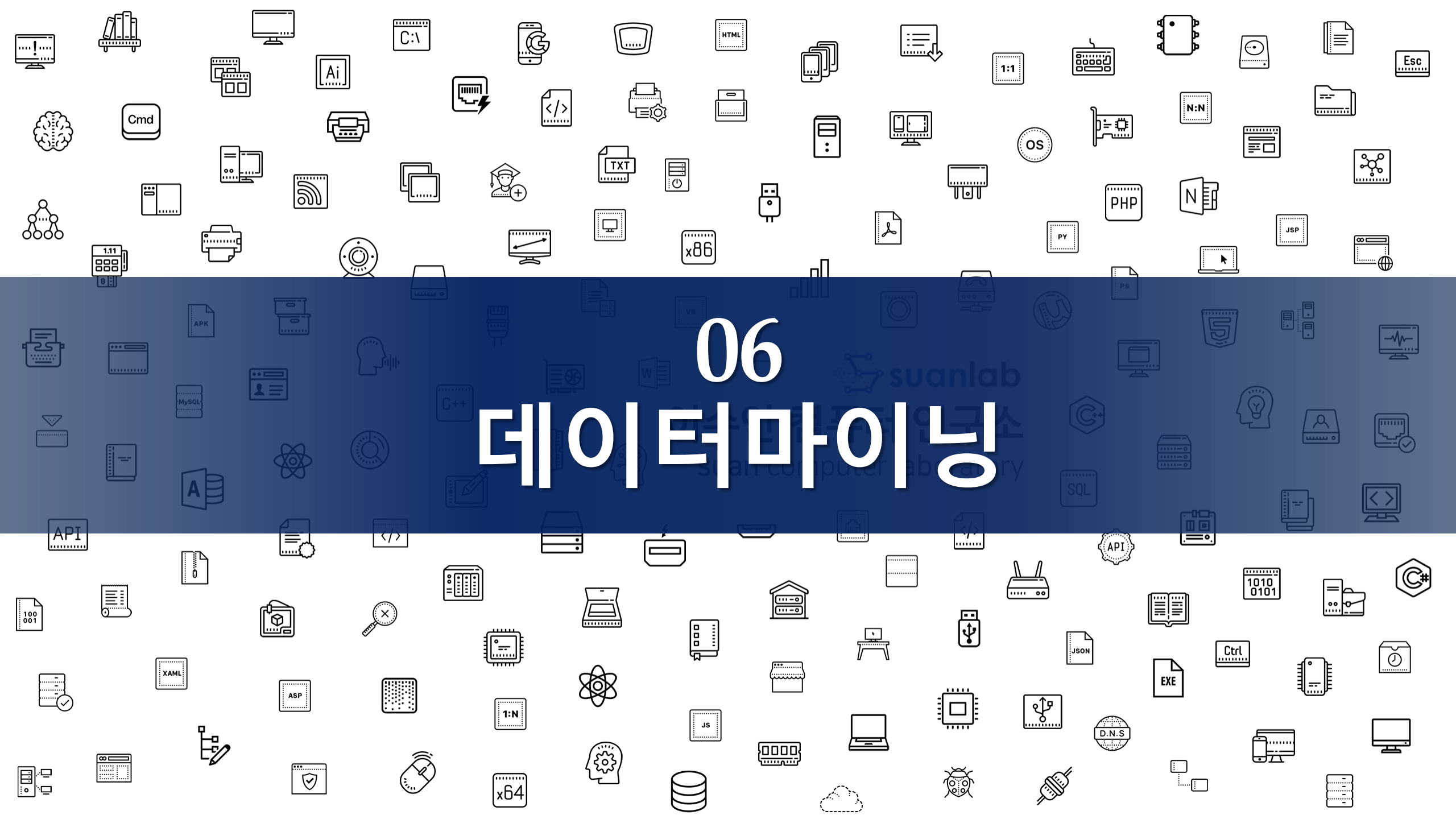
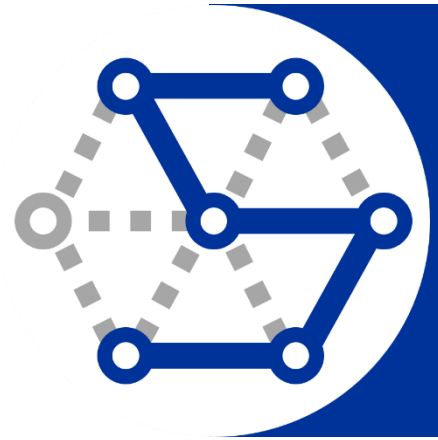
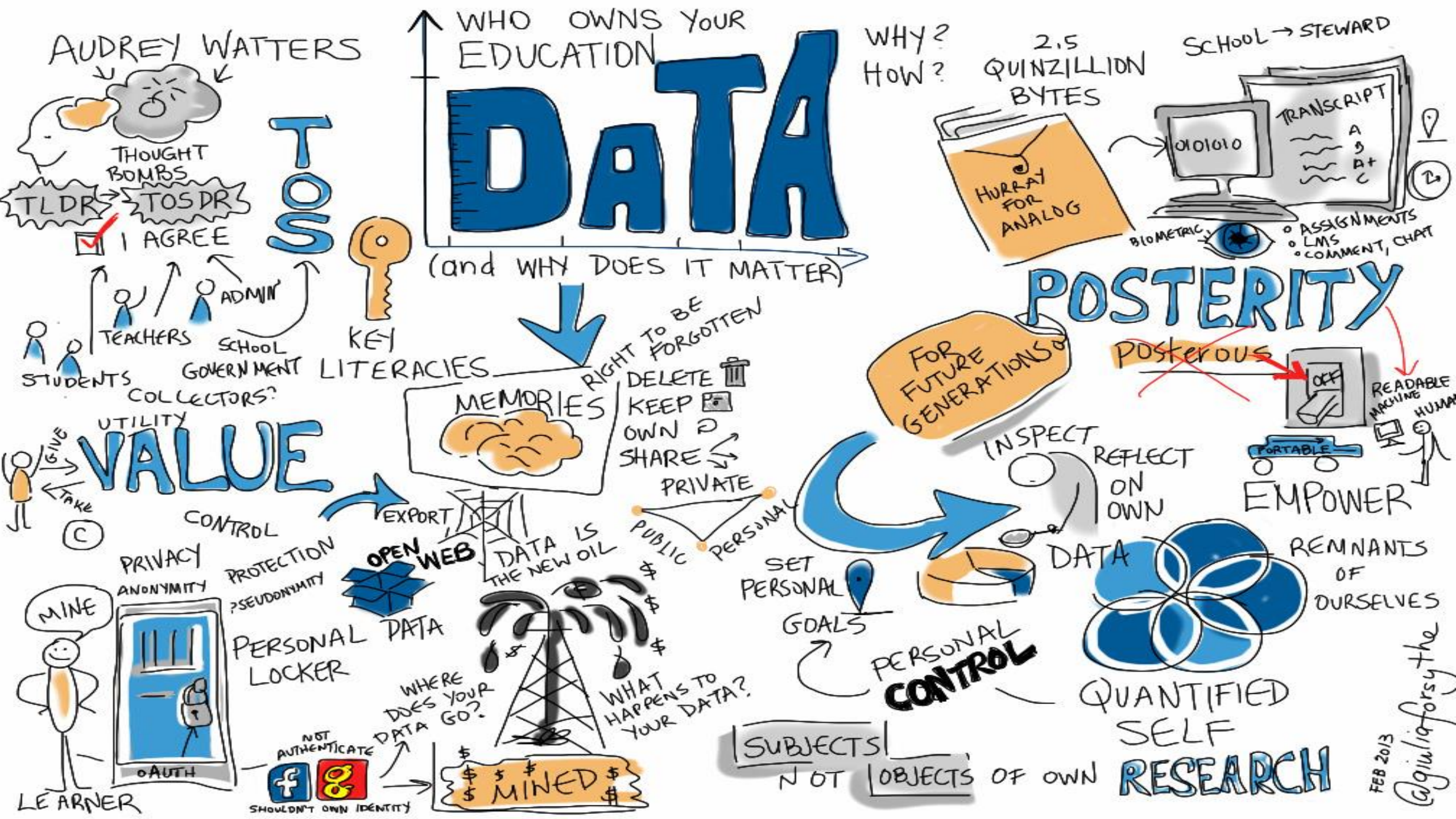

인공지능
Artificial Intelligence


06
데이터마이닝



1 데이터


AUDREY WATTERS
THOUGHT BOMBS
TLDR
TOSDR
I AGREE
TOS
TEACHERS
ADMIN
SCHOOL GOVERNMENT
STUDENTS
COLLECTORS?
KEY
WHO OWNS YOUR EDUCATION
DATA
(and WHY DOES IT MATTER)
WHY?
HOW?
2.5 QUINZILLION BYTES
HURRAY FOR ANALOG
SCHOOL → STEWARD
TRANSCRIPT
A
B
A+
C
0101010
BIOMETRIC
ASSIGNMENTS
LMS
COMMENT, CHAT
POSTERITY
FOR FUTURE GENERATIONS
POSTEROUS
OFF
PORTABLE
READABLE MACHINE
HUMAN
EMPOWER
INSPECT
REFLECT ON OWN DATA
SET PERSONAL GOALS
PERSONAL CONTROL
QUANTIFIED SELF RESEARCH
REMNANTS OF OURSELVES
SUBJECTS NOT OBJECTS OF OWN
UTILITY
VALUE
GIVE
TAKE
CONTROL
EXPORT
MEMORIES
RIGHT TO BE FORGOTTEN
DELETE
KEEP
OWN
SHARE
PRIVATE
PUBLIC
PERSONAL
OPEN WEB
DATA IS THE NEW OIL
PRIVACY
ANONYMITY
PROTECTION
PSEUDONYMITY
PERSONAL DATA LOCKER
MINE
OAUTH
LEARNER
NOT AUTHENTICATE
SHOULDN'T OWN IDENTITY
WHERE DOES YOUR DATA GO?
WHAT HAPPENS TO YOUR DATA?
MINED
FEB 2013
@juliaforsythe

데이터 단어 유래

https://en.wikipedia.org/wiki/Data
https://namu.wiki/w/데이터

- 데이터data는 라틴어 단어 Datum의 복수형인 Data에서 유래
- 라틴어에서 Datum의 뜻은 "present/gift, that which is given, debit"
- 현재에서도 기본적으로는 복수형 취급을 하나 가끔 하나의 고유명사화가 되어서 단수로 취급하는 경우도 있음

데이터 용어 정의

https://en.wikipedia.org/wiki/Data
https://namu.wiki/w/데이터

- 이론을 세우는 데 기초가 되는 사실. 또는 바탕이 되는 자료
- 관찰이나 실험, 조사로 얻은 사실이나 자료
- 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
- 데이터는 정보(information)가 아니고, 데이터를 가공해 얻는 것이 정보

데이터 세트 Data set

https://en.wikipedia.org/wiki/Data_set

- 데이터 모음
- 하나의 데이터베이스 테이블의 내용이나 하나의 통계적 자료 행렬과 일치
- 컬럼(column): 특정한 변수를 대표
- 로우(row): 주어진 멤버와 일치
- 변수 개개의 값들을 나열하고, 각각의 값은 데이터라고 부름
- 하나 이상의 멤버에 대한 데이터를 이루며, 로우의 수와 일치
- 웹에서 접근하고 다운로드 할 수 있는 다양한 형태의 데이터 세트가 존재

Id	Duration(hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

Google Dataset: https://toolbox.google.com/datasetsearch
Google AI Dataset: https://ai.google/tools/datasets/

데이터 세트Data set

- 데이터 세트(data set): 데이터 개체(data object)들의 집합
- 데이터 개체(data object): 레코드(record), 점(point), 벡터(vector), 패턴(pattern), 사례(case), 사건(event), 샘플(sample), 관찰(observation), 개체(entity) 등으로 불림
- 데이터 개체는 여러 개의 속성(attribute)으로 기술
- 속성(attribute) : 데이터 개체들 사이의 차이를 규정할 수 있는 특성이나 특징을 의미
  - 예) 사람을 기술할 때 눈동자의 색, 피부색, 키, 몸무게와 같은 속성을 사용
- 속성은 변수(variable), 특성(characteristic), 필드(field), 특징(feature), 차원(dimension) 등으로 불림

데이터 형태 – 실험설계 유무 기준

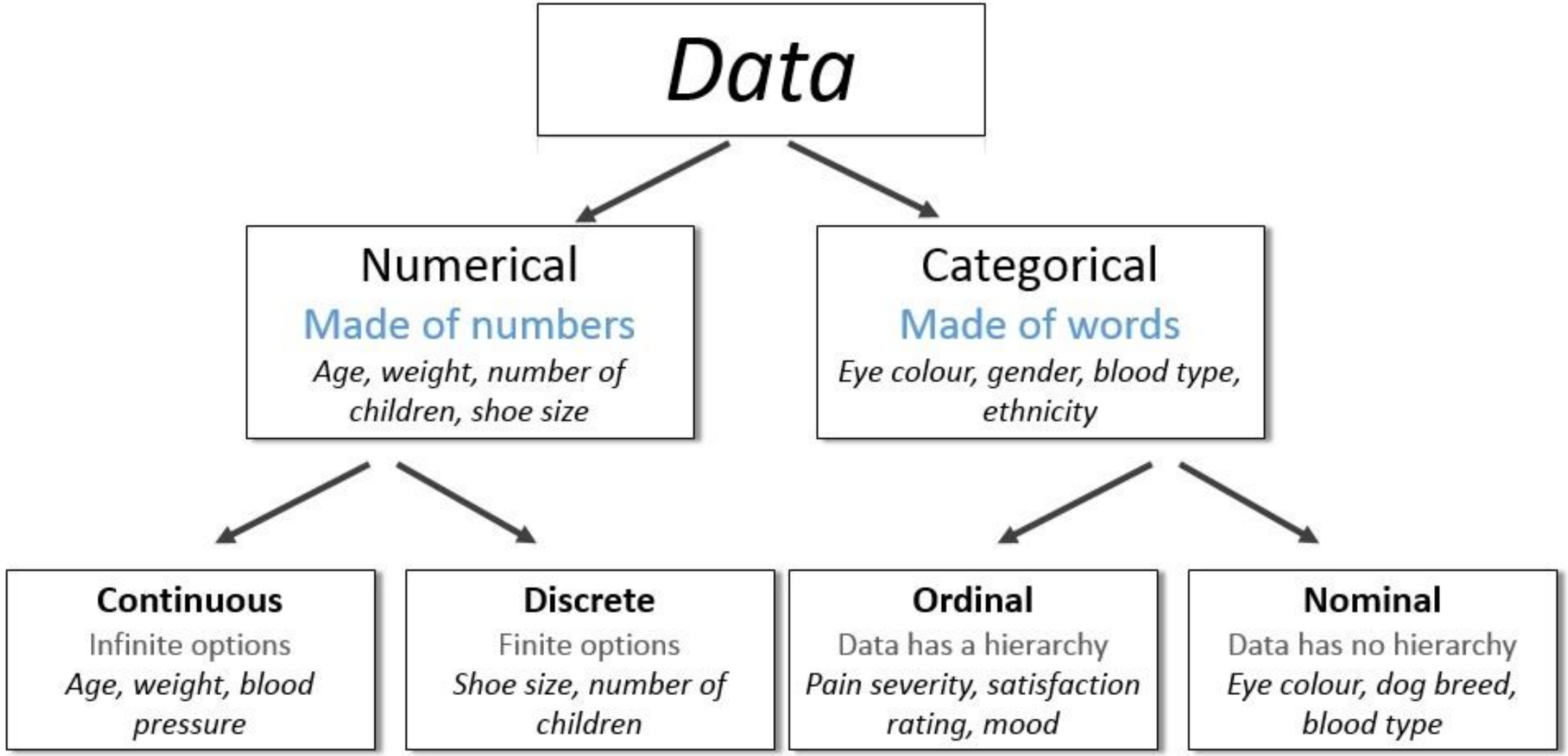
- **실험 데이터**(experimental data)
  - 생물학 실험, 약물 실험, 물리학 실험 등 **설정된 실험환경**에서 **수집**된 데이터
  - 미리 **가설**(hypothesis) 설정 후 가설에 따른 실험 설계
  - **가설 검증** 또는 **탐색적 분석**(exploratory analysis)
- **관측 데이터**(observational data)
  - 실험과정이 설계되지 않은 환경에서 **관측되어 수집**된 데이터
  - 센서 데이터, 유전체(genome) 데이터, 웹 로그, 거래(transaction) 데이터, 네트워크 트래픽 데이터
  - **주된 데이터마이닝 대상**

데이터 형태

- 질적자료(정성적자료, Qualitative or Categorical): 범주 또는 순서 형태의 속성을 가지는 자료
  - 범주형(명목형, nominal) 자료: 사람의 피부색, 성별
  - 순서형(서수형, ordinal) 자료: 제품의 품질, 등급, 순위
- 양적자료(정량적자료, Quantitative or Numeric): 관측된 값이 수치 형태의 속성을 가지는 자료
  - 범위형interval 자료: 화씨, 섭씨와 같이 수치 간에 차이가 의미를 가지는 자료.
  - 비율ratio 자료: 무게와 같이 수치의 차이 뿐만 아니라 비율 또한 의미를 가지는 자료

