



# 1주차: 데이터 사이언스 개요(1주 3회차 실시간 강의)

**ChulSoo Park**

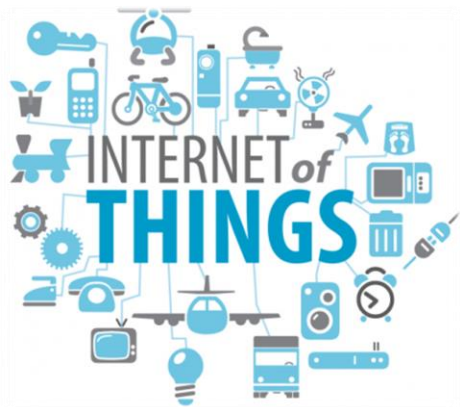
School of Computer Engineering & Information Technology

Korea National University of Transportation

E-Mail : pcs8321@naver.com

# 4차산업혁명 : 사물인터넷(IoT), 빅데이터(Big Data), 인공지능(AI) → 의사결정, 현업활용

출처: 제4차 산업혁명, 김진호, 최용주, p25



## 4차산업혁명 환경하의 디지털 경영혁신

이성열 · 강성근 · 김순신 지음



기업은 변화의 혁신을 통하여 성장한다. 4차 산업혁명과 디지털 혁신으로 바뀌고 있는 경영환경의 급격한 변화에서 전통기업들이 일어나 신속하게 디지털 경영혁신을 통하여 변화에 나갈 수 있는 가는 향후 기업생존의 핵심이 될 것이다.



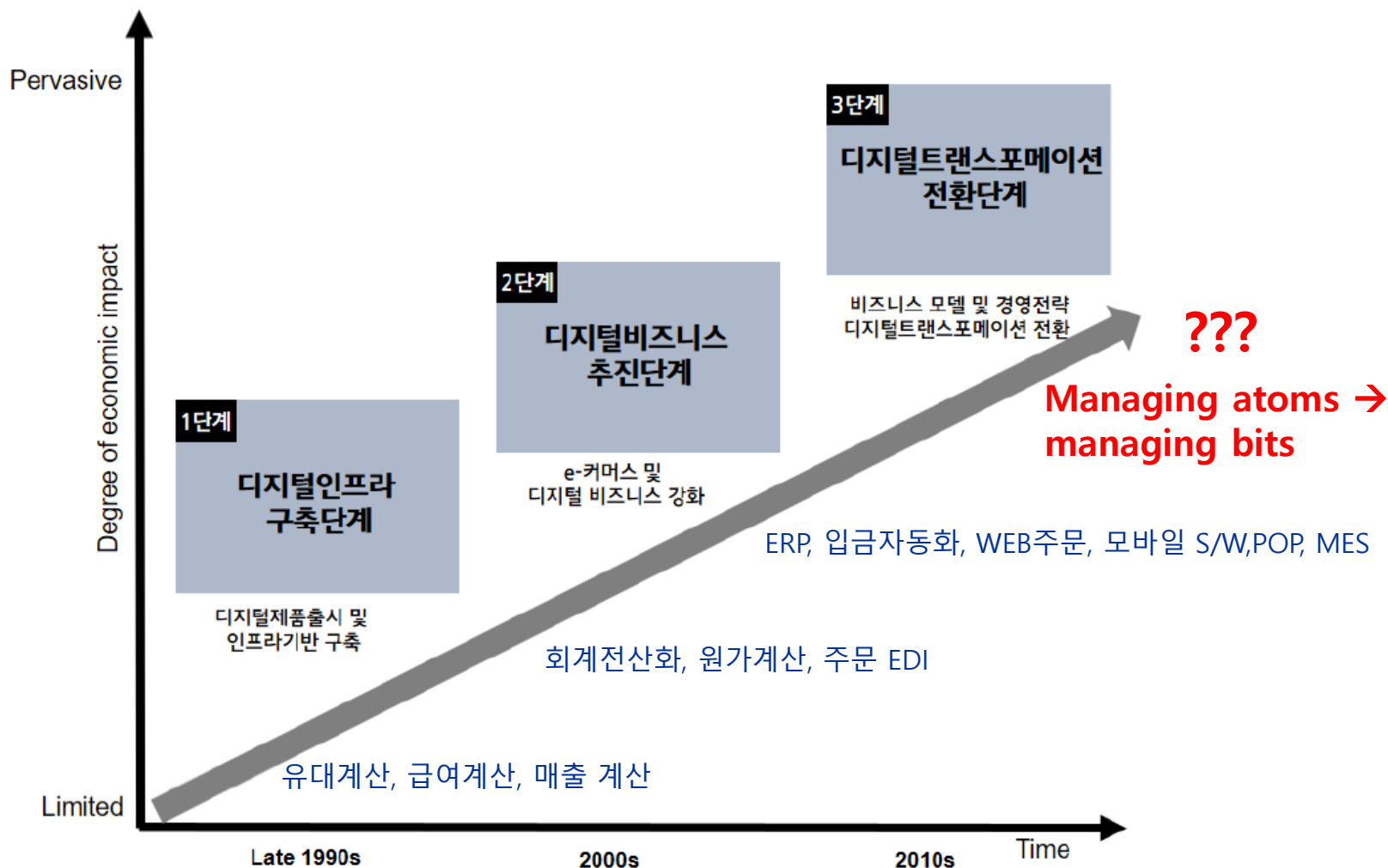
	제1차 산업혁명	제2차 산업혁명	제3차 산업혁명	제4차 산업혁명
시기	18세기	19~20세기 초	20세기 후반	20세기
특징	증기기관 기반의 '기계화 혁명'	전기 에너지 기반의 '대량생산 혁명'	컴퓨터와 인터넷 기반의 '디지털 혁명'	사물인터넷(IoT)과 빅데이터, 인공지능(AI) 기반의 '만물 초지능 혁명'
영향	수공업 시대에서 증기기관을 활용한 기계가 물건을 생산하는 기계화 시대로 변화	전기와 생산조립 라인의 출현으로 대량생산 체계 구축	반도체와 컴퓨터, 인터넷 혁명으로 정보의 생성·가공·공유를 가능케 하는 정보기술시대의 개막	사람, 사물, 공간을 연결하고 자동화·지능화되어 디지털·물리적·생물학적 영역의 경계가 사라지면서 기술이 융합되는 새로운 시대

