BERKELEY Institute for Data Science

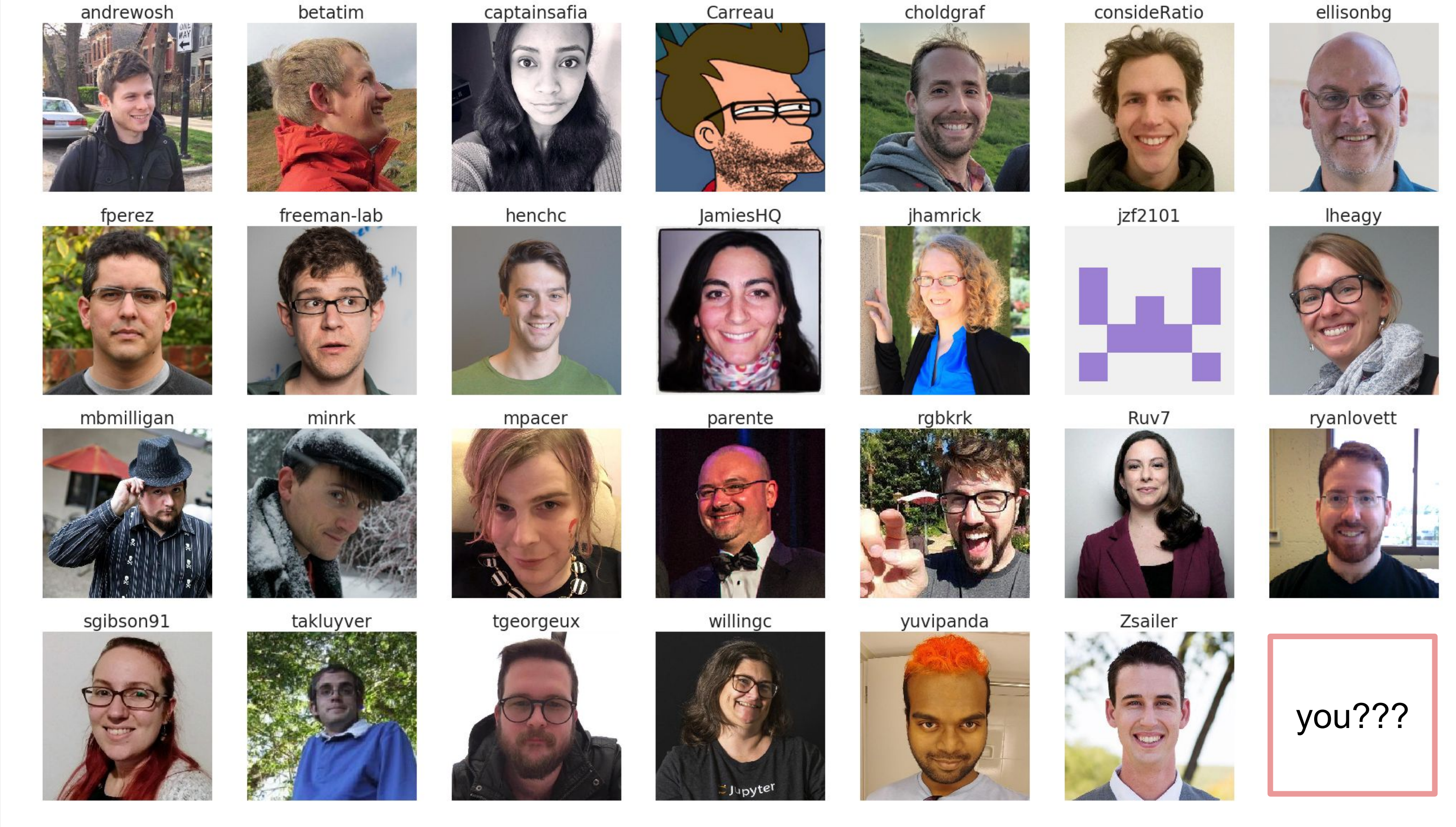
Berkeley Division of Data Sciences

Open Infrastructure for open science

How Binder Powers an Open Stack in the Cloud

Chris Holdgraf, UC Berkeley and Project Jupyter

@choldgraf


andrewosh
betatim
captainsafia
Carreau
choldgraf
consideRatio
ellisonbg
fperez
freeman-lab
henchc
JamiesHQ
jhamrick
jzf2101
lheagy
mbmilligan
minrk
mpacer
parente
rgbkrk
Ruv7
ryanlovett
sgibson91
takluyver
tgeorgeux
willingc
yuvipanda
Zsailer
you???

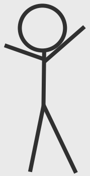
THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

ALFRED P. SLOAN FOUNDATION

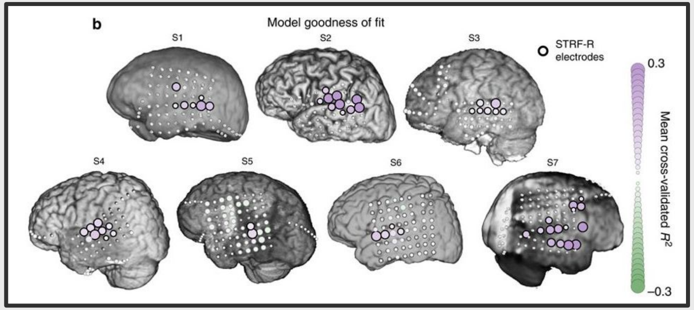
GORDON AND BETTY MOORE FOUNDATION

A bit about me then...

Berkeley Neuroscience
HELEN WILLS NEUROSCIENCE INSTITUTE



Cognitive Neuroscience



Open Source

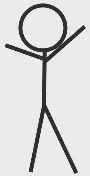
MNE
MEG + EEG ANALYSIS & VISUALIZATION

@choldgraf

A bit about me now...

Research and Open Source

BERKELEY
Institute for
Data Science



Education and Open Source

Berkeley
Division of
Data Sciences

jupyterhub

jupyter

Jupyter @ Berkeley

a *community* of people and an *ecosystem* of open tools and standards for interactive computing create things that are **language-agnostic** and **modular**.

Empower people to use **other open tools**.

Aside: Jupyter and the last mile problem


You
San Jose Coffee!

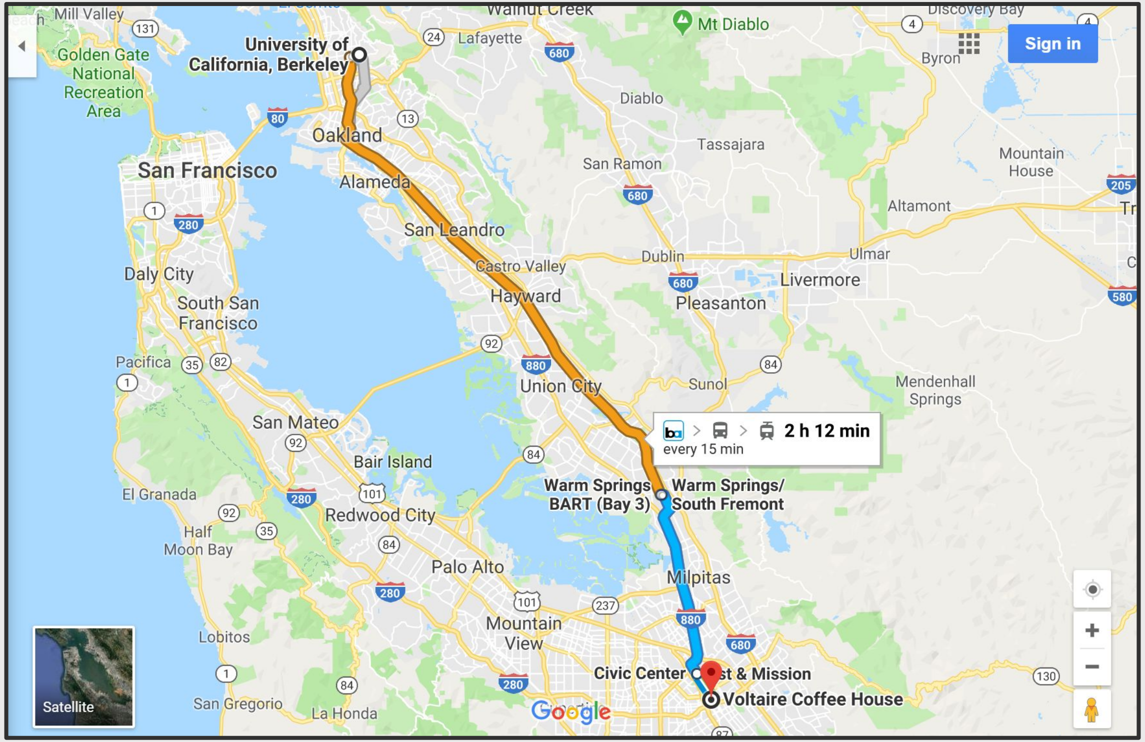
One option: walk there by myself

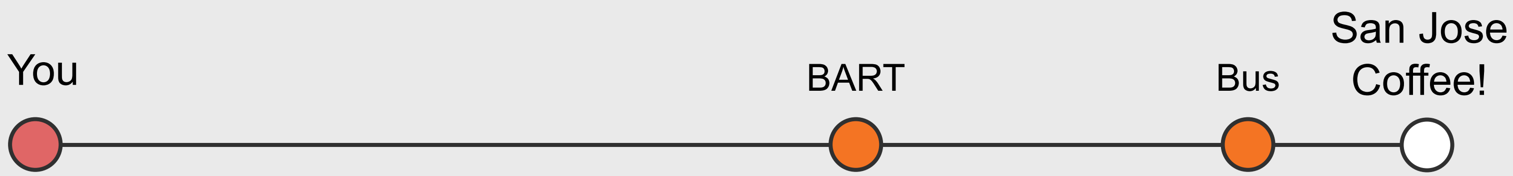
Another option: pay somebody to drive me

