

Binder for Fully Reproducible Research

Dr Andrew J. Stewart

E: drandrewjstewart@gmail.com

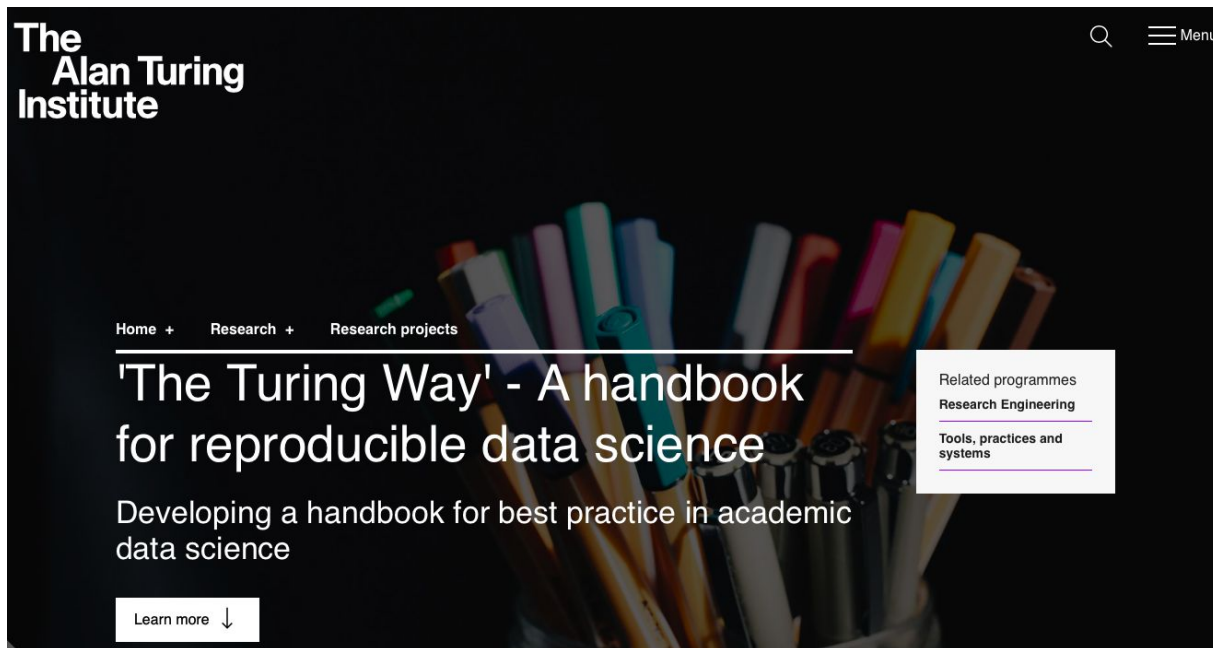
T: [@ajstewart_lang](https://twitter.com/ajstewart_lang)

G: [ajstewartlang](https://github.com/ajstewartlang)



The Turing Way

This workshop draws heavily upon The Turing Way Binder workshop which I attended in 2019 - The Turing Way Handbook is a fantastic resource on reproducible data science and I very much recommend you check it out!



<https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science>

Open and Reproducible Research

Shared Data

We already know this is important for reproducibility.

Shared Code

We already know this is important for reproducibility.

Shared Computational Environment

Why is this important and how do we do it?