자료 : 미래에셋대우 글로벌투자전략부



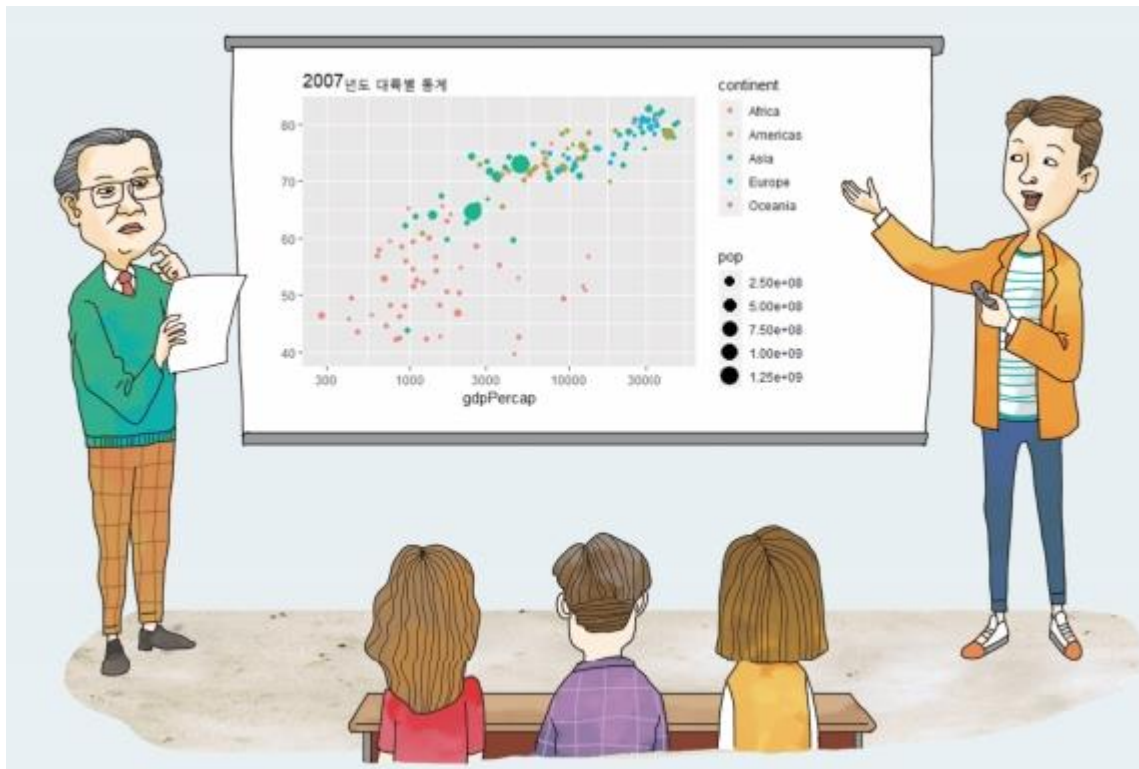
## 디지털 트랜스포메이션 단계별 진화

## 경쟁력 제고

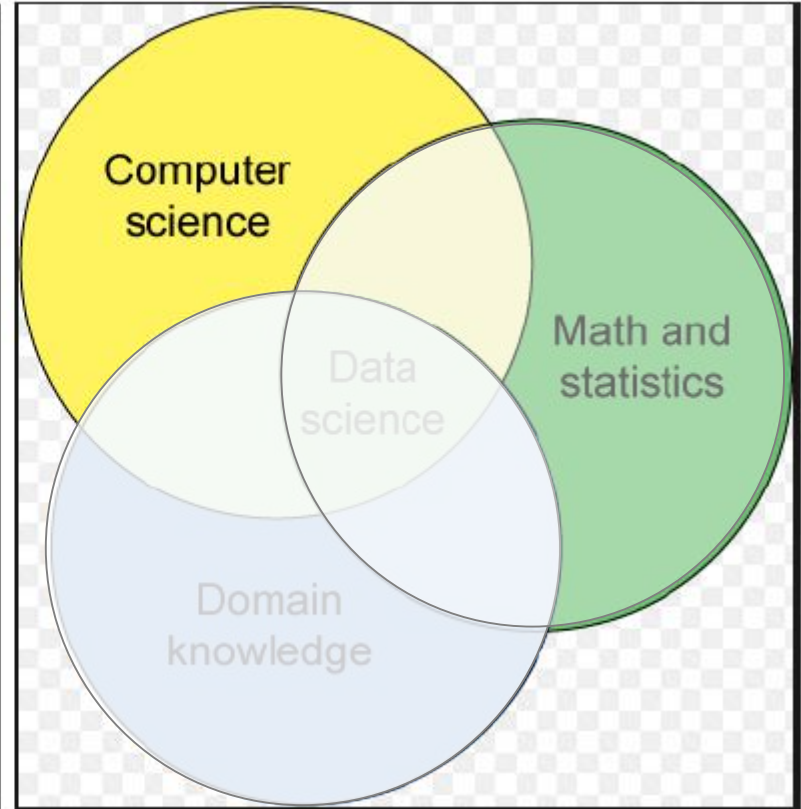
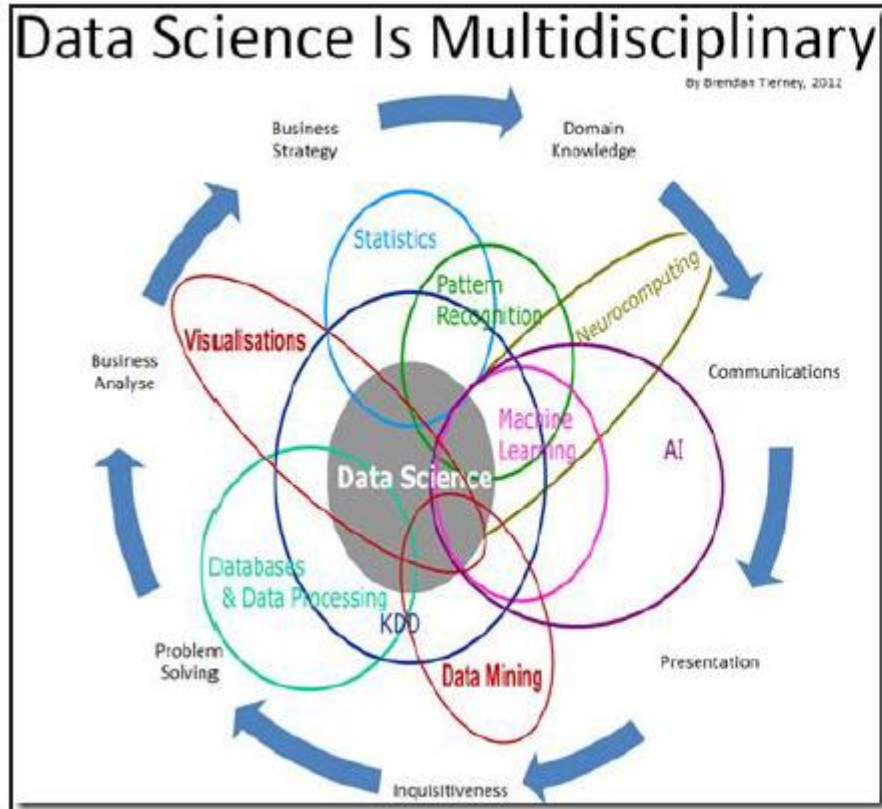


## ■ 데이터 과학

- 데이터로부터 유용한 정보와 통찰을 끄집어내고 합리적인 의사결정을 돕는 현대적인 학문 분야



## 1.3 데이터 과학이란?

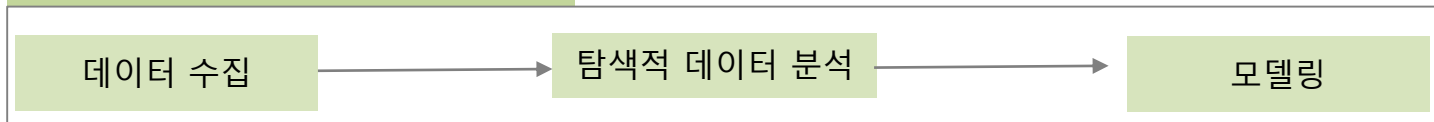


출처 : [www.oralytics.com](http://www.oralytics.com)

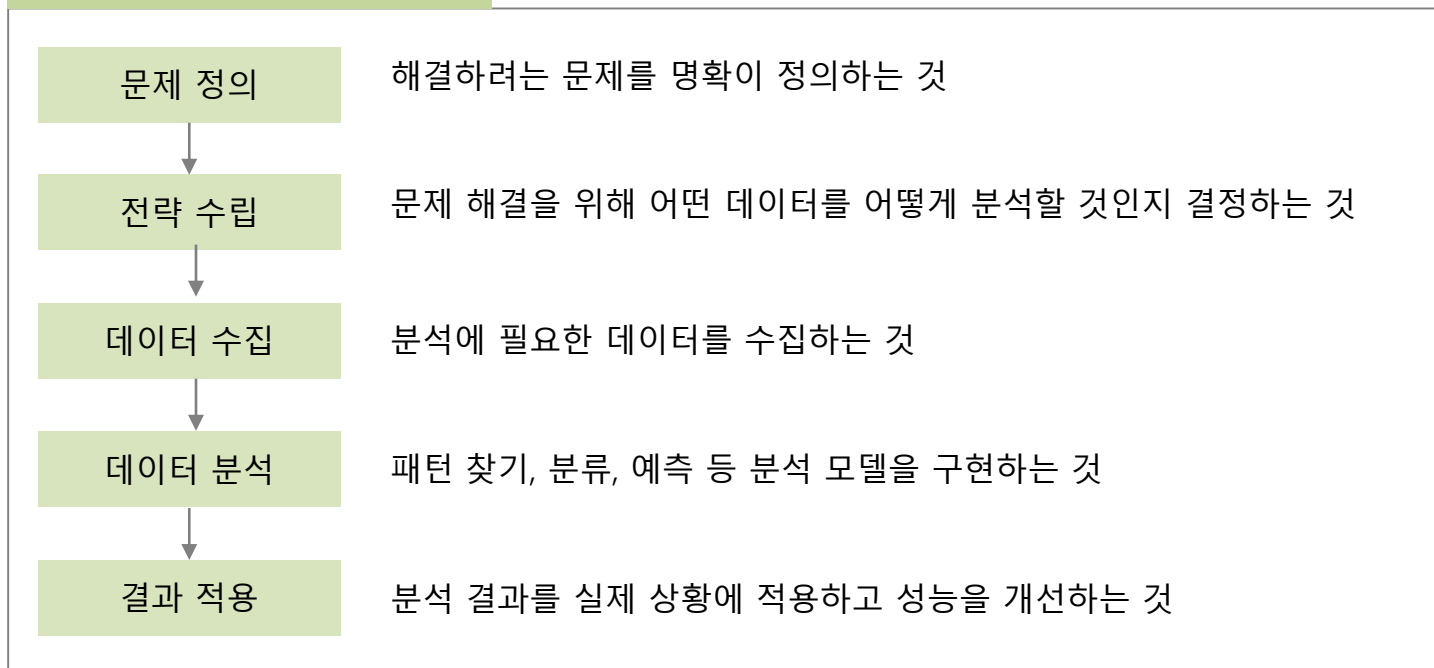
데이터 사이언스 학과(대학원)

