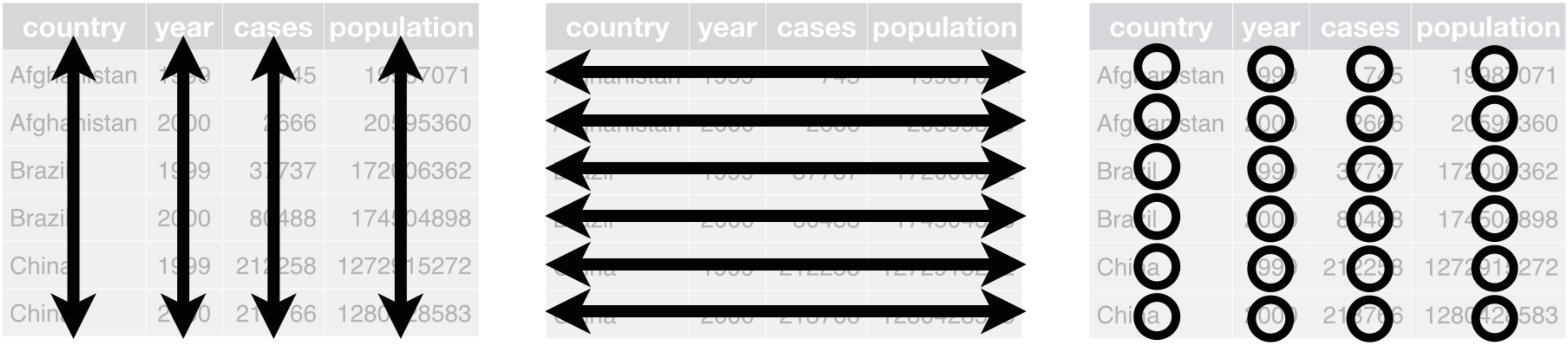
Tidy Data

From last time...

Does payroll differ between the American League and the National League?

- Load the `tidyverse`
- Load the `Lahman R` package
- Look at the `Salaries` and `Teams` data tables
- Devise a way to clearly compare the team payroll between the two leagues over the years

