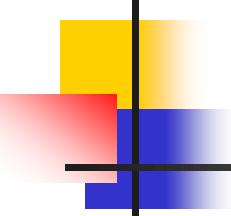


Become a Data Scientist

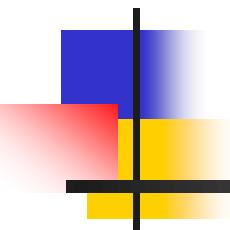
- <http://i1.wp.com/blog.datacamp.com/wp-content/uploads/2014/08/How-to-become-a-data-scientist.jpg>



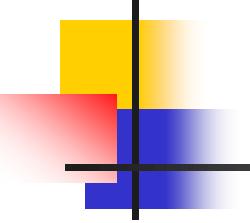


Useful Online Lectures

- 모두를 위한 머신러닝/딥러닝 강의
 - <http://hunkim.github.io/ml/>
- Learn real-world data science skills
 - <https://www.dataquest.io/>
- Data Science Specialization
 - by Johns Hopkins University
 - <https://www.coursera.org/specializations/jhu-data-science>
- Introduction to Recommender Systems
 - by University of Minnesota
 - <https://www.coursera.org/learn/recommender-systems/home/welcome>



Introduction to Data Mining



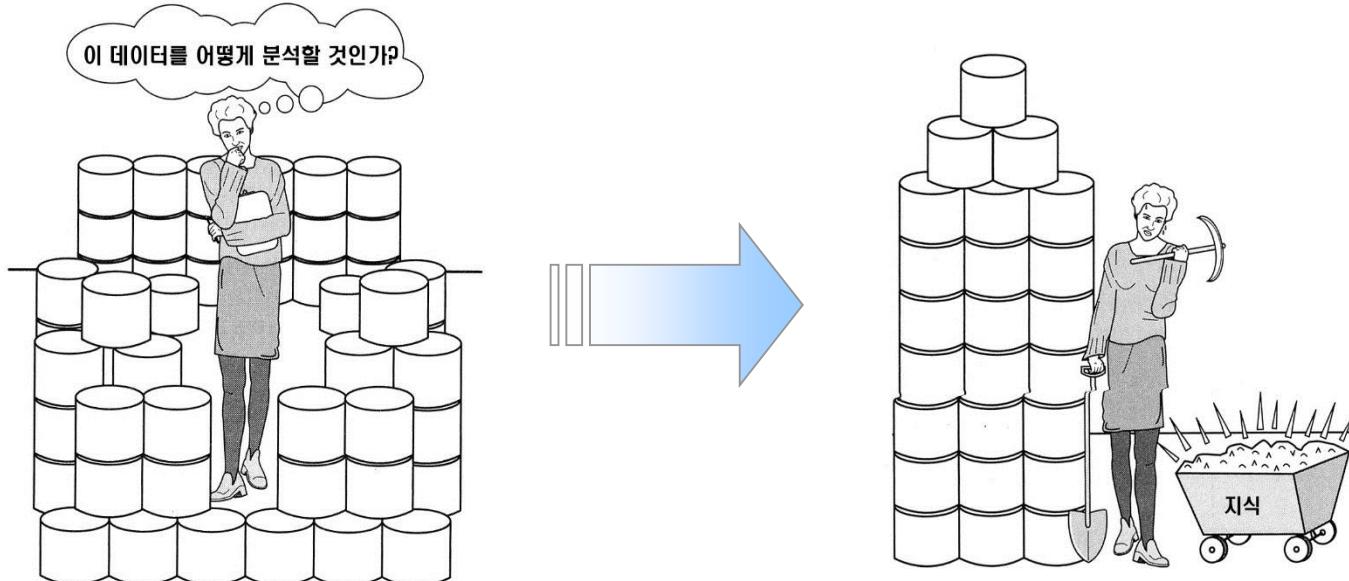
Machine Learning vs. Data Mining

- Limitations of explicit programming
 - Spam filter: many rules
 - Automatic driving: too many rules
- Machine learning
 - "Field of study that gives computers the ability to learn without being explicitly programmed", Arthur Samuel (1959)
- Data Mining
 - Data analysis processes that apply ML techniques to solving real world problems

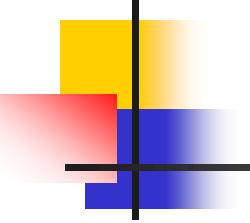
Concept of Data Mining

■ 데이터마이닝(Data Mining)의 정의

- 대량의 데이터로부터 새롭고 의미 있는 정보를 추출하여 의사결정에 활용하는 작업



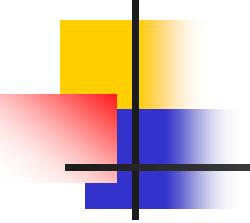
- 지식발견(KDD: Knowledge Discovery in Database)
- 정보발견(Information Discovery), 정보수확(Information Harvesting)
- 정보고고학 (Data Archeology), 자료패턴처리 (Data Pattern Processing)



Concept of Data Mining

■ 데이터마이닝의 다양한 정의

- 데이터베이스에서 지지발견은 데이터에 있는 유효하고, 잠재적으로 이용가능하며 궁극적으로 이해할 수 있는 패턴을 식별하는 중요한 프로세스 (Fayyad et al., “Advance in Knowledge Discovery and Data Mining,” 1996)
- 데이터 마이닝은 비즈니스 문제를 해결하기 위해 현재 조치를 취할 수 있고, 명시적이며 새로운 정보를 추출하기 위해 세부적인 데이터를 분석하는 프로세스이다.(NCR)
- 데이터 마이닝은 큰 데이터베이스로부터 이전에 알려지지 않고, 궁극적으로 이해가능한 정보를 추출 및 중요한 비즈니스 의사결정을 하는 프로세스이다.(IBM)
- 데이터 마이닝은 비즈니스 우위를 위해 이전에 알려지지 않은 패턴을 발견하기 위해 많은 양의 데이터를 선택하고, 탐색 및 모델링하는 프로세스이다. (SAS Institute)
- **데이터 마이닝은 기업의 경영 활동에서 발생하는 대용량 데이터에서 데이터 간의 관계·패턴·규칙 등을 찾아내고 모형화해 유용한 경영 정보로 변환시키는 일련의 과정이다.**



