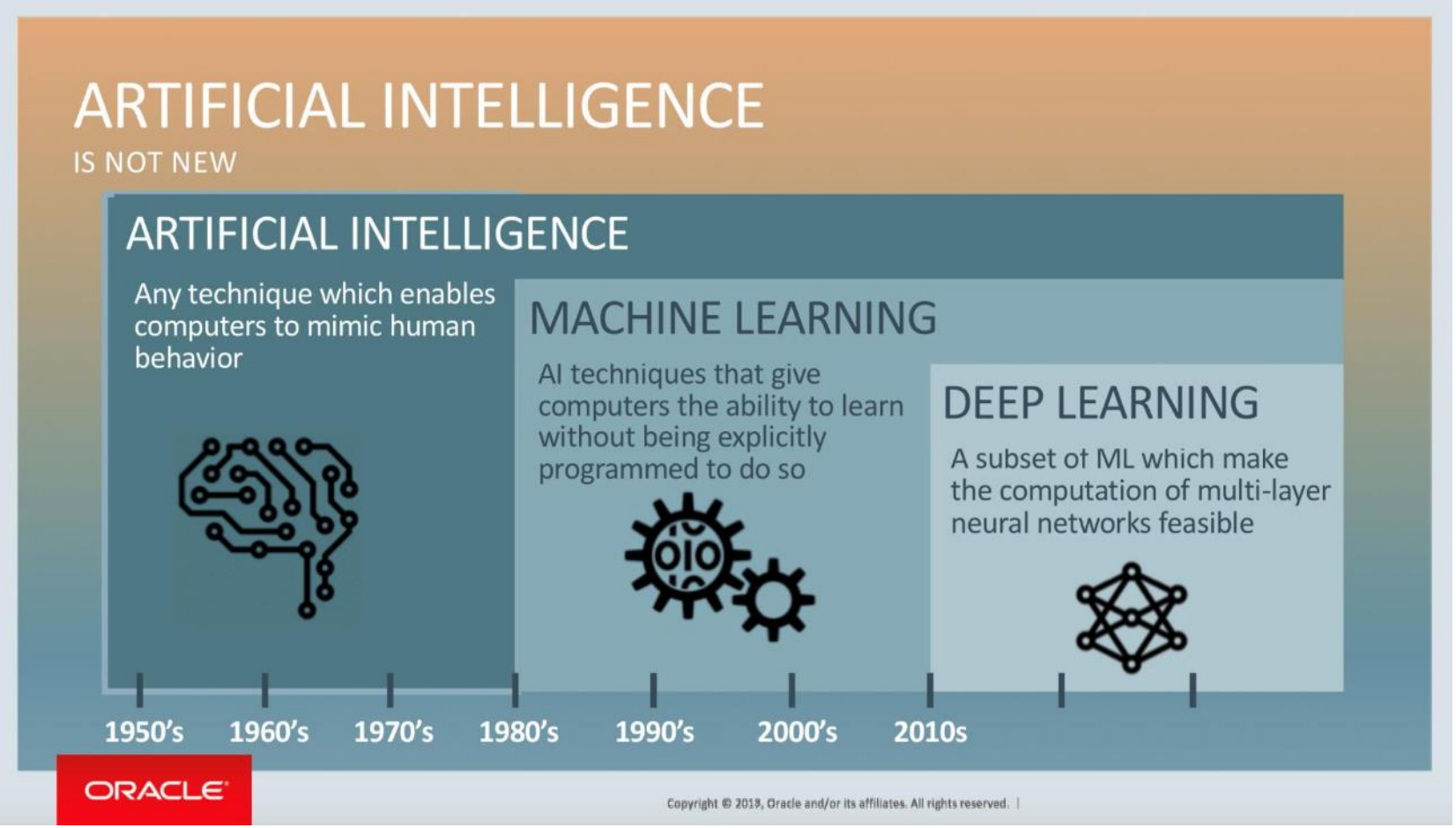
1.1 머신 러닝이란?

