The R package **bigstatsr**: Memory- and Computation-Efficient Tools for Big Matrices

useR!2017 lightning talk

Florian Privé (@privefl)

July 6, 2017

About

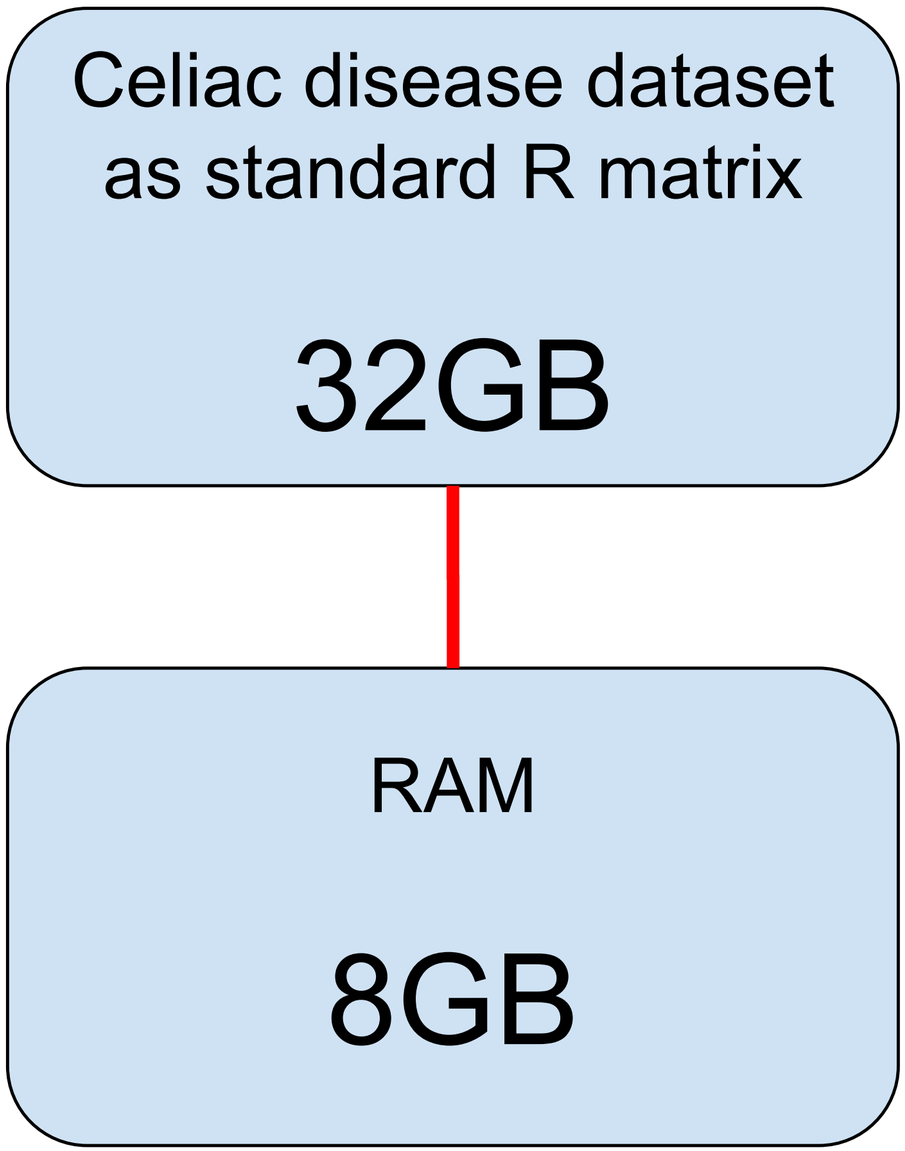
I'm a PhD Student (2016-2019) in **Predictive Human Genetics** in Grenoble.

