

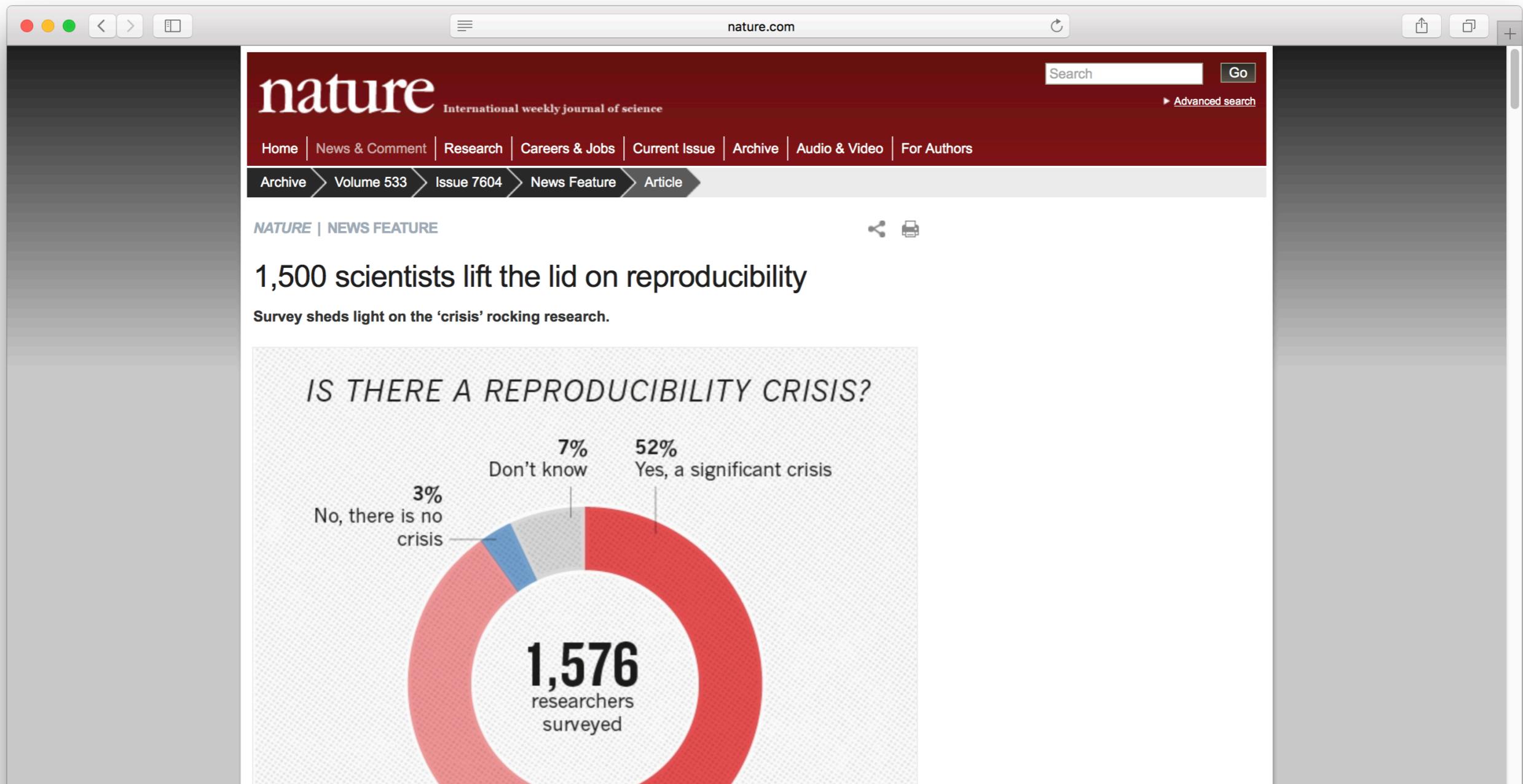
`liftr`: an R Package for Persistent Reproducible Research

JSM 2018

Nan Xiao
@road2stat

The Reproducibility Crisis

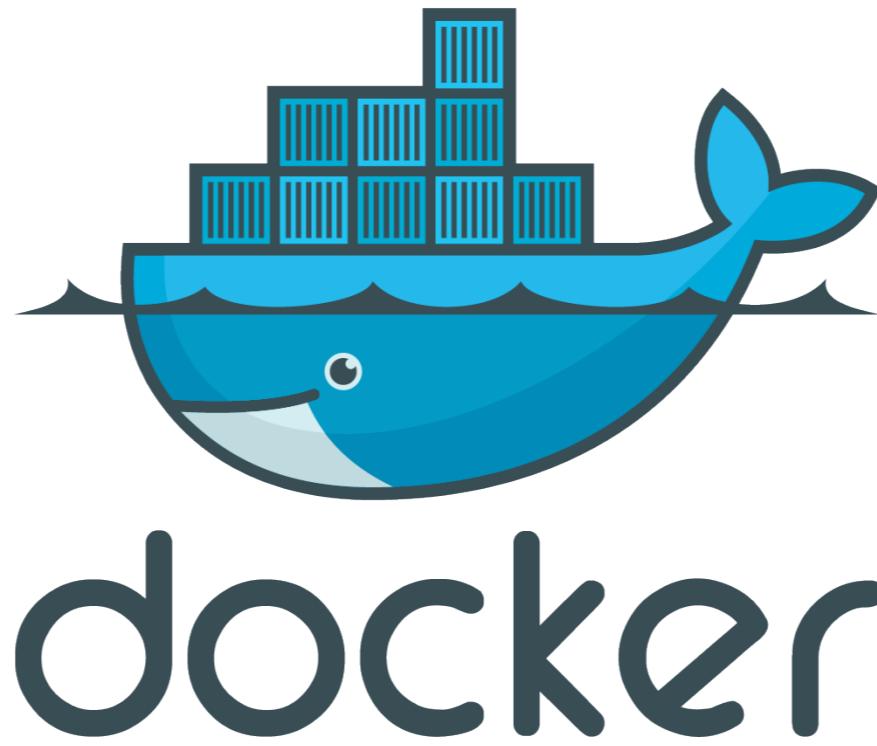
- Always a concern in both academia & industry.
- R Markdown + knitr pretty much saved the day.



