

# **Big Data Analytics Programming**

**Week-07. Data Preprocessing**

**Jungwon Seo, 2021-Spring**

# 데이터 확보 후..

## 무엇을 해야할까?



무작위로 수집된 데이터를 통제 가능한 상태로 변환하는 과정

# 데이터의 종류

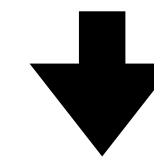
## 데이터의 타입에 따른 전처리 방식

- 비정형 데이터
  - 텍스트, 비디오, 오디오, 이미지
- 준정형 데이터
  - XML, JSON, HTML등의 형태
- 정형 데이터
  - Matrix 형태로 바로 표현/접근/연산 가능한 상태
- 어떠한 형태의 데이터든 결국에는 행렬 형태로 표현해야 분석 가능

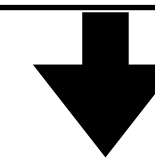
# 데이터 전처리

## 비정형데이터 - 텍스트 데이터

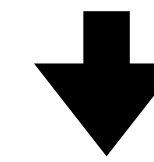
- Document Normalization (문서 정규화)
  - 인코딩 통일
  - 준정형 데이터로 변형 (XML, CSV, JSON, SQL 등)
- Text Parsing
  - 텍스트 추출 (예: <p>Hello!</p>)
  - Stemming or Lemmatization
- Text Filtering
  - 고빈출어 또는 저빈출어 제거
  - 불용어(stopwords) 제거
  - 개인정보 식별 정보 제거 (예. 주민번호, 핸드폰 번호)
- Transformation
  - One-hot encoding
  - Ordinal encoding



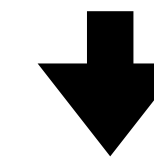
```
[{
  "doc_id": 1,
  "content": "<h1> 안녕하세요 </h1> <p>만나서 반갑습니다.</p> ....."
},
{
  "doc_id": 2,
  "content": "제 이름은 서중원이고, 나이는 31살입니다....."
},
{
  "doc_id": 3,
  "content": "I like playing tennis ....."
}]
```



```
[ {
  "doc_id": 1,
  "content": "안녕 만나다 반갑다"
},
{
  "doc_id": 2,
  "content": "제 이름 000 나이 00살"
},
{
  "doc_id": 3,
  "content": "like play tennis"
} ]
```



	안녕	만나다	반갑다	Tennis
doc1	1	1	1	0
doc2	0	0	0	0
doc3	0	0	0	1



```
[
  { "doc_id": 1, "encoded": [3, 32, 218, 12, 25, 2, 205, 337, 16, 2, 113, 4] },
  { "doc_id": 2, "content": [3, 225, 21, 55, 88, 278, 238, 73, 61, 21, 31, 4] },
  { "doc_id": 3, "content": [3, 22, 311, 72, 28, 34, 5, 7, 33, 65, 13, 4] }
]
```

# 데이터 전처리

## 비정형데이터 - 텍스트 데이터

- 원형복원
  - Stemming
    - 단어의 끝을 일정한 규칙으로 잘라내는 방식으로 원형 복원
  - Lemmatization
    - 사전을 기반으로, context를 고려한 원형 복원
- 불용어 제거
  - 불용어(stopwords)란?
    - 데이터셋에 자주 등장하지만, 큰 의미가 없는 단어
    - Task에 따라 제거 하기도 제거하지 않기도 함

Stemming	Adjustable Formality Formaliti Airliner Meeting	Adjust Formaliti Formal Airlin Meet
Lemmatization	Was Better Meeting	Be Good Meeting

Korean Stopwords		
아	어찌됐든	하기보다는
휴	그위에	차라리
아이구	게다가	하는 편이 낫다
아이쿠	점에서 보아	흐흐
아이고	비추어 보아	놀라다
어	고려하면	상대적으로 말하자
나	하게될것이다	면
우리	일것이다	마치
저희	비교적	아니라면

Default English stopwords list	
This list is used in our <a href="#">Page Analyzer</a> and <a href="#">Article Analyzer</a> for English text, when you let it use the default stopwords list.	
a	ourselves
about	out
above	over
after	own
again	same
against	shan't
all	she
am	she'd
an	she'll
and	she's
any	should
are	shouldn't
aren't	so

