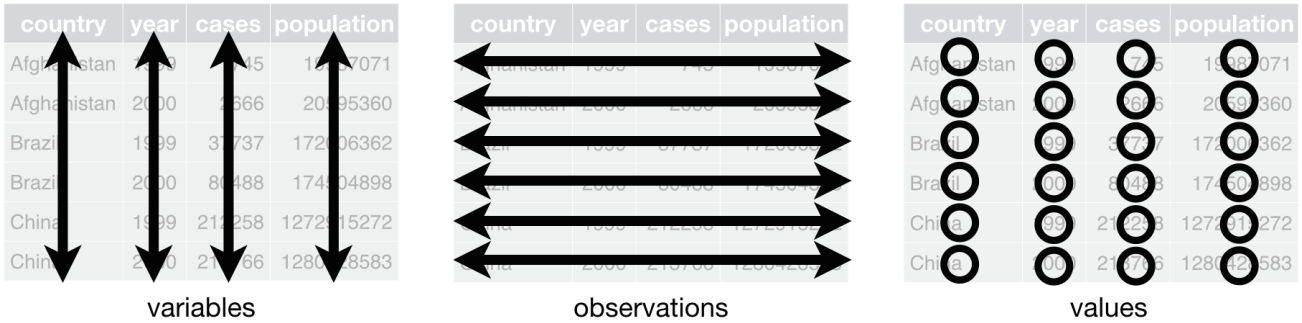
Lesson 10: Tidy data

Learning objectives

- Describe the concept of tidy data
- Determine whether a dataset is in tidy format
- Use tidyr::pivot_wider() and tidyr::pivot_longer() to reshape data frames
- Use tidyr::unite() and tidyr::separate() to merge or separate information from different columns

Three inter-related rules that make a dataset tidy:

- Each variable is a column; each column is a variable.
- Each observation is a row; each row is an observation.
- Each value is a cell; each cell is a single value.



Different ways to display the same data

Which structure is tidy?

	country	year	cases	population
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

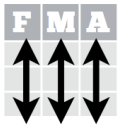
	country	year	type	count
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases	212258
10	China	1999	population	1272915272
11	China	2000	cases	213766
12	China	2000	population	1280428583

	country	year	rate
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

	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

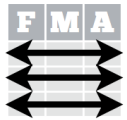
	country	1999	2000
	row names		
1	Afghanistan	19987071	20595360
2	Brazil	172006362	174504898
3	China	1272915272	1280428583

In a tidy data set:



Each **variable** is saved in its own **column**

&



Each **observation** is saved in its own **row**

Different ways to display the same data

Tidy data

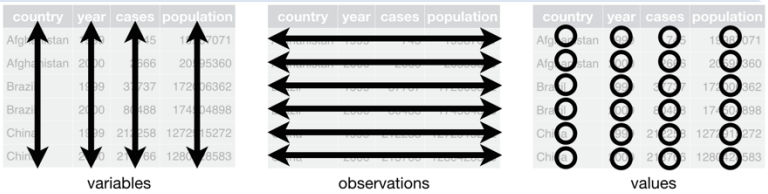
	country	year	cases	population
1	Afghanistan	1999	745	19987071
2	Afghanistan	2000	2666	20595360
3	Brazil	1999	37737	172006362
4	Brazil	2000	80488	174504898
5	China	1999	212258	1272915272
6	China	2000	213766	1280428583

	country	year	type	count
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
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases	212258
10	China	1999	population	1272915272
11	China	2000	cases	213766
12	China	2000	population	1280428583

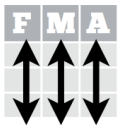
	country	year	rate
1	Afghanistan	1999	745/19987071
2	Afghanistan	2000	2666/20595360
3	Brazil	1999	37737/172006362
4	Brazil	2000	80488/174504898
5	China	1999	212258/1272915272
6	China	2000	213766/1280428583

	country	1999	2000
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

	country	1999	2000
1	Afghanistan	19987071	20595360
2	Brazil	172006362	174504898
3	China	1272915272	1280428583

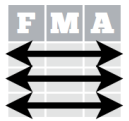


In a tidy data set:



Each **variable** is saved in its own **column**

&



Each **observation** is saved in its own **row**

Exercise

Compute the rate for table2, and table4a + table4b. You will need to perform four operations:

- Extract the number of TB cases per country per year.
- Extract the matching population per country per year.
- Divide cases by population, and multiply by 10000.
- Store back in the appropriate place.

