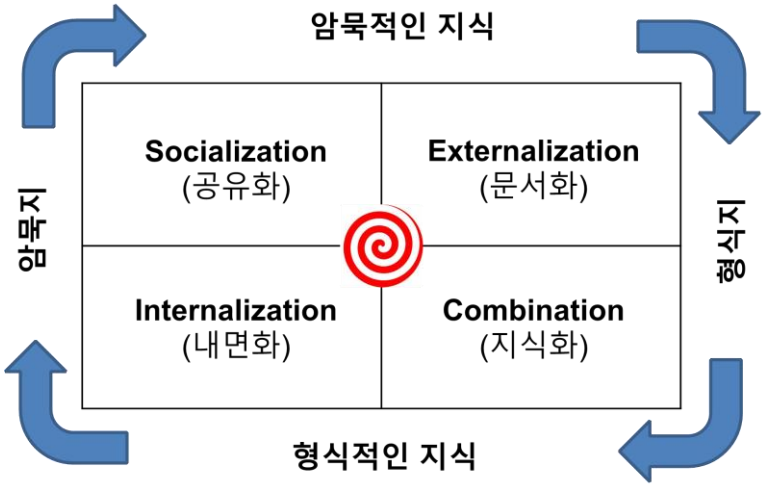
데이터 분석 개요

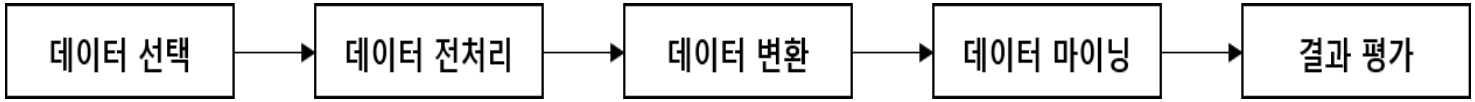
분석 모형 구축 절차 수립

- **분석 모형 구축 절차 수립**
  - 데이터 모델링을 위한 첫 단계 : 분석 모형 구축 절차 수립
  - 모델링을 위한 전체 과정을 몇 개의 단계로 나누고 각 단계 별 상세한 방법론, 도구나 기법, 산출물 등을 정의하는 작업
  - 개개인의 역량과 경험에 의존하지 않고 누가 수행하던 "일정수준의 질과 양"이 보장될 수 있는 체계(시스템)

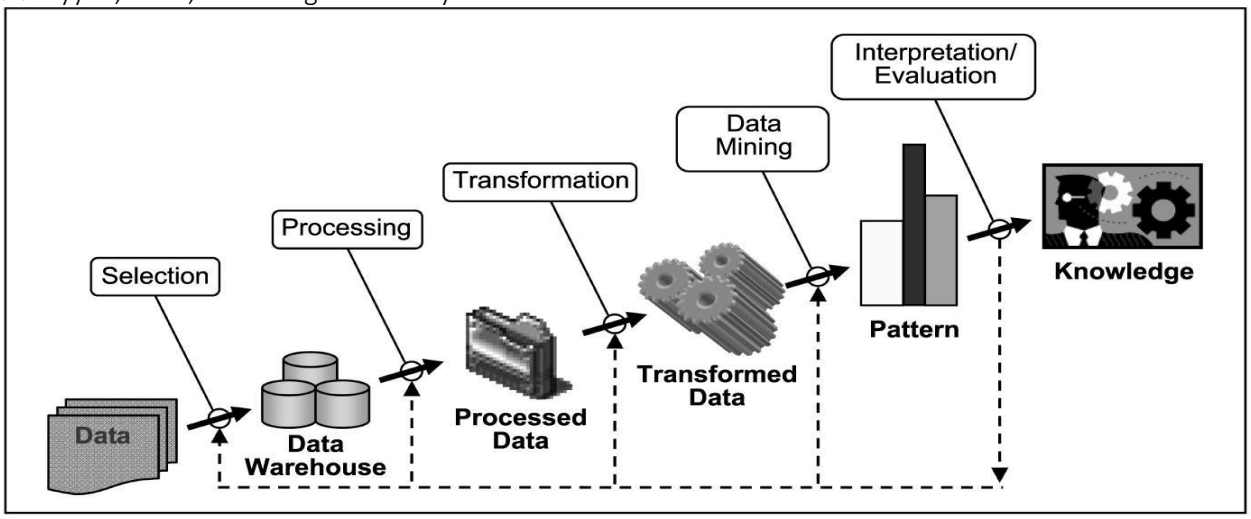


분석 모형 구축 절차 수립

- **KDD (Knowledge Discovery in Database)**
  - 1996년 Fayyad가 프로파일링 기술을 기반으로 개발한 방법론
  - 데이터로부터 통계적 패턴이나 지식을 찾기 위해 사용

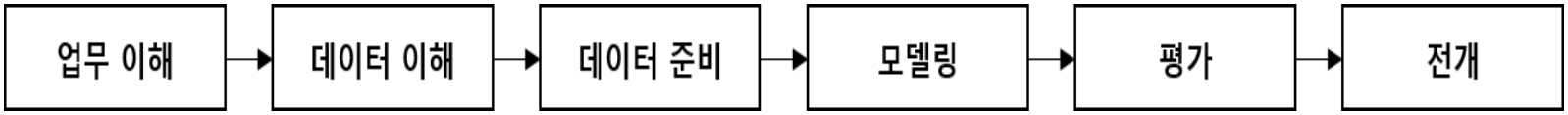


※ Fayyad, 1996, Knowledge Discovery in Database

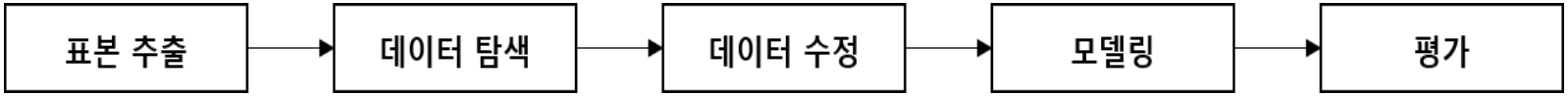


분석 모형 구축 절차 수립

- **CRISP-DM (Cross Industry Standard Process for Data Mining)**
  - 1996년 유럽 연합의 ESPRIT에 있던 프로젝트에서 시작
  - 계층적 프로세스 모델

