DATA MINING

2010.06.09

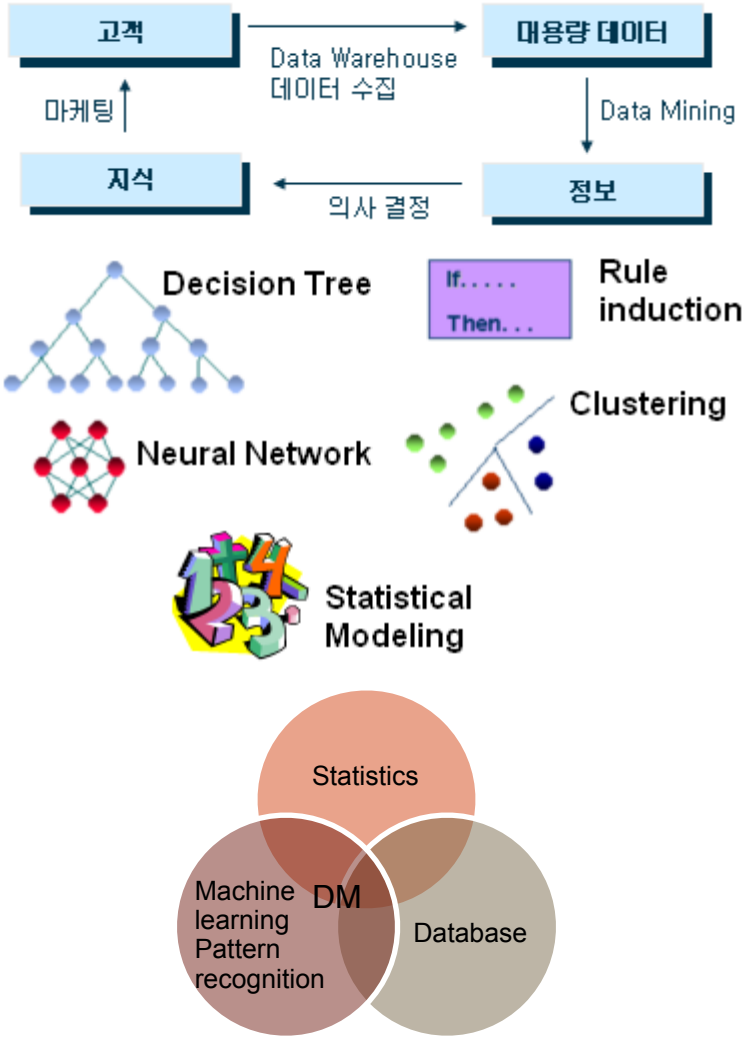
PROF. SEHYUG KWON

DEPT. OF STATISTICS
http://wolfpack.hnu.ac.kr

HAN NAM 한남대학교 Hannam University

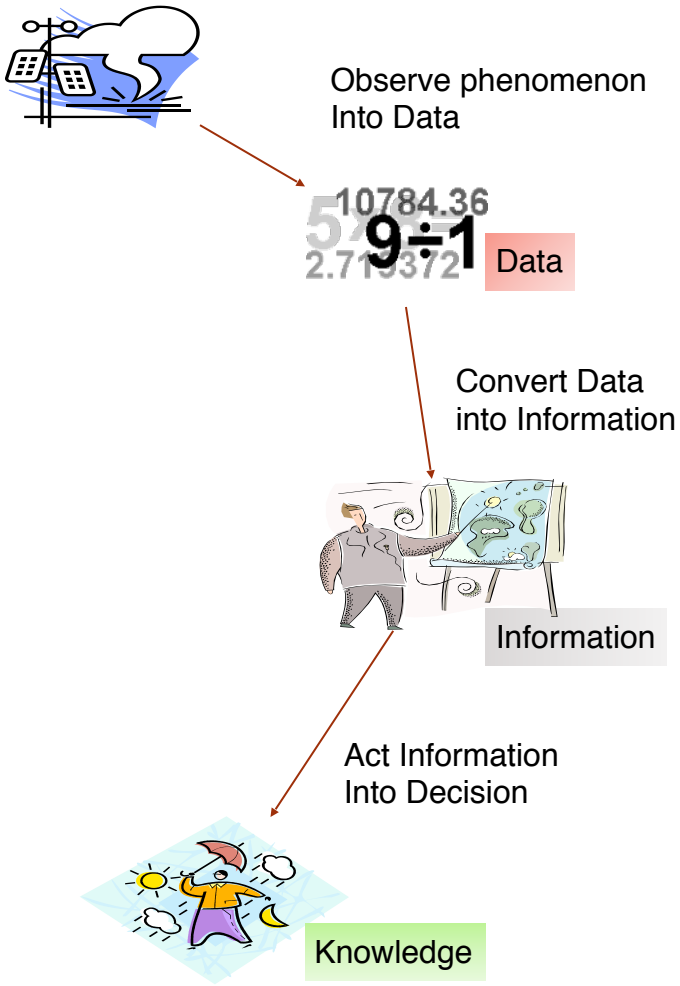
Introduction (definition)

