Managing very large-scale testing procedures with R

VJ Carey

DSC 2014, Bressanone

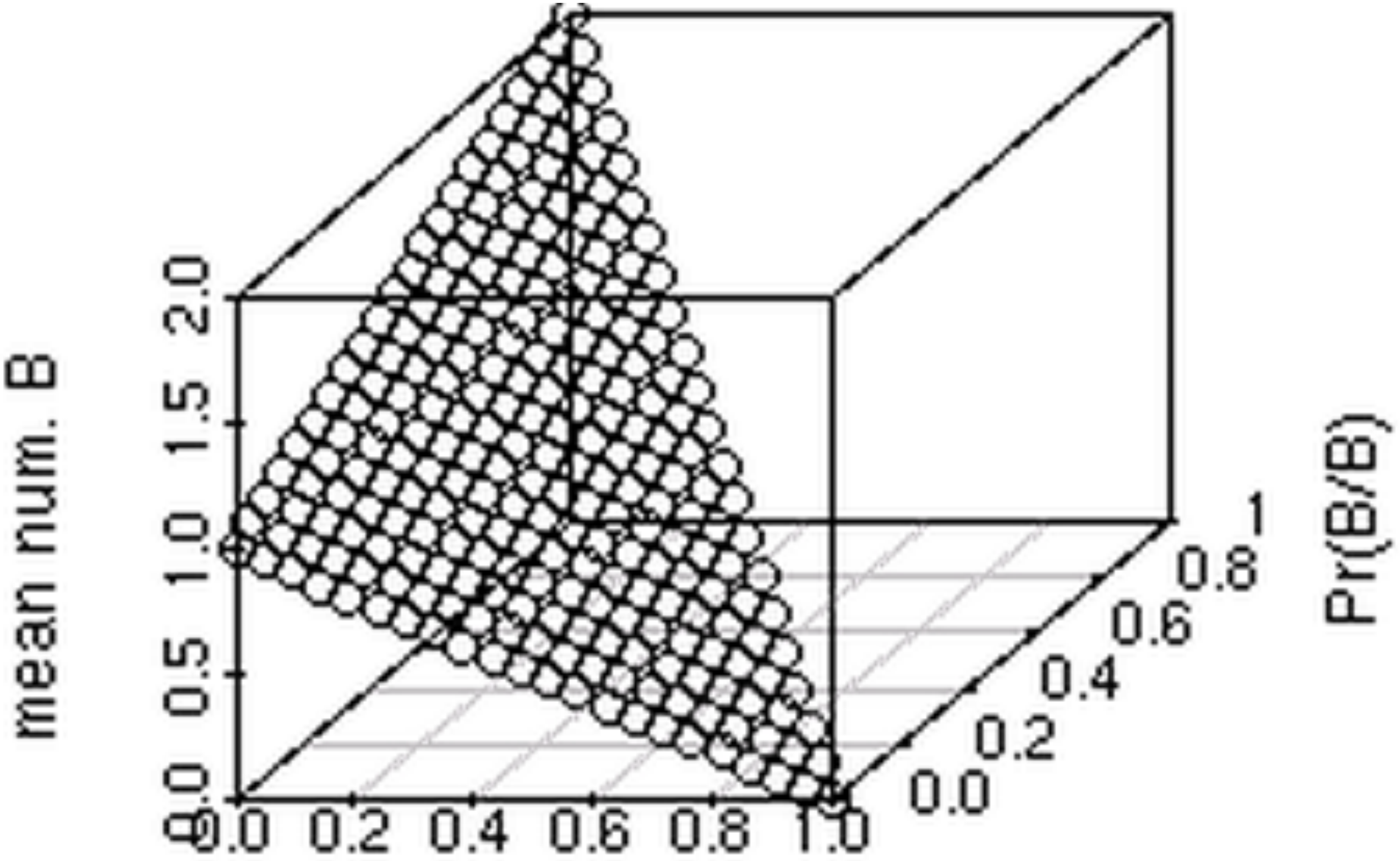
Task: genetics of gene expression

- 10^6 features x 10^9 variants
- Assay technologies allow consideration of associations that are
  - Tissue-specific
  - Condition-specific
- Slightly different from familiar “big data” problem: problem is not ingestion, but egestion and archiving for further use

