06강

빅데이터의 이해와 활용

데이터 시각화

한국방송통신대학교
Korea National Open University

학습목차

1 데이터 시각화
2 시간 시각화
3 텍스트 시각화
4 소셜 네트워크 시각화
5 데이터 시각화의 도구

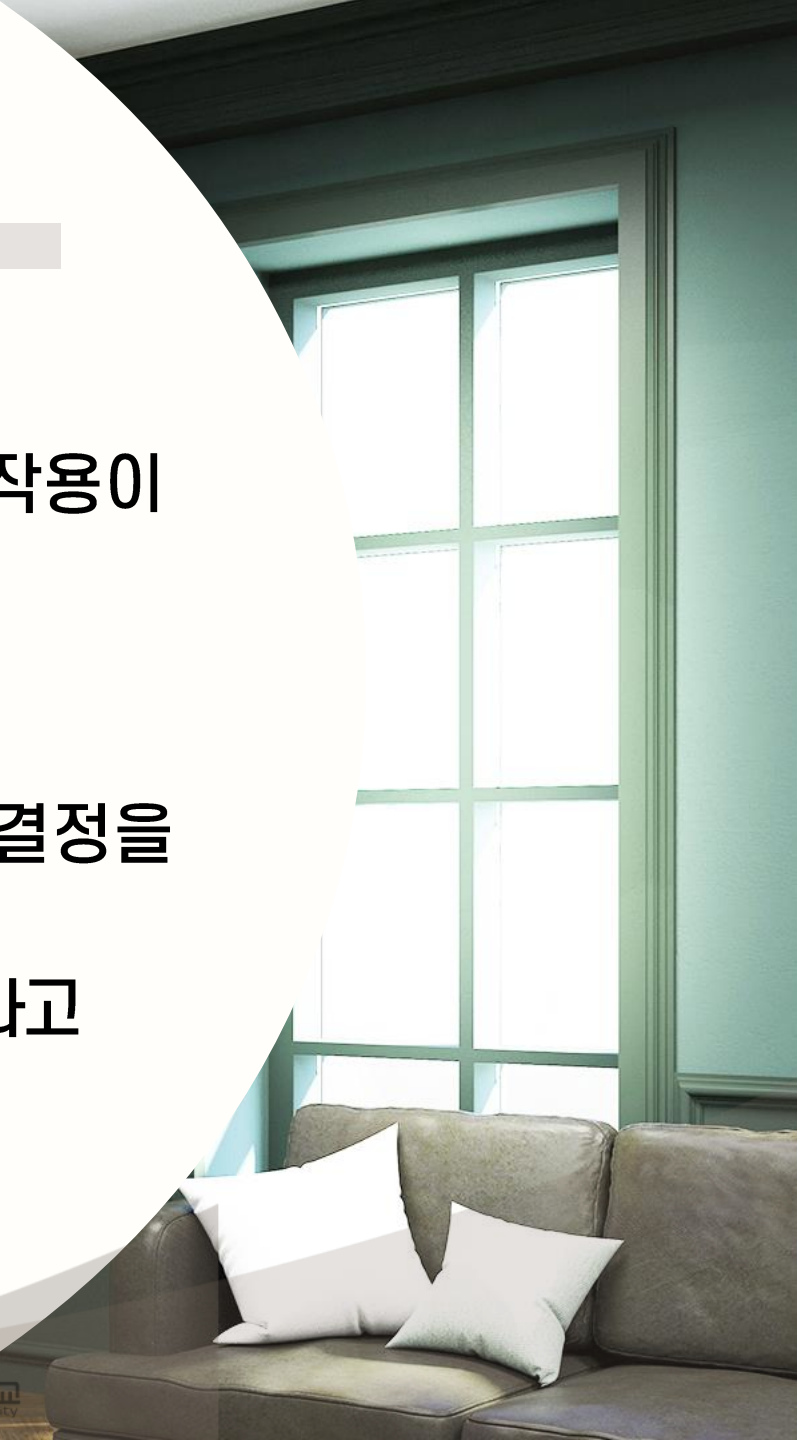
빅데이터의
이해와 활용

1 데이터 시각화

1. 데이터 시각화

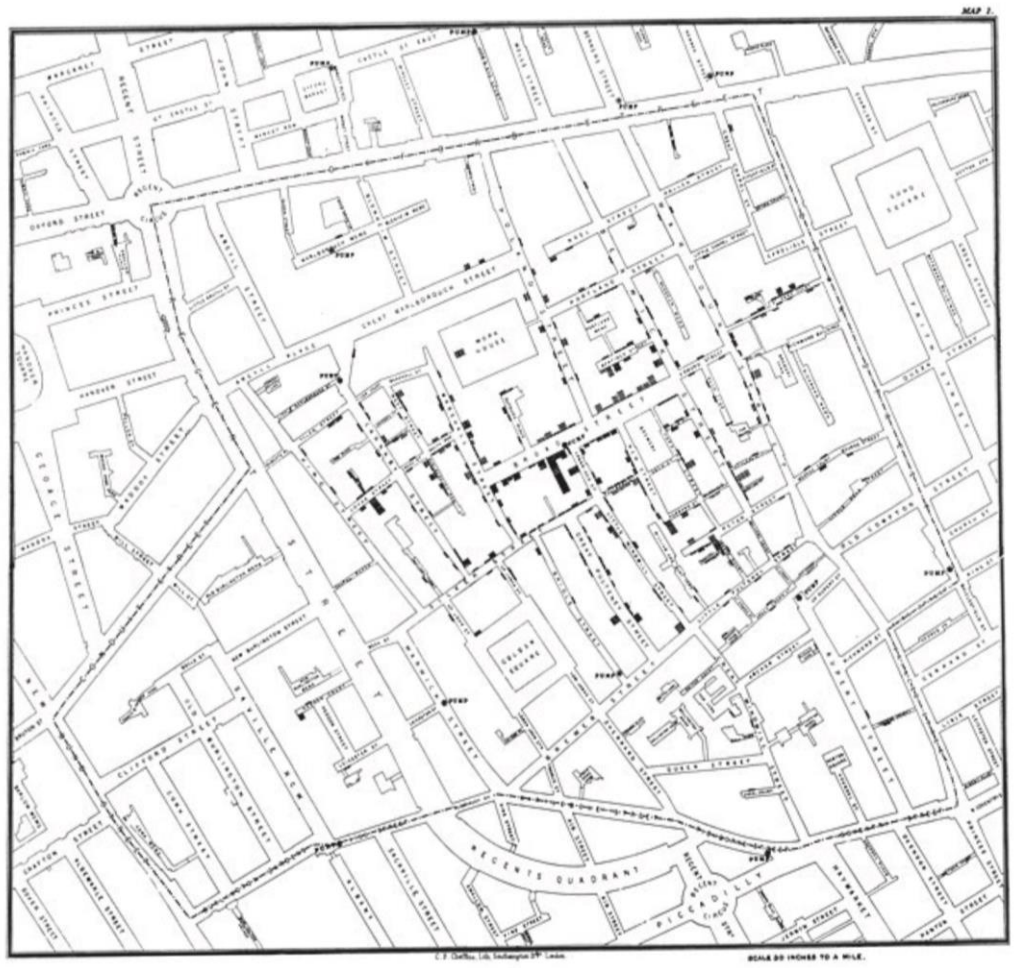
데이터 시각화란?

- “컴퓨터를 사용하여 인지를 넓힐 수 있도록 데이터를 상호작용이 가능한 시각적 형태로 만드는 것” (Card 등, 1998)
- 단순히 이미지를 만드는 과정이 아니라, 이미지를 통해 인지 과정을 도움
- 데이터의 숨은 의미를 발견하고, 설명하고, 이를 통해 의사결정을 내리는데 도움을 주는 통찰력(insight)을 가지게 함
- 이런 이유로 시각화는 외부인지보조(external condition aid)라고 불리우기도 함



1. 데이터 시각화

데이터 시각화란?

