Chap01

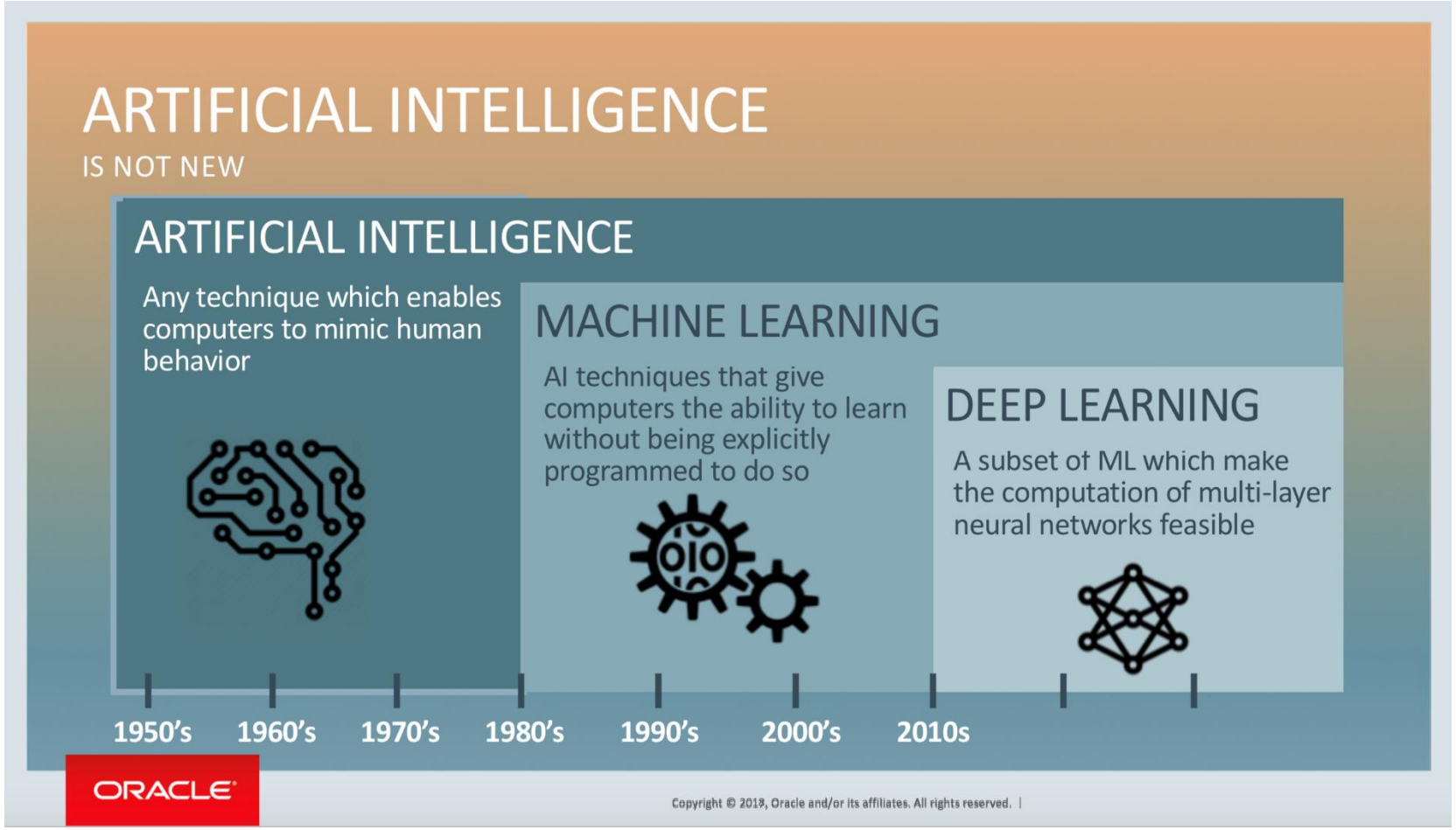
The Machine Learning Landscape

최종현

1.1 머신러닝이란?