# 데이터 전처리

## 비정형데이터 - 이미지 데이터

- 이미지 리사이즈

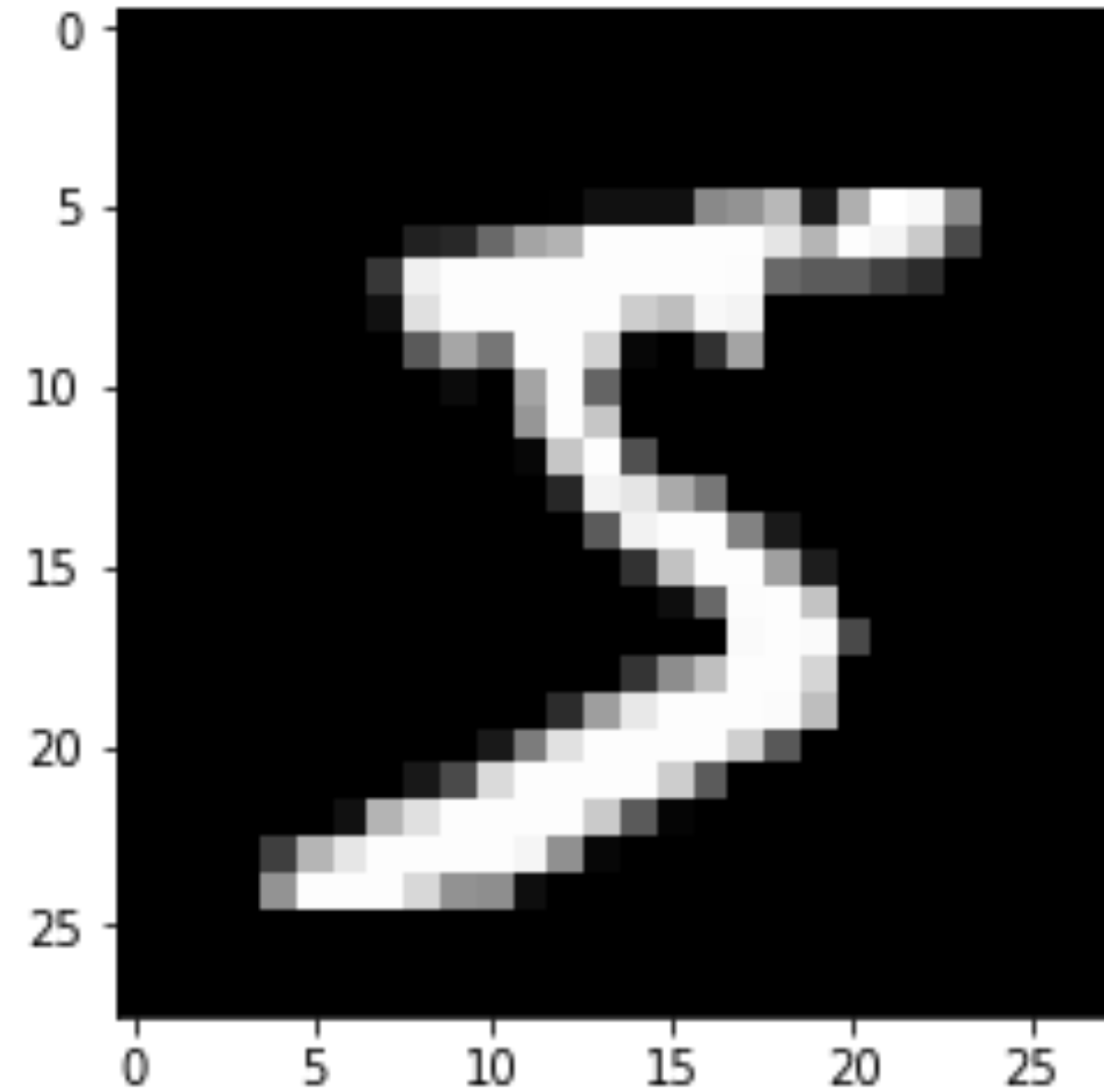
- 가로 세로
- 채널 수 (RGB)
- Cropping

- Noise 제거

- Blurring

- Segmentation

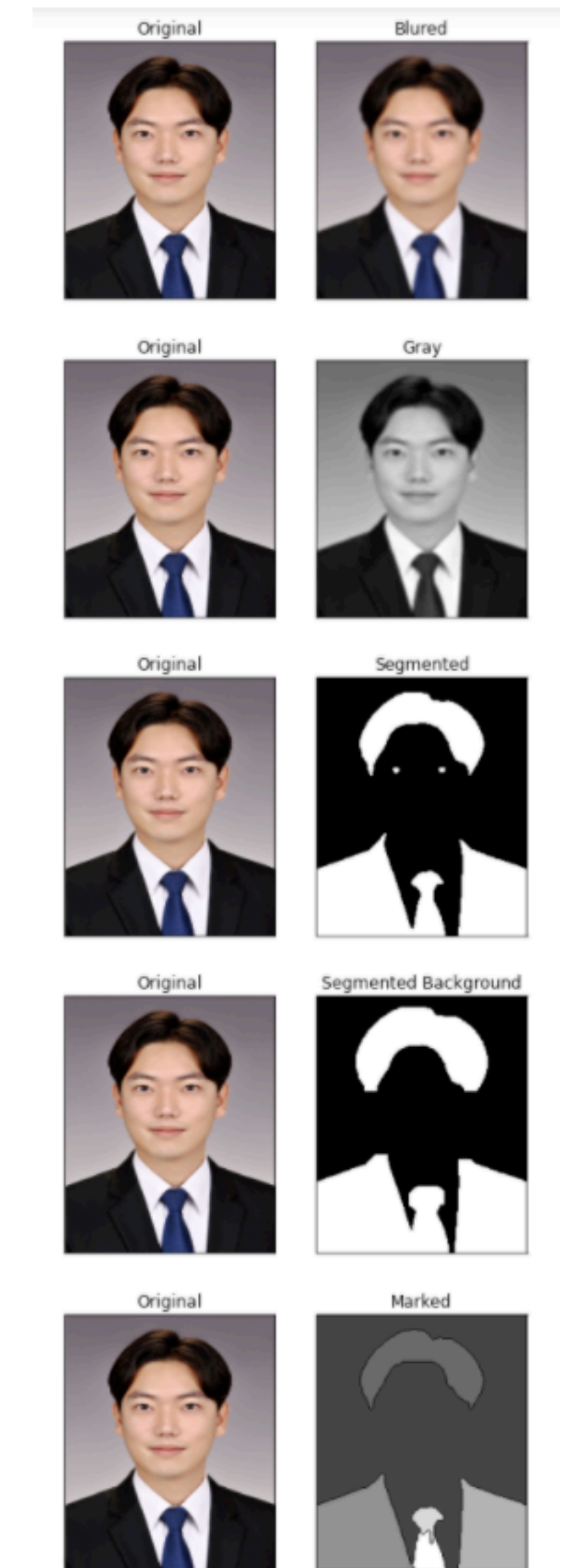
- Background vs Objects



```

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 3 18 18 18 126 136 175 26 166 255 247 127 0 0 0 0
0 0 0 0 0 0 0 0 0 30 36 94 154 170 253 253 253 253 225 172 253 242 195 64 0 0 0 0
0 0 0 0 0 0 0 49 238 253 253 253 253 253 253 253 251 93 82 82 56 39 0 0 0 0 0 0
0 0 0 0 0 0 0 18 219 253 253 253 253 253 198 182 247 241 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 80 156 107 253 253 205 11 0 43 154 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 14 1 154 253 90 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 139 253 190 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 11 190 253 70 