Machine Learning with Python

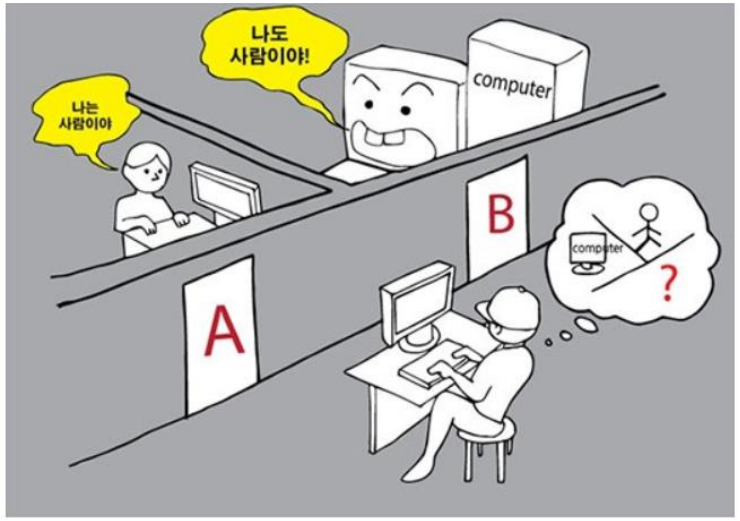
인공지능

- **지능 (知能, intelligence)**
  - 본능적이나 자동적으로 행동하는 대신에, 생각하고 이해하여 행동하는 능력

- **인공지능 (人工知能, Artificial Intelligence)**
  - 사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 구현하는 기술
  - 컴퓨터나 기계가 사람이 하는 것처럼 생각하고 행동할 수 있게 하는 기술
  - 튜링 테스트(Turing test)를 통과하면 (강한) 인공지능으로 판단

튜링 테스트(Turing test)

- **튜링 테스트(Turing test)**
  - 기계가 인간과 얼마나 비슷하게 대화할 수 있는지를 기준으로 기계에 지능이 있는지를 판별하고자 하는 테스트
  - 앨런 튜링이 1950년에 제안



강한 인공지능과 약한 인공지능

- 강한 인공지능 (strong AI)
  - 사람과 같은 지능, 인간을 완벽하게 모방한 인공지능
  - 마음을 가지고 사람처럼 느끼면서 지능적으로 행동하는 기계
  - 추론, 문제해결, 판단, 계획, 의사소통, 자아 의식(self-awareness). 감정(sentiment), 지혜(sapience), 양심(conscience)
  - 튜링 테스트

- 약한 인공지능 (weak AI, narrow AI)
  - 특정 문제를 해결하는 지능적 행동, 유용한 도구로 사용하기 위해 설계된 인공지능
  - 사람의 지능적 행동을 흉내 낼 수 있는 수준
  - 대부분의 인공지능 접근 방향
  - 중국인 방 사고 실험(Chinese room thought experiment)

강한 인공지능과 약한 인공지능

- **중국인 방 사고 실험(The Chinese Room Thought Experiment)**
  - John Searle (1980) 제시
    - 문 밑으로 중국어로 쓴 질문지를 전달
    - 방 안에서 중국어를 모르는 사람이 글자모양에 따른 중국어 단어 조합 방법 매뉴얼을 참조하여 답변에 대한 단어 조합
    - 조합된 단어들을 문 밖으로 내보냄
    - 문 밖 사람은 중국어를 이해하는 사람이 방안에 있다고 생각
    - 단지 흉내만 내고 이해하는 것은 아님
  - 이해하지 못하고 흉내 낼 수 있어도 지능적(intelligent) 행동

인공지능의 역사

인공지능(AI)의 역사

1943년
워런 맥컬록과 월터 피츠, 전기 스위치처럼 켜고 끄는 기초기능의 인공신경을 그물망 형태로 연결하면 사람의 뇌에서 동작하는 아주 간단한 기능을 흉내낼 수 있음을 증명

1950년

앨런 튜링, 기계가 인간과 얼마나 비슷하게 대화할 수 있는지를 기준으로 기계에 지능이 있는지를 판별하는 튜링 테스트 제안

1956년

다트머스 회의에서 인공지능 용어 처음 사용. "학습의 모든 면 또는 지능의 다른 모든 특성을 기계로 정밀하게 기술할 수 있고 이를 시뮬레이션할 수 있다"

1958년
프랭크 로젠블래트, 뇌신경을 모사한 인공신경 뉴런 '퍼셉트론' 제시

1970년대
AI 연구가 기대했던 결과를 보여주지 못하자 대규모 투자가 중단되며 암흑기 도래

1980년대
전문가들의 지식과 경험을 데이터베이스화해 의사결정 과정을 프로그래밍화한 '전문가 시스템' 도입. 그러나 관리의 비효율성과 유지·보수의 어려움으로 한계

