머신러닝의 개요

머신러닝이란?

머신러닝의 개요

머신러닝이란?

인공지능

머신러닝

딥러닝

빅데이터

머신러닝이란?

인공지능

Artificial intelligence can be classified into three different types of systems:
analytical, human-inspired, and humanized artificial intelligence.

analytical: generating a cognitive representation of the world and using learning based on past experience to inform future decisions.

human-inspired: understanding human emotions, in addition to cognitive elements, and considering them in their decision making.

humanized artificial intelligence: AI shows characteristics of all types of competencies (i.e., cognitive, emotional, and social intelligence), is able to be self-conscious and is self-aware in interactions with others.

출처: Wikipedia(https://en.wikipedia.org/wiki/Artificial_intelligence)

머신러닝이란?

머신러닝

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to **effectively perform a specific task without using explicit instructions**, relying on patterns and inference instead.

It is seen as a **subset of artificial intelligence.** Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

출처: Wikipedia(https://en.wikipedia.org/wiki/Machine_learning)

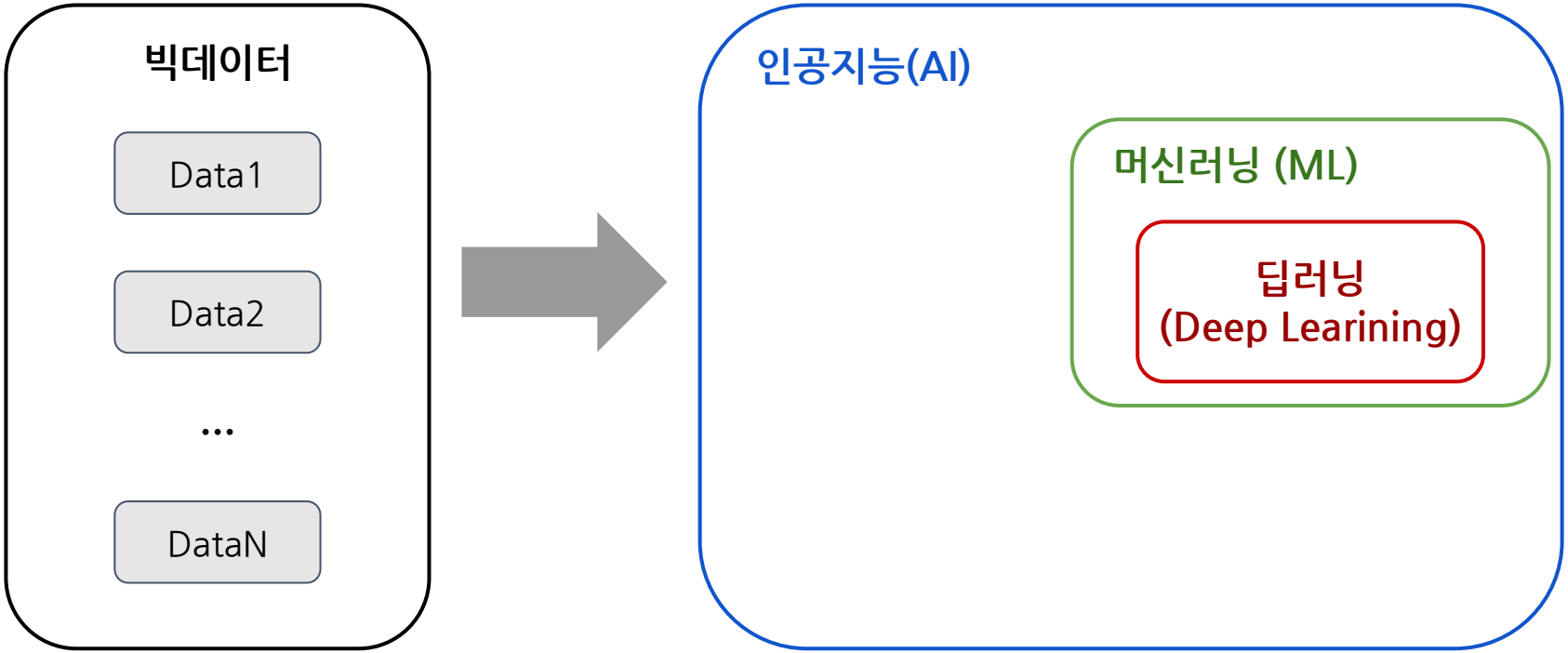
머신러닝이란?

딥러닝

Deep learning (deep structured learning or hierarchical learning)
is **part of a broader family of machine learning methods** based on the layers used in artificial **neural networks.**

출처: Wikipedia(https://en.wikipedia.org/wiki/Deep_learning)

머신러닝이란?



머신러닝이란?

A.I. vs ML vs DL

- 인공지능

외부 관찰자에게 인간처럼 스마트하게 소프트웨어를 작동시키는 폭넓은 방법, 알고리즘 및 기술

머신러닝, 컴퓨터 비전, 자연어 처리, 로봇 공학 및 그와 관련된 모든 주제를 포괄하는 개념

- 머신러닝

더 많은 데이터 축적을 통해 성능을 개선할 수 있도록 하는 다양한 알고리즘과 방법론

신경망, 서포트 벡터 머신, 결정 트리, 베이지안 신뢰 네트워크, k 최근접 이웃, 자기 조직화 지도, 사례 기반 추론,

인스턴스 기반 학습, 은닉 마르코프 모델, 회귀 기법

- 딥 러닝

신경망(Neural Network)을 부르는 다른 이름

여러 개의 히든 레이어를 통해 깊게 학습한다고 해서 붙여진 이름

머신러닝의 유명한 정의

"Not explicitily programmed & With task, performance measure, experience"

"Field of study that **gives computers the ability to learn without being explicitly programmed**" – Arthur Samuel (1959)

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" – T.Michell(1997)

Example1: A program for Go

T: Playing Go
P: Win or Lose
E: The number of Games

Example2: Spam mail Detection

T: Classifying emails as spam or ham
P: spam or ham
E: The number of classified emails

전통적인 프로그래밍과 머신러닝

<전통적 프로그래밍>

- 스팸 단어 파악
- 패턴 파악

- 스팸 단어 추가
- 패턴 추가
- 제어문(if문, for문)

전통적인 프로그래밍과 머신러닝

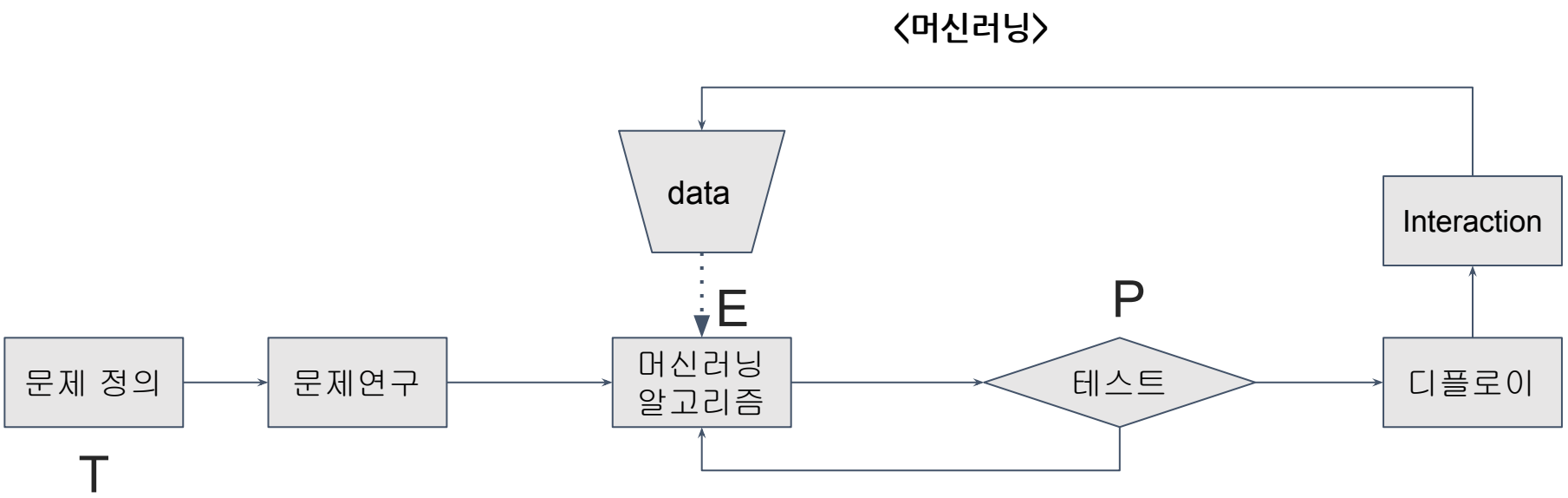
<전통적 프로그래밍>

- 스팸 단어 파악
- 패턴 파악

- 스팸 단어 추가
- 패턴 추가
- 제어문(if문, for문)

