

R and Big Data: Some Comments

Dirk Eddelbuettel

`dirk.eddelbuettel@R-Project.org`

Big Data Summit 2013

Research Park

University of Illinois at Urbana-Champaign

December 6, 2013

Outline

- 1 Big Data
- 2 R
- 3 Rcpp

Hype or Hope?

Big Data is the New New Thing

Twitter / BigDataBorat: 11 x

Twitter, Inc. [US] <https://twitter.com/BigDataBorat/status/21185>

Home @ # Person Search Search Mail Settings Compose

 **Big Data Borat**
@BigDataBorat Follow

I hear "Journal of Applied Statistics" will change name to "Journal of Applied #bigdata", list on NYSE for \$10B

[Reply](#) [Retweet](#) [Favorite](#) [More](#)

24 RETWEETS **3** FAVORITES

11:05 AM - 10 Jun 12

Hype or Hope?

Big Data is the New New Thing

Context:

- Ability to generate data grows at an ever faster rate
- Cost of storage and processing keeps decreasing
- Some highly successful business models and insights
- Leads to various expectations and promises
- http://en.wikipedia.org/wiki/Big_data

Outline

- 1 Big Data
- 2 R
- 3 Rcpp

About R

- “Dialect” of the S Language out of Bell Labs, home of C, C++, (large parts of) Unix
- “Designed by Statisticians”
- Its mantra is “Programming with Data”
- Designed in the 1970s, became feasible on 1980s workstations, continued growth in 1990s (as well as birth of R) – ready for wider adoption in last ten years
- *Lingua Franca* of statistical research, with unparalleled breadth: 5000+ CRAN packages
- Design model: Single-threaded, data in memory

pbdR Site

Currently most promising 'big data with R' initiative

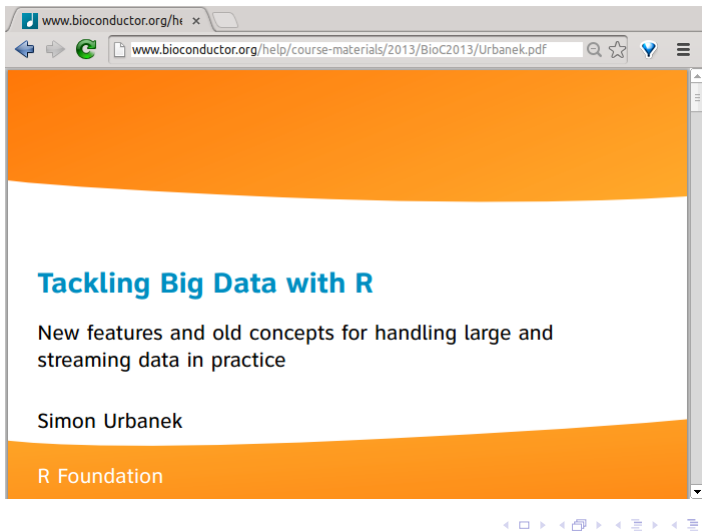


The screenshot shows a web browser window with the address bar displaying "www.r-pbd.org". The page content features the following elements:

- Header:** "pbdR: programming with big data in R" in large, multi-colored text. Below it, the tagline "Simplifying Scalability" is displayed.
- Navigation:** A horizontal menu with buttons for "About", "Packages", "Download", "Install", "Tutorials", "FAQ", and "Highlight".
- Section:** A section titled "Overview" with a dotted line separator.
- Text:** A paragraph describing the project: "The 'Programming with Big Data in R' project (pbdR) enables high-level distributed data parallelism in R, so that it can easily utilize large HPC platforms with thousands of cores, making the R language scale to unparalleled heights."

RCloud etc

Simon Urbanek et al at AT&T



The image shows a web browser window displaying a slide from a presentation. The browser's address bar shows the URL `www.bioconductor.org/help/course-materials/2013/BioC2013/Urbanek.pdf`. The slide content is as follows:

Tackling Big Data with R

New features and old concepts for handling large and streaming data in practice

Simon Urbanek

R Foundation

Navigation icons for the presentation are visible at the bottom right of the browser window.

