liftr: an R Package for Persistent Reproducible Research

JSM 2018

Nan Xiao
@road2stat
The Reproducibility Crisis

- Always a concern in both academia & industry.
- R Markdown + knitr pretty much saved the day.
The New Challenge

Even higher reproducibility for statistical computing: regardless of *time* or *environment*. 
Docker to the Rescue

- Docker allows applications and their dependencies to be packaged into discrete runtime environments, called containers.

- Apps packaged in this way can run from diverse infrastructures.
Our Solution: liftr

Persistent, OS-level reproducibility for R Markdown documents.
Containerize R Markdown Documents as Easy as 1-2-3

```
lift("foo.Rmd")  render_docker("foo.Rmd")  Containerized Report
```

R Markdown Documents with liftr options in metadata

Introduce additional R Markdown metadata for containerizing reports.

```
---
liftr:
syseps:
  - gfortran
cran:
  - glmnet
  - xgboost
---
```

 Parse metadata in the document, and generate Dockerfile.

Build Docker image, run the container, and render the RMD document.

Rendered HTML/PDF/DOCX Reports

PDF + YAML
IDE integration: RStudio addins (emojified): 📦 🎉 ✂️ 🗑️

```
---
title: "Explore tidyverse with liftr"
author: "Nan Xiao <me@nanx.me>"
date: "r Sys.Date()"
output:
  rmarkdown::pdf_document:
    toc: true
    number_sections: true
  liftr:
    from: "rocker/tidyverse:latest"
maintainer: "Nan Xiao"
email: "me@nanx.me"
pandoc: false
texlive: true
cran: - nycflights13
---
\clearpage

# ggplot2

The example is from: https://github.com/tidyverse/ggplot2.

```
Philosophies

- **Continuous reproducibility.** Reproducible research should be a continuous process, instead of simply archiving code/data.

- **Document first.** R Markdown documents should be the center. Everything should be driven by documents, not packages.

- **Minimal footprint.** Connect R Markdown and Docker wisely, achieve more flexibility by doing less.
Applications

- **Individuals**: off-the-shelf solution for achieving persistent, environment-irrelevant reproducibility for data analysis.

- **Institutions**: key backend component for automated, large-scale report compilation/orchestration services.
DockFlow

Basic Workflows

**Sequence Analysis**

R Markdown | liftr config

Import fasta, fastq, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.

**Oligonucleotide Arrays**

R Markdown | liftr config

Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

**Annotation Resources**

R Markdown | liftr config

Introduction to using gene, pathway, gene ontology, homology annotations and the AnnotationHub. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

**Annotating Genomic Ranges**

R Markdown | liftr config

Represent common sequence data types (e.g., from BAM, gff, bed, and wig files) as genomic ranges for simple and advanced range-based queries.

**Annotating Genomic Variants**

R Markdown | liftr config

Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding

**Changing Genomic Coordinate Systems with rtracklayer::liftOver**

R Markdown | liftr config

The liftOver facilities developed in conjunction with the UCSC browser track infrastructure are available for transforming data in GRanges formats. This is illustrated

Easily containerized ~20 complex R Markdown workflows from Bioconductor.
Feature Roadmap

- Automatic inference of document dependencies (packrat)
- New renderers for bookdown, xaringan, and blogdown
- Improve CLI message interface (cli + crayon)
- Better Docker integration (reticulate + Docker API)
Acknowledgements

A special thanks to:

- Prof. Qing-Song Xu (Central South University)
- Prof. Matthew Stephens (University of Chicago)
- Dr. Tengfei Yin (Seven Bridges Genomics)
- Dr. Miaozhu Li (Duke University)

for offering me all the freedom, advice, and encouragement to develop better software for the statistical and R community.
Thank you! Questions?