2. R 이해하기

2.1 R 환경설정

R

Why R ?

- 오픈 소스로 무료 (Open-Source) : FREE!
- 수많은 다양한 패키지(Packages) : R 공식 배포인 CRAN(open-source repository)을 통해 약14000개 (2019)
- Data Scientists가 가장 많이 사용하는 programming languages 중에 Python 과 더불어 최상위
- 학계나 과학자들을 통해 개발 Statistical 방법론에 특화 (예: 패널데이터 분석, 시계열 분석 등에 다양한 패키지)
- 다양한 커뮤티를 통해 지원

R. Community

- R-bloggers : https://www.r-bloggers.com
- The R Consortium : https://www.r-consortium.org/projects
- #rstats hashtag : https://twitter.com/search?q=%23rstats
- R-Ladies : https://rladies.org
- Local R meetup groups : https://jumpingrivers.github.io/meetingsR/r-user-groups.html
- Rweekly : https://rweekly.org
- DataCarpentry(http://www.datacarpentry.org) and Software Carpentry(https://software-carpentry.org)
- R Conferences : https://jumpingrivers.github.io/meetingsR/events.html
- Github : https://github.com/trending/developers/r?since=weekly

etc

R 환경설정

다운로드 R 설치 파일 (R 설치 파일 배포 : CRAN)

R 환경설정

R 설치 (OS: window)

R 설치완료 (실행)
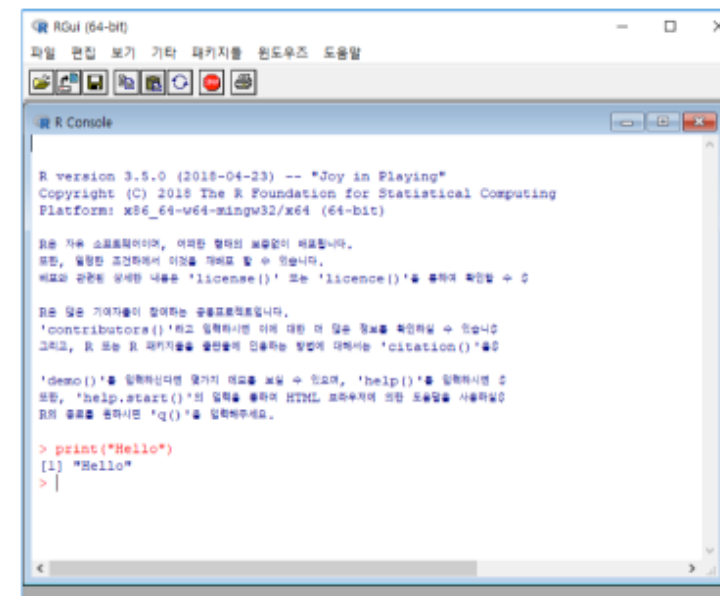
R 환경설정

통합 개발 환경IDE(Integrated Development Environment)

RStudio : https://www.rstudio.com/products/RStudio/

Visual Studio용 R : https://docs.microsoft.com/ko-kr/visualstudio/rtvs/installing-r-tools-for-visual-studio

jupyter notebook R kernel : https://irkernel.github.io/requirements/
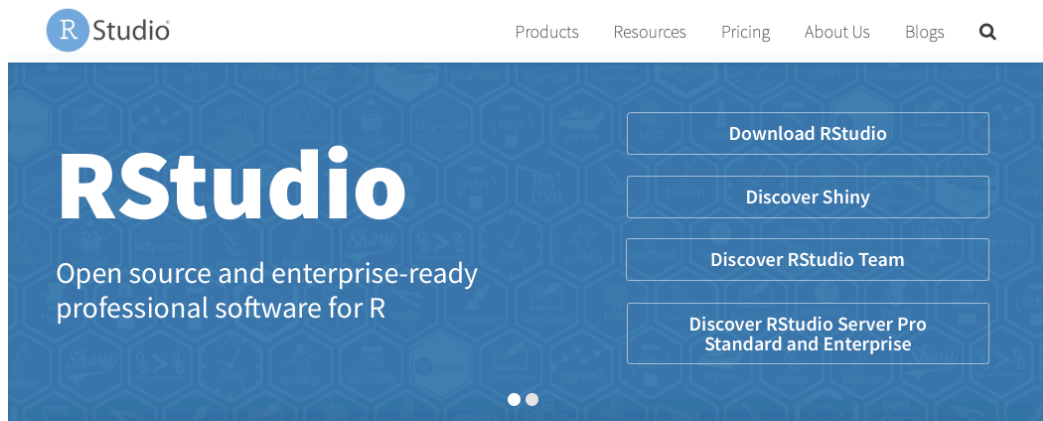
RStudio

