동국대학교
dongguk university

데이터 분석 단계

2020.07

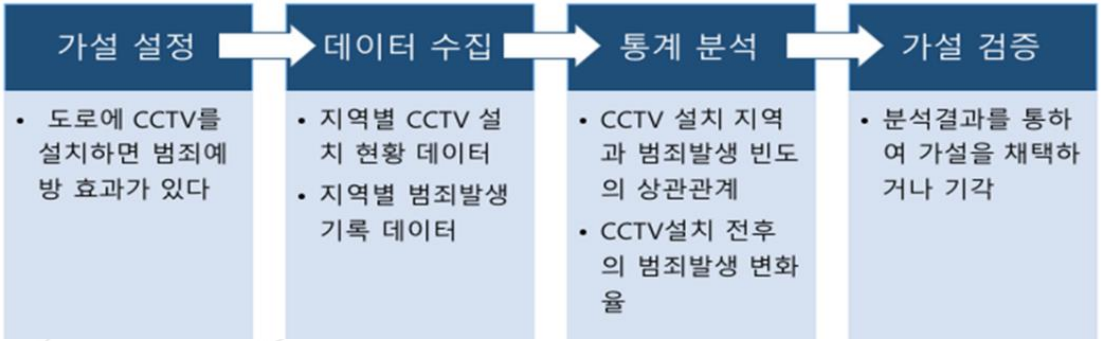
데이터 분석 이란?

- 데이터 분석
  - 데이터를 이용하여 크고 복잡한 현상에서 유의미한 패턴을 찾고 그로부터 의사결정에 필요한 통찰을 얻는 행위

- 확증적 분석 vs 탐색적 분석

  - 확증적 데이터 분석 (CDA: Confirmatory Data Analysis): 엄격하고 체계적인 방법으로 가설 검증
    가설을 설정한 후 수집한 데이터로 가설을 평가하고 추정하는 전통적인 분석 기법

    추론통계를 주로 사용하여 설문조사, 논문에 대한 내용을 입증하는데 많이 사용
    예) CCTV 범죄 예방효과를 가설로 설정, 관련 데이터 수집한 후, CCTV와 범죄 발생 빈도의 상관관계를 파악하여 가설을 검증하는 방식

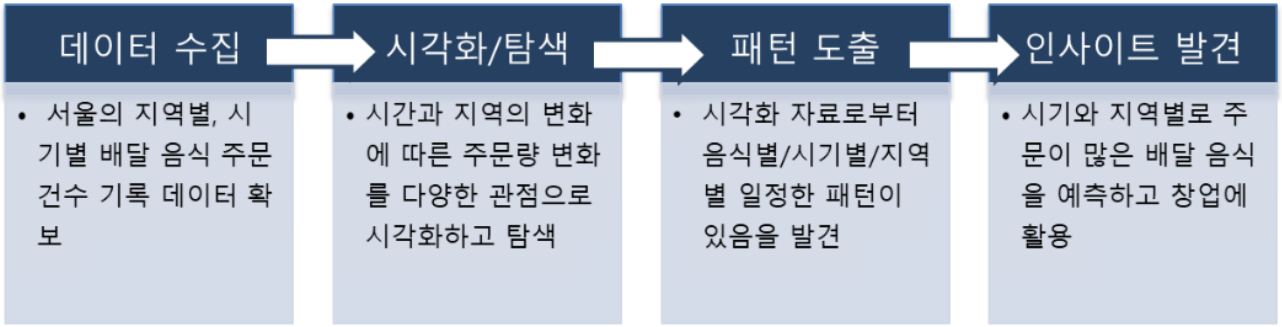


데이터 분석 이란?

- 확증적 분석 vs 탐색적 분석

  - 탐색적 데이터 분석 (EDA: Exploratory Data Analysis): 데이터 패턴에서 인사이트 발견
    데이터 시각화 기법을 통해 데이터의 특징과 구조로 부터 통찰을 얻는 귀납적 분석 기법

    선입견 없이 유연하게 데이터를 탐색하고 기술 통계 기법을 주로 사용하며 비교적 최근에 만이 사용하는 분석 방법
    예) 지역별/시기별 배달음식 주문 데이터를 시각화 하고 탐색하여 향후 매출이 높을 것으로 예측되는 장소와 시기에 창업



데이터 분석 이란?

