Reproducibility

with Revolution R Open
and the checkpoint package

David Smith

R Community Lead
Revolution Analytics, a Microsoft company
@revodavid

June 5 2015

Agenda

- What is Reproducibility?
- The checkpoint package
- Demonstration
- Revolution R Open
- Q&A

OUR COMPANY

REVOLUTION ANALYTICS

MOUNTAIN VIEW ▪ LONDON ▪ SINGAPORE

The leading provider of **advanced analytics software and services** based on open source R, since 2007

OUR PRODUCT

REVOLUTION R: The enterprise-grade predictive analytics application platform based on the R language

SOME KUDOS

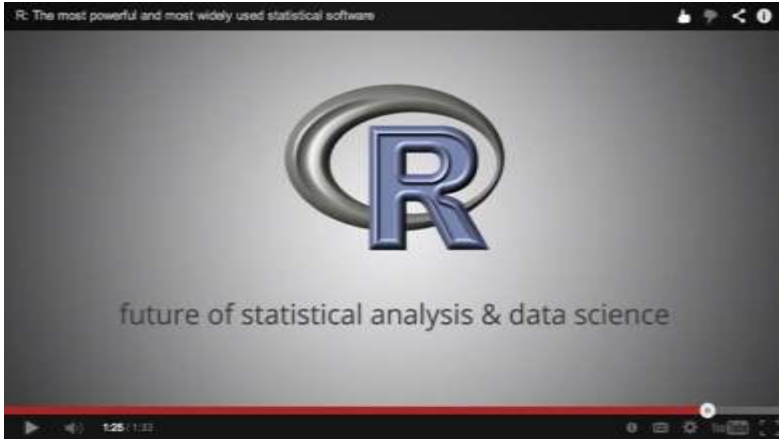
“This acquisition will help customers use advanced analytics within Microsoft data platforms“

-- Joseph Sirosh, CVP C+E

Microsoft

What is R?

- Most widely used data analysis software
  - Used by 2M+ data scientists, statisticians and analysts
- Most powerful statistical programming language
  - Flexible, extensible and comprehensive for productivity
- Create beautiful and unique data visualizations
  - As seen in New York Times, The Economist and FlowingData
- Thriving open-source community
  - Leading edge of analytics research
- Fills the talent gap
  - New graduates prefer R



