삼육대학교 SAHMYOOK UNIVERSITY

SU GLORY 사람을 참되게 / 세상을 환하게

[2022년 고교교육 기여대학 지원사업]

[빅데이터분석 특강]

삼육대학교 - 청원고등학교

Intro

- 주제 : 공공데이터 활용 빅데이터 분석 with 파이썬
  - 내가 관심있는 또는 해결하고 싶은 **문제를 정의**하고,
  - 이를 데이터 **기반**으로 **문제를 해결**한다.

- 일정
  - 1일차 : 데이터 분석 소개 / 실습 환경 준비 / 파이썬 실습
  - 2일차 : 데이터 분석 실습 (1) – 데이터 수집, 탐색적 데이터 분석
  - 3일차 : 데이터 분석 실습 (2) – 분석 모델 선정, 모델링, 문제해결
  - 4일차 : 프로젝트 설계
  - 5일차 : 프로젝트 발표

Airbnb 사례

- 2007년, 브라이언 체스키, 조 게비아의 아이디어
  - 샌프란시스코에 있는 자신의 집에서 여유공간을 여행객에서 유료로 제공
  - 여행객을 모집할 웹사이트 제작 –〉 에어비엔비의 시작 (2008년)
- 2010년, 뉴욕의 숙소 예약률이 매우 낮다는 사실 발견
  - 뉴욕 숙소 주인들과 인터뷰 –〉 광고에 사용된 사진의 질이 매우 낮음 –〉 고퀄리티 숙소 사진이 예약에 영향을 미칠 것이라는 가설 설정 –〉 전문 사진사를 고용하여 일부 숙소에서 실험 –〉 실험 결과로 뉴욕 매출이 두배로 증가 –〉 모든 숙소 주인에게 전문 사진사 무료 제공
- 2012년, 유럽에서 호스트 모집을 위한 온라인 광고 효과가 저조하다는 사실 발견
  - 온라인 마케팅 보다 오프라인 마케팅을 했을 때 참여 효과가 5배 증가 –〉 시장별로 다른 마케팅 전략 채택 후 결과 비교(A/B test)

