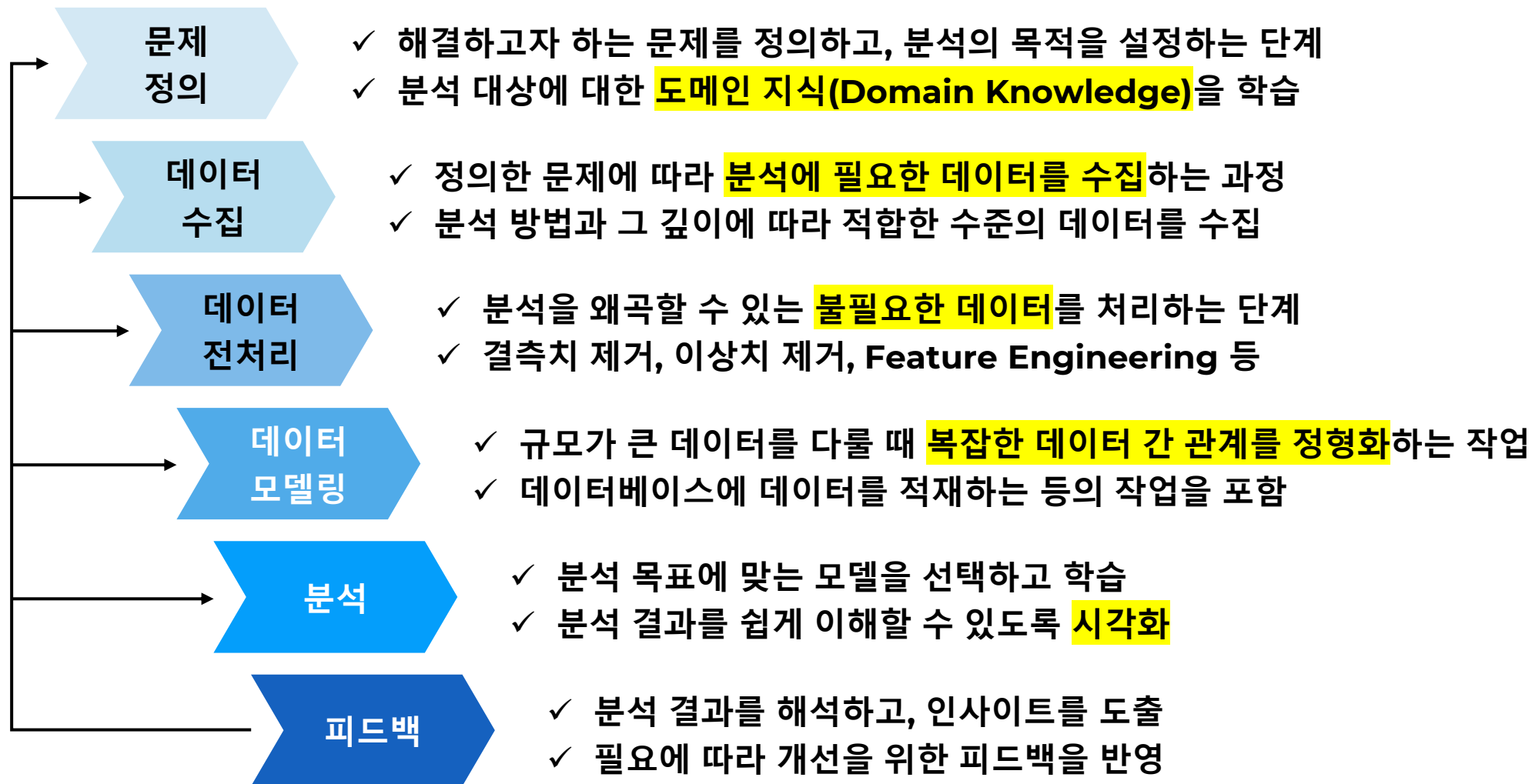


데이터 분석사례

Contents

1. 데이터 분석 프로세스
2. AI 데이터 분석 단계
3. 다양한 데이터(정형, 비정형) 분석 사례
4. 분석모형의 활용
5. 데이터 안심구역에서의 분석활용

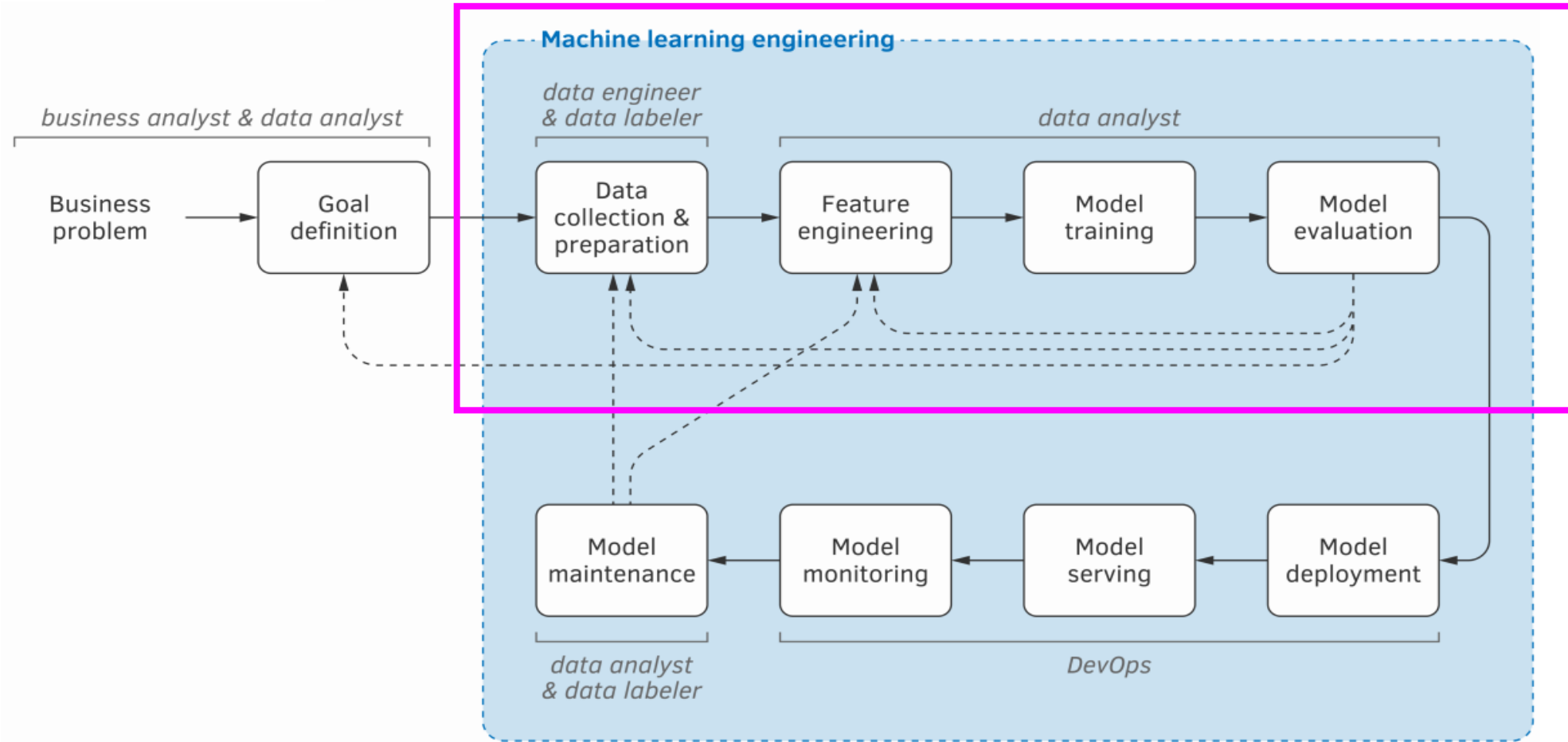
데이터분석 프로세스



1. 데이터분석 프로세스

3

데이터분석 프로세스



ref: Machine Learning Engineering (Burkov, 2020, p17 in chap1)

데이터분석 단계별 고려사항

문제 정의

- 여러 가지 제약사항을 고려한 문제 정의(데이터 수집, 자원 부족 등)
- 도메인 지식의 중요성은 생각보다 훨씬 높음
- 개발자는 클라이언트의 모호한 요청을 개발적인 언어로 해석할 수 있는 능력이 필요