# 1.4 데이터 과학의 절차

## 데이터 과학의 절차(그림 1- 5)



## 데이터 사이언스 프로세스





# 1.4 데이터 과학의 절차

축산 빅데이터 컨설팅 플랫폼 개요도



## 데이터 과학의 절차(그림 1- 5) 사례

데이터 수집

탐색적 데이터 분석

모델링

Q. Xie et al. / Journal of Cranio-Maxillo-Facial Surgery 44 (2016) 590–596

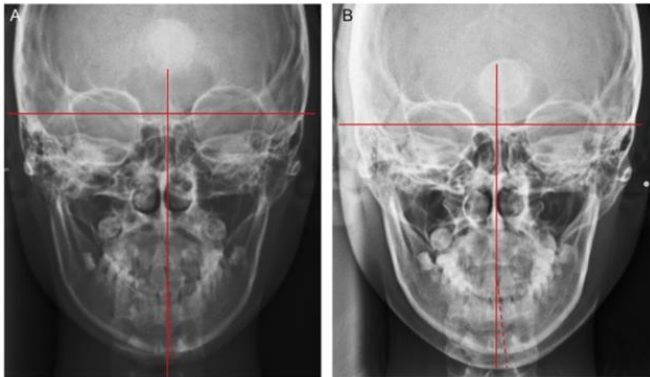


Fig. 6. Before and after PA image: A. first visit: no asymmetry was found, B. revisit: more than 5 mm skeletal asymmetry was noticed.

page 438

```
pkgs <- c("rpart", "rpart.plot", "party", "randomForest", "e1071")
install.packages(pkgs, depend=TRUE)
install.packages("nonparset")
## Load the package
library(nonparset)
library(package="xisc")
library(xisc)
library(lpsolve)
require(xisc) # load xisc package
```

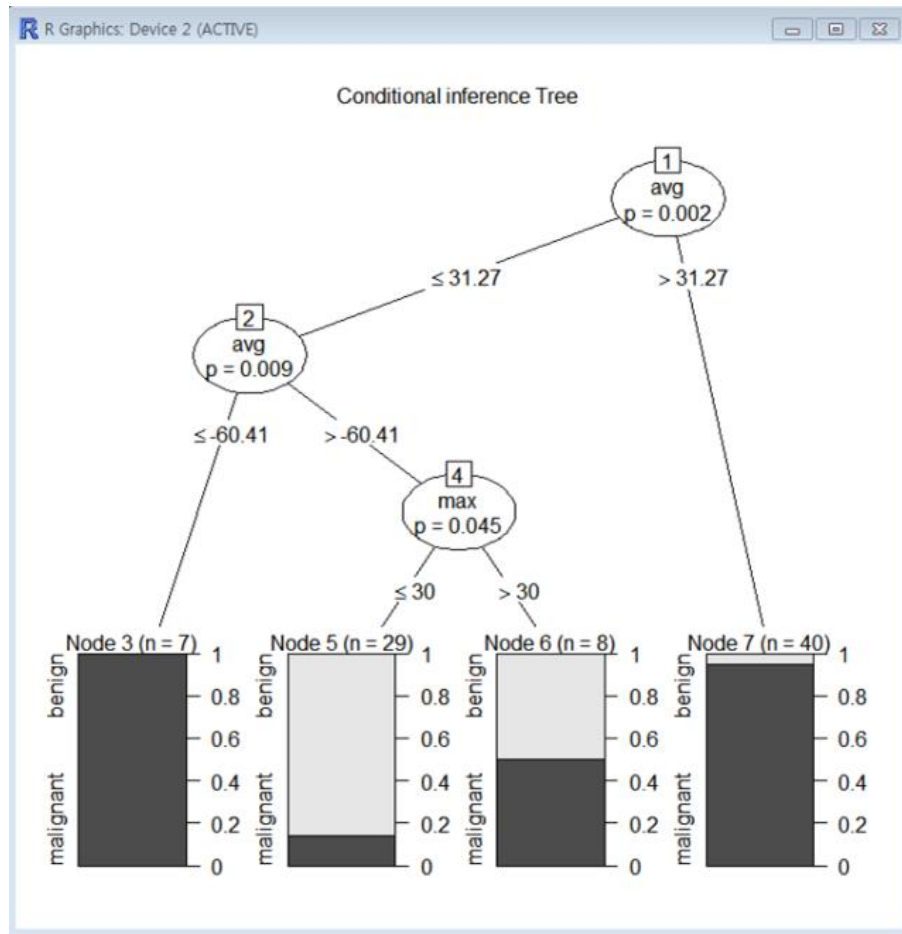
page 439

```
setwd("C:/자료관리/아동/논문/홍수영")
ds <- read.xisc("홍수영_분석_data_0316.xisc", 1)

df <- ds
df$class <- factor(df$class, levels=c(1,0), labels=c("benign", "malignant"))
set.seed(1234)
train <- sample(nrow(df), 0.7*nrow(df))
df.train <- df[train,]
df.validate <- df[-train,]
table(df.train$class)
table(df.validate$class)
```

page 443

```
library(rpart)
set.seed(1234)
```





# 1.4 데이터 과학의 절차

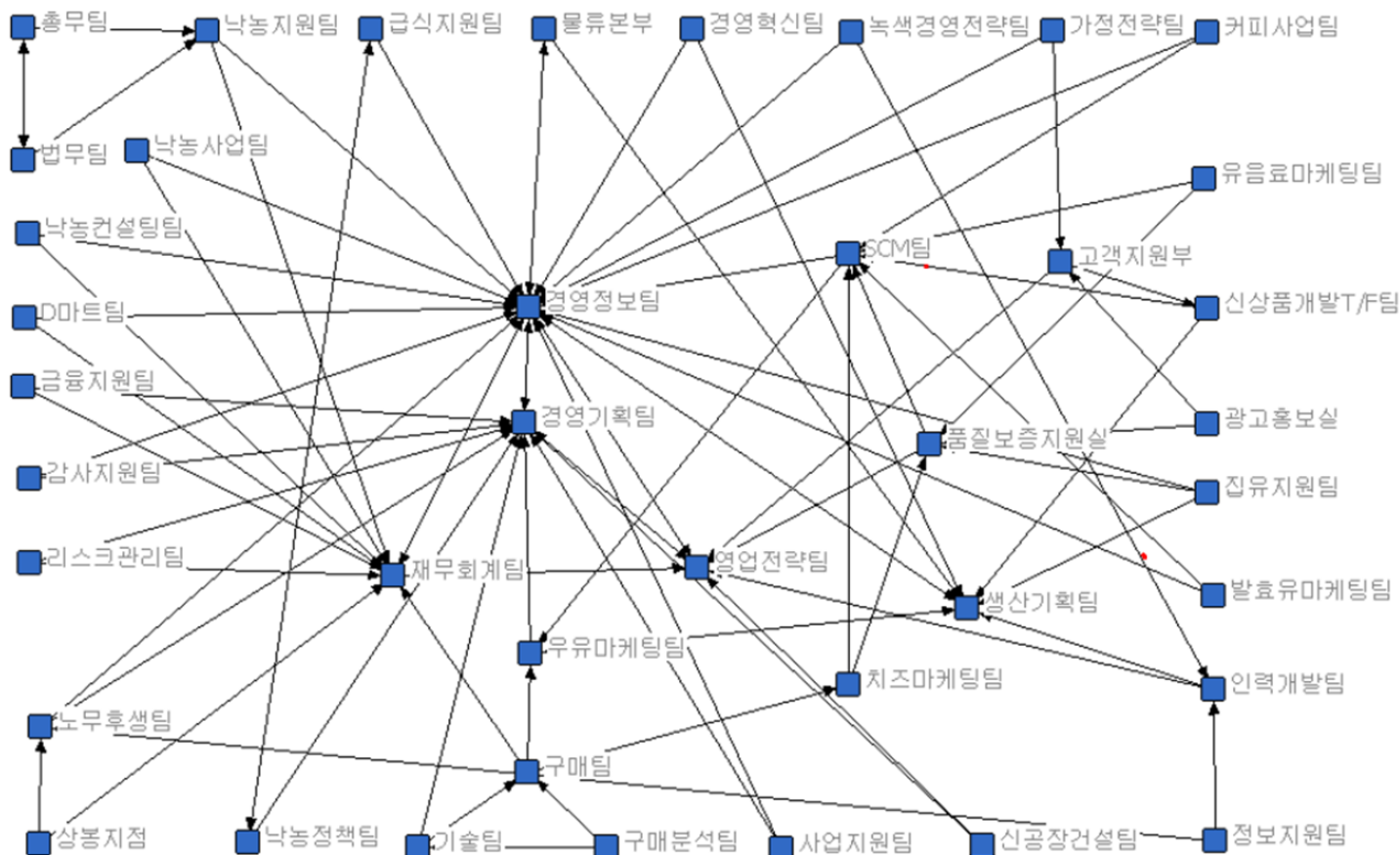
## 데이터 과학의 절차(그림 1- 5) 사례

데이터 수집

탐색적 데이터 분석

모델링

사회연결망(SNS)에서 연결정도 분석의 중앙성은 개인이나 조직의 업무 관련성이나 영향을 나타낸다.



