



# 빅데이터 분석 절차

경남대 전하용

# 빅데이터의 3요소

## 빅데이터(Big Data)

- 데이터 자원 확보
- 데이터 품질 관리

자원

빅데이터 3대 요소

기술

인력

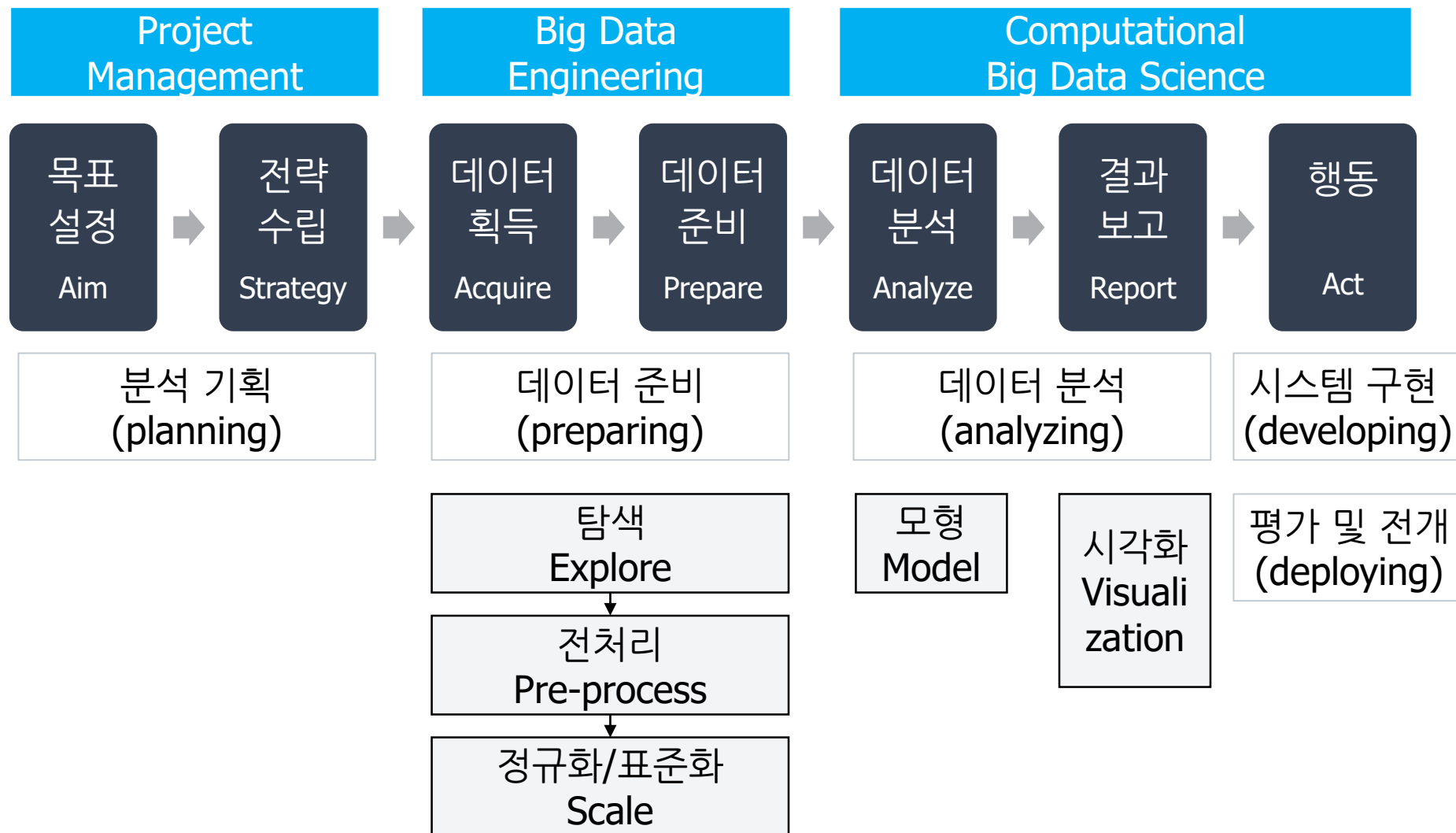
## 빅데이터 플랫폼 (Big Data Platform)

- 데이터 저장, 관리 기술(NoSQL, ETL)
- 대용량 데이터 처리(하둡, 맵리듀스)
- 빅데이터 분석
- 빅데이터 시각화

## 데이터 과학자(Data Scientist)

- 수학, 공학(IT기술과 엔지니어링) 능력
- 경제학, 통계학, 심리학 등 다문학적 이해
- 비판적 시각과 커뮤니케이션 능력
- 스토리텔링 등 시각화 능력

# 빅데이터 분석 절차(WorkFlow)



# 분석 기획: 목표 설정→전략 수립

- 분석 기획이란 실제 분석을 수행하기에 앞서 분석을 수행할 과제의 정의(목표 설정) 및 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획(전략 수립)하는 작업을 뜻한다.
- 분석 기획을 하기 위해서는 컴퓨터 사이언스, 수학&통계학 지식, 비즈니스 분석 능력에 대한 고른 역량과 시각이 필요하다.

비즈니스 이해 및 범위 설정

프로젝트 정의 및 계획 수립

프로젝트 위험 계획 수립

# 목표 설정

- 분석 주제의 4가지 유형

		분석 대상 (What)	
		known	Un-known
분석 방법 (How)	Known	최적화 (Optimization)	통찰력 (Insight)
	Un-known	해법 (Solution)	발견 (Discovery)

- ① 최적화(Optimization) : 분석 대상 및 분석 방법을 이해하고 현 문제를 최적화의 형태로 수행
- ② 솔루션(Solution) : 분석 과제는 수행되고, 분석 방법을 알지 못하는 경우 솔루션을 찾는 방식으로 분석 과제 수행
- ③ 통찰(Insight) : 분석 대상이 불분명하고, 분석 방법을 알고 있는 경우 인사이트 도출
- ④ 발견(Discovery) : 분석 대상, 방법을 모른다면 발견을 통하여 분석 대상 자체를 새롭게 도출

# 목표 설정: 올바른 질문 만들기

문제 정의  
(define the problem)



상황 평가  
(assess the situation)



목표 정의  
(define goals)

- 무엇을 분석할 것인가? 해결하려는 문제가 무엇인지 정의하는 단계로 비즈니스 가치에 연계할 수 있는 질문을 만든다.
- 결과가 어디에 왜 필요한가? 명확한 목표가 없다면, 문제를 해결해도 그것이 무엇인지 알지 못한다.

- 문제와 관련된 위험(Risks), 이익(Benefits), 만일의 사태(Contingencies), 규정(Regulations), 자원(Resources), 요구 사항(Requirement)을 주의 깊게 분석하는 단계

- 목적(objectives)과 목표를 정의하는 단계
- 명확한 목표 및 성공 기준(criteria)을 정의하면 프로젝트 평가에 도움이 됨

# 문제 정의 예시

- 신제품을 평가하기 위해서는 어떻게 판매 수치와 콜센터 기록을 결합할 수 있을까?
- 제조 과정에서 장비 고장을 감지하기 위해서는 장비의 여러 센서로부터 나오는 데이터를 어떻게 사용할 수 있을까?
- 효과적인 마케팅을 달성하기 위해 고객과 시장을 어떻게 더 잘 이해할 수 있을까?

# 상황 평가 예시

- 문제의 요구사항은 무엇인가?
- 가정과 제약 조건은 무엇인가?
- 어떤 빅데이터 자원(데이터, 기술, 인력)을 사용할 수 있나?
- 컴퓨터 시스템, 도구 등 인력과 자원은 사용할 수 있나?
- 주요 비용은 무엇인가?
- 잠재적 혜택은 무엇인가?
- 프로젝트 추진 시 어떤 위험이 있나?
- 잠재적 위험과 만일의 사태에 대비는 어떻게 할 것인가?



