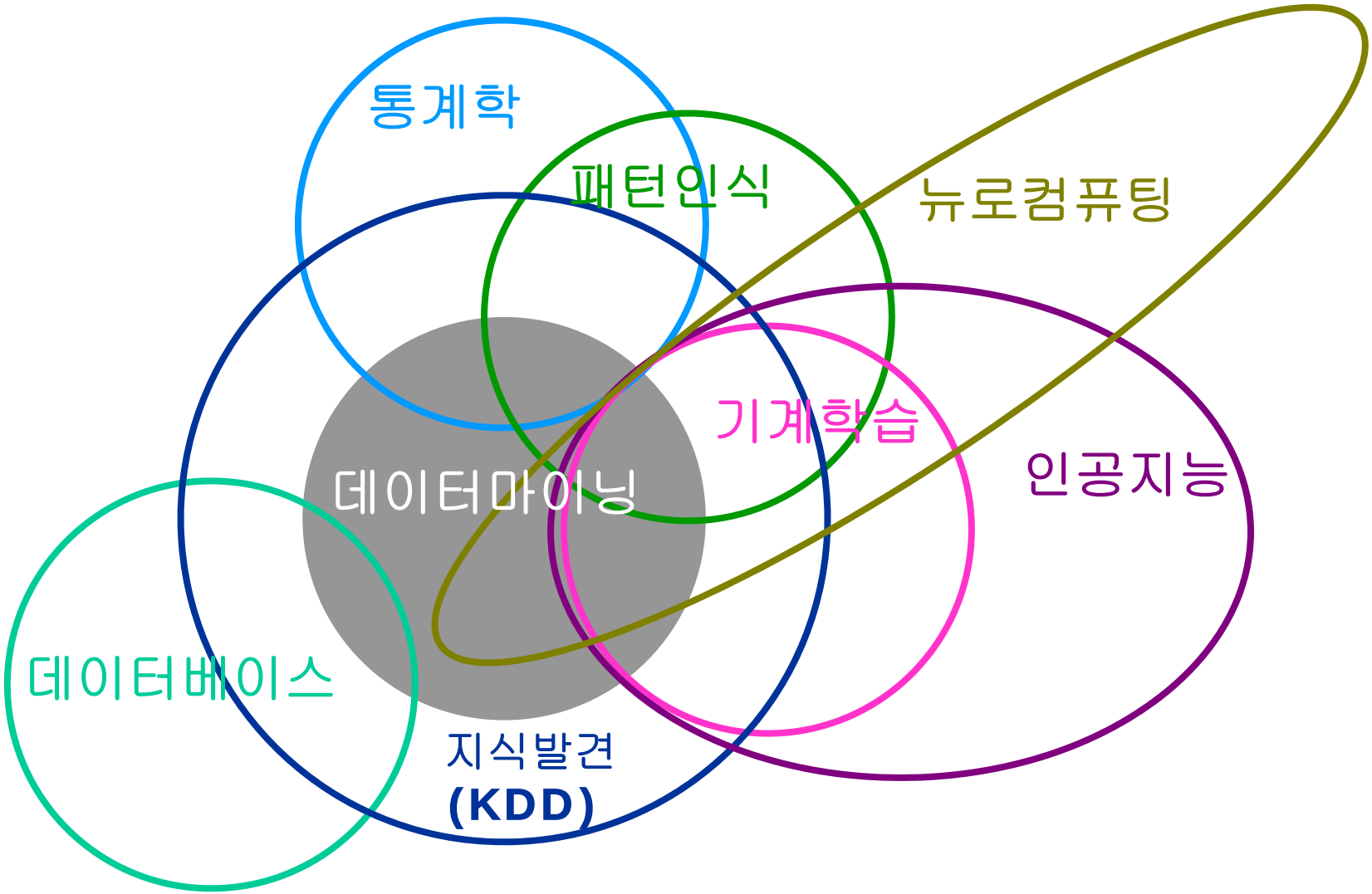
데이터마이닝

Chapter 1.



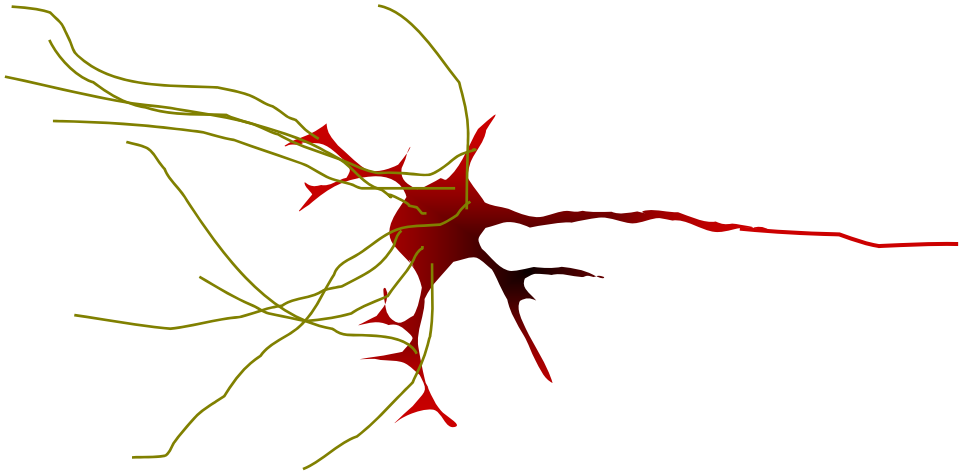
데이터마이닝과 연관된 분야들

데이터마이닝의 연구분야

- KDD (Knowledge Discovery in Databases)
- 기계학습 (Machine Learning)
- 패턴인식 (Pattern Recognition)
- 뉴로컴퓨팅 (Neurocomputing)
- 통계학 (Statistics)

