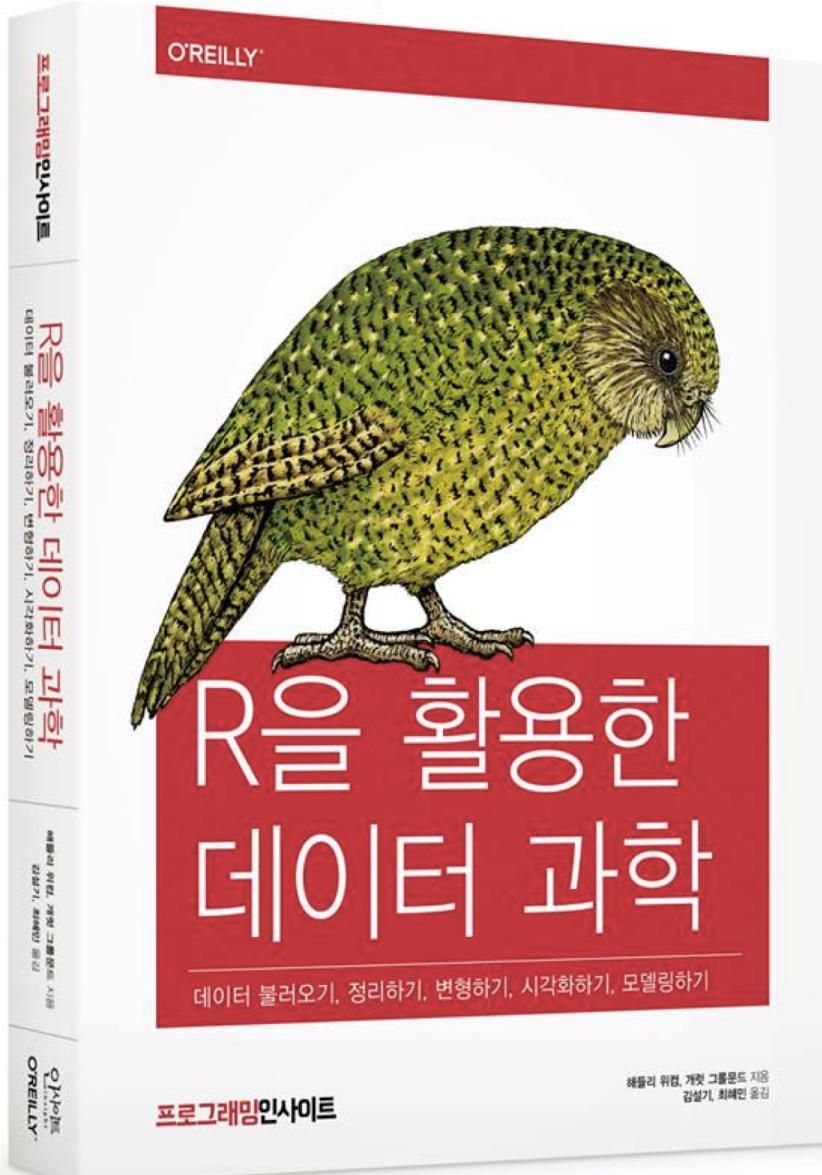


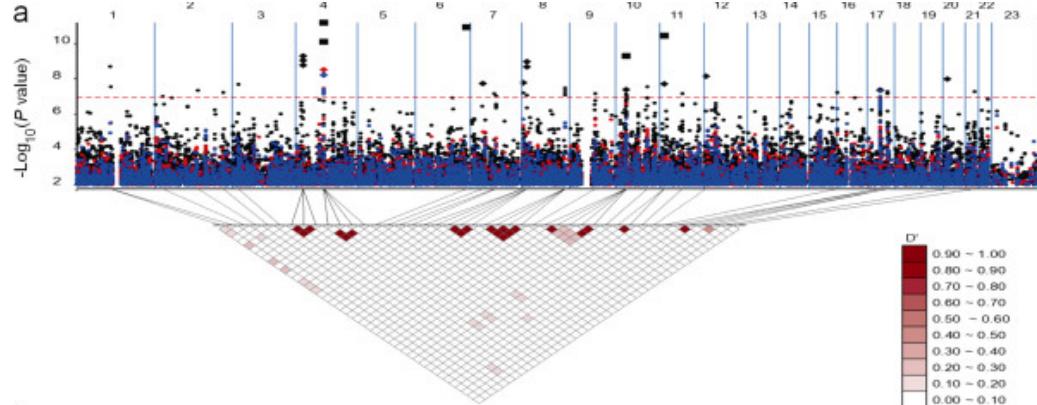
# R for Data Science

(번역서) “R을 활용한 데이터과학” 발간에 맞춰

김설기



# 유전통계학



# 에너지, IoT

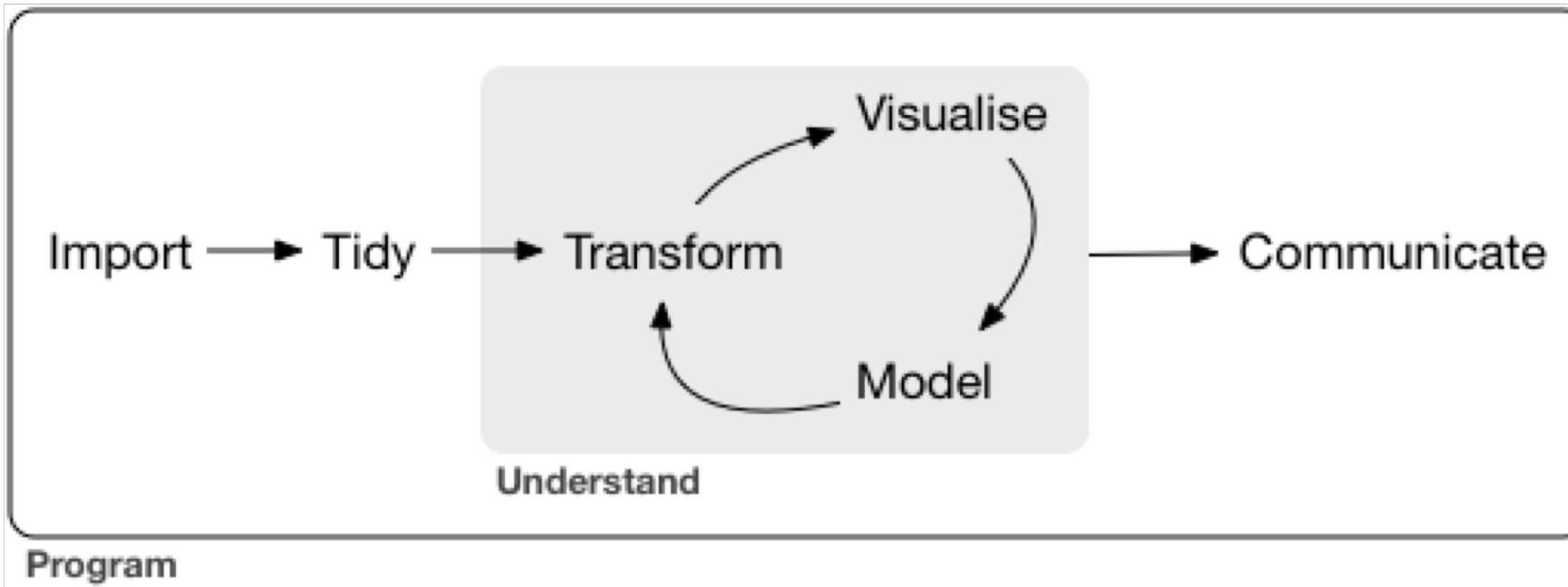


SAMSUNG  
SAMSUNG SDS



R 사용자 (2000년~)

이미지 출처: Smart Energy Efficiency



- 불러오기(Import)
- 정리하기, 타이디하게 하기(tidy)
- 변형하기(transform)
- 시각화하기(visualize)
- 모델링하기(model)
- 소통하기(communicate)

# Tidy 데이터

The diagram illustrates the process of transforming data from a wide, untidy format into a long, tidy format. It consists of three tables side-by-side, each with columns: country, year, cases, and population.

- 변수 (Variables):** The first table shows the data in a wide format where each country has multiple rows corresponding to different years. Each row contains the country name, year, cases, and population. Four vertical double-headed arrows point between the country, year, cases, and population columns, indicating they are variables grouped by country.
- 관측값 (Observations):** The second table shows the data in a long format where each row represents a single observation. The columns are country, year, cases, and population. Six horizontal double-headed arrows point between the country, year, cases, and population columns, indicating they are variables grouped by observation.
- 값 (Values):** The third table shows the data in a very long, flat format where each cell represents a single value. The columns are country, year, cases, and population. Each cell is represented by a black circle, indicating the data is now in a single, flat structure.

country	year	cases	population
Afghanistan	1999	745	1998071
Afghanistan	2000	2666	2059360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	128042583

country	year	cases	population
Afghanistan	1999	745	1998071
Afghanistan	2000	2666	2059360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	128042583

country	year	cases	population
Afghanistan	1999	745	1998071
Afghanistan	2000	2666	2059360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	128042583

