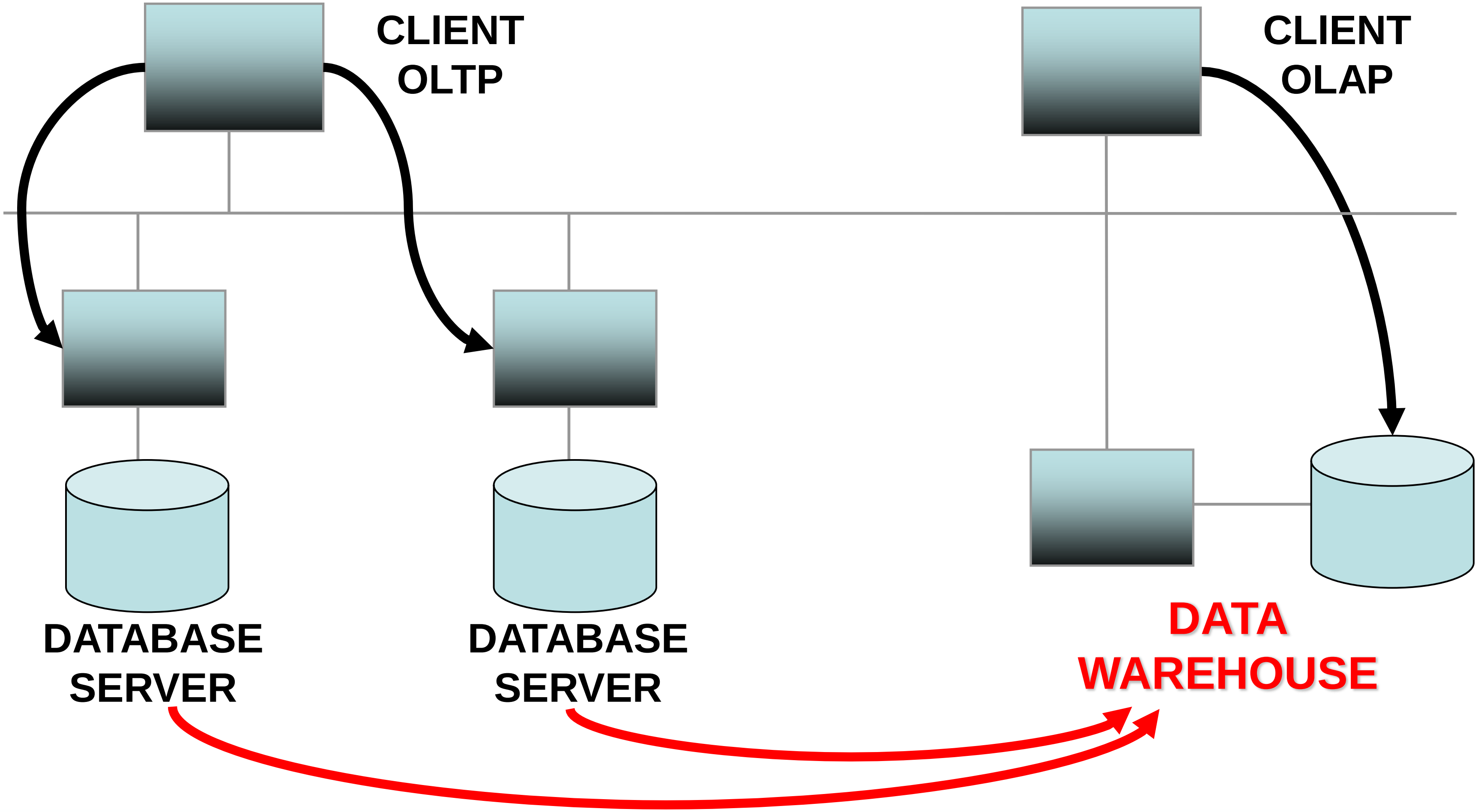
Advanced Databases

7 Data Warehouses

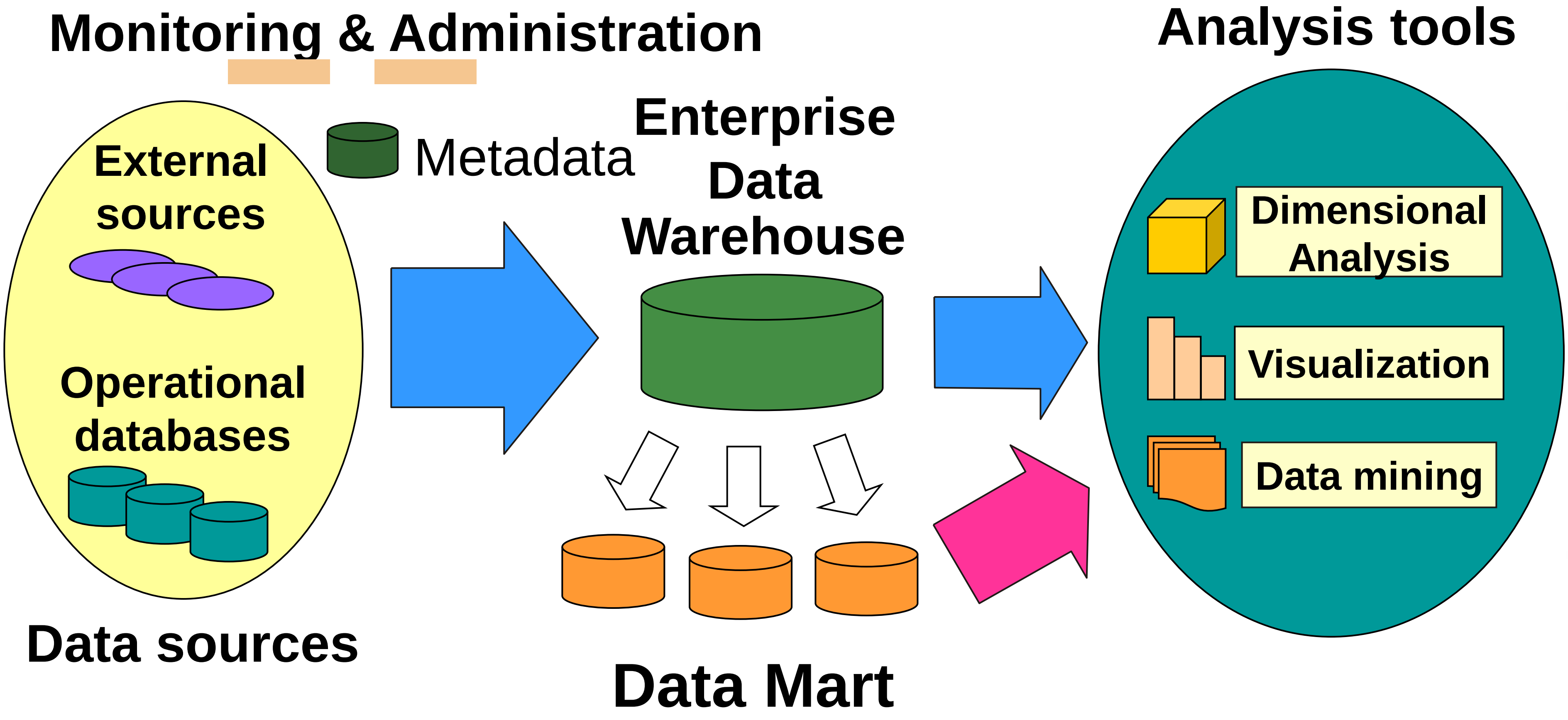
An Environment for Data Analysis

- **DATA WAREHOUSE**
  - A structured description of all those data that are necessary for a strategic analysis of the trends and the behaviour of a firm
- **ON-LINE ANALYTICAL PROCESSING (OLAP)**
  - The name given to analysis activities (it is contrasted to On Line Transaction Processing, OLTP)