데이터 분류Data Classification

http://survivestatistics.com/variables/



데이터의 형태 – 정형(structured) 데이터

- 일정한 구조 보유
- 데이터베이스 테이블(table, relation)
- 시장바구니 데이터(market basket data)
  - 매출별 구매 항목 목록에 대한 데이터
  - 행(row)이 항목(item)의 리스트 구성

거래번호	구매 항목
T1	빵, 우유
T2	기저귀, 달걀, 맥주, 빵
T3	기저귀, 맥주, 우유, 콜라
T4	기저귀, 맥주, 빵, 우유
T5	기저귀, 빵, 우유, 콜라

데이터의 형태 – 비정형(unstructured) 데이터

- **구조가 일정하지 않은 데이터**

- **텍스트**(text) **데이터** : 신문기사, SNS 메시지 등
- **스트림**(stream) **데이터** : 지속적으로 관측되어 생성되는 데이터
- **서열**(sequence) **데이터** : 염기 서열, 아미노산 서열 데이터

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC

- **클릭**(click) 데이터 : 홈페이지 방문자들의 순차적인 클릭
- **시스템 로그**(log) 데이터
- **그래프**(graph) 데이터

데이터의 형태 – 반정형(semi-structured) 데이터

- 구조화되어 있지만 관계형 데이터베이스의 테이블과 같은 형태로 저장되기 곤란한 데이터
- **XML(eXtensible Markup Language)** 등으로 표현

```
<title>The Transporter</title>
<year>2002</year>
<language>English</language>
<genre>Action</genre>
<genre>Crime</genre>
<genre>Thriller</genre>
<country>USA</country>
<actors>
   <actor>Jason Statham</actor>
   <actor>Matt Schulze</actor>
   <actor>François Berléand</actor>
   <actor>Ric Young</actor>
   <actress>Qi Shu/actress>
</actors>
<team>
   <director>Louis Leterrier</director>
   <director>Corey Yuen</director>
   <writer>Luc Besson</writer>
   <writer>Robert Mark Kamen</writer>
   <producer>Luc Besson</producer>
   <cinematographer>Pierre Morel</cinematographer>
</team>
```

데이터 종류

- Record data
  - Transaction or Market Basket Data
  - Data matrix (Pattern matrix)
  - Sparse Data Matrix
- Graph-based data
- Ordered data
  - Sequential data
  - Sequence data
  - Time series data
  - Spatial data

레코드 데이터(Record data)

- 데이터 마이닝에서 가장 많이 사용되는 데이터 형태로 대개 flat 파일 형태로 저장된 데이터 세트
- 레코드(Record)의 모음으로 구성
- 각 레코드는 고정된 수의 속성으로 구성

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

트랜잭션 데이터(Transaction Data)

- 구매자와 구매 물품목록 형태로 이루어진 데이터 세트
- 장바구니 데이터(Market Basket Data)라고도 불림

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

데이터 행렬(Data matrix)

- 모든 속성이 수치 형태의 값을 가지는 행렬 형태의 데이터 세트
- 일반적으로 데이터의 행은 개체, 열은 속성을 나타냄
- 패턴 행렬(Pattern matrix)이라고도 불림

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

희박한 데이터 행렬(Sparse Data Matrix)

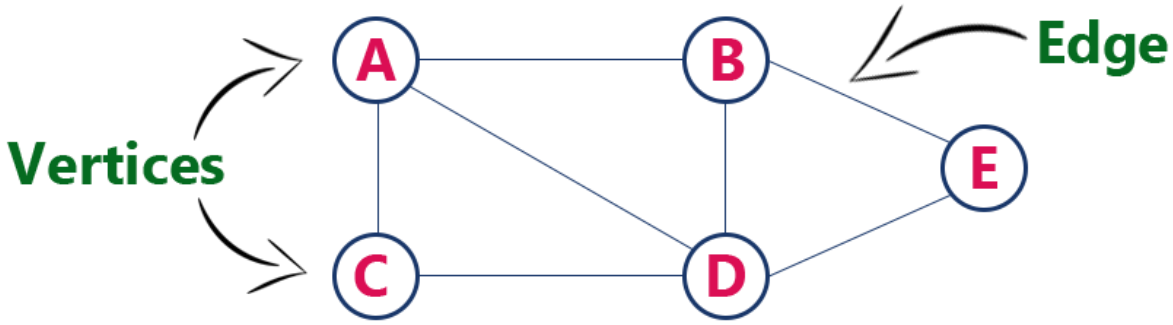
- Data matrix의 특별한 경우
- 예: 각 문서에서 용어가 출현하는 빈도수
- 문서의 경우에는 용어 벡터(term vector) 형태로 표현 가능

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

그래프 데이터(Graph-based data)

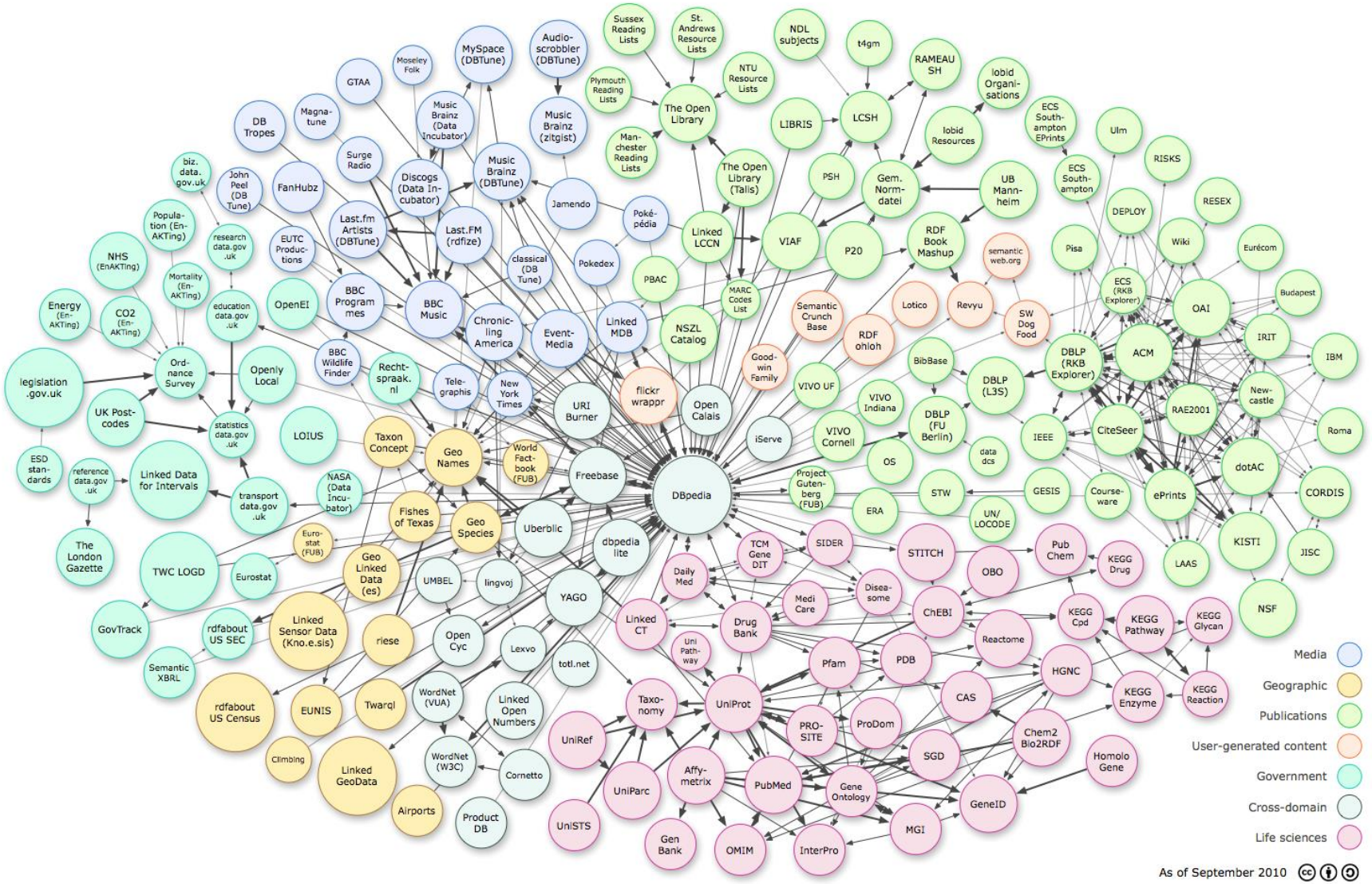
- 데이터 개체 간의 관계나 데이터 자체를 그래프로 표현하는 경우에 사용하는 데이터 세트
  (예: 웹 문서의 연결 관계나 화학 혼합물의 구조를 나타내는 경우에 사용)



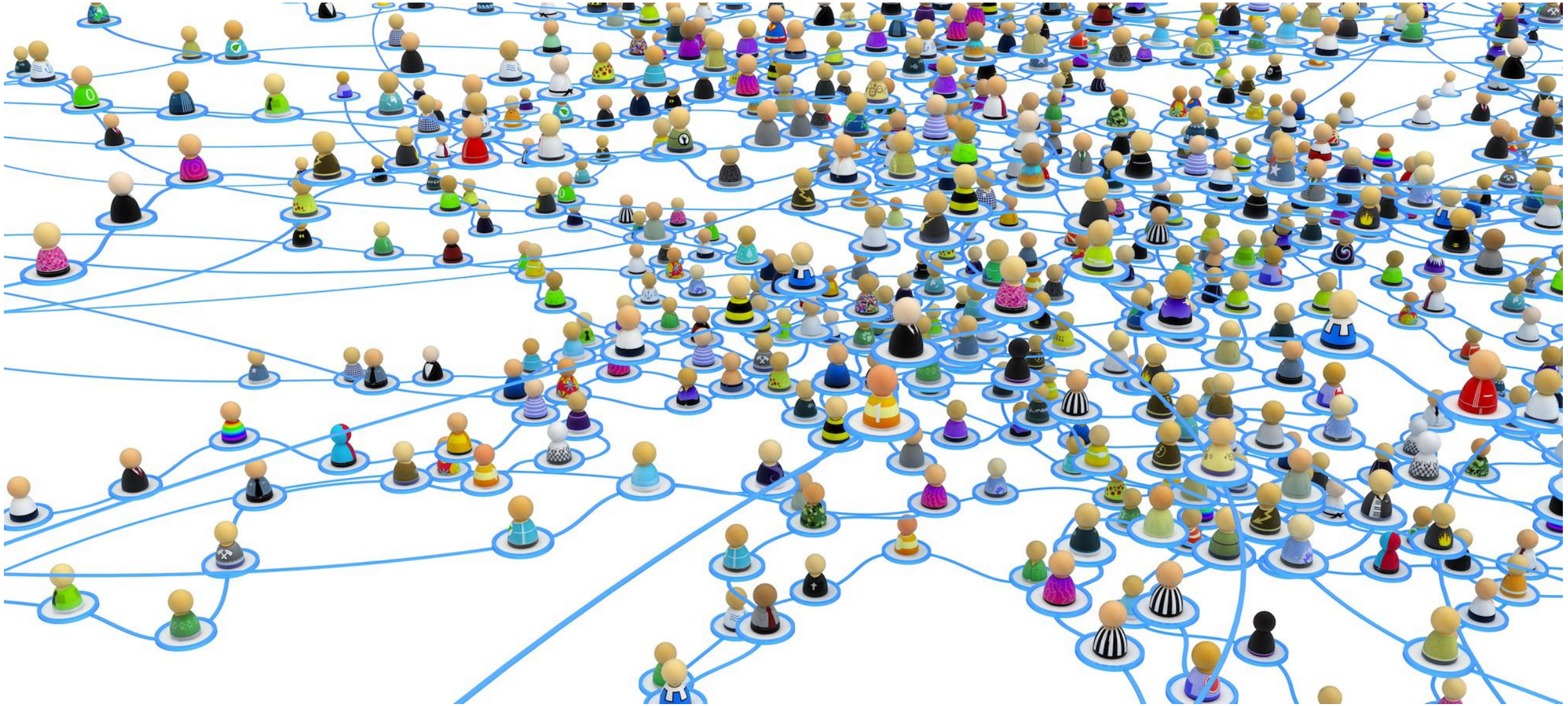
http://btechsmartclass.com/data_structures/introduction-to-graphs.html

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

그래프 데이터(Graph-based data)

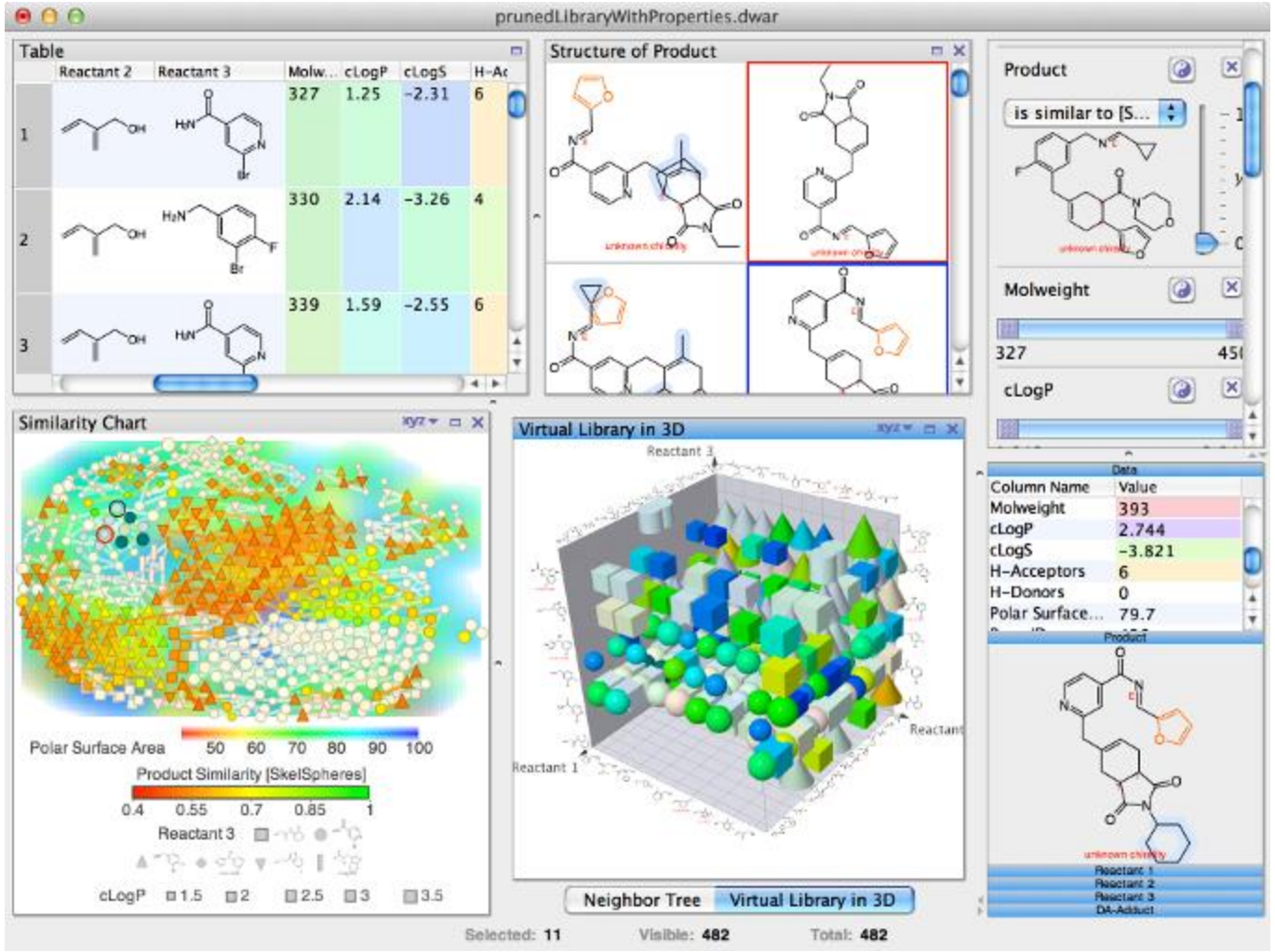


그래프 데이터(Graph-based data)



그래프 데이터(Graph-based data)

http://www.openmolecules.org



순서 데이터(Ordered data)

- 데이터 개체의 속성이 시간 또는 공간적인 순서와 연관되는 데이터 세트
- 순서 데이터의 종류
  - 연속 데이터(Sequential data)
  - 서열 데이터(Sequence data)
  - 시계열 데이터(Time series data)
  - 공간 데이터(Spatial data)

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

연속 데이터(Sequential data)

- 트랜잭션 데이터에서 시간 성분을 추가적으로 고려한 것
- 고객의 시간에 따른 구매 경향 예측과 같은 응용에서 사용 될 수 있음
- 예: CDP 구매 고객은 CD를 구매할 계획이 있음