R: The most powerful and most widely used statistical software

future of statistical analysis & data science

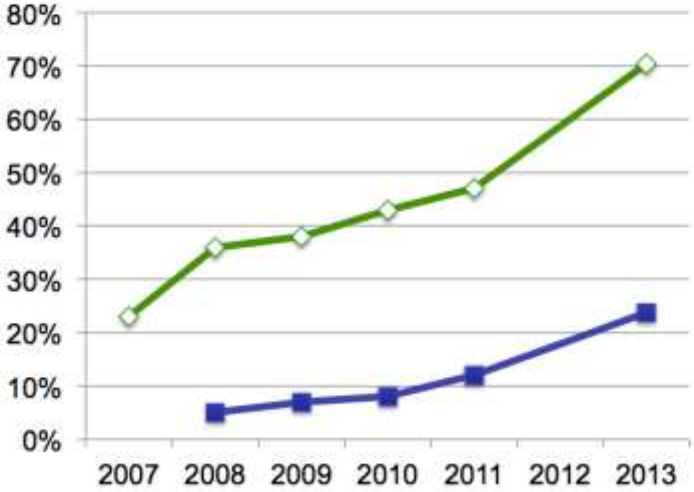
www.revolutionanalytics.com/what-is-r

R's popularity is growing rapidly

More at blog.revolutionanalytics.com/popularity

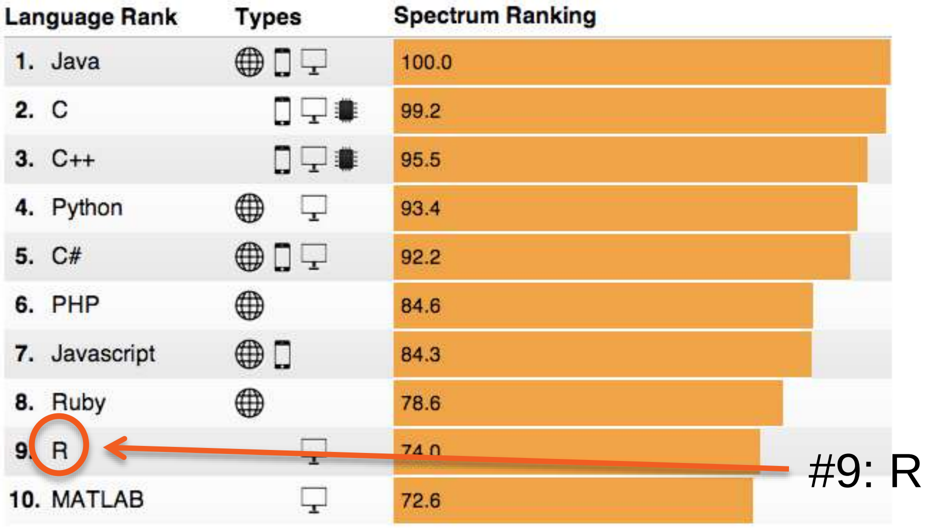
R Usage Growth

Rexer Data Miner Survey, 2007-2013



Language Popularity

IEEE Spectrum Top Programming Languages



Language Rank	Types	Spectrum Ranking
1. Java	🌐 📱 🖥	100.0
2. C	📱 🖥 ▦	99.2
3. C++	📱 🖥 ▦	95.5
4. Python	🌐 🖥	93.4
5. C#	🌐 📱 🖥	92.2
6. PHP	🌐	84.6
7. Javascript	🌐 📱	84.3
8. Ruby	🌐	78.6
9. R	🖥	74.0
10. MATLAB	🖥	72.6

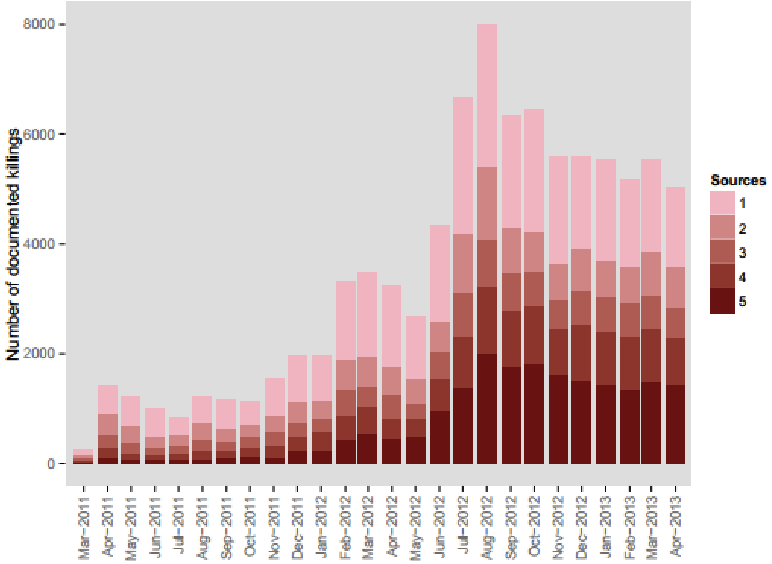
#9: R

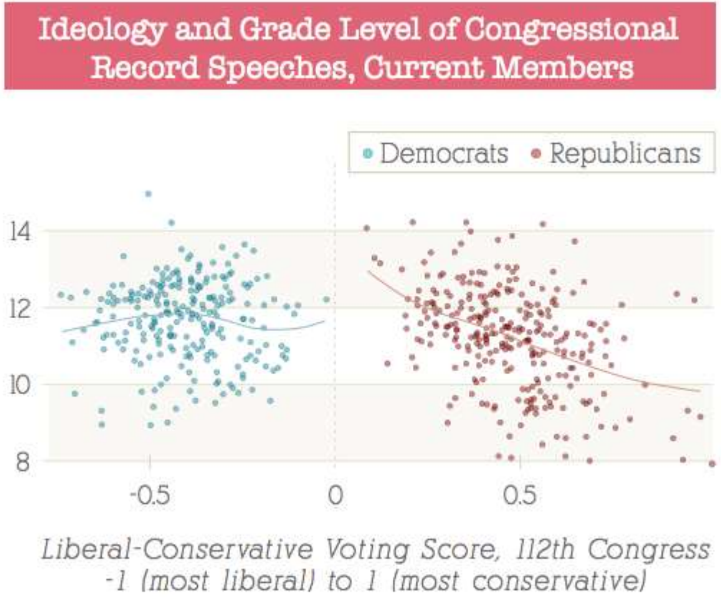
- Rexer Data Miner Survey
- IEEE Spectrum, July 2014

