

빅데이터 분석을 위한

**데이터마이닝 방법론**

*SAS Enterprise Miner 활용사례를 중심으로*

# <<제1장>> 데이터마이닝의 주요 개념

## Chapter 1 Concepts of Data Mining

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현

Update: 2014. 4. 1.

## 차례

---

1.1 데이터마이닝이란 무엇인가?

1.2 데이터마이닝 프로젝트의 수행 프로세스

1.3 데이터마이닝 예측기법

1.4 Enterprise Miner의 소개

1.5 맺음말

1.6 연습문제

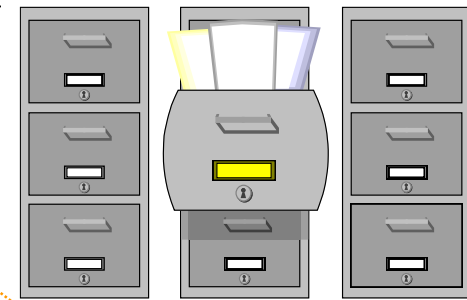
## 1.1.1 정보기술의 발달과 데이터마이닝

- 각 기업들의 운영계에는 이제 정보분석을 수행하기에 충분한 용량의 데이터가 축적되고 있다.

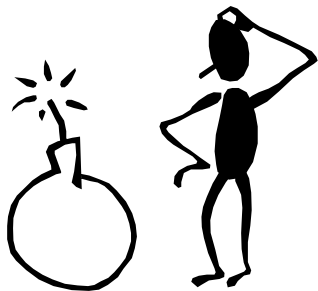
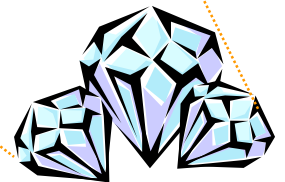


**Databases are too big**

**Terabyte**  
**=  $2^{40}$  bytes**

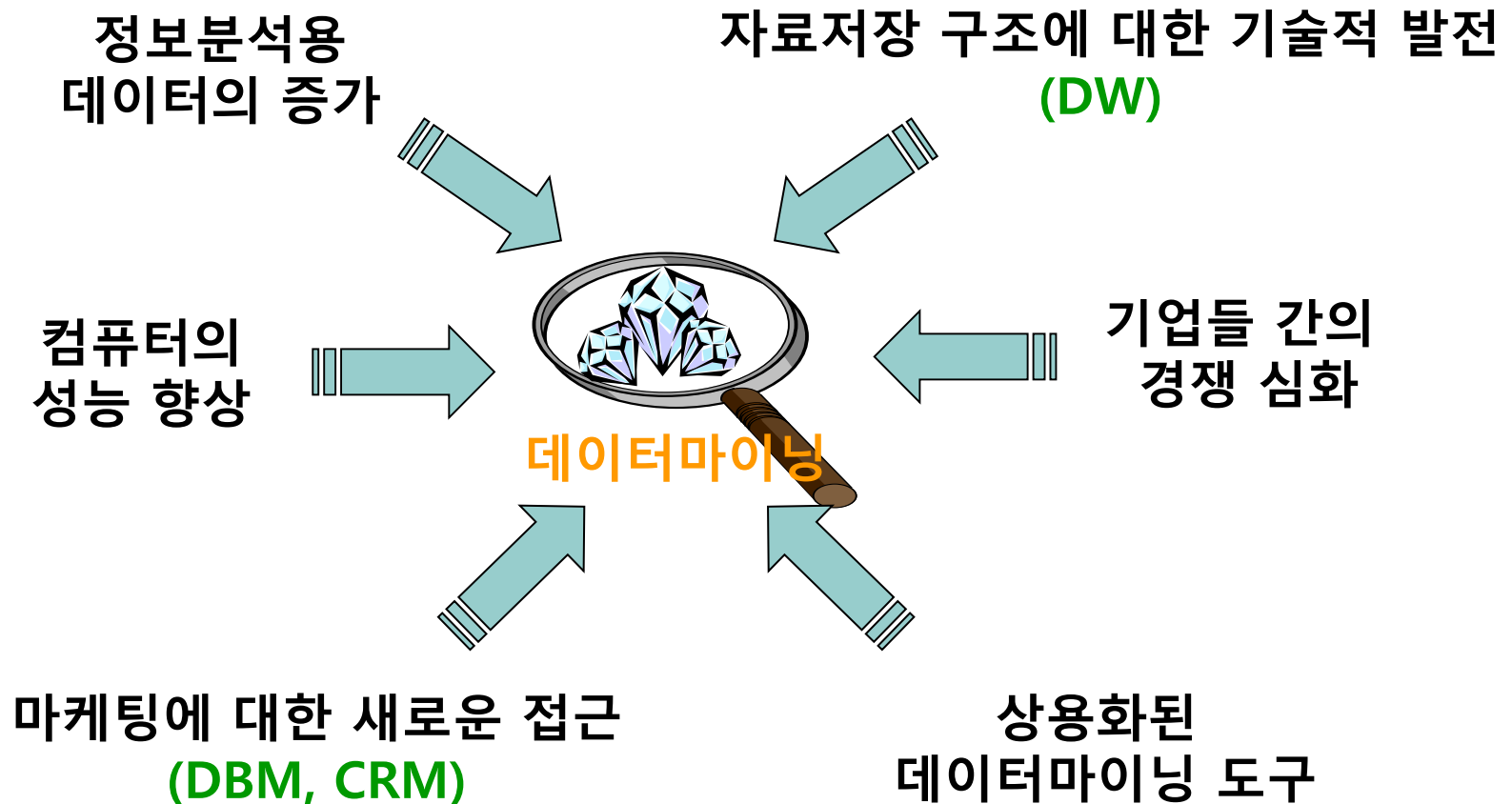


**Data Mining can help  
discover knowledge**

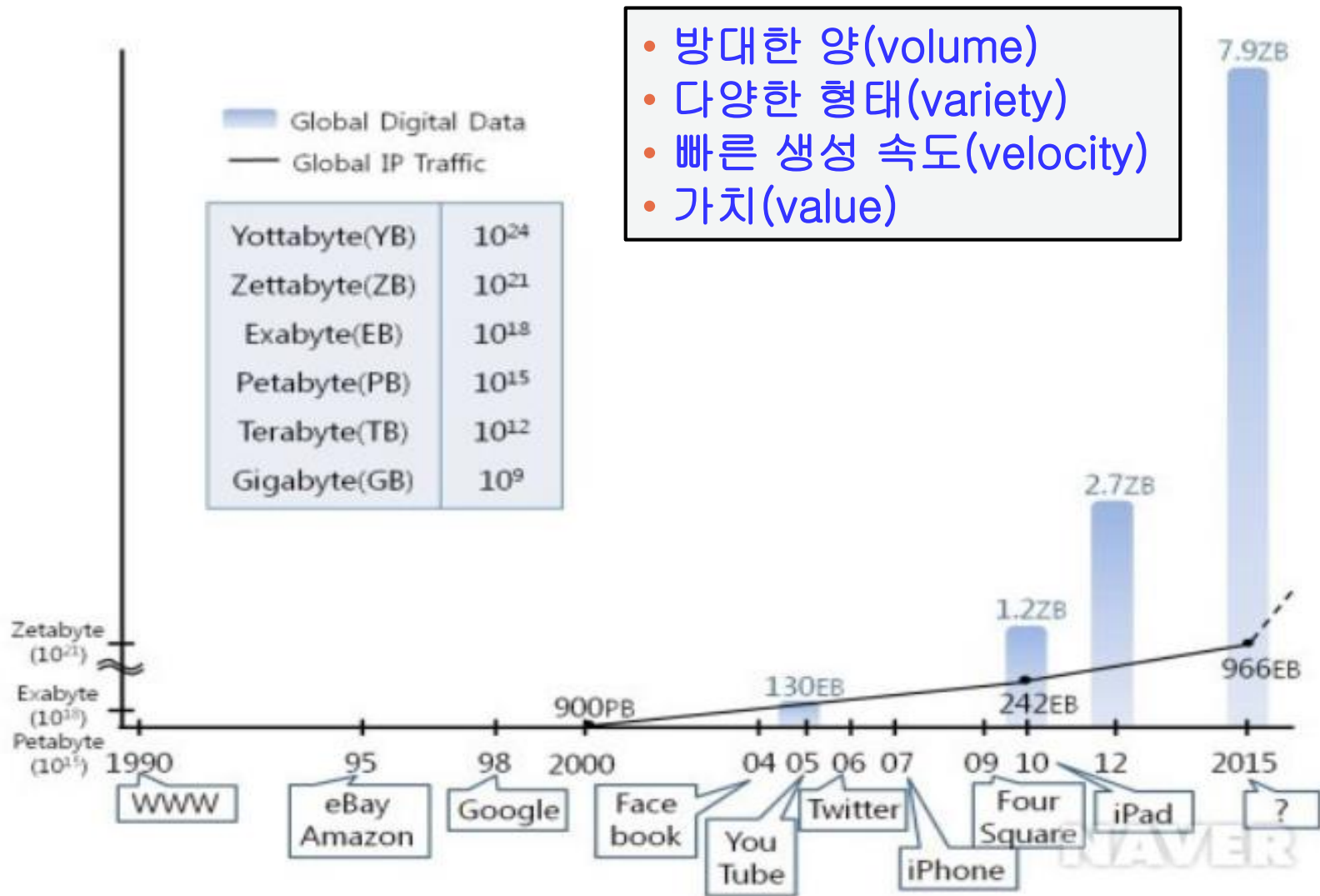


**Data Rich but Information Poor**  
**Terrorbytes**

## 데이터마이닝의 출현 배경



## 1.1.2 빅데이터(Big Data) 분석



### 1.1.3 고객관계관리(Customer Relationship Management)

- 시장의 포화 및 다자간 경쟁시대
- 고객 Needs의 증대 및 다양화
- 마케팅 매체의 다양화
- 체계적인 DB 구축
- 과학적 경영의 기업문화



Customer  
Relationship  
Management

Database  
Marketing

Data  
Mining

- ✓ 현 고객 중 이탈 가능성이 높은 고객은 누구인가?
- ✓ 현 고객 중 우량 고객들은 누구인가?
- ✓ 고객들의 상품 구매패턴은 어떠한가?
- ✓ 이탈한 고객의 이탈원인은 무엇인가?