Definition: tidy data



variables observations values

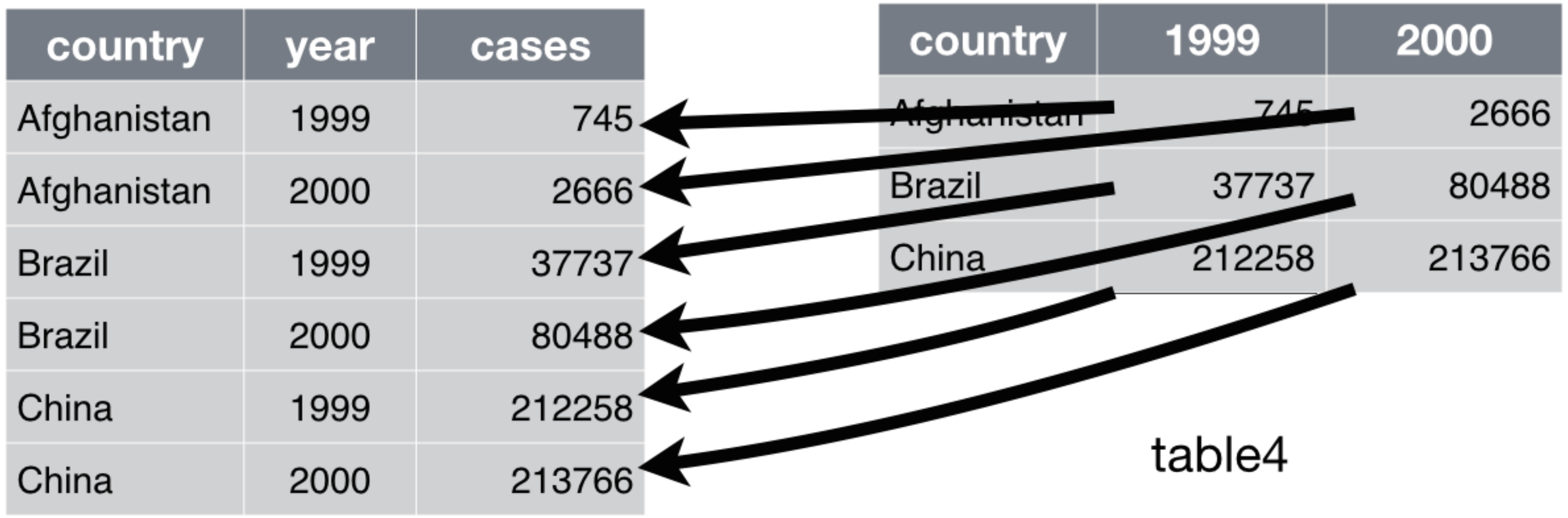
1. Each variable forms a column

2. Each observation (case) forms a row

3. Each type of observational unit forms a table

Gathering

```
table4a %>%
gather(key = "year", value = "cases", -country)
```



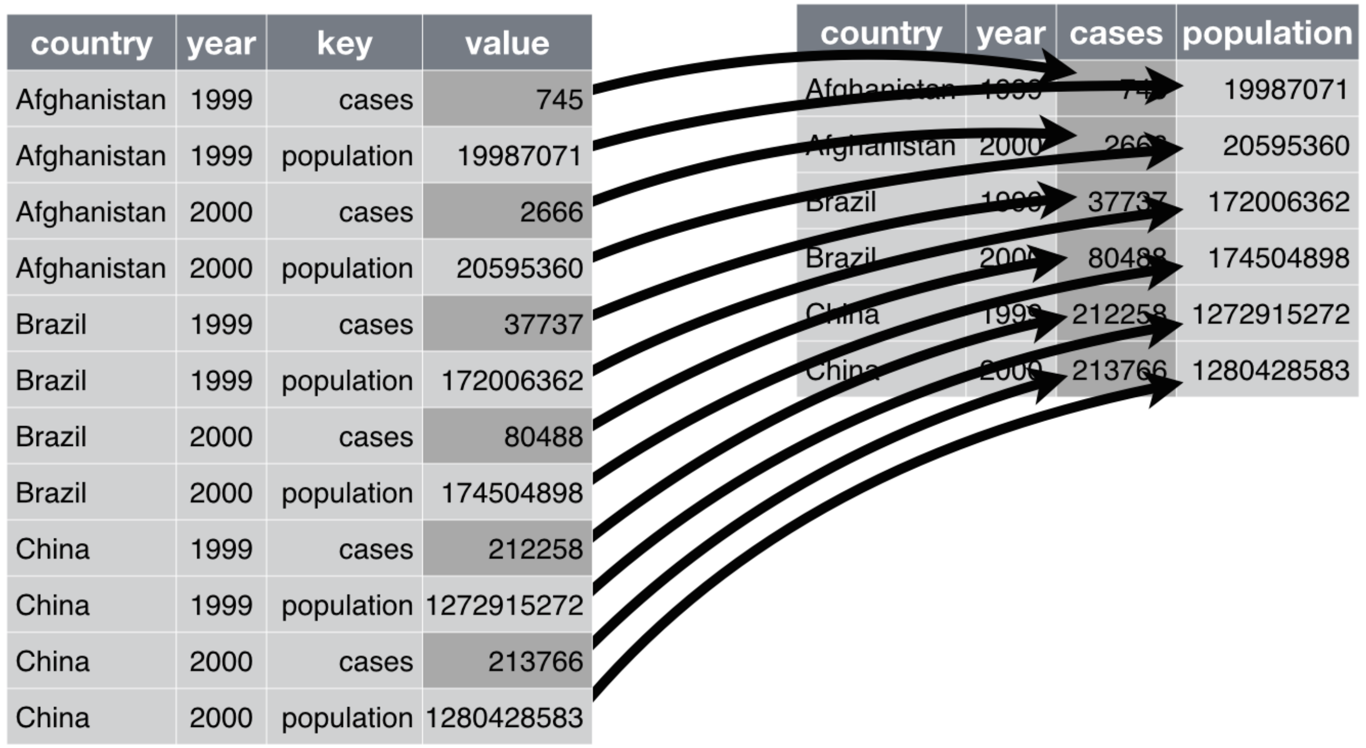
country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

