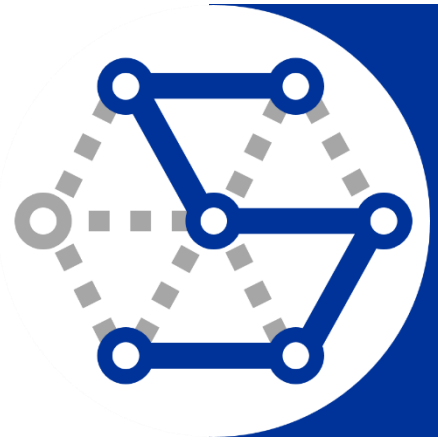
데이터 전처리
Data Preprocessing

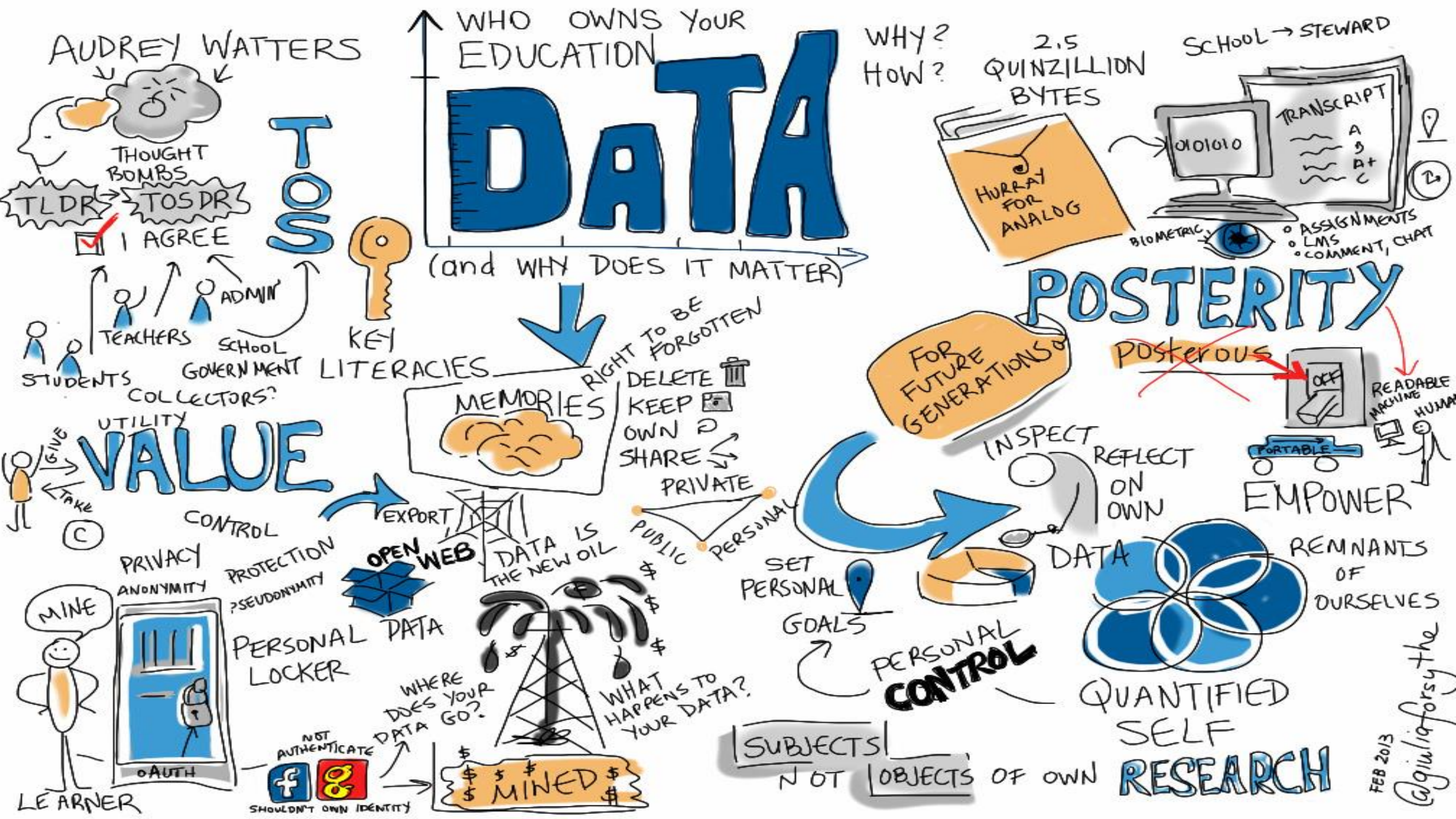

01
데이터 구조와 종류

목차

1. 데이터 개념
2. 데이터 구조
3. 데이터 종류



1. 데이터 개념


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

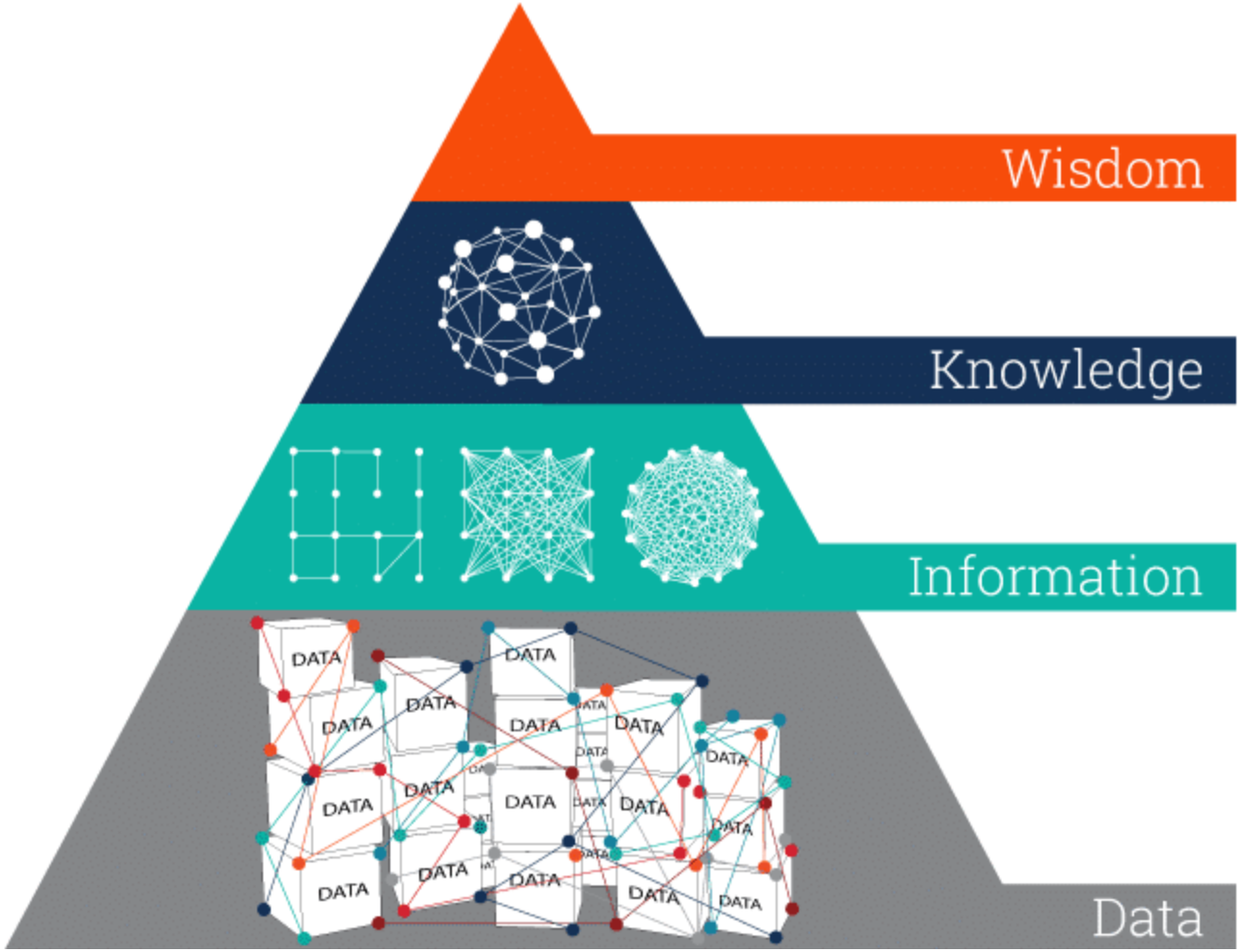
데이터 용어 정의

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 정보information가 아니고, 데이터를 가공해 얻는 것이 정보