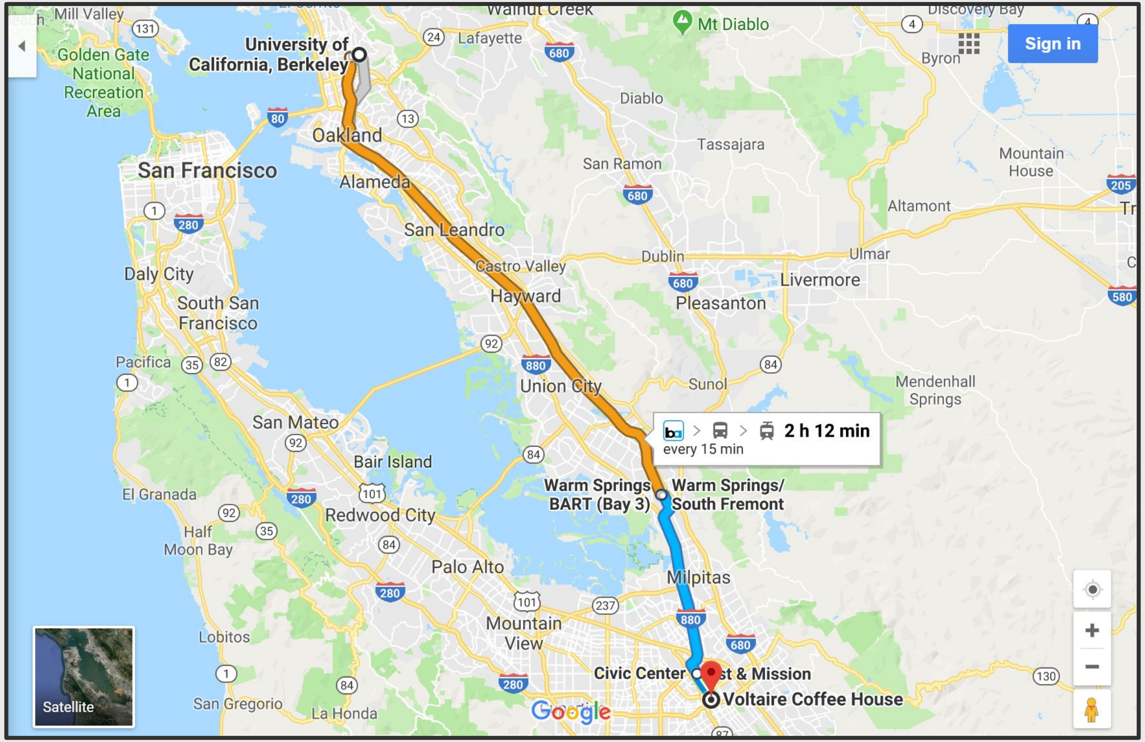
***My favorite option**: use public infrastructure*

You

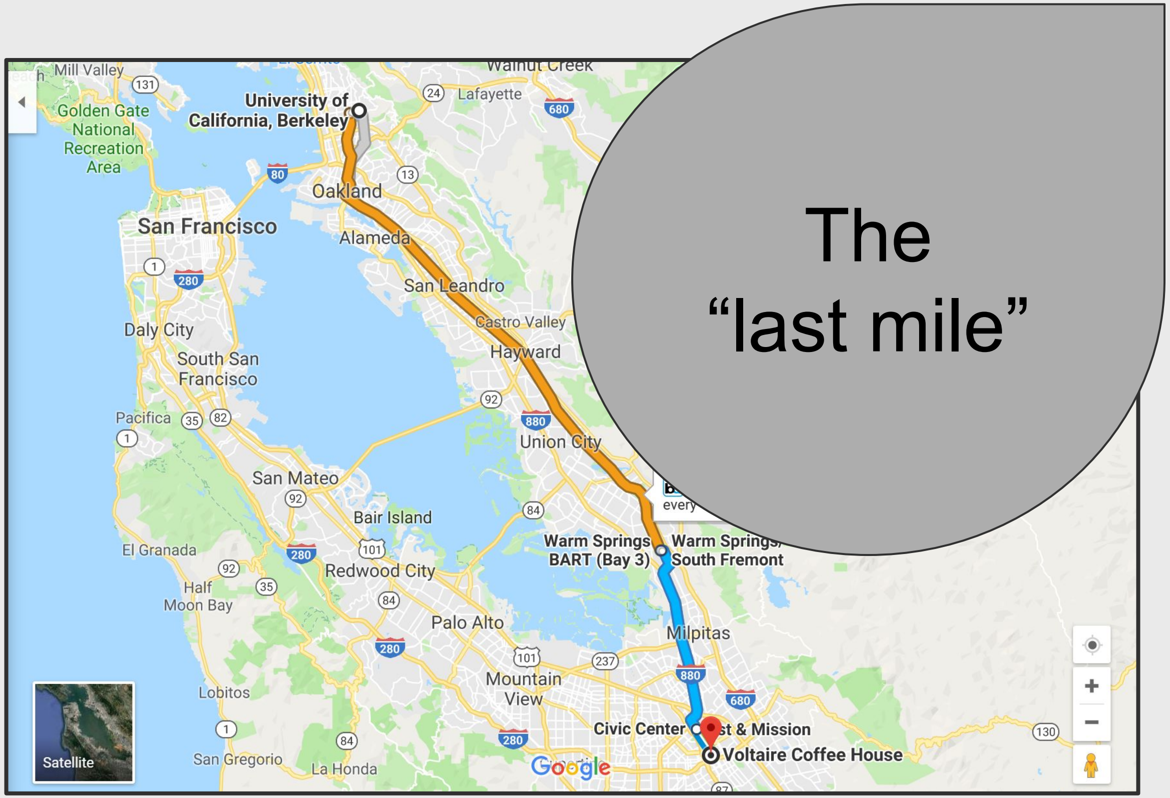
San Jose Coffee!




You
BART
Bus
San Jose
Coffee!


University of
California, Berkeley
Oakland
Alameda
San Leandro
Castro Valley
Hayward
Union City
San Francisco
Daly City
South San
Francisco
Pacifica
San Mateo
Bair Island
Redwood City
Palo Alto
Mountain
View
Milpitas
Warm Springs
BART (Bay 3)
Warm Springs/
South Fremont
Civic Center
st & Mission
Voltaire Coffee House
2 h 12 min
every 15 min
Golden Gate
National
Recreation
Area
Mill Valley
Lafayette
Mt Diablo
Diablo
Tassajara
San Ramon
Dublin
Pleasanton
Livermore
Ulmar
Altamont
Mountain
House
Byron
Sunol
Mendenhall
Springs
El Granada
Half
Moon Bay
Lobitos
San Gregorio
La Honda
Sign in
Satellite
Google


You
BART
Bus
San Jose
Coffee!


The
“last mile”
University of
California, Berkeley
Oakland
San Francisco
Alameda
San Leandro
Castro Valley
Hayward
Union City
Daly City
South San
Francisco
Pacifica
San Mateo
Bair Island
Redwood City
Palo Alto
Mountain
View
Milpitas
Warm Springs
BART (Bay 3)
Warm Springs/
South Fremont
Civic Center
Voltaire Coffee House
Mill Valley
Golden Gate
National
Recreation
Area
Walnut Creek
Lafayette
El Granada
Half
Moon Bay
Lobitos
San Gregorio
La Honda
Satellite

Public infrastructure gets us closer to our goal.

It makes the last mile shorter.

How does **Jupyter**
fit in to this?

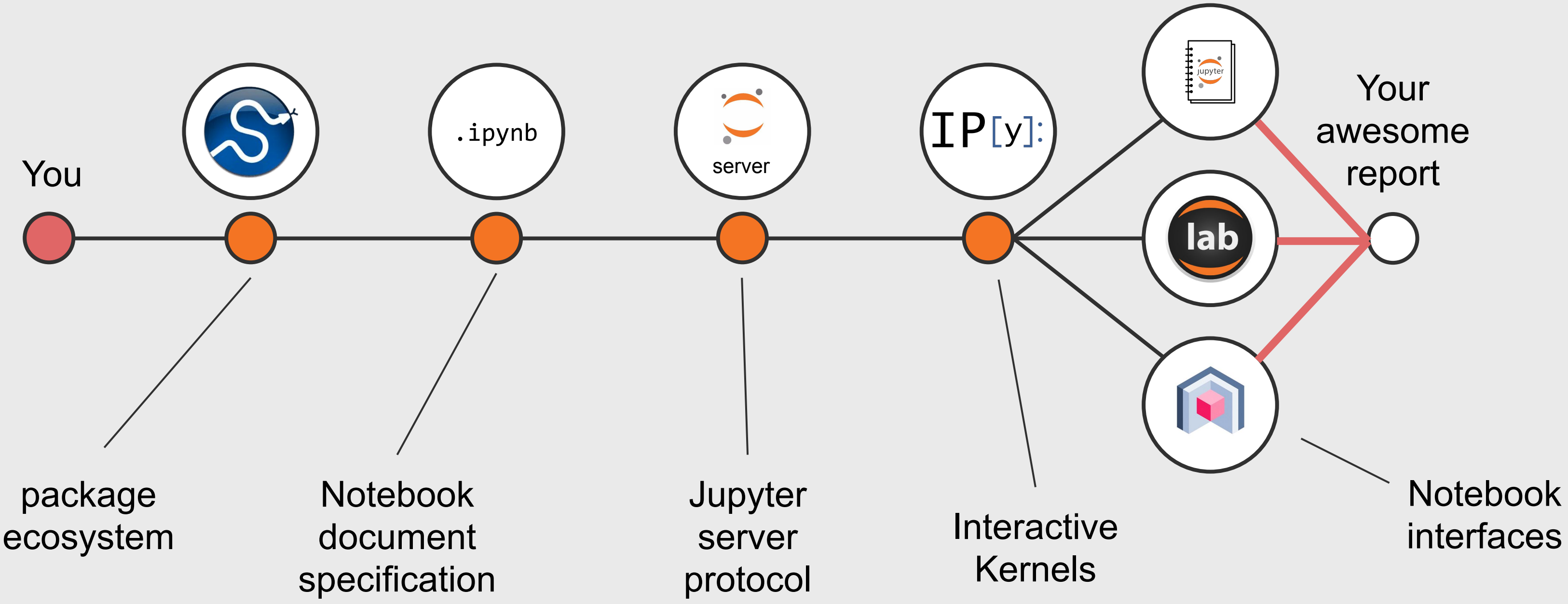

You
Your awesome report

One option: build it from scratch

Another option: pay somebody for a product

***My favorite option**: build on open, modular tools*

Jupyter shortens the last mile by creating and leveraging public infrastructure



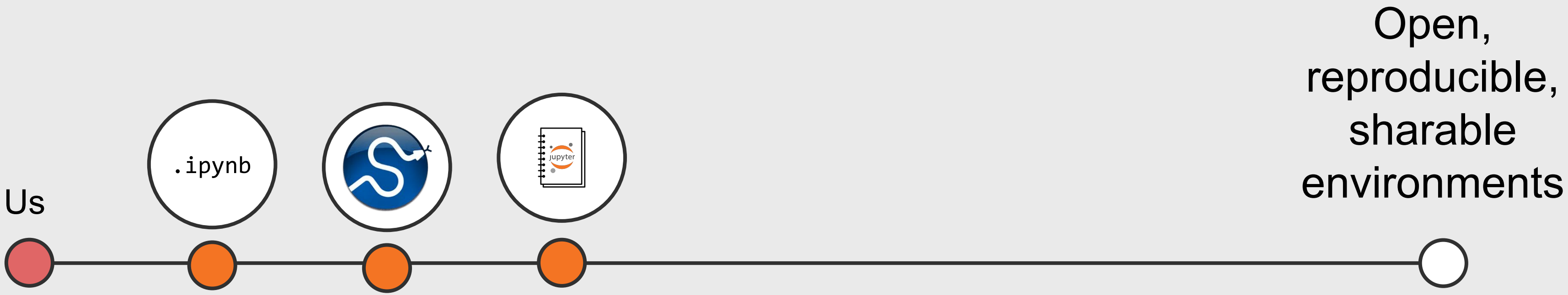
Back to our talk…

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

Our mission for this talk



Part 1: from your laptop to the cloud with **JupyterHub**

(some) data science should be taught to everyone

(no, really)

Here’s what this means at Berkeley...

How can we connect people with computation?

What is JupyterHub?