사회 연결망 분석툴 : ucinet 6.0

## 1.4.1 세상과 상호작용하는 데이터 과학

### ■ 데이터 과학은 세상과 활발히 상호작용

- EDA 단계에서 데이터가 부족하다 판단되면 데이터 수집 단계로 돌아가 추가 수집
- 변수를 추가하여 완전 새로 수집하는 상황도 발생
  - 예) 푸드트럭 예에서 골목상권의 영향을 반영하려면 음식점 수와 초밥집이 있는지 여부를 나타내는 변수를 추가하고 새로 데이터를 수집해야 함

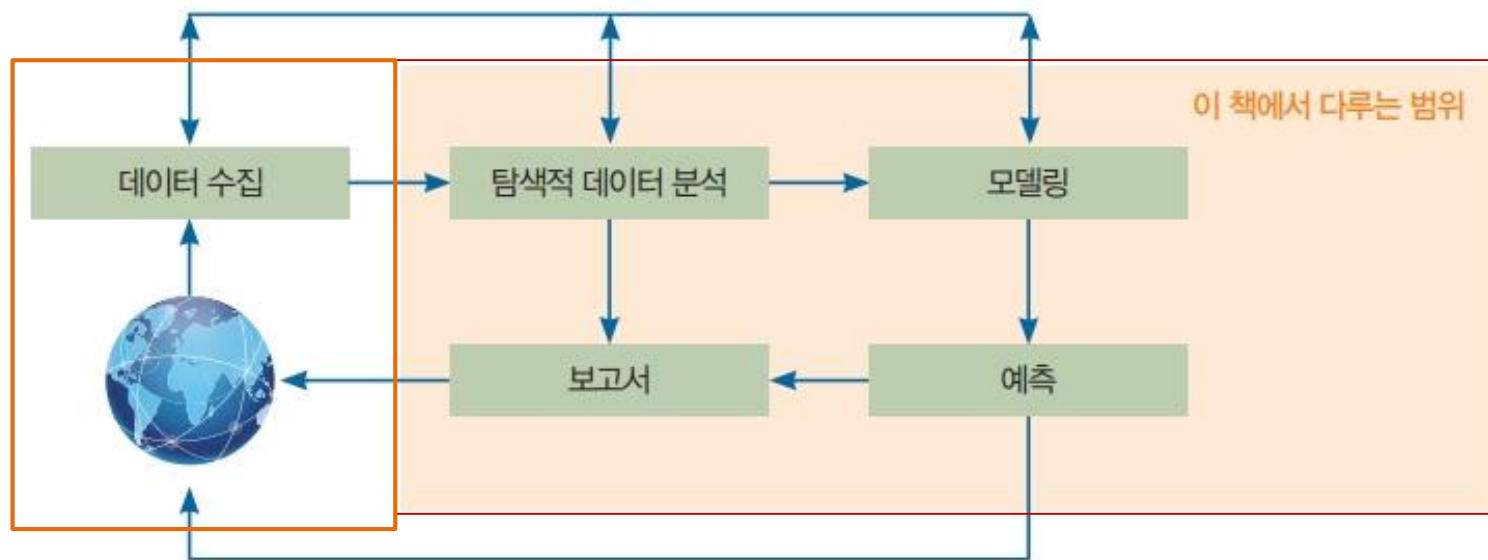
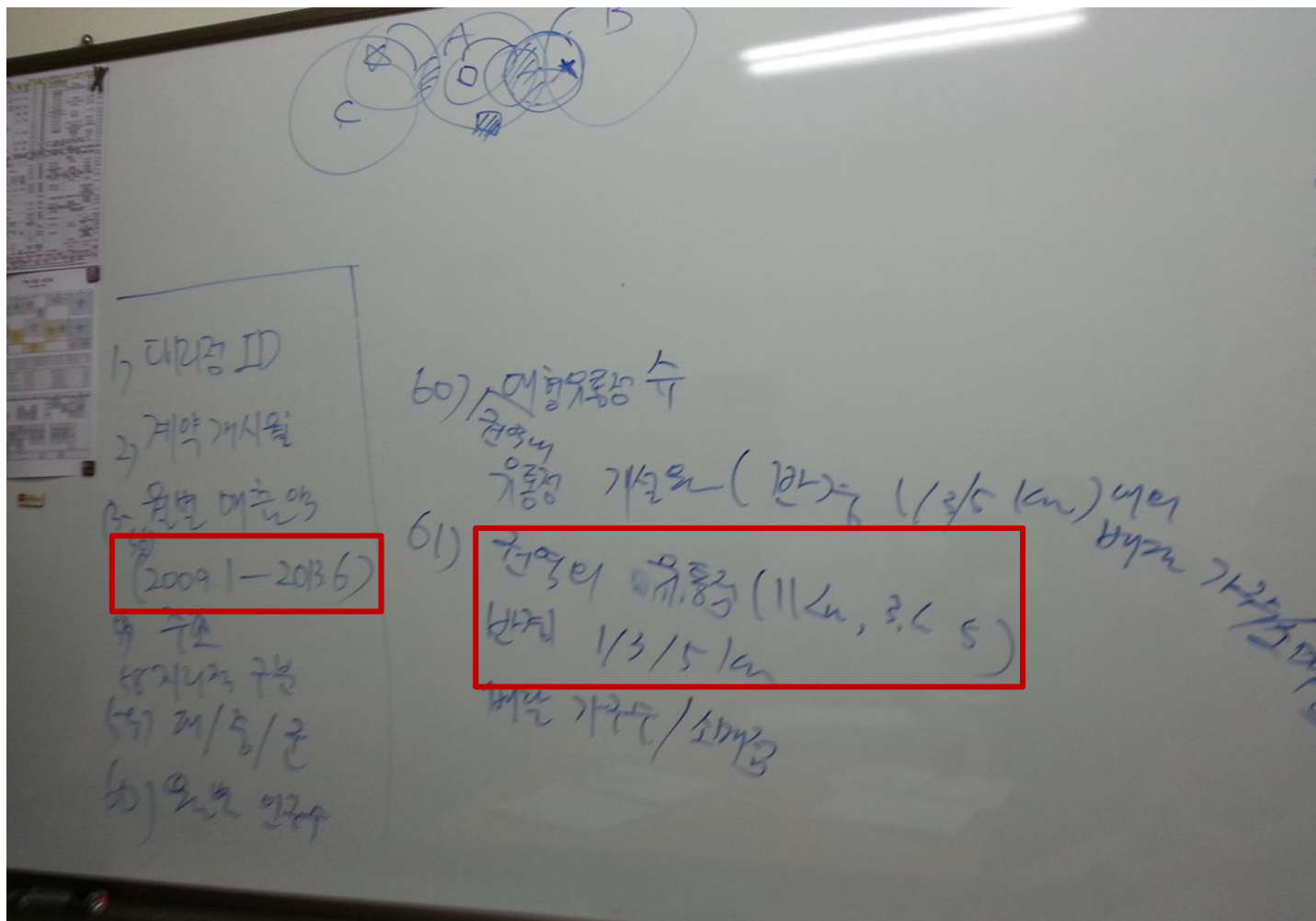


