LBNL CA Water Data Challenge

July 15th, 2018

Jupyter Team.

Presented by:
Matthias Bussonnier

bussonniermatthias@gmail.com
GitHub: @carreau
Twitter: @mbussonn

About Me

Matthias Bussonnier

- A Physicist/Bio-Physicist
- Core contributor of IPython/Jupyter since 2012
  - Co-founder, and Steering Council member
- Post doctoral Scholar on Jupyter at BIDS

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

IPython – 2001

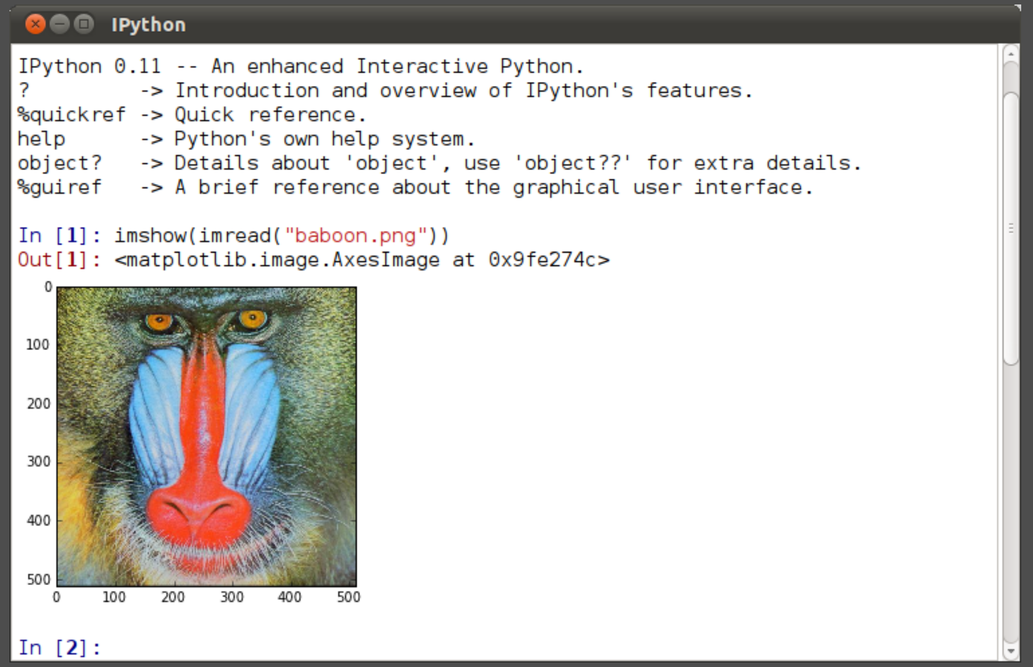
IPython

```
$ ipython
Python 3.6.0
Type 'copyright', 'credits' or 'license' for more information
IPython 6.0.0.dev -- An enhanced Interactive Python. Type '?' for help.

In [1]: from string import hexdigits
   ...: from random import choice
   ...:
   ...: def randhex(length=10):
   ...:     return '0x'+''.join([choice(hexdigits) for x in range(10)]).l
                                                                    ljust
                                                                    lower
                                                                    lstrip
```

(BTW, IPython is uppercase I)

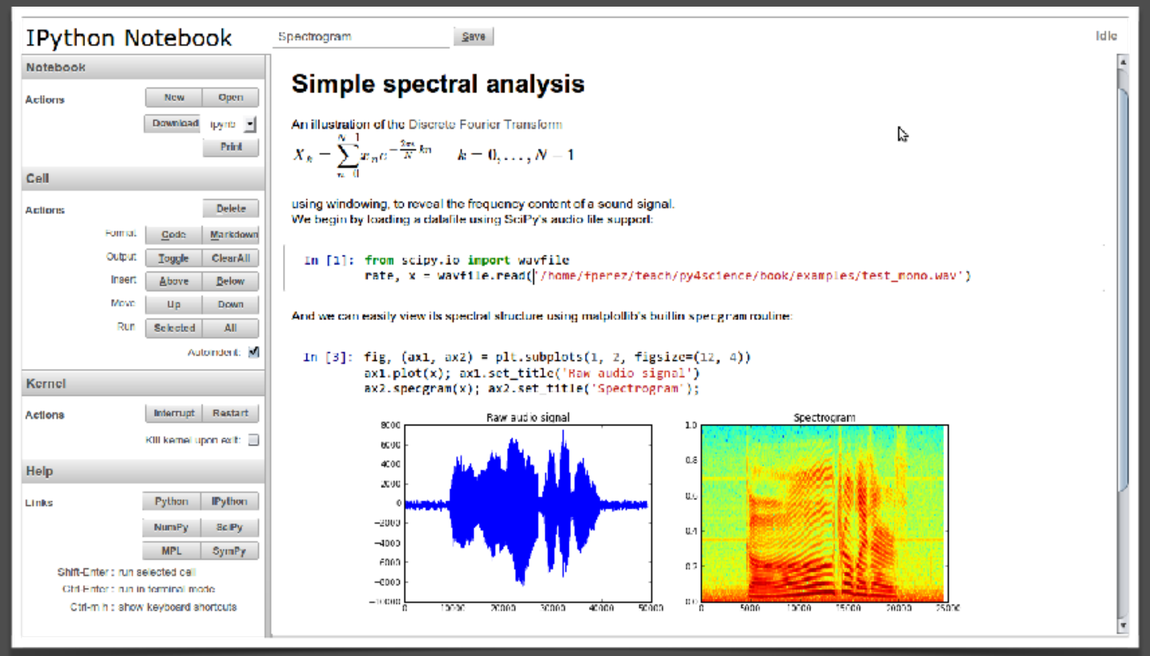
QtConsole 2010-2011



The IPython Notebook – 2012



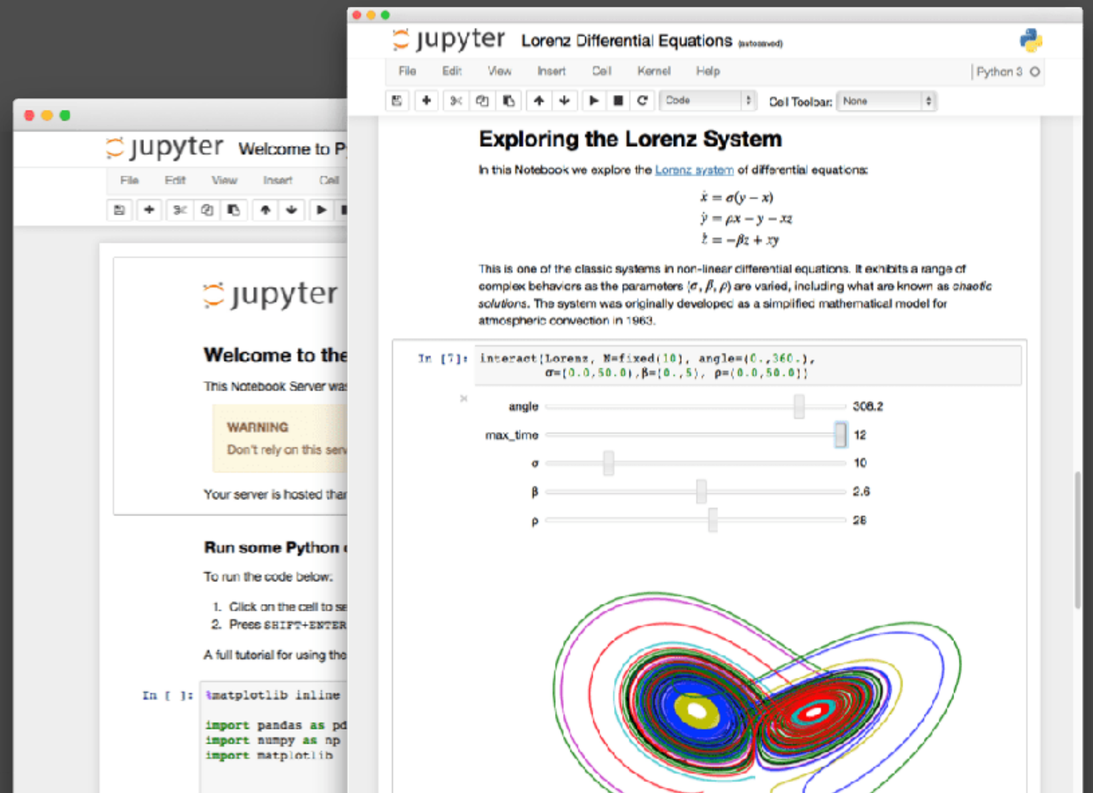




julia (2013)

Jupyter – 2014

Renames the Python Agnostic Part to “Jupyter” – an homage to Galileo first Notebooks.

