

R

WEEK 3

LAST WEEK

LAST WEEK

DATA IMPORT & EXPORT

IF / ELSE

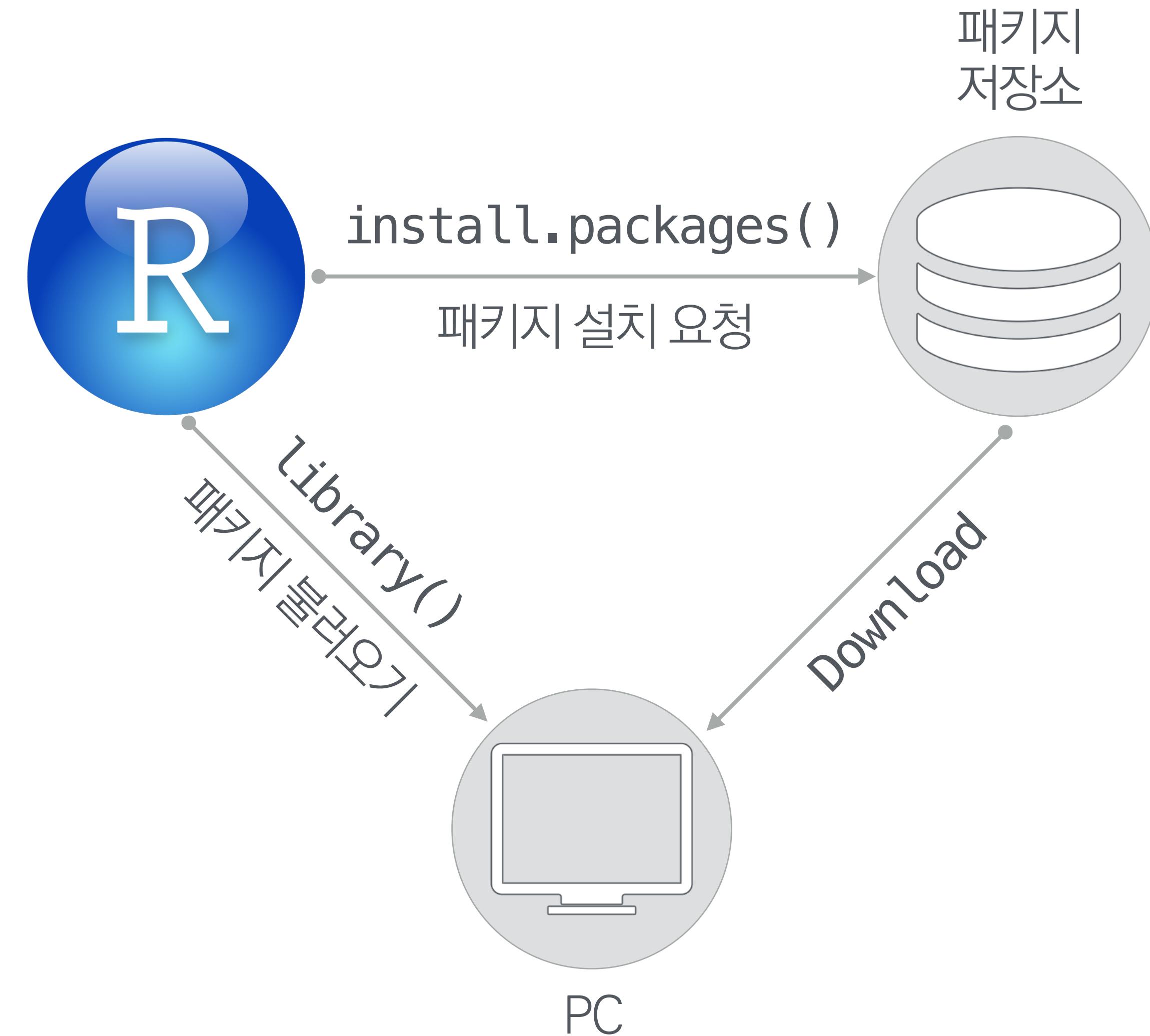
FOR

WHILE

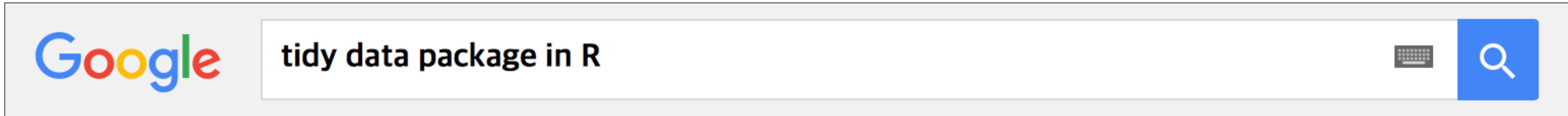
DATA MANIPULATION I
(APPLY FUNCTIONS)

INSTALLING R PACKAGES

WORK FLOW



WORK FLOW



Tidy data - CRAN

<https://cran.r-project.org/.../packages/tidyr/.../tidy-da...> ▾ 이 페이지 번역하기

2016. 2. 5. – Tidy data is a standard way of mapping the meaning of a dataset to its ... is particularly well suited for vectorised programming languages like R, ...

CRAN – Package tidyr

<https://cran.r-project.org/package=tidyr> ▾ 이 페이지 번역하기

2016. 2. 5. – tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions ... r-devel: tidyr_0.4.1.zip, r-release: tidyr_0.4.1.zip, r-oldrel: tidyr_0.4.1.zip.

