

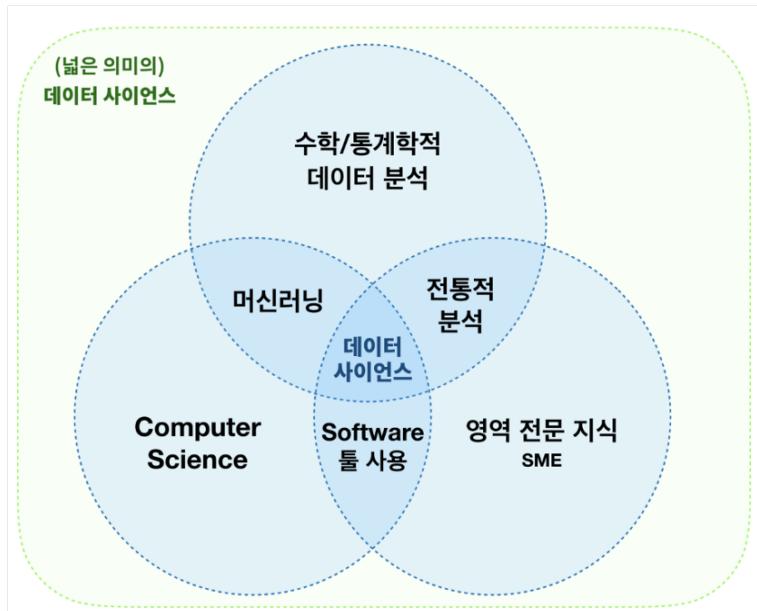
# **비즈니스를 위한 데이터마이닝**

가톨릭대학교 경영학과

이홍주

# 데이터 사이언스 Data Science

- 정형(structured), 비정형(unstructured) 형태를 포함한 다양한 데이터로부터 지식과 패턴, 인사이트(insight)를 추출하는 과학적 방법론이며, 통계적 분석 프로세스, 기계학습/ 딥러닝 알고리즘, 데이터 분석 시스템 등을 데이터 분석에 동원하는 융합분야 (위키피디아)
- 데이터 마이닝 Data Mining

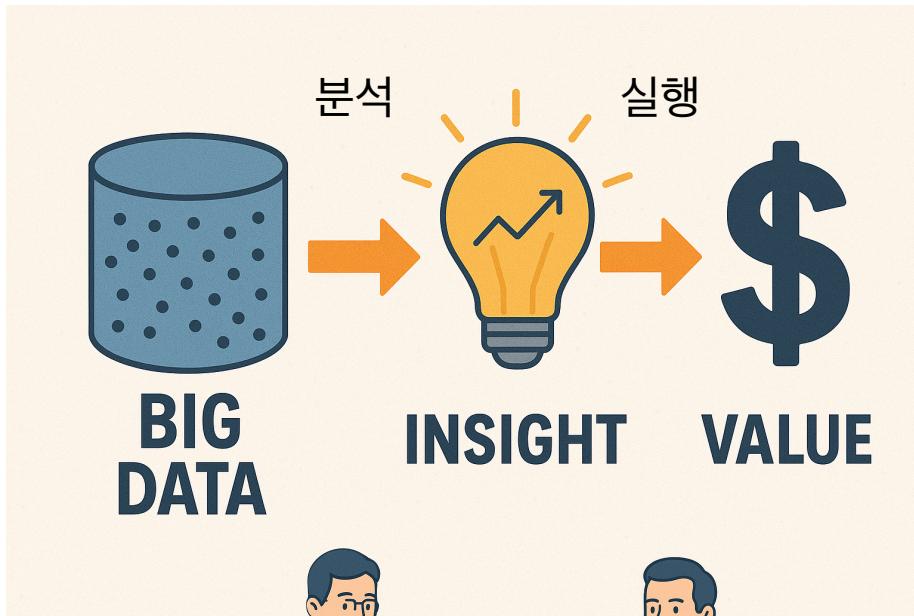


[https://ko.wikipedia.org/wiki/데이터\\_사이언스](https://ko.wikipedia.org/wiki/데이터_사이언스)

# 비즈니스 애널리틱스 Business Analytics

- 비즈니스 인텔리전스 Business Intelligence
  - 과거 상황과 현재 상황을 이해하기 위한 데이터 시각화 및 보고를 위해 차트, 표, 대시보드를 사용해 데이터를 표현, 검사, 탐색하는 방식으로 수행
- 비즈니스 애널리틱스
  - 비즈니스 인텔리전스뿐만 아니라 복잡한 데이터 분석 방법을 포함하는 용어
  - 복잡한 데이터 분석 방법의 예가 통계 모델과 데이터 마이닝 알고리즘으로, 데이터를 탐색하고, 변수 관계를 측정하거나 설명하며, 변수 값을 예측