뉴우런 **(neuron)**

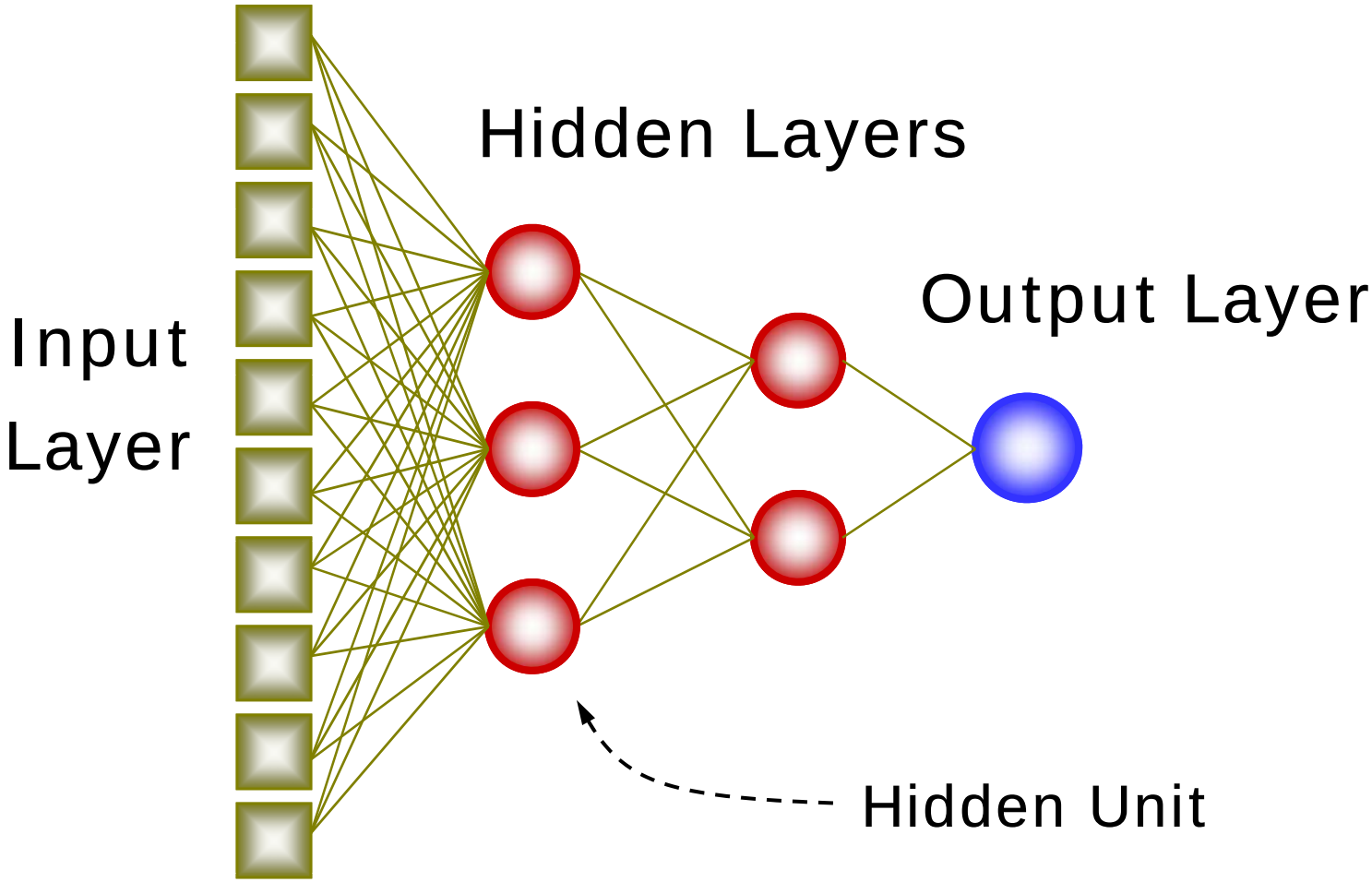
입력



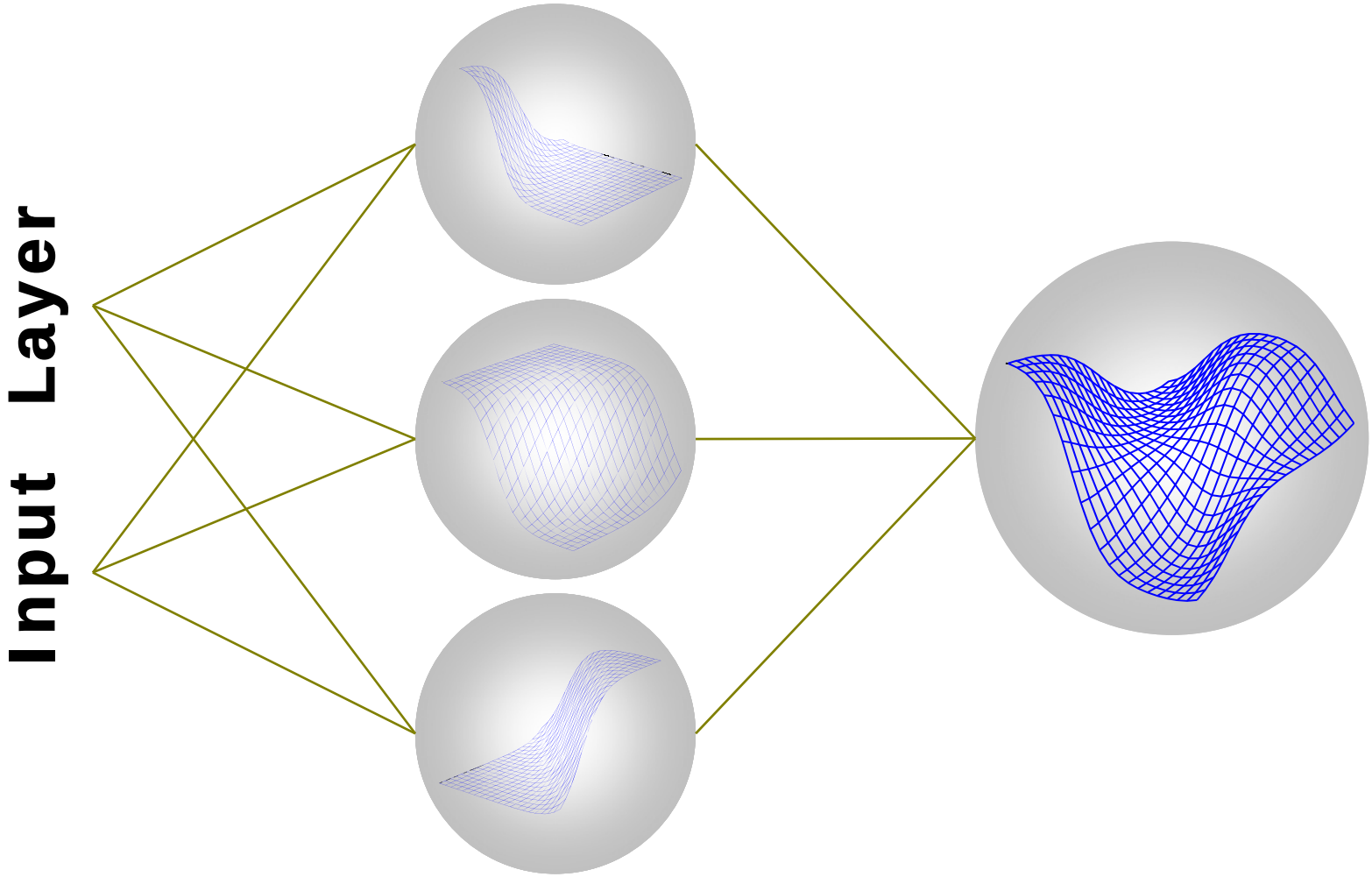
출력

인간의 신경전달 체계

신경망 구조



신경망에서의 함수추정



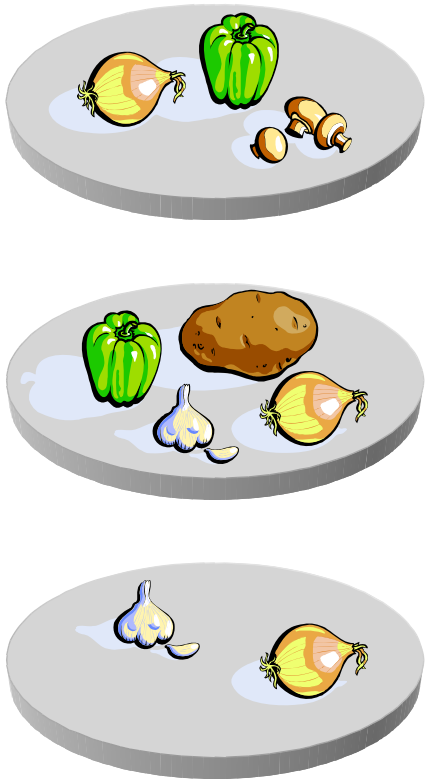
데이터마이닝의 활용분야

- 고객관계관리
- 신용평가
- 품질개선
- 부정행위 적발
- 이미지 분석
- 생명정보학

고객관계관리에서의 전략

- 목표 마케팅 (target marketing)
- 고객 세분화 (segmentation)
- 고객 이탈분석 (churn analysis)
- 교차분석 (cross selling)
- 시장바구니 분석 (market basket analysis)

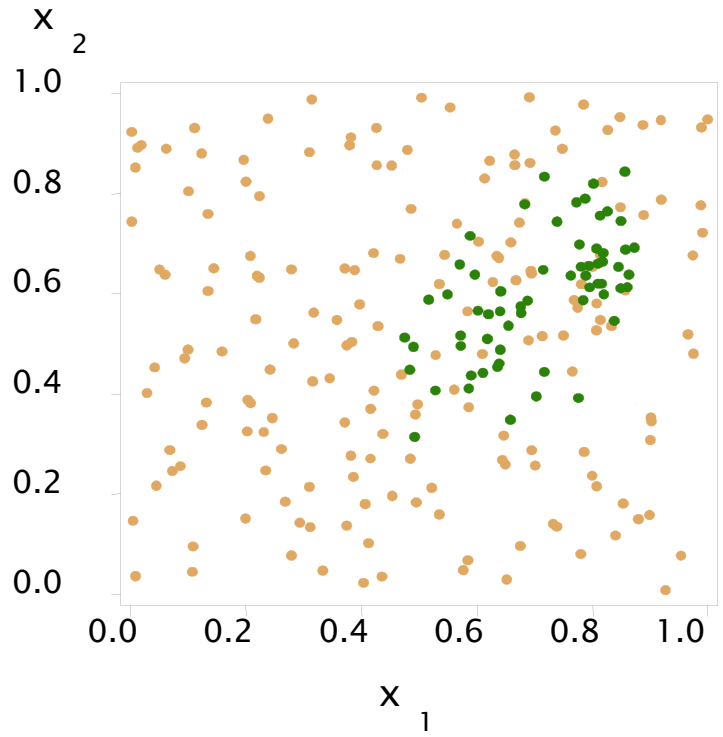
시장바구니 분석



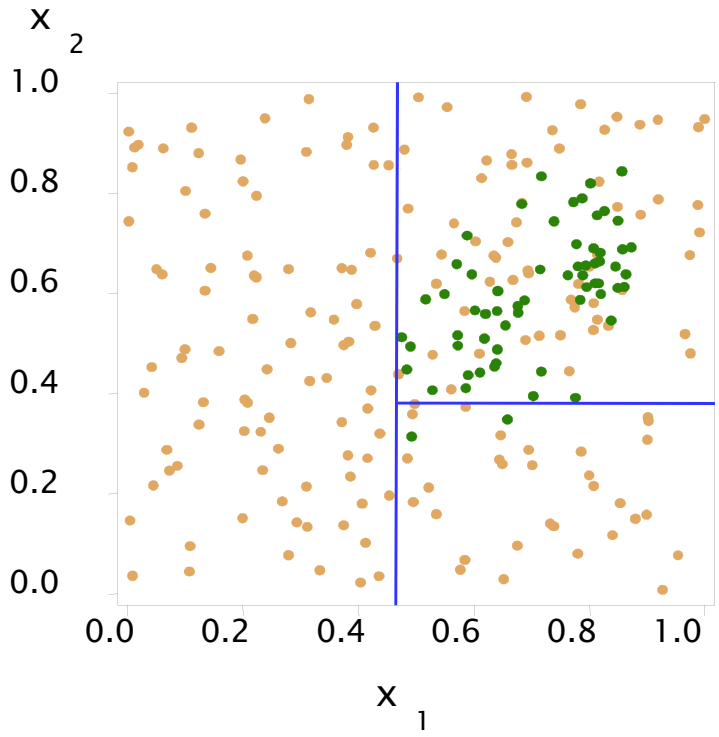
연관성 분석

Rule	Support	Confidence
A ⇒ D	2/5	2/3
C ⇒ A	2/5	2/4
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