- DM is …
  - A process of identifying and/or extracting previously unknown, non-trivial, unanticipated, important information from large sets of data. –Wolfgang Martin-
  - The exploration and analysis, by automatic or semiautomatic means, of large quantities of data to discover meaningful patterns or rules.
  - One of application tools for Data Warehousing to end-users for information.
  - A modern Exploratory Data Analysis, about looking at data to see what it seems to say.
  - A simply and automate the statistical process, decision supporter
  - the process of analyzing data from different perspectives and summarizing it into useful information



Introduction (glossary)

- Data
  - any facts, numbers, or text with context (story) that can be processed by a computer.
  - organizations are accumulating vast and growing amounts of data in different formats and different databases
- Information
  - The patterns, associations, or relationships among all this data can provide *information*
- Knowledge
  - Information can be converted into *knowledge* about historical patterns and future trends



Introduction (characteristics and vendors)

- DM 특징
  - Handling huge observational data
  - Computer intensive method
  - Ah-hoc and experience based method
  - Generalization
  - Obtaining Business information

- DM is not
  - Data Warehousing
    - a repository of an organization's electronically stored data
  - Structural Query Language
    - database computer language
  - Query
    - a form of questioning, in a line of inquiry
  - OLAP
    - On-Line Analytical Process
  - Data visualization

- DM vendors
  - SAS E-minor
  - SPSS Clementine
  - Insightful Minor
  - Oracle Darwin,
  - Angoss Knowledge studio

- Applications of DM is
  - CRM
  - Bio-informatics

http://wolfpack.hnu.ac.kr

Why DM?

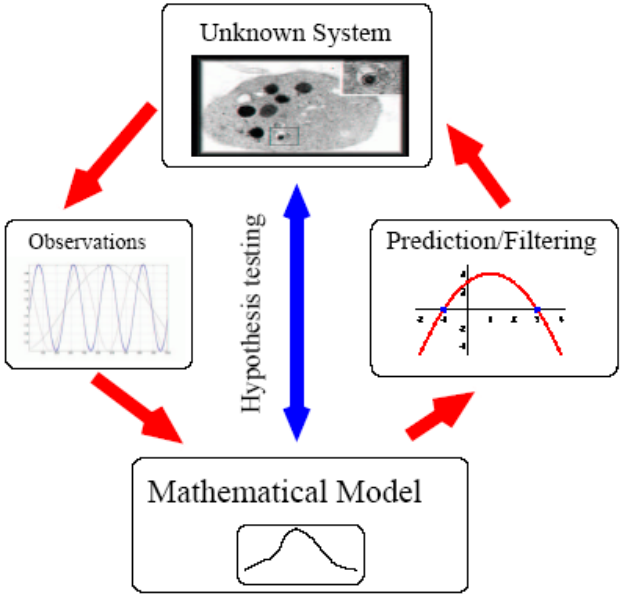
- Diaper and beer (시작)
  - Market basket theory (시장 바구니 이론)
- Information is a secrete weapon:
  - 고객은 기업과 직접 만날 수 없어도 자신들의 요구, 선호도, 만족도, 하물며 개인 사정까지도 알아 주기 원한다. 예전 구멍 가게 주인으로부터 받았던 그 관심으로… CRM (Customer Relationship Management 고객 관리 경영)
  - 고객들이 함께 사는 품목들을 조사하여 구매 동선을 줄인다.
  - 좋은 고객을 유지하고, 불량 고객은 정리하고, 불량 고객이 될 가능성이 보이는 고객을 미리 탐지하여 비용을 절감하고 이윤을 높인다. (예)신용 카드 회사
  - 떠나는 고객 원인 분석, 새로운 고객 창출 방법에 대한 정보를 얻는다. (예) AT&T 50$ 쿠폰

