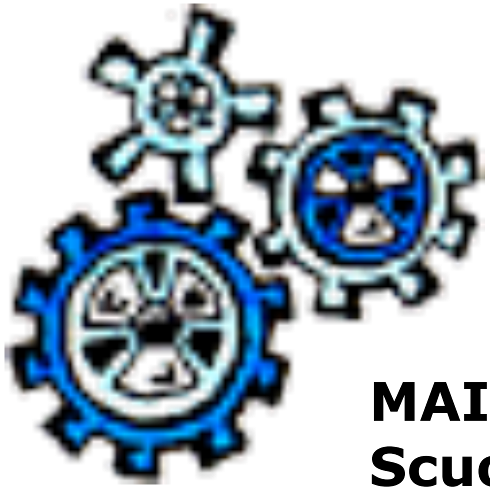
# Data Mining

## Knowledge Discovery in Databases

**Fosca Giannotti and Dino Pedreschi**

**Pisa KDD Lab, ISTI-CNR & Univ. Pisa**

**http://www-kdd.isti.cnr.it/**



**MAINS - Master in Management dell'Innovazione**
**Scuola Superiore Sant'Anna**

# Seminar 1 outline

- Motivations
- Application Areas
- KDD Decisional Context
- KDD Process
- Architecture of a KDD system
- The KDD steps in short
- Some examples in short

# Atherosclerosis prevention study

2nd Department of Medicine,
1st Faculty of Medicine of Charles University and Charles University Hospital,
U nemocnice 2, Prague 2
(head. Prof. M. Aschermann, MD, SDr, FESC)

# Atherosclerosis prevention study:

- **The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.**

- **Used for Discovery Challenge at PKDD 00-02-03-04**

# Atherosclerosis prevention study:

- **Study on 1400 middle-aged men at Czech hospitals**
  - **Measurements concern development of cardiovascular disease and other health data in a series of exams**
- **The aim of this analysis is to look for associations between medical characteristics of patients and death causes.**
- **Four tables**
  - **Entry and subsequent exams, questionnaire responses, deaths**

# The input data

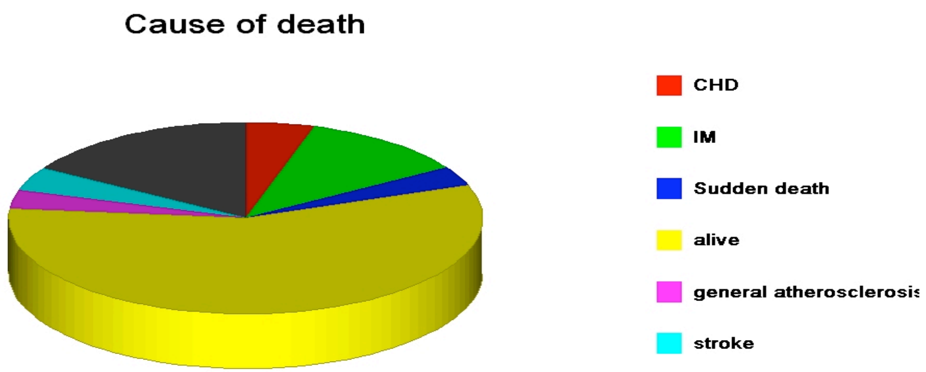
| Data from Entry and Exams | | |
|---|---|---|
| **General characteristics** | **Examinations** | **habits** |
| Marital status | Chest pain | Alcohol |
| Transport to a job | Breathlessness | Liquors |
| Physical activity in a job | Cholesterol | Beer 10 |
| Activity after a job | Urine | Beer 12 |
| Education | Subscapular | Wine |
| Responsibility | Triceps | Smoking |
| Age | | Former smoker |
| Weight | | Duration of smoking |
| Height | | Tea |
| | | Sugar |
| | | Coffee |

# The input data

| DEATH CAUSE | PATIENTS | % |
|---|---|---|
| myocardial infarction | 80 | 20.6 |
| coronary heart disease | 33 | 8.5 |
| stroke | 30 | 7.7 |
| other causes | 79 | 20.3 |
| sudden death | 23 | 5.9 |
| unknown | 8 | 2.0 |
| tumorous disease | 114 | 29.3 |
| general atherosclerosis | 22 | 5.7 |
| TOTAL | 389 | 100.0 |

# Data selection

- When joining “Entry” and “Death” tables we implicitely create a new attribute “Cause of death”, which is set to “alive” for subjects present in the “Entry” table but not in the “Death” table.
- We have only 389 subjects in death table.