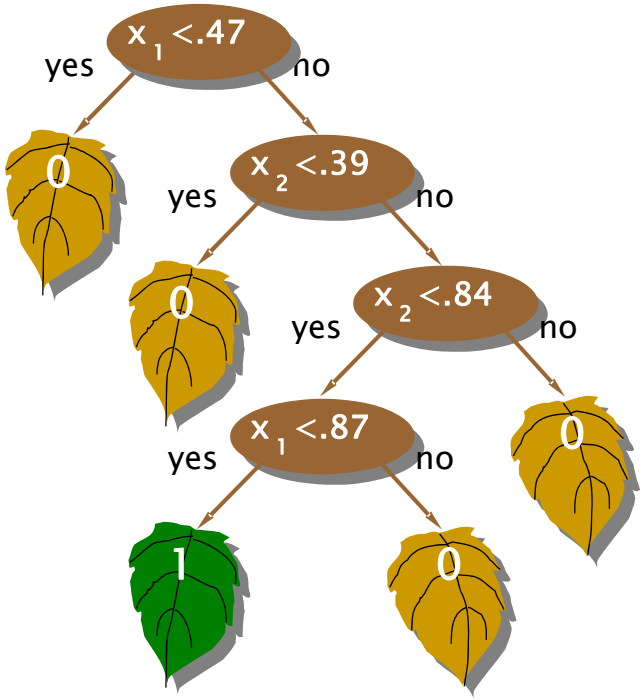
분류(classification) 대상의 자료

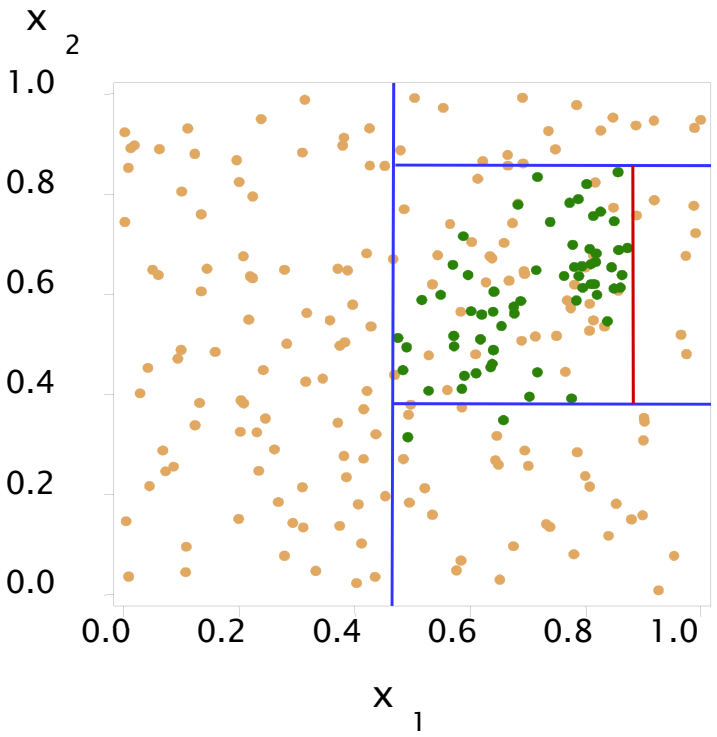


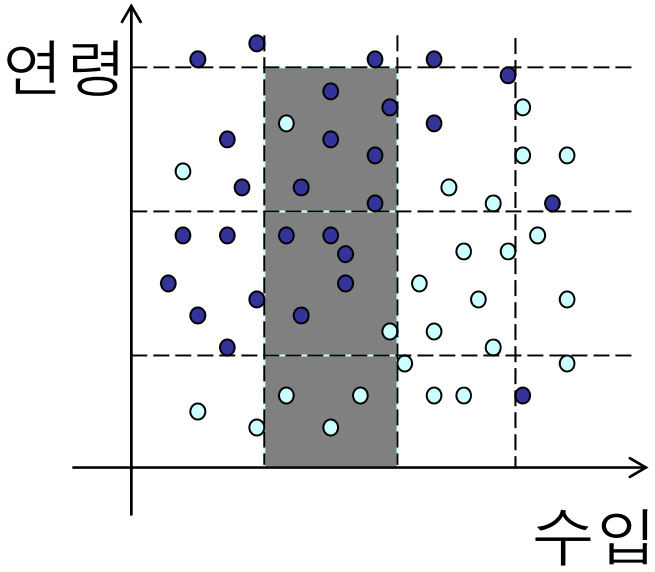
분할에 의한 분류



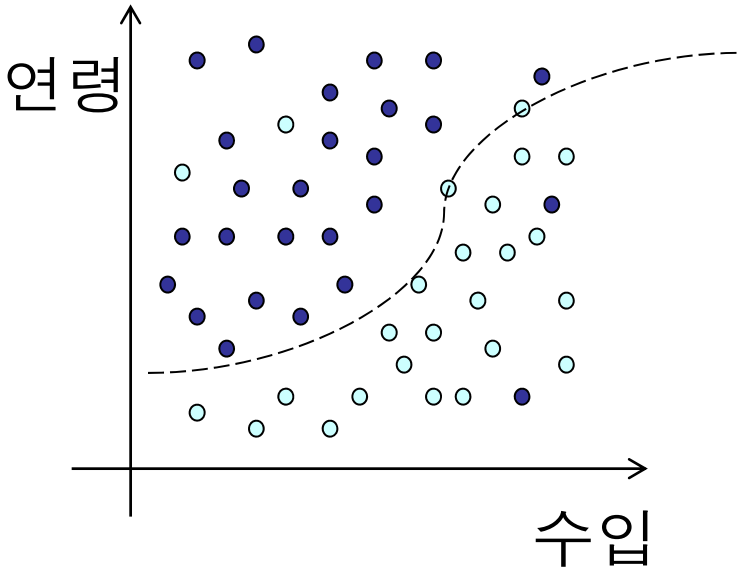
나무모형에 의한 분류


x1 <.47
yes
no
0
x2 <.39
yes
no
0
x2 <.84
yes
no
x1 <.87
yes
no
0
1
0


x2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
x1

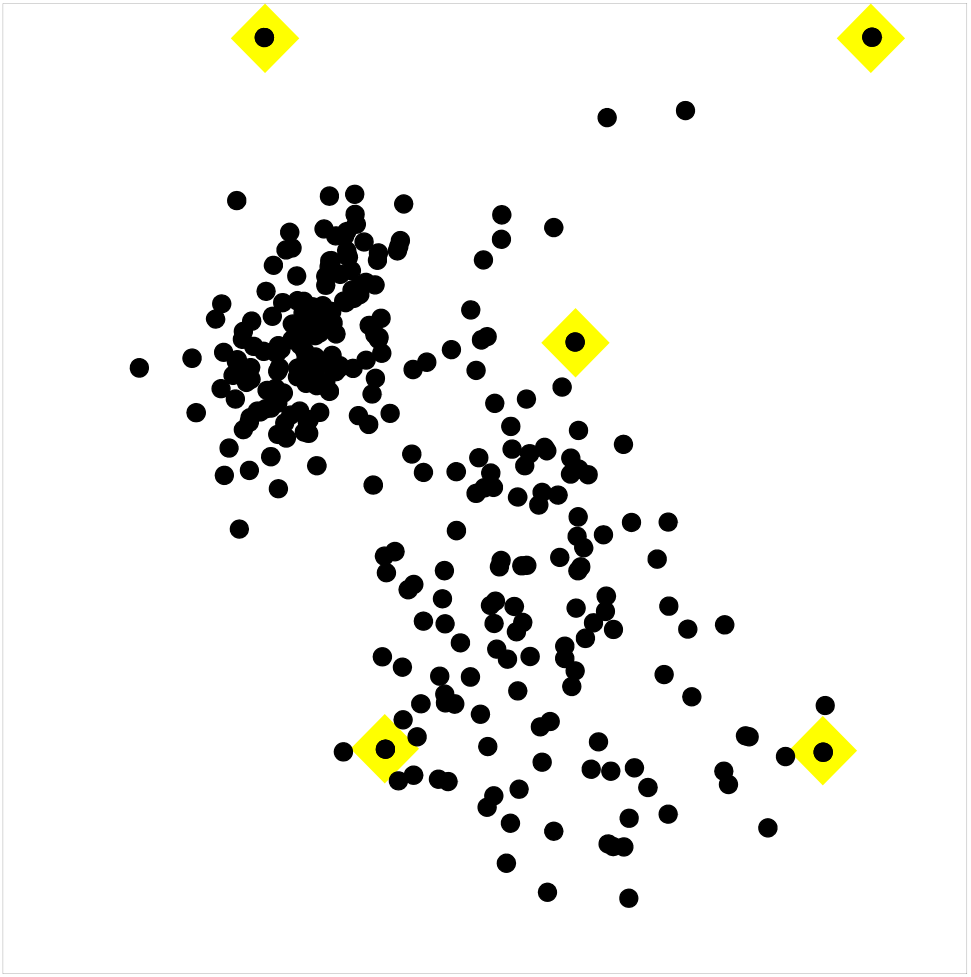


OLAP
=On-Line Analytic Processing



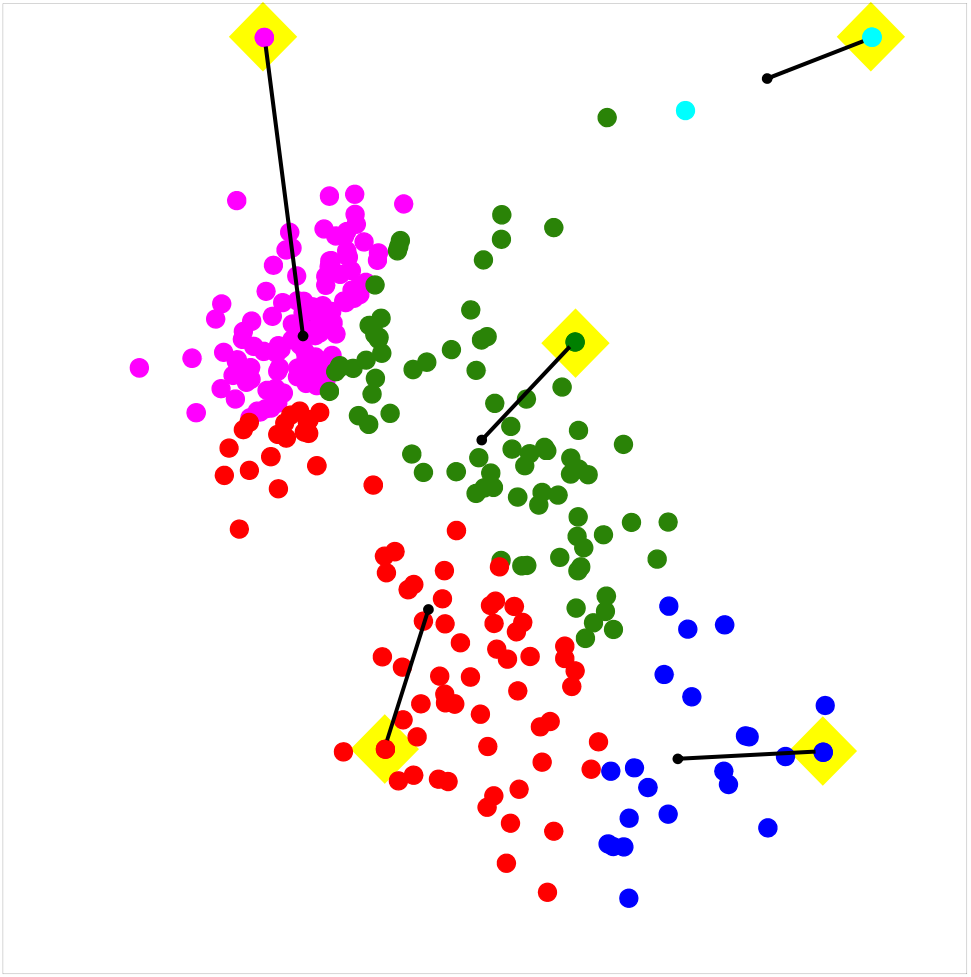
Data Mining

K-평균 군집화



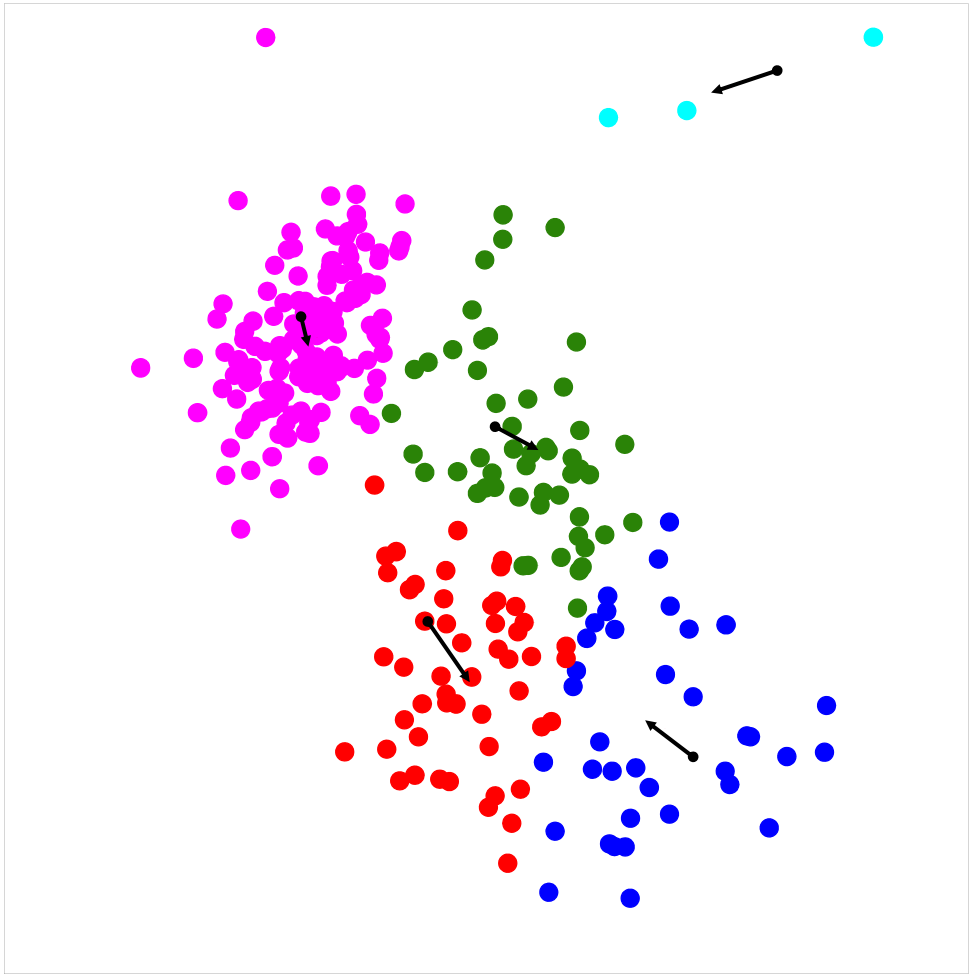
초기값 선정

K-평균 군집화



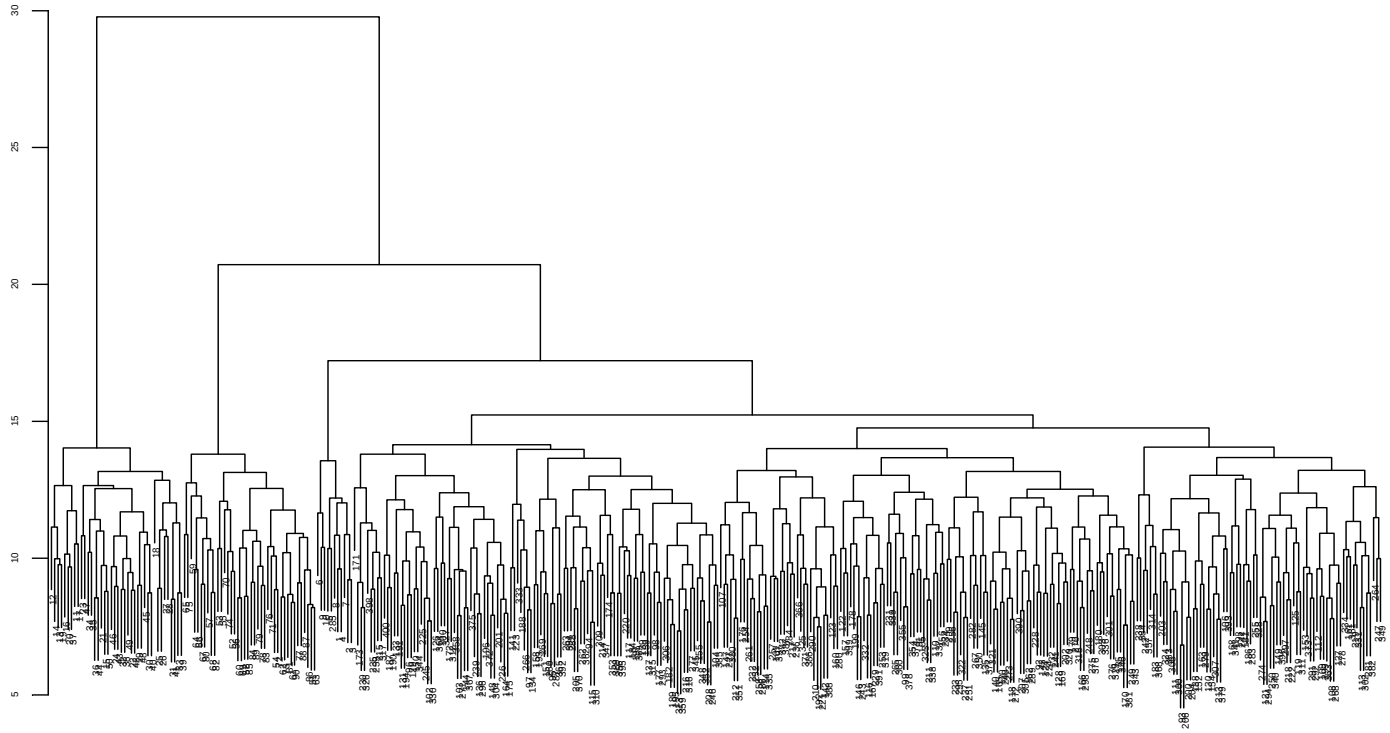
초기값의 이동

K-평균 군집화

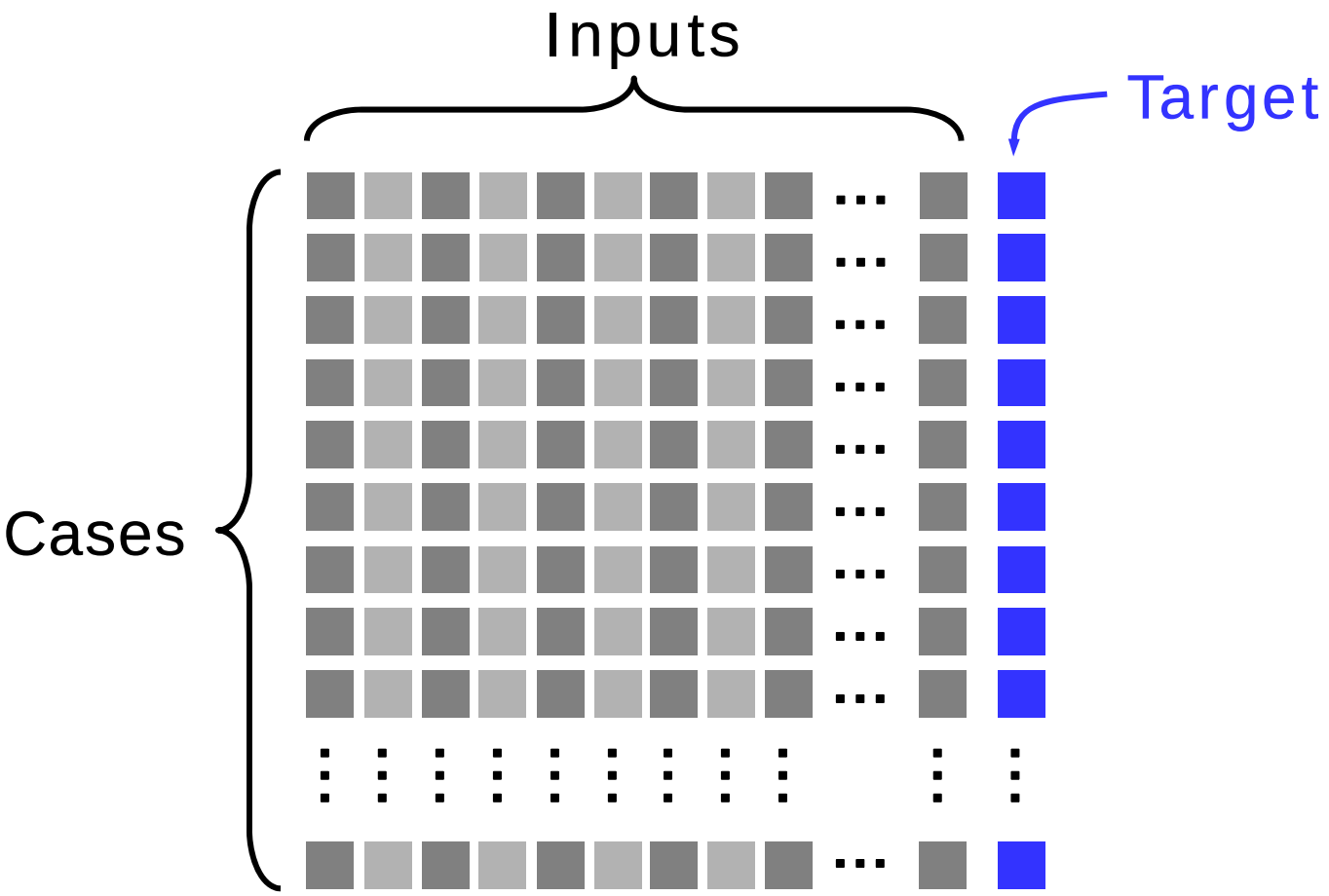


초기값의 이동 - 반복

계층적 군집화



예측 모델링을 위한 자료구조



지도학습 (Supervised Learning)

Training Data

case 1: inputs → prob → class
