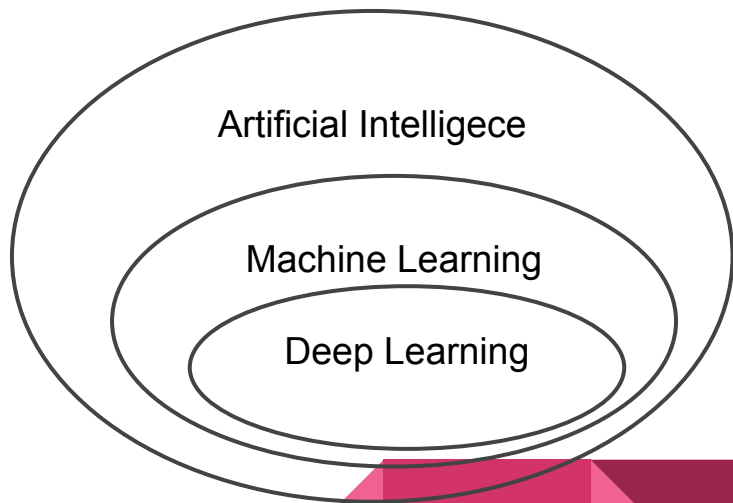


# 1. 한눈에 보는 머신 러닝

Hands-On Machine Learning

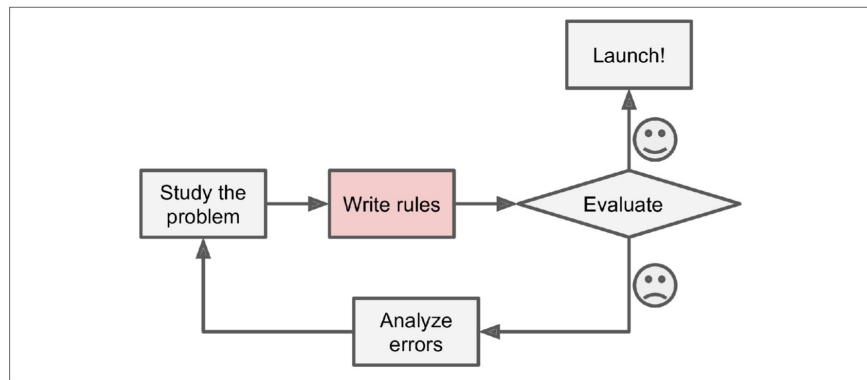
# 1. What Machine Learning

- 머신러닝은 데이터로부터 학습하도록 컴퓨터를 프로그래밍하는 것.
- 머신 러닝 성공 사례
  - 이미지 분석, 자동 분류
  - 뇌종양 진단
  - 신문 기사 자동 분류
  - 긴 문서 자동 요약
  - 챗봇, 개인 비서
  - 수익 예측
  - 음성 인식
  - 고객 분류
  - 상품 추천
  - 게임



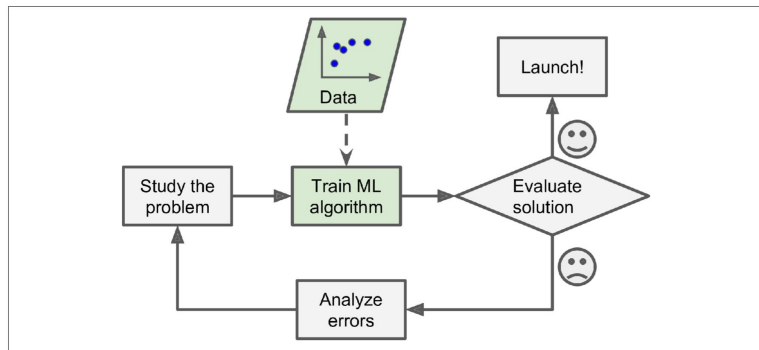
## 2. Why Machine Learning

- 스팸 필터
  - 스팸에 어떤 단어들이 주로 나타나는지 조사
  - 발견한 각 패턴을 감지하는 알고리즘을 작성하여 프로그램이 이런 패턴을 발견하는 스팸으로 분류하게 함
  - 프로그램을 테스트하고 충분한 성능이 나올 때까지 1단계와 2단계를 반복

