How the union of inferential thinking and computation are transforming research and education at Berkeley.

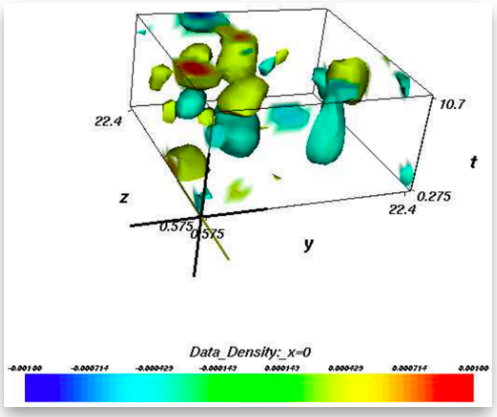
Fernando Pérez

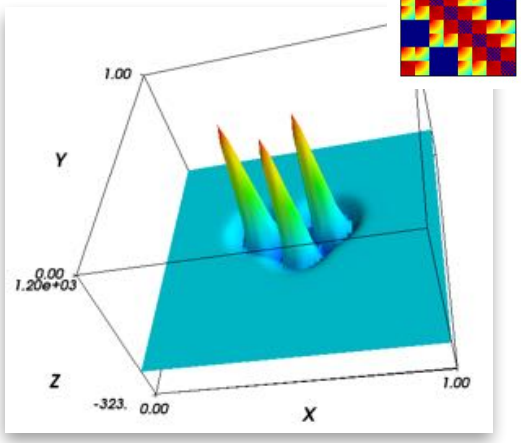
fernando.perez@berkeley.edu

A really odd career

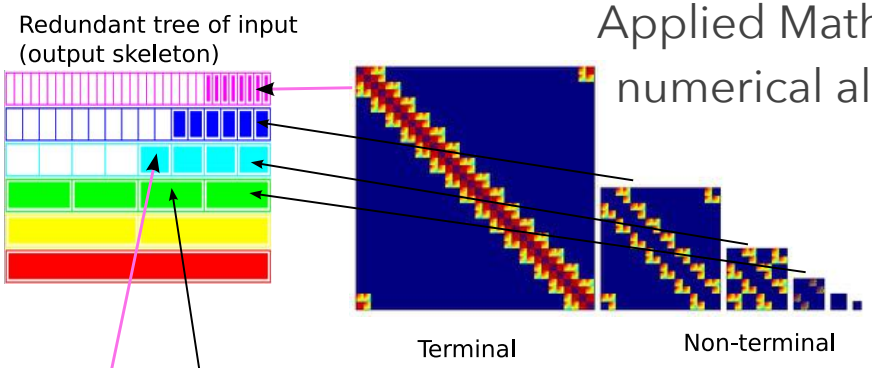
Boulder

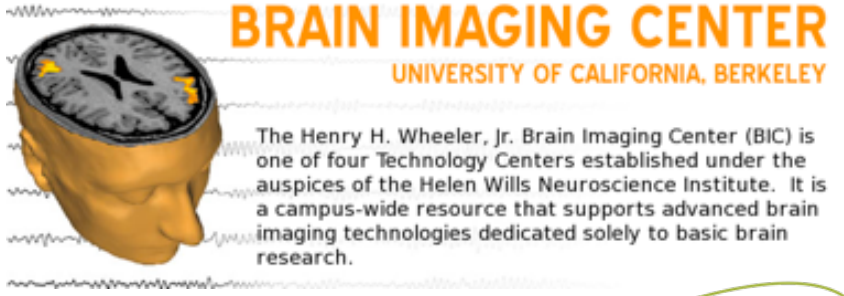
Physics PhD: Lattice QCD Simulations





Applied Math Postdoc: numerical algorithms



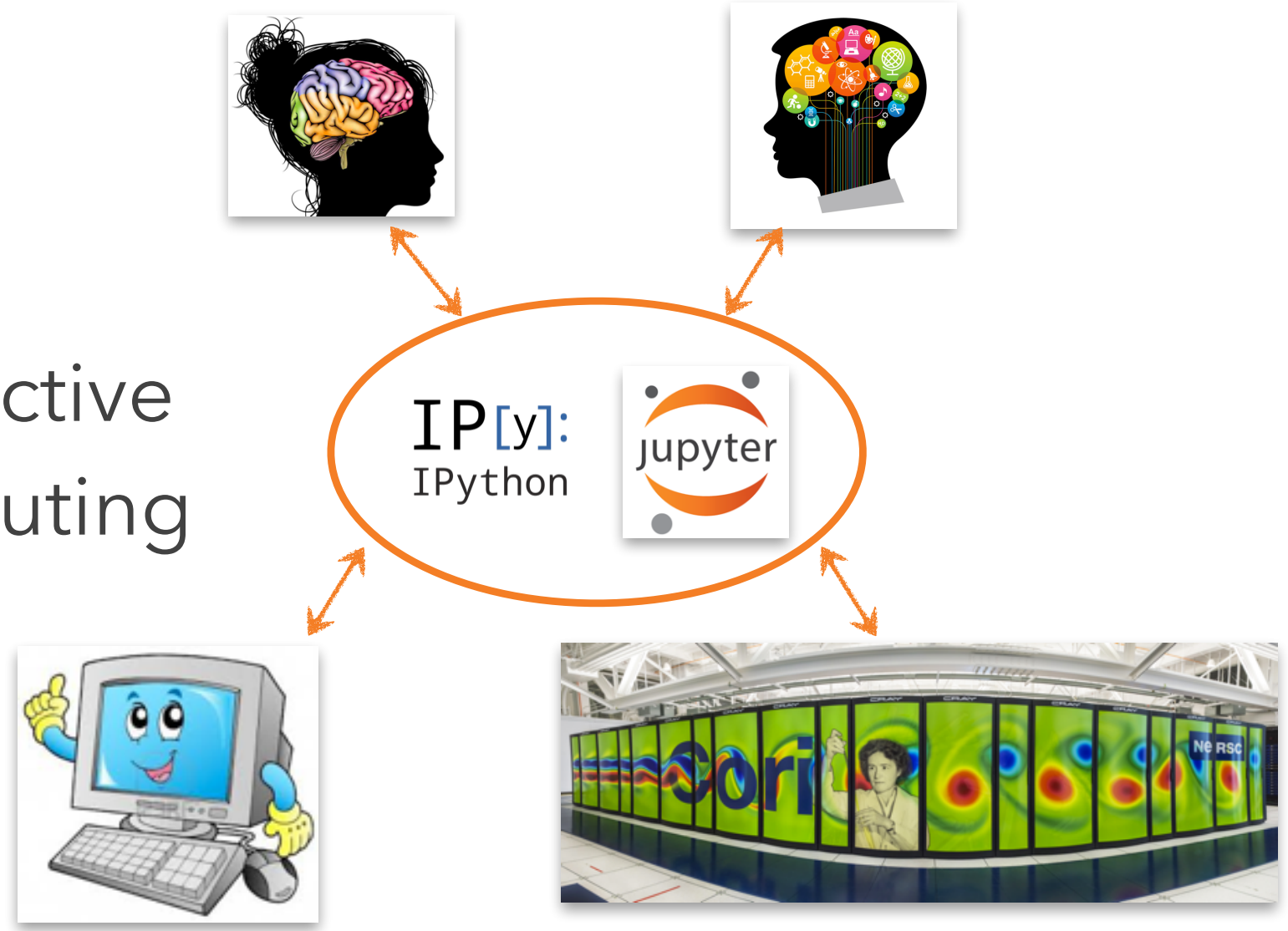


University of California, Berkeley
DEPARTMENT OF STATISTICS

“The purpose of computing is insight, not numbers”

–Hamming'62

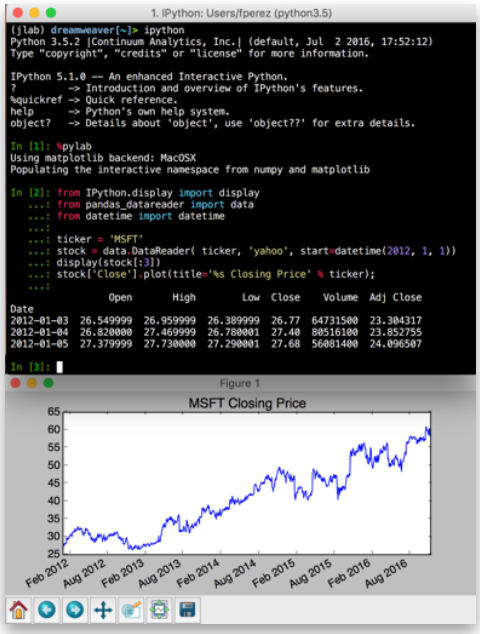
Interactive Computing

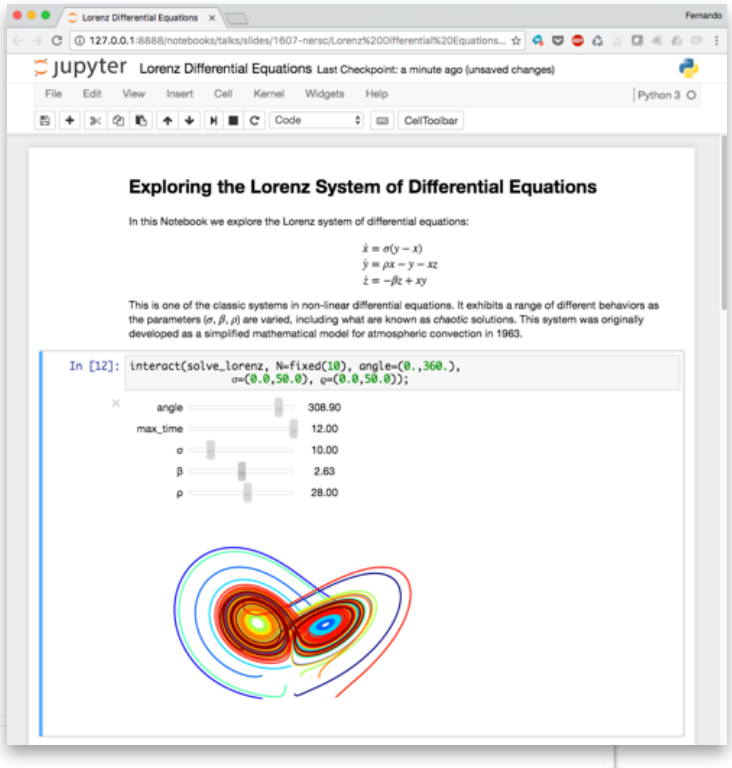


What is 

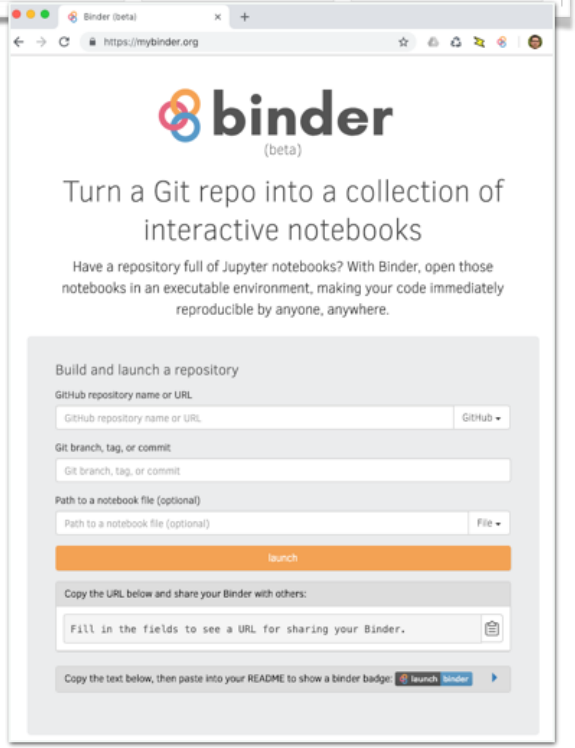
- Software
- Standards and Protocols
- Community

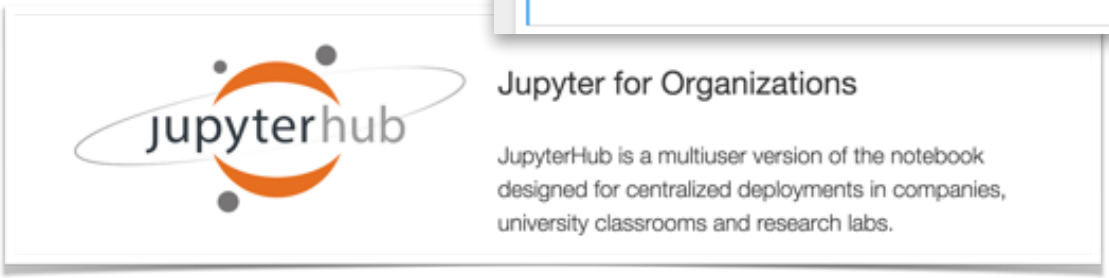
Software


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))
...: display(stock[:3])
...: 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
65
60
55
50
45
40
35
30
25
Feb 2012
Aug 2012
Feb 2013
Aug 2013
Feb 2014
Aug 2014
Feb 2015
Aug 2015
Feb 2016
Aug 2016

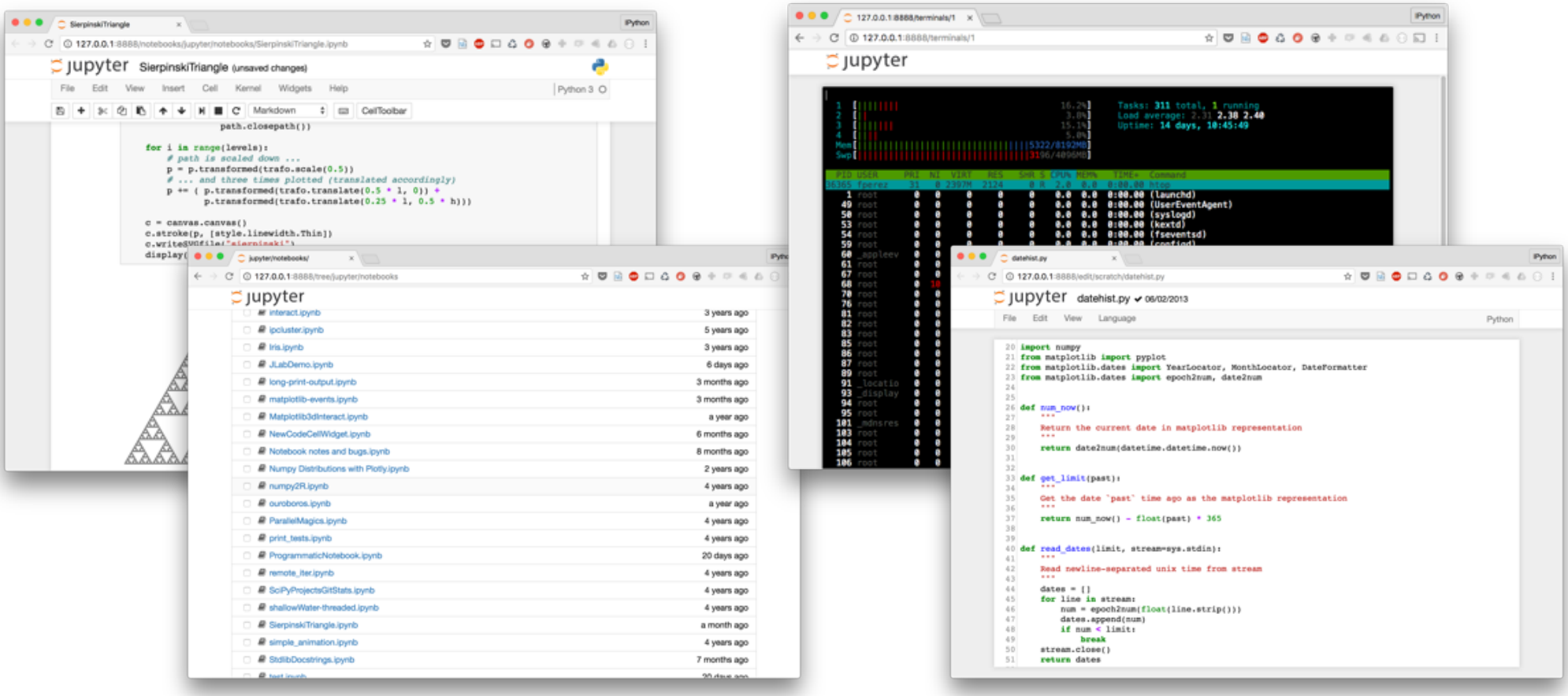

Lorenz Differential Equations
Fernando
127.0.0.1:8888/notebooks/talks/slides/1607-nersc/Lorenz%20Differential%20Equations...
jupyter Lorenz Differential Equations Last Checkpoint: a minute ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Python 3
Code
CellToolbar
Exploring the Lorenz System of Differential Equations
In this Notebook we explore the Lorenz system of differential equations:
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$
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, including what are known as chaotic solutions. This system was originally developed as a simplified mathematical model for atmospheric convection in 1963.
In [12]: interact(solve_lorenz, N=fixed(10), angle=(0.,360.),
σ=(0.0,50.0), ρ=(0.0,50.0));
angle 308.90
max_time 12.00
σ 10.00
β 2.63
ρ 28.00


jupyter
nbviewer
JUPYTER FAQ
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!
Programming Languages
IPython
IRuby
Julia
IP[y]: IPython
Interactive Computing
python
IRuby: Notebook
julia

