
# Apply函数家族与数据操纵

史冬波

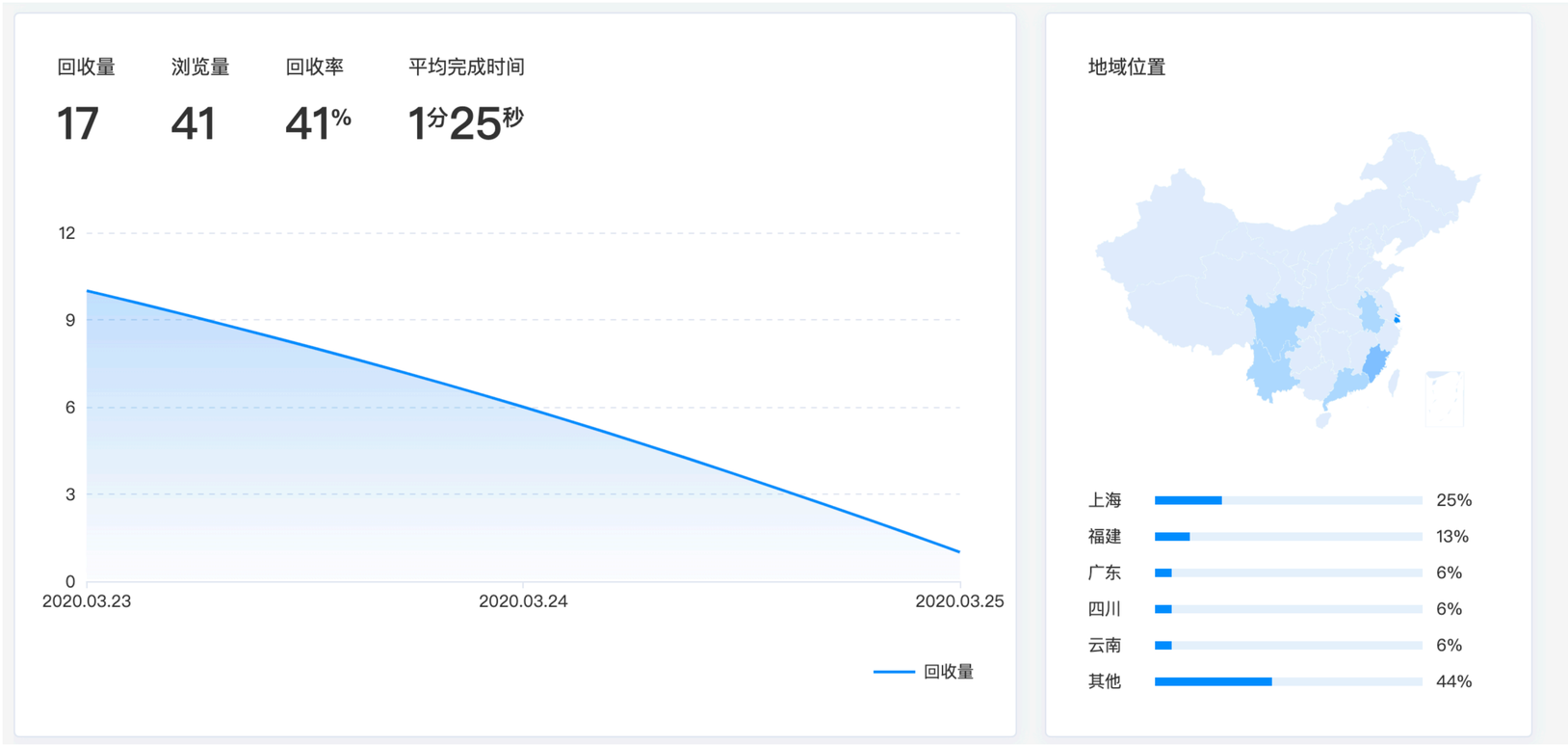
2020年3月30日

# 感谢大家使用Canvas

- 31个讨论话题、145条讨论
- 我们的学习社区正在形成
- 这是一种必将终身受益的学习习惯，开源精神
- 所有提问、回答、点赞都是对社区的贡献
- 从经济学来看，正外部性、同侪效应、激励反馈、learning by answering/ asking

