데이터, 변수

- 분류 Classification)

강의 계획 - 정형

날짜	강의 계획
3월 4일	# 데이터.변수_구분.pptx pandas
3월 5일	pandas 1.03..파이썬 자료형-기본.numpy.pandas
3월 6일	데이터 시각화 - Matplotlib, seaborn openCV
3월 10일	1.03.Data_Source 전처리
3월 11일	전처리
3월 12일	전처리
3월 13일	샘플링, 스케일링
3월 14일	2.05.MiningAlgorism
3월 17일	2.05.MiningAlgorism DT, Logistic, SVM 앙상블 Random Fores 클러스터링, 연관규칙
3월 18일	
3월 19일	
3월 20일	
3월 21일	
3월 24일	DL 모델 이해 (단층 신경망 구현/다층 신경망 구현)
3월 25일	DL 모델 적용 (신경망의 출력과 활성함수/역전파)
3월 26일	DL 모델 적용 및 이해 (활성화 함수, 경사하강법/오차역전파, optimizer, 과적
3월 27일	모델 결과 분석 (Confusion Matrix/Classification Report)
3월 28일	모델 결과 분석 (평가지표, 결과 시각화/모델 해석 도구, 적용)

수치 예측 지도학습 모델링 간이 프로젝트 진행
소스 - 오픈 데이터 소스

데이터, 변수

데이터 분류 (Data Classification)

1) 데이터 타입 (Type)

- int (정수형)
- float (실수형)
- str (문자열)
- bool (불리언)

2) 데이터 특성 (Data Format)

- 범주형 데이터 (Categorical Data)
  - 명목형(Nominal)
  - 서열형(Ordinal)
- 수치형 데이터 (Numerical Data)
  - 이산형(Discrete)
  - 연속형(Continuous)

3) 데이터 유형 - (Category)

- 정형
- 비정형
- 반정형

데이터, 변수

변수 분류 (Variable Classification)

1) 변수의 수에 따른 분류

- 단변량(Univariate)
- 다변량(Multivariate)

2) 변수 유형에 따른 분류

- 종속 변수(Dependent Variable)
- 독립 변수(Independent Variable)

Data Classification

데이터

데이터 분류(Data Classification)

Data Classification		
데이터 타입	int, float 등의 구분	Data Type
데이터 특성	categorical, numerical 구분	Data Attribute
데이터 유형	정형/비정형/반정형 데이터 구분	Data Category

데이터

데이터 분류 (Data Classification)

1) 데이터 타입 (Type)

- int (정수형)
- float (실수형)
- str (문자열)
- bool (불리언)

2) 데이터 특성 (Data Format)

- 범주형 데이터 (Categorical Data)
  - 명목형(Nominal)
  - 서열형(Ordinal)
- 수치형 데이터 (Numerical Data)
  - 이산형(Discrete)
  - 연속형(Continuous)

3) 데이터 유형 - (Category)

- 정형
- 비정형
- 반정형

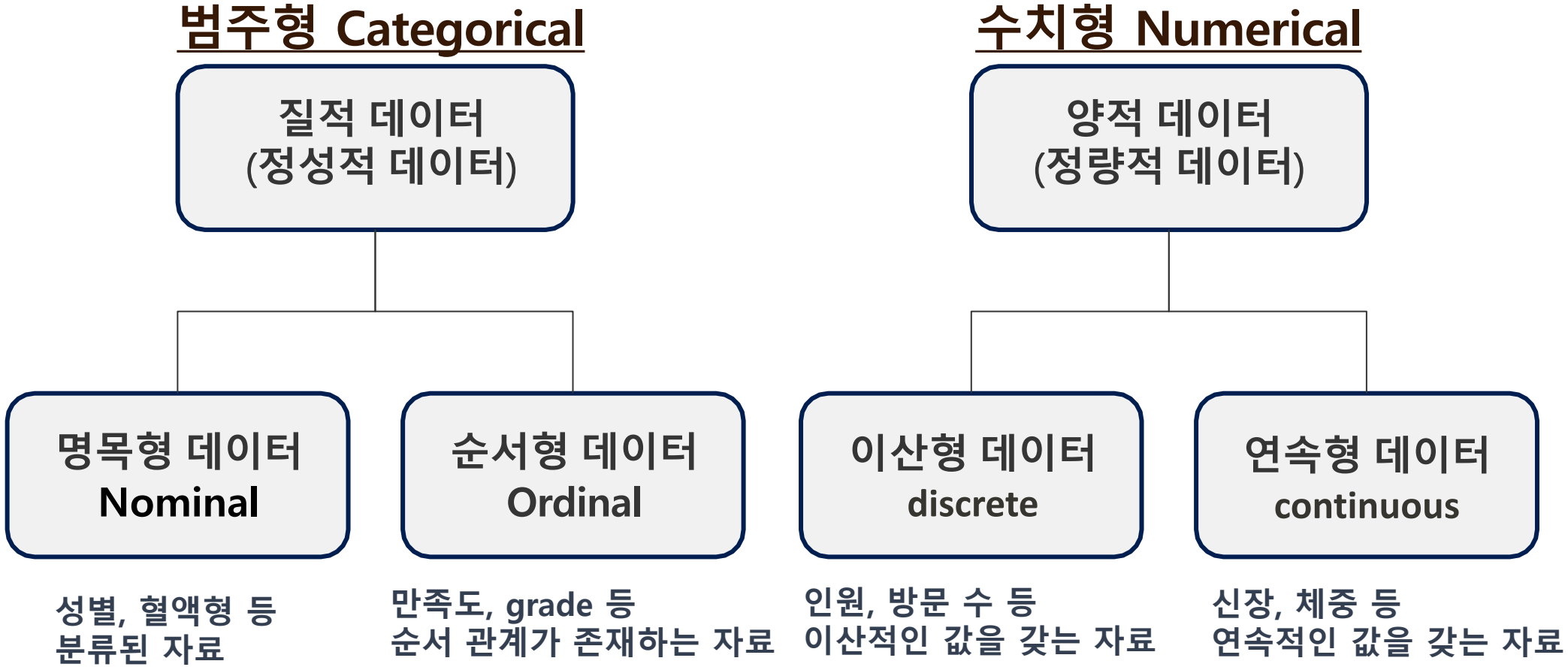
데이터

데이터 분류(Data Classification) - 데이터 타입 (Type)

- **파이썬 데이터 유형**
  - **int (정수형)** 이산형 데이터를 표현하는 데 사용
  - **float (실수형)** 연속형 데이터를 표현하는 데 사용
  - **str (문자열)** 범주형 데이터를 표현하는 데 사용
  - **bool (불리언)** 참/거짓을 나타내는 데이터 타입,
    이진형 데이터(binary data)를 표현할 때 사용

데이터

데이터 분류(Data Classification) - 데이터 특성 (Attribute)