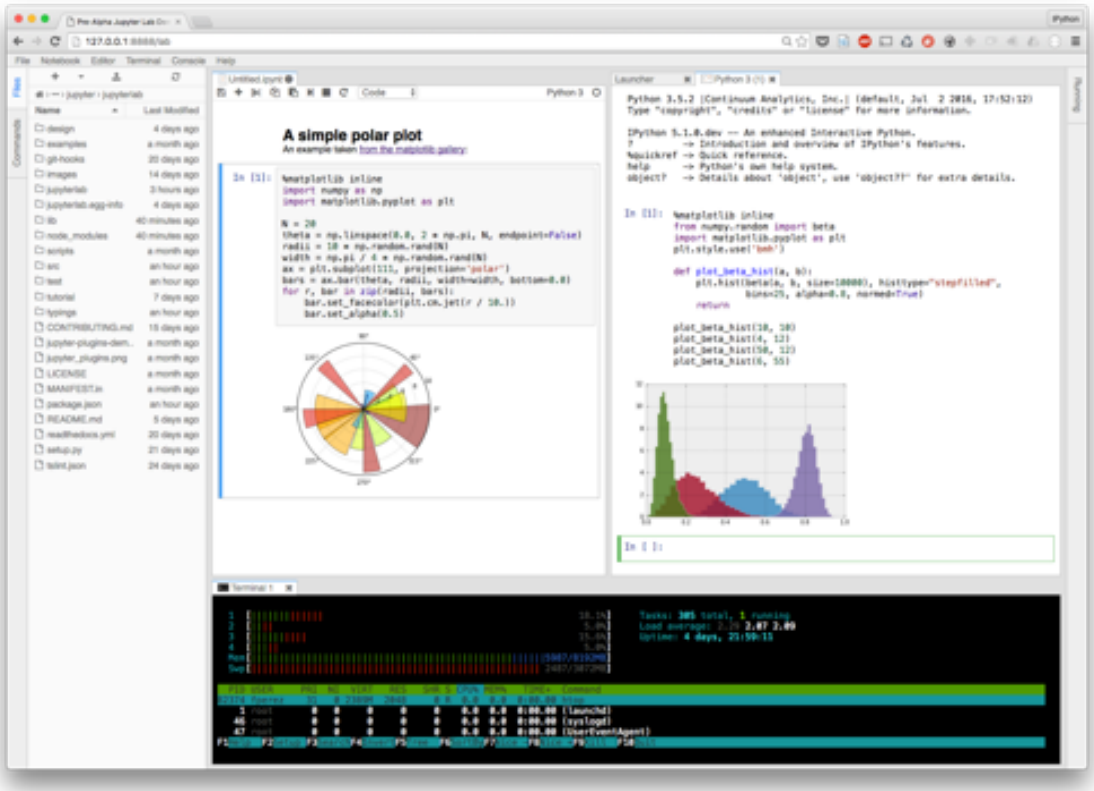

Binder (beta)
https://mybinder.org
binder
(beta)
Turn a Git repo into a collection of interactive notebooks
Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
Build and launch a repository
GitHub repository name or URL
GitHub repository name or URL
GitHub
Git branch, tag, or commit
Git branch, tag, or commit
Path to a notebook file (optional)
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

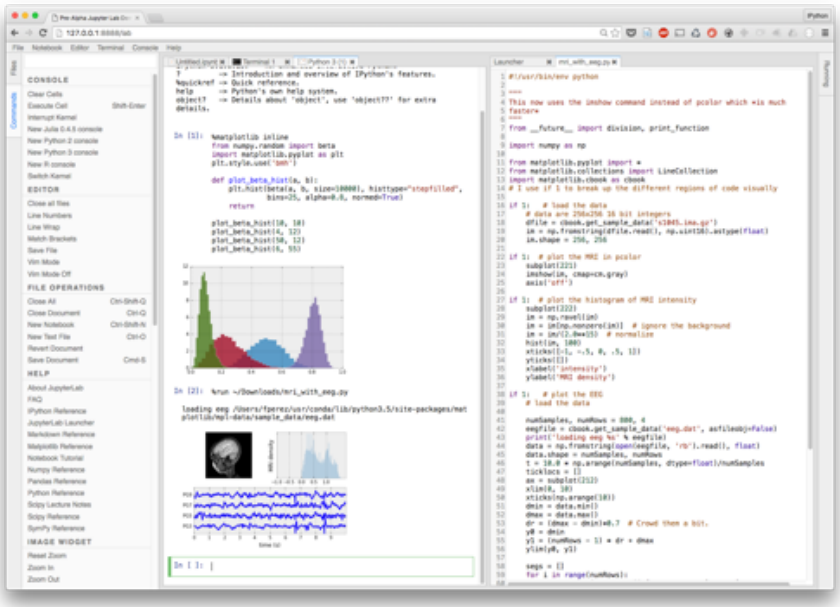

jupyterhub
Jupyter for Organizations
JupyterHub is a multiuser version of the notebook designed for centralized deployments in companies, university classrooms and research labs.

Classic ‘Notebook’…



JupyterLab: a grand unified theory of Jupyter





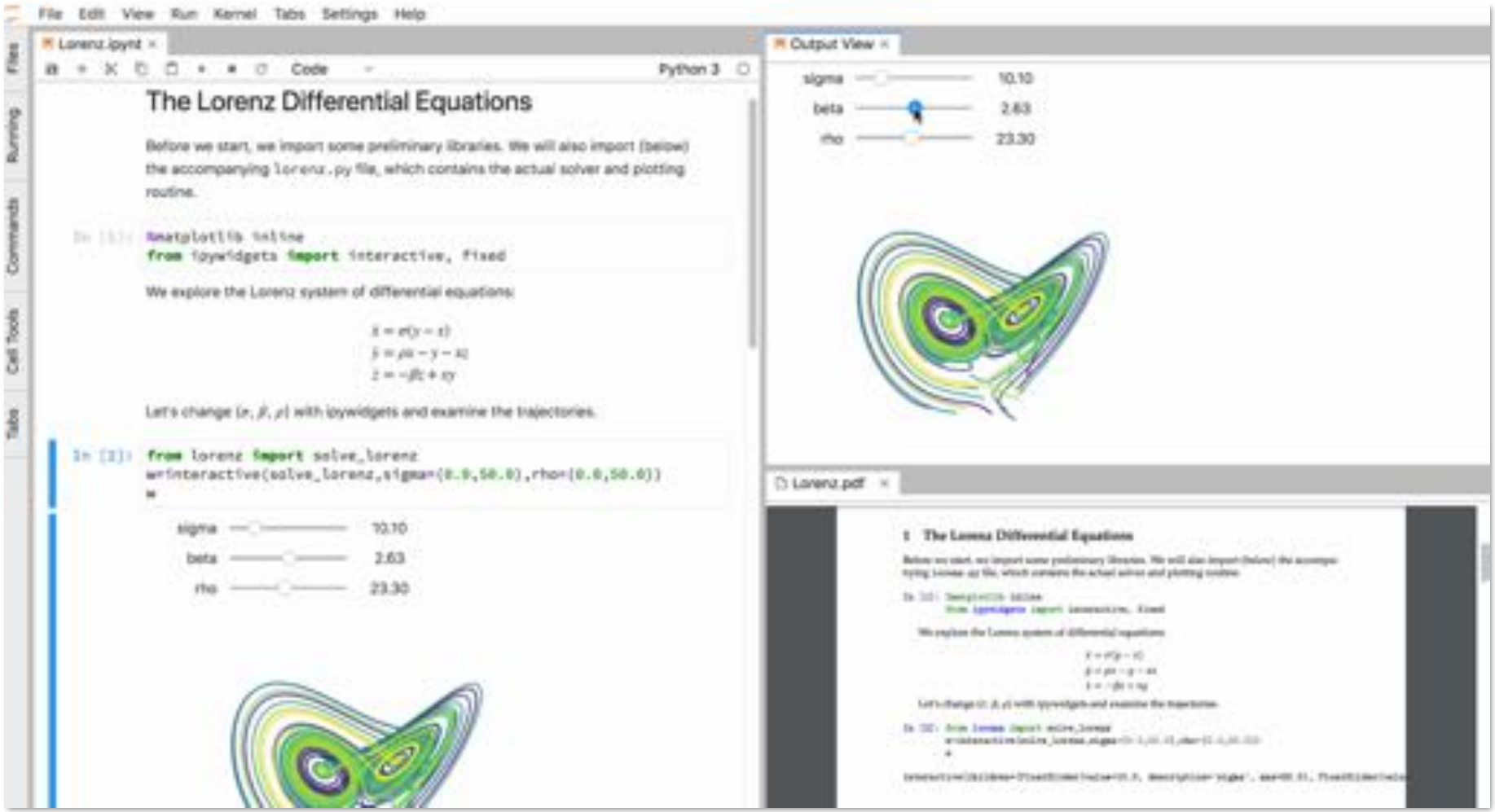
ANACONDA.

Bloomberg

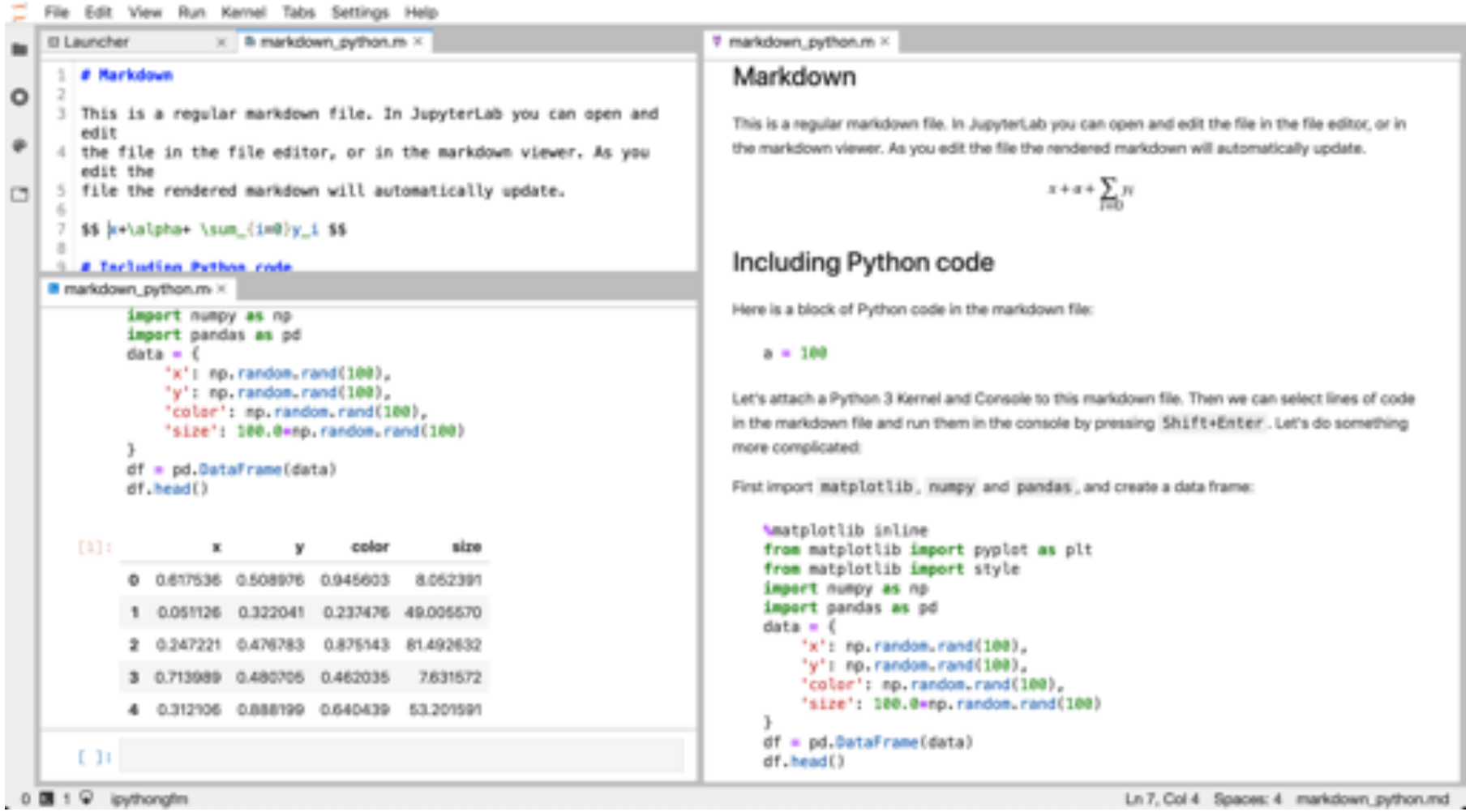
Huge Team Effort!