# 感谢大家反馈教学调研

- 帮助我及时调整教学节奏
- 我要你觉得，不是我觉得
- 把科研案例融入教学当中
- 从教学中丰富知识的理解
- 欲速则不达



# 答疑

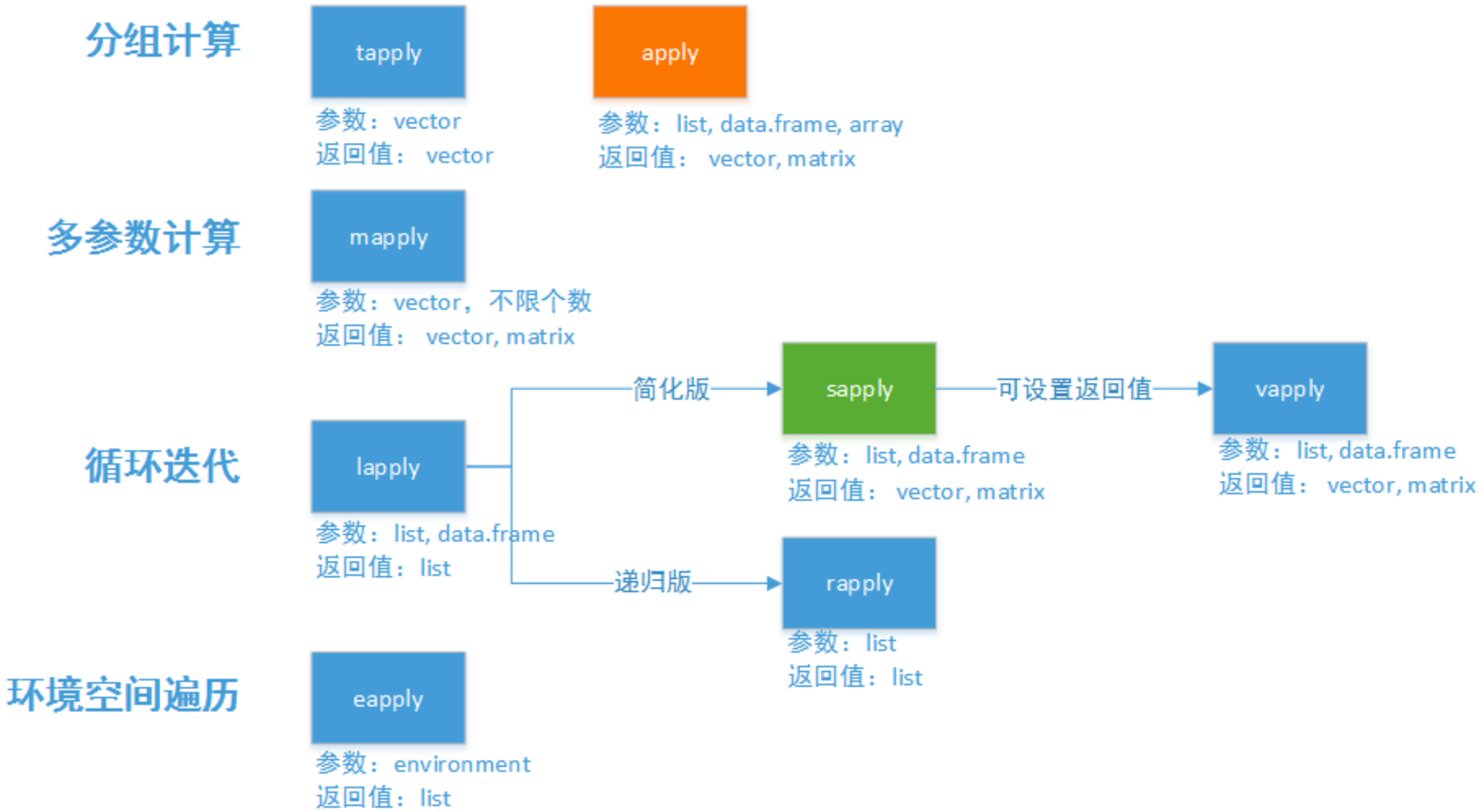
- str_detect不理解, 马上就讲到了
- 代码运行的问题（江）
- csv导入到excel修改编码方式
- .R文件是脚本文件，.RData文件是数据文件,有不同的打开方式

