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

useR!2017 lightning talk

Florian Privé (@privefl)

July 6, 2017

About

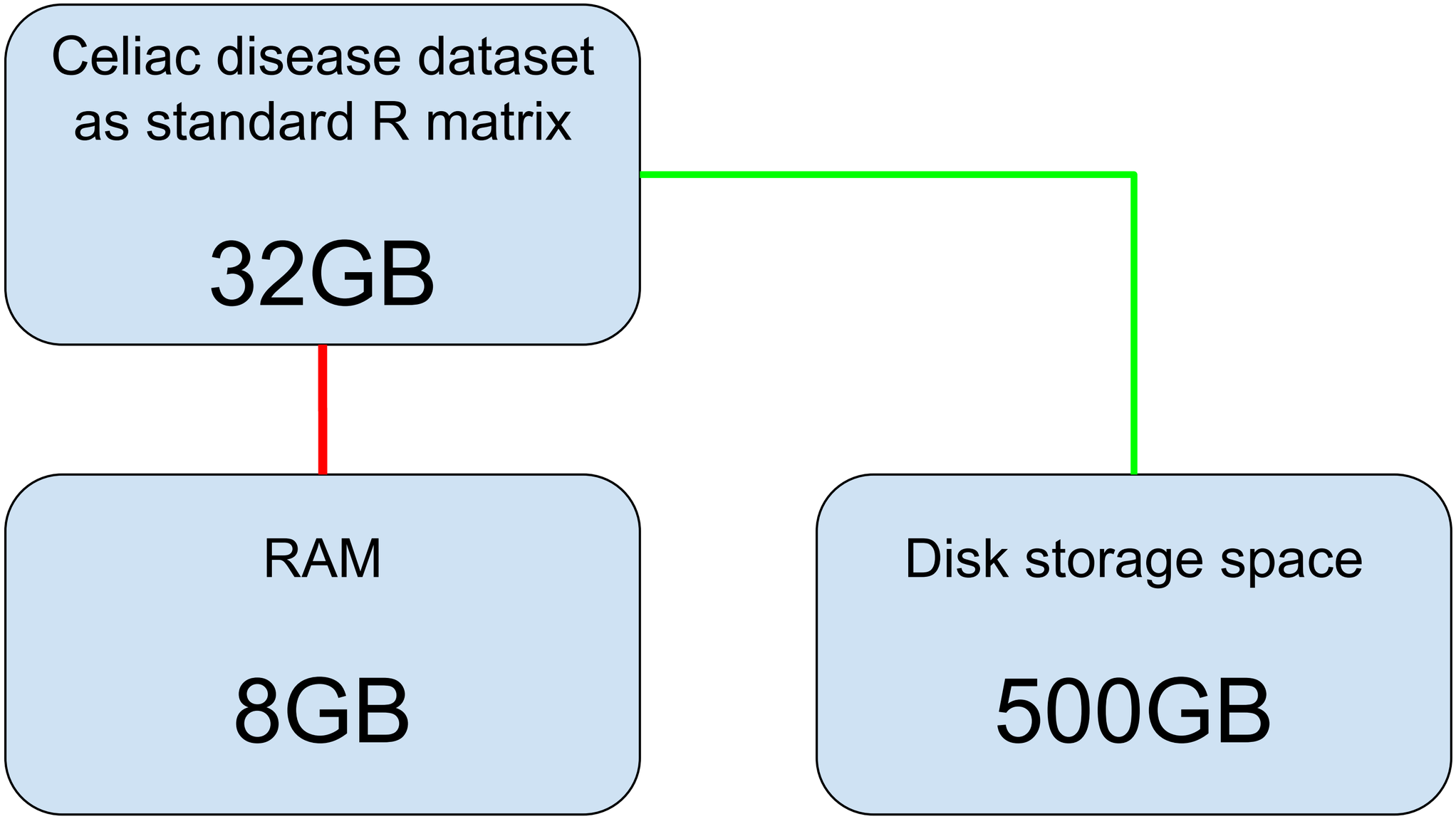
I'm a PhD Student (2016-2019) in **Predictive Human Genetics** in Grenoble.