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

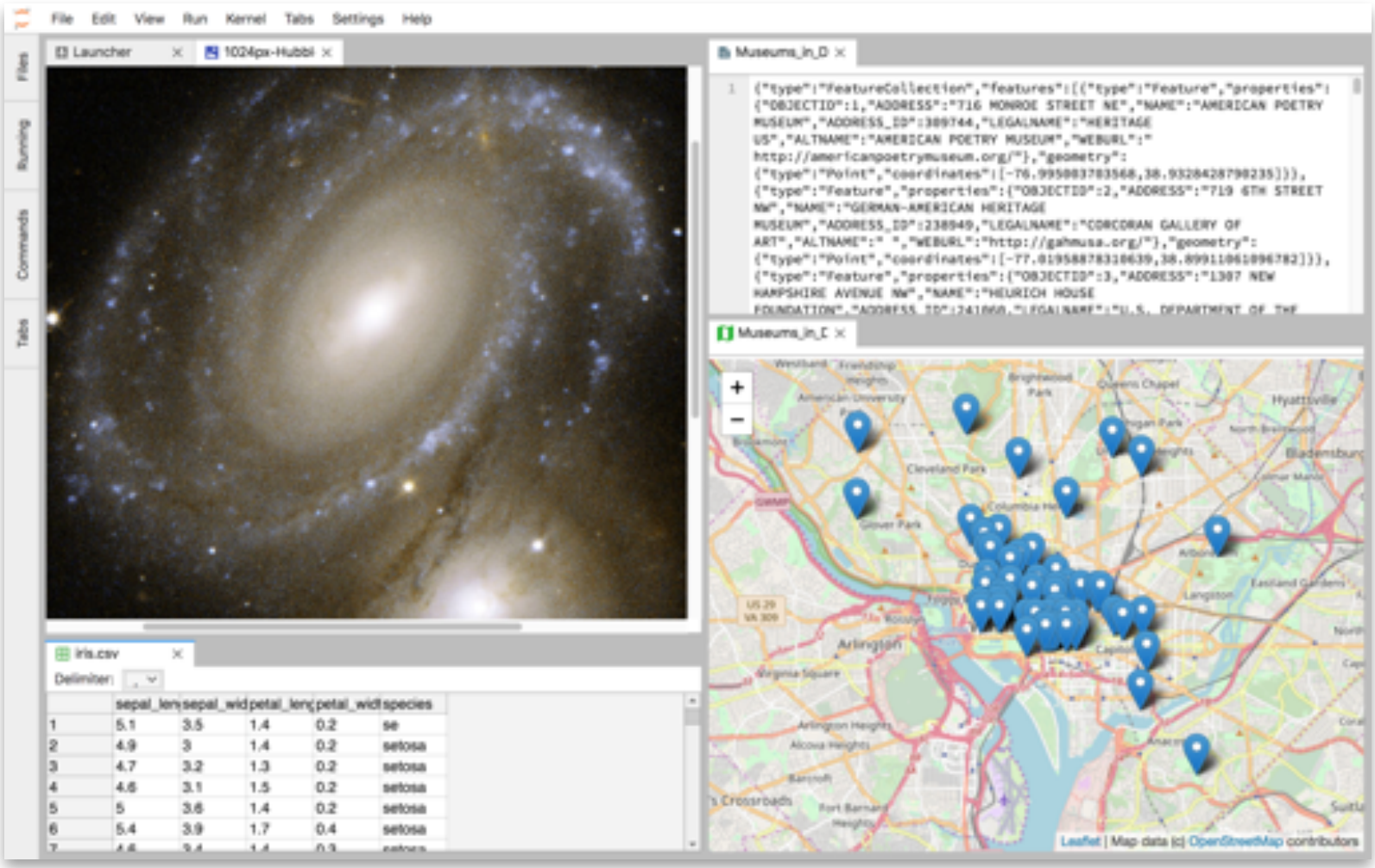
Notebooks++

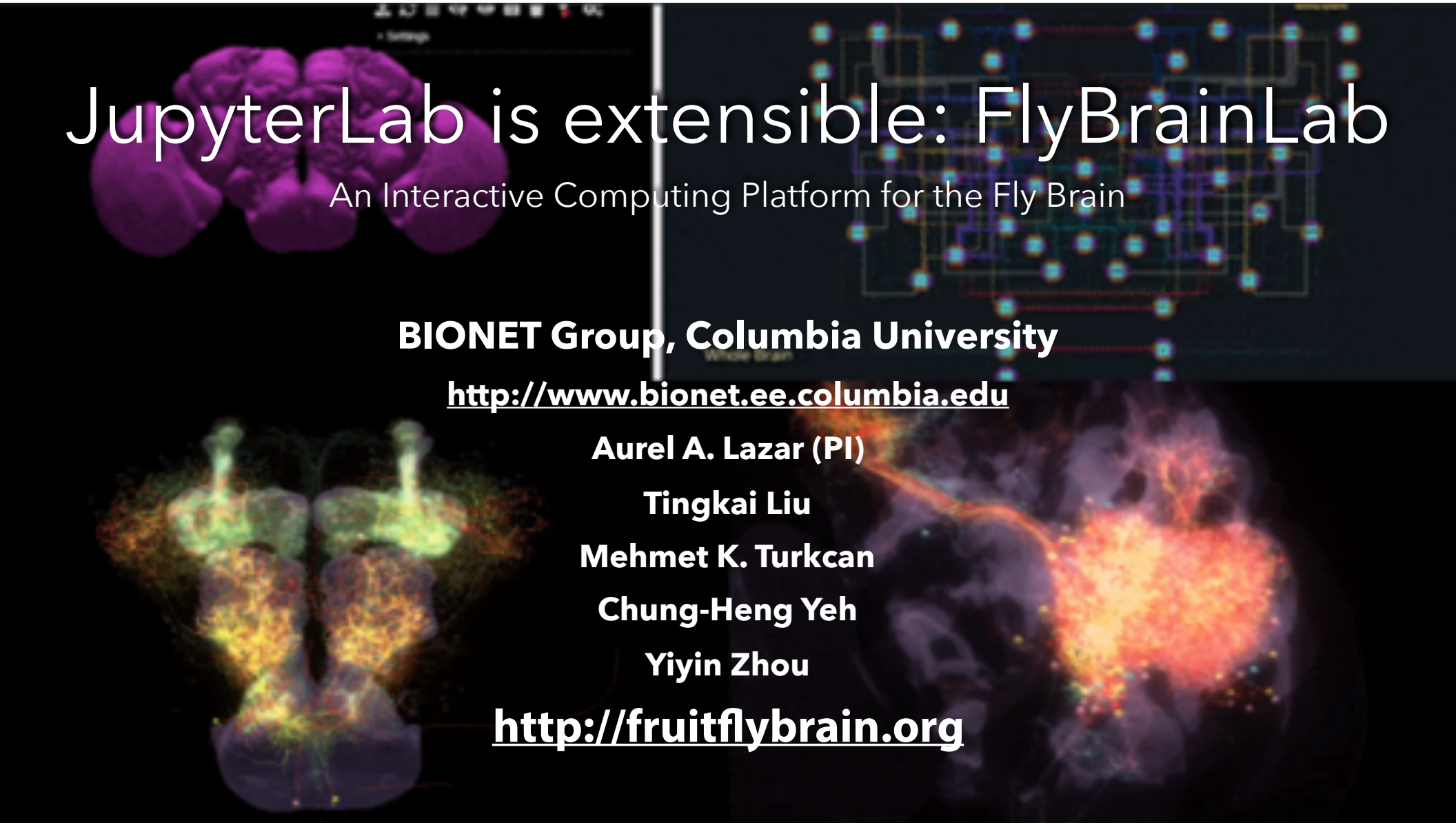

File Edit View Run Kernel Tabs Settings Help
Lorenz.ipynb
Code
Python 3
The Lorenz Differential Equations
Before we start, we import some preliminary libraries. We will also import (below) the accompanying lorenz.py file, which contains the actual solver and plotting routine.
%matplotlib inline
from ipywidgets import interactive, fixed
We explore the Lorenz system of differential equations:
Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.
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

Beyond notebooks



Data


File Edit View Run Kernel Tabs Settings Help
Launcher
1024px-Hubbl
Museums_in_D
{"type":"FeatureCollection","features":[{"type":"Feature","properties":
{"OBJECTID":1,"ADDRESS":"716 MONROE STREET NE","NAME":"AMERICAN POETRY
MUSEUM","ADDRESS_ID":389744,"LEGALNAME":"HERITAGE
US","ALTNAME":"AMERICAN POETRY MUSEUM","WEBURL":"
http://americanpoetrymuseum.org/"},"geometry":
{"type":"Point","coordinates":[-76.995083783568,38.93284287908235]}},
{"type":"Feature","properties":{"OBJECTID":2,"ADDRESS":"719 6TH STREET
NW","NAME":"GERMAN-AMERICAN HERITAGE
MUSEUM","ADDRESS_ID":238949,"LEGALNAME":"CORCORAN GALLERY OF
ART","ALTNAME":" ","WEBURL":"http://gahmusa.org/"},"geometry":
{"type":"Point","coordinates":[-77.01958878318639,38.899110610896782]}},
{"type":"Feature","properties":{"OBJECTID":3,"ADDRESS":"1307 NEW
HAMPSHIRE AVENUE NW","NAME":"HEURICH HOUSE
Museums_in_C
iris.csv
Delimiter:
sepal_len sepal_wid petal_len petal_wid species
1 5.1 3.5 1.4 0.2 se
2 4.9 3 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 3.6 1.4 0.2 setosa
6 5.4 3.9 1.7 0.4 setosa
Files
Running
Commands
Tabs
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)

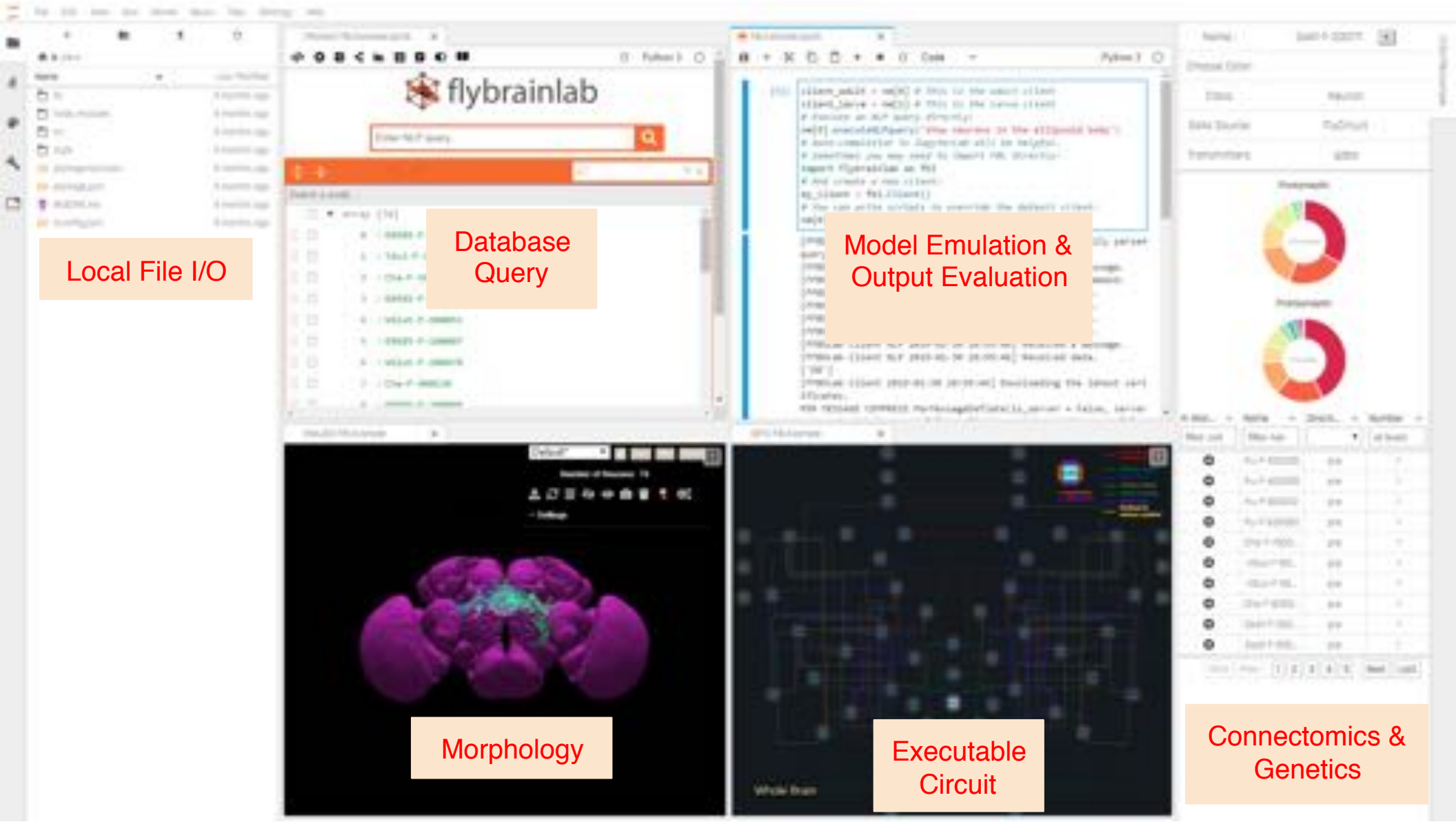
Tingkai Liu

Mehmet K. Turkcan

Chung-Heng Yeh

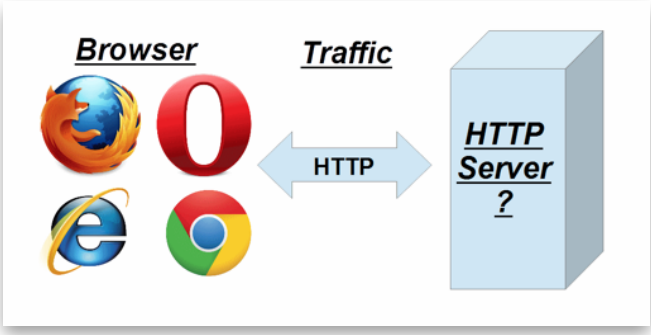
Yiyin Zhou

http://fruitflybrain.org


Local File I/O
flybrainlab
Database Query
Model Emulation & Output Evaluation
Morphology
Executable Circuit
Connectomics & Genetics

Standards and Protocols

Core ideas of the web: HTTP & HTML



HTTP: protocol to connect clients and servers

HyperText Transport Protocol

Image credit: eviltester.com

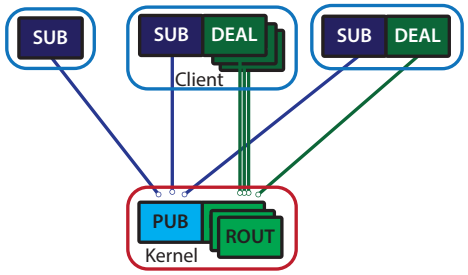
```
<div id="page-wrapper">
  <div id="main" class="clearfix">
    <div id="content" class="column" role="main"><div id="content-inner" class="clearfix">
                    <a id="main-content"></a>
                      <h1 class="title" id="page-title">Statistics at UC Berkeley</h1>
```

HTML: format to represent content

HyperText Markup Language

Core ideas of Jupyter

Interactive Computing Protocol



ØMQ + JSON

Document Format

We have already computed $P(X|A)$ above. On the other hand, $P(X| \sim A)$ is subjective: our code can pass tests but still have a bug in it, though the probability there is a bug present is reduced. Note this is dependent on the number of tests performed, the degree of complication in the tests, etc. Let's be conservative and assign $P(X| \sim A) = 0.5$. Then

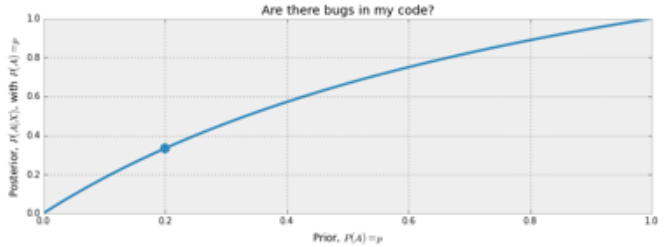
$$P(A|X) = \frac{1 \cdot p}{1 \cdot p + 0.5(1-p)}$$

$$= \frac{2p}{1+p}$$

This is the posterior probability. What does it look like as a function of our prior, $p \in [0, 1]$?

```
figsize(12.5, 4)
p = np.linspace(0, 1, 50)
plt.plot(p, 2 * p / (1 + p), color="#348ABD", lw=3)
# plt.fill_between(p, 2*p/(1+p), alpha=.5, facecolor=["#A60628"])
plt.scatter(0.2, 2 * (0.2) / 1.2, s=140, c="#348ABD")
plt.xlim(0, 1)
plt.ylim(0, 1)
plt.xlabel("Prior, $P(A) = p$")
plt.ylabel("Posterior, $P(A|X)$, with $P(A) = p$")
plt.title("Are there bugs in my code?")
```

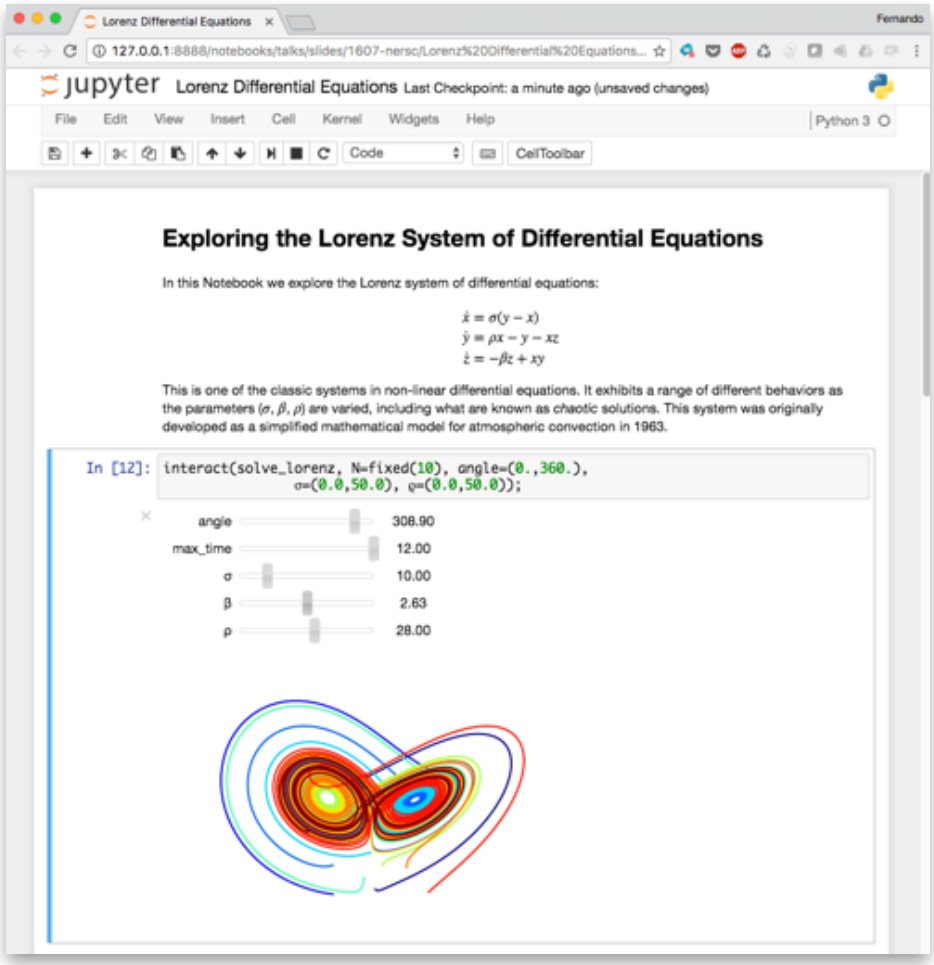
```
<matplotlib.text.Text at 0x1051de650>
```