# Apply 函数家族

参考资料：http://blog.fens.me/r-apply/

# 为什么要用Apply函数

- for, while 循环效率低
- R内置函数，效率高
- for是串行，apply是并行
- 不太好理解，尤其是记不住用法
- 记不住就记不住吧！



# apply-matrix，dataframe

- 按行按列进行循环计算
- apply(X, Margin, Fun,...)
  - X: 数组、矩阵、数据框
  - Margin：1表示按行计算，2表示按列计算
  - Fun：使用内置函数或自定义函数

# lapply-list,dataframe

- 用来对list、data.frame数据集进行循环，并返回和X长度同样的list结构作为结果集
- lapply(X, Fun,...)
  - X: 数组、矩阵、数据框
  - Fun：使用内置函数或自定义函数

# sapply-list,dataframe

- Simple lapply, 增加了2个参数simplify和USE.NAMES
- lapply(X, Fun,...,simplify=TRUE, USE.NAMES = TRUE)
  - X: 数组、矩阵、数据框
  - Fun：使用内置函数或自定义函数
  - simplify: 是否数组化，当值array时，输出结果按数组进行分组
  - USE.NAMES: 如果X为字符串，TRUE设置字符串为数据名，FALSE不设置

# vapply-list,dataframe

- vapply类似于sapply，提供了FUN.VALUE参数，用来控制返回值的行名，这样可以让程序更健壮
- vapply(X, FUN, FUN.VALUE, ..., USE.NAMES = TRUE)
  - X: 数组、矩阵、数据框
  - Fun：使用内置函数或自定义函数
  - FUN.VALUE: 定义返回值的行名row.names
  - USE.NAMES: 如果X为字符串，TRUE设置字符串为数据名，FALSE不设置

# mapply-list,dataframe

- mapply也是sapply的变形函数，类似多变量的sapply，但是参数定义有些变化
- mapply(FUN, ...,MoreArgs = NULL, SIMPLIFY = TRUE,USE.NAMES = TRUE)
  - ~~X: 数组、矩阵、数据框~~
  - Fun：使用内置函数或自定义函数
  - MoreArgs: 参数列表
  - SIMPLIFY: 是否数组化，当值array时，输出结果按数组进行分组
  - USE.NAMES: TRUE设置字符串为数据名，FALSE不设置

# 其他

- tapply
- rapply
- eapply # environment

# Tidyr

参考资料：bit.ly/wrangling-webinar

# 数据操纵

- 在第一单元，我们学习了R语言的基本知识
- 本单元，我们将学习如何清理数据
- 数据清理工作往往占到一个数据工程（除数据采集外）的50%-80%的工作量。
  - 易燃易爆炸
- 根据研究需求将数据清理、合并、变形，生成可用于下一步分析的数据
- 有时候，数据清理的下一步是作图。但大部分情况，数据清理的目标是生成一个可用于回归分析的dataframe
- 那什么样的数据是可用于回归分析的数据？

# TIDY DATA

- 每列储存且仅仅储存一个变量(variable)
- 每行储存且仅仅储存一个观测值(observation)
- 每一种信息储存在独立的表中，表与表之间有关联列
  - variable 对应的是回归中的变量
  - observation 对应回归单元，unit of analysis
  - 例如：研究下程序语言设计课程成绩与毕业第一份工作收入之间的关系

```
# devtools::install_github("rstudio/EDAWR")
library(EDAWR)
```

## storms

| storm | wind | pressure | date |
|---|---|---|---|
| Alberto | 110 | 1007 | 2000-08-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Allison | 65 | 1005 | 1995-06-04 |
| Ana | 40 | 1013 | 1997-07-01 |
| Arlene | 50 | 1010 | 1999-06-13 |
| Arthur | 45 | 1010 | 1996-06-21 |

## cases

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| **FR** | 7000 | 6900 | 7000 |
| **DE** | 5800 | 6000 | 6200 |
| **US** | 15000 | 14000 | 13000 |

## pollution

| city | particle size | amount ($\mu g/m^3$) |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

```
# devtools::install_github("rstudio/EDAWR")
library(EDAWR)
```

## storms

| storm | wind | pressure | date |
|---|---|---|---|
| Alberto | 110 | 1007 | 2000-08-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Allison | 65 | 1005 | 1995-06-04 |
| Ana | 40 | 1013 | 1997-07-01 |
| Arlene | 50 | 1010 | 1999-06-13 |
| Arthur | 45 | 1010 | 1996-06-21 |