DIKW Pyramid

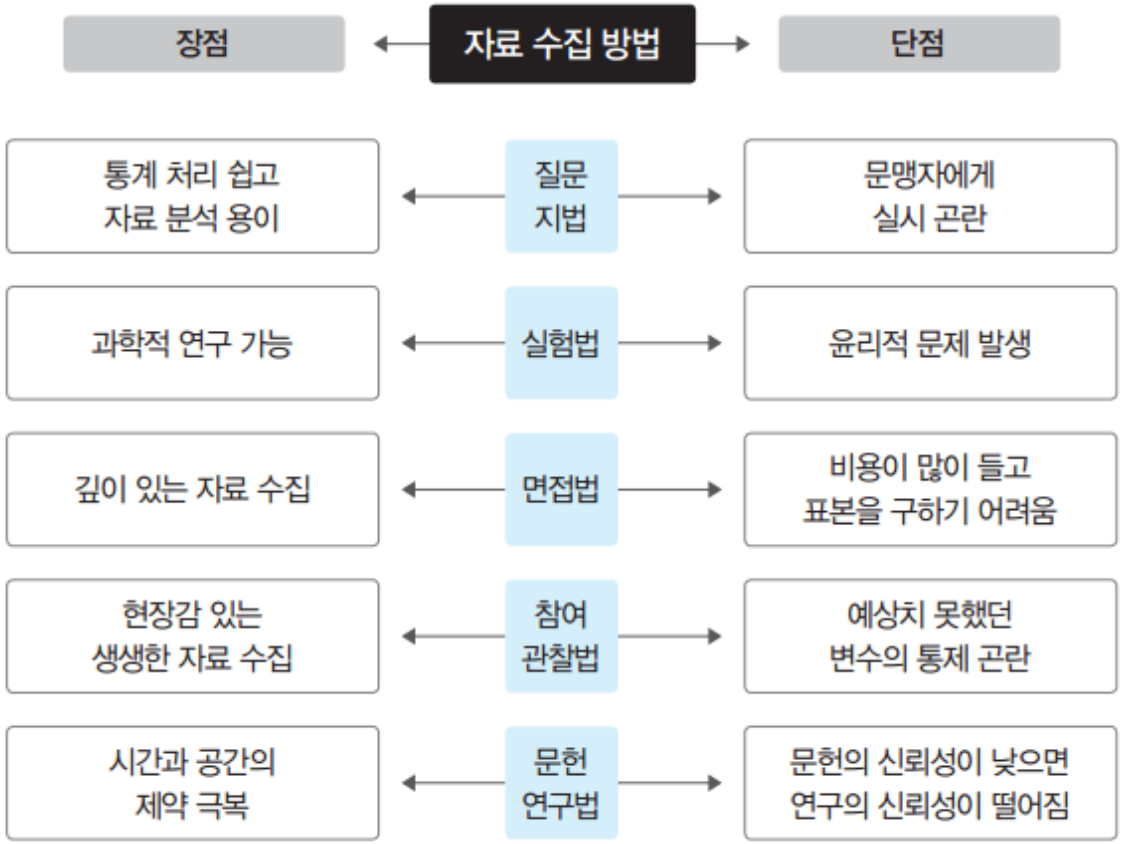
Data, Information, Knowledge, Wisdom (DIKW) Pyramid
https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/



Each step up the pyramid answers questions about and adds value to the initial data.

데이터 용어 (연구방법론)

- 연구에 직간접적으로 이용되는 일체의 자료
- 어떤 연구의 결과가 얼마나 유용할지는 그 자료의 질적 적절성이 중요
- 자료수집: 연구에 필요한 정보들을 수집하는 과정



데이터 종류 LOTS (연구방법론)

- **L 자료: 생애 데이터**
  - 한 대상의 통사적 정보를 알 수 있는 자료
  - 특히 특정 개인을 대상으로 한 임상 장면에서 많이 사용
  - 생활기록부, 범죄이력, 신용정보, 졸업증명, 병력 조회 등이 이에 해당
  - 객관화된 자료이지만, 이용에 한계가 존재

- **T 자료: 검사 데이터**
  - 실험적 절차를 거치거나 표준화된 검사를 통해 얻어진 데이터
  - 대중매체에서 과학자 인물들이 손에 들고 있는 도표들도 대부분 T-자료
  - 가장 객관적이고 질 좋은 자료이지만, 현실적으로 접해보기는 그다지 쉽지 않음
  - 자료를 확보하는 과정에서의 연구윤리 문제도 개입

- **O 자료: 관찰 데이터**
  - 숙련된 관찰자 혹은 대상을 잘 아는 관계자, 친지 등이 제공하는 자료
  - 면접법, 참여관찰법 등을 통해 확보 가능
  - 주변 사람들의 증언이나 CCTV 영상 자료 역시 O-자료에 속함

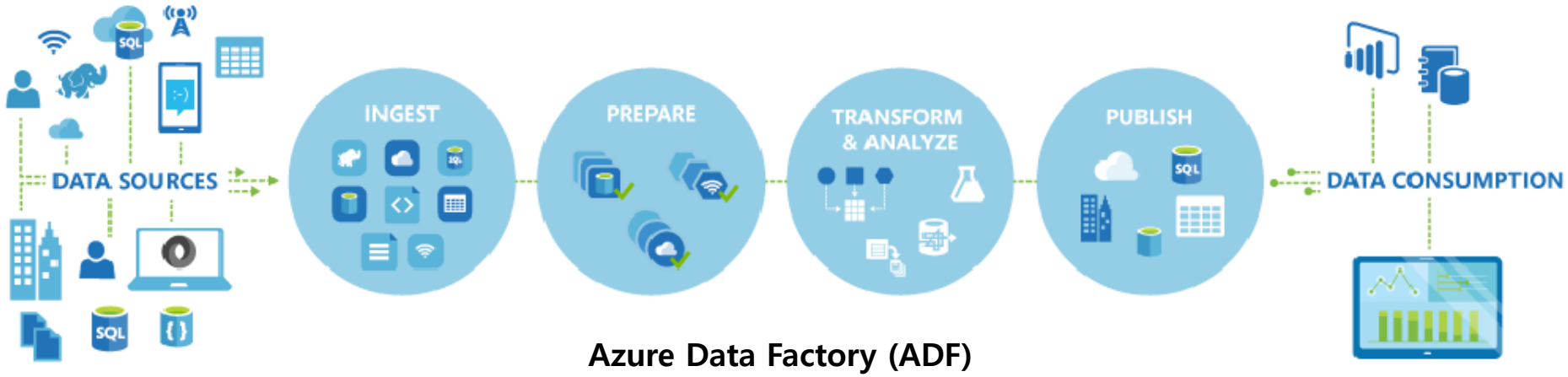
- **S 자료: 자기보고 데이터**
  - 어떤 대상에 대한 정보를 얻을 때 그 대상에게 직접 물어보아 얻은 자료
  - 당연히 사람을 대상으로 하므로, 그 분야는 심리학이나 사회학 등에 한정될 수밖에 없음
  - 매우 흔하게 접할 수 있는 자료로, 흔한 설문조사나 여론조사 등을 통해 얻어짐
  - "사람은 자신이 자신을 제일 잘 안다" 는 전제에 기초해 있으며, 사회적 선망에 의해 답변이 왜곡될 수 있음

데이터 용어 (컴퓨터)