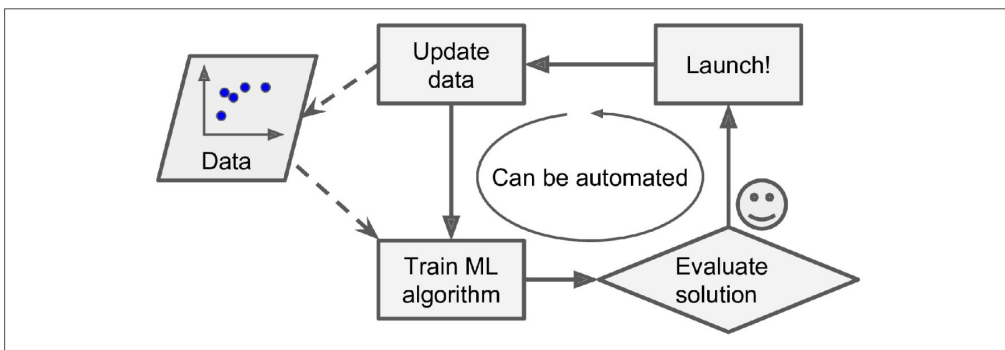


## 2. Why Machine Learning

- 전통적인 프로그래밍 방법의 한계 VS 머신 러닝
  - 규칙이 점점 길고 복잡해지므로 유지 보수하기 매우 힘들다.
  - → 스팸 메일을 판단하는 기준을 데이터에서 스스로 학습
  - 스팸 발송자가 스팸 필터에 계속 단어를 바꾸면, 새로운 규칙을 추가해야 한다.
  - → 사람의 별도의 작업 없이 새로운 데이터에서 새로운 규칙을 스스로 추가



머신러닝 접근 방법

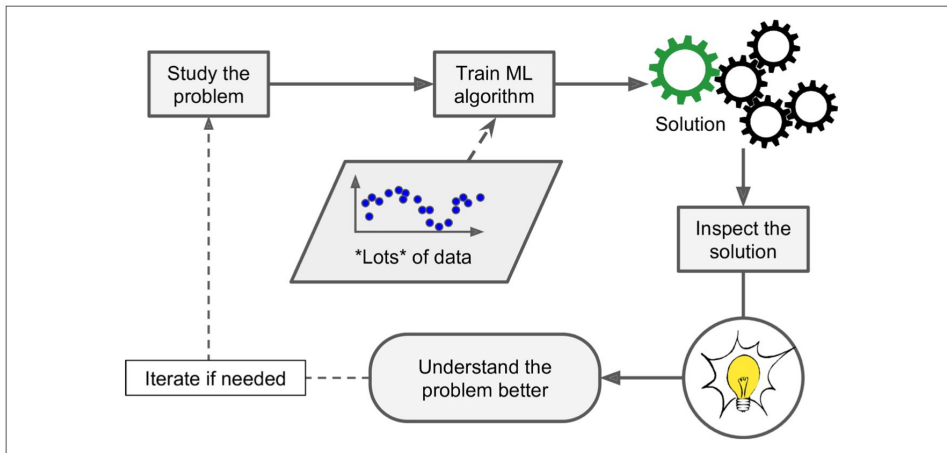


자동으로 변화에 적응

## 2. Why Machine Learning

- 데이터 마이닝(Data Mining)