그림 1-7 세상과 상호작용하는 데이터 과학

4.5 만점에 3.5점, 어느 생산 공장의 1달간 data 분석 결과

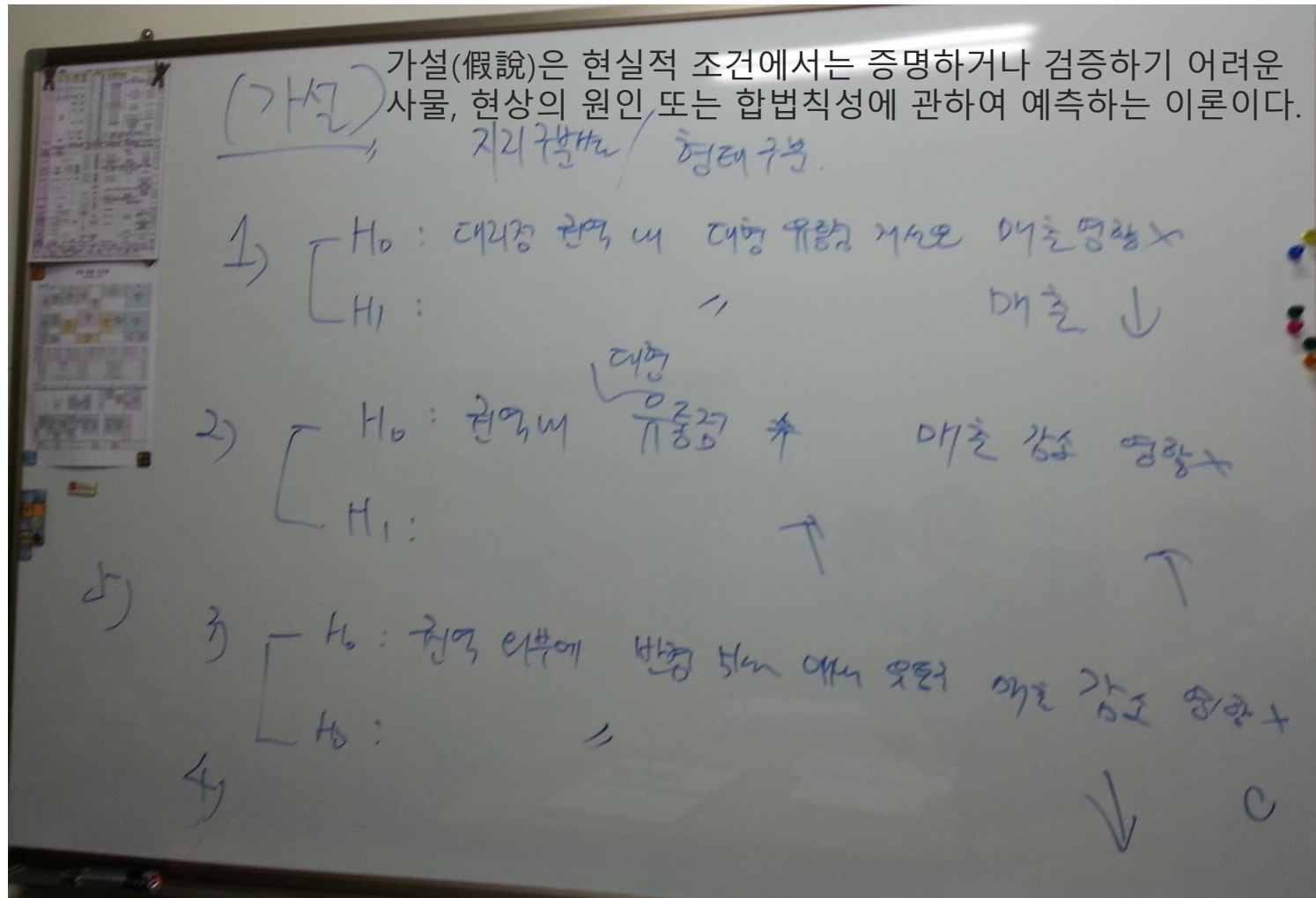
## 1.4.1 세상과 상호작용하는 데이터 과학

### ■ 데이터 수집 과정 사례



## 1.4.1 세상과 상호작용하는 데이터 과학

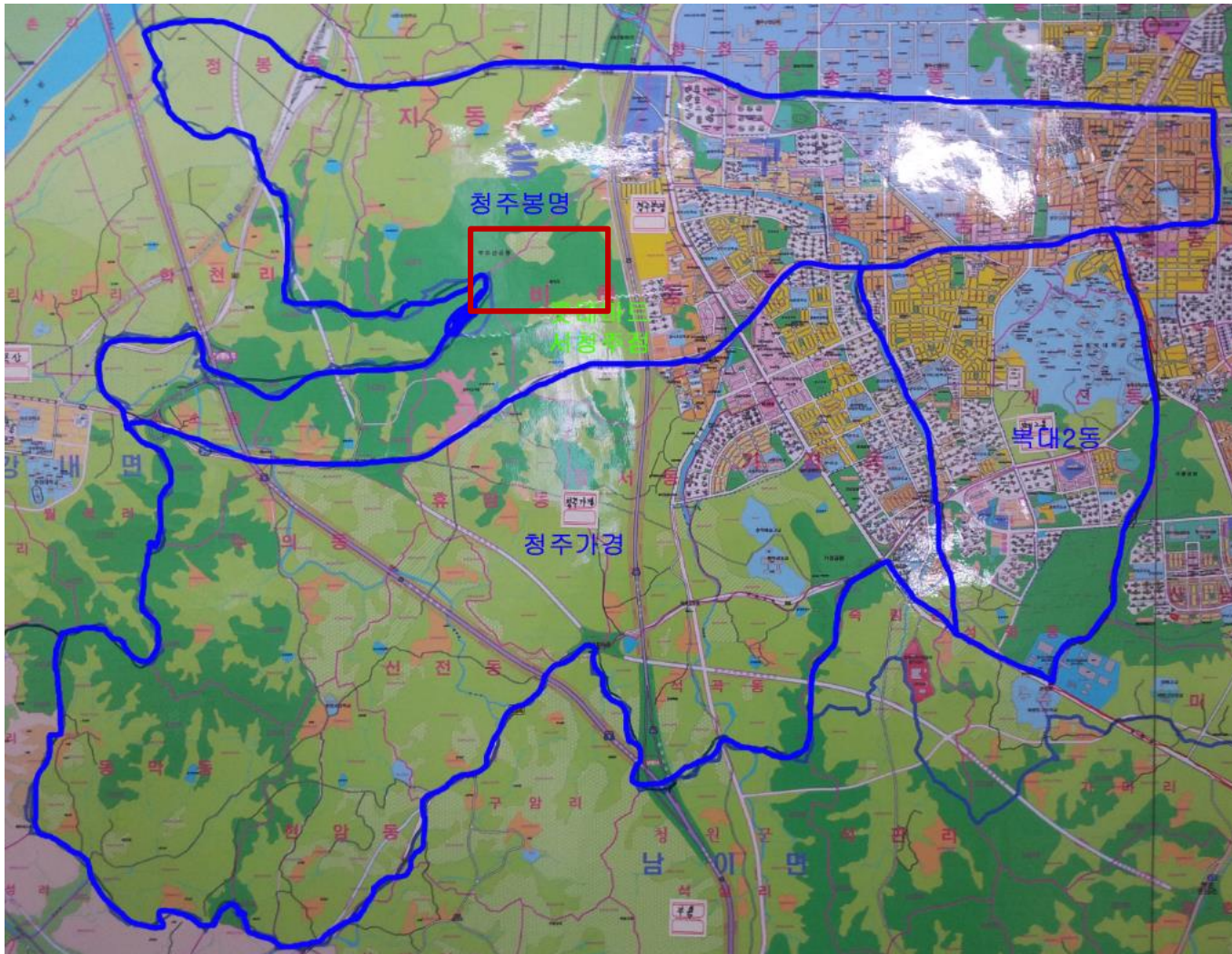
### ■ 데이터 수집 과정 사례





## 1.4.1 세상과 상호작용하는 데이터 과학

### ■ 데이터 수집 과정 사례

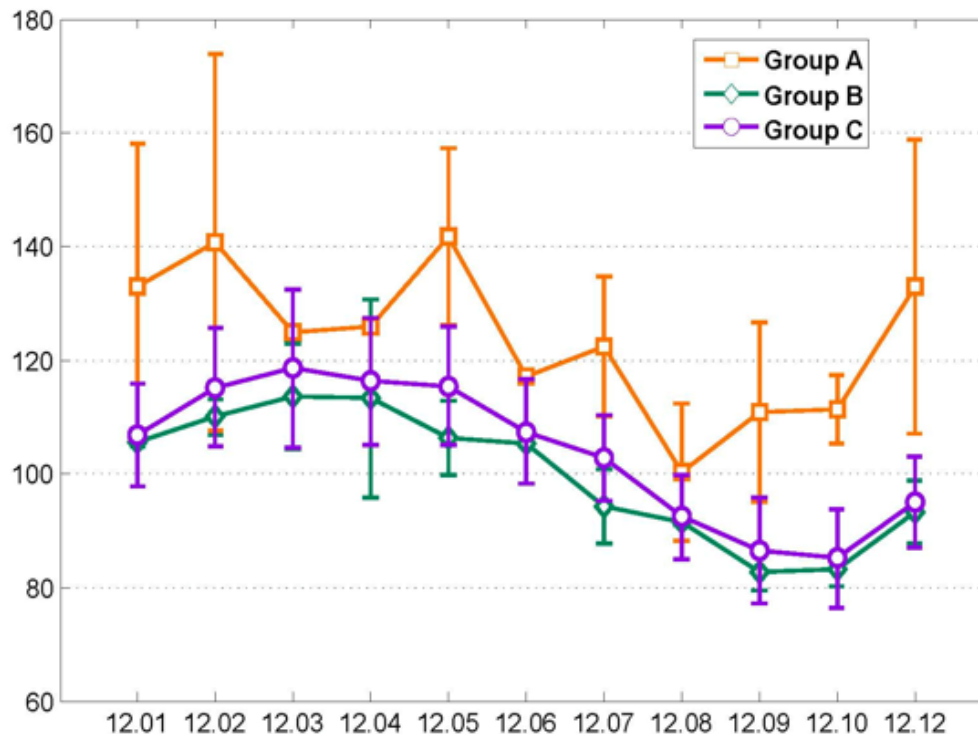


## 1.4.1 세상과 상호작용하는 데이터 과학

### ■ 데이터 분석 결과 사례

대형 유통점 입점과 연계된 Group A/B/C 대리점 수

Group	2012.01	2012.02	2012.03	2012.04	2012.05	2012.06	2012.07	2012.08	2012.09	2012.10	2012.12
A	2	3	1	1	4	1	2	2	3	5	4
B	3	5	3	3	9	4	4	2	7	11	10
C	977	974	978	978	969	977	976	978	972	969	969
계	982	982	982	982	982	982	982	982	982	985	983





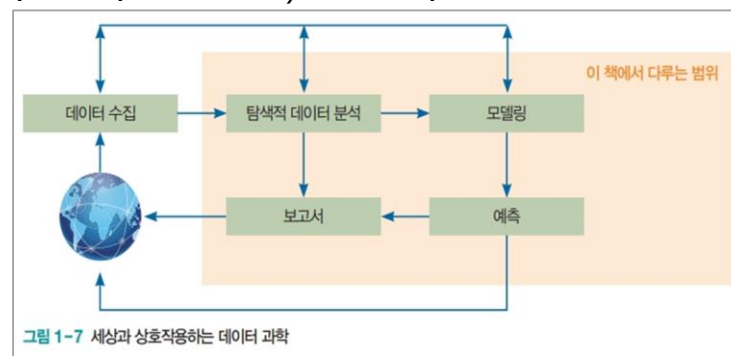
## 1.4.2 이 책에서 다루는 범위와 내용

### ■ 범위

- 데이터 수집은 제외(아주 많은 데이터가 인터넷에 공개. 1.6절에서 데이터 저장소 소개)
- 그림 1-7의 분홍색 표시한 부분(탐색적 데이터 분석+모델링)으로 국한

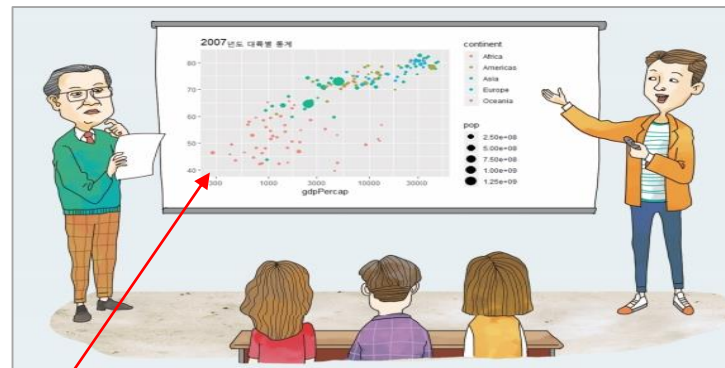
### ■ 내용

- 탐색적 데이터 분석: 2~6장 (중간 고사 전)
- 모델링과 예측: 7~11장 (중간 고사 후)



### ■ 프로그래밍 도구

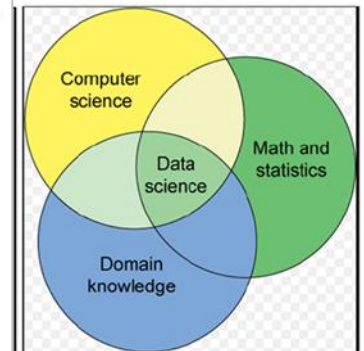
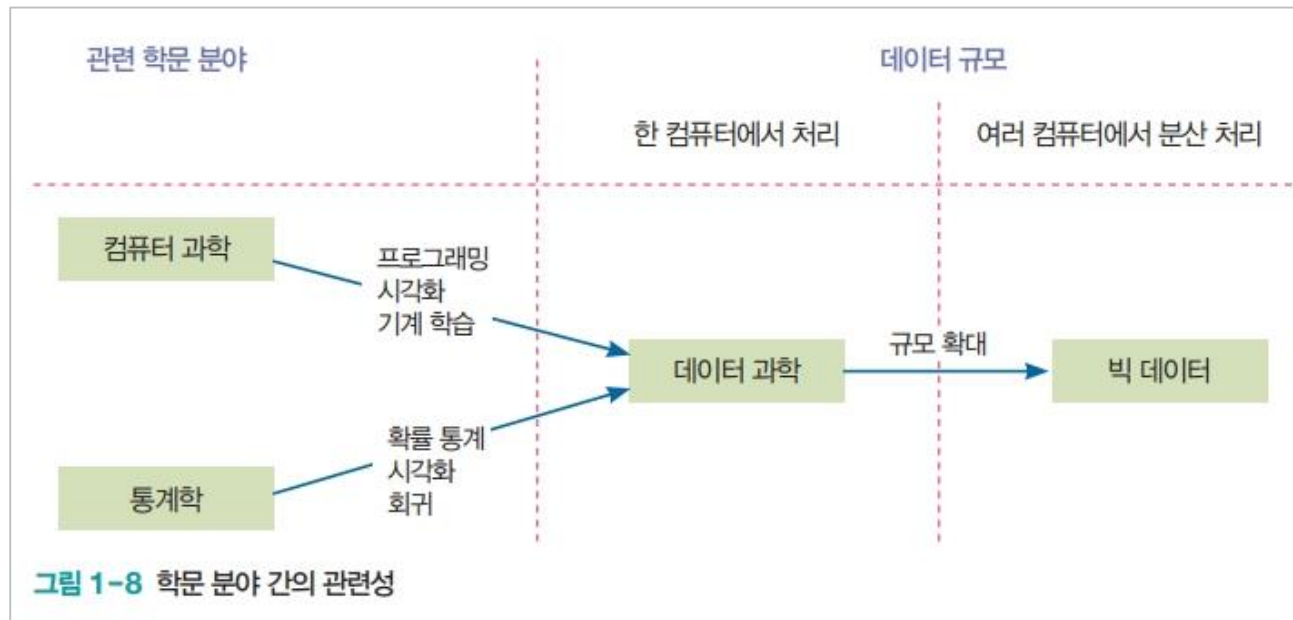
- 이 책은 R을 사용
- 아래 코드처럼 한 줄로 멋진 그림



```
> gapminder%>%filter(year==2007)%>%ggplot(aes(gdpPercap, lifeExp))+geom_
point(aes(size=pop, col=continent))+scale_x_log10()+ggtitle("2007년도 대륙별 통계")