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

table4

Spreading

```
table2 %>%
spread(key = type, value = 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

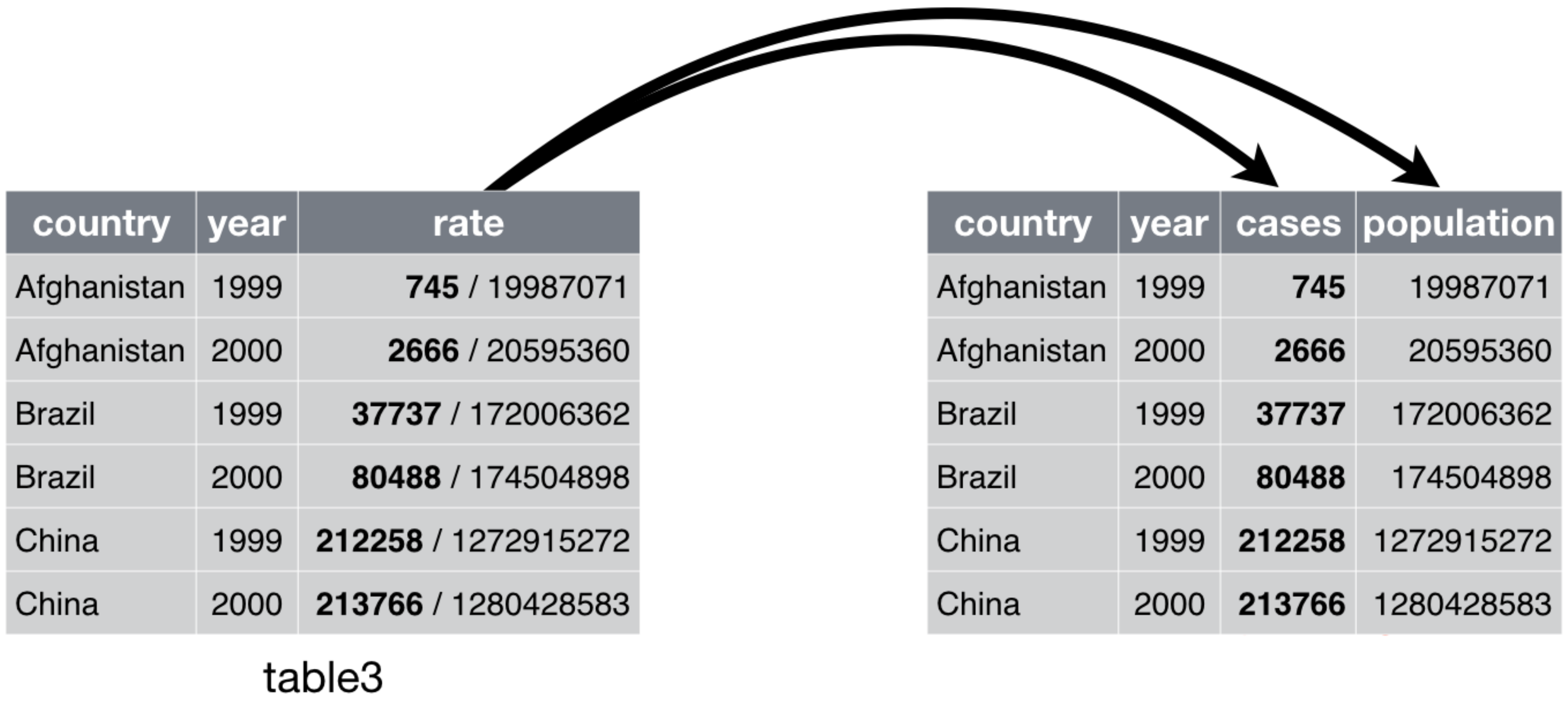
country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Your turn

- Data: `frenchfries.csv`
- 10 week sensory experiment, 12 individuals asked to assess taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do the fries taste?)
- French fries fried in 1 of 3 different oils, each week individuals had to assess 6 batches of french fries (all 3 oils, replicated 2x)
- Create boxplots of the numeric ratings by scale

Separate

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



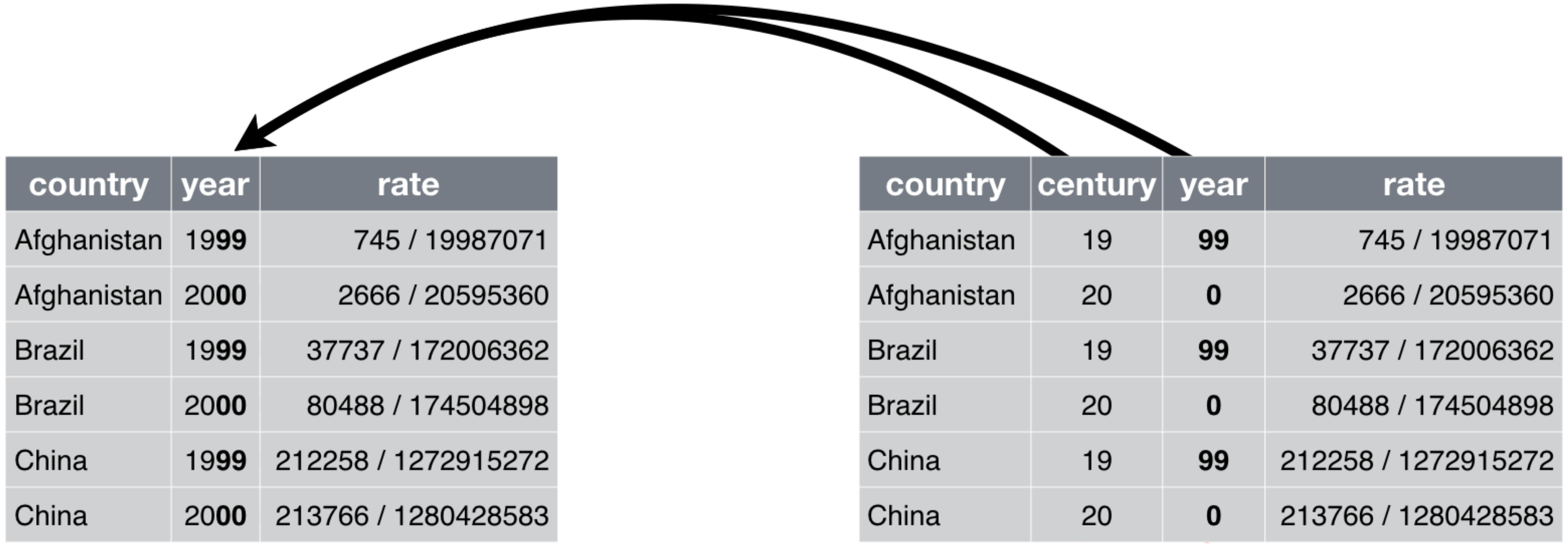
country	year	rate
Afghanistan	1999	**745** / 19987071
Afghanistan	2000	**2666** / 20595360
Brazil	1999	**37737** / 172006362
Brazil	2000	**80488** / 174504898
China	1999	**212258** / 1272915272
China	2000	**213766** / 1280428583

table3

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

Unite