- 확증적 분석 vs 탐색적 분석

  - 배달업종별 이용 통화량을 다운로드하여 시각화한 사례
    - 1년중 배달음식 주문이 적은 시기는 설날과 추석
    - 1년중 배달음식 주문이 가장 많은 시기는 12월24일
    - 점심에는 중국음식을 많이 주문하고 저녁에는 치킨을 많이 주문
    - 강서구 주민들이 배달 음식을 가장 많이 주문

탐색적 데이터 분석 사례 - 2017년 배달업종별 주문 건수 시각화

동국대학교 dongguk university

데이터 분석 이란?

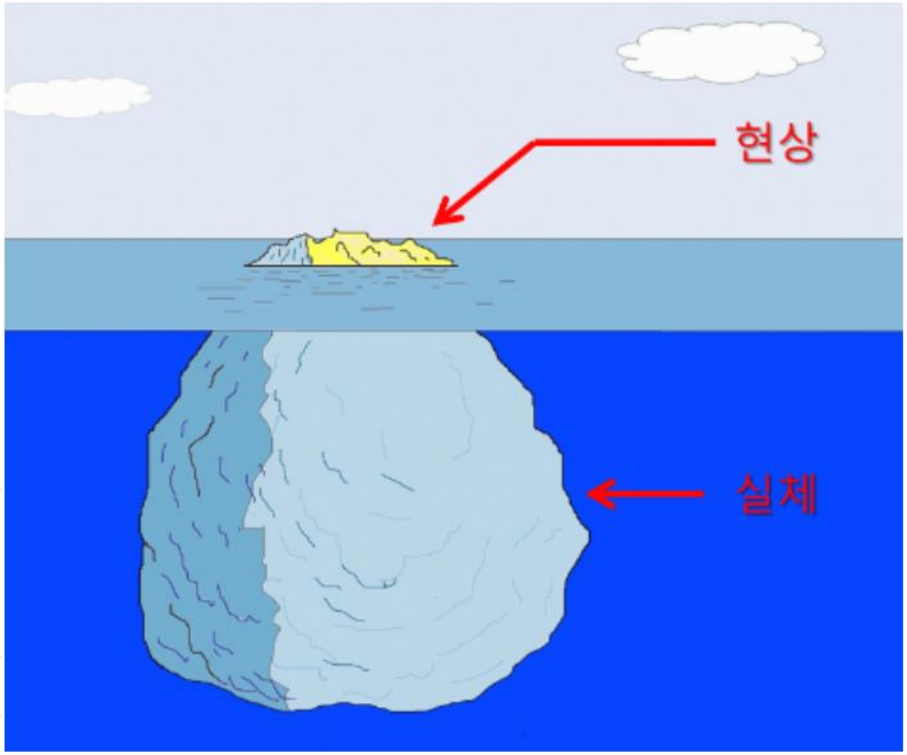
- 확증적 분석 vs 탐색적 분석 장단점

구분	확증적 데이터 분석	탐색적 데이터 분석
특징	• 기존 가설의 유효성 검증 • 기존 연구에 기반하여 수행 • 엄격한 절차와 방법	• 새로운 가설의 생성 • 시각화된 데이터로부터 패턴 발견 • 유연한 절차와 방법
장점	• 검증된 이론과 모형을 갖추고 있음	• 분석과정에서 유연하게 가설을 설정할 수 있음
단점	• 선입견이 개입되어 예상치 못한 결과의 사전 탐지가 어려울 수 있음	• 명확한 분석 목표가 없으면 방황할 가능성이 높음

데이터 분석 이란?

- 현상의 뒷면에 있는 실체를 찾아라

  - 현상은 발견하기 쉽지만, 현상의 이면에 있는 실체를 밝혀내려면 치열한 고민이 필요
  - 통찰 (Insight) 사전적 의미: "예리한 관찰력으로 사물은 꿰뚫어 보는 것"