R Tools for Visual Studio
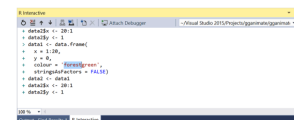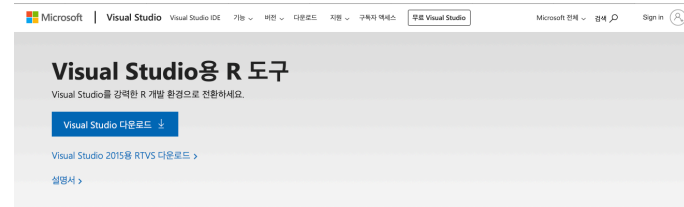
R Kernel for Jupyter Notebook
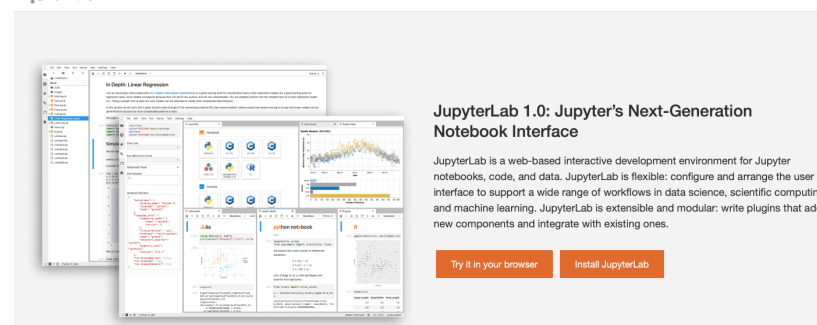
R 환경설정

RStudio 설치 (IDE)
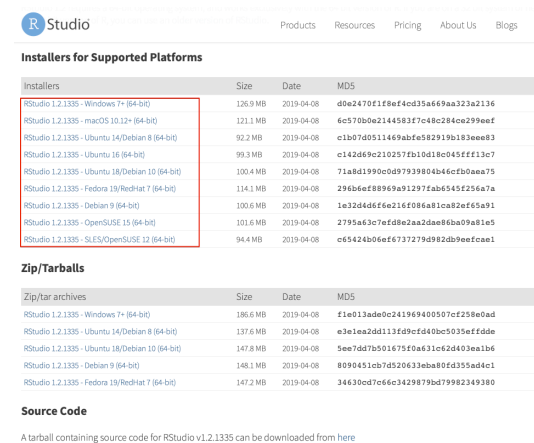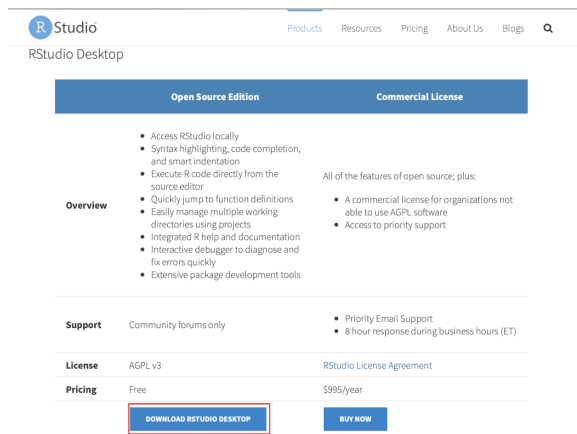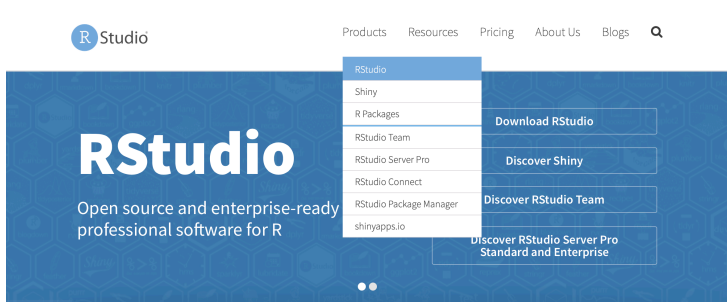
RStudio
관리자 실행으로 설정

