# Scaling Reproducible Research

**Carol Willing**

@WillingCarol

Workshop: Building reproducible workflows for earth sciences

ECMWF

October 15, 2019

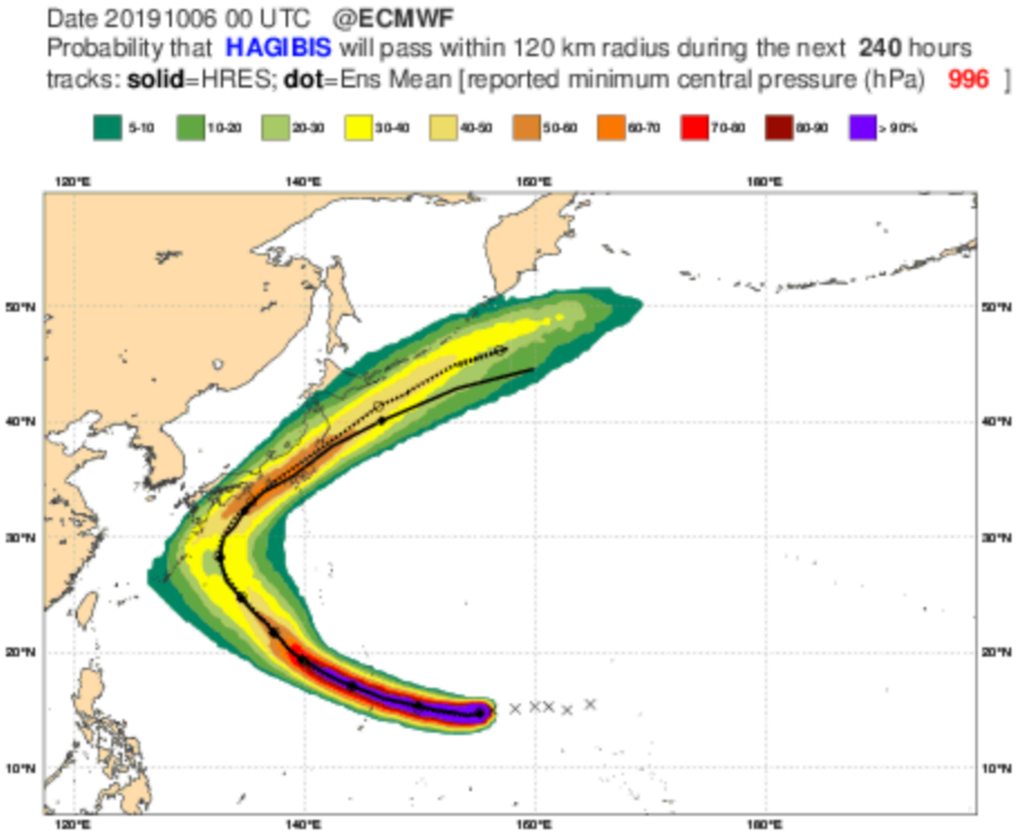
# San Diego, CA

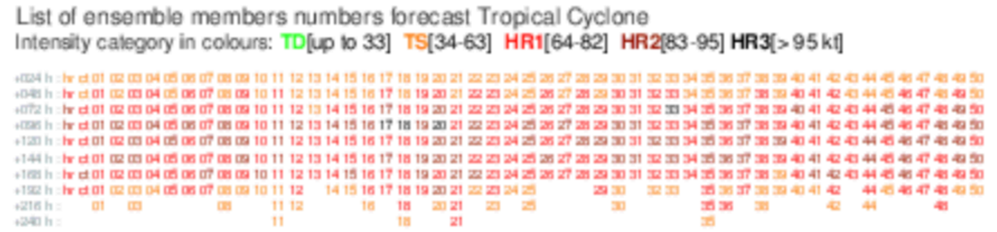
# Tokyo

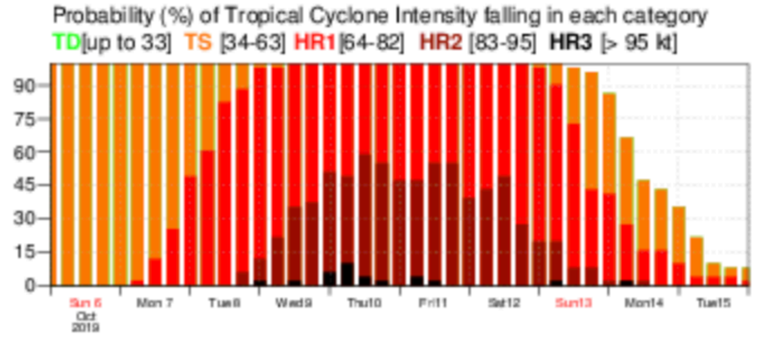
The **typhoon** came out of the sea first as a deep hollow roar.

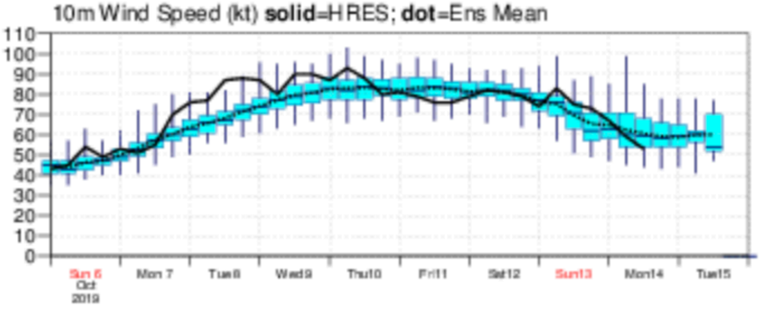
*–Pearl S. Buck*

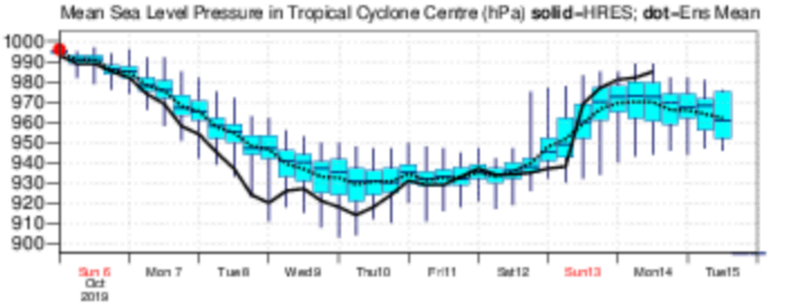
Base time: Sun 06 Oct 2019 00 UTC



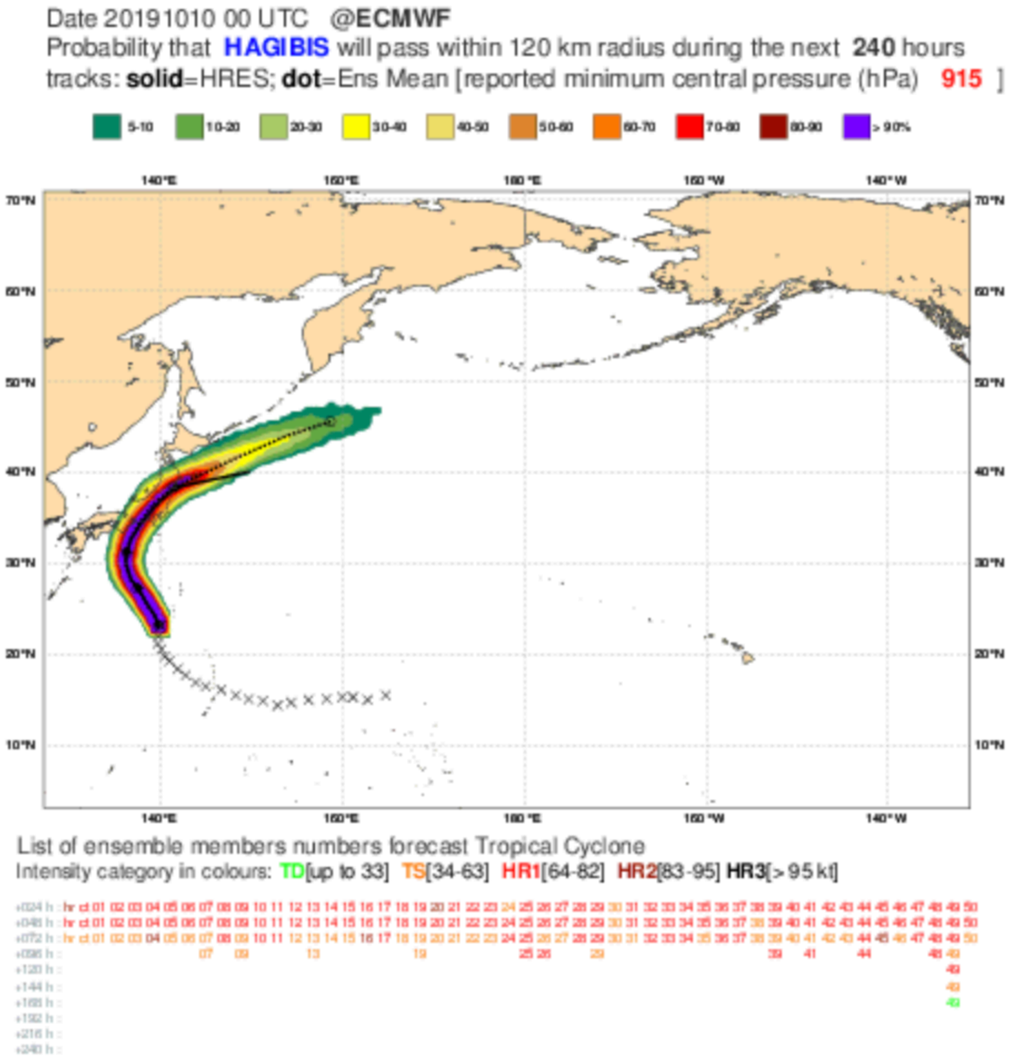



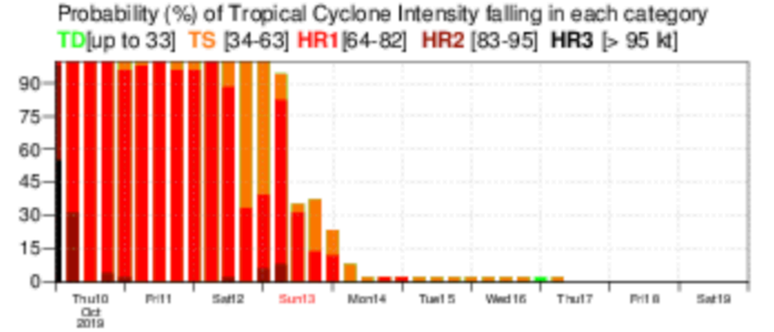



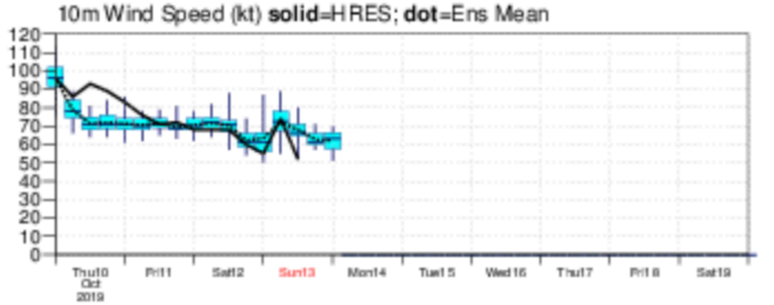



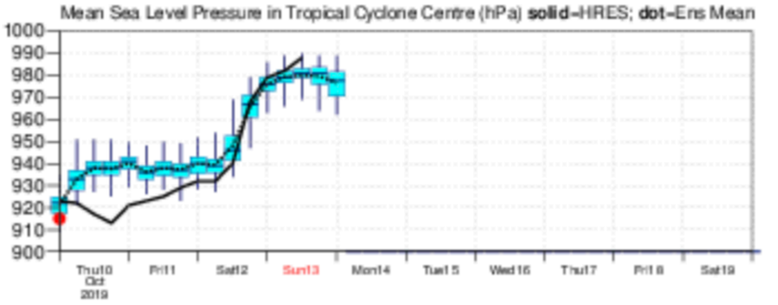

These charts show the evolution of the position and intensity of tropical cyclones (TCs) in the ECMWF high-resolution (HRES) and ensemble (ENS) forecasts. They provide an objective measure of the uncertainty in the current forecast. These products are generated for all TCs that have been officially observed at the initial time of the forecast.

Base time: Thu 10 Oct 2019 00 UTC




These charts show the evolution of the position and intensity of tropical cyclones (TCs) in the ECMWF high-resolution (HRES) and ensemble (ENS) forecasts. They provide an objective measure of the uncertainty in the current forecast. These products are generated for all TCs that have been officially observed at the initial time of the forecast.

**Linnea/リンネ**

Dunno how much the US news is covering typhoon 19, but it's pretty bad. Mandatory evacuation notices for the Hino districts near the river, but none for my area yet.

8:57 AM

**Linnea/リンネ**

Things are okay here so far, and my room's on the 4th floor with pretty thick wire-enforced windows so I think it's only the relatively safe side all things considered

8:58 AM

**Linnea/リンネ**

Might head to the closest evacuation center depending on how the news looks getting closer to when the typhoon hits, but safe for now + prepared to book it if I have to

9:00 AM

**Linnea/リンネ**

I'll try to keep you guys posted

9:00 AM

**Linnea/リンネ**

the peak of the typhoon expected to hit in about 30 min & last for about 3 hours before it passes and hino's out of the woods

10:22 AM

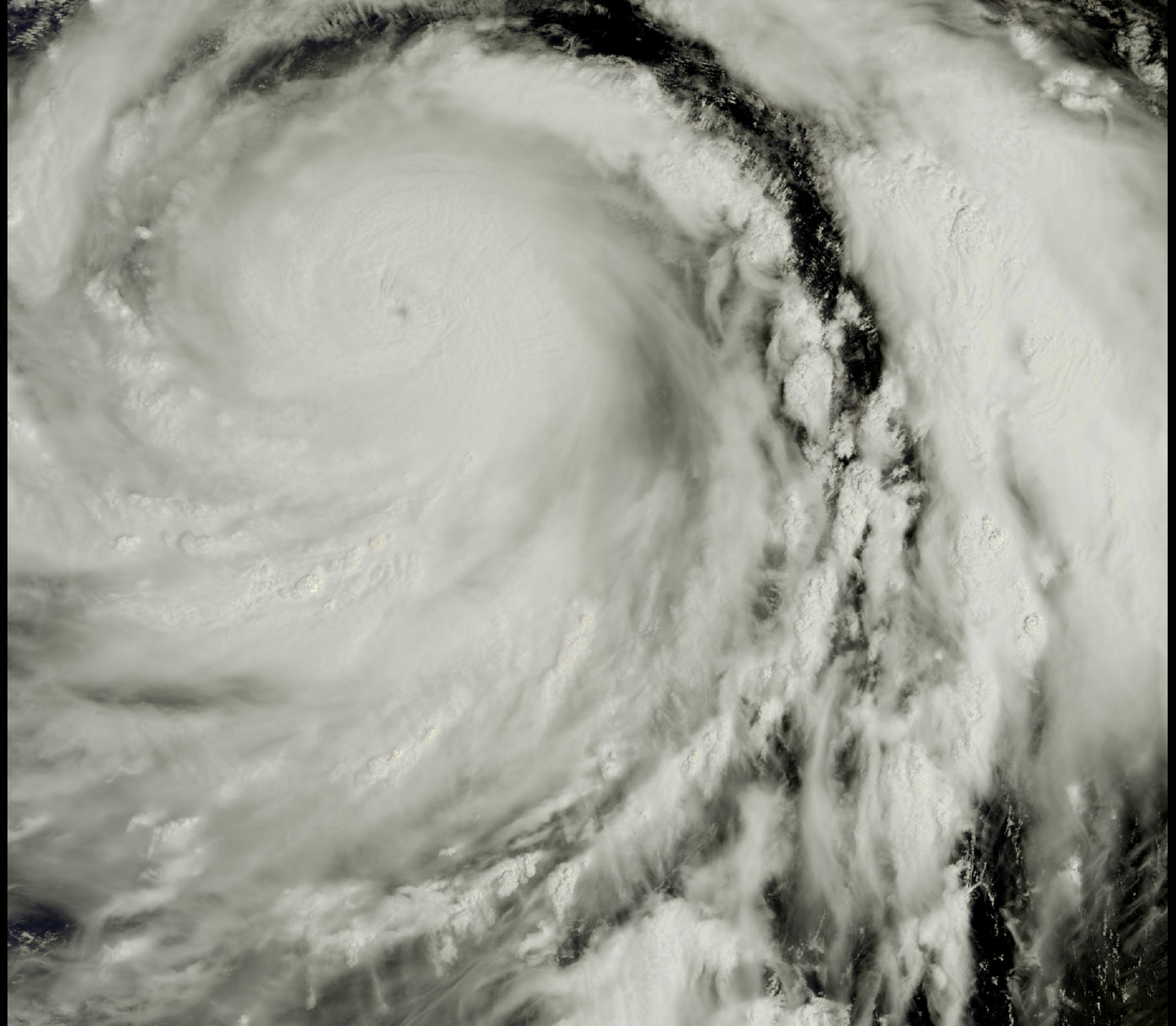
I was surrounded by the madness, the **unreason**, of **uncontrolled**, **undisciplined** energy.

