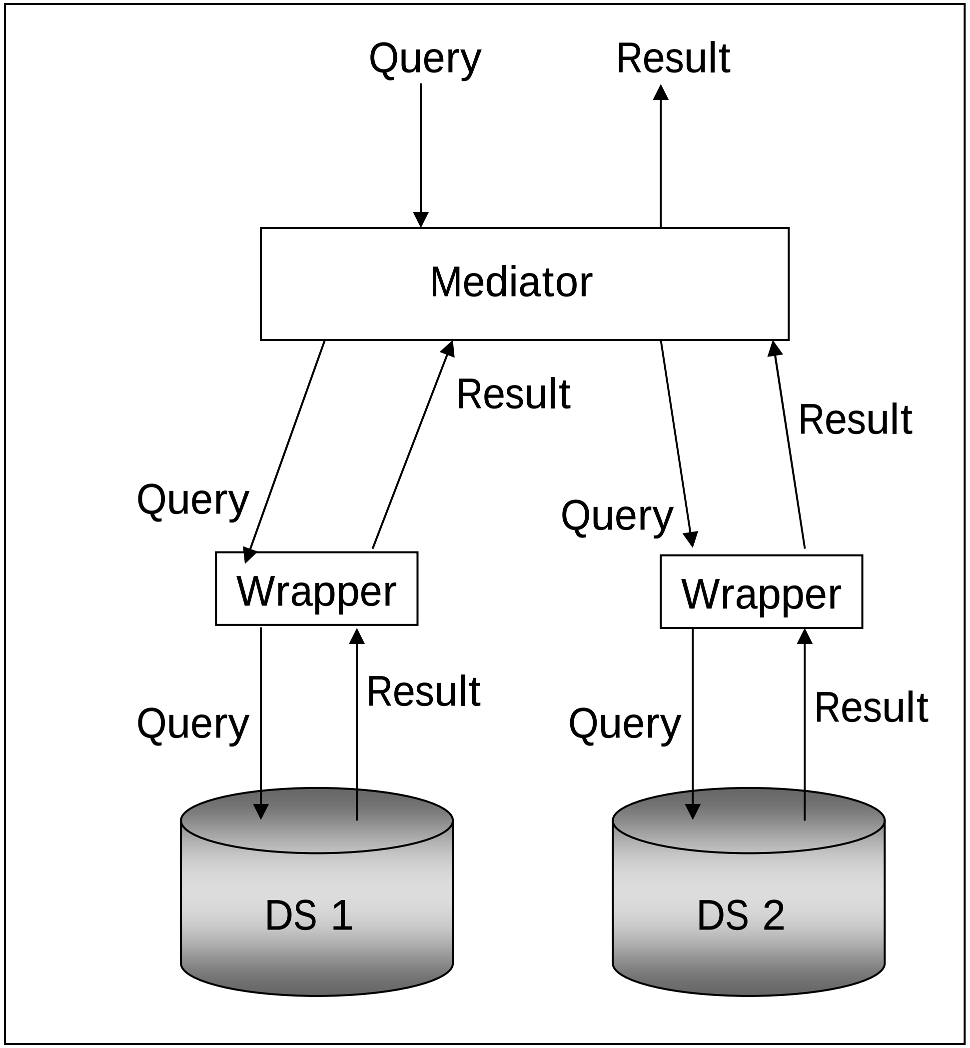
-- Federated Databases

- The sources are independent
- One source can call on others to supply information
- Advantage:
  - Easy to build.
- Disadvantage:
  - For n data sources, n(n-1) pieces of code is needed