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 35 241 225 160 108 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 81 240 253 253 119 25 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 45 186 253 253 150 27 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 16 93 252 253 187 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 249 253 249 64 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 46 130 183 253 253 207 2 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 39 148 229 253 253 253 250 182 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 24 114 221 253 253 253 253 201 78 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 23 66 213 253 253 253 253 198 81 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 18 171 219 253 253 253 253 195 80 9 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 55 172 226 253 253 253 253 244 133 11 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 136 253 253 253 212 135 132 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```



# 데이터 전처리

## 비정형데이터 - 이미지 데이터

- 이미지의 정의
  - 3차원 세상에 대한 2차원 화면(View)
  - 디지털 이미지는 **한정된 값의 묶음**으로 2차원 이미지를 숫자로 표현
  - 이러한 값을 **픽셀**이라고 부르며, 집합으로서 이미지를 나타냄
  - 다른말로 픽셀은 컴퓨터화면에 표시 될 수 있는 가장 작은 단위
  - 디지털 이미지는 픽셀 매트릭스로 컴퓨터에 표시
  - 이미지의 각 픽셀은 정수로 저장

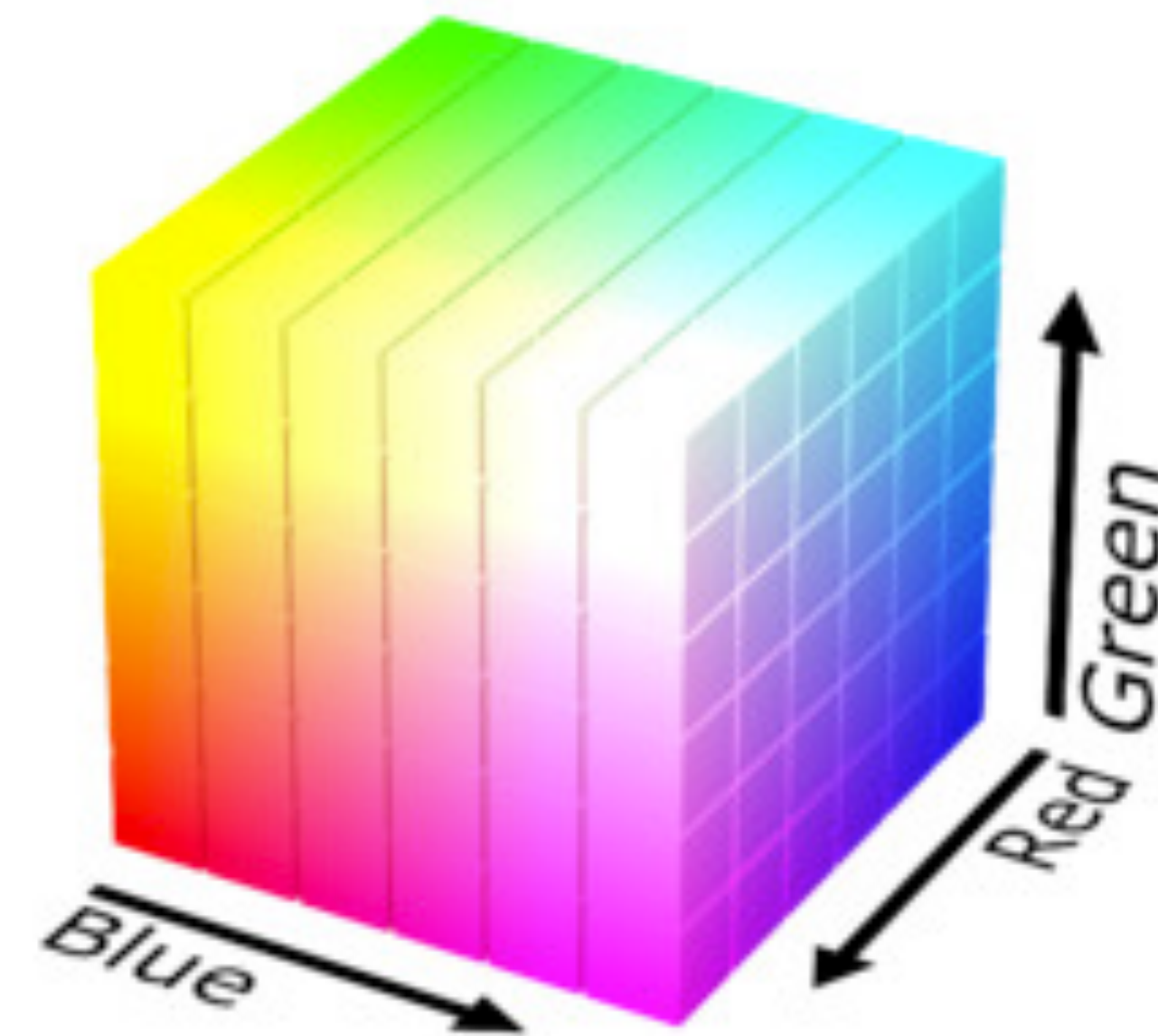


# 데이터 전처리

## 비정형데이터 - 이미지 데이터

- RGB

- 회색조 이미지를 처리하는 경우 0 (검은 색 픽셀)에서 최대 255 (흰색 픽셀)
  - 이 둘 사이의 숫자는 회색 음영
- 반면 컬러 이미지는 3 개의 행렬로 표현
  - 각 행렬은 채널이라고도하는 하나의 기본 색상을 나타냄
  - 가장 일반적인 색상 모델은 빨강, 녹색, 파랑 (RGB)
  - 이 세 가지 색상은 함께 혼합되어 광범위한 색상을 생성



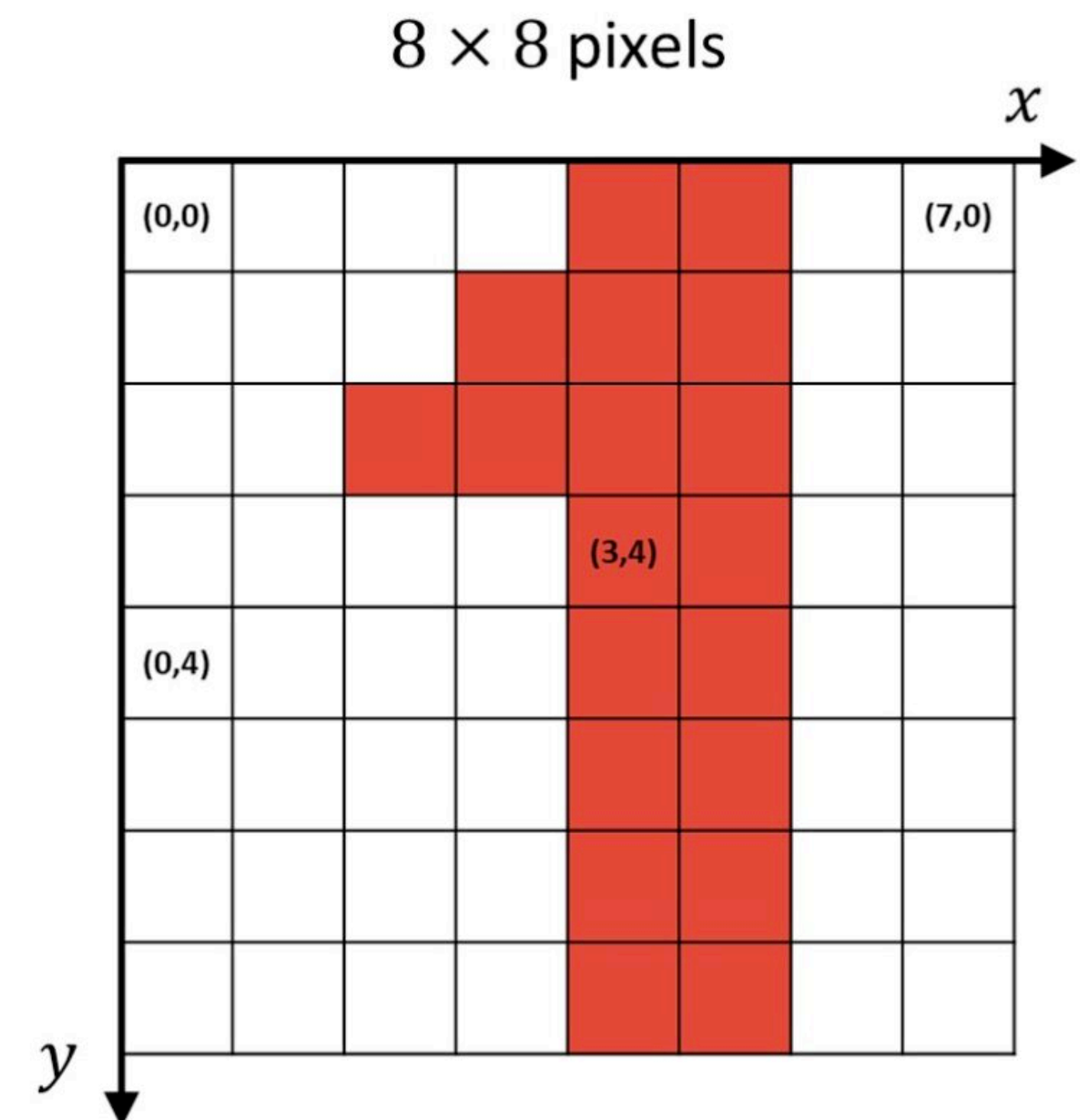


# 데이터 전처리

## 비정형데이터 - 이미지 데이터

- 픽셀좌표

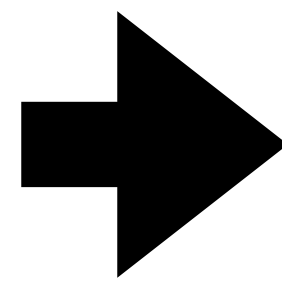
- 픽셀은 두 개의  $(x, y)$  좌표로 접근
  - $x$  값은 열을 나타내고  $y$  값은 행을 나타냄.