https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

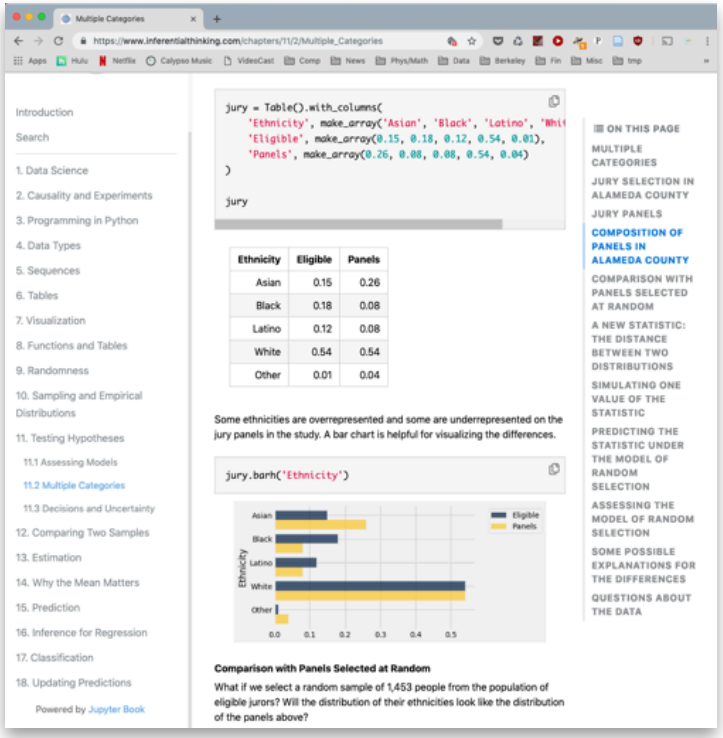
The Notebook as document & format

- JSON specification
- Machine readable
- Metadata-rich

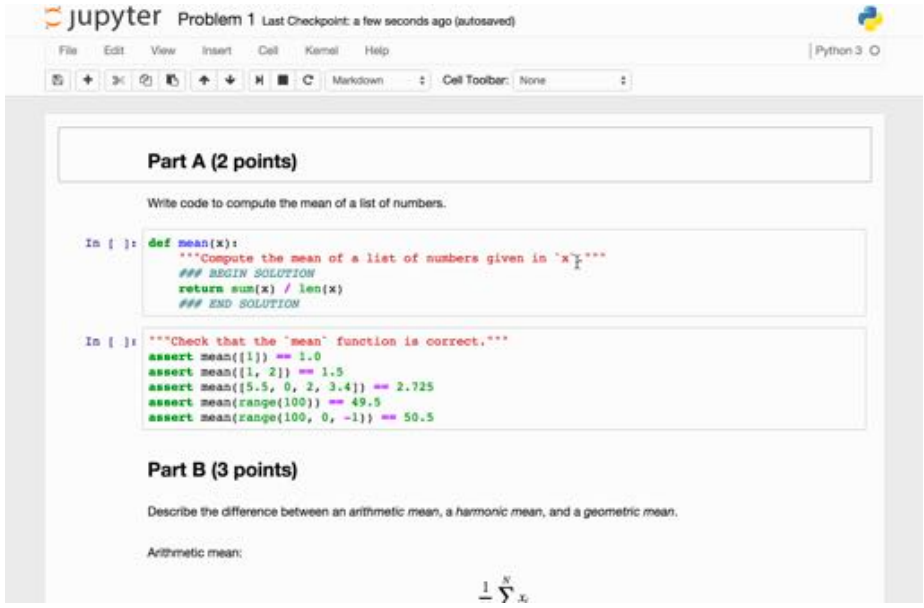


New tools built atop the format

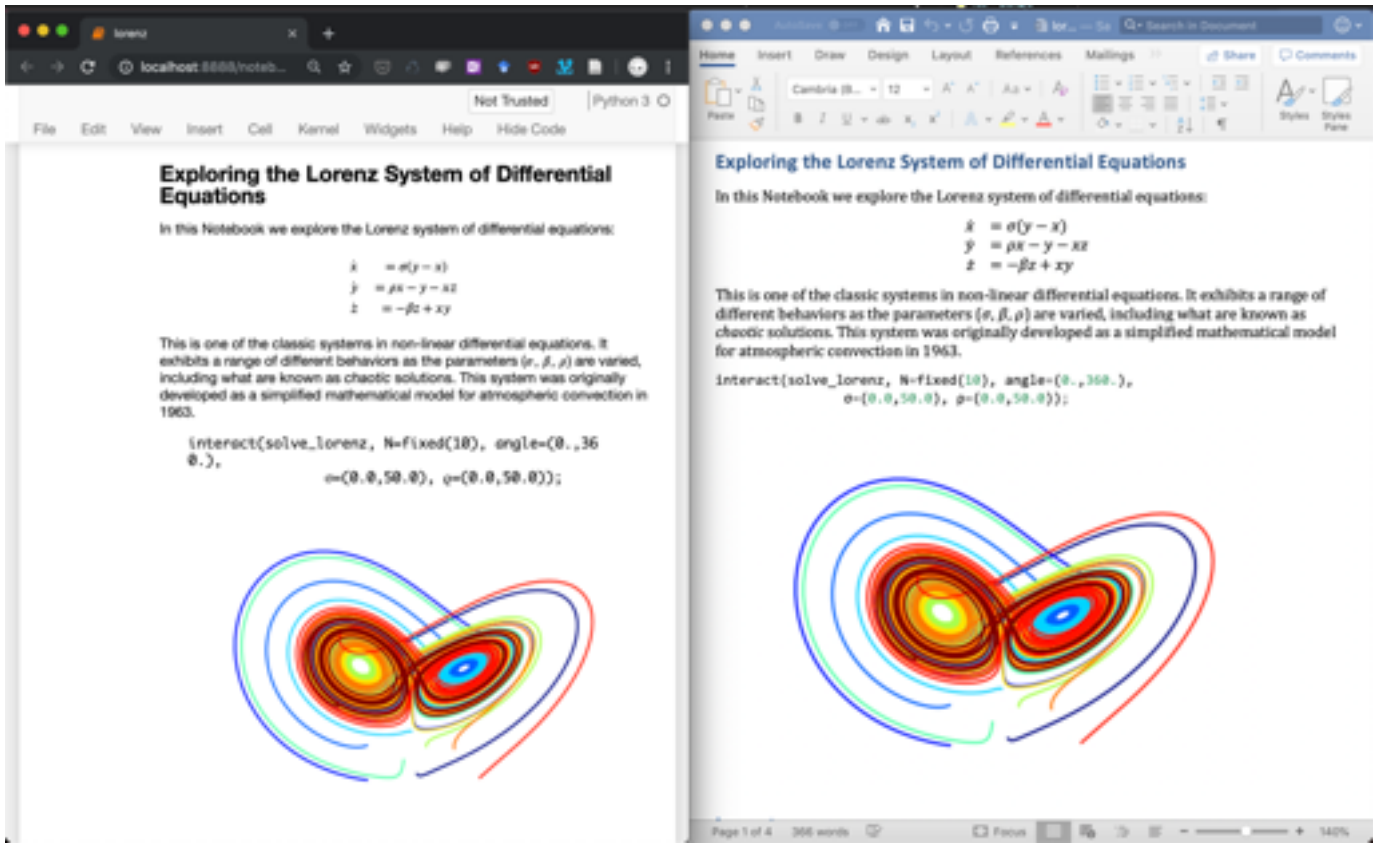
Jupyter Book



nbgrader: homework assignments



pandoc: ipynb to word (and more)



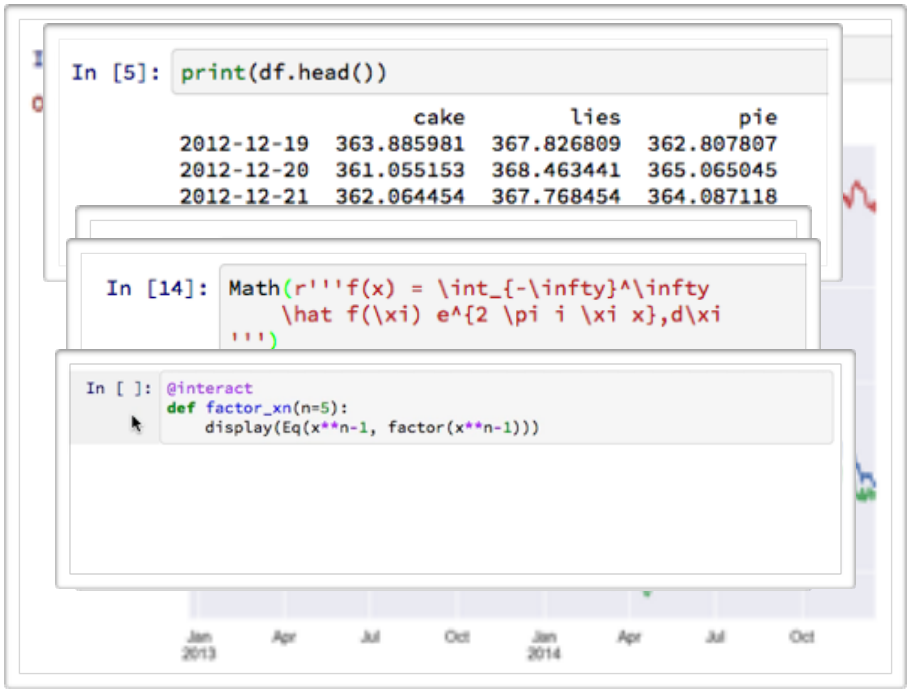

```
pandoc lorenz.ipynb -o lorenz.docx
```

Jupyter Protocol

web-age capture of the process of interactive computing

any mime-type output

- text
- svg, png, jpeg
- latex, pdf
- html, javascript
- interactive widgets

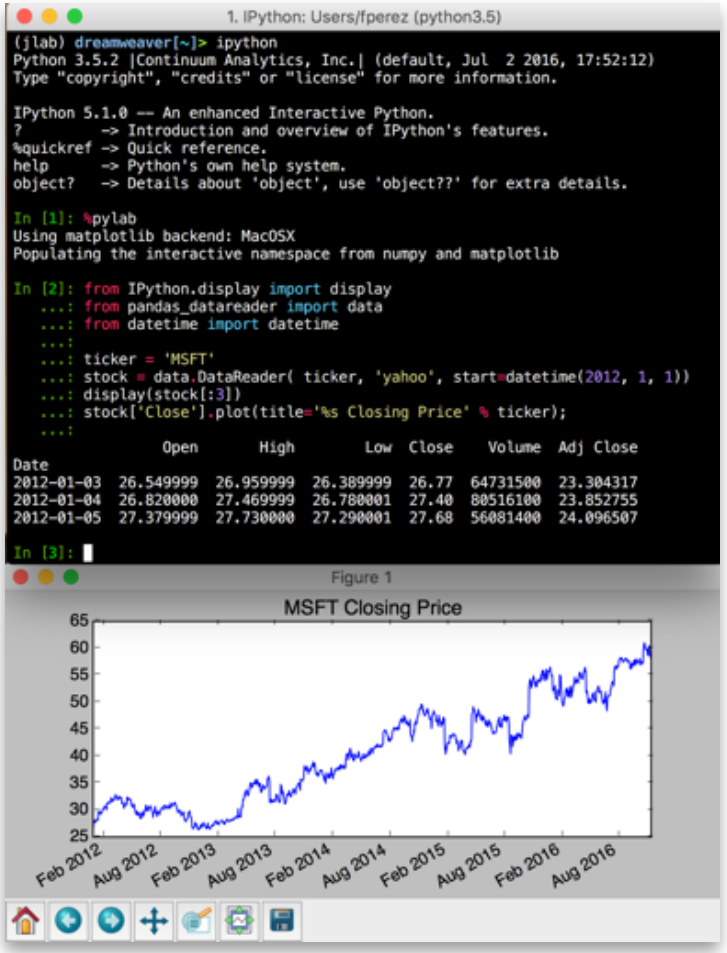


A language agnostic protocol

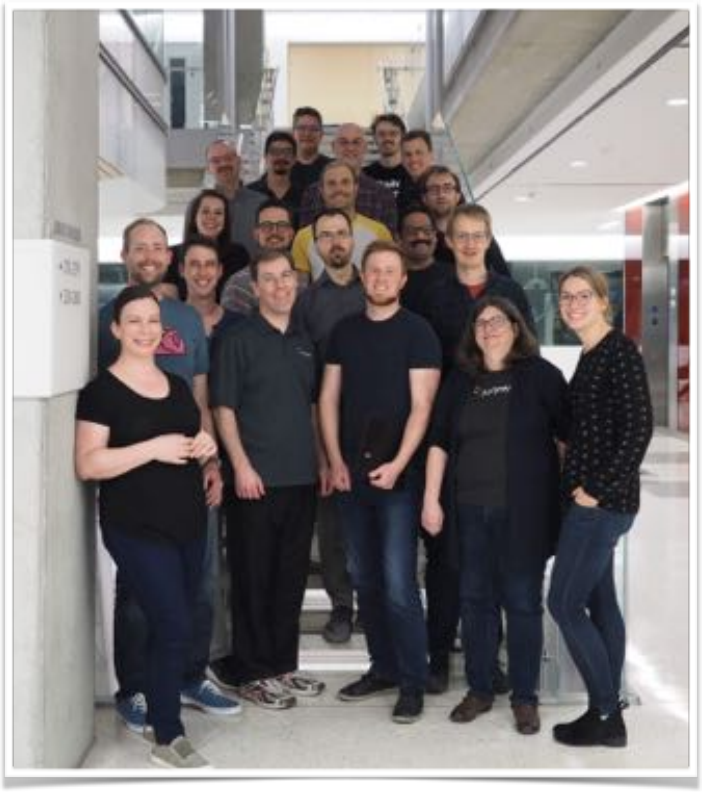
~100 different kernels: https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

Community

IPython: an afternoon hack, 2001



A true team effort



Plus ~ 1500 more Open source contributors!