# 데이터 과학자 Data Scientist

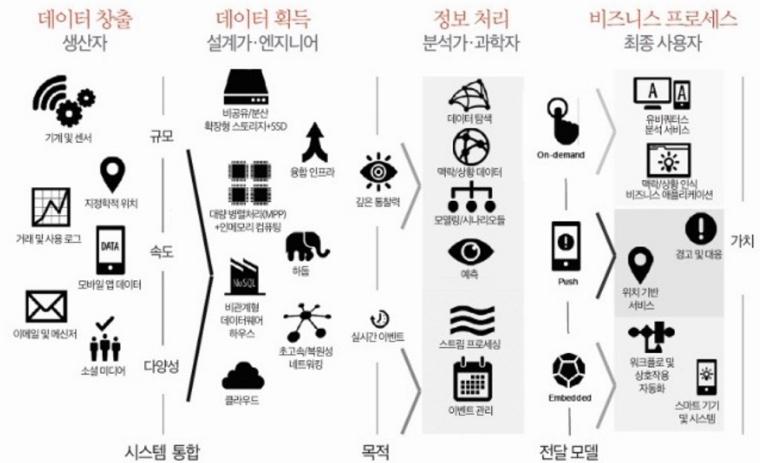


데이터 과학자 Data Scientist  
- 애널리틱스  
(통계, 시각화, ML, AI 등)

의사결정자 Decision Maker  
(마케터, 투자담당자,  
인사담당자, 엔지니어 등)

# 데이터 팀

- 데이터 엔지니어
  - 큰 규모의 데이터베이스 관리
  - 데이터 파이프라인 관리
  - 클라우드를 활용한 인프라 구축 및 관리
  - 실시간 A/B 테스트 실시
- 데이터 분석가
- 데이터 과학자



자료: IDC (2014). "Building a Datacenter Infrastructure to Support Your Big Data Plans"

# 데이터 팀

- 데이터 분석가
  - 주요 지표 측정 및 모니터링
  - 주요 지표 대쉬보드 관리
  - 내부 직원들의 데이터 분석 요청 응대
    - 임원들의 의사결정 도움
    - 간단하지만 많은 데이터로부터 답을 찾는 문제를 많이 다룸
- 주요 도구
  - SQL
  - Python/R//SAS/Matlab
  - 시각화 도구(Tableau, Microsoft Power BI)

# 데이터 팀

- 데이터 과학자
  - 알고리즘을 통해 고객 경험 개선
  - 제품 품질 향상을 위한 예측 모델 구축
  - 지속적인 성과 개선과 문제 설정
  - 주요 도구
    - 기계학습, 딥러닝
    - Python, Spark
    - SQL
    - R/SAS/Matlab

# 황금기 Golden Era

[Subscribe](#)[Sign In](#)[Latest](#)   [Magazine](#)   [Topics](#)   [Podcasts](#)   [Store](#)   [Data & Visuals](#)   [Case Selections](#)   [HBR Executive](#)

Analytics And Data Science

## Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)



# Data Science in Golden Era

## The Old Data Scientist Role Is Becoming Obsolete

From 2014 to 2020, companies hired “data scientists” to do *everything*:

- Clean messy data
- Build dashboards
- Write machine learning models
- Predict customer churn
- Build PowerPoint decks for leadership

It was a golden era. If you knew Python, pandas, and a bit of SQL, you were in.

# Data Science is Dead ?

[Subscribe](#)[Sign In](#)[Latest](#) [Magazine](#) [Topics](#) [Podcasts](#) [Store](#) [Data & Visuals](#) [Case Selections](#) [HBR Executive](#)[Analytics And Data Science](#)

