
데이터 과학
Data Science
Product
Brand
Success
Network
Information
Development
Analysis

데이터 단어 유래

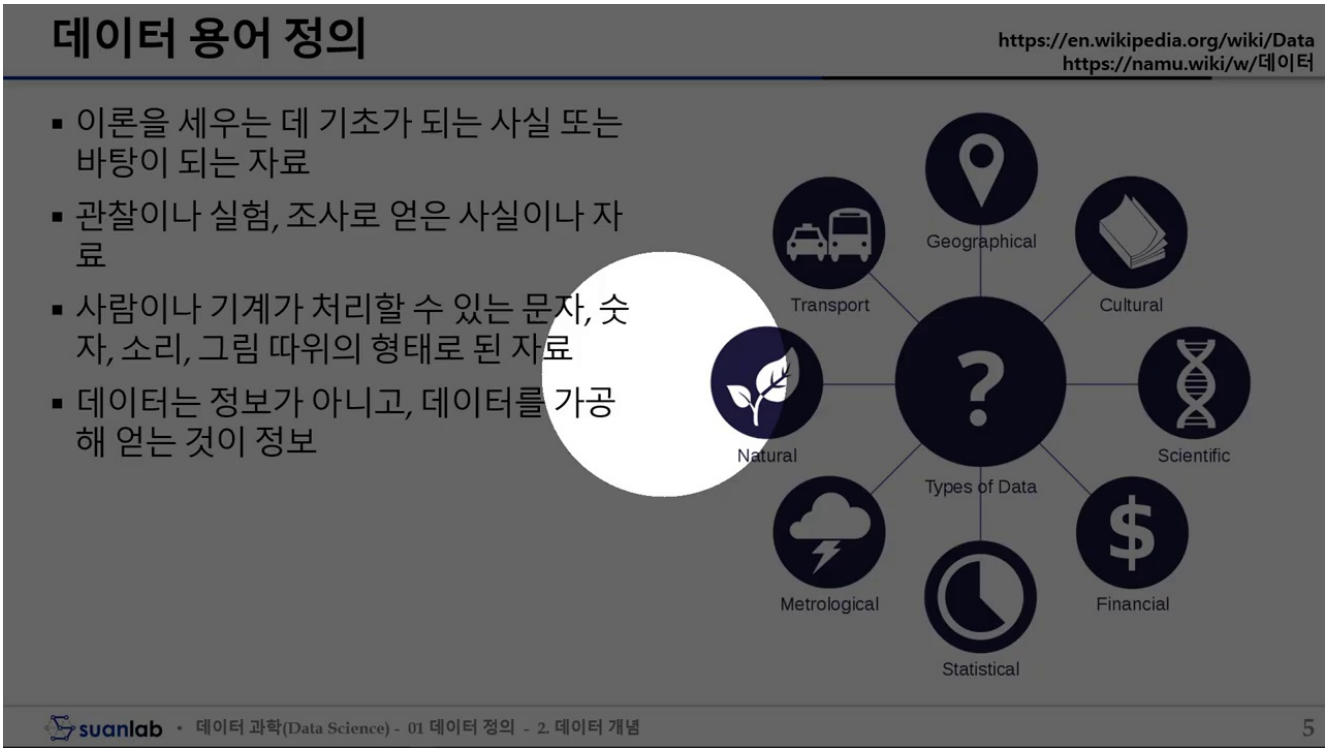
https://en.wikipedia.org/wiki/Data
https://namu.wiki/w/데이터

- 데이터는 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
- 사람이나 기계가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
- 데이터는 정보가 아니고, 데이터를 가공해 얻는 것이 정보


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
사람이나 기계가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
데이터는 정보가 아니고, 데이터를 가공해 얻는 것이 정보
Geographical
Transport
Cultural
Natural
Types of Data
Scientific
Metrological
Financial
Statistical
suanlab • 데이터 과학(Data Science) - 01 데이터 정의 - 2. 데이터 개념
5