Application: Public Affairs

HRDAG

SUNLIGHT FOUNDATION




Ideology and Grade Level of Congressional Record Speeches, Current Members

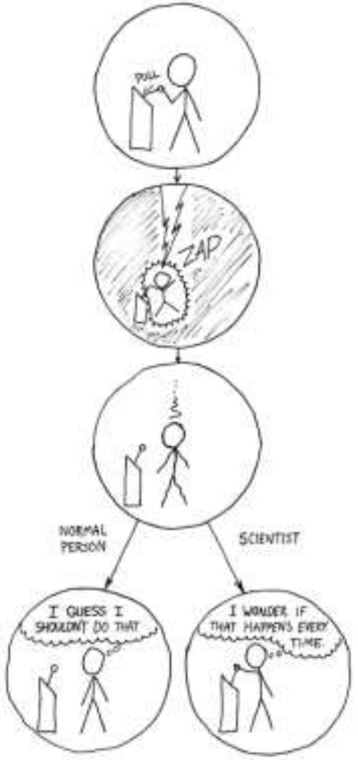
- Casualty estimation in Warzones
- Political Analysis

What is Reproducibility?

“The goal of reproducible research is to tie specific instructions to data analysis and experimental data so that scholarship can be recreated, better understood and verified.”
CRAN Task View on Reproducible Research (Kuhn)

- Method + Environment
  -> Results

- A **process** for:
  - Sharing the method
  - Describing the environment
  - Recreating the results


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME

xkcd.com/242/

Reproducibility – why do we care?