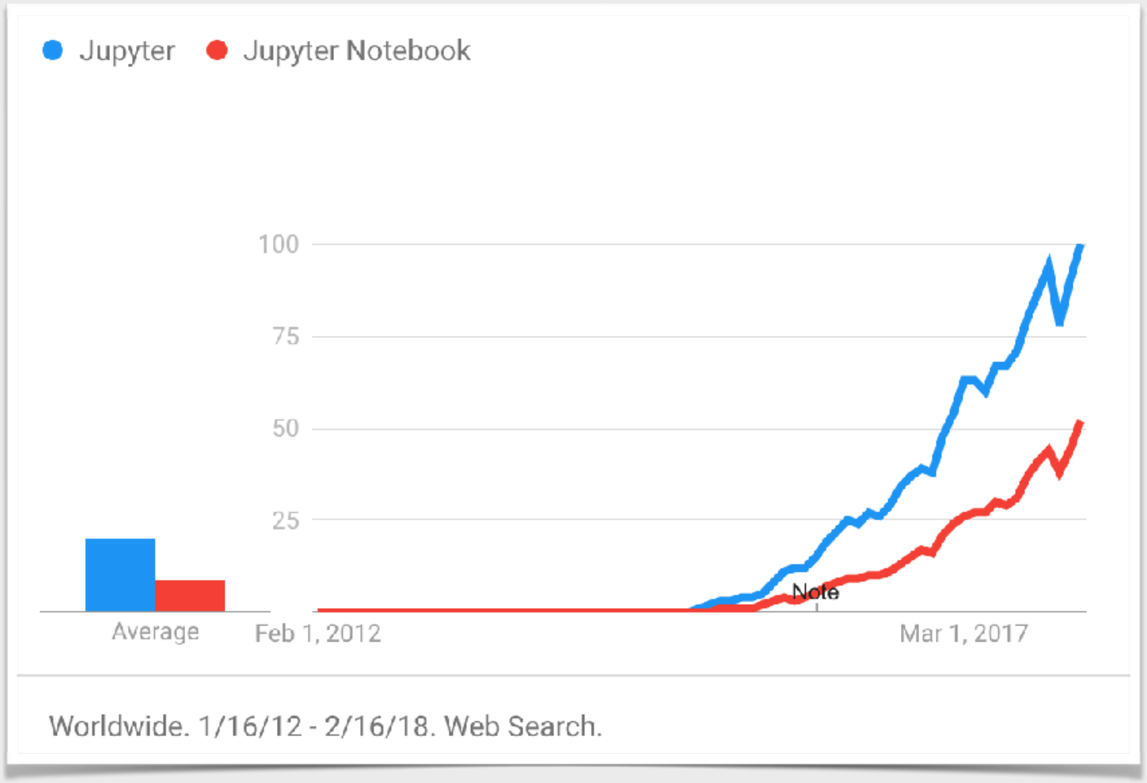


Jupyter: 2018

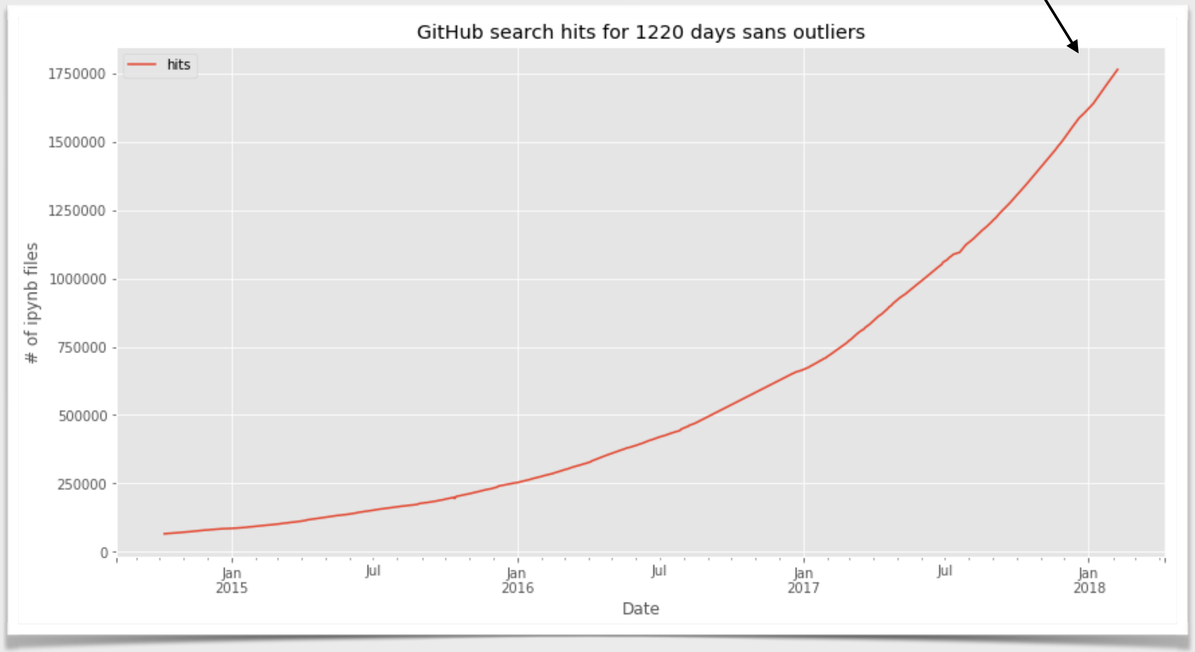
1000+ Contributors
Dozens of Projects

2017 ACM System Software Award

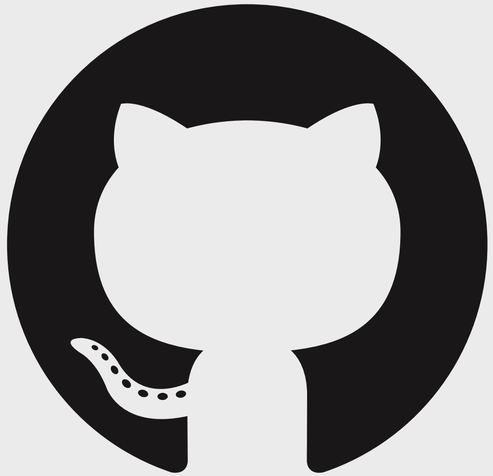
A few Numbers



~1.7M notebooks
on GitHub in Jan 2018



https://github.com/parente/nbestimate



150+ repositories across multiple organizations

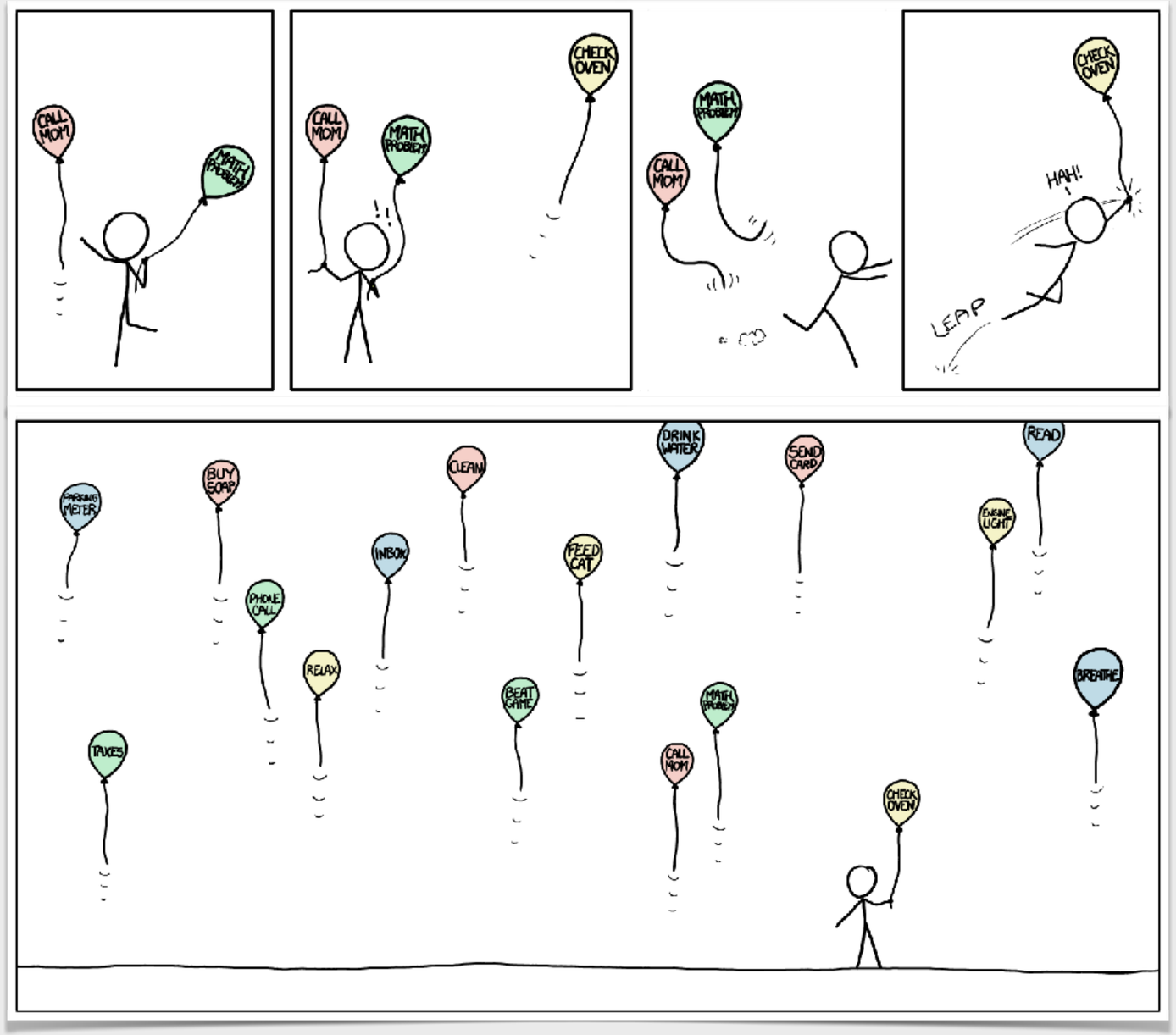
(IPython, Jupyter, JupyterHub, JupyterLab, ...)
at 2 release/year that's ~ 1 release per day

1000+ Contributors

8+ Millions Users,

(with conservative estimates)

Worldwide ~21M developers – North America ~4.4M
VS Code ~2.6 M Active Users
GitHub 24M Users



Randall Munroe
(https://xkcd.com/1106/)

Core Contributors

- 1000+ Open source contributors, Majority Volunteer
- Organisation with Open Governance

Sponsors


ALFRED P. SLOAN FOUNDATION


NUMFOCUS
OPEN CODE = BETTER SCIENCE


THE LEONA M. AND HARRY B.
HELMSLEY
CHARITABLE TRUST


GORDON AND BETTY
MOORE
FOUNDATION


O'REILLY


SIMONS FOUNDATION


U.S. DEPARTMENT OF
ENERGY


Google


CONTINUUM
ANALYTICS


IBM


POWERED BY
rackspace
the open cloud company


ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS


Bloomberg


Microsoft

How Jupyter came to be

Life cycle of a Scientific Idea

- Individual computational exploratory work
- Collaborative development
- Parallel production runs (HPC, cloud, ...)
- Publication & communication (reproducibly!)
- Education
- Goto 1

“The purpose of computing is insight, not numbers”

–Hamming’62

Life cycle of a Scientific Idea

- Individual exploratory work (Repl, Scripts)
- Collaborative development (Dropbox/ Google Doc / emails / git)
- Parallel production runs (MPI, rewrite C++, batch jobs)
- Publication & communication (Word, Latex, ppt...)
- Education
- Goto 1

Tools Overhead

Each Tool brings (cognitive) overhead, time to install, deploy, and master.

Can we create a (set of) tools, with minimal overhead end enough flexibility ?

Parallel with popular DataScience languages

Fortran/C/C++ are fast, but take significant development time and skills

Python/R/Julia are (usually) slower, but are useful immediately.