- 이미지의 왼쪽 상단 모서리에는 원점 좌표  $(0,0)$ 
  - $x$  좌표 값은 오른쪽으로 갈수록 증가
  - $y$  좌표 값은 아래로 갈수록 증가
- 각각의 픽셀에 접근해서 조작 가능



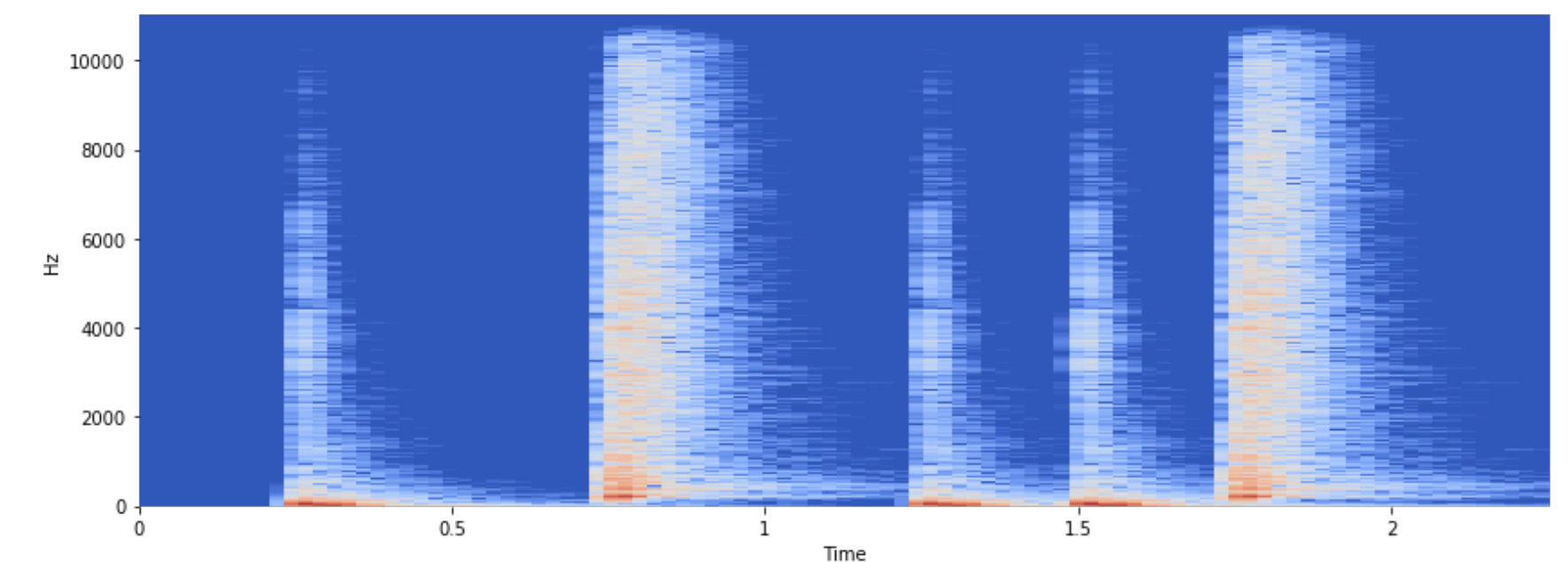
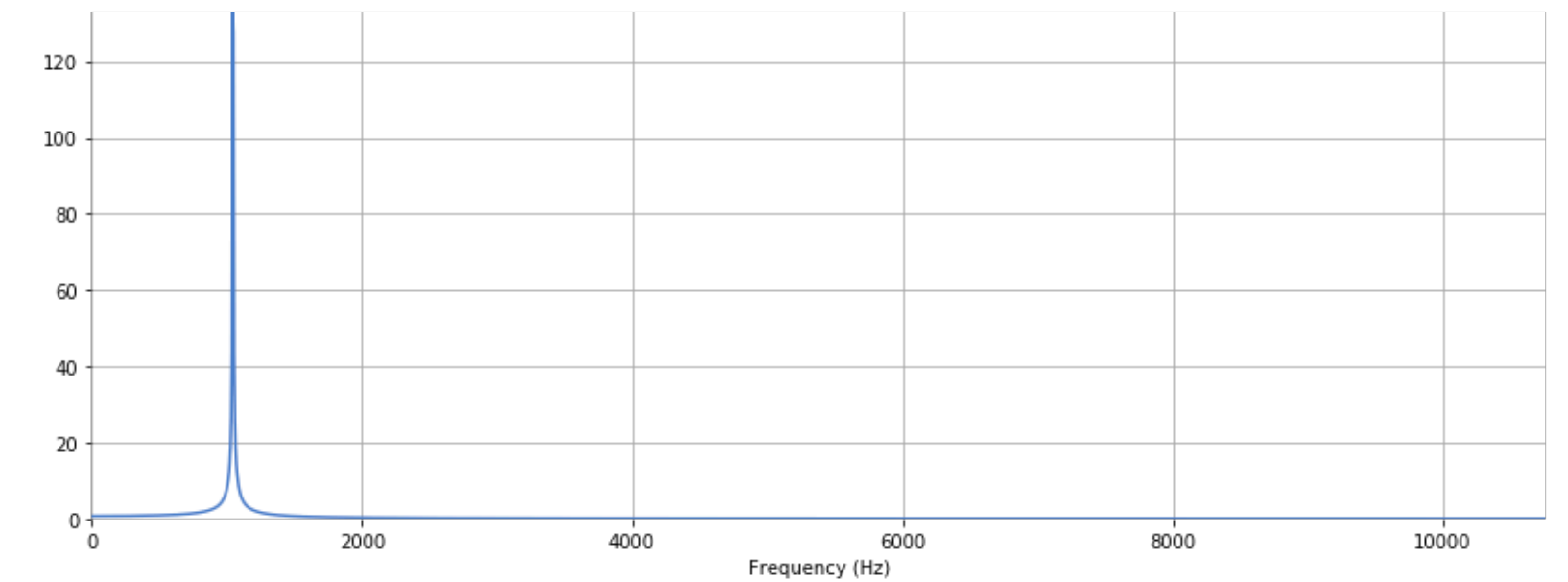
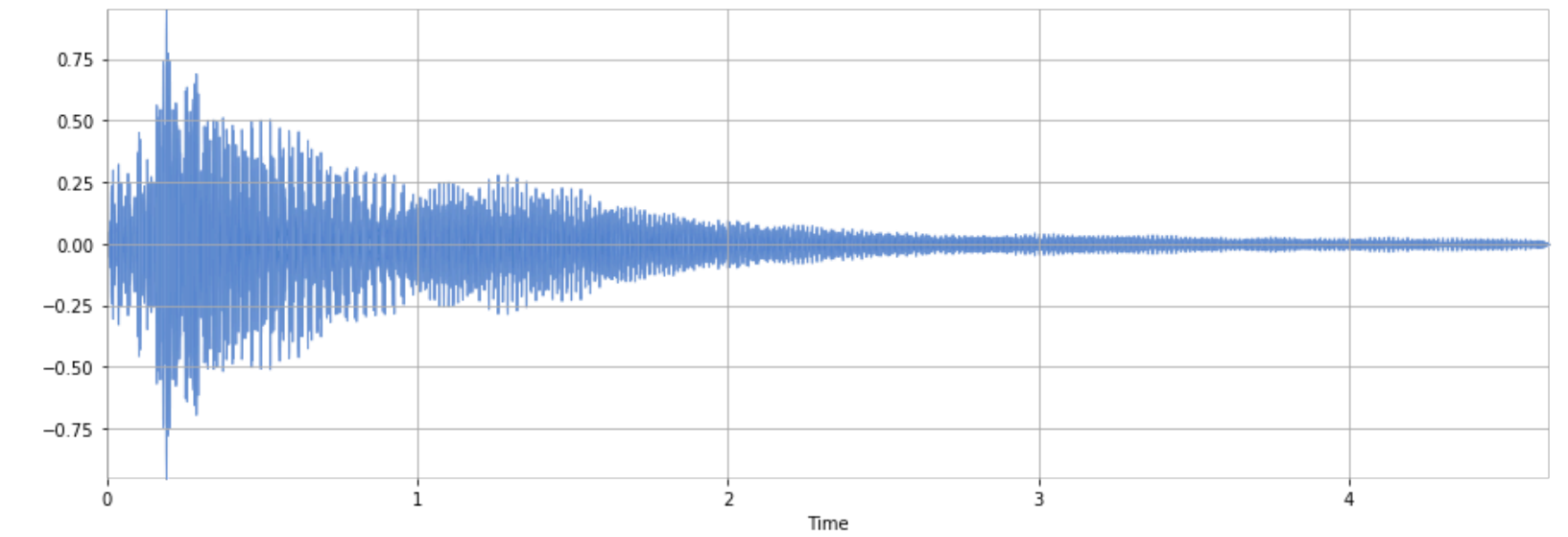
# 데이터 전처리

## 비정형데이터 - 오디오 데이터

- Audio 파일 불러오기
- Representation 찾기
  - 시간에 따른 세기 기반으로 표현
  - 주파수 기반으로 표현
  - 주파수와 시간으로 표현 (예: STFT)



[12,3004, 300, 0,0,300 ...]



# 데이터 전처리

## 비정형데이터 - 오디오 데이터

- 소리의 3요소
  - 주파수 : 음의 높낮이, Hz
    - 진동수가 높은 음을 고음
    - 진동수가 낮은 음을 저음
    - 돌고래: 초음파
    - 호랑이: 저주파
  - 진폭 : 음의 세기, dB
  - 파형 : 음색

물리량 (physical quantity)	심리량(subjective quantity)		
	소리 크기 (loudness)	소리 높낮이 (pitch)	음색 (timbre)
음압(pressure)	***	*	*
주파수(frequency)	**	***	**
스펙트럼(spectrum)	*	*	***
포락선(envelope)	*	*	**
지속시간(duration)	*	*	*

(상관정도: \*)

소리의 3요소

# 데이터 전처리

## 준정형데이터 - XML, JSON

- Row(행) 기반 데이터로 변형