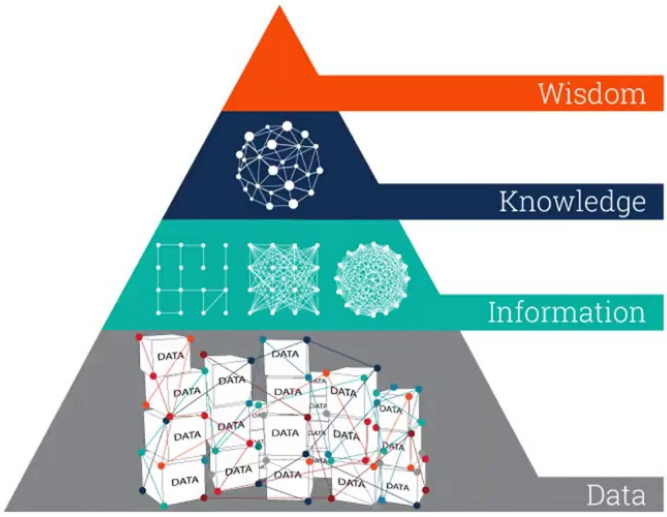
데이터 용어 정의

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

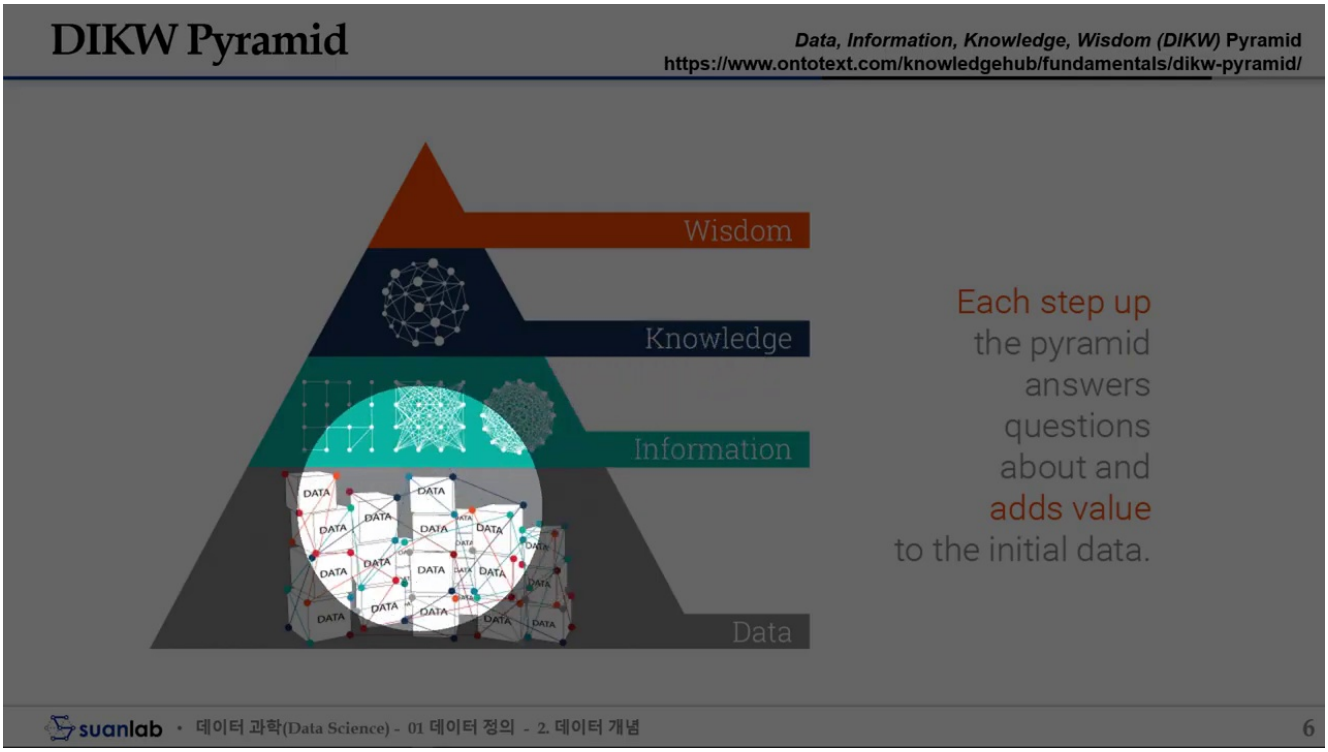
- 이론을 세우는 데 기초가 되는 사실 또는 바탕이 되는 자료
- 관찰이나 실험, 조사로 얻은 사실이나 자료
- 사람이나 기계가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
- 데이터는 정보가 아니고, 데이터를 가공해 얻는 것이 정보