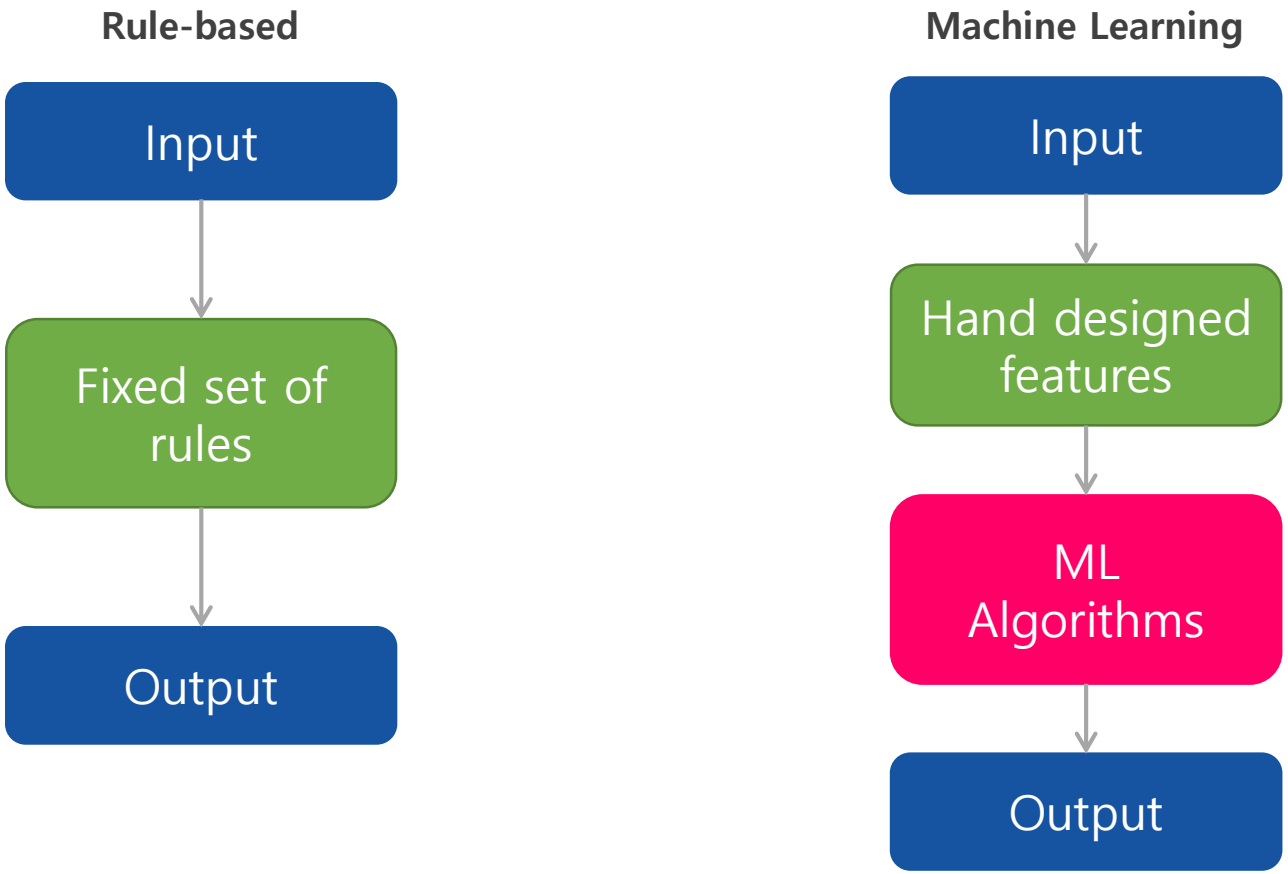
[머신러닝]은 **명시적인 프로그래밍 없이** 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야다.

- 아서 사무엘Athur Samuel , 1959 -



1.2 왜 머신러닝을 사용하는가?