1997년
IBM 딥블루, 체스 챔피언 게리 카스파로프와의 체스 대결에서 승리

2006년
제프리 힌튼 토론토대 교수, 딥러닝 알고리즘 발표

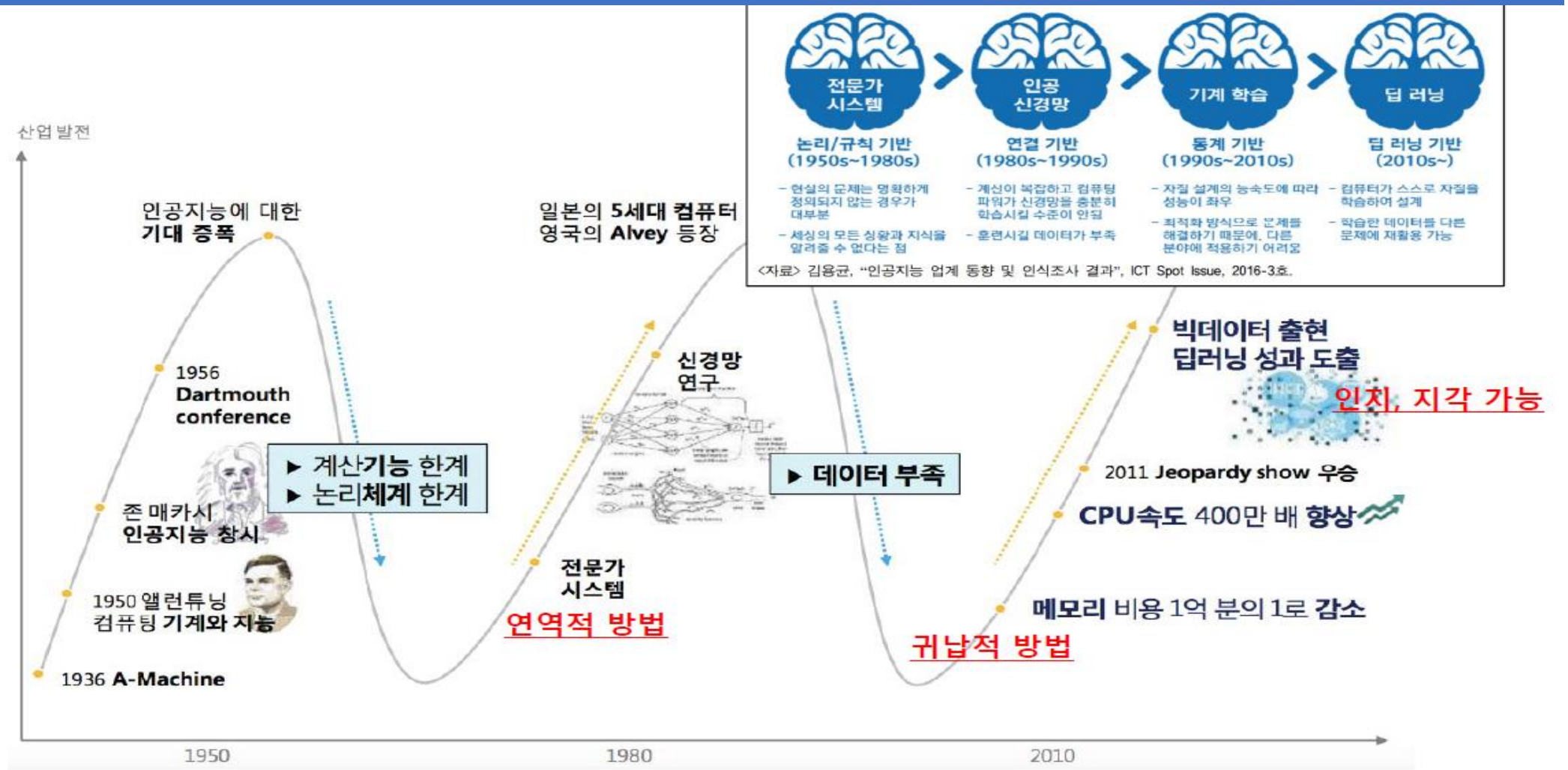
2012년

국제 이미지 인식 경진대회 '이미지넷'에서 딥러닝 활용한 팀이 우승하며 획기적 전환점

2014년
구글, 딥마인드 인수

2016년
구글 알파고, 이세돌에게 승리

인공지능의 역사


전문가 시스템
인공 신경망
기계 학습
딥 러닝
논리/규칙 기반 (1950s~1980s)
연결 기반 (1980s~1990s)
통계 기반 (1990s~2010s)
딥 러닝 기반 (2010s~)
- 현실의 문제는 명확하게 정의되지 않는 경우가 대부분
- 세상의 모든 상황과 지식을 알려줄 수 없다는 점
- 계산이 복잡하고 컴퓨팅 파워가 신경망을 충분히 학습시킬 수준이 안됨
- 훈련시킬 데이터가 부족
- 자질 설계의 능숙도에 따라 성능이 좌우
- 최적화 방식으로 문제를 해결하기 때문에, 다른 분야에 적용하기 어려움
- 컴퓨터가 스스로 자질을 학습하여 설계
- 학습한 데이터를 다른 문제에 재활용 가능
<자료> 김용균, "인공지능 업계 동향 및 인식조사 결과", ICT Spot Issue, 2016-3호.
산업 발전
인공지능에 대한 기대 증폭
1956 Dartmouth conference
존 매카시 인공지능 창시
1950 앨런튜닝 컴퓨팅 기계와 지능
1936 A-Machine
▶ 계산기능 한계
▶ 논리체계 한계
일본의 5세대 컴퓨터 영국의 Alvey 등장
신경망 연구
전문가 시스템
연역적 방법
▶ 데이터 부족
빅데이터 출현 딥러닝 성과 도출
인지, 지각 가능
2011 Jeopardy show 우승
CPU속도 400만 배 향상
메모리 비용 1억 분의 1로 감소
귀납적 방법
1950
1980
2010

