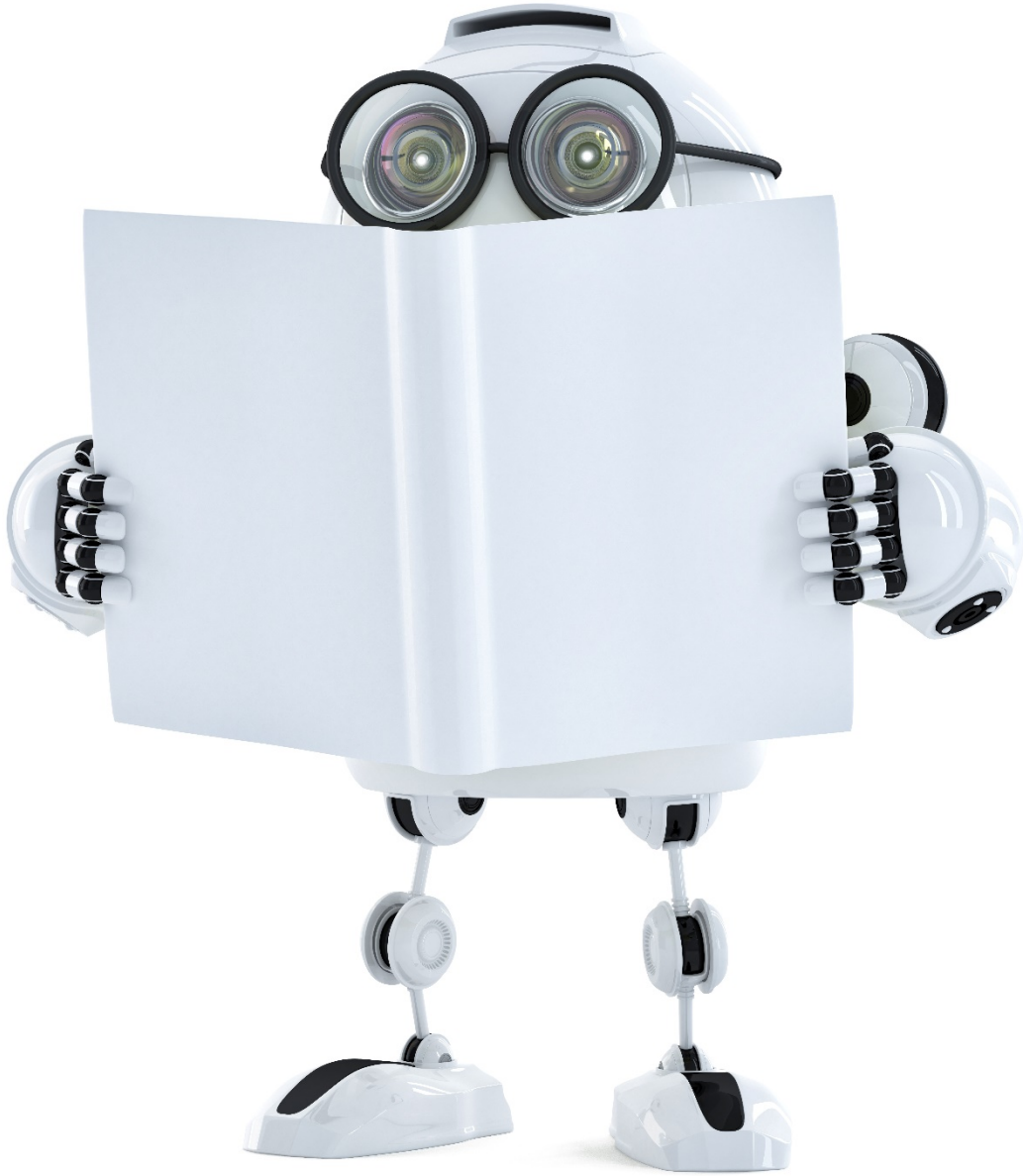
Performance measure

Linear Regression



Director of TEAMLAB
Sungchul Choi

만들어진 모델의 성능은 어떻게 평가할 것인가?