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

Very large genotype matrices

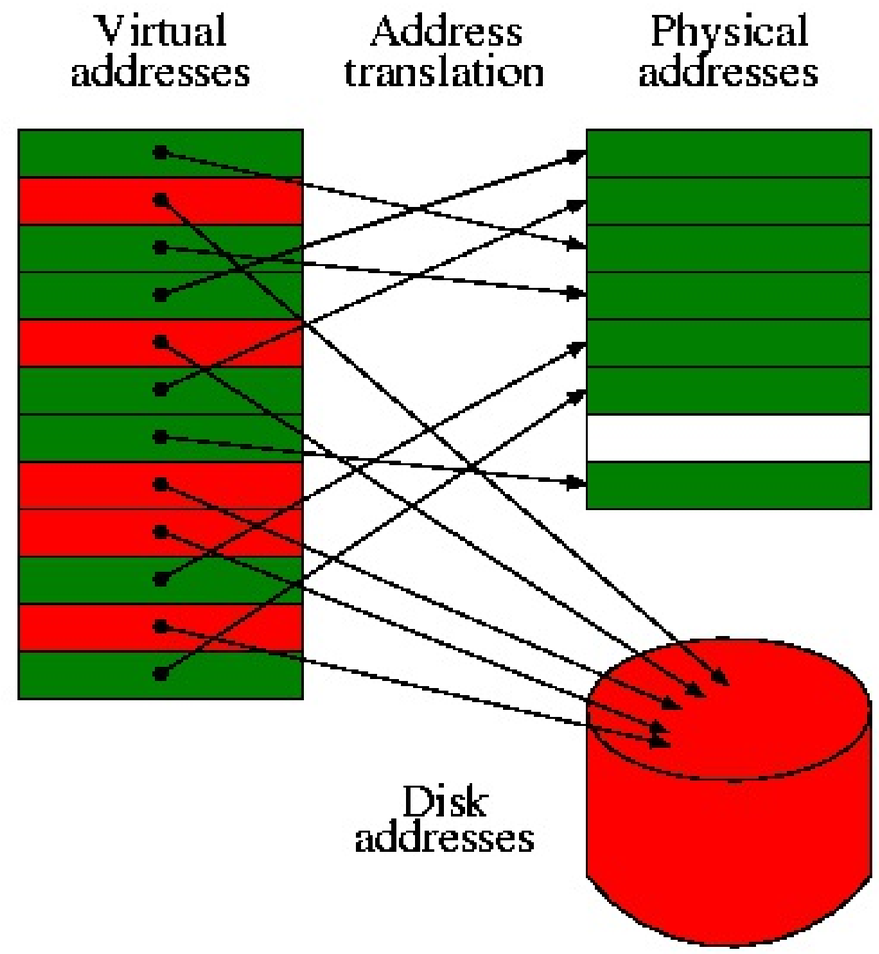
- currently: 15K x 300K, celiac disease
- soon: 500K x 800K, UK Biobank