고객 획득

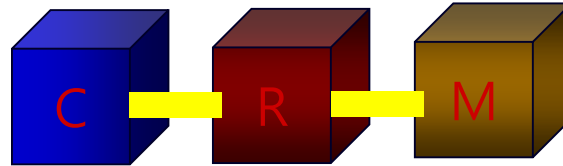


고객 이탈방지/유지



고객가치 증대

# CRM 분류



## Analytical CRM

### Extended DW or DBM

- Data Warehouse
- Data Mining
- OLAP

을 이용한 마케팅 의사 결정을 지원하는 마케팅 의사 지원 시스템(MDSS)

## Operational CRM

### Extended ERP

- ERP가 가지고 있는 기능 (거래처리, 재무, 인사 관리 등) 중 고객 접촉 관련 기능 강화
- ERP의 기능 확장 또는 CRM 모듈과 ERP를 통합
- 주로 영업과 서비스를 위한 시스템

## Collaborative CRM

### eCRM

- Internet을 기반으로 한 EC 및 Portal site의 급성장
- Offline 기업의 Online화 가속화
- Internet에 대응되는 신 개념의 CRM

# 고객관계관리(CRM)와 데이터마이닝: Analytical CRM



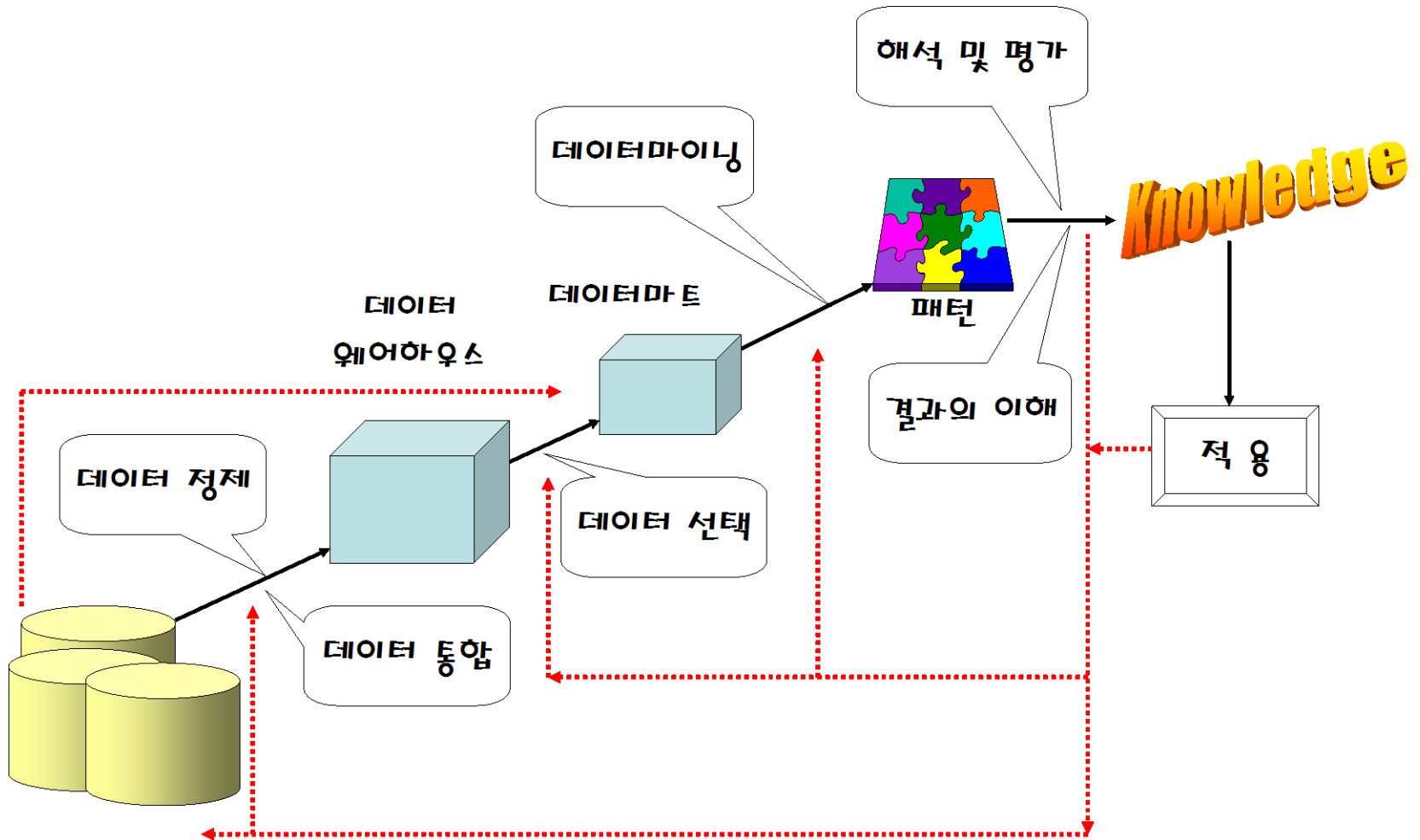


## 1.1.4 데이터마이닝 관련 분야

---

- **KDD (Knowledge Discovery in Databases)**
  - 데이터웨어하우징, 데이터마이닝 등을 포함하는 포괄적 의미
  - 데이터마이닝은 KDD(데이터베이스 지식탐색) 과정 중의 일부라고 말할 수 있다.
- **통계학**
  - 군집분석 (Cluster Analysis)
  - 판별분석 (Discrimination Analysis)
- **기계학습 (Machine Learning)**
- **패턴인식 (Pattern Recognition)**
- **뉴로컴퓨팅 (Neurocomputing) - ANN**

# 데이터베이스로 부터의 지식발견(KDD) 과정



## 1.1.5 데이터마이닝의 활용분야와 특징

---

- 데이터베이스 마케팅
  - 고객유치 (Customer Acquisition)
  - 고객유지 (Customer Retention)
  - 고객세분화 (Customer Segmentation)
  - 고객관계관리 (CRM, Churn Management)
  - 수요 및 판매 예측 (Forecasting)
  - 연관성규칙발견 (Association Rule Discovery)
  - Cross Selling / Up-Selling
  - Target Marketing
  - Telemarketing, Direct Marketing