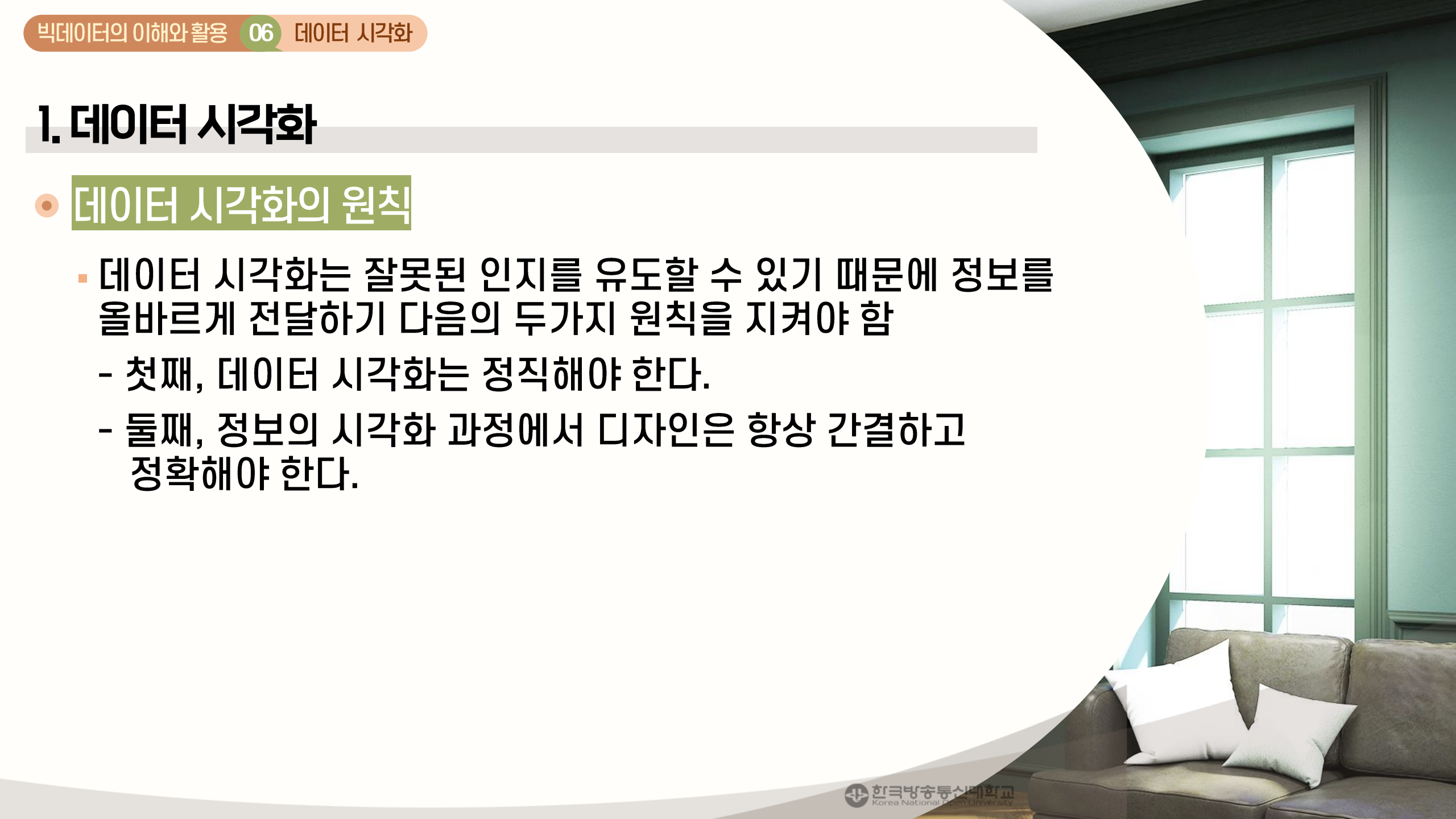


존 스노우의
런던의 콜레라 지도

1. 데이터 시각화

데이터 시각화의 원칙

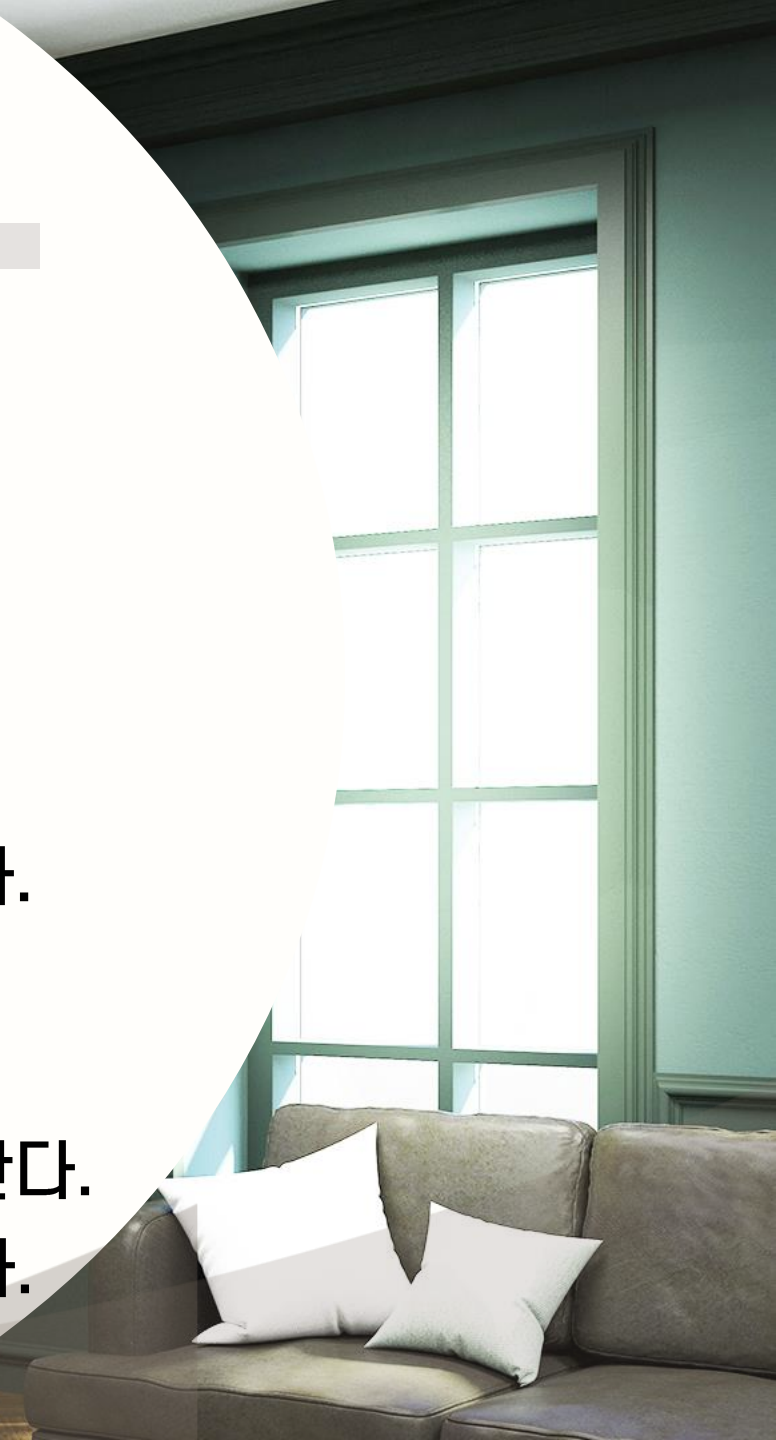
- 데이터 시각화는 잘못된 인지를 유도할 수 있기 때문에 정보를 올바르게 전달하기 다음의 두가지 원칙을 지켜야 함
  - 첫째, 데이터 시각화는 정직해야 한다.
  - 둘째, 정보의 시각화 과정에서 디자인은 항상 간결하고 정확해야 한다.



1. 데이터 시각화

데이터 시각화의 원칙

- 터프트(Tufte, 2001)의 데이터 시각화의 8가지 원칙
  - 첫째, 데이터 그 자체를 보여주는 것이 중요하다.
  - 둘째, 화려한 그래픽에 너무 집중하지 않게 한다.
  - 셋째, 데이터 자체가 말하고자 하는 바를 왜곡하지 말라.
  - 넷째, 너무 많은 정보를 작은 화면에 보여주려고 하지 말라.
  - 다섯째, 많은 양의 데이터가 일관성을 가져야 한다.
  - 여섯째, 서로 다른 데이터를 손쉽게 비교할 수 있게 한다.
  - 일곱째, 데이터는 깊이 들어가 자세히 살펴볼 수 있어야 한다.
  - 여덟째, 통계 결과, 설명을 데이터와 함께 보여주어야 한다.

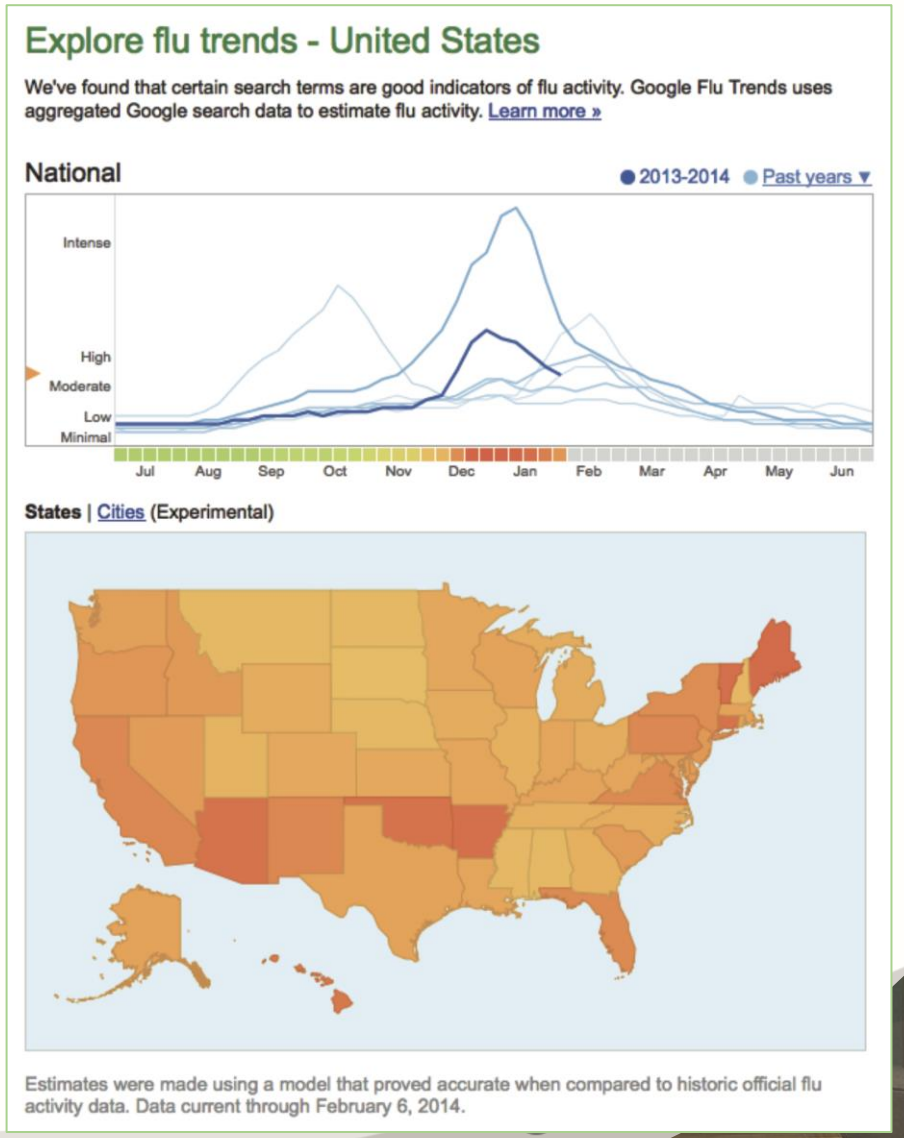


빅데이터의
이해와 활용

2 시간 시각화

2. 시간 시각화

- 시간 시각화
  - 시계열 데이터는 데이터 시각화의 가장 기본적인 형태
  - 이러한 데이터를 통해 우리가 파악할 수 있는 대표적인 것은 경향성(trend)

