
VROOM!
Life's Too Short To Drive Slow
Read
Jim Hester
RStudio
@jimhester
@jimhester_
SPEAKERDECK.COM/JIMHESTER/VROOM
Photo by Joe Neric on Unsplash

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Joshua Reddekopp on Unsplash

.1 second

1 second

10 seconds

SPEAKERDECK.COM/JIMHESTER/VROOM

source: https://www.nngroup.co/articles/response-times-3-important-limit

Amjad Masad 🎱
@amasad

Interactive programming environments should give you feedback faster than your impulse to open Twitter or HN.

3:34 PM · Jun 18, 2019 · Twitter Web Client

35 Retweets **183** Likes

SPEAKERDECK.COM/JIMHESTER/VROOM

source: https://twitter.com/amasad/status/1141066550180552704

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

dttm [1]: pickup_datetime

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Andre Benz on Unsplash

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

dttm [1]: pickup_datetime

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Andre Benz on Unsplash

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

dttm [1]: pickup_datetime

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Andre Benz on Unsplash

SPEAKERDECK.COM/JIMHESTER/VROOM

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

dttm [1]: pickup_datetime

Photo by Andre Benz on Unsplash

SPEAKERDECK.COM/JIMHESTER/VROOM

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

dttm [1]: pickup_datetime

Photo by Andre Benz on Unsplash

Taxi Trip Fare

Observations: 14,776,615

Variables: 11

File Size: 1.55G

chr [4]: medallion, hack_license, v

dbl [6]: fare_amount, surcharge,

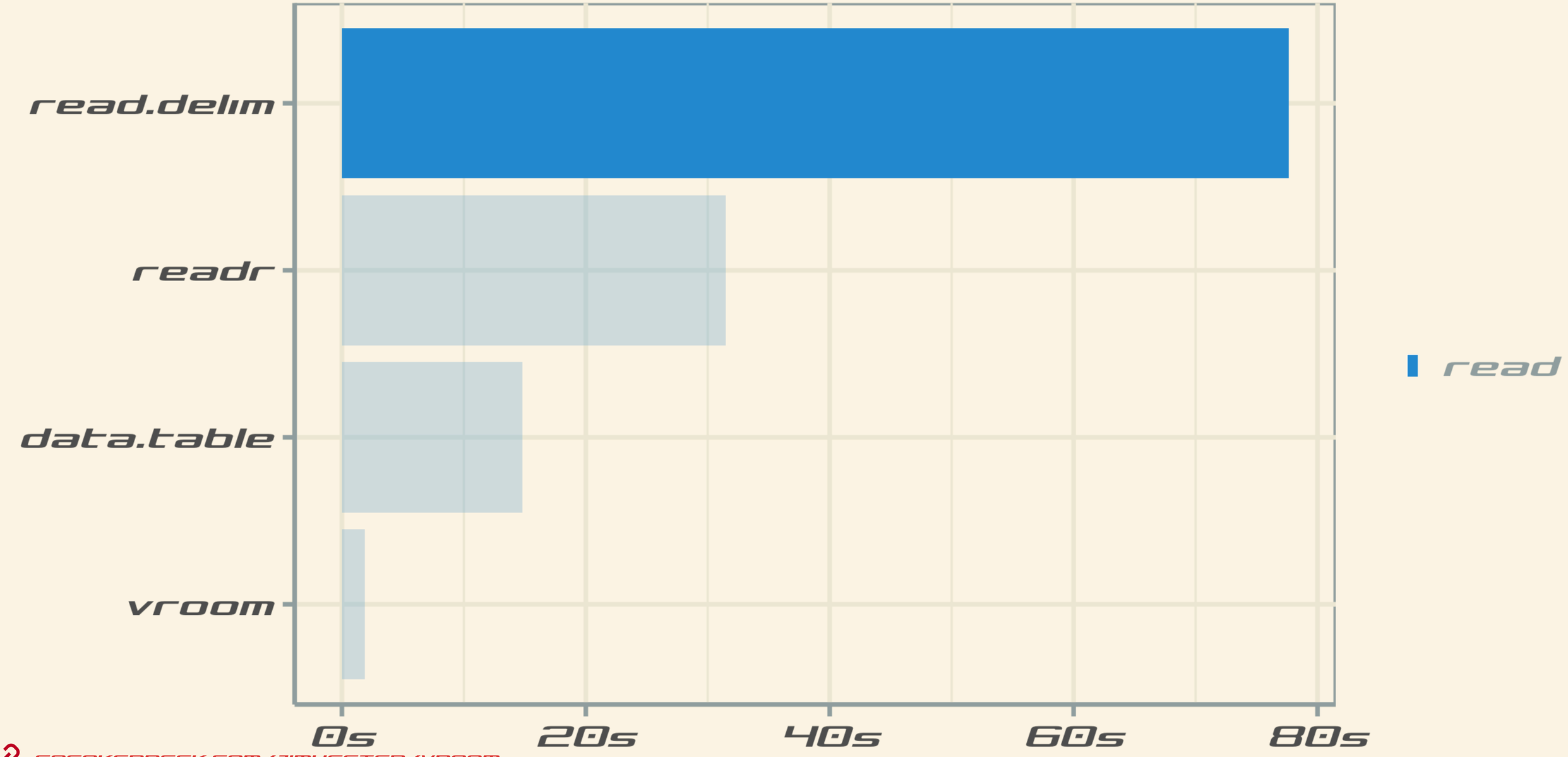
dttm [1]: pickup_datetime

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Andre Benz on Unsplash

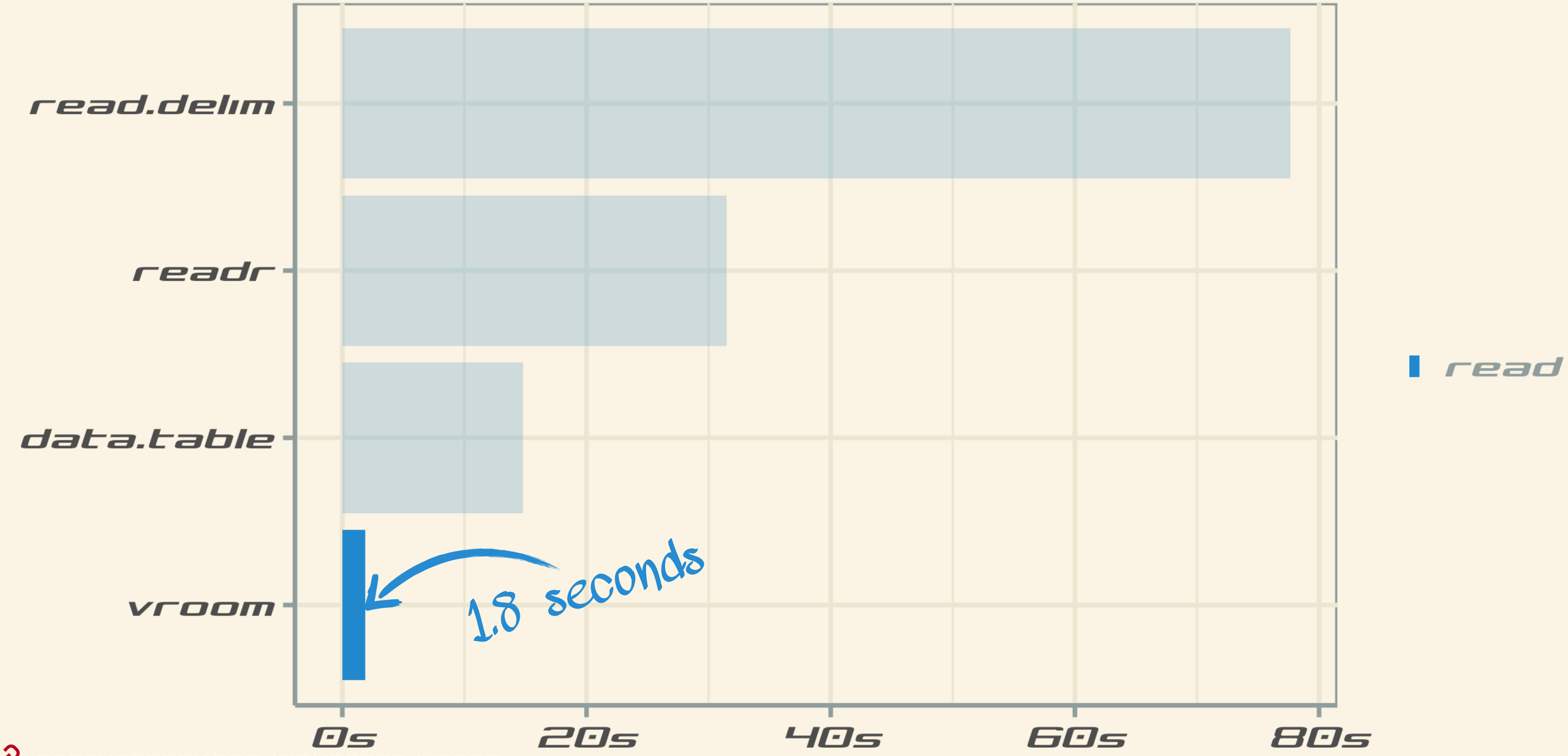
Taxi trip fare

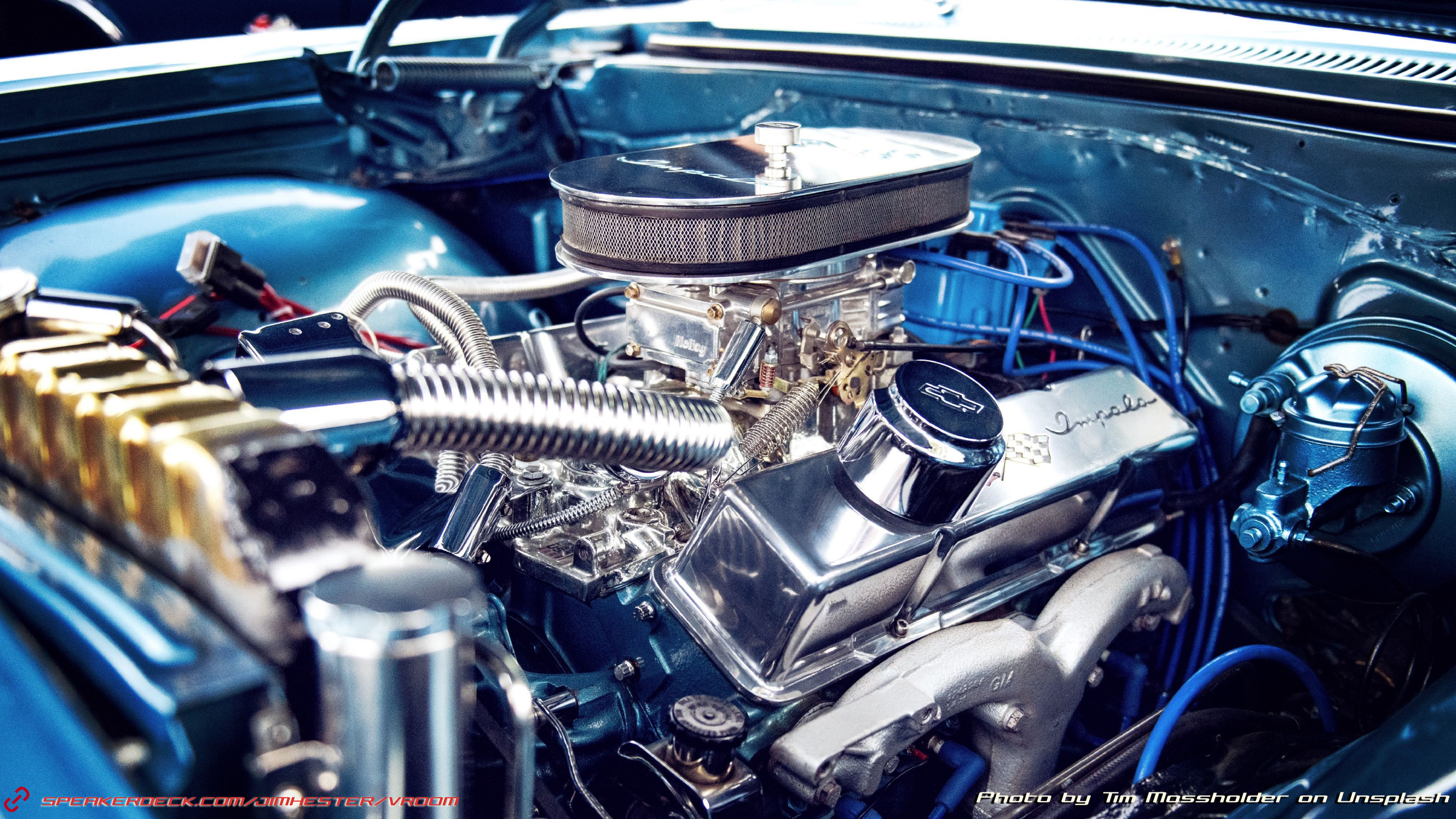
14,776,615 x 11 – 1.55GB



Taxi trip fare

14,776,615 x 11 – 1.55GB

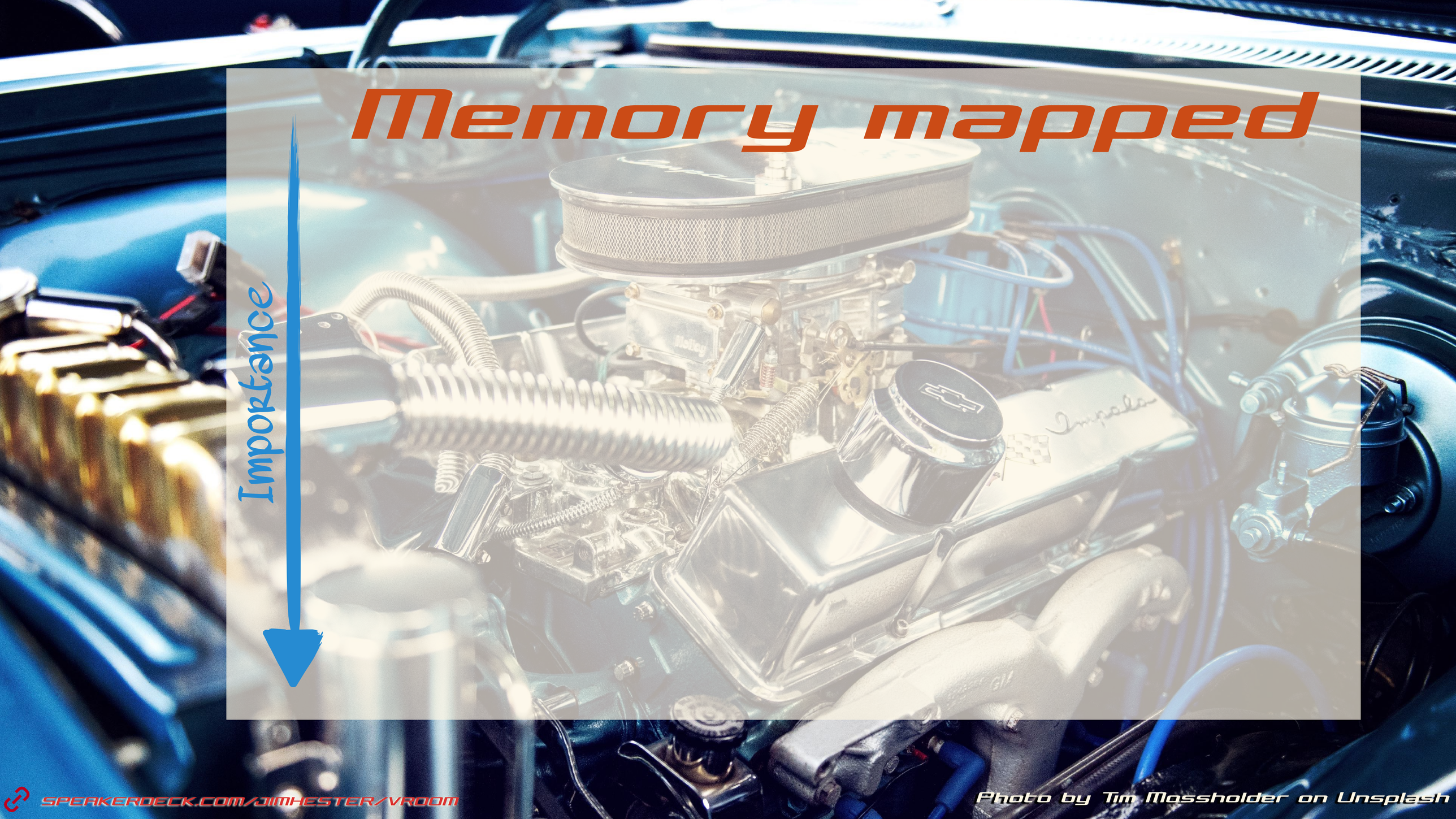




SPEAKERDECK.COM/JIMHESTER/VROOM

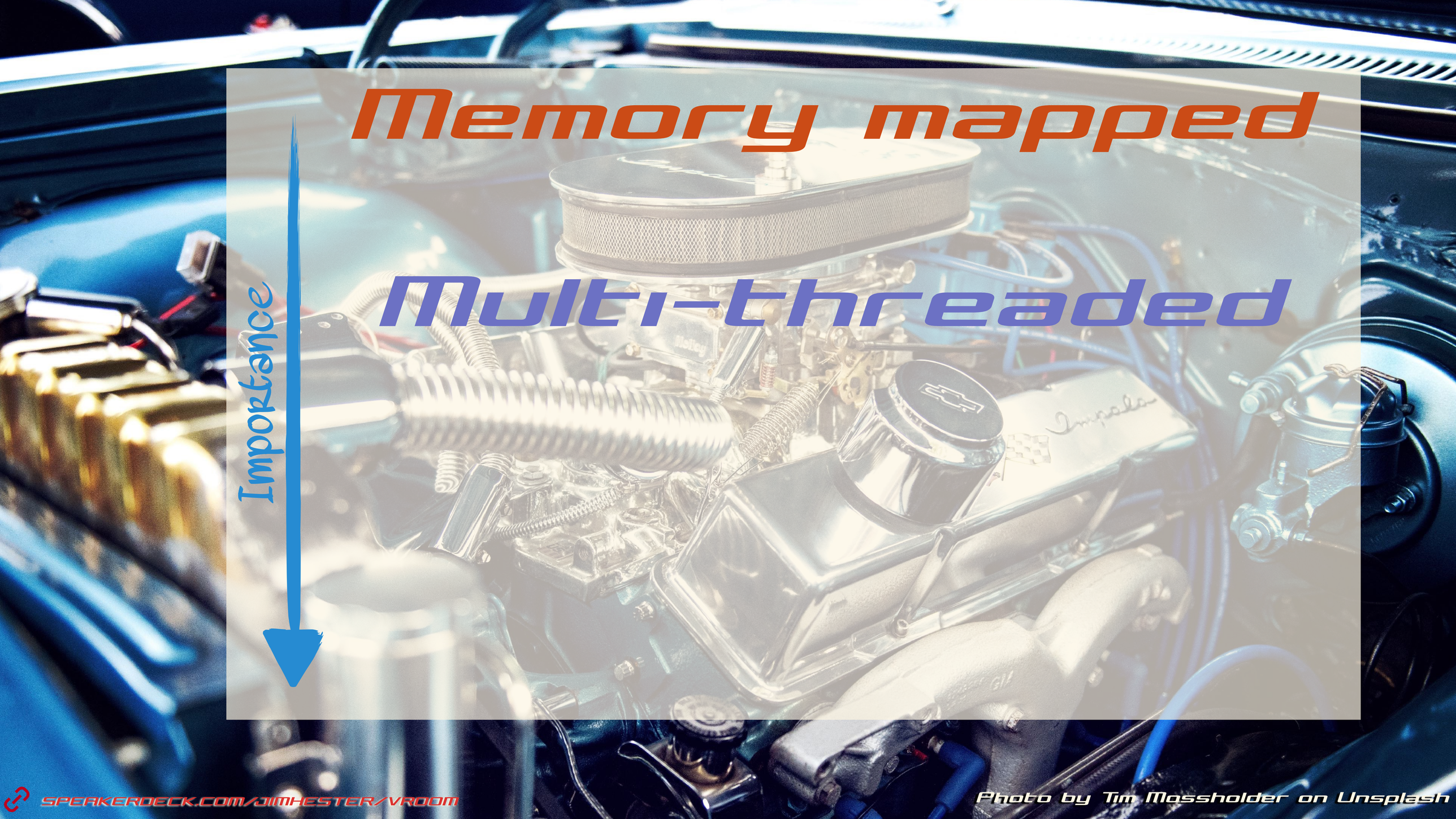
Photo by Tim Mossholder on Unsplash

Memory mapped



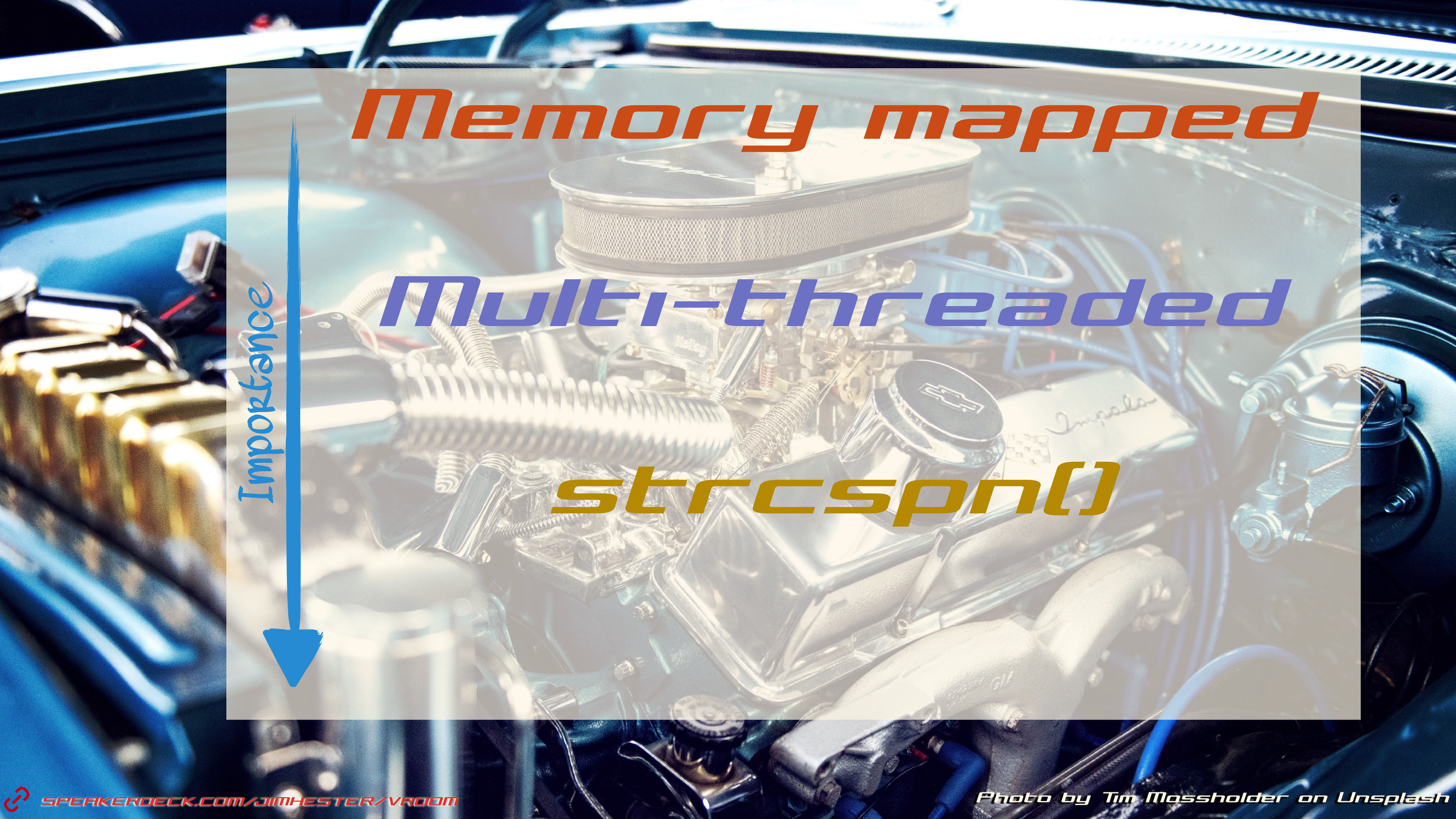
SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Tim Mossholder on Unsplash



