Lecture 1:

Introduction in Data Mining

Outline

- Why Data Mining? – a short motivation
- What is Data Mining? – basic concepts
- What is not Data Mining? – related topics
- Categories of data
- Main data mining tasks
- Course organization and evaluation rules

Why Data Mining?

Currently, a lot of data are collected:

- Commercial transactions (e.g. hypermarkets)
- Financial transactions (e.g. ATM – credit card transactions)
- Web usage data(e.g. e-commerce or other web services)
- Social interactions (e.g. social network - see Facebook, Twitter)
- Remote sensing data (e.g. data on Earth and atmosphere collected by sensors placed on satellites)
- Gene expression data (e.g. data collected using microarrays)
- Medical data (e.g. electronic health records)
- Electronic documents (e.g. scanned documents from libraries)

…

Why Data Mining?

These data incorporate a lot of knowledge which should be extracted in order to:

- Provide recommendations (e.g. guide the marketing activity, suggest products to customers)
- Detect anomalous behavior (e.g. fraudulent access to a bank account)
- Predict evolution (e.g. weather forecasting, stock market predictions)
- Identify patterns (e.g. infer the role of genes)
- Assist medical decision (e.g. provide suggestions of potential diagnostics)

Why Data Mining?

Example 1: **In vitro fertilization data**

[Witten, Frank, Hall – Data Mining. Practical Machine Learning Tools and Techniques - http://www.cs.waikato.ac.nz/ml/weka/book.html]

Given: embryos described by 60 features

Problem: selection of embryos with chance of survival

Data: historical records with features of embryos and their outcome (viable or not)

Example 2: **Processing loan applications**

Given: questionnaire with financial and personal information (e.g. age, employment status, health status, bank status etc.)

Problem: decide if the loan application is accepted or not

Data: historical records with other financial and personal information and the loan outcome (the loan has been reimbursed in time or there have been problems)

Why Data Mining?

Example 3: **Power load forecasting** (estimate future demand for power – useful for electricity supply companies)

[Witten, Frank, Hall – Data Mining. Practical Machine Learning Tools and Techniques - http://www.cs.waikato.ac.nz/ml/weka/book.html]

Given: a load model based on normal climatic conditions

Problem: forecast the minimum/maximum load for each hour

Data: historical records with meteorological data (temperature, humidity, wind speed, cloud cover) and the actual load

Example 4: **Market basket analysis**

Data: database of transactions (a transaction contains information about the products bought by each customer)

Problem: identify groups of items which tend to belong/occur together in a transaction (e.g. bread and water)

Why Data Mining?

Example 5: **Anomaly detection**

Given: transactions data

Problem: identify changes in the behavior of the customers

Example 6: **User profiles**

Given: access logs for a web server

Problem: identify user profiles (characterized by similar behaviors)

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

What is Data Mining?

There are various definitions:

Data mining = “collecting, cleaning, processing, analyzing data and gaining useful insights from them” [C.Aggarwal – Data Mining. The Textbook, 2015]

- Collecting: there are various data sources (sensors, written documents, web engines, gene microarrays, other devices)

- Cleaning: remove the noise (inconsistent or erroneous data)
- Processing: transform the data in a standardized format

- Analyzing: identify patterns, regularities, associations or relationships in data
- Gaining insight/ knowledge: formulate concise/ interpretable/ explainable and actionable rules (to be used by human decision makers)

What is Data Mining?