Problem I had



Solution I found: R package bigmemory

Store matrices on disk and access them from there



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

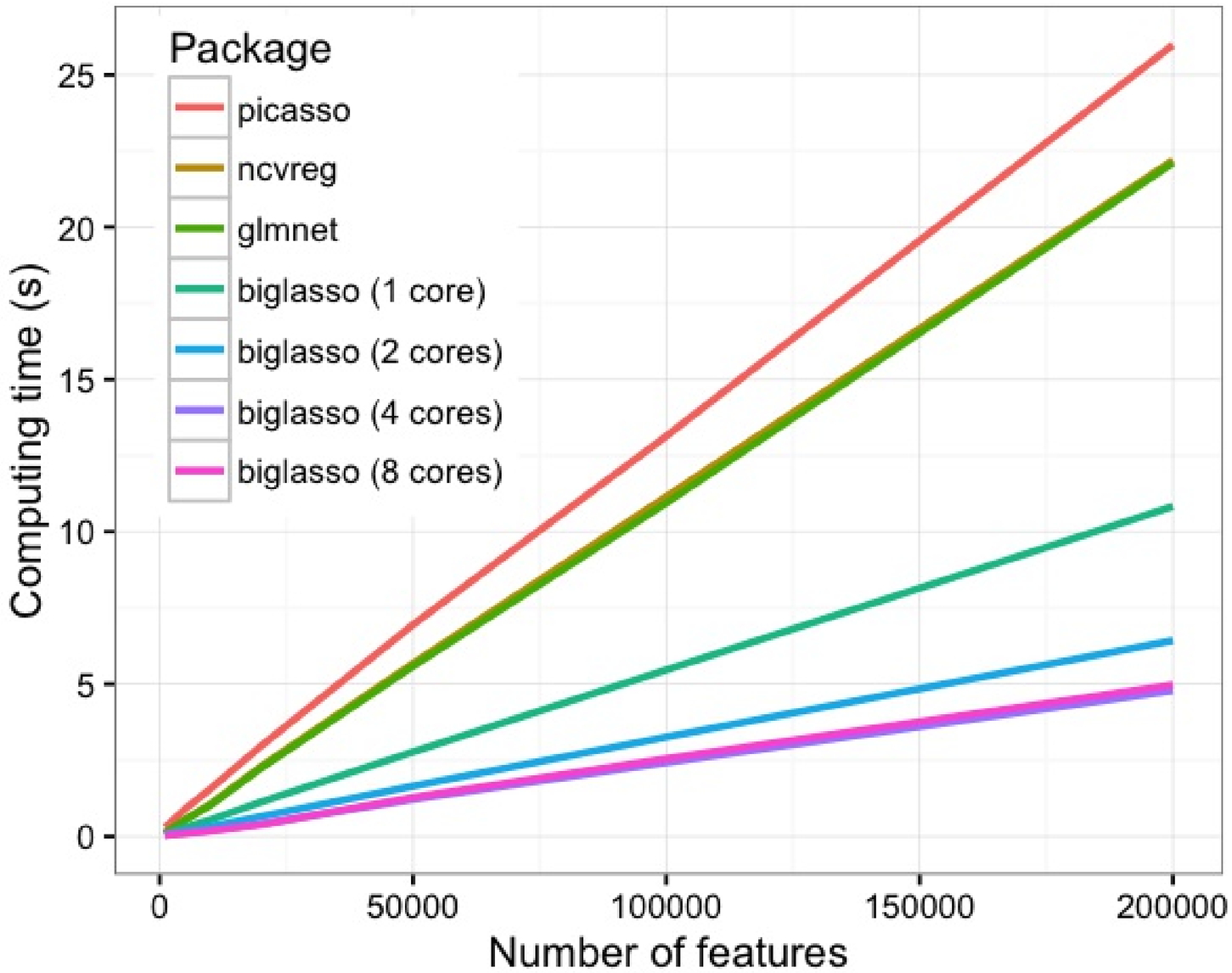
Access almost as if the matrix were in memory

- in **R**: accesses with `[`, as standard R matrices,
- in **Rcpp**: accesses single elements with `X(i, j)`, as standard Rcpp matrices.