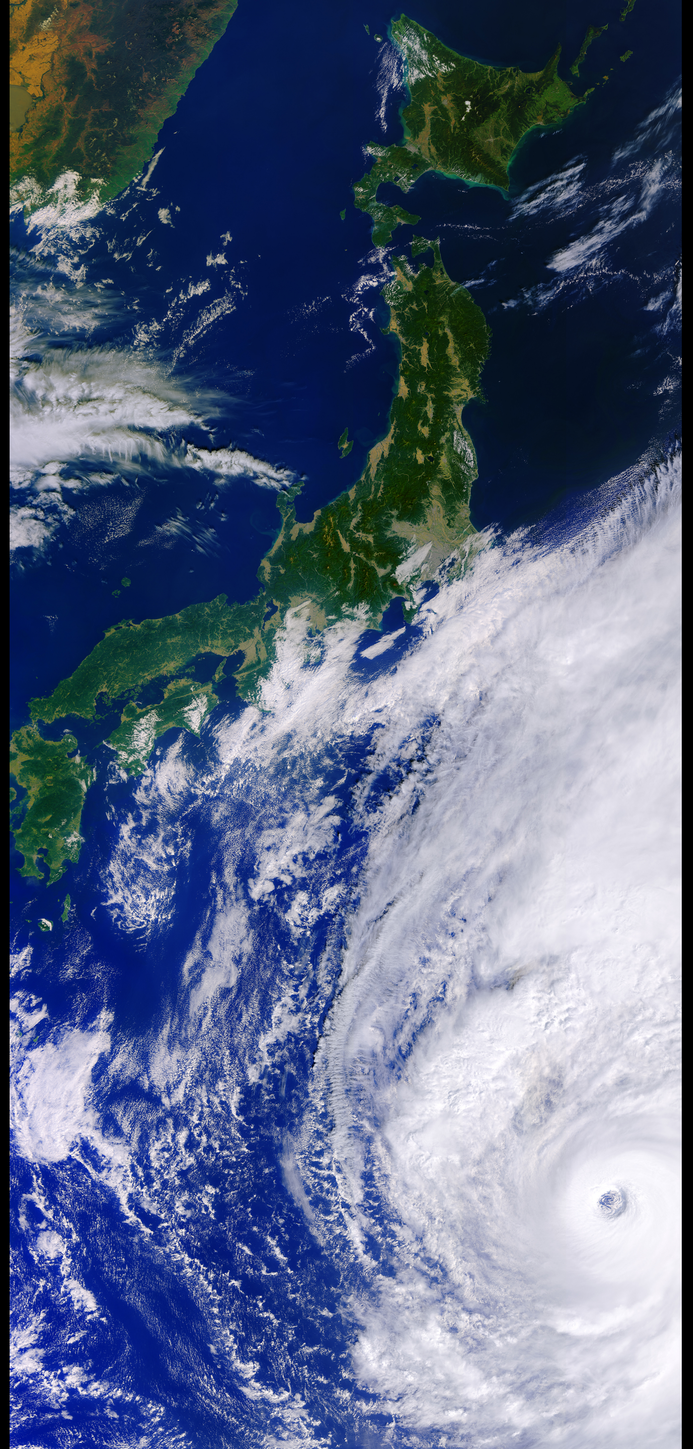
*–Pearl S. Buck*

# Super Typhoon Hagibis

View of Super Typhoon Hagibis south-west of Japan, as captured by the Copernicus Sentinel-3 satellite on 08 October at 00:16 UTC.

**Copyright: 2019 European Union, contains modified Copernicus Sentinel data 2019, processed by EUMETSAT**





Title Typhoon Hagibis

Released 10/10/2019 4:45 pm

Copyright contains modified Copernicus Sentinel data (2019), processed by ESA, CC BY-SA 3.0 IGO

Source:Twitter

A sign is partially submerged as the Tama River floods during Typhoon Hagibis. *Source:Getty Images*

**Linnea/リンネ**

In the end japan had a record typhoon, whirlwinds, tsunami, volcanic eruption, and earthquakes all at once

1:40 PM

Linnea/リンネ

Wind, water, earth, and fire. All that's left is for godzilla to crawl out of the bay and vaporize chiba

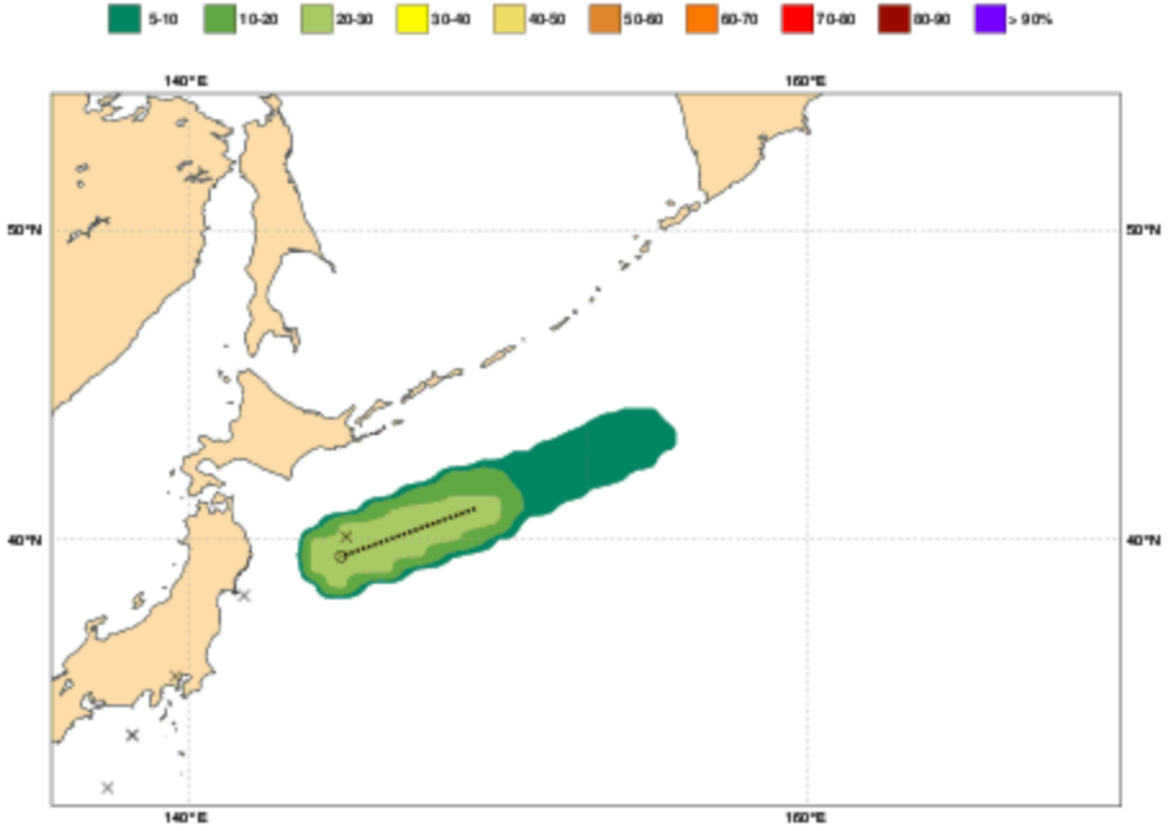
1:41 PM

That's quite the day. I am using that quote in my keynote

Read 2
1:41 PM

Base time: Sun 13 Oct 2019 00 UTC

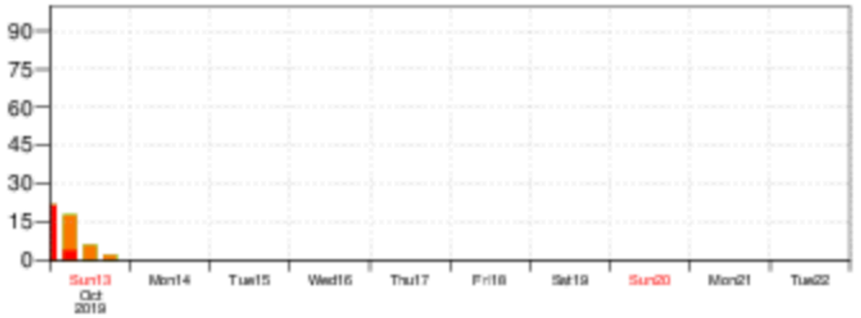
Date 20191013 00 UTC @**ECMWF**
Probability that **HAGIBIS** will pass within 120 km radius during the next **240** hours
tracks: **solid**=HRES; **dot**=Ens Mean [reported minimum central pressure (hPa) **975** ]



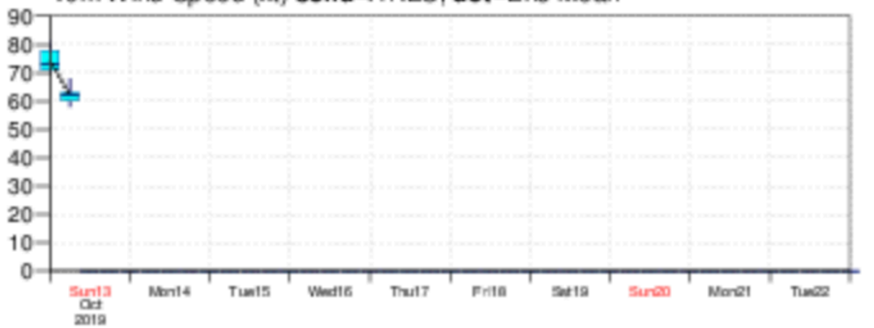

List of ensemble members numbers forecast Tropical Cyclone
Intensity category in colours: TD[up to 33] TS[34-63] HR1[64-82] HR2[83-95] **HR3**[> 95 kt]

+024 h :
+048 h :
+072 h :
+096 h :
+120 h :
+144 h :
+168 h :
+192 h :
+216 h :
+240 h :

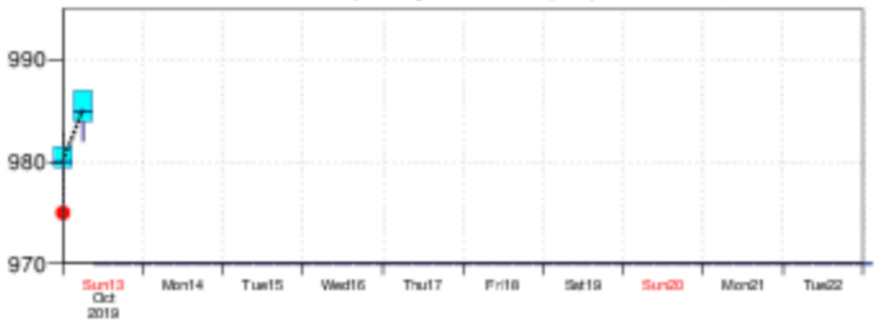
Probability (%) of Tropical Cyclone Intensity falling in each category
TD[up to 33] TS [34-63] HR1[64-82] HR2 [83-95] **HR3** [> 95 kt]




10m Wind Speed (kt) **solid**=HRES; **dot**=Ens Mean




Mean Sea Level Pressure in Tropical Cyclone Centre (hPa) **solid**=HRES; **dot**=Ens Mean