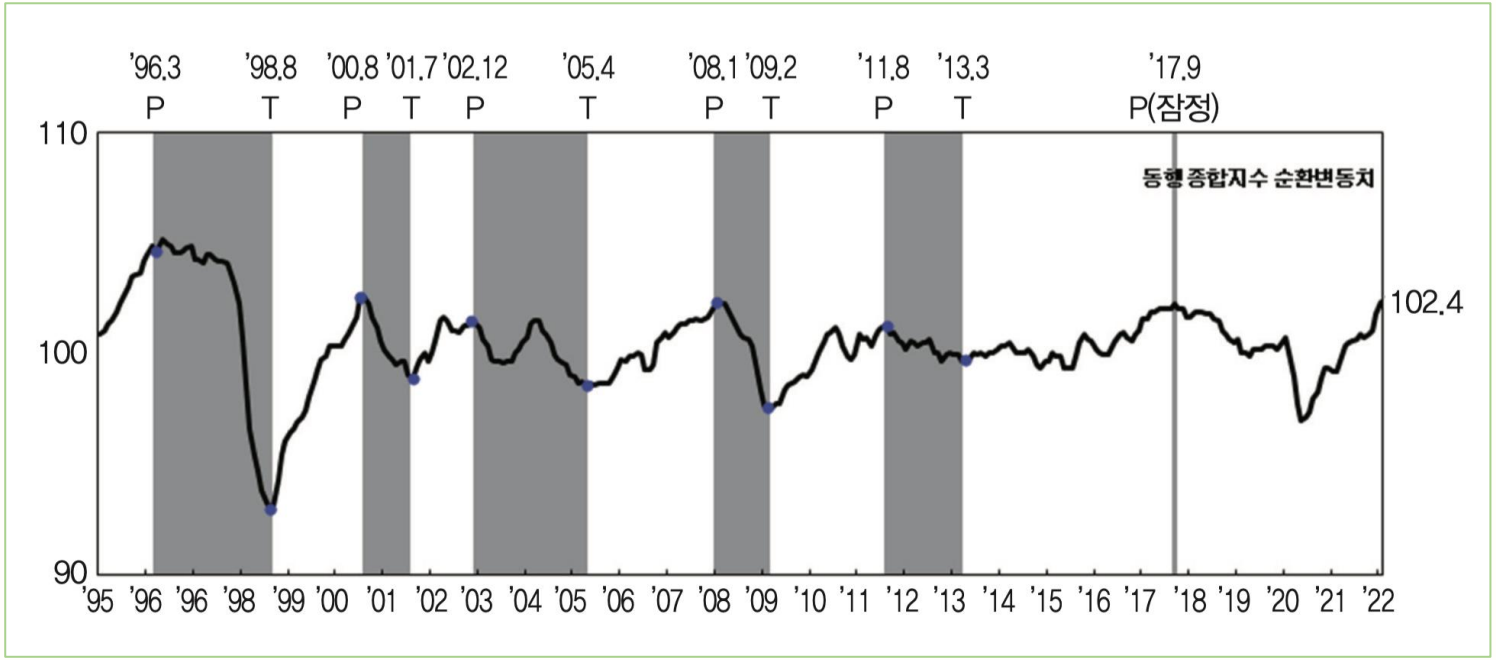


구글 독감 트렌드

2. 시간 시각화

선 그래프

- 대부분의 시계열 데이터를 표현할 수 있는 그래프
- 하나의 그래프에 여러개의 데이터를 그려서 비교할 수 있다는 장점이 있음

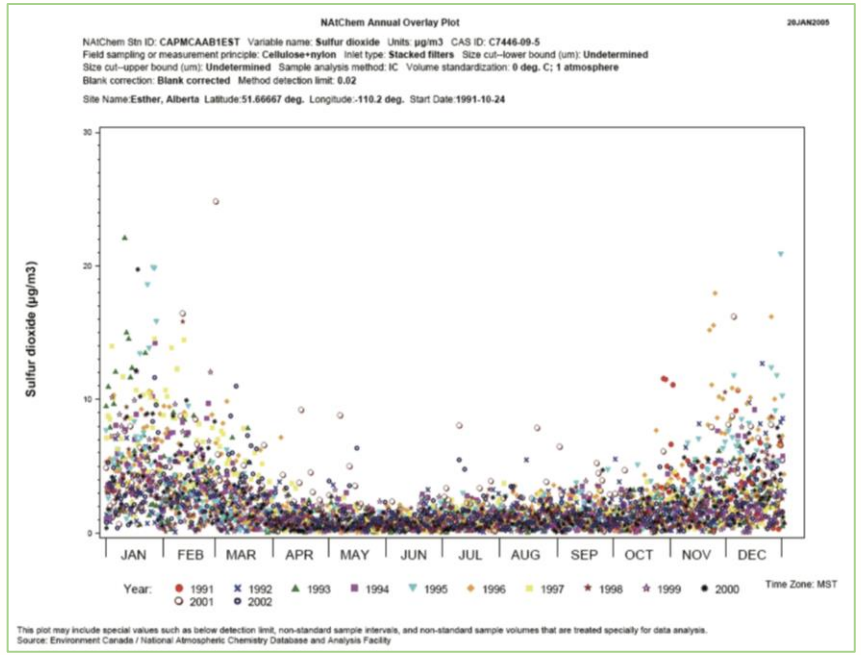


우리나라 경기동행지수 순환변동치의 추이

2. 시간 시각화

점 그래프

- 점 그래프(산점도, Scatter Plot)는 시간에 따라 데이터 포인트로 표현하는 그래프
- 데이터 포인트 수가 많을 때 적합

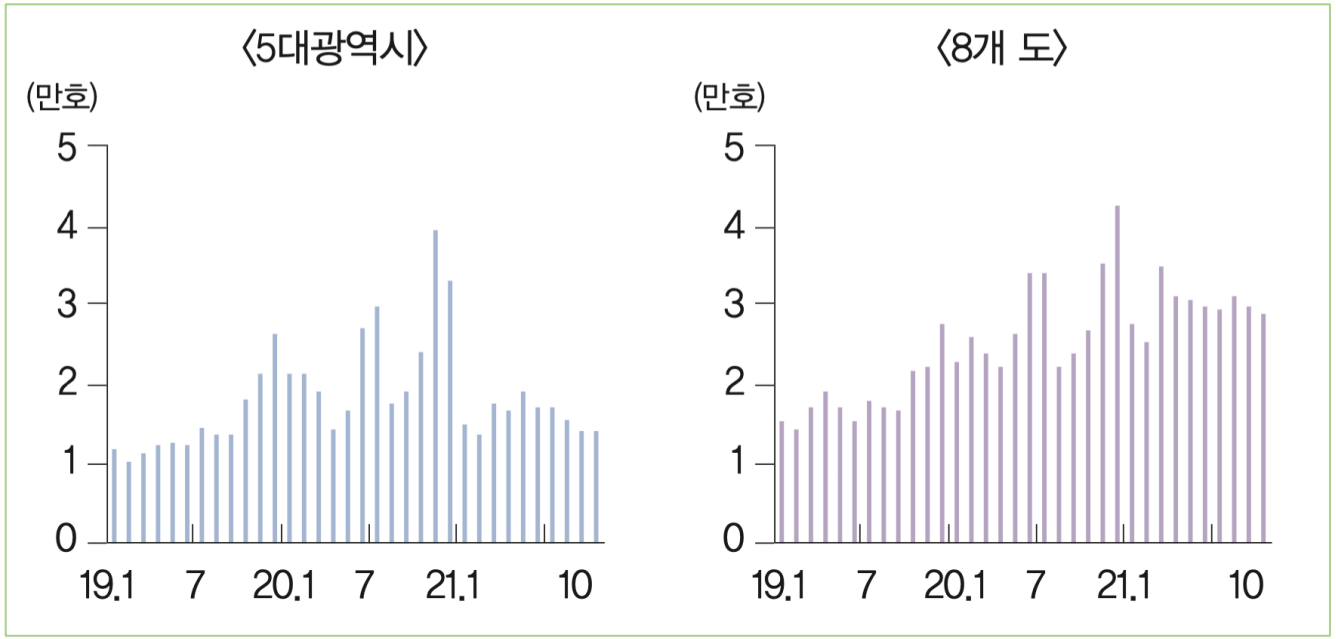


아황산 가스 추이 그래프

2. 시간 시각화

막대 그래프

- 연속형의 데이터가 아니라, 시간에 따라 명확하게 구분되는 데이터 포인트를 사용할 때 적합한 그래프
- 데이터 포인트가 일정하게 분포되어 있을 때 주로 사용