## 데이터마이닝 활용분야

---

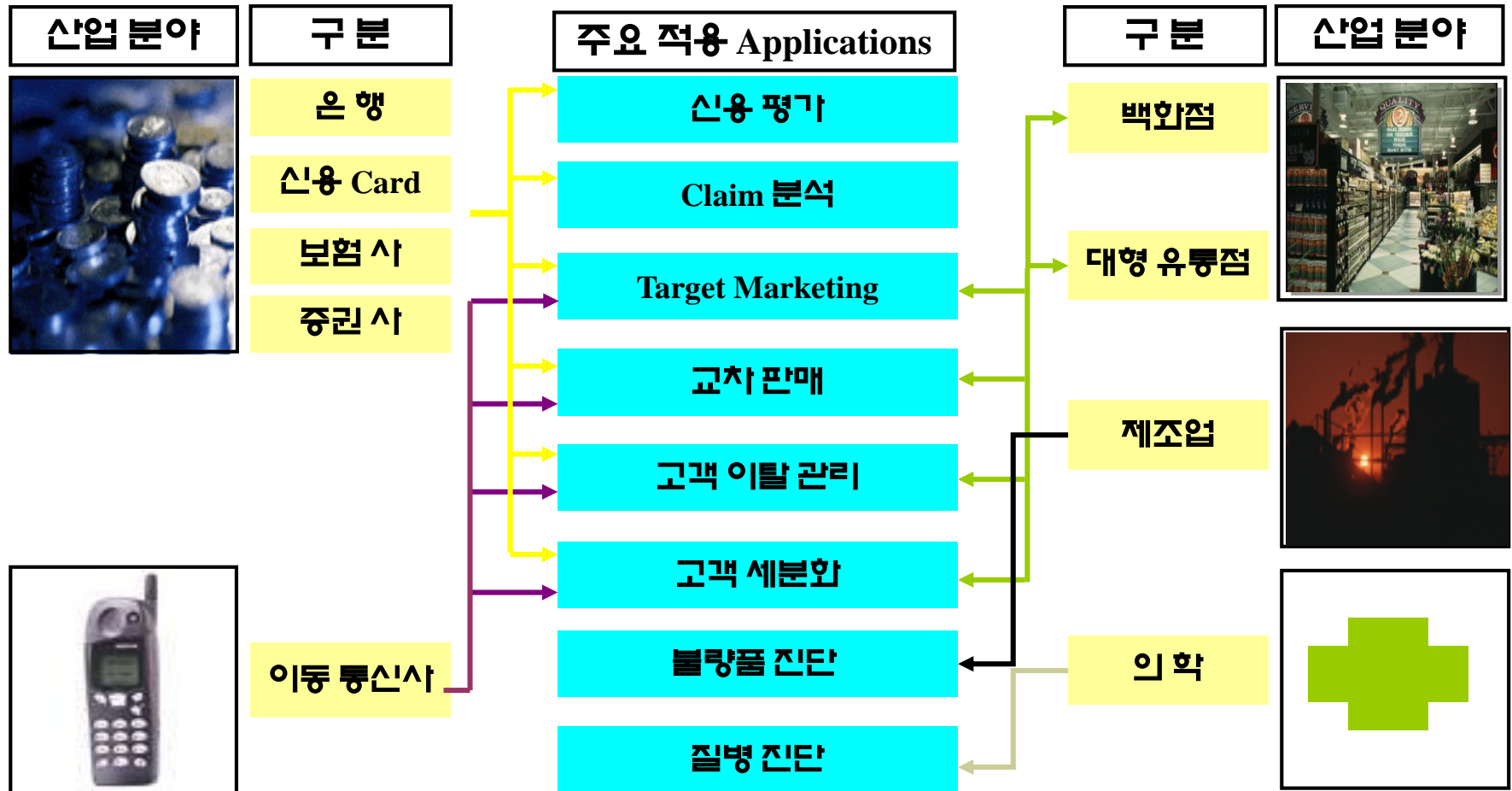
- Scoring

- 신용점수 (Credit Score)
- 우수고객점수 (Loyalty Score)
- 고객이탈 및 연체점수 (Attrition Score)
- 구매가능점수

- 기타

- 도용사고 방지 (Fraud Dection)
- 위험관리 (Risk Management)
- 고객불만관리 (Crime Prevention)
- 품질/제품관리 (Production & Process Management)

# 데이터마이닝 활용분야



## 데이터마이닝의 정의

---

**대용량의 데이터에서 유용한 정보와 관계를  
탐색하고 모형화 하여 지식을 발견하는 과정**

**Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.**

**(Gartner Group, [www.gartner.com](http://www.gartner.com))**

**Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases.”**

**(META Group, [www.metagroup.com](http://www.metagroup.com))**

 **There are many different definitions of data mining, but almost all involve finding or discovering useful relationships in large databases.**

## 데이터마이닝의 특징

---

- 운영계에 축적된 과거자료로부터 비계획적으로 수집된 대용량의 데이터를 다룬다. (Experimental Data vs Historical Data)
- 컴퓨터의 강력한 처리능력을 이용하여 실용화되고 있다.
- 대다수의 데이터마이닝 기법들은 수학적으로 증명되고 발전된 것이 아니라 경험적으로 개발되었다. (Exploratory vs Confirmatory)
- 데이터마이닝의 주요 관심은 통계적 추론과 검정보다는 예측모형의 일반화에 있다. (Underfitting vs Overfitting)
- 기업의 다양한 의사결정 활동에 활용하기 위해서 사용된다.
- 데이터마이닝은 통계학, 전산과학, 인공지능, 공학 분야에서 개발되기 시작하였다. 그러나 실제로 이를 활용하는 전문가들은 경영, 경제, 정보기술 분야에서 배출되고 있다. (Tangle of terminology)

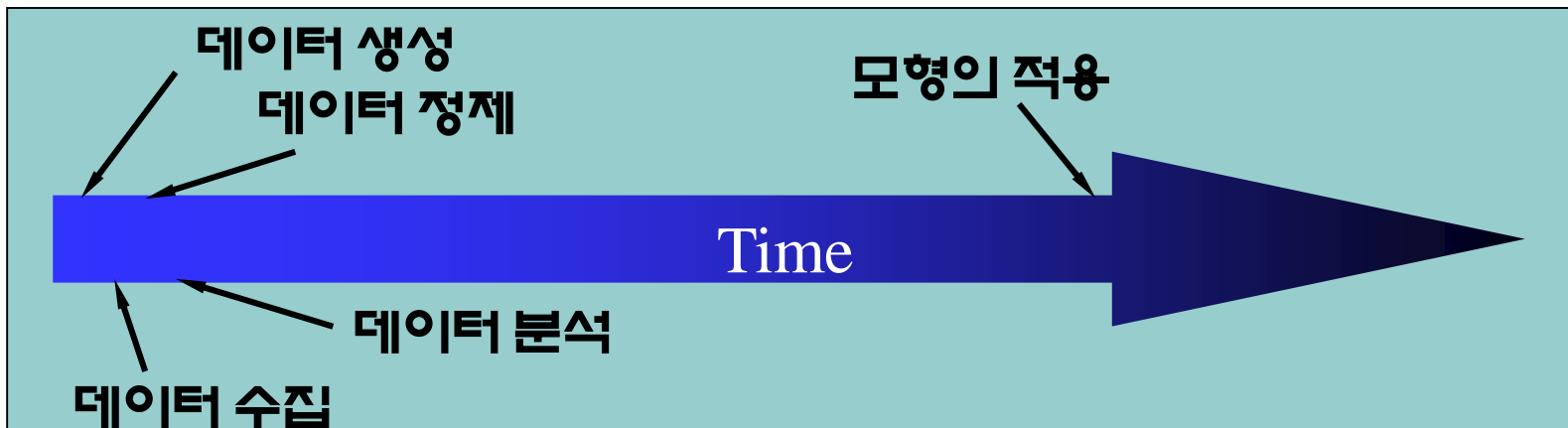
# 실험자료와 관측자료

## 실험자료

목적	연구
가치	과학
수집	통제된 현재 자료
크기	작다
정도	정제되어 있다
상태	정적

## 관측자료

업무활용
상업
관찰된 과거 자료
크다
정제되어 있지 않다
동적





## 1.1.6 데이터마이닝 적용 사례

The screenshot shows the SAS Korea website with the URL <http://www.sas.com/offices/asiapacific/korea/success/index>. The page features a navigation bar with links to HOME, PRODUCTS & SOLUTIONS, SUCCESS STORIES, PARTNERS, COMPANY INTRODUCTION, EDUCATION CENTER, and CUSTOMER SUPPORT. The main content area is titled '고객 성공 사례 / 솔루션별' (Customer Success Stories / By Solution). On the left, there are links to '성공 사례' (Success Stories) and '산업별 성공사례' (Success Stories by Industry). The main content is divided into two columns of solutions:

- 고객 관계 관리(CRM)**
  - General
  - 고객 경험 분석
  - 디지털 마케팅
  - 마케팅 리소스 관리
  - 마케팅 자동화
  - 마케팅 최적화
  - 실시간 의사 결정 관리
- 공급망 인텔리전스**
- 수요 예측**
- 품질 라이프사이클 분석**
- 공급업체 인텔리전스**
  - General
  - 비용 분석
- 금융 인텔리전스**
- 리스크 관리**
  - General
  - 신용 리스크 관리
  - 운영 리스크 관리
- 법규준수(Compliance)**
  - 바젤 II
  - 자금세탁방지
- 분석 보고 자료**
- 비즈니스 분석**
- 사기 행위 방지 및 적발**
- 서비스 인텔리전스**
  - General
  - 서비스 파트 최적화
  - 품질 보증 분석
- 성과 관리**
- 정보 관리**

At the bottom, there are two featured success stories:

- 1-800-FLOWERS.COM**: SAS CRM 솔루션 도입 후 새로운 전성기를 맞고 있는 1-800-FLOWERS.COM
- 1-800-FLOWERS.COM**: 찬사가 끊이지 않는 1-800-FLOWERS.COM의 탁월한 CRM 구현능력
- MARKTPLAATS.NL**: 이베이의 네덜란드 지사인 Marktplaats.nl은 SAS®을 통해 시장 및 고객 통찰력을 적극 활용하고 있습니다
- MAX NEW YORK LIFE**: SAS를 통해 Max New York Life는 시장 점유율을 높이고, 핵심 고객들을 보유할 수 있게 되었습니다

## 사례: 의류, 유통업체

---

- 목적

- 다량의 판매데이터를 이용하여 제품간의 연관관계를 발견

- 분석방법

- 연관성 규칙(Association Rule)

- 성과

- 분석결과 발견된 브랜드간이나 제품간의 연관규칙은 현업으로부터의 심도 있는 검증  
을 거쳐 다양한 판매전략에 활용

## 사례: 의류, 유통업체

---

- 활용 예

- 제품 카탈로그를 제작할 때
- 매장의 위치를 조정하고 제품을 배치할 때
- 한 제품을 구입한 고객에게 해당제품과 연관관계가 높은 타제품을 권하는 교차판매를 시도할 때

- 대표적 성공기업

- 미국의 아마존사(Amazon.com)

: A책을 조회할 때 나타나는 화면을 보면, 상단에는 책에 대한 간단한 정보를, 하단에는 이 책과 연관관계가 높은 책들의 리스트를 보여주므로써 추가 판매 기회를 극대화

