1주차:
데이터 사이언스 개요

ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation

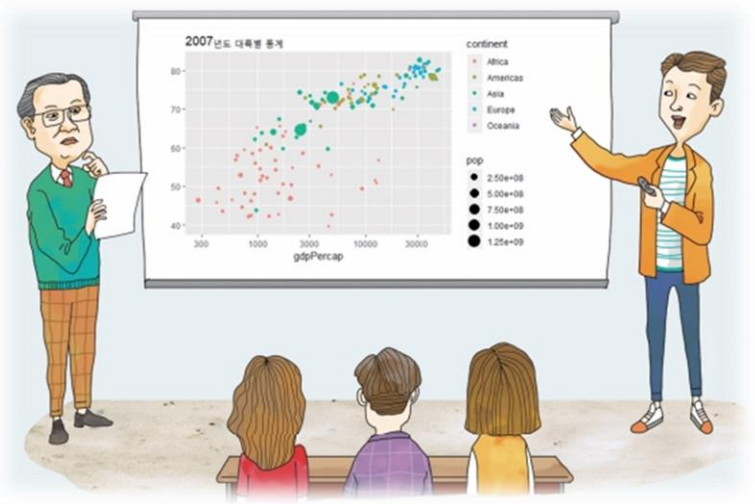
E-Mail : pcs8321@naver.com

학습목표 (1주차)