Open and Reproducible Research

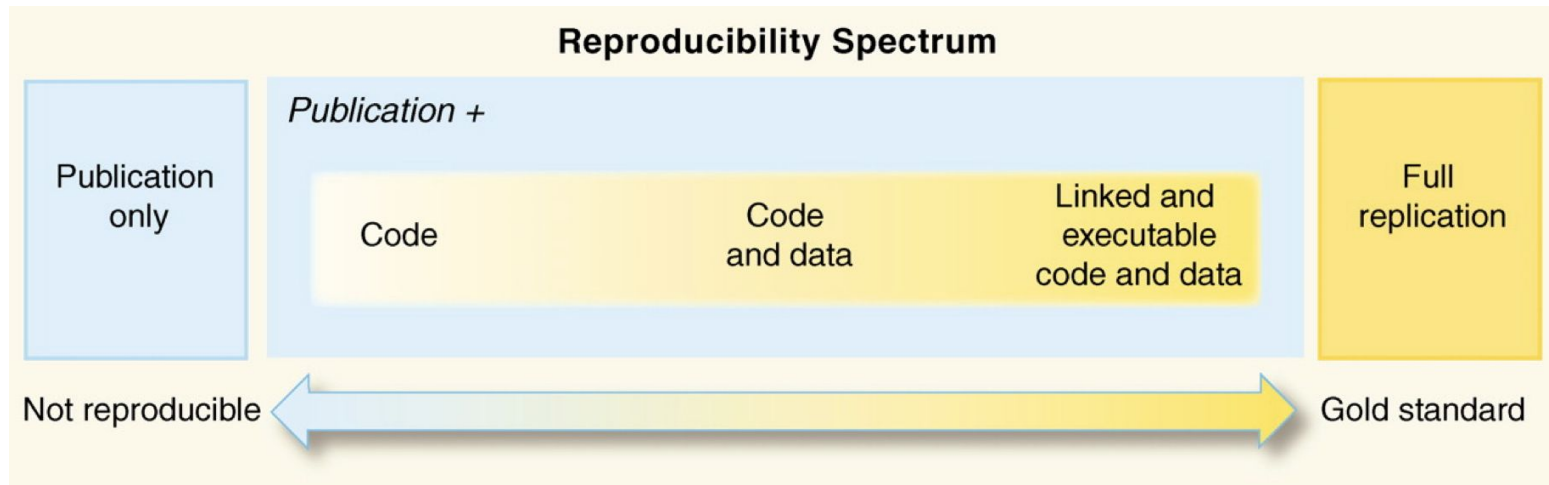
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

✦ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



Why do we need to reproduce the computational environment?

Quite often analysis code 'breaks' - often in one of two ways:

Code that worked previously now doesn't - maybe a function in an R package was updated (e.g., `lsmeans()` became `emmeans()` so old code using the `lsmeans()` function wouldn't now run).

Code that worked previously still works - but produces a slightly different result or now throws a warning where it didn't previously (e.g., convergence/singular fit warnings in `lme4` : : version 1.1-19 vs. version 1.1-20).

Why do we need to reproduce the computational environment?

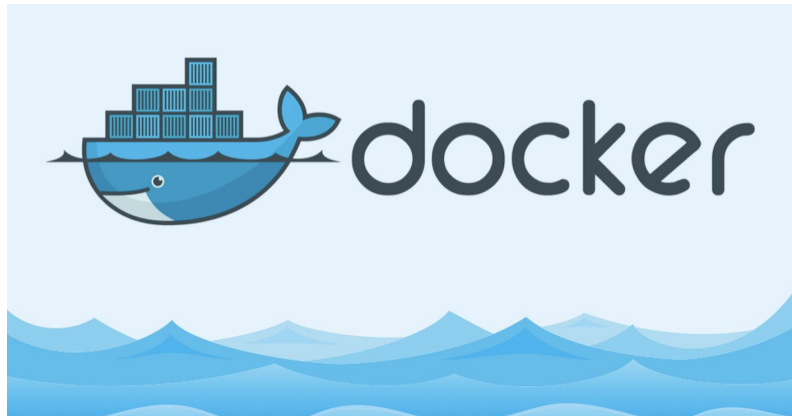
There have even been changes in the way that base R works - in R v3.6 the way in which `sample()` worked differed from how it had worked previously, and in R v4.0, when reading in data `stringsAsFactors = FALSE` (whereas previously `stringsAsFactors = TRUE`).

You need to capture the versions of the different R packages (plus their dependencies) and even the version of R you used in the original analysis.

Introduction to Docker

Docker packages your data, code and all its dependencies in the form called a docker container to ensure that your application works seamlessly in any environment.