## Is Data Scientist Still the Sexiest Job of the 21st Century?

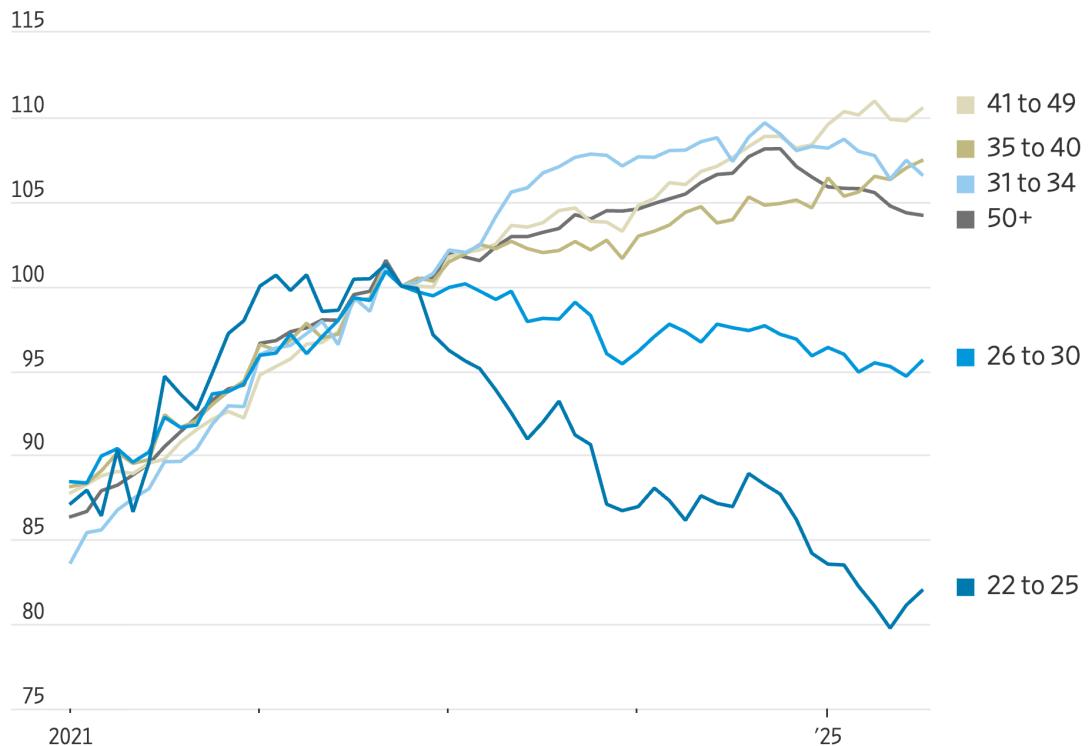
by Thomas H. Davenport and DJ Patil

July 15, 2022



# Data Science is Dead ?

Employee headcount among software developers, by age



Note: Indexed to 100 at October 2022

Source: Brynjolfsson, Chandar and Chen

# Data Science is Evolving

Old Role	New Roles in 2025
Data cleaning & ETL	<b>Analytics Engineer</b> (SQL + dbt + Airflow)
Business dashboards	<b>BI Developer</b> (Power BI, Tableau, Looker)
Product experimentation	<b>Data Analyst</b> (SQL + experimentation)
ML models	<b>ML Engineer</b> (MLOps, pipelines, deployment)
Causal inference & forecasting	<b>Research Scientist</b> (Stats + PhD-level modeling)

## What to Call Yourself Instead of “Data Scientist”

Here's what top companies are hiring for now:

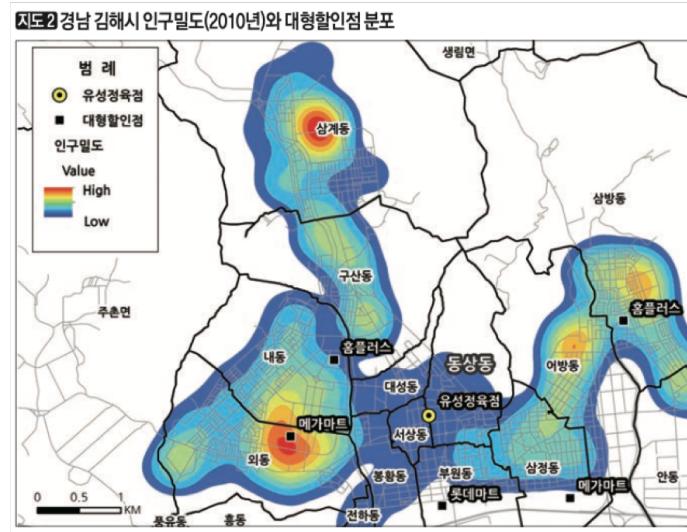
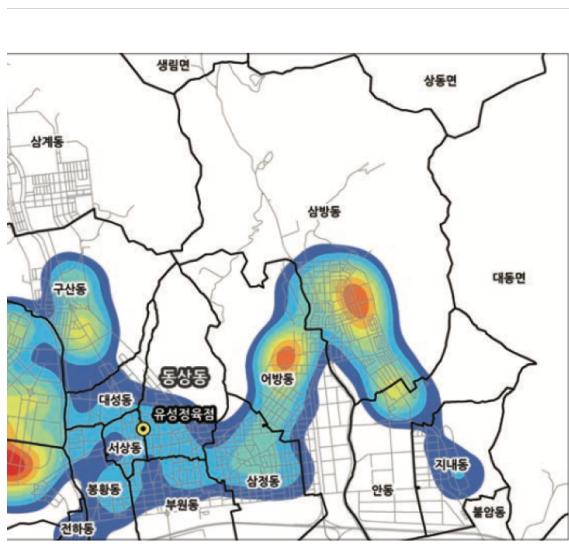
- Product Data Analyst
- Analytics Engineer
- Decision Scientist
- Machine Learning Engineer
- Quantitative Researcher
- Marketing Analyst
- Data Product Manager

# 데이터 분석의 주요 기능

- 데이터 탐색
- 시각화
- 분류
- 예측
- 데이터 및 차원 축소
- 연관규칙

# 데이터 탐색과 시각화

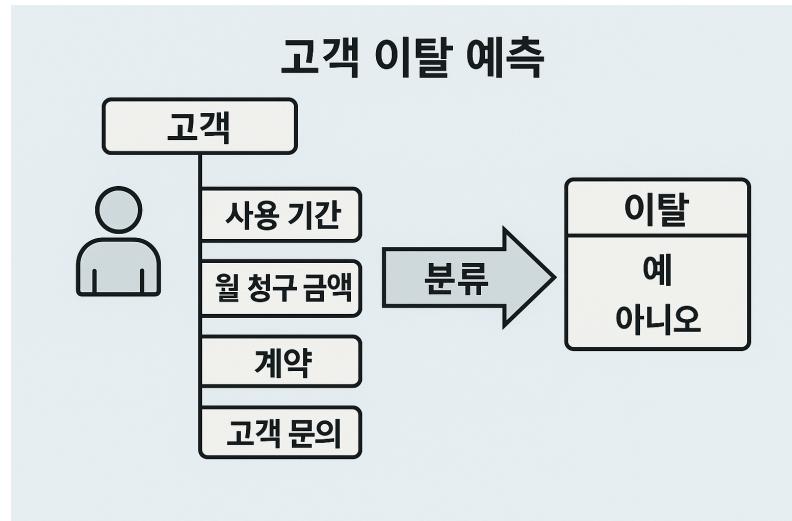
- 빈도 및 평균 등 기술통계를 중심으로 해당 집단의 특성을 요약 제공
- 데이터 이해를 위한 시각화 포함



(출처: 데이터, 아는 만큼 보인다. 조정래의 글쓰기에서 '분석비법' 배우자, 송규봉, 동아비즈니스리뷰, No. 148, 2014)