- 빅데이터 시대의 이해
- 데이터 사이언스의 정의
- 데이터 사이언스 절차
- 데이터 사이언스 분야
- 데이터 사이언티스트 정의

01 CHAPTER

데이터 과학 개요



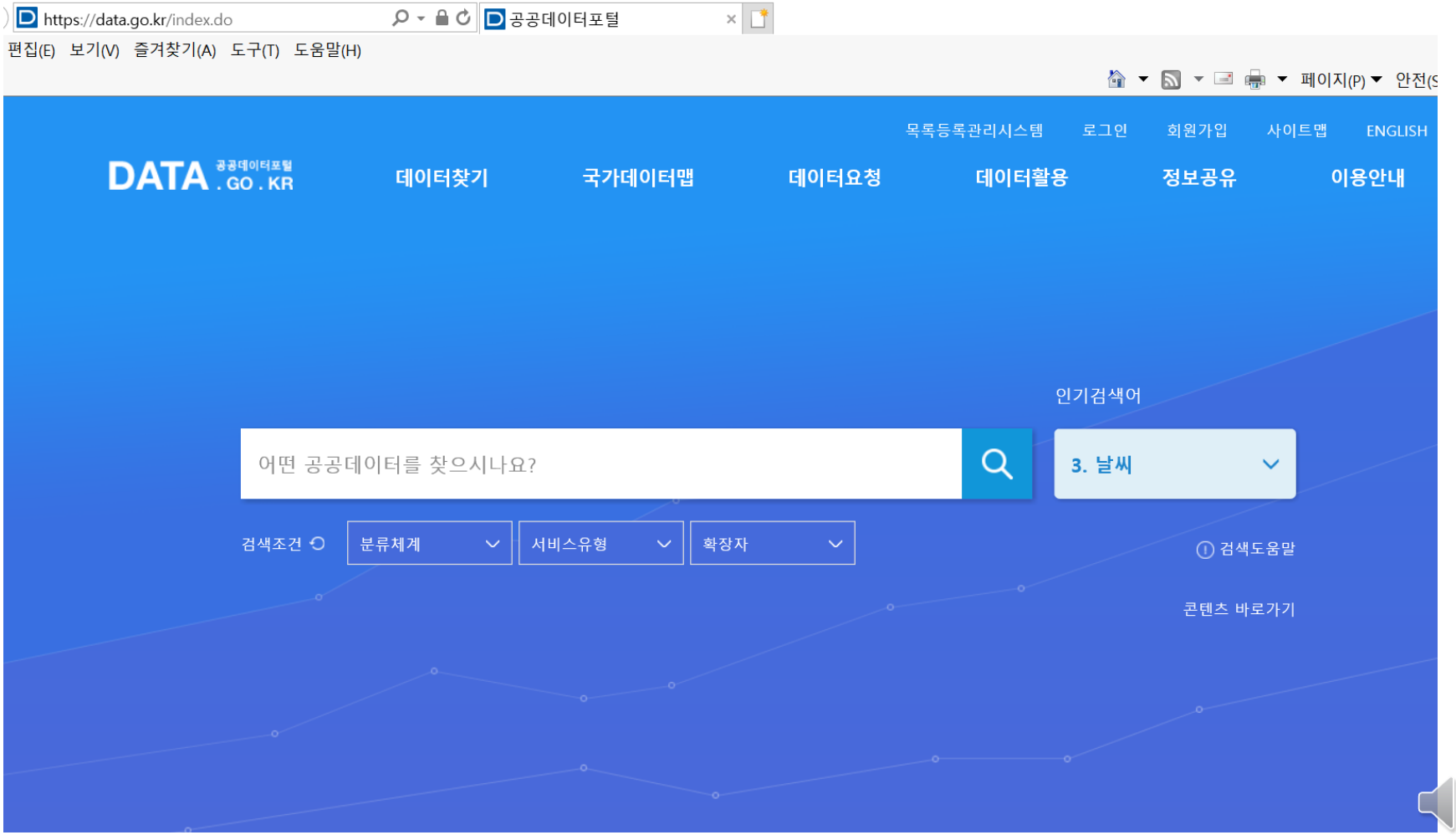
CONTENTS

1.2 데이터 과학 열풍

- 캐글은 가장 큰 데이터 과학 커뮤니티
  - 2017년 6월에 회원 100만명 돌파 (194개국)
  - 2019년 1만 3천개 이상의 데이터를 공개
  - 데이터 과학 경진 대회
    - 2018년 12월 기준으로 19개의 대회가 열리고 있음 (25,000~100,000$ 상금)
    - 가장 큰 상금은 100,000$이 걸린 "using news to predict stock movements" 대회 ← 2천 팀 이상이 참여
- 데이터 과학 분야는 공개 정신이 강함
  - 캐글 사이트: 데이터, 소스코드, 멘토링 등
  - 데이터 과학 언어: R과 파이썬
  - 우리나라는 공공데이터 포털 (data.go.kr)

1.2 데이터 과학 열풍

- 우리나라 정부가 제공하는 공공데이터 포털 (data.go.kr)



1.2 데이터 과학 열풍

- 우리나라 정부가 제공하는 공공데이터 포털 (data.go.kr)

데이터 1번가

전체 5,122건

등록일 순 | 조회수 순 | 좋아요 순

NO.		제목	요청자	조회수	좋아요	등록일
5122	답변대기	자동차주행거리 데이터	익명	2	0	2021-02-03
5121	답변대기	식품 성분(원재료, 복합 원재료의 원재료) 정보 api	익명	3	0	2021-02-03
5120	답변완료	전국 tmr사료 공장 현황	익명	19	1	2021-02-02
5119	답변완료	국군내 이슬람교도 수를 알고 싶습니다.	익명	55	1	2021-01-31
5118	답변완료	아파트명,전용면적, 공급면적 조회 data	김상*	27	1	2021-01-30
5117	답변완료	서울 중학교별 학년별 남녀 학생수가 알고 싶습니다.	익명	28	1	2021-01-29
5116	답변완료	eoa(ethylene oxide adduct 산화에틸렌유도체) 가격동향 요청	익명	14	1	2021-01-29
5115	답변완료	서울에 위치한 중소기업 리스트 요청	익명	62	1	2021-01-28
5114	답변완료	수도권 대학교 난방공급 시설현황 자료요청	백민*	26	1	20
5113	답변완료	지하철 보행자 사고 데이터 요청	한동*	27	1	202

개방 요청하기

1.2 데이터 과학 열풍

우리나라 정부가 제공하는 공공데이터 포털 (data.go.kr)

서울 중학교별 학년별 남녀 학생수가 알고 싶습니다.

요청자 익명 **제공기관** 기관알수없음 **조회수** 29 **등록일** 2021-01-29

서울 중학교별 학년별 남녀 학생수가 알고 싶습니다. 찾아보니 2014년도 버전뿐이 없어서요. csv 파일 있으면 부탁 드립니다.

답변

답변자명 공공데이터활용지원센터 **답변일** 2021-02-01

안녕하세요. 공공데이터활용지원센터입니다.

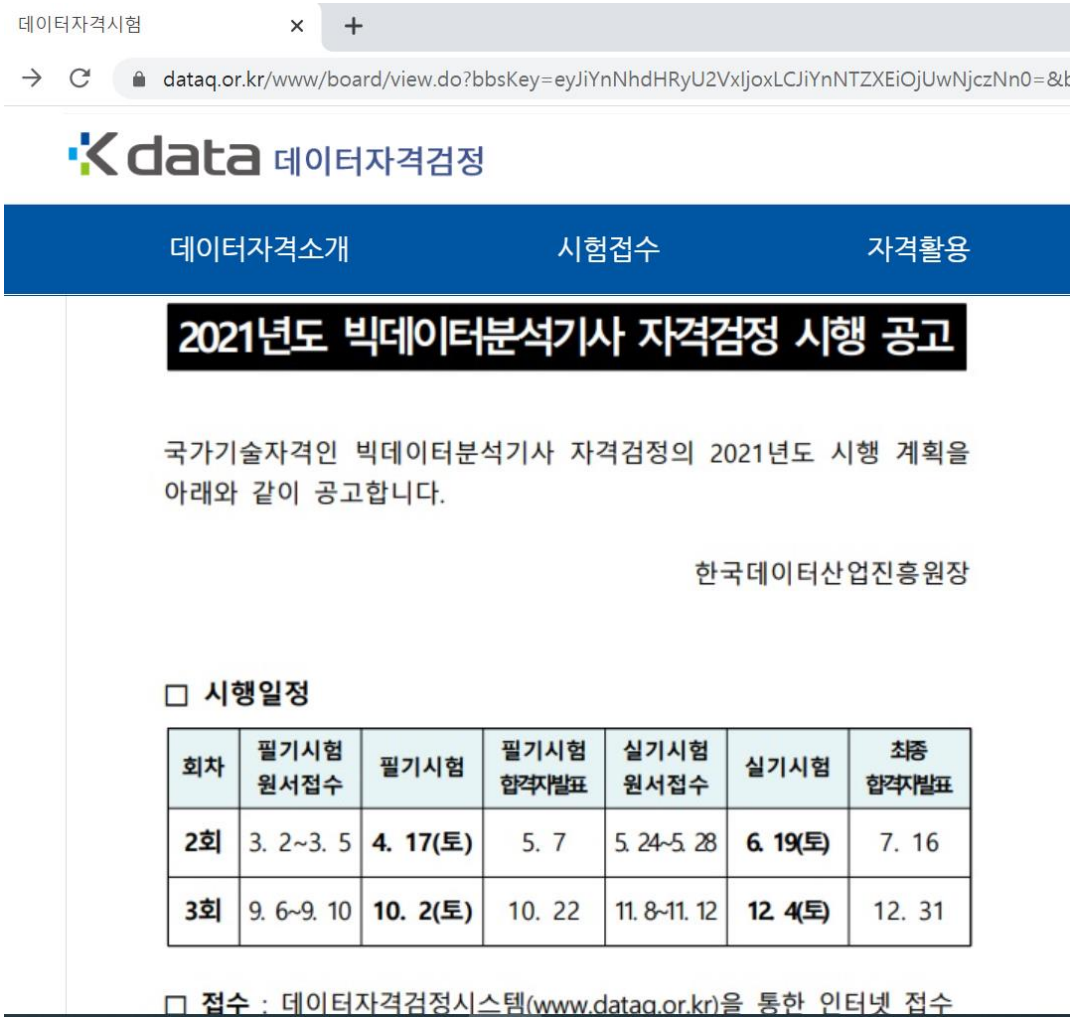
공공데이터포털은 공공기관이 생성 또는 취득하여 관리하고 있는 공공데이터를 한 곳에서 제공하는 통합 창구이며, 데이터 1번가 또한 대국민이 필요로 하는 공공데이터에 대한 의견을 자유롭게 나눌 수 있는 데이터 소통창구이므로 데이터를 직접 제공해드릴 수 없는 점 양해 부탁드립니다.

