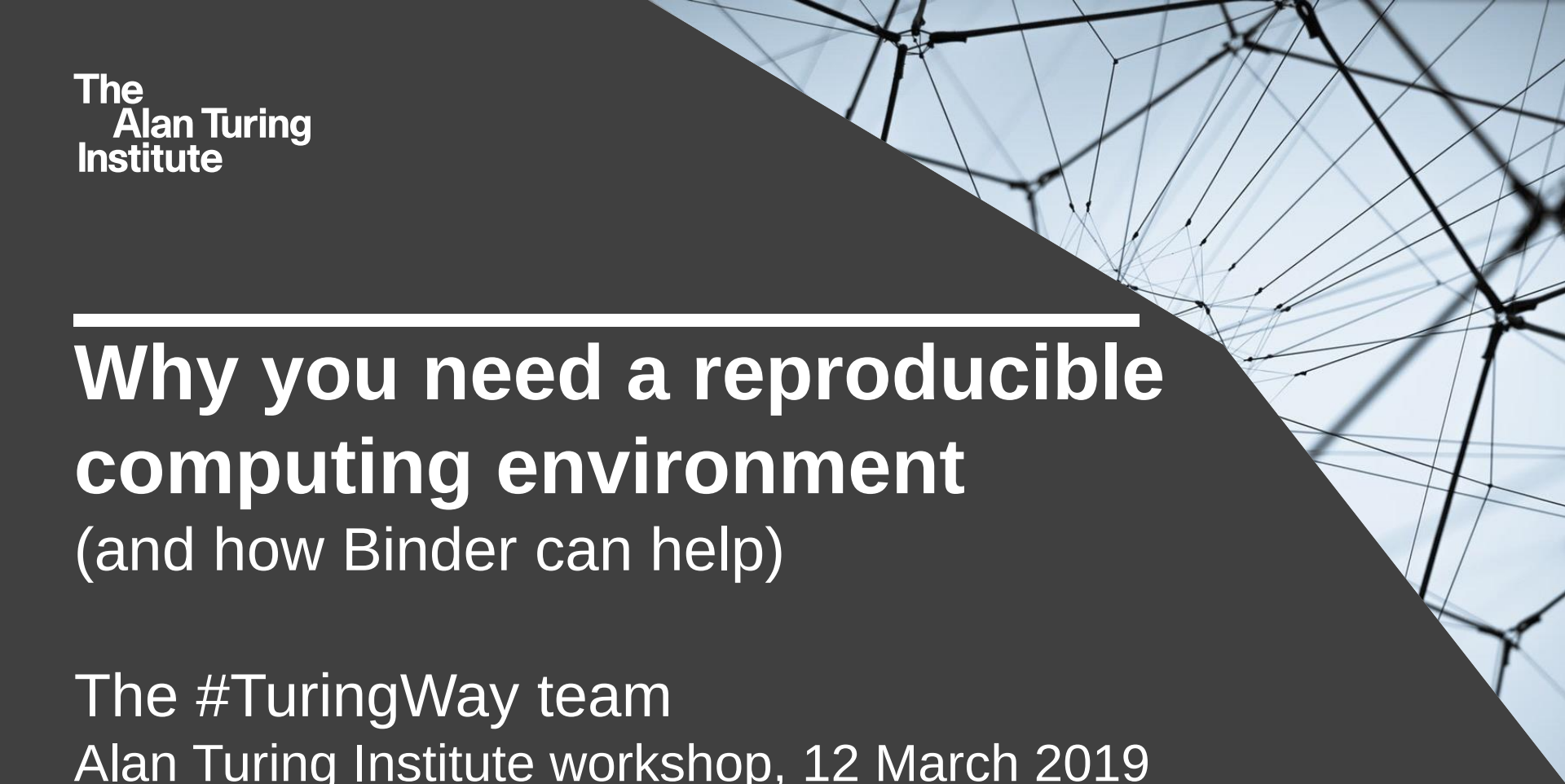
The Alan Turing Institute



Why you need a reproducible computing environment

(and how Binder can help)

The #TuringWay team

Alan Turing Institute workshop, 12 March 2019

Neurohackweek 2016

Photo credit: Chris Gorgolewski

The science is the code

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Buckheit and Donoho
(paraphrasing John Claerbout)
WaveLab and Reproducible Research, 1995

Slide courtesy of Chris Holdraf and the Jupyter Team

What does reproducible mean?

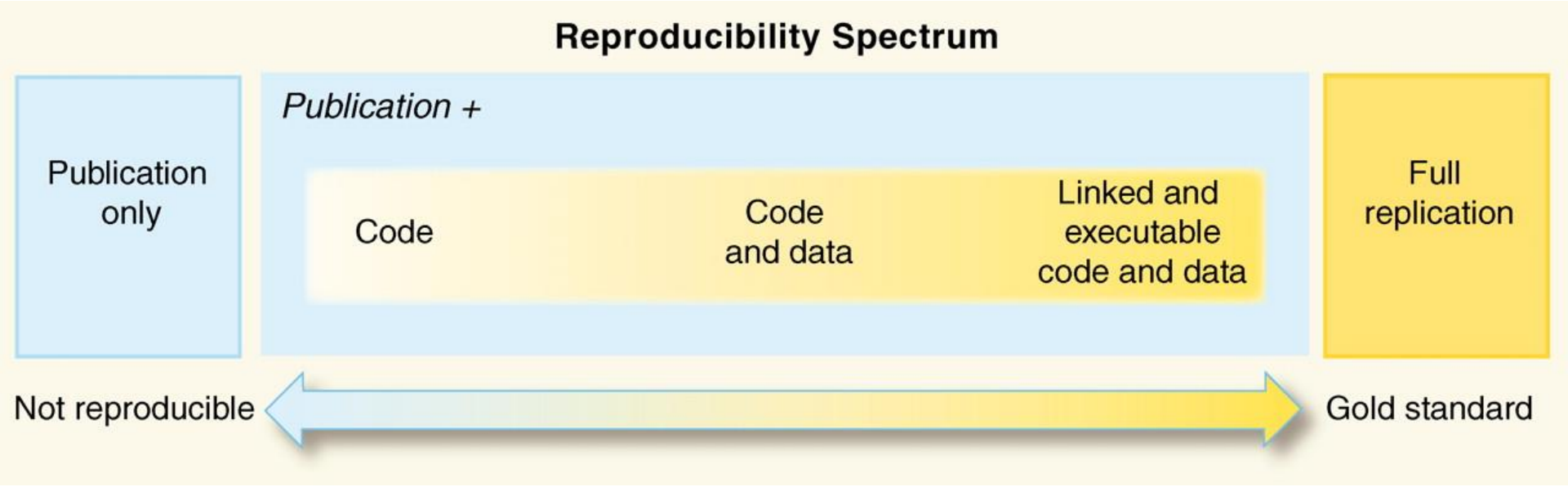
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://dx.doi.org/10.6084/m9.figshare.7140050

Upsetting take home message

Sharing your code and data isn't enough!