Statistical Analysis vs. Data Mining

- 전통적 통계분석

대상집단이 있으며, 모집단의 분포 혹은 모형 등 여러 가지 가정을 전제로 하게 되며 이 전제 조건하에서 분석을 실시

→ 표본의 관찰을 통해 모수 전체를 추론하는 과정

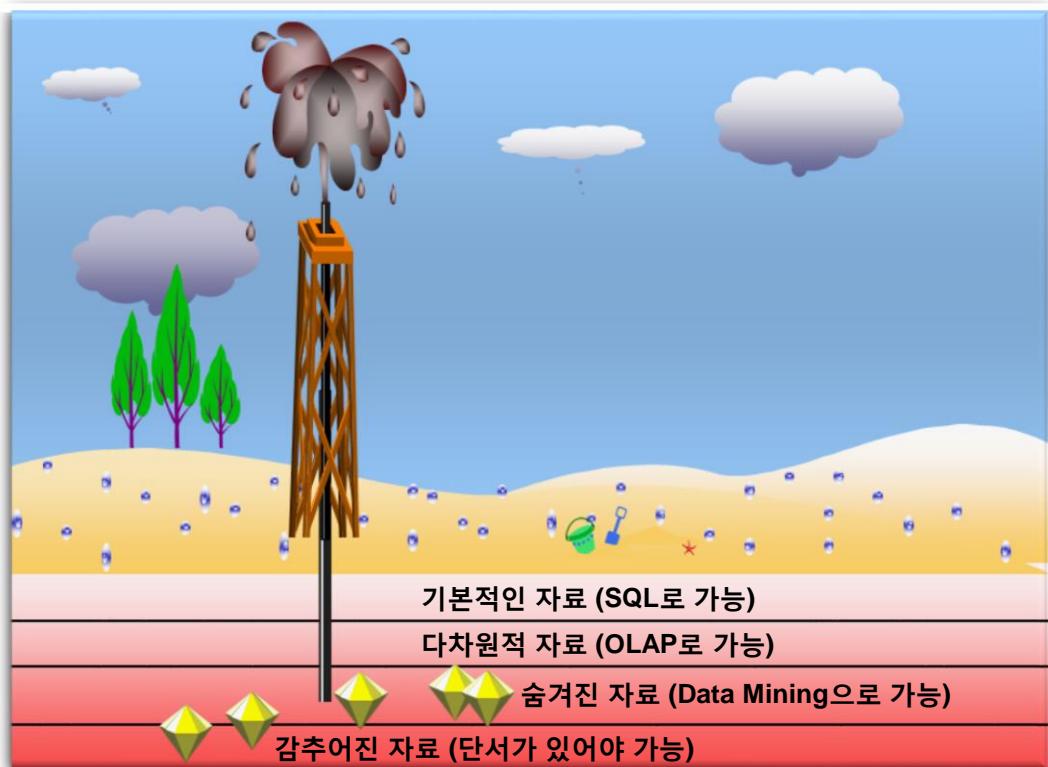
- 데이터마이닝

표본조사/실험에서 필연적으로 수반되는 분포라든가 모형에 대한 전제조건이 필요하지 않음

→ 모집단의 전체자료를 이용한 정보화하는 과정

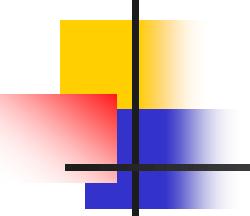
SQL/OLAP/Data Mining

- ✓ 데이터 마이닝은 분명한 미래의 분석 방향이다. 미래의 CEO들은 데이터 마이닝을 이용한 분석 결과를 첨부하지 않은 보고서는 검토의 가치가 없다고 판단할 것이다. – Elder Research



SQL/OLAP/Data Mining

- ✓ 2008년 4월의 매출건수는 ?
→ SQL
- ✓ 2008년 4월의 지역별 매출건수는 ?
→ OLAP
- ✓ 2008년 4월에 제품을 구매한 홍길동이 향후 6개월 이내에 추가 구매를 할 가능성은?
→ Data Mining



