문화 A0007

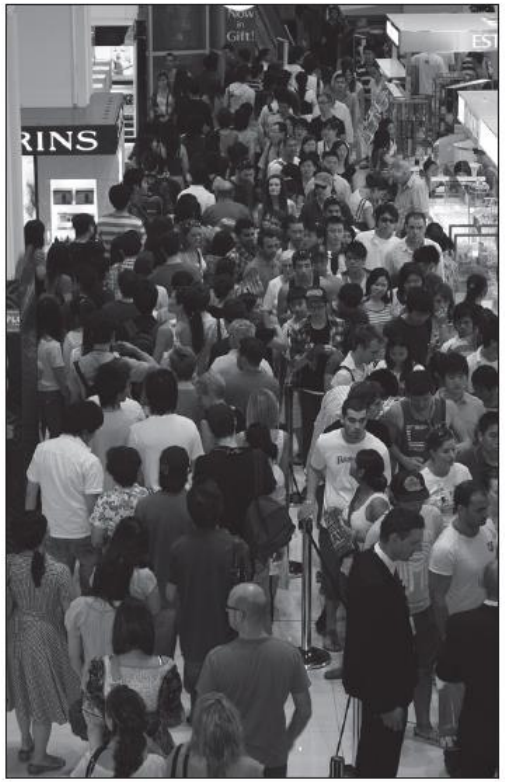
데이터사이언스입문

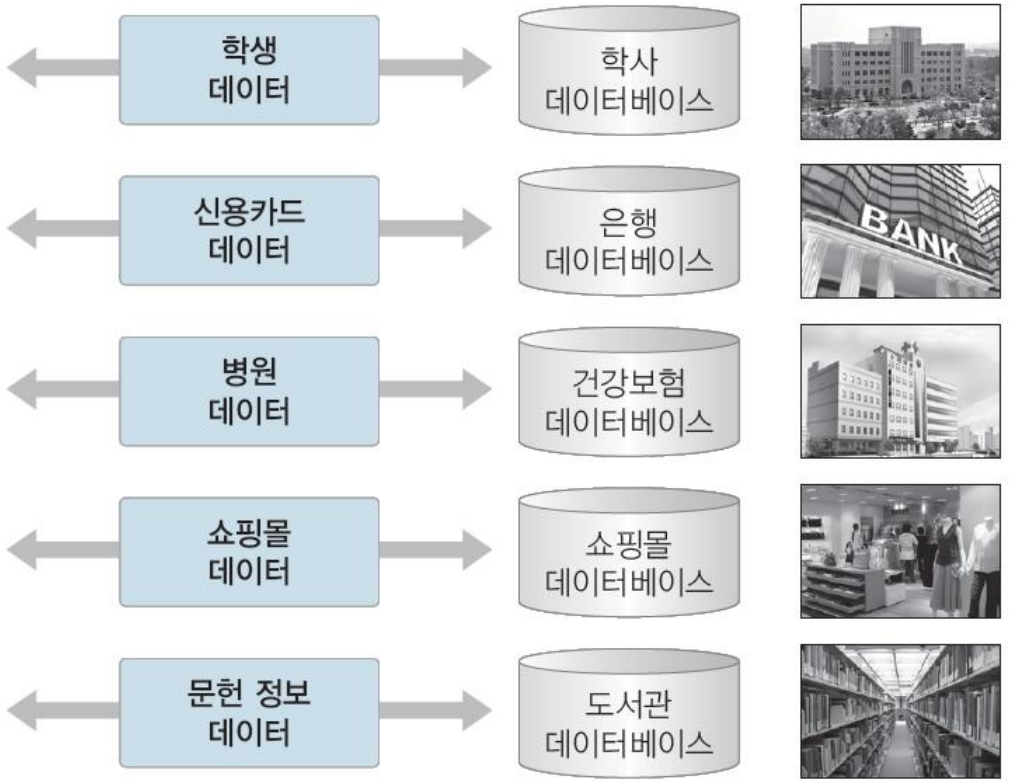
김 태 완

kimtwan21@dongduk.ac.kr

데이터베이스

- 데이터베이스
  - 조직에 필요한 정보를 얻기 위해 논리적으로 연관된 데이터를 모아 구조적으로 통합해 놓은 것
  - 사실 우리가 잘 인지 못하고 있지만 IT분야 뿐만 아니라 다른 분야에서도 보편적으로 사용
  - 우리의 삶이 데이터베이스와 직/간접적으로 연관되어 있다고 생각해도 무방





일상 생활의 데이터베이스

- 예시

CU Again

******* 최근영수증발행인쇄 *******
CU 강릉점
사업자등록번호:1751350895
강원도 강릉시 수촌로64 (지아동, 강릉아파트)
최 회 TEL:033-643-6434

정부 방침에 의해 12년 7월 1일부터 현금 결제 취소시, 영수증이 없으면 교환/환불이 불가합니다.

33177 2019-05-19(일) POS-01

달콤한대통과자1200	1	1,200
총 구 매 액	1	1,200
과세물품가액		1,091
부 가 세		109
*결 제 금 액		1,200
신 용 카 드		1,200

********* 신 용 카 드 *********
카드번호: 6161-07**-****-0303
카드회사: 001 비씨
할부개월: 00 승인번호: 43030033
결제금액: 1,200

*표시 상품은 부가세 면세 품목 임.
환불:30일내 영수증/카드지참시 가능
객층:14 담당:최 회 NO:6193 13:03

9511905193317709319 3

STARBUCKS
현금(소득공제)
사이렌오더

분당궁내DT점 T:1522-3232
경기 성남 대왕판교로 302
대표 : 송데이비드호섭 201-81-21515
[매장#3843, POS 03] 2022-02-20 08:13:25

김태완 (C-21)

BELT 샌드위치	5,900	1	5,900
ㄴ데우지 않음	0	1	0
ㄴ단가차액할인	-400	1	-400
합계	->		5,500
결제금액 (부가세포함)			5,500 (500)
결제			5,500
주문번호			32022022008121805633
신용카드			900
GIFTICON			4,600

현금영수증 발급 010-****-0915
승인금액 4,600
승인번호 177916305
현금영수증 문의 : 126-1-1

결제수단 변경은 구입하신 매장에서 가능하며, 반드시 구매 영수증과 원거래 결제수단을 지참하셔야 합니다.
(변경 가능 기간 : ~2022-03-06)

www.starbucks.co.kr

21210120200330834

KORAIL

열 차 승 차 권
Train Ticket

승차일자 YYMMDD 2019년 09월 22일 (금/Fri)

출발 From 송주 Yongsan
도착 To 용산 Yongsan

16:39 → 18:33

※승차일시와 이용구간을 반드시 확인하시기 바랍니다.

Train No. 4188 열차 KTX 산천

타는 곳 번호	호차번호	좌석번호
Tracks	Car No.	Seat No.
역 전광판 확인	17 호차	18A
	일반실	(순방향)

영수액₩ 42,100원

운임요금₩ 46,800원
할인금액₩ 4,700원
(부가세포함)
신용728720

발행일시 03/15 15:37

고객센터 전화번호 : 1544-7788
전화반환 전화번호 : 1544-8788

일상 생활의 데이터베이스

- 데이터베이스 활용 분야

분야	활용
생활과 문화	• 기상정보 : 날씨 관련 정보를 제공 • 교통정보 : 교통상황 관련 정보를 제공 • 문화예술정보 : 공연이나 인물에 관한 정보를 제공
비즈니스	• 금융정보 : 금융, 증권, 신용에 관한 정보를 제공 • 취업정보 : 노동부와 기업의 채용 관련 정보를 제공 • 부동산정보 : 공공기관이나 민간의 토지, 매물, 세금 정보를 제공
학술정보	• 연구학술정보 : 논문, 서적, 저작물에 관한 정보를 제공 • 특허정보 : 특허청의 정보를 기업과 연구자에게 제공 • 법률정보 : 법제처와 대법원의 법률에 관한 정보를 제공 • 통계정보 : 국가기관의 통계에 관한 정보를 제공

일상 생활의 데이터베이스

- 데이터베이스 시스템은 데이터의 검색과 변경 작업을 주로 수행함
- 변경이란 시간에 따라 변하는 데이터 값을 데이터베이스에 반영하기 위해 수행하는 삽입, 삭제, 수정 등의 작업을 말함

구축이 쉬움

구축이 어려움

유형	검색 빈도	변경 빈도	데이터베이스 예	특징
유형1	적다	적다	공룡 데이터베이스	• 검색이 많지 않아 데이터베이스를 구축할 필요 없음 • 보존가치가 있는 경우에 구축
유형2	많다	적다	도서 데이터베이스	• 사용자 수 보통 • 검색은 많지만 데이터에 대한 변경은 적음
유형3	적다	많다	비행기 예약 데이터베이스	• 예약 변경/취소 등 데이터 변경은 많지만 검색은 적음, 검색은 변경을 위하여 먼저 시도됨 • 실시간 검색 및 변경이 중요함
유형4	많다	많다	증권 데이터베이스	• 사용자 수 많음 • 검색도 많고 거래로 인한 변경도 많음

데이터베이스 개념

- 통합된 데이터(integrated data)
  - 데이터를 통합하는 개념으로, 각자 사용하던 데이터의 중복을 최소화하여 중복으로 인한 데이터 불일치 현상을 제거

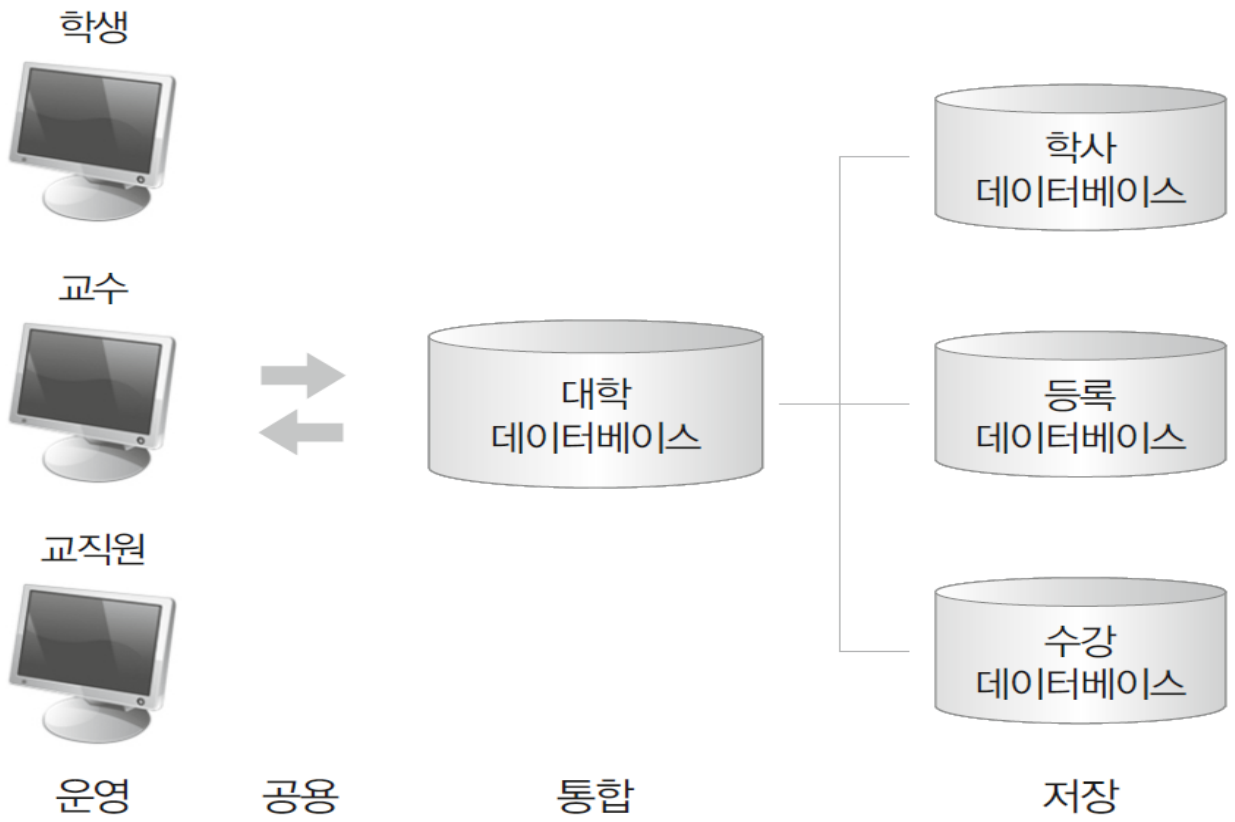
- 저장된 데이터(stored data)
  - 문서로 보관된 데이터가 아니라 디스크, 테이프 같은 컴퓨터 저장장치에 저장된 데이터를 의미

- 운영 데이터(operational data)
  - 조직의 목적을 위해 사용되는 데이터, 즉 업무를 위한 검색을 할 목적으로 저장된 데이터

- 공용 데이터(shared data)
  - 한 사람 또는 한 업무를 위해 사용되는 데이터가 아니라 공동으로 사용되는 데이터를 의미

데이터베이스 개념

- **데이터베이스는 운영 데이터를 통합하여 저장하며 공용으로 사용됨**



데이터베이스 특징

- 실시간 접근성 (real time accessibility)
  - 데이터베이스는 실시간으로 서비스
  - 사용자가 데이터를 요청하면 몇 시간이나 몇 일 뒤에 결과를 전송하는 것이 아니라 수 초 내에 결과를 서비스

- 계속적인 변화 (continuous change)
  - 데이터베이스에 저장된 내용은 어느 한 순간의 상태를 나타내지만, 데이터 값은 시간에 따라 항상 변화
  - 데이터베이스는 삽입, 삭제, 수정 등의 작업을 통하여 바뀐 데이터 값을 저장

- 동시 공유 (concurrent sharing)
  - 데이터베이스는 서로 다른 업무 또는 여러 사용자에게 동시에 공유
  - 동시는 병행 이라고도 하며, 데이터베이스에 접근하는 프로그램이 여러 개 있다는 의미

- 내용에 따른 참조 (reference by content)
  - 데이터베이스에 저장된 데이터는 데이터의 물리적인 위치가 아니라 데이터 값에 따라 참조

데이터베이스 시스템의 발전

- 예시 : 마당서점 데이터베이스
  - 1단계 마당서점 시작

- 도서 : 100권
- 고객 : 근처 학교의 학생, 지역 주민
- 업무 : 회계 업무(계산기 사용), 장부에 기록
- 고객 서비스 : 사장이 직접 도서 안내