# 분류

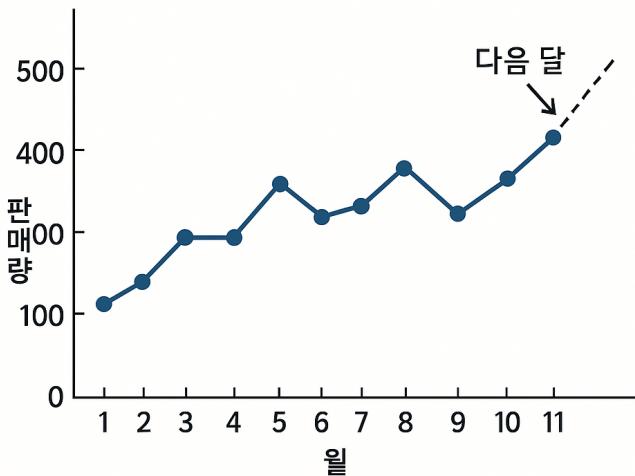
- 서로 다른 범주(category or class)에 속하는 데이터의 예가 주어지면 그 데이터의 속성을 사용하여 모델링하고, 이를 통해 새로운 데이터의 범주를 분류
- 고객 이탈 분류: 이탈 여부(예/아니오)



# 예측

- 과거와 현재 정보를 가지고 미래 시점에 대한 추정치를 예측
- 숫자로 표현된 연속형 변수의 미래 값 예측
  - 예) 판매, 수익, 성과
  - 판매 예측: 과거 12개월간의 판매 데이터를 기반으로 다음 달 아이스크림 판매량을 예측하세요.
  - 성과 예측: 광고비를 얼마 투자하면 매출이 얼마나 증가할지 예측하세요.

## 아이스크림 판매량 예측



과거 12개월간의 판매 데이터를 기반으로 다음 달 아이스크림 판매량을 예측하세요.

# 데이터 및 차원 축소

- 유사한 속성들을 갖는 데이터들을 묶어 전체 데이터들을 몇 개의 집합으로 나누거나 많은 속성들 중 유사한 속성들끼리 묶어 속성의 차원 축소
  - 예) 웹사이트 방문자 행동 분석: 방문자의 행동 패턴에 따라 그룹을 나눔
  - 입력 데이터: 페이지 체류 시간, 클릭 경로, 검색어, 장바구니 사용 여부 등
  - 활용: UX 개선, 추천 콘텐츠 제공, 이탈 방지 전략 수립
- 차원축소
  - 예) 고객당 30개가 넘는 변수(연령, 소득, 구매이력, 반품률, 웹사이트 방문 패턴 등)가 있을 때
  - 활용: 차원축소 기법(PCA 등)을 사용하여 2~3개의 주성분으로 줄여서 고객 세분화 결과를 시각화하거나 군집화 전처리로 사용
  - 고객 유형을 한눈에 파악 가능

# 연관규칙

- 데이터에 숨어있는 항목간의 연관도를 측정하여, 함께 구매하는 물건이나 시간차를 두고 같이 구매되는 물건 등의 연관 규칙을 발견함

Feb 16, 2012, 11:02am EST

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill Former Staff  
Tech  
*Welcome to The Not-So Private Parts where technology & privacy collide*

---

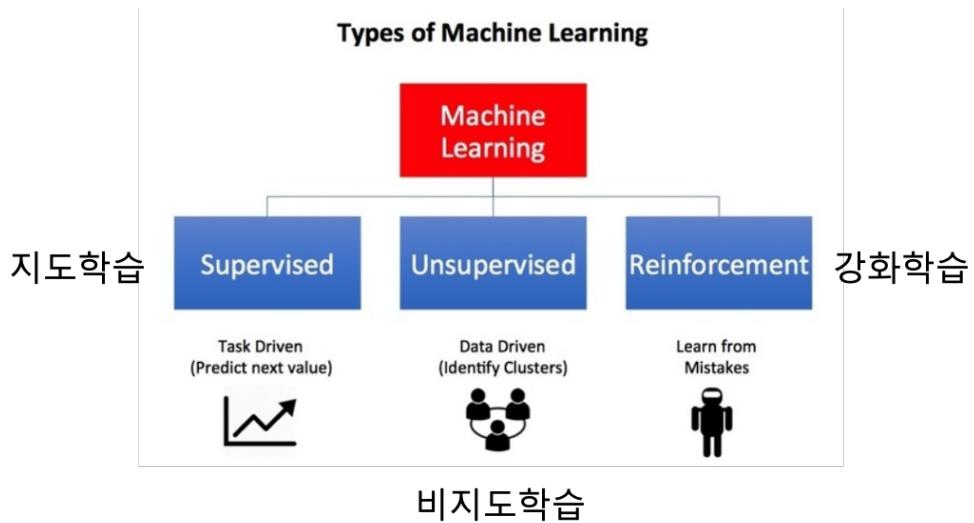
⌚ This article is more than 8 years old.

f Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those in retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.



