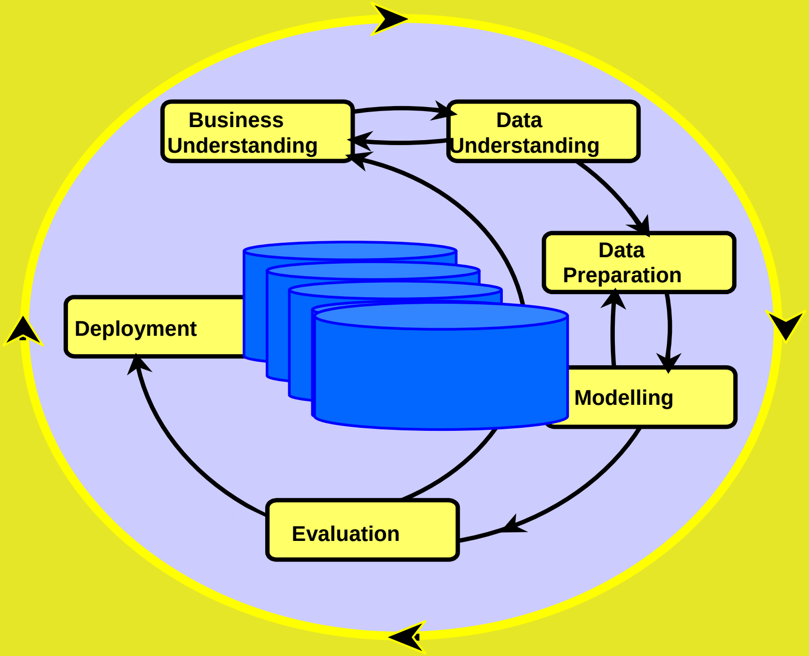
Statistic Methods in Data Mining



Data Mining Process

Professor Dr. Gholamreza Nakhaeizadeh

Short review of the last lecture

Introduction

- Literature used
- Why Data Mining?
- Examples of large databases
- What is Data Mining?
- Interdisciplinary aspects of Data Mining
- Other issues in recent data analysis: Web Mining, Text Mining
- Typical Data Mining Systems
- Examples of Data Mining Tools
- Comparison of Data Mining Tools
- History of Data Mining, Data Mining: Data Mining rapid development
- Some European funded projects
- Scientific Networking and partnership
- Conferences and Journals on Data Mining
- Further References

Examples of applications

- Optimal structure of a Data Mining Team
- Success factors of DM-Applications
- Predictive Modeling
- Data Mining in Business and Banking
- Data Mining in Quality Management

Data Mining Process

CRISP-DM :

- **Provides an overview of the life cycle of a data mining project**

- **Consists of six phases**

- **was partially funded by the European Commission**

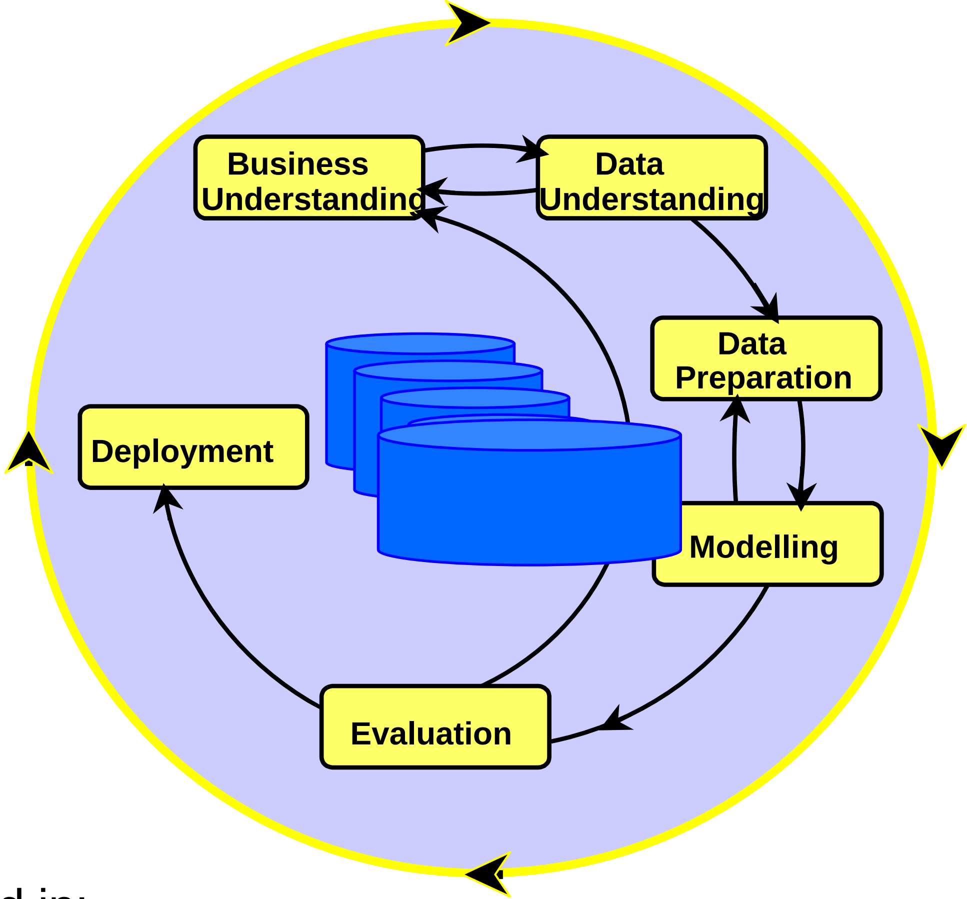
Project Partner:

Teradata
a division of NCR

DAIMLERCHRYSLER

SPSS

van de mensen van OHRA



- CRISP-DM Process Model is described in:

http://www.crisp-dm.org/CRISPwP-0800.pdf

Data Mining Process

CRISP-DM: Business Understanding

http://www.crisp-dm.org/CRISPwP-0800.pdf

- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan

Data Mining Process

CRISP-DM: Data Understanding

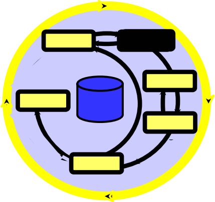
General aspects

- Collect initial data
- Describe data
- Explore data
- Verify data quality

Data Mining Process

CRISP-DM: Data Understanding

Collecting initial data



Can the data be accessed effectively and efficiently ?

- How big is the needed storage ?
- How long does it take to access the data ?

• Is there any restriction in collecting the data ?

- privacy issues,
- too expensive data,
- too expensive collecting process,..

•…………

Data Mining Process

CRISP-DM: Data Understanding

Collecting initial data

what are the needed data ? where are the data ?

Examples of data sources

UCI KDD Database Repository for large datasets used machine learning and knowledge discovery research.
UCI Machine Learning Repository.
Delve, Data for Evaluating Learning in Valid Experiments
FEDSTATS, a comprehensive source of US statistics and more
FIMI repository for frequent itemset mining, implementations and datasets.
Financial Data Finder at OSU, a large catalog of financial data sets
GeneSifter Data Center, access to microarray datasets through the GeneSifter microarray data analysis system.
GEO (GEO Gene Expression Omnibus), a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
Grain Market Research, financial data including stocks, futures, etc.
Investor Links, includes financial data
Microsoft's TerraServer, aerial photographs and satellite images you can view and purchase.
MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research.
National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America.
National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.
PubGene(TM) Gene Database and Tools, genomic-related publications database
SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.
SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site.
STATOO Datasets part 1 and part 2
UCR Time Series Data Mining Archive, offering datasets, papers, links, and code.
United States Census Bureau.

Source: http://www.kdnuggets.com/datasets/

Data Mining Process

CRISP-DM: Data Understanding

Collecting initial data

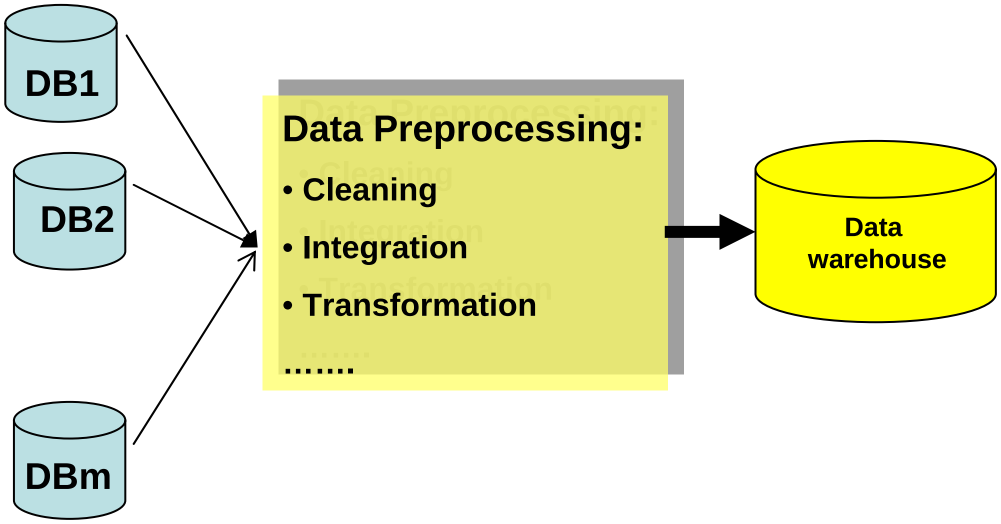
what are the needed data ?

- **where are the data ?**

- Flat Files
- Databases
- Heterogeneous Databases
- Connected autonomous databases
- Legacy Databases

  inherited from languages, platforms, and techniques earlier than current technology

- Data warehouse



Data Preprocessing:

- **Cleaning**
- **Integration**
- **Transformation**

…….

Data Warehouse (DWH)

Introduction

Development of DWH started in the beginning of 80s
DWH is an enterprise-wide *database* that serves as a databse for all kind of management support systems

Definition:

Several definition can be found for DW in the literature. One often used is due to W. H. Inmon:

„A Data Warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of Data in support of managements Decision support process.”