- 머신러닝 모델을 통해 코드를 간단하고 더 잘 수행할 수 있도록 함
- 머신러닝 기법으로 복잡한 문제를 해결할 수 있음
- 새로운 데이터에 대해 유동적으로 적응할 수 있음



1.3 머신러닝 시스템의 종류

- 사람의 감독 하에 훈련하는 것인지
  → *지도, 비지도, 준지도, 강화학습*

- 실시간으로 점진적인 학습을 하는지 아닌지
  → *온라인 학습/배치학습*

- 단순한 데이터 비교인지 패턴을 발견한 모델을 만드는지
  → *사례 기반 학습/모델 기반 학습*

1.3.1 지도 학습과 비지도 학습

- '학습하는 동안의 감독 형태나 정보량'에 따라 머신러닝 종류를 분류

Machine Learning

Supervised Learning
- 학습 데이터에 **레이블(label)** 포함

→ **Classification** → kNN, Logistic Regression, SVM, Decision Tree, ...

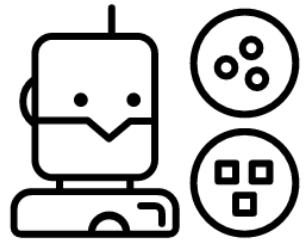
→ **Regression** → Linear Regression, SVR, Neural Networks, ...

Unsupervised Learning
- 학습 데이터에 **레이블(label)** 없음

→ **Clustering** → KMeans, HCA, EM, ...

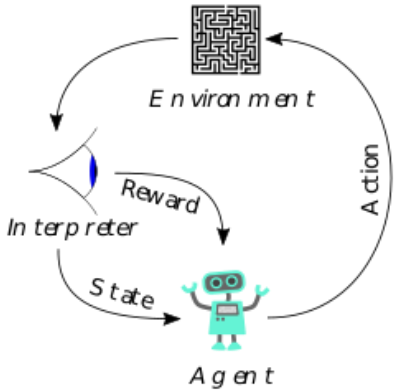
→ **Visualization & Dimensionality reduction** → PCA, Kernel PCA, LLE, t-SNE, ...