@Willir

# Lives depend on

# scaling reproducible research

# Tools

# Processes

# Communication

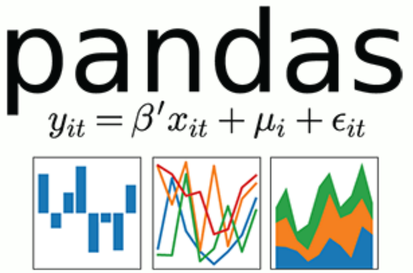
PyMC3

BLOSC

CONDA-FORGE

sunpy

xarray

scikit
learn

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

julia

jupyter

IP[y]:
IPython

QuantEcon

Econ
ARK

ROpenSci

将軍
sho
gun

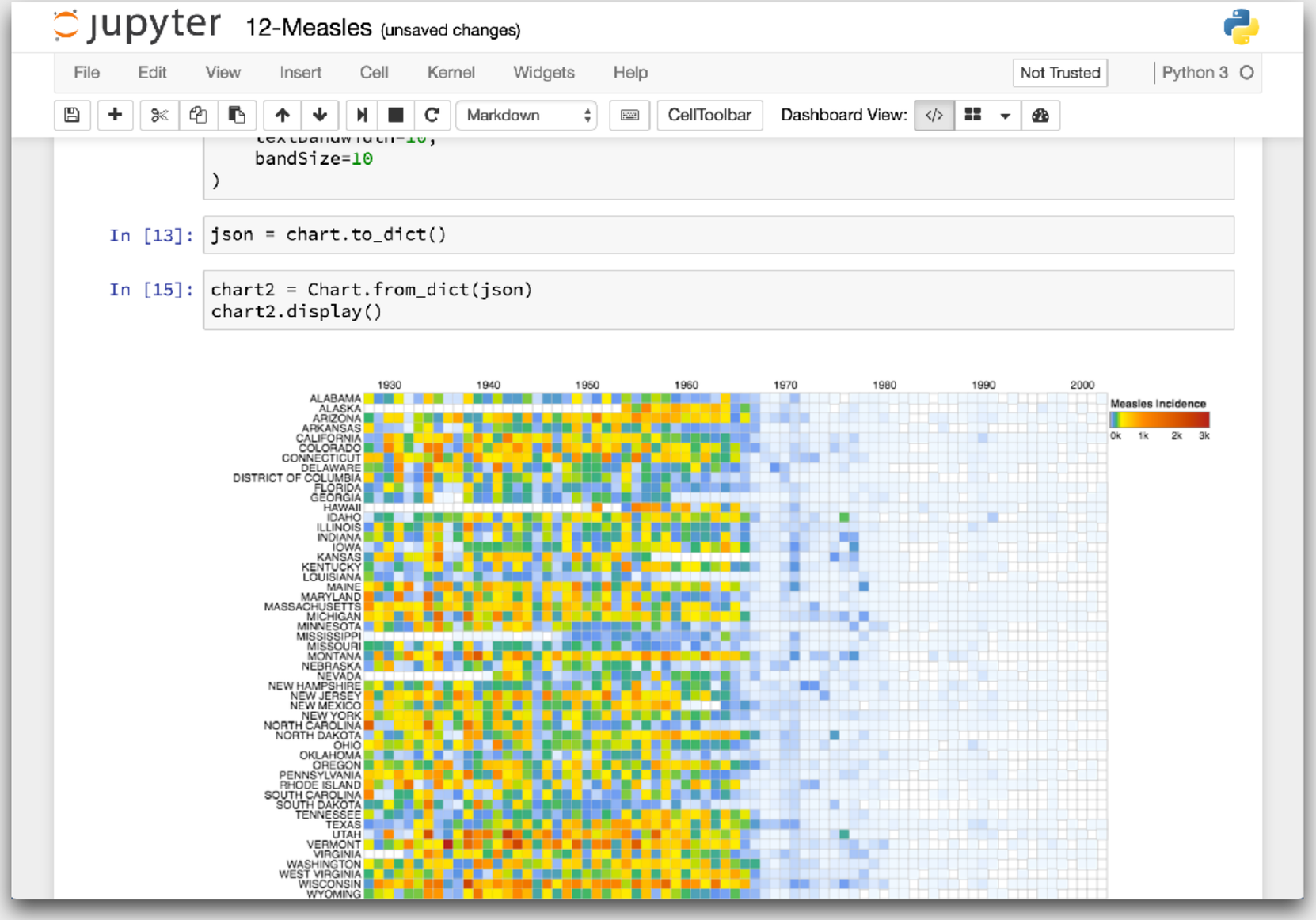
THE
CARPENTRIES

MathJax

PyTables

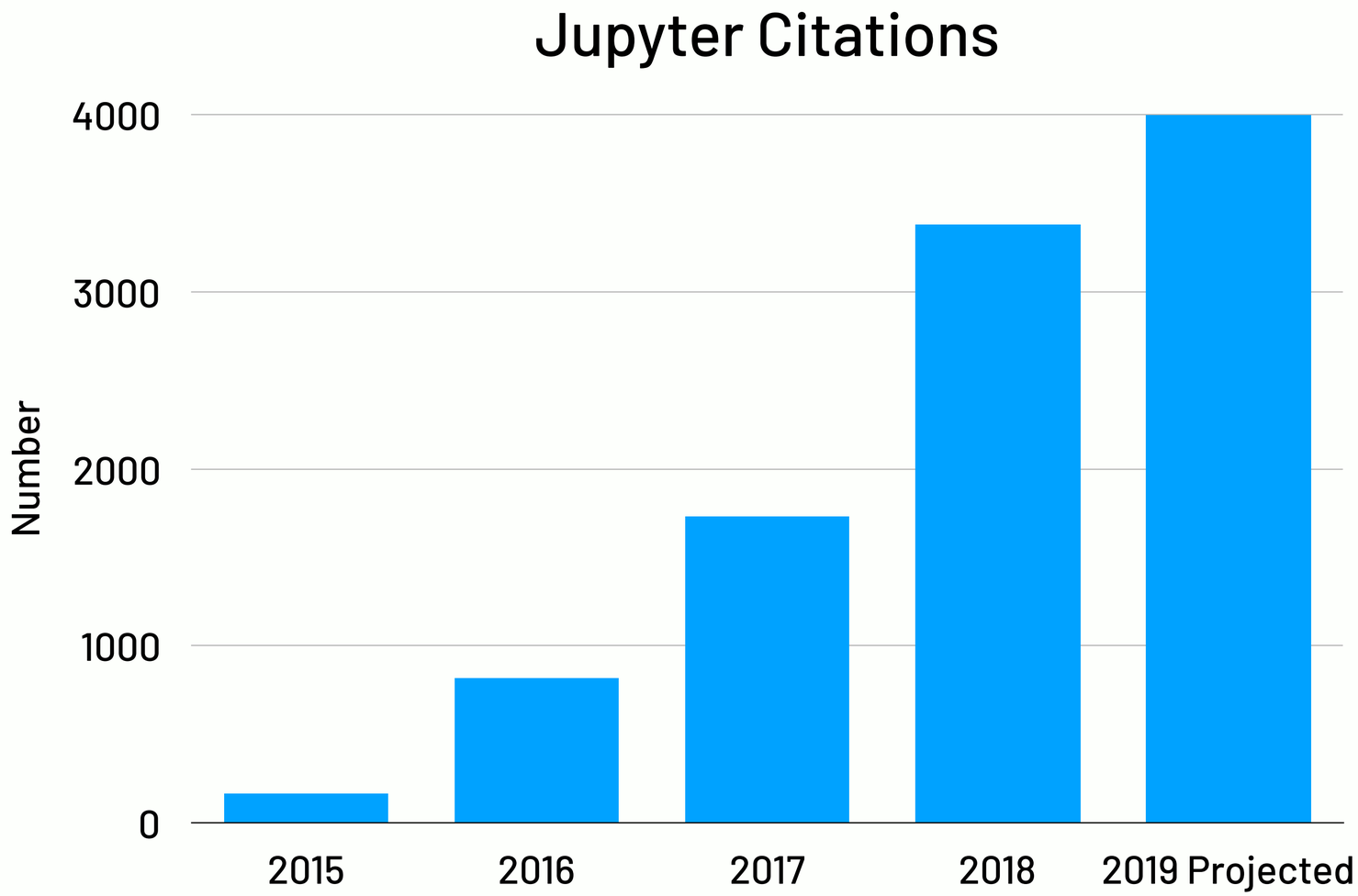
NUMFOCUS
OPEN CODE = BETTER SCIENCE

# Jupyter Notebook

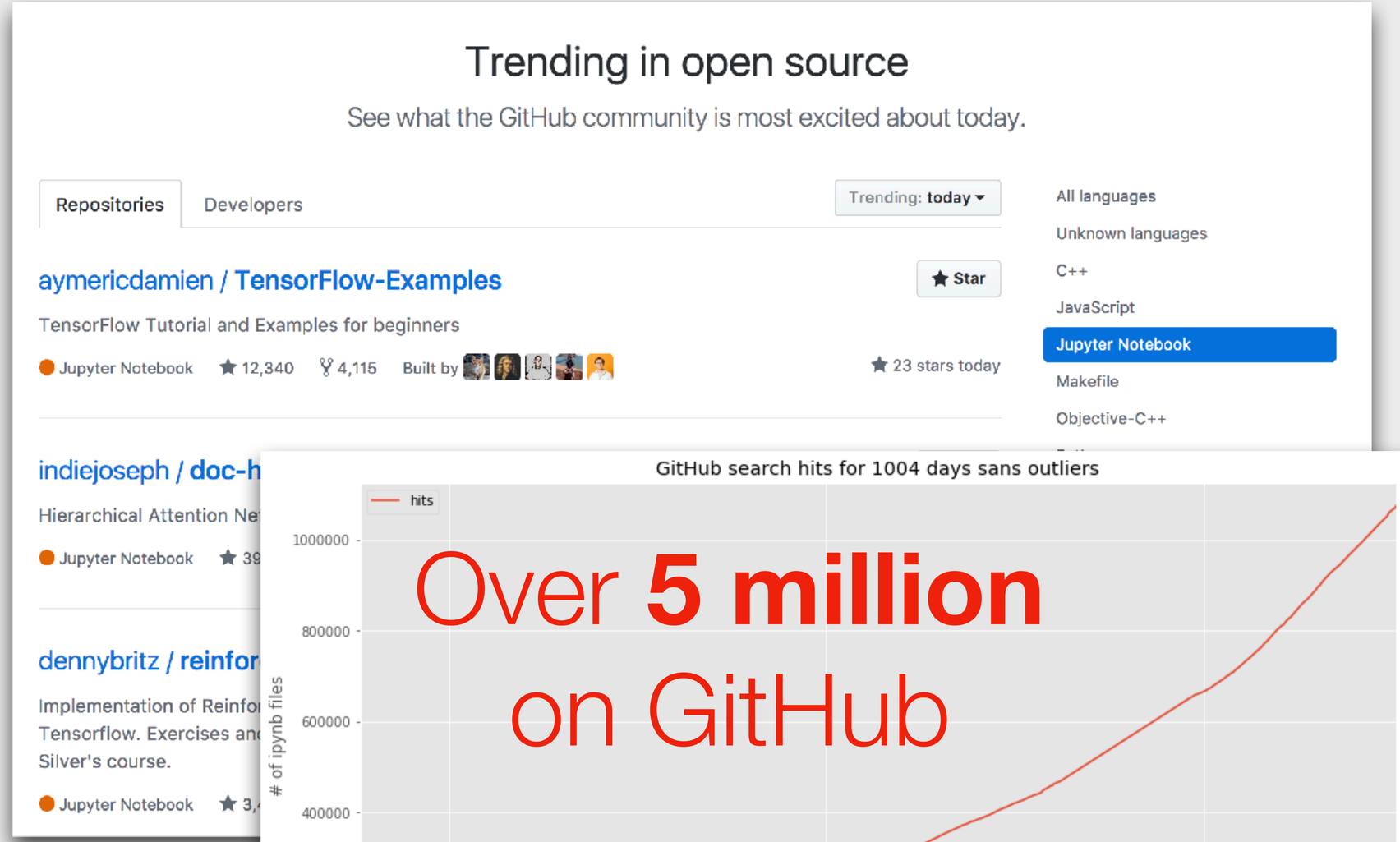



*A Jupyter Notebook document with a visualization of measles data.*

# Research


Jupyter Citations
Number
4000
3000
2000
1000
0
2015
2016
2017
2018
2019 Projected

# Millions of Notebooks




https://github.com/trending/jupyter-notebook

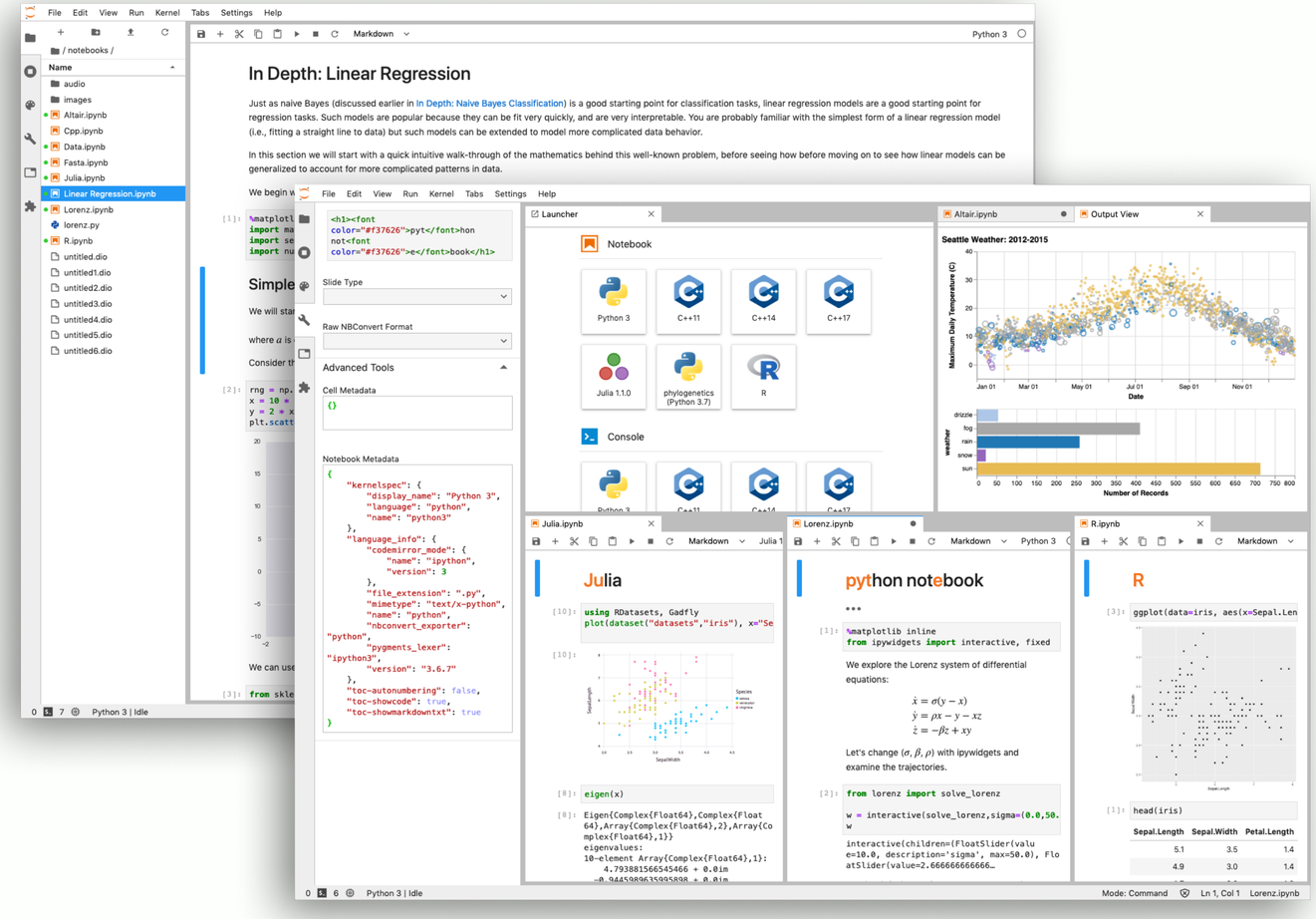
- Growth
- ACM Award
- Industry adoption
- Creative uses
- Open Source Book

## Jupyter receives the ACM Software System Award

It is our pleasure to announce that Project Jupyter has been awarded the 2017 ACM Software System Award, a significant honor for the project. We are humbled to join an illustrious list of projects that contains major highlights of computing history, including Unix, TeX, S (R's predecessor), the Web, Mosaic, Java, INGRES (modern databases) and more.

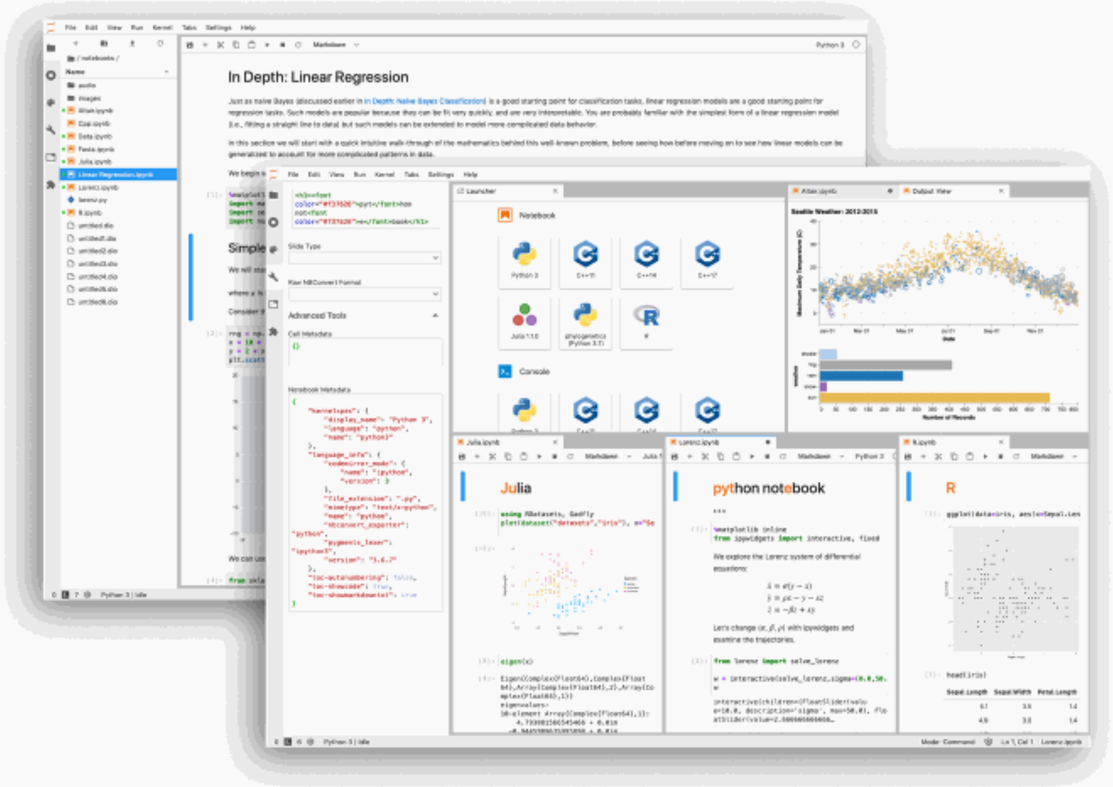
Officially, the recipients of the award are the fifteen members of the Jupyter steering council as of November 2016, the date of nomination (listed in chronological order of joining the project): Fernando Pérez, Brian Granger, Min Ragan-Kelley, Paul Ivanov, Thomas Kluyver, Jason Grout, Matthias Bussonnier, Damián Avila, Steven Silvester, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Carol Willing, Sylvain Corlay and Peter Parente.