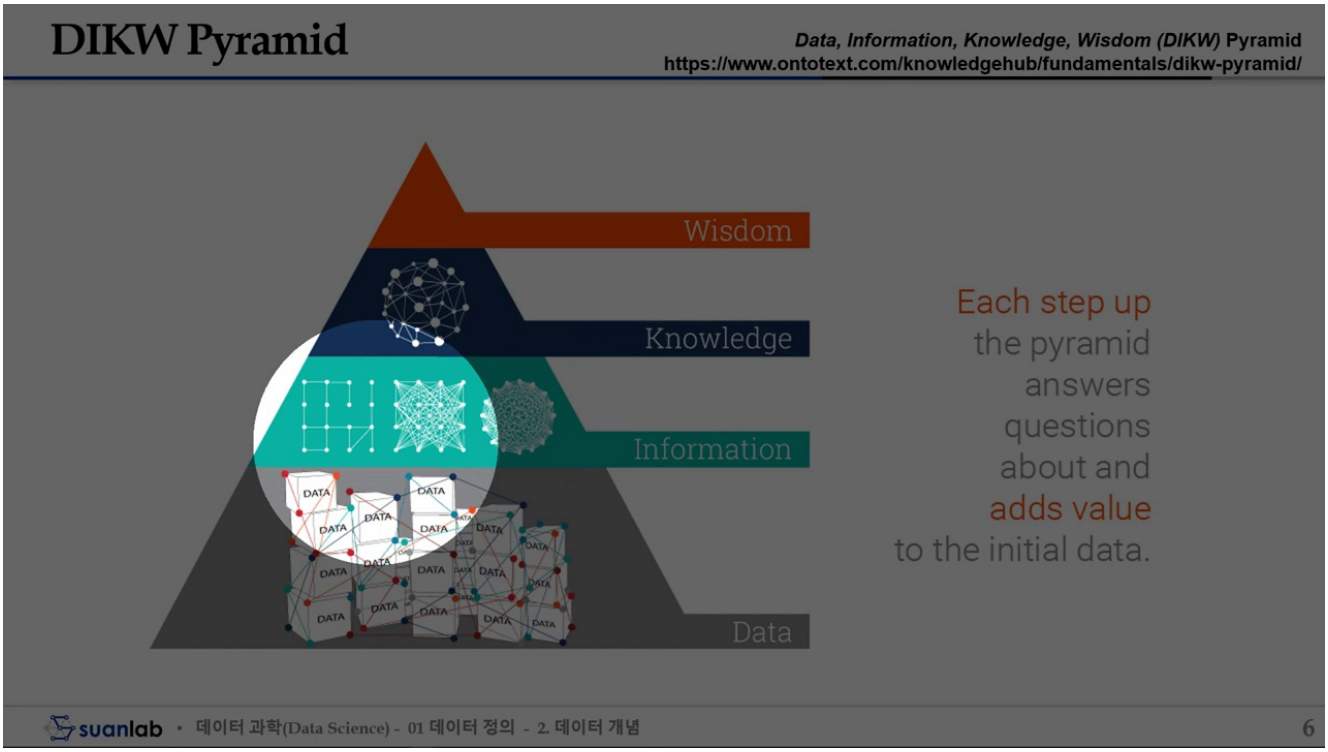
DIKW Pyramid

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



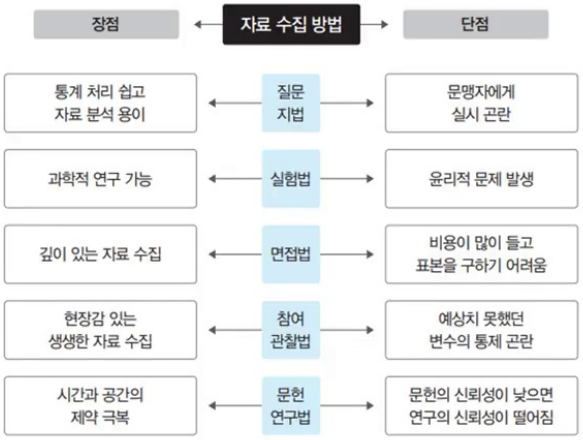
Each step up the pyramid answers questions about and adds value to the initial data.

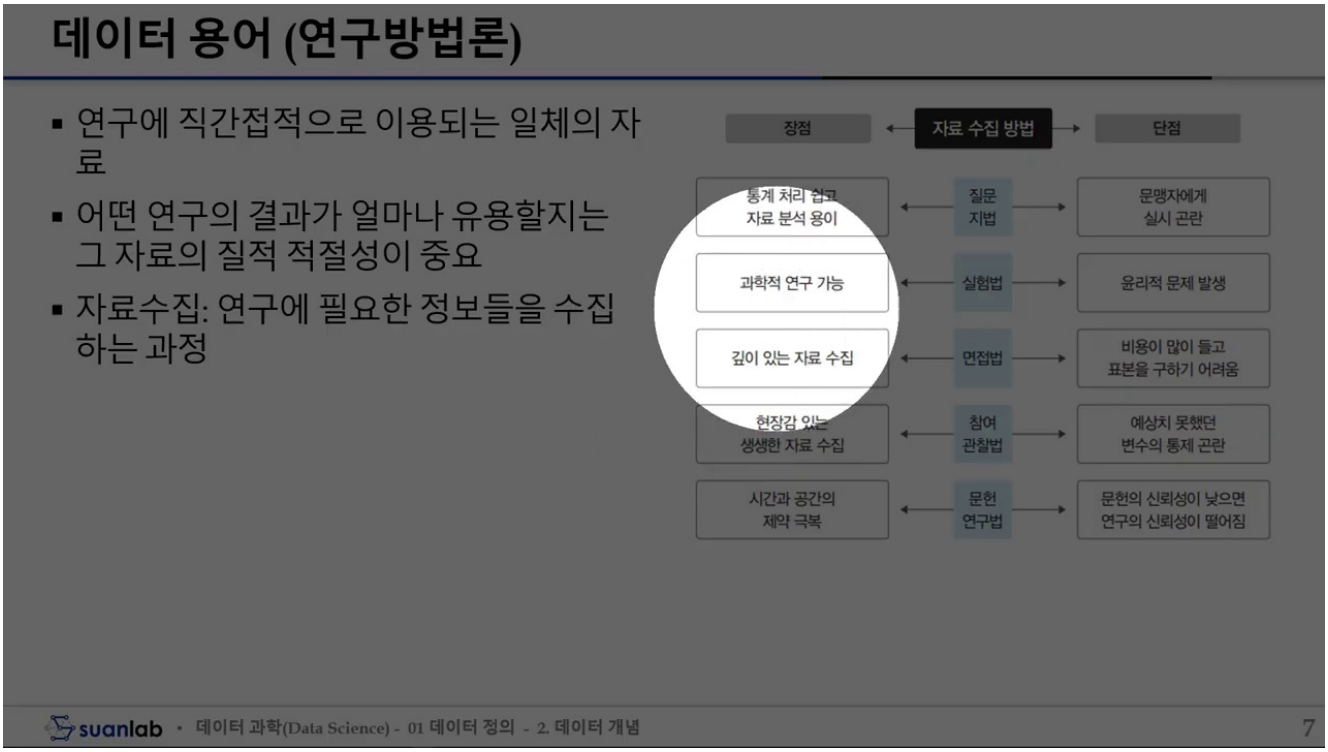

DIKW Pyramid
Data, Information, Knowledge, Wisdom (DIKW) Pyramid
https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/
Wisdom
Knowledge
Information
Data
DATA
Each step up the pyramid answers questions about and adds value to the initial data.
suanlab • 데이터 과학(Data Science) - 01 데이터 정의 - 2. 데이터 개념
6


