IP[y]:
IPython

Berkeley
Division of
Data Sciences

jupyter

Open Source Software in Science: Beyond the Code

Fernando Pérez

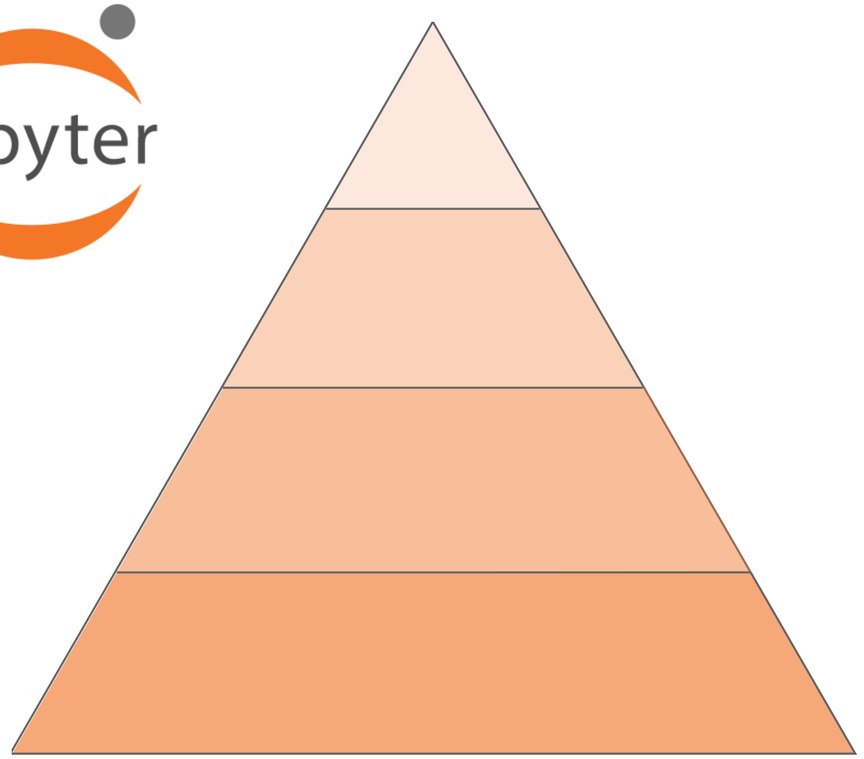
Lindsey Heagy

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

University of California, Berkeley
DEPARTMENT OF STATISTICS

BERKELEY LAB

OSS: more than software

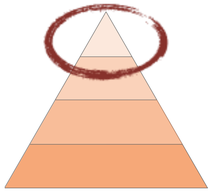


Services and content

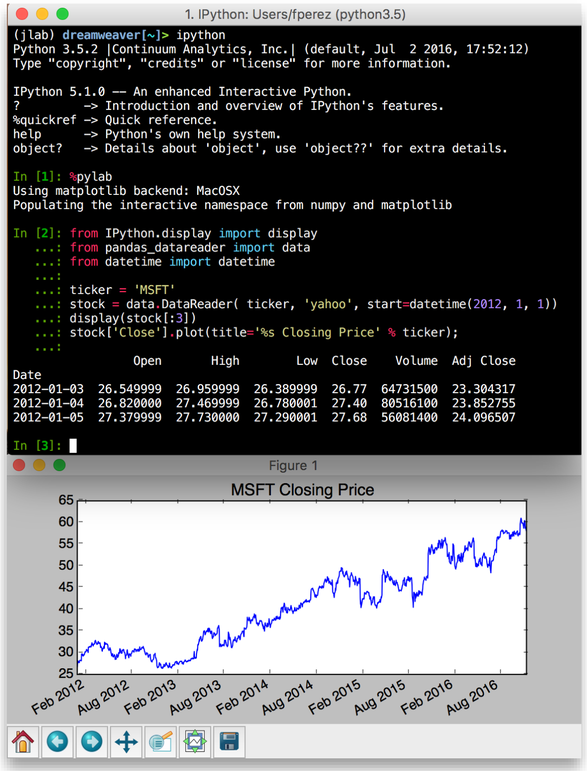
Software

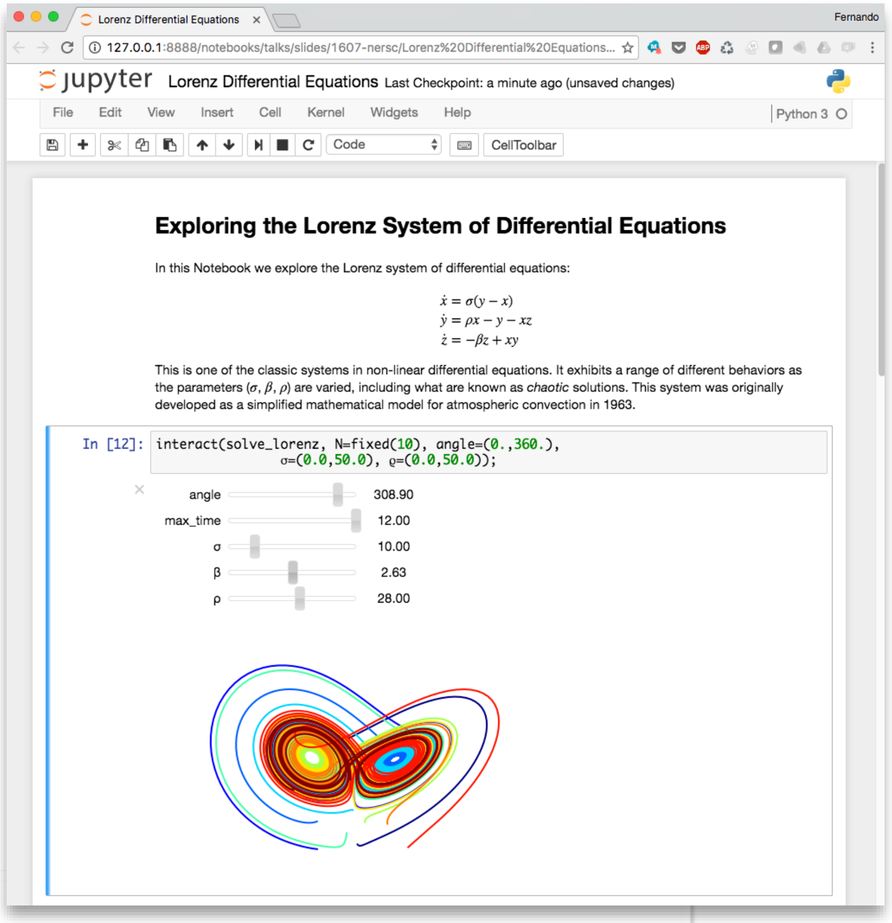
Standards and Protocols

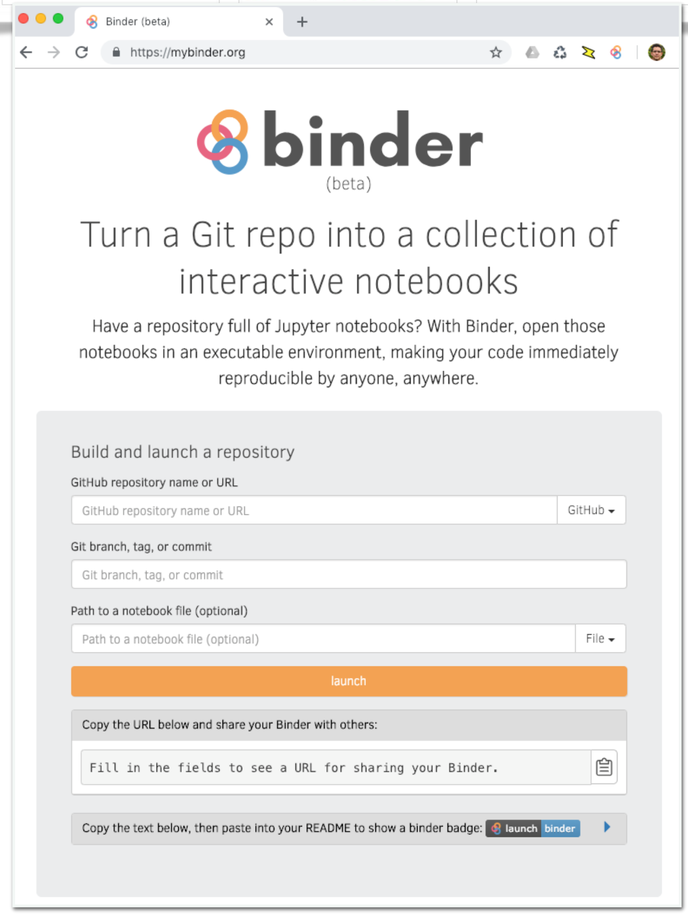
Community



Content/Services

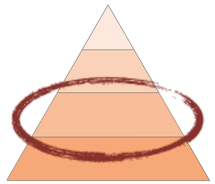




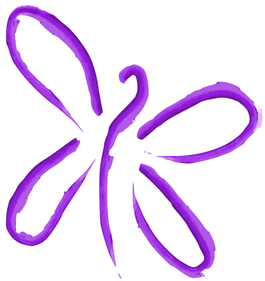


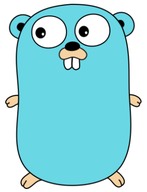
Jupyter for Organizations

JupyterHub is a multiuser version of the notebook designed for centralized deployments in companies, university classrooms and research labs.



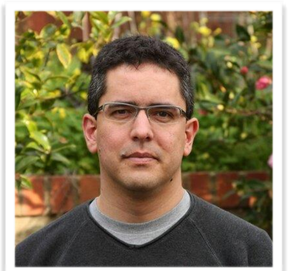
A language agnostic protocol





~100 different kernels: https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

Community: formalized governance



Me :)

Brian Granger

Cal Poly, Amazon

Institutional Partners

Institutional Partners are organizations that support the project by employing Jupyter Steering Council members. Current Institutional Partners include:

ANACONDA. · Bloomberg · NETFLIX · CAL POLY SAN LUIS OBISPO

Berkeley UNIVERSITY OF CALIFORNIA · QuantStack Scientific Computing · TWO SIGMA · JPMORGAN CHASE & CO.