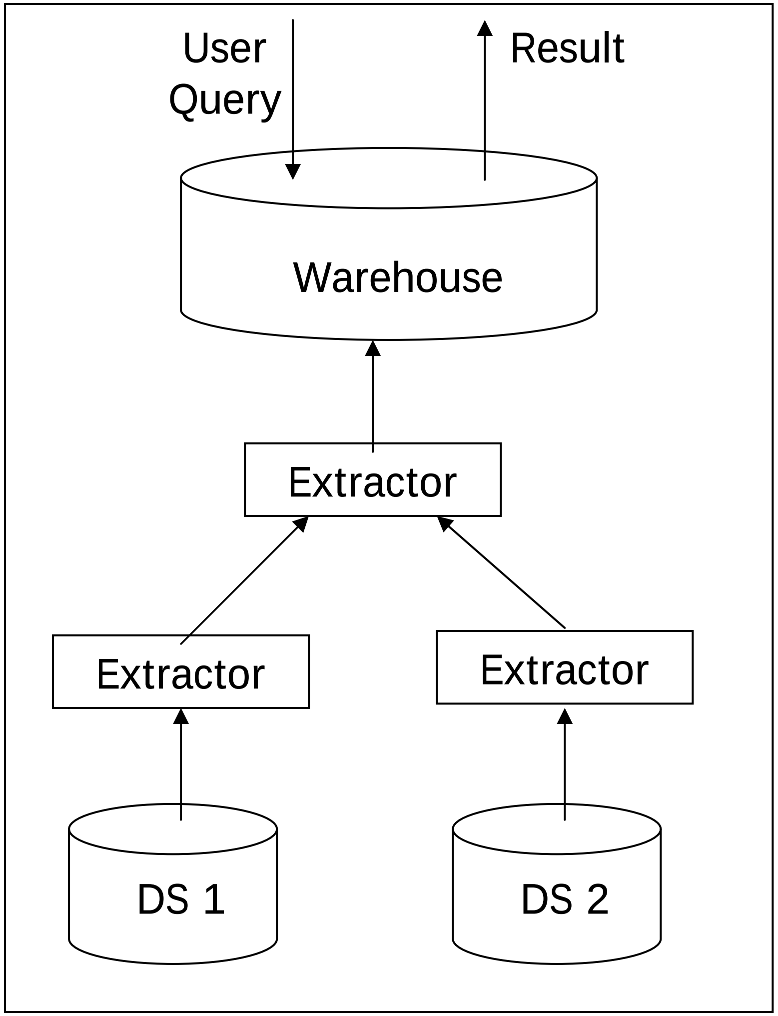
-- Mediators

- Supports a collection of views that integrate several sources.
- Unlike Data warehouse, the views are not materialized.
- The mediator sends queries to the corresponding wrappers.
- The results come back and are combined at the mediator



-- Data Warehouses

- Data from several sources is extracted and combined into a global schema.
- The data is stored in the warehouse
- User updates to the WH is generally forbidden, since they are not reflected in the source.
- Three ways of maintaining DW.
  - Periodic construction
  - Periodic update
  - Immediate update



- Problems of Information Integration

- Data in various Databases while having the same meaning can be represented in many different ways.
  - Data type difference
    - A field can be represented as character in one and integer in the other
  - Values difference
    - The same concept can be represented by different constants: example: sex can be represented as F and M or as 0 and 1.
  - Semantic difference
    - A relation in one DB excludes some entities while the same relation in another DB includes the same entities.
  - Missing values
    - A certain attribute in a relation in one DB may be missing from the corresponding relation in the other DB.

- On-Line Analytical Processing (OLAP)

- What is OLAP
- OLAP Applications
- A Multidimensional View of OLAP Data
- Star Schema
- Data Cubes
- OLAP Queries

-- What is OLAP

- The activity of querying a DW for patterns or trends of importance for an organization.
- Involve highly complex queries that use one or more aggregations.
  - These queries are often termed OLAP or decision support system (DSS) queries.
- In contrast to OLTP queries, OLAP queries typically examine large number of data.
- Example:

Shema

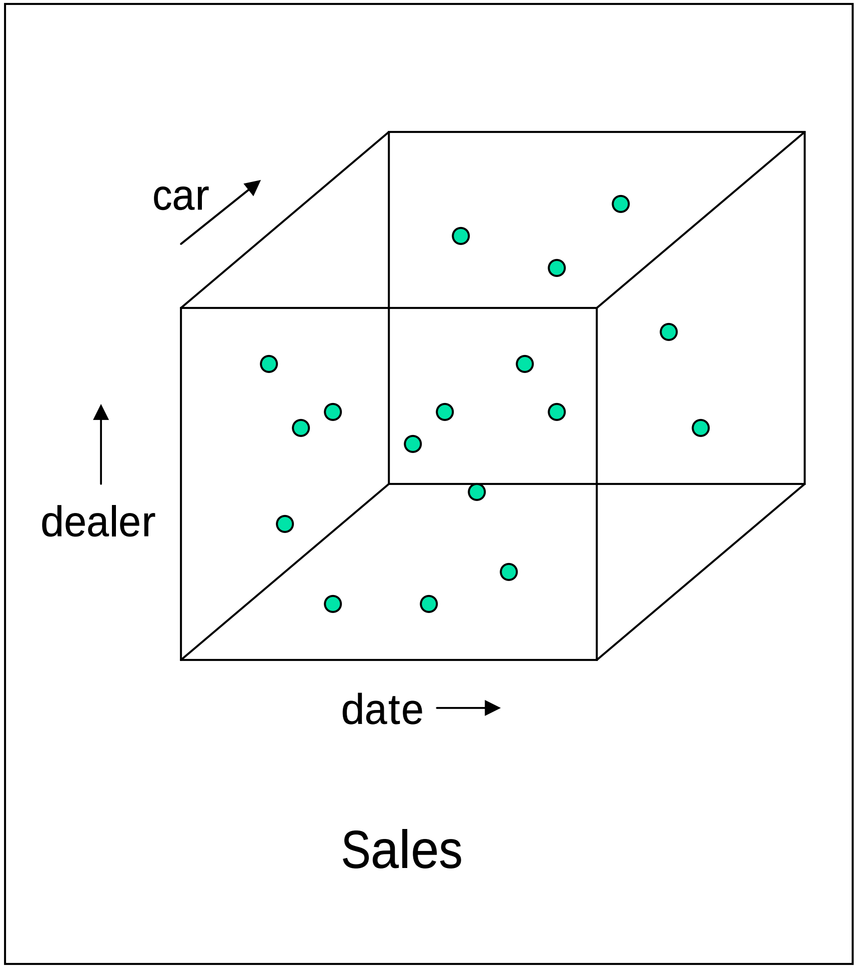
```
Sales(serialNo, date, dealer, price)
Autos(serialNo, model, color)
Dealers(name, city, state, phone)
```