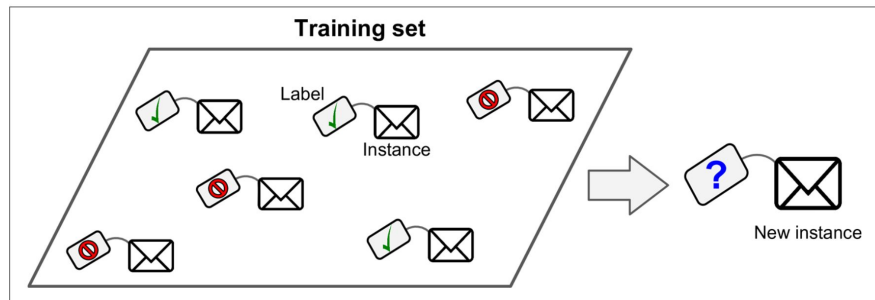
- 사람이 머신러닝 알고리즘이 학습한 것을 조사하다가 예상치 못한 연관 관계, 새로운 추세가 발견되기도 해서 해당 문제를 더 잘 이해하도록 도와준다.
- 머신러닝 기술을 적용해서 대용량의 데이터를 분석하면 겉으로 보이지 않던 패턴을 발견할 수도 있다.



### 3. 머신 러닝 종류

- 사람의 감독 하에 훈련하는 것인지 아닌지
  - 지도(Supervised) 학습
  - 비지도(Unsupervised) 학습
  - 준지도(Semi-supervised) 학습
  - 강화(Reinforcement) 학습
- 실시간으로 점진적인 학습을 하는지 아닌지
  - 온라인(Online) 학습
  - 배치(Batch) 학습
- 단순히 알고 있는 데이터 포인트와 새 데이터 포인트를 비교하는 것인지 아니면 훈련데이터 셋에서 과학자들처럼 패턴을 발견하여 예측 모델을 만드는지
  - 사례 기반(Instance-based) 학습
  - 모델 기반(Model-based) 학습

## 3-1. 지도 학습(Supervised Learning)

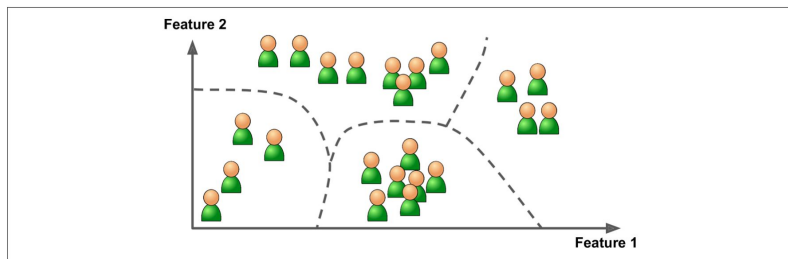


분류(Classification)

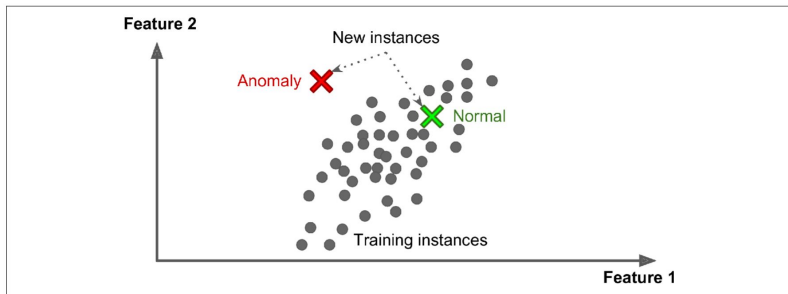


회귀(Regression)

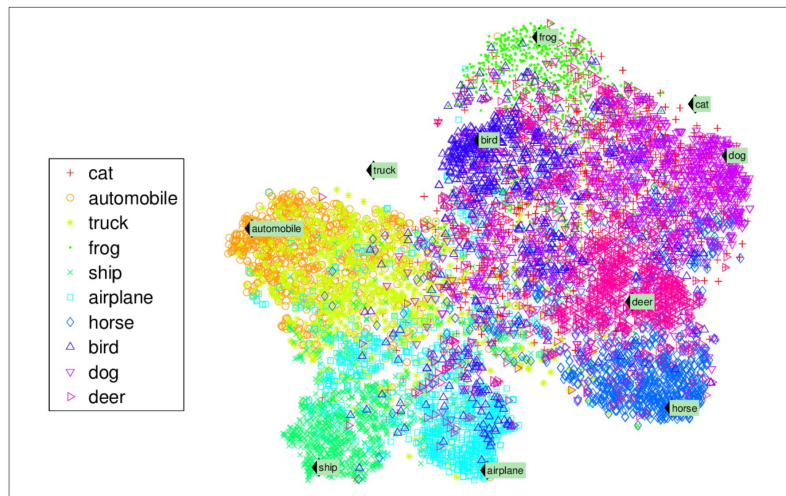
# 3-1. 비지도 학습(Unsupervised Learning)