DATA MANIPULATION

II

DATA AGGREGATION & DATA HANDLING

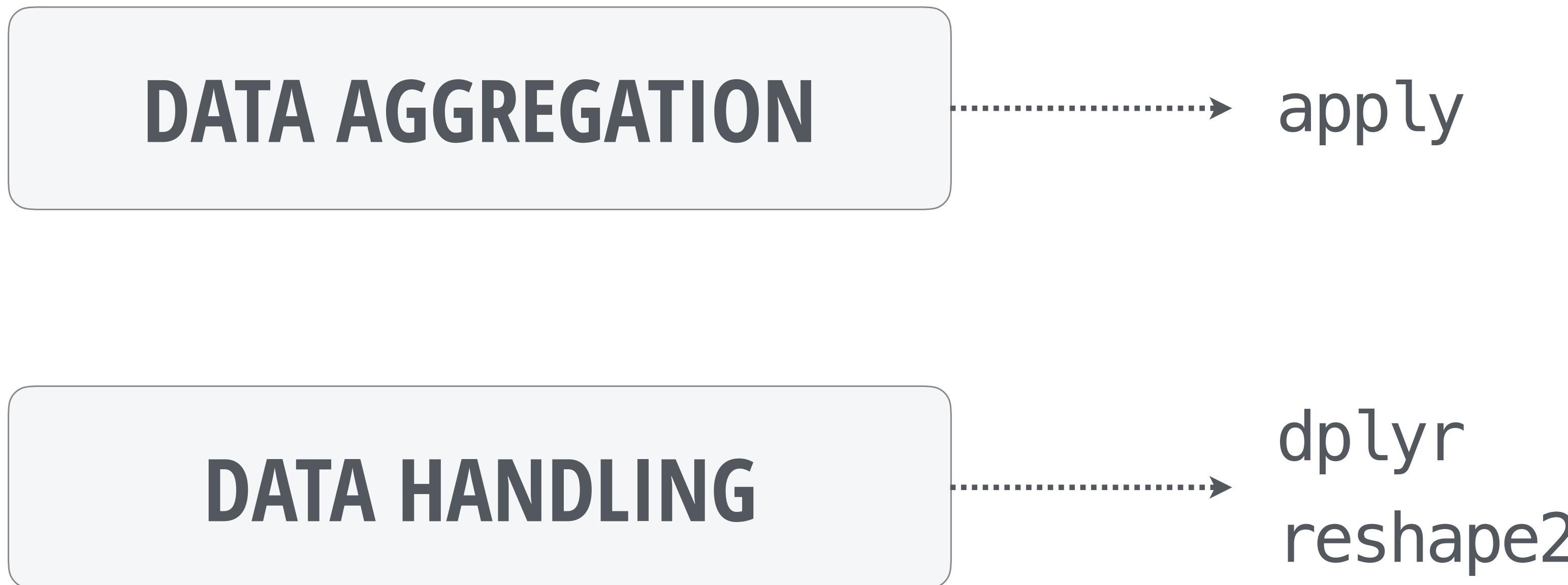
DATA AGGREGATION

데이터에 대하여 통계적으로 집단의 대표값을 구하는 것
(평균, 중앙값, 최댓값, 최솟값)

DATA HANDLING

데이터를 원하는 형태로 바꾸거나 매핑하는 과정

DATA AGGREGATION & DATA HANDLING



TIME TO HANDLE THE DATA



DPLYR

DPLYR

SO FAST

SO INTUITIVE

FOR DATA FRAME

SO CONVENIENT

DPLYR

TOO

DAMN

FAST

DPLYR

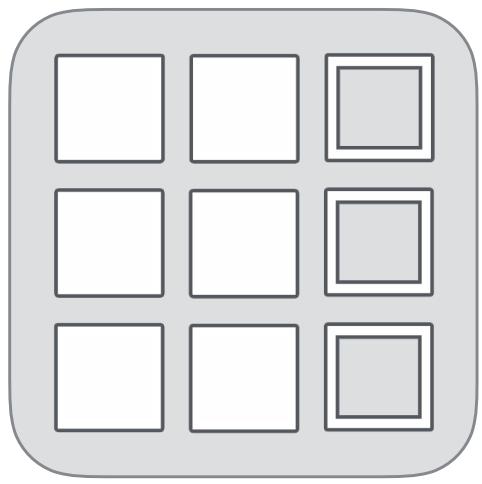


SWISS ARMY KNIFE FOR DATA HANDLING

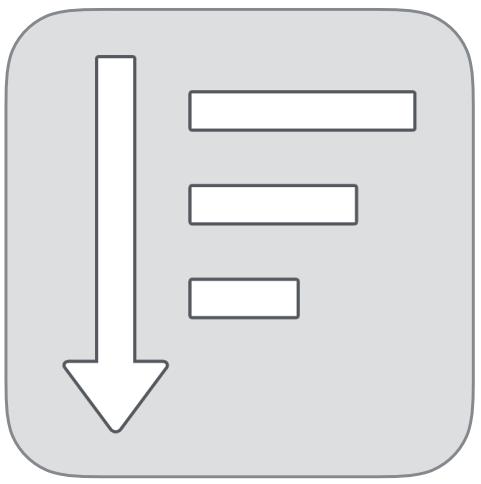
DPLYR



filter



select



arrange



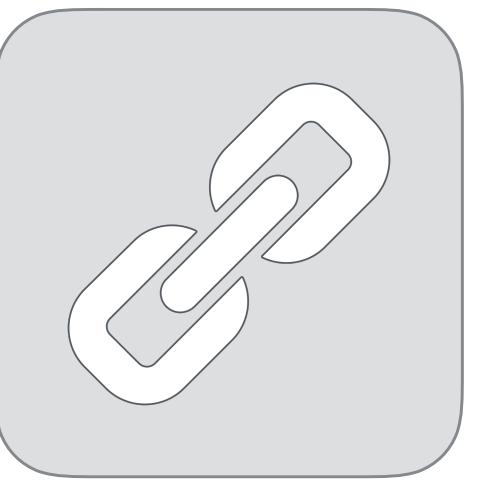
mutate



group_by



summarise



chaining

WITH THIS DATA

2013년 뉴욕 항공 데이터

336,776 x 19



READY!

```
install.packages("nycflights13")
install.packages("dplyr")
library(nycflights13)
library(dplyr)
data(flights)
```

GO!

flights

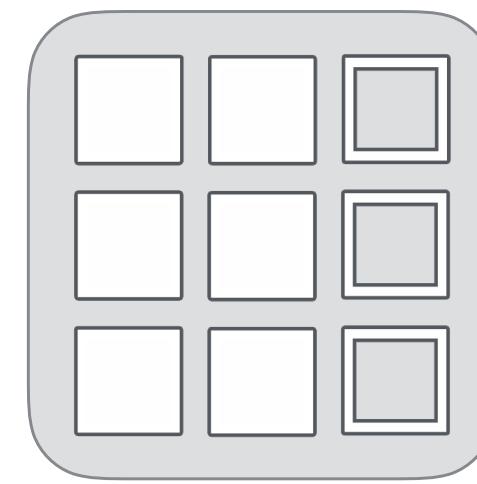
FILTER



`filter(data, conditions, ...)`

원하는 조건의 데이터를 쉽게 얻어낼 수 있다.