Interaction between OLTP and OLAP



An Architecture for Data Warehousing



Data Warehouse (DW) and Data Mart (DM)

- A Data Warehouse often integrates several Data Marts
- Users typically address one specific Data Mart
- Data Marts share common data
- Each Data Mart is responsible for one specific aspect of the firm business

Star model

- The ***star model*** is used for each Data Mart
  - Also known as ***multi-dimensional schema***
- It is a conceptual model which poses some restrictions
- Advantages:
  - Availability of suitable specific query interfaces
  - Good performance
  - Straightforward design of the relational schema

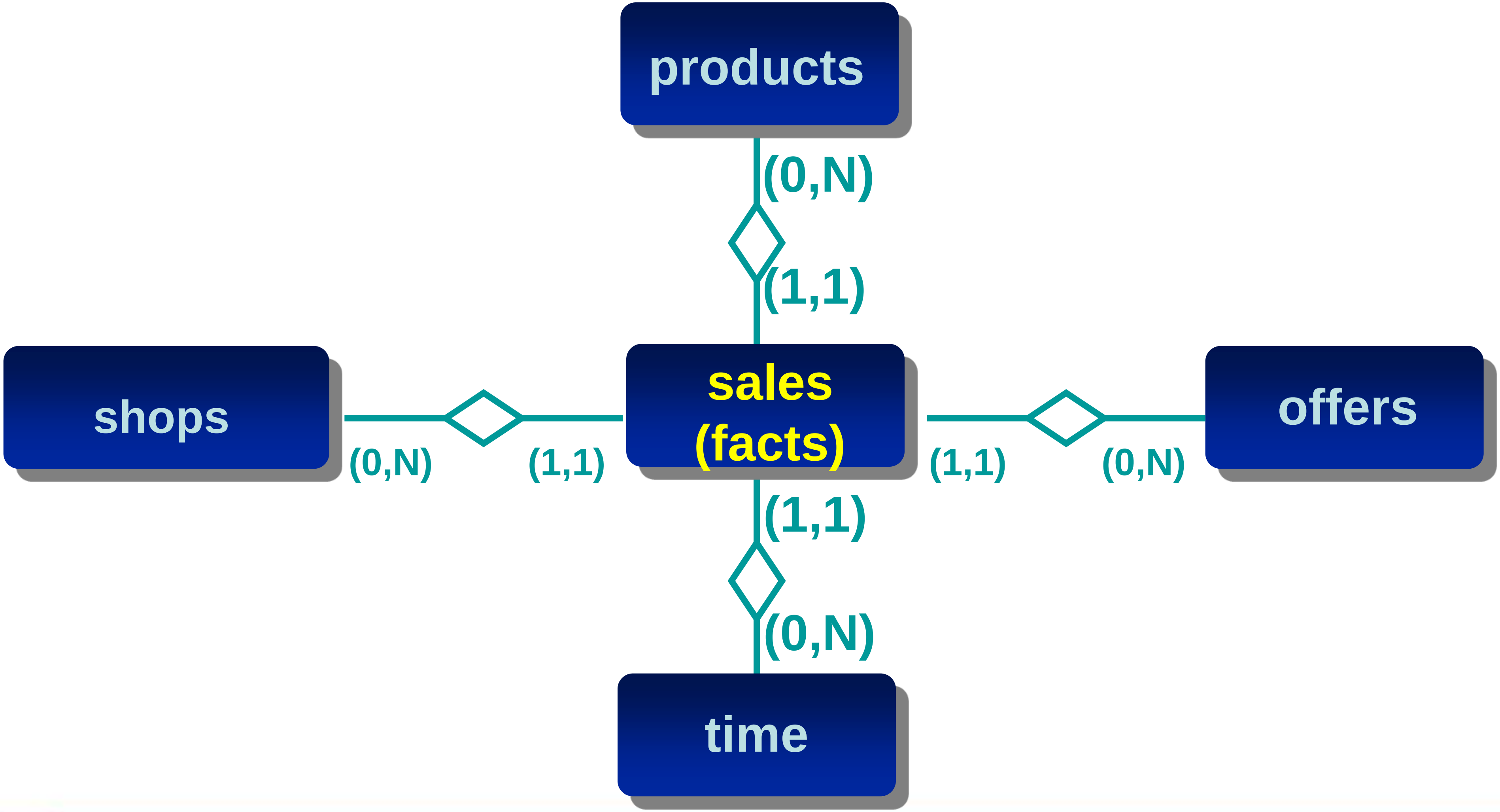
Multi-dimensional representation

- Relevant concepts:
  - **fact** — an aspect which is crucial for the analysis
  - **measure** — an atomic property of a fact
  - **dimension** — a specific perspective for the analysis