Technical potential benefits

- **Integrated database systems for management support**
- **Discharge operational data processing systems**
- **Quick queries and reports due to the integrated data**

Data Warehouse

Definition (continuous)

✓ **Subject-Oriented:**
Oriented to main subjects like Customer, Company, product, supplier,..
instead to concentrate on company's ongoing operations.

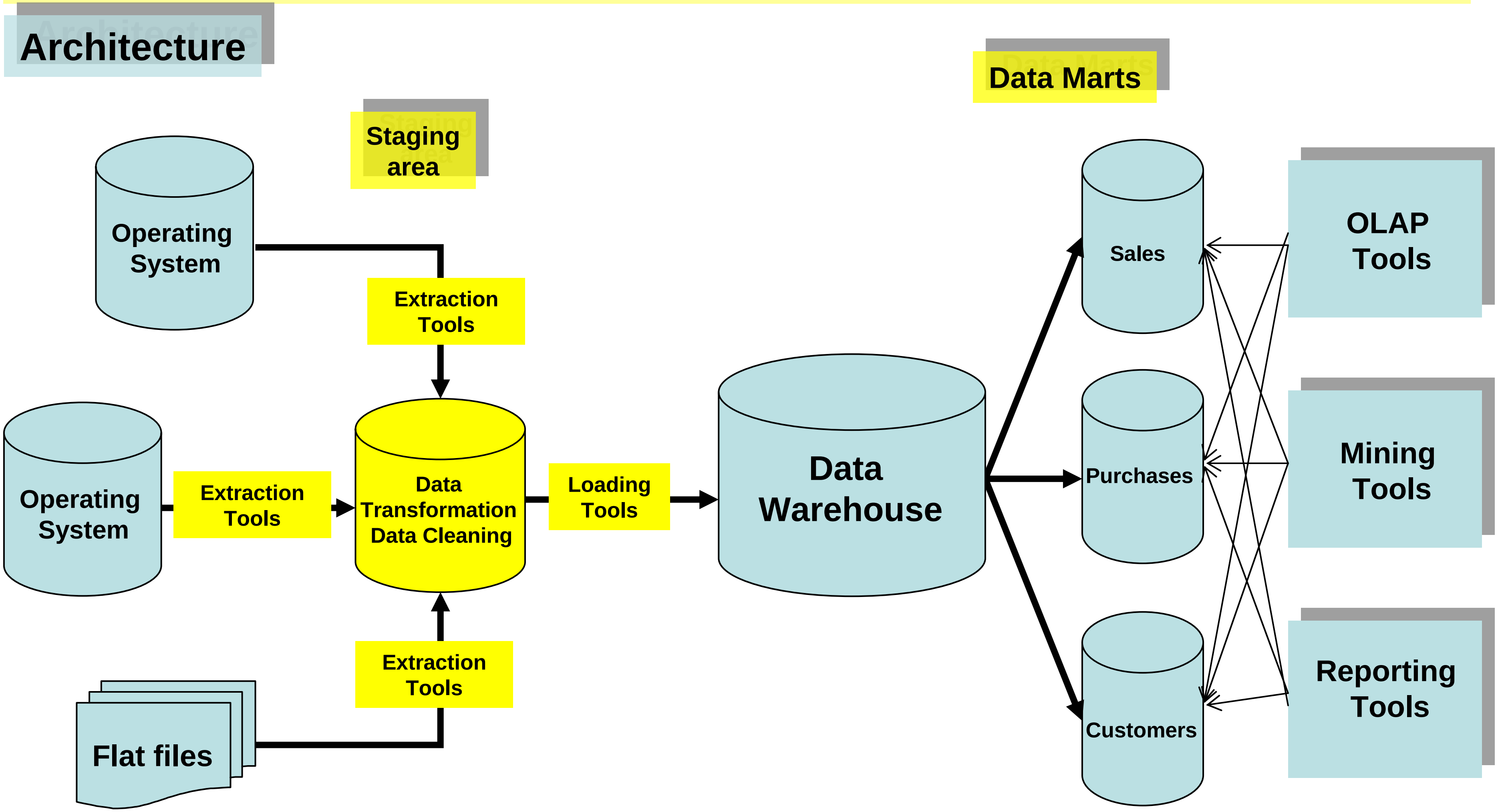
✓ **Integrated:**
Integrate data from different heterogeneous data sources
Relational databases flat files….
by application of data cleaning and data integration methods consistency in naming, encoding structure and attributes measures is fulfilled

✓ **Time-variant : Analysis on temporal changes and developments requires the long-term storage of data in DW; therefore “time” is a main dimension of DW**

✓ **Nonvolatile: The data once stored in a DW should not change ; otherwise it is not possible to perform a realistic data analysis**

Data Warehouse

Architecture

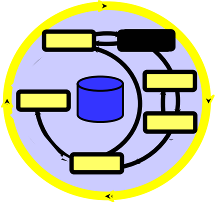


ETL: Extraction, Transformation, Loading

Data Mining Process

CRISP-DM: Data Understanding

Describing data



Some of data characterization measures

- number of observations
- number of attributes
- number of classes
- number of observations per class (balanced and unbalanced classes)
- …………

Data Characterizing Tool, DCT, was developed at DaimlerChrysler Data Mining Research Department in cooperation with the Universities of Karlsruhe and Leeds

Data Mining Process

CRISP-DM: Data Understanding

Describing data

- **Other measures to characterize data**

Initial Statistics	Example
Count	1000
Mean	1.407
Min	1
Max	4
Range	3
Variance	0.334
Standard Deviation	0.578
Standard Error of Mean	0.018

Data Mining Process

CRISP-DM: Data Understanding

Describing data

- **Other measures to characterize data**

Skewness
Is a measure that determines the degree of asymmetry of a distribution

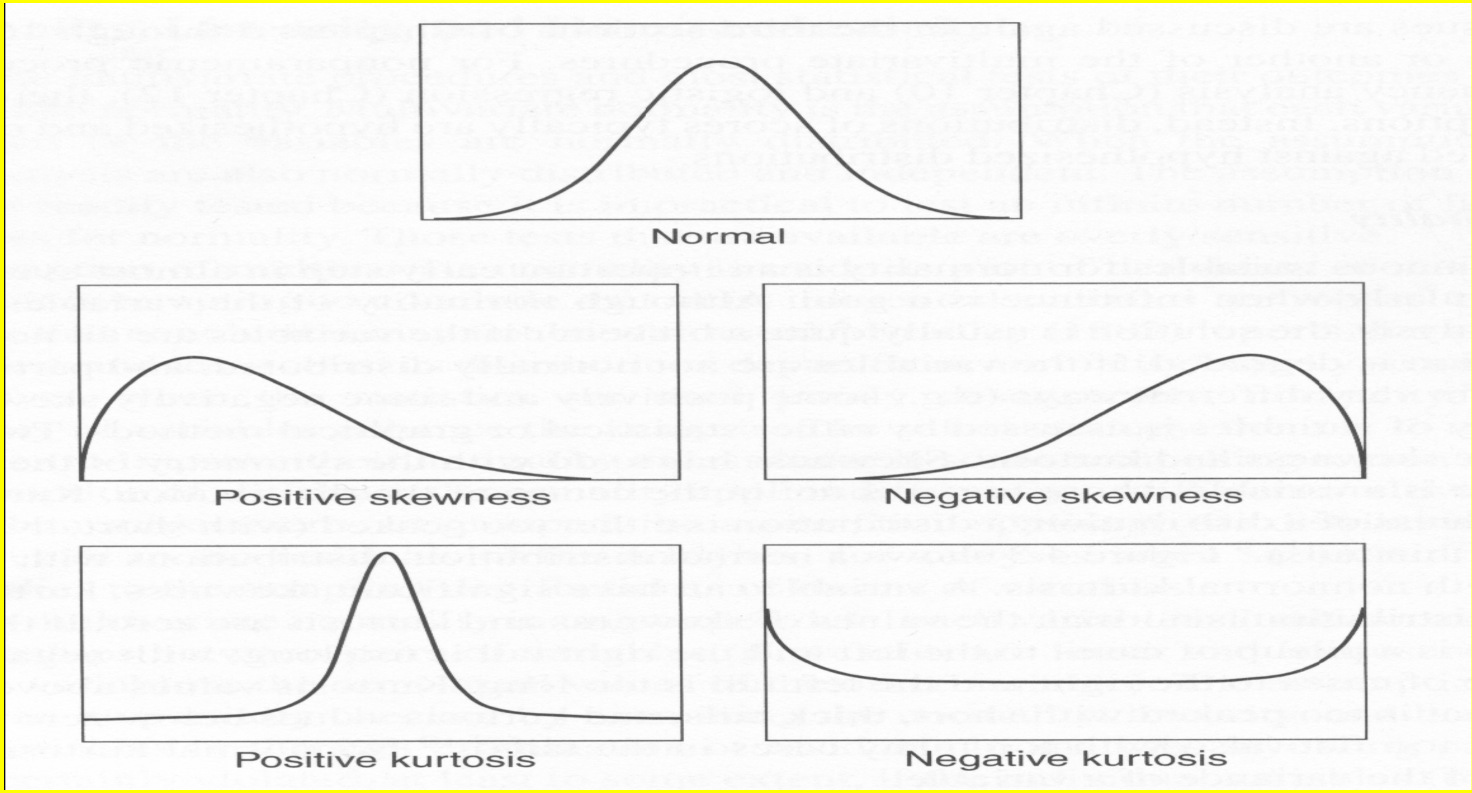
Kurtosis
Is a measure that determines the degree of peakedness or flatness of a distribution compared with normal distribution.

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Skewness and Kurtosis



http://www.csun.edu/~ata20315/psy524/docs/Psych%20524%20lecture%203%20DS.pdf

Data Mining Process

CRISP-DM: Data Understanding

Describing data

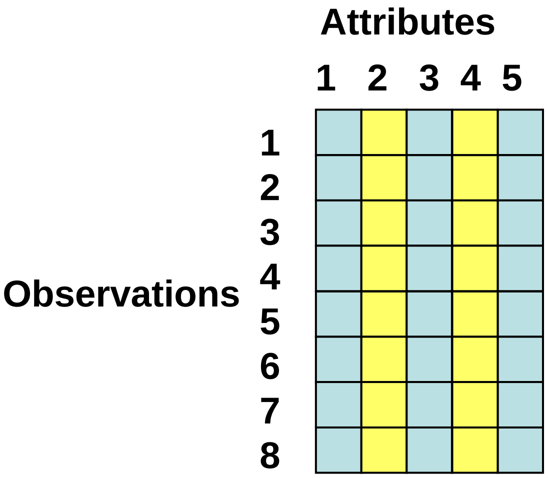
Dataset Structure

■ *Observations*

- A dataset can be considered as a collection of observations
- Other names for observation: case, data object, entity, event, instance, pattern, point, record, sample,..