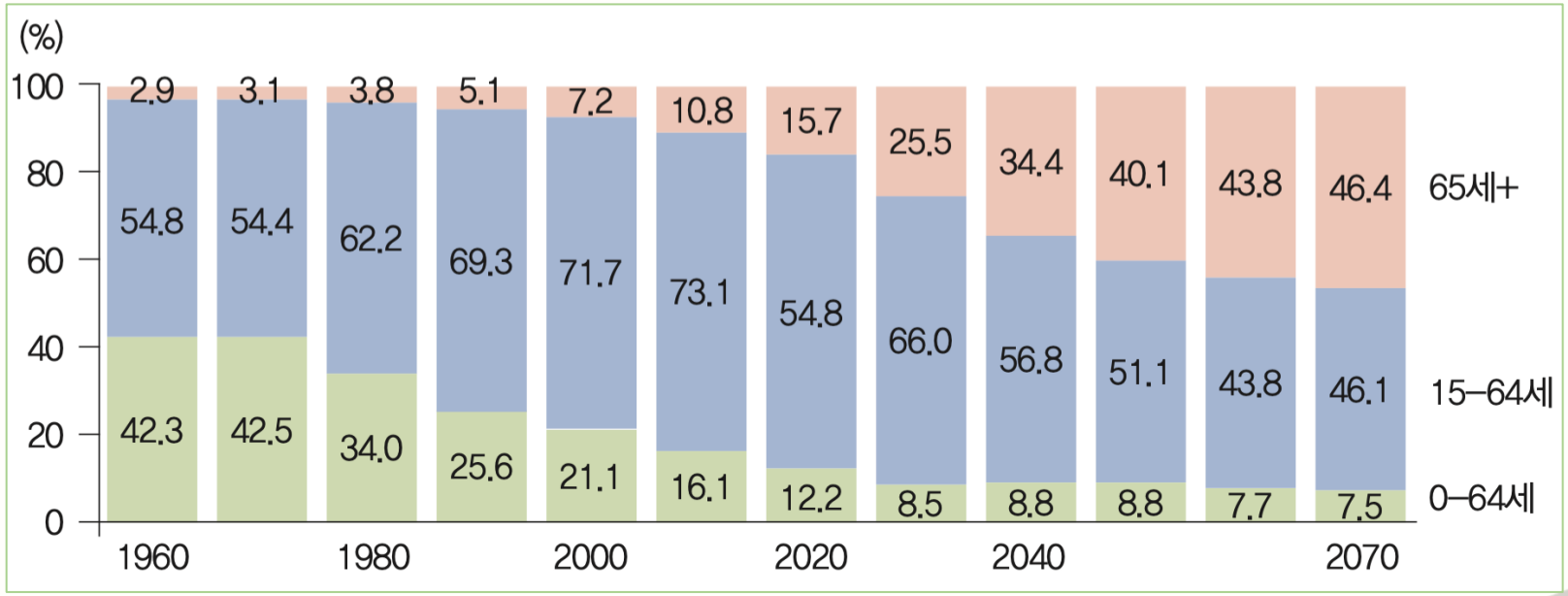


지역별 주택거래량

2. 시간 시각화

누적 그래프

- 막대 그래프와 동일한 경우에 사용 하나 범주가 여러 개일 경우, 그리고 단순히 범주를 서로 비교하는 것이 아니라 모두를 합친 것이 의미를 가질 때 사용

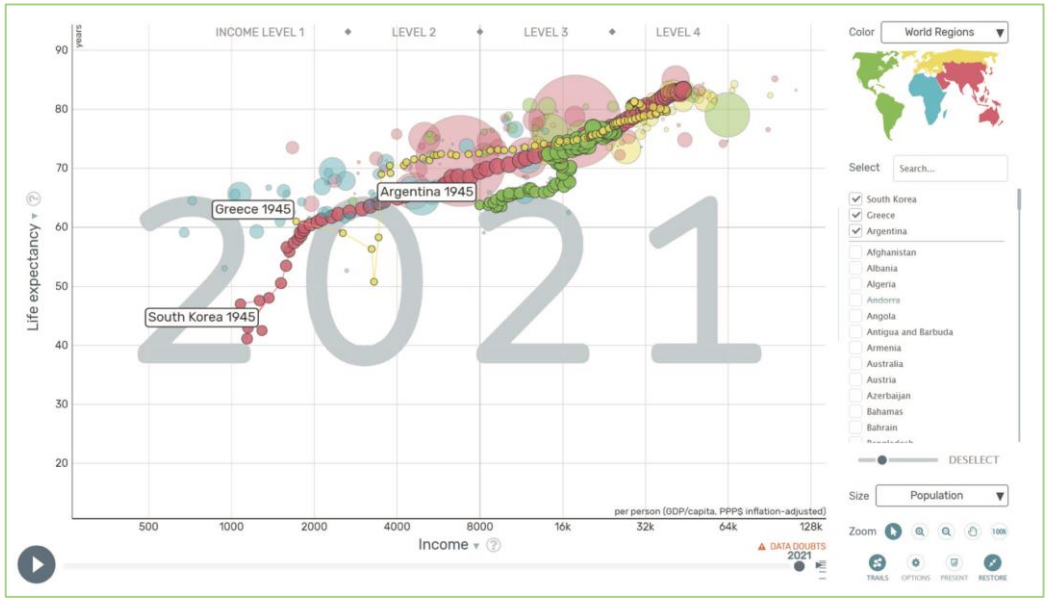


연령계층별 인구 구성비, 1960~2070년(중위)

2. 시간 시각화

버블 그래프

- 점 그래프가 단순히 데이터 포인트를 점으로 표현하는데 비해 버블(Bubble) 그래프는 원의 크기가 데이터의 어떠한 값을 표현
- 시간에 따라 변화하 는 두개의 값, 즉 버블의 x, y 위치, 크기 등을 모두 한번에 보여줄 수 있음

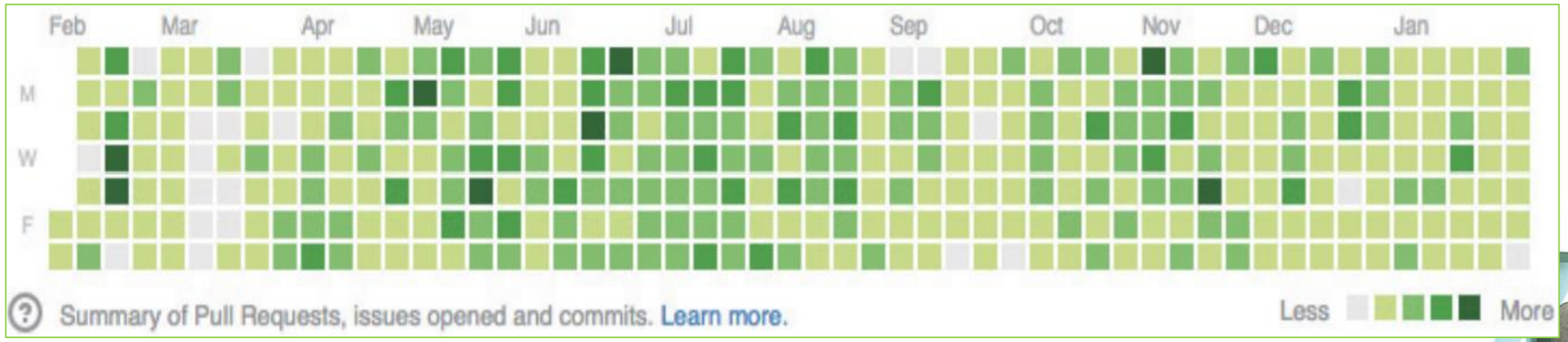


갭마인더 버블 그래프

2. 시간 시각화

칼라 스케일 그래프

- 막대 그래프의 높이 대신 시각적인 색으로 차이가 나게 보여주는 그래프
- 주로 x, y축을 활용한 패턴을 보여주어야 할 때 사용



github 사용자의 활동 그래프

빅데이터의
이해와 활용

3 텍스트 시각화

3. 텍스트 시각화

텍스트 시각화

- 문서의 핵심(gist)을 손쉽게 파악
- 많은 양의 문서를 분류해서 전체 문서를 조망
- 문서의 내용이나 문서의 집합이 어떻게 서로 다른지 비교
- 시간에 따라 텍스트가 어떻게 변해왔는지 분석
- 텍스트에 나타나는 패턴 파악

3. 텍스트 시각화

워드 클라우드(Word Cloud)

- 문서에 등장한 단어의 빈도수를 이용하여 어떠한 단어를 많이 사용했는지 시각화
- 무수히 많은 문서에서 중요한 단어를 추출하는데 유용

스티브 잡스의 스탠포드 대학교 졸업식 축사의 워드 클라우드

3. 텍스트 시각화

- 워드 트리(Word Tree)
  - 특정한 단어가 다른 단어들과 어떠한 구조로 연결되어 있는지를 시각화
  - 한글 구조에 적용이 쉽지 않음

오바마 대통령의 2009년 연설문을 워드트리로 분석한 예

빅데이터의
이해와 활용

4 소셜 네트워크 시각화

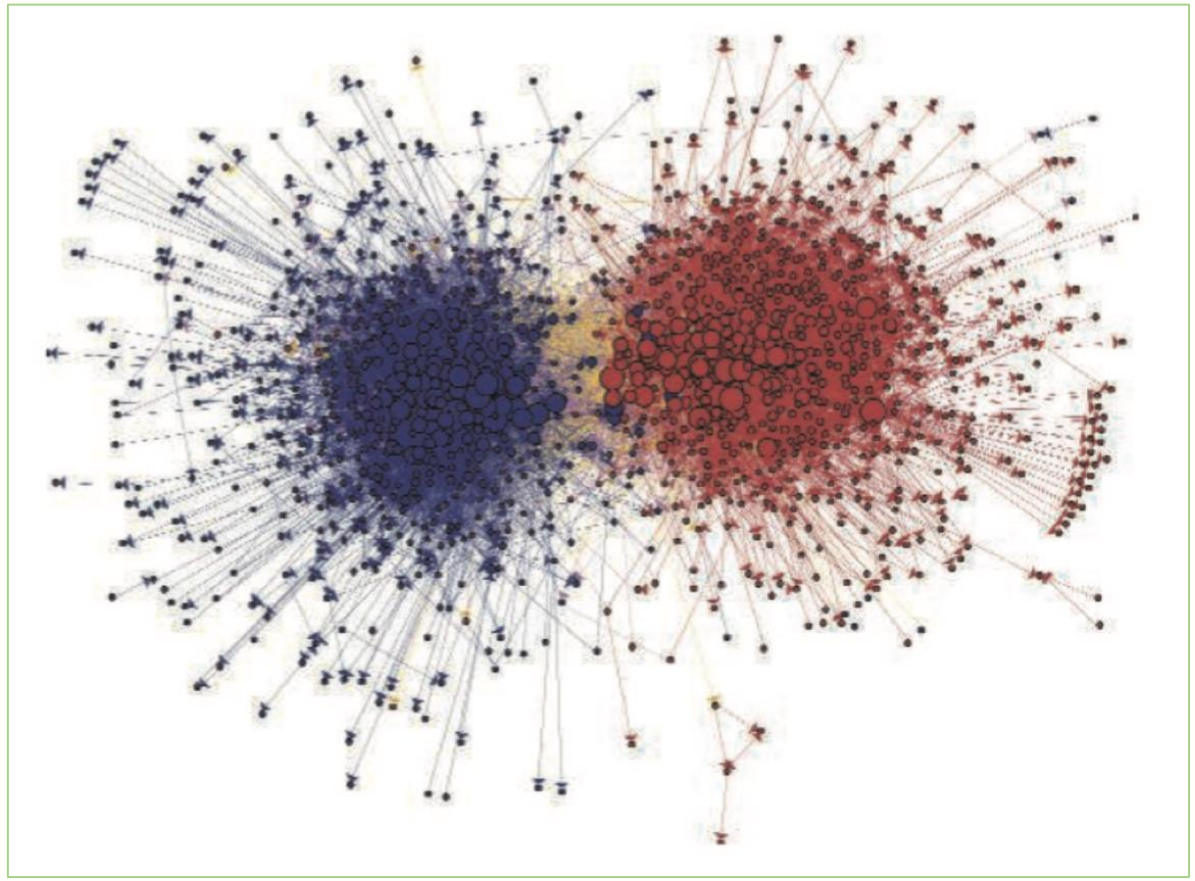
4. 소셜 네트워크 시각화

소셜 네트워크 분석

- 소셜 네트워크를 통한 정보 공유 활동이 활발해지면서 이 안에서 이루어지는 여러 사회현상을 탐색적으로 분석하고자 하는 노력이 시도됨
  - 타인과 어떤 관계를 맺는가
  - 정보는 어떻게 확산되는가
  - 누가 네트워크의 중심인가