OLAP (DSS) query

```
SELECT state, AVG(price)
FROM Sales, Dealer
WHERE Sales.dealer = Dealers.name
AND date > '2001-01-04'
GROUP BY state;
```

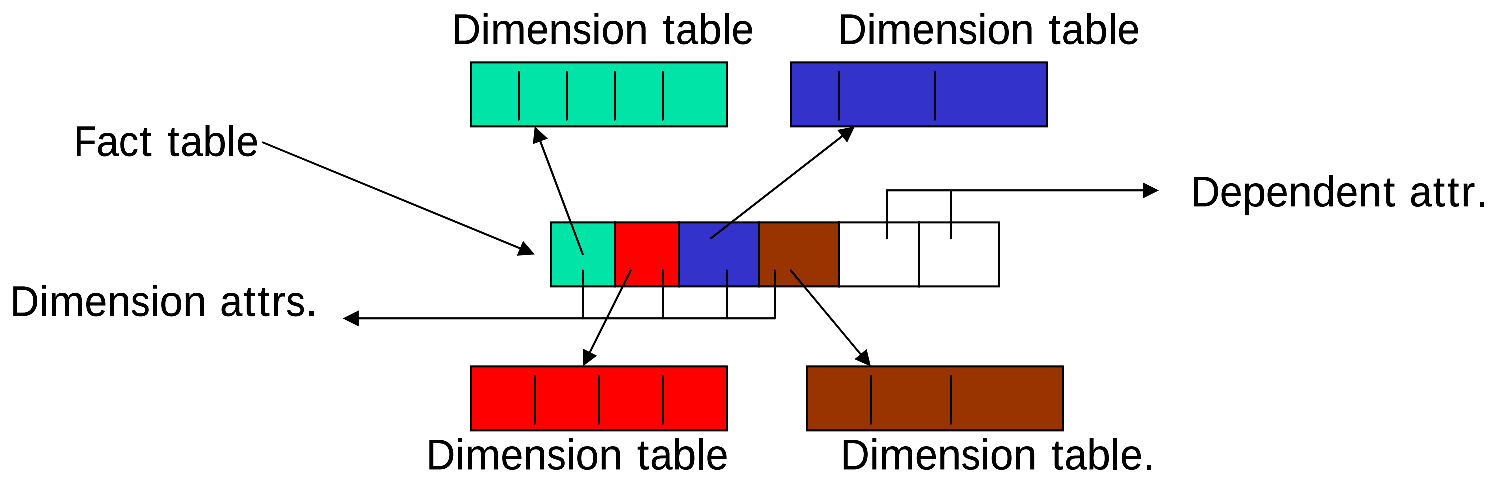
-- A Multidimensional View of OLAP Data

- In typical OLAP applications there is a central relation called fact table.
- Fact table represents events or objects of interest such as sales.
- It helps to envision the records in a fact table as arranged in a multidimensional space (cube).



-- Star Schema

- A star schema has 2 types of tables
  - **A fact table**
    - Is the center of the star and is linked to other relations
    - It normally has several attributes that represent dimension and one or more dependent attributes that represent the properties of interest.
  - **Dimension tables**: Smaller tables which are referenced by the fact table.



-- Data Cubes ...

Fact relation

sale	Product	Client	Amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Two-dimensional cube

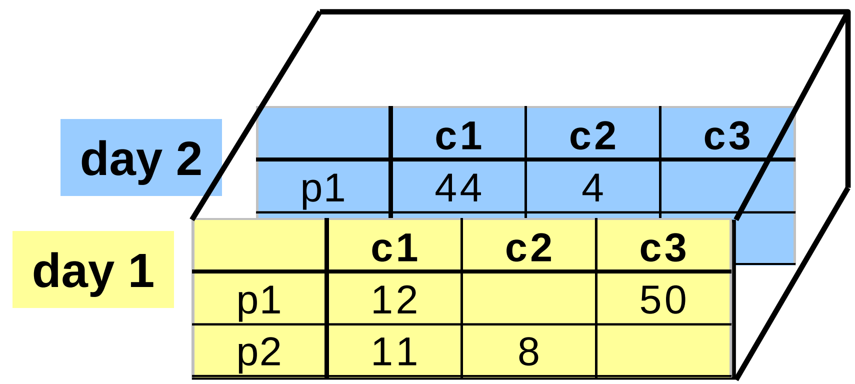
	c1	c2	c3
p1	12		50
p2	11	8	

... -- Data Cubes ...

