The Alan Turing Institute

binder

Build a BinderHub for hosting Reproducible Software in the Cloud

Sarah Gibson

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Turn code… into an environment!

binder-examples / requirements

Watch 3 | Star 20 | Fork 65

Code | Issues 0 | Pull requests 1 | Projects 0 | Wiki | Security | Insights

Simple requirements.txt based example

binder | binder-ready

32 commits | 1 branch | 2 releases | 4 contributors

Branch: master | New pull request | Create new file | Upload files | Find File

choldgraf adding pandas — Latest com

LICENSE	Create LICENSE
README.md	remove beta from link
index.ipynb	first move
requirements.txt	adding pandas
runtime.txt	Pin Python version to 3.5

jupyter index (unsaved changes)

File | Edit | View | Insert | Cell | Kernel | Widgets | Help | Not Trusted | Python 3

Run | Markdown

Welcome to an example Binder

This notebook uses `seaborn`, which we have because we included it in our `requirements.txt` file

Setup our plotting

In [1]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
```

Setup our imports

In [2]:
```
from numpy import random
from scipy.ndimage.filters import gaussian_filter
```

Make some plots!

In [3]:
```
x = random.randn(10,500)
x = gaussian_filter(x, [0, 10])
```

https://doi.org/10.5281/zenodo.3404774

This Workshop

– What it is:

– What it’s not:

– What we’ll do:

https://doi.org/10.5281/zenodo.3404774

This Workshop

– What it is: Challenging!

– What it’s not:

– What we’ll do:

https://doi.org/10.5281/zenodo.3404774

This Workshop

- What it is: Challenging!
- What it’s not: A cloud/Azure workshop
- What we’ll do:

https://docs.microsoft.com/en-gb/learn/azure/

https://doi.org/10.5281/zenodo.3404774

This Workshop

- What it is: Challenging!
- What it’s not: A cloud/Azure workshop
- What we’ll do: Build a BinderHub!

https://docs.microsoft.com/en-gb/learn/azure/

https://doi.org/10.5281/zenodo.3404774

Housekeeping

- Microsoft Azure: Please leave your email in #binderhub-workshop channel on RSE Slack
- Docker Hub: https://hub.docker.com/signup
- Code of Conduct: Be kind! https://rse.ac.uk/conf2019/code-of-conduct/
- HackMD: bit.ly/RSEConBinderHub
- post-its 🚦

Who?

Sarah
Research Data Scientist
Operator of mybinder.org

Tania
Microsoft Cloud
Developer Advocate

Anna
Research
Software Engineer

https://doi.org/10.5281/zenodo.3404774

(Rough) Agenda

Time	Activity
09:00 – 09:30	👋 Introduction