You need the computational environment too



Peng, Science, 2011

You need the computational environment too

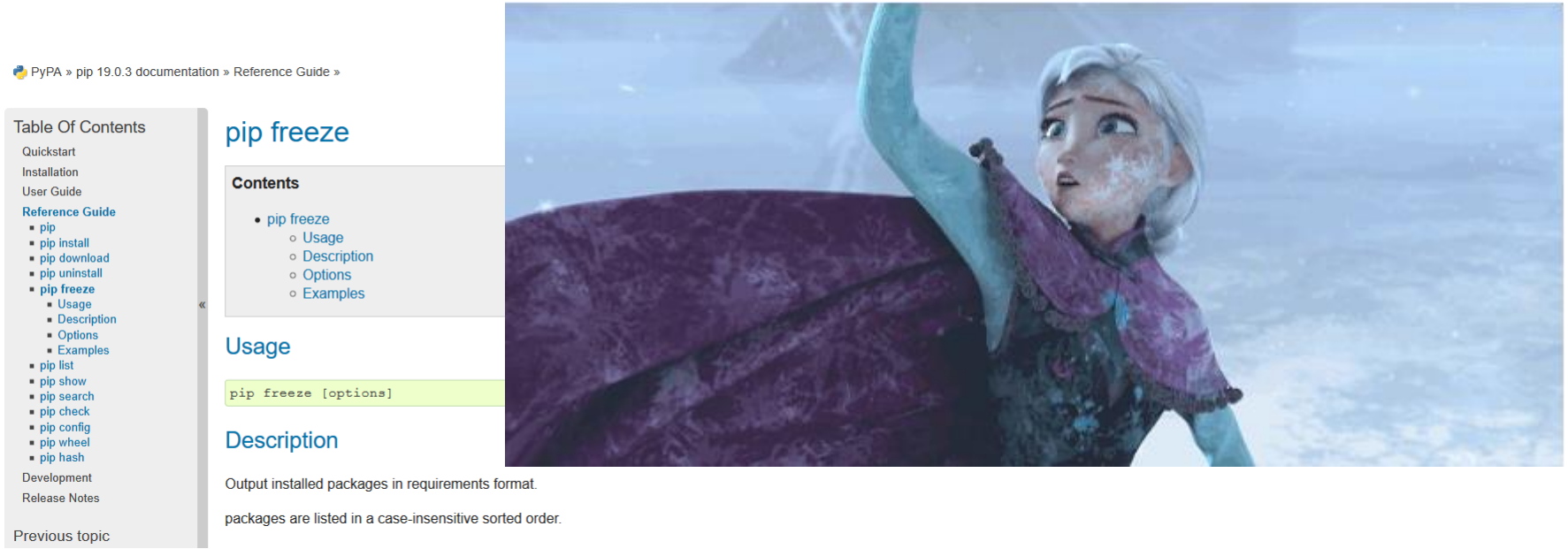
Peng, Science, 2011

You need the computational environment too

- Hardware (GPU, CPU)
- Operating system (mac, windows, linux)
- Software
  - Language version
  - Package versions

And all the interactions between the different layers

Capturing your local environment



PyPA » pip 19.0.3 documentation » Reference Guide »

Table Of Contents

- Quickstart
- Installation
- User Guide
- **Reference Guide**
  - pip
  - pip install
  - pip download
  - pip uninstall
  - **pip freeze**
    - Usage
    - Description
    - Options
    - Examples
  - pip list
  - pip show
  - pip search
  - pip check
  - pip config
  - pip wheel
  - pip hash
- Development
- Release Notes

Previous topic

pip freeze

Contents

- pip freeze
  - Usage
  - Description
  - Options
  - Examples

Usage

```
pip freeze [options]
```

Description

Output installed packages in requirements format.

packages are listed in a case-insensitive sorted order.

https://pip.pypa.io/en/stable/reference/pip_freeze

Capturing your local environment

PyPA » pip 19.0.3 documentation » Reference Guide »

Table Of Contents

- Quickstart
- Installation
- User Guide
- **Reference Guide**
  - pip
  - pip install
  - pip download
  - pip uninstall
  - **pip freeze**
    - Usage
    - Description
    - Options
    - Examples
  - pip list
  - pip show
  - pip search
  - pip check
  - pip config

pip freeze

Contents

- pip freeze
  - Usage
  - Description
  - Options
  - Examples

Usage

```
pip freeze [options]
```

Examples

1. Generate output suitable for a requirements file.

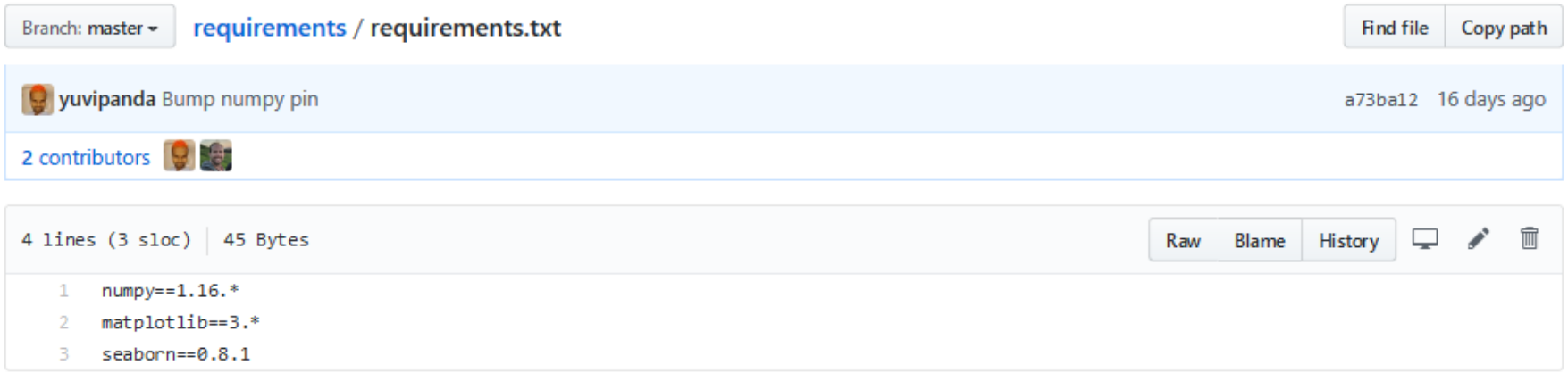
```
$ pip freeze
docutils==0.11
Jinja2==2.7.2
MarkupSafe==0.19
Pygments==1.6
Sphinx==1.2.2
```

2. Generate a requirements file and then install from it in another environment.

```
$ env1/bin/pip freeze > requirements.txt
$ env2/bin/pip install -r requirements.txt
```

https://pip.pypa.io/en/stable/reference/pip_freeze

Requirements.txt



https://github.com/binder-examples/requirements

Yet another markup language

Branch: master ▾ python-conda_pip / environment.yml

Find file | Copy path

choldgraf adding a notebook 10ba338 on 23 Nov 2017

1 contributor

12 lines (11 sloc) | 165 Bytes

Raw | Blame | History

```
1  name: example-environment
2  channels:
3    - conda-forge
4  dependencies:
5    - python
6    - numpy
7    - pip:
8      - nbgitpuller
9      - sphinx-gallery
10     - pandas
11     - matplotlib
```