When you run a docker container it's like running your analysis on a computer that has the same configuration as our own one at the point in time when you ran the original analysis.



<https://medium.com/the-andela-way/docker-for-beginners-61e8e0ce6a19>

So what's Binder?

Binder is powered by BinderHub, which is an open-source tool that deploys the Binder service in the cloud.

Binder works by pulling a repository that you set up on GitHub into a Docker container using `repo2docker`.

Think of a repository as a folder containing your R code, your data, and a few other small bits and pieces - but it sits in the cloud rather than on your computer.

It pulls the versions of the R packages you specify (on the date you specify) from MRAN.

My GitHub Repo

The screenshot shows the GitHub interface for the repository 'ajstewartiang / demo_binder'. The repository has 1 branch (master) and 0 tags. The file list shows the following files and their commit history:

File	Commit Message	Time Ago
data	first commit	23 minutes ago
.gitignore	first commit	23 minutes ago
README.md	Create README.md	now
binder_demo.Rproj	first commit	23 minutes ago
install.R	Create install.R	21 minutes ago
runtime.txt	Update runtime.txt	14 minutes ago
script.R	Create script.R	20 minutes ago

The README.md file is open, showing the text 'demo_binder'. An arrow points from the text 'This repo contains my data and R script (plus a few other files)' to the 'data' folder in the file list.

Repository details on the right:

- About: No description, website, or topics provided.
- Readme: Readme
- Releases: No releases published. [Create a new release](#)
- Packages: No packages published. [Publish your first package](#)
- Languages: R 100.0%

This repo contains my data and R script (plus a few other files).

When I link my GitHub repository to Binder and launch it I then get the following in my web browser.

This is RStudio running the cloud using my code, my data and the appropriate versions of the packages that I was using when I did the analysis originally!

The screenshot shows the RStudio web interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The top right shows the user 'jovyan' and a power button. Below the menu is a toolbar with icons for file operations and a 'Go to file/function' search bar. The main window is divided into three panes: Console, Terminal, and Jobs. The Console pane shows the R startup message and a prompt '>|'. The Environment pane on the right shows 'Global Environment' and 'Environment is empty'. The Files pane at the bottom shows a list of files and folders.

R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 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.