09:30 – 10:30	🚀 Deploy Kubernetes cluster*
10:30 – 11:00	☕ Coffee break
11:00 – 12:30	👩‍💻 Install BinderHub

*Don’t worry if you don’t know what this is yet, I’ll explain!

https://doi.org/10.5281/zenodo.3404774

The Turing Way

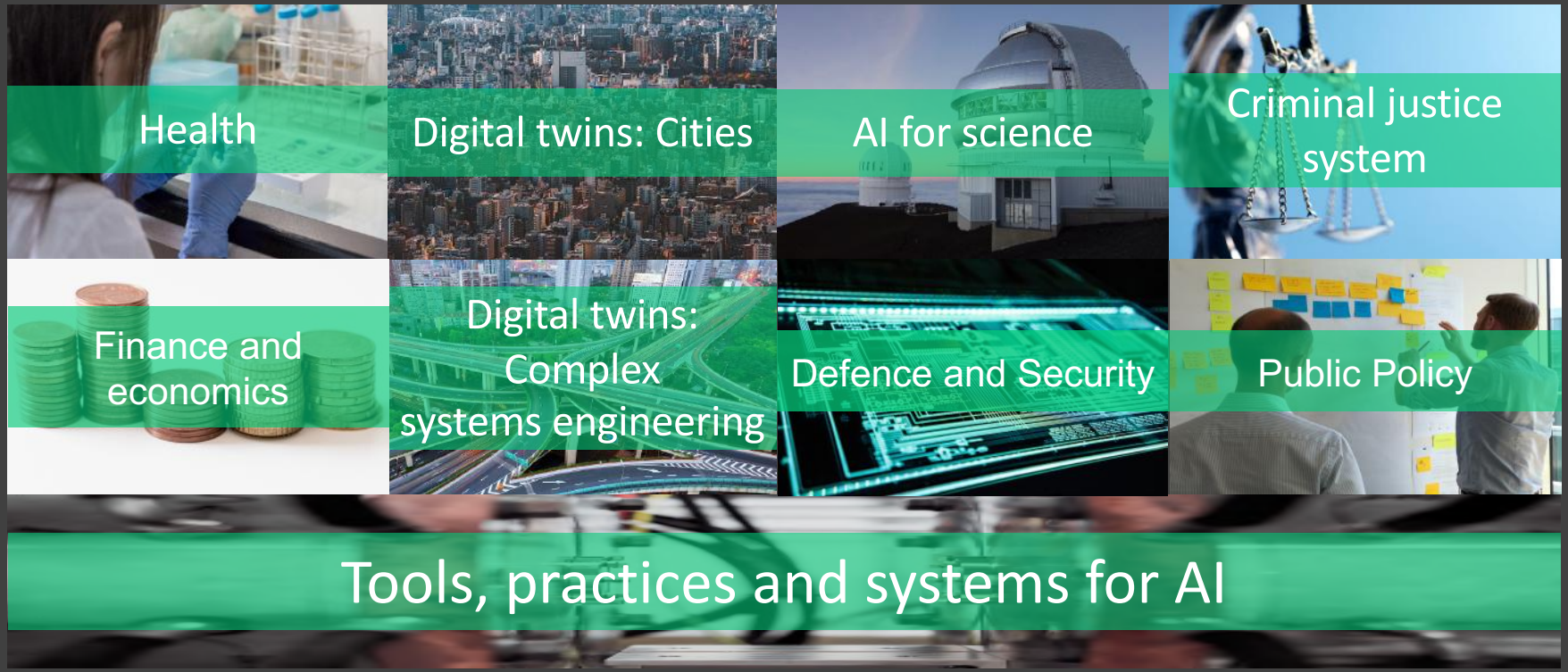
A Handbook for Reproducible Data Science

Making reproducibility too easy not to do!

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Where does The Turing Way fit in?



https://doi.org/10.5281/zenodo.3404774

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Kirstie Whitaker's talk at PyData LDN: https://youtu.be/IG3PcZ6EhiU

https://the-turing-way.netlify.com/reproducibility/03/definitions.html

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

		Data	
		Same	Different
Analysis	Same	Repeatable Reproducible	Replicable
	Different	Robust	Generalisable

Kirstie Whitaker's talk at PyData LDN: https://youtu.be/IG3PcZ6EhiU

https://the-turing-way.netlify.com/reproducibility/03/definitions.html

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

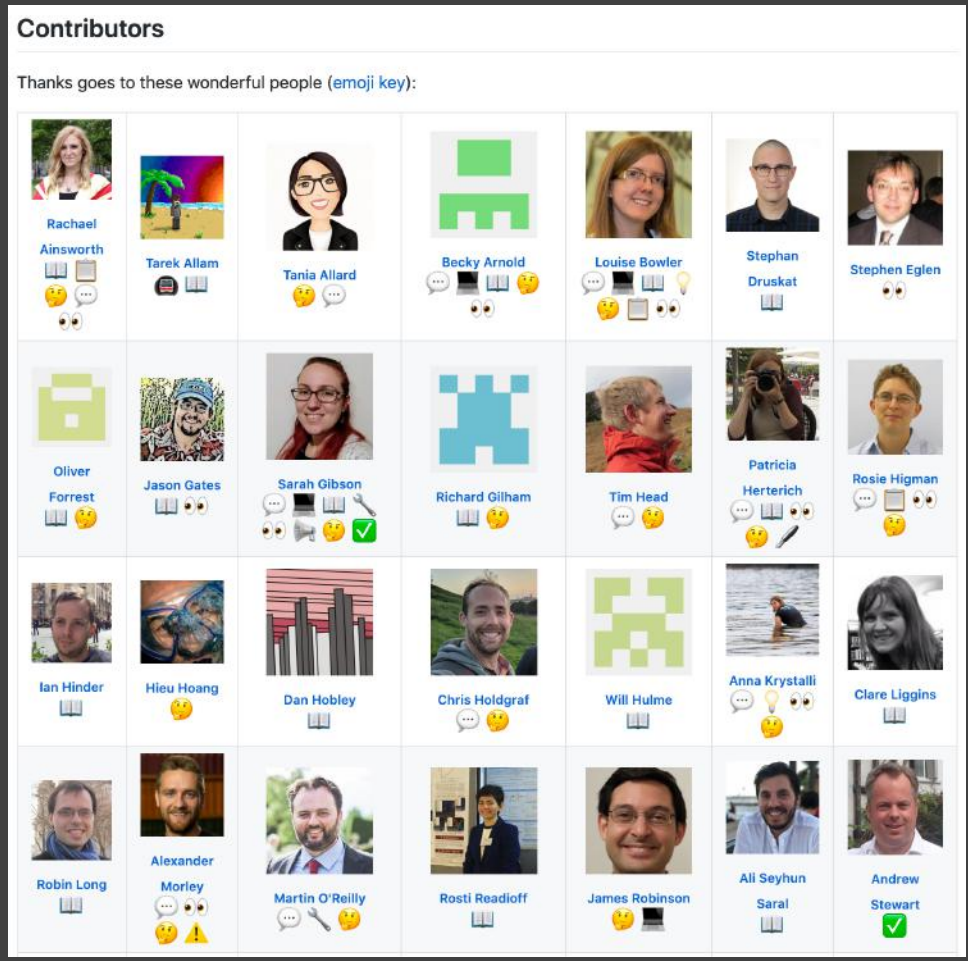
https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Built by a team… AND YOU!

– Please contribute!

github.com/alan-turing-institute/the-turing-way



https://doi.org/10.5281/zenodo.3404774

Barriers to reproducible research

Is not considered for promotion

Held to higher standards than others

Publication bias towards novel findings

Requires additional skills

Plead the 5th

Support additional users

Takes time

Kirstie Whitaker’s talk at PyData LDN: https://youtu.be/IG3PcZ6EhiU

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

Market Research

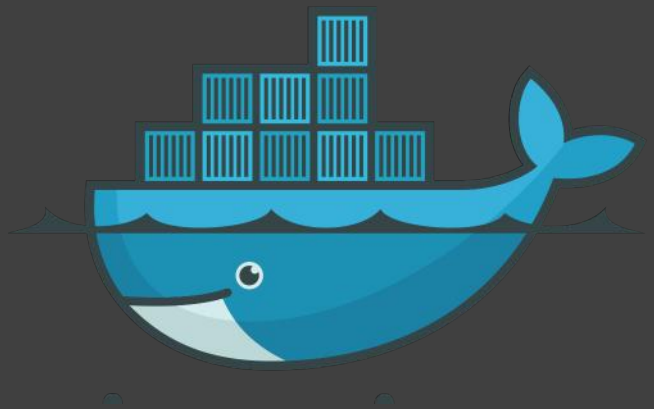
https://doi.org/10.5281/zenodo.3404774

Have you ever heard…?

“Oh, it worked on my computer?”

Have you ever heard…?

“Oh, it worked yesterday?”

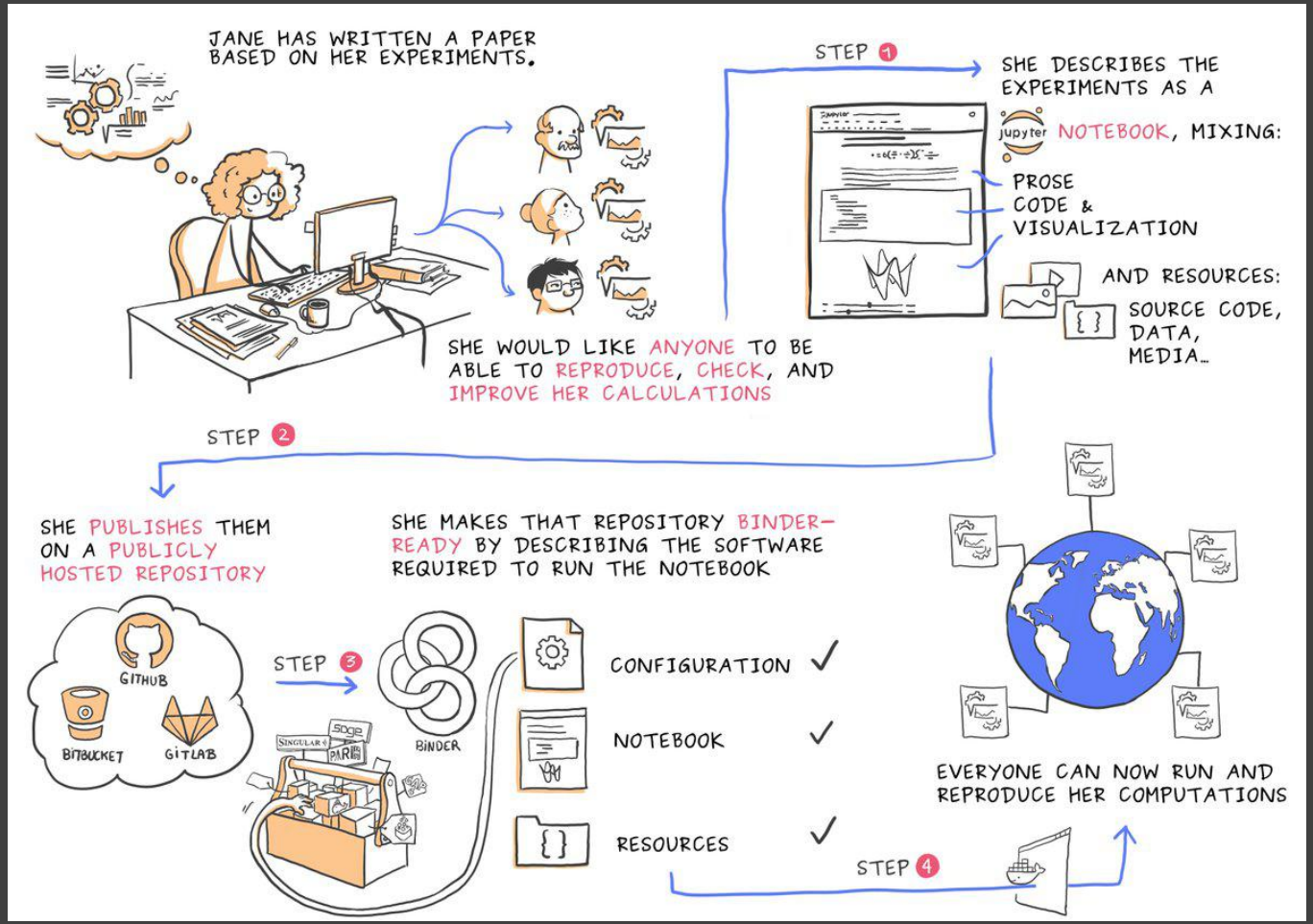


“Oh, it worked on my computer?”

+ CI

“Oh, it worked yesterday?”

https://doi.org/10.5281/zenodo.3404774



binder

mybinder.org

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

binder-examples / requirements

Watch 3 | Star 15 | Fork 60

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Security | Insights

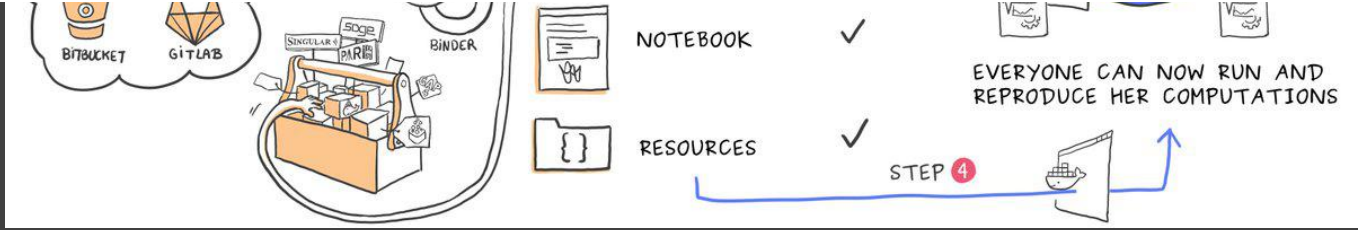
Branch: master | requirements / requirements.txt | Find file | Copy path

choldgraf Update requirements.txt 21a328d on 21 Jun

2 contributors

5 lines (3 sloc) | 46 Bytes | Raw | Blame | History

