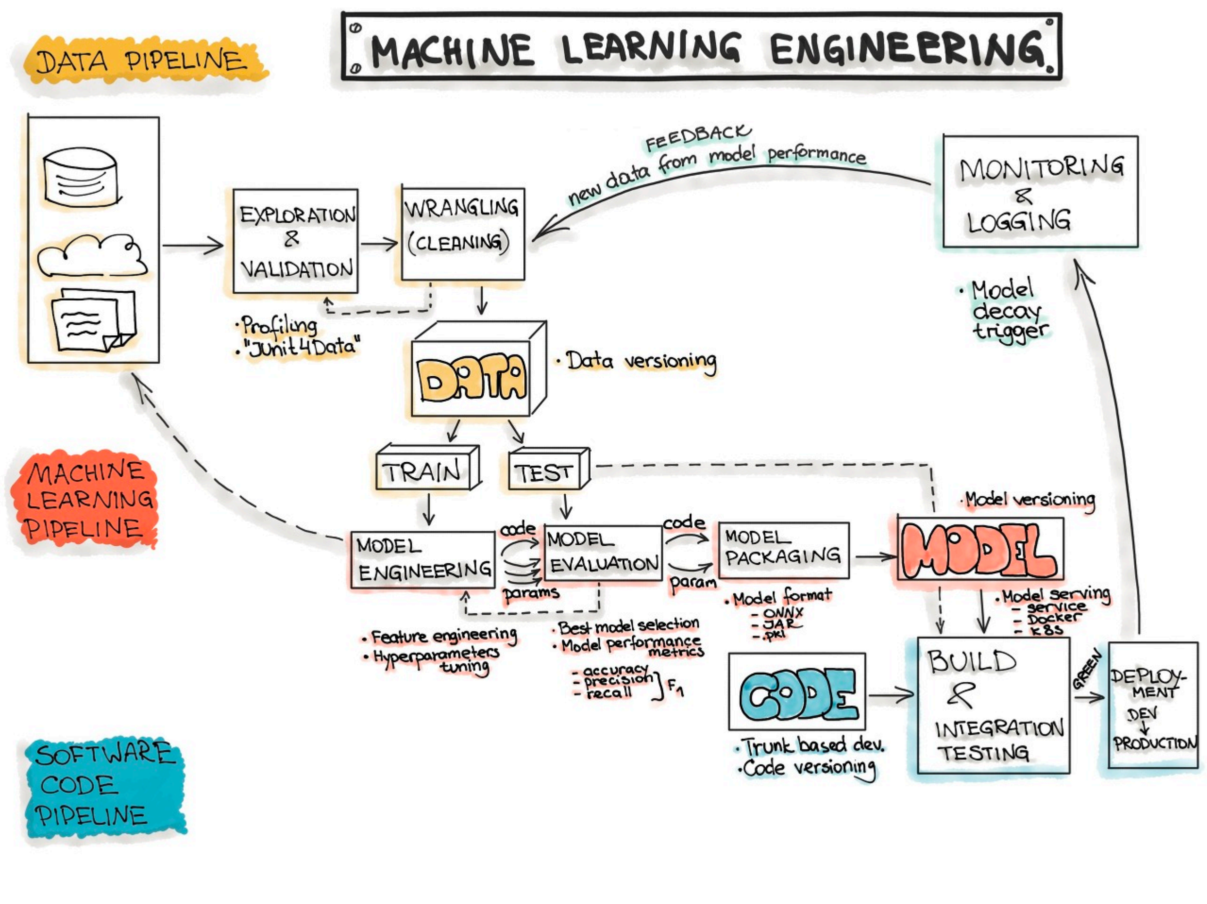
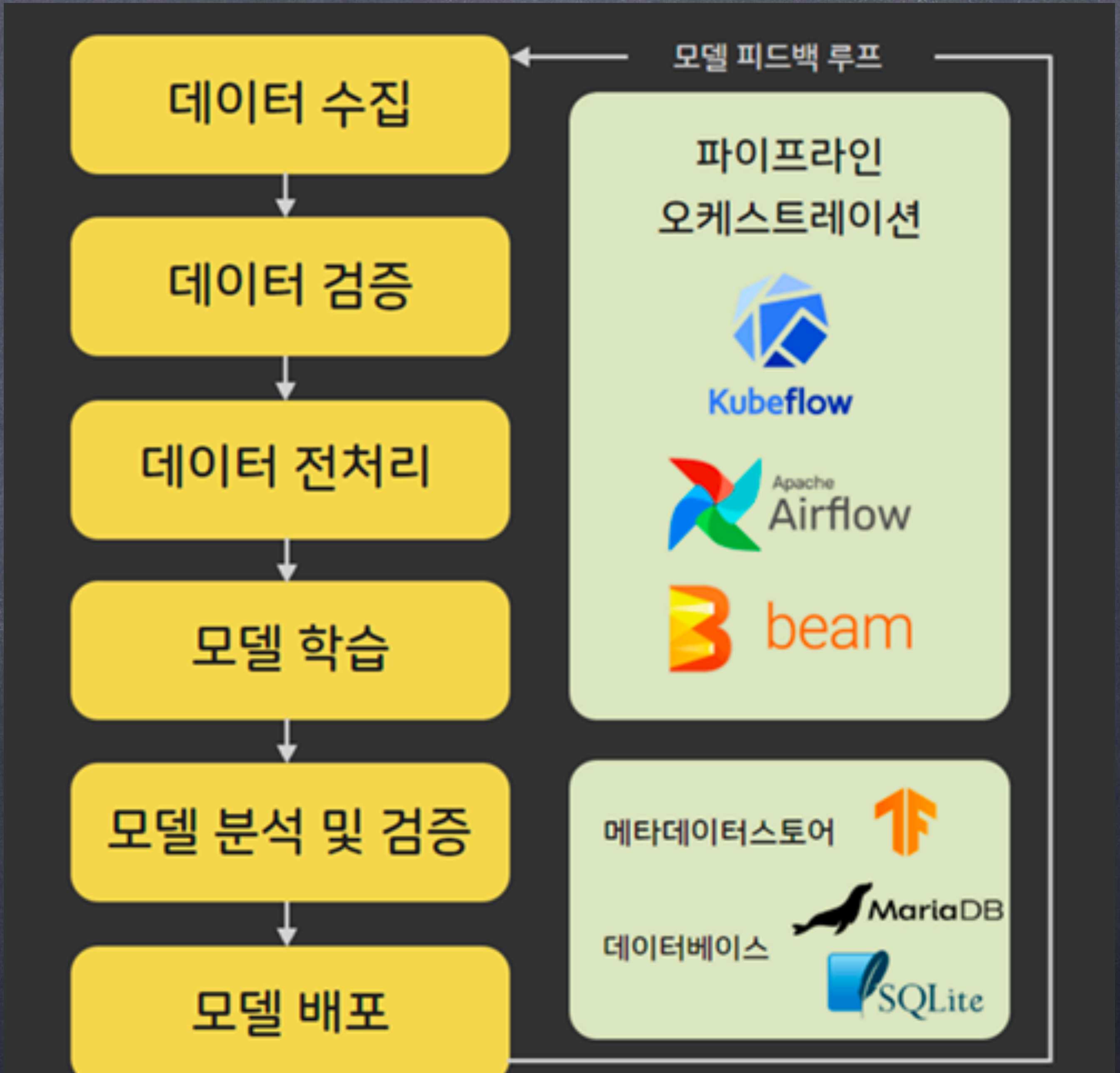
chapter 6
learning best practices for model evaluation and hyperparameter tuning