# 사례: 신용카드 회사

---

- 목적

- 카드사용의 부정 행위 적발 및 예방

- 분석방법

- 의사결정나무분석, 신경망 분석 등

- 성과

- 과거 정상적으로 거래된 데이터와 도용사고 경험이 있는 데이터를 기반으로 각각의 패턴을 분석하여 모형화하고, 구축된 모형을 카드승인시에 적용하여 만일 부정행위로 의심이 되면 승인을 거부함으로써 불법적인 카드사용을 적발하거나 사전에 예방함으로써 도용사고로 인한 손해액을 감소

## 사례: 통신회사

---

- 목적 : 고객의 이탈방지/감소

- 매년 전체 고객의 23%를 잃고 있음
- 고객을 새로 유치하는데 1인당 \$350의 비용지출

- 분석방법

- 고객성향변동관리(Churn management) 와 군집분석(Clustering)을 이용하여 이탈의 원인을 파악
- 고객의 이탈가능성을 예측할 수 있는 모델을 개발
- 이익분석(Profit analysis)

## 사례: 통신회사

---

- **결과**

- 이 회사의 관리자는 고객의 60% 정도는 경쟁업체로 옮겨갈 가능성이 적은 고객이고 나머지 40%는 이탈가능성이 높은 고객임을 알게 됨
- 이탈방지 노력이 이탈가능성이 매우 높은 고객에게는 별 효과가 없고 이탈가능성이 어느 정도 높은 고객에게는 큰 효과를 발휘한다는 것을 발견

- **성과**

- 무료 전화서비스 등을 제공하는 목표 마케팅(Target Marketing) 전략을 통해 고객 이탈율을 19.7%(전년도 23%)로 줄이고 큰 이익증가를 기록

## 사례: 의료, 병원

---

- 목적

- 종양의 악성/양성 판단에 의한 암 진단의 정확성 향상

- 분석방법

- 판별 및 분류(Discrimination and Classification)분석

- 분석과정

- 과거 환자들의 종양검사 결과를 근거로(즉, 종양의 크기, 모양, 색깔 등을 기반으로) 종양의 악성/양성 분류모형을 만든 후 새 환자로부터 채취한 종양분류시 적용

- 성과

각종 종양들에 대한 구별력을 향상시켰고 더욱 정확한 암진단과 치료에 이용

## 사례: 보험회사

---

- 목적

- 이탈/이탈 가능 고객 특성파악

- 분석방법

- 의사결정나무분석(Decision Tree Analysis)

- 성과

- 이탈고객의 특성파악 결과를 토대로 유사 특성을 지닌 기존고객(즉, 향후 이탈가능성이 높은 고객)을 대상으로 특별한 마케팅 활동을 펼쳐 이탈고객을 최소화 함으로써 기업의 이익을 증가



## 차례

---

1.1 데이터마이닝이란 무엇인가?

1.2 데이터마이닝 프로젝트의 수행 프로세스

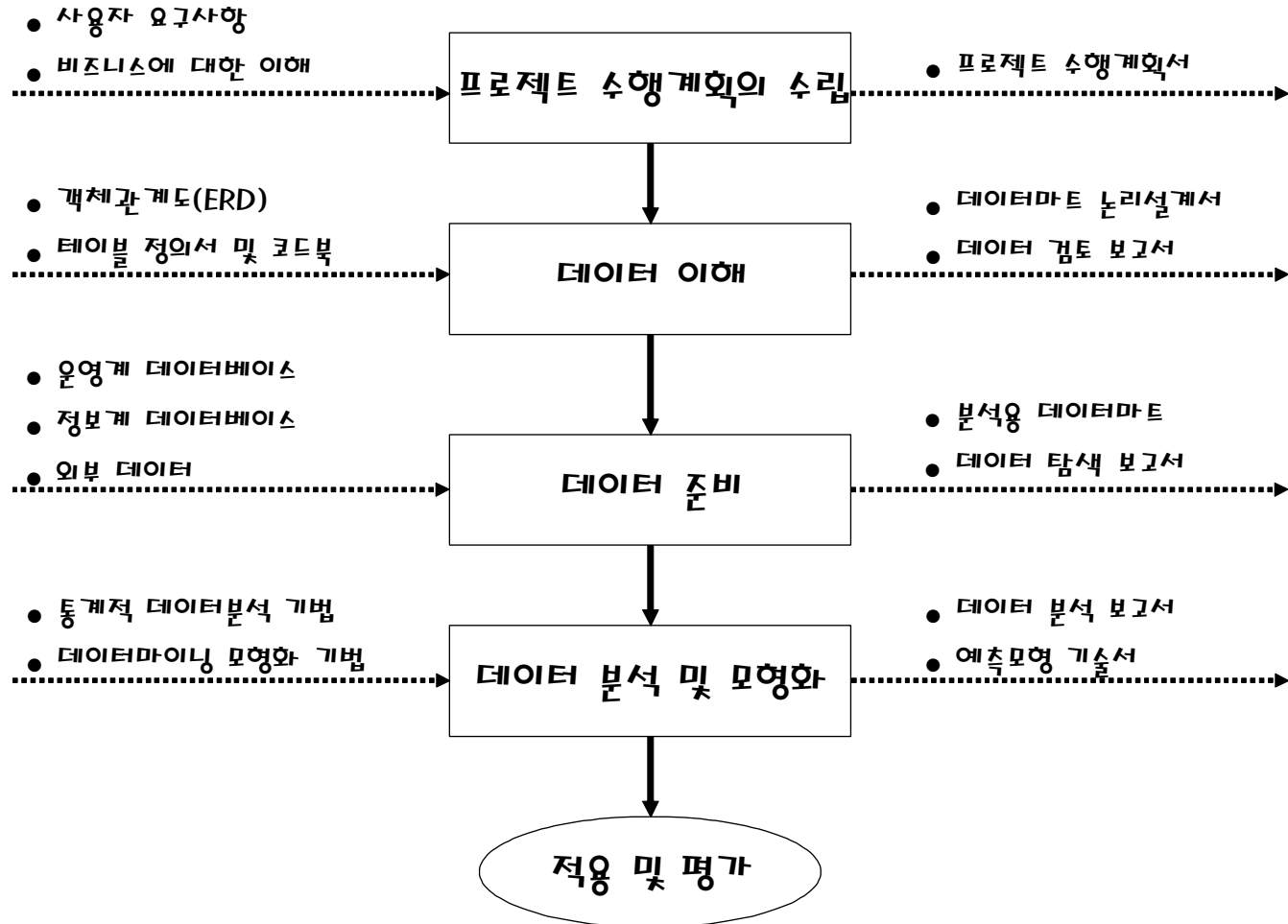
1.3 데이터마이닝 예측기법

1.4 Enterprise Miner의 소개

1.5 맺음말

1.6 연습문제

# 데이터마이닝 프로젝트의 수행 프로세스



## 1.2.1 프로젝트 수행계획의 수립

---

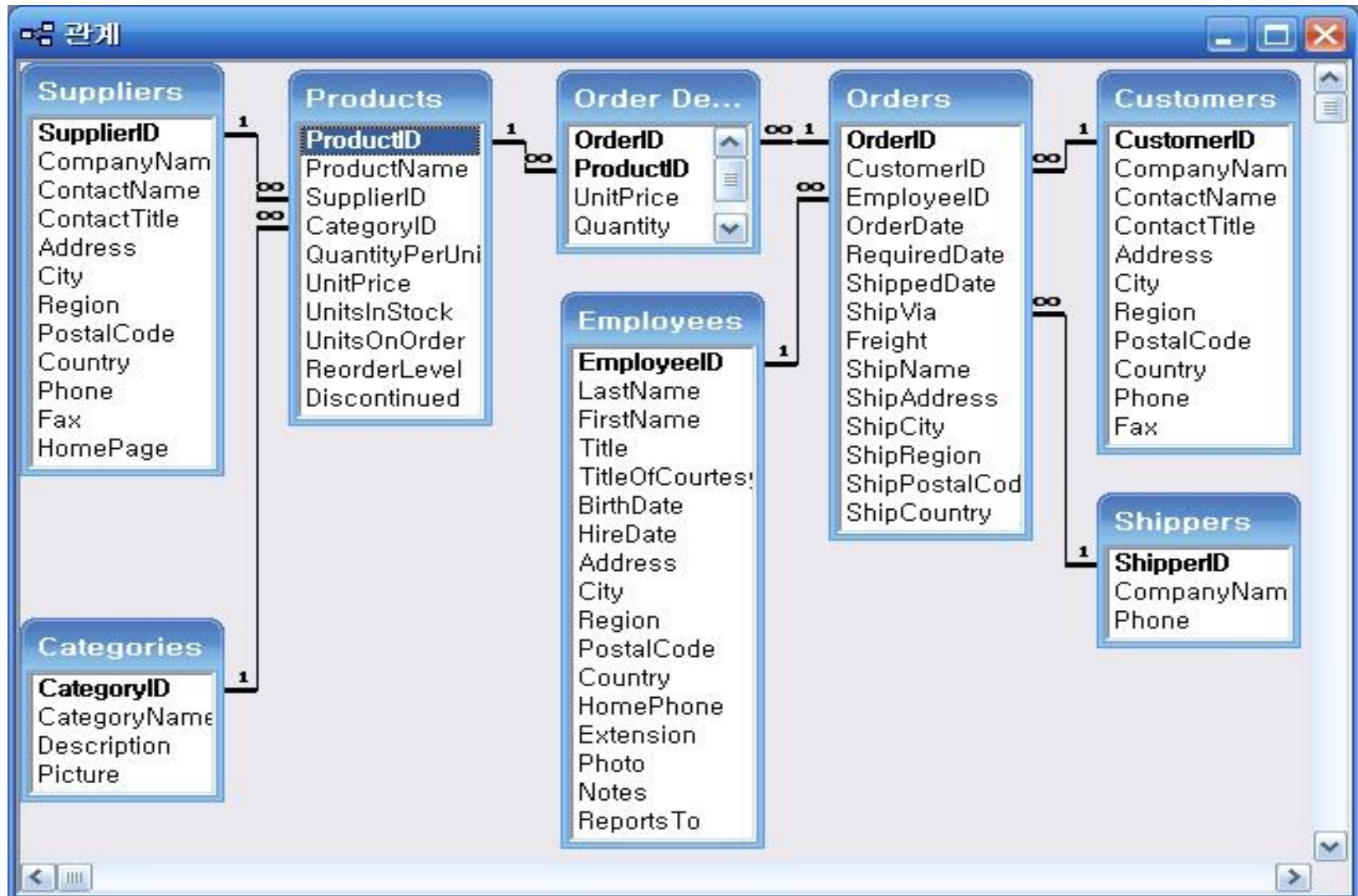
- 해당 비즈니스(업무)에 대한 충분한 이해
  - 필요한 데이터를 관리하고 추출할 수 있는 정보기술
  - 적절한 데이터 처리와 분석을 수행할 수 있는 데이터 분석 능력
- 
- ✓ 프로젝트의 범위와 산출물 정의
  - ✓ 비즈니스(업무)에 대한 이해 및 공유
  - ✓ 사용자 요구사항과 필요사항 검토
  - ✓ 참여 인력 및 역할에 대한 정의
  - ✓ 세부일정 정의 및 수행계획서 작성

