REVOLUTION ANALYTICS

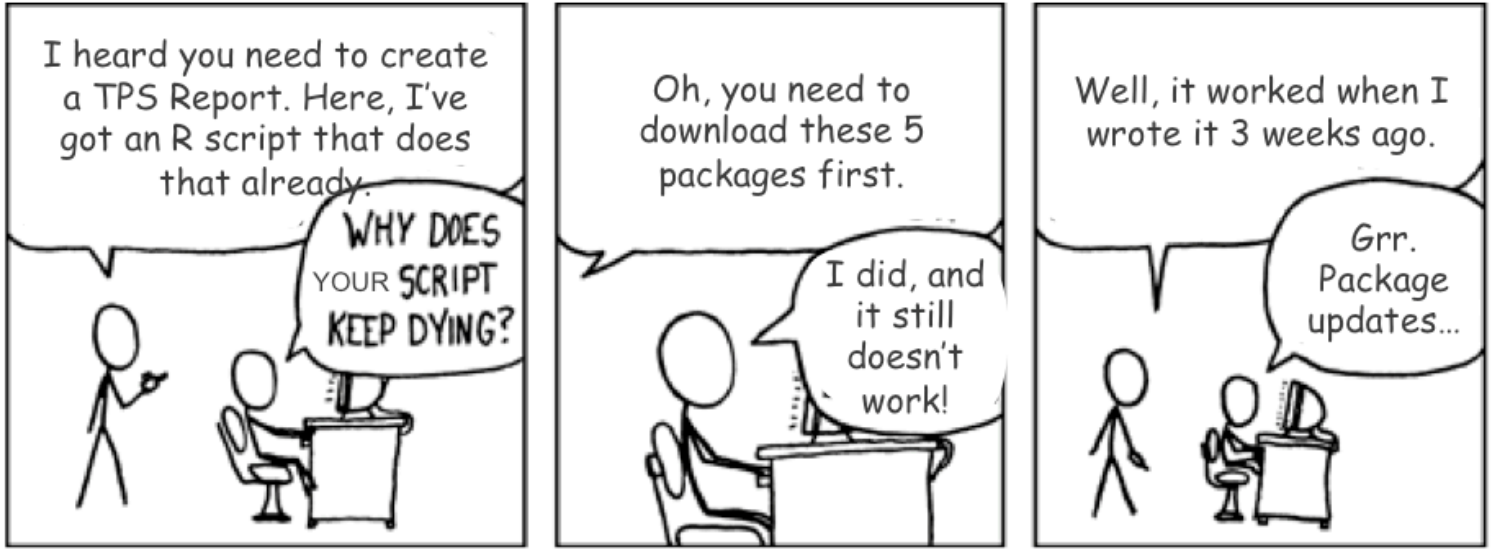
Simple Reproducibility
with the checkpoint package

David Smith

R Community Lead
Revolution Analytics, a Microsoft company
@revodavid

useR! 2015, July 1 2015

An R Reproducibility Problem



Package dependency explosion

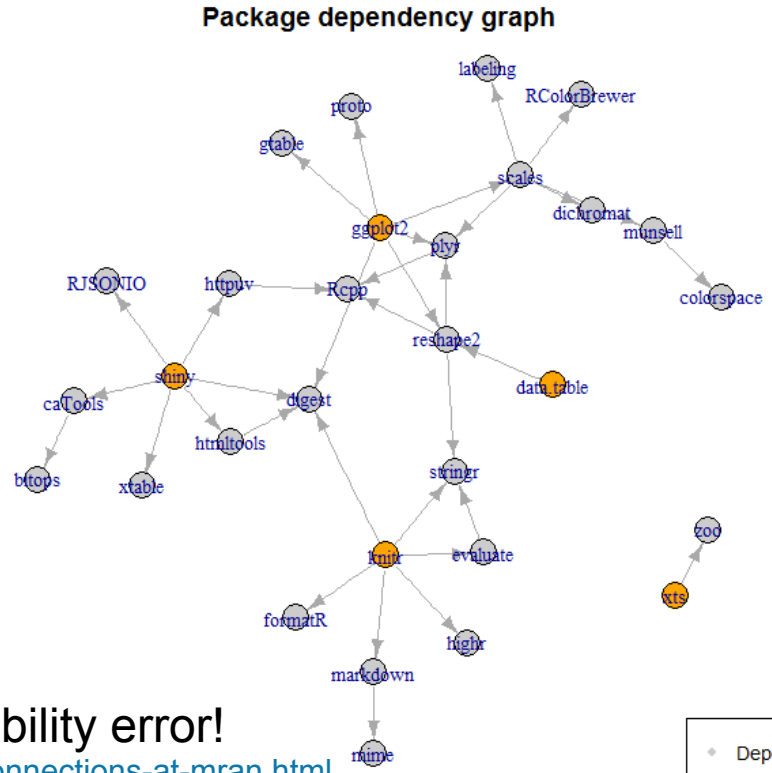
- R script file using 6 most popular packages

