

IP[y]:  
IPython



---

# Project Jupyter at BIDS

building the tools and  
the institutions for data  
science

---

Fernando Pérez  
([@fperez\\_org](https://twitter.com/fperez_org) & [fperez@lbl.gov](mailto:fperez@lbl.gov))

LBL & UC Berkeley





---

# A bit about me

---

- **Particle physics**, applied mathematics, neuroscience
  - Constant element: *computing in science*
- Building tools to use computers for **thinking and communicating (in science)**.
- Building projects to change the role of computers in science
  - **Open** tools for scientific computing: IPython & friends...
  - The Numfocus **foundation**
  - **BIDS**: the Berkeley Institute for Data Science



**“The purpose of computing is insight,  
not numbers”**

*–Hamming'62*



# IPython: CU Boulder, 2001

or how to best procrastinate on a Physics dissertation

```

/bin/bash

In [13]: run ~/scratch/error
reps: 5
-----
ValueError                                Traceback (most recent call last)
/home/fperez/scratch/error.py in <module>()
    70 if __name__ == '__main__':
    71     #explode()

----> 72     main()
    73     g2='another global'

/home/fperez/scratch/error.py in main()
    60 array_num = zeros(size,'d')
    61 for i in xrange(reps):
----> 62     RampNum(array_num, size, 0.0, 1.0)
    63     Rntime = time.clock()-t0
    64     print 'RampNum time:', Rntime

/home/fperez/scratch/error.py in RampNum(result, size, start, end)
    43     tmp = zeros(size+1)
    44     step = (end-start)/(size-1-tmp)
----> 45     result[:] = arange(size)*step + start
    46
    47 def main():

ValueError: shape mismatch: objects cannot be broadcast to a single shape

In [14]: □
```



# November 2001: "Just an afternoon hack"

- ❖ 259 Line Python script.
- ❖ `sys.ps1 -> In [N].`
- ❖ `sys.displayhook -> Out [N]`, caches results.
- ❖ Plotting, Numeric, etc.

## ~2014 (Openhub stats)

- ❖ 19,279 commits
- ❖ 442 contributors
- ❖ Total Lines: 187,326
- ❖ Number of Languages : 7 (JS, CSS, HTML, ...)



# A rapidly growing community



Plus ~ 500 more Open source contributors!



# Current and recent funding



**ALFRED P. SLOAN  
FOUNDATION**



SIMONS FOUNDATION

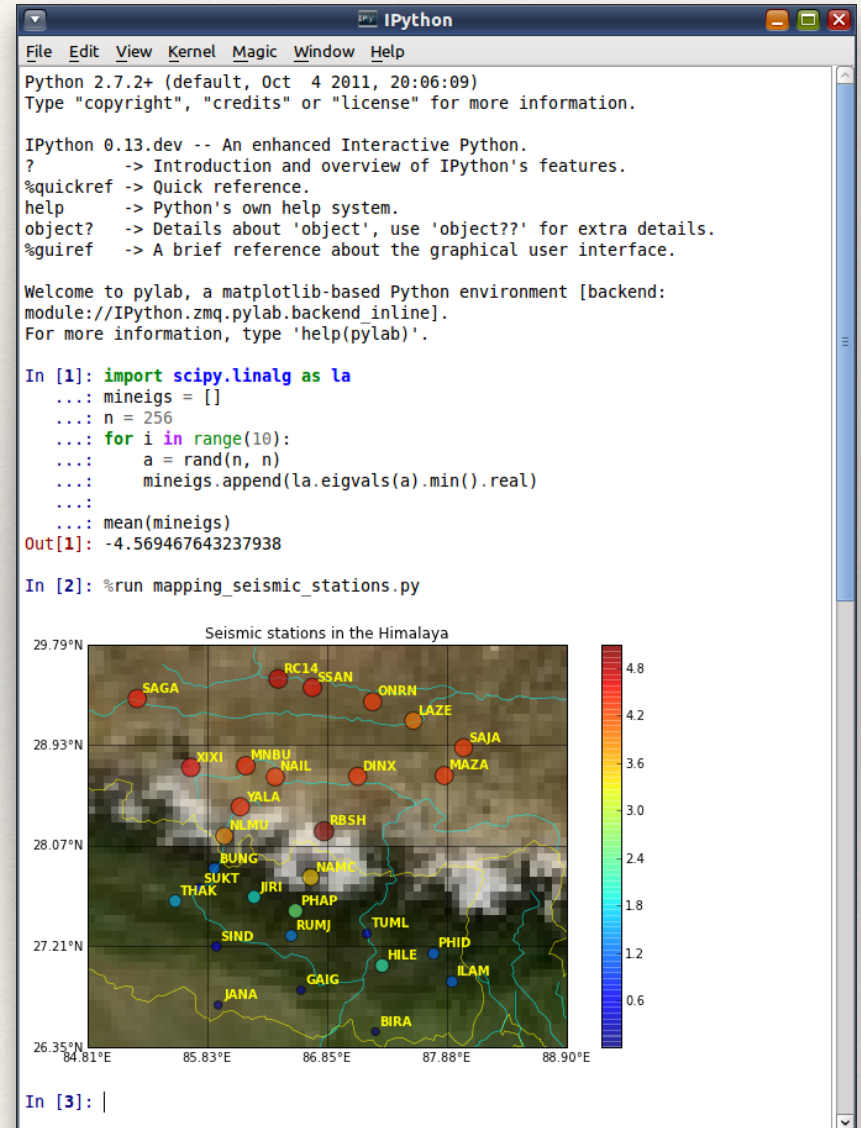




# Beyond the Terminal...

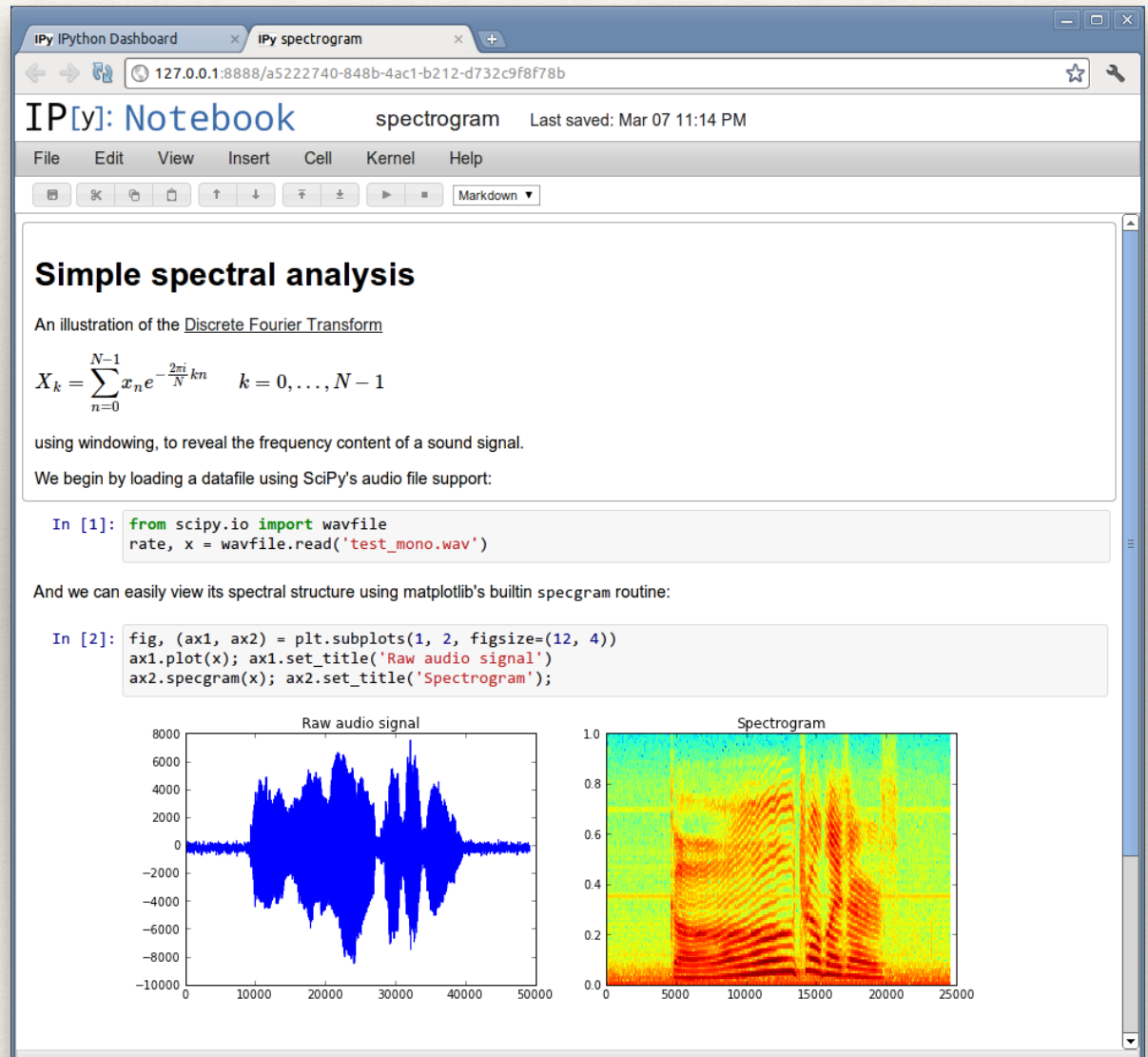
- ❖ The REPL as a network protocol
- ❖ Kernels
  - ❖ execute code
- ❖ Clients
  - ❖ Read input
  - ❖ Present output