Rise of Jupyter

- An increasing number of discipline have a fast growing amount of data
- Technology is a tool that should
  - Empower the User
  - Amplify Domain Knowledge and Expertise
  - Facilitate Sharing and Collaboration

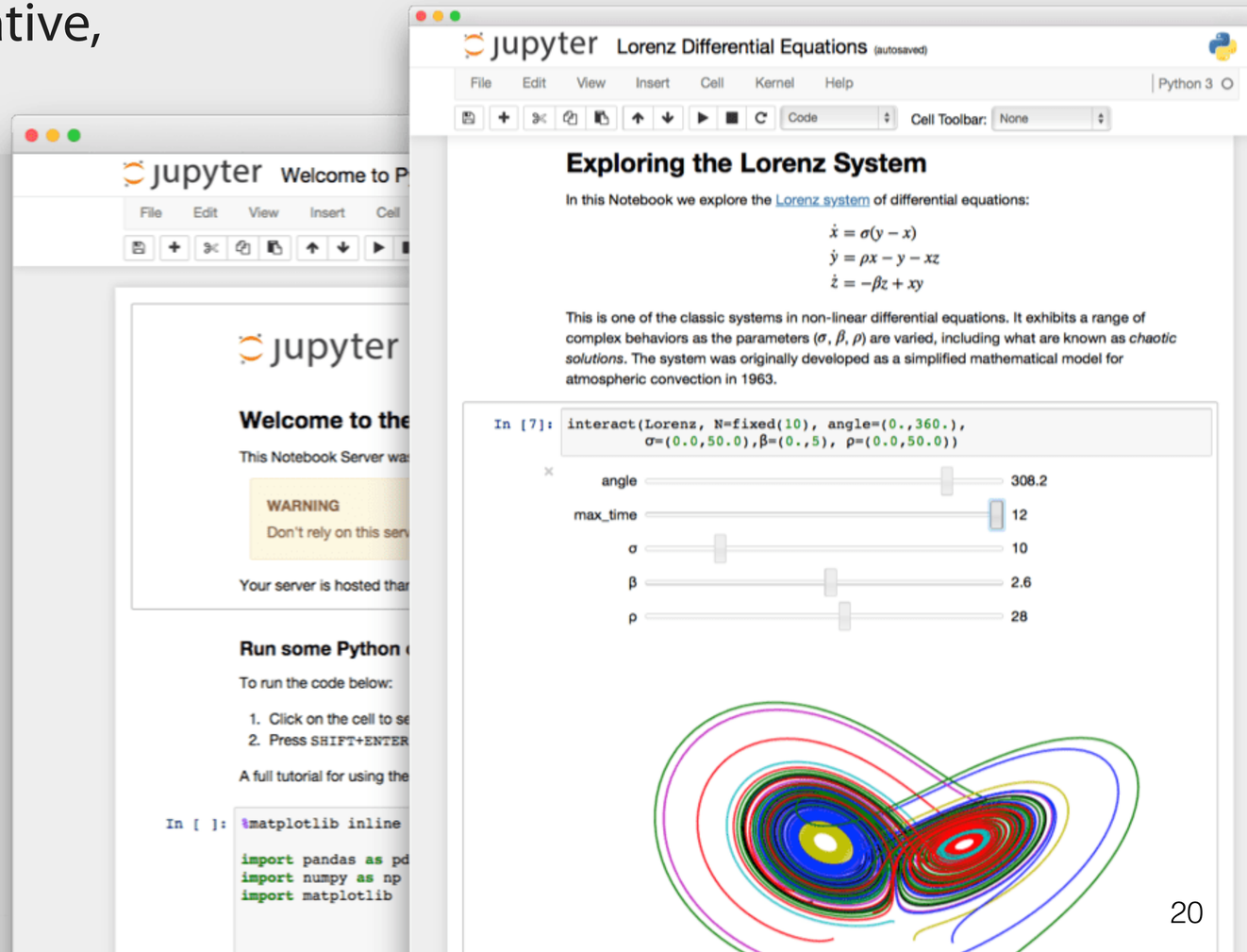
Jupyter provide a framework that can be use in all the step in the cycle of a scientific idea

- BSD Licensed (Free to use and redistribute even Commercially)
- Open Source, Community Maintained
- Important for sustainability, diversity, and equal access

What is Jupyter

Individual,Collaboration,Parallel,Publication,Education

- Mainly Known for **The Notebook**
  - Web server, a web app, containing code, narrative, math and results.
  - Attached to a **Kernel** doing computation.
- Results can be:
  - Static (Image)
  - Interactive (client-side scoll/pan/brush)
  - Dynamic (Call back into Kernel)



Narrative

Code

Result

jupyter Lorenz Differential Equations Last Checkpoint: 21 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Python 3

Code CellToolbar

Exploring the Lorenz System of Differential Equations

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Math

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters (σ, β, ρ) are varied.

```
In [11]: w = interactive(solve_lorenz, angle=(0.,360.), N=(0,50),σ=(0.0,50.0), ρ=(0.0,50.0))
         display(w)
```

N 10

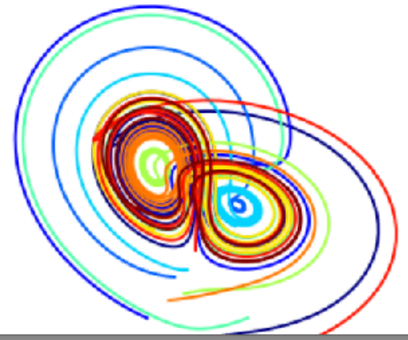
angle 0.00

max_time 4.00

σ 10.00

β 2.67

ρ 28.00

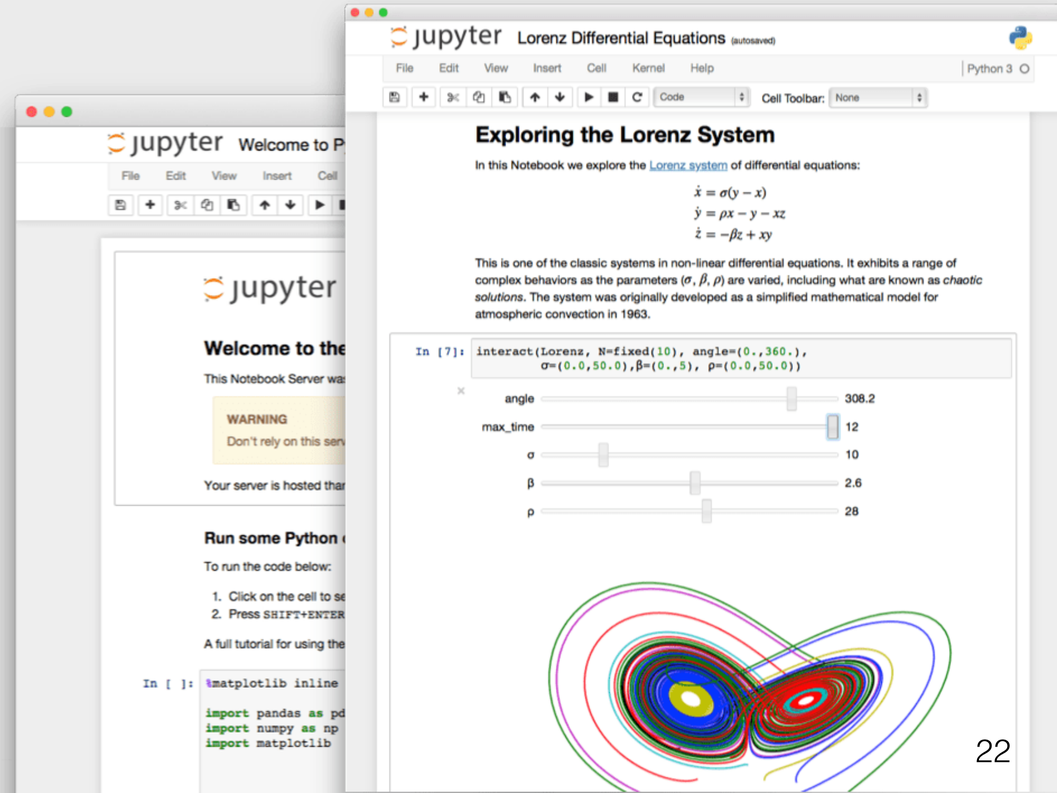


Dynamic Controls

aka “widgets”

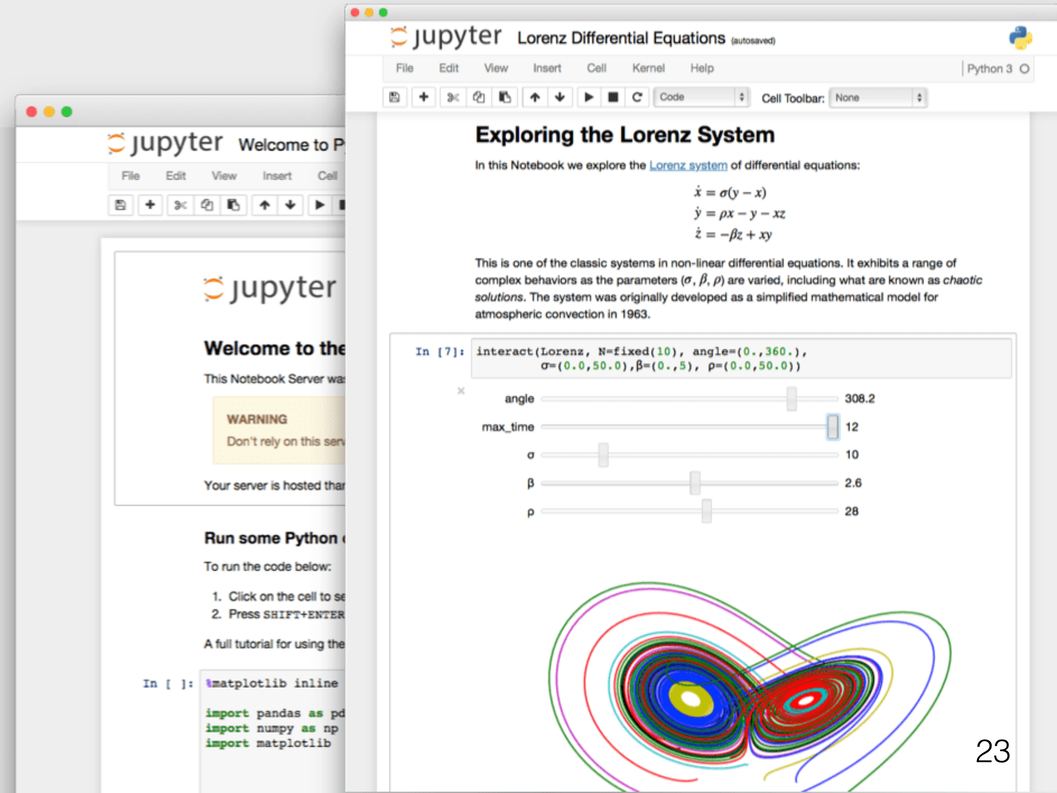
Web Based Notebook Application

- Web technologies are accessible.
  - Only need a web browser to control an HPC Cluster
  - Familiar to users
- Rapid increase in performance and functionality
  - V8, 3D, Wasm, ...
- Identical for local and remote use.
- Allow multiple domain collaboration

