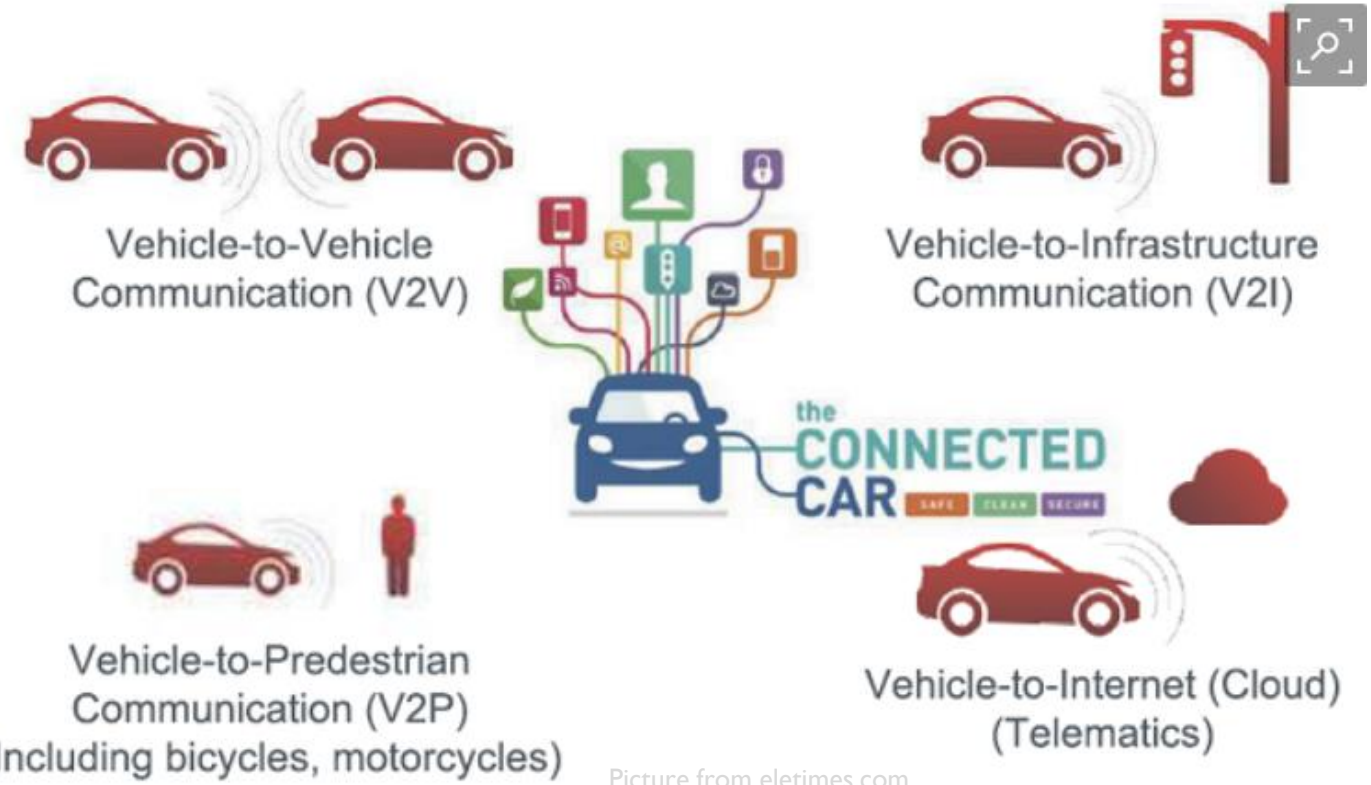
Brief Introduction to Data Mining

Philippe Fournier-Viger
http://www.philippe-Fournier-viger.com

Introduction

- Nowadays,
  - **storing data** on computers is cheap.
  - **transferring data** between computers is fast,
  - Small and cheap devices can collect a lot of data such as **smartphones** and other **sensors**.

Data collected from vehicles

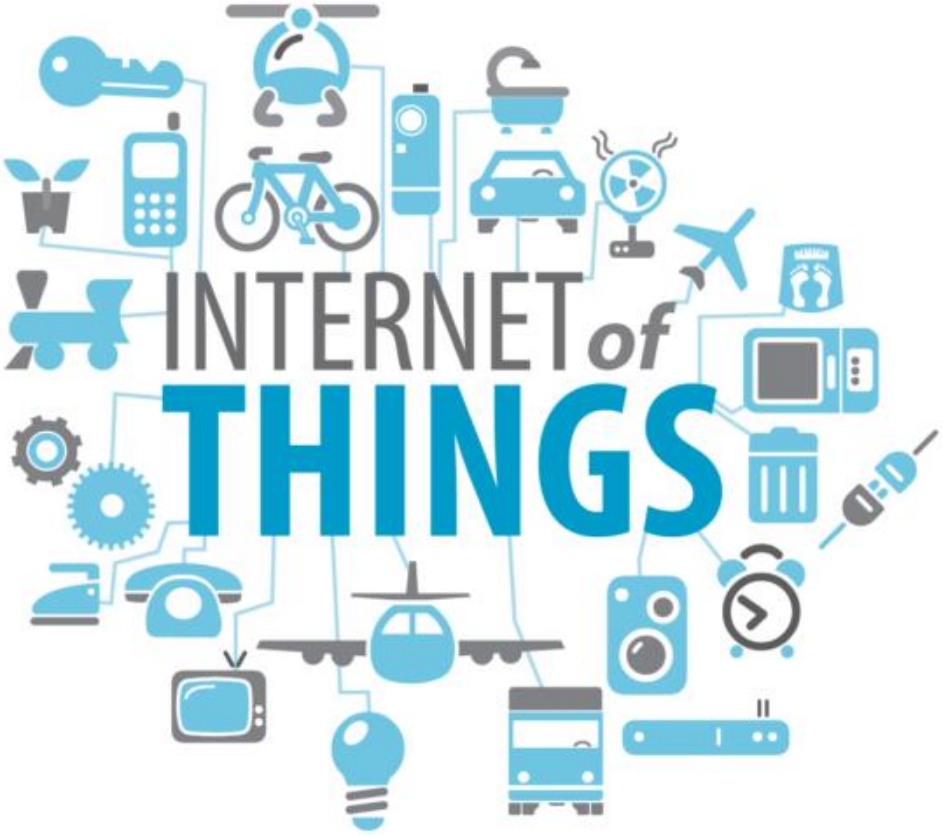


Picture from eletimes.com

Picture from ebay.co.uk

Internet of things (物联网)