## 1.2.2 데이터에 대한 이해

---

- 사용 가능한 내부 및 외부 데이터들의 원천 파악
- 데이터 원천들에 대한 위치와 구조(수집, 입력, 관리, 갱신 등의 경로) 파악
- 데이터 테이블들의 필드(field)와 그들의 코드(code) 파악
- 데이터들의 신뢰성, 정확성, 유용성에 대한 검토
- 분석용 데이터마트(data mart)를 구성하기 위한 논리설계서 작성

## ERD의 예: Northwind Data Base (Microsoft Access)



# 데이터에 대한 이해

## 〈테이블 정의서(Table Layout)와 코드(Code) 예시〉

Motor TABLE		
Field	Type	설명
PlcId	INTEGER(14)	증권번호
SmName	CHAR(8)	피보험자 이름
Ssn	CHAR(14)	주민등록번호
ZipCode	CHAR(6)	주소지 우편번호
Car	CHAR(2)	차종
Usage	CHAR(2)	차량용도
Displace	CHAR(1)	배기량
...	...	...

01 : 승용차  
02 : 승합차  
03 : 화물차  
04 : 이륜차  
...

1 : 1000cc 이하  
2 : 1000~1500cc  
3 : 1500~2000cc  
4 : 2000cc 이상  
...

# 데이터에 대한 이해

## 〈데이터마트 논리설계서의 예시〉

자동차보험 Mart						
구분	항목	변수명	설명	코드	소스	작업자
기본 사항	증권번호	Plc_id	기본키		Moter.cid	IT
	피보험자 ID	In_id	정의 계산방법 등에 관한 설명		계산에 필요한 소스 테이블 및 필드 이름	IT
	갱신여부	respond				IT
	피보험자 연령	In_age				DM
	피보험자 연령(R)	In_age_r		1)2b세 이하, 2)...		DW
	계약자 ID	Cn_id				IT
	계/피 동일인 여부	Meq		o)N, 1)Y		OLAP
	...	...	...	...		...
차량 사항	차종	Car		1)승용차, 2)...		IT
	배기량	Displace		1)1000cc이하, 2)		IT
	차량용도	Usage		1)사업용, 2)...		IT
	...	...	...	...	...	...

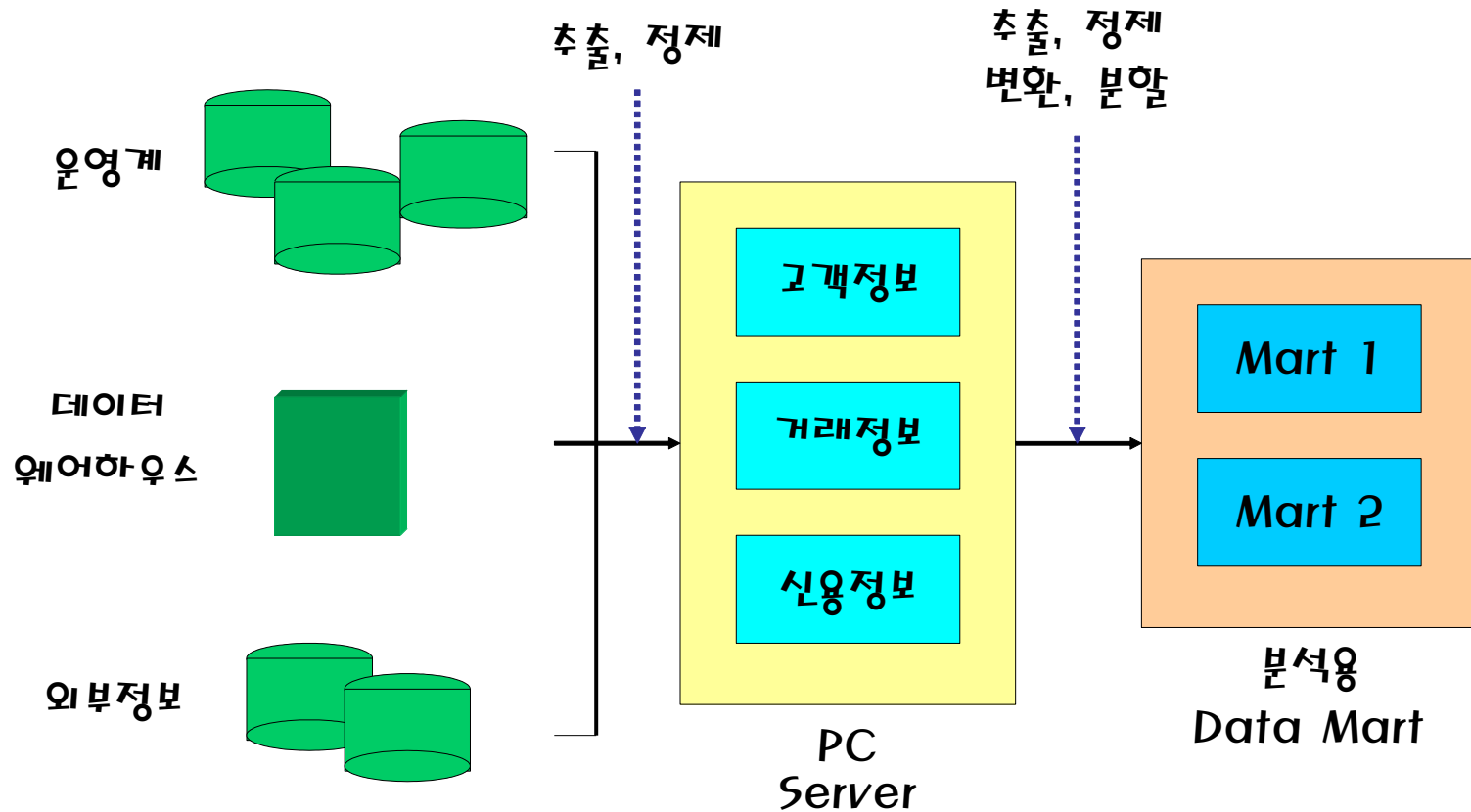
그룹 변수

파생 변수

- 기본사항 : 증권번호, 피보험자 ID, 갱신여부, 피보험자 연령, 계약자 ID, 계약자 연령, ...
- 차량사항 : 차종분류, 차량용도, 사용용도, 차량등록지, 배기량, 차량가액, 제조회사, ...
- 계약사항 : 계약일, 계약경로, 납입방법, 연령한정특약 가입여부, 자손 가입여부, ...
- 이력사항 : 계약년차, 1년전 가입사(자사/타사), ...
- 기타사항

## 1.2.3 데이터 준비

### 〈분석용 데이터마트 구축의 예시〉





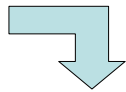
# 데이터 준비

- 데이터 사전처리 (Pre-processing of data)
  - 재배열 (Rearrangement)
  - 요약변수 (Summary Variable)
  - 파생변수 (Derived Variable)
  - 그룹화 (Grouping)

## 재배열의 예시

고객	구매일	상품	...
3135	970304	A01	...
3135	980715	B01	...
3135	991113	C01	...
2784	930508	C02	...
2784	980106	B01	...
8321	910305	A02	...
8321	930521	C02	...
8321	940627	D01	...
8321	981125	E03	...
8321	990305	F01	...
...	...	...	...

Long-Narrow (Transaction Table)

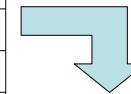


고객	P_A	P_B	P_C	P_D	P_E	P_F
3135	1	1	1	0	0	0
2784	0	1	1	0	0	0
8321	1	0	1	1	1	1

Short-Wide (Mart Table)

## 요약변수의 예시

고객	구매일	상품	금액	...
3135	970304	A01	160	...
3135	980715	B01	42	...
3135	991113	C01	212	...
2784	930508	C02	250	...
2784	980106	B01	122	...
8321	910305	A02	786	...
8321	930521	C02	458	...
8321	940627	D01	328	...
8321	981125	E03	27	...
8321	990305	F01	759	...
...	...	...	...	...