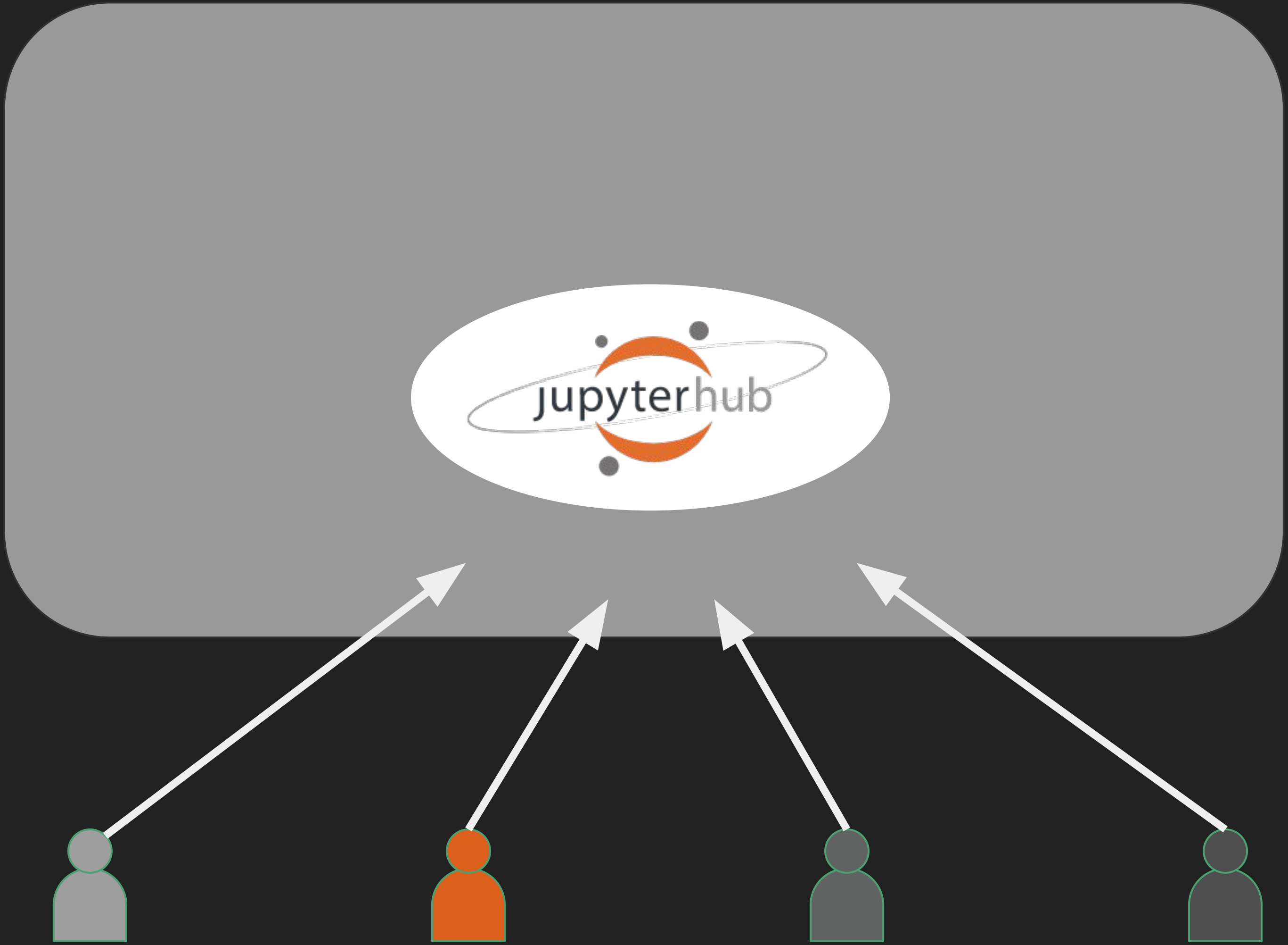
Host pre-configured data science environments on shared infrastructure

