

# 데이터 수집

2023년 1월  
현장프로젝트 교과  
허재석

# Data Preprocessing

## 데이터 과학의 현실

“데이터 과학의 80%는  
데이터 클리닝에 소비되고,  
나머지 20%는  
데이터 클리닝하는 시간을  
불평하는데 쓰인다.”

- Kaggle 창립자·CEO Anthony Goldbloom



진짜임.

[www.facebook.com/datagentoo](http://www.facebook.com/datagentoo)  
google에서 '데이터펍권'



# Data Collection

## Data Analysis Process

- Data collection
  - Given, purchase, experiment, open datasets, web crawling, ...
- Data preprocessing
  - Noise removal / outlier detection
  - Join / aggregation
  - Feature engineering / embedding / vectorization
  - Dimension reduction
- Data explorations
  - Data summary/visualization
  - Correlation analysis
- Problem solving
  - Descriptive: rule mining, clustering
  - Predictive: regression, classification

# Data Collection

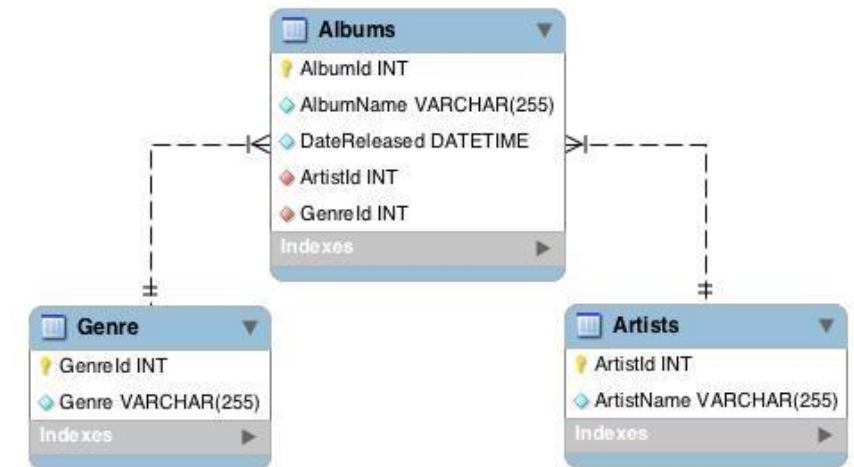
## Open Datasets

- Lots of open datasets
  - 모델 기초 학습을 위한 데이터로 사용 가능
- Domestic
  - AI hub <http://www.aihub.or.kr/>
  - 공공데이터포털: <https://www.data.go.kr/>
- Traditional
  - ML repository: <http://archive.ics.uci.edu/ml/index.php>
  - MNIST: <http://yann.lecun.com/exdb/mnist/>
- Others
  - Fashion-MNIST: <https://github.com/zalandoresearch/fashion-mnist>
  - Open Images Dataset: <https://opensource.google.com/projects/open-images-dataset>
  - Kaggle: <https://www.kaggle.com/>

# Data 속성

## Data Basics

- Data size
  - 건, 용량(GB)
  - # of instances
  - # of attributes
- Data schema(meta info.)
  - The skeleton structure that represents the logical view of the entire database  
(\*source: [https://www.tutorialspoint.com/dbms/dbms\\_data\\_schemas.htm](https://www.tutorialspoint.com/dbms/dbms_data_schemas.htm))
  - Attribute list and types
  - Relation between tables
  - Database → set of tables



# Data 속성

## Data Types

Data Type		Possible values	Example usage
Categorical	binary	0, 1 (arbitrary labels)	binary outcome ("yes/no", "true/false", "success/failure", etc.)
	categorical	1, 2, ..., K (arbitrary labels)	categorical outcome (specific blood type, political party, word, etc.)
	ordinal	integer or real number (arbitrary scale)	relative score, significant only for creating a ranking
Numerical	discrete	nonnegative integers (0, 1, ...)	number of items (telephone calls, people, molecules, births, deaths, etc.)
	continuous real-valued additive	real number	temperature, relative distance, location parameter, etc. (or approximately, anything not varying over a large scale)
	real-valued multiplicative	positive real number	price, income, size, scale parameter, etc. (especially when varying over a large scale)

➔ Categorical & Numeric

# Data 속성

## Data Summary

- Categorical
  - Frequencies
  - Mode
    - The mode of a set of data values is the value that appears most often
  - Cooccurrence
- Numerical
  - Most statistics
  - Mean, median, std., max/min, quartile
  - Correlation

