Persistent Reproducible Reporting

Nan Xiao

Genomic Data Scientist,
Seven Bridges

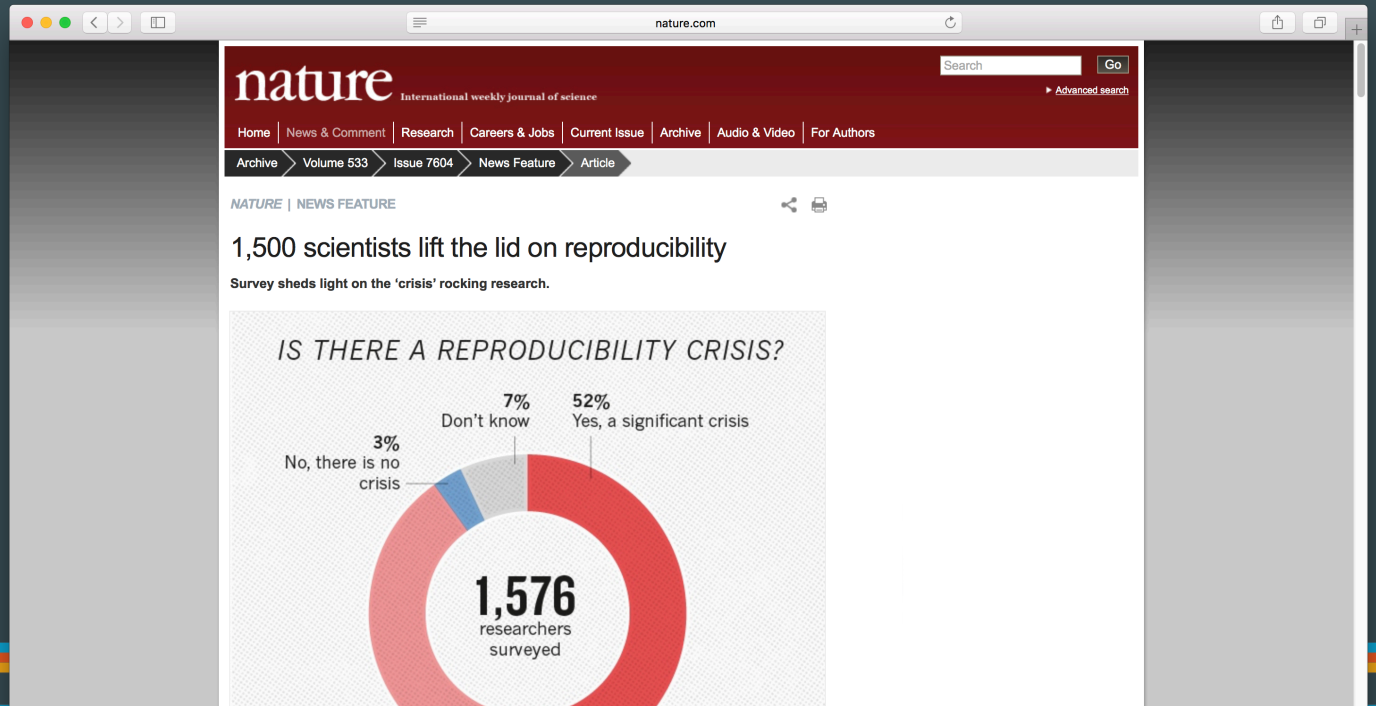
Reproducible Research