Formalized governance

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

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

Microsoft

QUANSIGHT

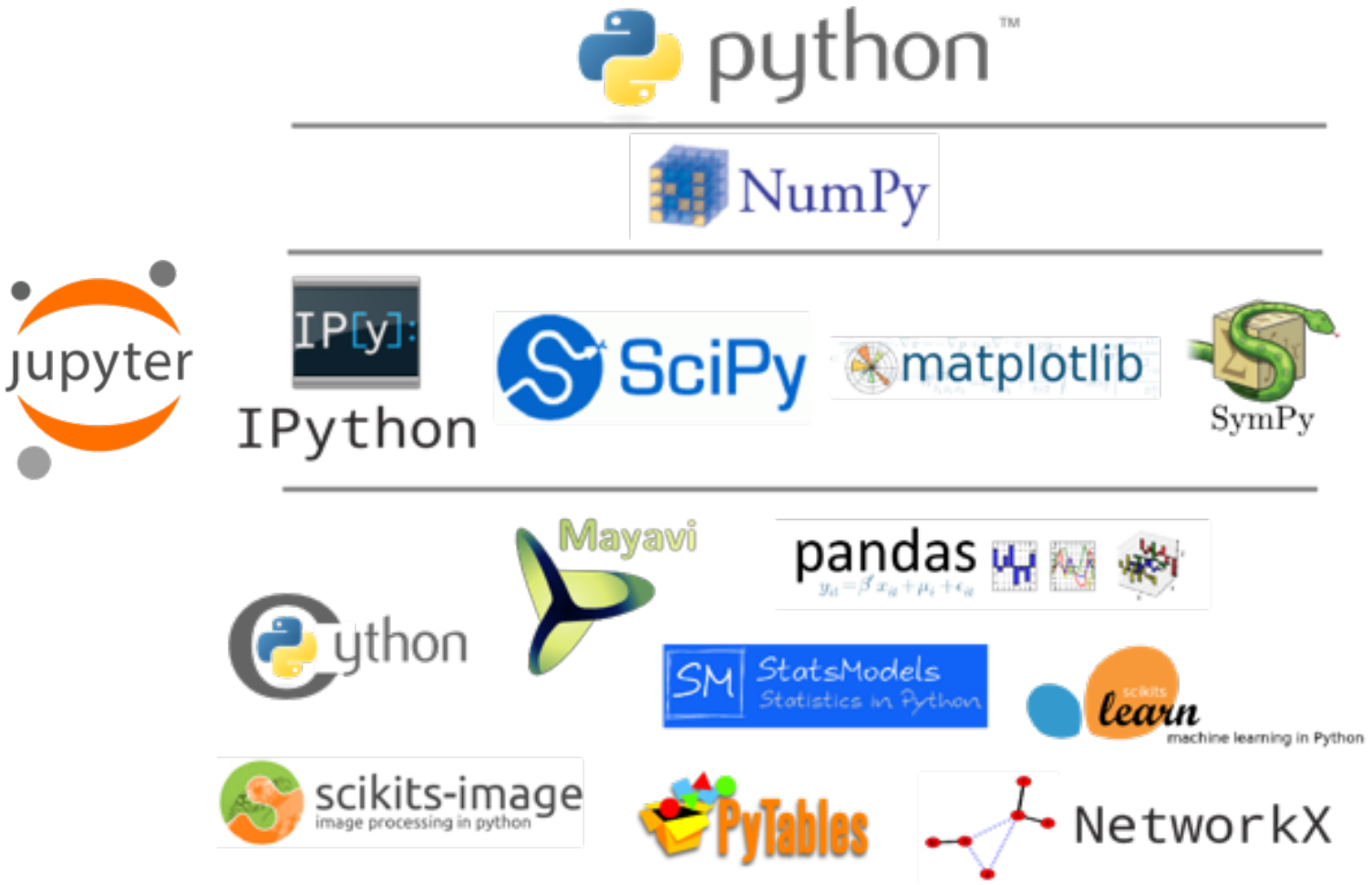
SCHMIDT FUTURES

Formal fiscal sponsorship

NUMFOCUS

OPEN CODE = BETTER SCIENCE

Jupyter: gateway to an ecosystem

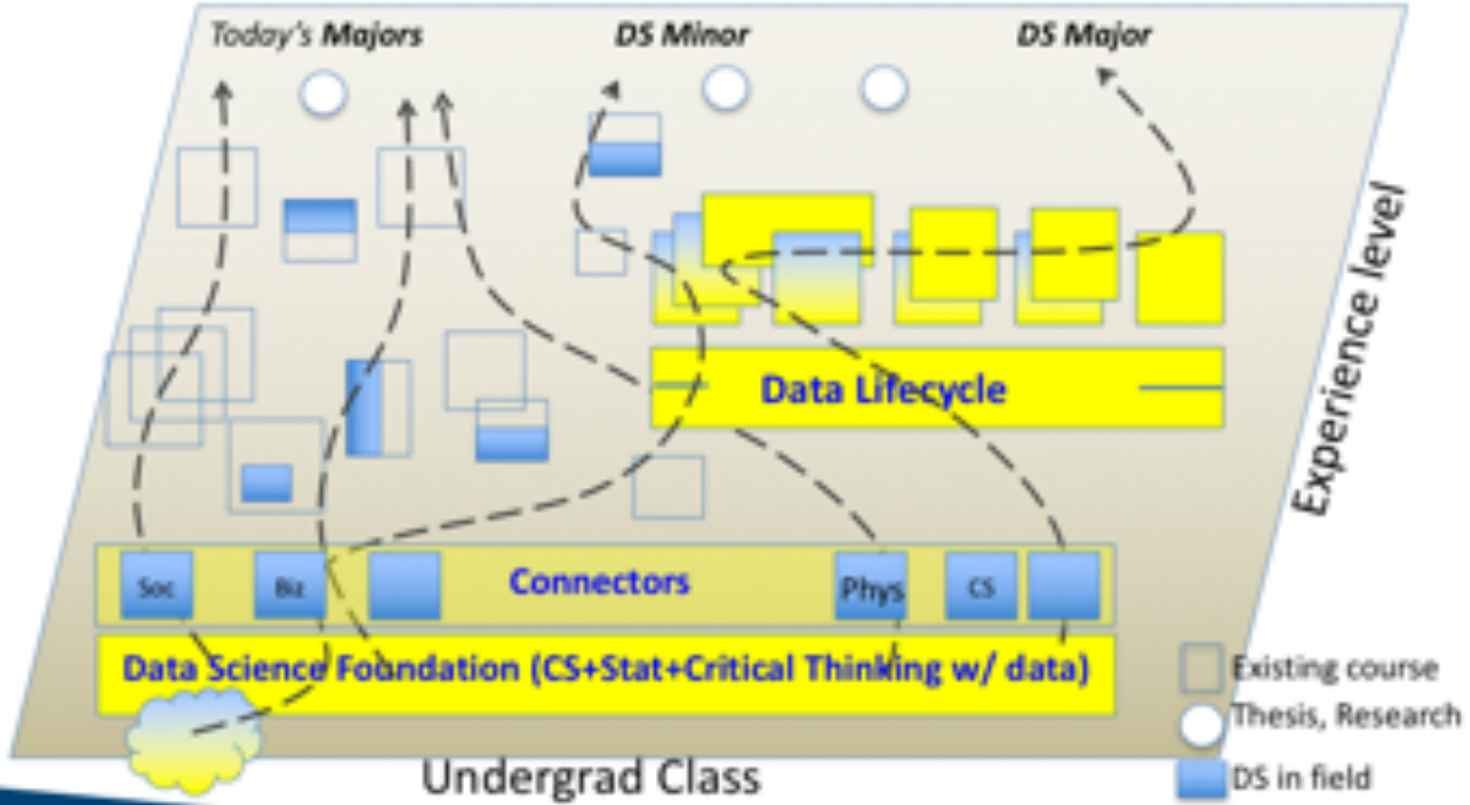


Educational impact:
a view from Berkeley

Chancellor's directive, June 2014

… rethinking at a fundamental level what **every educated person** must know about **quantitative reasoning**: how to effectively **understand, process and interpret information**, to inform decisions in their professional and personal lives and as citizens of the world in the 21st century.

Undergrad experience transformed

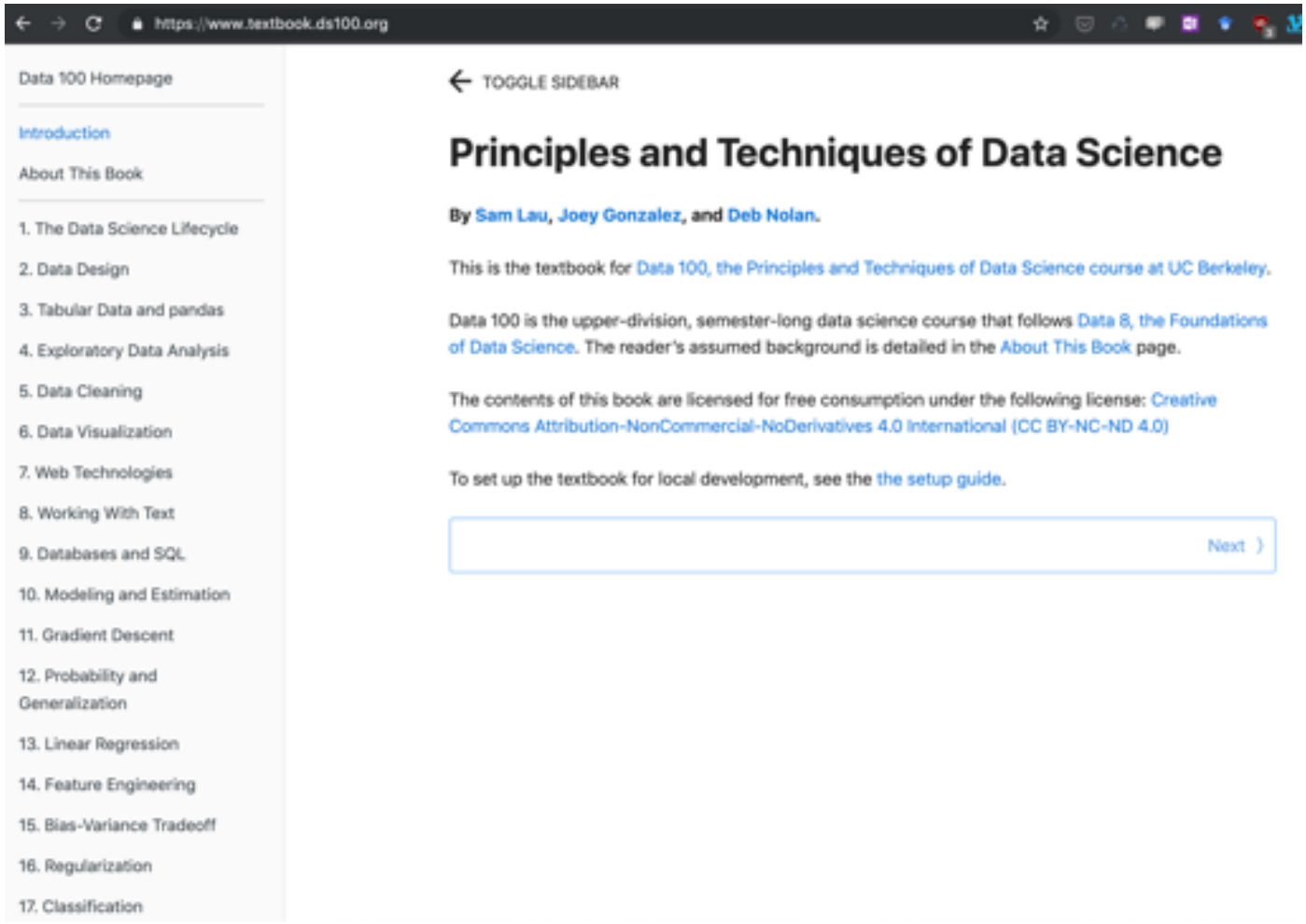


Data 8: Foundations of Data Science

- **Purpose:** "You will learn the core concepts of inference and computing, while working hands-on with real data."
- **A Data Science Course for Everyone:** "designed for entry-level students from any major, specifically for students who have not previously taken statistics or computer science courses."
- **Topics:** "all the key ideas of introductory statistics in a new, modern, hands-on way. It weaves in contextual issues like data privacy and bias. At the same time, it gives you a powerful understanding of key ideas in computing."

http://inferentialthinking.com

Data 100: ds100.org



Live textbooks

Visualizing Categorical Distributions

Interact

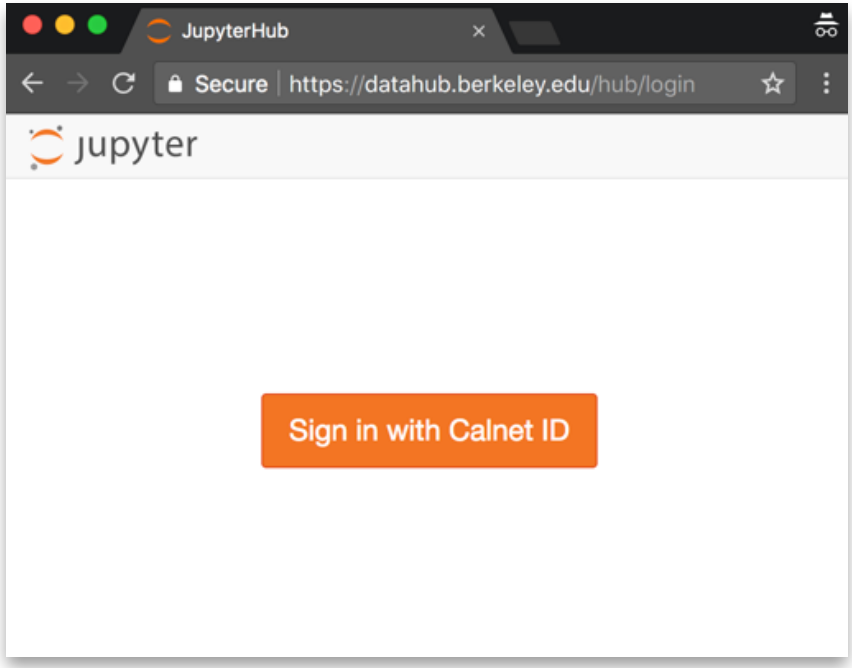
Data come in many forms that are not numerical. Data can be pieces of music, or places on a map. They can also be categories into which you can place individuals. Here are some examples of *categorical* variables.

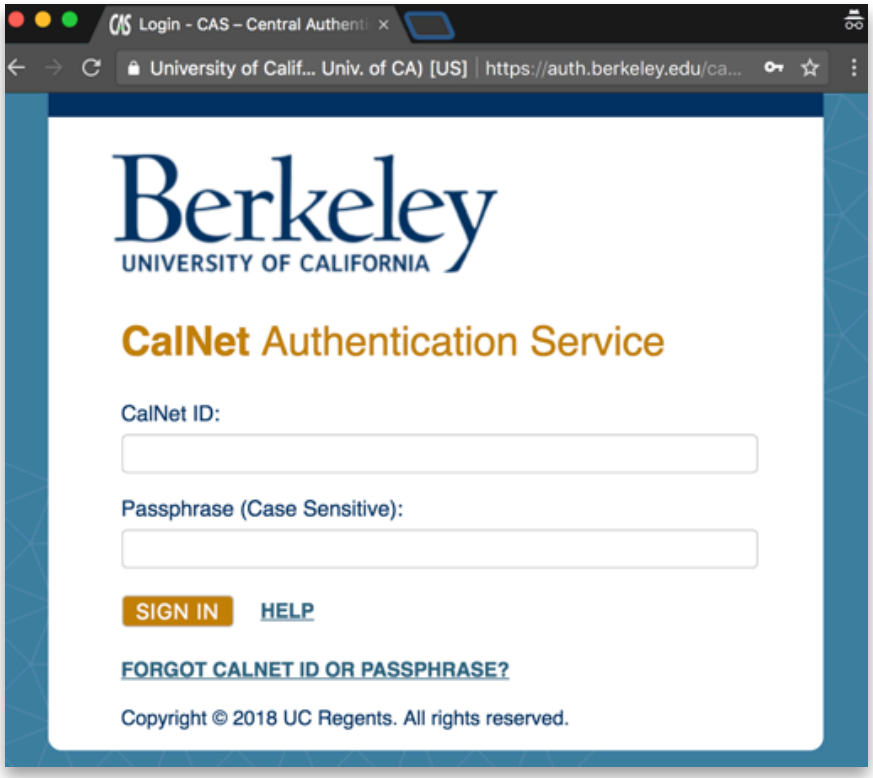
- The individuals are cartons of ice-cream, and the variable is the flavor in the carton.
- The individuals are professional basketball players, and the variable is the player's team.
- The individuals are years, and the variable is the genre of the highest grossing movie of the year.
- The individuals are survey respondents, and the variable is the response they choose from among "Not at all satisfied," "Somewhat satisfied," and "Very satisfied."

The table `icecream` contains data on 30 cartons of ice-cream.

```
icecream = Table().with_columns(
    'Flavor', make_array('Chocolate', 'Strawberry', 'Vanilla'),
    'Number of Cartons', make_array(16, 5, 9)
)
icecream
```

Berkeley: datahub.berkeley.edu





Fall 2018

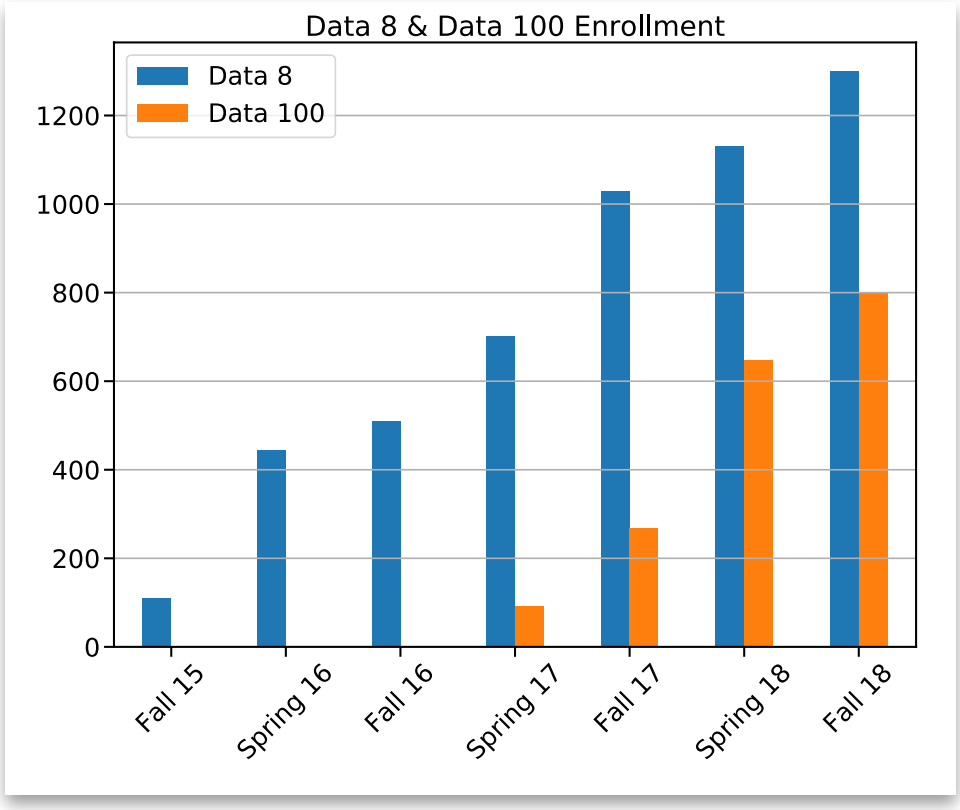
Data 8:

~1,300 students

Data 100:

~800 students

Fastest growing courses in Berkeley history



Data 8 in Fall 2018

- ~ 1,300 enrolled students
- ~ 200 waitlisted

Annual combined numbers

- Data 8: ~ 3,000 students
- UC Berkeley: ~ 7,500

At steady state, will easily reach ~50% of campus!

http://data8.org - http://ds100.org

LEGALST 88
Taking Measure of the Justice System

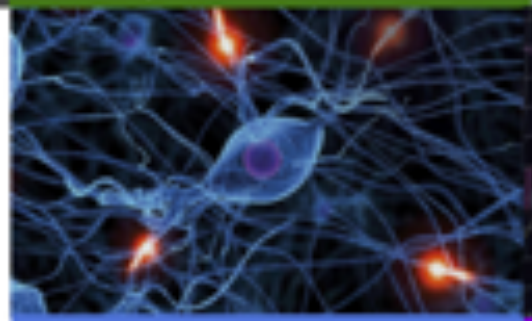
STAT 88
Probability & Mathematical Statistics

COMPSCI 88
Computational Structures in DS

COGSCI 88
Data Science & the Mind

DATA SCIENCE CONNECTOR COURSES

Companion Courses for Data 8



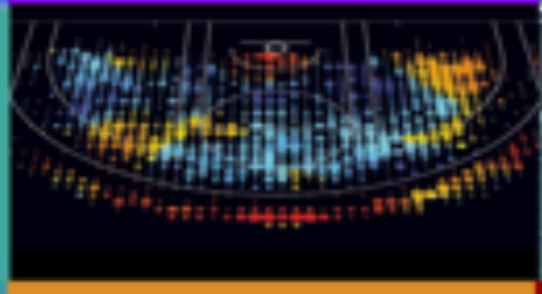
PSYCH 88
DS for Cognitive Neuroscience

UGBA 96-4 & 5
Data & Decisions

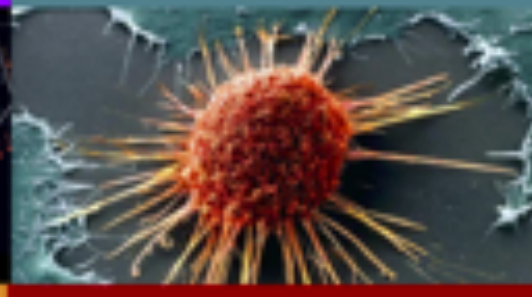
STAT 89A
Linear Algebra for Data Science