Natural language support but running in an English locale

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

Global Environment

Environment is empty

Files Plots Packages Help Viewer

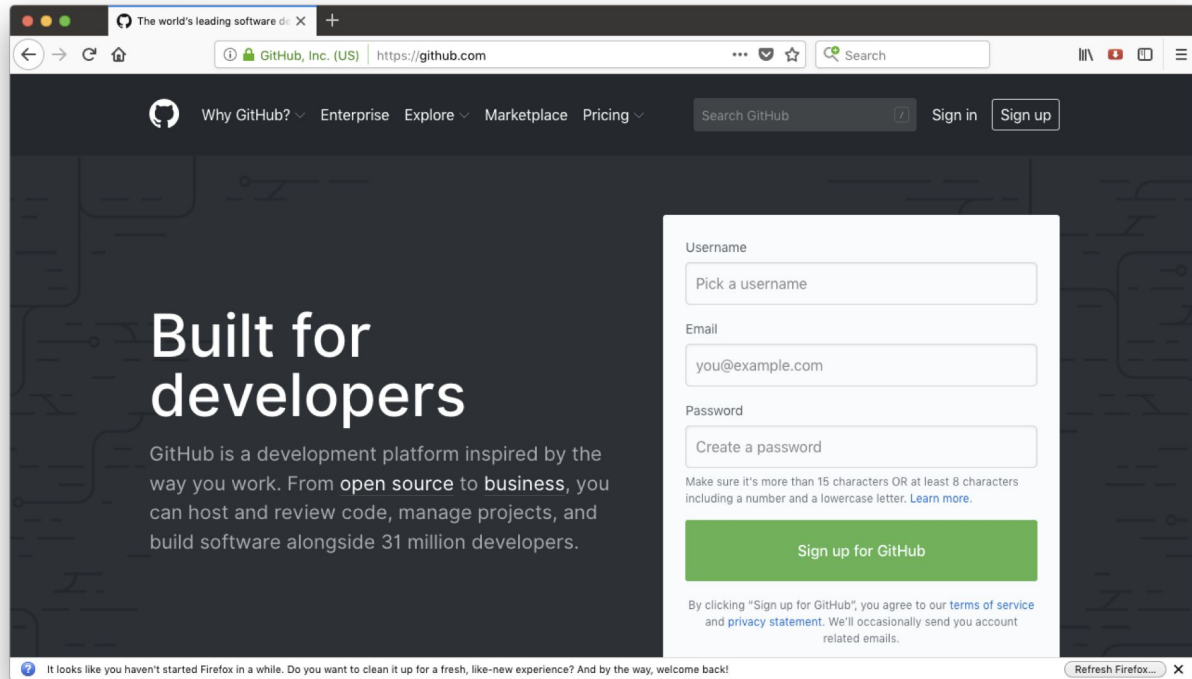
New Folder Upload Delete Rename More

Home

	Name	Size	Modified
<input type="checkbox"/>	.gitignore	40 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	binder_demo.Rproj	205 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	data		
<input type="checkbox"/>	install.R	57 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	README.md	14 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	runtime.txt	18 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	script.R	2 KB	Jul 20, 2020, 11:57 AM

https://mybinder.org/v2/gh/ajstewartlang/demo_binder/master?urlpath=rstudio

Step 1 – Set up a GitHub account



The screenshot shows the GitHub homepage in a web browser. The browser's address bar displays "https://github.com". The page header includes the GitHub logo, navigation links like "Why GitHub?", "Enterprise", "Explore", "Marketplace", and "Pricing", a search bar, and "Sign in" and "Sign up" buttons. The main content area features the text "Built for developers" and a description of GitHub as a development platform. A sign-up form is overlaid on the right side of the page, containing fields for "Username", "Email", and "Password". The "Username" field has the placeholder text "Pick a username". The "Email" field contains "you@example.com". The "Password" field has the placeholder text "Create a password". Below the password field, there is a note: "Make sure it's more than 15 characters OR at least 8 characters including a number and a lowercase letter. [Learn more.](#)". A green button labeled "Sign up for GitHub" is positioned below the password field. At the bottom of the form, a disclaimer states: "By clicking 'Sign up for GitHub', you agree to our [terms of service](#) and [privacy statement](#). We'll occasionally send you account related emails." The browser's status bar at the bottom shows a message: "It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!" and a "Refresh Firefox..." button.

Username

Pick a username

Email

you@example.com

Password

Create a password

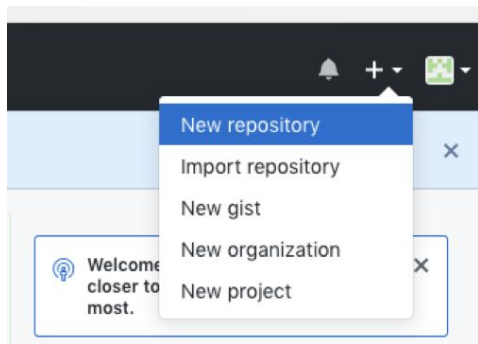
Make sure it's more than 15 characters OR at least 8 characters including a number and a lowercase letter. [Learn more.](#)

Sign up for GitHub

By clicking "Sign up for GitHub", you agree to our [terms of service](#) and [privacy statement](#). We'll occasionally send you account related emails.

Go to github.com to set up an account.

Step 2 – Create a new repository



Make it 'Public' and initialise it with a README.


Create a new repository

A repository contains all project files, including the revision history.

Owner

 andrewstewarttest

Repository name *

first_binder 

Great repository names are short and memorable. Need inspiration? How about **probable-funicular**?

Description (optional)

☒  **Public**

Anyone can see this repository. You choose who can commit.

☐  **Private**

You choose who can see and commit to this repository.

☒ **Initialize this repository with a README**

This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.