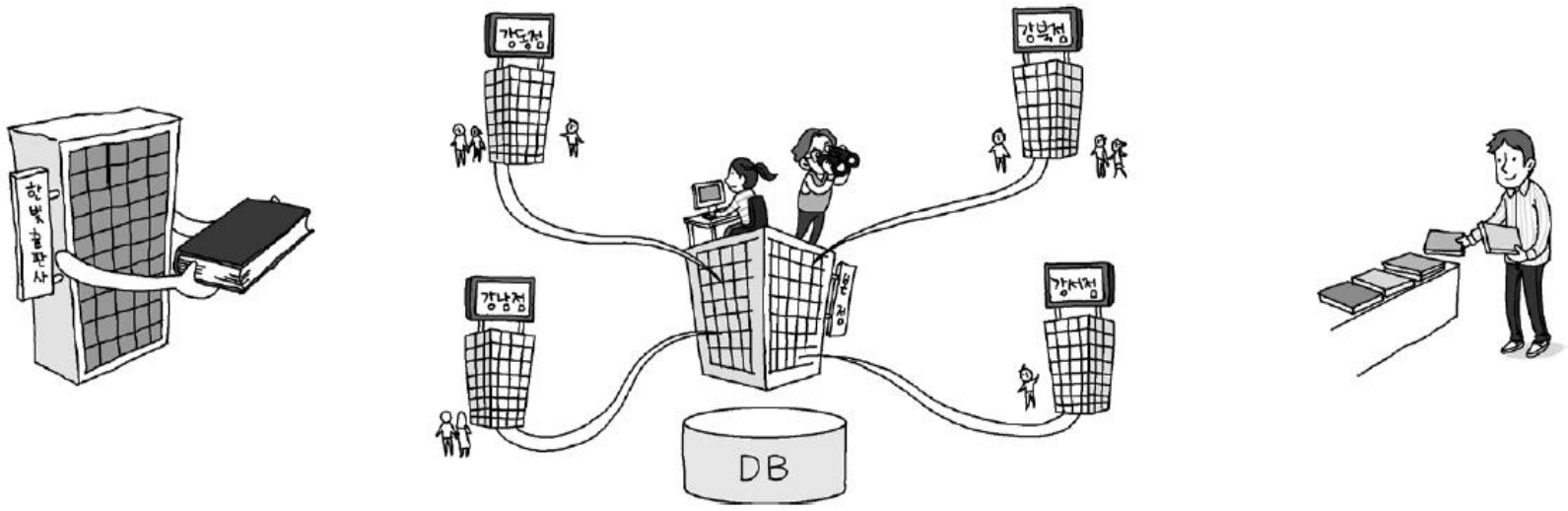
데이터베이스

- 예시 : 마당서점 데이터베이스
  - 2단계 컴퓨터 도입

- 도서 : 1,000권
- 고객 : 근처 학교의 학생, 지역 주민
- 업무 : 회계 업무(컴퓨터 사용), 파일 시스템
- 고객 서비스 : 컴퓨터를 이용하여 도서 검색, 직원 고용

데이터베이스

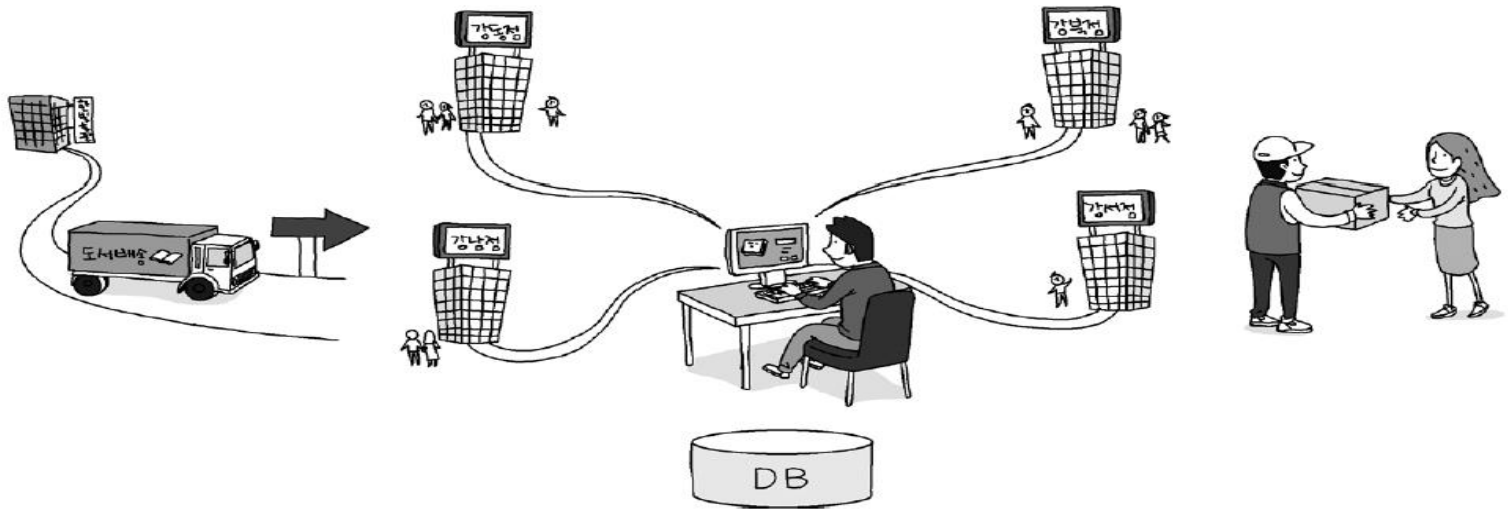
- 예시 : 마당서점 데이터베이스
  - 3단계 지점 개설 및 데이터베이스 구축



- 도서 : 10,000권
- 고객 : 서울 지역 고객
- 업무 : 회계 업무(컴퓨터 사용), 데이터베이스 시스템
- 고객 서비스 : 클라이언트/서버 시스템으로 지점을 연결하여 도서 검색 서비스 제공

데이터베이스

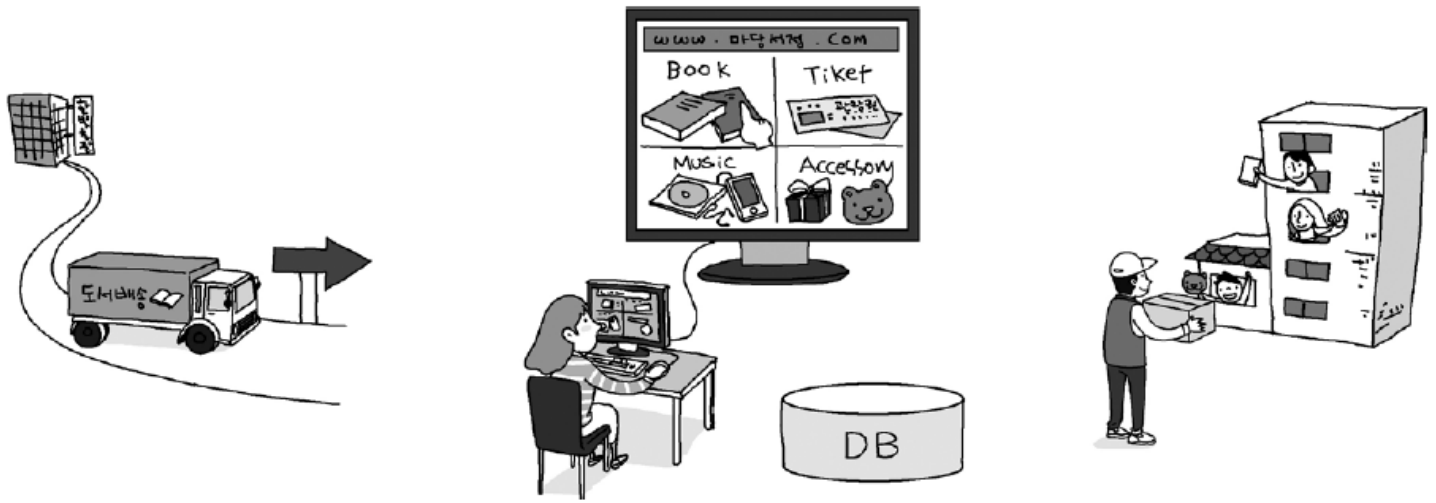
- 예시 : 마당서점 데이터베이스
  - 4단계 홈페이지 구축



- 도서 : 100,000권
- 고객 : 국민(전국으로 배송)
- 업무 : 회계/인사 업무(컴퓨터와 인터넷 사용), 웹 DB 시스템으로 지점 간 연계
- 고객 서비스 : 인터넷으로 도서 검색 및 주문

데이터베이스

- 예시 : 마당서점 데이터베이스
  - 5단계 인터넷 쇼핑몰 운영



- 도서 : 1,000,000권
- 고객 : 국민(전국으로 배송)
- 업무 : 회계/인사 업무(컴퓨터와 인터넷 사용), DB 서버 여러 개 구축
- 고객 서비스 : 인터넷 종합 쇼핑 서비스 제공

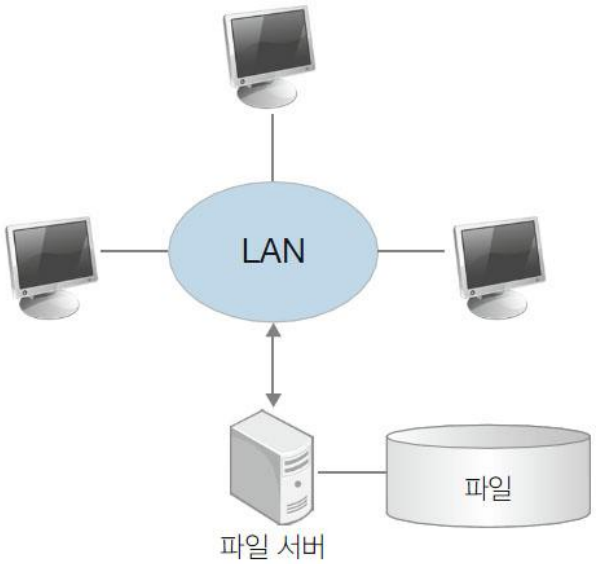
데이터베이스

- 예시 : 마당서점 데이터베이스

단계	시기 정보기술	주요 특징
1단계 마당서점	1970년대 컴퓨터 없음	• 사장이 모든 도서의 제목과 가격을 기억 • 매출과 판매가 컴퓨터 없이 관리됨 • 매출에 대한 내용이 정확하지 않음
2단계 초기전산화	1980년대 컴퓨터	• 컴퓨터를 이용한 초기 응용 프로그램으로 업무 처리 • 파일 시스템 사용 • 한 대의 컴퓨터에서만 판매 및 매출 관리
3단계 데이터베이스	1990년대 컴퓨터+원격통신	• 지점 간 클라이언트/서버 시스템을 도입하여 업무 처리 • 데이터베이스 관리 시스템(DBMS)을 도입
4단계 홈페이지 구축	2000년대 컴퓨터+인터넷	• 인터넷을 이용하여 도서 검색 및 주문 • 웹 DB 시스템으로 불특정 다수 고객 유치 • 고객이 지리적으로 넓게 분산됨
5단계 인터넷 쇼핑몰	2010년대 컴퓨터+인터넷	• 도서뿐만 아니라 음반, 액세서리, 문구, 공연 티켓까지 판매하는 인터넷 쇼핑몰로 확대 • 도서 외 상품의 매출 비중이 50% 이상으로 늘어남

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 파일시스템
    - 데이터를 파일로 관리하기 위해 파일을 생성·삭제·수정·검색하는 기능을 제공하는 소프트웨어
    - 응용 프로그램별로 필요한 데이터를 별도의 파일로 관리함
    - 각 컴퓨터는 LAN을 통해 파일 서버에 연결, 파일 서버에 저장된 데이터를 사용하기 위해 각 컴퓨터의 응용 프로그램에서 열기/닫기(open/close)를 요청
    - 각 응용 프로그램이 독립적으로 파일을 다루기 때문에 데이터가 중복 저장될 가능성이 있음
    - 동시에 파일을 다루기 때문에 데이터의 일관성이 훼손될 수 있음

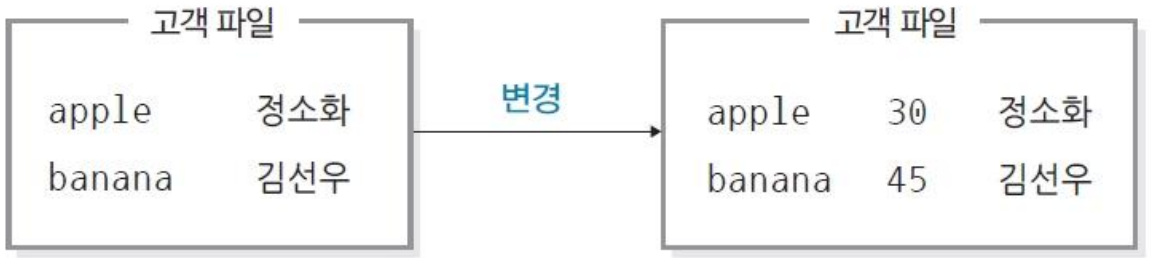


파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 파일시스템의 문제점
    - 같은 내용의 데이터가 여러 파일에 중복 저장
    - 응용 프로그램이 데이터 파일에 종속적
    - 데이터 파일에 대한 동시 공유, 보안, 회복 기능이 부족
    - 응용 프로그램을 개발하기 어려움

한 고객의 연락처를 고객 데이터 파일에서만 수정하고
실수로 주문 데이터 파일에서 수정하지 않았다면?

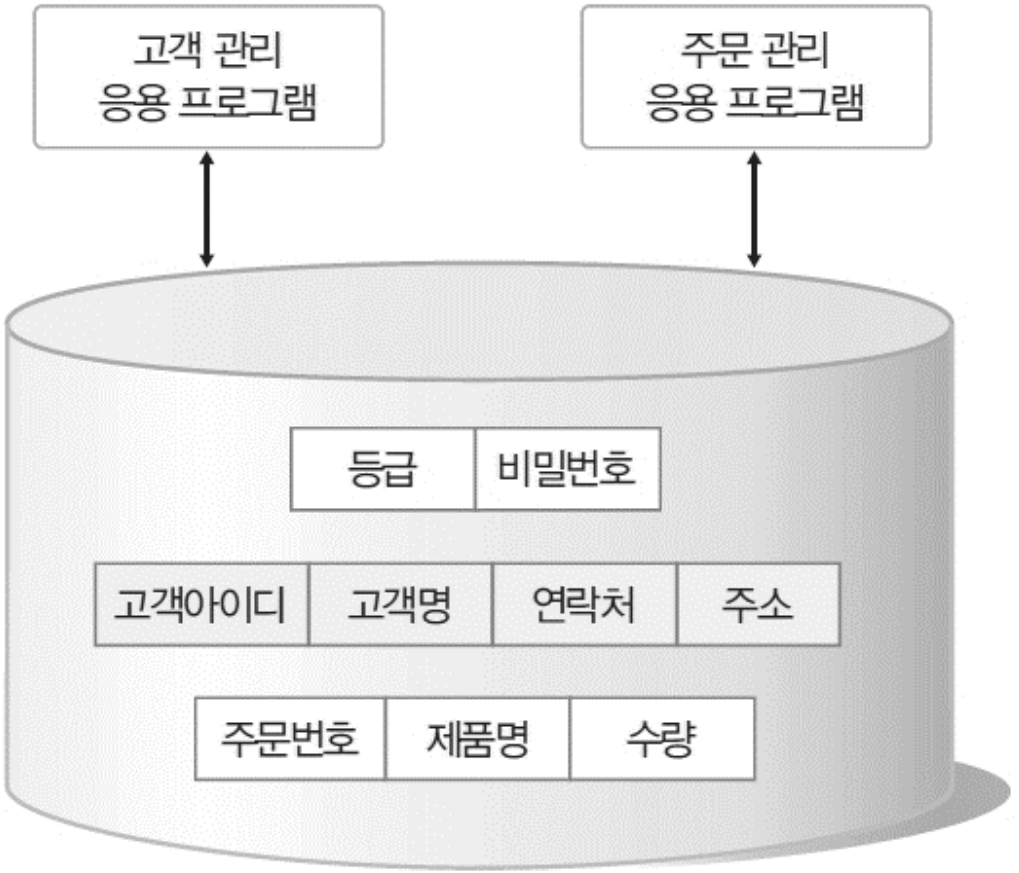
데이터 중복성



데이터 종속성

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 파일시스템의 문제점 해결 방안 : **데이터 통합**

