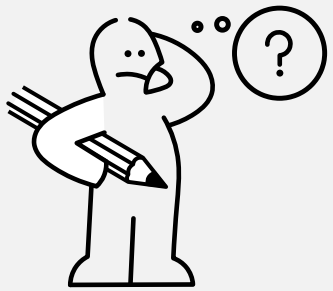
A (rough and incomplete) guide to making your data analysis (and a bunch of other work you do) *more* reproducible

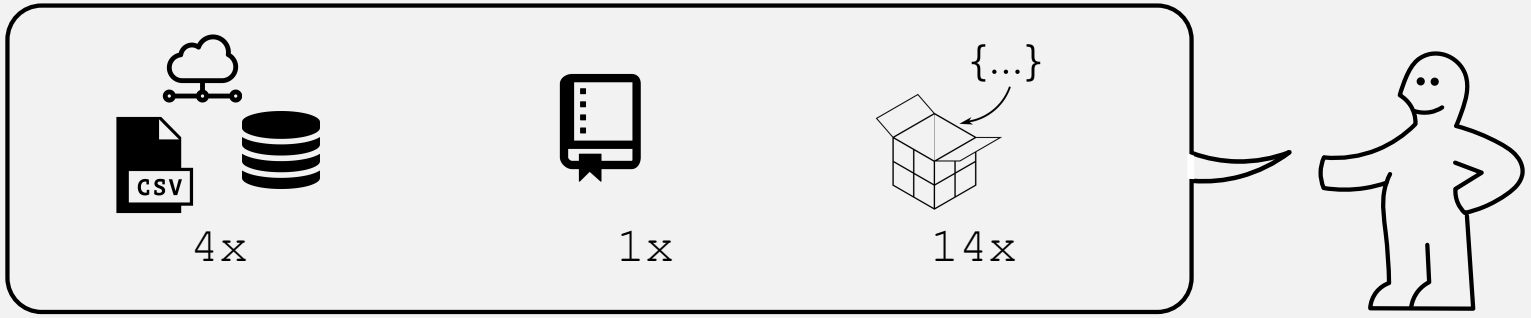
Karthik Ram • rOpenSci // UC Berkeley

KÖMPENDIUM

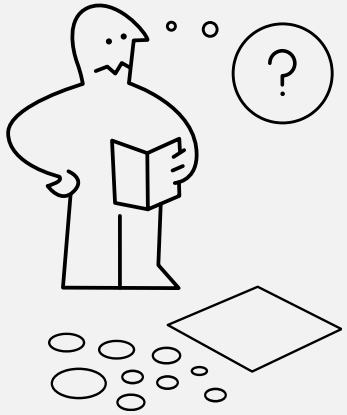
1.



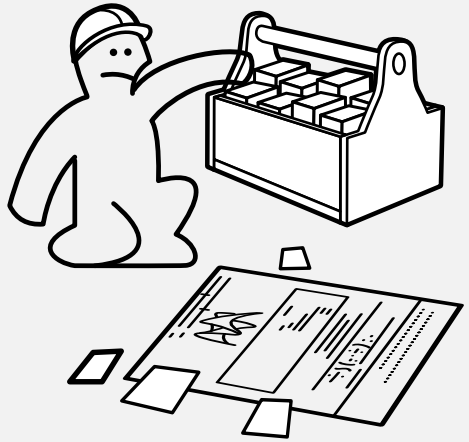
2.



3.

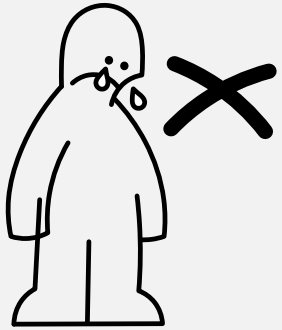


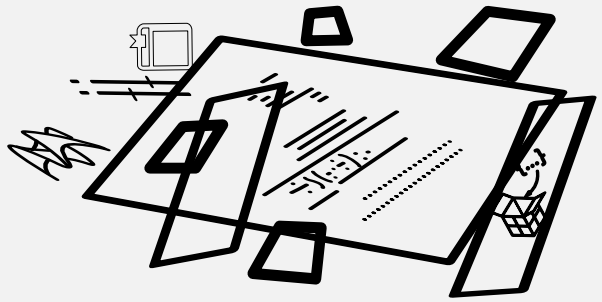
4.



5.

6.



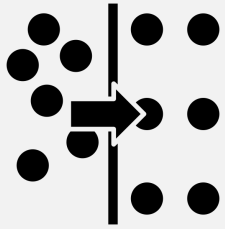


Research compendia

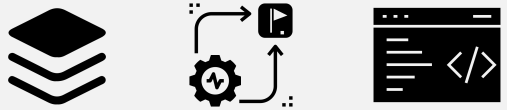
*" ...We introduce the concept of a **compendium** **as both a container** for the different elements that make up the **document** and its computations (i.e. **text**, **code**, **data**, ...), and as a means for **distributing**, **managing** and **updating** the collection.*

Gentleman and Temple Lang, 2004

Research compendium principles



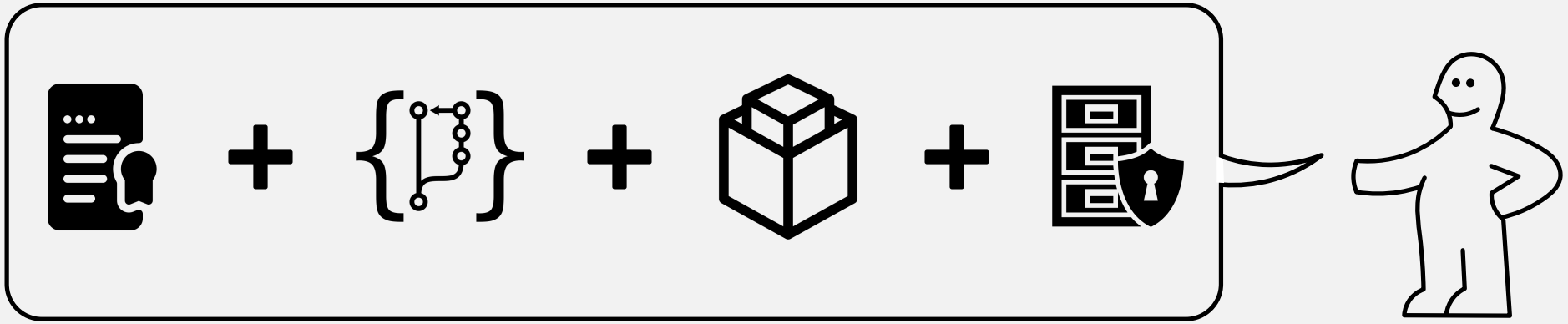
Stick with the conventions of your peers