GEOG 88
Data Science Applications in Geography

L&S 88-1
Children in the Developing World



L&S 88-2
Sports Analytics



MCB 88
Immunotherapy of Cancer

Modules

Fall 2017 Data Science Modules

INFO 290
Privacy & Security Lab

PSYCH 192
Special Topics

EDUC 223B
Special Problems in STEM Education

ART W23AC
Data Arts

IAS 150
International Area Studies

ETHSTD 21AC
Immigration & Detention

SOCIOL 130AC
Social Inequalities

XRHETOR R1A
The Craft of Writing

EPS C20
Earthquakes in Your Backyard

GWS 131
Gender and Science

ENGLISH 31AC
Literature of American Cultures

COGSCI 1B
Intro to Cognitive Science

ISF 189
Intro to Interdisciplinary Studies

NESTUD 190A
Text & Web Analysis for Social Sciences

LINGUIS 110
Intro to Phonetics & Phonology

GWS 101
Gender and Women's Studies

Repro. research course: STAT 159/259

- **Schedule:** 2 Lectures (80 min), 1 lab (2 h)
- **Prerequisites:** foundations in computation, probability and statistical modeling
- **Enrollment:** 50 undergrads, 10 grads
- **Graduate Student Instructor:** Eli Ben-Michael, Stats PhD student
- **Grading:** weekly readings, quizzes, homework and 3 projects.

http://bit.ly/stat159-f17

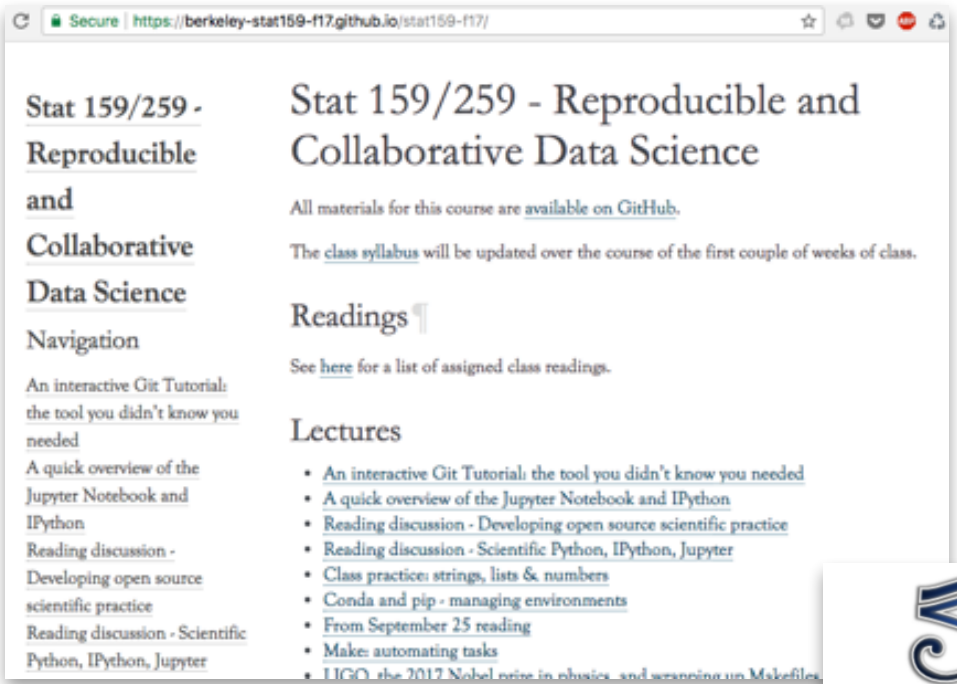
Goals

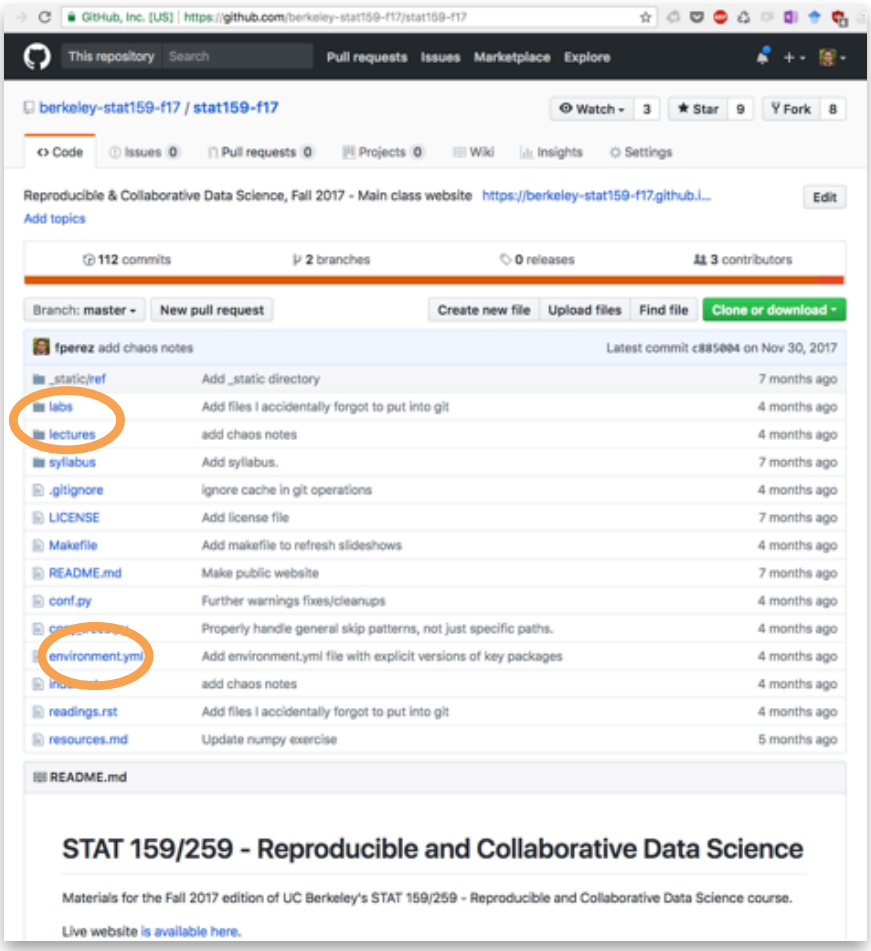
- **What?**
  - Core ideas: data access, computation, statistical analysis and publication.
- **Why?**
  - An essential concern of modern computational research.
  - Social and scientific implications of lack of reproducibility.
  - Frame the problem in terms practical, ethical and epistemological.
- **How?**
  - Skills and habits necessary to make a practice of reproducibility.
  - An everyday practice, not a "publication time" concept.

Core skills

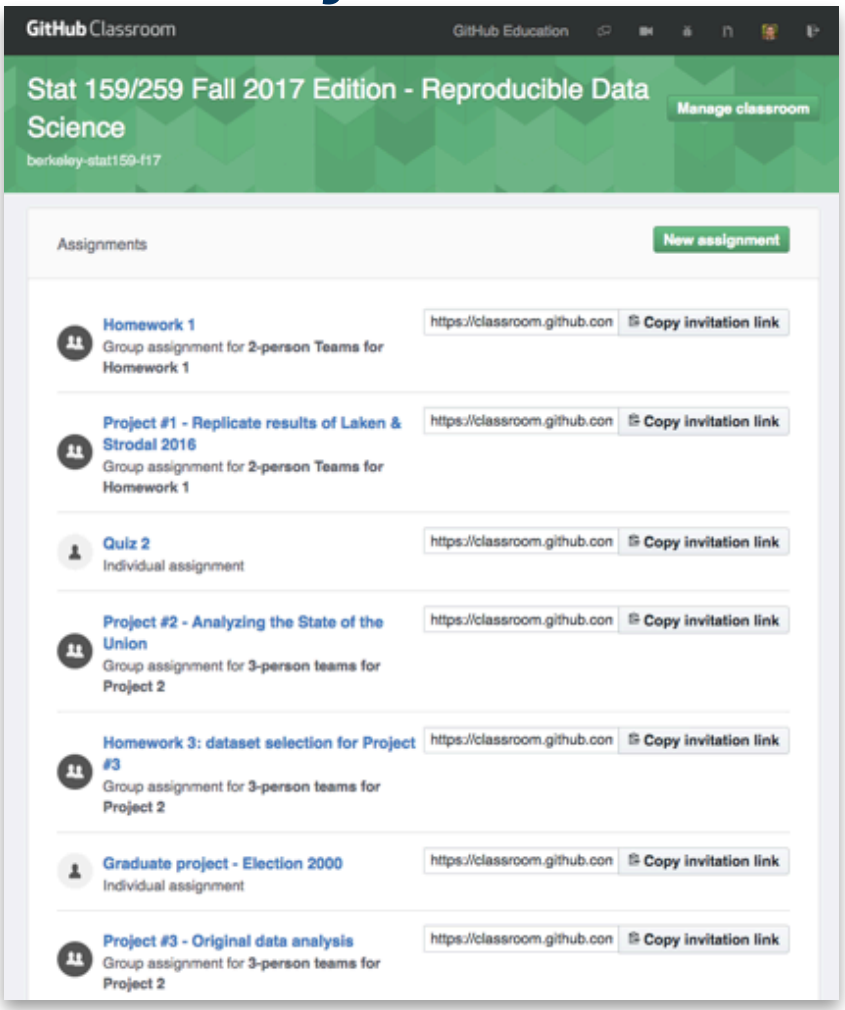
- **Version control:** Git and GitHub
- **Programming:** Python
- **Process automation:** Make
- **Data analysis:** Numpy, Pandas, Matplotlib, NLTK, Scikit-Learn, …
- **Documentation:** Sphinx
- **Software testing:** PyTest
- **Continuous Integration:** Travis
- **Reproducible containers:** Binder

Git and Python workflow everywhere





Computational hygiene: a daily habit



Final Project: an original analysis

- **Data:** included in repo or linked if too large.
- **Clean, tested code.**
- **Analysis notebooks and supporting code**
  - Break down your analysis into as many notebooks as is reasonable for convenient reading and execution.
- **Main narrative notebook:** summarizes and discusses results.
- **Reproducibility support:** `Makefile` and `environment.yml`
- **Good repository practices:** `README.md`, `LICENSE`, `.gitignore`.
  - Use Victoria Stodden's ENABLING REPRODUCIBLE RESEARCH: LICENSING SCIENTIFIC INNOVATION.

A "Standard Playbook"

Brief Analysis on the Marginal Effects of Studying

As students, we have often wondered what effect an extra hour of studying will have on our grades. When trying to determine whether staying up an extra hour to study for that final exam is truly worth it, we usually are limited by imperfect information and our own superstitions. In this project, we attempt to estimate the "true" marginal effect of studying on students' grades. We try to model the effects of studying first using OLS and then various instruments and 2 stage least squares. This repository is also meant to serve as an example of what a reproducible econometric analysis would look like.

Required Installations

The only installation needed to run this repo is Anaconda. Click here to learn about how to install Anaconda. Once installed, you should be good to go!

Using Binder

We've enabled Binder for this project which allows you to view jupyter notebooks in an executable environment. Feel free to click the link at the top of this README to launch the binder.

Getting Started

Download the repo onto your local machine and open your command prompt. Simply type in the following commands to run the analysis:

```
make clean
make env
source activate study
make run
```

After all your notebooks have run you should see new files in the results, fig, and data directories. Read about our approach and results in main.ipynb. All the figures from our analysis are saved in the fig directory and our regressions are saved in the results directory as dataframes. You can load in these dataframes and work with them as regression instances (i.e. you can call `.summary()`, `.params()` etc. click here for OLS documentation and here for 2SLS documentation)

Licensing

In an effort to enable reproducible, collaborative reserach our project is subject to the MIT License which allows you to modify and distribute the above code for both private and commercial usage. See LICENSE to learn more.

nadavtadelis Merge pull request #27 from berkeley-stat159-f17/nadav_actual_final ... Latest

data	added reproducibility aspects, split model fitting into 2 ntbks; NOTE...
fig	Fix typos in data_exploration.ipynb
results	Fix typos in model_fitting_2.ipynb
.gitignore	Add caches to .gitignore
.mailmap	adding mailmap to account for config issues
.travis.yml	Add pandas install to Travis
LICENSE	Added LICENSE
Makefile	added reproducibility aspects, split model fitting into 2 ntbks; NOTE...
README.md	added reproducibility sentence to README
data_exploration.ipynb	Minor add to data_exploration.ipynb
environment.yml	added reproducibility aspects, split model fitting into 2 ntbks; NOTE...
instructions.md	Add note about grades in team work
main.ipynb	correction to instruments justification
model_fitting_1.ipynb	Fix typos in model_fitting_1.ipynb
model_fitting_2.ipynb	Fix typos in model_fitting_2.ipynb
p3functions.py	Add two_way function
tests.py	Add two_way function

Analysis notebooks

While these histograms give us some information about the distributions of these individual variables, they don't help with understanding how these variables interact with our dependent variable G3. So lets look at some violin plots to visualize some of these interactions.

For the violin plots we split G3 into 5 bins to more clearly visualize the interactions. We also show the distributions relative to which school the students come from to determine whether there is a difference in the two schools.

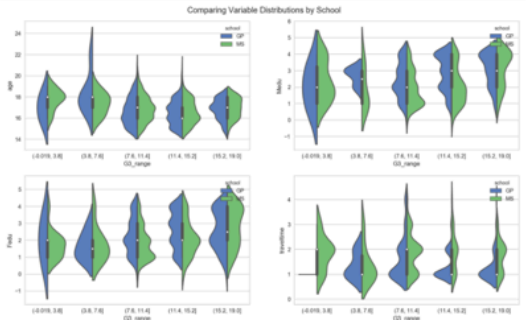
In [6]:
```
# Splitting G3 into ranges to get a cleaner visual
student_perf['G3_range'] = pd.cut(student_perf.G3, 5, retbins = True)[0]

# Creating the plots
plt.figure(figsize=(16, 36))
sns.set(style="whitegrid", palette="muted", color_codes=True)

plt.subplots_adjust(top=0.97)
plt.suptitle('Comparing Variable Distributions by School')

sns.despine()
for column_index, column in enumerate(['age', 'Medu', 'Fedu', 'traveltime', 'studytime',
'freetime', 'failures',
                                       'absences', 'famrel', 'goout', 'Dalc', 'Walc', 'he
alth', 'G1', 'G2']):
    if column == 'G3_range':
        continue
    plt.subplot(8, 2, column_index + 1)
    sns.violinplot(x='G3_range', y=column, hue = 'school', split = True, data=student_per
f)

plt.savefig('fig/distrbyschool.png');
```



Code and tests

Branch: master · project-3-p2-ka-jo-ta / p3functions.py

s-johnson Add two_way function

1 contributor

41 lines (34 sloc) 1.38 KB Raw

```
import pandas as pd
import numpy as np

def make_indicators(df, names):
    """Make indicator columns in dataframe df of whether existing columns are
    equal to given values.

    Args:
        df (pandas.DataFrame): Dataframe to be modified.
        names (dict)         : Dictionary containing:
                                   - Keys: Desired indicator column names
                                   - Values: Two item tuple containing:
                                       - Original dataframe column
                                       - Value to compare to column

    Returns:
        void: Dataframe df is modified in place.
    """
    for k,v in names.items():
        df[k] = 1*(df[v[0]] == v[1])
```