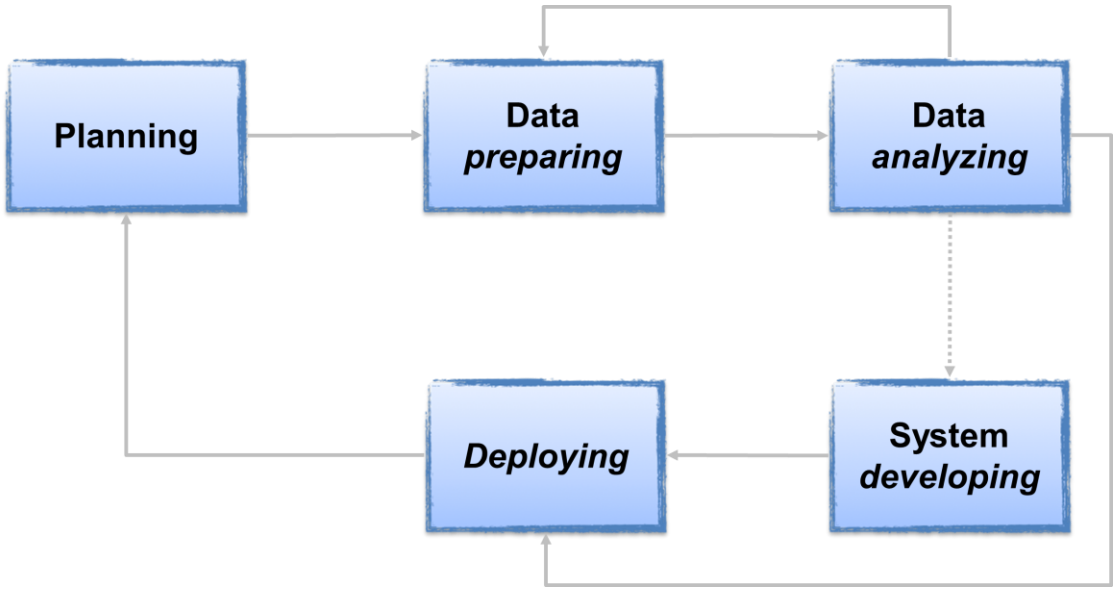


- **SEMMA (Sampling Explore Modify Model)**
  - SAS에서 개발한 데이터 마이닝 표준 가이드
  - 비즈니스에서 활용할 수 있는 패턴을 찾기 위해 사용



분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**

분석 기획 Plaaning	데이터 준비 Preparing	데이터 분석 Analysis	시스템 구현 Developing	평가 및 전개 Deploying
• 비즈니스 이해 및 범위 설정 • 프로젝트 정의 및 계획 수립 • 프로젝트 위험계획 수립	• 필요 데이터 정의 • 데이터 스토어 설계 • 데이터 수집 및 점검	• 분석용 데이터 준비 • 탐색적 분석 • 모델링 • 모형 평가 및 검증 • 모델적용 및 운영 방안 수립	• 설계 및 구현 • 시스템 테스트 및 운영	• 모델 발전계획 수립 • 프로젝트 평가 및 보고

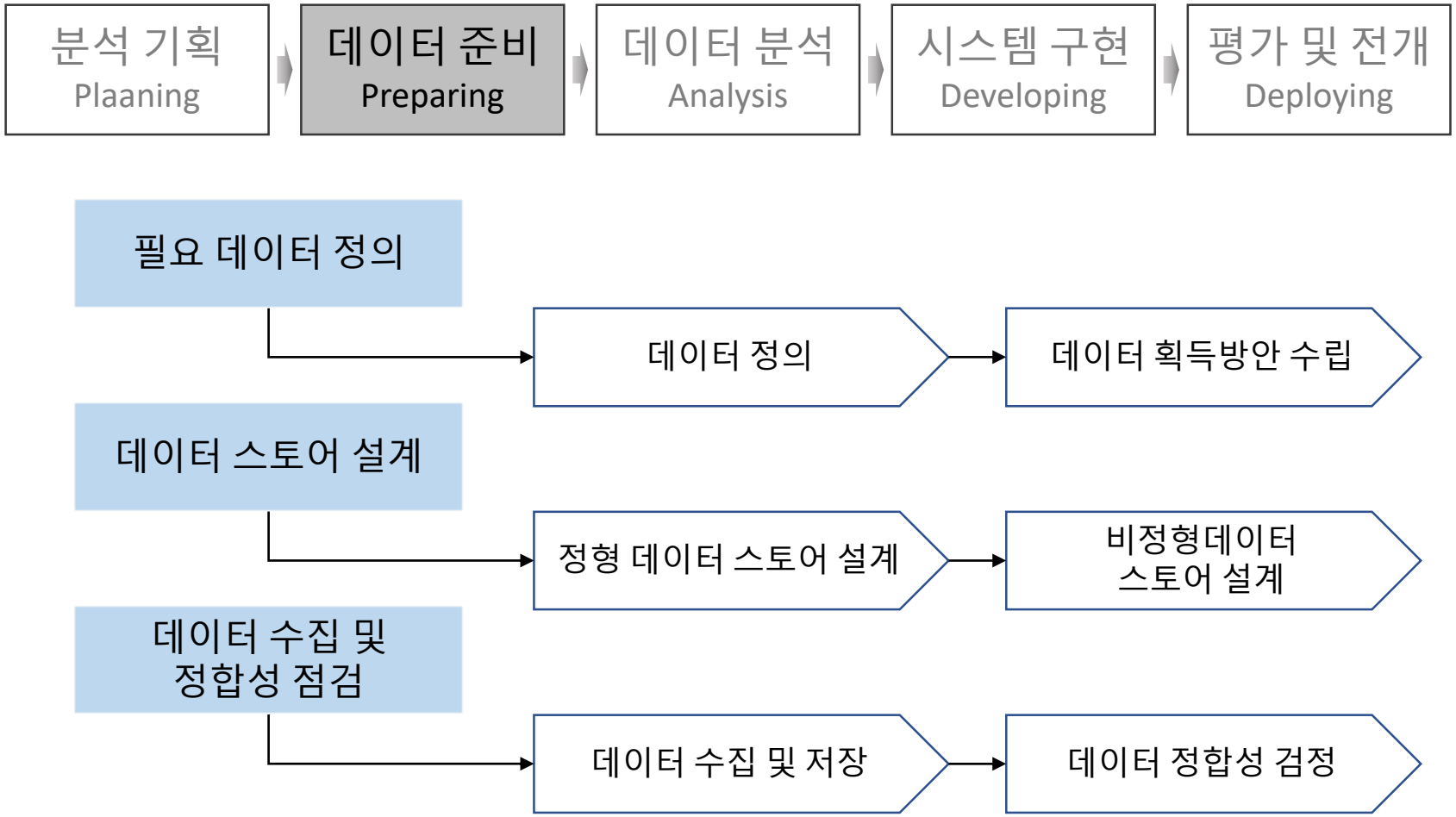
분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**

	KDD	CRISP-DM	SEMMA
분석 기획		업무 이해	
데이터 준비	데이터 선택 데이터 전처리 데이터 변환	데이터 이해 데이터 준비	표본 추출 데이터 탐색 데이터 수정
데이터 분석	데이터 마이닝 결과 평가	모델링 평가	모델링 평가
시스템 구현			
평가 및 전개		전개	