# 목표 정의

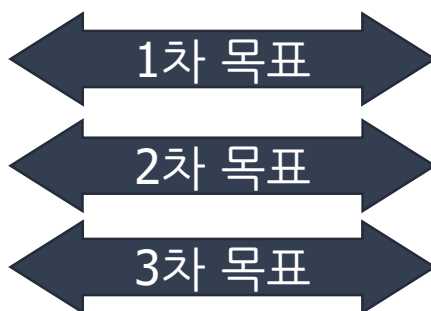
- 단기적 접근방식과 중장기적 마스터 플랜 접근 방식
- 분석의 가치를 증명하고 이해관계자들의 동의를 구하기 위해서는 분석을 통해 과거의 문제를 해결해서 분석의 가치를 높이고 공감대를 확산시키는 방식이 유용하다.

단기적 접근 방식

당면한 분석 주제 해결

(과제 단위)

Speed & Test
Quick-Win
Problem Solving



중장기적 마스터 플랜 접근 방식

지속적 분석문화 내제화

(마스터 플랜 단위)

Accuracy & Deploy
Long Term View
Problem Definition

# 목표 정의 예시

- 이 프로젝트가 끝날 때까지 무엇을 달성하기를 희망하나?
- 장기 목표(Long term)와 단기 목표(Short term)을 구분한다.

# 1단계 목표 설정

프로젝트 헌장(Project Charter)															
프로젝트 명 (Project Name)															
프로젝트 설명 (Project Description)															
프로젝트 매니저 (Project Manager, PM)		승인 날짜 (Date Approved)													
프로젝트 스폰서 (Project Sponsor)		서명 (Signature)													
비즈니스 케이스(Business Case)		목표(Goals) / 산출물(Deliverables)													
<div>팀 구성원(Team Member)</div> <table> <tr> <th>이름(Name)</th> <th>역할(Role)</th> </tr> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> <tr><td></td><td></td></tr> </table>				이름(Name)	역할(Role)										
이름(Name)	역할(Role)														
위험과 제약사항(Risk and Constraints)		주요 일정(Milestones)													

목표를 함께 정의하고 산출물과 일정 등의 계획에 합의

# 2단계 전략 수립

전략(strategy)이란 목표를 달성하기 위해 설계된 행동 계획 또는 정책을 의미한다. 빅데이터 전략을 수립할 때, 비즈니스 목표와 빅데이터 분석을 통합하는 것이 중요하다.



# 2단계 전략 수립

- 문제를 발굴한다.
- 목표에 대해 소통한다.
- 분석적 프로젝트를 위해 조직적 차원의 인수(buy-in)를 한다.
- 다양한 재능을 갖춘 팀을 구성하고 팀웍 마인드를 확립한다.
- 데이터 접근 및 통합의 장벽을 제거한다.
- 기술 발전에 대응하기 위해 반복한다.

# 2단계 전략 수립

- 프로젝트 팀으로 구성
  - 데이터 작업을 수행하고 새로운 아이디어를 테스트 하는 소규모 데이터 과학 전문가로 구성된 프로젝트 팀을 구성한다.
- 다양한 전문가로 구성
  - 데이터 과학자
  - 정보 기술자
  - 애플리케이션 개발자
  - 비즈니스 소유자

# 2단계 전략 수립

프로젝트(project)란 고유한 제품이나 서비스를 생산하기 위해 수행되는 일시적인 노력(endeavor)이다. 프로젝트는 한시성과 고유성이라는 특징을 갖는다.



# 2단계 전략 수립: 프로젝트 관리(PM)

## What is Project Management?

- The primary challenge is to achieve all of the project goals and objectives while honoring the preconceived constraints. Typical constraints are **scope**, **time**, and **cost**.
- The secondary—and more ambitious—challenge is to **optimize** the **allocation** and integrate the inputs necessary to meet pre-defined objectives.

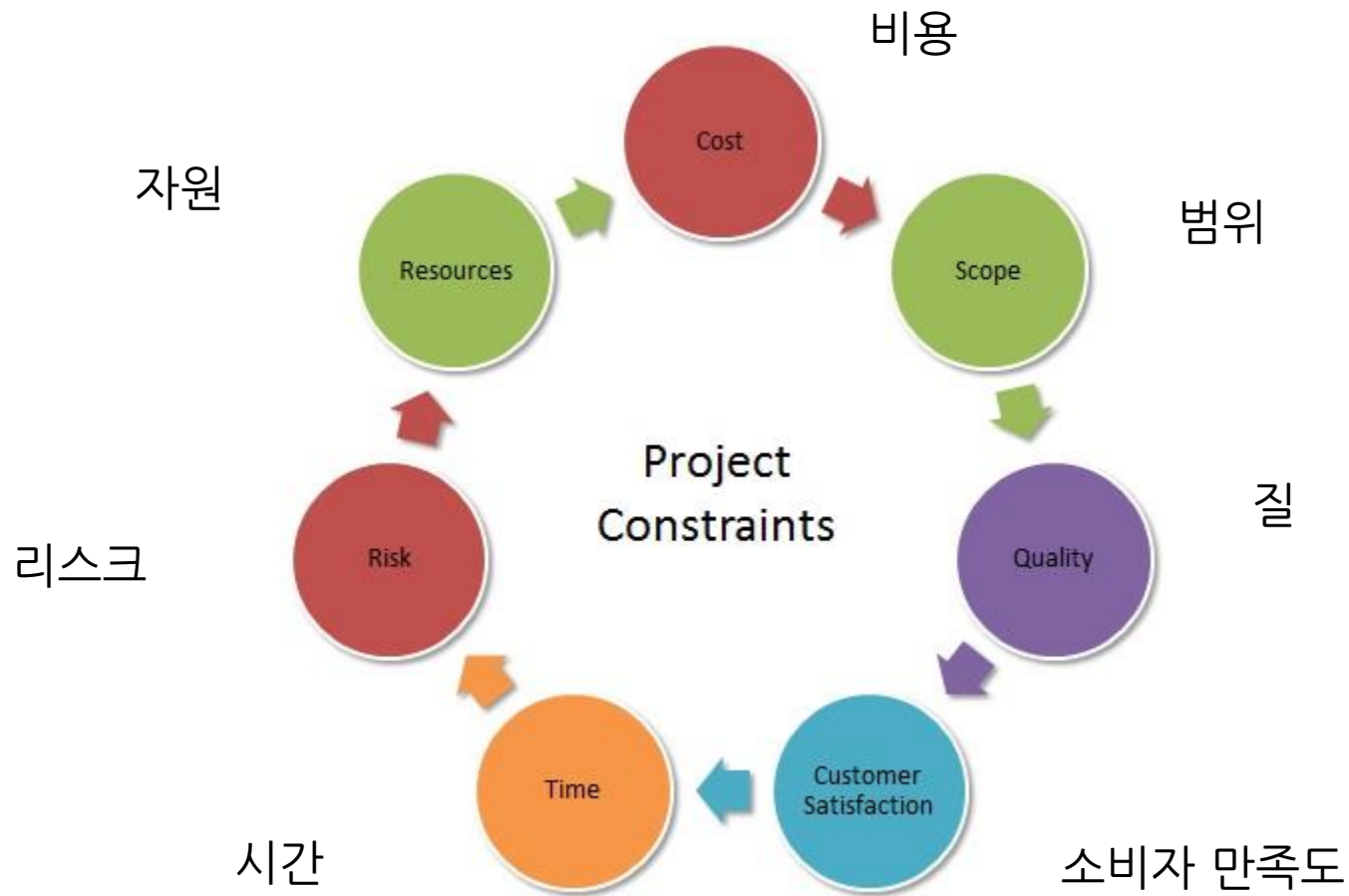


가장 중요한 과제는 제약 조건을 준수하면서 모든 프로젝트 목적과 목표를 달성하는 것이다. 일반적인 제약 조건은 **범위, 시간, 그리고 비용**이다.