```
numpy==1.16.*
matplotlib==3.*
seaborn==0.8.1
```



Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

binder-examples / conda

Watch 2 | Star 8 | Fork 156

Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki | Security | Insights

Branch: master | conda / environment.yml | Find file | Copy path

betatim Update environment.yml 89dd429 on 11 Dec 2018

4 contributors

14 lines (13 sloc) | 161 Bytes | Raw | Blame | History

```yaml
name: example-environment
channels:
  - conda-forge
dependencies:
  - numpy
  - psutil
  - toolz
  - matplotlib
  - dill
  - pandas
  - partd
  - bokeh
  - dask
```

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774

binder-examples / binder-r-description

Watch 1 | Star 2 | Fork 1

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Security | Insights

Branch: master | binder-r-description / DESCRIPTION | Find file | Copy path

gedankenstuecke first commit 70f8b8e on 18 Sep 2018

1 contributor

8 lines (7 sloc) | 282 Bytes | Raw | Blame | History

```
Package: binderdescription
Version: 0.1
Date: 2018-09-18
Title: Binder R DESCRIPTION support
Description: Test that automatically building R packages works
Author: Bastian Greshake Tzovaras <bgreshake@googlemail.com>
Maintainer: Bastian Greshake Tzovaras <bgreshake@googlemail.com>
```

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

binder-examples / r

Watch 13 | Star 71 | Fork 123

Code | Issues 3 | Pull requests 1 | Projects 0 | Wiki | Security | Insights

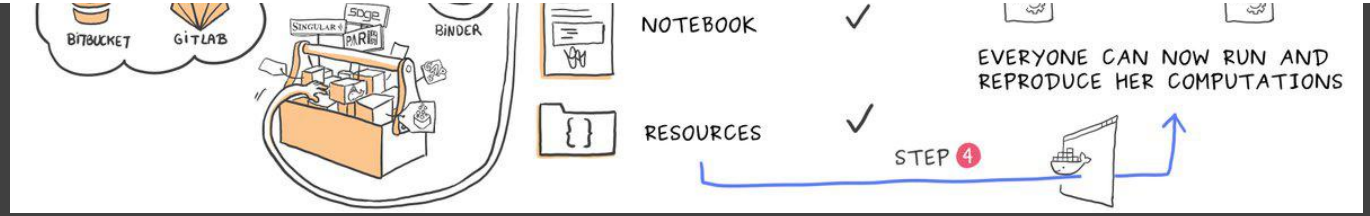
Branch: master | r / install.R | Find file | Copy path

betatim Add example Shiny app 8c01f0d on 31 May 2018

4 contributors

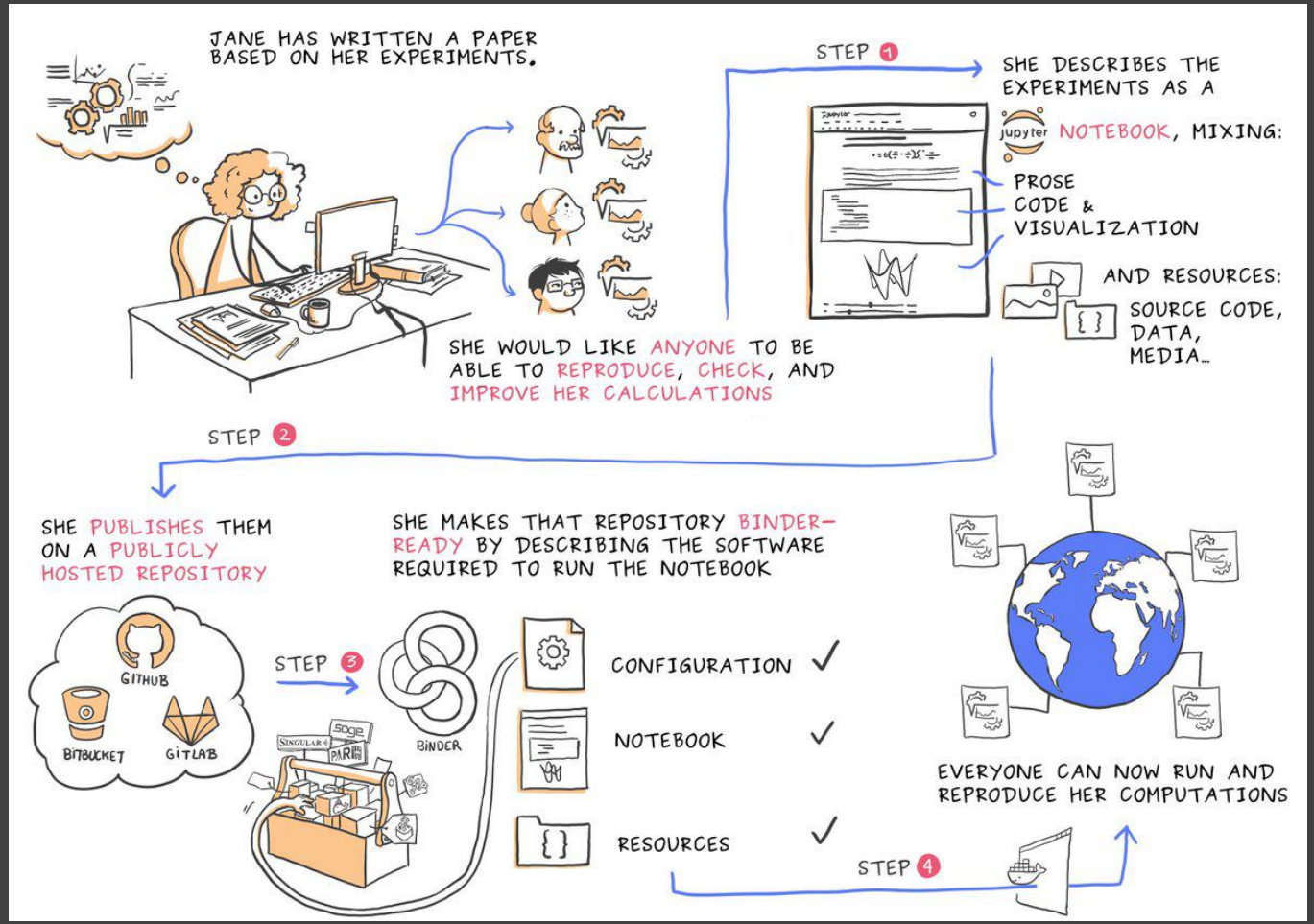
6 lines (5 sloc) | 148 Bytes | Raw | Blame | History

