STAT 209
Reshaping and "Tidy" Data

July 8, 2021

Colin Reimer Dawson

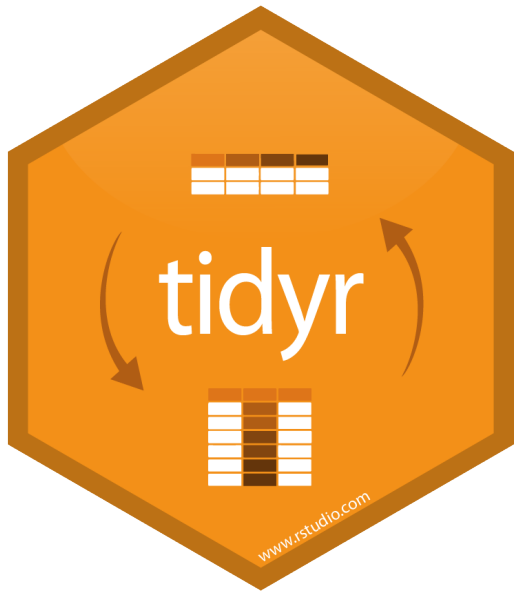
Outline

- Converting data between "wide" format and "long/narrow" (tidy) format
- New verbs:
  - `pivot_longer()`
  - `pivot_wider()`

Next Week...

- Some CS nuts and bolts (functions, loops) as applied to data wrangling and R
- Putting Together Data Wrangling
- Identify a Topic for Project 2

The tidyr package



Another piece of the "tidyverse". Here is a reference sheet that includes these verbs and others you may find useful.

Outline

Redundant Data

Tournament Winners

Tournament	Year	Winner	Winner Date of Birth
Indiana Invitational	1998	Al Fredrickson	21 July 1975
Cleveland Open	1999	Bob Albertson	28 September 1968
Des Moines Masters	1999	Al Fredrickson	21 July 1975
Indiana Invitational	1999	Chip Masterson	14 March 1977

- Cases here are **tournaments**, but `Winner Date of Birth` tells us about a **person** (the entry in `Winner`)
- It is *perfectly predictable* from the `Winner` column; i.e., it contains redundant information (if we know the `Winner` → `DOB` mapping)
- This is inefficient for storage and retrieval

Removing Redundancy

Tournament Winners

Tournament	Year	Winner
Indiana Invitational	1998	Al Fredrickson
Cleveland Open	1999	Bob Albertson
Des Moines Masters	1999	Al Fredrickson
Indiana Invitational	1999	Chip Masterson

Winner Dates of Birth

Winner	Date of Birth
Chip Masterson	14 March 1977
Al Fredrickson	21 July 1975
Bob Albertson	28 September 1968

- Here the data is separated into multiple tables
- Tables map case ID (`Tournament`, `Winner`) to pieces of information (variables) about that case
- Results in tables that are long (many rows) but narrow (few columns)
- More efficient database practice
- However, need to be able to `join` when needed

Generic Congressional Ballot

```
library(tidyverse)
url1 <- "https://raw.githubusercontent.com/TheUpshot"
url2 <- "leo-senate-model/master/fundamentals/data"
url3 <- "generic-approval/genericBallot.csv"
url <- paste(url1, url2, url3, sep = "/")
genericBallot <- read_csv(url)
```