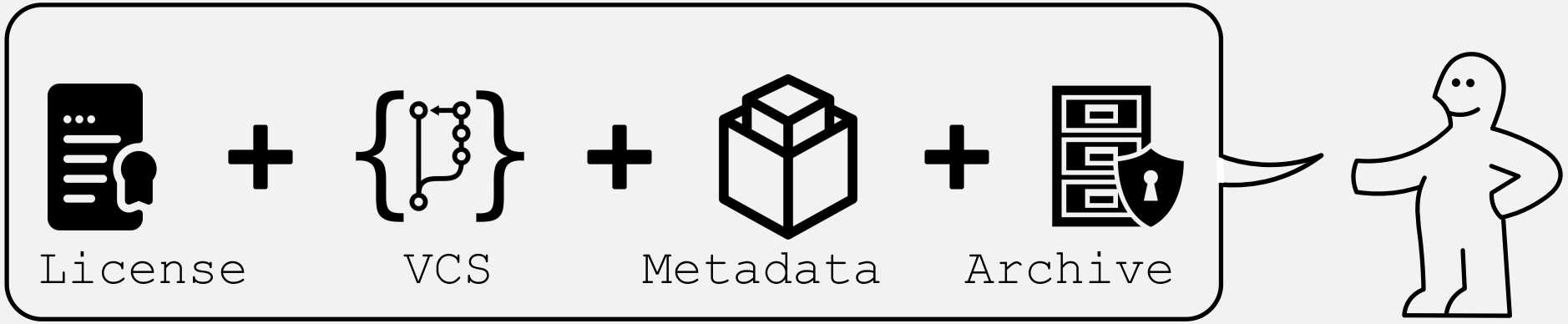
Keep data, methods and outputs separate

Specify your computational environment as clearly as you can

Key components you'll need for sharing a compendium



Key components you'll need for sharing a compendium



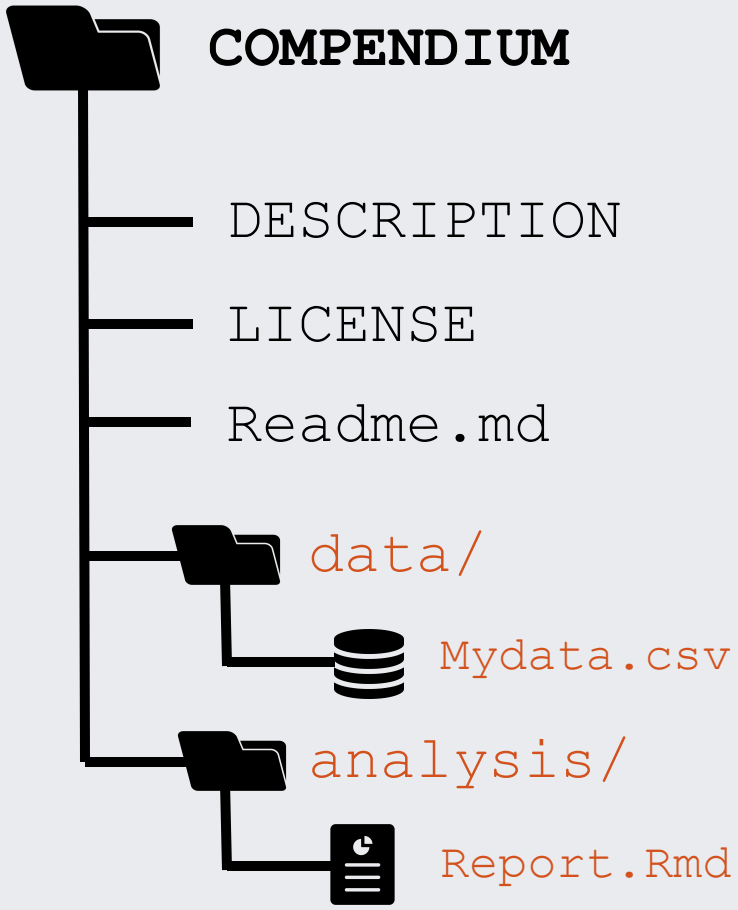
The R package structure is great way to organize and share a compendium!

```
Package: glue
Title: Interpreted String Literals
Version: 1.3.0.9000
Authors@R: person("Jim", "Hester", email = "james.f.hester@gmail.com", role =
c("aut", "cre"))
Description: An implementation of interpreted string literals, inspired by
  Python's Literal String Interpolation <https://www.python.org/dev/peps/pep-0498/>
and Docstrings
  <https://www.python.org/dev/peps/pep-0257/> and Julia's Triple-Quoted String
Literals
  <https://docs.julialang.org/en/stable/manual/strings/#triple-quoted-string-
literals>.
Depends:
    R (>= 3.1)
Imports:
  methods
Suggests:
    testthat,
    (and many more)
License: MIT + file LICENSE
Encoding: UTF-8
LazyData: true
RoxygenNote: 6.0.1
Roxygen: list(markdown = TRUE)
URL: https://github.com/tidyverse/glue
BugReports: https://github.com/tidyverse/glue/issues
VignetteBuilder: knitr
ByteCompile: true
```

```
Type: Compendium
Package: pomdpintro
Version: 0.1.0
Depends: nimble, tidyverse, sarsop, MDPtoolbox
Suggests: extrafont, hrbrthemes, Cairo, ggthemes
Remotes: boettiger-lab/sarsop
```

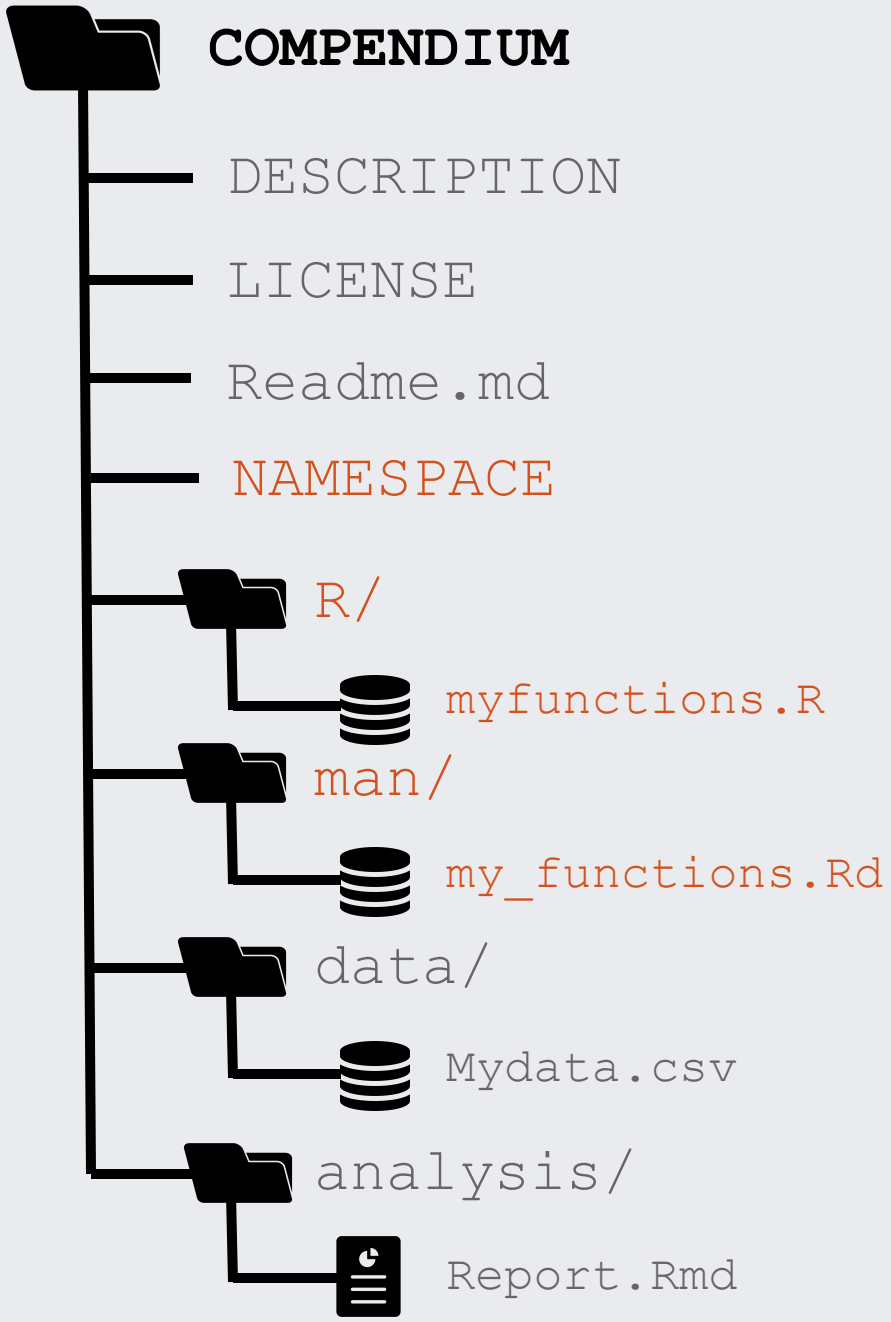
Packaging your analysis as a compendium gives you access to powerful developer tools

