데이터 시각화의 기초

Basic of data visualization

데이터 시각화의 기초

데이터의 차원: 1차원

- 하나의 차원 층위를 가진 데이터 의미
- 기울기 없는 직선 그래프, 막대 그래프, 파이 도표 등

2호선			
역 이름	거리(km)	누적 거리	시간
시청	0.0	0.0	
을지로 입구	0.7	0.7	2
을지로 3가	0.8	1.5	2
을지로 4가	0.6	2.1	1
동대문역사문화공원	1.0	3.1	2
신당	0.9	4.0	2
상왕십리	0.9	4.9	2
…	…	…	…

역 이름_순차 항목
[n번째 역]=[구간 거리]

데이터 시각화의 기초

데이터의 차원: 1차원 + 시간

- 어제, 한달 전, 1년 전 특별한 주기를 이용하여 데이터를 측정
- 하지만 데이터가 동일하게 측정되었다면 이는 정보라 부를 수 없음
- 시간에 따른 변화를 기록한 그래프
  > 주식 시장 그래프, 대통령 선거의 사전조사 변화 추이
  분기별 물가 상승 곡선 등…
- 시간과 함께 변화하는 데이터는 우리 일상 그 자체이며, "시계열 데이터"라 부름

시계열 데이터는 "값", "시간" 이라는 두 개의 축을 가지고 있으므로 2차원 데이터라고 볼 수 있음

시간이라는 축은 언제나 같은 속도로 어느 누구에게나 혹은 어떤 사물에나 공평하게 적용되는 예외가 없는 축

데이터 시각화의 기초

데이터의 차원: 1차원 + 시간

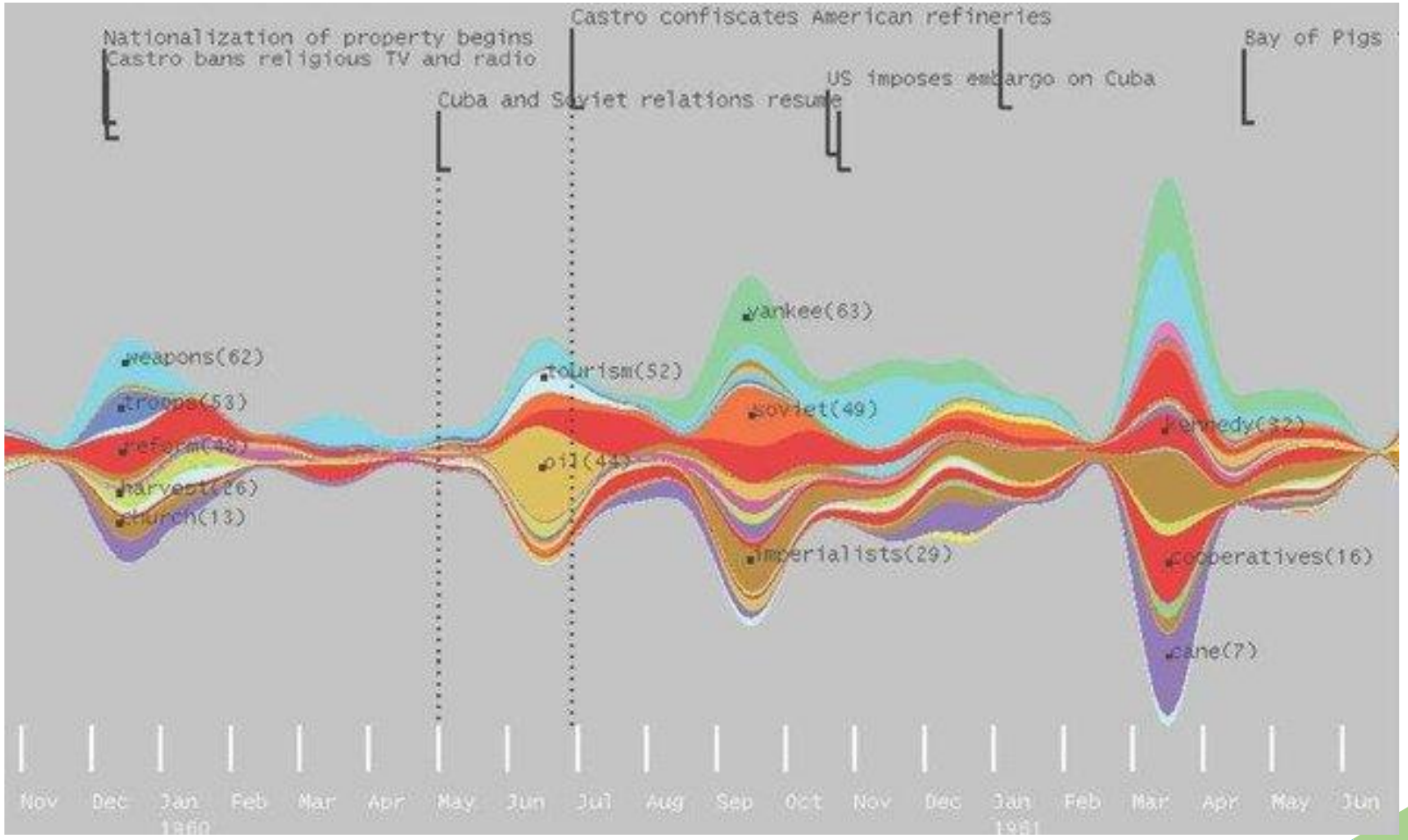
- 정보 이론의 엔트로피 (entropy) : 무작위성과 유사한 뜻
- 엔트로피가 높다 = 예측 불가능하다 = 정보가 많다
- 규칙성이 있거나 노이즈를 제거한 데이터에는 정보가 적다 라고 말함
- 그러므로 정보가 많고 = 엔트로피가 높고 = 예측 불가능성이 크면
  > 압축하기가 힘들고 암호를 깨기도 힘들다