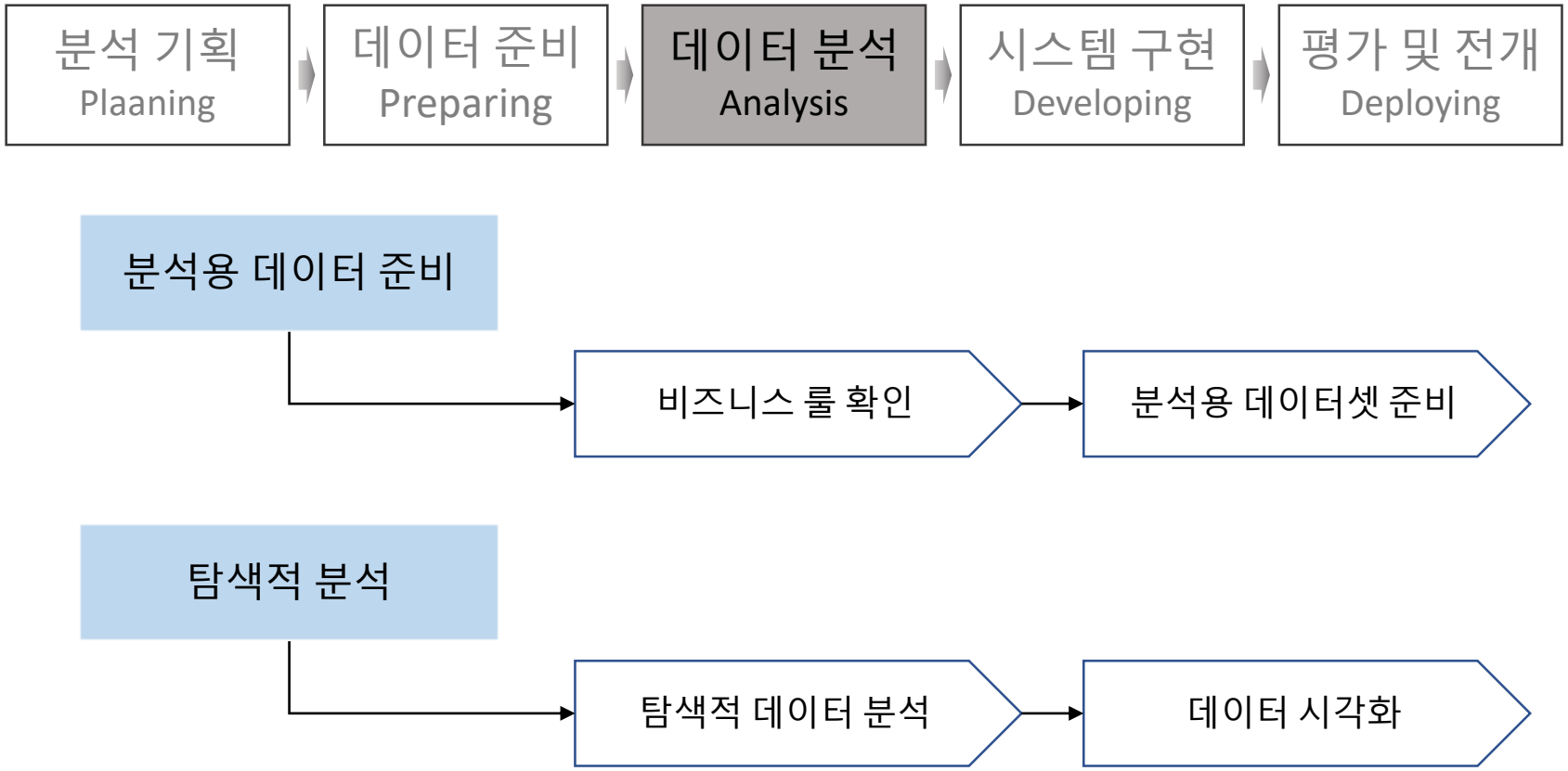
분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



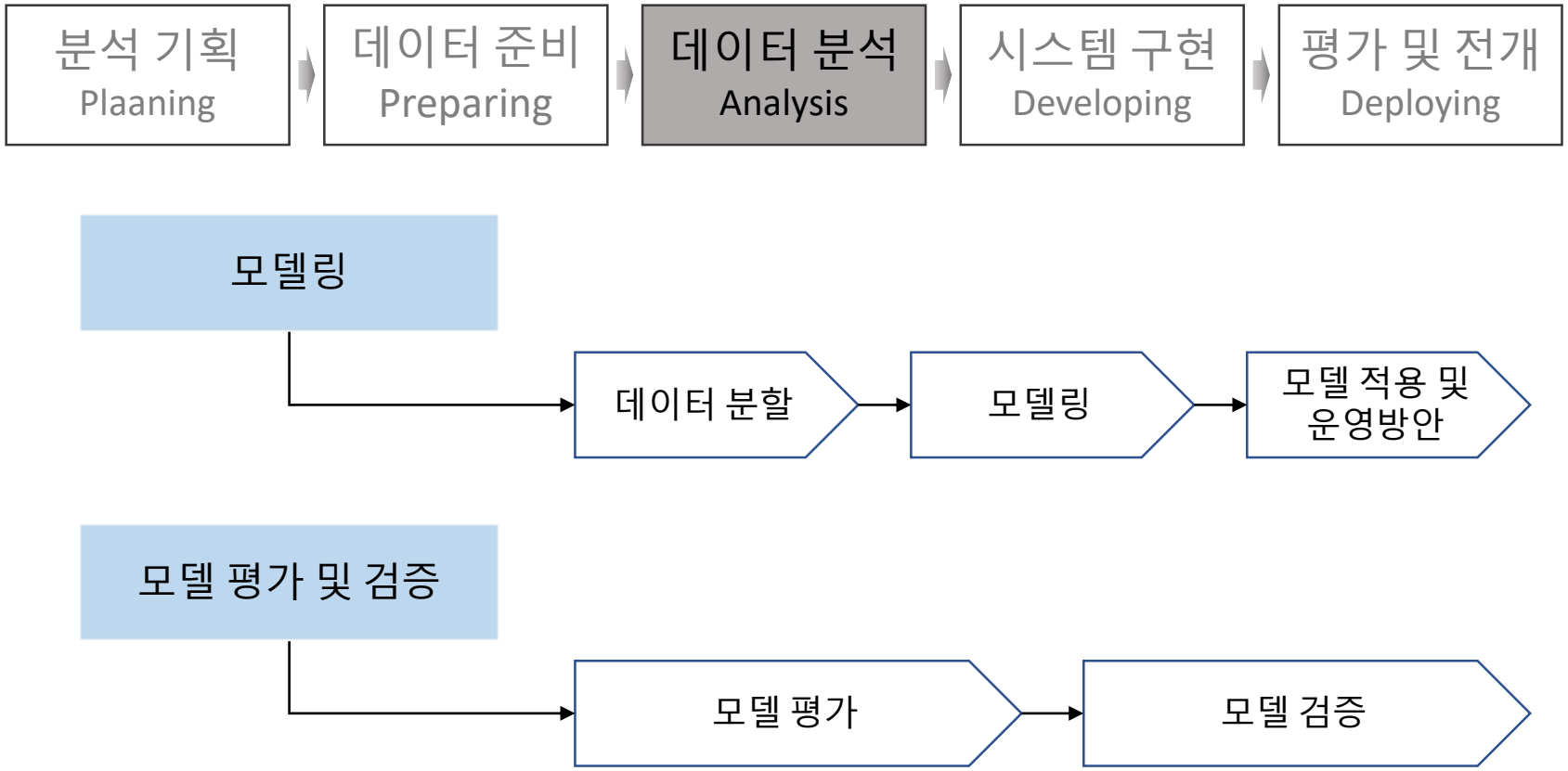
분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