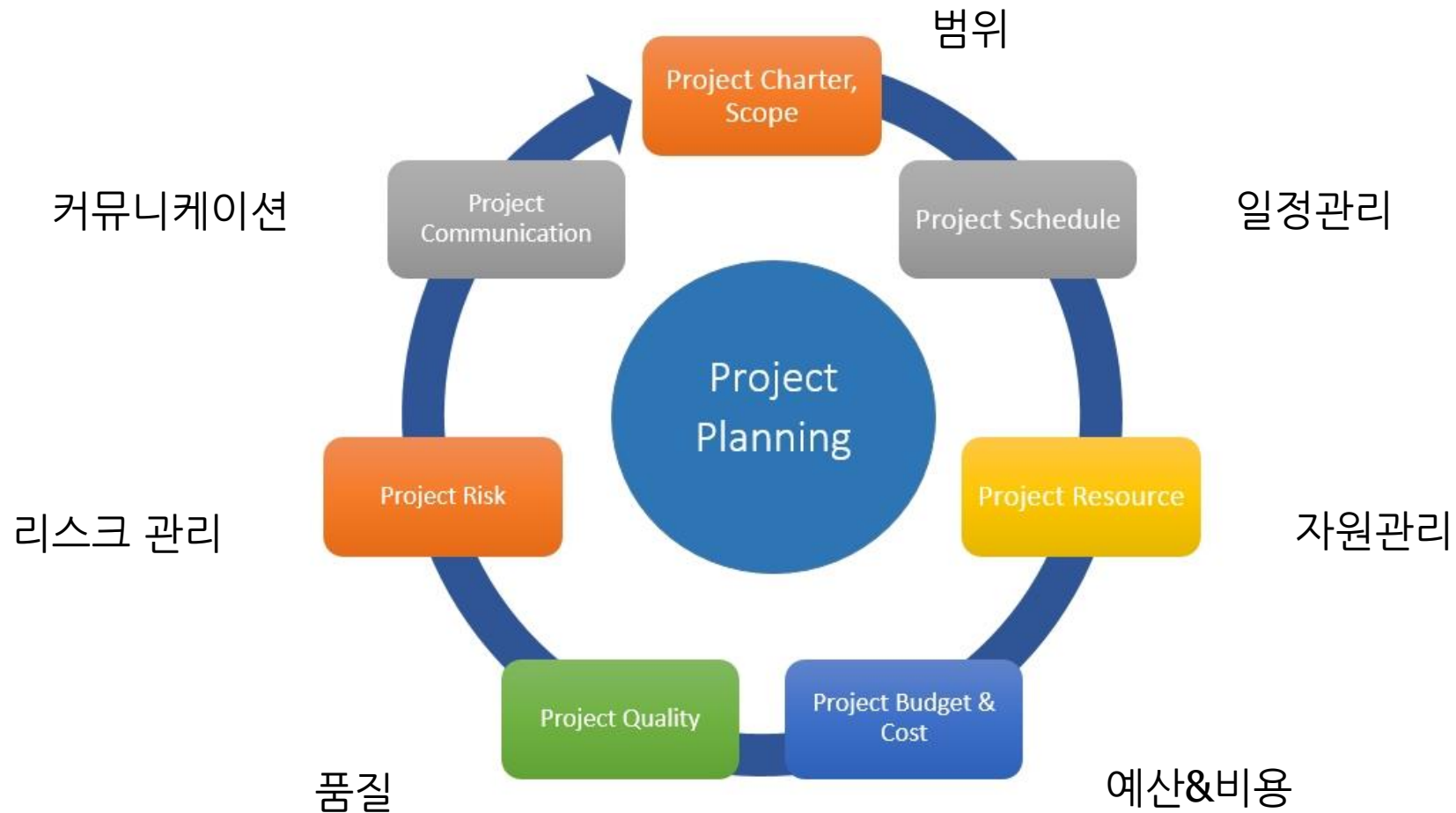
두 번째 과제는 **(자원)할당을 최적화**하고 사전 정의 된 목표를 달성하는 데 필요한 **(자원)투입을 통합**하는 것이다.



# 2단계 전략 수립: 프로젝트 제약조건



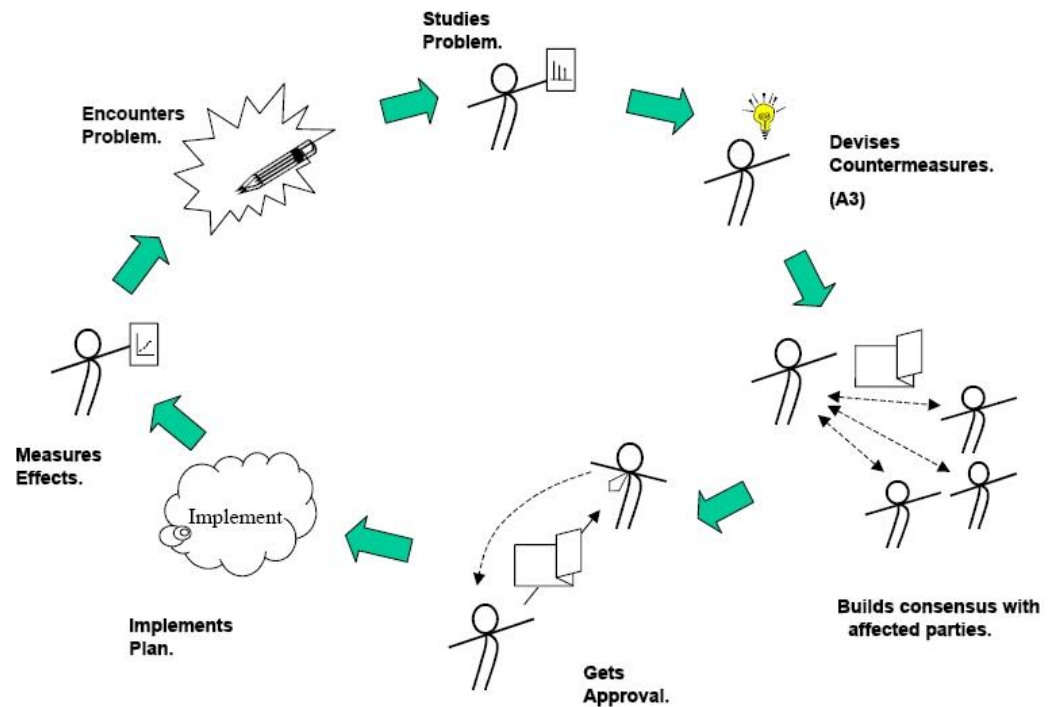
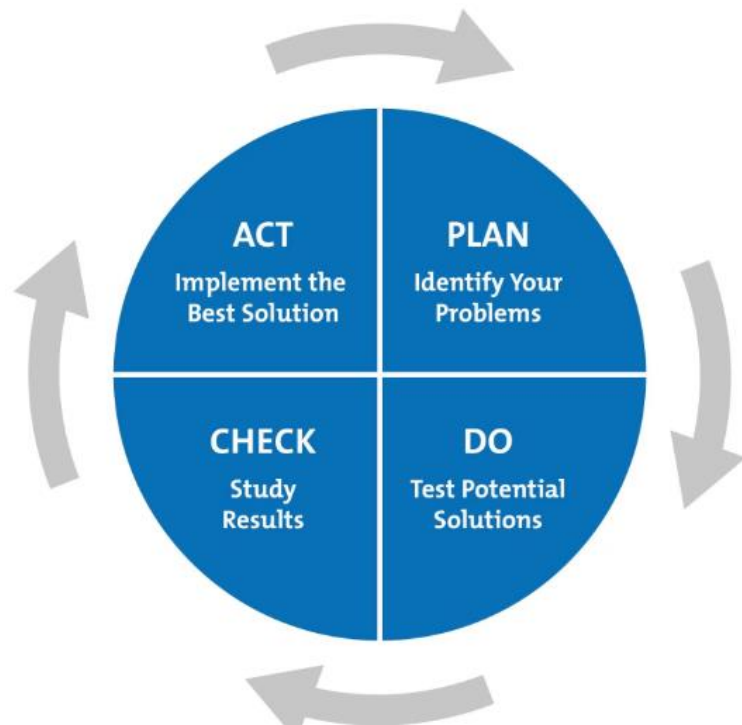
# 2단계 전략 수립: 프로젝트 관리 계획



# 2단계 전략 수립: 프로젝트 매니저와 팀원



# 2단계 전략 수립: Plan – Do – Check – Action



# 3단계 데이터 획득(acquire)

데이터 과학의 첫번째 단계는 데이터를 획득(수집)하는 것으로 원본 자료, 사용 가능한 데이터, 문제와 관련된 데이터, 즉 어떤 데이터가 필요한지 적절한 데이터를 식별하는 것을 의미한다.



