

Wk 2. R studio, Big Data

CH 1, CH2

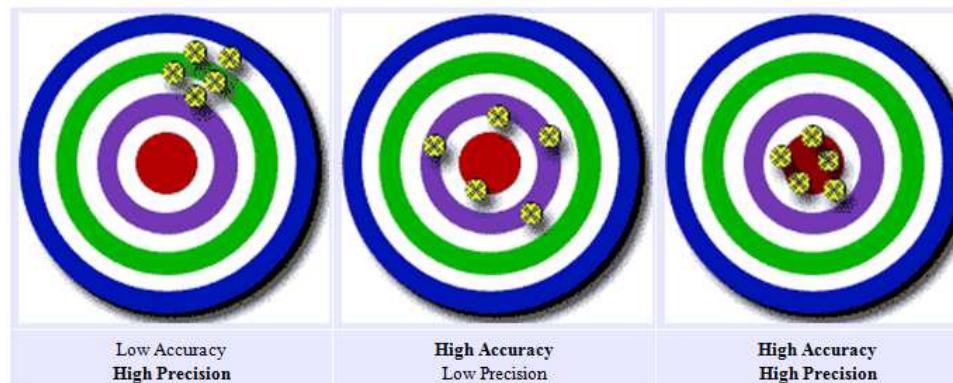


Terminology 1

Data
Science²

• Accuracy vs. Precision

- ▶ **Accuracy** is how close a measured value is to the **actual (true) value**.
=> Average ("Accuracy" in Korean?)
- ▶ **Precision** is how close the measured values are **to each other**.
=> Variance ("Precision" in Korean?)



Calibration?

<http://www.mathsisfun.com/accuracy-precision.html>

Terminology 1 (2)



- Accuracy VS. Precision

- ▶ eg 1. Thermometer:

Device 1: 37.**34** °C

Device 2: 37 °C

Q. Which one is more expensive? Do you need an expensive one?

Q. Is the fraction ".34" meaningful?

- ▶ eg 2. CO2 Sensor: $\pm 30 \text{ ppm} \pm 3\%$

$$400 - (30 + (400 * 0.03)) = 358 \text{ ppm} \quad \sim 400 + (30 + (400 * 0.03)) = 442 \text{ ppm}$$

(<https://www.co2meter.com/blogs/news/170700807-co2-measurement-range-why-it-matters>)

- ▶ eg 3. Speed Gun: $\pm 10\%$ ($40 \text{ km/hr} \Rightarrow 36 \text{ km/hr} \sim 44 \text{ km/hr}$)

Terminology 1 (3)



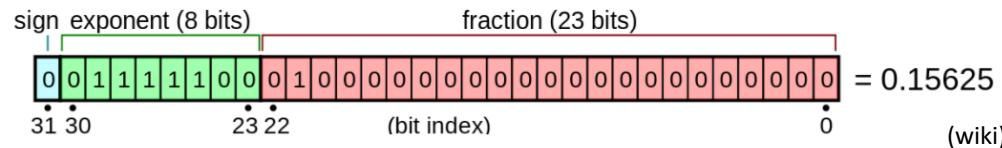
- Digital Data Precision ?

- ▶ Integer:

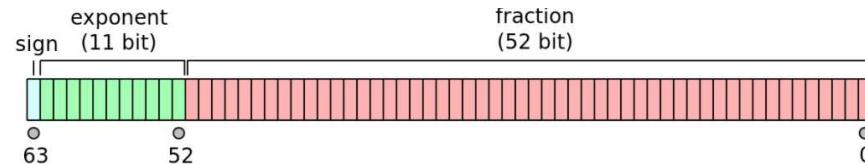
- ▶ Largest integer in R: 2,147,483,647 (How many bits?)

- ▶ Floating Point Number:

- ▶ Single Precision FP Number: 32bits



- ▶ Double Precision FP Number: 64bits



- ▶ Maximum Double number in R: 1.797693e+308

Compilation vs. Interpretation

COMPILATION

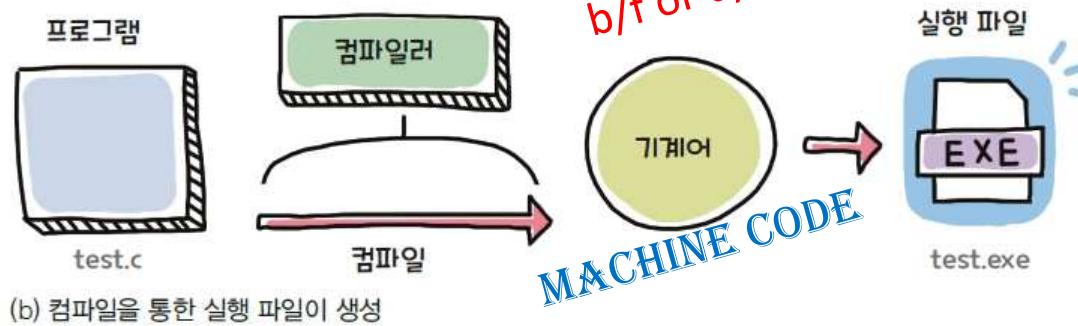


그림 1-7 프로그램이 만들어지는 과정

```
Include <stdio.h>
Int main() {
    printf("Hello World");
}
```