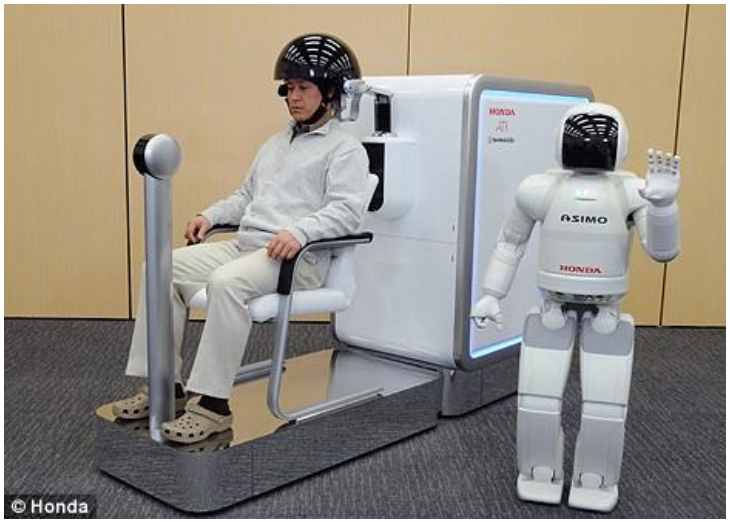
- Different objects can communicate and exchange data.
- ~30 billion connected objects in 2020 (Nordrum 2016)



Picture from teamarin.net

Data collected from humans

- Movements,
- Brain signals,
- Skin conductivity (皮肤电导率),
- Heart rate (心率),
- Blood pressure (血压),
- Eye movements,
- Spatial locations,
- …



Data collected from the industry

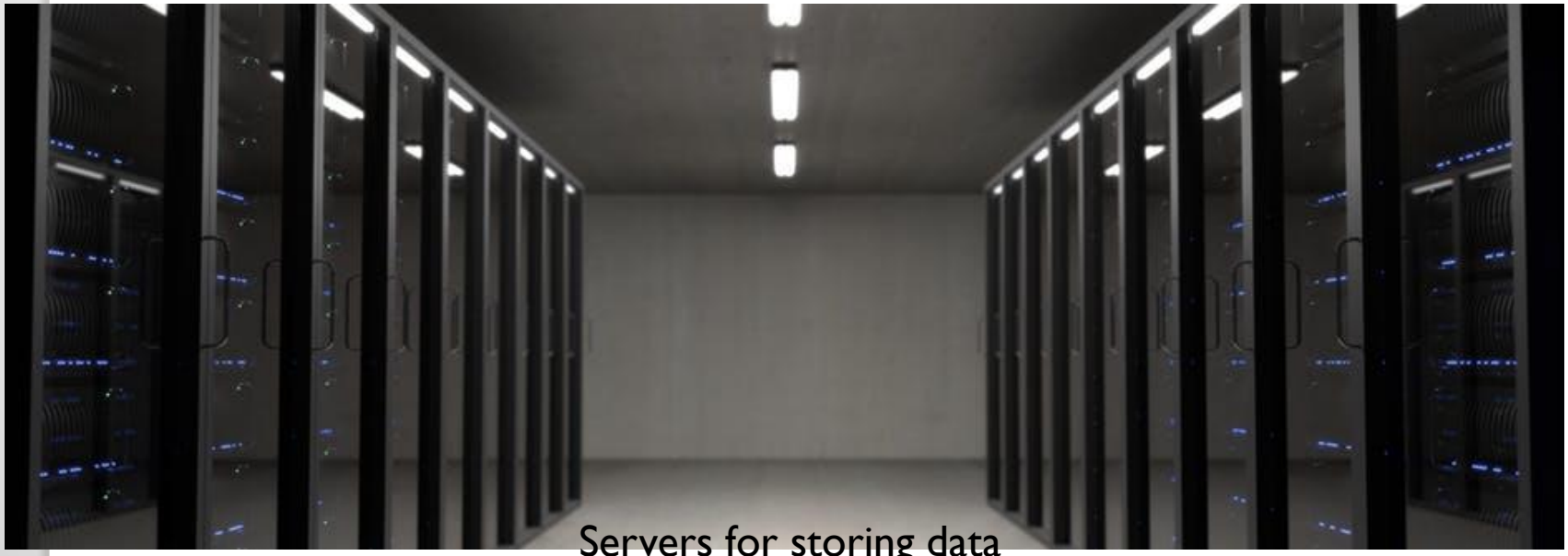
- Internal data
  - Data about employees, customers, market, etc.

- Banking:
  - Spending patterns, Income, social media, …
- Retail industry:
  - Purchases, reviews of products, customer behavior, surveys, advertisement campaign data…
- …

Introduction

As a result, **a huge amount of data** is collected and stored in databases.