- Storm name
- Wind Speed (mph)
- Air Pressure
- Date

## cases

| Country | | | |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14900 | 13000 |

- Country
- Year
- Count

## pollution

| city | particle size | amount ($\mu g/m^3$) |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

- City
- Amount of large particles
- Amount of small particles

```
# devtools::install_github("rstudio/EDAWR")
library(EDAWR)
```

## storms

| storm | wind | pressure | date |
|---|---|---|---|
| Alberto | 110 | 1007 | 2000-08-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Allison | 65 | 1005 | 1995-06-04 |
| Ana | 40 | 1013 | 1997-07-01 |
| Arlene | 50 | 1010 | 1999-06-13 |
| Arthur | 45 | 1010 | 1996-06-21 |

```
storms$storm
storms$wind
storms$pressure
storms$date
```

## cases

| Country | | | |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

```
cases$country
names(cases)[-1]
unlist(cases[1:3, 2:4])
```

## pollution

| city | particle size | amount ($\mu g/m^3$) |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

```
pollution$city[1,3,5]
pollution$amount[1,3,5]
pollution$amount[2,4,6]
```

# 如何把cases变成tidy data

cases

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |

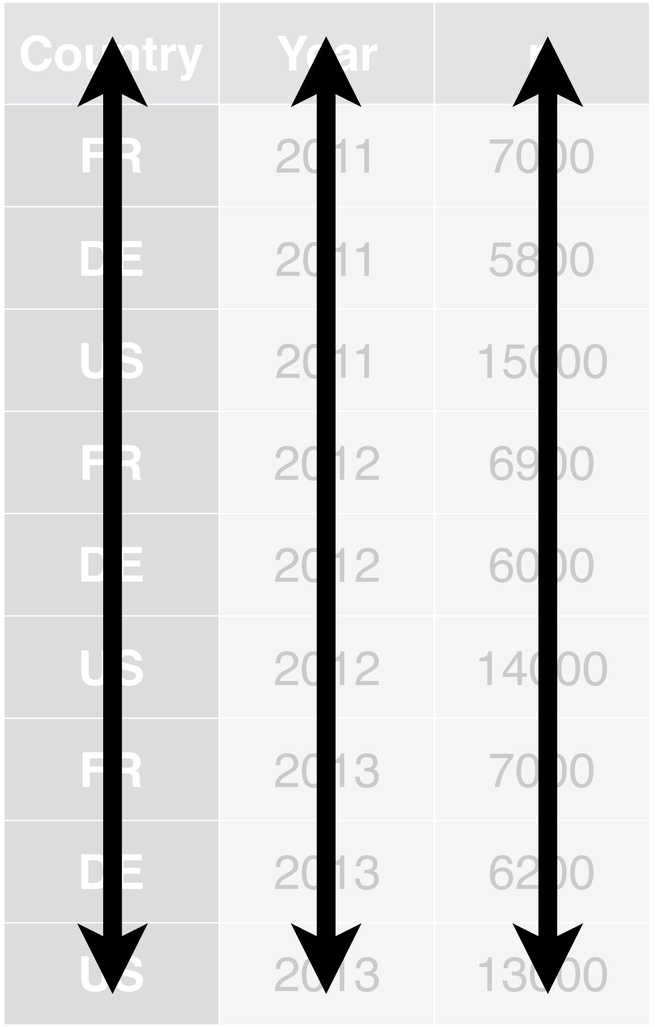
| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |
| US | 2013 | 13000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |



| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

宽数据


| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |
| US | 2013 | 13000 |

长数据

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |
| US | 2013 | 13000 |

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

**key** (former column names)

| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |
| US | 2013 | 13000 |

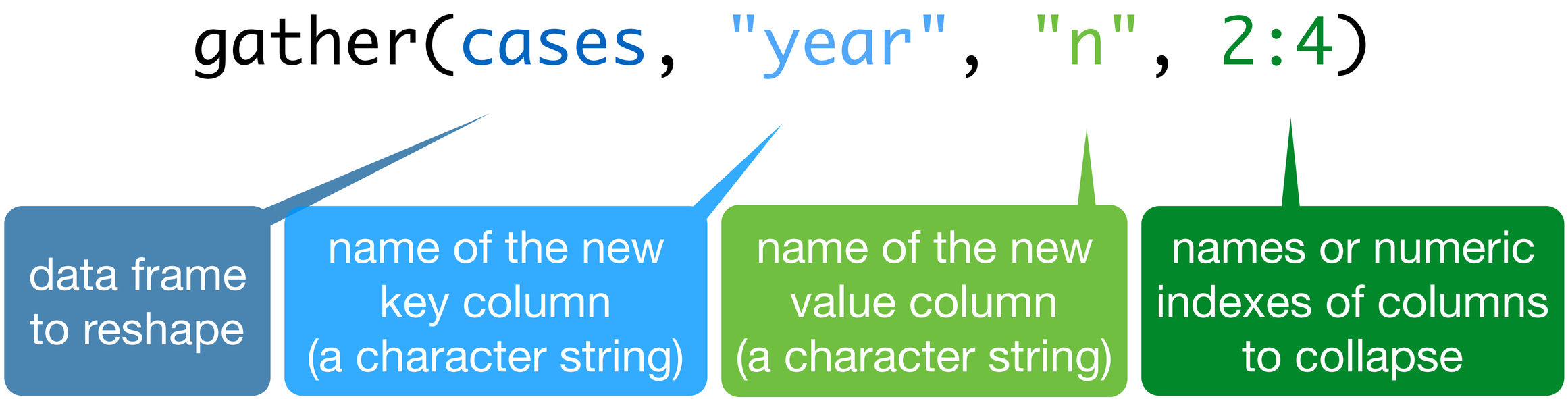
key **value** (former cells)

| Country | 2011 | 2012 | 2013 |
|---|---|---|---|
| FR | 7000 | 6900 | 7000 |
| DE | 5800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

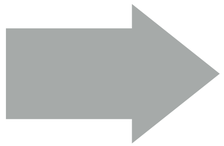
| Country | Year | n |
|---|---|---|
| FR | 2011 | 7000 |
| DE | 2011 | 5800 |
| US | 2011 | 15000 |
| FR | 2012 | 6900 |
| DE | 2012 | 6000 |
| US | 2012 | 14000 |
| FR | 2013 | 7000 |
| DE | 2013 | 6200 |
| US | 2013 | 13000 |

# gather()

- gather(cases, "year", "n", 2:4) # 可以实现这个功能
- key储存变量名，value储存值



```
##   country  2011  2012  2013
## 1      FR  7000  6900  7000
## 2      DE  5800  6000  6200
## 3      US 15000 14000 13000
```

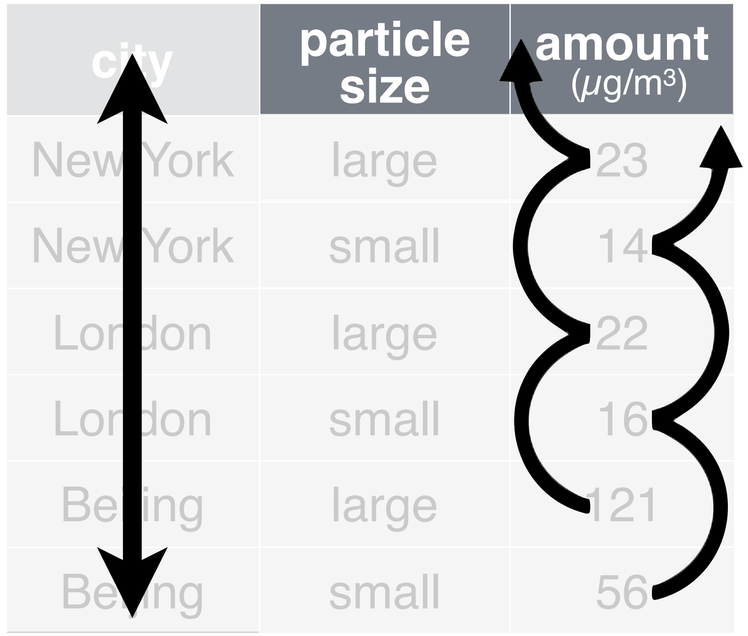


```
##    country year     n
## 1       FR 2011  7000
## 2       DE 2011  5800
## 3       US 2011 15000
## 4       FR 2012  6900
## 5       DE 2012  6000
## 6       US 2012 14000
## 7       FR 2013  7000
## 8       DE 2013  6200
## 9       US 2013 13000
```

```
gather(cases, "year", "n", 2:4)
```

pollution

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |



| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |


| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

**key** (new column names)

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

key **value** (new cells)