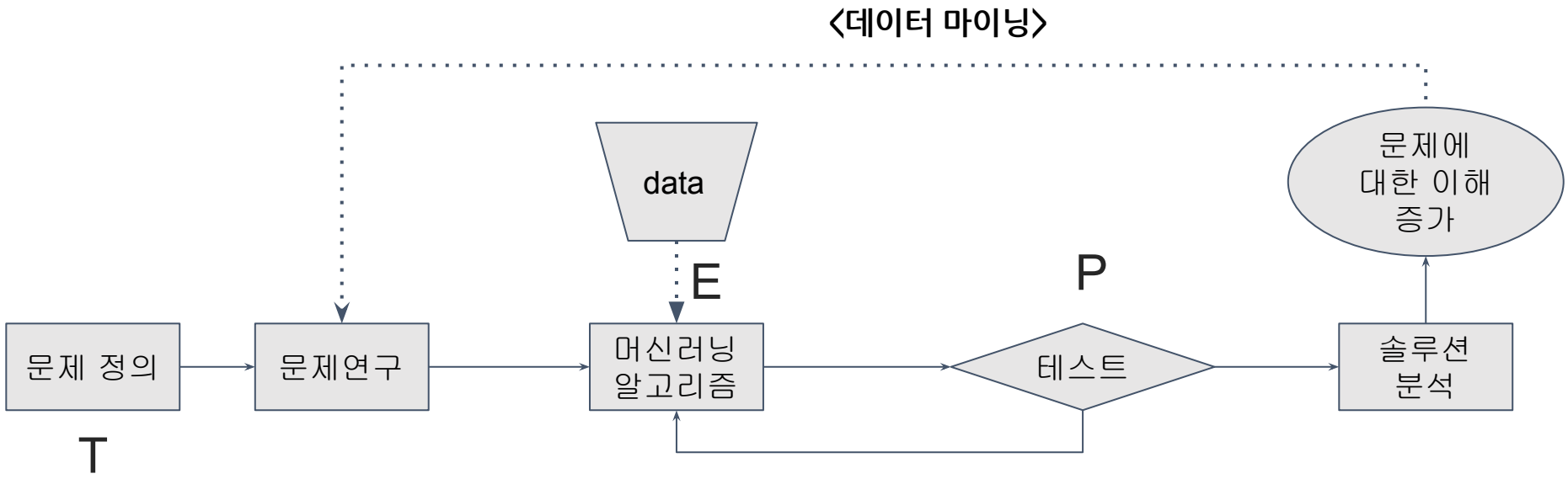
학습의 주체 = 인간

전통적인 프로그래밍과 머신러닝



T: Classifying emails as spam or ham
P: spam or ham
E: The number of classified emails

전통적인 프로그래밍과 머신러닝



학습의 주체 = 기계

T: Classifying emails as spam or ham
P: spam or ham
E: The number of classified emails

머신러닝 왜 필요한가

- 기존 솔루션으로는 많은 수동적인 규칙설정이 필요한 문제(스팸필터)
- 전통적인 방식으로는 해결 방법이 없는 문제 (음성인식)
- 실시간 규칙추가가 필요한 문제
- 데이터에 대한 추가적인 통찰력을 얻기 위함 (데이터 마이닝)

머신러닝 알고리즘의 종류

머신러닝의 개요

머신러닝의 종류 #1

〈훈련 방식에 따라〉

지도학습	비지도학습	강화학습	준지도학습
Supervised Learning	Unsupervised Learning	Reinforcement Learning	Semisupervised Learning

머신러닝의 종류 #2

<실시간인지 아닌지>

온라인 학습

- 가용한 데이터를 모두 사용해 훈련.
- 새로운 버전 = 처음부터 다시 훈련
- 실시간 처리가 중요한 문제 (단기 주가 예측)에 사용 불가

오프라인 학습 (배치학습)

- 데이터를 점차적으로 학습시킴
- 매 학습 단계가 빠름.
- 실시간 처리가 중요한 문제 (단기 주가 예측)에 사용 가능.

머신러닝의 종류 #3

〈일반화 접근법에 따라〉

사례 기반 학습

- 사례로 구분
- 유사도로 구분 (거리)

모델 기반 학습

- 사례로부터 하나의 모델을 구함
- 모델에 의해 일반화

머신러닝의 종류

〈훈련 방식에 따라〉 〈실시간인지 아닌지〉 〈일반화 접근법에 따라〉

배타적이지 않음

〉〉〉 **정답이 달린 데이터셋**을 이용해 **특정한 모델로 일반화 하는 머신러닝 알고리즘**을 온라인 상에서 **실시간으로 학습**시킴

⇒ **지도학습**이면서 **온라인학습**이고 **모델 기반학습** 일 수가 있음.

훈련방식에 따른 분류

1. Supervised Learning - 지도학습

감시되는 학습?

무엇을?

정답이 맞는지 아닌지.

- 정답이 있는 데이터셋.

  (정답 = Label = 레이블 = 라벨)

〈문제 종류〉

Regression & Classification

(회귀 & 분류)

훈련방식에 따른 분류

1. Supervised Learning - 지도학습

회귀(Regression)
- 연속적인 변수 예측
예) 새로운 붓꽃의 petal_width를 예측

분류(Classification)
- 범주형 변수 예측
예) 새로운 붓꽃의 종류를 예측

iris 데이터

Regression의 label Classification의 label

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	target_names
0	6.7	2.5	5.8	1.8	2.0	virginica
1	7.7	3.8	6.7	2.2	2.0	virginica
2	6.4	2.9	4.3	1.3	1.0	versicolor
3	6.4	2.8	5.6	2.1	2.0	virginica
4	5.2	3.4	1.4	0.2	0.0	setosa
5	5.7	2.8	4.5	1.3	1.0	versicolor
6	5.0	3.5	1.6	0.6	0.0	setosa
7	5.4	3.4	1.7	0.2	0.0	setosa
8	5.7	2.5	5.0	2.0	2.0	virginica
9	7.0	3.2	4.7	1.4	1.0	versicolor
10	6.2	2.2	4.5	1.5	1.0	versicolor
11	5.4	3.9	1.3	0.4	0.0	setosa

About 새로운 12번째 데이터

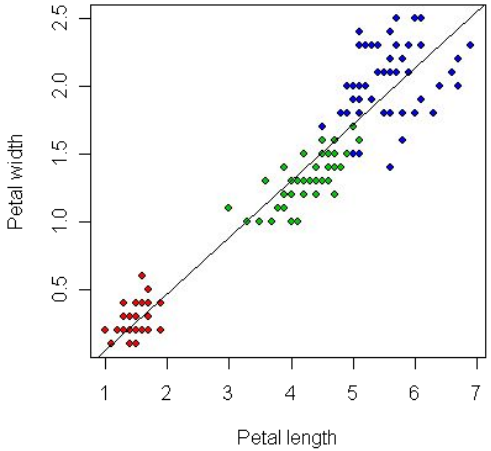
훈련방식에 따른 분류

1. Supervised Learning - 지도학습

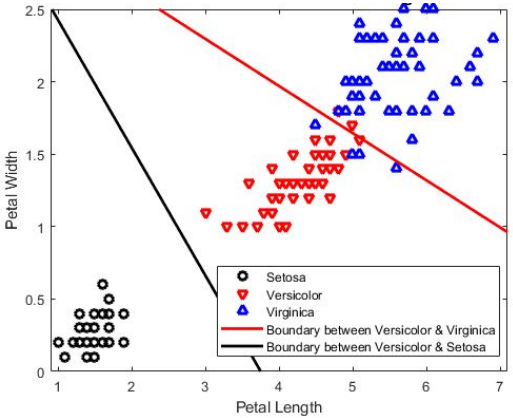
회귀(Regression)
- 연속적인 변수 예측
예) 새로운 붓꽃의 petal_width를 예측

분류(Classification)
- 범주형 변수 예측
예) 새로운 붓꽃의 종류를 예측

Regression



classification



훈련방식에 따른 분류

1-2. Unsupervised Learning - 비지도학습

- **정답이 없는 데이터셋**

  (정답 = Label = 레이블 = 라벨)

〈문제 종류〉

Clustering &
Dimensionality reduction

(군집화 & 차원축소)

훈련방식에 따른 분류

1-2. Unsupervised Learning - 비지도학습

군집화(Clustering)
- 그룹으로 세분화

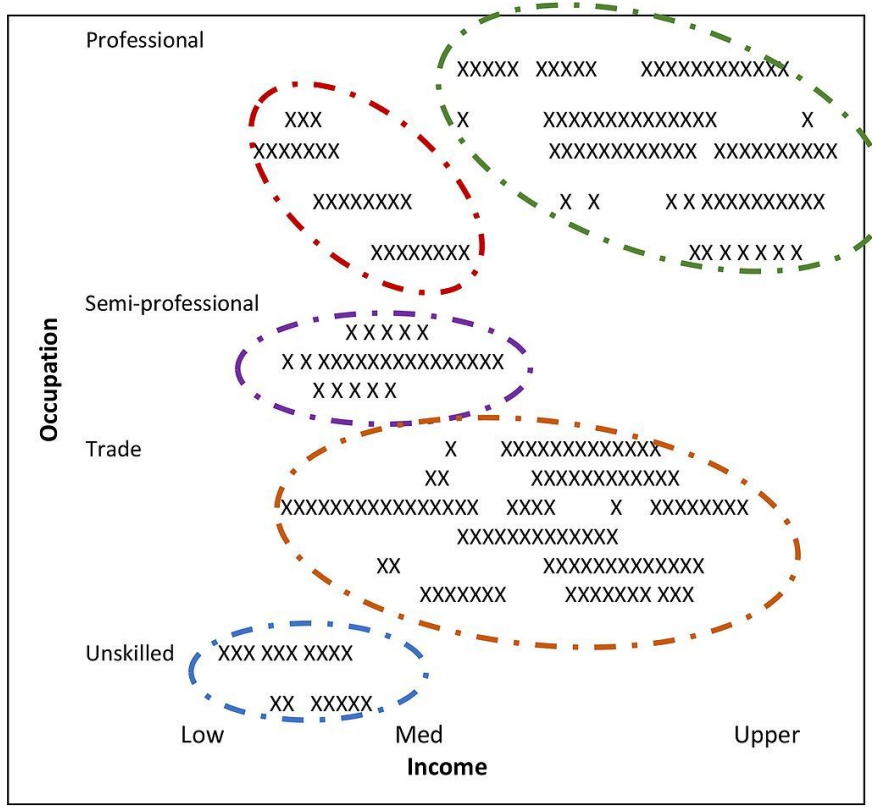
예)
- 수입에 따른 직업 군집화
- 페이스북 방문자 세분화 & 영화 추천 등