# The prepared data

| Patient | General characteristics | | Examinations | | Habits | | Cause of death |
|---|---|---|---|---|---|---|---|
| | Activity after work | Education | Chest pain | … | Alcohol | ….. | |
| 1 | moderate activity | university | not present | | no | | Stroke |
| 2 | great activity | | not ischaemic | | occasionally | | myocardial infarction |
| 3 | he mainly sits | | other pains | | regularly | | tumorous disease |
| …… | …….. | …….. | ……….. | .. | … | …… | alive |
| 389 | he mainly sits | | other pains | | regularly | | tumorous disease |

# Descriptive Analysis/ Subgroup Discovery /Association Rules

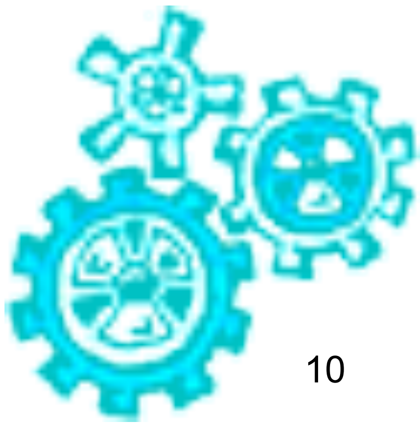
**Are there strong relations concerning death cause?**

**General characteristics (?) ⇒ Death cause (?)**

**Examinations (?) ⇒ Death cause (?)**

**Habits (?) ⇒ Death cause (?)**

**Combinations (?) ⇒ Death cause (?)**

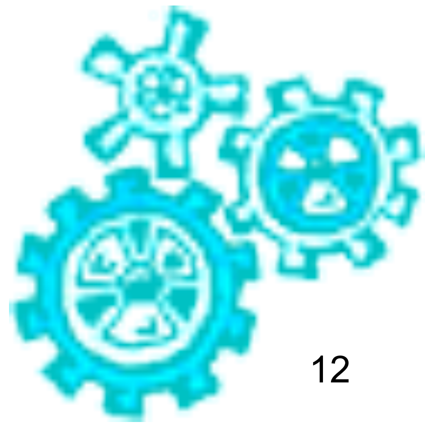


# Example of extracted rules

- **Education(university) & Height<176-180> ÞDeath cause (tumouros disease),** ***16 ; 0.62***
- **It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.**

# Example of extracted rules

- **Physical activity in work(he mainly sits) & Height<176-180> Þ Death cause (tumouros disease),** ***24; 0.52***
- **It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.**



