6주차:
데이터 분석 방법/데이터 마이닝의 이해

ChulSoo Park

School of Computer Engineering & Information Technology

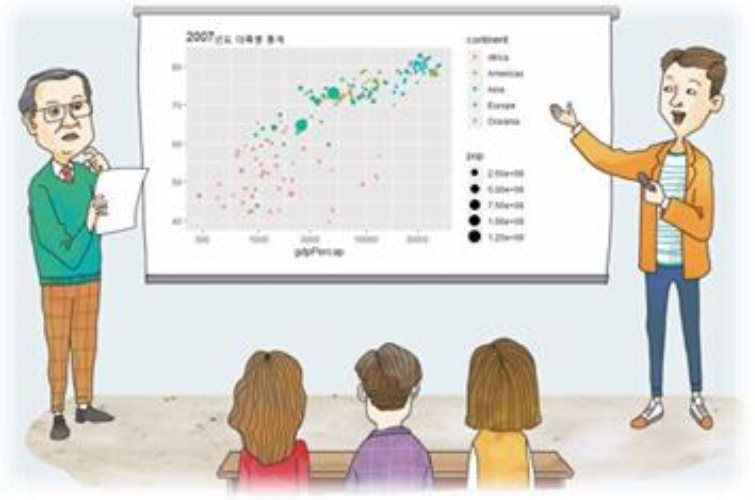
Korea National University of Transportation

학습목표 (6주차)

- 데이터 과학, 데이터 마이닝, 기계 학습 개념 이해
- 데이터 분석 유형의 이해
- 데이터 분석 모델 학습
- 데이터 분석 기법의 이해
- 데이터 분석 도구 파악
- 데이터 분석 사례 고찰

07 CHAPTER

데이터 분석 방법과 데이터 마이닝



데이터 사이언스 개론(김화종,홍릉과학출판사), 데이터 과학 입문(최대우외2명,한국방송통신대학교 출판부)

CONTENTS

Preview

- 데이터 과학은 다학제 분야
  - 컴퓨터 과학: 프로그래밍 언어, 현대적 시각화 기법, 기계 학습 등
  - 통계학: 요약 통계, 확률 통계 기법, 시각화 도구, 회귀 기법 등
  - 빅데이터: 분산 처리, 하둡 등
  - 데이터 마이닝 : 데이터로부터 유용한 지식을 찾아내는 과정
  - 이 책에서는 굳이 데이터 과학과 데이터 마이닝을 구분하지 않는다. (p.31)

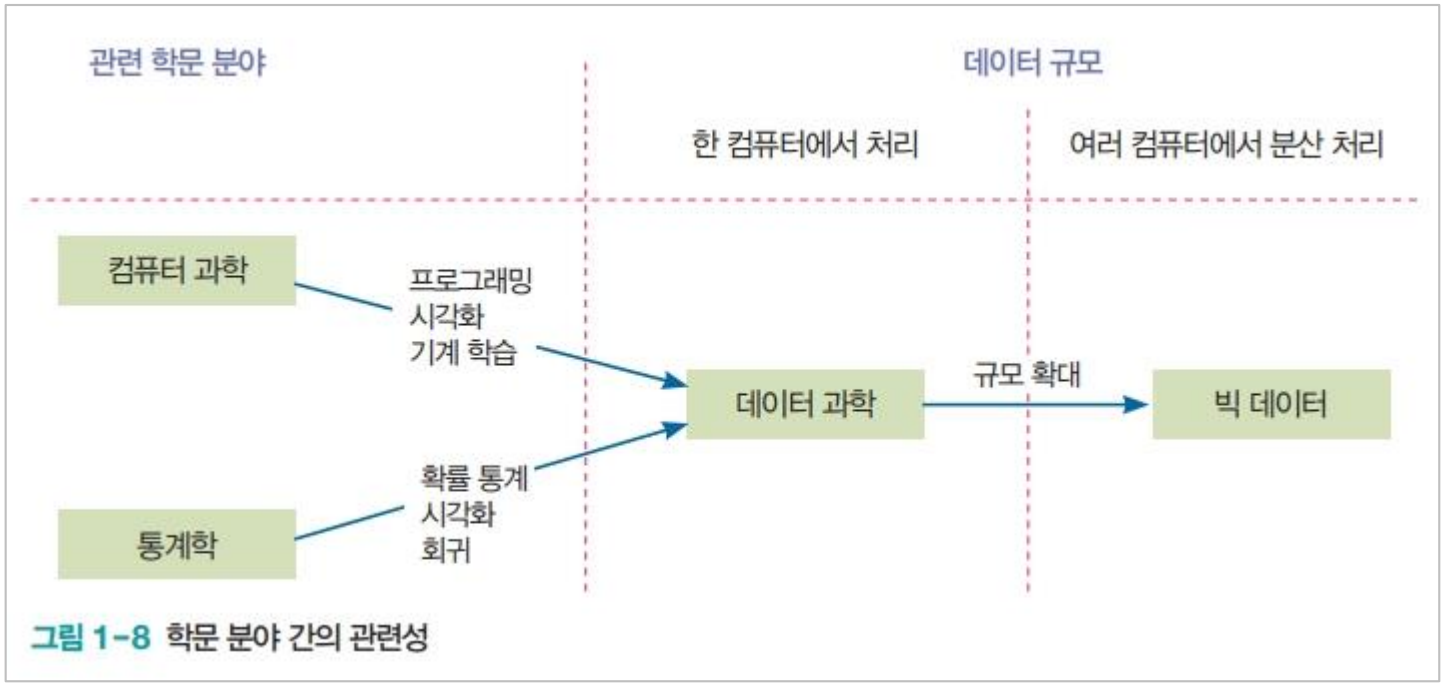


그림 1-8 학문 분야 간의 관련성

Preview

데이터 과학(data science)이란?

