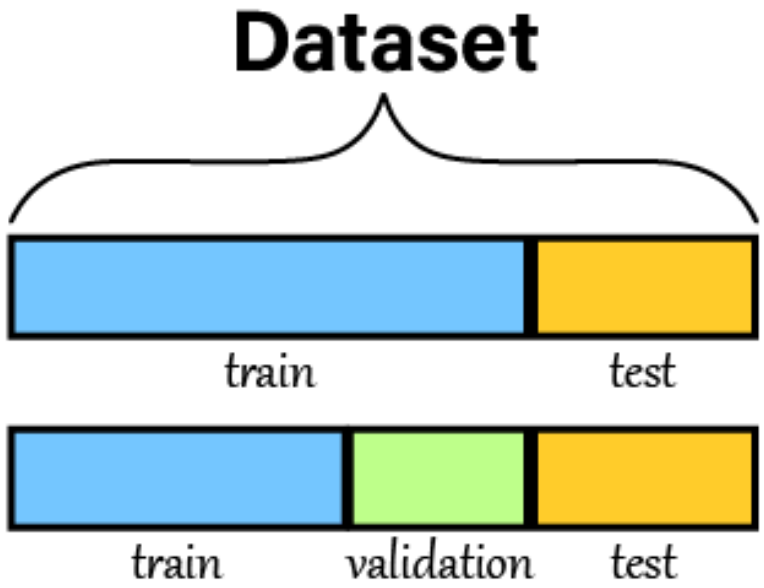
인공지능 모형의 정확도 검증 : 데이터 세트 분할

우리가 배운 머신러닝의 과정

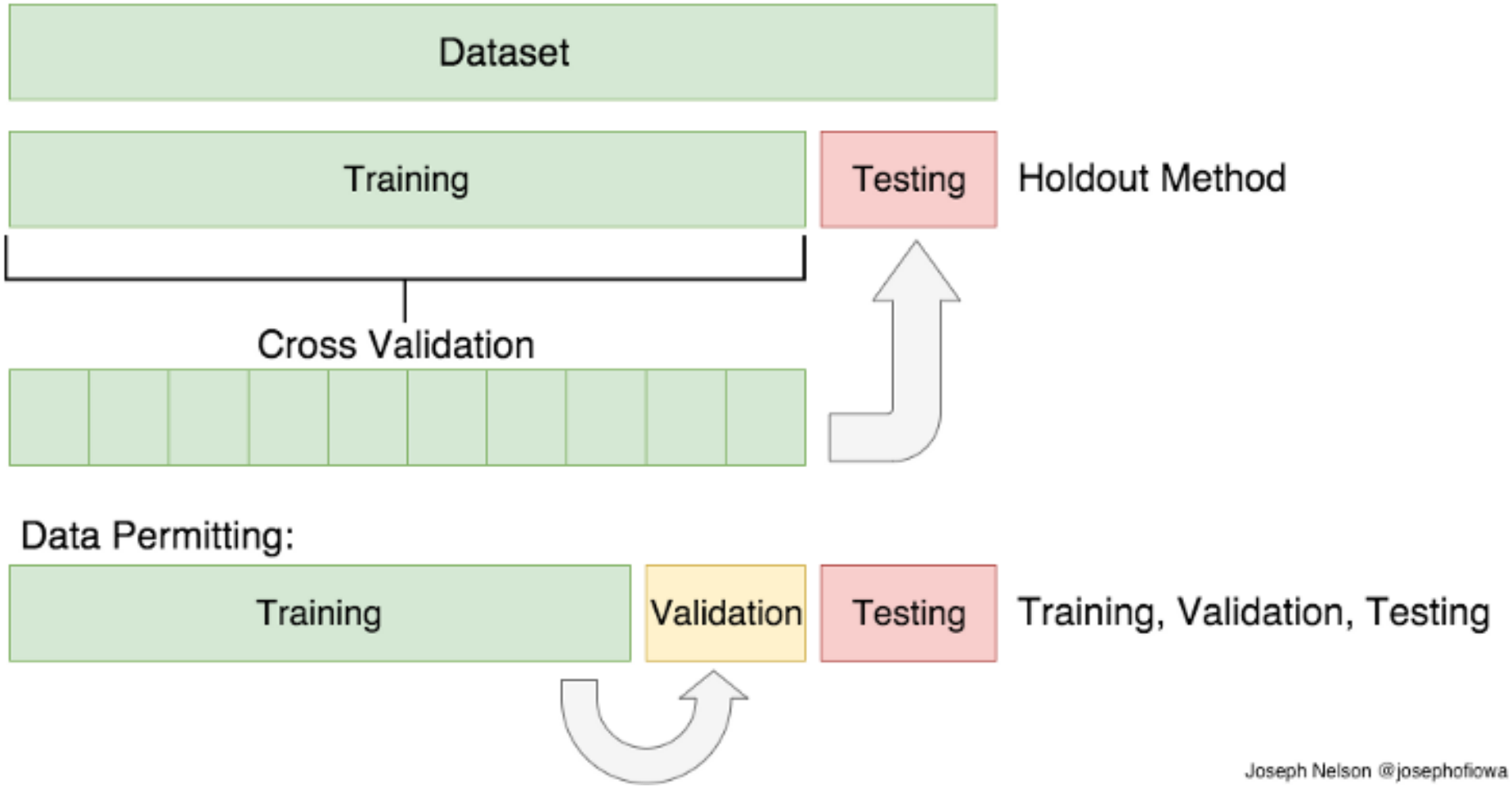
7:3?



