

Fig. 1.1 Steps in the KDD process

[3]

# KNOWLEDGE DISCOVERY in DATA

P A T T E R N module:

**pattern.web**

- url downloads
- interval requests
- search engine requests
- use google translate
- crawl
  - wikipedia articles
  - fb comments + reactions
  - dbpedia
  - twitter
  - rss
- parse HTML elements, PDFs
- retrieve emails via imap
- retrieve local information (eg. tweets)

[1]

P A T T E R N module:

**pattern.db**

- built database
- work with time/date

[1]

**sources**[1]: <http://www.clips.ua.ac.be/pages/pattern>

[2]: CLIPS - Guy de Pauw, Pattern workshop - Cqrelations, January 2015

[3]: Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

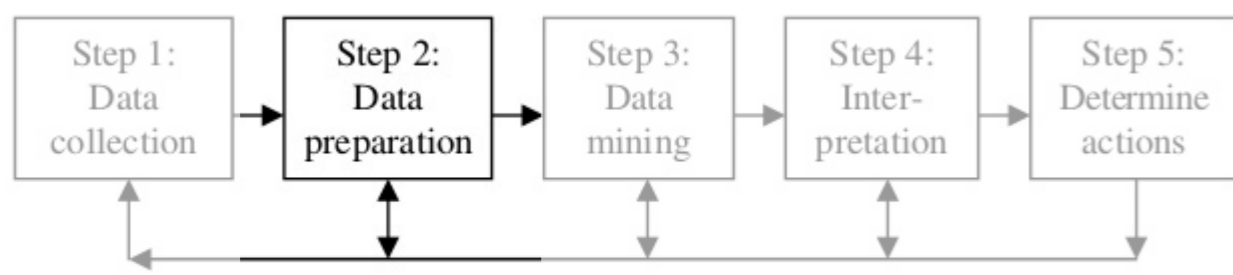


Fig. 1.1 Steps in the KDD process

[3]

# KNOWLEDGE DISCOVERY in DATA

Step 2: – text-preparation

**syntax:**  
 tokenization/normalization (98%)\*  
 simplest thing/important thing  
 identifying the units in your text  
 to read the punctuation, e.g.:  
 - dr.  
 - This is a sentence.

**lemmatization:**  
 reduce wordforms to their dictionary item  
 is/been/was/be  
 --> belongs to 'to be'  
 + plurals --> singulars

**syntactical:**  
 part-of-speech tagging  
 important elements for object text-mining  
 --> nouns  
 for subjective text-mining  
 --> adjectives  
 word sense disambiguation  
 bank / bank  
 --> river bank / money bank  
 semantic role labeling

**pragmatics: (?)**  
 named entity recognition  
 co-reference resolution (50%)\*  
 <-- meaning output

\*(% refers to accuracy)

[2]

P A T T E R N module:

```

pattern.en |es|de|fr|it|nl
- text preparation
- sentiment analysis tool
- WordNet interface
- wordlists interface
  
```

[1]

P A T T E R N module:

```

pattern.search
- a pattern matching system
  similar to regular expressions,
  that can be used to search a string
  by syntax (word function) or by
  semantics (word meaning).
- eg.: ('{NP} be * than {NP}')
```

[1]

## sources

[1]: <http://www.clips.ua.ac.be/pages/pattern>

[2]: CLIPS - Guy de Pauw, Pattern workshop - Cqrelations, January 2015

[3]: Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

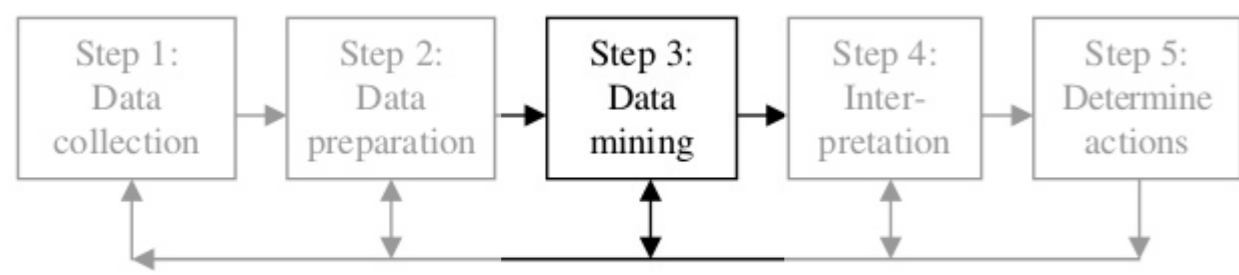


Fig. 1.1 Steps in the KDD process

[3]