데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 학문 미래 지향적.

데이터 마이닝(data mining)이란?

대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정이다.

개론7.5 데이터 분석 도구

데이터 분석 도구

- 프로그래밍 언어: R, Python, Java, C/C++ 등
- 스프레드시트(엑셀)
- 통계처리 패키지: STATA, SAS, SPSS
- 수치해석 도구: MATLAB
- **데이터 분석 도구의 선택**
  - **무료 or 유료**(라이선스의 숫자)
  - 도구의 가용성(조직 내 공유 도구의 경우)
  - 도구의 기능, 데이터 내보내기, 호환성, 사용성 등
  - 도구 벤더사의 지원(교육 등)

개론7.5 데이터 분석 도구

R

- 원래 통계 처리를 주 목적으로 개발된 언어로 공개 소프트웨어로서 free
  - 전 세계의 개발자들이 수많은 패키지를 개발하고 공유
- 현재는 데이터의 특성 조사, 기계학습 알고리즘 구현, 데이터 시각화 등에 널리 사용됨
- 데이터를 다루는데 특화된 언어
- 문법은 매우 간단하나 통계의 기본 개념 숙지 필요
- 하나의 함수 호출로 많은 작업 수행하므로 편리하나, 함수의 구현 내용에 대한 이해가 필요함
- 데이터를 보다 상세하게 다루거나 복잡한 기능을 구현하려면 범용 프로그래밍 언어를 사용해야 함

개론7.5 데이터 분석 도구

파이썬(Python)

- 범용 프로그래밍 언어로 데이터 분석에 필요한 상세한 함수들을 제공
- 세밀한 동작을 수행할 수 있고, 또한 스크립트형 언어로써 일부 영역만 실행하여 결과를 살펴볼 수 있는 장점이 있음
- 쉽게 알고리즘을 개발하고 통합할 수 있음(분석 아이디어 검증)
- 파이썬은 Java, C 보다는 속도가 느림
- NumPy(데이터 컨테이너), Pandas(데이터 처리 함수), Matplotlib(데이터 시각화), IPython(계산용 라이브러리), SciPy(계산 컴퓨팅 영역에 대한 라이브러리) 등 매우 편리한 라이브러리를 제공

개론7.5 데이터 분석 도구

SAS

- SAS(Statistical Analysis System)는 기업이나 대형 기관에서 널리 사용되는 통계 분석 도구
- 데이터베이스의 자료를 쉽게 접근할 수 있으며 자료의 정렬, 결합, 분류, 수정 등 자료 관리가 용이
- 의사 결정, 모델 훈련 및 예측 등의 데이터 분석 가능, 그래프와 통계표의 작성도 용이
- 대표 솔루션으로 Enterprise Guide(EG)와 Enterprise Miner(Eminer) 제공
- 초보 사용자를 위한 편의성과 전문가를 위한 알고리즘 개발 및 분석 도구 제공

개론7.5 데이터 분석 도구

SPSS

- SPSS(Statistical Package for the Social Sciences)는 가장 많이 사용되는 통계 처리 패키지의 하나
- 주요 4가지 기능:
  - 데이터의 입력
  - 데이터의 관리
  - 통계적 분석
  - 보고서 작성
- 두 가지 수행 방식:
  - 일반인을 위한 대화상자/메뉴 기반의 GUI 방식
  - 전문가를 위한 Syntax 방식(직접 명령문 스크립트 작성)