데이터 기반 문제해결 사례

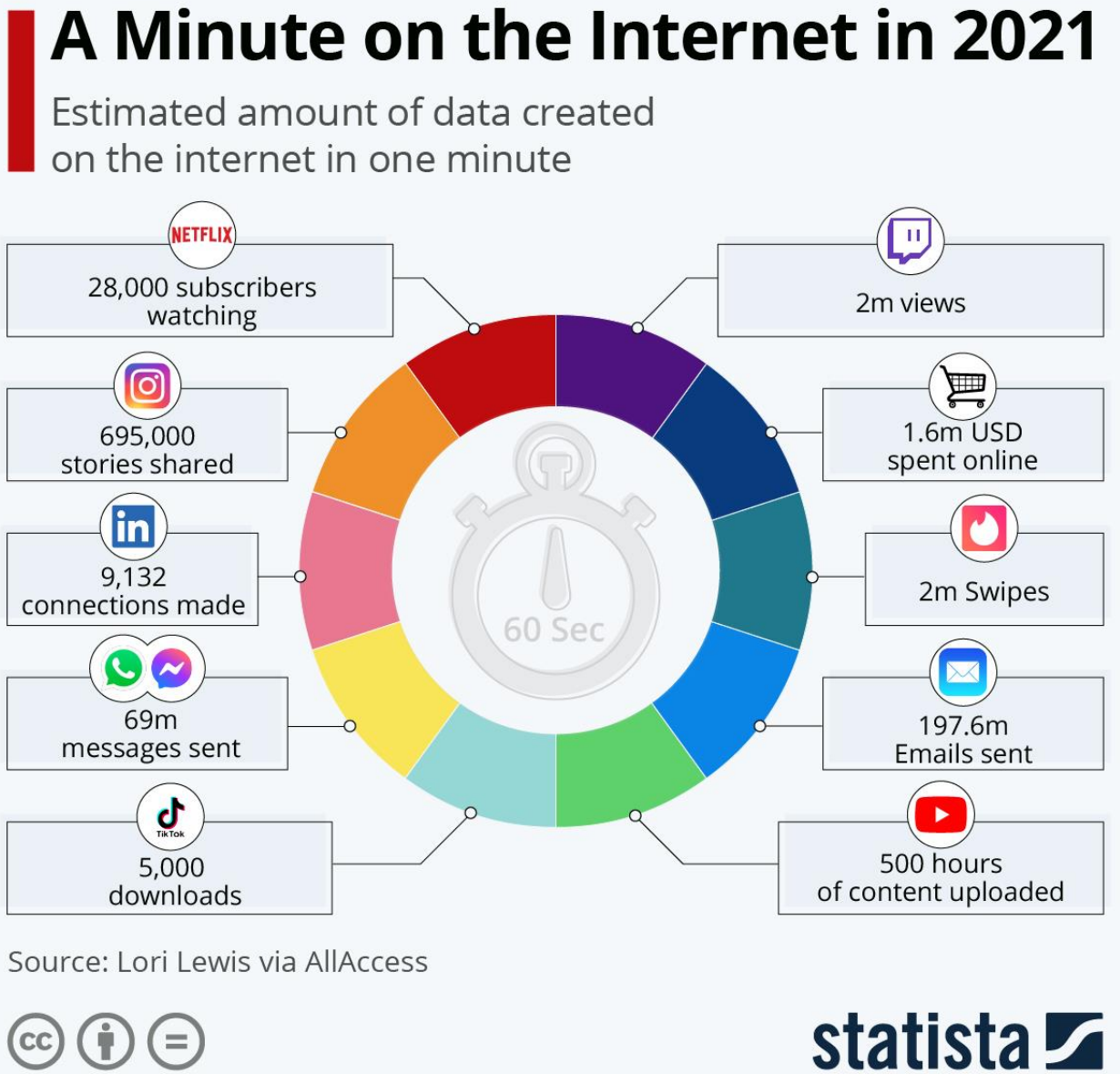
- (2017-08-04) 데이터 기반의 문제해결, 정부와 기업이 데이터를 활용하는 방법 [링크]

- (2021-10-09) 데이터 기반 글쓰기, 선정되면 100만원! 쏘프라이즈, [링크]

- 시빅해킹, 코드포코리아 [링크], 널채움 [링크]