Simple abstractions enable rich, sophisticated clients



# 2011: The IPython Notebook

- ❖ Rich web client
- ❖ Text & math
- ❖ Code
- ❖ Results
- ❖ Share, reproduce.



The screenshot shows a web browser window with the IPython Notebook interface. The browser address bar shows the URL `127.0.0.1:8888/a5222740-848b-4ac1-b212-d732c9f8f78b`. The notebook title is "spectrogram" and it was last saved on Mar 07 11:14 PM. The notebook content includes:

### Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N-1$$

using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```

The results are displayed as two side-by-side plots. The left plot, titled "Raw audio signal", shows a blue waveform of the audio signal over time, with the x-axis ranging from 0 to 50,000 and the y-axis from -10,000 to 8,000. The right plot, titled "Spectrogram", shows a heatmap of the signal's frequency content over time, with the x-axis ranging from 0 to 25,000 and the y-axis from 0.0 to 1.0.



---

# The Notebook: “Literate Computing”

---

## Computational Narratives

- ❖ Computers deal with *code and data*.
- ❖ Humans deal with narratives that *communicate*.

## Literate Computing (*not* Literate Programming)

narratives anchored in a live computation, that communicate a story based on data and results.

Cf: Mathematica, Maple, MuPad, Sage...

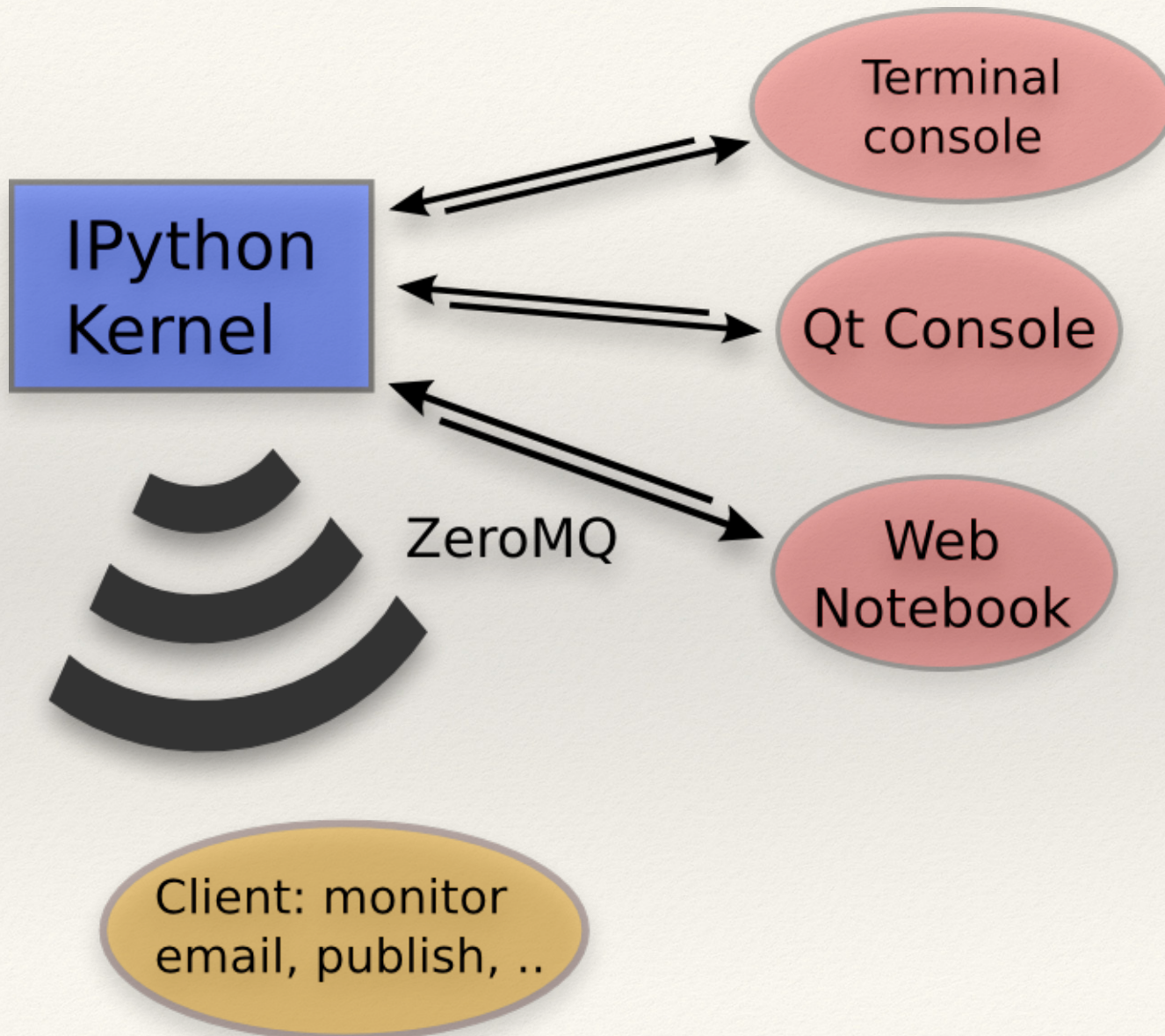
# From IPython to Project Jupyter

IP[y]:  
IPython





# A simple and generic architecture





---

# Not just about Python: Kernels in any language

---

- ❖ IPython "Official", we ship it.
- ❖ IJulia
- ❖ IRKernel
- ❖ IHaskell
- ❖ IFSharp
- ❖ Ruby
- ❖ IScala
- ❖ IErlang
- ❖ **Lots more! ~50 and counting**



“Why is it called IPython,  
if it can do Julia, R, Haskell, Ruby, ... ?”