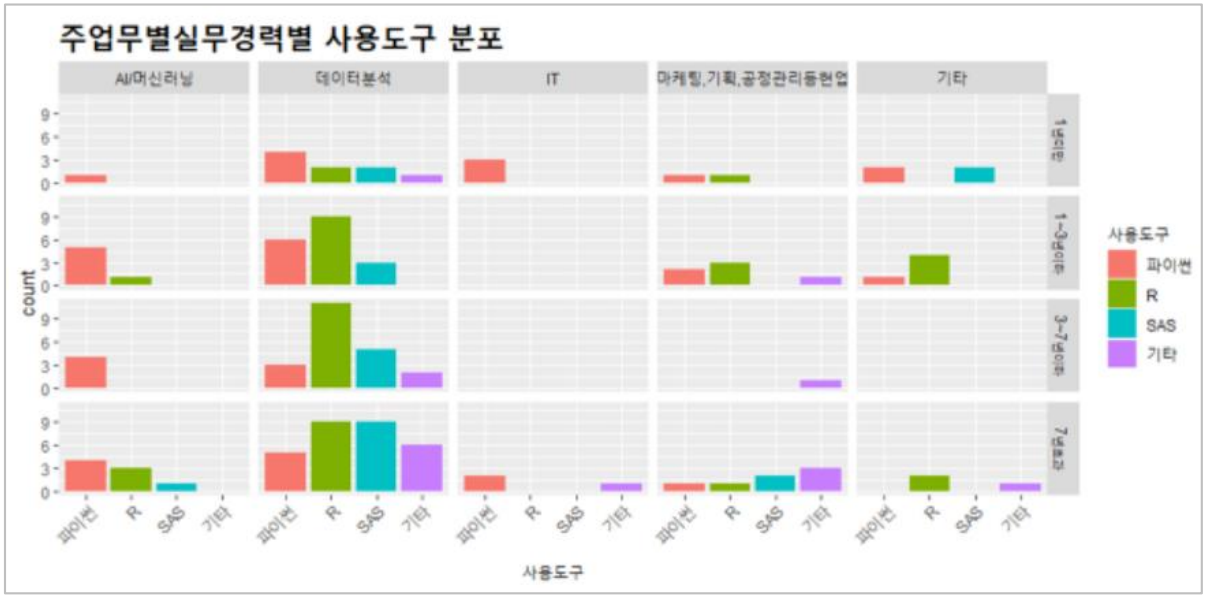
개론7.5 데이터 분석 도구

MATLAB

- MatLab(Matrix Laboratory)은 수치 해석, 행렬 연산, 신호 처리의 수치 계산을 편리하게 수행
- 다양한 그래픽 기능 제공
- 수치 데이터를 다루기가 편리하여 공학분야에서 널리 사용됨
- 다양한 형식의 데이터 가져오기 가능
- 데이터 관리, 필터링, 사전 처리 가능
- 자료 분석을 통하여 가설을 테스트하는 모델 제작 가능
- 데이터 분석 결과 시각화 라이브러리 제공

개론7.5 데이터 분석 도구

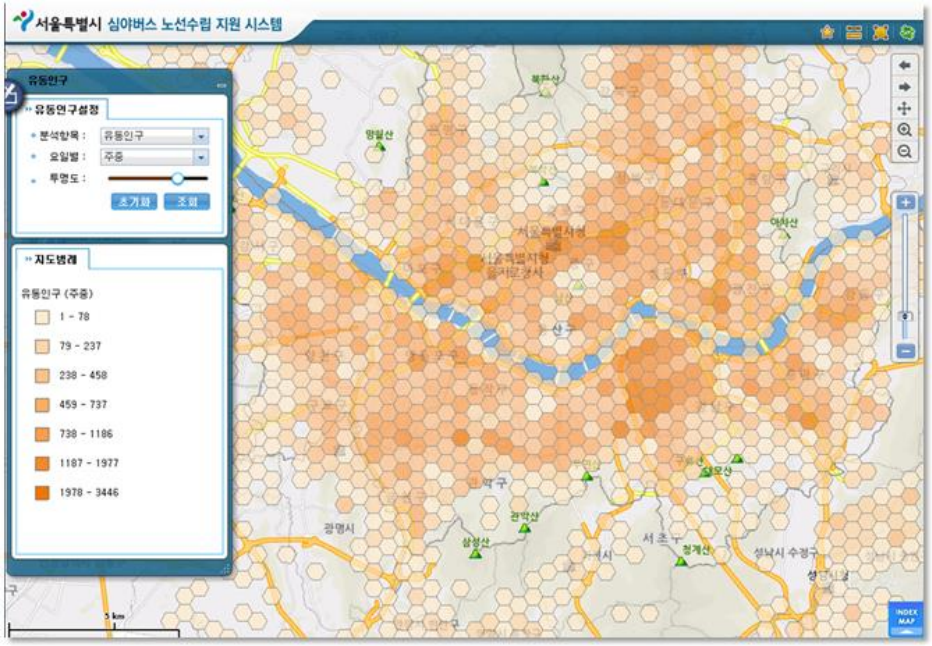
주로 사용하는 도구는 R과 Python (조사 표본이 치우쳐져있을 가능성 존재) 단, 데이터 분석에서 두 도구의 사용이 많은 것이 최근(2019년 기준)의 추세



입문7.1 데이터 과학에서 데이터 마이닝의 역할

데이터 과학에서 분석 과정은 전략적 통찰력을 창출하는 핵심

- 데이터에 내제된 스토리와 가치를 도출하기 위한 데이터 과학 분석 기법
- 시각화 기법 : 데이터의 성향 및 변수들의 연관성을 시각적으로 도출
- 기초통계 분석 : 평균, 상관관계 등
- Reporting
- OLAP(On-line Analytical Processing)
- 데이터베이스 조회