UNIVERSITY OF CALIFORNIA MERCED · aws

Sponsors

Project Jupyter receives direct funding from the following sources:

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST · ALFRED P. SLOAN FOUNDATION · GORDON AND BETTY MOORE FOUNDATION · Google

rackspace the #1 managed cloud company

fastly · Horizon 2020 European Union funding for Research & Innovation · Microsoft · QUANSIGHT

SCHMIDT FUTURES

Steering Council

The role of the Jupyter Steering Council is to ensure, through working with and serving the broader Jupyter community, the long-term well-being of the project, both technically and as a community. The Jupyter Steering Council currently consists of the following members (in alphabetical order).

Damian Avila

Anaconda, Inc.
@damianavila on GitHub

Matthias Bussonnier

UC Merced
@Carreau on GitHub

Sylvain Corlay

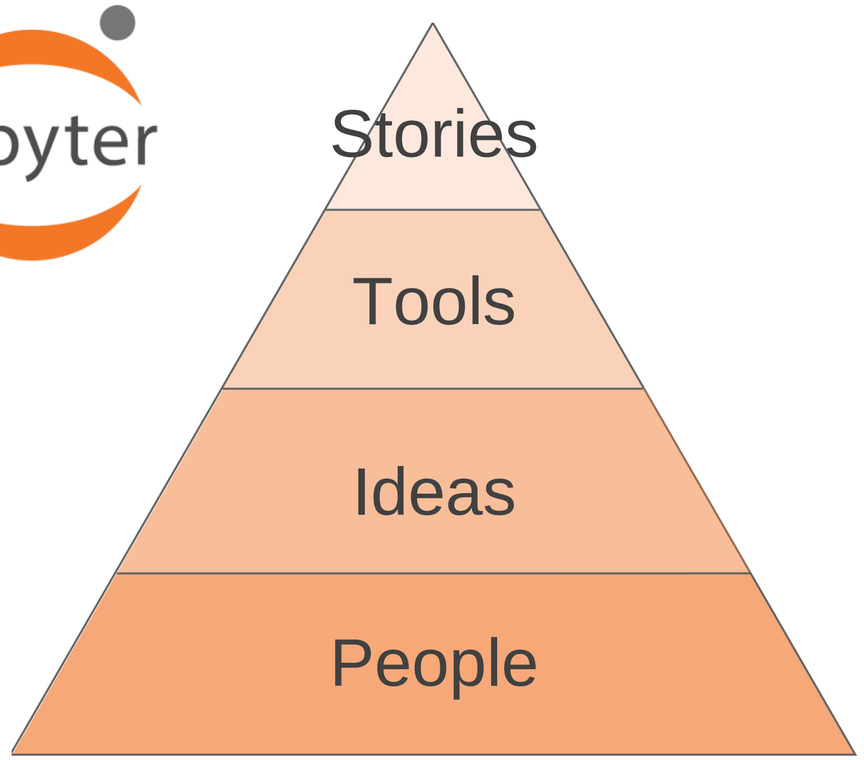
QuantStack
@sylvaincorlay on GitHub

Formal fiscal sponsorship

NUMFOCUS

OPEN CODE = BETTER SCIENCE

More than software, woven into science



Services and content: impact

Software

Standards and Protocols: ecosystem

Community: innovation & resiliency

OSS supports CORE Science*

Collaborative

Open

Reproducible

Extensible

Lindsey Heagy

* With a nod to the FAIR principles of open data

Collaborative?

Multiple stakeholders, team effort

- Academic scientists
- Educators
- Industry
- Government
- Media/journalism
- 1500+ community volunteers!