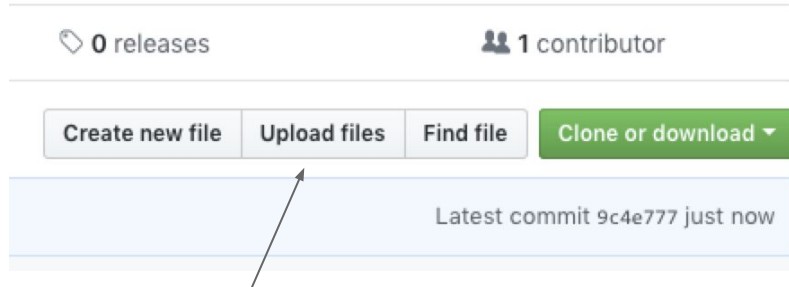
Add .gitignore: **None**

Add a license: **None**



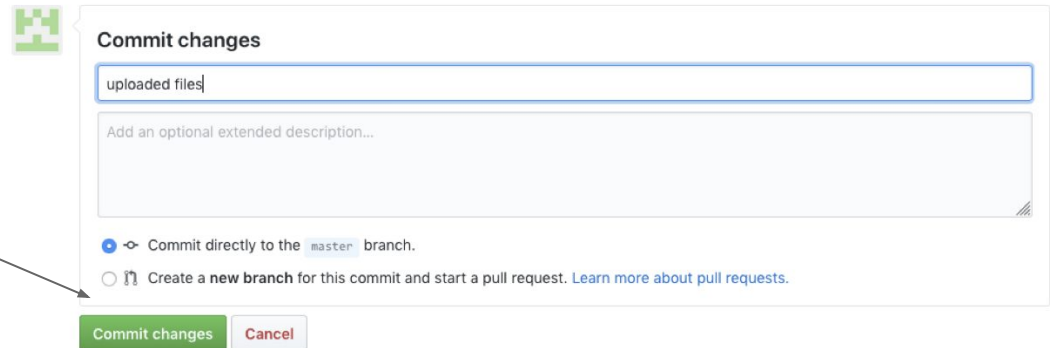
Create repository

Step 3 – Upload your R script and data and make your first “Commit”



Click here to upload.

Click here to commit.

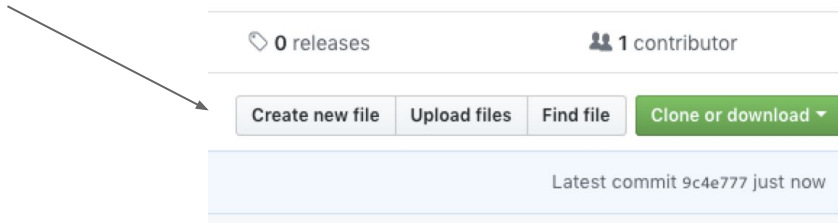


Step 4 – create the files “runtime.txt” and “install.R”

We need two other files at this point - one is called “runtime.txt” and contains the date and version of R and its associated packages that you want to simulate.

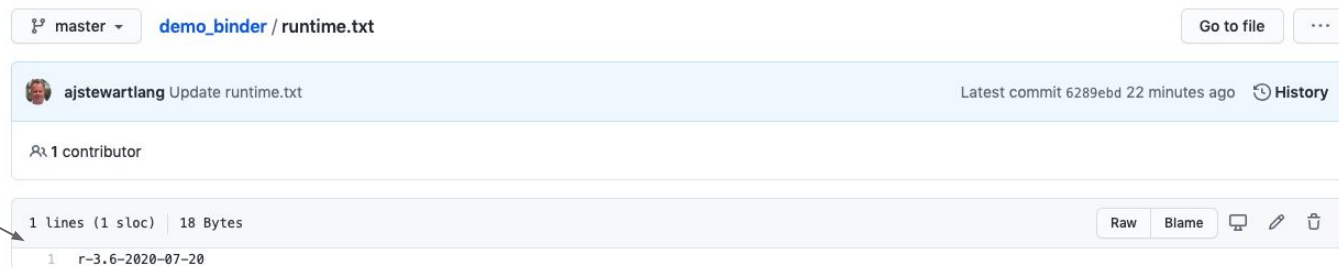
The other is called “install.R” and contains the list of R packages that need to be installed in order for your script to run.