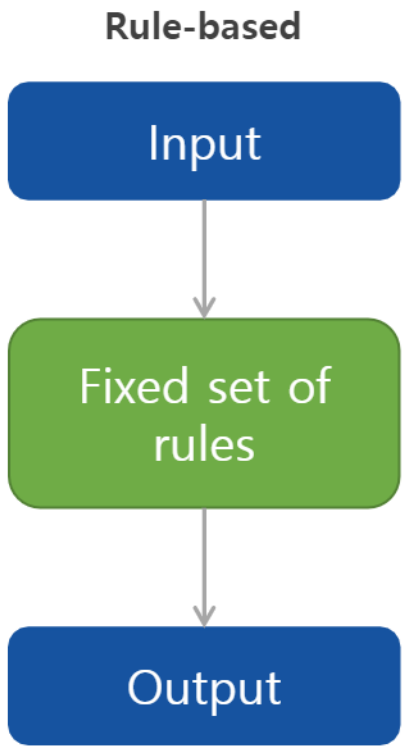
머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야다.

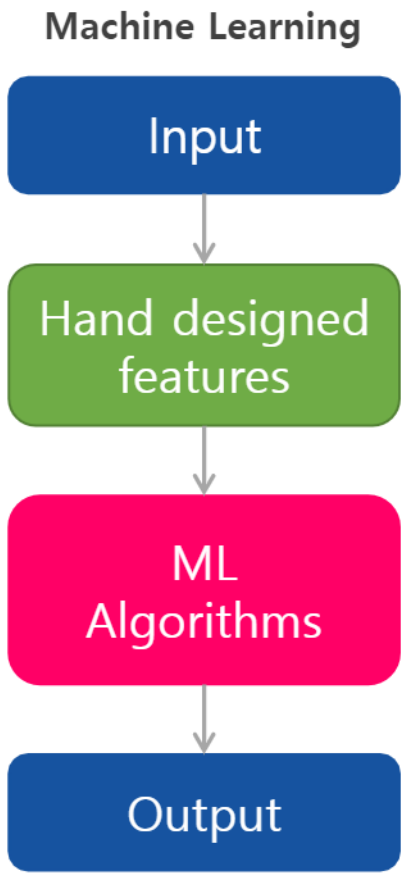
- 아서 새뮤얼(Arthur Samuel, 1959)



1.2 왜 머신러닝을 사용하는가?

- 기존의 방식은 많은 수동 조정과 규칙 필요 → 머신러닝 모델이 코드를 더 단순하게 만듦
- 전통적인 방식으로 해결 방법없는 복잡한 문제 → 머신러닝 기법을 통해 해결 가능
- 유동적인 환경(입력 내용이 계속 바뀜) : 머신러닝 시스템은 새로운 데이터에 적응 가능
- 복잡한 문제와 대량의 데이터에서 통찰 얻기 = 데이터마이닝(Datamining)





1.3 머신러닝 시스템의 종류

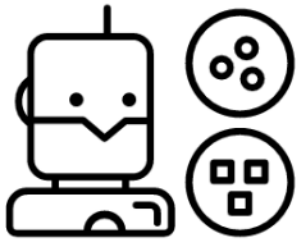
- 사람의 지도 여부 : 비지도 학습 vs 지도학습
- 시리간으로 점진적인 학습을 하는지 아닌지 : 온라인 학습 vs 배치 학습
- 단순한 데이터 비교인지 패턴 발견한 모델인지 : 사례 기반 학습 vs 모델 기반 학습