```

## ■ 데이터 과학은 다학제 분야

- 컴퓨터 과학: 프로그래밍 언어, 현대적 시각화 기법, 기계 학습 등 (1)
- 통계학: 요약 통계, 확률 통계 기법, 시각화 도구, 회귀 기법 등 (2)
- 도메인 지식 : 분석하고자하는 분야의 도메인 지식 (3)
- 빅데이터: 분산 처리, 하둡 등
- 데이터 마이닝 : 데이터로부터 유용한 지식을 찾아내는 과정
- 이 책에서는 굳이 데이터 과학과 데이터 마이닝을 구분하지 않는다. (p.31)



데이터 과학자 요구 기술 정리(2014년 8월 MarketingDistillery.com에서 소개함)

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

### PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

### DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

### COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



- 적용하려고 하는 영역을 이해하겠다는 열정
- 데이터에 대한 끝없는 호기심
- 전략적이면서 능동적이고 창의적이면서도 협업에 능함
- 상급자와의 원만한 관계형성 능력
- 스토리텔링 능력
- 예술적 감각

**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
- Sharding
- OLAP
- Multidimensional Data Model
- ETL
- Reporting Vs BI Vs Analytics
- JSON & XML
- Regex
- Vendor Landscape
- Env Setup
- Entropy
- Percentiles & Outliers
- Histograms
- Exploratory Data Analysis
- Descriptive Statistics (mean, median, range, SD, Var)
- Pick a Dataset (UCI Repo)
- 5%

**2. Statistics**

- ANOVA
- Skewness
- Continuous Distributions (Normal, Poisson, Gaussian)
- Cumul Dist Fn (CDF)
- Random Variables
- Bayes Theorem
- Probability Theory
- Factorial Analysis
- Install Packages
- 5%