■ *Attributes*

- Each observation is described by one or several attributes
- The attributes of an observation essentially define the properties of that observation
- Other names for attributes: feature, field, variable, ..



Data Mining Process

CRISP-DM: Data Understanding

Describing data

Dataset Structure

Example for a dataset: Annual Income

Attributes

Observations

	Income in three years ago	Education	Age	Income
1	24552	High School	32	27026
2	88282	BSc	52	93725
3	82902	PhD	41	82356
4	39838	High School	56	36828
5	53542	PhD	32	62542
6	63826	MS	28	64882
7	82783	MA	43	89025
8	72886	High School	33	74925
9	21383	BA	37	62572
10	63552	BA	41	66427
11	62522	High School	25	63552
12	65254	PhD	56	67252

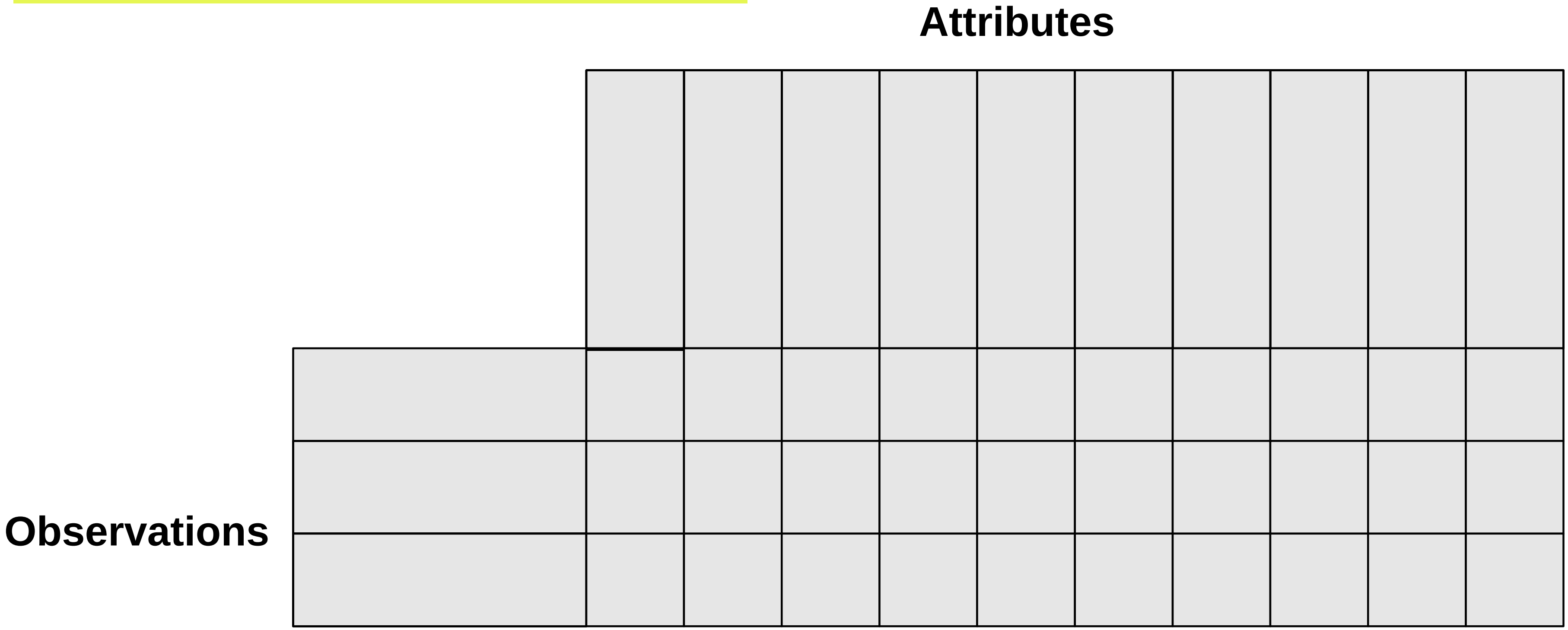
Data Mining Process

CRISP-DM: Data Understanding

Describing data

Dataset Structure

Example for representation of Document Data



Source: Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison wesley (May, 2005). Hardcover: 769 pages. ISBN: 0321321367

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Dataset Structure

Attribute Type: Attribute type is characterized by type of the values used to measure it

Level of Measurement: nominal, ordinal, interval, ratio

{nominal, ordinal} → categorical , qualitative
{interval, ratio} → continuous-valued , quantitative

The value of a *nominal-scaled* attribute does not have per se any evaluative distinction. It is just enough to distinguish one observation from another: A=B, or A ≠ B
Example: race, birthplace, religious, ID

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Dataset Structure

Attribute type

The value of a *ordinal-scaled* variable represents its rank order. It is enough to distinguish one observation from another: A=B, or A≠B and its rank: A>B or A<B.

Example (1): Mineral Hardness

Hardness	Mineral
1	Talc ($Mg_3Si_4O_{10}(OH)_2$)
2	Gypsum ($CaSO_4 \cdot 2H_2O$)
3	Calcite ($CaCO_3$)
4	Fluorite (CaF_2)
5	Apatite ($Ca_5(PO_4)_3(OH^-,Cl^-,F^-)$)
6	Orthoclase Feldspar ($KAlSi_3O_8$)
7	Quartz (SiO_2)
8	Topaz ($Al_2SiO_4(OH^-,F^-)_2$)
9	Corundum (Al_2O_3)
10	Diamond (C)

Source: http://en.wikipedia.org/wiki/Mohs_scale_of_mineral_hardness

Attribute type

Example 2: Ranking of German Soccer Teams (Bundesliga)

Rank	Club
1th	Bayern München
2nd	Hamburger SV
3rd	Bayer Leverkusen
4th	Werder Bremen
5th	FC Schalke 04
6th	VfB Stuttgart
7th	Eintracht Frankfurt
8th	VfL Wolfsburg
9th	Karlsruher SC
10th	Hannover 96

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Dataset Structure

Attribute type

Interval Attribute:

- Have all the features of ordinal attributes
- In addition equal differences between measurements can be viewed as equivalent intervals.
- **Differences** between arbitrary pairs of measurements can be meaningfully compared

It is meaningful: A=B, A>B (A<B), A-B

No absoult zero exists

Examples:

- Temperatur in Celsius or Fahrenheit (Equal differences represent equal differences in temperature, but 40 degrees is not twice as warm as 20 degrees).
- **Zero temperature does not mean no temperature**

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Attribute type

Ratio Attribute:

- Have all the features of interval attributes
- In addition ***ratios*** are meaningful

absoult zero exists

Examples:

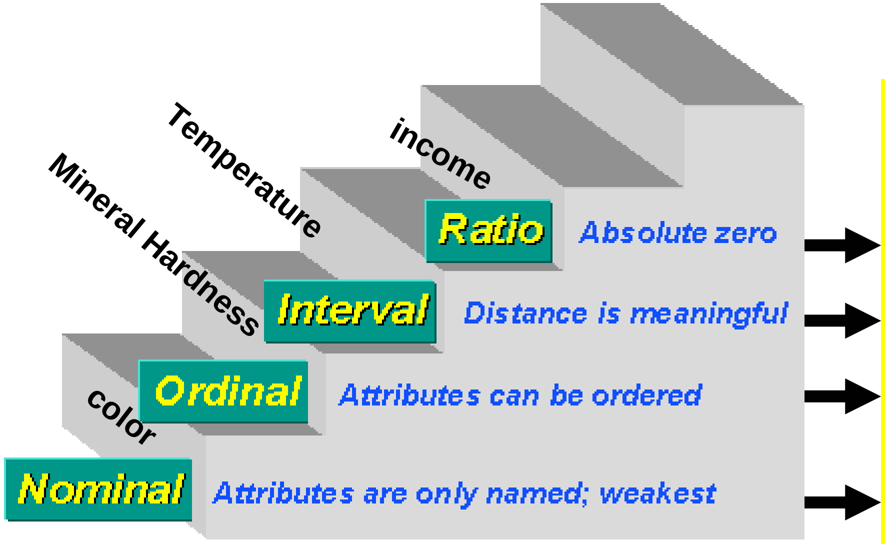
- Age, income , sales volume
- Zero Age is meaningful: absence of age or birth.
- A 60-year old person is twice as old as a 30-year old one
- Zero income means no income

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Attribute type



income

Temperature

Mineral Hardness

color

Ratio — Absolute zero

Interval — Distance is meaningful

Ordinal — Attributes can be ordered

Nominal — Attributes are only named; weakest

Source: http://www.socialresearchmethods.net/kb/measlevl.php

Meanigful are:

Multiplication, devision (*, /), (-), (> , <)(= ≠)

Difference (-), (> , <), (= ≠)

Greater, les (> , <), (= ≠)

Equality, inequality (= ≠)

Data Mining Process

CRISP-DM: Data Understanding

Describing data

Attribute type : another classification

- **Discrete Attributes**
  - Have a finite or countable infinite set of values
  - Examples: number of children , counts
  - Often represented as integer variables
  - Special case of discrete attributes : binary attributes

- **Continuous Attributes**
  - Have real numbers as attribute values
  - Examples: Income, sales , weight

Data Mining Process

CRISP-DM: Data Understanding

Data Type

- Cross-Section data
- Time Series data
- Panel data
- Sequences
  - Postman Routes
  - Web Click Streams

- Data Streams
  - Infinite volumes
  - Dynamically Changing
  - Real time processing
- Spatial data
- Spatiotemporal data
- Transaction data
- Text data
- web data
- Multimedia data