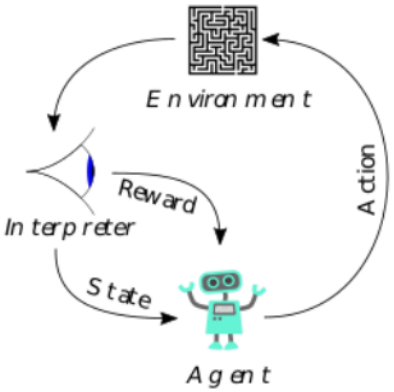
1.3.1 지도학습과 비지도 학습

: 학습하는 동안의 감독 형태나 정보량에 따른 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



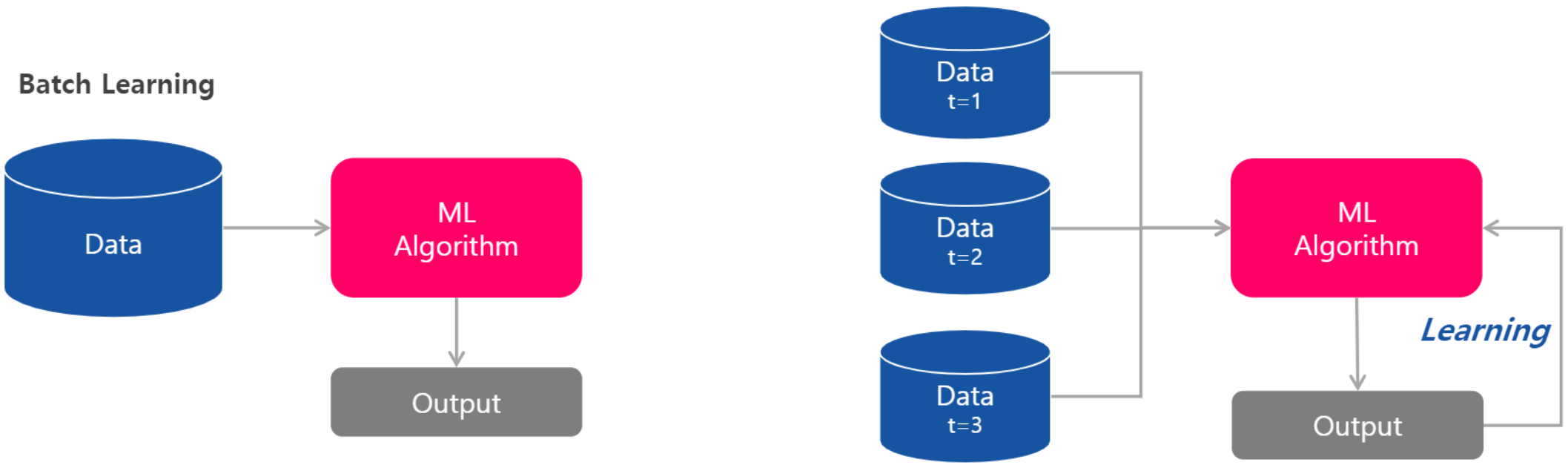


1.3.2 배치 학습과 온라인 학습

: 입력 데이터의 스트림(stream)부터 점진적으로 학습할 수 있는가 여부

(1) 배치 학습(Bath learning)
- 점진적 학습 불가능, 가용한 데이터를 모두 사용해서 훈련 → 시간과 자원 소모 多
- 먼저 시스템을 훈련 → 제품 시스템에 적용 = 더 이상의 학습 X → Offline learning

(2) 온라인 학습(Online learning)
- 순차적으로 데이터 사용하여 훈련, 데이터 도착 즉시 학습 → 빠르고 자원 소모 적음
- 연속적으로 데이터를 받거나 빠른 변화에 스스로 적응해야 하는 시스템에 적합(like 주식)
- 컴퓨팅 자원이 제한되었을 때도 좋음