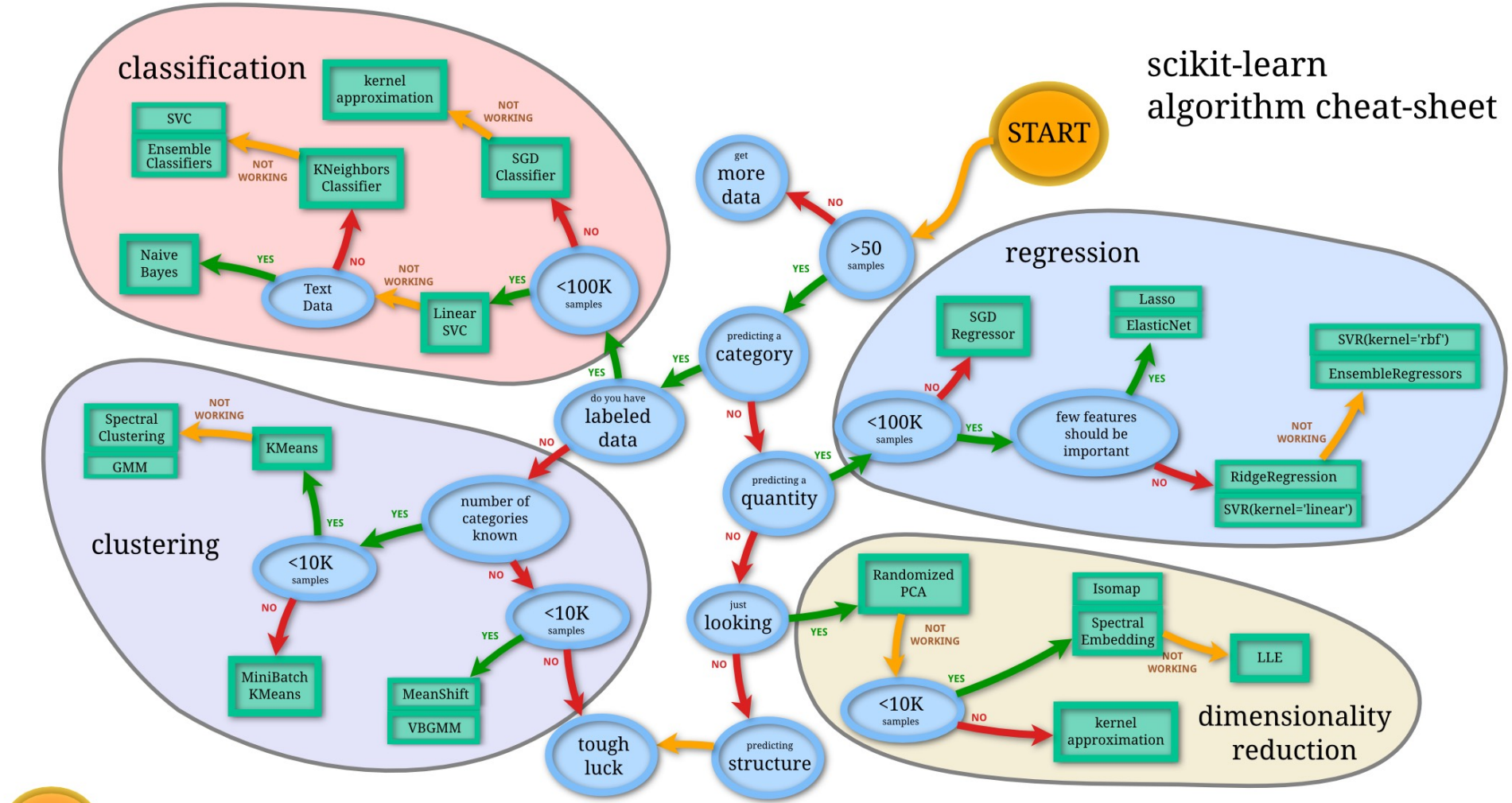
→ **Association Rule learning** → Apriori, Eclat, ...



Semi-Supervised Learning → Supervised & Unsupervised Learning의 조합으로 이루어짐
- 레이블이 있는 데이터 & 없는 데이터

Reinforcement Learning → 에이전트(Agent)가 상태(state)의 보상(reward)를 최대화 하기 위해서 행동함




scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
NOT WORKING
YES
Text Data
Linear SVC
<100K samples
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn

1.3.2 배치 학습과 온라인 학습

- 입력 데이터의 스트림stream 으로 부터 점진적으로 학습할 수 있는지 여부로 분류

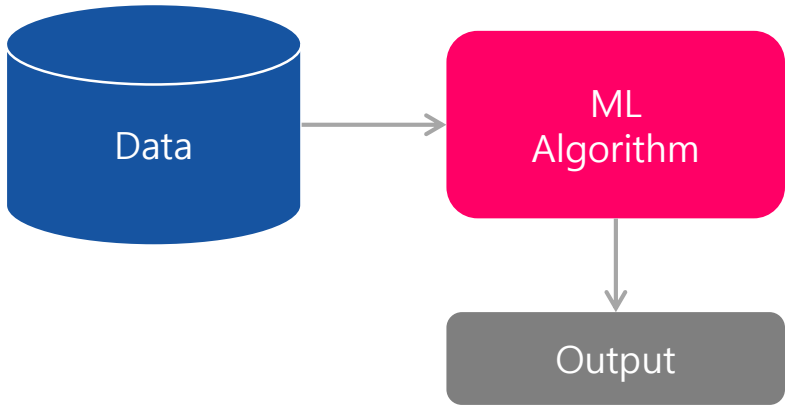
배치 학습 (Batch learning)