myScript.R

Source

```
1  ## Example script using packages
2  require(ggplot2)
3  require(data.table)
4  require(knitr)
5  require(xts)
6  require(shiny)
7
8  print(sessionInfo())
```

5:12 (Top Level) R Script



Any updated package = potential reproducibility error!

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

REVOLUTION ANALYTICS

Using the checkpoint package

- Install checkpoint package from CRAN
  - Or use Revolution R Open
- Add 2 lines to the top of your script

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

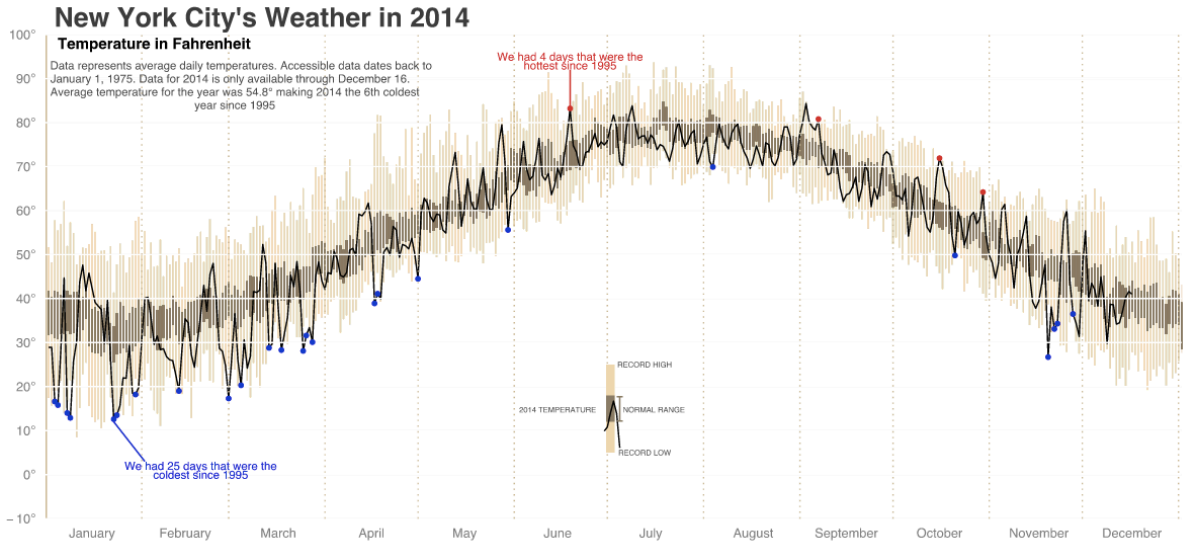
Use package versions as of this date

- Err, that's it.