# KNOWLEDGE DISCOVERY in DATA

P A T T E R N module:

**pattern.vector**

- machine learning tools:
  - word count functions
  - bag-of-word documents
  - a vector space model
  - latent semantic analysis (context analysis)
- algorithms for *clustering*
  - k-means (similar clusters)
  - hierarchical (nested clusters)
- and *classification*
  - NB (Naive Bayes)
  - KNN (k-nearest neighbor)
  - SLP (Single-layer perceptron)
  - SVM (Support vector machine)
- genetic algorithm

[1]

RANDOM POINTS IN 2D

POINTS BY DISTANCE TO CENTROID

[1]

COMPLEX IN LOW DIMENSION

SIMPLE IN HIGHER DIMENSION

[1]

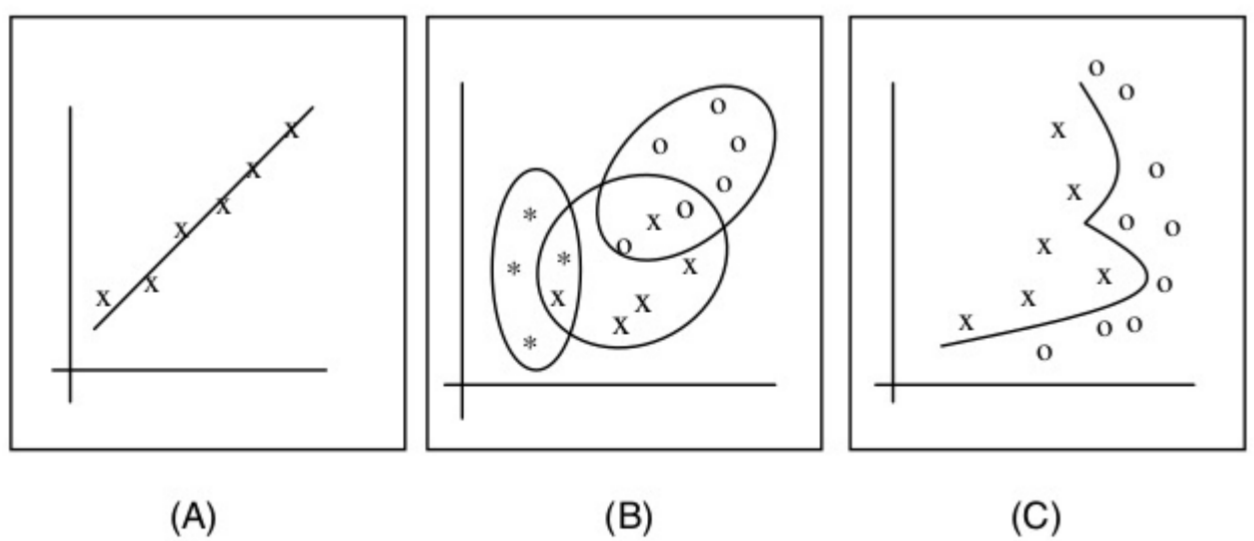
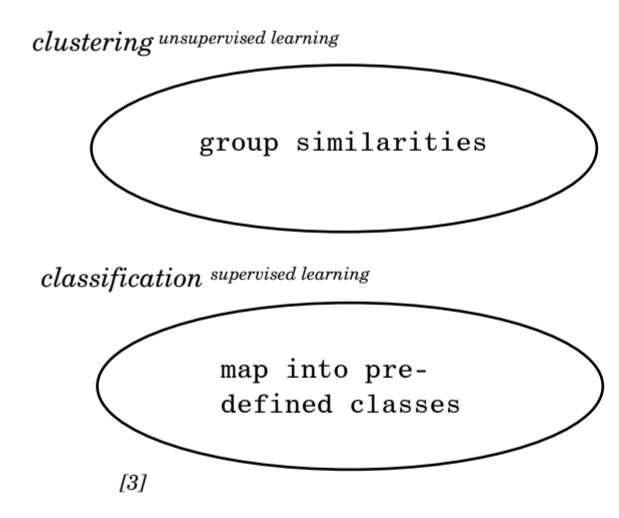


