Data Science

introduction

JUNIOR. SHEUN AHN

Orientation

- Direction : What is data? What is data Science?
- Study with : 김원 SW 중심세상 blog,
  The Data Science Handbook (Field Cady)
- With easy example

1. What is Data?

- Data : raw, valueless

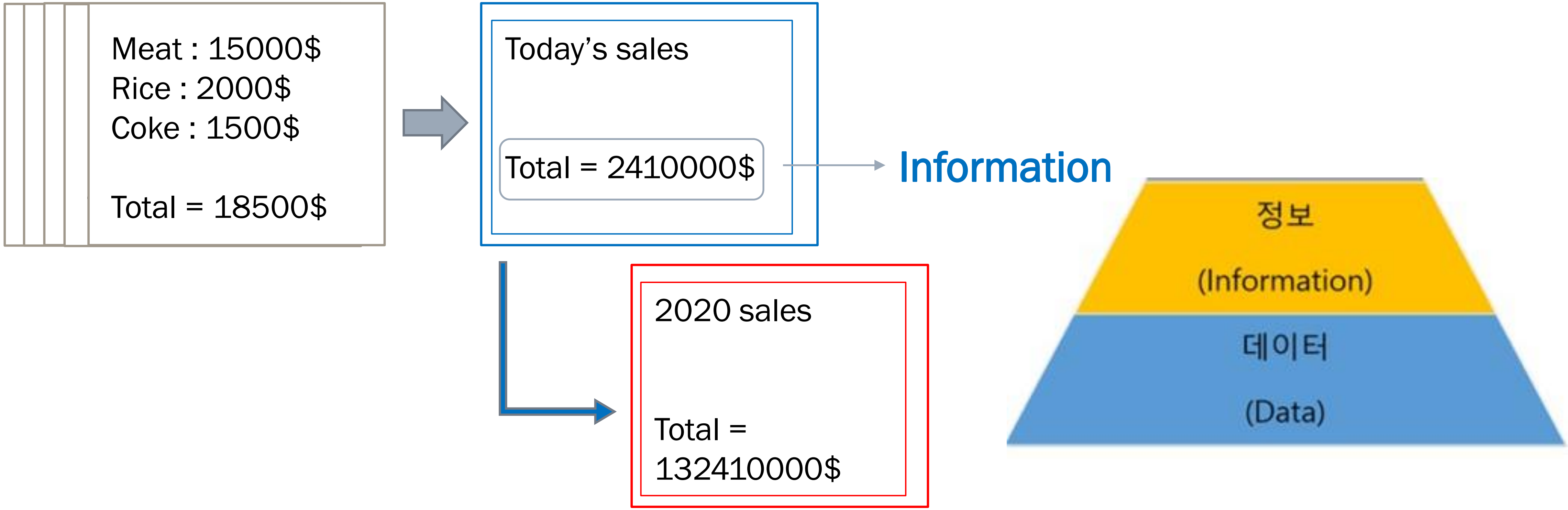
Meat : 15000$
Rice : 2000$
Coke : 1500$
.
.
.
.
.
.
Total = 18500$

데이터

(Data)