```
install.packages("tidyverse")
install.packages("rmarkdown")
install.packages("httr")
install.packages("shinydashboard")
install.packages('leaflet')
```



Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774



binder

mybinder.org

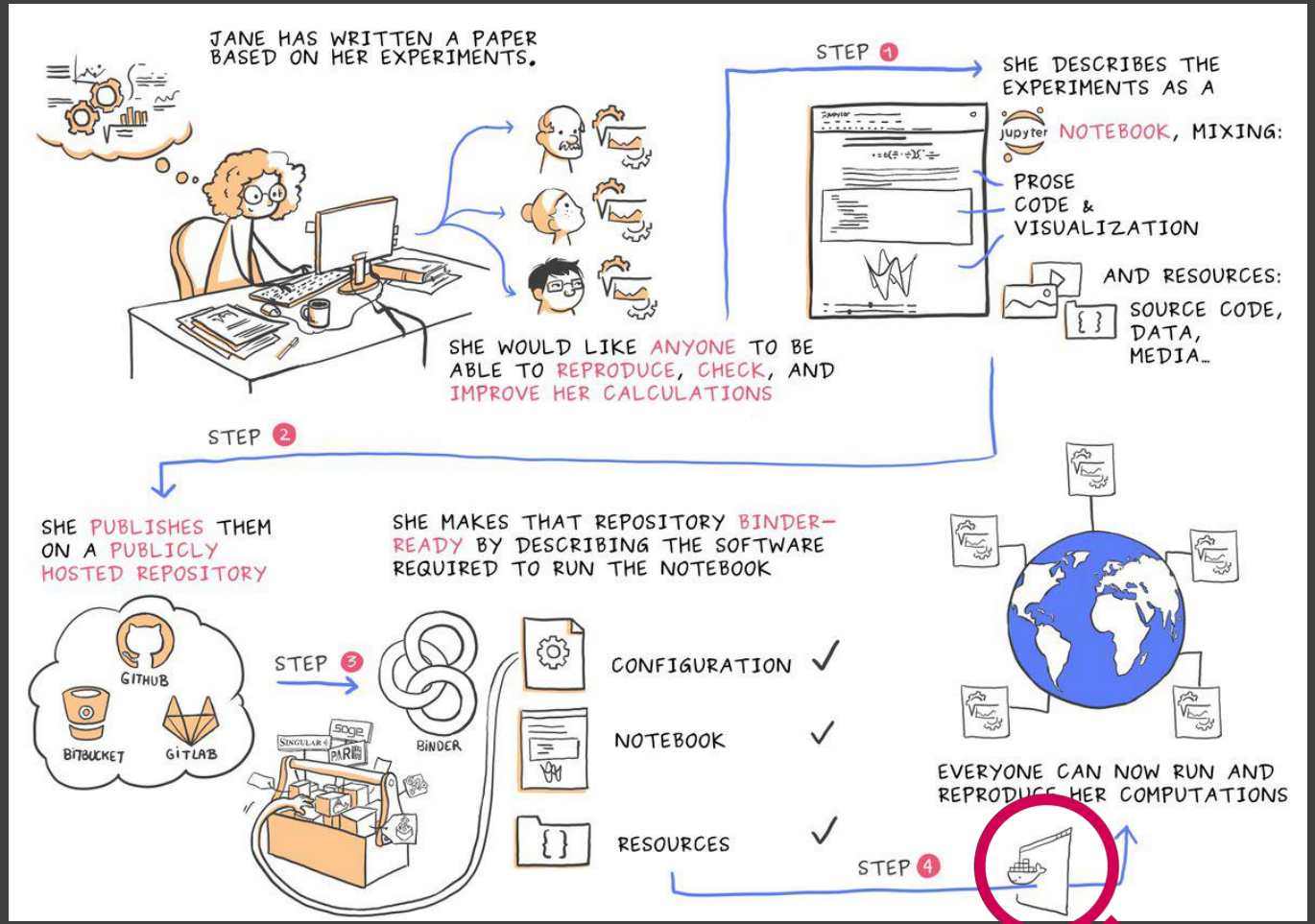
Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

binder

mybinder.org



Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19

What’s the difference?

mybinder.org

- Free to use service for everyone
- Must be public code
- Limited computational resources
- No GPU access

Private BinderHub

- Service can be limited to teams or institutions
- Can be public or private code
- Set your own computational limits
- Deploy onto any type of machine you need

The Vocab

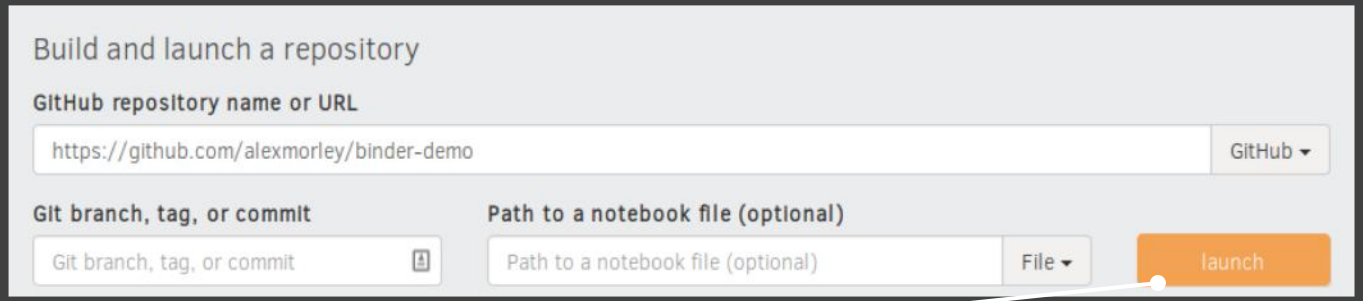
- Binder → user interface/experience
