

**Data sets
come in many
formats**

...but R prefers just one.

Tidy data

storms

| storm | wind | pressure | date |
|---------|------|----------|------------|
| Alberto | 110 | 1007 | 2000-07-12 |
| Alex | 45 | 1009 | 1998-07-30 |
| Anson | 65 | 1005 | 1995-07-04 |
| Ava | 40 | 1013 | 1997-07-01 |
| Annie | 30 | 1010 | 1999-07-13 |
| Arthur | 45 | 1010 | 1996-07-21 |

- 1 Each **variable** is saved in its own **column**.
- 2 Each **observation** is saved in its own **row**.
- 3 Each "type" of observation stored in a **single table** (here, storms).

tidyr

| 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 |

gather()

| 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 | Year | Revenue |
|---------|------|---------|
| 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 | 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 (former column names)

| 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 |

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()

Collapses multiple columns into two columns:

1. a **key** column that contains the former column names
2. a **value** column that contains the former column cells

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

gather()

Collapses multiple columns into two columns:

1. a **key** column that contains the former column names
2. a **value** column that contains the former column cells

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

data frame
to reshape

gather()

Collapses multiple columns into two columns:

1. a **key** column that contains the former column names
2. a **value** column that contains the former column cells

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

data frame
to reshape

name of the new
key column
(a character string)

gather()

Collapses multiple columns into two columns:

1. a **key** column that contains the former column names
2. a **value** column that contains the former column cells

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

data frame
to reshape

name of the new
key column
(a character string)

name of the new
value column
(a character string)

gather()

Collapses multiple columns into two columns:

1. a **key** column that contains the former column names
2. a **value** column that contains the former column cells

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

data frame
to reshape

name of the new
key column
(a character string)

name of the new
value column
(a character string)

names or numeric
indexes of columns
to collapse

| 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()

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)
```

data frame
to reshape

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)
```

data frame
to reshape

column to use for
keys (new columns
names)

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)
```

data frame
to reshape

column to use for
keys (new columns
names)

column to use for
values (new
column 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 |

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

`spread()`

`gather()`

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

unite() and separate()

There are three more variables hidden in storms:

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 |

- Year
- Month
- Day

separate()

Separate splits a column by a character string separator.

```
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 unites columns into a single column.

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

storms2

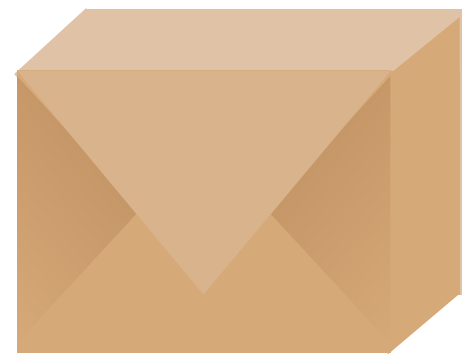
| storm | wind | pressure | year | month | day |
|---------|------|----------|------|-------|-----|
| Alberto | 110 | 1007 | 2000 | 08 | 12 |
| Alex | 45 | 1009 | 1998 | 07 | 30 |
| Allison | 65 | 1005 | 1995 | 06 | 04 |
| Ana | 40 | 1013 | 1997 | 07 | 1 |
| Arlene | 50 | 1010 | 1999 | 06 | 13 |
| Arthur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



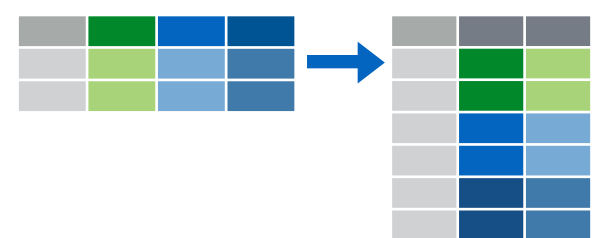
storms

| 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 |

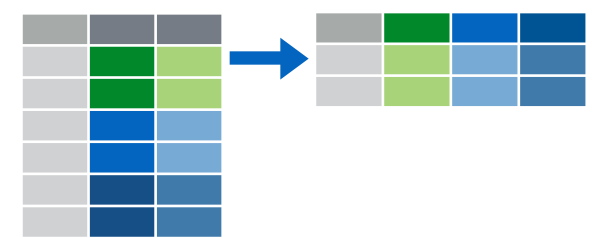
Recap: tidyr



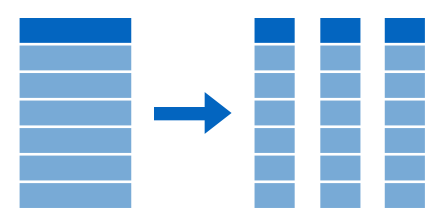
A package that reshapes the layout of data sets.