해당 정보는 서울 열린데이터 광장 사이트(http://data.seoul.go.kr/datalist/202/s/2/datasetview.do) > [통계] > [서울통계서비스] > [교육] > [초중등교육] > [중학교] 페이지에서 년도별로 해당 정보를 찾아보실 수 있습니다.

추가적인 내용의 데이터를 원하신다면 공식적인 절차인 공공데이터 '제공신청'을 해주시면 담당 기관으로부터 더욱 정확한 답변을 받으실 수 있습니다.
그 외 기타 문의사항이 있으신 경우 공공데이터포털 대표번호 (1566-0025)로 문의하여 주시기 바랍니다.

감사합니다.

1.2 데이터 과학 열풍



데이터자격시험 ×

dataq.or.kr/www/board/view.do?bbsKey=eyJiYnNhdHRyU2VxIjoxLCJiYnNTZXEiOjUwNjczNn0=&b

data 데이터자격검정

데이터자격소개 시험접수 자격활용

2021년도 빅데이터분석기사 자격검정 시행 공고

국가기술자격인 빅데이터분석기사 자격검정의 2021년도 시행 계획을 아래와 같이 공고합니다.

한국데이터산업진흥원장

□ 시행일정

회차	필기시험 원서접수	필기시험	필기시험 합격자발표	실기시험 원서접수	실기시험	최종 합격자발표
2회	3. 2~3. 5	**4. 17(토)**	5. 7	5. 24~5. 28	**6. 19(토)**	7. 16
3회	9. 6~9. 10	**10. 2(토)**	10. 22	11. 8~11. 12	**12. 4(토)**	12. 31

□ 접수 : 데이터자격검정시스템(www.dataq.or.kr)을 통한 인터넷 접수

□ **대상 직무분야 및 자격종목, 주무부처**

세부직무분야(전문위원회)	정보관리*
자격종목명	빅데이터분석기사
주무부처(담당과)	과학기술정보통신부(융합신산업과) 통계청(통계정책과)

* **(국가기술자격 정책심의위원회운영규정 제6조제2항)** 종목의 신설 등으로 해당하는 세부직무분야가 없는 경우에는 관련성이 큰 세부직무분야별전문위원회에 「국가기술자격법 시행규칙」 제7조에 따른 주무부장관의 추천을 받은 5명 이내의 해당 분야 전문가를 추가하여 심의할 수 있음

1.2 데이터 과학 열풍

- 데이터 과학 대회에 참가하자!

표 1-1 국내외 데이터 과학 대회

대회 이름	주최	비고
데이터 사이언스 컴피티션 Data Science Competition	서울대학교 통계연구소	네이버, 커넥스재단, 네이버 클라우드 플랫폼 후원
디지털 헬스 해커톤 Digital Heath Hackathon	삼성융합의과학원	digitalhealthhack.org
빅 콘테스트	한국정보통신진흥협회와 한국정보화 진흥원	bigcontest.or.kr
날씨 빅 데이터 콘테스트	기상청	big.kma.go.kr/contest
데이터 사이언스 경진대회	지퍼ZPER	dacon.io
캐글 컴피티션Kaggle Competition	캐글Kaggle	kaggle.com
Data Science Game	Paris-Saclay University	datasciencegame.com
Data Hackathon	Analytics Vidhya	datahack.analyticsvidhya.com
Open Data Hackathon	키프로스 공화국Republic of Cyprus 재정부	opendatacy.com
Asia Open Data Hackathon	Asia Open Data Hackathon	odhack.asia

1.3 데이터 과학이란?