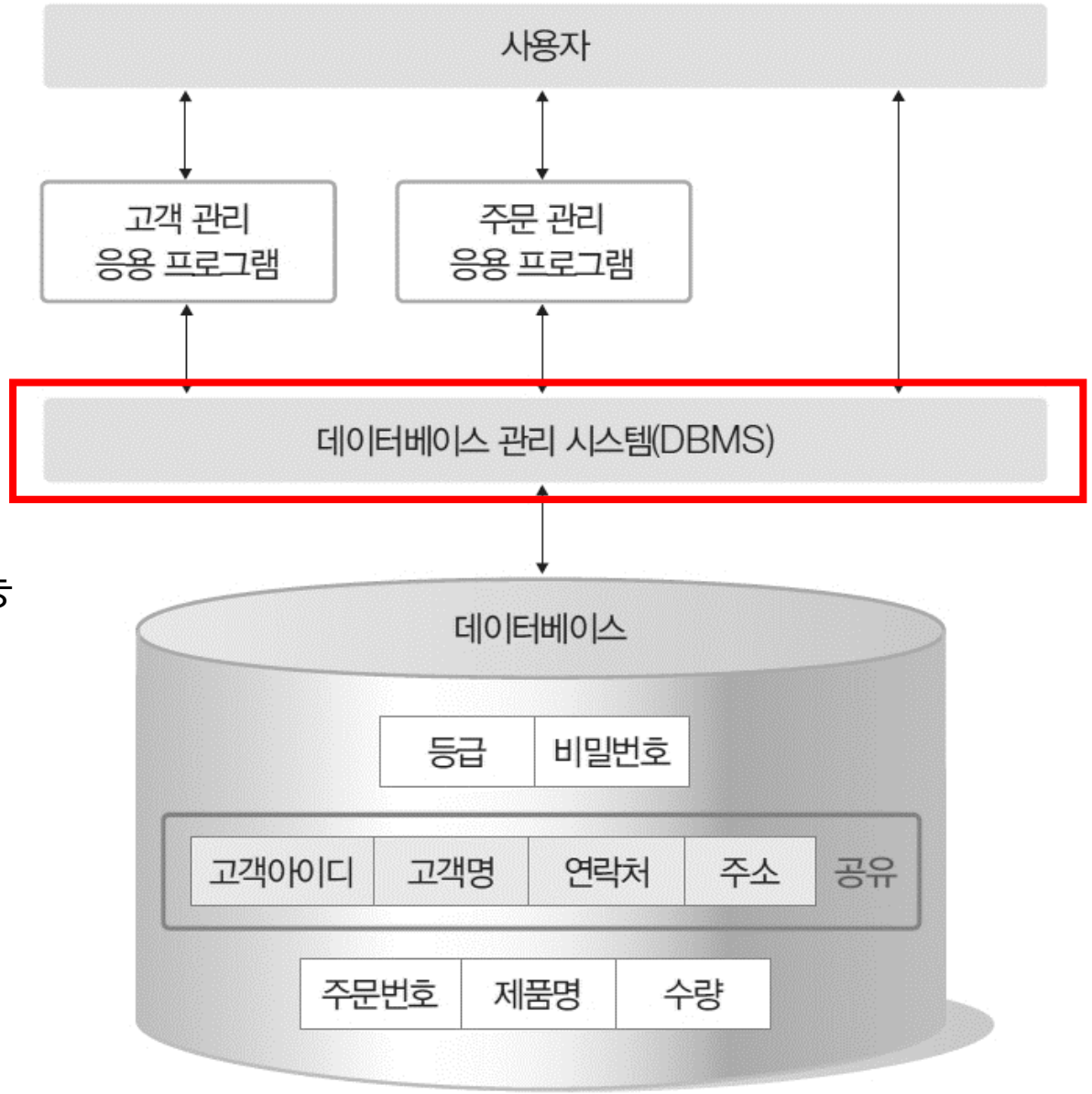


파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 데이터베이스 시스템
    - **DBMS (DataBase Management System)**
      - 파일 시스템의 문제를 해결하기 위해 제시된 소프트웨어
      - 데이터베이스를 관리·운영하는 소프트웨어
      - 사용자나 응용 프로그램은 DBMS가 관리하는 데이터에 동시에 접속하여 데이터를 공유함
    - DBMS를 도입하여 데이터를 통합 관리하는 시스템
    - DBMS가 설치되어 데이터를 가진 쪽을 서버(server), 외부에서 데이터 요청하는 쪽을 클라이언트(client)
    - DBMS 서버가 파일을 다루며 데이터의 일관성 유지, 복구, 동시 접근 제어 등의 기능을 수행
    - 데이터의 중복을 줄이고 데이터를 표준화 하며 무결성을 유지함

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 데이터베이스 시스템 주요 기능
    - **정의 기능**
      - 데이터베이스 구조를 정의하거나 수정가능
    - **조작 기능**
      - 데이터를 삽입, 삭제, 수정, 검색 가능
    - **제어 기능**
      - 데이터를 항상 정확하게 안전하게 유지 가능



파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 데이터베이스 시스템 장점
    - 데이터 중복 통제 가능
      - 데이터베이스에 데이터를 통합하여 관리하므로 데이터 중복 문제 해결
    - 데이터 독립성이 확보
      - 응용 프로그램 대신 데이터베이스에 접근하고 관리하는 모든 책임 담당

        → 응용 프로그램과 데이터베이스 사이에 독립성이 확보
    - 데이터를 동시 공유 가능
      - 동일한 데이터를 여러 응용 프로그램이 공유하여 동시 접근할 수 있게 지원

        → 동시 접근 제어 기술 보유
    - 데이터 보안 향상
      - 중앙 집중식으로 데이터를 관리하므로 효율적인 접근 제어 가능
      - 권한이 없는 사용자의 접근, 허용되지 않은 데이터와 연산에 대한 요청 차단 가능

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 데이터베이스 시스템 장점
    - 데이터 무결성을 유지
      - 데이터 삽입, 수정 등의 연산이 수행될 때마다 유효성을 검사하여 데이터 무결성(정확성)을 유지
    - 표준화
      - 데이터베이스 관리 시스템이 정한 표준화된 방식을 통해 데이터베이스에 접근
    - 장애 발생 시 회복 쉬움
      - 데이터 일관성과 무결성을 유지하면서 장애 발생 이전 상태로 데이터를 복구하는 회복 기능 지원
    - 응용 프로그램 개발 비용 감소
      - 데이터 관리 부담이 줄어 응용 프로그램 개발 비용 및 유지 보수 비용 감소

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템
  - 데이터베이스 시스템 단점
    - 고비용
      - 별도 구매 비용이 들고, 동시 사용이 허용되는 사용자 수에 따라 가격 증가
    - 백업과 회복 방법이 복잡
      - 장애 발생의 원인과 상태를 정확히 파악하기 어렵고 회복 방법도 복잡
    - 중앙 집중 관리로 인한 취약점이 존재
      - 데이터베이스나 데이터베이스 관리 시스템에 장애가 발생하면 전체 시스템의 업무 처리가 중단됨
      - 데이터베이스 의존도가 높은 시스템일수록 가용성과 신뢰성에 치명적임

파일 시스템과 데이터베이스 시스템

- 파일시스템과 데이터베이스 시스템

구분	파일 시스템	DBMS
데이터 중복	데이터를 파일 단위로 저장하므로 중복 가능	DBMS를 이용하여 데이터를 공유하기 때문에 중복 가능성 낮음
데이터 일관성	데이터의 중복 저장으로 일관성이 결여됨	중복 제거로 데이터의 일관성이 유지됨
데이터 독립성	데이터 정의와 프로그램의 독립성 유지 불가능	데이터 정의와 프로그램의 독립성 유지 가능
관리 기능	보통	데이터 복구, 보안, 동시성 제어, 데이터 관리 기능 등을 수행
프로그램 개발 생산성	나쁨	짧은 시간에 큰 프로그램을 개발할 수 있음
기타 장점	별도의 소프트웨어 설치가 필요 없음 (운영체제가지원)	데이터 무결성 유지, 데이터 표준 준수 용이

DBMS

- Database Management System (DBMS)
  - 데이터를 저장하고 유지보수 (수정, 삭제, 추가)하고 이를 검색하는 시스템
    - CRUD (Create, Retrieve, Update, Delete)
  - 대량의 데이터를 처리하는 시스템
  - 다양한 자료구조와 검색구조 (sorting, indexing, …)를 사용해 “빠른” 검색 가능
  - 대부분의 시스템은 R(검색) >>>>>> CUD (업데이트)의 빈도수가 많음
  - 검색에 최적화

DBMS – 정렬

- 정렬
  - 빠른 검색을 위해서는 데이터가 반드시 정렬 (Sorting)되어 있어야 함
  - 정렬되어 있지 않다면 (pile) 평균적으로 전체 데이터의 절반 필요
    - 최선 : 1, 최악 N, 평균 N/2
  - 정렬되어 있을 경우 데이터를 빠른 시간 안에 검색 가능
    - O(NlogN) ~ O(N^2)
  - 퀵정렬/힙정렬 계열이 주로 사용됨

DBMS – 인덱스

- 인덱스
  - 데이터 추가/수정/삭제할 때마다 정렬/인덱스 업데이트가 발생함
  - 인덱스 종류
    - 이진 검색 (Binary Search)
      - 최대 $\log_2 N$ 번 내에 검색 가능
      - 예시 : 데이터를 정렬 후 “TEST”단어 검색
        - 가운데 값을 확인 : “SAMPLE” → 뒤쪽 절반 / 뒤쪽 중 한 가운데 확인 : “VEIL” → 앞쪽 절반
        - 계속 반복해 “TEST” 단어 나올 때 까지 계속
        - 데이터 추가/삭제/변경 될 때마다 한가운데/왼쪽가운데/오른쪽가운데 값을 미리 계산 : 인덱스

      1,000개의 데이터가 있을 경우 10번만 찾으면 검색 가능 $2^N > 1000$
    - B-Tree 검색
      - 최대 $\log_3 N$ 번 내에 검색 가능
      - 상용 DBMS에서 가장 일반적으로 많이 사용됨
      - 이진 검색과 유사하지만 한 번에 비교를 2번 수행 (a, b : a < b)
      - 작은 값보다 작은 경우 (x < a) / 큰 값과 작은 값 사이 (a < x < b) / 큰 값보다 큰 경우 (b < x)
      - 데이터 추가/삭제/변경 될 때마다 a, b 업데이트 : 인덱스

DBMS

- DBMS 종류

문서 작성

표 계산

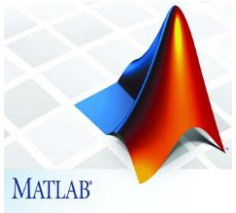


사진 편집

데이터베이스

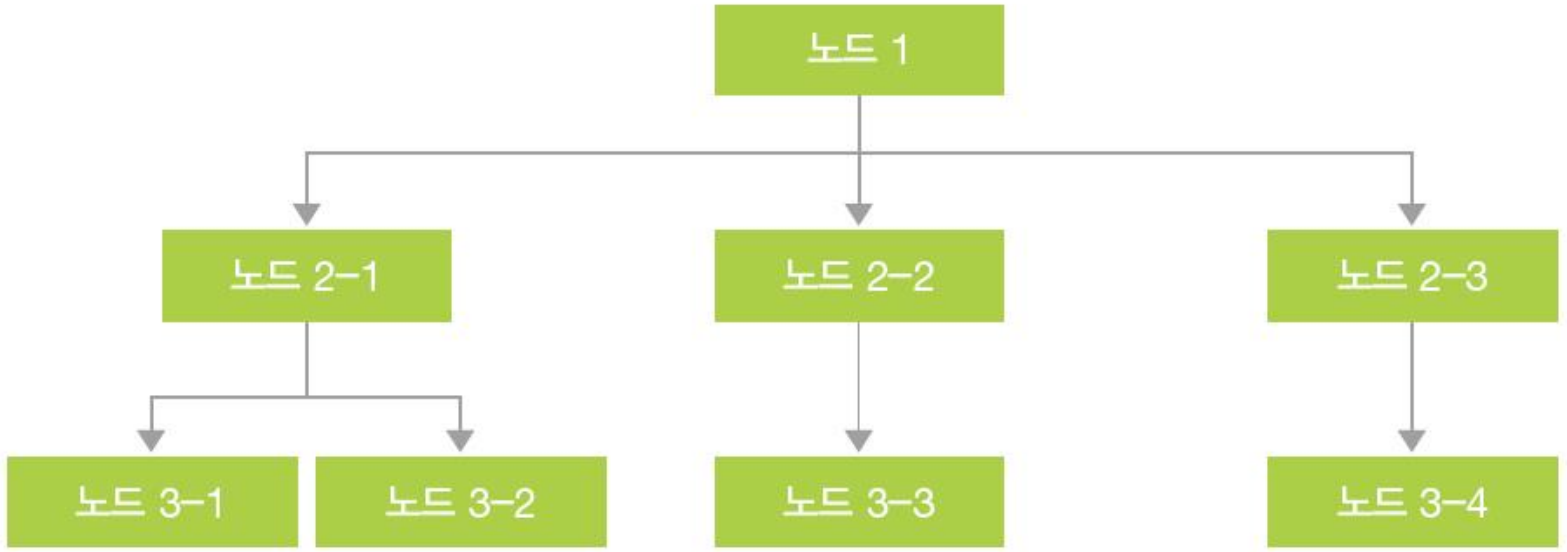
DBMS

- DBMS 종류

DBMS	제작사	운영체제	최신 버전	비고
MySQL	오라클	유닉스, 리눅스, 윈도우, 맥	8.0	오픈 소스(무료), 상용
MariaDB	마리아DB	유닉스, 리눅스, 윈도우	10.3	오픈 소스(무료)
PostgreSQL	PostgreSQL	유닉스, 리눅스, 윈도우, 맥	10.4	오픈 소스(무료)
Oracle	오라클	유닉스, 리눅스, 윈도우	18c	상용 시장 점유율 1위
SQL Server	마이크로소프트	리눅스, 윈도우	2017	
DB2	IBM	유닉스, 리눅스, 윈도우	10	메인프레임 시장 점유율 1위
Access	마이크로소프트	윈도우	2017	PC용
SQLite	SQLite	안드로이드, iOS	3.24	모바일 전용, 오픈 소스(무료)

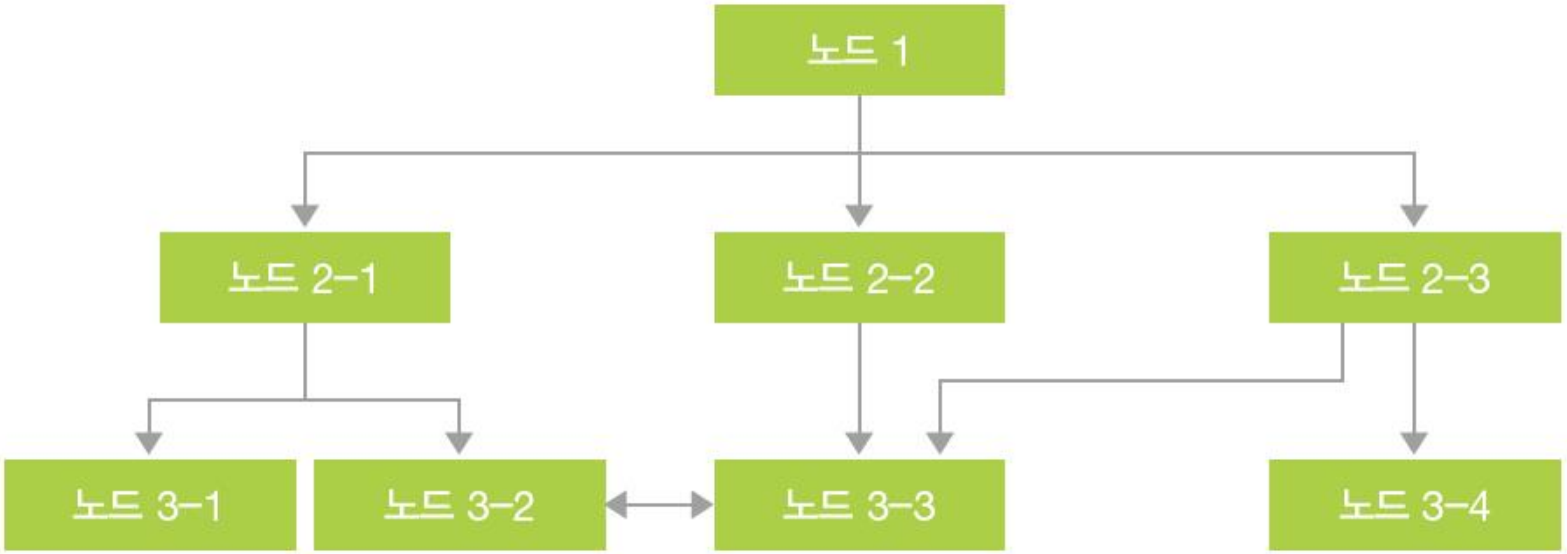
DBMS 분류

- 계층형 DBMS
  - 각 계층이 트리 형태를 띠고 1:N 관계를 가짐
  - 한번 구축하면 구조를 변경하기 까다로움
  - 접근의 유연성이 부족하여 임의 검색 시 어려움