Interactive statistical analysis very relevant

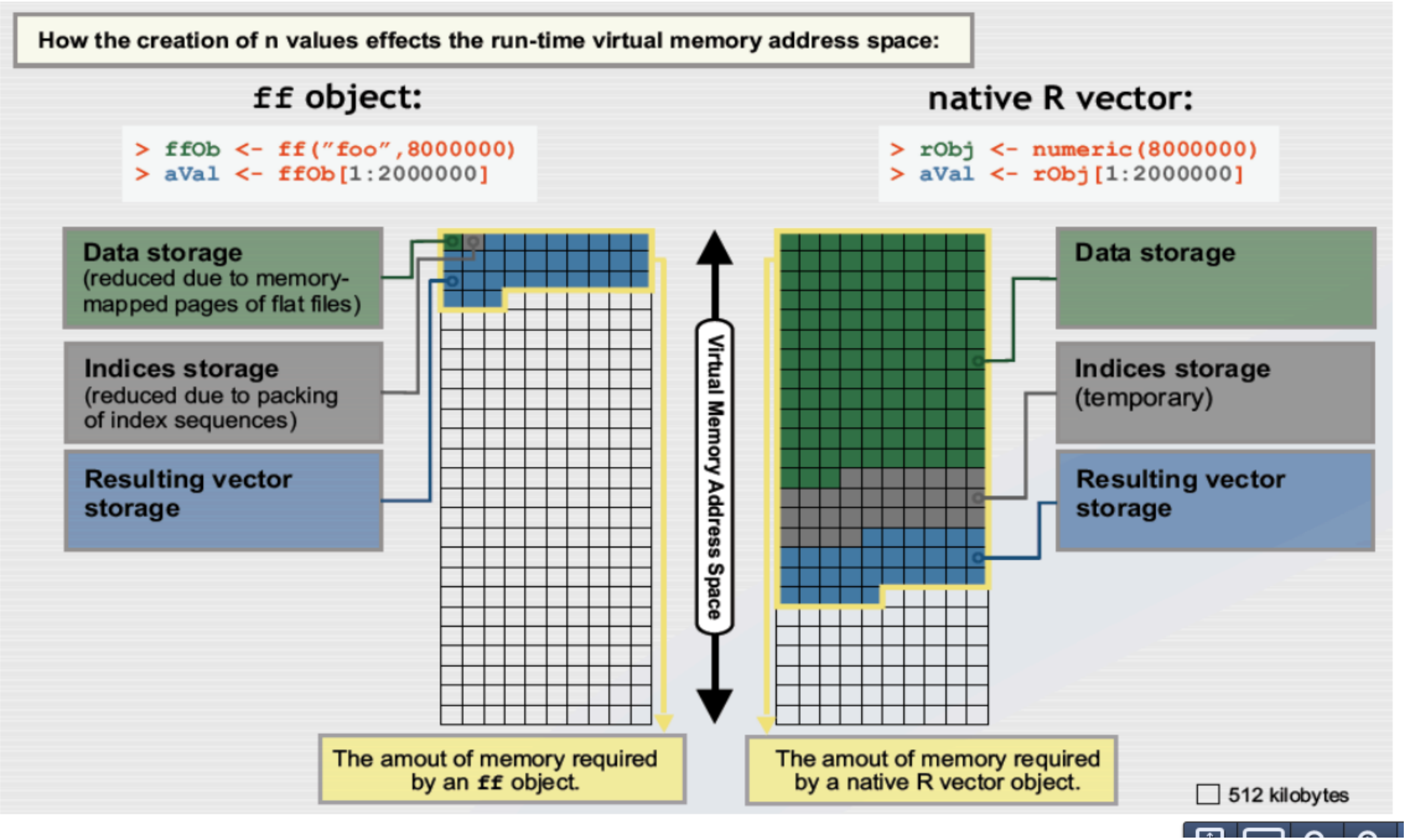
- QC, sanity checks
- Model criticism and elaboration
- Want good performance at
  - Storage/access to/modeling of voluminous assay data
  - Retrieval/updating of specific results

Data on IMPUTED genetic variants (‘reals’ in [0,2]) can be compressed: David Clayton



snpStats: also includes implementation of glm that uses this representation.

Segments of genome x transcriptome are tested and results are stored to ff as scaled short ints



How the creation of n values effects the run-time virtual memory address space:

ff object:

```
> ffOb <- ff("foo",8000000)
> aVal <- ffOb[1:2000000]
```

native R vector:

```
> rObj <- numeric(8000000)
> aVal <- rObj[1:2000000]
```

Data storage (reduced due to memory-mapped pages of flat files)

Indices storage (reduced due to packing of index sequences)

Resulting vector storage

Virtual Memory Address Space

Data storage

Indices storage (temporary)