Academic / Research

- Verify results
- Advance Research

Business

- Production code
- Reliability
- Reusability
- Collaboration
- Regulation

How Bright Promise in Cancer Testing Fell Apart

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
        fuTOHormap.values <= fuCut]));
> colnames(temp) <- c("Theirs", "Ours");
> temp
     Theirs          Ours
...
[3,] "1881_at"       "1882_g_at"
[4,] "31321_at"      "31322_at"
[5,] "31725_s_at"    "31726_at"
[6,] "32307_r_at"    "32308_r_at"
...
```

Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By GINA KOLATA
Published: July 7, 2011

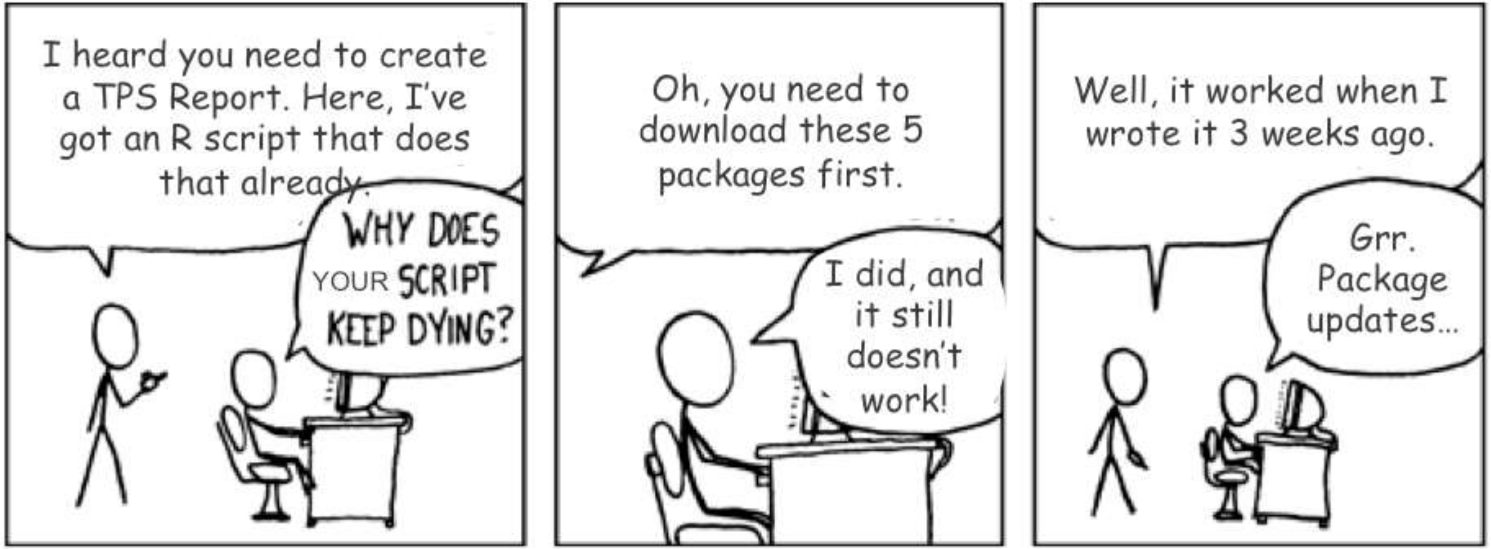
www.nytimes.com/2011/07/08/health/research/08genes.html
http://arxiv.org/pdf/1010.1092.pdf

REVOLUTION ANALYTICS

Observations

- R versions are pretty manageable
  - Major versions just once a year
  - Patches rarely introduce incompatible changes
- Good solutions for literate programming
  - Rstudio / knitr / Rmarkdown
- OS/Hardware not the major cause of problems
- The big problem is with **packages**
  - CRAN is in a state of continual flux

An R Reproducibility Problem


I heard you need to create a TPS Report. Here, I've got an R script that does that already.
WHY DOES YOUR SCRIPT KEEP DYING?
Oh, you need to download these 5 packages first.
I did, and it still doesn't work!
Well, it worked when I wrote it 3 weeks ago.
Grr. Package updates...

Adapted from http://xkcd.com/234/ CC BY-NC 2.5

Reproducible R Toolkit

`projects.revolutionanalytics.com/rrt/`

- Static CRAN mirror in Revolution R Open
  - CRAN packages fixed with each RRO update
- Daily CRAN snapshots
  - Storing every package version since September 2014
  - Hosted at mran.revolutionanalytics.com/snapshot
- Write and share scripts synced to a specific snapshot date
  - **checkpoint** package installed with RRO
  - Also available on CRAN

Using checkpoint

- Add 2 lines to the top of your script

```
library(checkpoint)
checkpoint("2015-01-28")
```

Or, whichever date you want

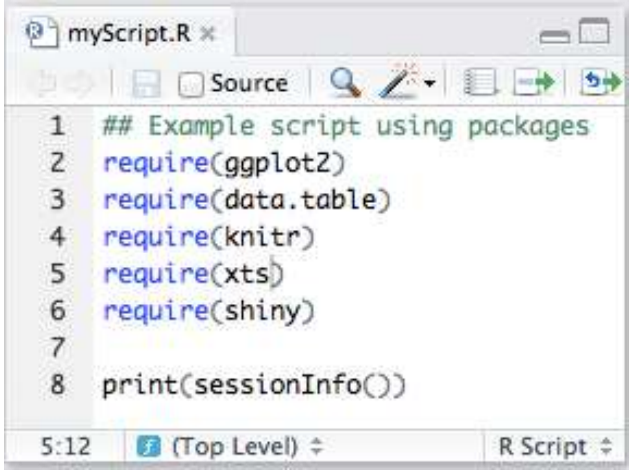
- Err, that's it.

- Optionally, check the R version as well