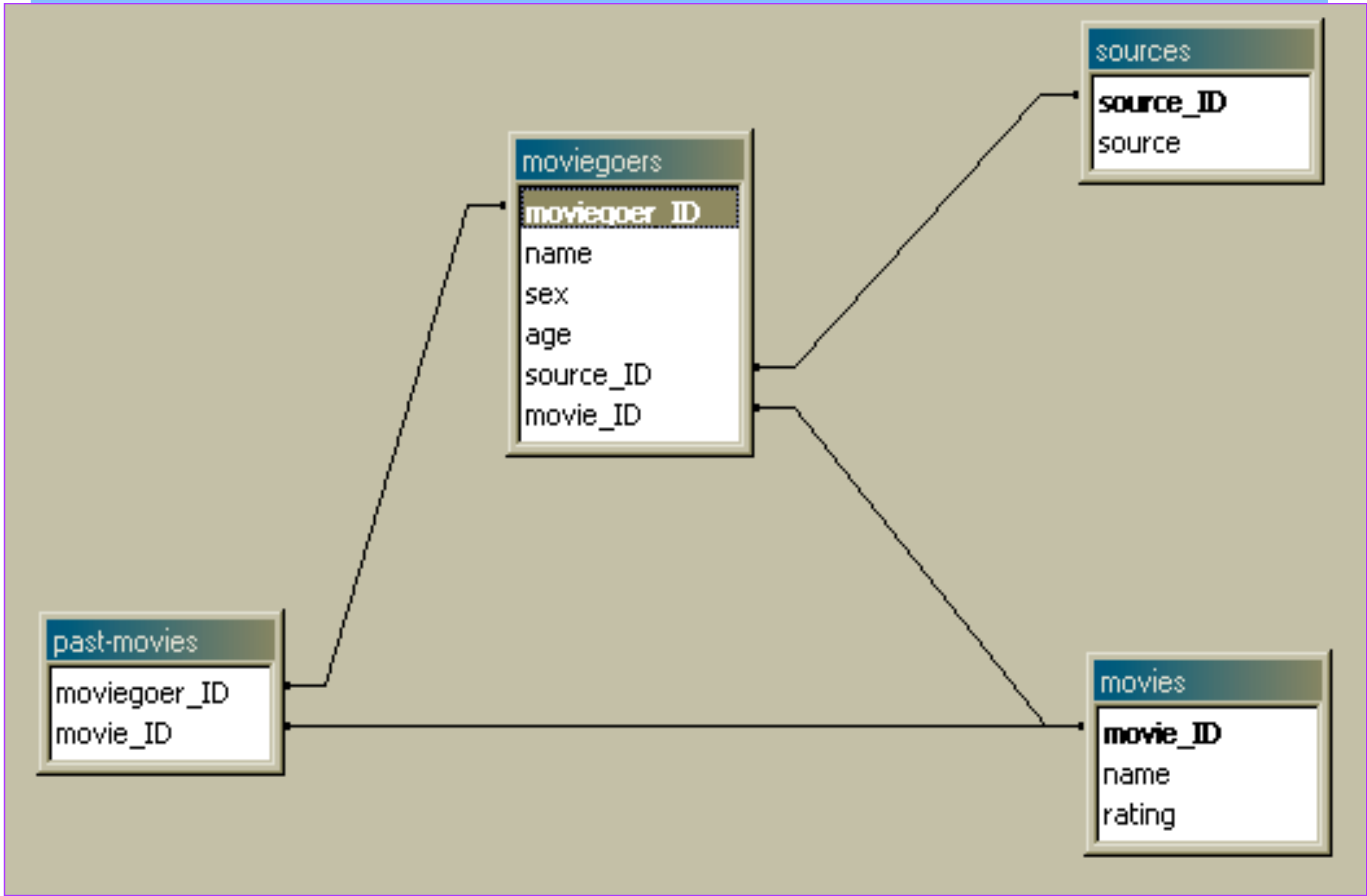
# Example of extracted rules

- Education(university) & Height<176-180> ÞDeath cause (tumouros disease),

  *16; 0.62; +1.1;*

- **the** relative frequency **of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients**

# Moviegoer Database :

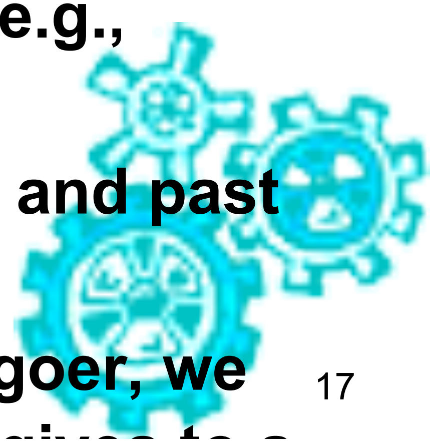

sources
source_ID
source
moviegoers
moviegoer_ID
name
sex
age
source_ID
movie_ID
past-movies
moviegoer_ID
movie_ID
movies
movie_ID
name
rating

```sql
SELECT moviegoers.name, moviegoers.sex, moviegoers.age,
sources.source, movies.name
FROM movies, sources, moviegoers
WHERE sources.source_ID = moviegoers.source_ID AND
movies.movie_ID = moviegoers.movie_ID
ORDER BY moviegoers.name;
```