- 감성적 표현
  - 데이터 과학은 데이터라는 장난감이 널려 있는 놀이터
  - 데이터 과학자는 이 놀이터에서 데이터를 가지고 노는 사람

- 과학적 정의

an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured

정형화 또는 비정형화된 여러 형태의 데이터로부터 지식과 직관을 추출하기 위해 과학적 방법, 과정, 알고리즘, 시스템을 활용하는 다학제 학문 분야

위키피디아에 있는 '데이터 과학' 정의

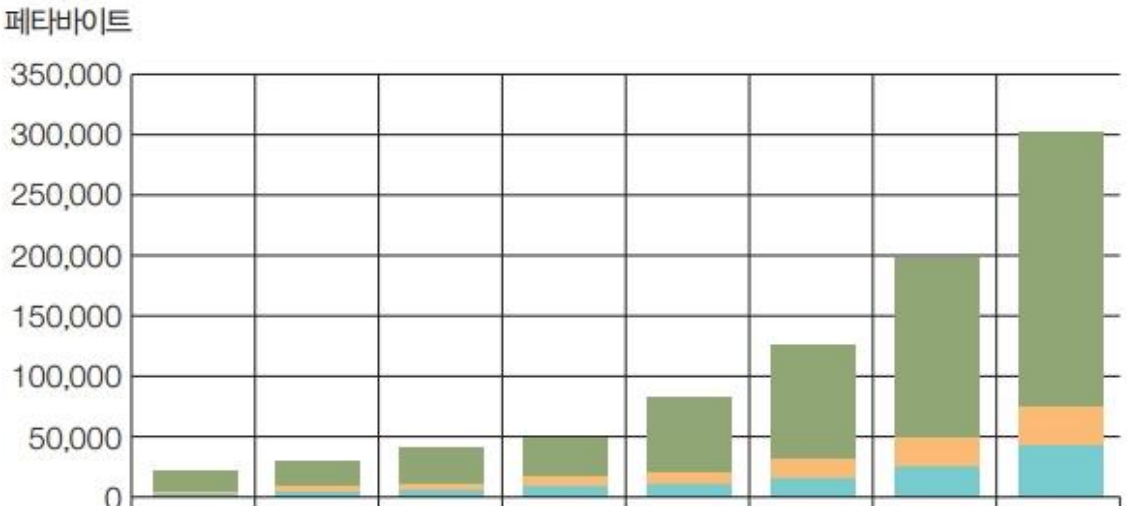
study of the generalizable extraction of knowledge from data

데이터로부터 일반화 가능한 지식을 추출하는 연구

뉴욕대학교 다르 교수의 '데이터 과학' 정의 [Vasant Dhar, "Data Science and Prediction," Communications of the ACM, 2013.]

1.3 데이터 과학이란?

- 정형 데이터와 비정형 데이터
  - 정형 데이터: 출석부, 대학의 학사와 교무 행정 데이터, 정부의 세무와 인구 데이터 등
  - 비정형 데이터: 이메일이나 편지, 영상, 동영상, 소리 등



	2008	2009	2010	2011	2012	2013	2014	2015
비정형	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
정형(데이터베이스)	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
정형(이메일)	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

그림 1-4 정형 데이터와 비정형 데이터의 증가 추세(2008년~2015년)

1.3 데이터 과학이란?

데이터 사이언스의 정의 및 필요성

- 데이터로부터 통찰력을 찾아 문제를 해결하는 것
- 데이터 사이언스의 필요성

기존의 연구

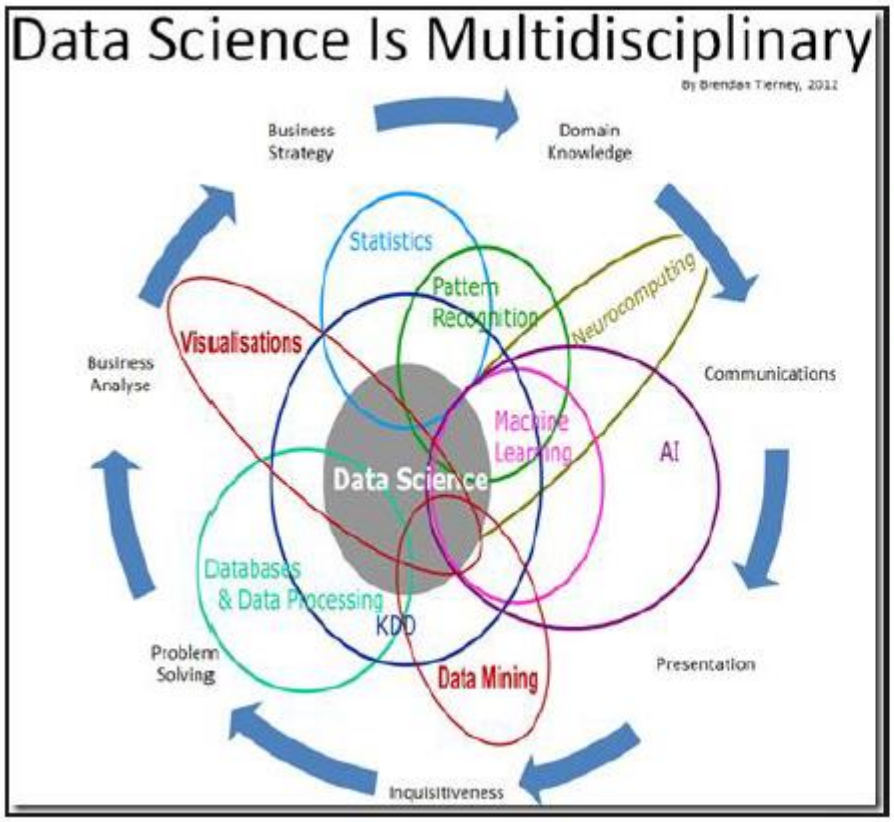
- 기존의 과학적 탐구는 원인 발견에 초점
- 어떤 현상의 원인을 연구하여 원리 발견
- 인과 관계 연구에 초점