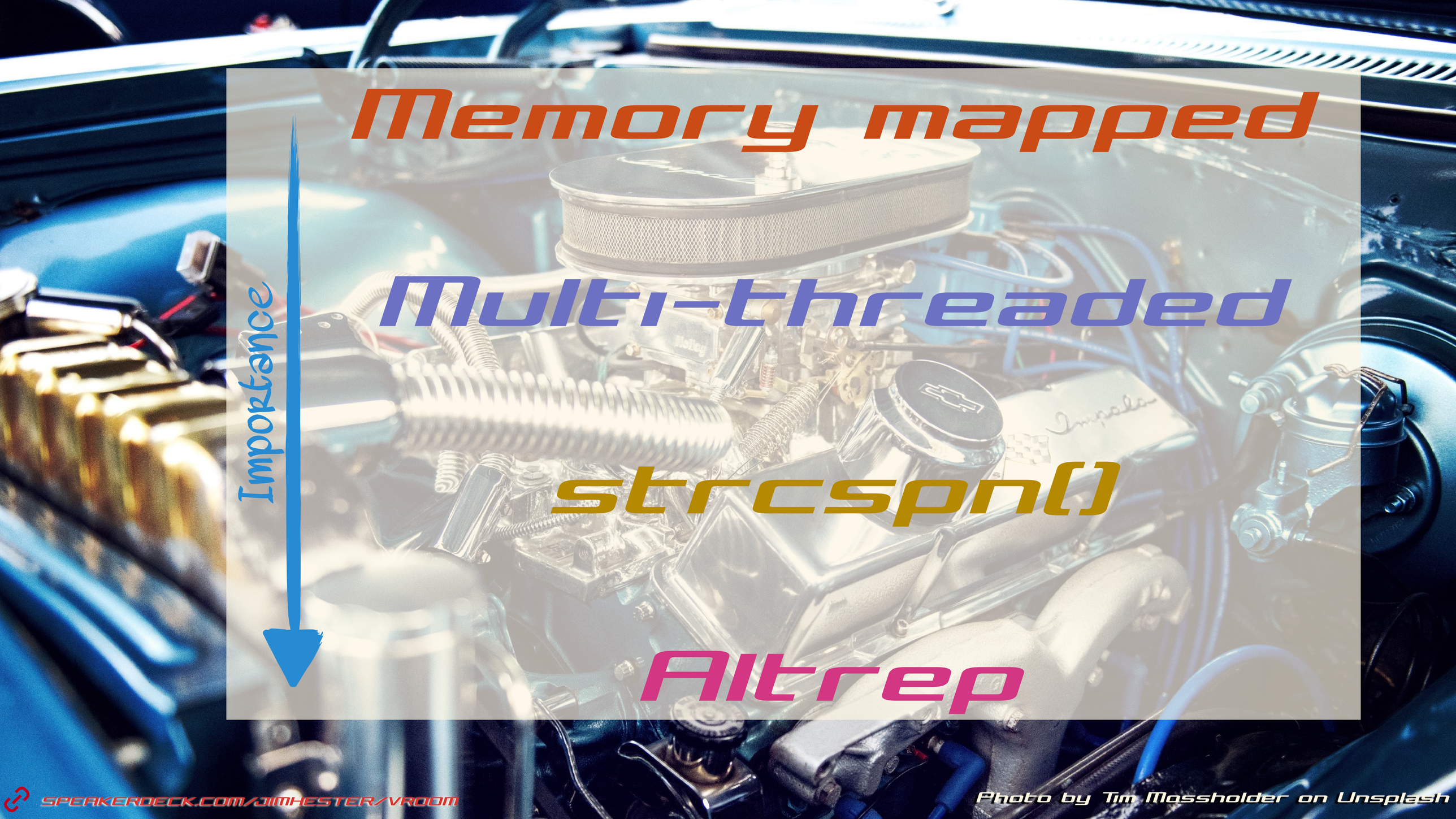
SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Tim Mossholder on Unsplash



SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Tim Mossholder on Unsplash



SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Tim Mossholder on Unsplash

ALTREP

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

ALTREP

Alternative representation

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

ALTREP

Alternative representation

R 3.5+

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

ALTREP

Alternative representation

R 3.5+

Custom memory storage

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

ALTREP

Alternative representation

R 3.5+

Custom memory storage

Transparent to R & C/C++

SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

ALTREP

Alternative representation

R 3.5+

Custom memory storage

Transparent to R & C/C++

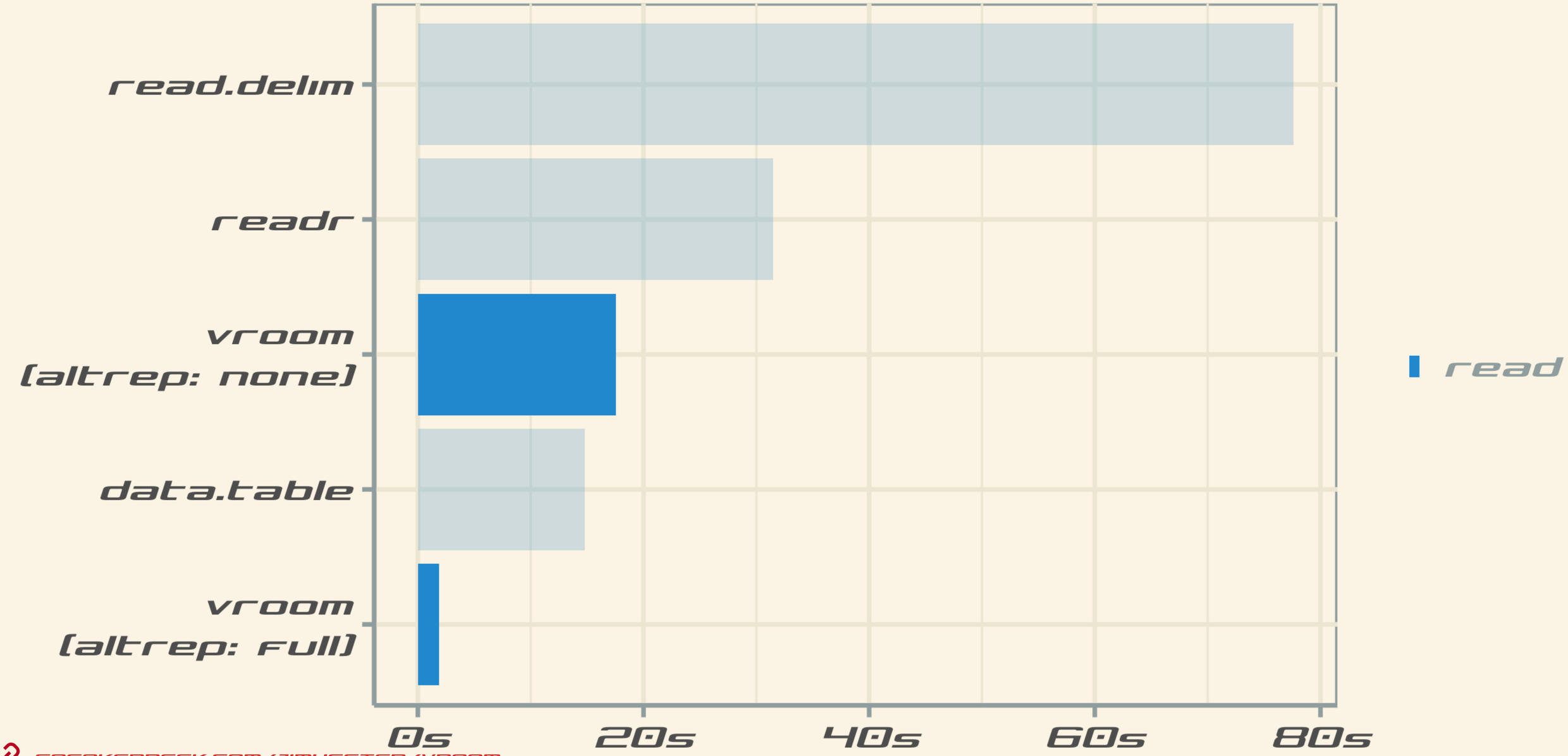
On-demand parsing

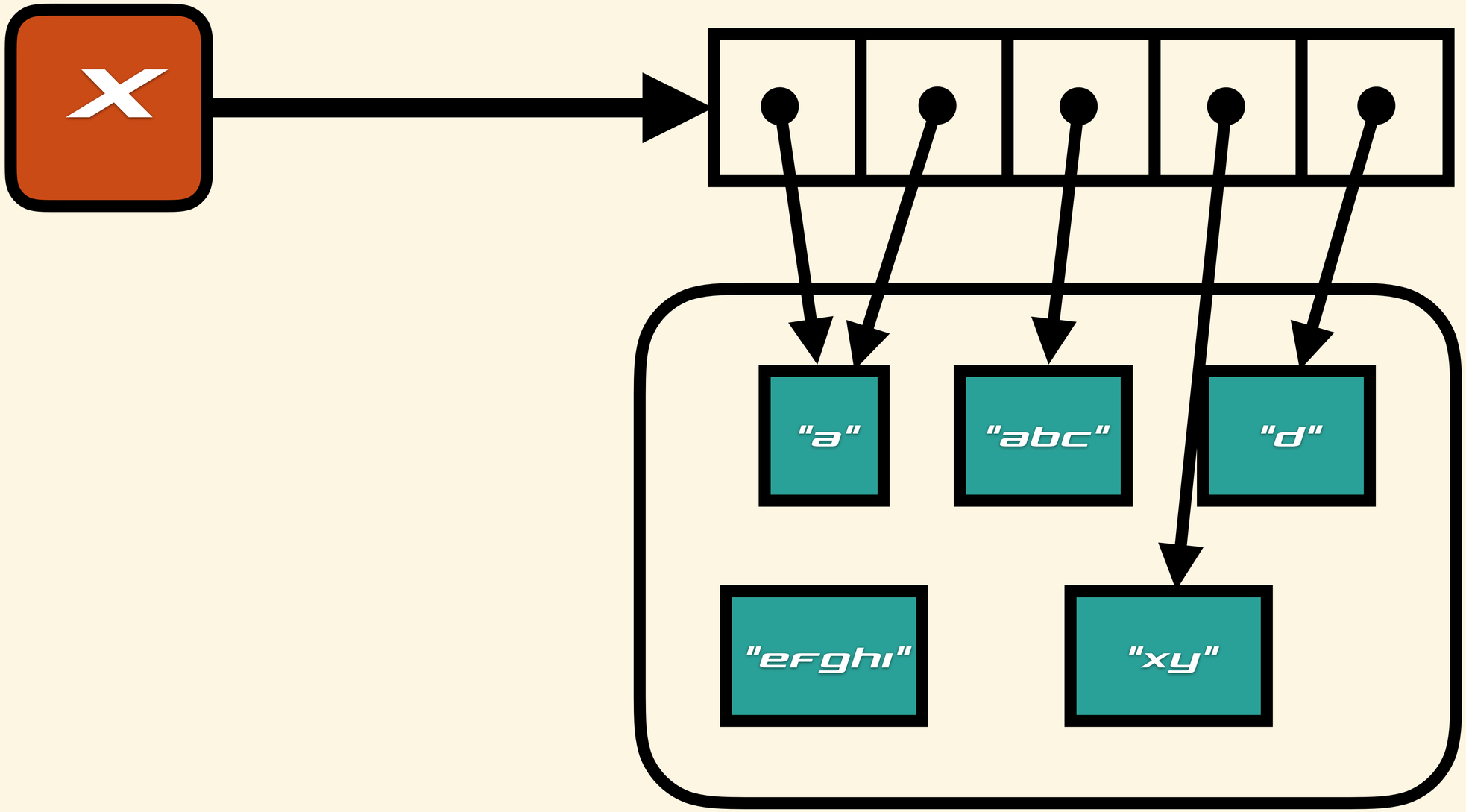
SPEAKERDECK.COM/JIMHESTER/VROOM

Photo by Callum Shaw on Unsplash

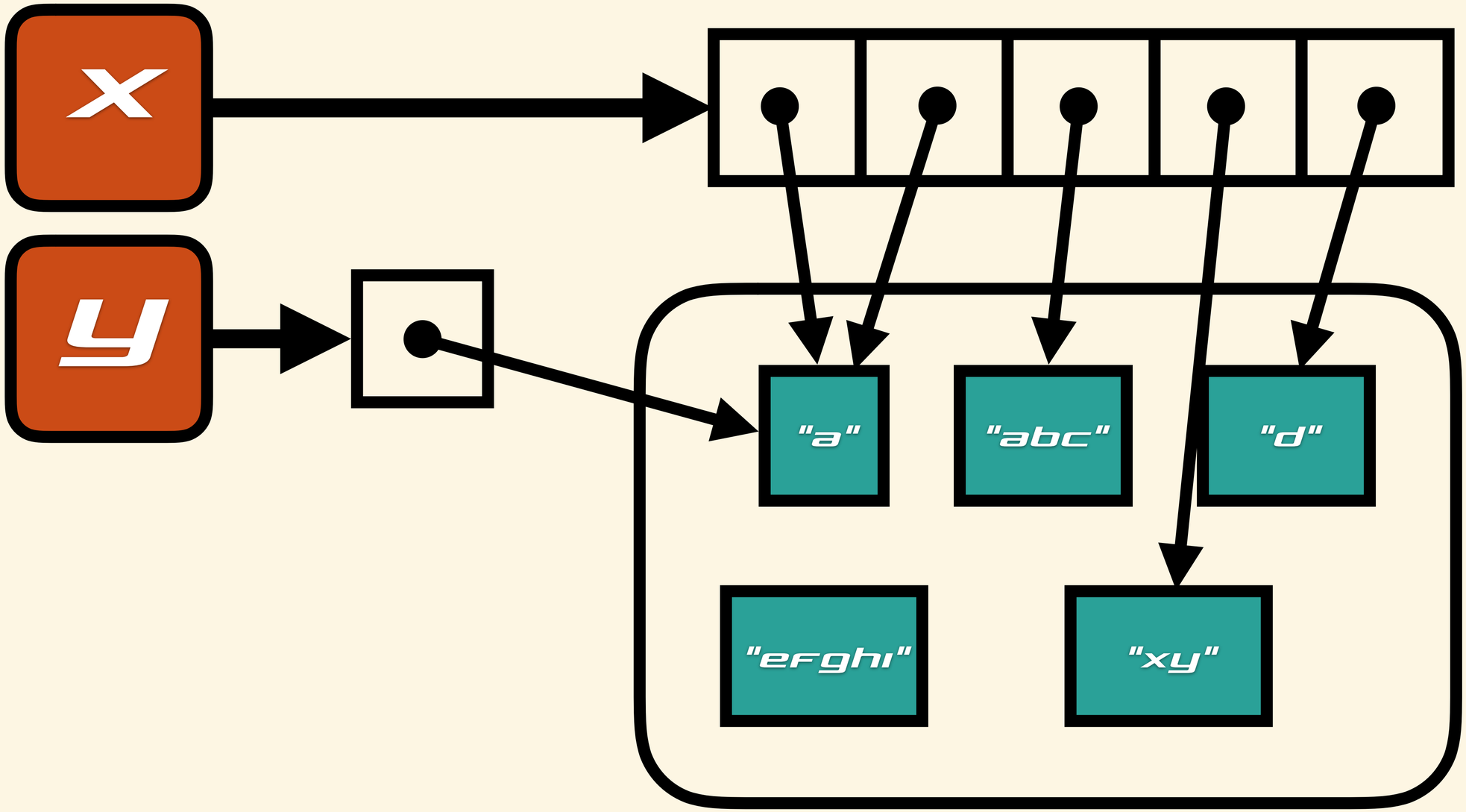
Taxi trip fare

14,776,615 x 11 – 1.55GB

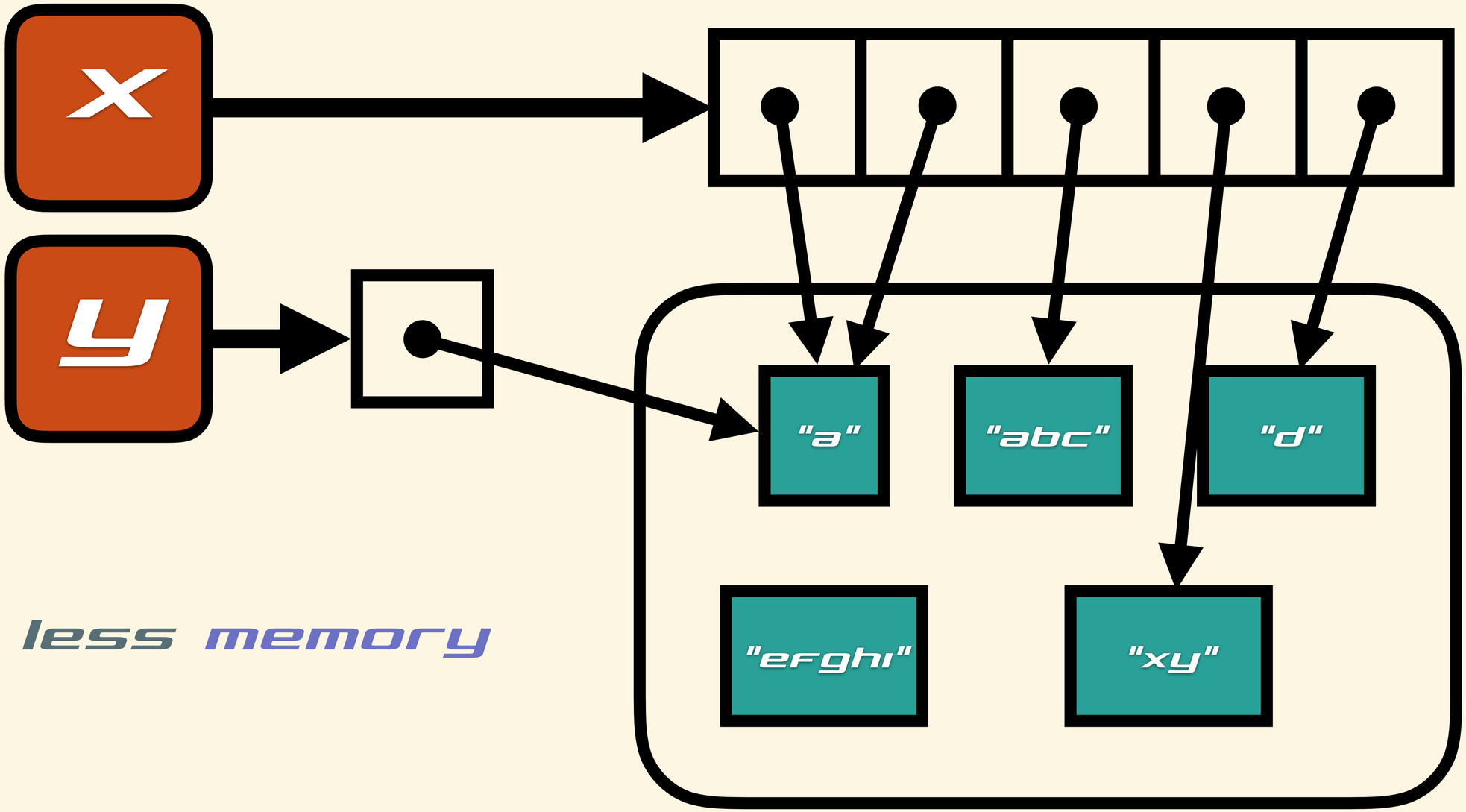




Global string pool

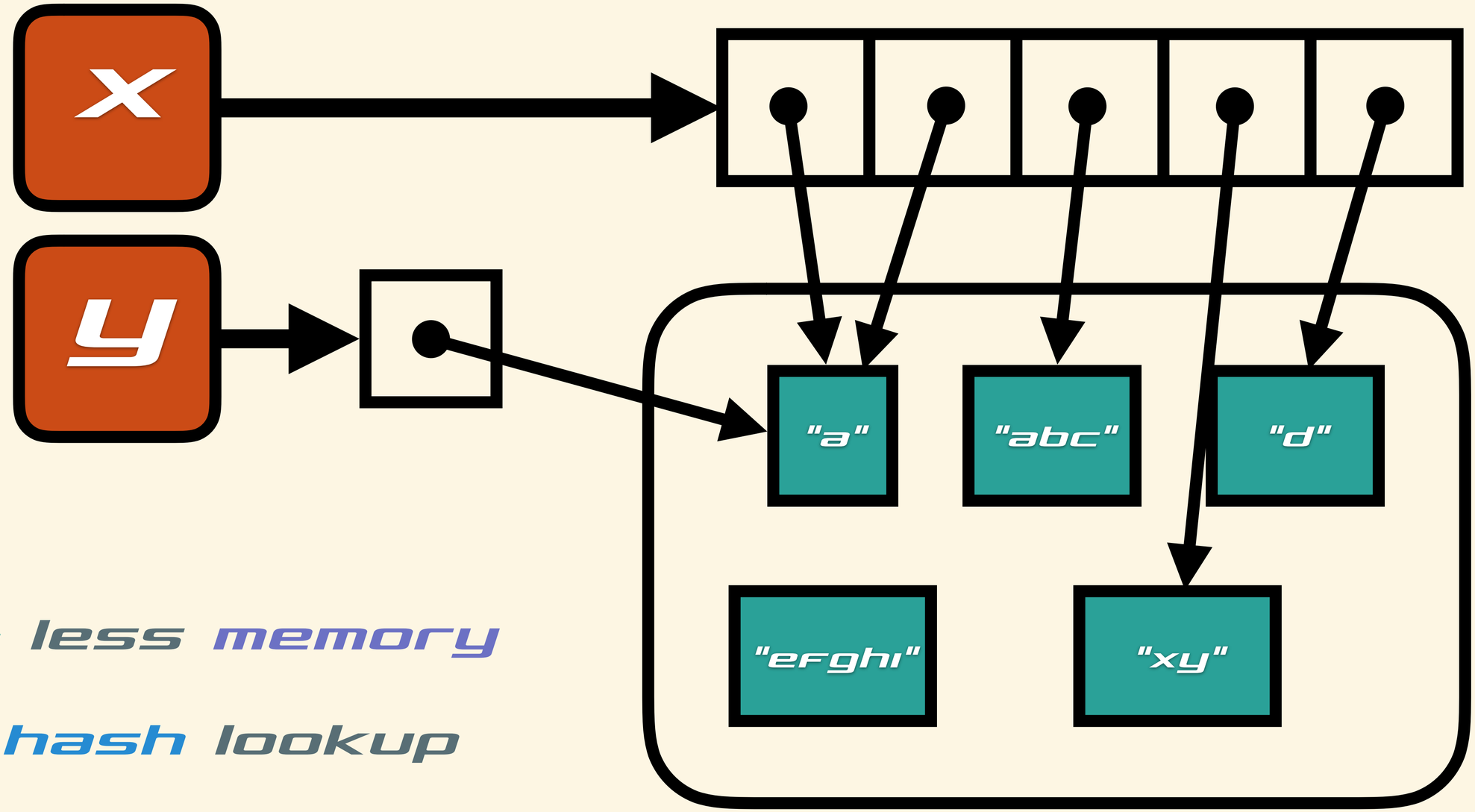


Global string pool



+++ less memory

