

**IP[y]:**  
IPython

Berkeley  
Division of  
Data Sciences

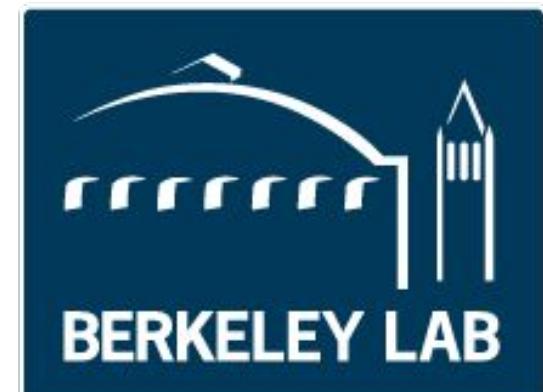


# Jupyter meets the Earth: an open, collaborative approach for Earth data science

Fernando Pérez  
Lindsey Heagy



University of California, Berkeley  
**DEPARTMENT OF STATISTICS**

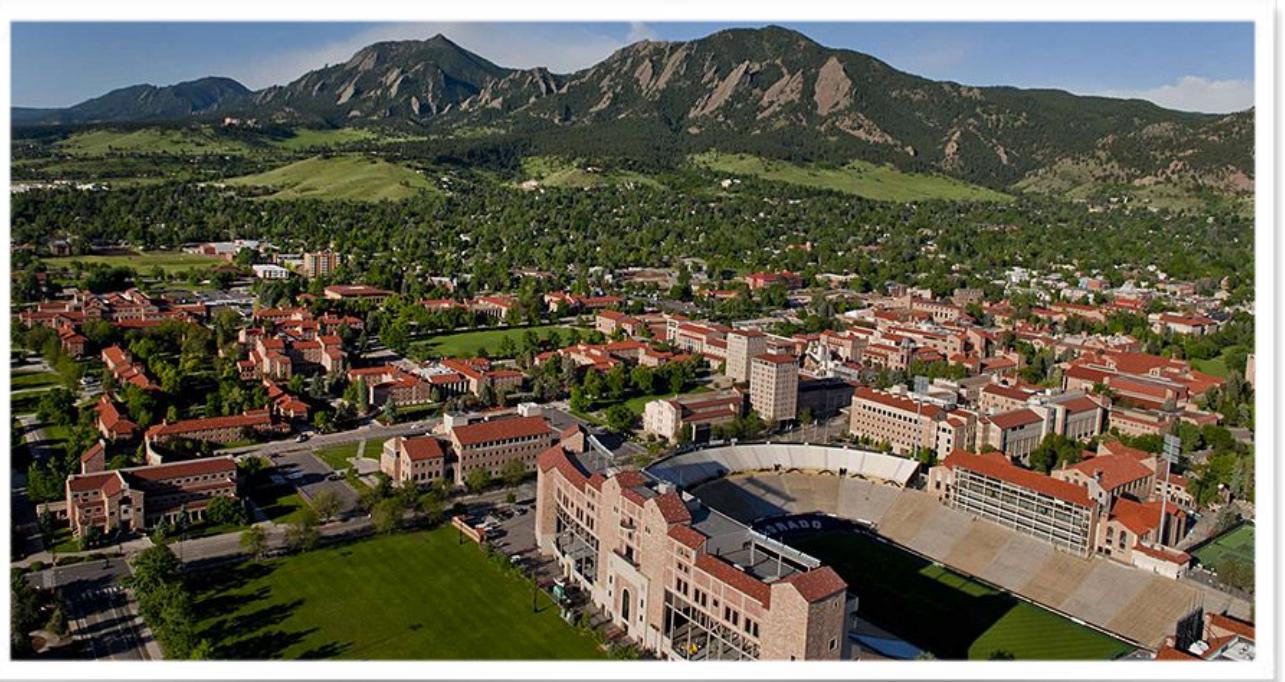




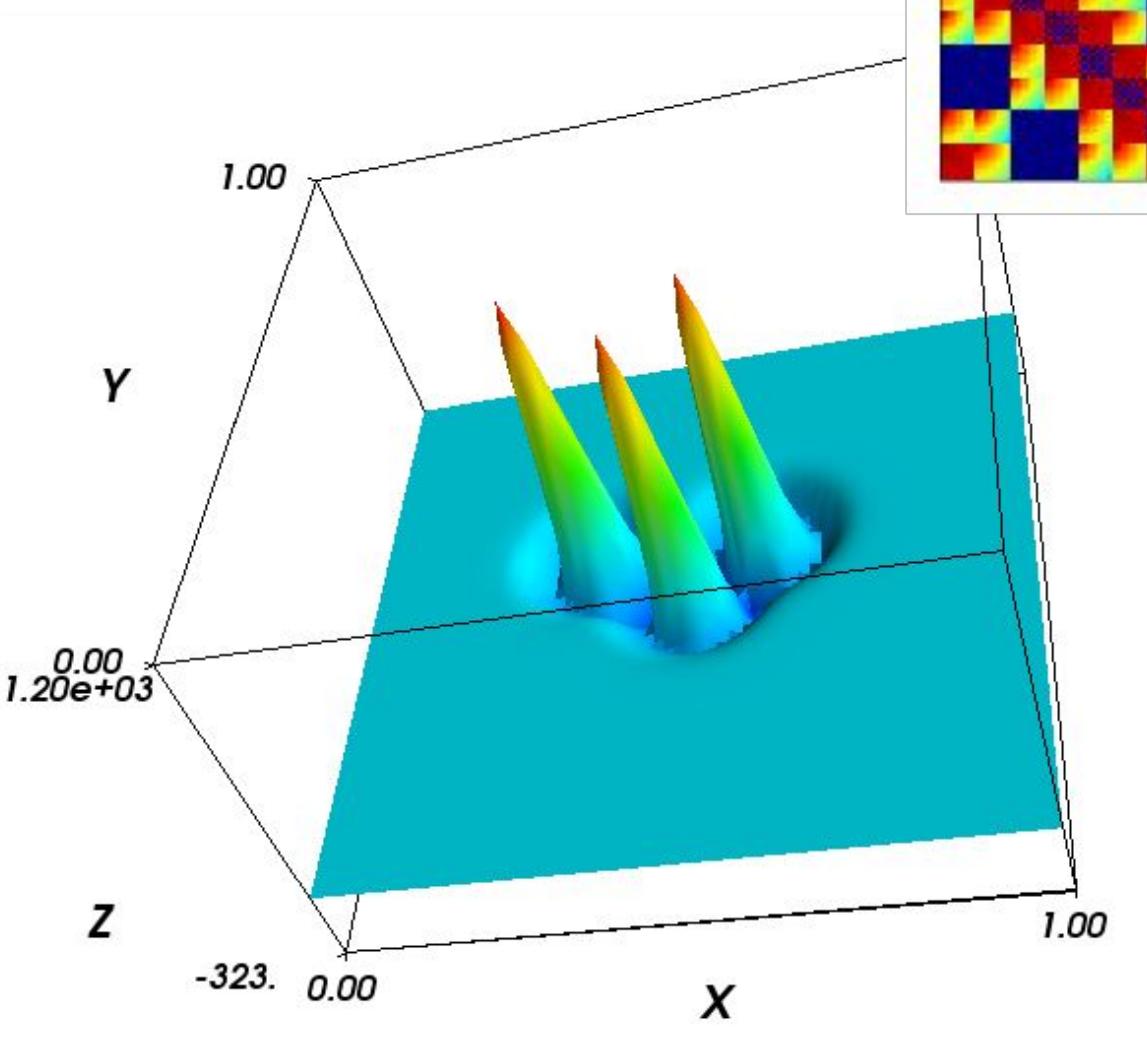
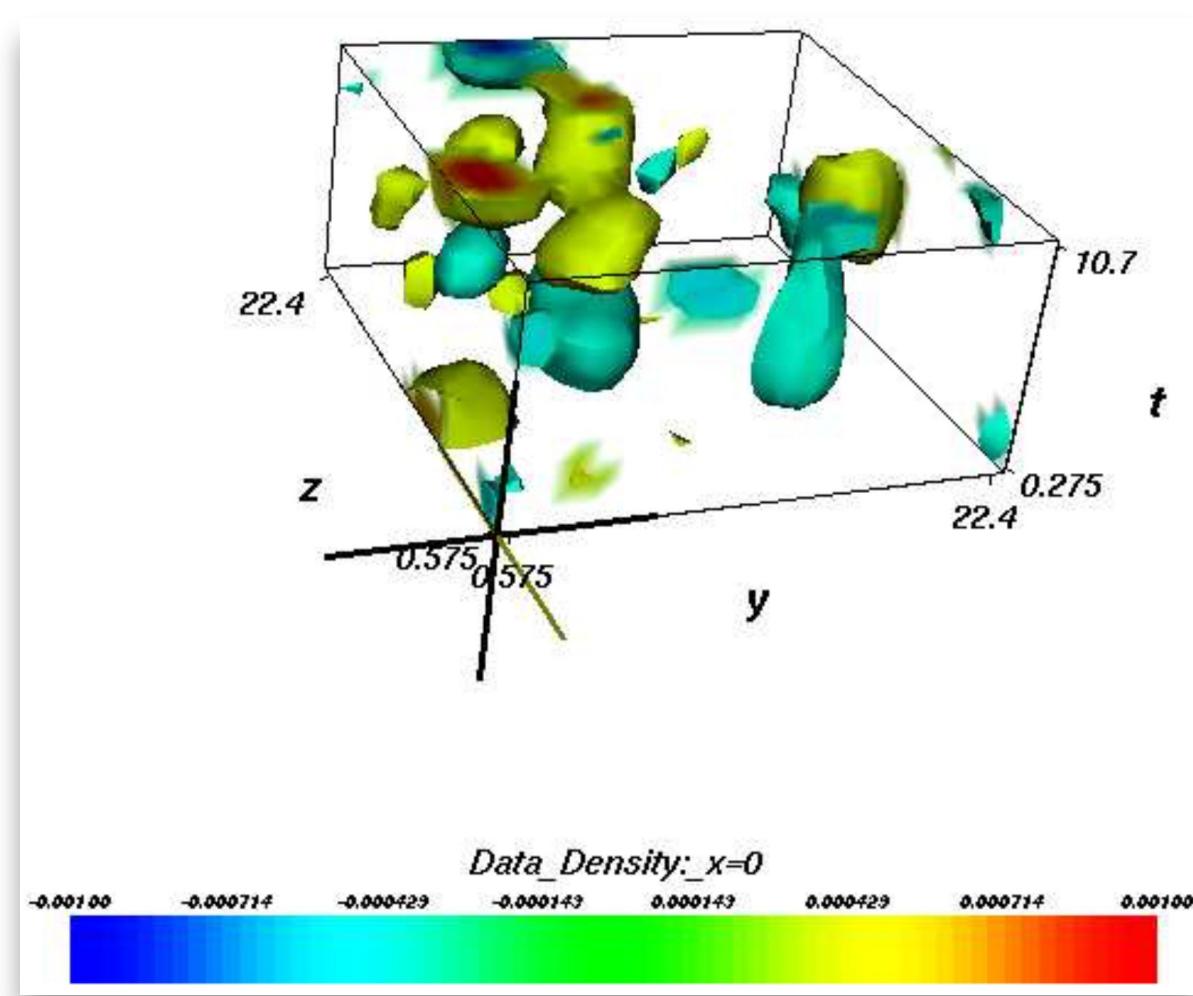
# ! a bit about me



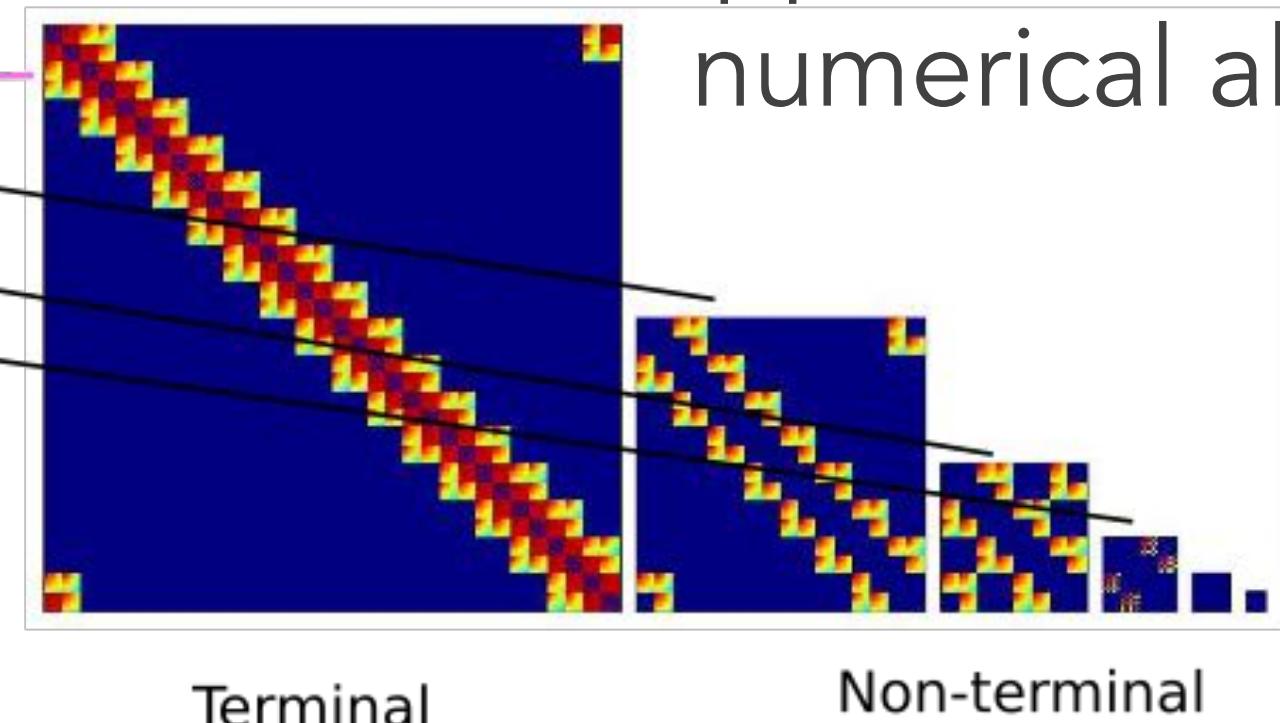
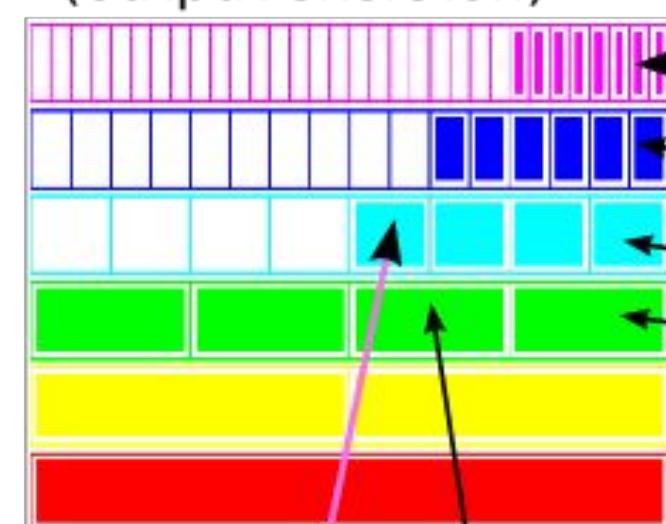
Boulder



Physics PhD: Lattice QCD  
Simulations



Redundant tree of input  
(output skeleton)



Applied Math Postdoc:  
numerical algorithms





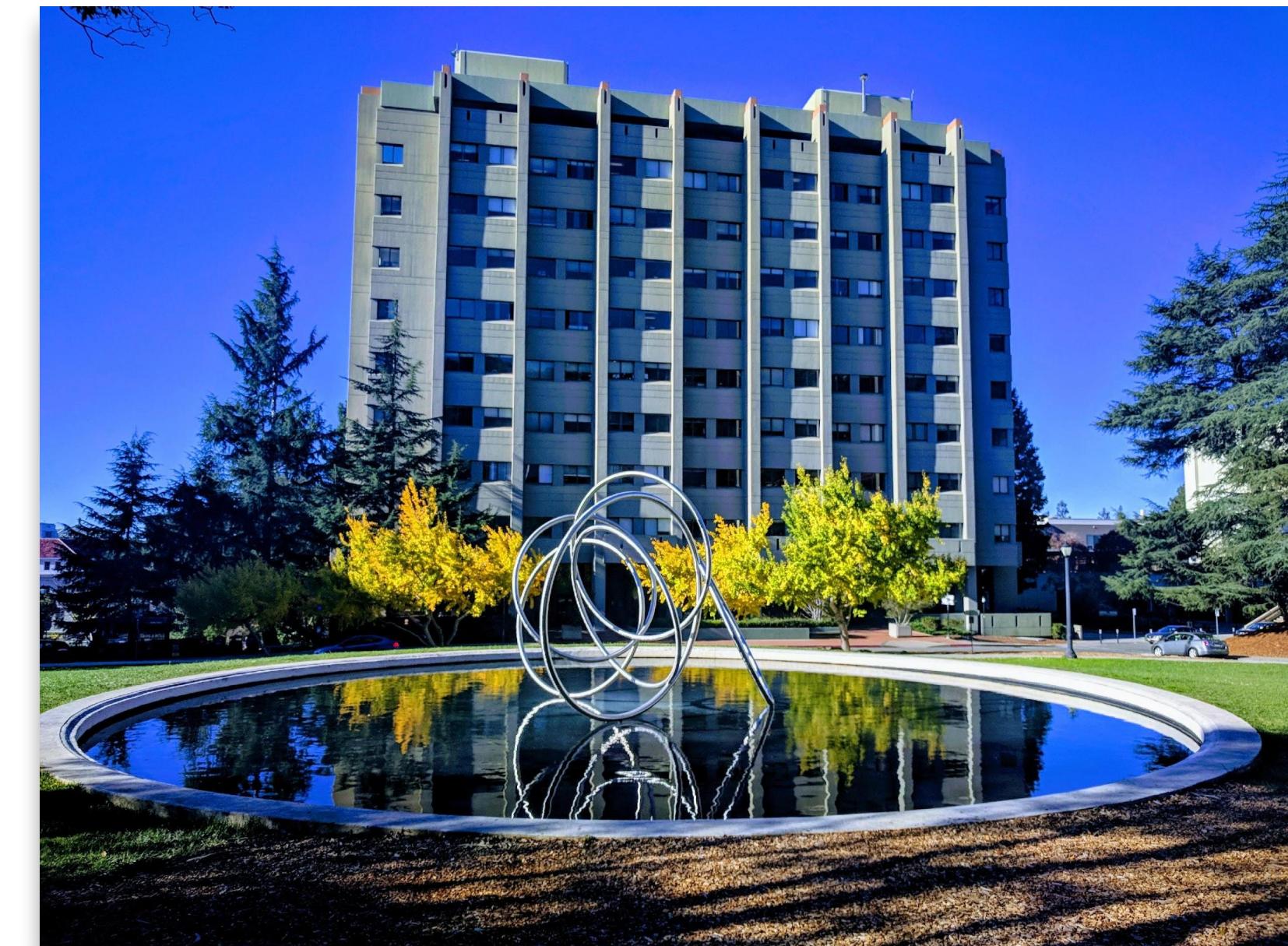
# BRAIN IMAGING CENTER

UNIVERSITY OF CALIFORNIA, BERKELEY

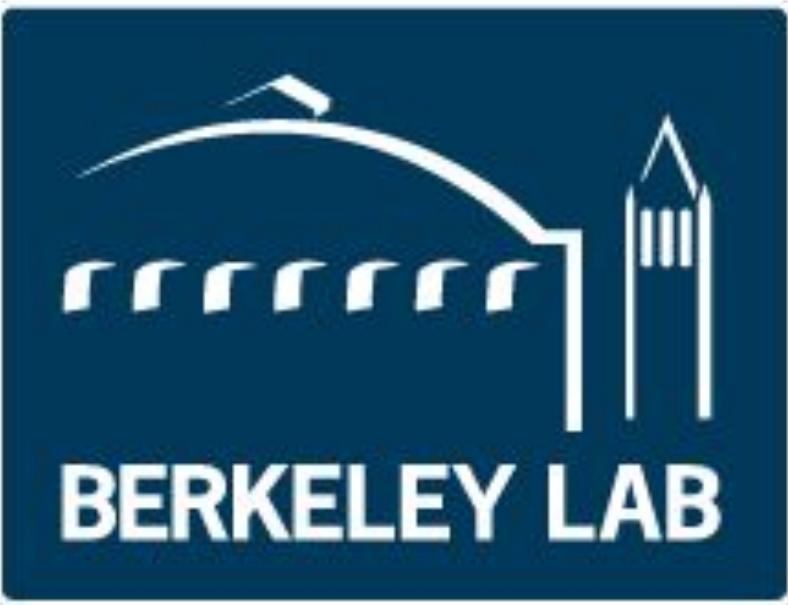
The Henry H. Wheeler, Jr. Brain Imaging Center (BIC) is one of four Technology Centers established under the auspices of the Helen Wills Neuroscience Institute. It is a campus-wide resource that supports advanced brain imaging technologies dedicated solely to basic brain research.