# JupyterLab


In Depth: Linear Regression
Just as naive Bayes (discussed earlier in In Depth: Naive Bayes Classification) is a good starting point for classification tasks, linear regression models are a good starting point for regression tasks. Such models are popular because they can be fit very quickly, and are very interpretable. You are probably familiar with the simplest form of a linear regression model (i.e., fitting a straight line to data) but such models can be extended to model more complicated data behavior.
In this section we will start with a quick intuitive walk-through of the mathematics behind this well-known problem, before seeing how before moving on to see how linear models can be generalized to account for more complicated patterns in data.
Launcher
Notebook
Python 3
C++11
C++14
C++17
Julia 1.1.0
phylogenetics (Python 3.7)
R
Console
Altair.ipynb
Output View
Seattle Weather: 2012-2015
Julia.ipynb
Julia
Lorenz.ipynb
python notebook
R.ipynb
R
Slide Type
Raw NBConvert Format
Advanced Tools
Cell Metadata
Notebook Metadata

jupyter

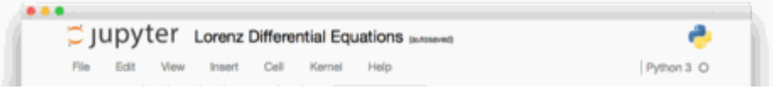
Install About Us Community Documentation NBViewer JupyterHub Widgets Blog




## JupyterLab 1.0: Jupyter's Next-Generation Notebook Interface

JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. JupyterLab is extensible and modular: write plugins that add new components and integrate with existing ones.

Try it in your browser  Install JupyterLab

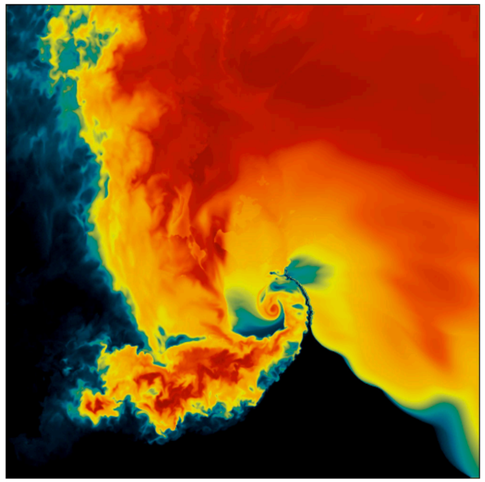


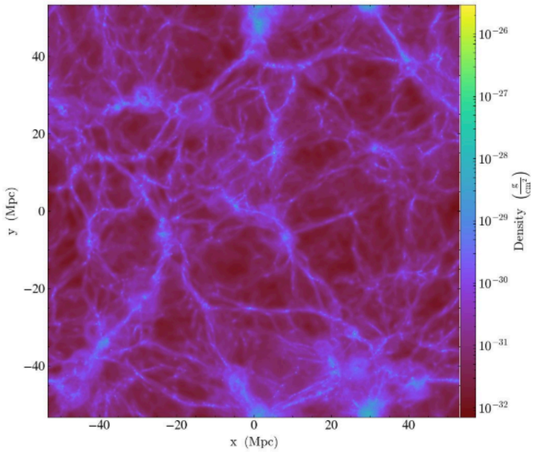

jupyter.org demo

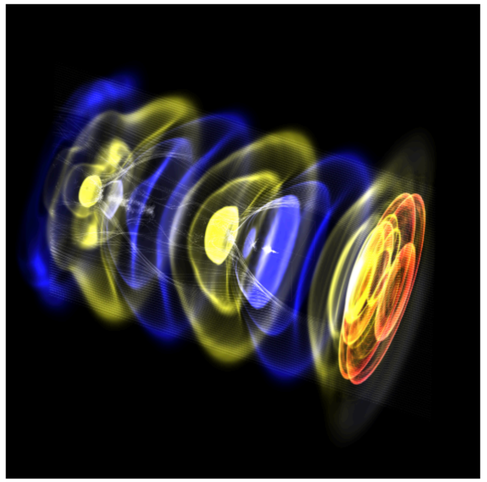
jupyter.org demo

# yt and jupyter widgets

## VISUALIZING OUR DATA





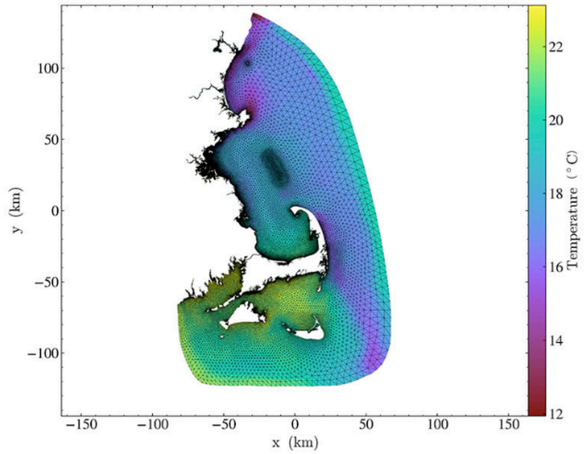



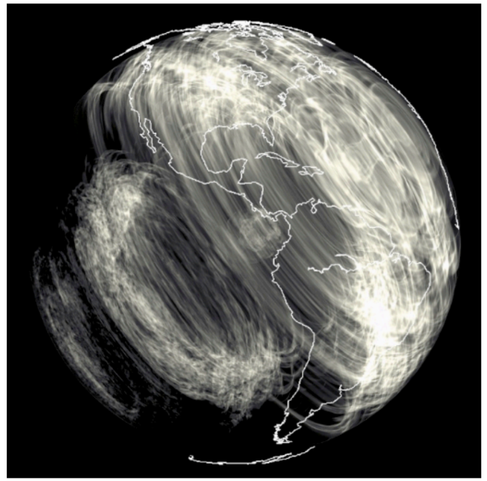
https://github.com/munkm/widgyts

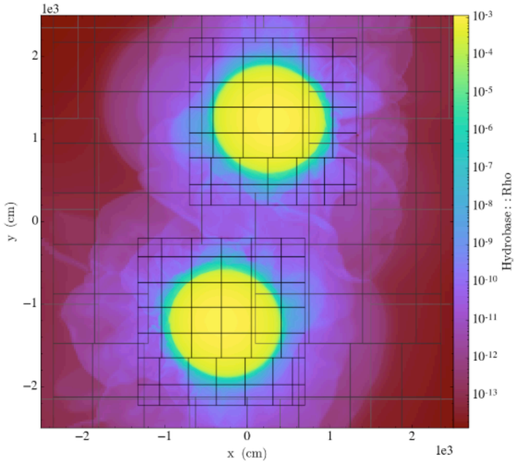
npm package @data-exp-lab/yt-tools

https://github.com/data-exp-lab/rust-yt-tools/

Irber Junior LC. Oxidizing Python: writing extensions in Rust [version 1; not peer reviewed]. F1000Research 2018, 7(ISCB Comm J):955 (poster) (https://doi.org/10.7490/f1000research.1115726.1)



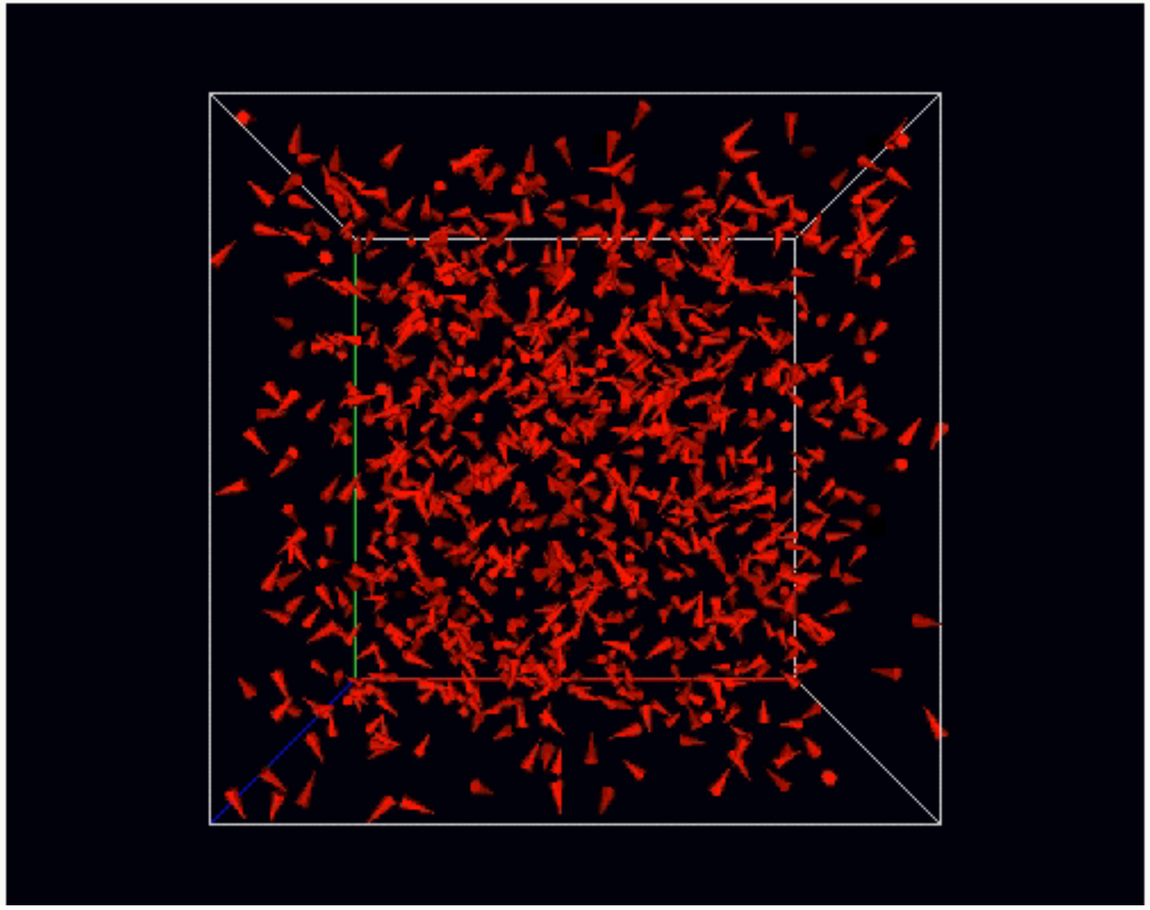


# ipyvolume

```
hdz = ipyvolume.datasets.hdz2000.fetch()
```

```
p3.clear()
```

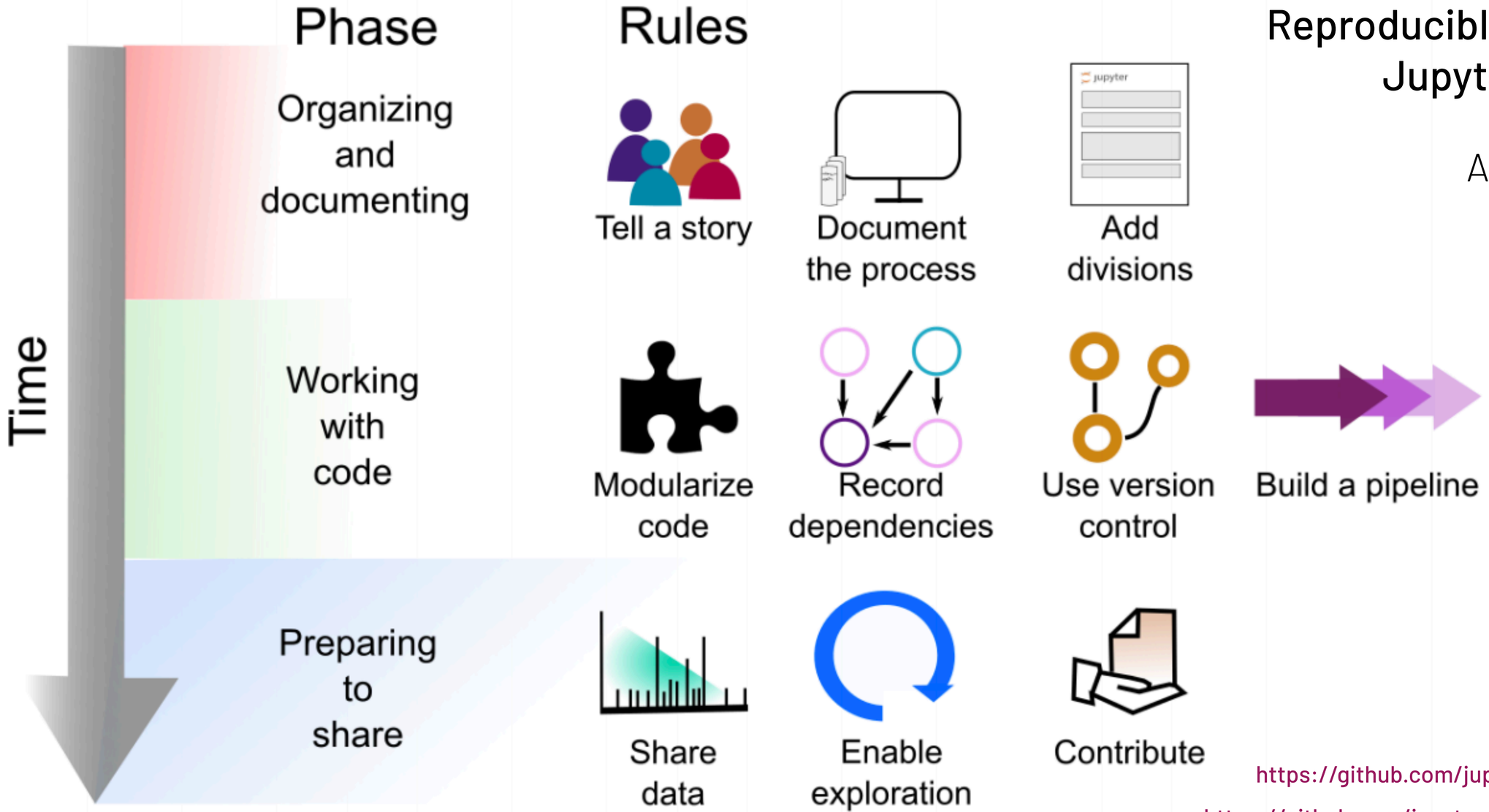
```
p3.figure(controls=False)
p3.volshow(hdz.data.T, level=[0.6, 0.8, 0.9],
           opacity=[0.01, 0.03, 0.05], controls=False)
x, y, z, u, v, w = np.random.random((6, 1000))*2-1
quiver = p3.quiver(x, y, z, u, v, w, s=0.05, ss=0.1)
p3.show()
```



https://towardsdatascience.com/multivolume-rendering-in-jupyter-with-ipyvolume-cross-language-3d-visualization-64389047634a

# Healthy ~~Best~~ Practices

# Simplified Notebook Cycle




## Ten Simple Rules for Reproducible Research in Jupyter Notebooks

Adam Rule et al.

https://github.com/jupyter-guide/jupyter-guide

https://github.com/jupyter-guide/ten-rules-jupyter

# Keep up with changes

# Proceed cautiously with **pseudo-open** projects

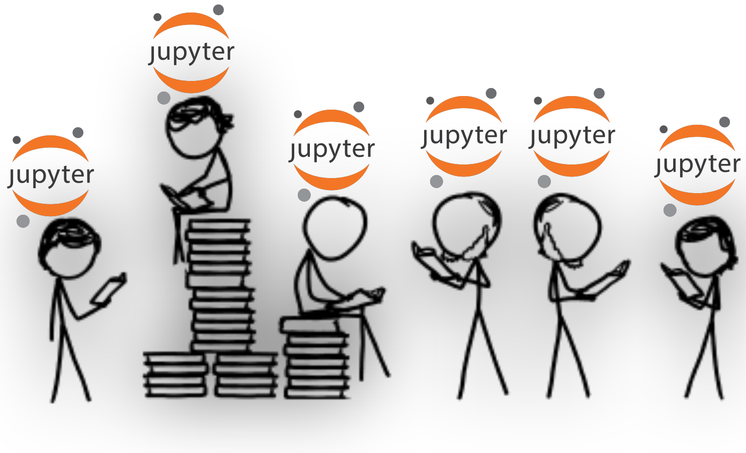
# Ask why

Tools

**Processes**

Communication

# jupyterhub






jupyterhub

**Zero to JupyterHub with Kubernetes**

A tutorial to help install and manage JupyterHub with Kubernetes

**Quick search**

## Zero to JupyterHub

JupyterHub is a tool that allows you to quickly utilize cloud computing infrastructure to manage a hub that enables users to interact remotely with a computing environment that you specify. JupyterHub offers a useful way to standardize the computing environment of a group of people (e.g., for a class of students or an analytics team), as well as allowing people to access the hub remotely.