**3. Programming**

- Python Basics
- Working in Excel
- Expressions
- R Basics
- R Setup
- R Studio
- 10%

**4. Machine Learning**

- Prob Den Fn (PDF)
- ANOVA
- Central Limit Theorem
- Monte Carlo Method
- Hypothesis Testing
- P-Value
- Chi Test
- Equation
- Conf Int (CI)
- MLE
- K-fold Cross Validation
- Regression
- Covariance & Correlation
- Decision Tree
- 10%

**5. Text Mining / NLP**

- Named Entity Recognition
- Text Analysis
- Term Document Matrix
- Support Vector Machines
- Association Rules
- Marked Based Analysis
- Using ETL
- How much Data?
- Google OpenRefine
- Feature Extraction
- Using Mahout
- Using Weka
- Using NLTK
- 30%

**6. Visualization**

- Scatter Plot (Bi)
- Line Charts (Bi)
- Spatial Charts
- Survey Plot
- Timeline
- Decision Tree
- 10%

**7. Big Data**

- Job & Task Tracker
- Map Programming
- Setup Hadoop (IBM / Cloudera / HortonWorks)
- Data Replication Principles
- HDFS
- Hadoop Components
- Map Reduce Fundamentals
- 10%

**8. Data Ingestion**

- 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
- 10%

**9. Data Munging**

- Sampling
- Stratified Sampling
- Principal Component Analysis
- Transformation & Enrichment
- Data Fusion
- Data Integration
- Data Sources & Acquisition
- Data Discovery
- Summary of Data Formats
- 10%

**10. Toolbox**

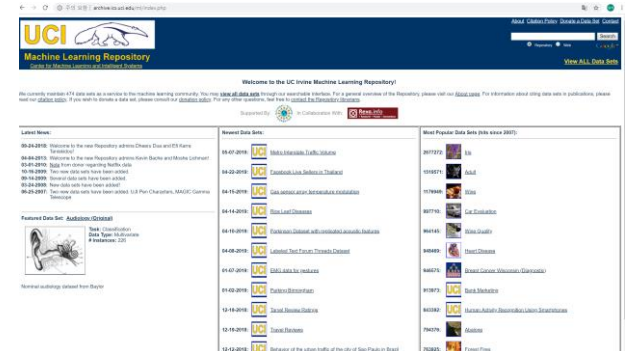
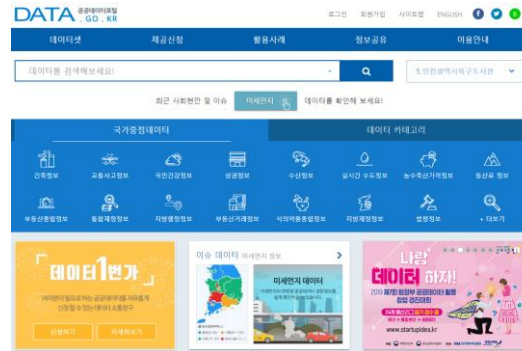
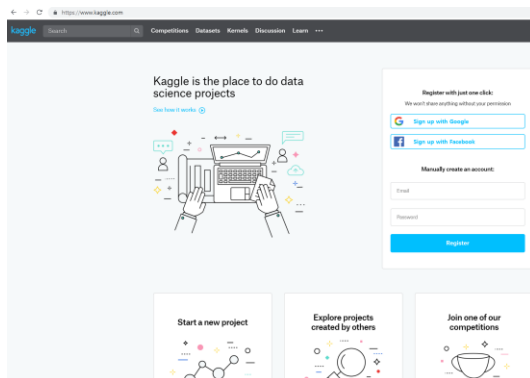
- MS Excel w/ Analysis ToolPak
- Java, Python
- R, R-Studio, Rattle
- Weka, Knime, RapidMiner
- Hadoop Dist of Choice
- Spark, Storm
- Flume, Scribe, Chukwa
- Nutch, Talend, ScraperWiki
- Webcrawler, Flume, Sqoop
- tm, RWeka, NLTK
- RHPIPE
- D3.js, ggplot2, Shiny
- Cassandra, MongoDB
- 10%

**Author: Swami Chandrasekar**

- ① 기본 기술
- ② 통계학
- ③ 프로그래밍
- ④ 기계학습
- ⑤ 텍스트마이닝/자연어 처리
- ⑥ 데이터 시각화
- ⑦ 빅데이터
- ⑧ 데이터 가공 및 통합
- ⑨ 데이터 표준화와 변수 선택 등
- ⑩ 기본 도구 활용

## ■ 데이터 저장소

- 캐글 (<https://www.kaggle.com/>)
- 대한민국 공공데이터 포털 (<https://www.data.go.kr/>)
- UCI machine learning repository (<http://archive.ics.uci.edu>)
- 위키 "list of datasets for machine learning research" (<https://www.wikipedia.org/>)
- ...



## 온라인 교육 사이트

- 데이터 사이언스 아카데미 (<http://datascienceacademy.com/free-data-science-courses>)
- 에덱스 (<https://www.edx.org/course/subject/data-analysis-statistics>)
- 코세라 (<https://www.coursera.org/courses?query=data%20science>)
- Data camp (<https://www.datacamp.com/>)
- ...

## Data Analysis & Statistics Courses

Learn about data analysis and statistics and more from the best universities and institutions around the world.

Home > All Subjects > Data Analysis & Statistics

Data is the foundation of the Digital Age. Learn how to organize, analyze and interpret these new and vast sources of information. Online courses from top institutions cover topics such as machine learning, business analytics, probability, randomization, quantitative methods and much more.

Related Topics - Computer Programming | Artificial Intelligence | Big Data | Business Intelligence | C Programming | Data Analysis | Data Mining | Data Science | Data Visualization | Excel | Information Technology | Java | Machine Learning | Predictive Analytics | R Programming | Master's in Data Science | Master's in Analytics | Python | SQL



GalileoX  
Analisis estadístico con Excel

Current  
Colf, Baced



EPFLx  
Programming Reactive Systems

Current  
Colf, Baced



UCSanDiegoX  
Python for Data Science

Current  
Colf, Baced

The screenshot shows the Coursera website with a search bar containing 'data science'. Below the search bar, several course listings are visible:

- IBM Data Science Professional Certificate** (SPECIALIZATION) by IBM, 4.6 rating (41,263 reviews), 160K students, Beginner level.
- SQL for Data Science** (COURSE) by University of California, Davis, 4.6 rating (1,797 reviews), 87K students, Beginner level.
- What is Data Science?** (COURSE) by IBM, 4.7 rating (11,450 reviews), 80K students, Beginner level.
- Applied Data Science with Python** (SPECIALIZATION) by University of Michigan, 4.5 rating (20,076 reviews), 330K students, Intermediate level.
- Introduction to Data Science in Python** (COURSE) by University of Michigan, 4.5 rating (10,795 reviews), 260K students, Intermediate level.