Data Mining Process

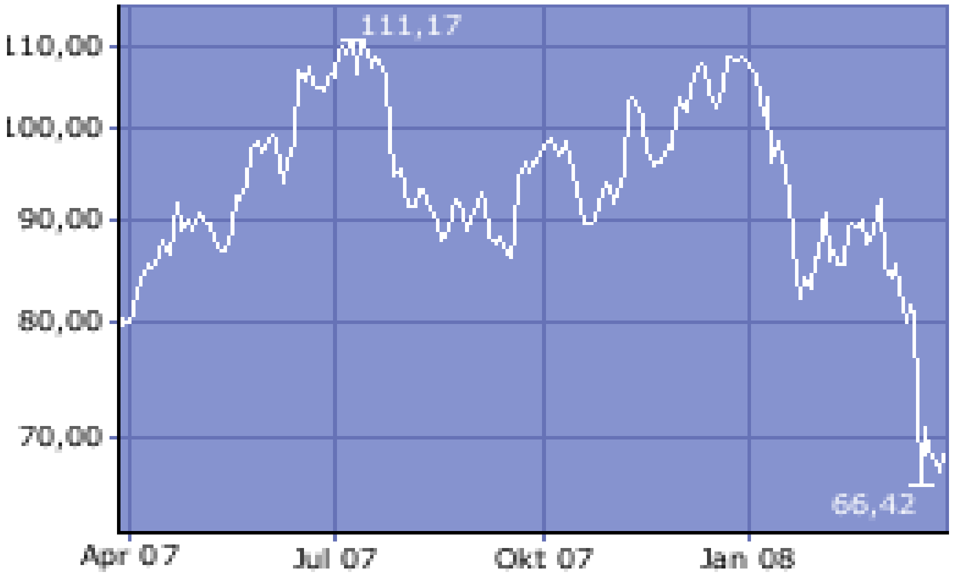
CRISP-DM: Data Understanding

Data Type

Example for cross-section data: Annual Income

	Income in three years ago	Education	Age	Income
1	24552	High School	32	27026
2	88282	BSc	52	93725
3	82902	PhD	41	82356
4	39838	High School	56	36828
5	53542	PhD	32	62542
6	63826	MS	28	64882
7	82783	MA	43	89025
8	72886	High School	33	74925
9	21383	BA	37	62572
10	63552	BA	41	66427
11	62522	High School	25	63552
12	65254	PhD	56	67252

Example for time-series data: Siemens share



Data Mining Process

CRISP-DM: Data Understanding

Data Type

Example for the source of panel-data

SOEP

The German Socio-Economic Panel Study

A Representative Longitudinal Study of Private Households in the Entire Federal Republic of Germany

- The SOEP is a wide-ranging representative longitudinal study of private households.
- It provides information on all household members, consisting of Germans living in the Old and New German States, Foreigners, and recent Immigrants to Germany.
- **The Panel was started in 1984. In 2006, there were nearly 11,000 households, and more than 20,000 persons sampled.**
- Some of the many topics include household composition, occupational biographies, employment, earnings, health and satisfaction indicators.
- The data are available to researchers in Germany and abroad in SPSS, SAS, Stata, and ASCII format for immediate use. Extensive documentation in English and German is available online.

Source: http://www.diw.de/deutsch/soep/29012.html

Data Mining Process

CRISP-DM: Data Understanding

Data Type

Spatial Data

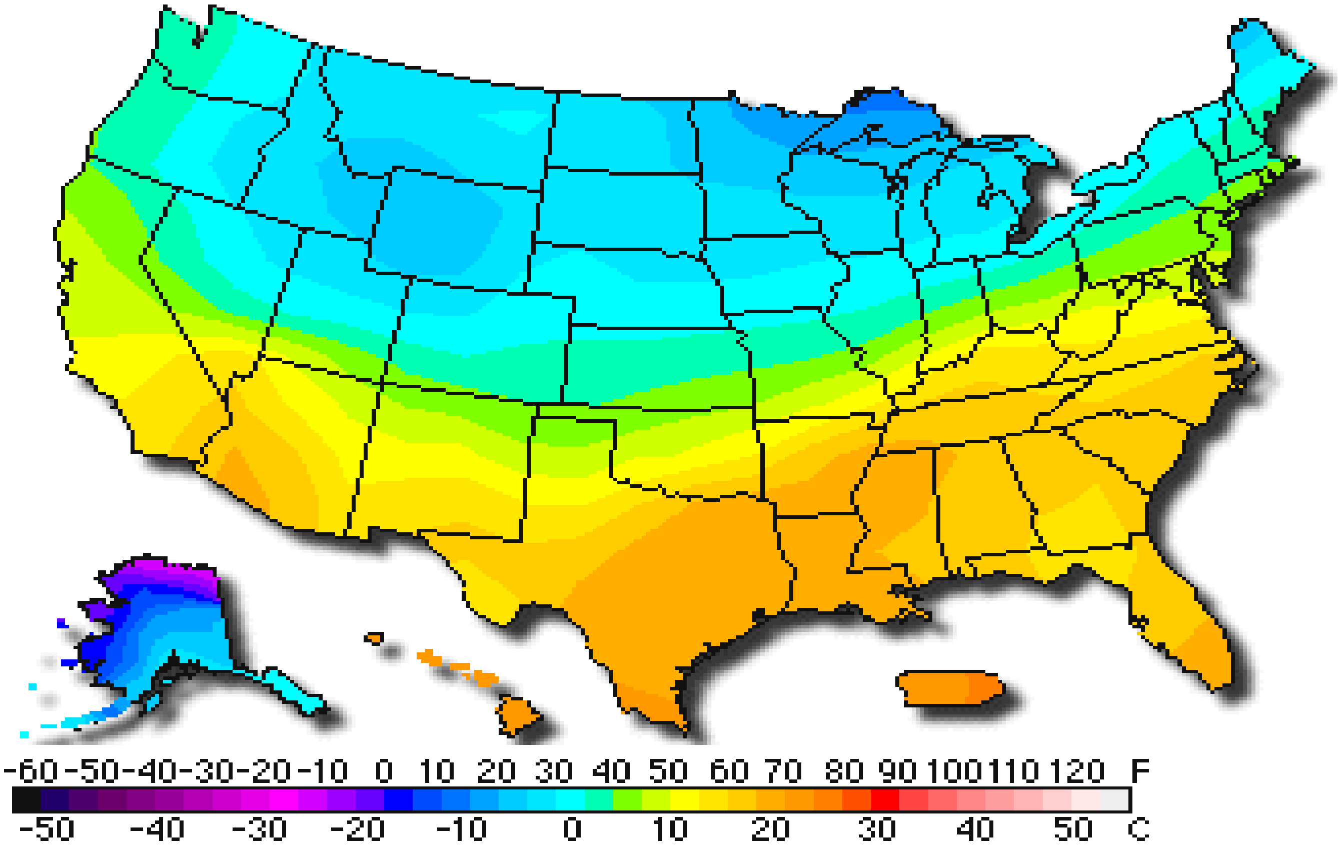
- known also as ***geospatial data or geographic*** information
- describes the ***geographic location of features and boundaries on Earth***
- usually stored as ***coordinates and topology***
- can be mapped represented as ***2D or 3D images***
- can be often accessed or analyzed through ***GIS*** (Geographic Information systems)

Data Mining Process

CRISP-DM: Data Understanding

Data Type

Example for Spatial Data: US Temperature Map


-60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100 110 120 F
-50 -40 -30 -20 -10 0 10 20 30 40 50 C

Source: http://www.wunderground.com/US/Region/US/Temperature.html

Letzter Stand 05:00 AM GMT am 28. März 2008

Data Mining Process

CRISP-DM: Data Understanding

Data Type

Spatiotemporal Data

- Spatiotemporal data describes the development and changes of Spatial data over the time

Examples:
GPS-Data,
Satallite images
Traffic Data
Telecommunication Data
….