데이터 수집

- 문제 정의 단계에서 논의된 사항에 따라 분석 방법을 결정

데이터 전처리

- 가장 많은 시간을 쏟아야 하는 단계
- 도메인 지식을 활용하는 등 꼼꼼하게 전처리를 할 수록 유의미한 분석이 됨

데이터분석 단계별 고려사항

데이터 모델링

- 의미 있는 모델을 만들기 위해 필수적인 단계
- 분석 결과를 개선할 때 이 단계에서의 수정이 들어가는 경우가 있음

분석

- 분석 결과를 보고 가장 적합한 방식으로 결과를 설명
- 데이터 시각화의 중요성은 매우 높음

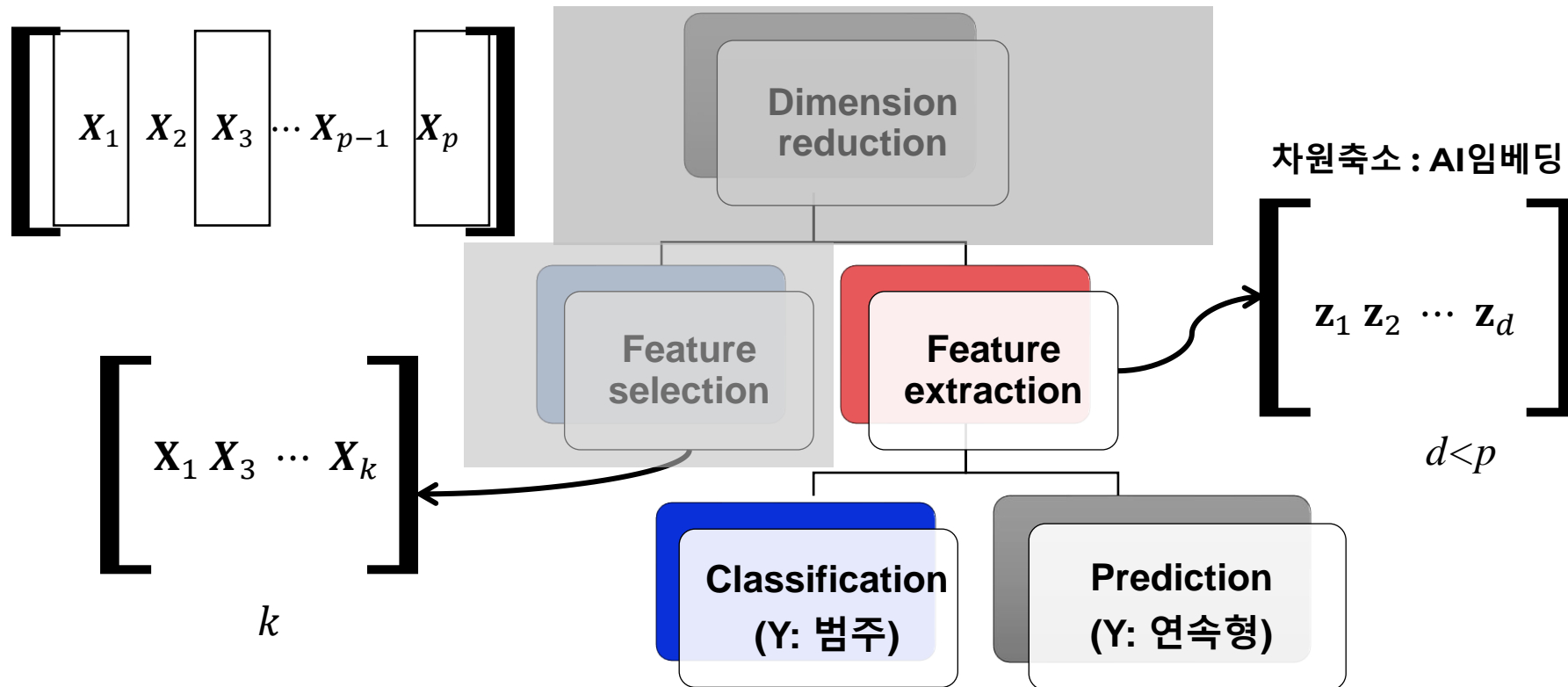
피드백

- 클라이언트에게 분석 결과를 설명하기 위해 데이터 시각화 등의 방법을 활용
- 간결하고 명확한 메시지로, 상대방이 이해할 수 있는 언어로 전달
- 어떤 분야에서 개선이 필요한지 확인하고, 어느 단계로 돌아갈 것인지를 결정

2. AI 데이터 분석단계

6

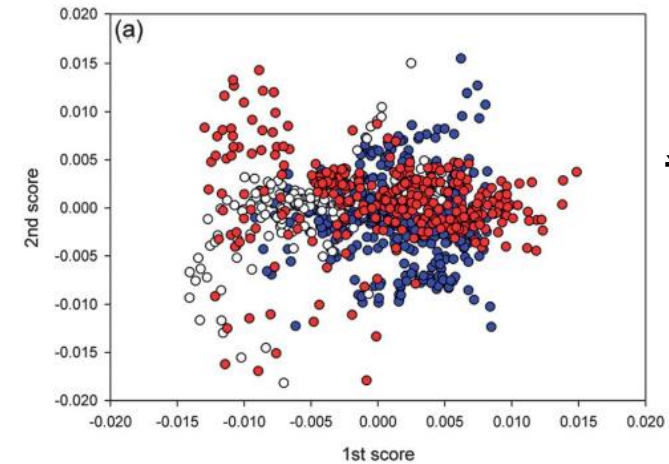
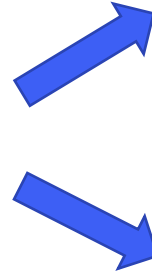
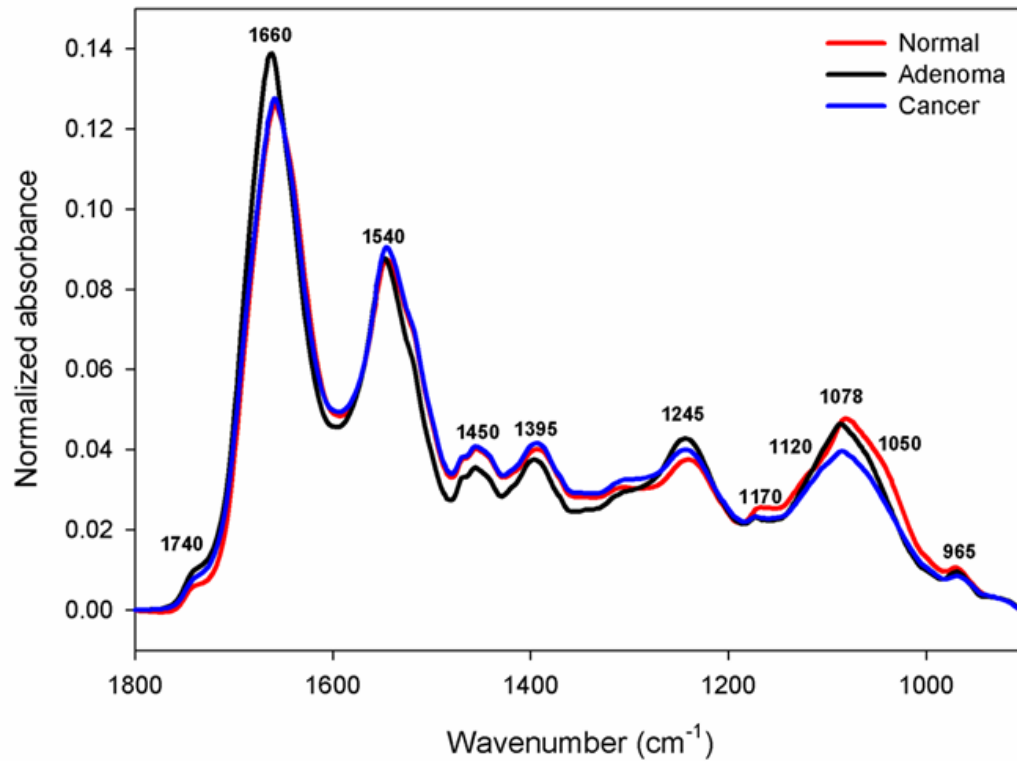
(1) AI 기술 활용의 분석체계



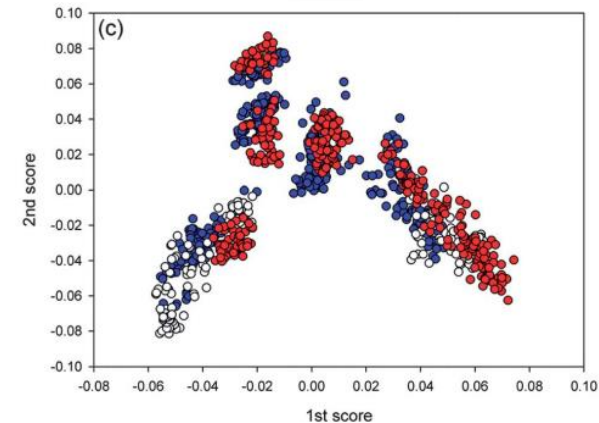
2. AI 데이터 분석단계

(2) AI데이터 분석 예시 : 근적외선 흡광도 데이터를 이용한 분류 (정상/전이단계/비정상)

흡광도 데이터



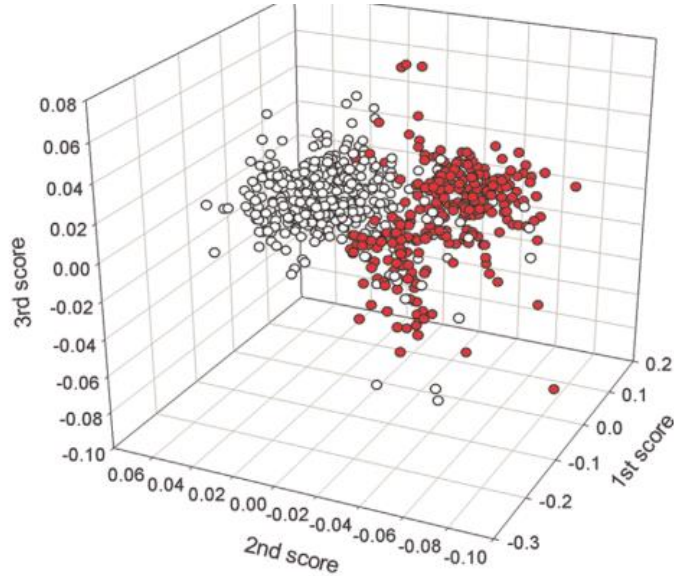
주성분분석



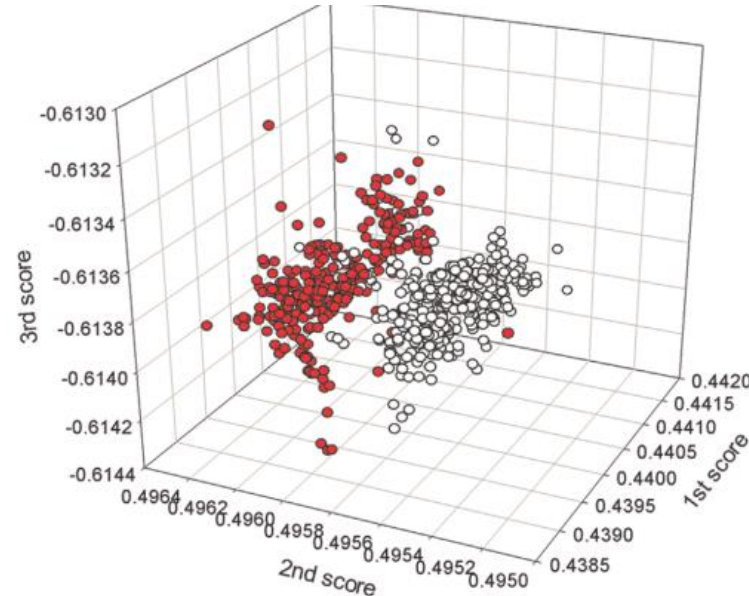
인접보존기법

2. AI 데이터 분석단계

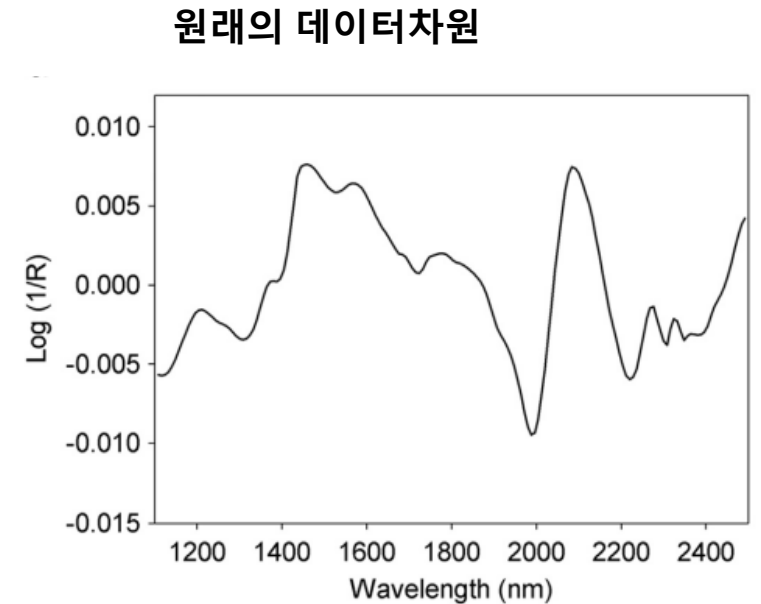
(3) AI데이터 분석 예시 : 근적외선 스펙트럼 데이터의 저차원 시각화 – 농산물(당귀)의 원산지 분류