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |

| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

# spread()

Generates multiple columns from two columns:

1. each unique value in the **key** column becomes a column name
2. each value in the **value** column becomes a cell in the new columns

```
spread(pollution, size, amount)
```

# spread()

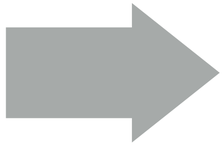
```
spread(pollution, size, amount)
```

data frame to reshape

column to use for keys (new columns names)

column to use for values (new column cells)

```
##       city  size amount
## 1 New York large     23
## 2 New York small     14
## 3   London large     22
## 4   London small     16
## 5  Beijing large    121
## 6  Beijing small     56
```



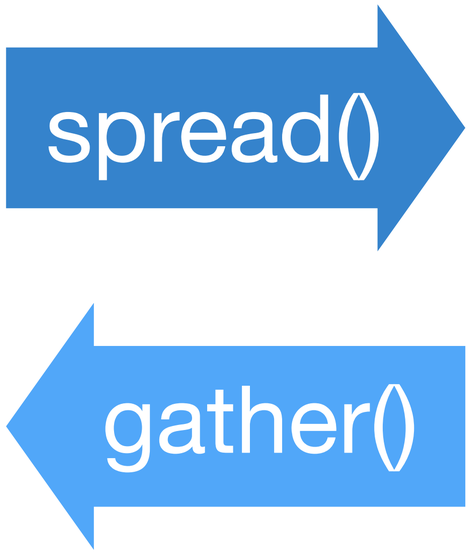
```
##       city large small
## 1  Beijing   121    56
## 2   London    22    16
## 3 New York    23    14
```

```
spread(pollution, size, amount)
```

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |


| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

| city | size | amount |
|---|---|---|
| New York | large | 23 |
| New York | small | 14 |
| London | large | 22 |
| London | small | 16 |
| Beijing | large | 121 |
| Beijing | small | 56 |




| city | large | small |
|---|---|---|
| New York | 23 | 14 |
| London | 22 | 16 |
| Beijing | 121 | 56 |

# separate()

```
separate(storms, date, c("year", "month", "day"), sep = "-")
```

storms

| storm | wind | pressure | date |
|---|---|---|---|
| Alberto | 110 | 1007 | 2000-08-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Allison | 65 | 1005 | 1995-06-04 |
| Ana | 40 | 1013 | 1997-07-01 |
| Arlene | 50 | 1010 | 1999-06-13 |
| Arthur | 45 | 1010 | 1996-06-21 |

storms2

| storm | wind | pressure | year | month | day |
|---|---|---|---|---|---|
| Alberto | 110 | 1007 | 2000 | 08 | 12 |
| Alex | 45 | 1009 | 1998 | 07 | 30 |
| Allison | 65 | 1005 | 1995 | 06 | 04 |
| Ana | 40 | 1013 | 1997 | 07 | 1 |
| Arlene | 50 | 1010 | 1999 | 06 | 13 |
| Arthur | 45 | 1010 | 1996 | 06 | 21 |

# unite()

```
unite(storms2, "date", year, month, day, sep = "-")
```

storms2

| storm | wind | pressure | year | month | day |
|---|---|---|---|---|---|
| Alberto | 110 | 1007 | 2000 | 08 | 12 |
| Alex | 45 | 1009 | 1998 | 07 | 30 |
| Allison | 65 | 1005 | 1995 | 06 | 04 |
| Ana | 40 | 1013 | 1997 | 07 | 1 |
| Arlene | 50 | 1010 | 1999 | 06 | 13 |
| Arthur | 45 | 1010 | 1996 | 06 | 21 |

storms

| storm | wind | pressure | date |
|---|---|---|---|
| Alberto | 110 | 1007 | 2000-08-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Allison | 65 | 1005 | 1995-06-04 |
| Ana | 40 | 1013 | 1997-07-01 |
| Arlene | 50 | 1010 | 1999-06-13 |
| Arthur | 45 | 1010 | 1996-06-21 |