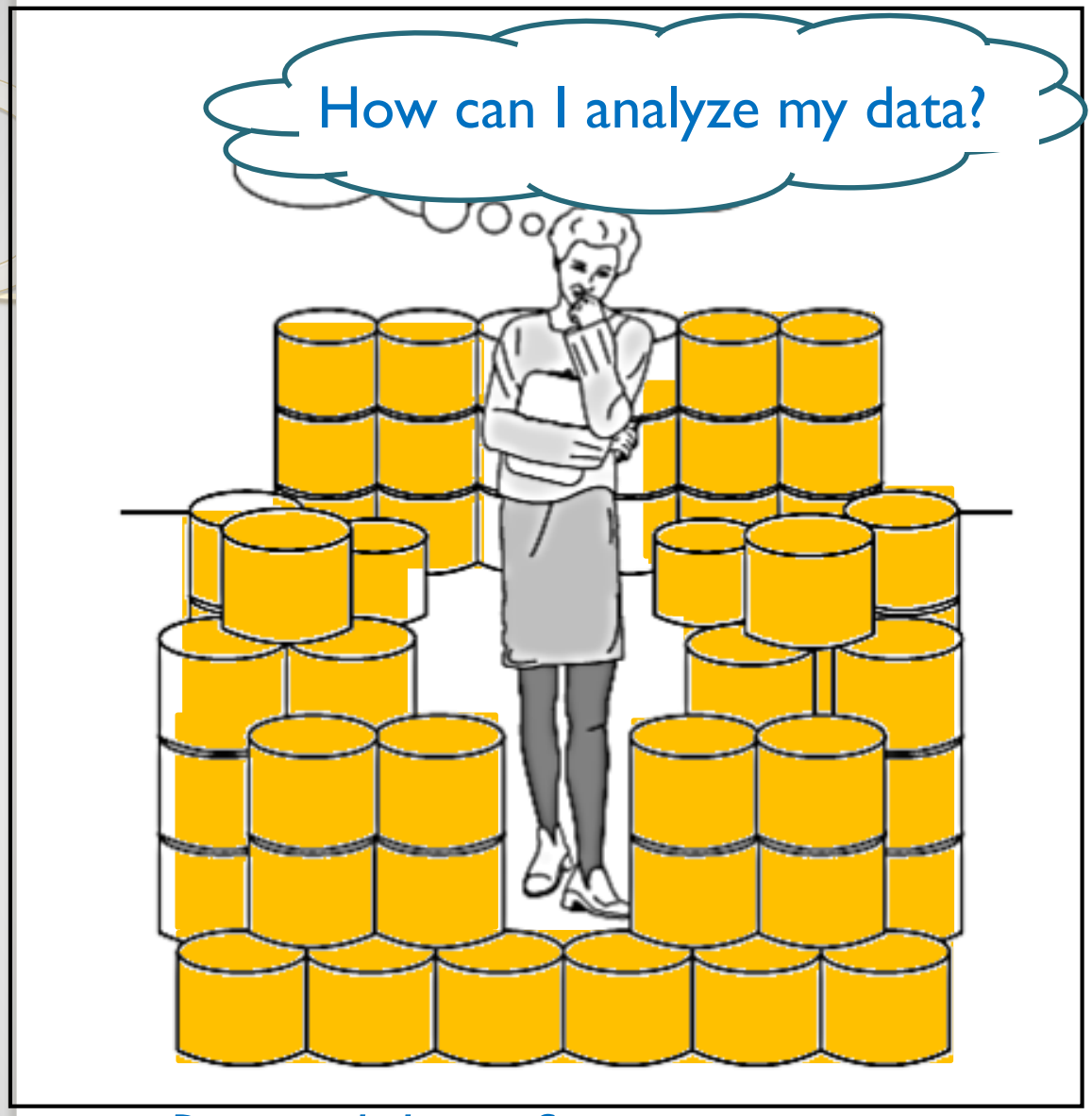
« Big Data » (大数据)



Servers for storing data

Introduction

- Having a lot of data is **great** ☺
- But we want to be able to **understand** the data.
- We also want to discover **new knowledge** that can help us understand the data and support **decision making (决策).**
- If we cannot do that, the data is useless…



« *Data rich but information poor* »

Illustration: Han & Kamber (2006)

Analyzing data by hand ?

- time-consuming
- may miss important information
- Not suitable for "big data".

What is data mining (数据挖掘) ?

- **Data mining** consists of **techniques** to automatically **discover interesting patterns in data (discover knowledge)**.

- **Two goals:**
  - Understand the past

    e.g. Why there was an earthquake (地震) last year ?
  - Predict the future

    e.g. Will there be an earthquake (地震) tomorrow?

    e.g. Will this customer pay back his debt (债务)?

How to do data mining?

- To do data mining, a **process** is followed, consisting of seven steps →

- This process is often called « Knowledge Discovery »
  (数据库中的知识发现)

- Data mining is only one step of this process.

The Knowledge Discovery process (数据库中的知识发现)

Preparing Data (数据预处理)

1. **Data cleaning** (数据清洗) (remove noisy data and fix inconsistencies)
2. **Data integration** (数据集成) (integrate data from multiple sources)
3. **Data selection** (数据选择) (select relevant data)
4. **Data transformation** (数据转换)

Discovering patterns

5. **Discovering patterns (*data mining*)**

Evaluating patterns and using them

6. **Evaluate the patterns** found using interestingness measures
7. **Visualize** the discovered knowledge

Data Mining techniques

- In general, there are many **techniques for analyzing data**.
- Data mining techniques are generally applicable to large volumes of data.
- Many different techniques:
  - to analyze different types of data,
  - to discover different types of knowledge, to be used in different ways.

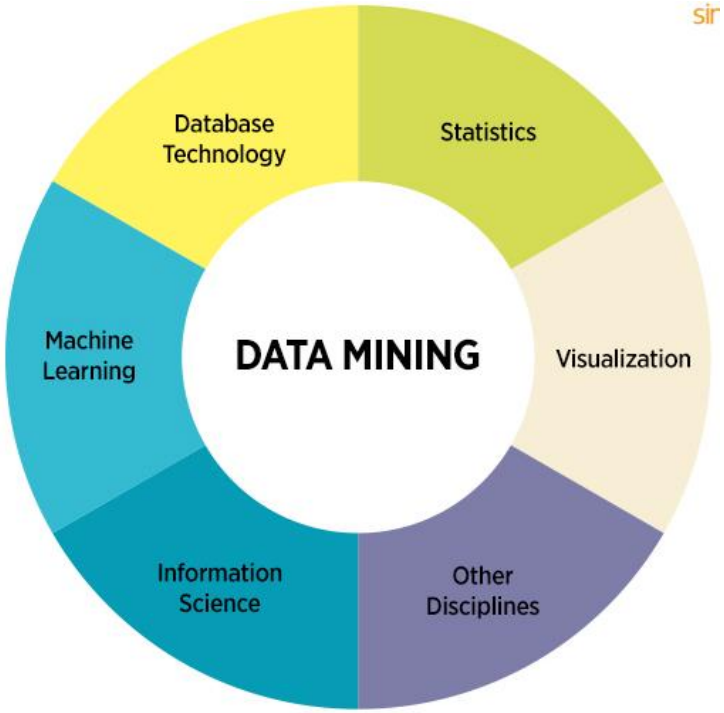
What are the applications of data mining?