데이터, 변수

데이터 분류(Data Classification) - 데이터 특성 (Attribute)

1.범주형 데이터 (Categorical Data)

- **정의** 데이터가 특정 카테고리나 그룹에 속하는지 나타내는 데이터
- **세부 분류**
  - **명목형(Nominal)** 순서가 없는 카테고리 (예 성별, 혈액형)
  - **서열형(Ordinal)** 순서가 있는 카테고리 (예 교육 수준, 고/중/저 만족도)

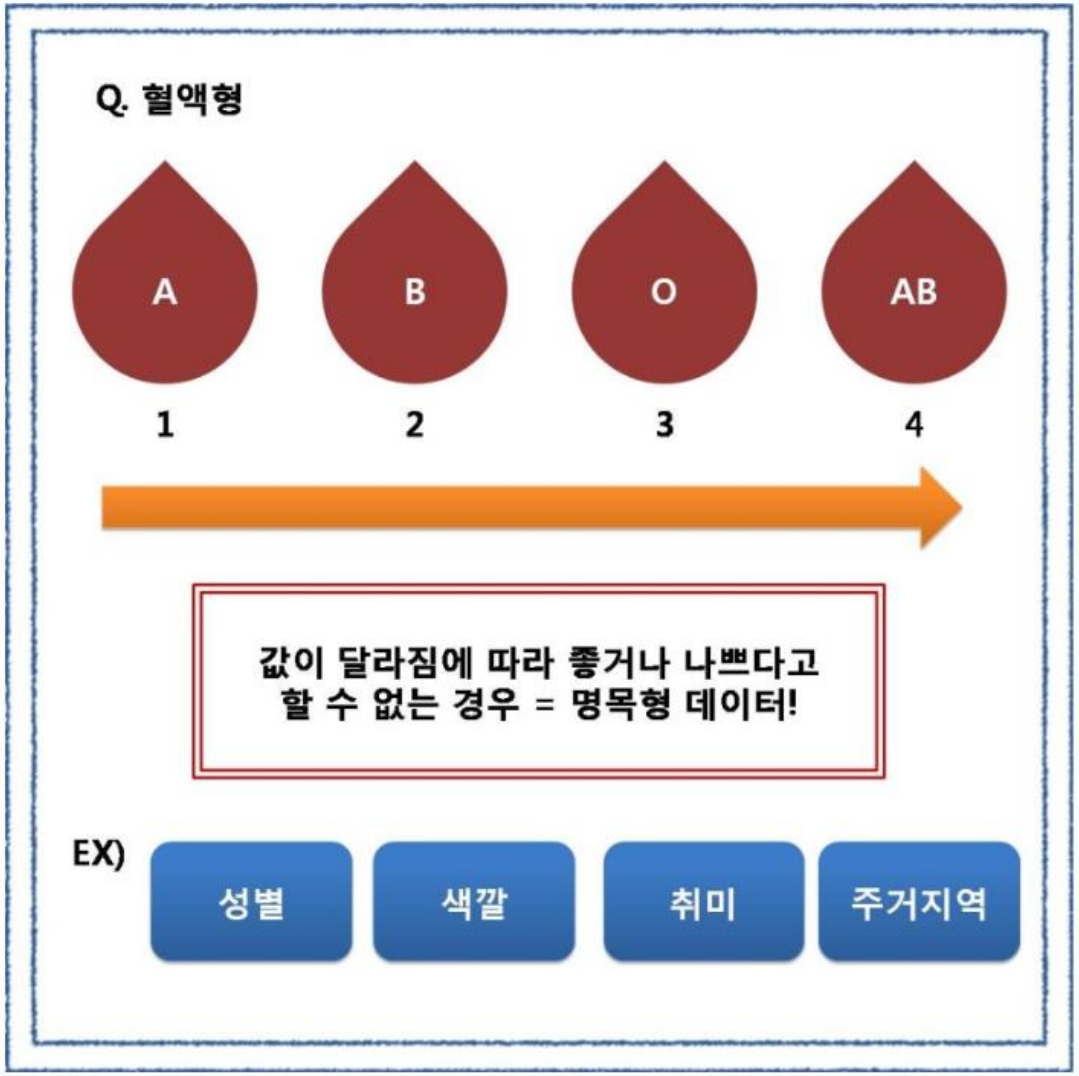
2.수치형 데이터 (Numerical Data)

- **정의** 수치로 표현되는 데이터, 산술 연산 가능
- **세부 분류**
  - **이산형(Discrete)** 셀 수 있는 정수 값 (예 학생 수, 사건 발생 횟수)
  - **연속형(Continuous)** 셀 수 없는 무한한 값 (예 키, 몸무게, 온도)

데이터, 변수

데이터 분류(Data Classification) - 데이터 특성 (Attribute)