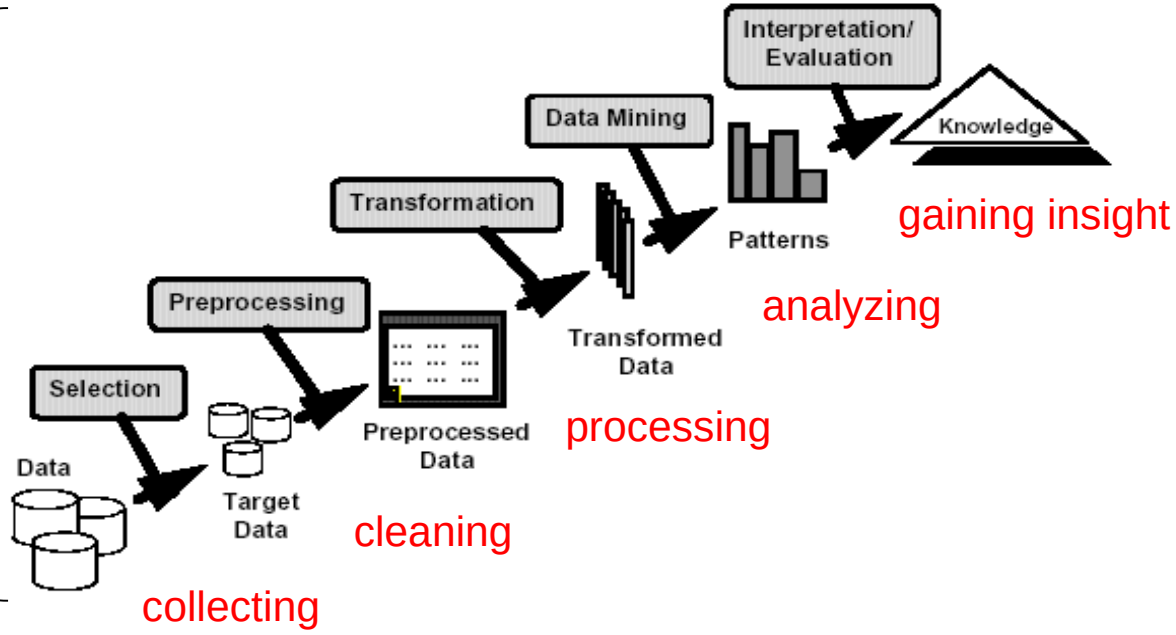
There are various definitions:

Data mining =

"extraction of implicit, previously unknown and potentially useful information from data" [http://www.cs.waikato.ac.nz/ml/weka/book.html] or

"exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns" [Tan, Steinbach, Kumar – Introduction to Data Mining, 2004]

Sometimes this process is called knowledge discovery and data mining is just the analysis step



What is not Data Mining?

Example: Given a database containing information on customers of a bank:

- Searching for all customers who live in a given town is not a data mining task
- Searching for all customers who have in their account an amount larger / smaller than a given threshold is not a data mining task

… such problems can be solved by a simple query in the database

On the other hand:

- Identify customers who are reliable for a loan
- Identify anomalous operations on an account

… are problems requiring human expertise and/or data mining tools

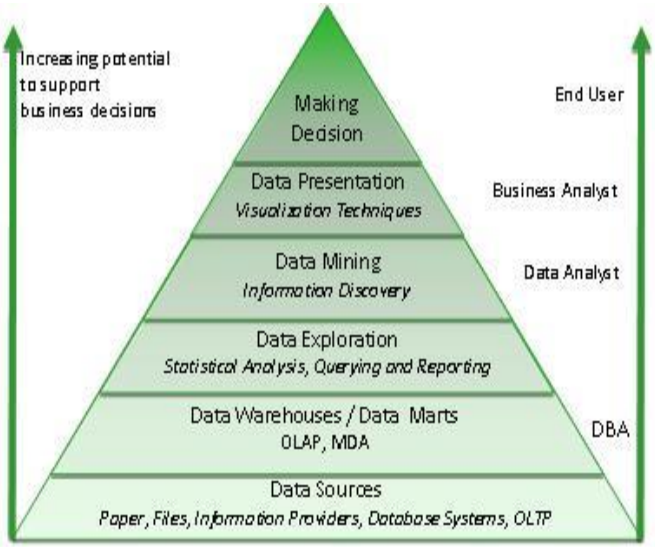
Related fields

Data mining is related to:

- Statistics – some data mining methods have roots and rely on statistical models and methods
- Machine learning = extracting models from data through a learning process – most of models used in data mining rely on machine learning tools
- Database technology – most data are in databases
- Other:
  - Visualization: various tools for visualizing data
  - Optimization: extracting some models from data is based on solving optimization problems
  - Linear algebra: data are frequently organized as matrices, thus operations/transformations on them are frequently used

Other definition

Data mining = application of machine learning methods to extract knowledge from real data



Some related buzzwords: data science, big data

Categories of data

Structured data = set of records/instances/items containing a fixed number of fields/ attributes/ features

