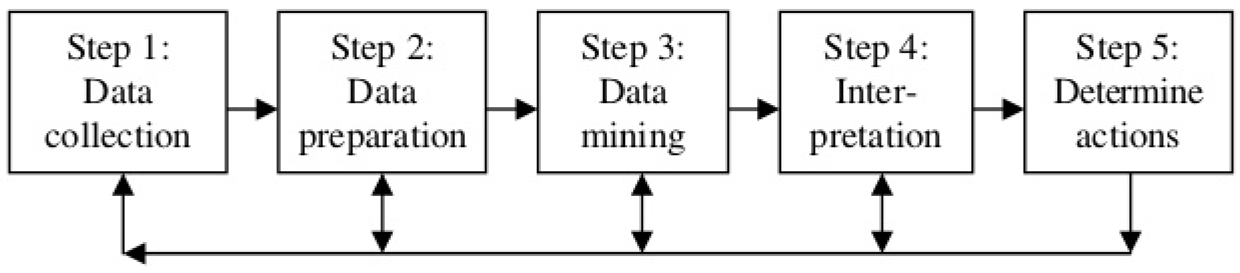


Fig. 1.1 Steps in the KDD process
(**k**nowledge **d**iscovery in **d**ata)

[3]

P A T T E R N

# data mining (step 3)
- a Google, Twitter and Wikipedia API
- a web crawler
- a HTML DOM parser

# natural language processing (step 2)
- part-of-speech taggers
- n-gram search
- sentiment analysis
- WordNet

# machine learning (step 4)
- vector space model
- clustering
- SVM

# network analysis
# <canvas> visualization

[1]

Step 2: – text-preperation

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]

Step 3: – data mining

PATTERN modules:

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)

pattern.db
- built database
- work with time/date

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

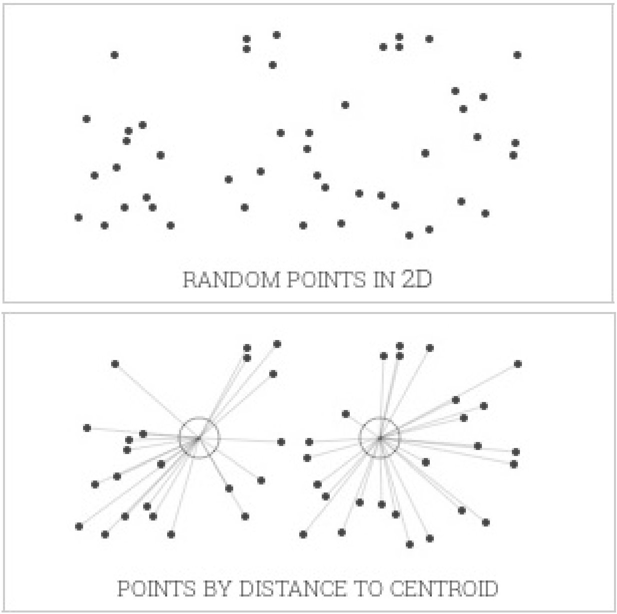
pattern.search
- a pattern matching system similar to regular expressions, that can be used to search a string by *syntax* (word function) or by *semantics* (word meaning).
- eg.:('{NP} be * than {NP}')

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)

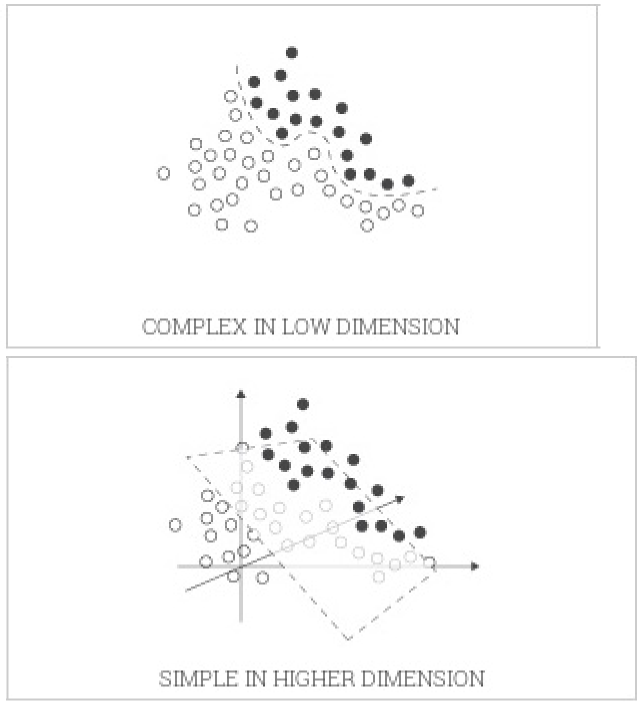
    and *classification*
    - NB (Naive Bayes)
    - KNN (k-nearest neighbor)
    - SLP (Single-layer perceptron)
    - SVM (Support vector machine)

pattern.graph

[1]



[1]



[1]

clustering unsupervised learning

group similarities

classification supervised learning

map into pre-defined classes

[3]

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]

Step 5: – Pattern in practise (actions)

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

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]

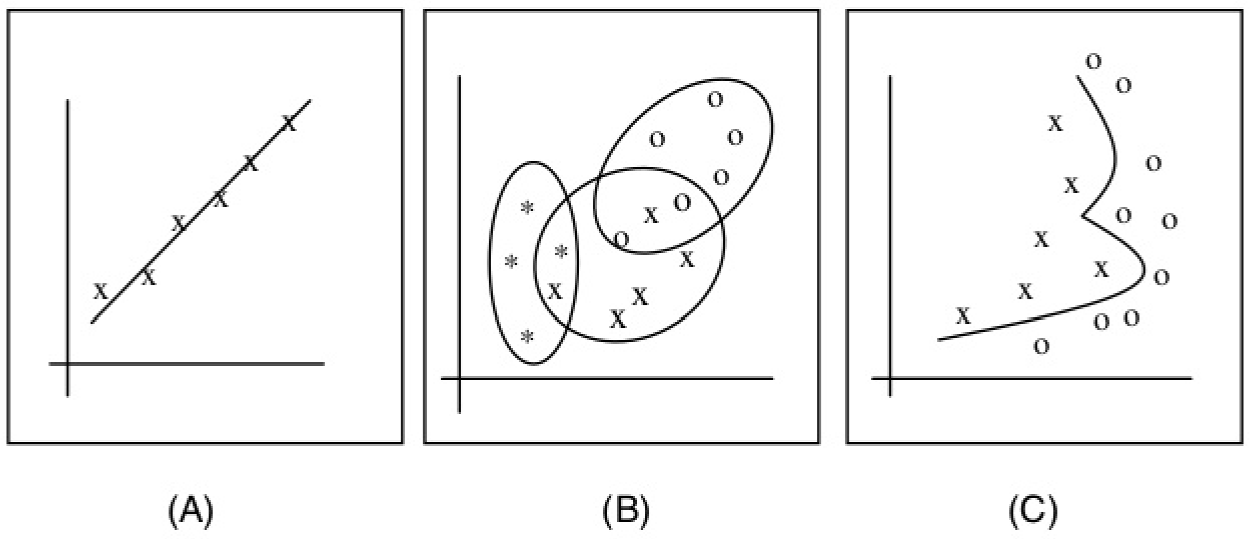


Fig. 2.2 Examples of different types of discovery algorithms: Pattern mining with a linear regression function (A), clustering (B), and classification (C)

[3]

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