차원축소(Dimensionality Reduction)
- 상관관계가 있는 여러 특성을 하나로 합침

예) 차의 주행거리와 연식은 매우 연관 -> 합칠 수 있음

→ 알고리즘의 실행 속도가 훨씬 좋아진다.
(시간, 공간적 측면에서)

수입에 따른 직업 군집화



훈련방식에 따른 분류

1-2. Unsupervised Learning - 비지도학습

군집화(Clustering)
- 그룹으로 세분화

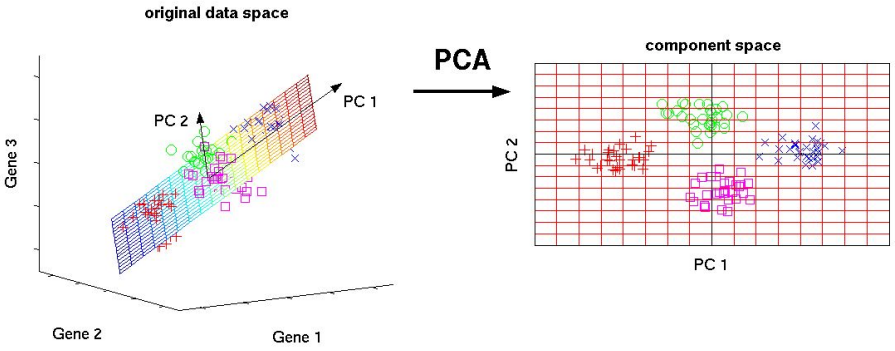
예)
- 수입에 따른 직업 군집화
- 페이스북 방문자 세분화 & 영화 추천 등

차원축소(Dimensionality Reduction)
- 상관관계가 있는 여러 특성을 하나로 합침

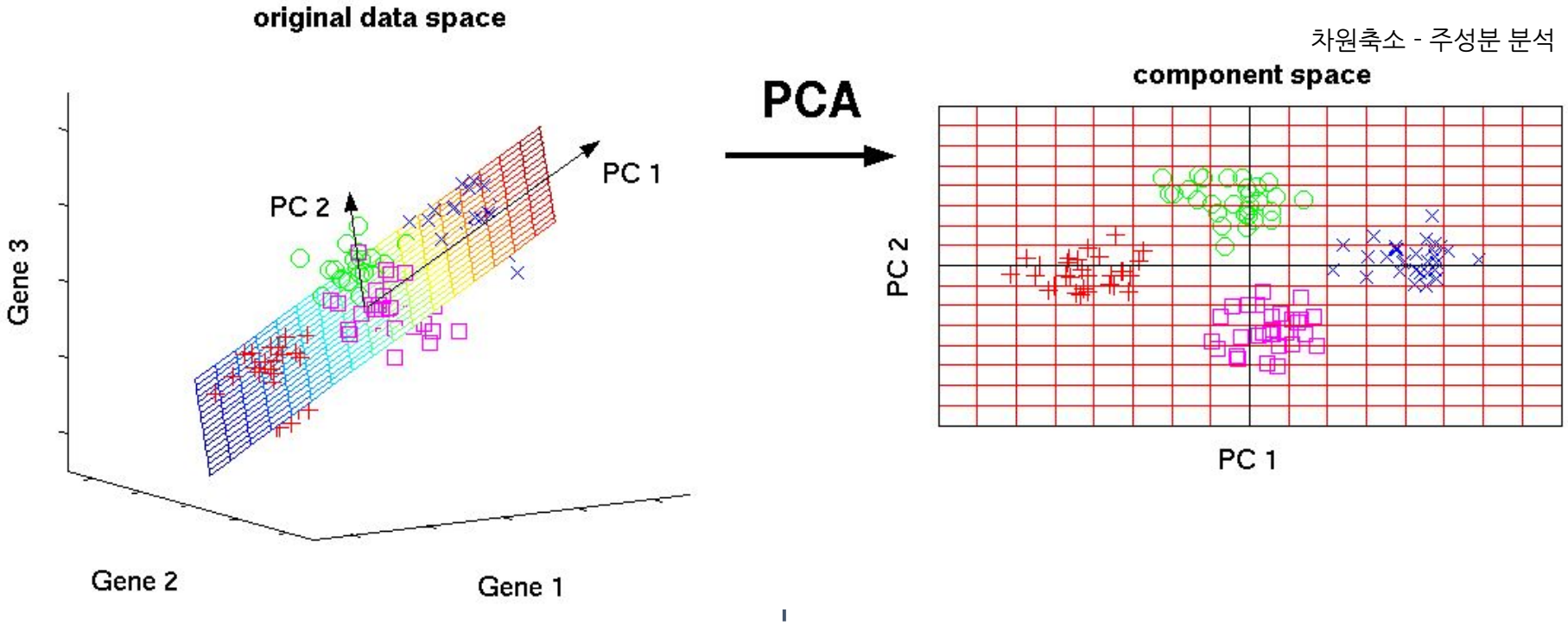
예) 차의 주행거리와 연식은 매우 연관 -> 합칠 수 있음

→ 알고리즘의 실행 속도가 훨씬 좋아진다.
(시간, 공간적 측면에서)

차원축소 - 주성분 분석



훈련방식에 따른 분류



차원축소 - 주성분 분석

훈련방식에 따른 분류

1-3. Semisupervised Learning - 준지도학습

- **정답이 있는 것도 있지만, 대부분 없는 데이터셋**
- **데이터의 정답을 얻기 위해서는 훈련된 사람의 손을 거쳐야 함 => 비용문제**

(정답 = Label = 레이블 = 라벨)

지도학습 + 비지도학습

예) 페이스북 사진 (***가 나온 사진)
- 클러스터링 후 레이블 입력

훈련방식에 따른 분류

1-4. Reinforcement Learning - 강화학습

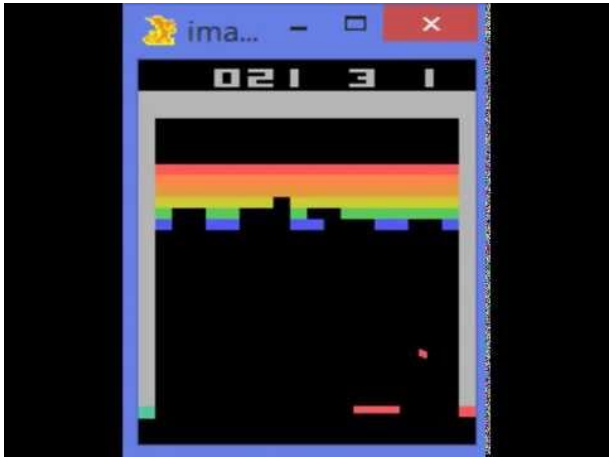
마르코프 결정과정

1. Agent(시스템)가 환경을 관찰하고 행동을 실행함
2. 행동에 대한 결과로 reward(보상) 혹은 penalty(벌점)를 줌.
3. 큰 보상을 얻기 위한 policy를 채택하기 위한 학습

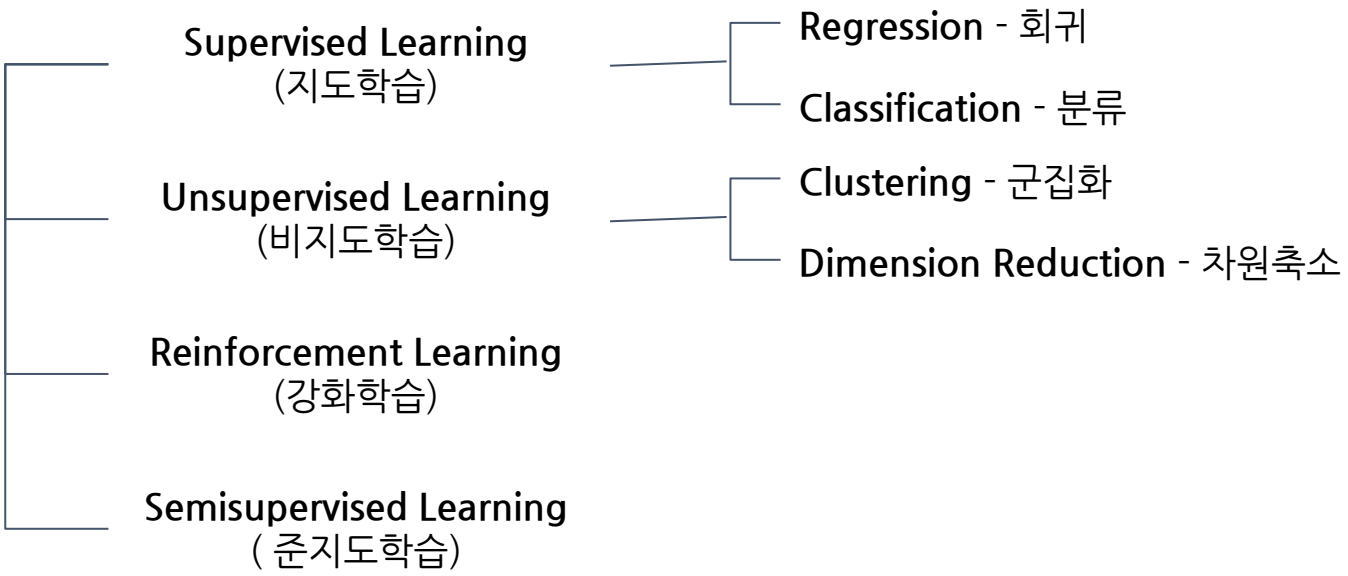
예) 알파고

1. 형세를 판단한다.
2. n번씩 두었을 때, 형세를 판단한다.
3. 형세가 좋으면 reward, 안 좋으면 penalty.