시간.. 정량적으로 증가,, 변화가 없는 것과 마찬가지 = 정보가 없다 라고 할 수 있음

즉, 엄밀히 말해 시간 데이터는 정보를 주는 것이 아니므로 정보성이 떨어진다라고 정리할 수 있음

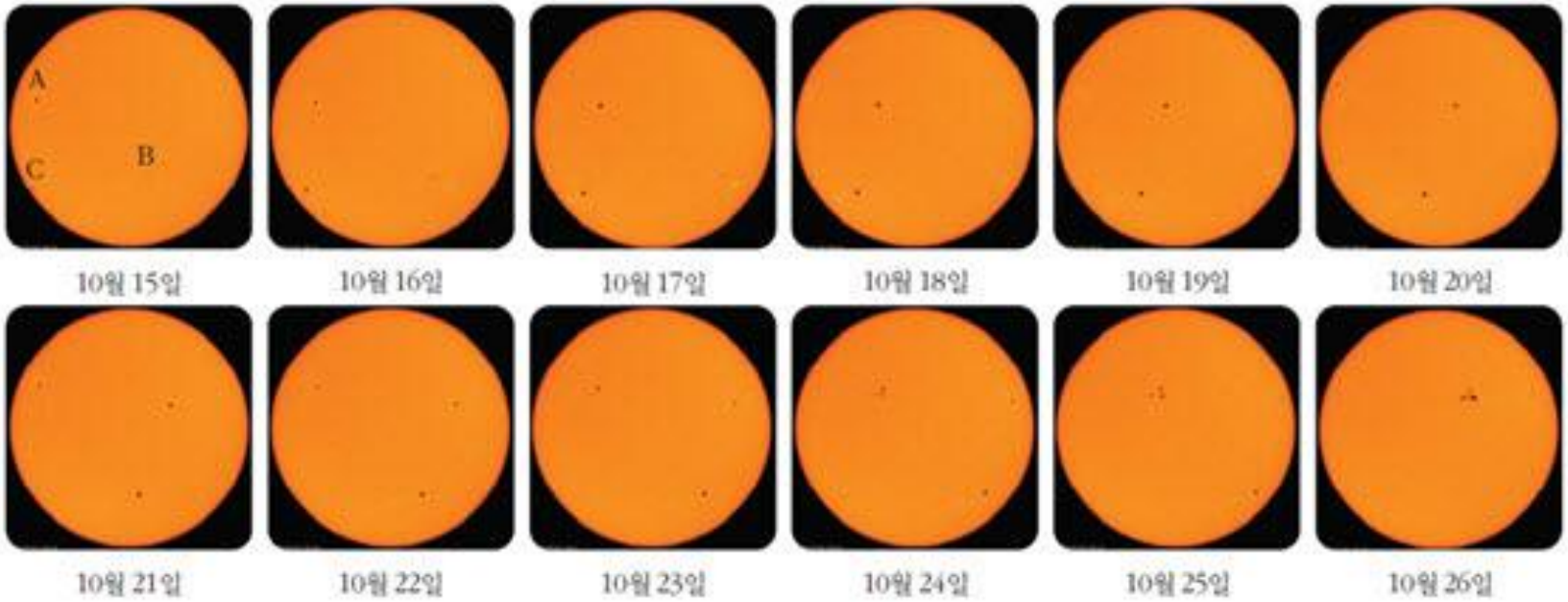
데이터 시각화의 기초

데이터의 차원: 1차원 + 시간



데이터 시각화의 기초

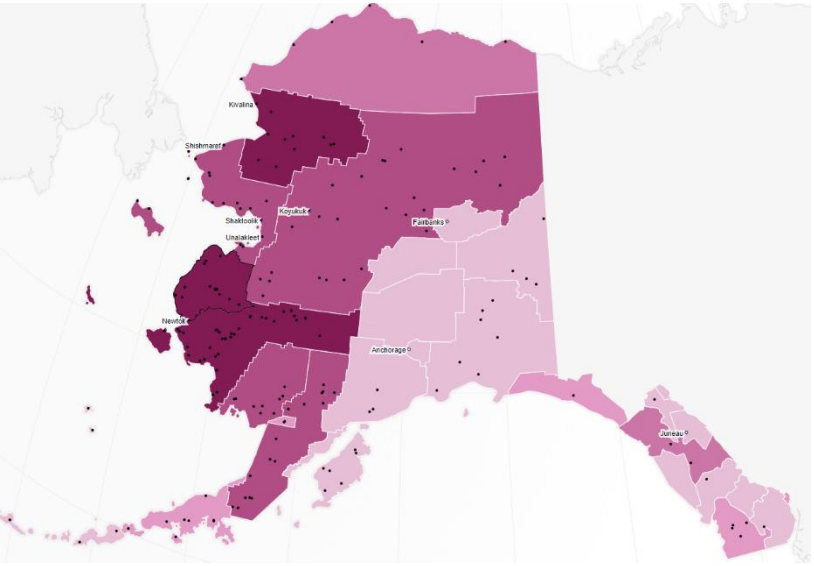
데이터의 차원: 1차원 + 시간



데이터 시각화의 기초

데이터의 차원: 2차원

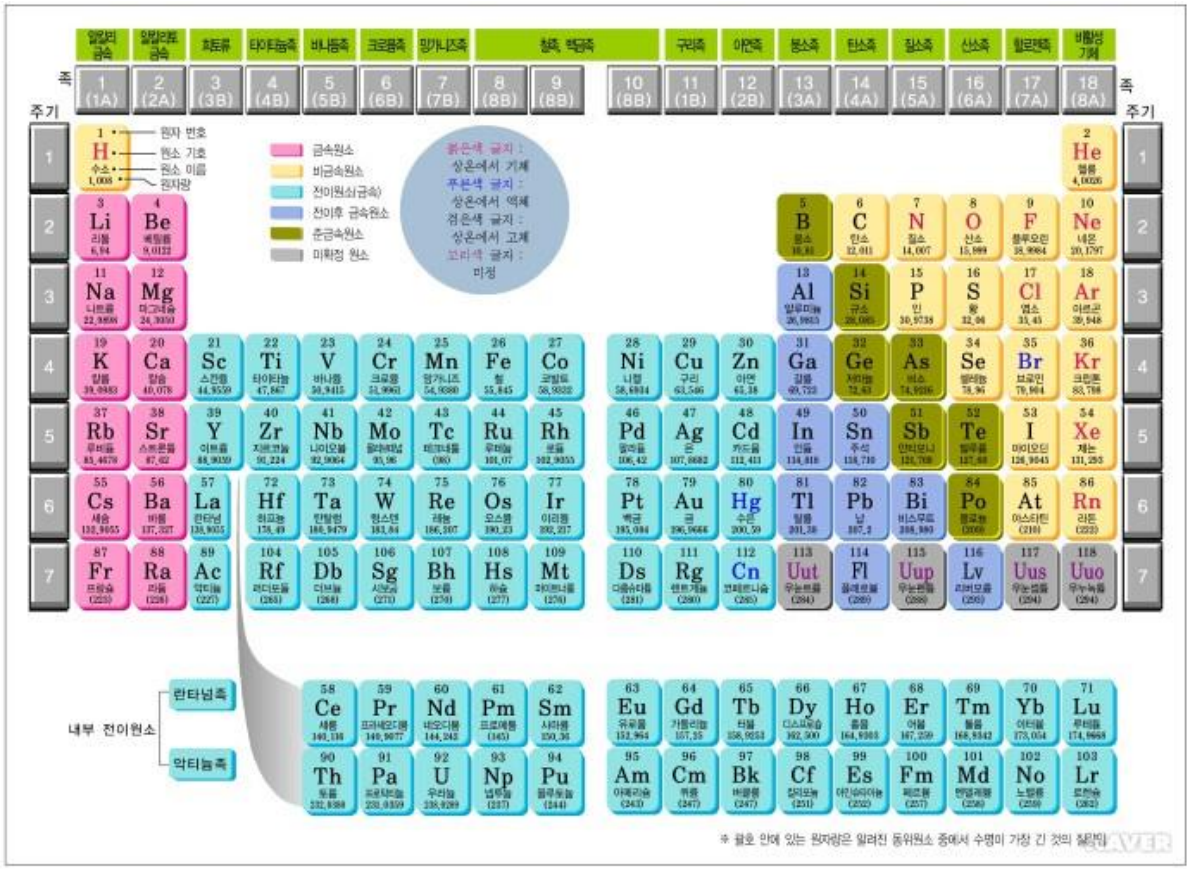
- 서로 다른(독립된) 두 가지 차원을 가지고 있음
  예) 위, 경도라는 지구 위치 정보 데이터
- 2차원 시각화 원조 – 지도 (카토그라피, cartography)
- 우리나라 카토그라피 - 대동여지도