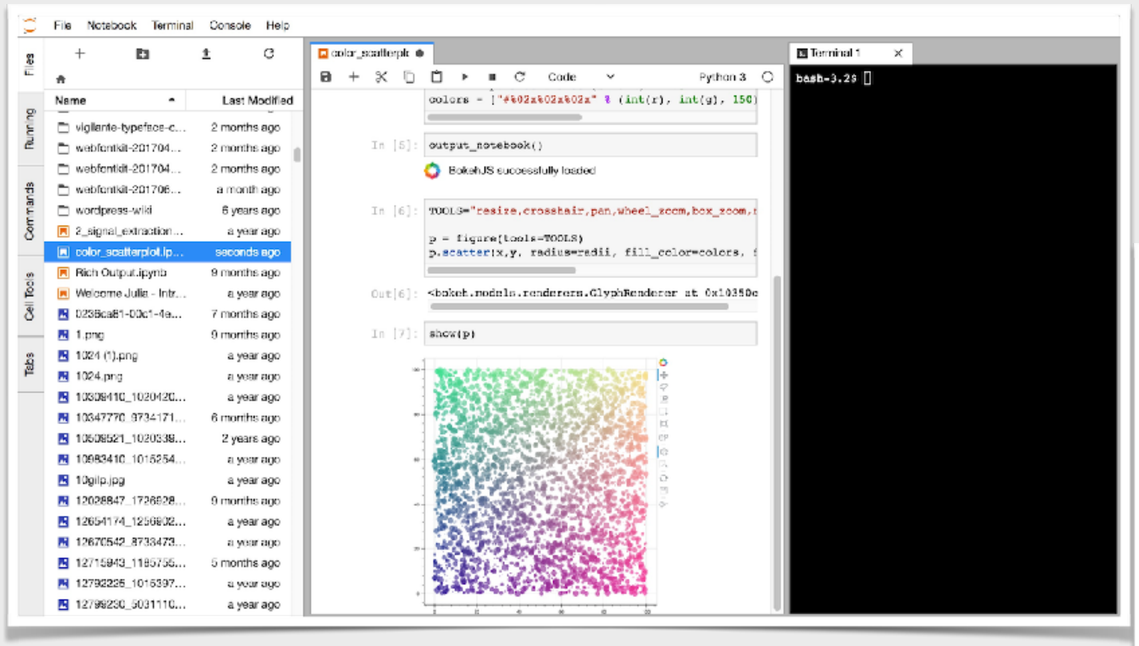


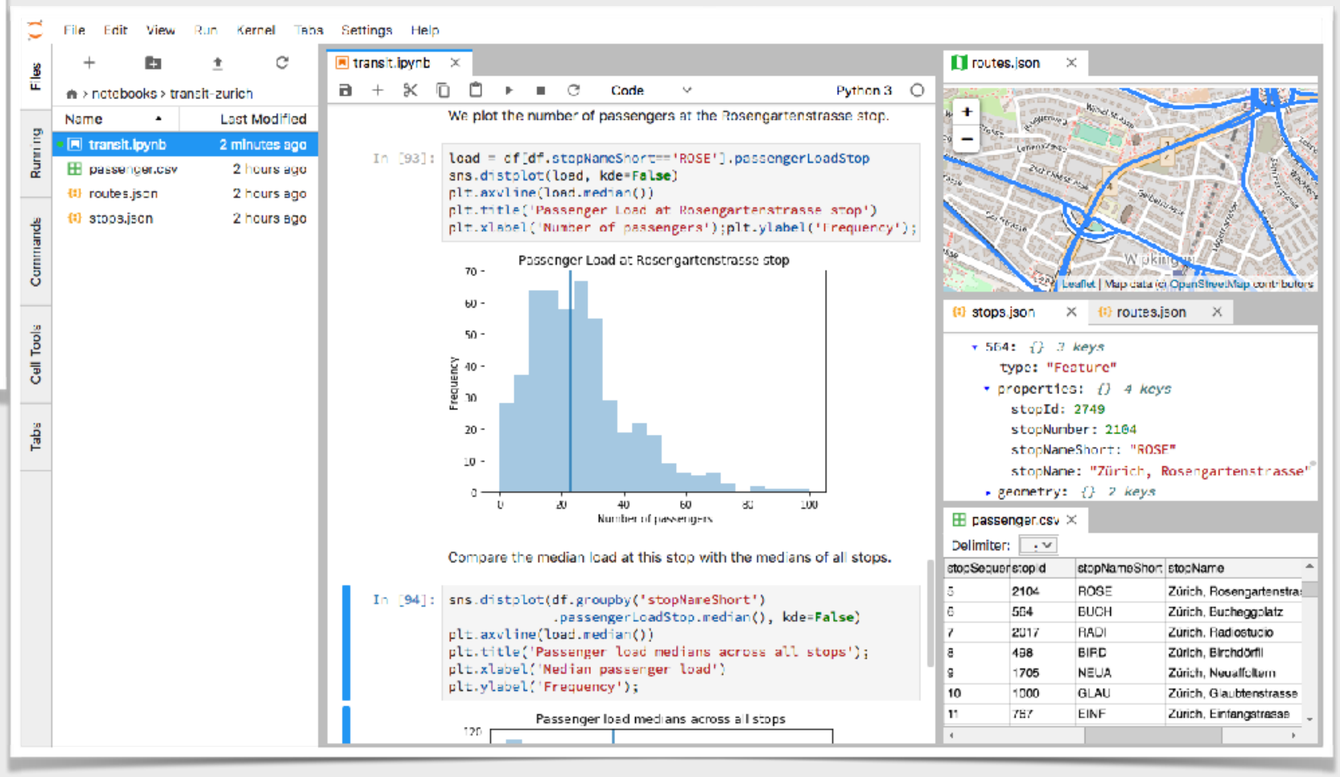
Open Notebook Document Format

- Notebooks get saved as JSON documents, which contain **narrative, code, and results**
  - Ubiquitous, JSON is readable in ~all languages.
  - Result embedding ensure trust (no Copy Past errors)
  - Make it easy to share and modify (Nbviewer, Binder)



JupyterLab

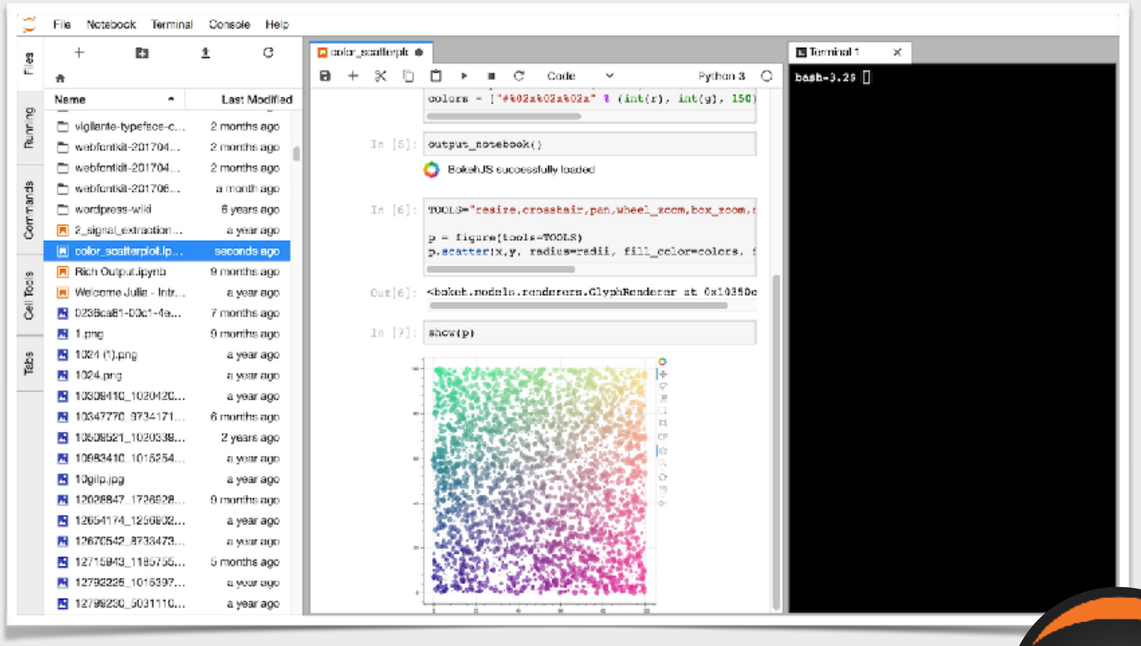


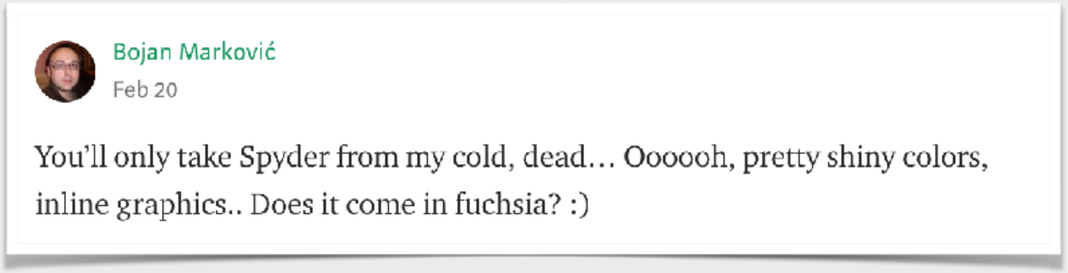


https://blog.jupyter.org/jupyterlab-is-ready-for-users-5a6f039b8906

JupyterLab

- Install Side by Side with Classic Notebook
- No Change in File Format, or protocol
  - Better Architecture (all extensions are first class)

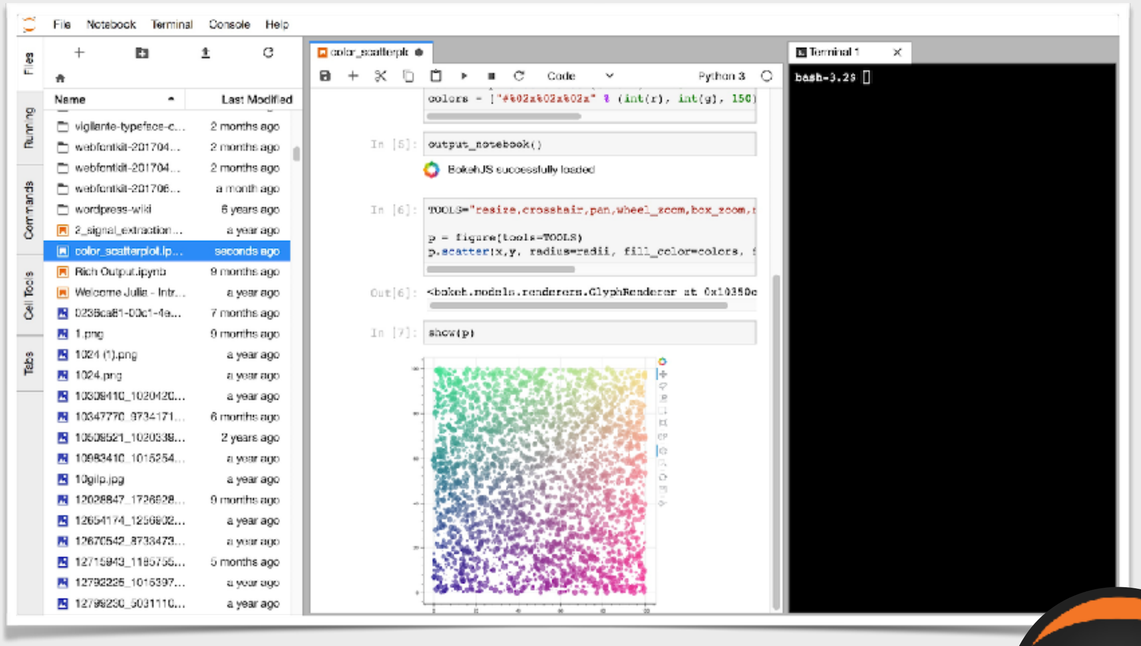




- Classic Notebook will be deprecated at some point

JupyterLab

- Ready for users
- Ready for Classic Extensions to be ported
- Will have all the new features