VS.

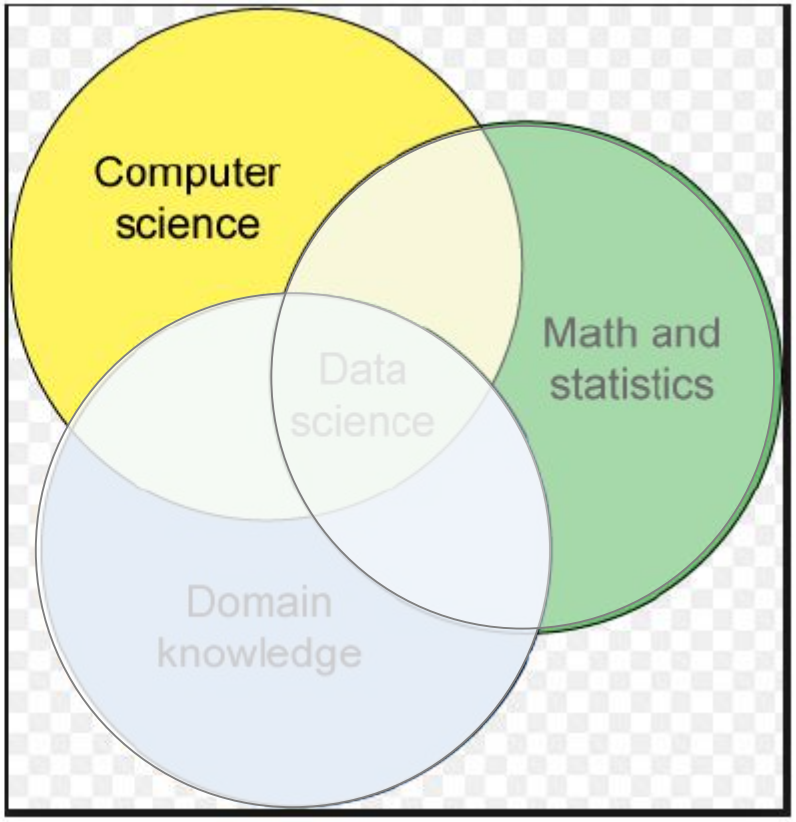
빅 데이터 분석

- 원인 탐구 보다는 관계 분석이 중요(원인 분석은 후 순위)
- 근본 원인은 모르더라도 미래 예측 연구에 가치를 둠
- 문제 해결에 필요한 패턴이나 상관 관계 연구에 초점

1.3 데이터 과학이란?



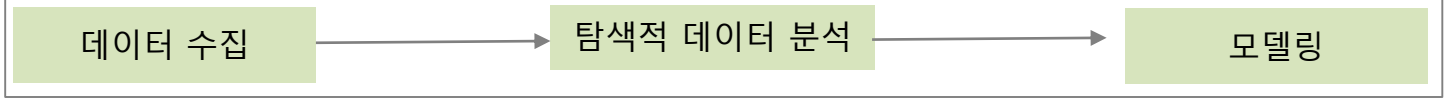
출처 : www.oralytics.com



데이터 사이언스 학과(대학원)

1.4 데이터 과학의 절차

데이터 과학의 절차(그림 1- 5)