- 데이터셋 정의(Identify data sets)
- 데이터 검색(Retrieve data)
- 데이터 수행 요청(Query data)

# 하향식 접근법

- 문제 해결 방법을 찾기 위해 필요한 데이터를 수집 및 분석하는 방식
- 문제 해결을 위해 근본 원인을 파악하고 분석 과제를 도출한 뒤 해결 방안을 도출
- 도출된 해결 방안에 대한 실현 가능성과 우선순위를 결정하기 위해 데이터를 수집, 가공, 분석하는 접근법
- 분석 과제를 도출하기 위해 ‘수요 기반 분석 과제 도출 방식’을 사용
- 데이터 분석은 문제 해결을 가능하게 하는 실행 동인 역할

# 상향식 접근법

- 현재 보유하고 있는 데이터를 분석하여 의미 있는 관계나 패턴을 찾아 지식을 발견하고 문제를 해결하는 방식
- 정형 데이터는 물론이고 다양한 원천의 비정형 데이터를 조합 하고 시각화를 통해 의미 있는 패턴을 파악한 뒤 이를 적용하여 문제를 해결하는 데이터 기반의 접근
- 분석 과제를 도출하기 위해 ‘데이터 주도 분석 과제 도출 방식’을 사용
- 데이터는 추진 동인 역할

# 프로토타이핑 접근법

- 빅데이터 환경의 불확실성을 고려한 방식
- 소비자의 요구 사항이나 데이터를 규정하기가 어렵고 데이터 원천도 명확히 파악하기 어려운 경우 사용
- 일단 프로토타입을 만들어 분석을 시도한 뒤 결과를 확인하고 개선하고 이를 반복



# 데이터 준비(preparing)

- 데이터 준비 단계는 초기의 데이터로부터 최종 데이터셋을 구성하기 위한 모든 활동

필요 데이터 정의

데이터 스토어 설계

데이터 수집 및 정합성 점검

# 3단계 데이터 획득(acquire)

- 필요한 데이터가 어디에 있을까?
  - 적합한 데이터 식별
  - 사용 가능한 모든 데이터 획득
- 데이터는 여러 곳에서 다양한 방법으로 가져온다.
  - 전통적인 데이터베이스: SQL & query browsers(MySQL, Oracle SQL, ...)
  - 텍스트 파일: 스크립팅 언어(python, R, JavaScript, php, ...)
  - 원격 데이터 & 웹서비스 : HTML, XML, RSS, JSON, ...
  - API & 웹서비스: NoSQL storage

# 3단계 데이터 획득(acquire)

- 화재 데이터 획득 예
  - 과거 날씨 : SQL
  - 현재 날씨: WebSocket
  - 실시간 화재 현황: Twitter

# 4단계 데이터 준비(prepare)

## 4-1 탐색 (Explore)

- 데이터의 본질을 이해하는 단계
- 예비 분석(산포도, 히스토 그램)

## 4-2 전처리 (Pre-process)

- 정제(Clean)
- 통합(Integrate)
- 패키징(package)

# 4-1 단계 데이터 탐색(explore)

- 데이터 탐색은 데이터의 특성보다 잘 이해하기 위해 예비 조사를 하는 단계이다.
- 데이터와 변수 간의 관계나 상호작용, 데이터의 분포, 편차, 패턴 존재 여부를 확인하는 탐색적 데이터 분석(Exploratory Data Analysis, EDA)이라고도 한다.

## 탐색적 데이터 분석(Exploratory Data Analysis, EDA)



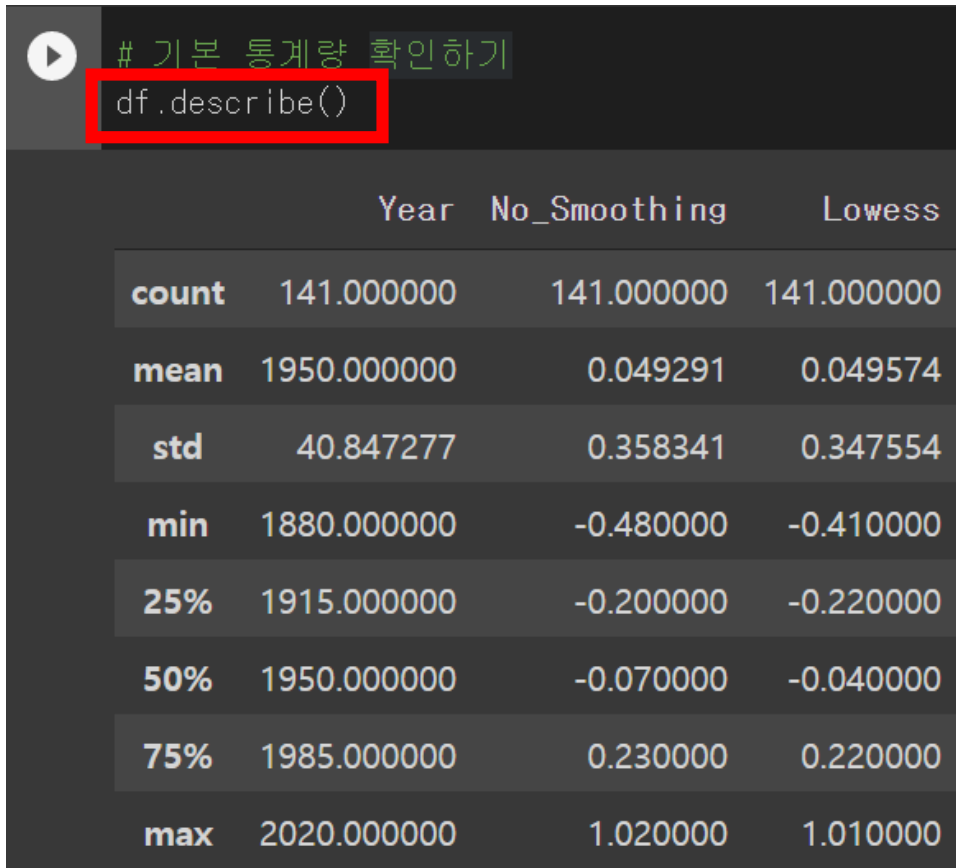
## 확증적 데이터 분석(Confirmatory Data Analysis, CDA)



# 4-1 단계 데이터 탐색(explore)

- 이 단계에서는 기초 통계량(statistic), 상관 관계(correlation), 데이터 분포(histogram, scatter plot), 데이터 추세 및 이상치(boxplot)와 같은 항목을 조사한다.
  - 요약 통계(평균, 중간값, 범위 및 표준편차 등)로 데이터의 성격을 확인하고 데이터에 문제가 없는지 여부를 판단할 수 있다.
  - 히트맵(heatmap)으로 상관 관계를 분석하여 변수 간 종속성을 탐색할 수 있다.
  - 추세 그래프(trend graph)는 일관된 방향이 있는지 여부를 탐색할 수 있다.
  - 산포도(scatter plot)나 히스토그램(histogram)으로 데이터 분포의 왜곡이나 비정상적인 분포를 탐색할 수 있다.
  - 박스 플롯(box plot)으로 이상치(outlier)를 표시하여 데이터의 오류를 찾거나 희귀한 사건을 탐색할 수 있다.
- 이 단계가 없으면 데이터를 효과적으로 사용할 수 없다.