Fact relation

sale	Product	Client	Date	Amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

3-dimensional cube



… -- Data Cubes …

Example: computing sums

…

day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3
p1	56	4	50
p2	11	8	

	c1	c2	c3
sum	67	12	50

	sum
p1	110
p2	19

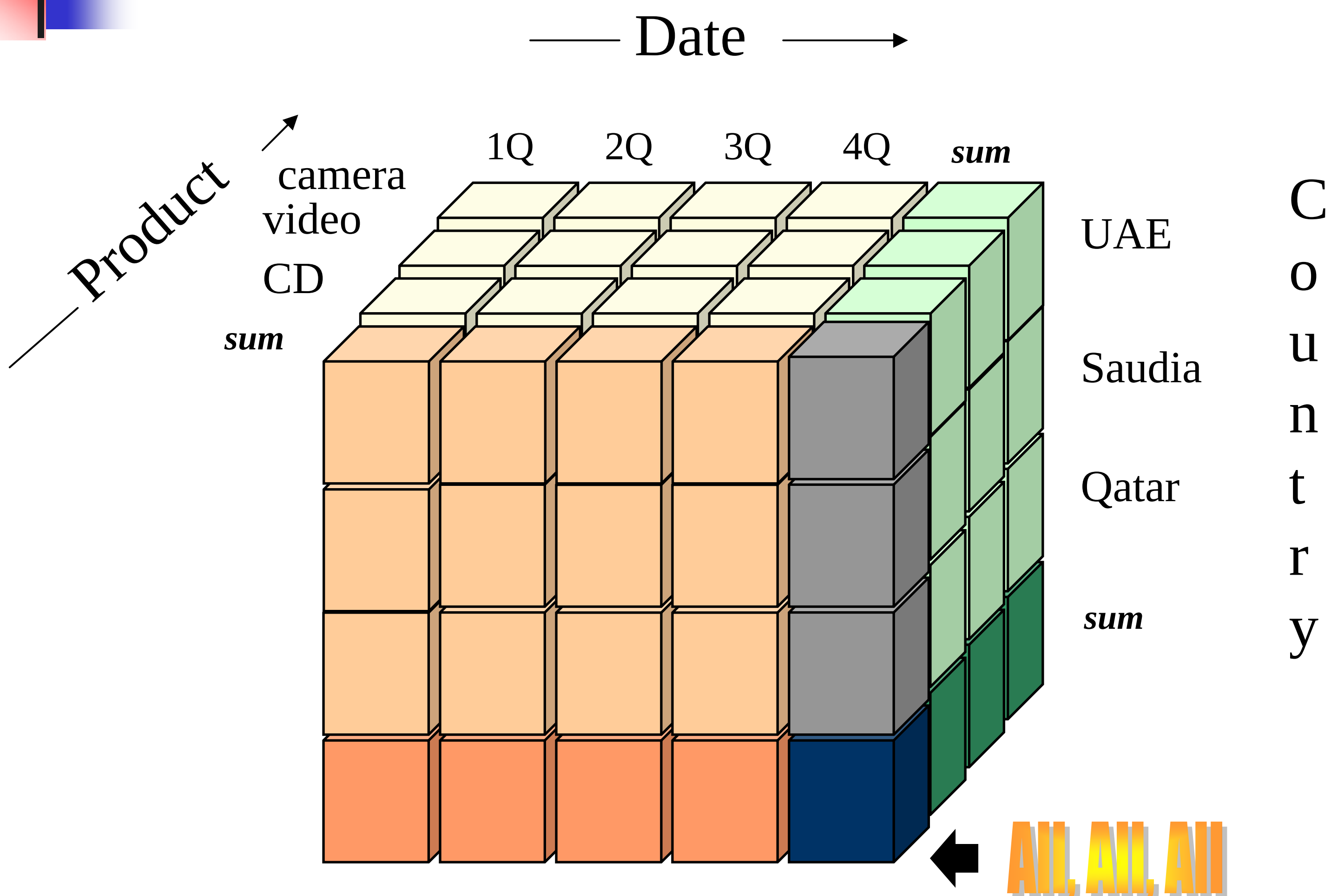
129

... -- Data Cubes ...