- 프로그램에 부속된 파일, 특히 사용자가 해독할 수 없는 형태의 이진 파일
- 컴퓨터에 의해 특정한 방법으로 처리되거나 해석될 목적으로 순서를 가지고 나열된 기호Symbol가 모여있는 것
- 수치화된 크기/규모Magnitude, 개수Quantity, 문자, 또는 컴퓨터에 의해 해석되어 처리되거나 다른 기계, 다른 컴퓨터를 제어할 수 있는 명령어를 나타내는 심볼 등
- 보통 자기 저장매체(플로피디스크, 하드디스크, 카세트 테이프, 오픈릴 테이프, DAT, OMR카드 등), 메모리 저장매체(RAM, ROM, 플래시 메모리, SSD 등), 광학 저장매체(CD, DVD, 블루레이, OCR카드, 펀치카드 등), 기계적 저장매체 등에 저장되며 전기 신호의 형태로 전송 가능
- 프로그램은 컴퓨터가 해석하여 실행할 수 있는 명령을 나타내는 심볼 데이터의 모임
  근본적으로 컴퓨터라는 기계는 데이터의 형태로 표현된 일련의 명령어에 따라 동작하도록 설계

데이터 용어 (경영학)

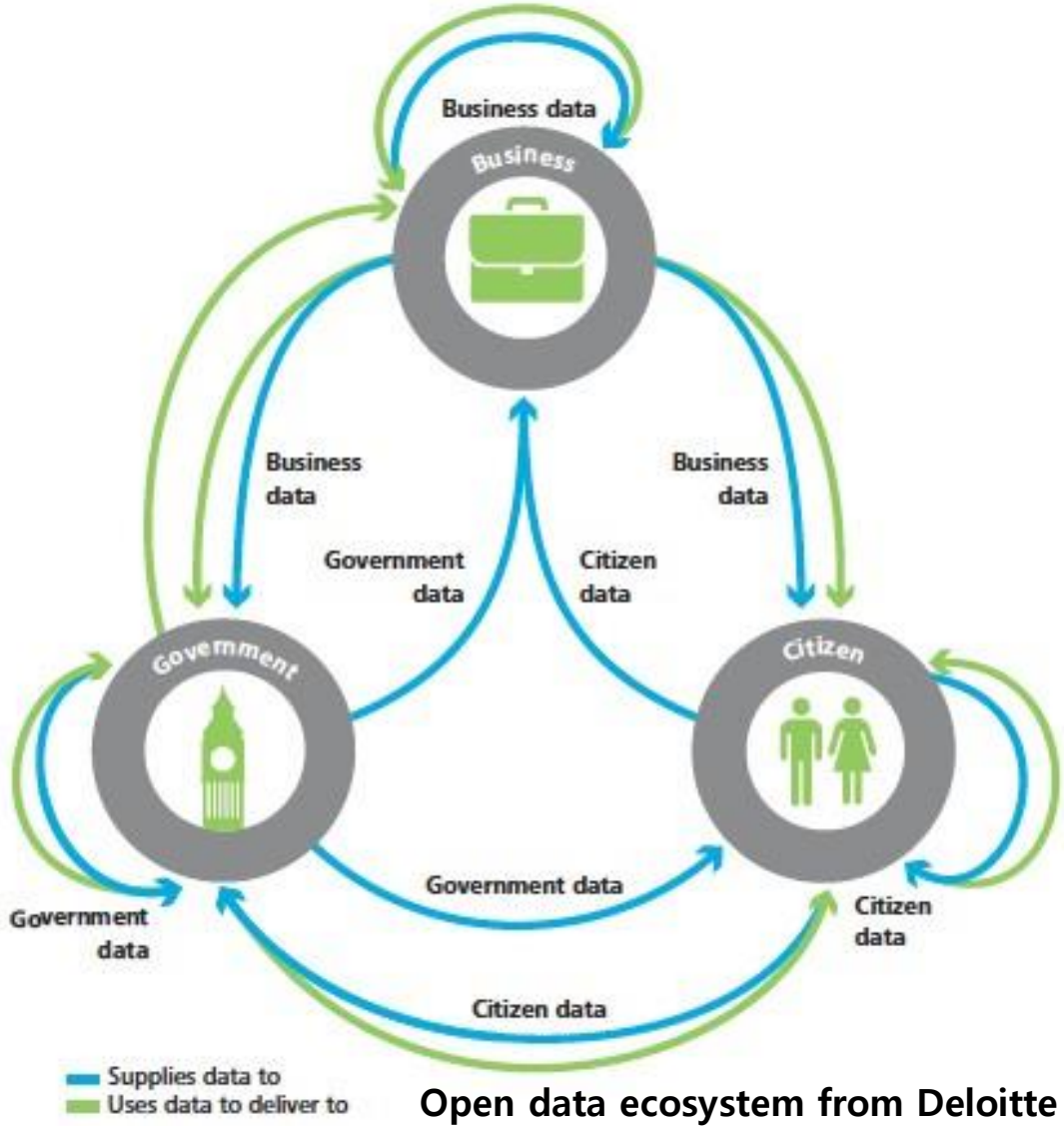
- 2010년 이후 데이터의 시대라고 부르기도 하며, 일부는 심지어 산업혁명 4.0이라고 부르기함
- 데이터 유통 분야
  - 데이터 팩토리data factory라는 새로운 개념의 회사들이 생겨났는데, 다른 말로는 데이터 뷰로data bureau라고 불리기도 함
  - 가치 있는 데이터들을 수집, 저장, 가공, 통합하여 재판매하는 일을 주로 하고 있음
  - 엡실론Epsilon, 액시엄Acxiom, 이퀴팩스Equifax 같은 회사들이 유명
  - 국내에도 KCB, NICE, SK 지오비전, 네이버 등이 데이터 팩토리로 불릴 수 있음



Azure Data Factory (ADF)

데이터 용어 (경영학)

- 금융 분야
  - 데이터 생태계라 하여 콜렉터, 브로커, 유저로 나누어지는 순환구조를 가정
  - 데이터는 판매자가 과거 판매했던 데이터가 이후 다시 특정 "사인sign"을 달고 판매자에게 되돌아오는 식으로 구성
  - 데이터 소비자는 구입한 데이터에 자신의 내부 데이터를 융합시켜서 활용하고, 그러한 경제활동을 통해서 데이터 판매자에게 가치 있는 데이터가 다시 전달되는 형태



Open data ecosystem from Deloitte

데이터 유형과 형식

https://guides.library.oregonstate.edu/research-data-services/data-management-types-formats

관측 및 관찰 데이터	실험 데이터	파생 또는 컴파일 데이터	시뮬레이션	참조 또는 표준
• 현장에서 캡쳐 • 다시 캡쳐하거나 재생산 및 교체 불가 • 예) 센서, 인간 관찰, 설문 조사 등	• 현장 또는 실험실 기반의 통제된 조건 속에서 수집된 데이터 • 재현이 가능하지만 비쌈 • 예) 유전자 서열, 크로마토 그램, 분광 데이터, 현미경 데이터 등	• 재현가능하지만 비쌈 • 예) 텍스트 및 데이터 마이닝, 파생 변수, 컴파일된 데이터베이스, 3D 모델 등	• 모델을 사용하여 실제 또는 이론적 시스템의 동작 및 성능을 연구한 결과 • 모델 및 메타데이터는 입력 데이터가 출력 데이터보다 더 중요 • 예) 기후 모델, 경제 모델, 생지화학 모델 등	• 정적 또는 유기적 컬렉션 데이터 세트 • 예) 유전자 서열 데이터뱅크, 화학 구조, 공간 데이터 포털 등

데이터 집합 특성

Dimensionality

- 데이터 집합의 차원은 각 데이터 개체가 가지는 속성의 개수를 의미
- 데이터에 따라서는 속성의 수가 너무 많아 분석의 어려움이 발생할 수 있는데 이를 '차원의 저주Curse of Dimensionality'라 표현