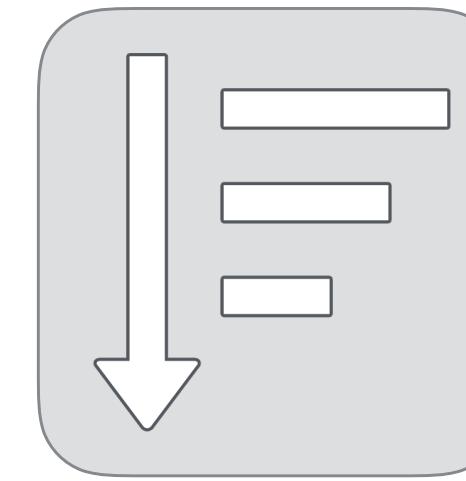
SELECT



`select(data, column.name, ...)`

데이터에서 원하는 칼럼만 추출한다.

ARRANGE



`arrange(data, column.name, desc(column.name))`

데이터에서 원하는 칼럼을 기준으로 정렬한다.

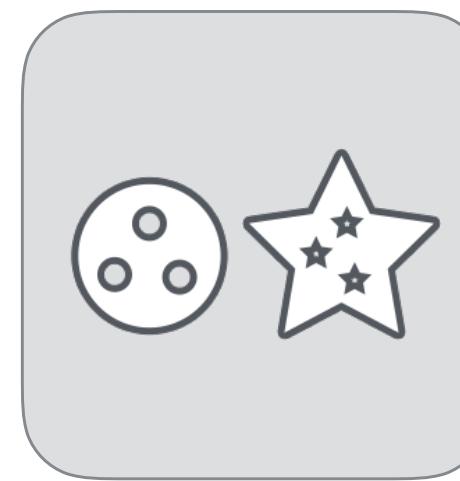
MUTATE



`mutate(data, new.column)`

기존 데이터를 이용해서 새로운 칼럼을 생성한다.

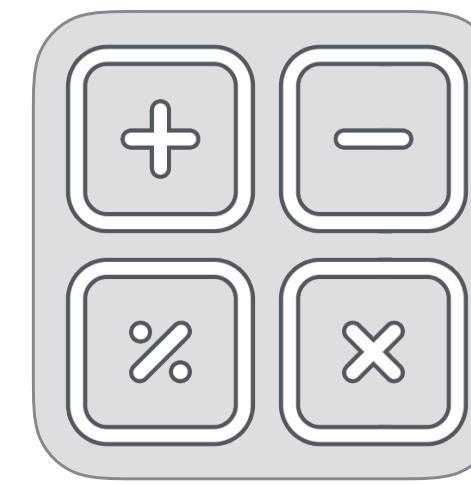
GROUP_BY



`group_by(data, column.name)`

데이터 집계를 위해서 데이터를 그룹핑한다.

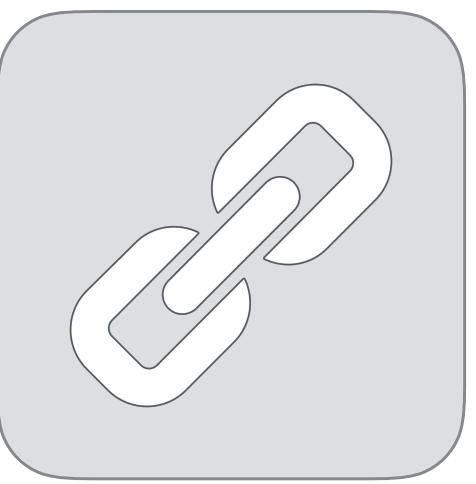
SUMMARISE



`summarise(data, new.column)`

그룹핑한 데이터를 집계한다.

CHAINING



%>%

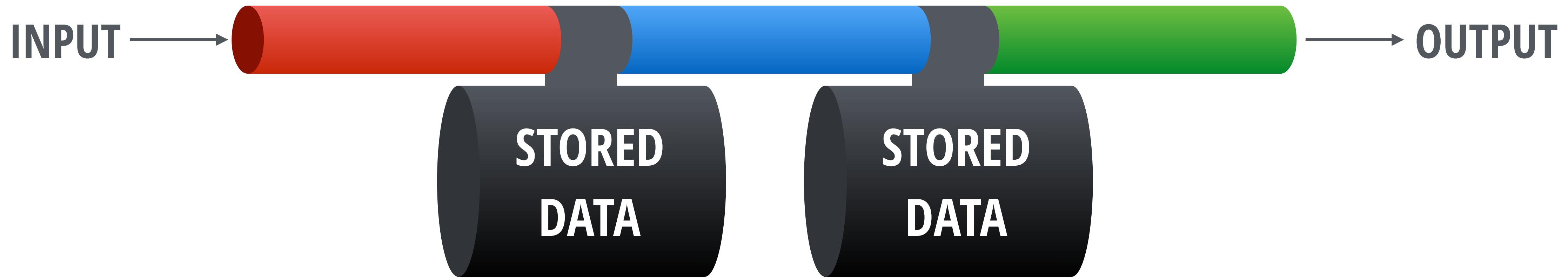
CHAINING

%>%

PIPE OPERATOR

일련의 데이터 핸들링 과정을
읽기 쉽고 더 빠르게
작성할 수 있도록 돕는다.

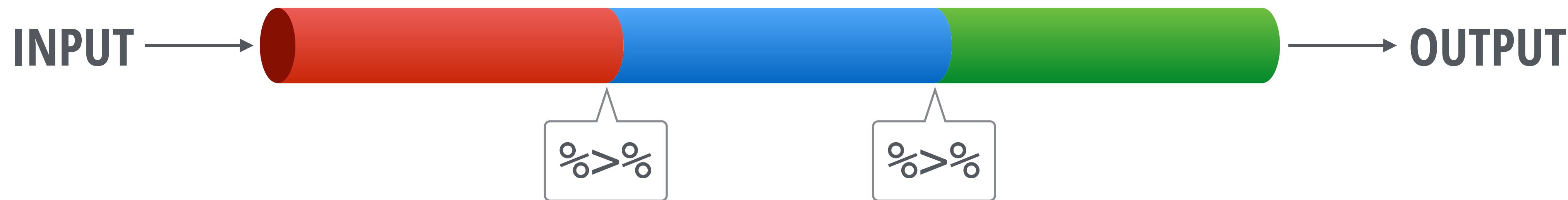
WITHOUT PIPE OPERATOR



저장된 데이터들이 메모리를 잡아먹는다.