- 점진적으로 학습할 수 없으며, 데이터를 모두 사용해 훈련 시킴
- 주로 오프라인에서 수행 → ***offline learning***

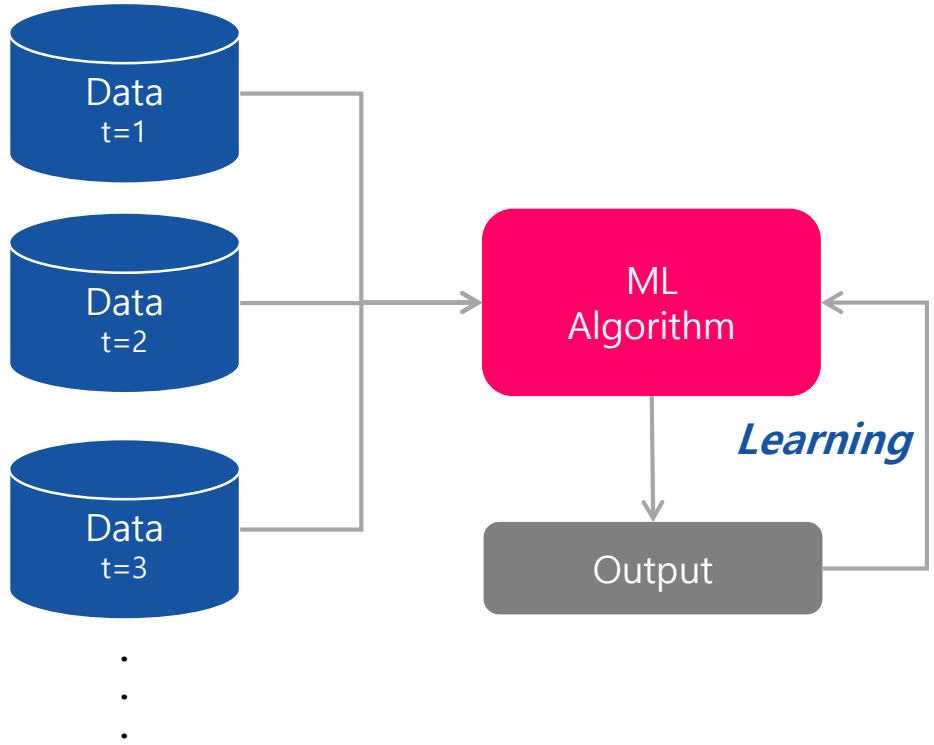
온라인 학습 (Online learning)

- 데이터를 순차적 또는 작은 묶음 단위(미니 배치)로 학습
- 빠른 변화에 적응해야 하는 시스템에 적합(ex. 주식가격)

Batch Learning



On-line Learning



1.3.3 사례 기반 학습과 모델 기반 학습

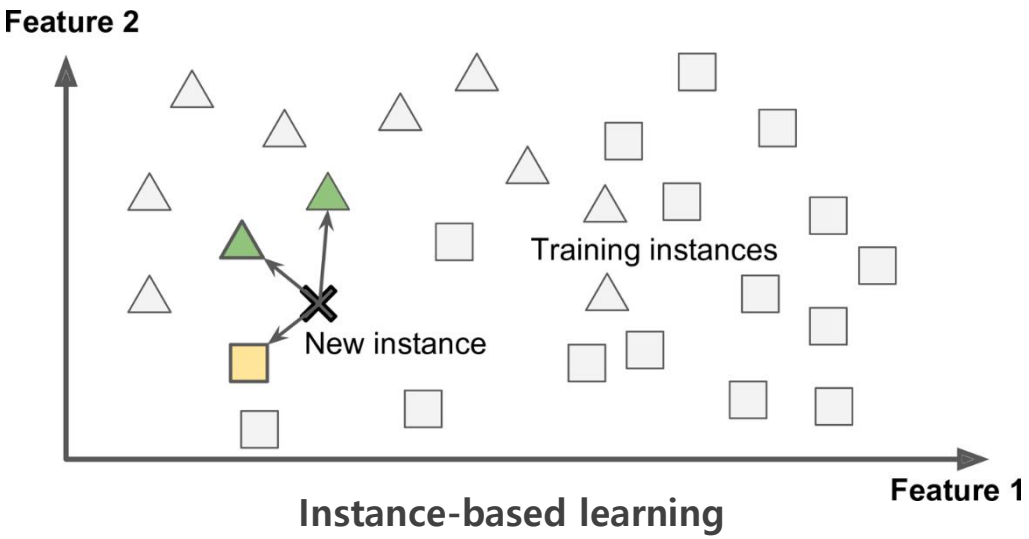
- 어떻게 **일반화** 되는가에 따라 분류
- **일반화**란 학습데이터가 아닌 새로운 데이터를 예측하는데 전반적으로 잘 예측할 수 있도록 하는 것을 말함