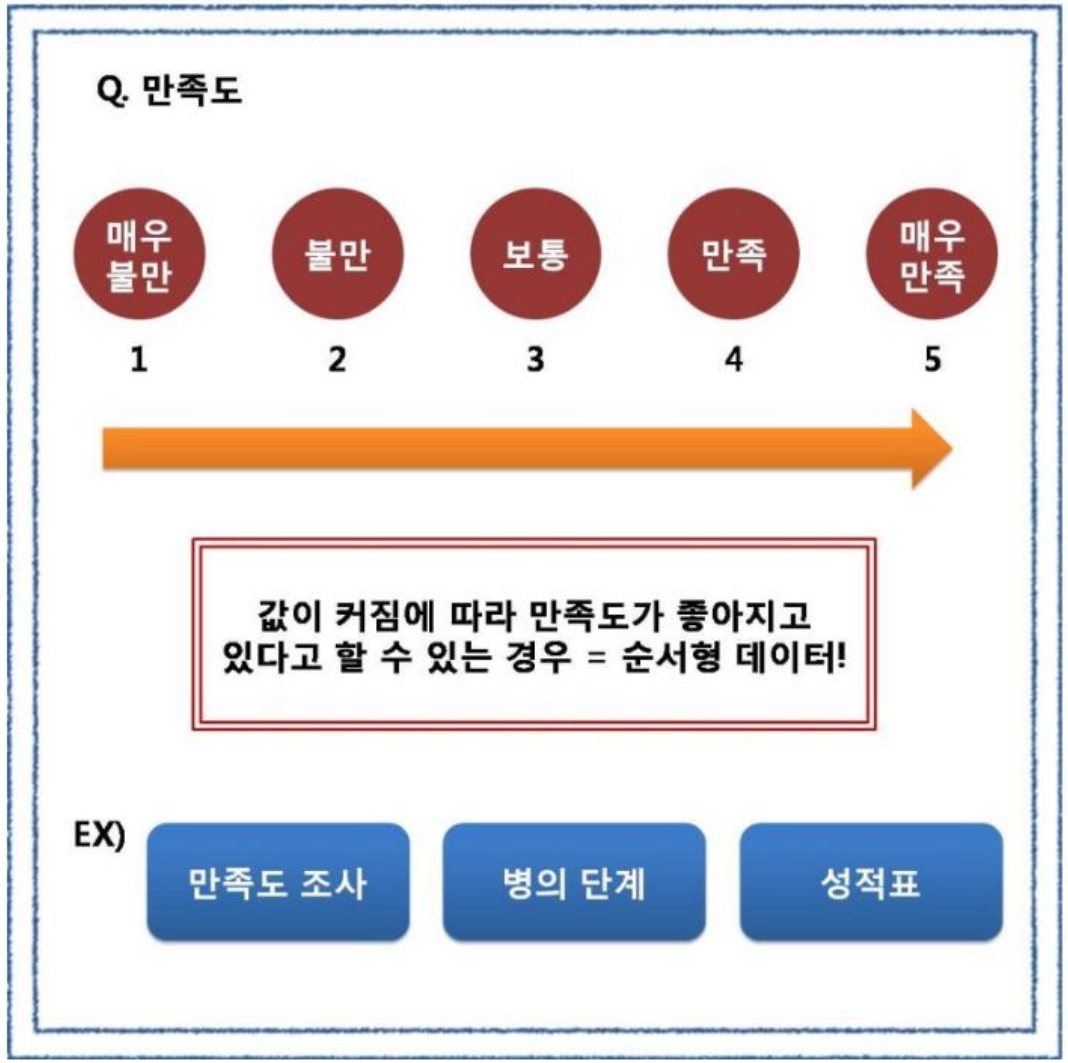
- 범주형
  - L 명목형



데이터

데이터 분류(Data Classification) - 데이터 특성 (Attribute)

- 범주형
  - L 순서형



데이터

데이터 분류(Data Classification) - 데이터 특성 (Attribute)

- 수치형
  - L 이산형

Q. 하루동안 은행에 방문하는 고객의 수

1명, 2명, 3명

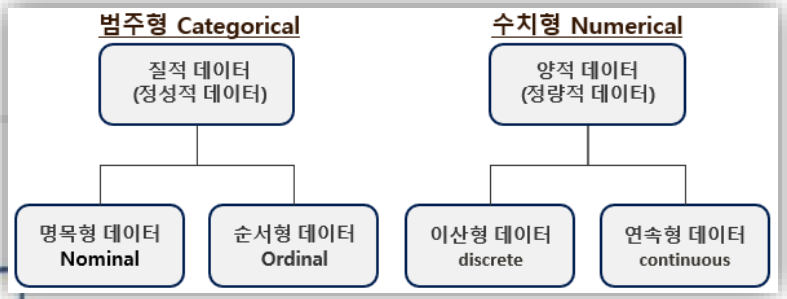
3.5명 ?
소수점의 의미가
없음

수치적인 의미는 가지고 있으나
소수점으로 표현되지 않는 경우
= 이산형 데이터!

EX) 불량품 수 / 사고 건수 / 차량대수

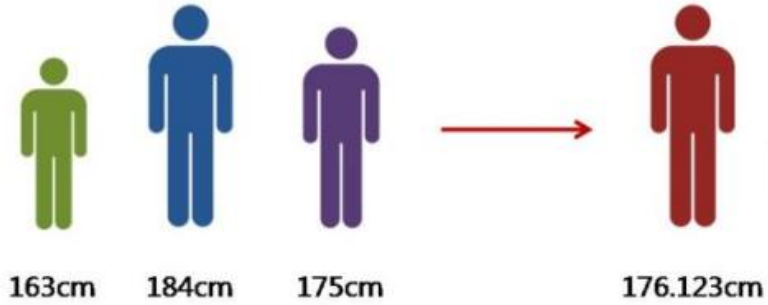
데이터

데이터 분류(Data Classification) - 데이터 특성 (Attribute)



- 수치형
  - L 연속형

Q. 키, 몸무게 등



가질 수 있는
값이 실수인 경우

수치적인 의미가 있으며
소수점으로 표현되는 경우
= 연속형 데이터!

EX) 몸무게 시간 길이

데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

정형 데이터

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/6/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/6/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

attribute

item

cell

데이터

용어 정리

데이터 분류(Data Classification) - 데이터 유형 (Category)

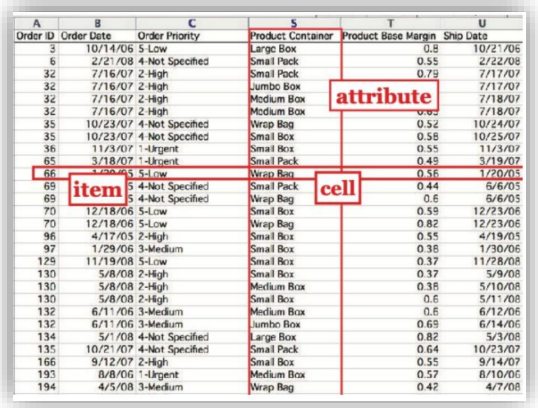
정형 데이터

Rows:

- Record
- Observation
- Instance
- Sample
- Tuple
- Entry
- Case
- Item
- DataPoint

Columns:

- Attribute
- Field
- Feature
- Variable
- Dimension
- Property



데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

정형 데이터

- 타이타닉 데이터 셋

	passengerId	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S

데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

시계열 데이터

: 일정한 시간 간격으로 측정된 데이터

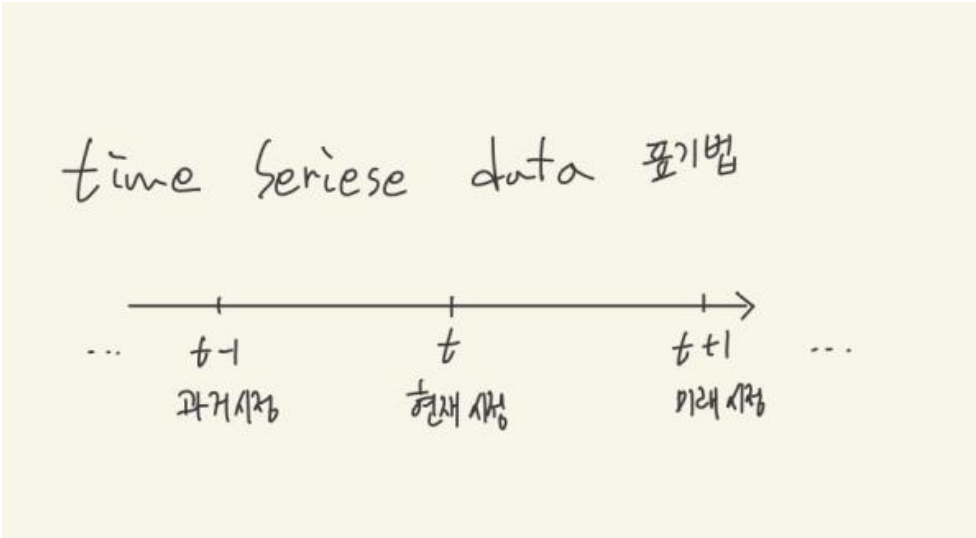
데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