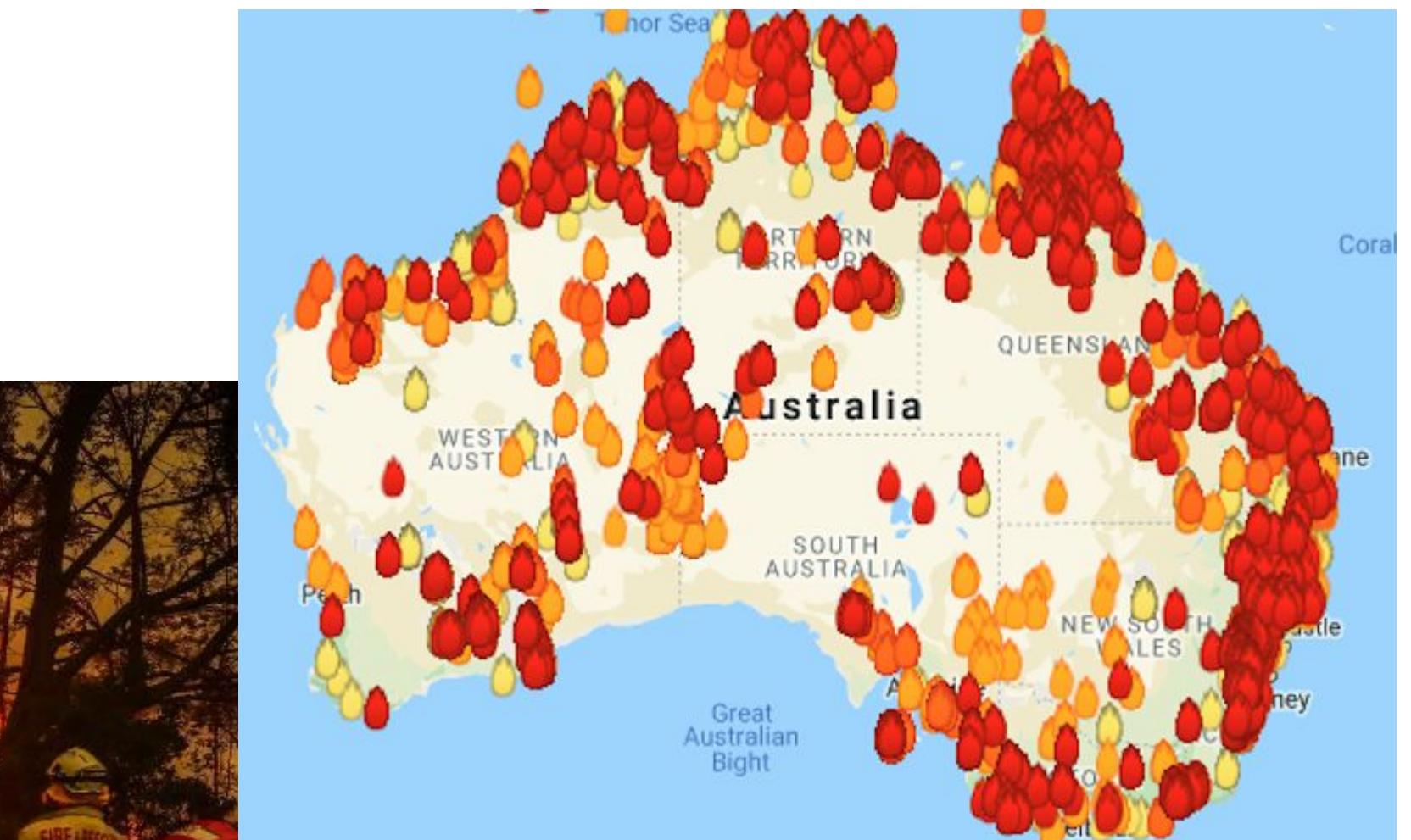
BERKELEY INSTITUTE  
FOR DATA SCIENCE



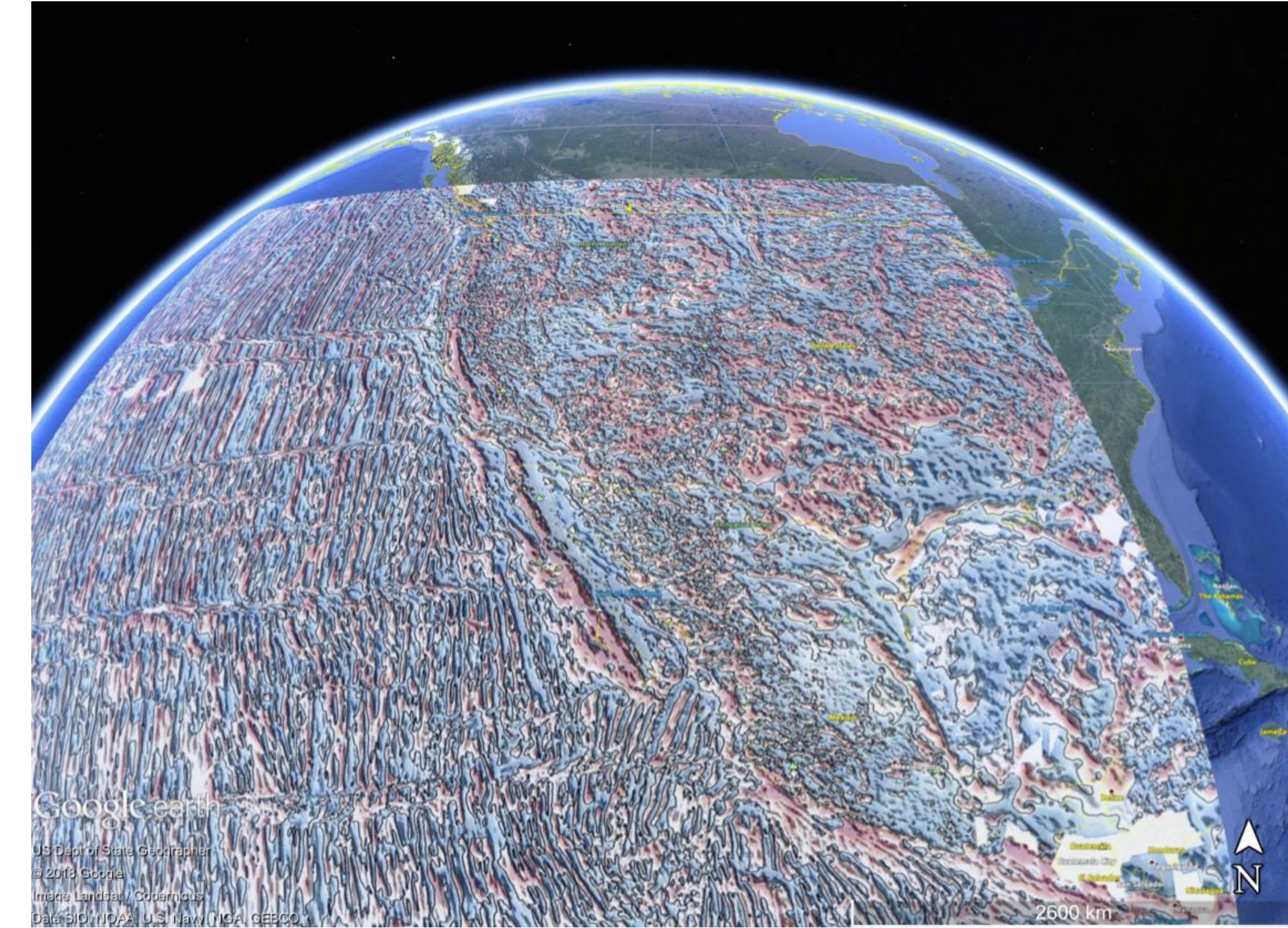
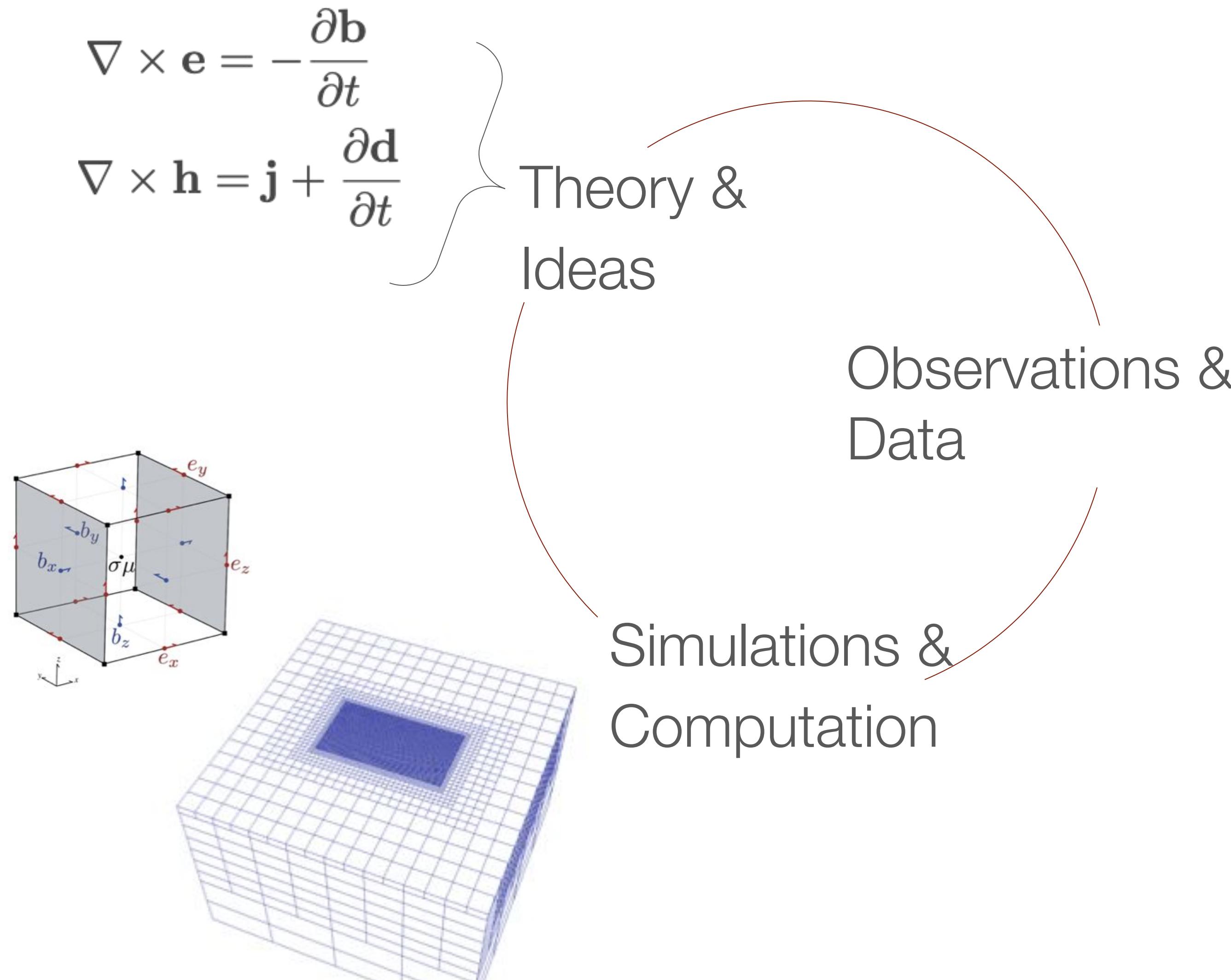
University of California, Berkeley  
**DEPARTMENT OF STATISTICS**



# Wait, why geoscience??



# what drives progress in (geo)science?



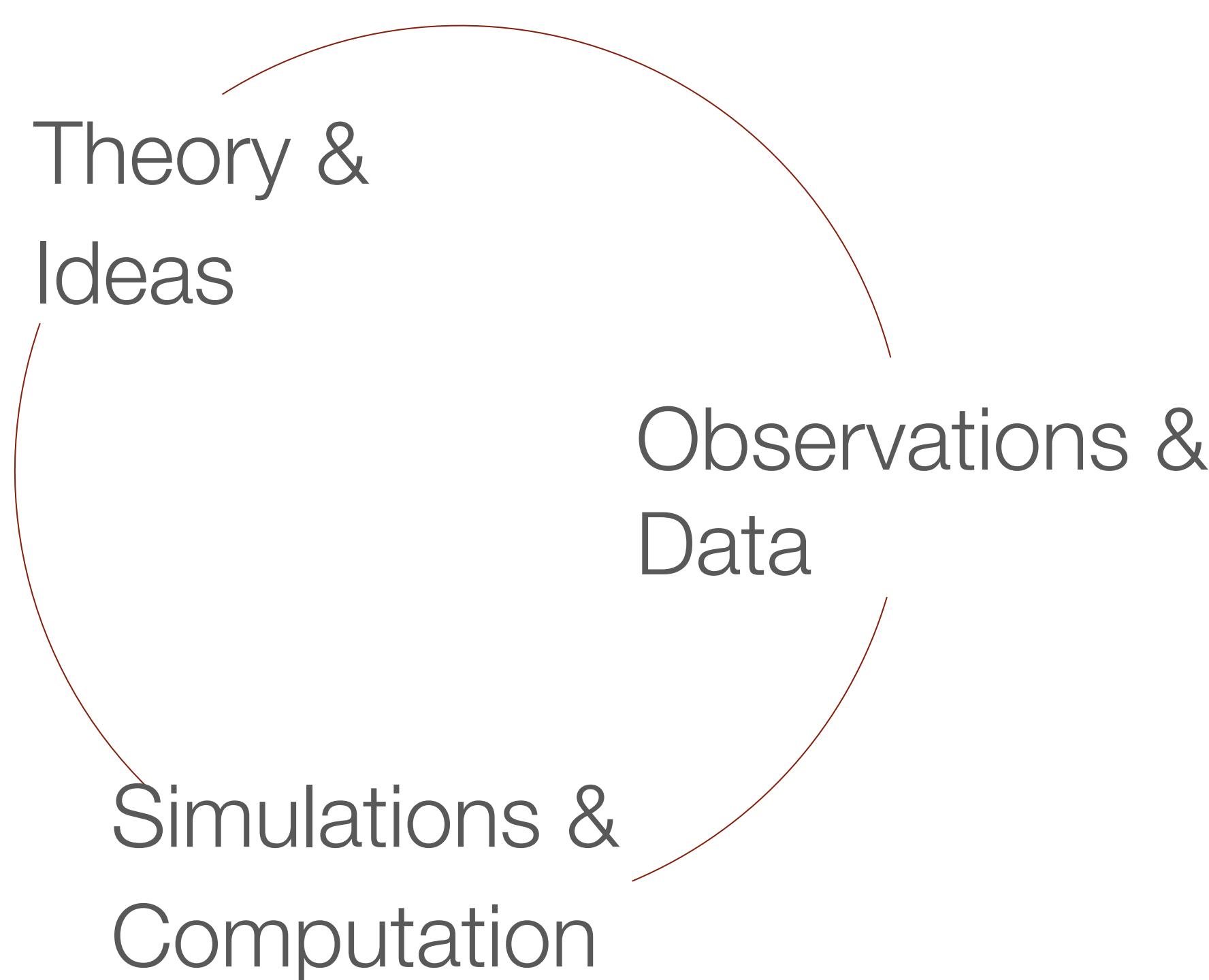
[EMAG2](#): Earth Magnetic Anomaly Grid (2-arc-minute resolution). Image credit: Dom Fournier ([toolkit.geosci.xyz](http://toolkit.geosci.xyz))

After [Hamman, 2018](#)

# "The gears of the engine are starting to grind"

Improved  
multiscale,  
nonlinear, noisy  
models

Exascale computing  
Cloud engineering  
skills  
Machine Learning  
pipelines



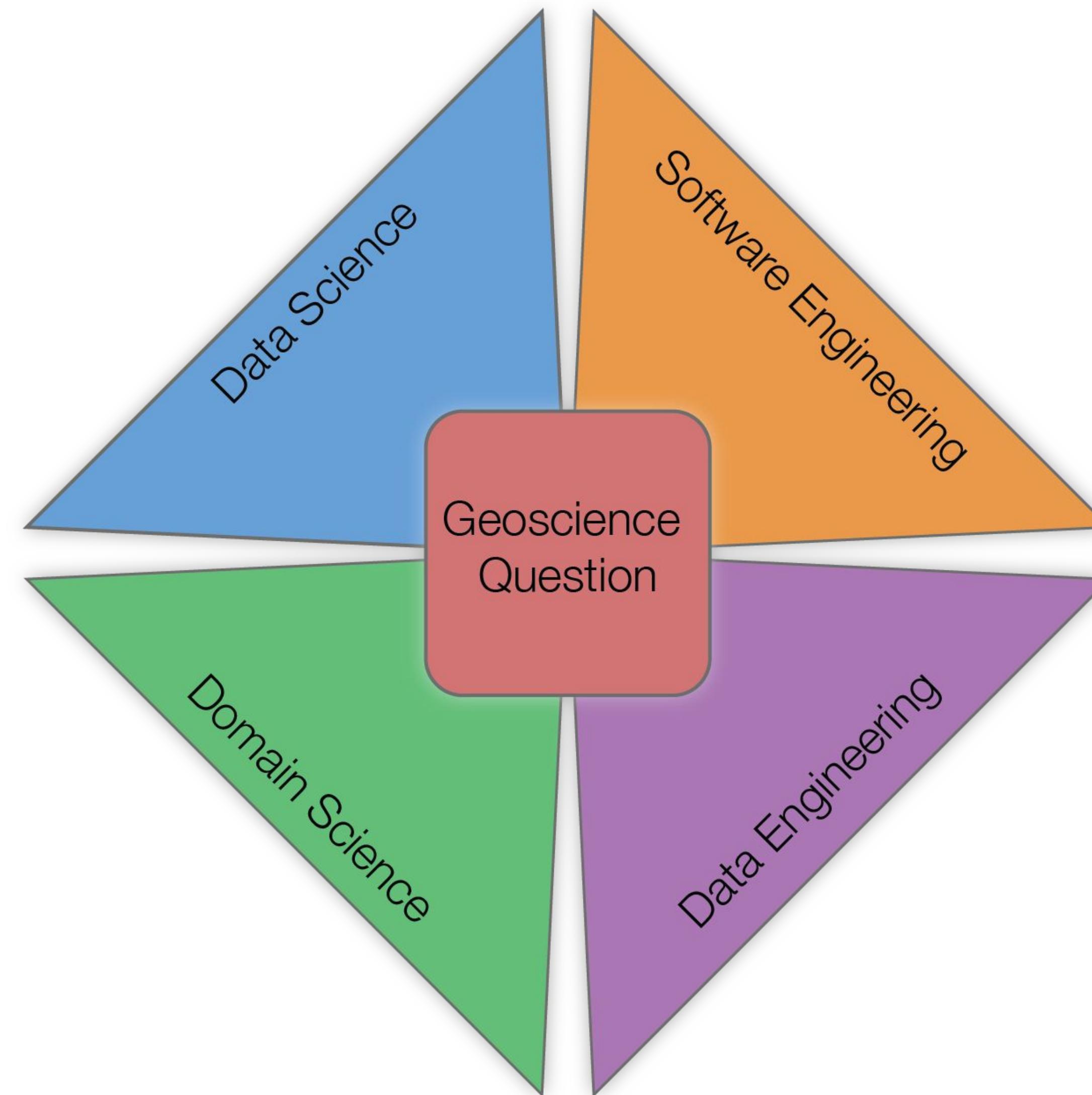
CMIP6 ~15-30 PB

Experimental Data  
rates of TB/day+

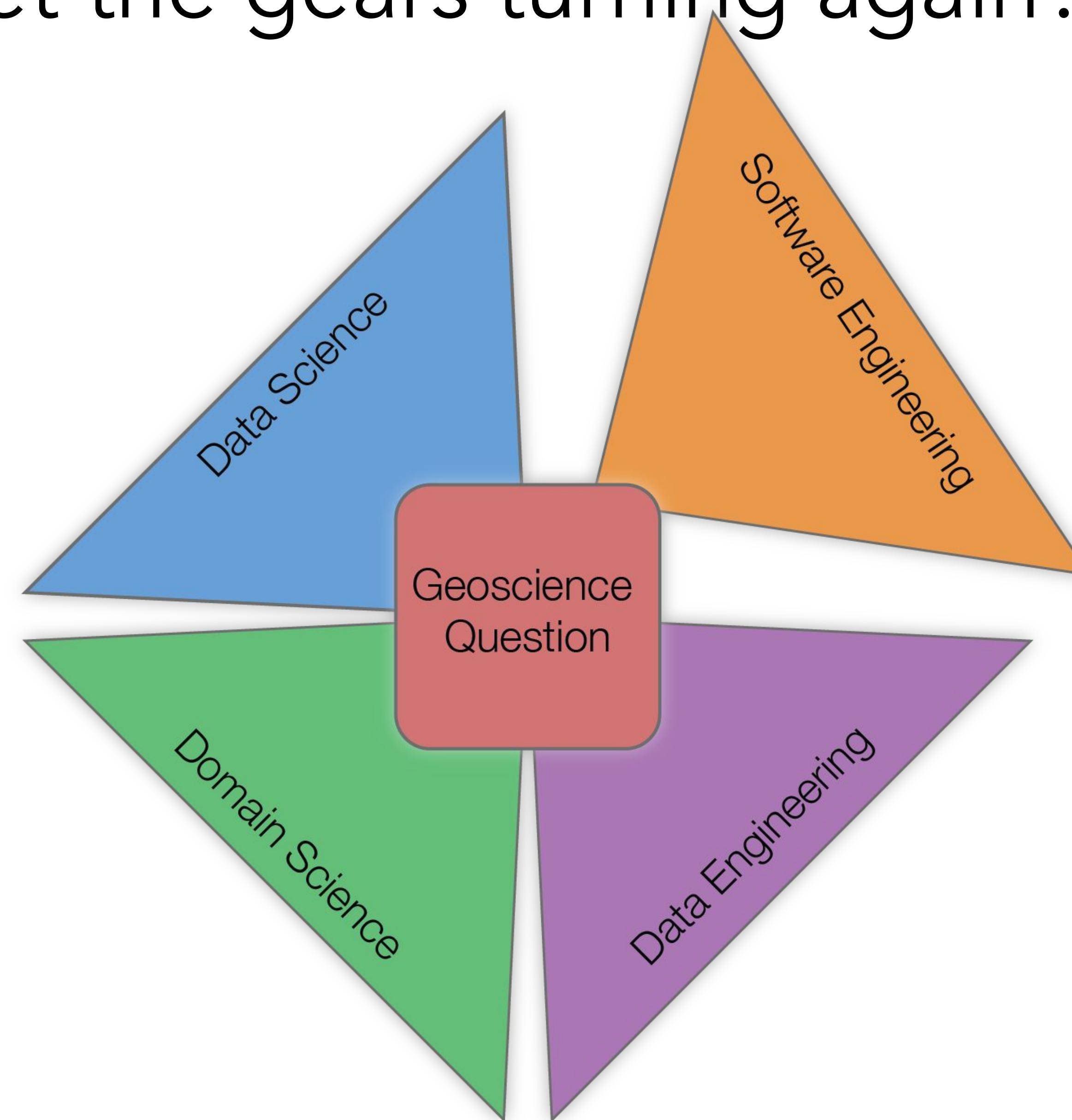
Heterogeneous,  
multimodal

After [Hamman, 2018](#)