https://github.com/binder-examples/python-conda_pip

Yet another markup language

Branch: master ▾ python-conda_pip / environment.yml

Find file | Copy path

choldgraf adding a notebook — 10ba338 on 23 Nov 2017

1 contributor

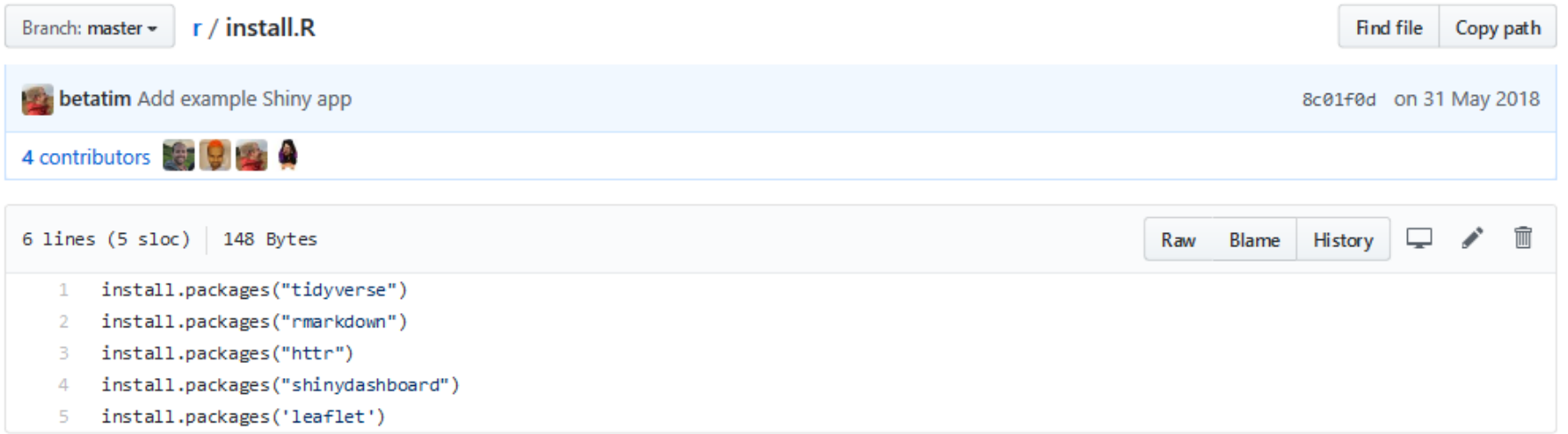
12 lines (11 sloc) | 165 Bytes — Raw | Blame | History

```
1  name: example-environment
2  channels:
3    - conda-forge
4  dependencies:
5    - python
6    - numpy
7    - pip:
8      - nbgitpuller
9      - sphinx-gallery
10     - pandas
11     - matplotlib
```

(This is actually a json file in the background)

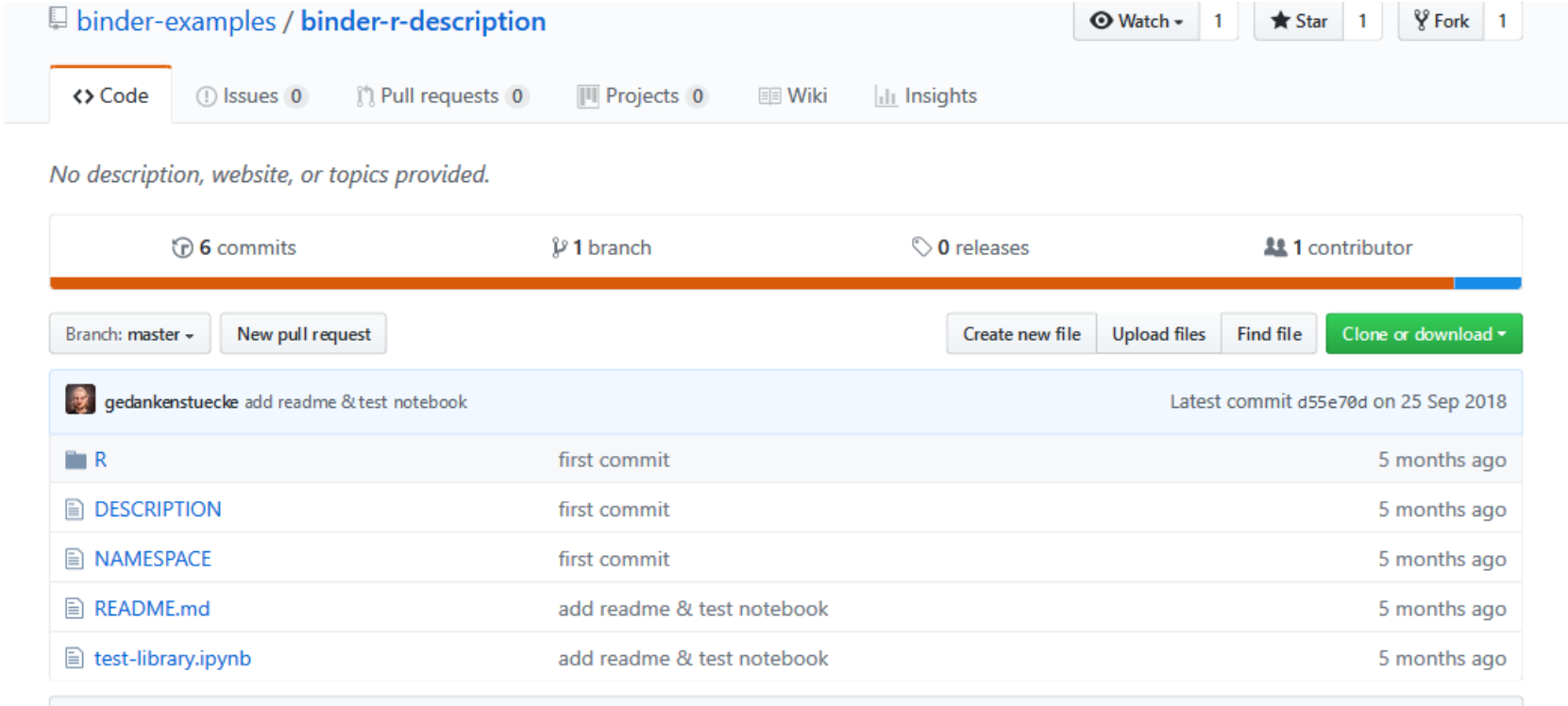
https://github.com/binder-examples/python-conda_pip

R and RStudio



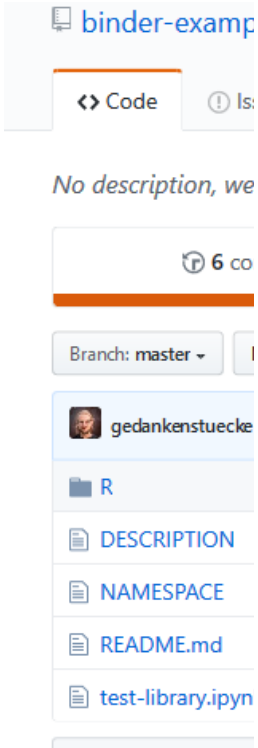
https://github.com/binder-examples/r

Pinning to a version on MRAN



https://github.com/binder-examples/binder-r-description

Pinning to a version on MRAN



README.md

Specifying an R environment by having a DESCRIPTION file

Jupyter+R: launch binder

RStudio: launch binder

Binder supports using R and RStudio, with libraries pinned to a specific snapshot on MRAN.

If you specify a `runtime.txt` file that is formatted like:

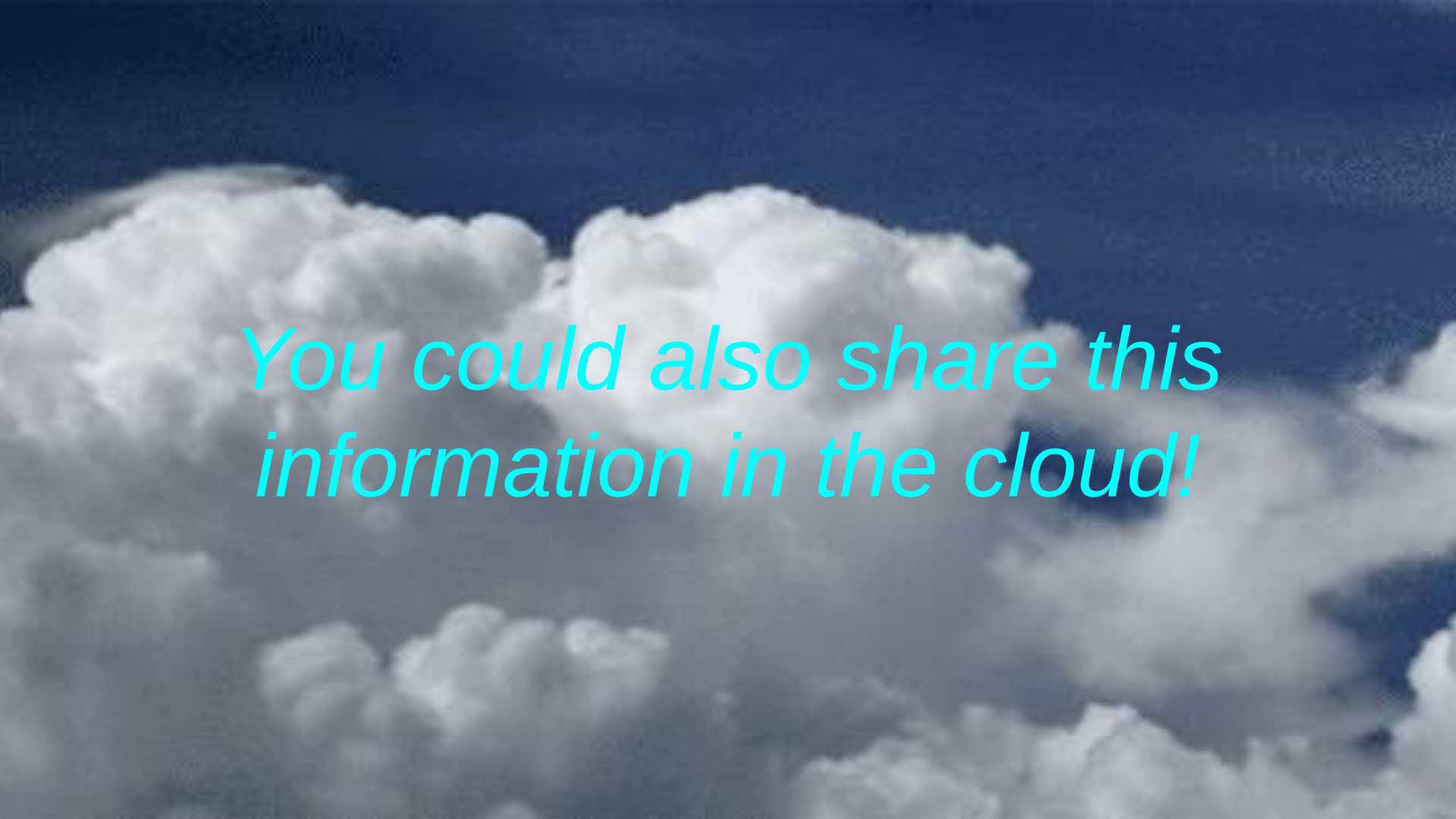
```
r-<YYYY>-<MM>-<DD>
```

where YYYY-MM-DD it will use the MRAN snapshot of that day for setting up the R runtime.

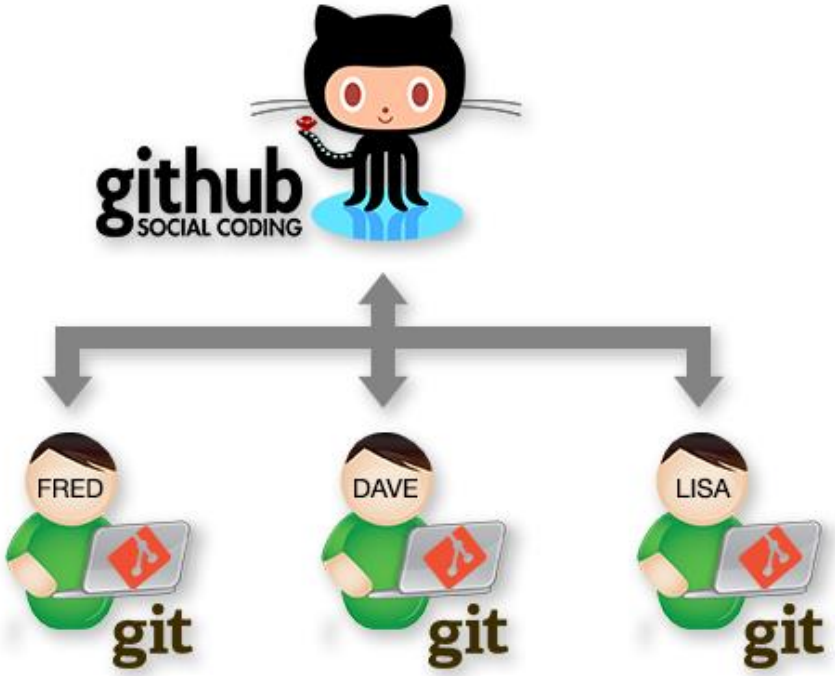
Without specifying a `runtime.txt` it will use a 2-day old snapshot of MRAN.

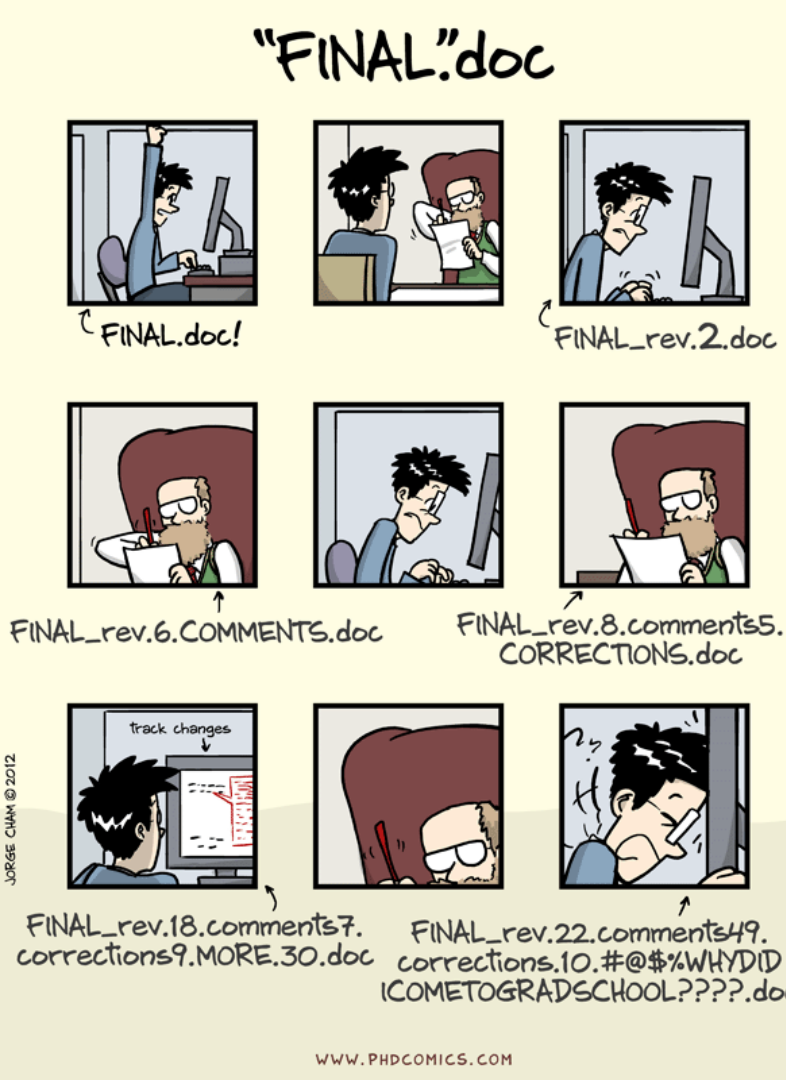
Both RStudio and IRKernel are installed by default, so you can use either the Jupyter notebook interface or the RStudio interface.

