Big Data

데이터 분석 기획

류영표 강사
youngpyoryu@dongguk.edu

강사 소개

류영표

Youngpyo Ryu

現 동국대학교 수학과/응용수학 석사수료

現 SD아카데미 국비과정 강사

現 Upstage AI X 네이버 부스트캠프 멘토

前 메가 IT아카데미(파이썬,빅데이터) 강사

한국파스퇴르연구소 Image Mining 인턴(Deep learing)

前 ㈜셈웨어(수학 컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- 딥러닝 집중 교육과정 강사
- (재)윌튼블록체인 6일 과정 (파이썬기초, 크롤링,머신러닝)
- 서울특별시 X AI 양재허브 X 모두의연구소 (중급 NLP과정) 보조강사
- SK아카데미_HLP(임원) 1차/2차 보조강사
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- LG전자 / LG 인화원 보조강사
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의

주요 프로젝트 및 기타사항

- 제1회 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 최우수상 수상(Q&A 챗봇)
- 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

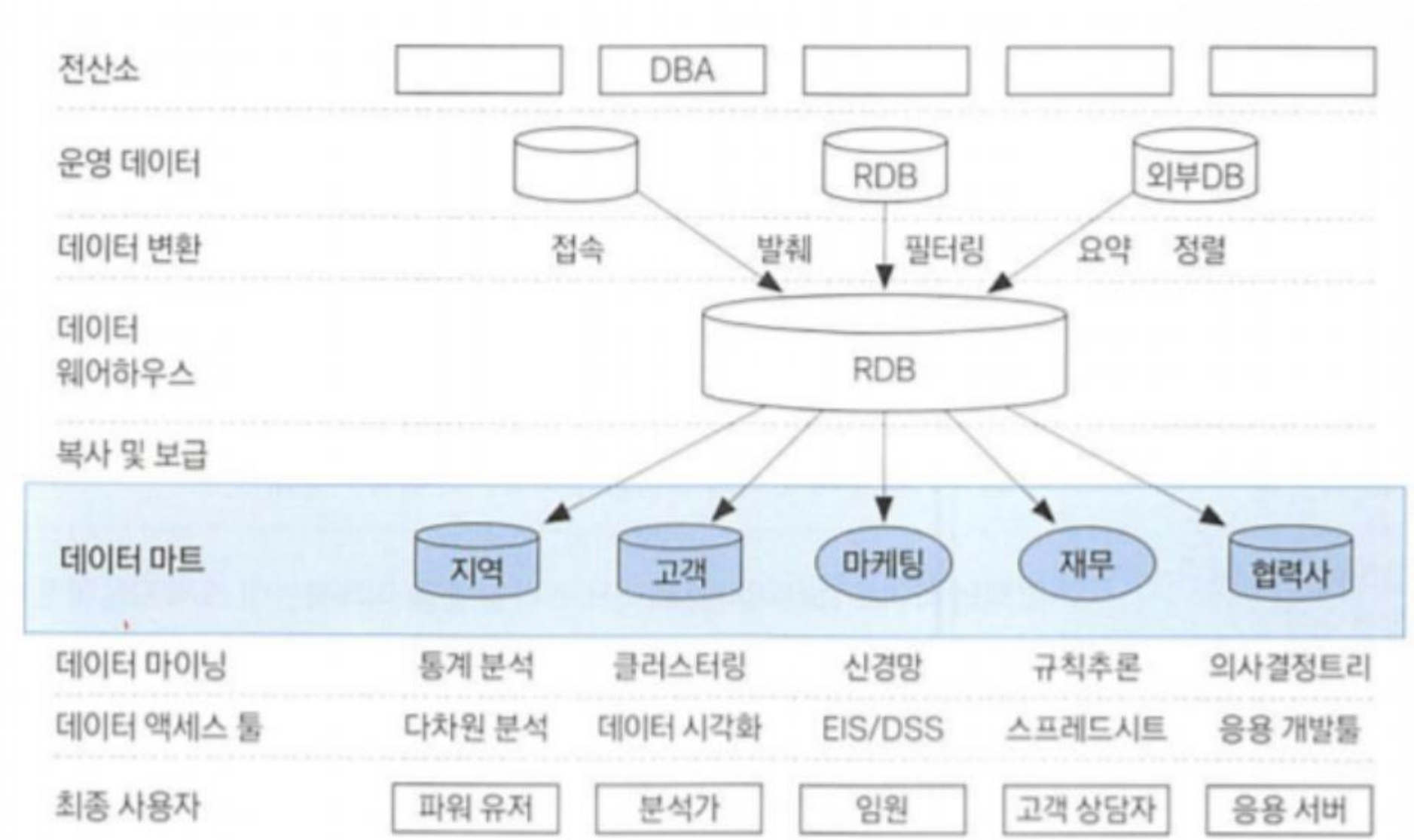
데이터 마트