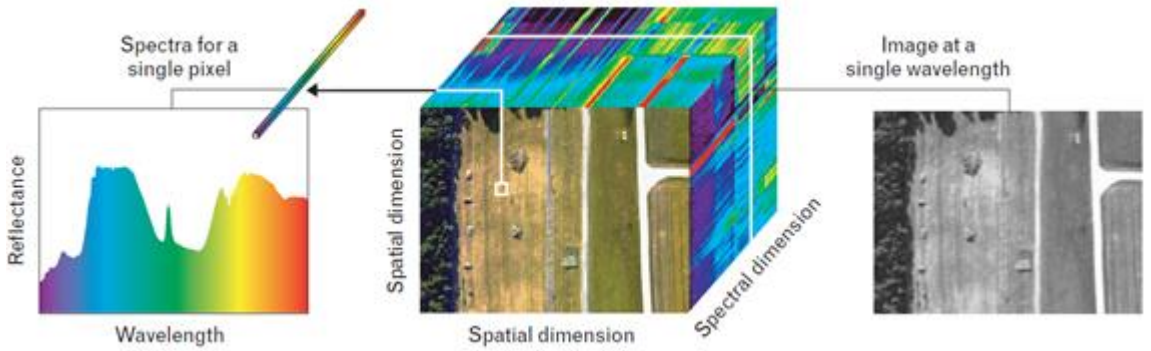
Remarks:

- Each instance corresponds to one of the objects/entities to be analyzed (e.g. customer, patient, transaction, day etc.)
- Each attribute corresponds to a measurable characteristic of the object (e.g. age, weight, income, temperature etc.)

Examples:

- Bi-dimensional table (i.e. data matrix)
  - Relational database
  - Spreadsheet
- Multi-dimensional table
  - Multi-spectral images

http://www.tankonyvtar.hu/en/tartalom/tamop425/0032_terinformatika/ch04s04.html



Categories of data

Structured data = set of records/instances/items containing a fixed number of fields/ attributes/ features (correspond to each characteristic of the instance)

Example: Car Evaluation Database [http://archive.ics.uci.edu/ml/datasets.html]

1728 instances 6 attributes

Aim: classify a car in one of four categories: unacceptable, acceptable, good, very good

Attributes

Instance	Buying price	Mainte-nance price	Number of doors	Capacity	Size of luggage boot	Safety	Class
1	Very high	Very high	2	2	small	low	unaccept.
2	Very high	high	4	4	big	medium	unaccept.
3	Very high	medium	5more	4	big	medium	accept
4	low	low	5more	4	big	medium	good

Categories of data

Semi - structured data = data which do not have a standard structure (i.e. not all instances have the same attributes); however they contain some elements (e.g. tags) which help to identify some structure in data

Example: XML version of a CV file [http://www.eife-l.org]

Aim: automatic processing of CVs for expertise retrieval (typical task for HR departments)

```
....
<Address type="Residence">
  <oa:AddressLine sequence="1">myaddress</oa:AddressLine>
  <oa:CityName>mycity</oa:CityName>
  <CountryCode>FR</CountryCode>
  <oa:PostalCode>29630</oa:PostalCode>
  <UserArea> <europass:CountryLabel
                    xml:lang="fr">France</europass:CountryLabel>
  </UserArea>
 </Address>
...
```

Remark: the semi-structured data are usually transformed in structured data before applying data mining tasks

Categories of data

Unstructured data = data which are not organized in a pre-defined manner (no model of the data) – they are mainly as free text and the useful information should be extracted from the text.

Example: text documents

Tasks:

- document summarization (extract keywords, main ideas)
- named entity recognition (e.g. identify names of persons, institutions, geographical places etc)
- sentiment analysis, opinion mining (extract and quantify subjective information from text)

Challenges:

- The data might be ambiguous (e.g. the name of a person can appear in different versions: Ioan Popescu, I. Popescu, Popescu Ioan)
- Processing textual data requires methods specific to natural language processing (e.g. part-of-speech tagging – identification of substantives, verbs, adjectives …) – [which is not the object of this course]

Main data mining tasks

Predictive tasks

Aim: predict unknown or future values of some attribute(s) based on the values of other attributes or previous values of the same attribute

Variants:

- Classification = identify the class (category) to which an instance should belong (based on the values of its attributes)

  Examples: in vitro fertilization data, evaluation of loan applications

- Regression = estimate the value of an attribute based on the values of other attributes

  Example: power load forecasting

Main data mining tasks

Descriptive tasks

Aim: find human interpretable patterns that describe/explain the data

Variants:

- Clustering = identify natural groups in data

  Example: user profiles

- Association rule discovery = find dependency rules describing data co-occurrency

  Example: market basket analysis

- Outlier analysis = identify entities (records, instances) which seems to be unusual in some sense (significantly different from the other data)

  Example: anomaly behavior/fraud detection

Classification

What is known?