Generic Congressional Ballot

```
genericBallot %>% slice_head(n = 3)
```

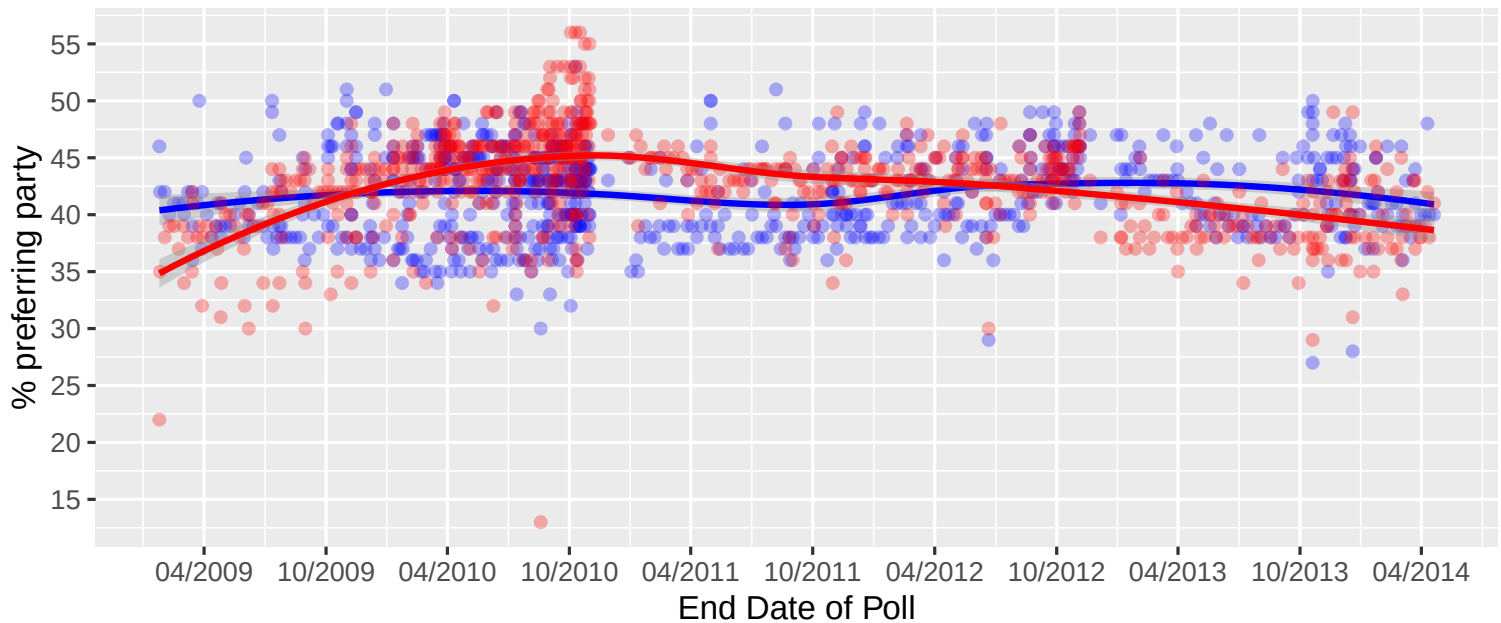
```
# A tibble: 3 x 7
  startdate  enddate    Democrats Republicans Sample  Poll       knownDate
  <date>     <date>         <dbl>       <dbl> <chr>   <chr>      <date>
1 2014-04-14 2014-04-20        40          41 3500 LV Rasmussen~ 2014-04-2
2 2014-04-07 2014-04-13        40          38 3500 LV Rasmussen~ 2014-04-1
3 2014-04-07 2014-04-10        48          42 1036 RV McClatchy~ 2014-04-1
```

```
genericBallot <- genericBallot %>%
    filter(enddate >= "2009-01-20", enddate < "2017-01-20") %>%
    rownames_to_column(var = "pollID")
genericBallot %>%
    slice_head(n = 3)
```

```
# A tibble: 3 x 8
  pollID startdate  enddate    Democrats Republicans Sample  Poll
  <chr>  <date>     <date>         <dbl>       <dbl> <chr>   <chr>
1 1      2014-04-14 2014-04-20        40          41 3500 LV Rasmussen Rep
2 2      2014-04-07 2014-04-13        40          38 3500 LV Rasmussen Rep
3 3      2014-04-07 2014-04-10        48          42 1036 RV McClatchy/Mar
# ... with 1 more variable: knownDate <date>
```

Plotting Party Preference Over Time

```
genericBallot %>%
    ggplot(aes(x = enddate)) +
    geom_point(aes(y = Democrats), color = "blue", alpha = 0.3) +
    geom_smooth(aes(y = Democrats), color = "blue") +
    geom_point(aes(y = Republicans), color = "red", alpha = 0.3) +
    geom_smooth(aes(y = Republicans), color = "red") +
    scale_x_date(
        name = "End Date of Poll", breaks = "6 months", date_labels = "%m/%Y") +
    scale_y_continuous(name = "% preferring party", breaks = seq(10,60,5))
```



Tidy Data?

```
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

- Dataset contains two variables recording the same quantity at different times.

Tidy Data?