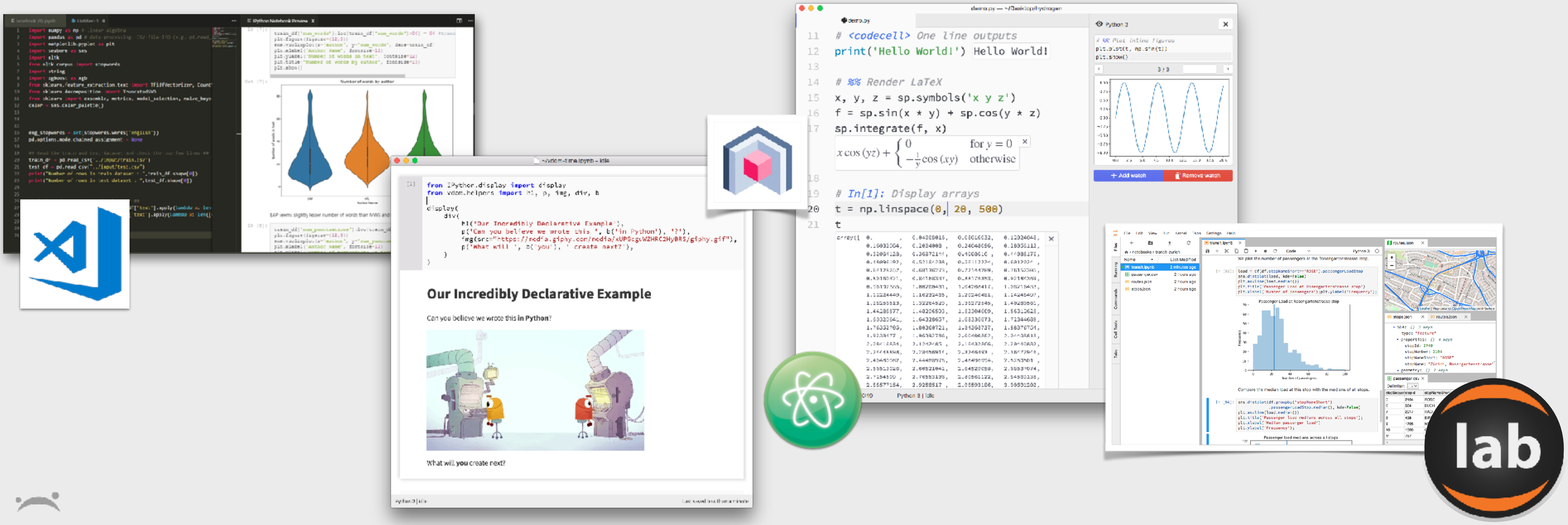


Many languages

...

Tools Integrations

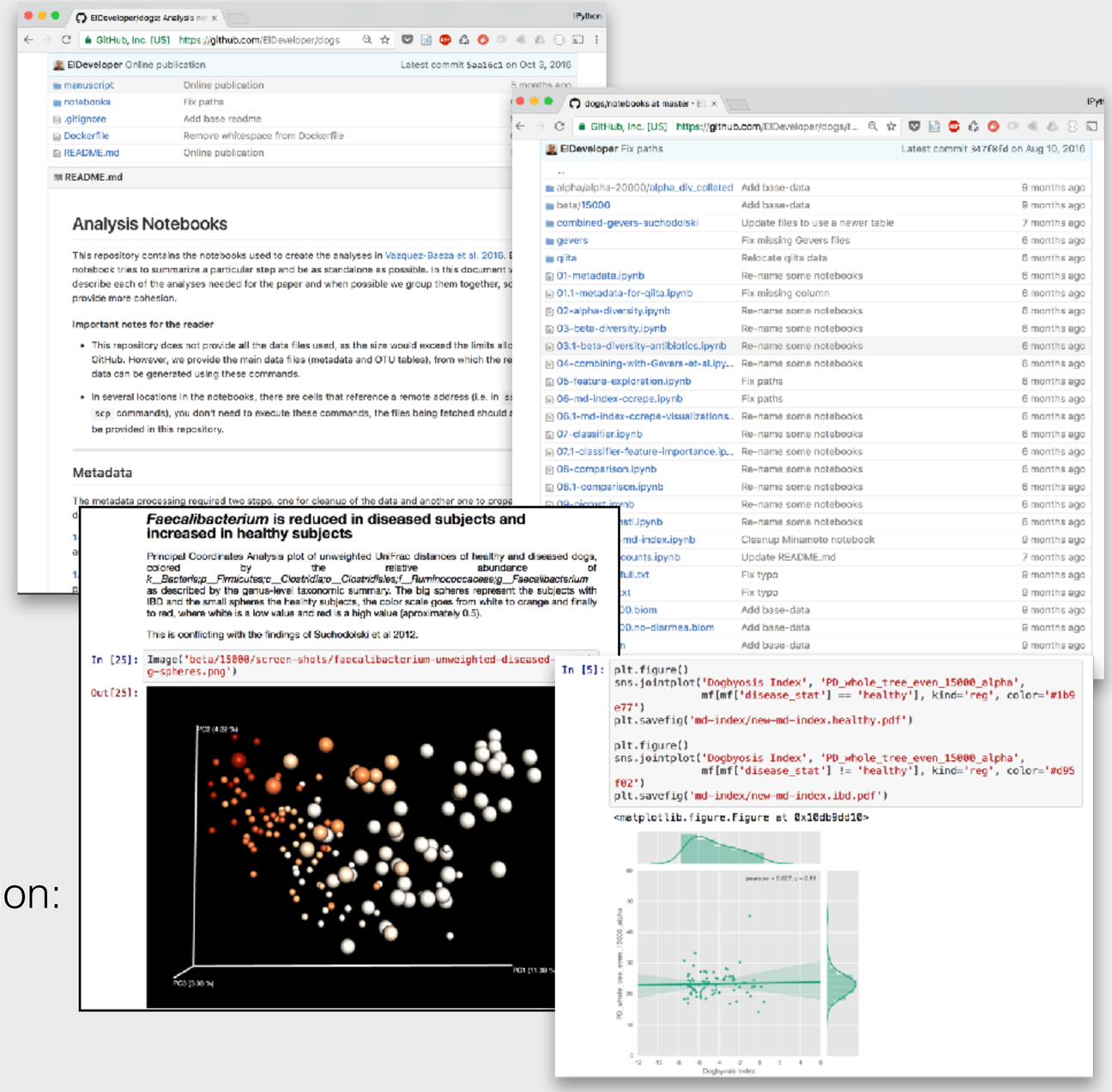
Frontends: Notebook, JupyterLab, CLI, Vim, Emacs, Visual Studio Code, Atom, Nteract, Juno...



Notebooks as supporting material and interactive companions

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, WaveLab and Reproducible Research, ***1995***



Advertising

Full code and specific discussion: executable supplementary materials:

“computational companions”

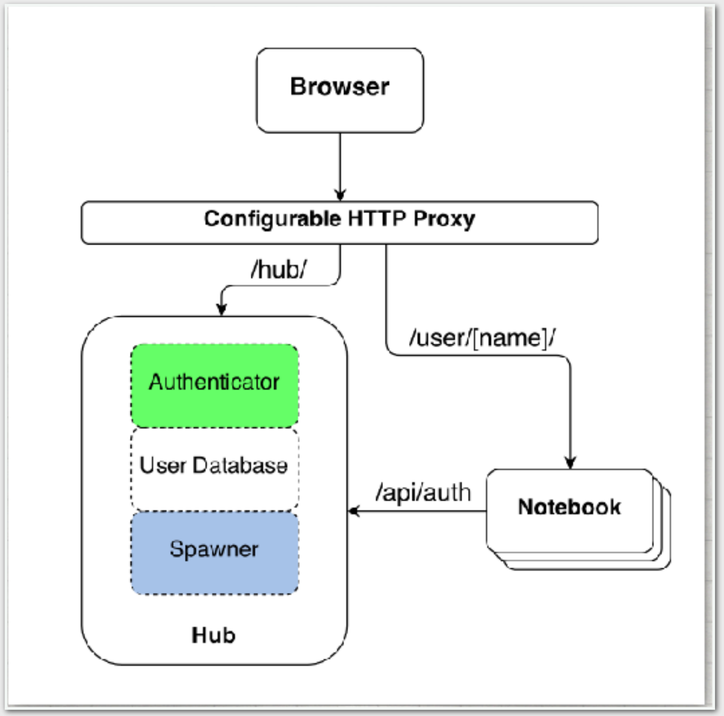
https://github.com/EIDeveloper/dogs

Executable books

Easy, Scalable Deployment

JupyterHub

- A notebook application is a **Single User** application
- Quick and easy multi-user deployments are critical to lower overhead.
- JupyterHub Provides way a simple way to deploy Jupyter at scale.
  - https://z2jh.jupyter.org/ for a guide.



Amazon SageMaker

R-Brain

Google Colaboratory

GRYD

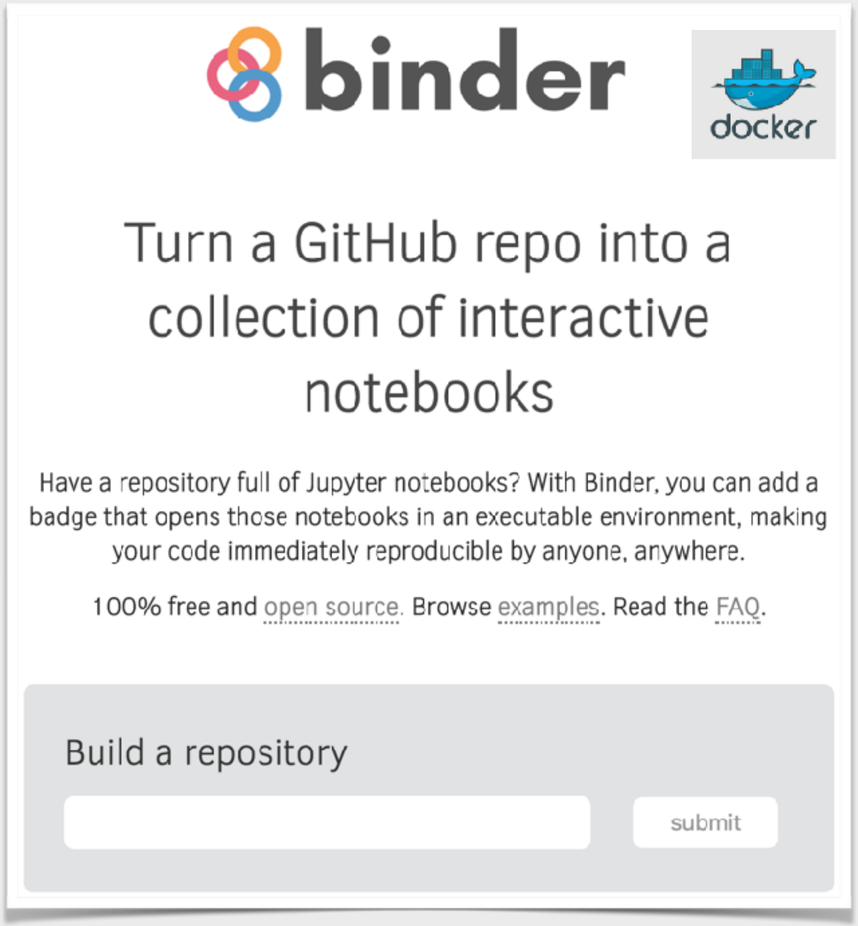
CODE OCEAN

Binder

- Technology which takes any GitHub repository with Jupyter notebooks
- Turn it into a Docker image to ensure reproducibility and quick deployment.
- Starts an isolated, ephemeral server in a few seconds, for user to interact with.