시계열 데이터

시계열 : 일정 **시간 간격**으로 배치된 데이터들의 수열
시계열 해석 : 시계열을 해석하고 이해하는 데 쓰이는 여러 가지 방법을 연구하는 분야. - wiki

Time-series data : **시간 순서**에 따라 관측된 데이터
주식 가격, 기상 정보, 웹사이트의 사용자 트래픽 등 다양한 분야에서 사용
시계열 데이터의 주요 특징은 시간적 순서를 따르는 점과 연속성이 있음



데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

시계열 데이터

Time-series data : **시간 순서**에 따라 관측된 데이터

Air Passengers

Month	#Passengers
Jan-49	112
Feb-49	118
Mar-49	132
Apr-49	129
May-49	121
Jun-49	135
Jul-49	148
Aug-49	148
Sep-49	136
Oct-49	119

Temperature Data

1960_3_m	1960_4_m	1960_5_m	1960_6_m	1960_7_m	1960_8_m
7.9	3.6	13.5	17.3	24	27.1
8.5	4.2	17	19	24.1	27.9
6.4	7.4	21.5	19.2	24.5	28.6
8.4	8.2	17.6	18	25.4	28.8
8.7	6	15.6	19.9	25.2	26.2
12.7	9.2	13.4	20.2	24.3	28
13.7	6.1	14.8	21.4	23.3	28.4
12.1	9.3	14.5	23.5	24.8	28.1

데이터

시계열 데이터

Time-series data : **시간 순서**에 따라 관측된 데이터

Q. 시계열 데이터 분석?

데이터

시계열 데이터

Sequence Data

- 시퀀스 데이터란 **순서대로 정렬된 데이터의 연속**이다.
- 이 데이터 유형은 특정 순서에 따라 배열된 항목들로 구성되며, 각 항목의 순서는 데이터 해석에 있어 중요한 역할을 한다.
- 시퀀스 데이터는 다양한 분야에서 나타나며, 예를 들어 생물학에서는 DNA 서열,
  컴퓨터 과학에서는 문자열이나 시간에 따른 이벤트 로그, 금융에서는 시간에 따른 주식 가격 변동 같은 형태로 나타난다.

  - wiki

데이터

시계열 데이터

Sequence Data

- 시퀀스 데이터란 **순서대로 정렬된 데이터의 연속**이다.

“The movie is not fun”

```
 9 embedding_layer_weights[4]

Embedding layer weights shape: (10000, 64)
Initial vector for the first word:
 [-0.00259084  0.01449068 -0.0360497   0.04829049  0.03736104  0.0439479
  0.04184875  0.03720633 -0.00947944 -0.02207879  0.01567849  0.01323043
 -0.00623485 -0.02969201  0.02015174  0.0419557   0.03964961  0.03541067
 -0.04929231 -0.02532177 -0.01945633  0.0294967  -0.02734786 -0.00235792
  0.04007456  0.01345647 -0.02578268  0.02072446 -0.01814718 -0.00438038
 -0.00013449  0.04992953  0.00172102 -0.02740901 -0.04531569  0.01248584
  0.03048811 -0.01235138 -0.03553003 -0.0029698  -0.00998558 -0.0273586
 -0.01493003 -0.00096365 -0.0060817   0.01747804 -0.00963595  0.04506803
 -0.02521282 -0.04632554 -0.0016393  -0.02918473  0.02854191 -0.00418777
 -0.01391535 -0.01575425 -0.00093541 -0.00559113 -0.02639507  0.0006377
  0.01263886 -0.00140343  0.00334054  0.03750087]
```

데이터

Data Category

- 정형 데이터 (Structured Data)
- 시계열 데이터
  - Timeseries Data
  - Sequence Data
- 비정형 데이터 (Unstructured Data)

데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

비정형 데이터

데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

비정형 데이터

정형 데이터

```
#longitude,latitude,city name
145.768,-16.915,"Cairns"
146.801,-19.265,"Townsville"
150.501,-23.365,"Rockhampton"
139.485,-20.715,"Mount Isa"
150.893,-34.423,"Wollongong"
151.785,-32.932,"Newcastle"
141.451,-31.965,"Broken Hill"
145.951,-30.082,"Bourke"
150.932,-31.091,"Tamworth"
149.581,-33.417,"Bathurst"
153.118,-30.315,"Coffs Harbour"
```

CSV

```
<?xml version="1.0"?>
- <job>
  - <production>
      <ApprovalType>WebCenter</ApprovalType>
      <Substrate>carton 150 gr</Substrate>
      <SheetSize>220-140</SheetSize>
      <press>SuperFlat2</press>
      <finishing>standard</finishing>
      <urgency>normal</urgency>
    </production>
  - <customer>
      <name>FruitCo</name>
      <number>2712</number>
      <currency>USD</currency>
    </customer>
</job>
```