주성분분석 : 특징1, 특징2, 특징3



인접보전기법 : 특징1, 특징2, 특징3



Ref: *Talanta*, 144, 960–968 (2015)

2. AI 데이터 분석단계

(4) AI데이터 분석 예시 : 캐나다 경전철 지연 문제

- 캐나다 토론토 지역은 기후, 보수 공사 등의 요인으로 경전철 서비스가 지연되는 경우가 잦음
- 미래의 지연을 예측하고, 이를 사람들에게 전달하기 위한 서비스를 배포하는 것이 목적



- 데이터 수집

토론토 교통국(TTC)으로부터 경전철 지연 데이터를 수집(2014-2019)

(4) AI데이터 분석 예시 : 캐나다 경전철 지연 문제

- 데이터 전처리
 - 데이터 병합
 - 여러 개로 나누어진 스프레드시트를 **단일 데이터프레임으로** 변환
 - 누락된 값 고치기 (결측치 확인)
 - 데이터 유형의 문제 수정
 - '지연 시간'로 표시되어야 할 열이 일부 스프레드시트에서 '지연'으로 표시되는 문제
 - 하나로 통일하는 방법을 통해 해결
 - '연-월-일' 형식의 데이터를 연, 월, 일 열로 분리하여 **수치형 데이터로 저장**
 - 경로, 위치, 방향, 차량 열의 잘못된 값 정리
 - 현재 운행하지 않는 경로, 차량이 포함되어 있는 경우가 존재 → 삭제
 - **방향, 위치 데이터가 통일되지 않은 형식으로 존재**
 - 방향의 경우 {eastbound, westbound, southbound, northbound}, {e, w, s, n}, {E, W, S, N} 등으로 존재
 - 위치의 경우 주소를 표현하는 방식이 다양하며, 대문자와 소문자의 사용이 불규칙함

(4) AI데이터 분석 예시 : 캐나다 경전철 지연 문제

- 데이터 모델링

- 데이터 수가 많지 않아 단일 데이터프레임으로 해결 가능

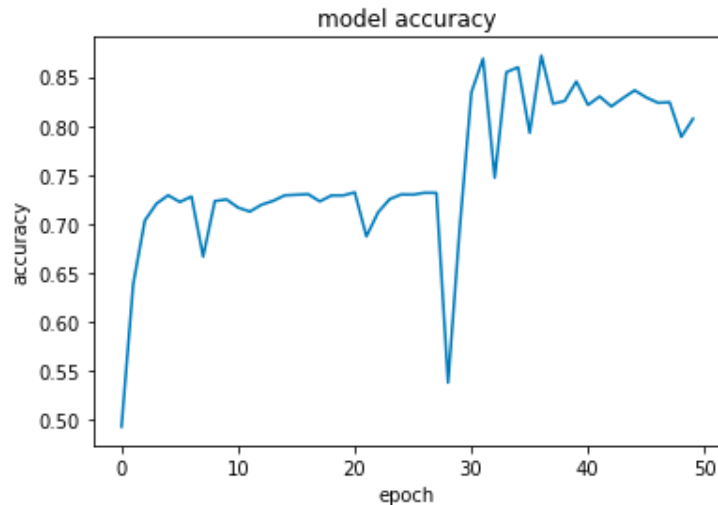
- 분석

- 간단한 딥러닝 모델로 경전철 지연을 예측하는 것이 목표
 - 데이터 형태를 단순화한 후, 타입(수치형, 범주형, 텍스트 등)에 따라 열을 분리
 - 적절한 모델을 설계 후 학습

- 모델의 매개변수(파라미터)를 적절히 조절하여 최대한의 성능을 내도록 함

- 피드백

- 모델의 학습 정도를 시각화하여, 모델의 학습 성공 여부를 판단
 - 최신 데이터를 반영하여 검증과정
 - 도메인 지식을 활용하여 성능을 개선할 수 있을 것으로 기대



3. 다양한 데이터 (정형·비정형) 분석 사례

- 정형 데이터(Structured Data)

- 미리 정해 놓은 형식과 구조에 따라 저장되도록 구성된 데이터
- 일반적으로 행과 열로 구성된 **테이블 형태**로 저장
- 데이터베이스, 스프레드시트와 같은 관계형 데이터베이스에서 관리
- **손쉽게** 데이터에 대한 부분 검색, 선택, 갱신, 삭제 등의 연산을 수행할 수 있음

- 비정형 데이터(Unstructured Data)

- 정의된 구조가 없는 **정형화되지 않은** 데이터
- 동영상, 오디오, 이미지, 텍스트 등 다양한 형태로 존재
- 일반적인 분석 도구나 기술로는 처리가 어려움
- 추가적인 처리를 통해 **정형 데이터로 변환**하는 처리 과정이 필요

3. 다양한 데이터 (정형·비정형) 분석 사례

- 일상의 모든것이 데이터화(Datafication)된 세상



Activity, Invisible

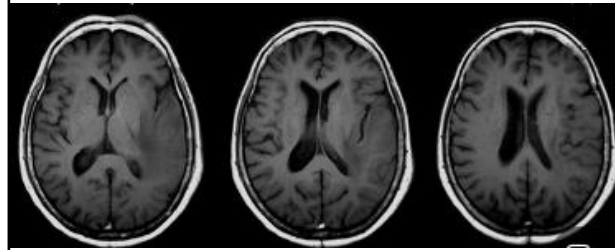


Data화 되는것

(Schoenberger and Cukier, 2014)



이미지(영상)



Data
(숫자, 벡터로 변환)

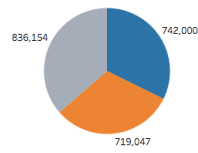
3. 다양한 데이터 (정형·비정형) 분석 사례

(1) 정형 데이터

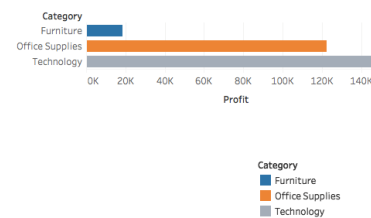
- 손쉽게 데이터를 분석할 수 있음
- 다양한 방식의 분석론과 시각화가 요구됨
- Tableau, PowerBI 등의 tool이 존재

Category Analysis

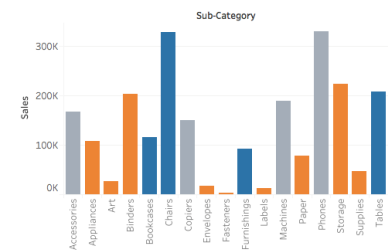
Net Sales Based on Category



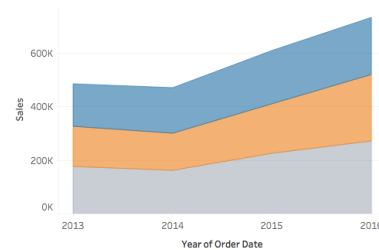
Net Profit



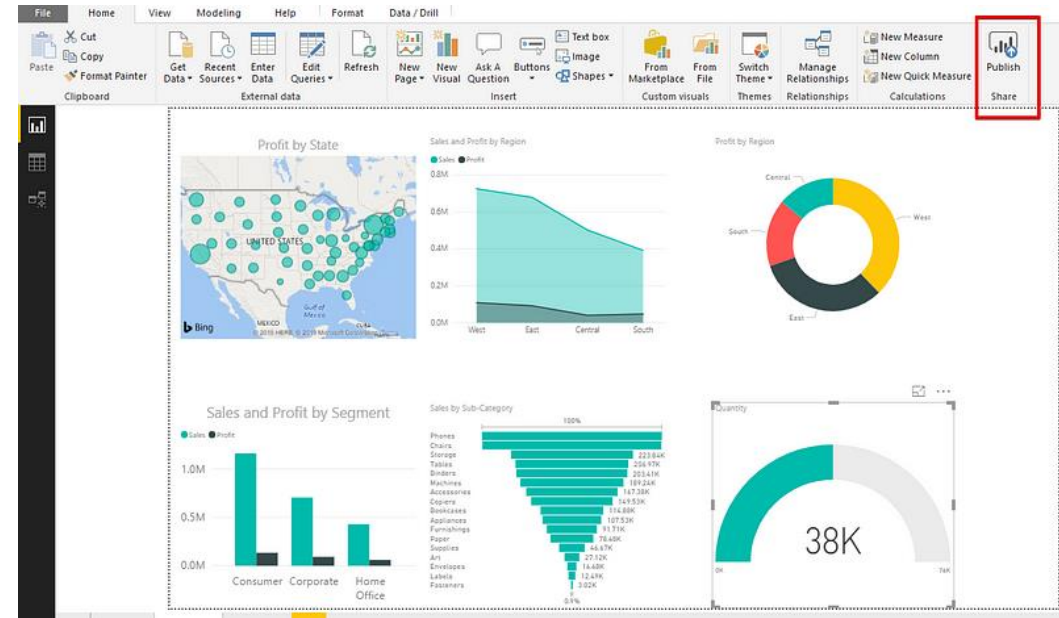
Sales by Sub - Category



Sales Growth over the years



Tableau



PowerBI

3. 다양한 데이터 (정형·비정형) 분석 사례

(1) 정형 데이터 : 주택가격 동향분석 예시

- 10년간(2007~2017) **시 아파트 거래 데이터(from. Kaggle)
- 30개의 설명변수와 5892개의 데이터
 - 가격: 'SalePrice'
 - 준공일자, 거래일자: 'YearBuilt', 'YrSold', 'MonthSold'
 - 규격: 'Size(sqf)', 'Floor'
 - 그 외: 복도 유형, 난방 유형, 판매 유형, 엘리베이터 수, 경비원 수
 - 주변 데이터: 지하철 역 & 버스 정류장까지의 거리, 주변 주차장, 학교, 공공 시설, 백화점, 쇼핑몰 수