Branch: master · project-3-p2-ka-jo-ta / tests.py

s-johnson Add two_way function

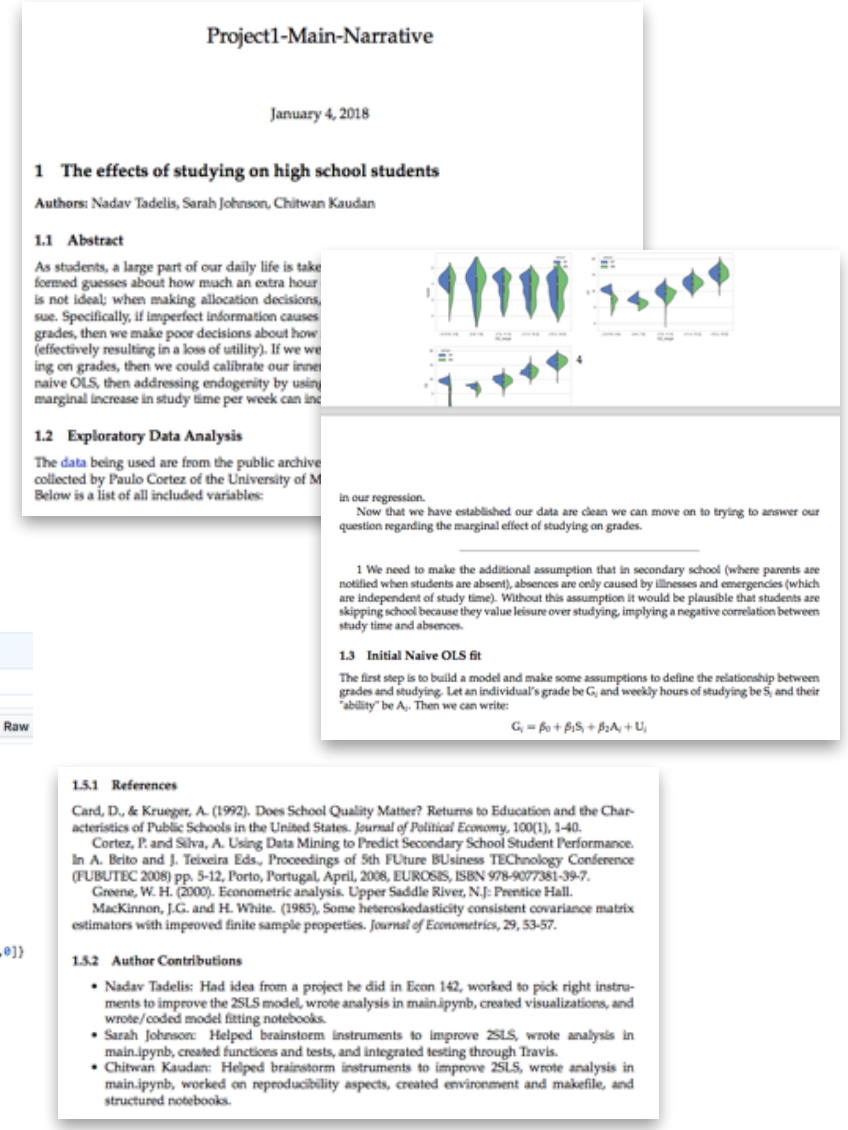
1 contributor

31 lines (24 sloc) 906 Bytes Raw

```
import pandas as pd
import numpy as np
import numpy.testing as npt

from p3functions import *


def test_make_indicators():
        d = {'col1': [1, 2], 'col2': [3, 4]}
        df = pd.DataFrame(data=d)
        names = {'ind1': ('col1', 2), 'ind2': ('col2', 3)}
        make_indicators(df,names)
        exp_d = {'col1': [1, 2], 'col2': [3, 4], 'ind1': [0, 1], 'ind2': [1,0]}
        exp = pd.DataFrame(data=exp_d)
        obs = df
        assert obs.equals(exp)
```



Project1-Main-Narrative

January 4, 2018

1 The effects of studying on high school students

Authors: Nadav Tadelis, Sarah Johnson, Chitwan Kaudan

1.1 Abstract

As students, a large part of our daily life is take
formed guesses about how much an extra hour
is not ideal; when making allocation decisions,
sue. Specifically, if imperfect information causes
grades, then we make poor decisions about how
(effectively resulting in a loss of utility). If we we
ing on grades, then we could calibrate our inve
naive OLS, then addressing endogeneity by usin
marginal increase in study time per week can in

1.2 Exploratory Data Analysis

The data being used are from the public archive
collected by Paulo Cortez of the University of M
Below is a list of all included variables:

in our regression.

Now that we have established our data are clean we can move on to trying to answer our question regarding the marginal effect of studying on grades.

1 We need to make the additional assumption that in secondary school (where parents are notified when students are absent), absences are only caused by illnesses and emergencies (which are independent of study time). Without this assumption it would be plausible that students are skipping school because they value leisure over studying, implying a negative correlation between study time and absences.

1.3 Initial Naive OLS fit

The first step is to build a model and make some assumptions to define the relationship between grades and studying. Let an individual's grade be G_i and weekly hours of studying be S_i and their "ability" be A_i. Then we can write:

$$G_i = \beta_0 + \beta_1 S_i + \beta_2 A_i + U_i$$

1.5.1 References

Card, D., & Krueger, A. (1992). Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy*, 100(1), 1-40.

Cortez, P. and Silva, A. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Greene, W. H. (2000). Econometric analysis. Upper Saddle River, N.J: Prentice Hall.

MacKinnon, J.G. and H. White. (1985), Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 53-57.

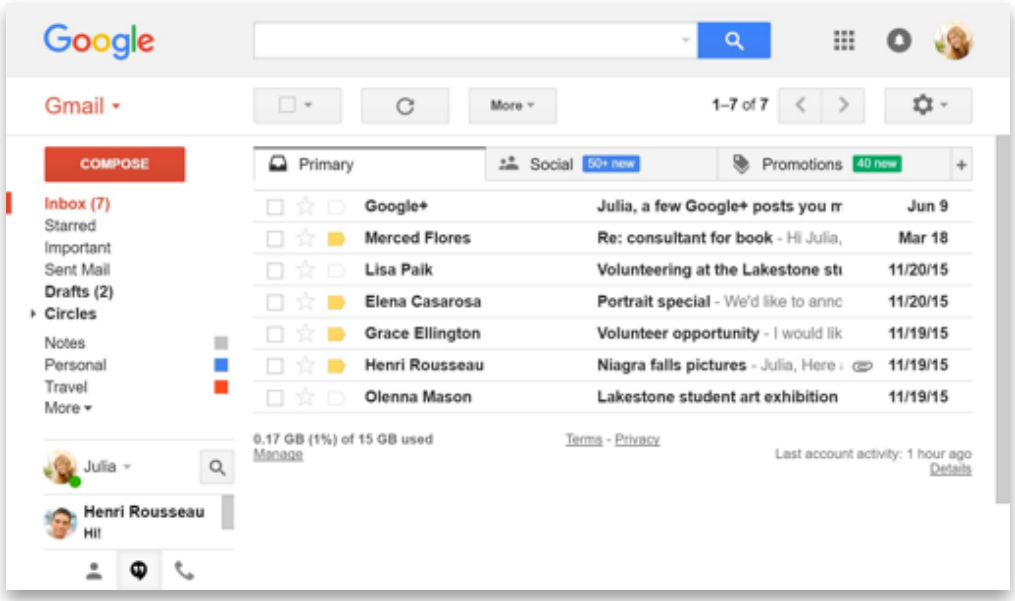
1.5.2 Author Contributions

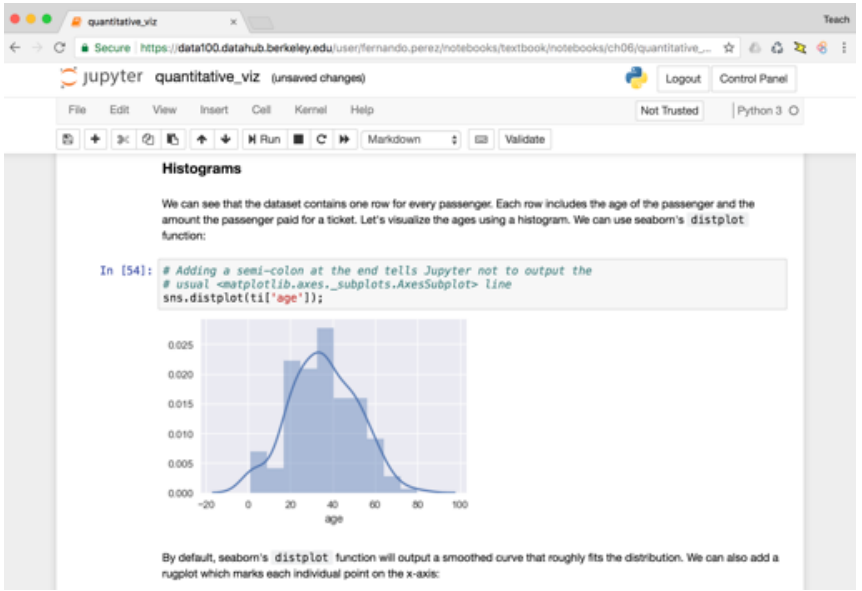
- Nadav Tadelis: Had idea from a project he did in Econ 142, worked to pick right instruments to improve the 2SLS model, wrote analysis in main.ipynb, created visualizations, and wrote/coded model fitting notebooks.
- Sarah Johnson: Helped brainstorm instruments to improve 2SLS, wrote analysis in main.ipynb, created functions and tests, and integrated testing through Travis.
- Chitwan Kaudan: Helped brainstorm instruments to improve 2SLS, wrote analysis in main.ipynb, worked on reproducibility aspects, created environment and makefile, and structured notebooks.

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.
  - (*) typically from non computational/statistical domains)

Berkeley in a few years…





Gmail by Google

jupyter

python™

In the words of a Berkeley scientist

We are witnessing a monumental phase shift in data science knowledge on campus - undergrads are extremely well trained.

[…]

There is a **knowledge gap** between the trained data science **undergrads** on campus and the **upper level scientists and PIs** who need their expertise for their projects. The labs are ill equipped on reproducible data science methods and therefore are **incapable of knowing how to manage the full potential of undergrad help**.

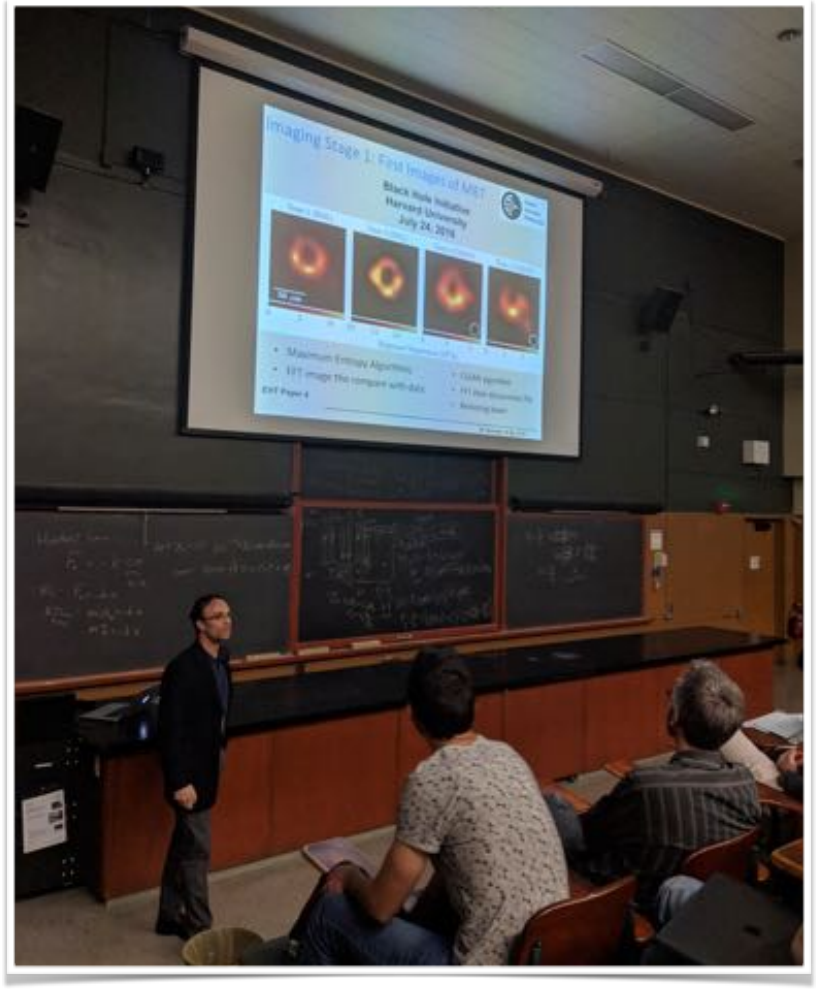
Ciera Martinez, @CieraReports

Postdoctoral Fellow, Molecular and Cell Biology & BIDS Fellow

Transforming research

Event Horizon Telescope

April 18, 2019: Shep Doeleman



THE ASTROPHYSICAL JOURNAL LETTERS

OPEN ACCESS

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

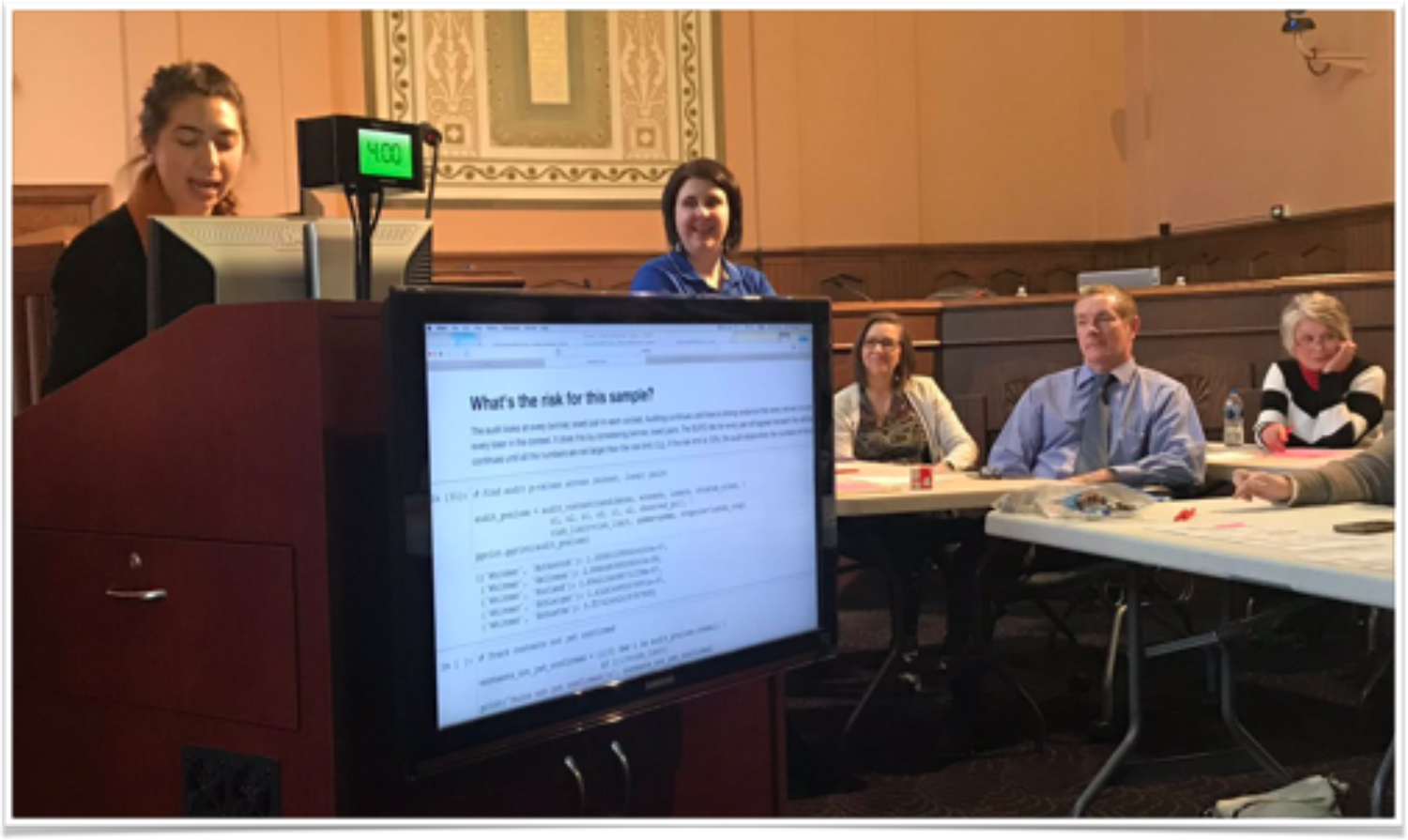
The Event Horizon Telescope Collaboration, Kazunori Akiyama[1,2,3,4], Antxon Alberdi[5], Walter Alef[6], Keiichi Asada[7], Rebecca Azulay[8,9,6], Anne-Kathrin Baczko[6], David Ball[10], Mislav Baloković[4,11], John Barrett[2] +Show full author list

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).

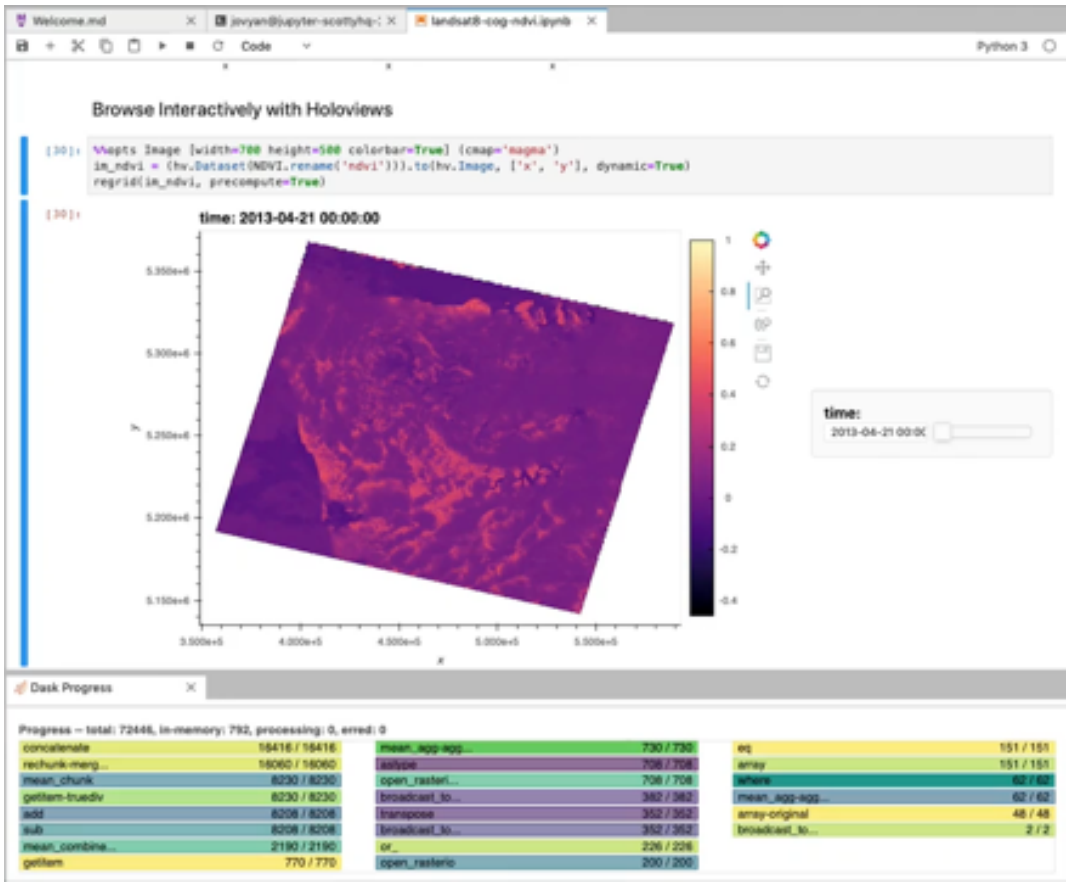
K. Ottoboni/P. Stark: Defenders of Democracy

Michigan election audit, Dec. 2018



Credit: @ginvdr

Pangeo: open geosciences (and more!)