Which representation is easiest to work with? Which is hardest? Why?

If I had one thing to tell biologists learning bioinformatics, it would be “write code for humans, write data for computers”
— Vince Buffalo (@vsbuffalo)

Common problems

- One variable might be spread across multiple columns and sometimes values have ended up in column names

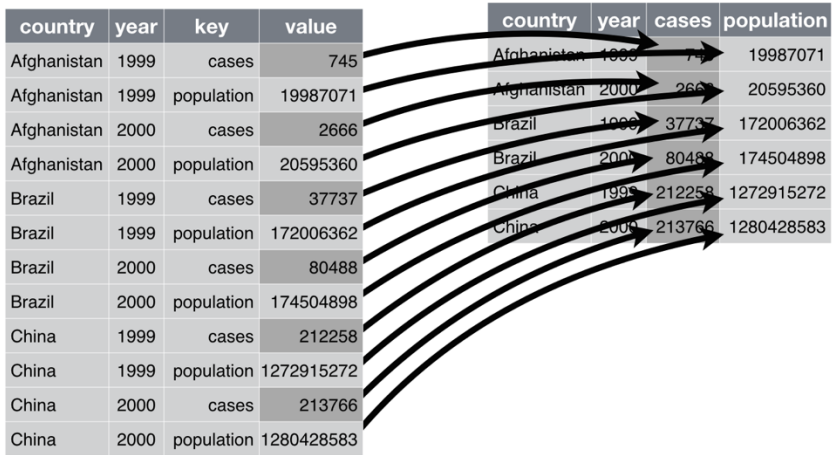
- One observation is spread across multiple rows

pivot_longer()

```
table4a %>%
  pivot_longer(c(`1999`, `2000`), names_to = "year", values_to = "cases")
```

pivot_wider()

```
table2 %>%
    pivot_wider(names_from = type, values_from = count)
```



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

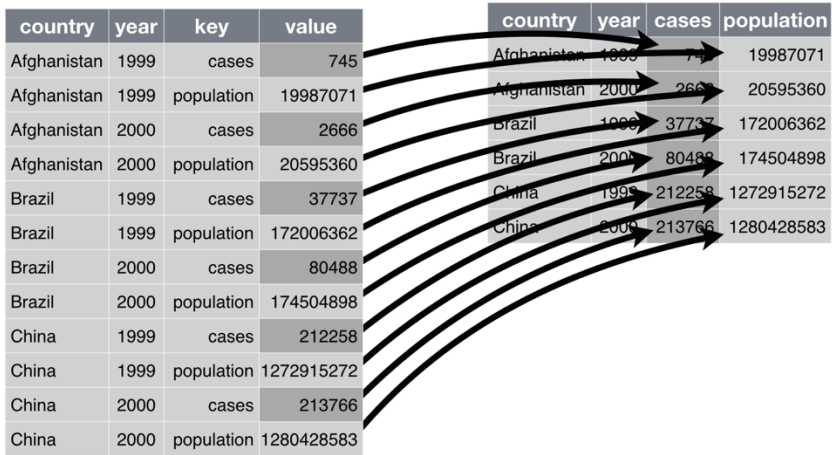
table2

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

separate()

```
table3 %>%
  separate(rate, into = c("cases", "population"))
```

By default, separate() will split values wherever it sees a non-alphanumeric character (i.e. a character that isn't a number or letter)



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Exercise – LOTR data

1. After tidying the data and completing your analysis, you may want to output a table that has each race in its own column. Let's use the pivot_wider() function to make such a table and save it as "lotr_wide"

2. OPTIONAL: Use the pivot_longer() function to transform you lotr_wide back to tidy format.

Exercise

- Make a barchart that shows how many words are spoken by males and females in each of the movies.
  - First with the tidy dataset
  - Then with the wider dataset you just created (with Male and Female word counts in different columns)

Exercise – coronavirus data

- Convert the coronavirus dataset to a wider format where the confirmed cases, deaths and recovered cases are shown in separate columns

- With this wide format data, make a bar chart of the total number of confirmed cases, deaths, and recoveries per day for the US