전설

1. streamlining workflows with pipelines

pipelines





데이터 수집

데이터 검증

데이터 전처리

모델 학습

모델 분석 및 검증

모델 배포

모델 피드백 루프

파이프라인 오케스트레이션

Kubeflow

Apache Airflow

beam

메타데이터스토어

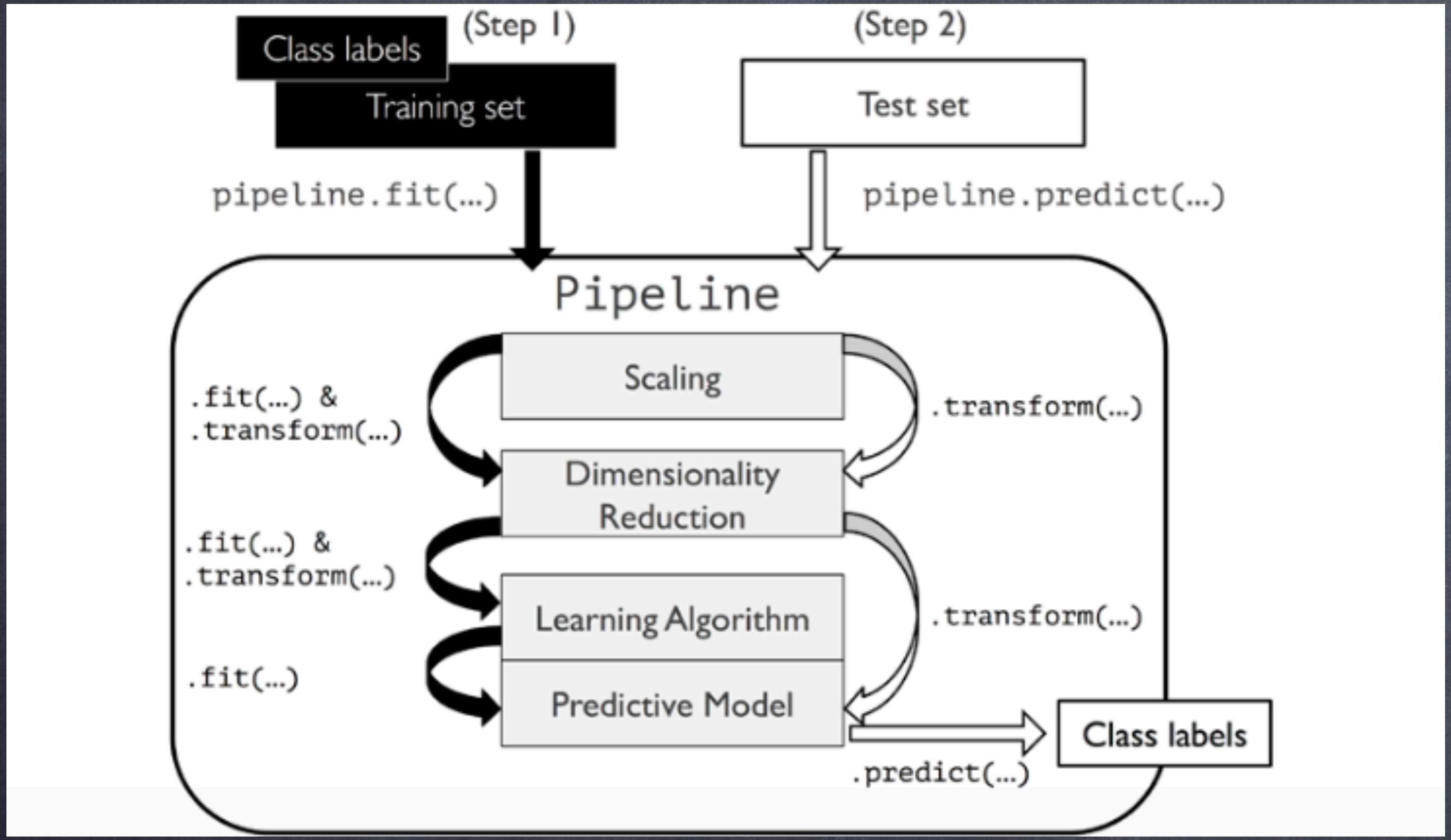
데이터베이스

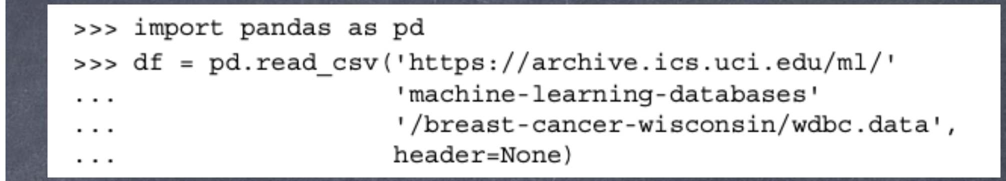
MariaDB

SQLite

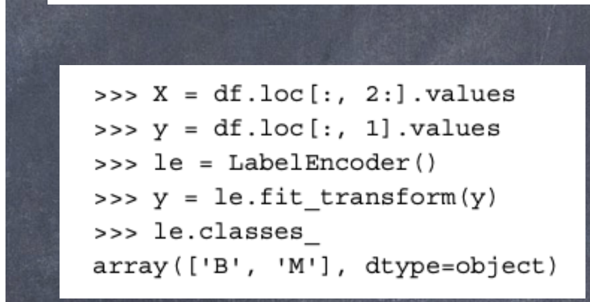
*출처: 한빛미디어, <살아 움직이는 머신러닝 파이프라인 설계>

sklearn.pipeline

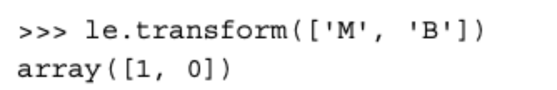




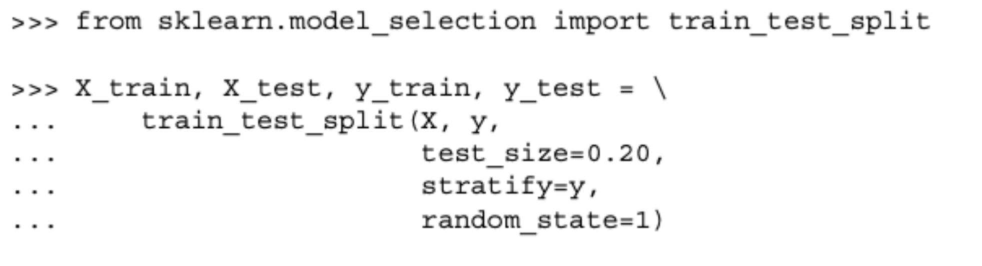
```
>>> import pandas as pd
>>> df = pd.read_csv('https://archive.ics.uci.edu/ml/'
...                  'machine-learning-databases'