- a collection of records for which it is known to which class they belong (training data set)
- each record contains a set of attributes and one of the attributes is the class label

What is desired?

- a model which captures the relationship between the class attribute and the other attributes (the model is inferred using a training set through a process which is called supervised learning)

Which is the final aim?

- Use the inferred model to identify the right class for a previously unseen record

Remark: a useful model should be accurate; the model accuracy should be estimated using data which have not been used during the training (test set)

Classification

Example:

- Medical diagnosis = associate to a health record a class associated with the presence/absence of an illness

Example of a data subset (breast-cancer-wisconsin - arff format – see Lab 1)

```
@relation wisconsin-breast-cancer
@attribute Clump_Thickness integer [1,10]
@attribute Cell_Size_Uniformity integer [1,10]
@attribute Cell_Shape_Uniformity integer [1,10]
@attribute Marginal_Adhesion integer [1,10]
@attribute Single_Epi_Cell_Size integer [1,10]
@attribute Bare_Nuclei integer [1,10]
@attribute Bland_Chromatin integer [1,10]
@attribute Normal_Nucleoli integer [1,10]
@attribute Mitoses integer [1,10]
@attribute Class { benign, malignant}
@data
5,1,1,1,2,1,3,1,1,benign
5,4,4,5,7,10,3,2,1,benign
3,1,1,1,2,2,3,1,1,benign
8,10,10,8,7,10,9,7,1,malignant
1,1,1,1,2,10,3,1,1,benign
```

Classification

Example:

- Spam filter = identify the class (spam/ham) for an e-mail or sms

Example of a data subset (SMS spam collection dataset from UCI Machine Learning Repository)

ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham Siva is in hostel aha:-.

spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop

spam Sunshine Quiz! Win a super Sony DVD recorder if you can name the capital of Australia? Text MQUIZ to 82277. B

spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

….

Regression

What is known?

- a collection of records characterized by numerical attributes (data set) – some of them are predictors and the other ones (in many cases only one) are the response/ target attribute(s)

What is desired?

- a model of dependence between one of the attributes (response attribute) and the other ones (predictors)

Which is the final aim?

- Predict a value of the attribute of interest based on known values of other attributes

Remark

- a linear or nonlinear model of dependency should be previously assumed; the model could be fixed, as in statistical regression, or it could be more flexible (e.g. neural networks or other models from computational intelligence)

Regression

Examples:

- Predicting sales amounts of a new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series forecasting in the case of stock market indices.

Data set: prediction of the gasoline consumption based on car characteristics

@relation autoMpg
@attribute cylinders { 8, 4, 6, 3, 5} @attribute displacement real
@attribute horsepower real @attribute weight real @attribute acceleration real
@attribute model { 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82}
@attribute origin { 1, 3, 2}
@attribute MilesPerGallon real
@data
8,307,130,3504,12,70,1,18
8,350,165,3693,11.5,70,1,15
8,318,150,3436,11,70,1,18
8,304,150,3433,12,70,1,16
….

Clustering

What is known?

- A set of data (not necessarily structured)
- A similarity/dissimilarity measure between data (it is specific to the problem)

What is desired?

- A model describing the grouping of data in clusters such that data belonging to the same cluster are more similar than data belonging to different clusters

Which is the final aim?