DBMS 분류

- 망형 DBMS
  - 1:1, 1:N, N:M(다대다) 관계가 지원되어 효과적이고 빠른 데이터 추출이 가능
  - 매우 복잡한 내부 포인터 사용
  - 프로그래머가 모든 구조를 이해해야 만 프로그램을 작성할 수 있음

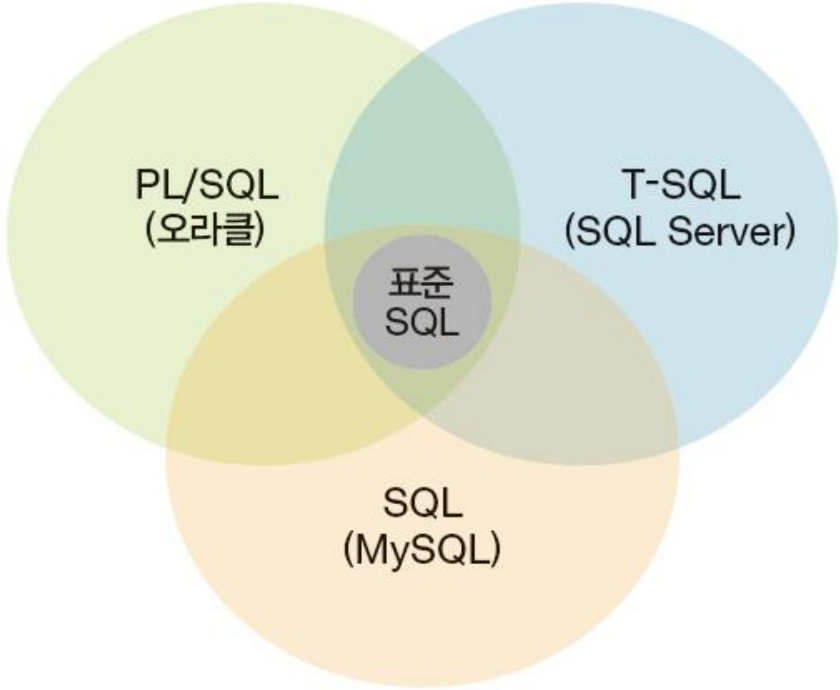


DBMS 분류

- 관계형 DBMS (RDBMS)
  - 모든 데이터는 **테이블**에 저장
  - 테이블 간의 관계는 기본키(PK)와 외래키(FK)를 사용하여 맺음(부모-자식 관계)
  - 다른 DBMS에 비해 업무 변화에 따라 바로 순응할 수 있고 유지·보수 측면에서도 편리
  - 대용량 데이터를 체계적으로 관리할 수 있음
  - 데이터의 무결성도 잘 보장됨
  - 시스템 자원을 많이 차지하여 시스템이 전반적으로 느려지는 단점이 있음

DBMS에서 사용되는 언어 : SQL

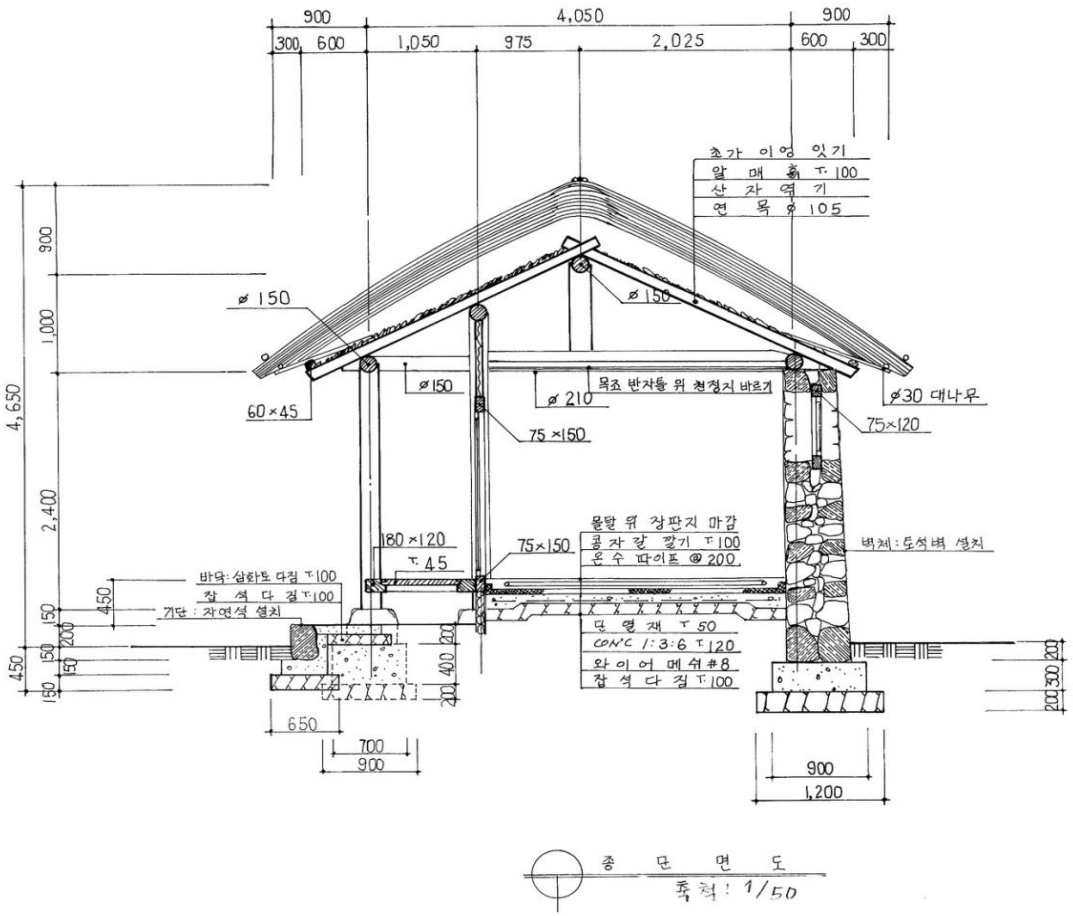
- SQL (Structured Query Language)
  - 데이터베이스를 조작하는 언어
- SQL의 특징
  - DBMS 제작 회사와 독립적임
  - 다른 시스템으로의 이식성이 좋음
  - 표준이 계속 발전함
  - 대화식 언어임
  - 클라이언트/서버 구조 지원함



- 표준 SQL과 각 회사의 SQL
  - 많은 회사가 되도록 표준 SQL을 준수하려고 노력하지만 각 회사의 DBMS마다 특징이 있기 때문에 현실적으로 완전히 통일되기는 어려움
  - 각 회사의 제품은 모두 표준 SQL을 공통으로 사용하면서 자기 제품의 특성에 맞춘 호환되지 않는 SQL 문 사용

데이터베이스 프로젝트

- 프로젝트 진행 단계
  - 프로젝트 : 현실 세계에서 일어나는 업무를 컴퓨터 시스템으로 옮겨 놓는 과정
  - 더 쉽게는 대규모 소프트웨어를 작성하기 위한 전체 과정

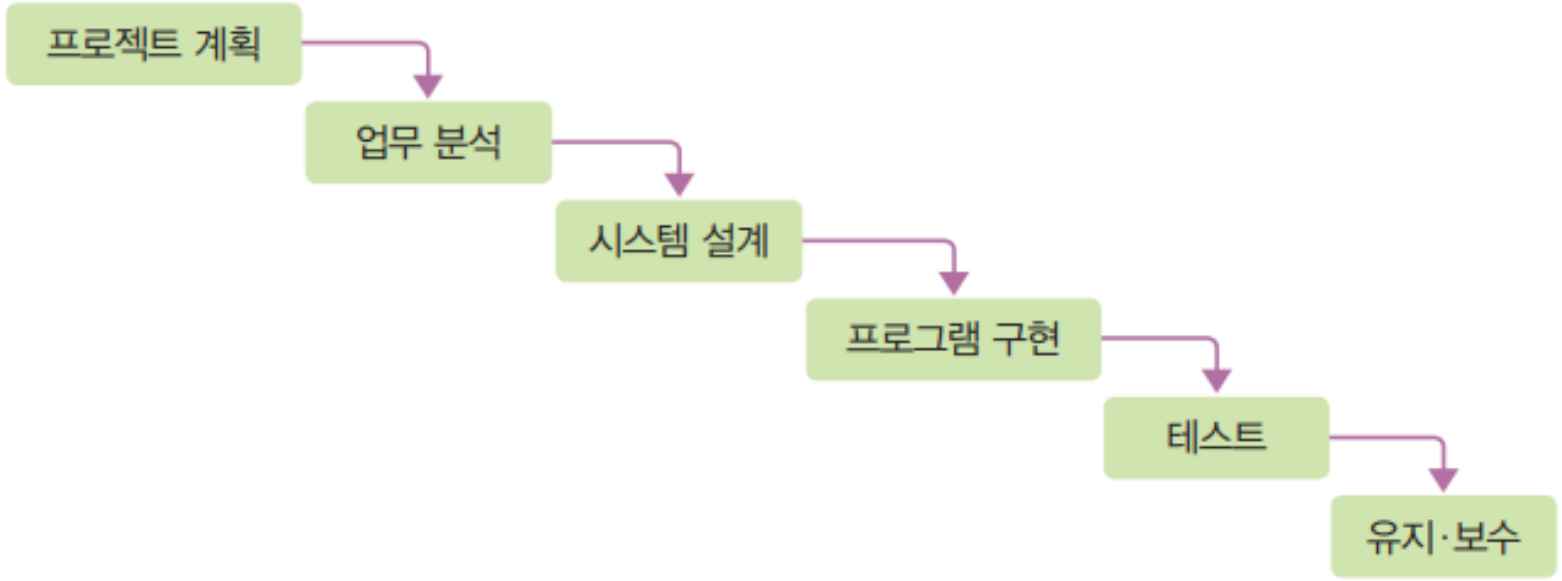


건물 설계
(소프트웨어 설계)

지반 설계
(데이터베이스 설계)

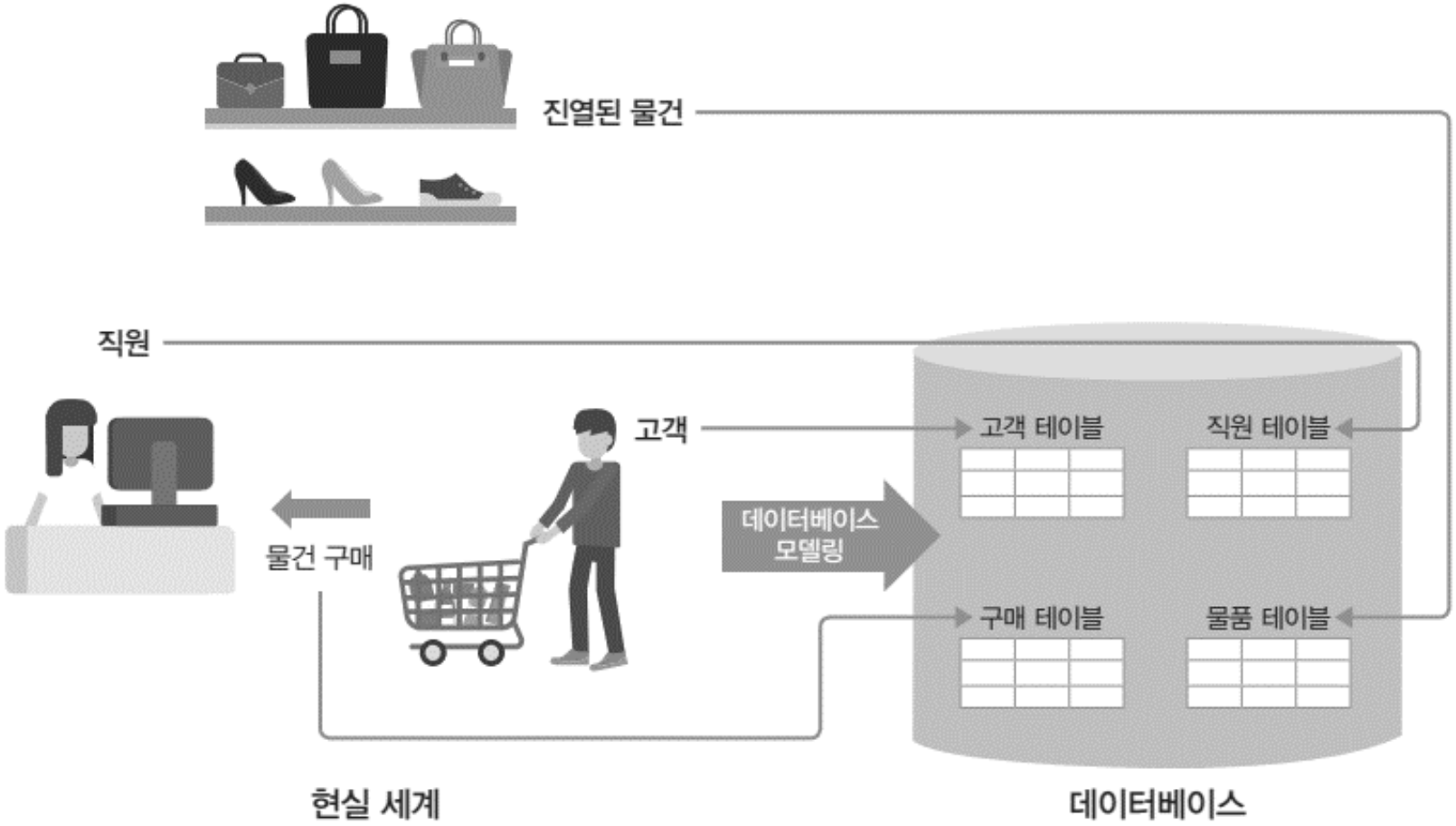
데이터베이스 프로젝트

- 폭포수 모델(waterfall model)
  - 폭포가 떨어지듯이 각 단계가 끝나면 다음 단계로 진행
  - 각 단계가 명확히 구분되어 프로젝트의 진행 단계가 명확
  - 앞 단계에서 문제가 발생했을 때 되돌아가기 어려움
  - 폭포수 모델에서 가장 핵심적인 단계는 업무 분석과 시스템 설계
  - 앞으로 살펴볼 데이터베이스 모델링은 분석과 설계 단계에서 가장 중요한 작업



데이터베이스 모델링

- 데이터베이스 모델링
  - 우리가 살고 있는 세상에서 사용되는 사물이나 작업을 DBMS의 데이터베이스 개체로 옮기기 위한 과정
  - 다시 말해, 현실에서 쓰이는 것을 테이블로 변경하는 작업