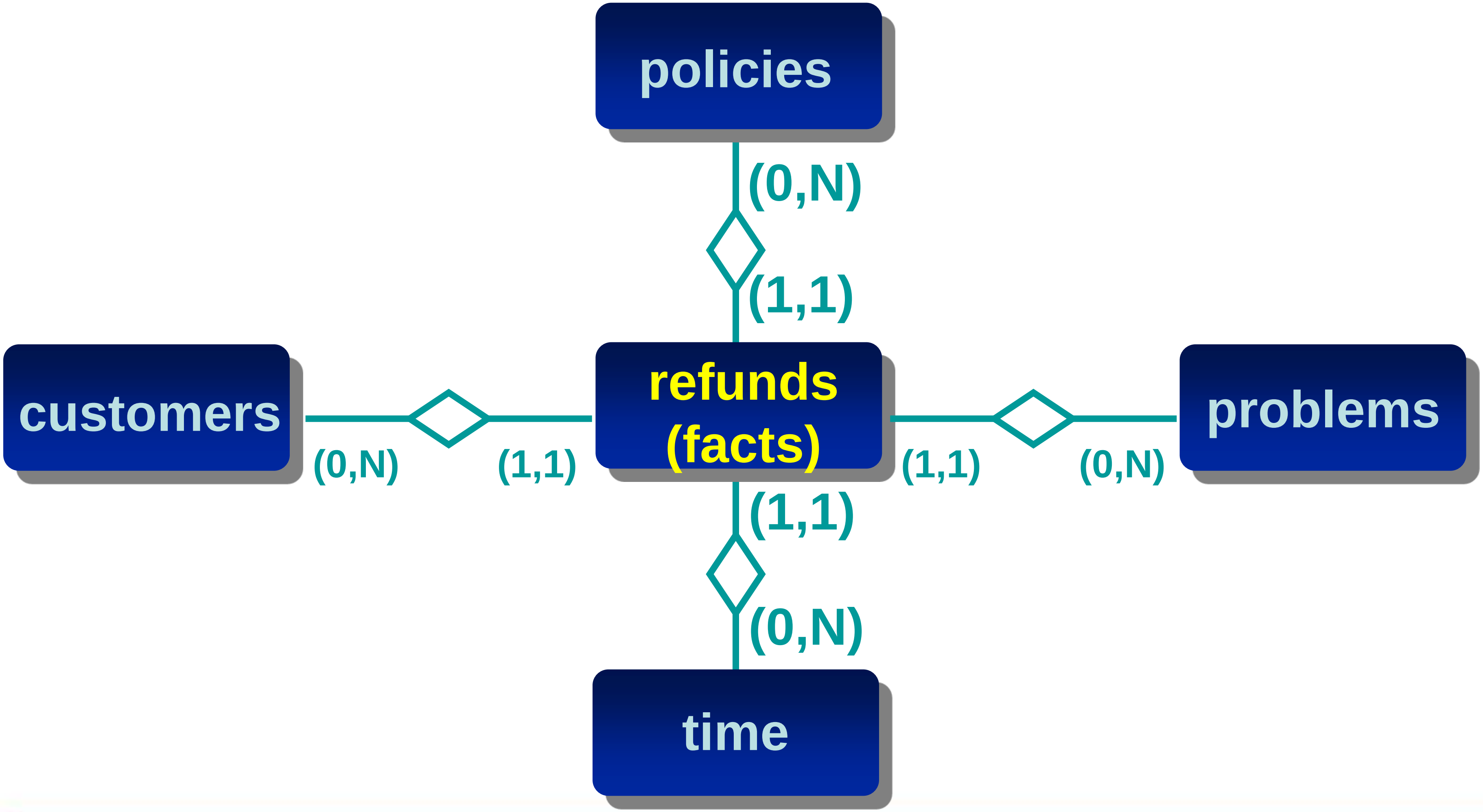
Examples of facts/measures/dimensions

- Retail shops:
  - Sales
  - Quantity, price
  - Product, time, zone
- Telephone service:
  - Phone call
  - Cost, duration
  - Caller, answerer, time