Dependencies on bigmemory

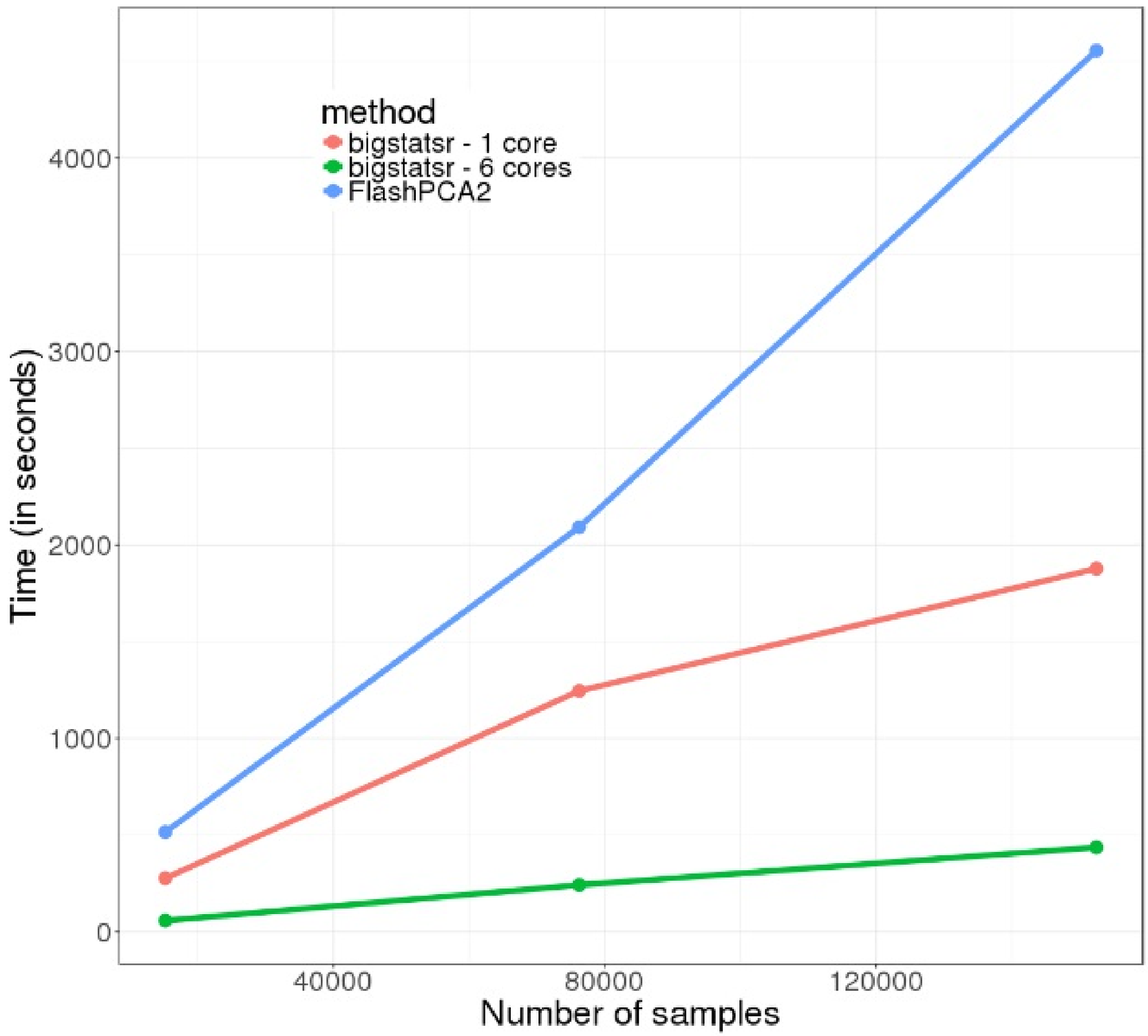
Reverse depends:	bigalgebra, biganalytics, bigFastlm, bigKRLS, biglasso, bigpca, bigtabulate, GHap, oem
Reverse imports:	BGData, bigstep, CollapsABEL, cooccurNet, geneSLOPE, kangar00, mbest, misclassGLM, multiplyr, Rdsm, s2dverification, slimrec, startR
Reverse linking to:	bigalgebra, biganalytics, bigFastlm, bigKRLS, biglasso, bigtabulate, oem, sgd
Reverse suggests:	bio3d, filematrix, matpow, mlDNA, nat.nblast, NetRep, NMF, PopGenome, rsgcc, sgd