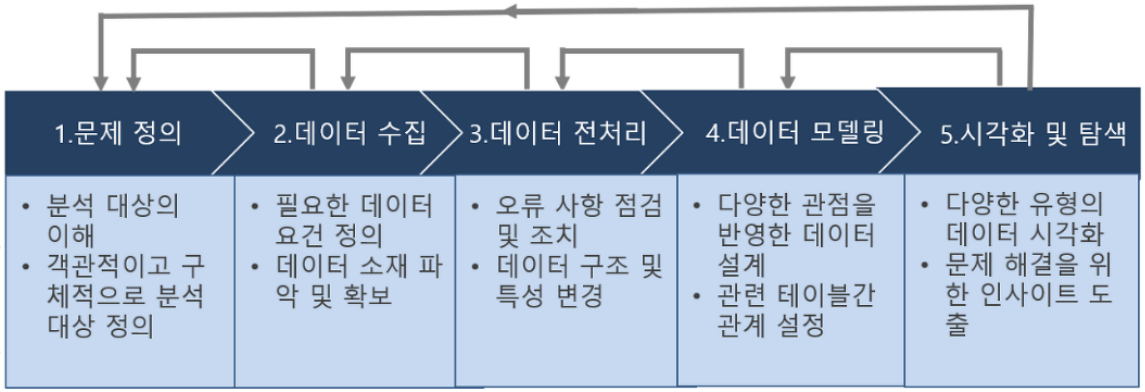
현상의 뒷면에 있는 실체

데이터 분석 이란?

- 확증적 분석 기법과 탐색적 분석 기법의 장점을 적용한 확증적 데이터 분석 기법

  - 확증적 데이터 분석 기법 단계

    - 문제의 정의 단계: 분석하고자 하는 분야를 이해하고, 해결해야 할 문제를 객관적이고 구체적으로 정의
    - 데이터 수집 단계: 분석에 필요한 데이터 요건을 정의하고 데이터를 확보
    - 데이터 전처리 단계: 수집한 데이터에 존재하는 결측값이나 오류를 수정/보완 경우에 따라서 데이터 구조나 특성을 변경
    - 데이터 모델링 단계: 하나의 테이블(데이터 셋)이 아닌 다수의 테이블을 이용하여 분석을 하는 경우 데이터 모델링이 필요
    - 시각화 및 탐색 단계: 다양한 도구를 이용하여 데이터를 시각화 하고 탐색을 통하여 문제를 해결



공공데이터 분석 절차

데이터 분석 이란?

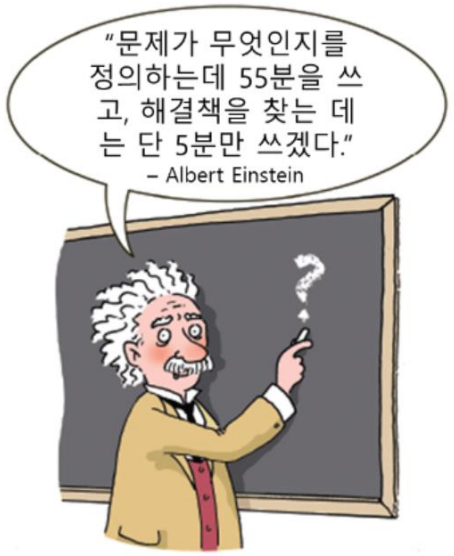
- 문제의 정의 단계: 가장 중요하지만 가장 어려운 단계

  - 문제는 분석의 대상이며 분석의 목적
  - 문제가 제대로 설정되지 않으면 분석 목표가 불분명 해짐

  - 공공 분야에서 문제 정의가 어려운 이유
    - 많은 사람들이 공감할 만한 가치가 있는 문제를 찾아야 한다.
    - 향후 정의된 문제 해결을 위한 구체적인 행동이 수반되어야 한다.
    - 데이터의 제약사항(데이터 확보 가능성 등)을 극복해야 한다.
    - 분석을 위한 전문가와 분석 기간을 확보하여야 한다.

문제 정의를 잘 하려면, 무엇보다 잘 알거나 관심이 많은 분야를 선택, 모든 사람들이 명료하게 이해 할 수 있도록 구체적이어야 한다

예) 서울의 교통문제는 심각한가? -> 서울 시민의 평균 출퇴근 시간은? 등



아인시타인은 이렇게 말했다.

데이터 분석 이란?