---

# IPython

---

- ❖ Interactive Python shell at the terminal
- ❖ Kernel for this protocol in Python
- ❖ Tools for Interactive Parallel computing
- ❖ Network protocol for interactive computing
- ❖ Clients for protocol
  - ❖ Console
  - ❖ Qt Console
  - ❖ Notebook
- ❖ Notebook file format & tools (nbconvert...)
- ❖ Nbviewer



---

# IPython ... Jupyter

---

- ❖ Interactive Python shell at the terminal
- ❖ Kernel for this protocol in Python
- ❖ Tools for Interactive Parallel computing

- ❖ Network protocol for interactive computing
- ❖ Clients for protocol
  - ❖ Console
  - ❖ Qt Console
  - ❖ Notebook
- ❖ Notebook file format & tools (nbconvert...)
- ❖ Nbviewer



**Language Agnostic**



---

# What's in a name?

---

- ❖ *Inspired* by the open languages of science:
  - ❖ Julia, Python & R
  - ❖ *not* an acronym: *all languages* equal class citizens.
- ❖ *Astronomy* and Scientific Python:
  - ❖ A long and fruitful collaboration
- ❖ *Galileo's* notebooks:
  - ❖ the original, open science, data-and-narrative papers
  - ❖ Authorea: “Science was Always meant to be Open”



# The Jupyter Notebook Ecosystem

# nbviewer: seamless notebook sharing

- ❖ Zero-install reading of notebooks
- ❖ Just share a URL
- ❖ [nbviewer.ipython.org](http://nbviewer.ipython.org)

The screenshot shows the nbviewer website homepage. At the top, there are navigation links for 'nbviewer', 'FAQ', 'IPython', and 'Jupyter'. The main heading is 'nbviewer' in a large, bold font, followed by the subtitle 'A simple way to share Jupyter Notebooks'. Below this is a search bar with the placeholder text 'URL | GitHub username | GitHub username/repo | Gist ID' and a 'Go!' button. The page is organized into several sections:

- IPython**: A section featuring the IPython logo and a snippet of code: 

```
In [13]: from IPython.display import SVG
In [13]: from IPython.display import SVG
Out[13]:
```
- Programming Languages**: A section with three sub-sections:
  - IRuby**: Features the Ruby logo and a link to 'IRuby: Notebook'.
  - IJulia**: Features the Julia logo and a link to 'An IJulia Preview'.
- Books**: A section with three book covers:
  - 'Python for Signal Processing' by O'Reilly.
  - 'Mining the Social Web' (2nd Edition) by O'Reilly.
  - 'Probabilistic Programming & Bayesian Methods for Hackers' by John D. Cook.
- Misc**: A section with three sub-sections:
  - 'Data Visualization with Lightning' featuring various plots.
  - 'Interactive data visualization with Bokeh' featuring a collage of Bokeh plots.
  - 'Interactive plots with Plotly' featuring various Plotly plots.



# Reproducible Research (2012): Paper, Notebooks and Virtual Machine

The screenshot shows the ISME Journal website. The header includes the journal title 'The ISME Journal' and 'Multidisciplinary Journal of Microbial Ecology'. A search bar is visible. The main content area displays the article title 'Collaborative cloud-enabled tools allow rapid, reproducible biological insights' and its publication details: 'The ISME Journal (2013) 7, 461-464; doi:10.1038/ismej.2012.123; published online 25 October 2012'. A 'FULL TEXT' button is highlighted.

```
This notebook is intended to calculate the positions of primers in an alignment, using functions from PrimerProspector.

Import the needed functions, and define the primer sequences

In [8]: # Code modified from PrimerProspector library slice_aligned_region.py (development version)

# Imports and definitions
from string import lower, upper
from operator import itemgetter

from cogent import LoadSeqs, DNA
from cogent.core.alphabet import AlphabetError
from cogent.align.align import make_dna_scoring_dict, local_pairwise
from cogent.parse.fasta import MinimalFastaParser
from cogent.core.moltype import IUPAC_DNA_ambiguities

DNA_CODES = ['A', 'C', 'T', 'G', 'R', 'Y', 'M', 'K',
             'W', 'S', 'B', 'D', 'H', 'V', 'N']

# Note that these are all written 5'->3', the reverse primers are reverse complemented for
the local alignment

# If one wanted to test different primers, they would be defined here.

# 27f/338r = V2 (also includes V1, but generally just referred to as V2)
# 349f/534r = V3
# 515f/806r = V4
# 967f/1045r = V6
# 1391f/1492r = V9

primer_seqs = {
    '27f': 'AGAGTTTGATCMTGGCTCAG',
    '338r': DNA.rc('GCTGCTCCCGTAGAGT'),
    '349f': 'GYGASCAGKCGMGAAN',
    '534r': DNA.rc('ATTACCGCGCTGCTGG'),
    '515f': 'GTGCGAGKCCCGCGTA',
    '806r': DNA.rc('GGACTACVSGGGATCTAAT'),
    '967f': 'CAACGCGAAGACCTTACC',
    '1048r': DNA.rc('CGRCRCGATGYACWC'),
    '1391f': 'TGYACACACCGCCGTC',
    '1492r': DNA.rc('GCTACCTTGTTAGACT'),
    '1391r': 'TGYACACACCGCCGTC' # Need this rather than forward primer to get proper
3' position of reverse version
}

reference_aligned_file = '/home/ubuntu/qiime_software/gg_otus-4feb2011-release/rep_set/gg_
76_otus_4feb2011_aligned.fasta'
```

agan-Kelley<sup>1,12</sup>, William Anton Walters<sup>2,12</sup>,  
Donald<sup>3,5,12</sup>, Justin Riley<sup>4</sup>, Brian E Granger<sup>5</sup>,  
nzalez<sup>6</sup>, Rob Knight<sup>7,8</sup>, Fernando Perez<sup>9</sup> and J  
poraso<sup>10,11</sup>

Group in Applied Science and Technology, University of  
t Berkeley, Berkeley, CA, USA  
nt of Molecular, Cellular and Developmental Biology,  
f Colorado at Boulder, Boulder, CO, USA  
s Institute, University of Colorado at Boulder, Boulder, CO,  
ducational Innovation and Technology, Massachusetts  
Technology, Cambridge, MA, USA  
partment, California Polytechnic State University, San Luis  
USA  
nt of Computer Science, University of Colorado at Boulder,  
, USA

The screenshot shows the 'FULL TEXT' section of the article. It includes a 'Table of contents' with links for 'Download PDF', 'Send to a friend', and 'View interactive PDF in'. Below this, there are instructions for reproducing the analysis, including a list of supporting files and a code block for setting up a virtual machine. The code block is as follows:

```
[plugin icluster]
setup_class = starcluster.plugins.icluster.ICluster
enable_notebook = true
# If you leave notebook_passwd out, a random password
# will be generated instead.
notebook_passwd = YOUR-PASSWORD

[cluster qiime-ipynon]
node_image_id = ami-9f69c1f6
cluster_user = ubuntu
keyname = YOUR-KEY
cluster_size = 4
node_instance_type = m2.4xlarge
plugins = icluster
volumes = qiime-ipynon-data

[volume qiime-ipynon-data]
VOLUME_ID = YOUR-VOLUME-ID
MOUNT_PATH = /home/ubuntu/data
```

# Today: mybinder.org



Turn a GitHub repo into a collection of interactive notebooks powered by Jupyter and Kubernetes.



[github.com/freeman-lab](https://github.com/freeman-lab)

1

## Tell us your GitHub repo

user/project OR github url

2

## Configure dependencies

- `> none` for basic Python projects
- `requirements.txt` for pip Python projects
- `environment.yml` for conda Python projects
- `Dockerfile` for custom builds

3

## Attach services

- Postgres
- Spark

make my binder



# Scientific Blogging

SCIENTIFIC AMERICAN™

Sign In | Register

Search ScientificAmerican.com

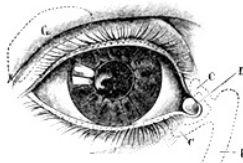
Subscribe News & Features Topics Blogs Videos & Podcasts Education Cit

SA en español

Blogs

About

Like 0 Tweet 2 +1 3 in Share 3 reddit this!