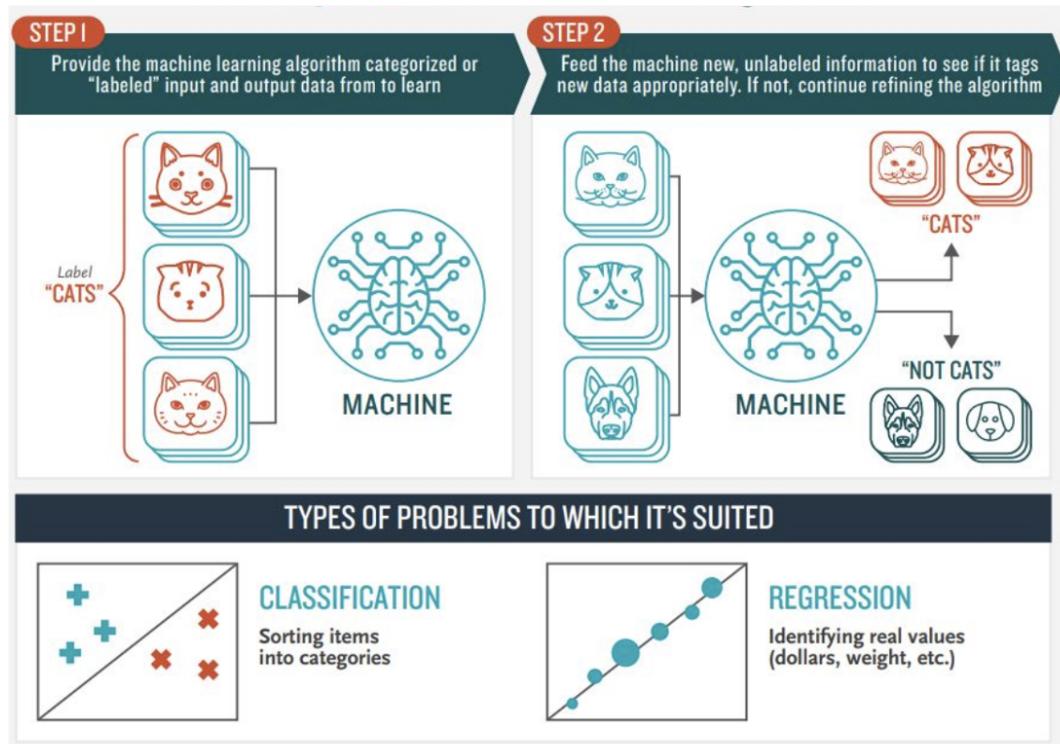
**TARGET**

# 기계학습 Machine Learning 의 분류



(출처: <https://www.stoodnt.com/blog/best-online-courses-on-machine-learning-deep-learning-ai-and-big-data-analytics>)