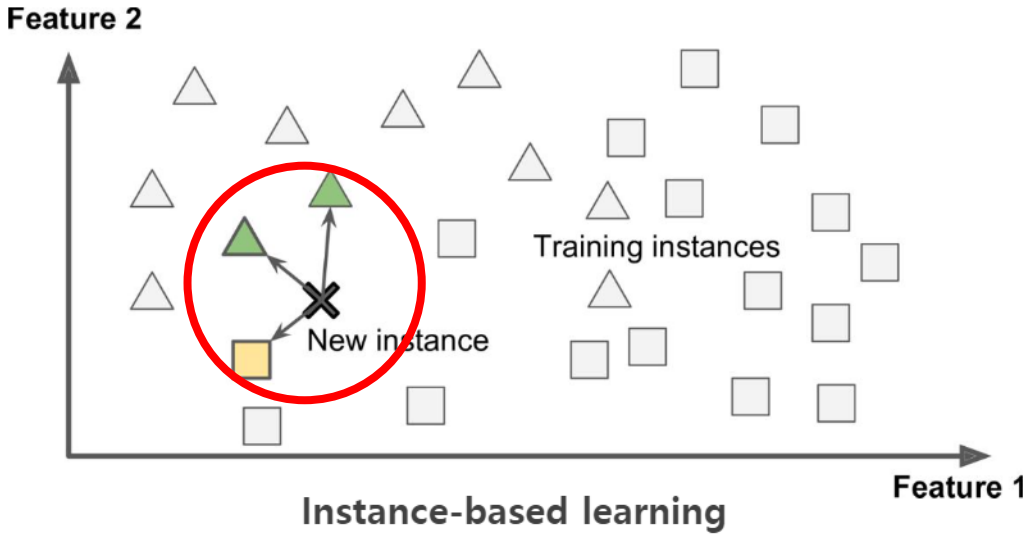


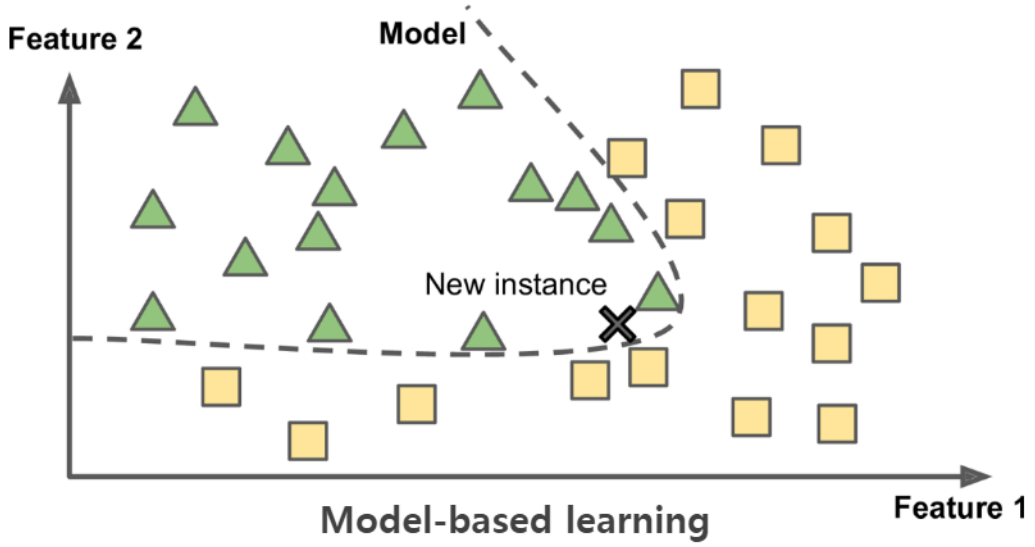
1.3.3 사례 기반 학습과 모델 기반 학습

: 어떻게 일반화(generalize)되는가에 따른 분류

(1) 사례 기반 학습(Instance-based learning)
- 훈련 샘플 기억 → 유사도 측정을 사용하여 새로운 데이터와 학습한 샘플 간의 비교를 통해 일반화