군집(Clustering)



이상치 찾기(Anomaly detection)

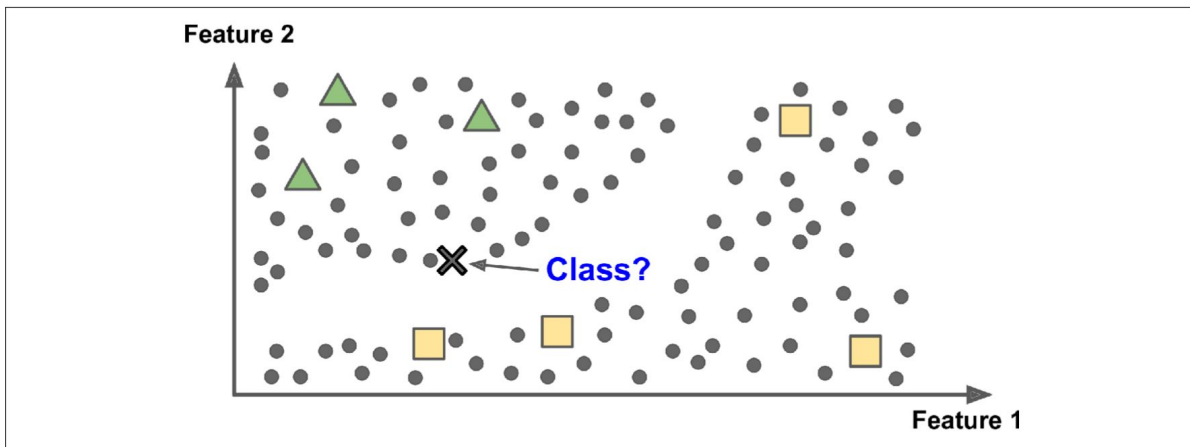


시각화(t-SNE visualization)



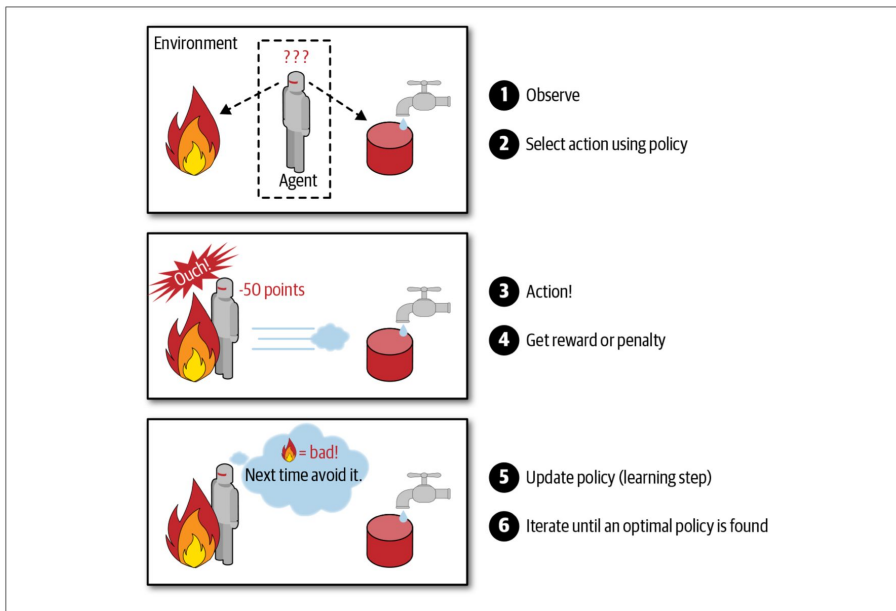
## 3-1. 준지도 학습(Semi-supervised Learning)