8:2?

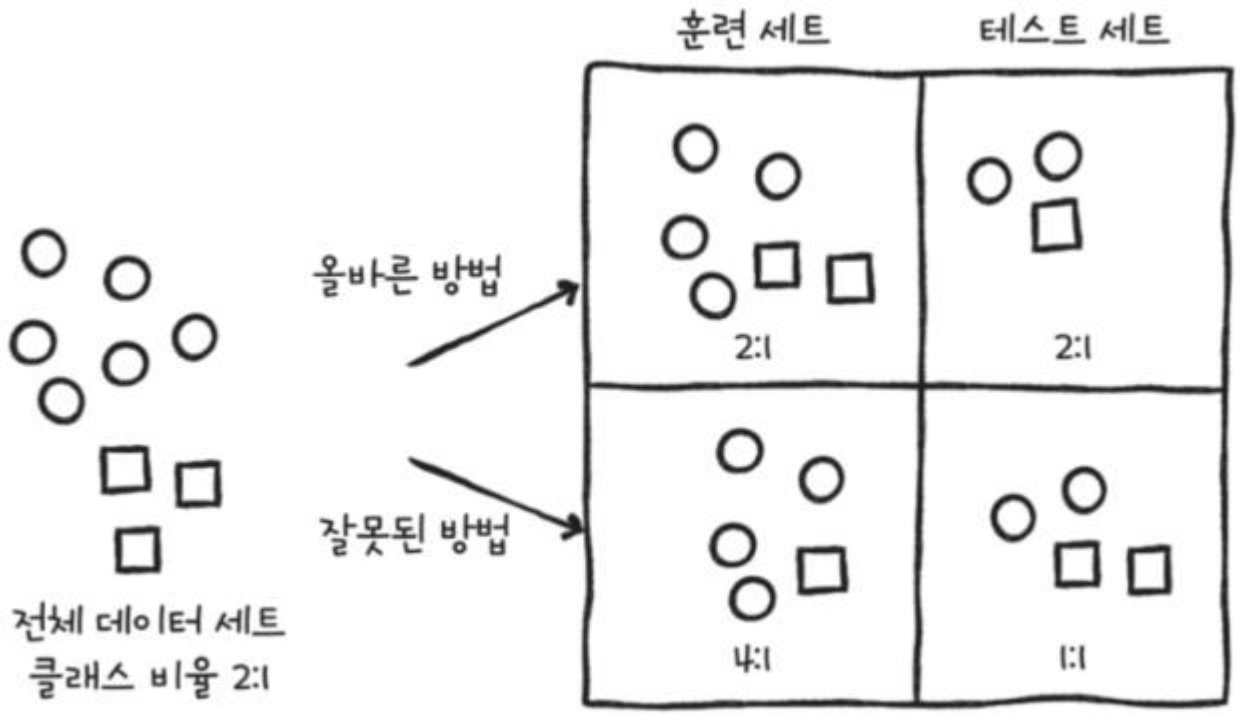
DATADOCTOR

우리가 배운 머신러닝의 과정

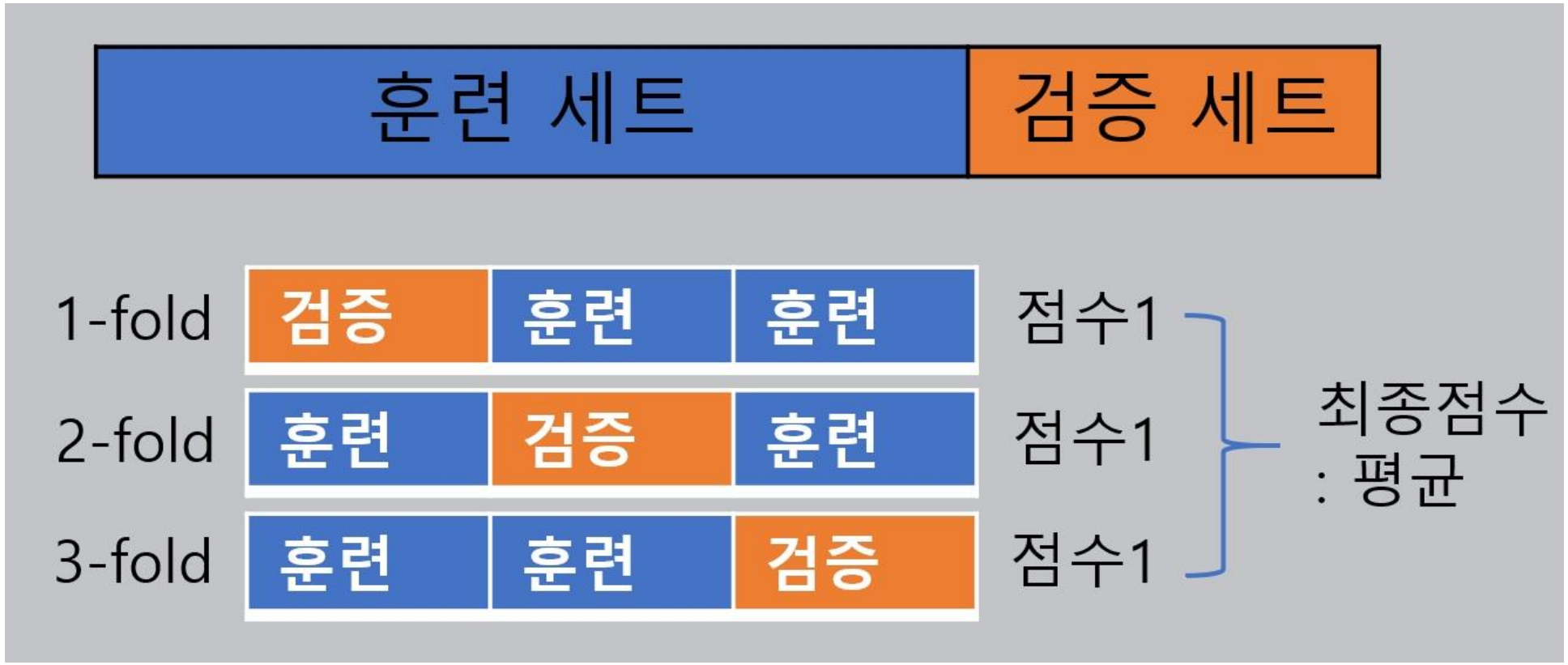