RStudio 실행

2.2 R 데이터 타입

R 기본적 이해

데이터 타입 이해
- 수치형
- 문자형
- 논리형

데이터 구조이해
- 벡터(vector)
- 행렬(matrix)
- 데이터프레임(dataframe)
- 배열(Array)
- 리스트(list)

선언문 (Statement)
- 조건문(if)
- 반복문(loop)
- 함수(function)
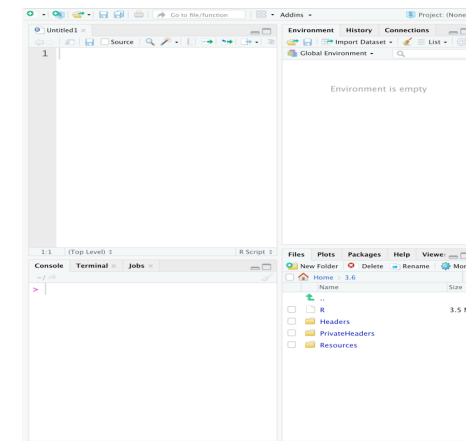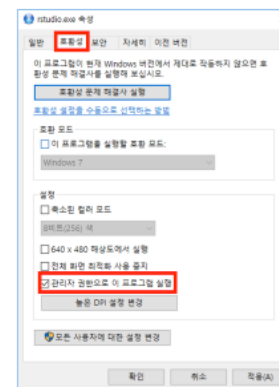