# 4-1 단계 데이터 탐색(explore)



```
# 기본 통계량 확인하기
df.describe()
```

	Year	No_Smoothing	Lowess
count	141.000000	141.000000	141.000000
mean	1950.000000	0.049291	0.049574
std	40.847277	0.358341	0.347554
min	1880.000000	-0.480000	-0.410000
25%	1915.000000	-0.200000	-0.220000
50%	1950.000000	-0.070000	-0.040000
75%	1985.000000	0.230000	0.220000
max	2020.000000	1.020000	1.010000

## 기본 통계량(요약 통계)

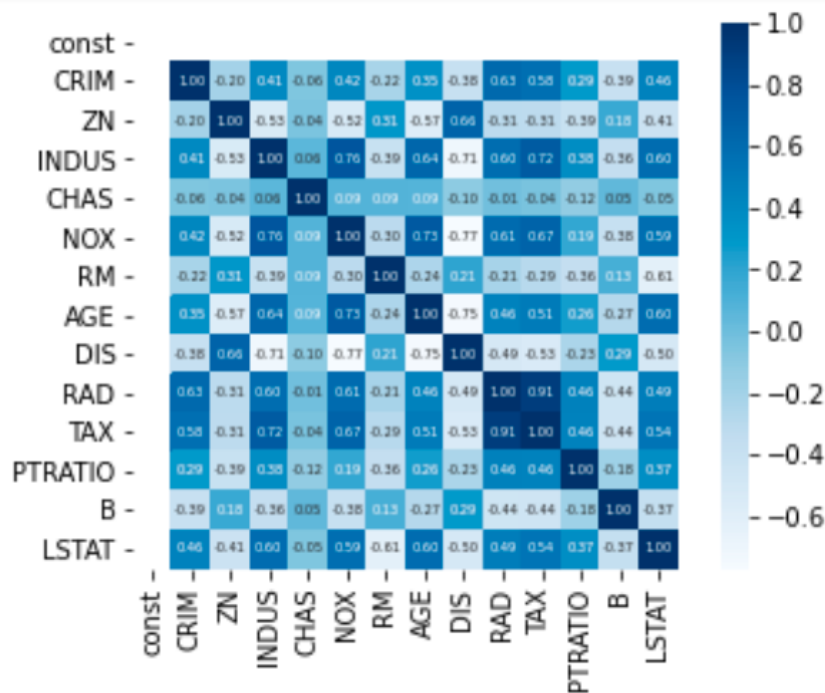
count	표본수
mean	평균
std	표준편차
min	최솟값
25%	제1사분위수
50%	중앙값
75%	제3사분위수
max	최댓값

# 4-1 단계 데이터 탐색(explore)

# 독립변수 간 상관관계 계수 확인

```
corr = df_X.corr(method = 'pearson')
```

```
df_heatmap = sns.heatmap(corr, cbar = True, annot = True, annot_kws={'size' : 5}, fmt = '.2f', square = True, cmap = 'Blues')
```



상관관계 계수가 0.7 이상

TAX, RAD = 0.91

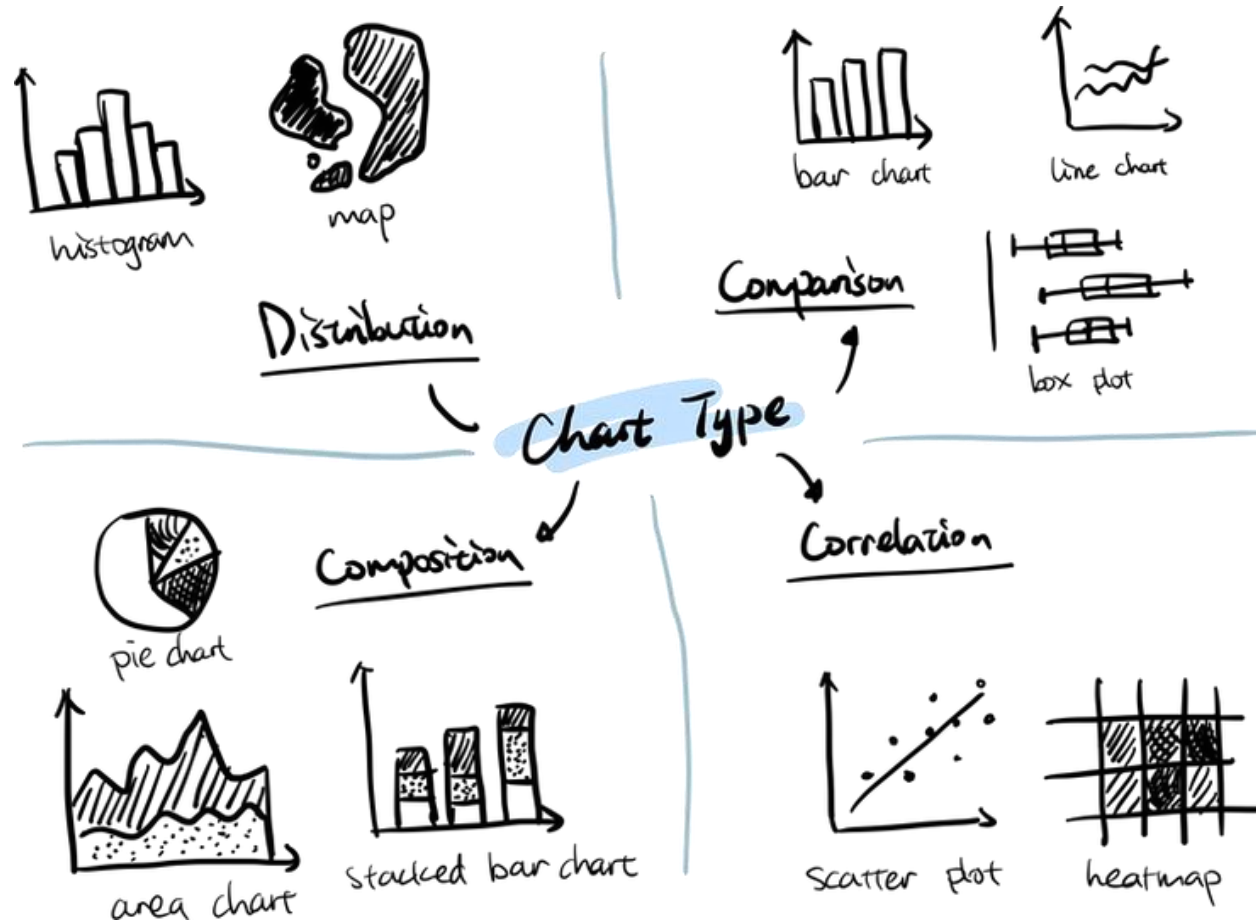
TAX, INDUS = 0.72

NOX, INDUS = 0.76

NOX, AGE = 0.73



# 4-1 단계 데이터 탐색(explore)



# 4-1 단계 데이터 탐색(explore)

## 맷플롯립(Matplotlib)으로 데이터 시각화

Matplotlib.org 웹사이트 갤러리에는 Matplotlib로 그릴 수 있는 다양한 샘플 차트(Line Plot, Bar Chart, Pie Chart, Histogram, Box Plot, Scatter Plot 등)의 이미지와 소스 코드를 제공하고 있다.



- 맷플롯립(Matplotlib)은 파이썬의 데이터 시각화 패키지(Data Visualization)로 데이터를 차트(chart)로 그려준다(plot).
- 차트는 그래프(graph), 도표(diagram), 지도(map), 테이블 형식을 포함한다.
- 레이어의 형태로 겹치는 방식으로 그린다

막대 차트

Bar Chart



크기를 기준으로 요소를 범주화하는 데 사용합니다. 차원의 특성에 따라 순서를 정하거나 정하지 않을 수 있습니다.