- 데이터 수집 단계: 가장 중요하지만 가장 어려운 단계

  - 주변에서부터 분석에 필요한 데이터 수집
  - 공공기관을 중심으로 데이터 공개

[전체]

- 공공데이터 포털 : https://www.data.go.kr/
- 서울시 열린데이터 광장 : http://data.seoul.go.kr/

[행정]

- 주민등록 인구통계 : http://27.101.213.4/
- 지방행정 데이터 : http://localdata.kr/

[지도]

- 국가 공간정보 포털 : http://www.nsdi.go.kr/

[건축]

- 건축데이터 민간 개방 시스템 : http://open.eais.go.kr/
- 국가공간정보포털 : http://data.nsdi.go.kr/dataset
- 등기정보광장 : https://data.iros.go.kr/

[기상]

- 기상 자료 개방 포털 : https://data.kma.go.kr/

[관광]

- TourAPI : http://api.visitkorea.or.kr

[농림]

- 농림축산부 : http://www.mafra.go.kr/mafra/322/subview.do

[금융]

- 금융빅데이터 개방 시스템 : https://credb.kcredit.or.kr/
- 금융데이터 거래소 : https://www.findatamall.or.kr/

[치안]

- 경찰청 공공 데이터 개방 : https://www.police.go.kr/portal/main/contents.do?menuNo=200527

[문화]

- 문화 데이터 광장 : https://www.culture.go.kr/data/

[복지]

- 보건복지 데이터 포털 : https://data.kihasa.re.kr

[교통]

- 국가 교통 DB : https://www.ktdb.go.k
- 교통사고 분석 시스템 : http://taas.koroad.or.kr/

[전기]

- 전력데이터 개발 포털시스템 : https://bigdata.kepco.co.kr/

[기타]

- 데이터 스토어 : https://www.datastore.or.kr/
- SKT 빅데이터 허브 : https://www.bigdatahub.co.kr/

데이터 분석 이란?

- 데이터 수집 단계: 가장 중요하지만 가장 어려운 단계
  - 서울시 빅데이터캠퍼스 (https://bigdata.seoul.go.kr)
  - 통계 빅데이터센서 (https://data.kostat.go.kr)
  - 빅카인즈(언론 데이터) (https://www.bigkinds.or.kr/)
  - 데이터산업진흥원(데이터안심구역) (http://datakorea.datastore.or.kr/)

데이터 분석 이란?

- 데이터 전처리 단계: 가장 많은 수고가 필요한 단계

  - 데이터 과학의 80%는 데이터 클리닝에 소비,나머지 20%는 데이터 클리닝하는 시간을 불평하는데 쓰인다. (Kaggle 창립자 Anthony Goldbloom)

  - 분석을 위하여 수집한 데이터가 바로 분석에 쓰이는 경우는 거의 없다.
  - 누락된 항목이 있거나 분석에 부적합한 구조 등 전처리가 필요한 경우가 대부분

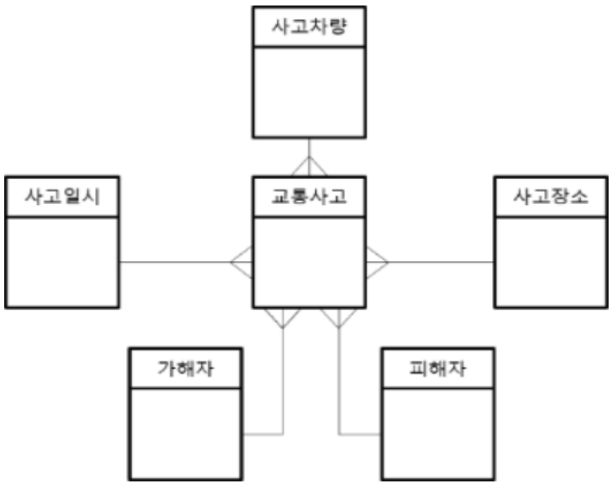
  - 중복값 제거
  - 결측값 보정
  - 데이터 연계/통합
  - 데이터 구조 변경

데이터 분석 이란?

- 데이터 모델링 단계: 관점별로 나누고 쪼개어 보기

  - 다수의 테이블을 연계하는 행위를 관계설정이라고 하고 모델링이라고 부른다.
  - 모델일 기법으로 많이 알려진 방법중 하나는 스타 스키마이다.
  - 스타 스키마는 한 개의 사실(fact) 테이블과 여러 개의 차원 (dimension)으로 구성

