R package development cla

Toby Dylan Hoc
Assistant Profe
Northern Arizona U
toby.hocking@r-pro

talk overvi

1. data.table NSF POSE project
2. rOpenSci Statistical Software Pe
3. Google Summer of Code

who am i

- BA, MS, PHD in Stats/Math (mach
- Assistant Professor of computer sc
- Using R since 2003! 20+ years! Au
- data.table user since 2015, contrib
- Principal Investigator, NSF Pathwa

  Open-Source Ecosystems (POSE)

1/3
data.table NSE
project

data.frame

- 2D *columnar* data structure
  - rows and columns
- subset rows — DF[DF$id != “a”,]
- select columns — DF[, “val”]
- subset rows & select columns — DF[DF$id != “a”, “val”]
- that’s pretty much it…

data.table

- Like data.frame, but with more powerful R code syntax, and C code implementation
- R package on CRAN since 2006
- Created by Matt Dowle, co-author Arun Srinivasan since 2013, 50+ contributors
- 1463 other CRAN packages require data.tab (in most popular 0.05% of all CRAN packag

Comparing data.table

- tidyverse R package 1.0 on CRAN in 2016
- tidyverse packages tibble + readr + tidyr + dplyr ~= c
- tidyverse uses DF |> ... |>, data.table uses DT[...][...]
- tidyverse is verbose (lots of code), data.table is conc

  example: tibble |> filter(x=="a") |> group_by(z) |> sum

  vs: DT[x=="a", .(m=mean(y)), by=z]
- tidyverse has many dependencies, data.table has no
- tidyverse has frequent breaking changes, data.table

  (easier for users to upgrade to new data.table version

Why is data.t
popular/pow
(efficien

two kinds of data tab

- **Efficient R code syntax (saves programmi**
- Efficient C code implementation (saves time

  data sets can be analyzed using smaller com

data table R cod

- think in terms of — rows, what to do with co
- Matt's 2014 useR talk https://youtu.be/qL

General form: DT[i, j, by

data.frame(DF) vs da

```
sum(DF[DF$code != “abd”, “valA”])
```

```
DT[code !
```

- Consider subset of rows with "abd" ir compute sum of values in **valA** colum
- DF needs to be repeated, no repetitic
- sum can be placed in the square brack than outside with DF.