인공지능 vs. 머신러닝 vs. 딥러닝

인공지능
Artificial Intelligence

사고방식이나 학습 등
인간이 가지는 지적 능력을
컴퓨터를 통해 구현하는 기술

머신러닝
Machine Learning

컴퓨터가 스스로 학습하여
인공지능의 성능을
향상 시키는 기술 방법

딥러닝
Deep Learning

인간의 뉴런과 비슷한
인공신경망 방식으로
정보를 처리

머신러닝 분류

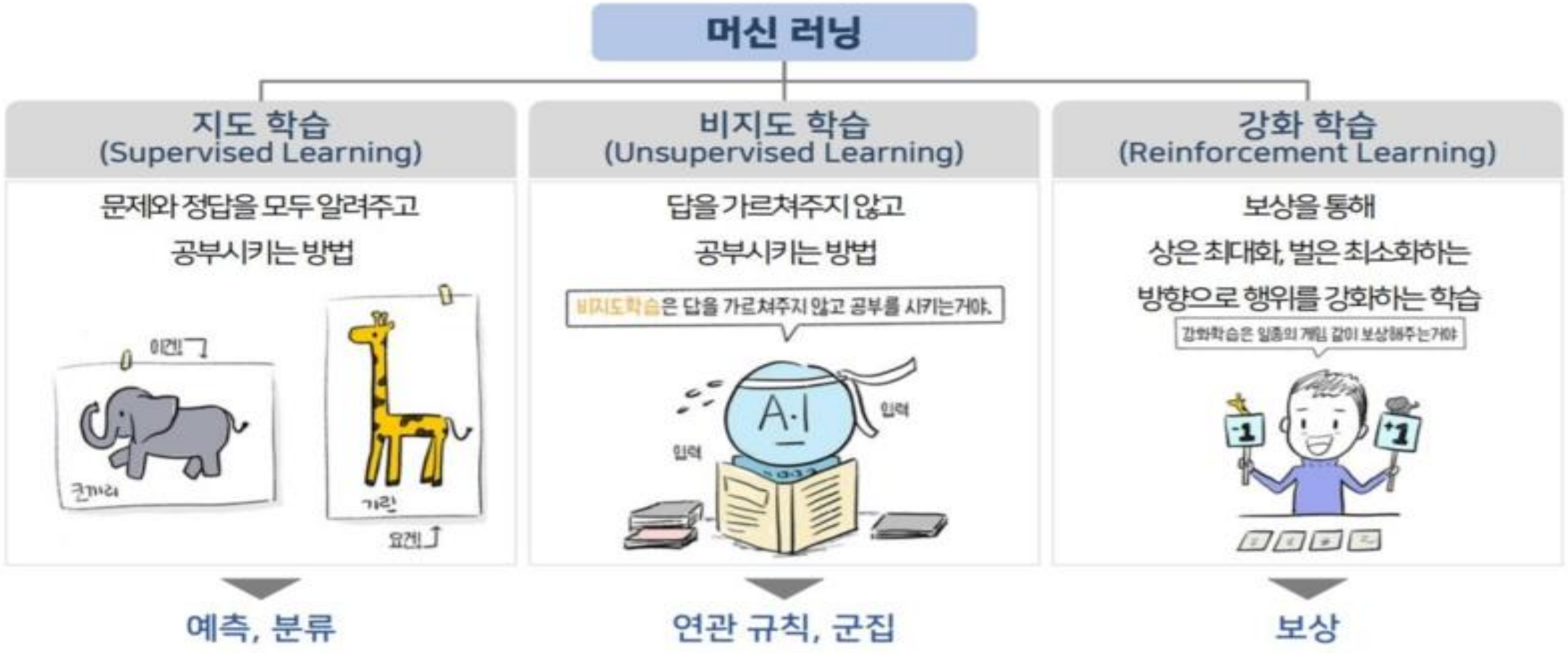
비지도학습
Unsupervised Learning
레이블 없이 학습
즉각적인 피드백 없음
데이터의 숨겨진 특징 발견

