



# readr read\_csv & read\_tsv



### Overview

- Before: utils package
- Specific R packages
  - readr
  - data.table



## readr

- Hadley Wickham
- Fast, easy to use, consistent
- utils: verbose, slower

- > install.packages("readr")
- > library(readr)



### CSV files

```
> read.csv("states.csv", stringsAsFactors = FALSE)
                   capital pop_mill area_sqm
         state
1 South Dakota
                   Pierre
                              0.853
                                        77116
      New York
                   Albany
                             19.746
                                        54555
                     Salem
        Oregon
                              3.970
                                        98381
       Vermont Montpelier
                              0.627
                                         9616
5
        Hawaii
                  Honolulu
                              1.420
                                        10931
  read_csv("states.csv")
# A tibble: 5 \times 4
                   capital pop_mill area_sqm
         state
                              <dbl>
                                        <int>
         <chr>
                     <chr>
 South Dakota
                              0.853
                    Pierre
                                        77116
                    Albany
      New York
                             19.746
                                        54555
                     Salem
                                        98381
                              3.970
        Oregon
       Vermont Montpelier
                              0.627
                                         9616
                 Honolulu
        Hawaii
5
                              1.420
                                        10931
```



#### states.csv

state,capital,pop\_mill,area\_sqm South Dakota, Pierre, 0.853, 77116 New York, Albany, 19.746, 54555 Oregon, Salem, 3.970, 98381 Vermont, Montpelier, 0.627, 9616 Hawaii, Honolulu, 1.420, 10931



### TSV files

```
> read.delim("states.txt", stringsAsFactors = FALSE)
                  capital pop_mill area_sqm
         state
                 Pierre
1 South Dakota
                             0.853
                                       77116
                   Albany
      New York
                            19.746
                                       54555
                   Salem
                             3.970
                                       98381
        Oregon
       Vermont Montpelier
                            0.627
                                        9616
5
        Hawaii
                 Honolulu
                              1.420
                                       10931
> read_tsv("states.txt")
# A tibble: 5 \times 4
         state
                  capital pop_mill area_sqm
                    <chr>
                              <dbl>
         <chr>
                                       <int>
1 South Dakota
                   Pierre
                             0.853
                                       77116
      New York
                   Albany
                            19.746
                                       54555
                    Salem
                                       98381
                              3.970
        Oregon
       Vermont Montpelier
                                        9616
                              0.627
        Hawaii
                 Honolulu
5
                              1.420
                                       10931
```