XML

JSON

비정형 데이터

유연성과 확장성

분석 용이성

데이터

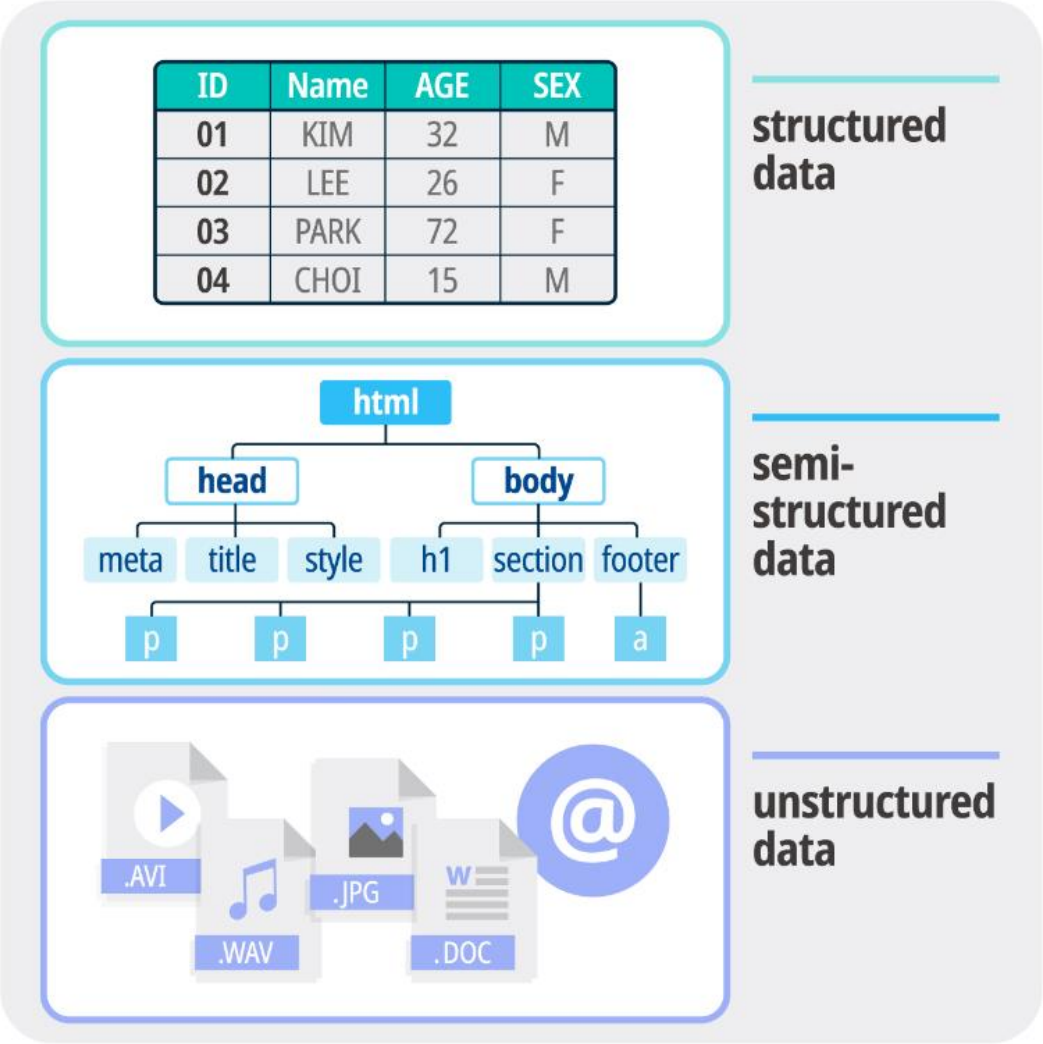
데이터 분류(Data Classification) - 데이터 유형 (Category)

특징	정형 데이터	비정형 데이터
정의	잘 정의된 데이터 모델에 따라 구성 일반적으로 테이블 형태로 저장 ex) 데이터베이스, 엑셀 스프레드시트, CSV 파일 등	고정된 구조가 없는 데이터 ex) 텍스트 문서, 이미지, 오디오, 비디오, 이메일 본문, 웹 페이지 콘텐츠 등
데이터 특성	양적 데이터	질적 데이터
데이터 포맷	제한된 포맷	다양한 포맷
분석	분석 용이 분류, 회귀, 클러스터링 등 다양한 분석 가능	분석 어려움 텍스트 분석, 자연어 처리(NLP), 이미지 인식, 음성 인식 등의 고급 분석 기술 필요
DB	관계형 데이터 베이스	NoSQL 데이터 베이스

데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

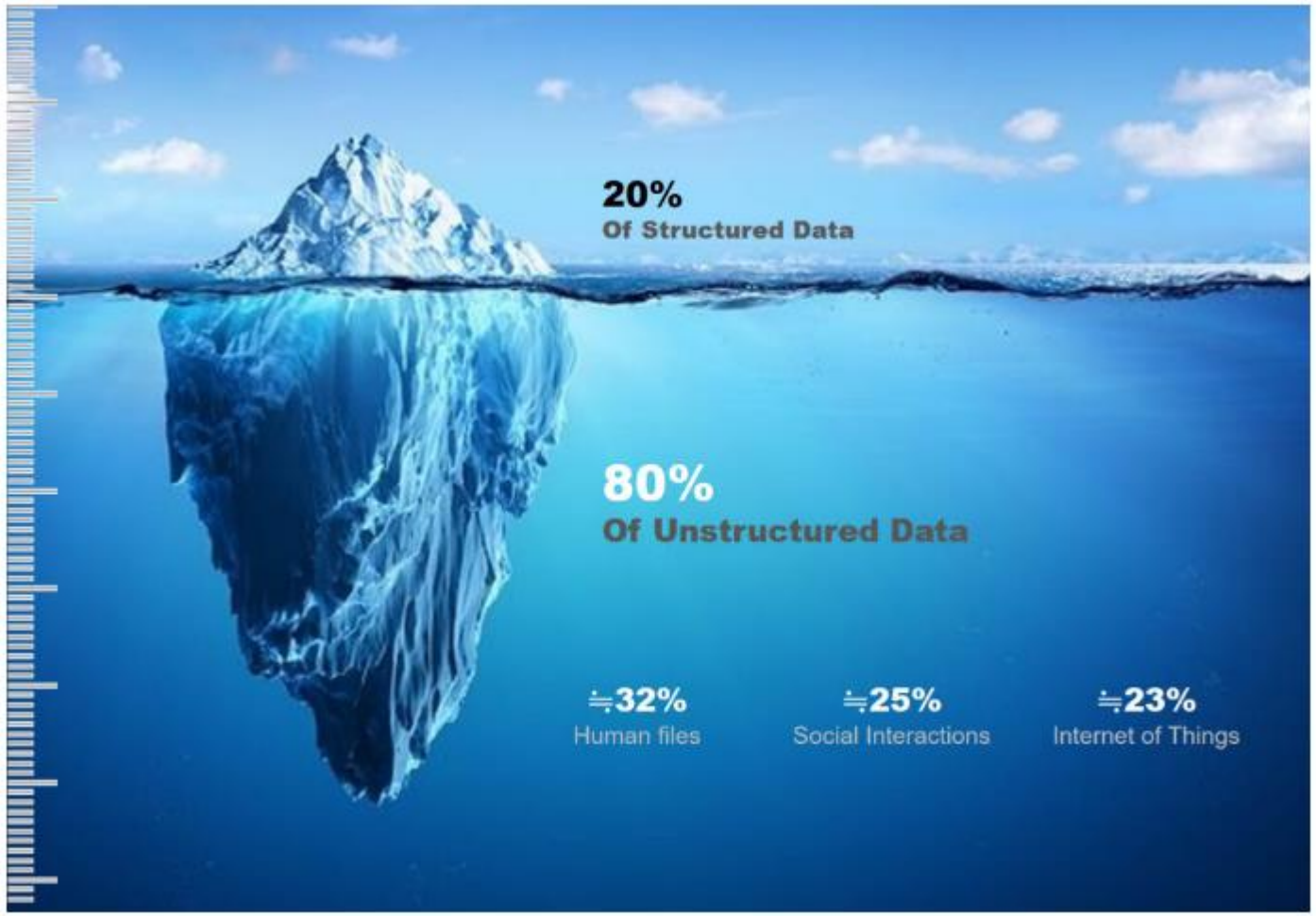
비정형 데이터



데이터

데이터 분류(Data Classification) - 데이터 유형 (Category)