지도학습
Supervised Learning
레이블된 데이터
즉각적 피드백
미래 데이터 예측

기계학습
Machine Learning

강화학습
Reinforcement Learning
의사결정을 위한
최적의 액션 선택
보상시스템으로 학습

머신러닝 분류

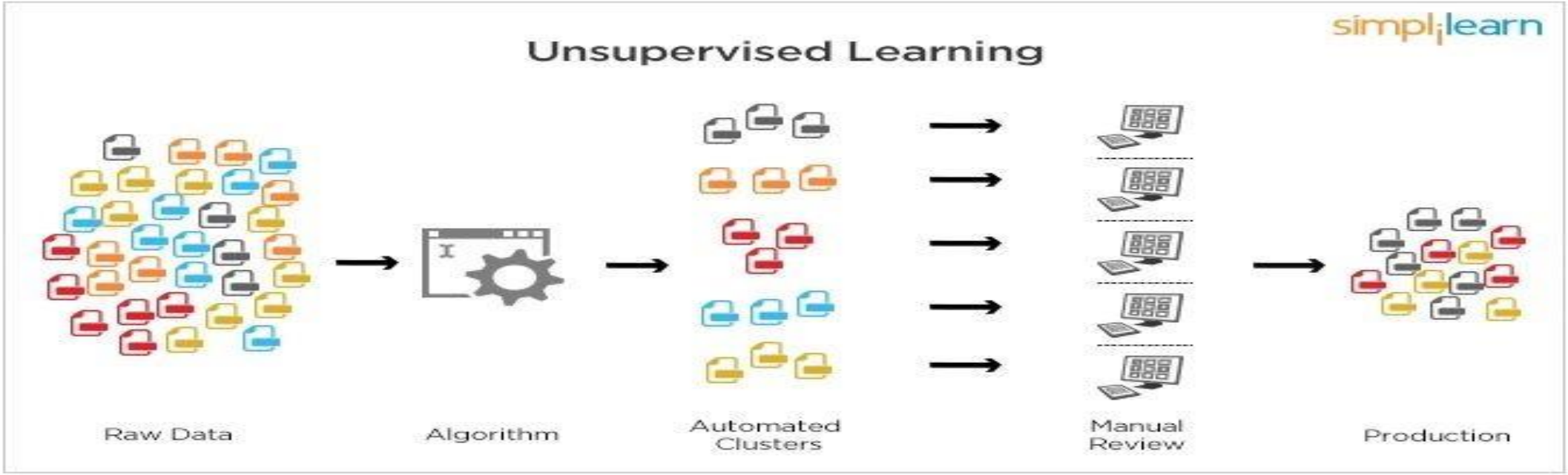

머신 러닝
지도 학습
(Supervised Learning)
문제와 정답을 모두 알려주고
공부시키는 방법
비지도 학습
(Unsupervised Learning)
답을 가르쳐주지 않고
공부시키는 방법
비지도학습은 답을 가르쳐주지 않고 공부를 시키는거야.
A.I
입력
입력
강화 학습
(Reinforcement Learning)
보상을 통해
상은 최대화, 벌은 최소화하는
방향으로 행위를 강화하는 학습
강화학습은 일종의 게임 같이 보상해주는거야
-1
+1
예측, 분류
연관 규칙, 군집
보상

지도학습(Supervised Learning)

- 지도학습이란 **정답이 주어진 상태에서 학습을 하는 알고리즘**을 말한다.

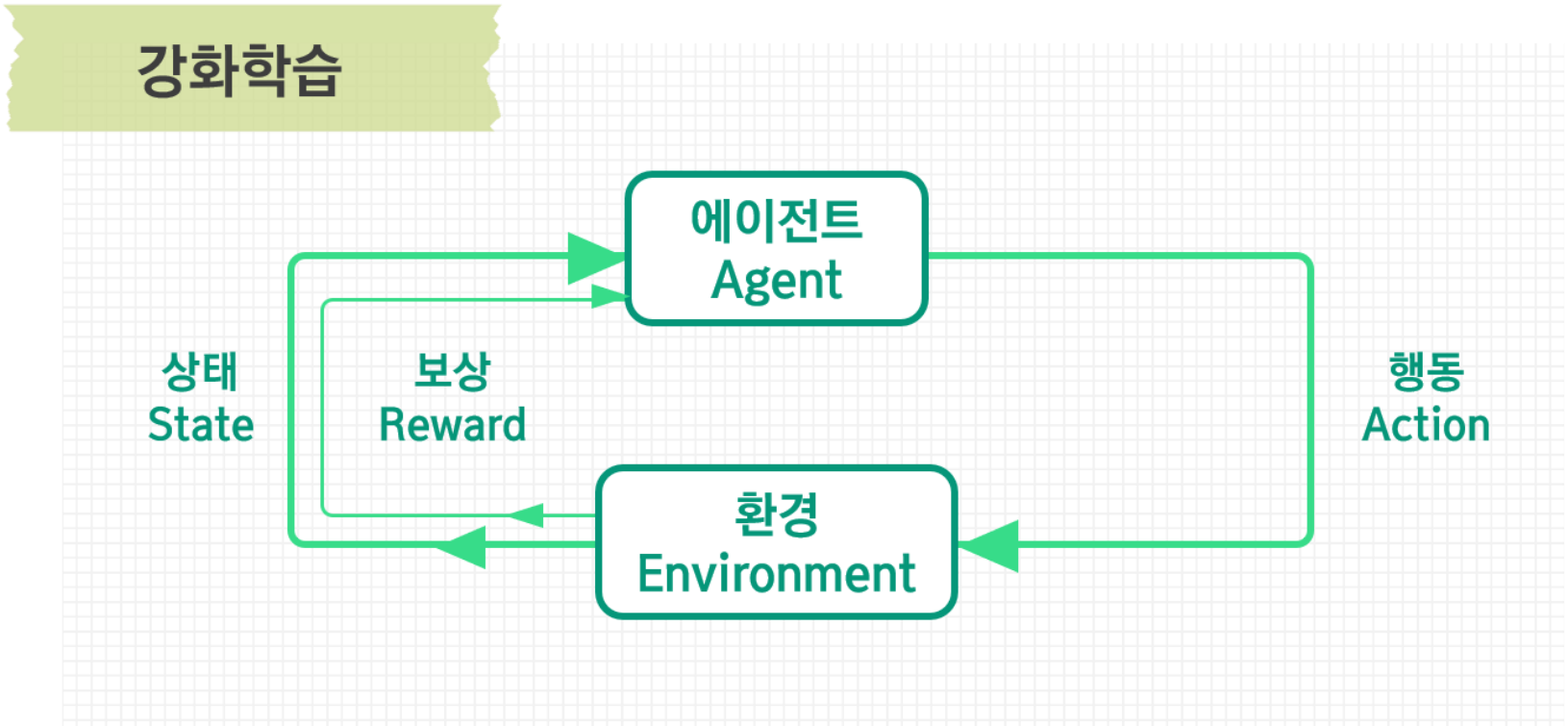
비지도 학습(Unsupervised Learning)