데이터 시각화의 기초

데이터의 차원: 2차원

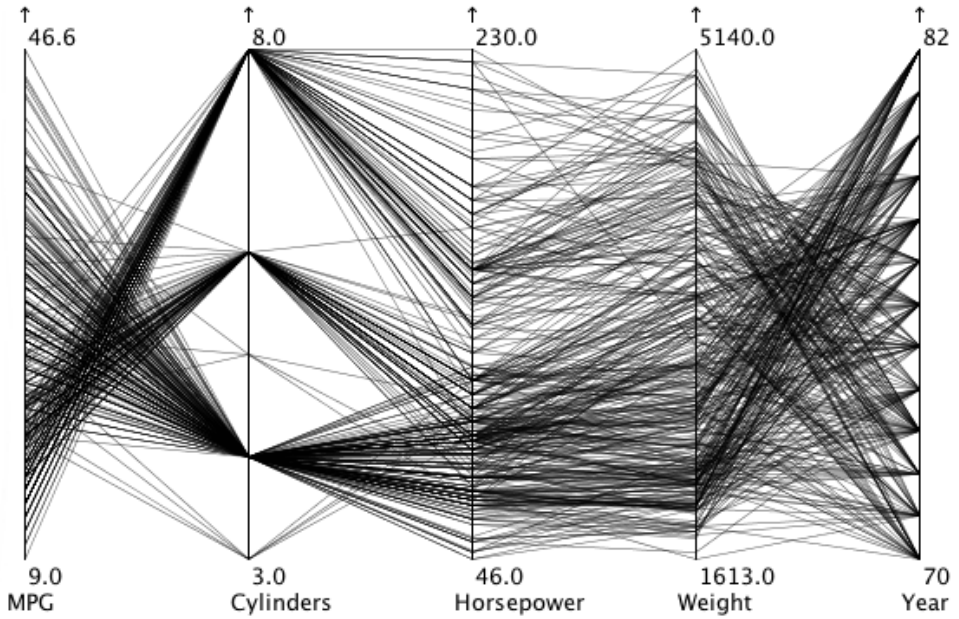
- 전체적인 데이터를 1차원으로 해석하는 것이 자연스러운 상태에서 이를 2차원 배열로 바꾸었을 때, (원소 주기율표)



데이터 시각화의 기초

데이터의 차원: 다차원

- 실제 세상은 훨씬 복잡
- 실제 대부분 데이터 집합은 세 개 이상 속성으로 구성
- 다차원(또는 다변수) 데이터 사례
  > 관계 데이터베이스 테이블
- 표현 했지만 읽는 법을 모르면 마찬가지로 혼란스러움
  (해당 분야에 대한 기본 지식)
- 네트워크, 상호 참조 그래프, 산포도 (3차원) 등 형태



데이터 시각화의 기초

데이터의 관계

- 데이터 관계?
  > 데이터 사이 연결

- 데이터 차원은 비교적 독립적, But 특성의 경우 이를 독립적으로 보기 힘듬

- 원시 자료(Raw data)의 경우 데이터를 어떻게 추출하느냐에 따라
  전혀 다른 시각화와 해석 존재

- 차원 또한 데이터에 따라 수도 없이 달라질 수 있으므로 “고정된 차원”을
  설정하기 어려움

- 그렇기 때문에 전체 데이터 특성을 쉽게 파악할 수 있는 방법은 데이터 사이의
  “관계”를 살펴보는 것

데이터 시각화의 기초

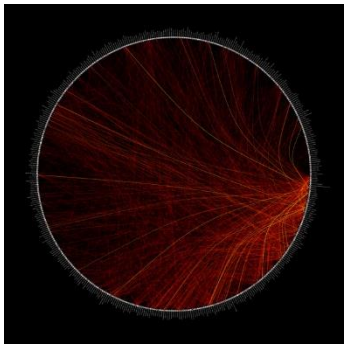
데이터의 관계: 수평적 참조

- 수평적 참조란, 위계가 없는 참조를 얘기함
  예) 하이퍼텍스트, 텍스트

- 텍스트에 관련된 시각화 특성
  > 순차적인 경우 (예: 소설)
  > 데이터 참조인 경우(예: 논문 인용)
  > 독립적 데이터 (예: 사전)

- 주로 다른 정보의 조각과 관계를 맺음

- 연결 그래프 활용 (노드 & 엣지)



Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202: 18-27.

Guo, H., L. Wang and D. Liang. 2016. Big Earth Data from space: a new engine for Earth science. Science Bulletin 61(7): 505-513.