https://github.com/binder-examples/binder-r-description

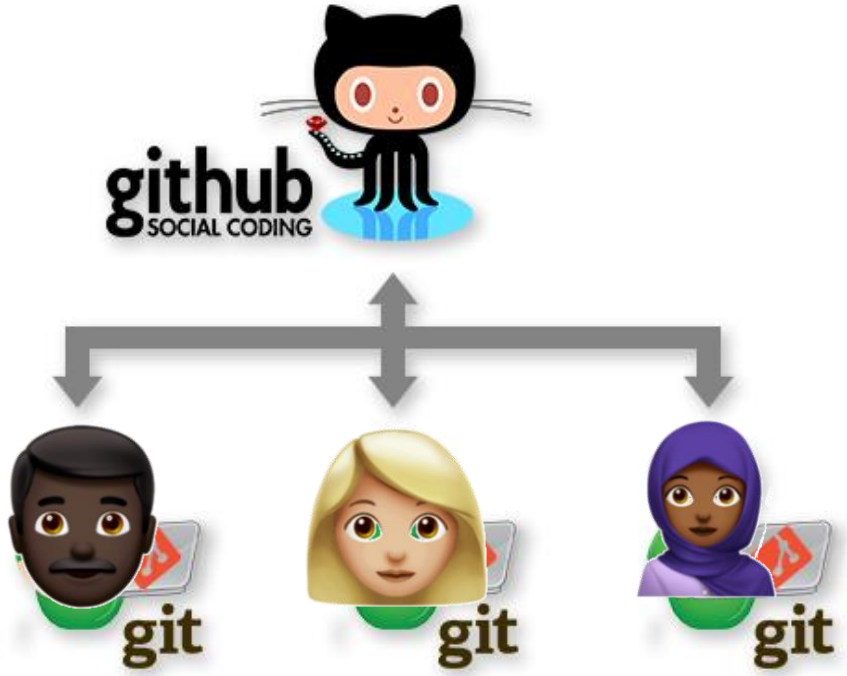

You could also share this
information in the cloud!