데이터

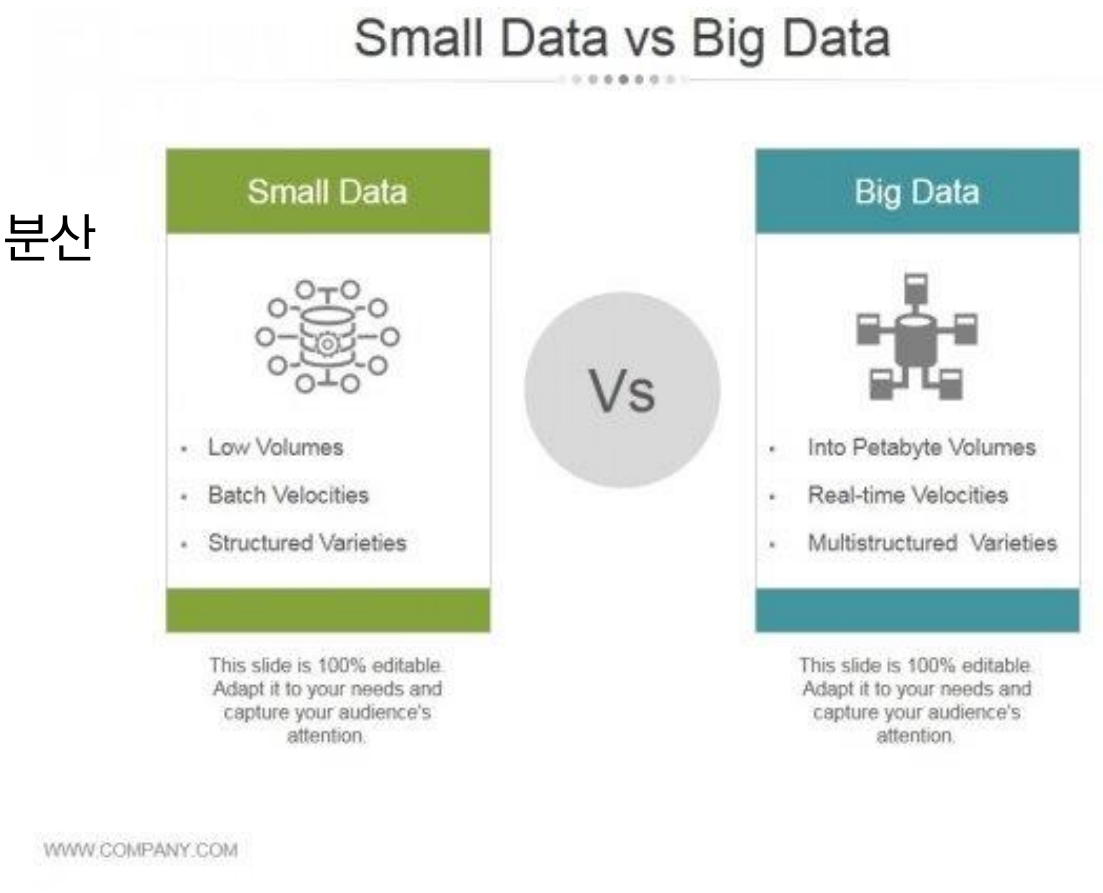
- 데이터 홍수의 시대
  - 수 많은 데이터가 생산되고 수집
    - ex) 웹사이트 내에서 사용자의 클릭 추적
  - 일 분 동안 온라인에서 생성되는 데이터수
- 과거에도 수 많은 데이터가 존재했지만,
  - 인식하고, 수집하기가 어려웠음
- 컴퓨팅 기술의 발달은,
  - 데이터를 수집하고 처리하고, 분석하는 과정의 자동화가 진행됨
  - 새로운 종류의 데이터가 생기고, 그 수가 기하급수적으로 증가함 -> 빅 데이터!