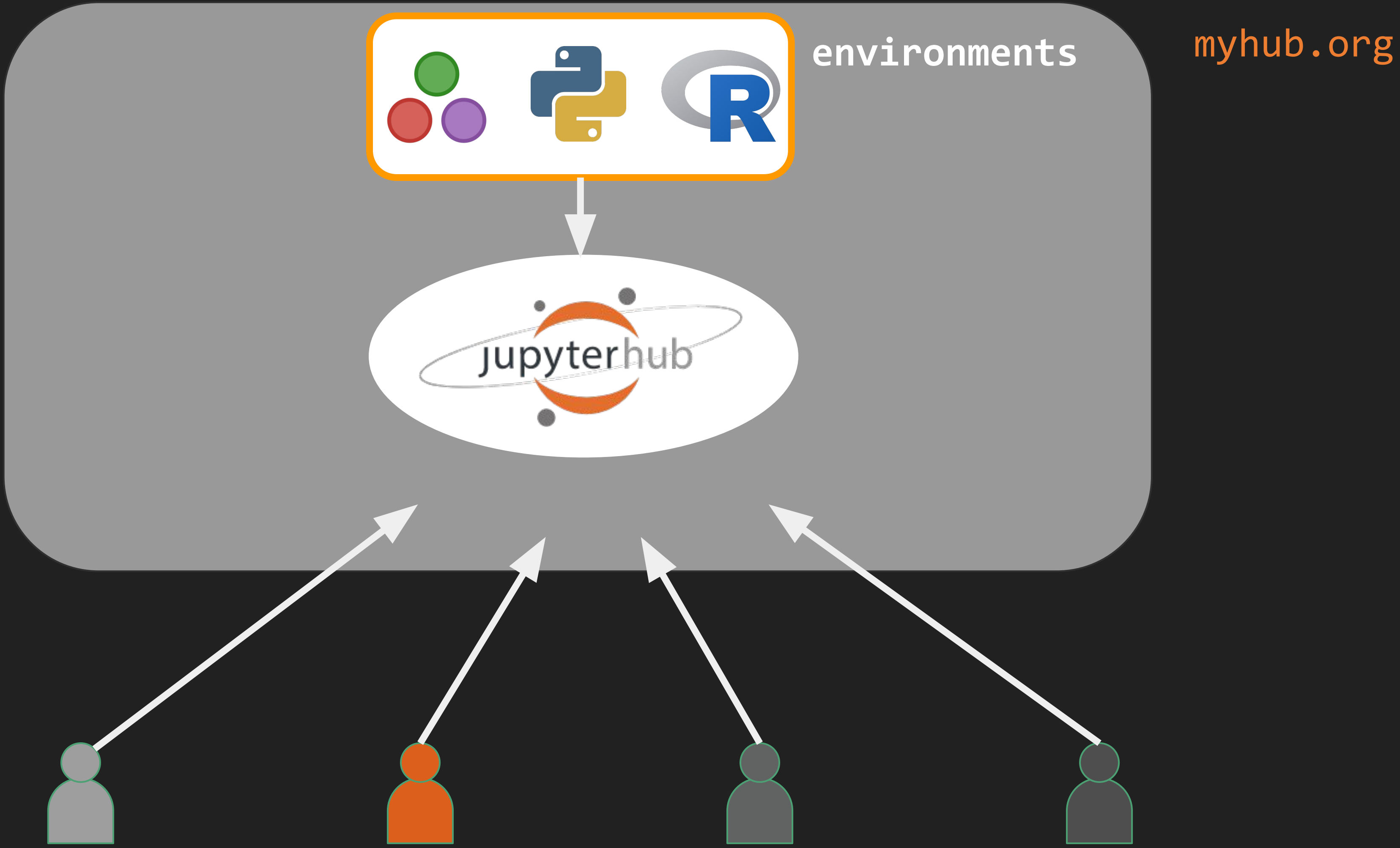
jupyter.org/hub

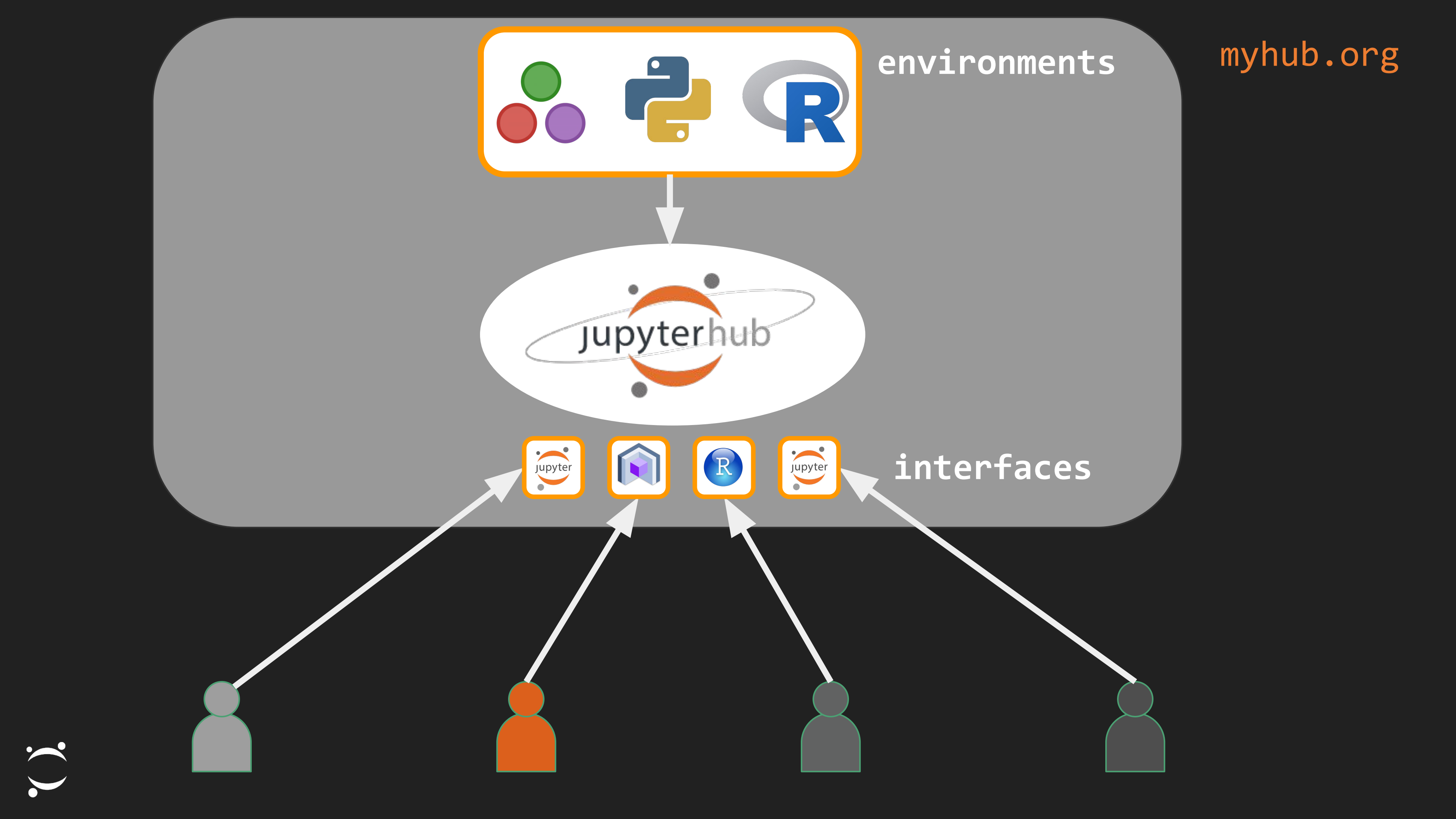
myhub.org

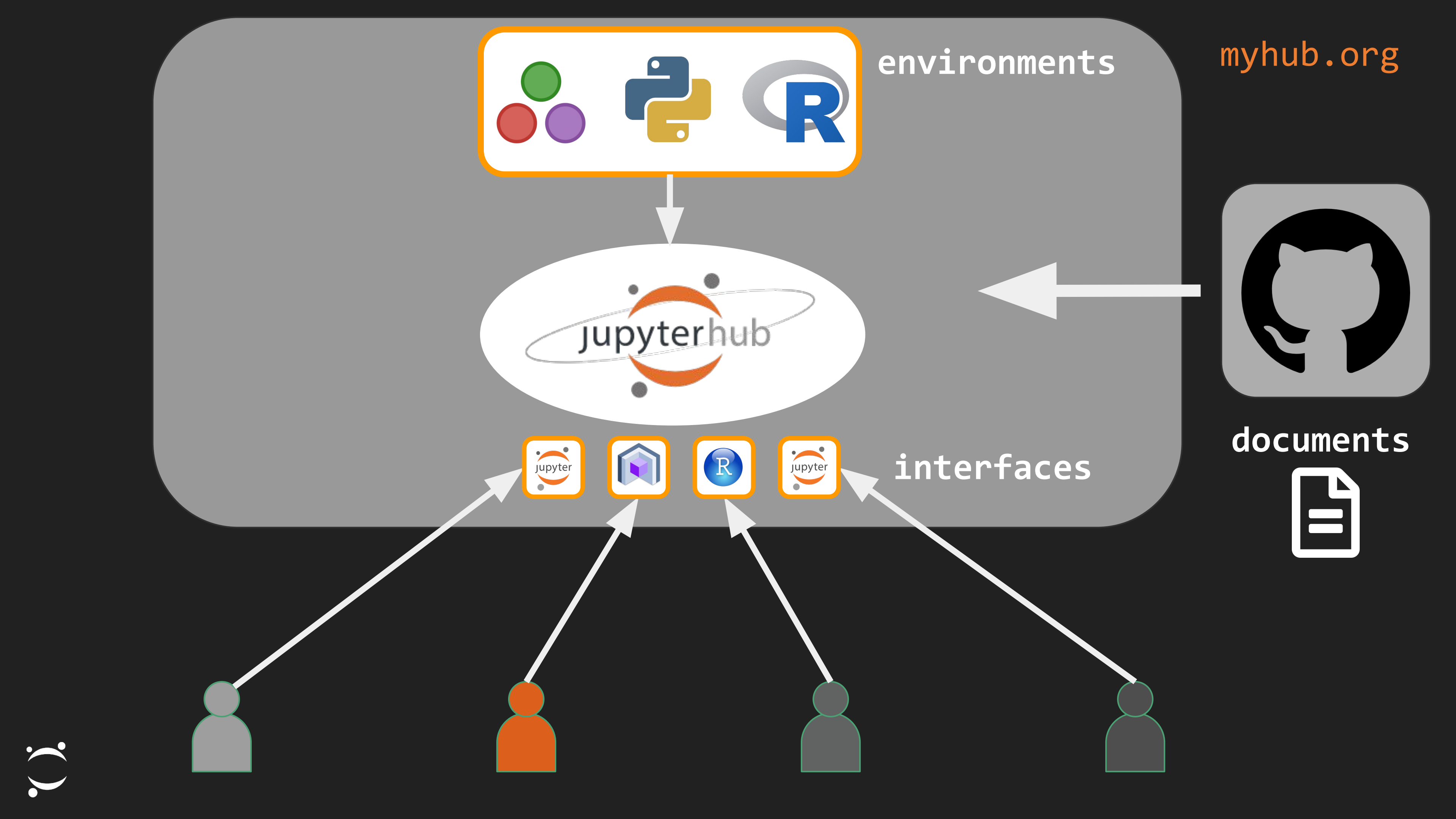
My fancy machine in the cloud

myhub.org

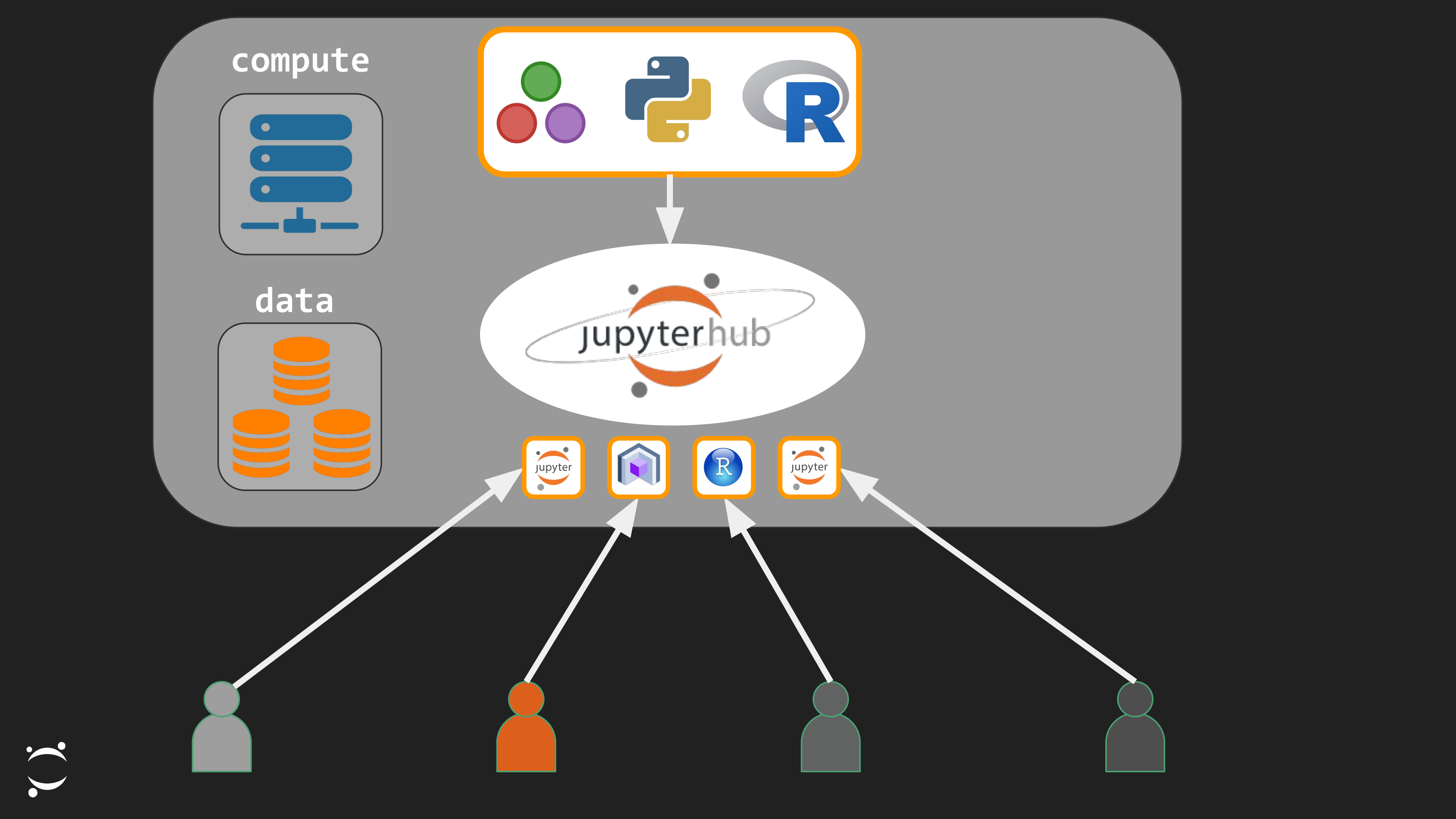


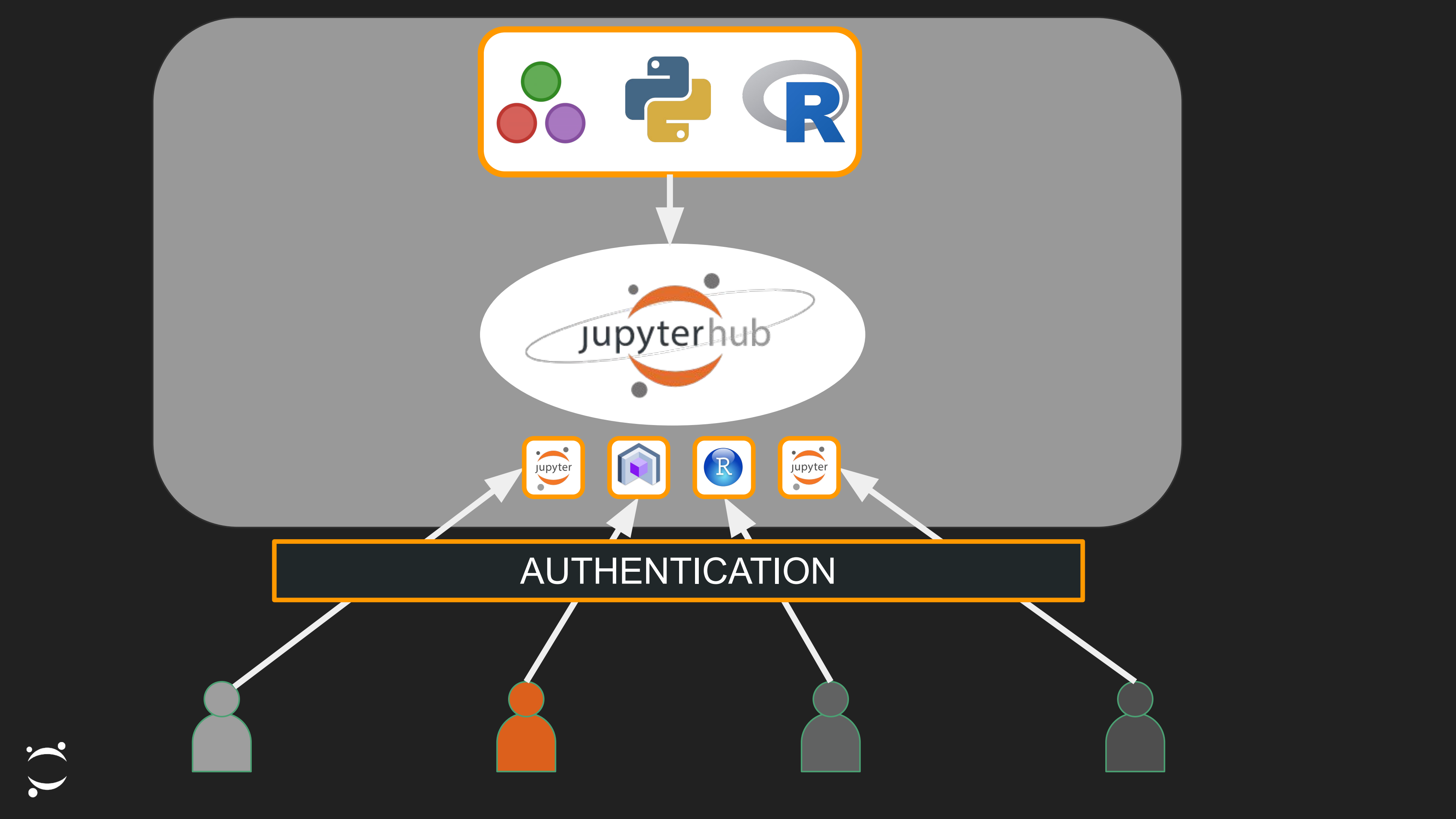

environments
myhub.org
jupyterhub


environments
jupyterhub
interfaces
jupyter
R
jupyter
myhub.org



environments
myhub.org
jupyterhub
documents
interfaces
jupyter
R
jupyter


compute
data
jupyterhub
jupyter
R
jupyter


jupyterhub
jupyter
jupyter
AUTHENTICATION

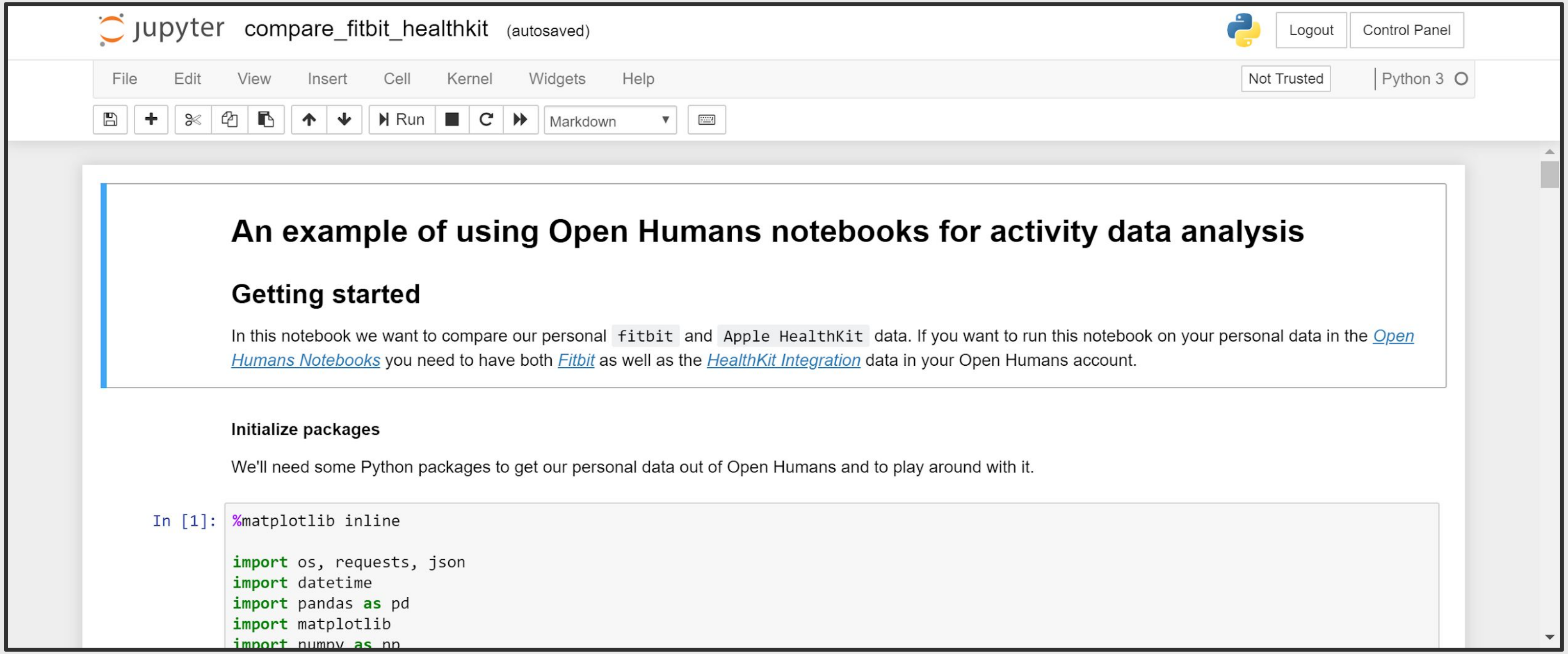
Chris Is Trying A Live Demo

Hopefully he doesn't embarrass himself too badly.

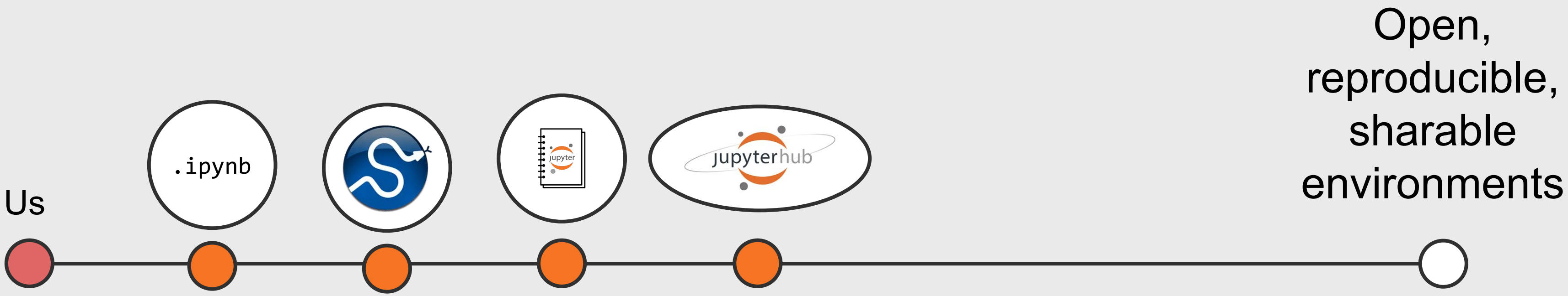
* Press ESC to cancel and return to Windows.
* Press ENTER to close this application that is not responding. You will lose any unsaved information in this application.
* Press CTRL+ALT+DEL again to restart your computer. You will lose any unsaved information in all applications.

[Textbook link]

[Interact link]



notebooks.openhumans.org


Us
.ipynb
jupyter
jupyterhub
Open,
reproducible,
sharable