# Data 속성

## Data Types (Format)

- Structured data
  - Has predefined data format (e.g. matrix: instance x feature)
  - Relational Database
- Unstructured data
  - No predefined data format
  - Text data
  - Image data
  - Sequential data
  - Network data
- Semi-structured data
  - Has changeable data format (e.g. Json, XML)



# Data 속성

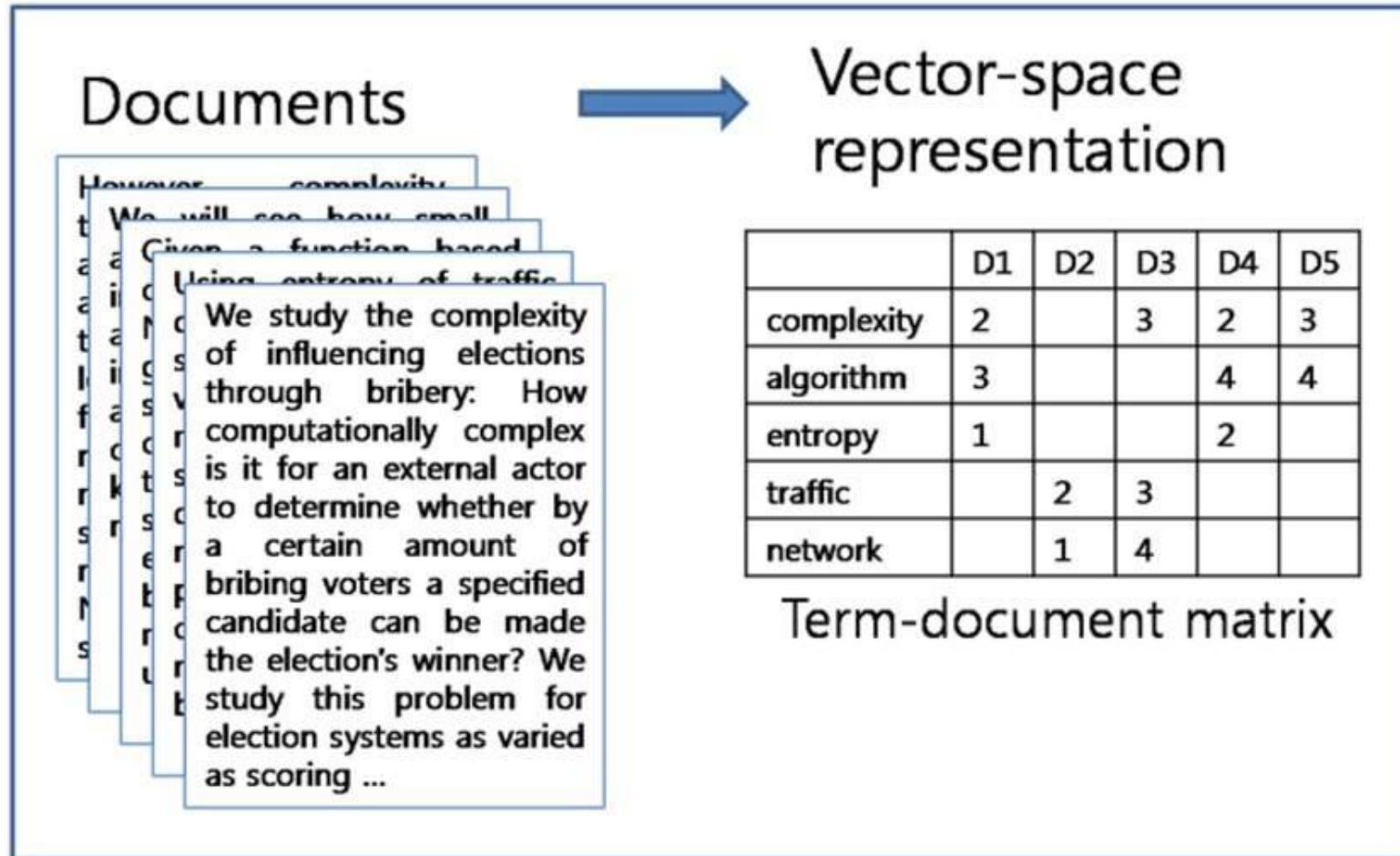
## Unstructured Data

- Text data
  - Composed of strings
  - Sometimes long document
  - Word vector, w2v
- Image / video data
  - List of vectors of (R, G, B) indicating each pixel
- Sequential data
  - Data collected according to time  $t$
  - Signal data, audio signal, brain signal
  - Aggregated features (e.g. mean/std. over 5min.)
  - Few models deals with sequences (e.g. HMM, RNN)
- Network data
  - Node-edge