R Markdown + knitr to the rescue

+

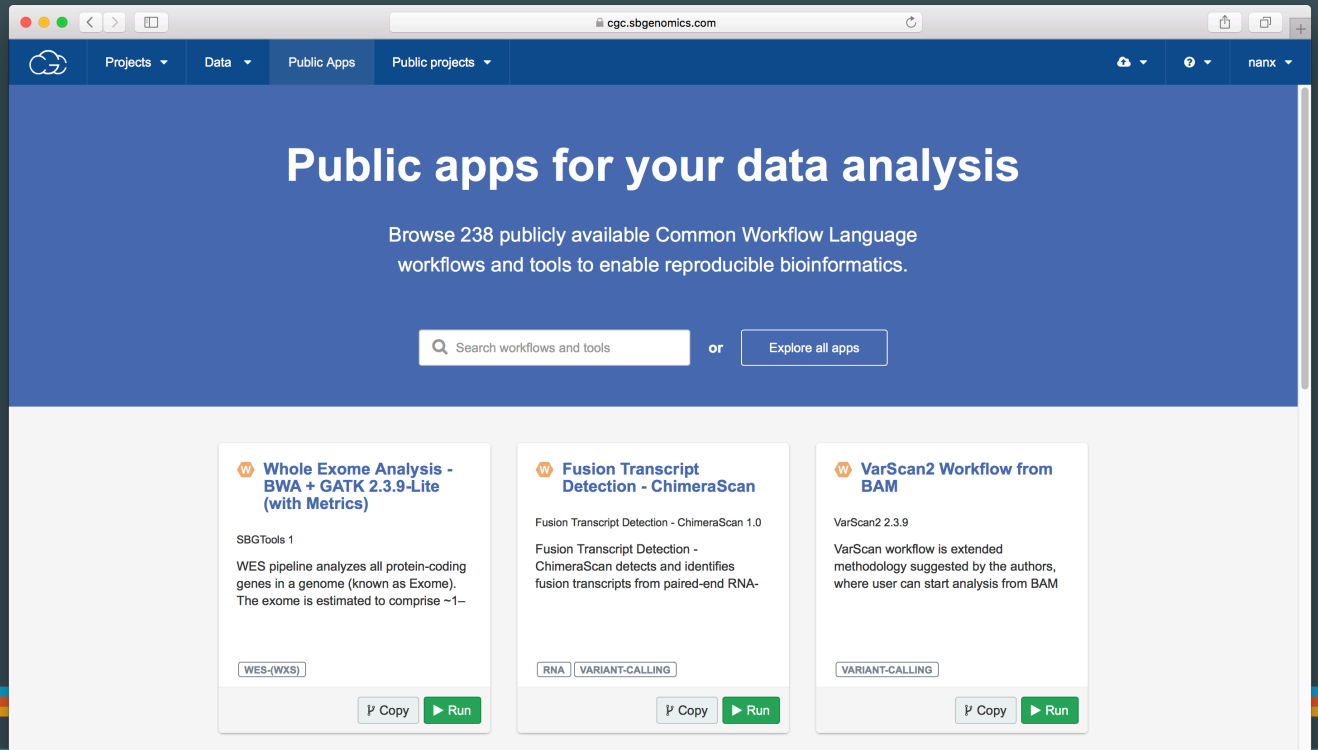
Reproducibility

... has always been a concern in both academia & industry.