1. What is Data?

- Information : valuable data



Meat : 15000$
Rice : 2000$
Coke : 1500$

Total = 18500$

Today's sales

Total = 2410000$

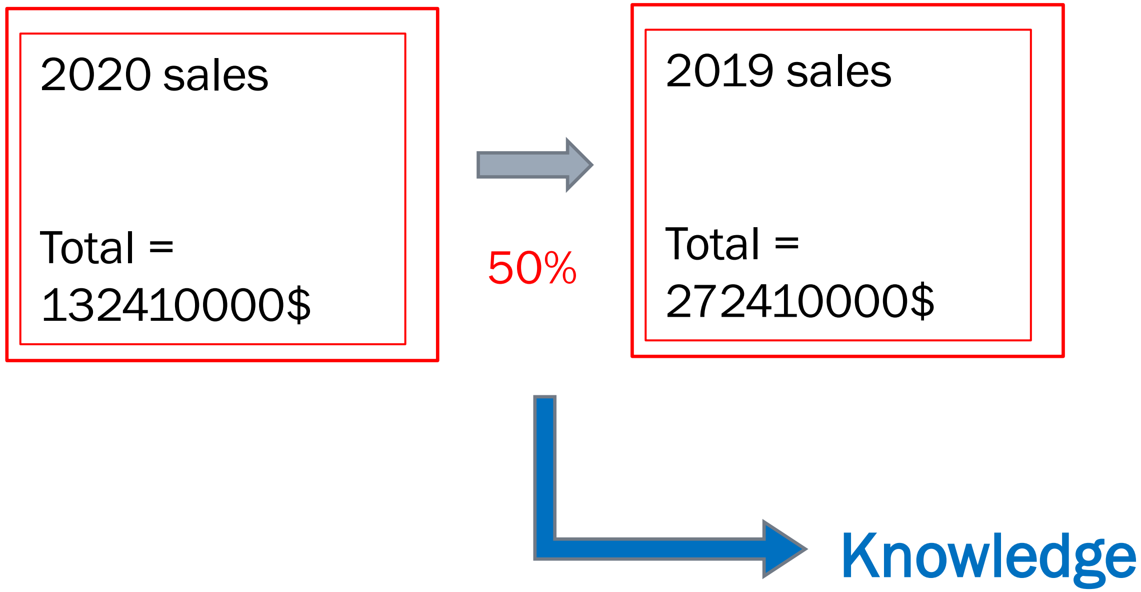
Information

2020 sales

Total = 132410000$

1. What is Data?

- Knowledge : valuable information



2020 sales

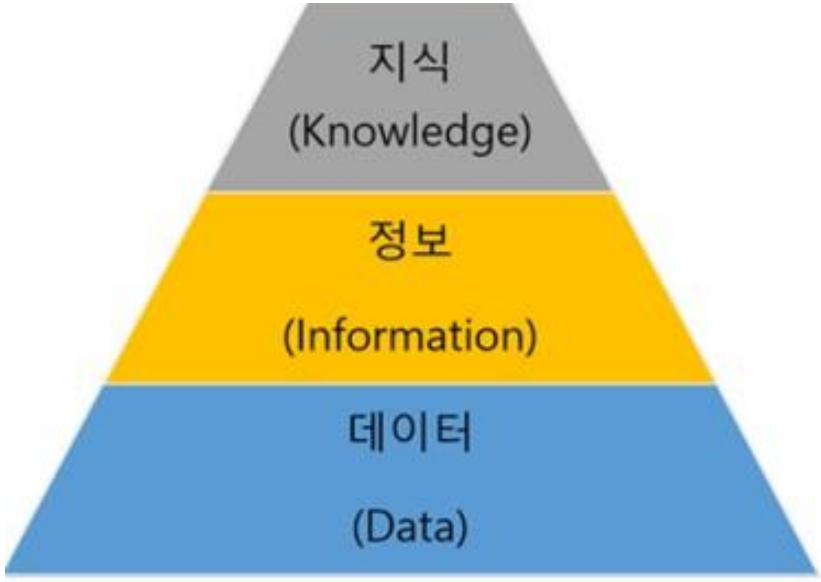