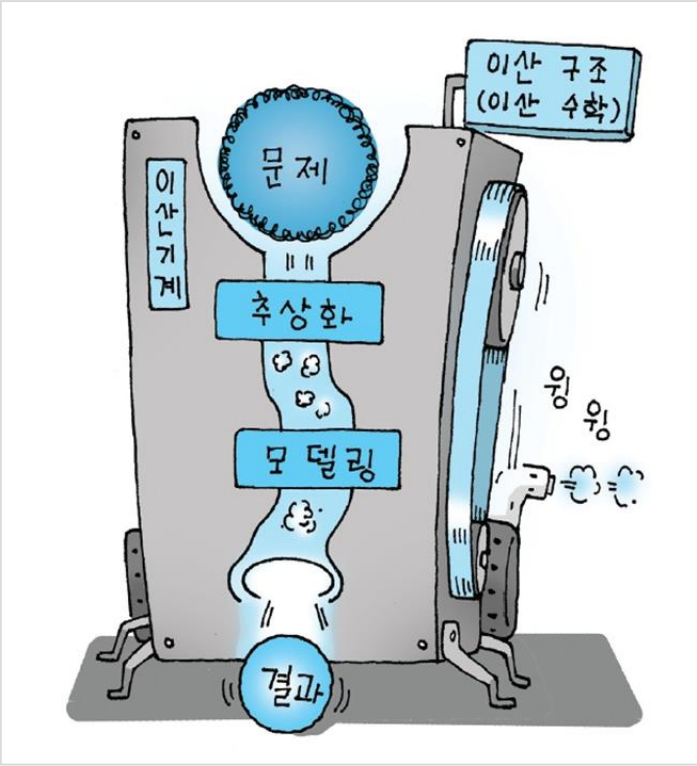


데이터 사이언스 프로세스

문제 정의	해결하려는 문제를 명확이 정의하는 것
전략 수립	문제 해결을 위해 어떤 데이터를 어떻게 분석할 것인지 결정하는 것
데이터 수집	분석에 필요한 데이터를 수집하는 것
데이터 분석	패턴 찾기, 분류, 예측 등 분석 모델을 구현하는 것
결과 적용	분석 결과를 실제 상황에 적용하고 성능을 개선하는 것

출처 : 데이터 사이언스 개론, p27

1.4 데이터 과학의 절차



출처: 4차 산업혁명 시대의 이산수학(p,22)

1. Define and understand the purpose of data mining project
2. Formulate the data mining problem
3. Obtain/verify/modify the data
4. Explore and customize the data
5. Build data mining models
6. Evaluate and interpret the results
7. Deploy and monitor the model