  - 일반적으로 XML과 JSON으로 이루어진 데이터의 경우 불규칙한 차원수를 포함하기 때문에 데이터를 분해하여, 2차원의 매트릭스로 표현하는 작업이 필요
  - Row 기반의 데이터가 되어 빈도수나 평균/합과 같은 Aggregation을 통해 추가적인 유의미한 정보 추출 가능 (Feature Engineering)
  - CSV의 경우에는 이미 행 기반

	WHO	WEEK
0	Joe	{{'NUMBER': 3, 'EXPENSE': [{'WHAT': 'Beer', 'A...
1	Beth	{{'NUMBER': 3, 'EXPENSE': [{'WHAT': 'Beer', 'A...
2	Janet	{{'NUMBER': 3, 'EXPENSE': [{'WHAT': 'Car', 'AM...

[원본 데이터]

WHAT	AMOUNT	WHO	AMOUNT
Beer	183.0	Beth	80.0
Car	68.0	Janet	129.0
Food	136.0	Joe	178.0

[Aggregate 데이터]

	WHO	WEEK	WHAT	AMOUNT
0	Joe	3	Beer	18.0
1	Joe	3	Food	12.0
2	Joe	3	Food	19.0
3	Joe	3	Car	20.0
4	Joe	4	Beer	19.0
5	Joe	4	Beer	16.0
6	Joe	4	Food	17.0
7	Joe	4	Food	17.0
8	Joe	4	Beer	14.0
9	Joe	5	Beer	14.0
10	Joe	5	Food	12.0
11	Beth	3	Beer	16.0
12	Beth	4	Food	17.0
13	Beth	4	Beer	15.0
14	Beth	5	Food	12.0
15	Beth	5	Beer	20.0
16	Janet	3	Car	19.0
17	Janet	3	Food	18.0
18	Janet	3	Beer	18.0
19	Janet	4	Car	17.0