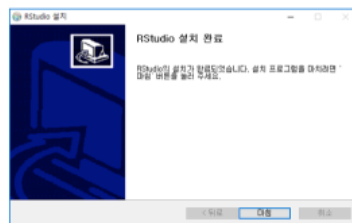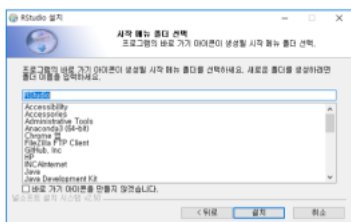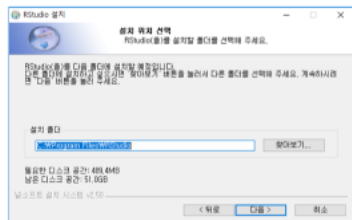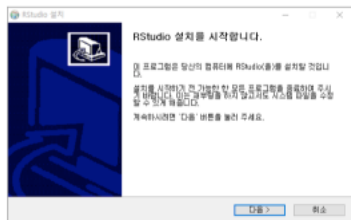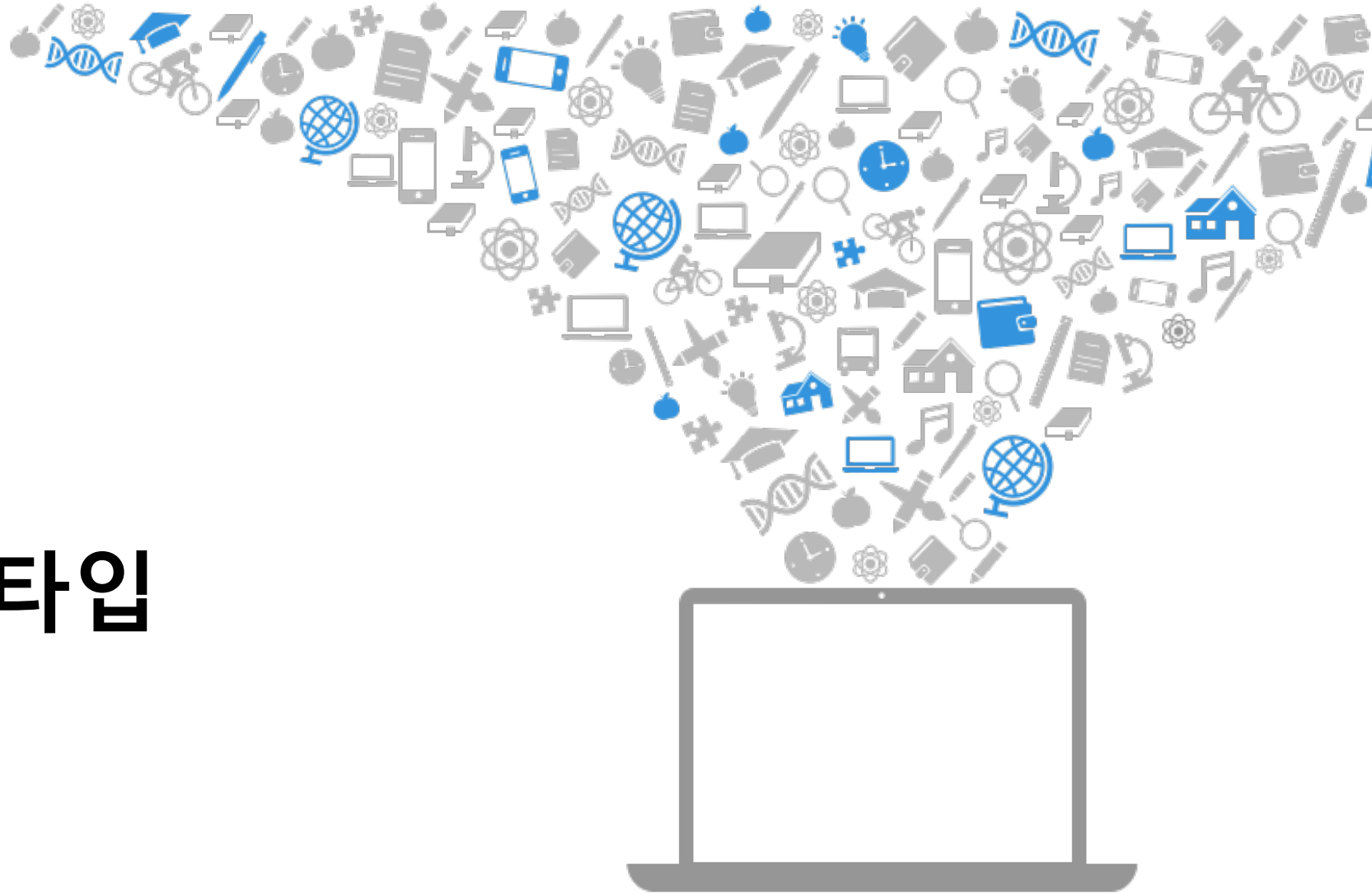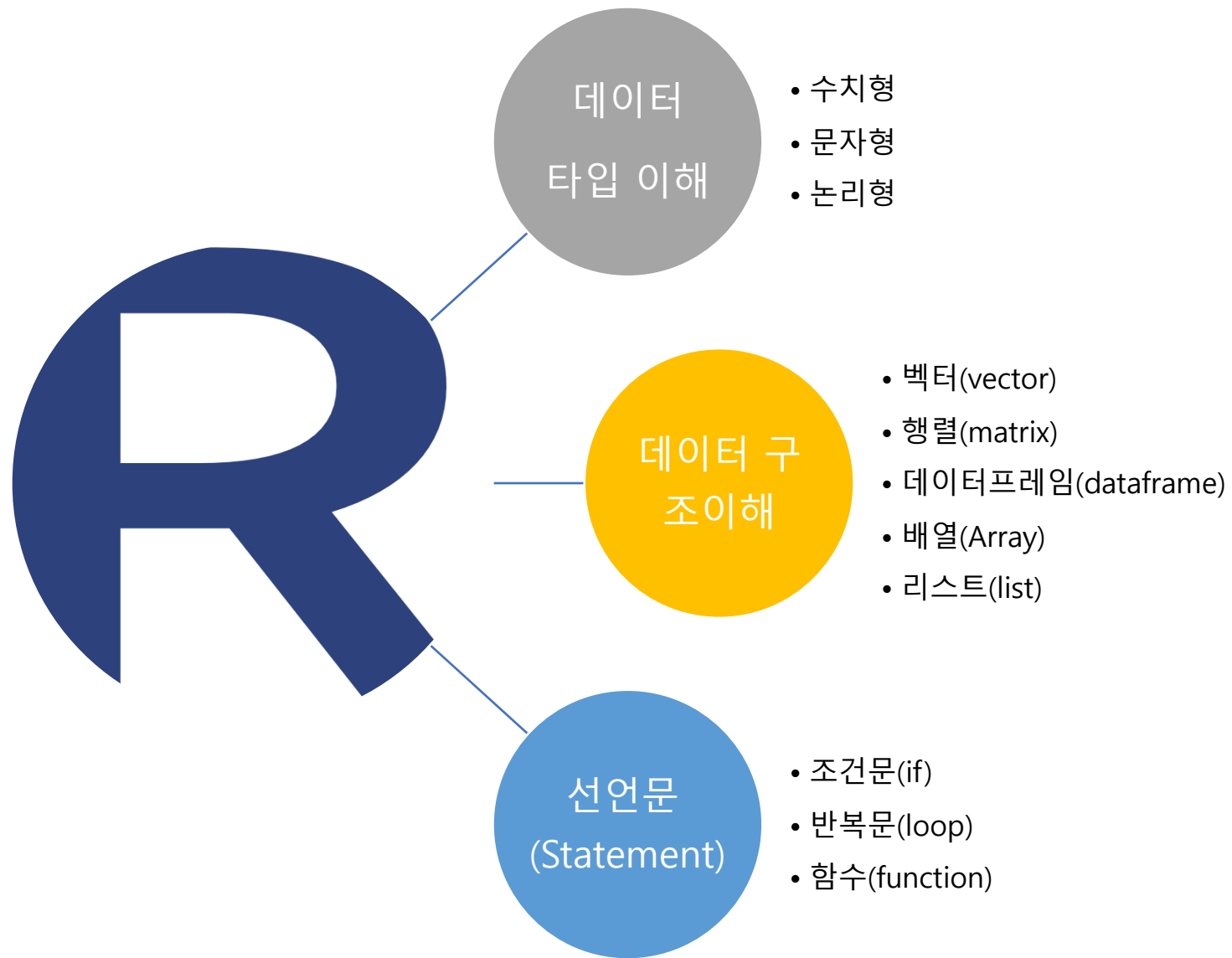
Ⅰ R 데이터 타입