DIKW Pyramid
Data, Information, Knowledge, Wisdom (DIKW) Pyramid
https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/
Wisdom
Knowledge
Information
Data
DATA
Each step up the pyramid answers questions about and adds value to the initial data.
suanlab • 데이터 과학(Data Science) - 01 데이터 정의 - 2. 데이터 개념
6

데이터 용어 (연구방법론)

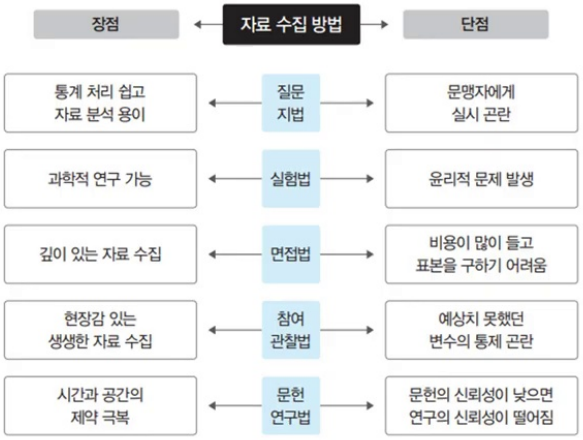
- 연구에 직간접적으로 이용되는 일체의 자료
- 어떤 연구의 결과가 얼마나 유용할지는 그 자료의 질적 적절성이 중요
- 자료수집: 연구에 필요한 정보들을 수집하는 과정




데이터 용어 (연구방법론)
연구에 직간접적으로 이용되는 일체의 자료
어떤 연구의 결과가 얼마나 유용할지는 그 자료의 질적 적절성이 중요
자료수집: 연구에 필요한 정보들을 수집하는 과정
장점
자료 수집 방법
단점
통계 처리 쉽고 자료 분석 용이
질문지법
문맹자에게 실시 곤란
과학적 연구 가능
실험법
윤리적 문제 발생
깊이 있는 자료 수집
면접법
비용이 많이 들고 표본을 구하기 어려움
현장감 있는 생생한 자료 수집
참여 관찰법
예상치 못했던 변수의 통제 곤란
시간과 공간의 제약 극복
문헌 연구법
문헌의 신뢰성이 낮으면 연구의 신뢰성이 떨어짐
suanlab • 데이터 과학(Data Science) - 01 데이터 정의 - 2. 데이터 개념
7

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

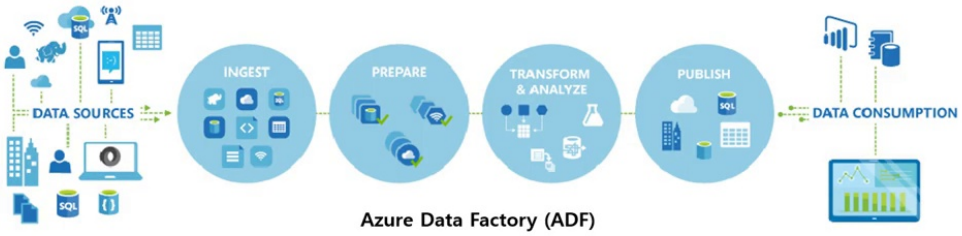
- **S 자료: 자기보고 데이터**
  - 어떤 대상에 대한 정보를 얻을 때 그 대상에게 직접 물어보아 얻은 자료
  - 당연히 사람을 대상으로 하므로, 그 분야는 심리학이나 사회학 등에 한정될 수밖에 없음
  - 매우 흔하게 접할 수 있는 자료로, 흔한 설문조사나 여론조사 등을 통해 얻어짐
  - "사람은 자신이 자신을 제일 잘 안다" 는 전제에 기초해 있으며, 사회적 선망에 의해 답변이 왜곡될 수 있음