우리가 배운 머신러닝의 과정

훈련 세트 / 테스트 세트로 나누는 과정이 중요



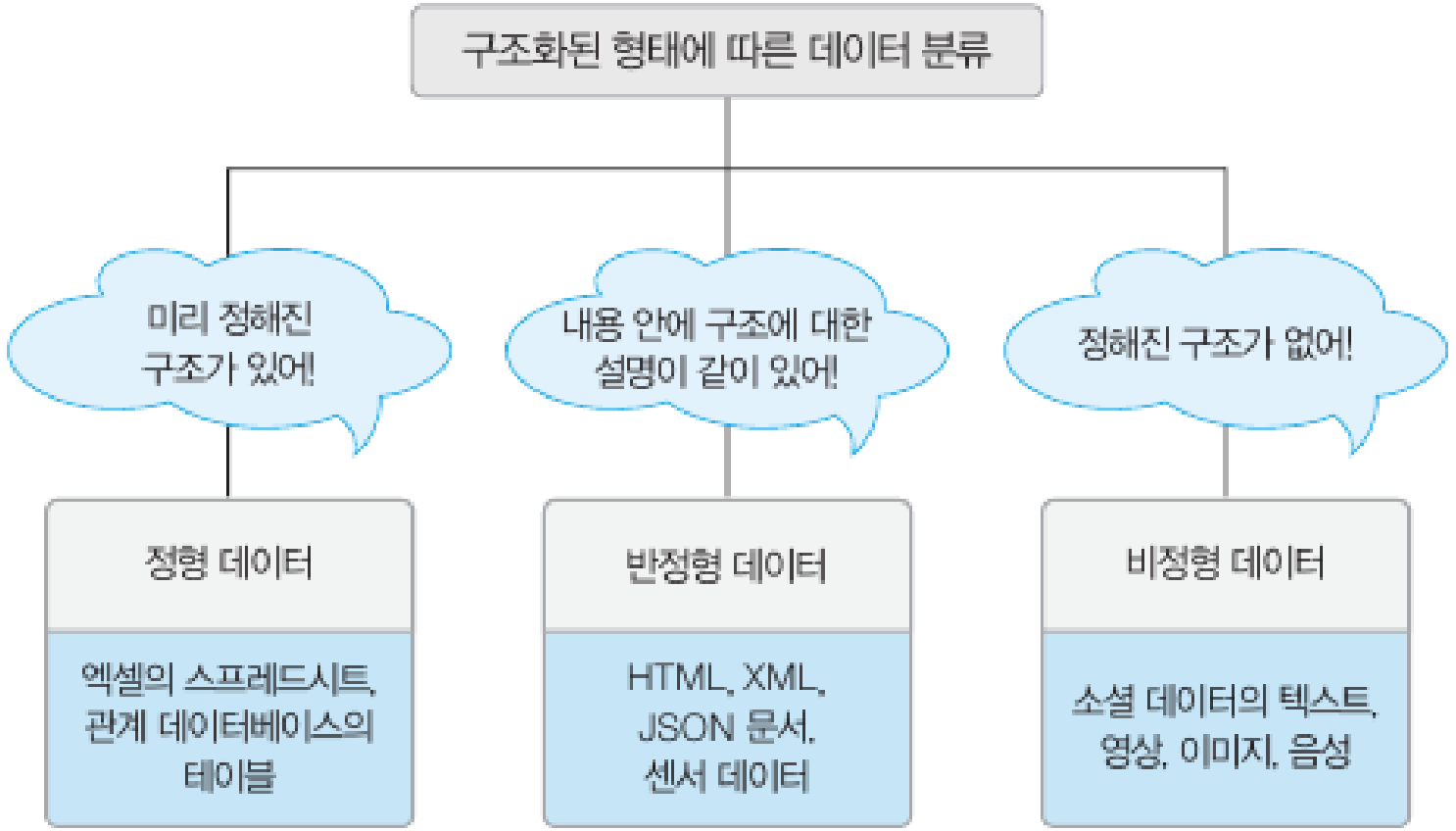
정확도는 매번 달라질 수 밖에 없다.