Cancer Genomics Cloud

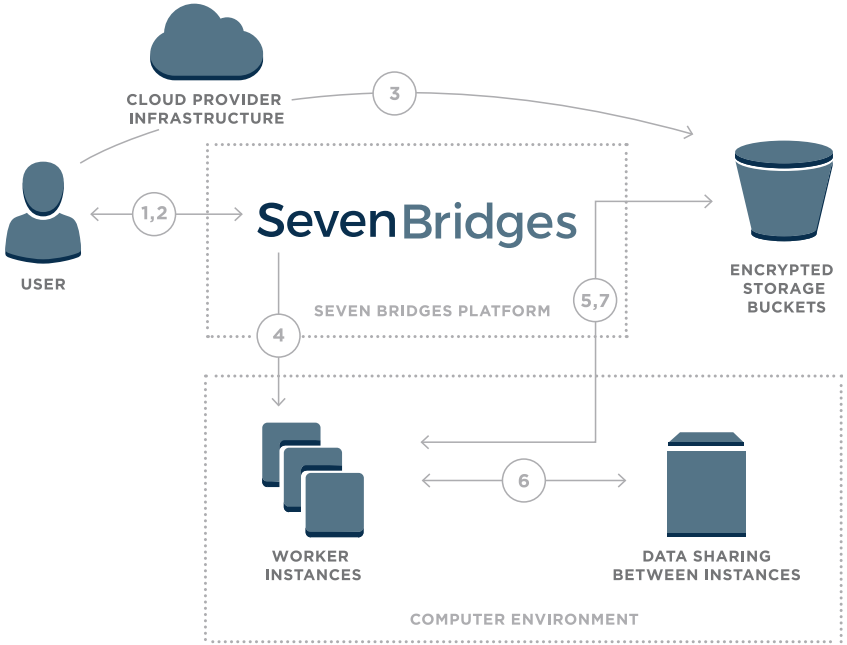
www.cancergenomicscloud.org



Hundreds of automated analysis workflows for petabyte-scale data from The Cancer Genome Atlas.

docker17 con

Product & Tech Innovations in CGC



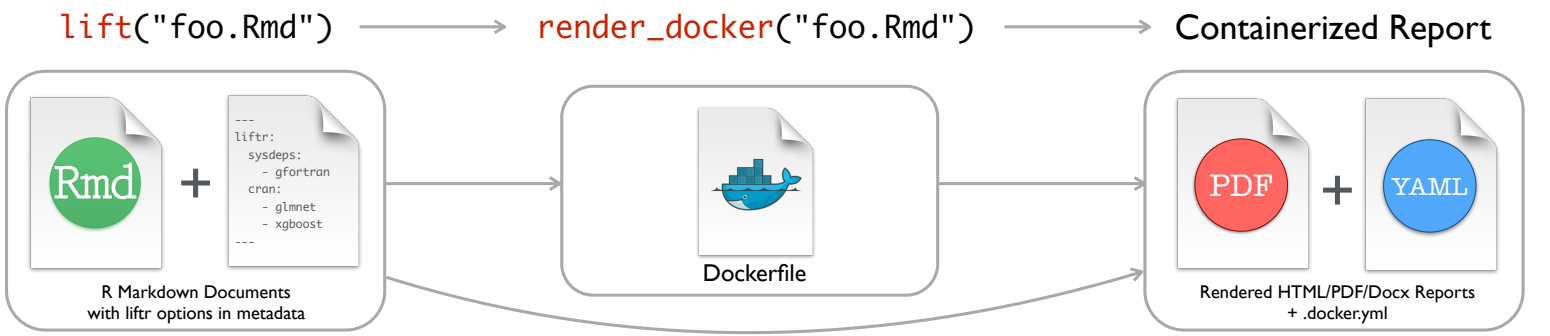
COMMON WORKFLOW LANGUAGE

Rabix

liftr

OS-level reproducibility & persistency for reports.

Dockerize documents as easy as 1-2-3



liftr extends the R Markdown metadata format, introducing additional options for containerizing and rendering reports.

By running lift() on the RMD file, liftr parses the metadata fields appeared in the R Markdown document; then generates the Dockerfile.

By running render_docker(), liftr will build the Docker image, run the container, and render the R Markdown document.

Dockerize documents as easy as 1-2-3

```
library("liftr")
input = "demo.Rmd"

lift(input)              # Generate Dockerfile
render_docker(input)     # Render report with Docker

purge_image(input)       # Clean up Docker image
push_image(input)        # Push image to registry (devel)
```

Demo: RNA-Seq Data Analysis

Example workflow from Bioconductor

- RNA-Seq: biotechnology for measuring the expression of genes. It can help identify key genes in cancer.
- TBs of RNA-Seq data are generated. Computational tools and workflows are developed to analyze such data.
- How to ensure such reports are reproducible through time, when datasets, analysis tools are both evolving?
- Code available from: `bit.ly/liftrdemo`

Step 1

Add liftr metadata to the R Markdown document:

base image, system dependencies, package dependencies, etc.