```
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

- Dataset contains two variables recording the same quantity at different times.
- No actual "time" variable in the data

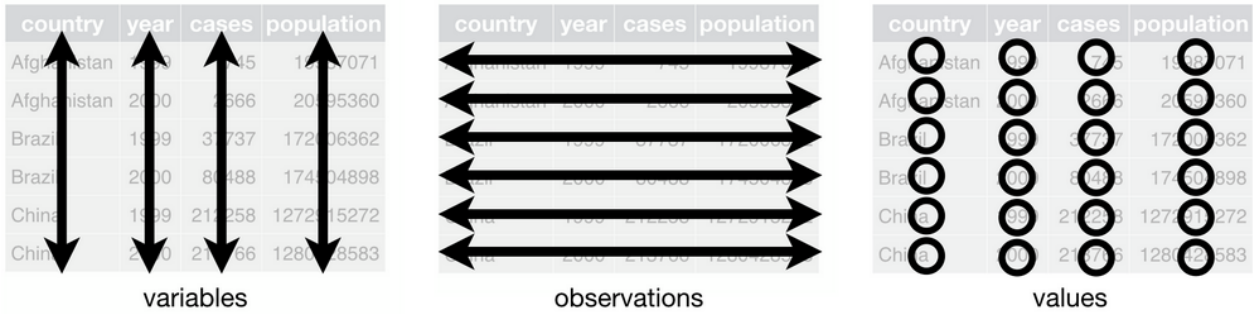
Tidy Data?

```
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

- Dataset contains two variables recording the same quantity at different times.
- No actual "time" variable in the data
- We can make this data "tidier"

Tidy Data



Data is **tidy** when each row represents a distinct **entity**, and each variable records a **property** of that entity.[1]

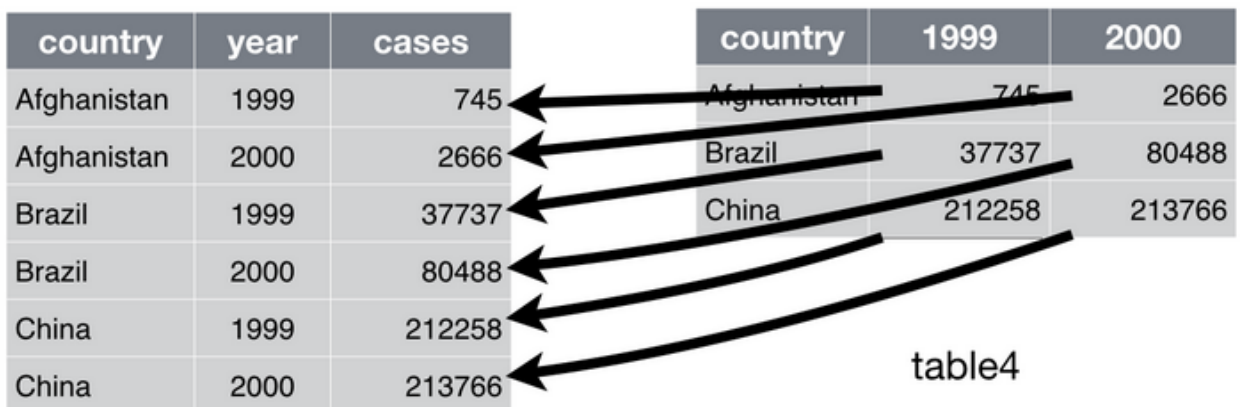
[1] Your textbook talks about data being "tidy" or "not tidy". While there are certainly formats that *cannot* be considered "tidy", there may not be one unique format which is "tidy": it depends what you consider an entity and a property.

Outline

pivot_longer(): Convert "wide" to "long"

Here we merge the 1999 and 2000 columns to a single column, creating a year column to distinguish the values.