- 데이터 분석은 통계에 기반을 두고 있지만, 통계지식과 복잡한 가정이 상대적으로 적은 실용적인 분야.
- 신규 데이터나 DW에 없는 데이터는 기존 운영시스템(legacy)에서 직접 가져오거나 운영데이터저장소(ODS)에서 정제된 데이터를 가져와서 DW의 데이터와 결합하여 활용
- 시각화 기법

  - 가장 낮은 수준의 분석이지만, 잘 사용하면 복잡한 분석보다 더 효율적, 대용량 데이터를 다룰 때와 탐색적 분석을 할 때 시각화는 필요
- 공간분석(GIS)

  - 공간적 차원과 관련된 속성들을 시각화하는 분석으로 지도 위에 관련된 속성들을 생성하고 크기, 모양, 선 굵기 등으로 구분하여 인사이트를 얻음.

데이터 마트


전산소
DBA
운영 데이터
RDB
외부DB
데이터 변환
접속
발췌
필터링
요약
정렬
데이터
웨어하우스
RDB
복사 및 보급
데이터 마트
지역
고객
마케팅
재무
협력사
데이터 마이닝
통계 분석
클러스터링
신경망
규칙추론
의사결정트리
데이터 액세스 툴
다차원 분석
데이터 시각화
EIS/DSS
스프레드시트
응용 개발툴
최종 사용자
파워 유저
분석가
임원
고객 상담자
응용 서버

요약 변수

- 데이터 분포 등을 잘 나타내는 변수
- 수집된 정보를 분석에 맞게 종합한 변수.
- 많은 모델들이 공통으로 사용하기에 재활용성이 높다.
- 합계, 횟수 같은 간단한 구조이므로 자동화프로그램으로 구축 가능
- 단점 : 정해놓은 기준치(기준값, 기준횟수)를 뛰어 넘을 경우 기준의 의미해석이 애매해진다. 연속형 변수를 그룹핑해 사용하자!