RStudio

Go to file/function | Addins

bioc-rnaseq.Rmd

Knit | Insert | Run

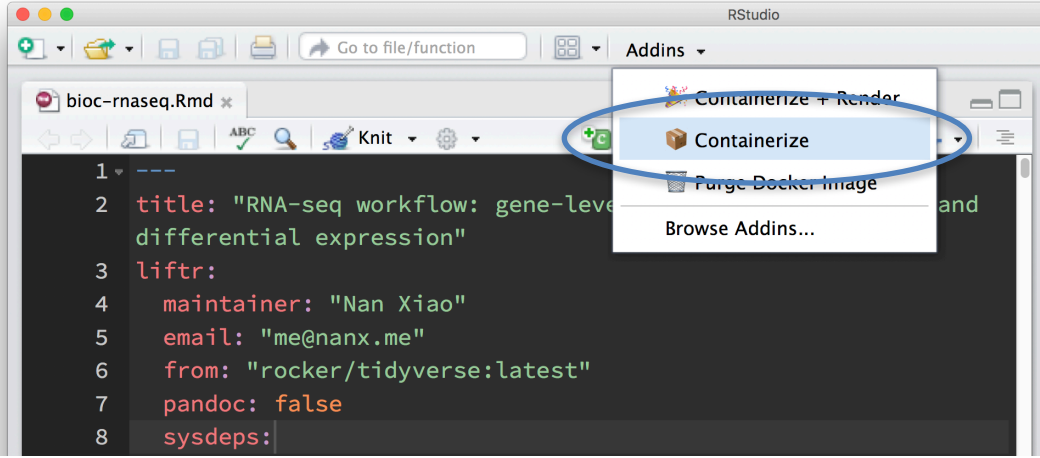
```
1  ---
2  title: "RNA-seq workflow: gene-level exploratory analysis and
   differential expression"
3  liftr:
4    maintainer: "Nan Xiao"
5    email: "me@nanx.me"
6    from: "rocker/tidyverse:latest"
7    pandoc: false
8    sysdeps:
9      - samtools
10   cran:
11     - pheatmap
12     - RColorBrewer
13     - PoiClaClu
14     - ggbeeswarm
15   bioc:
16     - BiocStyle
17     - airway
18     - Rsamtools
19     - GenomicFeatures
20     - GenomicAlignments
```

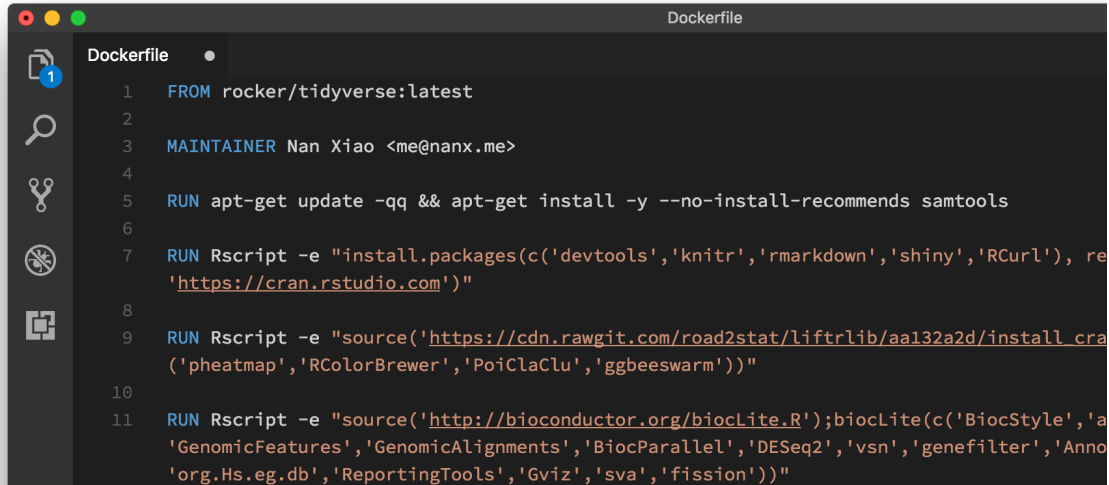
44:30 | RNA-seq workflow: gene-level exploratory analysis and differential expression | R Markdown