가격		일자		규격		그 외			주변 데이터									
SalePrice	YearBuilt	YrSold	MonthSold	Size(sqf)	Floor	HallwayType	HeatingType	AptManageType	N_Parkinglot(Ground)	N_Parkinglot(Basement)	TimeToBusStop	TimeToSubway	N_APT	N_manager	N_elevators	SubwayStation		
141592	2006	2007	8	814	3	terraced	individual_heating	management_in_trust	111		184 5min~10min	10min~15min	3	3	0	Kyungbuk_uni_hospital		
51327	1985	2007	8	587	8	corridor	individual_heating	self_management	80		76 0~5min	5min~10min	1	2	2	Daegu		
48672	1985	2007	8	587	6	corridor	individual_heating	self_management	80		지하철 역과 거리						2	Daegu
380530	2006	2007	8	2056	8	terraced	individual_heating	management_in_trust	249		536 0~5min	15min~20min	6	5	11	Sin-nam		
221238	1993	2007	8	1761	3	mixed	individual_heating	management_in_trust	523		536 0~5min	15min~20min	8	주변 시설 수				
35840	1992	2007	8	355	5	corridor	individual_heating	management_in_trust	200		0 5min~10min	10min~15min	3					
78318	1992	2007	8	644	2	mixed	individual_heating	self_management	142		79 5min~10min	15min~20min	3	4	8	Myung-duk		
61946	1993	2007	8	644	10	mixed	individual_heating	management_in_trust	523		536 0~5min	15min~20min	8	8	20	Myung-duk		
84070	1993	2007	8	644	3	mixed	individual_heating	management_in_trust	523		536 0~5min	15min~20min	8	8	20	Myung-duk		
									142		79 5min~10min	15min~20min	3	4	8	Myung-duk		
									713		0 0~5min	10min~15min	7	8	27	Kyungbuk_uni_hospital		

3. 다양한 데이터 (정형·비정형) 분석 사례

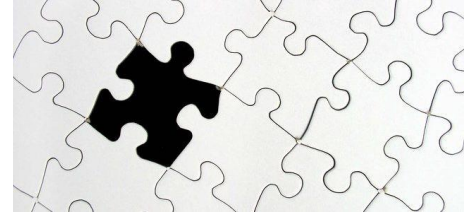
16

(1) 정형 데이터 : 주택가격 동향분석 예시

S1. Check Data

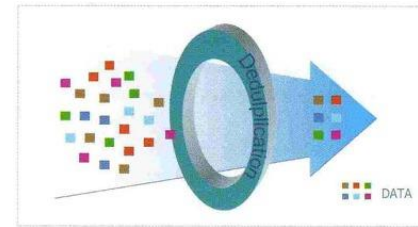
- 전체 data수 확인
- Missing data, 중복 데이터 확인

```
def div_cols(df):  
    df_division = {'number': [], 'string': []}  
    for i in df.columns:  
        # Numeric data  
        if df[i].dtype == 'int64':  
            df_division['number'].append(i)  
        elif df[i].dtype == 'float64':  
            df_division['number'].append(i)  
        # Categorical data  
        else:  
            df_division['string'].append(i)  
    return df_division  
  
df_division = div_cols(df)
```



S2. Data 및 Column 제거

- 필요 없는 column 제거
- Ex. elevator 수, 판매 유형



S3. 기존 data 변환 및 추가

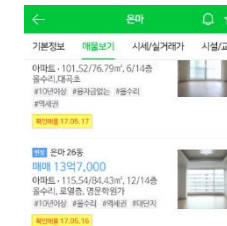
- 기존 data 형태 변환
- 필요한 column 추가 (평당 가격, 주차장 유형..)

SalePrice	YearBuilt	YrSold	MonthSold	PP	Size(sqft)	Size(sqft)	Floor	HallwayType	HeatingType	AptComplex	%Parking
141592	2008	2007	8	5740	814	25	3	terrace	individual	managem	111
51327	1985	2007	8	2886	587	18	8	comdor	individual	self_mana	80
48672	1985	2007	8	2736	587	18	6	comdor	individual	self_mana	80
305330	2006	2007	8	6108	2056	62	8	terrace	individual	managem	248
221238	1993	2007	8	4146	1761	53	3	mixed	individual	managem	523
35840	1992	2007	8	3332	355	11	5	comdor	individual	managem	200
78318	1992	2007	8	4013	644	20	2	mixed	individual	self_mana	142
61846	1993	2007	8	3174	644	20	10	mixed	individual	managem	523
84070	1993	2007	8	4308	644	20	3	mixed	individual	managem	523
83185	1992	2007	8	4283	644	20	13	mixed	individual	self_mana	142
168141	1986	2007	8	4030	1377	42	4	terrace	central	has_mana	713
153982	1986	2007	8	5560	914	28	11	terrace	central	has_mana	713
200884	2007	2007	8	7637	868	26	18	terrace	individual	managem	0
60176	1985	2007	9	3383	587	18	7	comdor	individual	self_mana	80
93362	1986	2007	9	3386	910	28	7	comdor	individual	managem	100
185840	1993	2007	9	4761	1286	39	24	mixed	individual	managem	523



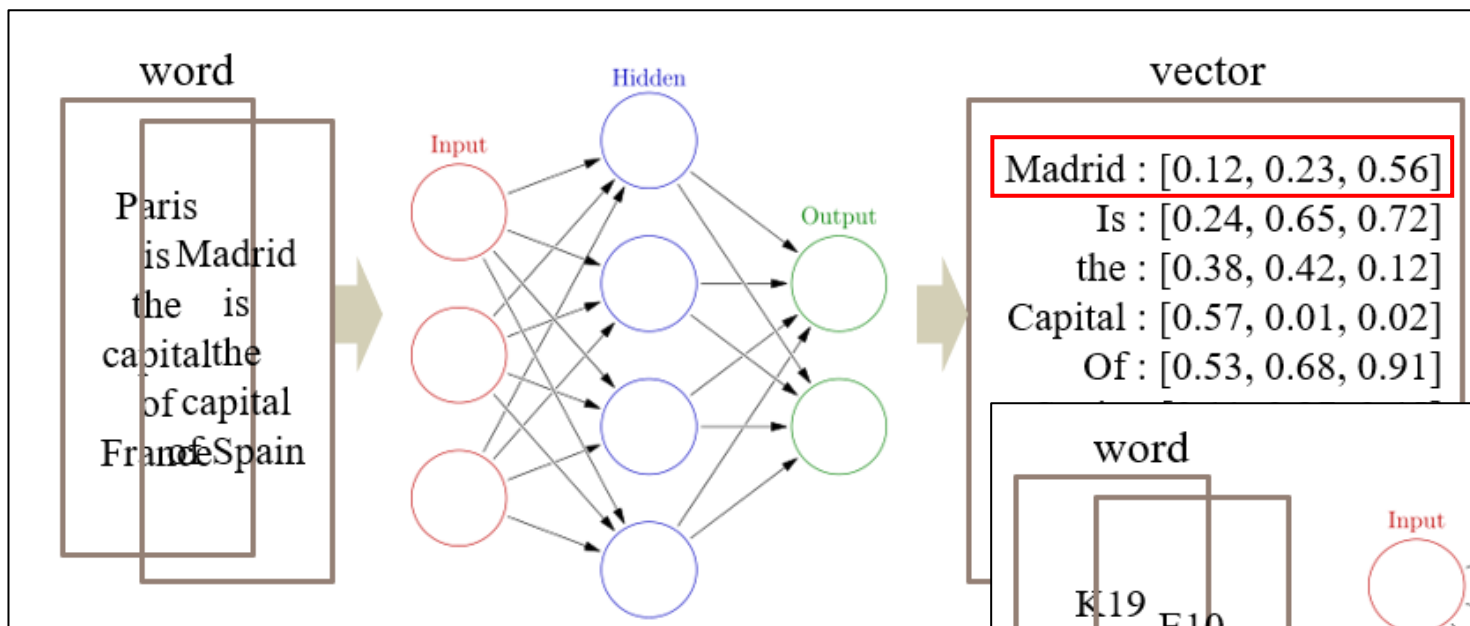
S4. Data detail 사항 확인

- Data type(numeric, categorical)
- Data detail 사항 확인(주변 data의 범위..)



3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 텍스트 데이터 => 텍스트마이닝

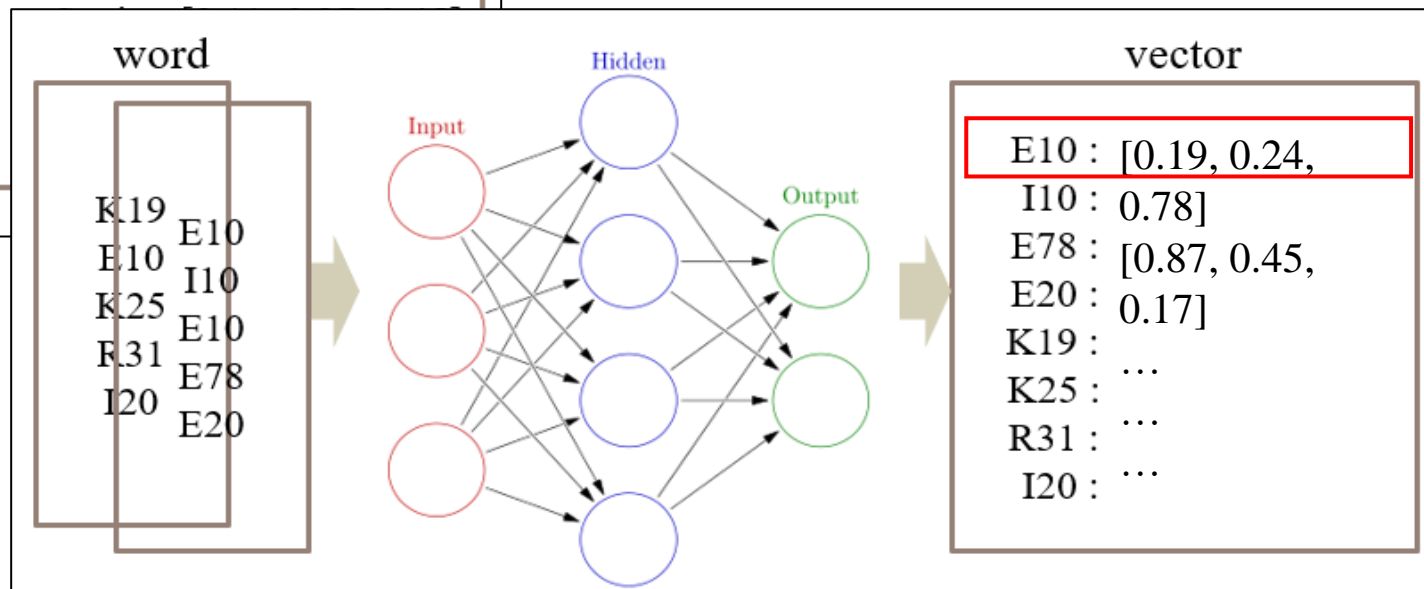


도시, 국가이름의 벡터화

자연어처리

word2vec(2013, google연구)원

(2) 질병코드의 벡터화 (E10 : 당뇨병 관련)



3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 텍스트 데이터 => 텍스트마이닝

- 텍스트 마이닝(text mining)이란, 다양한 알고리즘을 이용하여 대용량의 텍스트 문서로부터 트렌드와 관심어를 찾아내는 기법이다.