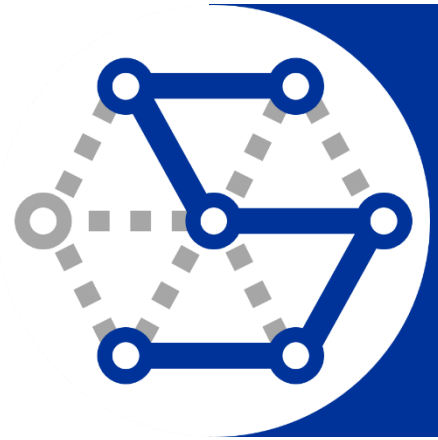
국내 연구진, 통계학 난제 '차원의 저주' 해결
http://www.hankookilbo.com/News/Read/201808081515040760

Sparsity

- 어떤 데이터 집합은 대부분의 데이터 개체에서 속성들이 0의 값을 가지며, 1% 미만의 데이터 개체에서만 0이 아닌 값을 가지는 경우가 있음
- 일반적으로 이러한 데이터의 경우 저장에 있어 0이 아닌 값만을 사용함으로써 데이터의 저장과 분석을 용이하게 할 수 있음
- 예를 들어 4 x 4 행렬에서 (2, 3) 원소의 값만이 0이 아닌 값이라면 이 행렬의 저장은 16개의 모든 원소를 저장하는 것이 아니라 (2, 3, 값) 이라는 정보만으로도 행렬을 표현할 수 있음

Resolution

- Resolution에 따라서 획득되는 데이터의 특성이 달라질 수 있음
- Resolution이 너무 높은 경우에는 잡음과 같은 간섭 요인에 영향을 많이 받을 수 있으며, 반대로 너무 낮은 경우에는 정보가 사라질 수도 있음
- 예를 들어 해수 온도 측정에 있어 1년마다 측정을 한다면 계절별 온도 변화 패턴을 찾기는 어려울 것
- 그러므로 적절한 수준의 Resolution을 사용 하는 것이 필요하며, 이는 실험 계획법과도 연관



2. 데이터 구조


1992 (11)
1991 (11)
2003 (11)
2002 (11)
2001 (11)


FIRE EXIT
FIRE EXIT
AMERICA · AFRICA
ASIA · AUSTRALASIA


{"timestamp":"2017-06-03T18:42:18.018", "deltaStartMillis":"0", "level":"INFO", "webURL":"/app/page/report", "sessionID":"144o2n620jm9trnd3s3n7wg0k", "webParams":"file=chartdata_new.json", "class":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"789d89cb-bfa8-4e7d-8047-498454af885d", "sizeChars":"48455", "message":"Duration Log", "durationMillis":"7"}{"timestamp":"2017-06-03T18:46:921.000", "deltaStartMillis":"0", "level":"INFO", "webURL":"/app/page/analyze", "sessionID":"144o2n620jm9trnd3s3n7wg0k", "webParams":"null", "class":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"8249868e-afd8-46ac-9745-839146a20f09", "sizeChars":"5022", "message":"Duration Log", "durationMillis":"36"}{"timestamp":"2017-06-03T18:42:18.018", "deltaStartMillis":"0", "level":"INFO", "webURL":"/app/rest/json/file", "sessionID":"144o2n620jm9trnd3s3n7wg0k", "webParams":"file=chartdata_new.json", "class":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sizeChars":"10190", "message":"Duration Log", "durationMillis":"23"}{"timestamp":"2017-06-03T18:43:335.030", "deltaStartMillis":"0", "level":"INFO", "webURL":"/app/page/analyze", "sessionID":"144o2n620jm9trnd3s3n7wg0k", "webParams":"null", "class":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID":"8249868e-afd8-46ac-9745-839146a20f09", "sizeChars":"5022", "message":"Duration Log", "durationMillis":"508"}

데이터 세트Data set

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

- 데이터 모음
- 하나의 데이터베이스 테이블의 내용이나 하나의 통계적 자료 행렬과 일치
- 컬럼column: 특정한 변수를 대표
- 로우row: 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

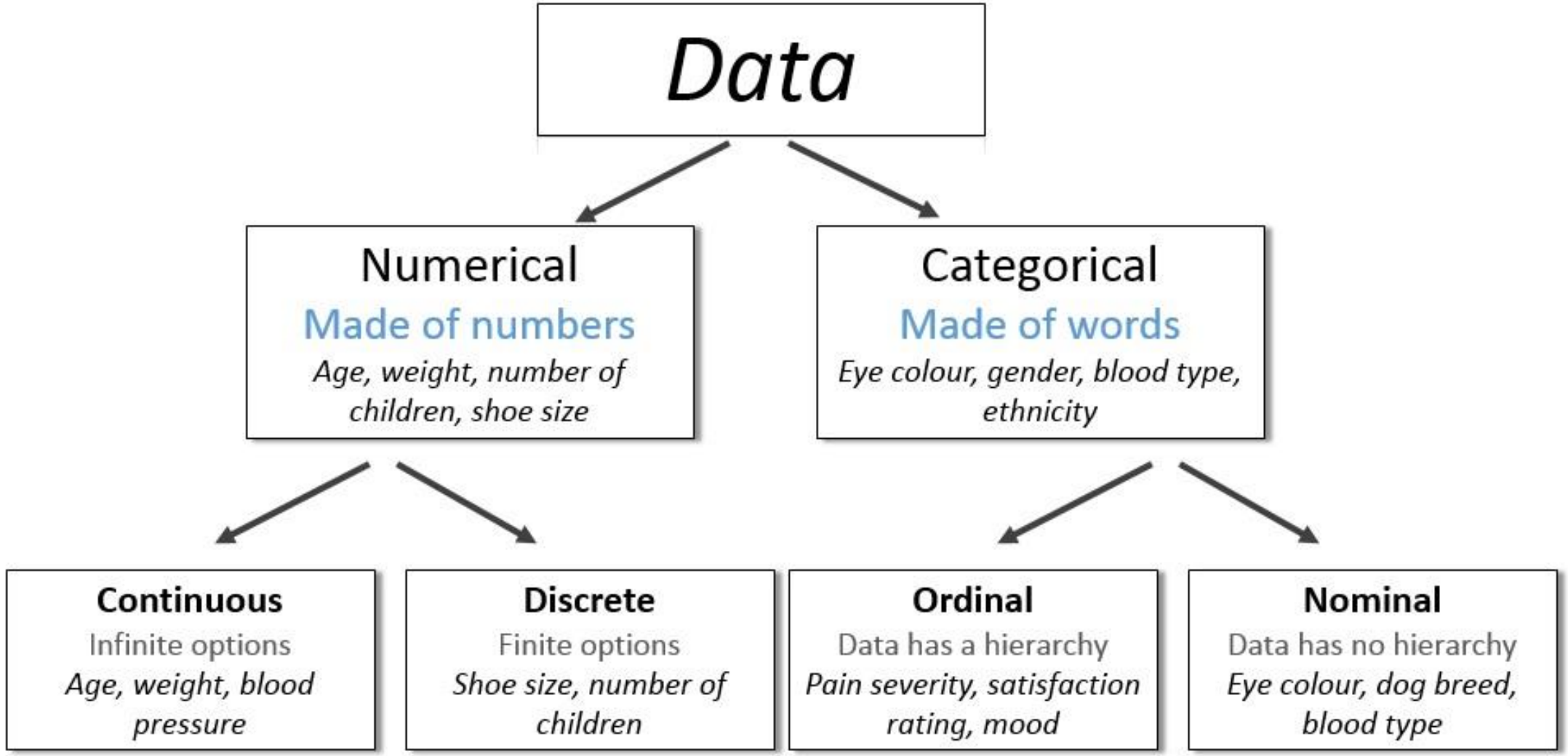
- 데이터 세트data set: 데이터 개체data object들의 집합
- 데이터 개체data object: 레코드record, 점point, 벡터vector, 패턴pattern, 사례case, 사건event, 샘플sample, 관찰observation, 개체entity 등으로 불림
- 데이터 개체는 여러 개의 속성attribute으로 기술
- 속성attribute : 데이터 개체들 사이의 차이를 규정할 수 있는 특성이나 특징을 의미
  - 예) 사람을 기술할 때 눈동자의 색, 피부색, 키, 몸무게와 같은 속성을 사용
- 속성은 변수variable, 특성characteristic, 필드field, 특징feature, 차원dimension 등으로 불림