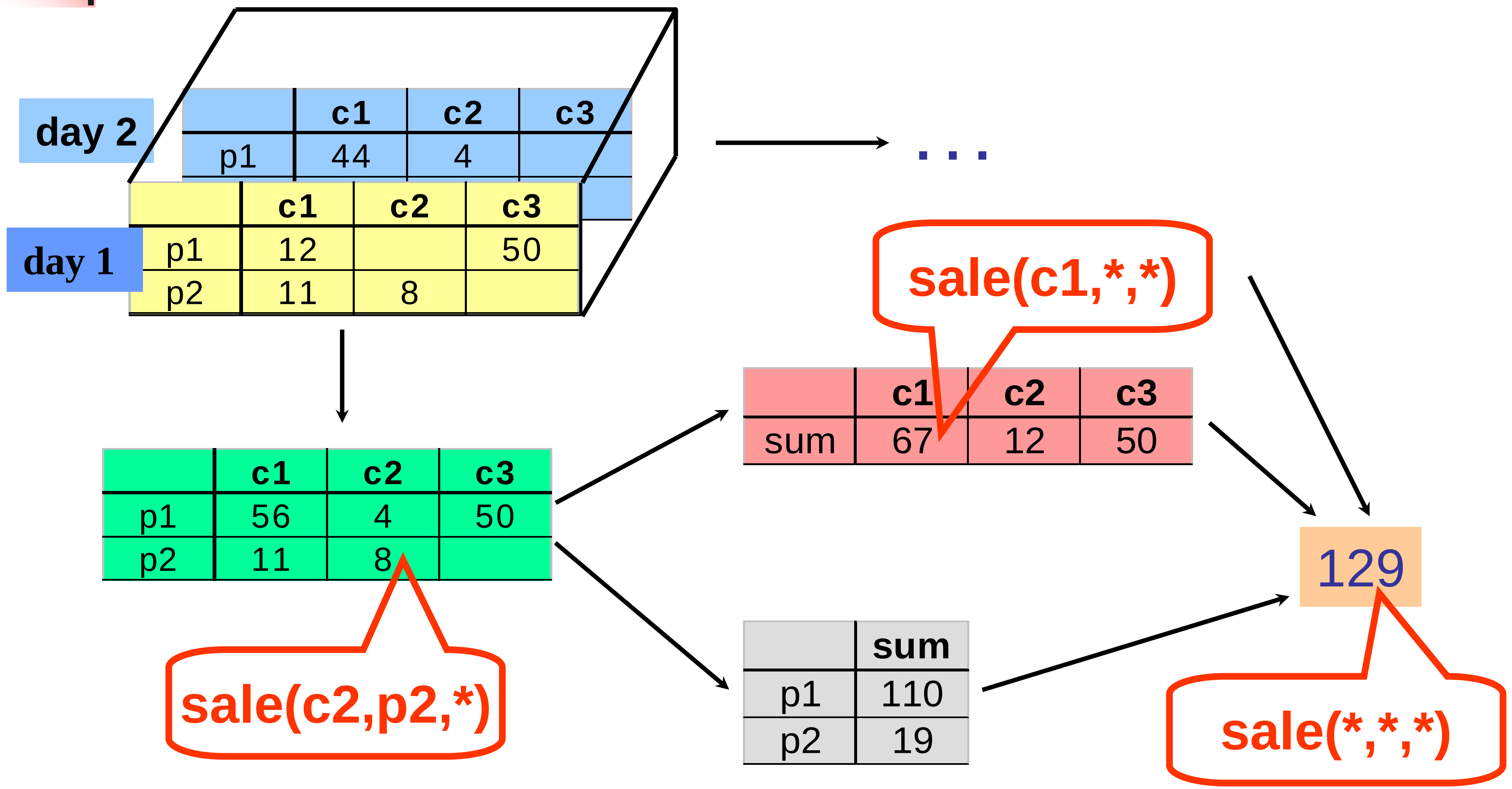
- In multidimensional data model together with measure values usually we store summarizing information (aggregates)

	c1	**c2**	**c3**	**Sum**
p1	56	4	50	110
p2	11	8		19
Sum	67	12	50	**129**

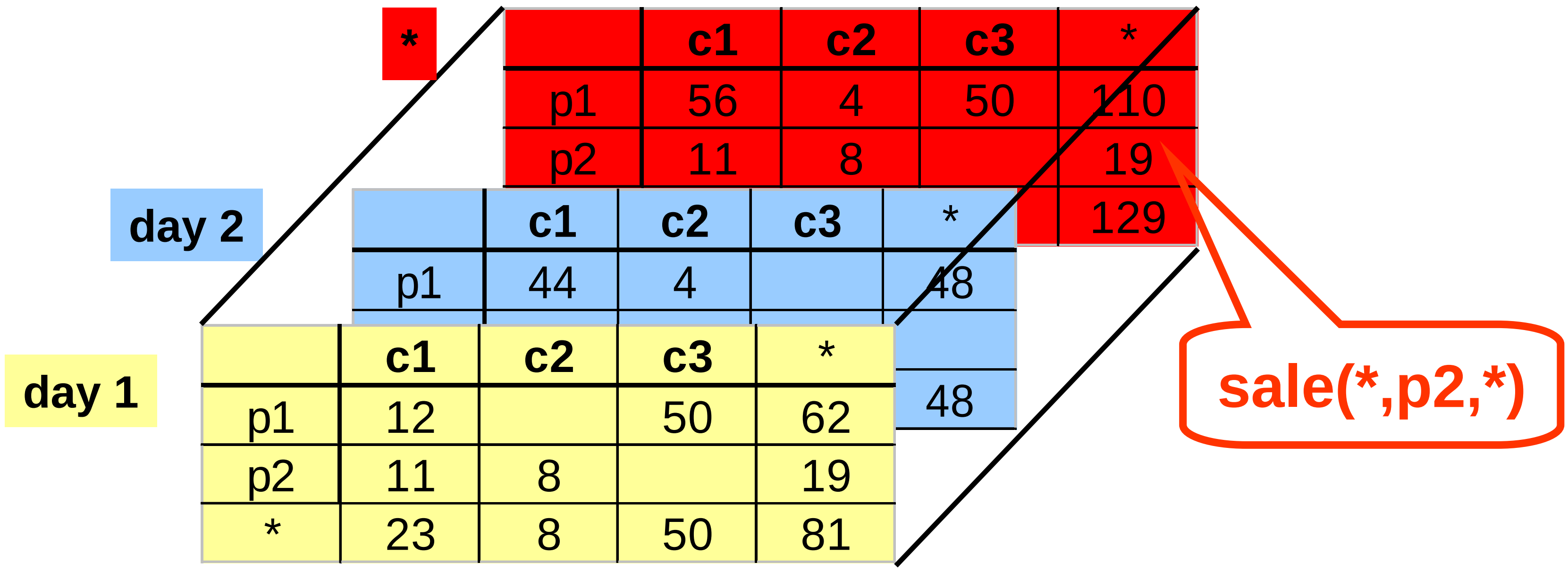
... -- Data Cubes


Date
Product
1Q
2Q
3Q
4Q
sum
camera
video
CD
sum
UAE
Saudia
Qatar
sum
Country
All, All, All