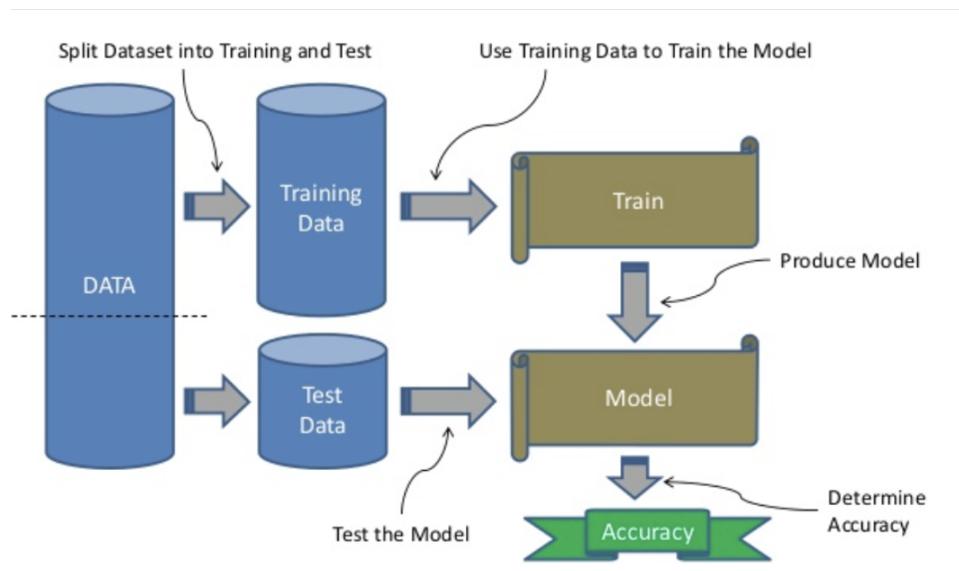
# 지도학습 Supervised Learning



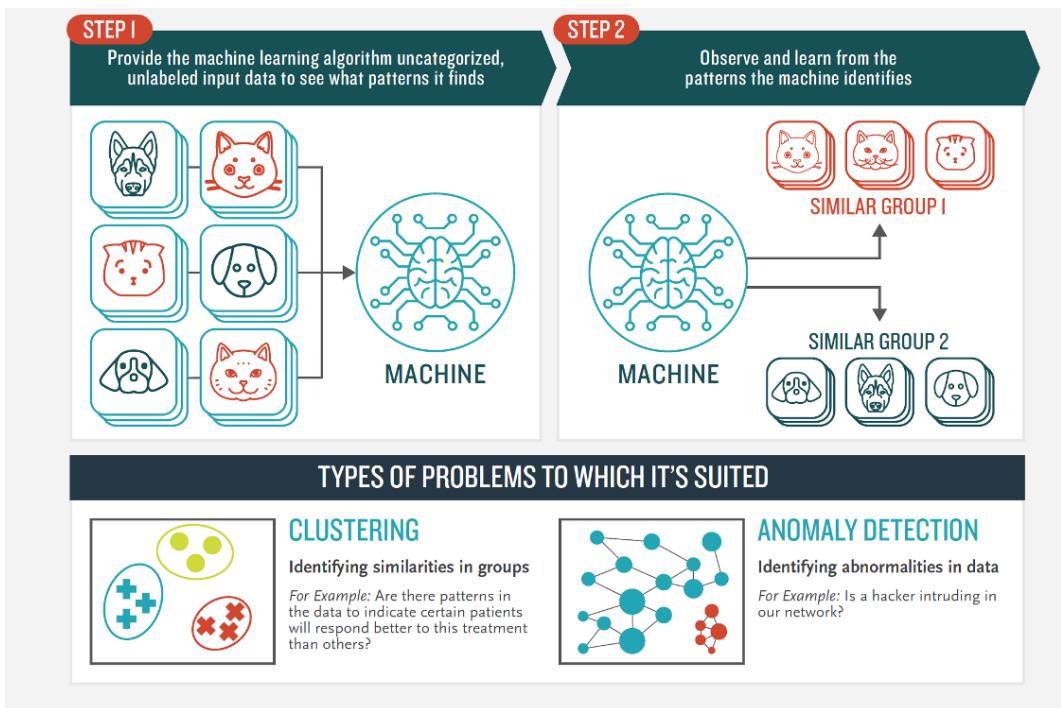
(출처: <https://okanbulut.github.io/bigdata/modeling-big-data.html>)

# 지도학습

- 데이터 분할: 학습 데이터 / 테스트 데이터



# 비지도 학습 Unsupervised Learning



(출처: <https://okanbulut.github.io/bigdata/modeling-big-data.html>)

# 강화 학습 Reinforcement Learning

- 에이전트 스스로 학습
  - 예) 재고 및 공급망 관리: 유통/물류 기업의 자동 발주 시스템
  - 주문 시점과 수량을 강화학습 기반으로 결정 → 재고 과잉이나 부족 최소화
  - 보상: 비용 절감, 납기 준수율 향상 등



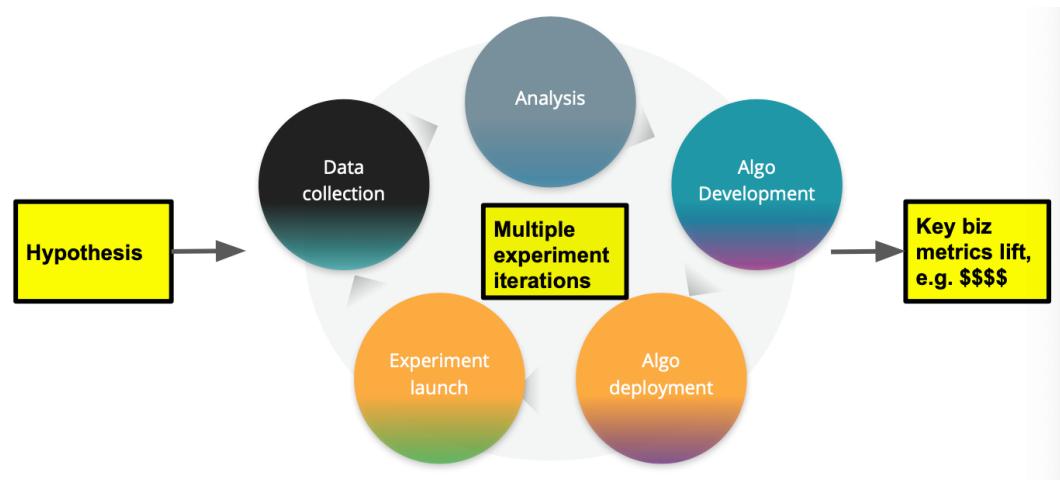
# 데이터 마이닝 수행 단계

- SEMMA 단계(SAS 개발)
  - Sample(추출): 데이터셋을 학습/검증/테스트 데이터셋으로 나눈다.
  - Explore(탐색): 데이터셋을 통계적 혹은 시각적으로 분석한다.
  - Modify(수정): 변수를 변환하고 결측치를 처리한다.
  - Model(모델링): 예측 모델을 구축한다(의사결정 트리, 인공 신경망).
  - Assess(평가): 검증 데이터셋을 이용해 후보 모델을 비교한다.
- CRISP-DM(IBM SPSS Modeler)
  - 비즈니스 이해
  - 데이터 이해
  - 데이터 준비
  - 모델링
  - 평가
  - 배포

# 데이터 마이닝 수행 단계

- ① 데이터 마이닝 프로젝트의 목적 정확히 설정하기
- ② 분석에 필요한 데이터셋 획득하기
- ③ 데이터의 탐색/정제/전처리하기
- ④ 필요시 데이터 축소하기
- ⑤ 데이터 마이닝 문제 결정하기(분류, 예측, 군집 등)
- ⑥ 데이터 분할하기(지도 학습의 경우)
- ⑦ 사용할 데이터 마이닝 기법 선택하기(회귀 분석, 인공 신경망, 계층 군집 등)
- ⑧ 알고리즘을 사용해 과제 수행하기
- ⑨ 알고리즘 결과 해석하기
- ⑩ 모델 적용하기