...                  '/breast-cancer-wisconsin/wdbc.data',
...                  header=None)
```



```
>>> X = df.loc[:, 2:].values
>>> y = df.loc[:, 1].values
>>> le = LabelEncoder()
>>> y = le.fit_transform(y)
>>> le.classes_
array(['B', 'M'], dtype=object)
```



```
>>> le.transform(['M', 'B'])
array([1, 0])
```

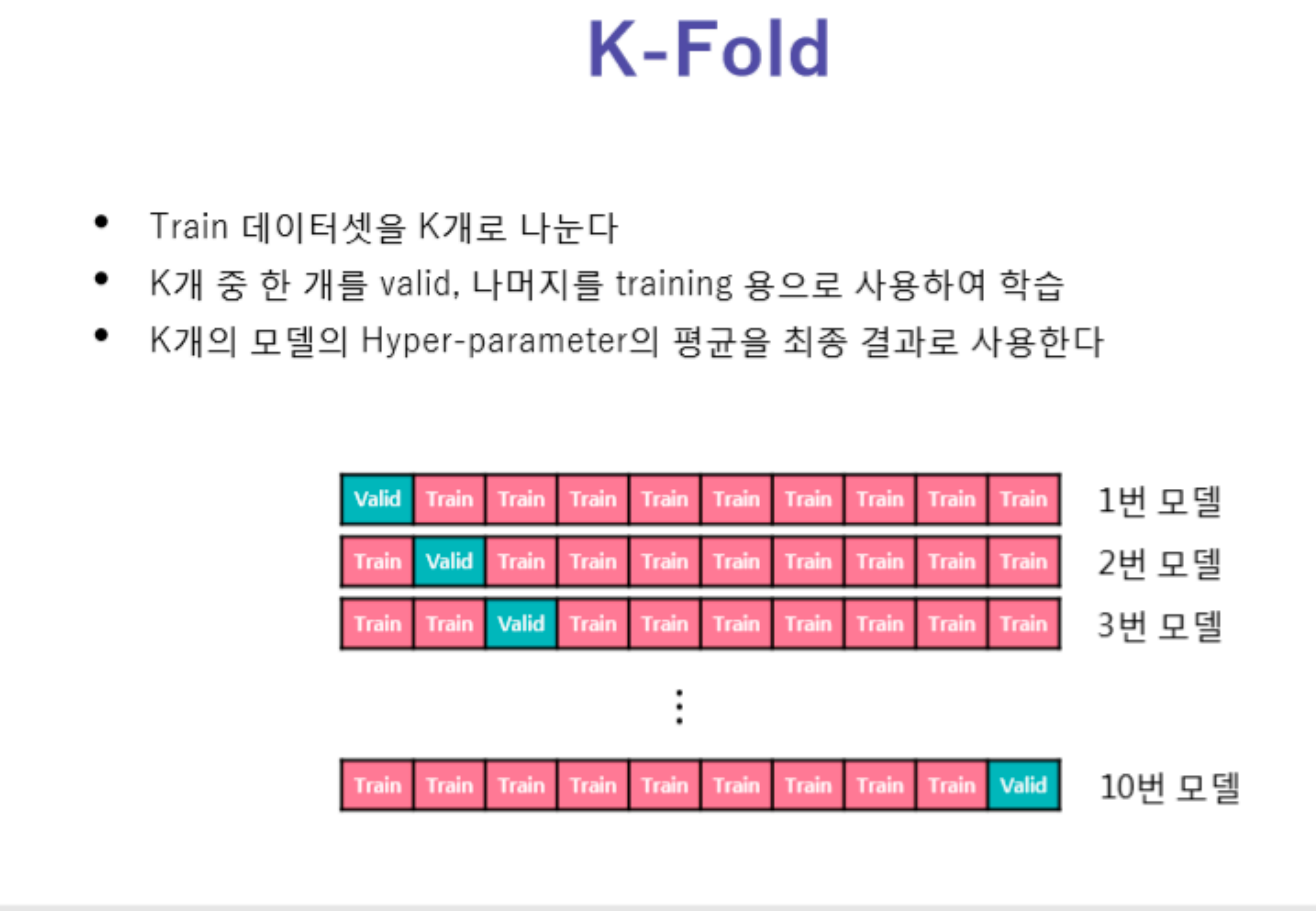


```
>>> from sklearn.model_selection import train_test_split

>>> X_train, X_test, y_train, y_test = \
...     train_test_split(X, y,
...                      test_size=0.20,
...                      stratify=y,
...                      random_state=1)
```

