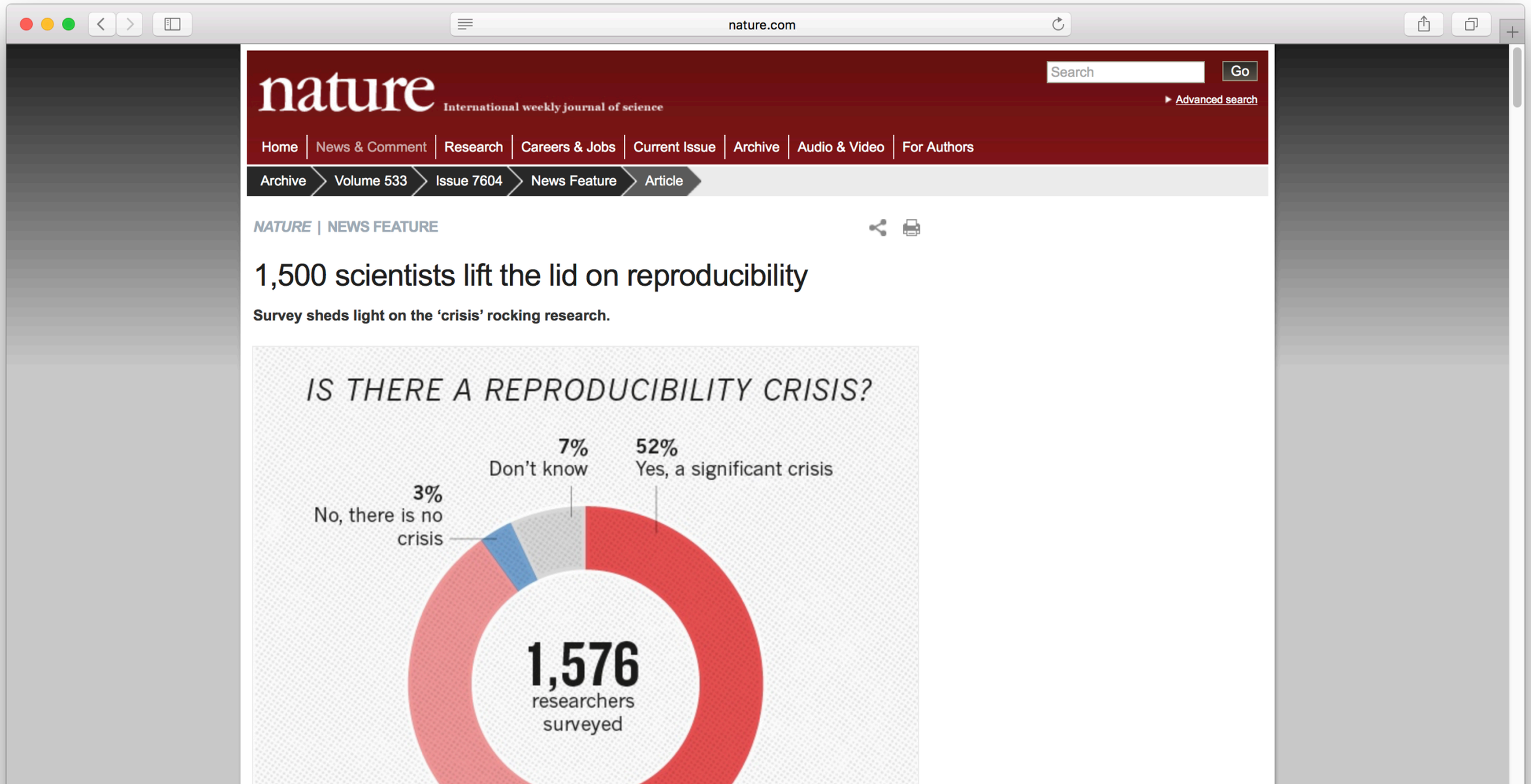
liftr: an R Package for Persistent Reproducible Research