```
library(checkpoint)
checkpoint("2015-01-28", R.version="3.1.3")
```

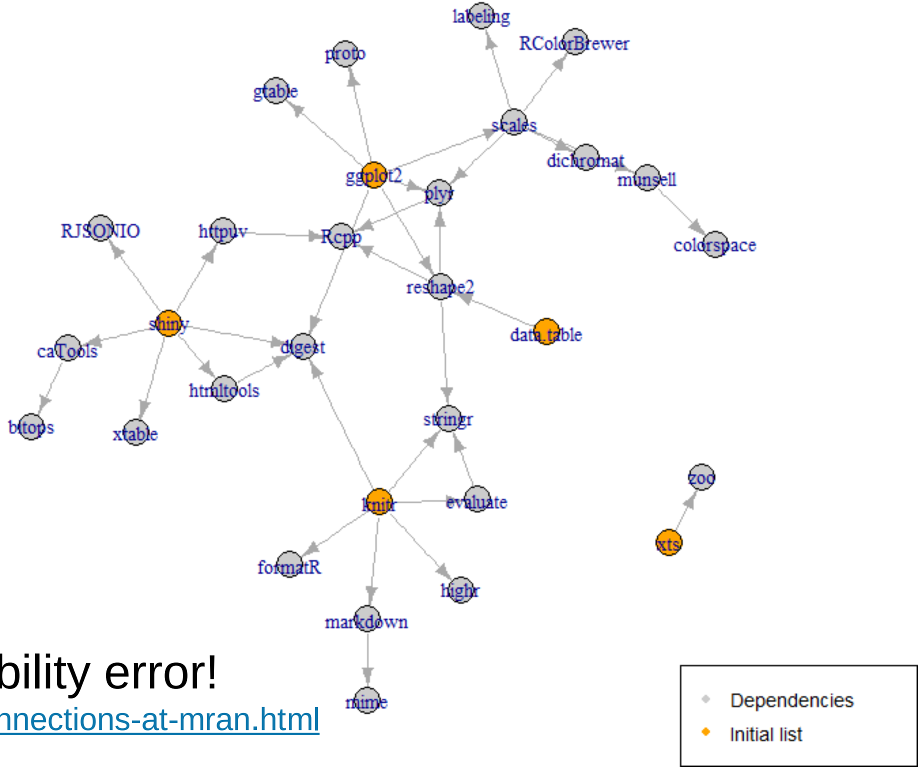
Package dependency explosion

- R script file using 6 most popular packages



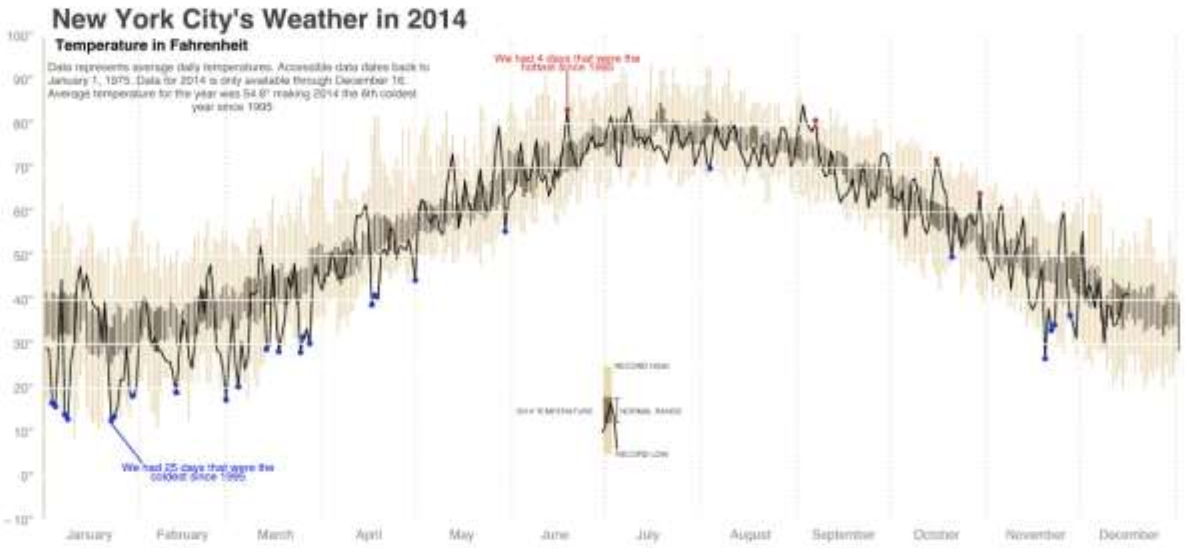
```
## Example script using packages
require(ggplot2)
require(data.table)
require(knitr)
require(xts)
require(shiny)