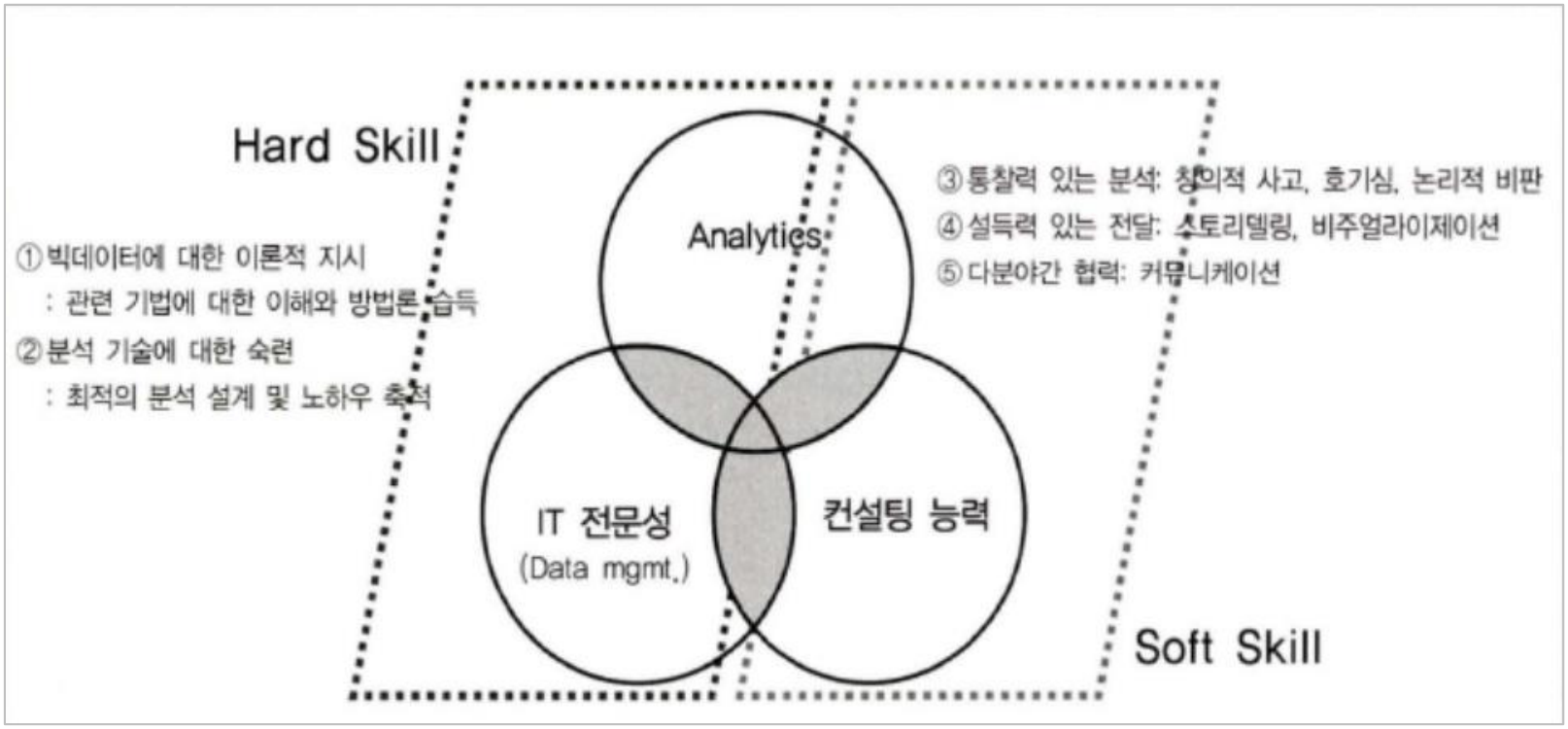
입문7.1 데이터 과학에서 데이터 마이닝의 역할

데이터 과학자와 전통적인 자료분석가의 차이점(p,209)

- 기존의 분석 접근법은 대규모의 data 이용이 가능한 빅데이터 시대에 사용하기에는 한계가 있다.(data 량,data 유형 등)
- 대규모의 자료 분석이 필요한 정보통신, 금융정보, 경영정보,의료정보, 생명정보, 사회관계망 등의 데이터베이스에는 수천 개의 변수가 존재함.
- 다차원 혹은 고차원 데이터를 분석하려면 전통적인 통계처리로는 해결할 수 없는 여러가지 문제점이 야기 될 수 있음
- 데이터 과학자는 데이터마이닝에서 제공하는 데이터로부터 흥미로운 패턴을 발견하고 미래를 예측하는 작업으로 마이닝 기법을 정확히 숙지하고 적절히 활용할 수 있는 능력을 함양하여야 함.
- 데이터 과학자와 전통적인 자료분석가의 차이점은, 자료분석가는 기술적인 자료처리 및 통계, 데이터마이닝 기법 등의 하드 스킬(hard skill)을 활용하는 반면, 최근에 요구되는 **데이터 과학자의 차별화된 역량**은 하드 시킬을 넘어 사고방식, 비즈니스 이슈에 대한 감각, 고객들에 대한 공감 능력, 인문학적 감성 도출 등 전략적 통찰과 소프트 스킬(soft skill)을 활용하는 것이다.