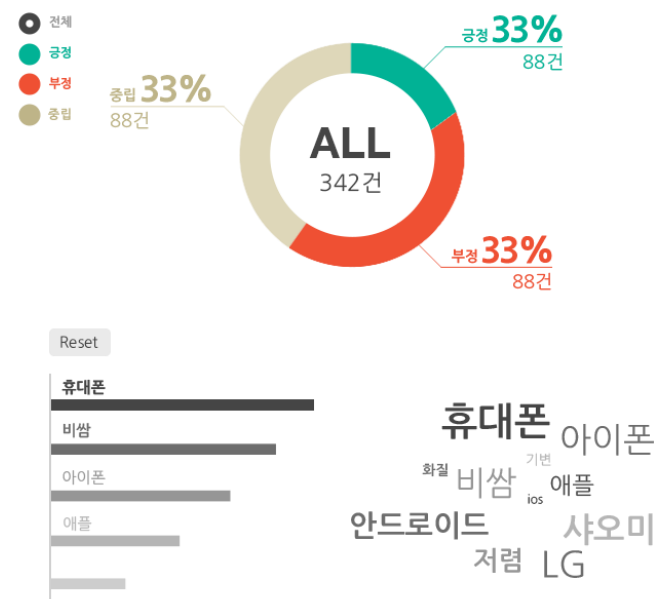


Image source: TREUM Co, Ltd.

기법	설명
키워드 분석	단어 <u>출현 빈도</u> 추출 후 시각화
단어간 관계 분석	문서 내 출현 빈도 높은 단어 파악 후 <u>단어간 상관관계</u> 계산
감성 분석	긍정 단어와 부정 단어 빈도 추출 후 <u>문서 감성</u> 수치화

3. 다양한 데이터 (정형·비정형) 분석 사례

19

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

- 웹(web)에 존재하는 정보(텍스트, 숫자 등)를 수집하는 작업
- 웹크롤링을 이용하여 가져올 수 있는 정보는 html기반의 웹사이트, 이미지, 텍스트 문서



3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

- <http://metacritic.com/movie/eternals/user-reviews> (예시 : 영화 <이터널스> 리뷰)
- 다른 영화의 리뷰를 수집하고자 하는 경우, 해당 영화의 주소 확인
예) 영화 : 해리포터와 마법사의 돌 : [harry-potter-and-the-sorcerers-stone](https://www.metacritic.com/movie/harry-potter-and-the-sorcerers-stone)

User Reviews

USER SCORE BY DATE MOST HELPFUL

8 grandpajoe6191 Sep 23, 2011
Director Chris Columbus has created "Harry Potter and the Sorcerer's Stone" a dazzling movie that stays true to its book and adding itself with pure zeal and passion with impressive visuals.
18 of 21 users found this helpful 18 3 All this user's reviews

7 Jaredc324 Jul 18, 2019
Both deeply enriching and wondrously enchanting. Chris Columbus puts the "m" in magic and makes Harry Potter the introductory chapter it needs to be and more.
3 of 3 users found this helpful 3 0 All this user's reviews

9 JPK Jun 14, 2019
Very Good Start
Philosopher's Stone proves that you can adapt a book and massively succeed.
2 of 2 users found this helpful 2 0 All this user's reviews

<https://www.metacritic.com/movie/harry-potter-and-the-sorcerers-stone/user-reviews>

3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

① 크롤링 및 데이터 전처리에 필요한 패키지 설치 또는 라이브러리 불러오기

```
#install.packages("NLP") #영어 자연어 처리
library(NLP)
#install.packages("xml2") #html 처리
library(xml2)
#install.packages("rvest") #html 처리
library(rvest)
#install.packages("stringr") #string 처리
library(stringr)
#install.packages("dplyr") #데이터 전처리
library(dplyr)

#install.packages("tm") #정보 추출 도구
library(tm)
#install.packages("wordcloud") #워드 클라우드
library(wordcloud)
```

② 크롤링 대상 URL 설정

```
url_base <- 'http://www.metacritic.com/movie/eternals/user-reviews?page='
```

- `http://www.metacritic.com/movie/eternals/user-reviews?page=`
- 위 URL에서 영화명 변경 시 다른 영화의 user review 수집 가능

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

4 크롤링 수행

```
for(page in 0:5){  
  url <- paste(url_base, page, sep='') # 1~5페이지까지 페이지 수집  
  htxt <- read_html(url)  
(1) comment <- html_nodes(htxt, 'div.review_body') # user review 위치  
  review <- html_text(comment) # 실제 리뷰의 text 파일만 추출  
  review <- repair_encoding(review, from = 'utf-8') # 인코딩 변경  
(2) review <- str_replace(review, "This review contains spoilers, click expand to view.", "")  
(3) review <- str_trim(review)  
  reviews <- c(reviews, review) # 결과값 저장  
}
```

- 140자 평의 위치를 찾아 **comment**에 저장
- 스포일러 포함 경고가 있는 경우 제거하고 저장
- **문장 앞뒤 공백 제거하여 저장**
➔ 크롤링 수행하면서 **전처리 수행** 가능!

3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

5 크롤링 결과 확인

```
> head(reviews)
[1] "A spoiler free review Go watch it knowing it's not the same as any other m
arvel movie and it's more of a family drama then action and you will love it.
The diversity of the cast is great and the movies culture representation is t
heA spoiler free review Go watch it knowing it's not the same as any other marv
el movie and it's more of a family drama then action and you will love it. The
diversity of the cast is great and the movies culture representation is the gr
eatest mcu one I ever seen so go watch it and you will love the story, the char
acters, and the plot.... Expand"
[2] "Loves it. Marvel movies were starting to get a bit boring to me tbh, but e
ternals changes that. It' marvel movie with all the action and lores but also a
health dose of complexity, drama and intrigue. Hope the studio keep the directi
```

User Reviews

USER SCORE

BY DATE

MOST HELPFUL

10

Finsherad

Nov 6, 2021

A spoiler free review Go watch it knowing it's not the same as any other marvel movie and it's more of a family drama then action and you will love it. The diversity of the cast is great and the movies culture representation is the... Expand

131 of 141 users found this helpful 131 10

All this user's reviews

10

Lylithania

Nov 5, 2021

Loves it. Marvel movies were starting to get a bit boring to me tbh, but eternals changes that. It' marvel movie with all the action and lores but also a health dose of complexity, drama and intrigue. Hope the studio keep the direction they... Expand

원래의 웹문서

3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

텍스트 데이터의 전처리

- **토큰화(tokenization)** : 주어진 **말뭉치(corpus)**에서 **토큰(token)**이라 불리는 단위로 나누는 작업
 - 토큰의 단위는 상황에 따라 문장, 단어, 어절, 형태소 등으로 달라짐
 - 보통 의미있는 단위로 토큰을 정의함
- **코퍼스(corpus, 말뭉치)** : 텍스트 집합을 의미함
 - 예) 신문기사(html, text), SNS(트위터, 페이스북)
- **불용어(stopword)** : 분석 시 의미를 갖지 않는 단어 리스트. 제거하기 위함.
- **공백 제거, 문장부호 제거** ⑥ ⑦

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

8 전처리 결과 확인

```
> inspect(corpus[[1]])  
<<PlainTextDocument>>  
Metadata: 7  
Content: chars: 317  
  
spoiler free review go watch knowing marvel family drama action will  
diversity cast great culture representation thea spoiler free review g  
tch knowing marvel family drama action will love diversity cast great
```

```
> inspect(corpus[[2]])  
<<PlainTextDocument>>  
Metadata: 7  
Content: chars: 315  
  
loves marvel starting get bit boring tbh eternal's changes marvel action  
res also health dose complexity drama intrigue hope studio keep directio  
theyloves marvel starting get bit boring tbh eternal's changes marvel ac
```

3. 다양한 데이터 (정형·비정형) 분석 사례

26

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

```
## TD matrix 기반 워드클라우드
windowsFonts(font=windowsFont("맑은 고딕")) #폰트 설정

set.seed(1234) #동일한 워드클라우드 생성 (난수 고정)

wordcloud(words=rownames(doc),          #키워드
           freq=doc$word_order,         #빈도
           min.freq=2,                  #최소 출현 빈도
           max.words=50,                 #출력 키워드 수
           random.order=FALSE,           #고빈도 키워드 중앙 배치
           scale = c(5,1),              #키워드 크기 범위
           rot.per=0.35,                 #회전 키워드 비율
           family="font", colors=brewer.pal(8,"Dark2"))
```



텍스트분석과 시각화 (Wordcloud)



3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

텍스트분석 : 빈도와 상관분석

```
#특정 횟수 이상 언급된 단어들 찾기
findFreqTerms(tdm, 50)
findFreqTerms(tdm, 100)
findFreqTerms(tdm, 150)
findFreqTerms(tdm, 200)
findFreqTerms(tdm, 300)
```

특정빈도 이상의 단어 찾기 : findFreqTerms

영화평에 100번 이상 언급된 단어

```
> findFreqTerms(tdm, 100)
[1] "action"      "bad"          "characters"   "eternals"    "even"
[6] "expand"      "film"         "good"         "great"       "just"
[11] "like"        "marvel"       "mcu"          "movie"       "movies"
[16] "new"         "one"          "que"          "really"      "see"
[21] "story"       "time"

> findFreqTerms(tdm, 150)
[1] "characters" "eternals"    "expand"      "film"        "good"
[6] "great"      "just"        "like"        "marvel"      "mcu"
[11] "movie"      "movies"      "one"         "really"      "story"
```