- 비지도 학습의 핵심은 '스스로'라고 말할 수 있다.
- 비지도학습이란 **정답이 주어지지 않은 상태에서 데이터의 특성을 학습하여 스스로 패턴을 파악하는 알고리즘**을 말한다. 사람들이 정답을 하나하나 입력하는 수고를 덜어주기 때문에 최근에 집중적으로 연구되는 분야이다. 가장 오랜 시간 동아 연구되어온 것이 군집화(Clustering)이다

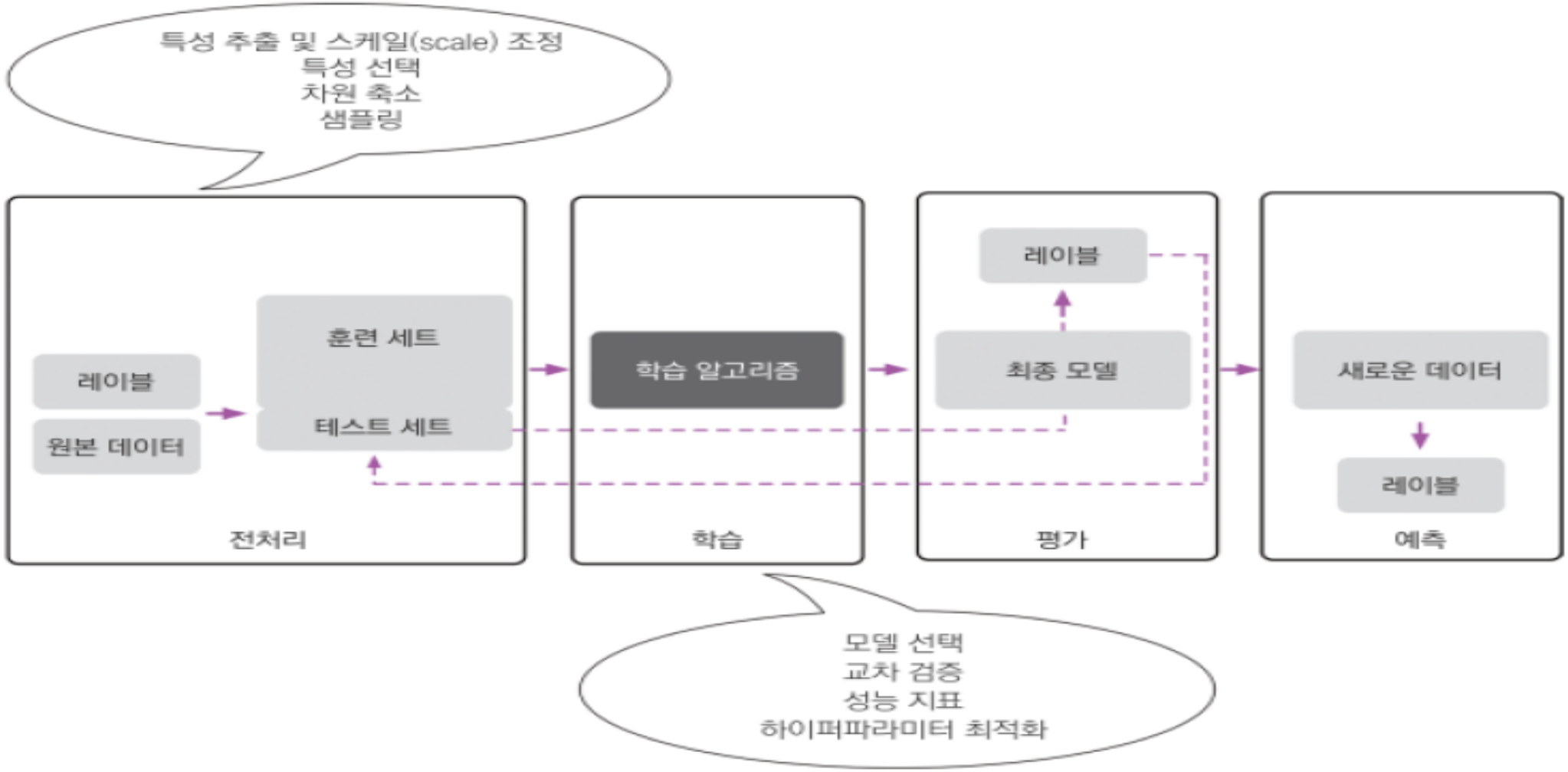


강화학습(Reinforcement Learning)

- 강화학습은 기계학습의 한 영역으로 행동심리학에서 영감을 받았다. 어떤 환경(Environment)안에서 정의되는 에이전트(Agent)가 현재의 상태(State)를 인식하여, **선택 가능한 행동(Action) 중 보상(Reward)을 최대화 하는 행동 혹은 행동 순서**를 선택하는 방법이다.