AlphaGo



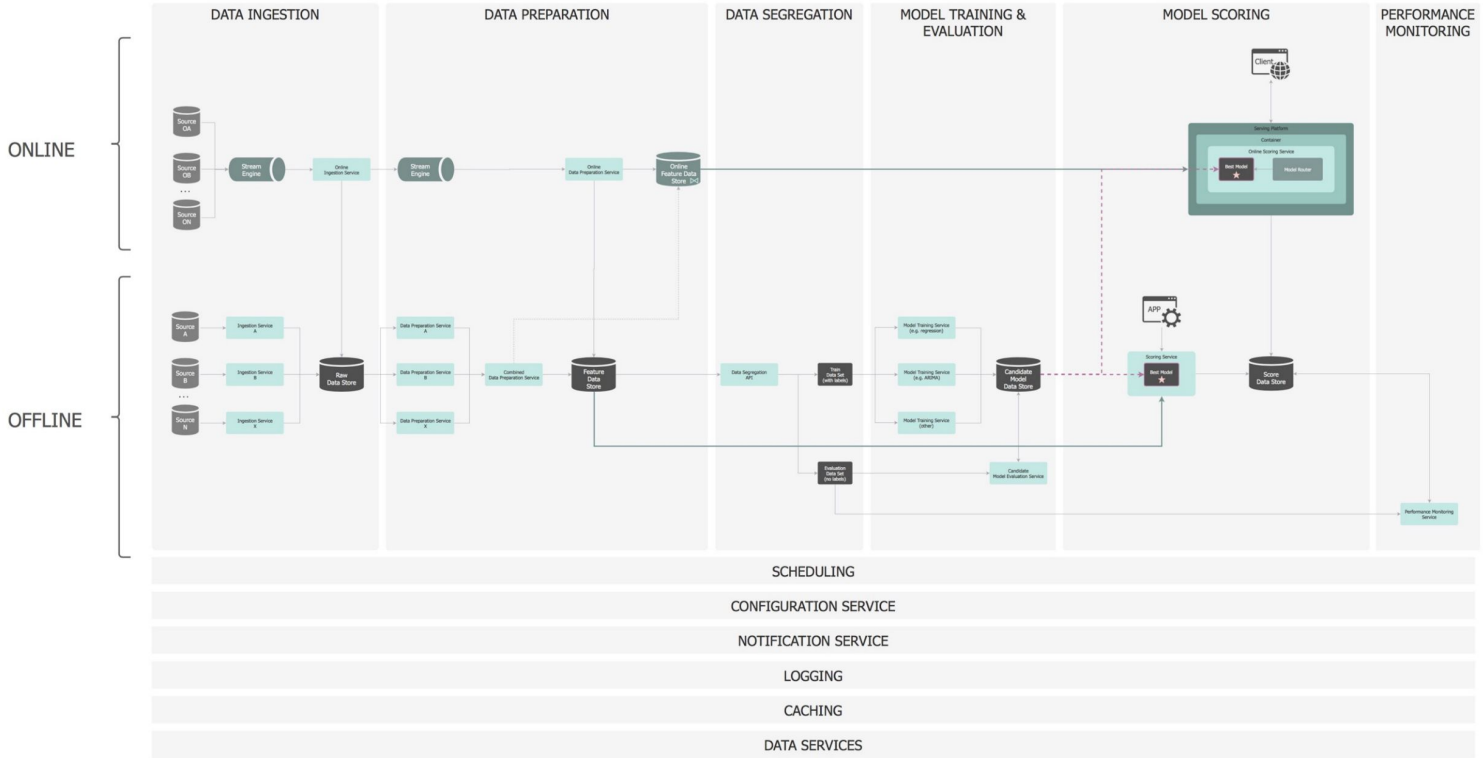
머신러닝의 종류



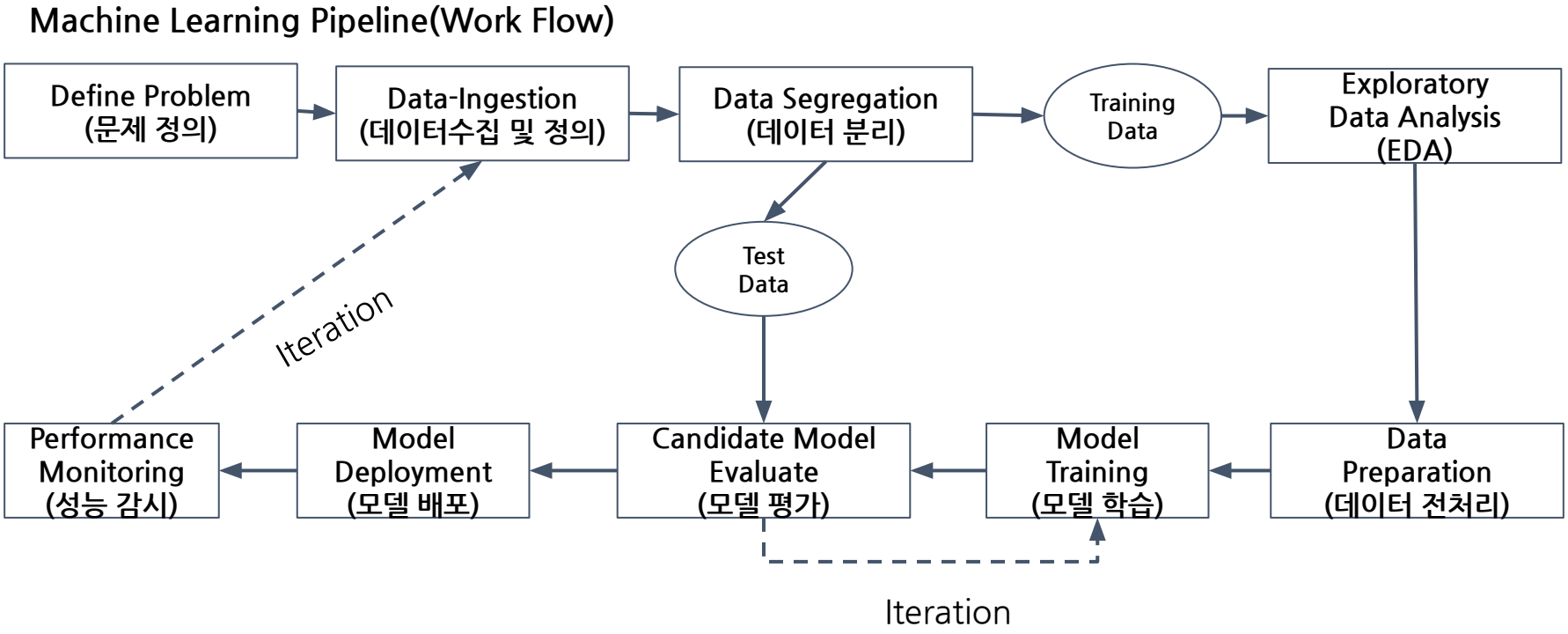
머신러닝 프로세스

머신러닝의 개요

머신러닝 프로세스



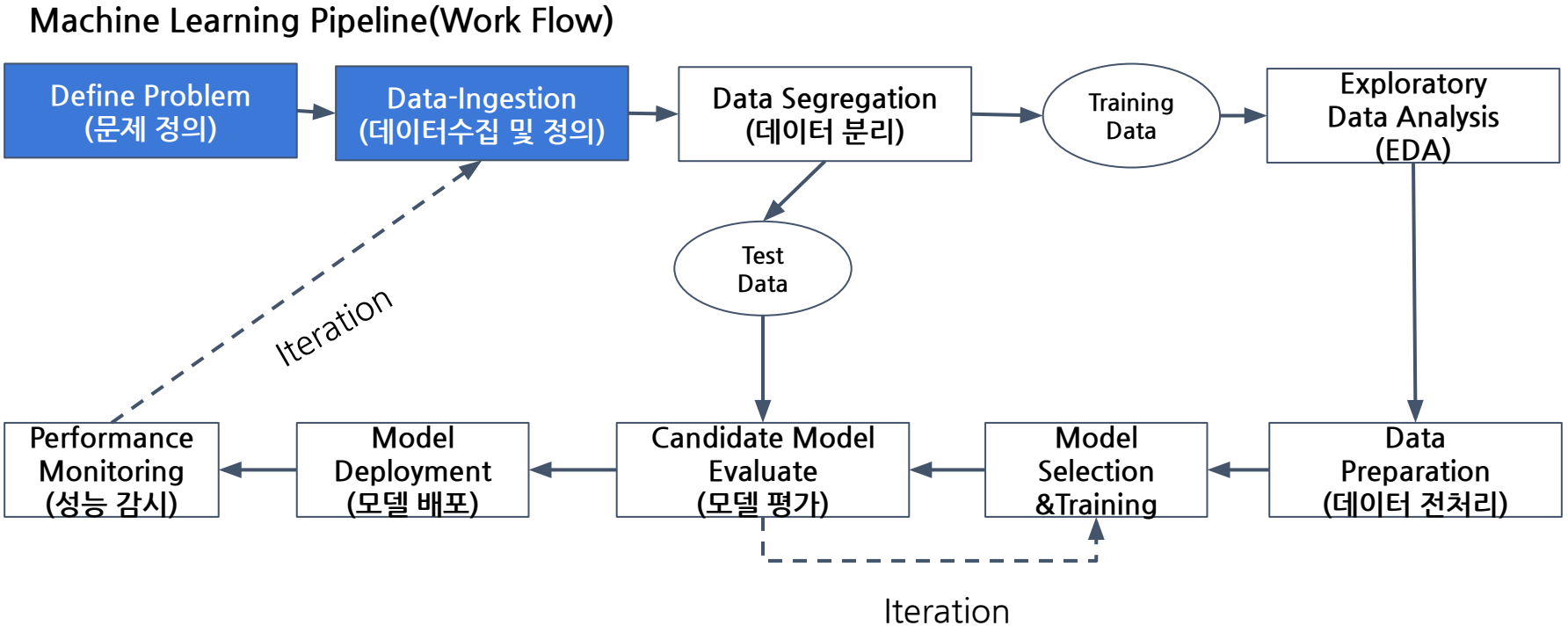
머신러닝 프로세스



사고 실험

머신러닝의 개요

1. Define Problem



1. Define Problem

우리는 부동산 업자

집 가격에 민감하게 받아들일 수 밖에 없는데,

이 집이 시세보다 싼지 혹은 비싼지

집가격을 어떻게 예측할 수 있을까?

1. Define Problem

부동산 투자를 위한 머신러닝

머신러닝의 목적은?

부동산 투자가치 분석하기

2. Data Ingestion

어떤 데이터가 필요할까?