# Tidy 하지 않은 데이터

```
table2  
  
#> # A tibble: 12 x 4  
  
#>   country     year type      count  
#>   <chr>       <int> <chr>      <int>  
#> 1 Afghanistan 1999 cases        745  
#> 2 Afghanistan 1999 population 19987071  
#> 3 Afghanistan 2000 cases       2666  
#> 4 Afghanistan 2000 population 20595360  
#> 5 Brazil       1999 cases       37737  
#> 6 Brazil       1999 population 172006362
```

```
table3  
  
#> # A tibble: 6 x 3  
  
#>   country     year rate  
#>   <chr>       <int> <chr>  
#> 1 Afghanistan 1999 745/19987071  
#> 2 Afghanistan 2000 2666/20595360  
#> 3 Brazil       1999 37737/172006362  
#> 4 Brazil       2000 80488/174504898  
#> 5 China        1999 212258/1272915272  
#> 6 China        2000 213766/1280428583
```

```
# 티블 두 개로 나누어짐  
  
table4a # cases  
  
#> # A tibble: 3 x 3  
#>   country     `1999` `2000`  
#>   * <chr>      <int>  <int>  
#> 1 Afghanistan    745   2666  
#> 2 Brazil          37737  80488  
#> 3 China           212258 213766  
  
table4b # population  
  
#> # A tibble: 3 x 3  
#>   country     `1999`     `2000`  
#>   * <chr>      <int>      <int>  
#> 1 Afghanistan  19987071  20595360  
#> 2 Brazil        172006362 174504898  
#> 3 China         1272915272 1280428583
```

# 데이터 사이언스 도구

## 1. 데이터프레임 기반: python pandas, spark dataframe

### **Announcement: DataFrame-based API is primary API**

**The MLlib RDD-based API is now in maintenance mode.**

As of Spark 2.0, the [RDD](#)-based APIs in the `spark.mllib` package have entered maintenance mode. The primary Machine Learning API for Spark is now the [DataFrame](#)-based API in the `spark.ml` package.

# 데이터 사이언스 도구

1. 데이터프레임 기반: python pandas, spark dataframe
2. pipe/pipeline

```
foo_foo <- hop(foo_foo, through = forest)
foo_foo <- scoop(foo_foo, up = field_mice)
foo_foo <- bop(foo_foo, on = head)
```

VS

```
foo_foo %>%
  hop(through = forest) %>%
  scoop(up = field_mice) %>%
  bop(on = head)
```

- 디버깅이 쉽다
- 동사(함수)가 주목된다
- 읽기가 쉽다

# 데이터 사이언스 도구

1. 데이터프레임 기반: python pandas, spark dataframe
2. pipe/pipeline
3. “많은 주제를 넓고 얕게 살펴보기다는 깊게 파면 더 빨리 할 수 있을 것이다.”

# 빅데이터와 R

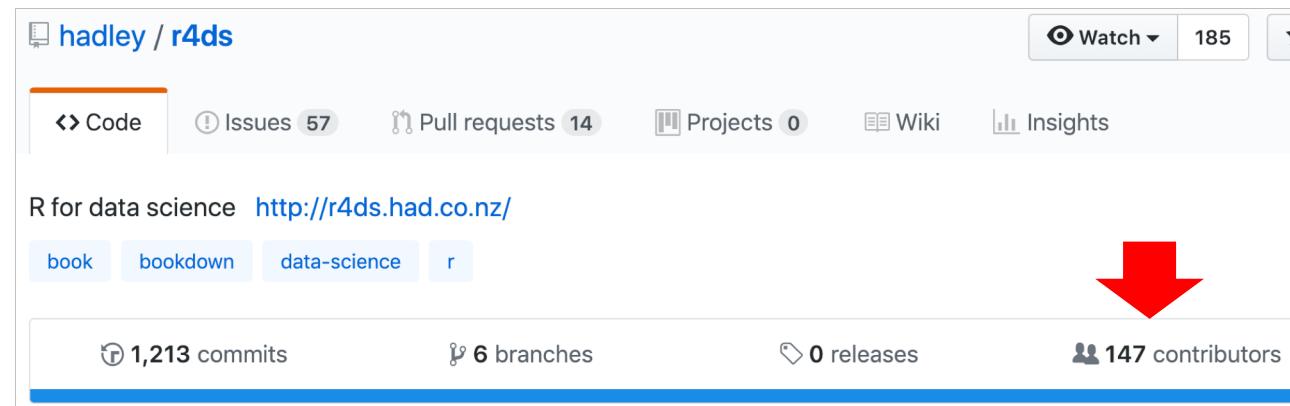
- "전체 데이터는 클 수도 있지만, 특정 문제에 답을 얻는데 필요 한 데이터는 작은 경우가 많다."
- data.table, sparklyr 등