머신러닝 분석절차


특성 추출 및 스케일(scale) 조정
특성 선택
차원 축소
샘플링
레이블
원본 데이터
훈련 세트
테스트 세트
전처리
학습 알고리즘
학습
레이블
최종 모델
평가
새로운 데이터
레이블
예측
모델 선택
교차 검증
성능 지표
하이퍼파라미터 최적화

머신러닝에 사용되는 라이브러리

- 아나콘다(Anaconda) : 배포판
  - 주피터 노트북: IDE 프로그램 코드를 브라우저에서 실행
- Numpy(넘파이): 과학 계산용 패키지, 다차원 배열, 선형 대수 연산 등
- Matplotlib (맷플롯립): 그래프 패키지, 선 그래프, 히스토그램 등
- Pandas (판다스)
  - 데이터 처리와 분석, R의 data.frame과 유사한 DataFrame 사용
  - 엑셀 파일, CSV 파일 같은 다양한 파일 처리
- Sklearn(싸이킷런): 기계학습 패키지
- Tensorfow(텐서플로)
  - 그글에서 개발한 딥러닝용 고성능 패키지
  - GPU지원
- Keras(케라스)
  - 딥러닝용 고급패키지
  - 텐서플로1.1.0부터 통합

Machine Learning with Python

기초(기술) 통계량

기술통계량

- Pandas는 데이터를 보다 좀 더 편하게 다룰 수 있게 하는 데이터 구조 측면에서의 장점을 가진 패키지로 Pandas에서 제공하는 통계분석은 기본적인 기술통계 및 데이터요약이다.
- 고급통계 기법은 scikit-learn 이나 다른 통계패키지를 이용하여 수행할 수 있다.

함 수	작용원리
count	NA 를 제외한 개수
min, max	최소, 최대값
sum	합
cumprod	누적합
mean	평균
median	중앙값
quantile	분위수
Var	표본분산
std	표본 정규분산
Describe	요약통계량

변수

- **연속변수**
  - 연속적인 값을 가지는 변수
  - 예) 나이, 점수, 무게, 가격 등

- **범주변수(이산변수)**
  - 서로 다른 것으로 구분되는 변수
  - 예) 성별(남자/여자), 애완동물(강아지/고양이)

척도

- **척도(scale)** : 측정된 변수의 값을 표현하는 수준(levels of measurement)을 의미
- **명명척도(nominal scale)**
  - 측정값이 같고 다름을 말할 수 있음
  - 측정값들 사이에 순서가 없음
  - 사칙연산이 불가능
  - 종류에 따른 빈도만 계산
  - 예) 혈액형
- **서열척도(ordinal scale)**
  - 측정값들 사이에 순서가 있음
  - 측정값들의 간격이 동일하지 않음
  - 사칙연산은 불가능
  - 예) 직급(부장, 과장, 대리, …)
    - 부장이 과장보다, 과장이 대리보다 높음
    - 부장과 과장의 차이가 과장과 대리의 차이와 같지 않음

척도

- 등간척도(interval scale)
  - 측정값들 사이에 순서가 있고 간격이 일정
  - 영점(0)의 의미가 임의적 (영점을 옮겨도 무방함)
  - 덧셈, 뺄셈이 가능
  - 예) 섭씨온도
    - 섭씨 20도는 섭씨 10도보다 수치로는 2배
    - 그러나 2배 따뜻한 것이 아님
    - 화씨로 바꾸면 각각 50도와 68도가 되어 1.36배에 불과
    - 영점의 기준이 임의적이기 때문 (섭씨 0도 = 화씨 32도)
- 비율척도(ratio scale)
  - 등간척도 + 절대영점
  - 사칙연산 모두 가능
  - 절대영점이란, 영점의 의미가 아무 것도 존재하지 않는 상태를 말함
  - 예) 길이
    - 20미터는 10미터보다 수치로도 2배이고
    - 실제로도 2배 길
    - 미터를 피트로 바꿔도 32.8피트와 65.6피트로 2배
    - 영점의 기준이 절대적 (0 미터 = 0 피트)

도수분포표와 히스토그램

- 도수분포표(frequency table) : 데이터를 구간으로 나누어, 각 구간의 빈도를 나타낸 표
- 히스토그램(histogram) : 도수분포표를 그래프로 그린 것

중심경향치

- **중심경향치(central tendency measures)**
  - 자료의 중심을 나타내는 숫자
  - 자료 전체를 대표
  - 평균, 중간값, 최빈값 등이 있다.

- **평균(mean)** : 자료의 합을 자료의 개수로 나눈 값

- **중간값(median)** : 자료를 크기 순으로 정렬했을 때 정 가운데에 있는 값
  - 자료의 상위 50%와 하위 50%를 가르는 지점
  - '중앙값' 또는 '중위수'라고도 한다.