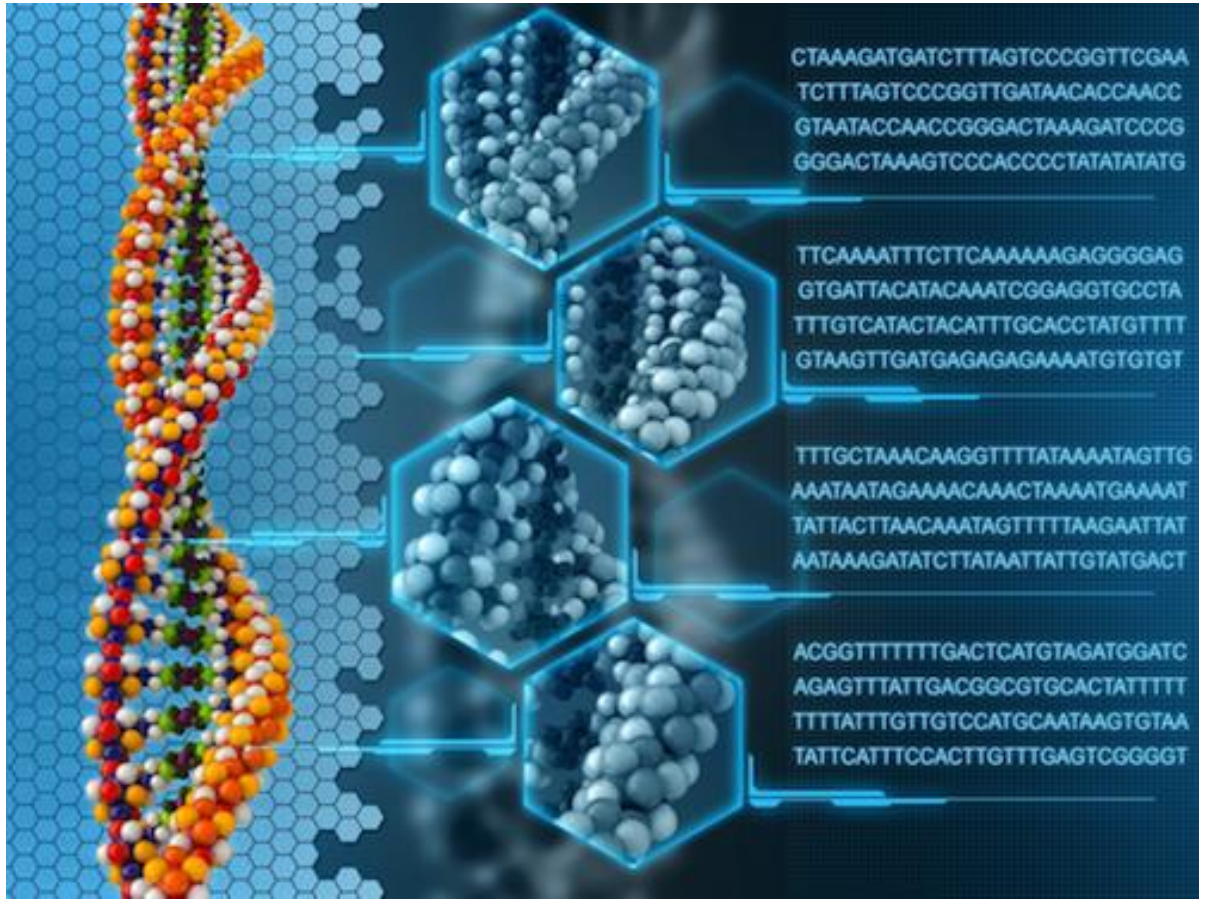
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A, B) (t2: C, D) (t5: A, E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

서열 데이터(Sequence data)

- 데이터 개체들 사이에 순서가 존재하는 데이터
- 예: DNA 서열
  A(아데닌), T(티아민), G(구아닌), C(사이토신)의 염기로 이루어져 있는 이중 나선형의 물질

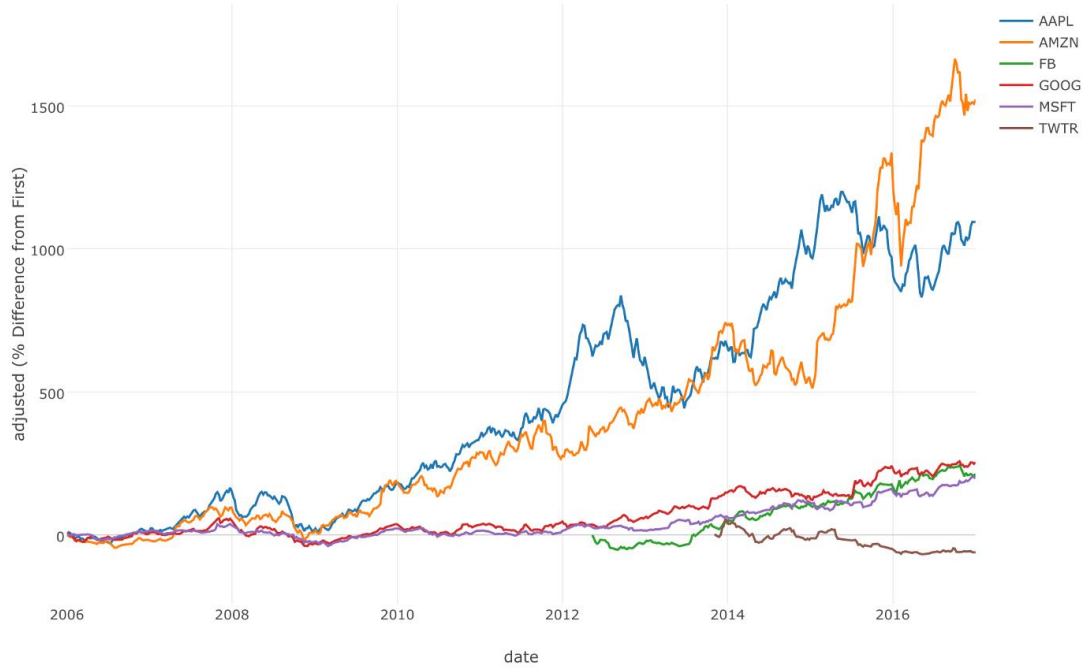


https://florence20.typepad.com/renaissance/2013/02/the-big-data-of-plant-genomics.html

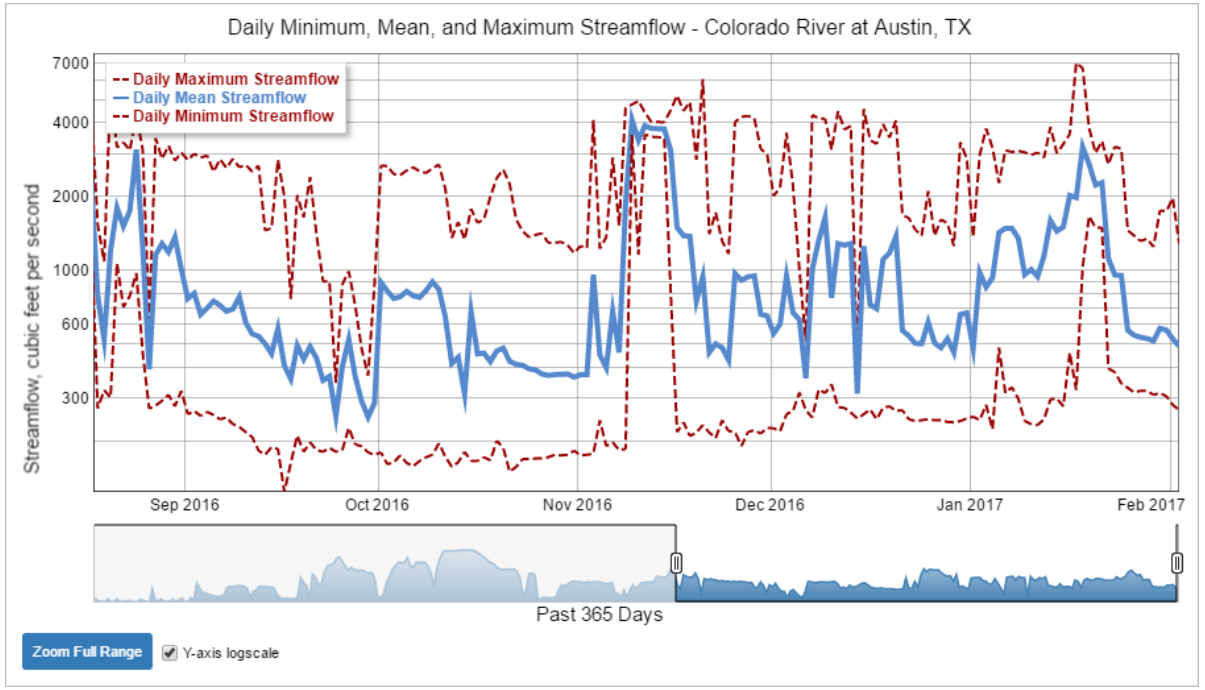
Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

시계열 데이터(Time series data)

- 연속 데이터(Sequential data)의 특수한 경우
- 시간에 따른 속성의 변화를 관찰한 데이터 집합
- 예: 주가 지수, 시간별 기온 변화



https://blog.exploratory.io/introduction-to-tidyquant-quantitative-financial-analysis-for-tidyverse-habitats-e5f72a023ce2

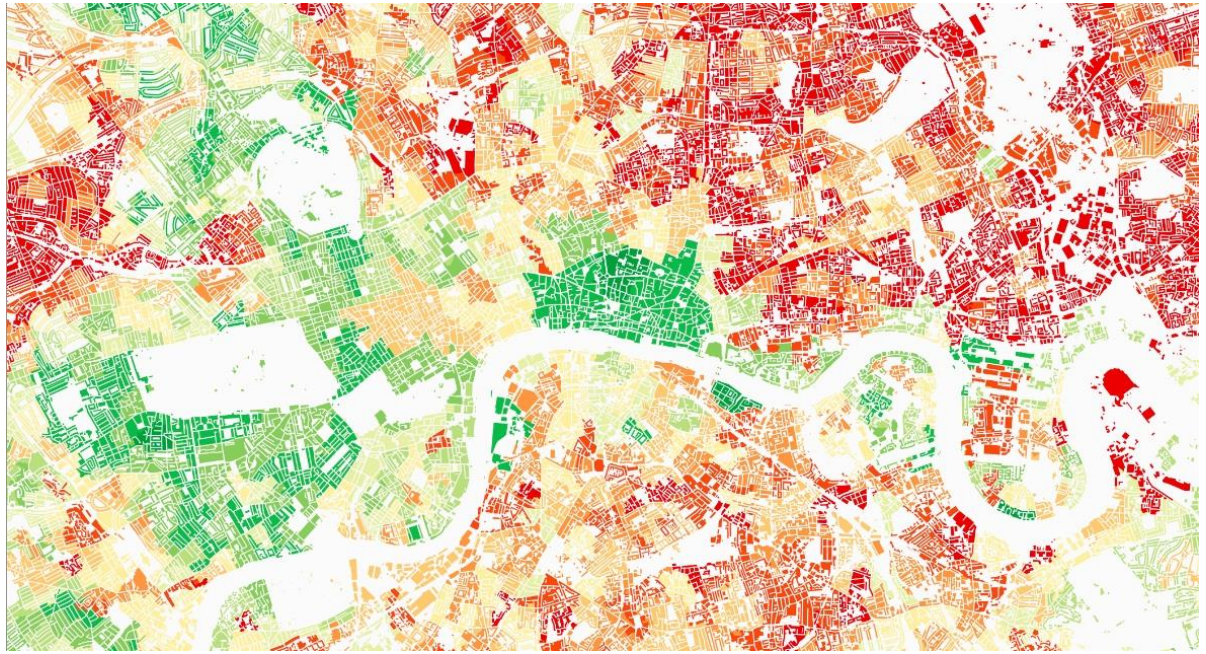


https://www.usgs.gov/media/images/time-series-data-usgs-station-colorado-river-austin

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

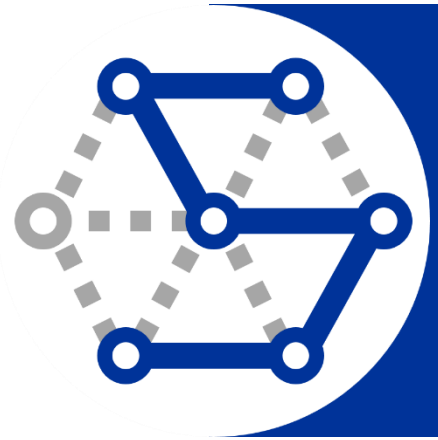
공간 데이터(Spatial data)

- 위성 사진 분석 데이터와 같이 각 데이터 개체가 공간 상의 위치 정보와 연관이 되는 데이터 집합
- 예: 지구 상의 지점에 따른 온도



http://spatial.ly/2013/08/big-open-data-mining-synthesis/

Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005



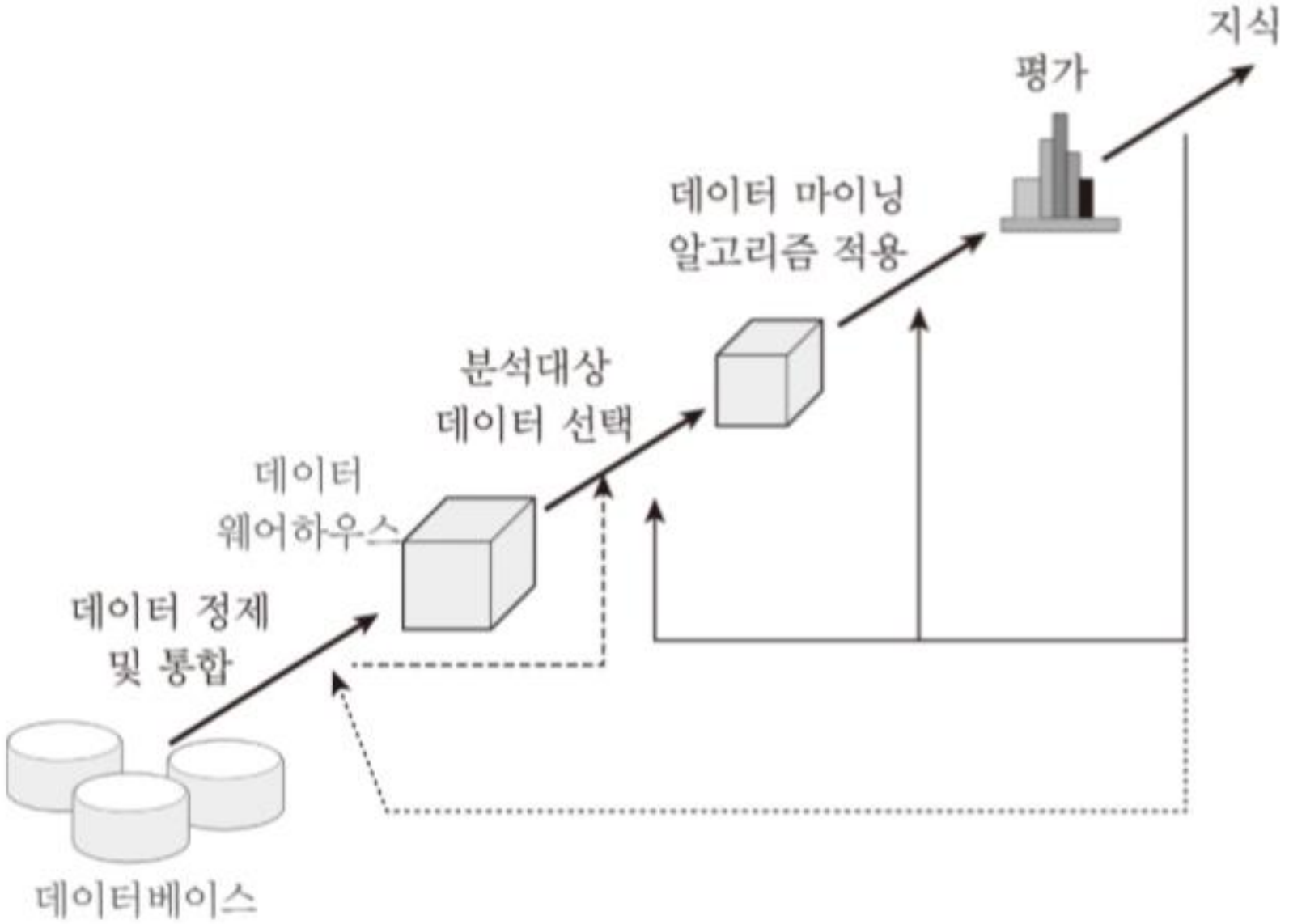
2 데이터마이닝

데이터마이닝(data mining)