case 2: inputs → prob → class
case 3: inputs → prob → class
case 4: inputs → prob → class
case 5: inputs → prob → class

자율학습 (Unsupervised Learning)

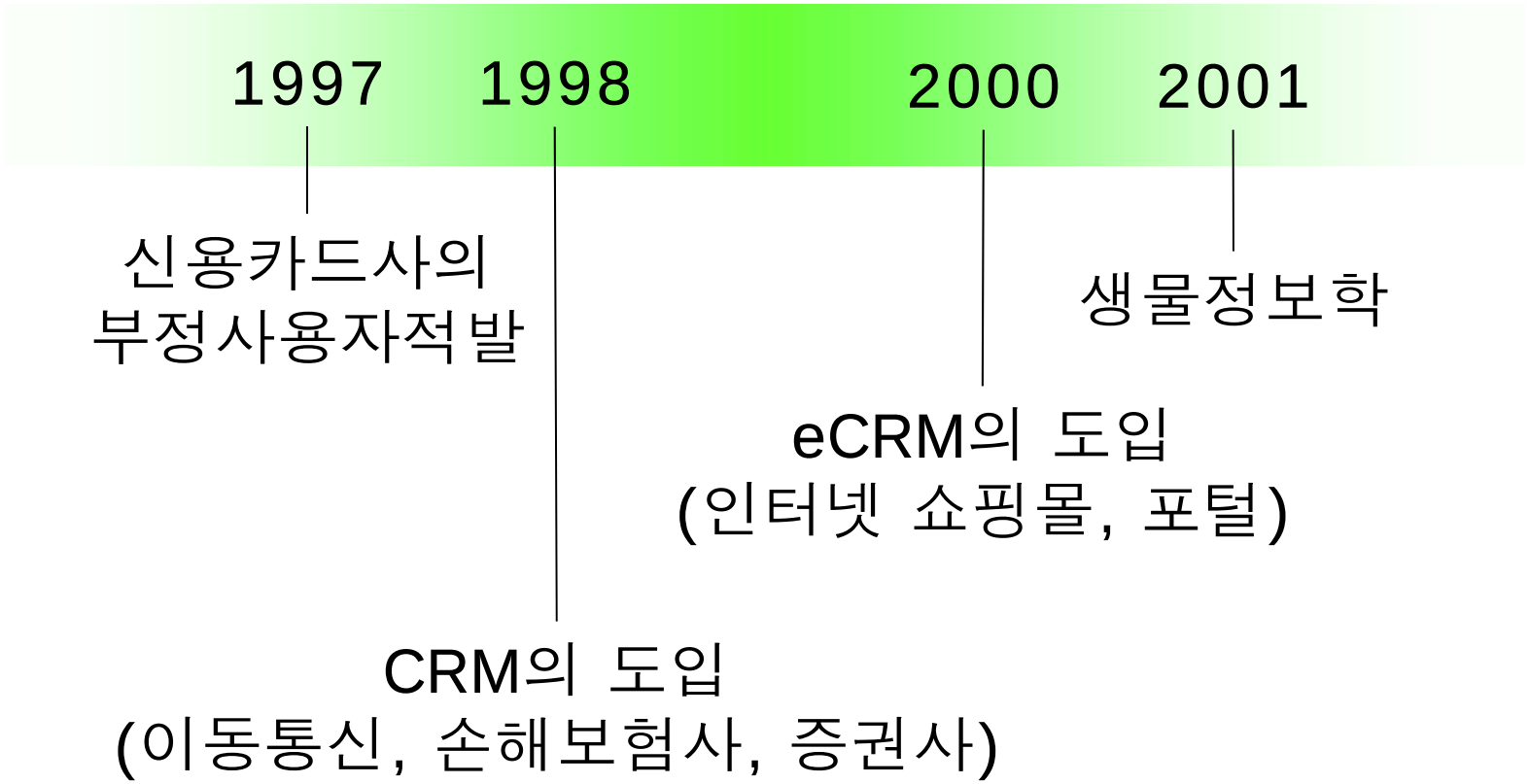
Training Data
case 1: inputs, ?
case 2: inputs, ?
case 3: inputs, ?
case 4: inputs, ?
case 5: inputs, ?

Training Data
case 1: inputs, cluster 1
case 2: inputs, cluster 3
case 3: inputs, cluster 2
case 4: inputs, cluster 1
case 5: inputs, cluster 2

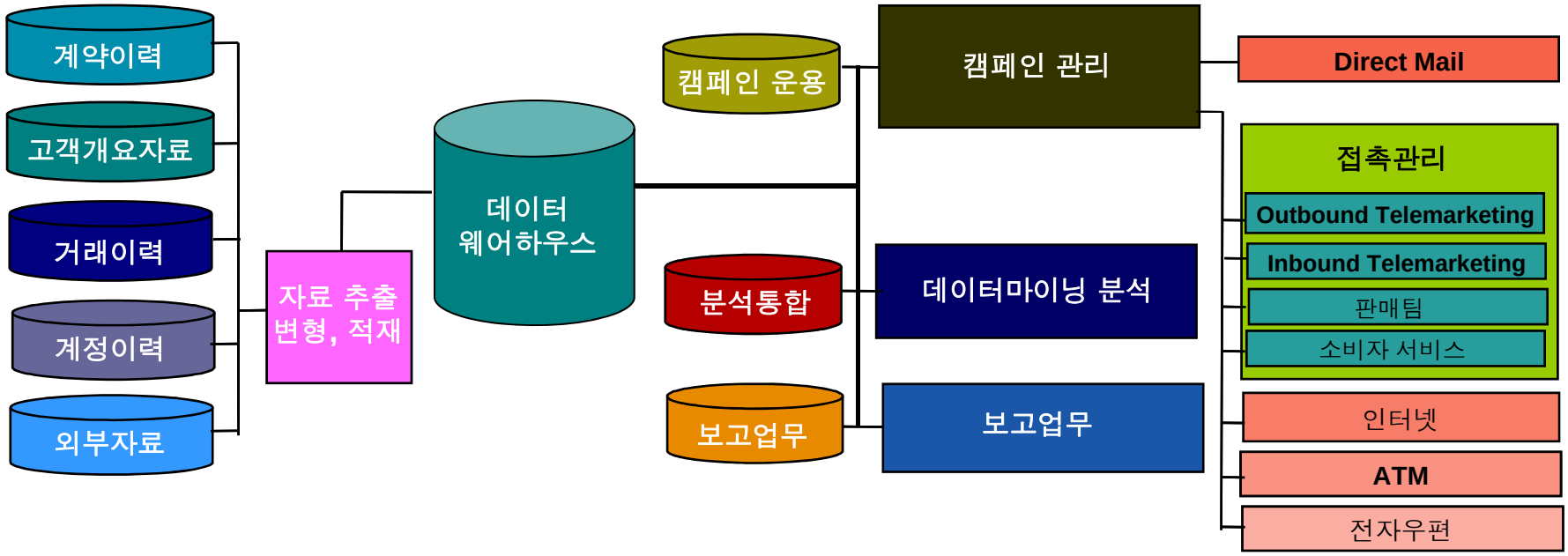
데이터마이닝 기법

- 자율학습
  - 연관성 분석
  - 군집분석
- 지도학습
  - 분류
  - 회귀분석

국내 데이터마이닝의 역사



CRM 시스템 구조


계약이력
고객개요자료
거래이력
계정이력
외부자료
자료 추출
변형, 적재
데이터
웨어하우스
캠페인 운용
분석통합
보고업무
캠페인 관리
데이터마이닝 분석
보고업무
Direct Mail
접촉관리
Outbound Telemarketing
Inbound Telemarketing
판매팀
소비자 서비스
인터넷
ATM
전자우편