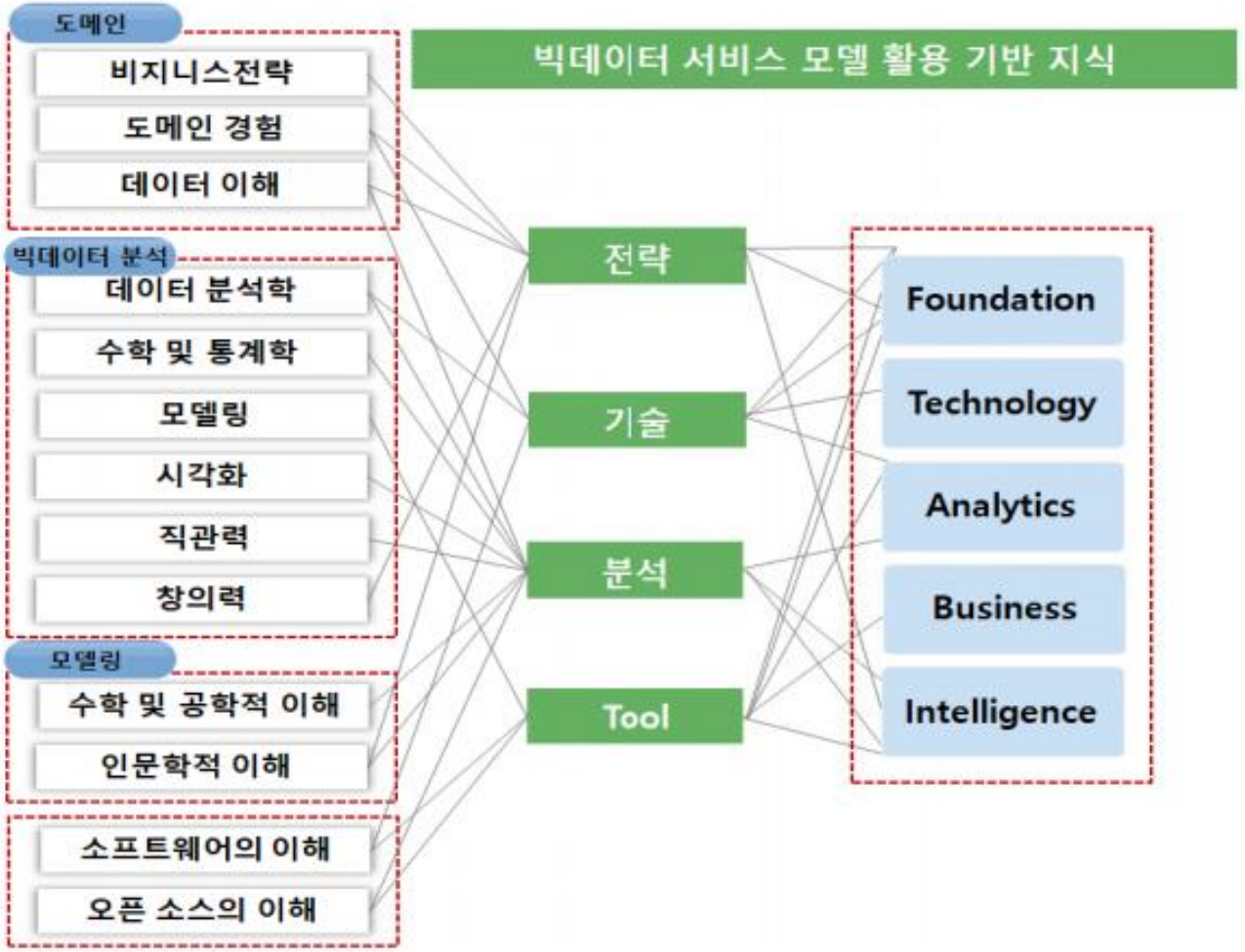
출처: 박사학위 데이터 마이닝 강의 자료

1.4 데이터 과학의 절차

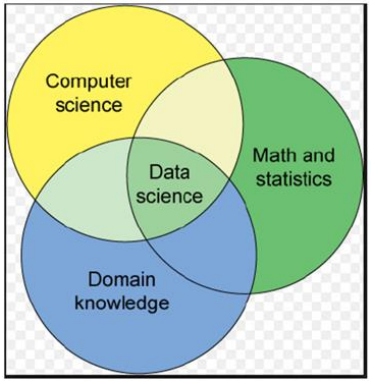
[그림 3-2] 서비스 적용을 위한 데이터 흐름도

출처 : https://www.ncs.go.kr

1.4 데이터 과학의 절차



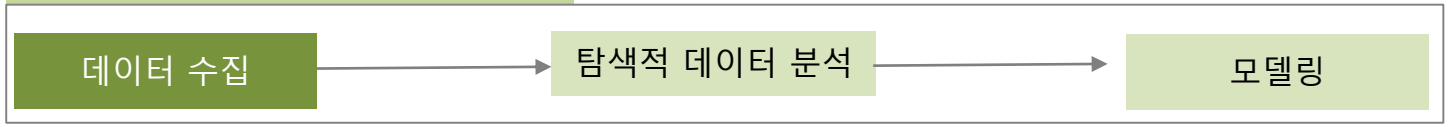
[그림 1-6] 빅데이터 서비스 모델 활용을 위한 관련 요소



출처 : https://www.ncs.go.kr

1.4 데이터 과학의 절차

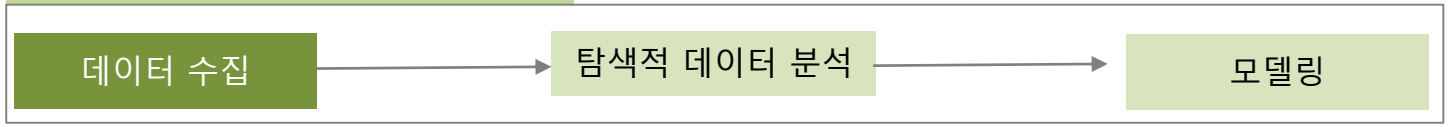
데이터 과학의 절차(그림 1- 5)



- 데이터 수집
  - 주어진 문제와 현장에 맞는 수집 계획 수립, 실제 현장에서 수집하고 기록
  - 예) 초밥 아이템으로 푸드트럭 창업
    - 매일 신선 재료 공급이 매우 중요 (덜 주문하면 덜 팔아, 더 주문하면 폐기에 따른 손해)
    - 초기 6개월 간, 매일 날씨, 미세먼지 수치, 기온, 습도, 요일, 도시락 판매 개수를 수집함

1.4 데이터 과학의 절차

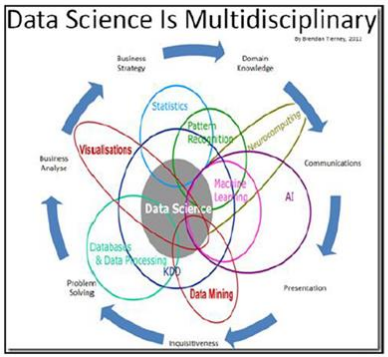
데이터 과학의 절차(그림 1- 5)

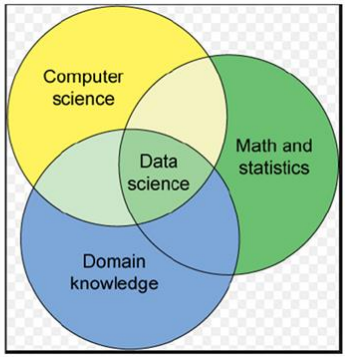


데이터 수집

- 데이터 수집은 많은 비용과 시간을 요하는 단계
- 실제로 분석 자체에 걸리는 시간보다 분석 데이터 분비에 많은 시간 필요
- 분석 전체 과정에서 70~80%의 시간은 데이터를 모으고 준비하는데 소요(수년 이상 기간이 필요 할 수 있음)
- 데이터 수집 후 데이터 전처리 단계 필요

출처 : 데이터 사이언스 개론, p32





2000년 CRM 프로젝트 실패 경험담

1.4 데이터 과학의 절차

데이터 과학의 절차(그림 1- 5)

데이터 수집 → 탐색적 데이터 분석 → 모델링

스마트 축사 ICT 도입 활용 요약

- 축산분야 ICT활용은 ICT장비도입부터 활용까지 3단계 지원사업으로 구성
- 1. ICT장비도입(컨설팅) > 2. 축산 빅데이터 플랫폼 구축 > 3. 데이터 활용(컨설팅)

ICT 융복합 확산 = ICT장비도입 / 플랫폼(데이터수집&활용서비스) / 데이터 활용 컨설팅

ICT 장비도입 지원컨설팅

- ICT 시설장비 도입 신청
- (1단계) 장비선정 및 계획서 작성
- (사업자 선정) 지자체 선정
- (2단계) 장비설치 확인 점검
- (3단계) 장비 활용, 생산성 향상 점검