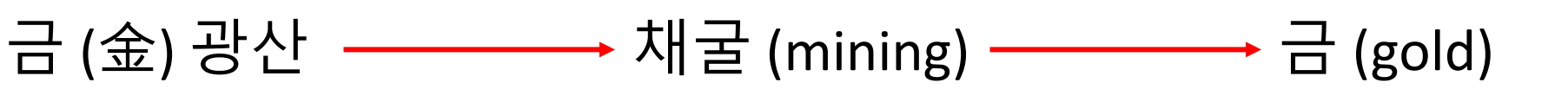


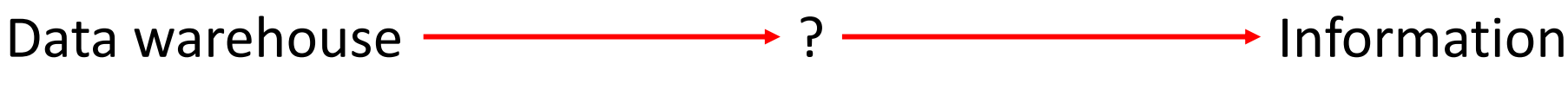
데이터베이스 모델링

- 데이터베이스 모델링에 정답은 없음
- 다만 좋은 모델링과 나쁜 모델링은 분명히 존재하며, 모범답안 형태의 모델링이 존재
- 이는 다양한 학습과 실무 경험에서 도출

데이터사이언스 : 데이터마이닝

- 데이터 마이닝
  - database, data warehouse, data mart 등 자료저장소에 저장되어 있는 방대한 양의 데이터로부터
  - 의사결정에 도움이 되는 유용한 정보를 발견하는 일련의 작업들의 집합.
  - 고객관계관리 (Customer Relationship Management)
    - 기업경영에서는 고객의 요구가 무엇인지 지속적으로 파악하는 것이 중요
    - 고객정보를 지속적으로 축적하고 분석해야 하며, 그 도구로서 데이터마이닝의 사용이 요구





데이터사이언스 : 데이터마이닝

- 데이터 마이닝
  - 1995년 지식발견 및 데이터마이닝(KDD：Knowledge Discovery and Data Mining) 국제학술대회가 처음 개최된 이후, 현재 데이터마이닝에 대한 정의는 다양하게 제시

    - “대량의 데이터 집합으로 부터 유용한 정보를 추출하는 것” (Han et al 2001)

    - “데이터 마이닝이란 의미 있는 패턴과 규칙을 발견하기 위해서 자동화 되거나 반자동화 된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정 (Berry and Linoff, 1977, 2000)

    - “데이터 마이닝은 통계 및 수학적 기술 뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정“ (Gartner Group 2004)

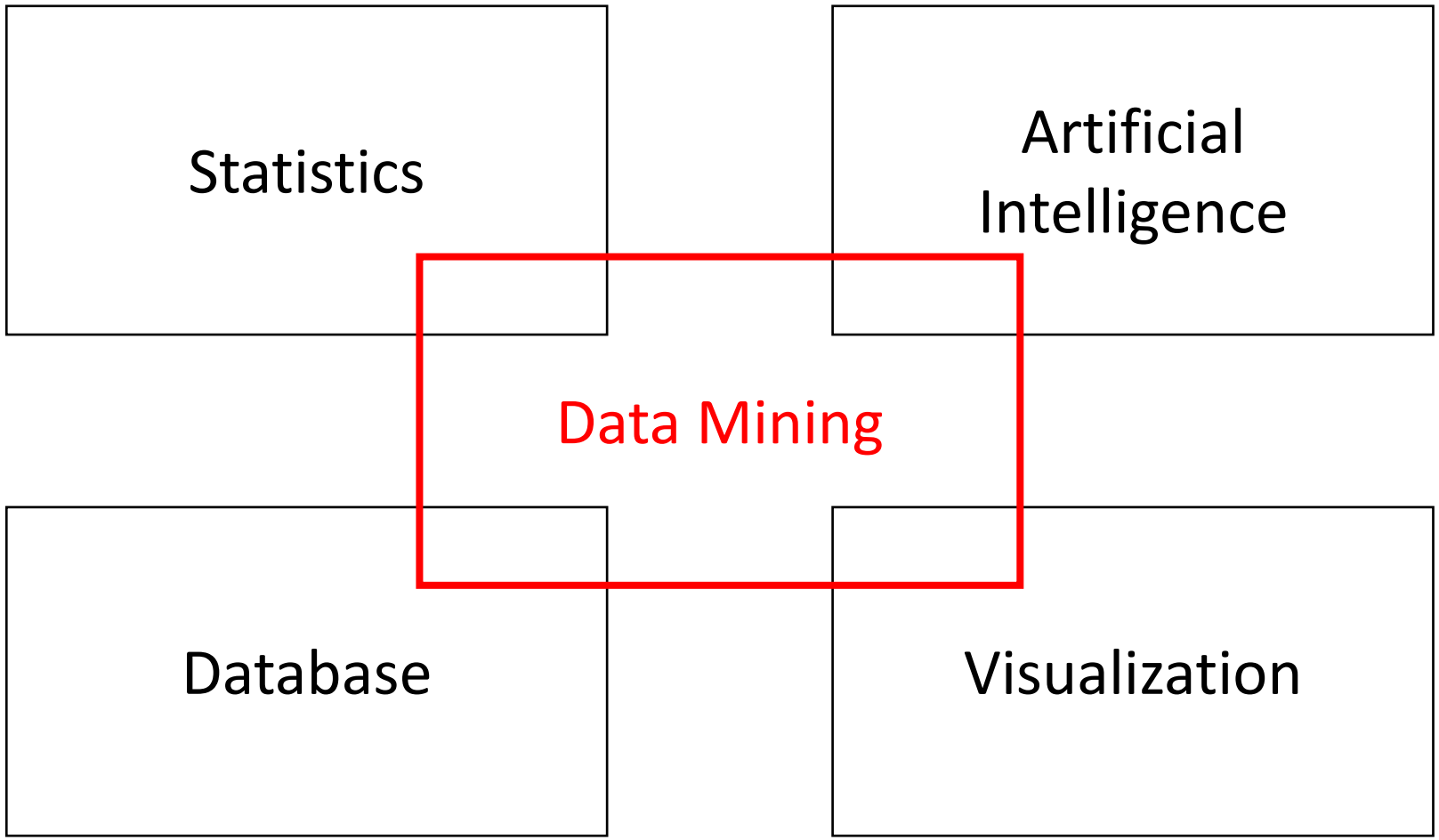
데이터마이닝의 역사

- 통계학 + 기계학습 / 인공지능
  - 오늘날에는 통계학과 기계학습 또는 인공지능이 함께 융합

- 통계와 데이터마이닝의 차이점
  - 기존의 통계적 분석 도구는 세워진 모형이나 가설에 의거해 이를 검증하거나 요약 보고하는데 중점
  - 데이터마이닝의 목적은 궁극적으로 예측에 중점
  - 데이터마이닝에 사용되는 인공지능 기법은 그 어떠한 기법보다 모형의 예측 성과를 높이는데 가장 우수함

전통적인 통계	데이터 마이닝
• 현실에 적응하기 부적합한 가정 (assumption) • 모수 추정이 주 목적 • 제안된 가설에 대한 검증이 주 목적 • 알고리즘에 선형성에 기반을 두고 있음	• 현실적인 noisy한 데이터에 대한 가정이 없음 • 알고리즘이 비선형성에 기반을 두고 있음 • 미래를 예측하는 것이 주 목적 • 모형에 대해 robust한 결과를 제공함 • 예측 성과가 통계기법보다 우수한 것으로 많은 실증 연구에서 검증 됨

데이터마이닝 관련 학문


Statistics
Artificial Intelligence
Data Mining
Database
Visualization

데이터마이닝과 비즈니스 인텔리전스

- 비즈니스 인텔리전스 (business intelligence)
  - 최종 사용자 질의 및 보고 (end user query and reporting)를 포괄하는 의미
  - ’90년대 초 Gartner 그룹의 Howard Dresner에 의해 만들어진 용어
  - 경영진과 경영 분석가들이 데이터를 통해 합리적 의사결정을 내릴 수 있도록 데이터를 수집, 저장, 처리, 분석하는 일련의 기술, 응용시스템을 의미
  - 아래의 모든 요소들을 포함하는 포괄적 의미로 사용
    - 데이터웨어하우스 (data warehouse)
    - 데이터 질의 및 보고도구 (data query and reporting tools)
    - 데이터 마이닝 (data mining)
    - 비즈니스 성과관리 (BPM : business performance management)

  - 즉 데이터마이닝은 비즈니스 인텔리전스의 일부로서 경영자와 경영분석가들이 다양한 비즈니스 의사결정문제를 해결해 주는 일련의 데이터 분석과정이라고 할 수 있음

데이터마이닝 특징

- 데이터마이닝의 특징
  - 대용량의 관측 가능한 자료를 다룬다.
  - 관측자료는 시간의 흐름에 따라 비계획적으로 축적되며, 자료분석을 염두에 두고 수집되지는 않는다.
  - 컴퓨터 중심적 기법이다.
  - 수리적으로 밝혀지지 않는 경험적 방법에 근거하고 있다.
  - 일반화에 초점을 두고 있다.
  - 경쟁력 확보를 위한 의사결정을 지원하기 위해 활용된다.

데이터마이닝 사용 예

- 수많은 가망고객 목록 중 어느 고객이 반응할 가능성이 가장 높은가?
  - 인구통계학 데이터 및 기타 데이터들을 이용하여 기존의 최고 우량고객들과 가장 일치하는 개인들을 파악하기 위해서 다양한 분류기법들(로지스틱 회귀분석, 분류나무 또는 다른 기법들)을 사용할 수 있음

- 가장 부정거래를 할 가능성이 높거나 이미 부정거래를 하였을 것 같은 고객은 누구인가?
  - 예를 들어 부정거래 가능성이 가장 높은 의료보상 청구신청을 식별하고, 이러한 청구신청에 대해 좀더 세심한 주의를 기울이기 위해 분류기법을 사용할 수 있음

- 어떤 대출 신청자가 파산할 것 같은가?
  - 파산가능성이 높은 대출신청자를 식별하기 위해 분류기법을 사용할 수 있음
  - 즉, 파산확률값을 부여하기 위해 로지스틱 회귀분석이 사용될 수 있음

- 전화, 잡지 등의 가입서비스를 포기할 것 같은 고객들은 누구인가?
  - 이탈고객들을 식별하기 위해 '이탈확률'값을 부여하는 로지스틱 회귀분석 등의 분류기법을 사용할 수 있음
  - 이 경우 이탈고객관리(churn management)를 통해 할인 또는 다른 유인책들을 선별적으로 내놓을 수 있음

데이터마이닝 프로세스

- 데이터마이닝 프로세스
  - 비즈니스 이해
    - 비즈니스 목표를 이해하고 이를 데이터 수집 목표로 정의 (비즈니스에 영향을 주는 중요한 항목 도출)
  - 데이터 이해
    - 초기 데이터 수집, 데이터 품질 체크, 가설을 위한 데이터 셋 정의
  - 데이터 준비
    - 분석 모델링에 필요한 데이터 추출
  - 모델링
    - 분석 기법을 선택, 분석에 필요한 최적의 변수 설정, 분석 모델 구축
  - 평가
    - 분석 모델 평가, 비즈니스 목표를 달성할 분석 모델 선정, 전체 프로세스를 재검토하고, 다음 단계 결정
  - 적용
    - 분석 모델링을 통해 획득한 지식 가공, 보고서 작성 및 시각화

데이터마이닝에서 사용하는 데이터

- 데이터 유형
  - 서술적 데이터 (descriptive data)
    - 개인이나 가구의 특성 묘사, 보통 요약 데이터의 형태
    - 고객의 기본 정보라서 자주 변하지 않음
    - 안정적이라 예측 모형을 구축하는데 유용
    - 응답자가 쉽게 응답하지 않거나 거짓 정보 제공
  - 행동특성 데이터 (behavioral data)
    - 기업이 고객과 상호 교류함으로써 발생하는 데이터
    - 고객의 행동이나 행위를 측정한 것이라 예측모형에 유용
    - 시간에 따라 빠르게 변화하며 쉽게 갱신가능
  - 태도특성 데이터 (attitudinal data)
    - 고객의 태도 또는 심리적 특성을 측정한 데이터
    - 정확한 데이터 수집이 어려움

데이터마이닝에서 사용하는 데이터

- 데이터 원천
  - 운영계 데이터베이스 (Operational Database)
    - 기업의 운영과 관련된 업무처리를 위해 구축된 것
    - 최근의 데이터를 저장하며 대량의 데이터 저장에 용이
  - 정보계 데이터베이스 (Informational Database)
    - 정보분석을 위해 구축된 것 /수집된 데이터를 요약, 가공하여 저장
  - 데이터 웨어하우스 (Data Warehouse)
    - 조직 전체를 통해 데이터에 대한 통합된 관점 제공
    - 좀 더 적절하고 유용한 정보를 만들 수 있도록 데이터 요약
  - 데이터 마트 (Data Mart)
    - 하나의 데이터마이닝 주제 또는 고객분석을 위해 통합된 데이터로 구성된 보조적인 데이터 저장소
    - 특정한 목적의 사용자 그룹을 위해 특정 주제영역의 데이터로 구성
  - 메타 데이터 (Meta Data)
    - 데이터에 대한 데이터이며, 데이터 관리를 위해 매우 상세하고 이해하기 쉽게 작성되어야 함

데이터마이닝 기법

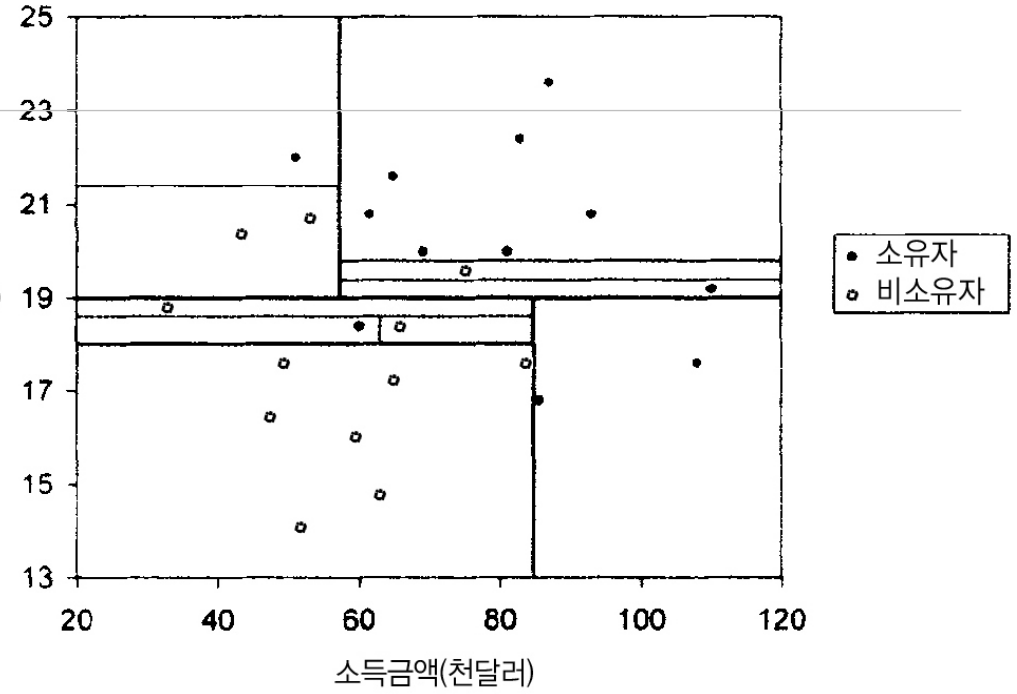
- 왜 다양한 데이터마이닝 기법들이 존재할까?
  - 각 기법 나름대로의 장.단점 존재