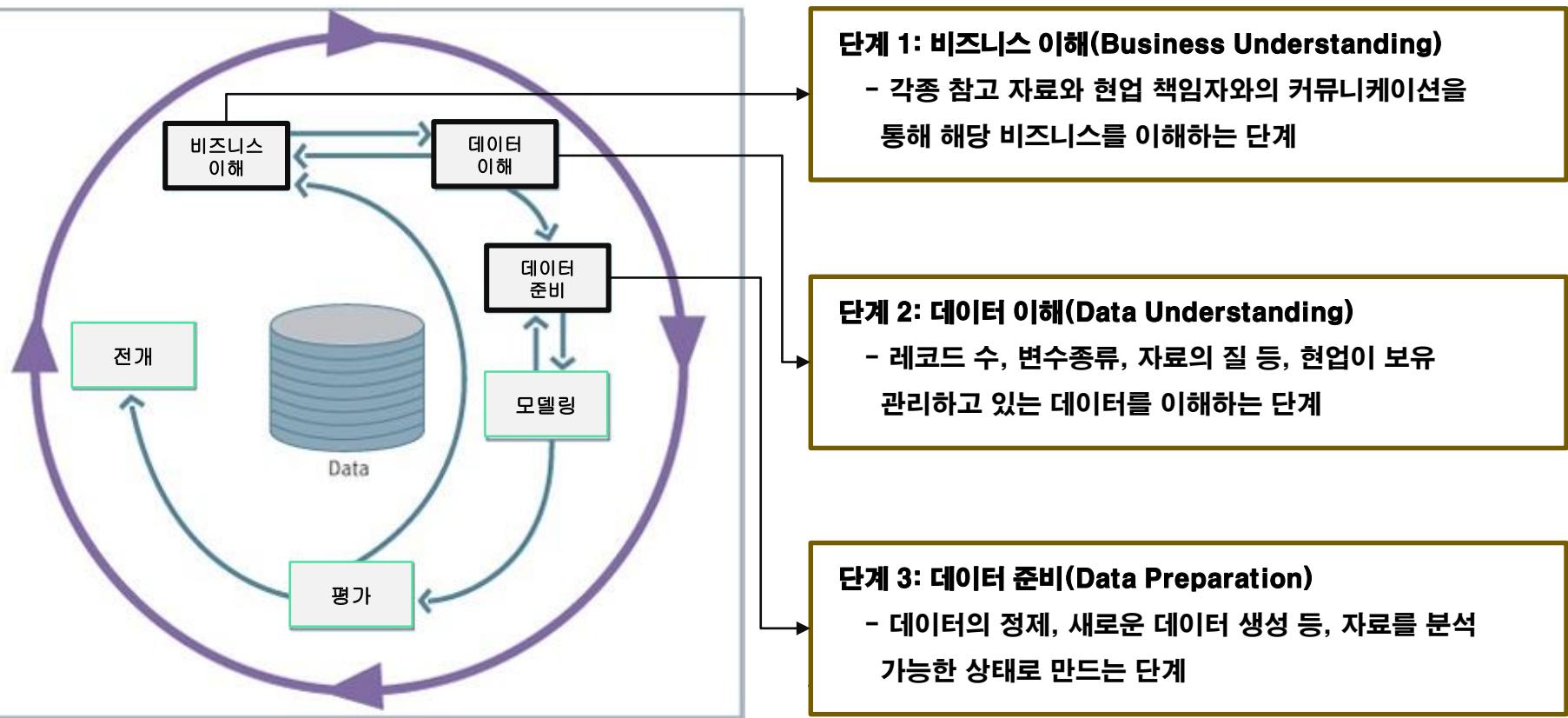
Data Mining S/W

구분	SAS Enterprise Miner	IBM SPSS Modeler	R	Python
특징	데이터마이닝 프로세스 전반을 지원이 가능하고, 사용이 간편한 GUI를 통해 모델 구축의 가속화가 가능하다.	Data 핸들링에 강하고 사용하기 쉬운 사용자 인터페이스를 가지고 있다.	오픈 소스로서 누구나 자유롭게 실행하고 복사하고 수정하고 배포할 수 있다.	오픈 소스로서 누구나 자유롭게 실행하고 복사하고 수정하고 배포할 수 있다.
장점	대용량 데이터분석이 가능하고 활용영역이 다양하다.	자동 모델링 기능, 텍스트 분석 기능, 개체 분석 기능이 강화되었다.	코딩을 이용하기 때문에 다른 도구들에 비해 자유롭게 분석할 수 있다.	코딩을 통해 딥러닝 등 최신의 머신러닝 라이브러리를 사용할 수 있다.
단점	초반 작업 설정과 사용법을 습득하는데 다소 시간이 걸린다.	분석을 위한 설정과 연결 과정에 대한 프로세스가 다소 많은 편이다.	코딩의 어려움 때문에 전문가가 아닌 일반인은 이용하기 힘든 편이다.	코딩의 어려움 때문에 전문가가 아닌 일반인은 이용하기 힘든 편이다.
평가판 이용가능 기간	없음.	14일	무료이용가능	무료이용가능
홈페이지	http://www.sas.com/korea/	http://www.spss.co.kr/	http://www.r-project.org/	http://www.python.org/

Data Mining Process

▫ CRISP-DM : SPSS에서 제시하는 데이터마이닝 프로세스 (1/2)

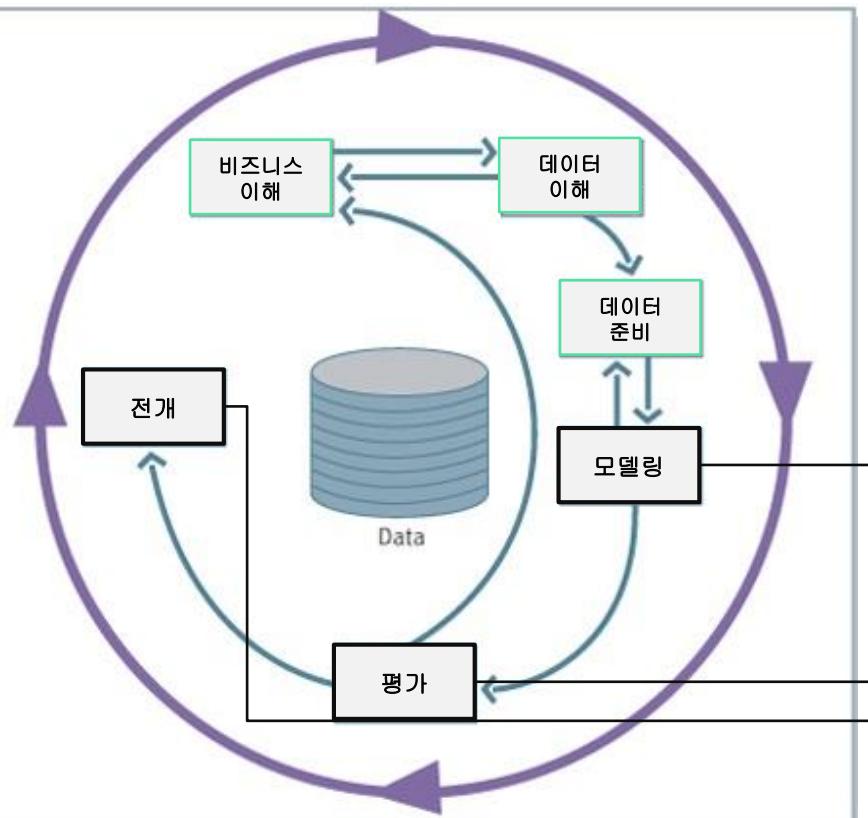
CRISP-DM(cross-industry standard process for data mining)은 데이터마이닝에 관련된 광범위한 업무의 범위를 다루고 있음.