영화평에 150번 이상 언급된 단어

3. 다양한 데이터 (정형·비정형) 분석 사례

(2) 비정형 데이터 : 웹크롤링 => 텍스트마이닝

텍스트분석 : 특정 단어와의 상관분석

```
#특정 단어와 관련된 단어 찾기(상관 관계)
findAssocs(tdm, "marvel", 0.3)
findAssocs(tdm, "characters", 0.3)
findAssocs(tdm, "mcu", 0.4)
```



'marvel' 와 상관성이 0.3 이상인 용어

```
> findAssocs(tdm, "marvel", 0.3)
$marvel
expand      get pulling
  0.35      0.31    0.30
```

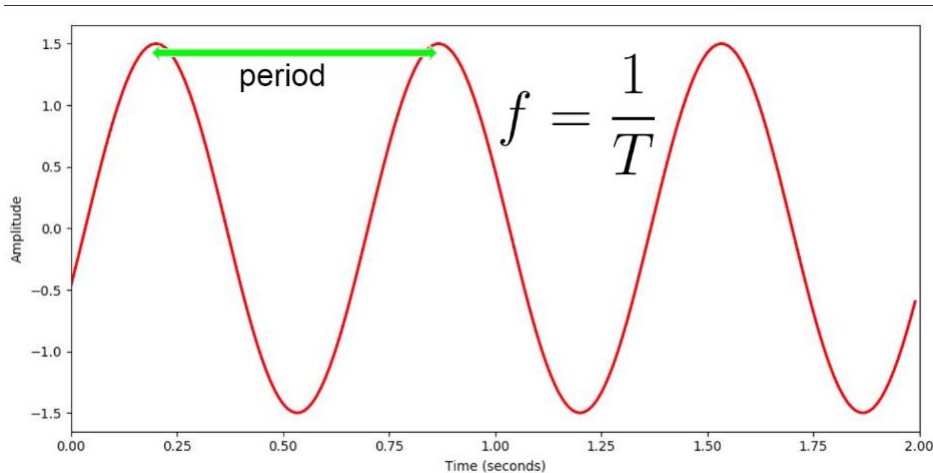
'characters' 와 상관성이 0.3 이상인 용어

```
> findAssocs(tdm, "characters", 0.3)
$characters
      expand      eternal      many      care      guardians      main
      0.39      0.36      0.35      0.34      0.34      0.34
interesting      new      galaxy      get      movie
      0.32      0.32      0.31      0.31      0.30
```

(3) 비정형 데이터 : 오디오 데이터

waveform 형태를 띄게 되어 우리가 흔히 보는 그래프 형태

Y축은 Amplitude(진폭), X축은 Time(sec), 특정 지점에서 다음 등장하는 그 값까지를 period(주기), 즉 Frequency(주파수)이며 단위는 Hz를 사용. 1초에 100번 진동하는 소리를 100Hz로 정의. 따라서 $f=1/T$ 가 성립. Amplitude는 소리의 크기와 관련, 대체로 Amplitude가 크면 소리가 크다.



<https://hyunlee103.tistory.com/54>

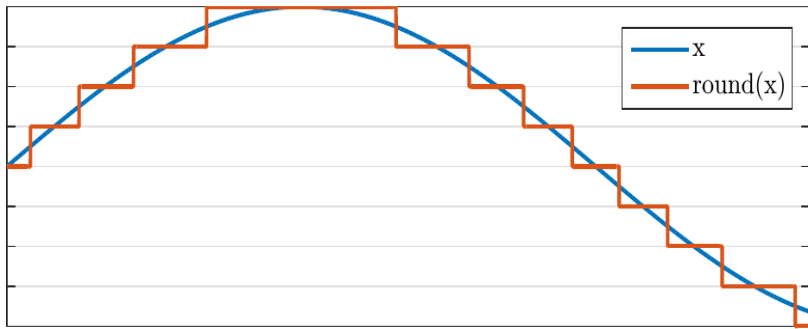
3. 다양한 데이터 (정형·비정형) 분석 사례

30

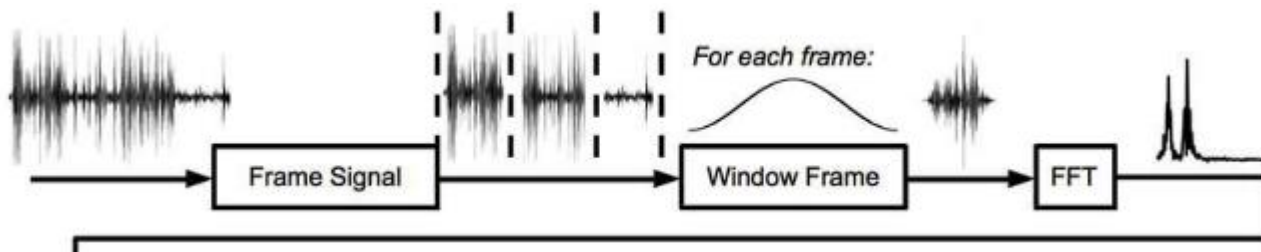
(3) 비정형 데이터 : 오디오 데이터

- Analog Digital Conversion을 거쳐 이산 벡터를 생성-> 딥러닝 모델 적용

오디오 데이터는 연속형 데이터 => 모델링하기 위해서는 discrete한 벡터로 만들어야 한다.



- Feature extraction : 신호의 특성을 반영하는 feature(특징)을 도출하여 분석
- Windowing : 오디오 데이터는 연속적이고 시간 의존성이 있으므로 time invariant 가정을 충족하는 구간으로 나눔

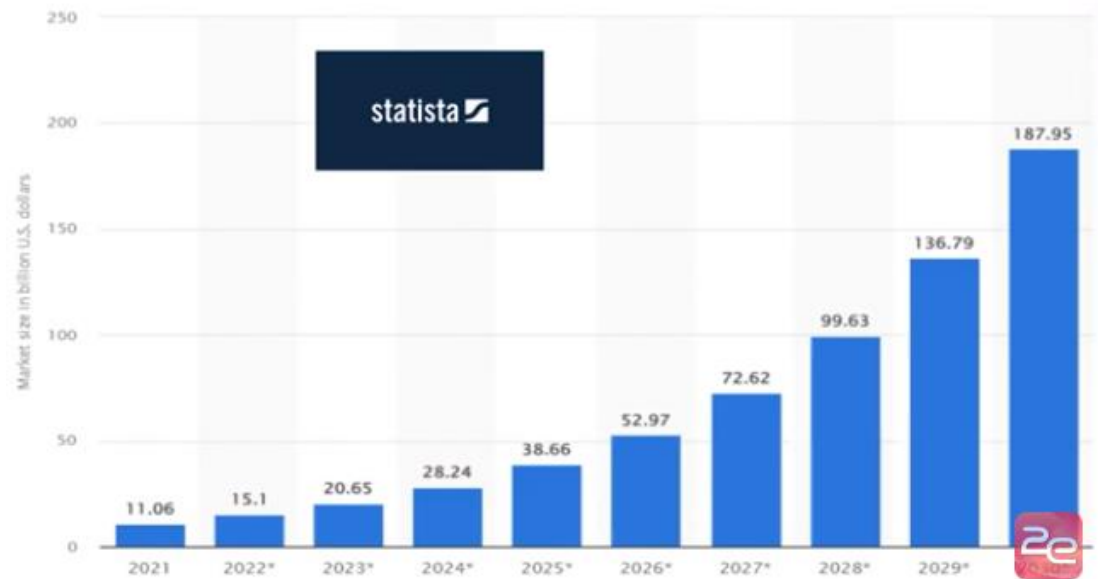


<https://hyunlee103.tistory.com/54>

4. 분석모형의 활용

(1) 의료분야 AI플랫폼 : 생성형AI, 텍스트마이닝 기술 활용

의사들의 업무 부담을 덜어주는 헬스케어 코파일럿 서비스, DAX Copilot



<https://www.2e.co.kr/news/articleView.html?idxno=302942>

Dax Express는 환자와 대화내용을 AI가 인지하고 EMR(의료기록부)에 자동으로 입력하는 서비스

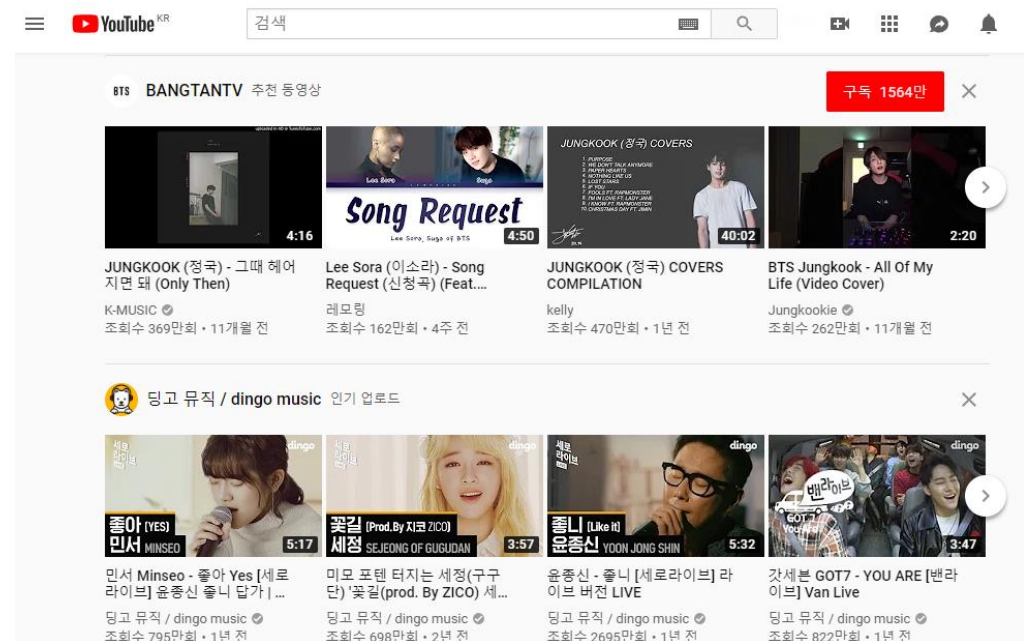
4. 분석모형의 활용

(2) 예측모형의 구현 시나리오 : 영화, 음악

영화 추천



음악 추천



어떻게 원하는 콘텐츠를 예측하여 추천하는가?