Small compendia



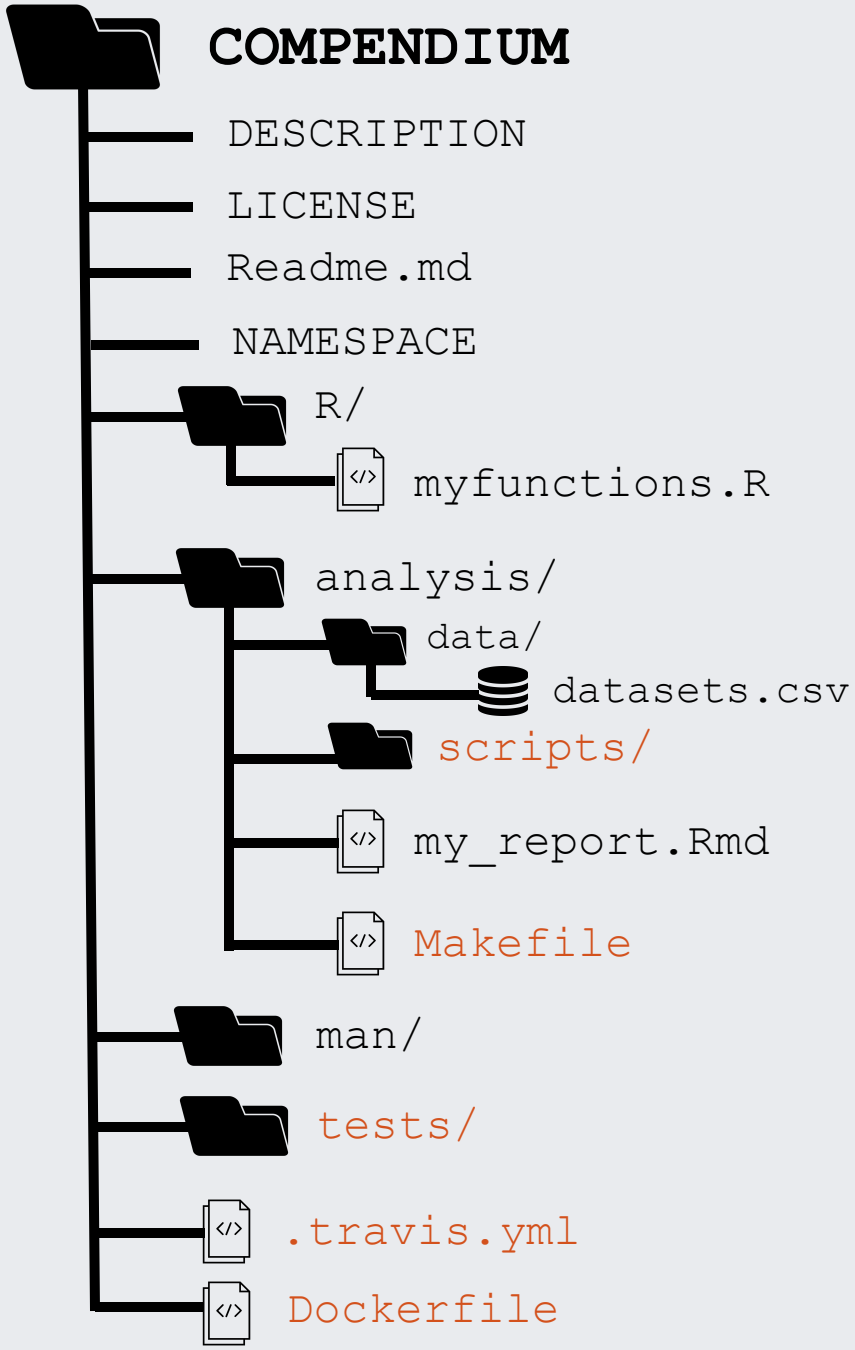
Marwick et al 2017

Medium compendia



Marwick et al 2017

Large/complex compendia



Marwick et al 2017

Data (Small → Medium)