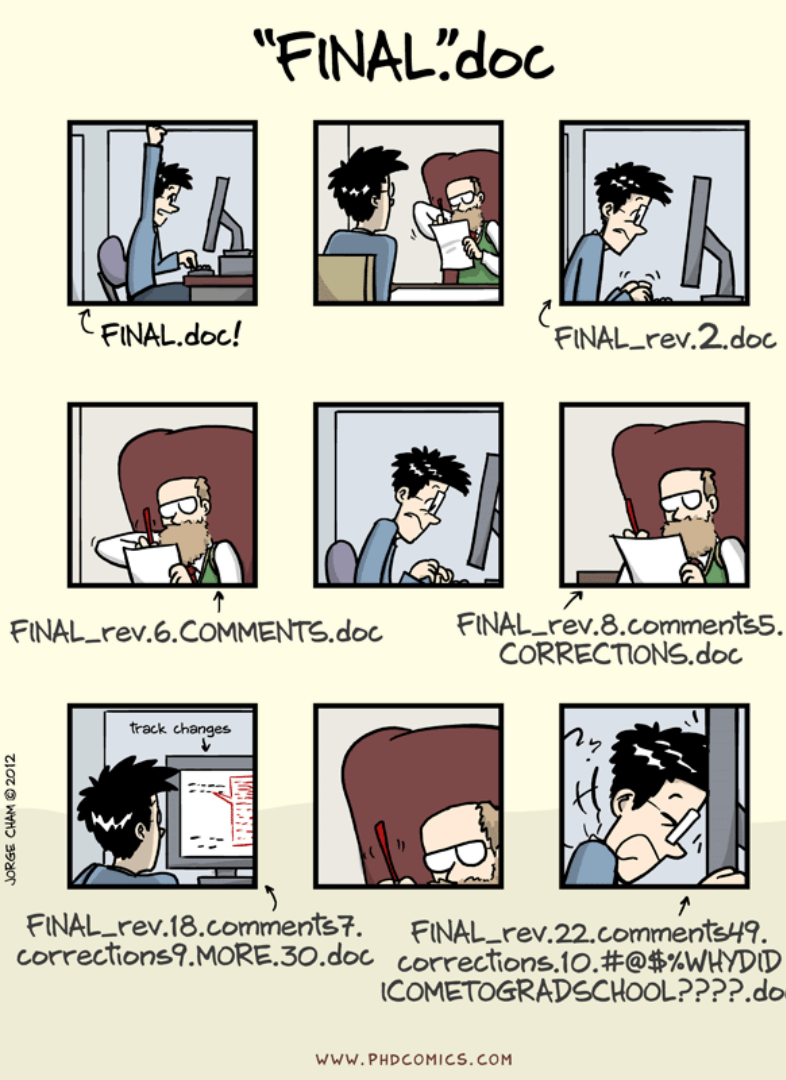
Put your code in the cloud



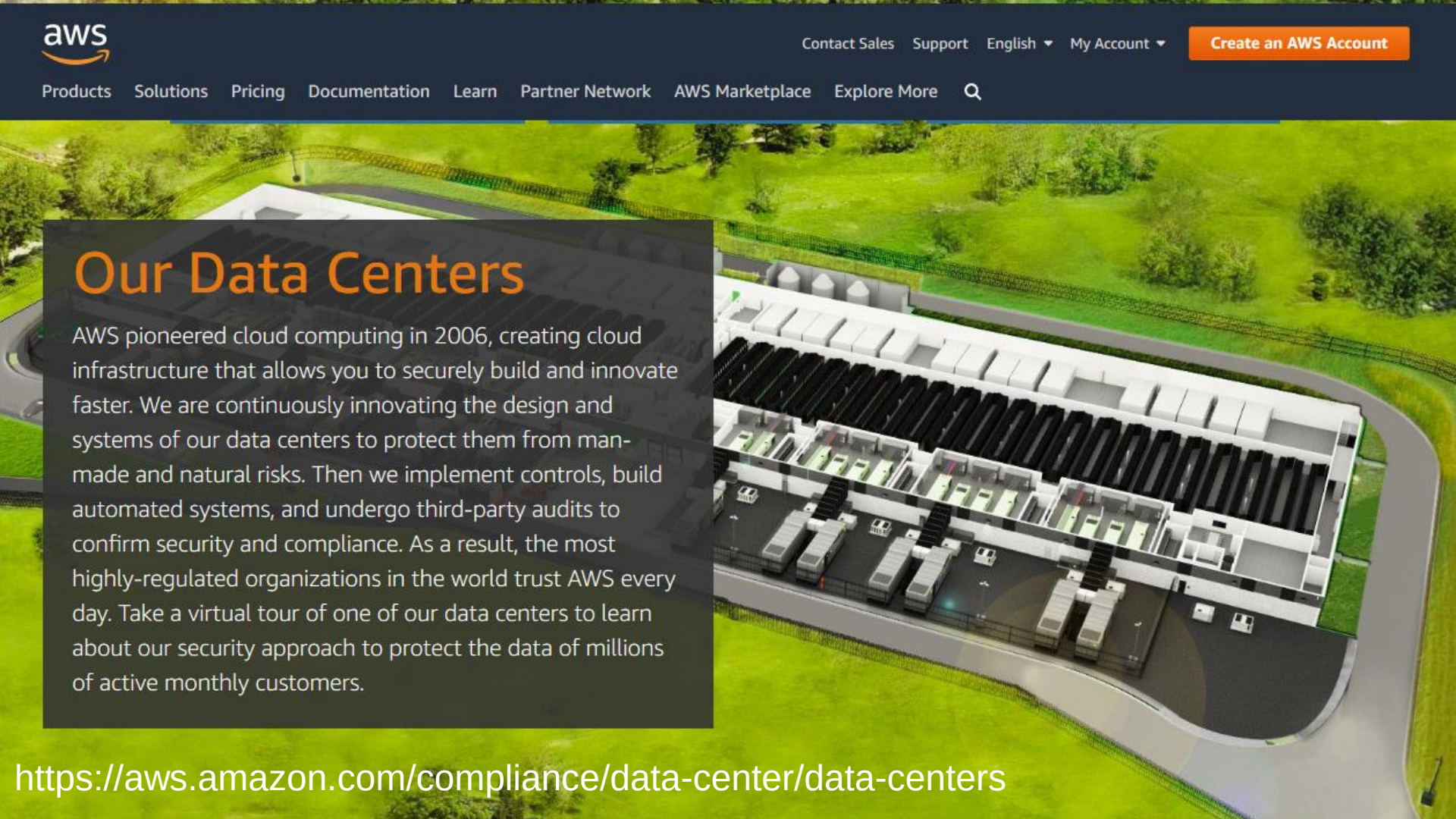


Put your code in the cloud





A server is someone else’s computer



Our Data Centers

AWS pioneered cloud computing in 2006, creating cloud infrastructure that allows you to securely build and innovate faster. We are continuously innovating the design and systems of our data centers to protect them from man-made and natural risks. Then we implement controls, build automated systems, and undergo third-party audits to confirm security and compliance. As a result, the most highly-regulated organizations in the world trust AWS every day. Take a virtual tour of one of our data centers to learn about our security approach to protect the data of millions of active monthly customers.

https://aws.amazon.com/compliance/data-center/data-centers

aws

Contact Sales | Support | English | My Account | Create an AWS Account

Products | Solutions | Pricing | Documentation | Learn | Partner Network | AWS Marketplace | Explore More

Our Data Centers

AWS pioneered cloud computing in 2006, creating cloud infrastructure that allows you to securely build and innovate faster. We are continuously innovating the design and systems of our data centers to protect them from man-

PERIMETER LAYER

AWS data center physical security begins at the Perimeter Layer. This layer includes a number of security features depending on the location, such as security guards, fencing, security feeds, intrusion detection technology, and other security measures.

EXPLORE »

DATA LAYER

The Data Layer is the most critical point of protection because it is the only area that holds customer data. Protection begins by restricting access and maintaining a separation of privilege for each layer. In addition, we deploy threat detection devices and system protocols, further safeguarding this layer.

EXPLORE »

INFRASTRUCTURE LAYER

The Infrastructure Layer is the data center building and the equipment and systems that keep it running. Components like back-up power equipment, the HVAC system, and fire suppression equipment are all part of the Infrastructure Layer.

EXPLORE »

ENVIRONMENTAL LAYER

The Environmental Layer is dedicated to environmental considerations from site selection and construction to operations and sustainability. AWS carefully chooses our data center locations to mitigate environmental risk, such as flooding, extreme weather, and seismic activity.

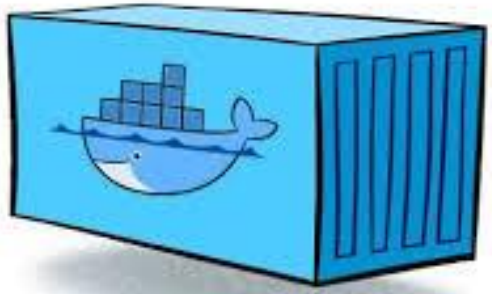
EXPLORE »

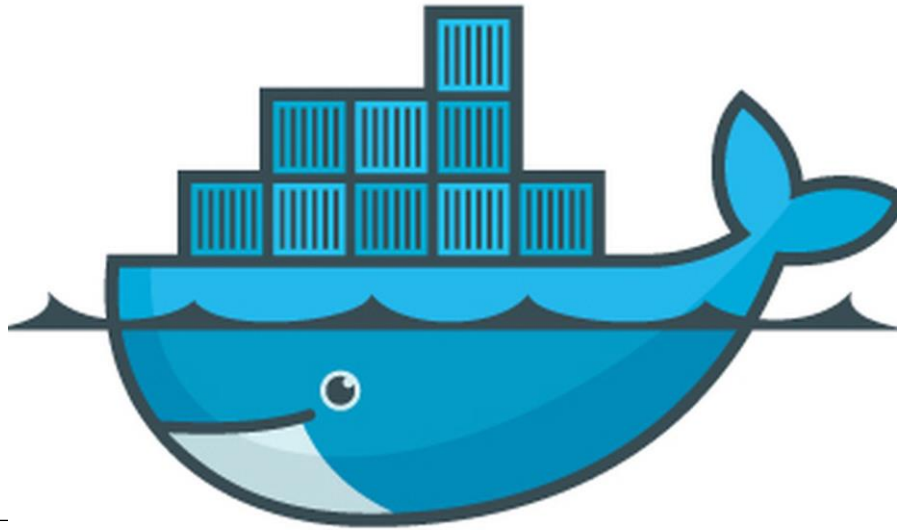
https://aws.amazon.com/compliance/data-center/data-centers

These computers run software

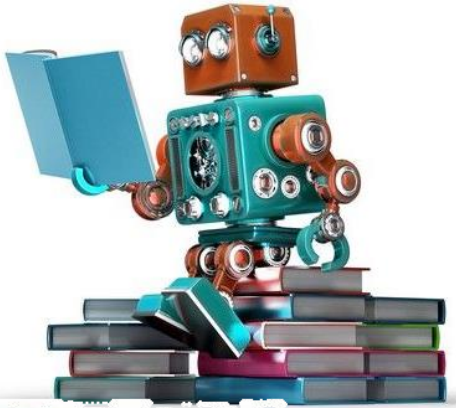
- Docker is a container that bundles all the infrastructure and software together.

- You don’t have to worry about all the different moving parts, just use the same set up and you’ll be fine.