- Increase computing power (development of computer and software)
  - 컴퓨터 대용량, 초고속화, 관련 통계소프트웨어 등장
  - OLTP과 Data warehouse 발달
  - DM 관련 소프트웨어 발달: SAS E-minor, SPSS Clementine
  - DM을 넘어 Text mining이다. Data Integration (예) 소비자 불만 처리 관련 게시판, 인터넷 연구 자료 수집
- Statistical and learning algorithms
  - KDD (Knowledge Discovery in Database) DB로부터 지식을 추출하는 과정
  - Machine Learning 인공지능(AI)의 한 분야 자동적인 학습기번 설계
  - Patter Recognition: 공학, 문자 인식 또는 이미지 분류
  - Bioinformatics: 생명정보학 (생물, 공학, 통계학)

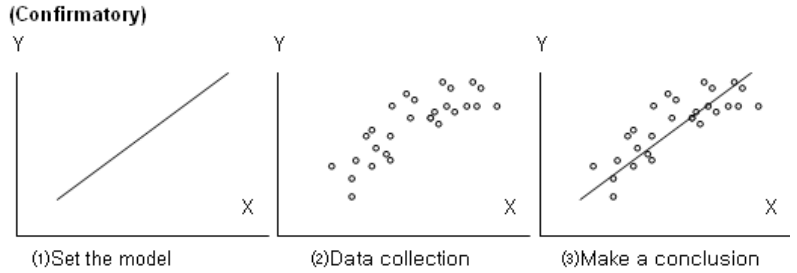
http://wolfpack.hnu.ac.kr

Statistics

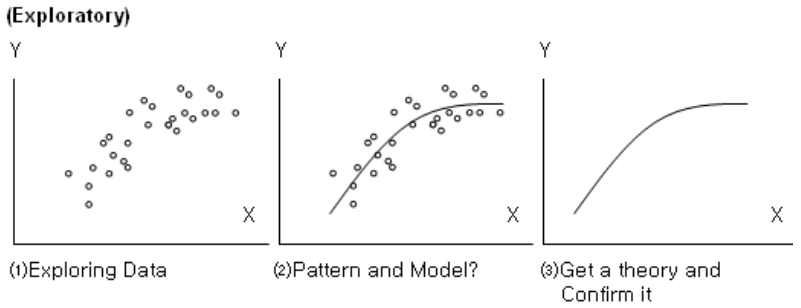
- Definition
  - Statistics is about data.
  - Statistics is a guide to the unknown.



- CDA
  - Set statistical Hypothesis or model
  - Data Collection
  - Confirm the theory based on the statistical results



- EDA
  - Data Collection
  - Summarize and Represent the data graphically
  - Get a (tentative) theory (Pattern) and confirm it.