Computing environment

Workflows

1. Data

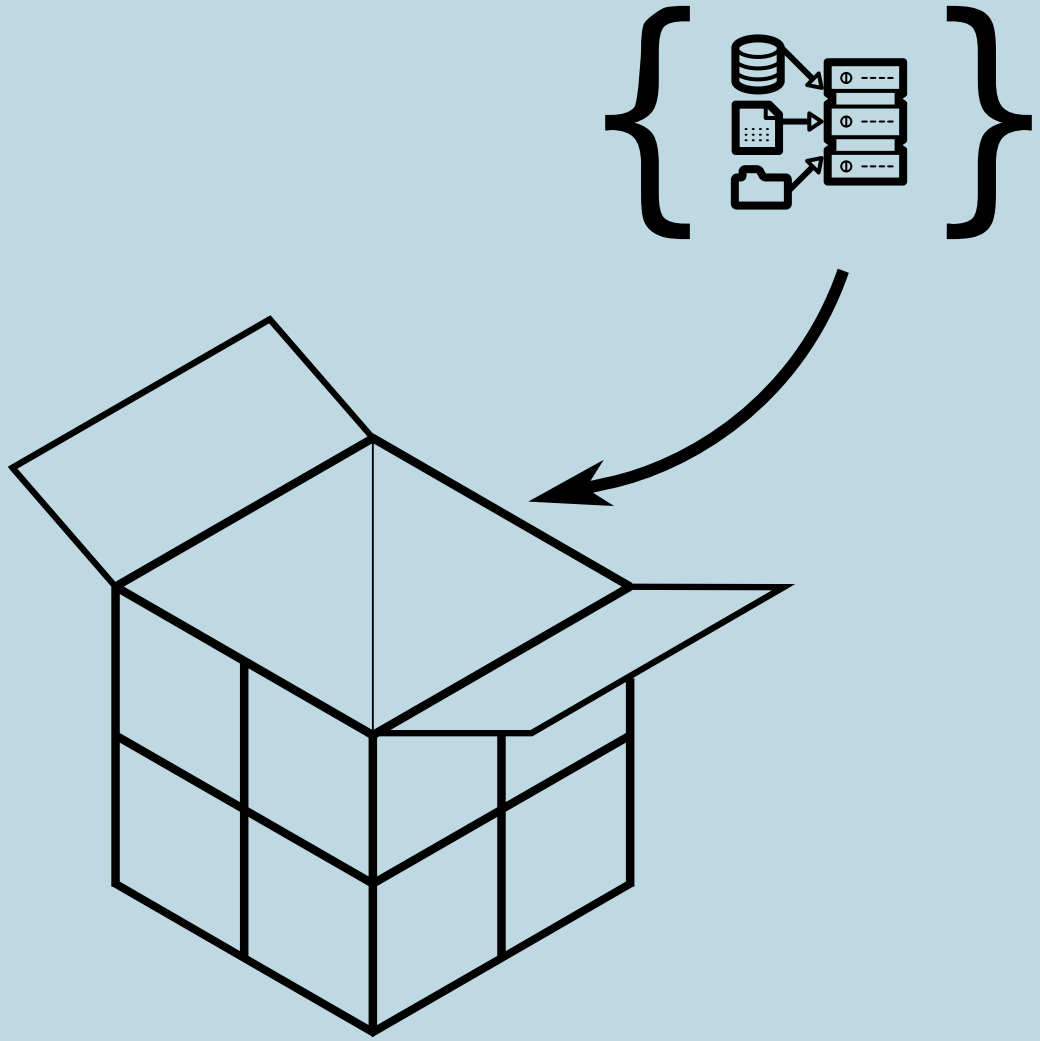
How does one manage small to medium data in the context of a research compendium?

Small data

Put small data inside packages, especially if you ship a methods package with your analysis

CRAN = < 5 mb.

37% of the 13K packages on CRAN have some form of data.



piggyback

Attach large [data] files to Github repositories

github.com/ropensci/piggyback

Leveraging Github releases to share medium sized files

github.com/ropensci/piggyback

```
pb_new_release("user/repo", "v0.0.5")
pb_upload("datasets.tsv.xz", "user/repo")
# Access them in your scripts with
# pb_download
```

Latest release

v-1.0

91e9dbe

Verified

Compendium release v-1.0

karthik released this 3 minutes ago

Assets 3

mtcars.tsv.xz 604 Bytes

Source code (zip)

Source code (tar.gz)

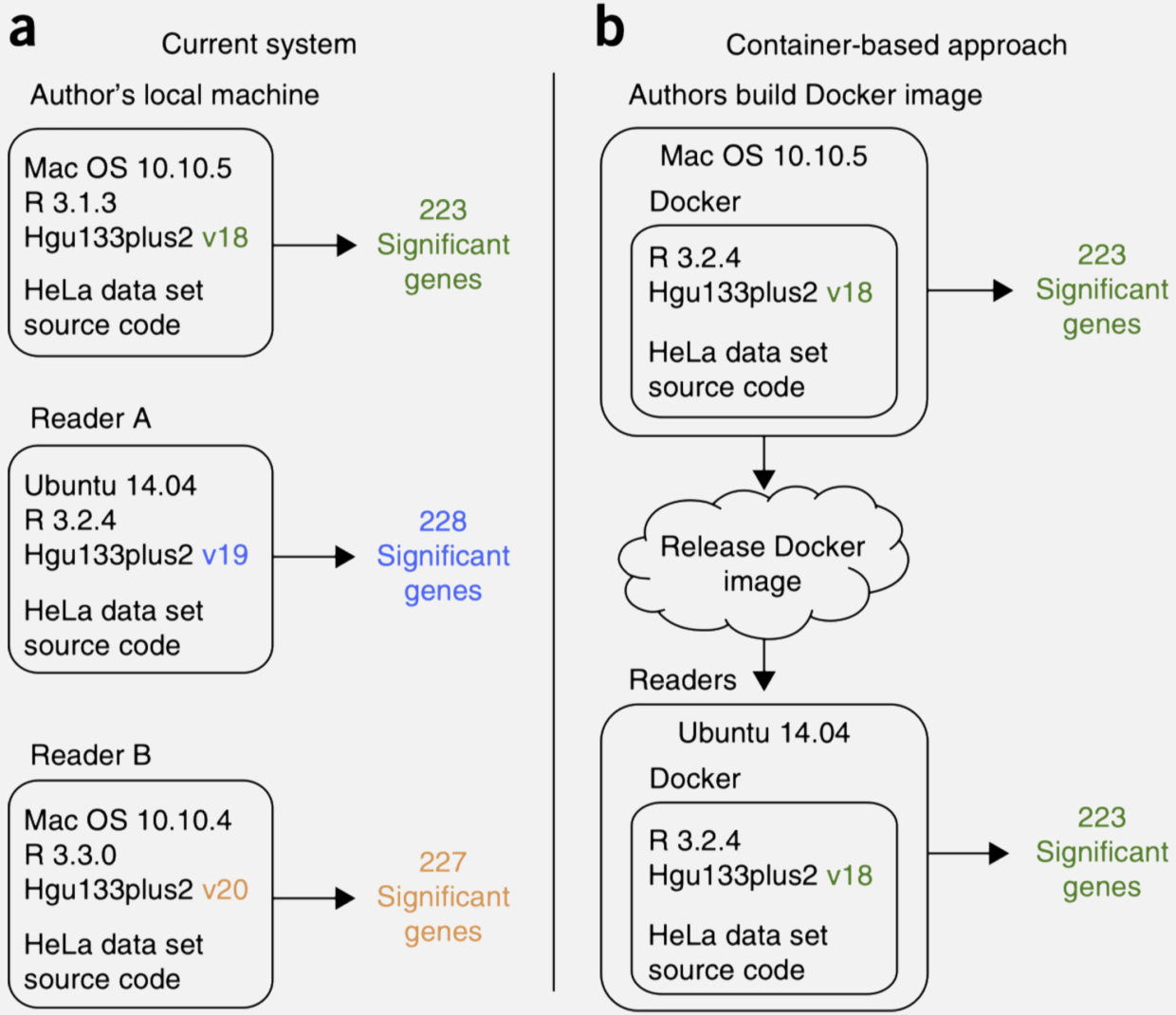
Initial release of project datasets

Medium data

github.com/ropensci/arkdb

2. Isolate your computing environment

It's important to isolate the computing environment **so that changes in software dependencies don't break your analysis.**



doi:10.1038/nbt.3780

Adding a Dockerfile to your compendium

Many ways to write a Dockerfile for your R project

o2r/containerit
jupyter/repo2docker

Binder

mybinder.org

Binder is an open source project that is designed to make it **really easy to share analyses that are in notebooks.**