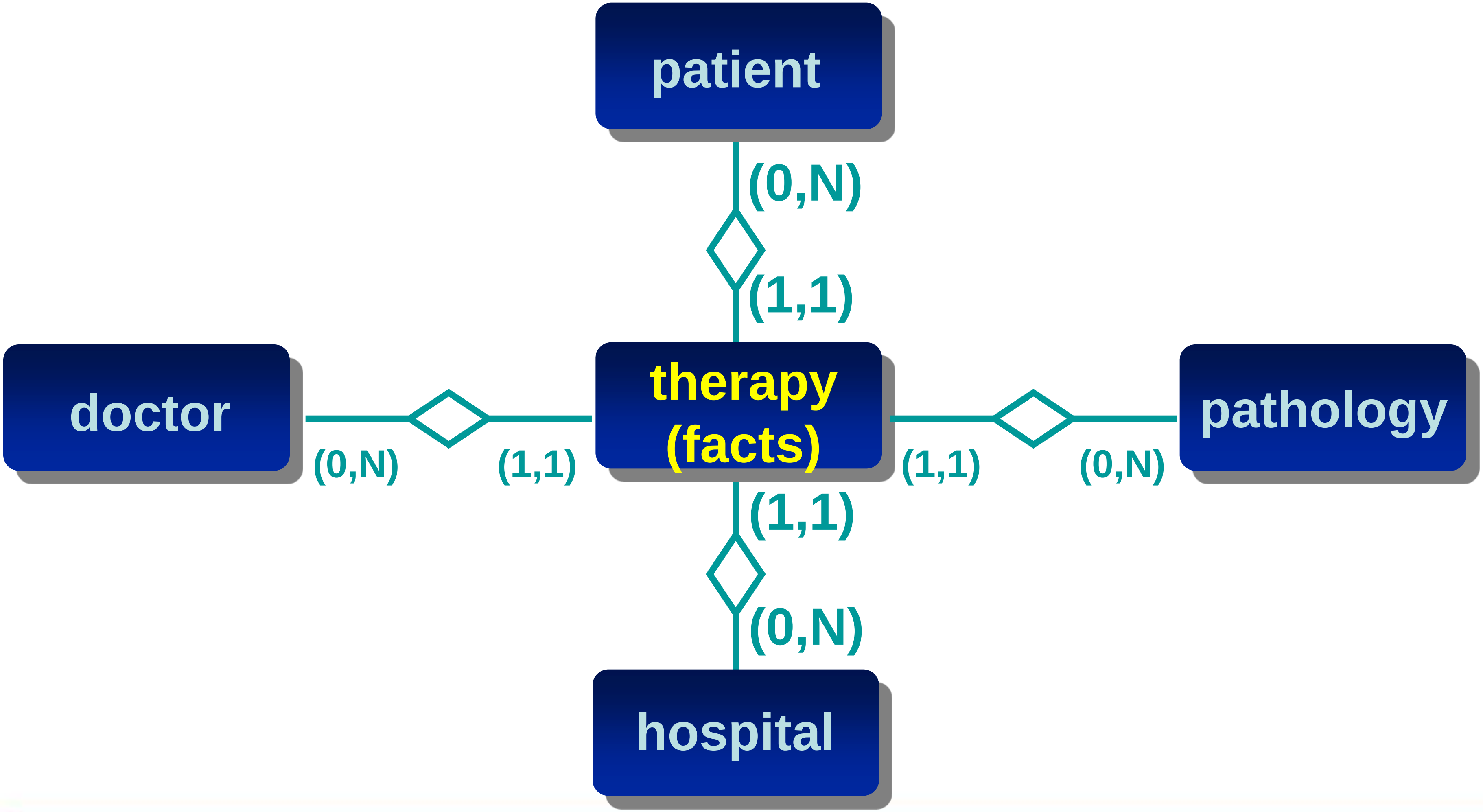
An example: sales management



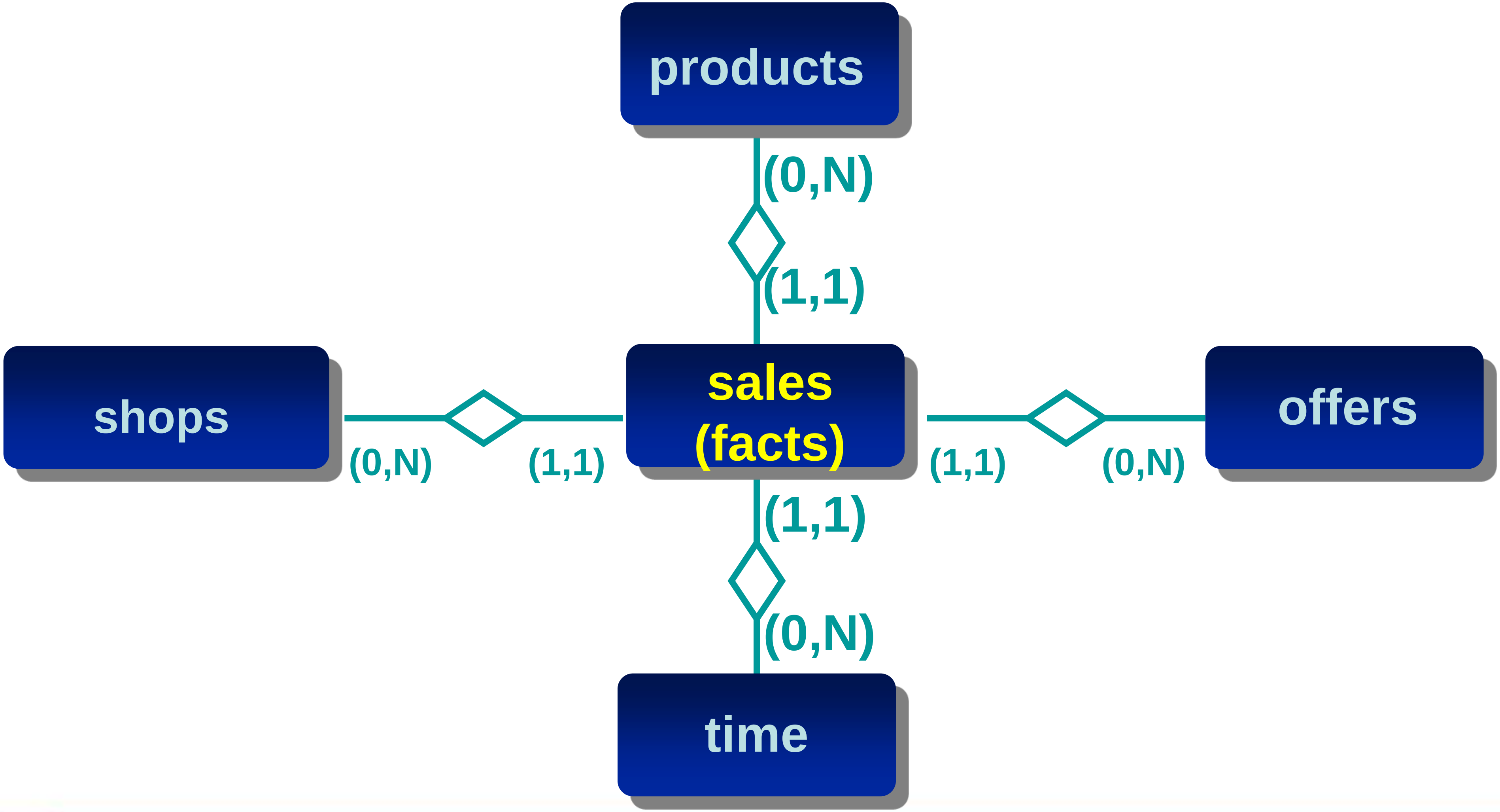
Another example: reimbursements



Another example: therapies



Consider again the sales management schema



Facts: sales

Product-ID
Shop-ID
Time-ID
Offer-ID
Total-proceeds
Quantity
Unit-proceeds

First dimension: products

Product-ID
Category
Sub-Category
Brand
Packing
Weight
Size
Provider

Second dimension: shops

<u>Shop-ID</u>
Name
Address
City
Sales-District
Phone
Manager-Name
Size
Logistics

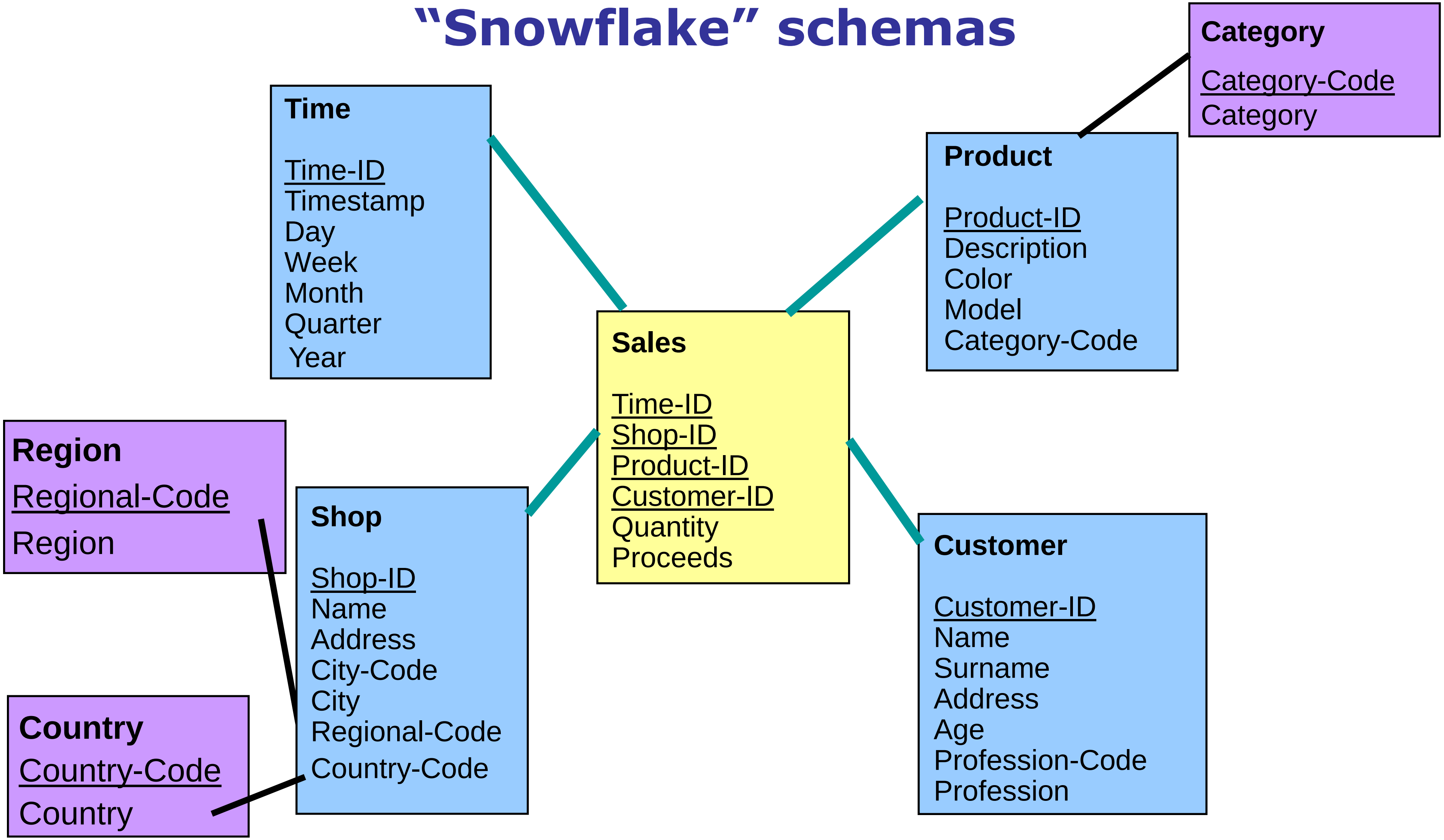
Third dimension: time

Time-ID
Day-in-Week
Day-in-Month
Day-in-Year
Week-in-Month
Week-in-Year
Month-in-Year
Season
Flag-WorkingDay
Flag-Sunday

Fourth dimension: offers

<u>Offer-ID</u>
Offer-name
Discount-Type
Discount-Percentage
Advertisement
Flag-Coupon
Start-Date
End-Date
Cost
Agency