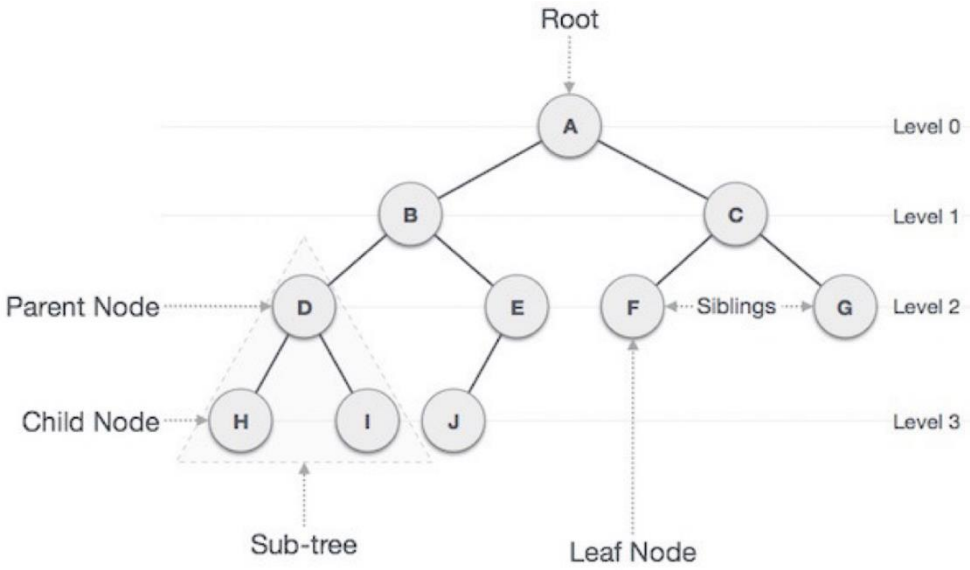
Hu, F., M. Xu, J. Yang, Y. Liang, K. Cui, M. M. Little, C. S. Lynnes, D. Q. Duffy and C. Yang. 2018. Evaluating the Open Source Data Containers for Handling Big Geospatial Raster Data. ISPRS International Journal of Geo-Information 7: 144.

Lewis, A., L. Lymburner, M. B. J. Purss, B. Brooke, B. Evans, A. Ip, A. G. Dekker, J. R. Irons, S. Minchin, N. Mueller, S. Oliver, D. Roberts, B. Ryan, M. Thankappan, R.

데이터 시각화의 기초

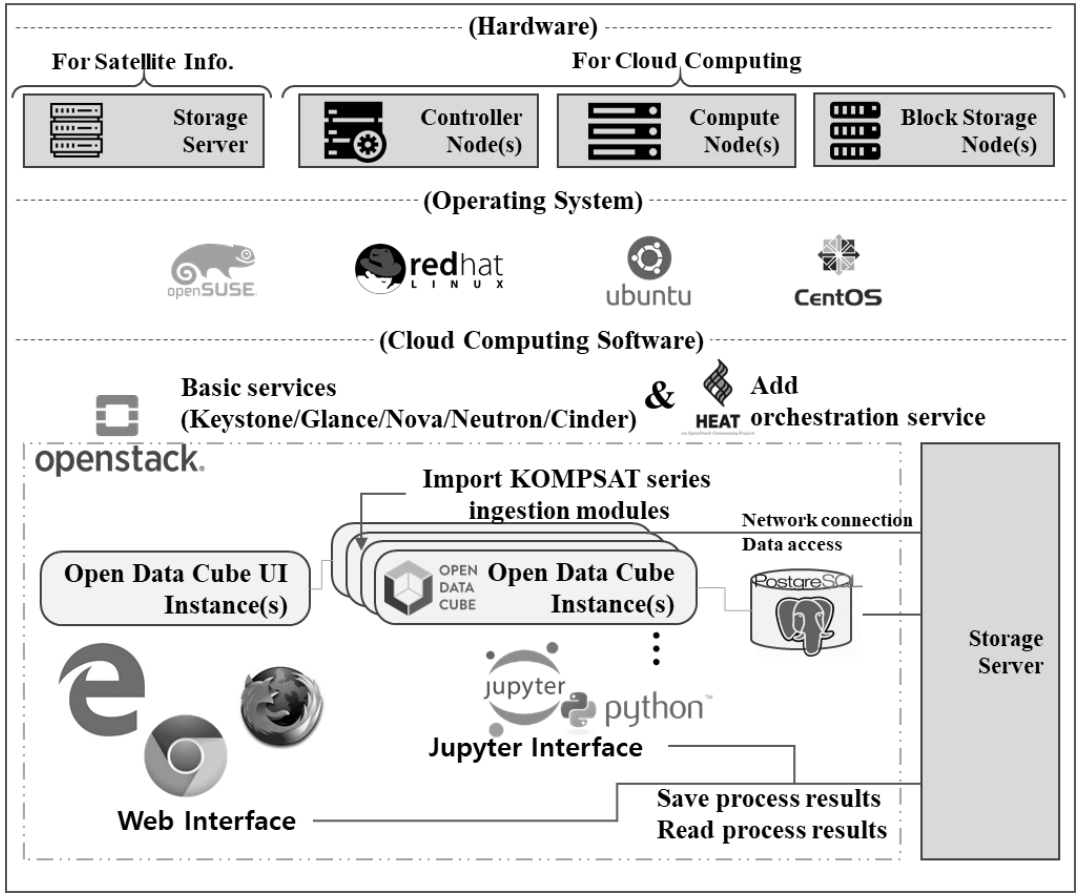
데이터의 관계: 수직적 참조

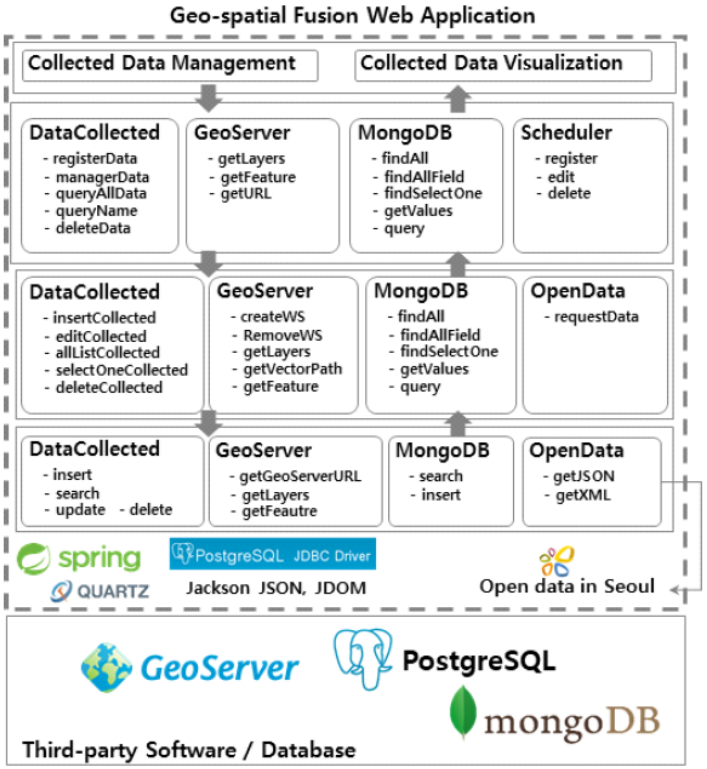
- 명백한 위계가 존재할 경우
- 컴퓨터 공학에서 데이터 구조의 가장 손쉬운 탐색 형태인 “트리 구조”가 대표적인 예
- 시각화시 계층의 명시적 표현이 매우 중요