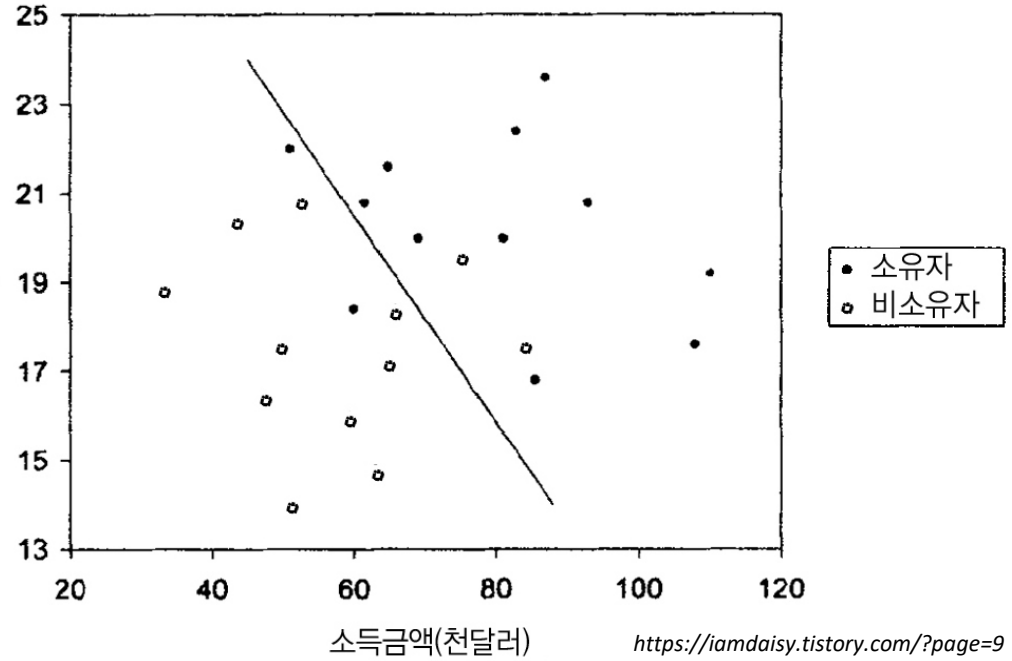
  - 어느 한 기법의 유용성은 데이터

집합의 크기, 데이터에 존재하는 패턴의 유형,
해당 기법이 요구하는 몇 가지 기본가정을 해당 데이터가
충족시키는지 여부, 데이터 잡음의 정도, 그리고 특수한
분석목적 등 다양한 요인들에 의해 영향을 받음

- 다양한 데이터마이닝 기법들을 적용해 보고,

그 중에서 목적에 맞는 가장 유용한 한 가지 기법을 선택





데이터마이닝 기법

- 연관성 규칙 (association)
  - 하나의 거래나 사건에 포함되어 있는 둘 이상의 품목들의 상호 관련성을 발견하는 것

- 분류 규칙 (classification)
  - 수집된 데이터의 패턴 및 속성으로 결합하여 트리형태의 모델로 변형, 의사결정 및 예측

- 군집 분석 (clustering)
  - 서로 유사한 정도에 따라 다수의 객체를 군집으로 나누는 작업 또는 이에 기반한 분석

연관 규칙 (association)

- 한 데이터와 다른 데이터 사이의 관련성이 있음을 찾는 규칙
- 나열되어 있는 원본 데이터들에 대해서 데이터들간의 연관관계를 탐색하는 무방향성 데이터마이닝 기법 중 하나
- 장바구니 분석 : 보통 마트의 고객들이 구매한 상품 간의 연관관계를 찾을 때 많이 사용되는 분석기법
- 수학 및 통계학의 확률과 기대값에 대한 기념을 기반
  - 원인과 결과의 직접적인 인과관계로 생각해서는 안됨
  - 두 개 또는 그 이상의 품목들 사이의 상호의 관련성으로 해석해야 함
- 예시
  - 주말을 위해 목요일에 기저귀를 사러 온 고객들은 맥주도 함께 구매
  - 최근 통장정리와 금리상담을 요구한 고객은 이 후 한달 이내에 거래를 중단할 가능성이 일반 고객의 2배
  - 이전에 동일한 제조사의 전자제품을 주로 구매했던 고객은 신제품 구매에서도 동일한 회사의 제품을 구매

- 데이터를 통해 얻어지는 모든 연관성이 의미 있다고 보기는 어려우며, 수많은 품목 들의 관계 속에서 의미 있는 관련성을 찾기 위해서는 결과해석에 앞서 연관성의 내용이 일반화할 수 있는 내용인가를 판단할 수 있는 기준이 필요

연관 규칙 (association)

- 연관규칙의 특징
  - 장바구니 분석
    - 구매 내역 분석을 통해 동시에 구매될 가능성이 있는 상품의 연관관계를 찾음
  - 비지도 예측 / 자율 예측 (unsupervised prediction)
    - 목표변수, 결과값을 모르는 상태에서 데이터를 분석해서 데이터간의 관계를 분석해서 결과값을 도출
    - 비지도 예측의 가장 대표적인 것이 연관 규칙 분석
  - 탐색적 기법
    - 조건반응(if then else)으로 표현
      - 만일 A가 일어난다면 B가 일어난다. / 상품 A를 구매하면 상품 B도 구매를 한다.
  - 목적변수 없음
    - 목적변수(Target Variable) 없이 특성의 조합으로 규칙을 표현,
    - 특정한 변수가 아닌 모든 변수 또는 특성에 대하여 예측
  - 규칙간의 독립성
    - 규칙들 간에는 서로 영향을 주지 않기 때문에 하나의 고객이 여러 개의 규칙에 해당 될 수 있음

연관 규칙 (association)

- 지지도 (support) 와 신뢰도 (confidence)
  - 품목들 간의 연관성의 정도를 평가하는 중요한 평가도구

  - 지지도 (support)
    - 전체 자료에서 관련성이 있다고 판단되는 항목 A, B가 동시에 일어날 확률
    - 지지도는 상호 대칭적이므로 지지도(A→B) = 지지도(B→A)
    - 지지도(A→B) = $\Pr(A \cap B) = \frac{n(A,B)}{N}$

  - 신뢰도 (confidence)
    - 항목 A가 구매되었을 때 항목 B가 추가로 구매될 확률 (조건부 확률)
    - 신뢰도는 상호 대칭적이지 않으므로 신뢰도(A→B)≠신뢰도(B→A)
    - 신뢰도(A→B) = $\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{n(A,B)}{n(A)}$

연관 규칙 (association)

- 예시 : 피자 가게에서 토핑을 무엇으로 하는가?

요약된 거래 데이터 (추가토핑) → **재구성 데이터**

항목	거래 수
버섯	100
페페로니	150
치즈	200
버섯 + 페페로니	400
버섯 + 치즈	300
페페로니 + 치즈	200
버섯 + 페페로니 + 치즈	100
추가 토핑 X	550
합계	2000

항목	항목이 포함된 거래 수	포함률
버섯	100+400+300+100=900	45%
페페로니	150+400+200+100=850	42.5%
치즈	200+300+200+100=800	40.0%
버섯 + 페페로니	400+100=500	25.0%
버섯 + 치즈	300+100=400	20.0%
페페로니 + 치즈	200+100=300	15.0%
버섯 + 페페로니 + 치즈	100	5.0%

연관 규칙 (association)

- 예시 : 피자 가게에서 토핑을 무엇으로 하는가?

지지도와 신뢰도

규칙 (A→B)	지지도(A→B)	신뢰도(A→B)
버섯 → 페페로니	25%	25/45 = 55.6%
버섯 + 페페로니 → 치즈	5%	5/25 = 20.0%
버섯 + 치즈 → 페페로니	5%	5/20 = 25.0%
페페로니 + 치즈 → 버섯	5%	5/15 = 33.3%
…	…	…

지지도와 신뢰도는 확률의 개념이기에 0과 1사이의 값을 가지며, 1에 가까울 수록 연관성이 높다고 판단

연관 규칙 (association)

- 향상도 (Lift)
  - 우연에 의한 연관성의 정도 측정
  - 향상도(A→B) $= \frac{\Pr(B|A)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A)\Pr(B)}$
  - 향상도 > 1 : 두 항목 A, B가 양의 연관성을 가짐 (예: 빵과 버터)
  - 향상도 = 1 : 두 항목 A, B가 독립에 가까움 (과자와 후추)
  - 향상도 < 1 : 두 항목 A, B가 음의 연관성을 가짐 (지사제와 변비약)
  - 향상도가 1에 가깝다면 신뢰도가 높다 하더라도 우연에 의해 연관성이 높게 나타났을 가능성이 있음

규칙 (A→B)	Pr(A)	Pr(B)	지지도(A→B)	신뢰도(A→B)	향상도(A→B)
버섯 → 페페로니	45%	42.5%	25%	0.556	1.31
버섯 + 페페로니 → 치즈	25%	40%	5%	0.2	0.5
버섯 + 치즈 → 페페로니	20%	42.5%	5%	0.25	0.588
페페로니 + 치즈 → 버섯	15%	45%	5%	0.333	0.74

연관 규칙 (association)

- 연관 규칙의 응용
  - 둘 또는 그 이상의 항목들 사이의 수많은 연관성 규칙 중에서

  지지도, 신뢰도, 향상도에 근거해 일반화 할 수 있는 의미 있는 규칙을 탐색하는 방법
  - 교차판매, 매장진열, 카탈로그 디자인, 소프트웨어 번들링, 첨부 우편물, 보험의 부정행위 적발 등

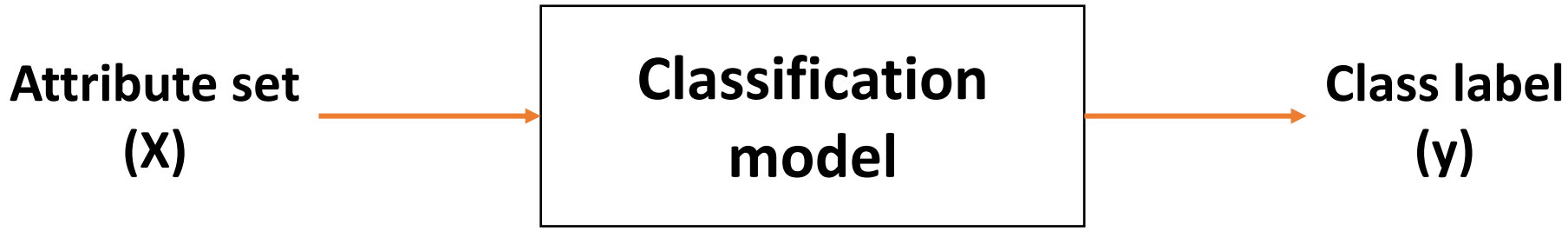
- 예시
  - 어떤 제품과 관련된 선택사양들을 고객의 성향에 따라 정리한다면
    - 비슷한 성향을 갖는 신규고객에게 동시에 그 둘을 제공해 구매의사를 높임
  - 예금구좌, CD, 투자 서비스, 카드 대출 등의 은행 상품의 관련성을 파악한다면
    - 고객들이 원하는 또 다른 상품의 형태를 파악해 고객의 구매를 유도
  - 보험금 청구 내용이 이전의 내용과 비교해 관련성이 적은 특이한 상황이라면
    - 보험사기를 의심하고 좀 더 면밀한 조사를 요구

연관 규칙 (association)

- 연관 규칙의 장단점
  - 장점
    - 탐색적인 기법
      - ‘조건 => 반응’ 의 규칙의 형태를 가지고 있어 이해가 쉽고 적용이 용이
    - 강력한 비목적성 분석기법
      - 대부분의 데이터마이닝 기법과 달리 뚜렷한 목적변수 없이도 적용이 쉬움
    - 사용 편리한 분석데이터의 형태
      - 특별한 변환 없이 간단히 사용이 가능한 데이터구조를 갖고 있음
    - 계산의 용이성
      - 대용량의 데이터인 경우 계산의 수가 크게 증가하기는 하지만 분석을 위한 계산은 아주 간단함
  - 단점
    - 연관성을 관찰하고자 하는 항목이 증가하면 계산의 수가 크게 증가
    - 적절한 항목의 결정 : 실제 불필요한 항목들이 많이 존재
    - 거래량이 적은 품목의 경우 거래수가 적기 때문에 연관성규칙발견 과정 중 제외될 가능성 있음

분류 규칙 (classification)

- 분류 규칙 (classification)
  - 데이터마이닝의 가장 기본적인 기법 중 하나로 범주형 자료 또는 이산형 자료에 사용
  - 각 속성 집합 x를 미리 정의된 클래스 레이블 y중 하나에 매핑하는 모델을 학습
  - 테스트 집합 (test set)은 학습한 모델의 정확도를 결정하는 역할
  - 예시
    - Categorizing email message
      - X : features extracted from email message header and context
      - Y : spam or non-spam



분류 규칙 (classification)

- 혼동 행렬 (confusion matrix)

		예측 클래스	
		Class = 0	Class = 1
실제 클래스	Class = 0	a	b
	Class = 1	c	d

  - 정확도
    - 정확한 예측 개수 / 총 예측 개수 = (a+d) / (a+b+c+d)
  - 오류율
    - 부정확한 예측 개수 / 총 예측 개수 = (b+c) / (a+b+c+d)

분류 규칙 (classification)

- 분류성능평가지표 - Precision(정밀도), Recall(재현율) and Accuracy(정확도)

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

- True Positive(TP) : 실제 True인 정답을 True라고 예측 (정답)
- False Positive(FP) : 실제 False인 정답을 True라고 예측 (오답)
- False Negative(FN) : 실제 True인 정답을 False라고 예측 (오답)
- True Negative(TN) : 실제 False인 정답을 False라고 예측 (정답)

분류 규칙 (classification)

- 분류성능평가지표 - Precision(정밀도), Recall(재현율) and Accuracy(정확도)

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

- Precision (정밀도) : 정밀도란 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율

$$(Precision) = \frac{TP}{TP + FP}$$

- Recall (재현율) : 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율

$$(Recall) = \frac{TP}{TP + FN}$$

분류 규칙 (classification)

- 분류성능평가지표 - Precision(정밀도), Recall(재현율) and Accuracy(정확도)
  - Recall (재현율)
    - 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율
    - 통계학에서는 sensitivity으로, 그리고 다른 분야에서는 hit rate라는 용어로도 사용
  - Precision이나 Recall은 모두 실제 True인 정답을 모델이 True라고 예측한 경우에 관심이 있음
  - 하지만 Precision은 모델의 입장에서, Recall은 실제 정답(data)의 입장에서 정답은 정답으로 맞춘 경우 고려
  - Precision-Recall Trade-off

		H_0	
		True	False
Test result	Accept		**Type 2 error**
	Reject	**Type 1 error**	

분류 규칙 (classification)

- 분류성능평가지표 - Precision(정밀도), Recall(재현율) and Accuracy(정확도)
  - Accuracy(정확도)
    - Precision 과 Recall은 모두 True를 True라고 옳게 예측한 경우에 대해서만 다룸
    - 하지만, False를 False라고 예측한 경우도 옳은 경우이며, 이를 함께 고려하는 지표가 바로 정확도(Accuracy)
  - 정확도는 가장 직관적으로 모델의 성능을 나타낼 수 있는 평가 지표

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

- F1 Score
  - F1 score는 Precision과 Recall의 조화평균
  - F1 score는 데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있음