- BinderHub → computational infrastructure
- mybinder.org → public BinderHub for everyone

https://doi.org/10.5281/zenodo.3404774

~~Magic!~~ Technology

https://doi.org/10.5281/zenodo.3404774

BinderHub

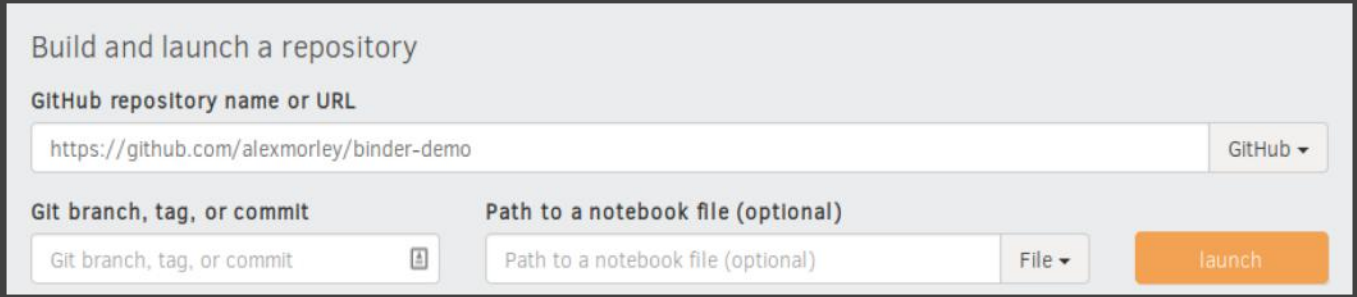


Clone GitHub Repo

1

https://doi.org/10.5281/zenodo.3404774

BinderHub



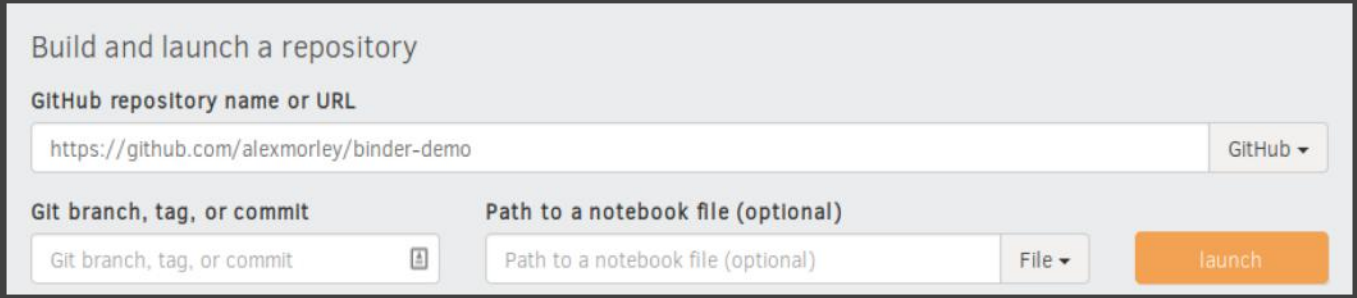
1 Clone GitHub Repo

2 Build image according to instructions contained within the repo

repo2docker

https://doi.org/10.5281/zenodo.3404774

BinderHub



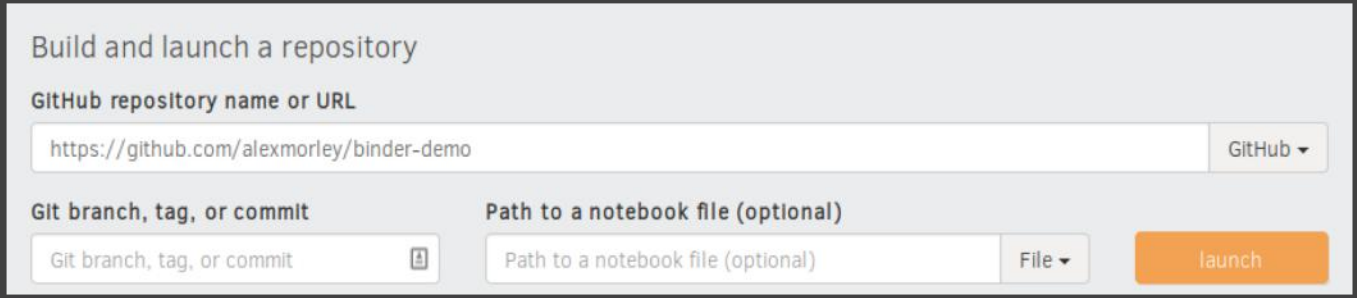
1 Clone GitHub Repo

2 Build image according to instructions contained within the repo

3 Execute image

https://doi.org/10.5281/zenodo.3404774

BinderHub



1 Clone GitHub Repo

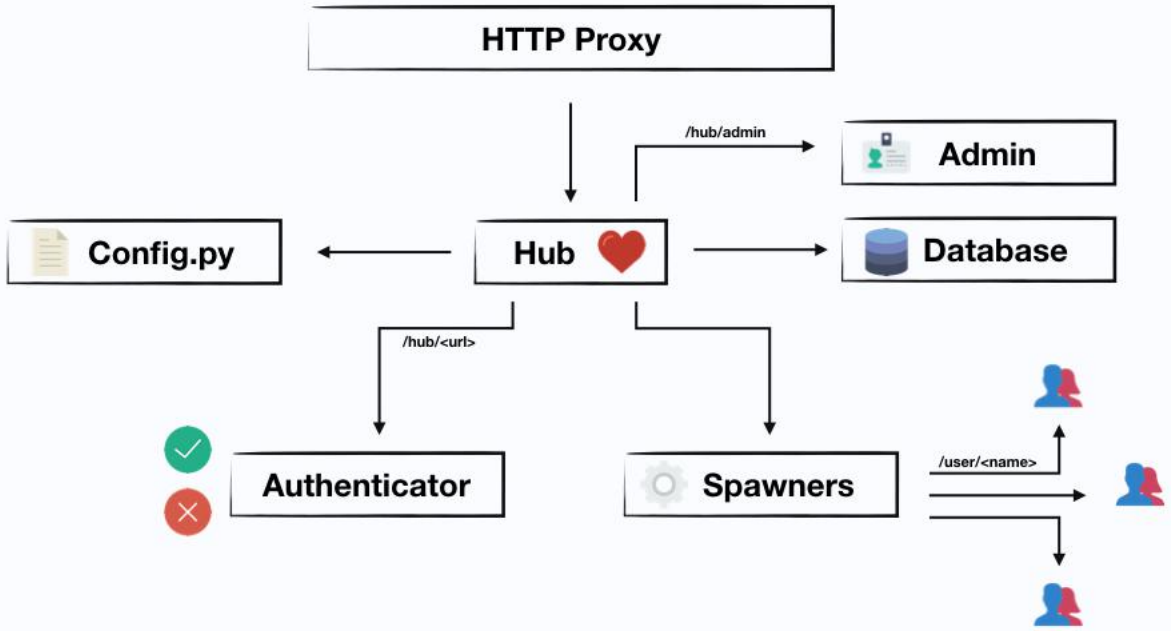
2 Build image according to instructions contained within the repo