분석 모형 정의하기

- **목적에 따른 분류**

기술 분석 (Descriptive analysis)

탐색적 분석 (Exploratory analysis)

추론 (Inferential analysis)

예측 (Predictive analysis)

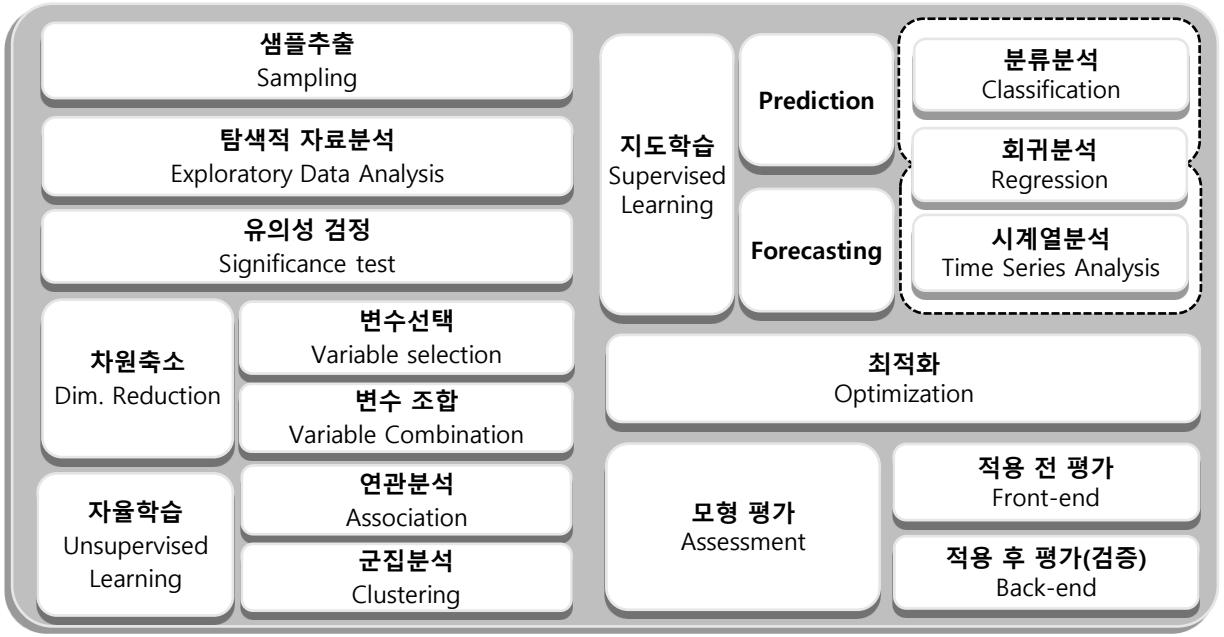
분류 (classification analysis)

분석 모형 정의하기

- **방법론에 따른 분류**
  - 고전적인 통계분석(classical methodology)
  - 컴퓨터의 알고리즘에 의해 분석을 하는 기계학습(machine learning) 방법
  - 목적과 Data의 유형에 따라 다음의 분석 방법들을 고려

통계분석/ 기계학습 방법론 인벤토리

(Inventory for Statistical Analysis / Machine Learning Methodology)



분석 모형 정의하기

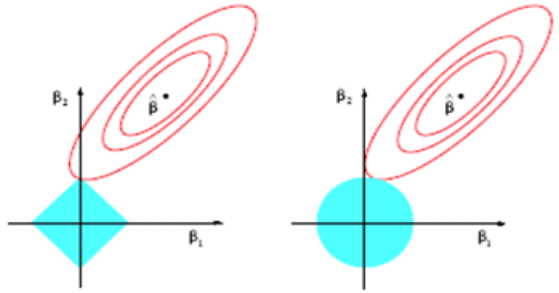
방법론에 따른 분류

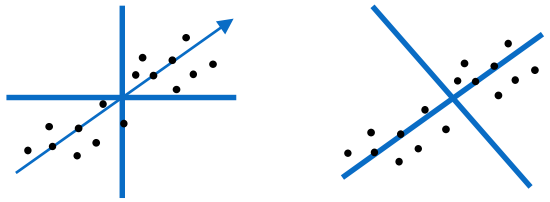
분석 목적	반응변수 유형	설명변수 유형	분석 알고리즘
분포가정		연속형, 이산형	• 적합도 검정 : 카이제곱 검정, 콜모고로프-스미노로프 검정, 우도비 검정
연관관계 분석 (반응변수 없음)		연속형	• 변수 2개 : (피어슨)상관분석, 스피어만 순위분석, 켄달의 타우 분석 • 변수 2개 이상 : 부분상관분석, 주성분분석, 요인분석, 정준상관분석
		이산형	• 변수 2개 : 피셔의 적확 검정, 카이제곱 검정, 맥네머 검정, 연관성 척도(파이 계수, 크래머 브이, 람다계수 등) • 변수 2개 이상 : 반응분석, 로그 선형분석
		연속형, 이산형	• 군집분석 : K평균군집분석, SOM, 계층적 군집분석, 모형기반 군집분석 등 • 연관분석 : Apriori 알고리즘 등
인과관계 분석 (반응변수 있음)	연속형	연속형 (이산형 포함)	• 회귀분석 : 선형 회귀분석, 비선형 회귀분석, 회귀나무, 주성분 회귀분석, 부분 최소제곱 회귀분석, 앙상블(Bagging, Boosting), Elastic net(LASSO, 릿지 회귀분석), Neural Network(Deep Learning 포함) • 시계열 분석 : 분해법, 평활법, ARIMA, GARCH, 누적 합 검정 등 • 생존분석 : Cox 회귀분석 등 • 이산형과 연속형 설명변수의 혼합 : 다변수 공분산 분석(MANCOVA)
		이산형	• 이산형 설명변수 1개, 1-2수준 : t-검정, 만-휘트니 검정 • 이산형 설명변수 1개, 다수준 : 분산분석, 크루스칼-왈리스 검정 • 다수의 이산형 설명변수 : 다변량 분산분석 (MANOVA),
	이산형	연속형 (이산형 포함)	• 분류분석 : 로지스틱 회귀분석, 선형 판별분석, 이차 판별분석, SVM, Neural Network(Deep Learning 포함), 분류 나무, 앙상블(Bagging, Boosting), Elastic net(LASSO, 릿지 회귀분석),
		이산형	• 로그 선형분석

분석 모형 정의하기

- 일반적으로 독립 변수의 수가 매우 많거나, 다중 공선성의 여지가 보일 경우 변수를 선택하거나 조합을 실시함

구분	상세구분
데이터 처리	
탐색적 자료분석	추정 및 검정
	시각화
차원축소	변수 선택
	변수 조합
자율학습	연관분석
	군집분석
지도학습	회귀분석
	시계열분석
	분류분석
	생존분석
최적화	
모형평가	
	전,후진/단계적선택법
	지표기준 선택법
	앙상블
	LASSO
	베이지안 선택법
	주성분분석
	요인분석
	정준상관분석
	독립성분분석
	부분 최소제곱