Make observations from variables with `gather()`



Make variables from observations with `spread()`



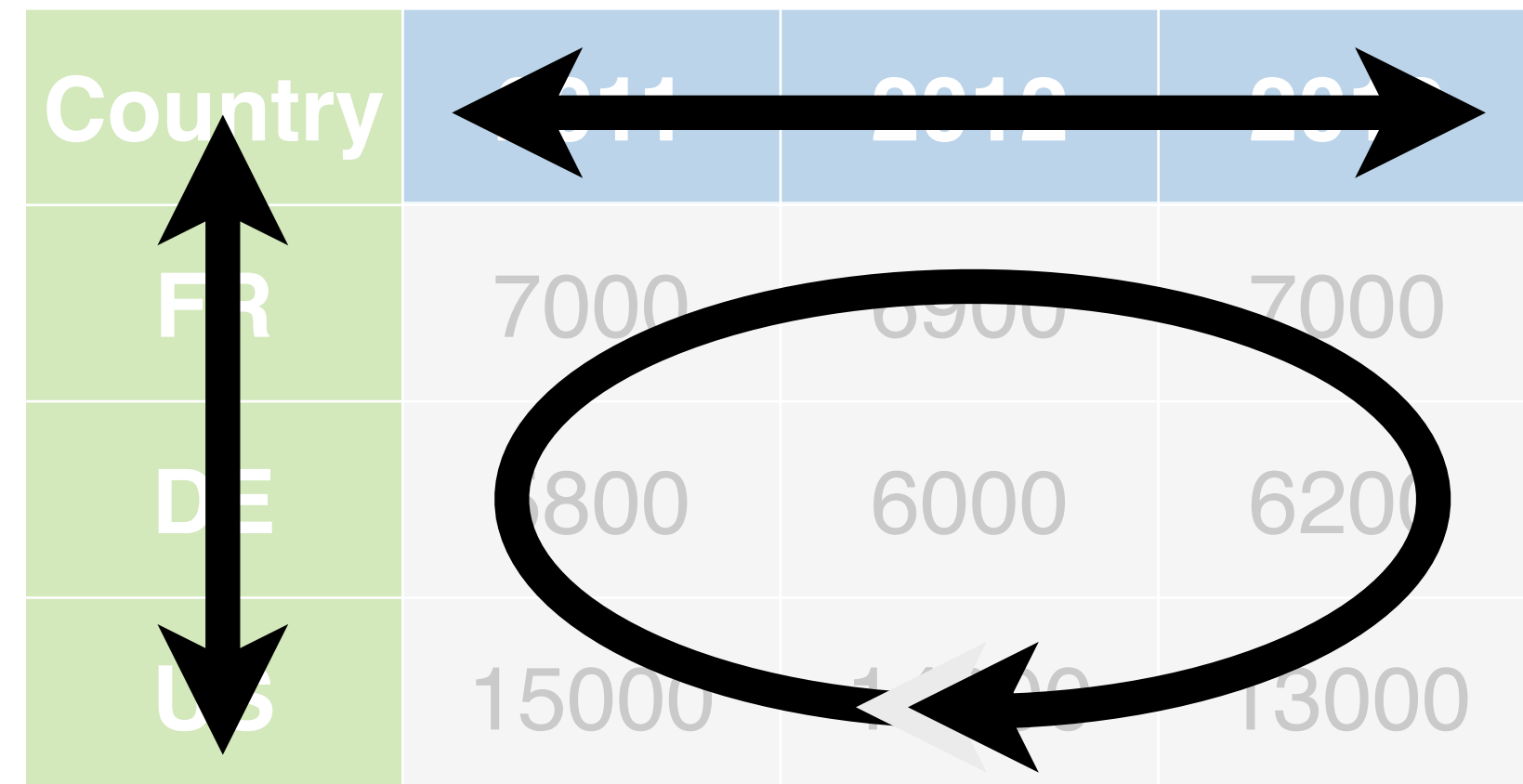
Split and merge columns with `unite()` and `separate()`

Data Wrangling with R

How to work with the
structures of your data

Slides at:

bit.ly/wrangling-webinar



The diagram shows a table with columns for years (2011, 2012, 2013) and rows for countries (FR, DE, US). A vertical double-headed arrow on the left indicates row operations, and a horizontal double-headed arrow on top indicates column operations. A large oval with arrows at the top and bottom indicates a loop or transformation across the data.

| Country | 2011 | 2012 | 2013 |
|---------|-------|-------|-------|
| FR | 7000 | 6900 | 7000 |
| DE | 6800 | 6000 | 6200 |
| US | 15000 | 14000 | 13000 |

Garrett Grolemond

Data Scientist and Master Instructor
January 2015
Email: garrett@rstudio.com