축산농가+지자체+농식품부
농정원, 컨설턴트, 장비기업

축산 빅데이터 플랫폼

- (농가) 한눈에 농장 현황 파악
- (컨설턴트) 데이터 활용 컨설팅
- (민간기업) 제품 및 서비스 개발
- (유관기관) 정책지원 활용

축산농가 : 수기관리 정보입력
ICT장비기업 : 정보연계
유관기관 : 개체정보, 출하정보 등 연계

데이터 활용 컨설팅

- (공통) 시설, 생산성, 장비, 역량 분석
- (수의) 질병(방역,백신,구충) 진단 처치
- (사양) 사료, 번식, 환경 진단 처치
- (경영) 손익/원가 분석, 중장기 경영 예측

축종별 분야(수의사,양경영,컨설턴트
축산농가

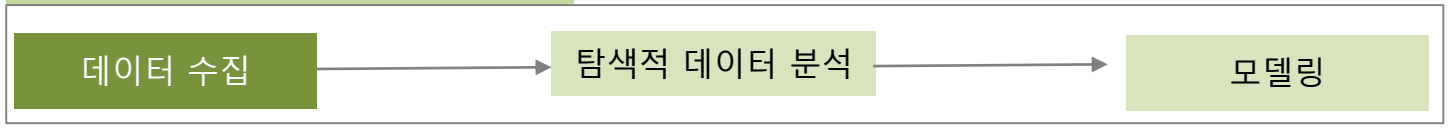

급이기


양돈 선별기


음수관리기

1.4 데이터 과학의 절차

데이터 과학의 절차(그림 1- 5)



대한민국산업현장교수
기술지원 결과

(정보통신) 분야 기술지원 결과를 아래와 같이 제출합니다.

기관대표 : 대표이사 임칠규 (인)

□ 지원제목 : 국가직무표준(NCS) 이해 및 빅데이터 기획. 분석 기술 전수와 실 사례 도출(정보통신)

□ 지원기간 : 2020. 06. 15. ~ 2020 . 07. 20.

□ 지원시간 : 총 18 시간

□ 지원교수 : 대한민국산업현장교수 박 철 수 (인)

2020년 07월 21일

1.4 데이터 과학의 절차

데이터 과학의 절차(그림 1- 5)

- 탐색적 데이터 분석 (EDA: exploratory data analysis)
  - 변수 값의 분포, 변수 사이의 상관관계 등을 살펴 데이터 특성을 파악함
  - 요약 통계량 계산, 시각화visualization 등
  - 예) 푸드트럭
    - "이 골목은 맑은 날이 80% 이상이군", "미세먼지 나쁨 수준인 날이 5%에 불과하군", "월요일은 유독 판매량이 많군", "미세먼지가 치솟으면 판매량이 줄어드는군" 등의 분석 → "자릿세가 조금 오르더라도 이 곳을 떠나지 말아야지 " 라는 의사결정
    - 요일과 일기예보를 보고 내일 주문량을 늘릴지 줄일지 가늠하는 일이 가능해짐

1.4 데이터 과학의 절차

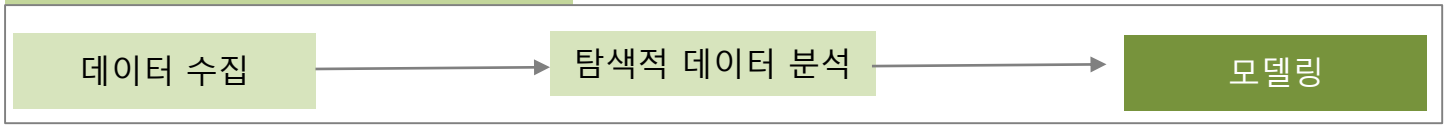
데이터 과학의 절차(그림 1- 5)

- 수집한 데이터로부터 어떤 의미를 찾기 위해서는 데이터를 분석 해야 한다. 데이터 분석에는 간단히 평균치를 구해보는 것부터 미래를 예측하는 등 다양한 형태의 데이터 분석 유형이 있다.
  - 기초 통계 분석 – 평균,중앙값,최빈값 등 데이터 분포를 파악
  - 클러스터링 – 비슷한 성격의 항복을 그루핑하기
  - 연관관계 분석 – 자주 발생하는 패턴 찾기
  - 분류 – 미리 정해진 카테고리 중 어디에 속하는지 판별하기
  - 예측 – 주가, 매출, 손익, 생산성 등 예측하기

출처 : 데이터 사이언스 개론, p32

1.4 데이터 과학의 절차

데이터 과학의 절차(그림 1- 5)



모델링

- 데이터를 가장 잘 설명하는 모델을 찾는 과정
- 모델은 변수 사이의 관계를 수학식으로 표현
- 예) 푸드트럭
  - 날씨, 미세먼지, 기온, 습도, 요일이 도시락 판매 개수에 미치는 영향을 수식으로 표현
  - 모델링을 하면, 기상청에서 알아낸 날씨, 미세먼지, 기온, 습도를 요일과 함께 모델에 입력하여 주문량을 예측할 수 있음

Thank you


2007년도 대륙별 통계
80
70
60
50
40
300
1000
3000
10000
30000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09