비정형 데이터

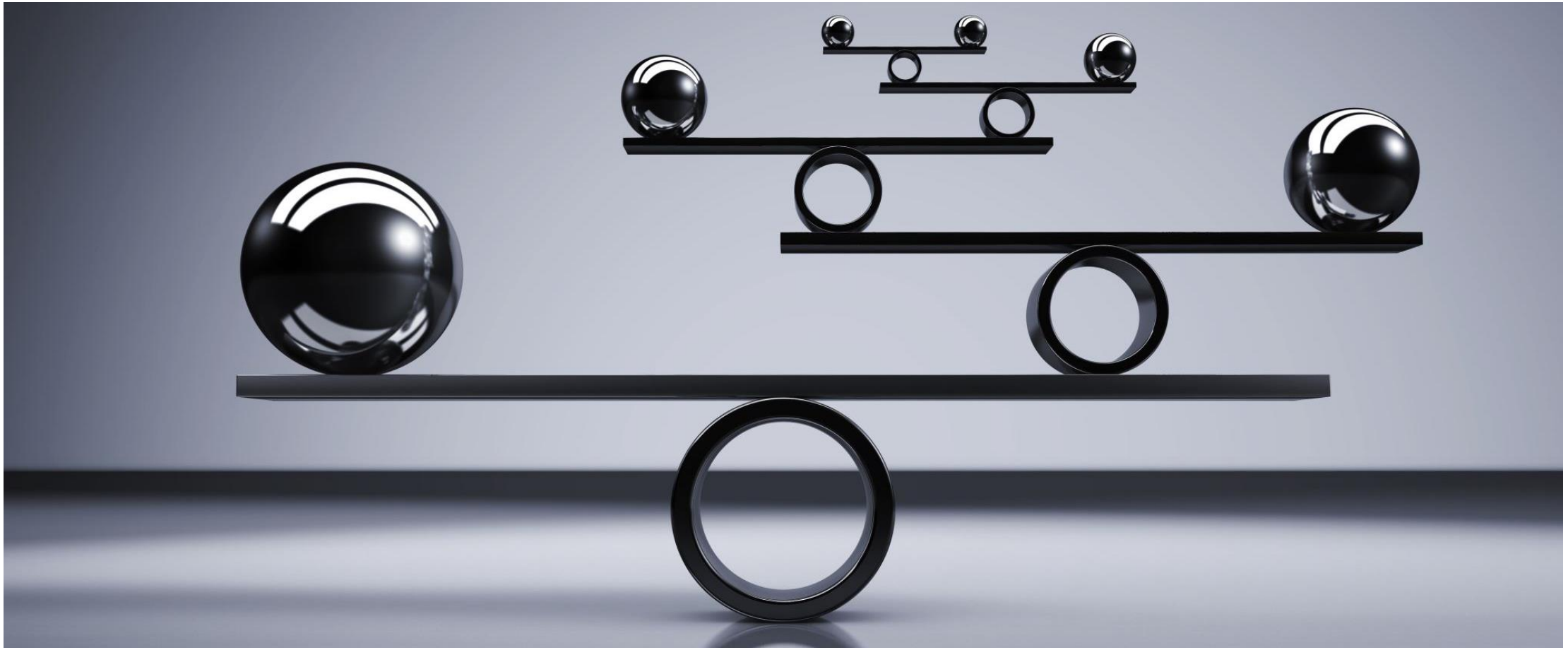


데이터, 변수

- 실습

  # Data.Classification.ipynb

Variable Classification



변수

변수 분류(Variable Classification) - 변수의 수에 따른 분류

- **단변량(Univariate)**
  - 하나의 변수만 분석하는 경우
  - ex, 특정 반의 학생들의 키(만) 분석
  - **분석 방법 :** 히스토그램, 박스플롯 등

- **다변량(Multivariate)**
  - 두 개 이상의 변수를 동시에 분석하는 경우
  - ex, 학생들의 키와 몸무게를 함께 분석
  - **분석 방법** : 산점도, 상관 행렬, 다변량 회귀 분석 등

상세 내용 2.01.EDA.4th.pdf 참조

변수

변수 분류(Variable Classification) - 변수의 유형에 따른 분류

- **종속 변수(Dependent Variable)**
  - 다른 변수에 의해 영향을 받는 변수, 일반적으로 예측하려는 대상, 정답
  - ex, 집값을 예측하는 모델에서의 집값

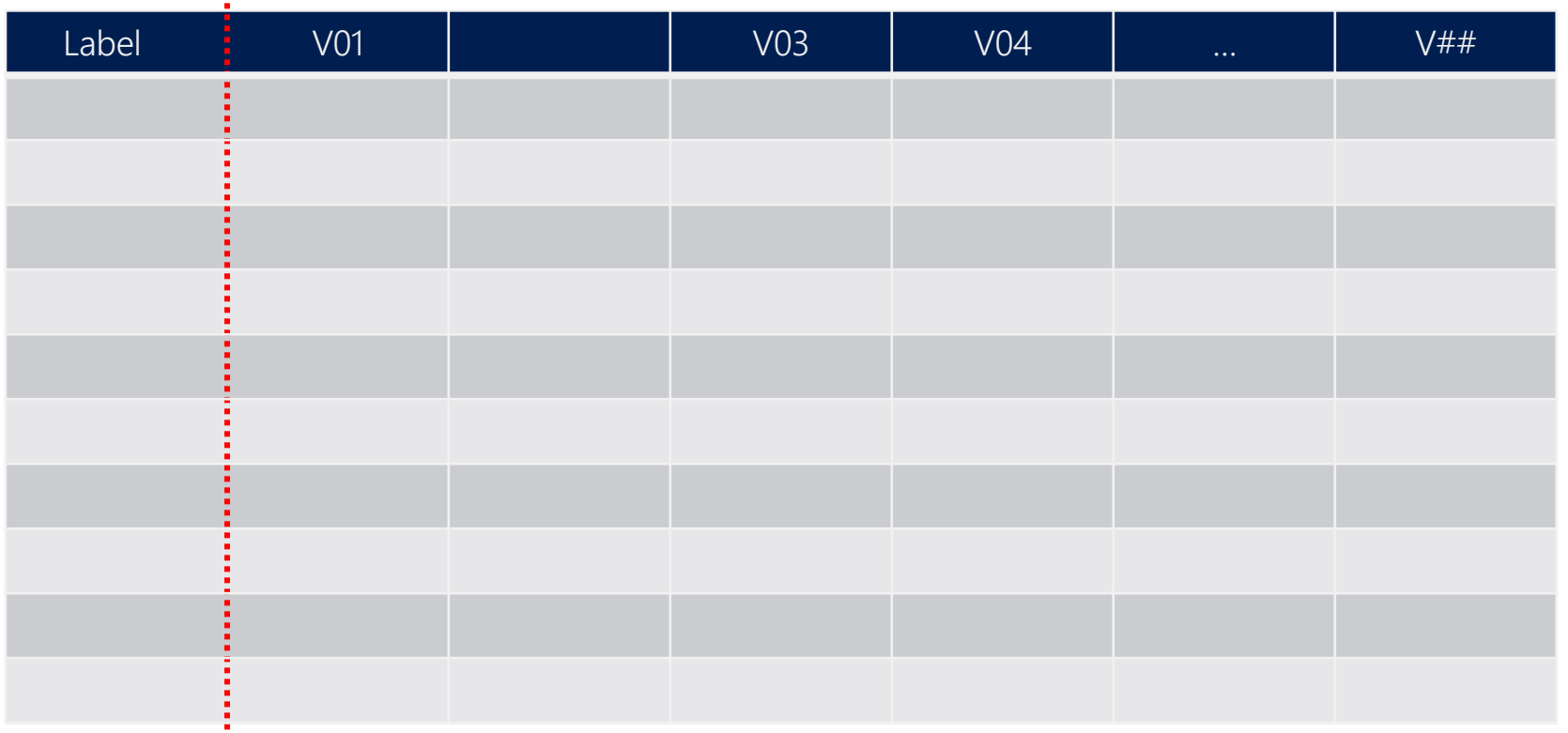
- **독립 변수(Independent Variable)**
  - 종속 변수에 영향을 미치는 변수, 독립적으로 조작되거나 변화할 수 있는 변수
  - ex, 집값을 예측하는 모델에서 집의 크기, 위치 등