%>%

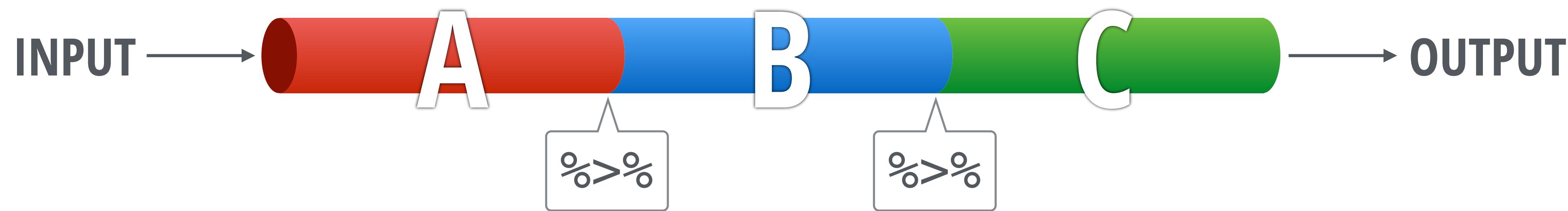
WITH PIPE OPERATOR



중간에 데이터 저장이 없기 때문에
메모리 관점에서 효율적이다.

%>%

WITH PIPE OPERATOR



%>%

百問不如一打

RESHAPE2

RESHAPE2

데이터의 형태 바꾸기

RESHAPE2

데이터의 형태 바꾸기

우선 R로 데이터를 만들어보자.

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

RESHAPE2

데이터의 형태 바꾸기

학생	국어 점수	수학 점수	영어 점수
WIDE LAYOUT			
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

LONG LAYOUT

LAYOUTS

WIDE LAYOUT

하나의 행에 여러 데이터가 포함되어 있다.

각각의 칼럼이 각각의 변수를 의미하지 않는다.

WIDE LAYOUT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

LONG LAYOUT

LAYOUTS



하나의 행에는 하나의 관측치만 포함되어 있다.

각 변수는 개별의 칼럼으로 존재한다.

LONG LAYOUT

학생	국어 점수	수학 점수	영어 점수
WIDE LAYOUT			
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

RESHAPE2

자유자재로 바꾸고 싶다.

학생	국어 점수	수학 점수	영어 점수
WIDE LAYOUT			
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

LONG LAYOUT

BASIC CONCEPTS OF RESHAPE2

데이터를 녹여서 원하는 모양의 거푸집에 봇는 과정

`melt()`

`cast()`

MELT

```
melt(data, id.vars, measure.vars,  
      variable.name = "name",  
      na.rm = FALSE, factorsAsStrings = TRUE)
```

MELT

```
melt(data, id.vars, measure.vars,  
      variable.name = "name",  
      na.rm = FALSE, factorsAsStrings = TRUE)
```

Uhh... What?

MELT

학생	국어 점수	수학 점수	영어 점수
학생			
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80

MELT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68

MELT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72

MELT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94

MELT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B		

MELT

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

CAST

dcast(data, formula, ...)

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80		
B			

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80		
B	68		

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80	72	
B	68		

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80	72	
B	68	94	

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	

MELT

학생	과목	점수
A	국어	80
B	국어	68
A	수학	72
B	수학	94
A	영어	77
B	영어	82

학생	국어 점수	수학 점수	영어 점수
A	80	72	77
B	68	94	82

MORE... MORE!

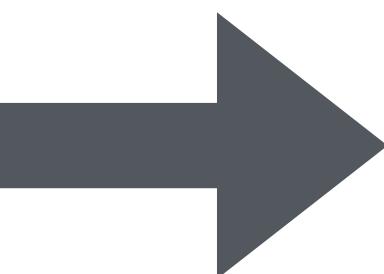
뉴욕 도시의 대기질 측정 데이터 (1973년 5월 ~ 1973년 9월)



OBSERVATION

```
wide_airq <- dcast(molten_airq, month + day ~ climate_variable,  
value.var = "climate_value")
```

month	day	variable	value
5	1	ozone	41
5	2	ozone	36
5	3	ozone	12
5	4	ozone	18
5	5	ozone	NA
5	6	ozone	28



month	day	ozone	solar.r	wind	temp
5	1	41	190	7.4	67
5	2	36	118	8.0	72
5	3	12	149	12.6	74
5	4	16	313	11.5	62
5	5	NA	NA	14.3	55
5	6	28	NA	14.9	66

DATA VISUALIZING

DATA VISUALIZING



우리는 정보 과부하나 자료 과다로부터
비롯된 모든 고로움에 있는 것을 느낀다.
좋은 소식은
그것에 대한 쉬운 해결이
있을지도 모른다는 것이고,
그것은 우리의 눈을 더 사용하는 것이다.

— David McCandless —

DATA VISUALIZING

차트와 그래프는
단순히 분석을 위한 도구가 아니라
생각의 소통을 위한 전달체이며,
농담거리를 전달하는 매개이기도 하다.