environments

How can users package+share their work?

Part 2: packaging and sharing your environment with **repo2docker**

What is repo2docker?

Convert a repository into a Docker image that runs the code inside.

repo2docker.readthedocs.io

github.com/minrk/ligo-binder

repo2docker

what does it do?

```
$ jupyter repo2docker \
>       https://github.com/minrk/ligo-binder
```

repo2docker *what does it do?*

```
$ jupyter repo2docker \
>       https://github.com/minrk/ligo-binder

Cloning into
'/var/folders/.../T/repo2dockermu6z66sd'...
Using CondaBuildPack builder
Step 1/31 : FROM buildpack-deps:bionic
 ---> 29f4eef41002
Step 2/31 : ENV DEBIAN_FRONTEND=noninteractive
 ---> Using cache
 ---> ee1ba7c4f5f4
Step 3/31 : RUN apt-get update &&     apt-get
install --yes --no-install-recommends locales &&
apt-get purge &&     apt-get clean &&     rm -rf
/var/lib/apt/lists/*
```

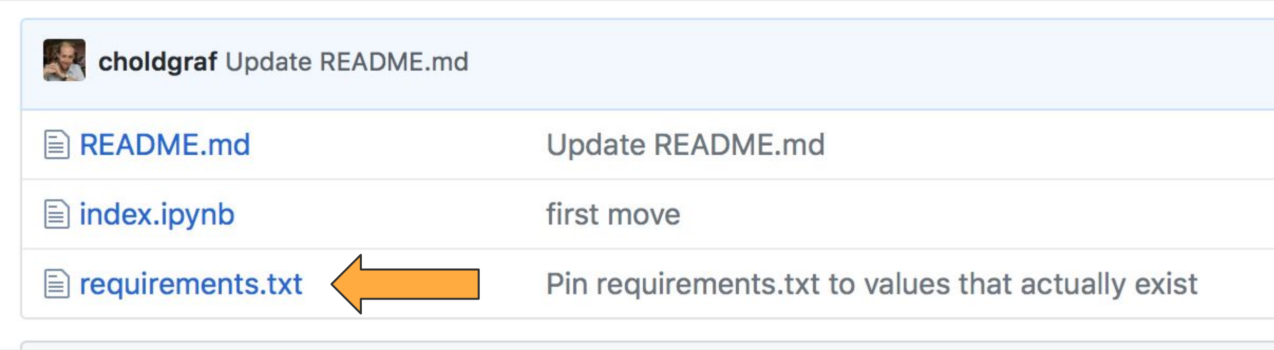
repo2docker
what does it do?

Step 1: get the **repo**

```
git clone https://github.com/me/myproject
```

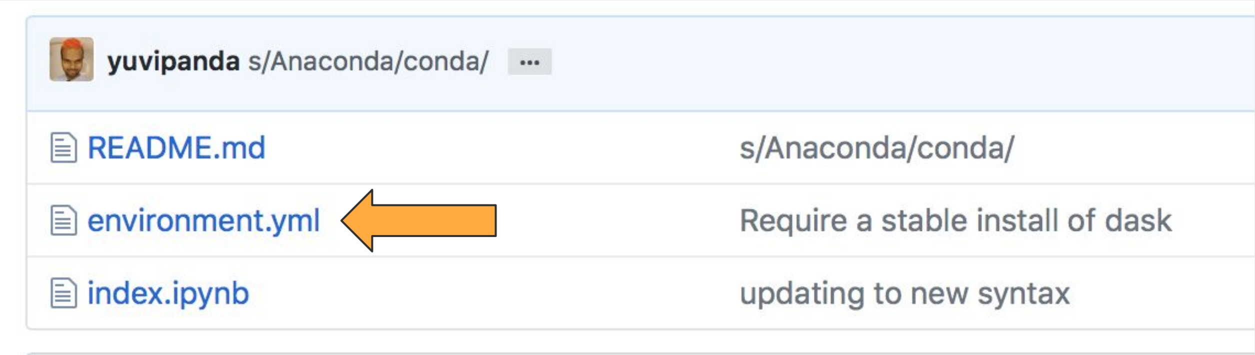
repo2docker
what does it do?