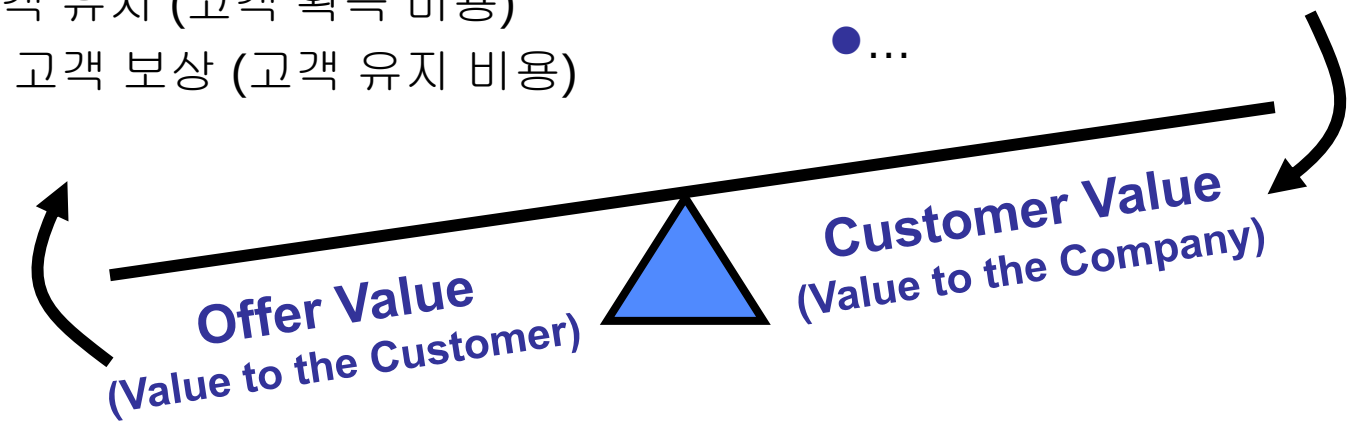
국내 CRM의 변화


현재 위치
Data Warehousing의 태동
Tool 중심의 CRM 구축
기능중심의 CRM 확산
Value-balancing 전략중심의 CRM
최적화 중심의 전략 CRM

Value Balance

- 상품 (상품 판매 비용)
- 서비스 (서비스 비용)
- 고객 유치 (고객 획득 비용)
- 대 고객 보상 (고객 유지 비용)
- …

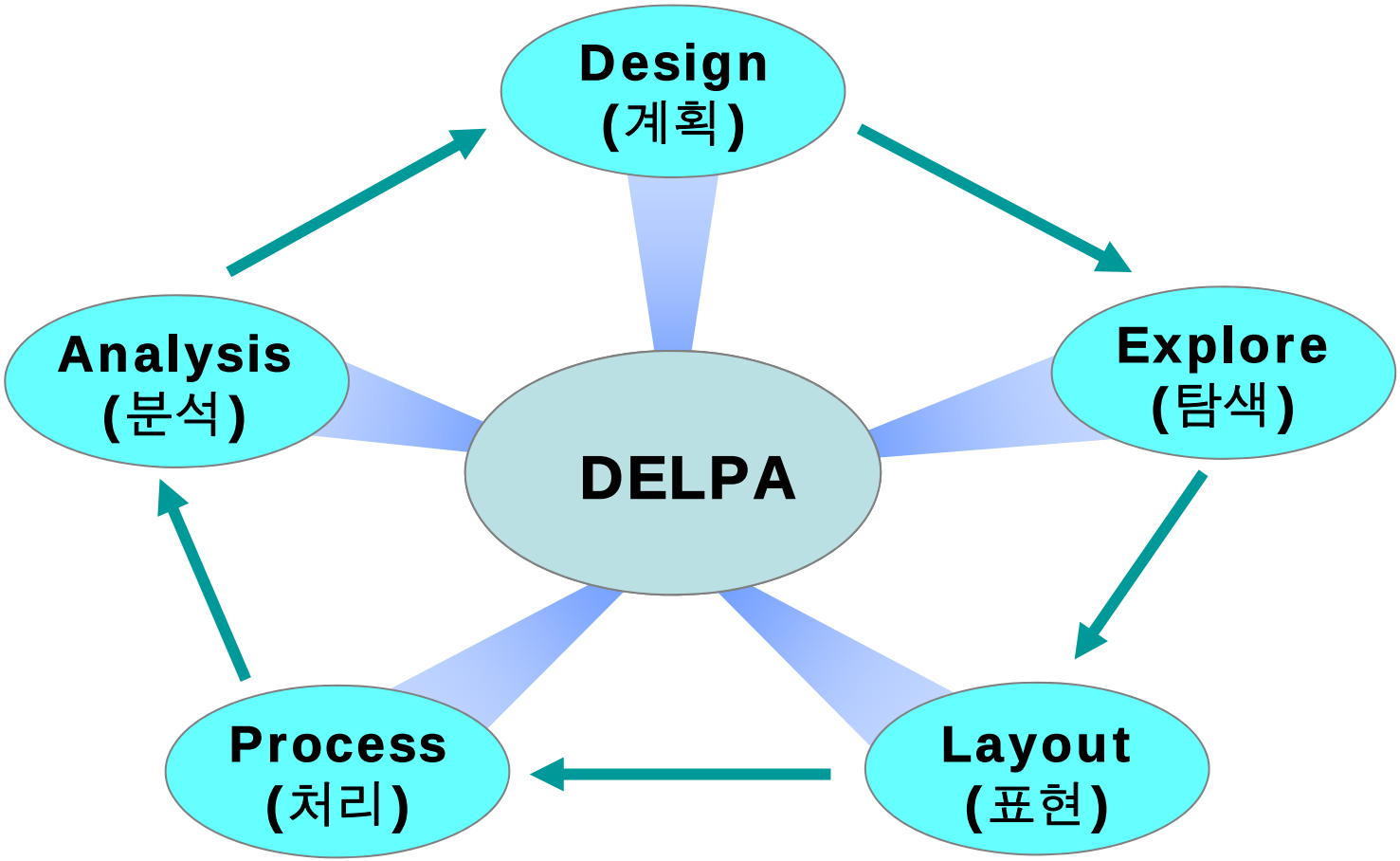
- 상품 판매로 인한 수입
- 추가 발생 수입
- 향상된 주주 가치
- …



신용카드 부정사용의 종류

- 분식 혹은 도난
- 배달사고
- 허위신청
- 카드위조
- 주변인의 사기
- 불법현금 융통

DELAP 과정



신용카드 사기적발을 위한 모델링 단계

- 단계1. 문제의 정의 및 이해
  - Design
- 단계2. 파생변수 생성 및 자료탐색
  - Exploration
  - Layout
  - Process
- 단계3. 모형적합 및 평가
  - Analysis

신용카드 사기적발을 위한 모델링 단계

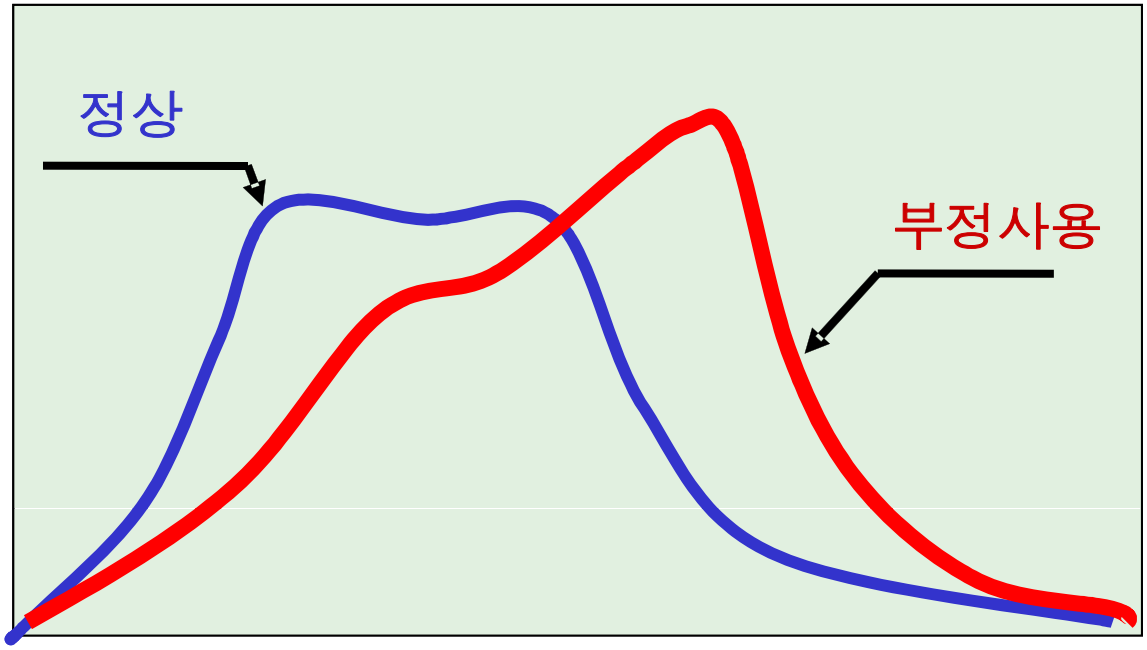
단계1. 문제의 정의 및 이해

- 현장의 전문가와 인터뷰하여 지식을 습득하라.
- 그 지식을 수량화할 수 있는 자료를 파악하라.
- 현장의 업무와 프로세스를 이해하라.
- 나 자신이 탐색 대상이 되어 취할 행동을 상상하라.