_____ code?

_____ readable

000110 0101010 0011100 1101101
000101 0101001 0011111 1101110
000011 0101111 0011001 1101000
001111 0100011 0010101 1100100
010111 0111011 0001101 1111100

■) 기계어의 예

_____ readable

Compilation vs. Interpretation



INTERPRETATION

- ▶ **Interpretation:** Machine Code generation is postponed until execution time!
- ▶ Compilation based languages: _____
- ▶ Interpretation based languages: _____
- ▶ Pros and Cons of Interpretation ?

	Complied Lang	Interpreted Lang
Performance (Speed)		
Maintenance / Upgrade		
Platform Independence		

Installing R & R-Studio

- Installing R: <https://www.r-project.org/>
- Installing R-Studio: <https://www.rstudio.com/>

Data
Science
⁷



[Home]

[Download](#)

[CRAN](#)

[R Project](#)

[About R](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.



a/b R

LET'S FIND OUT WHAT **R** LANGUAGE IS!



■ Characteristics of R

- R is a relatively **new** programming language
- R was developed by Ross Ihaka and Robert Gentleman of Auckland University, New Zealand in 1993 as a Statistical Programming languages called “**S-PLUS**” (Free Version)



CHARACTERISTICS OF R



■ Specialized for Data Analysis

- R is a language developed for data analysis & statistics.
- R can be run immediately without compilation to see the results.
- What you write in R is called a “**script**”, not a “**program**”.

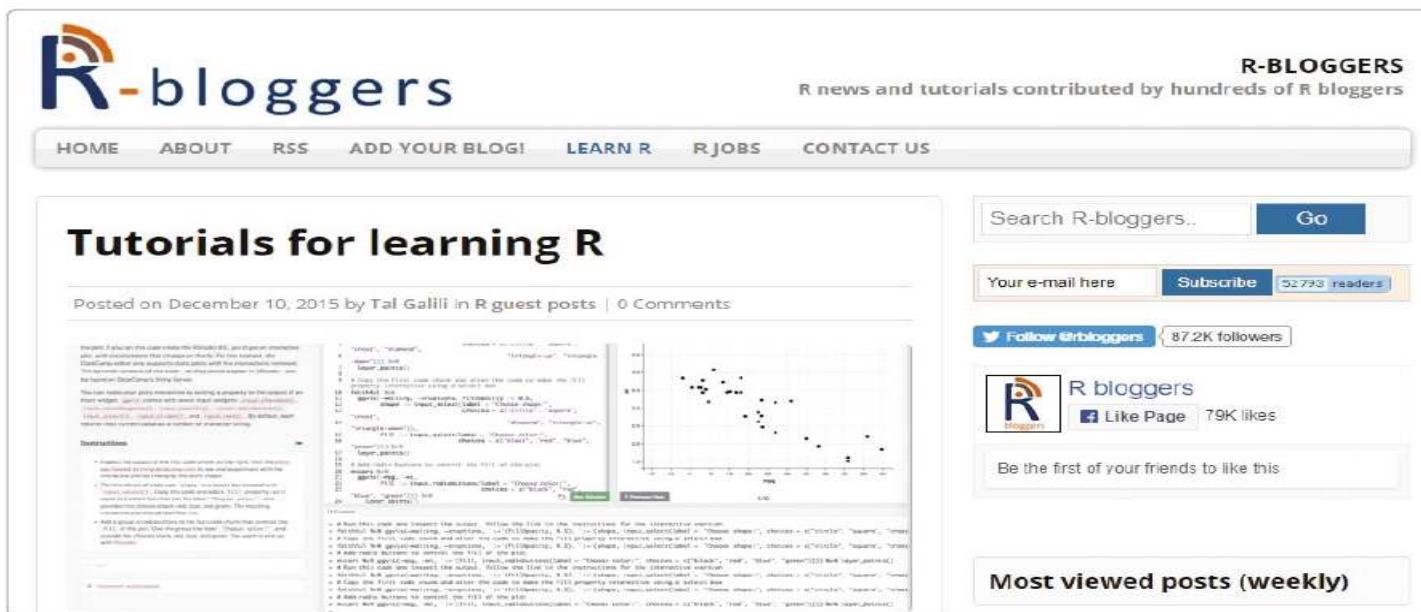
■ A strong user community

- R has a strong user base, so there are a variety of communities.
- There are also plenty of learning materials for beginners.
- The number of Korean-language online materials / sites are increasing too.