print(sessionInfo())
```

Package dependency graph



Any updated package = potential reproducibility error!

http://blog.revolutionanalytics.com/2014/10/explore-r-package-connections-at-mran.html



New York City's Weather in 2014

Temperature in Fahrenheit

Data represents average daily temperatures. Accessible data dates back to January 1, 1975. Data for 2014 is only available through December 16. Average temperature for the year was 54.8° making 2014 the 6th coldest year since 1995

We had 4 days that were the hottest since 1995

We had 25 days that were the coldest since 1995

Demo

Weather Map

Checkpoint tips for script authors

- Work within a **project**
  - Dedicated folder with scripts, data and output
  - eg `/Users/david/R/weather`
- Create a master .R script file beginning with

```
library(checkpoint)
checkpoint("DATE")
```

  - package versions used will be as of this date
- Don’t use `install.packages` directly
  - Use `library()` and checkpoint does the rest
  - You **can** have different package versions installed for different projects at the same time!

Sharing projects with checkpoint

- Just share your script or project folder!
- Recipient only needs:
  - compatible R version
  - checkpoint package (installed with RRO)
  - Internet connection to MRAN (at least first time)
- Checkpoint takes care of:
  - Installing CRAN packages
    - Binaries (ease of installation)
    - Correct versions (reproducibility)
    - Dependencies (ease of installation)
  - Eliminating conflicts with other installed packages

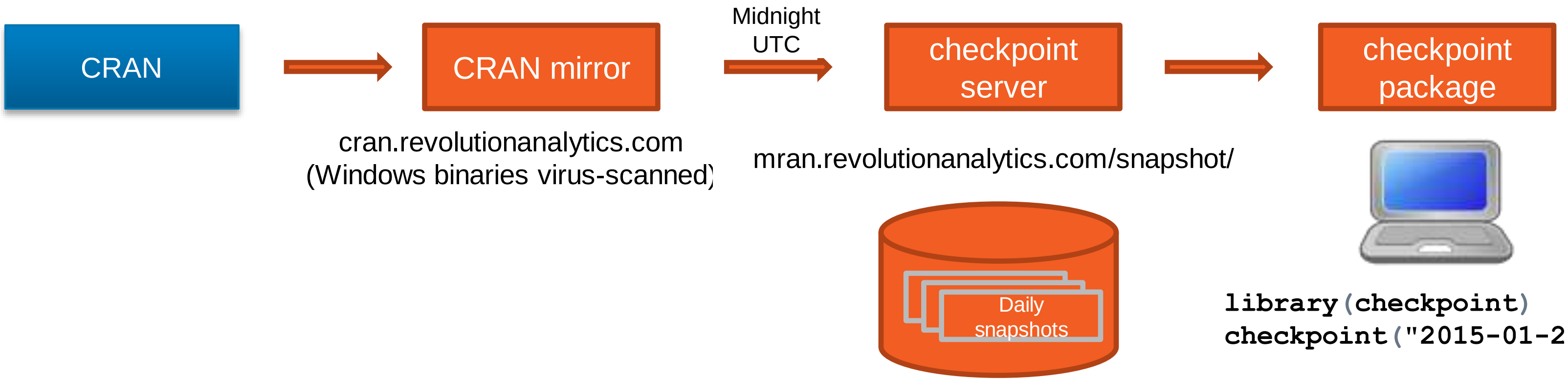
The checkpoint magic

The `checkpoint()` call does all this:

- Scans project for required packages
- Installs required packages and dependencies
  - Packages installed specific to project
  - Versions specific to checkpoint date
    - Installed in `~/.checkpoint/`DATE
    - Skips packages if already installed (2nd run through)
- Reconfigures package search path
  - Points only to project-specific library

MRAN checkpoint server

Checkpoint uses MRAN’s downstream CRAN mirror with daily snapshots.



CRAN

CRAN mirror

cran.revolutionanalytics.com
(Windows binaries virus-scanned)

Midnight
UTC

checkpoint
server

mran.revolutionanalytics.com/snapshot/

Daily
snapshots

checkpoint
package