라인 그래프

Line Graph



시간에 따른 변화를 표현하는 데 가장 적합합니다. 여러 개의 라인을 사용하여 데이터 집합을 비교할 수 있습니다.

이중 축 차트



축 간 또는 차원 간의 데이터 관계를 알아보기 위해 기호, 막대 또는 라인을 결합한 차트입니다.

분산형 차트

Scatter Chart



독립 축과 두 측정값 간의 상관관계를 알아보는 데 사용합니다. 흔히 추세선과 결합하여 사용합니다.

Gantt 차트

Gantt Chart



시간 세그먼트를 표시하는 데 사용하는 틈새가 있는 막대 차트입니다. 시간 사용량을 나타내거나 기간을 시각화하는 데 유용합니다.

파이 차트

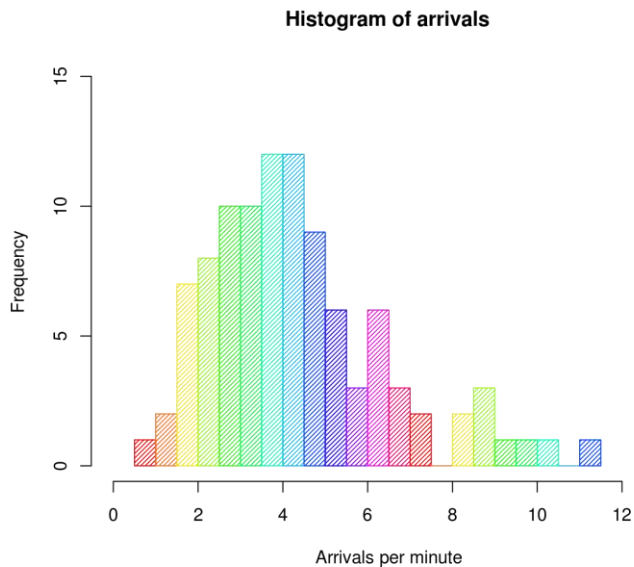
Pie Chart



몇 가지 차원을 서로 그리고 전체와 비교할 때 사용하는, 많이 사용되지만 제한적인 비주얼리제이션입니다.

# 4-1 단계 데이터 탐색(explore)

## 히스토그램(Histogram)

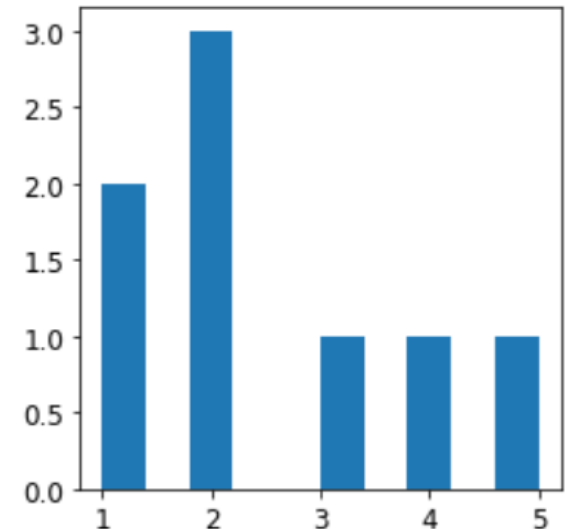


- 히스토그램(Histogram)은 도수 분포를 그림으로 나타낸 차트다.
- 가로축은 계급, 세로축은 도수를 뜻한다.
- 계급은 보통 변수의 구간이고, 서로 겹치지 않는다.
- 막대 차트처럼 보이지만 연속형 측정값의 값을 구간차원으로 그룹화한 것이다.
- $N$ 을 모든 관측값의 수라 하고,  $n$ 을 구간 개수라 하면, 히스토그램  $h_k$ 는  $N = \sum_{k=1}^n h_k$  조건을 만족한다.

# 4-1 단계 데이터 탐색(explore)

## 히스토그램(Histogram)

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
value = [1, 1, 2, 2, 2, 3, 4, 5]
plt.hist(value)
plt.show()
```



# 4-1 단계 데이터 탐색(explore)

- 이상치(outlier)의 유형
  - 실수로 잘못 입력한 경우
  - 분석 목적에 부합하지 않아 제거해야 하는 경우
  - 의도하지 않은 이상 현상이지만, 분석에 포함해야 하는 경우
  - 의도된 이상치인 경우(fraud, 불량)

# 4-1 단계 데이터 탐색(explore)

- 이상치의 인식 방법

- ① **EDS(Extreme Studentized Deviation):**

- 평균에서 3표준편차 떨어진 값(3시그마)

- ② **기하평균과 표준편차 활용법**

- $(\text{기하평균} - 2.5 \times \text{표준편차}) < data < (\text{기하평균} + 2.5 \times \text{표준편차})$

- ③ **사분위수 활용법**

- 사분위수의 Q1, Q3로부터  $1.5 \times IQR(Q3 - Q1)$  이상 떨어져 있는 데이터

# 4-1 단계 데이터 탐색(explore)

- 이상치의 처리 방법

- ① 극단값 절단(trimming) 방법

- 기하평균을 이용한 이상치 제거
    - 하단, 상단 백분율 이용한 이상치 제거: 상, 하위 5%에 해당되는 데이터 제거

- ② 극단값 조정(winsorizing) 방법

- 상한값과 하한값을 벗어나는 값들을 상한값, 하한값으로 바꿈
    - 박스플롯에서 IQR의 약 1.5배 벗어난 데이터를 이상치로 분류



# 4-1 단계 데이터 탐색(explore)

- 이상치 활용 예

- ① 사기탐지 : 정상시의 신용카드 구매 패턴과 다른 패턴을 조사하여 도난 여부 확인
- ② 침입탐지: 컴퓨터 네트워크에 대한 예외적인 행위를 감시하는 경우 탐지
- ③ 의료: 환자에게 보이는 예외적인 이상 증세를 발견함으로써 건강 이상 발견
- ④ 기계: 기계 장비의 작동에 이상 증세를 발견함으로써 고장 탐지

## 4-2 단계 데이터 전처리(pre-process)

종류	설명
데이터 여과	• 오류 발견, 보정, 삭제, 중복성 확인 등의 과정을 통해 데이터 품질을 향상시킨다.
데이터 정제	• 결측치는 채워 넣고 이상치는 식별 또는 제거하고 잡음이 섞인 데이터는 평활화하여 데이터 불일치성을 교정한다.
데이터 통합	• 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터(또는 데이터베이스)를 통합한다.
데이터 축소	• 분석 시간을 단축하기 위해 분석에 사용하지 않는 항목은 제거한다.
데이터 변환	• 데이터 분석에 용이한 형태로 데이터 유형을 변환한다. • 정규화normalization, 집합화aggregation, 요약summarization, 계층 생성 등의 방법을 활용한다. • ETLExtraction, Transformation, Loading 도구를 제공한다.

## 4-2 단계 데이터 전처리: 여과 및 정제

- 데이터의 품질 문제를 해결하기 위해 정제 후 사용 가능한 형태로 가공하는 단계이다.
- 데이터 정제는 부정확한 값, 결측치(missing value), 불일치(inconsistency), 잡음(noise) 등을 제거하고 데이터의 범위를 벗어난 데이터 및 특이값을 제거하는 단계이다.
  - 결측치 제거
  - 중복 레코드 병합
  - 유효하지 않은 값에 대한 최상의 추정치 생성
  - 이상치 제거
  - 데이터 형태 변환(벡터화)
- 도메인 지식(domain knowledge)이 필요하다.

## 4-2 단계 데이터 전처리: 데이터 먼징(Data Munging)

- 데이터 먼징(Data Munging)이란 원데이터를 쉽고 효율적으로 가공하고 분석할 수 있도록 변환하는 과정을 의미한다.
  - 차원 축소(Dimensionality reduction)
  - 사전 처리(Data manipulation)
  - 변환(Transformation)
  - 속성 선택(Feature selection)
  - 스케일링(scaling)

## 4-2 단계 데이터 전처리: 결측치 처리

- 단순 대치법
  - Completes analysis: 결측값이 존재하는 레코드를 삭제해 버림
  - 평균 대치법: 관측/실험을 통해 얻어진 데이터의 평균으로 메꿈
    - 조건부 평균대치법: 결측값 있는 변수를 종속변수로하는 회귀분석 활용
    - 비조건부 평균대치법: 관측데이터의 단순 평균 사용
- 단순확률 대치법
  - 평균대치법에서 추정량 표준오차의 과소추정문제를 보완
  - Hot-Deck방법, nearest-neighbor 방법 등이 있음
- 다중 대치법
  - 단순대치법을 m번 반복해서 m개의 완전한 자료를 가상적으로 만들어내는 방법
  - 1단계 대치 > 2단계 분석 > 3단계 결합

## 4-2 단계 데이터 전처리: 결측치 처리

### 결측치(Missing Value)

- **NA**(Not Available)

for missing or undefined data

- **NULL**

for empty object(e.g. empty lists)

- **None**

- **NaN(Not a Number)** for results that cannot be reasonably defined

	A	B	C
1	1	4	7
2	2		8
3	3	6	9

0	1	4	7
0	2	NaN	8
1	3	6.0	9

## 4-2 단계 데이터 전처리: 스케일링(scaling)

### 특성 스케일링(Feature Scaling)

특성(크기, 범위 및 단위)이 서로 다른 변수 간 연산을 위해 특성을 통일함

#### 정규화 (normalization)

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

가장 큰 값은 1로, 가장 작은 값은 0으로 변환하여  
모두 [0, 1]의 범위를 갖도록 함.

MinMaxScaler

#### 표준화 (Standardization)

$$x' = \frac{x - \mu}{\sigma}$$

평균  $\mu$ 와 표준편차  $\sigma$ 를 기준으로 전체평균  
0과 표준편차 1을 갖도록  $x'$ 를 변환

Z-score

## 4-2 단계 데이터 전처리: 스케일링(scaling)

단위가 서로 다른 두 변수의 데이터 프레임 만들기