Step 2: Identify **requirements**



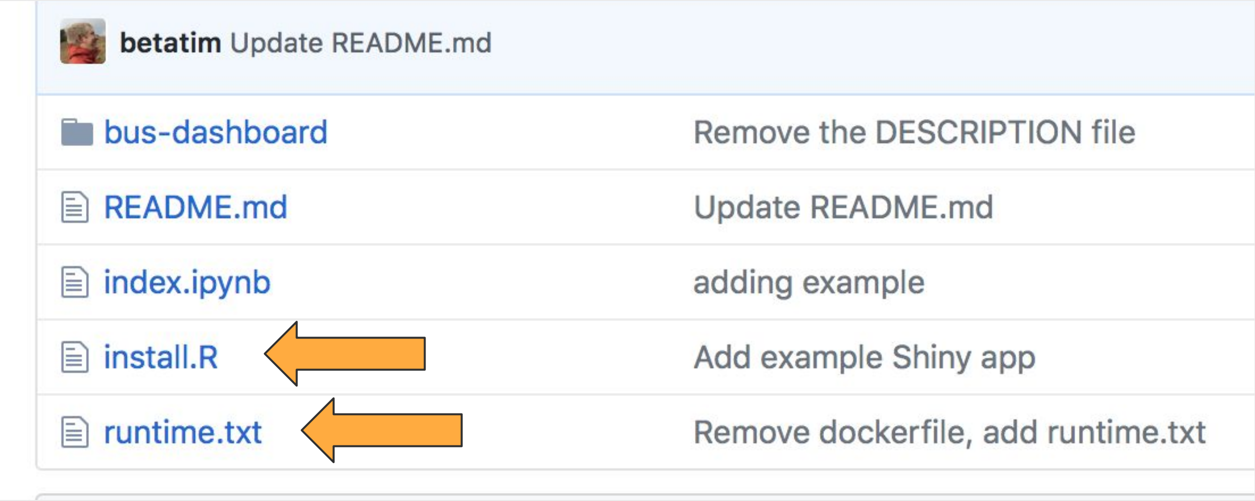
Step 2: Identify **requirements**

repo2docker
what does it do?



repo2docker
what does it do?

Step 2: Identify **requirements**



repo2docker
what does it do?

Step 3: generate Dockerfile

setup the runtime environment

```
...
COPY conda/install-miniconda.bash /tmp/install-miniconda.bash
COPY conda/environment.py-3.6.frozen.yml /tmp/environment.yml
RUN bash /tmp/install-miniconda.bash && \
rm /tmp/install-miniconda.bash /tmp/environment.yml
...
```

repo2docker
what does it do?

Step 3: generate **Dockerfile**

assemble the environment for the repo

```
...
# Copy and chown stuff.
COPY src/ ${HOME}
RUN chown -R ${NB_USER}:${NB_USER} ${HOME}

# Run assemble scripts! These will actually build the spec
# in the repository into the image.
USER ${NB_USER}
RUN ${KERNEL_PYTHON_PREFIX}/bin/pip install --no-cache-dir \
    -r "requirements.txt"
...
```

repo2docker *what does it do?*

Step 4: **build** (& **push**) image

```
docker build -t myimage
docker push myimage
```

repo2docker *what does it do?*

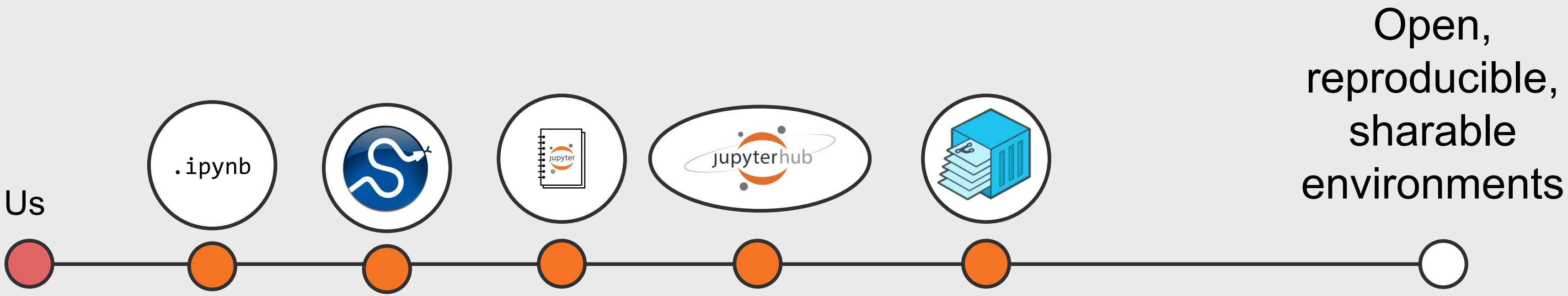
```
$ jupyter repo2docker \
>       https://github.com/minrk/ligo-binder

Cloning into
'/var/folders/.../T/repo2dockermu6z66sd'...
Using CondaBuildPack builder
Step 1/31 : FROM buildpack-deps:bionic
 ---> 29f4eef41002
Step 2/31 : ENV DEBIAN_FRONTEND=noninteractive
 ---> Using cache
 ---> ee1ba7c4f5f4
Step 3/31 : RUN apt-get update &&     apt-get
install --yes --no-install-recommends locales &&
apt-get purge &&     apt-get clean &&     rm -rf
/var/lib/apt/lists/*
```

Some supported configuration files

- `environment.yml` - conda
- `requirements.txt` - pip
- `REQUIRE` - julia
- `install.R` - R
- `apt.txt` - apt
- `setup.py` - pip
- `postBuild` - shell
- `runtime.txt`
- `Dockerfile` - docker
- **`yournewbuildpack.txt`**

Guiding principles of repo2docker

- Repos should be **human** and **machine readable**
- Use **existing specifications and standards**
- Support **many languages** and interfaces
- Be lightweight and **tightly-scoped,** but **extendable**

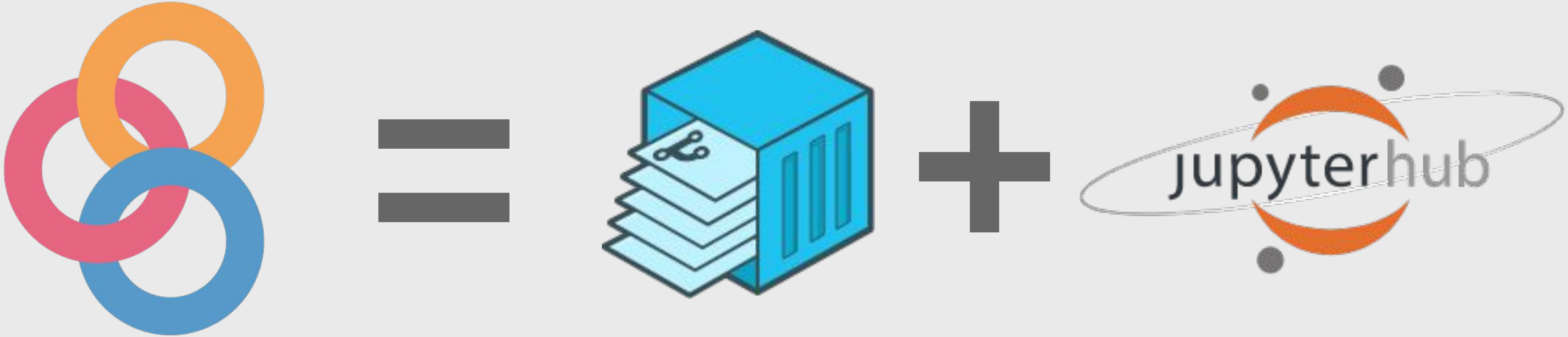

Us
.ipynb
jupyter
jupyterhub
Open,
reproducible,
sharable
environments

Part 3: tying this together with **BinderHub**

What is BinderHub?

One-click sharable, interactive, reproducible environments from your public git repository

`mybinder.org`
`binderhub.readthedocs.io`


jupyterhub

BinderHub is open tech...

- Built on Kubernetes
- Cloud-agnostic
- Scalable
- **Community driven and deployable by anyone**

One example: mybinder.org

Chris Is Trying A Live Demo

Hopefully he doesn't embarrass himself too badly.

* Press ESC to cancel and return to Windows.
* Press ENTER to close this application that is not responding. You will lose any unsaved information in this application.
* Press CTRL+ALT+DEL again to restart your computer. You will lose any unsaved information in all applications.

mybinder.org

binder-examples/requirements

Project Jupyter Follow
Project Jupyter exists to develop open-source software, open standards, and services for interactive and reproducible computing.
Feb 20 · 5 min read

JupyterLab is Ready for Users

We are proud to announce the beta release series of JupyterLab, the next-generation web-based interface for Project Jupyter.

tl;dr: **JupyterLab is ready for daily use (installation, documentation, try it with Binder)**

Mybinder.org is an open service…

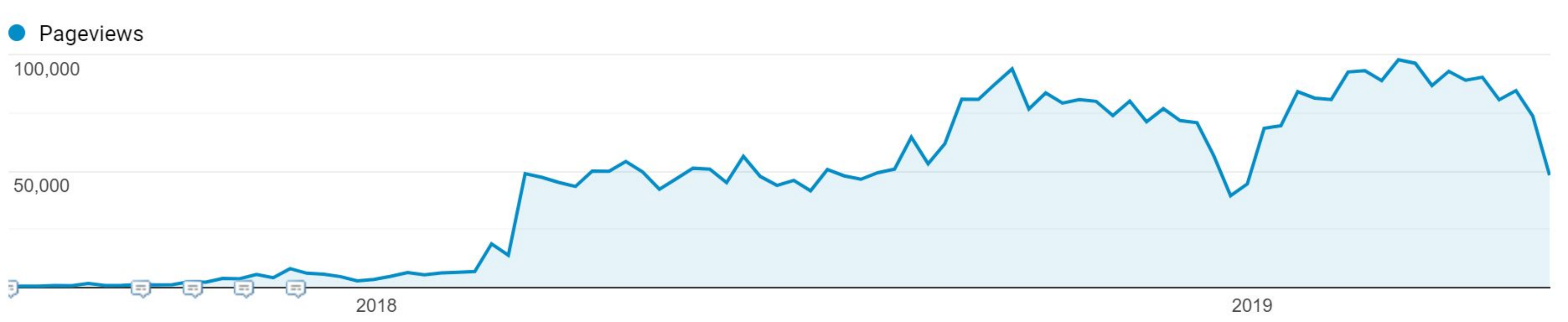
Mybinder.org activity and status:
grafana.mybinder.org/