  - 교통사고 분석을 위한 데이터 모델링 예시



스타스키마 모델링 기법

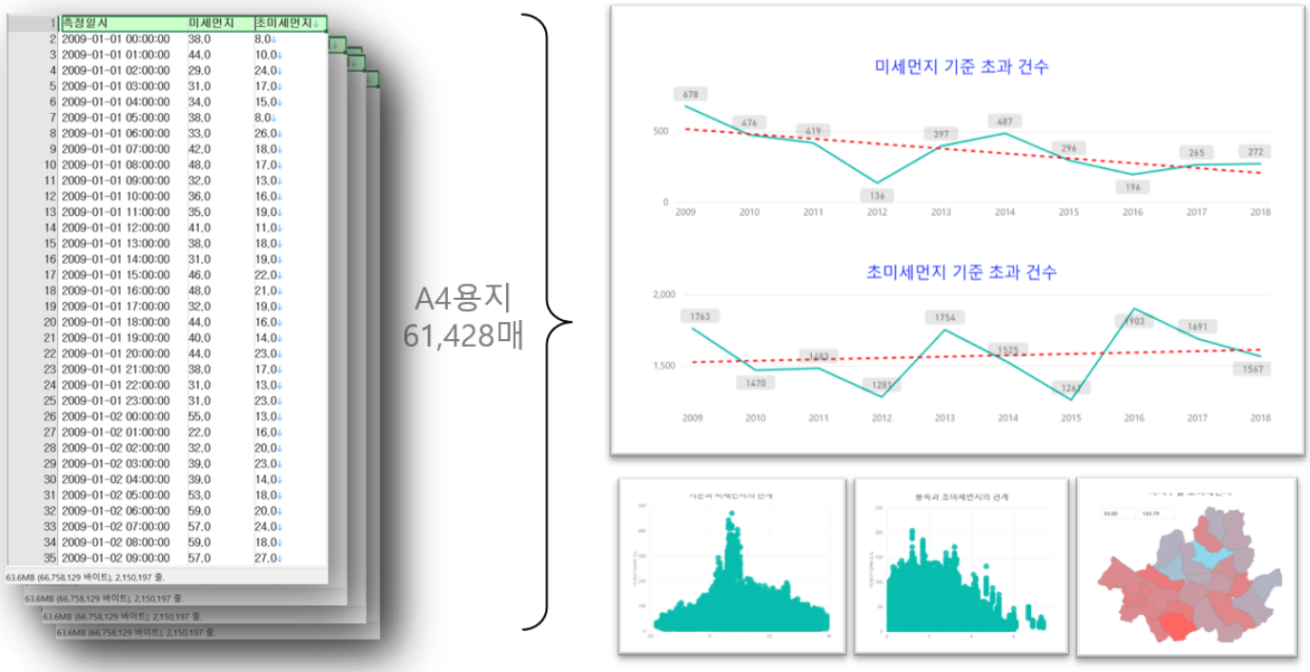
동국대학교
dongguk university

데이터 분석 이란?

- 시각화 및 탐색 단계: 패턴을 찾고 인사이트를 얻기

  - 데이터 시각화는 대량의 데이터를 요약하고 사람이 판단하기 쉬운 형태의 이미지로 표현하여 데이터 안에 숨겨진 유의미한 인사이트를 발견할 수 있도록 제공