| moviegoers.name | sex | age | source | movies.name |
|---|---|---|---|---|
| Amy | f | 27 | Oberlin | Independence Day |
| Andrew | m | 25 | Oberlin | 12 Monkeys |
| Andy | m | 34 | Oberlin | The Birdcage |
| Anne | f | 30 | Oberlin | Trainspotting |
| Ansje | f | 25 | Oberlin | I Shot Andy Warhol |
| Beth | f | 30 | Oberlin | Chain Reaction |
| Bob | m | 51 | Pinewoods | Schindler's List |
| Brian | m | 23 | Oberlin | Super Cop |
| Candy | f | 29 | Oberlin | Eddie |
| Cara | f | 25 | Oberlin | Phenomenon |
| Cathy | f | 39 | Mt. Auburn | The Birdcage |
| Charles | m | 25 | Oberlin | Kingpin |
| Curt | m | 30 | MRJ | T2 Judgment Day |
| David | m | 40 | MRJ | Independence Day |
| Erica | f | 23 | Mt. Auburn | Trainspotting |

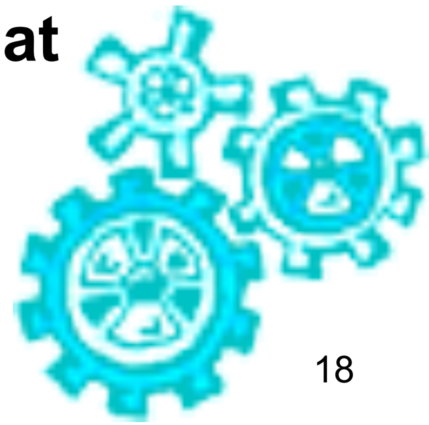
# Example: Moviegoer Database

- **Classification**
  - **determine** sex **based on age, source, and movies seen**
  - **determine** source **based on sex, age, and movies seen**
  - **determine** most recent movie **based on past movies, age, sex, and source**
- **Estimation**
  - **for predict, need a continuous variable (e.g., "age")**
  - **predict** age **as a function of source, sex, and past movies**
  - **if we had a "rating" field for each moviegoer, we**



# Example: Moviegoer Database

- **Clustering**
  - **find groupings of movies that are often seen by the same people**
  - **find groupings of people that tend to see the same movies**
  - **clustering might reveal relationships that are not necessarily recorded in the data (e.g., we may find a cluster that is dominated by people with young children; or a cluster of movies that correspond to a particular genre)**



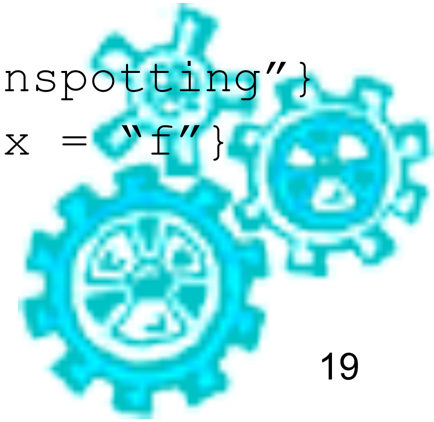
# Example: Moviegoer Database

- ## Association Rules
  - **market basket analysis (MBA): “which movies go together?”**
  - **need to create “transactions” for each moviegoer containing movies seen by that moviegoer:**

| name | TID | Transaction |
|---|---|---|
| Amy | 001 | {Independence Day, Trainspotting} |
| Andrew | 002 | {12 Monkeys, The Birdcage, Trainspotting, Phenomenon} |
| Andy | 003 | {Super Cop, Independence Day, Kingpin} |
| Anne | 004 | {Trainspotting, Schindler's List} |
| ... | ... | ... |