A few examples:

- Fraud detection (欺诈检测)
- Analyzing trends on the stock market (股票市场趋)
- Analyzing the behavior of customers in terms of what they buy (市场篮分析).
- Recommending products to customers on online retail stores (e.g. JD.com)
- Identifying people in a crowd or at store

Data Mining is an interdisciplinary research field

- Database systems (数据库系统),
- Algorithmic (算法),
- Computer Science (计算机科学),
- Machine Learning (机器学),
- Data vizualization (数据可视化),
- Image and signal processing,
- Statistics (统计),
- etc.
- Applications: design…



Data mining vs Statistics

What is the difference between **data mining** and **statistics?**

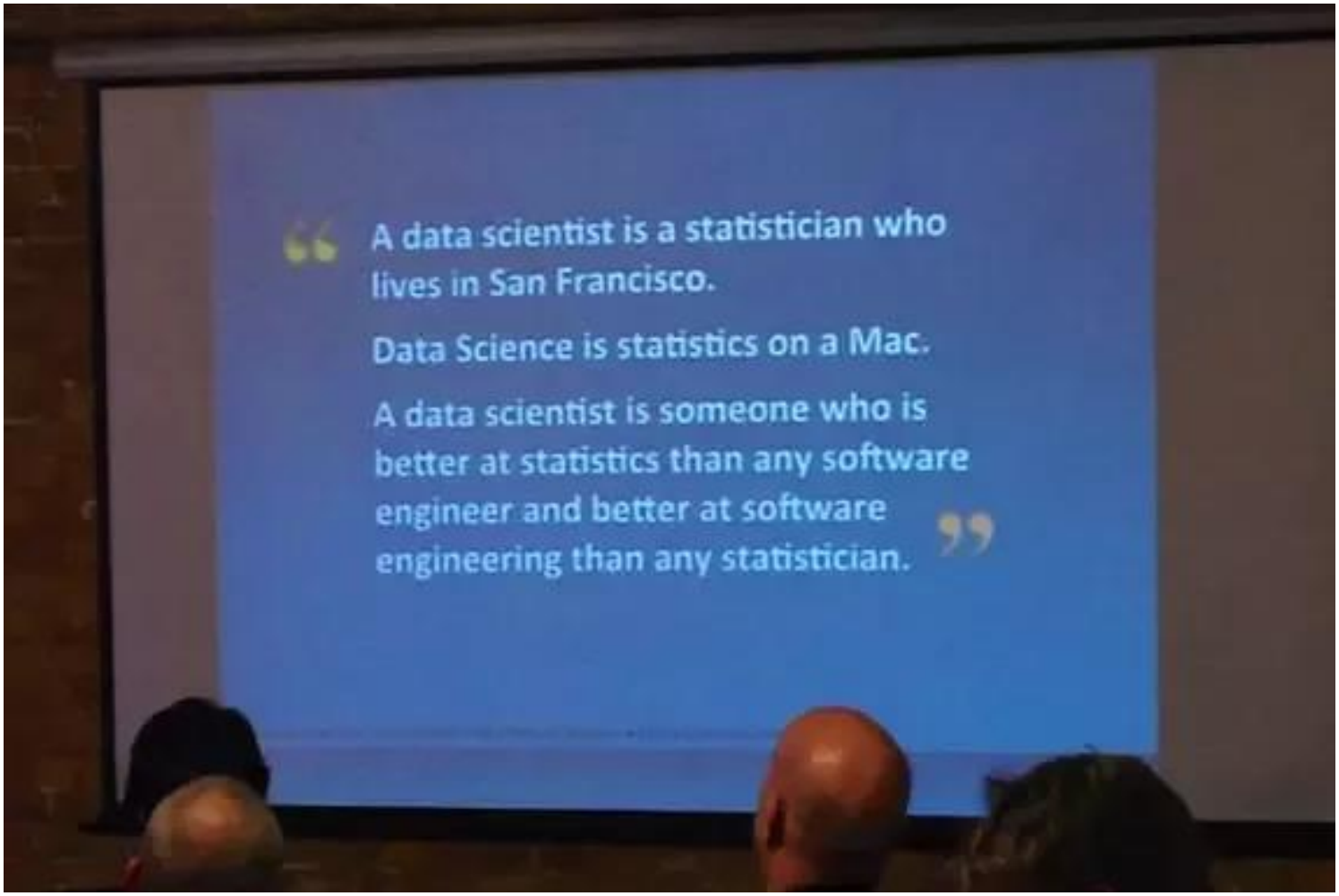
- **Descriptive statistics** (描述性统计) is about describing data.
- **Inferential statistics** (推论统计) is about testing hypothesis,
  - the goal is to draw significant conclusions

Data mining vs Statistics

- **Data mining** focuses on the automated discovery of unknown properties of the data (trends, anomalies, correlations…).
  - the end result is what is important.
- **Statistical learning** (统计学习): this term is sometimes used to describe data mining techniques.

"Data science"

"**Data science**" is another recent term, often use to refer to statistics and/or data mining, but generally with a more multidisciplinary meaning.



Why using data mining?

- To take **decision based** on **facts rather than** based on **intuition** (直觉).
- To **avoid analyzing data by hand**, as it is time-consuming and may result in errors.

Data mining software (数据挖掘软件)

Some popular software programs are:

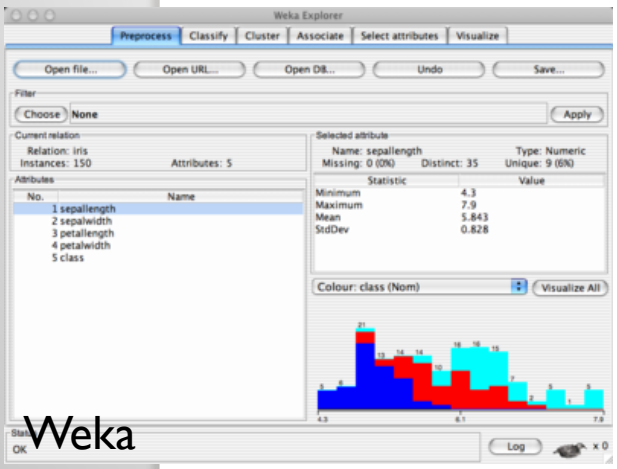
- **Weka**: free, open-source
- **Knime**: free/commercial, open-source
- **R**: a language widely used for data mining and statistics
- **SPMF**: free, open-source (my software)
- **SAS**: commercial software for statistics
- … and many others