R HAS STRONG USER COMMUNITIES!

Data
Science
11

■ “R-bloggers” Community



The screenshot shows the homepage of R-bloggers. At the top, there's a navigation bar with links for HOME, ABOUT, RSS, ADD YOUR BLOG!, LEARN R, R JOBS, and CONTACT US. The main title "R-bloggers" is displayed with a blue "R" icon. To the right, it says "R-BLOGGERS" and "R news and tutorials contributed by hundreds of R bloggers". Below the navigation, there's a search bar with "Search R-bloggers.." and a "Go" button. A sidebar on the right includes a "Follow @rbloggers" button with "87.2K followers", a "Like Page" button with "79K likes", and a link to "Be the first of your friends to like this". The central content area features a section titled "Tutorials for learning R" with a sub-section "How to make an interactive plot using rCharts". It includes a screenshot of an R script and a scatter plot generated by the code. Below this, there's a "Most viewed posts (weekly)" section.

그림 1-9 대표적 R 커뮤니티 R-bloggers(<https://www.r-bloggers.com/>)

CHARACTERISTICS OF R

Data
Science 12

■ Variety of Packages

- R bundles the functions used for data analysis into **packages**
- Almost all the features for **Data Analysis!**
- When the **latest theory** is published, an R package is created immediately, so it is possible to **quickly** utilize the latest theory for data analysis

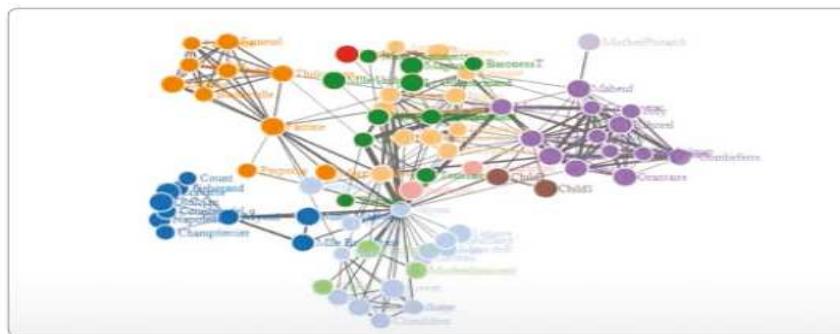


그림 1-10 networkD3 패키지 : 네트워크 형태의 데이터를 시각화하는 데 사용됩니다.

CHARACTERISTICS OF R

■ Providing Aesthetic and Functional statistical graphs

- Visual Representation of results is critical in Data Analysis.
- R makes it easy to create beautiful and functional graphs with a package called “ggplot”.

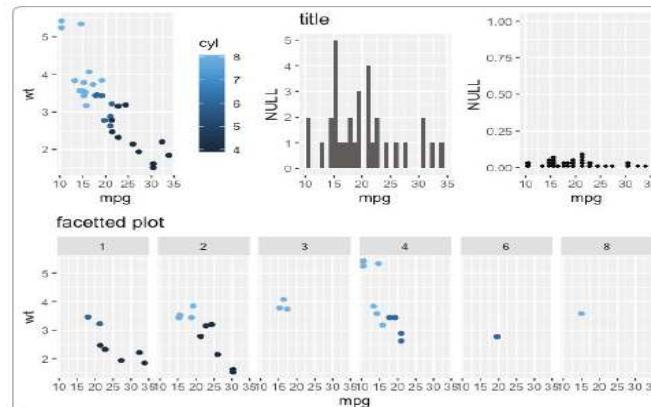


그림 1-11 ggplot 그래프의 예

■ User-friendly Programming Environment

- **R-Studio**, integrated development environment (IDE) is provided to handle all tasks of R programming!