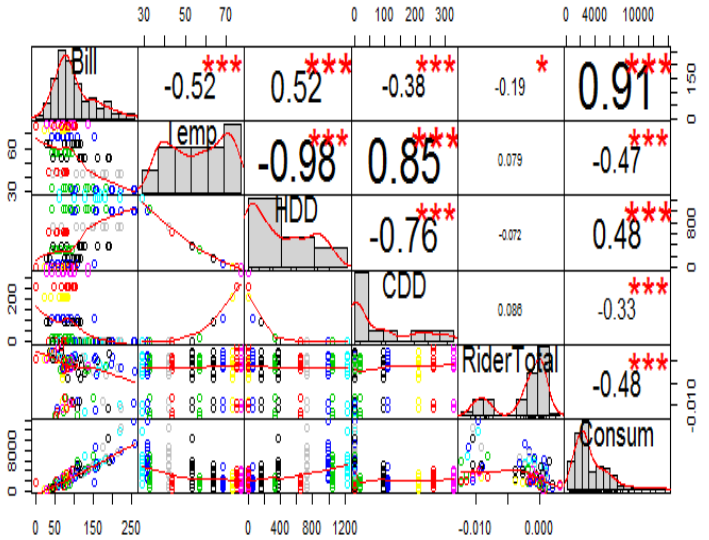




분석 모형 정의하기

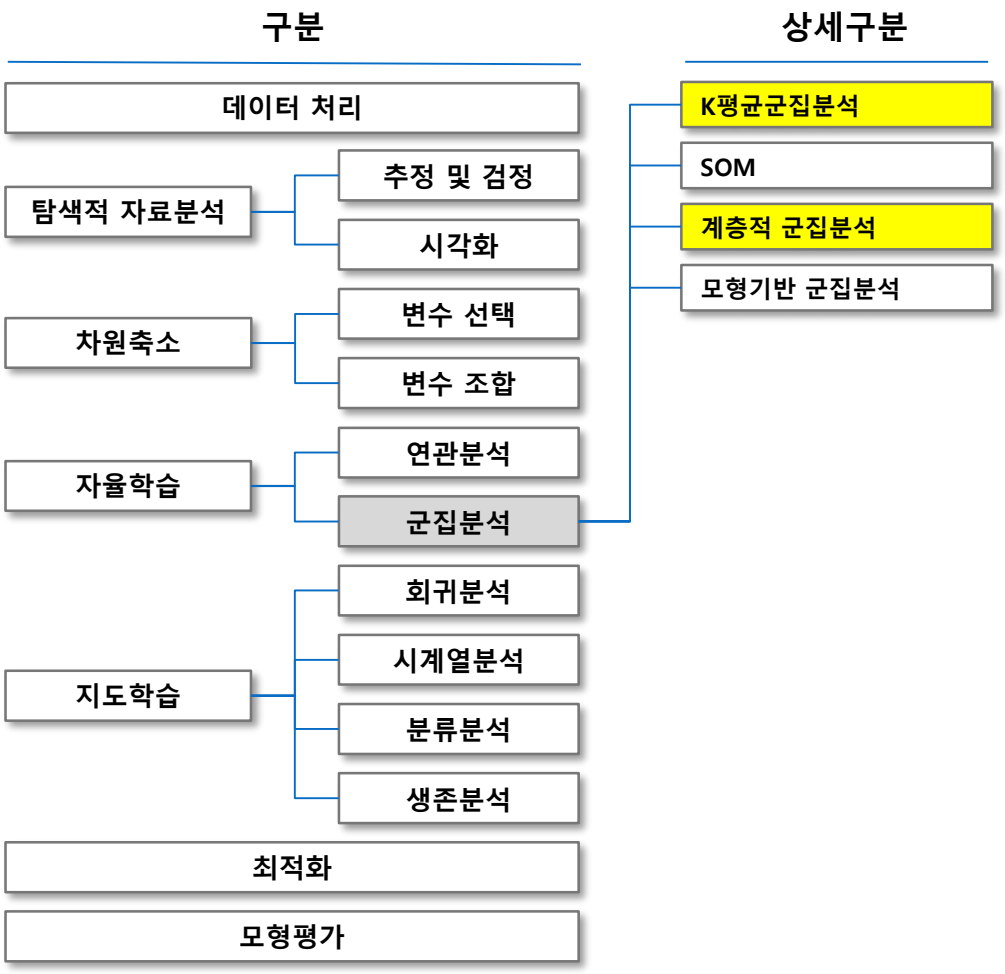
- 독립 변수들 간의 관계를 살펴보는 연관분석을 실시함

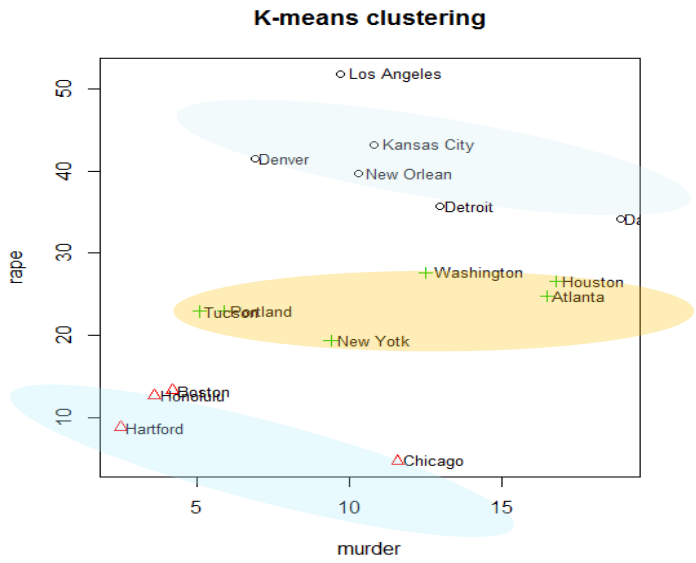
구분	상세구분
데이터 처리	(피어슨)상관분석
탐색적 자료분석	스피어만 순위분석
추정 및 검정	켄달의 타우 분석
시각화	부분상관분석
차원축소	주성분분석
변수 선택	요인분석
변수 조합	연관분석
자율학습	정준상관분석
군집분석	피셔의 정확 검정
지도학습	카이제곱 검정
회귀분석	맥네머 검정
시계열분석	연관성 척도
분류분석	반응분석
생존분석	로그 선형분석
최적화	
모형평가	