Data Mining Process

CRISP-DM: Data Understanding

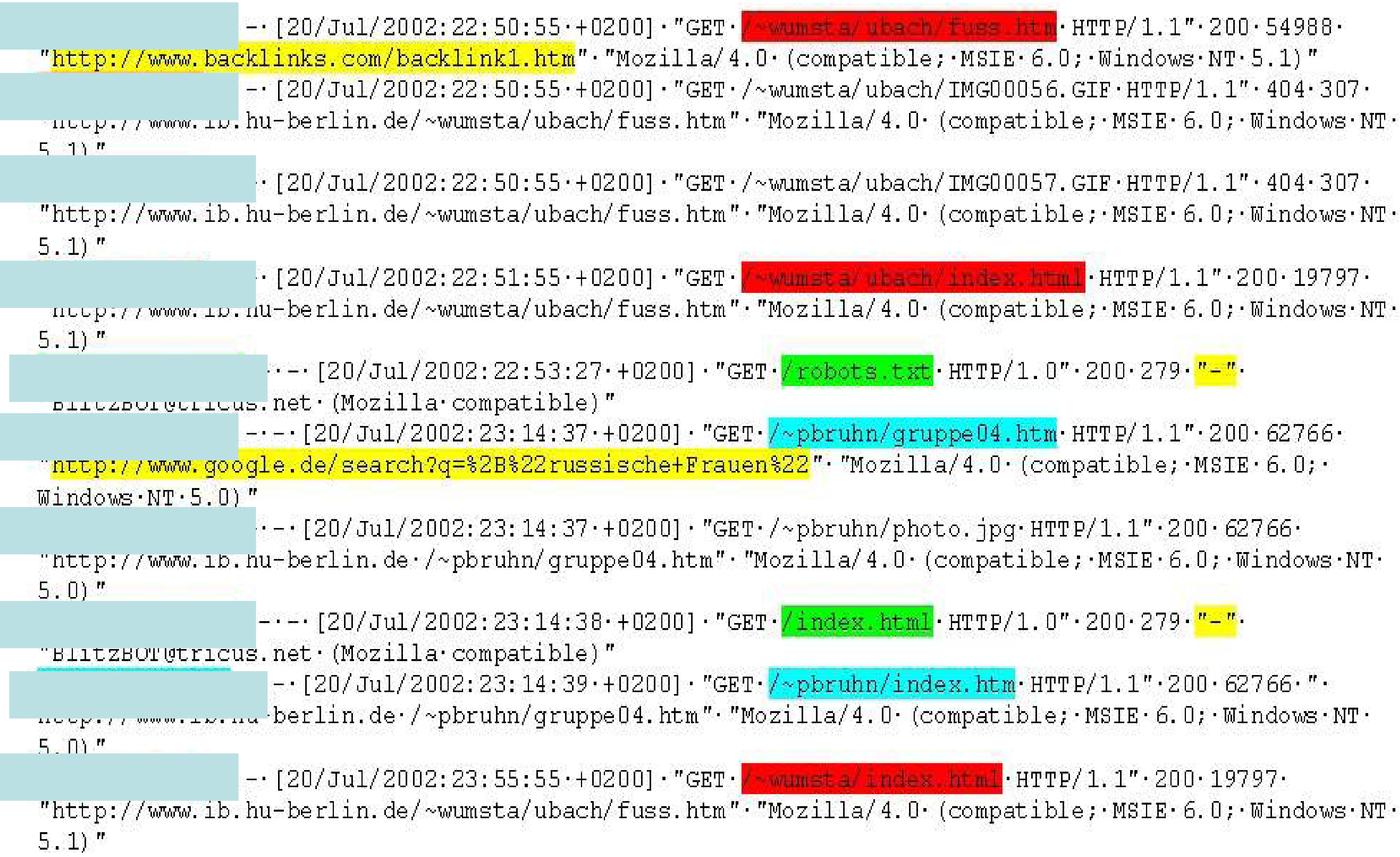
Data Type

Example for the source of spatial data

- USGS : U.S.Geological Survey
- Geospatial Data One-Stop
- Geodata Explorer
- National Mapping Information
  - Products, Information, and Services
  - Data Standards
- FGDC : Federal Geographic Data Committee
  - Manual of Federal Geographic Data Product
- SDTS : Spatial Data Transfer Standard
- NGDC : National Geospatial Data Clearinghouse
  - Popular Digital Geospatial Data Set Collections
  - Digital Geospatial Data Set by Theme
- GLIS : Global Land Infomation System
  - 1:100,000-Scale Digital Line Graphs
  - 1:200,000-Scale Digital Line Graphs
  - 30 Arc-Sec. DCW Digital Elevation Models
  - 5 Minute Gridded Earth Topography Data
  - Conterminous U.S. AVHRR
  - MultiSpectral Scanner Landsat Data
  - Space Shuttle Earth Observation Program
  - Thematic Mapper Landsat Data
  - USGS Land Use and Land Cover Data
- EDC : EROS Data Center
  - Earth Explorer
  - Seamless Data Distribution Center">
- Publications and Data Products
  - Cartographic Data
  - Geologic Data
  - Water Resources Data
- U.S. GeoData FTP file access - DEM, DLG, LULC
- CENSUS BUREAU
- TIGER Database
- 2000 U.S. Census Data
- 1990 U.S Census Data
- 1980 Census Data (SEEDIS)
- Data Maps
- TIGER Map Services
- Census State Data Centers
- NOAA : National Oceanic and Atmospheric Administration
- NOAA Data Set Catalog
- National Geophysical Data Center (NGDC)
  - World Data Center System
- National Climatic Data Center (NCDC)
- National Hurricane Center
- National Oceanographic Data Center (NODC)
- Environmental Research Laboratories

http://ncl.sbs.ohio-state.edu/5_sdata.html

Example of Web Data: A log file sample



```
[redacted] -·[20/Jul/2002:22:50:55·+0200]·"GET·/~wumsta/ubach/fuss.htm·HTTP/1.1"·200·54988·
"http://www.backlinks.com/backlink1.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·5.1)"
[redacted] -·[20/Jul/2002:22:50:55·+0200]·"GET·/~wumsta/ubach/IMG00056.GIF·HTTP/1.1"·404·307·
hu-berlin.de/~wumsta/ubach/fuss.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.1)"
[redacted]·[20/Jul/2002:22:50:55·+0200]·"GET·/~wumsta/ubach/IMG00057.GIF·HTTP/1.1"·404·307·
"http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.1)"
[redacted]·[20/Jul/2002:22:51:55·+0200]·"GET·/~wumsta/ubach/index.html·HTTP/1.1"·200·19797·
hu-berlin.de/~wumsta/ubach/fuss.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.1)"
[redacted]·-·[20/Jul/2002:22:53:27·+0200]·"GET·/robots.txt·HTTP/1.0"·200·279·"-"·
.net·(Mozilla·compatible)"
[redacted] -·-·[20/Jul/2002:23:14:37·+0200]·"GET·/~pbruhn/gruppe04.htm·HTTP/1.1"·200·62766·
google.de/search?q=%2B%22russische+Frauen%22"·"Mozilla/4.0·(compatible;·MSIE·6.0;·
Windows·NT·5.0)"
[redacted]·-·[20/Jul/2002:23:14:37·+0200]·"GET·/~pbruhn/photo.jpg·HTTP/1.1"·200·62766·
hu-berlin.de·/~pbruhn/gruppe04.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.0)"
[redacted]-·-·[20/Jul/2002:23:14:38·+0200]·"GET·/index.html·HTTP/1.0"·200·279·"-"·
.net·(Mozilla·compatible)"
[redacted] -·[20/Jul/2002:23:14:39·+0200]·"GET·/~pbruhn/index.htm·HTTP/1.1"·200·62766·"·
berlin.de·/~pbruhn/gruppe04.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.0)"
[redacted] -·[20/Jul/2002:23:55:55·+0200]·"GET·/~wumsta/index.html·HTTP/1.1"·200·19797·
"http://www.ib.hu-berlin.de/~wumsta/ubach/fuss.htm"·"Mozilla/4.0·(compatible;·MSIE·6.0;·Windows·NT·
5.1)"
```

Source: http://eprints.rclis.org/archive/00004887/01/kx05-poster_mayr.pdf

Example of Web Data: A log file sample

```
fcrawler.looksmart.com - - [26/Apr/2000:00:00:12 -0400] "GET /contacts.html HTTP/1.0" 200 4595 "-"
"FAST-WebCrawler/2.1-pre2 (ashen@looksmart.net)"

fcrawler.looksmart.com - - [26/Apr/2000:00:17:19 -0400] "GET /news/news.html HTTP/1.0" 200 16716 "-"
"FAST-WebCrawler/2.1-pre2 (ashen@looksmart.net)"

ppp931.on.bellglobal.com - - [26/Apr/2000:00:16:12 -0400] "GET /download/windows/asctab31.zip HTTP/1.0" 200 1540096
"http://www.htmlgoodies.com/downloads/freeware/webdevelopment/15.html" "Mozilla/4.7 [en]C-SYMPA (Win95; U)"


123.123.123.123 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/wpaper.gif HTTP/1.0" 200 6248 "http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"

123.123.123.123 - - [26/Apr/2000:00:23:47 -0400] "GET /asctortf/ HTTP/1.0" 200 8130
"http://search.netscape.com/Computers/Data_Formats/Document/Text/RTF" "Mozilla/4.05 (Macintosh; I; PPC)"

123.123.123.123 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/5star2000.gif HTTP/1.0" 200 4005 "http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"

123.123.123.123 - - [26/Apr/2000:00:23:50 -0400] "GET /pics/5star.gif HTTP/1.0" 200 1031 "http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"

123.123.123.123 - - [26/Apr/2000:00:23:51 -0400] "GET /pics/a2hlogo.jpg HTTP/1.0" 200 4282 "http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"

123.123.123.123 - - [26/Apr/2000:00:23:51 -0400] "GET /cgi-bin/newcount?jafsof3&width=4&font=digital&noshow HTTP/1.0" 200 36
"http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
```

Source: http://www.jafsoft.com/searchengines/log_sample.html

Data Mining Process

CRISP-DM: Data Understanding

Data exploration

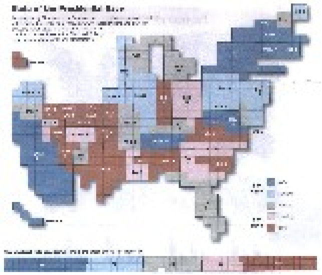
Data exploration

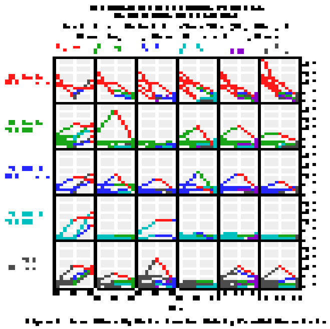
May be useful

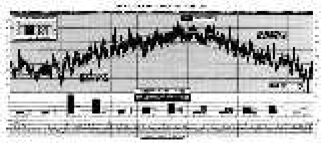
- to get the first insights into the structure of data
- to identify noisy data or outliers

Data exploration Tools

- **Using descriptive data summarization**
- **Using Visualization**







Source: http://www.math.yorku.ca/SCS/Gallery/

Data Mining Process

CRISP-DM: Data Understanding

Data exploration

Tools for descriptive data summarization

- **Measures of Location (Central Tendency):**

summarize an attribute by a "typical" value
common measures: *mean, median , mode*

- **Measures of Spread (Dispersion):**

summarize how much the observations of an attribute differ from each other
common measures of spread: *range, variance, average absolute deviation*

Data Mining Process

CRISP-DM: Data Understanding

Data exploration

Measures of Location:

Mean (Average):

$$\overline{X} = 1/n \sum_{i=1}^{n} X_i$$

Median (Middel Number):

(The observations should be arranged in ascending order)

n odd → $X_{Med} = X_{(n + 1)/ 2}$

n even → $X_{Med} = 1/ 2 \; (\; X_{n / 2} + X_{n / 2 \; +1} \;)$

Mode (Modal Number) :
The most frequently occurring attribute value

Warning: If there is in observation an outlier, the ***mean*** understates (overstates) the true value. In this case the ***median*** is a better measure

Data Mining Process

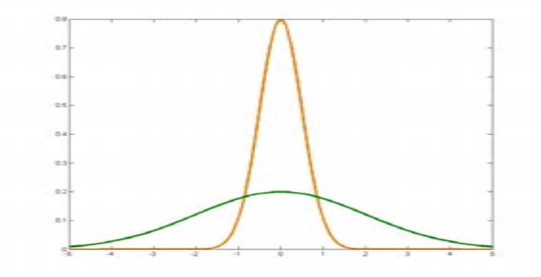
CRISP-DM: Data Understanding

Data exploration

Measures of Spread

Unbiased Sample Variance:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2.$$