  - **may result in association rules such as:**

```
{“Phenomenon”, “The Birdcage”} ==> {“Trainspotting”}
{“Trainspotting”, “The Birdcage”} ==> {sex = “f”}
```



# Example: Moviegoer Database

- **Sequence Analysis**
  - **similar to MBA, but order in which items appear in the pattern is important**
  - **e.g., people who rent “The Birdcage” during a visit tend to rent “Trainspotting” in the next visit.**

# On the road to knowledge:
# mining 21 years of UK traffic accident reports

**Peter Flach et al.**

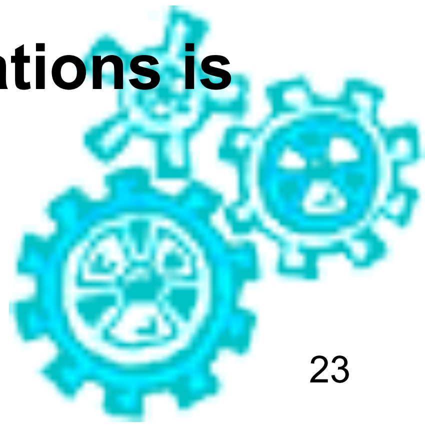
**Silnet Network of Excellence**

# Mining traffic accident reports

- The Hampshire County Council (UK) wanted to obtain a better insight into how the characteristics of traffic accidents may have changed over the past 20 years as a result of improvements in highway design and in vehicle design.
- The database, contained police traffic accident reports for all UK accidents that happened in the period 1979-1999.

# Business Understanding

- **Understanding of road safety in order to reduce the occurrences and severity of accidents.**
  - **influence of road surface condition;**
  - **influence of skidding;**
  - **influence of location (for example: junction approach);**
  - **and influence of street lighting.**
- **trend analysis: long-term overall trends, regional trends, urban trends, and rural trends.**
- **the comparison of different kinds of locations is interesting: for example, rural versus metropolitan versus suburban.**



# Data understanding

- **Low data quality. Many attribute values were missing or recorded as unknown.**
- **Different maps were created to investigate the effect of several parameters like accident severity and accident date.**

# Modelling

- **The aim of this effort was to find interesting associations between road number, conditions (e.g., weather, and light) and serious or fatal accidents.**
- **Certain localities had been selected and performed the analysis only over the years 1998 and 1999.**