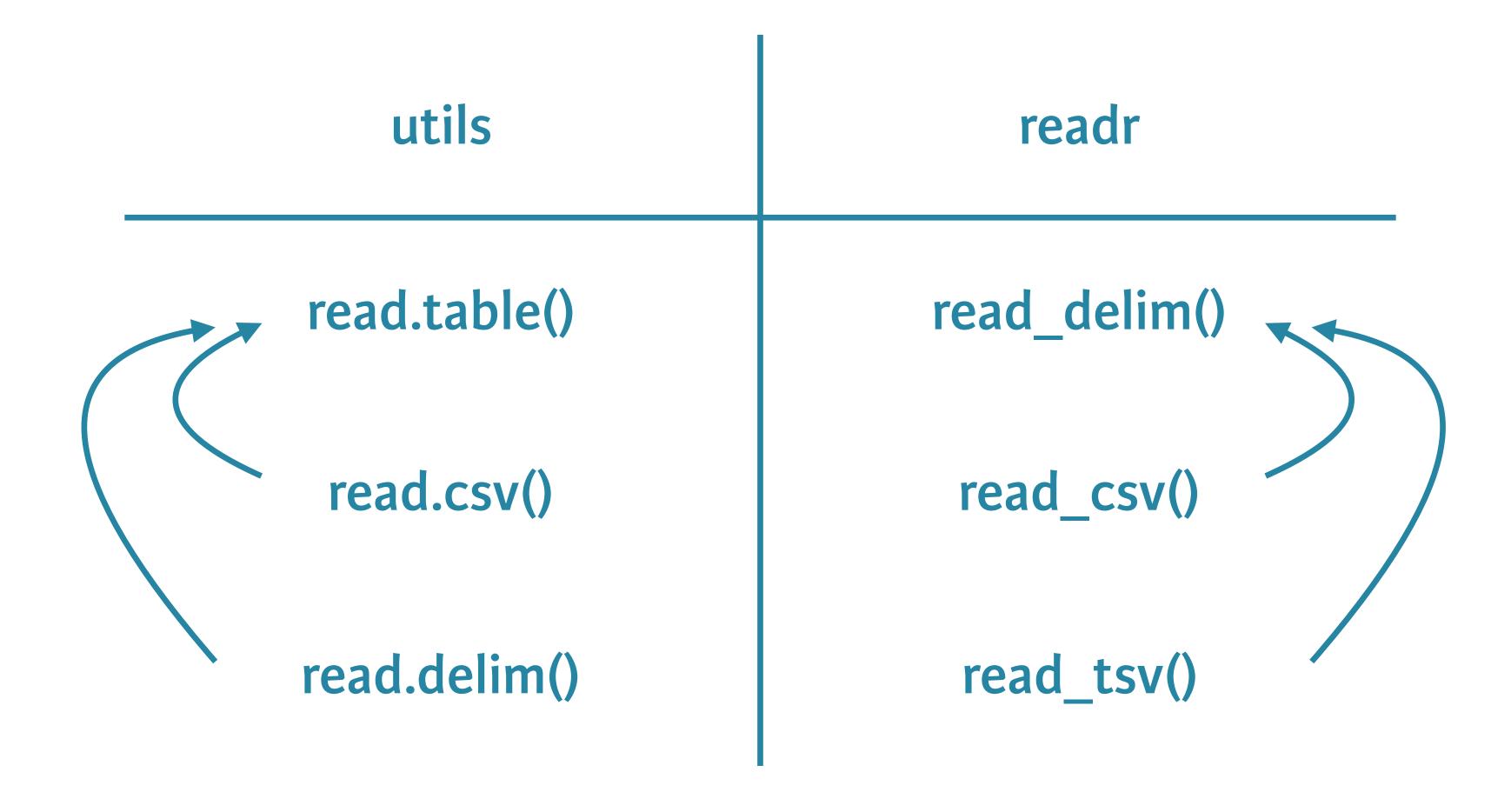


#### states.txt

state capital pop\_mill area\_sqm
South Dakota Pierre 0.853 77116
New York Albany 19.746 54555
Oregon Salem 3.970 98381
Vermont Montpelier 0.627 9616
Hawaii Honolulu 1.420 10931



# Wrapping in utils and readr







# Let's practice!





# readr read\_delim



### states2.txt

```
> read.table("states2.txt", header = TRUE, sep = "/",
             stringsAsFactors = FALSE)
                  capital pop_mill area_sqm
         state
 South Dakota
                   Pierre
                              0.853
                                        77116
      New York
                 Albany
                             19.746
                                       54555
                    Salem
                              3.970
                                       98381
        Oregon
       Vermont Montpelier
                              0.627
                                        9616
        Hawaii
5
                 Honolulu
                              1.420
                                        10931
> read_delim("states2.txt", delim = "/")
                                            col names
# A tibble: 5 x 4
                                             col_types
                  capital pop_mill area_sqm
         state
                     <chr>
                              <dbl>
         <chr>
                                        <int>
1 South Dakota
                   Pierre
                              0.853
                                        77116
      New York
                   Albany
                                        54555
                             19.746
                     Salem
        Oregon
                              3.970
                                        98381
       Vermont Montpelier
                              0.627
                                         9616
                 Honolulu
        Hawaii
5
                              1.420
                                        10931
```



#### states2.txt

state/capital/pop\_mill/area\_sqm South Dakota/Pierre/0.853/77116 New York/Albany/19.746/54555 Oregon/Salem/3.970/98381 Vermont/Montpelier/0.627/9616 Hawaii/Honolulu/1.420/10931



## col names

```
> read_delim("states3.txt", delim = "/", col_names = FALSE)
          X1
                    X2
                           X3
                                 X4
        <chr>
              <chr> <dbl> <int>
              Pierre 0.853 77116
1 South Dakota
     New York Albany 19.746 54555
       Oregon Salem 3.970 98381
      Vermont Montpelier 0.627
                               9616
5
       Hawaii
                Honolulu 1.420 10931
> read_delim("states3.txt", delim = "/",
            col_names = c("state", "city", "pop", "area"))
                   city
       state
                          pop area
                <chr> <dbl> <int>
        <chr>
1 South Dakota
              Pierre 0.853 77116
     New York
                  Albany 19.746 54555
                   Salem 3.970 98381
       Oregon
      Vermont Montpelier 0.627
       Hawaii
                Honolulu
                        1.420 10931
```