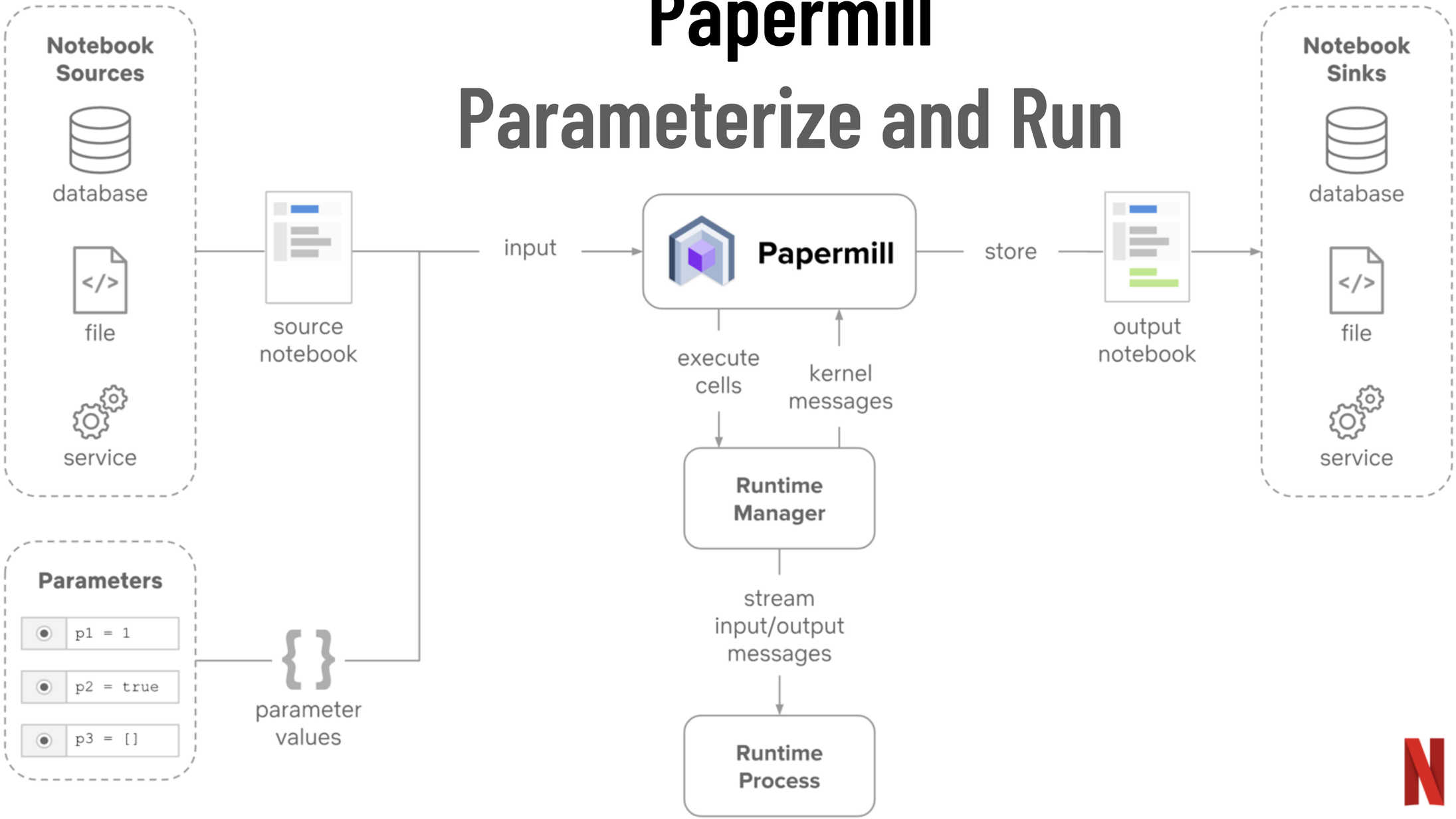
This growing collection of information will help you set up your own JupyterHub instance. It is in an early stage, so the information and tools may change quickly. If you see anything that is incorrect or have any questions, feel free to reach out at the issues page.

## Creating your JupyterHub

This tutorial starts from "step zero" and walks through how to install and configure a complete JupyterHub deployment in the cloud. Using Kubernetes and the JupyterHub Helm chart provides sensible defaults for an initial deployment.