Data Mining Process

▫ CRISP-DM : SPSS에서 제시하는 데이터마이닝 프로세스 (2/2)

CRISP-DM(cross-industry standard process for data mining)은 데이터마이닝에 관련된 광범위한 업무의 범위를 다루고 있음.



단계 4: 모델링 (Modeling)

- 자료 기술 및 탐색을 포함하여 필요한 각종 모델링을 하는 단계

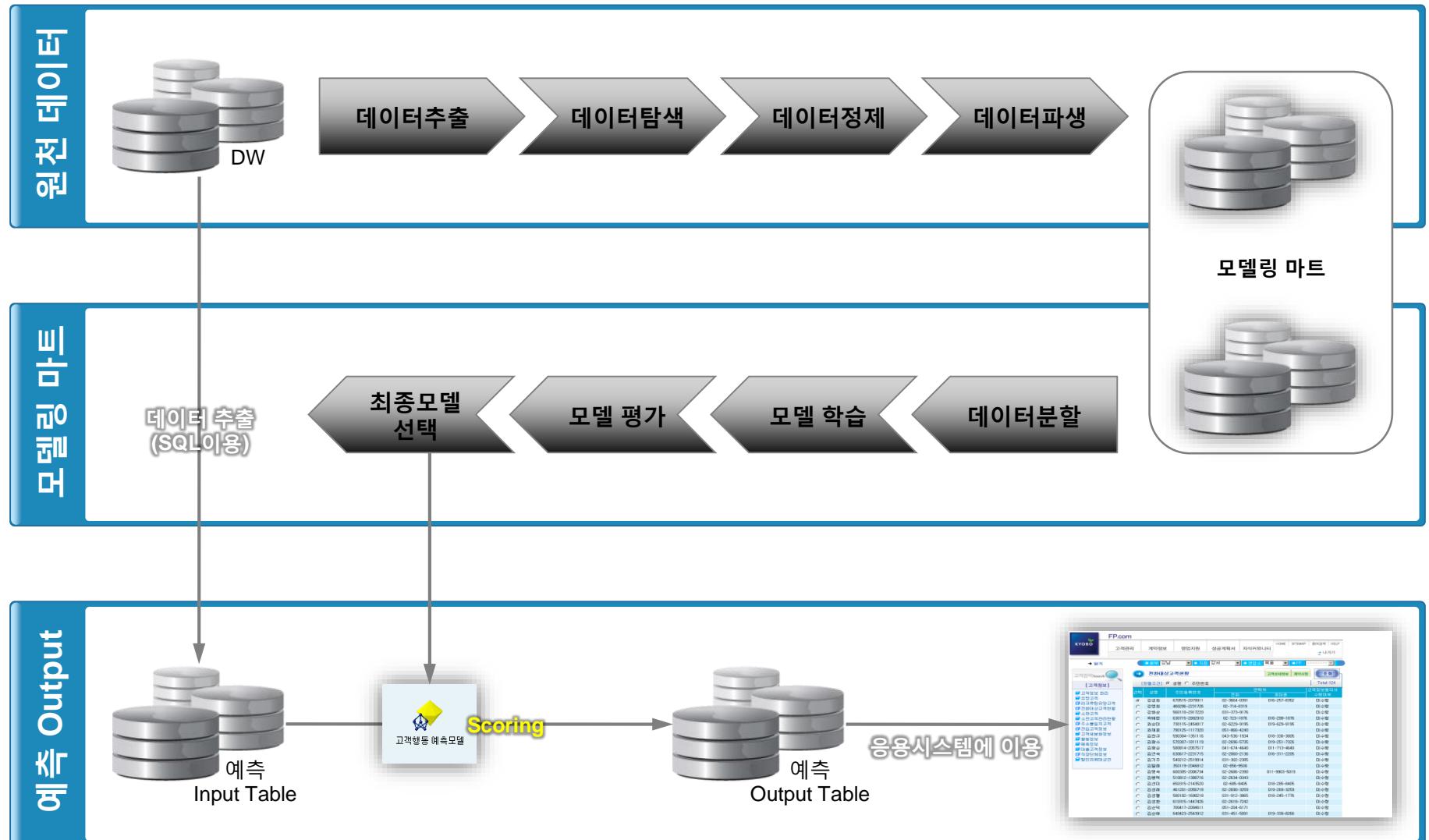
단계 5: 평가 (Evaluation)

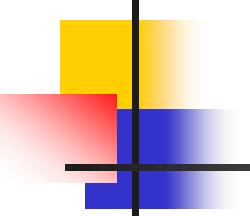
- 모형의 해석 가능 여부, 독립적인 새 자료에 적용되는 경우에도 재현 가능한지를 검토하는 단계

단계 6: 전개 (Deployment)

- 각 관리자에게 전달하여 필요한 조치를 취하는 등 검토가 끝난 모형을 실제 현업에 적용하는 단계

Predictive Modeling Process



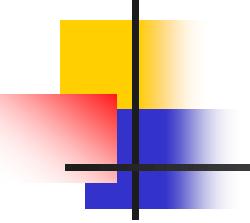


데이터마이닝 기법의 분류

- Supervised Modeling (지도학습, Predictive Analytics)
 - Estimation / Prediction (추정/예측: 연속형)
 - Linear Regression, Neural Network
 - Classification / Prediction (분류/예측: 이산형)
 - Decision Tree (C5.0) , Neural Network, SVM