### ■ 소프트웨어 도구

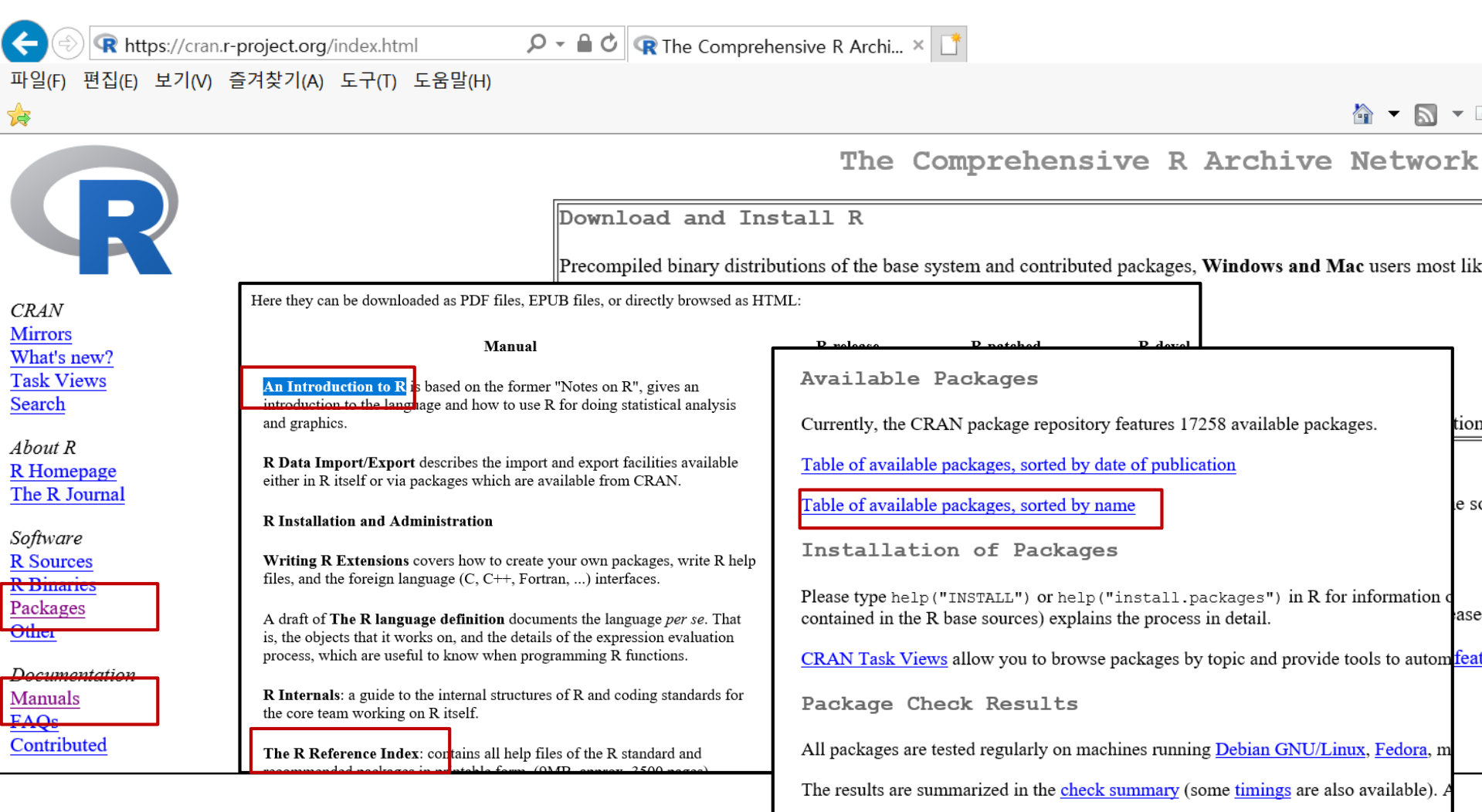
- SPSS, SAS, STATA 등의 통계 패키지, 엑셀 ← 프로그래밍 기능 약해서 요구사항에 딱 맞는 맞춤 처리 약함
- R과 파이썬 ← 무료. 프로그래밍 기능 강함. 오픈 커뮤니티를 통한 무수한 라이브러리. 배우기 쉬워 비전공자에 유리
- 이번 학기 분석 도구는 R을 사용 ← 데이터 과학도 배우고 프로그래밍도 배우고

### ■ R은 선형 및 비선형 모델링, 통계 테스트, 시계열 분석, 분류, 클러스터링 등 다양한 통계 및 그래픽 기술을 제공하는 통계 컴퓨팅 및 그래픽을위한 무료 언어 및 환경 (<https://cran.r-project.org/>)

### ■ R 공식 문서

- CRAN 사이트 (<https://cran.r-project.org>): 소프트웨어와 문서를 제공하는 R 공식 사이트
- 두 가지 중요 문서 : <An Introduction to R> <The R Reference Index>

## ■ R의 공식 문서 CRAN(The Comprehensive R Archive Network)



← → https://cran.r-project.org/index.html The Comprehensive R Archi... ×

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

### The Comprehensive R Archive Network

#### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

**Manual**

[An Introduction to R](#) is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.

**R Data Import/Export** describes the import and export facilities available either in R itself or via packages which are available from CRAN.

**R Installation and Administration**

**Writing R Extensions** covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.

A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.

**R Internals**: a guide to the internal structures of R and coding standards for the core team working on R itself.

**The R Reference Index**: contains all help files of the R standard and recommended packages in a stable form. (CIP: January 2000 pages)

**Available Packages**

Currently, the CRAN package repository features 17258 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

**Installation of Packages**

Please type `help("INSTALL")` or `help("install.packages")` in R for information contained in the R base sources) explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automate

**Package Check Results**

All packages are tested regularly on machines running [Debian GNU/Linux](#), [Fedora](#), m

The results are summarized in the [check summary](#) (some [timings](#) are also available). A

**CRAN**

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

**About R**

[R Homepage](#)

[The R Journal](#)

**Software**

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

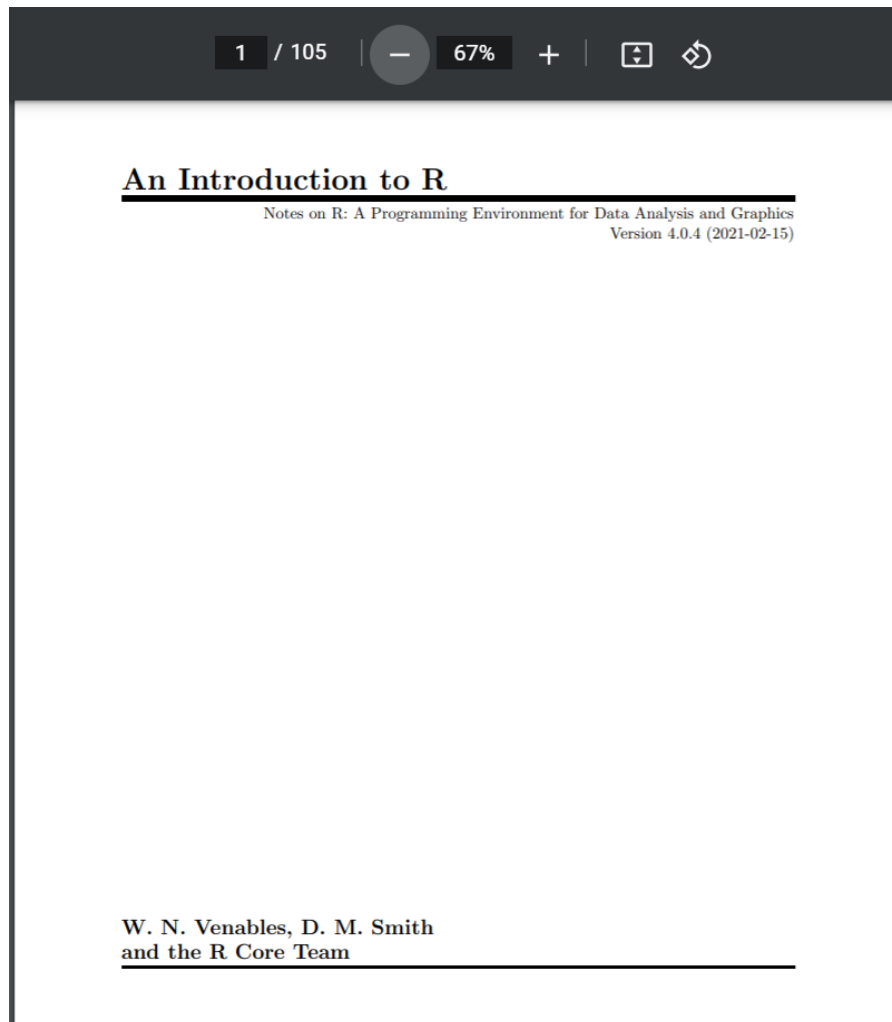
**Documentation**

[Manuals](#)

[FAQs](#)

[Contributed](#)

## ■ An Introduction to R



## R: A Language and Environment for Statistical Computing

### Reference Index

The R Core Team

Version 4.0.4 (2021-02-15)

Copyright (©) 1999–2012 R Foundation for Statistical Computing.  
Permission is granted to make and distribute verbatim copies of this manual provided the copyright notice and this permission notice are preserved on all copies.  
Permission is granted to copy and distribute modified versions of this manual under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.  
Permission is granted to copy and distribute translations of this manual into another language, under the above conditions for modified versions, except that this permission notice may be stated in a translation approved by the R Core Team.

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under the terms of the GNU General Public License. For more information about these matters, see <https://www.gnu.org/copyleft/gpl.html>.

### ■ 케플러의 행성 운동 법칙

- 브라헤 **천동설 가설하에** 데이터 분석 계속되는
- 케플러 천동설을 버리고 지동설 가설하에 데이터 분석 브라헤의 모든 데이터 설명

### ■ 존 스노의 콜레라 발생 지도

- 마이스론 이론 : 전염병 원인 공기(당시 믿음 강한 마이스론 이론)
- 존 스노 : 사망자 지도에 막대 길기로 표시 물 공급 펌프와 연관성 발견

### ■ 세종 대왕

- **70만호를 일일이 찾아 데이터 수집**
- 토지 비옥정도 6등급, 풍흉에 따른 9등급 조율 조정

1. 데이터 홍수 시대
2. 데이터 과학 열풍
3. 데이터 과학이란?
4. 데이터 과학의 절차
5. 데이터 과학 관련 분야 및 학습 범위
6. 데이터 과학 자원

# Thank you