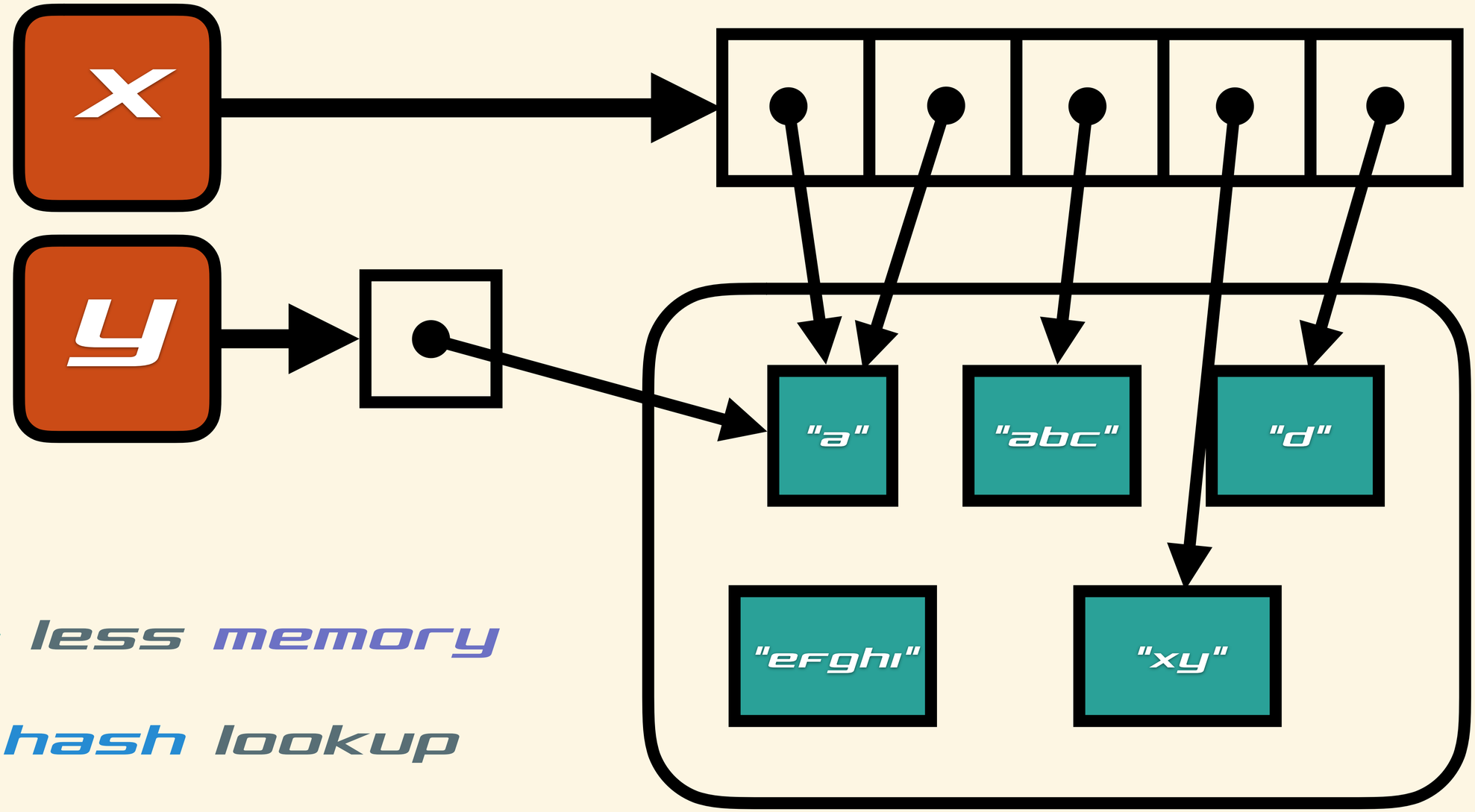
Global string pool



+++ less memory

--- hash lookup

Global string pool



+++ less memory

--- hash lookup

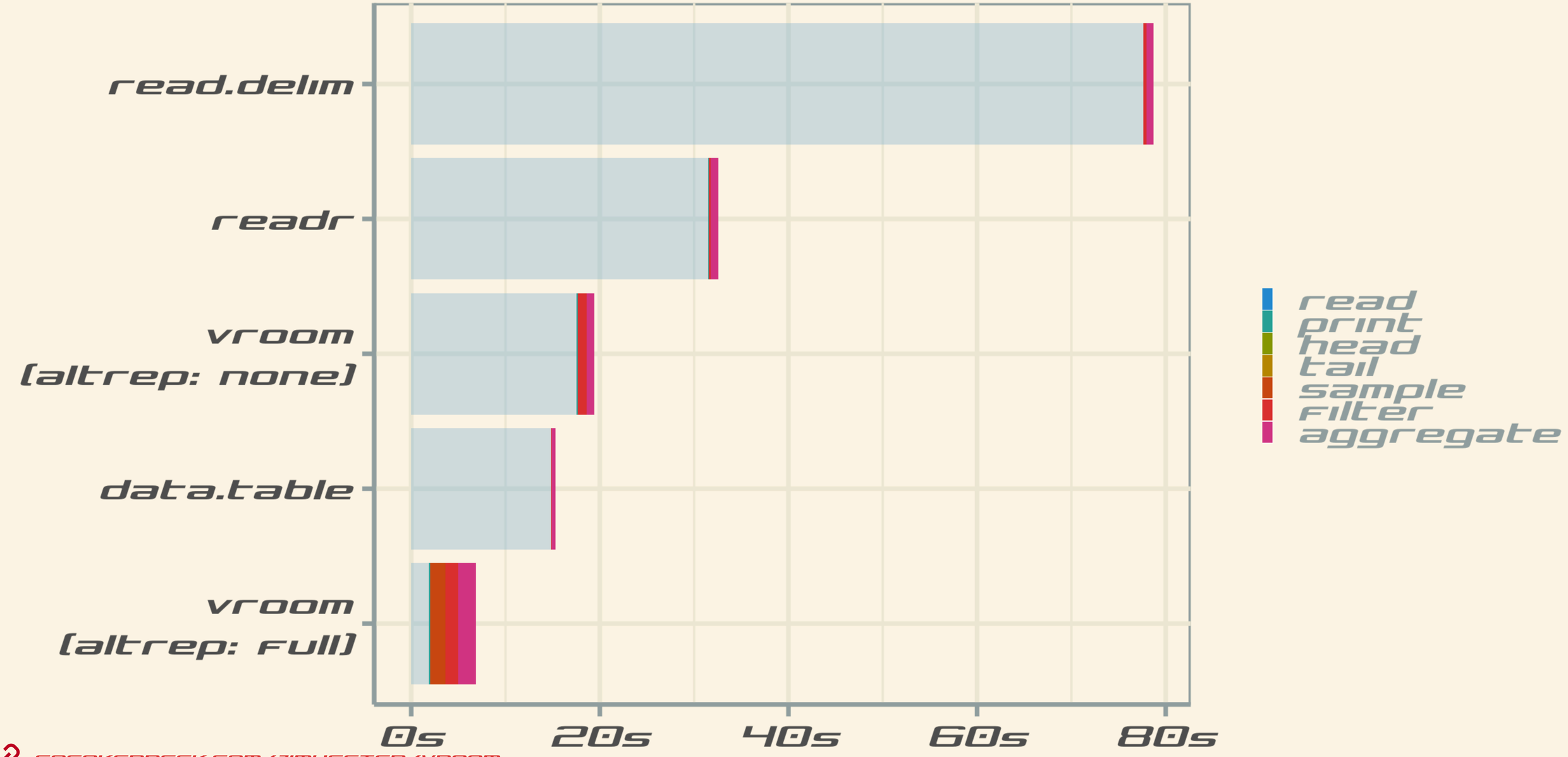
--- single threaded

Global string pool

COST OF LAZINESS

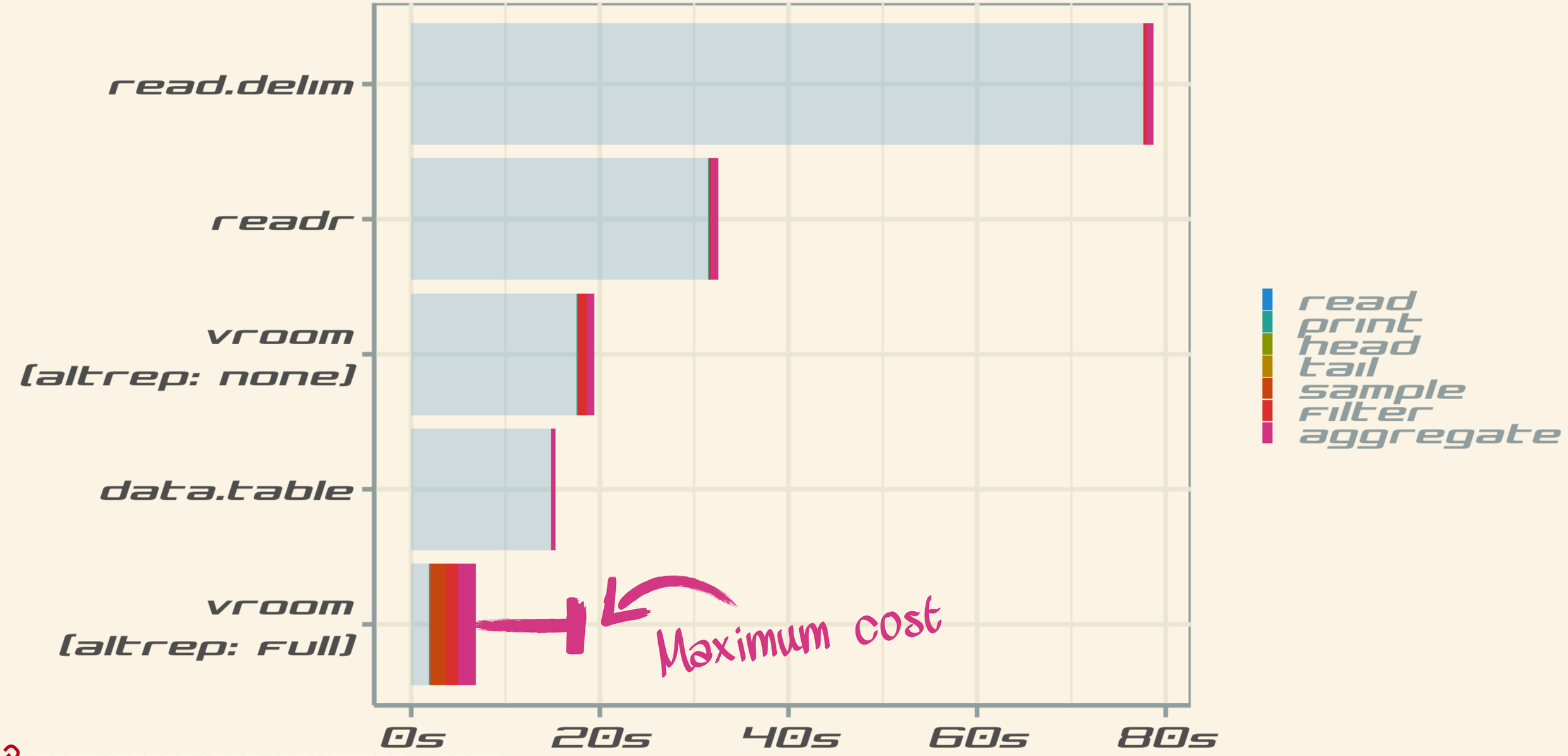
Taxi trip fare

14,776,615 x 11 – 1.55GB



Taxi trip fare

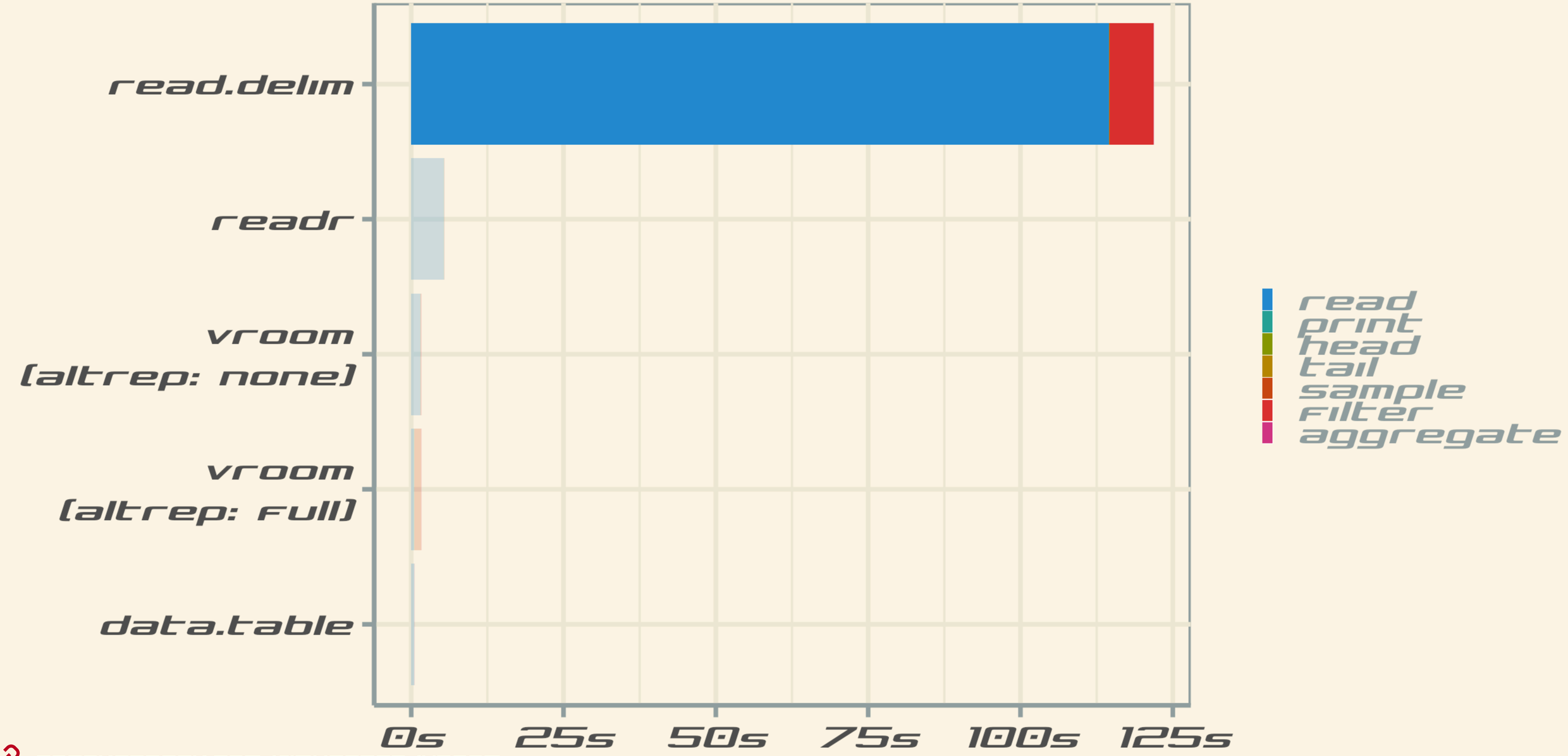
14,776,615 x 11 – 1.55GB



ALL DOUBLES

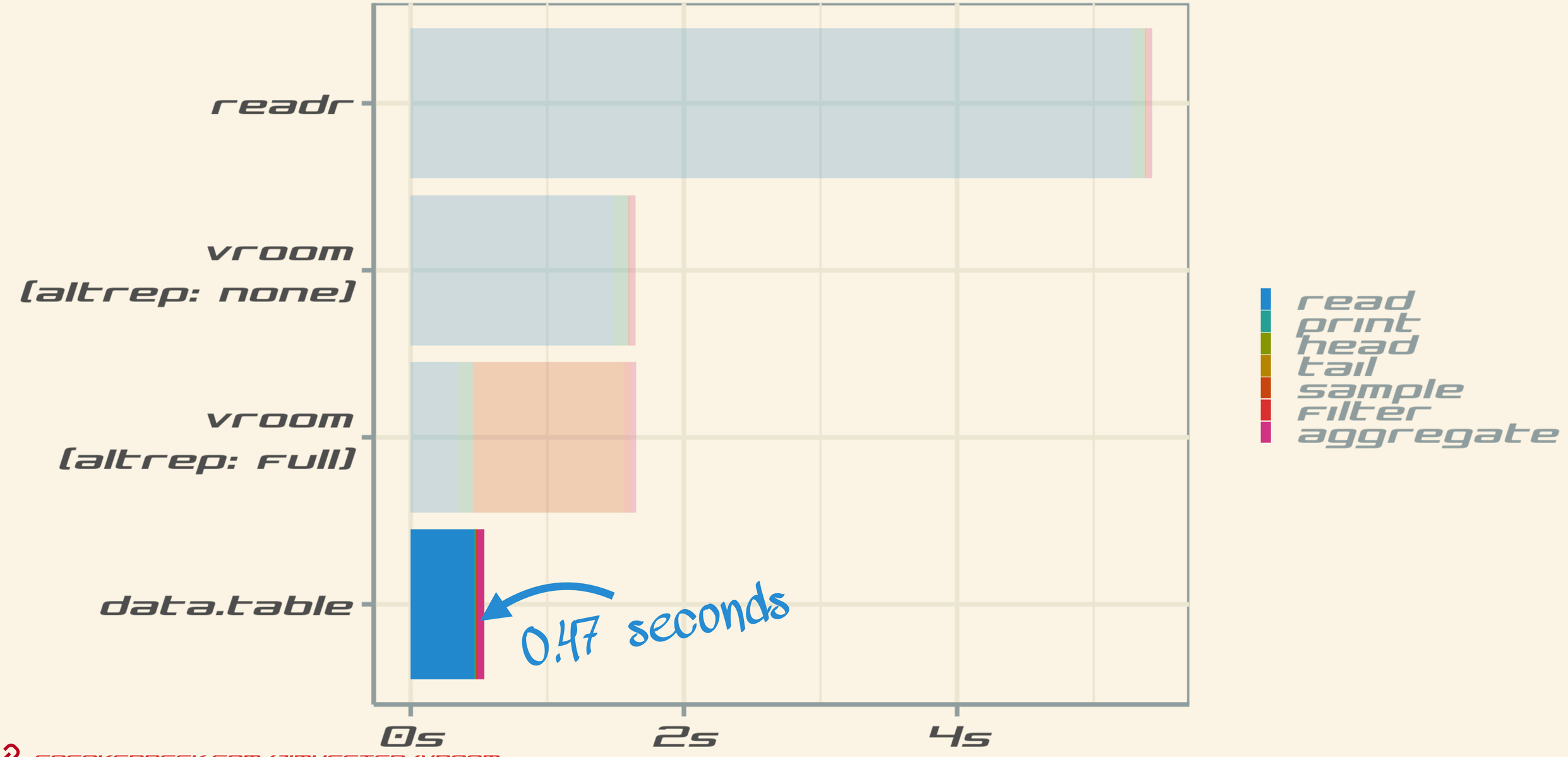
All double

1,000,000 x 25 - 468MB



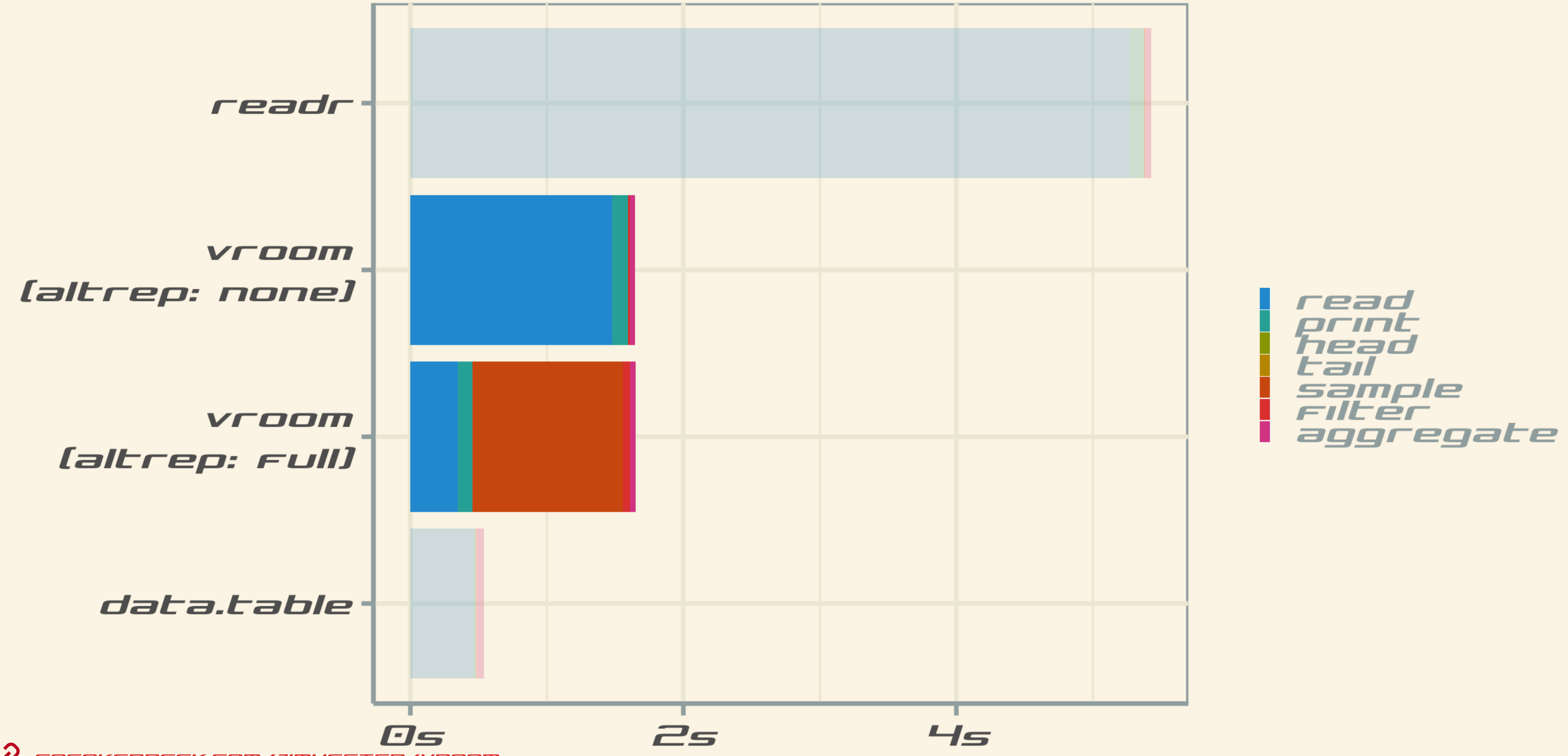
All double