변수

변수 분류(Variable Classification) - 변수의 유형에 따른 분류

X

Target,
Label,
Output,
종속 변수



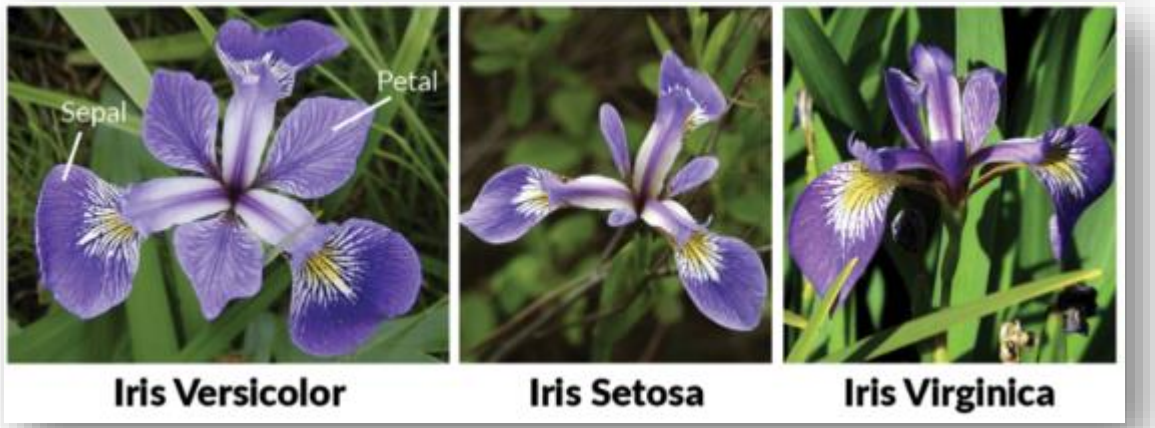
y

Feature,
colums,
input,
요인,
독립 변수

변수

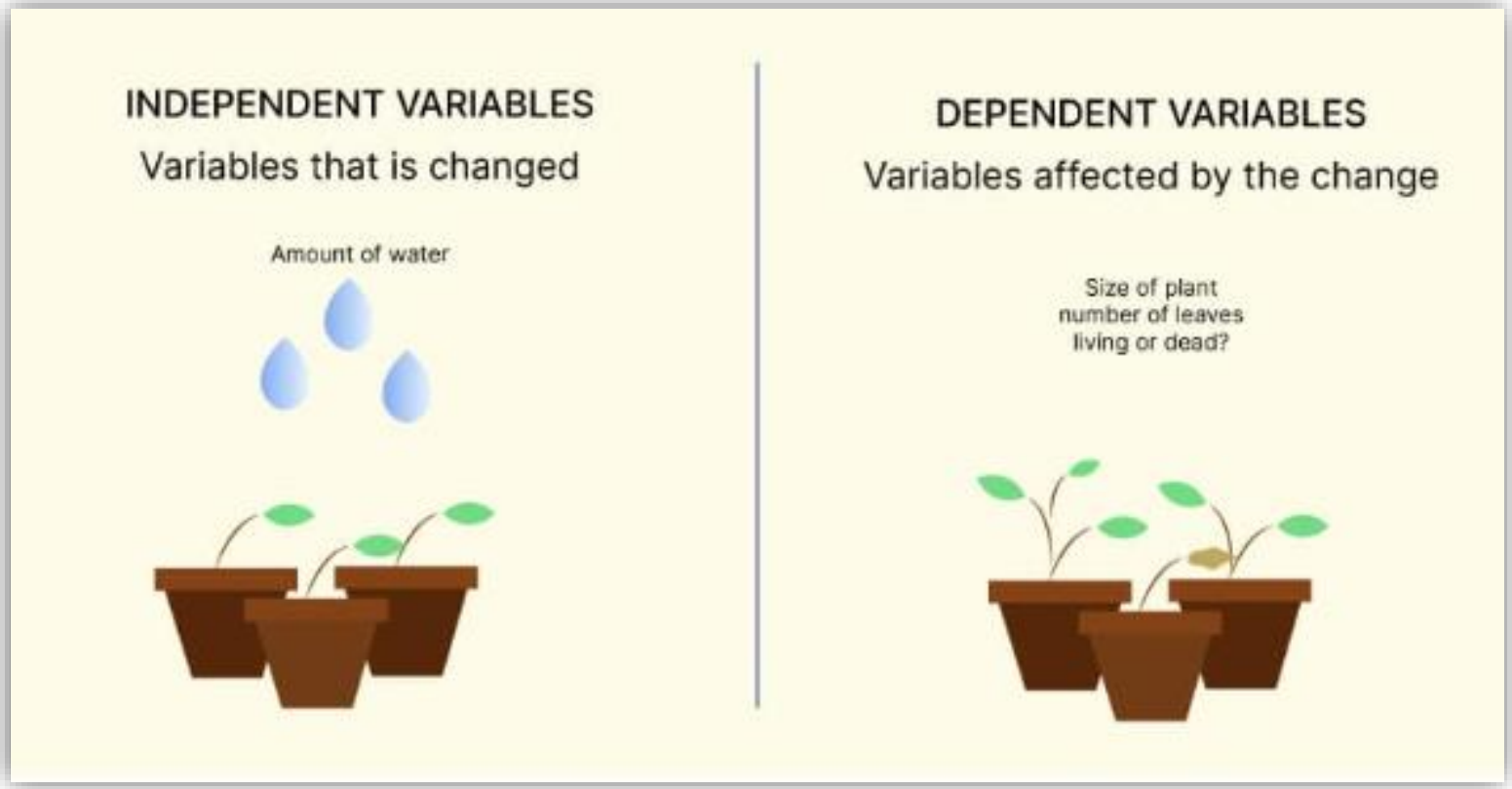
변수 분류(Variable Classification) - 변수의 유형에 따른 분류

```
X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.3, random_state=42)
```