분할 하고자 하는 데이터 세트의 형식

데이터의 분류

데이터의 정의



데이터의 분류

🏦 형태에 따른 데이터의 분류

정형데이터 : 미리 정해진 구조에 따라 저장된 데이터

	A	B	C	D
1	일자	배송 업체	배송 건수	전일대비 상승률
2	2022-03-02	빠르다 택배	100	0%
3	2022-03-02	한빛 택배	200	10%
4	2022-03-02	안전 택배	50	3%
5	2022-03-02	당일 택배	30	-10%

NOTE 미리 정해진 데이터 구조를 스키마schema 라 한다.

데이터의 분류

형태에 따른 데이터의 분류

반정형데이터 : 구조에 따라 저장된 데이터 + 데이터 안에 구조에 대한 설명이 존재

```
{
  "이름" : "오형준",
  "나이" : 23,
  "성별" : "남"
}
```

(a) JSON

```
<친구정보>
  <이름> 오형준 </이름>
  <나이> 23 </나이>
  <성별> 남 </성별>
</친구정보>
```

(b) XML

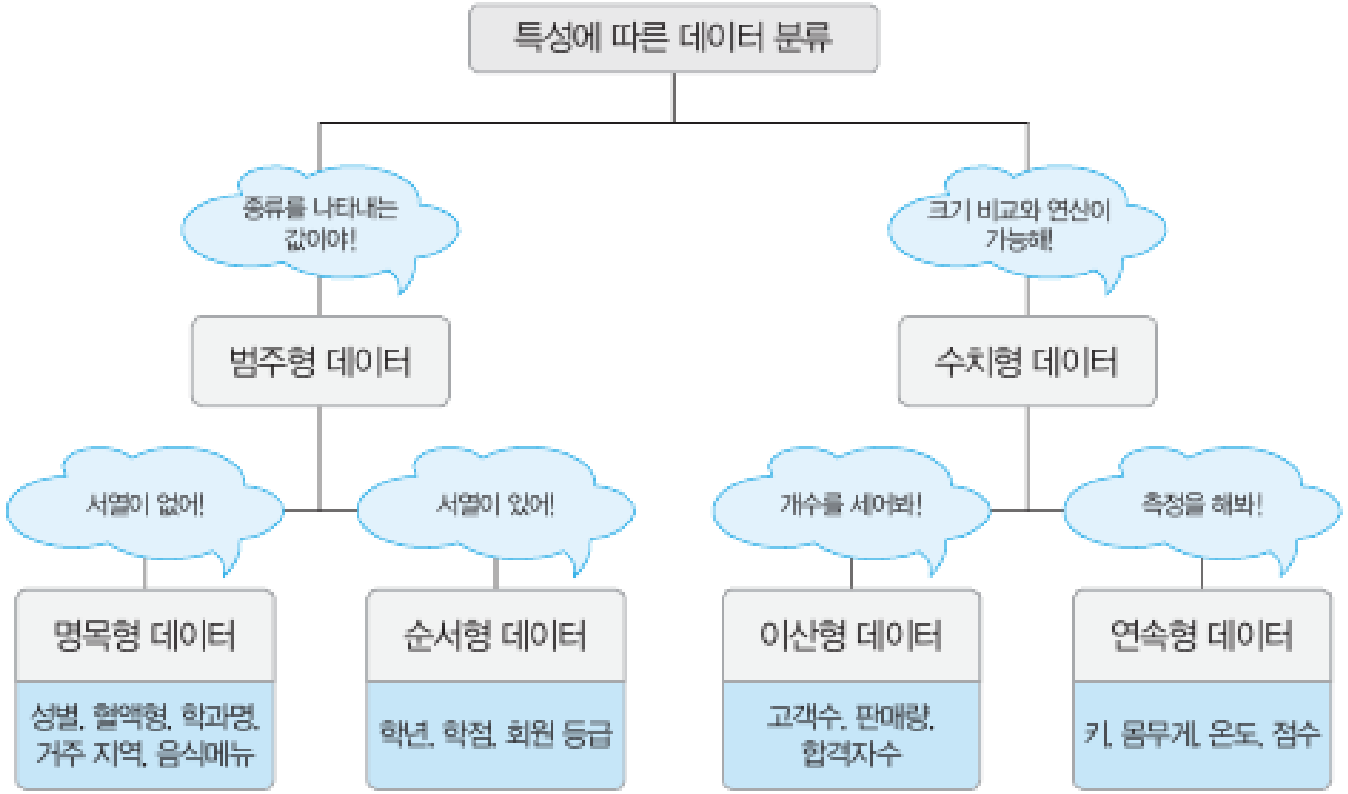
데이터의 분류

🏦 형태에 따른 데이터의 분류

비정형데이터 : 정해진 구조가 없이 저장된 데이터

데이터의 분류

특성에 따른 데이터의 분류



데이터의 분류

🏛 특성에 따른 데이터의 분류

- 범주형 자료(질적, 비계량)
  - 명목변수(Nominal Variable) : 개체나 사람의 특성만 (성별, 종교 등)
  - 순서(서열)변수(Ordinal Variable) : 측정대상 간 선호를 부여 (선호도, 만족도)