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

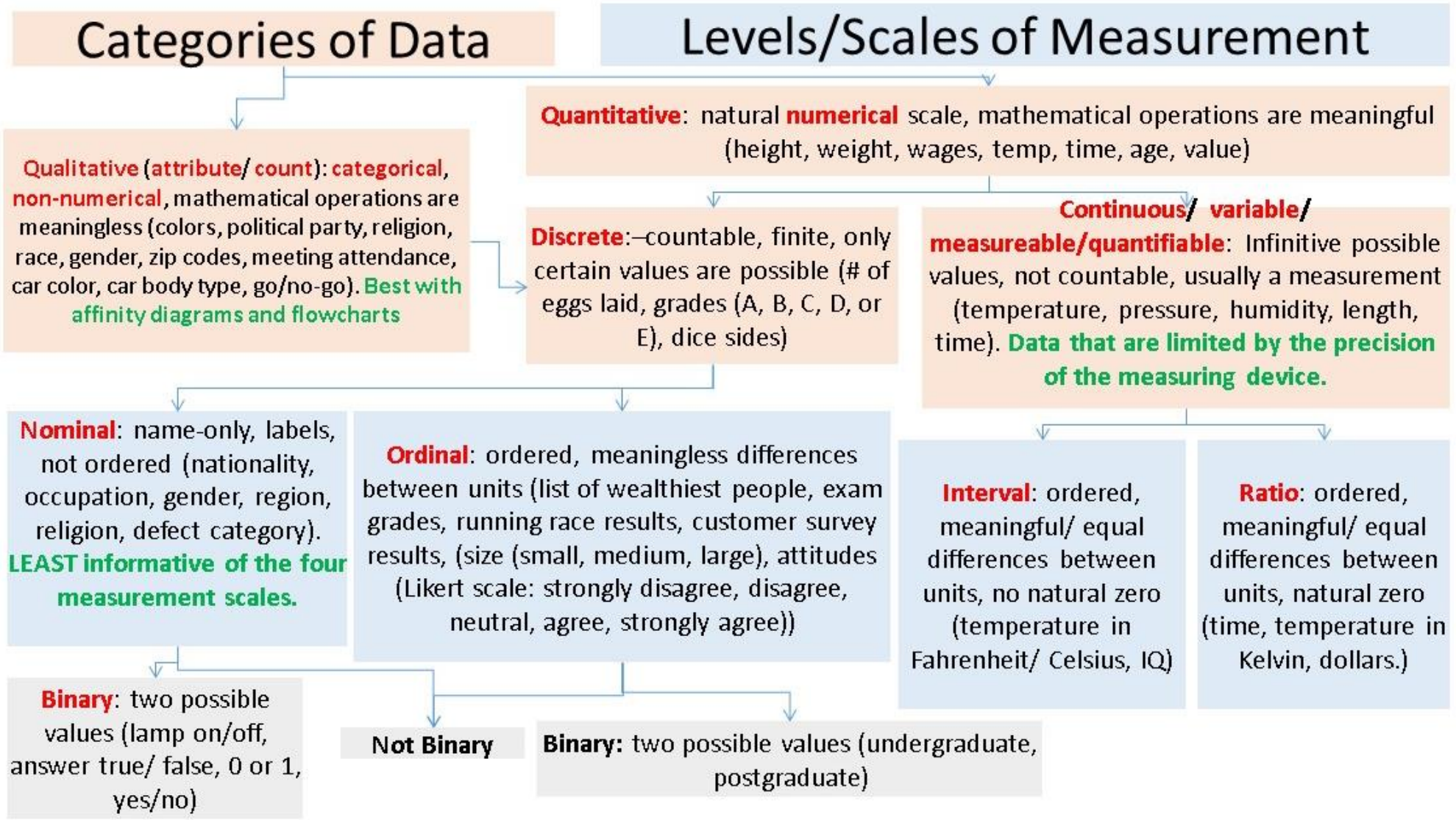
데이터 분류Data Classification

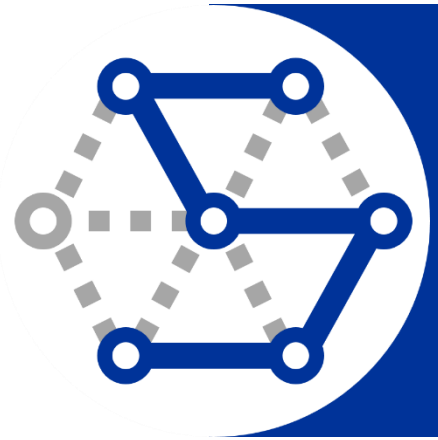
http://survivestatistics.com/variables/



데이터 분류Data Classification

https://i.stack.imgur.com/J8Ged.jpg





3. 데이터 종류

데이터 종류


Data
data
categories
meaningful
twice
nominal
Ordinal
items
interval
Nominal
non-equidistant
observed
distance
Quantitative
zero
equidistant
include
Interval
ratio
Ratio
types
important
make
arbitrary
ordered
measured
Qualitative
statistical
properties
Examples
following
adjacent
median
mode
discrete
zero-point
intervals
ordinal
powerful
just
agree
etc
may
scale
Tumour
categorised
central
long
difference
value
know
blog
disagree
Favourite
Infographic


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

레코드 데이터Record data

- 데이터 마이닝에서 가장 많이 사용되는 데이터 형태로 대개 flat 파일 형태로 저장된 데이터 세트
- 레코드Record의 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

트랜잭션 데이터Transaction Data

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

- 구매자와 구매 물품목록 형태로 이루어진 데이터 세트
- 장바구니 데이터Market Basket Data라고도 불림

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

데이터 행렬Data matrix

- 모든 속성이 수치 형태의 값을 가지는 행렬 형태의 데이터 세트
- 일반적으로 데이터의 행은 개체, 열은 속성을 나타냄
- 패턴 행렬Pattern matrix이라고도 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

희박한 데이터 행렬Sparse Data Matrix

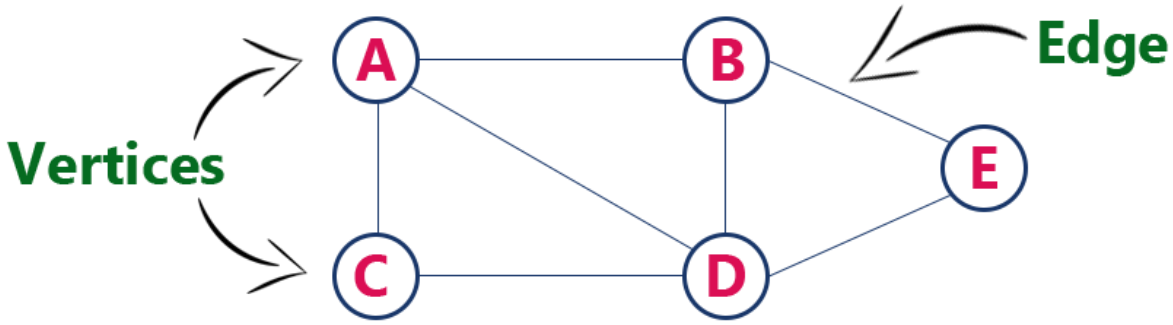
Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesley, 2005

- Data matrix의 특별한 경우
- 예: 각 문서에서 용어가 출현하는 빈도수
- 문서의 경우에는 용어 벡터term 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