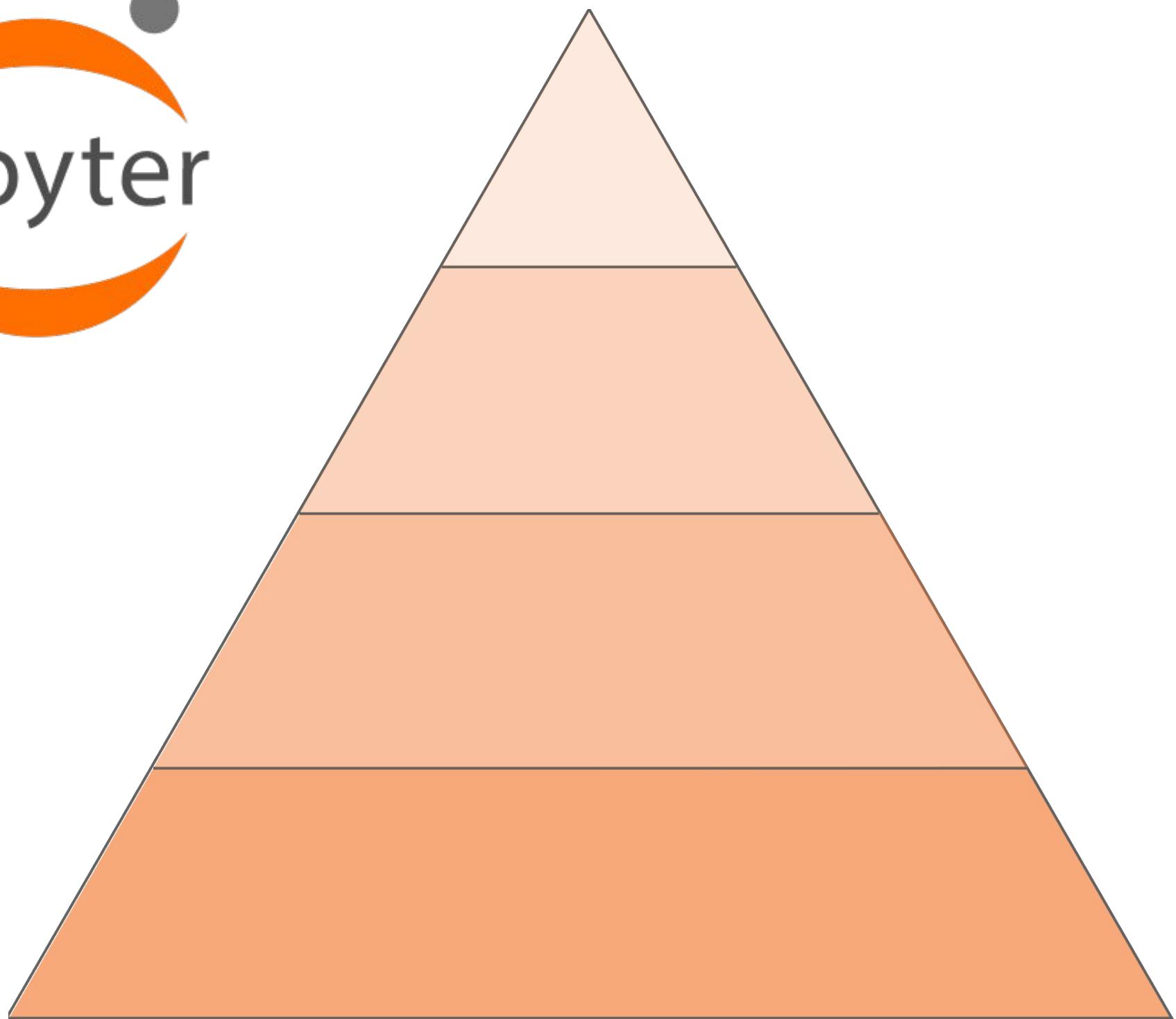
# how do we get the gears turning again?



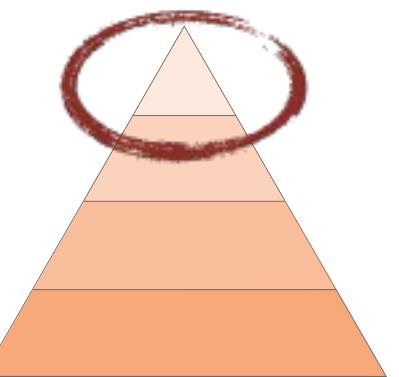
# how do we get the gears turning again?



# OSS: more than software



Services and content  
Software  
Standards and Protocols  
Community



# Content/Services

```

1. IPython: Users/fperez (python3.5)
(jlab) dreamweaver[~]> ipython
Python 3.5.2 |Continuum Analytics, Inc.| (default, Jul 2 2016, 17:52:12)
Type "copyright", "credits" or "license" for more information.

IPython 5.1.0 -- An enhanced Interactive Python.
?           -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help        -> Python's own help system.
object?    -> Details about 'object', use 'object??' for extra details.

In [1]: %pylab
Using matplotlib backend: MacOSX
Populating the interactive namespace from numpy and matplotlib

In [2]: from IPython.display import display
...: from pandas_datareader import data
...: from datetime import datetime
...:
...: ticker = 'MSFT'
...: stock = data.DataReader( ticker, 'yahoo', start=datetime(2012, 1, 1))
...: stock['Close'].plot(title='%s Closing Price' % ticker);
...:

      Open      High       Low     Close      Volume   Adj Close
Date
2012-01-03  26.549999  26.959999  26.389999  26.77  64731500  23.304317
2012-01-04  26.820000  27.469999  26.780001  27.40  80516100  23.852755
2012-01-05  27.379999  27.730000  27.290001  27.68  56081400  24.096507

In [3]:

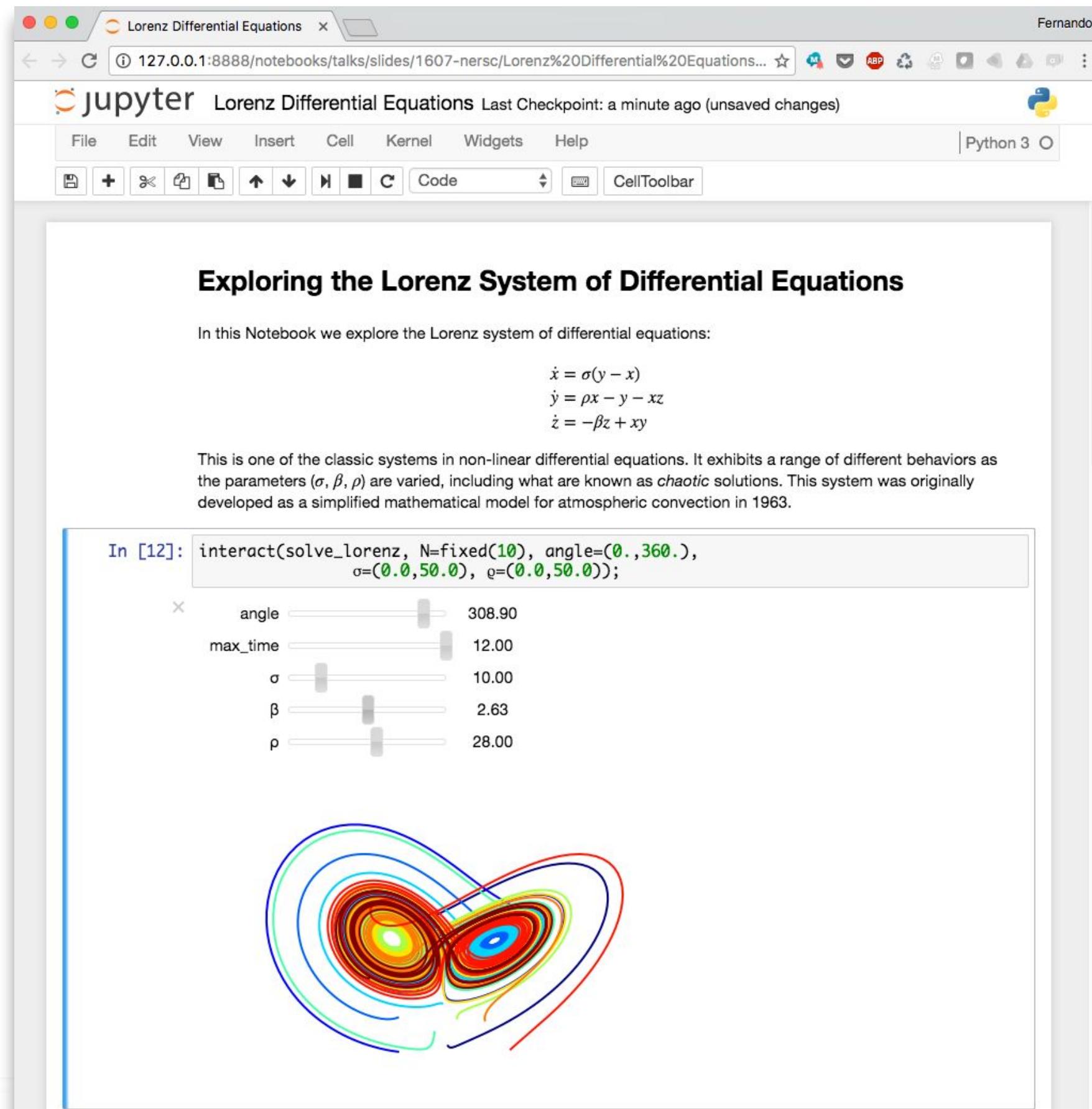
```

Figure 1  
MSFT Closing Price



## Jupyter for Organizations

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



**jupyter nbviewer**

A simple way to share Jupyter Notebooks

Enter the location of a Jupyter Notebook to have it rendered here:

URL | GitHub username | GitHub username/repo | Gist ID Go!

**binder** (beta)

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

GitHub repository name or URL

Git branch, tag, or commit

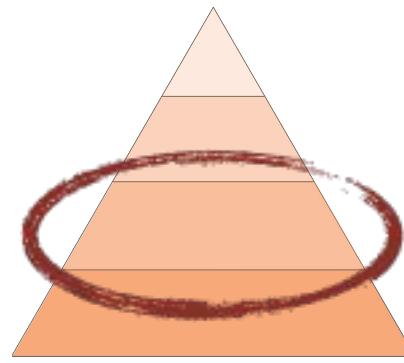
Path to a notebook file (optional)

**launch**

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder.

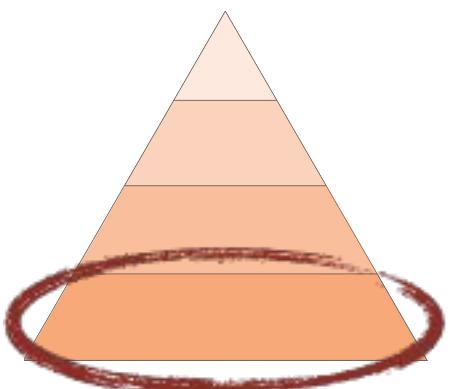
Copy the text below, then paste into your README to show a binder badge:



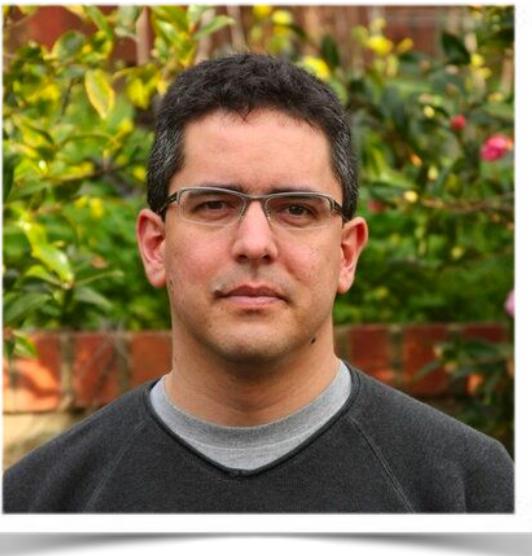
# A language agnostic protocol



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



# Community: formalized governance



Brian Granger

Me :)



Steering Council

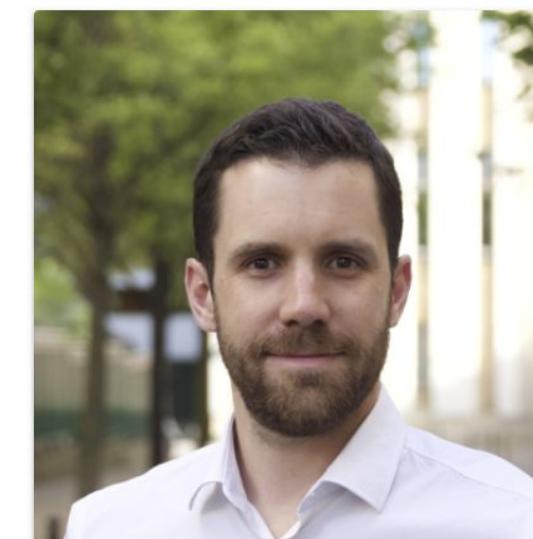
The role of the Jupyter Steering Council is to ensure, through working with and serving the broader Jupyter community, the long-term well-being of the project, both technically and as a community. The Jupyter Steering Council currently consists of the following members (in alphabetical order).



Damian Avila  
Anaconda, Inc.  
[@damianavila](https://github.com/damianavila) on GitHub



Matthias Bussonnier  
UC Merced  
[@Carreau](https://github.com/Carreau) on GitHub

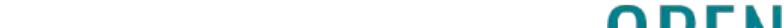
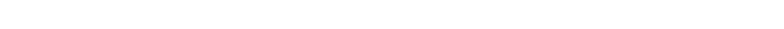
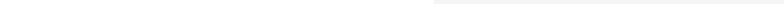
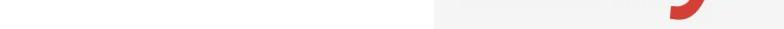
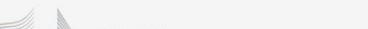
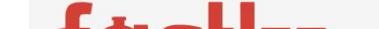
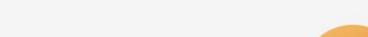
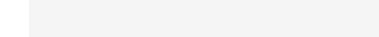


Sylvain Corlay  
QuantStack  
[@sylvaincorlay](https://github.com/sylvaincorlay) on GitHub

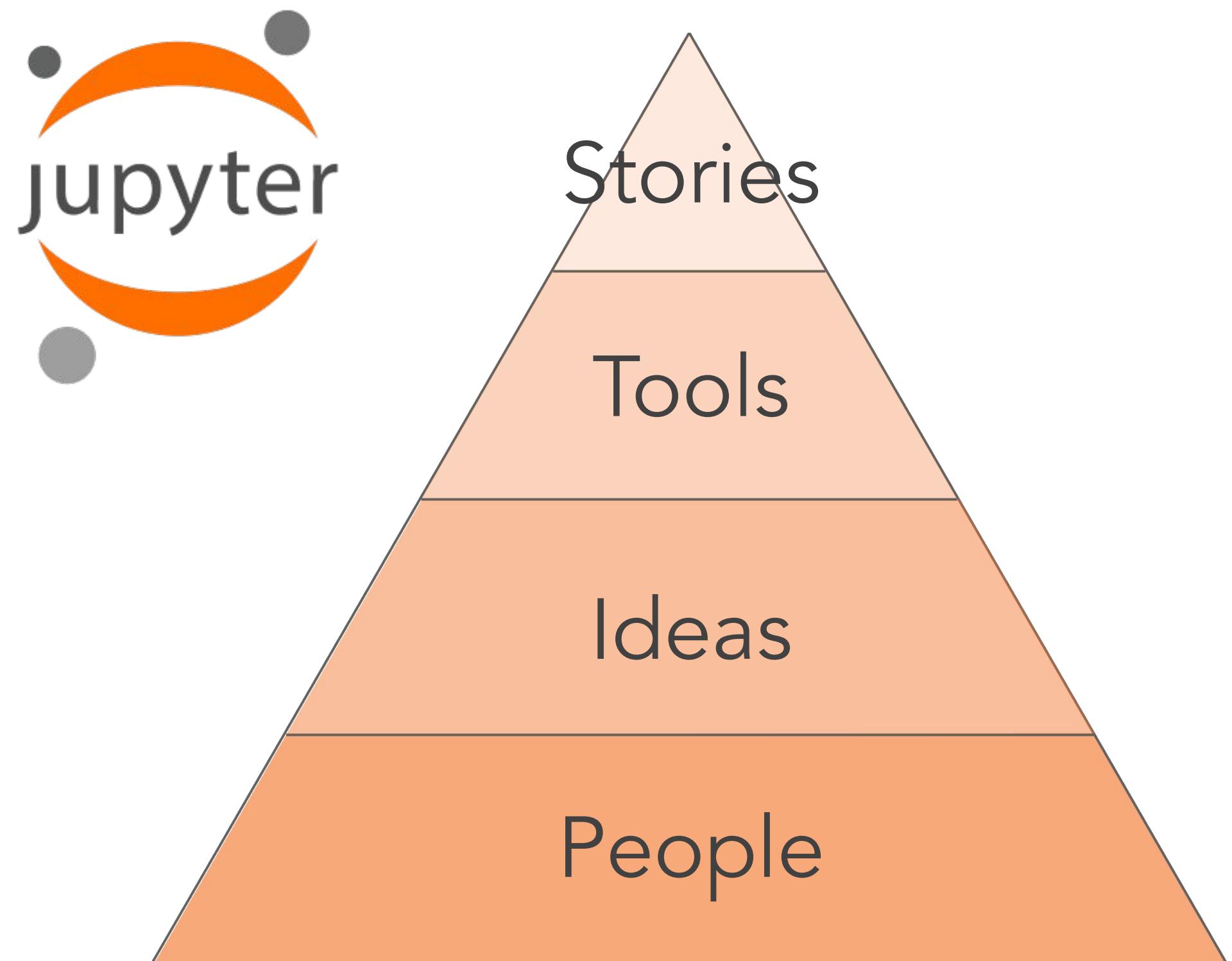


## Institutional Partners

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



# 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

\* With a nod to the FAIR principles of open data

Collaborative?