용어의 유래: 어린아이가 말을 배우는 과정 (엄마가 *Supervisor* 역할)

- Unsupervised Modeling (비지도학습, Descriptive Analytics)
 - Clustering (군집화)
 - K-Means, SOM
 - Association rule mining (연관규칙탐사)
 - Apriori
 - Sequential rule mining(연속규칙탐사)

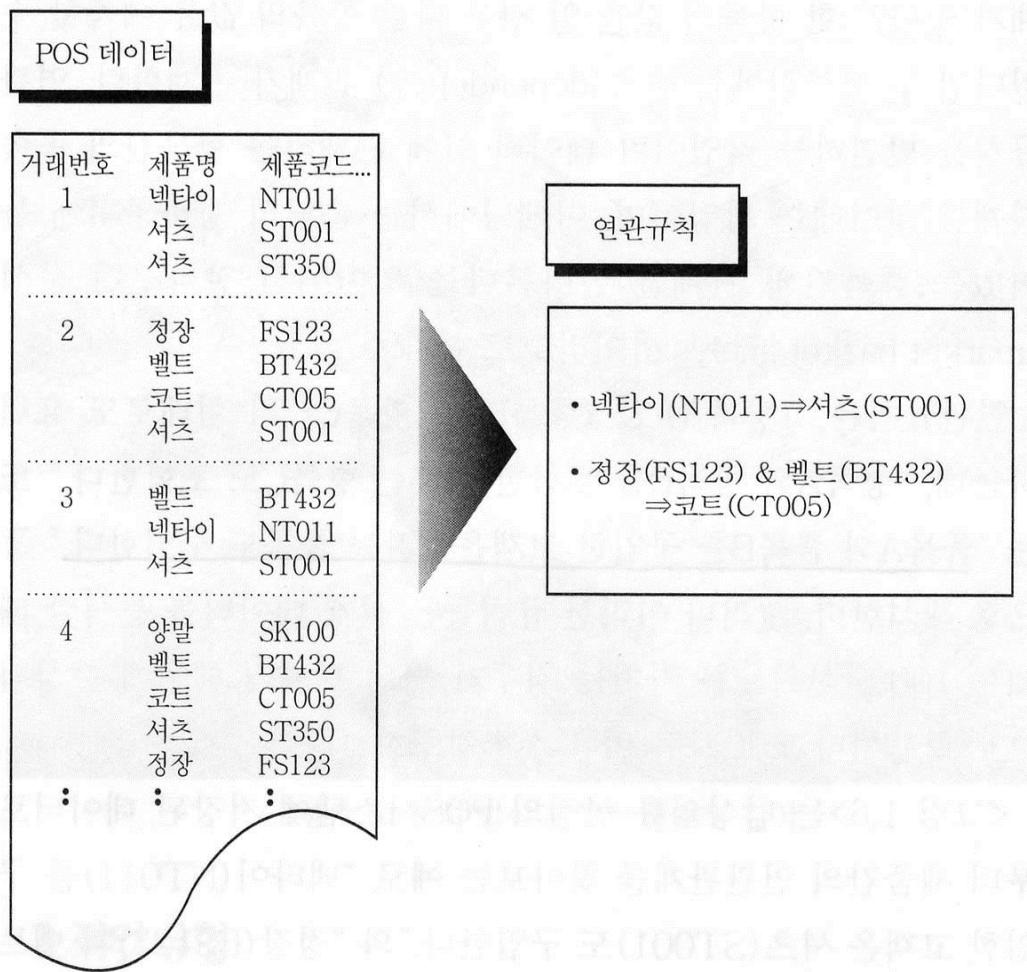


연관규칙탐사(Association Rule Mining)

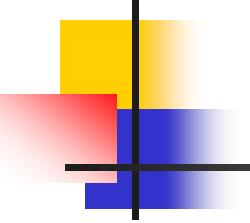
- 정의
 - 데이터 안에 존재하는 항목간의 종속 관계를 찾아내는 작업
- 장바구니 분석(market basket analysis)
 - 고객의 장바구니에 들어있는 품목 간의 관계를 발견
- 규칙의 표현
 - 항목 A와 품목 B를 구매한 고객은 품목 C를 구매한다.
 - (품목 A) & (품목 B) \Rightarrow (품목 C)
- 연관규칙의 활용
 - 제품이나 서비스의 교차판매
 - 매장진열, 첨부우편
 - 사기적발

연관 규칙

■ 연관 규칙의 예



[Source: 데이터마이닝, 장남식 외, 1999]



연속규칙탐사(Sequential Rule Mining)

- 정의
 - 연관 규칙에 시간 관련 정보가 포함된 형태
- 규칙의 표현
 - 새 냉장고를 구입한 고객 중 한달 이내에 새 오븐을 구입하는 경향이 많다.
- 연속규칙의 활용
 - 타겟 마케팅
 - 일대일 마케팅
- 전제조건
 - 고객의 구매내역(history) 정보가 반드시 필요함

연속 규칙

■ 연속 규칙의 예

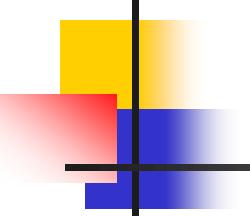
POS 데이터

회원번호	거래일	구입품목
1	99-02-01	B, C
	99-02-05	A
	99-02-19	D, E, H
2	99-02-07	A
	99-02-10	H
3	99-02-12	G
	99-02-20	A, C, D
	99-02-23	F
4	99-02-08	A, C
	99-02-18	B, H
5	99-02-21	A

연속규칙

A품목을 구입한 회원이 향후
H품목을 구입할 가능성은 75%이다.