The New Challenge

Even higher reproducibility for statistical computing:
regardless of *time* or *environment*.

Docker to the Rescue



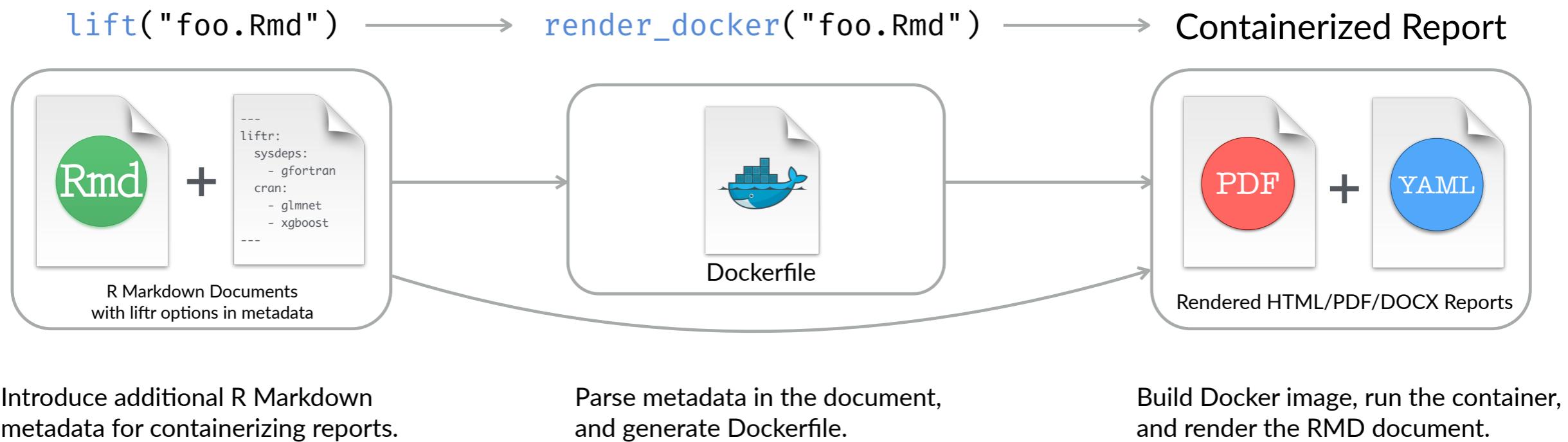
- Docker allows applications and their dependencies to be packaged into discrete runtime environments, called **containers**.
- Apps packaged in this way can run from diverse infrastructures.

Our Solution: liftr

Persistent, OS-level reproducibility for R Markdown documents.



Containerize R Markdown Documents as Easy as 1-2-3



~/liftr - master - RStudio

liftr-tidyverse.Rmd

```
1 ---  
2 title: "Explore tidyverse with liftr"  
3 author: "Nan Xiao <me@nanx.me>"  
4 date: `r Sys.Date()`  
5 output:  
6   rmarkdown::pdf_document:  
7     toc: true  
8     number_sections: true  
9 liftr:  
10   from: "rocker/tidyverse:latest"  
11   maintainer: "Nan Xiao"  
12   email: "me@nanx.me"  
13   pandoc: false  
14   texlive: true  
15   cran:  
16     - nycflights13  
17 ---  
18 \clearpage  
19 # ggplot2  
20  
21 The example is from: https://github.com/tidyverse/ggplot2.  
22  
23 ```{r}  
24 library("ggplot2")  
25  
26  
27
```

9:7 # Explore tidyverse with liftr

R Markdown

Addins ▾

CLIPR

Output to clipboard

Value to clipboard

LIFTR

Containerize

Render

Prune Dangling

Remove Image

PKGDOWN

Build pkgdown

RHANDSONTABLE

Edit a Data Frame

SEVENBRIDGES

Tool UI

Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > liftr > inst > examples

	Name	Size	Modified
..	..		
	bioc-rnaseq.bib	49 KB	Apr 10
	bioc-rnaseq.Rmd	74.5 KB	Apr 14
	liftr-minimal.Rmd	1018 B	Apr 10
	liftr-tidyverse.Rmd	1.5 KB	Dec 12

Console

IDE integration: RStudio addins (emojified):



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.org

The screenshot shows a web browser window with the URL `dockflow.org` in the address bar. The page title is "DockFlow". The navigation bar includes "Home" (which is highlighted in blue), "About", and "GitHub". The main content area is titled "Basic Workflows" and contains six sections, each with an "R Markdown" and "liftr config" button:

- Sequence Analysis**: 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**: 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**: 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**: 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**: 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**: 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. Miao Zhu 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?

