Fast automatic indexing with data.table, for beginners

Matt Dowle

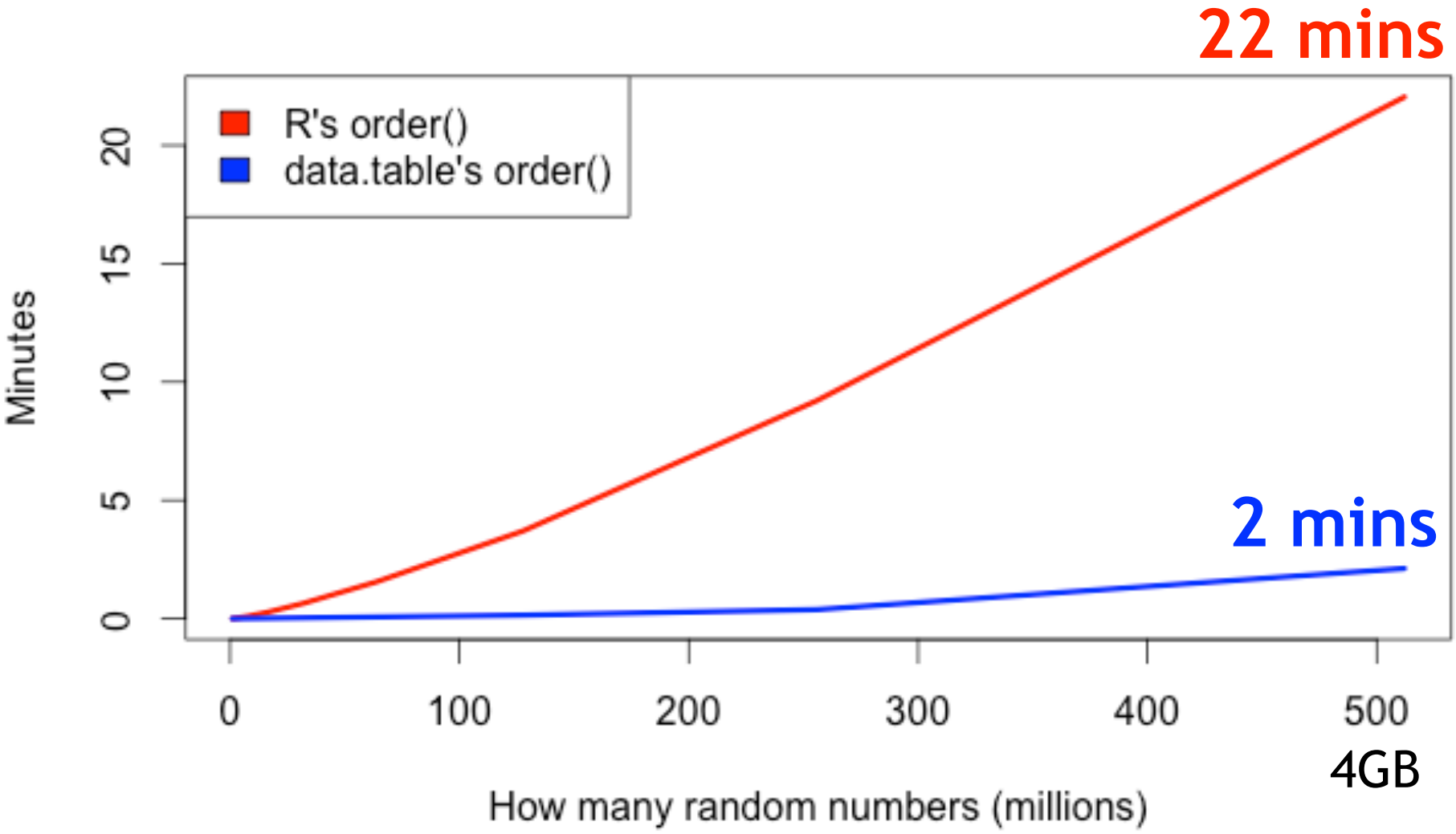
H_2O.ai
Machine Intelligence

Mainly via demo in RStudio. See video:

http://livestream.com/accounts/10932136/events/3958345

and R script :

https://github.com/h2oai/h2o-meetups/tree/master/2015_04_09_data.table_indexes



MacBook Pro 2.8GHz Intel Core i7 16GB

R 3.1.3 data.table 1.9.4

What is a data.table index?

- The ordering vector. That's it.
- The key column names and their order in the key determines the ordering vector.
- Either automatic or call `set2key(DT,col1,col2,...)`
- NB: `setkey()` reorders the data so an index vector isn't needed: primary key.

Pros

- Index storage is small and fixed: nrow * 4|8 bytes
- No collisions in hash table (no hash table)
- Building new indexes may be able to reuse existing indexes
- Rolling joins and overlapping range joins

Cons

- Insert and delete of rows requires memmove
- Binary search vs direct hash table lookup (note though collisions)

+ your thoughts very welcome.

References

- Terdiman, 2000

http://codercorner.com/RadixSortRevisited.htm

- Herf, 2001

http://stereopsis.com/radix.html

- Arun Srinivasan implemented forder() in data.table entirely in C for integer, character and double
- Matt Dowle changed from LSD (backwards) to MSD (forwards)