Total = 132410000$

50%

2019 sales

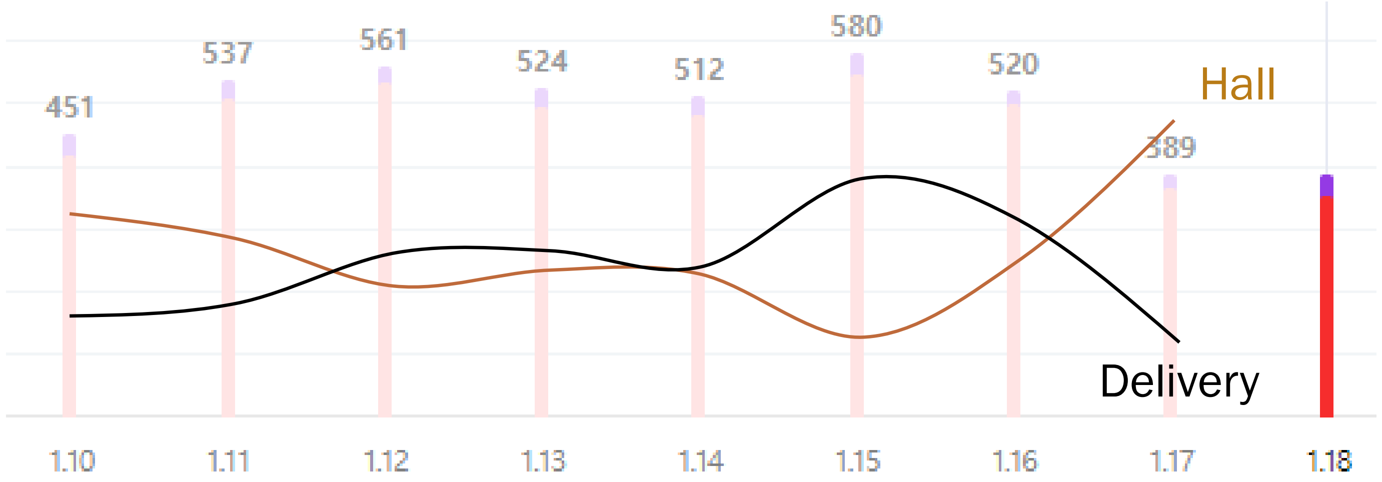
Total = 272410000$

Knowledge

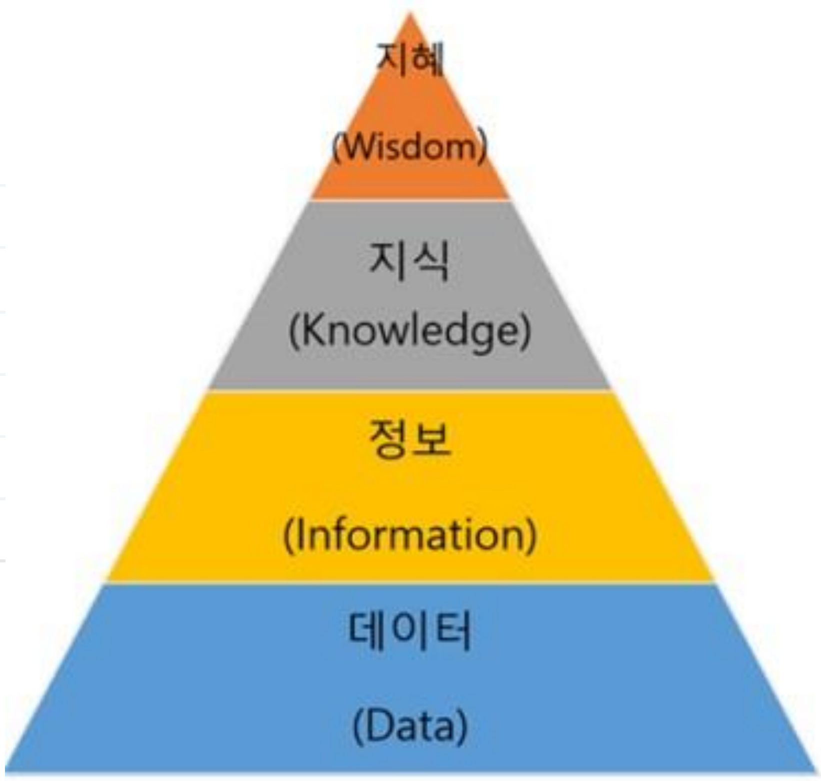


1. What is Data?

- Wisdom : patterned knowledge



Wisdom



지혜

(Wisdom)

지식

(Knowledge)

정보

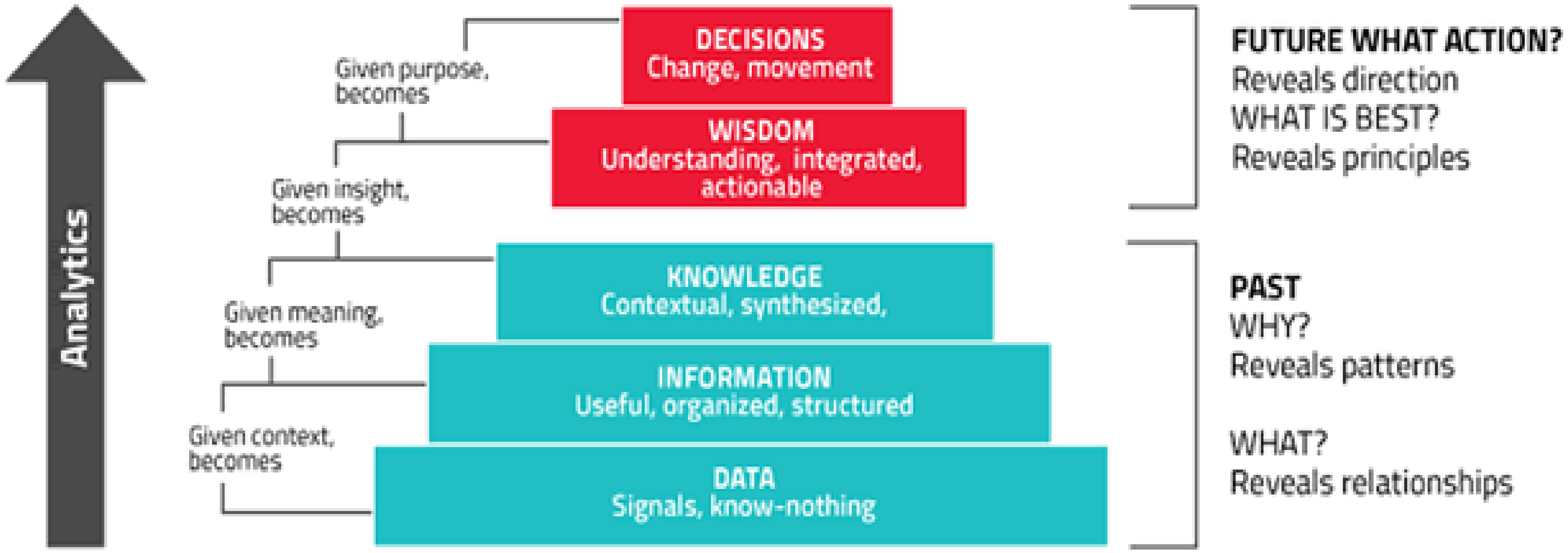
(Information)