Same mean different variance

Standard Deviation:
is the positive square root of the variance

Average Absolute Deviation

$$AA = 1/n \sum_{i=1}^{n} \left| X_i - m(X) \right|$$

m(x): Mean, Median or Mode

Range:

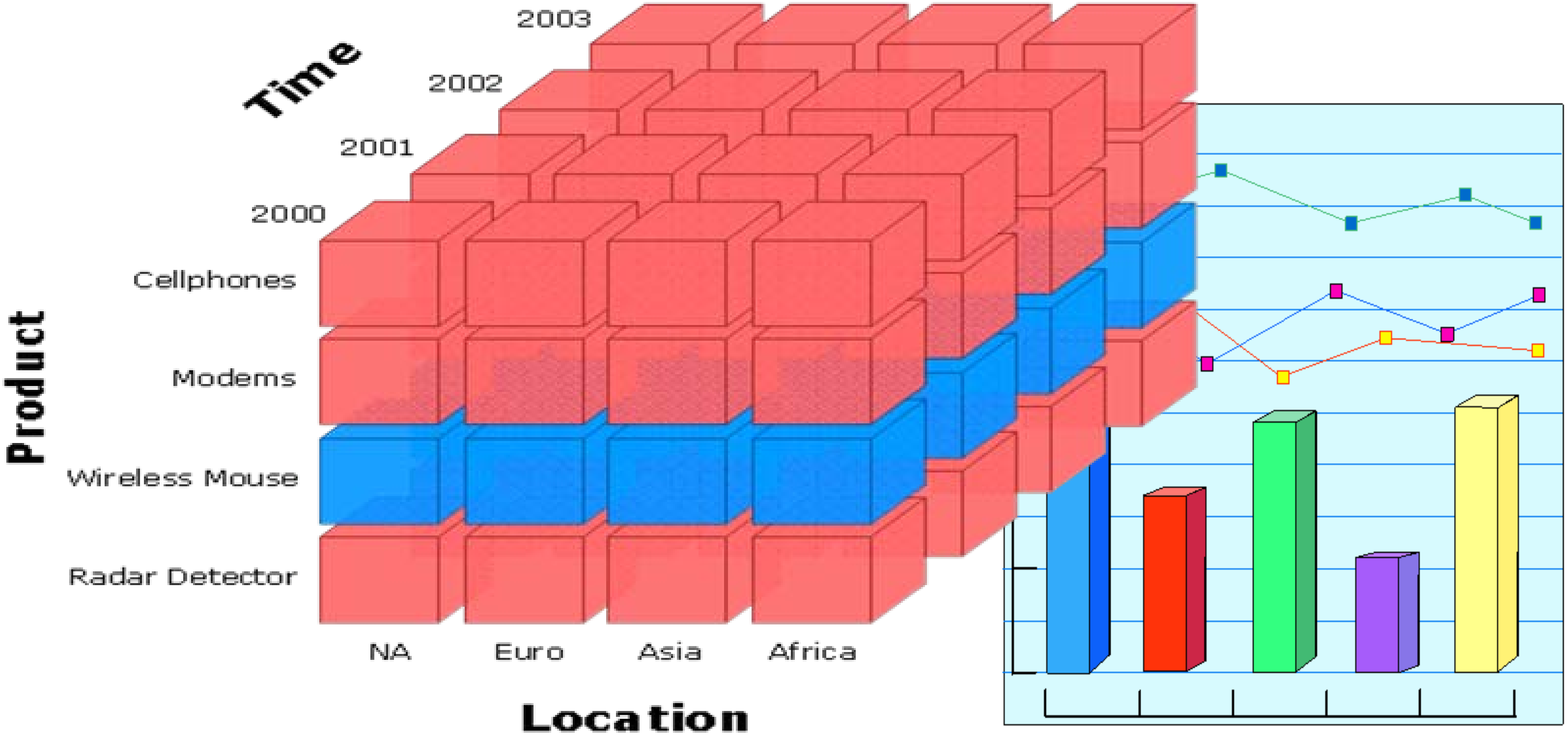
$$R = X_{max} - X_{min}$$

Data Mining Process

CRISP-DM: Data Understanding

Data exploration

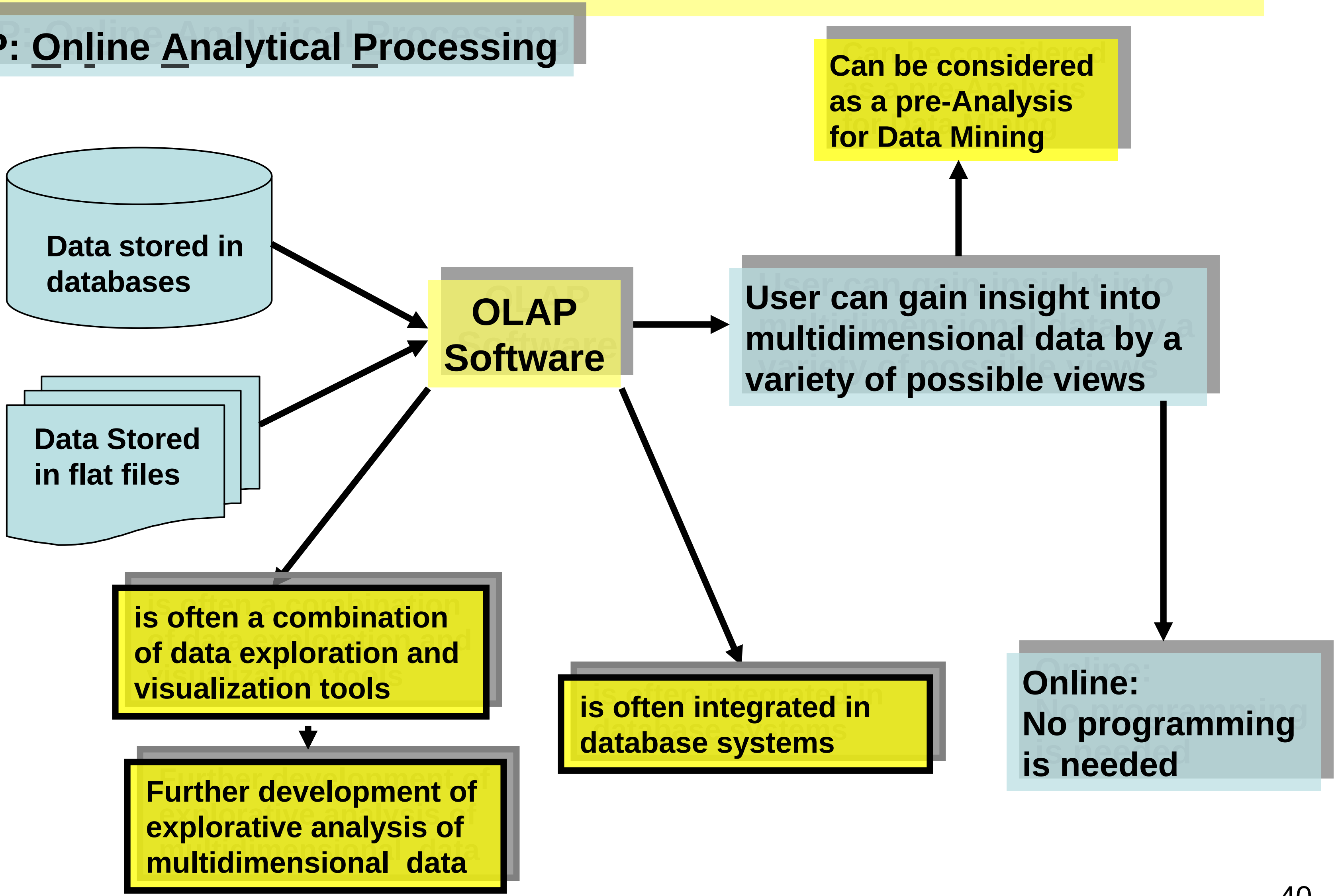
OLAP: Online Analytical Processing



Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

OLAP: Online Analytical Processing



Data stored in databases

Data Stored in flat files

OLAP Software

User can gain insight into multidimensional data by a variety of possible views

Can be considered as a pre-Analysis for Data Mining

Online: No programming is needed

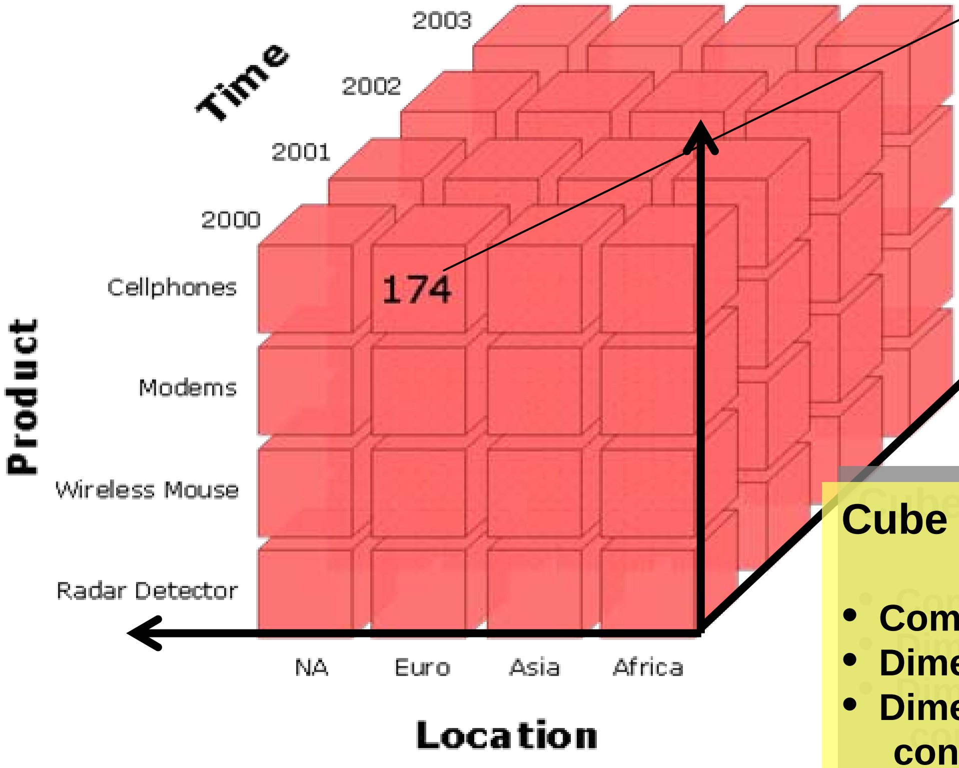
is often a combination of data exploration and visualization tools