위치, 집의 크기, 층 수, 리모델링 여부, 집 가격,
침실 개수, 화장실 개수, 정부 정책, 대체투자상품,
소득수준, 생활수준, 거시경제 흐름, ...

도메인 영역의 중요성

2. Data Ingestion

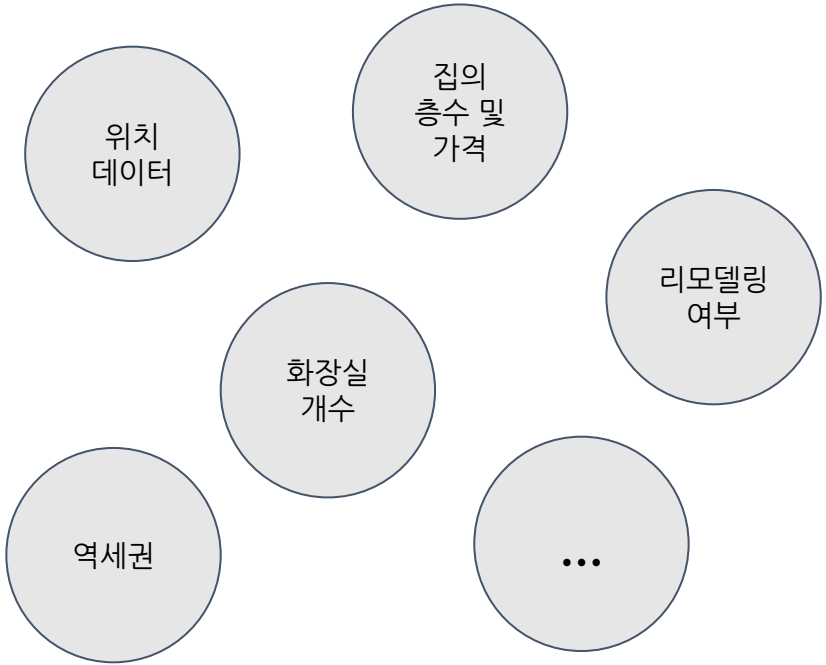
어떻게 해당 데이터를 수집할 수 있을까?

통계청 자료, 공공데이터, 온라인자료, 구매, 크롤링
등등..

데이터 정의 및 수집

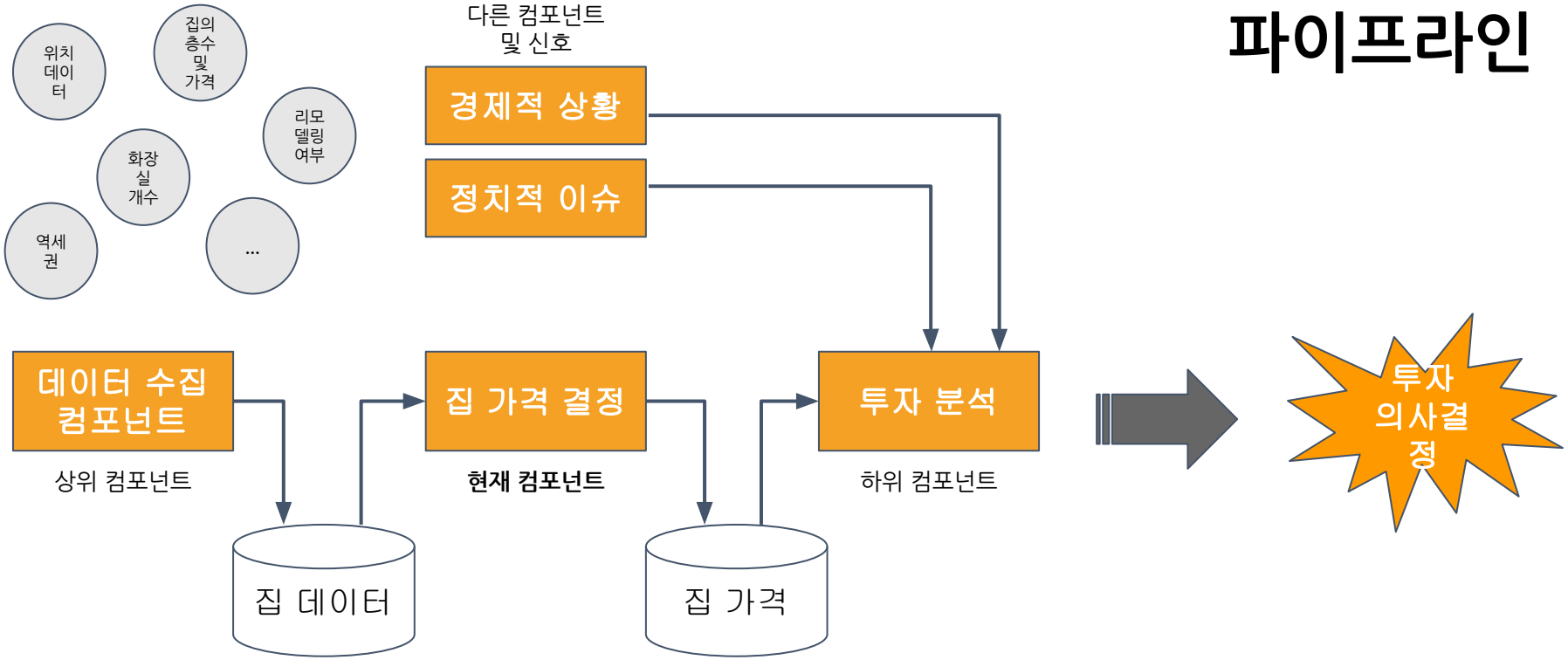
2. Data Ingestion

데이터의 파편화



2. Data Ingestion

파이프라인



파이프라인

컴포넌트가 연속적으로 연결되어 있는 것

컴포넌트: 하나의 Task 단위

크롤링 컴포넌트, 머신러닝 컴포넌트, 스케쥴링 컴포넌트...