Our billing:
github.com/jupyterhub/binder-billing

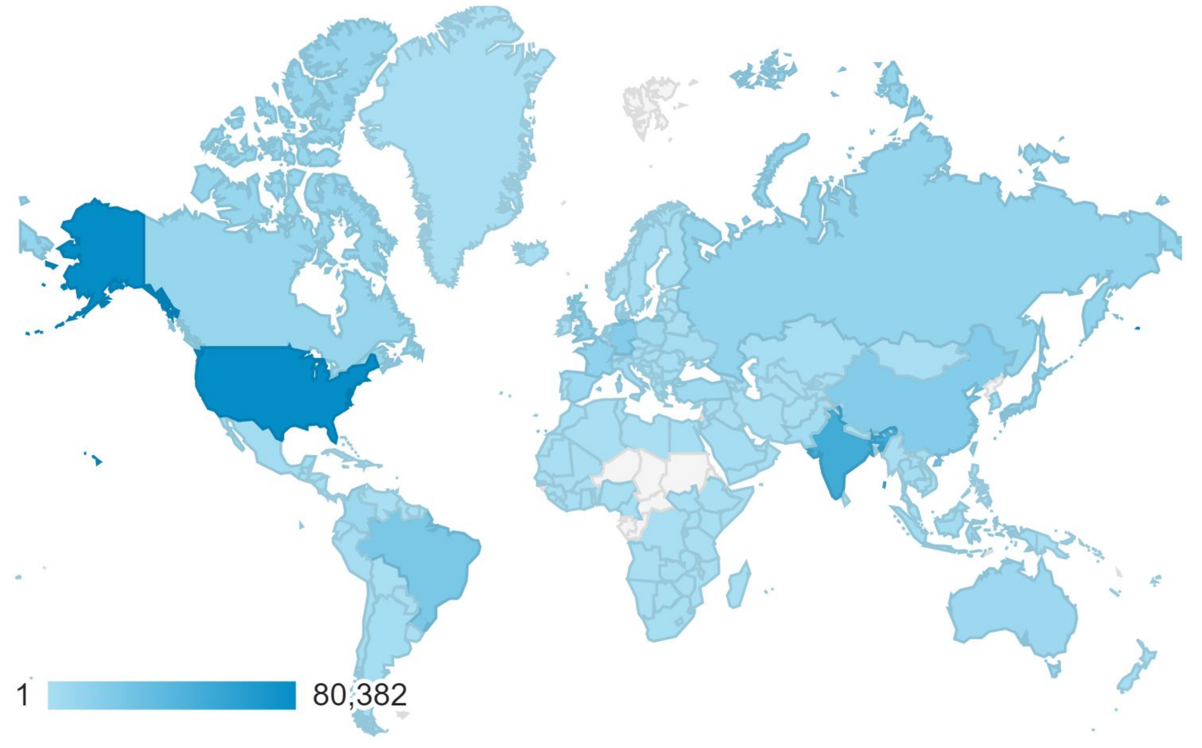
Our operations guide:
mybinder-sre.readthedocs.io/

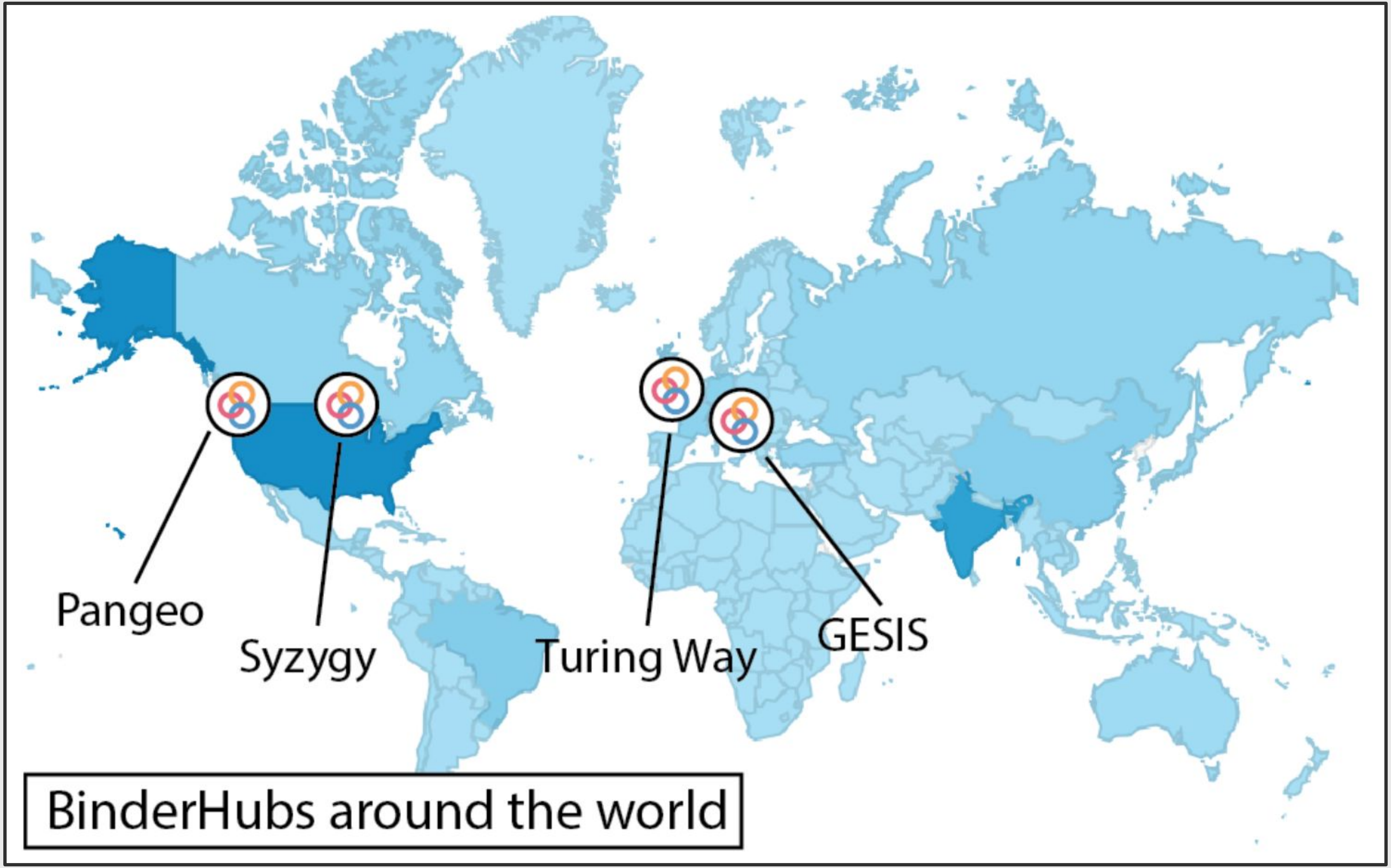
Our fires:
mybinder-sre.readthedocs.io/en/latest/#incident-reports

mybinder.org weekly sessions, last ~year



mybinder.org sessions, last month



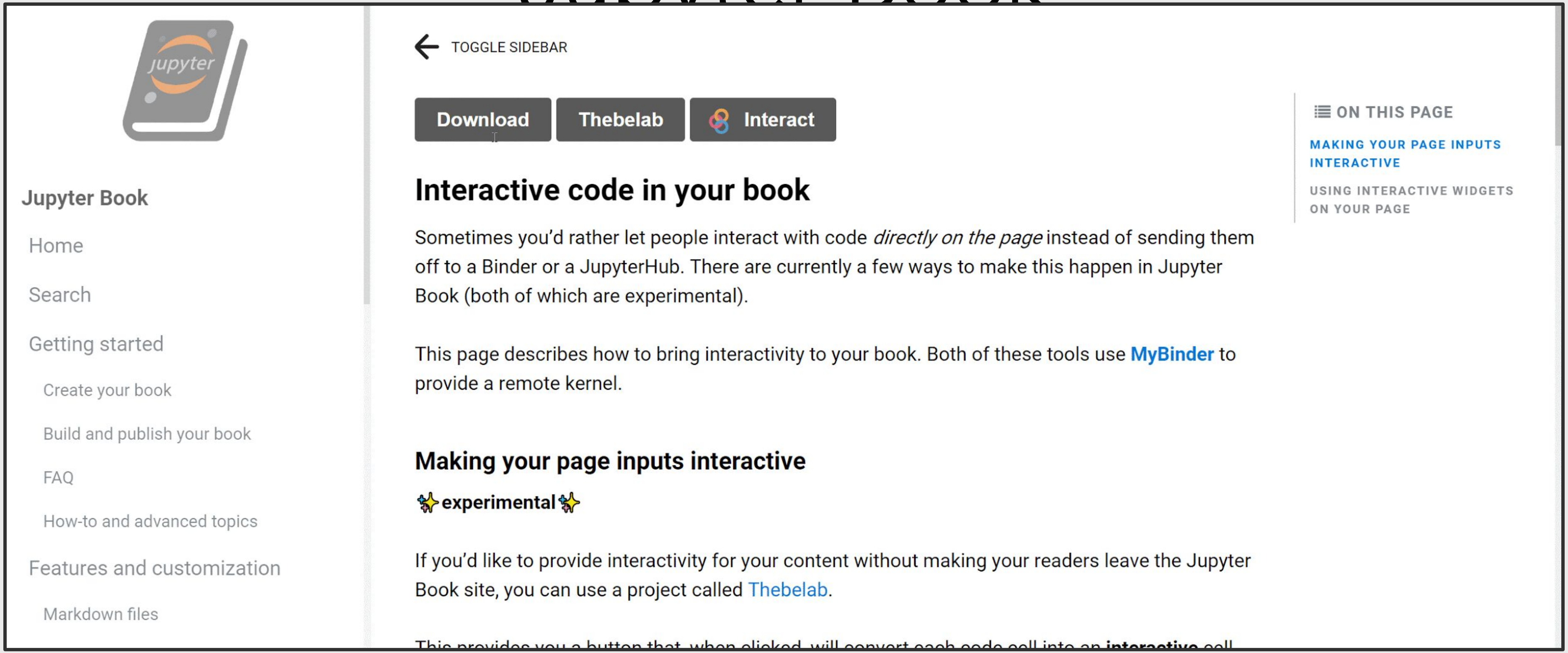

Pangeo
Syzygy
Turing Way
GESIS
BinderHubs around the world

Some cool new projects

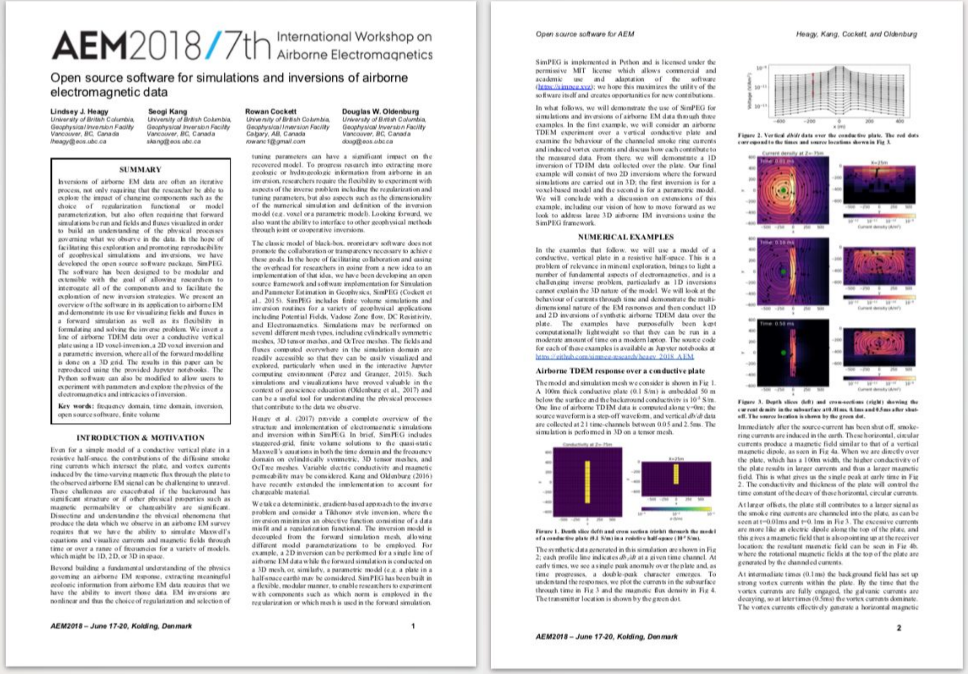
(or, stuff you can help out with)

Interactive books with Jupyter-Book

github.com/jupyter/jupyter-book



Publishable documents (with pandoc?)