데이터 시각화의 기초

데이터의 관계: 수직과 수평의 혼재





데이터 시각화의 기초

시각화 기법의 종류: 표

- 시각화 영역에 넣기 힘든 기본적인 데이터 표현법

- 행, 열로 조직된 구조화된 형식으로 정의
  > 행(row) = 기록(record) = 튜플(tuple) = 벡터(vector)
  > 열(column) = 필드(field) = 파라미터(parameter) = 속성(attribute) = 특성(property)

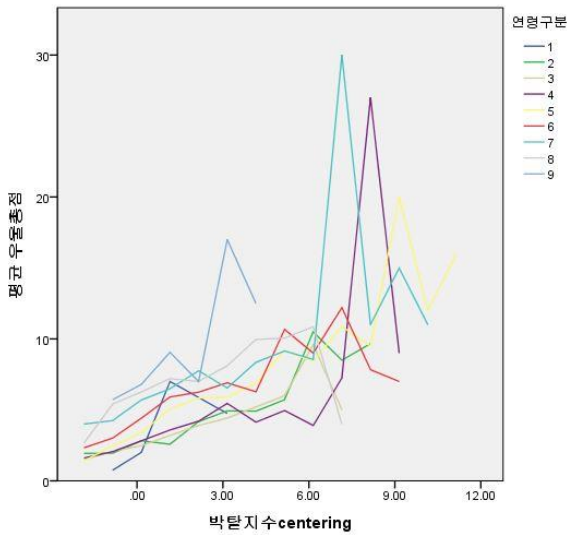
- 순서가 있을 수도 있고, 없을 수도 있음
  → 이는 데이터와 데이터 속성에 의해 결정

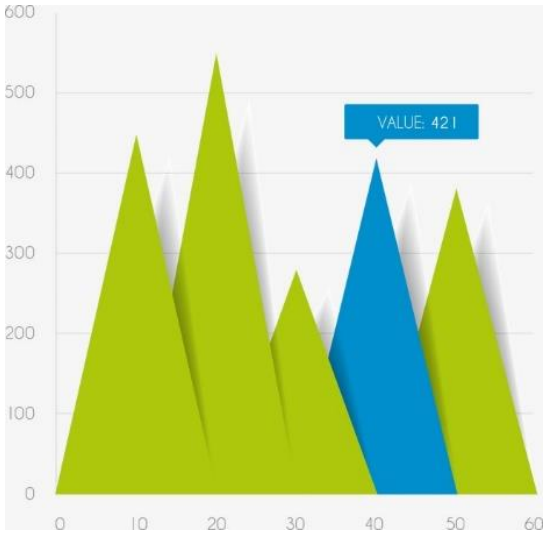
데이터 시각화의 기초

시각화 기법의 종류: 선도표, 면적 도표

- 선도표(line chart): 연결된 점으로 정보를 나타내는 그래프
  > 산포도(scatter plot) 확장된 형태
  > 주로 시간적 간극이 있는 데이터의 트렌드 시각화에 많이 사용 (시계열 데이터)
  > 원형 선 그래프, 표면 그래프, 밀도 플롯, 벡터 그래프, 평행 좌표 등

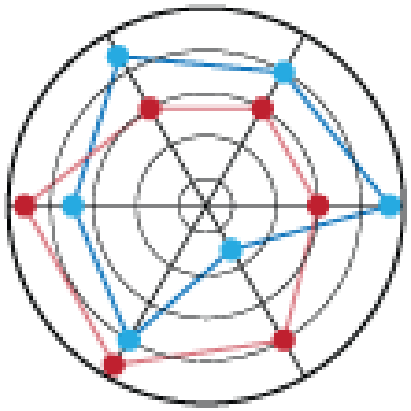
- 면적 도표(area chart): 양적 데이터를 그래픽적으로 표현하는 용도
  >면적 도표의 경계 영역은 선도표를 기초 로함