입문7.1 데이터 과학에서 데이터 마이닝의 역할

- 데이터 과학자와 전통적인 자료분석가의 차이점(p,209)



입문7.2 데이터 마이닝의 개념

데이터 마이닝의 정의, 특징, 데이터 마이닝의 과정

데이터마이닝의 정의

- 데이터마이닝은 다량의 가공되지 않은 데이터로부터 가치 있는 정보 혹은 지식을 축출하는 과정
- 방대한 데이터를 정제하여 통계 및 수학적 기술, 패턴 인식 기술 등을 사용하여 의미 있는 연관성, 패턴 그리고 추세를 발견하는 총칭
- 이상의 내용과 여러 정의를 종합하면 데이터마이닝이란 패턴인식 기술 뿐만 아니라 통계적· 수학적 기법을 이용하여 다양한 형태로 저장된 거대한 데이터로부터 우리에게 유용하고 흥미 있는 새로운 관계, 성향, 패턴 등 다양하고 가치 있는 정보를 찾아내는 일련의 과정

입문7.2 데이터 마이닝의 개념

데이터 마이닝의 정의, 특징, 데이터 마이닝의 과정

- 데이터마이닝의 특징(p.213)

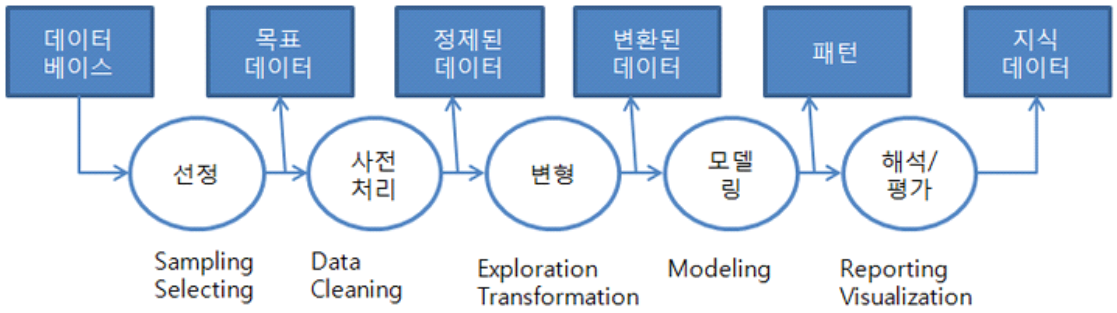
특징	비고
대용량의 관측 가능한 자료	▪ 시간의 흐름에 따른 축적 ▪ 데이터 분석을 업무에 두지 않은 경우가 많음
컴퓨터 집약적 기법 (computer-intensive method)	▪ 컴퓨터의 강력한 처리 속도와 능력 활용 ▪ 기존 분석 기법의 한계 극복
경험적 방법 (adhockery method)	▪ 경험에 기초하여 기법 개발 ▪ 수리적 특성이 규명 되지 않은 기법도 존재
일반화 (generalization)	▪ 새로운 데이터에 얼마나 잘 적용되는지가 성공적인 데이터 마이닝 기법의 판단 기준
업무활용성 (business applications)	▪ 다양한 경영 상황에서 경쟁력 확보를 위한 의사결정을 지원

입문7.2 데이터 마이닝의 개념

■ 데이터 마이닝의 정의, 특징, 데이터 마이닝의 과정

▪ 데이터마이닝의 절차(p.214)

- 다음 그림은 데이터마이닝의 수행 단계임
- 각 단계들은 상호 배타적으로 이루어지지 않으며, 한 방향으로 적용되기 보다는 상호 보완적으로 반복수행된다.