20	Janet	5	Beer	14.0
21	Janet	5	Car	12.0
22	Janet	5	Beer	19.0
23	Janet	5	Food	12.0

[행기반 데이터]

# 데이터 전처리

## 정형데이터

- Data Cleaning
  - 결측값(Missing Value) 처리
  - Noise와 Outlier 처리
  - 중복 데이터 처리
  - 기타: 다른 화폐단위, 다른 온도 단위 등등
- Data Normalization (Numeric Feature)
  - Min-Max Scaling
  - Z-score Normalization
- Data Encoding (Categorical Feature)
  - Ordinal Encoding
  - One-hot Encoding
  - Binary Encoding

# 데이터 전처리

## 정형데이터 - 결측값

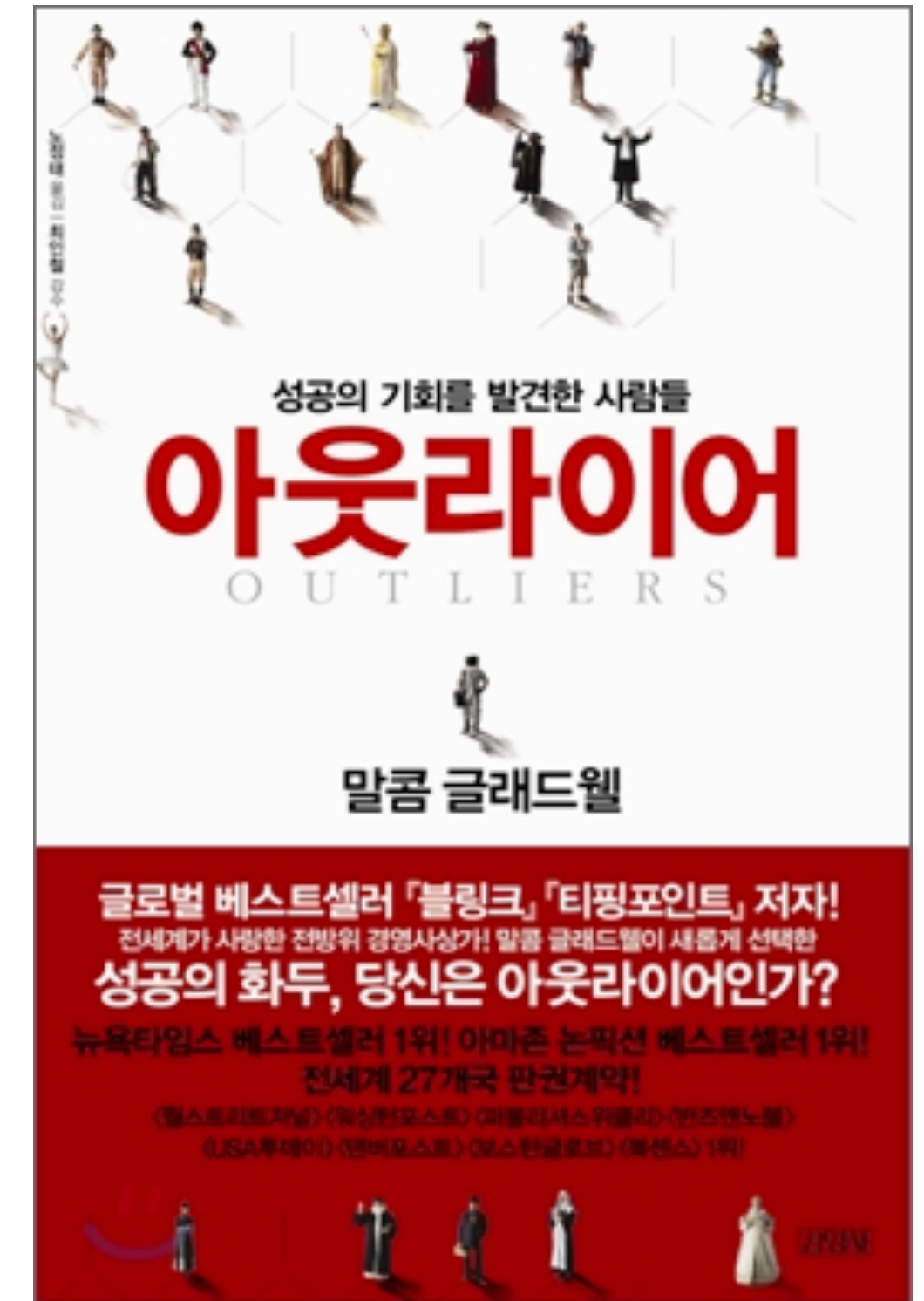
- 결측값이 생기는 이유
  - 모든 정보가 반드시 수집되는 것은 아님
    - 코로나로 인한 방명록 작성
    - 사람들은 나이나 체중을 숨기려는 경향이 있음
  - 일부 속성들은 특정 대상에게는 적용되지 않음
    - 연간소득, 어린이는?
- 결측값 처리
  - 데이터 객체 제거
  - 데이터 속성 제거
  - 결측값 추정
  - 분석 중 결측값 무시

1	183	70	25
2	176	60?	22
3	154	40	35
4	168	70	40

# 데이터 전처리

## 정형데이터 - Noise, Outlier

- 잡음 (Noise)
  - 잘못된 관측 또는 무작위적 오류
  - 예) 오타, 결측값, 의미없는 값
- 이상치 (Outlier)
  - 대부분의 데이터와는 상당히 다른 형태를 보이는 데이터