고객	총금액	평균금액	건수	...
3135	414	138	3	...
2784	372	186	2	...
8321	2,358	471	5	...

최근 6개월 구매건수  
 최근 12개월 구매건수  
 최근 6개월 구매금액  
 최근 12개월 구매금액  
 최근 6개월 평균 구매금액  
 최근 12개월 평균 구매금액  
 최근 12개월 외주 구매금액  
 최근 12개월 식품 구매금액  
 최근 12개월 가전 구매금액  
 ...

# 데이터 준비

- 데이터에 대한 탐색 및 보완

- 오류값(Error)

- : 변수가 가질 수 없는 값, 변수값의 불가능한 조합, 일관성 없는 코드값, 잘못된 코드값.

- 이상치(Outlier)

- : 정상이 아닌 자료값. 특이값은 오류값일 수도 있고 그렇지 않을 수도 있다.

- 결측값(Missing)

- : 원인과 기록방법을 정밀하게 조사하여 자료값을 정정하고 기록방법을 변경해야 하며, 필요 시에는 자료값을 보정해야 한다.

Garbage in, garbage out !

- 연구와 분석의 목적을 명확히 해야 한다.
- 분석의 목적에 부합하는 데이터를 수집해야 한다.
- 데이터는 정밀하게 검사되고 분석에 적합하도록 정리되어야 한다

## 차례

---

1.1 데이터마이닝이란 무엇인가?

1.2 데이터마이닝 프로젝트의 수행 프로세스

1.3 데이터마이닝 예측기법

1.4 Enterprise Miner의 소개

1.5 맺음말

1.6 연습문제

# 데이터마이닝 예측기법

---

- **Supervised Prediction (지도예측)**

- 신경망 (Artificial Neural Network)
- 판별분석 (Discrimination Analysis)
- 일반화선형모형 (GLM, Generalized Linear Model)
  - 선형회귀분석 (Regression Analysis)
  - 로지스틱 회귀분석 (Logistic Regression)
- 사례기반추론 (Case-Based Reasoning)

- **Unsupervised Prediction (자율예측)**

- OLAP (On-Line Analytic Processing)
- 연관성규칙발견 (Association Rule Discovery, Market Basket)
- 군집분석 (k-Means Clustering)
- 인자분석(Factor Analysis), 주성분분석(Principal Component)
- k-Nearest Neighbor
- SOM (Self Organizing Map, Kohonen Network)

## 1.3.1 지도예측(Supervised Prediction)

: Supervised Learning, Directed Learning

- **목표변수 (Target Variable)**

: response, outcome, dependent variable

- **입력변수 (Input Variable)**

: predictors, explanatory variables, independent variables

Obs.	입력변수			목표변수	
	Sex	Age	Region	y	
1	F	18	A	1	<div>예측확률</div> <div>↑</div> <div>P ( y = 1 )</div>
2	M	25	D	0	
3	F	67	D	1	
4	F	43	B	1	
5	F	28	A	0	
6	M	53	C	0	
7	F	42	A	0	

Obs.	입력변수			목표변수	
	Sex	Age	Region	y	
1	F	18	A	125	<div>예측값</div> <div>↑</div> <div><math>\hat{y}</math></div>
2	M	25	D	35	
3	F	67	D	150	
4	F	43	B	45	
5	F	28	A	13	
6	M	53	C	38	
7	F	42	A	20	

$$\hat{P}(y=1) = \frac{\exp(a + b_1x_1 + b_2x_2 + \dots + b_px_p)}{1 + \exp(a + b_1x_1 + b_2x_2 + \dots + b_px_p)}$$

(예측모형: 로지스틱 회귀분석)

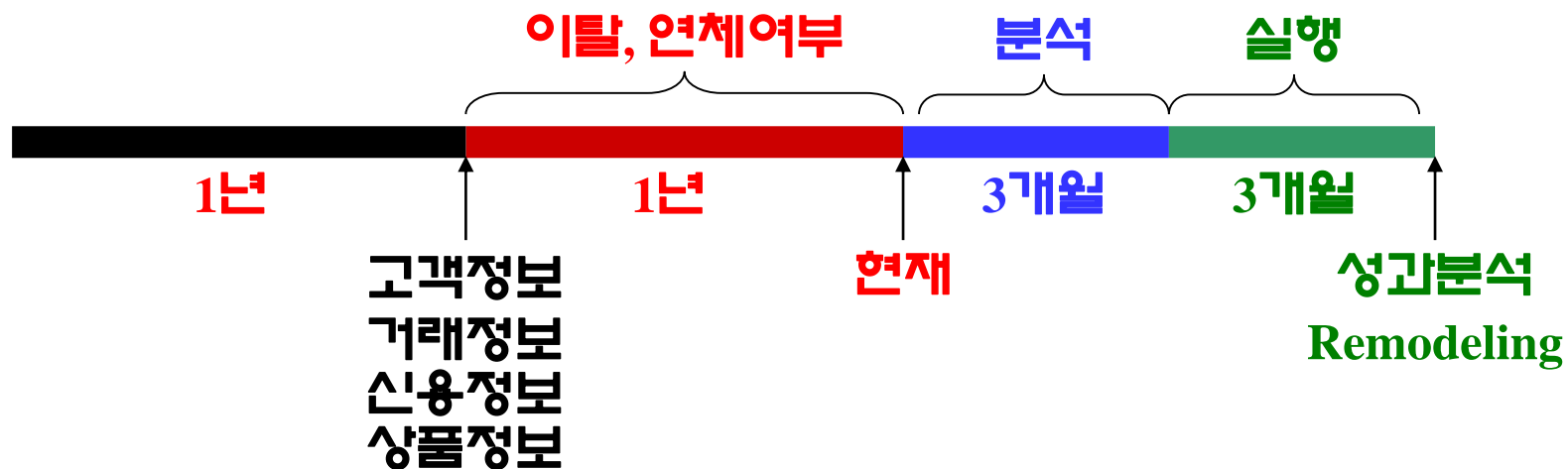
$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_px_p$$

(예측모형: 선형 회귀분석)

## 지도예측(Supervised Prediction)

### ● Scoring

- 신용점수 (Credit Score)
- 우수고객점수 (Loyalty Score)
- 고객이탈 및 연체점수 (Attrition Score)
- 구매가능점수



# 지도예측(Supervised Prediction)

---

- 이탈예측

- 대상: 현재의 고객 (?)
- 입력변수: 고객정보, 거래정보, 상품정보 ...
- 목표변수: 이탈, 연체 ...
- 활용: 고객의 이탈을 방지하고 충성도를 높임

- 신용점수

- 대상: 과거의 대출신청자
- 입력변수: 대출신청 당시의 고객정보, 신용정보
- 목표변수: 채무불이행, 연체 ...
- 활용: 새로운 고객의 대출신청에 대한 판단

# 지도예측(Supervised Prediction)

---

- Target Marketing

- 대상: 거래실적이 있는 고객
- 입력변수: 고객정보, 거래정보(RFM), 상품정보 ...
- 목표변수: 구매여부, DM/TM에 대한 반응여부
- 활용: 캠페인 또는 판촉 등의 영업활동

- 부정거래 적발

- 대상: 거래실적이 있는 고객
- 입력변수: 거래정보, 고객정보
- 목표변수: 부정거래, 카드의 도용
- 활용: 부정거래 방지, 카드의 도용사고 방지



## 1.3.2 자율예측(Unsupervised Prediction)

: Clustering, Unsupervised Learning, Undirected Learning

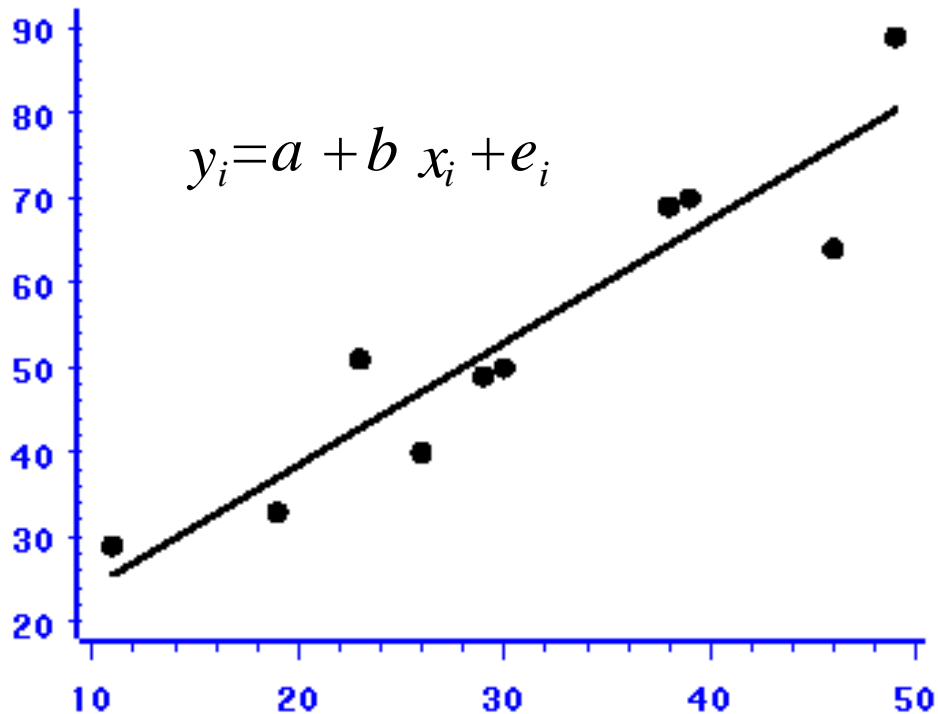
- **목표변수(Target Variable):** 정해져 있지 않음
- **세분화(Segmentation):** 고객 세분화, 시장 세분화

	설명	해당 Life Style	특징	주관심 품목
Seg 1	성별:남 연령:40 대 구매금액 상	거래 편의성과 제품 구매측면의 대안평가 과정 중시	보안문제 중시 점심시간 및 퇴근시간 조회 집중	Computer 관련제품 가전제품
Seg 2	성별:여 연령:3~40 세 신 APT 지역 구매건수 상	상품검색 편리성 /인터넷 접근 용이성 중시	주문/배달/결제 의 일괄처리 신속한반품처리	생활필수품 김치, 쌀등
.	.	.	.	.