“Snowflake” schemas



Time

Time-ID
Timestamp
Day
Week
Month
Quarter
Year

Product

Product-ID
Description
Color
Model
Category-Code

Category

Category-Code
Category

Sales

Time-ID
Shop-ID
Product-ID
Customer-ID
Quantity
Proceeds

Region

Regional-Code
Region

Shop

Shop-ID
Name
Address
City-Code
City
Regional-Code
Country-Code

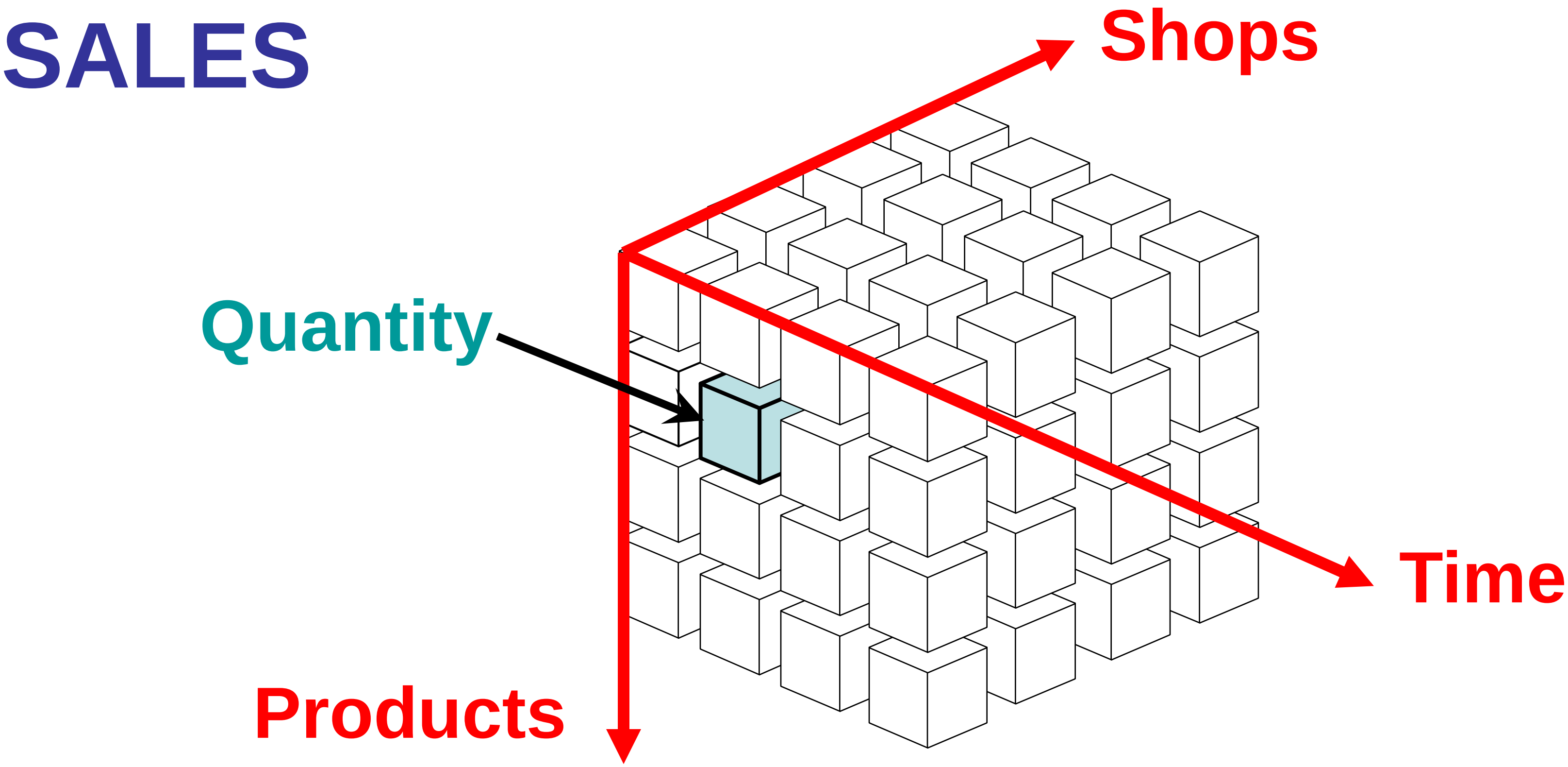
Country

Country-Code
Country

Customer

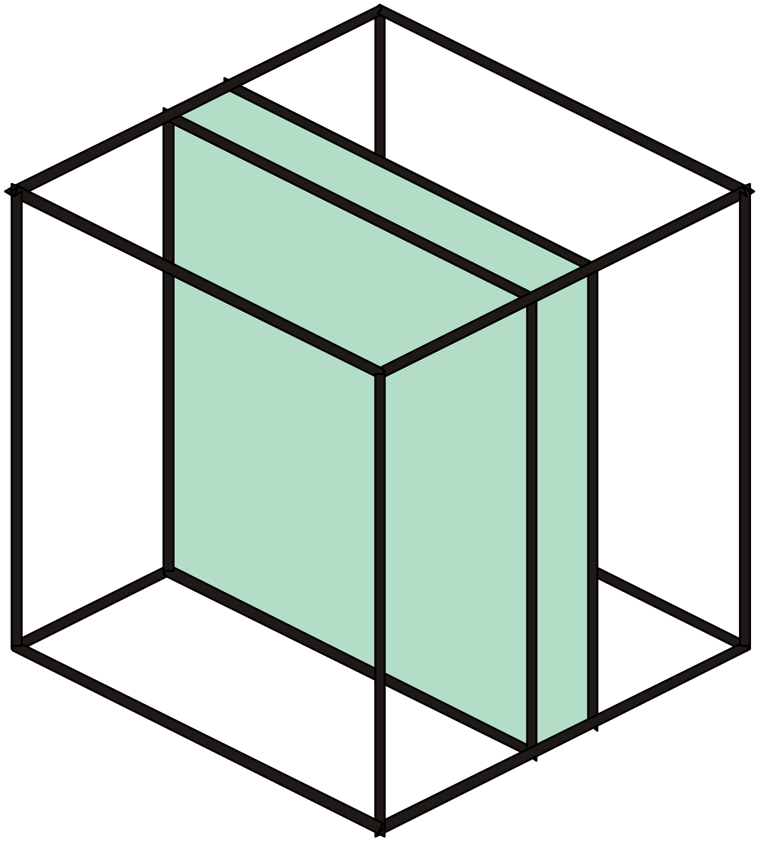
Customer-ID
Name
Surname
Address
Age
Profession-Code
Profession