# IPython: an afternoon hack, 2001

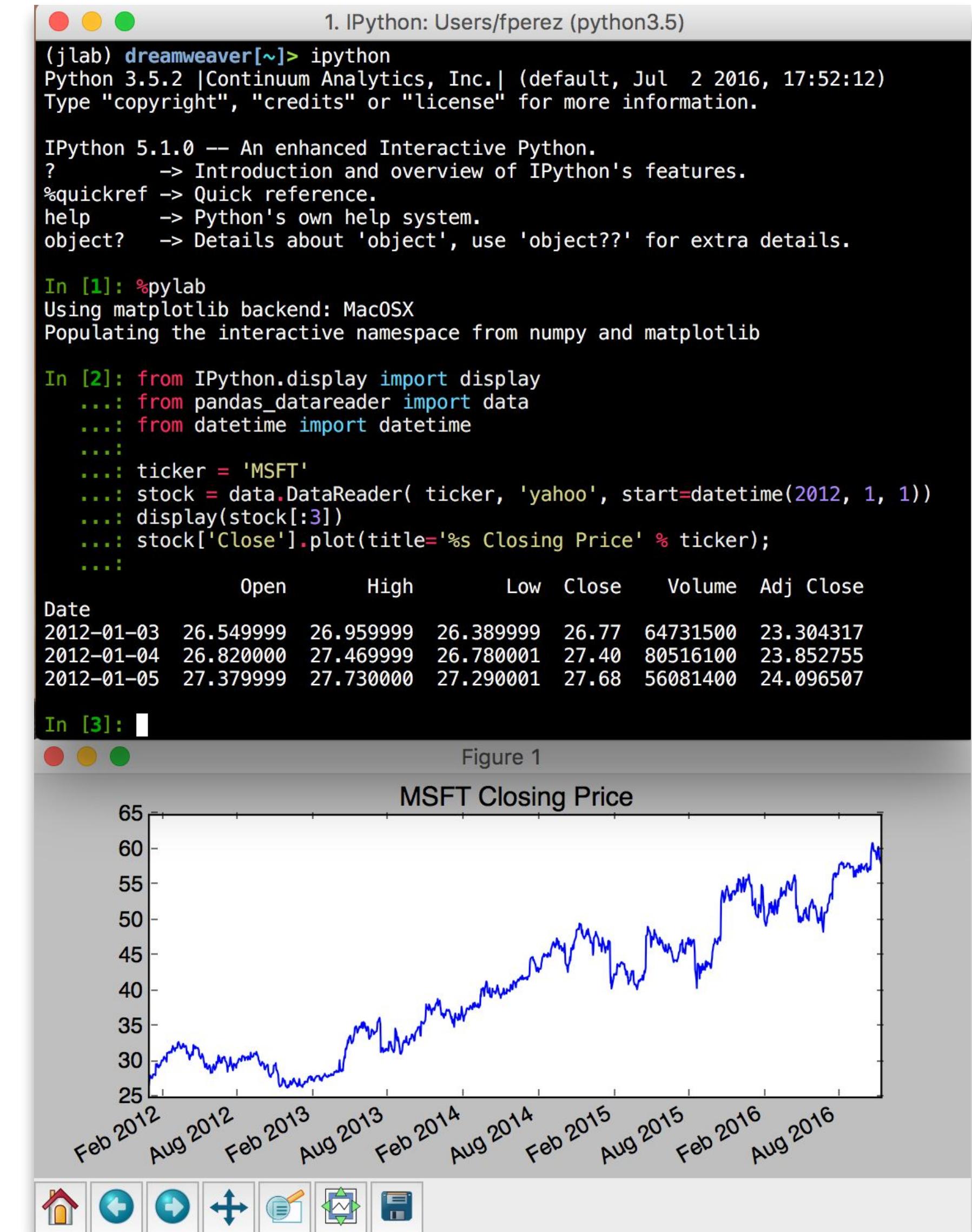


Boulder



```
ipython-0.0.1.py
```

```
32     Globals for SI units (including g=9.8)      : _load_units    = %(_load_units)s
33     Starting number for prompt counter        : _prompt_ini   = %(_prompt_ini)s
34     Number of history items to store in cache : _cache_size   = %(_cache_size)s
35
36     ****
37     # Configure here
38     _load_Numeric   = 1
39     _load_Gnuplot   = 1
40     _load_gracePlot = 1
41     _load_units     = 1
42     _cache_size     = 1000
43     _prompt_ini     = 1
44
45     # *** Don't modify below unless you know what you're doing. ***
46
47     # Crude first version, with minimal object structure. This could be done much
48     # better, by defining a Cache class (probably using weak references or
49     # generators). But it seems to work ok. Haven't checked for memory circularity
50     # problems, though.
51
52     ****
53     # Copyright (C) 2001 Fernando Pérez. <fperez@pizero.colorado.edu>
54
55     # Distributed under the terms of the GNU General Public License.
56
57     # The full text of the GPL is available at:
58
59     #         http://www.gnu.org/copyleft/gpl.html
60
61     __author__ = 'Fernando Pérez. <fperez@pizero.colorado.edu>'
62     __version__ = '0.1'
63
64     ****
65     definitions
66
67     prompt1:
68     e interactive prompt like Mathematica's."""
69     r__(self):
70     rn '\nIn[' +`_prompt_count`+']:= '
71
72     prompt2:
73     e interactive continuation prompt."""
74     r__(self):
75     rn '... '+ `*(len('In[' +`_prompt_count`+']:= ') - 3)`
76
77     ****
78     definitions
79
80     print(arg):
81     ing with history cache management.
82
83     invoked everytime the interpreter needs to print, and is activated
84     by setting the variable sys.displayhook to it."""
85
86     global _p,_pp,_ppp,_cache,_prompt_count
```



# Multiple stakeholders, team effort

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



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!

# the science more than the paper

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 Claerbout)  
WaveLab and Reproducible Research, 1995

(and a place to run the code?)



# binder

shareable, interactive, reproducible  
environments from your public git repository

Binder (beta) https://mybinder.org

## binder (beta)

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

GitHub repository name or URL

GitHub repository name or URL GitHub ▾

Git branch, tag, or commit Path to a notebook file (optional)

Git branch, tag, or commit Path to a notebook file (optional) File ▾ launch

Copy the URL below and share your Binder with others:

Fill in the fields to see a URL for sharing your Binder. ⌂

Copy the text below, then paste into your README to show a binder badge: [Launch binder](#) ▶

### How it works

1 Enter your repository information

Provide in the above form a URL or a GitHub repository that contains Jupyter notebooks, as well as a branch, tag, or commit hash. Launch will build your Binder repository. If you specify a path to a notebook file, the notebook will be opened in your browser after building.

[mybinder.org](https://mybinder.org)

# Black holes! LIGO, Sept 14, 2015

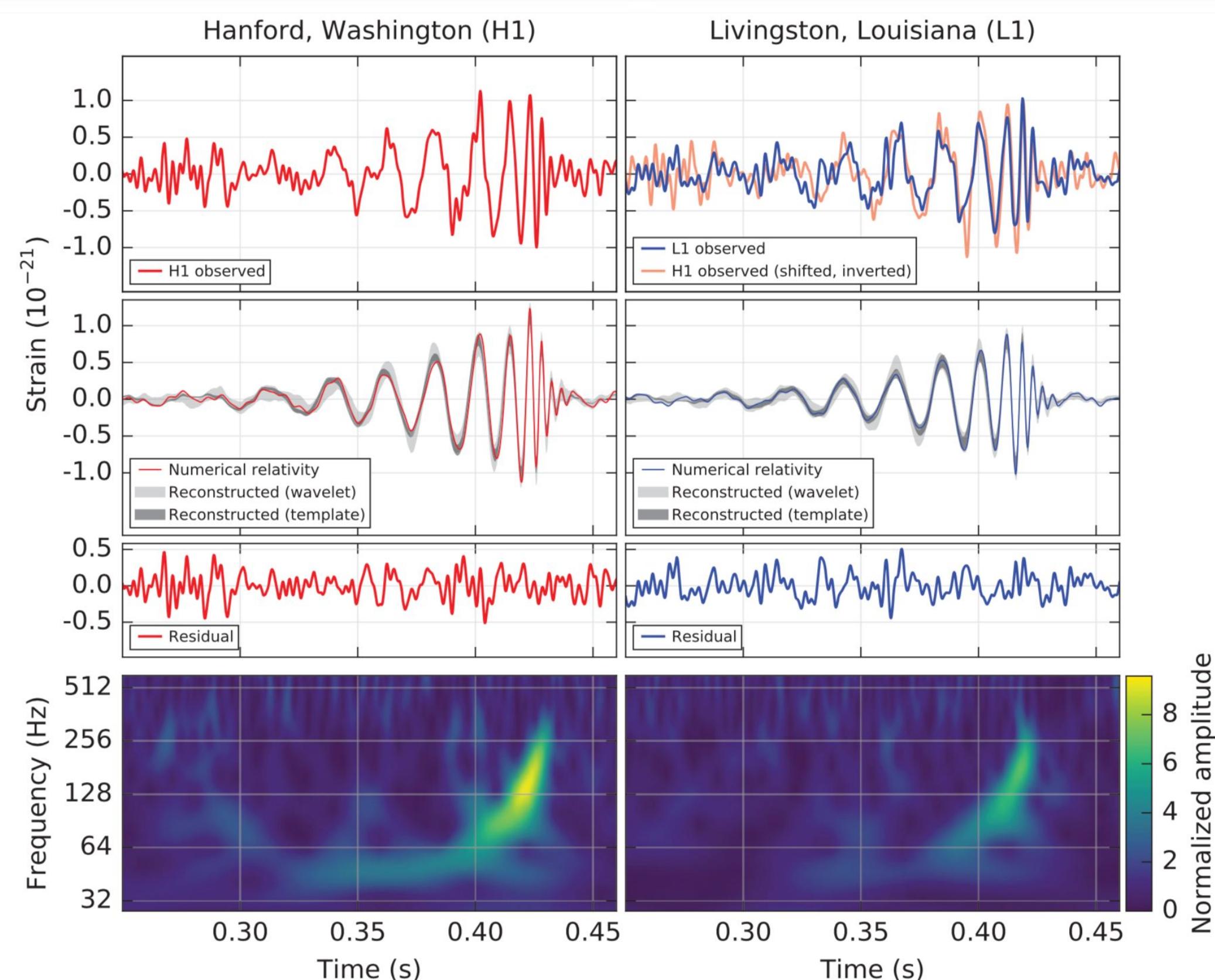
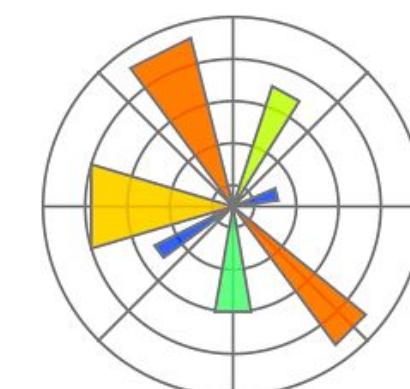
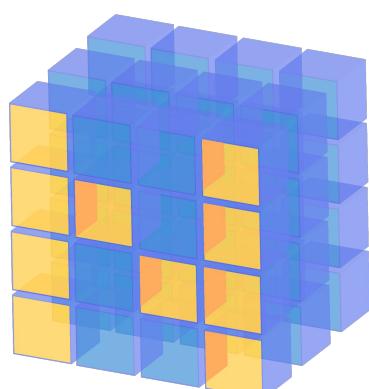


FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress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



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

Audio

# Extensible?

# JupyterLab: a grand unified theory of Jupyter

The screenshot shows the Pre-Alpha Jupyter Lab interface. It includes a terminal window displaying system monitoring data (CPU usage, memory usage, and a process list), a code editor with Python code for generating a polar plot, and a notebook cell showing the resulting polar plot. The interface is designed to be highly integrated and modern.

The screenshot shows the Pre-Alpha Jupyter Lab interface. It includes a terminal window displaying MRI analysis code, a code editor with Python code for generating histograms, and a notebook cell showing an EEG visualization. The interface is designed to be highly integrated and modern.

Huge Team Effort!

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

# Notebooks++

File Edit View Run Kernel Tabs Settings Help

Lorenz.ipynb ×

Files Running Commands Cell Tools Tabs

## The Lorenz Differential Equations

Before we start, we import some preliminary libraries. We will also import (below) the accompanying lorenz.py file, which contains the actual solver and plotting routine.

In [1]:

```
%matplotlib inline
from ipywidgets import interactive, fixed
```

We explore the Lorenz system of differential equations:

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

Let's change  $(\sigma, \beta, \rho)$  with ipywidgets and examine the trajectories.

In [2]:

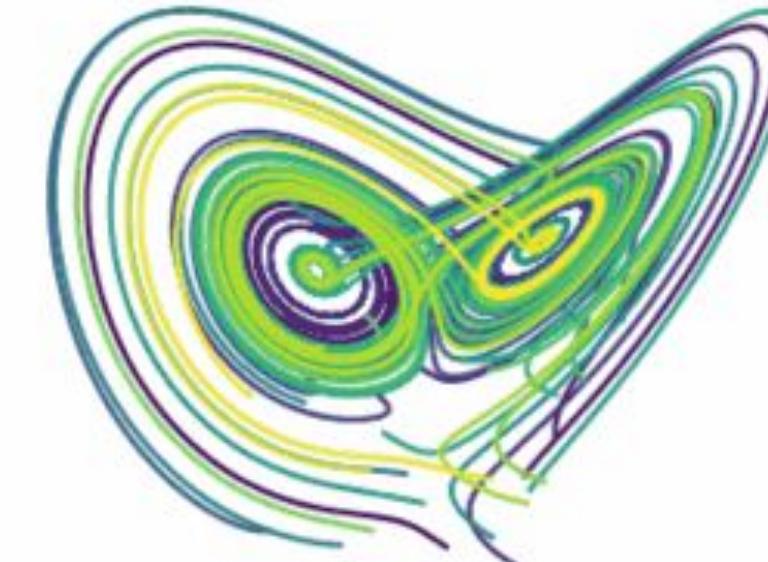
```
from lorenz import solve_lorenz
w=interactive(solve_lorenz,sigma=(0.0,50.0),rho=(0.0,50.0))
w
```

sigma 10.10  
beta 2.63  
rho 23.30



Output View ×

sigma 10.10  
beta 2.63  
rho 23.30



Lorenz.pdf ×

## 1 The Lorenz Differential Equations

Before we start, we import some preliminary libraries. We will also import (below) the accompanying lorenz.py file, which contains the actual solver and plotting routine.

In [1]:

```
%matplotlib inline
from ipywidgets import interactive, fixed
```

We explore the Lorenz system of differential equations:

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

Let's change  $(\sigma, \beta, \rho)$  with ipywidgets and examine the trajectories.

In [2]:

```
from lorenz import solve_lorenz
w=interactive(solve_lorenz,sigma=(0.0,50.0),rho=(0.0,50.0))
w
```

```
interactive(children=(FloatSlider(value=10.0, description='sigma', max=50.0), FloatSlider(valu
```

# Beyond notebooks

The screenshot shows the JupyterLab interface with two main panes. The left pane is a file editor containing a Markdown file named `markdown_python.m`. The right pane is a markdown viewer showing the rendered content of the same file.

**File Editor Content (markdown\_python.m):**

```
1 # Markdown
2
3 This is a regular markdown file. In JupyterLab you can open and
4 edit the file in the file editor, or in the markdown viewer. As you
5 edit the file the rendered markdown will automatically update.
6
7 $$ \alpha + \sum_{i=0}^n y_i $$
8
9 # Including Python code
```

**File Tree:**

- Launcher
- markdown\_python.m

**Markdown Viewer Content:**

## Markdown

This is a regular markdown file. In JupyterLab you can open and edit the file in the file editor, or in the markdown viewer. As you edit the file the rendered markdown will automatically update.

$$\alpha + \sum_{i=0}^n y_i$$

## Including Python code

Here is a block of Python code in the markdown file:

```
a = 100
```

Let's attach a Python 3 Kernel and Console to this markdown file. Then we can select lines of code in the markdown file and run them in the console by pressing `Shift+Enter`. Let's do something more complicated:

First import `matplotlib`, `numpy` and `pandas`, and create a data frame:

```
%matplotlib inline
from matplotlib import pyplot as plt
from matplotlib import style
import numpy as np
import pandas as pd
data = {
    'x': np.random.rand(100),
    'y': np.random.rand(100),
    'color': np.random.rand(100),
    'size': 100.0*np.random.rand(100)
}
df = pd.DataFrame(data)
df.head()
```

**Console Output:**

```
[1]:
```

	x	y	color	size
0	0.617536	0.508976	0.945603	8.052391
1	0.051126	0.322041	0.237476	49.005570
2	0.247221	0.476783	0.875143	81.492632
3	0.713989	0.480705	0.462035	7.631572
4	0.312106	0.888199	0.640439	53.201591

```
[ ]:
```

0 1 ipythongfm Ln 7, Col 4 Spaces: 4 markdown\_python.md

# Data

File Edit View Run Kernel Tabs Settings Help

Launcher 1024px-Hubble

Museums\_in\_D

iris.csv

Delimiter: ,

	sepal_length	sepal_width	petal_length	petal_width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa

1 {"type": "FeatureCollection", "features": [{"type": "Feature", "properties": {"OBJECTID": 1, "ADDRESS": "716 MONROE STREET NE", "NAME": "AMERICAN POETRY MUSEUM", "ADDRESS\_ID": 309744, "LEGALNAME": "HERITAGE US", "ALTNAMES": "AMERICAN POETRY MUSEUM", "WEBURL": "http://americanpoetrymuseum.org/"}, "geometry": {"type": "Point", "coordinates": [-76.995003703568, 38.9328428790235]}}, {"type": "Feature", "properties": {"OBJECTID": 2, "ADDRESS": "719 6TH STREET NW", "NAME": "GERMAN-AMERICAN HERITAGE MUSEUM", "ADDRESS\_ID": 238949, "LEGALNAME": "CORCORAN GALLERY OF ART", "ALTNAMES": " ", "WEBURL": "http://gahmusa.org/"}, "geometry": {"type": "Point", "coordinates": [-77.01958878310639, 38.89911061096782]}}, {"type": "Feature", "properties": {"OBJECTID": 3, "ADDRESS": "1307 NEW HAMPSHIRE AVENUE NW", "NAME": "HEURICH HOUSE FOUNDATION", "ADDRESS\_ID": 241060, "LEGALNAME": "U.S. DEPARTMENT OF THE"}]

Museums\_in\_C

Leaflet | Map data (c) OpenStreetMap contributors

# 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

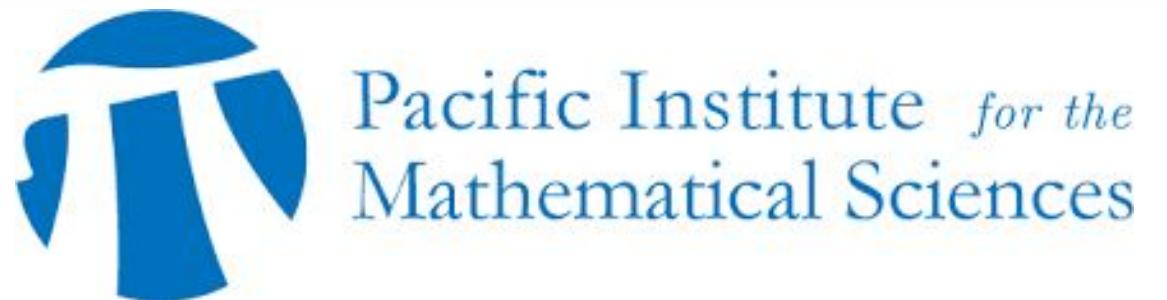
Yiyin Zhou

<http://fruitflybrain.org>



# National in A structure, from K-12 to HPC

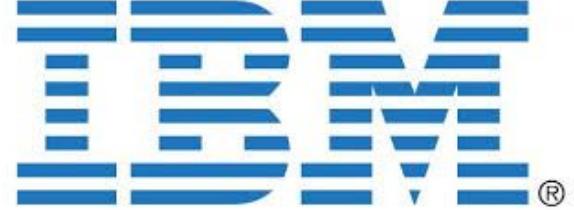
The screenshot shows the Cybera website with a black header bar containing navigation links: CYBERA, NETWORK, SERVICES, PROJECTS, NEWS & EVENTS, MEMBERSHIP, and CONTACT US. Below the header, the text "Jupyter 'All-in-One' Science Platform" is displayed. On the left, there is a logo for "jupyter" consisting of three colored dots (orange, yellow, and grey) connected by lines. To the right of the logo, the text "Learning and sharing in a flexible, collaborative and interactive way." is followed by a paragraph about Jupyter. Below this text are two images: one showing people working in a modern office space, and another showing two men looking at a computer screen next to a 3D printer. At the bottom, there are two sections: "Who Is Jupyter Useful For?" and "Use Cases".



J. Colliander,  
I. Allison,  
B. Carra

The screenshot shows a news article from compute canada | calcul canada. The title is "compute canada and pims launch jupyter service for researchers". The date is 15/03/2017, and the category is Featured, News. The article text discusses the collaboration between compute canada and the Pacific Institute for the Mathematical Sciences (PIMS) to launch a Jupyter service. It includes a colorful graphic of many small arrows pointing upwards and outwards from a central point. Below the graphic, there is a sidebar with links: Home, About, Renewing Canada's Advanced Research Computing Platform, Research Data Management, and News.

The screenshot shows the SYZYGY.CA landing page. The background features a large, swirling orange and brown pattern resembling a celestial body or a nebula. In the center, the text "SYZYGY.CA" is prominently displayed in large white letters, with "SYZYGY" being larger than ".CA". Below this, the text "Launch Jupyter at your university, school or company?" is shown. On the right side, there is a form with input fields for "Your name", "Your email", and "Your message", along with a "SEND" button. The top navigation bar includes links for ABOUT, PARTNERS, INTRO, LAUNCH (which is highlighted in blue), EN, and FR.



# Microsoft



# Azure



**co**  
CODE OCEAN

The screenshot shows the Microsoft Azure Notebooks interface. It features a header with 'Microsoft Azure Notebooks' and 'PREVIEW'. Below the header are links for 'Overview', 'Libraries', 'FAQ/Support', and 'What's New'. A 'Sign in' button is also present. The main content area is titled 'jupyter' and displays the text 'Notebooks hosted on Microsoft Azure'. It includes two buttons: 'Go to my Notebook Server' and 'Show me some samples'. A large image on the right shows a Jupyter notebook interface with code and output. Below this, a section titled 'WHAT IS JUPYTER?' lists three items: 'Interactive Notebooks for Data Science and Technical Computing', 'Browser-based REPL with Markdown and inline interactives', and 'Support for Python 2, Python 3 and R'. At the bottom, there are buttons for 'Create notebook instance', 'View jobs', 'View models', and 'View endpoints'.