평가할 수 있는
Measure가 필요

Regression metrics

- Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i| \,.$$

잔차의 절대값의 Sum

```
from sklearn.metrics import median_absolute_error
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
median_absolute_error(y_true, y_pred)
```

Regression metrics

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

잔차 제곱의 sum의 루트

```
from sklearn.metrics import mean_squared_error
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
mean_squared_error(y_true, y_pred)
```

Regression metrics

- R squared

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}.$$

0과 1사이 숫자로 크면 클 수록 높은 적합도를 지님

```
from sklearn.metrics import r2_score
y_true = [3, -0.5, 2, 7]
y_pred = [2.5, 0.0, 2, 8]
r2_score(y_true, y_pred)
```

모델을 만드는 데이터와 평가받는 데이터를 나누자

Why?

Training & Test data set

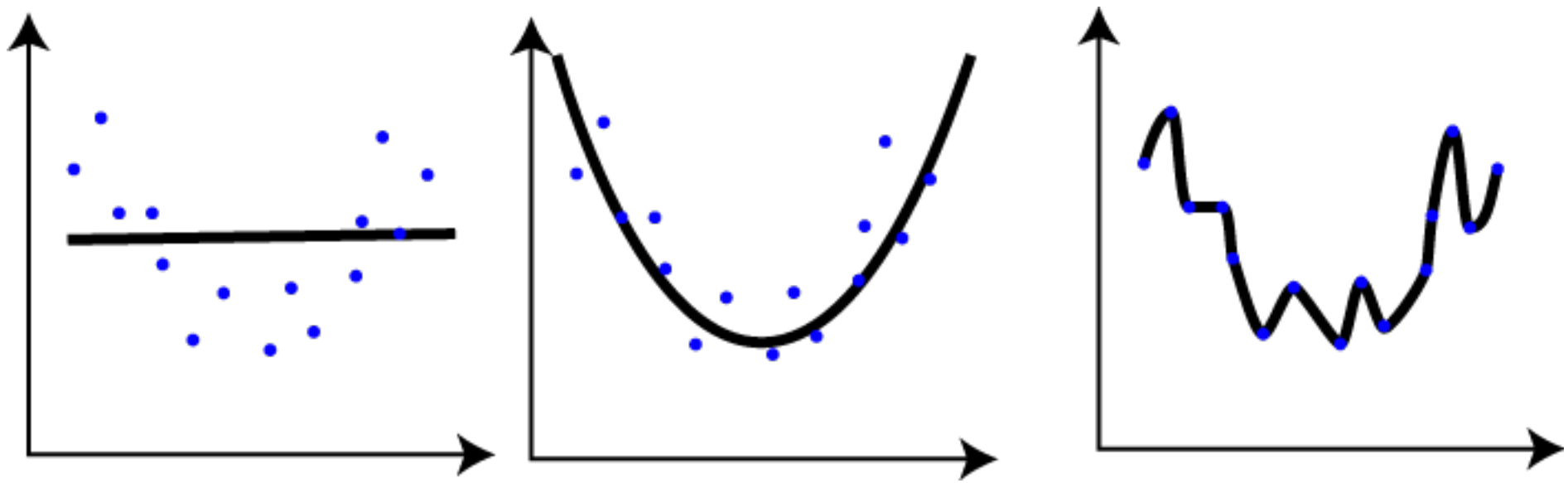
- Training한 데이터로 다시 Test를 할 경우,
  Training 데이터에 과도하게 fitting 된 모델을 사용될 수 있음