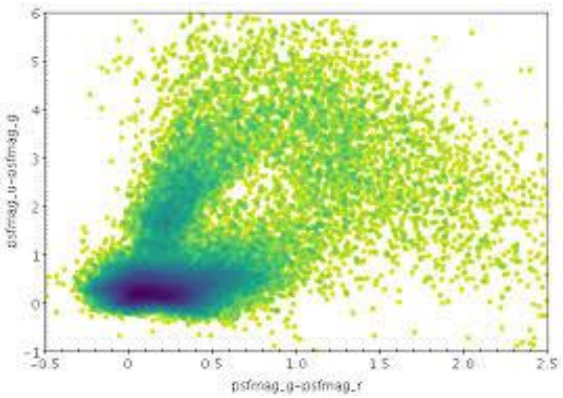


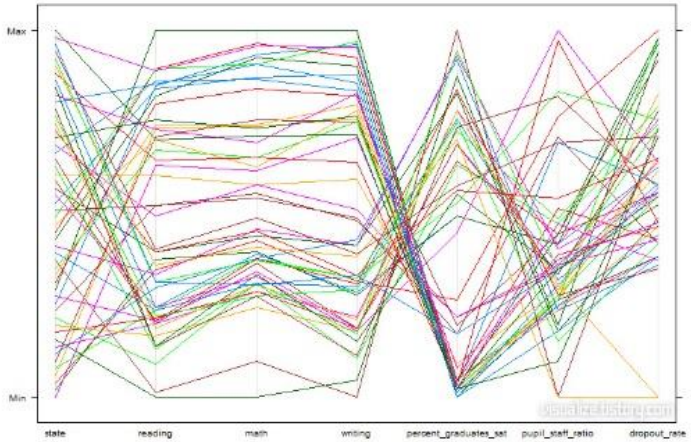


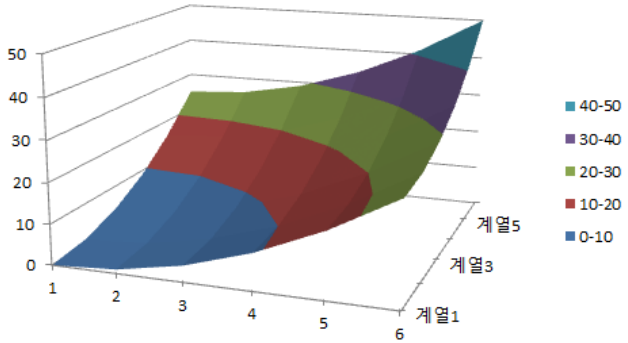
데이터 시각화의 기초

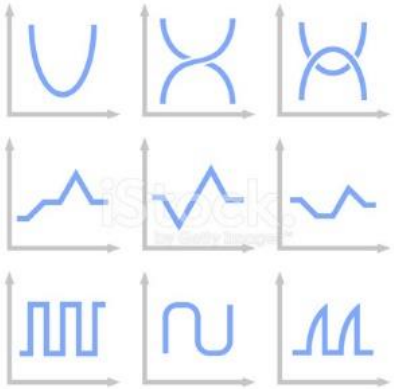
시각화 기법의 종류: 선도표, 면적 도표







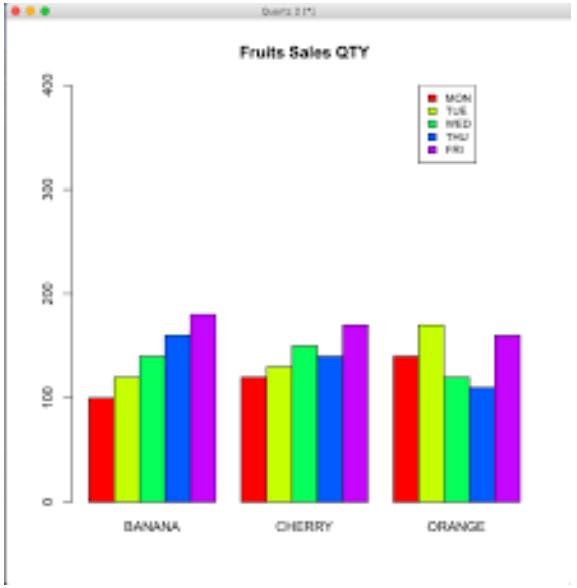


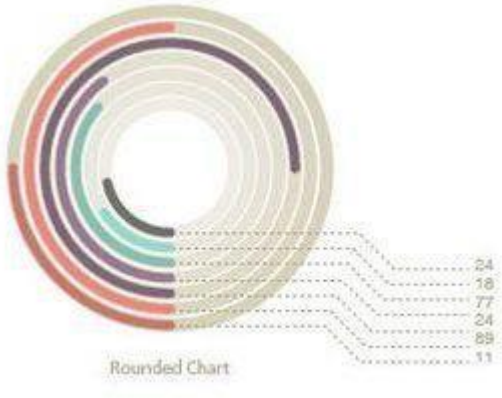


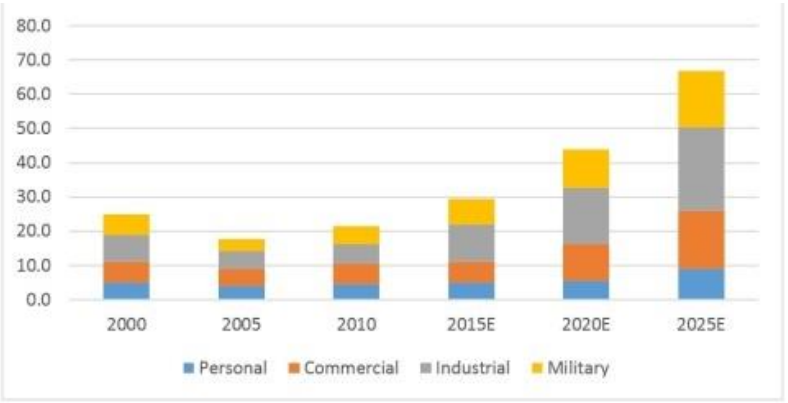
데이터 시각화의 기초

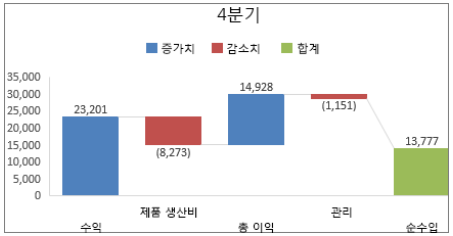
시각화 기법의 종류: 막대 도표

- 데이터 시각화 기법 중 가장 많이 활용되는 것 중 하나
- 세로 단 도표라고 일컫기도 함
- 대부분 연속 데이터가 아닌 분산 데이터 사용
- 막대 방향에 따라 수평, 수직 막대로 표현 가능하며, 막대 길이를 데이터 값으로 표현
- 원형, 누적 막대, 폭포 도표와 같은 다양한 종류도 있음







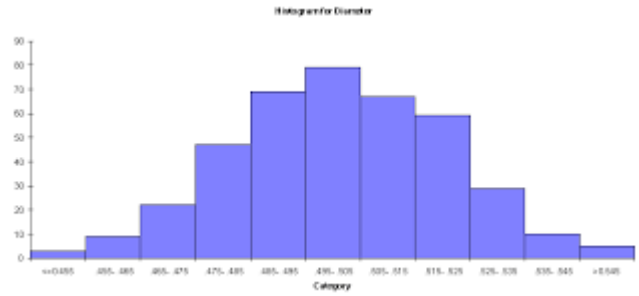


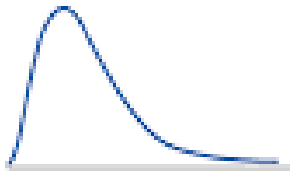
데이터 시각화의 기초

시각화 기법의 종류: 막대 도표

- 장점 : 다양한 값을 상대적으로 비교하고 쉬움
  눈에 띄는 값을 찾기 쉬움
  매우 흔한 표현 방법으로 누구나 쉽게 이해
  단점 : 단순화된 표현 → 때에 따라 막대의 실제 값 등 부가 설명 필요
  패턴을 발견하기 힘듬

- 히스토그램의 경우 막대 도표지만 독립적인 항 x
  > 주로 데이터 분석과 통계에 매우 중요하게 활용
  > 가로축 : 계급 / 세로축 : 도수 (간혹 반대가 될 때도 있음)
  > 막대끼지 붙어 있어야함
  > 대칭형 / 좌우 첨도왜도형 / 쌍봉형 / 다봉형





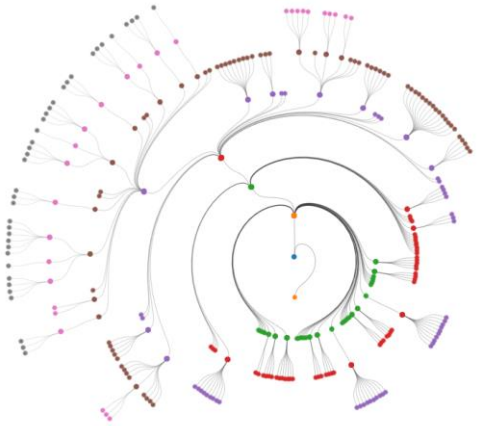
평균이 중앙값보다 크다.