$$\boxed{\text{Disease} \sim \text{DNA mutations}}$$

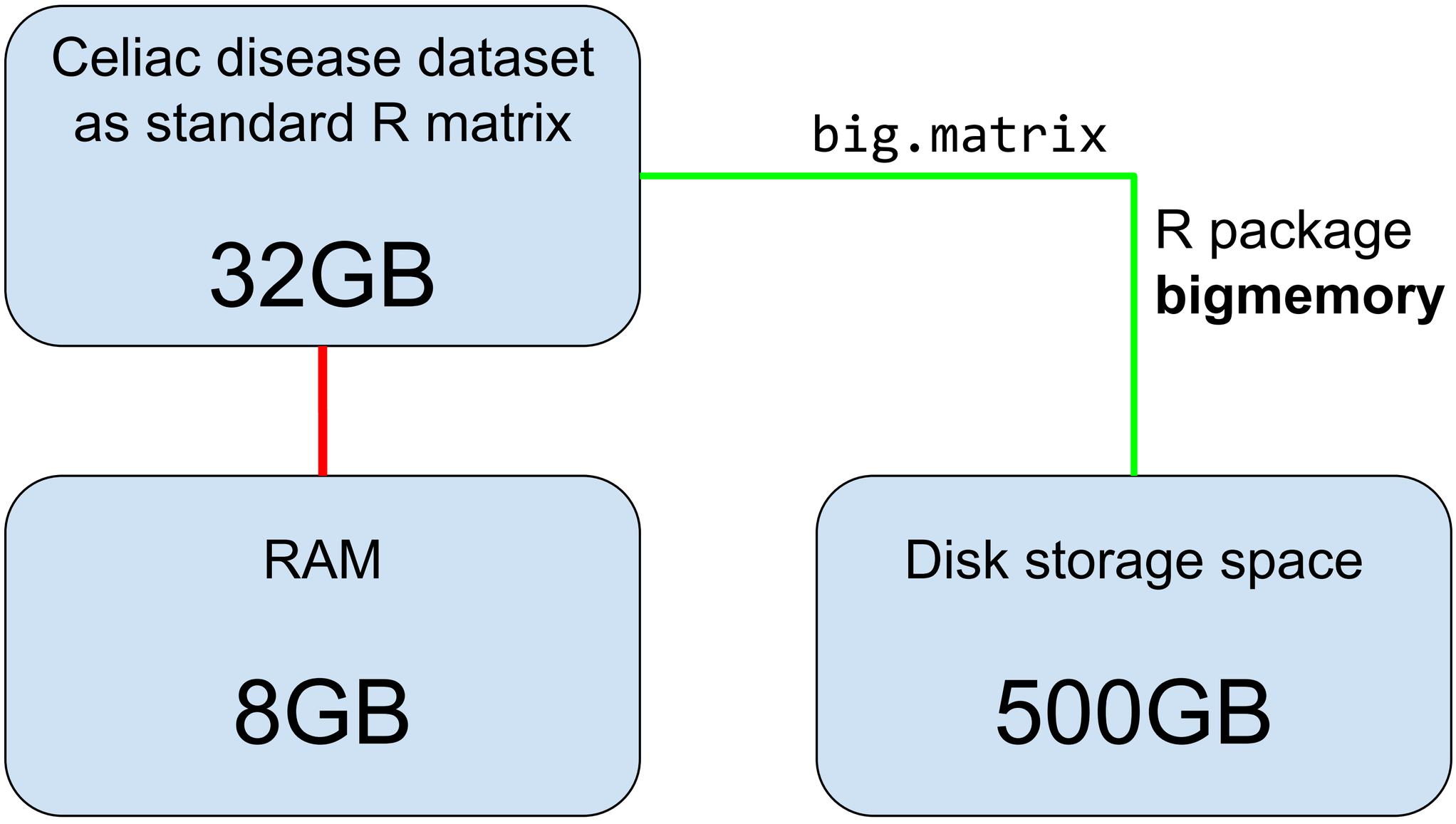
Very large genotype matrices

- currently: 15K x 300K, celiac disease
- soon: 500K x 800K, UK Biobank

Problem I had



Solution I found