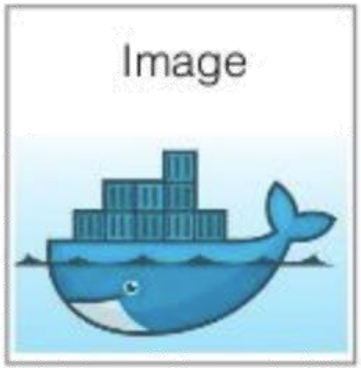




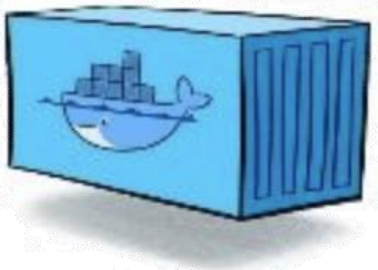
Human and machine readable files



build



run



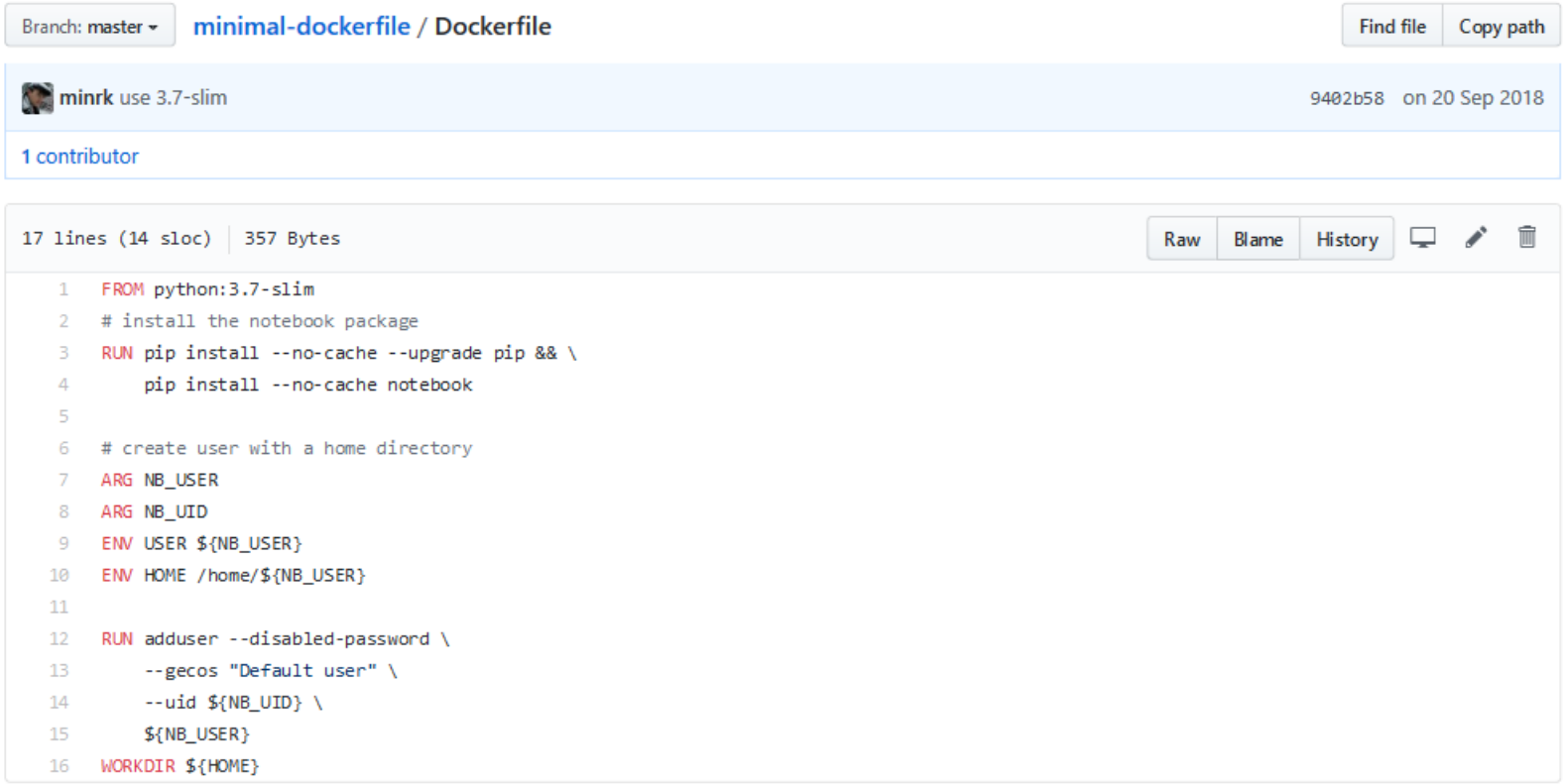
Dockerfile

Docker Image

Docker Container

https://medium.com/platformer-blog/practical-guide-on-writing-a-dockerfile-for-your-application-89376f88b3b5

Docker container



Branch: master ▾ minimal-dockerfile / Dockerfile

Find file | Copy path

minrk use 3.7-slim — 9402b58 on 20 Sep 2018

1 contributor

17 lines (14 sloc) | 357 Bytes — Raw | Blame | History

```
1  FROM python:3.7-slim
2  # install the notebook package
3  RUN pip install --no-cache --upgrade pip && \
4      pip install --no-cache notebook
5
6  # create user with a home directory
7  ARG NB_USER
8  ARG NB_UID
9  ENV USER ${NB_USER}
10 ENV HOME /home/${NB_USER}
11
12 RUN adduser --disabled-password \
13     --gecos "Default user" \
14     --uid ${NB_UID} \
15     ${NB_USER}
16 WORKDIR ${HOME}
```

https://github.com/binder-examples/minimal-dockerfile

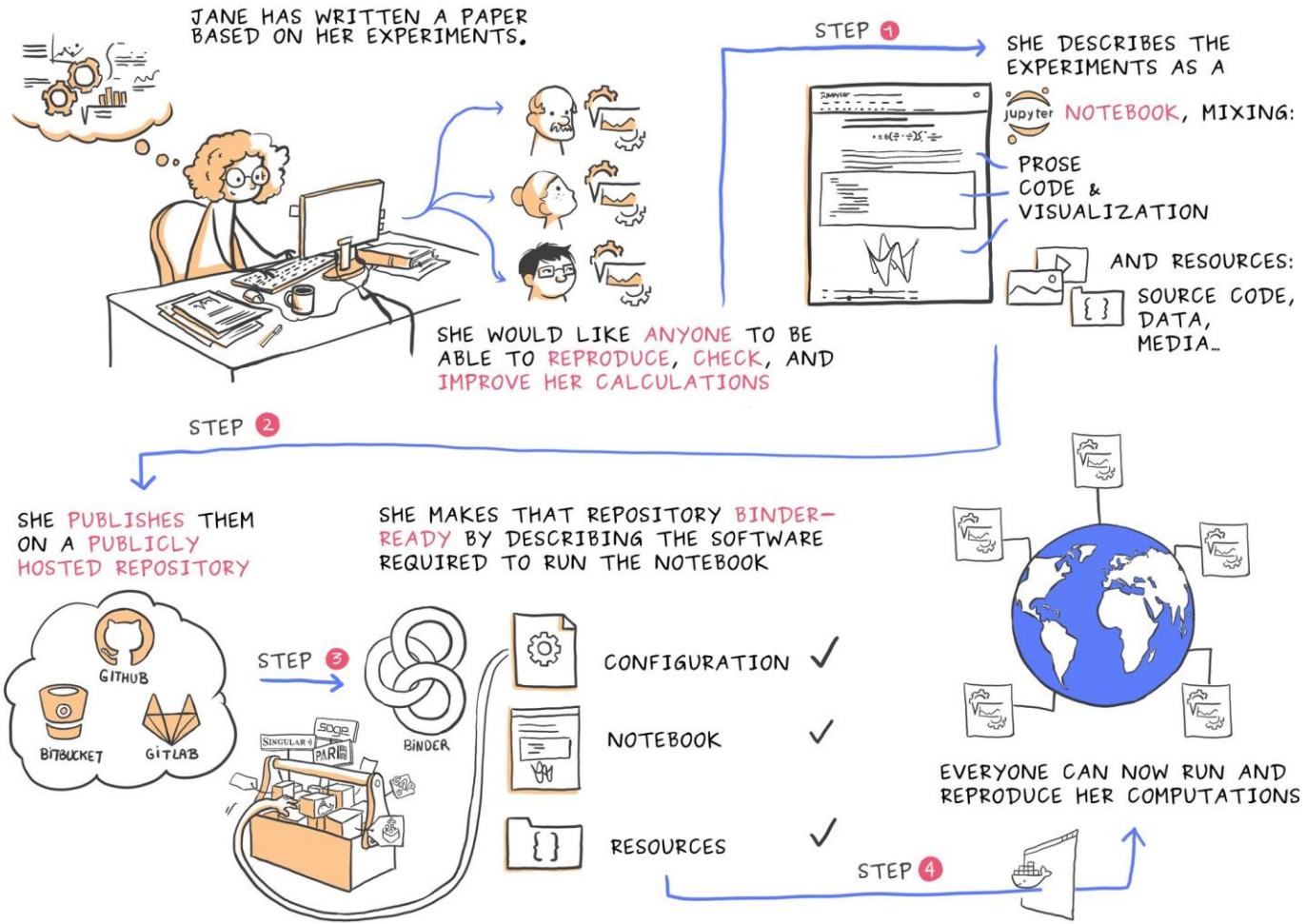
Small group exercise

Please get into groups of 3-4 and explore the examples below.

