Data Mining and Big Data

m5211143 - Tomohiro Saito

Database Systems Laboratory, 141-F

Data Mining (DM)

> Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
>
> Quoted from: Data mining - wikipedia

Variations

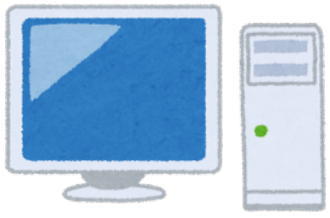
- Text mining -> text
- Web mining -> web resource

Big Data

Big data is a term for data sets
that are so large or complex that traditional data
processing application software is inadequate
to deal with them.

Quoted from: Big data - wikipedia

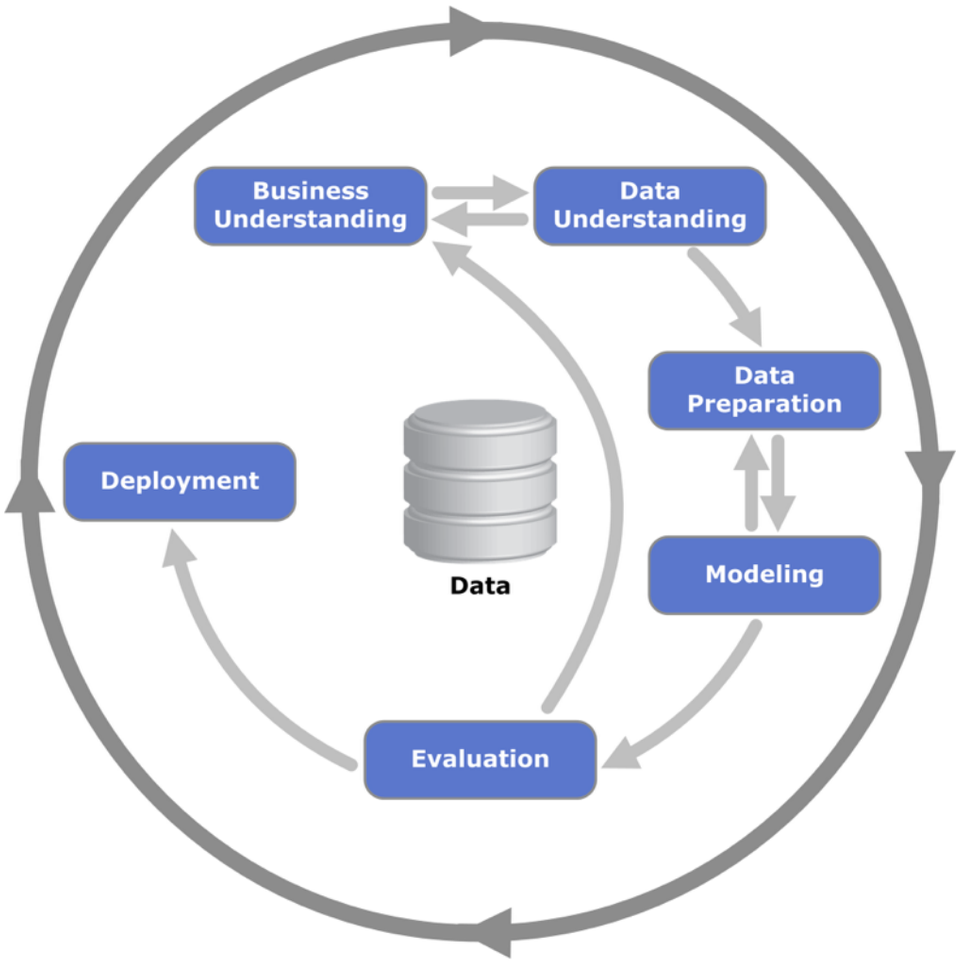
customer data



DB, server data

SNS

Cross Industry Standard Process for DM(CRISP-DM)



Why do we need?

- Framework for recording experience
- Aid to project planning and management
- “Comfort factor” for new adopters

We can advance the project according to this process.

Data mining and its advantage

Lots of data exist today, for example,

- Customer data in company
- Web data

Advantages to apply data mining to such big data is

- Reduce the cost for analyzing data
- Saving time and get more information from data

Data mining techniques (Example)

- Frequent pattern extraction

  → A method for extracting patterns appearing in a data set

- Regression analysis

  → A statistical process for estimating the relationships among variables

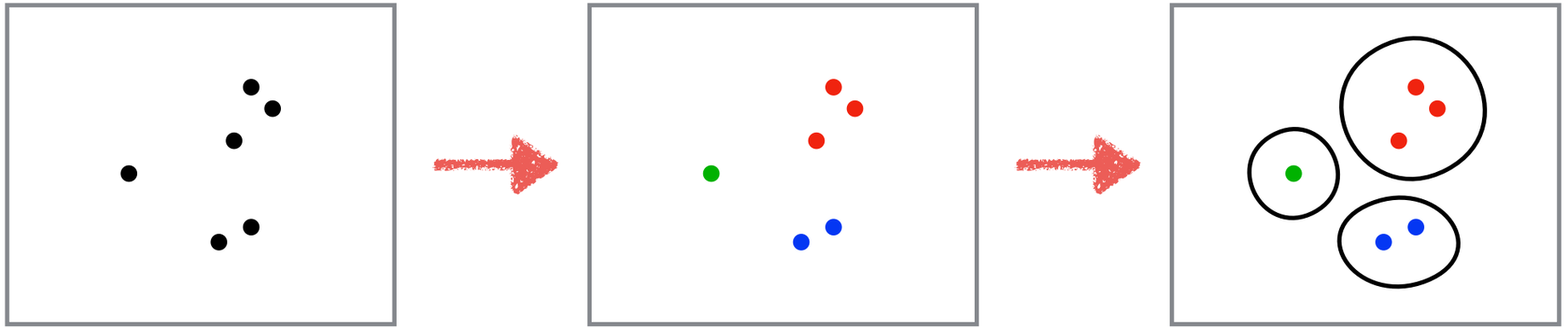
- **Clustering**

  → A method for extracting patterns appearing in a data set

What is Clustering?

- Clustering

  → A method for extracting patterns appearing in a data set



For example, three clusters are created according to Euclidian distance in above dataset.

Example. Clustering information resource in AOJ

There are many source codes in AOJ. By using Data mining method, we can analyze

- Characteristics of source codes
  - typical mistakes for solving problems
  - writing habit of user
  - etc…
- Problems in AOJ
  - Data manipulation
  - String processing
  - etc…

Problems applying data mining to big data

When we use data mining, we need to decide before clustering:

- Suitable method
  - Which algorithm is applicative for the dataset?
- Data selection
  - Which data would be usable for analysis?
- Accuracy
  - How to calculate the correctness of results which obtained by data mining?