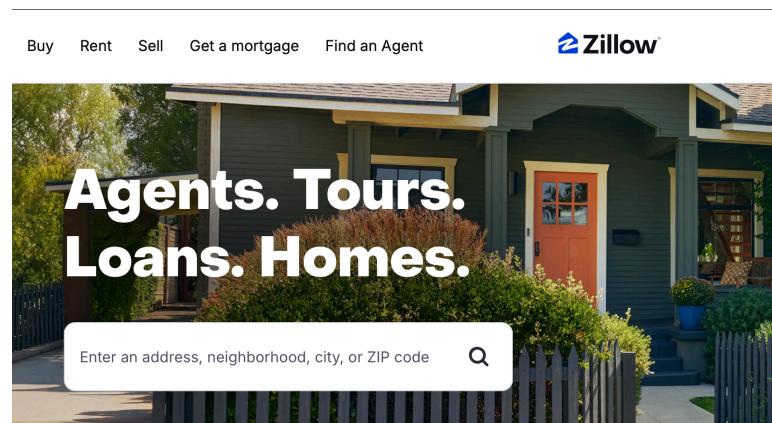
# 데이터 마이닝 수행 단계



출처: <https://arxiv.org/pdf/1602.05142.pdf>

# 주택 가치 예측 사례

- Zillow – 부동산 중개 업체
- 주택 데이터 기반 주택 가격 예측
- 국내 유사 업체 XAI Land (<https://xai.land/>) – 아파트 가격 예측
  - XAI Land는 근처 유사 매물의 가격에 기반하여 가격 예측



① 데이터 마이닝 프로젝트의 목적 정확히 설정하기

# 주택 가치 예측 사례

- 웨스트 록스베리 데이터
- 보스턴시에서 공개한 지역 부동산 데이터

표 2-1 웨스트 록스베리(보스턴 지역) 주택 가격 데이터 변수 설명

TOTAL VALUE	주택 가격(단위: 1,000달러)
TAX	세금, 주택 가격에 세율을 곱한 값에 근거한 세금 계산서 금액(단위: 달러)
LOT SQFT	총 부지 면적(단위: 제곱피트)
YR BUILT	건축 연도
GROSS AREA	총 바닥 면적
LIVING AREA	주거 공간 총 면적(단위: 제곱피트)
FLOORS	층 수
ROOMS	총 방 수
BEDROOMS	총 침실 수
FULL BATH	총 욕실 개수
HALF BATH	총 보조 욕실 개수
KITCHEN	총 주방 개수
FIREPLACE	총 벽난로 개수
REMODEL	리모델링 시기(최근/오래 전/안 함)

## ② 분석에 필요한 데이터셋 획득하기

# 주택 가치 예측 사례

- 데이터 구조 파악
  - 데이터 열(column) 이름 변경
  - 데이터 샘플링
  - 데이터 전처리와 정제
    - 변수 종류 변환
    - 변수 선택
    - 이상치 탐지
    - 결측치 탐지 및 처리
    - 데이터 표준화
- ③ 데이터의 탐색/정제/전처리하기  
④ 필요시 데이터 축소하기

# 분류 문제에서 희소사건에 대한 오버샘플링

- 관심 데이터가 희귀할 경우(메일에 응답하여 특정 상품을 구매하는 고객, 사기 카드 거래) 데이터 부족으로 모델 구축이 어려움
- 샘플링 시 소수 클래스에 더 큰 가중치를 주어 불균형 해결 또는 각 클래스의 오분류에 가중치를 주어 해결
- 다수 클래스와 소수 클래스의 불균형을 고려하지 않은 모델의 경우 전반적인 정확도는 높을 수 있으나 실제 문제에서 적용하기 어려움
- 모델 구축 시 클래스별 오분류의 중요도를 고려해 가중치를 다르게 부여해서 보다 중요한 클래스의 오분류를 줄이는 것이 중요

# 데이터 전처리와 데이터 정제

- 변수 종류
  - 변수는 여러 기준으로 분류 가능
  - 일반적인 분류: 숫자 혹은 범주형 변수
  - 숫자: 연속형 실수 혹은 정수
  - 범주형: 숫자 (0/1/2 등) 혹은 문자로 표현
    - 북아메리카, 유럽, 아시아와 같이 특별히 순위가 없는 경우를 명목형 변수
    - 큰 값, 작은 값 등 순위로 표현할 수 있는 경우를 순서형 변수
  - 데이터 마이닝에서는 나이브 베이즈 분류기(Naive Bayes Classifier)와 같이 범주형 변수만 사용하는 특별한 경우를 제외하면 범주형 변수를 이진 더미(dummy) 변수로 변환하여 사용 (One hot encoding이라고도 함)

# 데이터 전처리와 데이터 정제

- `housing_df2 = pd.get_dummies(housing_df, prefix_sep='_', drop_first=True)`

	REMODEL_Old	REMODEL_Recent
0	0	0
1	0	1
2	0	0
3	0	0
4	0	0

- `housing_df3 = pd.get_dummies(housing_df, prefix_sep='_')`

	REMODEL_None	REMODEL_Old	REMODEL_Recent
0	1	0	0
1	0	0	1
2	1	0	0
3	1	0	0
4	1	0	0

# 변수 선택

- TAX는 TOTAL VALUE 에 세율을 곱해서 정해짐
- 아직 모르는 TOTAL VALUE를 예측하는 문제에서 TOTAL VALUE에 의해  
서 결정되는 TAX가 입력변수가 될 수 없음
- 사람의 몸무게를 예측하는데 입력변수로 키와 BMI 지수를 입력하면될까?
- 데이터에 포함된 많은 변수들 중에서 분류나 예측 정확도에 기여하지 못하  
는 변수들도 많음