— Nathan Yau —

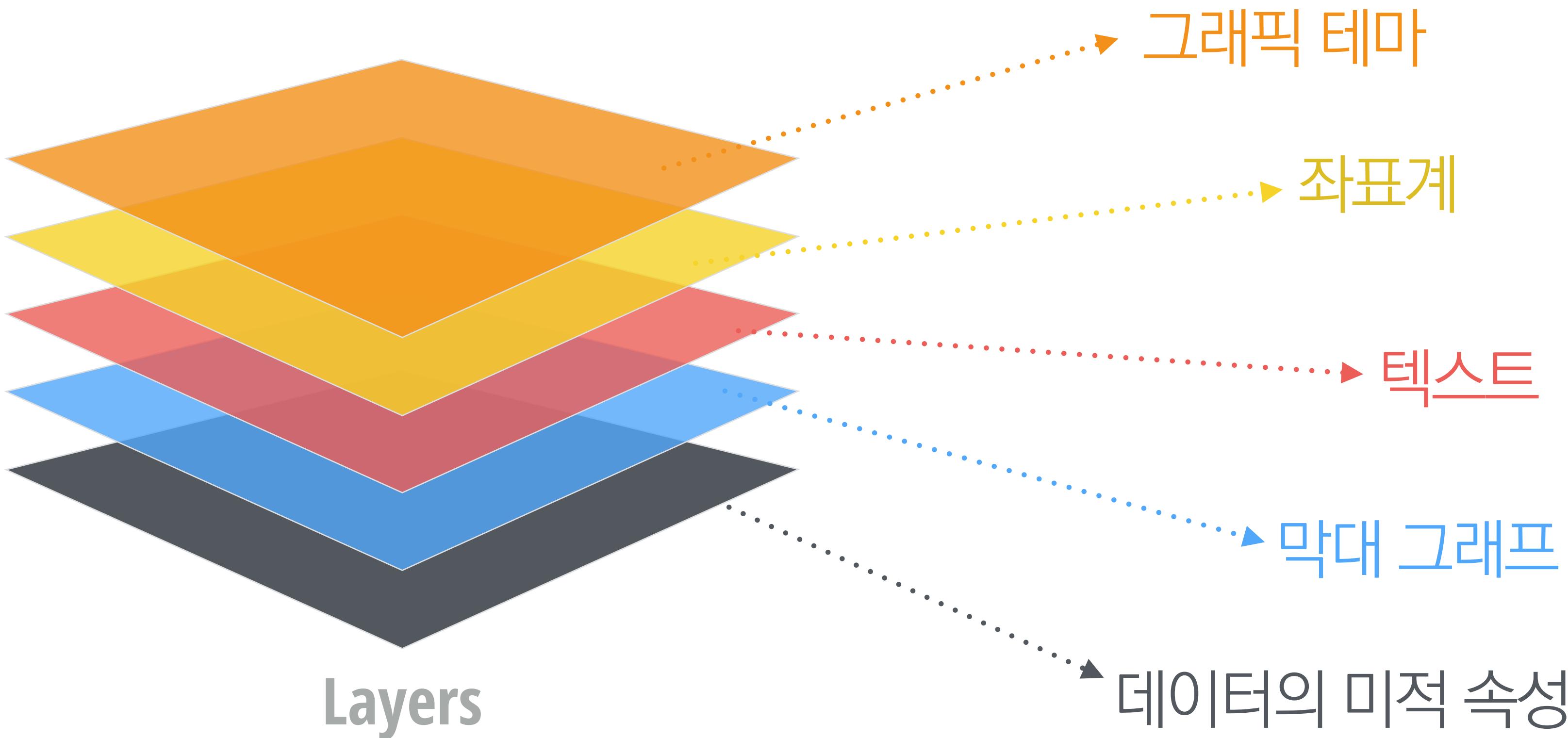


WITH R

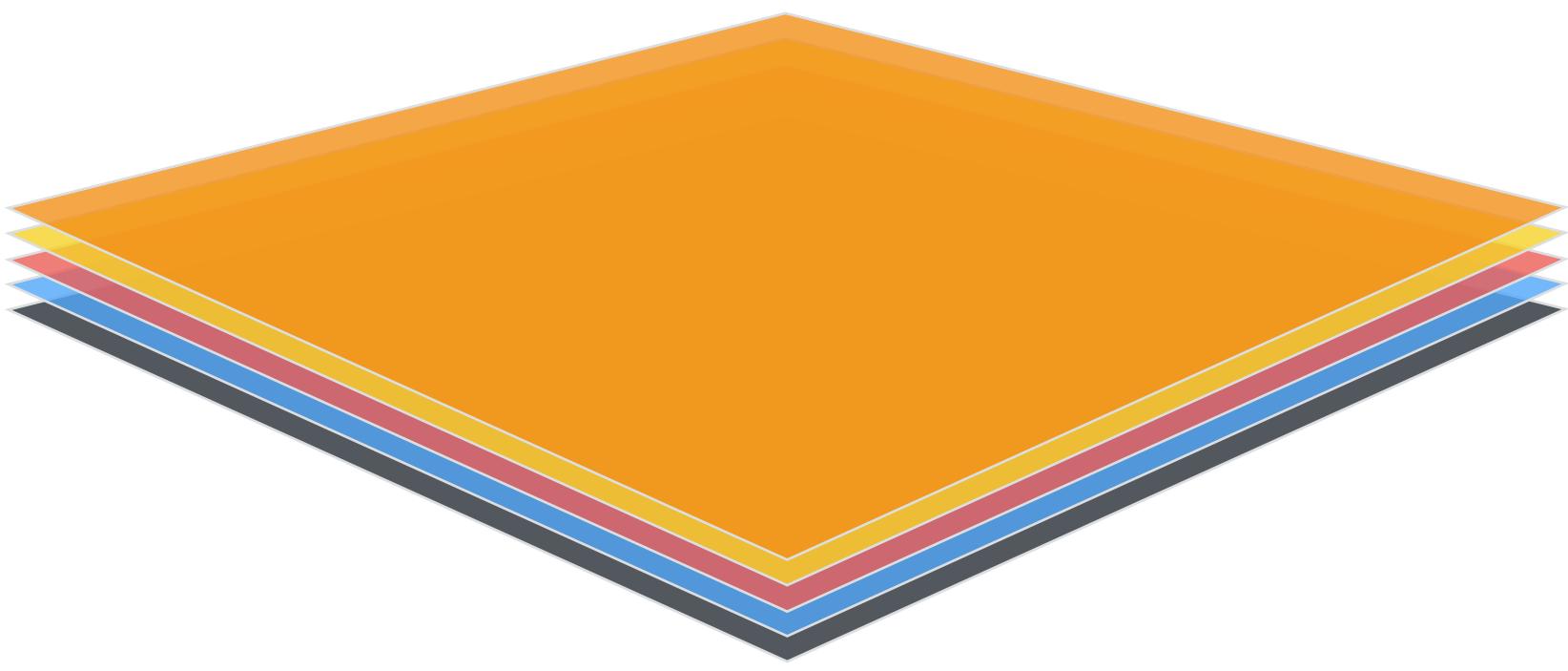
ggplot2

The Grammar of Graphics

ggplot2



ggplot2



NEW GRAPHIC OBJECT

ggplot2