Bigmemory

Kane and Emerson: Using external pointer interface

The screenshot shows a web browser window with the address bar at `www.bigmemory.org`. The page title is "BIGMEMORY.ORG". The main heading is "The bigmemory Project" with a search bar below it. A left sidebar contains a navigation menu with categories: "bigmemory" (expanded), "Research", and "Sitemap". The "bigmemory" category lists sub-items: "bigalgebra", "biganalytics", "bigmemory", "bigtabulate", and "synchronicity". The "Research" category lists "Data Expo" and "Documentation". The "Sitemap" category is also listed. The main content area features the heading "bigmemory" and a paragraph describing the project's purpose: extending the R statistical programming environment for massive matrices. It lists the packages *biganalytics*, *bigtabulate*, *synchronicity*, and *bigalgebra*. A "NEWS (July 2011)" section mentions version 4.2.11 and its availability on CRAN. At the bottom right of the page, there are navigation icons for back, forward, and search.

BIGMEMORY.ORG

www.bigmemory.org

The bigmemory Project

Search this site

- bigmemory
 - bigalgebra
 - biganalytics
 - bigmemory
 - bigtabulate
 - synchronicity
- Research
 - Data Expo
 - Documentation
- Sitemap

bigmemory

This project extends the *R* statistical programming environment. Package *bigmemory* supports the creation, storage, access, and manipulation of massive matrices. These matrices are allocated to shared memory and may use memory-mapped files. Packages *biganalytics*, *bigtabulate*, *synchronicity*, and *bigalgebra* (please see [the bigalgebra page](#) for 32-bit/64-bit library information) provide advanced functionality. We provide a short overview with examples in the [Documentation](#) area.

NEWS (July 2011). Version 4.2.11 has fixed a few minor problems and has been unloaded to CRAN. We note some problem with newer macos

Links

Outline

- 1 Big Data
- 2 R
- 3 Rcpp

Simple to use

Via `evalCpp()`, `cppFunction()`, and `sourceCpp()`

```
## evaluate a C++ expression, retrieve result
evalCpp("2 + 2")

## [1] 4

## a little fancier
evalCpp("std::numeric_limits<double>::max()")

## [1] 1.798e+308

## create ad-hoc R functions 'accu' using STL
cppFunction('double accu(NumericVector x) {
  return(std::accumulate(x.begin(), x.end(), 0.0));
}')
accu(1:100)

## [1] 5050
```

70+ fully documented examples

Open for contributions

Rcpp Gallery - Google Chrome

Rcpp Gallery x

gallery.rcpp.org

Rcpp Projects Gallery Book Events More -

Featured Articles

[Quick conversion of a list of lists into a data frame](#) — John Merrill
This post shows one method for creating a data frame quickly

[Passing user-supplied C++ functions](#) — Dirk Eddelbuettel
This example shows how to select user-supplied C++ functions

[Using Rcpp to access the C API of xts](#) — Dirk Eddelbuettel
This post shows how to use the exported API functions of xts

[Timing normal RNGs](#) — Dirk Eddelbuettel
This post compares drawing $N(0,1)$ vectors from R, Boost and C++11

[A first lambda function with C++11 and Rcpp](#) — Dirk Eddelbuettel
This post shows how to play with lambda functions in C++11

[First steps in using C++11 with Rcpp](#) — Dirk Eddelbuettel
This post shows how to experiment with C++11 features

[Using Rcout for output synchronised with R](#) — Dirk Eddelbuettel
This post shows how to use Rcout (and Rcerr) for output

[Using the Rcpp sugar function clamp](#) — Dirk Eddelbuettel
This post illustrates the sugar function clamp

[Using the Rcpp Timer](#) — Dirk Eddelbuettel
This post shows how to use the Timer class in Rcpp

[Calling R Functions from C++](#) — Dirk Eddelbuettel
This post discusses calling R functions from C++

[More »](#)

Recently Published

Apr 12, 2013 » [Using the RcppArmadillo-based Implementation of R's sample\(\)](#) — Christian Gunning and Jonathan Olmsted

Apr 8, 2013 » [Dynamic Wrapping and Recursion with Rcpp](#) — Kevin Ushey

Mar 14, 2013 » [Using bigmemory with Rcpp](#) — Michael Kane

Mar 12, 2013 » [Generating a multivariate gaussian distribution using RcppArmadillo](#) — Ahmadou Dicko

Mar 1, 2013 » [Using Rcpp with Boost.Regex for regular expression](#) — Dirk Eddelbuettel

The Rcpp Book