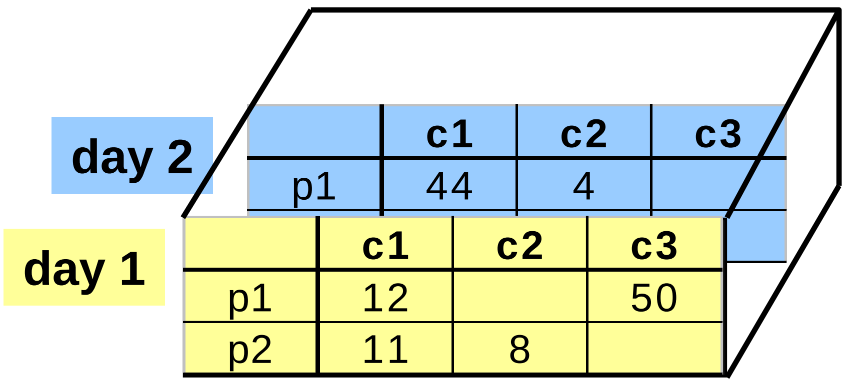
-- The Cube Operator ...



... -- The Cube Operator ...



-- Aggregation Using Hierarchies ...



	region A	region B	
p1	12	50	
p2	11	8	

customer

region

country

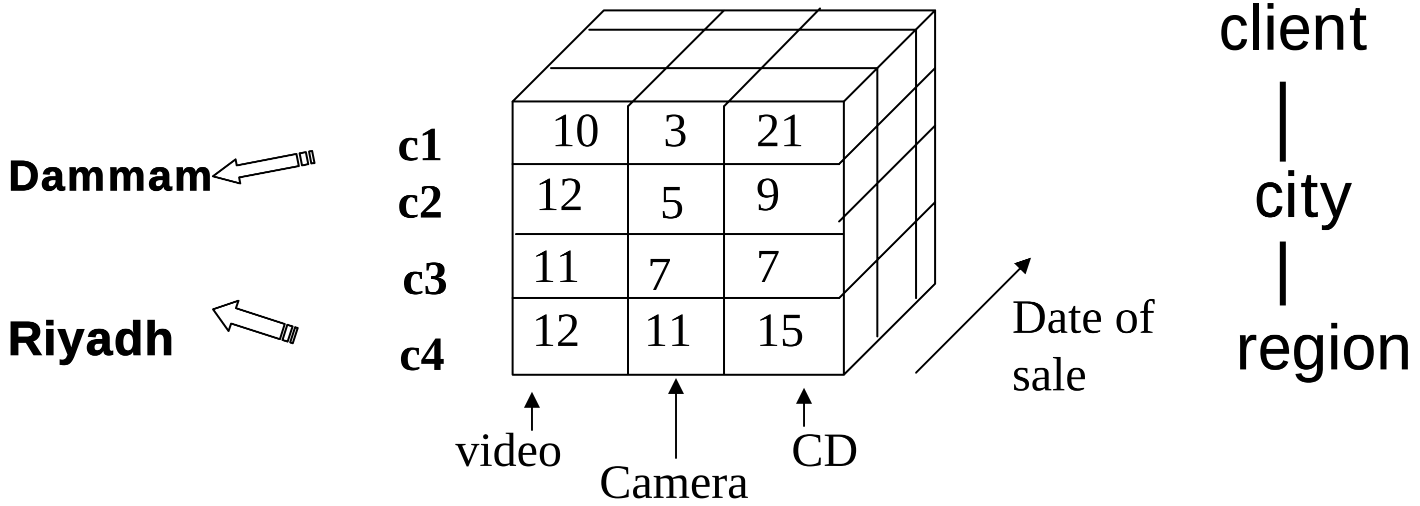
(customer c1 in Region A;
customers c2, c3 in Region B)

-- OLAP Servers

- Relational OLAP (ROLAP)
  - Extended relational DBMS that maps operations on multidimensional data to standard relations operations.
  - Store all information, including fact tables, as relations
- Multidimensional OLAP (MOLAP)
  - Special purpose server that directly implements multidimensional data and operations
  - Store multidimensional datasets as arrays.

-- OLAP Queries: Roll Up

- Summarizes data along dimension.



Roll up
aggregation with respect to city

	Video	Camera	CD
Dammam	22	8	30
Riyadh	23	18	22

-- OLAP Queries: Drill Down ...