마케팅 전략수립의 기초

# 모델링: 선형 회귀분석(Linear Regression)

판매대수



$$y_i = a + b x_i + e_i$$

예약대수

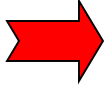
$x$	$y$	$\hat{y}$
11	29	25.5
19	33	37.1
23	51	42.8
26	40	47.1
29	49	51.5
30	50	52.9
38	69	64.4
39	70	65.9
46	64	76.0
49	89	80.3



$$y_i = a + b x_i = 9.74 + 1.44 x_i$$

## 모델링: 로지스틱 회귀분석(Binary Response)

$$y = 0.1 + 0.01x$$

$x$	$y$		$\hat{y}$
10	0		<b>0.2</b>
100	1		<b>1.1</b>
1000	1		<b>10.1</b>

- 로지스틱 회귀분석

$$\log \frac{P(y=1|x)}{1-P(y=1|x)} = \alpha + \beta x + \varepsilon$$

## 로지스틱 회귀분석에 의한 예측모형 예시

대출금	대출금잔액	담보금	대출사유	직업	근무년수	신용거래수	신용상태	최초신용	P(나쁨)	P(좋음)
2300	102370	120953	Homelmp	Office	2	13	0	91	0.04	0.96
2400	34863	47471	Homelmp	Mgr	12	21	1	70	0.14	0.86
2400	98449	117195	Homelmp	Office	4	13	0	94	0.03	0.97
2900	103949	112505	Homelmp	Office	1	13	0	96	0.03	0.97
2900	104373	120702	Homelmp	Office	2	13	0	102	0.03	0.97
2900	7750	67996	Homelmp	Other	16	8	1	122	0.68	0.32
2900	61962	70915	DebtCon	Mgr	2	37	1	283	0.19	0.81
3000		14500	Homelmp	Other	3	2	1	9		
3000		14100	Homelmp	Other	1	19	1	104		
3200	74864	87266	Homelmp	ProfExe	7	12	0	251	0.08	0.92
3200	23159		Homelmp	Mgr	20	9	1	118		
3800		73189					0			
3300	130518	164317	DebtCon	Other	9	33	1	192	1.00	0.00
3600	52337	63989	Homelmp	Office	20	20	0	204	0.00	1.00
3700	17857	21144	Homelmp	Other	5	9	1	130	0.03	0.97
3800	51180	63459	Homelmp	Office	20	20	0	204	0.00	1.00
3900	29896	45960	Homelmp	Other	11	14	1	146	0.02	0.98
4000	105164	112774	Homelmp	Office	1	13	0	95	0.03	0.97
4000	54543	61777	Homelmp	Office	21	19	0	206	0.01	0.99
4000	26572	31960	Homelmp	Office	11	8	1	118	0.10	0.90
4100	57992	63797	DebtCon	ProfExe	7	31	0	166	0.22	0.78

$$P(\text{신용상태} = \text{좋음}) = \frac{\exp(1.7 + 2.3X_1 - 0.45X_2 + \dots)}{1 + \exp(1.7 + 2.3X_1 - 0.45X_2 + \dots)}$$

# 사후확률에 의한 예측

## 〈사례 : 손해보험회사의 이탈고객분석의 예〉

목표변수

	연납	성별	납입방법	접금방법	대부유무	연령	납입비율	가입일자	계약상태	예측변수	이탈확률	이탈집단
1	10	여자	1	신용카드	무	39	10.00	10	정상	정상	.124	0
2	5	남자	1	신용카드	무	43	21.67	8	정상	정상	.491	0
3	5	남자	2	신용카드	무	35	21.67	9	정상	정상	.012	0
4	5	남자	2	방문수금	무	37	21.67	9	정상	정상	.024	0
5	5	남자	2	신용카드	무	31	21.67	12	정상	정상	.125	0
6	5	남자	1	신용카드	무	51	23.33	11	해지	정상	.125	0
7	5	남자	4	신용카드	무	44	11.67	13	해지	해지	.562	1
8	10	여자	3	신용카드	무	49	.83	12	해지	해지	.989	1
9	10	남자	1	신용카드	무	34	7.50	8	해지	해지	.510	1
10	7	남자	2	신용카드	무	37	7.14	9	해지	해지	.780	1
11	10	여자	2	신용카드	무	59	4.17	12	해지	해지	.887	1
12	10	여자	2	신용카드	무	34	10.83	9	정상	정상	.123	0
13	10	남자	2	신용카드	무	31	10.83	9	정상			

예측결과

# 의사결정나무분석의 예시

계약상태

Cat.	%	n
0	74.93	6328
1	25.07	2117
Total (100.00)		8445

0:정상  
1:해지

집금방법

P-value=0.0000; Chi-square=446.8867; df=1

직원집금

Cat.	%	n
0	61.39	1857
1	38.61	1168
Total (35.82)		3025

자동이체

Cat.	%	n
0	82.49	4471
1	17.51	949
Total (64.18)		5420

납입방법

P-value=0.0000; Chi-square=60.1157; df=1

월납;3개월납

Cat.	%	n
0	60.39	1776
1	39.61	1165
Total (34.83)		2941

6개월납;연납

Cat.	%	n
0	96.43	81
1	3.57	3
Total (0.99)		84

연납

P-value=0.0000; Chi-square=41.7403; df=2

3개월

Cat.	%	n
0	84.11	98
1	15.89	17
Total (1.27)		107

5개월

Cat.	%	n
0	62.21	1032
1	37.79	627
Total (19.64)		1659

7;10개월

Cat.	%	n
0	55.66	654
1	44.34	521
Total (13.91)		1175

나이

P-value=0.0000; Chi-square=62.1392; df=3

[20,36]

Cat.	%	n
0	79.11	2034
1	20.89	537
Total (30.44)		2571

(36,53]

Cat.	%	n
0	84.66	2091
1	15.34	379
Total (29.25)		2470

(53,78]

Cat.	%	n
0	89.88	293
1	10.12	33
Total (3.86)		326

성별

P-value=0.0024; Chi-square=9.2446; df=1

남

Cat.	%	n
0	78.02	1654
1	21.98	466
Total (25.10)		2120

여

Cat.	%	n
0	84.26	380
1	15.74	71
Total (5.34)		451

## 모델링: 군집분석(Clustering Analysis)

- 개인 또는 개체 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석방법

(예) 수입과 상표충성도 기준으로 고객세분화 (Segmentation)



## 모델링: 연관성 분석(Association Analysis)

- 하나의 거래나 사건에 포함되어 있는 항목들의 관련성을 파악해서 둘 이상의 항목들로 구성된 연관성 규칙을 통한 탐색적 자료분석 방법

예) Products in Shop Cart (One trip, Together)



- 1) “오렌지 주스와 식기세제” 구입시 “윈도우 클리너” 를 같이 구입하는가?
- 2) “우유” 를 “바나나” 구입시 함께 구입하는가?  
또한 구입 할 때 특정 브랜드를 구입하는가?
- 3) “식기세제” 를 어느 곳에 위치시켜야지만 판매고를 최대화하는가?