변수, 객체(object) : 할당하여 생성

Rule of R language
- 첫 글자 반드시 문자
- 대소문자 구분
- 할당연산자 : <-, ->, =, 변수를 생성할 때 사용

자료구조

수치형
- 실수(numeric) : 1.3
- 정수(integer) : 3, 3L,
- 복소수(complex) : 2i

문자형
- 문자열(character) : "string"
- 범주형(factor) : levels(male, female)
- 날짜(date) : "2019-09-23", format ('y-m-d')
- 시간(time) : format = (h:m:s)

논리값 : TRUE, FALSE, T, F

산술연산자 : + , -, *, /, %/%, %%, ^
논리연산자 : ==, !=, >=, <=, >, <
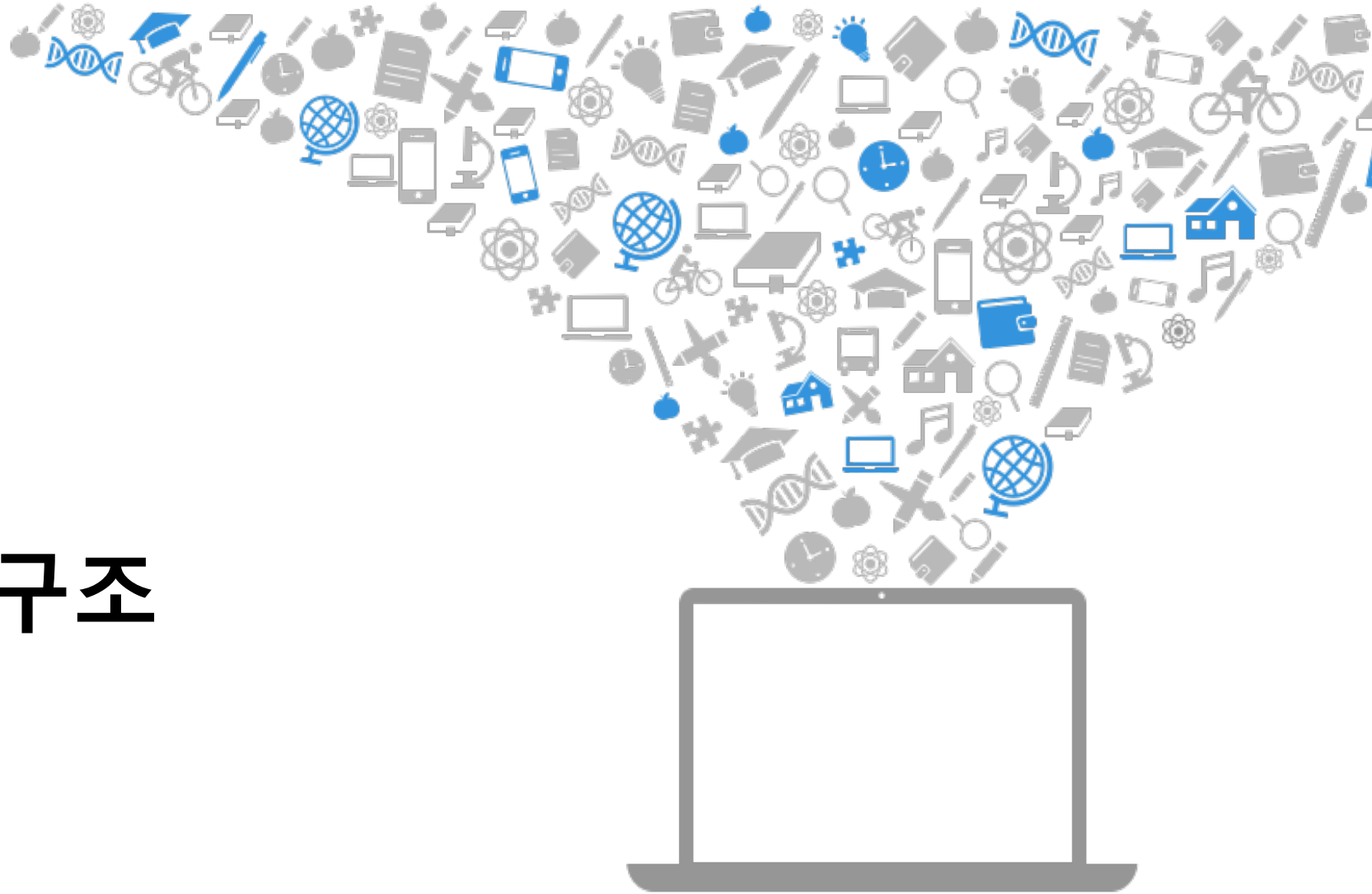
2.3 R 데이터 구조