Console

Step 2

Use `liftr::lift` to generate `Dockerfile`



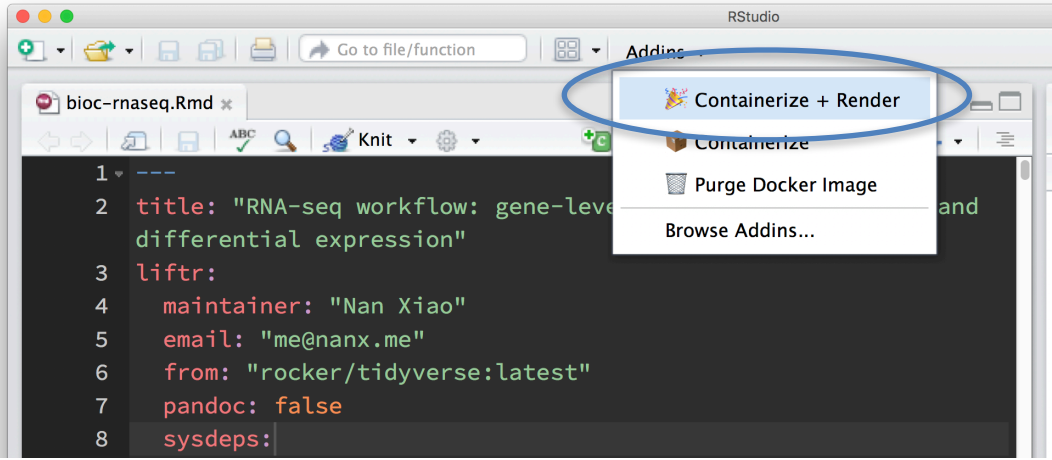


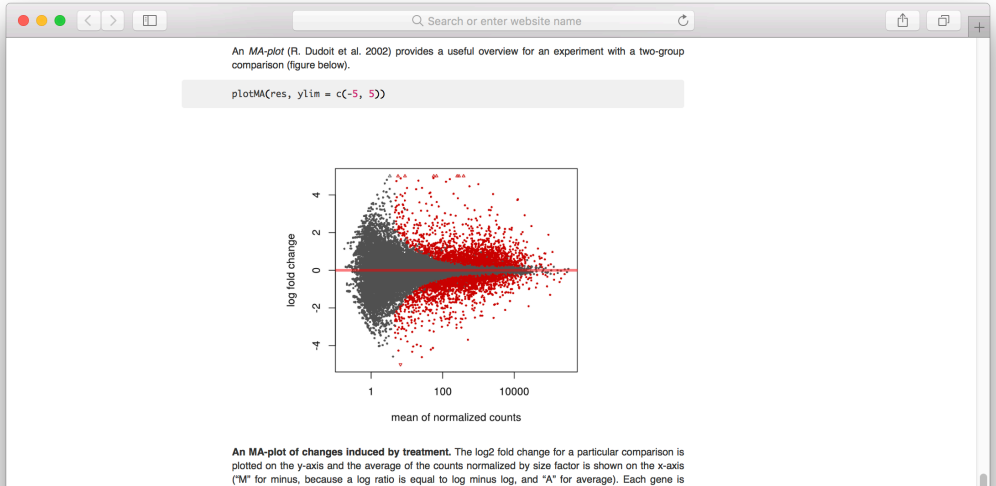
Step 3

`liftr::render_docker` will build the image, run the container, and render into PDF/HTML/Docx.

Re-compilation: cached image layers are used to improve speed.

Remove the used image, or push to registry.





Future works

- Cloud-based rendering and containerization services for dynamic documents
- Democratize reproducible report creation/sharing

Thank You!

liftr.me

@road2stat

#dockercon #liftr