4. 분석모형의 활용

(2) 예측모형의 구현 시나리오 : 영화, 음악

과거의 구매패턴을 분석하여 미래의 구매를 예측(추천)

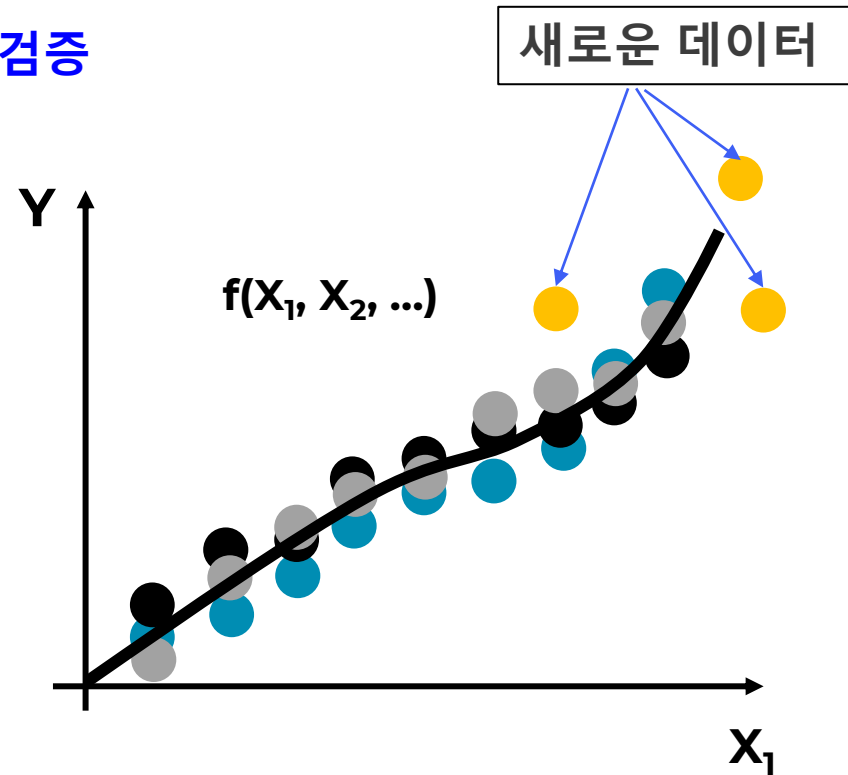
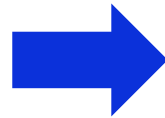


4. 분석모형의 활용

34

(2) 예측모형의 구현 시나리오 : 영화, 음악

- 주어진 데이터 => 예측모형 구현 => 새로운 데이터로 검증

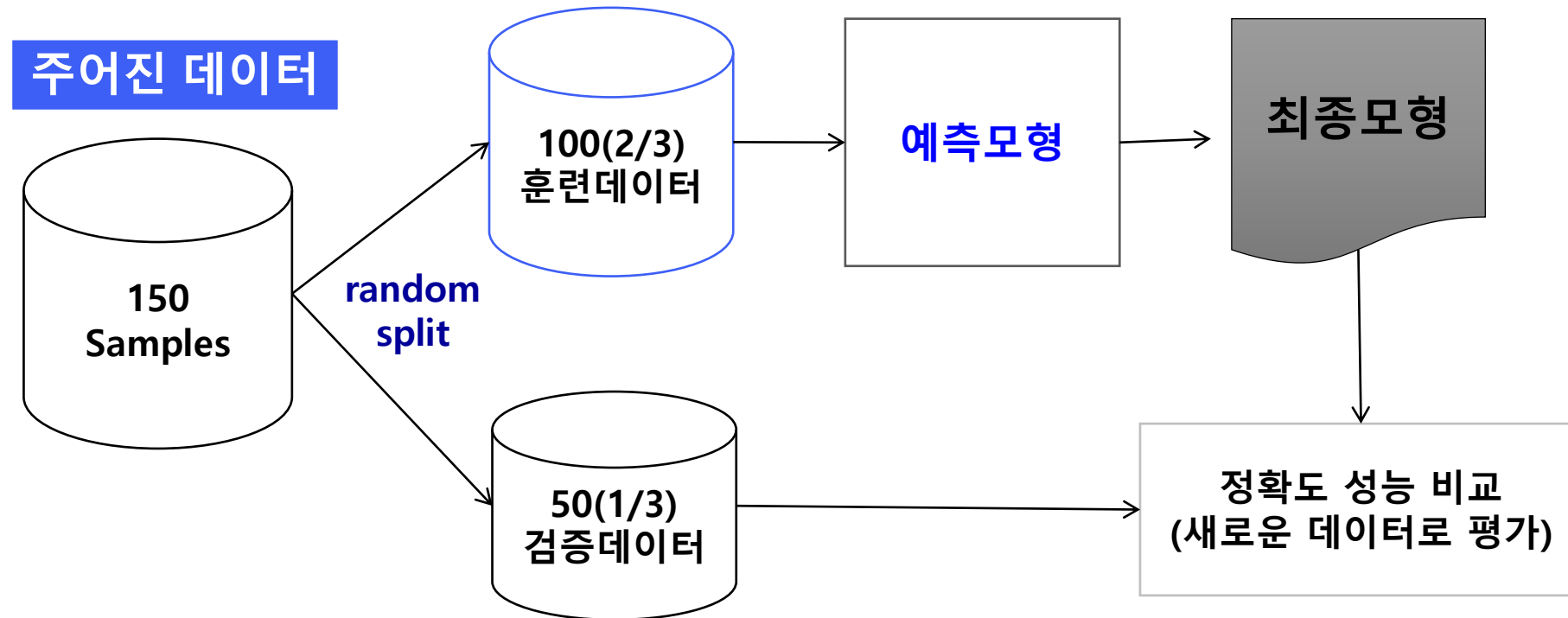


모형이 적절한지 어떻게 판단하는가?

4. 분석모형의 활용

35

(3) 예측모형의 훈련데이터와 검증데이터

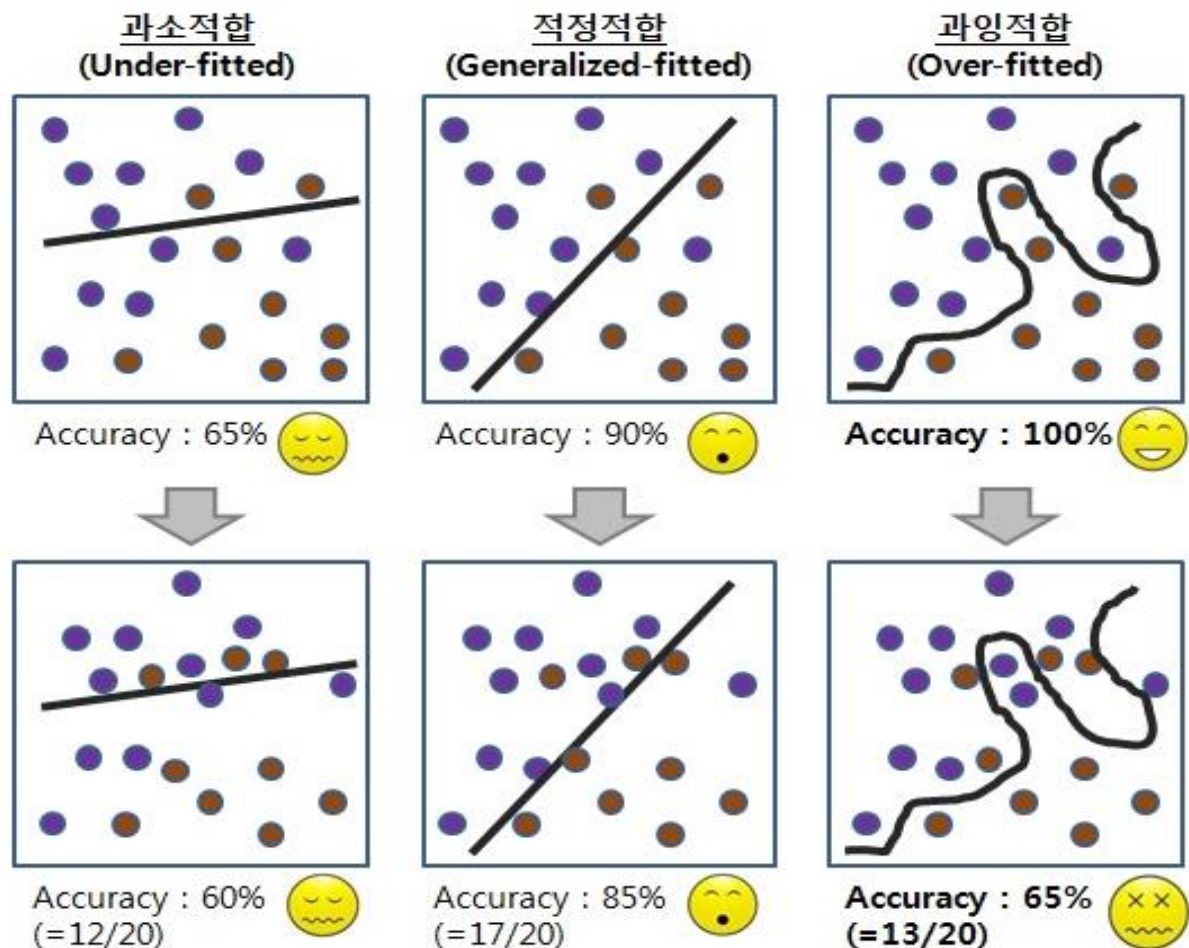


3- fold cross-validation (n=150)

(4) 예측모형 결과 모니터링과 피드백

- 예측모형의 오차수준
- 예측모형의 과적합을 경계
- 정확도가 높은 모형이 과연 맞을까?

예측모형의 과적합(overfitting)



4. 분석모형의 활용

(4) 예측모형 결과 모니터링과 피드백



예측모형에 대한 중요



과적합을 주의하여
안정성있는 예측모형개발

훈련데이터와 검증데이터 기반 예측모형 구현/평가 필요

5. 데이터 안심구역에서의 분석 활용

- 데이터 안심구역은 미개방 데이터를 안정적으로 분석 활용할 수 있는 플랫폼
- 과학기술정보통신부와 한국데이터산업진흥원이 구축
- 공공기관 및 민간기업의 미개방 데이터를 누구나 안전한 분석환경에서 무료로 활용 가능

데이터 안심구역 제공데이터




데이터 안심구역 사이트 : <https://dsz.kdata.or.kr/svc/page/intro/safezone.do>

5. 데이터 안심구역에서의 분석 활용

39


- 데이터 안심구역 : 서울센터(한국데이터산업진흥원), 대전센터(충남대학교 정보화본부)

데이터안심구역
서울센터



상세보기


데이터안심구역
대전센터



상세보기

• 분석환경 정보


인기



Studio


R studio
R 프로그래밍 언어 기반의
오픈소스 소프트웨어로 통계
분석에 보다 특화된
라이브러리를 활용한 데이터
분석 등에 유용

인기




Jupyter

Jupyter
Python 프로그래밍 언어 기반의
오픈소스 소프트웨어로서 각종
라이브러리를 활용한
데이터분석 및 머신러닝 등에
유용



Tableau

빠르게 정보를 탐색하여
즉각적인 통찰을 얻을 수 있으며,
전문적인 데이터 분석 기술
없이도 직접 고품질 데이터
시각화 활용 가능



rapidminer

Rapidminer
대표적인 NO 프로그래밍,
GUI기반의 예측적 데이터 분석
솔루션으로 예측 분석, 데이터
관리, 시각화 등에 활용 가능

데이터 안심구역 사이트 : <https://dsz.kdata.or.kr/svc/page/intro/safezone.do>

5. 데이터 안심구역에서의 분석 활용

40

- 데이터 안심구역 이용 신청서



이용절차안내

안심구역 이용신청

연계기관 이용신청

데이터 안심구역 사이트 : <https://dsz.kdata.or.kr/svc/page/intro/safezone.do>

데이터 제공기관 현황



통계청
열린 통계허브 구축을 통해 국가정책을
선도하고 국민의 미래를 설계하는
기관입니다.