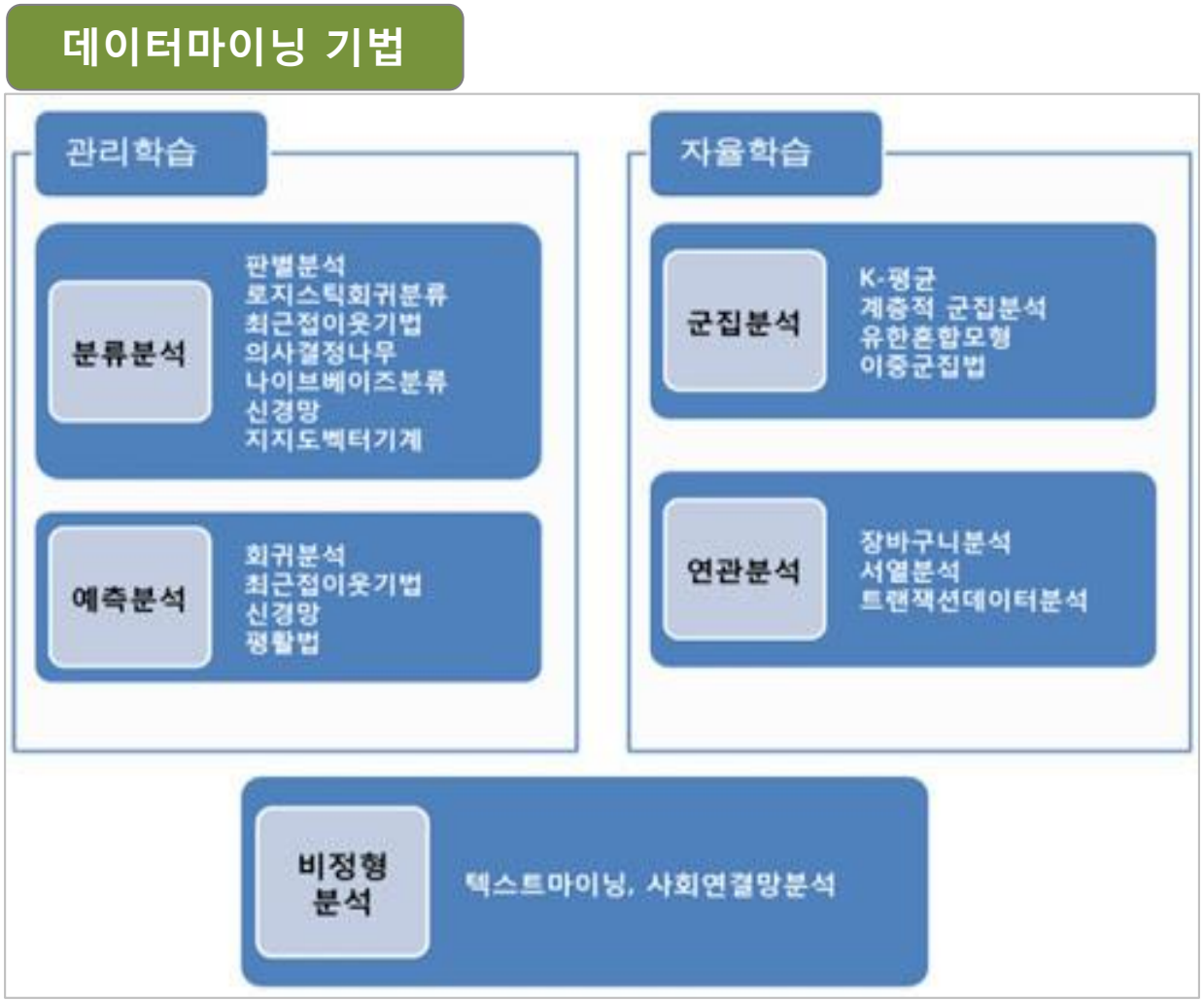
입문7.3 데이터 마이닝 관련 분야

- 데이터마이닝은 거의 모든 기업과 다양한 업종에서 의사 결정 문제에 많이 활용 되는데 주요 활용 분야는 다음과 같다.

분야	사례
금융	카드 현금서비스, 연체 등의 데이터를 패턴화 하여 우량/불량 고객 분석, 신용도를 분석
유통	매장 진열 위치에 따른 판매량을 조사하고 분석 하여 매장 진열 전략을 수립 하거나, 광고 시간대 와 구매율 분석을 통해 광고효과 분석
마케팅	고객 소비패턴분석으로 신규고객발굴, 고객 세분화, 선호도분석이 가능
통신	통화상세내역분석, 고객 충성도 분석 및 스팸 이나 부정이용 등을 발견
치안	범인의 행동패턴이나 심리분석으로 범인 추적
의료	환자의 특성에 따른 의약품의 효과, 유전자 패턴에 따른 질병 관계를 파악하여 신약이나 새로운 치료법 개발 가능
제조	생산 시기나 공급량이 판매에 미치는 영향 등을 분석하여 제품 출하 제품별 수량과 시기를 조절 가능

입문7.4 데이터마이닝 기법 및 도구

- 다음은 데이터마이닝을 사용 목적에 따라 분류한 것이며 각 기법에 대해서는 별도 학습 필요(p.220~221)



입문7.4 데이터마이닝 기법 및 도구

- 마이닝 기법

기술	설명
연관성 분석 Association Analysis	- 여러 개의 트랜잭션들 중에서 동시에 발생하는 트랜잭션의 **연관관계를 발견**하는 것 [사례] 넥타이를 구매한 고객이 셔츠를 구매할 확률은 50% 이다.
연속성 규칙 Sequence	- 개인별 트랜잭션 이력 데이터를 시계열 적으로 분석, 트랜잭션의 **향 후 발생 가능성** 예측 [사례] A 품목을 구입한 회원이 향후 H 품목을 구입할 가능성은 75% 이다. A품목을 고객이 구매 했으면 이 고객은 나중에 H 품목을 구입할 것이다.
분류 규칙 Classification	- 이미 알려진 특정 **그룹의 특징을 부여**하고 정의된 분류에 맞게 구분 - 분류 규칙의 형태 표기는 의사결정 트리, 신경망 형식으로 표현 [사례] 신용카드 신규 가입자를 낮음/중간/높음 신용 위험 집단으로 구분
데이터 군집화 clustering	- 상호간에 **유사한 특성**을 갖는 데이터들을 **집단화** 하는 과정임 [사례] A~D 의 데이터를 집단화 하는 과정에서 고객 군집 별 특성을 파악함 A 군집은 소득이 300만원 이상이고, 자녀가 2~3 명 이고 평균 연령이 30대
특성화 Characterization	- 데이터 집합의 일반적인 특성을 분석 하는 것으로 데이터의 요약 과정을 통하여 **특성 규칙을 발견**하는 기법
신경망 Neutral Net	- 인간 신경계를 모방한 개념으로 **반복학습** 과정을 통한 분석 [사례] 지난 사례의 패턴을 추출한 카드 연체자 추측, 주가 예측 등
사회연결망분석 social network Analysis	개인과 집단들 간의 관계를 모델링하여 그것의 위상 구조와 확산 진화 과정을 계량적으로 분석하는 데이터마이닝 방법론