- Roll down, drill down: go from higher level summary to lower level summary or detailed data
  - For a particular product category, find the detailed sales data for each salesperson by date
  - Given total sales by state, we can ask for sales per city, or just sales by city for a selected state

... -- OLAP Queries: Drill down

day 2

	c1	**c2**	**c3**
p1	44	4	

day 1

	c1	**c2**	**c3**
p1	12		50
p2	11	8	

	c1	**c2**	**c3**
p1	56	4	50
p2	11	8	

	c1	**c2**	**c3**
sum	67	12	50

	sum
p1	110
p2	19

129

rollup →

← drill-down

-- Other OLAP Queries ...

- Slice and dice: select and project
  - Sales of video in USA over the last 6 months
  - Slicing and dicing reduce the number of dimensions
- Pivot: reorient cube
  - The result of pivoting is called a cross-tabulation
  - If we pivot the Sales cube on the Client and Product dimensions, we obtain a table for each client for each product value

... Other OLAP Queries

- Pivoting can be combined with aggregation

sale	prodId	clientid	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

day 2

	c1	c2	c3
p1	44	4	

day 1

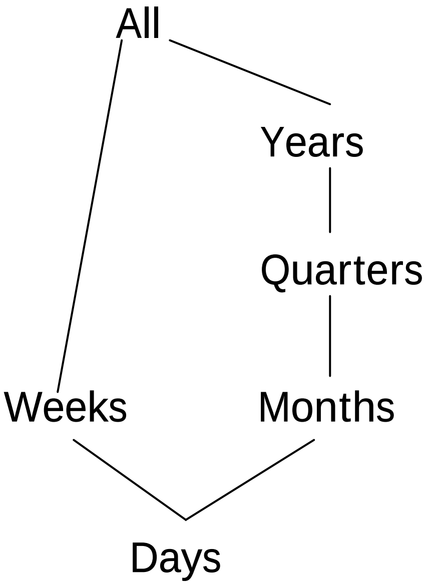
	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3	Sum
1	23	8	50	81
2	44	4		48
Sum	67	12	50	129

	c1	c2	c3	Sum
p1	56	4	50	110
p2	11	8		19
Sum	67	12	50	129

-- Cube Implementations

- Data cubes are implemented by materialized views
- A materialized view is the result of some query, which we chose to store its output table in the database.
- For the data cube, the views we would choose to materialize will typically be aggregations of the full data cube.
- Lattice of views are created for performance reasons



Lattice

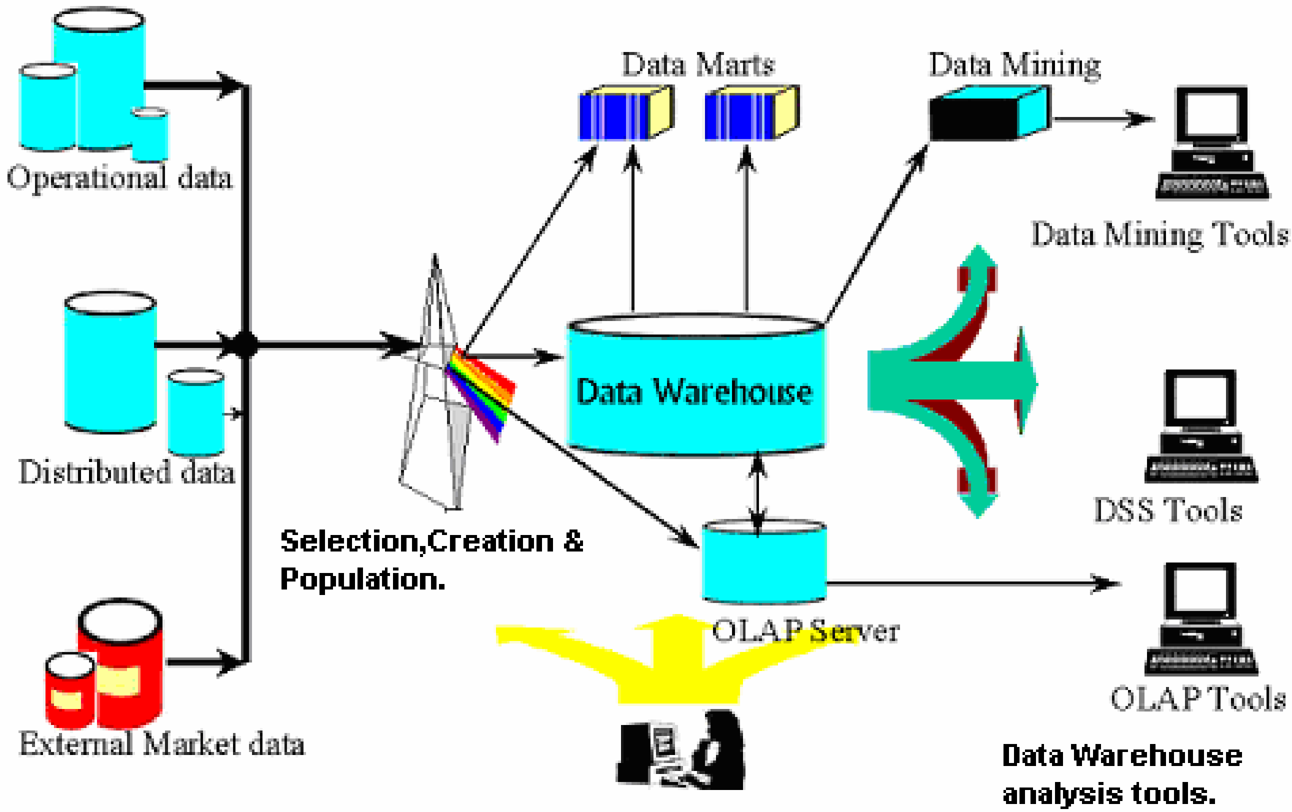
- Data Mining

- Data mining: an introduction
- Goals of data mining
- Knowledge discovery during data mining
- Applications of data mining