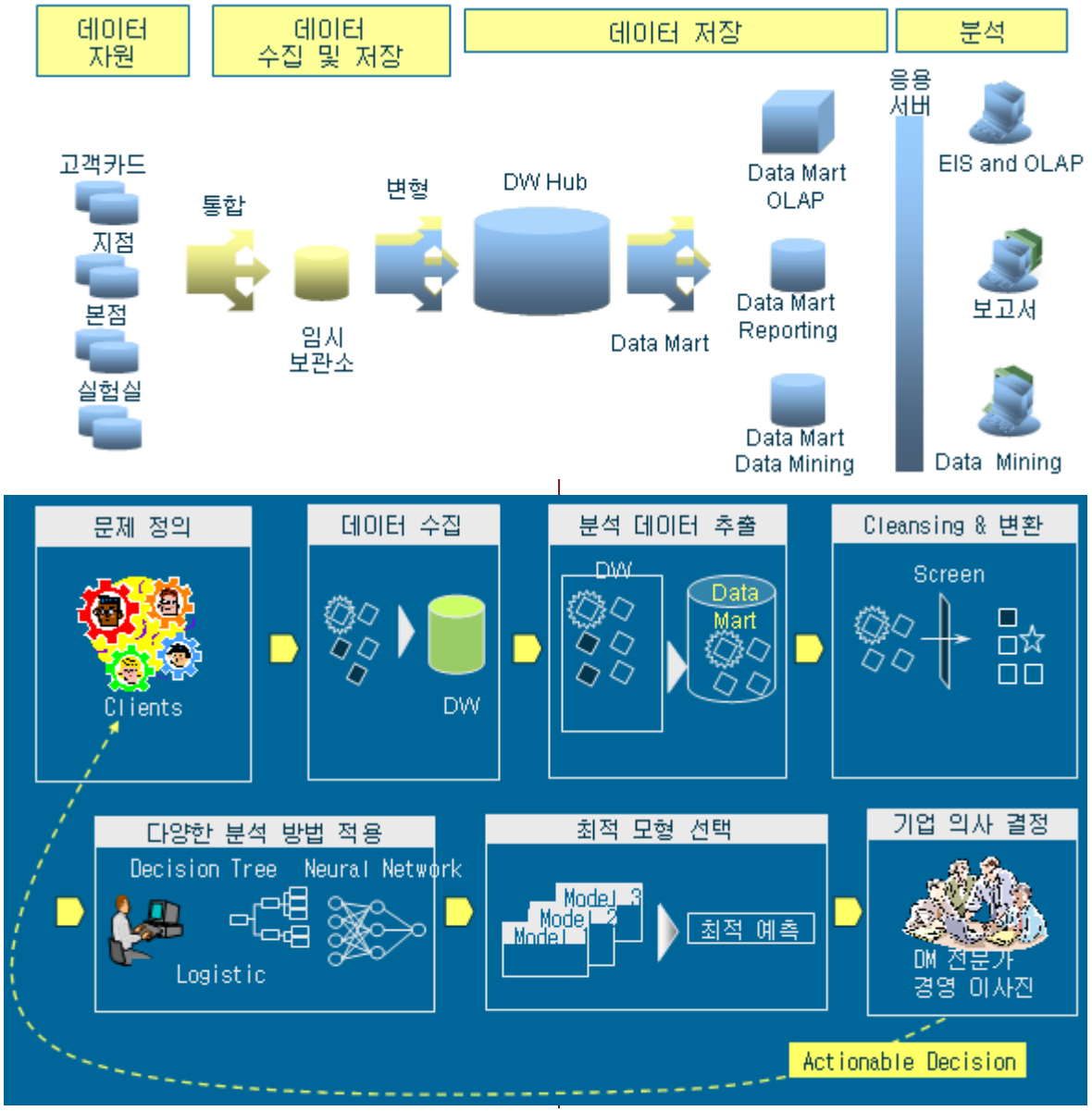
http://wolfpack.hnu.ac.kr

Where DM?

- Biz
  - DM 마케팅: Grocery Safeway & Pepsi
    - 목표 마케팅
    - 고객 세분화(segmentation): 충성고객, CRM, Direct Mail Marketing
    - 고객 성향변동 (churn): 이탈 고객 attrition
    - 교차판매 (cross sales)
    - Market basket theory
  - 신용 평가
    - Scoring
  - Credit card fraud
    - 판별분석
- Government
  - FBI (criminal)
  - IRS (tax evasion)
  - National Statistics
- Sports
  - statistics 4 game and players
- Web
  - text mining
  - web log analysis
- Clinic
  - Bioinformatics
  - Pre-diagnosis
  - Actuarial statistics