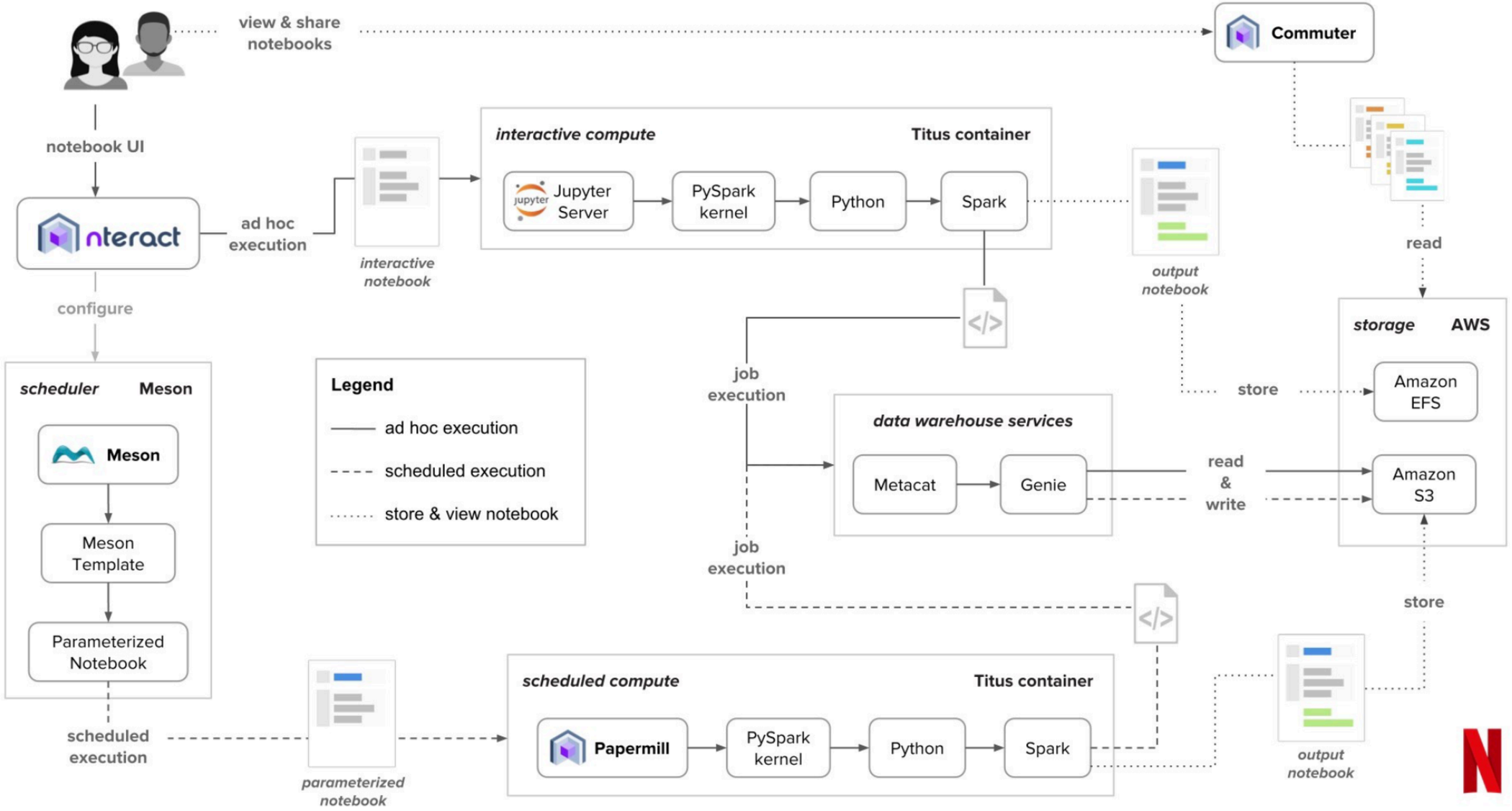
**zero-to-jupyterhub.readthedocs.io**

# Papermill
## Parameterize and Run



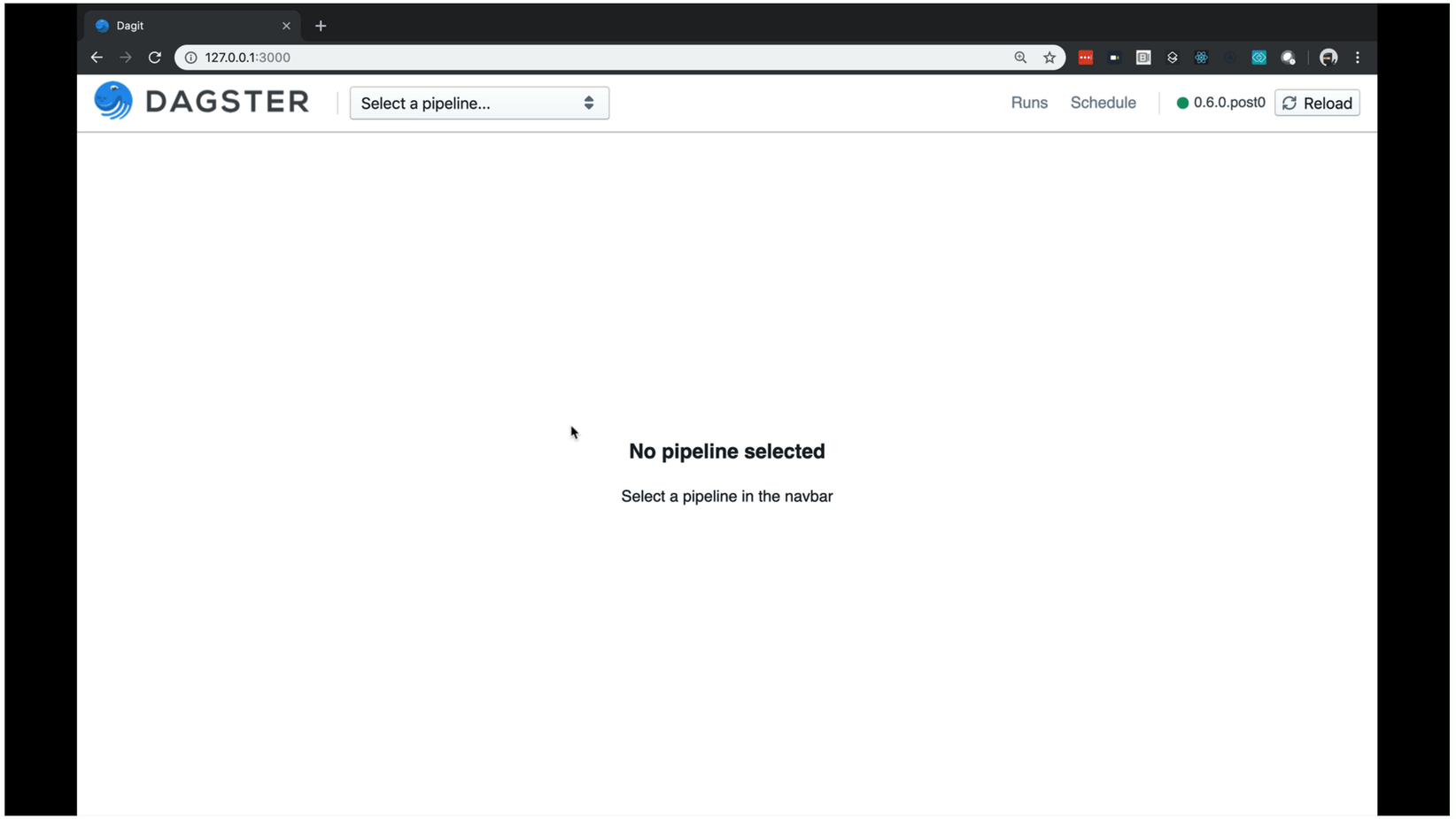
# Data at scale - Netflix

nteract
Papermill
Scrapbook
Bookstore
Commuter




https://medium.com/netflix-techblog/notebook-innovation-591ee3221233

# Pipelines



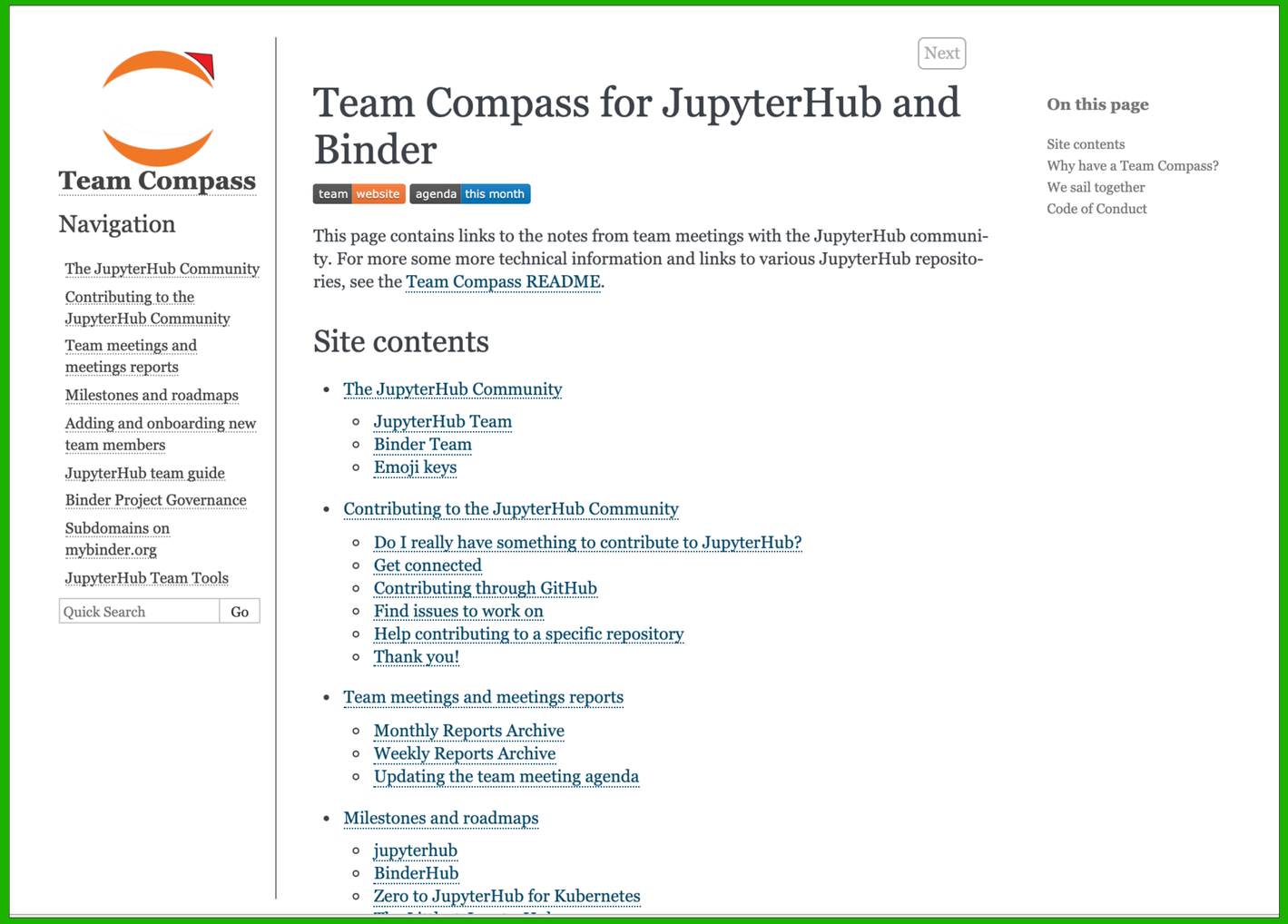

DAGSTER

https://medium.com/dagster-io/dagster-0-6-0-impossible-princess-898b459375e0

# Create a Reproducibility Pipeline

# Decouple steps for flexibility

# Plan
# Execute
# Change




https://jupyterhub-team-compass.readthedocs.io

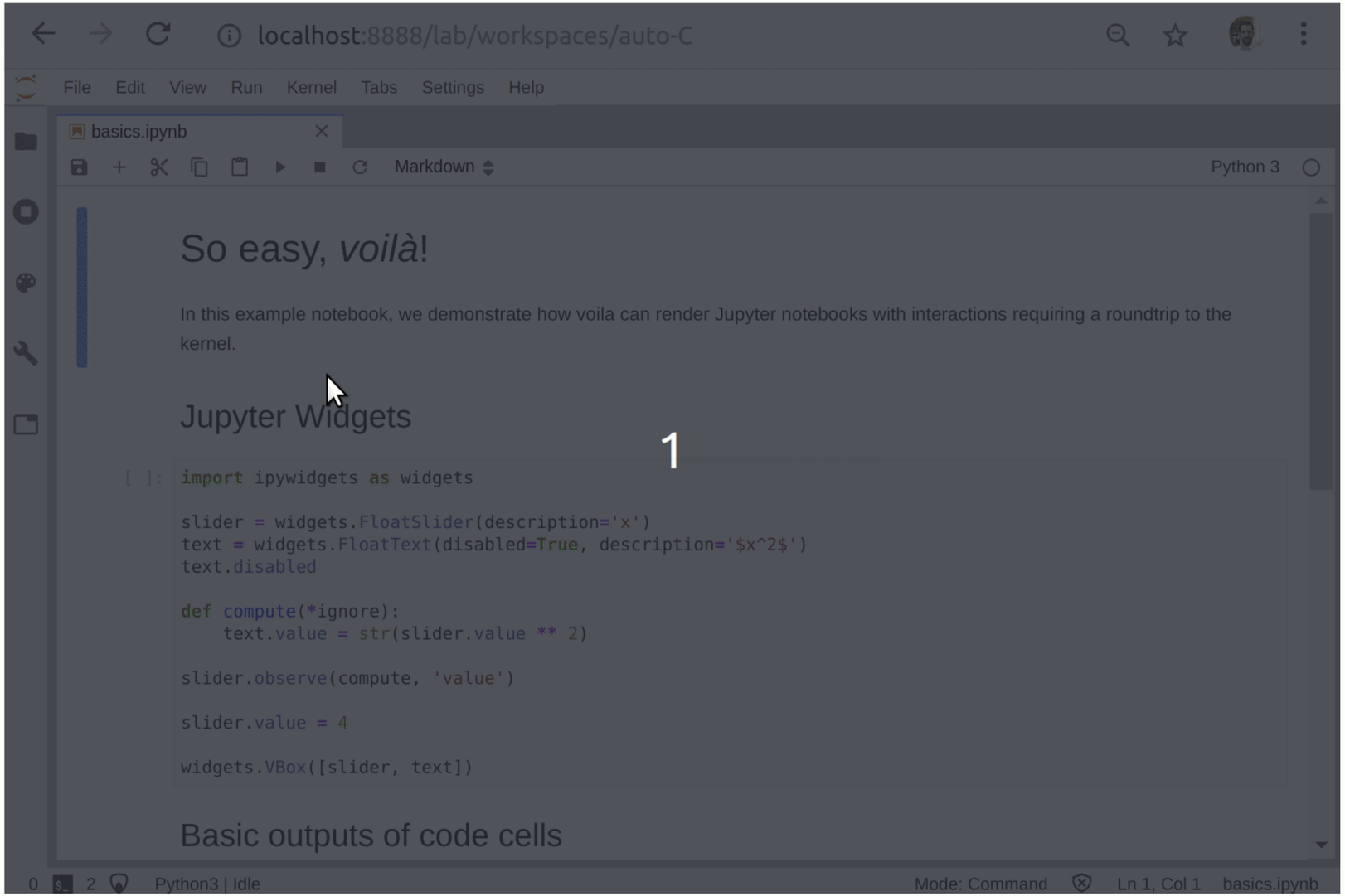
https://github.com/jupyterhub/team-compass

Tools

Processes

**Communication**

# Notebooks to web




voilà

https://blog.jupyter.org/and-voil%C3%A0-f6a2c08a4a93

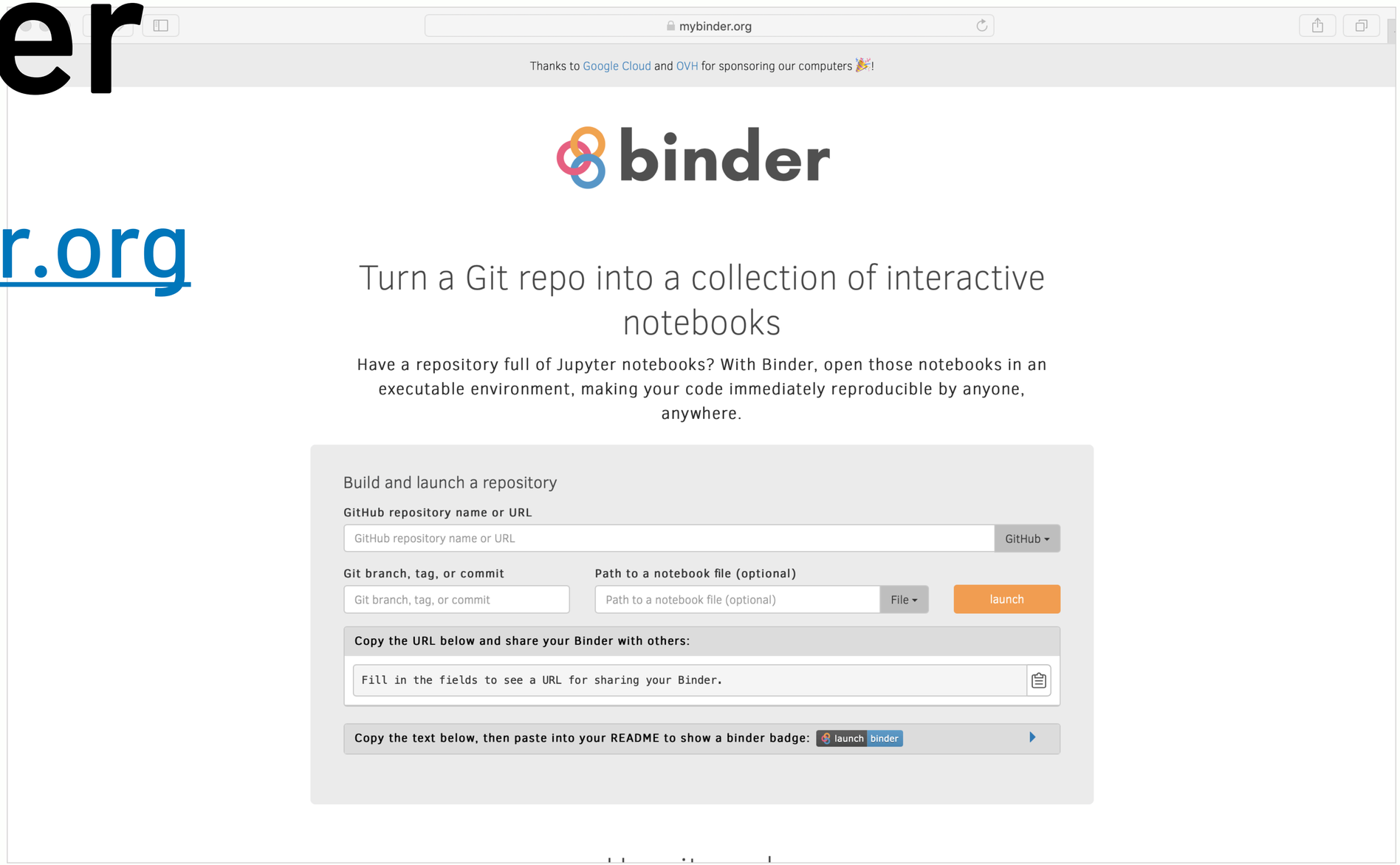
# Binder

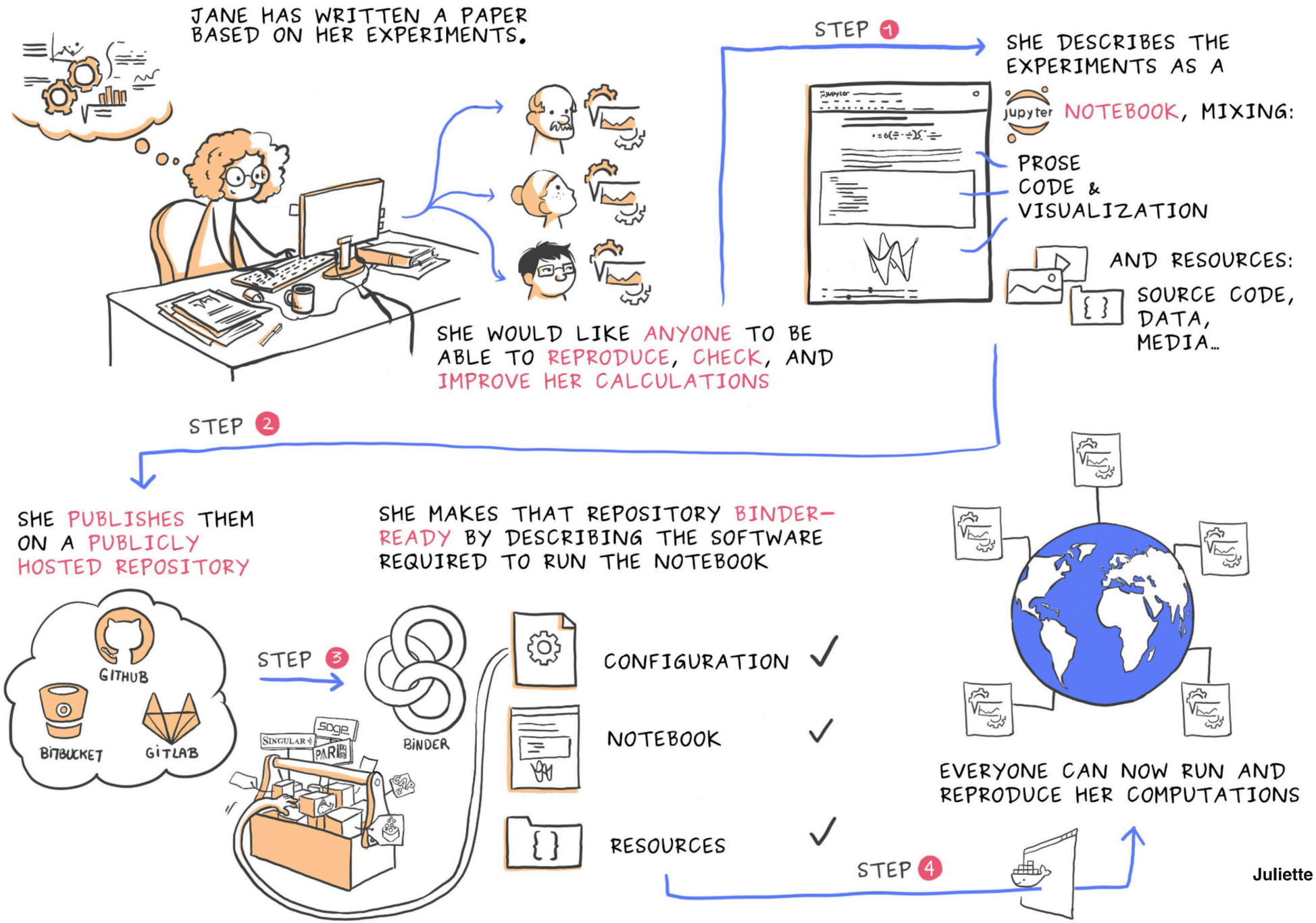
[mybinder.org](mybinder.org)

Binder 2.0 blog post

elifesciences: Share your interactive research environment

Nature article about Binder

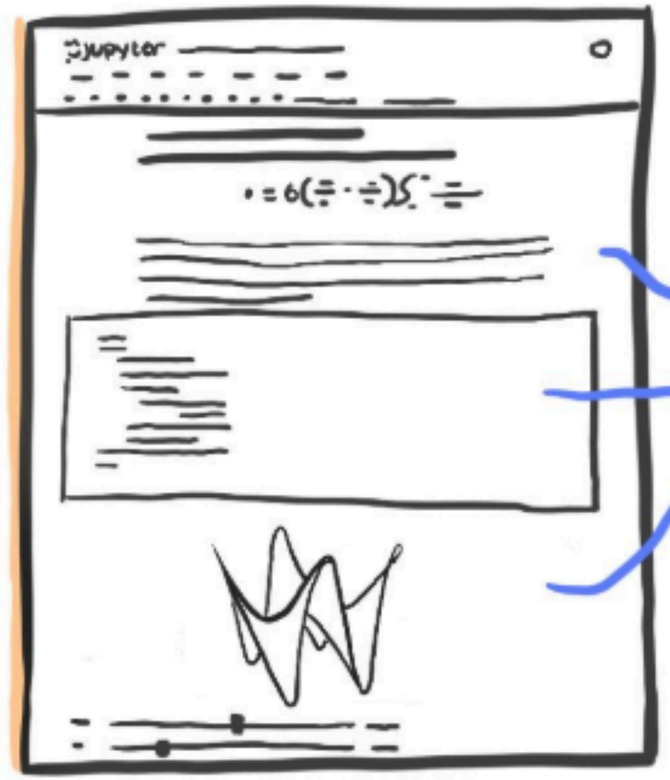



JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE, DATA, MEDIA…
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
BINDER
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS
Juliette Taka