R 데이터 구조

스칼라(Scala)

벡터(vector)
- 생성
- indexing(slicing)
- 연산

행렬(Matrix)
- 생성
- indexing(slicing)
- 연산

배열(Array)
- 생성
- indexing(slicing)
- 연산

데이터 프레임(data frame)
- 생성
- indexing(slicing)
- 연산

리스트(list)
- 생성
- indexing(slicing)
- 연산

Vector

Matrix

rows

columns

Array

Data Frame
(Table)

Lists

source: http://venus.ifca.unican.es/Rintro/dataStruct.html

2.4
조건문(If),
반복문(loops),
함수(function)

2.5 시각화 (Visualization)

R 주요 패키지 예시)

20 BEST LIBRARIES FOR DATA SCIENCE IN R

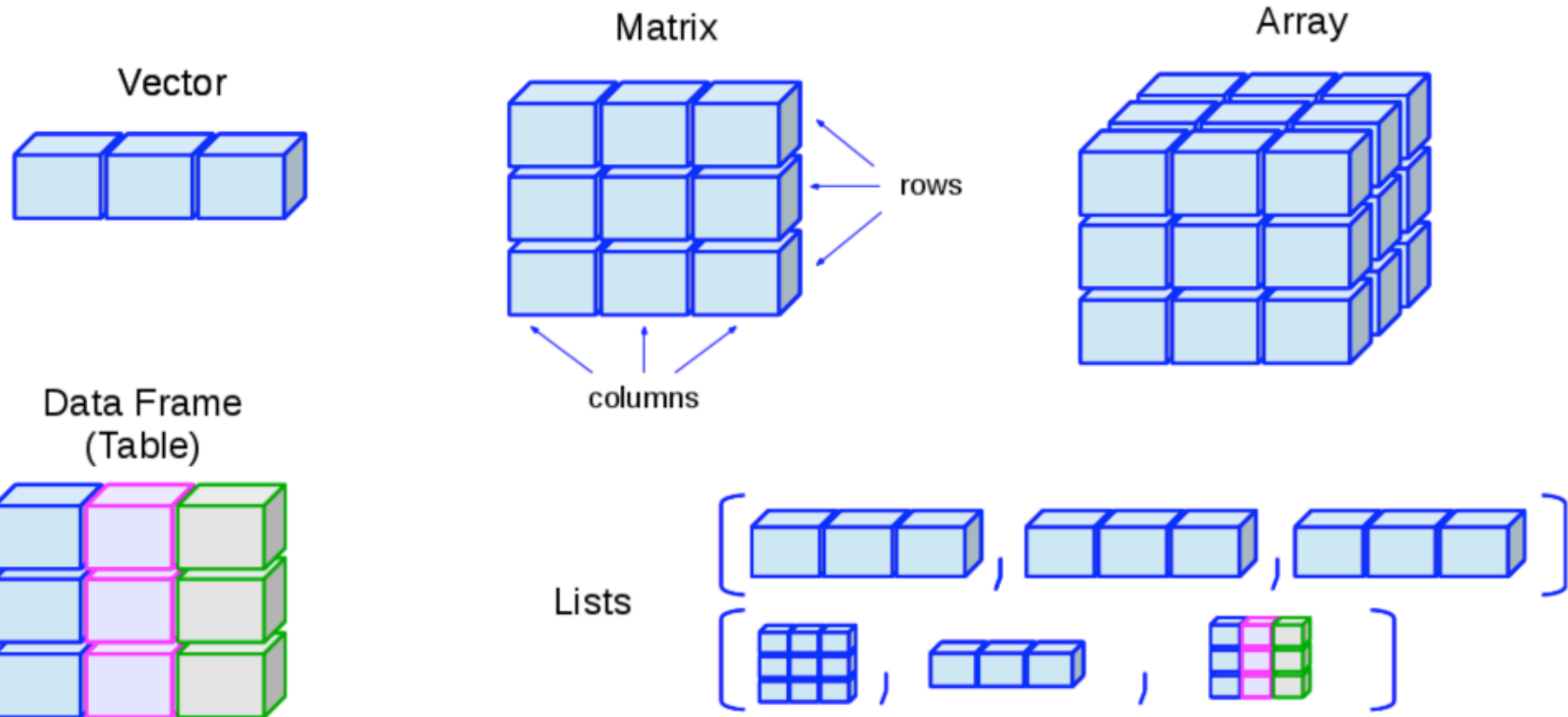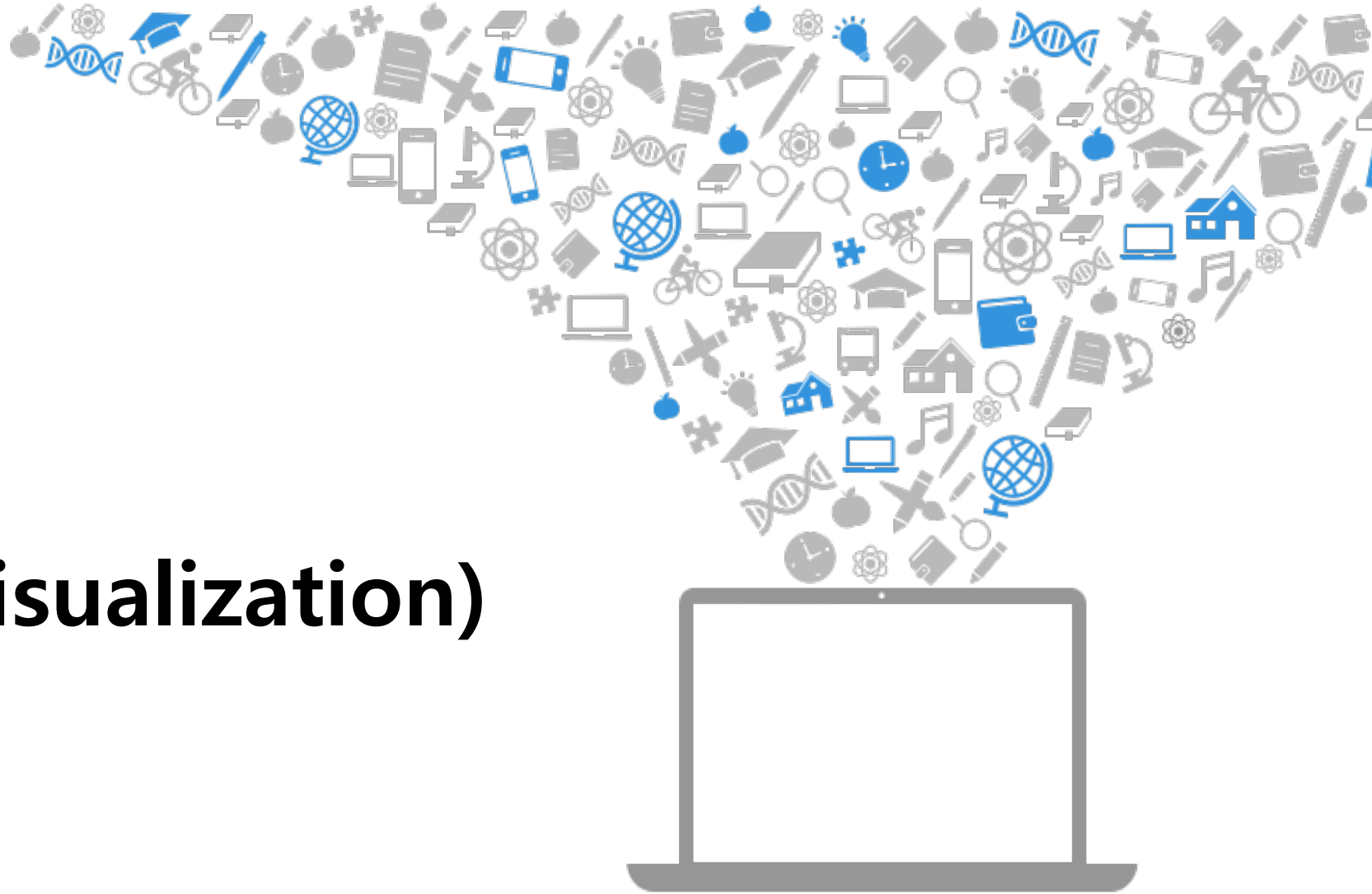|		COMMITS	CONTRIBUTORS	FEATURES
DATA MANIPULATION	dplyr	4 354	136	• powerful library for data wrangling • works with local data frames and remote database tables • precise and simple command syntax
	data.table	3 211	43	• quick aggregation of large data • laconic flexible syntax and a wide suite of useful functions • friendly file reader and parallel file writer
	lubridate	1 427	45	• a set of functions to work with date and time format • easy and fast parsing of date-time data • expanded mathematical operations on time data
	jsonlite	908	11	• robust and quick parsing JSON objects in R • great tool for interacting with web APIs and building pipelines • functions to stream, validate, and prettify JSON data
GRAPHIC DISPLAYS	ggplot2	3 903	133	• powerful implementation of the grammar of graphics visualization • developed static graphics system • takes care of plot specifications
	Corrplot	299	8	• abilities to visualize correlation matrices and confidence intervals • contains algorithms to do matrix reordering • flexible appearance details settings
	lattice	132	0	• high-level visualization system • emphasis on multivariate data • efficiently copes with nonstandard requirements
HTML WIDGETS	plotly	2 989	26	• rich features and plenty of available charts • web-based toolbox for building visualizations • abilities to make ggplot2 graphics interactive
	ggvis	2 159	21	• implementation of an interactive grammar of graphic • incorporates shiny reactive programming model and dplyr grammar of data transformation
	DT DataTables	1 919	21	• displays R matrices and data frames as interactive HTML tables • creates sortable tables with a minimum of code • many useful features and styling options for tables
	rCharts	638	11	• interactive JS charts from R • tools for creation, customization, and sharing