# Unstructured Data

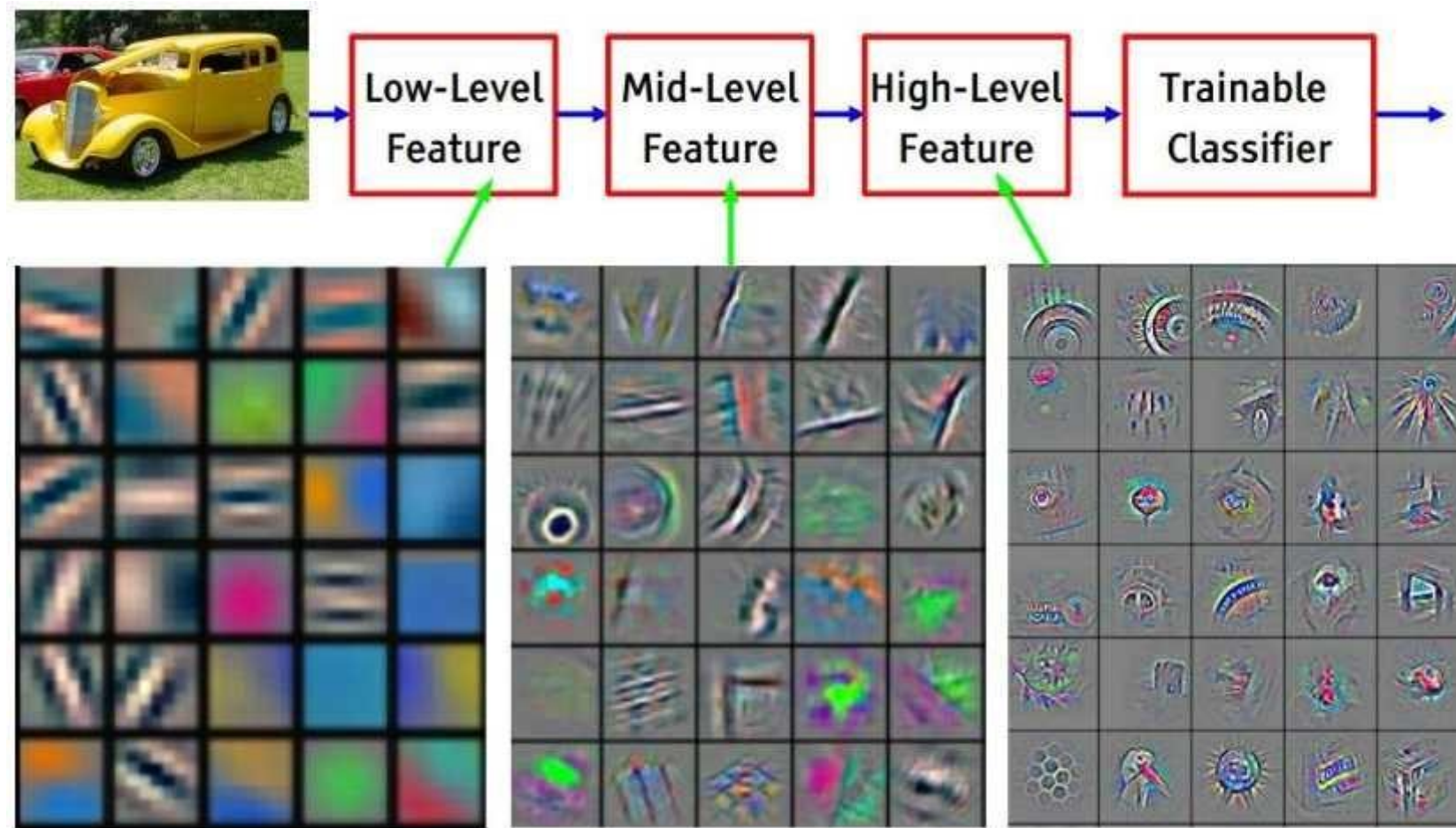
## Text Data

- Document classification



# Unstructured Data

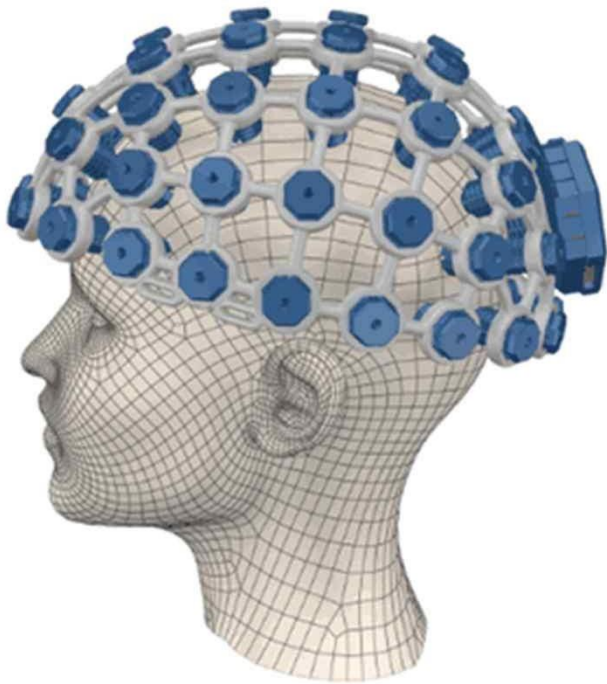
## Image Data



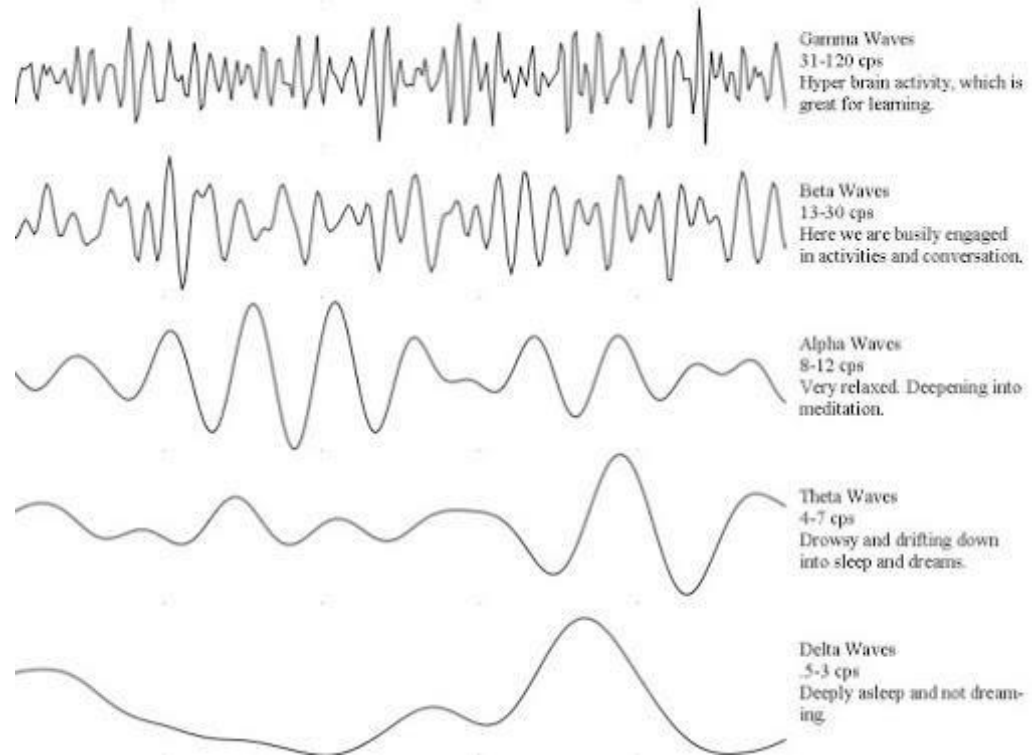
# Unstructured Data

## Sequential Data

- Brain signal processing



Brain Waves Graph






# Unstructured Data

## Others

- Recommender systems
  - Netflix: 10,000,000 users x 17,770 movies matrix

### Frequently Bought Together



Price for all three: **\$74.20**

[Add all three to Cart](#) [Add all three to Wish List](#)

[Show availability and shipping details](#)

☒

**This item:** Beginning Ruby: From Novice to Professional (Expert's Voice in Open Source) by Peter Cooper Paperback **\$27.78**

☒

Learn to Program, Second Edition (The Facets of Ruby Series) by Chris Pine Paperback **\$16.94**