$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

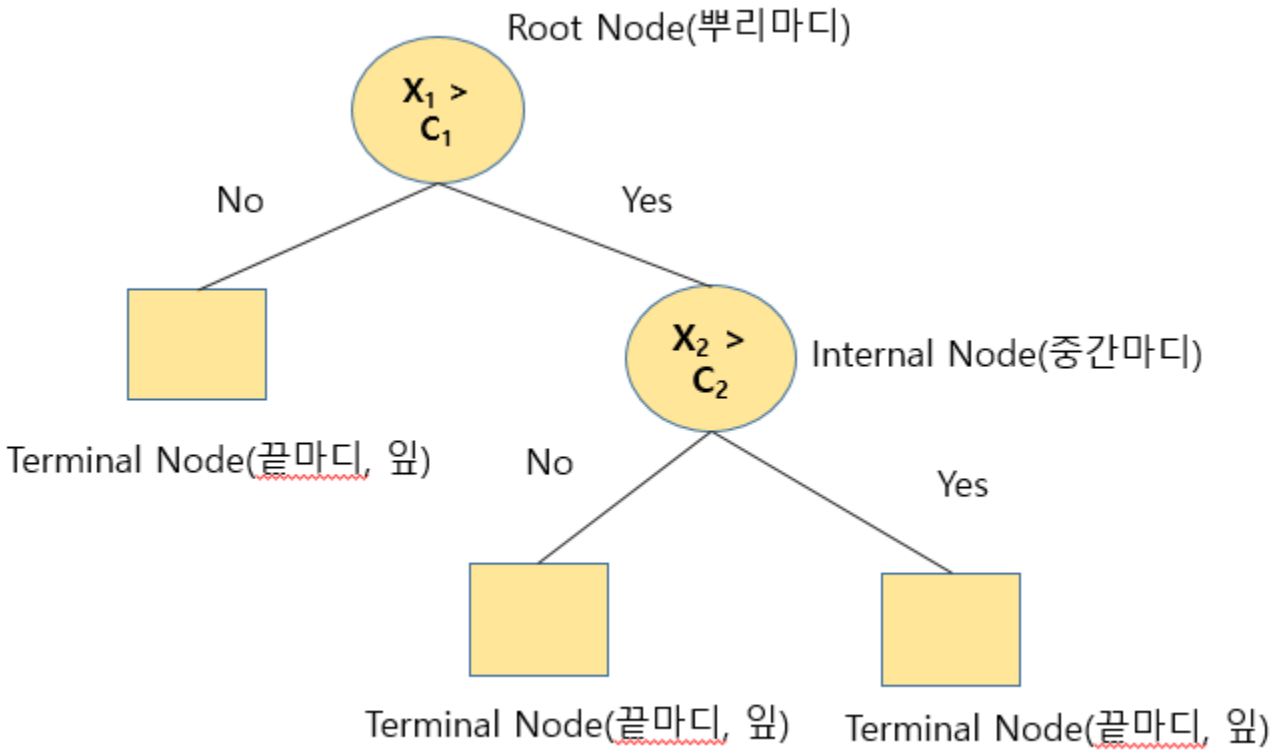
분류 규칙 (classification)

- 분류 기법

기본 분류기 (Base classifier)	앙상블 분류기 (Ensemble classifier)
의사 결정 트리 (Decision tree)	Bagging, Random Forest
규칙 기반 (Rule-base method)	Boosting, AdaBoost, …
인접 이웃 (Nearest-neighbor)	Stacking
신경망 (Neural Networks)	
심화신경망 (Deep Learning)	
베이지안 (Naiive Bayes)	
지지도 벡터 머신 (Support Vector Machine)	

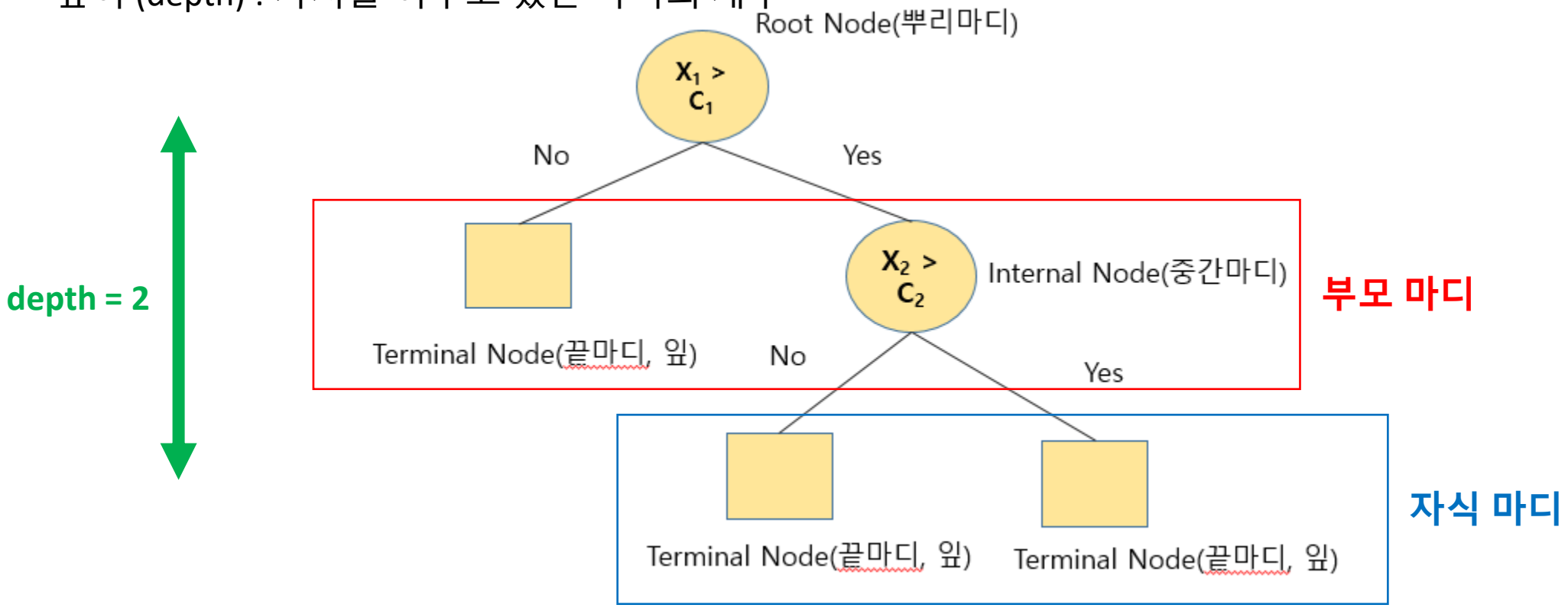
분류 규칙 (classification)

- 의사 결정 나무 (decision tree)
  - 뿌리 마디 (root node) : 나무구조가 시작되는 마디
  - 끝 마디 (terminal node, leaf) : 각 나무줄기의 끝에 위치한 마디
  - 중간 마디 (internal node) : 중간에 있는 끝 마디가 아닌 마디



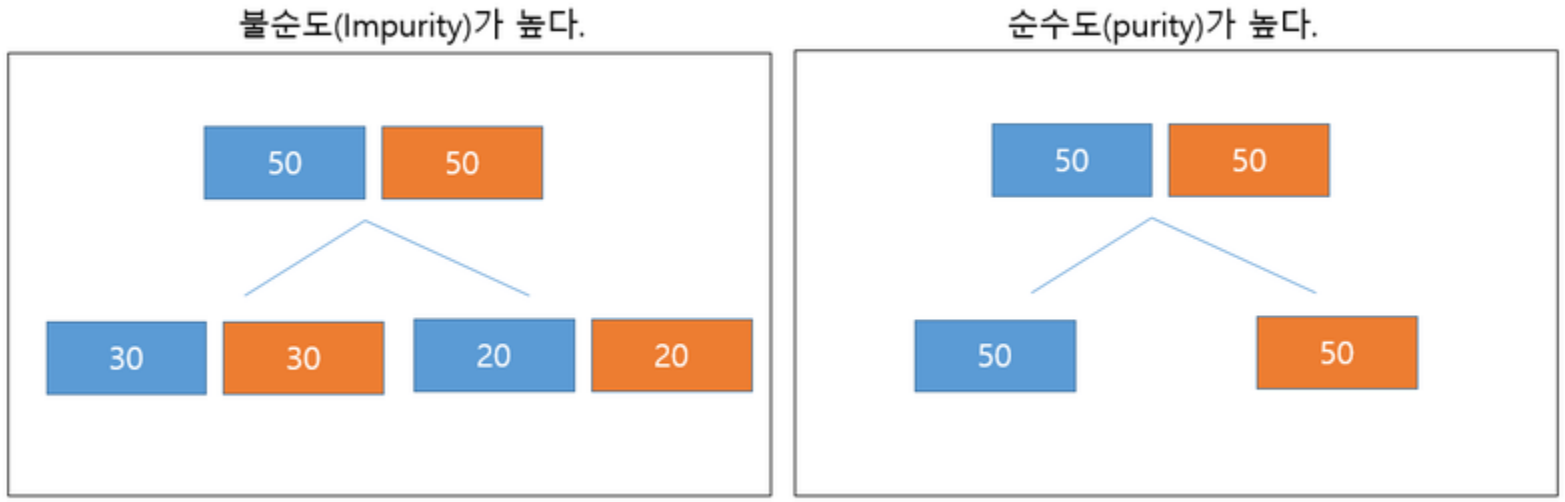
분류 규칙 (classification)

- 의사 결정 나무 (decision tree)
  - 자식 마디 (child node) : 하나의 마디로부터 분리되어 나간 마디
  - 부모 마디 (parent node) :자식 마디의 상위 마디
  - 가지 (branch) : 하나의 마디로부터 끝 마디까지 연결된 마디들
  - 깊이 (depth) : 가지를 이루고 있는 마디의 개수



분류 규칙 (classification)

- 의사 결정 나무 (decision tree)의 형성
  - 목적과 자료구조에 따라 적절한 분리기준과 정지규칙을 지정하여 의사결정나무를 얻음
  - 분리 기준 (split criterion)
    - 어떤 입력변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지 그 기준
    - 목표변수의 분포를 구별하는 정도 : **순수도 / 불순도**
      - 순수도 : 목표변수의 특정 범주에 개체들이 포함되어 있는 정도
    - 부모마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식마디를 형성함



분류 규칙 (classification)

- 의사 결정 나무 (decision tree)의 형성
  - 정지 규칙 (stopping rule)
    - 더 이상 분리가 일어나지 않고, 현재의 마디가 끝 마디가 되도록 하는 규칙
    - 의사결정나무에서는 초기에 트리를 키우는 단계에서 정지규칙을 적용하지 않으면 끝마디가

    하나의 범주만을 가질 때까지 계속하여 최대 트리를 형성하기 때문에 과적합(overfitting) 문제가 발생 가능

  - 가지 치기 (pruning)
    - 적절하지 않은 마디를 제거하여, 적당한 크기의 부나무(subtree) 구조를 가지도록 하는 규칙
    - 과적합 문제를 줄이기 위해 정지규칙(Stopping Rule) 최대 트리를 적절히 가지치기(Pruning)을 이용하여

    예측력이 좋은 분류규칙을 도출

분류 규칙 (classification)

- 의사 결정 나무 (decision tree)의 장단점
  - 장점
    - 해석의 용이성
      - 나무구조에 의해서 모형이 표현되기 때문에 해석이 쉬움
      - 새로운 자료에 모형을 적합 시키기 쉬우며, 어떤 입력변수가 중요한지 파악이 쉬움
    - 비모수적 모형
      - 선형성, 정규성, 등분산성의 가정이 필요 없음
      - 단지 순위만 분석에 영향을 주므로 이상치에 민감하지 않음
  - 단점
    - 비연속성
      - 연속형 변수를 비연속적인 값으로 취급하여 예측오류가 클 가능성이 있음
    - 비안정성
      - 분석용 자료에만 의존하므로 새로운 자료의 예측에 불안정

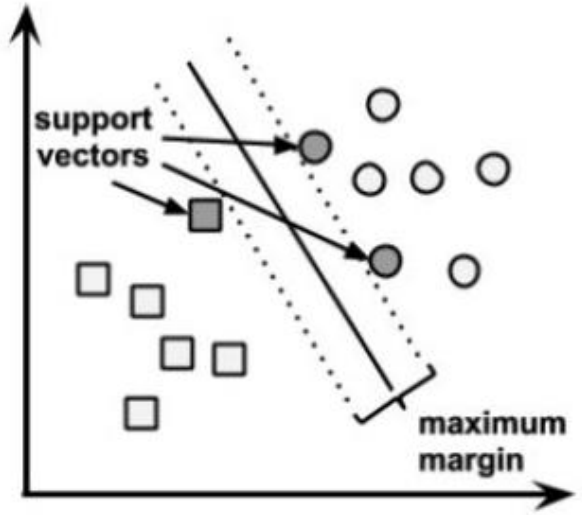
분류 규칙 (classification)

- 지지도 벡터 머신 (Support Vector Machine)
  - 데이터의 특성을 이용하여 입체 공간에 데이터를 분류할 수 있는 초평면(hyperplane)을 만들어 주는 알고리즘
    - Hyperplane: 데이터를 나누는 평면
      - $\mathrm{W^T X} + \mathrm{b}=0$
      - Support vector: 초평면에서 가장 가까운 데이터
      - Margin: 양쪽 support vector와 hyperplane 간의 거리의 합
  - 분류 (classification)

$$y_i = +1 \text{ when } \mathrm{W^T x_i} + \mathrm{b} \geq +1$$
$$y_i = -1 \text{ when } \mathrm{W^T x_i} + \mathrm{b} \leq -1$$

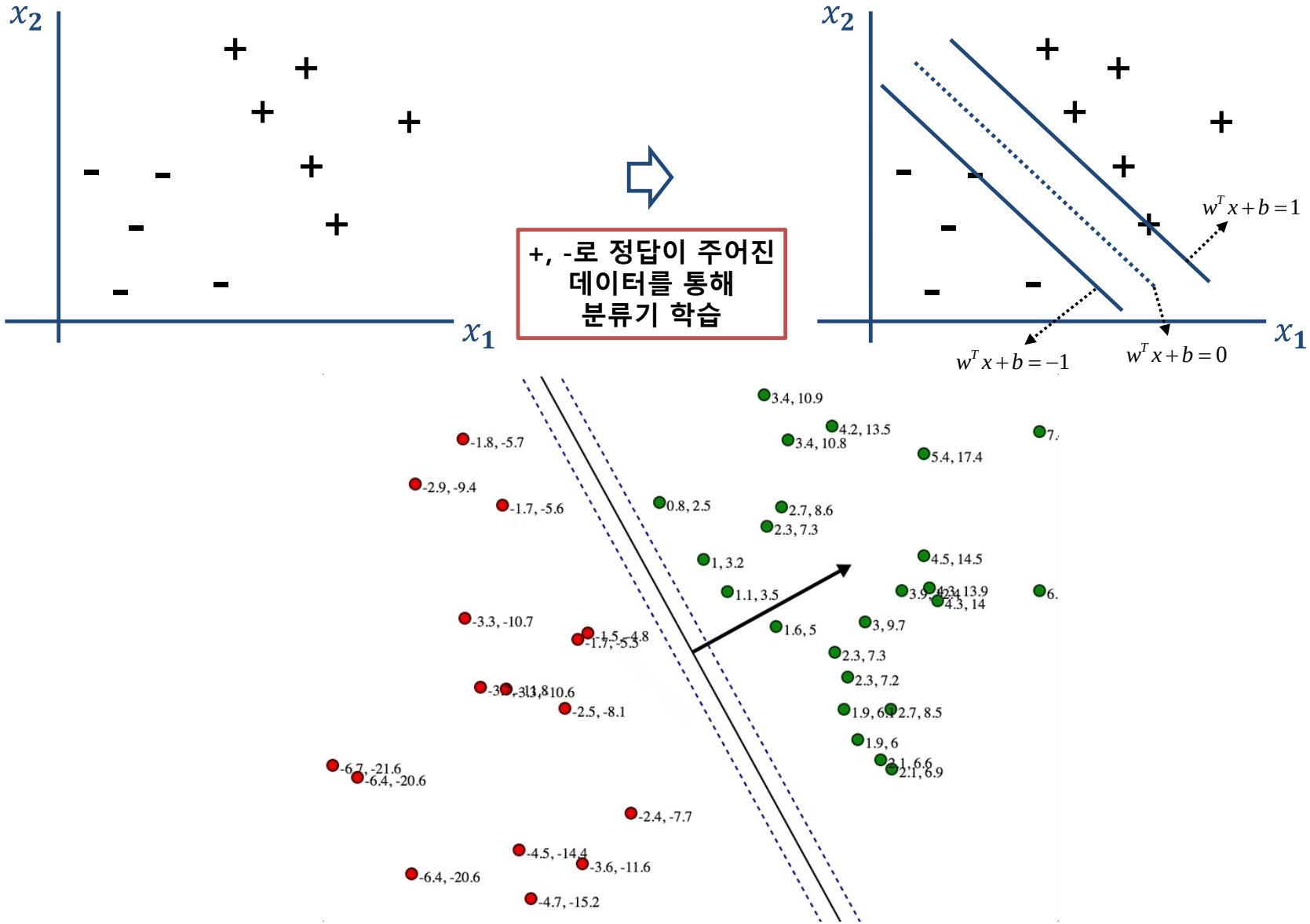
  - 예측 (prediction)

$$\mathrm{y} = \mathrm{W^T x_i} + \mathrm{b}$$



분류 규칙 (classification)