Further development of explorative analysis of multidimensional data

is often integrated in database systems

OLAP

OLAP-CUBE:
Analysis in OLAP is done by using OLAP-CUBES



CUBE Measure: content of a cell can be

- **a Number (number of cell phones produced in Europe in 2000)**
- **an amount (total sales in $ of cell phones produced in Europe in 2000)**
- **Sometimes called “target quantity”**

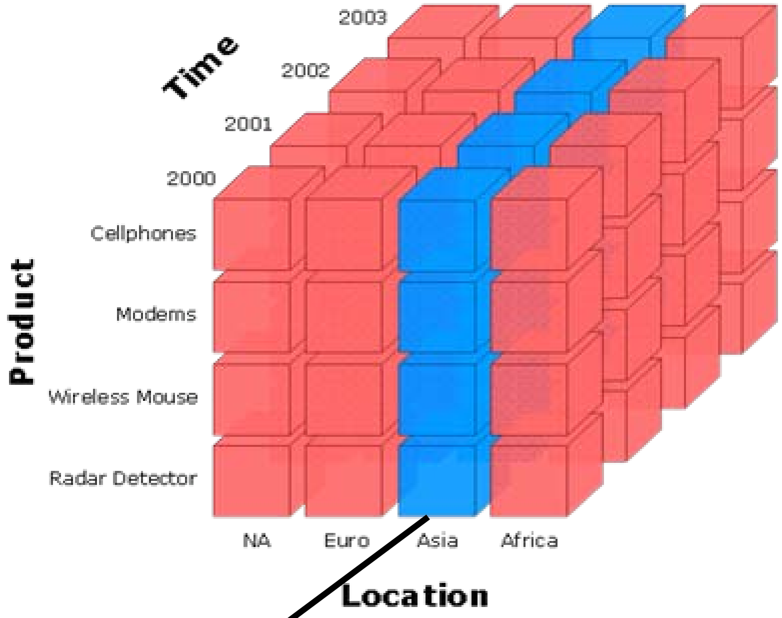
Cube Dimensions:

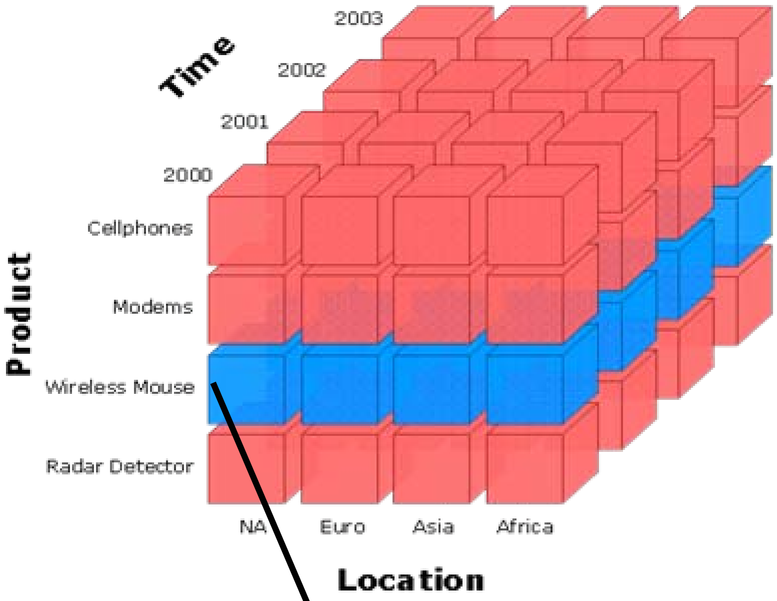
- **Comparable with attributes in Data Mining**
- **Dimensions have nominal values (called categories)**
- **Dimension with continuous categories have to be converted to nominal categories**
- **In the reality, the number of Dimensions is often more than 3 (Hypercube)**

Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

Slicing: Selecting a value of a dimensional and consider all the cells belong to other dimensions