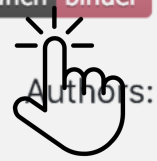
Binder

mybinder.org

README.md

Resolving the measurement uncertainty paradox in ecological management

launch binder build passing



- Authors: Milad Memarzadeh, Carl Boettiger

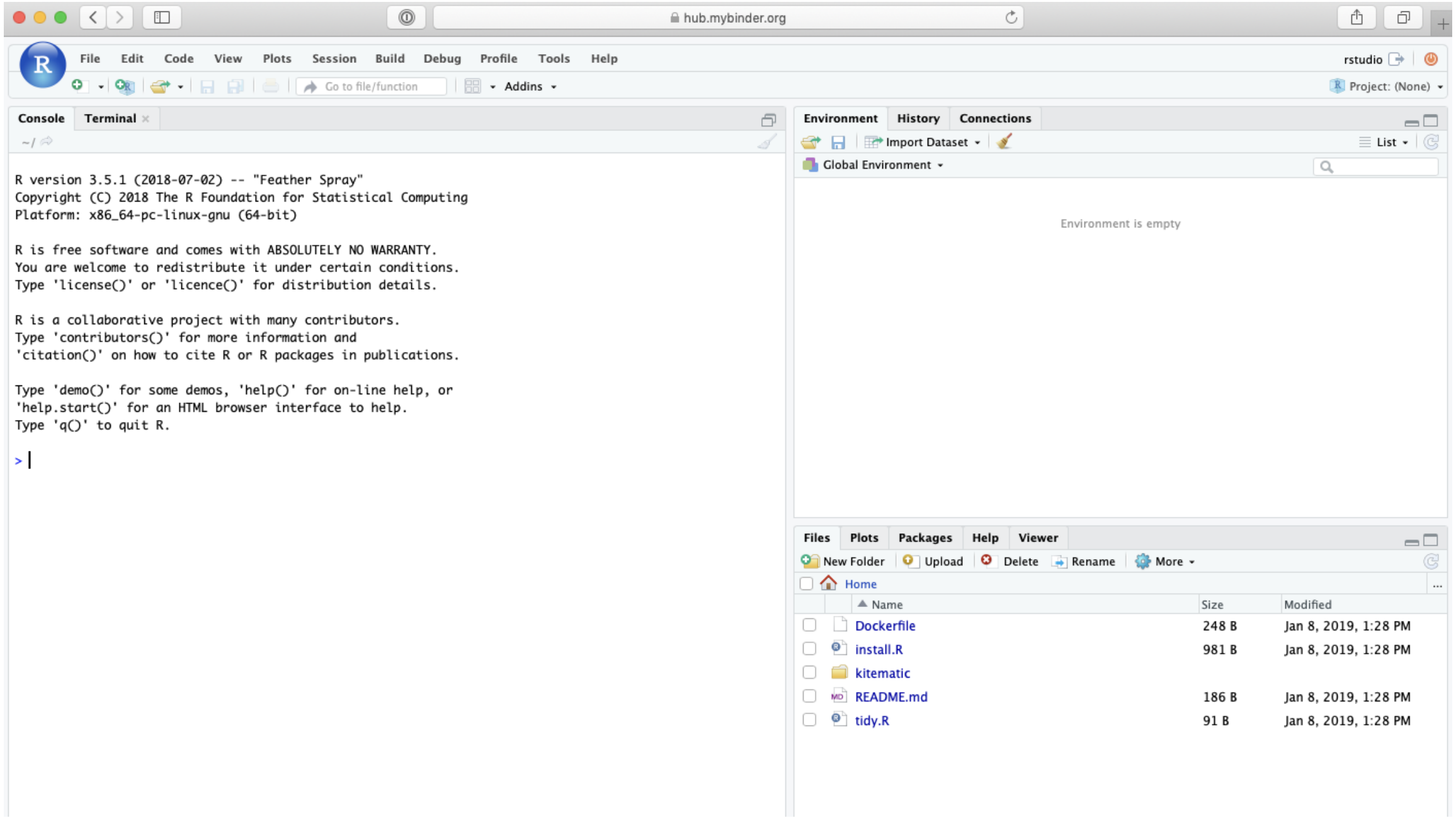
Contents

- Manuscript: R Markdown source document for manuscript. Includes code to reproduce for figures from tables generated by the analysis.
- Appendix: R Markdown source documents for both appendices, containing all necessary R code to generate all

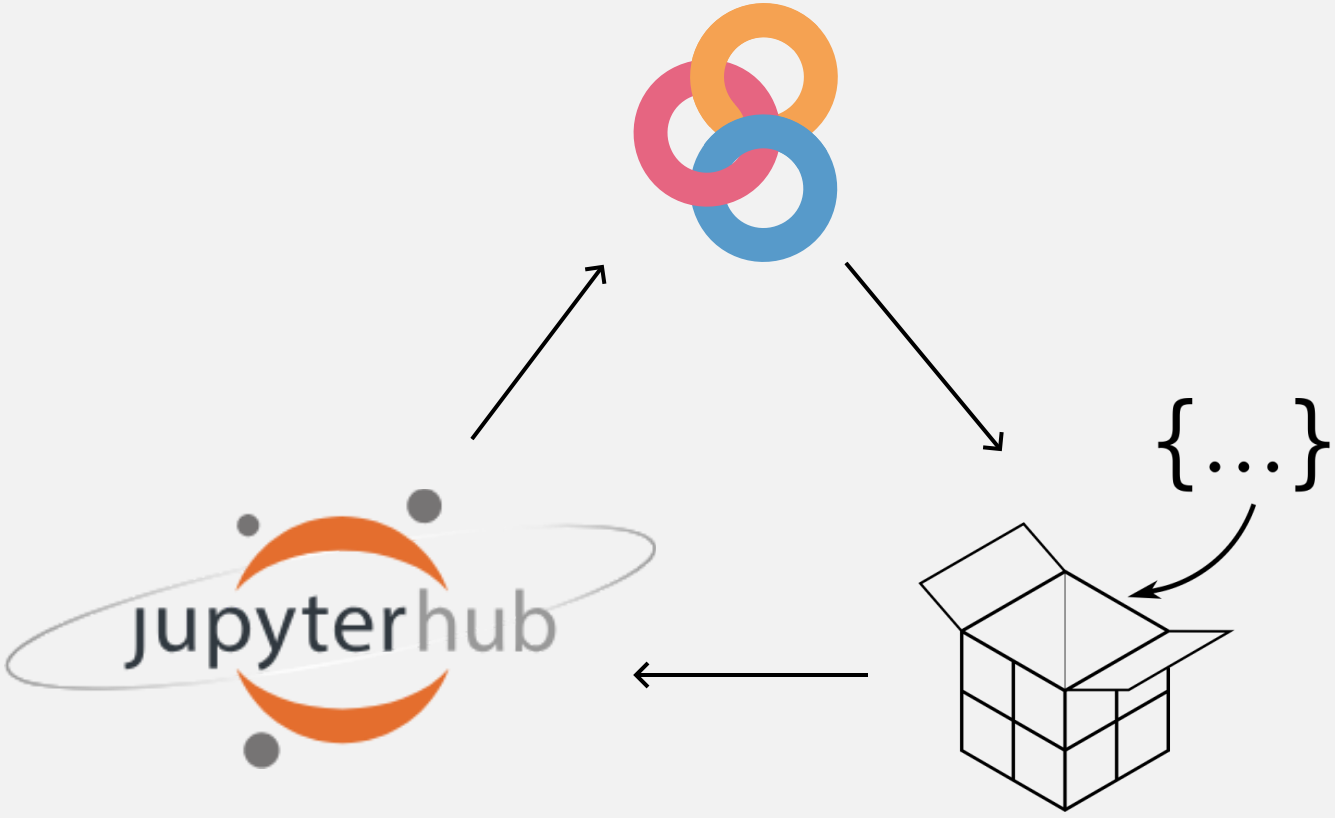
Loading repository: karthik/binder-test-fastest/master