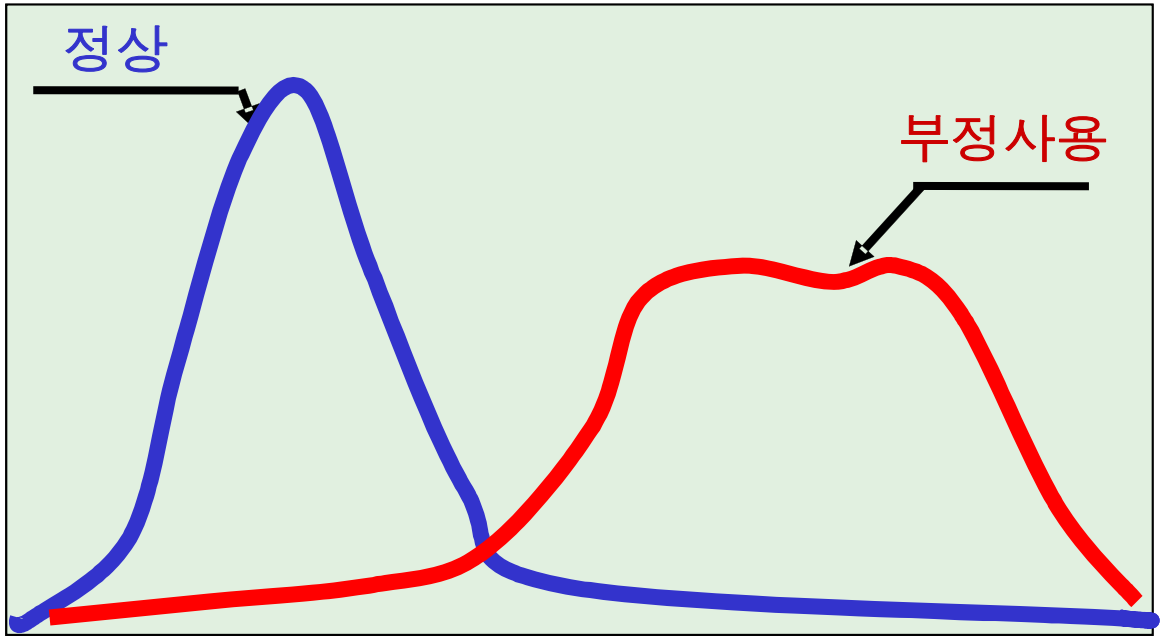
신용카드 사기적발을 위한 모델링 단계

단계2. 파생변수 생성 및 자료탐색

- 누적된 거래패턴을 표현하는 파생변수를 생성
- 정상고객을 보호할 수 있는 방안을 고려
- 세그먼트 별로 모형을 도출하고 적용



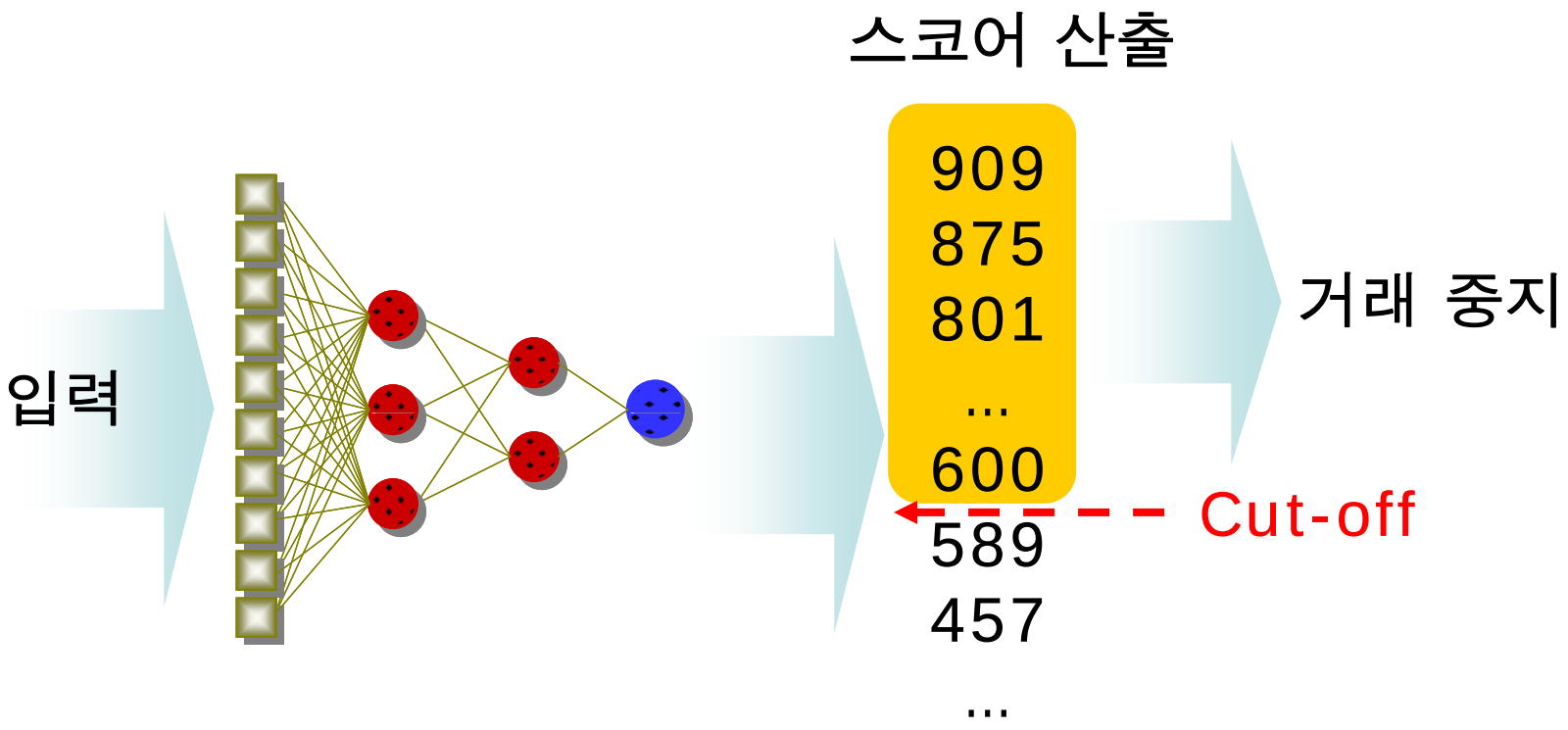
승인금액의 로그 값에 대한 분포

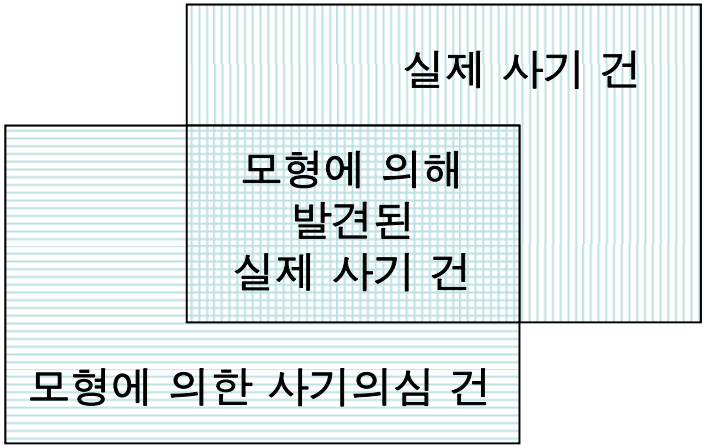


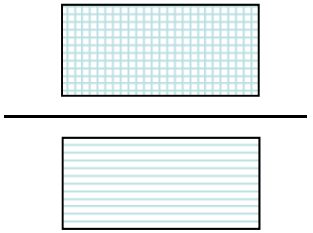
당일 사용횟수가 **4**회 이상인 경우

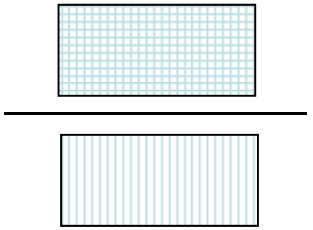
신용카드 사기적발을 위한 모델링 단계

단계3. 모형적합 및 평가

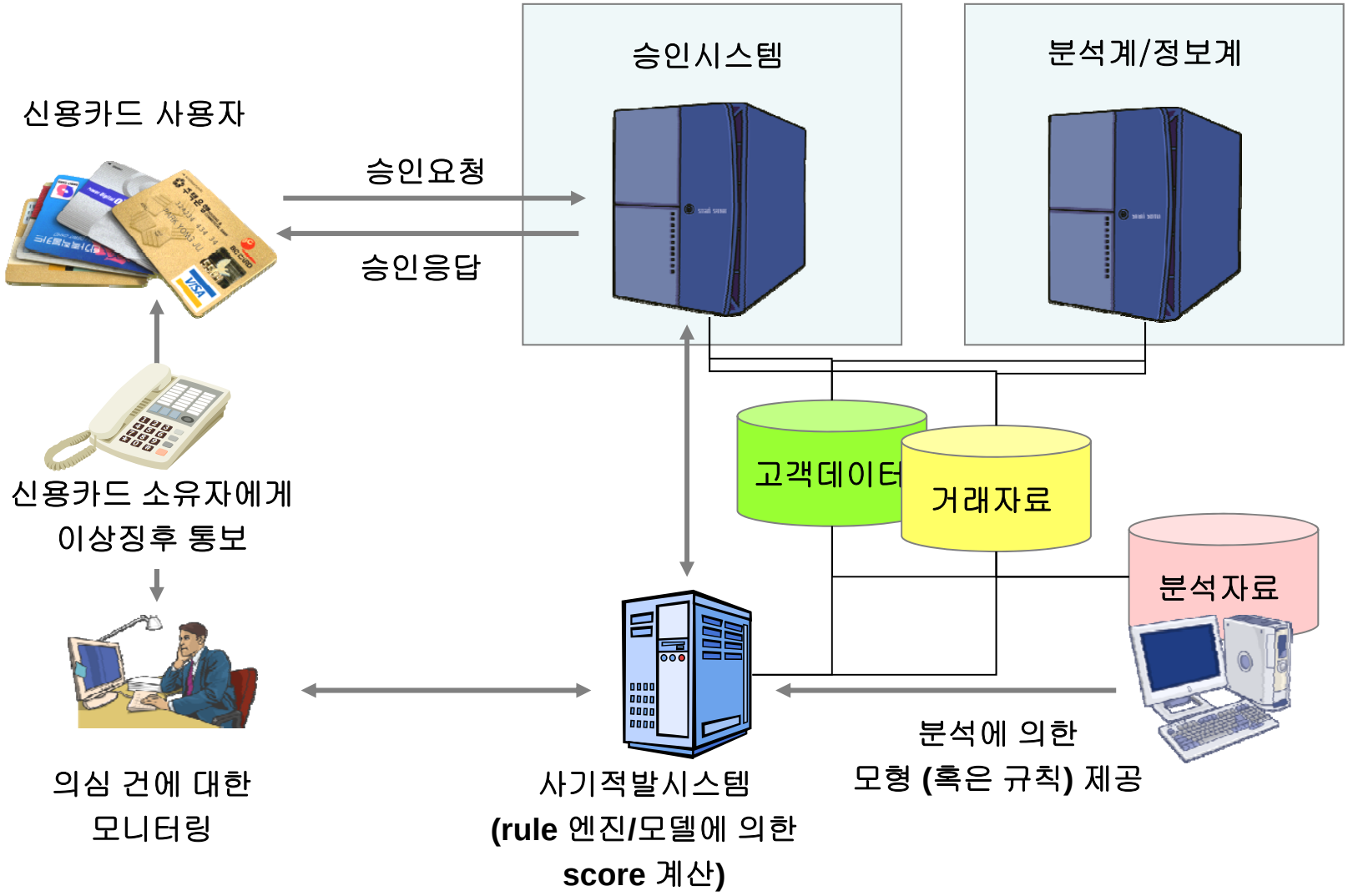




검출효율 = 

검출력 = 

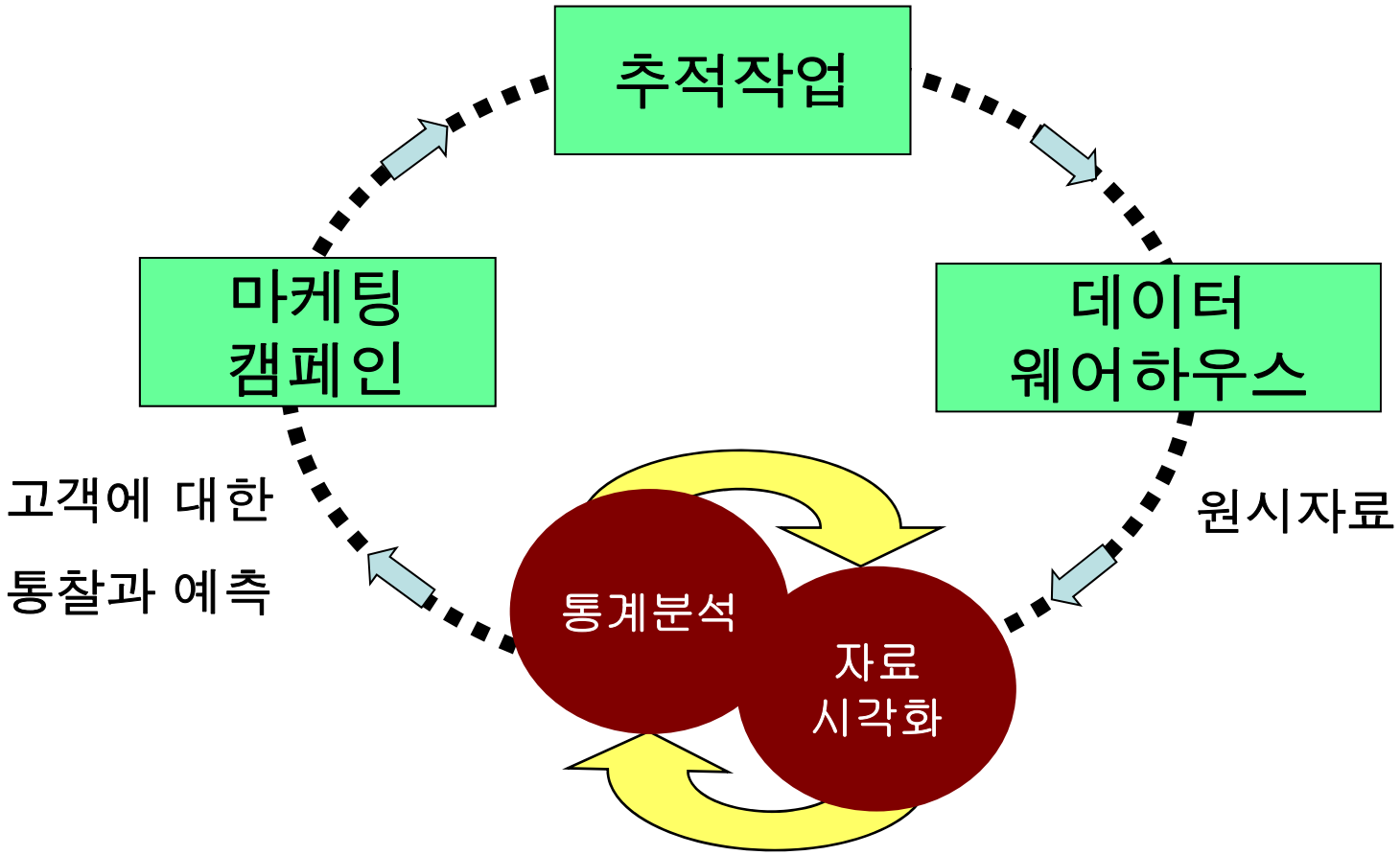
검출효율과 검출력



신용카드 사기거래 적발을 위한 시스템 구성도

시장 및 마케팅의 변화

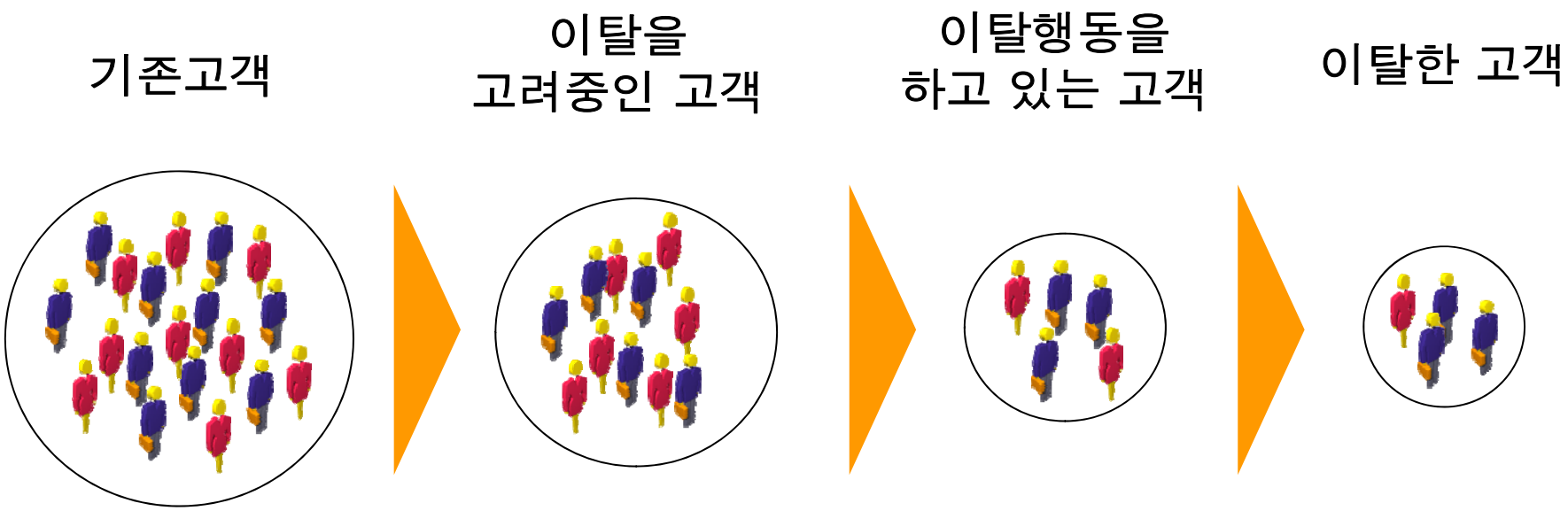
분석 중심의 CRM 과정



CRM 분석을 이용한 마케팅 전략

- 고객확보 전략 (acquisition)
- 이탈방지 전략 (churn management)
- 되찾기 전략 (win-back)
- 교차판매 전략 (cross-selling)