Jupyter meets the Earth: newly funded NSF grant - $2M/3y

Research use-cases

- Climate data analysis
- Hydrology
- Geophysics

PANGEO

Tech Developments

- Data discovery
- Interactivity
- Cloud/HPC infrastructure

Fernando Perez

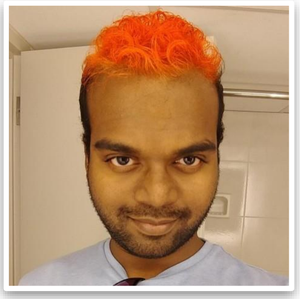
Joe Hamman

Laurel Larsen

Kevin Paul

Lindsey Heagy

Chris Holdgraf



Yuvi Panda

Open?

Dimensions of Openness

- Open source code
- Open (FAIR) data
- Open access publications & artifacts
- Open standards: interoperability (even with proprietary tools)
- Open community: all welcome (and mean it!)
- …

Reproducible?
The foundation of collaboration!

mybinder.org: shareable reproducibility

Origins:

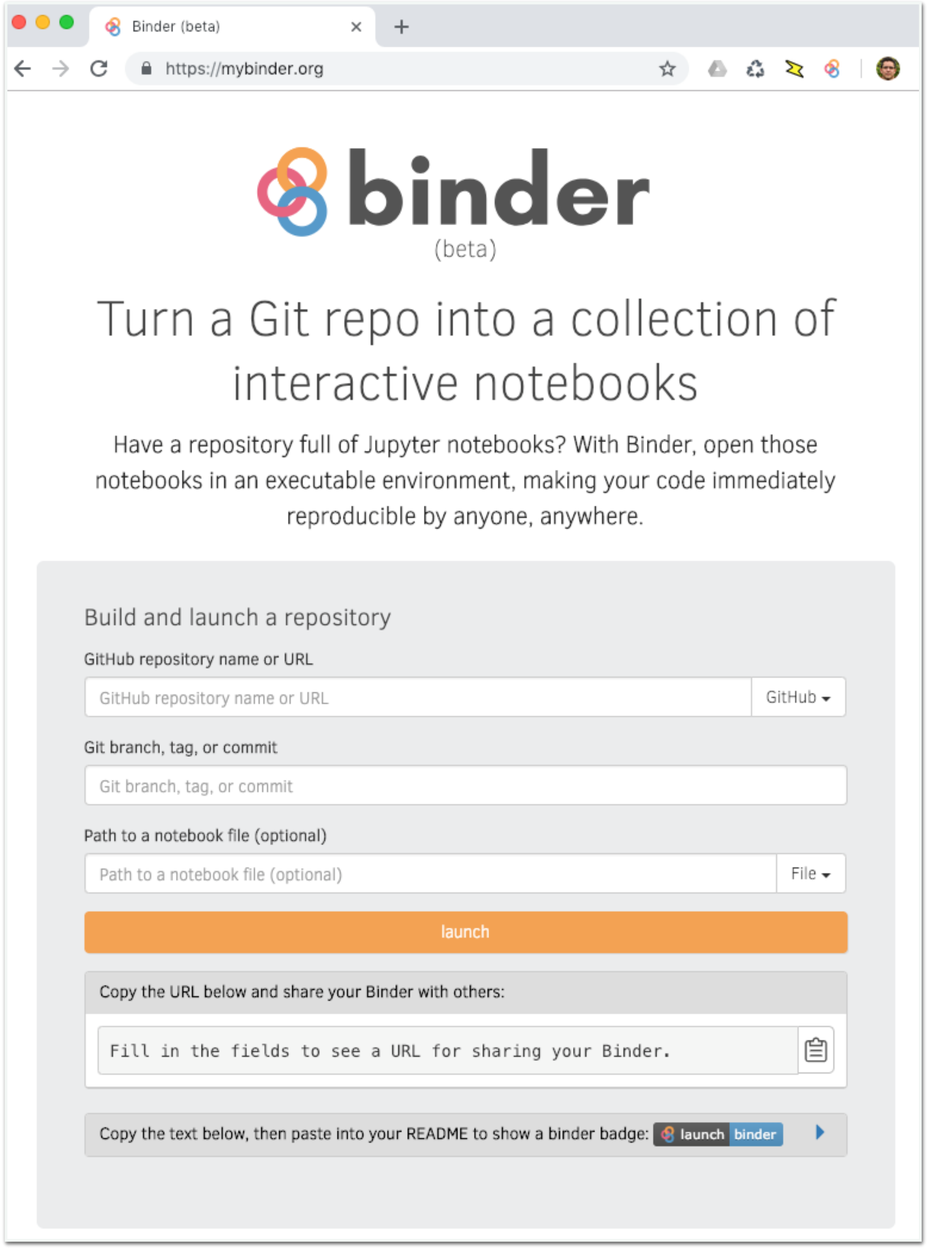
Jeremy Freeman's lab at Janelia farm.

That "incentives" business…

github.com/freeman-lab

cнkey contributor!

Tim Head @betatim



jupyter

GitHub

+

Explicit Dependencies

+

docker

kubernetes

Black holes! LIGO, Sept 14, 2015

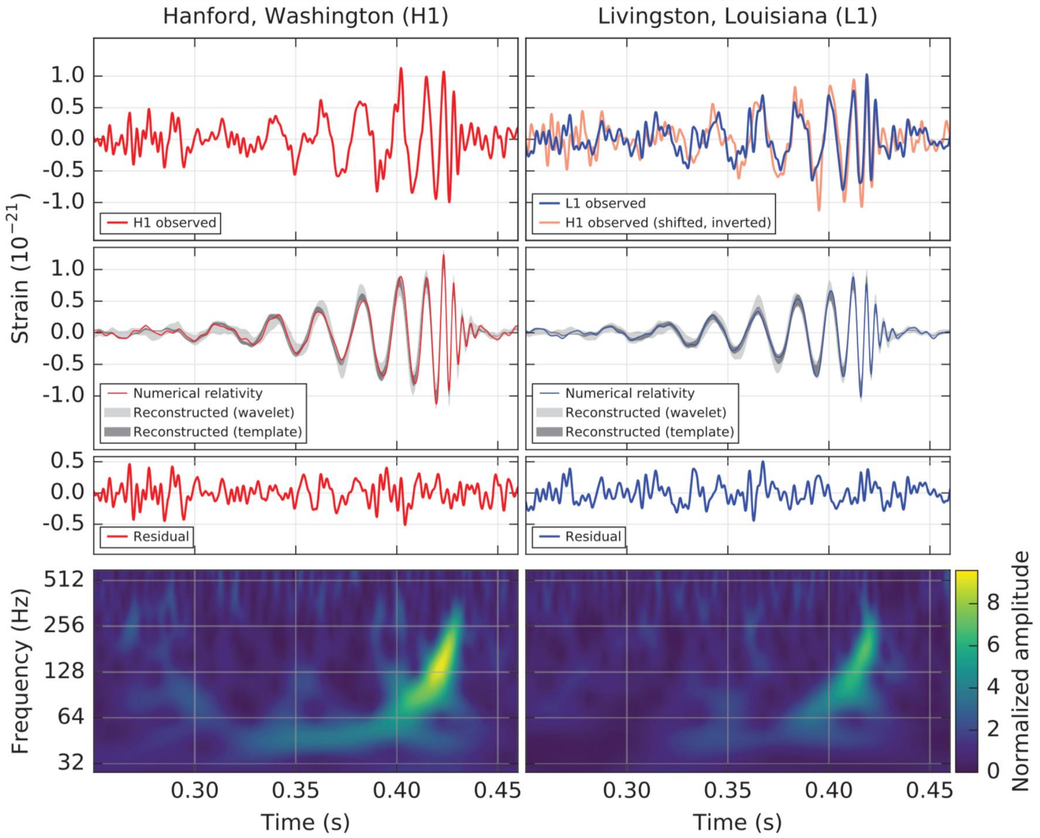
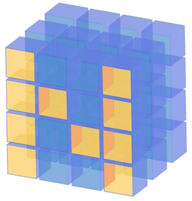


FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors' most sensitive frequency band, and band-reject



Make sound files

Make wav (sound) files from the filtered, downsampled data, +-2s around the event.

```
# make wav (sound) files from the whitened data, +-2s around the event.
from glob import glob
from IPython.display import display, Audio

from scipy.io import wavfile

# function to keep the data within integer limits, and write to wavfile:
def write_wavfile(filename,fs,data):
    d = np.int16(data/np.max(np.abs(data)) * 32767 * 0.9)
    wavfile.write(filename,int(fs), d)

tevent = 1126259462.422          # Mon Sep 14 09:50:45 GMT 2015
deltat = 2.                      # seconds around the event

# index into the strain time series for this time interval:
indxt = np.where((time >= tevent-deltat) & (time < tevent+deltat))

# write the files:
write_wavfile("GW150914_H1_whitenbp.wav",int(fs), strain_H1_whitenbp[indxt])
write_wavfile("GW150914_L1_whitenbp.wav",int(fs), strain_L1_whitenbp[indxt])
write_wavfile("GW150914_NR_whitenbp.wav",int(fs), NR_H1_whitenbp)

for wav in glob('*whitenbp.wav'):
    display(wav)
    display(Audio(filename=wav))
```

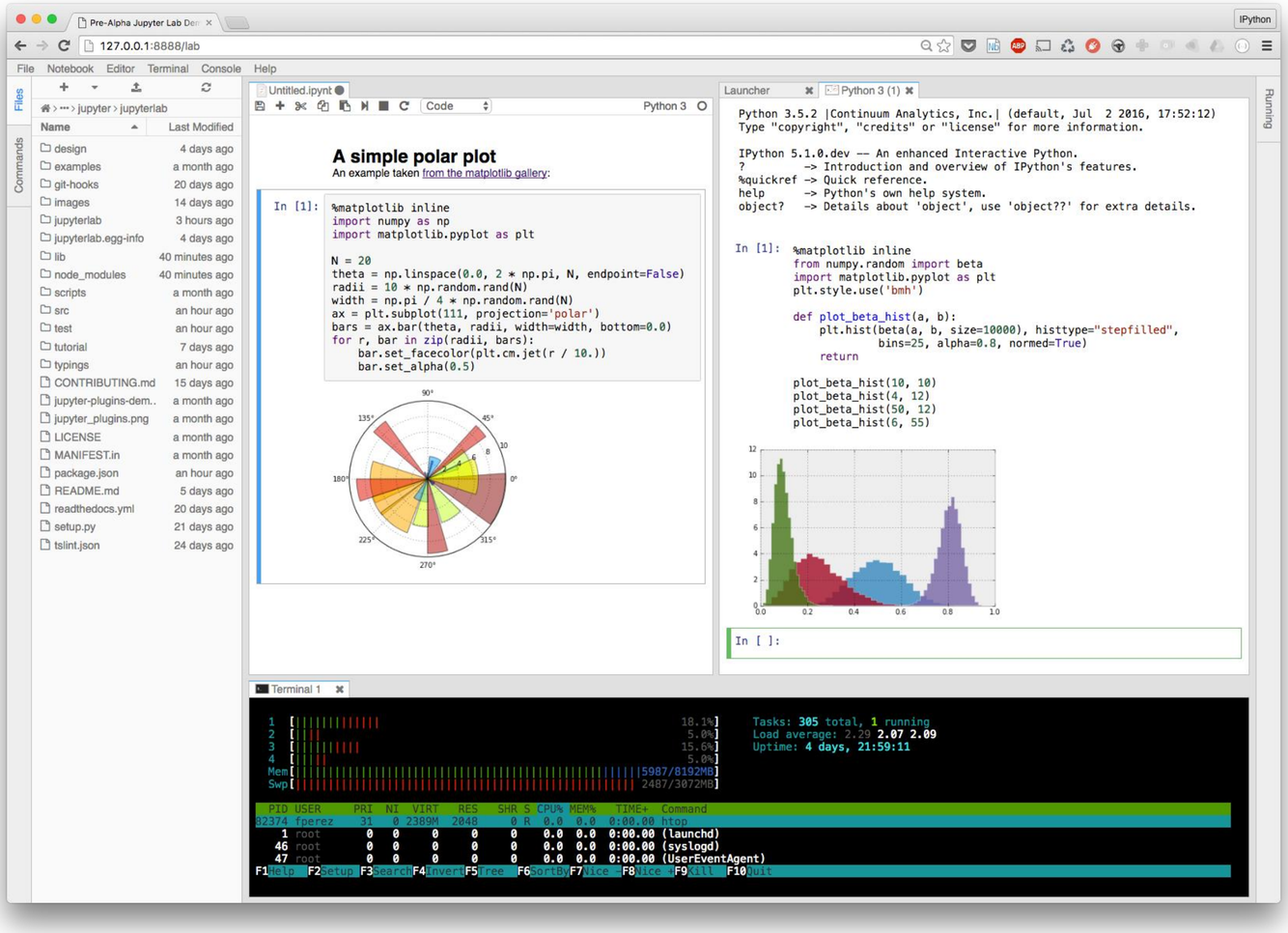
```
'GW150914_H1_whitenbp.wav'
```

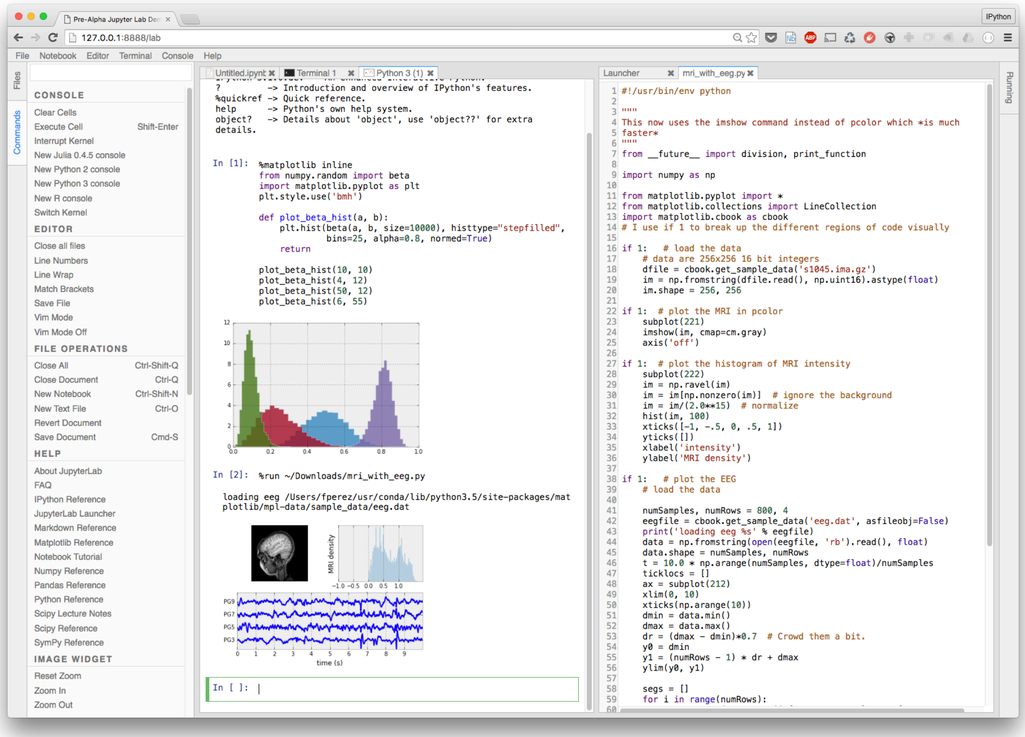
0:00

http://bit.ly/black-holes-woop

Extensible?

JupyterLab: a grand unified theory of Jupyter





Huge Team Effort!

C. Colbert, S. Corlay, A. Darian, B. Granger, J. Grout, P. Ivanov, I. Rose, S. Silvester, C. Willing, J. Zosa-Forde …

ANACONDA.

Bloomberg

JupyterLab is extensible: FlyBrainLab

An Interactive Computing Platform for the Fly Brain

BIONET Group, Columbia University

http://www.bionet.ee.columbia.edu

Aurel A. Lazar (PI)