Build logs hide

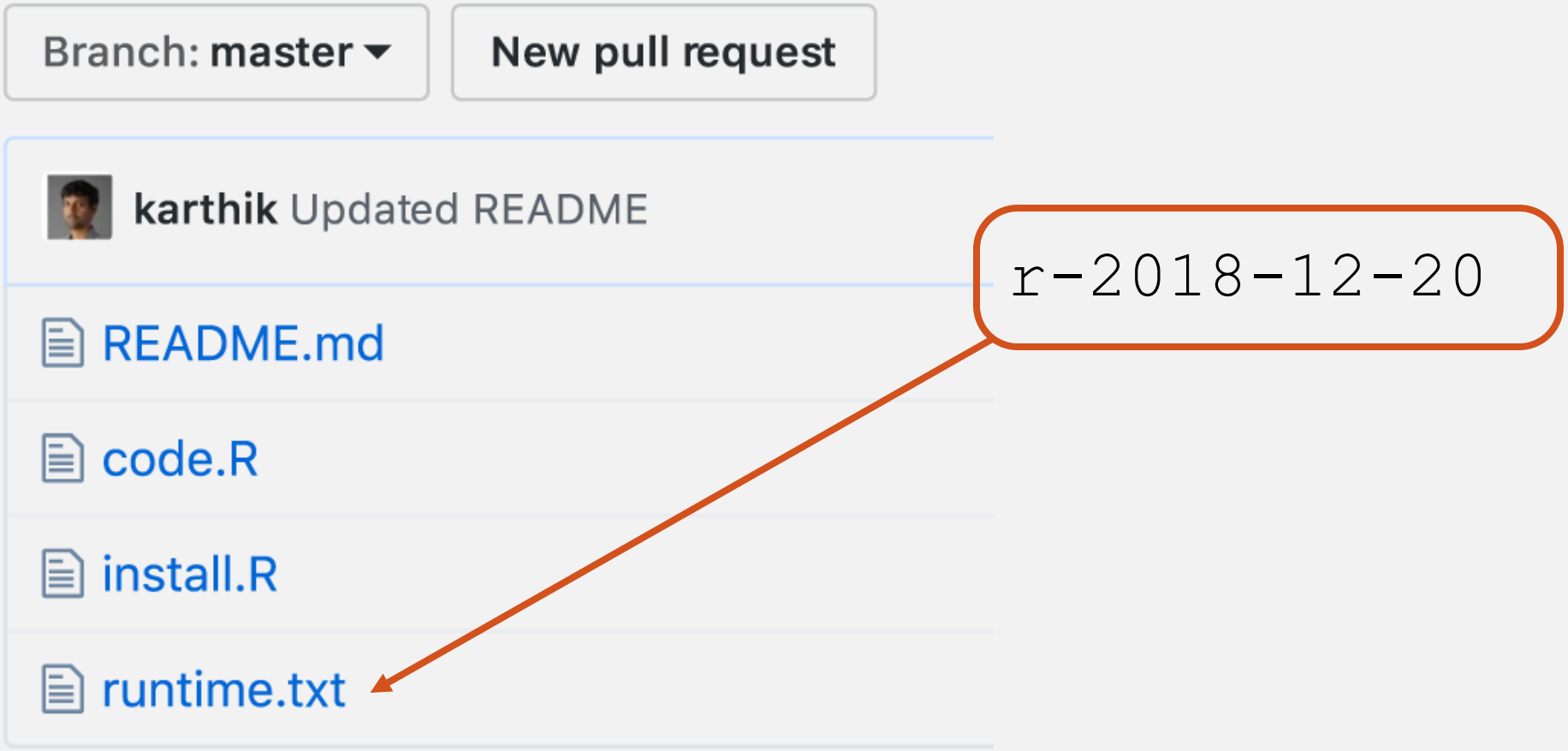
```
 ---> 8d5595591b60
Step 4/6 : RUN chown -R ${NB_USER} ${HOME}
 ---> Running in c0bec64ca779
Removing intermediate container c0bec64ca779
 ---> ae982cee19d0
Step 5/6 : USER ${NB_USER}
 ---> Running in dfe91fd9bcfa
Removing intermediate container dfe91fd9bcfa
 ---> 114a4cd0b227
Step 6/6 : RUN wget https://github.com/karthik/binder-test-fastest/raw/master/DESCRIPTION && R -e "devtools::install_deps()"
 ---> Running in 2dad32d8d3e2
--2019-01-13 02:04:34--  https://github.com/karthik/binder-test-fastest/raw/master/DESCRIPTION
Resolving github.com (github.com)... 192.30.253.112, 192.30.253.113
Connecting to github.com (github.com)|192.30.253.112|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/karthik/binder-test-fastest/master/DESCRIPTION [following]
--2019-01-13 02:04:34--  https://raw.githubusercontent.com/karthik/binder-test-fastest/master/DESCRIPTION
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.0.133, 151.101.64.133, 151.101.128.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.0.133|:443... connected.
HTTP request sent, awaiting response...
```


hub.mybinder.org
File Edit Code View Plots Session Build Debug Profile Tools Help
rstudio
Go to file/function
Addins
Project: (None)
Console Terminal
~/
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
Dockerfile 248 B Jan 8, 2019, 1:28 PM
install.R 981 B Jan 8, 2019, 1:28 PM
kitematic
README.md 186 B Jan 8, 2019, 1:28 PM
tidy.R 91 B Jan 8, 2019, 1:28 PM