데이터 용어 (컴퓨터)

- 프로그램에 부속된 파일, 특히 사용자가 해독할 수 없는 형태의 이진 파일
- 컴퓨터에 의해 특정한 방법으로 처리되거나 해석될 목적으로 순서를 가지고 나열된 기호(Symbol)가 모여있는 것
- 수치화된 크기/규모(Magnitude), 개수(Quantity), 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, OMR 카드 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)

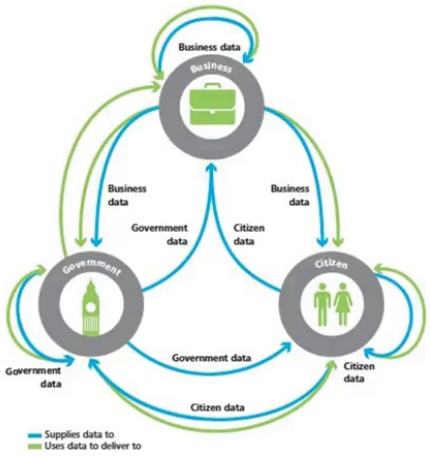
- 2010년 이후 데이터의 시대라고 부르기도 하며, 일부는 심지어 산업혁명 4.0이라고 부르기함
- 데이터 유통 분야
  - 데이터 팩토리(data factory)라는 새로운 개념의 회사들이 생겨났는데, 다른 말로는 데이터 뷰로(data bureau)라고 불리기도 함
  - 가치 있는 데이터들을 수집, 저장, 가공, 통합하여 재판매하는 일을 주로 하고 있음
  - 엡실론(Epsilon), 액시엄(Acxiom), 이퀴팩스(Equifax) 같은 회사들이 유명
  - 국내에도 KCB, NICE , SK 지오비전, 네이버 등이 데이터 팩토리로 불릴 수 있음



Azure Data Factory (ADF)

데이터 용어 (경영학)

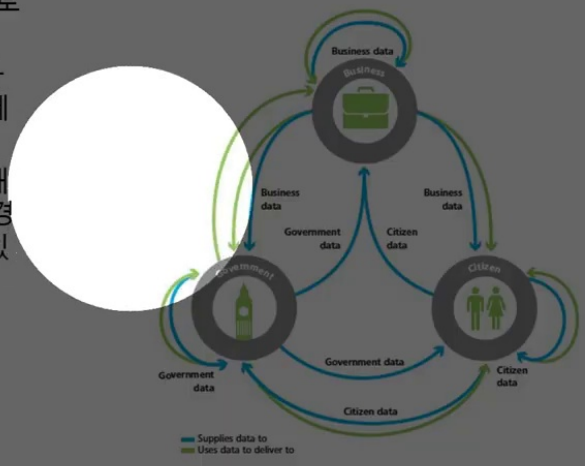
- 금융 분야
  - 데이터 생태계라 하여 콜렉터, 브로커, 유저로 나누어지는 순환구조를 가정
  - 데이터는 판매자가 과거 판매했던 데이터가 이후 다시 특정 "사인(sign)"을 달고 판매자에게 되돌아오는 식으로 구성
  - 데이터 소비자는 구입한 데이터에 자신의 내부 데이터를 융합시켜서 활용하고, 그러한 경제활동을 통해서 데이터 판매자에게 가치 있는 데이터가 다시 전달되는 형태



Open data ecosystem from Deloitte

데이터 용어 (경영학)

- 금융 분야
  - 데이터 생태계라 하여 콜렉터, 브로커, 유저로 나누어지는 순환구조를 가정
  - 데이터는 판매자가 과거 판매했던 데이터가 이후 다시 특정 "사인(sign)"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, 컴파일 된 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 저주(Curse 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


suanlab
이수안 컴퓨터 연구소
suan computer laboratory