- 양적인 자료(양적, 계량)
  - 등간변수(Interval Variable) : 측정대상 간의 순서 + 값 사이의 간격이 일정 (IQ,온도)
    절대 영점이 없음(100도는 50도의 2배 뜨겁다?)
  - 비율변수(Ratio Variable) : 측정대상 간에 비율 계산이 가능함(연령, 무게, 거리, 시간)
  - 이산형 변수(Discrete Variable) : 점수, 빈도수
  - 연속형 변수(Continuous Variable) : 실수, 키, 몸무게

데이터의 분류

🏦 특성에 따른 데이터의 분류

연속형 변수와 이산형 변수

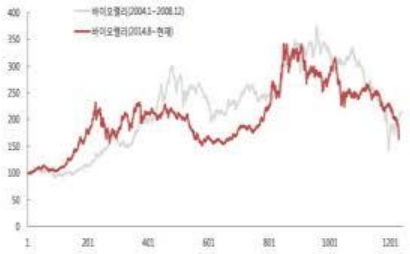
- **연속형 변수 : 시간에 대해서 행동과 상태가 유한차원 벡터공간**

  : 주가 수익률, 자동차의 운행거리,

- **이산형 변수 : 시간에 대해서 행동과 상태가 원소형태의 값(Value)를 갖는다.**

  : {라이트를 켠다, 라이트를 끈다, 매도한다, 매수한다}

연속형 변수 : 수익률



이산형 변수 : 전원 상태

데이터의 분류

특성에 따른 데이터의 분류

테이블 정의서			작성자	김설계		승인자		김관리	
			작성일	2018.08.30		버전		1.0	
단계	설계		업무명	스마트영업지원시스템		페이지수		4	
순번	테이블명	테이블ID	컬럼명	컬럼ID	타입/길이	PK여부	FK여부	NULL여부	비고
1	영업일지	TB_BUSI_REPT	영업일지 아이디	C_BUSI_REPT_ID	CHAR(10)	Yes	No	NOT NULL	
2	영업일지	TB_BUSI_REPT	사원번호	C_EMP_NO	CHAR(5)	Yes	No	NOT NULL	
3	영업일지	TB_BUSI_REPT	부서번호	C_DEPT_NO	CHAR(5)	Yes	No	NOT NULL	
4	영업일지	TB_BUSI_REPT	고객사 아이디	C_CUST_ID	CHAR(10)	No	No	NOT NULL	
5	영업일지	TB_BUSI_REPT	영업일지	S_BUSI_REPT	VARCHAR2(4000)	No	No	NULL	
6	영업일지	TB_BUSI_REPT	결재상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No		
7	영업일지	TB_BUSI_REPT	방문 시작 시간	D_BUSI_START_DT	DATE	No	No		
8	영업일지	TB_BUSI_REPT	방문 종료 시간	D_BUSI_END_DT	DATE	No	No	NULL	
9	영업일지	TB_BUSI_REPT	방문 위치 X좌표	N_VISIT_LOC_X	NUMBER(10,2)	No	No	NULL	
10	영업일지	TB_BUSI_REPT	방문 위치 Y좌표	N_VISIT_LOC_Y	NUMBER(10,2)	No	No	NULL	
11	영업일지	TB_BUSI_REPT	작성자 아이디	S_INST_ID	VARCHAR2(10)	No	No	NULL	
12	영업일지	TB_BUSI_REPT	작성일시	D_INST_DT	DATE	No	No	NULL	
13	영업일지	TB_BUSI_REPT	수정자 아이디	S_UPDT_ID	VARCHAR2(10)	No	No	NULL	
14	영업일지	TB_BUSI_REPT	수정일시	D_UPDT_DT	DATE	No	No	NULL	
1	영업비용	TB_BUSI_COST	영업비용 아이디	C_BUSI_COST_ID	CHAR(10)	Yes	No	NOT NULL	
2	영업비용	TB_BUSI_COST	영업일지 아이디	C_BUSI_REPT_ID	CHAR(10)	Yes	Yes	NOT NULL	
3	영업비용	TB_BUSI_COST	영업비용	N_BUSI_COST	NUMBER(10)	No	No	NOT NULL	
4	영업비용	TB_BUSI_COST	결재상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No	NULL	
5	영업비용	TB_BUSI_COST	처리상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No	NULL	
6	영업비용	TB_BUSI_COST	작성자 아이디	S_INST_ID	VARCHAR2(10)	No	No	NULL	
7	영업비용	TB_BUSI_COST	작성일시	D_INST_DT	DATE	No	No	NULL	