입문7.4 데이터마이닝 기법 및 도구

- 데이터마이닝 도구(p.225)

구분	SAS	SPSS	R
프로그램 비용	유료, 고가	유료, 고가	오픈 소스
설치 용량	대용량	대용량	모듈화로 간단
다양한 툴 지원 및 비용	별도 구매	별도 구매	비용 없음
최근 알고리즘 및 기술 반영	느림	다소 느림	매우 빠름
학습자료 구입의 편의성	유료 도서 위주	유료 도서 위주	공개 논문 및 자료 많음

출처 : ECG Analysis

입문7.5 데이터마이닝 적용 사례

1. 신용카드사의 부정 사용자 적발을 위한 데이터마이닝

- 분실 혹은 도난 : 분실이나 도난에 의해 타인이 카드를 도용하는 경우
- 배달사고 : 배달과정에서 타인이 카드를 받아 사용하는 경우
- 허위신청 : 카드신청 단계부터 서류 등을 허위로 작성하여 카드를 발급받는 행위
- 카드위조 : 신용카드 뒷면의 마그네틱 부분에 정상적으로 발급된 신용카등의 정보를 입력하는 경우
- 주변인의 사기
- 불법 현금 융통 : 현금 확보를 위해 실제 구입하지 않은 물건을 구매한 것으로 위장하는 행위

입문7.5 데이터마이닝 적용 사례

2. 이동통신사 고객 이탈방지를 위한 데이터마이닝

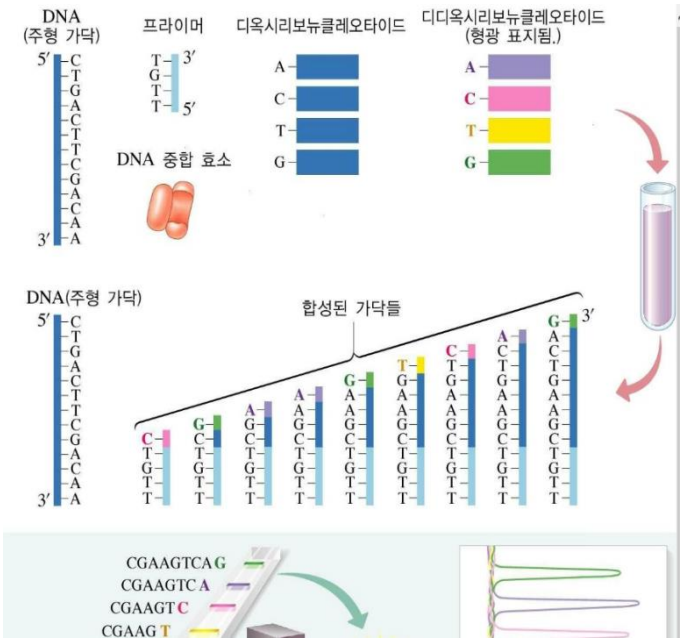
- 고객관계관리(CRM)
- 고객이탈방지
  - 분석 주제 정의 및 방향 설
  - 파일럿 분석의 점검
  - 변수 생성 및 자료 탐색
  - 분류 예측 모형 설정

- 데이터마이닝을 이용한 이동통신사 고객 이탈 등급 예측 모형 연구
  http://share.ewha.ac.kr/content/?p=000000069639
- 신한카드, 빅데이터 분석으로 '고객 이탈 예측 적중률 68.4%'
  http://www.ittoday.co.kr/news/articleView.html?idxno=49405
- 데이터마이닝으로 게임 이탈 예측하기
  http://www.kocca.kr/knowledge/abroad/indu/__icsFiles/afieldfile/2012/11/05/PredictingChurnData-MiningYourGame.pdf