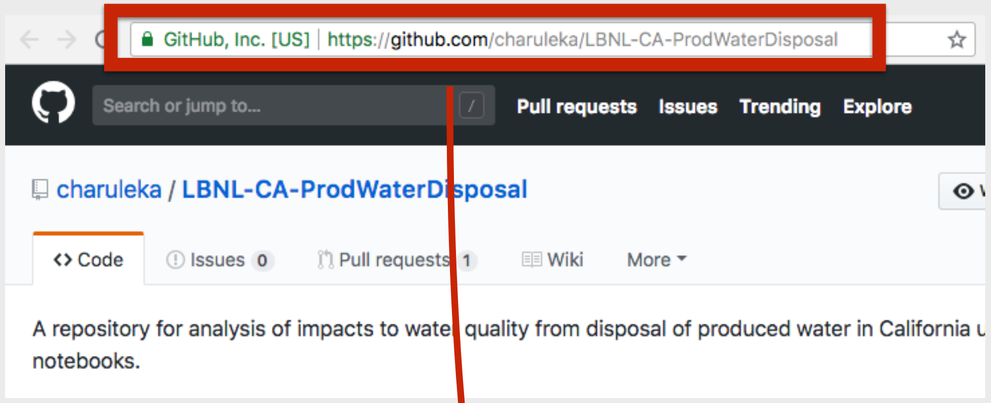
* Not limited to GitHub, Notebooks, Jupyter, Docker, or Ephemeral

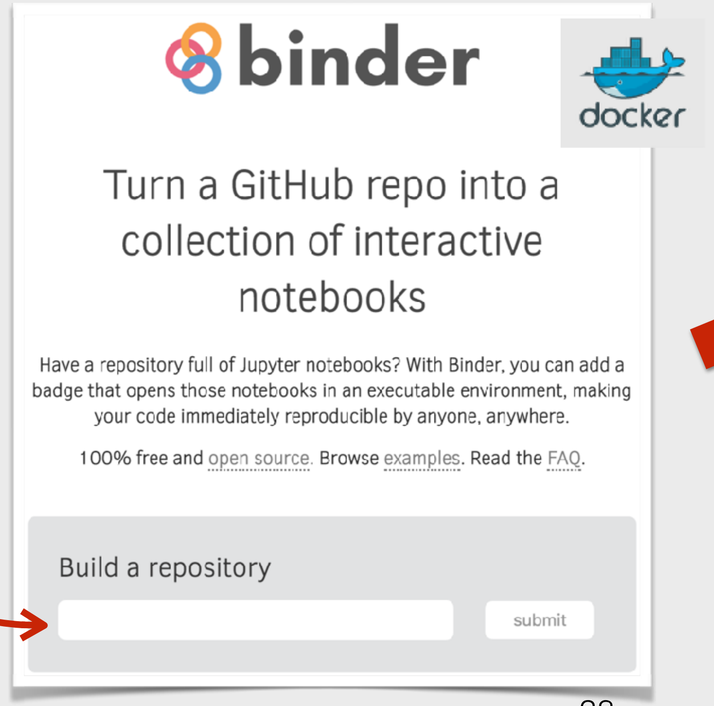
MyBinder.org

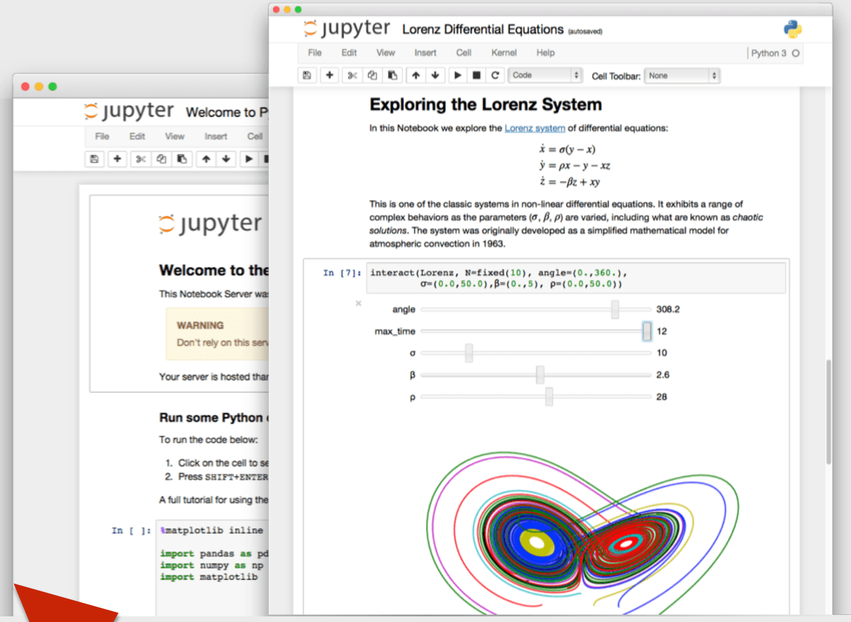


- One Public instance of Binder
  - mybinder.org
  - (stats at grafana.mybinder.org)
  - Limited CPU/Memory/network
- Anonymous Login
- Ephemeral (2h) and restricted to 50 parallel launch
- Build on demand
  - Caches images for fast launch

MyBinder.org

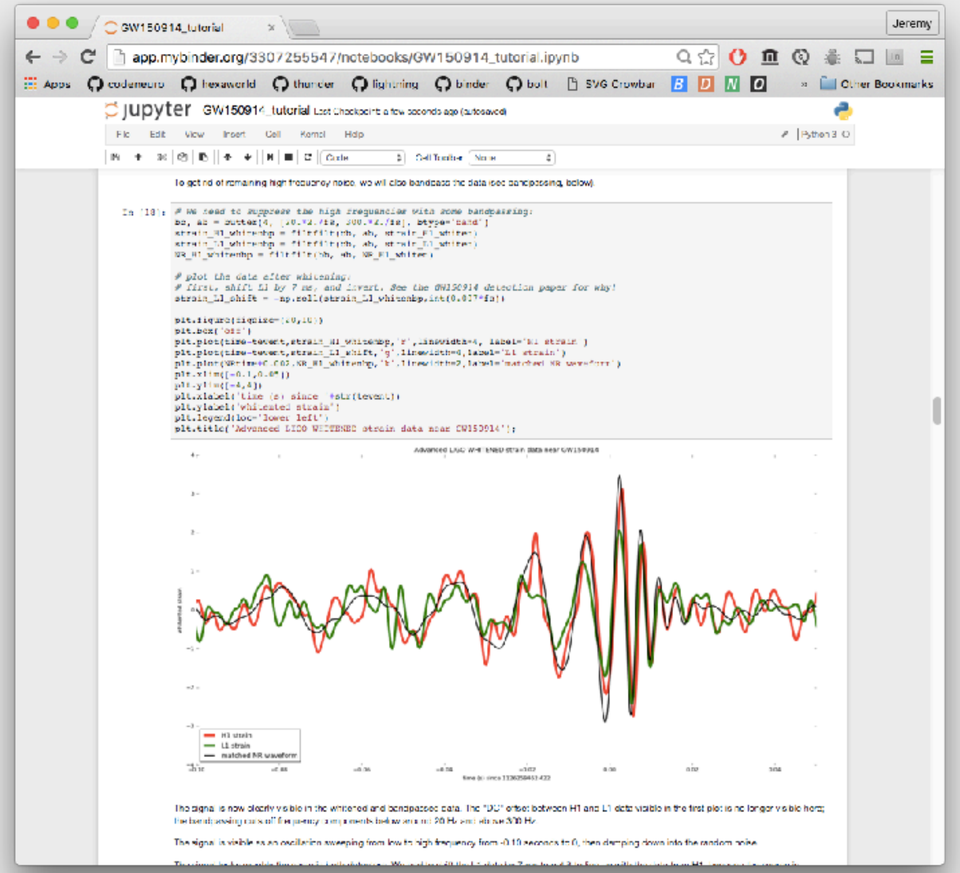






launch binder

Ligo binder



NbInteract

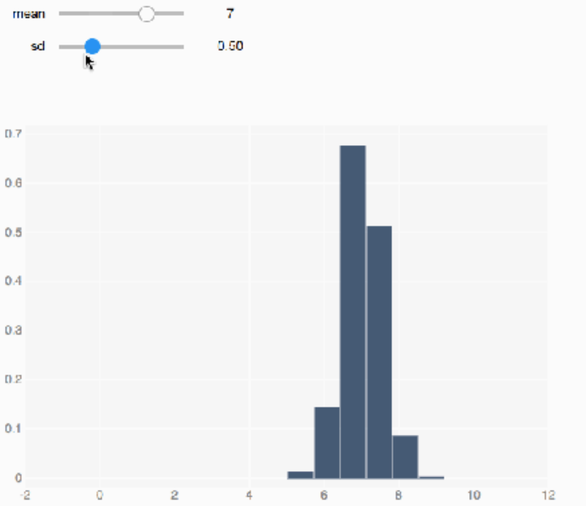
nbinteract.hist ¶

`hist` generates a histogram that allows interaction with the parameters for the response function.

`hist` takes in a single response function. The response function returns the array of numerical values that will be shown in the histogram. The `hist` function allows interaction with the response function's parameters by specifying them as keyword arguments in the same format as `ipywidgets.interact`. Any argument that can be used for `ipywidgets.interact` can be used for `hist`.

```
def hist_response_function(mean, sd, size=1000):
    '''
    Returns 1000 values picked at random from the normal
    distribution with the mean and SD given.
    '''
    return np.random.normal(loc=mean, scale=sd, size=1000)
```

```
nbi.hist(hist_response_function, mean=(0, 10), sd=(0, 2.0, 0.2))
```



dataChallenge

Not Secure | 0.0.0.0:49230/notebooks/dataChallenge.ipynb

jupyter dataChallenge Autosave Failed!