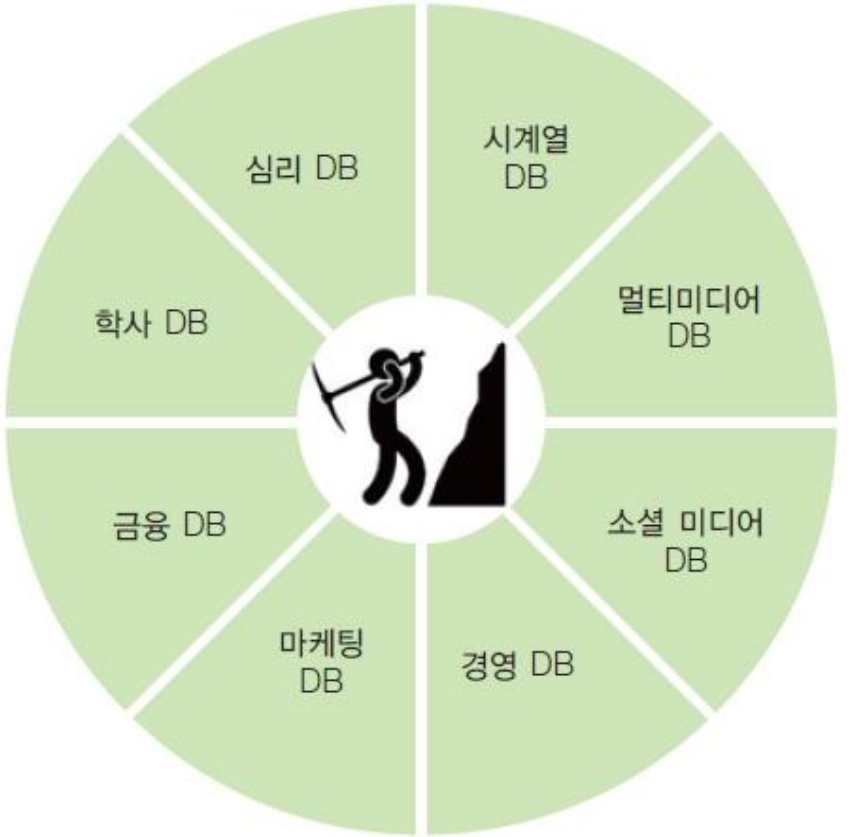
- 실제 대규모 데이터에서 암묵적인, 이전에 알려지지 않은, 잠재적으로 유용할 것 같은 정보를 추출하는 체계적인 과정
- a process of
  nontrivial extraction of implicit, ; 단순 DB 검색 수준이 아니고
  previous unknown, and ; 잘 알려진 지식을 찾는 것도 아니고
  potentially useful information ; 응용 분야에 따른 유용성이 있고
  from large volume of ; 처리 속도에 대한 고려가 필요하고 (디스크 접근)
  actual data ; 결손(missing)되거나 오류가 있는 데이터
- 데이터베이스에서 지식 발견(Knowledge Discovery in Database, KDD) 라고도 함
- 데이터 마이닝 기법은 통계학 분야에서 발전한 탐색적 데이터 분석, 가설 검증, 다변량 분석, 시계열 분석, 일반선형모형 등 방법론과 데이터베이스 분야에서 발전한 OLAP(On-Line Analytic Processing), 인공지능 분야에서 발전한 SOM(Self-Organizing Map), 신경망, 전문가 시스템 등 기술적 방법론 등에 사용

데이터마이닝(data mining)

데이터마이닝 과정

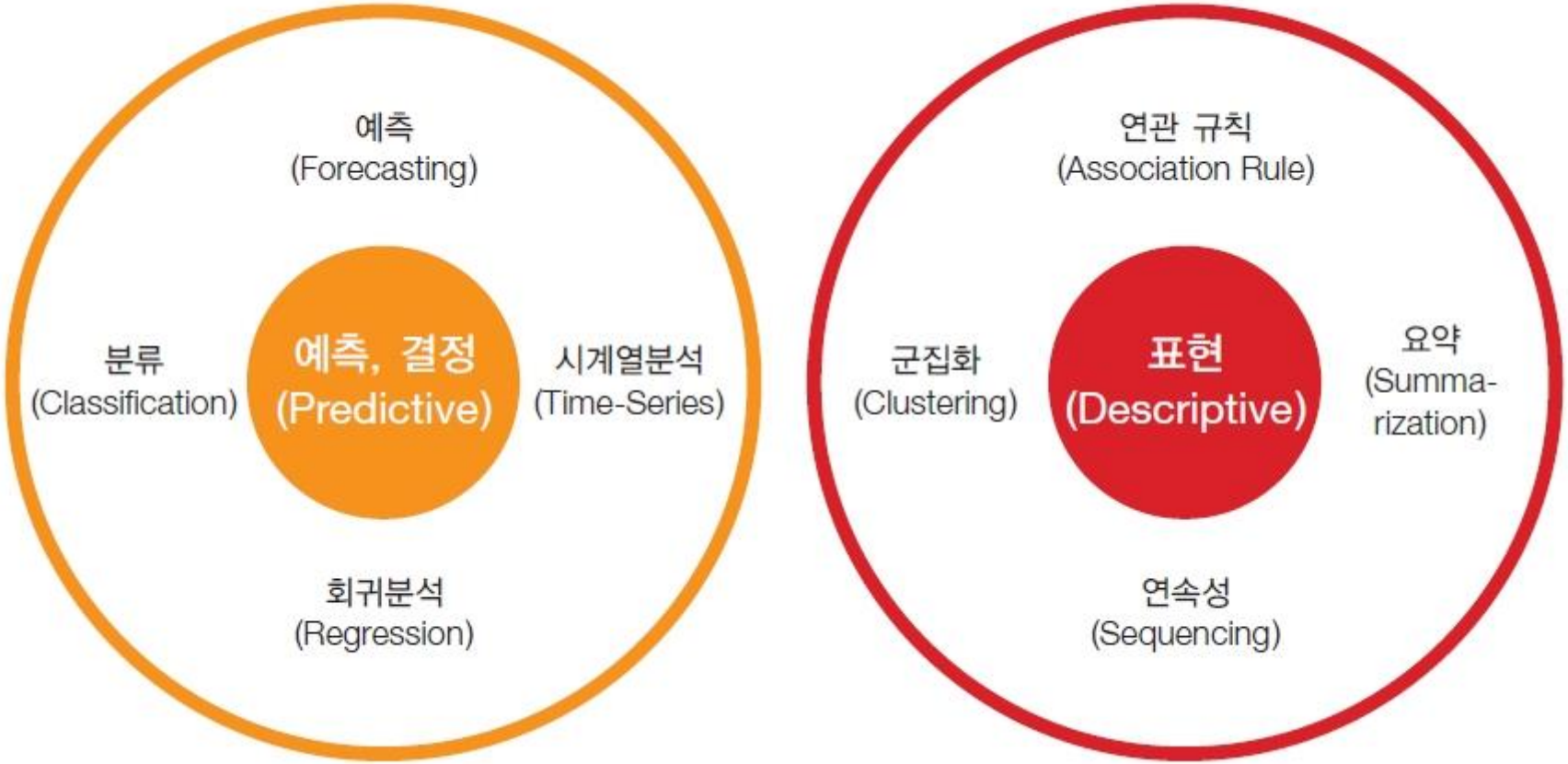


데이터마이닝(Data Mining)



각종 DB 속에 내재된 의미 있는 결과를
분석하고 이를 현실에 적용

데이터마이닝(Data Mining)

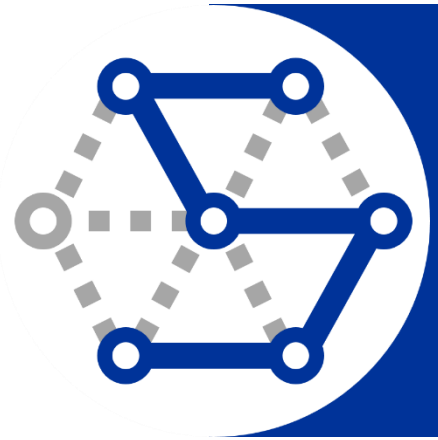


데이터마이닝(Data Mining)

- **회귀분석(Regression)** : 하나 이상의 변수 간의 영향이나 관계를 분석 및 추정하는 기술
- **연관 규칙(Association Rule)** : 동시에 발생한 사건 간의 관계를 정의
  - 예를 들어, 장바구니 안에 동시에 들어가는 상품들의 관계를 규명하는 기술
- **분류(Classification)** : 일정한 집단에서 특정한 정의를 이용하여 분류 및 구분을 추론
  - 예를 들어, 경쟁자에게로 이탈한 고객을 분류해 내는 기술
- **군집화(Clustering)** : 구체적인 특성을 공유하는 군집을 찾음. 군집화는 미리 정의된 특성의 정보가 없다는 점에서 분류와 다름
  - 예를 들어, 비슷한 행동 집단을 구분해 내는 기술

데이터마이닝(Data Mining)

- **시계열(Time-Series) 분석** : 시간의 변화에 따라 일정한 간격으로 연속적인 통계 숫자를 저장한 시계열 데이터에 바탕을 둔 분석 방법
  - 예를 들어, 매일 주식의 값을 저장하는 시계열 데이터를 분석하는 기술
- **연속성(Sequencing)** : 시간에 따라 순차적으로 나타나는 사건의 종속성을 말함
  - 예를 들어, A 제품을 구입한 고객이 향후 B 제품을 구입할 확률이라든가 작년의 계절적 매출 변동 요인과 올해의 매출 등을 알아내는 기술
- **요약(Summarization)** : 데이터의 일반적인 특성이나 특징의 요점을 간략히 정리하는 기술
- **예측(Forecasting)** : 방대한 양의 데이터 집합의 패턴을 기반으로 미래를 예측
  - 예를 들어, 수요를 예측하는 기술



3 연관 규칙 마이닝

연관규칙(Association Rule)

- 시장바구니 데이터에 내재된 항목 출현의 연관성을 찾아 표현한 규칙
- **항목집합**(itemset): 한 개 이상의 항목들의 모음
- **k-항목집합**(k-itemset): k개의 항목을 포함하는 항목집합 ex. 2-itemset : {기저귀, 맥주}
- **지지회수**(support count): 항목집합이 나타난 거래 데이터의 수
- **지지도**(support): 전체 데이터에 대한 항목집합 를 포함한 거래 데이터의 비율

$$s(A) = \frac{\text{항목집합 } A\text{의 지지 회수}}{\text{전체 거래 데이터의 수}}$$

거래번호	구매 항목
T1	빵, 우유
T2	기저귀, 달걀, 맥주, 빵
T3	기저귀, 맥주, 우유, 콜라
T4	기저귀, 맥주, 빵, 우유
T5	기저귀, 빵, 우유, 콜라

$s(\{\text{기저귀, 우유}\}) = 3/5 = 0.6$

연관규칙(Association Rule)

- **빈발 항목집합**(frequent itemset): 미리 지정된 값 이상의 지지도를 갖는 항목집합
- 연관 규칙 A → B 의 **신뢰도**(confidence) c(A → B)

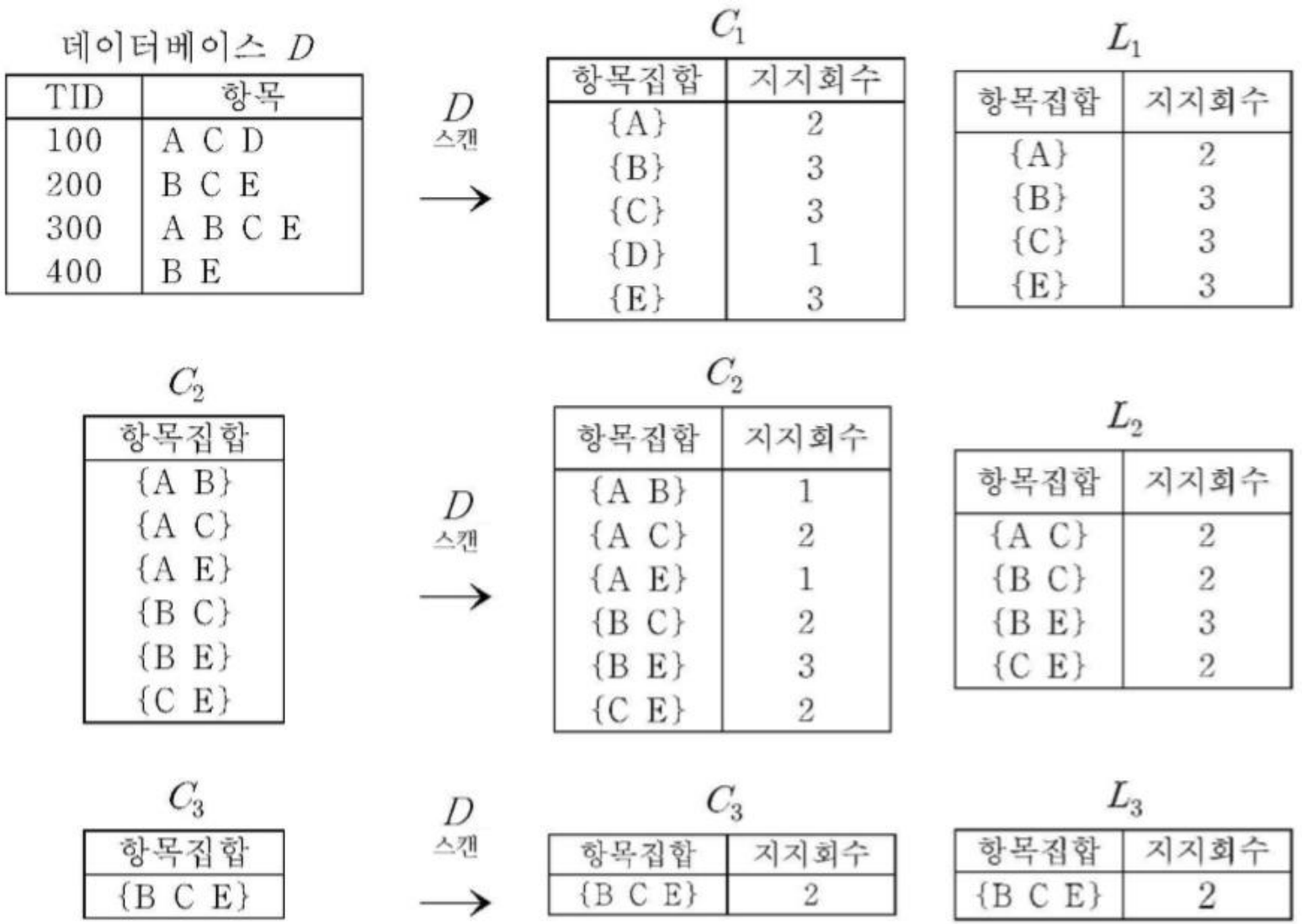
$$c(A \rightarrow B) = \frac{s(A \cup B)}{s(A)}$$

거래번호	구매 항목
T1	빵, 우유
T2	기저귀, 달걀, 맥주, 빵
T3	기저귀, 맥주, 우유, 콜라
T4	기저귀, 맥주, 빵, 우유
T5	기저귀, 빵, 우유, 콜라

$$c(\{기저귀, 우유\} \rightarrow \{콜라\}) = \frac{s(\{기저귀, 우유, 콜라\})}{s(\{기저귀, 우유\})} = \frac{2}{3} = 0.667$$

Apriori 알고리즘

- **대표적인 연관규칙 마이닝 알고리즘**
- 빈발 항목집합들 선택 ➔ 이들로 부터 신뢰도가 높은 연관 규칙 생성
- 빈발 항목집합을 점진적으로 생성



연관 규칙 마이닝의 응용 분야

비즈니스 분야

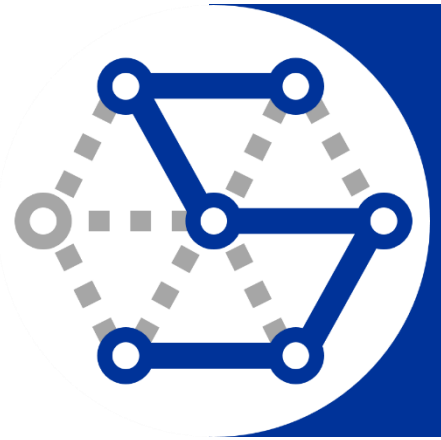
- 상품 카탈로그 설계, 끼워팔기 전략 수립, 매장 상품배치, 쿠폰 설계, 구매 패턴에 따른 고객 세분화 등