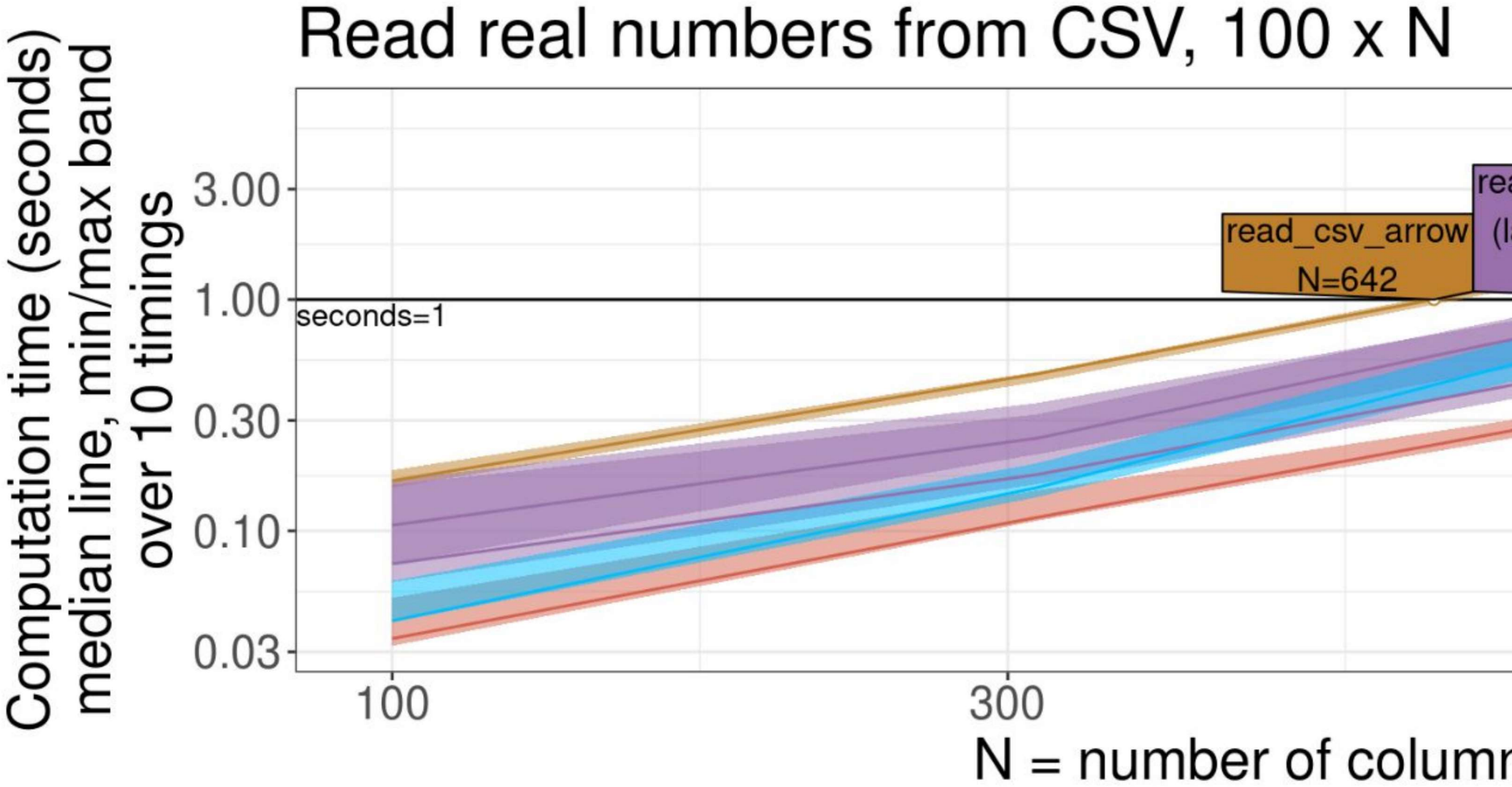
two kinds of data tab

- Efficient R code syntax (saves programming
- **Efficient C code implementation (saves ti**

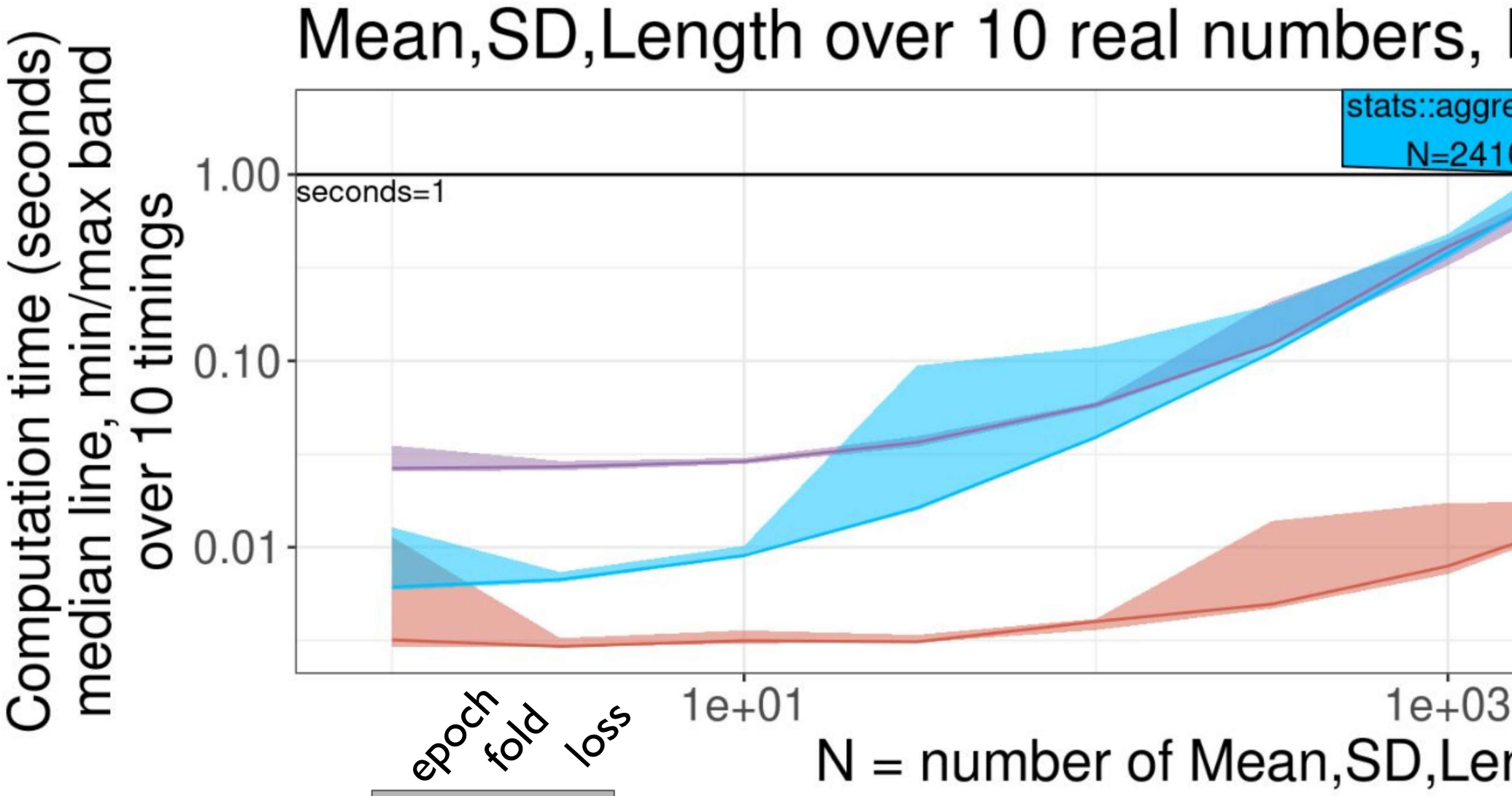
  larger data sets can be analyzed using sm