SA Visual

Illustrating science since 1845

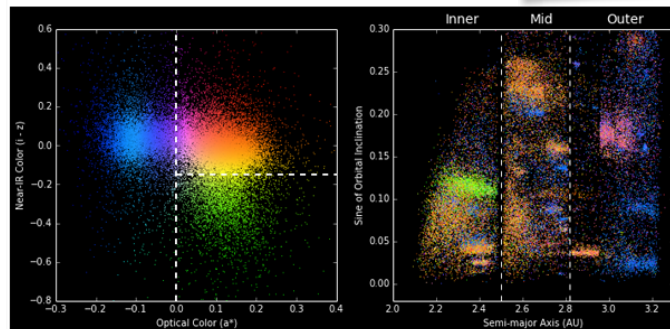
SA Visual Home About Contact

## Visualizing 4-Dimensional Asteroids

By Jake VanderPlas | September 16, 2014

### Multicolor plot

Let's put these all together. Rather than using two separate color scales to identify these asteroid groups, we can define a single two-dimensional color reflecting the asteroid chemistry and use these colors when plotting the same points in orbital space. The result is a plot very similar to the one that appeared in where this work was first reported:

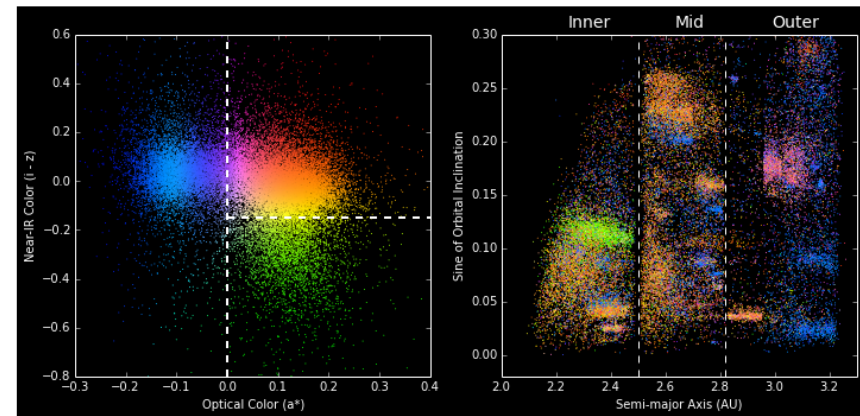


nbviewer.ipython.org/github/jakevdp/SciAmBlogPost/blob/master/AsteroidVis.ipynb

### Multicolor plot

Let's put these all together. Rather than using two separate color scales to identify these asteroid groups, we can define a single two-dimensional color scale reflecting the asteroid chemistry and use these colors when plotting the same points in orbital space. The result is a plot very similar to the one that appeared in Parker et al., 2008, where this work was first reported:

```
In [13]: plot_multicolor()
```



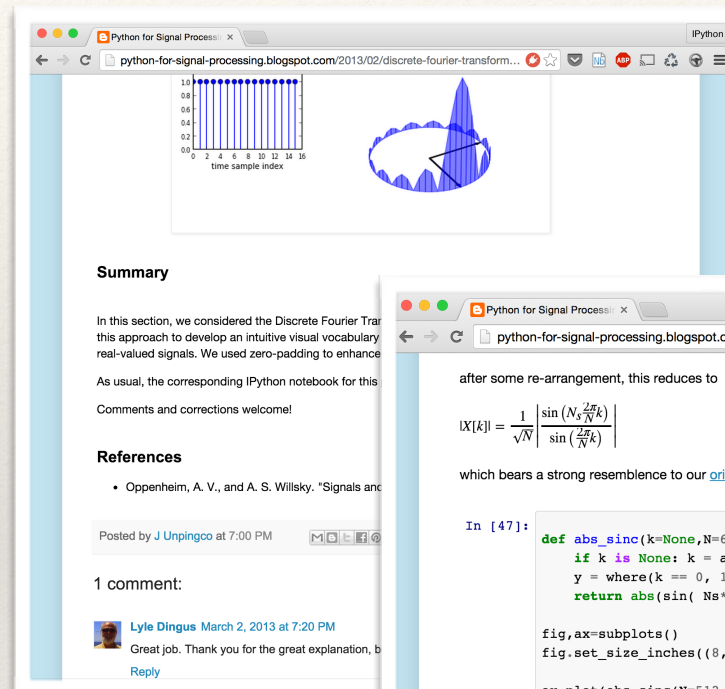
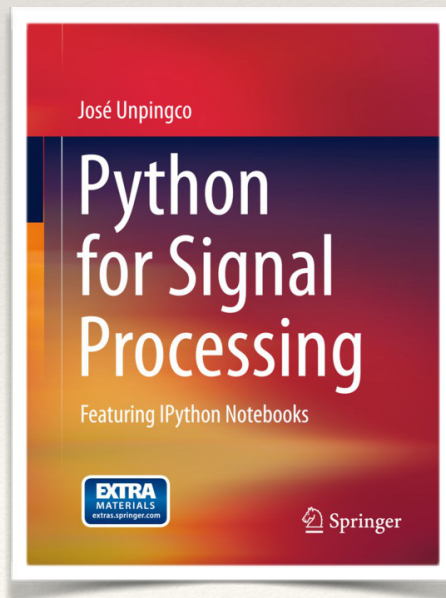
Jake van der Plas @ UW

<http://blogs.scientificamerican.com/sa-visual/2014/09/16/visualizing-4-dimensional-asteroids>

# Executable books

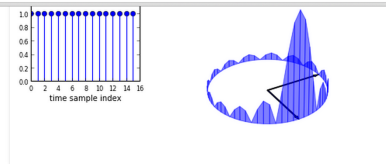
Python for Signal Processing, by José Unpingco

- ❖ Springer hardcover book
- ❖ Chapters: IPython Notebooks
- ❖ Posted as a blog entry
- ❖ All available as a Github repo



Python for Signal Processing

python-for-signal-processing.blogspot.com/2013/02/discrete-fourier-transform...



**Summary**

In this section, we considered the Discrete Fourier Transform. This approach to develop an intuitive visual vocabulary for real-valued signals. We used zero-padding to enhance the resolution of the DFT.

As usual, the corresponding IPython notebook for this section is available. Comments and corrections welcome!

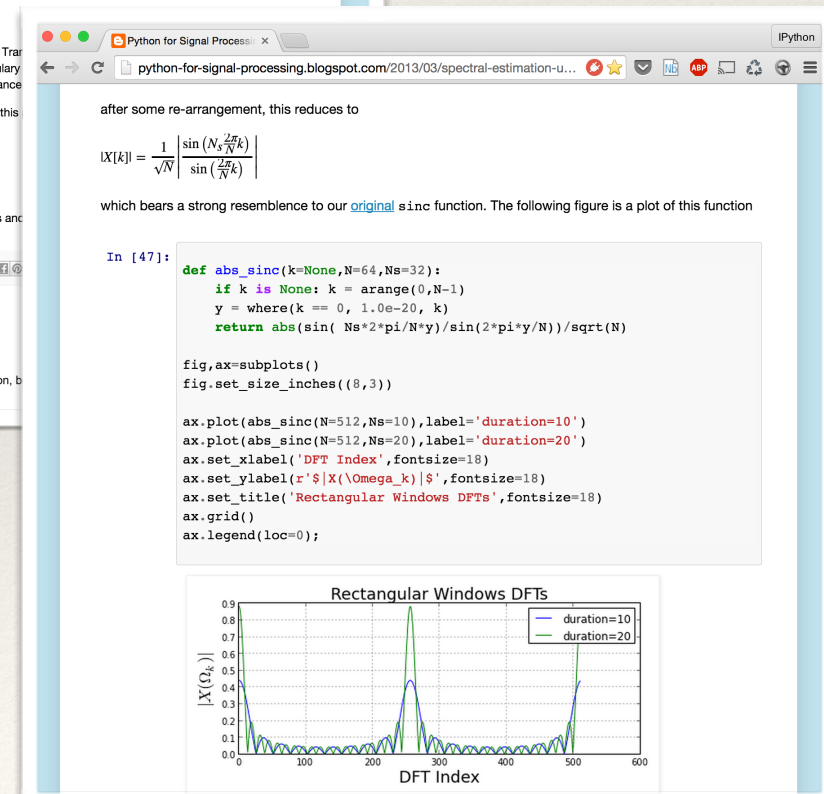
**References**

- Oppenheim, A. V., and A. S. Willsky. "Signals and Systems." Prentice Hall, 1989.

Posted by J Unpingco at 7:00 PM

1 comment:

Lyle Dingus March 2, 2013 at 7:20 PM  
Great job. Thank you for the great explanation, book!



Python for Signal Processing

python-for-signal-processing.blogspot.com/2013/03/spectral-estimation-u...