states3.txt

South Dakota/Pierre/0.853/77116

New York/Albany/19.746/54555 Oregon/Salem/3.970/98381

Vermont/Montpelier/0.627/9616

Hawaii/Honolulu/1.420/10931





## col\_types

```
> read_delim("states2.txt", delim = "/")
                   capital pop_mill area_sqm
         state
                               <dbl>
                                         <int>
         <chr>
                     <chr>
1 South Dakota
                    Pierre
                               0.853
                                         77116
      New York
                   Albany
                              19.746
                                         54555
                     Salem
                               3.970
                                         98381
        Oregon
                               0.627
       Vermont Montpelier
                                          9616
        Hawaii
5
                  Honolulu
                               1.420
                                         10931
> read_delim("states2.txt", delim = "/", col_types = "ccdd")
                   capital pop_mill area_sqm
         state
                               <dbl>
                                         <dbl>
         <chr>
                     <chr>
                                                    c = character
1 South Dakota
                                         77116
                    Pierre
                               0.853
                                                    d = double
      New York
                    Albany
                              19.746
                                         54555
                                                    i = integer
                     Salem
                                         98381
                               3.970
        Oregon
                                                    I = logical
       Vermont Montpelier
                               0.627
                                          9616
                                                     _{=} skip
        Hawaii
                  Honolulu
5
                                         10931
                               1.420
```

### states2.txt

South Dakota/Pierre/0.853/77116
New York/Albany/19.746/54555
Oregon/Salem/3.970/98381
Vermont/Montpelier/0.627/9616
Hawaii/Honolulu/1.420/10931



## skip and n\_max

```
> read_delim("states2.txt", delim = "/",
            skip = 2, n_max = 3)
# A tibble: 3 x 4
 New York Albany 19.746 54555
    <chr> <chr> <dbl> <int>
   Oregon Salem 3.970 98381
  Vermont Montpelier 0.627 9616
   Hawaii
           Honolulu 1.420 10931
3
> read_delim("states2.txt", delim = "/",
            col_names = c("state", "city", "pop", "area"),
            skip = 2, n_max = 3)
# A tibble: 3 x 4
    state
             city
                      pop area
           <chr> <dbl> <int>
    <chr>
1 New York
             Albany 19.746 54555
           Salem 3.970 98381
   Oregon
  Vermont Montpelier 0.627 9616
```



#### states.csv

state, capital, pop\_mill, area\_sqm South Dakota, Pierre, 0.853, 77116 New York, Albany, 19.746, 54555 Oregon, Salem, 3.970, 98381 Vermont, Montpelier, 0.627, 9616 Hawaii, Honolulu, 1.420, 10931





# Let's practice!





### data.table: fread



### data.table

- Matt Dowle & Arun Srinivasan
- Key metric: speed
- Data manipulation in R
- Function to import data: fread()
- > install.packages("data.table")
- > library(data.table)
- Similar to read.table()



## fread()



state, capital, pop\_mill, area\_sqm

South Dakota, Pierre, 0.853, 77116 New York, Albany, 19.746, 54555

Oregon, Salem, 3.970, 98381

Vermont, Montpelier, 0.627, 9616

Hawaii, Honolulu, 1.420, 10931

#### > fread("states.csv")

	•		<b>,</b>		
		state	capital	pop_mill	area_sqm
1:	South [	)akota	Pierre	0.853	77116
2:	Nev	v York	Albany	19.746	54555
3:	(	regon	Salem	3.970	98381
4:	Ve	ermont	Montpelier	0.627	9616
5:	H	Hawaii	Honolulu	1.420	10931



states2.csv

South Dakota, Pierre, 0.853, 77116 New York, Albany, 19.746, 54555 Oregon, Salem, 3.970, 98381 Vermont, Montpelier, 0.627, 9616 Hawaii, Honolulu, 1.420, 10931

#### > fread("states? csv")

/ Tread( Statesz.Csv )								
	V1	V2	V3	V4				
1:	South Dakota	Pierre	0.853	77116				
2:	New York	Albany	19.746	54555				
3:	Oregon	Salem	3.970	98381				
4:	Vermont	Montpelier	0.627	9616				
5:	Hawaii	Honolulu	1.420	10931				

## fread()

- Infer column types and separators
- It simply works
- Extremely fast
- Possible to specify numerous parameters
- Improved read.table()
- Fast, convenient, customizable





# Let's practice!