Sparse linear models: biglasso



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

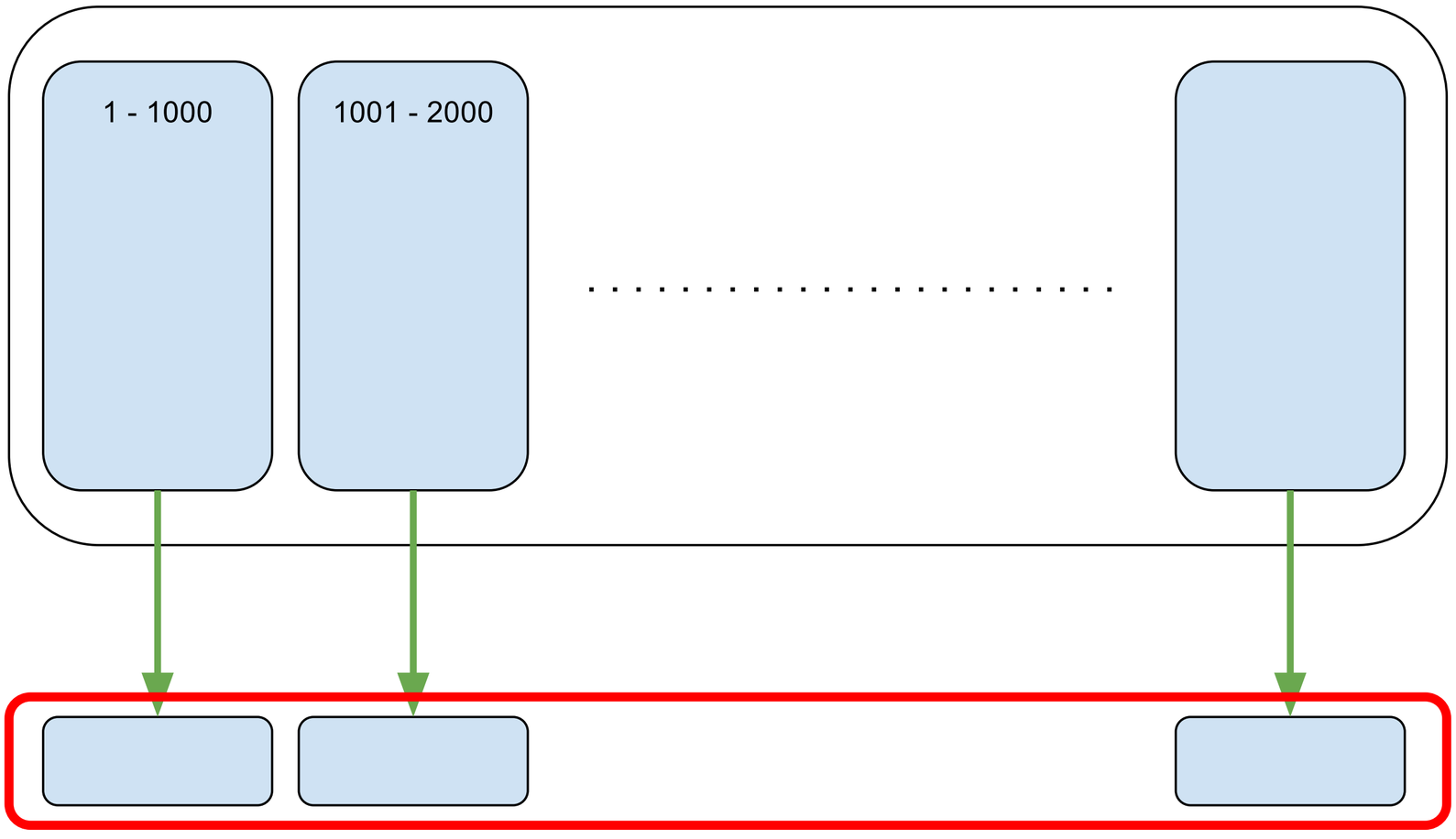
Partial Singular Value Decomposition



based on R package **RSpectra**

Split-(par)Apply-Combine Strategy

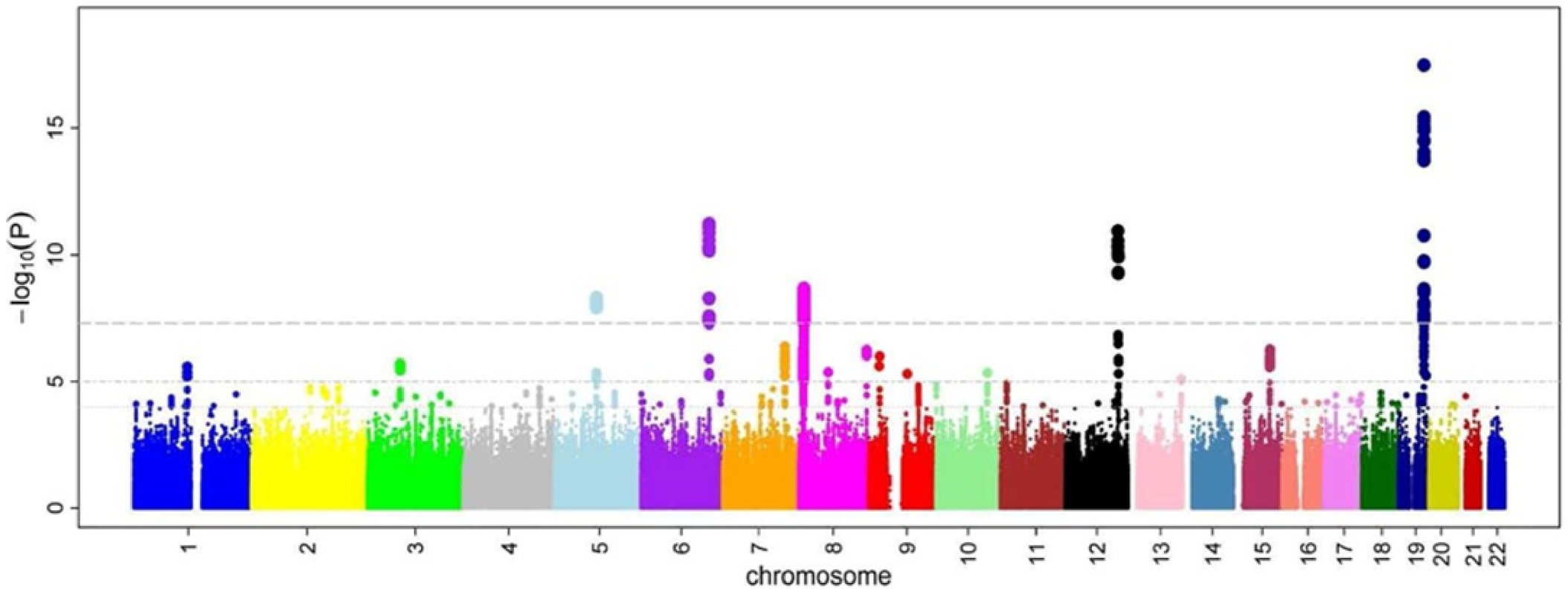
Apply standard R functions to big matrices (in parallel)



strategy coined by Hadley Wickham (2011)

Test association of each column with an outcome

In genetics, this is called a Genome-Wide Association Study (GWAS)



Manhattan plot

R Packages

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

I'm now able

to run algorithms

on hundreds of

Gigabytes of data

Any contributor is 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**.