- 새로운 데이터가 출현했을 때, 기존 모델과의 차이 존재

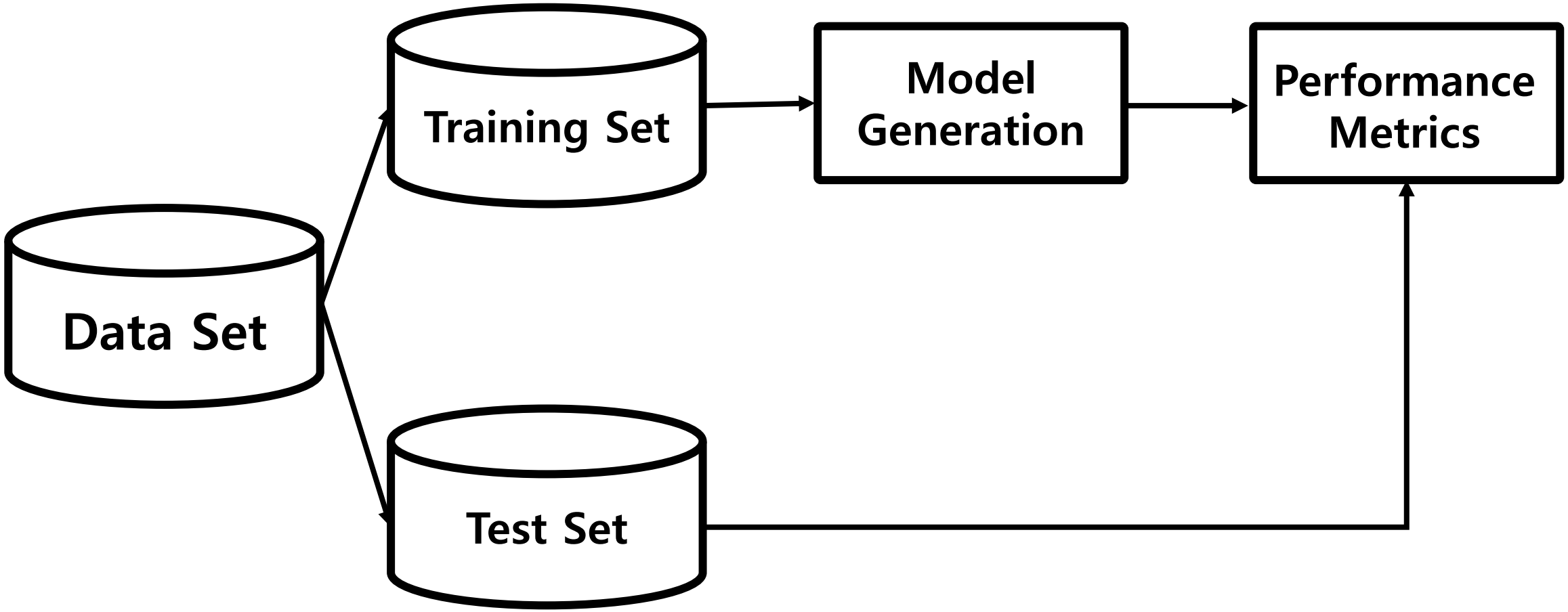
- 모델은 새로운 데이터가 처리가능하도록 generalize되야함

- 이를 위해 Training Set과 Test Set을 분리함

Training & Test data set



General ML Process
Training / Test Set


Data Set
Training Set
Test Set
Model Generation
Performance Metrics

Hold-out Method

Holdout Method (Sampling)

- 데이터를 Training과 Test와 나눠서 모델을 생성하고 테스트하는 기법
- 가장 일반적인 모델 생성을 위한 데이터 램덤 샘플링 기법
- Training과 Test를 나누는 비율은 데이터의 크기에 따라 다름
- 일반적으로 Training Data 2/3 , Test Data 1/3를 활용함

```
import numpy as np
from sklearn.model_selection import train_test_split

X, y = np.arange(10).reshape((5, 2)), range(5)



X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.33, random_state=42)
```

Human knowledge belongs to the world.