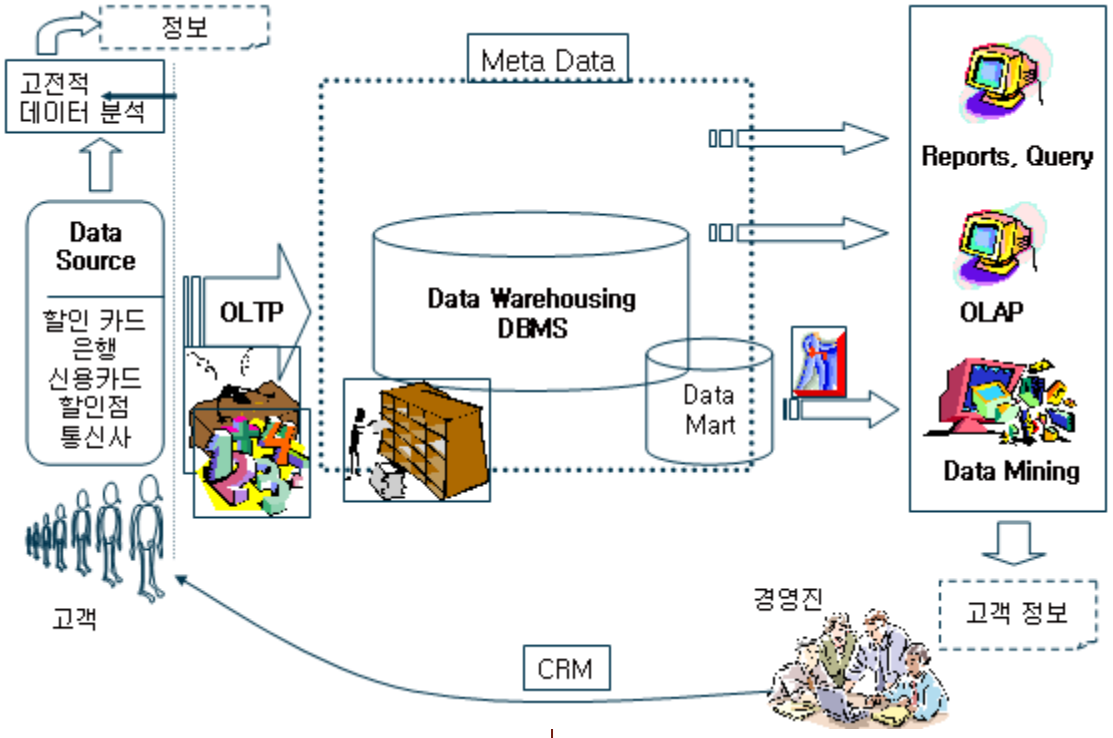
http://wolfpack.hnu.ac.kr

DM Flow



http://wolfpack.hnu.ac.kr

DM Architecture



http://wolfpack.hnu.ac.kr

DM Architecture (glossary)

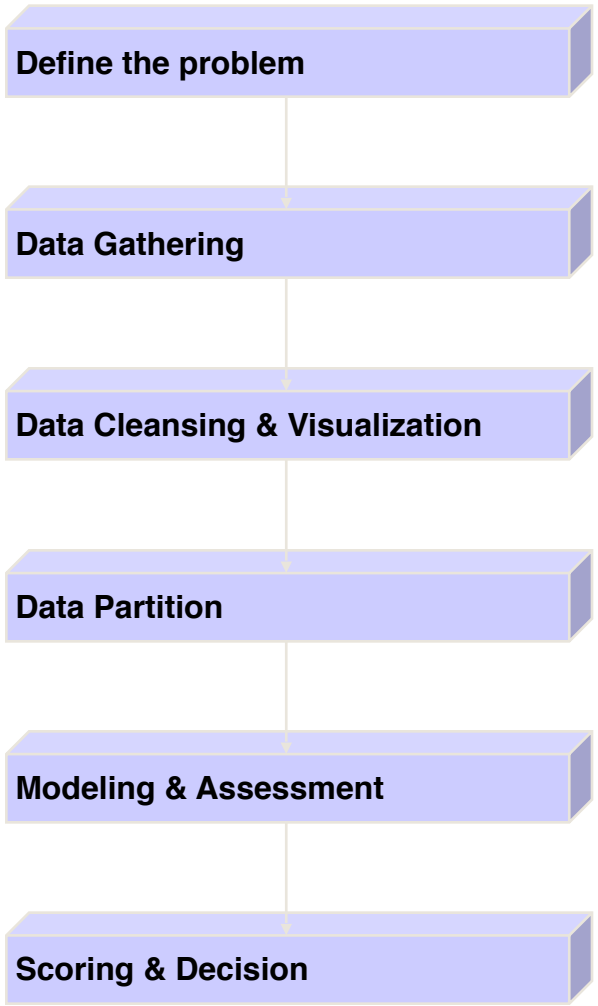
- [OLAP: On-Line Analytical Process]
  - 사용자로 하여금 대용량 데이터로부터 원하는 정보를 한눈에 파악할 수 있도록 표나 그래프를 제공한다.
- [Data Mart]
  - 전사적인 데이터베이스 혹은 자사 DW 다른 회사로부터 넘겨 받은 database로부터 원하는 정보를 얻기 위한 분석을 목적으로 변형시킨 데이터를 의미한다.
- [Database]
  - 사용자가 데이터에 쉽게 접근하여 원하는 작업을 처리할 수 있도록 구성된 데이터의 집합체이다.
- [DBMS: Data-Base Management Server]
  - 데이터베이스 관리 시스템. 다수의 컴퓨터 사용자들이 데이터베이스 안에 데이터를 기록하거나 접근할 수 있도록 해주는 프로그램이다.