Resulting vector storage

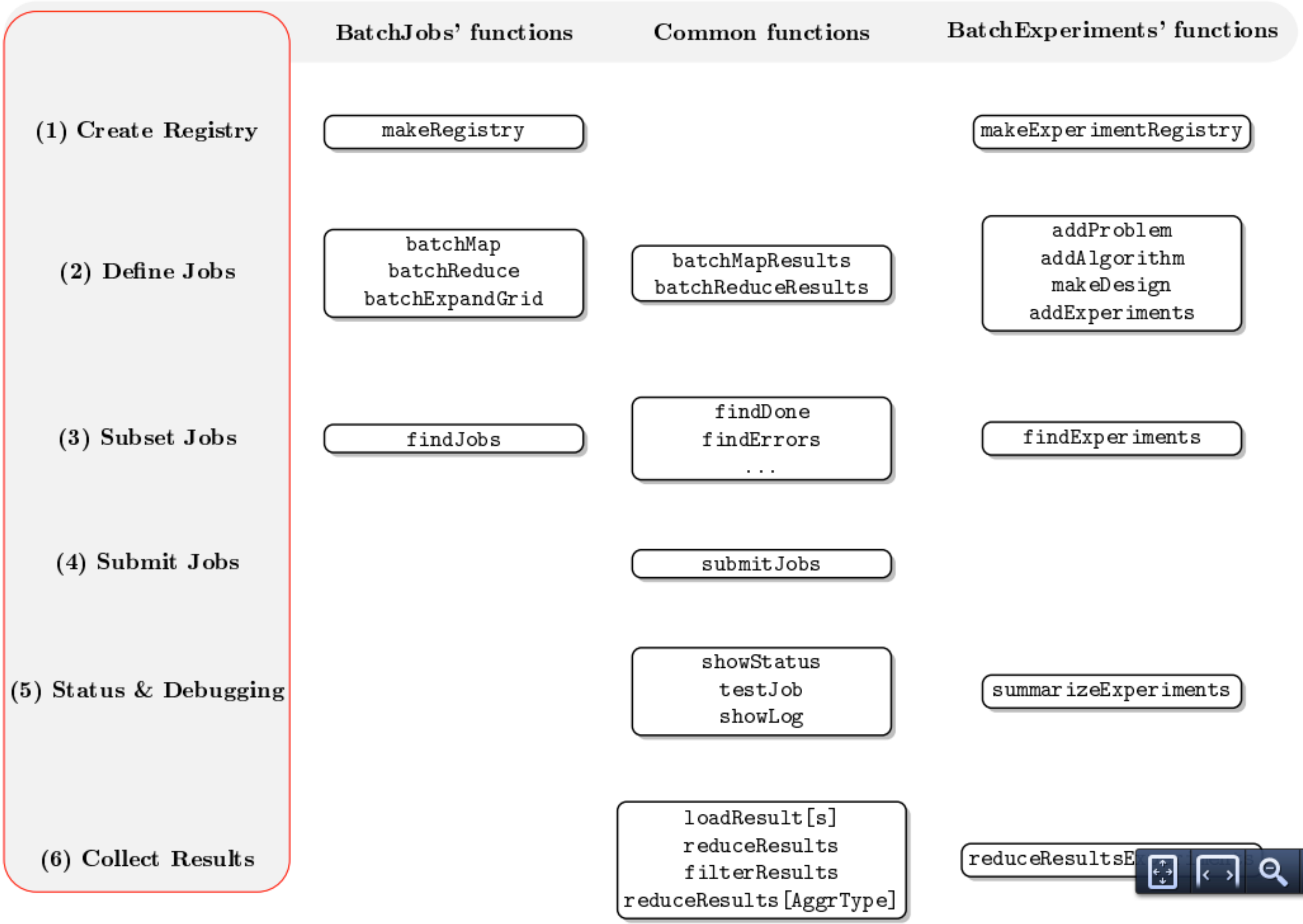
The amout of memory required by an ff object.

The amout of memory required by a native R vector object.