They have different features, advantages and limitations…

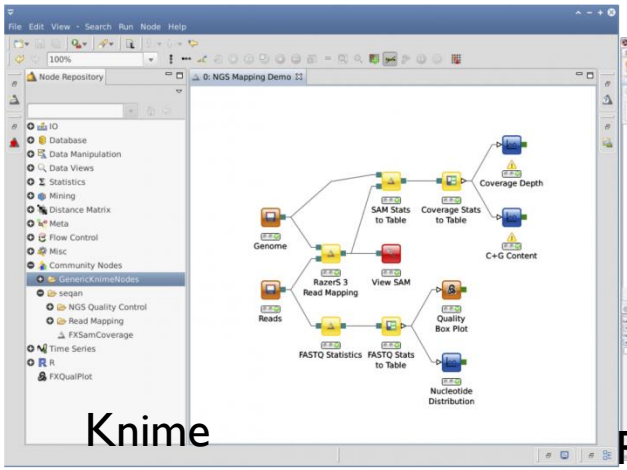
Data mining software (数据挖掘软件)

Typical features of a data mining software:

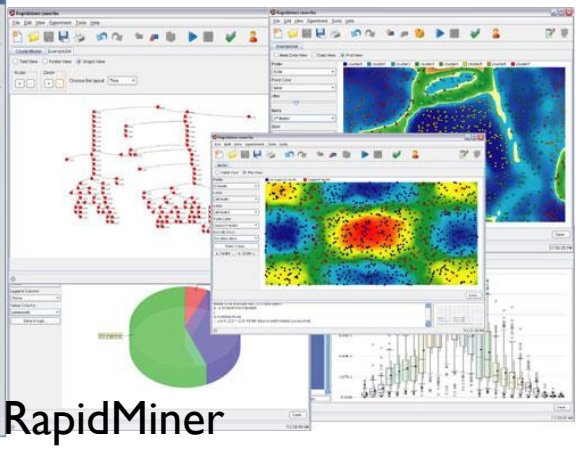
- User interface（用户界面)
- Read different types of data (files, databases…)
- Prepare the data for analysis
- Provide several algorithms to analyze the data
- Data visualization