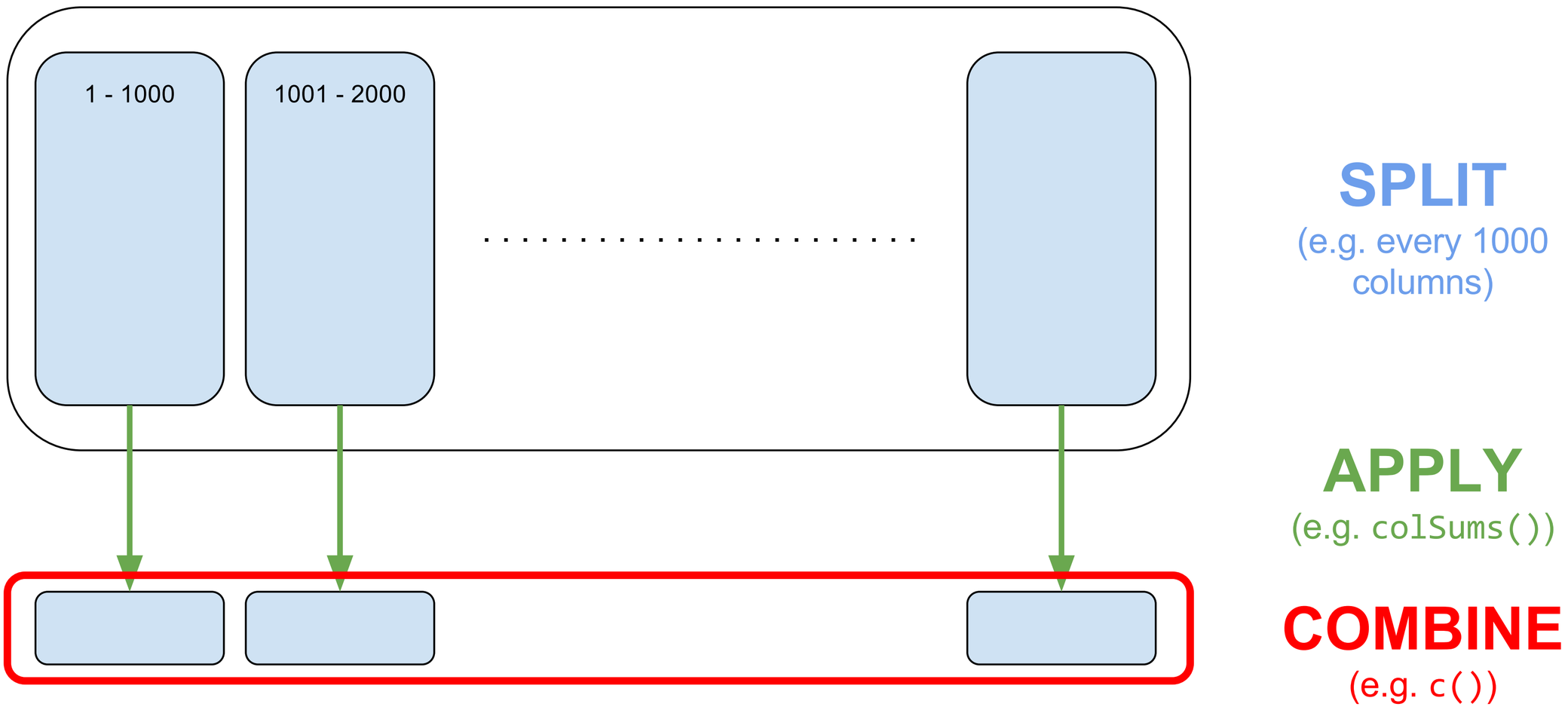


Michael J. Kane, John Emerson, Stephen Weston (2013).

Similar accessor as R matrices

Split-(par)Apply-Combine Strategy

Apply standard R functions to big matrices (in parallel)



strategy coined by Hadley Wickham (2011)