- 지지도 벡터 머신 (Support Vector Machine)

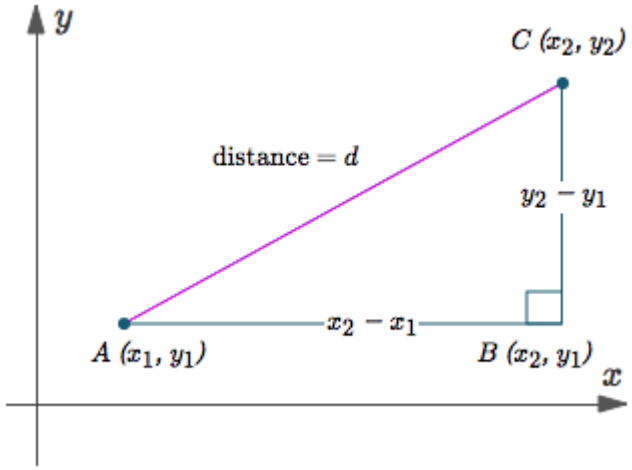


군집 분석 (clustering)

- 군집 분석
  - 변수 또는 개체(item)들이 속한 모집단 또는 범주에 대한 사전정보가 없는 경우에 관측 값들 사이의 유사성을 이용하여 변수 또는 개체들을 자연스럽게 몇 개의 그룹 또는 군집으로 나누는 분석방법
  - 각 객체(대상)의 유사성을 측정하여 유사성이 높은 대상 집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체 간의 상이성을 규명하는 분석방법
  - 군집의 개수나 구조에 대한 가정 없이 데이터들 사이의 거리를 기준으로 군집화를 유도
  - 마케팅에서 소비자들의 상품 구매활동이나 life style에 따른 소비자군을 분류하여 시장 전략 수립 등에 활용
  - 군집분석은 통계적 방법이기도 하지만 최근 관심이 높아진 머신러닝의 대표적인 방법
  - 군집을 묶는 데에는 정답이 없으며, 결국 연구자의 판단에 달려있음

군집 분석 (clustering)

- 군집 분석 특징
  - 종속 변수 (y변수)가 없는 데이터 마이닝 기법 (비지도 예측)
  - 유클리드 거리 기반 유사 객체 묶음 (유사성 = 유클리드 거리)
    - 유클리드 거리 : 두 점 사이의 거리를 계산하는 방법



$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

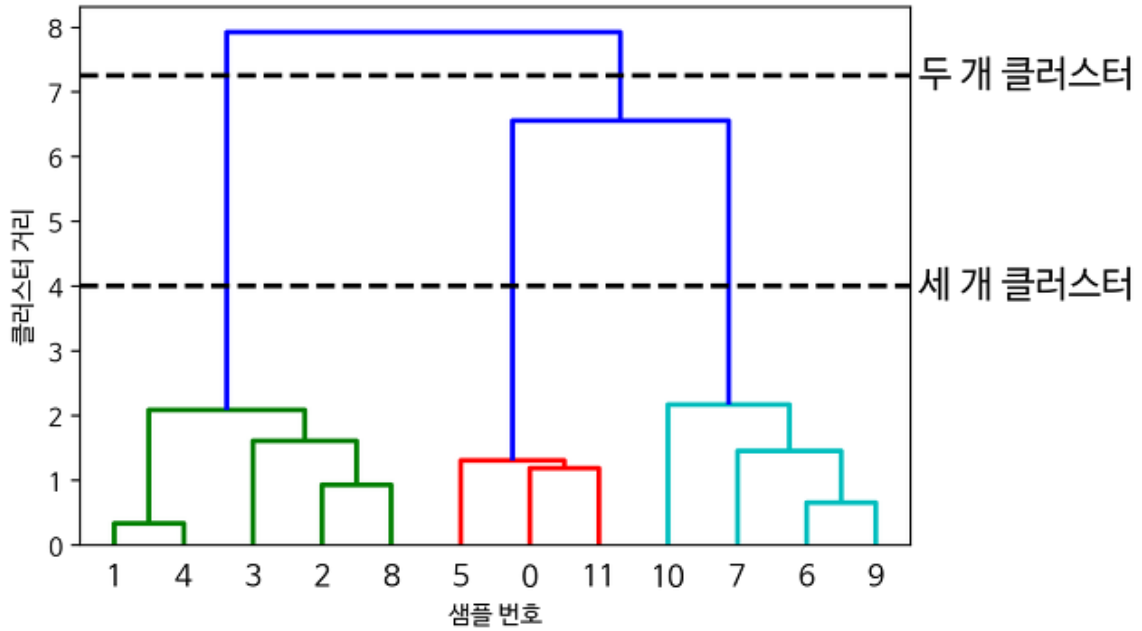
연속형 변수의 거리	수학적 거리 (math matical distance)
	유클리드 거리 (euclidian distance)
	맨하튼 거리 (manhatten distance)
	민코프스키 거리 (minkowski distance)
통계적 거리	표준화 거리 (standartized distance)
	마할라노비스 거리 (mahalanobis distance)

군집 분석 (clustering)

- 군집 분석 종류
  - 계층적 군집 분석 (hierarchical) / 비계층적 군집분석
  - 계층적 군집 분석 (탐색적 군집 분석)
    - N개의 군집으로 시작하여 점차 군집의 개수를 줄여 나가는 방법
    - 단일기준결합방식
      - 각 군집에서 중심으로부터 거리가 가까운 것 1개씩 비교하여 가장 가까운 것끼리 군집화
    - 완전기준결합방식
      - 각 군집에서 중심으로부터 가장 먼 대상끼리 비교하여 가장 가까운 것끼리 군집화
    - 평균기준결합방식
      - 한 군집 안에 속해 있는 모든 대상과 다른 군집에 속해 있는 모든 대상의 쌍 집합에 대한 거리를 평균 계산하여 가장 가까운 것끼리 군집화

군집 분석 (clustering)

- 계층적 군집 분석 (탐색적 군집 분석)
  - 덴드로그램 (Dendrogram) : 표본들이 군을 형성하는 과정을 나타내는 나무 형식의 그림
  - 과정
    - 거리행렬을 기준으로 덴드로그램을 그림
    - 덴드로그램의 최상단부터 세로축의 개수에 따라 가로선을 그어 군집의 개수를 선택
    - 각 객체들의 구성을 고려해서 적절한 군집수를 선정
      - 군집의 수는 분석 목적에 따라 선정할 수 있지만 대부분 5개 이상의 군집은 잘 활용 x

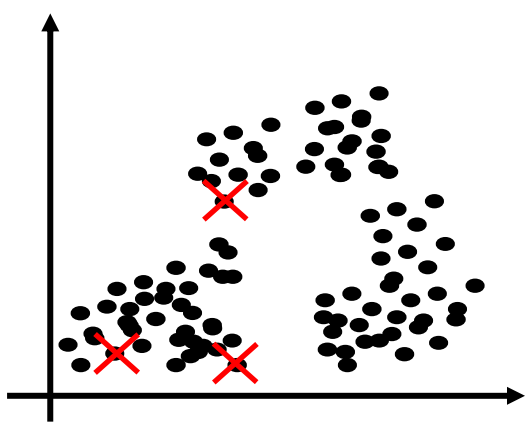


군집 분석 (clustering)

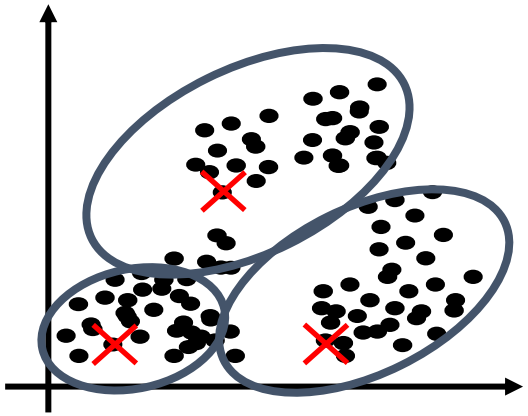
- 비계층적 군집 분석
  - N개의 개체를 n개의 군집으로 나눌 수 있는 모든 가능한 방법을 점검해 최적화한 군집을 형성
  - K-means clustering
    - 데이터를 k개의 클러스터로 묶는 알고리즘, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작
    - 과정
      - 원하는 군집의 개수와 초기값(seed)들을 정해 seed중심으로 군집을 형성
      - 각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류
      - 각 군집의 seed 값을 다시 계산
      - 모든 개체가 군집으로 할당될 때까지 위 과정들을 반복

군집 분석 (clustering)

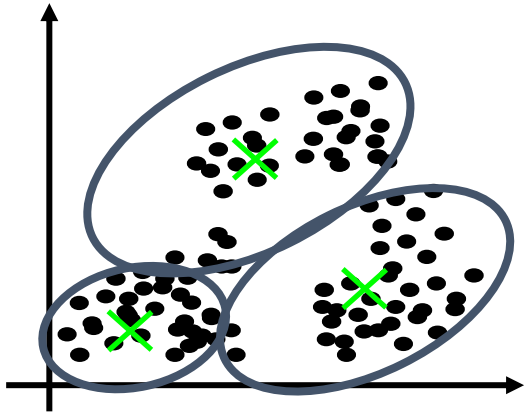
- K-means clustering
  - 과정



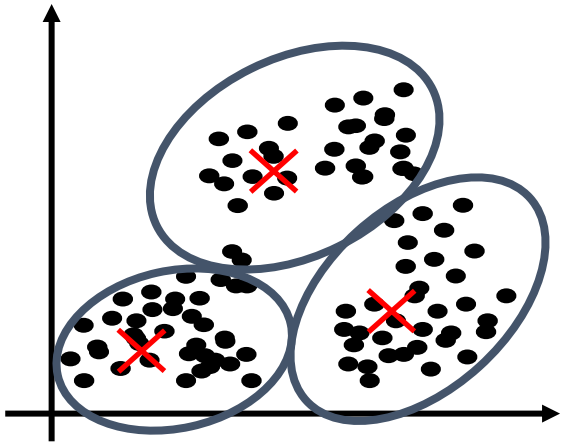
1. K개의 중심을 무작위로 선택



2. 클러스터 배정



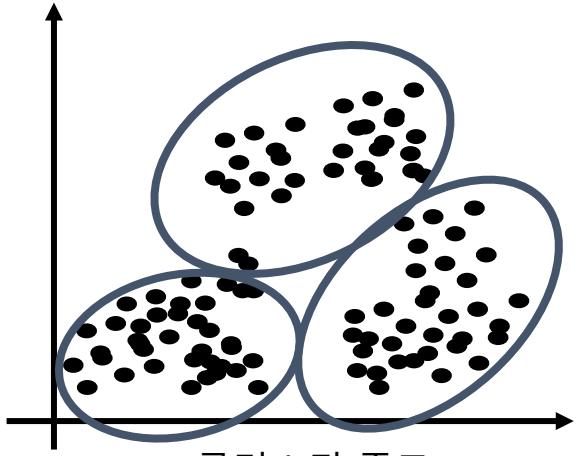
3. 중심점 다시 계산



4. 클러스터 재배정

2~4 반복

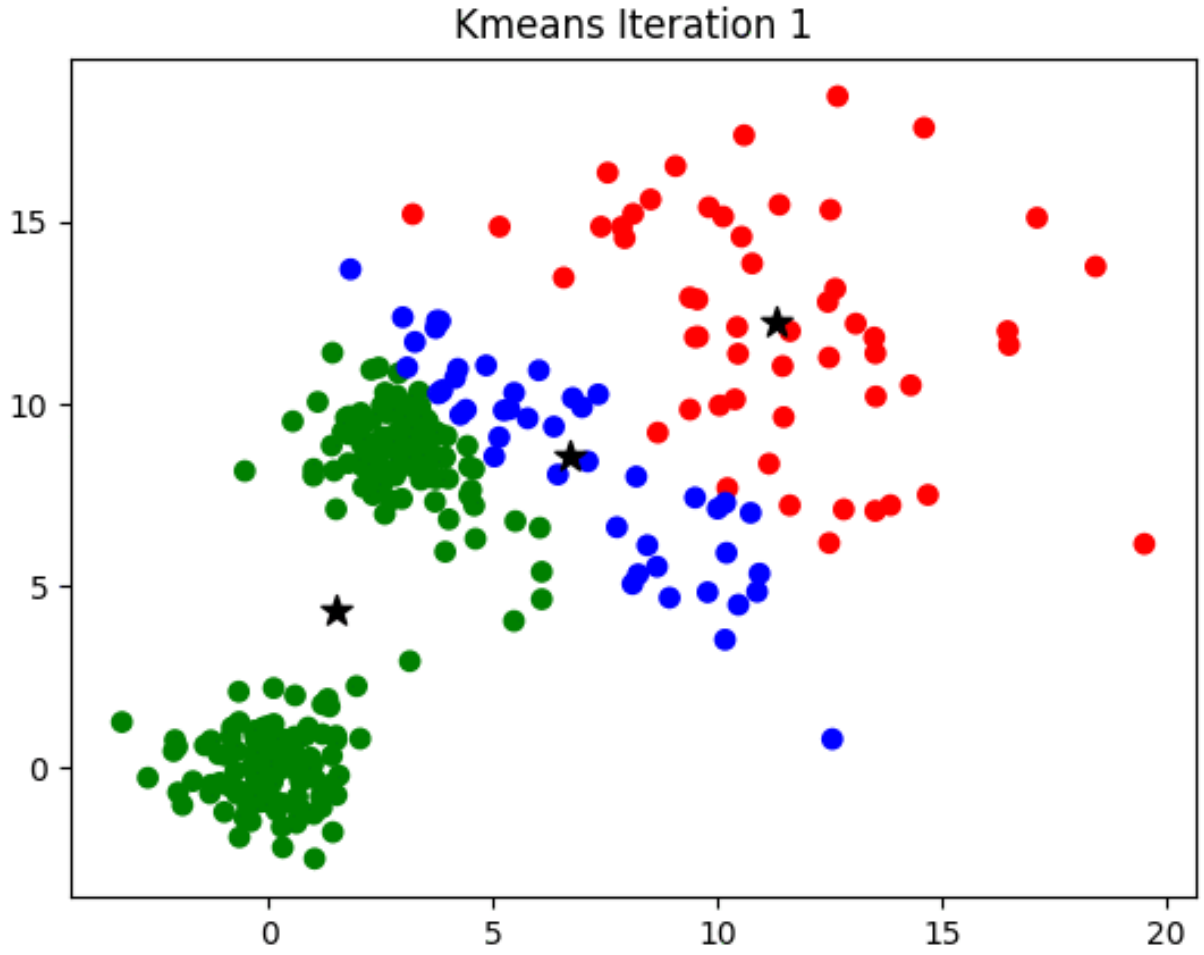
• • •



클러스터 종료

군집 분석 (clustering)

- K-means clustering
  - 과정



군집 분석 (clustering)

- K-means clustering
  - 장점
    - 비교적 쉬운 알고리즘
    - 구현이 쉬움
  - 단점
    - 데이터의 평균 값이 정의될 수 있는 데이터에만 사용 가능
    - K는 사용자가 정해야함
    - 최적의 K를 찾기 어려움
    - 아웃라이어(outlier)에 민감함
    - 아웃라이어란 데이터들에서 매우 동 떨어져 있는 데이터

감사합니다

kimtwan21@dongduk.ac.kr

김 태 완