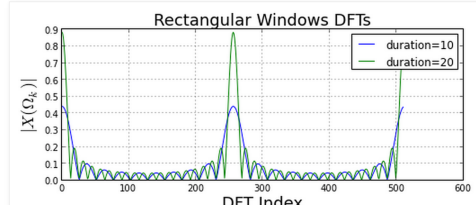
after some re-arrangement, this reduces to

$$|X[k]| = \frac{1}{\sqrt{N}} \left| \frac{\sin(N_s \frac{2\pi}{N} k)}{\sin(\frac{2\pi}{N} k)} \right|$$

which bears a strong resemblance to our [original](#) sinc function. The following figure is a plot of this function

In [47]:

```
def abs_sinc(k=None, N=64, Ns=32):  
    if k is None: k = range(0, N-1)  
    y = where(k == 0, 1.0e-20, k)  
    return abs(sin(Ns*2*pi/N*y)/sin(2*pi*y/N))/sqrt(N)  
  
fig, ax = subplots()  
fig.set_size_inches((8, 3))  
  
ax.plot(abs_sinc(N=512, Ns=10), label='duration=10')  
ax.plot(abs_sinc(N=512, Ns=20), label='duration=20')  
ax.set_xlabel('DFT Index', fontsize=18)  
ax.set_ylabel(r'$|X(\Omega_k)|$', fontsize=18)  
ax.set_title('Rectangular Windows DFTs', fontsize=18)  
ax.grid()  
ax.legend(loc=0);
```





# University Courses

	Course	University	Instructor
0	Data Science and Visualization with Python	Santa Clara	Brian Granger
1	Python for Data Science	UC Berkeley	Josh Bloom
2	Introduction to Data Science	UC Berkeley	Michael Franklin
3	Working with Open Data	UC Berkeley	Raymond Yee
4	Introduction to Signal Processing	UC Berkeley	Miki Lustig
5	Data Science (CS 109)	Harvard University	Pfister and Blitzstein
6	Practical Data Science	NYU	Josh Attenberg
7	Scientific Computing (ASTR 599)	University of Washington	Jake Vanderplas
8	Computational Physics	Cal Poly	Jennifer Klay
9	Introduction to Programming	Alaskan High School	Eric Matthes
10	Aerodynamics-Hydrodynamics (MAE 6226)	George Washington University	Lorena Barba

11	HyperPython: hyperbolic conservation laws	KAUST	David Ketcheson
12	Quantitative Economics	NYU	Sargent and Stachurski
13	Practical Numerical Methods with Python	4 separate universities + MOOC	Barba, et al.
14	Data Science: Algorithms	Columbia - Lede Program	Chris Wiggins
15	Data Science: Databases	Columbia - Lede Program	Chris Wiggins
16	Data Science: Foundations	Columbia - Lede Program	Chris Wiggins
17	Data Science: Platforms	Columbia - Lede Program	Chris Wiggins

These are just some we are aware of!

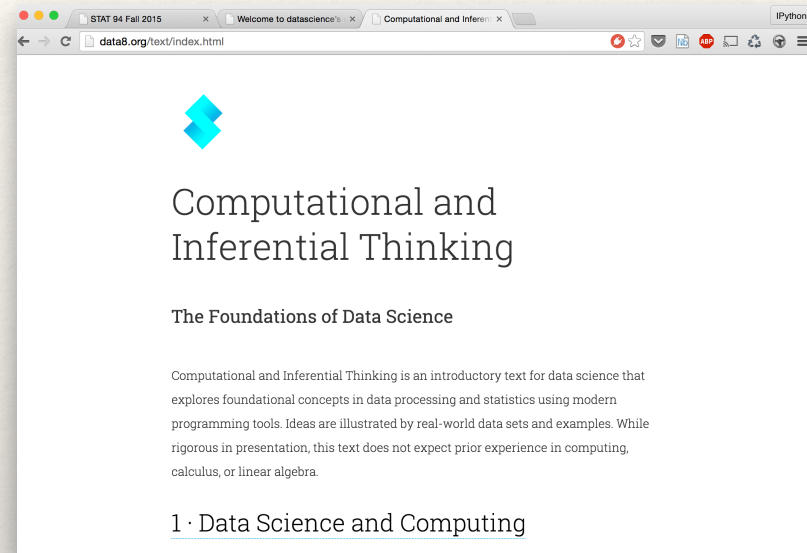
New Jupyter in Education Mailing List:

<https://groups.google.com/forum/#!forum/jupyter-education>

# Berkeley's *Foundations of Data Science*

- ❖ New curriculum aimed at all freshmen at UC Berkeley
- ❖ Interactive textbook is Jupyter Notebooks
- ❖ Course deployment is JupyterHub

<http://data8.org>



## Arrays

[INTERACT](#)

Many experiments and data sets involve multiple values of the same type. An *array* is a collection of values that all have the same type. The `numpy` package, abbreviated `np` in programs, provides Python programmers with convenient and powerful functions for creating and manipulating arrays.

An array is created using the `np.array` function, which takes a list or tuple as an argument.

```
temps = np.array([8.1, 8.3, 8.7, 9.4])
temps
array([ 8.1,  8.3,  8.7,  9.4])
```

Arrays differ from lists and tuples because they can be used in arithmetic expressions to compute over their contents. When two arrays are combined together using an arithmetic operator, their individual values are combined.

## Example: Plotting the Classics

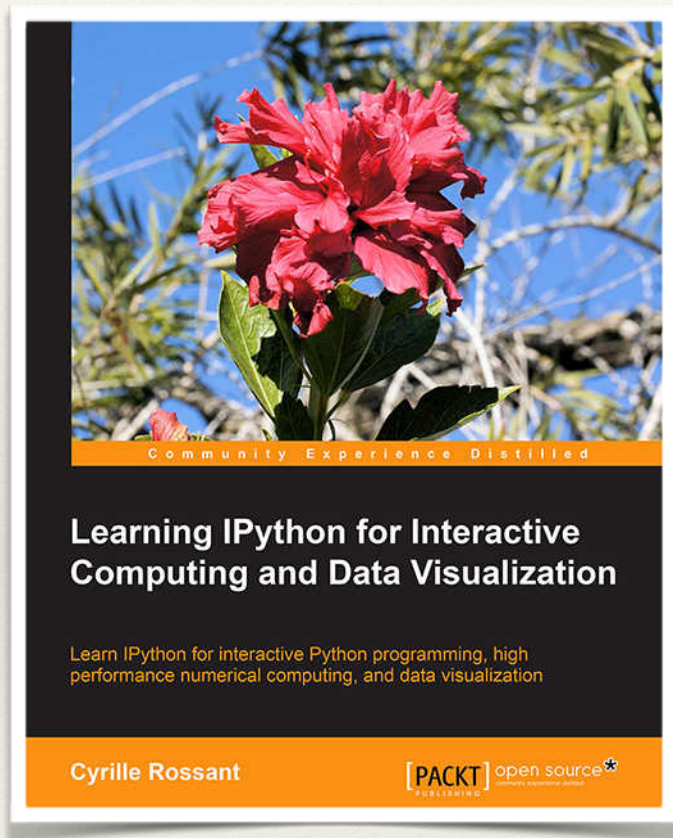
[INTERACT](#)

In this example, we will explore statistics for two classic novels: *The Adventures of Huckleberry Finn* by Mark Twain, and *Little Women* by Louisa May Alcott. The text of any book can be read by a computer at great speed. Books published before 1923 are currently in the *public domain*, meaning that everyone has the right to copy or use the text in any way. [Project Gutenberg](#) is a website that publishes public domain books online. Using Python, we can load the text of these books directly from the web.