[Source: 데이터마이닝, 장남식 외, 1999]

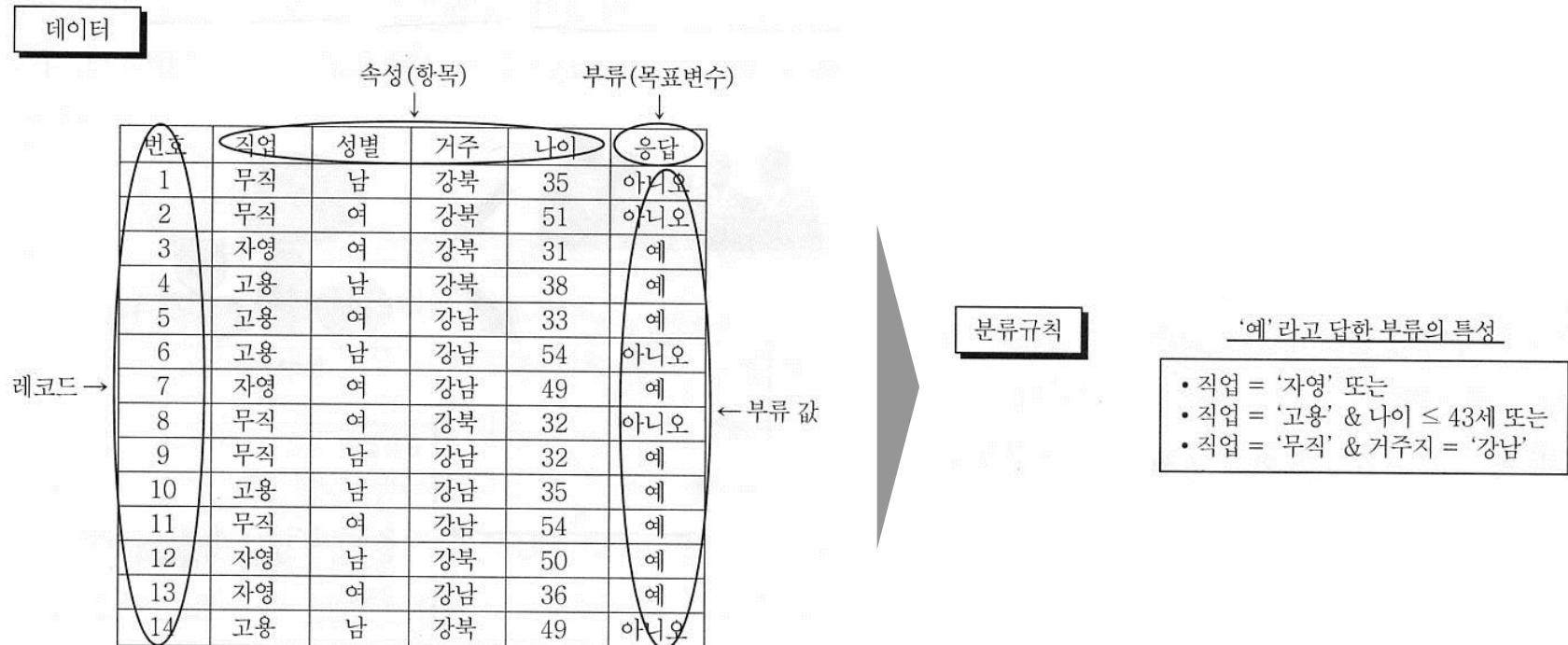


분류(Classification)

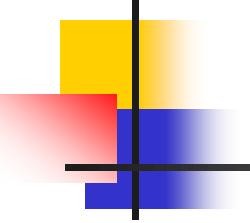
- 분류 프로세스
 - 과거의 데이터를 부류로 구분
 - 부류별 특성을 발견
 - 분류 모형 생성
 - 모형을 토대로 새로운 레코드의 분류 값 예측
- 분류의 활용
 - 고객의 신용등급 분류
 - 기업의 도산 예측
 - 프로모션 대상고객 선정
- 분류 기법
 - 의사결정나무(Decision Tree)
 - 인공신경망(Neural Network)
 - SVM(Support Vector Machine)

분류(Classification)

■ 분류의 예



[Source: 데이터마이닝, 장남식 외, 1999]