3 Execute image

4 Allocate computational resources

https://doi.org/10.5281/zenodo.3404774

What is a JupyterHub?



All icons were obtained from Flaticon (https://www.flaticon.com/packs/essential-collection)

GitHub

JupyterHub is a way to help your humans use your computers. With notebooks!

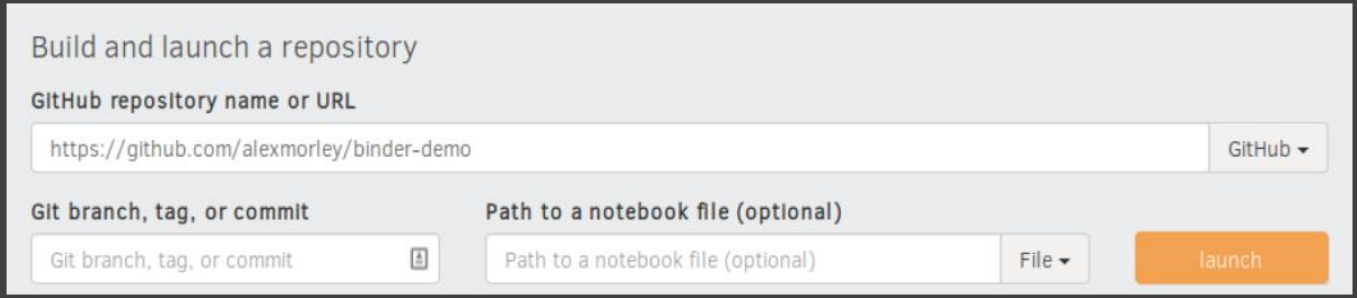
hin the

resources

Execute image

https://doi.org/10.5281/zenodo.3404774

BinderHub



1 Clone GitHub Repo

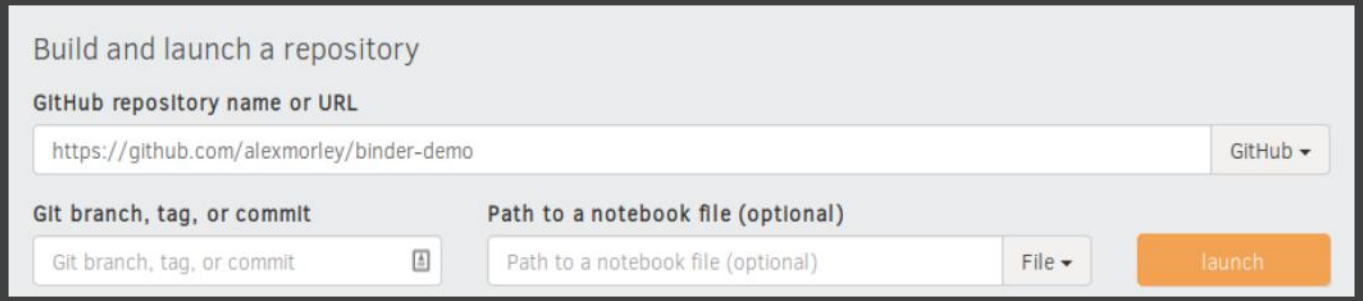
2 Build image according to instructions contained within the repo

3 Execute image

4 Allocate computational resources

https://doi.org/10.5281/zenodo.3404774

BinderHub



1 Clone GitHub Repo

2 Build image according to instructions contained within the repo

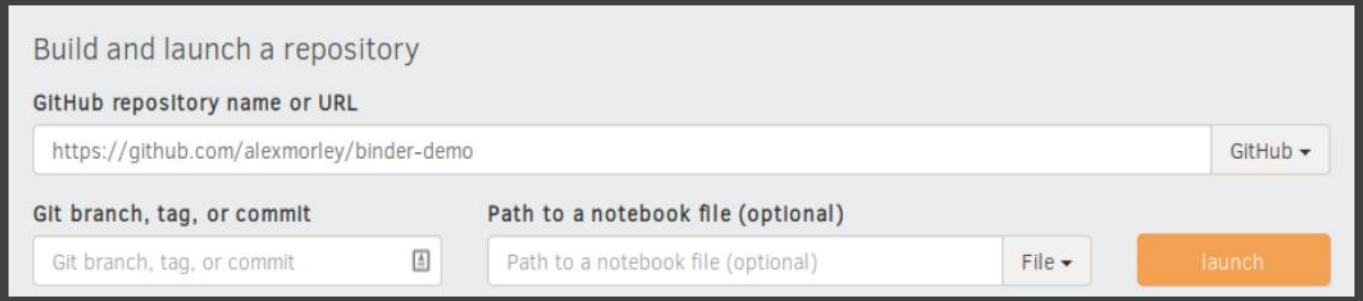
3 Execute image