통신 서비스

- 고객의 사용 패턴 추출, 동시에 발생하는 장애 패턴 분석 등

사회 안전

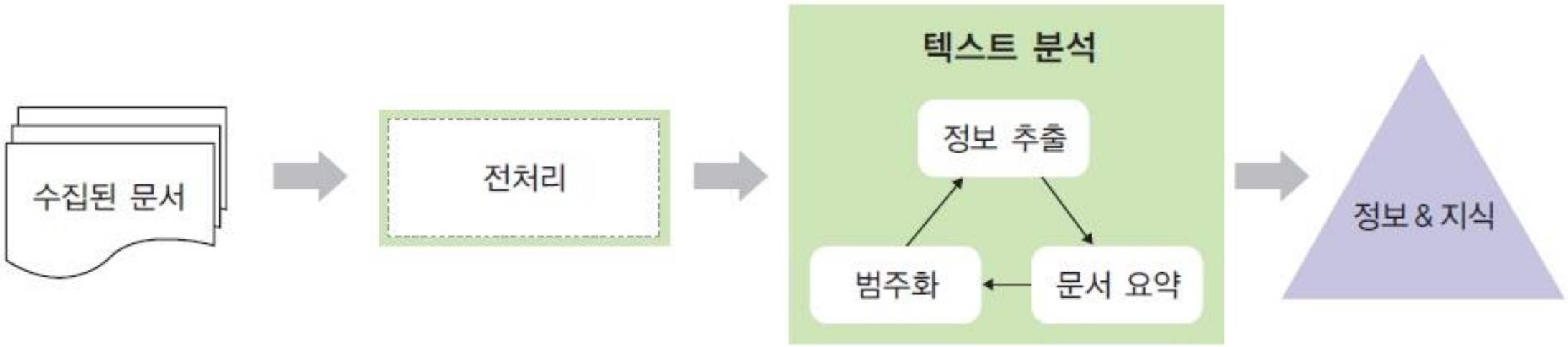
- 범죄 및 테러 모의, 불법 유통, 범죄 네트워크 식별, 범죄 추적 등



4 텍스트 마이닝

텍스트 마이닝(text mining)

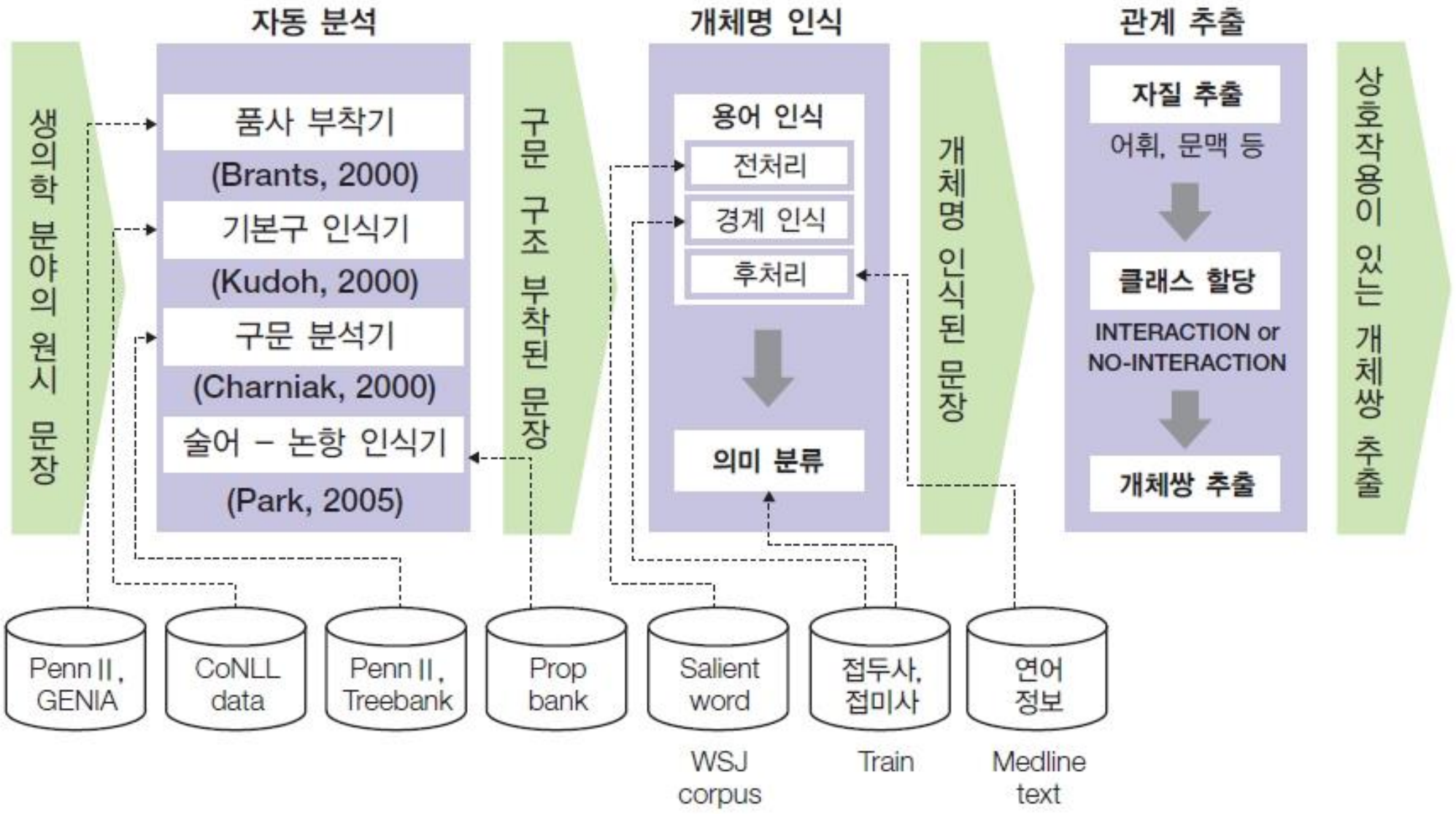
- 대규모 텍스트 데이터로부터 의미 있는 정보를 추출하는 것
- 비정형인 텍스트 데이터를 **자연어 처리** 기법으로 **문장 분석**
- **의미요소**를 **추출**하고 구조화된 **정형 데이터**로 변환
- 변환된 정형 데이터에 **데이터 마이닝** 기법을 적용하여 의미 있는 패턴을 추출



텍스트 마이닝 절차

- 정보 수집 : 비 ·반정형의 텍스트 데이터를 수집하는 단계
- 정보 처리 : 대용량의 데이터에서 특정 키워드나 일부 의미 있는 요소를 추출하려고 전처리를 하는 단계
- 정보 추출
  - 수학적인 모델이나 알고리즘을 이용하여 유용한 정보를 추출
  - 텍스트 마이닝을 위한 정보추출 방법에는 다양한 목적, 조건, 환경 등이 존재
  - 정보 추출 방법은 텍스트 마이닝에서 가장 중요한 부분 중 하나
  - 정보 추출 방법에는 수많은 수학적 알고리즘과 방법이 있으며, 그 중 간단하면서 가장 강력한 방법인 TF-IDF(Term Frequency-Inverse Document Frequency) 방식을 많이 사용
- 정보 분석 : 최종 키워드나 의미 있는 요소의 우선순위를 도출하는 단계

텍스트 마이닝 기술


자동 분석
개체명 인식
관계 추출
생의학 분야의 원시 문장
품사 부착기
(Brants, 2000)
기본구 인식기
(Kudoh, 2000)
구문 분석기
(Charniak, 2000)
술어 – 논항 인식기
(Park, 2005)
구문 구조 부착된 문장
용어 인식
전처리
경계 인식
후처리
의미 분류
개체명 인식된 문장
자질 추출
어휘, 문맥 등
클래스 할당
INTERACTION or
NO-INTERACTION
개체쌍 추출
상호작용이 있는 개체쌍 추출
Penn II, GENIA
CoNLL data
Penn II, Treebank
Prop bank
Salient word
WSJ corpus
접두사, 접미사
Train
연어 정보
Medline text

텍스트 마이닝 대상

텍스트 분류
(text classification)

텍스트 군집화
(text clustering)

개념 추출
(concept extraction)

개체 관계 모델
(entity relation modeling)

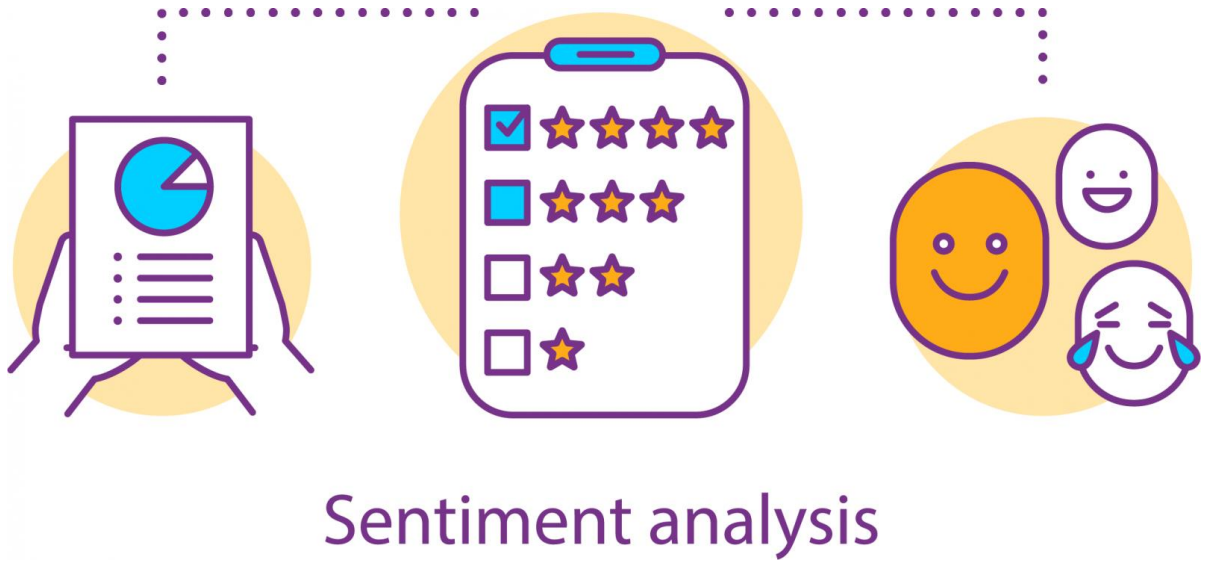
문서 요약
(document summarization)

토픽 모델링
(topic modeling)

감성분석
(sentiment analysis)

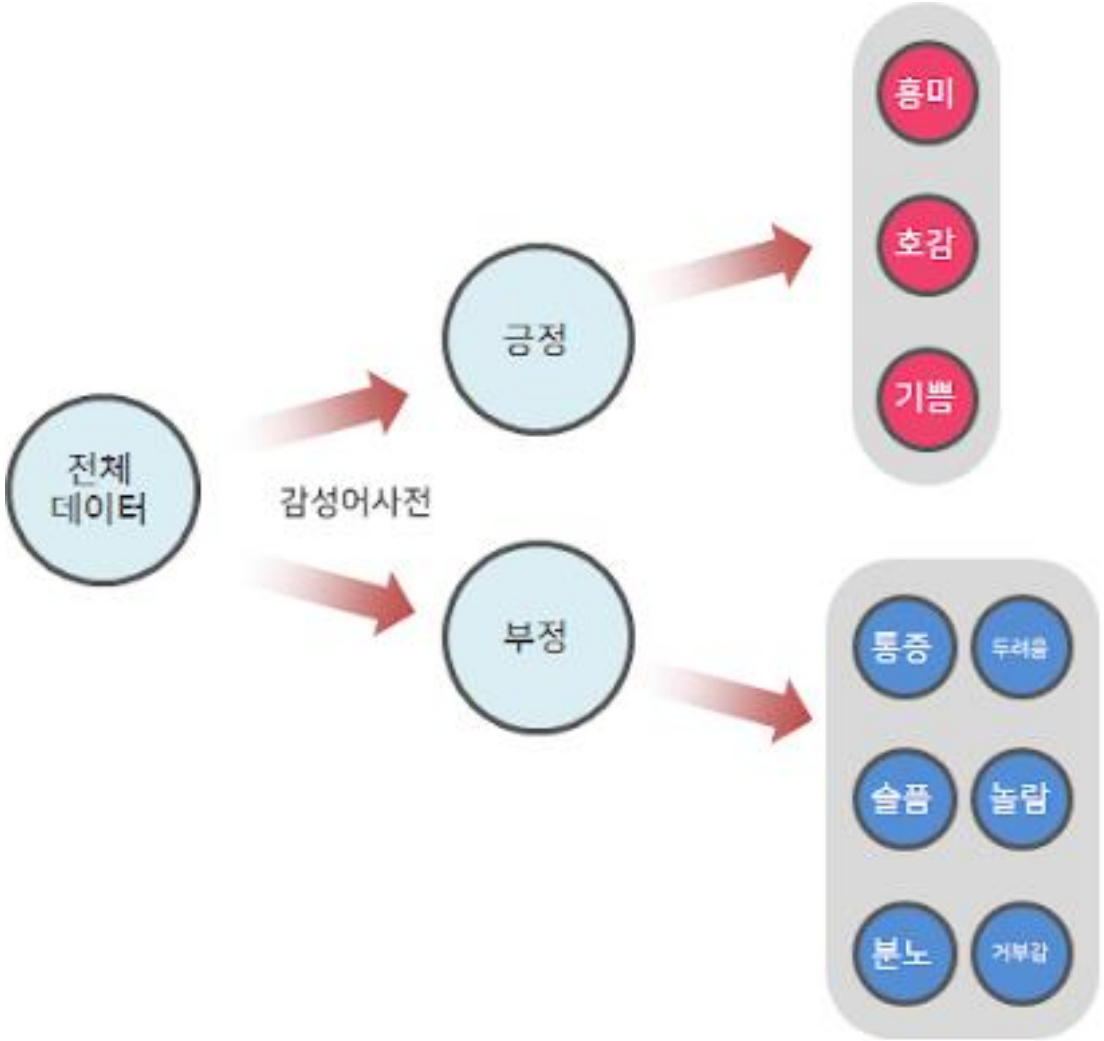
감성분석(sentiment analysis, 오피니언 마이닝: opinion mining)

- 텍스트를 작성한 사람 또는 문장의 주어에 해당하는 사람의 기분, 긍정 또는 부정 의견을 추출
- 특정 이슈, 인물, 이벤트에 대한 사람들의 평가, 태도, 감정을 분석하는 것



감성분석 단계

- **특징 추출 단계**
  - 감성사전 구축
- **긍정 부정 의견 분류 단계**
  - 문장 단위로 세부 평가 요소와 긍정 부정 표현 식별
  - 해당 문장이 긍정적인 의견인지 부정적인 의견인지 결정
- **오피니언 요약 및 전달 단계**
  - 각 세부 평가 요소별로 긍정 표현과 부정 표현의 차이 비교
  - 긍정 부정의 평가 정도 표현
  - 평가 요소별로 오피니언에 대한 요약정보를 문장으로 표현



오피니언 마이닝 기술

- 긍정 및 부정을 표현하는 단어 정보 추출
  - 기존에 구축된 사전 등 리소스를 이용하거나 수작업으로 해당 도메인의 고빈도 긍정/부정을 표현하는 단어들을 확인할 수 있음
  - 학습 데이터에서 유용한 통계 정보를 활용하여 자동으로 어휘 정보를 얻을 수도 있음
  - 한국어로 작성한 사용자 별점이 아주 높은 리뷰에서는 고빈도 단어를 긍정 표현으로, 사용자 별점이 아주 낮은 리뷰에서는 고빈도 단어를 부정 표현으로 추출할 수 있음
- 세부 평가 요소와 그것이 가리키는 오피니언의 연결 관계를 포함한 문장 인식
  - 첫 번째 단계에서 구축된 어휘 정보를 사용하여 세부 평가 요소와 긍정/부정 표현을 찾으며 긍정적인 오피니언 인지, 부정적인 오피니언 인지 문장 단위로 분류하려고 여러 방법을 적용할 수 있음
  - 대량의 레이블이 부착된 학습 데이터를 생성한 후 Naive Bayes, ME(Maximum Entropy) 모델, SVM(Support Vector Machine)과 같은 알고리즘을 적용하여 기계 학습을 수행