- **최빈값(mode)** : 가장 빈번하게 관찰/측정되는 값

변산성 측정치(분산, 표준편차, 범위, 사분위간 범위)

- 변산성(variability) : 자료가 흩어져 있는 정도, 혹은 개체에 따라 변할 수 있는 정도
- 중심경향치가 자료가 **무엇을 중심으로** 모여 있는가(혹은 흩어져 있는가)를 나타내는 것이라면,
- 변산성 측정치는, 그 **모여 있는 정도**(혹은 흩어져 있는 정도)를 의미함

범위(Range)

- 자료가 갖는 최대값과 최소값 사이의 거리, 즉 자료가 얼마나 퍼져 있는가를 나타냄
- 범위 = 최댓값 – 최솟값

분산

- 평균에서 데이터가 벗어난 정도를 수치화한 값
- 각각의 데이터에서 평균값을 빼고, 그것을 제곱하여 평균을 구함
- 분산이 크면 : 데이터가 평균에서 많이 벗어나 있다
- 분산이 작으면 : 데이터가 평균 주변에 모여 있다

```
numpy.var(x)
```

```
3.2399999999999998
```

사분위수

- 사분위간 범위(IQR, InterQuartile Range)는 제3사분위수에서 제1사분위수 간의 범위
- 사분위수란 전체 데이터를 작은 값부터 큰 값까지 순서대로 나열한 후 4등분 하였을 때, 각 지점에 해당하는 값
  - 제1사분위수(Q1): 25% 지점
  - 제2사분위수(Q2): 50% 지점 = 중간값
  - 제3사분위수(Q3): 75%
- 제1사분위수와 제3사분위수 사이의 구간에는 항상 전체 데이터의 50%가 포함 됨
- 사분위는 임의로 정하는 기준이므로 필요에 따라 십분위 등으로 변경가능

가설검정

통계 분석

- **모집단과 표본**
  - 모집단 : 우리가 알고자 하는 대상 전체. 조사 대상의 범위
  - 표본 : 모집단으로부터 조사하기 위해 선택된 조사 대상
- **전수조사와 표본조사**
  - 전수조사 : 모집단을 구성하는 대상 전부를 조사하는 것
    - 가장 정확하지만 비용과 시간이 많이 든다
    - 전수조사가 불가능한 경우도 있음(예: 감기약의 경우 모두 복용을 해야만 효과를 알 수 있다.)
  - 표본조사 : 표본을 대상으로 조사

통계 분석 기법

- **통계 분석 기법**
  - 어떤 그룹, 집단, 형태 등의 **차이를 검정**
    - 1개, 2개 또는 그 이상의 데이터 차이가 있다고 볼 수 있는지를 검정하는 것
    - 독립표본 T-검정, 대응표본 T-검정, ANOVA 등
      - 대응표본 : 한 집단으로부터 두 번 반복해 샘플 추출
      - 독립표본 : 서로 독립된 집단에서 각각 샘플 추출
  - 요소와 요소 간의 인과관계(상관관계)를 파악
    - 상관분석 – 변수와 변수 사이의 직선 관계를 상관계수를 이용해서 분석
    - 회귀분석 – 종속변수와 독립변수 간의 관계를 모형화 하여 분석

가설이란

- 가설의 정의
  - 모집단의 특성, 특히 모수에 대한 가정 혹은 잠정적인 결론
  - 분석을 통해 확인하고자 하는 명제
  - 과학분야에서의 증명 : **반증법**에 의거해 증명
    - ✓ “모든 사람이 정직하다 ” 라는 명제가 있을 때 이 명제가 참인지 거짓인지를 확인하는 접근법에는 2가지가 존재
    - ✓ 모든 사람을 일일이 조사해서 정직한지 확인하는 방법
    - ✓ 정직하지 못한 사람(사례)을 하나 찾아내 명제가 거짓임을 입증하는 방법

- 귀무가설(Null Hypothesis)
  - 대립가설과 반대되는 가설

- 대립가설(Alternative Hypothesis)
  - 표본을 통해 확실한 근거를 가지고 입증하고자 하는 가설로 연구 가설이라고도 한다.

가설검정

- 가설검정
  - 모집단에 대한 통계적 가설을 세우고 표본을 추출한 다음, 그 표본을 통해 얻은 정보를 이용하여 통계적 가설의 진위를 판단하는 과정
  - 표본을 활용하여 모집단에 대입해보았을 때 새롭게 제시된 대립가설이 옳다고 판단할 수 있는지를 평가하는 과정
  - 대부분 귀무가설이 참이라는 전제하에서 표본을 통하여 귀무가설이 옳지 않다는 것을 보임으로써 귀무가설을 기각시키고 대신 대립가설을 채택하게 되는 것

- p-value
  - 통계의 유의의성을 대표하는 지표
  - 일반적으로 0.05 이하일 때 통계적으로 의미가 있다고 인정
  - p-value $\geq$ 0.05 : 대립가설 기각, 귀무가설 채택
  - p-value $<$ 0.05 : 대립가설 채택, 귀무가설 기각

T-검정

- **독립표본 T-검정**
  - 서로 독립된 두 집단간의 평균의 차이가 통계적으로 유의미한 지 비교하고자 할 때 사용
  - 서로 독립된 두 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지 없는지를 통계적으로 검정할 때 사용하는 기법
  - 예 : 전체 응답자 중 남자와 여자 사이의 연령은 차이가 있는가?

- **대응표본 T-검정(Paired-sample t-test)**
  - 서로 동일한 모집단에서 추출된 두 표본에 대해 특정 연속성 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법
  - 예 : 한 회사에서 자사가 개발한 한 달 간의 식이요법 프로그램이 효과가 있는지 여부를 분석

- **일원배치 분산분석(One-way ANOVA)**
  - 세 개 이상의 집단간의 평균의 차이가 통계적으로 유의미한 지 비교하고자 할 때 사용
  - 예 : 학력수준에 따라 직무만족도의 수준은 차이가 있는가?