데이터 시각화의 기초

시각화 기법의 종류: 파이 도표

- '원형 도표' 라고도 함

- 대부분 경우 파이 도표는 퍼센트 (%)를 나타냄



- 변형이 일어나 나타내기도 함
  > 부채꼴 그래프
  > 도넛 도표
  > 다층 파이 도표
  > 방사 트리 도표
  > 링 파이 도표

- 파이 도표에서 색상 활용은 데이터 이해와 해석에 큰 도움을 주므로 매우 중요

데이터 시각화의 기초

시각화 기법의 종류: 다이어그램

- 2차원 평면에 도형 등 기호로 정보를 표현하는 수단을 통칭
- 과거 무지한 사람들을 위한 계몽의 수단으로 사용
- 다양한 표현법이 존재, 그 중 “순서도”, “타임라인”, “벤 다이어그램” 등이 가장 많이 공통적으로 사용

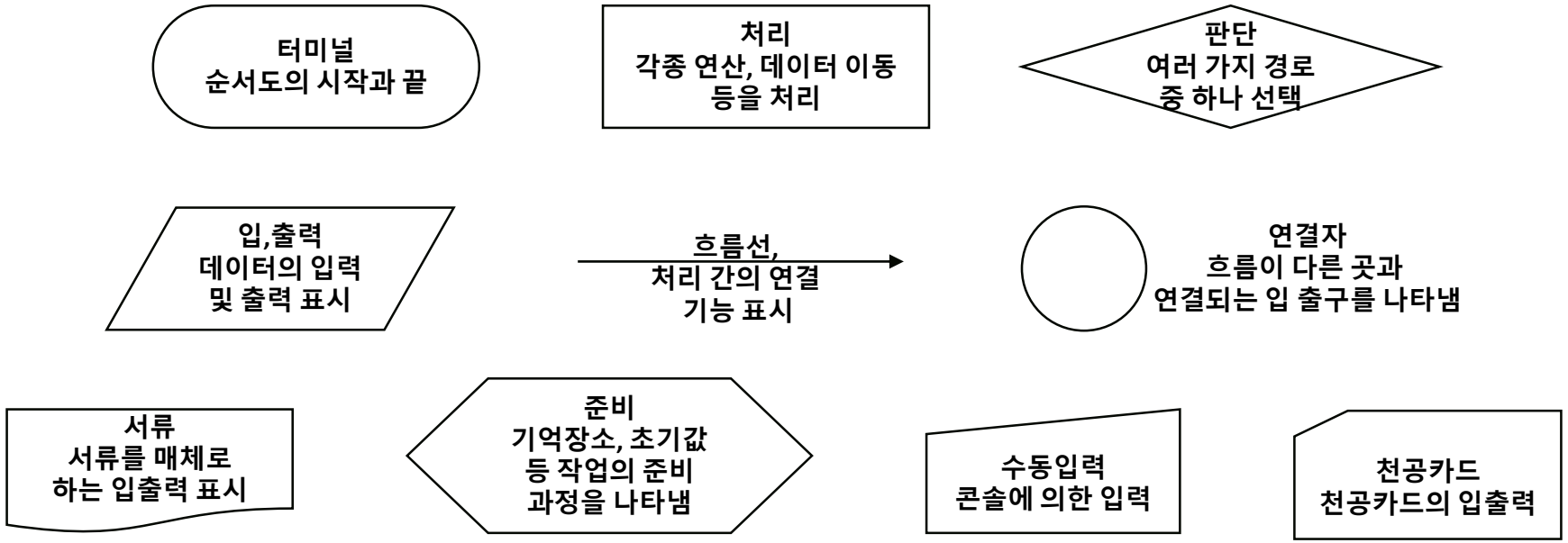
순서도

- 하나의 과정에 관련한 단계의 그래픽적 또는 상징적 재현으로 데이터 흐름의 순서, 지시와 통제의 흐름을 화살표로 표현
- 1921년 기계공학 분야에서 처음 제안
- 순차적 단계(procedural step), 컴퓨터 알고리즘 개발할 때 위력 발휘

데이터 시각화의 기초

순서도

- 순차적 단계의 경우 공통 약속된 표현으로 정해 사용되고 있으며, 현재, 다양한 목적에 따라 변형, 확장되어 제공되고 있음



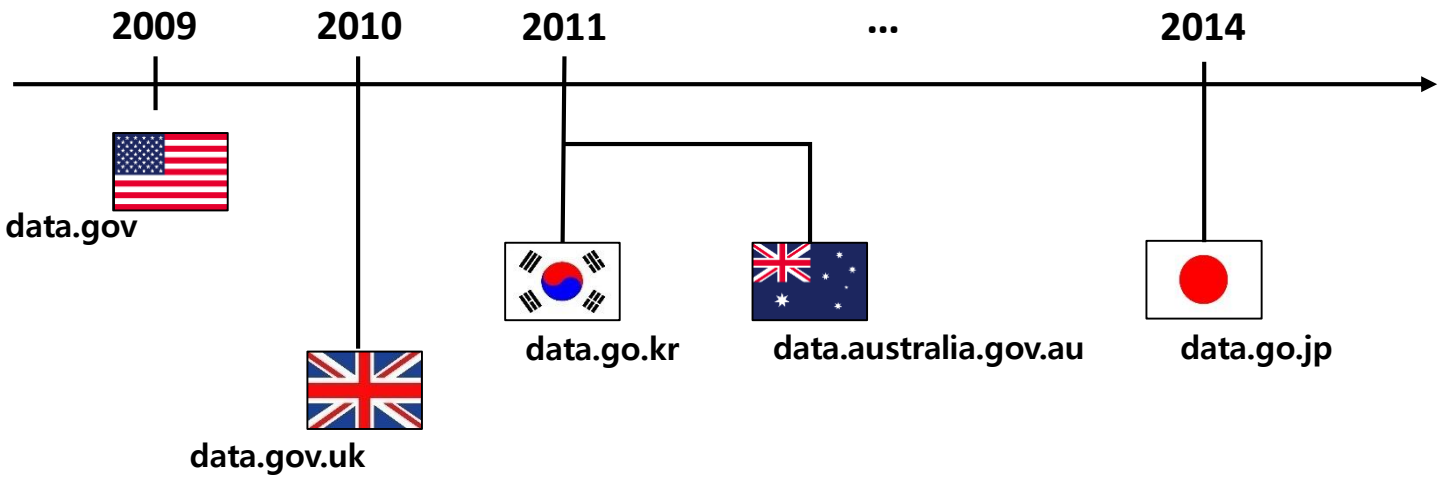
- UML 활동 다이어그램 등이 변경 확장된 사례

데이터 시각화의 기초

타임라인

- 각기 다른 사건 사이의 관계를 쉽게 이해할 수 있도록 시간 순으로 나열한 그래프 (연대표와 비슷한 개념)