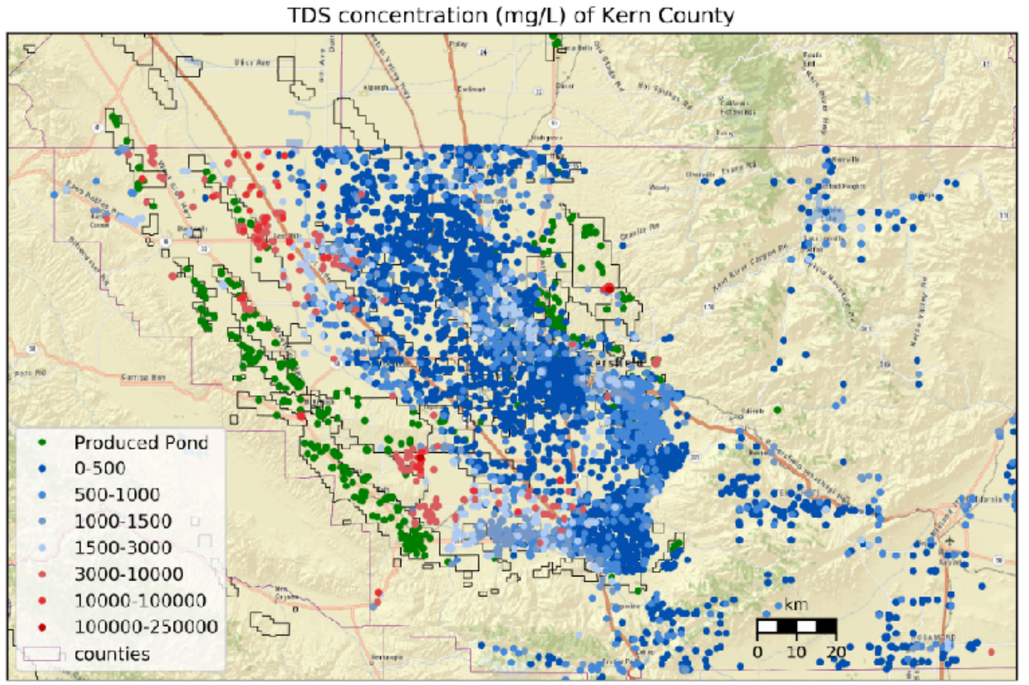
File Edit View Insert Cell Kernel Widgets Help

```
#make the figure pretty
plt.legend()
plt.title('TDS concentration (mg/L) of Kern County')
```

```
http://server.arcgisonline.com/ArcGIS/rest/services/ESRI_StreetMap_World_2D/MapServer/export?bbox=-13391734.74243081,4136735.049397425,-13135599.913606282,4300621.372044271&bboxSR=3785&imageSR=3785&size=1500,960&dpi=96&format=png32&f=image
```

Out[10]: `Text(0.5,1,'TDS concentration (mg/L) of Kern County')`

```
<Figure size 432x288 with 0 Axes>
```



We can see that TDS is generally higher to the west. This is most likely due to geological processes. The western side of the valley is composed of marine sediments which undergo geochemical reactions and produce higher salinity water.

Mann-Kendall (MK) Statistical Test

Now that we know where our ponds and TDS measurements are, we can begin to analyze them. We've chosen to use a Mann-Kendall test on locations with 3 or more data points.

In the Classroom

DataHub

datahub.berkeley.edu

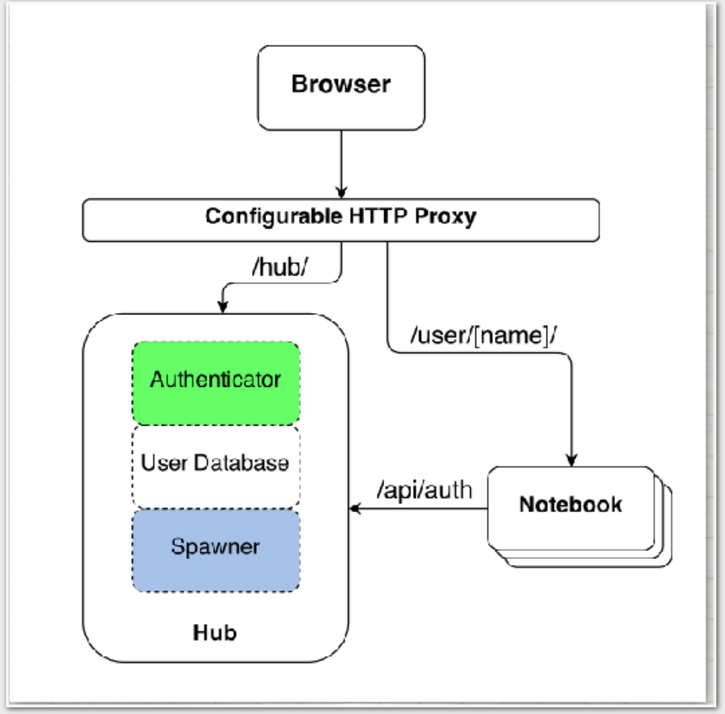
http://www.ds100.org/

Zero setup*

- Campus Wide deployment
- Login with Cal ID
- Can focus on Domain Knowledge

Students can still optionally install

Jupyter on their machine later on.



* At least for students

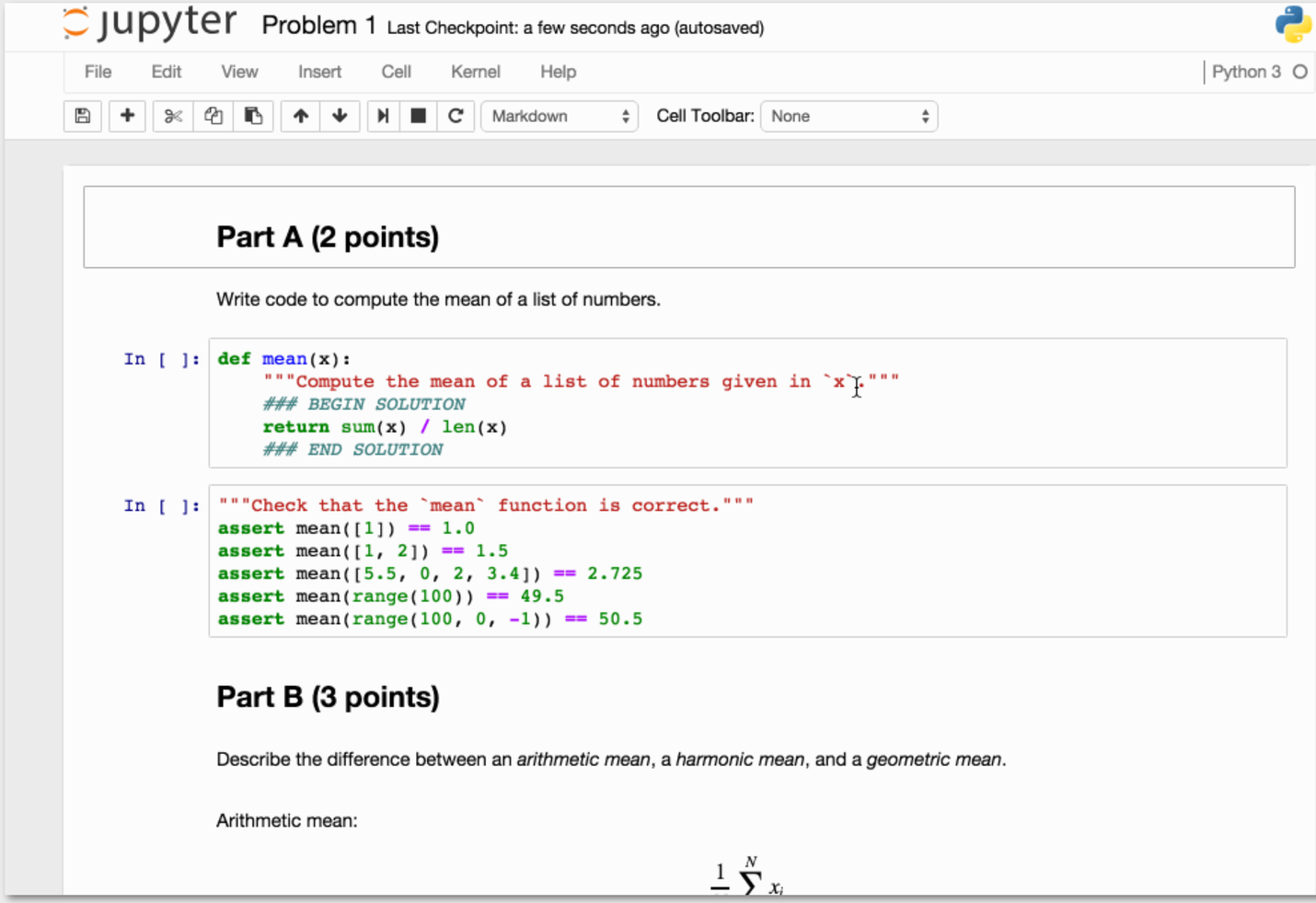
Autograding

- Via Nbgrader
- Generate “Students” version of the notebook
- Partial auto grading
  - Frees up a lo of TA time.

J. Hamrick
Deep Mind

B. Granger
Cal Poly

Fernando Pérez
UC Berkeley

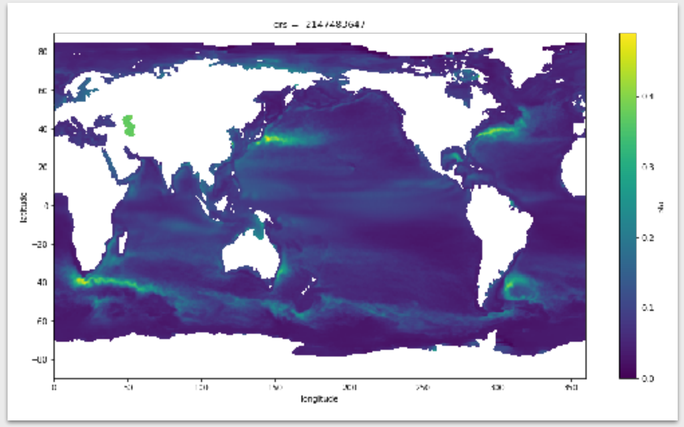


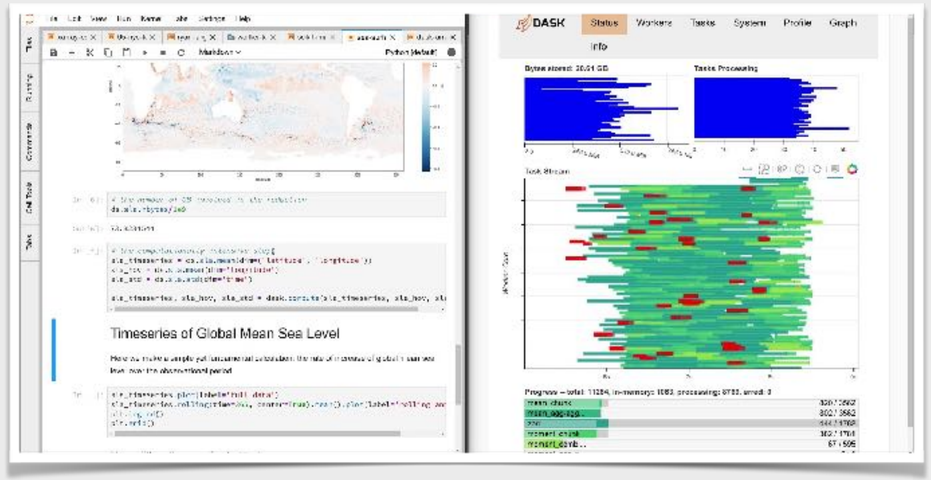
In the Cloud

1. Foster collaboration around the open source scientific python ecosystem for ocean / atmosphere / land / climate science.
2. Support the development with domain-specific geoscience packages.
3. Improve scalability of these tools to to handle petabyte-scale datasets on HPC and cloud platforms.