(2) 모델 기반 학습(Model-based learning)
- 샘플로 일반화 시키는 방법 → 샘플로 모델을 만들어 예측하는 방식





1.5 머신러닝의 주요 도전 과제

- 데이터

(1) 데이터가 충분히 있어야 한다 : 좋은 모델만큼 데이터가 중요

(2) 훈련 데이터가 데이터를 잘 대표하기 있어야 한다

(3) 데이터의 에러, 이상치(Outlier), 잡음을 잘 정제해야 한다 → 가장 많은 시간을 쓰는 파트

(4) 훈련에 사용할 가장 관계 높은 특성을 찾아야한다 → 특성공학

- 특성 선택(Feature Selection) : 가지고 있는 특성 중에서 훈련에 가장 유용한 특성을 선택
- 특성 추출(Feature Extraction) : 특성을 결합하여 더 유용한 특성을 만든다
- 새로운 데이터 수집을 통해 새 특성 만듦

1.5 머신러닝의 주요 도전 과제

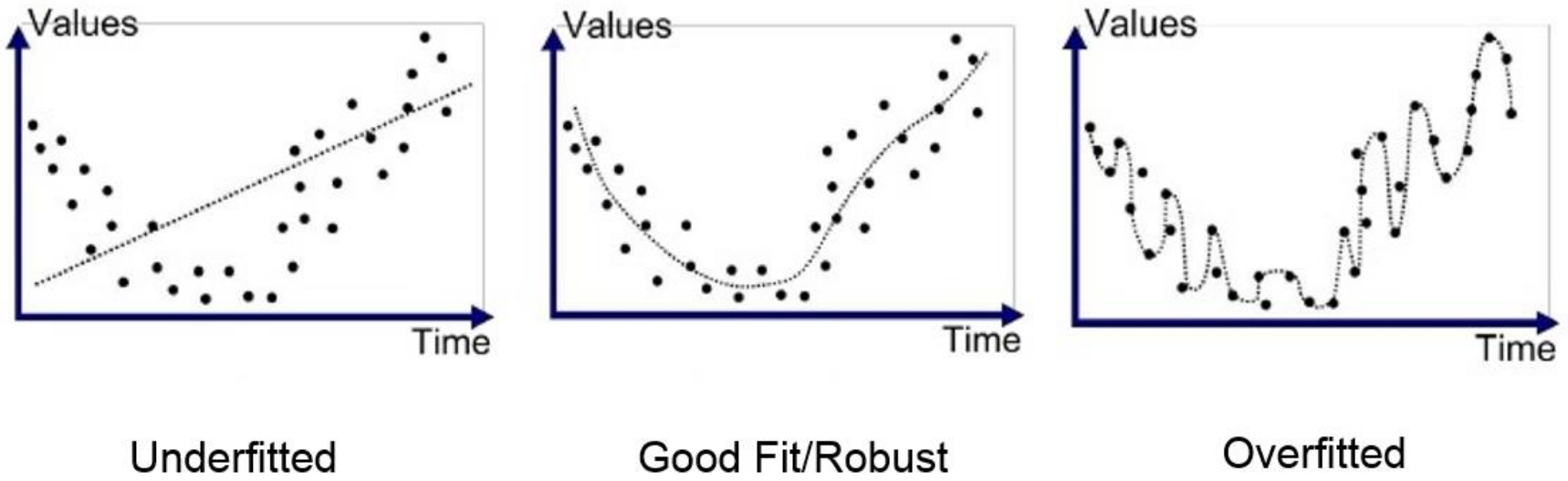
- 알고리즘

(1) Overfiting 경계 : 모델이 학습데이터에 지나치게 과대 적합

- 파라미터 수를 줄임
- 학습 데이터의 특성 줄임
- 모델에 제약 추가
- 학습 데이터 추가
- 학습 데이터의 노이즈 제거

(2) Underfiting 주의

- 파타미터의 수 추가
- 더 좋은 데이터의 특성 추가
- 모델의 제약 감소



1.6 테스트와 검증

- 학습 세트(Train set) : 머신러닝 모델 학습을 위해 사용하느 데이터
- 검증 세트(Validation set) : 모델학습에 필요한 하이퍼파라미터를 찾기 위해 사용하는 데이터 셋
- 데스트 세트(Test set) : 학습된 모델을 평하기 위한 데이터 셋