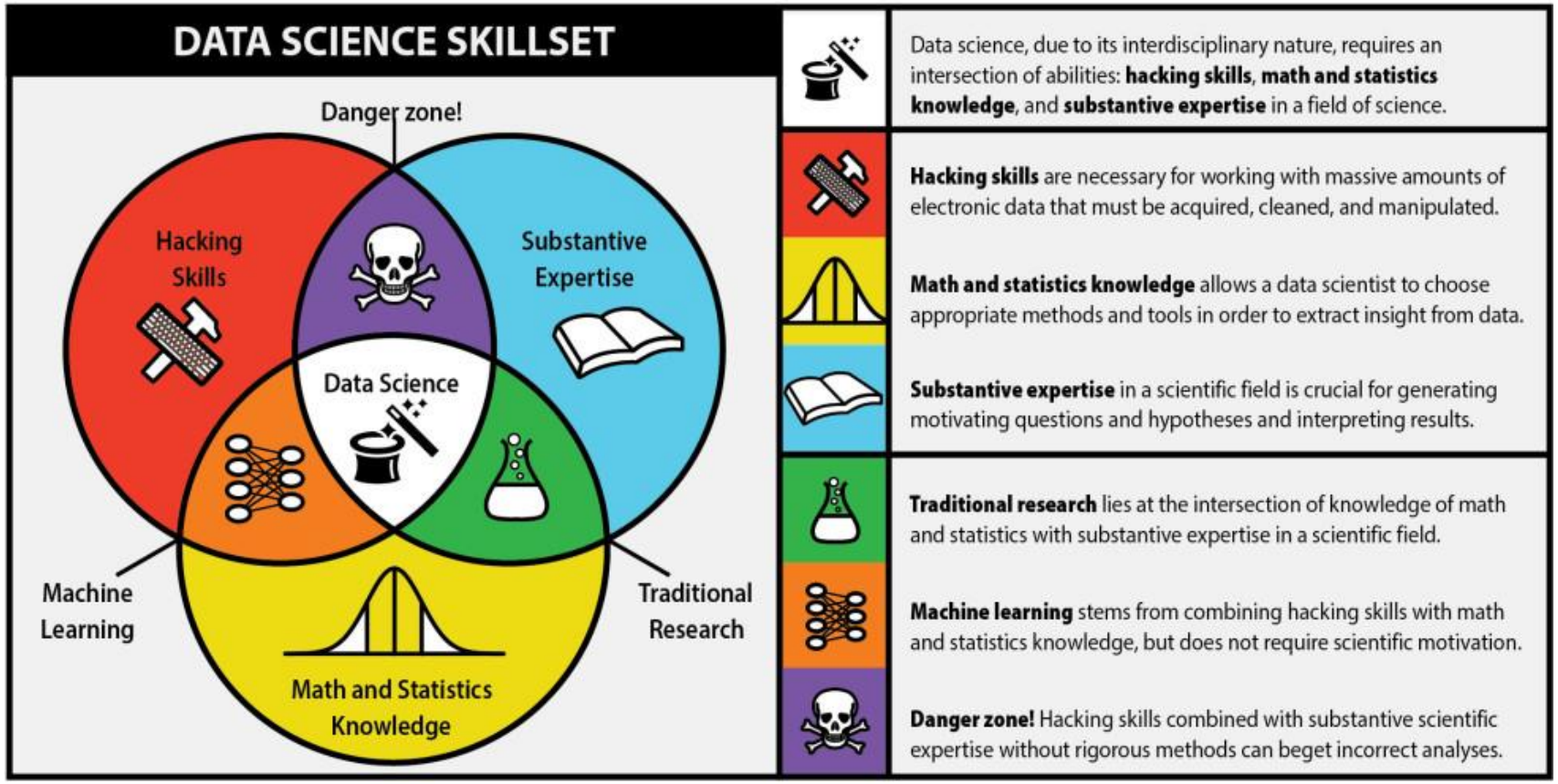
source: https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute/

데이터 vs 빅데이터

- 데이터와 빅데이터는 어떤 차이가 있을까?
  - 데이터의 크기 : GB, TB vs PB, EB
  - 데이터 구성 방법 : 정형 데이터 vs 비정형 데이터
  - 데이터 관리를 위한 아키텍처 : 중앙 집중 시스템 vs 분산 시스템