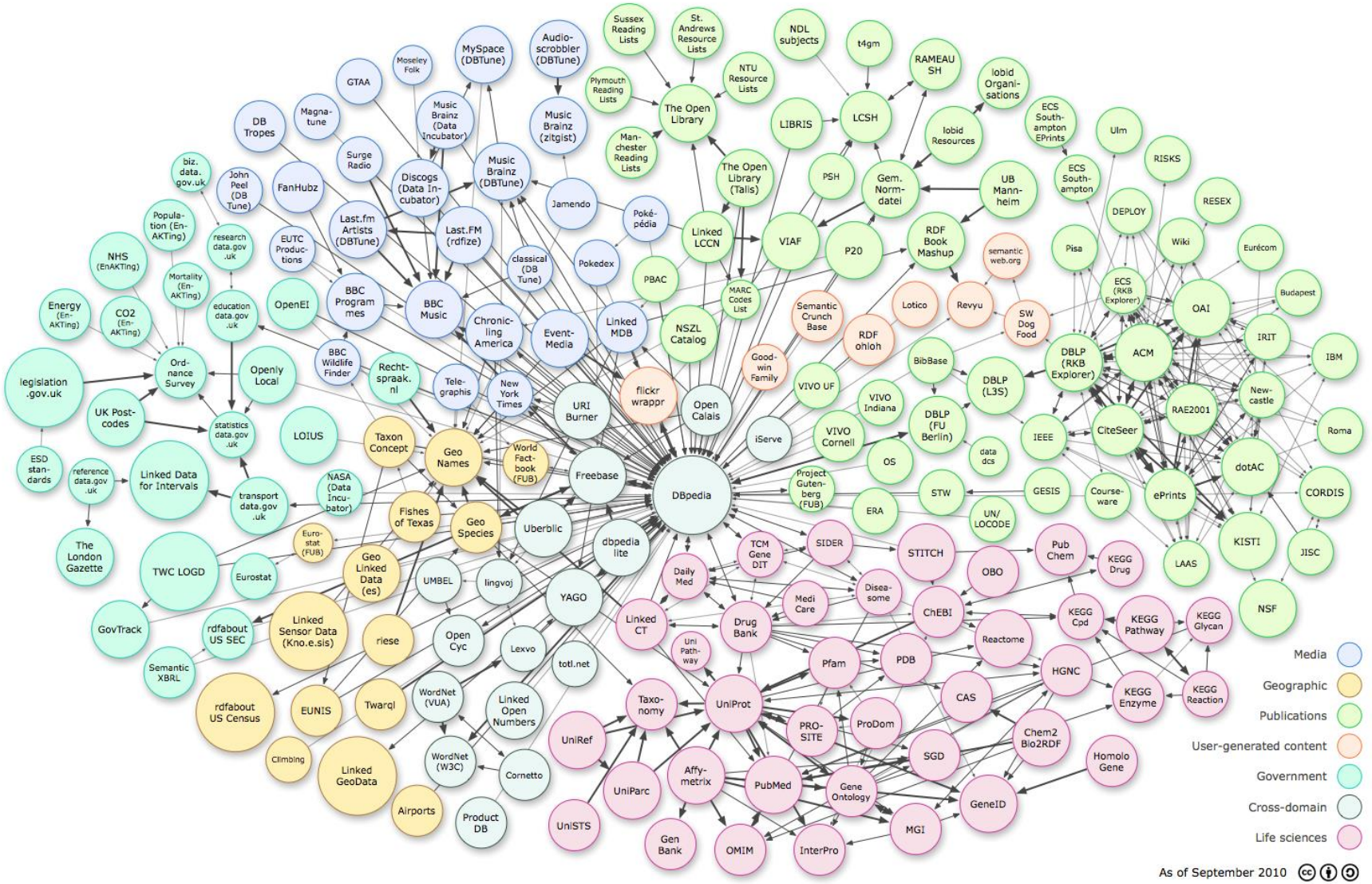
그래프 데이터Graph-based data

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

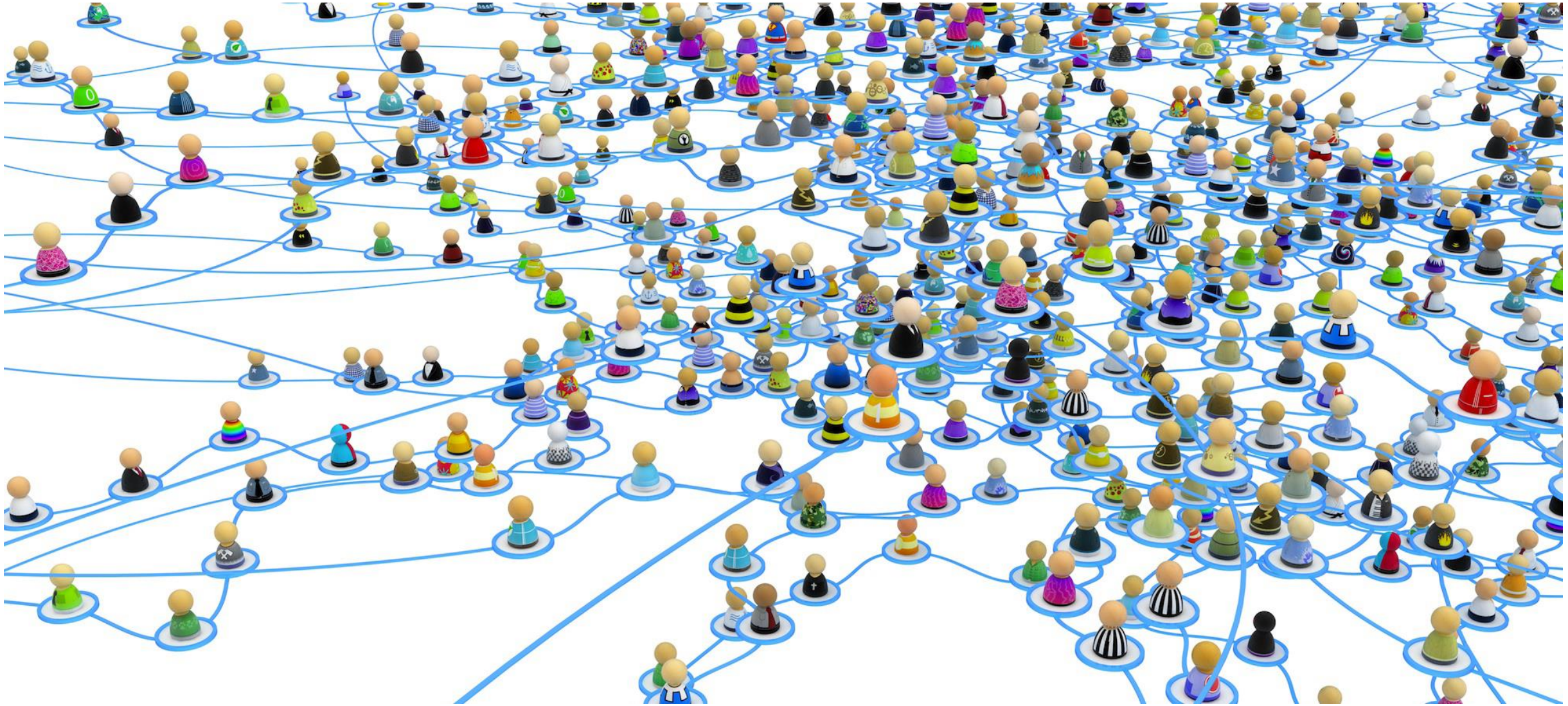


http://btechsmartclass.com/data_structures/introduction-to-graphs.html

그래프 데이터Graph-based data

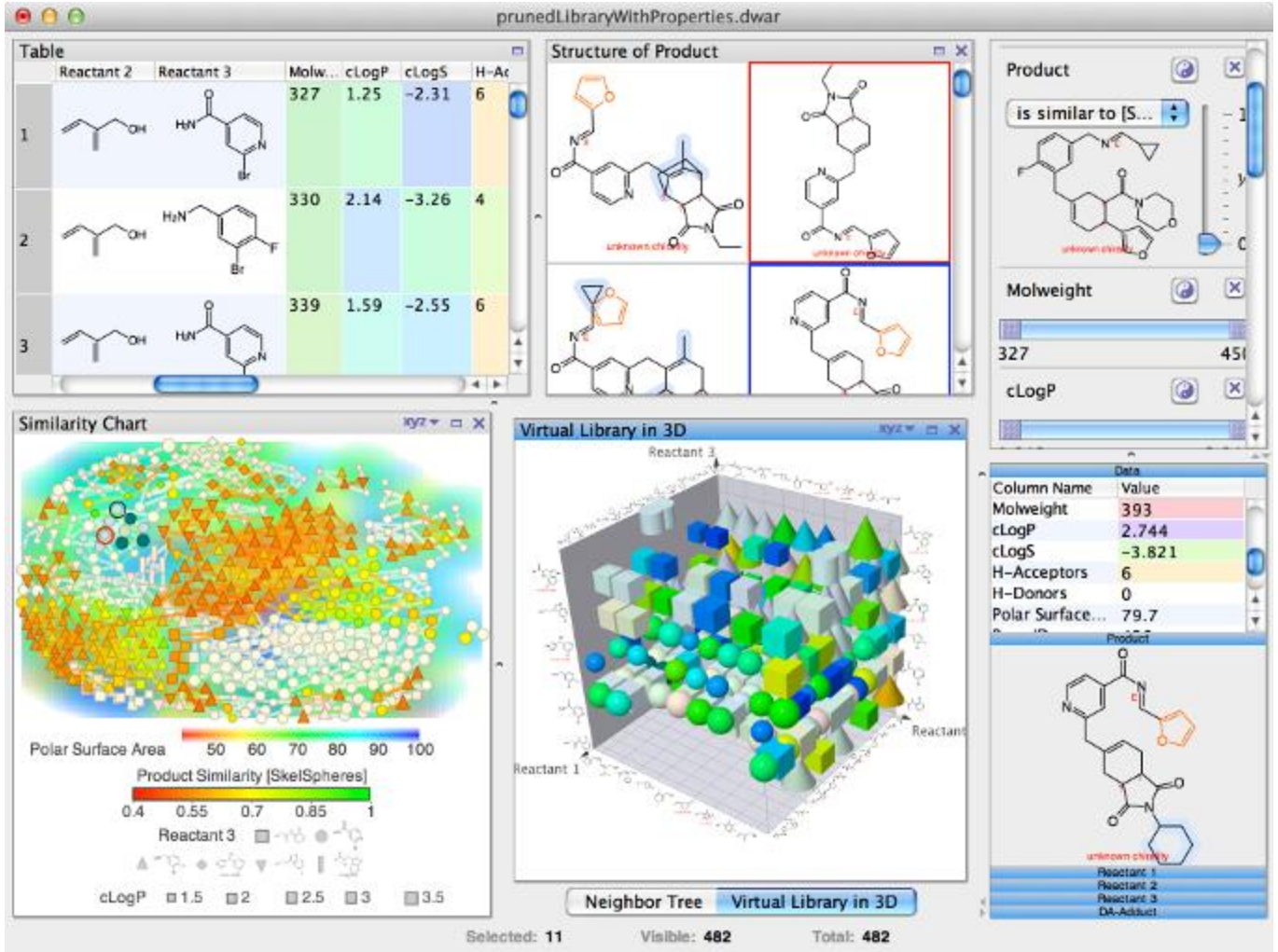


그래프 데이터Graph-based data



그래프 데이터Graph-based data

http://www.openmolecules.org