Tingkai Liu

Mehmet K. Turkcan

Chung-Heng Yeh

Yiyin Zhou

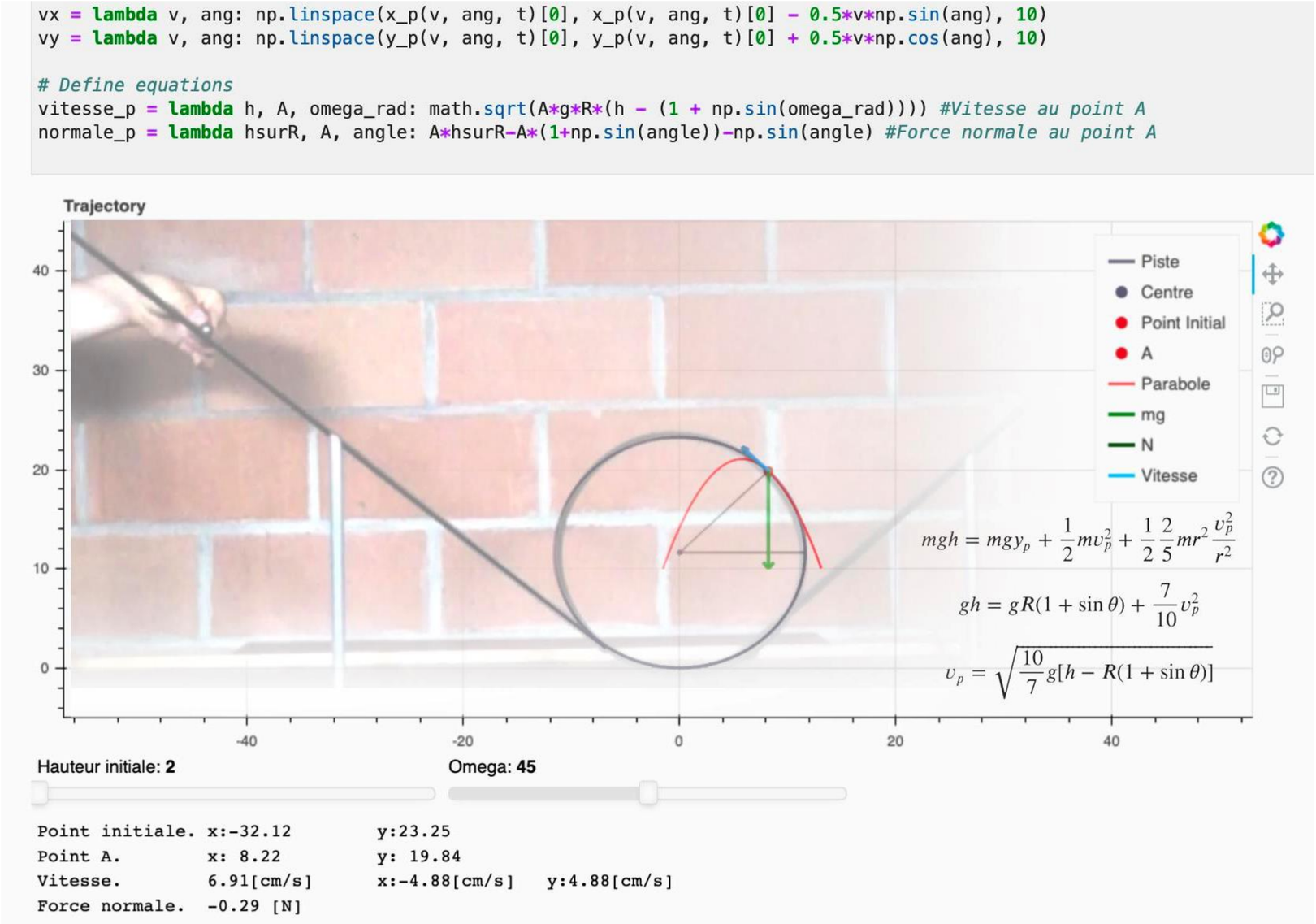
http://fruitflybrain.org

Teaching with Programmable Notebooks

EPFL

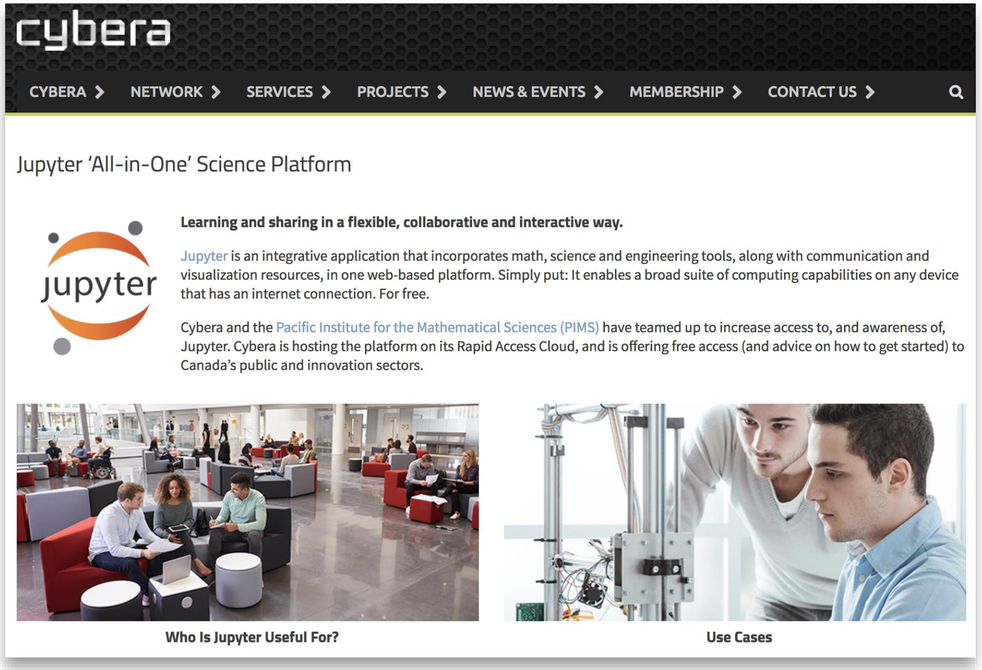
Launched in September, NOTO (http://noto.epfl.ch), EPFL's new JupyterLab platform for education, allows teachers and students to create and share programmable notebooks.