  - Noise와는 다르게 관심 갖고 지켜봐야 하는 경우가 있음
    - Anomaly Detection
    - 예) 넷플릭스 계정 해외접속



말콤 글래드웰, 아웃라이어, 김영사(2009)



# 데이터 전처리

## 정형데이터 - Normalization

- 수치 범위가 넓은 속성(feature)에 대한 편향(bias)을 방지하기 위해
  - 예 : 1.7m vs 60kg
- Min-Max Scaling
  - $x_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$
- z-score Normalization
  - $z = \frac{x - \mu}{\sigma}$
- 효과
  - 경사하강법 사용시 더 빠르게 최적의 해를 찾을 수 있음
  - 정규화가 필요하지 않은 모델들도 있다: 의사결정트리

# 데이터 전처리

## 정형데이터 - Encoding

- 목적
  - Categorical type의 데이터는 수학적 연산이 불가능 하기 때문에, 이를 숫자로 표현할 필요성이 있음
- Ordinal Encoding
  - 일반적으로 값들 사이의 대소관계가 존재하는 경우에 사용
  - 하지만, 딥러닝을 이용한 NLP 분야에서는 단어들을 Ordinal Encoding을 사용함
- One-hot Encoding
  - 값들 사이의 대소관계가 존재하지 않는 경우에 사용
  - Sparse matrix
- Binary Encoding
  - ID 값을 부여후 2진법으로 표현

혈액형	ID	이진 인코딩	원-핫 인코딩
A	1	0 0 1	1 0 0 0
B	2	0 1 0	0 1 0 0
AB	3	0 1 1	0 0 1 0
O	4	1 0 0	0 0 0 1

**E.O.D**