-- Data Mining: An Introduction

- Data mining refers to the discovery of new information in terms of patterns or rules from vast amounts of data
- Data warehousing and Data mining
  - Data mining can be used in conjunction with a data warehouse to help with certain decisions
  - Data mining can be applied to operational databases but to make it more efficient and meaningful it is applied to data warehouses
- Data mining applications should be considered early during the design of a data warehouse

-- Data Warehouse Architecture



-- Goals of Data Mining

- **Prediction** --- data mining can show how certain attributes within the data will behave in the future
- **Identification** --- data patterns can be used to identify the existence of an item, event, or an activity
- **Classification** --- data mining can partition the data so that different classes or categories can be identified based on combinations of parameters
- **Optimization** --- one eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials

-- Knowledge Discovery During Data Mining

- Deductive knowledge vs. inductive knowledge
- Data mining addresses inductive knowledge
- The knowledge discovered during data mining can be described as
  1. Association rules
  2. Classification hierarchies
  3. Sequential patterns
  4. Patterns within time series
  5. Categorization and segmentation

-- Types of Knowledge Discovered During Data Mining

- **Association rules** --- correlate the presence of a set of items with another range of values for another set of variables
- **Classification hierarchies** --- create hierarchies of classes
- **Sequential patterns** --- sequence of actions or events
- **Pattern with time series** --- similarities detected within positions of the time series
- **Categorization and segmentation** --- partition a given population of events or items into sets of “similar” elements.

--- Association Rules ...

- An **association rule** is of the form $X \Rightarrow Y$ where $X = \{x_1, x_2, \ldots., x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$ are sets of distinct items. The rule states that if a customer buys X, he is also likely to buy Y
- **Support** for the rule LHS $\Rightarrow$ RHS is the percentage of transactions that hold all the items in the union, the set LHS $\cup$ RHS.
- **Confidence** for the rule LHS $\Rightarrow$ RHS is the percentage (fraction) of all transactions that include items in LHS and out of these the ones that include items of RHS.

... --- Association Rules ...

- Example:

Transaction id	Time	items bought
101	6:35	milk, bread, cookies, juice
792	7:38	milk, juice
1130	8:05	milk, eggs
1735	8:40	bread, cookies, coffee

Milk ➔ Juice, 50% support, 66.7% confidence
Bread ➔ Juice, 25% support, 50% confidence

... --- Association Rules ...

- The goal of mining association rules is to generate all possible rules that exceed some minimum user-specified support and confidence thresholds.
- The problem of mining association rules is thus decomposed into two sub-problems:
  - Generate all item sets that have a support that exceeds the threshold. These sets of items are called **large itemsets.**
  - For each large item set, all the rules that have a minimum confidence are generated as follows:
    for a large itemset X and Y $\subset$ X, let Z = X - Y;
    then if support (X)/support (Z) $\Rightarrow$ minimum confidence, the rule Z $\Rightarrow$ Y (i.e., X - Y $\Rightarrow$ Y) is a valid rule.

... --- Association Rules ...

Basic Algorithms for Finding Association Rules

- The current algorithms (Apriori Algorithm) that find large itemsets are designed to work as follows:
  - Test the support for itemsets of length 1, called 1-itemsets, by scanning the database. Discard those that do not meet minimum required support.
  - Extend the large 1-itemsets into 2-itemsets by appending one item each time, to generate all candidate itemsets of length two. Test the support for all candidate itemsets by scanning the database and eliminate those 2-itemsets that do not meet the minimum support.
  - Repeat the above steps; at step *k*, the previously found (*k* - 1) itemsets are extended into *k*-itemsets and tested for minimum support.
  - The process is repeated until no large itemsets can be found.

... --- Association Rules

- Apriori Algorithm:
  - Is based on the following 2 properties:
    1. Antimonotonicity
    2. Downward closure
- Several other algorithms have been proposed to mine association rules:
  - Sampling algorithms
  - Frequent-pattern tree algorithm
  - Partition algorithm

-- Approaches to Other Data Mining Problems

- Discovery of sequential patterns
- Discovery of Patterns in Time Series
- Discovery of Classification Rules
- Regression
- Neural Networks
- Genetic Algorithms
- Clustering and Segmentation

-- Applications of Data Mining

- Data mining can be applied to a large variety of decision-making contexts in business like
  - Marketing
  - Finance
  - Manufacturing
  - Health care

- Reading list

- All Chapter 20 except sections 20.2 and 20.3

END