```
table4a %>%
    pivot_longer(
        cols      = c("1999", "2000"), #which columns to 'stack'
        names_to  = "year",  # new variable to distinguish values
        values_to = "cases") # new name for the values variable
```



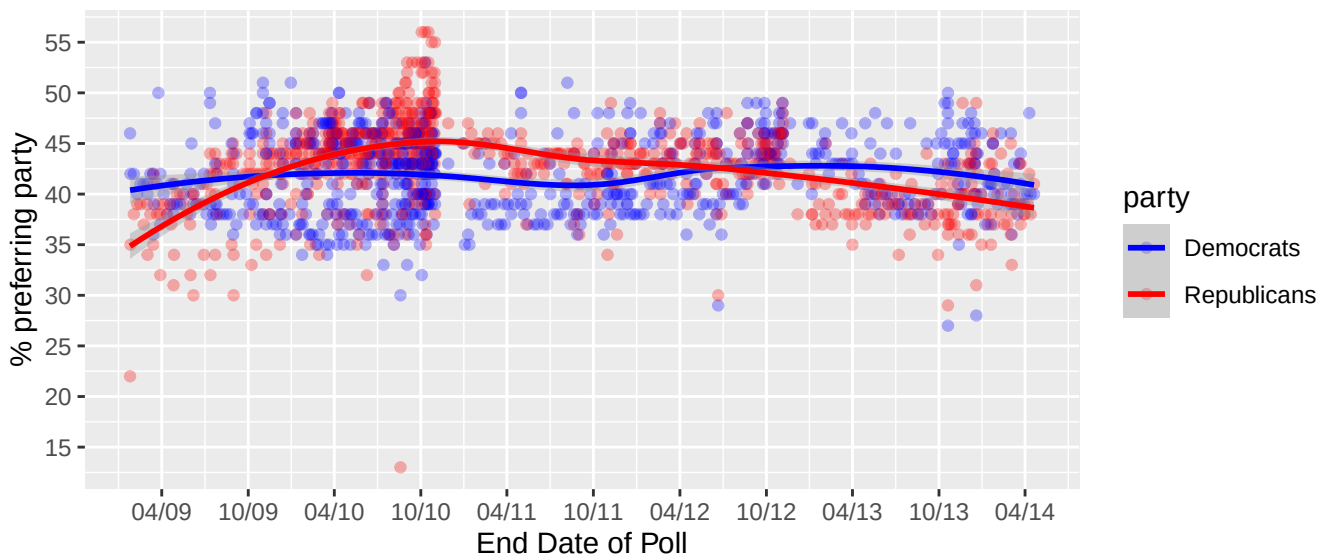
Reshaping the Generic Ballot Data

```
genericBallot_long <- genericBallot %>%
    pivot_longer(
        cols      = c("Democrats", "Republicans"),
        names_to  = "party",
        values_to = "share")
genericBallot_long %>%
    select(pollID, Poll, enddate, party, share) %>%
    slice_head(n=8)
```

```
# A tibble: 8 x 5
  pollID Poll              enddate    party       share
  <chr>  <chr>             <date>     <chr>       <dbl>
1 1      Rasmussen Reports 2014-04-20 Democrats      40
2 1      Rasmussen Reports 2014-04-20 Republicans    41
3 2      Rasmussen Reports 2014-04-13 Democrats      40
4 2      Rasmussen Reports 2014-04-13 Republicans    38
5 3      McClatchy/Marist  2014-04-10 Democrats      48
6 3      McClatchy/Marist  2014-04-10 Republicans    42
7 4      Rasmussen Reports 2014-04-06 Democrats      40
8 4      Rasmussen Reports 2014-04-06 Republicans    39
```

Cleaner Plotting