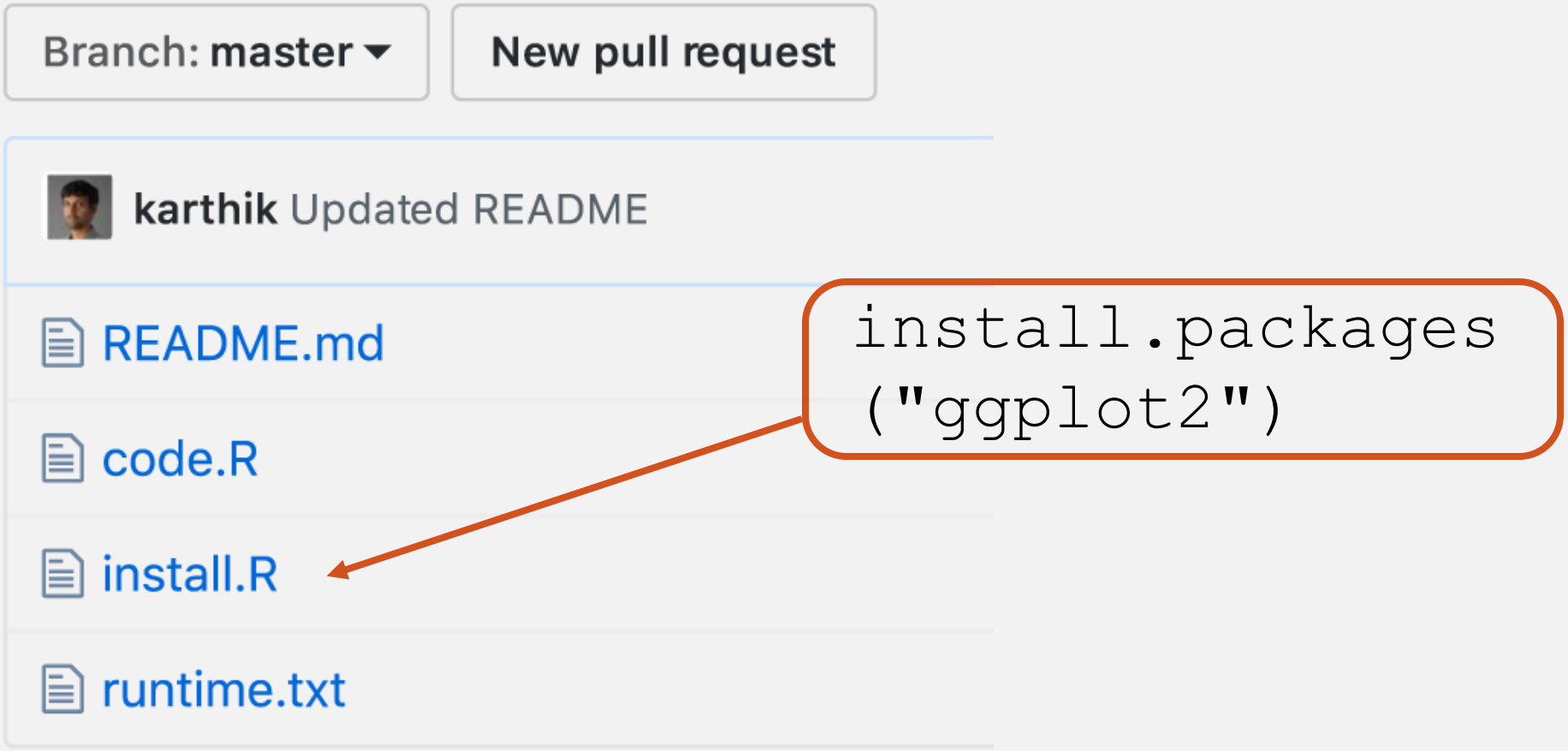
Git + Docker + RStudio



Setting up Binder



Setting up Binder



Binder

Build and launch a repository

GitHub repository name or URL

GitHub repository name or URL | GitHub

Git branch, tag, or commit

Git branch, tag, or commit

Path to a notebook file (optional)

Path to a notebook file (optional) | File

launch

Copy the URL below and share your Binder with others:

```
Fill in the fields to see a URL for sharing your Binder.
```

Copy the text below, then paste into your README to show a binder badge: launch binder

Basic *free*

install.r
runtime.txt
apt.txt

Slow but easy to setup. Recommended for beginners

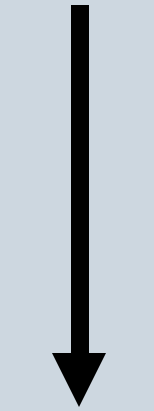
Basic *free*	Premium *free*
install.r ***runtime.txt*** ***apt.txt***	***Dockerfile*** ***install.r***
Slow but easy to setup. Recommended for beginners	Faster launch
	

Basic *free*	Premium *free*	Pro *free*
install.r ***runtime.txt*** ***apt.txt***	***Dockerfile*** ***install.r***	***Dockerfile*** ***DESCRIPTION***
Slow but easy to setup. Recommended for beginners	Faster launch	Best for compendia
		

A fast set up binder

`Dockerfile`

`DESCRIPTION`



`Pull a base image from Rocker (e.g.` **`rocker`**`:binder/latest)`

`rocker-project.org`

The versioned stack

image	description	size
r-ver	Specify R version in docker tag. Builds on `debian:stable`	239.7MB 9 layers
rstudio	Adds rstudio	356.6MB 21 layers
tidyverse	Adds tidyverse & devtools	661.2MB 22 layers
verse	Adds tex & publishing-related packages	1GB 24 layers
geospatial	Adds geospatial libraries	1.4GB 26 layers

3. Workflow

Include a workflow to manage relationships between data output and code.

drake

general purpose workflow manager & pipeline toolkit for reproducibility and high-performance computing.