CHARACTERISTICS OF R

R-Studio

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays R code for generating datasets and visualizations.
- Console:** Shows the output of the R code, including salary data and a scatter plot.
- Environment:** Shows the global environment with objects like `dat_raw`, `iris`, `Rdat`, `Rtable`, `values`, and `salary`.
- Plots:** Displays a scatter plot of `iris$PetalLength` versus `Index`.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
analytic.Rmd | adj_ifield.R | experiment.R |Untitled | dataset_info.R | 30
Source on Save Run Source
1 salary <- c(400,350,500,700,600)
2 salary
3 salary - 100
4 salary + 50
5 beaver1
6 beaver2
7 CO2
8 co2
9 euro
10
11 accident <- c(31.26,42.47,50.54,70.66,43.32,32.22)
12 names(accident) <- c("M1","M2","M3","M4","M5","M6","M7","M8",
13 "M9","M10","M11","M12")
14
15 accident
16 sum(accident)
17 accident/0.9
18
19 library(ggplot2)
20 ggplot(iris,aes(x=Index,y=IrisPetalLength)) + geom_point()
21
22 (Top Level) <-
>
FRF   ATS    BEF    DEM    ESP    FIM
13.760300 40.339900 1.955830 166.386000 5.945730 6.5
59570 IEP   ITL    LUF    NLG    PTE
0.787564 1936.270000 40.339900 2.203710 200.482000
>

```

그림 1-12 R 스튜디오 실행 화면

CHARACTERISTICS OF R

Data
Science
16

■ Free!

- R is a **free** open source software!
- Regular updates are made one or two times a year to continuously improve functionality.
- R can be installed and used not only in Windows environments, but also in Linux and MacOS environments.

02. Why learn R?



- **The 4th Industrial Revolution :**

Newly emerged technologies such as AI, Big Data, Block Chain, Robotics, and the IoT hugely impact the entire political, economic, and cultural aspects of our lives.

- The core of the 4th Industrial Revolution is **DATA!**
- Companies and Individuals who manage/control DATA will be the winner!
- **Computational Thinking** is crucial regardless of majors.

CHARACTERISTICS OF R

■ Why Learn R?

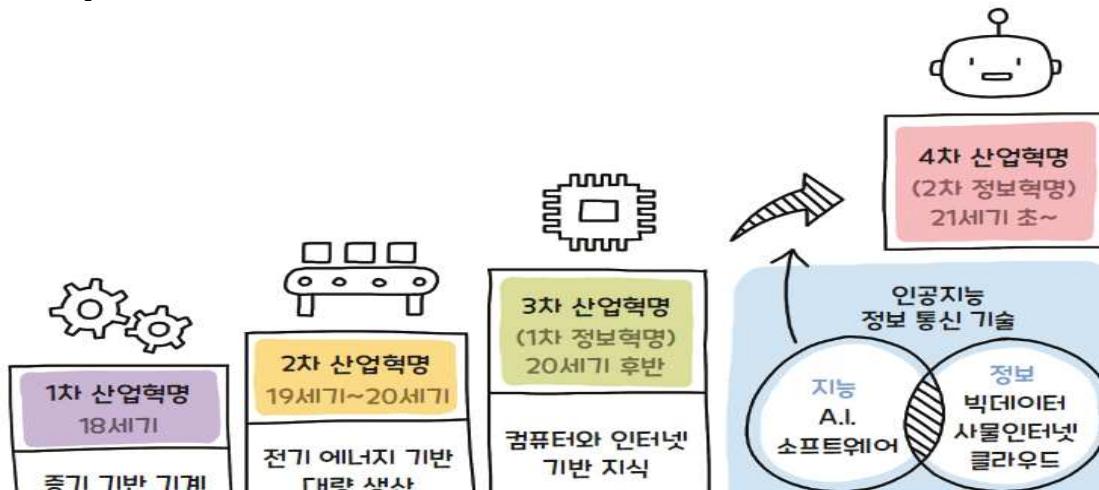


그림 1-13 산업혁명의 발전과 핵심 기술

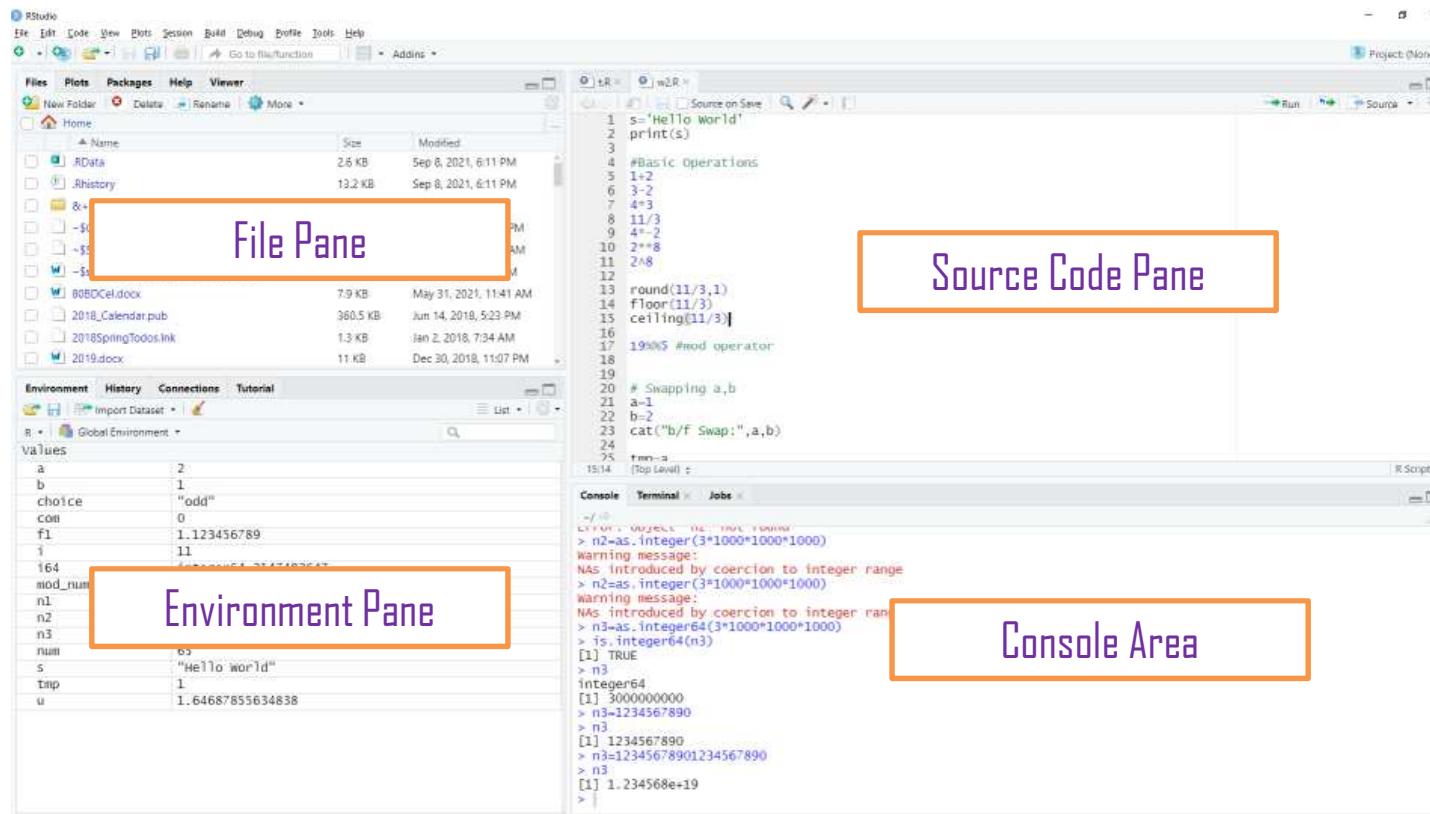
- R provides easy-to-learn yet powerful **data processing** and **data analysis** capabilities.
- Suitable for learning **computing thinking** by programming .

01

R-Studio Menu & the Screen Layout

R-STUDIO MENU & THE SCREEN LAYOUT

Data
Science
20



02

Let's try simple operations

ARITHMETIC OPERATIONS



[Code 2-4]