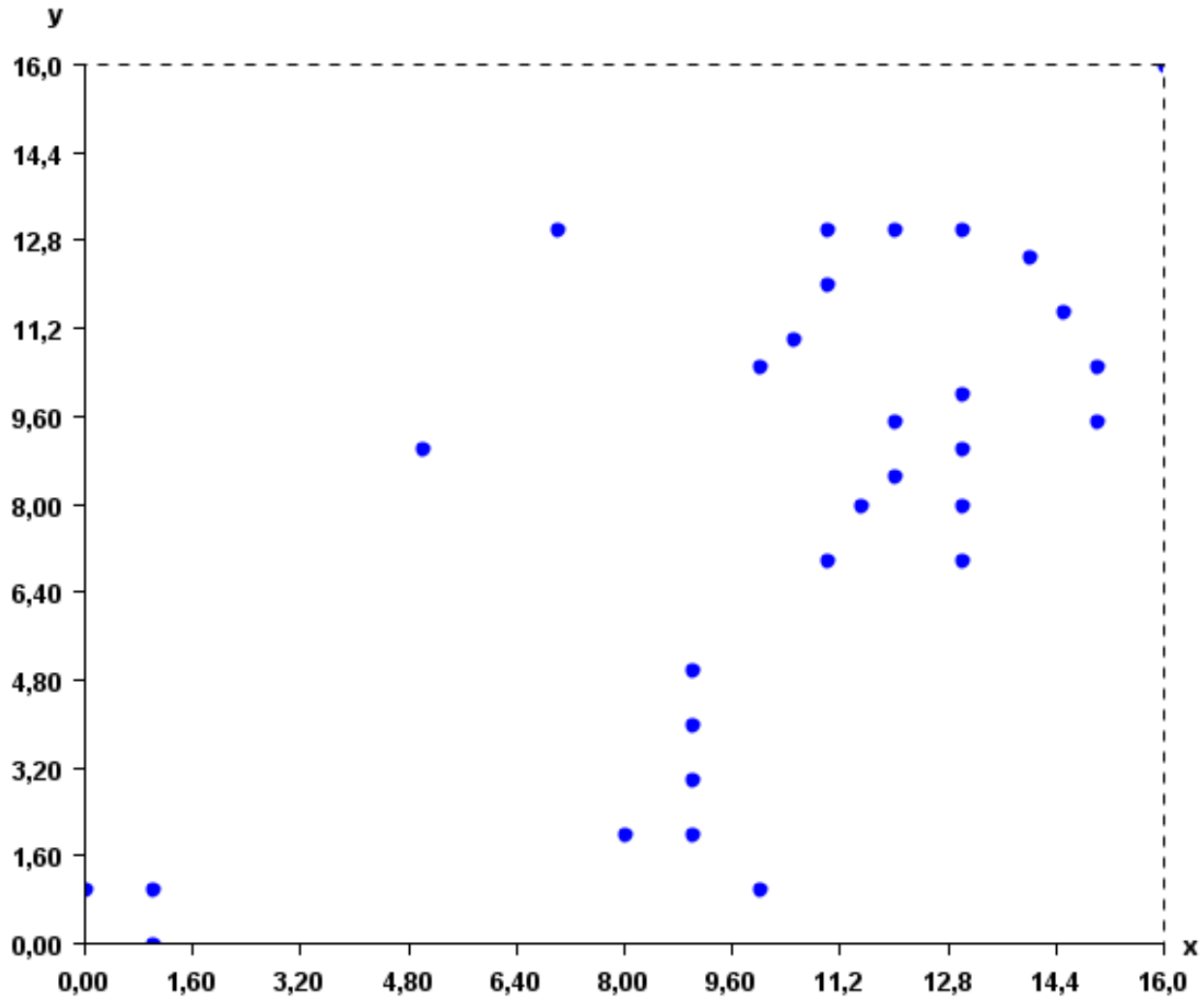


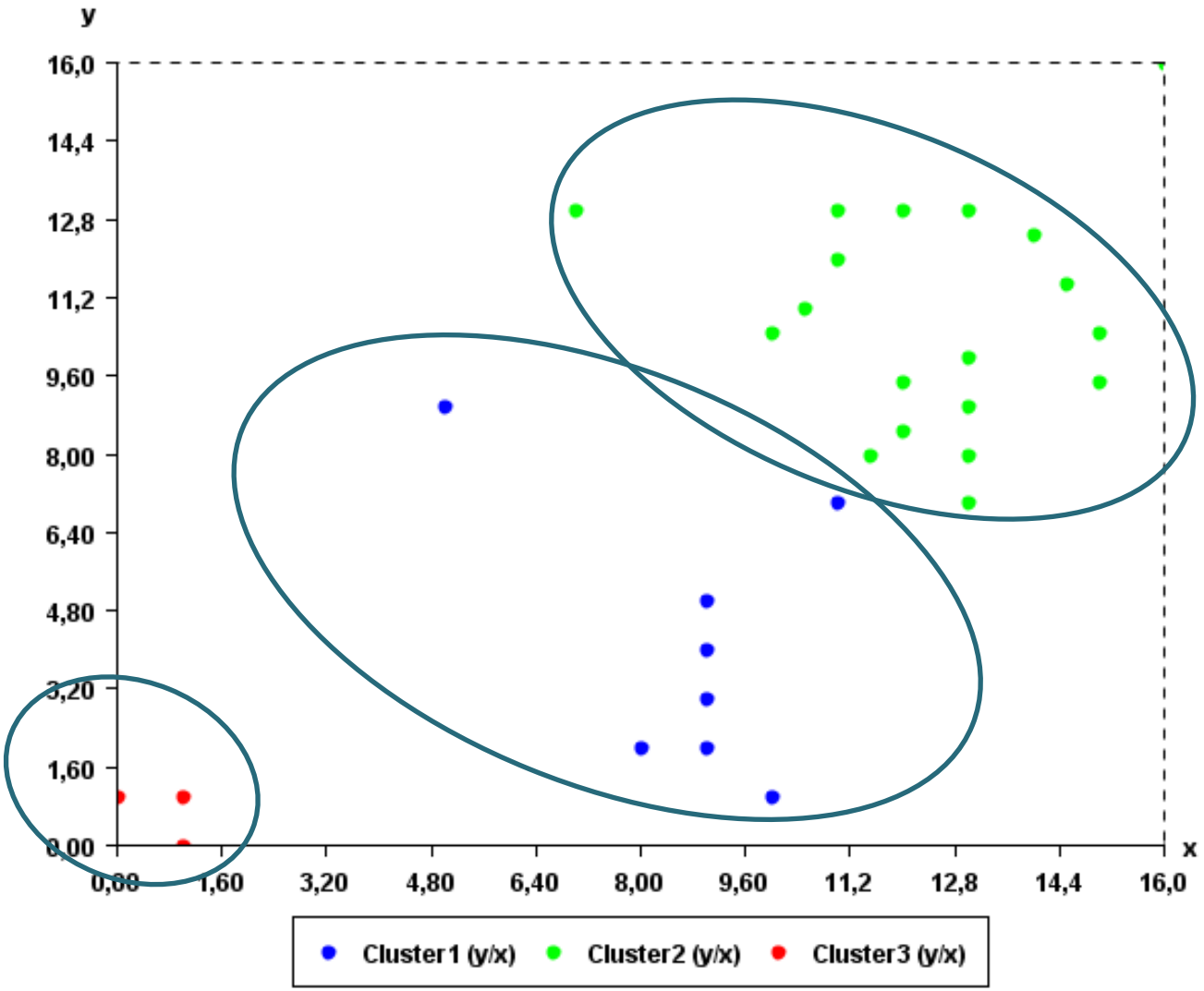
Weka



Knime


RapidMiner


y
16,0
14,4
12,8
11,2
9,60
8,00
6,40
4,80
3,20
1,60
0,00
x
0,00
1,60
3,20
4,80
6,40
8,00
9,60
11,2
12,8
14,4
16,0



clustering (群集)

Data mining software

- Several data mining techniques are designed to be applied on **huge databases**.
- But they can also be applied on **small databases**.

- Data mining techniques can be applied on **various types of data** →

VARIOUS TYPES OF DATA (数据类型)

Relational database (关系数据库)

In a typical **database system (数据库系统)**, data is organized as tables:

Patient Table

Patient Id	Name	D.o.B	Gender	Phone	Doctor Id
134	Jeff	4-Jul-1993	Male	7876453	01
178	David	8-Feb-1987	Male	8635467	02
198	Lisa	18-Dec-1979	Female	7498735	01
210	Frank	29-Apr-1983	Male	7943521	01
258	Rachel	8-Feb-1987	Female	8367242	02

Doctor Table

Doctor Id	Doctor	Room
01	Dr Hyde	03
02	Dr Jekyll	06

Traditional database systems allows to search information in databases (**e.g.** finding all patients that are male and >20 years old)

Data mining allows do to more. It allows to find correlations, trends, and other types of complex knowledge in data (**e.g.** finding that young patients are more likely to cure using a given treatment…)

Transactional data (事务型数据库)

- A database of customer transactions (客户交易).
- A transaction is a list of items bought by customers.
- **Example:**

TID	Bread	Milk	Noodles	Eggs	…
1	X		X		
2		X	X		
3	X	X	X		

- May contain additional information
  - **e.g.** purchase quantities, unit price, time, location…

Temporal data

Sequences:

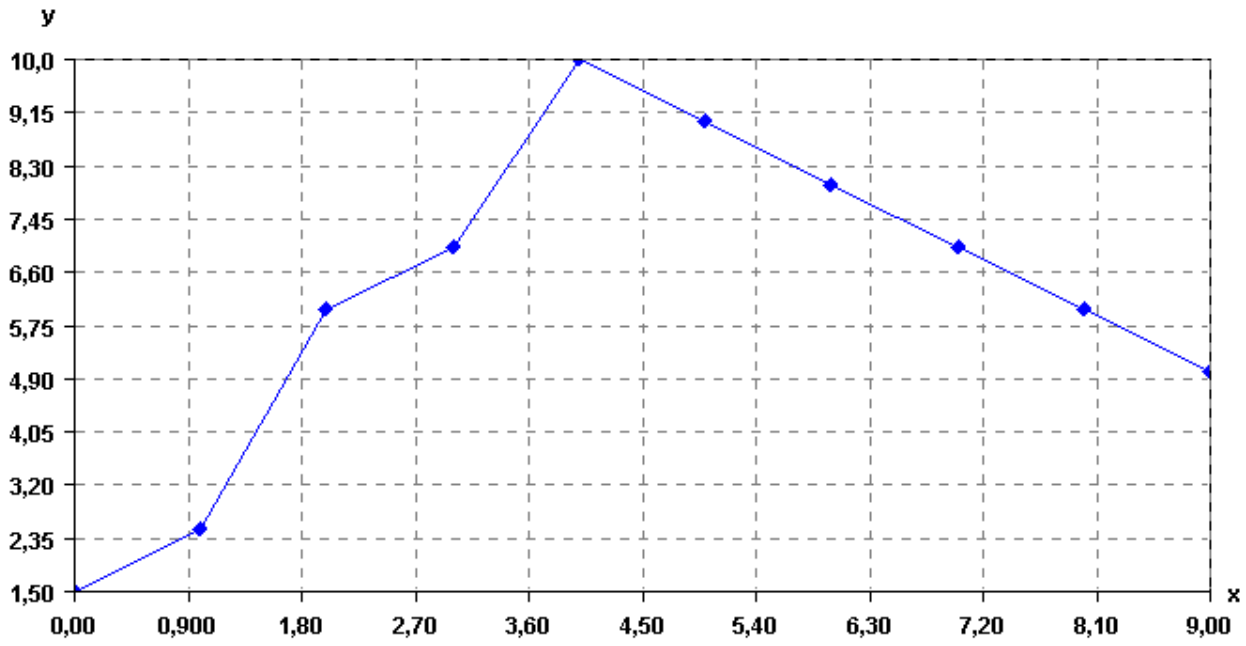
series of symbols: a, b, c, b, a, c, d ,a