```
# 단위가 서로다른 두 변수의 데이터 프레임 만들기
df = pd.DataFrame([[100, 500, 300, 400, 150], [1.5, 0.6, 0.8, 0.1, 0.3]]).T
df.columns = ['코끼리', '개미']
print(df)
```

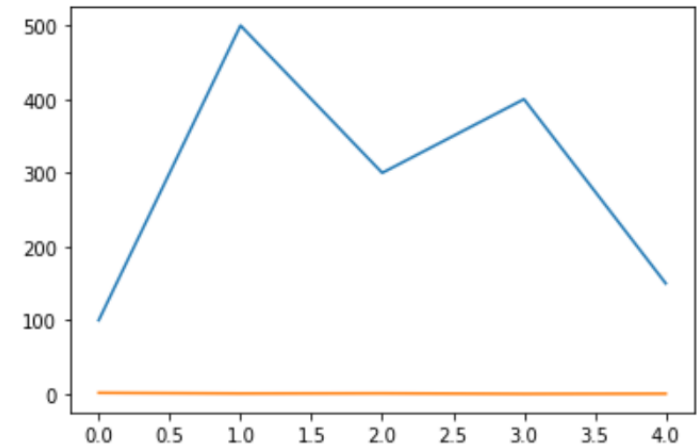
```
↗
```

	코끼리	개미
0	100.0	1.5
1	500.0	0.6
2	300.0	0.8
3	400.0	0.1
4	150.0	0.3

```
# 그래프로 원데이터 확인하기
import matplotlib.pyplot as plt
plt.plot(df)
```

```
↗
```

[<matplotlib.lines.Line2D at 0x7fc8048f79d0>,  
<matplotlib.lines.Line2D at 0x7fc8048f7bd0>]





## 4-2 단계 데이터 전처리: 스케일링(scaling)

정규화  
(normalization)

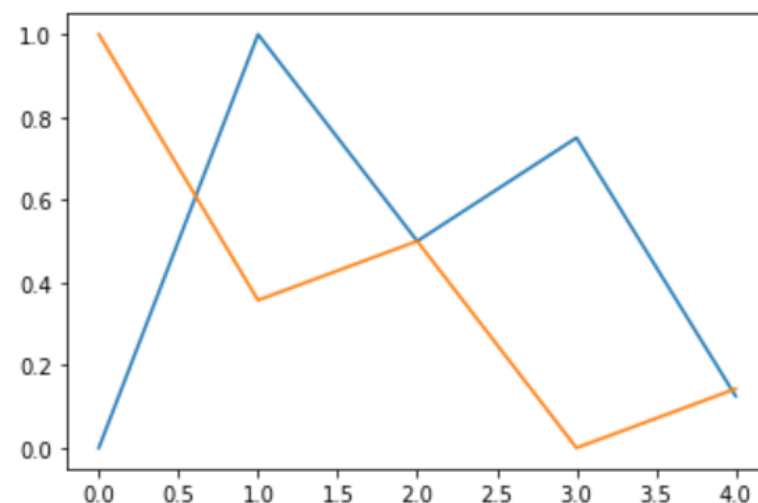
$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

```
# 정규화(normalization)
# MinMaxScaler
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df_minmax= scaler.fit_transform(df)
print(df_minmax)
```

```
[ [0.      1.      ]
  [1.      0.35714286]
  [0.5     0.5     ]
  [0.75    0.      ]
  [0.125   0.14285714]]
```

```
# 그래프로 정규화 데이터 확인하기
import matplotlib.pyplot as plt
plt.plot(df_minmax)
```

```
[<matplotlib.lines.Line2D at 0x7fc80486ee10>,
 <matplotlib.lines.Line2D at 0x7fc80486efd0>]
```



## 4-2 단계 데이터 전처리: 스케일링(scaling)

### 표준화 (Standardization)

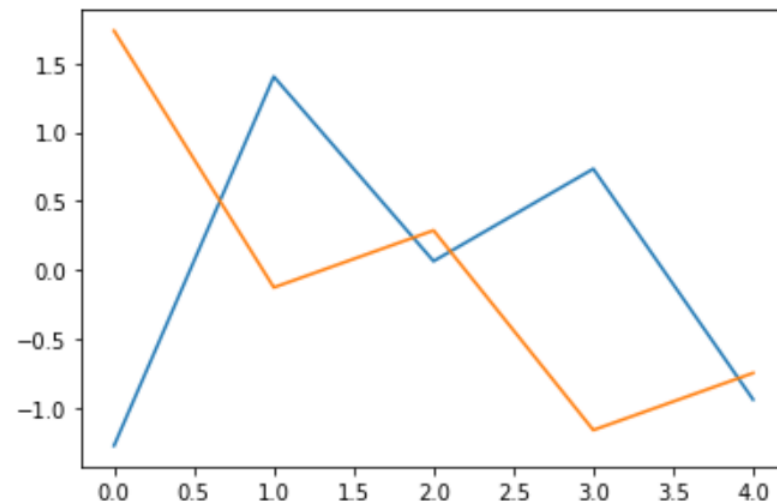
$$x' = \frac{x - \mu}{\sigma}$$

```
▶ # 표준화(standardization)
# z-score
from sklearn.preprocessing import StandardScaler
df_zscore = StandardScaler().fit_transform(df)
print(df_zscore)
```