```
2+3  
(3+6)*8  
2^3          # 2 to the power of 3
```

[Result]

```
> 2+3  
[1] 5  
> (3+6)*8  
[1] 72  
> 2^3          # 2 to the power of 3  
[1] 8
```

ARITHMETIC OPERATORS

Operators

표 2-2 R에서 자주 사용하는 산술연산자

연산자	의미	사용 예
+	덧셈	$3+5+8$
-	뺄셈	$9-3$
*	곱셈	$7*5$
/	나눗셈	$8/3$
%%	나눗셈의 나머지	$8\%3$
^	제곱	2^3

COMMENTING IN R

■ Comments

[Code 2-5]

```
7+4  
# 2^3
```

```
> 7+4  
[1] 11  
> # 2^3  
>
```

* Try **Ctrl + Shift + C** !

03

R Packages

R PACKAGES

■ Packages:

- A Package is the collection of related functions for series of Analysis

공구함



미술도구함



조리도구함



그림 2-5 패키지의 개념

■ B/f using a Package

- **Loading** is required b/f using packages
- Packages should be installed in a fold of your computer
- If there is none, download and install it first

INSTALLING / USING PACKAGES

■ Install and Use

- ① **Install** the right package for the functions you want to use
- ② **Load** the package

- The package needs to be installed only once, but the package load is required each time R Studio starts.



PACKAGE

■ Package Installation and Usage

- Install a package using command prompt
 - `install.packages()`

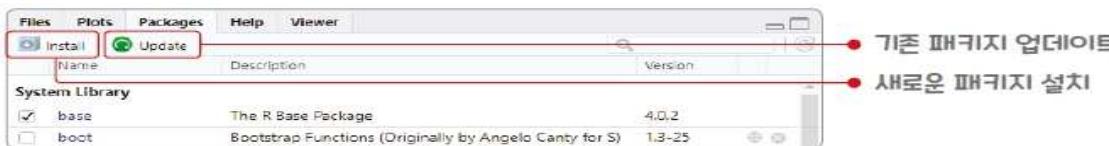
[Code 2-7]