# 데이터 과학 컬쳐

- Communication
  - "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
  - rmarkdown, bookdown

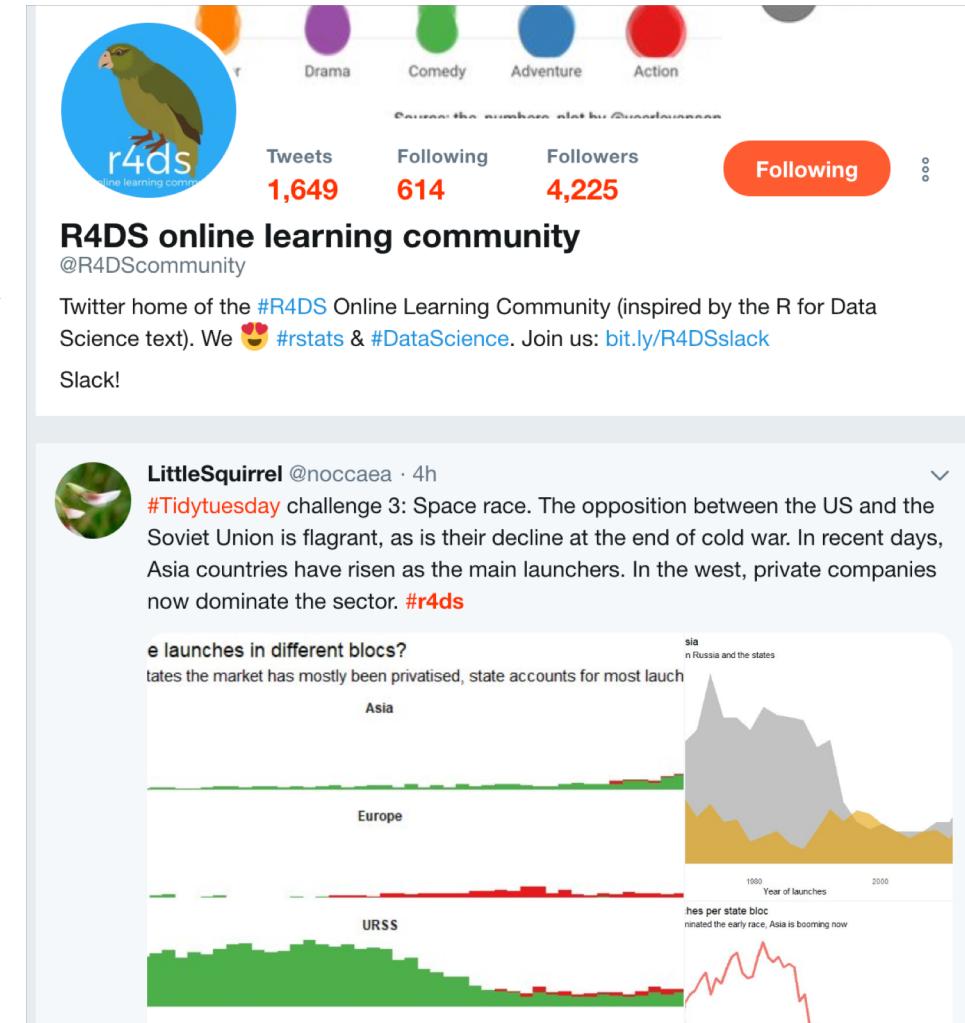
# 데이터 과학 컬쳐

- Communication
  - "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
  - rmarkdown, bookdown
- open project for book



# 데이터 과학 컬쳐

- Communication
  - "다른 사람에게 설명할 수 없다면 분석이 얼마나 훌륭한지는 중요하지 않다."
  - rmarkdown, bookdown
- open project for book
- twitter #rstats #r4ds #TidyTuesday



# R 도움받기 – 문제 발생시

- 구글 검색
  - 오류메시지
  - 'R' 추가
- stackoverflow



# R 도움받기 – 문제예방

- 블로그 (RStudio, R-bloggers)
- 트위터 (@hadleywickham, @statgarrett, @rstudiotips, #rstats, #r4ds) 팔로우

# R 도움받기 – 한국어

- 페이스북 그룹
- 온라인, 오프라인 강의, 멋업
- 책
- <https://sulgik.github.io/r4ds/>