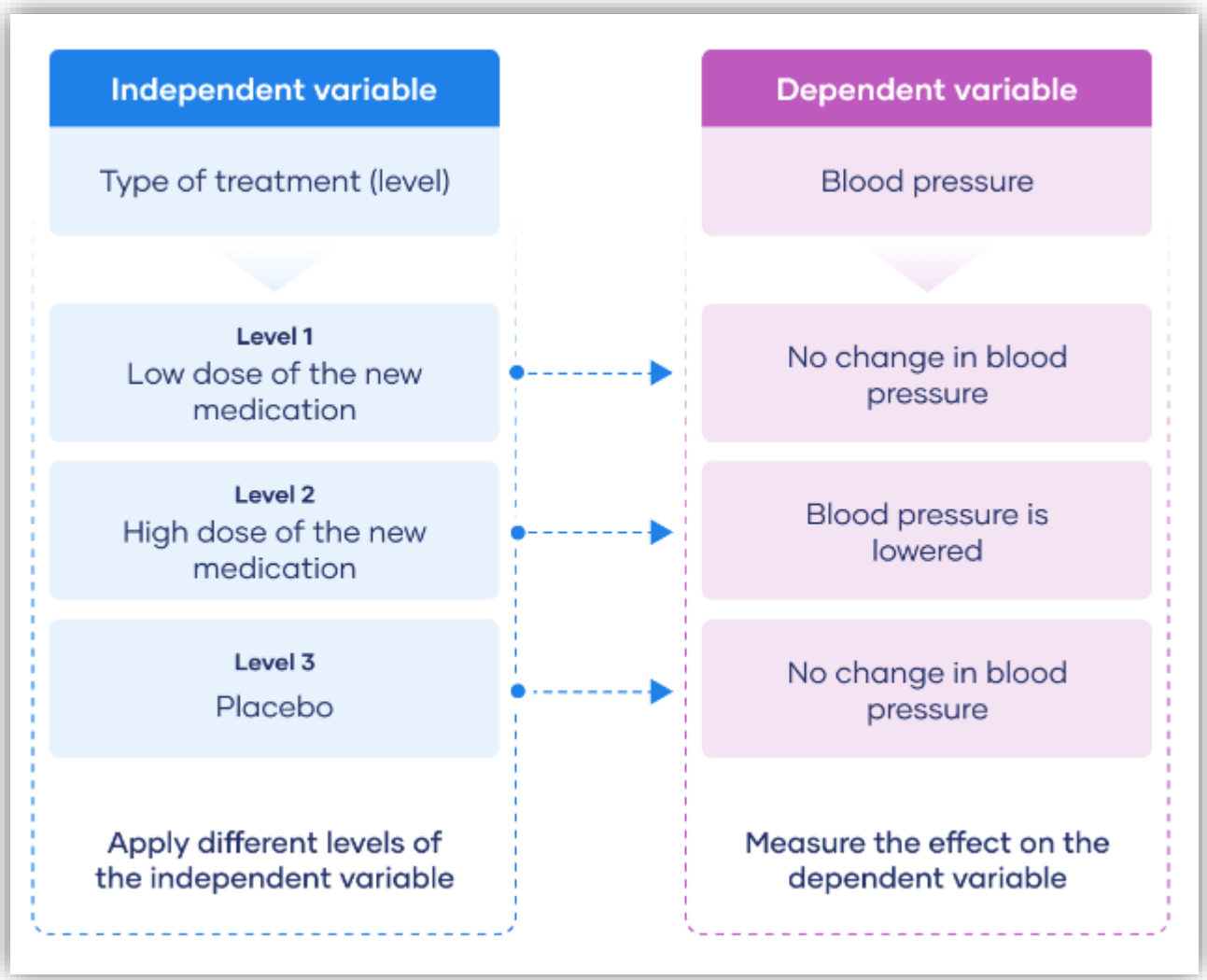
변수

변수 분류(Variable Classification) - 변수의 유형에 따른 분류



변수

변수 분류(Variable Classification) - 변수의 유형에 따른 분류



변수

용어 정리

변수 분류(Variable Classification) - 변수의 유형에 따른 분류

▪ 독립 변수 (InDependent Variable)

- **Feature(특성)** : 기계 학습 및 데이터 과학
- **Colums**: 행, 열
- **요인**: 결과에 영향을 미치는 변수
- **예측 변수**: 회귀 분석 및 기계 학습
- **설명 변수**: 회귀 분석 및 기타 통계 모델
- **입력 변수**: 모델 입력 변수
- **회귀 변수**: 회귀 모델
- **조작 변수**: 연구자가 독립 변수를 제어하는 실험 연구

변수

용어 정리

변수 분류(Variable Classification) - 변수의 유형에 따른 분류

▪ 종속 변수 (Dependent Variable)

- **Target :** 목표 변수
- **Label :** 이미지 분석
- **응답 변수**: 통계 모델링 (ex, 회귀 분석)
- **결과 변수**: 임상시험, 역학, 사회과학
- **대상 변수**: 기계 학습 (ex, 지도 학습)
- **예측변수**: 모델에 의해 예측 또는 추정되는 변수
- **기준 변수**: 심리학 연구, 사회 과학
- **설명 변수**: 독립변수에 의해 설명되는 변수
- **Output :** 결과

데이터, 변수

AI 개발

"Applied machine learning is basically feature engineering"
— Andrew Ng

When working with a paucity of data, or less feature-rich data, which is all too common for data scientists tasked with coming up with predictions based on just a dozen or so features, *feature engineering* is essential to **eke and tease out all the available 'signal' that's present in the limited data**; as well as to **overcome the limitations of popular machine-learning algorithms**, for example, difficulty in separating data based on multiplicative or divisive feature interactions.

Artificial Neural Network Modeling

If we can **train a model to map X to Y** based on a labelled dataset then it can be used to predict ….

데이터, 변수

Artificial Neural Network Modeling

If we can **train a model to map X to Y** based on a labelled dataset then it can be used to predict ….

THANK YOU