Persistent Reproducible Reporting

Nan Xiao, Seven Bridges

2017/05/20 @ China R Conference Beijing
DOCUMENT-LEVEL REPRODUCIBILITY
R MARKDOWN + KNITR TO THE RESCUE
REPRODUCIBILITY

... has always been a concern in both academia & industry.
CANCER GENOMICS CLOUD (CGC)

- [www.cancergenomicscloud.org](http://www.cancergenomicscloud.org)
- Hundreds of automated analysis workflows for petabyte-scale data from The Cancer Genome Atlas
PRODUCT & ENGINEERING INNOVATIONS IN CGC

1. USER
2. CLOUD PROVIDER INFRASTRUCTURE
3. SevenBridges
4. SEVEN BRIDGES PLATFORM
5. ENCRYPTED STORAGE BUCKETS
6. WORKER INSTANCES
7. DATA SHARING BETWEEN INSTANCES

COMMON WORKFLOW LANGUAGE

Rabix
CHALLENGE: OS-LEVEL REPRODUCIBILITY

How to ensure your reports are reproducible across time and environments, when the data, analysis tools, operating systems are all evolving?
Docker allows applications and their dependencies to be packaged into discrete runtime environments, called containers. Applications packaged in this way can be run from many diverse infrastructures.
liftr

OS-level reproducibility & persistency for reports.
**DOCKERIZE DOCUMENTS AS EASY AS 1-2-3**

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

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

**R Markdown Documents**

+ lifr: sysdeps:
  - gfortran
  - curl
  - glmnet
  + xgboost

**Dockerfile**

**Rendered HTML/PDF/Docx Reports**

+ PDF
+ YAML

**Containerized Report**

lifr extends the R Markdown metadata format, introducing additional options for containerizing and rendering reports.
DOCKERIZE DOCUMENTS AS EASY AS 1-2-3

```r
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.org

- RNA-Seq: biotechnology for measuring the expression of genes. It can help identify potential key genes in cancer.
- TBs of RNA-Seq data are generated. Computational tools and workflows are developed to analyze such data.
- We need to ensure such reports are reproducible through time, when datasets, analysis tools are both evolving.
STEP 1

Add liftr metadata to the R Markdown document:

- Base image
- System dependencies
- Package dependencies
- ...
STEP 2

Generate Dockerfile with liftr::lift
STEP 3

- `liftr::render_docker` will build the Docker image, run the container, and render into PDF/HTML/DOCX.
- Re-compile: cached Docker image layers are used to improve speed.
- Remove the used image, or push to Docker registry.
We aim to expand the R Markdown tool chain by exploring the next frontier: system-level reproducibility, and democratize reproducible report creation/sharing.

To achieve this, we need:

- Standard renderers + independent YAML configuration file
- Better IDE support (RStudio Addins)
- Better on-boarding experience: automatic dependency parsing
- Cloud-based rendering and containerization services for dynamic documents
Q & A

Visit liftr.me for more info

Contact: me@nanx.me