```
genericBallot_long %>%
    ggplot(aes(x = enddate, y = share, color = party)) +
    geom_point(alpha = 0.3) +
    geom_smooth() +
    scale_color_manual(values = c("Democrats" = "blue", "Republicans" = "red")) +
    scale_y_continuous(name = "% preferring party", breaks = seq(10,60,5)) +
    scale_x_date(
        name        = "End Date of Poll",
        date_breaks = "6 months",
        date_labels = "%m/%y")
```



Outline

The Opposite Problem

```
head(table2)
```

```
## # A tibble: 6 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
```

- What is a case here?

The Opposite Problem

```
head(table2)
```

```
## # A tibble: 6 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
```

- What is a case here?
- What does count represent?

The Opposite Problem

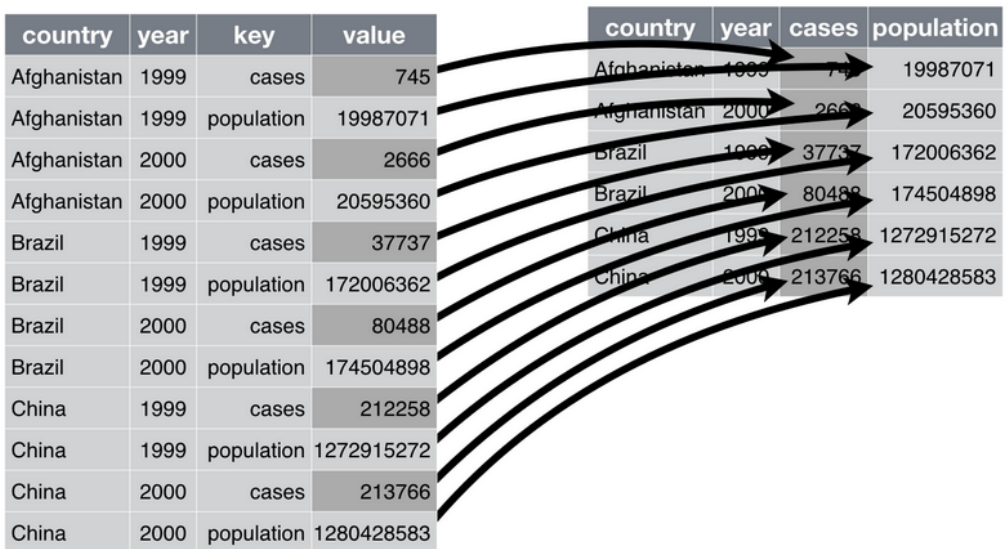
```
head(table2)
```

```
## # A tibble: 6 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
```

- What is a case here?
- What does count represent?
- Not tidy: doesn't map cases to properties

pivot_wider(): Convert narrow to wide

```
table2 %>%
    pivot_wider(
        id_cols    = c("country", "year"), # optionally give minimal columns
                                           # that uniquely specify a case
        names_from = "key",     # the key column tells us
                                # which variable 'value' is about
        values_from = "value")  # column of values to be 'unstacked'
```



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

spread() also reverses gather()

```
genericBallot_long %>%
    select(pollID, Poll, enddate, party, share) %>%
    slice_head(n = 4)


     # A tibble: 4 x 5
       pollID Poll                enddate    party       share
       <chr>  <chr>               <date>     <chr>       <dbl>
     1 1      Rasmussen Reports 2014-04-20 Democrats      40
     2 1      Rasmussen Reports 2014-04-20 Republicans    41
     3 2      Rasmussen Reports 2014-04-13 Democrats      40
     4 2      Rasmussen Reports 2014-04-13 Republicans    38

genericBallot_wide <- genericBallot_long %>%
    pivot_wider(
        names_from  = party,
        values_from = share)
```

spread() also reverses gather()

```
genericBallot_wide %>%
    select(pollID, Poll, enddate, Democrats, Republicans) %>%
    slice_head(n = 2)

    # A tibble: 2 x 5
      pollID Poll              enddate    Democrats Republicans
      <chr>  <chr>             <date>         <dbl>       <dbl>
    1 1      Rasmussen Reports 2014-04-20        40          41
    2 2      Rasmussen Reports 2014-04-13        40          38
```