분석 모형 정의하기

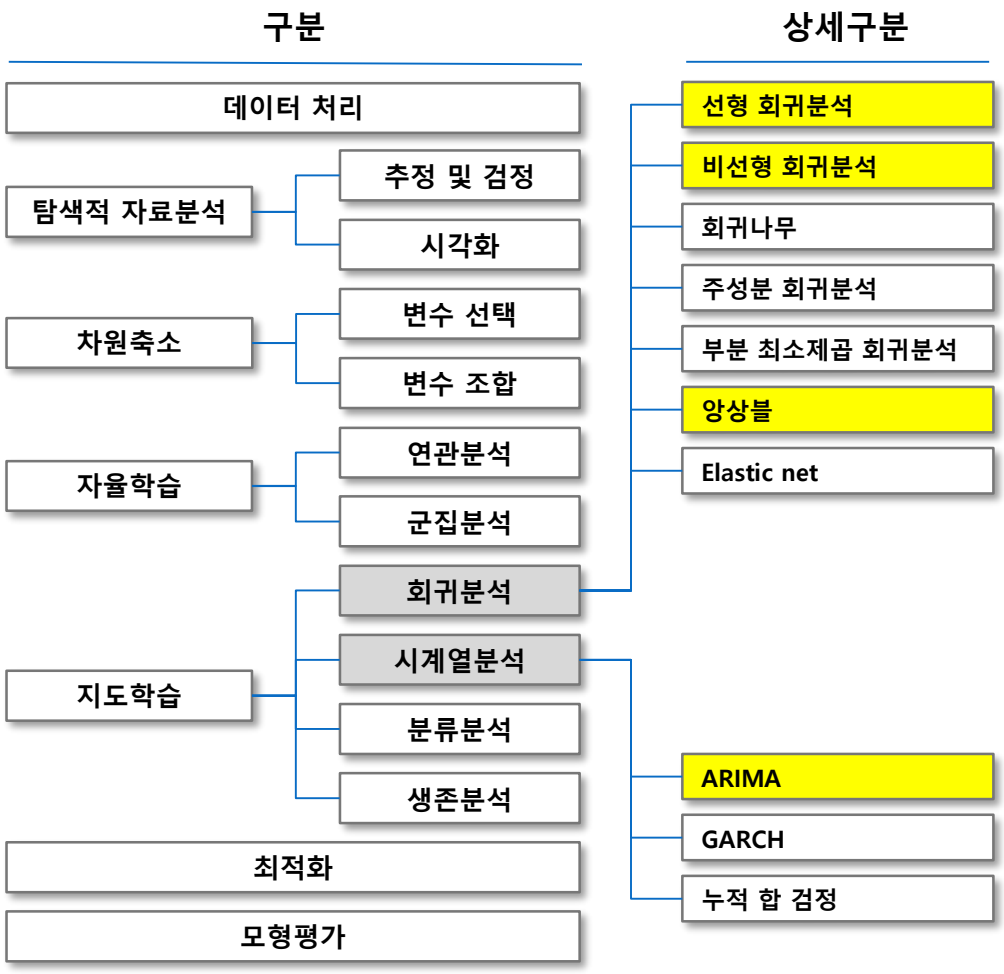
- 독립 변수들 간의 거리를 고려한 군집을 형성하는 군집분석을 실시함

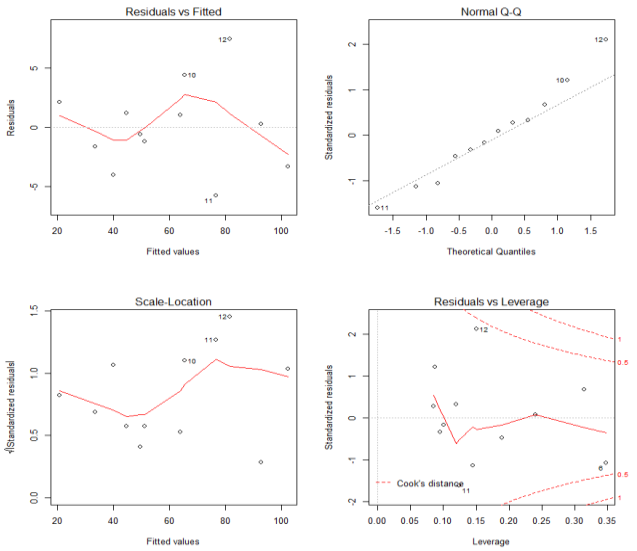




분석 모형 정의하기

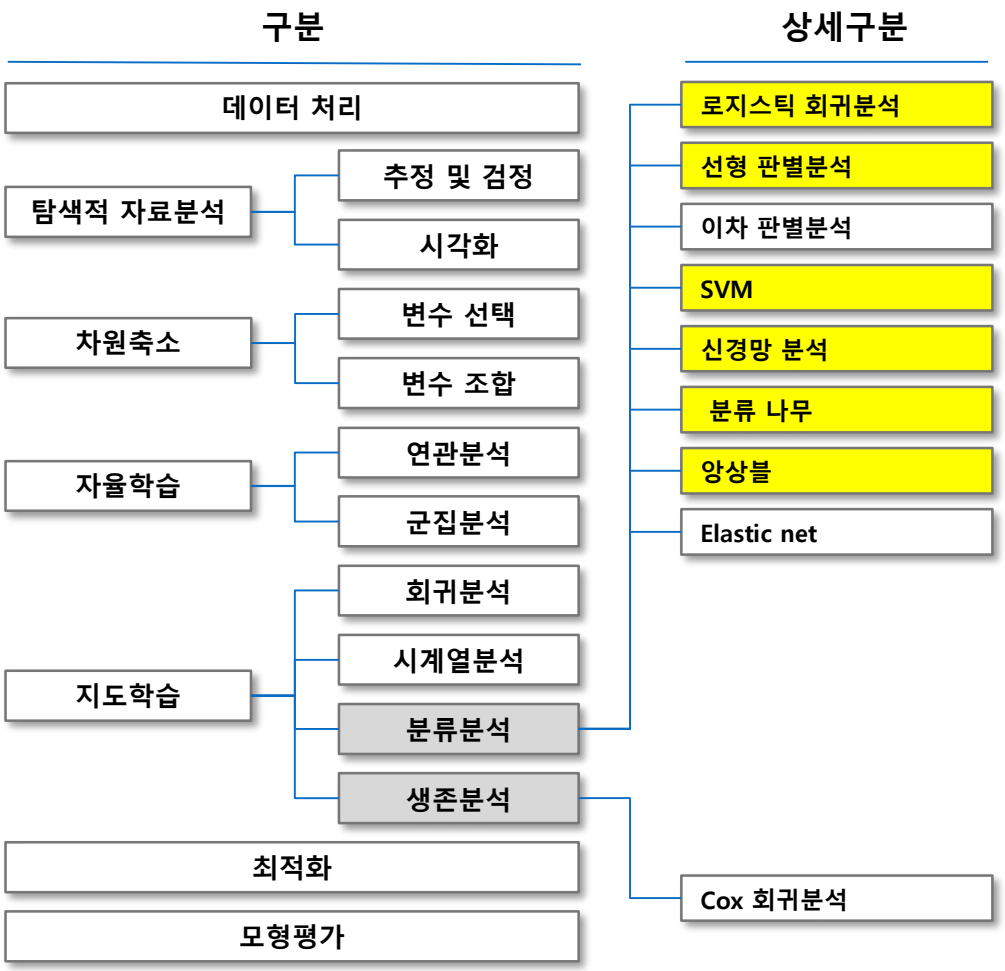
- 반응 변수의 값이 연속일 경우 회귀 분석을 실시하고, 특히 시간에 영향을 많이 받는 경우 시계열 분석을 실시함