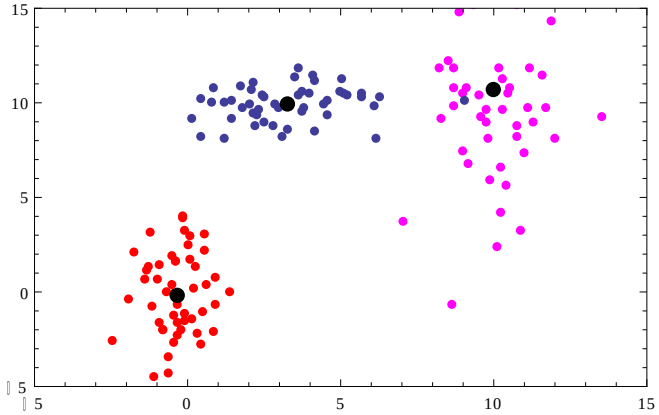
- Check if two data belong to the same cluster
- Find the most appropriate cluster for a new data

Remark: for some clustering methods it is enough to know the matrix of (dis)similarity values

Clustering

Examples:

- Customer segmentation = identify groups of customers with similar shopping behaviors
- Data summarization / document clustering = identify groups of electronic documents based on their content
- User profiles extraction = identify groups of users of an e-commerce system characterized by similar behaviors
- Image Segmentation = identify homogeneous regions in an image



Outlier Analysis

What is known?

- A set of data (not necessarily structured)
- A similarity/dissimilarity measure between data (they are specific to the problem)

What is desired?

- A model capturing the normal/ typical distribution of data (rmk: not necessarily normal distribution from a statistical point of view)

Which is the final aim?

- Identify exceptions, i.e. data which significantly deviates from the model

Remark: outlier analysis is somehow complementary to cluster analysis

Outlier Analysis

Examples:

- Intrusion detection systems
  - Anomalous operating system calls or network traffic might suggest the presence of malicious activity
- Credit card fraud
  - Unusual patterns in the usage of a credit card (e.g. card used in unusual geographical locations/ at unusual times) might suggest a possible fraud
- Medical diagnosis
  - Abnormal structures observed on MRI (magnetic resonance imaging), PET (positron emission tomography) images or EKG time series might reflect some disease conditions

Association Rule Discovery

What is known?

- a set of records each of which contain a list of items from a given collection (e.g. a transaction, list of shopping items)

What is desired?

- Find a dependence model which can predict the occurrence of an item based on occurrences of other items

Which is the final aim?

- Identify patterns of associations between items

Association Rule Discovery

Example: market basket analysis (each instance is a transaction = list of products/items)

T1: {milk, bread, meat, water}
T2: {bread, water}
T3: {bread, butter, meat, water}
T4: {water}

Outcomes:

- Frequent itemset: {bread, water} - a support of 75% (the pair of items appears in 3 out of 4 transactions) –> "bread and water are frequently bought together"
- Association rule: bread -> water (a 100% confidence: "in all cases when bread is bought also water is bought")

Course structure

1. Introduction to knowledge discovery (this lecture)
2. Data pre-processing
3. Classification methods
4. Clustering methods
5. Association rules
6. Nonlinear regression and time series analysis
7. Ensemble methods
8. Text mining, web mining, network analysis

Course materials: http://staff.fmi.uvt.ro/~daniela.zaharie/dm2020/EN
- lectures
- lab

Lab structure

1. Data sets and repositories. Introduction in Rattle (R) and Scikit-learn (Python)
2. Data visualization and pre-processing
3. Classification using rules, decision trees, instance-based models, probabilistic models etc
4. Clustering using partitional and hierarchical methods
5. Association rules and regression models
6. Time series analysis
7. Ensemble methods, text mining

References

- C.C. Aggarwal, *Data Mining – The Text Book*, Springer, 2015
- M. H. Dunham. *Data Mining. Introductory and Advanced Topics*, Pearson Education 2003
- F. Gorunescu, *Data Mining. Concepts, Models and Techniques*, Springer, 2011
- C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.
- I.H. Witte, E. Frank, M.A. Hall. *Data Mining – Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2011

Evaluation

- Open book written exam (20%) – 20 questions/ 90 minutes

- Project (60%-80%):
  - Report (6-12 pages)
  - Application (in R, Python, Java (e.g. Weka) or any other language)
  - Slides for presentation during the exam (for around 10 minutes)

- Lab activity (20%)
  - Lab participation
  - Homework

Examples of applications

(Store Product Placement) A merchant has a set of d products together with previous transactions from the customers containing baskets of items bought together. The merchant would like to know how to place the product on the shelves to increase the likelihood that items that are frequently bought together are placed on adjacent shelves.

Which task is appropriate?