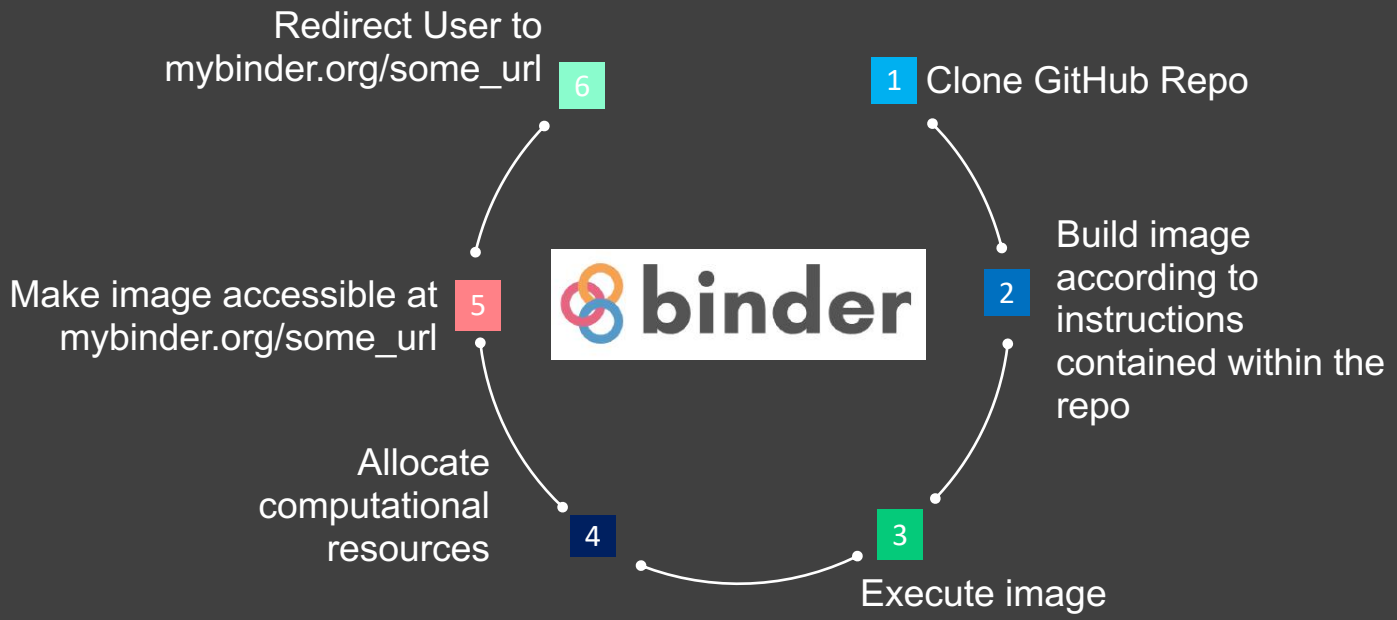
4 Allocate computational resources

5 Make image accessible at mybinder.org/some_url

https://doi.org/10.5281/zenodo.3404774

BinderHub

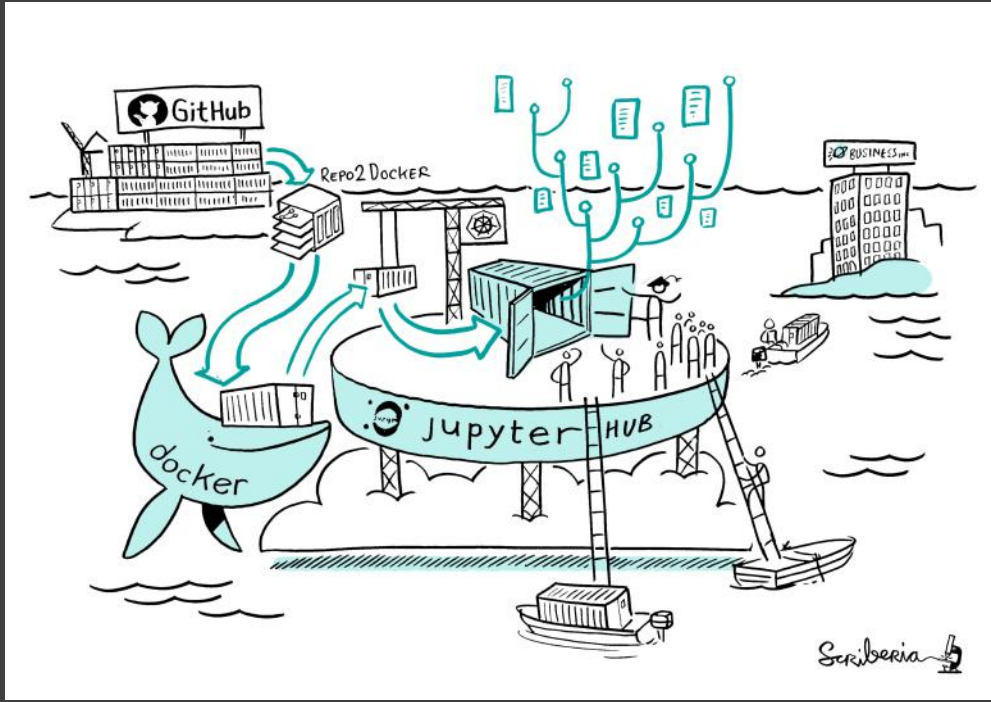




https://doi.org/10.5281/zenodo.3404774

BinderHub

– Collection of tools working in harmony which BinderHub orchestrates



https://doi.org/10.5281/zenodo.3404774

Scaling a BinderHub for multiple users

Problems if you run this on one computer:

– Resource intensive

– Resource control

– Security

Solution: Kubernetes!

– Resource intensive → Cluster management
– Resource control → Container management
– Security → Container isolation

https://doi.org/10.5281/zenodo.3404774

Solution: Kubernetes!

– Resource intensive → Cluster management
– Resource control → Container management
– Security → Container isolation