순서 데이터Ordered data

- 데이터 개체의 속성이 시간 또는 공간적인 순서와 연관되는 데이터 세트
- 순서 데이터의 종류
  - 연속 데이터Sequential data
  - 서열 데이터Sequence data
  - 시계열 데이터Time series data
  - 공간 데이터Spatial data

연속 데이터Sequential data

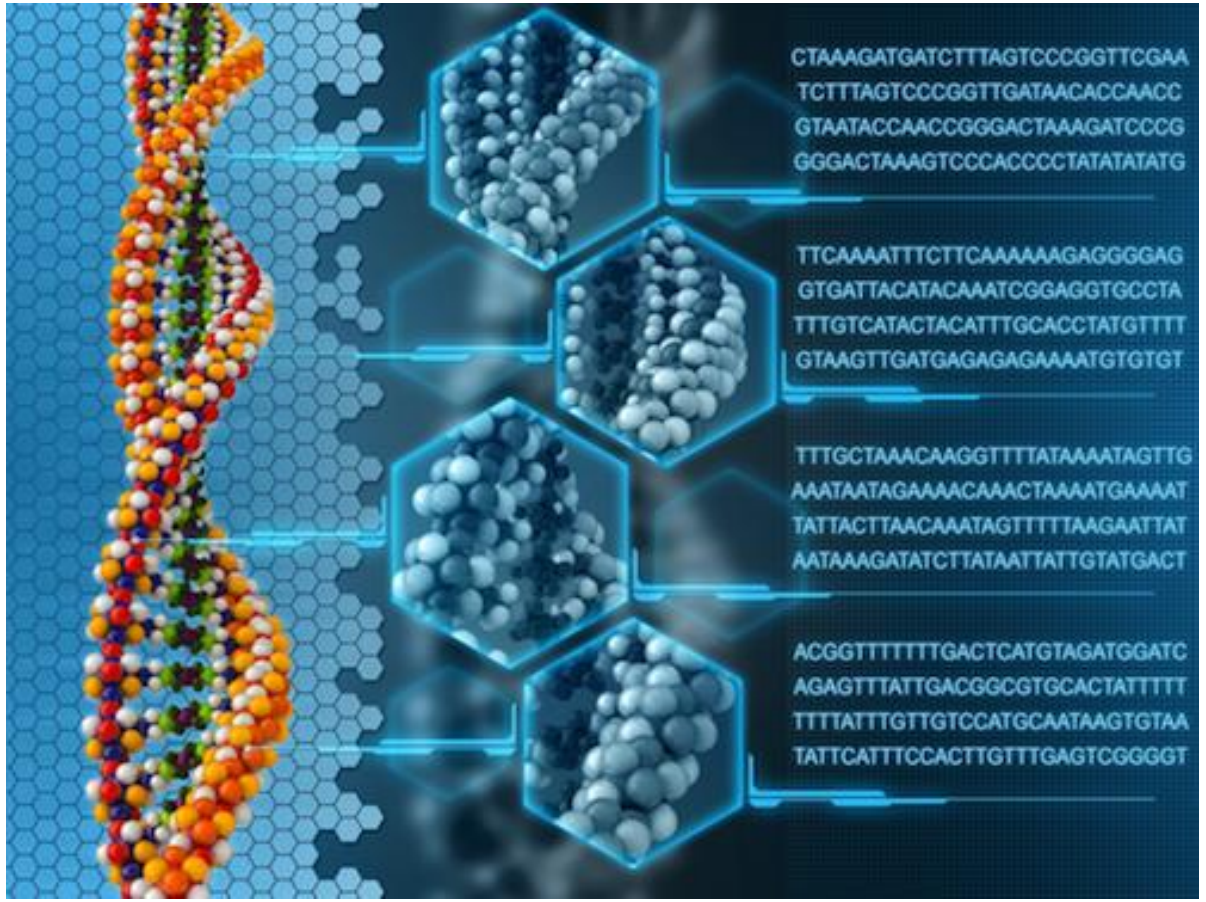
- 트랜잭션 데이터에서 시간 성분을 추가적으로 고려한 것
- 고객의 시간에 따른 구매 경향 예측과 같은 응용에서 사용 될 수 있음
- 예: CDP 구매 고객은 CD를 구매할 계획이 있음

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)