4. 소셜 네트워크 시각화

소셜 네트워크 시각화

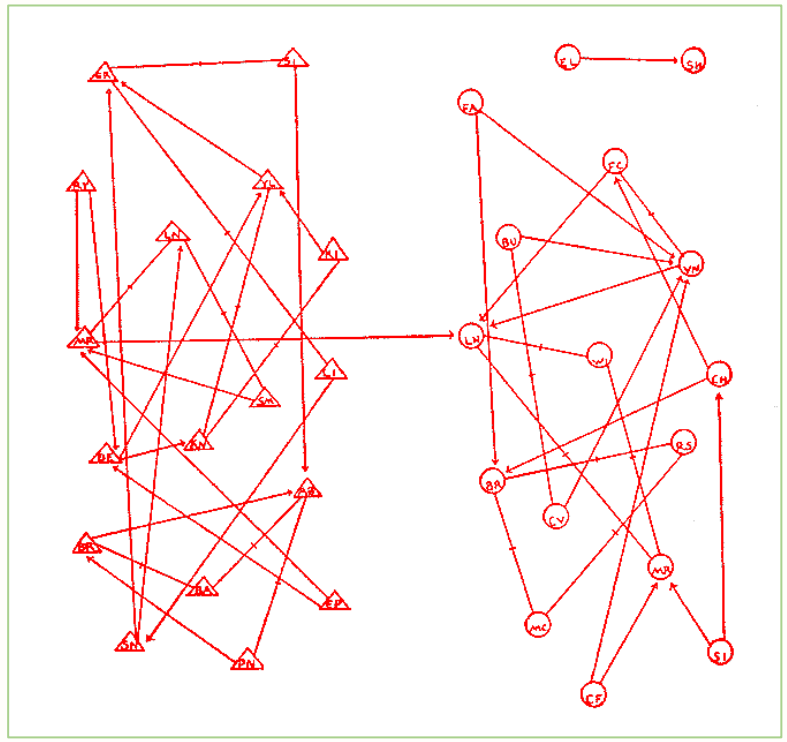


미국 정치 블로거의 소셜 네트워크 연결망 구조 시각화

4. 소셜 네트워크 시각화

소셜 네트워크 시각화

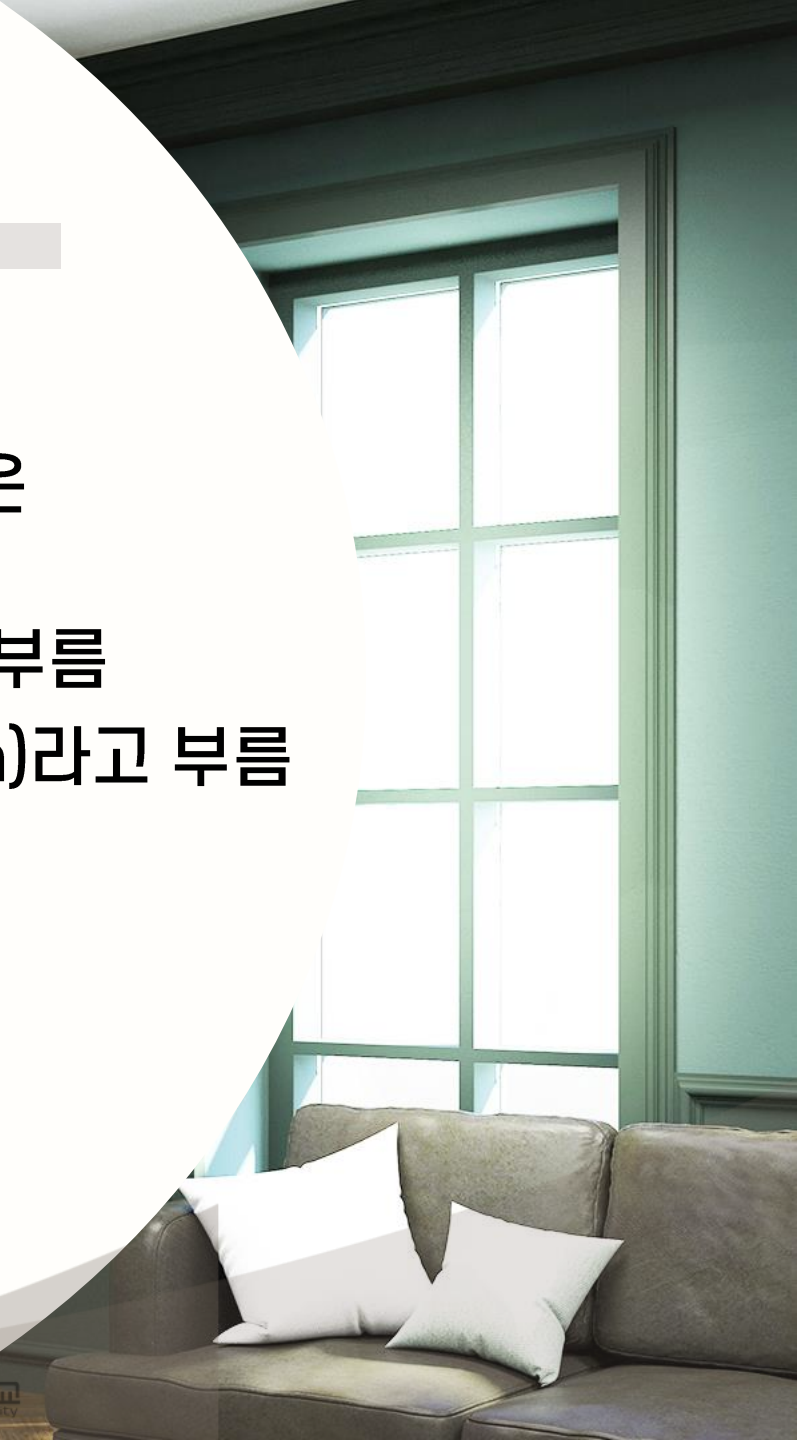
- 야코프 모레노(Jacob Moreno)가 원형을 제시



초등학교 4학년 학생들의 친구관계 시각화

4. 소셜 네트워크 시각화

- 그래프(Graph)
  - 소셜네트워크 시각화에서 행위자들은 버티스(vertice) 혹은 노드(node)로 표현됨
  - 두 노드를 연결하는 선을 엣지(edge) 혹은 링크(link)라고 부름
  - 노드와 엣지의 관계를 그림으로 표현한 것을 그래프(graph)라고 부름