DataSet - 집 데이터

California Housing Price Prediction

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-118.20	33.94	42.0	618.0	163.0	680.0	179.0	3.3472	154200.0	<1H OCEAN
1	-122.03	37.62	32.0	2964.0	547.0	1472.0	527.0	4.2468	221200.0	NEAR BAY
2	-122.19	37.82	32.0	1835.0	264.0	635.0	263.0	8.3170	365900.0	NEAR BAY
3	-118.00	33.77	24.0	1324.0	267.0	687.0	264.0	3.4327	192800.0	<1H OCEAN
4	-118.31	34.17	12.0	3188.0	931.0	2118.0	850.0	3.1823	218300.0	<1H OCEAN
5	-120.46	37.31	26.0	3170.0	572.0	1524.0	565.0	3.4800	95300.0	INLAND
6	-118.40	34.24	35.0	2552.0	545.0	1850.0	503.0	4.7750	179500.0	<1H OCEAN
7	-119.70	36.30	10.0	956.0	201.0	693.0	220.0	2.2895	62000.0	INLAND
8	-121.93	37.72	26.0	2806.0	459.0	1453.0	444.0	4.9107	213800.0	<1H OCEAN
9	-121.18	39.23	8.0	2112.0	360.0	782.0	344.0	3.7125	175000.0	INLAND
10	-121.47	38.56	51.0	2083.0	559.0	874.0	524.0	2.0221	95800.0	INLAND

02. Califonia (machine_learning_process).ipynb
H1

2. ML 문제 정의

California Housing Price Prediction

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-118.20	33.94	42.0	618.0	163.0	680.0	179.0	3.3472	154200.0	<1H OCEAN
1	-122.03	37.62	32.0	2964.0	547.0	1472.0	527.0	4.2468	221200.0	NEAR BAY
2	-122.19	37.82	32.0	1835.0	264.0	635.0	263.0	8.3170	365900.0	NEAR BAY
3	-118.00	33.77	24.0	1324.0	267.0	687.0	264.0	3.4327	192800.0	<1H OCEAN
4	-118.31	34.17	12.0	3188.0	931.0	2118.0	850.0	3.1823	218300.0	<1H OCEAN

지도학습 **VS** **비지도학습** **VS** **강화학습** **VS** **준지도학습**

온라인 학습 **VS** **오프라인 학습 (배치학습)**

우리가 예측해야할 값은?

ML 문제 정의

California Housing Price Prediction

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-118.20	33.94	42.0	618.0	163.0	680.0	179.0	3.3472	154200.0	<1H OCEAN
1	-122.03	37.62	32.0	2964.0	547.0	1472.0	527.0	4.2468	221200.0	NEAR BAY
2	-122.19	37.82	32.0	1835.0	264.0	635.0	263.0	8.3170	365900.0	NEAR BAY
3	-118.00	33.77	24.0	1324.0	267.0	687.0	264.0	3.4327	192800.0	<1H OCEAN
4	-118.31	34.17	12.0	3188.0	931.0	2118.0	850.0	3.1823	218300.0	<1H OCEAN

지도학습 VS 비지도학습 VS 강화학습 VS 준지도학습

회귀문제

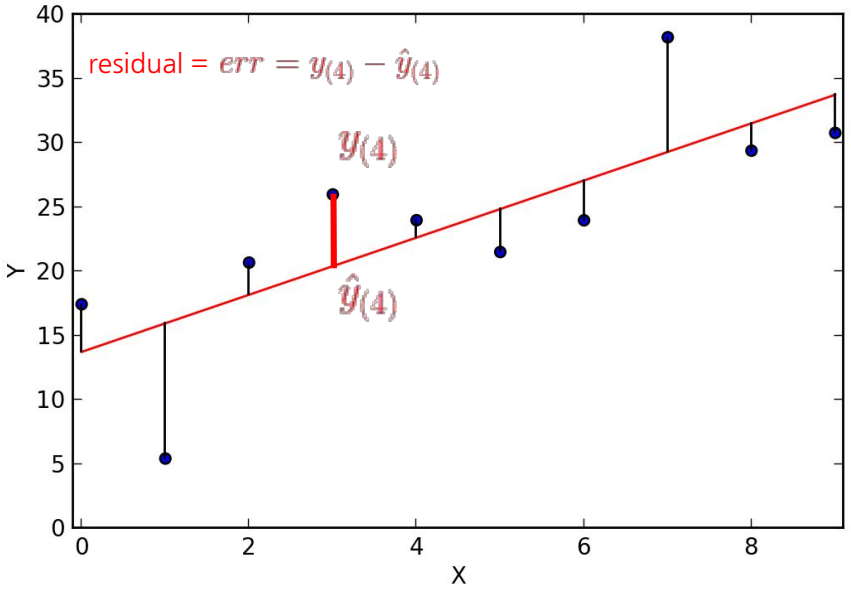
온라인 학습 VS 오프라인 학습 (배치학습)

Performance Measure Index (Cost Function)

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R Squared (R^2)
- Adjusted R Squared (R^2)
- Mean Square Percentage Error (MSPE)
- Mean Absolute Percentage Error (MAPE)
- Root Mean Squared Logarithmic Error (RMSLE)

부록: MAE(Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}|$$



Performance Measure Index (Cost Function)

RMSE(Root Mean Squared Error)

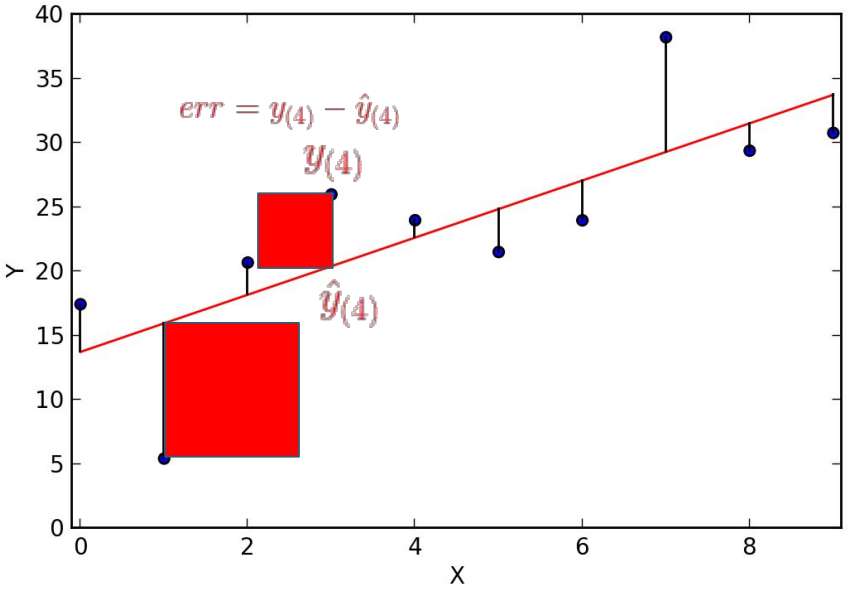
평균 제곱근 오차

$$RMSE(X,h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x^{(i)}) - y^{(i)})^2}$$

RMSE(Root Mean Squared Error)

$$RMSE(X,h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_{(i)}) - y_{(i)})^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{(i)})) - y_{(i)})^2}$$



X = 데이터셋 행렬.
x(i) = 데이터셋의 i번째 데이터 - 벡터
y(i) = i번째 정답
h(x(i)) = i번째 x에 대해 예측한 값
h: hypothesis → 가설. x에 대해 y를 예측하는 모델

RMSE 사용 이유

거리가 먼 것(차이가 큰 것)에
더 큰 페널티

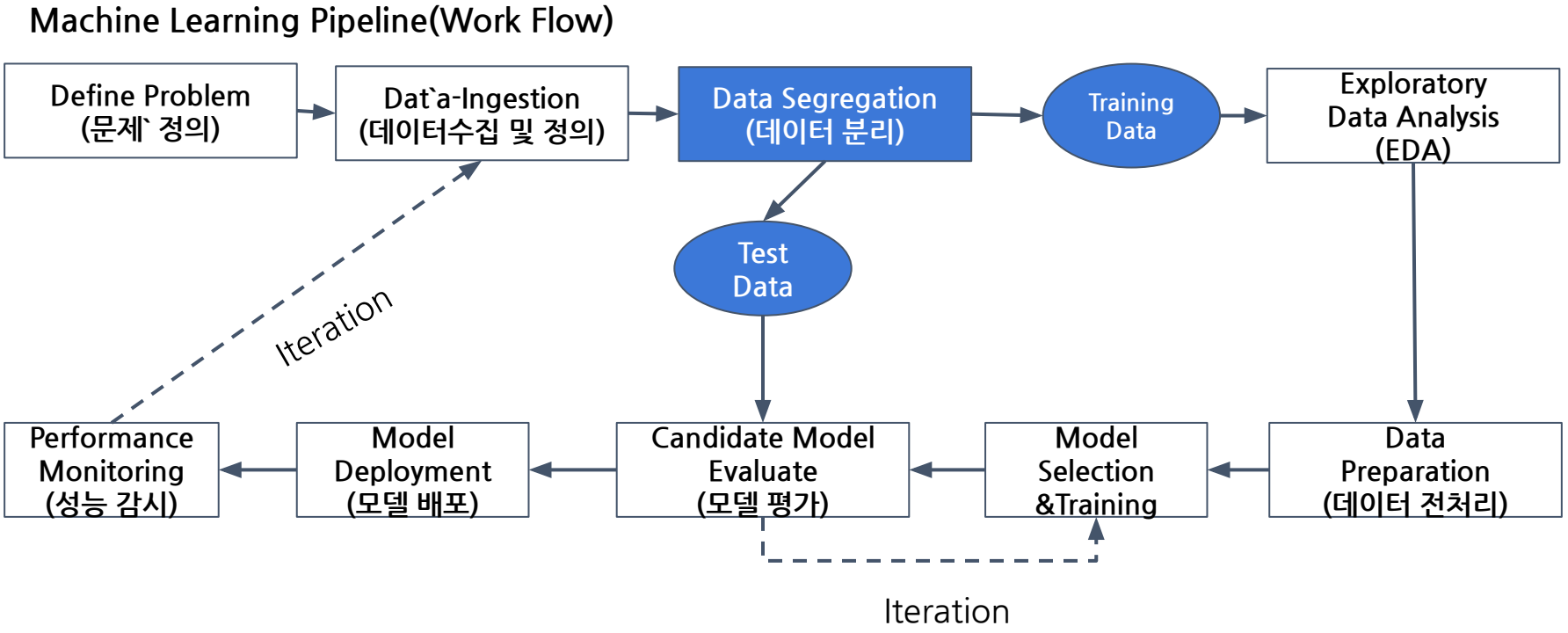
큰 에러를 최대한 줄이자!