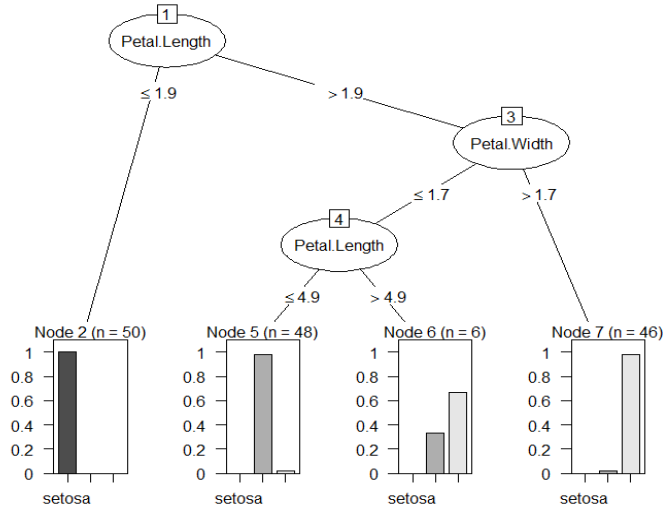


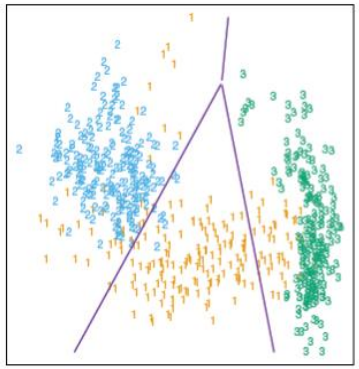


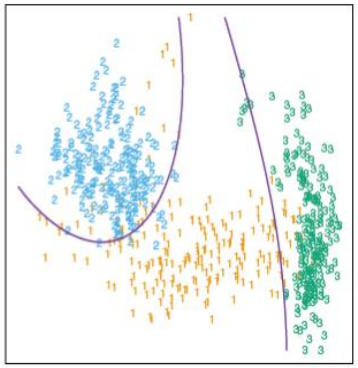
분석 모형 정의하기

- 반응 변수의 값이 이산일 경우 분류 분석을 실시하고, 특히 시간도 고려해야 할 경우 생존분석을 실시함

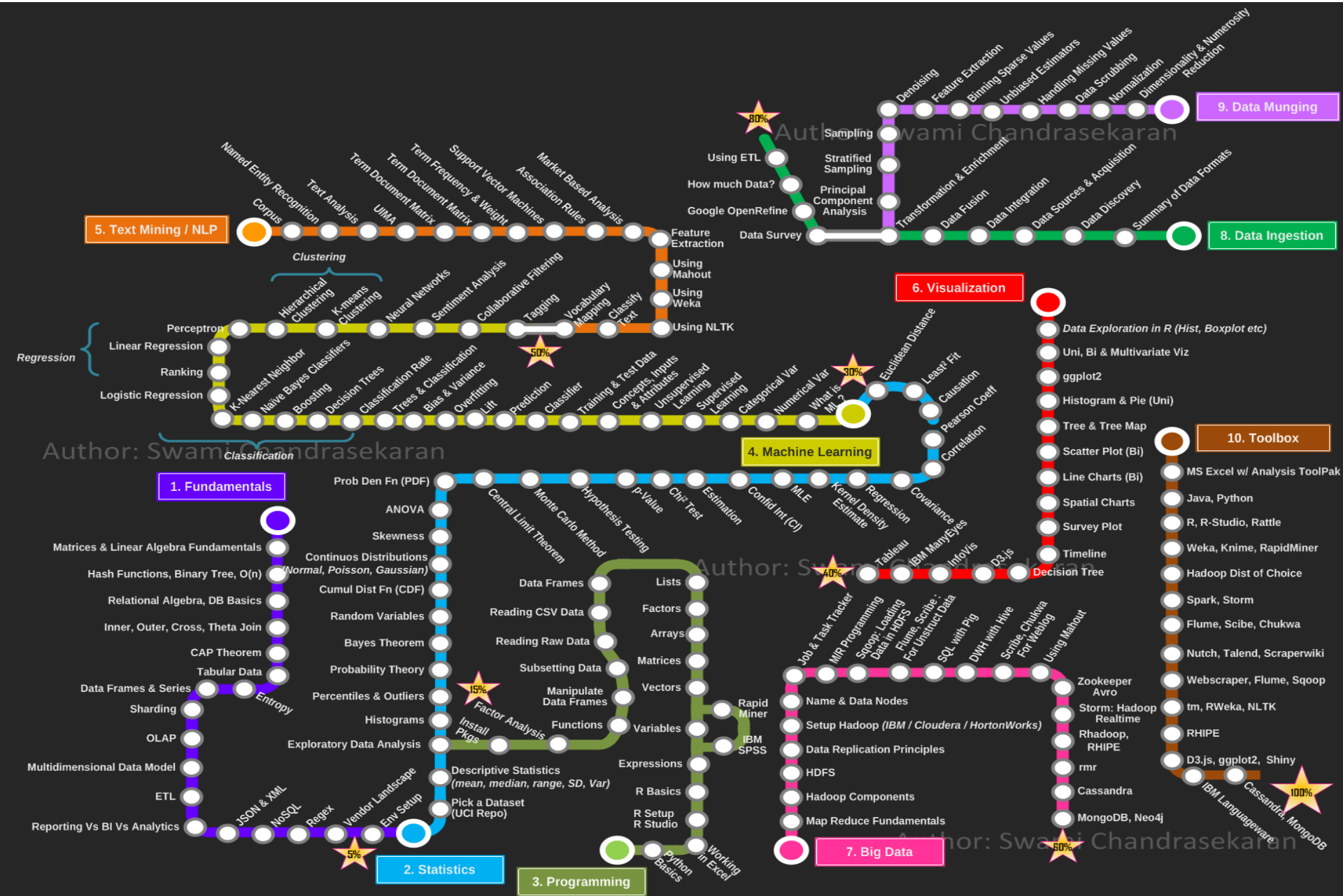






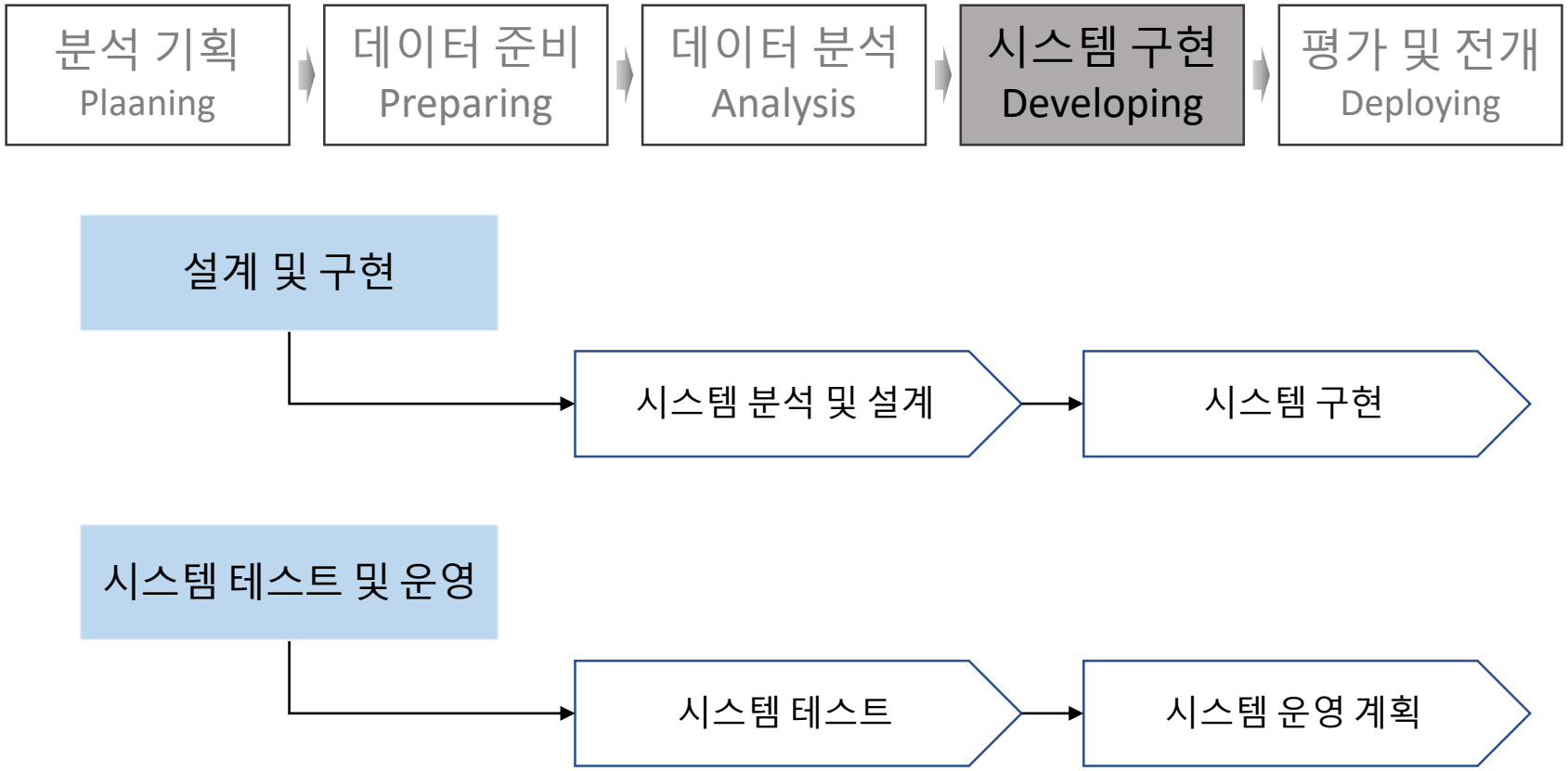


분석 모형 정의하기


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions
(Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
Descriptive Statistics
(mean, median, range, SD, Var)
Pick a Dataset
(UCI Repo)