  - 예) 서울의 미세먼지 농도 변화 추이



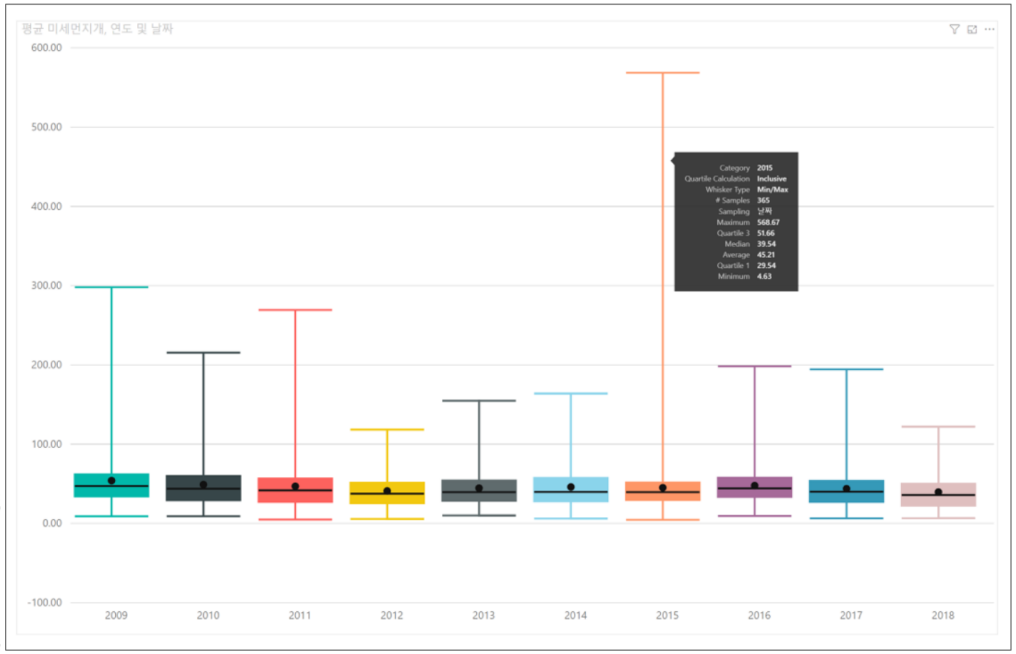
서울의 10년간(2009년~2018년) 미세먼지 농도 변화 추이

동국대학교
dongguk university

데이터 분석 이란?

- 시각화 및 탐색 단계: 패턴을 찾고 인사이트를 얻기

  - 데이터 시각화 및 탐색 단계에서 데이터를 요약하고 설명하는 방법으로 기술 통계를 많이 사용
  - 수집한 데이터를 요약, 묘사, 설명하는 통계 기법으로 데이터의 대표값(평균, 중위값, 최빈값 등) 및 분포 등을 이용
  - 예) 서울의 미세먼지 농도 분포를 박스플롯으로 시각화