```
table5 %>%
unite(year, century, year, sep = "")
```



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

country	century	year	rate
Afghanistan	19	**99**	745 / 19987071
Afghanistan	20	**0**	2666 / 20595360
Brazil	19	**99**	37737 / 172006362
Brazil	20	**0**	80488 / 174504898
China	19	**99**	212258 / 1272915272
China	20	**0**	213766 / 1280428583

table6

Why do we care?

- In the `tidyverse` input and outputs of all functions are encouraged to follow the tidy data format
- You might not be able to analyze the data in wide/long format depending on the type of analysis you want to run, or the plot you want to create

Your turn

`polls.csv` contains the results of various presidential polls conducted during July 2016, and was scraped from RealClear Politics.

1. Briefly describe why it is not considered to be tidy data and what changes need to be made to tidy it.
2. Use `separate` and `gather` to tidy the data set.

Your turn

under5mortality.csv contains the child mortality rate per 1,000 children born for each country from 1800 to 2015.

1. Briefly describe why it is not considered to be tidy data and what changes need to be made to tidy it.
2. Create a tidy data set with columns country, year and mortality.

Hint: Use `parse_numeric` to ensure that the year column is numeric (see `?parse_numeric` for help).

Your turn

`mlb2016.csv` contains the salary information presented by USA Today for all 862 players in Major League Baseball.

1. Briefly describe why it is not considered to be tidy data and what changes need to be made to tidy it.
2. Tidy this data set.