The screenshot shows the Google Cloud Platform Cloud Datalab interface. It features a header with 'Google Cloud Platform' and a search bar. Below the header, there are links for 'Why Google', 'Products', 'Solutions', 'Launcher', 'Pricing', 'Customers', 'Documentation', 'Support', and 'Partners'. A 'Free Trial' button is also present. The main content area is titled 'CLOUD DATALAB BETA' and describes it as 'An easy to use interactive tool for large-scale data exploration, analysis, and visualization.' A 'TRY IT FREE' button is at the bottom. To the right, there is a sidebar with a table of contents for 'Overview of Colaboratory Features' including sections like 'Cells', 'Working with python', 'System aliases', 'Magics', 'Tab-completion and exploring code', 'Rich, interactive outputs', 'Integration with Drive', and 'Commenting on a cell'. Below the sidebar, there is a preview of a chart titled 'Fills and Alpha Example'.

# Google

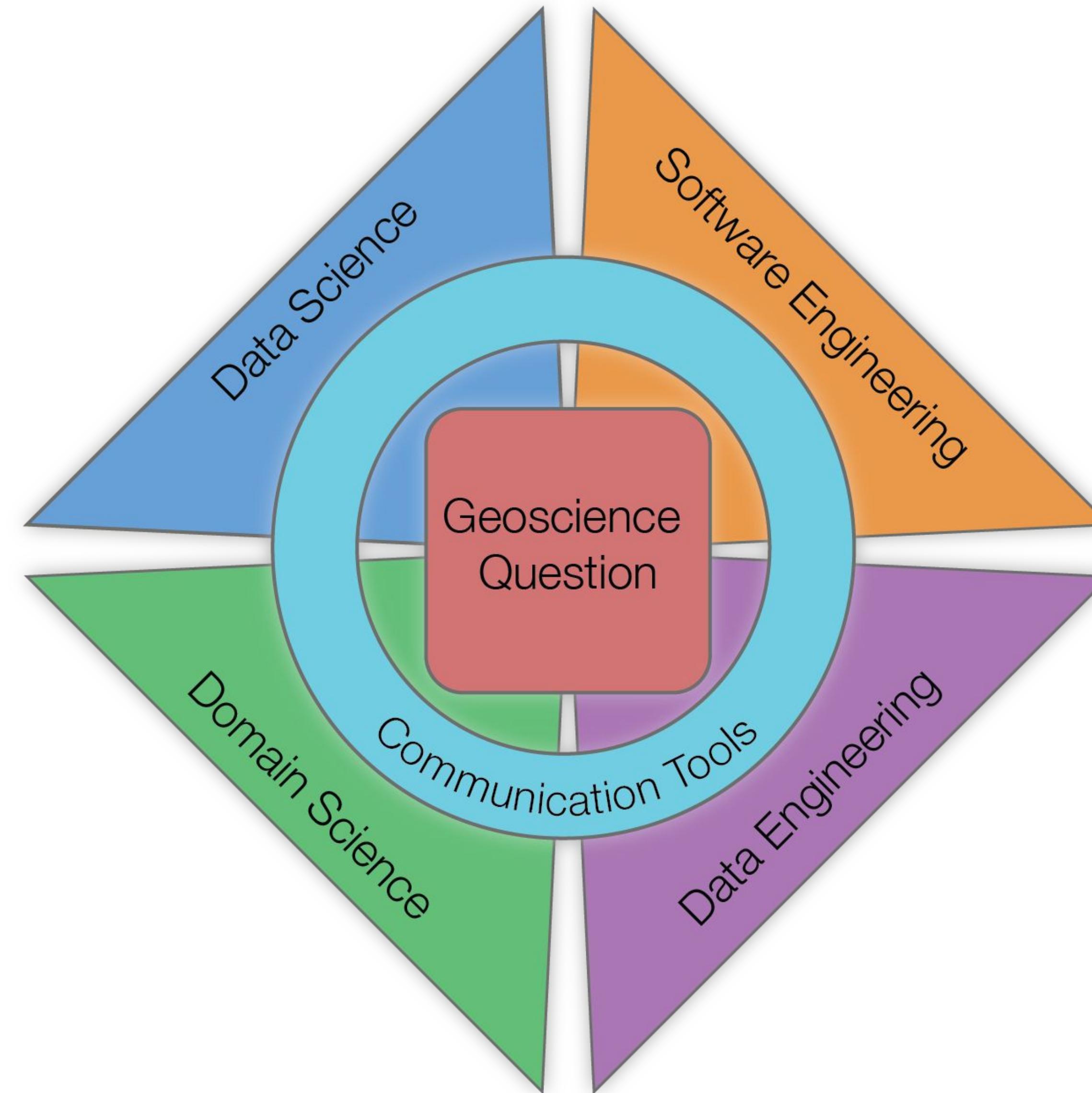
The screenshot shows the Amazon SageMaker interface. It features a large blue 3D cube icon and the text 'Amazon SageMaker'. Below the icon, there is a diagram illustrating the workflow: 'Notebook instance' leads to 'Jobs', which leads to 'Models', which finally leads to an 'Endpoint'. Below the diagram, there are four buttons: 'Create notebook instance', 'View jobs', 'View models', and 'View endpoints'. The background shows a blurred view of a Jupyter notebook interface.

The screenshot shows the IBM Data Science Experience interface. It features a header with 'Data Science Experience' and a search bar. Below the header, there are sections for 'Jupyter Notebooks', 'RStudio', and 'Machine Learning (Coming Soon)'. The 'Jupyter Notebooks' section shows a Jupyter notebook interface with a table and a scatter plot. The 'RStudio' section shows an RStudio interface with a similar dataset. The 'Machine Learning' section is labeled '(Coming Soon)'. At the bottom, there is a 'See All Features' button.

The screenshot shows the Anaconda interface. It features a large green circular logo with a grid pattern. Below the logo, the word 'ANACONDA' is written in a bold, green, sans-serif font. The interface itself is a Jupyter notebook environment with a code cell containing Python code to generate a filled area plot, and a corresponding plot titled 'Fills and Alpha Example'.

# Impact: Research and Education

# Impact: Research and Education



# Fall 2018



Data 8:  
~1,300  
students

Data 100:  
~800 students

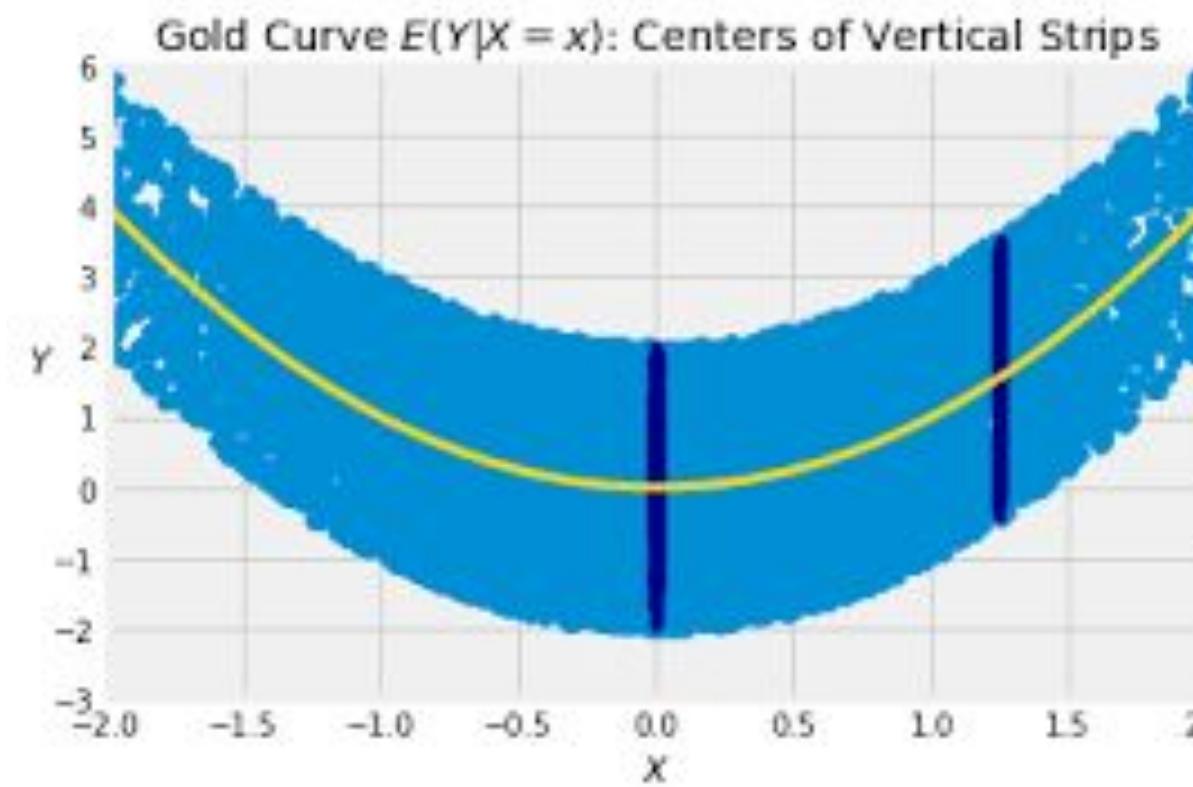


# Prob 140 and Data 102

## PROB 140



Probability for Data Science



<http://prob140.org>: Probability for Data Science

Berkeley **Division of Data Science and Information**

About ▾ Academics ▾ Research ▾ News ▾ Events ▾ Support Us

Home » New Course Takes Data Science to the Next Level

## New Course Takes Data Science to the Next Level



<http://data102.org>: Data, Inference and Decisions

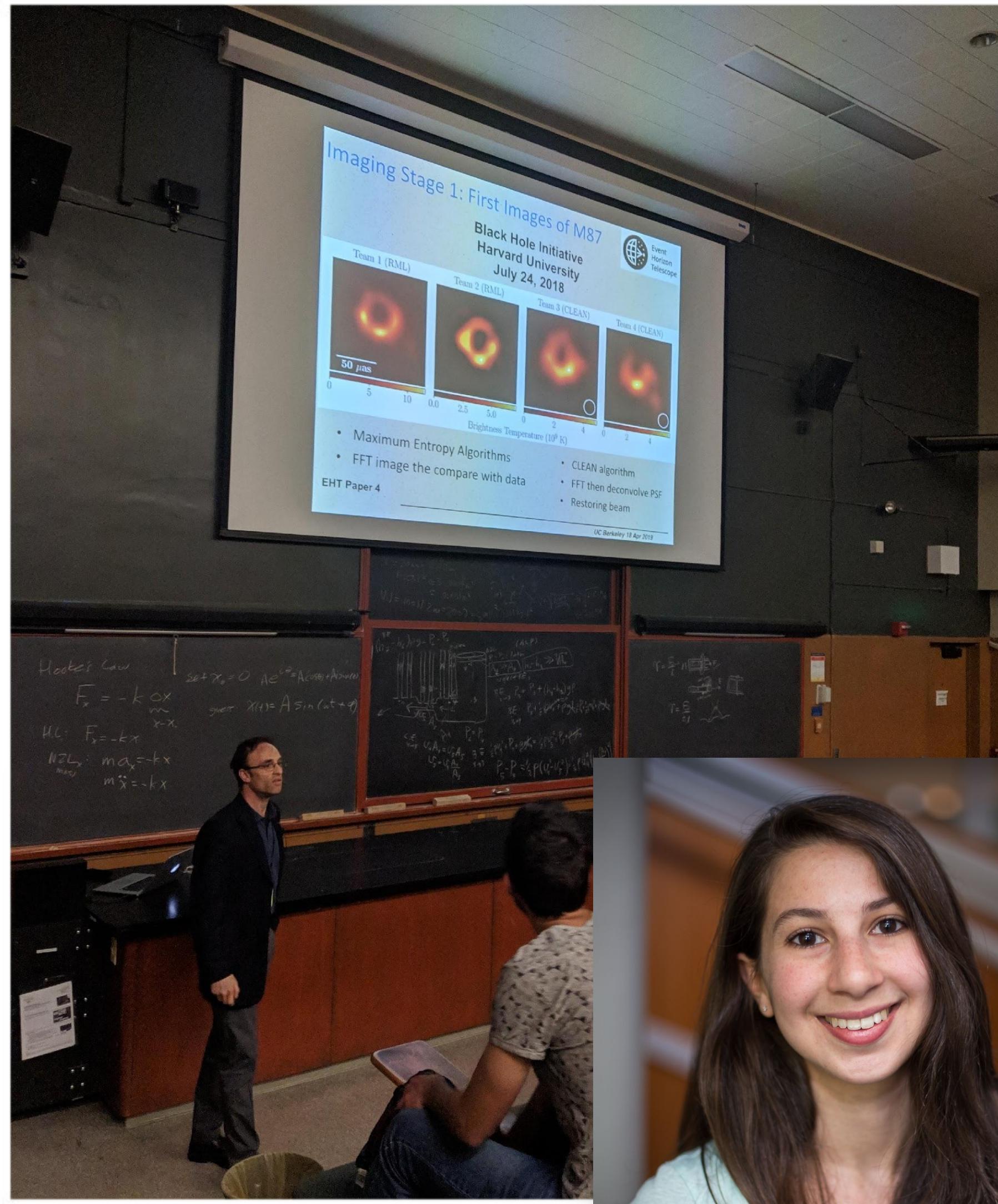
---

# With these tools, we provide

---

- Broad disciplinary reach and impact of statistical thinking.
- Drastically lowered barriers to student access - intellectual and economic.
- Lowered barriers for faculty\* to engage with statistical and computational ideas.
- (\*) often from non computational/statistical domains

# 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<sup>1,2,3,4</sup> , Antxon Alberdi<sup>5</sup> , Walter Alef<sup>6</sup>, Keiichi Asada<sup>7</sup>, Rebecca Azulay<sup>8,9,6</sup> , Anne-Kathrin Baczko<sup>6</sup> , David Ball<sup>10</sup>, Mislav Baloković<sup>4,11</sup> , John Barrett<sup>2</sup>  +Show full author list

Published 2019 April 10 • © 2019. The American Astronomical Society.

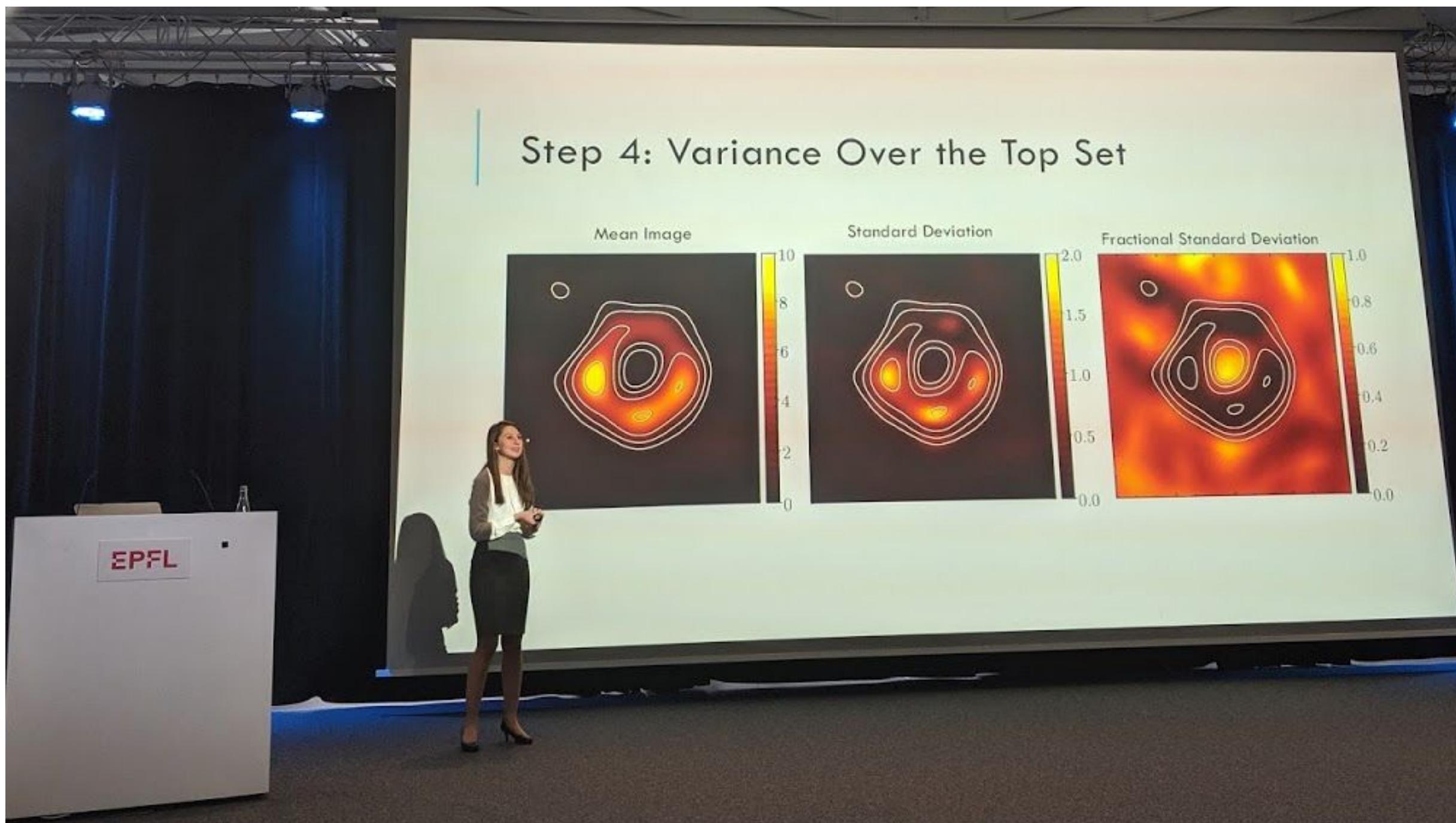
[The Astrophysical Journal Letters, Volume 875, Number 1](#)

**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

# K. Bouman @ EPFL, Oct'19

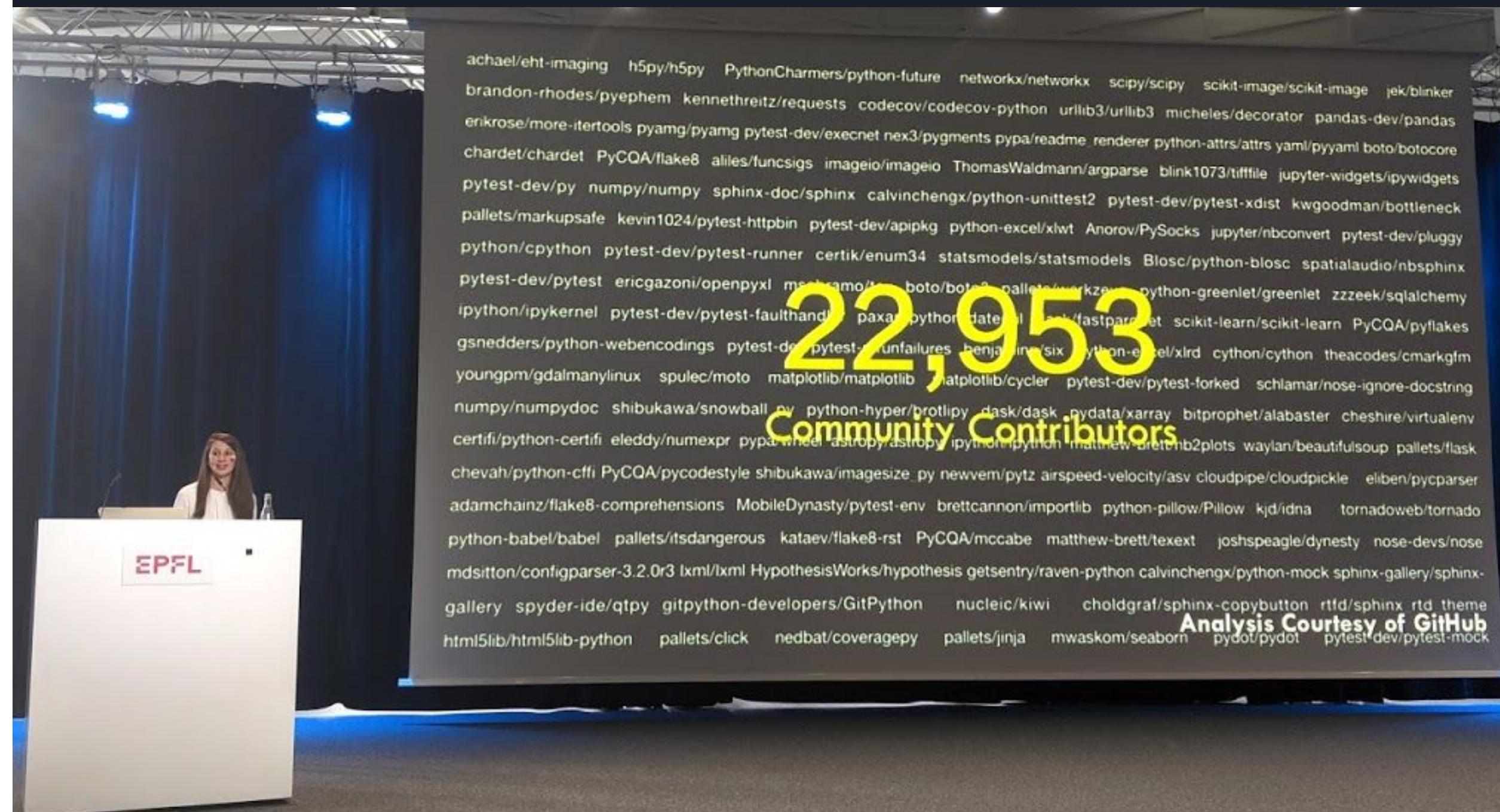


Video of talk: <https://youtu.be/TSgpliktkwc>

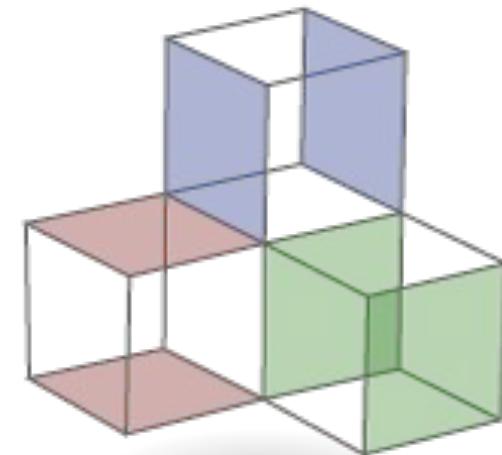


EPFL Open Science  
@EPFLOpenScience

"By continuing this tradition of open and reproducible science, I know we'll be able to continue doing things that may at first glance seem impossible." - Katie Bouman on capturing the first picture of a black hole. Her full #EPFLOpenScienceDay talk here: [youtu.be/TSgpliktkwc](https://youtu.be/TSgpliktkwc)