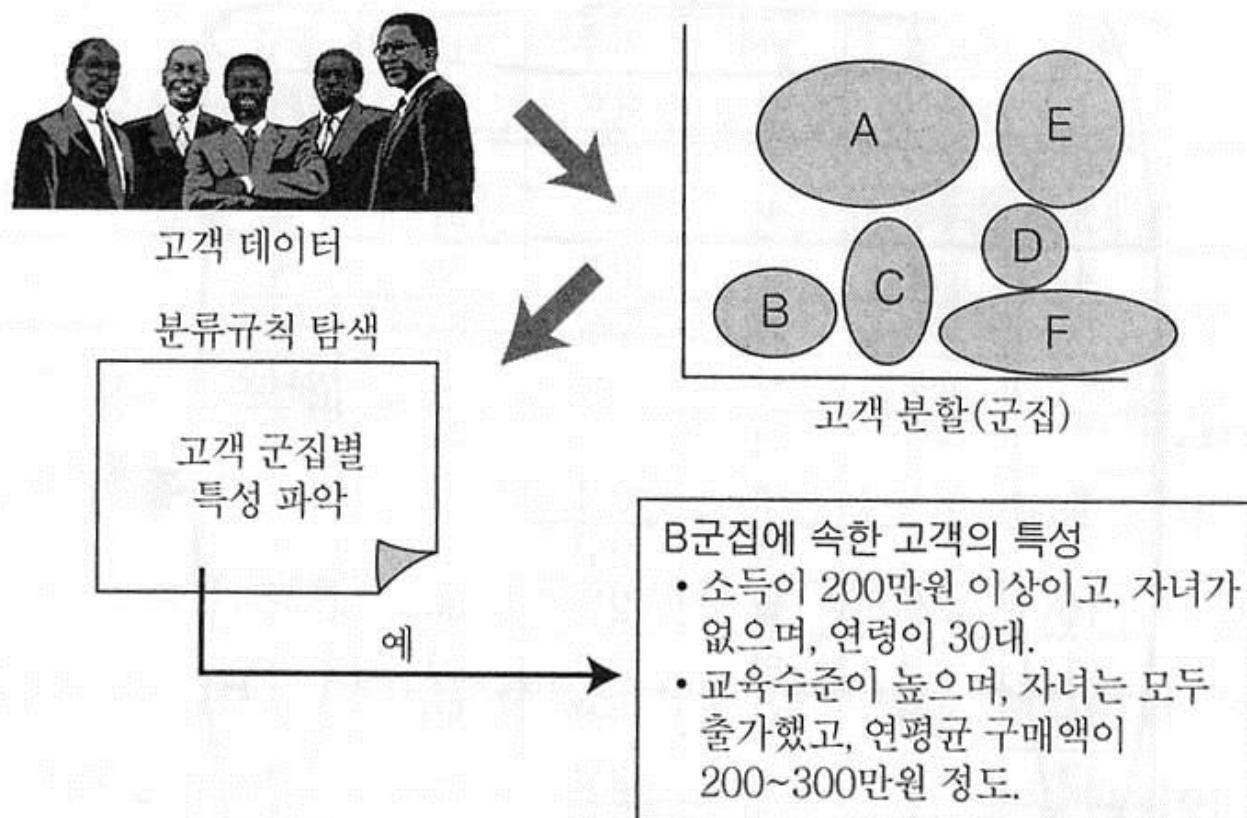


군집화(Clustering)

- 정의
 - 레코드들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업
- 군집화의 활용
 - 다른 데이터마이닝 기법의 선행 작업으로써 많이 이용
- 분류 vs 군집화
 - 분류 값의 유무
- 군집화 기법
 - 계층적 군집분석
 - 비계층적 군집분석: K-means, EM Algorithm, SOM(Self-Organizing Map)

군집화(Clustering)

■ 군집화의 예



[Source: 데이터마이닝, 장남식 외, 1999]

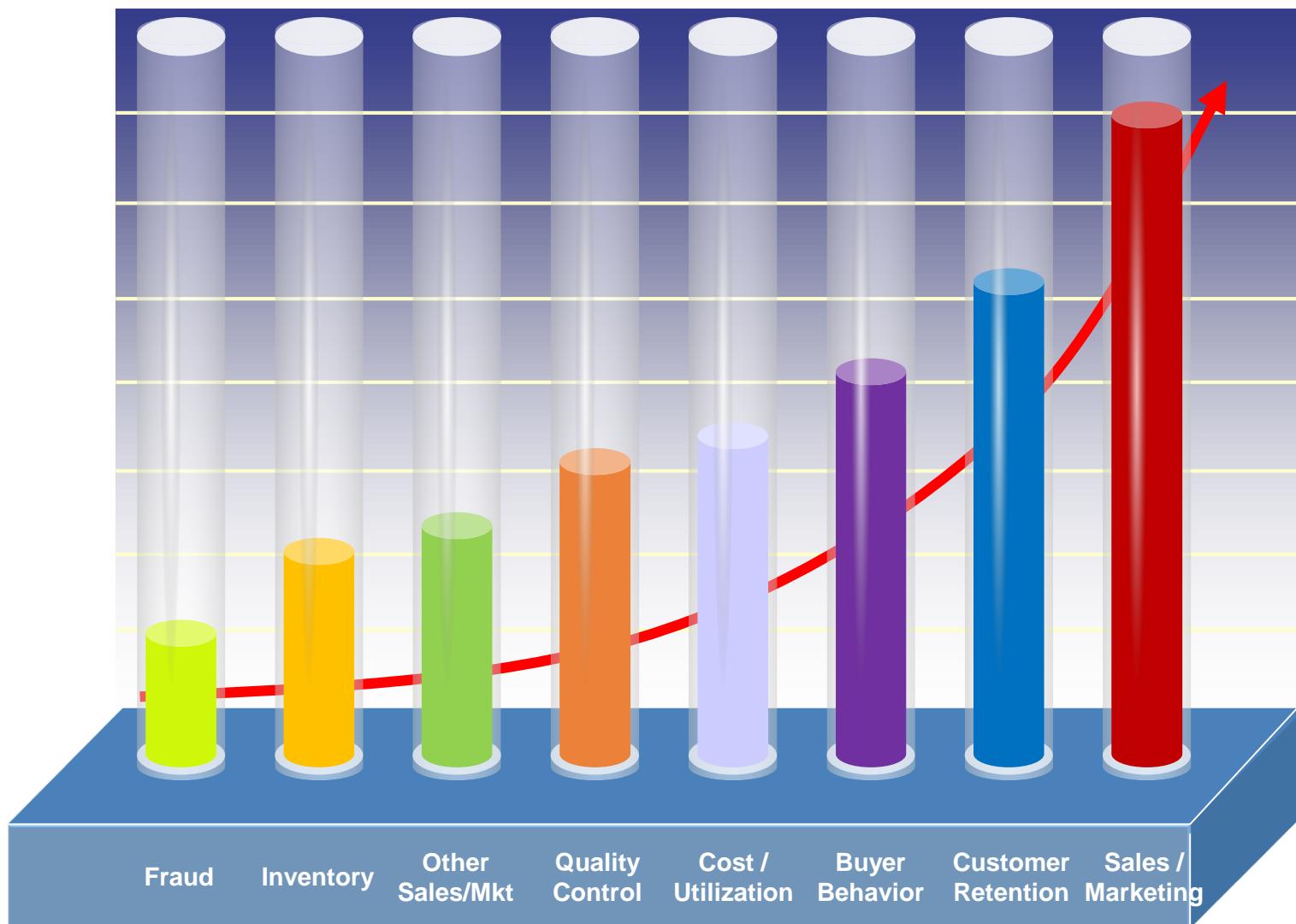
데이터 분할 (Data Partitioning)