NOTO



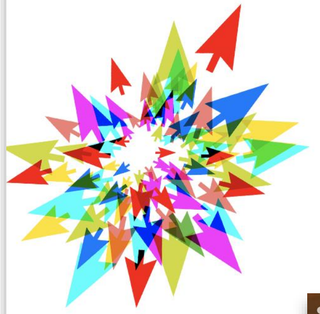
https://actu.epfl.ch/news/teaching-with-online-programmable-notebooks

CA National infrastructure, from K-12 to HPC



Pacific Institute *for the* Mathematical Sciences

compute canada | calcul canada



Home

About

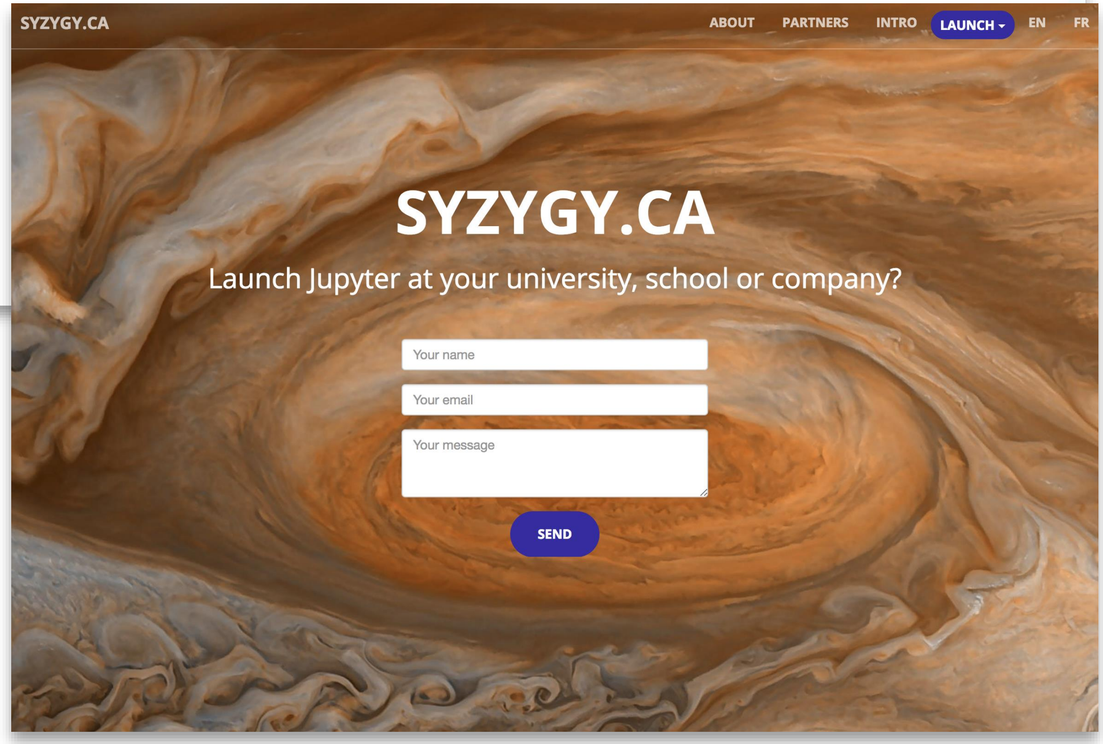
Renewing Canada's Advanced Research Computing Platform