PANGEO

http://pangeo-data.org/

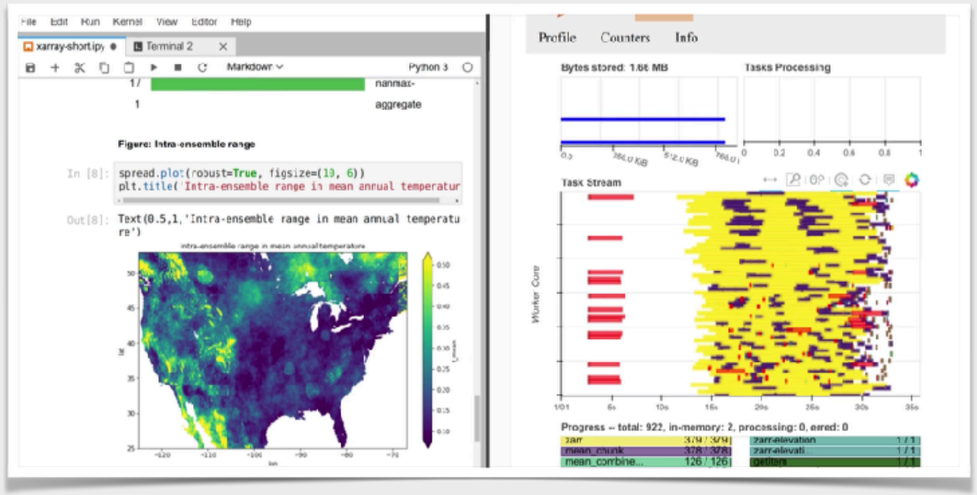




In the Cloud

http://pangeo-data.org/

- Completely managed JupyterHub on Kubernetes
  - http://pangeo.pydata.org/
- Login via GitHub
- Customized for GeoScience
- Persisting servers on Google Cloud,
  - Large amount of Ram/CPU/Nodes
  - Dynamically scalable

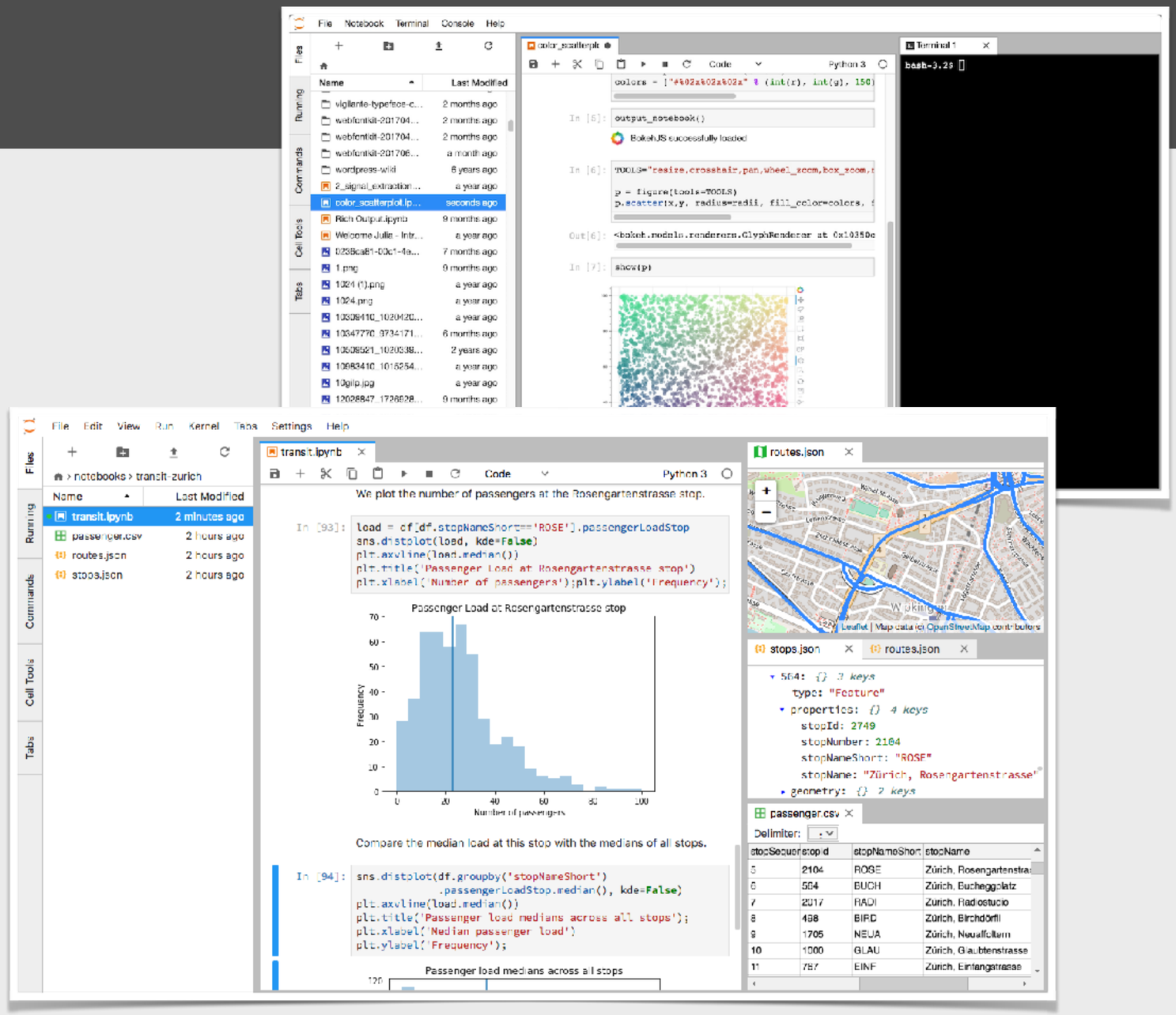


Demo at the end if time ?

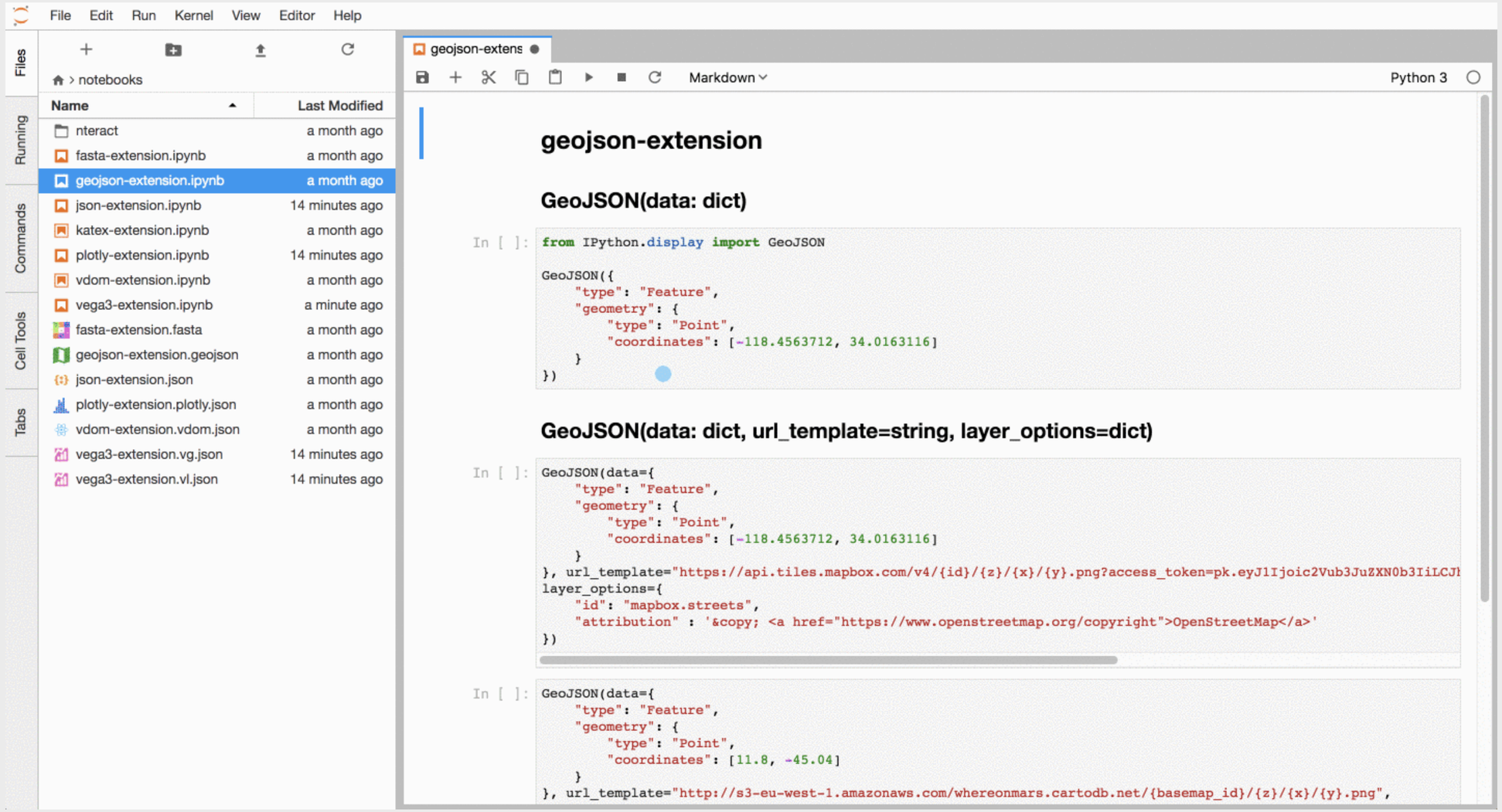
The Future

JupyterLab

- Extends the notebook interface with text editor, shell, ...etc
- Plugins are first class citizen
- Custom Layouts (Education-oriented layout ?)
- Realtime collaboration
  - Broadcast the teacher's notebook ?



https://blog.jupyter.org/jupyterlab-is-ready-for-users-5a6f039b8906


File Edit Run Kernel View Editor Help
Files
Running
Commands
Cell Tools
Tabs
notebooks
Name
Last Modified
nteract a month ago
fasta-extension.ipynb a month ago
geojson-extension.ipynb a month ago
json-extension.ipynb 14 minutes ago
katex-extension.ipynb a month ago
plotly-extension.ipynb 14 minutes ago
vdom-extension.ipynb a month ago
vega3-extension.ipynb a minute ago
fasta-extension.fasta a month ago
geojson-extension.geojson a month ago
json-extension.json a month ago
plotly-extension.plotly.json a month ago
vdom-extension.vdom.json a month ago
vega3-extension.vg.json 14 minutes ago
vega3-extension.vl.json 14 minutes ago
geojson-extens
Markdown
Python 3
geojson-extension
GeoJSON(data: dict)
In [ ]:
from IPython.display import GeoJSON
GeoJSON({
"type": "Feature",
"geometry": {
"type": "Point",
"coordinates": [-118.4563712, 34.0163116]
}
})
GeoJSON(data: dict, url_template=string, layer_options=dict)
In [ ]:
GeoJSON(data={
"type": "Feature",
"geometry": {
"type": "Point",
"coordinates": [-118.4563712, 34.0163116]
}
}, url_template="https://api.tiles.mapbox.com/v4/{id}/{z}/{x}/{y}.png?access_token=pk.eyJ1Ijoic2Vub3JuZXN0b3IiLCJ
layer_options={
"id": "mapbox.streets",
"attribution" : '© OpenStreetMap'
})
In [ ]:
GeoJSON(data={
"type": "Feature",
"geometry": {
"type": "Point",
"coordinates": [11.8, -45.04]
}
}, url_template="http://s3-eu-west-1.amazonaws.com/whereonmars.cartodb.net/{basemap_id}/{z}/{x}/{y}.png",


O'REILLY
jupytercon
Brought to you by NumFOCUS Foundation and O'Reilly Media Inc.
August 21-25, 2018
New York, NY
jupytercon.com

Pangeo Demo and Questions...