4. 소셜 네트워크 시각화

- 인접테이블을 이용한 소셜네트워크 그래프

	A	B	C
A	0	1	1
B	1	0	0
C	1	0	0

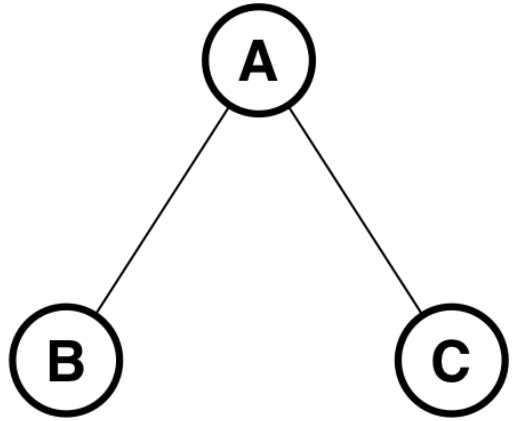
인접테이블

A : B, C

B : A

C : A

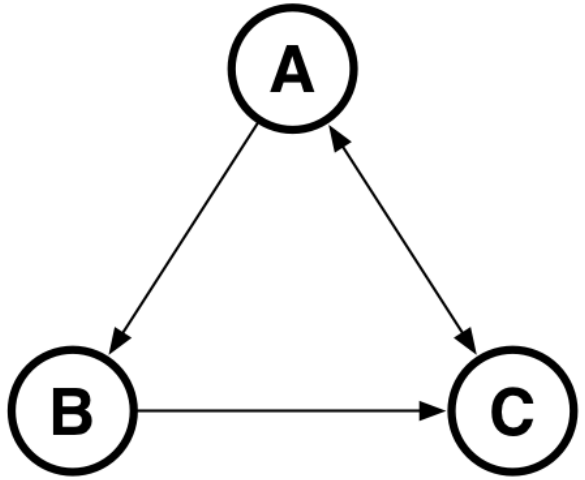
인접 리스트



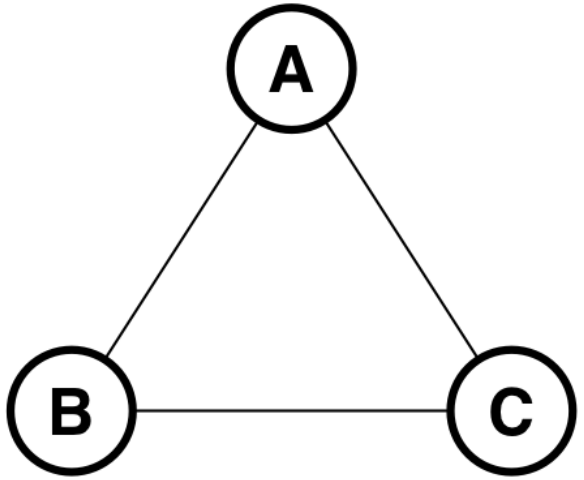
그래프 네트워크

4. 소셜 네트워크 시각화

- 디렉티드와 언디렉티드 그래프



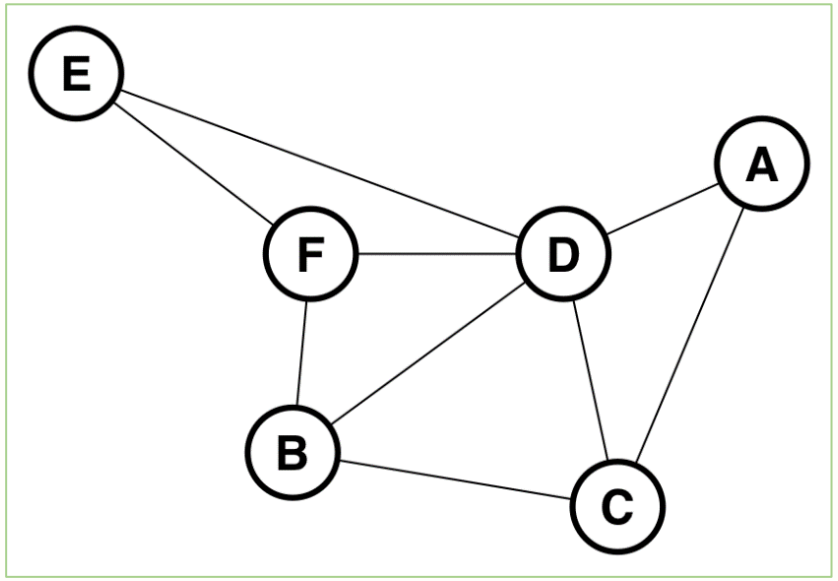
디렉티드(directed)



언디렉티드(undirected)

4. 소셜 네트워크 시각화

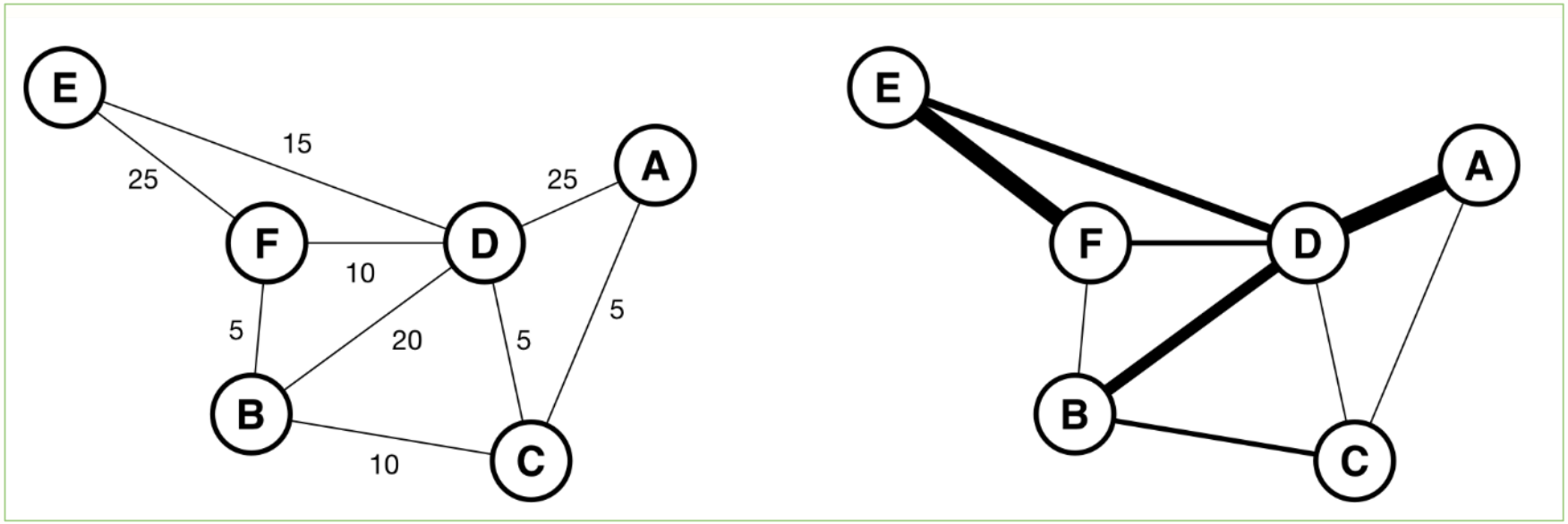
엣지의 디그리(Degree of Edge)



- 노드 D는 5개의 디그리를 가짐 (네트워크의 중심으로 해석)
- 디렉티드 그래프:
  인디그리(in-degree)와 아웃디그리(out-degree)로 구분

4. 소셜 네트워크 시각화

- 가중 그래프(Weighted Graph)