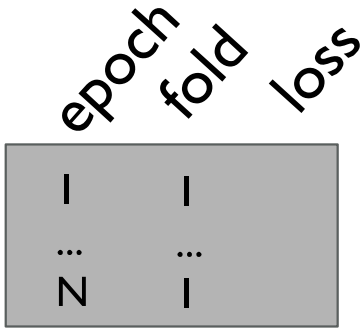
  resources)

data.table::fread is an extremely effic



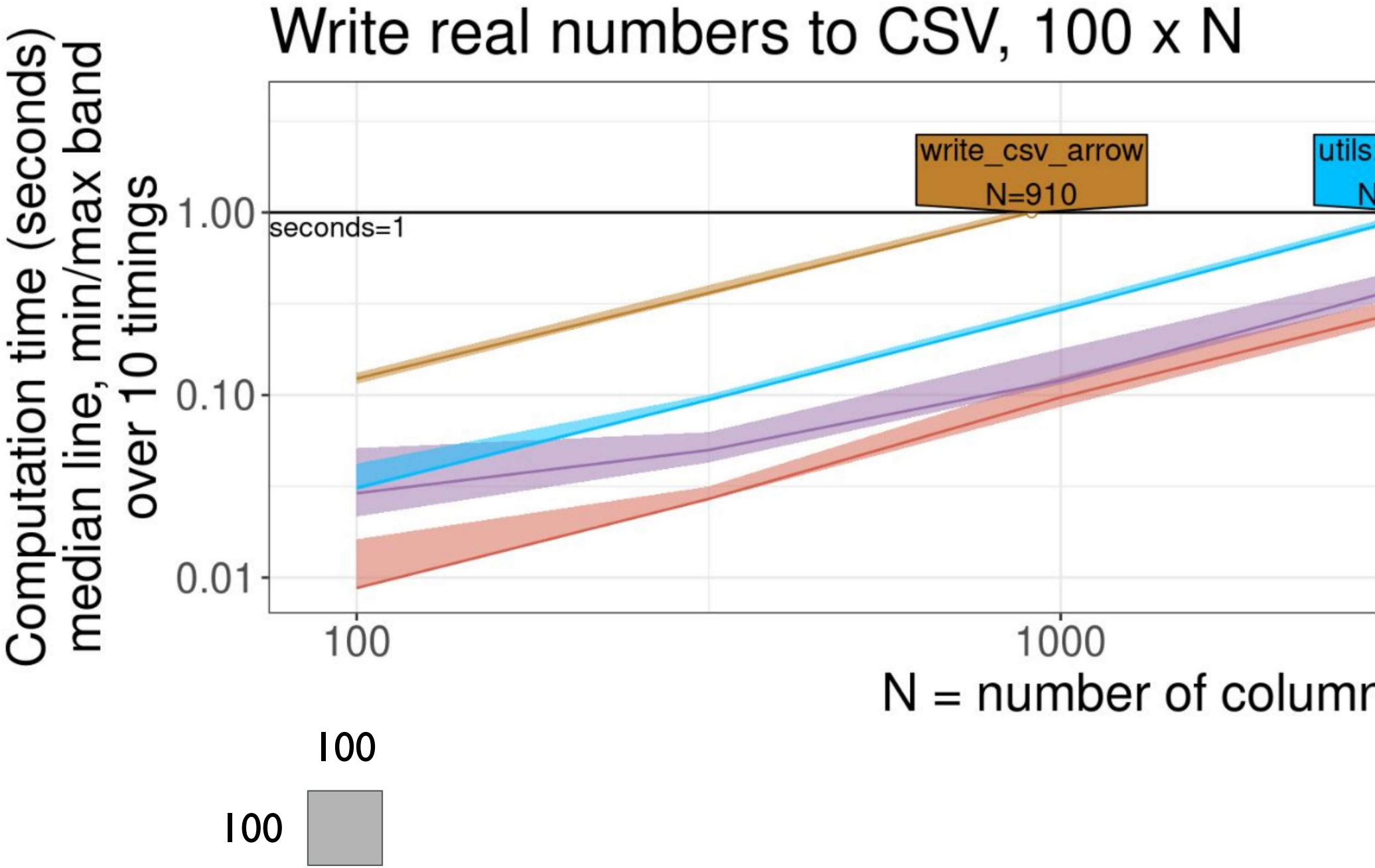
data.table computes summaries 100





In mach
estimat

data.table::fwrite is an extremely effi



most underra
package

Conor Nash
@conornash

Data.table is the most underrate
has saved me *days* in waiting
complete.