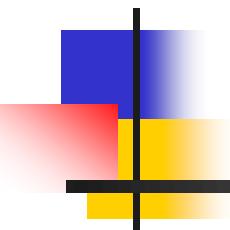


- ❖ 데이터를 용도에 따라 분할
- ❖ 학습데이터 (training data) → 모델 적합
- ❖ 검증데이터 (test data) → 모델 평가
- ❖ 50% – 50% 분할
- ❖ 대안 : 60% – 40% 분할, 75% – 25% 분할

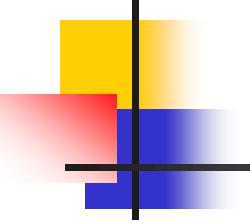
학습데이터 = Training data
검증데이터 = Test data

Data Mining Applications



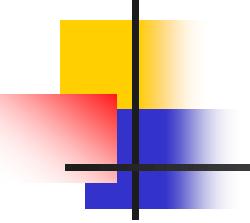


Data Mining Process Demo



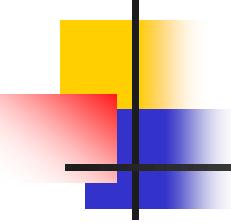
Step 1: Business Understanding

- 한 은행이 새로운 개인연금상품(PEP)을 신설하여 기존 고객들을 대상으로 가능한 많은 계좌를 유치하고자 한다
- 고객의 금융상품(PEP: Personal Equity Plan, 연금보험) 구매 여부 예측에 의한 신규고객 창출
 - 고객 프로파일 개발
 - 다이렉트 메일 광고 효율성 제고
 - 타겟 메일링에 의한 응답률 제고
- 분석 절차
 - 1) 기존고객 DB로부터 시험메일 발송을 위한 표본고객목록을 추출
 - 2) 새로운 금융상품(PEP)의 제안 메일을 발송
 - 3) 고객의 반응을 기록
 - 4) R을 이용하여 캠페인 결과를 분석



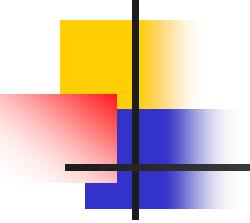
Step 2: Data Understanding

- 학습용 데이터 300건 (`pepTrainSet.csv`)
- 검증용 데이터 300건 (`pepTestSet.csv`)
- 신규고객 데이터 200건 (`pepNewCustomers.csv`)



Step 3: Data Preparation

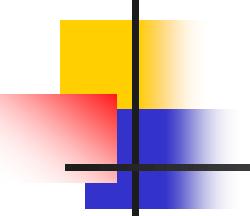
```
> setwd("d:/datamining")  
  
> train <- read.csv("pepTrainSet.csv", stringsAsFactors=F)  
> train <- subset(train, select=-c(id))  
> test <- read.csv("pepTestSet.csv", stringsAsFactors=F)  
> newd <- read.csv("pepNewCustomers.csv", stringsAsFactors=F)  
  
> train$pep <- factor(train$pep)  
> test$pep <- factor(test$pep)
```



Step 4: Modeling

```
> install.packages("caret")
> install.packages("ROCR")
> install.packages("C50")

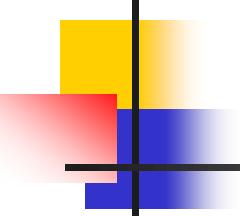
> library(caret)
> library(ROCR)
> library(C50)
```



Step 4: Modeling (*Cont.*)

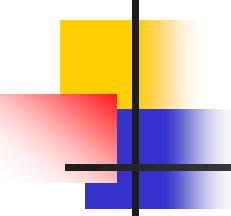
```
> # first candidate model: Decision Tree (C5.0)
> c5_options <- C5.0Control(winnow = FALSE, noGlobalPruning = FALSE)
> c5_model <- C5.0(pep ~ ., data=train, control=c5_options, rules=FALSE)
> summary(c5_model)
> plot(c5_model)

> # second candidate model: Logistic Regression
> lm_model <- glm(pep ~ ., data=train, family = binomial)
> summary(lm_model)
```



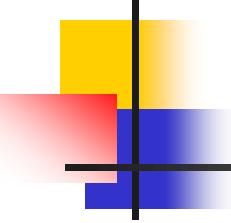
Step 5: Evaluation

```
> # Model evaluation by Confusion Matrix  
  
> test$c5_pred <- predict(c5_model, test, type="class")  
> test$c5_pred_prob <- predict(c5_model, test, type="prob")  
> confusionMatrix(test$c5_pred, test$pep)  
  
> test$lm_pred <- ifelse(predict(lm_model, test, type="response") > 0.5, "YES",  
  "NO")  
> test$lm_pred_prob <- predict(lm_model, test, type="response")  
> confusionMatrix(test$lm_pred, test$pep)
```



Step 5: Evaluation (*Cont.*)

```
> # Model evaluation by ROC chart  
  
> c5_pred <- prediction(test$c5_pred_prob[, "YES"], test$pep)  
> c5_model.perf <- performance(c5_pred, "tpr", "fpr")  
  
> lm_pred <- prediction(test$lm_pred_prob, test$pep)  
> lm_model.perf <- performance(lm_pred, "tpr", "fpr")  
  
> plot(c5_model.perf, col="red")  
> plot(lm_model.perf, col="blue", add=T)  
> legend(0.7, 0.7, c("C5","LM"), cex=0.9, col=c("red", "blue"), lty=1)
```



Step 6: Deployment

```
> newd$c5_pred <- predict(c5_model, newd, type="class")
> newd$c5_pred_prob <- predict(c5_model, newd, type="prob")
> target <- subset(newd, c5_pred=="YES" & c5_pred_prob[, "YES"] > 0.8)
> write.csv(target[order(target$c5_pred_prob[, "YES"], decreasing=T), ],
  "dm_target.csv", row.names=FALSE)
```