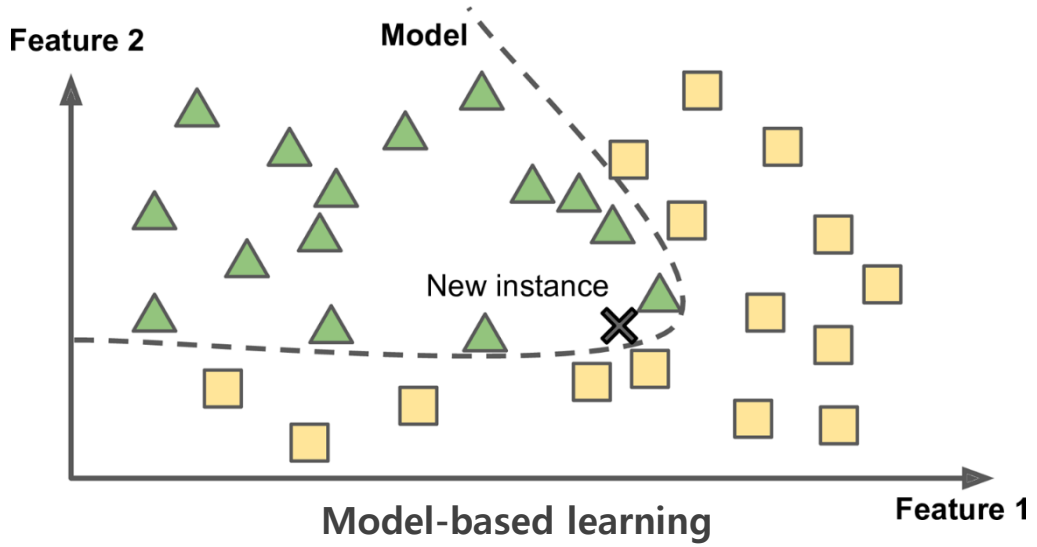
사례 기반 학습 (Instance-based learning)

- 이미 알고 있는 데이터와 새로운 데이터 간의 유사한 정도를 통해 예측하는 방법
- 대표적인 알고리즘으로는 k-NN(k-Nearest Neighbors)가 있음

모델 기반 학습 (Model-based learning)