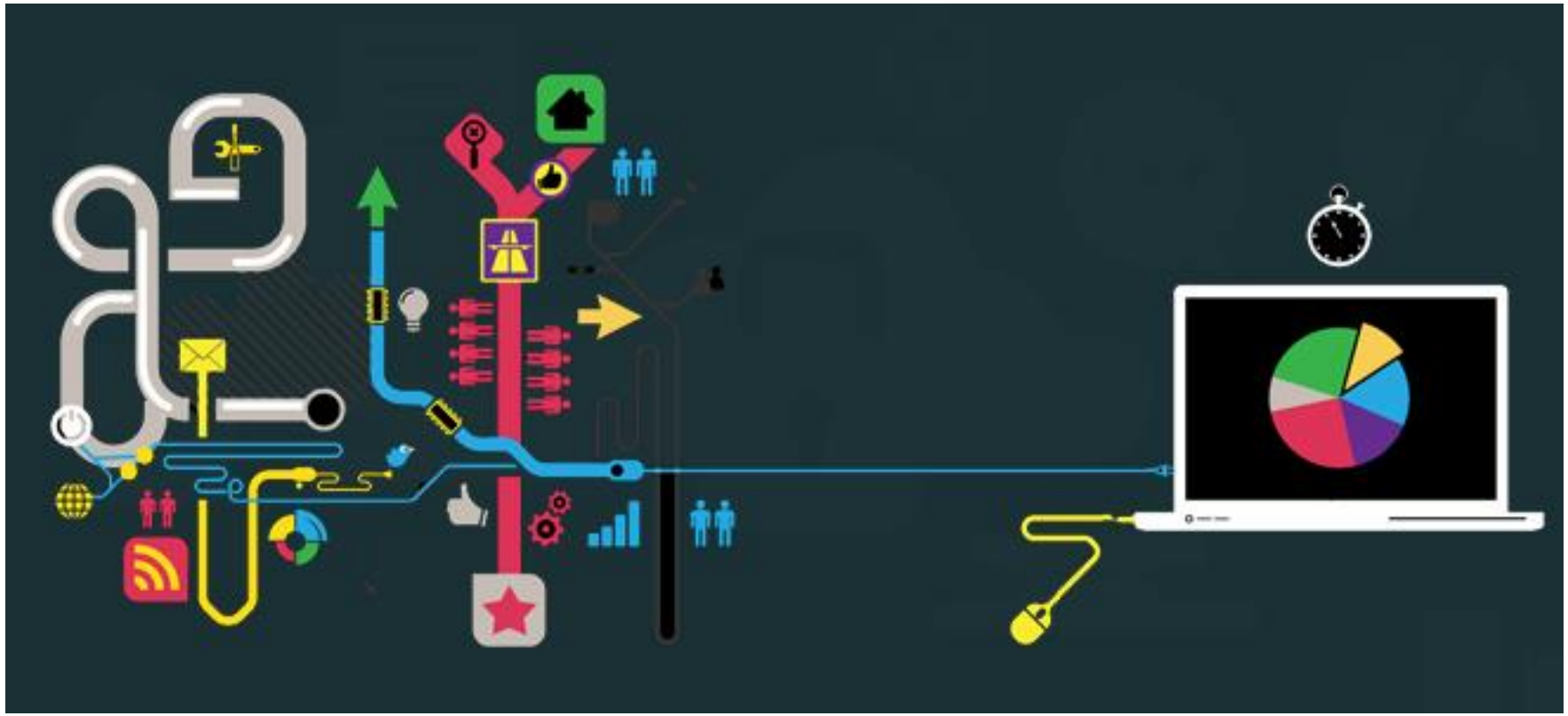
데이터의 분류

특성에 따른 데이터의 분류

테이블 정의서			작성자	김설계		승인자		김관리	
			작성일	2018.08.30		버전		1.0	
단계	설계		업무명	스마트영업지원시스템		페이지수		4	
순번	테이블명	테이블ID	컬럼명	컬럼ID	타입/길이	PK여부	FK여부	NULL여부	비고
1	영업일지	TB_BUSI_REPT	영업일지 아이디	C_BUSI_REPT_ID	CHAR(10)	Yes	No	NOT NULL	
2	영업일지	TB_BUSI_REPT	사원번호	C_EMP_NO	CHAR(5)	Yes	No	NOT NULL	
3	영업일지	TB_BUSI_REPT	부서번호	C_DEPT_NO	CHAR(5)	Yes	No	NOT NULL	
4	영업일지	TB_BUSI_REPT	고객사 아이디	C_CUST_ID	CHAR(10)	No	No	NOT NULL	
5	영업일지	TB_BUSI_REPT	영업일지	S_BUSI_REPT	VARCHAR2(4000)	No	No	NULL	
6	영업일지	TB_BUSI_REPT	결재상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No		
7	영업일지	TB_BUSI_REPT	방문 시작 시간	D_BUSI_START_DT	DATE	No	No		
8	영업일지	TB_BUSI_REPT	방문 종료 시간	D_BUSI_END_DT	DATE	No	No	NULL	
9	영업일지	TB_BUSI_REPT	방문 위치 X좌표	N_VISIT_LOC_X	NUMBER(10,2)	No	No	NULL	
10	영업일지	TB_BUSI_REPT	방문 위치 Y좌표	N_VISIT_LOC_Y	NUMBER(10,2)	No	No	NULL	
11	영업일지	TB_BUSI_REPT	작성자 아이디	S_INST_ID	VARCHAR2(10)	No	No	NULL	
12	영업일지	TB_BUSI_REPT	작성일시	D_INST_DT	DATE	No	No	NULL	
13	영업일지	TB_BUSI_REPT	수정자 아이디	S_UPDT_ID	VARCHAR2(10)	No	No	NULL	
14	영업일지	TB_BUSI_REPT	수정일시	D_UPDT_DT	DATE	No	No	NULL	
1	영업비용	TB_BUSI_COST	영업비용 아이디	C_BUSI_COST_ID	CHAR(10)	Yes	No	NOT NULL	
2	영업비용	TB_BUSI_COST	영업일지 아이디	C_BUSI_REPT_ID	CHAR(10)	Yes	Yes	NOT NULL	
3	영업비용	TB_BUSI_COST	영업비용	N_BUSI_COST	NUMBER(10)	No	No	NOT NULL	
4	영업비용	TB_BUSI_COST	결재상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No	NULL	
5	영업비용	TB_BUSI_COST	처리상태코드	C_SIGN_STAT_CD	CHAR(3)	No	No	NULL	
6	영업비용	TB_BUSI_COST	작성자 아이디	S_INST_ID	VARCHAR2(10)	No	No	NULL	
7	영업비용	TB_BUSI_COST	작성일시	D_INST_DT	DATE	No	No	NULL	