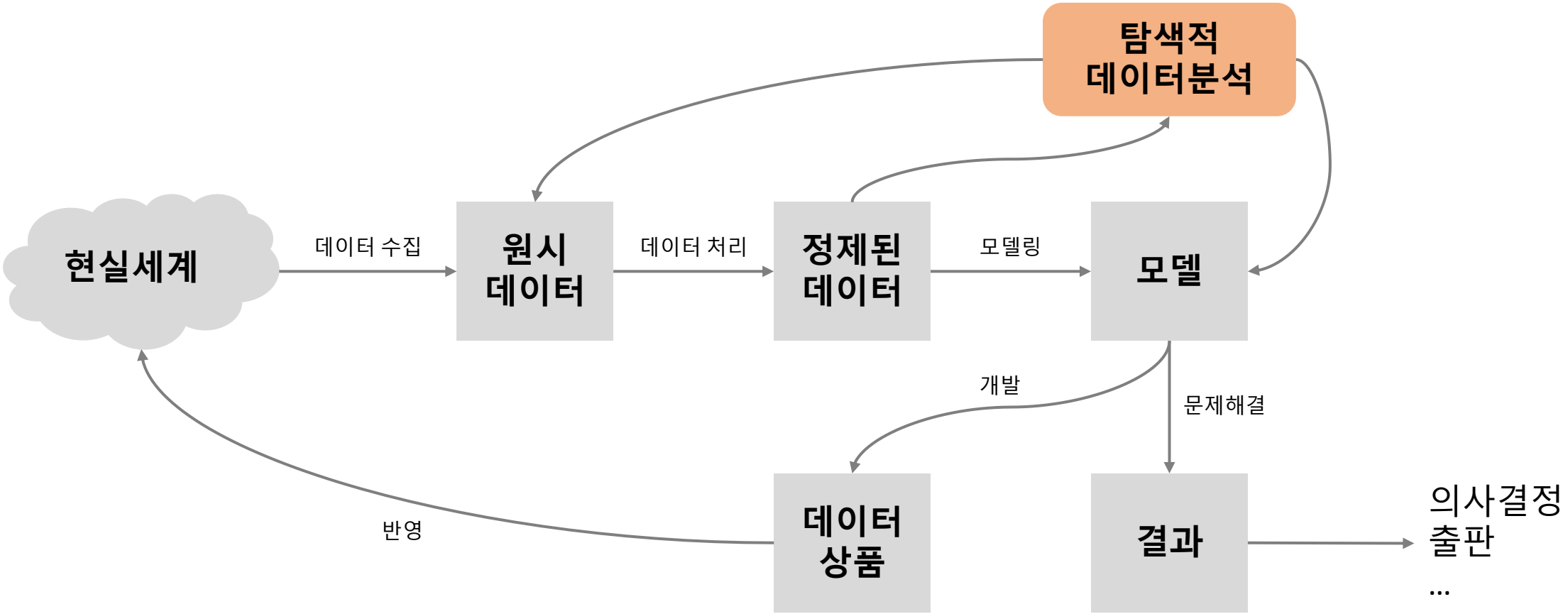


데이터 과학자의 역량

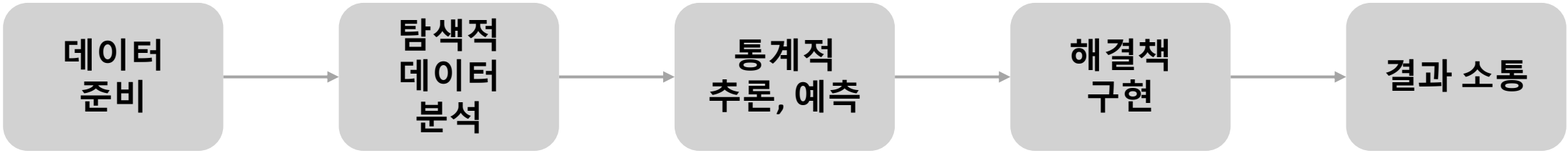


source: http://berkeleysciencereview.com/article/first-rule-data-science/

데이터 분석 단계


탐색적
데이터분석
현실세계
데이터 수집
원시
데이터
데이터 처리
정제된
데이터
모델링
모델
개발
문제해결
반영
데이터
상품
결과
의사결정
출판
...