- Optionally, check the R version as well

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

REVOLUTION ANALYTICS



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

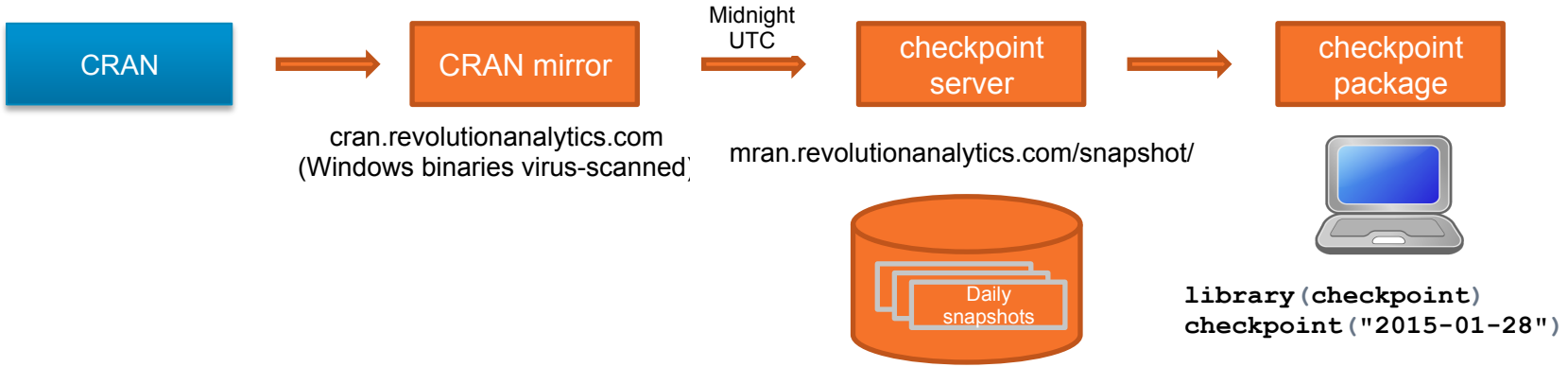
The checkpoint magic

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.



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

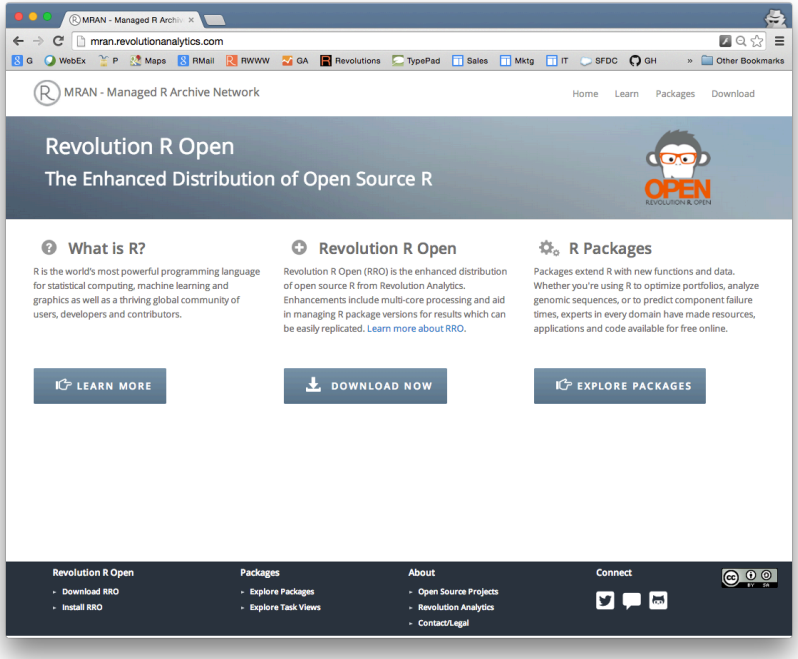
rstudio.github.io/packrat/

- 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
- 100% compatible with all R-related software
- Multi-threaded for performance
- Built-in 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



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

mran.revolutionanalytics.com

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.

REVOLUTION ANALYTICS

Thank you

Contribute:
github.com/RevolutionAnalytics/checkpoint

Download Revolution R Open
mran.revolutionanalytics.com/download

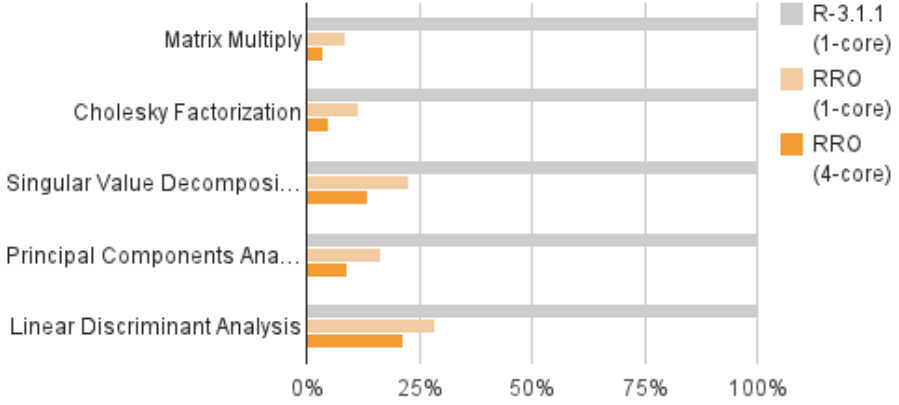
David Smith
R Community Lead
Revolution Analytics
@revodavid
davidsmi@microsoft.com

More at: blog.revolutionanalytics.com
Slides: http://www.slideshare.net/RevolutionAnalytics/checkpoint-user-2015

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

Performance comparison



More at Revolutions blog