JSM 2018

Nan Xiao
@road2stat

The Reproducibility Crisis

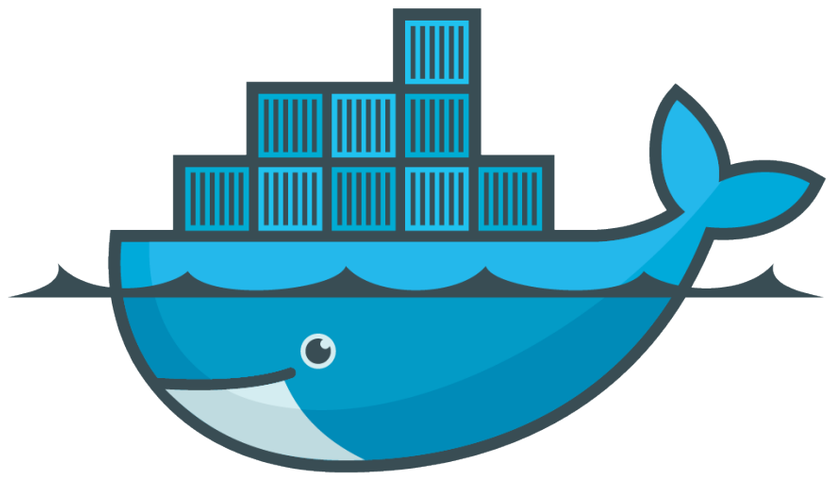
- Always a concern in both academia & industry.
- R Markdown + knitr pretty much saved the day.



The New Challenge

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

Docker to the Rescue



docker

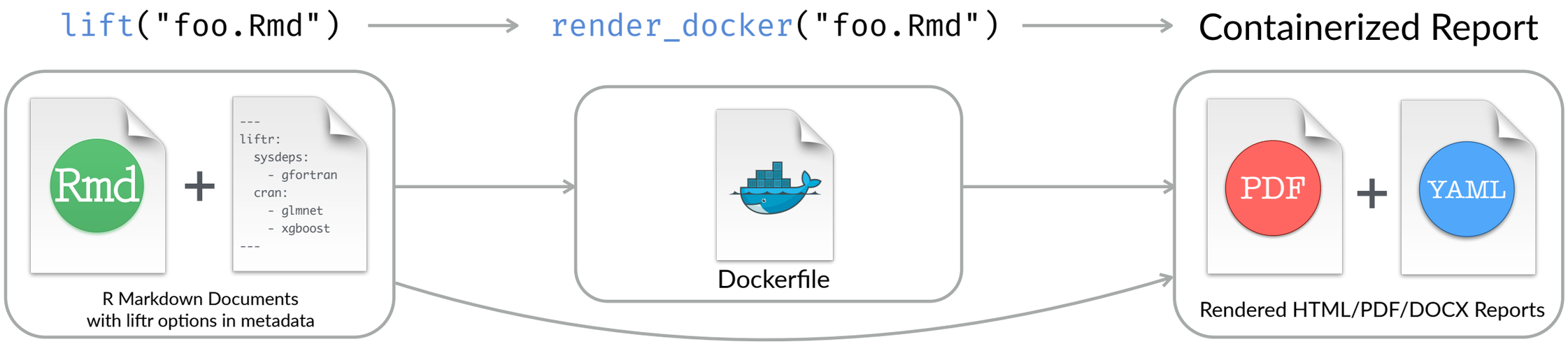
- Docker allows applications and their dependencies to be packaged into discrete runtime environments, called containers.

- Apps packaged in this way can run from diverse infrastructures.

Our Solution: liftr

Persistent, OS-level reproducibility for R Markdown documents.