AEM2018/7th International Workshop on Airborne Electromagnetics

Open source software for simulations and inversions of airborne electromagnetic data

Lindsey J. Heagy, Seogi Kang, Rowan Cockett, Douglas W. Oldenburg

SUMMARY

INTRODUCTION & MOTIVATION

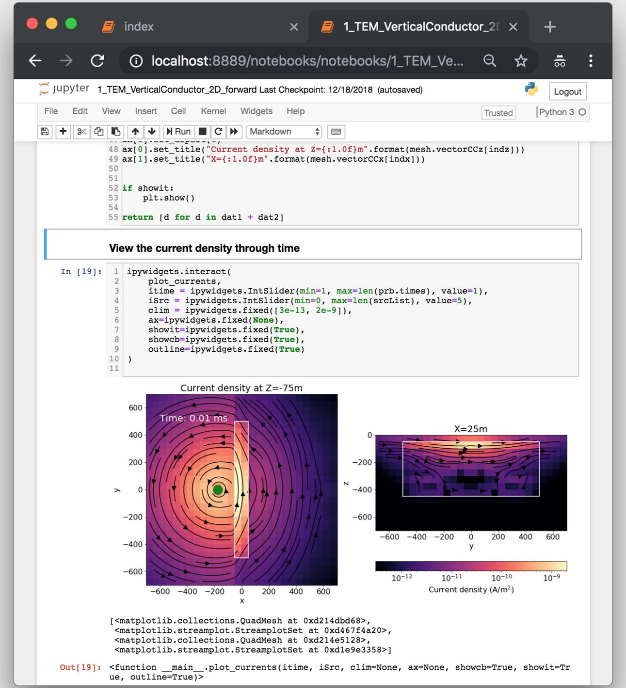
AEM2018 – June 17-20, Kolding, Denmark 1

Open source software for AEM — Heagy, Kang, Cockett, and Oldenburg

NUMERICAL EXAMPLES

Airborne TDEM response over a conductive plate

AEM2018 – June 17-20, Kolding, Denmark 2



https://github.com/simpeg-research/heagy-2018-aem

HTML dashboards with voila

github.com/QuantStack/voila

jupyter nbviewer

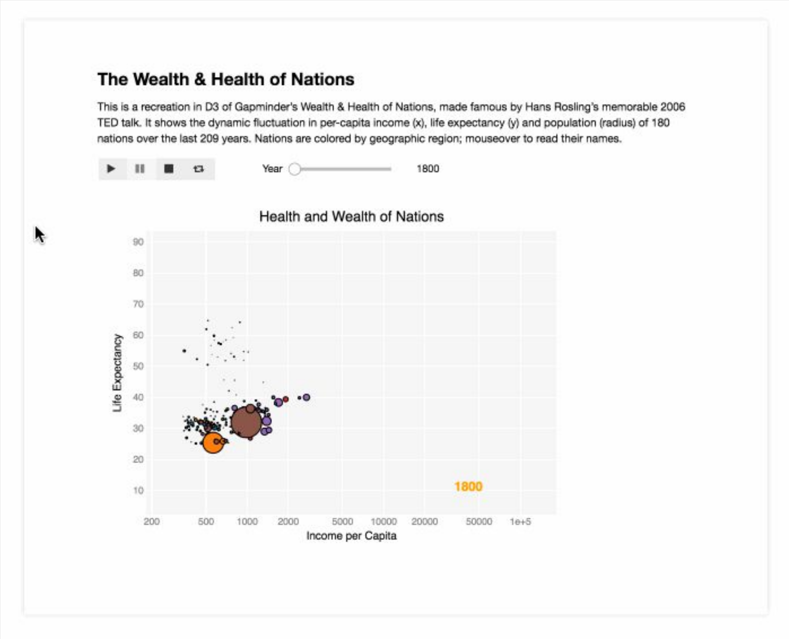
voila / notebooks

In [ ]:

```
import warnings
warnings.filterwarnings('ignore')
```

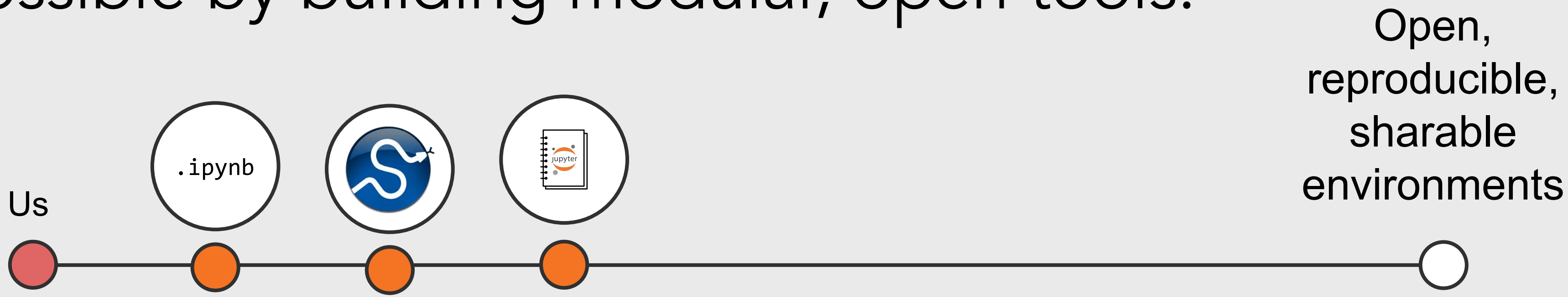
In [ ]:

```
import numpy as np
from bqplot import pyplot as plt

plt.figure(1, title='Line Chart')
np.random.seed(0)
n = 200
x = np.linspace(0.0, 10.0, n)
y = np.cumsum(np.random.randn(n))
plt.plot(x, y)
plt.show()
```



In summary

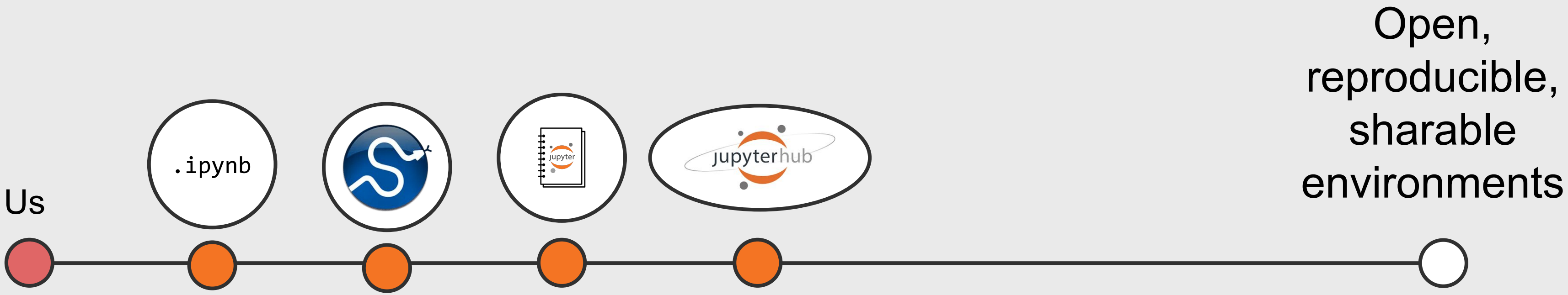
Jupyter makes the "last-mile" problem as small as possible by building modular, open tools.



jupyter.org

In summary

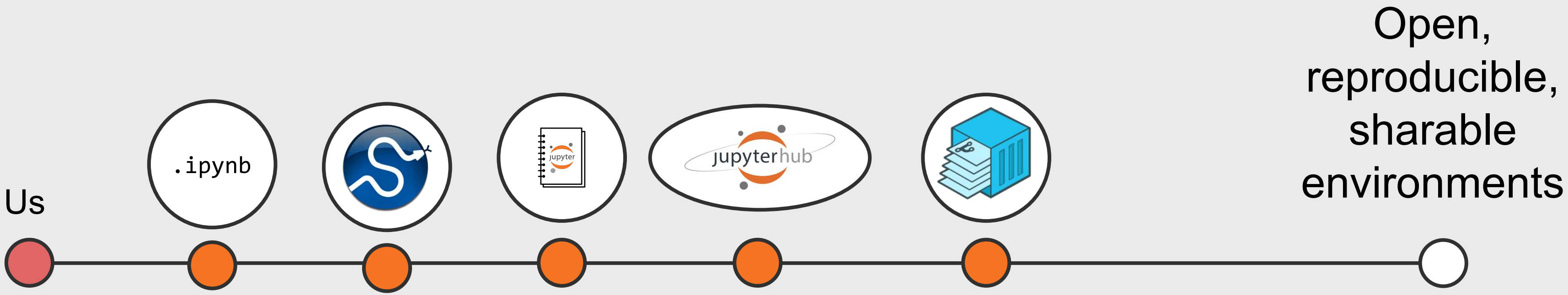
JupyterHub lets you create a shared, interactive analytics environment



mybinder.org

In summary

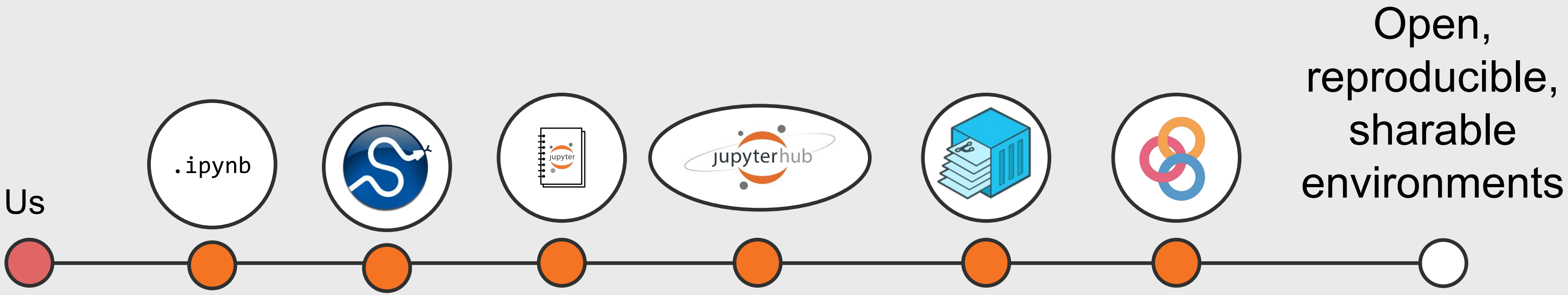
repo2docker creates reproducible Docker images from a repository



`mybinder.org`

In summary

BinderHub is an open web application to create shareable, reproducible coding environments



Mybinder.org

Get involved with Jupyter

- All of these projects are open source, run by open communities
- Jupyter is a place where *anybody* can participate
- If you'd like to get involved:

jupyterhub-team-compass.readthedocs.io

discourse.jupyter.org

@choldgraf