Slice Asia

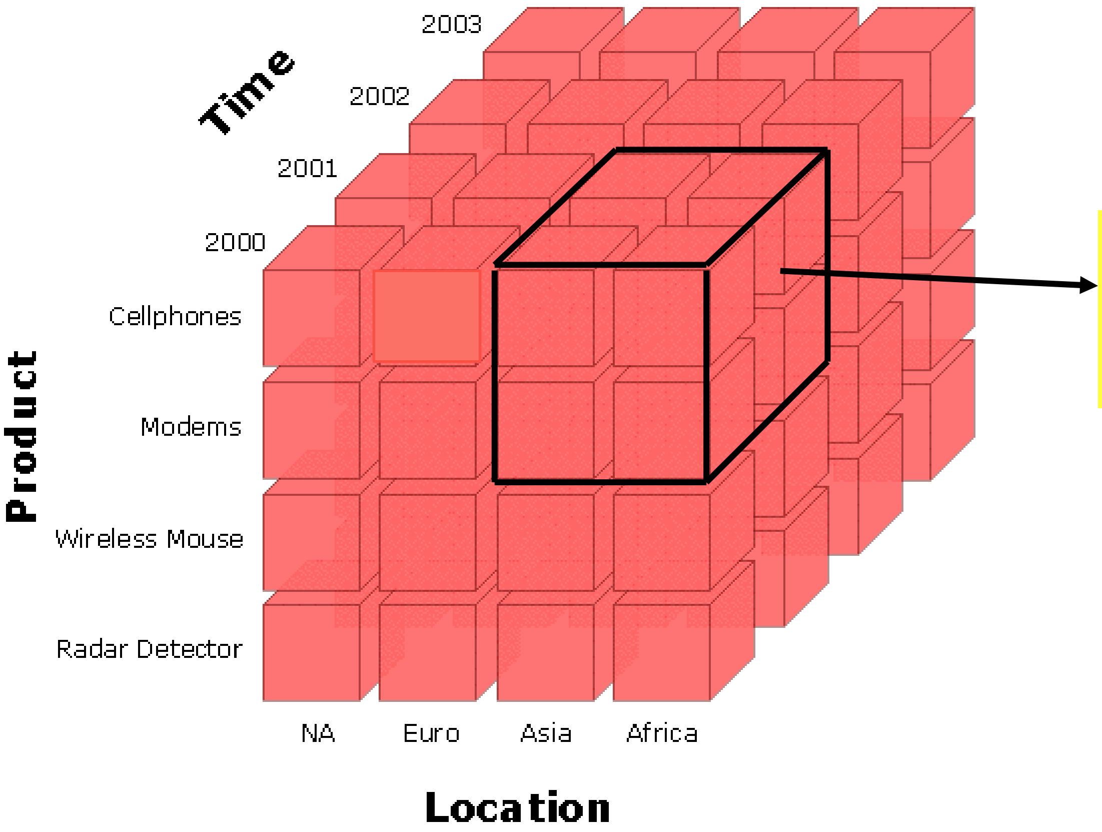
Slice Wireless Mouse

Consist of 16 cells and 16 measures

Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

Dicing: selecting a subset of a cube on two or more dimensions

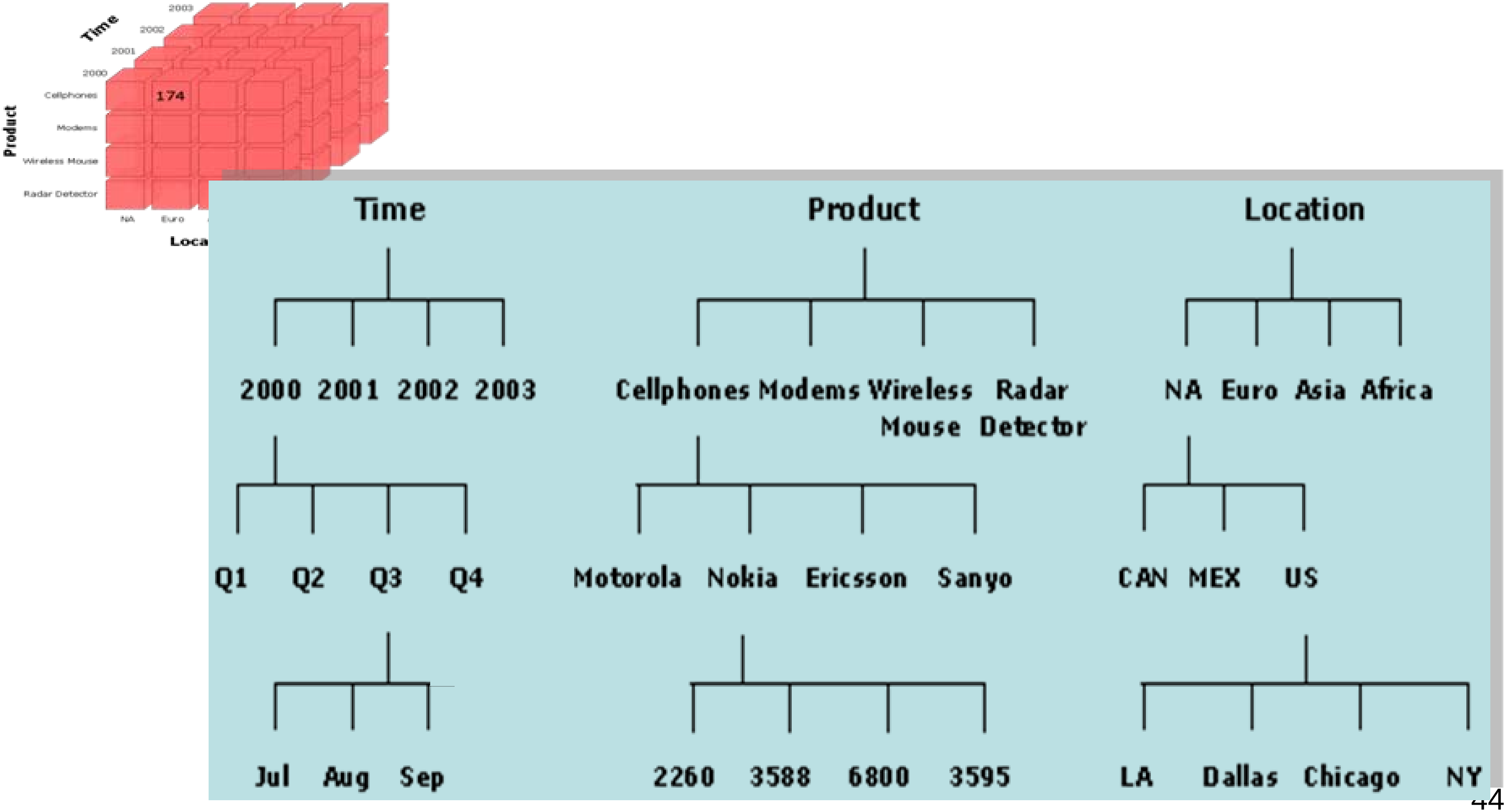


Dice operation involving 3 Dimensions:
(Location: Asia, Africa),
(Product: Modems, Cell phones)
and (Time: 2000, 2001)

Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

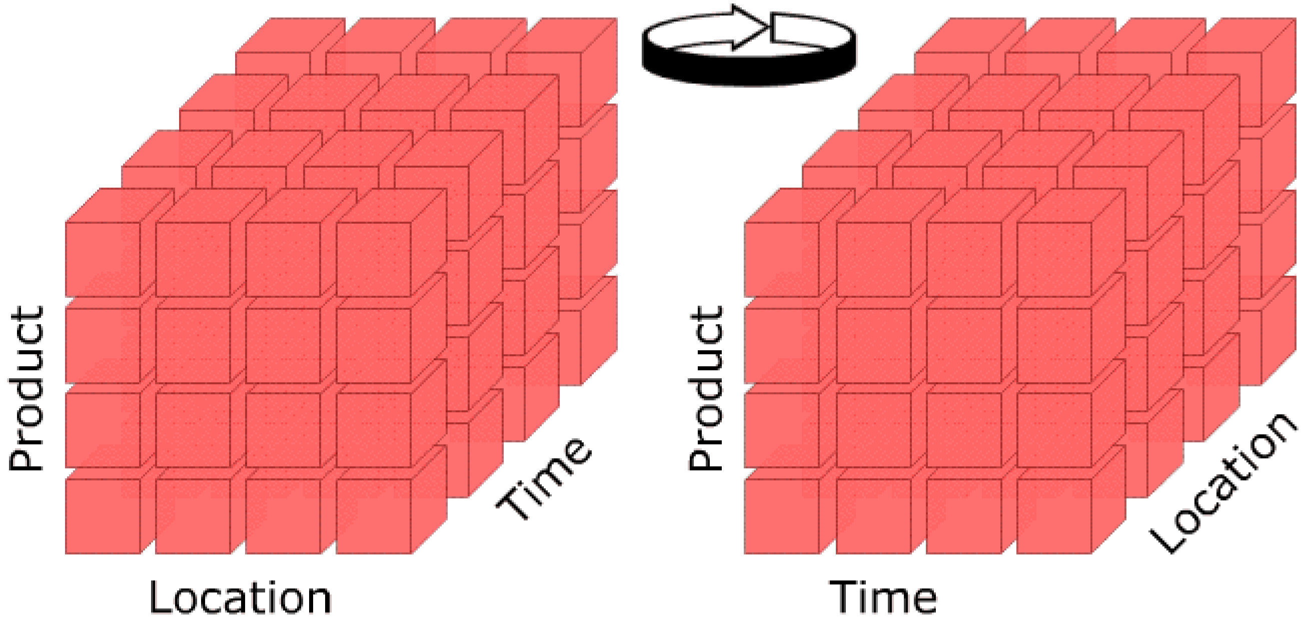
More about Dimensions: Each category of a Dimension may have subcategories



Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

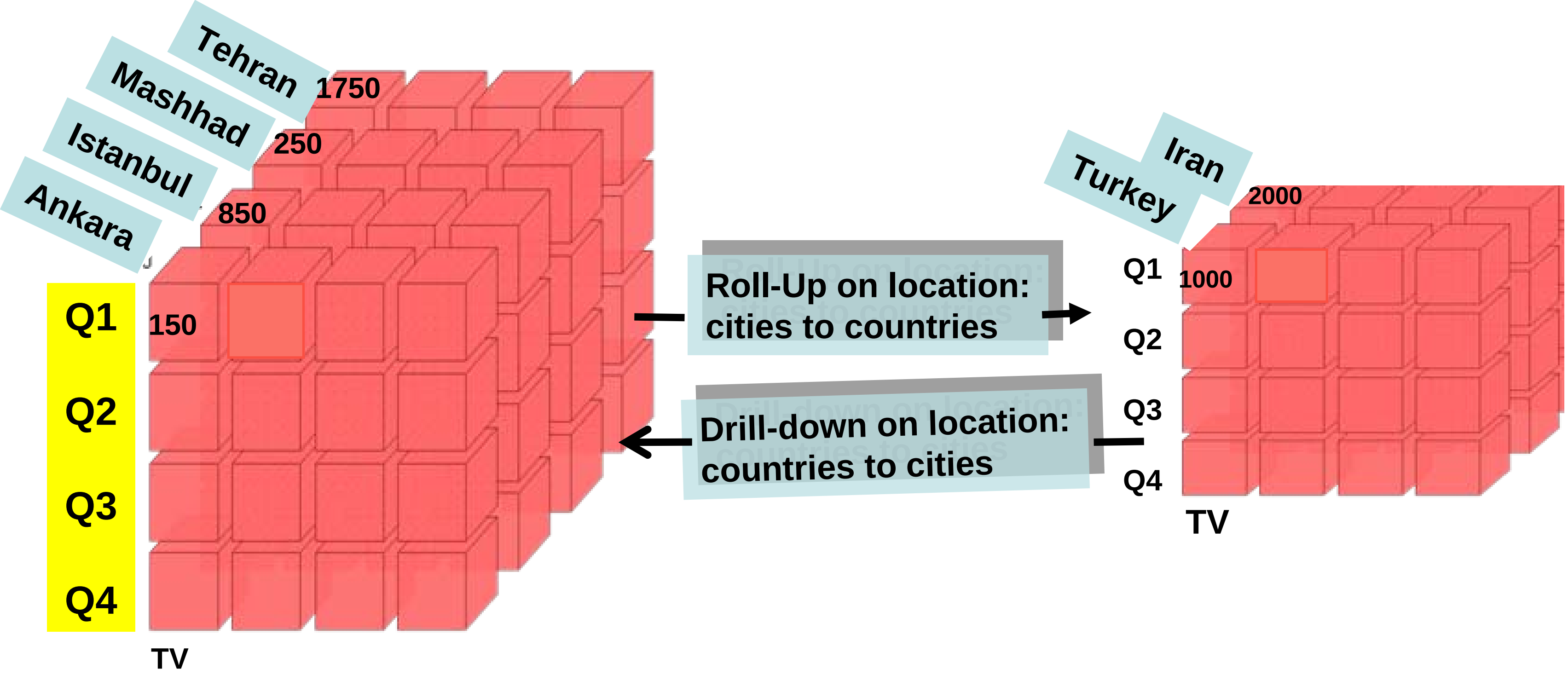
Rotating (Pivoting): Rotating the axes in order to generate an alternative presentation of the data



Source of the cube fig. in this and the following pages: http://training.inet.com/OLAP/Cubes.htm

OLAP

Roll-up : Aggregation by climbing up a category hierarchy



Roll-Up on location: cities to countries

Drill-down on location: countries to cities

Drill-down : Going to more detailed data by stepping down a category hierarchy

Source of the cube : http://training.inet.com/OLAP/Cubes.htm

OLAP

Other capabilities and functionalities

- Calculation Engine for
  - Ratios
  - Mean
  - Variance
  - …..
- Supporting functional modeling for:
  - Forecasting
  - Trend analysis
  - Other statistical computations and tests

OLAP

Other systems

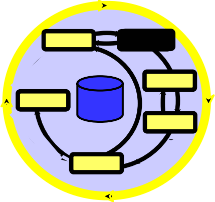
- ROLAP: Relational OLAP
  - OLAP software based on relational data bases
  - They have greater scalability than MOLAP but less efficiency
- MOLAP: Multidimensional OLAP
  - OLAP software based on multidimensional data models
  - Mapping multidimensional views directly to data cube array structures

- HOLAP: Hybrid OLAP
  - Such systems combine ROLAP and MOLAP technologies
  - They benefit from the high scalability of ROLAP systems and faster computation of MOLAP systems
- OLAM: Online Analytical Mining
  - Integration of OLAP with Data Mining
    - Related to the concept “in-database Mining”

Data Mining Process

CRISP-DM: Data Understanding

Verifying data quality



The real world data are often “dirty”, data “Cleaning” is needed

• Are data accurate ?

- noisy data

• Are data complete ?

- missing values

•Are data consistent ?

- Coding Errors

Short review of business and data understanding

- **Collect initial data**
  - Can the data be accessed effectively and efficiently ?
  - Is there any restriction in collecting the data ?
  - what are the needed data ? where are the data ?
  - Examples of data sources
  - Data warehouse

- **Describe data**
  - Some of data characterization measures
  - Data Structure

Observation, attribute type (nominal, ordinal, interval, ratio, qualitative, quantitative, discrete)
Data Type: Cross-section data, time series data, panel data, spatial data…

- **Explore data**

- Data exploration Tools
Using descriptive data summarization (mean, median, modus, variance,…)
- Using Visualization
- OLAP

- **Verify data quality**

- Are data accurate ?
- Are data complete ?
- Are data consistent ?