Multi-dimensional data representation

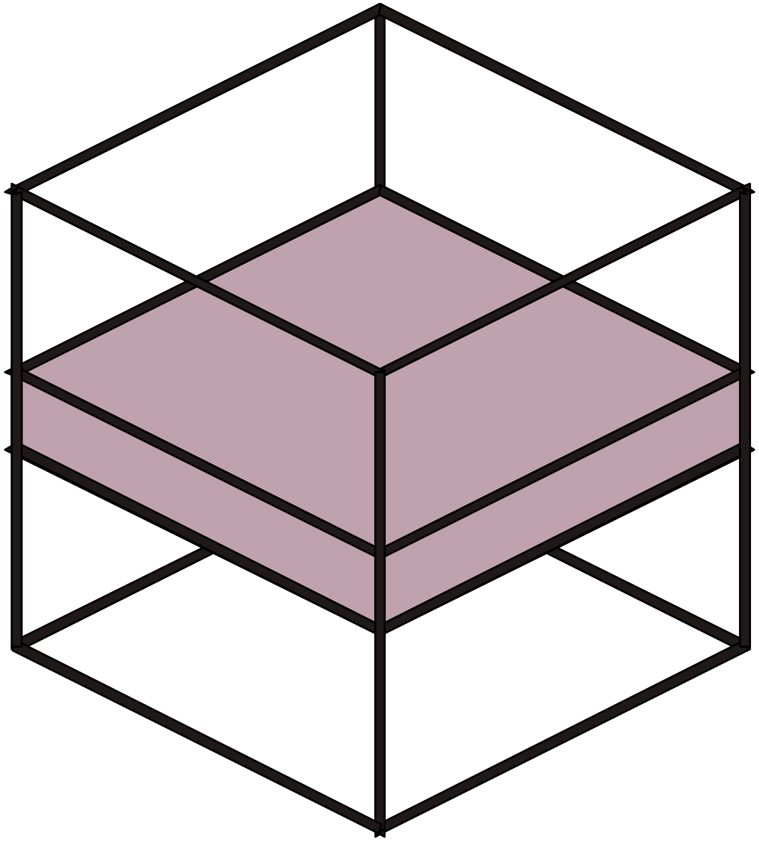


Data Warehouses

- The regional manager studies the sales of all products in all periods w.r.t. the shops of his region

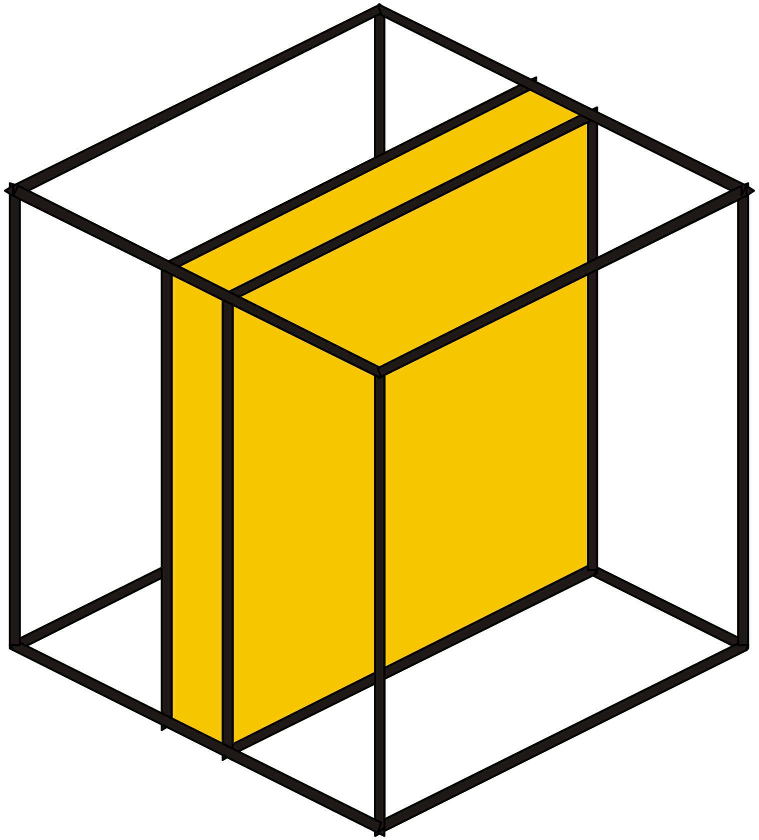


- The product manager studies the sales of a product in all periods and in all shops



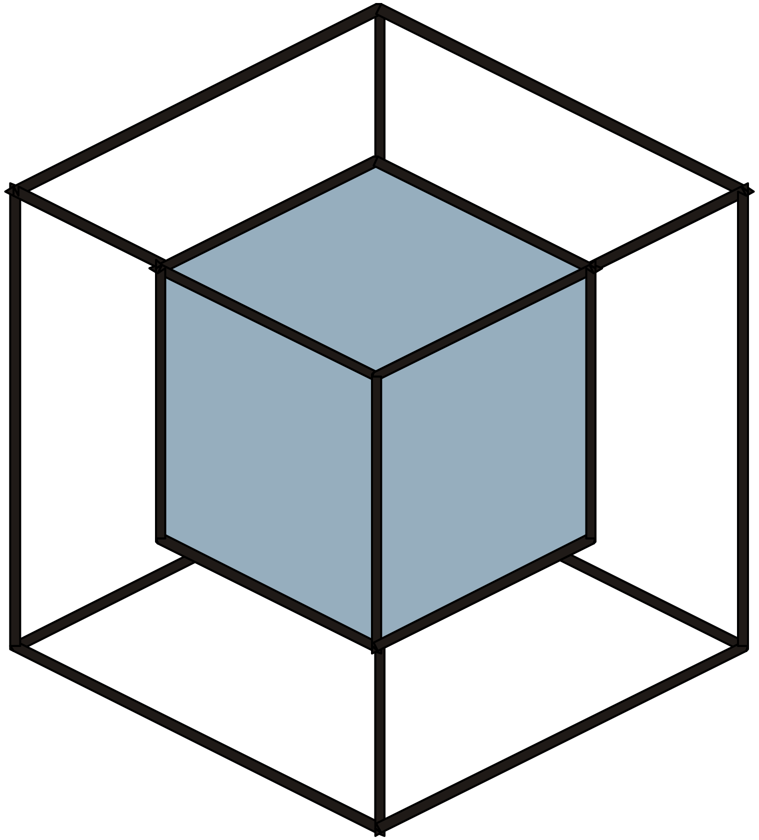
Data Warehouses

- The financial manager studies the sales of all products in all shops, comparing the current period with the previous one



Data Warehouses

- The strategic manager focuses on one category of products, one area and a limited period of time



Data Visualization

- Data are visualized and rendered graphically, so as to be easy to understand

- Common means of visualization:
  - Tables
  - Pie/doughnut charts
  - Column/bar histograms
  - Line charts
  - 3D surfaces
  - Bubble charts
  - Area blocks
  - Cylinders/cones/pyramids
  - …

Example of query with a browser

Offer	period	zone	product	dimension of analysis
Pay 2 & buy 3	March	north	milk	total-proceeds
40% discount	April	east	bread	quantity
20% discount	May	west	pasta	unit-proceeds
1-free-mug (...)	....	....	....	
…..				