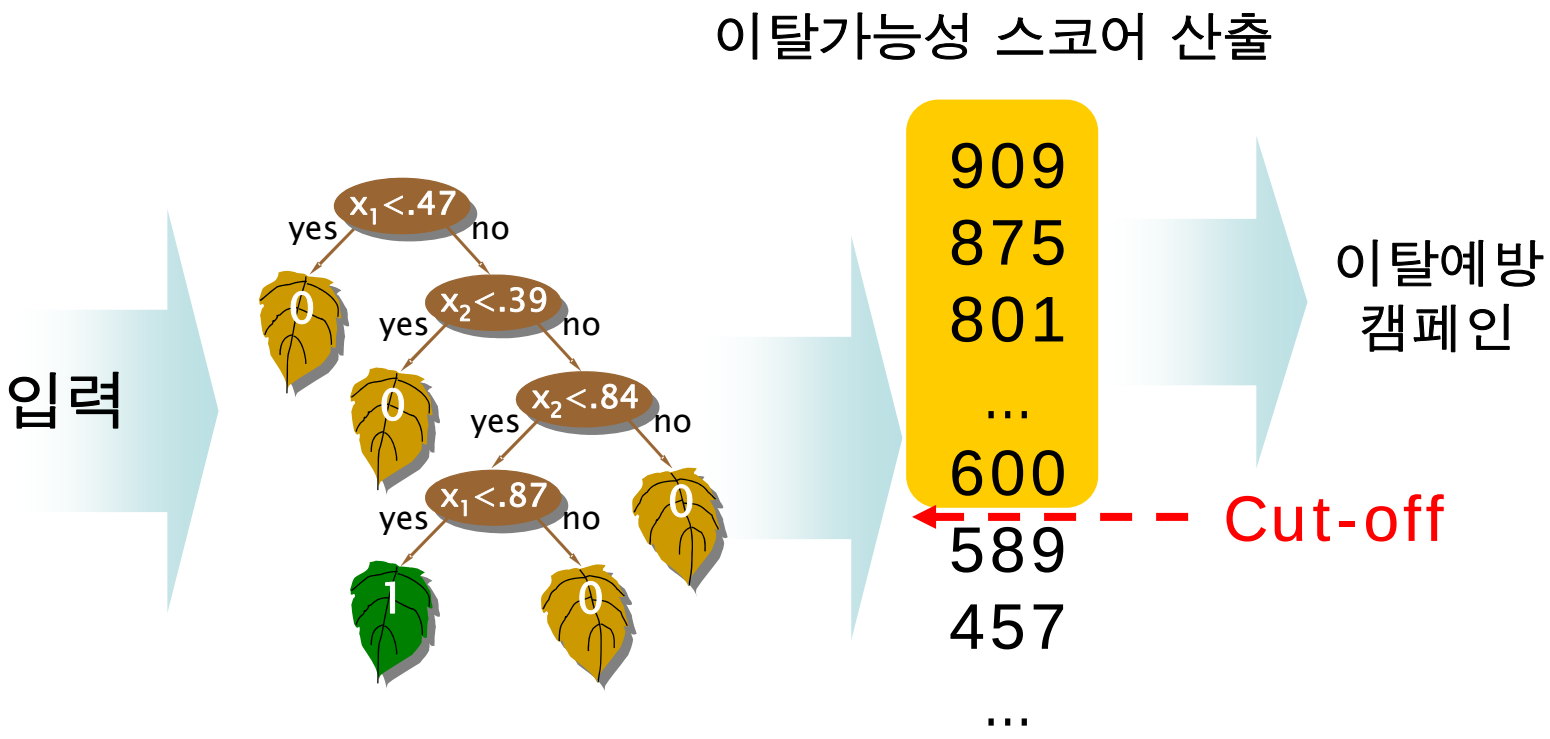
이동통신사의 고객이탈방지



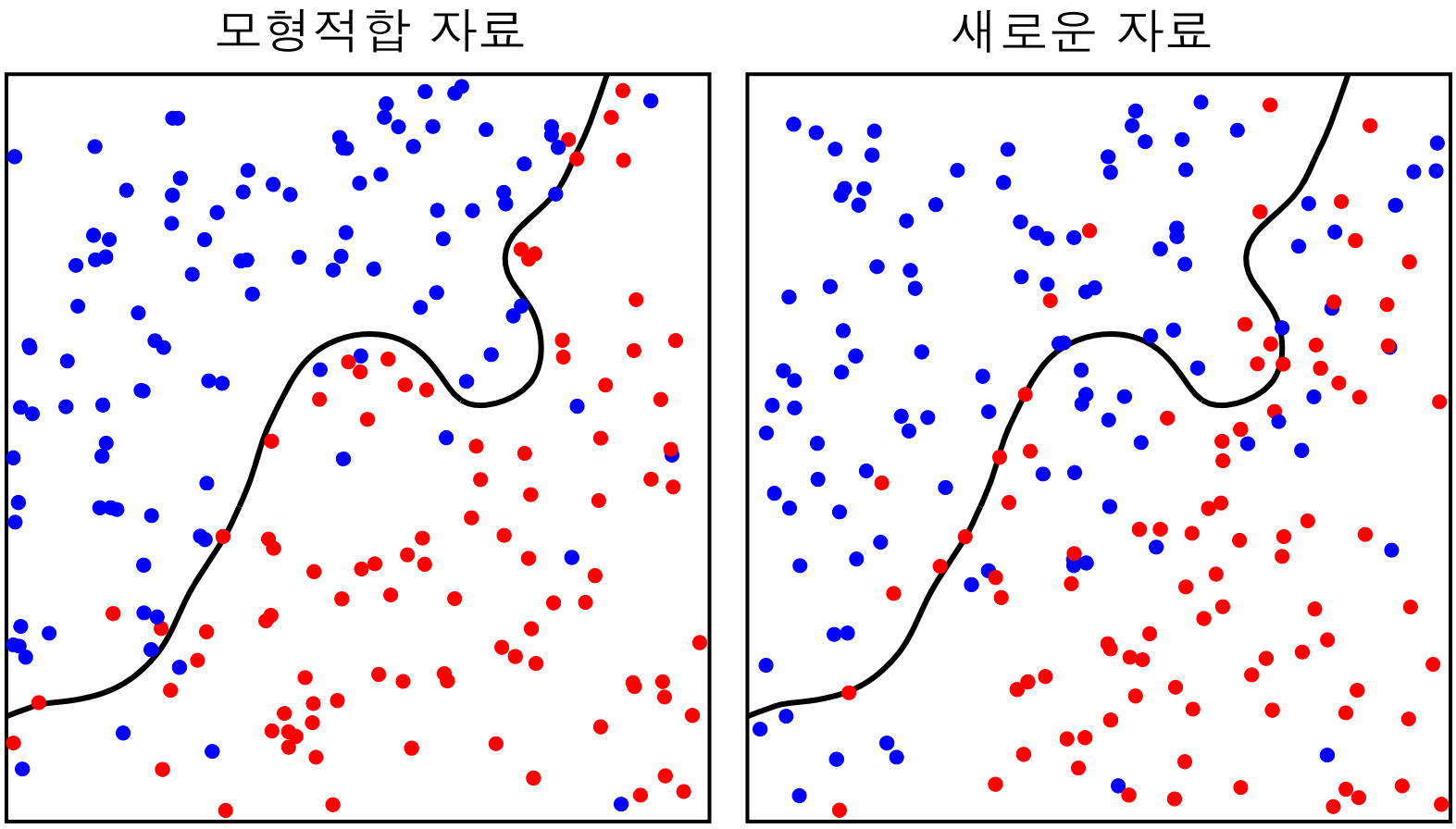
이탈고객의 행동패턴

이동통신사의 고객이탈방지

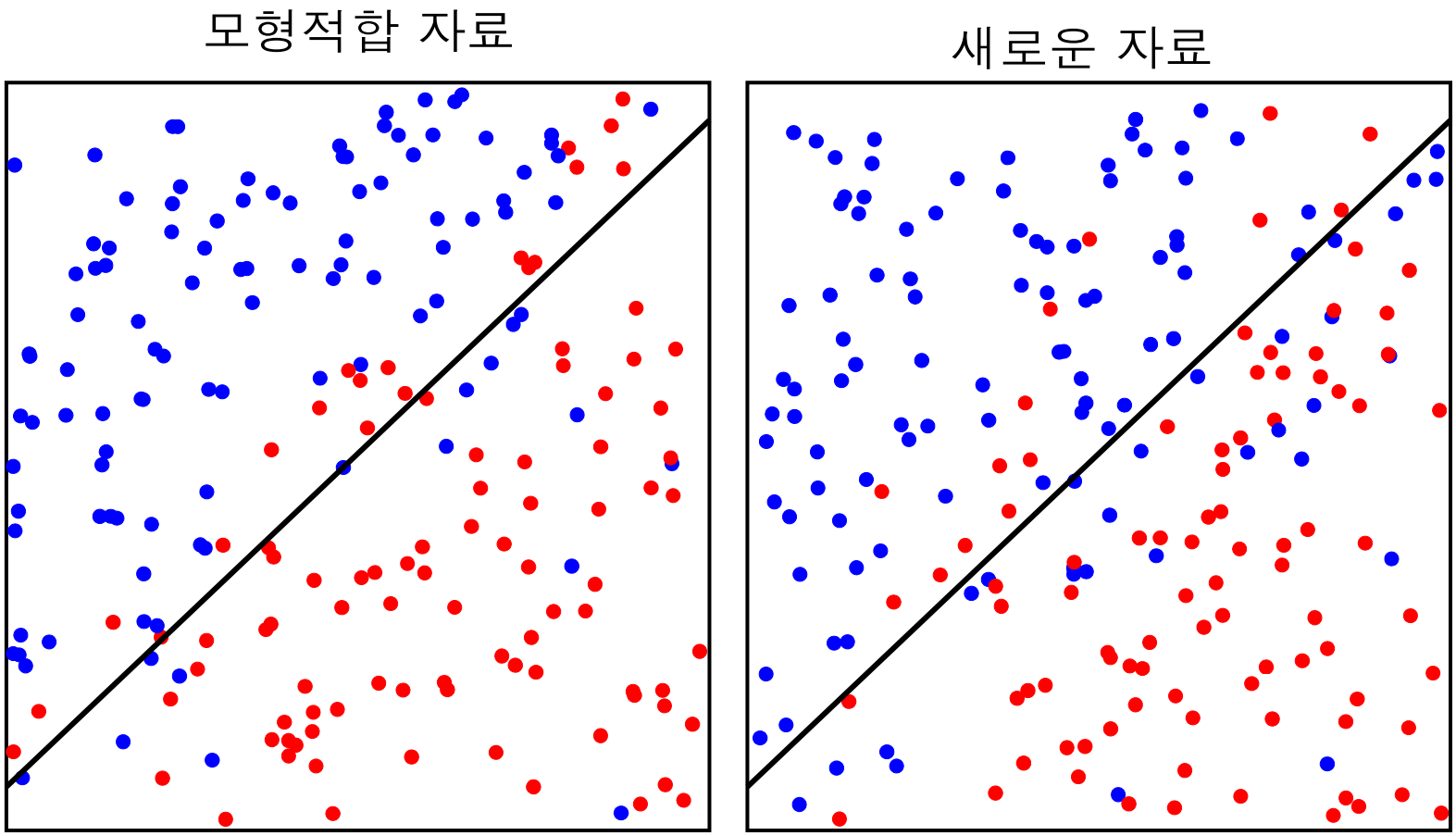
모형의 활용



과대적합



향상된 적합



다차원의 저주

- 각 변수가 0과 1 사이에 고루 분포하고 총 변수가 10개라고 하자.

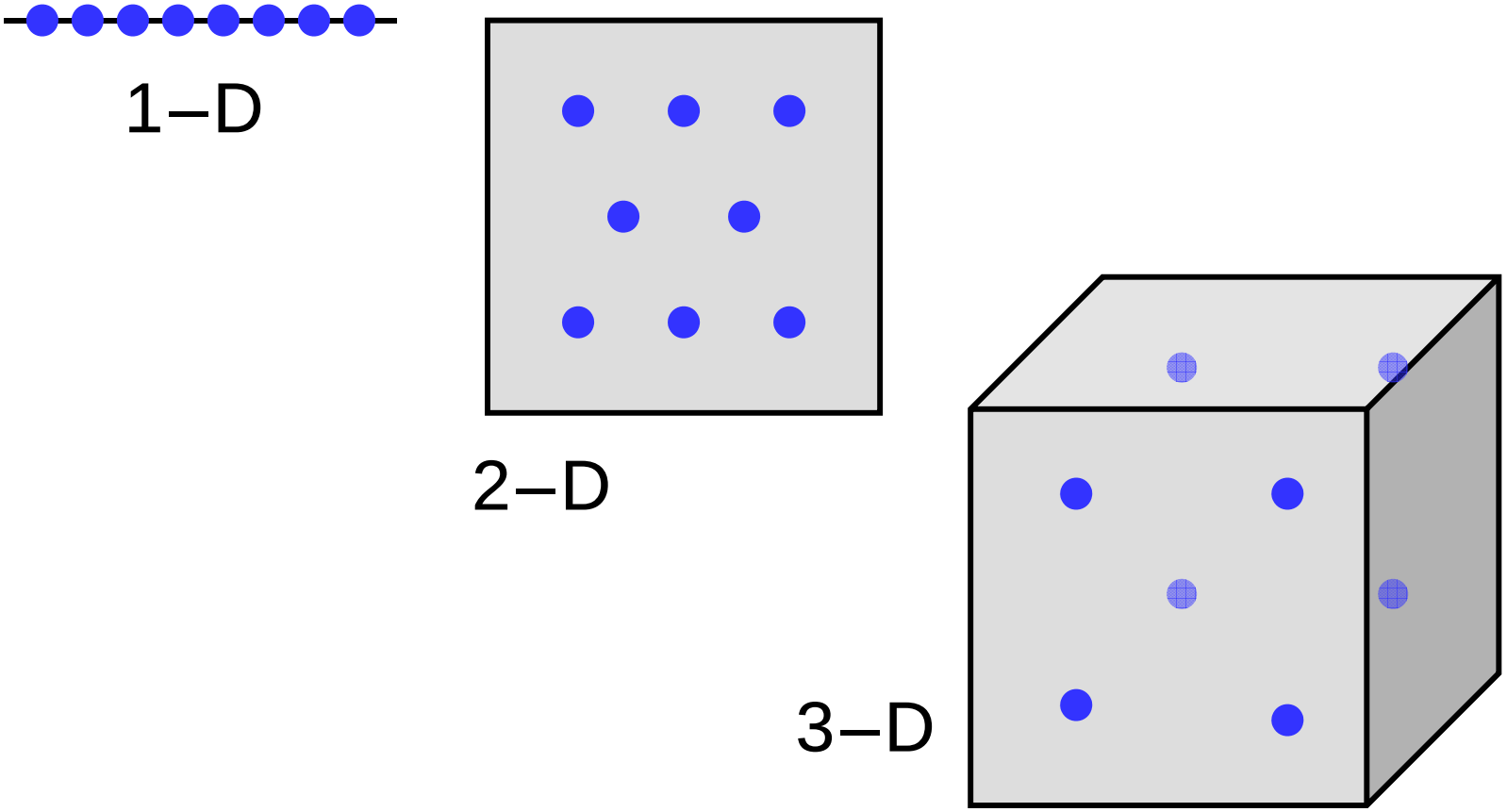
- Question
  - 만약 전체 자료 중 5%를 관찰하고 싶다면 각 변수별 필요한 범위는 얼마나 될까?

- Answer
  - 변수의 범위를 x라고 하면 다음과 같이 표현된다.

$$x^{10} = 0.05$$

  - 결국 필요한 한 변의 길이는 0.74가 된다.

“다차원의 저주”의 이해



개인신용평가시스템이란?

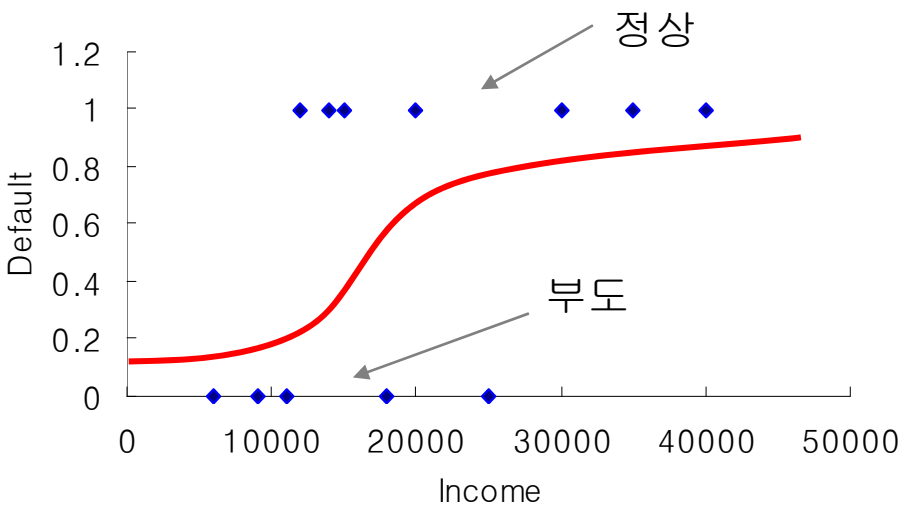