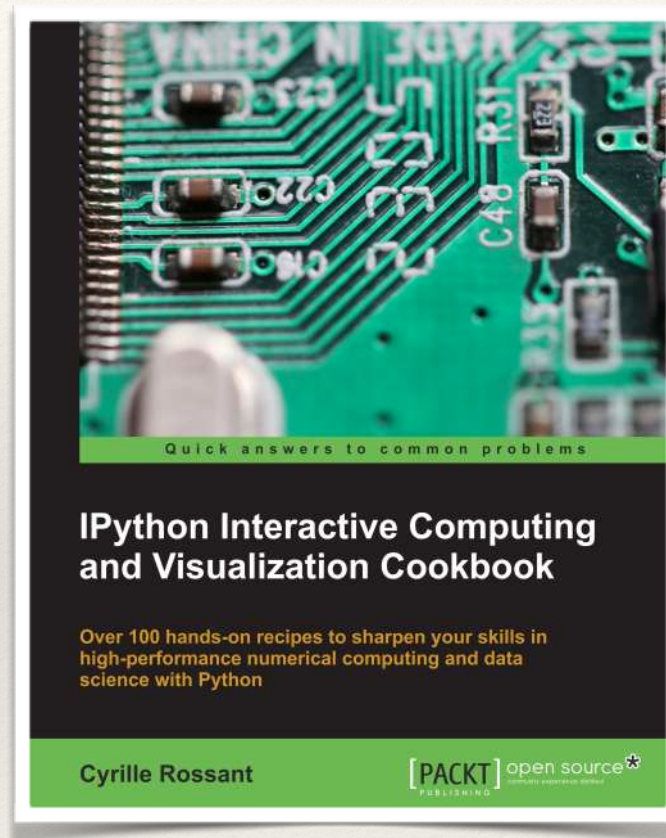
The features of Python used in this example will be explained in detail later in the course. This example is meant to illustrate some of the broad themes of this text. Don't worry if the details of the program don't yet make sense. Instead, focus on interpreting the images generated below. The "Expressions" section later in this chapter will describe most of the features of the Python programming language used below.



# Books about IPython



Learning IPython for Interactive Computing and Data Visualization



IPython Interactive Computing and Visualization Cookbook



Cyrille Rossant  
[cyrille.rossant.net](http://cyrille.rossant.net)



# Changing the scientific culture

The screenshot shows the Nature journal website interface. At the top, the 'nature' logo is displayed with the tagline 'International weekly journal of science'. A search bar and 'Go' button are on the right. Below the logo is a navigation menu with links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Authors. A secondary navigation bar includes Archive, Volume 515, Issue 7525, Toolbox, and Article. The main content area features the article title 'Interactive notebooks: Sharing the code' by Helen Shen, dated 05 November 2014. A large illustration depicts hands interacting with a digital notebook displaying various scientific data visualizations like graphs and charts. To the right, a 'Top story' section highlights a dinosaur article: 'Beloved Brontosaurus makes a comeback', with a sub-headline 'Jurassic giant's taxonomic status is restored.' Below this is a 'Recent' section with a list of four articles: 'History: Women at the edge of science', 'Scientific instrumentation: The aided eye', 'Books in brief', and 'Antibody shows promise as...'. Social media sharing options for E-alert, RSS, Facebook, and Twitter are visible at the top right of the article content.

**nature** International weekly journal of science

Search   [Advanced search](#)

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 515 > Issue 7525 > Toolbox > Article

NATURE | TOOLBOX

## Interactive notebooks: Sharing the code

The free IPython notebook makes data analysis easier to record, understand and reproduce.

**Helen Shen**

05 November 2014

**Top story**

**USA 25**  
*Brontosaurus*

**Beloved *Brontosaurus* makes a comeback**

Jurassic giant's taxonomic status is restored.

**Recent**

- History: Women at the edge of science**  
*Nature* | 08 April 2015
- Scientific instrumentation: The aided eye**  
*Nature* | 08 April 2015
- Books in brief**  
*Nature* | 08 April 2015
- Antibody shows promise as**

*Illustrations by The Project Twins*

<http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261>



# Executable papers: the future?

nature.com | Sitemap | Login | Register | Close

IPython Notebook Nature (autosaved) | IPython (Python 3)

File Edit View Insert Cell Kernel Help

Markdown Cell Toolbar: None

**nature** **rackspace**

### Introduction

Welcome! You have just launched a live example of an IPython Notebook. The notebook is an open-source, interactive computing environment that lets you combine live code, narrative text, mathematics, plots and rich media in one document. Notebook documents provide a complete reproducible record of a computation and its results and can be shared with colleagues (through, for example, email, web-hosting services such as GitHub, Dropbox, and [nbviewer](#)).

You can edit anything in this temporary demonstration notebook, including the text you are reading. To see it full-screen, click on the 'Expand' icon in the lower right corner of the frame around this notebook.

This notebook showcases some of IPython's capabilities for researchers.

This demonstration is hosted by [Rackspace](#) and is running on its bare metal offering, [OnMetal](#). Try out these cloud services yourself through [Rackspace's developer+ page](#).

### Basic Python code and plotting

The box below (known as a code cell) contains the Python code to plot  $y = x^2$  over the range  $[0, 5]$ . The blue comments preceded by # explain what the code does.

To run the code:

1. Click on the cell to select it.
2. Press SHIFT+ENTER on your keyboard or press the play button (▶) in the toolbar above.

A full tutorial for using the notebook interface is available [here](#).

```
In [ ]: # Import matplotlib (plotting) and numpy (numerical arrays).
# This enables their use in the Notebook.
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

# Create an array of 30 values for x equally spaced from 0 to 5.
x = np.linspace(0, 5, 30)
```

```
ax.imshow(image_gray, interpolation='nearest', cmap=gray_r
circle_color = 'red'
else:
ax.imshow(image, interpolation='nearest')
circle_color = 'yellow'
for blob in blobs:
y, x, r = blob
c = plt.Circle((x, y), r, color=circle_color, linewidth=2,
ax.add_patch(c)

# Use interact to explore the galaxy detection algorithm.
interact(plot_blobs, max_sigma=(10, 40, 2), threshold=(0.005, 0.02,
```

x max\_sigma 30  
threshold 0.017  
gray

Galaxies in the Hubble Deep Field

# Notebook Workflows: The Big Picture

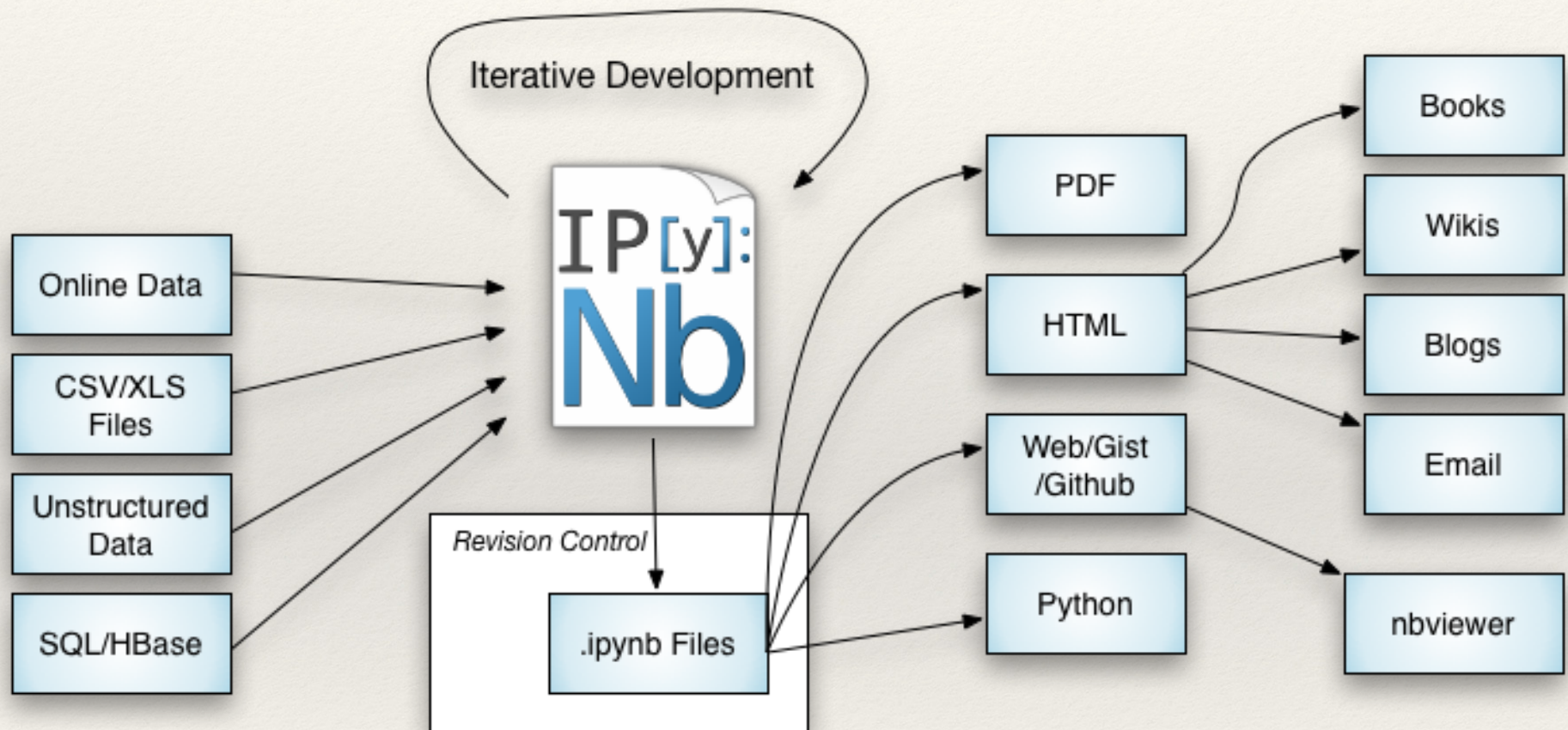


Image credit: [Joshua Barratt](#)