```
↳ [[-1.2694909  1.73500401]
     [ 1.40312152 -0.12392886]
     [ 0.06681531  0.28916733]
     [ 0.73496842 -1.15666934]
     [-0.93541435 -0.74357315]]
```

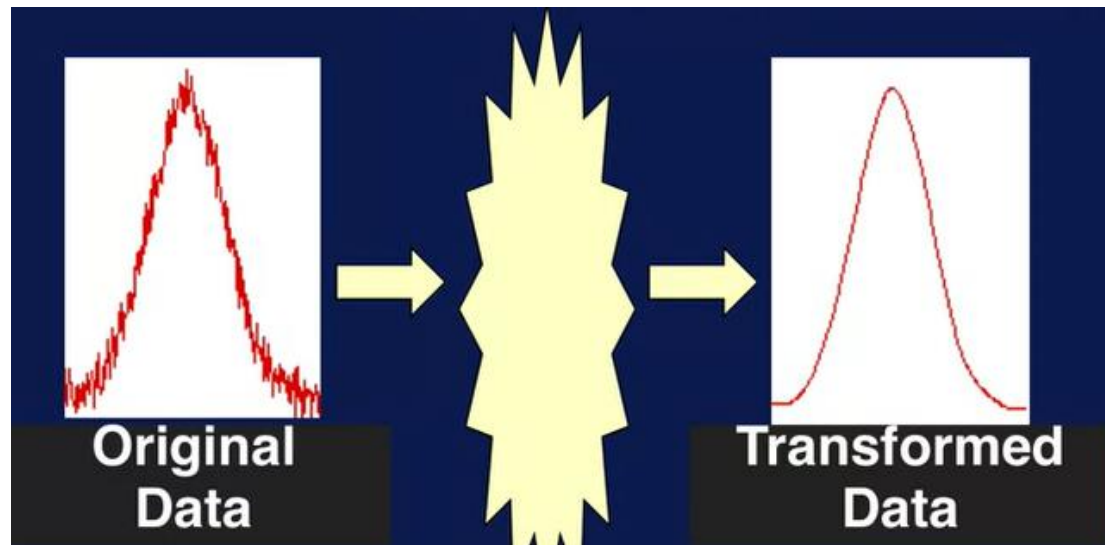
```
▶ # 그래프로 표준화 데이터 확인하기
import matplotlib.pyplot as plt
plt.plot(df_zscore)
```

```
↳ [<matplotlib.lines.Line2D at 0x7fc8047dd050>,
    <matplotlib.lines.Line2D at 0x7fc8047dda50>]
```



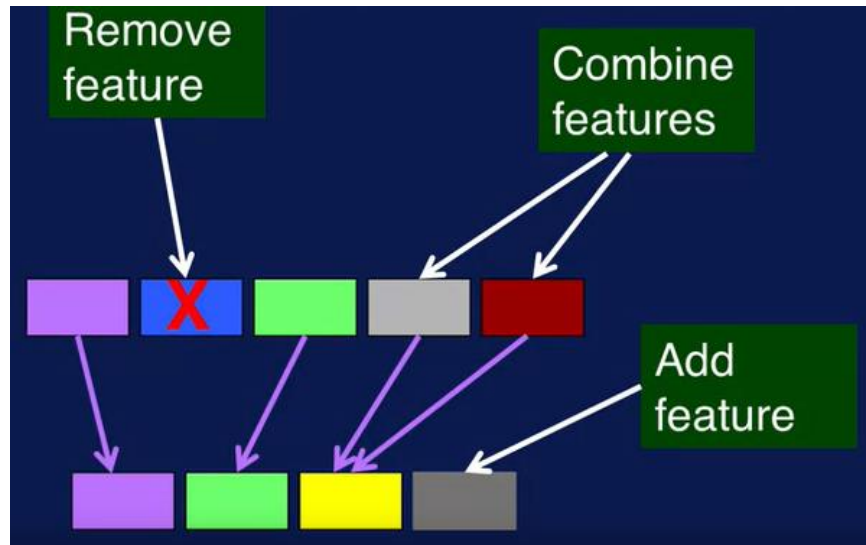
## 4-2 단계 데이터 전처리: 변환(Transformation)

- 데이터 변환 단계에서는 이미 존재하는 필드로부터 새로운 데이터 필드를 생성하거나 더 많은 정보를 포함하도록 몇 개의 필드를 하나의 필드로 변환하는 등의 작업을 수행하고 선택된 데이터가 특정 데이터 마이닝 알고리즘 수행에 적당하도록 데이터 값을 변형한다.
- 일일 단위의 데이터 변동이 심한 경우, 주 단위, 월 단위의 집계 데이터(Aggregate data)로 변환



## 4-2 단계 데이터 전처리(pre-process)

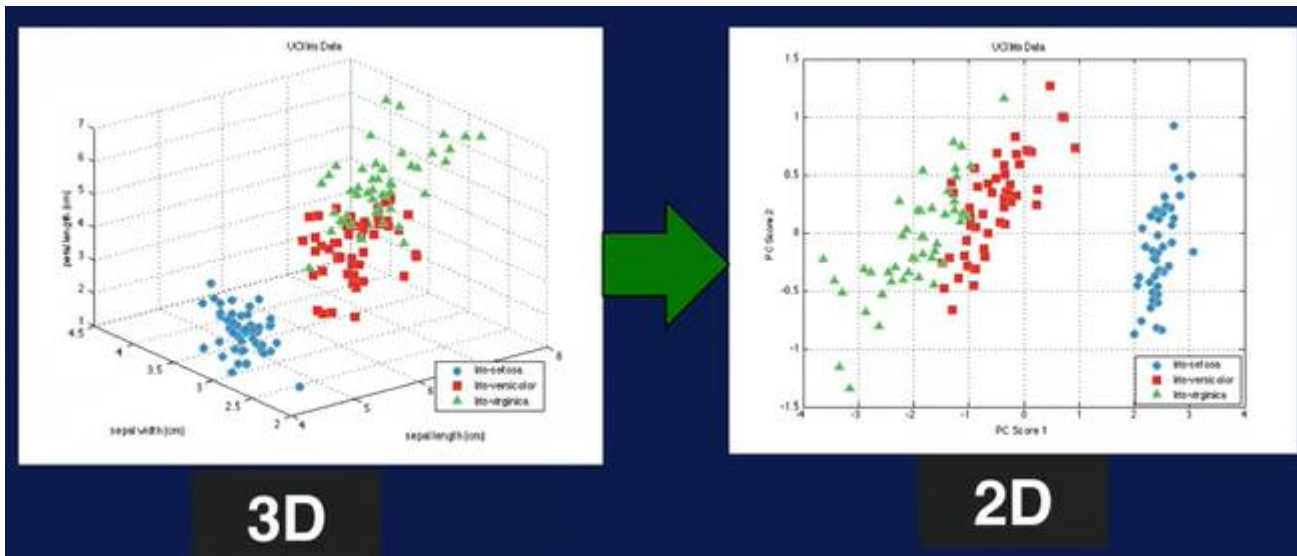
속성 선택(feature selection)



중복되거나 관련성이 없는 속성을 제거하면 후속 분석이 훨씬 간단해진다.

## 4-2 단계 데이터 전처리: 차원 축소

차원 축소(Dimensionality reduction)



데이터 세트에 많은 수의 치수가 있는 경우에 유용하다.

일반적으로 사용되는 기술을 주성분 분석 (principle component analysis, PCA)이다.

## 4-2 단계 데이터 전처리: 차원축소

- 차원축소(Dimension reduction)는 고차원의 데이터를 정보의 손실을 최소화하면서 저차원으로 변화하는 것을 말한다.
- 차원이 커질수록 학습에 필요한 데이터가 기하급수적으로 증가한다.
- 각 부분공간에 적어도 하나의 학습데이터가 필요하기 때문이다
  - 특이값 분해(SVD)
  - 주성분분석(Principle Component Analysis, PCA)

# 경청 해 주셔서 감사합니다.



본 자료는 교육을 목적으로 제작된 것으로  
다른 목적으로의 사용 및 무단 복사 행위를 금합니다.

경남대 전하용  
hayongj@kyungnam.ac.kr