# 연관성 분석의 예제

## 고객의 구매 상품 List

ID	판매 상품
1	소주 , 쿨라 , 맥주
2	소주 , 쿨라 , 와인
3	소주 , 주스
4	쿨라 , 맥주
5	소주 , 쿨라 , 맥주 , 와인
6	주스



## 지지도가 50% 이상인 연관성 규칙

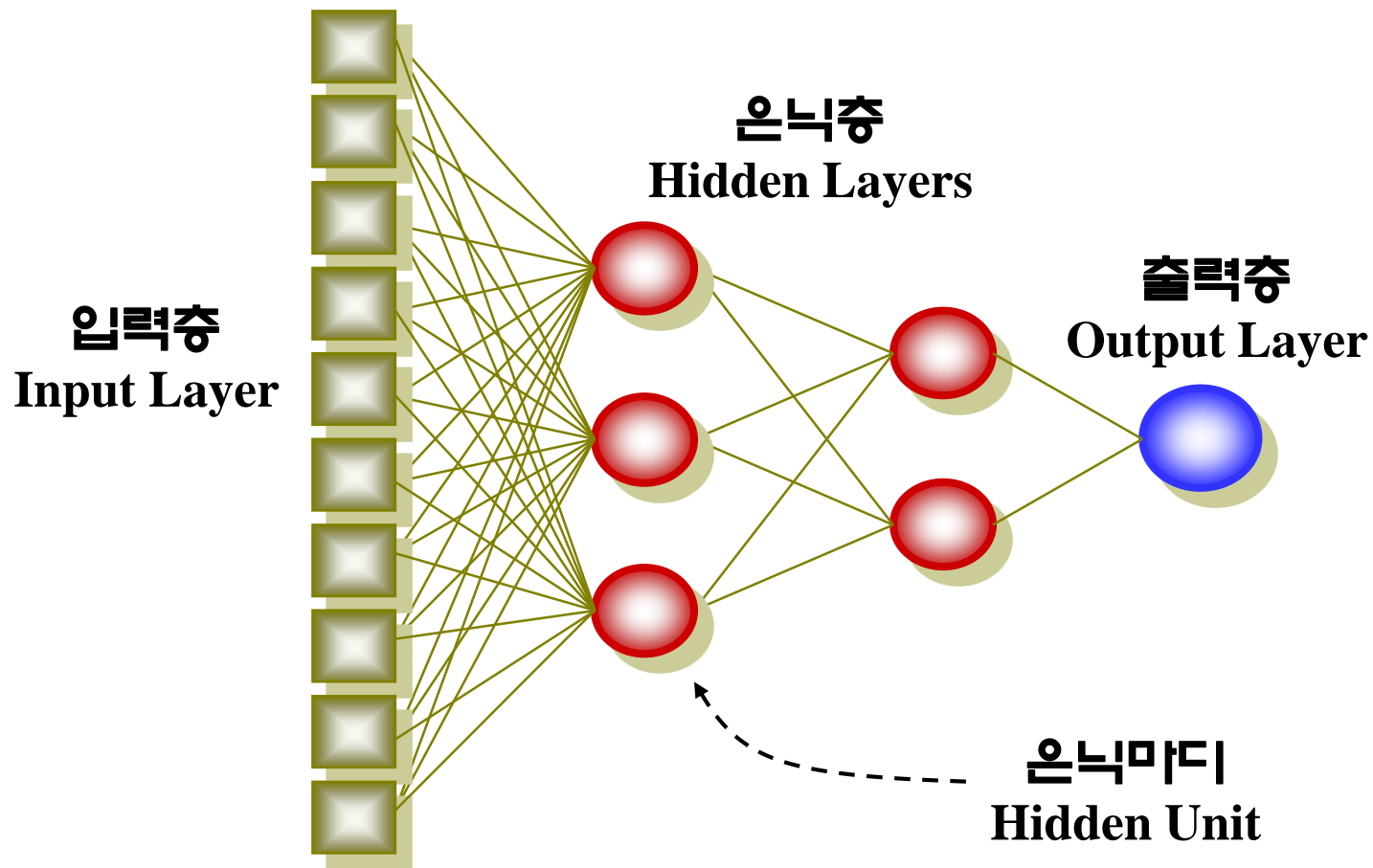
연관성 규칙 (지지도 50%이상)	해당 거래	신뢰도
소주 → 쿨라	1, 2, 5	75 %
쿨라 → 맥주	1, 4, 5	75 %
맥주 → 쿨라	1, 4, 5	100 %

\* 연관성 규칙 : 맥주를 구입한 사람들 모두는(100%) 쿨라도 구매한다.

$$\text{Lift} = P(\text{쿨라}|\text{맥주}) / P(\text{쿨라}) = 1 / (4/6) = 1.5$$

- 이러한 경향을 가지는 사람들은 전체의 절반(50%) 정도
- 맥주 구매 시 쿨라를 구입하게 될 가능성은 맥주 구매가 전제되지 않았을 경우보다 1.5배나 높아진다.

## Multilayer Perceptron



# 모형 평가

## The Two-Class Problem

		Predicted Class		
		0	1	
Actual Class	0	True Neg Neg	False Pos Pos	Total Negative
	1	False Neg Neg	True Pos Pos	Total Positive
		Total Negative	Total Positive	

## ...Two-Class Problem

		Predicted		
		0	1	
Actual	0	40	4	44
	1	20	86	106
		60	90	150

Mosaic display

40% 60%

오류율 (Error rate)

$$= (\text{false negative} + \text{false positive}) / (\text{grand total}) = (20 + 4) / 150 = 16\%$$

정확도 (Accuracy)

$$= (\text{true negative} + \text{true positive}) / (\text{grand total}) = (40 + 86) / 150 = 84\%$$

민감도 (Sensitivity)

$$= (\text{true positive}) / (\text{total actual positive}) = 86 / 106 = 81\%$$

특이도 (Specificity)

$$= (\text{true negative}) / (\text{total actual negative}) = 40 / 44 = 91\%$$

## 차례

---

1.1 데이터마이닝이란 무엇인가?

1.2 데이터마이닝 프로젝트의 수행 프로세스

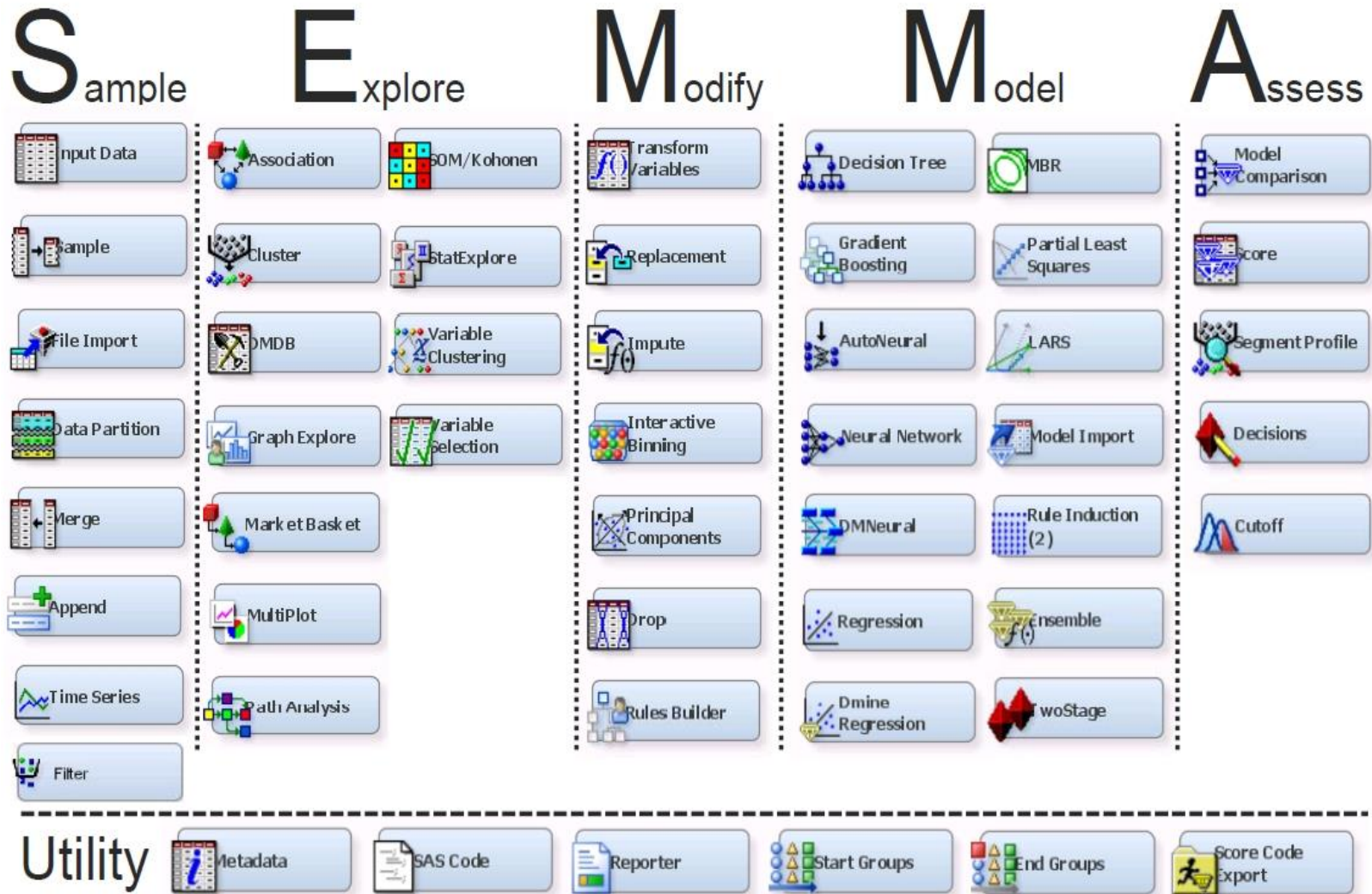
1.3 데이터마이닝 예측기법

1.4 Enterprise Miner의 소개

1.5 맺음말

1.6 연습문제

# SEMMA



## 차례

---

1.1 데이터마이닝이란 무엇인가?

1.2 데이터마이닝 프로젝트의 수행 프로세스

1.3 데이터마이닝 예측기법

1.4 Enterprise Miner의 소개

1.5 맺음말

1.6 연습문제

## 데이터마이닝 프로젝트 수행의 어려움

- 장기적이고 구체적인 계획의 부족
- 데이터에 대한 준비 부족
- 시간차이 문제
- 적용상의 문제
- 부서 및 프로젝트들 간의 비협조체제