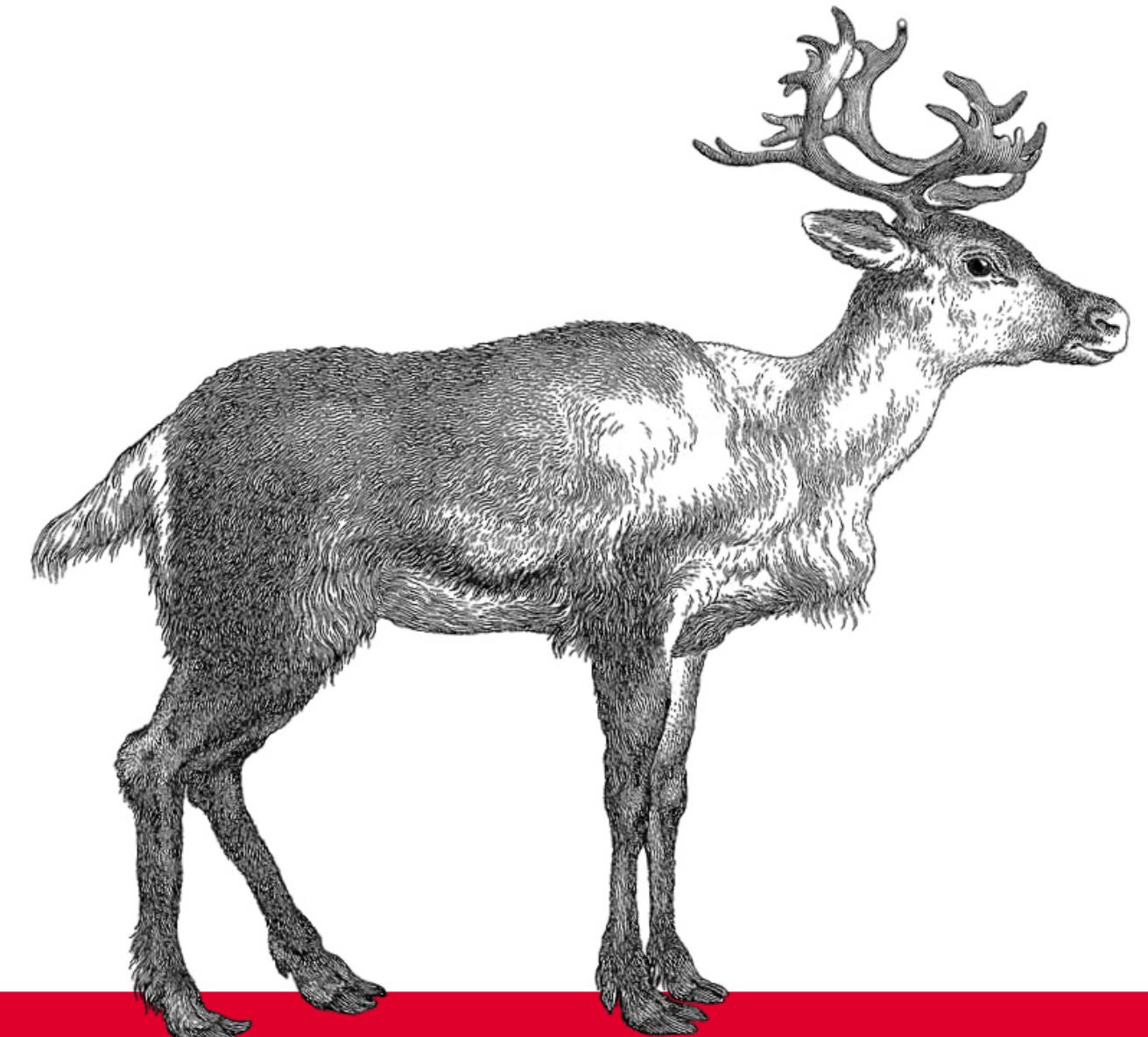
미려한 그래픽

재사용성

구조화된 코드

ggplot2

데이터 시각화를 위한 실용 레시피



R Graphics Cookbook

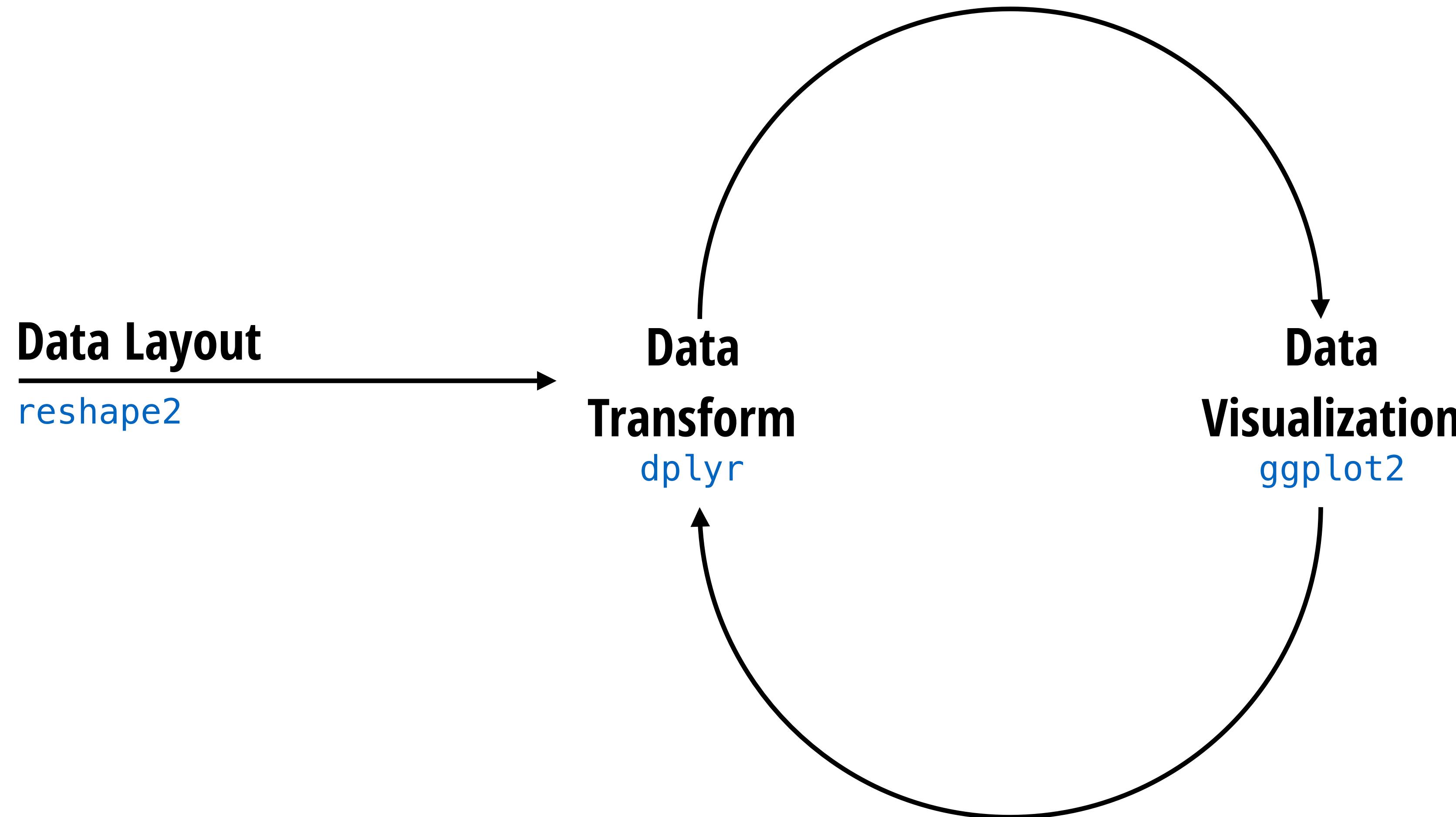
 ProgrammingInsight
O'REILLY® | 온사이트
insight