```
library(checkpoint)
checkpoint("2015-01-28")
```

checkpoint server - implementation

Checkpoint uses MRAN’s downstream CRAN mirror with daily snapshots.

- rsync to mirror CRAN daily
  - Only downloads changed packages
- zfs to store incremental snapshots
  - Storage only required for new packages
- Organizes snapshots into a labelled hierarchy
  - `mran.revolutionanalytics.com/snapshot/`YYYY-MM-DD
- MRAN hosted by high-performance cloud provider
  - Provisioned for availability and latency

`https://github.com/RevolutionAnalytics/checkpoint-server`

Using non-CRAN packages Reproducibly

- Today, checkpoint only manages packages from CRAN

- **GitHub**: use install_github with a specific checkin hash

```
install_github("ramnathv/rblocks",
ref="a85e748390c17c752cc0ba961120d1e784fb1956")
```

- **BioConductor**: use packages from a specific BioConductor release
  - Not as easy as it seems!

- **Private packages / behind the firewall**
  - use miniCRAN to create a local, static repository

Comparison with packrat

rstudio.github.io/packrat/

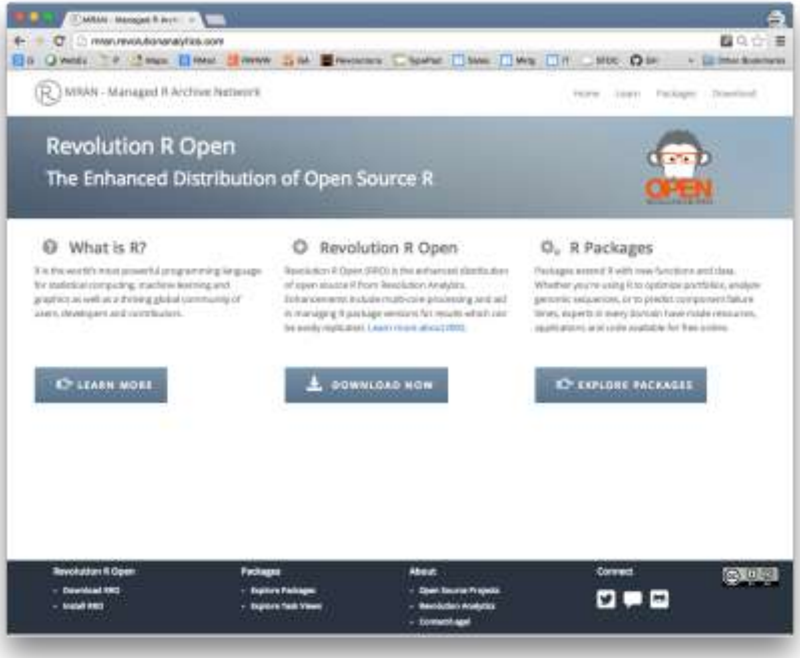
- Packrat is flexible and powerful
  - Supports non-CRAN packages (e.g. github)
  - Allows mix-and-matching package versions
  - Requires shipping all package source
  - Requires recipients to build packages from source
- Checkpoint is simple
  - Reproducibility from one script
  - Simple for recipients to reproduce results
  - Only allows use of CRAN packages versions that have been tested together
  - Requires Web access (and availability of MRAN)

Revolution R Open includes checkpoint

- Enhanced Open Source R distribution
- Compatible with all R-related software
- Multi-threaded for performance
- Focus on reproducibility
- Open source (GPLv2 license)
- Available for Windows, Mac OS X, Ubuntu, Red Hat and OpenSUSE
- Free download at `mran.revolutionanalytics.com`

MRAN

The Managed R Archive Network


mran.revolutionanalytics.com
MRAN - Managed R Archive Network
Home
Learn
Packages
Download

Revolution R Open

The Enhanced Distribution of Open Source R

What is R?

R is the world's most powerful programming language for statistical computing, machine learning and graphics as well as a thriving global community of users, developers and contributors.

Revolution R Open

Revolution R Open (RRO) is the enhanced distribution of open source R from Revolution Analytics. Enhancements include multi-core processing and aid in managing R package versions for results which can be easily replicated. Learn more about RRO.