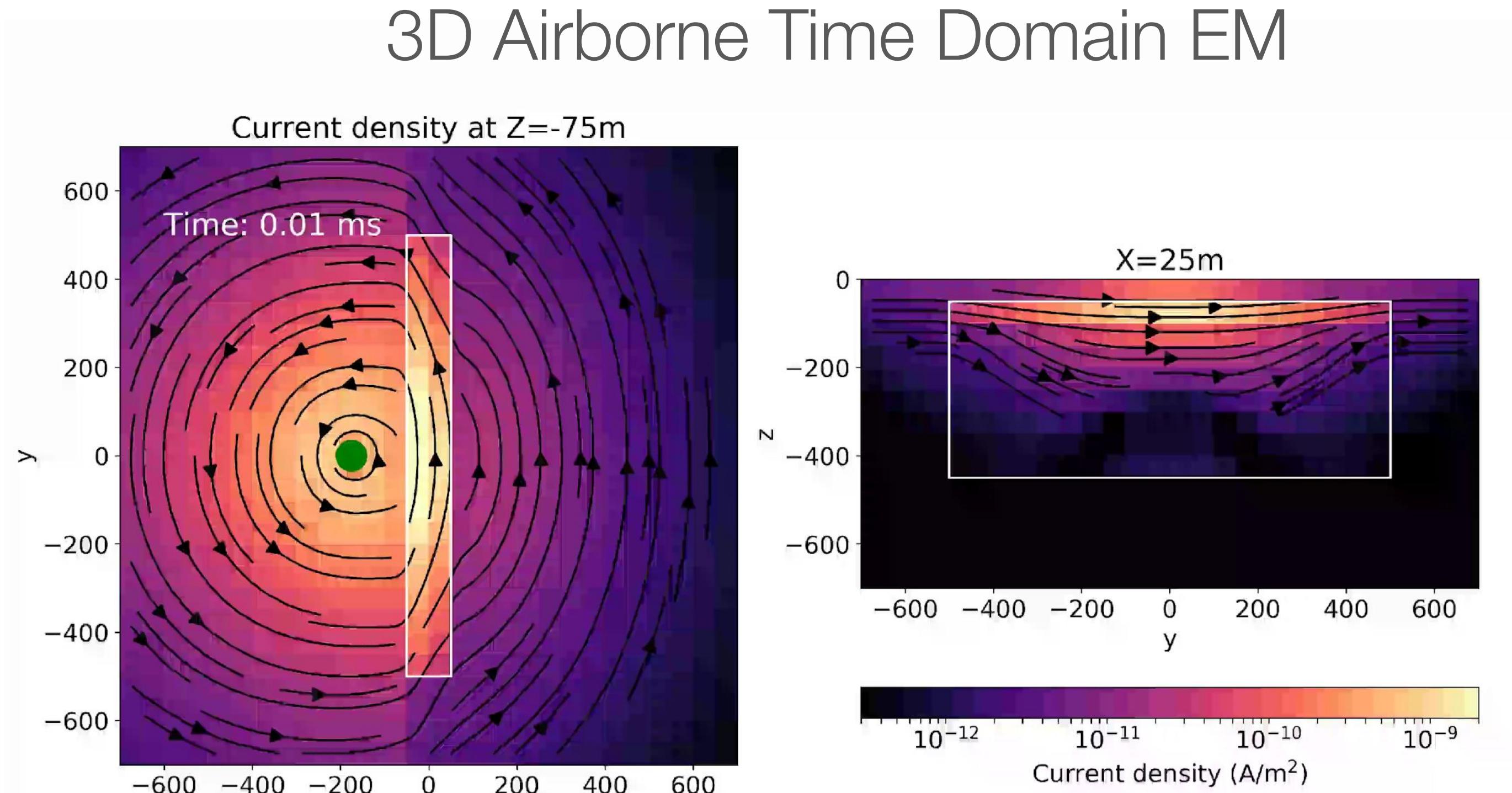
CORE Geoscience?



# simpeg

## tools by and for researchers

- Modular, multi-physics
  - Gravity
  - Magnetics
  - Direct current resistivity
  - Induced Polarization
  - Electromagnetics
    - Frequency Domain
    - Time Domain
  - Fluid Flow
    - Richards Equation



Lindsey



Seogi



Rowan

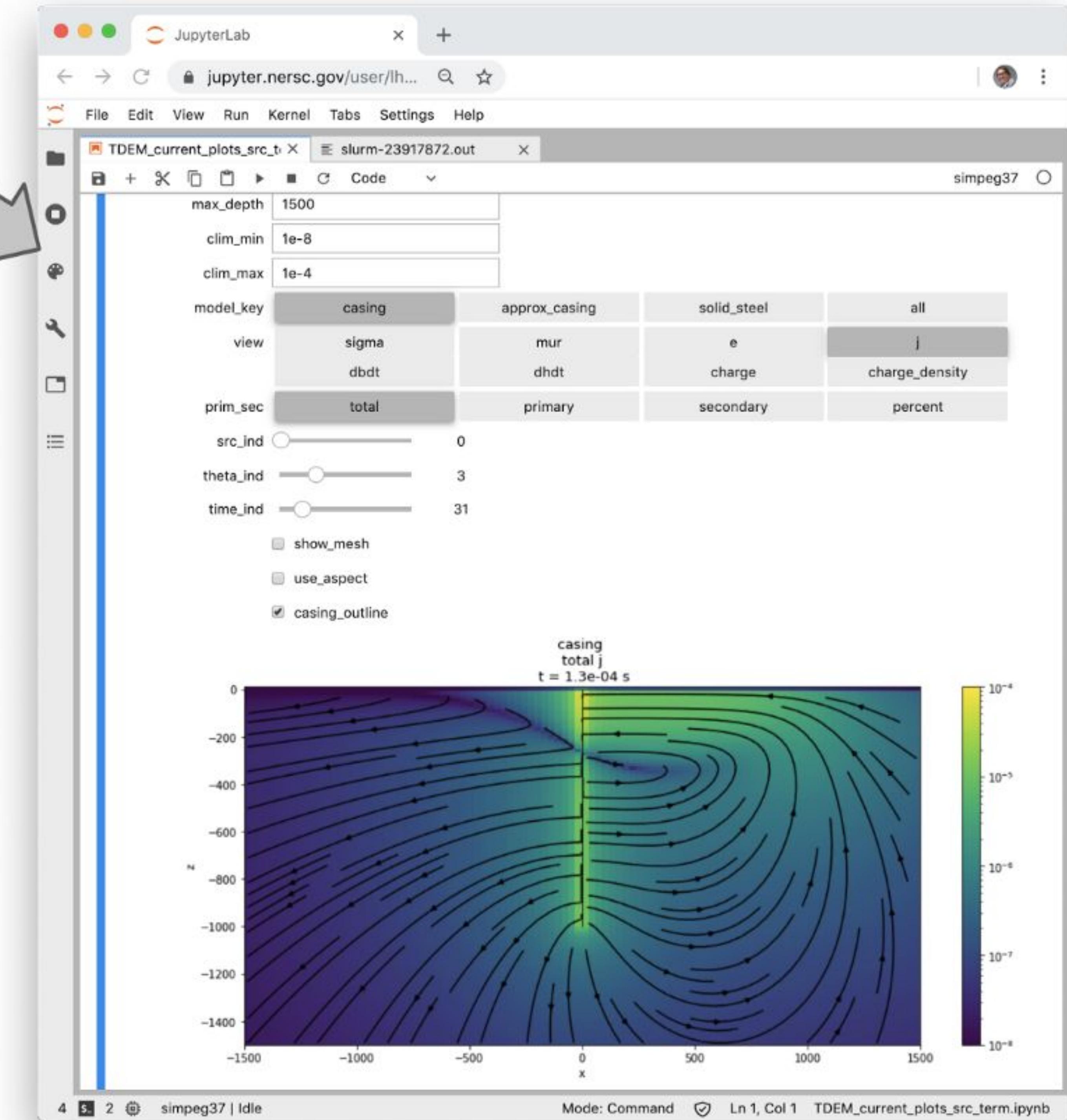
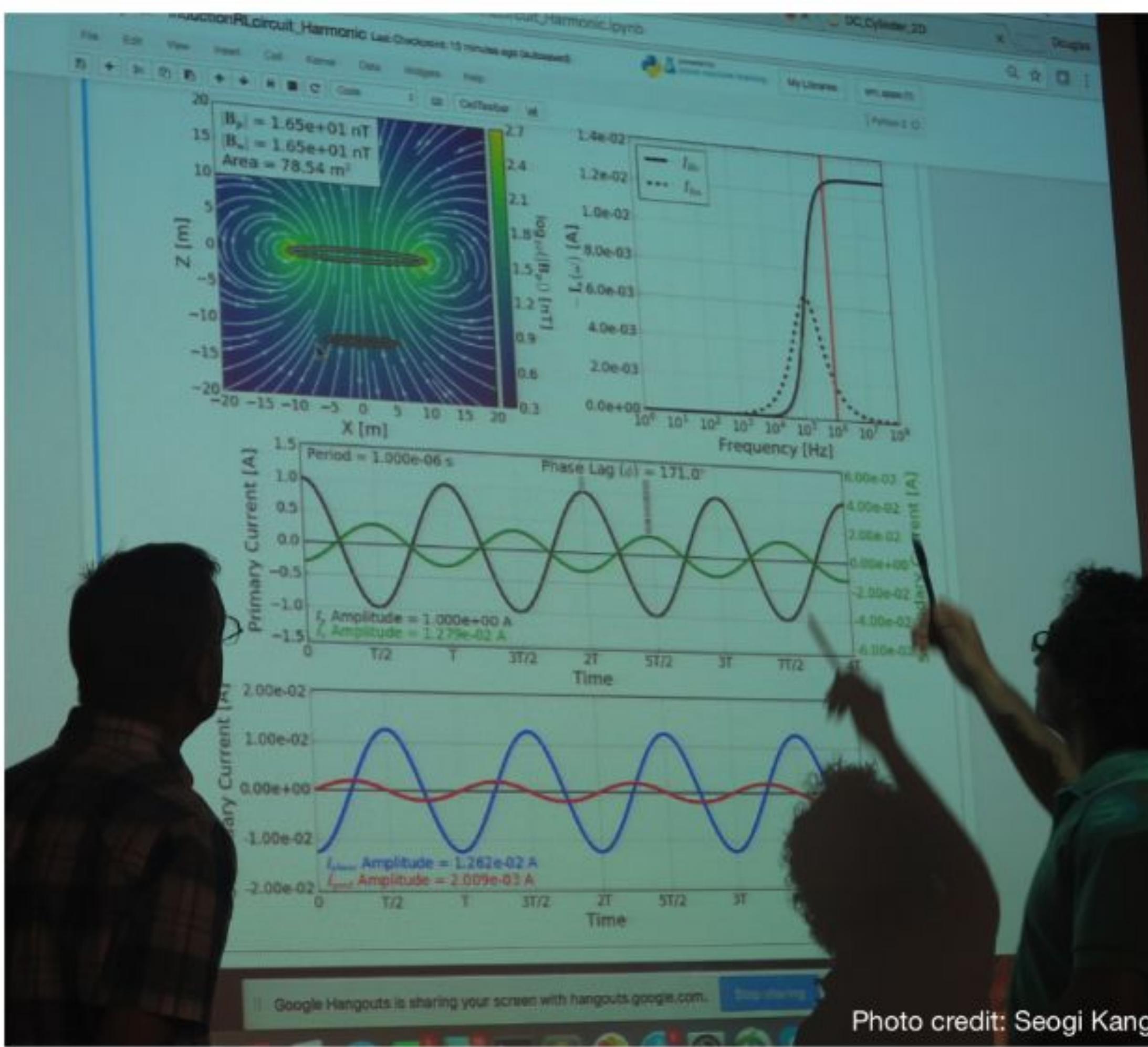


Doug

<https://simpeg.xyz>

$$\nabla \times \vec{E} + i\omega \vec{B} = 0$$

$$\nabla \times \mu^{-1} \vec{B} - \sigma \vec{E} = \vec{J}_s$$



Users  
30K  
↑28%  
vs last year

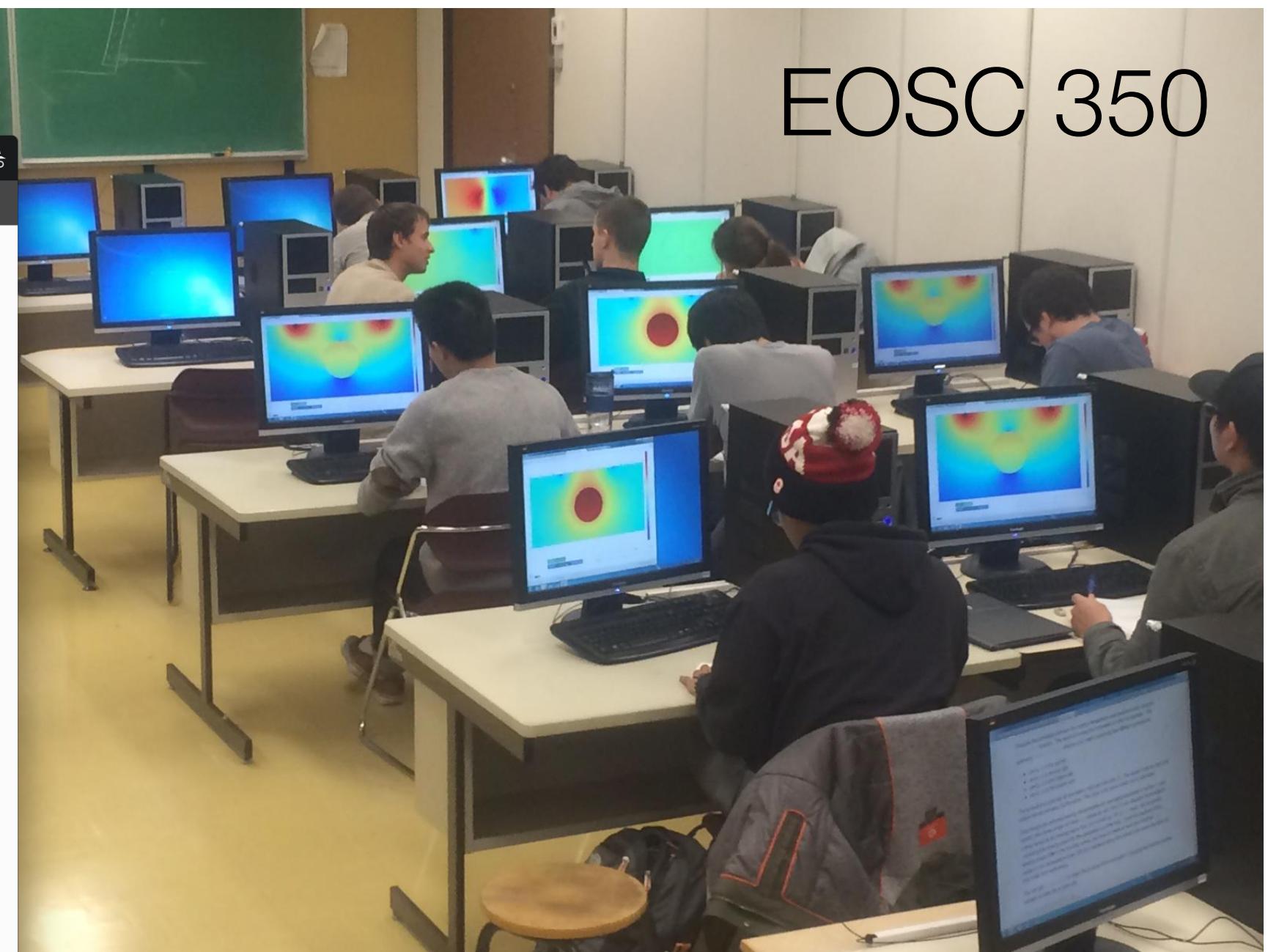
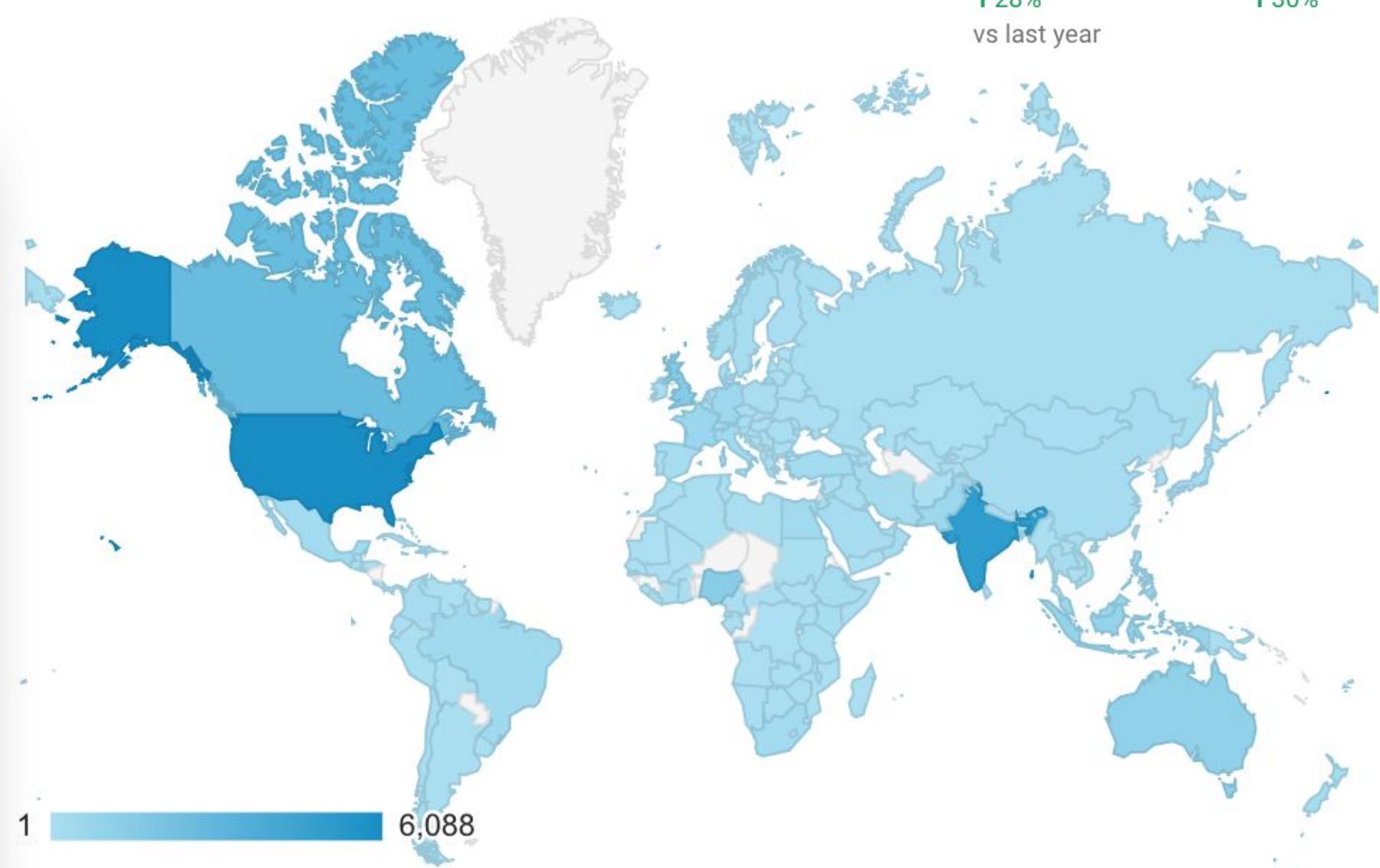
Sessions  
48K  
↑30%

# GeoSci.xyz

<https://geosci.xyz>



26 locations worldwide



EOSC 350



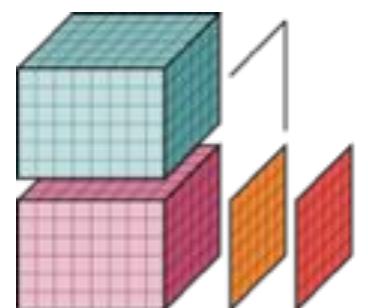
Harnessing the power of cloud computing to study the whole Earth interactively