엣지에 값을 표현

엣지 굵기나 색으로 값을 표현

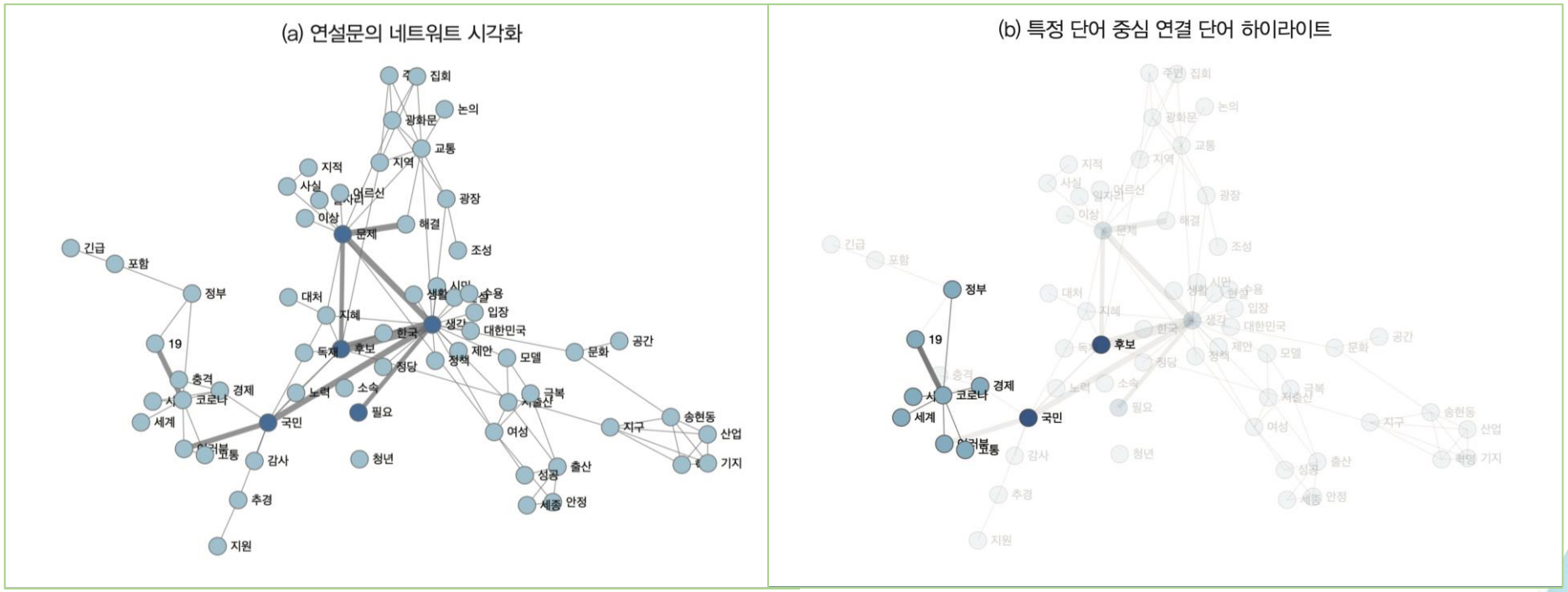
4. 소셜 네트워크 시각화

- 소셜 네트워크 시각화의 활용: 정치인의 메시지
  - 주요 메시지 파악
  - 주요 메시지와 연결된 핵심 개념 파악

  -> 단어의 공출현(co-occurrence) 네트워크 시각화

4. 소셜 네트워크 시각화

- 소셜 네트워크 시각화의 활용: 정치인의 메시지



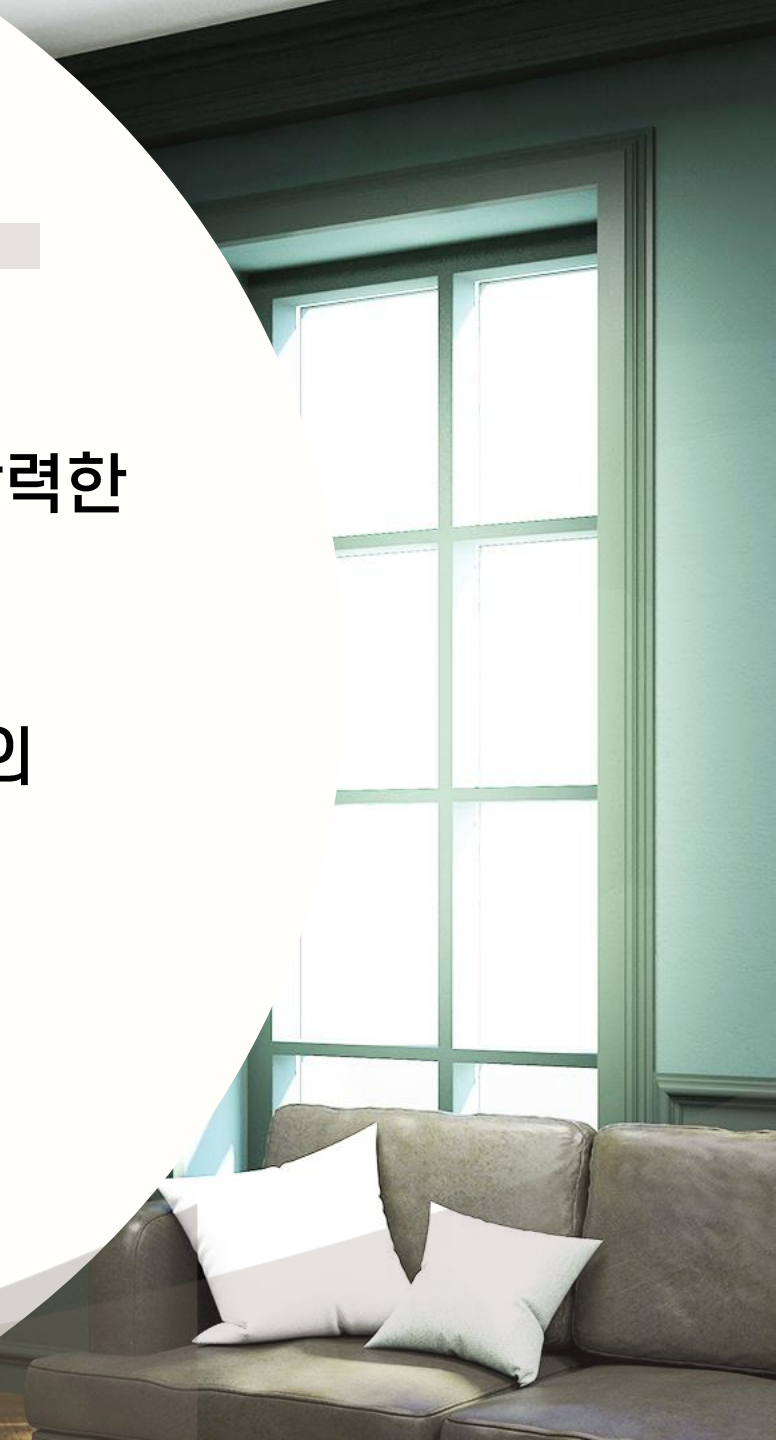
정치인의 발언을 중심으로 본 소셜 네트워크 시각화 예제

빅데이터의
이해와 활용

5 데이터 시각화의 도구

5. 데이터 시각화의 도구

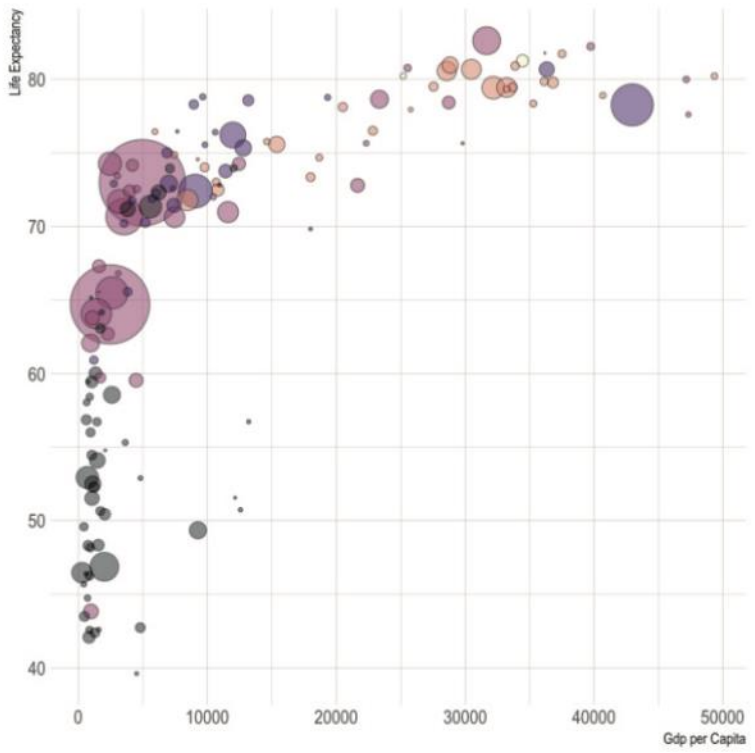
- R과 Python
  - 통계처리와 빅데이터 분석에 주로 쓰이는 R과 Python은 강력한 시각화 도구를 제공
  - R은 ggplot2라는 라이브러리를 통해 다양한 시각화 가능
  - Python의 경우 Pandas, Matplotlib, Seaborn, Plotly 등의 시각화 라이브러리를 제공
  - R과 Python이 제공하는 시각화도구로는 인터랙티브한 시각화 결과물을 만들 수 없다는 한계가 있음
  - 인터랙티브 시각화를 위해서는 JavaScript기반의 시각화 도구를 주로 사용