github.com/ropensci/drake

Drake: *Data Frames in R for Make*

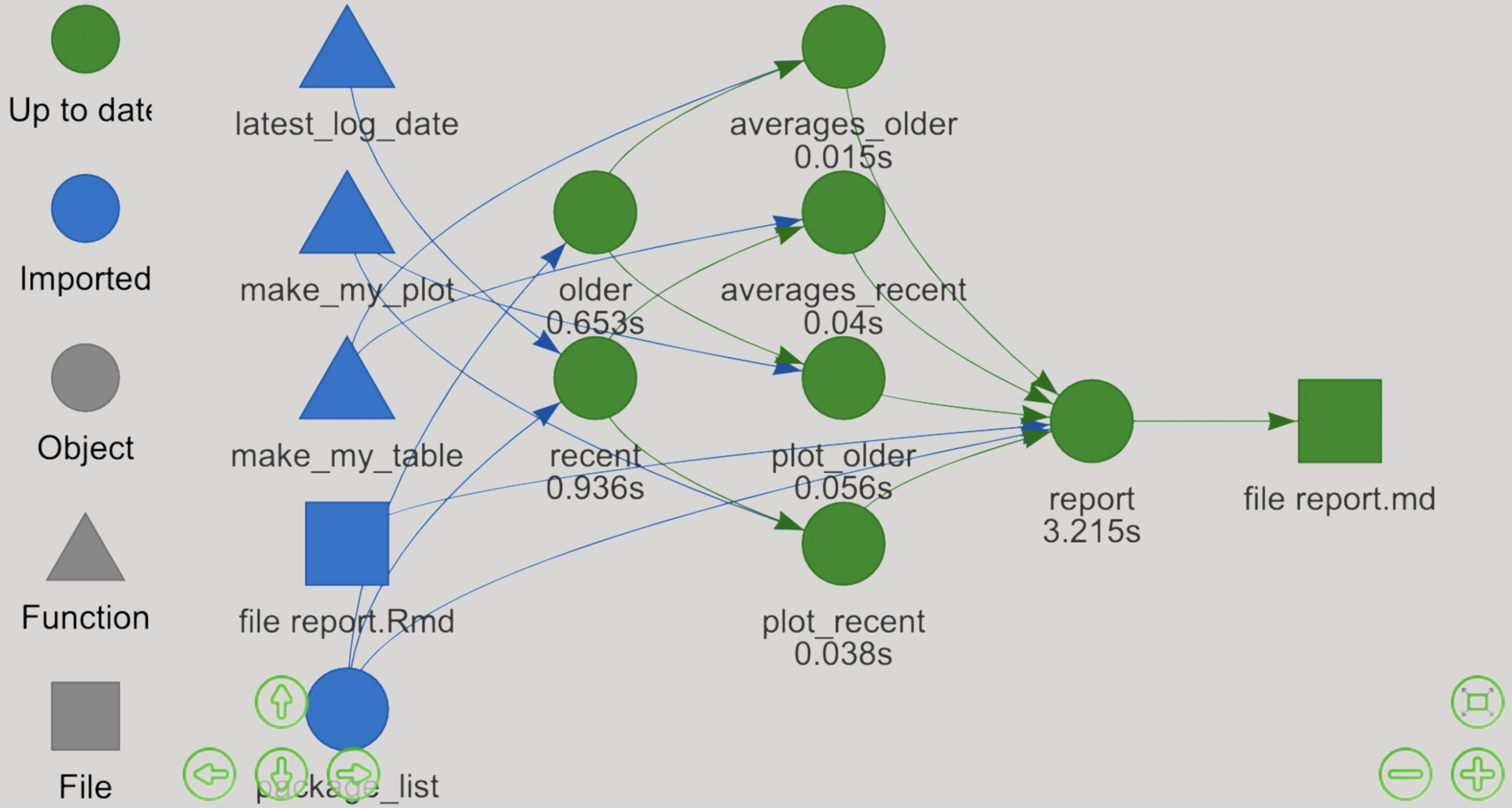
No cumbersome Makefiles

Vast arsenal of **parallel computing** options

Visualize dependency graph and estimate run times

Convenient **organization of output**.

Drake: visualize dependency graph



Take home

Leverage the R package structure and support tools/services as much as possible

Take home

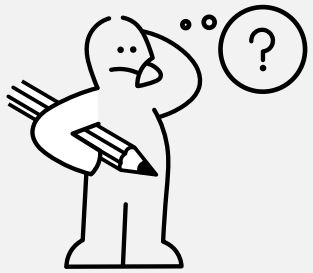
Use modern tools to make your compendia more accessible, but don't forget long-term archives and simpler formats

`github.com/topics/research-compendium`

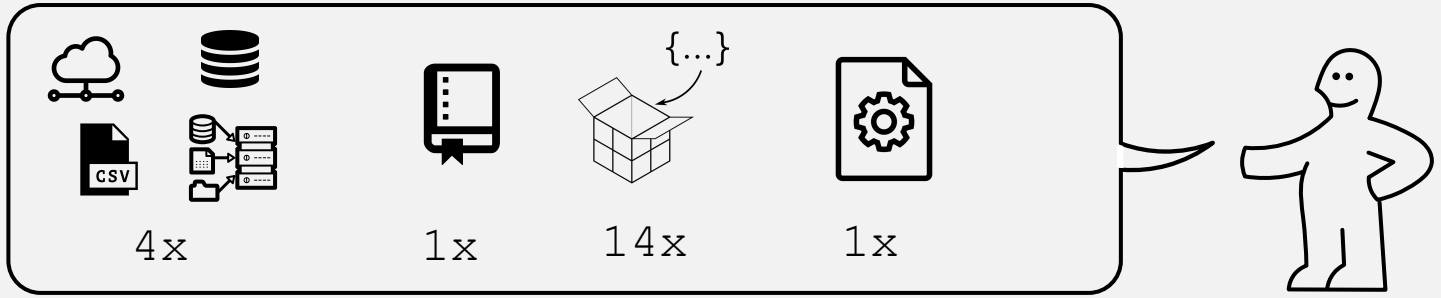
	Near term	*Long term*
data	piggyback, data packages	Zenodo and friends
environment	Binder and friends	Dockerfile
workflow	Drake	Core R tools, Make

KÖMPENDIUM

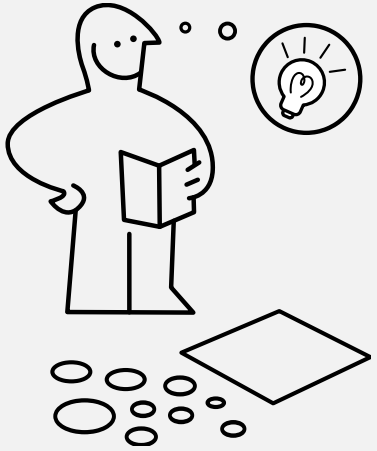
1.



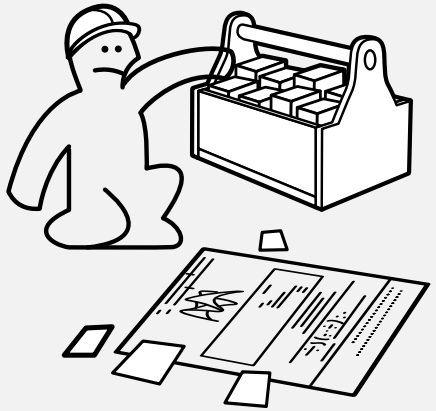
2.



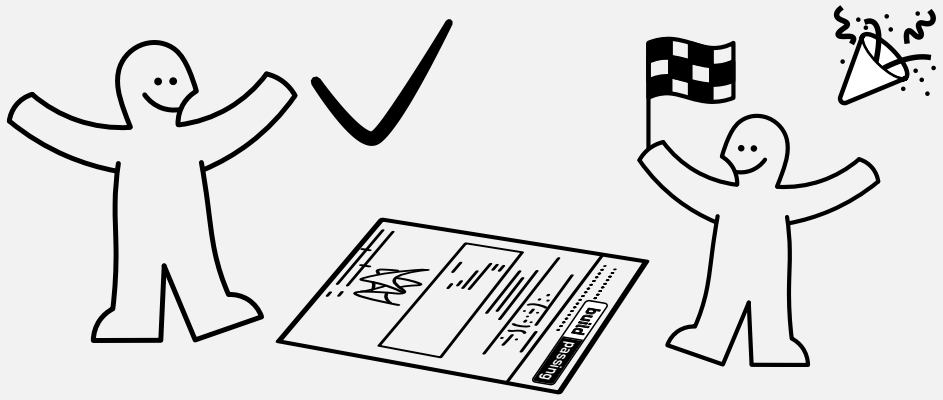
3.



4.



5.



/karthik/rstudio2019

git repo has links to slides and all resources mentioned in the talk