Regression의 전형적인 성능지표.

단점: 이상치가 많을 경우 문제발생.

→ MAE 사용 or 이상치 제거 필요

3. Data Segregation



Data Segregation

Test Data 와
Training Data의 분리

Data Segregation

Why Do This?

평가를 위함(“머신러닝이 잘 되었나?”)

- Overfitting 경계 → 일반화

Why Do Now?

사용자의 Model 선택 편견을 줄이기 위해

테스트 세트의 패턴의 속아 Model 선택 가능성

Data Segregation

70 Vs 30

OR

80 Vs 20

Method

Random Sampling
무작위 분리

Stratified Sampling
층화(계층)적 분리

Stratified Sampling

- male, full-time: 90
- male, part-time: 18
- female, full-time: 9
- female, part-time: 63
- total: 180

Stratified Sampling

- % male, full-time = 90 ÷ 180 = 50%
- % male, part-time = 18 ÷ 180 = 10%
- % female, full-time = 9 ÷ 180 = 5%
- % female, part-time = 63 ÷ 180 = 35%

Stratified Sampling

- male, full-time = 90 × (126 ÷ 180) = 63
- male, part-time = 18 × (126 ÷ 180) = 13
- female, full-time = 9 × (126 ÷ 180) = 6
- female, part-time = 63 × (126 ÷ 180) = 44

Stratified Sampling

California Housing Price Prediction

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-118.20	33.94	42.0	618.0	163.0	680.0	179.0	3.3472	154200.0	<1H OCEAN
1	-122.03	37.62	32.0	2964.0	547.0	1472.0	527.0	4.2468	221200.0	NEAR BAY
2	-122.19	37.82	32.0	1835.0	264.0	635.0	263.0	8.3170	365900.0	NEAR BAY
3	-118.00	33.77	24.0	1324.0	267.0	687.0	264.0	3.4327	192800.0	<1H OCEAN
4	-118.31	34.17	12.0	3188.0	931.0	2118.0	850.0	3.1823	218300.0	<1H OCEAN
5	-120.46	37.31	26.0	3170.0	572.0	1524.0	565.0	3.4800	95300.0	INLAND
6	-118.40	34.24	35.0	2552.0	545.0	1850.0	503.0	4.7750	179500.0	<1H OCEAN
7	-119.70	36.30	10.0	956.0	201.0	693.0	220.0	2.2895	62000.0	INLAND
8	-121.93	37.72	26.0	2806.0	459.0	1453.0	444.0	4.9107	213800.0	<1H OCEAN
9	-121.18	39.23	8.0	2112.0	360.0	782.0	344.0	3.7125	175000.0	INLAND
10	-121.47	38.56	51.0	2083.0	559.0	874.0	524.0	2.0221	95800.0	INLAND

NEAR BAY: **30%**
<1H OCEAN: **20%**
INLAND: **50%**

NEAR BAY: **0.3*TOTAL*SPLIT_RATIO**
<1H OCEAN: **0.2*TOTAL*SPLIT_RATIO**
INLAND: **0.5*TOTAL*SPLIT_RATIO**

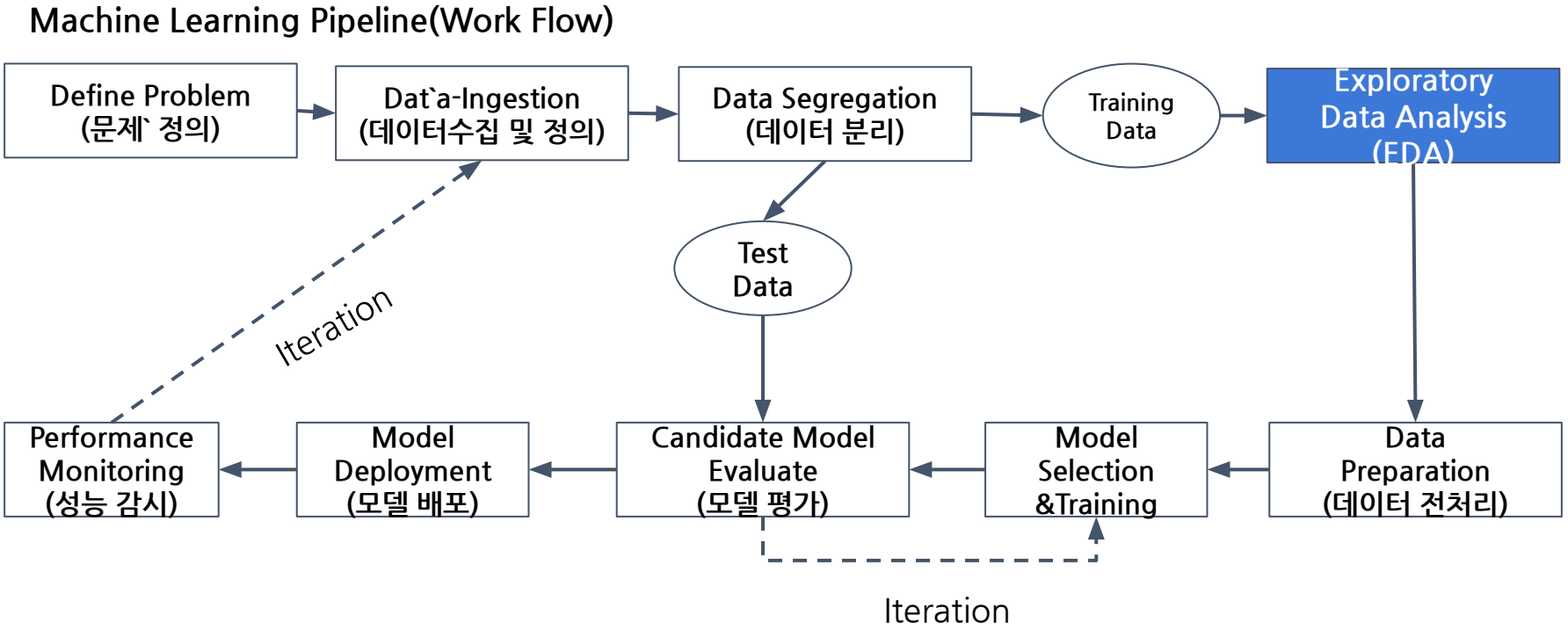
Stratified Sampling

California Housing Price Prediction

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-118.20	33.94	42.0	618.0	163.0	680.0	179.0	3.3472	154200.0	<1H OCEAN
1	-122.03	37.62	32.0	2964.0	547.0	1472.0	527.0	4.2468	221200.0	NEAR BAY
2	-122.19	37.82	32.0	1835.0	264.0	635.0	263.0	8.3170	365900.0	NEAR BAY
3	-118.00	33.77	24.0	1324.0	267.0	687.0	264.0	3.4327	192800.0	<1H OCEAN
4	-118.31	34.17	12.0	3188.0	931.0	2118.0	850.0	3.1823	218300.0	<1H OCEAN
5	-120.46	37.31	26.0	3170.0	572.0	1524.0	565.0	3.4800	95300.0	INLAND
6	-118.40	34.24	35.0	2552.0	545.0	1850.0	503.0	4.7750	179500.0	<1H OCEAN
7	-119.70	36.30	10.0	956.0	201.0	693.0	220.0	2.2895	62000.0	INLAND
8	-121.93	37.72	26.0	2806.0	459.0	1453.0	444.0	4.9107	213800.0	<1H OCEAN
9	-121.18	39.23	8.0	2112.0	360.0	782.0	344.0	3.7125	175000.0	INLAND
10	-121.47	38.56	51.0	2083.0	559.0	874.0	524.0	2.0221	95800.0	INLAND

높음
중간
낮음

4. Data Segregation



EDA (Exploratory Data Analysis)

<탐색적 데이터 분석>

데이터를 **탐색하여 방향성을 파악**.
가설의 시작이 데이터.

데이터를 시각화 등을 통해서 탐색 및 확인을 해보았는데, 집값은 동네 주민들의 소득수준과 상관관계가 있어보이는 그림이 나왔다.
이를 바탕으로 집값은 소득수준에 영향을 받을 것이라는 가설이 생겼다.