원스턴 챕 지음
이제원 옮김

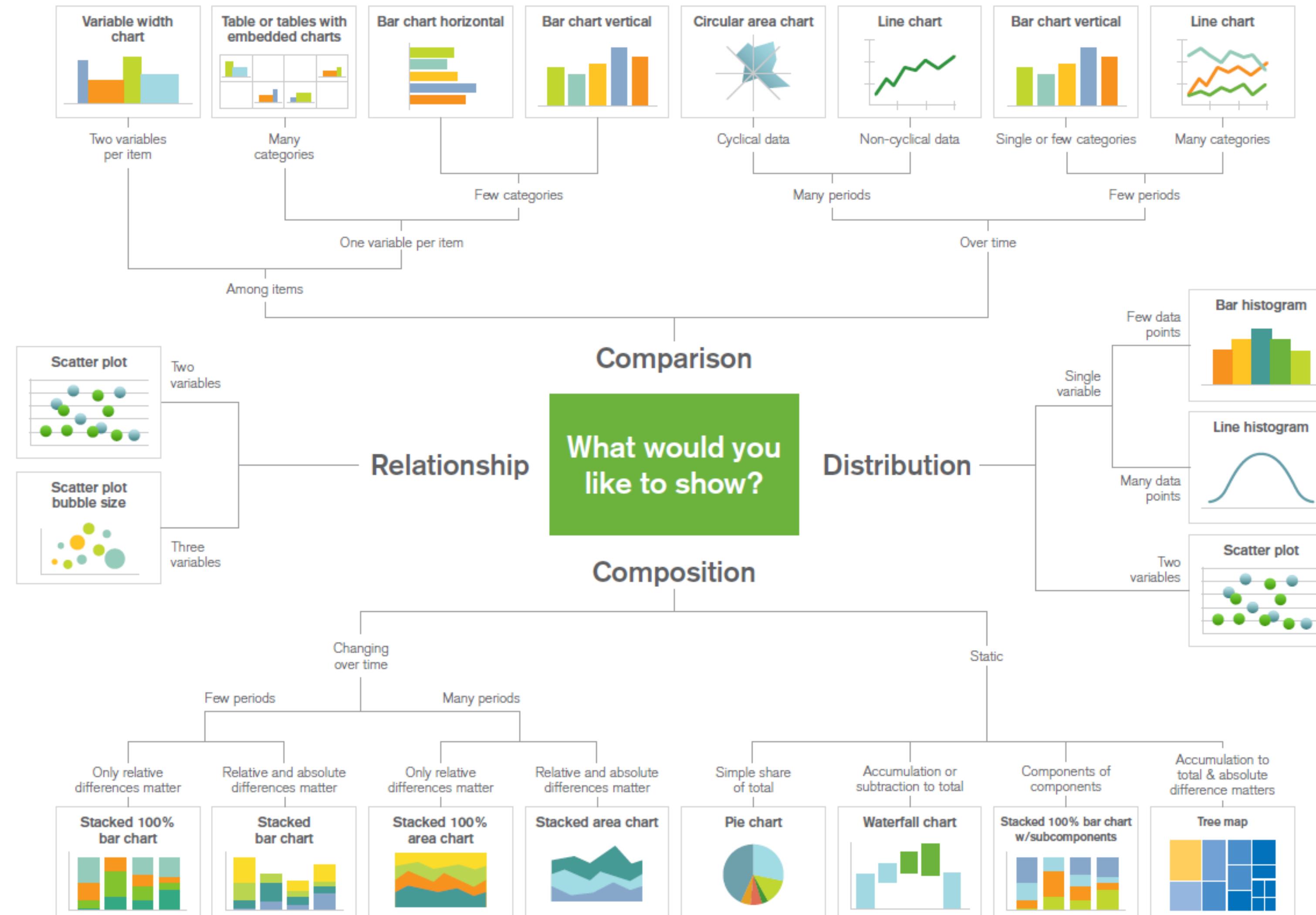
ggplot2

사용자가 원하는
미려한 시각화 결과물은
적합한 데이터에서 비롯된다.