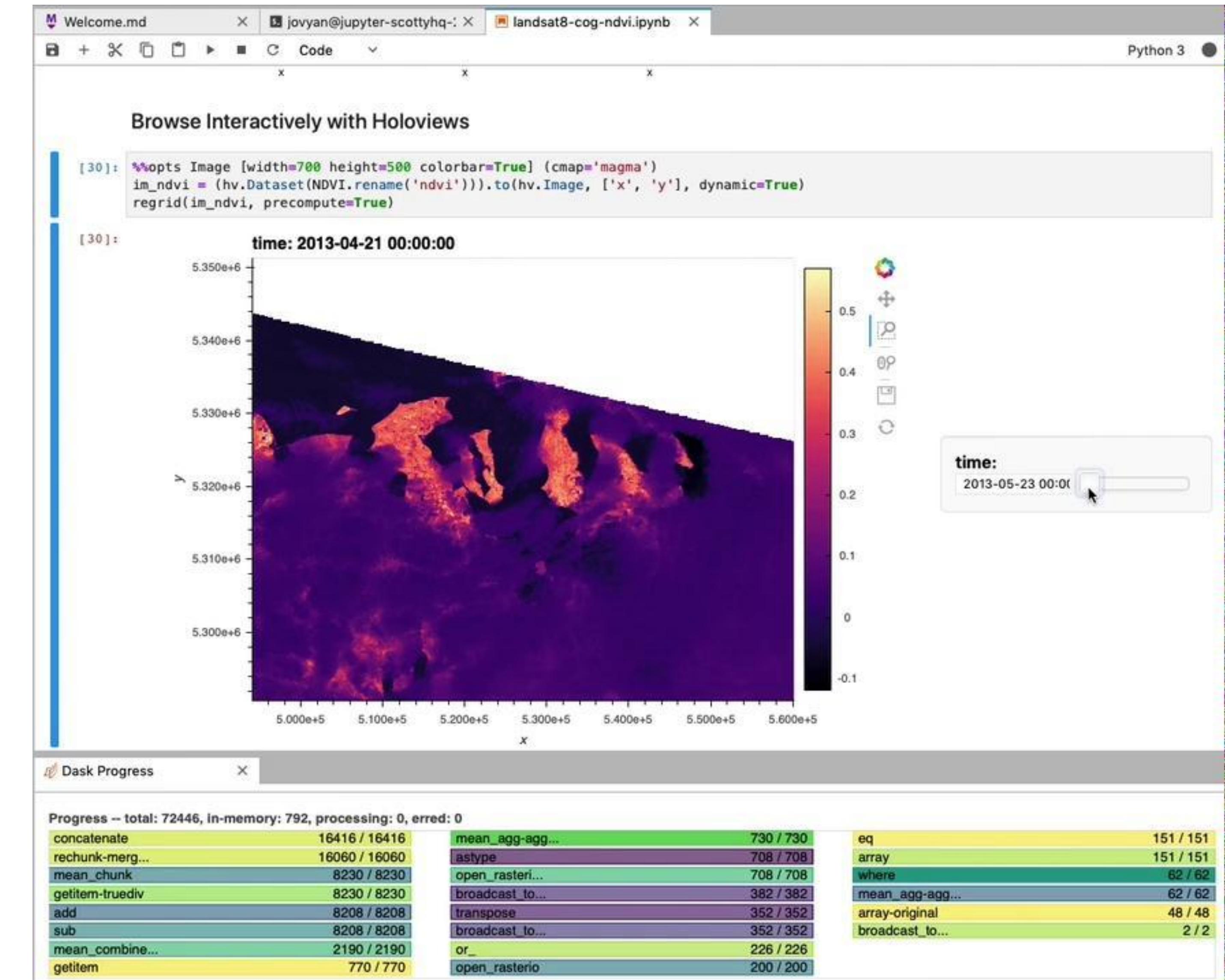
Interactivity



Distributed computing



xarray



Scott Henderson

[Follow](#)

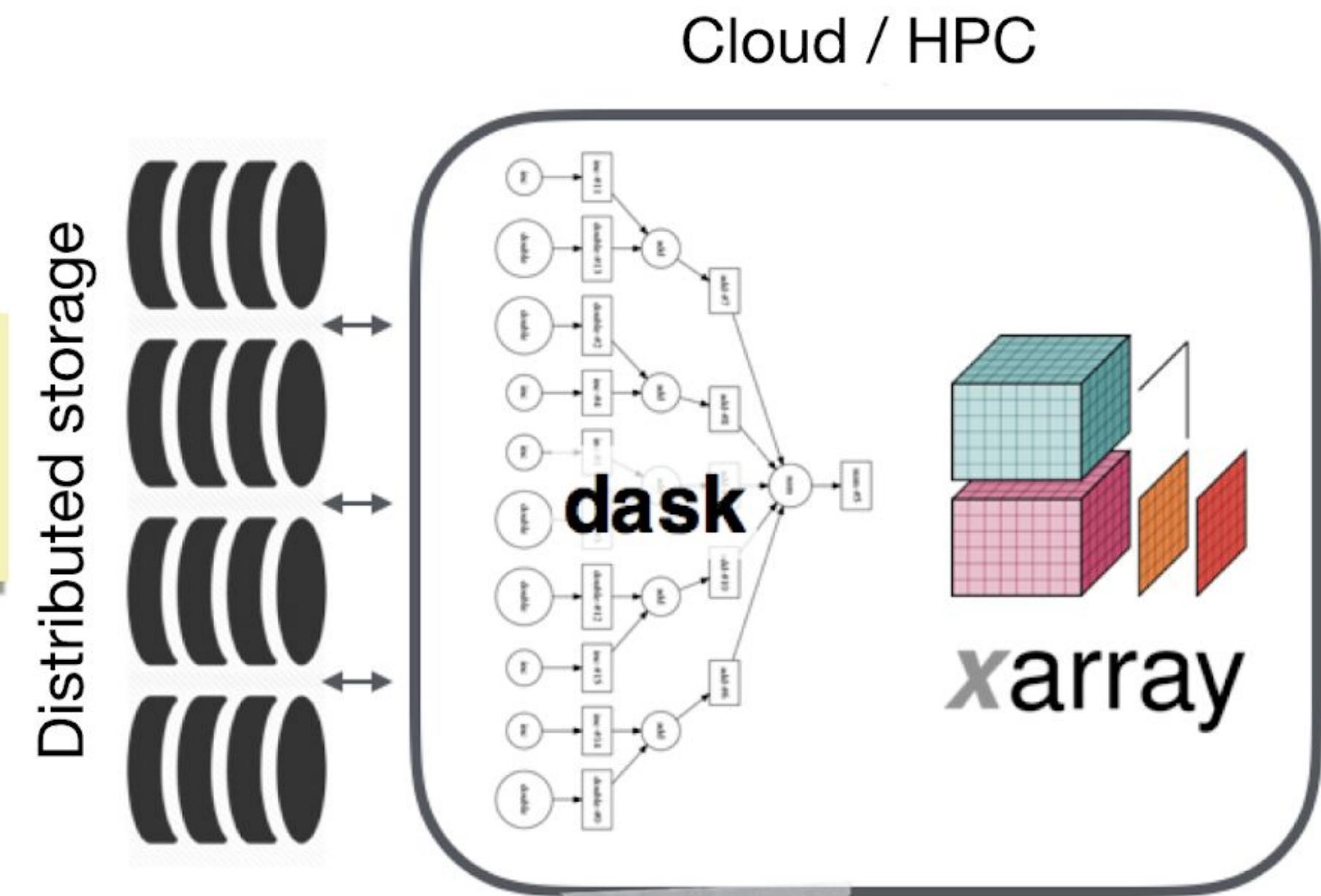
Research geophysicist at University of Washington eScience Institute

Oct 1 · 7 min read



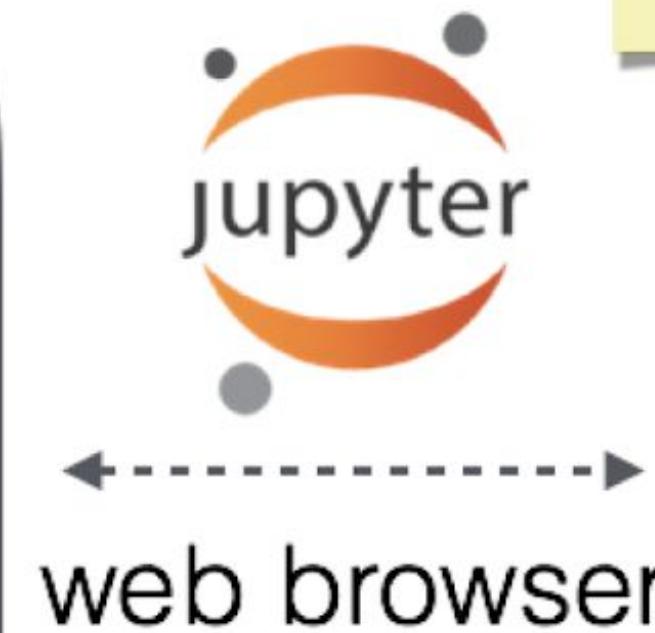
# Pangeo Architecture

**“Analysis Ready Data”**  
stored on globally-available  
distributed storage.



Parallel computing system allows users deploy clusters of compute nodes for data processing.

Dask tells the nodes what to do.



Jupyter for interactive access remote systems

end user



Xarray provides data structures and intuitive interface for interacting with datasets

# What is Pangeo?

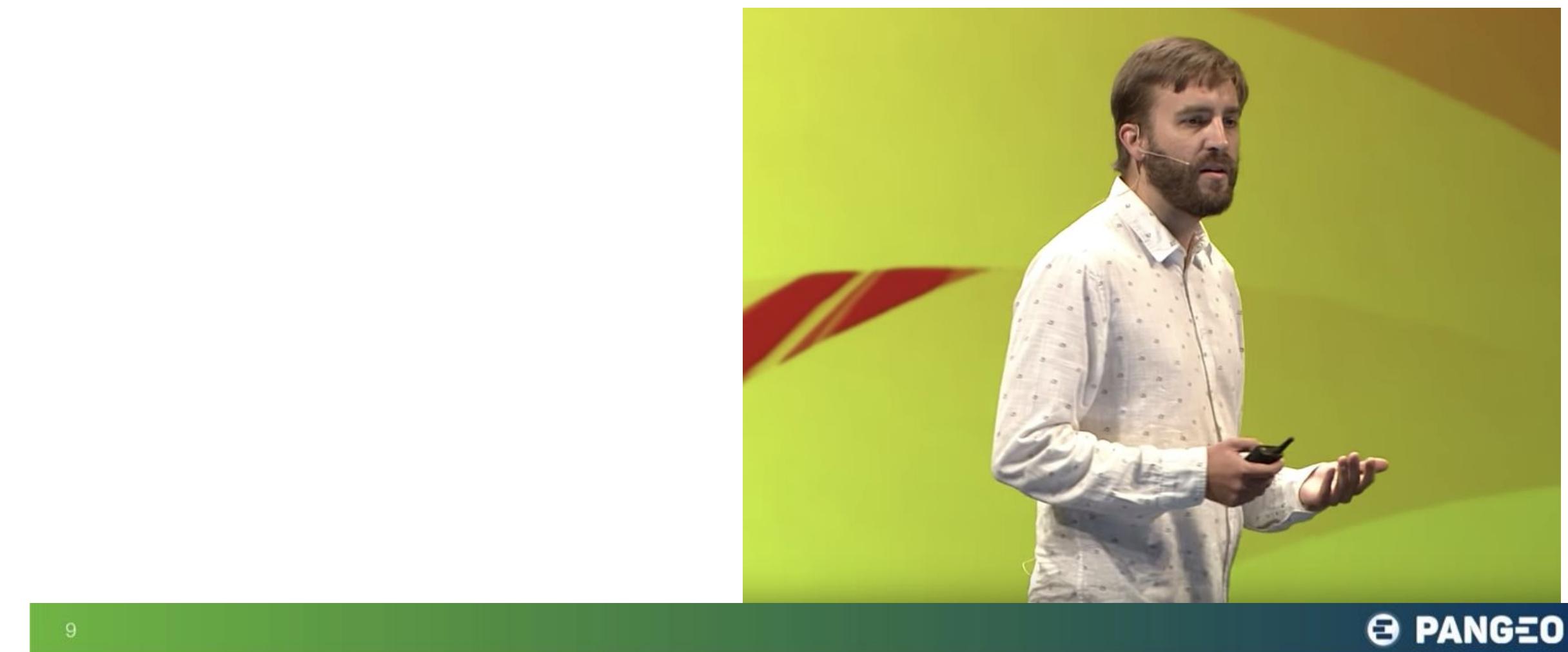
*The Future of Data-driven Discovery in the Cloud*  
Ryan Abernathey (Columbia), JupyterCon 2018  
<https://www.youtube.com/watch?v=7kDYfUe0Zhw>

WHAT IS PANGEO?

Pangeo is a community working to develop software and infrastructure to enable big-data geoscience.

- **Mission:** To cultivate an ecosystem in which the next generation of open-source analysis tools for the big-data geosciences can be developed, distributed, and sustained.
- **Vision:**
  - ▶ Open and collaborative development
  - ▶ Tools for scaling computations from small to very large datasets
  - ▶ Frameworks for moving scientific analysis to the data
  - ▶ Welcoming and inclusive development culture