데이터

(Data)

1. What is Data?

- DIKW model



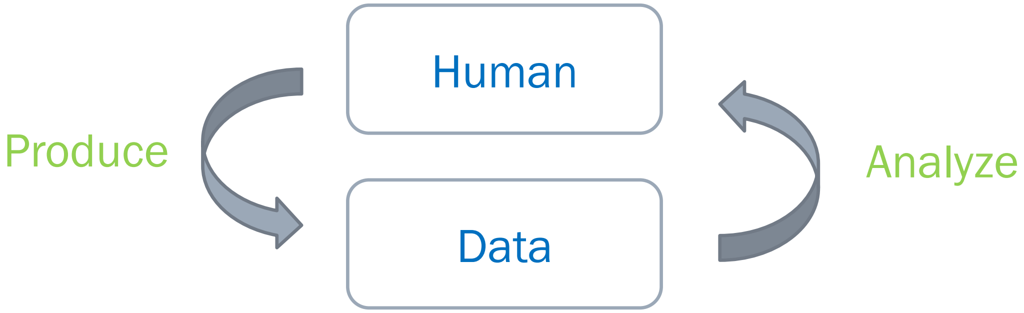
Data mining → Data science

Data Science

- Data ↑ Data Science ↑

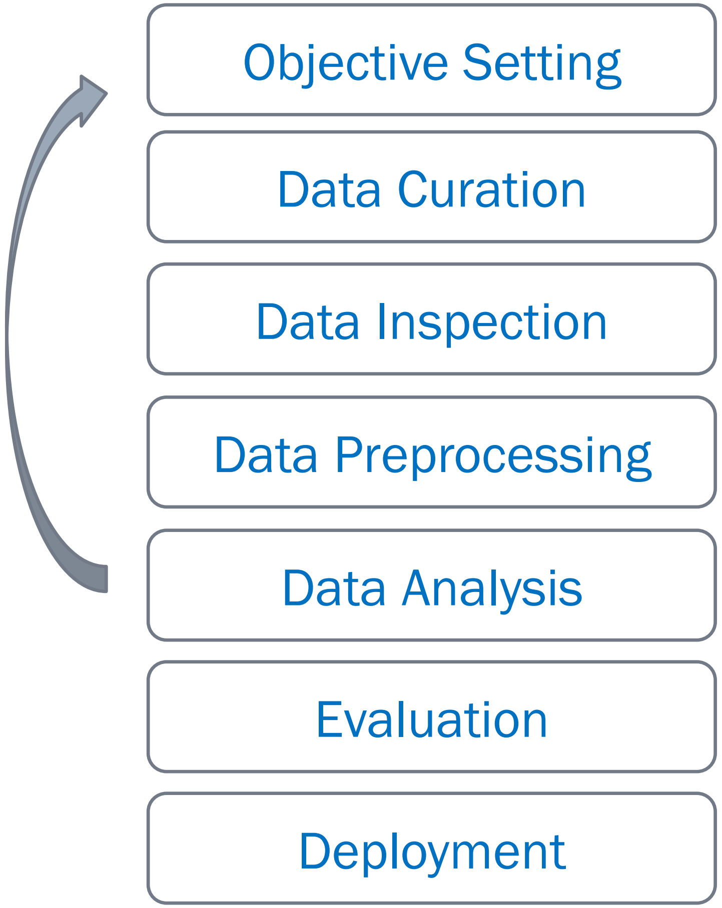
-> Wisdom & Insight

Engineering ✚ Humanistic



End to End Process

- Process are not strictly sequential
- Each process can be repeated



2. Objective Setting

- Business problem -> Engineering problem

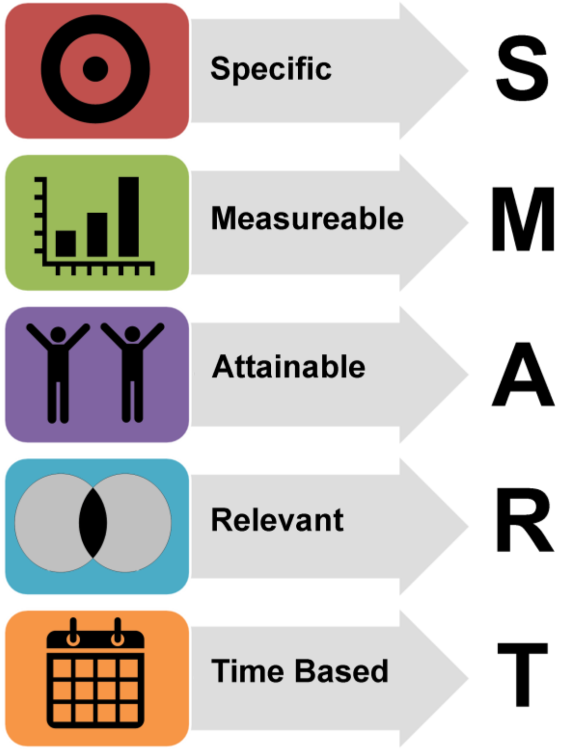
(Important)

- How to define the end?

- How much performance should?

- What is the conditions for success?

-> Documentation



Specific S

Measureable M

Attainable A

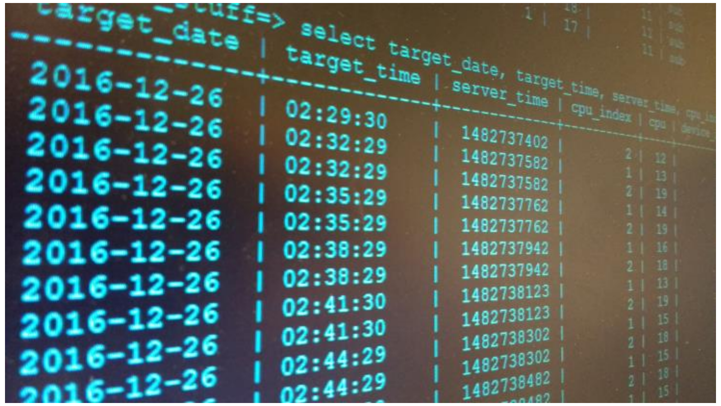
Relevant R

Time Based T

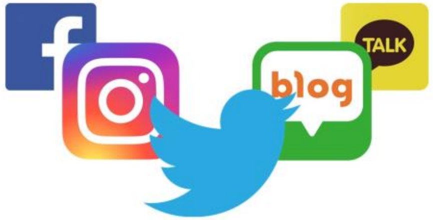
3. Data Curation

- Determine the data needed to best meet goal

- Collect and store the data in the computer



```
stuff=> select target_date, target_time, server_time, cpu_i
target_date | target_time | server_time | cpu_index | cpu | device_
------------+-------------+-------------+-----------+-----+--------
2016-12-26  | 02:29:30    | 1482737402  |         2 |  12 |
2016-12-26  | 02:32:29    | 1482737582  |         1 |  13 |
2016-12-26  | 02:32:29    | 1482737582  |         2 |  19 |
2016-12-26  | 02:35:29    | 1482737762  |         1 |  14 |
2016-12-26  | 02:35:29    | 1482737762  |         2 |  19 |
2016-12-26  | 02:38:29    | 1482737942  |         1 |  16 |
2016-12-26  | 02:38:29    | 1482737942  |         2 |  18 |
2016-12-26  | 02:41:30    | 1482738123  |         1 |  13 |
2016-12-26  | 02:41:30    | 1482738123  |         2 |  19 |
2016-12-26  | 02:44:29    | 1482738302  |         1 |  15 |
2016-12-26  | 02:44:29    | 1482738302  |         2 |  18 |
```