서열 데이터Sequence 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

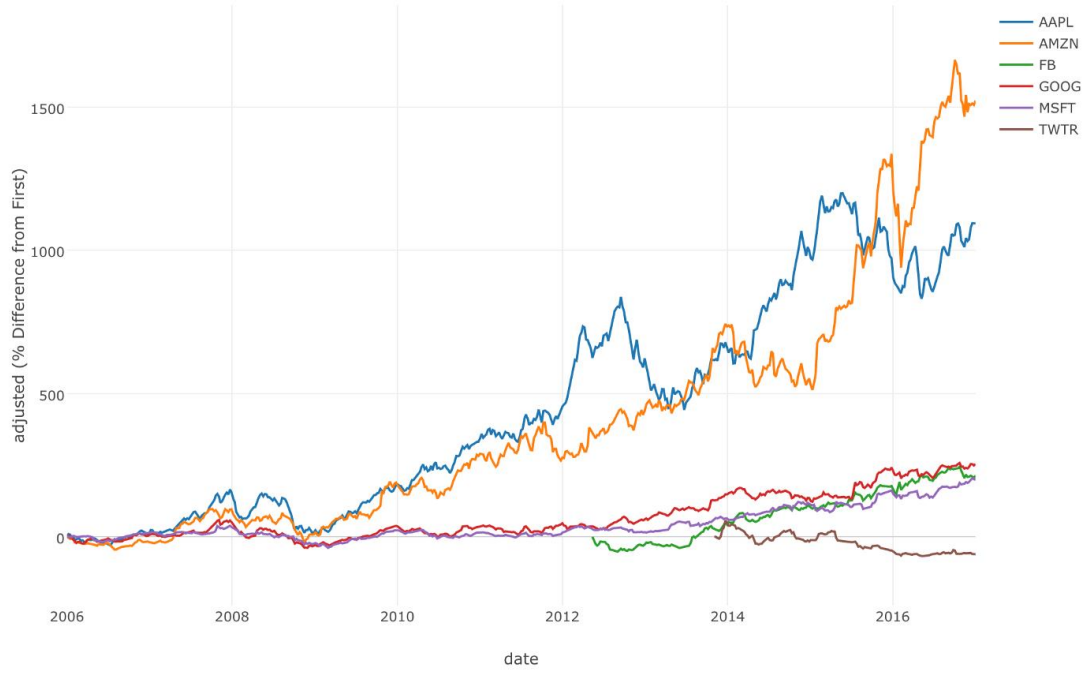


https://florence20.typepad.com/renaissance/2013/02/the-big-data-of-plant-genomics.html

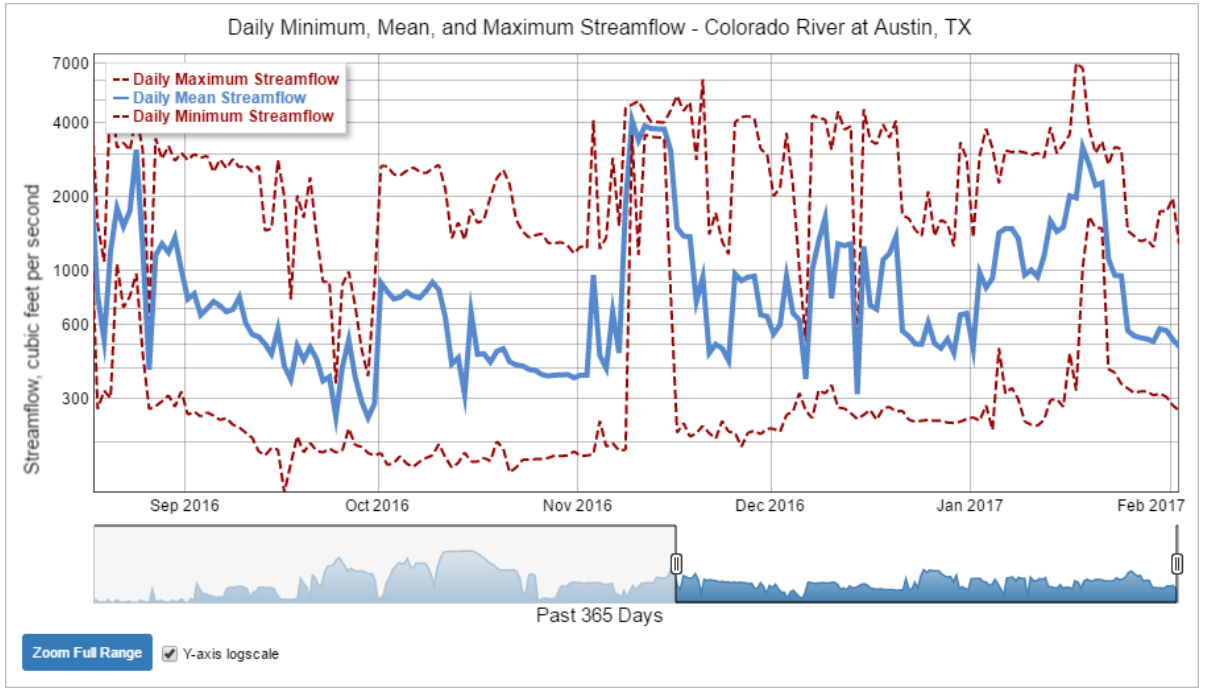
시계열 데이터Time series data

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

- sequential data의 특수한 경우
- 시간에 따른 속성의 변화를 관찰한 데이터 집합
- 예: 주가 지수, 시간별 기온 변화



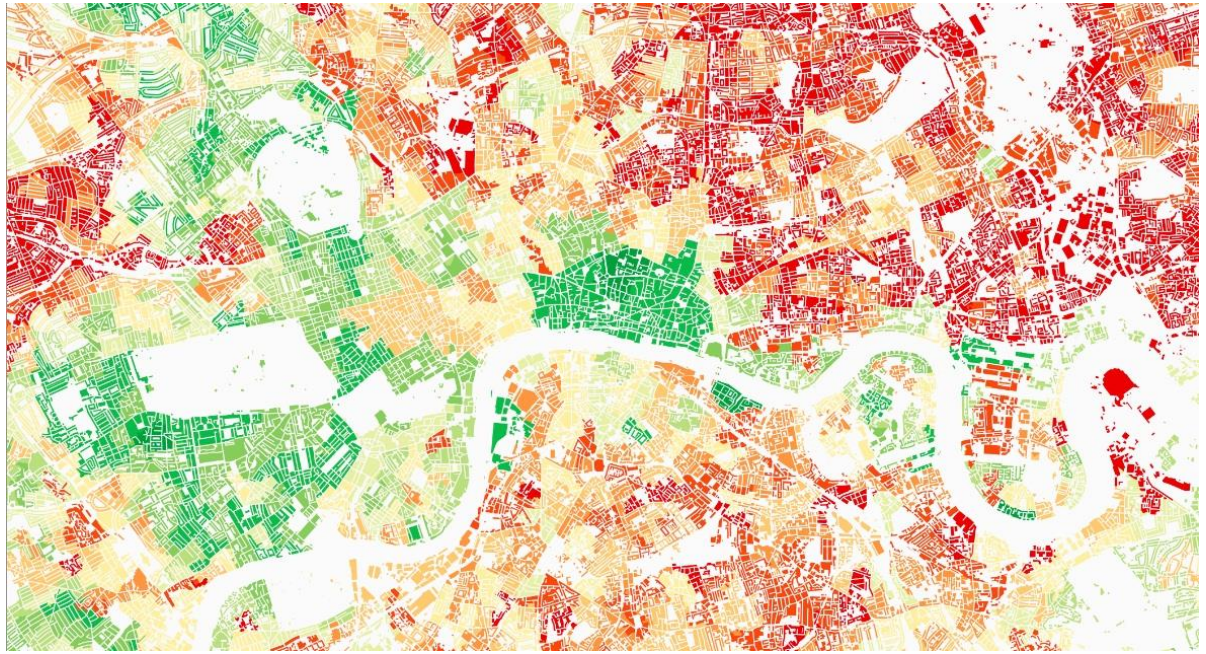
https://blog.exploratory.io/introduction-to-tidyquant-quantitative-financial-analysis-for-tidyverse-habitats-e5f72a023ce2



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

공간 데이터Spatial 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/

Q & A