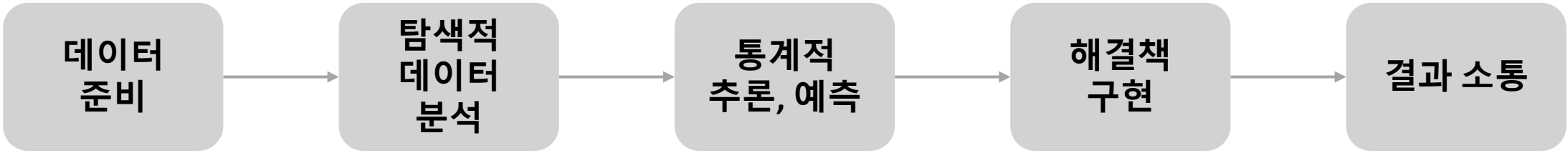
데이터 분석 단계



- 데이터 준비
  - 직접 수집
  - 간접 수집

- 탐색적 데이터 분석
  - 데이터 값 확인 -> 이상치 확인, 새로운 패턴 발견 / 집계, 시각화 기능 필요
  - 통계적 분석

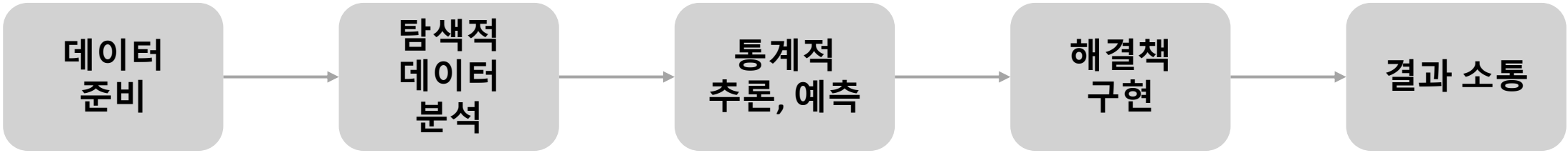
데이터 분석 단계



- 통계적 추론, 예측
  - 통계적인 결론, 예측 모델 만들기

- 모델 종류
  - 통계 모델
    - 기술통계 : 데이터가 가진 일반적인 특성(빈도수, 비율, 평균, 표준편차 등)
    - 추론통계 : 표본분석을 토대로 모집단의 특성을 추론(분산, 상관, 회귀 등)
  - 데이터 마이닝 모델 : 대규모 데이터에 숨어있는 패턴 발견 및 규칙 발견(군집, 연관, 분류, 예측 등)
  - 텍스트 마이닝 모델 : 텍스트 기반 데이터에서 정보 검색, 추출, 체계화, 분석을 포함하는 방법
  - 소셜 네트워크 분석 모델 : 언어 분석 기반의 정보 추출을 통해 대용량의 소셜 미디어 데이터에서 이슈를 탐지하고, 이슈가 유통되는 전 과정을 분석하는 방법

데이터 분석 단계



- 해결책 구현
  - 실제 제품화, 기존 시스템 개선

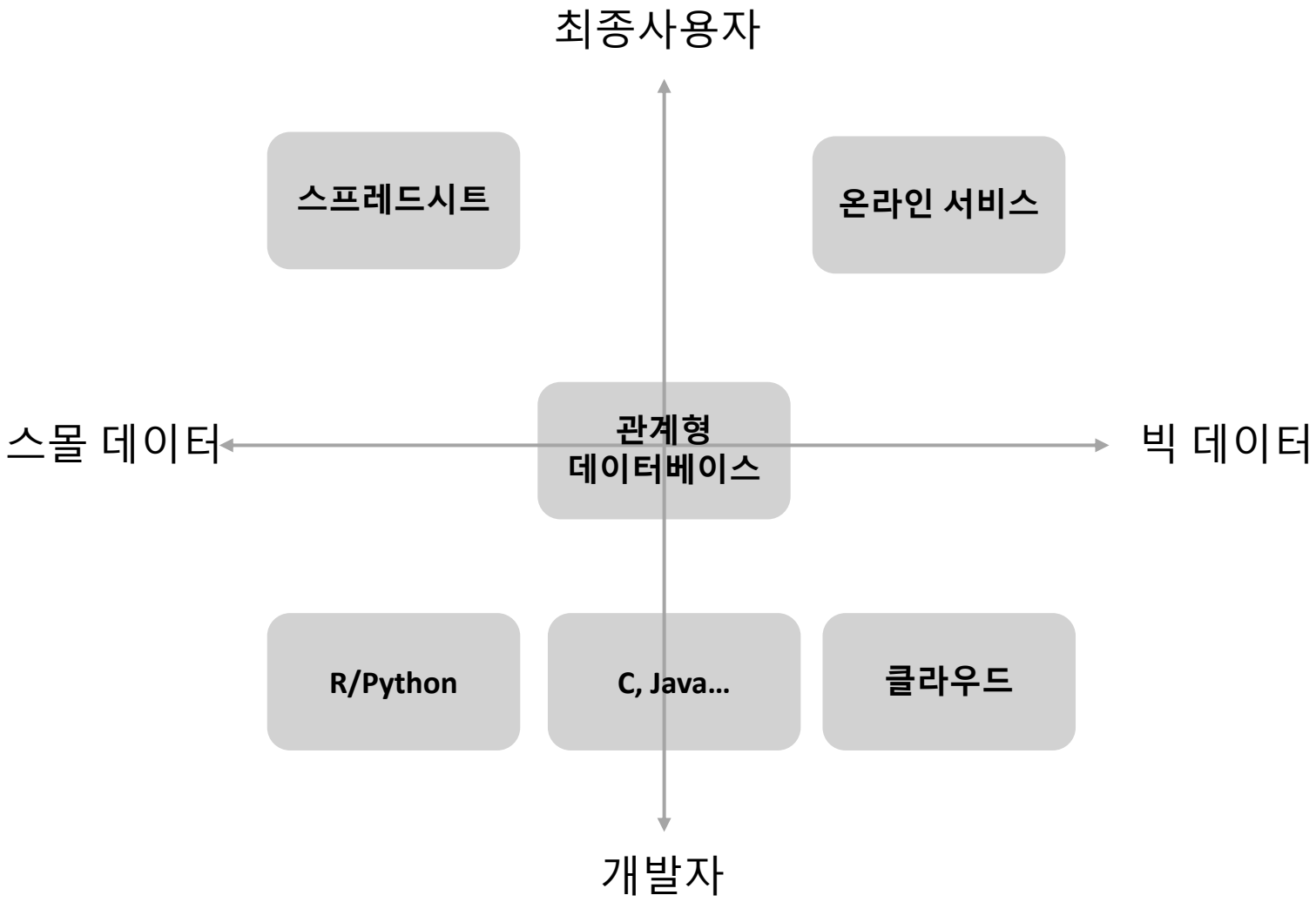
- 결과 소통
  - 프로젝트(문제해결) 결과 발표
  - 시각화, 대시보드 활용