SciPy 2018

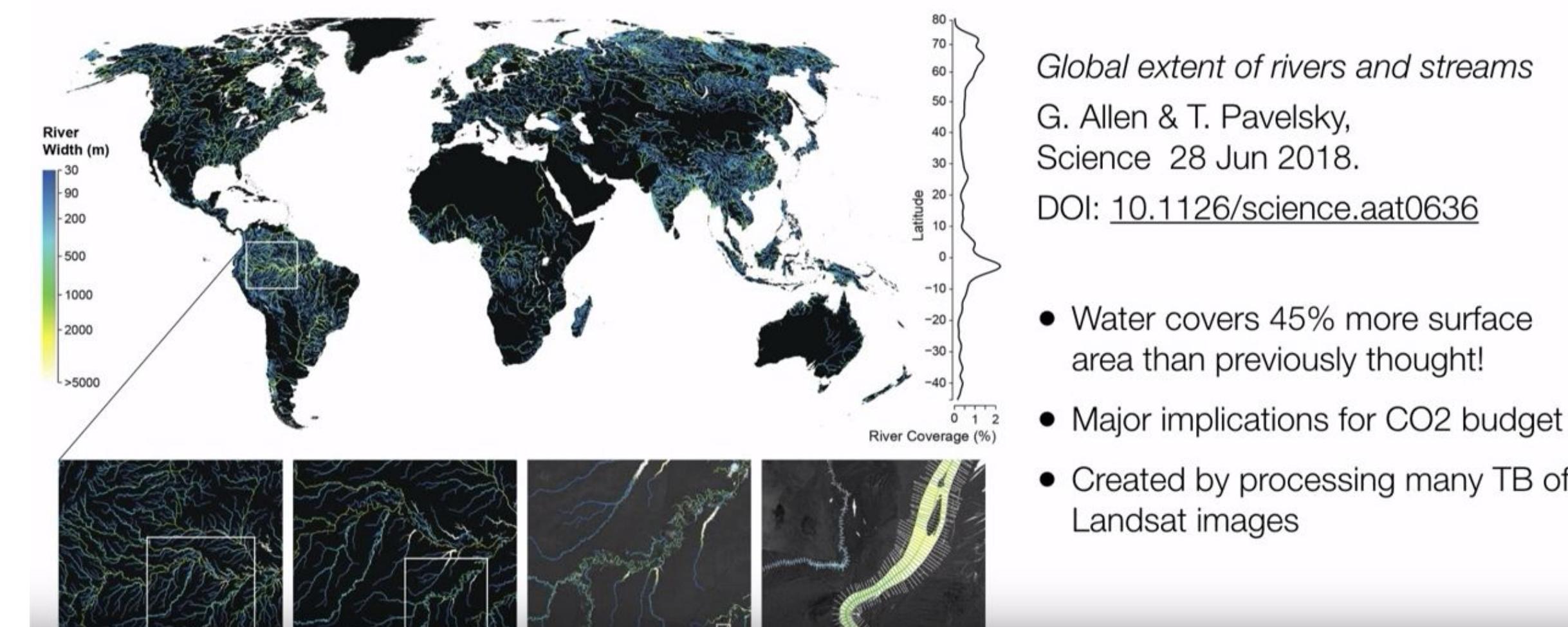


## Pangeo: A Big Data Ecosystem for Scalable Earth System Science

Joe Hamman (NCAR), SciPy 2018

<https://www.youtube.com/watch?v=2rgD5AJsAbE>

## BIG SCIENCE FROM BIG DATA!



# Jupyter meets the Earth: an NSF grant (2M / 3Y)!



## Research use-cases

- Climate data analysis
- Hydrologyd
- Geophysics



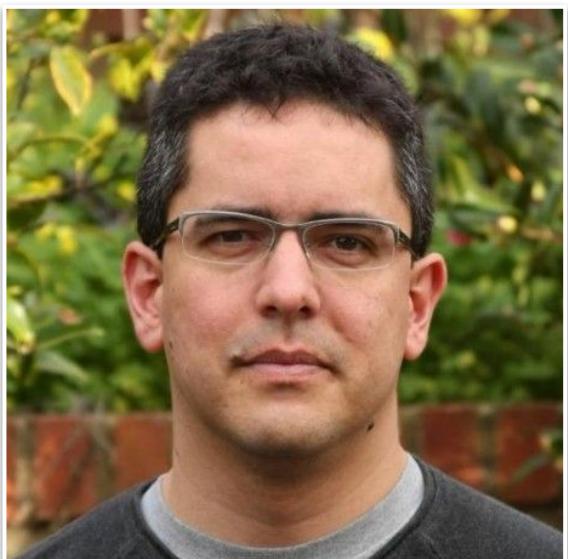
PANGEOT



## Tech developments

- Data discovery
- Interactivity
- Cloud/HPC infrastructure

For more: <http://bit.ly/jupytearth>



Fernando  
Pérez



Joe  
Hamman



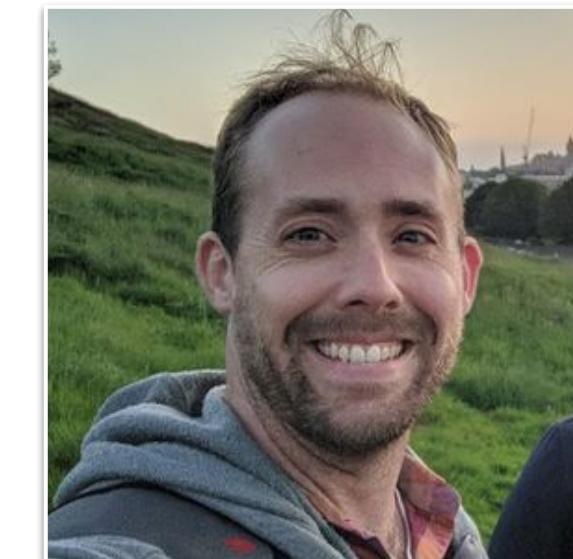
Laurel  
Larsen



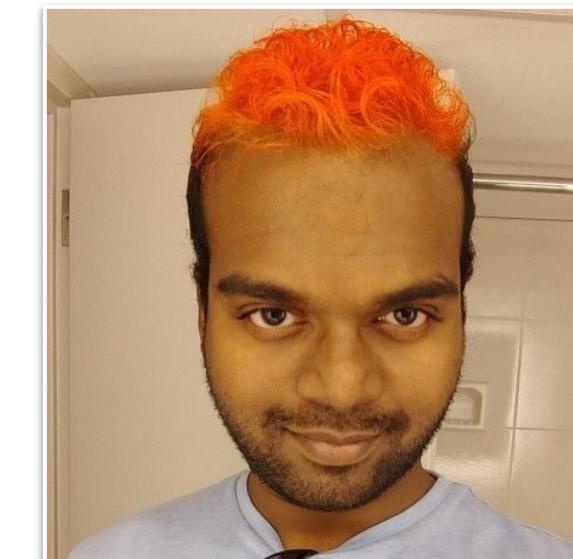
Kevin  
Paul



Lindsey  
Heagy



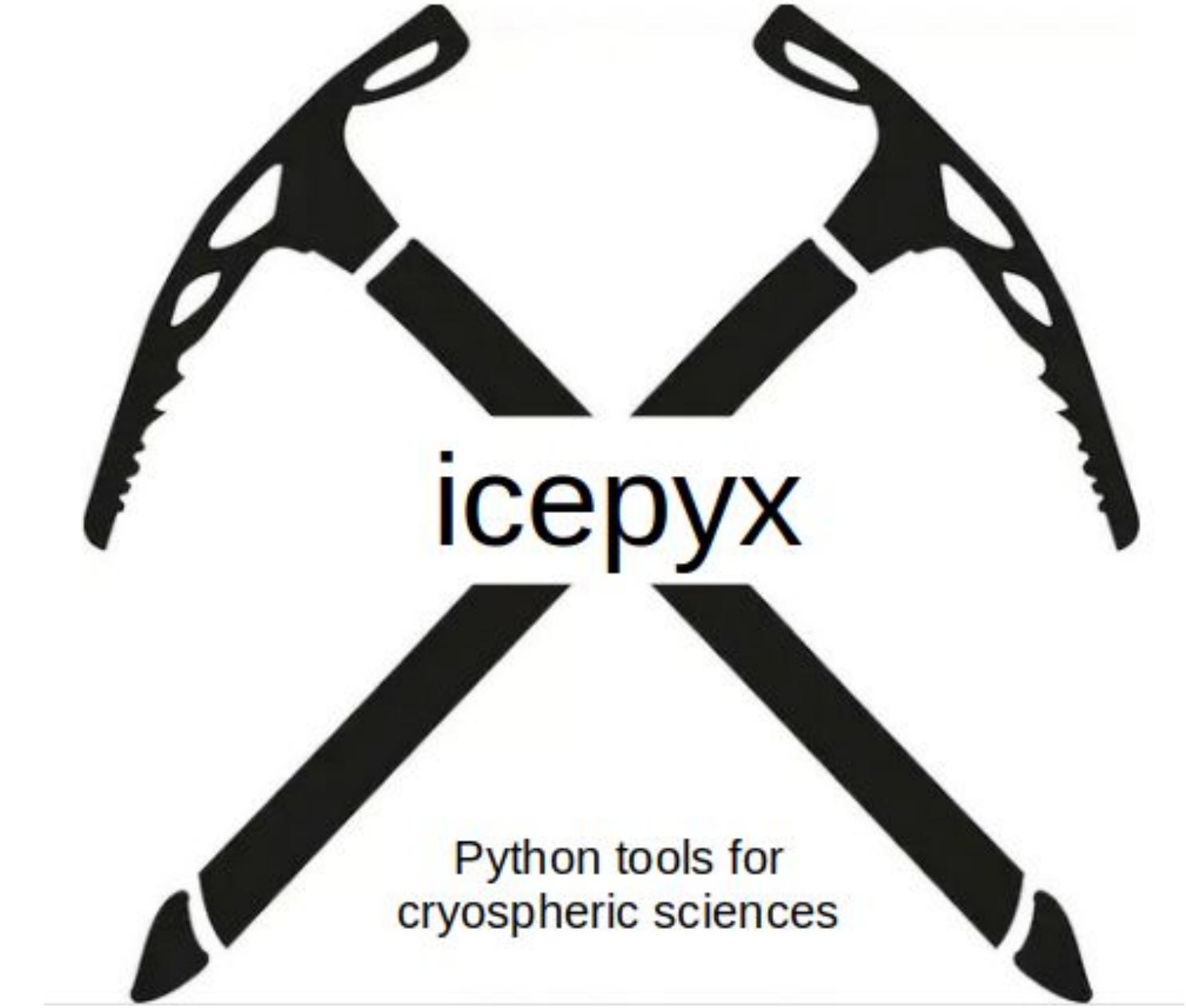
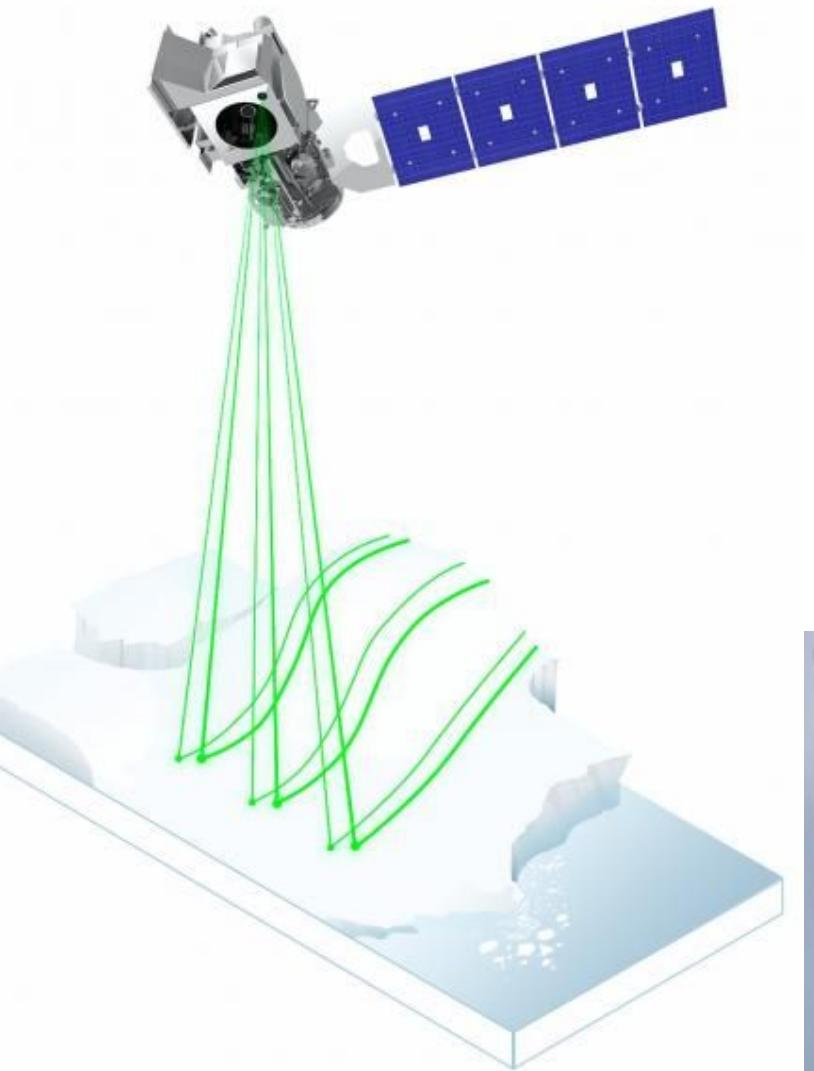
Chris  
Holdgraf



Yuvi  
Panda



ICE, CLOUD, AND LAND ELEVATION SATELLITE-2



**CRYOSPHERIC SCIENCE WITH ICESAT-2 HACKWEEK 2019**

WORKSHOP ON ICESAT-2 DATASETS FOR CRYOSPHERIC STUDIES  
UNIVERSITY OF WASHINGTON  
JUNE 17-21, 2019

[WIKI](#)   [GITHUB](#)   [SCHEDULE](#)   [LOCATION](#)   [PROJECTS](#)

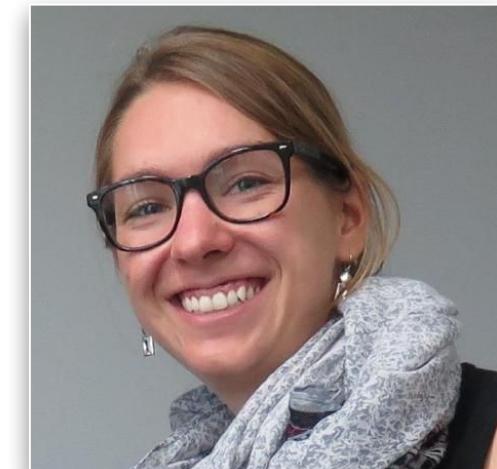
<https://icesat-2.gsfc.nasa.gov>



A. Arendt



J. Scheik



L. Heagy

# Jupyter at AGU 2019

**Jupiter with Jupyter** *Lessons from Teaching Data Visualization & Statistics in Geosciences*

Abigail Azari ([azari@umich.edu](mailto:azari@umich.edu))  
Co-Authors & Co-Instructors: M. W. Liemohn & B. M. Swiger  
Acknowledgements: A. Huang-Saad

Presented at Fall AGU, December 13<sup>th</sup>, 2019

Photo Credit: NASA / SwRI / MSSS / Gerald Eichstädt / Séan Doran

AGU100 ADVANCING EARTH AND SPACE SCIENCE

M COLLEGE OF ENGINEERING CLIMATE AND SPACE SCIENCES AND ENGINEERING UNIVERSITY OF MICHIGAN

Australian National University

**Underworld in the Cloud — a research code for all to use**  
a.k.a. Research-driven education tools ...

Louis Moresi  
The UNDERWORLD team  
The AuScope AVRE team

The ENKI project seeks to unify these efforts within an open-source user-friendly environment

[www.enki-portal.org](http://www.enki-portal.org)

**WELCOME TO ENKI**  
ENabling Knowledge Integration

Model at the speed of thought

S. Wolf - Thermodynamic Modeling Using ENKI: | Motivation | ENKI | Applications | AGU, Dec 2019

seismo-live.org

A Jupyter Notebook Training Platform for Seismology

Heiner Igel, Sebastian Noe, Tobias Megies, Lion Krischer  
<sup>1</sup>LMU Munich, Germany  
<sup>2</sup>ETH Zürich, Switzerland

Seismo-Live  
Live Jupyter Notebooks for Seismology

Help - What is this? Launch More Information

48 of 50 containers are currently available

python NumPy xarray matplotlib

## Interactive Climate Modeling and Reproducible Workflows in the Classroom

Brian E. J. Rose  
University at Albany (SUNY)

ED52A - Linking Education and Research with Jupyter | AGU Fall Meeting 2019

NSF

Using Python and Jupyter Notebook to Teach Undergraduate Climate Data Analysis

Karen M. Shell

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University  
[Karen.Shell@oregonstate.edu](mailto:Karen.Shell@oregonstate.edu)  
<https://github.com/karensshell/climate-data-class>

Big thanks to:

UCAR COMMUNITY PROGRAMS unidata Data Services and Tools for Geoscience

Oregon State UNIVERSITY OSU College of Earth, Ocean, and Atmospheric Sciences

So you want to build Data Science tools  
in academia...

# Jupyter - funding and resources



ALFRED P. SLOAN  
FOUNDATION

GORDON AND BETTY  
**MOORE**  
FOUNDATION

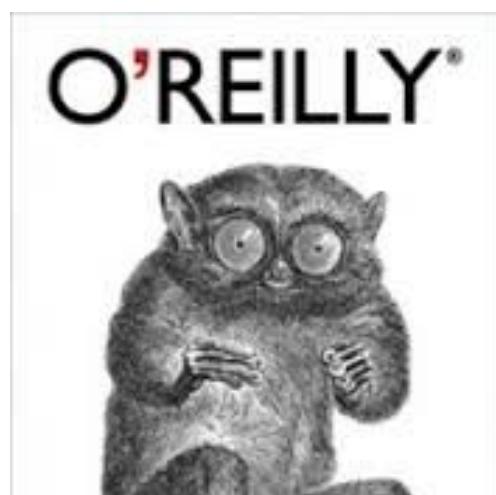
THE LEONA M. AND HARRY B.  
**HELMSLEY**  
CHARITABLE TRUST

U.S. DEPARTMENT OF  
**ENERGY**

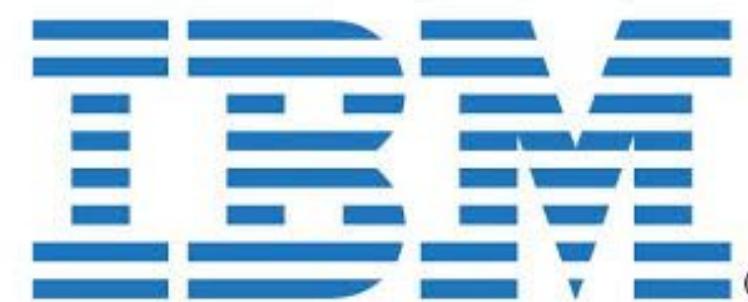


CHAN  
ZUCKERBERG  
INITIATIVE

SIMONS FOUNDATION



ANACONDA®

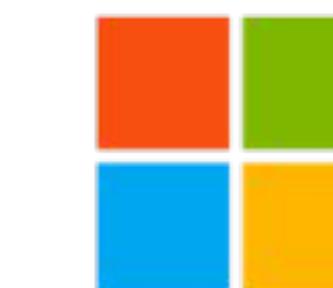


**NETFLIX**

POWERED BY  
**rackspace**  
the open cloud company

**ENTHOUGHT**  
SCIENTIFIC COMPUTING SOLUTIONS

**Google**

 Microsoft

**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.



 CoCalc Blog



## 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

---

# Scientific Open Source: *Despite* (direct) federal \$\$ support

---

- “Indirectly”, lots of \$ have supported Scientific OSS projects/tools.
  - Under the cover of domain-focused work.



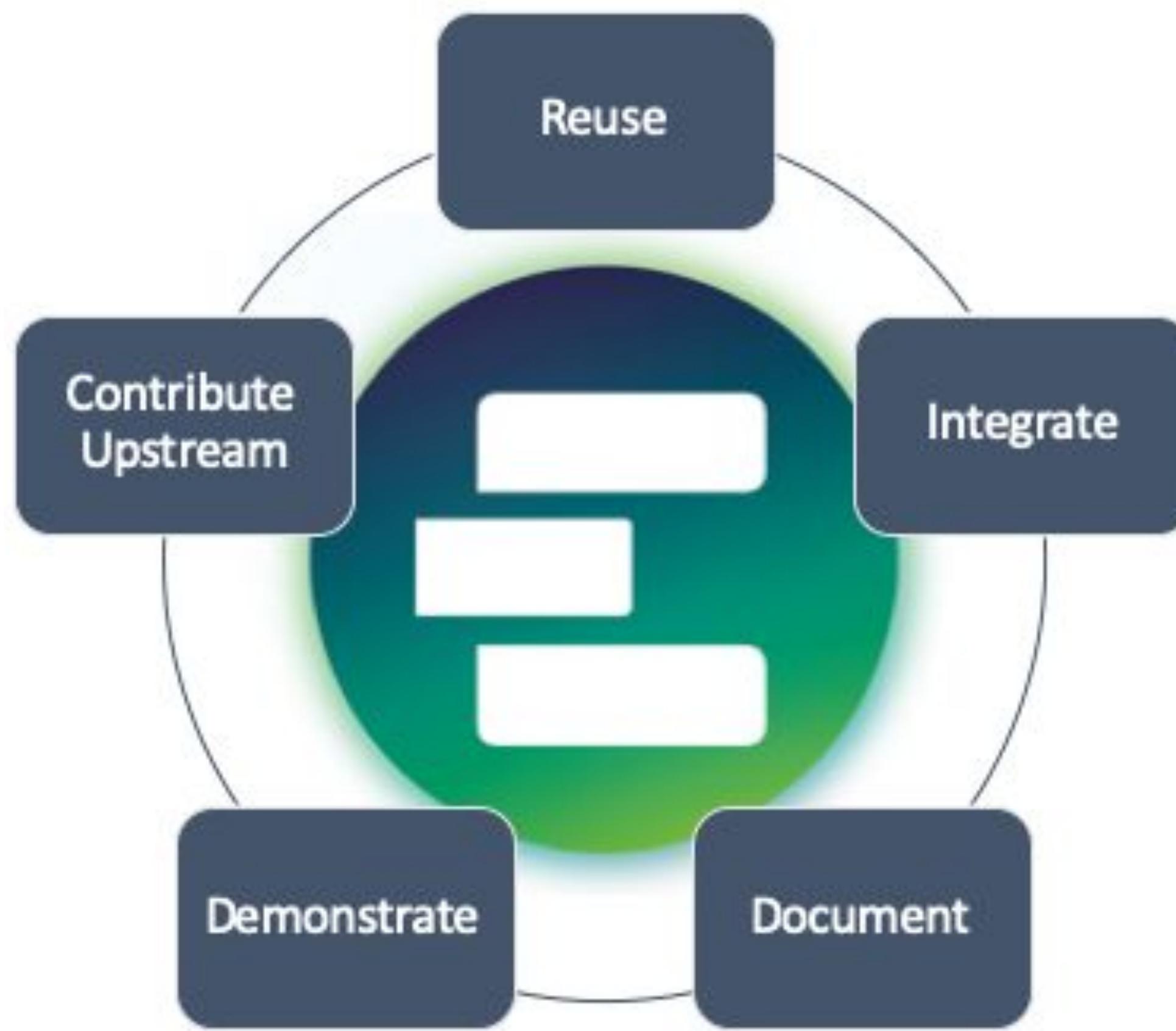


# Traditional software infrastructure funding

Yes, it's true, the budget is gone again... But you can't deny that now, we get here in an instant!

Quino (Argentinian cartoonist)

# The Pangeo Pattern

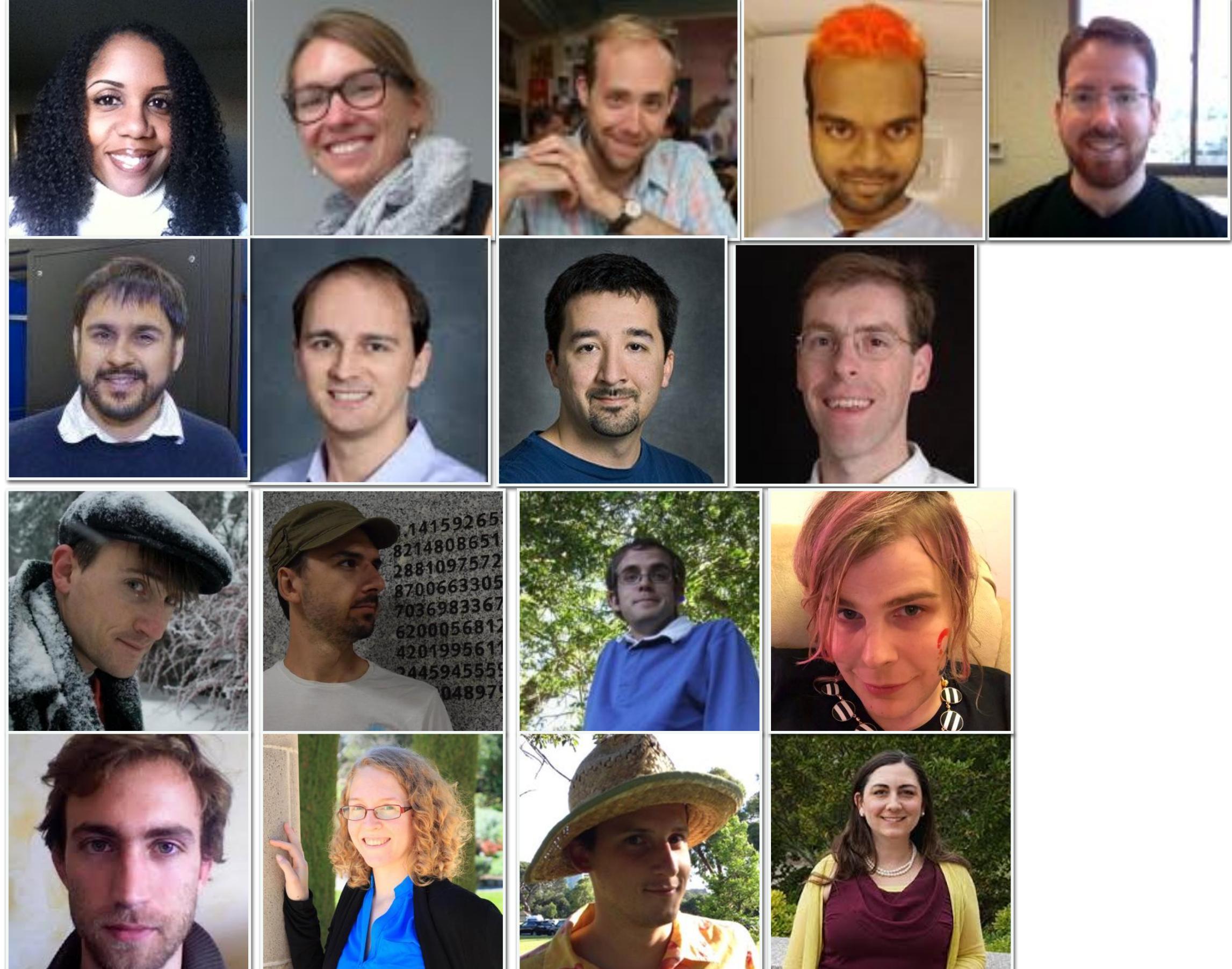


1. *Reuse*: resist the temptation to build something new
2. *Integrate*: use the modular ecosystem as intended
3. *Document*: explain concrete use cases that impact your community
4. *Demonstrate*: offer public infrastructure ([binder.pangeo.io](https://binder.pangeo.io))
5. *Contribute upstream*: help sustain your own ecosystem

# Challenges: contrasts in cultures and incentives

	<b>Open Source</b>	<b>Academia</b>
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”

# Thank you (Bay Area team)



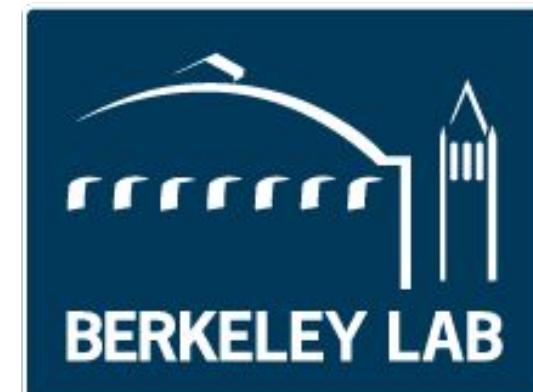
Current (Berkeley, LBNL, Bloomberg)  
Stacey Dorton, Lindsey Heagy, Chris Holdgraf, Yuvi  
Panda, Ryan Lovett, Shreyas Cholia, Shane Canon,  
Rollin Thomas, Jason Grout

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



**Bloomberg**

# Open, collaborative geoscience

- Meet scientific/technical challenges
  - and interdisciplinary ones
- Research/education partnership
- Engage with society/policy makers
- Have a better community experience!

Thank You!



@fperez\_org, @lindsey\_jh



{fernando.perez, lheagy}@berkeley.edu