Similar accessor as Rcpp matrices

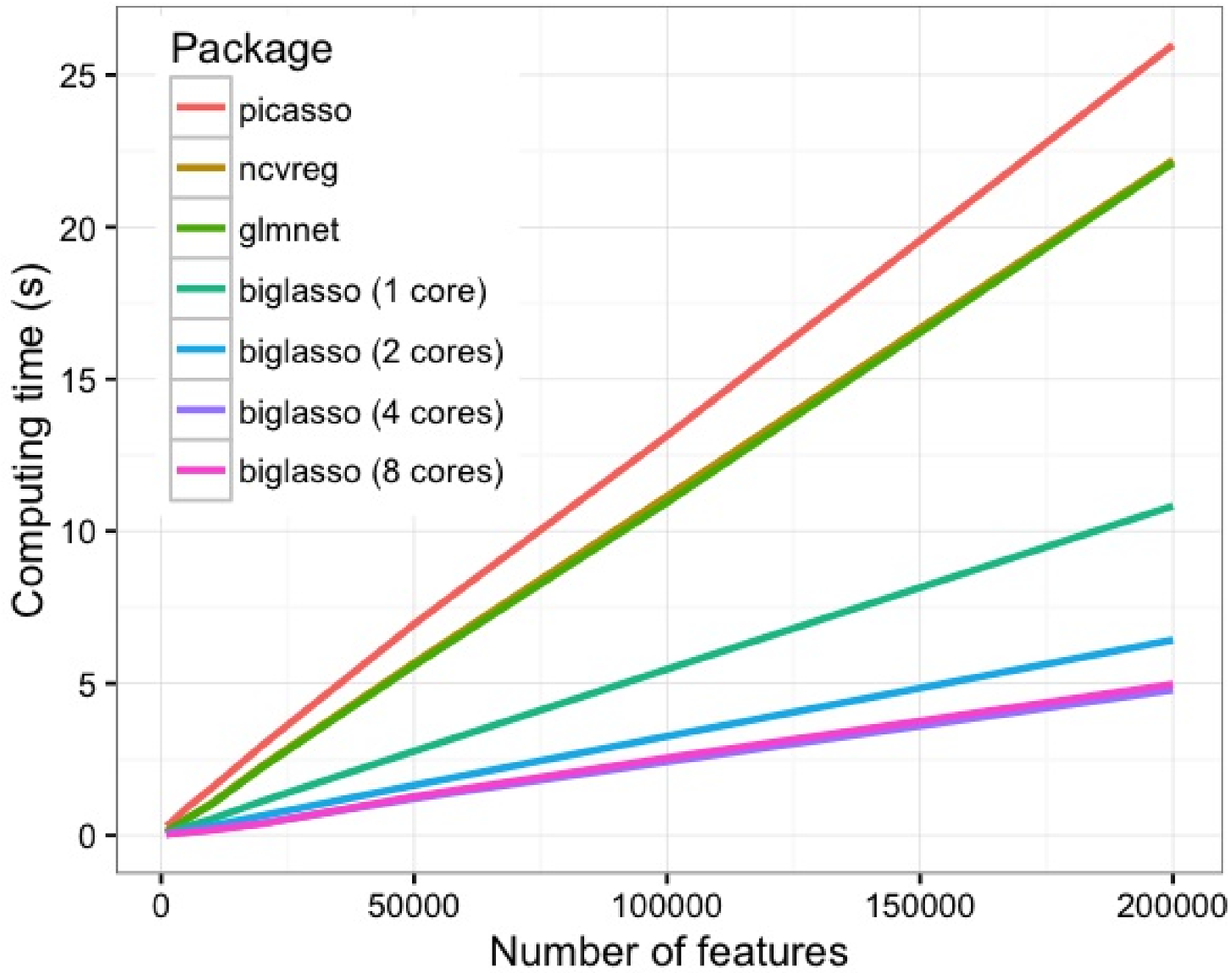
In Rcpp we trust

Partial Singular Value Decomposition

15K x 100K `big.matrix`, 6 cores, K = 10, **1 min** (vs 2h in base R)

based on R package **RSpectra**

Sparse linear models: **biglasso**



Zeng, Y., and Breheny, P. (2017).

Other functions

- matrix operations (Split-Apply-Combine strategy)
- association of each variable with an output (RcppArmadillo)
- plotting functions (ggplot2)
- read from text files
- others..

I'm now able

to run algorithms

on 100GB of data

R Packages

bigmemory	`big.matrix` object
bigstatsr	Statistical functions for `big.matrix` objects **to be used by any field**
bigsnpr	Specific functions for my field of research

Paper in preparation: "Efficient management and analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr".

Contributors are welcomed!

Thanks!

Package's website: https://privefl.github.io/bigstatsr/

Twitter and GitHub: @privefl

Presentation available online: https://goo.gl/nNg0hw

Slides created via the R package **xaringan**.