	February/
	April
			pasta	sum(proceeds)
				sum(quantity)

The “same” query in SQL

```
select  c1, c2, aggr(c3), aggr(c4)
from facts, dim1, dim2
where join-predicate(facts,dim1)
  and join-predicate(facts,dim2)
  and selection-predicate(dim1)
  and selection-predicate(dim2)
group by c1, c2
order by c1, c2
```

Result

Month	product	Sum of proceeds	Sum of quantity
February	pasta	110.000	45.000
March	pasta	95.000	50.000
April	pasta	105.000	51.000

Operations over multi-dimensional data

- ***Roll up*** — aggregates data
  - Sums up the sale quantity over last year per each region and product category
- ***Drill down*** — disaggregates data
  - For one particular product category in a region, “unrolls” and shows in detail the sale quantities of each day in each shop
- ***Slice & dice*** — selection and projection
- ***Pivot*** — change the orientation of the data cube

Drill Down: adding one dimension (Zone)

Month	Product	Zone	Sum of quantity
February	pasta	north	15.000
February	pasta	east	17.000
February	pasta	west	13.000
March	pasta	north	18.000
March	pasta	east	18.000
March	pasta	west	14.000
April	pasta	north	18.000
April	pasta	east	17.000
April	pasta	west	16.000

Roll-up: removing one dimension (Month)

Product	Zone	Sum of quantity
pasta	north	51.000
pasta	east	52.000
pasta	west	43.000

Aggregate queries

- **Examples**:
  - Total proceeds for each product category in each shop in each day
  - Total monthly proceeds in each shop
  - Total monthly proceeds for each product category in each shop
  - Average monthly proceeds for each category (calculated over all shops)

Aggregates in SQL: data cube

- Expresses all possible aggregations of the tuples of a table
- Uses a new purpose-specific *polymorphic* value: **ALL**

Data cube in SQL

```
select Model, Year,
       Color, sum( Quantity )
from Sales
where Model in ('Fiat','Ford')
  and Color = 'Red'
  and Year between 1994 and 1995
group by Model, Year, Color
with cube
```

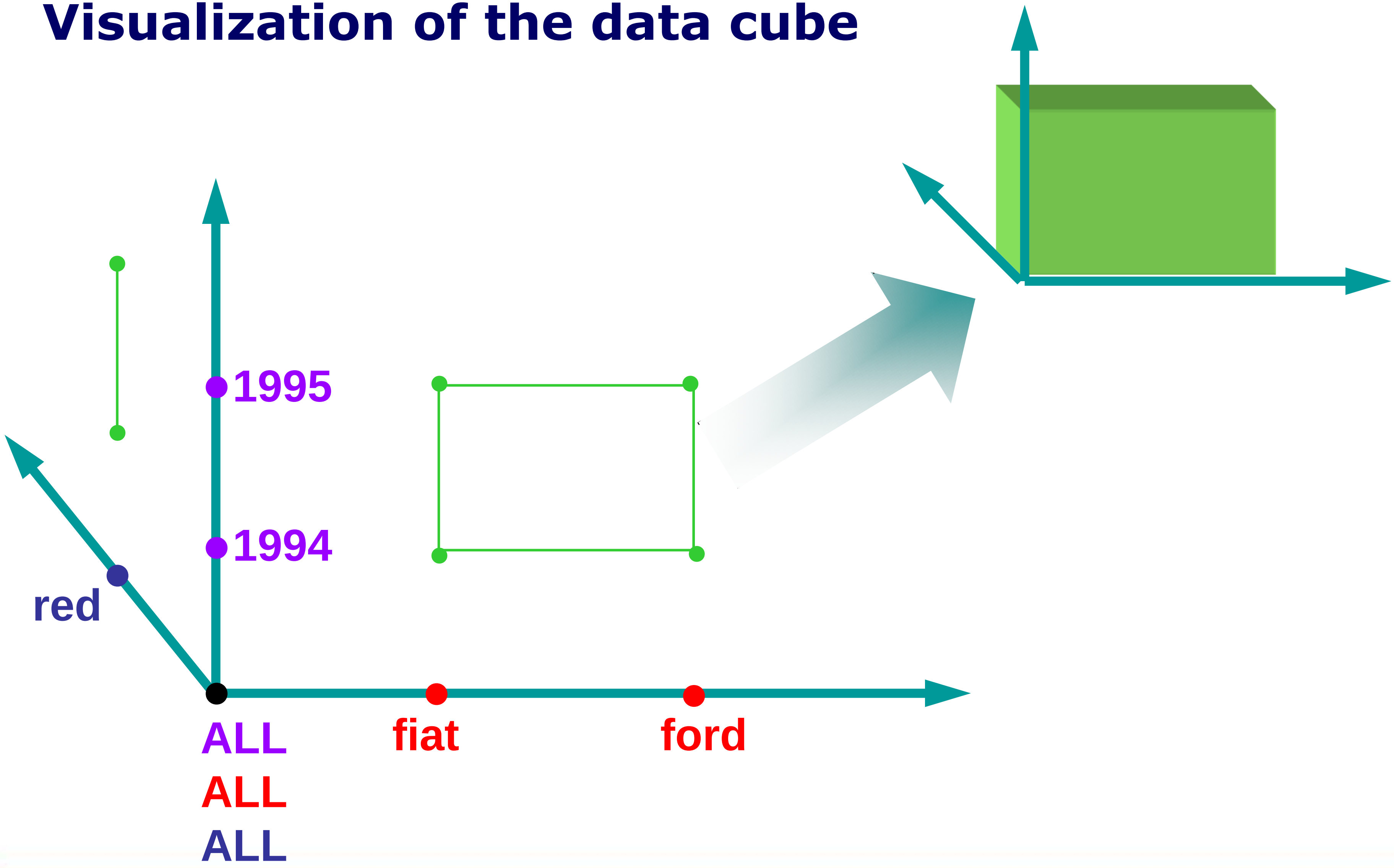
Relevant facts

Model	Year	Color	Quantity
fiat	1994	red	50
fiat	1995	red	85
ford	1994	red	80

Data cube results:

model	year	color	sum (quantity)
fiat	1994	red	50
fiat	1995	red	85
fiat	1994	ALL	50
fiat	1995	ALL	85
fiat	ALL	red	135
fiat	ALL	ALL	135
ford	1994	red	80
ford	1994	ALL	80
ford	ALL	red	80
ford	ALL	ALL	80
ALL	1994	red	130
ALL	1995	red	85
ALL	ALL	red	215
ALL	1994	ALL	130
ALL	1995	ALL	85
ALL	ALL	ALL	215

Visualization of the data cube



Roll up in SQL

```
select Model, Year,
       Color, sum( Quantity )
from Sales
where Model in ('Fiat','Ford')
  and Color = 'Red'
  and Year between 1994 and 1995
group by Model, Year, Color
with roll up
```

Roll up results:

Model	Year	Color	sum(Quantity)
fiat	1994	red	50
fiat	1995	red	85
ford	1994	red	80
fiat	1994	ALL	50
fiat	1995	ALL	85
ford	1994	ALL	80
fiat	ALL	ALL	135
ford	ALL	ALL	80
ALL	ALL	ALL	215

Typical Size of a Data Warehouse

time: **730 days**
shops: **300**
products: **30.000**
daily sales: **3.000**
offers: **at most one per product on sale**

sales: **730 * 300 * 3000 * 1 = 657 millions**

Size: **657 millions * 8 attributes * 4 Byte = 21 GB**

Classification of OLAP System

- **M**OLAP (**M**ulti-dimensional OLAP)

  as alternative to

- **R**OLAP (**R**elational OLAP)

  - MOLAP: the internal data storage is not relational, so as to guarantee better performance

  - ROLAP: the relational storage guarantees the capability of managing large volumes of data

Specific OLAP Technologies

- **Bitmap Indexes**
  - Allow for efficient evaluation of OR and AND combinations of simple comparison predicates
- **Join Indexes**
  - Pre-computed joins between the table of facts and the tables representing the dimensions
- **Materialized views**
  - Those views are pre-computed, which can be used to answer most frequently asked queries

7 Data Warehouses

Data Mining

Data mining

- Objective
  - Extract information *hidden* into data so as to support strategic decisions