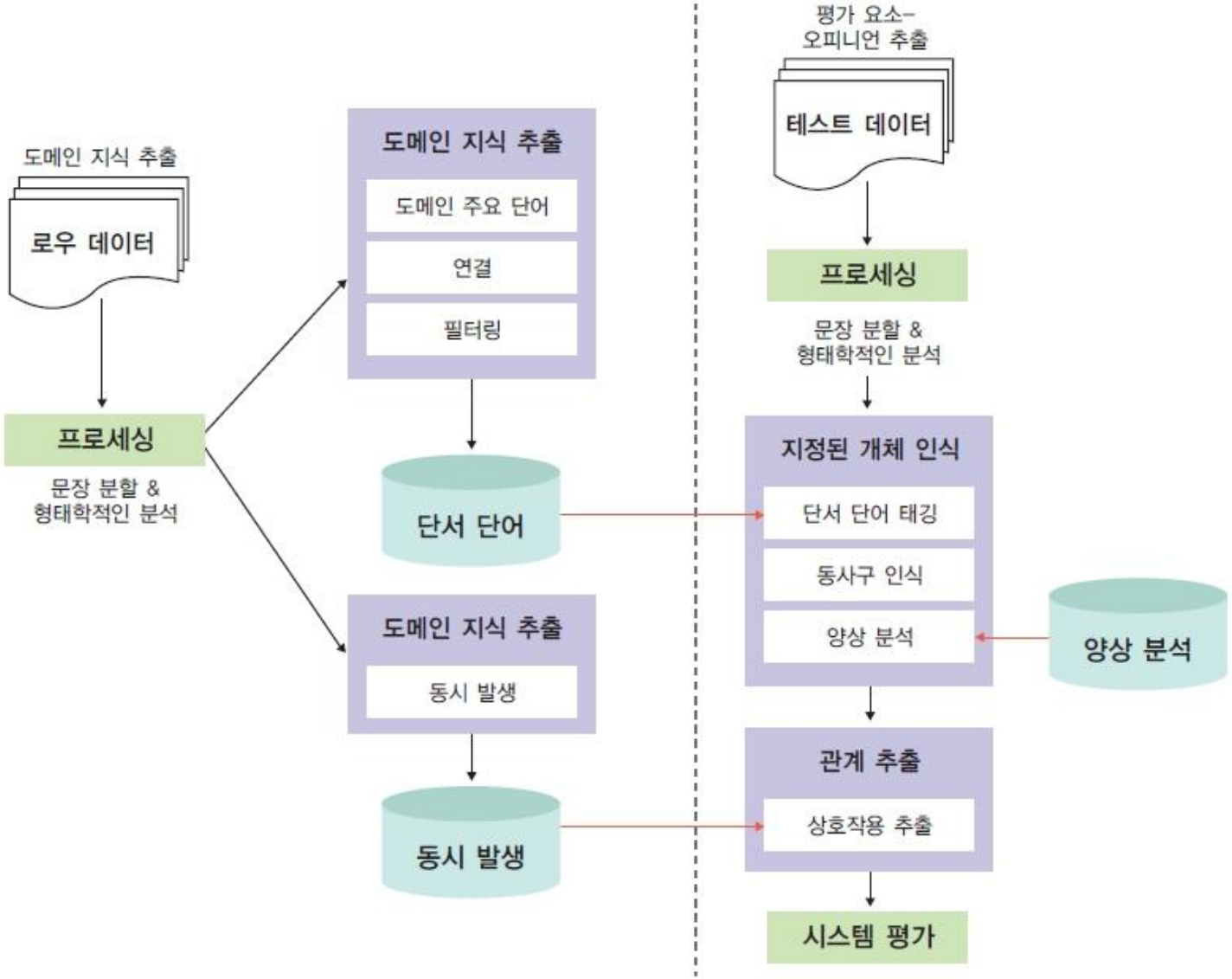
오피니언 마이닝 기술

- 긍정/부정 표현의 수 및 유용한 문장을 추출하여 리뷰 요약 생성
  - 각 세부 평가 요소에서 긍정 표현과 부정 표현의 차이를 이용하여 사용자들의 선호도를 제시할 수 있음
  - 오피니언 마이닝 결과를 이용하여 특정 맛집의 세부 평가 요소에서 좋아하거나 싫어하는 정도를 얻을 수 있음

전체 평가	맛	서비스	분위기	가격
• 긍정 : 80% • 부정 : 20%	• 긍정 : 60% • 부정 : 40%	• 긍정 : 70% • 부정 : 30%	• 긍정 : 55% • 부정 : 45%	• 긍정 : 80% • 부정 : 20%
식재료	**양**	**위생**	**주차**	**대표 메뉴**
• 긍정 : 60% • 부정 : 40%	• 긍정 : 70% • 부정 : 30%	• 긍정 : 55% • 부정 : 45%	• 긍정 : 70% • 부정 : 30%	• 긍정 : 55% • 부정 : 45%

맛집의 평가 요소 – 가격	
긍정 평가	▪ 순수함 그 자체. 육수도 서비스로 한 사발 더 주신 주인아저씨도 친절하시고 가격대비 대만족^^ ▪ 아줌마들이 모임하기에 안성맞춤이고, 가격도 적당하며 무엇보다 자극적이지 않은 음식... ▪ 분위기, 서비스, 가격 등 레스토랑으로서 손색이 없어 글을 올립니다.
부정 평가	▪ 굉장히 유명한 편인데— 가격도 싼 편은 아닌데— 그에 비해 맛은 쏘쏘— 서비스는 중하 정도? ▪ 가격만 비싸지고 맛은 조미료 범벅. 어렸을 때 먹었던 그 맛이 그립네요. ▪ 가격대비 별로 먹을 것 정말 없고요.

오피니언 마이닝 기술

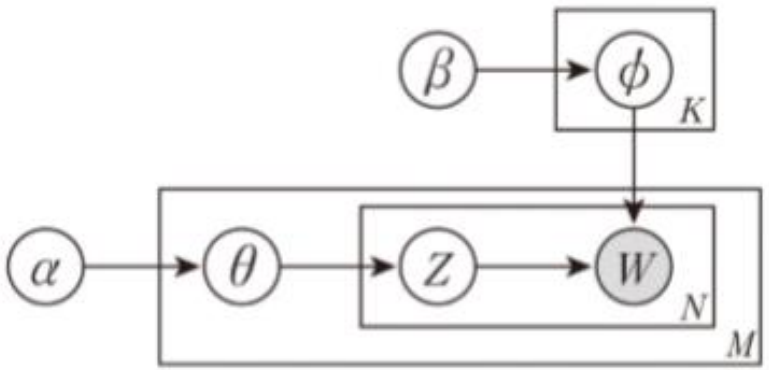

도메인 지식 추출
로우 데이터
프로세싱
문장 분할 &
형태학적인 분석
도메인 지식 추출
도메인 주요 단어
연결
필터링
단서 단어
도메인 지식 추출
동시 발생
동시 발생
평가 요소-
오피니언 추출
테스트 데이터
프로세싱
문장 분할 &
형태학적인 분석
지정된 개체 인식
단서 단어 태깅
동사구 인식
양상 분석
양상 분석
관계 추출
상호작용 추출
시스템 평가

오피니언 마이닝 기술

- 소셜 네트워크 서비스에는 사람들의 네트워크 관계와 대화 내용 속에 오피니언이 포함
- 주제나 대상, 인물이 특정 부분에 국한되지 않고 다양한 오피니언이 많기 때문에 오피니언 마이닝 기술을 활용

토픽 모델링(topic modeling)

- 문서들에서 어떤 주제들이 다루어지고 있는지, 각 문서는 어떤 주제들을 다루고 있는지를 비지도학습적인 방법으로 찾아내는 작업
  - 다수의 문서가 주어질 때 각 문서가 여러 주제(topic)를 다루고 있다고 전제하고, 이들 주제의 반영비율 결정
  - 주제를 구성하는 단어들의 확률 분포 결정
- **LDA(Latent Dirichlet Allocation) 알고리즘**
  - 대표적인 토픽 모델링 알고리즘

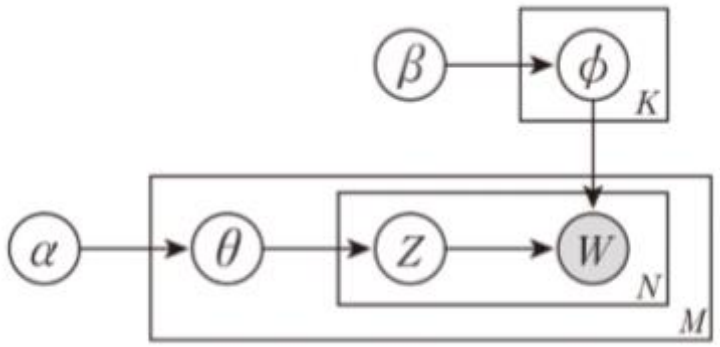


사각형 안의 원: 확률변수(random variable)
사각형 밖의 원: 하이퍼파라미터(hyperparameter)
사각형 우측 하단의 기호 : 확률변수의 반복 개수

K : 전체 주제의 개수
M : 전체 문서의 개수
N : 각 문서에 있는 단어의 개수

LDA(Latent Dirichlet Allocation) 알고리즘

- **주제(topic) ϕ의 표현**
- **다항 분포**로 표현된다고 가정



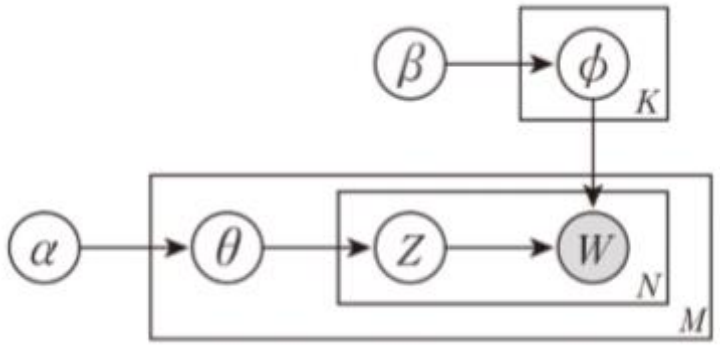
야구 = {(KBO, 0.1), (류현진, 0.05), (결승, 0.05), (선발, 0.1), (마무리, 0.01), (홈런, 0.3), (이닝, 0.1), (홈, 0.05), (송구, 0.05), (승부, 0.09), (우중간, 0.02), (적시타, 0.03), (선취점, 0.05)}

자동차 = {(시속, 0.05), (사고, 0.03), (타이어, 0.1), (출력, 0.05), (드라이브, 0.2), (모델, 0.1), (F1, 0.02), (전설, 0.01), (엔진, 0.1), (출전, 0.05), (라이업, 0.1), (마력, 0.03), (가솔린, 0.1), (연비, 0.1), (정숙성, 0.04), (변속기, 0.1)}

- 하이퍼파라미터 β
  - 디리쉬리 분포(Dirichlet distribution; 디리클레 분포)의 특성을 지정하는 파라미터
  - 디리쉬리 분포 : 다항분포들에 대한 확률분포

LDA(Latent Dirichlet Allocation) 알고리즘

- θ: 문서의 주제 분포를 나타내는 확률변수
  - 각 문서는 여러 주제를 부분적으로 다루는 것으로 가정

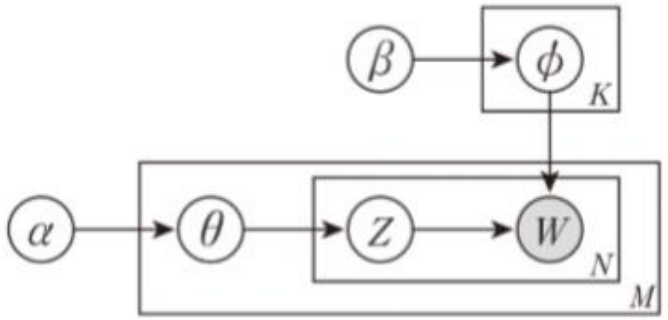


문서1 = {(정치, 0.3), (연예, 0.4), (정보기술, 0.1), (미래예측, 0.2)}
문서2 = {(자동차, 0.2), (야구, 0.3), (연예, 0.1), (경제, 0.2), (법률, 0.2)}

  - 확률변수 Z : 문서의 주제 분포에서 표본추출된 주제 하나를 표현
  - W : 실제 문서에 나타나는 단어
- Z 가 가리키는 주제 ϕ의 단어 분포로부터 표본추출하여 단어 W 가 결정된다고 가정

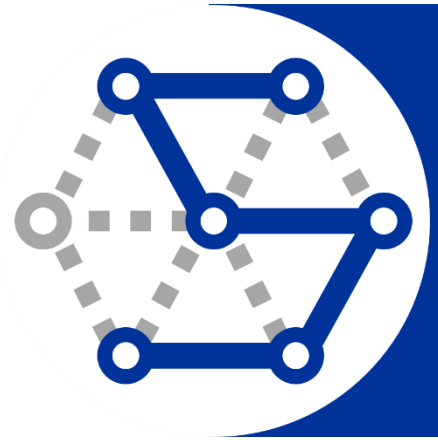
LDA(Latent Dirichlet Allocation) 알고리즘

- 학습의 목적
  - 문서의 주제 분포와 주제의 단어 분포 결정



  - Z 값을 결정되면 문서의 주제 분포 및 주제의 단어 분포 결정 가능

- Z의 값을 결정하는 대표적인 방법
  - **깁슨 표본추출**(Gibbs sampling) 이용 방법



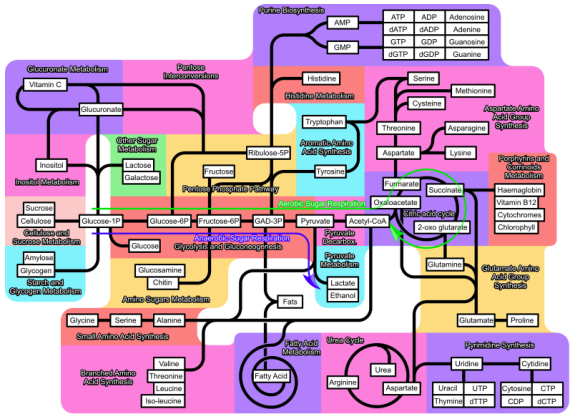
5 그래프 마이닝

그래프 마이닝(graph mining)

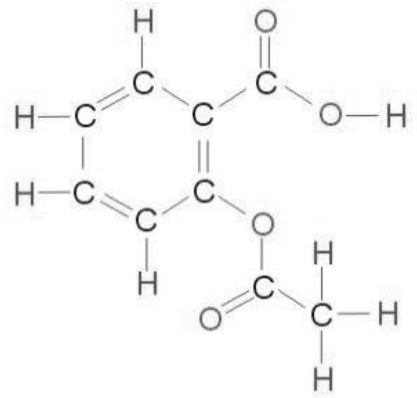
- 그래프 데이터들로부터 의미 있는 패턴을 추출하는 것

컴퓨터 네트워크

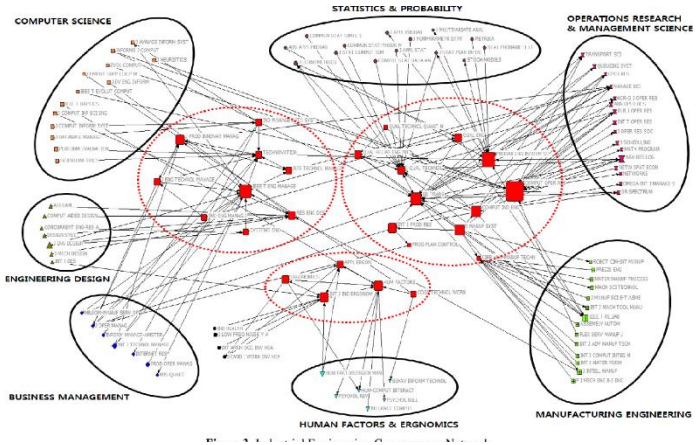
소셜 네트워크



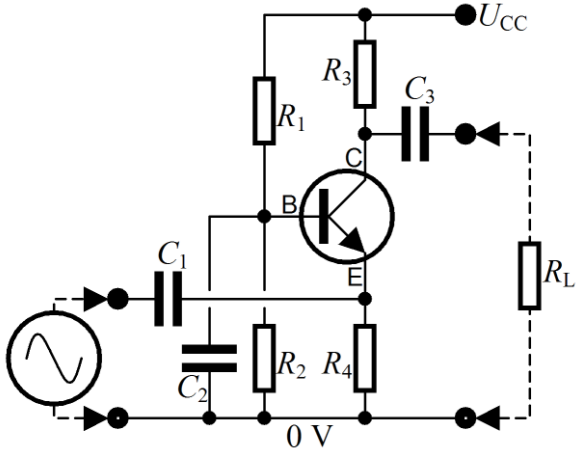
대사경로 네트워크



화학 구조식



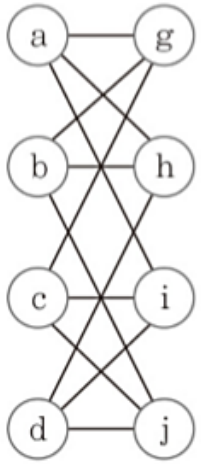
인용 네트워크



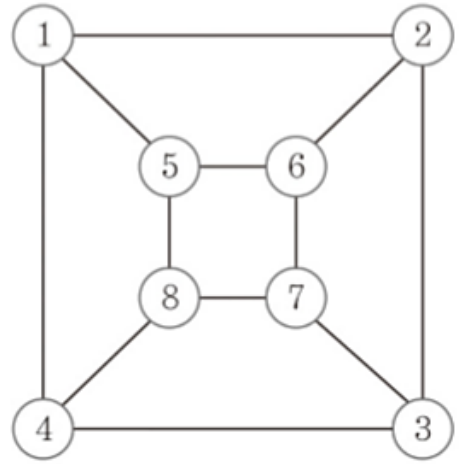
전자 회로도

빈발 부분그래프(frequent subgraph) 마이닝

- 그래프 데이터들의 집단에서 자주 나타나는 **부분그래프**(subgraph)를 찾는 것
- **부분그래프**
  - 원래 그래프에 속하는 (즉, 노드와 링크가 포함되는) 그래프
- **동형**(同型, isomorphism)
  - 한 그래프의 노드들의 이름을 다른 그래프의 노드들의 이름으로 변환하여 같은 그래프로 만들 수 있는 것
  - 동형 확인
    - NP-hard 문제로서 비용이 많이 드는 연산



(a)