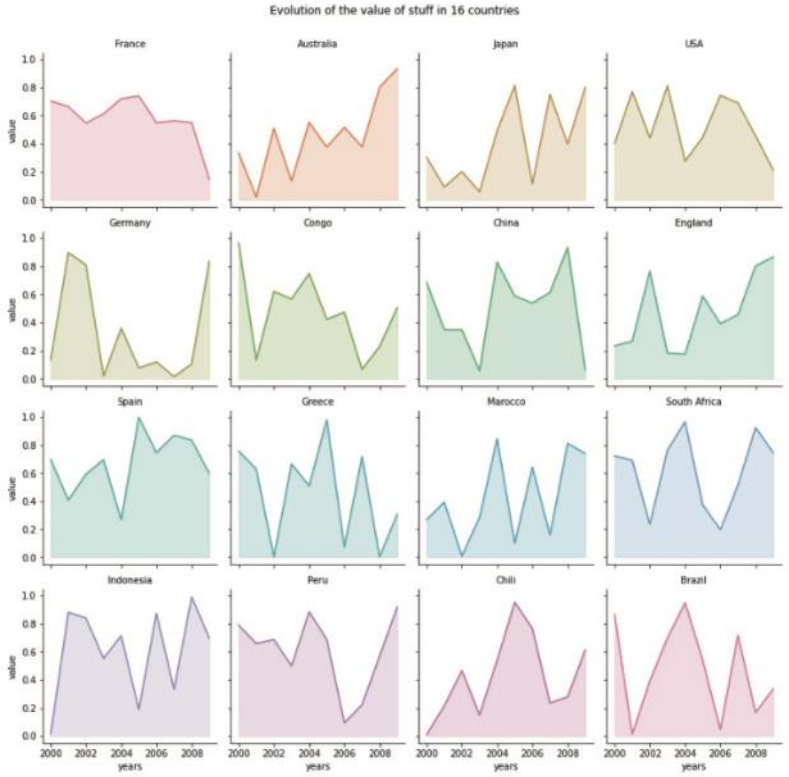
5. 데이터 시각화의 도구

- R과 Python

(a) R ggplot2를 이용한 버블 차트



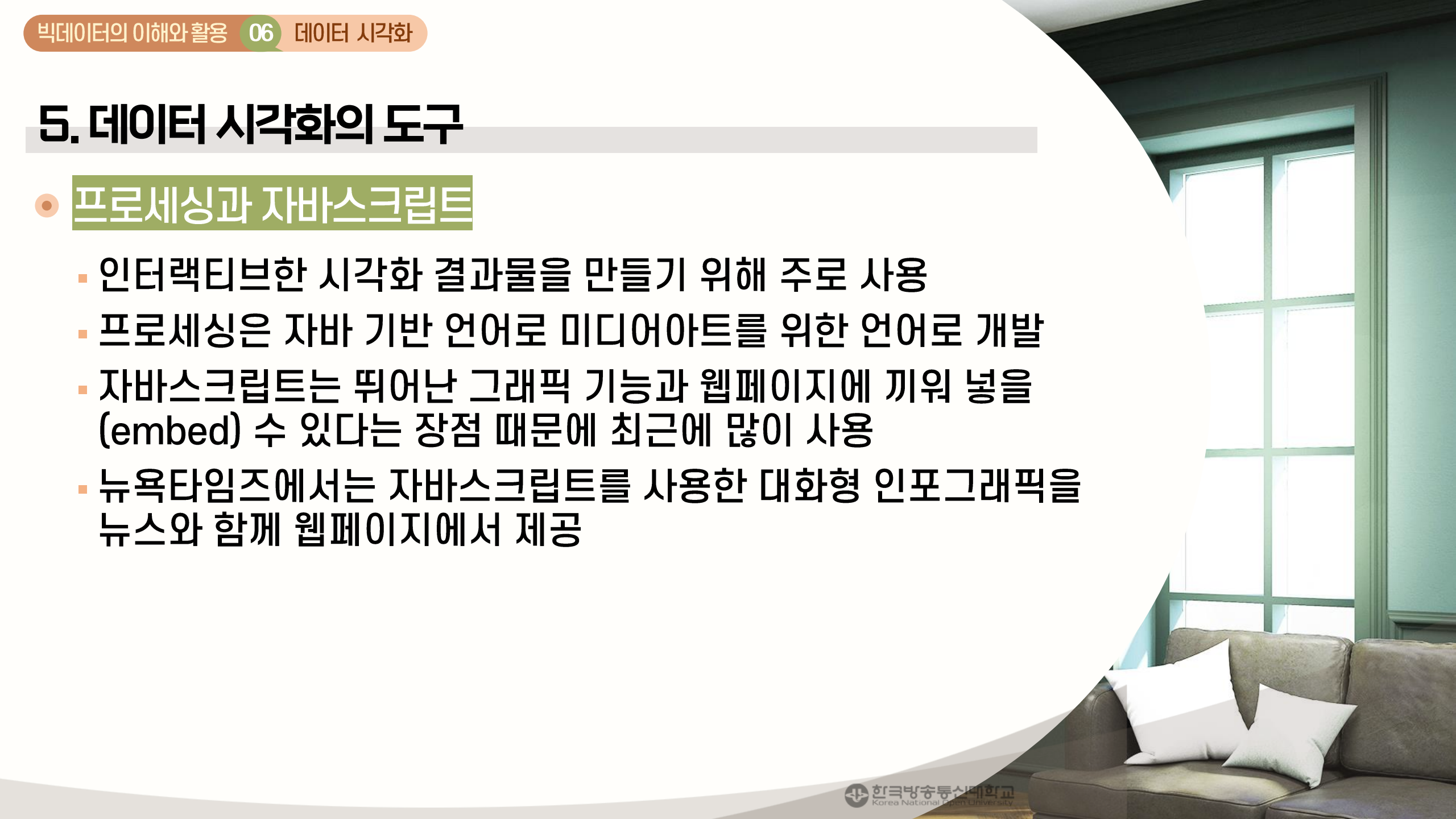
(b) matplotlib를 이용한 선그래프



5. 데이터 시각화의 도구

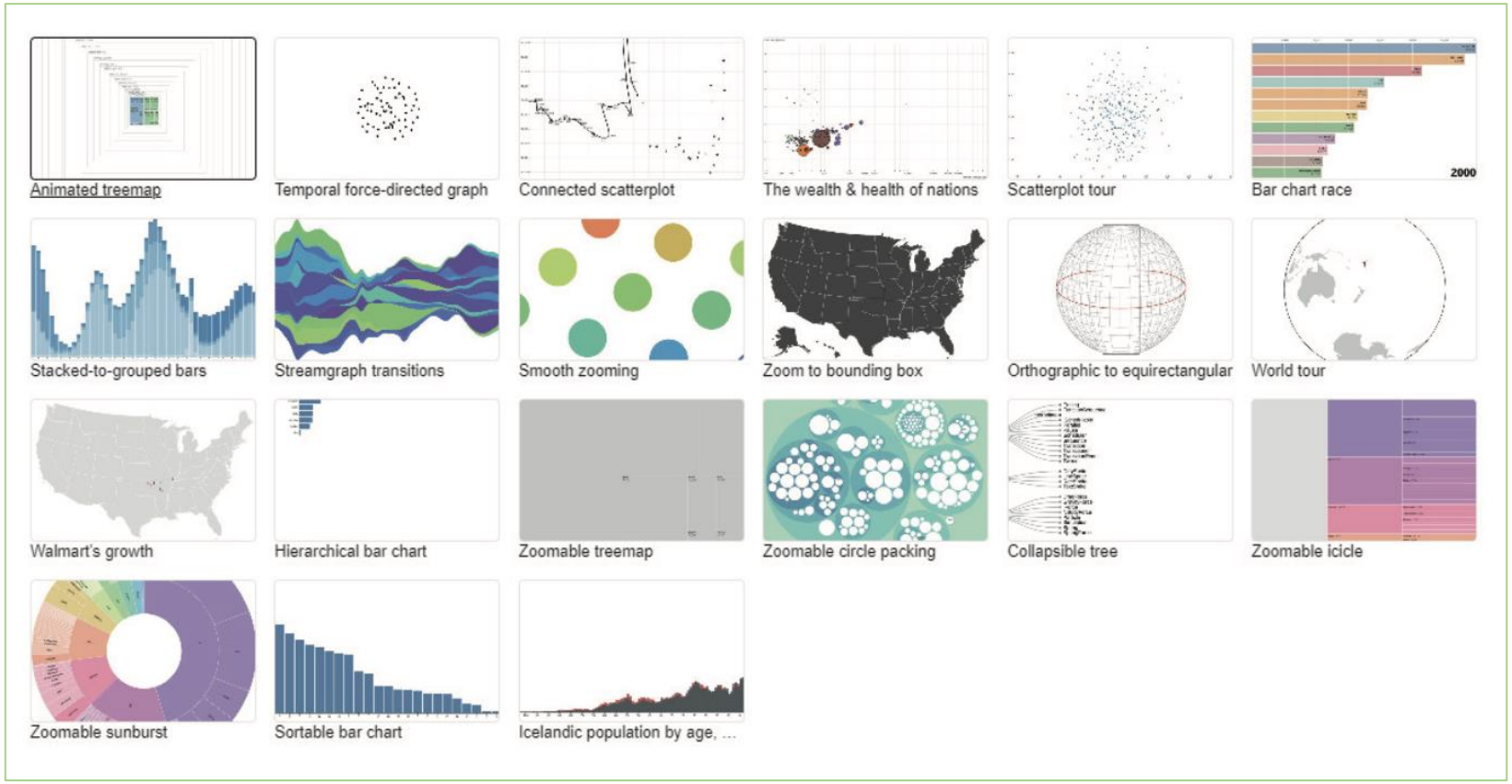
프로세싱과 자바스크립트

- 인터랙티브한 시각화 결과물을 만들기 위해 주로 사용
- 프로세싱은 자바 기반 언어로 미디어아트를 위한 언어로 개발
- 자바스크립트는 뛰어난 그래픽 기능과 웹페이지에 끼워 넣을 (embed) 수 있다는 장점 때문에 최근에 많이 사용
- 뉴욕타임즈에서는 자바스크립트를 사용한 대화형 인포그래픽을 뉴스와 함께 웹페이지에서 제공



5. 데이터 시각화의 도구

- 프로세싱과 자바스크립트



D3.js로 만든 다양한 시각화

5. 데이터 시각화의 도구

- 태블로(Tableau)
  - 스탠포드 대학의 정보시각화그룹(Information Visualization Group)에서 개발
  - 쿼리, 분석 툴을 가지고 있으며 손쉽게 다양한 시각화구현 가능
  - 사용자가 프로그래밍 없이 데이터를 손쉽게 조작하여 분석할 수 있는 인터페이스를 가지고 있는 인터렉티브 시각화 도구
  - 최근 빅데이터 분석 시각화 도구로 많이 사용되고 있음

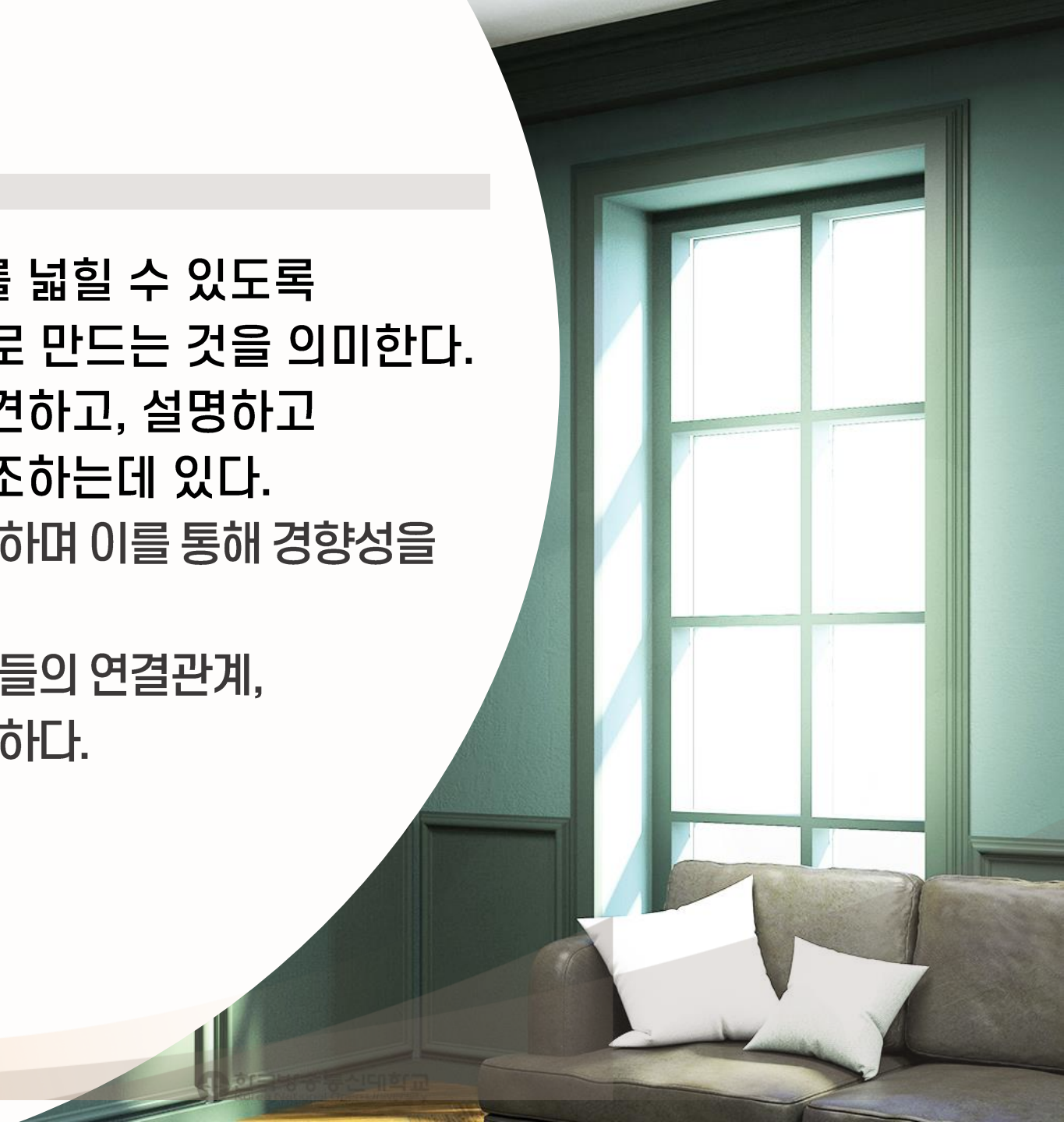
5. 데이터 시각화의 도구

- **Power BI**
  - 마이크로소프트(Microsoft)가 개발한 대화형 데이터 시각화 소프트웨어
  - 마이크로소프트가 제공하는 빅데이터 솔류션인 Azure와 연결하여 빅데이터 시각화에 활용되고 있음

Power BI를 이용한 데이터 시각화

정리하기

- 데이터 시각화는 컴퓨터를 사용하여 인지를 넓힐 수 있도록 데이터를 상호작용이 가능한 시각적 형태로 만드는 것을 의미한다.
- 시각화의 목적은 데이터의 숨은 의미를 발견하고, 설명하고 그걸 통해 의사결정을 내리는 통찰력을 보조하는데 있다.
- 시간 시각화는 시계열 데이터의 시각화를 의미하며 이를 통해 경향성을 파악할 수 있다.
- 소셜 네트워크 시각화는 네트워크 상에서 사람들의 연결관계, 정보의 흐름 등을 탐색적으로 파악하는데 유용하다.



07 강

다음시간 안내

추천시스템

수고하셨습니다!