2. Using k-fold cross-validation to assess model performance

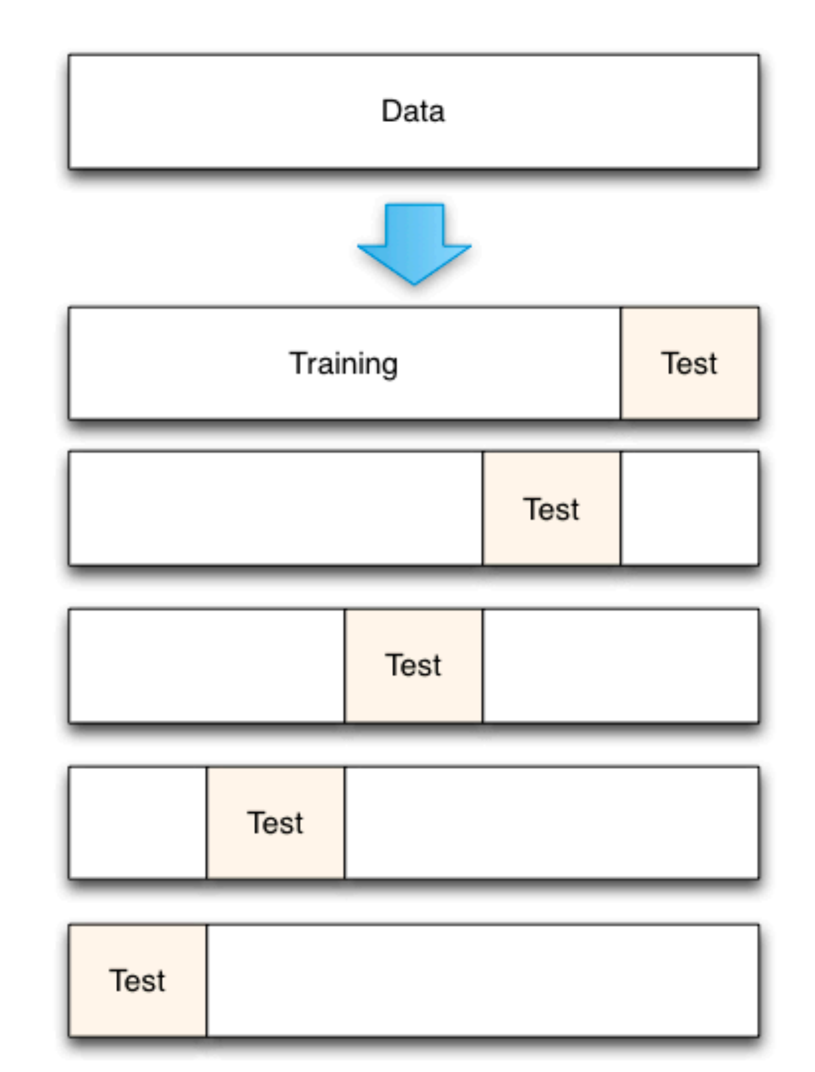
k-fold cross-validation



K-Fold

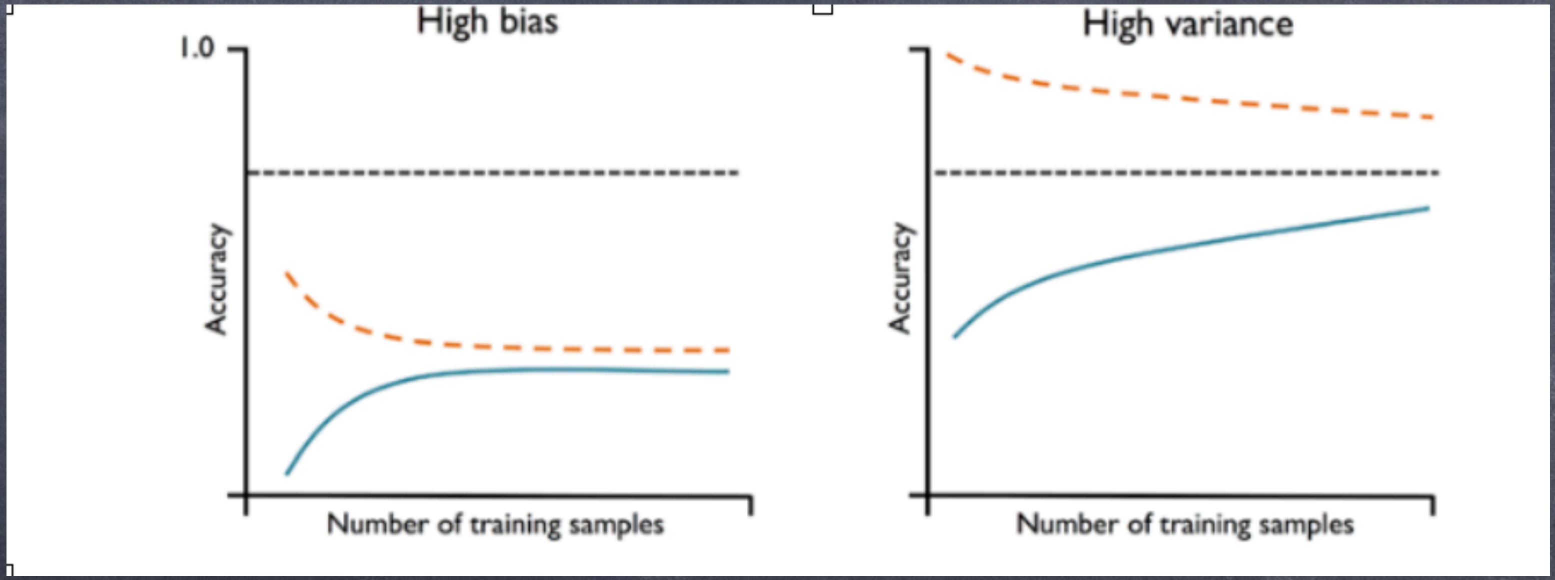
- Train 데이터셋을 K개로 나눈다
- K개 중 한 개를 valid, 나머지를 training 용으로 사용하여 학습
- K개의 모델의 Hyper-parameter의 평균을 최종 결과로 사용한다

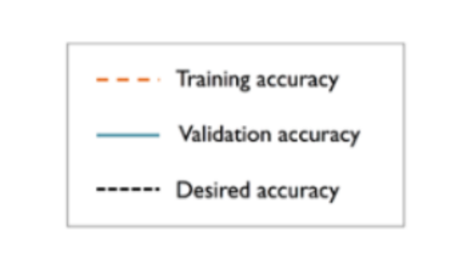
Valid	Train	Train	Train	Train	Train	Train	Train	Train	Train	1번 모델
Train	Valid	Train	Train	Train	Train	Train	Train	Train	Train	2번 모델
Train	Train	Valid	Train	Train	Train	Train	Train	Train	Train	3번 모델
				⋮						
Train	Train	Train	Train	Train	Train	Train	Train	Train	Valid	10번 모델



3. debugging algorithms with learning and validation curves

Diagnosing bias and variance problems with learning curves


High bias
Accuracy
Number of training samples
High variance
Accuracy
Number of training samples


Training accuracy
Validation accuracy
Desired accuracy