Research Data Management

News

15/03/2017 Featured, News

compute canada and pims launch jupyter service for researchers



J. Colliander,
I. Allison,
B. Carra

Microsoft

Microsoft Azure

COCALC

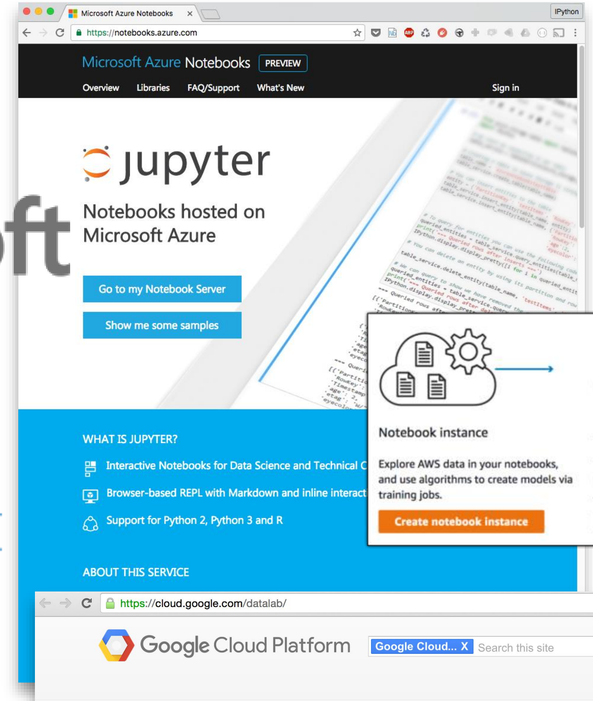
CODE OCEAN

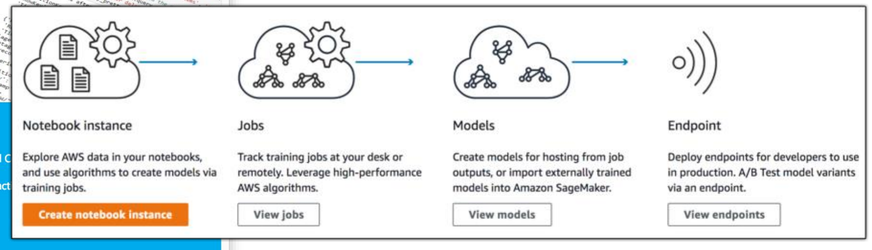
Google

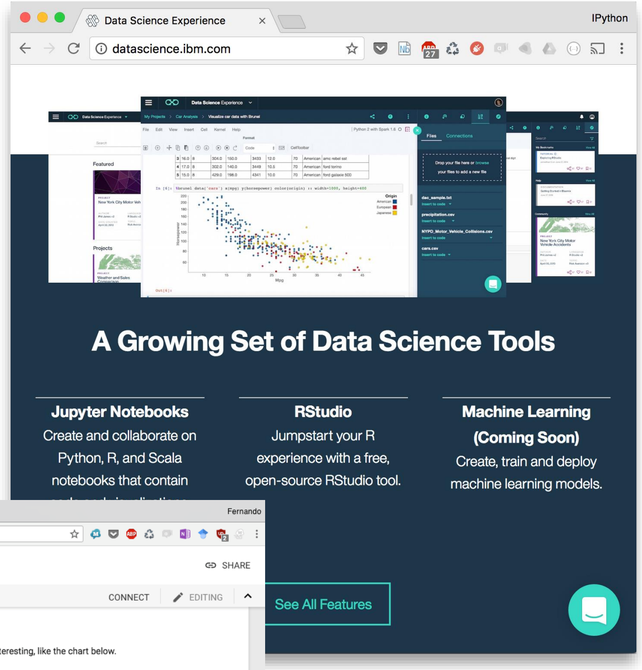
IBM

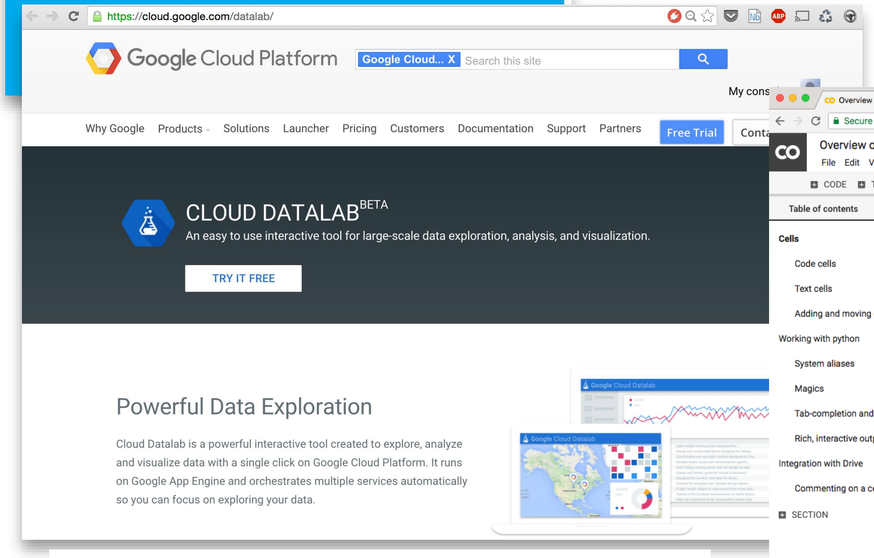
Amazon SageMaker

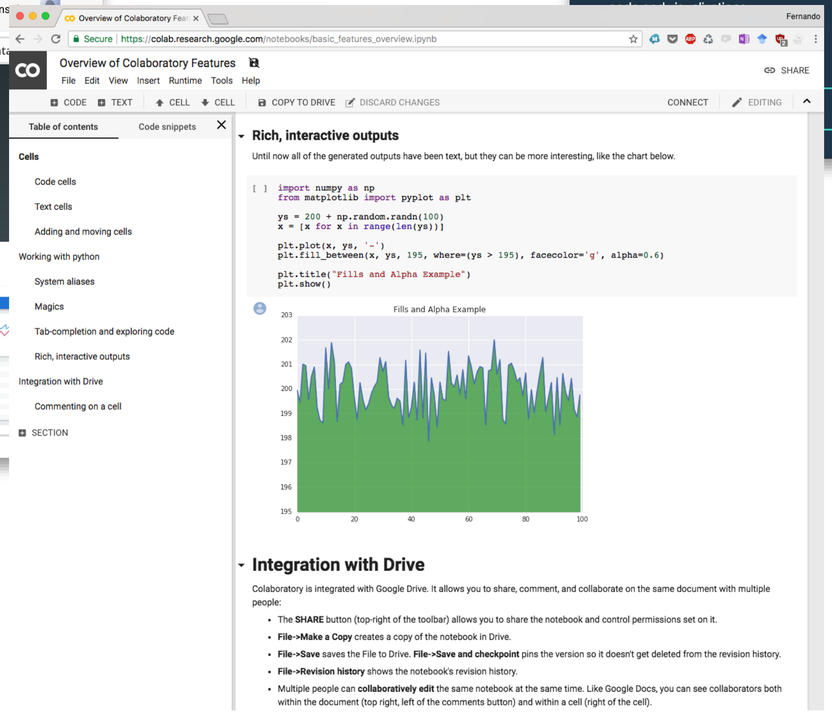
ANACONDA®











Impact: Research and Education

Data 8: Foundations of Data Science

Berkeley
Division of
Data Sciences

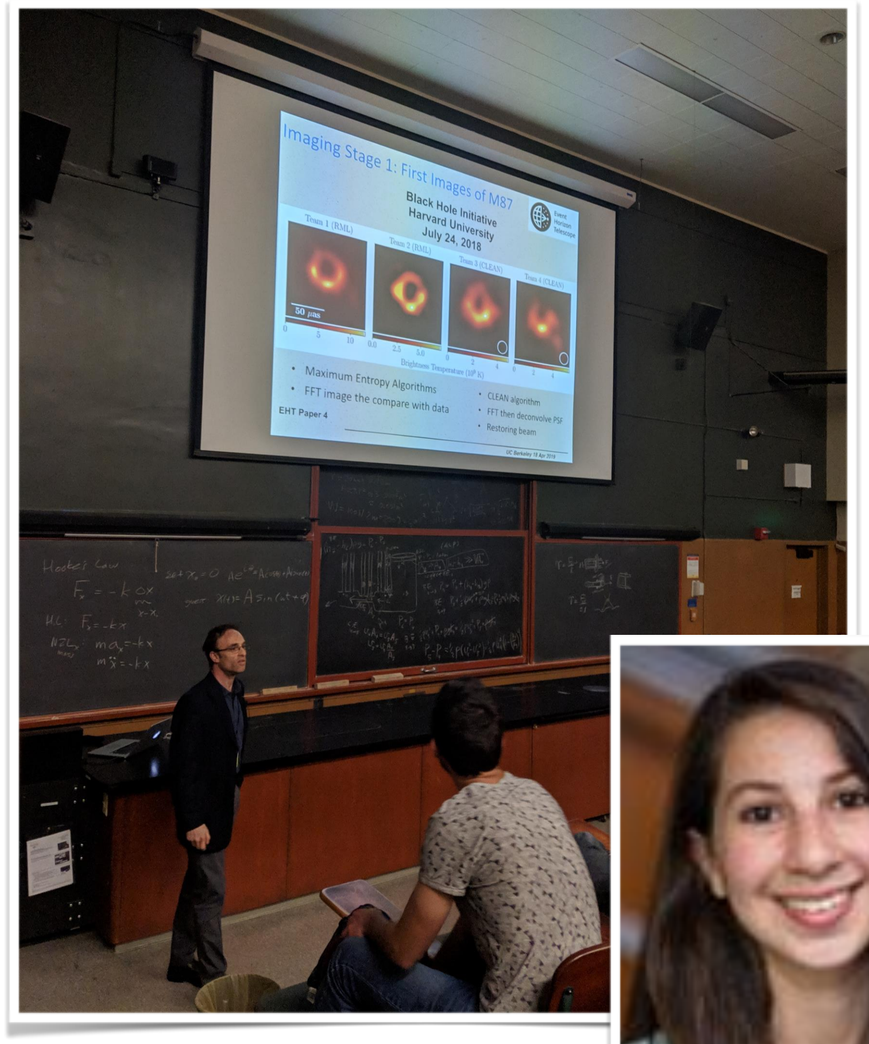
Ani Adhikari

John De Nero

Cathryn Carson

+ Data 100,
Prob 140,
Data 102,
…
+ a large
team!

April 18/19, 2019: Shep Doeleman & Katie Bouman



THE ASTROPHYSICAL JOURNAL LETTERS

First M87 Event Horizon Telescope Results. III. Data Processing and Calibration

The Event Horizon Telescope Collaboration, Kazunori Akiyama[1,2,3,4], Antxon Alberdi[5], Walter Alef[6], Keiichi Asada[7], Rebecca Azulay[8,9,6], Anne-Kathrin Baczko[6], David Ball[10], Mislav Baloković[4,11], John Barrett[2] +Show full author list

Published 2019 April 10 • © 2019. The American Astronomical Society.
The Astrophysical Journal Letters, Volume 875, Number 1

Software: DiFX (Deller et al. 2011), CALC, PolConvert (Martí-Vidal et al. 2016), HOPS (Whitney et al. 2004), CASA (McMullin et al. 2007), AIPS (Greisen 2003), ParselTongue (Kettenis et al. 2006), GNU Parallel (Tange 2011), GILDAS, eht-imaging (Chael et al. 2016, 2018), Numpy (van der Walt et al. 2011), Scipy (Jones et al. 2001), Pandas (McKinney 2010), Astropy (The Astropy Collaboration et al. 2013, 2018), Jupyter (Kluyver et al. 2016), Matplotlib (Hunter 2007).

Event Horizon Telescope

So you want to build Data Science tools in academia…

Jupyter - funding and resources

ALFRED P. SLOAN FOUNDATION

GORDON AND BETTY MOORE FOUNDATION

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

U.S. DEPARTMENT OF ENERGY

CHAN ZUCKERBERG INITIATIVE

SIMONS FOUNDATION

ANACONDA.

NETFLIX

POWERED BY rackspace® the open cloud company

IBM®

Microsoft

Google

Bloomberg

Should I Resign From My Full Professor Job To Work Fulltime On Cocalc?

William Stein • Apr 12, 2019 •

Nearly 3 years ago, I gave a talk at a Harvard mathematics conference announcing that "I am leaving academia to build a company". What I really did is go on *unpaid leave* for three years from my tenured Full Professor position. No further extensions of that leave is possible, so I finally have to decide whether or not to go back to academia or resign.