```
# ggplot2 package
install.packages('ggplot2')
```

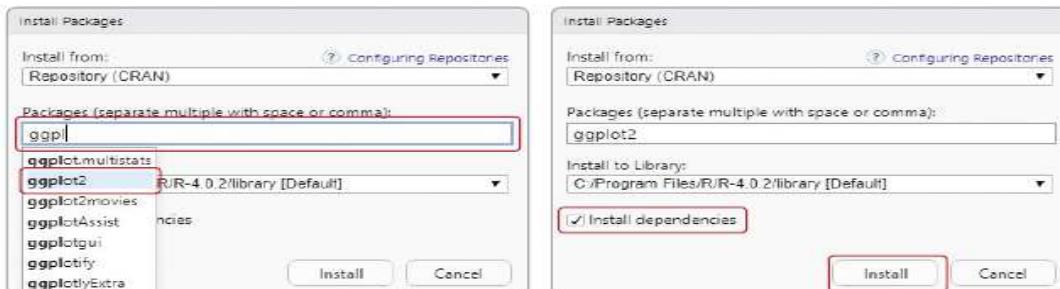
```
# Load and use
library(ggplot2)
ggplot(data=iris,aes(x=Petal.Length,y=Petal.Width))+geom_point()
```

PACKAGE INSTALLATION BY MENU

- Package Installation by Using R-studio



(a) 패키지창 화면



(b) 패키지 설치 화면

그림 2-7 R 스튜디오 메뉴로 패키지 설치하기

Today's Proverb

- “Give me **six hours** to chop down a tree and I will use the **first four** sharpening the axe”

(Abraham Lincoln)



Gmshtcdat alwstffsta

<https://wise4edu.com/init/#s4>

Programming 101: Truth Table

Data
Science

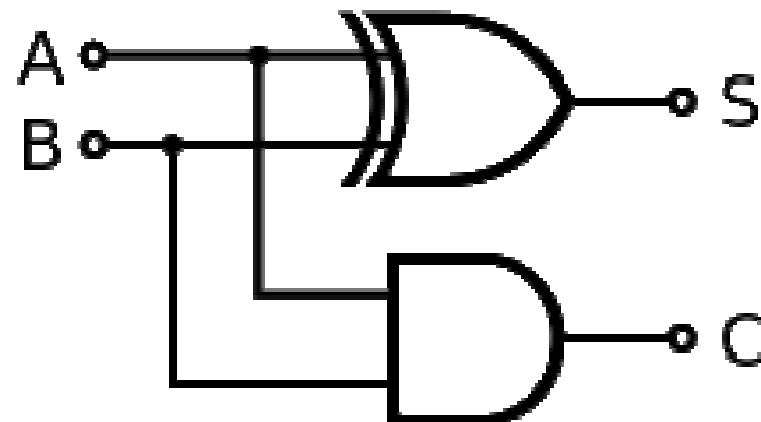
32

Q. Can you make a linear model for XOR ?



* <https://deepdatascience.wordpress.com/2017/10/25/linear-regression-vs-and-or-xor-logic/>

Adder



- [https://en.wikipedia.org/wiki/Adder_\(electronics\)](https://en.wikipedia.org/wiki/Adder_(electronics))

Programming 101: Truth Table (2)

- p: “Today is Thursday” => T/F?
- q: “Sam is handsome” => T/F?
- p: “Sam is a bike rider” => T
- q: “Sam is the president of Korea” => F
- $\sim p$?
- $r = p \text{ AND } q$ vs. $r = p \text{ OR } q$

Programming 101: Truth Table (3)

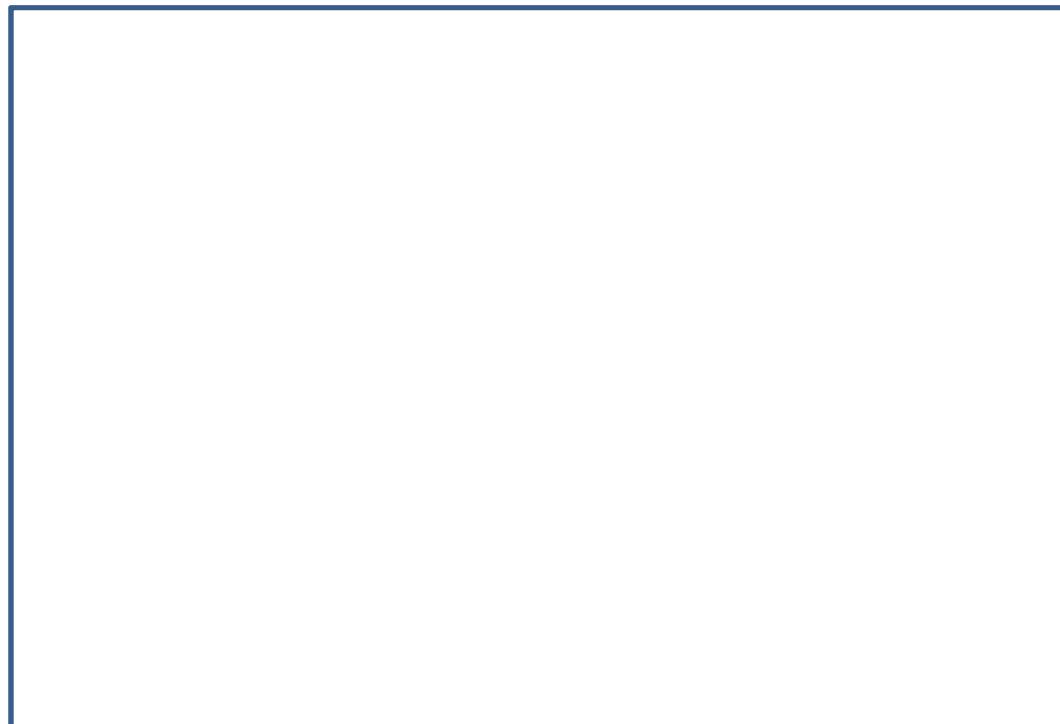
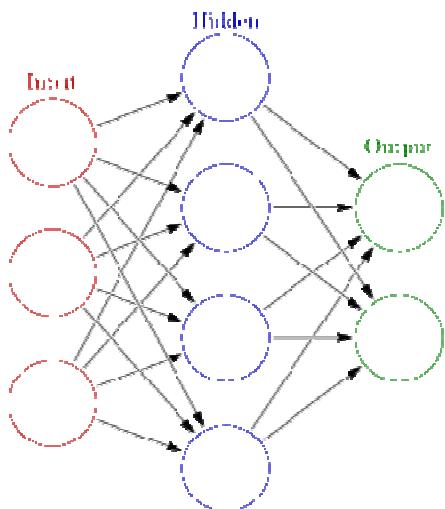
- Quantifier:

- p: “For all integer n, $n > 0$ ” $\Rightarrow T/F?$
- $\sim p$?
- w/ for some?

- Conditional Proposition:

- If p, then q ($p \rightarrow q$) eg: _____
- Truth table of $p \rightarrow q$ (Advanced)

Q. What a/b ANN for XOR?



* <https://stackoverflow.com/questions/41712420/can-an-ann-of-2-neurons-solve-xor>

Big Data: Definition

- “an accumulation of data that is too large and complex for processing by traditional database management tools”
(merriam-webster.com)
- Various Types: Photos, Audios, Videos, SNS, GPS, IoT, Shopping list, and even Clicks on web browsers (probably soon where you look at)

Big Data: Characteristics

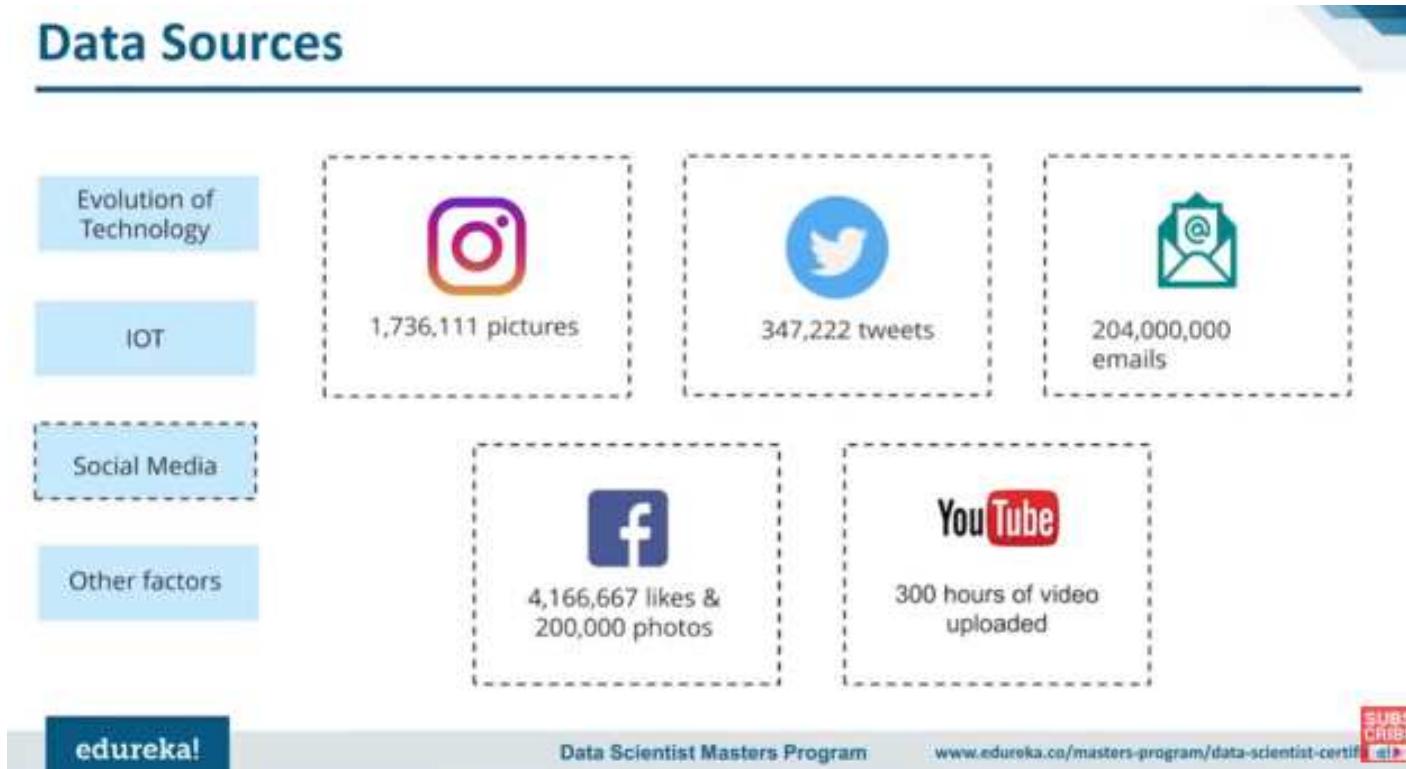
Data
Science
39

- **3Vs (+ more Vs):**
 1. **Volume:** Giga-bytes ~ Zeta-bytes
 2. **Velocity:** 5G, 6G~
 3. **Variety:** texts, transactions, logins , videos and so on
 4. **Veracity:** quality of data
 5. **Value:** worth of processed data

.... **What about Privacy?**

Big Data generation per minute

Data Sources



Big Data: IoT



* Pixabay.com

2021 Fall Data Science, Sang Jin Han

Big Data: Applications



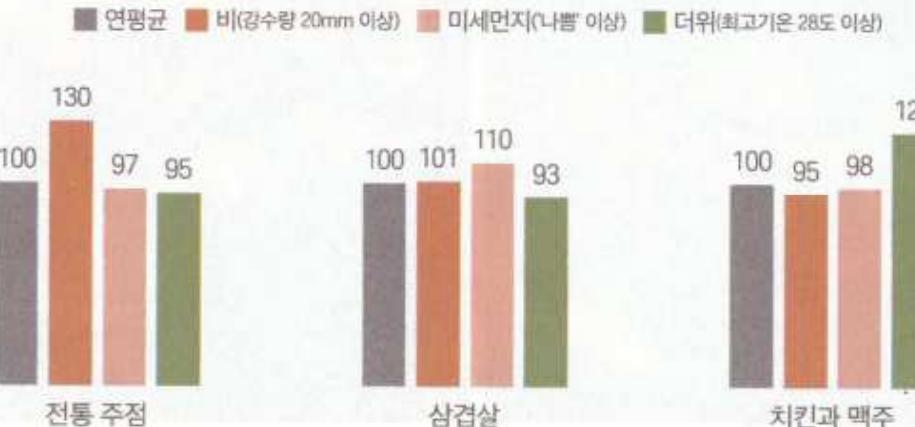
- Government
 - Business
 - Healthcare, Science
 - Education
 - Technology, Military
 - Culture: Media, Entertainment, Sports, Arts, ..
 - International Development
- and indeed all areas, fields, and aspects of lives

Big Data: Cases / Stories (1)

- **Netflix:** Data Collected from over 200M Subscribers to deliver customized Streaming service beating giants like Disney 
- **Target:** If a woman starts buying unscented lotion, they identifies her as “Pregnant Customer” 

Big Data: Cases / Stories (1)

업종별 날씨에 따른 매출 지수(2019년 기준)



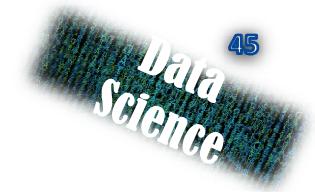
출처: BC카드

여름에는 “**치맥지수**”, 미세먼지 많은 날에는 “**삼겹살지수**”

* 빅데이터 생활을 바꾸다 (BC카드 빅데이터 센터)

Next Week Topics

- Variable, Vector, Function
- Open Data



To-dos for Next Class (Flipped Learning)



- **Try enough R codes!**
(Think a/b how to become a **good chef**)
 - Self-Game project #1: “**RSP**” Game
- **Textbook**
 - Read ch3

Any Question?



Thank you for your attention!