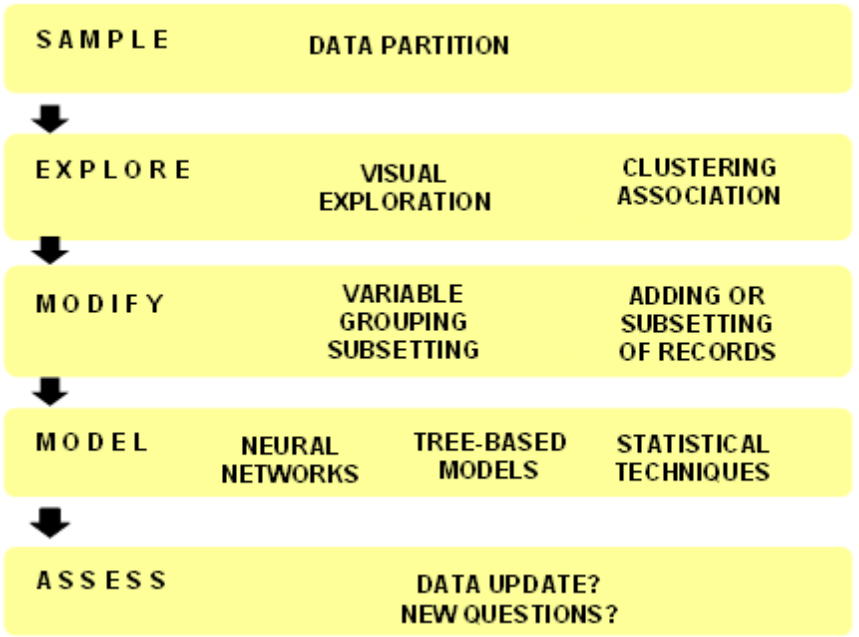
- RDBDS[Relational DBMS]
  - 관계형 데이터베이스, 1970년 IBM의 E. F. Codd가 개발, 일련의 정형화된 테이블로 구성된 데이터 항목들의 집합체로서 테이블을 재구성하지 않아도 데이터에 다양한 방법으로 접근하거나 조합 가능
- 용어 차이

Statistics	RDB	Data Mining
Data set	Table, Database	Data set
Case	Row, Record	Record
Variable	Column, Field	Field
Independent	Column, Field	Predictor
Dependent	Column, Field	Prediction
Observation	Value	Value

http://wolfpack.hnu.ac.kr

DM Process



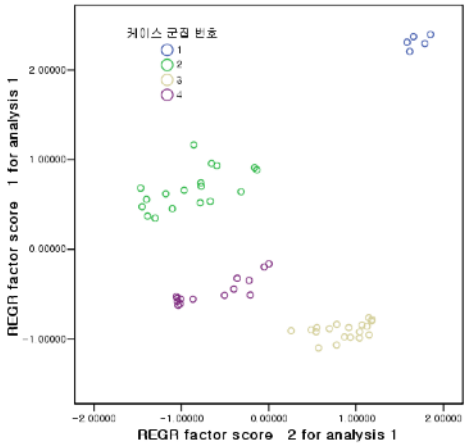


http://wolfpack.hnu.ac.kr

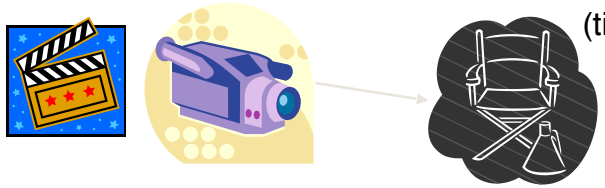
DM Technique (Descriptive methods)

- Association
  - Find the rule and patterns of individuals
- Clustering
  - Classify individuals using the similarity of them
  - Similarity? Euclidian distance of individuals
- Sequential Rule
  - Find rules that predict strong sequential dependency among different events

Market Basket Analysis
(cross-sectional, association)



Clustering
(Hierarchical, average)