데이터로부터 인사이트

EDA (Exploratory Data Analysis)

<탐색적 데이터 분석>

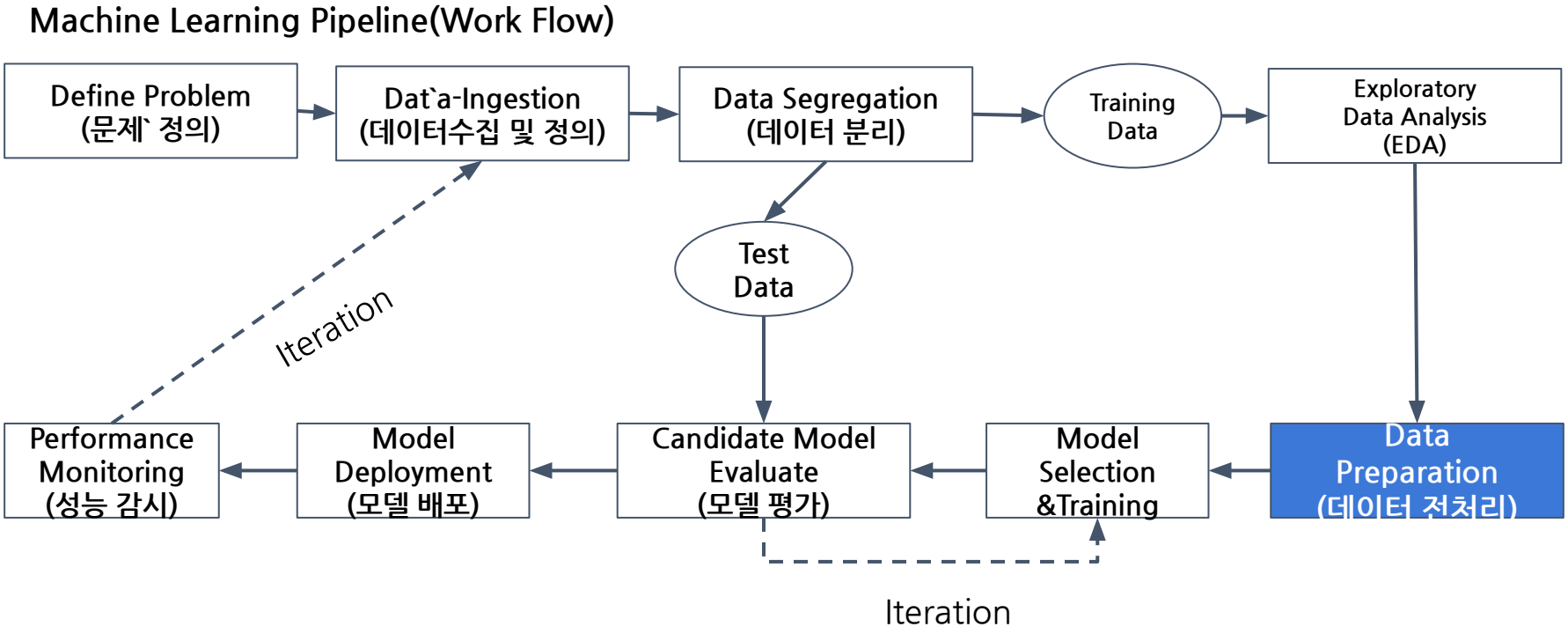
데이터 분포

데이터 시각화

상관분석

특성조합

4. Data Segregation



데이터 전처리 - Data Preprocessing

일반적인 처리 과정

데이터 셋 확인	결측값(N/A) 처리	Feature Engineering
EDA	- 삭제 - 대표값 대체 - 예측값 삽입	- Scaling - Binning - Transform - Dummy

결측값 처리 (MissingValue treatment)

>> hounsing.info( )

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16512 entries, 5250 to 13976
Data columns (total 9 columns):
longitude             16512 non-null float64
latitude              16512 non-null float64
housing_median_age    16512 non-null float64
total_rooms           16512 non-null float64
total_bedrooms        16344 non-null float64
population            16512 non-null float64
households            16512 non-null float64
median_income         16512 non-null float64
ocean_proximity       16512 non-null object
dtypes: float64(8), object(1)
memory usage: 1.3+ MB
```

삭제

대표값으로 대체

예측값 삽입

02. Califonia (machine_learning_process).ipynb
H5

Feature Engineering #2

Categorical Data 처리
(범주형 → 수치형)

머신러닝 알고리즘은 대부분 수치형 자료만 다룸

Encoding의 필요성

Feature Engineering #3

Feature Scaling

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
count	16512.000000	16512.000000	16512.000000	16512.000000	16344.000000	16512.000000	16512.000000	16512.000000
mean	-119.566444	35.622168	28.668362	2621.941194	535.850649	1422.233164	498.051962	3.869586
std	2.001586	2.127758	12.592234	2155.624205	417.154776	1130.392675	378.459720	1.891238
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900
25%	-121.800000	33.930000	18.000000	1446.000000	297.000000	788.000000	279.000000	2.566950
50%	-118.490000	34.250000	29.000000	2126.000000	435.000000	1167.000000	410.000000	3.536000
75%	-118.010000	37.700000	37.000000	3136.000000	646.000000	1722.000000	603.000000	4.747050
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100

Feature 간에 스케일이 매우 다르다.

스케일을 맞춰주는 것

Feature Engineering #3

Feature Scaling

min-max normalization

0 <= x <= 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

큰 이상치가 존재시
문제 발생 가능성

Standardization

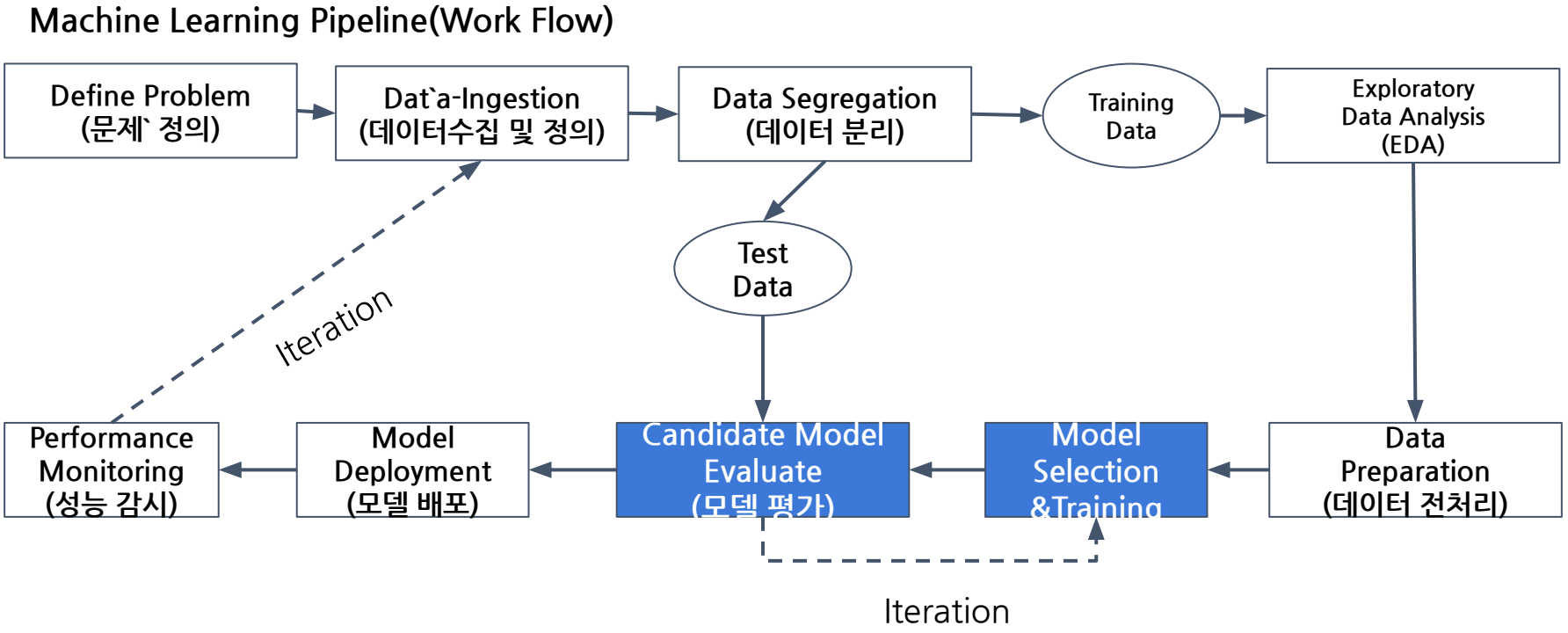
평균:0 / 분산:1

$$x' = \frac{x - \bar{x}}{\sigma}$$

이상치에 영향을 잘 안 받음,
상한과 하한이 없기 때문

02. Califonia (machine_learning_process).ipynb
H5-3

4. Data Segregation



Model Selection & Training

Model Selection

Supervised Learning

- KNN
- Linear Regression
- Logistic Regression
- SVM
- Decision Tree / Random Forest
- Neural Network

Unsupervised Learning

- K-means
- HCA
- PCA
- …

Model Selection & Training

Model Selection

Supervised Learning

- KNN
- Linear Regression
- Logistic Regression
- SVM
- Decision Tree / Random Forest
- Neural Network

수많은 알고리즘 중 어떤 걸 선택?

Model Selection & Training

Model Selection

Supervised Learning

- KNN
- Linear Regression
- Logistic Regression
- SVM
- Decision Tree / Random Forest
- Neural Network

수많은 알고리즘 중 어떤 걸 선택?

여러개 해보자!

Model Selection & Training

Model Selection

Supervised Learning

- KNN
- Linear Regression
- Logistic Regression
- SVM
- Decision Tree / Random Forest
- Neural Network

수많은 알고리즘 중 어떤 걸 선택?

여러개 해보자!

Model Selection & Training

Model Training

머신러닝 알고리즘 적용 및 평가

[Optimization]

Model Selection & Training

The goal of ML is never to make "perfect" guesses, because ML deals in domains where there is no such thing. The goal is to make guesses that are good enough to be useful.