☒

Ruby on Rails Tutorial: Learn Web Development with Rails (2nd Edition) (Addison-Wesley Professional Ruby ... by Michael Hartl Paperback **\$29.48**

### Customers Who Bought This Item Also Bought



Learn to Program, Second Edition (The Facets of...  
Chris Pine  
★★★★★ 42  
Paperback  
**\$16.94** ✓Prime



The Well-Grounded Rubyist  
David A. Black  
★★★★★ 39  
Paperback  
**\$32.49** ✓Prime



Ruby on Rails Tutorial: Learn Web Development...  
Michael Hartl  
★★★★★ 70  
Paperback  
**\$29.48** ✓Prime



The Ruby Programming Language  
David Flanagan  
★★★★★ 74  
Paperback  
**\$26.35** ✓Prime



The Well-Grounded Rubyist  
David A. Black  
★★★★★ 19  
**#1 Best Seller** in Ruby Programming Computer  
Paperback  
**\$29.67** ✓Prime



# Web Data Crawling

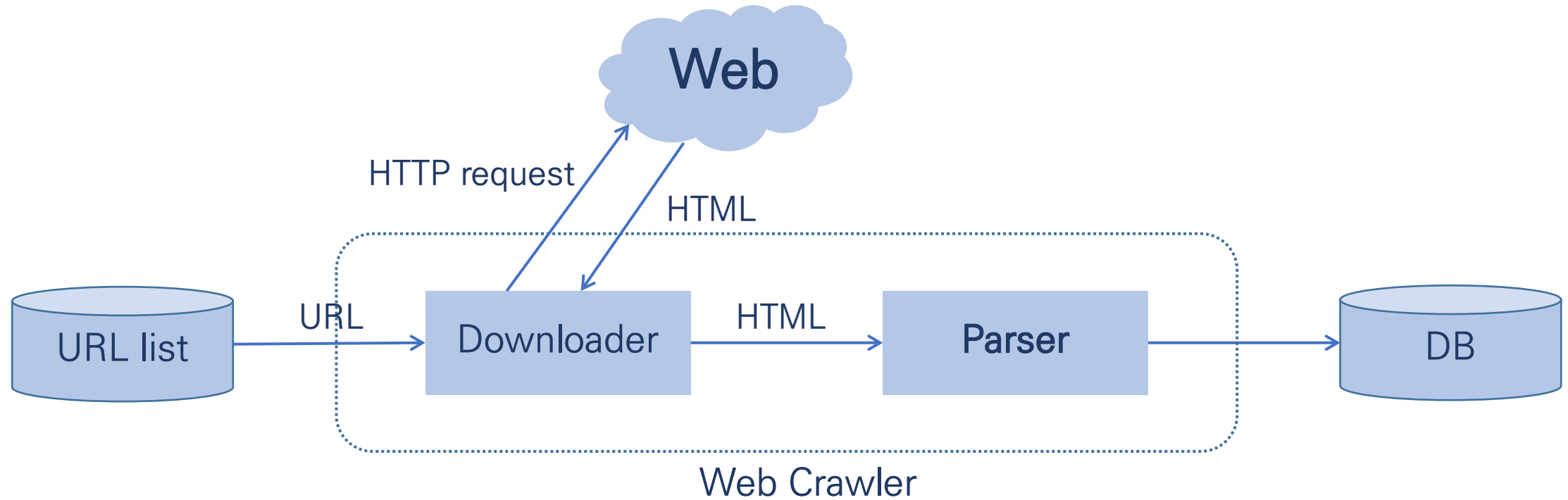
## Crawler

- 정의
  - 웹상의 다양한 정보를 자동으로 검색하고 색인하기 위해 검색 엔진을 운영하는 사이트에서 사용하는 소프트웨어
  - 스파이더(spider), 봇(bot), 지능 에이전트라고도 함
  - 사람들이 일일이 해당 사이트의 정보를 검색하는 것이 아니라 컴퓨터 프로그램의 미리 입력된 방식에 따라 끊임없이 새로운 웹 페이지를 찾아 종합하고, 찾은 결과를 이용해 또 새로운 정보를 찾아 색인을 추가하는 작업을 반복 수행함
  - 방대한 자료를 검색하는 특징은 있으나 로봇의 검색 기능을 역이용하여 순위를 조작하거나 검색을 피할 수 있는 단점도 있음