# Lots more! The IPython Gallery

## A gallery of interesting IPython Notebooks

Fernando Perez edited this page 8 days ago · 229 revisions

This page is a curated collection of IPython notebooks that are notable for some reason. Feel free to add new content here, but please try to only include links to notebooks that include interesting visual or technical content; this should *not* simply be a dump of a Google search on every ipynb file out there.

**Important contribution instructions:** If you add new content, please ensure that for any notebook you link to, the link is to the rendered version using [nbviewer](#), rather than the raw file. Simply paste the notebook URL in the nbviewer box and copy the resulting URL of the rendered version. This will make it much easier for visitors to be able to immediately access the new content.

Note that [Matt Davis](#) has conveniently written a set of [bookmarklets and extensions](#) to make it a one-click affair to load a Notebook URL into your browser of choice, directly opening into nbviewer.

## Table of Contents

1. [Entire books or other large collections of notebooks on a topic](#)
  - [Introductory Tutorials](#)
  - [Programming and Computer Science](#)
  - [Statistics, Machine Learning and Data Science](#)
  - [Mathematics, Physics, Chemistry, Biology](#)
  - [Earth Science and Geo-Spatial data](#)
  - [Linguistics and Text Mining](#)
  - [Signal Processing](#)
2. [Scientific computing and data analysis with the SciPy Stack](#)
  - [General topics in scientific computing](#)
  - [Social data](#)
  - [Psychology and Neuroscience](#)
  - [Machine Learning](#)
  - [Physics, Chemistry and Biology](#)
  - [Economics](#)
  - [Earth science and geo-spatial data](#)

## Reproducible academic publications

This section contains academic papers that have been published in the peer-reviewed literature or pre-print sites such as the [ArXiv](#) that include one or more notebooks that enable (even if only partially) readers to reproduce the results of the publication. If you include a publication here, please link to the journal article as well as providing the nbviewer notebook link (and any other relevant resources associated with the paper).




1. [Reply to 'Influence of cosmic ray variability on the monsoon rainfall and temperature': a false-positive in the field of solar-terrestrial research](#) by Benjamin Laken, 2015. Reviewed article will appear in JASTP. The [IPython notebook](#) reproduces the full analysis and figures exactly as they appear in the article, and is available on Github: [link via figshare](#).
2. [The probability of improvement in Fisher's geometric model: a probabilistic approach](#), by Yoav Ram and Lilach Hadany. (Theoretical Population Biology, 2014). An [IPython notebook](#), allowing figure reproduction, was deposited as a [supplementary file](#).
3. [Stress-induced mutagenesis and complex adaptation](#), by Yoav Ram and Lilach Hadany (Proceedings B, 2014). An [IPython notebook](#), allowing figures reproduction, was deposited as a [supplementary file](#).
4. [Automatic segmentation of odor maps in the mouse olfactory bulb using regularized non-negative matrix factorization](#), by J. Soelter et al. (Neuroimage 2014, Open Access). The [notebook](#) allows to reproduce most figures from the paper and provides a deeper look at the data. The [full code repository](#) is also available.
5. [Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss](#), by A. Gross et al. (Nature Genetics 2014). The full collection of notebooks to replicate the results.
6. [powerlaw: a Python package for analysis of heavy-tailed distributions](#), by J. Alstott et al.. [Notebook of examples in manuscript](#), [ArXiv link](#) and [project repository](#).
7. [Collaborative cloud-enabled tools allow rapid, reproducible biological insights](#), by B. Ragan-Kelley et al.. The [main notebook](#), the [full collection of related notebooks](#) and the [companion site](#) with the Amazon AMI information for reproducing the full paper.
8. [A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data](#), by C.T. Brown et al.. [Full notebook](#), [ArXiv link](#) and [project repository](#).
9. [The kinematics of the Local Group in a cosmological context](#) by J.E. Forero-Romero et al.. The [Full notebook](#) and also all the data in a [github repo](#).

Growth: opportunity and  
challenge



NEWS & MEDIA

- News
- CS In the News
- InTheLoop

-  Facebook
-  Google
-  Twitter

# Project Jupyter gets \$6M to expand collaborative data science software

JULY 7, 2015

Tags: [CRD](#)

PALO ALTO, Calif. July 7, 2015 — Three foundations pledged \$6M over the next three years to Project Jupyter, an open-source software project that supports scientific computing and data science across a wide range of programming languages via a large, public, open and inclusive community.

Fernando Perez of University of California, Berkeley and [Lawrence Berkeley National Laboratory \(Berkeley Lab's\) Computational Research Division](#) and Brian Granger of California Polytechnic University, San Luis Obispo will lead the project at their institutions. Perez and Granger's efforts with Project Jupyter are the result of their work developing IPython, a popular user interface for interactive computing across multiple programming languages.

With this award from the Leona M. and Harry B. Helmsley Charitable Trust, Alfred P. Sloan Foundation, and Gordon and Betty Moore Foundation, these researchers will expand and improve the capabilities of the Jupyter Notebook, a web-based platform that allows scientists, researchers and educators to combine live code, equations, narrative text and rich media into a single, interactive document.



**Fernando Perez and Brian Granger discuss the architecture of Project Jupyter, as its scope expands to reach data science applications in over 40 programming languages. Photo credit: Adriana Restrepo**

# JupyterHub: multiuser support



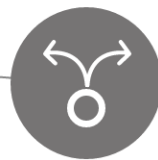
## Jupyter for Organizations

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



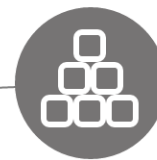
### Pluggable authentication

Manage users and authentication with PAM, OAuth or integrate with your own directory service system. Collaborate with others through the Linux permission model.



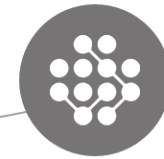
### Centralized deployment

Deploy the Jupyter Notebook to all users in your organization on centralized servers on- or off-site.



### Container friendly

Use Docker containers to scale your deployment and isolate user processes using a growing ecosystem of prebuilt Docker containers.



### Code meets data

Deploy the Notebook next to your data to provide unified software management and data access within your organization.