모델 검증을 위한 데이터 분할 : 실제 비즈니스 사례

데이터가 실시간으로 업데이트 될 경우에는?



데이터가 실시간으로 업데이트 될 경우에는?

Real time 데이터일 경우에는 어떻게 데이터를 분할해야 하는가?

- 과거 데이터와 현재 데이터의 가중치?
- 과거 데이터와 현재 데이터의 가중치는 불변인가?
- 과거 데이터 가중치를 어떻게 설정?

실제 사례 예시

모델의 평가 기준 : 정확도? KPI?

모델링의 목적	목표 변수 유형	관련 모델	평가 방법
예측 / 회귀 (Prediction)	연속형	선형 회귀	MSE, RMSE, MAE, MAPE 등
분류 (Classification)	범주형	- 로지스틱 회귀 - 의사결정나무 - 서포트벡터머신	정확도, 정밀도, 재현율, F1 -score

모델 검증을 위한 데이터 분할 :Colab(파이썬)의 함수

1. 데이터의 분할

함수 : train_test_split -> 데이터 세트를 분할

```
 0.]
```

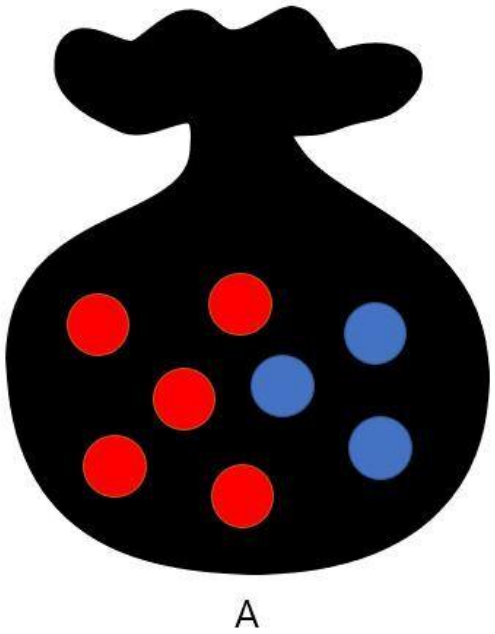
```
from sklearn.model_selection import train_test_split
#train_test_split함수 -> 적정한 비율로 훈련세트와 테스트 세트를 나누어 준다.
```

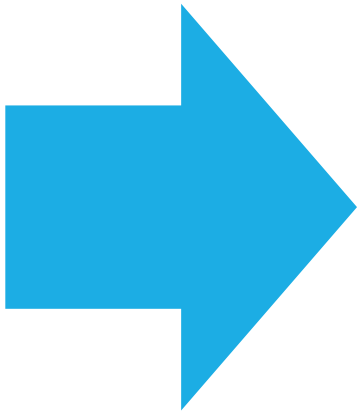
```
[ ]  train_input, test_input, train_target, test_target = train_test_split(fish_data, fish_target, random_state=42)
     #4개 값으로 된 데이터 셀을 묶어서 훈련(train)데이터 셀 테스트 데이터 셀으로 나눈다. 그리고 랜덤하게 바꾸는데 42라는 특정한 패턴으로 진행
```

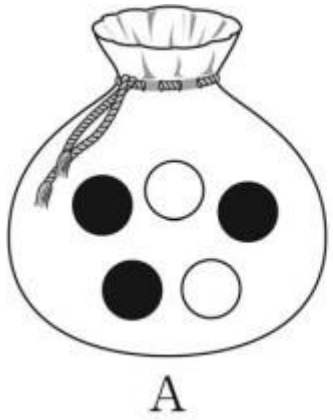
```
print(train_input.shape, test_input.shape)
#훈련(train)과 테스트할 데이터의 특징
#각각 36,13개 값, 공통적으로  2개 변수(길이와 무게)
```