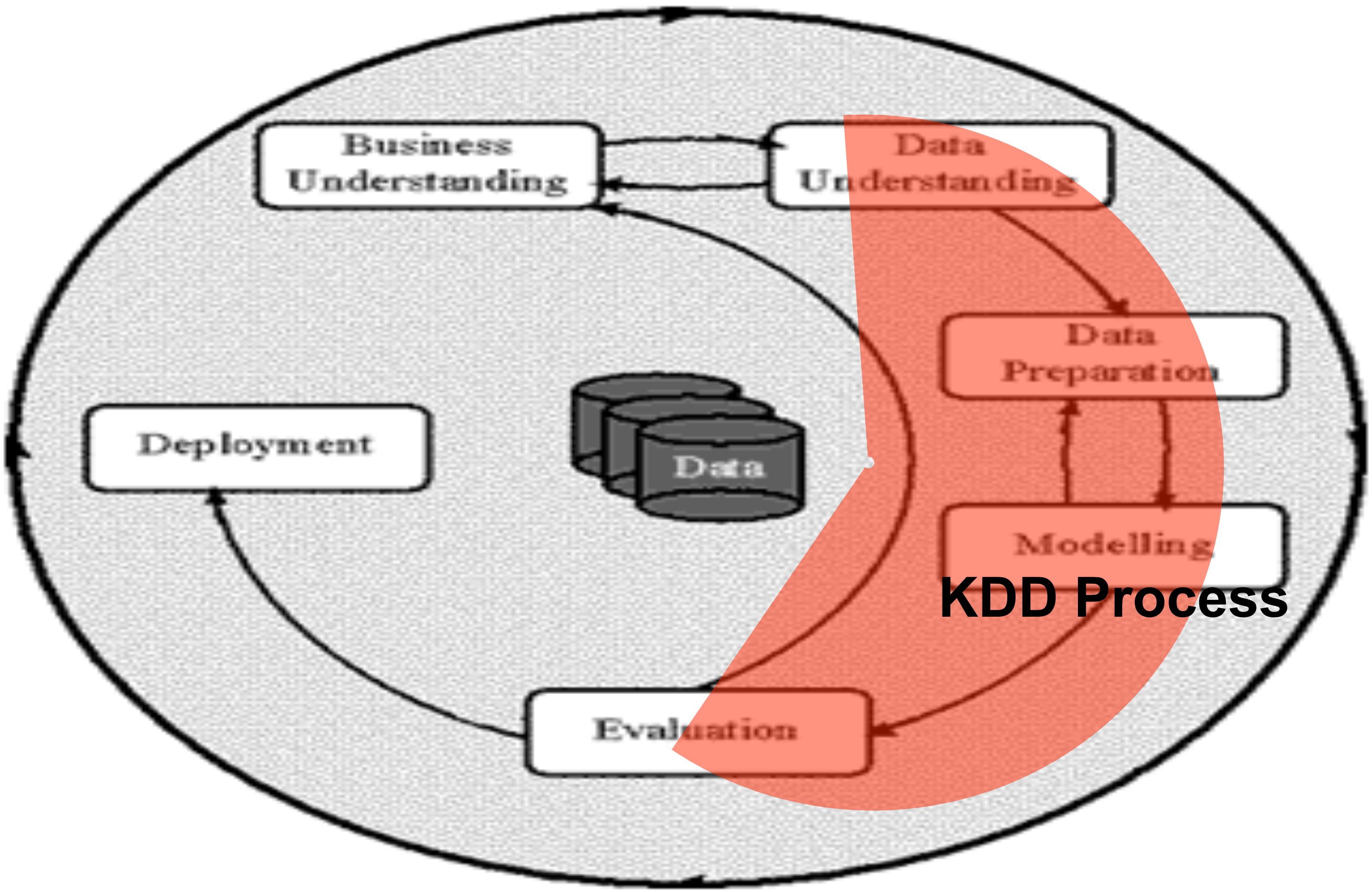
# Extracted rule

| | FATAL | Non FATAL | TOTAL |
|---|---|---|---|
| Road=V61 AND Weather=1 | 15 | 141 | 156 |
| NOT (Road=V61 AND Weather=1) | 147 | 5056 | 5203 |

- The relative frequency of fatal accidents among all accidents in the locality was 3%.
- The relative frequency of fatal accidents on the road (V61) under fine weather with no winds was 9.6% — more than 3 times greater.

# How to develop a Data Mining Project?

# CRISP-DM: **The life cicle of a data mining project**



# Business understanding

- **Understanding the project objectives and requirements from a business perspective.**
- **then converting this knowledge into a data mining problem definition and a preliminary plan.**
  - **Determine the Business Objectives**
  - **Determine Data requirements for Business Objectives**
  - **Translate Business questions into Data Mining Objective**

# Data understanding

- **Data understanding: characterize data available for modelling. Provide assessment and verification for data.**

# Modeling

- In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
- Therefore, stepping back to the data preparation phase is often necessary.

# Evaluation

- **At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective.**
- **Evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.**
- **A key objective is to determine if there is some important business issue that has not been sufficiently considered.**

# Deployment

- The knowledge gained will need to be organized and presented in a way that the customer can use it.
- It often involves applying “live” models within an organization’s decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases.

# Deployment

- It can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.
- In many cases it is the customer, not the data analyst, who carries out the deployment steps.