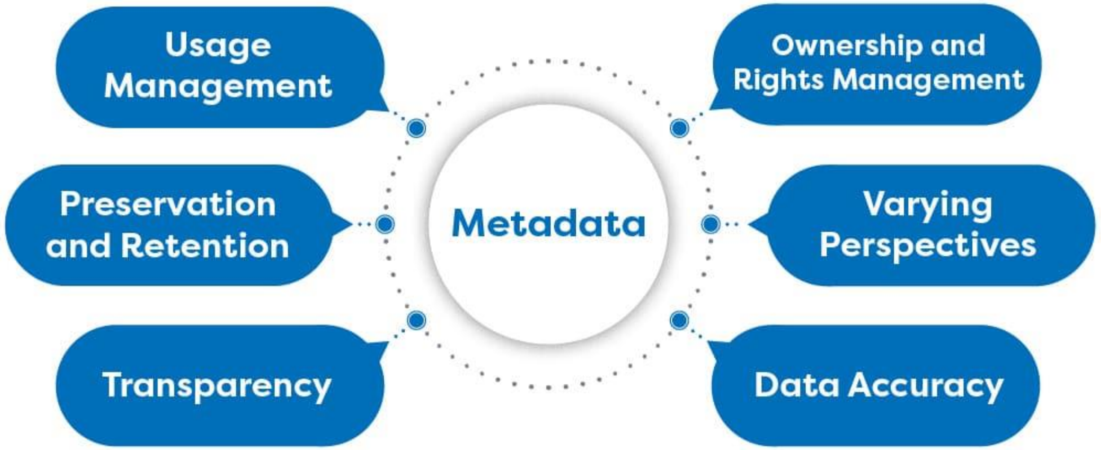


4. Data Inspection

- Check the collected data (suitability & quality)
- 2nd phase of data curation
- Requires expert in certain field
- Requires software tool for browsing the metadata

Meta data

- Data about data
- ex) [DB] entity, attribute, relationship, index
- To represent data, to find quickly
- [Process] result, start/end time