Juliette Taka

STEP 1

SHE DESCRIBES THE EXPERIMENTS AS A




jupyter NOTEBOOK, MIXING:

PROSE

CODE &

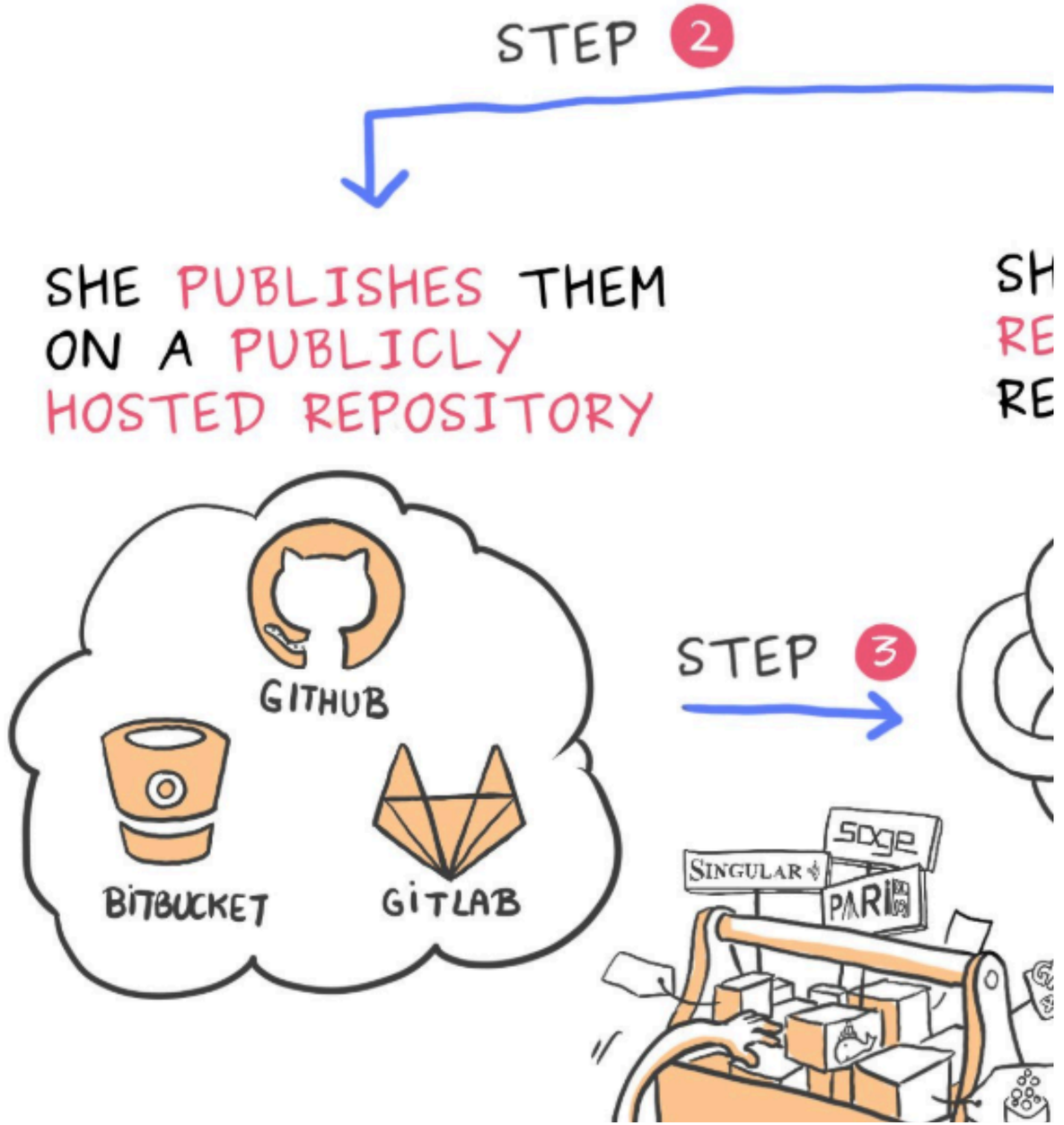
VISUALIZATION

AND RESOURCES:

SOURCE CODE,
DATA,
MEDIA...

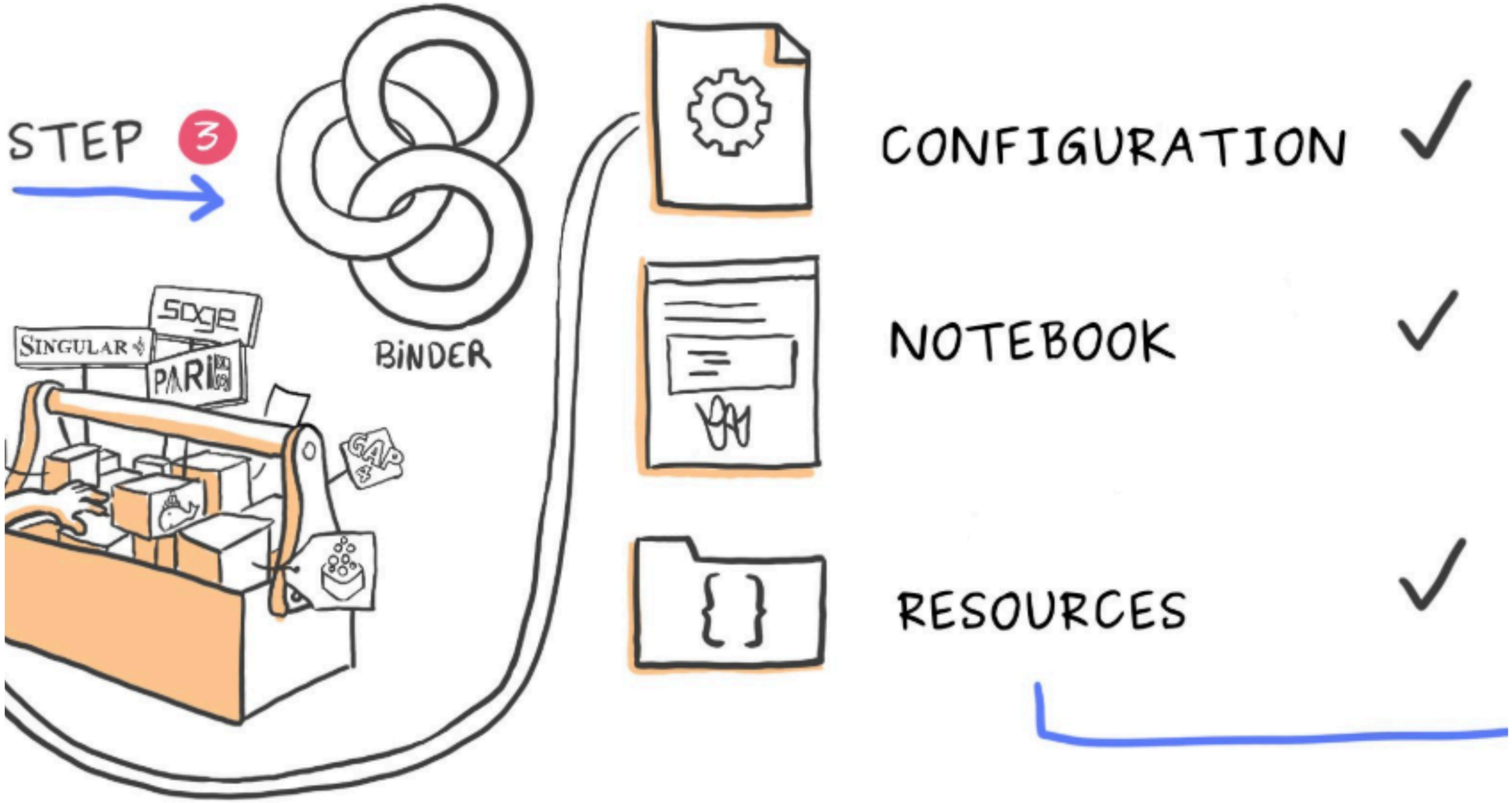
ANYONE TO BE
CE, CHECK, AND

Juliette Taka

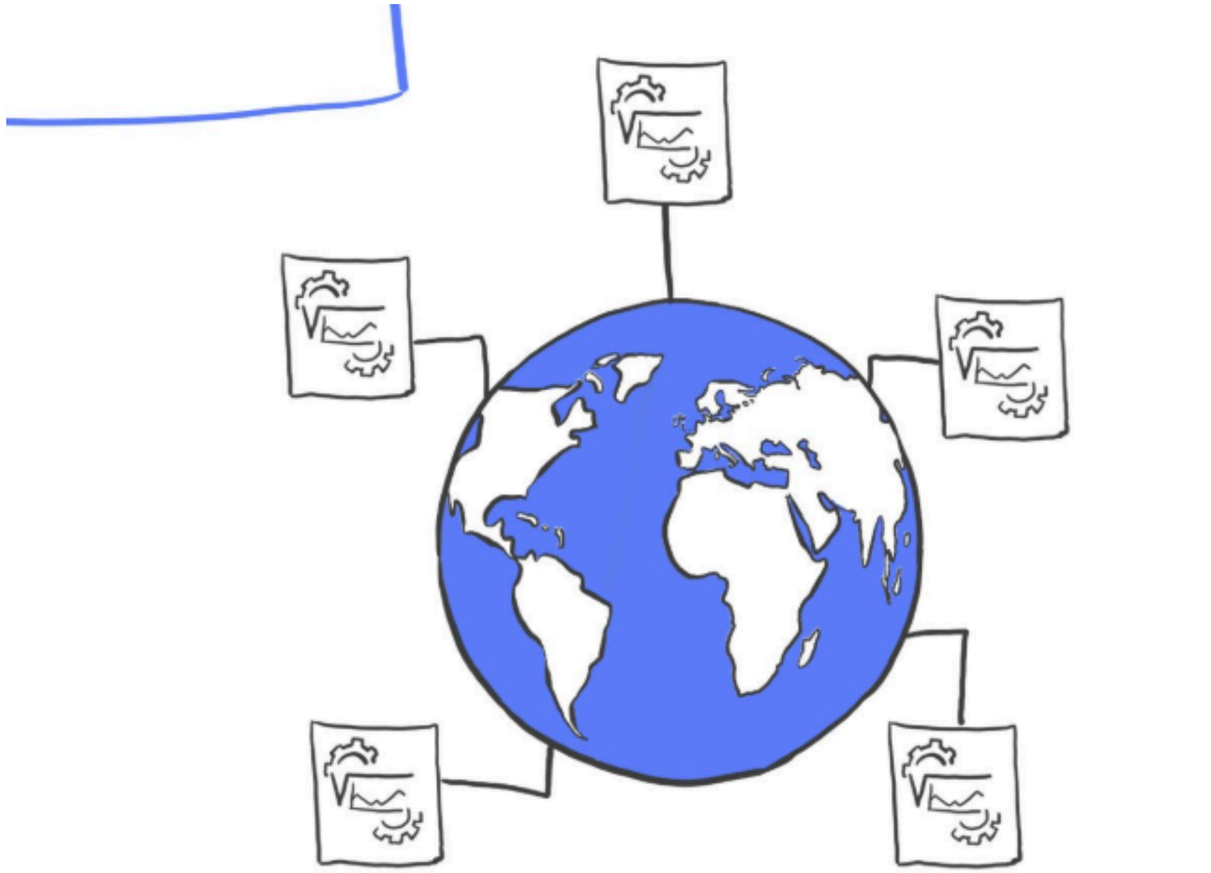



Juliette Taka

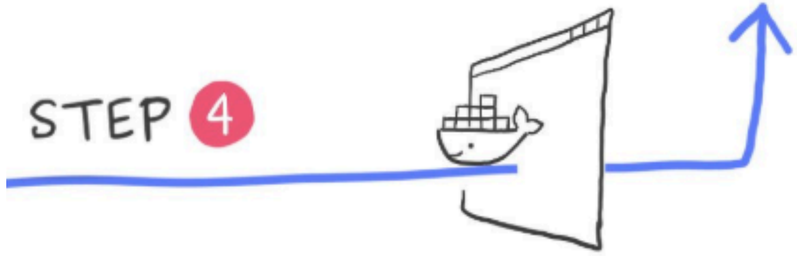
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK




Juliette Taka



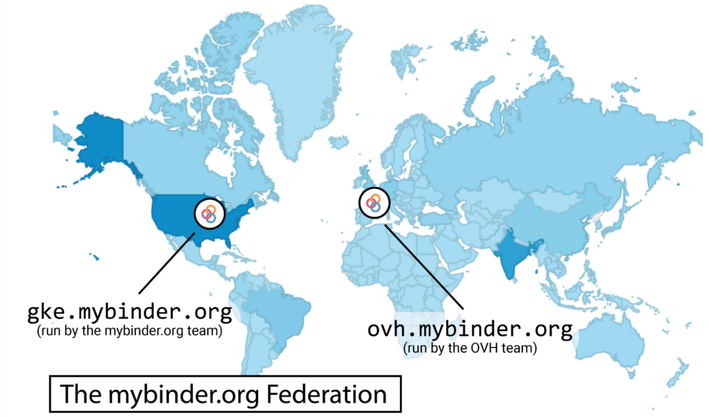
EVERYONE CAN NOW RUN AND
REPRODUCE HER COMPUTATIONS

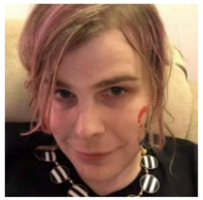
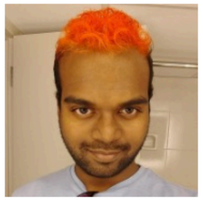



Juliette Taka

# Binder

Zenodo DOIs now work with BinderHub




| | | | |
|---|---|---|---|
| **Jessica Forde**<br>UC Berkeley<br>team red<br>📖 | <br>**Sarah Gibson**<br>The Alan Turing Institute<br>team blue<br>💬 ,📖 ,✅ | <br>**Tim Head**<br>Wild Tree Tech<br>team red | <br>**Lindsey Heagy**<br>UC Berkeley<br>team blue<br>🤔 ,💡 |
| <br>**Chris Holdgraf**<br>Berkeley Institute for Data Science<br>team red<br>💻 ,🤔 ,📖 ,💬 |  <br>**M Pacer**<br>Netflix<br>team blue |  <br>**Yuvi Panda**<br>UC Berkeley<br>team blue<br>💻 ,🚇 | <br>**Min Ragan-Kelley**<br>Simula<br>team lead<br>data,💻 |
| <br>**Zach Sailer**<br>Project Jupyter<br>team blue<br>💻 ,🤔 ,💬 | <br>**Erik Sundell**<br>Sandvik CODE<br>team blue<br>💻 ,🚇 | <br>**Carol Willing**<br>Project Jupyter<br>team red<br>Python,Community | |

# Binder

## mybinder.org

README.md

### BrainViz

build unknown | launch binder

This project was created as part of Neurohackademy 2019.

**Contributors**: Anna Jafarpour, Matan Mazor, Liberty Hamilton. Thanks also to Noah Benson, developer of neuropythy.
**Description of project**: With BrainViz we can visualize neural activity in selected ECOG contacts.
**How to get involved**: Please get in touch with one of the contributors.

This is a tool to create an interactive 3D brain in your browser. You can select points (or electrodes) on the brain to look at activity from those points. A dropdown menu will allow you to select the stimulus for which you are plotting a response.