- 개인에 대한 대출의 실행여부, 대출 한도 등을 과학적인 분석과정을 결정하는 리스크 관리 시스템
- CSS (Credit Scoring System) 사용처
  - 은행
  - 신용카드사
  - 캐피탈사

개인신용평가시스템에서 스코어링의 원리

Step 1: 기존고객으로부터 자료를 수집한다.
(불량고객이면 Y = 0, 아니면 Y=1)

Step 2: 모형일 적용

$Y = f$(수입, 직업 등.)



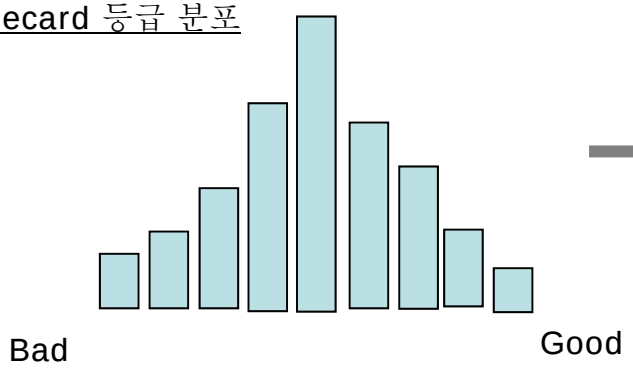
Step 3: 신규 대출신청자에 대하여 불량고객일 가능성을 계산한다.

평점표 (scorecard)

특성				
나이	30대	40대	50,60대	70대 이상
	16	20	25	30
부동산 소유	5억 이하	30억 이하	30억 이상	
	10	20	25	
거래기간	6개월 이하	2년 이하	2년 이상	
	25	0	-8	
평균잔액	A	B	C	D
	34	45	0	-5

스코어 등급에 따른 부도확률 추정

Scorecard 등급 분포



등급별 PD 추정

등급	부도확률
1	0.03
2	0.17
….	……