카이제곱 검정(Chi-Square Test)

- 두 범주형 변수(범주로만 분류되고 수치적으로 측정되지 않는 자료)가 서로 상관이 있는지 판단하는 통계적 검정 방법
- 예: 학력, 성별, 직업의 만족도 등
- 귀무가설 : 두 데이터는 차이가 없다.
- 대립가설 : 두 데이터는 차이가 있다.

- 피서 검정(Fisher's exact Test) : 표본수가 적거나 데이터의 분포가 치우친 경우에 적용한다.

Shapiro-Wilk 검정

- T-검정에 앞서 데이터의 분포가 정규분포인지 아닌지를 확인하는 작업이 우선되어야 한다.
- 정규분포 여부를 판단하는 검정방법
- 귀무가설 : 정규분포를 따른다.
- 대립가설 : 정규분포를 따르지 않는다.

가설검정 절차

- 가설 설정
- 유의수준 설정
- 검정 통계량 산출
- 기각/채택 판단

A 학원은 성적 향상에 도움이 됐을까?

```
# 학원에 다니기 전의 학생 점수
before_study = [34,76,76,63,73,75,67,78,81,53,
                58,81,77,80,43,65,76,63,54,64,
                85,54,70,71,71,55,40,78,76,100,
                51,93,64,42,63,61,82,67,98,59,
                63,84,50,67,80,83,66,86,57,48]

# 학원에 다닌 후의 학생 점수
after_study = [74,87,89,98,65,82,70,70,70,84,
               56,76,72,69,73,61,83,82,89,75,
               48,72,80,66,82,71,49,54,70,65,
               74,63,65,101,82,75,62,83,90,76
               87,90,78,63,59,79,74,65,77,74]
```

Scikit-learn

Scikit-learn

- Scikit-learn은 머신러닝을 위한 파이썬 패키지
- Sample dataset, Data preprocessing 기능, Supervised learning, Unsupervised learning, 모델 평가 기능 등을 담고 있다.
- Scikit-learn의 특징은 다양한 머신러닝 알고리즘을 하나의 패키지 안에서 모두 제공해준다는 점이다.
- Scikit-learn을 사용하기위해서는 원하는 기능의 클래스 객체를 생성해야 한다.
- Scikit-learn이 가진 대표적인 클래스는 자료변환을 위한 Transformer클래스, 회귀분석, 분류, 클러스터링을 위한 Regressor, Classifier, Cluster클래스, 여러 개의 전처리와 모델을 연결하여 하나의 모델처럼 활용하게 하는 Pipeline 클래스 등이 있다
- 공식사이트
  - http://scikit-learn.org
  - http://scikit-learn.org/stable/documentation : 머신러닝 알고리즘 설명 문서
  - http://scikit-learn.org/stable/user_guide.html : scikit-learn 사용자 가이드

Scikit-learn 패키지에서 제공하는 머신러닝 알고리즘

Supervised learning	Unsupervised learning
Generalized Linear Models	Gaussian mixture models
Linear and Quadratic Discriminant Analysis	Manifold learning
Kernel ridge regression	Clustering
Support Vector Machines	Biclustering
Stochastic Gradient Descent	Decomposing signals in components (matrix factorization problems)
Nearest Neighbors	Covariance estimation
Gaussian Processes	Novelty and Outlier Detection
Cross decomposition	Density Estimation
Naive Bayes	Neural network models (unsupervised)
Decision Trees	
Ensemble methods	
Multiclass and multilabel algorithms	
Feature selection	
Semi-Supervised	
Isotonic regression	
Probability calibration	
Neural network models (supervised)	

Scikit-learn 샘플 데이터 사용법과 전처리

- Scikit-learn의 서브패키지 sklearn.datasets는 실습을 위한 샘플용 dataset을 제공하고 있다.
- 샘플용 Dataset은 기본적으로 Scikit-learn 패키지안에 내장되어 있는 형태(load명령으로 import), 인터넷에서 다운로드하여 사용하는 형태(fetch명령으로 import), 그리고 새로운 dataset을 생성시켜 사용하는 형태(make명령으로 생성)로 접근할 수 있다.
- Scikit-learn 패키지에는 데이터 전처리를 위한 preprocessing, feature_extraction 서브 패키지가 있다. 패키지를 활용하여 스케일링(Scaling), 인코딩(Encoding), 결측값 처리(Imputation)가 가능하다.

Scikit-learn 샘플 데이터 소개

[표2] sklearn 샘플 Dataset

load 계열	fetch 계열	make 계열
load_boston() : 보스턴 집값 데이터	fetch_covtype() : 토지 조사 데이터	make_regression() : regression용 데이터 생성
load_diabetes() : 당뇨병 관련 데이터	fetch_20newsgroups() : 뉴스 텍스트 데이터	make_classification() : classification용 데이터 생성
load_iris() : iris 데이터	fetch_rcv1() : 로이터 뉴스 말뭉치	make_blobs() : clustering용 데이터 생성
	fetch_california_housing : 주택 데이터	

```
# sklearn 샘플 Dataset 객체는 다음과 같은 속성으로 구성되어 있다.
# data: 독립 변수의 ndarray 배열 형태
# target: 종속 변수의 ndarray 배열 형태
# feature_names: 독립 변수 이름의 리스트 형태
# target_names: 종속 변수 이름의 리스트 형태
```