powerful

Alexander Flyax
@aflyax

somebody should just write a ve
data.table for #python. end of st
powerful exists at the moment.

data.table data

Joey Reid
@JoeyPReid

data.table

data.table
data.table
data.table
data.table
ggplot2
rstan
knitr

great sadne

Jim Savage
@khakieconomist

With great sadness I was force
data.table today.

Using data.ta
efficient big data

See https://github.com/tdhock/2023-10-Latin
hour tutorial presentation slides, with code,

Contributing to c
we need you

Community blog a

- data.table mascot is a sea lion, which ba
- data.table community has a new blog, T

  https://rdatatable-community.github.io/T

  sea lions often float together on the ocea

  called "rafts." - Marine Mammal Center

GitHub repos

- data.table has an active issue/Pull Request(PR) t
  https://github.com/Rdatatable/data.table/
- 1000+ open issues, 100+ open PRs
- if you have any time/interest, we could use your h
- easy first contribution: try reproducing an issue
  (very helpful to know if an issue is reproducible)
- very inclusive community -- after you submit you
  join the github group!
- now is a very exciting time to get involved, as we

Translation Aw

- In 2023-2025, National Science Foundation ha
  expanding the ecosystem of users and contrib
- 20 translation awards, US$500 each, in order
  messages more accessible, ideas:
- Translate errors/warnings/messages (potools
- Translate most important vignettes (intro, impo
- Translate other documentation (cheat sheets,
- Priority: Portuguese, Spanish, Chinese, Frenc

Travel awar

- In 2023-2025, National Science Foundation has
  expanding the ecosystem of users and contribu
- Eight travel awards, US$2700 each
- Candidates should give a talk about data.table
  relevant audience (potential data.table users or
- Call for proposals on https://rdatatable-commu

Summary of da

- concise, consistent syntax
- time and memory efficient
- No dependencies (easy to install)
- No breaking changes (easy to upgrade)
- Inclusive user/developer communi

  to contribute:

- translation awards, US$500 each

2/3

rOpenSci Sta
Software Peer

Sharing R package

- R package is the formal unit of code shar
- Package includes documentation, tests, v
- CRAN is the Comprehensive R Archive N most widely used package repository
- CRAN regularly checks every package to correctly (and compatible with all other pa benefit for the R community.

Example of review f

```
Please rather use the Authors@R field and declare Ma
and Contributors with their appropriate roles with p
e.g. something like:
Authors@R: c(person("Alice", "Developer", role = c("
email = "alice.developer@some.domain.net"),
person("Bob", "Dev", role = "aut") )

Please always write package names, software names an
programming interface) names in single quotes in tit
e.g: --> 'C'; 'PCRE'; ...
Please note that package names are case sensitive.
```

Peer-reviewed resea

- Peer-reviewed paper is the formal unit of
- Peer reviewers typically read the manuscr substantive feedback/suggestions about
- But typically peer reviewers do not read t the computations in those papers. There example Journal of Statistical Software.

Example of peer

(**) Section 1 and abstract: The proposed model uses
along the data to constrain the inference during the
programming algorithm. I don't understand the reason
discussing train and test data. Training is typicall
coefficients of variables. In this case, the learnin
intervals with the presence or absence of a change.
by experts using already known data (training set).
ring to apriori information would be less confusing.

...

(**) Page 10 Algorithm and line 48: "... by the Deco
We would need more details about the recovering of v
tracking step the reason why the authors used notati
description of the costs? Do we have vector equality

rOpenSci Guide an

- Goal is to provide thorough peer review o packages (code/documentation/etc), ana do for peer-reviewed research papers.
- https://stats-devguide.ropensci.org/
- case study of canaper R package peer re https://github.com/ropensci/software-rev

3/3

Google Summe

What is GSO

- 3 month free/open-source coding project
- You don't have to be a coding expert to p
- Goal is to teach you how to contribute to
  software projects
- I have been co-administrator of R project
  Code for 10+ years, and I have mentored
  (typically college students)

Guide and exa

- List of all organizations

  https://summerofcode.withgoogle.com/p

  ons

- R project idea list:

  https://github.com/rstats-gsoc/gsoc2024

  posed%20coding%20projects

Thank you
Questions

Toby Dylan Hoc
Assistant Profe
Northern Arizona U
toby.hocking@r-pro