- An inter-disciplinary task (and subject)
  - Statistics
  - Algorithmics
  - Neural networks
  - Fractals
  - …

Applications of Data Mining

- Market analysis
  - Which products are purchased together or one before another? (basket analysis)
- Analysis of behaviours
  - Identify fraudulent credit card usage
- Forecasts
  - Foreseeing the cost of medical treatments
- Control
  - Industrial production errors

An example: sales analysis

Transaction	Date	Item	Qty	Price
1	12/17/95	ski-pants	1	140 €
1	12/17/95	ski-boots	1	180 €
2	12/18/95	T-shirt	1	25 €
2	12/18/95	jacket	1	300 €
2	12/18/95	ski-boots	1	70 €
3	12/18/95	jacket	1	300 €
4	12/19/95	T-shirt	3	25 €
4	12/19/95	jacket	1	300 €

Association Rules

- Association rules look for *regularity* within data
  - When a customer buys ski-boots, she also buys skis

- Structure of association rules:

Body $\Rightarrow$ ***Head***

- *Body:* premise of the rule
- *Head:* consequence of the rule

An example of Association Rule

Diaper $\Rightarrow$ Beer

- 2% of all transactions contain both items
- 30% of transactions containing Diaper also contain Beer

Characteristics of Association Rules

- **Support**
  - Probability that both Head and Body are in the same transaction [P(H,B)]
- **Confidence**
  - Probability that the Head is in a transaction t, given that the Body **is** in t [P(H|B), *conditional probability*]
- **Problem statement**
  - Extract from a dataset all association rules with support and confidence over given thresholds

Examples of association rules

Body	Head	Support	Confidence
ski-pants	ski-boots	0.25	1
ski-boots	ski-pants	0.25	1
T-shirt	ski-boots	0.25	0.5
T-shirt	jacket	0.25	1
ski-boots	T-shirt	0.25	0.5
ski-boots	jacket	0.25	1
jacket	T-shirt	0.5	0.66
jacket	ski-boots	0.25	0.33
{ T-shirt, ski-boots }	jacket	0.25	1
{ T-shirt, jacket }	ski-boots	0.25	0.5
{ ski-boots, jacket }	T-shirt	0.25	1

Other Examples

- Items sold in the same special offer
- Items frequently purchased together in summer but not in winter
- Items frequently purchased together as in the shop they are arranged in a particular layout (adjacent, near, ...)
- Items purchased in consecutive transactions by the same customer

Sequential Patterns

- Input dataset:
  - All the transactions of a given customer

- Objective:
  - Find those sequences of items which are frequently contained into corresponding sequences of transactions, such that the frequency is over a given threshold

Examples

- “5% of customers bought a CD player in a transaction and some CDs in the following two transactions”
- “10% of the purchases of a television set is followed by the purchase of a video-recorder”
- Applications
  - Measure of the customer satisfaction
  - Special offers tailored for specific customer classes
  - Medicine (sequences of symptoms ⇒ disease)

Discretization

- A continuous domain can be represented by means of a sequence of suitable intervals
  - Example: blood-pressure
    - High: >250
    - Medium: >130, <250
    - Low: <130
- *Objective*: find the correlation between the risk of infarction and blood-pressure with a given statistical significance
- *Advantages*:
  - Compact value representation
  - Determination of critical values
  - Facilitation of future data analysis

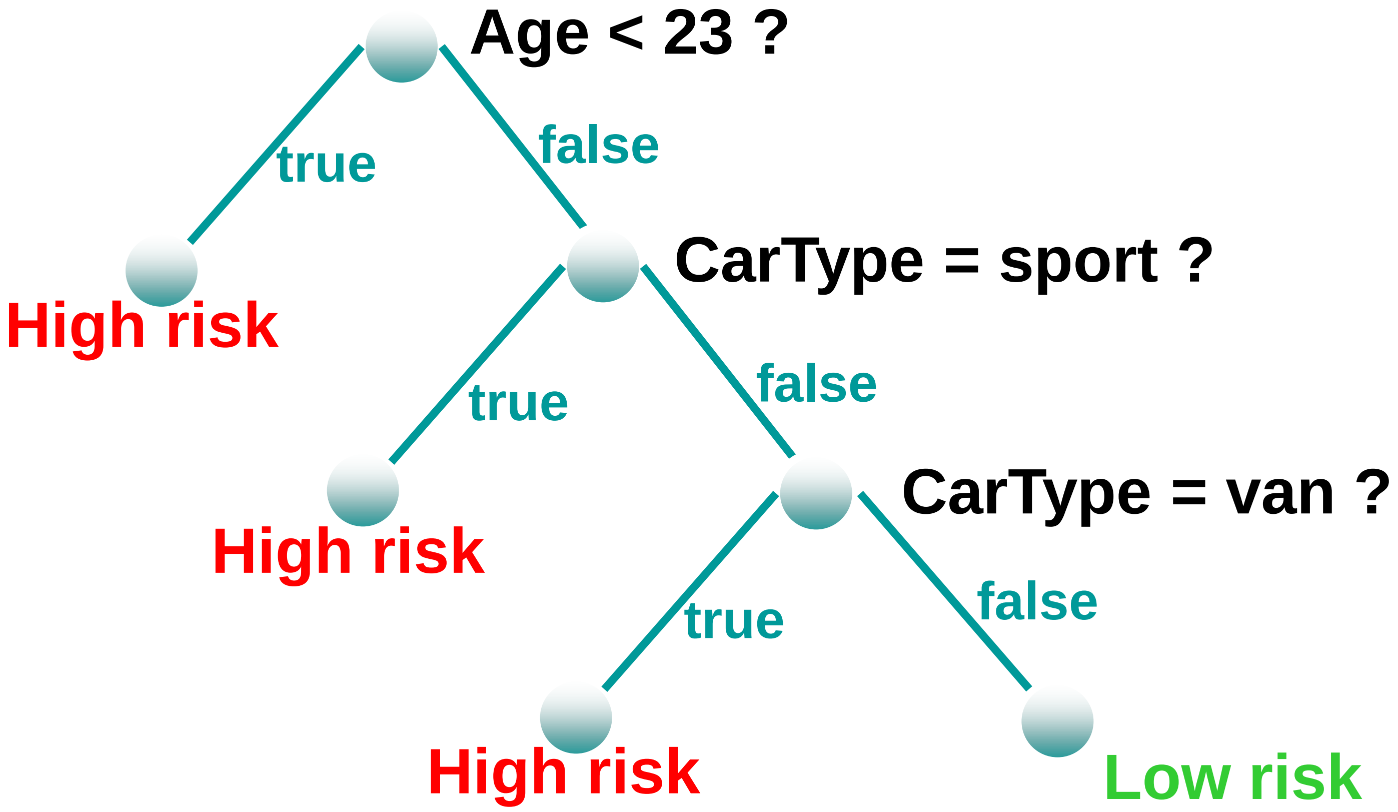
Classification

- Cataloguing a fact, concept, or phenomenon into a pre-defined class

- The phenomenon is described by elementary facts (atomic data, within **tuples**)
- The classifier is constructed and trained over a set of training data (**training set**)
- Classifiers are represented as **decision-trees**

Example: identify risky policies

POLICY (DrivingLicense, Age, CarType)


Age < 23 ?
true
false
High risk
CarType = sport ?
true
false
High risk
CarType = van ?
true
false
High risk
Low risk