R Packages

Packages extend R with new functions and data. Whether you're using R to optimize portfolios, analyze genomic sequences, or to predict component failure times, experts in every domain have made resources, applications and code available for free online.

mran.revolutionanalytics.com

- Download Revolution R Open
- Learn about R and RRO
- Explore R Packages
- Explore Task Views
- R tips and applications
- Daily CRAN snapshots

RRO: 100% Compatibility

- Built on latest R engine
  - Currently R 3.2.0

- Drop-in replacement for R
- 100% compatible with
  - R scripts
  - R packages
  - Applications with R connections

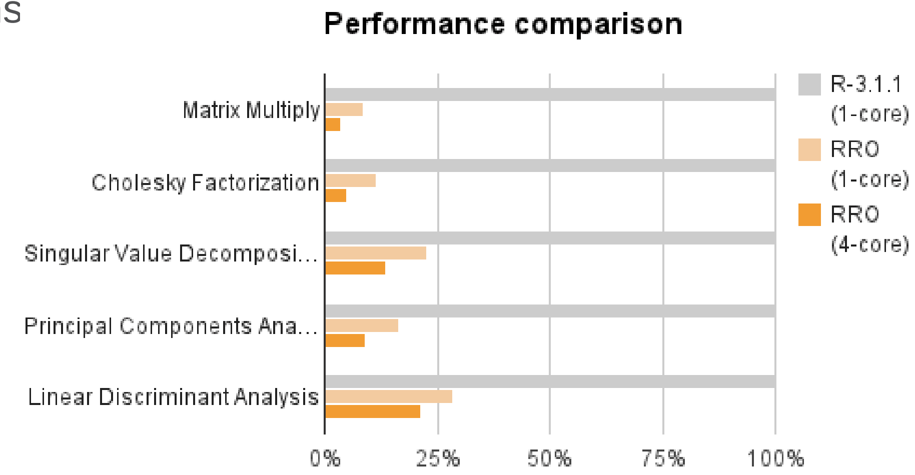
- Designed to work with RStudio
  - No configuration required

CRAN mirrors and RRO

- Revolution R Open ships with a fixed default CRAN mirror
  - Currently: 1 May 2015 (v 3.2.0)
  - Soon: 1 July 2015 (v 3.2.1)
  - (RRO updates released within 3 weeks of CRAN R)
- All users of same RRO version get same CRAN package versions by default
  - regardless when “install.packages” is run
- Use checkpoint to access newer package versions

Multi-threaded performance

- Intel MKL replaces standard BLAS/LAPACK algorithms (Windows/Linux)
- Pipelined operations
  - Optimized for Intel, works for all archs
- High-performance algorithms
- Sequential ➔ Parallel
  - Uses as many threads as there are available cores
  - Control with: `setMKLthreads(<value>)`
- No need to change any R code
- Included in RRO binary distribution



More at Revolutions blog

Why use checkpoint?

- Write and share code R whose results can be reproduced, even if new (and possibly incompatible) package versions are released later.
- Share R scripts with others that will automatically install the appropriate package versions (no need to manually install CRAN packages).
- Write R scripts that use older versions of packages, or packages that are no longer available on CRAN.
- Install packages (or package versions) visible only to a specific project, without affecting other R projects or R users on the same system.
- Manage multiple projects that use different package versions.

Thank you

Contribute:

github.com/RevolutionAnalytics/checkpoint

Download Revolution R Open

mran.revolutionanalytics.com/download

www.revolutionanalytics.com
Twitter: @RevolutionR

David Smith
R Community Lead
Revolution Analytics
`@revodavid`
`davidsmi@microsoft.com`