20 BEST LIBRARIES FOR DATA SCIENCE IN R

		COMMITS	CONTRIBUTORS	FEATURES
REPRODUCIBLE RESEARCH	knitr	5 467	96	• transparent tool for easy dynamic report generation in R • enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents
	rmarkdown	2 297	56	• next generation implementation of R Markdown based on pandoc • many static and dynamic output formats • abilities to define new formats for custom publishing requirements
	Slidify	302	7	• generates reproducible html5 slides from r markdown • allows embedded code chunks and mathematical formulas • rich sharing and customizing opportunities
MACHINE LEARNING	mlr	3 915	55	• extensible framework for classification, regression, survival analysis, and clustering • easy extension mechanism through S3 inheritance
	dmlc XGBoost	3 188	259	• implementation of the Gradient Boosted Decision Trees algorithm • reach tools for regression, classification, and ranking problems • high speed and performance
	caret	1 659	59	• many models for classification and regression • powerful tools and algorithms for creating predictive models
	gbm	731	26	• represents Generalized Boosted Regression Models • includes plenty of regression methods • tools variable selection and final stage precision modeling
	Prophet	190	20	• high-quality forecasts for time series data • manages data that has multiple seasonality with linear or non-linear growth • robust to missing data, shifts in the trend, and large outliers
	randomForest	56	0	• implements Breiman's random forest algorithm for classification and regression • builds multiple decision trees and gives back the mean prediction of the individual trees

Updated: December 2017