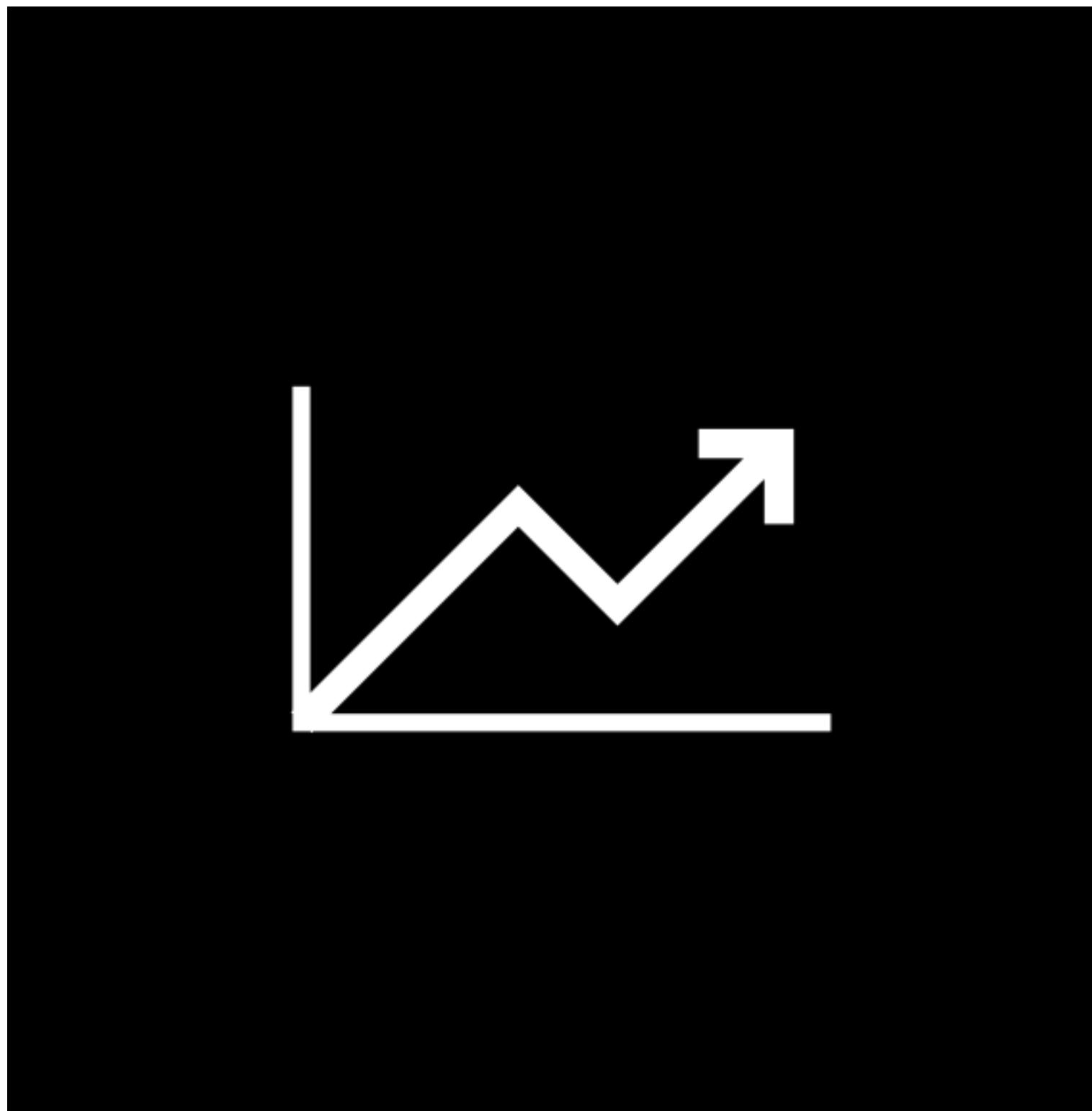
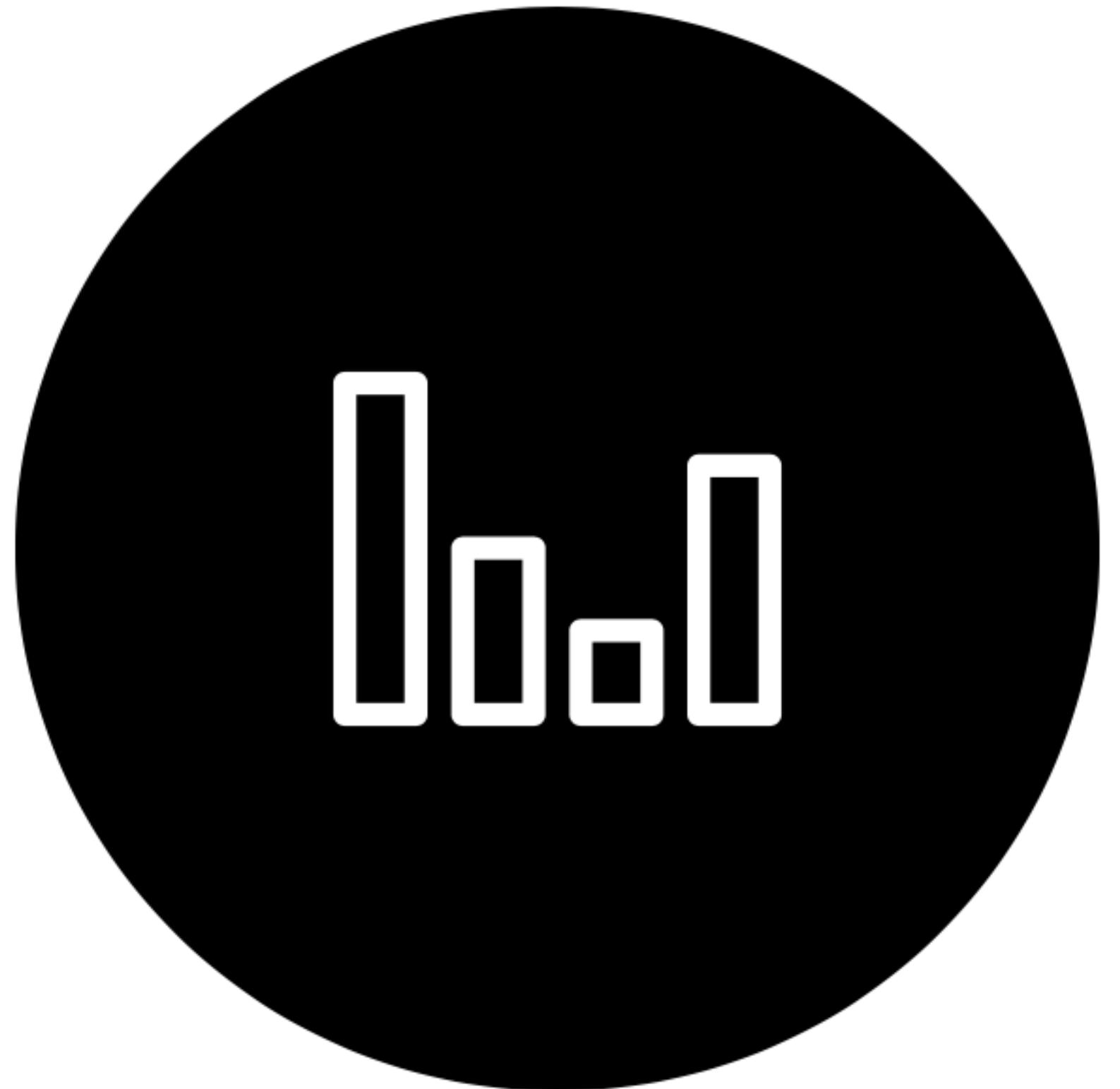
Ecosystem



ROADMAP



TODAY



DO IT

```
install.packages("ggplot2")
install.packages("gcookbook")
library(ggplot2)
library(gcookbook)
```

:)