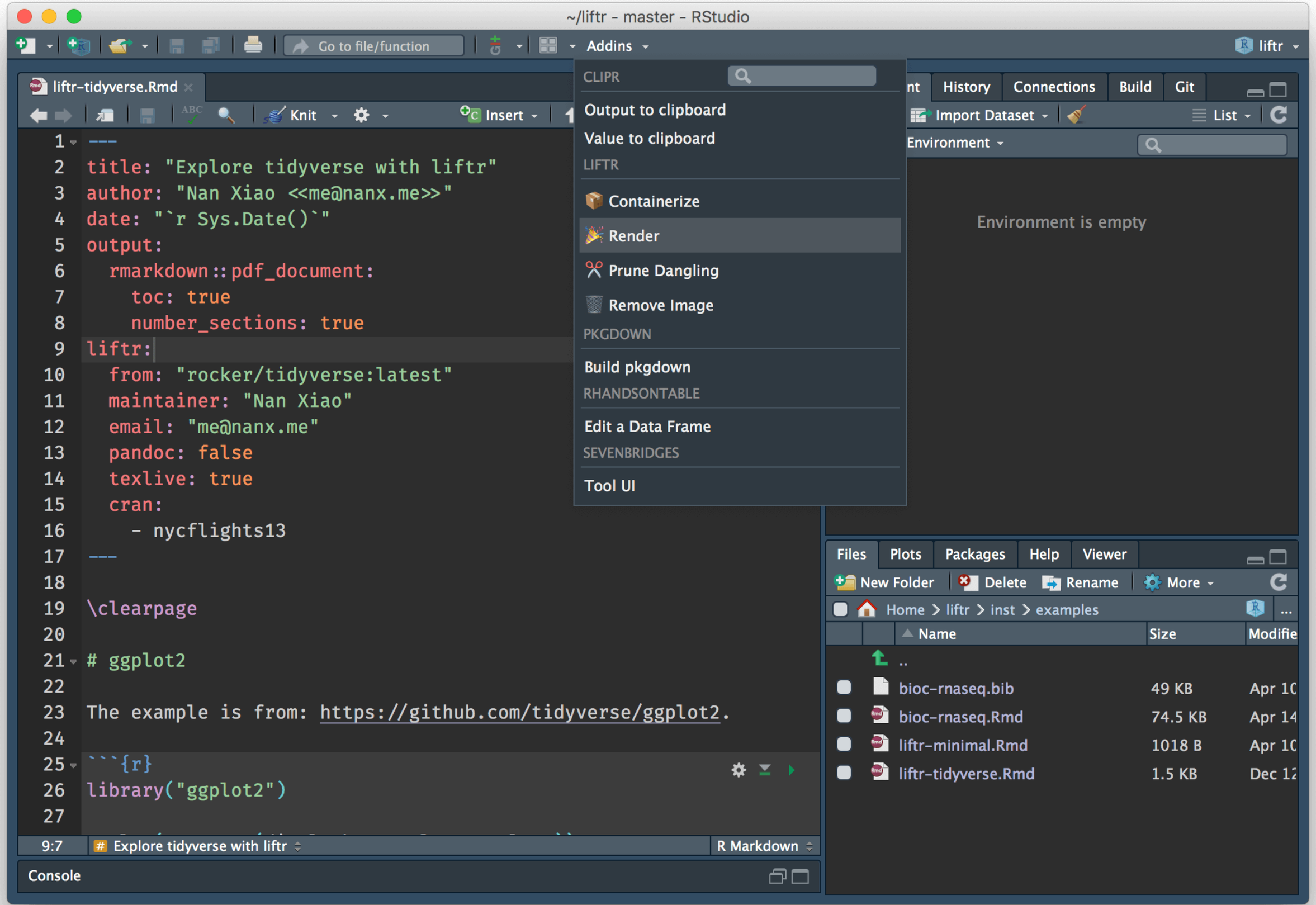
Containerize R Markdown Documents as Easy as 1-2-3



Introduce additional R Markdown metadata for containerizing reports.

Parse metadata in the document, and generate Dockerfile.

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



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.