어떤 도구를 사용할까?

	스프레드시트	관계형 데이터베이스	R or Python	Cloud	커스텀 코드
처리용량	메모리 용량에 제한	디스크 용량에 제한	메모리 용량에 제한	거의 무제한	구현 방식에 의해 결정
응답시간	실시간 ~ 수 분	환경설정에 따라 결정	실시간 ~ 수 분	수 분 ~ 수 시간	구현 방식에 의해 결정
지원데이터 형태	테이블	테이블	테이블 / 벡터 / 행렬	거의 무제한(key-value/테이블/비정형)	무제한
프로그래밍 지원	VBScript	내장 프로시저	R function / python	다양한 언어 지원	무제한
통계 및 기계학습 기능	제한적인 통계 / 학습 모델 지원	지원하지 않음	대부분의 통계 / 학습 모델 지원	대부분의 통계 / 학습 모델 지원	무제한
데이터 시각화 기능	제한적인 시각화 지원	지원하지 않음	다양한 시각화 지원	대부분 지원하지 않음	무제한

source: 김진영(2016). 헬로 데이터 과학. 서울: 한빛미디어.

어떤 도구를 사용할까?



source: 김진영(2016). 헬로 데이터 과학. 서울: 한빛미디어.

실습 환경 구축

아나콘다

파이썬 실습

구글 코랩 [링크]

데이터 수집

공공데이터 수집하기

공공데이터 open data

- 공공데이터란,
  - 공공기관에서 업무를 수행하면 만들었거나 관리하고 있는 다양한 형태의 데이터

- 공공데이터 제공 사이트
  - 공공데이터 포털 : https://www.data.go.kr/
  - 국가통계포털 : https://kosis.kr/index/index.do
  - 기상 자료 개방 포털 : https://data.kma.go.kr/cmmn/main.do
  - 국가 공간정보 포털 : http://www.nsdi.go.kr/lxportal/?menuno=2679
  - 문화 공공데이터 : https://www.culture.go.kr/data/main/main.do
  - ...