(Product Recommendations) A merchant has an n × d binary matrix D representing the buying behavior of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.

Which task is appropriate?

Examples of applications

(Store Product Placement) A merchant has a set of d products together with previous transactions from the customers containing baskets of items bought together. The merchant would like to know how to place the product on the shelves to increase the likelihood that items that are frequently bought together are placed on adjacent shelves.

Association rules

(Product Recommendations) A merchant has an n × d binary matrix D representing the buying behavior of n customers across d items. It is assumed that the matrix is sparse, and therefore each customer may have bought only a few items. It is desirable to use the product associations to make recommendations to customers.

Clustering

Applications

(Medical ECG Diagnosis) Consider a set of ECG time series that are collected from different patients. It is desirable to determine the anomalous series from this set.

Which task is appropriate?

(Web Log Anomalies) A set of Web logs is available. It is desired to determine the anomalous sequences from the Web logs.

Which task is appropriate?

Applications

(Medical ECG Diagnosis) Consider a set of ECG time series that are collected from different patients. It is desirable to determine the anomalous series from this set.

(Web Log Anomalies) A set of Web logs is available. It is desired to determine the anomalous sequences from the Web logs.

Outlier detection

Classification

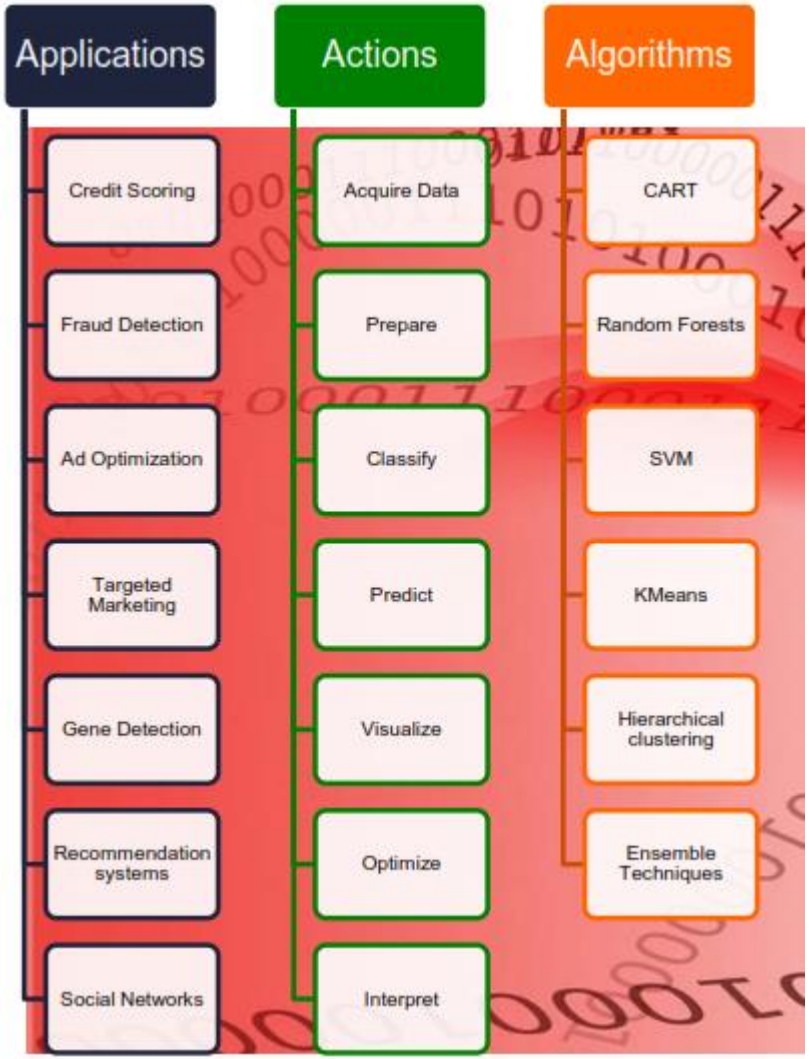
Summary

Data mining:

- Application
- Task (action)
- Method (algorithm)

Key elements:

- Ask the right question
- Identify the appropriate data
- Prepare the data
- Select the appropriate algorithms
- Interpret the results of algorithms



Summary: a roadmap for a data scientist

"A data scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician."