



AN  
INTRODUCTION  
TO  
**MACHINE  
LEARNING**  
WITH **R**

DAY 3

The background is a dark blue landscape at night, featuring silhouettes of mountains against a lighter sky. A bright, star-like light source is visible on the horizon, casting a glow over the water in the foreground.

DAY 3

# Data Visualization



데이터  
시각화



우리는 **정보 과부하나 자료 과다**로부터  
비롯된 모든 괴로움에 있는 것을 느낀다.  
좋은 소식은 그것에 대한 쉬운 해결이  
있을지도 모른다는 것이고,  
그것은 **우리의 눈을 더 사용하는 것**이다.

— David McCandless  
Data Journalist



[데이터 시각화]

DATA VISUALIZATION

데이터 분석 결과를 **쉽게 이해**할 수 있도록 **시각적으로 표현하고 전달**하는 과정

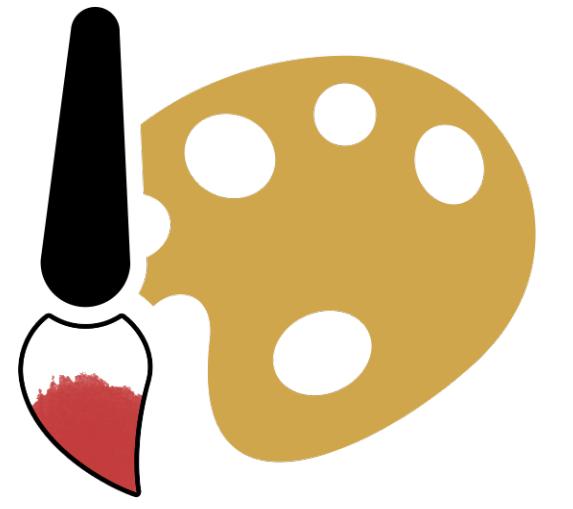


[데이터 시각화]

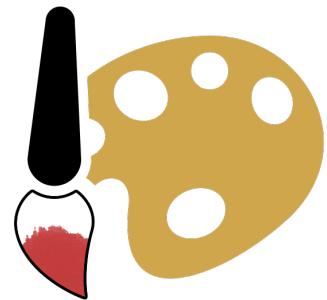
DATA VISUALIZATION

데이터 분석 결과를 **쉽게 이해**할 수 있도록 **시각적**으로 **표현하고 전달**하는 과정

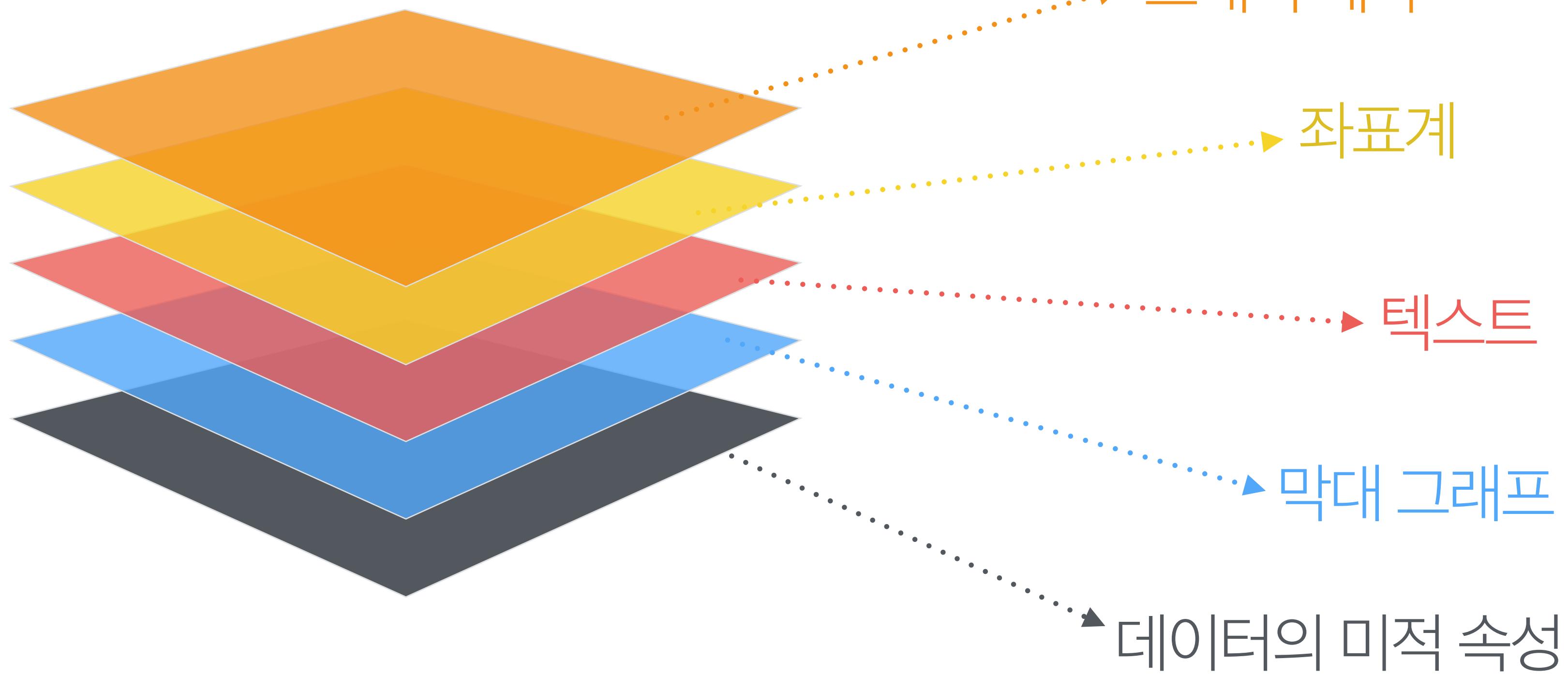
예측 모델 구축 과정에서 간과할 수 없는 부분

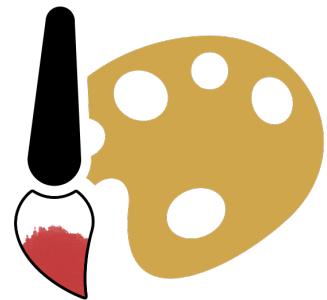


**ggplot2**

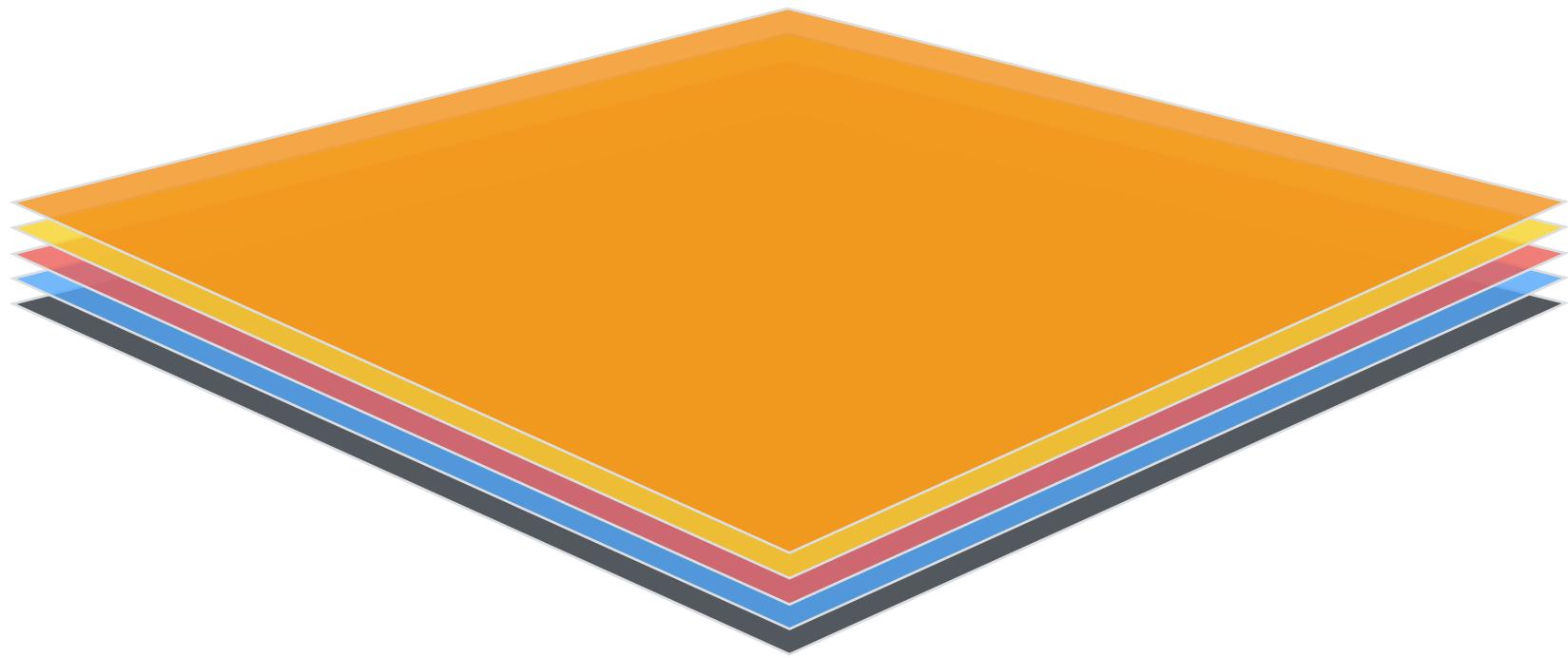


GGPLOT2

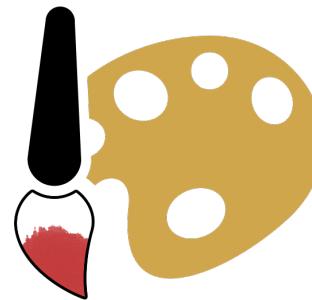




GGPLOT2



NEW GRAPHIC OBJECT



GGPLOT2

## Data Layout

`reshape2`

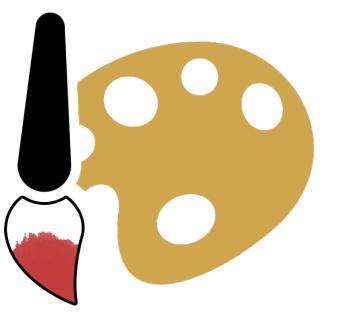
## Data Transform

`dplyr`

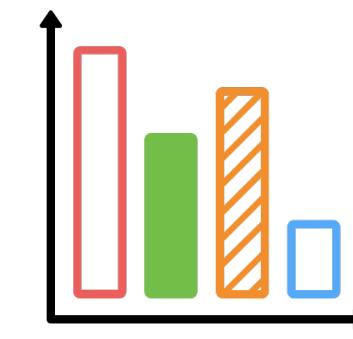
## Data Visualization

`ggplot2`

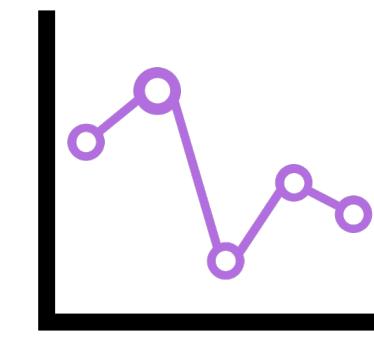




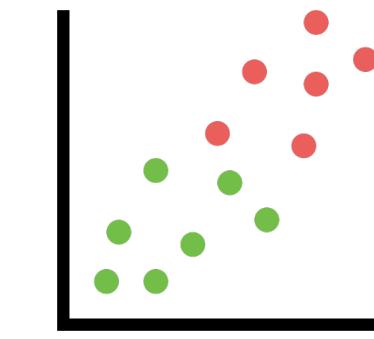
GGPLOT2



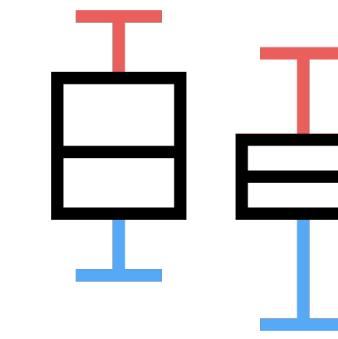
BAR CHART



LINE GRAPH



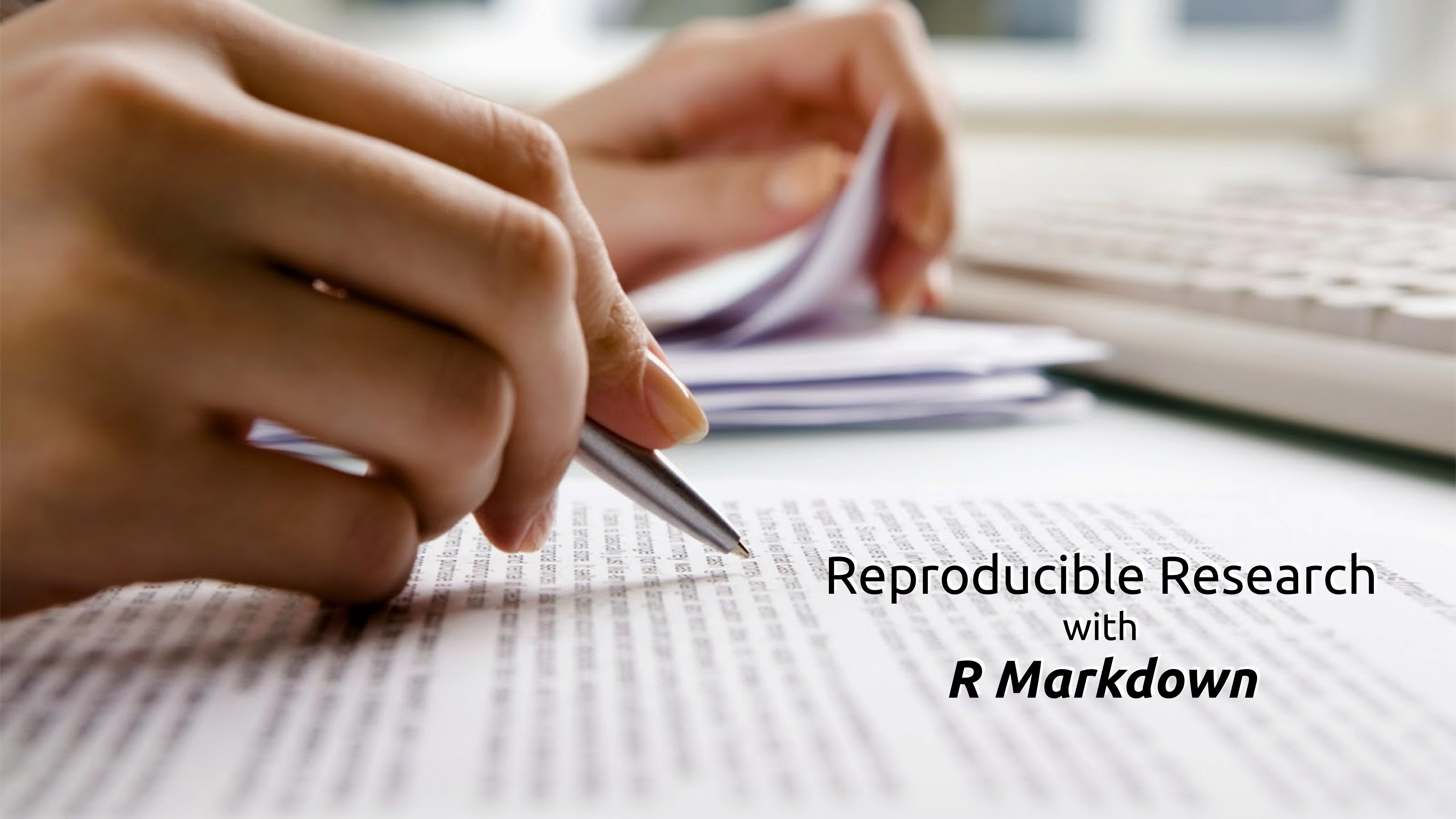
SCATTER PLOT



DATA  
DISTRIBUTION



MAP



# Reproducible Research with *R Markdown*

# R Markdown

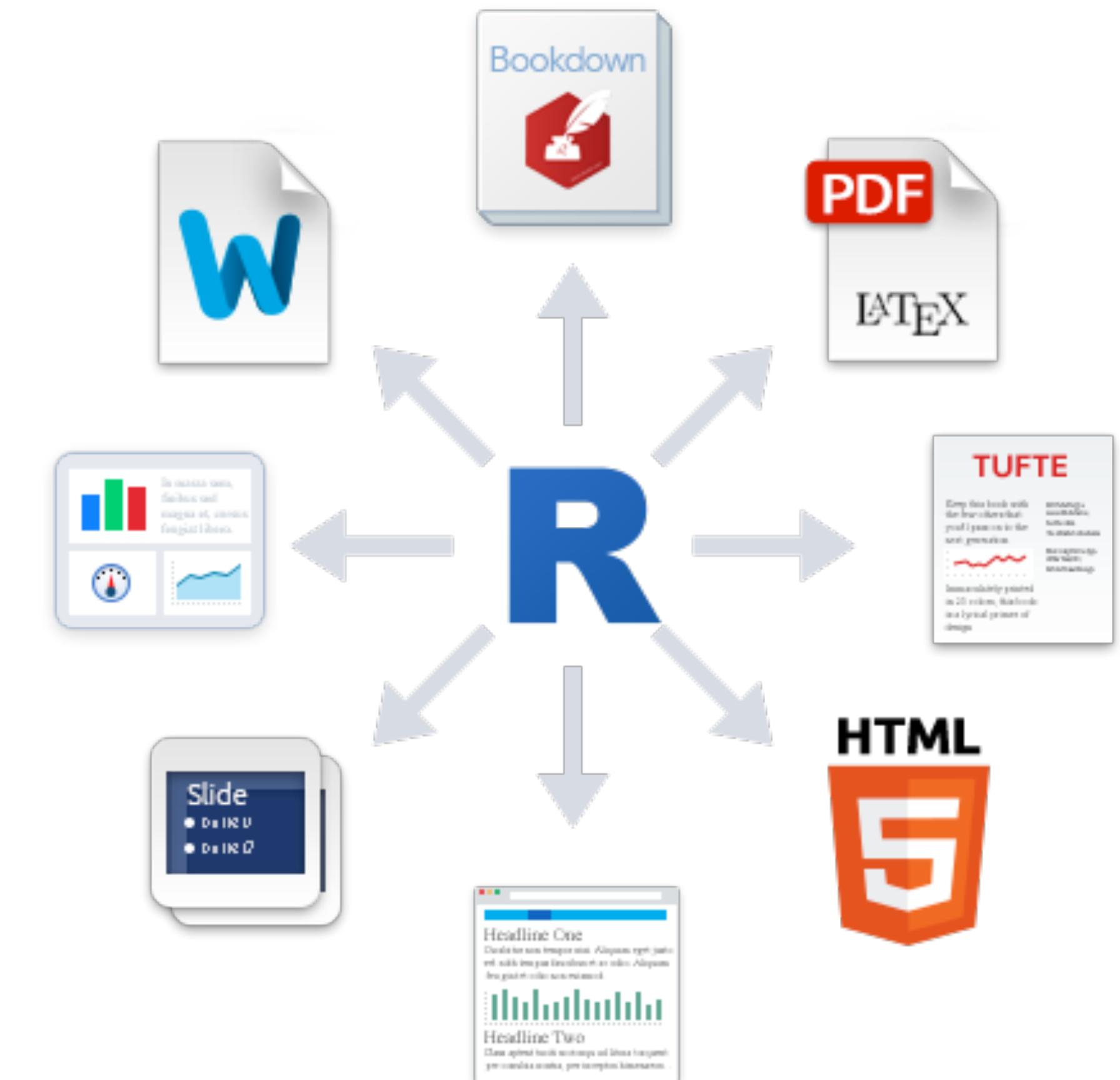
<http://rmarkdown.rstudio.com>

Authoring Format

Easy To Write with R Codes

Fully Reproducible

Various Types of Output Formats  
(HTML, PDF, MS Word, HTML5 Slides, etc.)



LR.Rmd

```

95
96  ``{r}
97  data(cars)
98  head(cars)
99  ```
100
101 `cars` 데이터에서 `speed`는 차량의 속도, `dist`는 정지거리를 의미합니다. 선형회귀는
    `lm()` 함수를 사용해서 모델링할 수 있습니다. 사용법은 다음과 같습니다.
102
103 lm(formula, data, ...)
104
105 `formula`는 선형회귀 모델에 대한 공식을, `data`는 사용할 데이터를 넣으면 됩니다.
    일반적으로 정지거리는 속도가 빠를 수록 증가하므로, 다음과 같은 모델을 생각할 수
    있습니다.
106 $$
107 dist \approx \beta_0 + \beta_1 \times speed. \tag{3}
108 $$
109
110 위 공식을 `lm()` 함수의 `formula` 자리에 넣고 모델링을 해보도록 하겠습니다.
111
112 ``{r}
113 model <- lm(dist ~ speed, data = cars)
114 model
115
116 모델링의 결과에 따르면  $\beta_0 = -17.579$ ,  $\beta_1 = 3.932$ 가 나왔습니다. 다시
    말해서 식 (3)은 아래 식 (4)와 같습니다.
117
118 $$
119 dist \approx -17.579 + 3.932 \times speed.
120 $$
121

```

Line 105, Column 27

Spaces: 2

R Markdown

UNREGISTERED

file:///Users/Han/Google%20Drive/N

처음 해 볼 실습은 차량 관련 데이터를 활용한 단순선형회귀입니다. cars 데이터는 R에 내장되어 있는 데이터로, 차량의 속도에 따른 정지거리를 기록한 데이터입니다. 데이터에 대한 자세한 설명은 ?cars를 실행하여 확인할 수 있습니다.

`data(cars)`

`head(cars)`

speed	dist
4	2
4	10
7	4
7	22
8	16
9	10

cars 데이터에서 speed는 차량의 속도, dist는 정지거리를 의미합니다. 선형회귀는 lm() 함수를 사용해서 모델링할 수 있습니다. 사용법은 다음과 같습니다.

`lm(formula, data, ...)`

formula는 선형회귀 모델에 대한 공식을, data는 사용할 데이터를 넣으면 됩니다. 일반적으로 정지거리는 속도가 빠를 수록 증가하므로, 다음과 같은 모델을 생각할 수 있습니다.

$$dist \approx \beta_0 + \beta_1 \times speed. \tag{3}$$

위 공식을 lm() 함수의 formula 자리에 넣고 모델링을 해보도록 하겠습니다.

`model <- lm(dist ~ speed, data = cars)`

메뉴 표시

```
LR.Rmd UNREGISTERED
38
39 # 선형회귀법(Linear Regression)[^1] -
40
41 [^1]: 회귀(regression)라는 용어는 영국의 우생학자인 프랜시스 갈튼(Sir Francis Galton, 1822-1911)으로부터 유래했습니다. 갈튼은 부모의 키와 아이들의 키 사이의 연관 관계를 연구하면서 아이들의 키는 부모의 키 평균으로 '돌아가려는 경향'이 있다는 가설을 세우고 이를 분석하는 방법으로 회귀분석이라고 하였죠. 회귀라는 뜻의 영어 단어 regress에서 파생되었습니다.
42
43 단순해보이지만 굉장히 깊이 있는 통계학 기반의 기계학습 방법인 선형회귀법에 대해서 알아보도록 하겠습니다. 선형회귀법은 *독립변수(independent variable)*와 *종속변수(dependent variable)* 사이의 관계를 모델링하는 기법을 말합니다. 한 변수(독립변수)의 변화로부터 다른 변수(종속변수)가 어떻게 변하는지 예측하는 기법이 선형회귀법입니다. 이 때 독립변수의 개수가 하나인 경우 **단순 선형 회귀(Simple Linear Regression)**, 독립변수가 두 개 이상인 경우 **다중 선형 회귀(Multiple Linear Regression)**라고 부릅니다. 독립변수  $\$Y\$$ 와 종속변수  $\$X_1, X_2, \dots, X_n\$$ 에 대하여 다음과 같이 나타낼 수 있습니다.
44 $$
45 Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n. \tag{1}
46 $$
47
48 예를 들어,  $\$X = (1, 2, 3, 4, 5, 6, 7)\$$ 에 대하여  $\$Y = (3, 4, 5, 6, 7, 8, 9)\$$ 의 관계를 나타낸다면
49 $$
50 Y = X + 2
51 $$
52 로 나타낼 수 있습니다.
53
54 ````{r}
55 library(dplyr)
```

→ Header (Level 1 ~ Level 6)

→ Footnote

→ *Italic*

→ **Bold**

→ Math Equation

```
95
96  ``-{r}——
97  data(cars)
98  head(cars)
99  ````
100
101 `cars` 데이터에서 `speed`는 차량의 속도, `dist`는 정지거리를 의미합니다. 선형회귀는
    `lm()` 함수를 사용해서 모델링할 수 있습니다. 사용법은 다음과 같습니다.
102
103     lm(formula, data, ...)
104
105 `formula`는 선형회귀 모델에 대한 공식을, `data`는 사용할 데이터를 넣으면 됩니다.
    일반적으로 정지거리는 속도가 빠를 수록 증가하므로, 다음과 같은 모델을 생각할 수
    있습니다.
106 $$
107 dist \approx \beta_0 + \beta_1 \times speed. \tag{3}
108 $$
109
110 위 공식을 `lm()` 함수의 `formula` 자리에 넣고 모델링을 해보도록 하겠습니다.
111
112 ``-{r}
113 model <- lm(dist ~ speed, data = cars)
114 model
115 ````
116
117 모델링의 결과에 따르면  $\beta_0 = -17.579$ ,  $\beta_1 = 3.932$ 가 나왔습니다. 다시
    말해서 식 (3)은 아래 식 (4)와 같습니다.
118 $$
119 dist \approx -17.579 + 3.932 \times speed.
120 $$
121
```

# Code Chunk

# Inline Code

# Blocked Code

# 당위성

(當爲性, appropriateness)

[명사]

마땅히 그렇게 하거나 되어야 할 성질.

# 설득

(說得, persuasion)

[명사]

상대편이 이쪽 편의 이야기를 따르도록 여러 가지로 깨우쳐 말함.



마신러닝이란?

비교 검색어 ▾

Big Data  
검색어

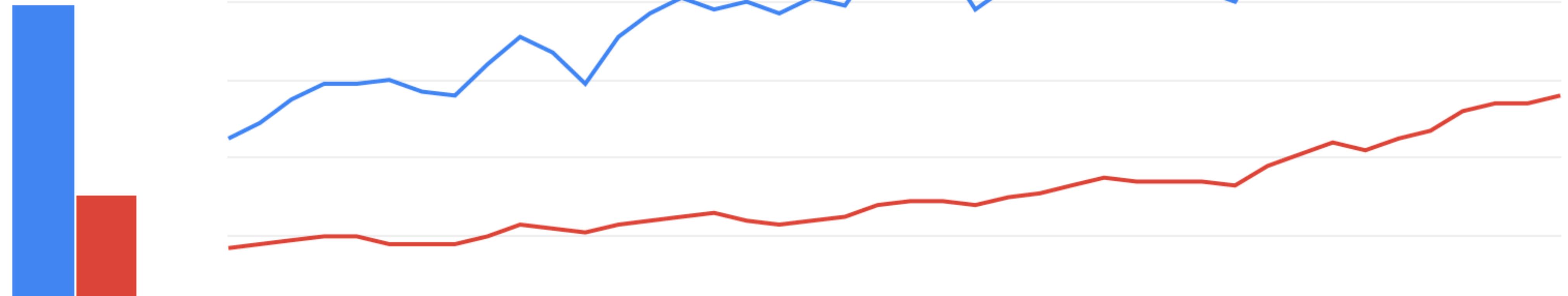
Machine Learning  
검색어

+검색어 추가

## 시간 흐름에 따른 관심도 변화



■ 뉴스 제목 ? ■ 예측 ?



평균

2013년 7월

2014년 1월

2014년 7월

2015년 1월

2015년 7월

2016년 1월

</>

~~Big Data?~~

Machine Learning?

# Machine Learning

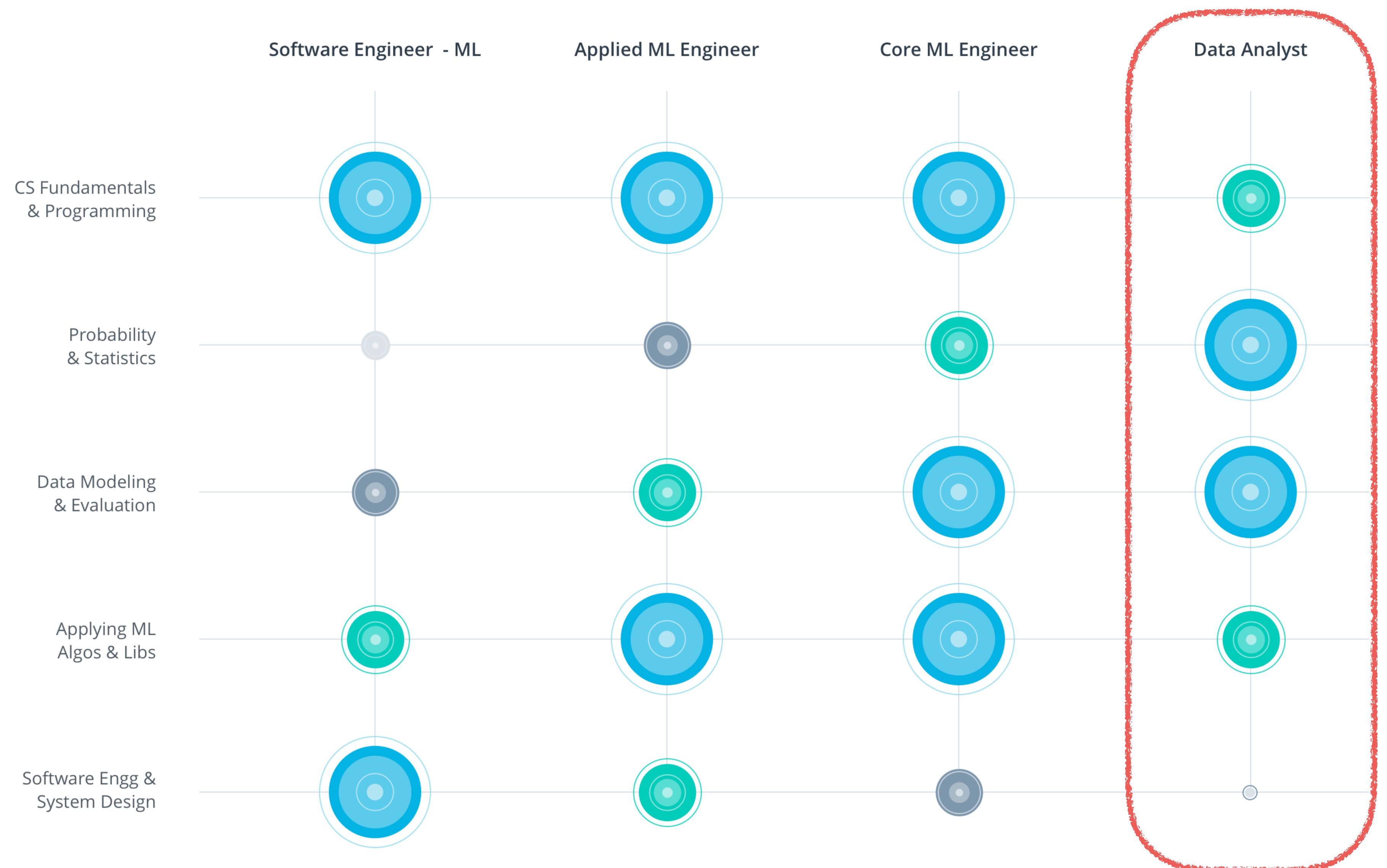
**컴퓨터**에게 직접 답을 알려주지 않고,  
**데이터**를 통해 컴퓨터가 **학습**을 하여  
**문제를 해결**하도록 하는 것

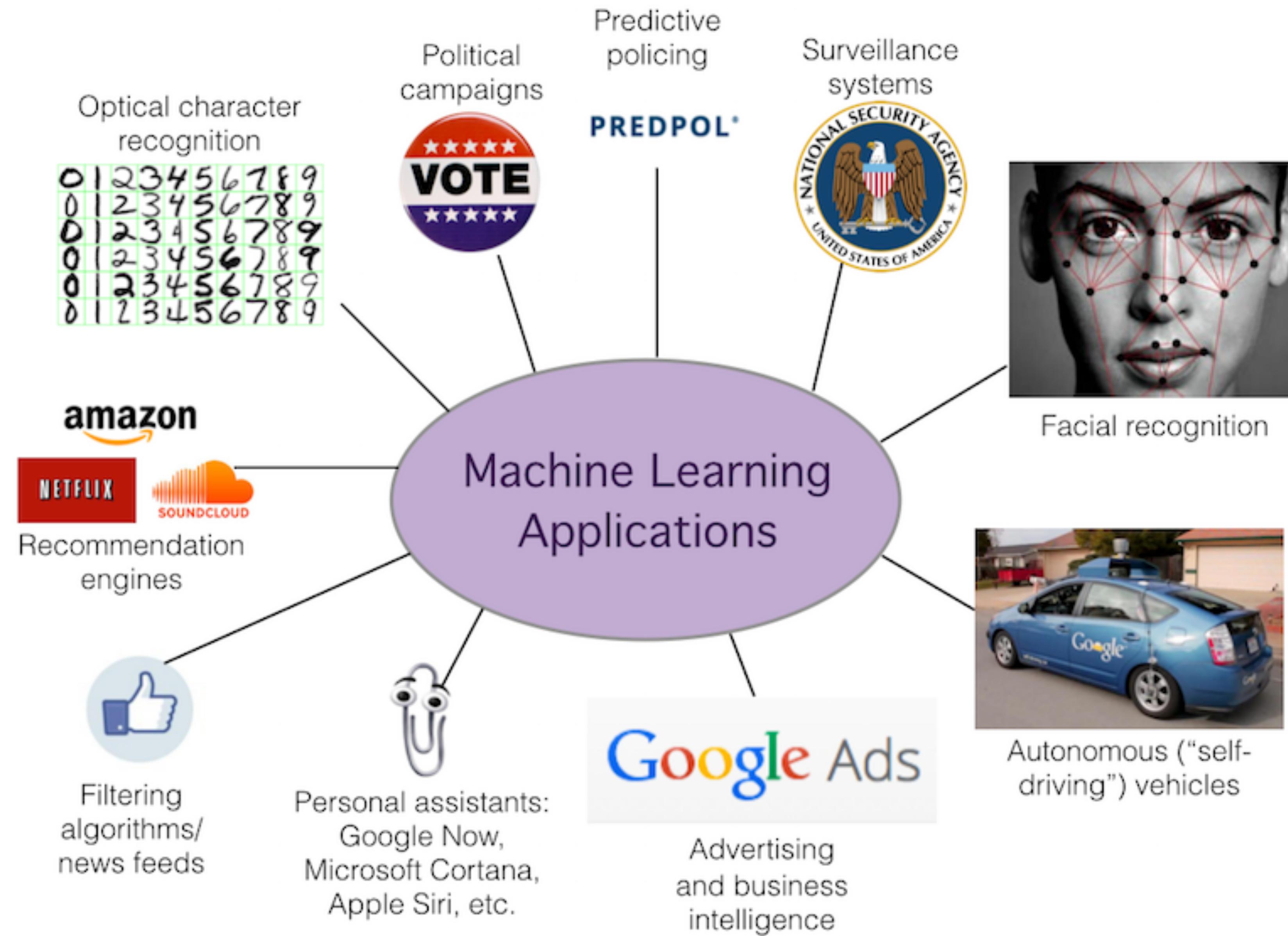
# Machine Learning

*COMPUTER SCIENCE and PROGRAMMING*

+

*MATHEMATICS*







*things you should know*



THINGS YOU SHOULD KNOW

예측?

관찰?



THINGS YOU SHOULD KNOW

# Supervised Learning

# Unsupervised Learning

# Supervised Learning

가지고 있는 데이터에 이름이 붙여져 있다.

Cat



Dog



위 사진들을 학습하고 새로운 사진들에서 답을 찾는다.

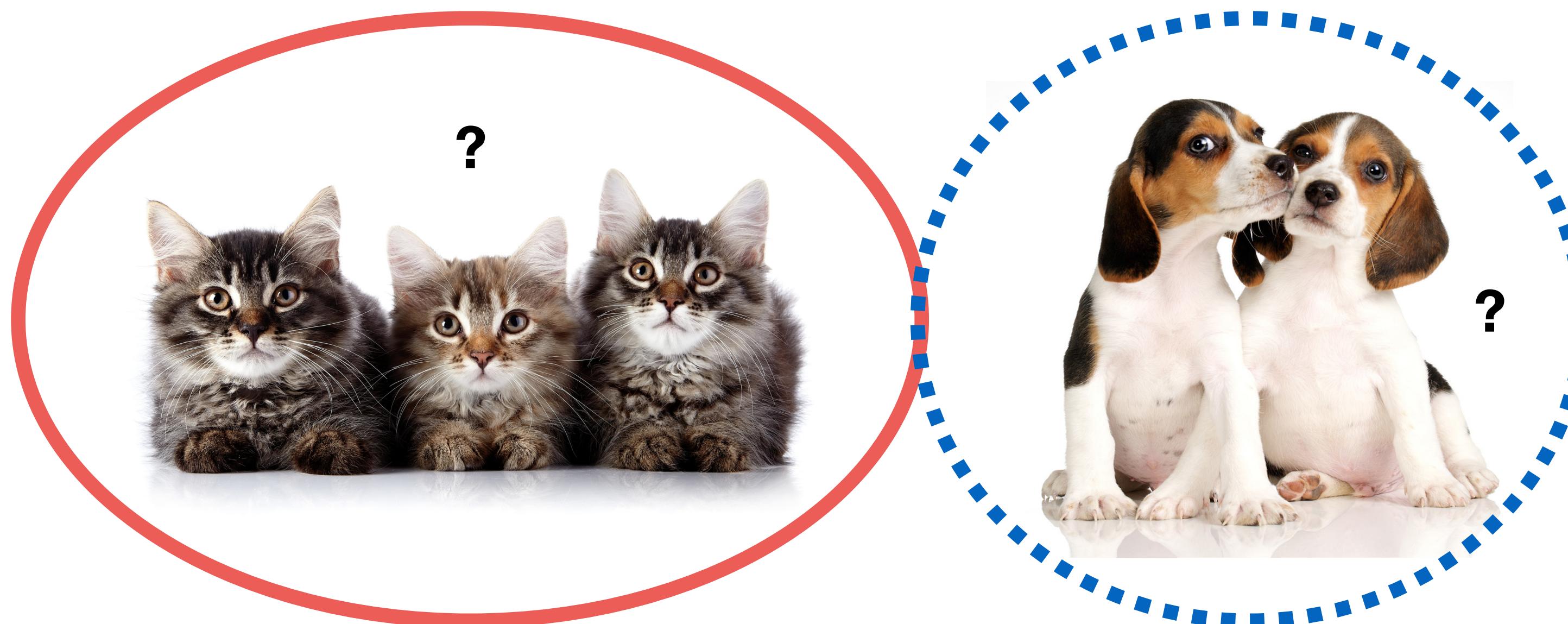
# Supervised Learning

Training  
Data

Test  
Data

7 : 3

# Unsupervised Learning



동물들이 무슨 동물인지는 모른다.

단지 비슷해보이는 것들끼리 묶으면 된다.



*Different* **DATA**

*Different* **OBJECTIVES**

*Different* **ALGORITHMS**

*Different* **ALGORITHMS**

Linear Regression

Logistic Regression

k-Nearest Neighbor

Clustering

CART

Random Forest

Boosting

Support Vector Machine

# ***Supervised*** Learning

Linear Regression

Logistic Regression

k-Nearest Neighbor

Clustering

CART

Random Forest

Boosting

Support Vector Machine

# ***Supervised Learning***

## ***Regression***

Linear Regression

Logistic Regression

k-Nearest Neighbor

Clustering

CART

Random Forest

Boosting

Support Vector Machine

# ***Supervised*** Learning *Classification*

Linear Regression

Logistic Regression

k-Nearest Neighbor

Clustering

CART

Random Forest

Boosting

Support Vector Machine

# ***Unsupervised*** Learning

Linear Regression

Logistic Regression

k-Nearest Neighbor

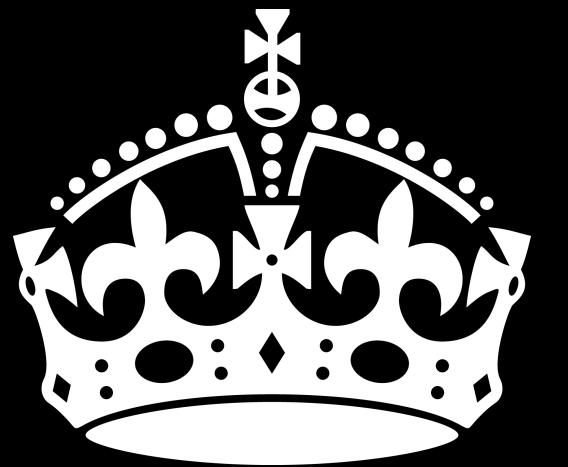
Clustering

CART

Random Forest

Boosting

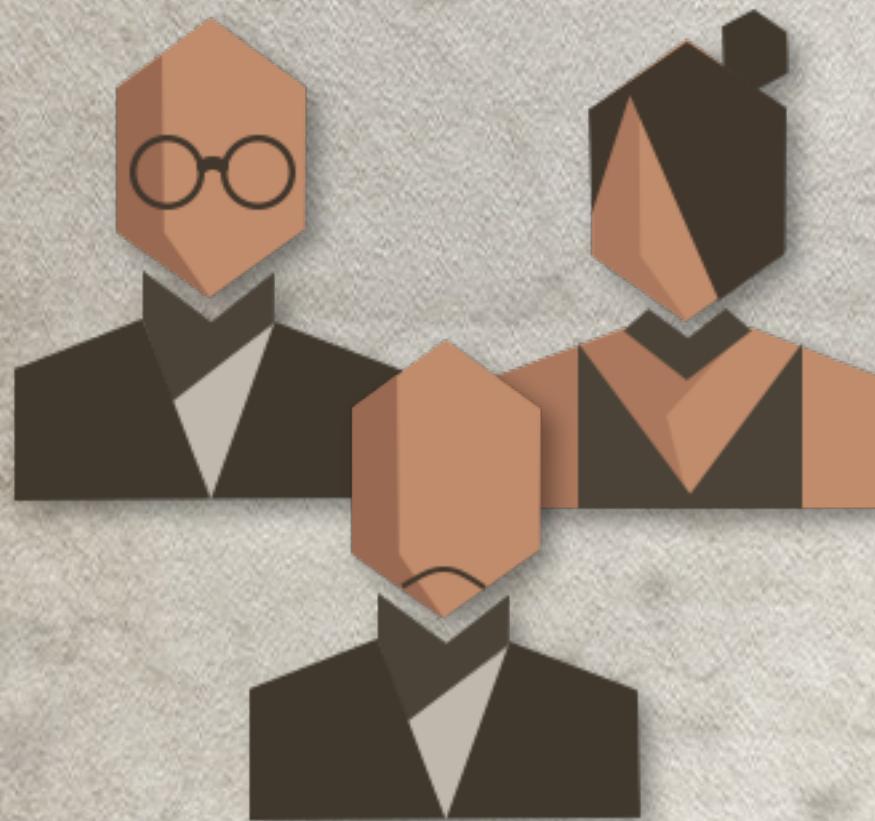
Support Vector Machine



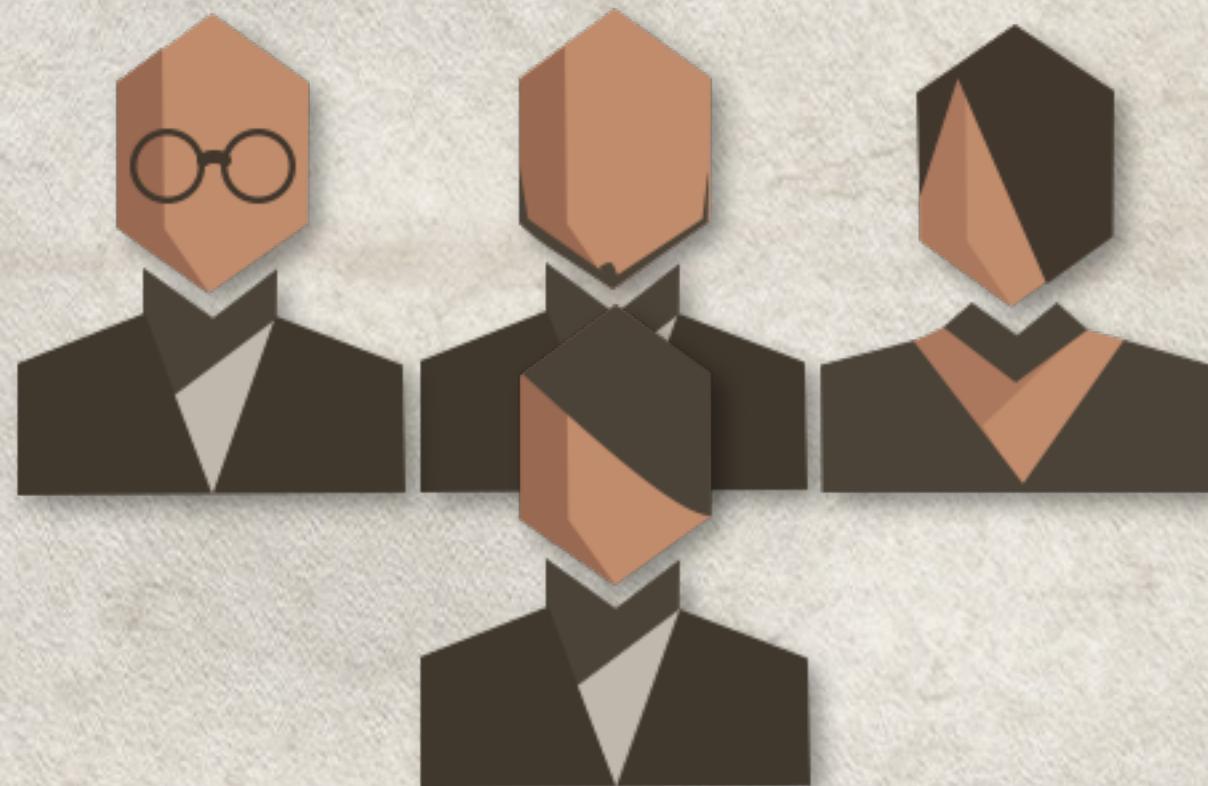
KEEP  
CALM  
AND  
THIS IS  
A COMPETITION