# 이상치 outlier 탐지

- 데이터 형태와 부합하지 않은 데이터가 포함된 경우
  - LOT\_SIZE (대지 면적) 컬럼에 문자열 "unknown"이나 "N/A"가 포함된 경우 → 이 컬럼은 원래 숫자(float) 형태여야 하므로 부적절한 데이터
- 데이터가 가질 수 없는 범위의 값을 가진 경우
  - AGE 컬럼에 음수 값 -5가 입력된 경우 → 나이는 0 이상의 정수여야 하므로 데이터 타입은 맞더라도 의미상 부합하지 않음
  - ROOMS (전체 방 개수)보다 BEDRMS 값이 더 큰 경우 → ROOMS=5인데 BEDRMS=6이면 말이 되지 않음
- 데이터가 가질 수 있는 일반적인 범위의 값보다 매우 크거나 작은 값
  - LOT\_SIZE가 100,000 (평방 피트) 이상인 경우 → 대다수 주택은 5,000~10,000 사이인데, 10배 이상의 값은 극단적인 대형 부지
  - TAX가 20,000 이상인 경우 → 대부분 세금이 1,000~5,000 사이에 분포한다면 이는 이례적으로 높은 세금

# 결측치 탐지 및 처리

- 다양한 이유로 데이터에 결측치 missing value 존재
- 정보 누락, 응답 회피, 여러 파일을 연결하면서 오류로 인한 누락 등
- 전체 데이터에서 결측치가 포함된 행의 수를 파악하여 비율이 문제가 안되는 경우에는, 결측치가 포함된 행 제거
- 전체 데이터에서 결측치를 포함한 행의 비율이 높은 경우
  - 결측치를 많이 포함한 변수의 중요도 파악, 중요하지 않은 변수가 결측치를 많이 포함하고 있다면 변수 제거
  - 결측치를 많이 포함한 변수의 중요도 파악, 중요한 변수가 결측치를 많이 포함하고 있다면 결측치를 다른 값으로 대체 (중앙값, 평균 등)

# 데이터 표준화 data standardization

- 변수가 갖는 수치 값의 범위가 서로 다르기 때문에, 범위가 큰 변수의 영향력이 높을 수 있음
  - LOT SQFT는 BEDRMS 보다 값의 범위가 큼
- 범위의 차이로 인한 영향력을 없애기 위해서 표준화 시행
  - StandardScaler( ):  $N(0,1)$  표준 정규분포로 변환 
$$z = \frac{x - \mu}{\sigma}$$
  - MinMaxScaler( ): 
$$\frac{x - \min}{\max - \min}$$

0 <

< 1

# 데이터 분할

- 학습 데이터
- 검증 데이터
  - 학습과정에서 구축된 모형의 성능 평가(중간 시험)
  - 파라미터 튜닝, 조기 종료 필요한 알고리즘이 활용
- 테스트 데이터
- 새 데이터
  - ⑤ 데이터 마이닝 문제 결정하기(분류, 예측, 군집 등)
  - ⑥ 데이터 분할하기(지도 학습의 경우)



그림 2-4 데이터 마이닝 프로세스에서 학습/검증/테스트 데이터의 역할

# 주택 가치 예측 사례

```
model = LinearRegression()
```

```
model.fit(train_X, train_y)
```

Mean Error (ME): 0.1463

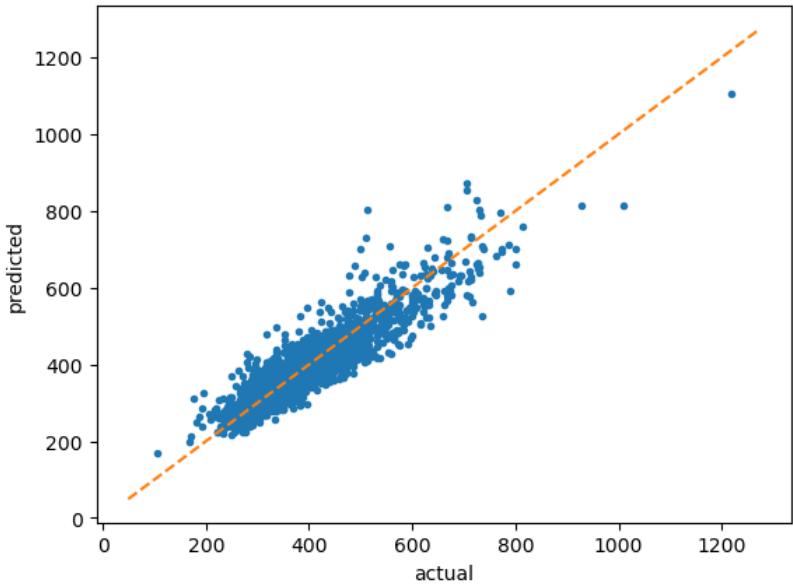
Root Mean Squared Error (RMSE): 42.7292

Mean Absolute Error (MAE): 31.9663

Mean Percentage Error (MPE): 1.09%

Mean Absolute Percentage Error (MAPE): 8.33%

R-squared: 0.8171



- ⑦ 사용할 데이터 마이닝 기법 선택하기(회귀 분석, 인공 신경망, 계층 군집 등)
- ⑧ 알고리즘을 사용해 과제 수행하기
- ⑨ 알고리즘 결과 해석하기
- ⑩ 모델 적용하기