1,000,000 x 25 - 468MB



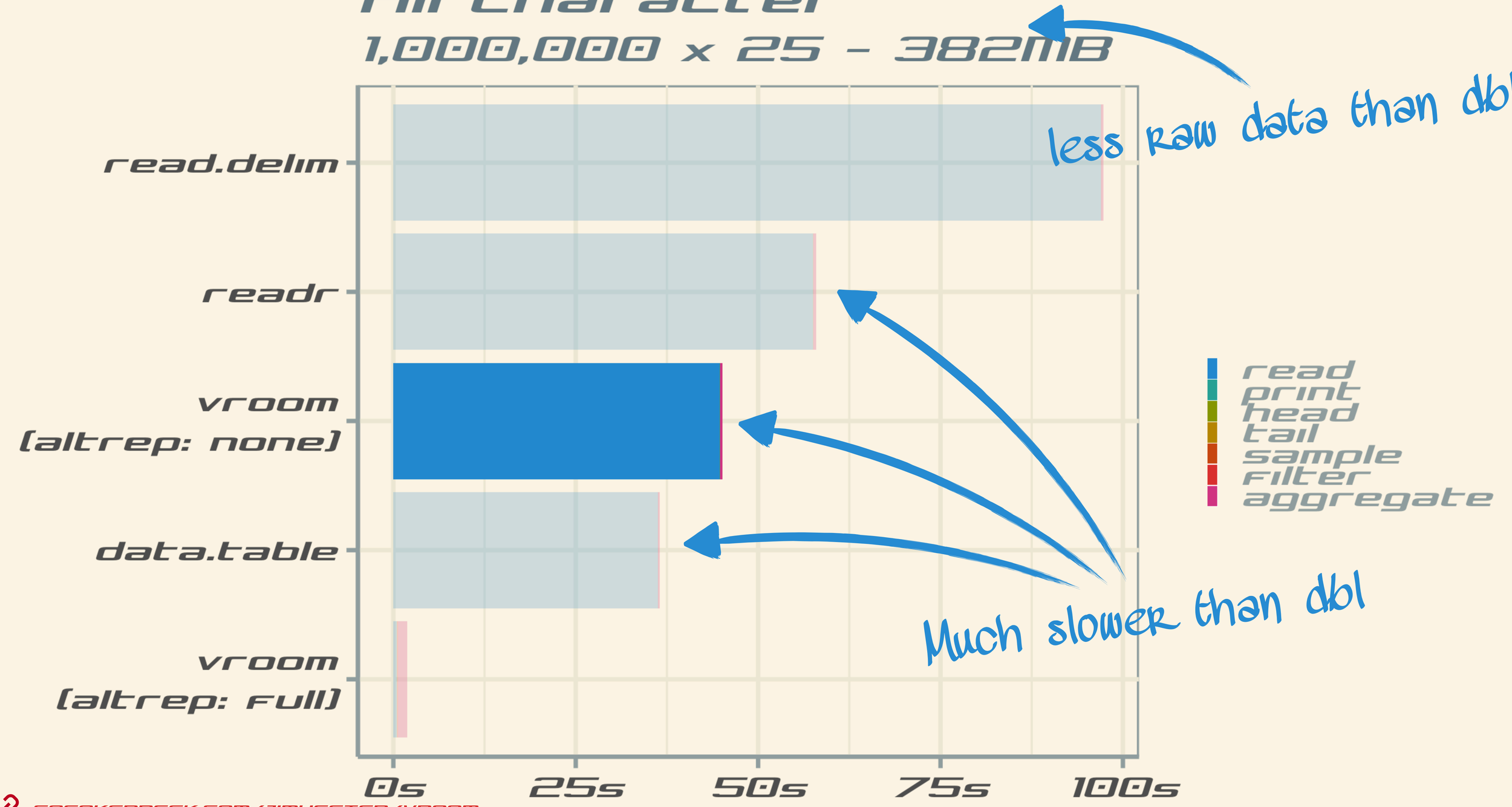
All double

1,000,000 x 25 - 468MB



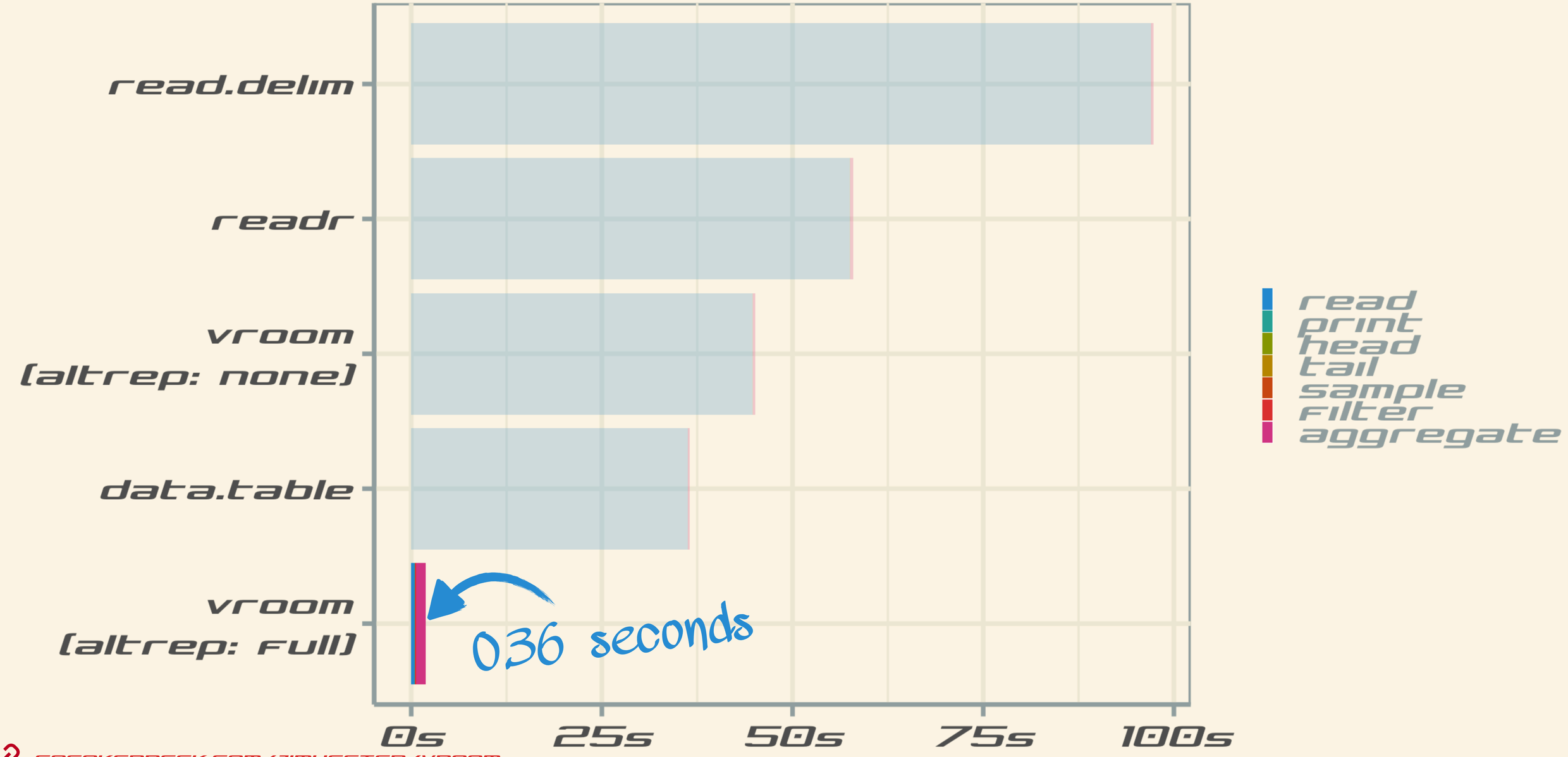
ALL CHARACTERS

All character
1,000,000 x 25 - 382MB



All character

1,000,000 x 25 - 382MB



OTHER FEATURES

SELECTION

```
vroom("taxi.csv") %>%
  dplyr::select(medallion, pickup_datetime, ends_with("amount"))

# select
vroom("taxi.csv",
  col_select = list(medallion, pickup_datetime, ends_with("amount")))

# remove
vroom("taxi.csv", col_select = -hack_license)

# rename
vroom("taxi.csv", col_select = list(taxi = medallion, everything()))
```