MainMatch

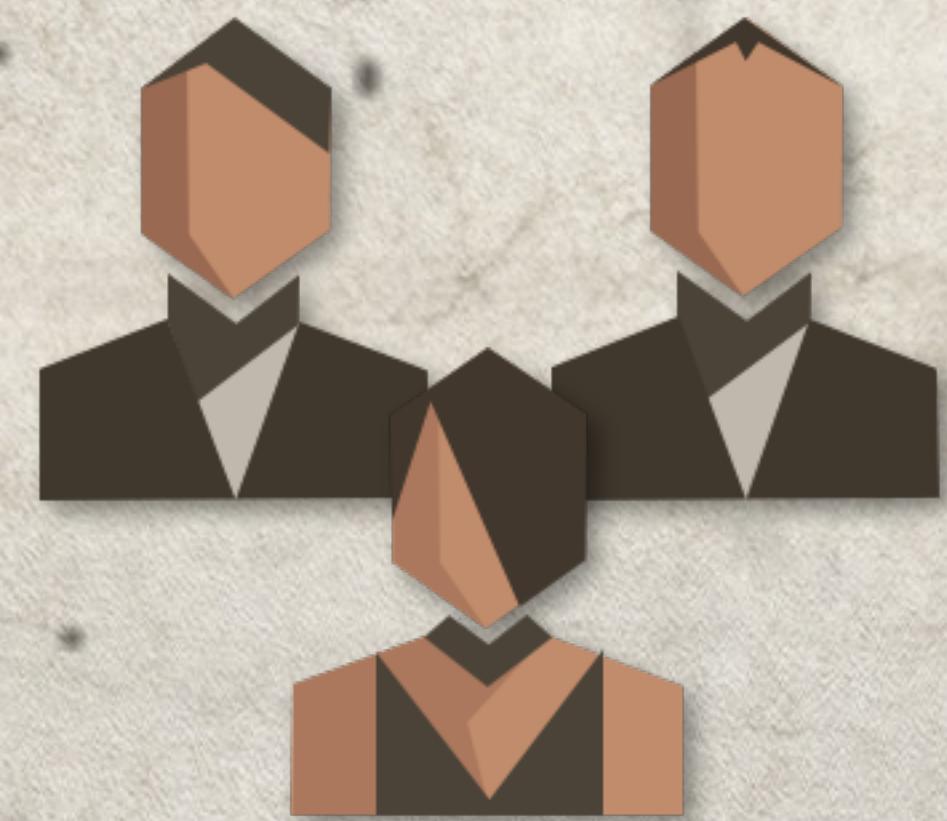
머신 러닝



Team 1



Team 2



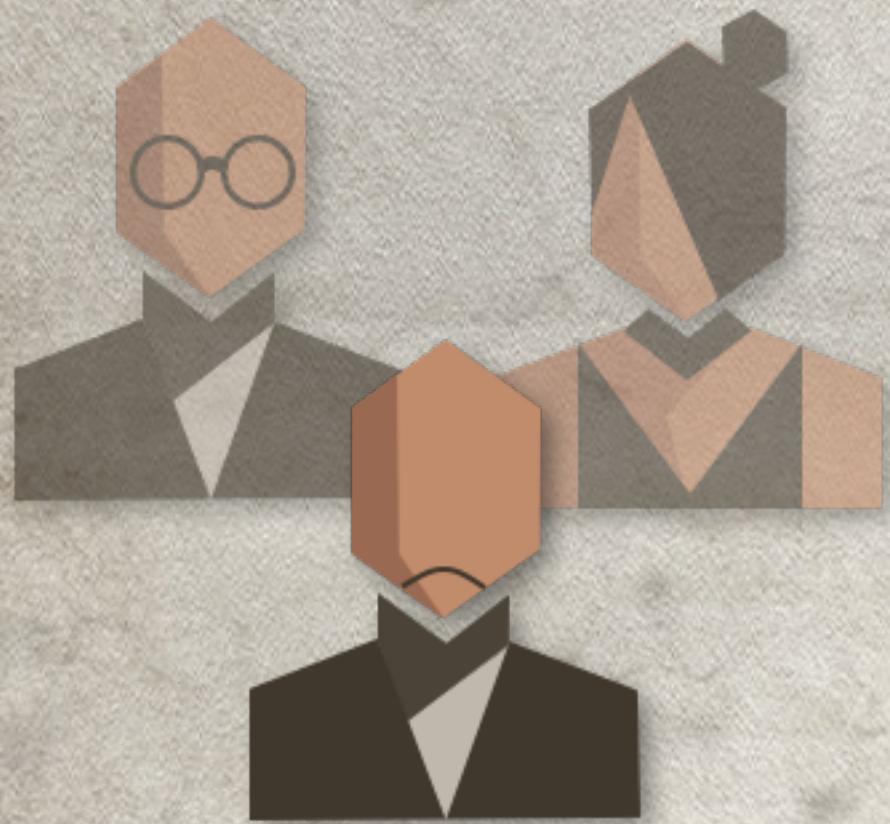
Team 3



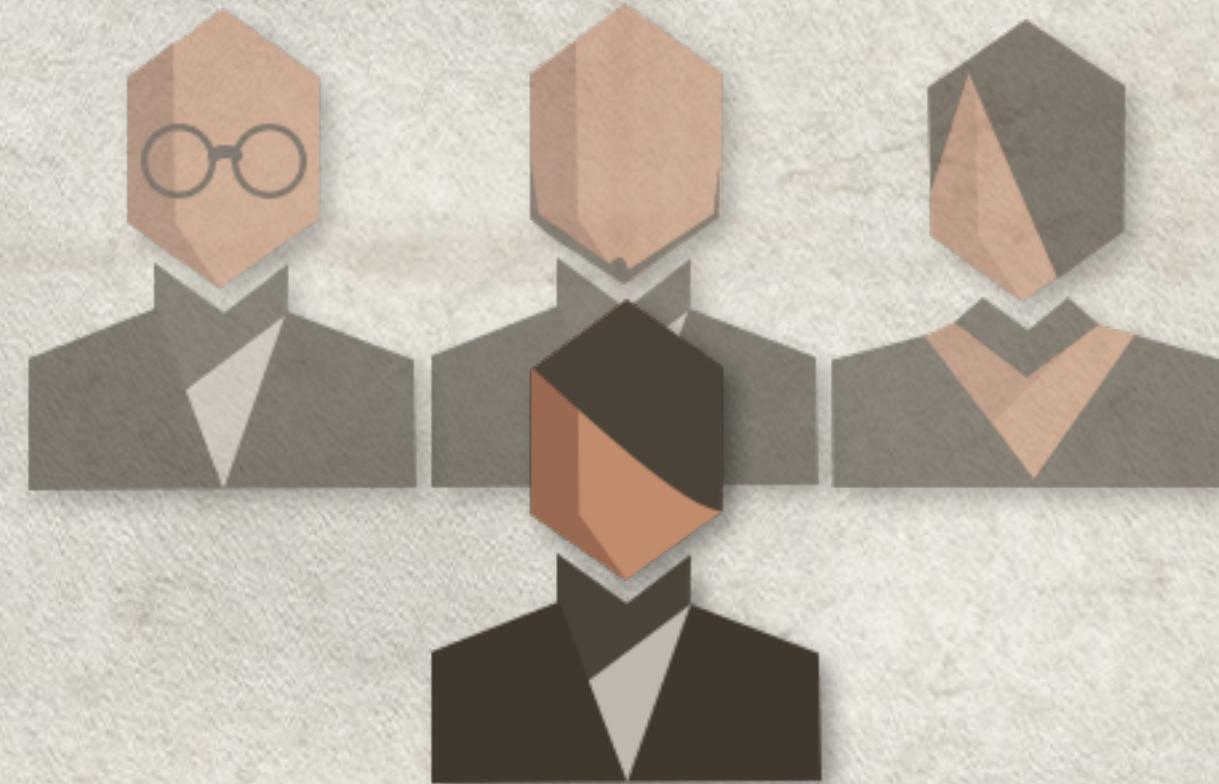
# Linear Regression

MainMatch

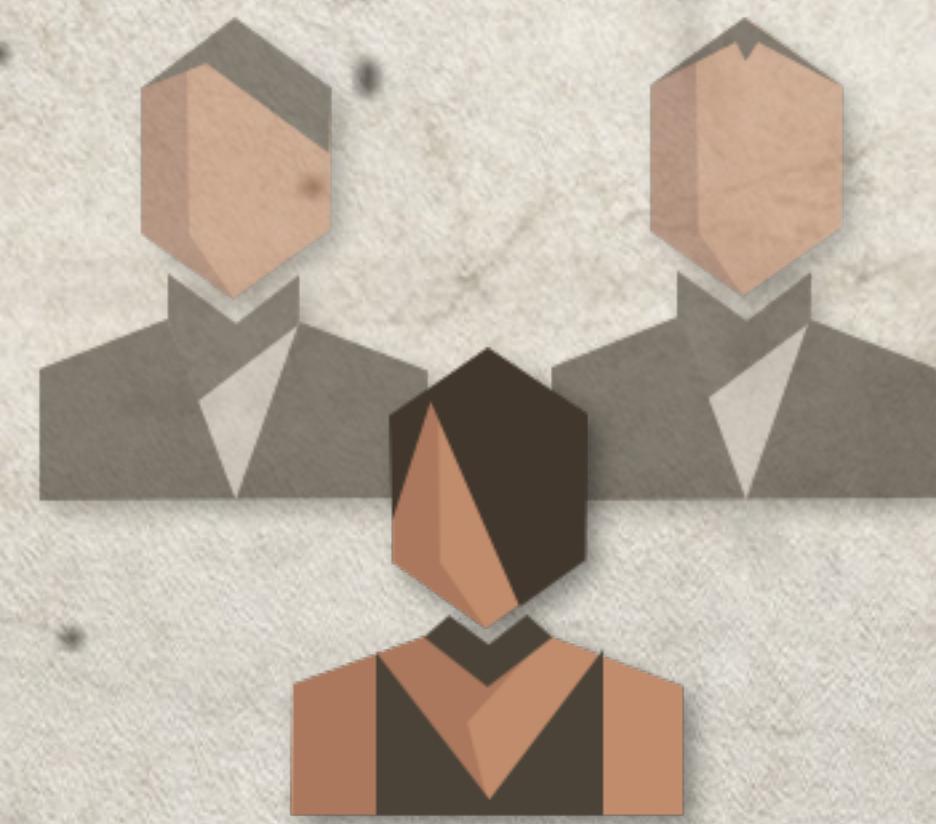
# 머신 러닝



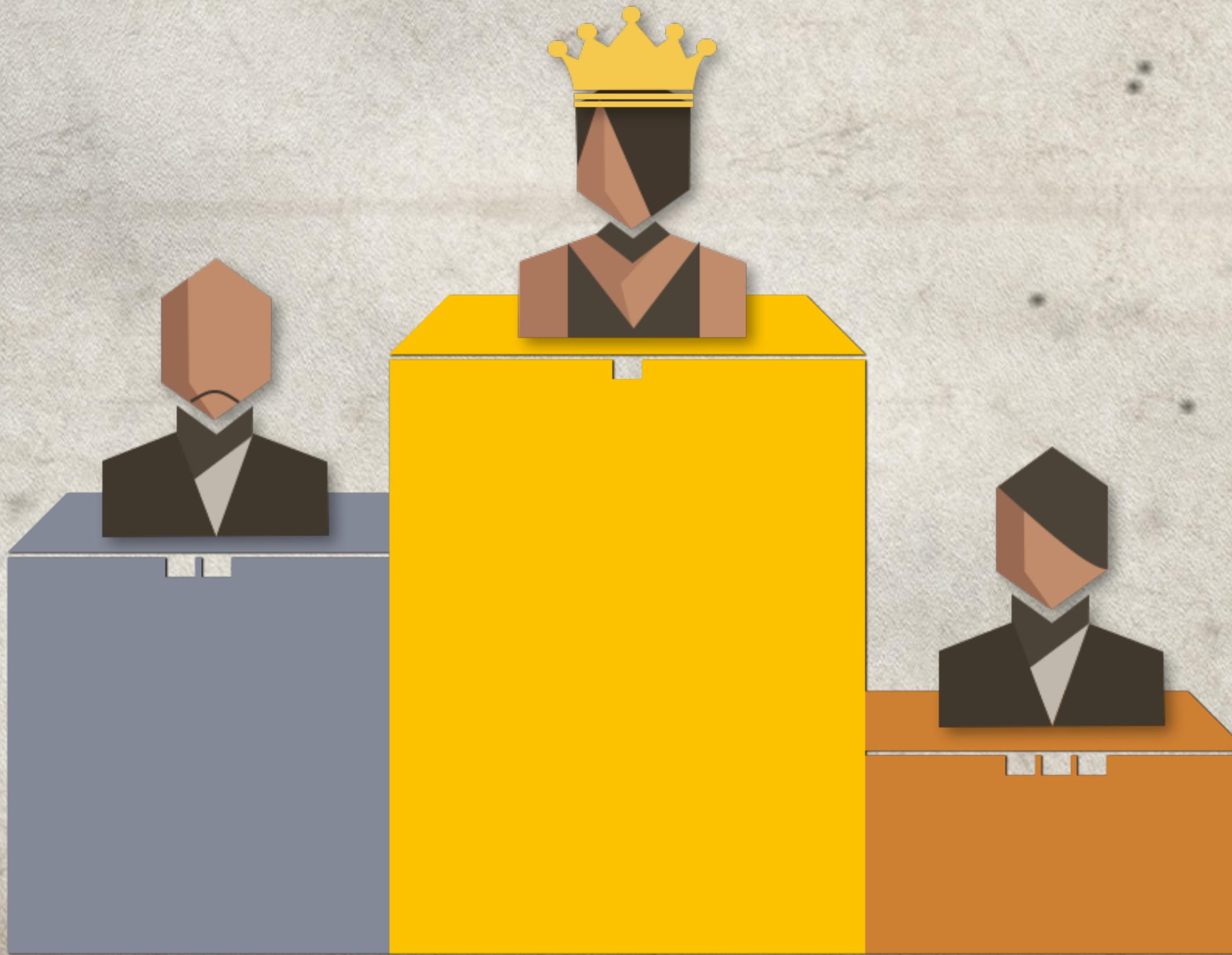
0.85

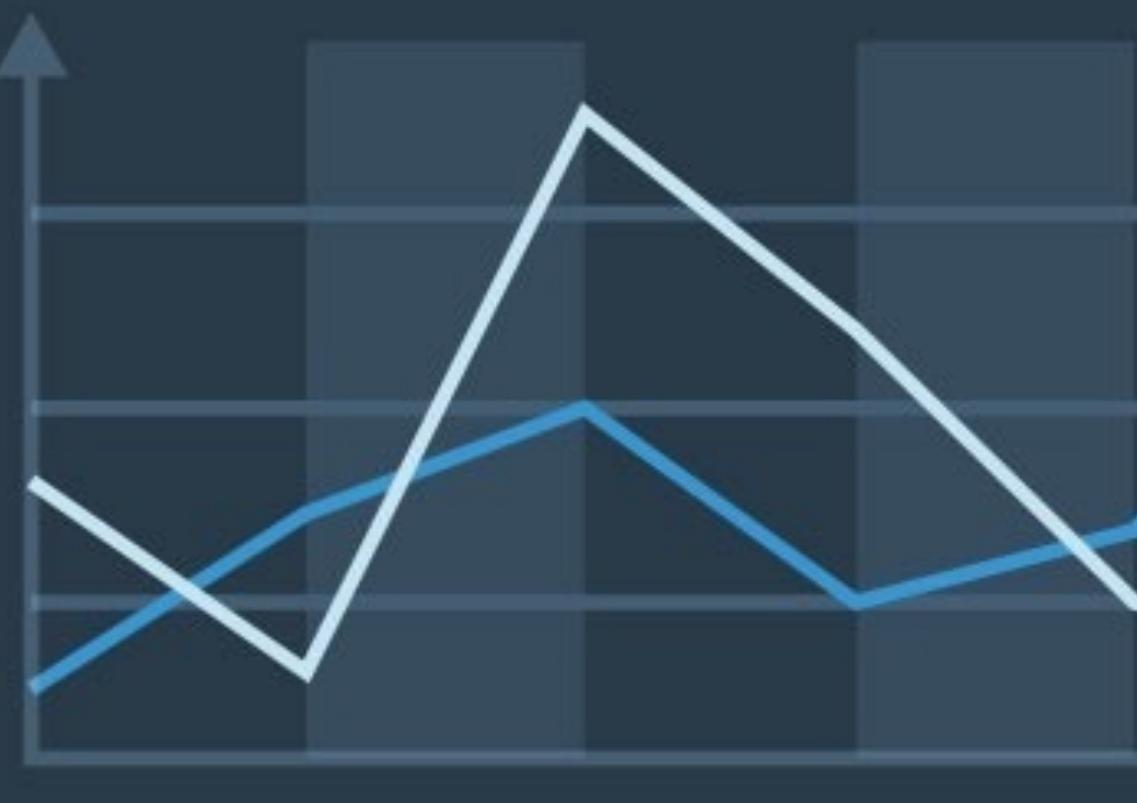


0.88



0.79





과제  
데이터 시각화

주어진 데이터에 대해서  
다양한 시각화를 해보자.



- Test A
- Test B

과제

## 데이터 시각화

주어진 데이터에 대해서  
다양한 시각화를 해보자.

R은 다양한 데이터를 내장하고 있다.  
`data()`를 입력하면 확인 가능하다.

30 out of 100

65 out of 100

아무 데이터를 활용해서  
데이터를 **표현**할 수 있거나  
데이터를 **요약**할 수 있는  
훌륭한 시각화 결과물을 만들어보자.

● Test A  
Test B

A B C D E F G H I J K L M



THX :)