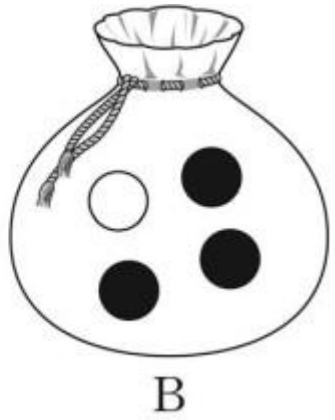
1. 데이터의 분할

함수 : train_test_split -> 데이터 세트를 분할



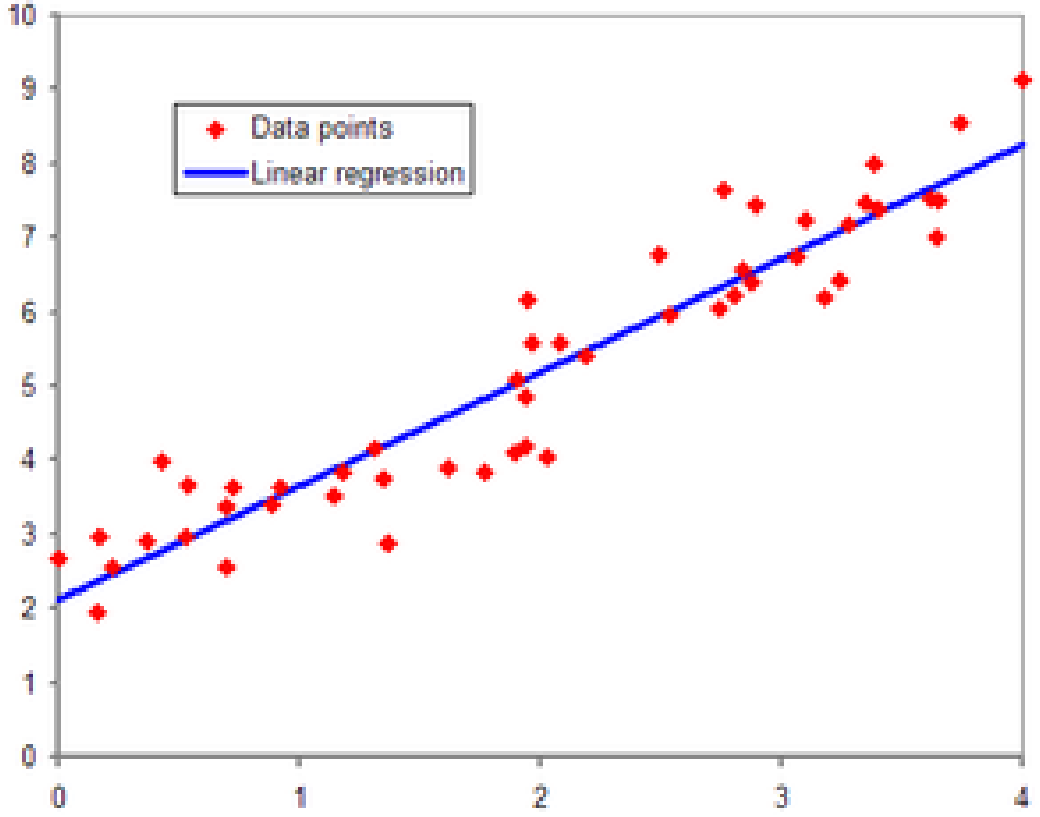


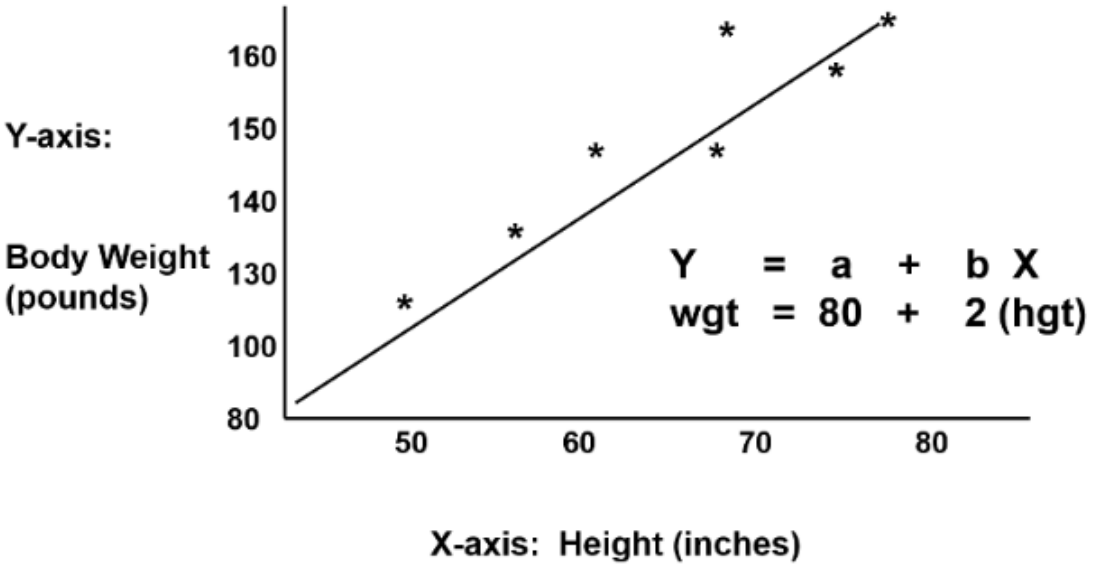




Chapter03 회귀모형

https://colab.research.google.com/drive/110LF1UKWDWHgYxT0k3UFQQLCReb3_psn?usp=sharing


Data points
Linear regression
0
1
2
3
4
0
1
2
3
4
5
6
7
8
9
10


Y-axis:
Body Weight
(pounds)
160
150
140
130
100
80
50
60
70
80
Y = a + b X
wgt = 80 + 2 (hgt)
X-axis: Height (inches)