- 학습 데이터로부터 일반화 할 수 있는 모델을 만들어 **예측**하는 방법





1.3.3 사례 기반 학습과 모델 기반 학습 - 실습

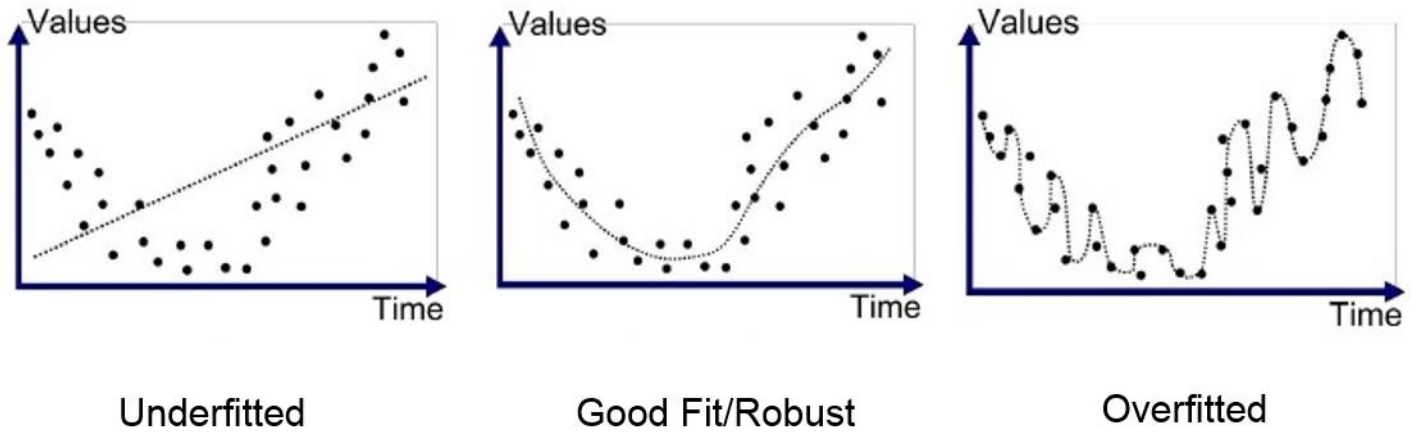
1.4 머신러닝의 주요 도전 과제 - 데이터

- 머신러닝 알고리즘이 잘 작동 하기 위해서는 데이터가 많아야 한다.
- 일반화가 잘되려면 새로운 데이터를 학습 데이터가 잘 대표하는 것이 중요하다.
- 학습 데이터 정제에 많은 시간을 투자해야 한다.
  - 학습 데이터에 에러, 이상치, 노이즈가 많으면 제대로 학습되지 않음
- 학습에 사용할 좋은 특성을 찾는 것이 중요하다. → 특성 공학(feature engineering)
  - **특성 선택(feature selection)**: 가지고 있는 특성 중에서 유용한 특성을 선택
  - **특성 추출(feature extraction)**: 특성을 결합하여 유용한 특성을 만듦

1.4 머신러닝의 주요 도전 과제 - 알고리즘

- 머신러닝 알고리즘이 학습데이터에 **과대적합(overfitting)** 되지 않아야 한다.
  - 파라미터 수가 적은 모델을 선택(고차원 다항 모델 보다 선형모델)
  - 학습 데이터에서 특성 수를 줄이거나, 모델에 제약을 가함
  - 학습 데이터를 더 많이 모은다.
  - 학습 데이터의 노이즈를 줄인다.

- 또한, **과소적합(underfitting)** 되지 않아야 한다.
  - 파라미터가 더 많은 모델을 선택
  - 더 좋은 특성을 추가
  - 모델의 제약을 줄임