입문7.5 데이터마이닝 적용 사례

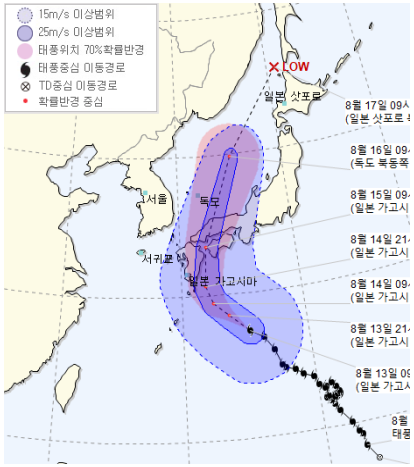
3. DNA 칩 자료분석에서의 데이터마이닝

- 칩의 원리와 활용
- 데이터마이닝 분석 적용



입문7.5 데이터마이닝 적용 사례

선택항목[총4개선택]			종합검진: [A선택]1항목+[B선택]3항목=총4개항목선택 또는 백신특화: [B선택]1항목 + [C선택]1항목= 총2개
구 분	검 사 항 목	관 련 질 환	종합검진
MR (자기공명영상)	뇌 MRI	뇌신경계질환	A항목 9종검사 中 선택1
	뇌 MRA	뇌신경계질환	
	요추 MRI	후경부 등증,상지등증 등	
	경추 MRI	요통, 좌골신경통, 대퇴신경 등	
DNA 유전자 검사	여성암 5종	여성 5종암 유전자 분석 [유방암, 위암, 갑상선암, 대장암, 폐암]	
	남성암 5종	남성 5종암 유전자 분석 [전립선암, 위암, 간암, 대장암, 폐암]	
	일반질환 5종	치매, 당뇨, 심혈관질환, 뇌졸중, 파킨슨병	
	피부유전자 검사	피부의 색소침착,탄력,노화,비타민C 농도 분석	
	탈모유전자 검사	탈모와 관련된 질환의 발생 가능성 예측	



2011년 애플 창업자 스티브 잡스가 자신이 앓고 있는 췌장암의 원인을 밝히기 위해 유전체 분석을 실시했다. 당시만 해도 개인의 유전체를 분석하는 데 드는 비용은 10만달러, 우리 돈으로 1억원이 넘었다. 여러 시장조사기관에 따르면 지놈 산업 규모는 2020년 20조원 수준이 될 것으로 예상된다. 특히 **인공지능, 빅데이터 기술이 융합되면서** 유전체 분석 기술은 빠르게 상용화되고 있다.

요약

1. 데이터 분석 방법
2. 데이터 분석 절차
3. 데이터마이닝

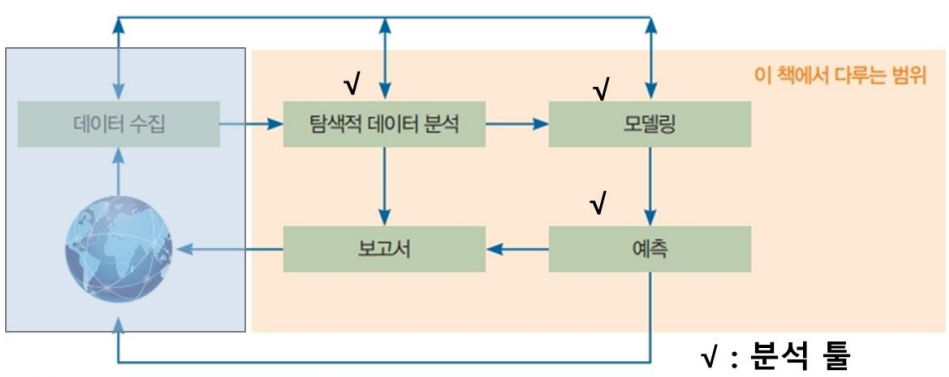
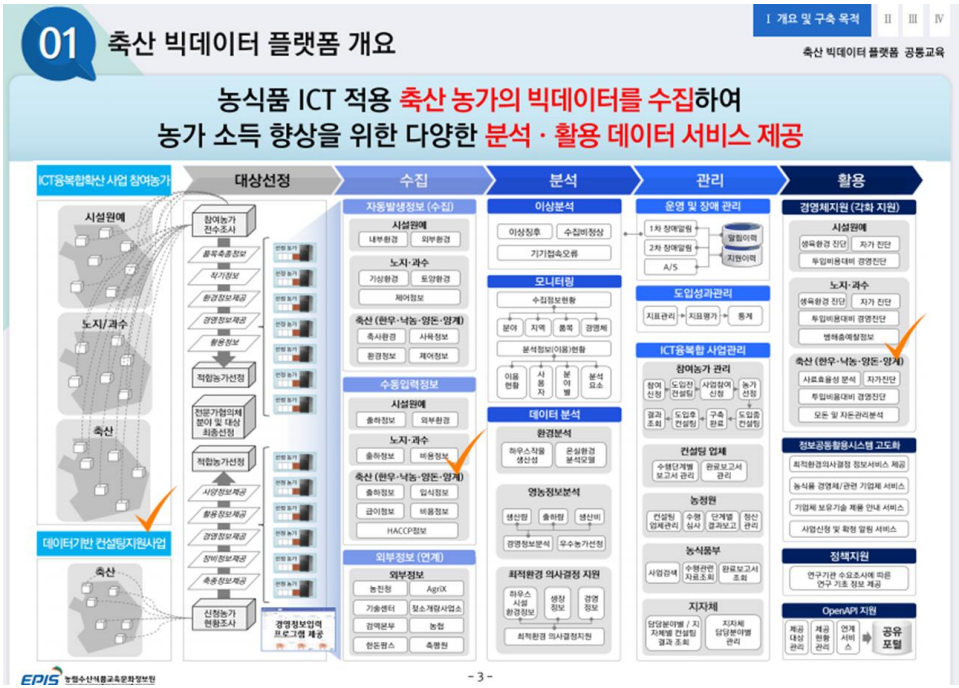


그림 1-7 세상과 상호작용하는 데이터 과학



Thank you


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