- Sequences of clicks on a website:
  Page1, Page2, Page4, Page1…
- Protein sequences (蛋白质序列)
- Sequences of moves when playing chess (国际象棋)
- Sequence of GPS locations

Temporal data

Time series:

- a series of numeric values,
- usually obtained at a regular interval.
- **e.g.** stock market data, EEG data, temperature data, student grades over time…



Spatial data

- **Spatial or geographic data**
  - **e.g.** forestry (林业), ecology (生态学), infrastructure management (基础设施管理)
- **Spatio-temporal data**
  - spatial and temporal data
  - **e.g.** meteorological data (气象数据), crowd movement (人群的运动), bird migration (鸟的迁徙)

Text data

Text documents:

A type of unstructured data (非结构化数据): documents that have no clear structure, or are not organized in predefined manner.

Examples:

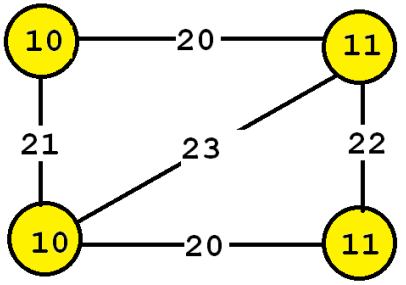
- Predicting if someone will like a movie or product
- Analyzing an anonymous text to find the likely author. How old he is ? What is the author profile?
- Sentiment analysis
- Automatic summarization of a document

Web data

- **Web**:
  - a set of documents (webpages)
  - links between documents
- **Examples:**
  - Predicting the next webpage that someone will visit
  - Automatically grouping webpages by topics into categories.
  - Analyzing the time spent on webpages
  - Analyzing data from attacks by hackers on a website.

Graphs

- **Social networks (社交网络),**
  - Finding communities
  - Analyzing the relationships between people
  - Predict who will be your friend
  - Observe how communities evolve
  - Find who has the most influence
  - Find the location of a person
  - Infer the psychological profile of a person
  - Study trends
- **Chemical molecules…**



Heterogeneous data (异构数据)

- Sometimes, we need to analyze data combining multiple types of data (**e.g.** spatial, temporal, time series, text, GPS, etc.)
- We may also need to analyze data stored using different technologies and file format (**e.g.** Excel files, text files, Word documents, pictures, videos, GPS data, audio).

Data streams (数据流)

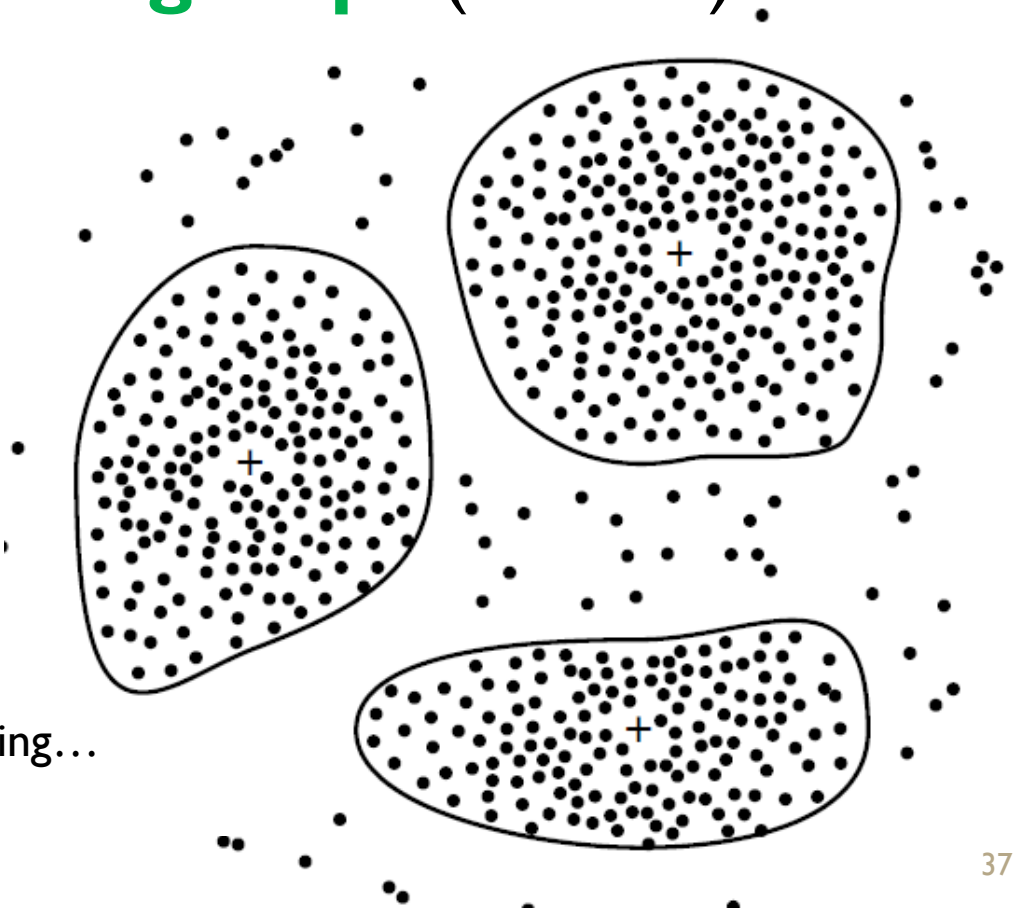
- **Data stream**: a high-speed and non-stop stream of data that is potentially infinite
- **Ex.:** satellite data, video cameras, environmental data.
- **Challenge**: must be analyzed in real-time
- **Needs**:
  - extract summaries of data
  - detect changes (**ex.:** trends, detect changes),
  - evaluate the state of a stream

We may want to extract different types of "patterns" (模式) from data.