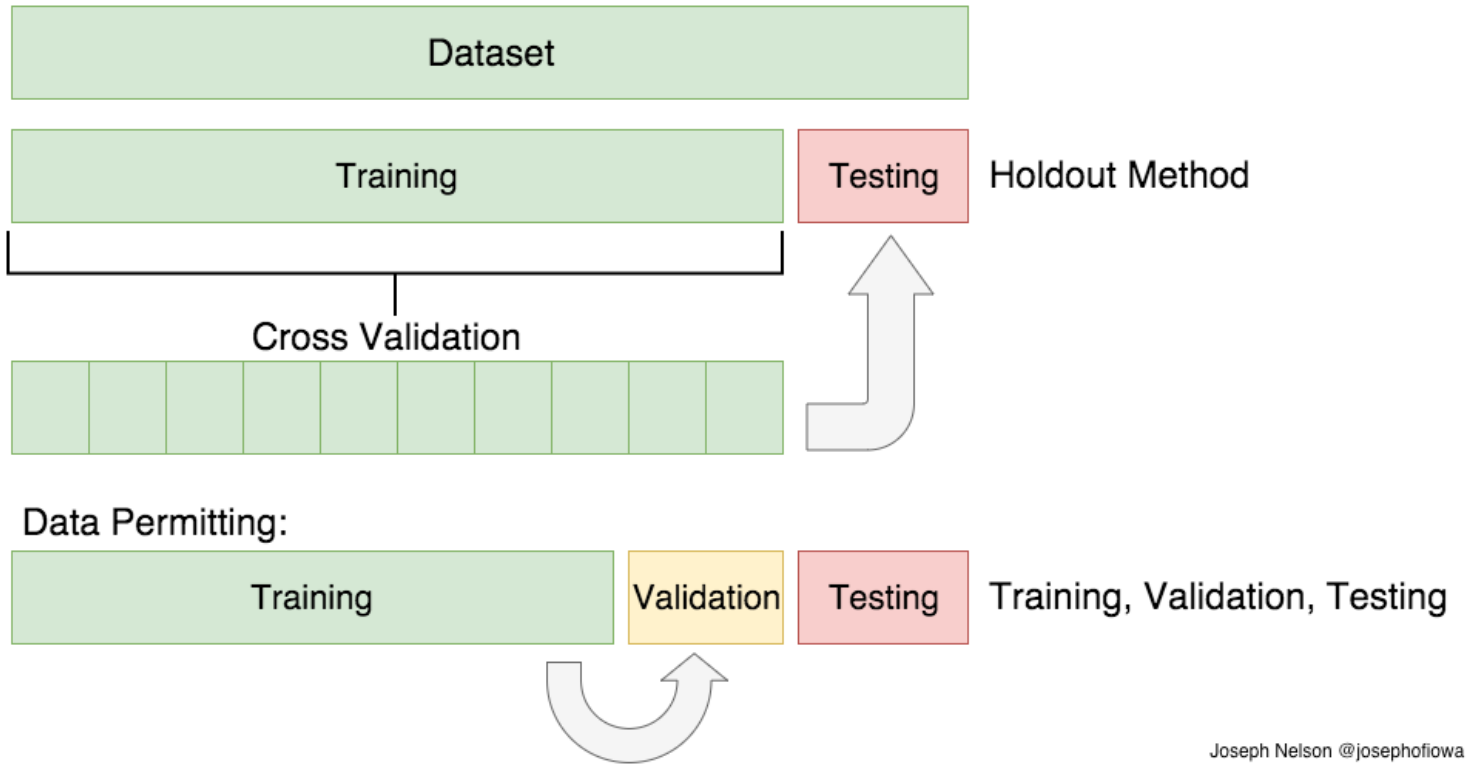
1.5 지금 까지의 정리

- 머신러닝은 명시적인 규칙을 코딩하지 않고 기계가 데이터로부터 학습하여 어떤 작업을 더 잘하도록 만드는 것이다.
- 머신러닝 시스템은 지도/비지도 학습, 배치/온라인 학습, 사례 기반/모델 기반 학습으로 나눌 수 있다.
- 머신러닝은 훈련(학습) 세트에 데이터를 모아 학습 알고리즘에 넣어준다.
  - 모델 기반일 경우, 훈련 세트에 모델을 맞추기 위해 파라미터를 조정한다.
  - 사례 기반일 경우, 학습 데이터를 기억하고 새로운 데이터를 일반화 하기 위해 유사도 측정을 사용한다.

- 훈련 세트가 너무 적거나, 대표성이 없는 데이터, 노이즈가 많은 데이터는 학습이 잘 이루어 지지 않는다.

1.6 테스트와 검증

- **학습 세트(Train Set)**: 머신러닝 모델을 학습할 때 사용하는 데이터 셋
- **검증 세트(Validation Set)**: 모델학습에 필요한 하이퍼파라미터를 찾기위해 사용하는 데이터 셋
- **테스트 세트(Test Set)**: 학습된 모델을 평가하기 위한 데이터 셋



THANK YOU