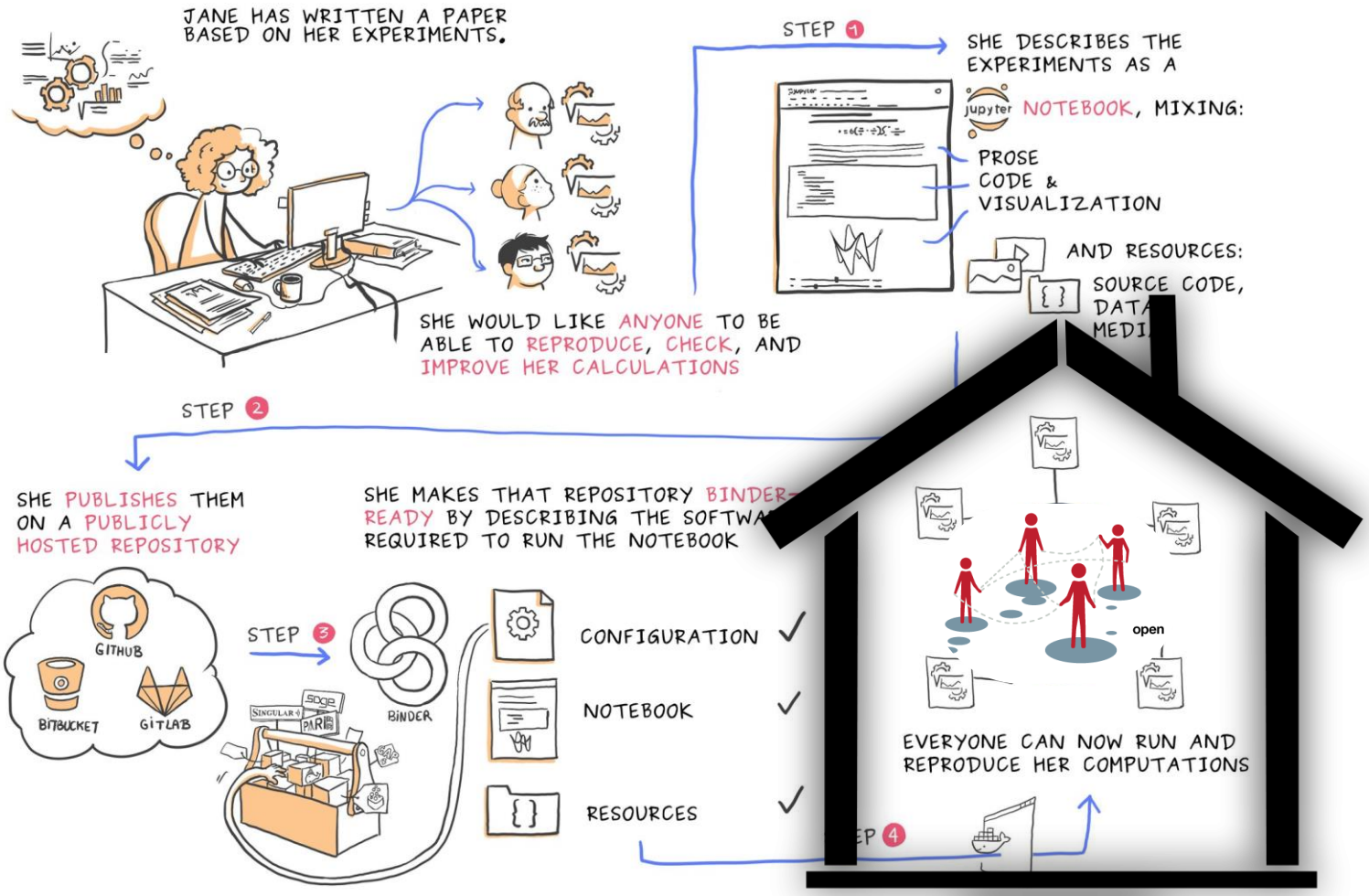
Try to answer:

- *Are there differences between different branches?*
- *Does that give different results?*
- *Did you get what you'd expect?*

- https://github.com/alan-turing-institute/CompEnv-Ex1
- https://github.com/alan-turing-institute/CompEnv-Ex2
- https://github.com/alan-turing-institute/CompEnv-Ex3
- https://github.com/alan-turing-institute/CompEnv-Ex4



Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960

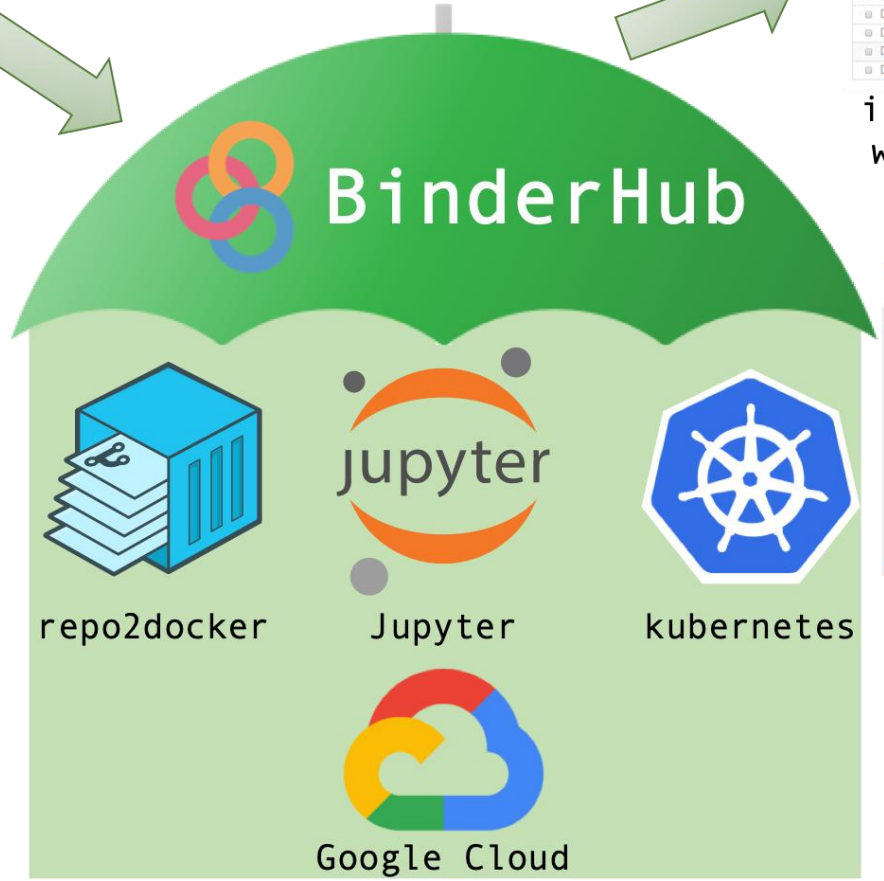


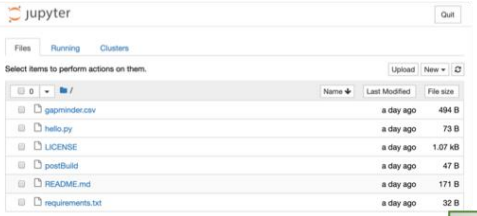
Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960

The public BinderHub instance

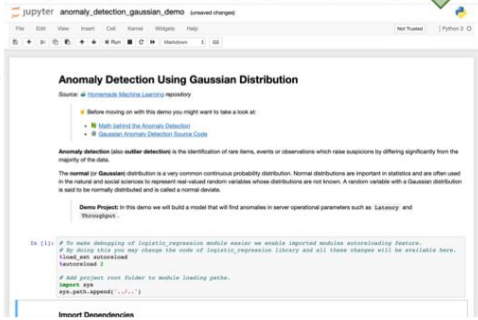
your GitHub repo

mybinder.org





interactive browser with your code and computational environment



The Turing Way


#TuringWay


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


gitter.im/alan-turing-institute/the-turing-way