To create a new file select “Create new file”



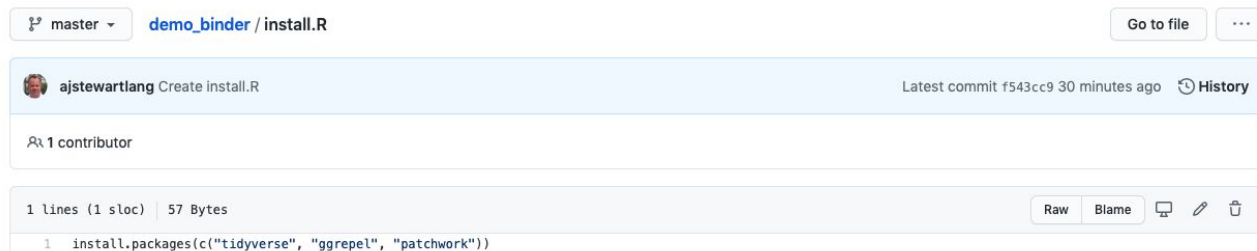
Step 4 – create the files “runtime.txt” and “install.R”

In the runtime.txt file
type the R version and
date of the packages on
MRAN that you want in
the format
r-version-YYYY-MM-DD



This screenshot shows a GitHub commit for the file `runtime.txt` in the `demo_binder` repository. The commit was made by `ajstewartlang` and is titled "Update runtime.txt". The commit message indicates it was updated 22 minutes ago. The file content is shown as a single line: `r-3.6-2020-07-20`. The file is 18 bytes long and contains 1 line of code.

List your packages
like this in the
install.R file



This screenshot shows a GitHub commit for the file `install.R` in the `demo_binder` repository. The commit was made by `ajstewartlang` and is titled "Create install.R". The commit message indicates it was created 30 minutes ago. The file content is shown as a single line: `install.packages(c("tidyverse", "ggrepel", "patchwork"))`. The file is 57 bytes long and contains 1 line of code.

Don't forget to click
“Commit” after you’ve
created each file!

Step 5 – Now we need to link our repo to Binder (mybinder.org)

New to Binder? Get started with a Zero-to-Binder tutorial in [Julia](#), [Python](#) or [R](#).

Build and launch a repository

GitHub repository name or URL

GitHub

Git branch, tag, or commit

URL to open (optional)

URL

launch

Copy the URL below and share your Binder with others:

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

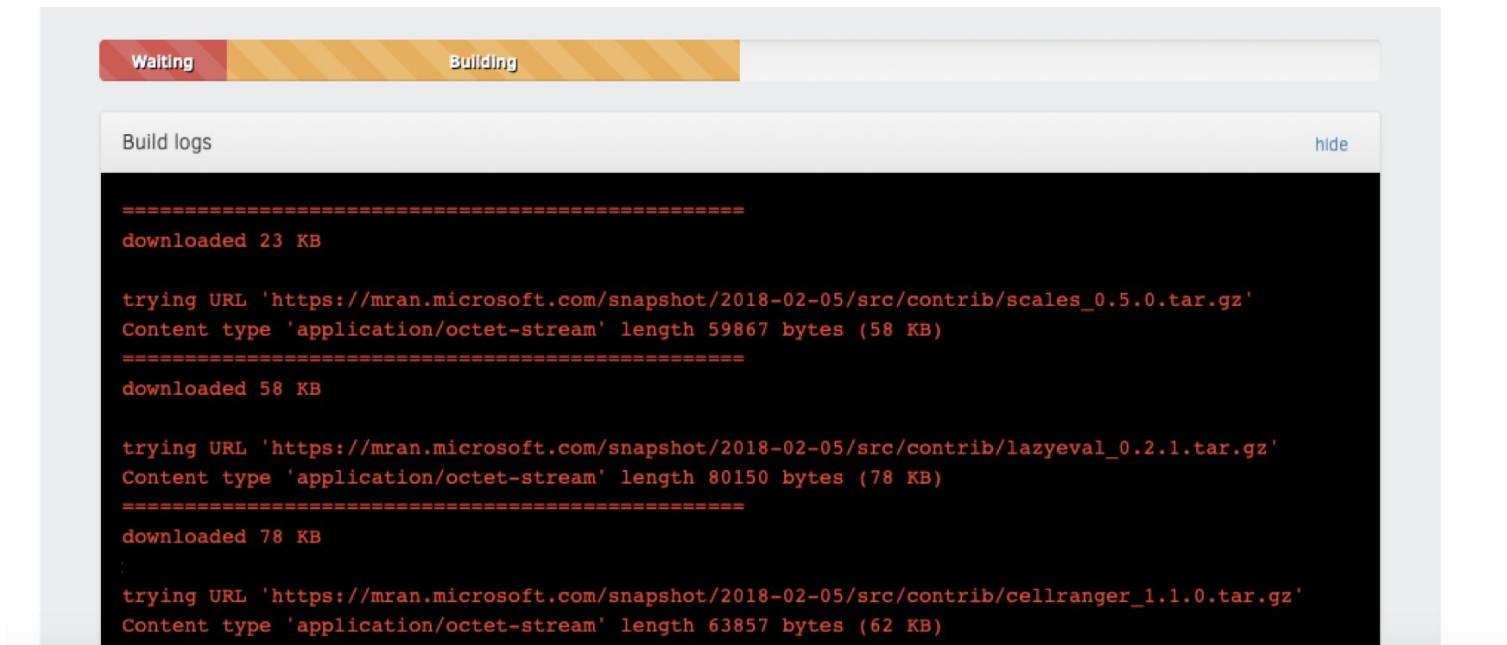
Paste the address of your GitHub repo you want to Binderise.

Type “RStudio” and select “URL”

Click on “Launch”

And this is the link to share with others.

When you first click on “Launch”



You can check the progress of the build by clicking on the “Build logs” bar.

If Binder can find an image that you've built previously, it will simply launch that. If you've made changes to your GitHub repo, it will rebuild the Docker image and create a new Binder. Either way, once Binder launches you get the following in your browser (even on mobile devices so you can even R away on your phone)...

The screenshot displays the RStudio web interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The user is logged in as 'jovyan'. The left sidebar has tabs for Console, Terminal, and Jobs. The Console shows the R version 3.6.3 (2020-02-29) and standard startup messages. The right sidebar has tabs for Environment, History, and Connections. The Environment tab shows 'Global Environment' and 'Environment is empty'. Below it, the Files tab shows a file explorer with a table of files.

	Name	Size	Modified
<input type="checkbox"/>	.gitignore	40 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	binder_demo.Rproj	205 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	data		
<input type="checkbox"/>	install.R	57 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	README.md	14 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	runtime.txt	18 B	Jul 20, 2020, 11:57 AM
<input type="checkbox"/>	script.R	2 KB	Jul 20, 2020, 11:57 AM

A few other things

Installing the entire Tidyverse in a Binder can take a long time - better to install only the packages you use (e.g., `ggplot2`, `dplyr`, `readr` etc.) - this will also ensure the individual packages are consistent with the date in your `runtime.txt` file.

Even with just a couple of packages it can take ~15 minutes or so for your Binder to be built. While your Binder builds you can close your computer as the build will continue in the cloud.

To change the version of R that Binder builds (to 3.5 say) change the R version in `runtime.txt` like this `"r-3.5-2020-07-20"`

For Ultimate Reproducibility

Make sure you have updated all your packages before you run your script locally.

Build your Binder and specify the day you ran your analysis in the runtime.txt file - and add a version of R if you don't want it to go with the default.

Patience while your Binder builds.



Your turn

Start from Slide 11, take an R script you've already written and Binderise it!