SELECTION

```
vroom("taxi.csv") %>%
  dplyr::select(medallion, pickup_datetime, ends_with("amount"))

# select
vroom("taxi.csv",
  col_select = list(medallion, pickup_datetime, ends_with("amount")))

# remove
vroom("taxi.csv", col_select = -hack_license)

# rename
vroom("taxi.csv", col_select = list(taxi = medallion, everything()))
```

SELECTION

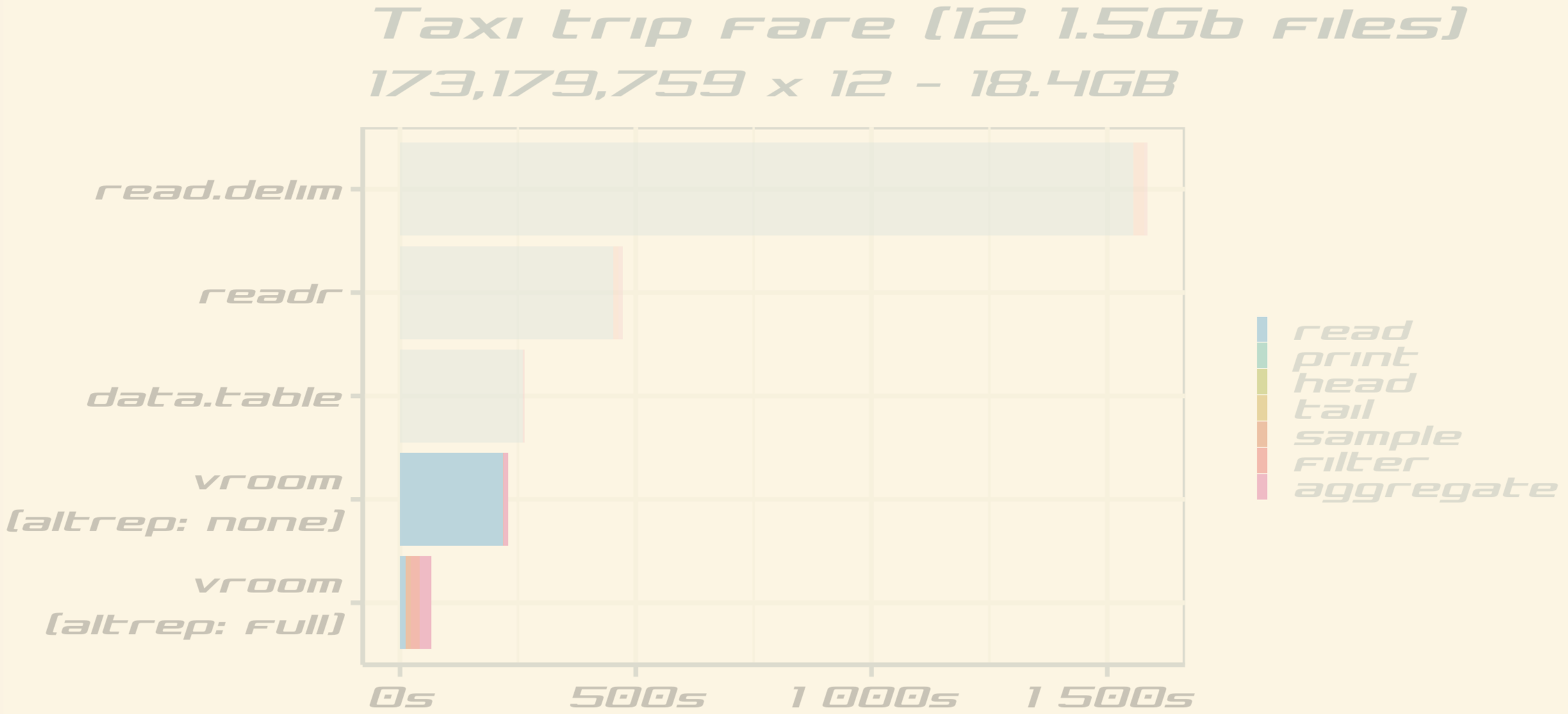
```
vroom("taxi.csv") %>%
  dplyr::select(medallion, pickup_datetime, ends_with("amount"))

# select
vroom("taxi.csv",
  col_select = list(medallion, pickup_datetime, ends_with("amount")))

# remove
vroom("taxi.csv", col_select = -hack_license)

# rename
vroom("taxi.csv", col_select = list(taxi = medallion, everything()))
```

SELECTION

```
vroom("taxi.csv") %>%
  dplyr::select(medallion, pickup_datetime, ends_with("amount"))

# select
vroom("taxi.csv",
  col_select = list(medallion, pickup_datetime, ends_with("amount")))

# remove
vroom("taxi.csv", col_select = -hack_license)

# rename
vroom("taxi.csv", col_select = list(taxi = medallion, everything()))
```

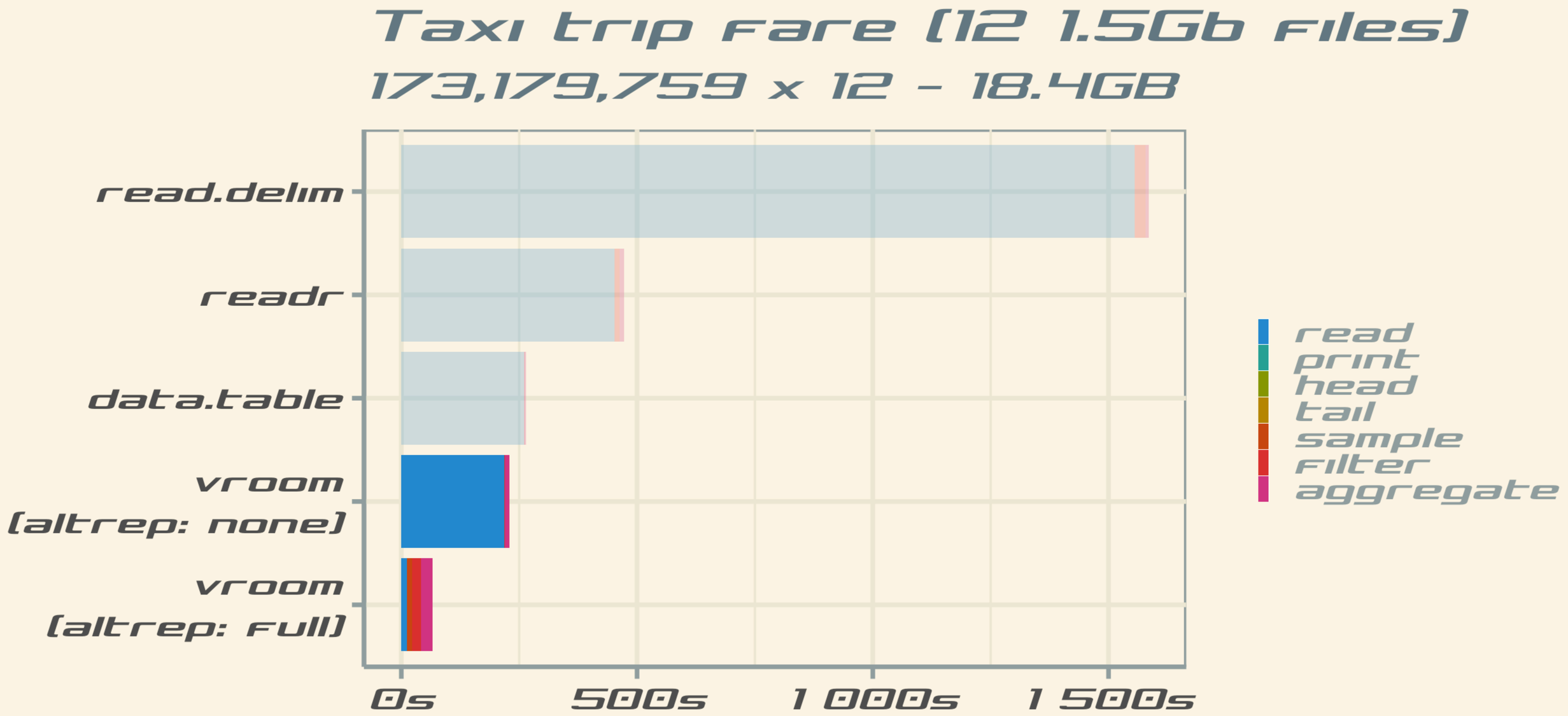
Multiple files

```
vroom(c("taxi_1.csv", "taxi_2.csv"), id = "path")
```



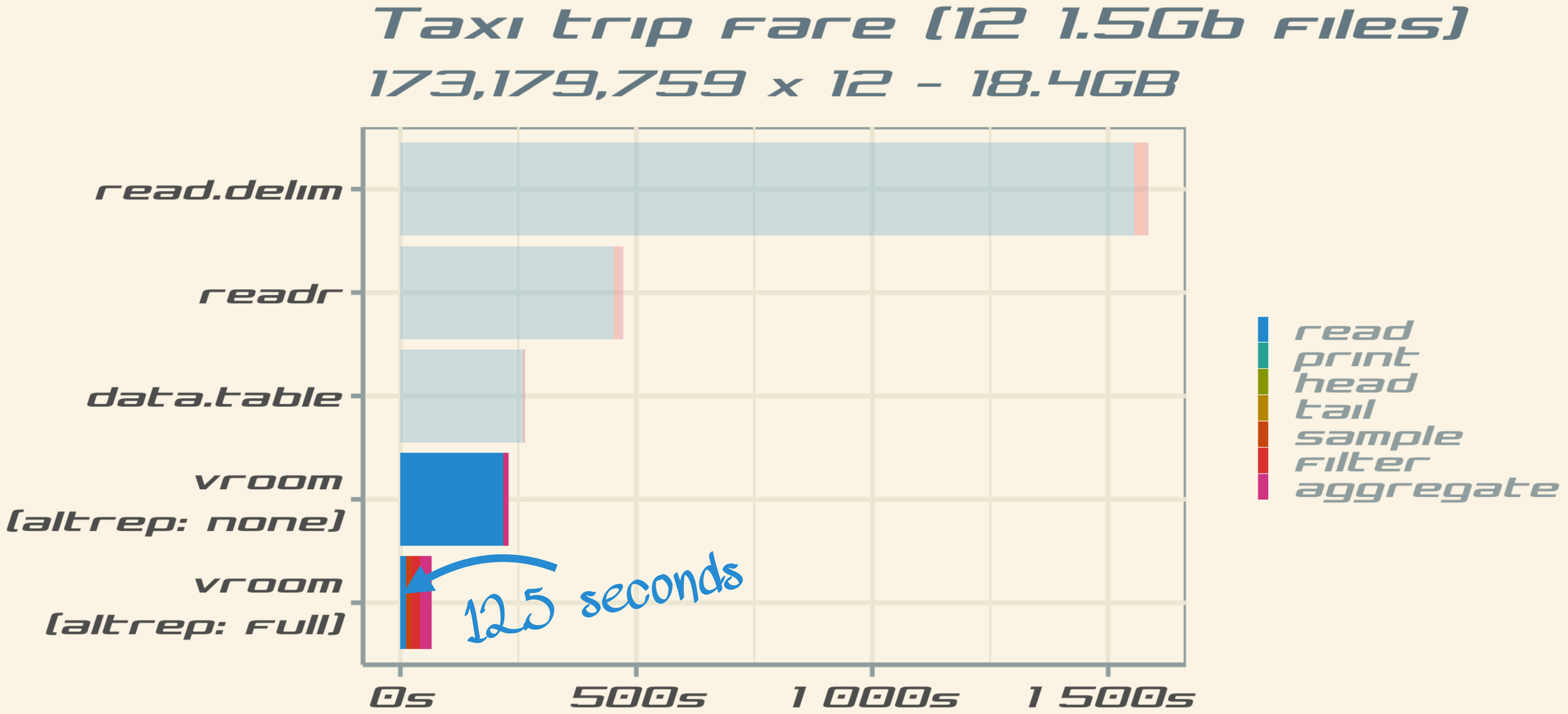
MULTIPLE FILES

```
vroom(c("taxi_1.csv", "taxi_2.csv"), id = "path")
```