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Named Entity Recognition
Corpus
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
5. Text Mining / NLP
What is ML ?
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Google OpenRefine
How much Data?
Using ETL
Sampling
Stratified Sampling
Principal Component Analysis
80%
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran

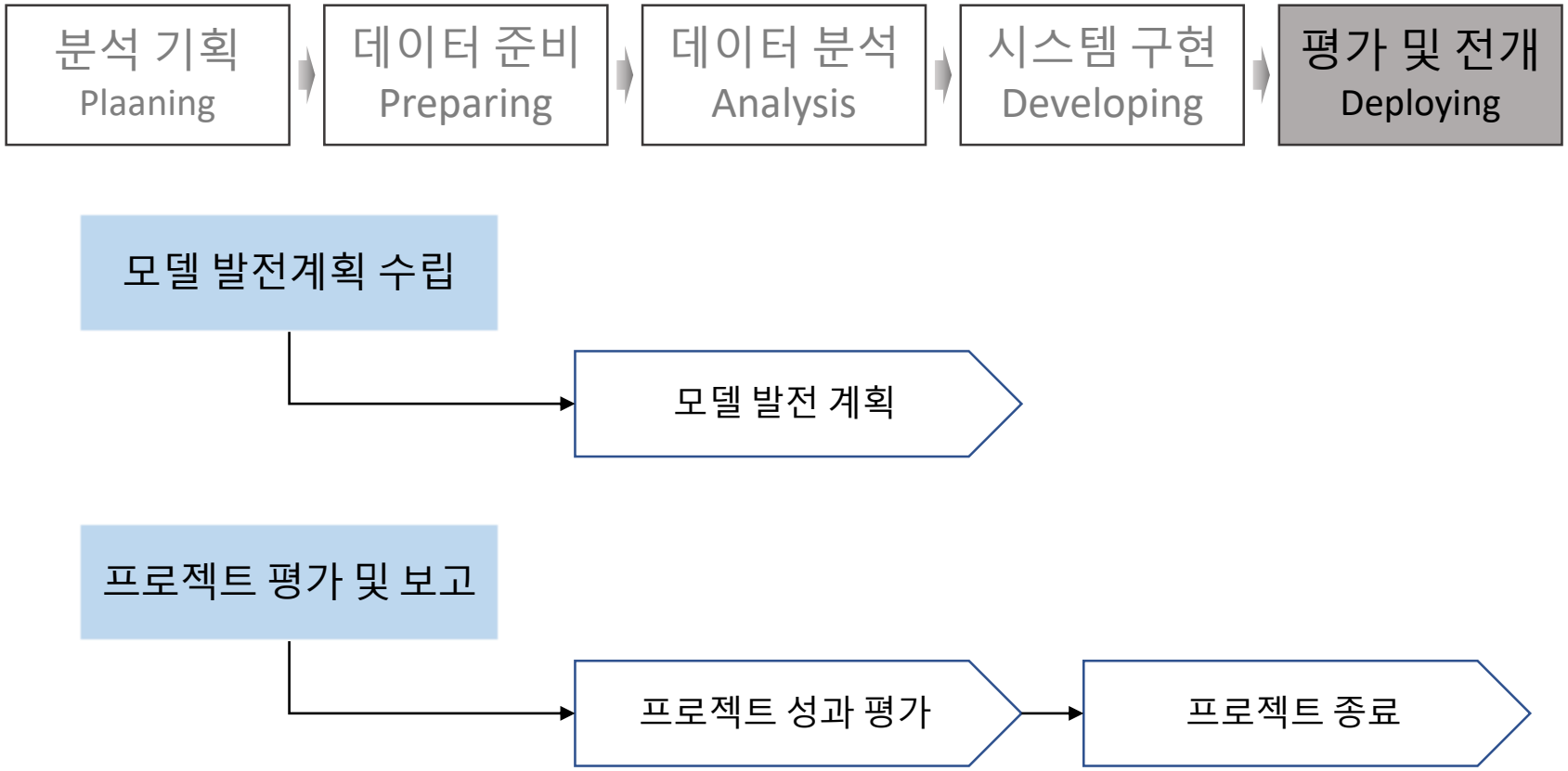
분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



분석 모형 구축 절차 수립

- **빅데이터 분석 방법론**



End of Document