Problem: Also Kubernetes… 😢

https://doi.org/10.5281/zenodo.3404774

This Workshop

- What it is: Challenging!
- What it’s not: A cloud/Azure workshop
- What we’ll do: Build a BinderHub!

https://docs.microsoft.com/en-gb/learn/azure/

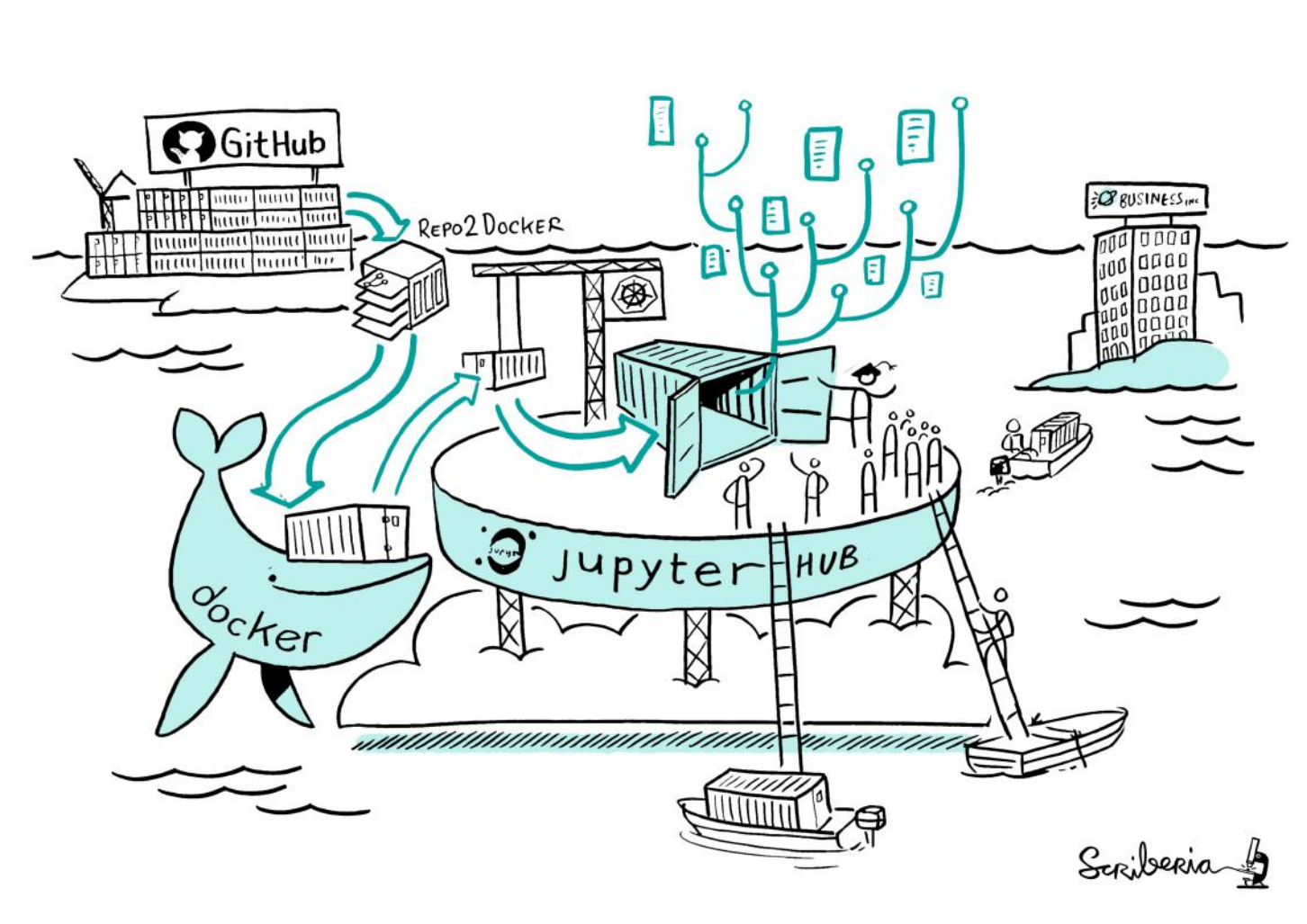
https://doi.org/10.5281/zenodo.3404774

This Workshop

bit.ly/zero-to-binderhub-workshop

HackMD: bit.ly/RSEConBinderHub

https://doi.org/10.5281/zenodo.3404774

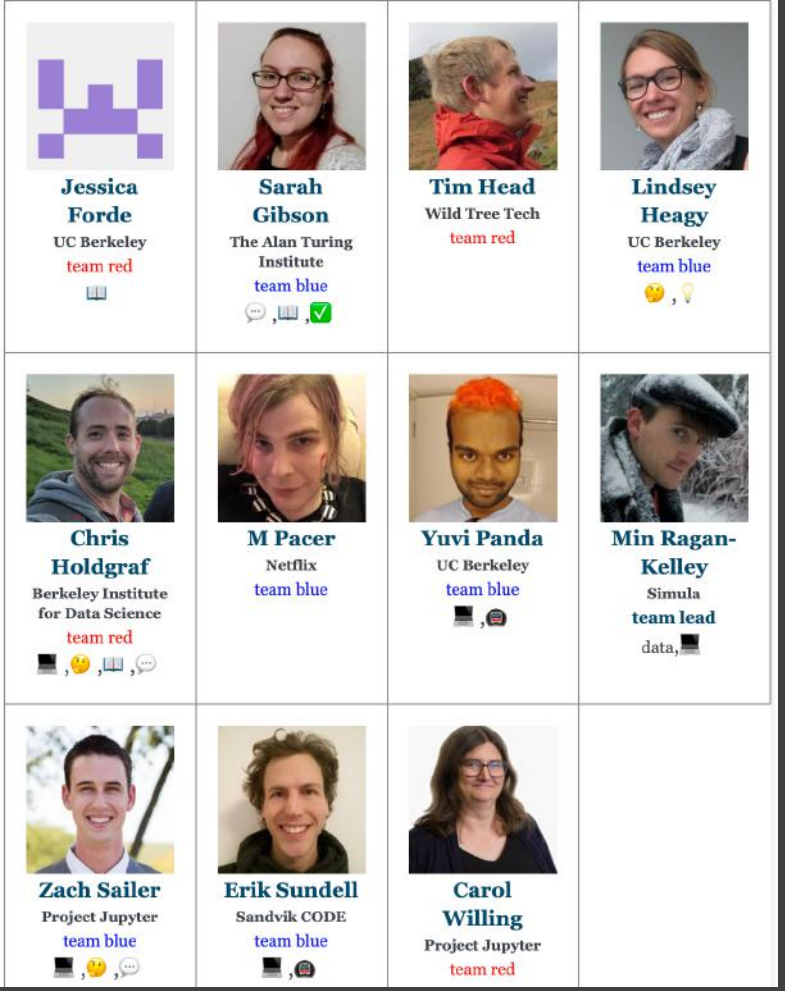


https://doi.org/10.5281/zenodo.3404774

- You have successfully built a BinderHub! ☝️
- Now check out this repo: github.com/alan-turing-institute/binderhub-deploy
- Please leave feedback in the HackMD: bit.ly/RSEConBinderHub

https://doi.org/10.5281/zenodo.3404774

Thank You!



https://doi.org/10.5281/zenodo.3404774

#TuringWay #ukrse19