건강보험심사평가원
건강하고 안전한 의료문화를 열어가는
국민의료평가기관입니다.



한국교통안전공단
교통안전관리의 효율화를 도모하기
위하여 설립된 기관입니다.



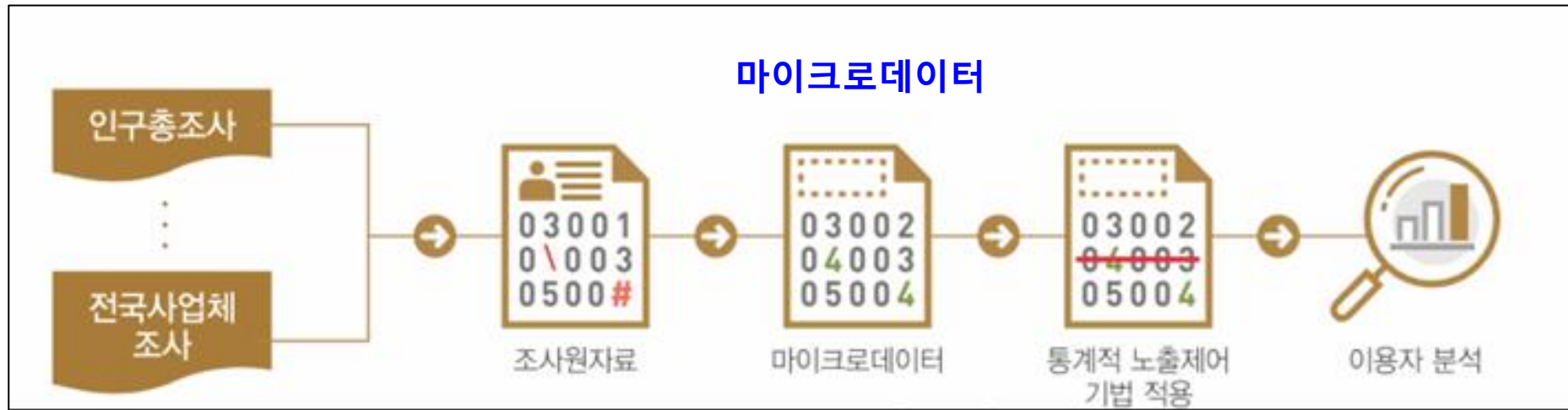
한국전력공사
깨끗하고 안전한 전기를 만들고 더
편리한 '에너지플랫폼' 서비스를
제공합니다.

데이터 안심구역 사이트 : <https://dsz.kdata.or.kr/svc/page/intro/safezone.do>

5. 데이터 안심구역에서의 분석 활용

국내 공공데이터의 분석활용 : 데이터센터

- 마이크로데이터 : 통계조사 원자료(raw data)에서 개인정보, 입력오류, 논리오류 등을 수정한 조사개별 단위(개인, 가구, 사업체별 등) 자료.
- 응답자 정보의 노출을 통제한 기법을 적용하여 제공

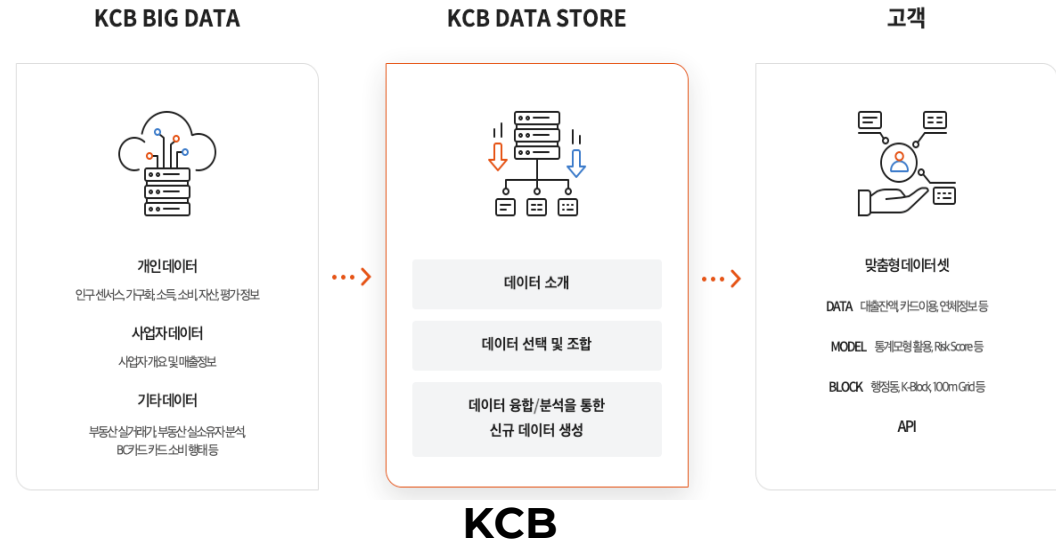
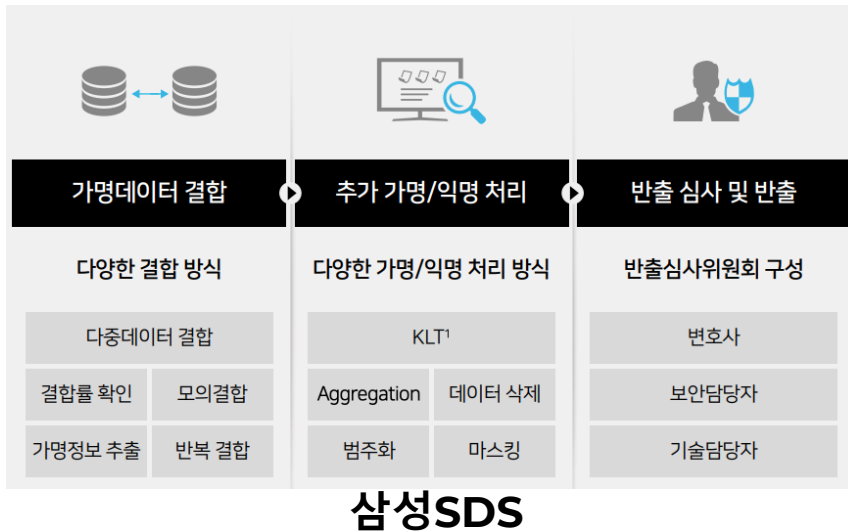


5. 데이터 안심구역에서의 분석 활용

43

• 데이터센터(데이터전문기관)

- 금융사끼리, 또는 금융사와 비금융사간 **가명정보 데이터를 결합**해주는 기관
- 결합데이터의 **익명성**이 보장되는지 평가
- **신뢰받는 결합전문기관**이 관련 법률에 따라 결합을 수행
 - 한국신용정보원, 금융결제원, 삼성SDS, KCB 등
- 데이터 결합을 통해 **더욱 가치 있는 데이터**를 생성할 것을 기대



5. 데이터 안심구역에서의 분석 활용

• 편의점 AI 상품추천모델 구축 사례(이마트24)

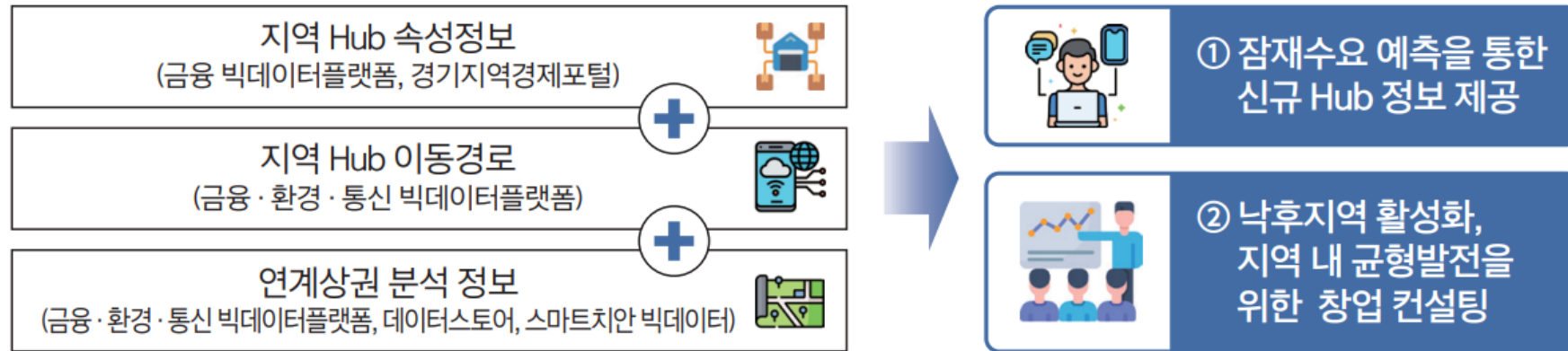
- 고객 결제정보를 보유한 카드사의 데이터와 그 카드를 보유한 고객의 구매 데이터를 결합
- 고객 맞춤형 상품구성을 제공
- 고객 편의성을 높이고, 브랜드 이미지를 제고해 고객 확대 및 점포의 가치상승을 기대



5. 데이터 안심구역에서의 분석 활용

• 데이터 기반 창업 및 의사결정 지원 사례

- 소규모 유동인구 집중시설(지역 Hub)에 대한 속성 데이터, 이동 경로 분석 데이터 등을 결합
- 데이터 기반 컨설팅의 일환으로 지역·업종별 맞춤형 추천 서비스를 제공
- 데이터에 기반한 창업 및 의사결정 문화가 확산될 것으로 기대



감사합니다