□ 512 kilobytes

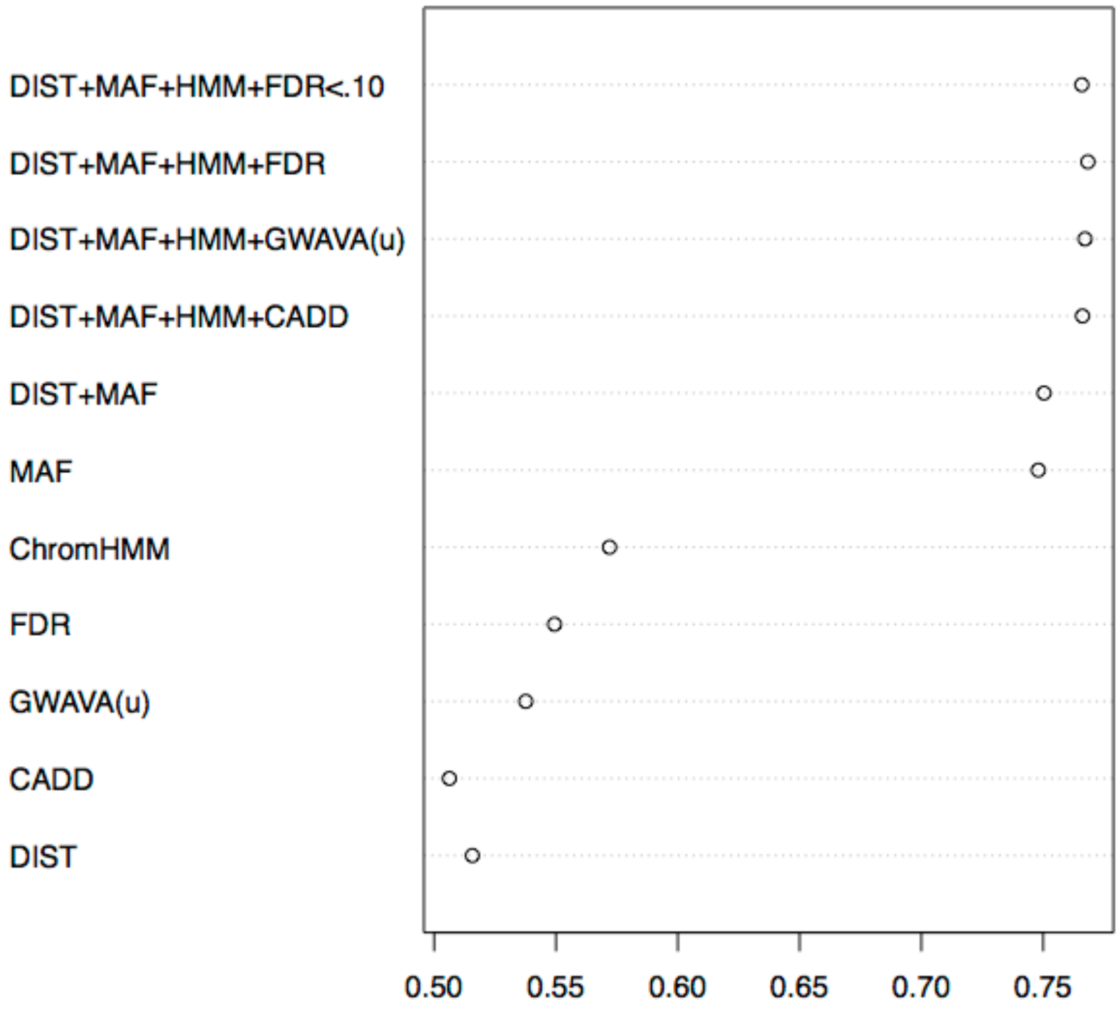
Deployment on generic cluster of multicore machines

```
library(BatchJobs)
csplreg6 =
  makeRegistry(id="mar3",
    seed=123, file.dir="mar3f")
batchMap(csplreg6, doCisChunk,
  1:length(configList) )
ids = getJobIds(csplreg6)
submitJobs(csplreg6, ids)
```



	BatchJobs' functions	Common functions	BatchExperiments' functions
(1) Create Registry	makeRegistry		makeExperimentRegistry
(2) Define Jobs	batchMap batchReduce batchExpandGrid	batchMapResults batchReduceResults	addProblem addAlgorithm makeDesign addExperiments
(3) Subset Jobs	findJobs	findDone findErrors ...	findExperiments
(4) Submit Jobs		submitJobs	
(5) Status & Debugging		showStatus testJob showLog	summarizeExperiments
(6) Collect Results		loadResult[s] reduceResults filterResults reduceResults[AggrType]	reduceResultsExperiments

Estimation of SNP tendency to be associated with trait variation: bigglms on data.table of 2-20 million records – ROC AUCs, apply over a list of formulas



Upshots

- Aims: achieve feasibility, limit use of resources, facilitate model comparison
- Data acquisition, statistical aggregates, results archiving “transparently” chunked and performed asynchronously
- Constraints: didn’t want/need standard data representations (doubles, .Rdata)

Queries

- “External memory algorithms” seem worthwhile even in the presence of huge quantities of RAM
  - a natural aspect of R software design? A prominent documentation/training objective?
- “Triply agnostic” modeling deployments:
  - Data origins (internal vs. external)
  - Data format (assumed vs. improvised/template)
  - Execution plan (selectable parallelism)