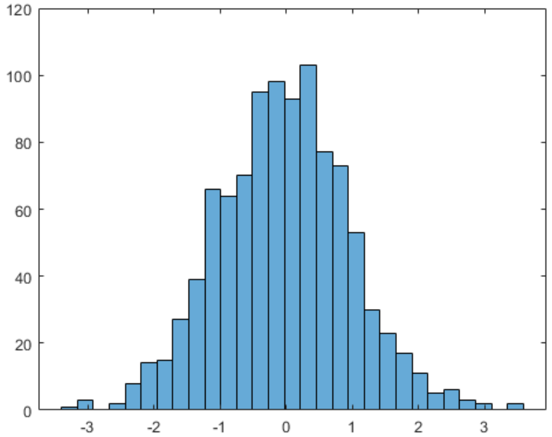


Data exploration

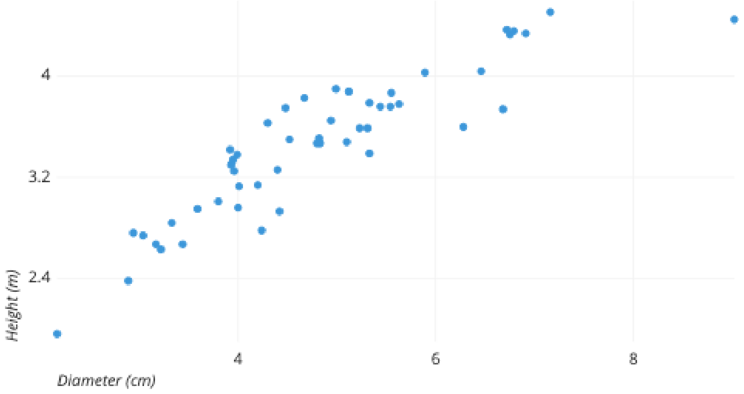
- Find the properties of data

-> Visualization

- Check data tendency, distributions, outliers, correlation

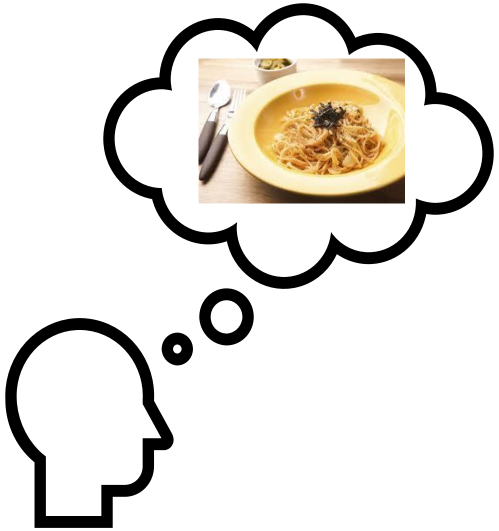


Histogram



Scatter plot

1st~3rd Procedure



Objective Setting

Data Curation &
Data Inspection

Data Preprocessing

5. Data Preprocessing

- Data Preparation

-> important process

- 80% of the time and efforts needed

- Require experts who use software tools

- Include 4 major steps

Data Restructuring

Data Value Change

Feature Engineering

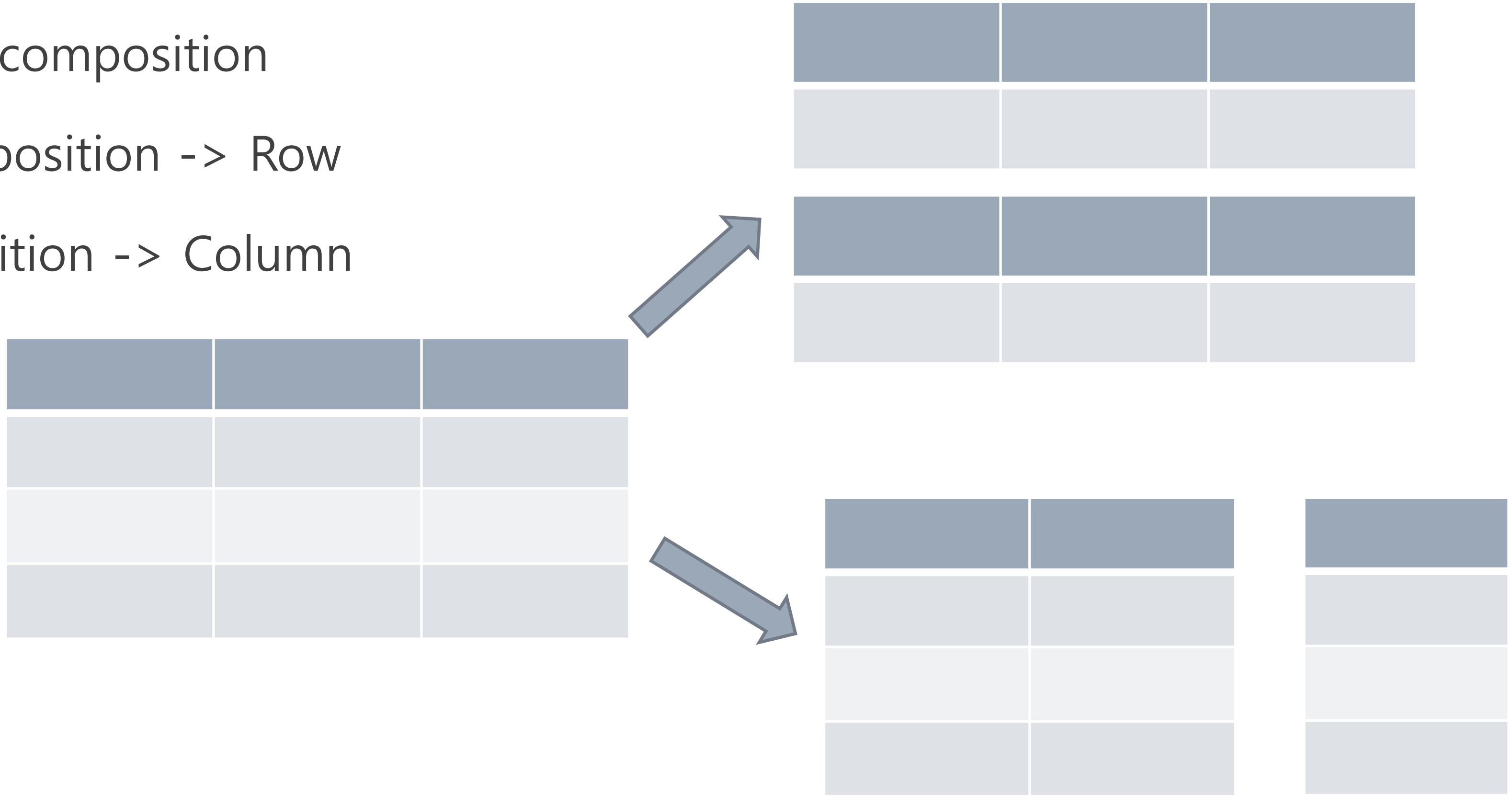
Data Reduction

5-1. Data Restructuring

- Table Merge or Decomposition

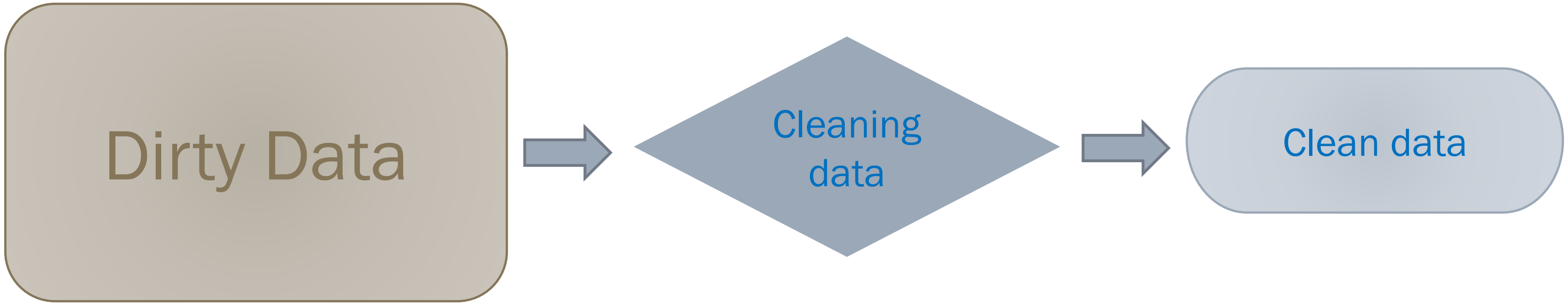
- Horizontal Decomposition -> Row

- Vertical Decomposition -> Column



5-2 Data value change

- There should be no missing or incorrect information



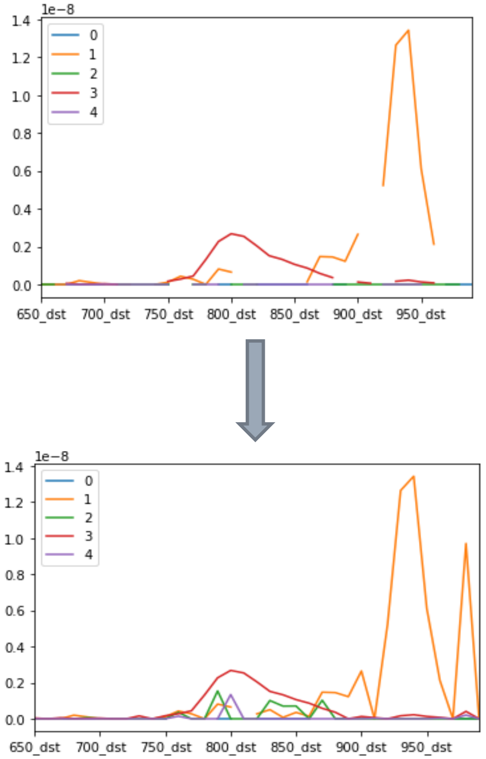
Dirty Data

- Data with no value (NULL) = [Missing data]

- Drop : **pros** : simple / **cons** : data become small, be biased

- Replace : with mean, median, sampling, regression

-> Python : import pandas as pd (dropna, fillna, interpolate)



Dirty Data

[Wrong data]

- Invalid data from data generator
- User do not specify integrity constraint
- Non-Primitive data (e.g. compound data, categorical data)

Resident Registration Number
XXXXXX-XXXXXXX

Dirty Data

[Outlier]

- Data that does not belong in a group of similar data

- Caused by input error

- Must be detected ∵ potential errors, understanding data distribution

- Outlier = value > mean ± 3 * standard deviation

Dirty Data

[Unusable Data]

- Data with ambiguous meanings ex) homonym
- Data that do not conform to standards ex) version, type
- Redundant data

Dealing with Dirty Data

[Prevention]

- Type checking, Integrity constraints

[Cleaning]

- missing data, wrong data, outlier, unusable data