Chapter 1: Data Mining Process

1. Data Sources
2. Definitions
3. Process

Data Sources, Examples

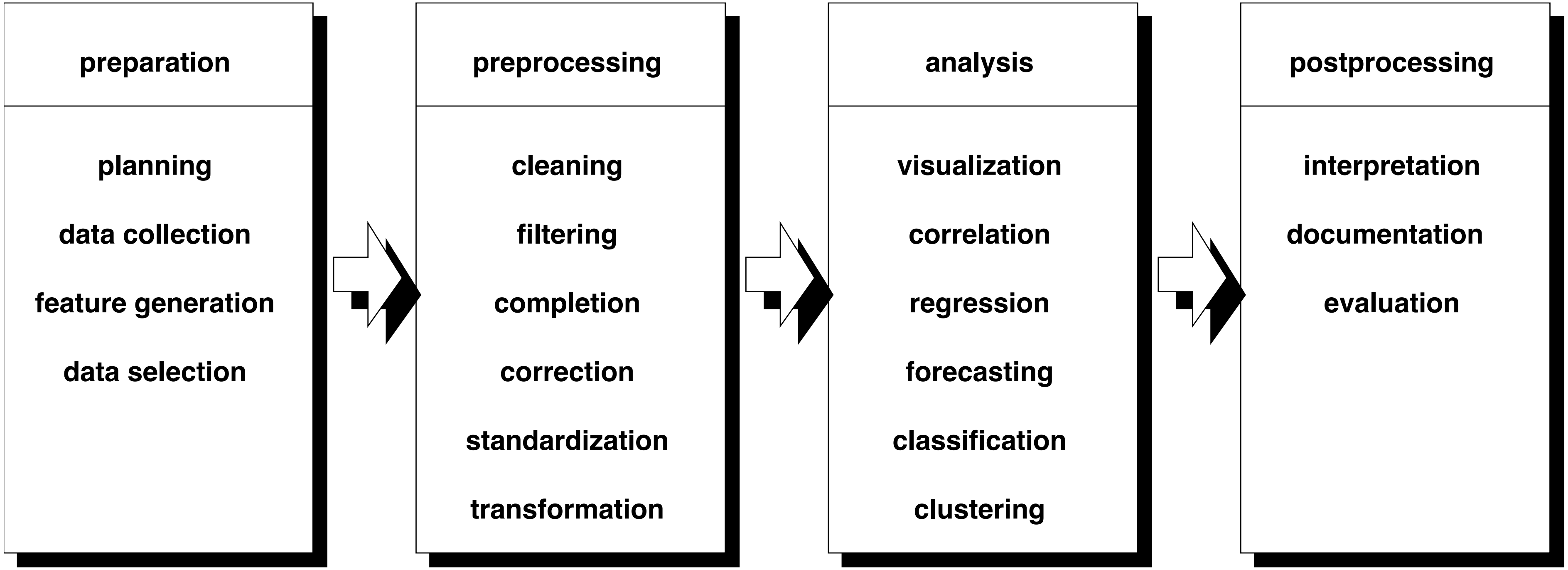
- industrial process data
  - field and controller level
  - operator and management level
- business data
  - shopping basket analysis
  - customer segmentation
- text data
  - text documents, text messages
  - web documents
- image data
  - smartphone cameras
  - satellite data
- biomedical data
  - genome data
  - lab data

Definitions

- Data Mining (DM): extract knowledge from data
- knowledge: interesting patterns
- interesting: general, nontrivial, new, useful, comprehensive
- Knowledge Discovery (KDD): preprocessing (a priori knowledge), knowledge extraction, postprocessing (evaluation)
- Data Analytics (DA) application of computer systems to the analysis of large data sets for the support of decisions
- DM, KDD, DA: feedback processes involving experts
- related areas: statistics, signal theory, pattern recognition, computational intelligence, machine learning, operations research

Knowledge Discovery Process

4 phases



preparation

- planning
- data collection
- feature generation
- data selection

preprocessing

- cleaning
- filtering
- completion
- correction
- standardization
- transformation

analysis

- visualization
- correlation
- regression
- forecasting
- classification
- clustering

postprocessing

- interpretation
- documentation
- evaluation

Chapter 2: Data and Relations

1. Example
2. Scales
3. Matrix Representation
4. Relations
5. Dissimilarity/Distance Measures
6. Similarity/Proximity Measures
7. Relations for Sequences and Text
8. Sampling and Quantization

Iris Data Set (Anderson 1935)

- data set with $n = 150$
  vectors of dimension $p = 4$
- objects: iris plants
- classes (50 instances each):
  Iris Setosa
  Iris Versicolor
  Iris Virginica
- components:
  sepal length
  sepal width
  petal length
  petal width

Iris Data Set (Part)

Setosa				Versicolor				Virginica			
sepal		petal		sepal		petal		sepal		petal	
length	width	length	width	length	width	length	width	length	width	length	width
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1

Typical Questions to answer in data analytics

1. Which of the data might contain errors or false class assignments?
2. What is the error caused by rounding the data off to one decimal place?
3. What is the correlation between petal length and petal width?
4. Which pair of dimensions is correlated most?
5. None of the flowers in the data set has a sepal width of 1.8 centimeters. Which sepal length would we expect for a flower that did have 1.8 cm as its sepal width?
6. Which species would an Iris with a sepal width of 1.8 centimeters belong to?
7. Do the three species contain sub-species that can be identified from the data?