My unpaid leave is up – what am I going to do?

My third year of unpaid leave from UW is up. I have to decide whether to return to UW or resign. If I return, it turns out that I would have to have at least a 50% appointment. I currently have 50% of one year of teaching in "credits", which means I wouldn't be required to teach for the first year I go back as a 50% appointment. Moreover, the current department chair (John Palmieri) understands and appreciates Sage – he is among the top 10 all time contributors to the source code of Sage!

I have decided to resign. I'm worried about issues of intellectual property; it would be extremely unfair to my employees, investors and customers if I took a 50% UW

Contrasts in culture and incentives

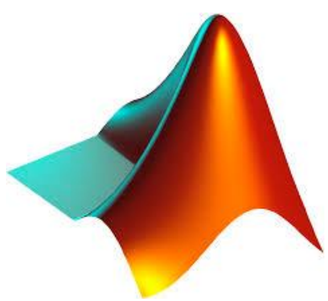
	Open Source	Academia
Credit	Distributed	PI & hierarchy
Output/artifacts	Continuous & Project-specific	Discrete papers
Collaborators	Fluid: professionals, volunteers, …	Structured, funding-dependent
Governance/decision making	Open, community based	Top-down, PI
Authorship	Fluid, roles can evolve, no clear “first/senior” author	Need to say more?
Peer review	Continuous, open, pervasive, friendly	The opposite
Value metric	Utility, need, impact	“Novel and transformative”

Catastrophic Success: an economic problem

3.5. *Python vs. IDL?*

A recent shift in astronomy has been the favored choice of interpreted programming language for day-to-day analysis work. In the previous section we showed that Python has overtaken IDL in popularity. This may not have been true three to five years ago, but today Python is, by a wide margin, the most popular interpreted language in astronomy (at least insofar as this survey is representative). Still, there is a significant over-

(2015 data) https://arxiv.org/abs/1507.03989



- **MathWorks:** 4,000+ employees
- **Wolfram:** 800 employees
- **IDL/Harris:** 17,000 employees

Thank you (Bay Area team)

Current (Berkeley, LBNL, Bloomberg)

Stacey Dorton, Lindsey Heagy, Chris Holdgraf, Yuvi Panda, Ryan Lovett, Shreyas Cholia, Shane Canon, Rollin Thomas, Jason Grout

Former Berkeley

Min Ragan-Kelley, Paul Ivanov, Thomas Kluyver, M Pacer, Matthias Bussonnier, Jessica Hamrick, Ian Rose, Jamie Whitacre.

University of California, Berkeley
DEPARTMENT OF STATISTICS

BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE

BERKELEY LAB

NERSC

Bloomberg

Scientific OSS at scale: complex challenges

- Economic incentives & sustainability
- Governance models
- Roles and professional career paths
- Multi-stakeholder organizations

Thank You!

No scientist is trained for any of this!!