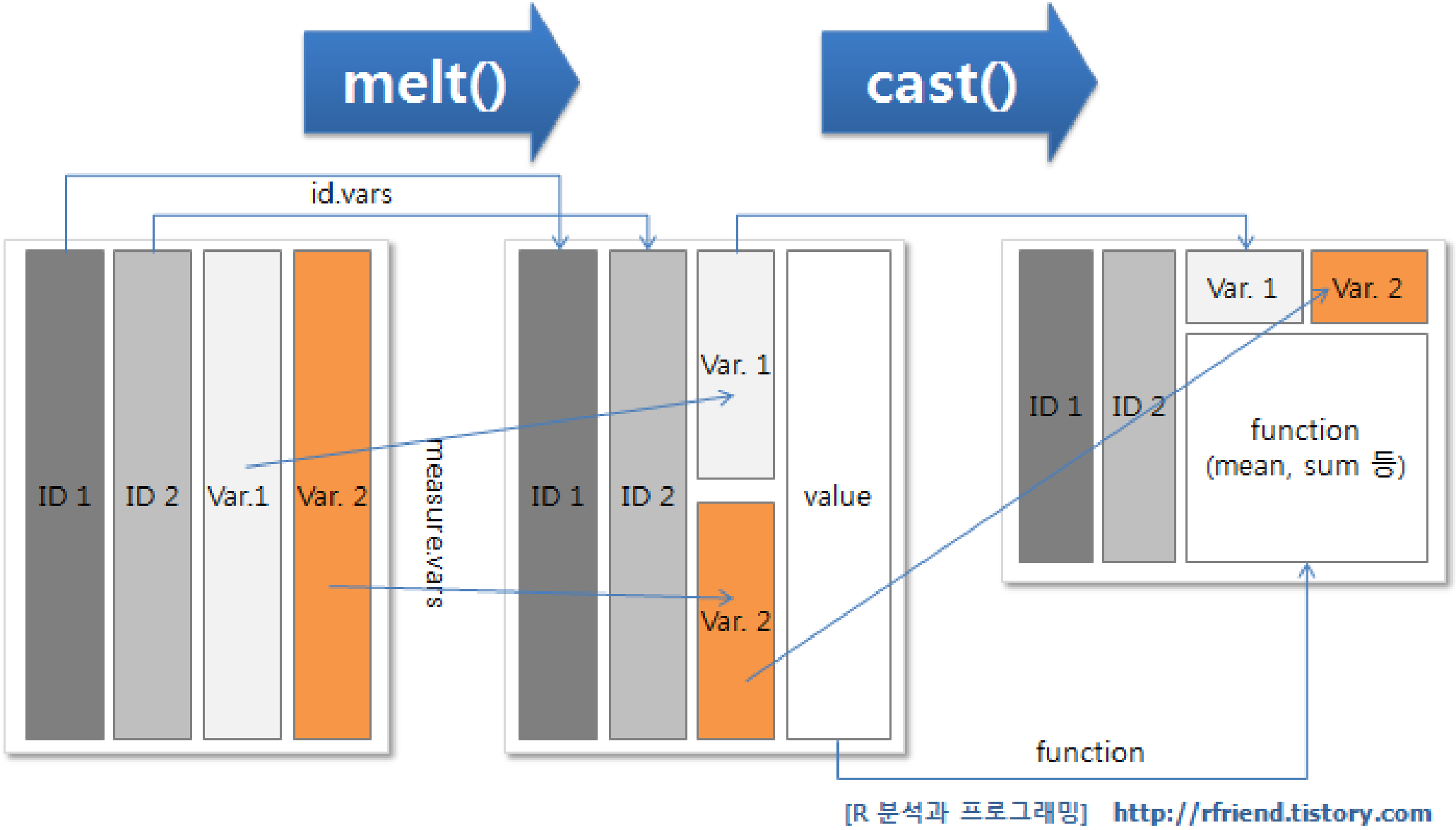
파생 변수

- 수집된 정보를 분석에 맞게 종합한 변수.
- 많은 모델들이 공통으로 사용하기에 재활용성이 높다.
- 합계, 횟수 같은 간단한 구조이므로 자동화프로그램으로 구축 가능
- 단점 : 정해놓은 기준치(기준값, 기준횟수)를 뛰어 넘을 경우 기준의 의미해석이 애매해진다. 연속형 변수를 그룹핑해 사용하자!

Reshape

- DB에 들어있는 데이터를 내렸을때, 혹은 설문조사의 데이터를 받았을때, 또는 분석 과정 중의 데이터 셋이 분석가가 하고자 하는 통계 분석, 데이터마이닝 분석이나, ggplot2등의 그래프 / 시각화를 위해 필요한 데이터 구조로 딱 맞아 떨어지는 경우가 거의 없음.
- 이 때 필요한 것이 데이터를 분석 목적, 기법에 맞게 자유자재 변형하여 재구조화 하는 일.
- 엑셀에서는 Pivot table을 생성하는 것이 예제.
- Melt() : 데이터를 원하는 방식대로 분석하기 위해 식별(ID), 측정 변수(value) 형태로 자료를 재구성하는 단계.
- Cast() : melt()로 녹인 데이터들을 원하는 형태로 계산, 라벨링 하는 함수.
-

Reshape



[R 분석과 프로그래밍] http://rfriend.tistory.com

Thank you.

빅데이터 기초 / 류영표 강사
youngpyoryu@dongguk.edu