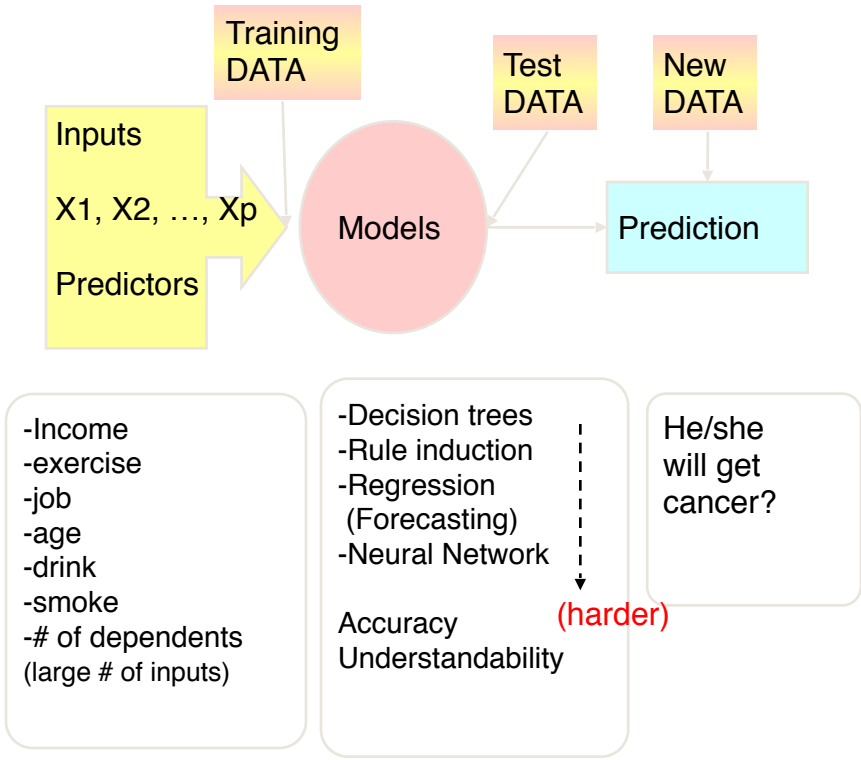


Sequence
(time chain)

http://wolfpack.hnu.ac.kr

DM Technique (Prediction methods; concept)

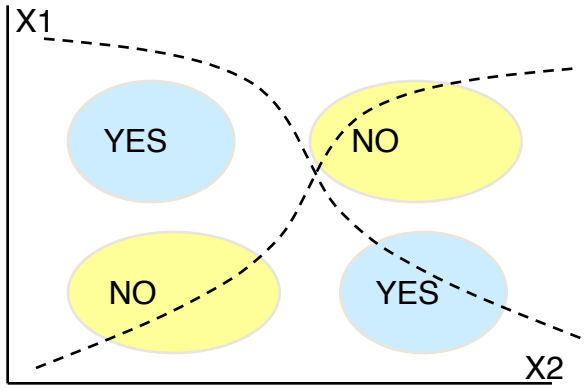
- Methodology
  - Classification: cancer or not
  - Forecasting: prob. of getting cancer
  - Association rule: combinations of inputs
  - Sequential detection: bankruptcy => smoking => get cancer)
  - Clustering/Discrimination: characteristics of inputs for personal bankruptcy



DM Technique (Prediction methods, Techniques)

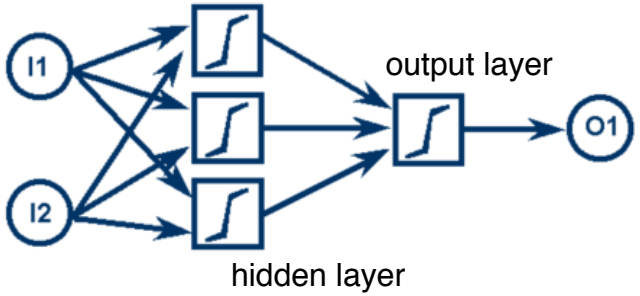
Logistic Regression

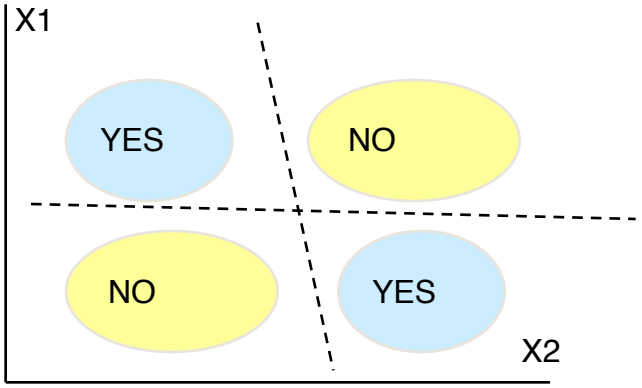
- Which inputs affect the target variable?
- If so, how much the effect is?
- Model: $Y=f(x_1, x_2, \ldots, x_p)$
- Prediction Y: binary, ordinal, metric
- function? linear function
- $x_1, x_2, \ldots, x_p$: metric / qualitative (non-metric)



Neural Network

- Based on biology
- Inputs is transformed via a network into a output
- difficult to understand, no intuitive understanding of results

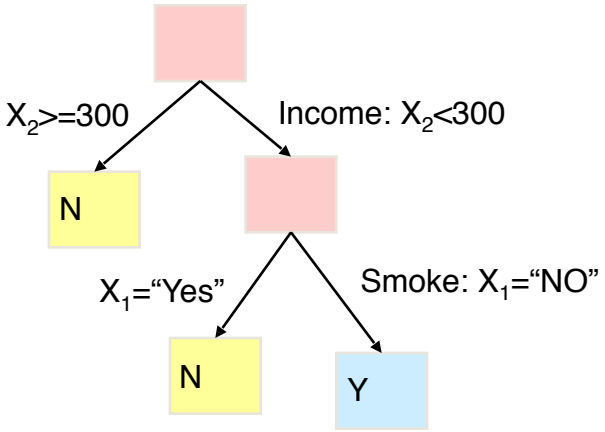




DM Technique (Prediction methods, Techniques)