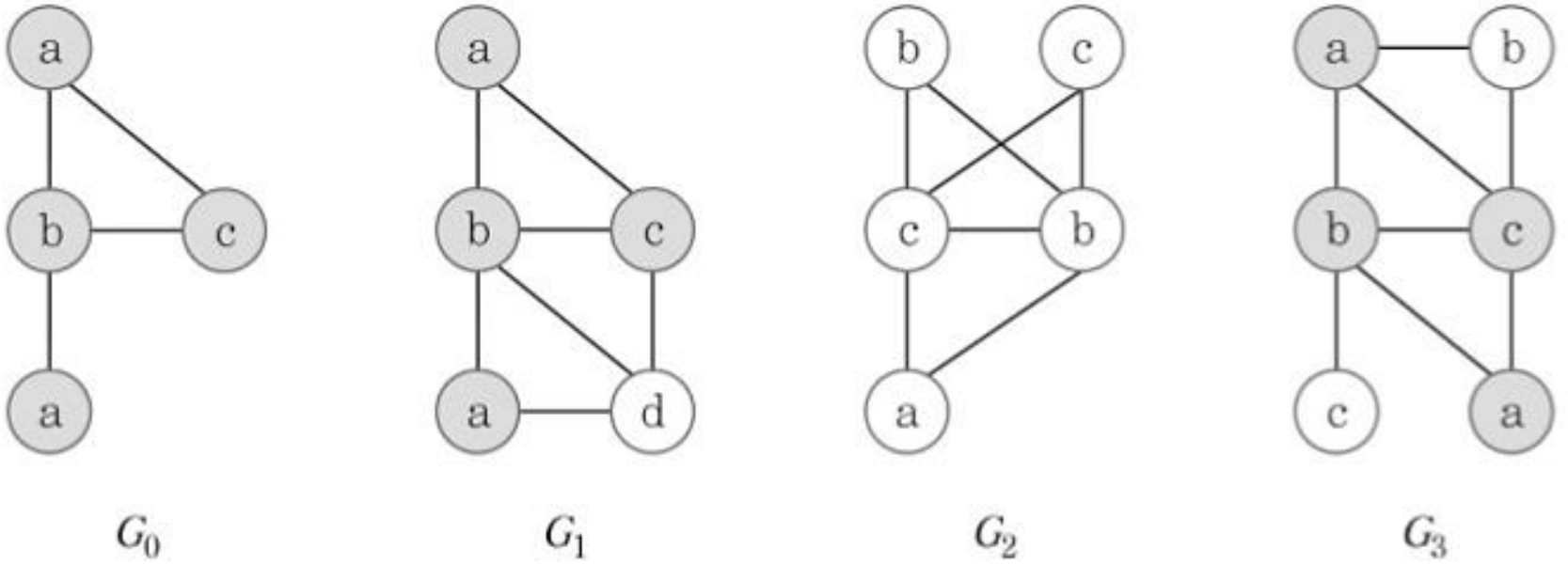
(b)

$f(a)=1$
$f(b)=6$
$f(c)=8$
$f(d)=3$
$f(g)=5$
$f(h)=2$
$f(i)=4$
$f(h)=7$

(c)

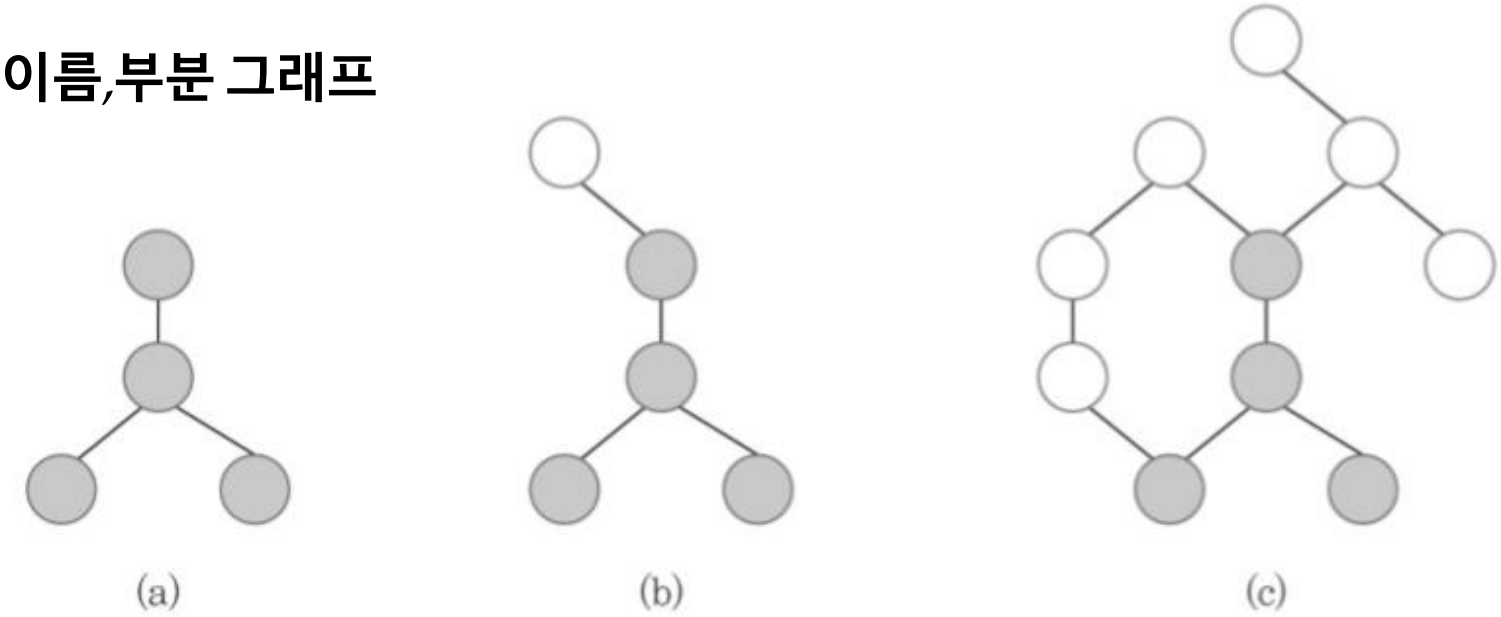
빈발 부분그래프(frequent subgraph) 마이닝

- **후보 부분그래프**를 만들어가면서, 후보 부분그래프가 각 그래프 데이터에 포함되는지 **동형 확인을** 통해 빈발한 것들을 찾는 것
- 대표적인 빈발 부분그래프 마이닝 알고리즘
  - FSG, gSpan



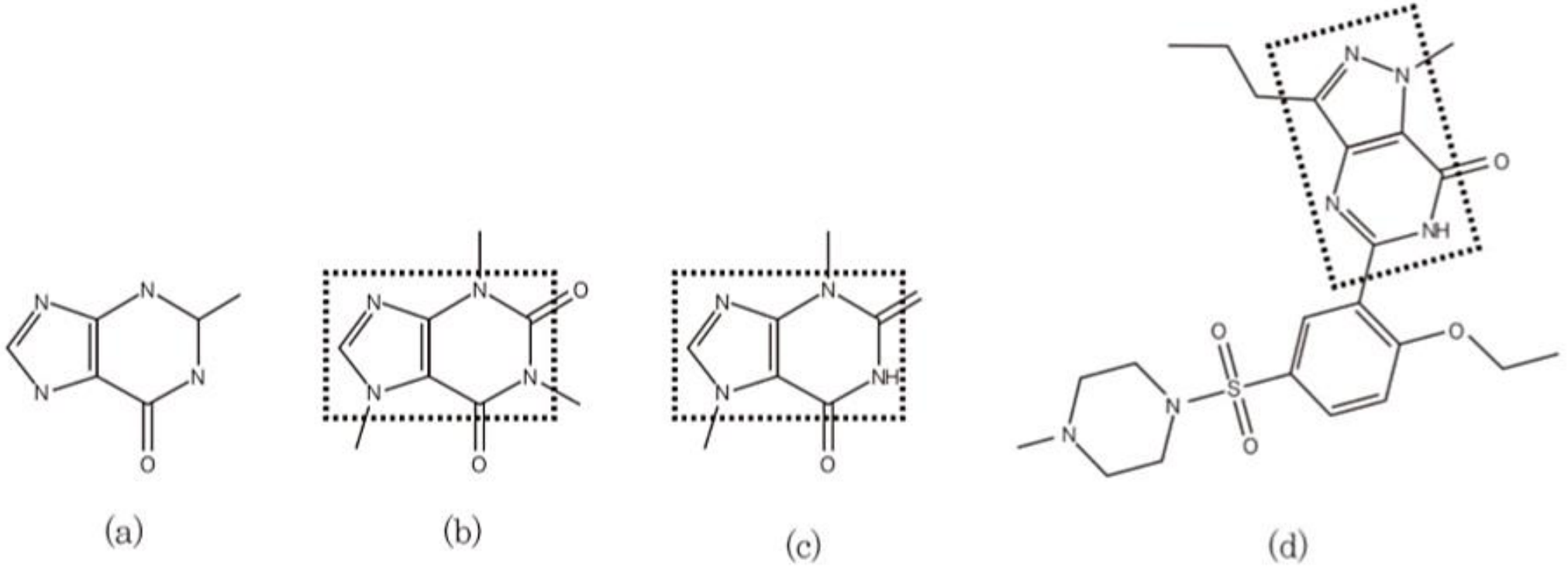
그래프 검색(graph search)

- **그래프 데이터베이스**와 **질의 그래프**(query graph)가 주어질 때,
  질의 그래프를 부분그래프로 포함하는 그래프들을 데이터베이스에서 모두 찾는 문제
- 검색 문제를 효과적으로 다루기 위해 **인덱싱**(indexing) 기법 사용
- **인덱싱:** 어떤 대상을 쉽게 찾을 수 있도록, 기준이 되는 특징으로 대상의 위치를 쉽게 찾을 수 있게 하는 정보를 관리하는 기법
- 인덱싱에 사용할 수 있는 **특징**
  - **경로**, **부분구조**, **노드** 또는 **링크**의 **이름**,**부분 그래프**
  - 주로 **빈발 부분그래프** 사용



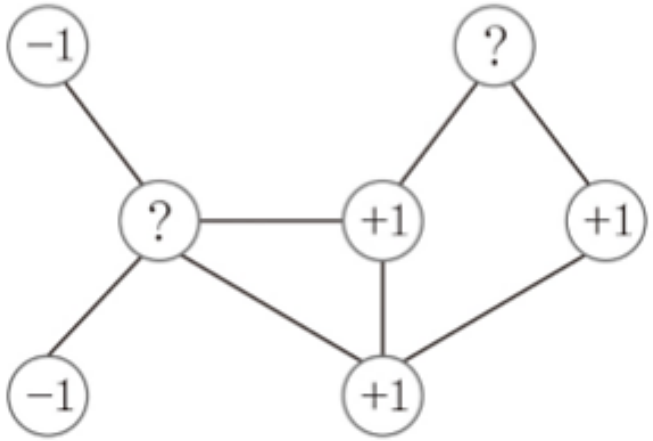
유사도 검색(similarity search)

- 인덱싱을 할 때 여러 가지 특징을 사용하여 각 그래프 데이터를 **특징벡터**로 표현
- **질의 그래프**에 대해서 **인덱싱**할 때 사용한 **특징**을 추출하여 **벡터**로 표현
- 질의 그래프의 특징벡터와 특징벡터의 **유사도**가 **일정값 이상**인 그래프들을 **후보**로 선택



그래프 분류(classification)

- 주어진 그래프에 대해서 학습 데이터를 이용하여 **라벨**(label)을 모르는 노드의 라벨을 **결정**하는 문제

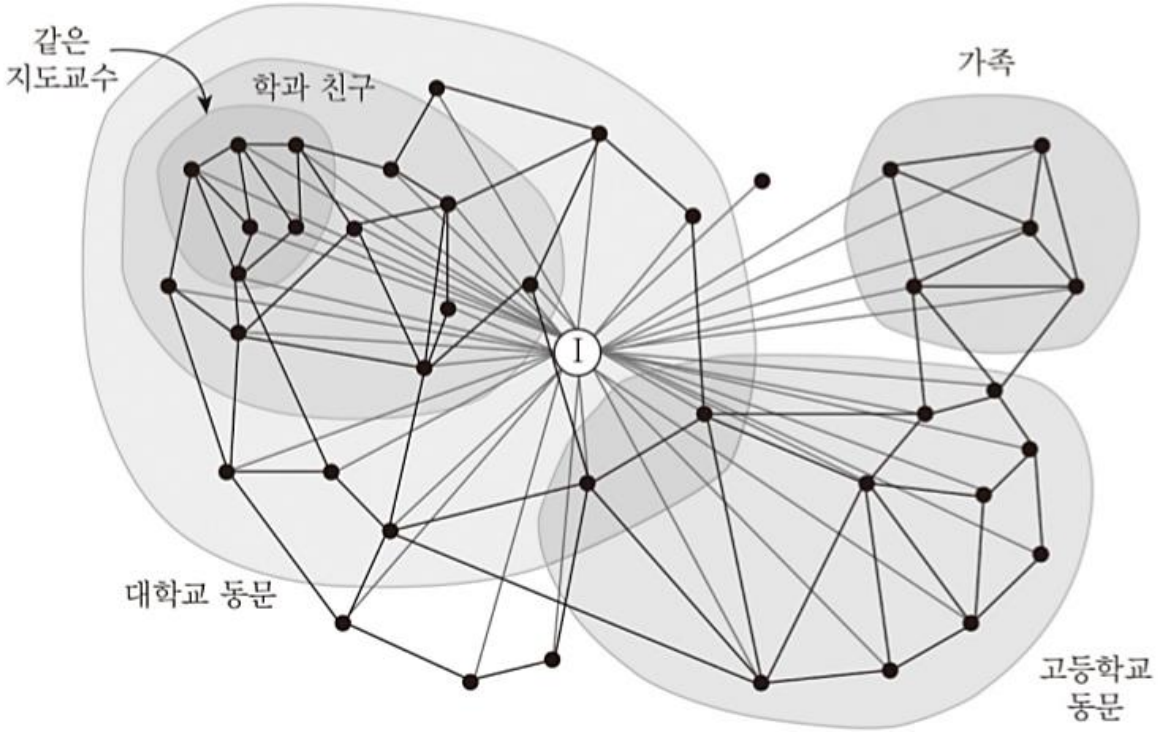


- 그래프 데이터에 대한 **부류**를 결정하는 문제
  - 예. 화합물을 나타내는 그래프들이 주어질 때, **특정한 성질**의 **유무**를 결정하는 문제

그래프 군집화(graph clustering)

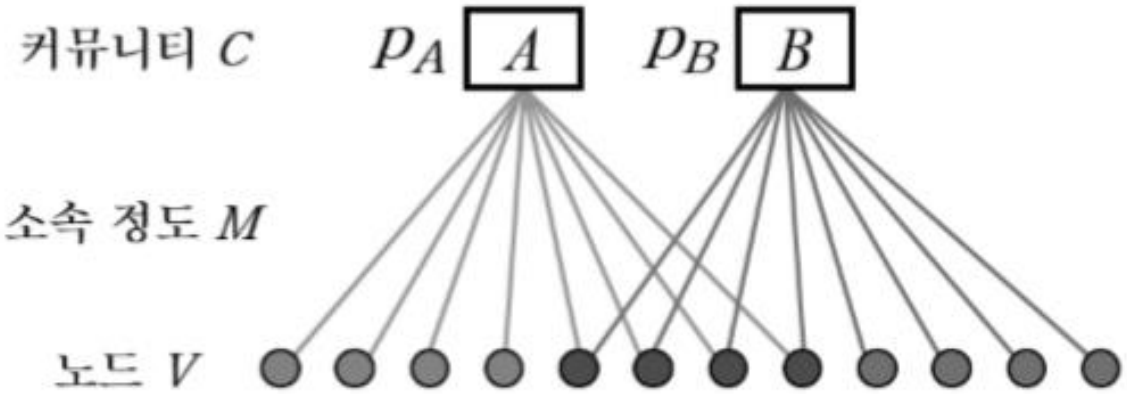
- 하나의 그래프에서 **특정 성질**을 **만족**하는 부분그래프들을 찾거나,
  많은 그래프 데이터들에서 **비슷한 것**들을 **군집**으로 묶는 것
  - **밀집군집 식별**(dense cluster identification)
  - **그래프 분할**(graph partitioning)

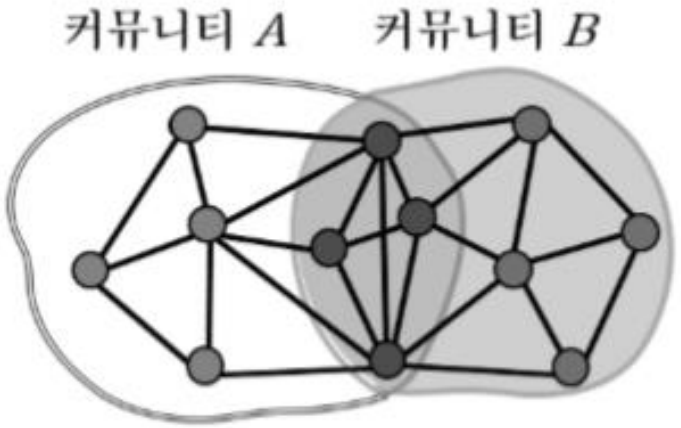
- **밀집군집 식별**
  - **커뮤니티**(community, 관심사나 관계에 의해 만들어지는 집단) 찾는 문제