Harnessing the power of cloud computing to study the whole earth interactively.

https://pangeo.io

Scott Henderson Follow
Research geophysicist at University of Washington eScience Institute
Oct 1 · 7 min read

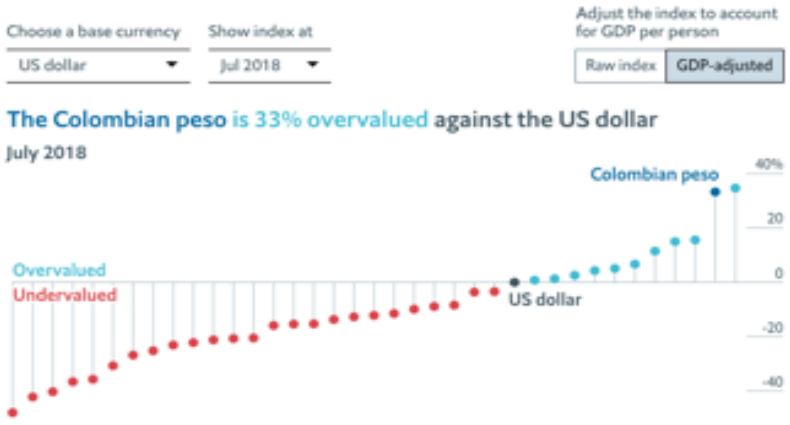
The Economist

Evan Hensleigh Follow
Visual journalist at The Economist; accidental Londoner.
Oct 9 · 3 min read

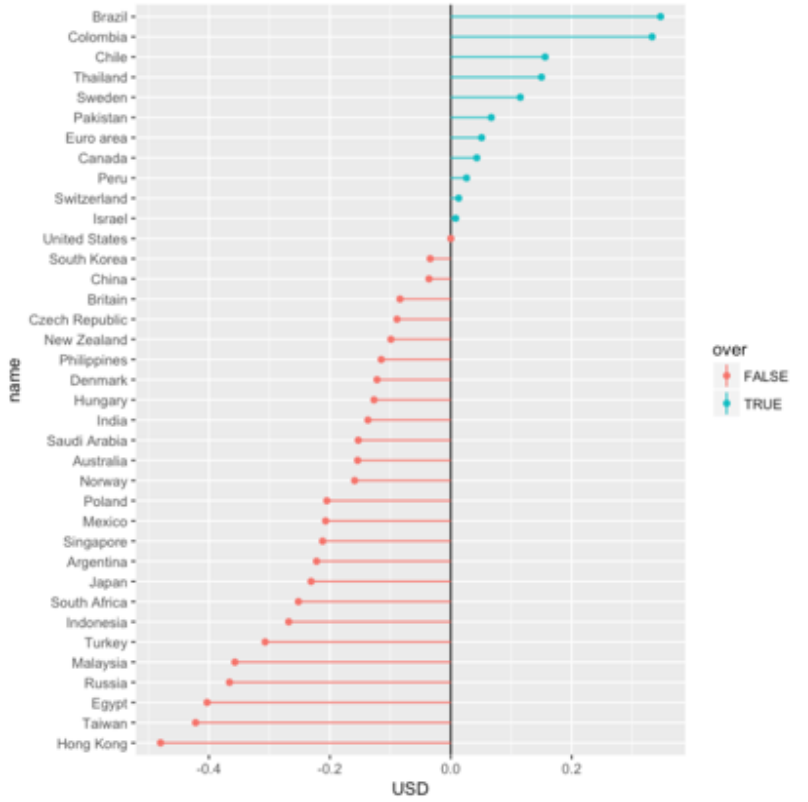
Peeling back the curtain

How the Economist is opening the data behind our reporting

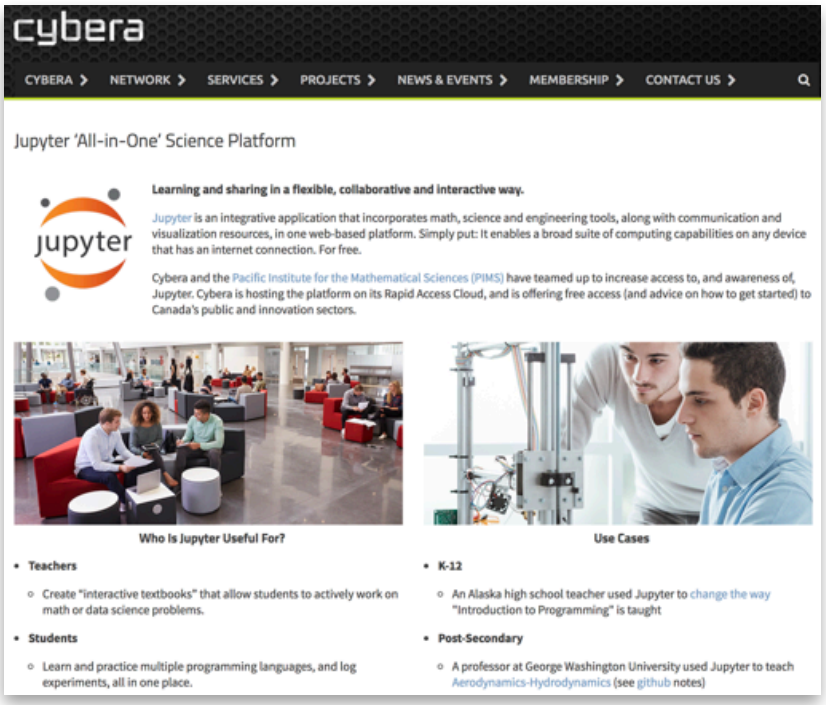
We started calculating the Big Mac index in 1986, and until this year it has been compiled and calculated manually. There are still a lot of hands-on parts to the process—in particular, compiling the list of prices—but we have now converted much of the calculation into code. We published these calculations in a Jupyter notebook, a tidy format for breaking scripts into small blocks and annotating them.

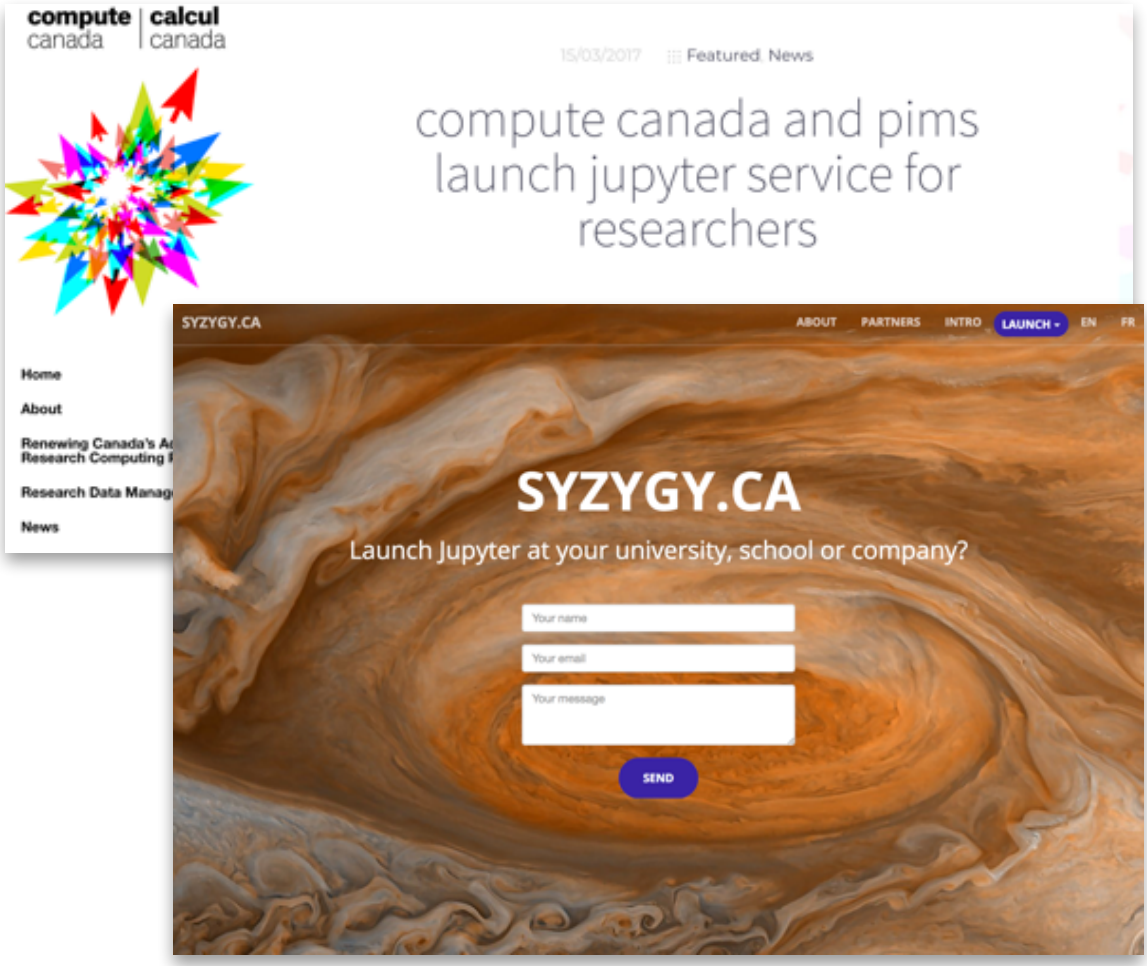


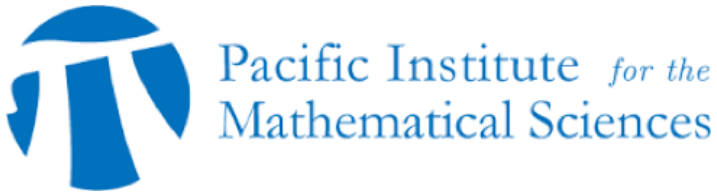

```
In [16]: to_plot = big_mac_adj_index[date == latest_date]
         to_plot$name = factor(to_plot$name, levels=to_plot$name[order(to_plot$USD)])
         ggplot(to_plot[, over := USD > 0], aes(x=name, y=USD, color=over)) +
             geom_hline(yintercept = 0) +
             geom_linerange(aes(ymin=0, ymax=USD)) +
             geom_point() +
             coord_flip()
```



National infrastructure, from K-12 to HPC







Microsoft

Microsoft Azure

COCALC

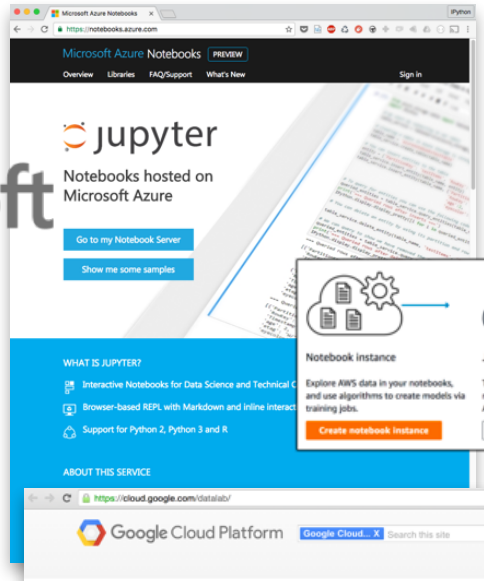
CODE OCEAN

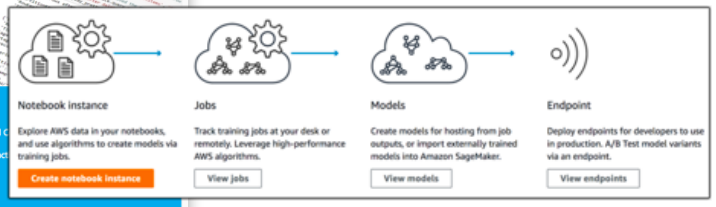
Google

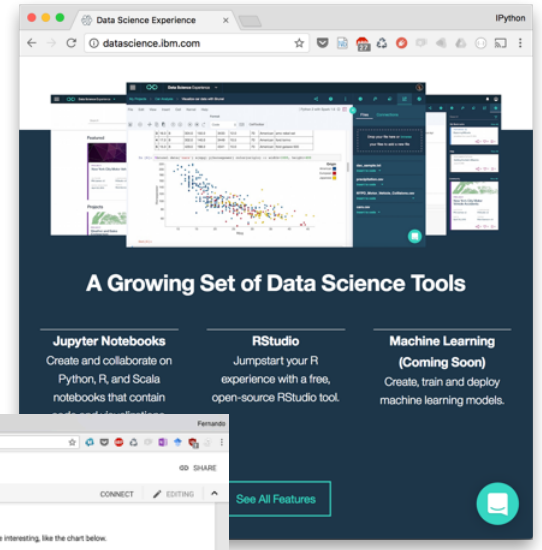
Amazon SageMaker

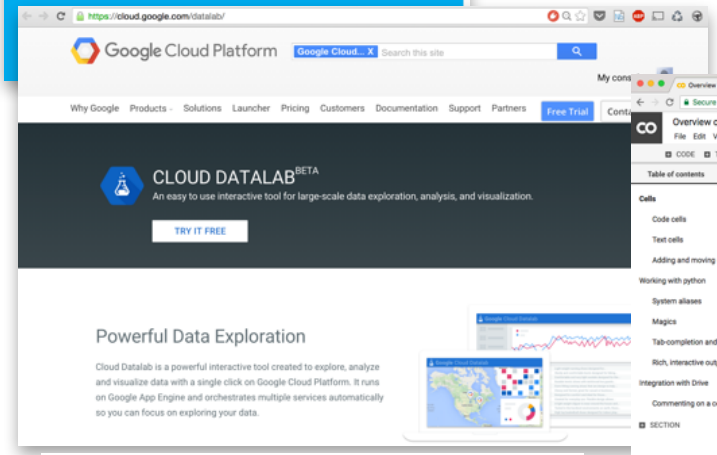
IBM

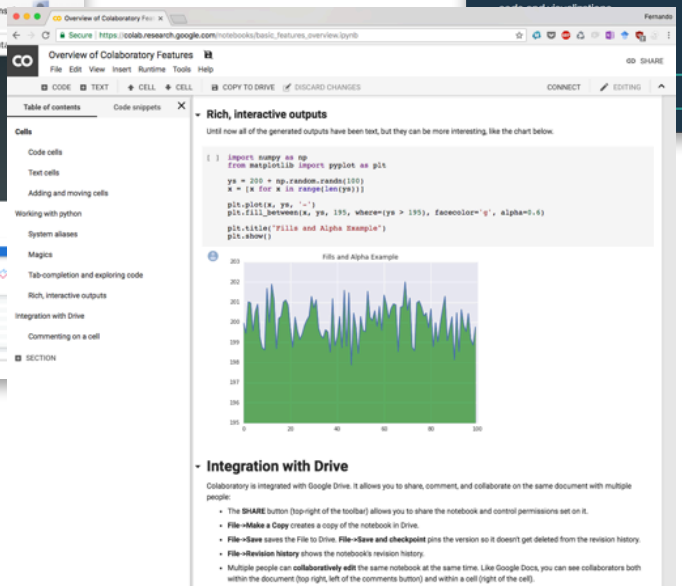
ANACONDA.











So you want to build Data Science tools in academia…

Career paths?

matplotlib

NumPy

sage

THE UNIVERSITY OF CHICAGO

BrighamYoung.University

QUANSIGHT

W

TradeLink

ANACONDA.

COCALC

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

An awkward space for academics

- These problems sit at the **interface of research and engineering**.
  - Some go to the heart of statistical thinking.
  - Yet they may require heavy-duty software engineering.
- This type of work is extremely **hard to fund and reward**.
- **Industry** alone won't do what we need.
- **Sustainability** of these efforts is today extremely challenging
  - (even funded ones like Jupyter)

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

SIMONS FOUNDATION

NETFLIX

POWERED BY rackspace® the open cloud company

ANACONDA®

IBM®

Microsoft

ENTHOUGHT SCIENTIFIC COMPUTING SOLUTIONS

Google

Bloomberg