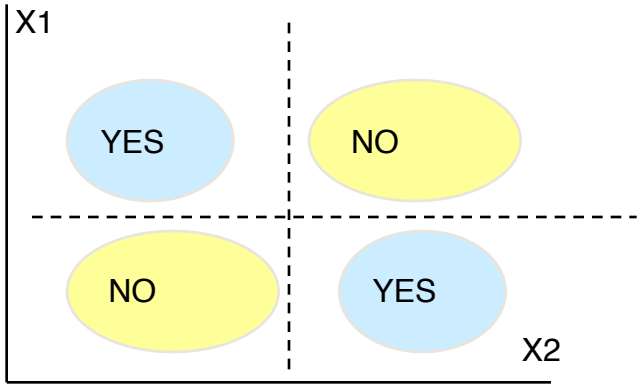
Decision Tree

- Hierarchical classification of individuals
- Algorithms: CHAID, CART
- Chi-square Automatic Interaction Detection: n-way split, categorical variables
- Classification And Regression Tree: binary split, continuous variables
- Similar: Fisher Discriminant Analysis, Logistics Regression



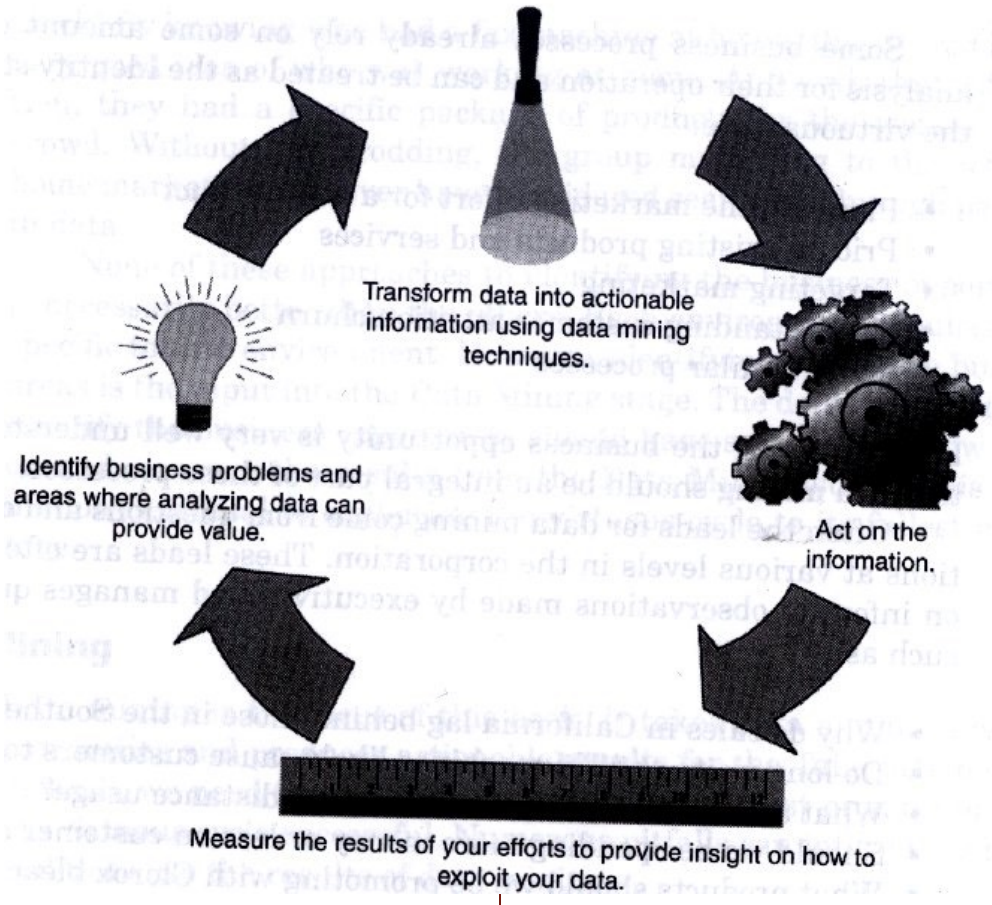
Rule Induction

- Find the rule of "If A, then B"
- Look at all possible variable combinations (huge) n^p



http://wolfpack.hnu.ac.kr

Virtuous cycle of Data Mining



http://wolfpack.hnu.ac.kr