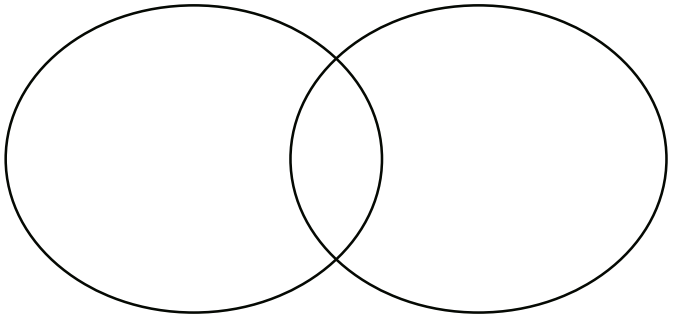
- 선형 / 비교 타임라인으로 구분
  > 선형: 특정 기간 동안 발생한 사건의 순서
  > 각기 다른 장소에서 발생한 두 가지 사건 집합

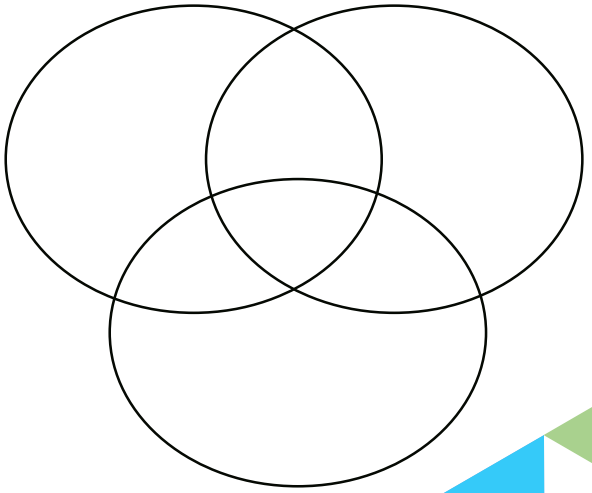


데이터 시각화의 기초

벤 다이어그램

- 1880년 존 벤이 처음 소개
- 2개 또는 그 이상의 집합 간 관계를 설명하는데 사용
- 서로 중첩된 영역이 있는 데이터의 경우 그 특성을 잘 나타낼 수 있지만, 완벽하게 독립적이거나 일부 요인의 영향력이 있는 경우 이를 표현하기 적합하지 않음

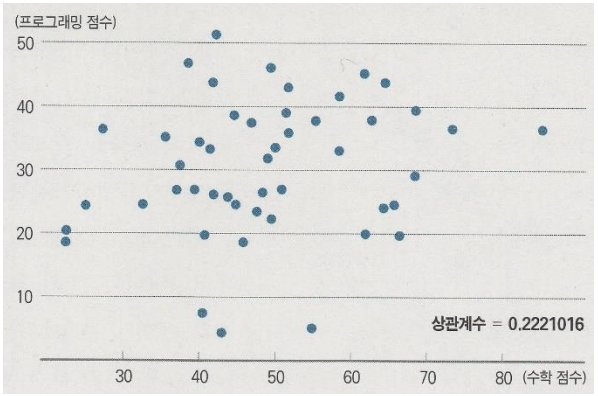


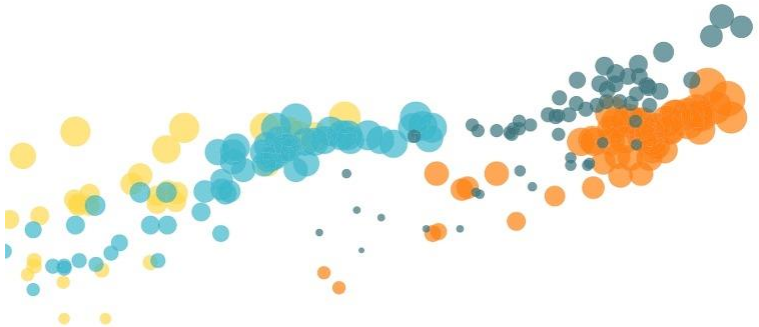


데이터 시각화의 기초

시각화 기법의 종류: 산포도

- 플롯, 플롯 도표, 산점 도표, 분산도, 점도표, 산란 그래프
- 데카르트 좌표 데이터 집합의 그래픽적 표현으로서 두 데이터 간의 상관관계를 나타냄
- 변수들 사이 관계의 강도를 나타내고, 데이터 내부 특이 요소의 존재 여부를 결정 데이터가 분산되어 있는 방식도 보여줌 (군집을 확실히 보여줌)
- 전체 데이터 패턴을 발견하므로 데이터 간 관계 분석에 용이 (최댓값, 최솟값, 노이즈 값 등 쉽게 발견)





데이터 시각화의 기초

시각화 기법의 종류: 텍스트 기반, 의미망

- 매우 폭넓게 활용되며, 새롭게 대두되는 시각화 주제로 관심이 커지고 있는 추세

- '워드 클라우드'는 텍스트 내부의 등장 빈도를 글자 크기로 나타내는 중요 키워드를 표현함으로써 하나의 문서 또는 문서의 집합을 시각화 (자동으로 생성해주는 여러 서비스가 존재)

- 구문망은 용어 사이의 다양한 관계를 드러내려는 목적으로 활용되는 시각화 (위계적 데이터/네트워크 데이터 표현법과도 연결)

- 의미망은 반드시 텍스트일 필요가 없지만, 가장 흔히 쓰이는 분야 중 하나가 언어
  > 서로 다른 개념 사이의 논리적 관계에 대한 그래픽적 표현 기법

데이터 시각화의 기초

시각화 기법의 종류: 텍스트 기반, 의미망

데이터 시각화의 기초

시각화 기법의 종류: 텍스트 기반, 의미망