#### input

Example Brain image is from https://zenodo.org/record/996814#.XU29UZNKi_s for 3D visualization. ECOG data is synthetic for demonstration, that is from a hf5 format file containing: data['/ecog/condition1'] = numpy array of dim [elecs x time] data['/ecog/condition2'] = numpy array of dim [elecs x time] data['/times'] = vector of times

#### interactive visualization

The electrodes of interest can be selected by mouse or dropdown menu. There is also a dropdown menu to select conditions.

#### output

3D brain with electrodes overlay. The color of selected electrodes changes.

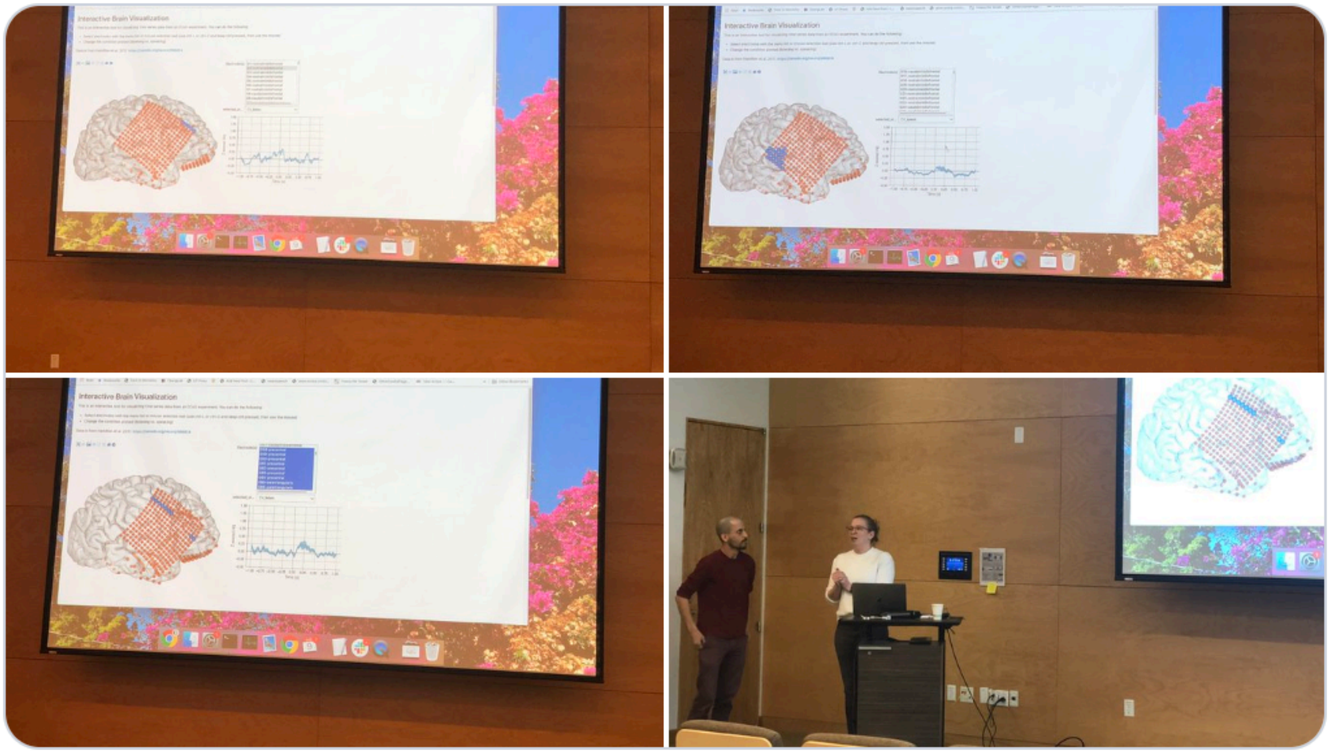
Binder Team Retweeted

**Kirstie Whitaker** @kirstie_j · Aug 10

The first #NeuroHack19 project team built an awesome brain visalisation too! It's so BEAUTIFUL 😍😍😍

Check it out for yourself via @mybinderteam at mybinder.org/v2/gh/matanmaz....

Code at: github.com/matanmazor/Bra...



@WillingCarol

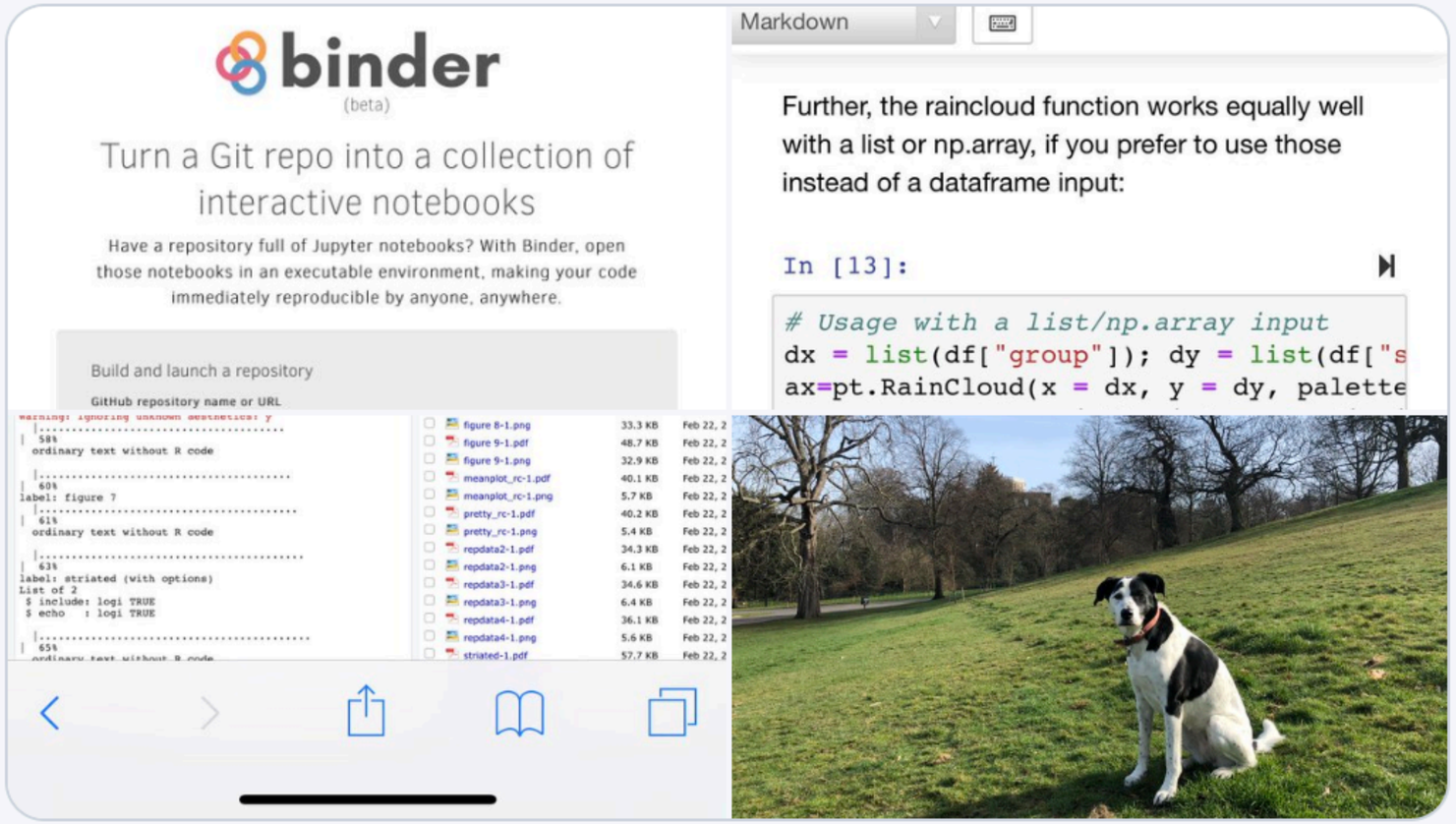
# From a phone in the park!

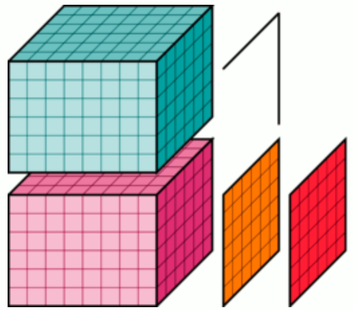
Binder Team Retweeted

**Kirstie Whitaker** @kirstie_j · Feb 22

WHAT A GREAT FRIDAY! #RainCloudPlots in a @turinginst hosted @mybinderteam instance!! FROM MY PHONE in the park! 😱👾🚀✨💖🥳😊🤩 Way to go Sarah!

#TuringWay #BinderHub #ComeToOurWorkshops: eventbrite.co.uk/o/the-turing-w... 😉😄🙌





xarray

jupyter

# Pangeo

Iris

https://pangeo.io

DASK

# Jupyter meets the Earth: we got an NSF grant!

## Research use-cases

- CMIP6 Climate data analysis
- Large scale hydrological modelling
- Geophysical simulations and inversions

## Tech developments

- Data discovery through JupyterLab
- Interactivity: Widgets & Dashboards
- JupyterHub: Using and managing shared computational infrastructure

Fernando Perez

Joe Hamman

Laurel Larsen

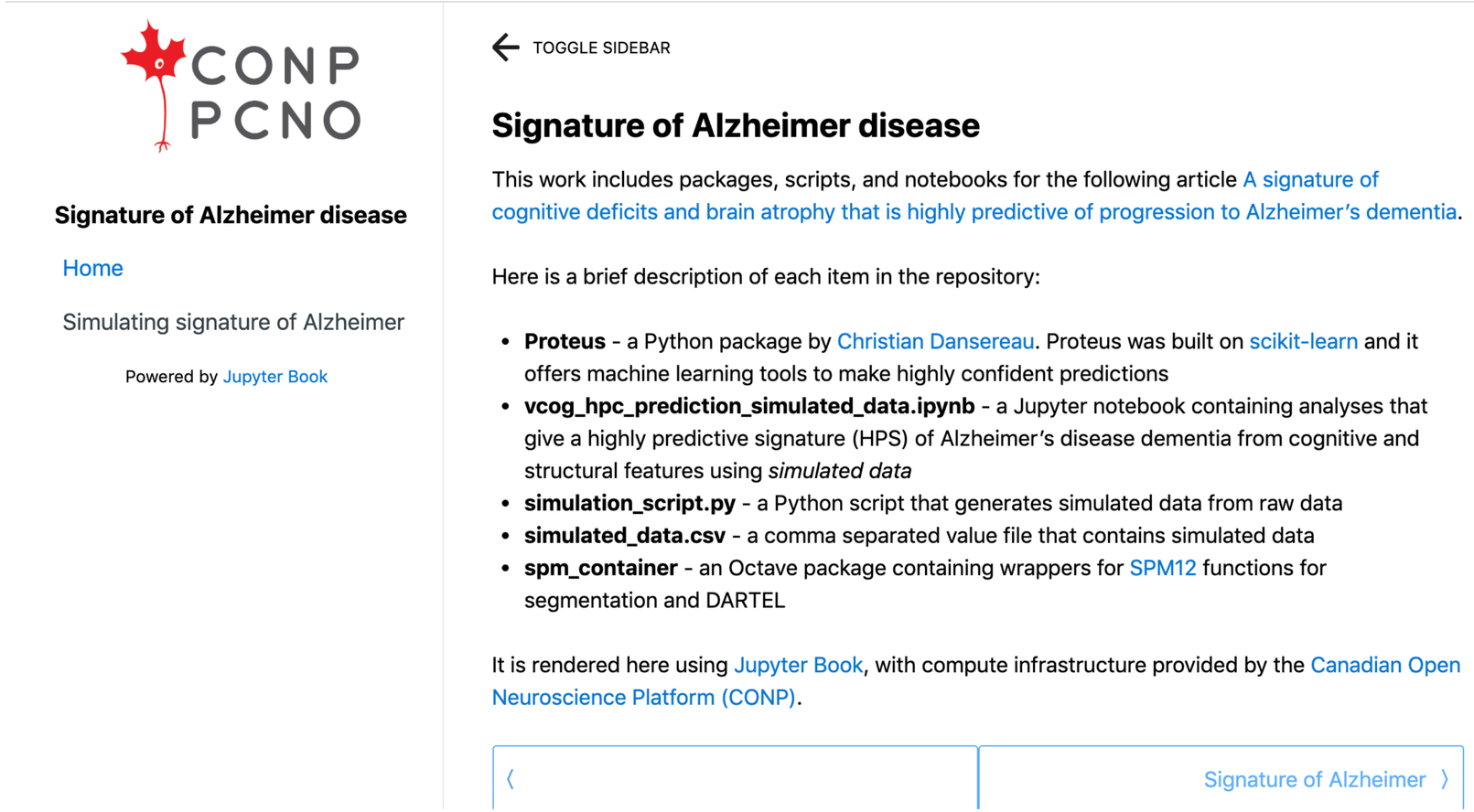
Kevin Paul

Lindsey Heagy

Chris Holdgraf

Yuvi Panda

# Canadian Open Neuroscience Platform




https://simexp.github.io/vcog_hps_ad_book/intro.html

Jupyter Book
Binder
Jupyter
pandas
scipy
scikit learn
matplotlib
numpy
seaborn

# Build Communities

jupyter.org

# Leverage solutions across disciplines

# Share binders.
# Foster scientific research.

# Tools
# Processes
# Communication

# Why strive for reproducible research?

# Reproducible research improves prediction

# prediction =

# impact

# Scaling reproducible research improves science and our world

後長岡
まつり

# Thank you

ECMWF Workshop Organizers
Claudia Vitolo
Project Jupyter Team
Min Ragan-Kelly

GORDON AND BETTY MOORE FOUNDATION

Alfred P. Sloan FOUNDATION

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

# Attributions

**References to published research, projects, and drawings (and marked on slides)**


[2] Statistics: https://fivethirtyeight.com/features/which-city-has-the-most-unpredictable-weather/
[7, 11] A Bridge for Passing, Pearl S. Buck
[8, 9, 18] ECMWF
[12] Copyright: 2019 European Union, contains modified Copernicus Sentinel data 2019, processed by EUMETSAT
[13] Copyright contains modified Copernicus Sentinel data (2019), processed by ESA, CC BY-SA 3.0 IGO
[30] Madicken Munk
[31] Maarten Breddels
[33] Adam Rule et al.
[46] Quantstack - Voila
[48-53] Juliette Taka
[57] Pangeo
[58] Lindsey Heagy
[59] Canadian Open Neuroscience Platform


**Photos**


[2-6, 10, 16-17, 69, 70] Source: Carol Willing and Linnea Willing
[14] Twitter
[15] Getty Images
[55, 56] Kirstie Whitaker
[23-29, 38, 44, 47, 54, 61] Project Jupyter
[39-40] nteract and Netflix
[41] Nick Shrock, Dagster

**15 October - 9:10**

**Carol Willing (Willing Consulting / Project Jupyter; Twitter: @WillingCarol)**

**Scaling reproducible research with Project Jupyter**

Jupyter notebooks have become the de-facto standard as a scientific and data science tool for producing computational narratives. Over five million Jupyter notebooks exist on GitHub today. Beyond the classic Jupyter notebook, Project Jupyter's tools have evolved to provide end to end workflows for research that enable scientists to prototype, collaborate, and scale with ease. JupyterLab, a web-based, extensible, next generation interactive development environment enables researchers to combine Jupyter notebooks, code and data to form computational narratives. JupyterHub brings the power of notebooks to groups of users. It gives users access to computational environments and resources without burdening the users with installation and maintenance tasks. Binder builds upon JupyterHub and provides free, sharable, interactive computing environments to people all around the world.