커뮤니티 소속 그래프 기반 커뮤티니 식별 알고리즘

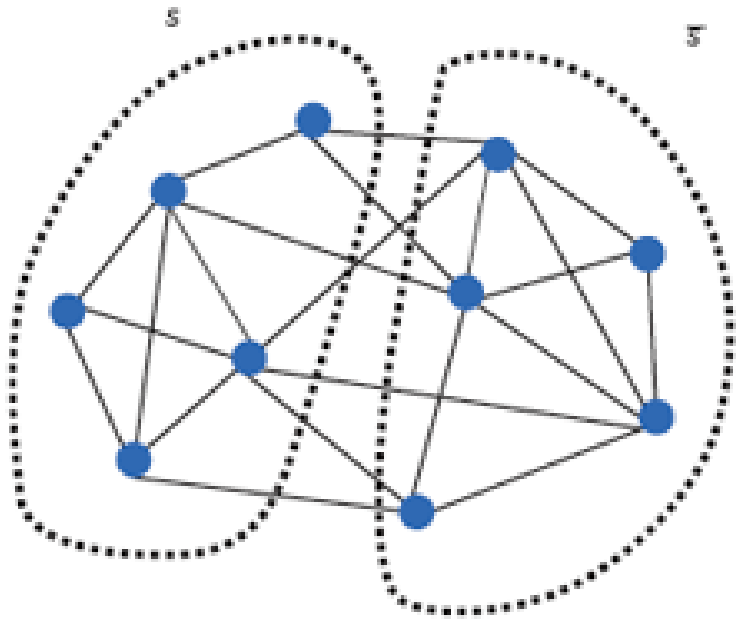
- **커뮤니티 소속 그래프**(community affiliation graph)
  - 노드와 커뮤니티의 소속 관계를 확률적으로 표현하는 그래프
- 커뮤니티 소속 그래프 모델이 **주어진 그래프를 생성할 확률**이 **최대**가 되도록 커뮤니티 소속 그래프 모델의 **파라미터 결정**





그래프 군집화(graph clustering)

- **그래프 분할**
  - 하나의 그래프를 **몇 개의 부분그래프**로 **분할**하는 문제
  - 분할될 때 제거되는 링크의 개수나 링크의 가중치의 합이 최소가 되도록 만드는 것과 같은 요건을 만족



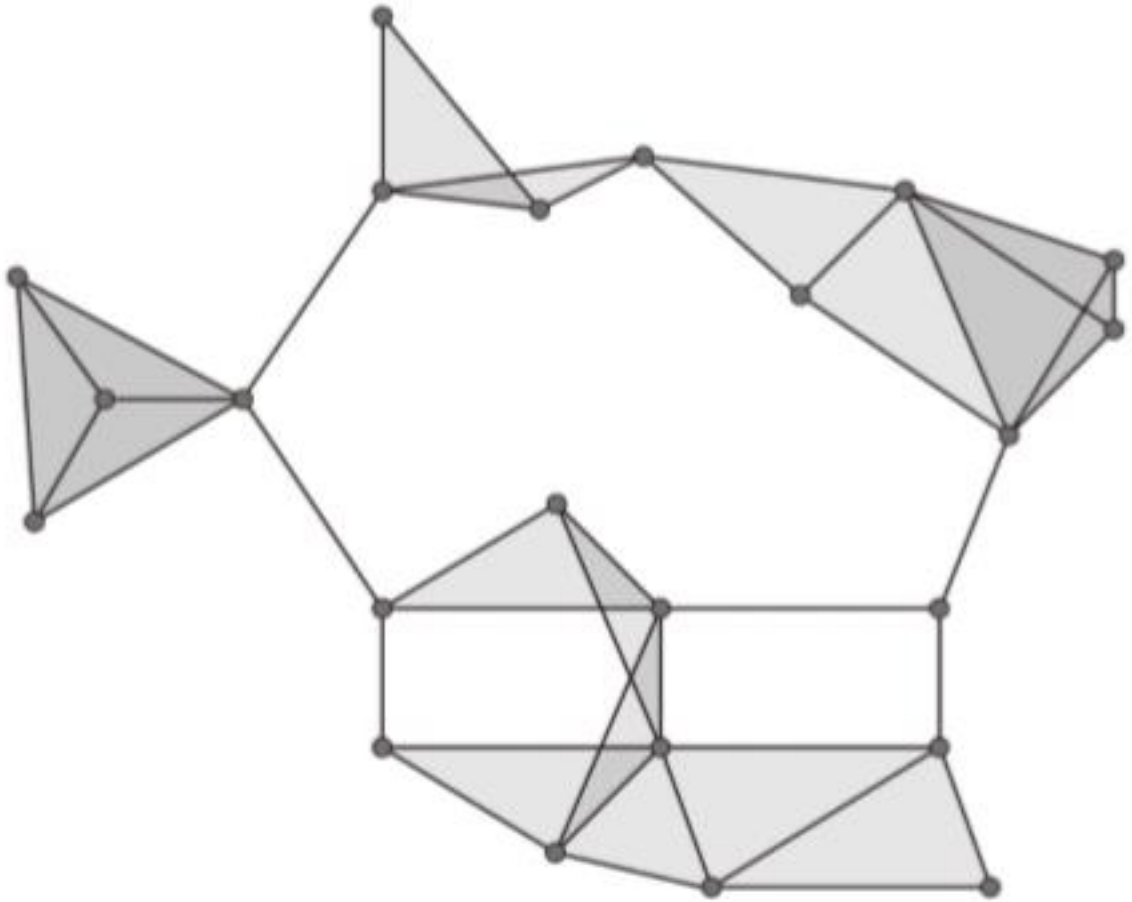
그래프의 키워드 검색

- **데이터**가 **그래프**로 **표현**되어 있을 때, **키워드**를 **사용**하여 필요한 정보를 **검색**
- Rusell Crowe가 출연한 영화의 감독이 연출한 영화 중에 Helena Carter가 주연한 것

The King's Speech
Directed by **Tom Hooper**

그래프 데이터 문제

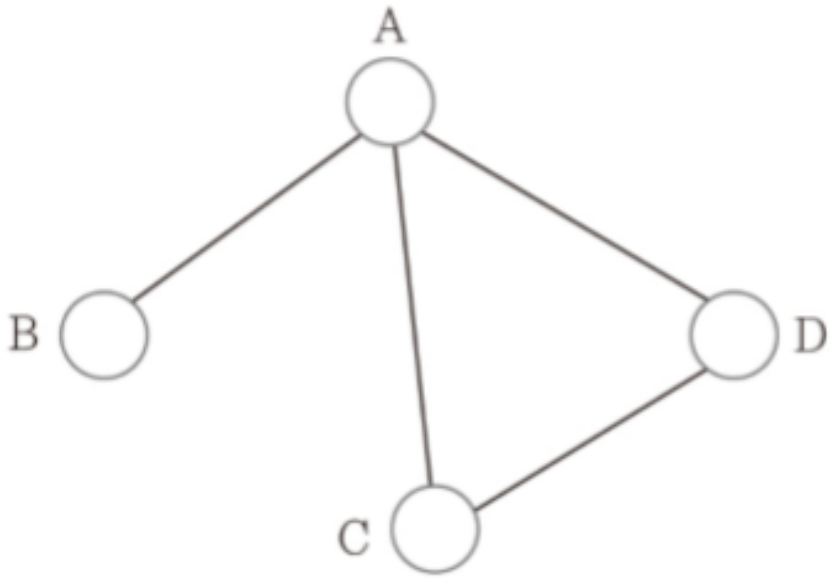
- 작은 문제에서 간단한 문제도 규모가 커지면 곤란
  - **최단경로** 찾기, **허브**(hub, 많은 노드와 연결된 노드) 찾기, **최소 신장 트리**(minimum spanning tree) 찾기
- 허브 찾기, **클릭**(clique, 모든 노드 쌍 사이에 링크가 존재하는 부분 그래프) 찾기, 최소 비용 그래프 절단(graph cut)
  - **클릭(clique)**: 클릭에 속하는 노드들은 서로 간에 링크가 존재
  - 준-클릭(quasi-clique)을 찾는 것에 더 관심



군집화 상수와 구조 유사도

- **군집화 상수:** 그래프 에서 노드들이 뭉쳐 있는 정도가 얼마나 강한지 측정하는 척도
- 값이 클수록 높은 응집도 의미

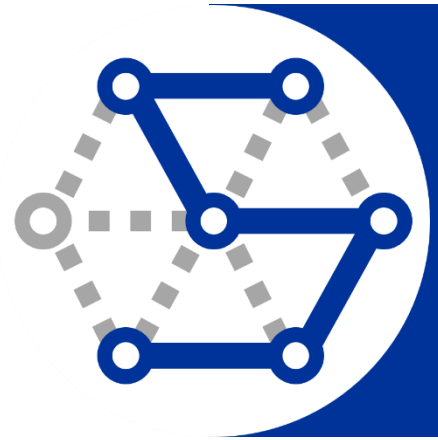
$$c(v) = \frac{\text{노드 } v\text{의 인접 노드 간의 간선의 개수}}{\text{노드 } v\text{의 인접 노드 간에 만들 수 있는 간선의 개수}}$$



- **구조 유사도: 노드 쌍**에 대해서 측정
- 값이 클수록 해당 노드들이 동일한 클릭이나 커뮤니티에 속할 가능성 높음

$$\sigma(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)||\Gamma(v)|}}$$

$\Gamma(u)$: 노드 u에 이웃한 노드들의 집합



6 추천

추천 (Recommendation)

- 개인별로 맞춤형 정보를 제공하려는 기술
- 사용자에게 맞춤형 정보를 제공하여 정보 검색의 부하를 줄여주는 역할
- **등수 매기기**(ranking) 방법 사용

PageRank 알고리즘

- 구글 CEO인 **Larry Page**가 스탠포드 대학 박사과정 중에 개발
- 웹페이지의 중요도를 평가하는 알고리즘
- **무작위 서퍼**(random surfer) 개념 사용
  - 무작위로 웹 페이지를 돌아다니는 이 서퍼가 각 페이지를 방문하는 비율만큼 중요도가 높다고 가정
  - 특정 페이지에서 다른 페이지로 갈 때, 똑같은 확률로 다른 페이지를 방문

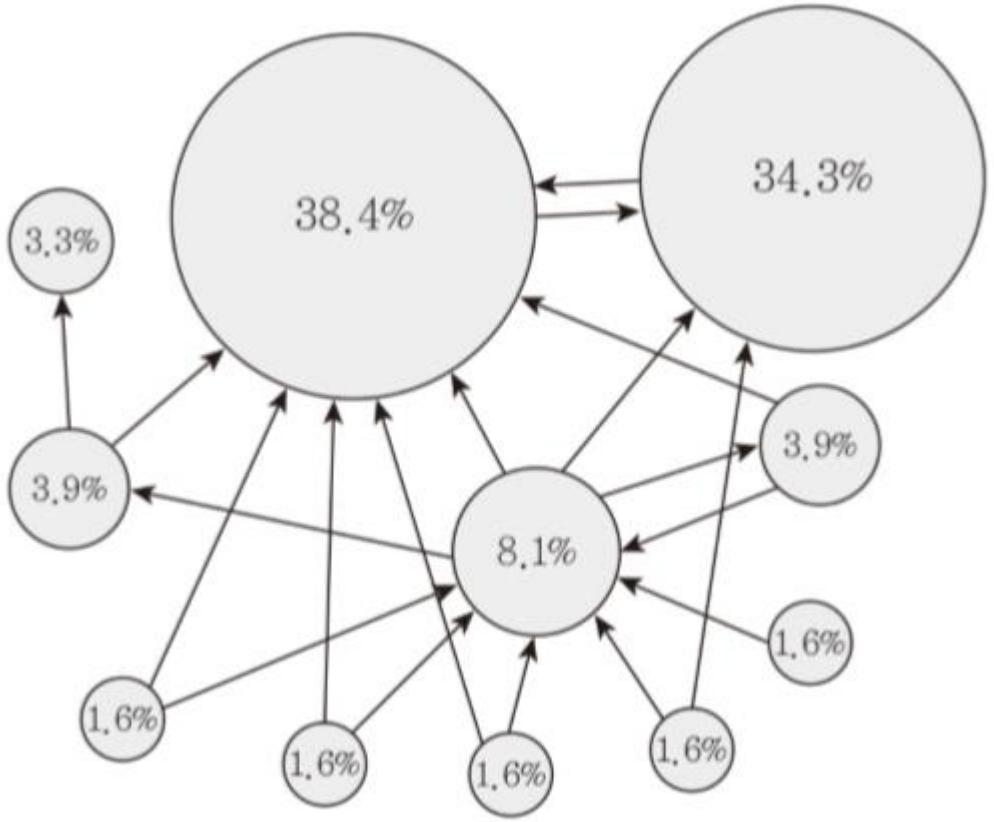
PageRank 알고리즘

- **페이지 중요도 계산**

$$r_j = \sum_{i \to j} \frac{r_i}{n_i}$$

r_j : 페이지 j의 중요도
n_i : 페이지 i의 하이퍼링크의 수

- **텔레포트(teleport) 기능**
  - 임의의 노드로 점프하는 기능



Netflix 상(Netflix prize)

- 미국의 DVD 대여와 비디오 스트리밍 서비스를 하는 회사
- 고객들의 취향을 반영하여 영화를 추천하는 시스템을 자체적으로 개발하여 활용
- 자사의 추천 시스템보다 10% 이상 예측 성능이 뛰어난 시스템을 개발한 첫 번째 팀에게 백만 달러의 상금
- 2000년에서 2005년 사이에 48만여 명의 자사 고객이 **1만 7천여 개의 영화**에 대해서 **별점 평가**를 한 **약 1억 건의 데이터**
  - (고객 ID, 영화 ID, 등급부여 날짜, 별점 등급)으로 표현
- 2006년 10월 2일에 시작
- 2009년 9월 21일에 10.06%의 성능 향상을 시킨 **BellKor팀** 우승

추천 데이터

- **희소 행렬(sparse matrix) 형태**
  - 많은 원소가 비어 있음
  - 비어 있는 부분을 채우는 것이 추천에 해당

고객

영화	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

내용 기반 추천(content-based recommendation)

- 고객이 이전에 높게 평가했던 것과 유사한 내용을 갖는 대상을 추천

- **항목 프로파일**(item profile)
  - 추천 대상 항목에 대한 특징 기술
  - 영화의 예: 영화의 장르, 시대, 지역, 역사적 배경, 배우와 같은 특징 정보
- **사용자 프로파일**(user profile)
  - 고객별로 선호 대상 정보 기술

- 추천 방법
  - **사용자** 프로파일과 **항목 프로파일**이 **유사**하면, 해당 항목에 대한 사용자의 선호도가 높은 것으로 추천

협력 필터링(collaborative filtering)

- **사용자간 협력 필터링**
  (user-user collaborative filtering)
  - 추천 대상 사용자와 비슷한 평가를 한 사용자 집합 이용
  - 추천 데이터 행렬의 열(column)이 비슷한 것 선택
  - 유사도 평가 : 코사인 거리 사용

$$\cos(U_i, U_j) = \frac{U_i \cdot U_j}{|U_i||U_j|}$$

  - 빈 원소의 추천정도 계산
    - 유사 사용자 집단의 평가 평균값 또는 가중 평균값

- **항목간 협력 필터링**
  (item-item collaborative filtering)
  - 항목간의 유사도를 구하여 유사 항목을 선택
  - $\hat{r}(x, I_a)$: 사용자 u의 특정 항목 I_a에 대한 예측 등급

$$\hat{r}(x, I_a) = \frac{\sum_{I_b} s(I_a, I_b) r(x, I_b)}{\sum_{I_b} r(x, I_b)}$$

$r(x, I_b)$: u의 항목 I_b에 대한 평가 등급
$s(I_a, I_b)$: 항목 I_a와 I_b의 유사도

은닉 요소 모델(latent factor model)

- 행렬 분해에 기반한 방법
  - 추천 데이터 행렬을 2개의 항렬의 곱으로 표현할 수 있다고 가정
  - 추천 데이터 행렬에 주어진 값에 맞도록 분해된 행렬들의 원소를 결정
  - 분해된 행렬의 곱을 사용하여 추천 행렬의 빈 원소 결정
  - 대표적인 알고리즘
    - **ALS**(alternating least square) **알고리즘**

고객

영화	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

≈

영화 \ 요소			
1	.1	−.4	.2
2	−.5	.6	.5
3	−.2	.3	.5
4	1.1	2.1	.3
5	−.7	2.1	−2
6	−1	.7	.3

요소 \ 고객												
1	1.1	−.2	.3	.5	−2	−.5	.8	−.4	.3	1.4	2.4	−.2
2	−.8	.7	.5	1.4	.3	−.1	1.4	2.9	−.7	1.2	−.1	.5
3	2.1	−.4	.6	1.7	2.4	.9	−.3	.4	.8	.7	−.6	−.4


suanlab
이수안 컴퓨터 연구소
suan computer laboratory