MULTIPLE FILES

```
vroom(c("taxi_1.csv", "taxi_2.csv"), id = "path")
```



Fixed Width Files

`vroom_fwf(altrep = TRUE)`	740 ms
`vroom_fwf(altrep = FALSE)`	10.1sec
`readr::read_fwf()`	24.3sec
`read.fwf()`	17.8min

480,174 x 156 - 1.05 Gb file

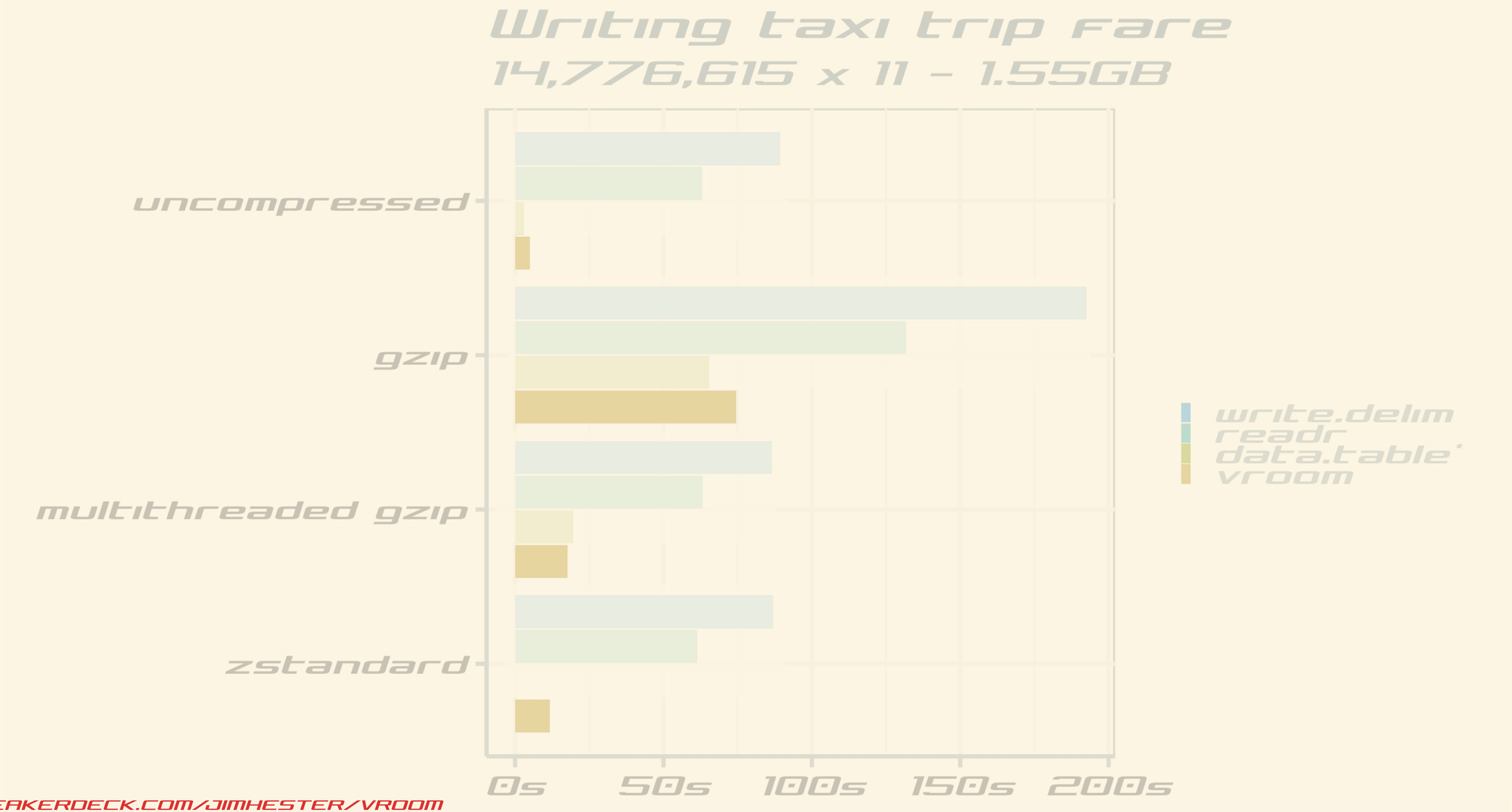
COUNTING LINES

```
length(vroom_lines(file))                                  501 ms
length(data.table::fread(file, sep = "\n",
                         header = FALSE)[1]])             9.19sec
length(readr::read_lines(file))                          11.82sec
length(readLines(file))                                  20.72sec
```

14,776,616 lines - 1.55 Gb file

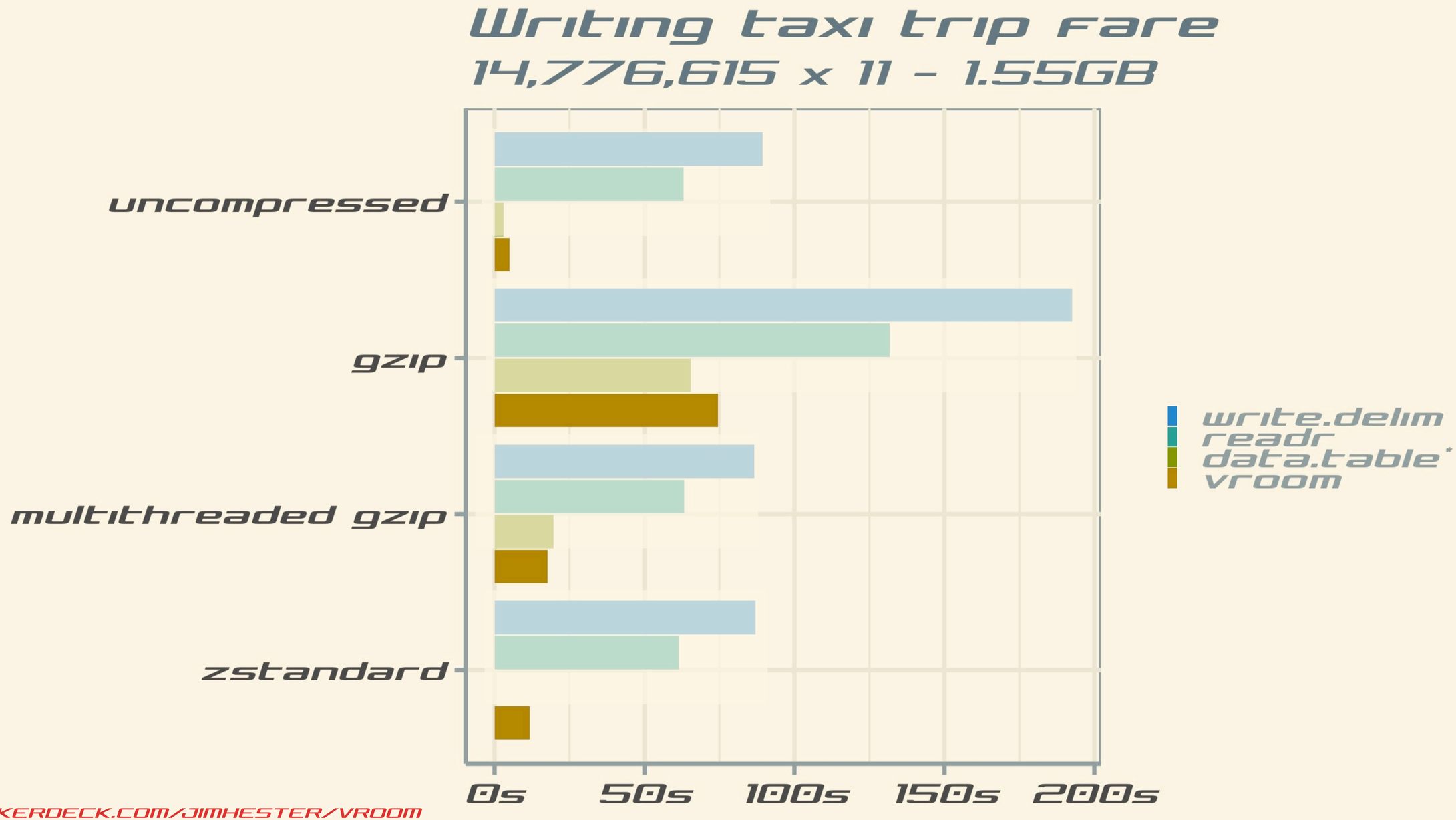
WRITING

```
vroom_write(df, "taxi.tsv.gz", delim = "\t")
```



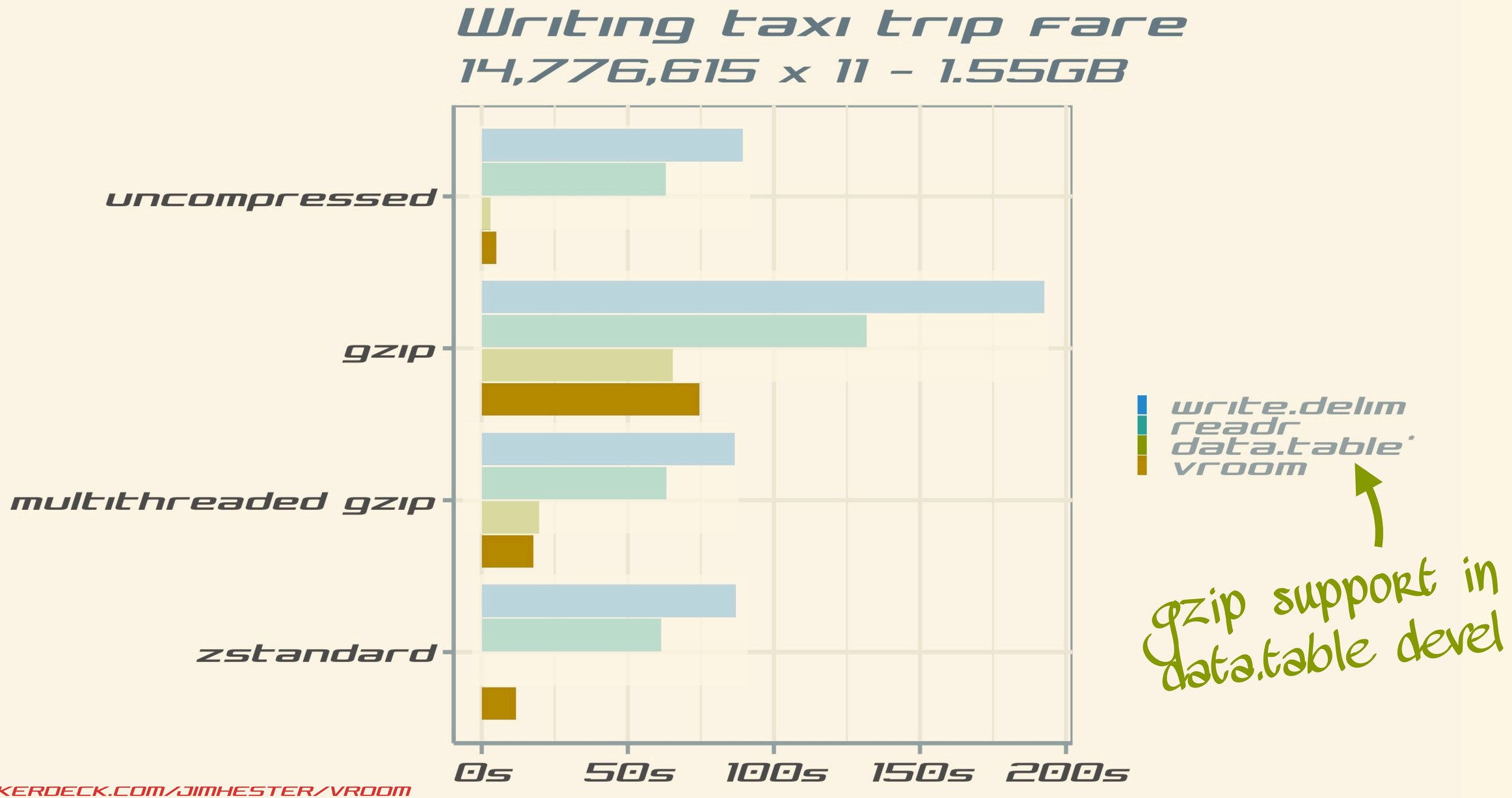
WRITING

```
vroom_write(df, "taxi.tsv.gz", delim = "\t")
```



WRITING

```
vroom_write(df, "taxi.tsv.gz", delim = "\t")
```



VROOM!

install.packages("vroom")

vroom.r-lib.org

bit.ly/vroom-yt

github.com/r-lib/vroom/issues

@jimhester
@jimhester_

SPEAKERDECK.COM/JIMHESTER/VROOM