TYPES OF PATTERNS

Clusters

Clustering (群集): consists of automatically grouping similar objects/instances into **groups** (*clusters*) of similar instances.



Examples:

- Hospital patients having a similar profile
- Individuals who are likely to develop dependencies to gambling
- taxonomy of animals
- Students with similar learning profile

Use to summarize data, for decision making…
We want to discover « natural » clusters

Classification (分类)

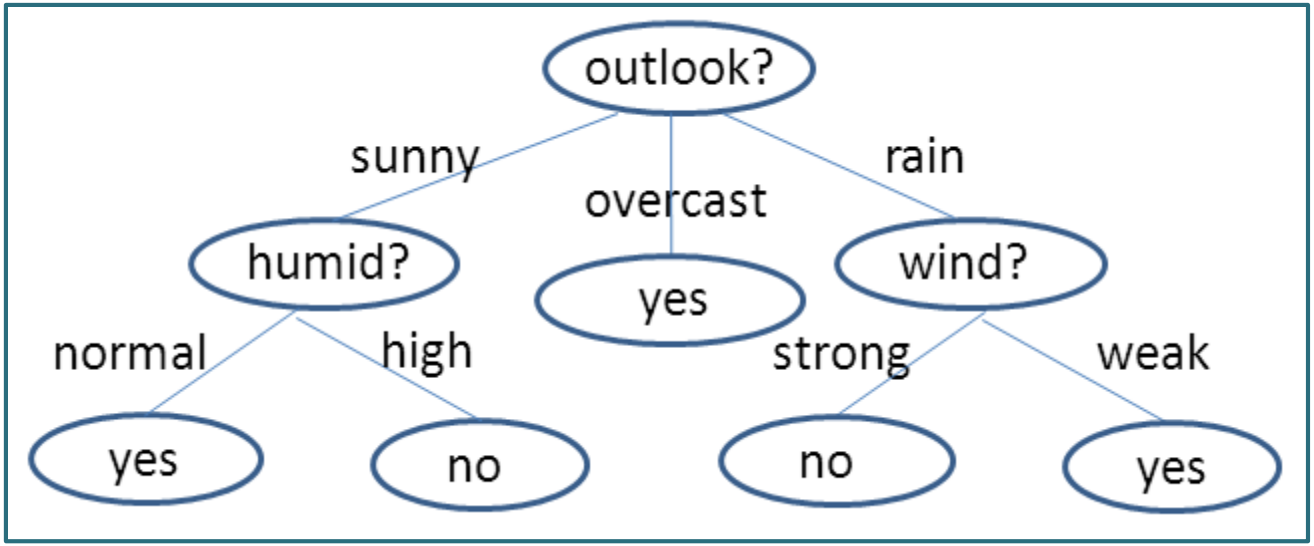
- **Classification**: build a model that can automatically classify instances into different categories/classes.
- **Several applications**
  - predict who will pay back their debt and who will not,
  - predicting who will fail/pass a course,
  - Handwriting character recognition (手写字符识别)
- **Several techniques**:
  - Neural networks (神经网络), SVM, decision trees (决策树), Naïve Bayes classifier (素贝叶斯分类器), etc.

e.g. ID3 decision tree (决策树)

Training data (训练数据)

outlook	temp	humid	wind	play?
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

A decision tree to predict the « play? » attribute



A prediction:

sunny	hot	normal	weak	?

⇨ yes

Understand groups of instances

- **Find what is common to a given group of instances**
  - Characteristics of successful movies
  - Characteristics of tourists in Shenzhen
  - Characteristics of people who spend the most money in a given restaurant.
- **Find what distinguish two groups of instances**
  - What distinguishes successful movies from unsuccessful ones.
  - What distinguishes poisonous mushrooms from those that are not.

Discovering patterns

- **Discovering values that appear frequently together in the data:**
  - 30 % of the tourists visiting Shenzhen are less than 30 years old and have a university degree.
- **Discovering strong associations in data:**
  - There is a 60 % conditional probability that tourists visiting the Guangdong province will also visit Shenzhen.

Anomalies, outliers (离群值)

Detecting what is abnormal (**anomalies, outliers**) is interesting and has many applications.

e.g.

- detecting hackers attacking a computer system,
- identifying potential terrorists based on suspicious behavior,
- detecting fraud on the stock market

Trends, regularities, periodic patterns ….

Several applications:

- study patterns in the stock-market
  - to predict stock prices and take investment decisions.
  - to understand the past.
- discovering regularities to predict earthquake aftershocks,
- find cycles in the behavior of a system,
- discover the sequence of events that lead to a system failure.

FINDING INTERESTING PATTERNS IN DATA

Data mining techniques are designed to extract interesting patterns and knowledge from data

How to evaluate the patterns of knowledge found in data to ensure that it is interesting and useful?

Finding interesting patterns

- Data mining techniques can find **millions of patterns** in data.
- As humans, we do not want to analyze millions of patterns.
- Thus, we need to filter patterns to obtain a set of **patterns** that is **interesting** or **useful**.
- To evaluate patterns, different measures are used in data mining.
- Evaluating patterns can be during data mining or **after** (as post-processing).

What is an interesting pattern?

- **A pattern is interesting if:**
  - it easy to understand,
  - it is still valid for new data;
  - It is useful;
  - It is novel or unexpected.
- **Several measures:**
  - **objective measures:**
    - **e.g.:** how frequently a pattern appears
  - **subjective measures**:
    - **e.g.** how interesting a pattern is for a person

Conclusion

In this part, I have introduced :

- the topic of **data mining,**
- different **types of data,**
- different **types of patterns,**

References

Some content from these sources:

- Data Mining: The Textbook by Aggarwal (2015)
- Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014)
- Han and Kamber (2011), Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann Publishers,
- Tan, Steinbach & Kumar (2006), Introduction to Data Mining, Pearson education, ISBN-10: 0321321367.