- Google Photos

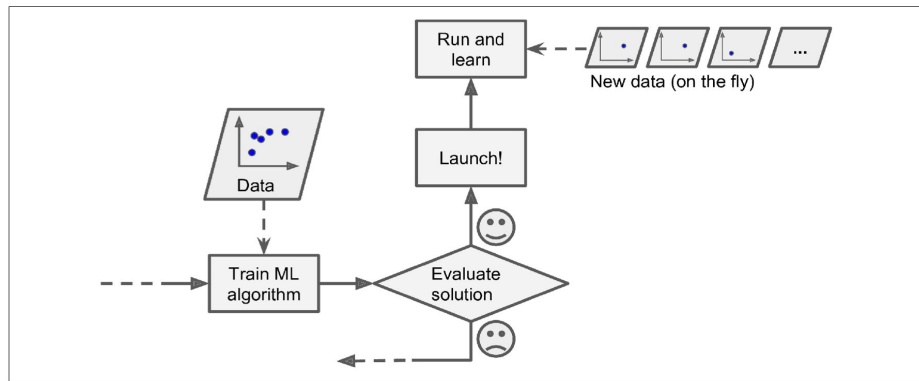


# 3-1. 강화 학습(Reinforcement Learning)

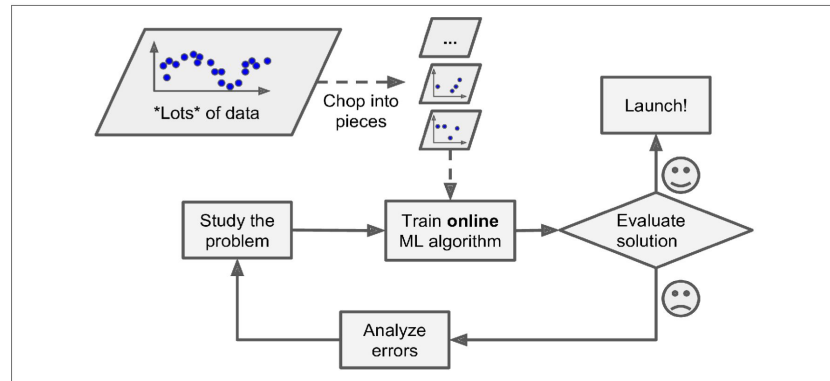
- AlphaGo, AlphaGo Zero



## 3-2. Batch vs Online Learning

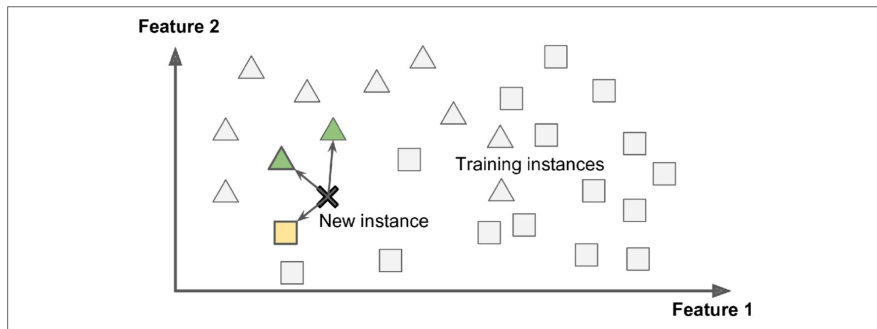


온라인 학습

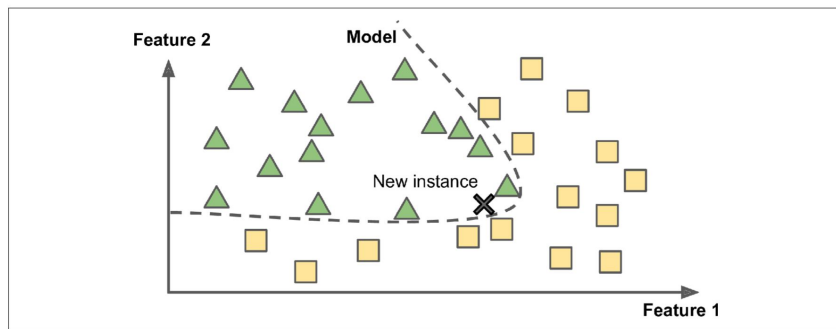


온라인 학습을 사용한 대용량 데이터의 점진적 처리

### 3-3. Instance-based, Model-based Learning

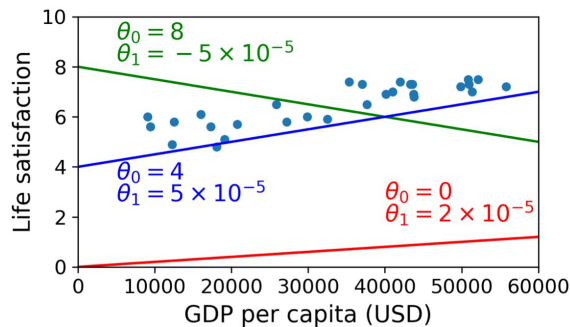
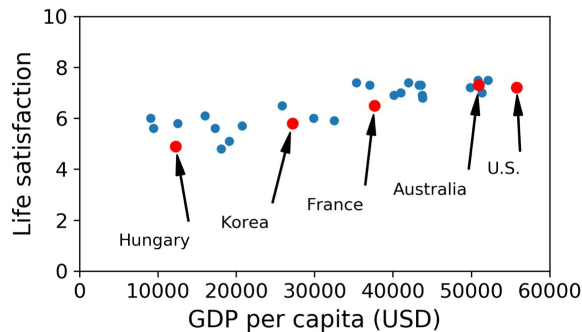


사례 기반 학습



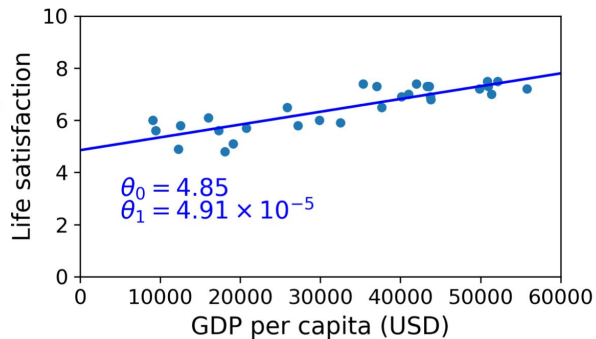
모델 기반 학습

# 3-3. Model-based Learning




Simple Linear Model:

$$\text{life\_satisfaction} = \theta_0 + \theta_1 \times \text{GDP\_per\_capita}$$



## 4. 나쁜 데이터, 나쁜 알고리즘

- 충분하지 않은 양의 훈련 데이터
  - 대표성이 없는 훈련 데이터
    - 샘플링 잡음(sampling noise), 샘플링 편향(sampling bias)
  - 낮은 품질의 데이터
    - 결측치(missing values, NA), 이상치(outliers)
  - 관련 없는 특성
    - Gargage in, garbage out
  - 훈련 데이터 과대적합(overfitting)
  - 훈련 데이터 과소적합(underfitting)
- 

## 5. Testing and Validating

- Training set
  - 다양한 하이퍼 파라미터들을 갖는 여러개의 모델을 학습시킴.
- Validation set
  - validation set에서 최상의 성능을 내는 모델을 선택.
- Test set
  - test set에서 모델을 평가하고 일반화 에러를 추정.