샘플 문제

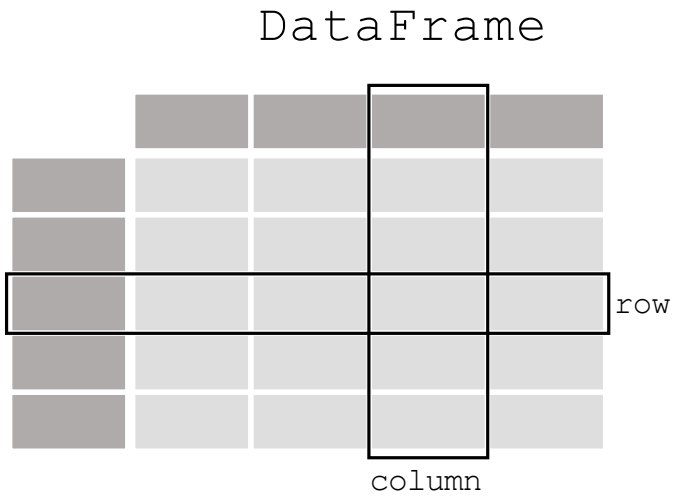
- 코로나 사태 처럼 재난 및 감영병 발생시 우리나라 국민은 누구나 충분한 의료 서비스를 받을 수 있을까?
  - 김정회, 이정면, 이용갑(2020). 공공의료 확충의 필요성과 전략, 건강보험연구원 이슈리포트. [링크]
  - "코로나19, 공공병원 확충 필요성 재확인"...핵심은 지역완결성? [링크]
- 우리나라의 어느 지역에 공공보건의료기관을 확충해야 할까?
- 우리나라의 행정구역별로 인구수 대비 공공보건의료기관의 비율은 얼마인가?

데이터 수집

- 필요한 데이터는 무엇인가?
  - 행정구역별 인구 수 -〉 국가통계포털 사이트
  - 행정구역별 공공보건의료기관 수 -〉 공공데이터포털 사이트
  - 행정구역별로 인구 수 대비 공공보건의료기관 비율 = 기관 수 / 인구 수

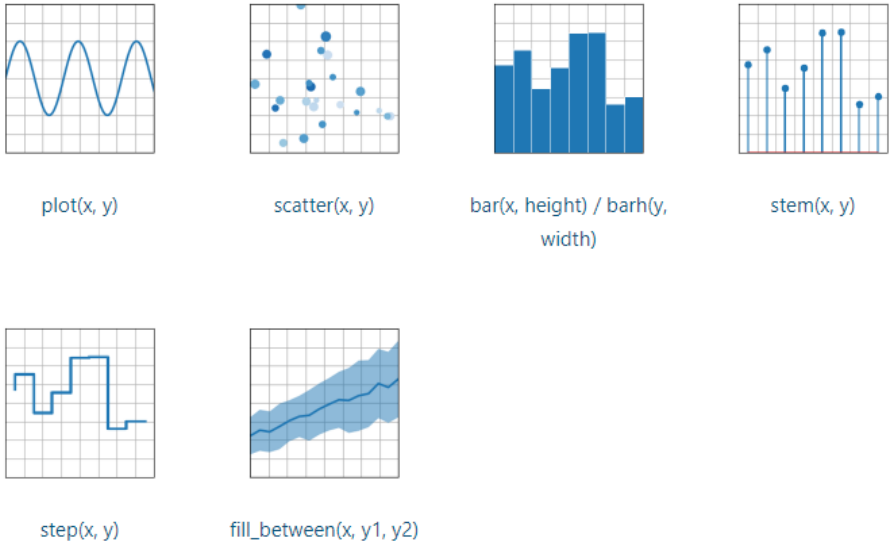
Pandas 기초 실습

- 판다스 Pandas
  - https://pandas.pydata.org/
  - 테이블 형태를 다룰 수 있는 파이썬 라이브러리
  - 1차원 구조인 Series, 2차원 구조인 DataFram, 3차원 구조인 Panel 을 지원



matplotlib 기초 실습

- 맷플롯립 matplotlib
  - https://matplotlib.org/
  - 데이터를 차트나 플롯으로 시각화하는 파이썬 패키지



추가 문제

- 인구수 대비 비율을 계산할 수 있는 문제 만들기

- 예를 들면, 서울시 자치구별 인원 대비 백신 접종 비율을 파악해서, 코로나 예방 전략 세우기
  - 서울 열린데이터 광장 〉 서울시 자치구별 백신 접종자수 현황