Once again, I’m dipping outside of the tidyverse, but this package and its functions have been really useful in getting data quickly in (and out) of R.

For work, I have to pull in data from a few different sources, and manipulate and work with them to give me the final dataset that I use for much of my analysis. So that I don’t have to go through all of that joining, recoding, and calculating each time, I created a final merged dataset as a CSV file that I can load when I need to continue my analysis. The problem is that the most recent version of that file, which contains 13 million+ records, was so large, writing it (and subsequently reading it in later) took forever and sometimes timed out.

That’s when I discovered the data.table library, and its fread and fwrite functions. Tidyverse is great for working with CSV files, but a lot of the memory and loading time is used for formatting. fread and fwrite are leaner and get the job done a bit faster. For regular-sized CSV files (like my reads2019 set), the time difference is pretty minimal. But for a 5GB datafile, it makes a huge difference.

```
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------- tidyverse
1.3.0 --
```

```
##  ggplot2 3.2.1      purrr   0.3.3
##  tibble  2.1.3      dplyr   0.8.3
##  tidyr   1.0.0      stringr 1.4.0
##  readr   1.3.1      forcats 0.4.0
```

```
## -- Conflicts ----------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
system.time(reads2019 <- read_csv("~/Downloads/Blogging A to
Z/SaraReads2019_allchanges.csv",
                      col_names = TRUE))
```

```
## Parsed with column specification:
## cols(
##   Title = col_character(),
##   Pages = col_double(),
##   date_started = col_character(),
##   date_read = col_character(),
##   Book.ID = col_double(),
##   Author = col_character(),
##   AdditionalAuthors = col_character(),
##   AverageRating = col_double(),
##   OriginalPublicationYear = col_double(),
##   read_time = col_double(),
##   MyRating = col_double(),
##   Gender = col_double(),
##   Fiction = col_double(),
##   Childrens = col_double(),
##   Fantasy = col_double(),
##   SciFi = col_double(),
##   Mystery = col_double(),
##   SelfHelp = col_double()
## )
```

```
##    user  system elapsed
##    0.00    0.10    0.14
```

```
rm(reads2019)
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
system.time(reads2019 <- fread("~/Downloads/Blogging A to
Z/SaraReads2019_allchanges.csv"))
```

```
##    user  system elapsed
##       0       0       0
```

But let’s show how long it took to read my work datafile. Here’s the elapsed time from the system.time output.

read_csv:
user system elapsed
61.14 11.72 90.56

fread:
user system elapsed
57.97 16.40 57.19

But the real win is in how quickly this package writes CSV data. Using a package called wakefield, I’ll randomly generate 10,000,000 records of survey data, then see how it takes to write the data to file using both write_csv and fwrite.

```
library(wakefield)
```

```
## Warning: package 'wakefield' was built under R version 3.6.3
```

```
##
## Attaching package: 'wakefield'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, minute, month, second, year
```

```
## The following object is masked from 'package:dplyr':
##
##     id
```

```
set.seed(42)
```

```
reallybigshew <- r_data_frame(n = 10000000,
                              id,
                              race,
                              age,
                              smokes,
                              marital,
                              Start = hour,
                              End = hour,
                              iq,
                              height,
                              died)
```

```
system.time(write_csv(reallybigshew, "~/Downloads/Blogging A to
Z/bigdata1.csv"))
```

```
##    user  system elapsed
##  134.22    2.52  137.80
```

```
system.time(fwrite(reallybigshew, "~/Downloads/Blogging A to Z/bigdata2.csv"))
```

```
##    user  system elapsed
##    8.65    0.32    2.77
```