Fig. 2.2 Examples of different types of discovery algorithms: Pattern mining with a linear regression function (A), clustering (B), and classification (C)

[3]



**sources**

[1]: <http://www.clips.ua.ac.be/pages/pattern>  
 [2]: CLiPS - Guy de Pauw, Pattern workshop - Cqrelations, January 2015  
 [3]: Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

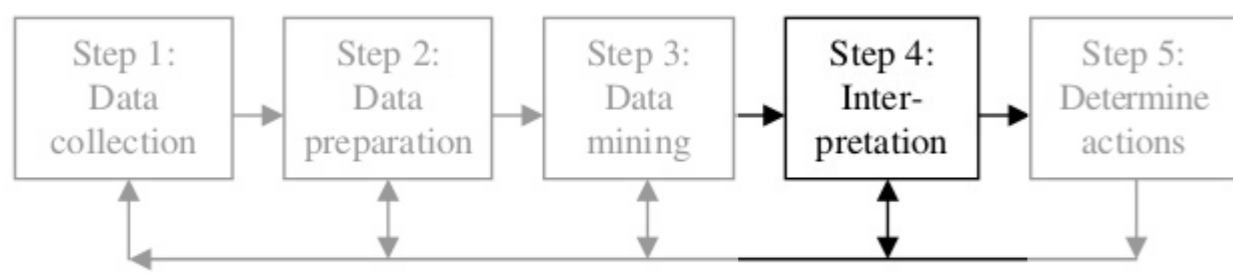


Fig. 1.1 Steps in the KDD process

[3]

# KNOWLEDGE DISCOVERY in DATA

Step 4: – interpretation

**gold standard**

--> annotated test set

**10-fold cross validation**

taking 1000 tweets  
 training 800 tweets  
 test 100 tweets  
 val 100 tweets

**compare to baseline scores**

- frequency-baseline  
 you expect 80% of the tweets to be neutral(?)

- informative baseline  
 i have a 60% chance that it will rain tomorrow  
 --> your result need to be higher  
 otherwise --> why do all the work?

[2]

**sources**

[1]: <http://www.clips.ua.ac.be/pages/pattern>

[2]: CLIPS - Guy de Pauw, Pattern workshop - Cqrelations, January 2015

[3]: Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)

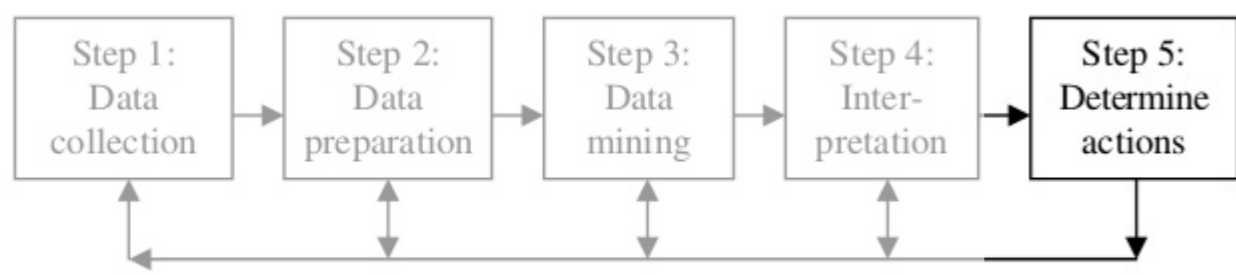


Fig. 1.1 Steps in the KDD process

[3]

# KNOWLEDGE DISCOVERY in DATA

Step 5: – Pattern in practise (actions)

> ... as a maker of such tools, it has a site of application that has not been intended.

How do you see this, and where will there be talked about certain issues?

< **three levels:**

- fundamental research  
without any concrete applications
- IWT\* research center (applied)  
develop research to improve problems in society (AMiCA)
- industry sponsored efforts

\* IWT, Agentschap voor Innovatie door Wetenschap en Technologie

[2]

P A T T E R N module:

**pattern.graph**

graph analysis (shortest path, centrality) and graph visualization in the browser. A graph is a network of nodes connected by edges. It can be used for example to study social networks or to model semantic relationships between concepts.

[1]

#### sources

[1]: <http://www.clips.ua.ac.be/pages/pattern>

[2]: CLIPS - Guy de Pauw, Pattern workshop - Cqrelations, January 2015

[3]: Data Mining and Profiling in Large Databases, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.) (2013)