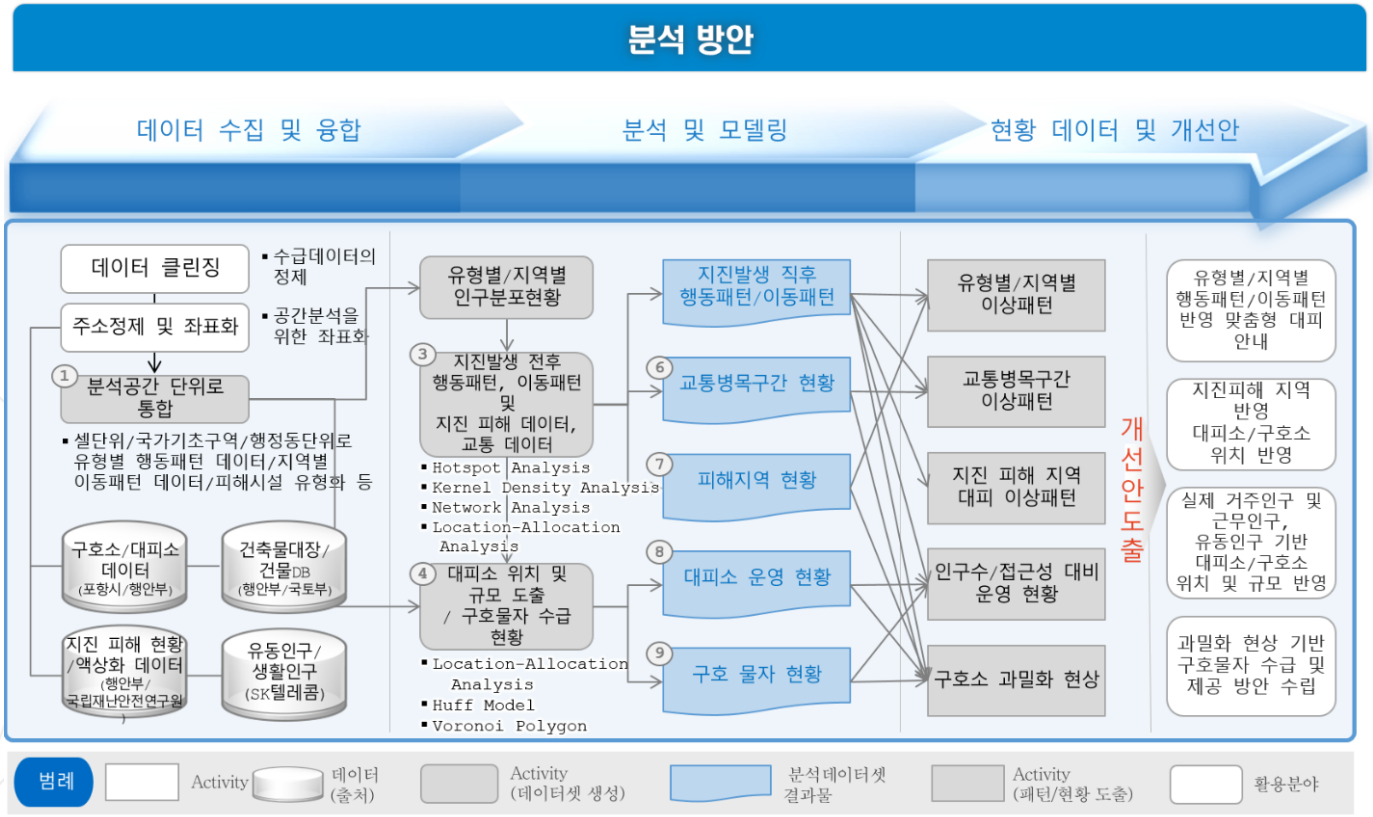


서울의 연도별 미세먼지 분포

동국대학교 dongguk university

데이터 분석 이란?

- 데이터 분석
  - 목적 : 지진 발생 직후 국민행동분석 및 이동 패턴분석을 통해 개선된 지진 대응 행동 요령을 제공하고 데이터 기반의 최적 대피소 위치와 규모를 산출하는 등의 지진대응체계 개선이 필요



동국대학교 dongguk university

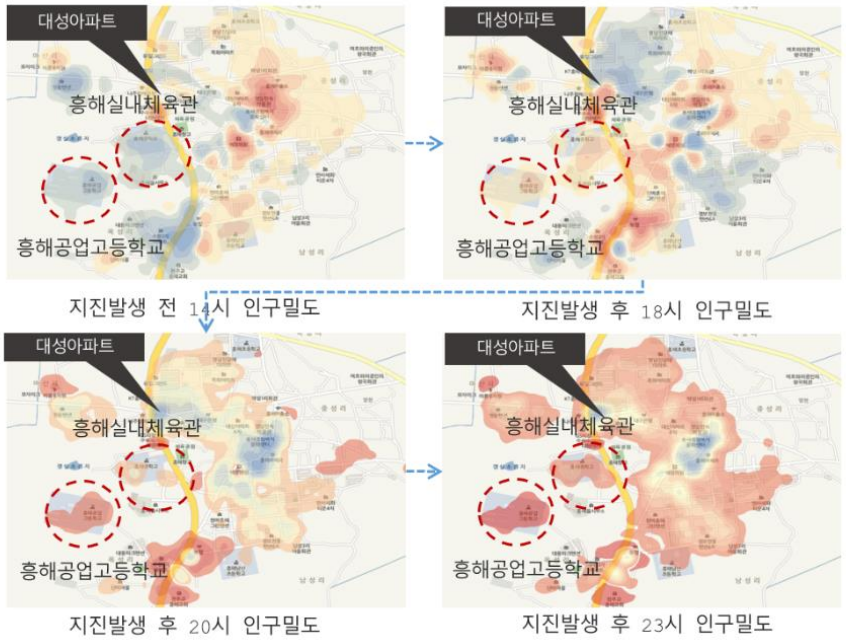
빅데이터 분석 사례 (http://www.bigdata.go.kr/)

- 빅데이터 기반의 맞춤형 지진대응체계 과학화

  - 목적 : 지진 발생 직후 국민행동분석 및 이동 패턴분석을 통해 개선된 지진 대응 행동 요령을 제공하고 데이터 기반의 최적 대피소 위치와 규모를 산출하는 등의 지진대응체계 개선이 필요

분석 결과

* 흥해실내체육관 구호소 쏠림 현상으로 인한 과밀화현상 분석 결과 예시



흥해실내체육관 주변
쏠림/과밀화 현상

1. 지진 전 전주대비 감소

2. 지진 4시간 후
   - 실내체육관과 고등학교 주변 인구밀도 증가
   - 주거지역 인구밀도 감소

3. 지진 6시간 후
   - 주거지역 인구밀도 감소
   - 대피소 인구밀도 증가

4. 지진 9시간 후
   - 대피소 인구밀도 증가
   - 주거지역 인구밀도 증가 또는 감소

5. 시간이 지날수록 처음에는 인구가 타 지역으로 이동을 하지만 시간이 지나면 다시 주거지 근처로 인구밀도가 높아지는 것을 알 수 있으며 특정지역에서 특히 많은 인원이 감소하는 것이 보임