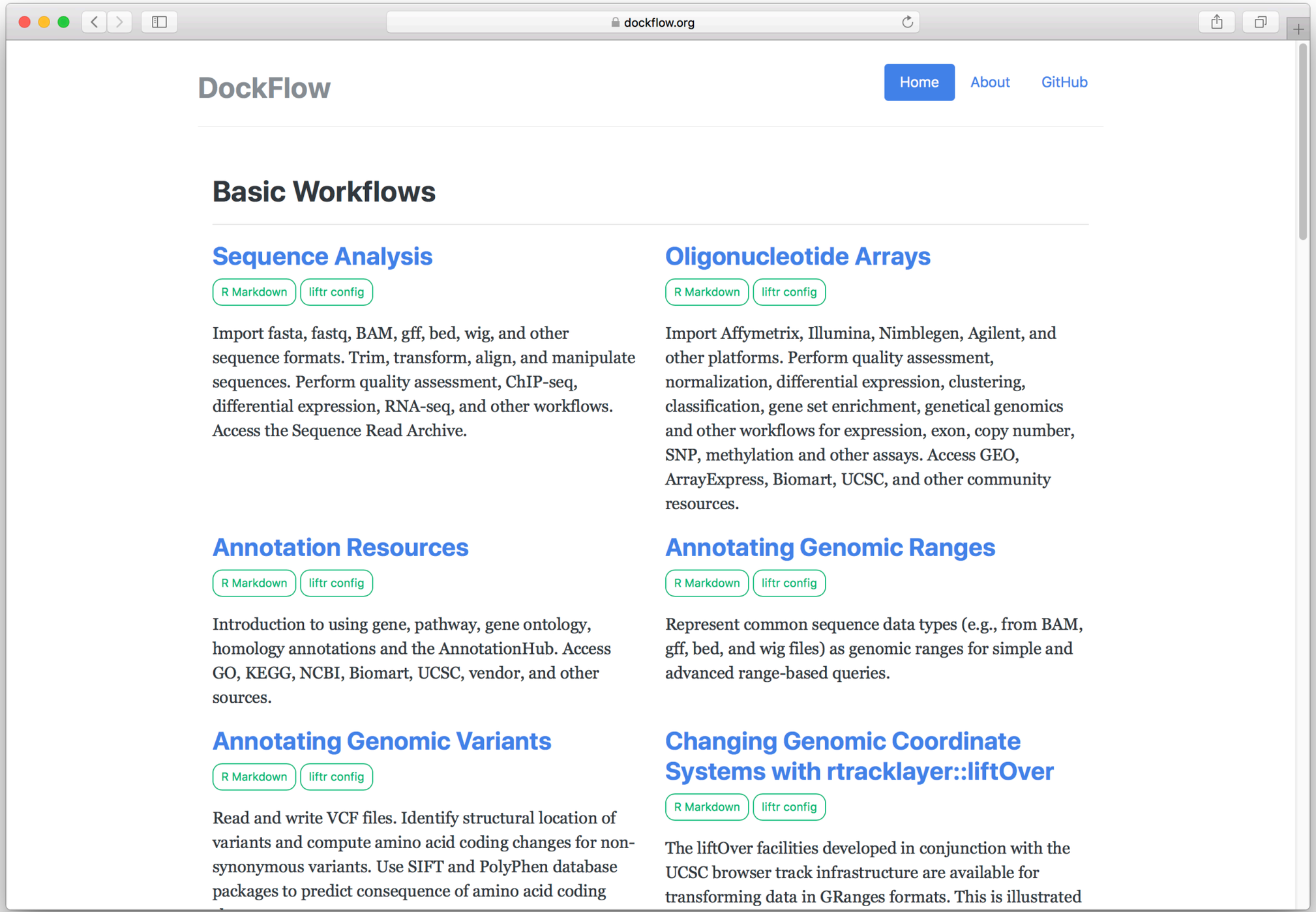
- **Minimal footprint.** Connect R Markdown and Docker wisely, achieve more flexibility by doing less.

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



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?