# Web Data Crawling

## Crawler

- Web crawler 작동 환경



# Web Data Crawling

## HTML (Hypertext Markup Language)

- 하이퍼텍스트\*를 표기하는 언어
- 웹 브라우저를 통해 번역됨
- HTML element로 구성
  - `<tag attribute="value">(content)</tag>`

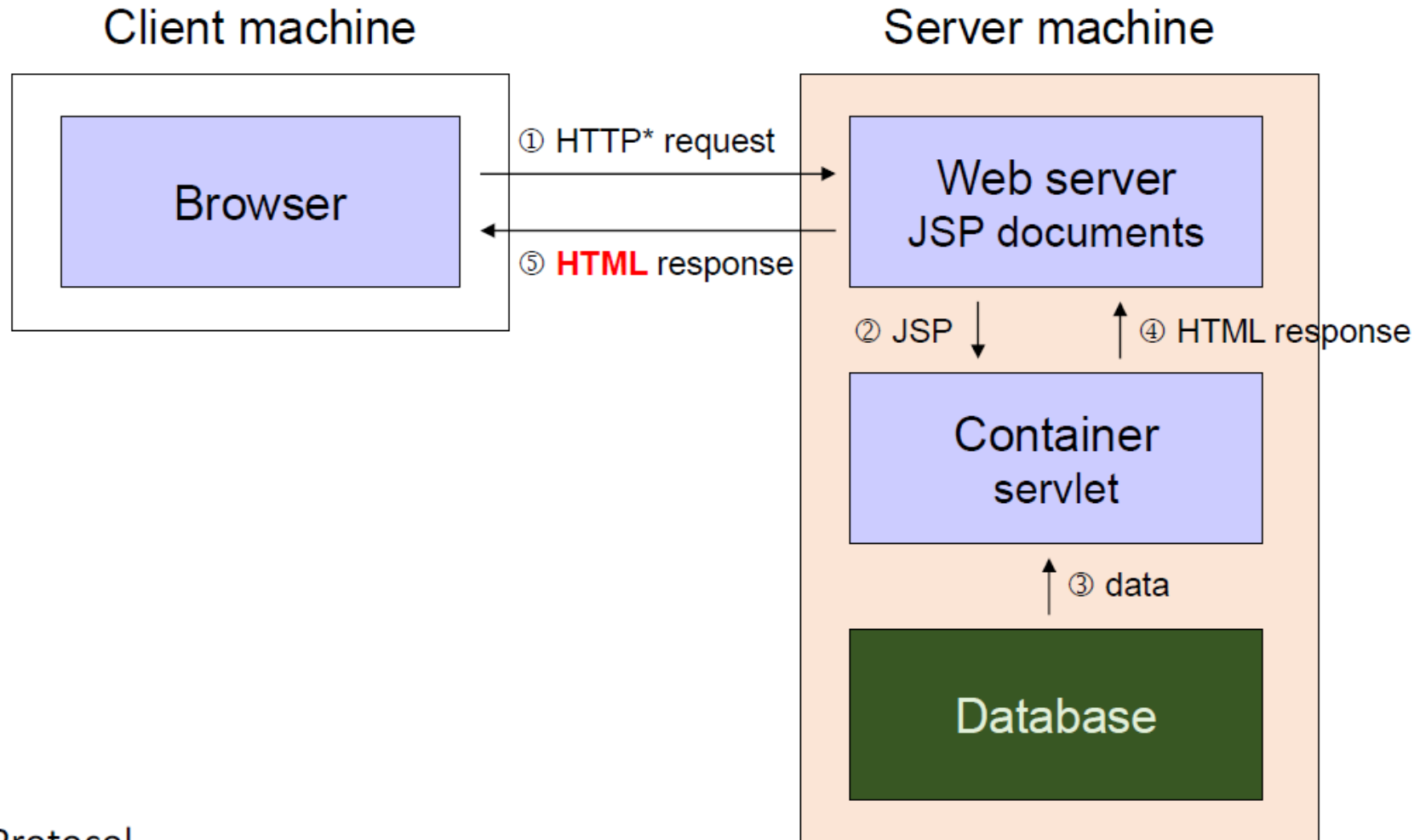
\*Hypertext is text which is not constrained to be linear. Hypertext is text which contains links to other texts. The term was coined by Ted Nelson around 1965 (source : <https://www.w3.org/WhatIs.html>).

```
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p>Hello world!</p>
  </body>
</html>
```



# Web Data Crawling

## HTML 작동 모습



\*Hypertext Transfer Protocol