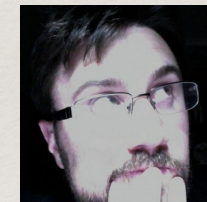
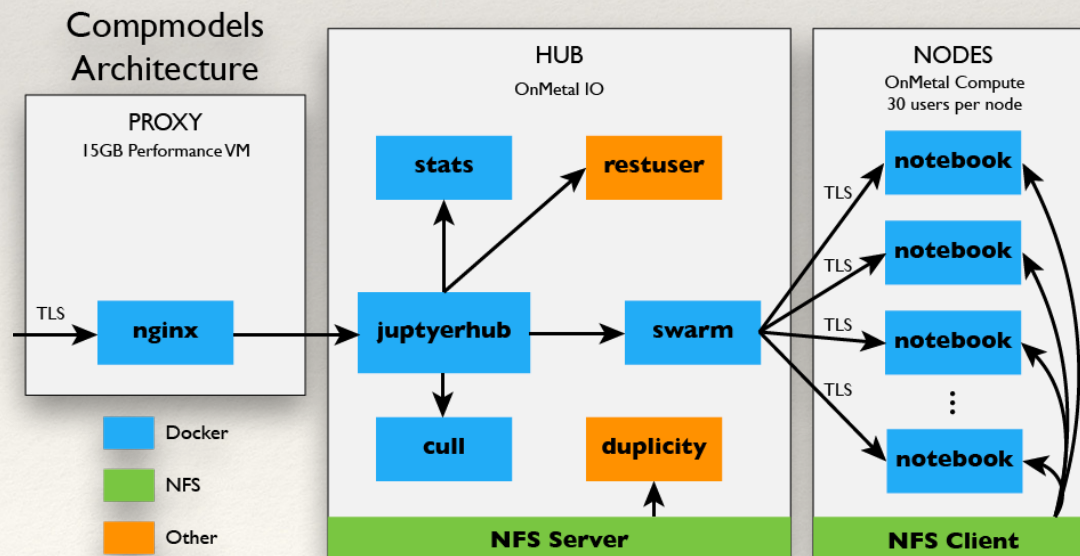


# JupyterHub in Education @ Berkeley

- ❖ Computationally intensive course, ~220 students
- ❖ Fully hosted environment, zero-install, spring 2015.
- ❖ Homework management and grading (w B. Granger)
- ❖ Now powers [data8.org](http://data8.org) - Cal's new *Foundations of Data Science*, (fall 2015).



Jess Hamrick @ Cal



K. Kelley  
Rackspace



M. Ragan-Kelley  
Cal



B. Granger  
Cal Poly



<https://developer.rackspace.com/blog/deploying-jupyterhub-for-education>

# Industry: Microsoft, IBM, Google, ...

**Machine Learning Blog**

## Introducing Jupyter Notebooks in Azure ML Studio

ML Blog Team | 24 Jul 2015 10:00 AM | 10

Posted by *Shahrokh Mortazavi*, Partner Director of Program Management at Microsoft.

Azure ML Studio is a powerful canvas for the composition of machine learning experiments and their subsequent operationalization and consumption. Although the Studio provides an easy to use, yet powerful, drag-drop style of creating experiments, you sometimes need a good old "REPL" to have a tight loop where you enter some script code and get a response. I am delighted to announce that we've now integrated this functionality into ML Studio through Jupyter Notebooks:

NAME	LANGUAGE	LAST MODIFIED
Sample-1.ipynb	Python 2	5/15/2015 4:39:52 PM
Basic Notebook from AzureMLStudio	Python 2	5/15/2015 2:02:34 PM
Simple Notebook	Python 2	5/15/2015 2:04:07 PM
Movie Recommendation	Python 2	5/15/2015 9:54:12 AM
Sample-2	Python 2	5/15/2015 9:52:11 AM

**Related Links**

- Microsoft Azure ML
- Microsoft Data Platform Insider Blog
- Microsoft Big Data Solutions
- Data Science Dojo
- Azure Big Data Blog

**Tags**

- ADF
- asa
- Azure ML
- Azure Stream Analytics
- Cortana Analytics Suite
- Customers Data Science Machine Learning Partners
- Python r Webinar

More ▾ Less ▲

## Data Scientist Workbench

Prepare data. Analyze data. Get answers.

Prepare data effortlessly.

Explore Data.  
Find and explore large data sets with ease.

Clean and Transform Data  
Easily clean messy data and transform formats.

Reconcile and Match Data  
Link and extend your datasets with web services.

Analyze data interactively.

**Powerful Notebook Environment**  
Use IPython/Jupyter notebooks to combine code execution, text, plots and rich media.

**Re-use and Extend**  
Use preinstalled Python and R libraries. Install others as needed.

**Collaborate and Share**  
Build on what others have done and easily share your analysis.

```
p = ml.get_dataframe('x', 2, 'predicted', 'occident', 'liverwinib')  
# surface plot with color gradient and color bar  
ml = FigML4Dmatplotlib('x', 'y', 'predicted', 'liverwinib')  
p = ml.plot_dataframe('x', 'y', 'predicted', 'occident', 'liverwinib', 'ml4dmatplotlib')  
ml = FigML4Dmatplotlib('x', 'y', 'predicted', 'liverwinib')
```

## Google Cloud Platform

### CLOUD DATALAB<sup>BETA</sup>

An easy to use interactive tool for large-scale data exploration, analysis, and visualization.

**TRY IT FREE**

### Powerful Data Exploration

Cloud Datalab is a powerful interactive tool created to explore, analyze and visualize data with a single click on Google Cloud Platform. It runs on Google App Engine and orchestrates multiple services automatically so you can focus on exploring your data.

## Google Cloud Datalab

Notebooks Sessions

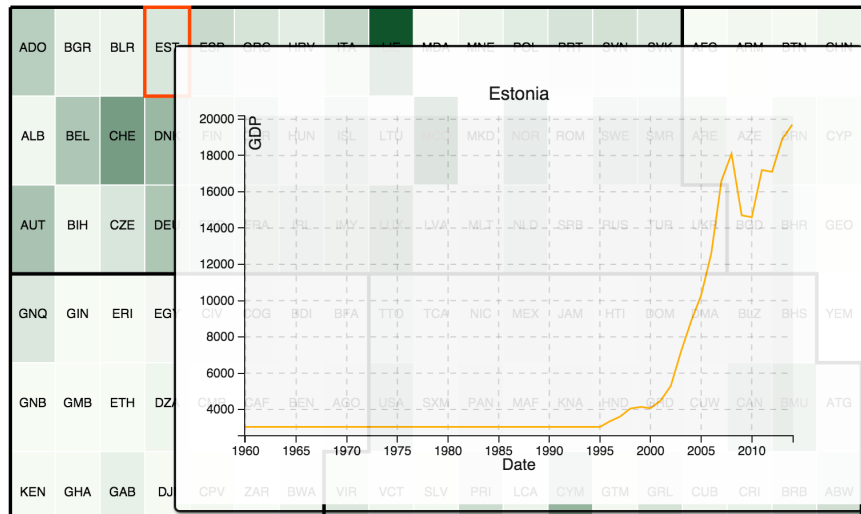
+ Notebook + Folder Upload git Repository

- datalab / intro
- ..
- Introduction to Notebooks.ipynb Running
- Introduction to Python.ipynb
- Working with Datalab.ipynb

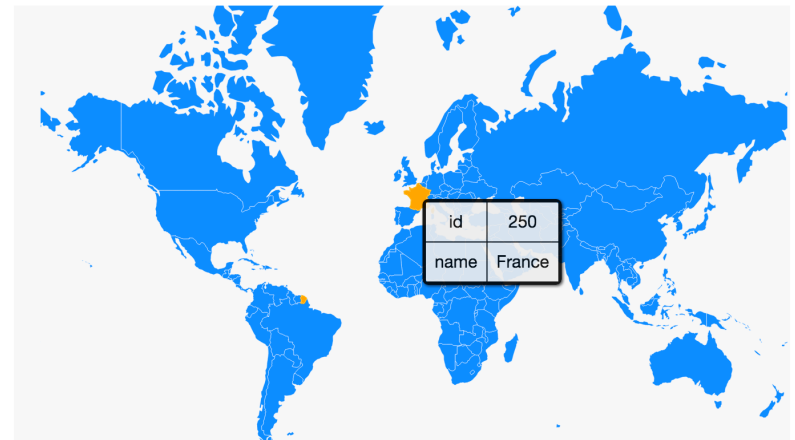


# bqplot: interactive plotting from Bloomberg

```
# Custom msg sent when a particular cell is hovered on  
market_map.on_hover(hover_handler)  
display(market_map)
```



```
def_tt = Tooltip(fields=['id', 'name'])  
map_mark = Map(map_data=topo_load('WorldMapData.json'),  
               scales={'projection': Mercator(), tooltip=def_tt})  
map_mark.interactions = {'click': 'select', 'hover': 'tooltip'}  
Figure(marks=[map_mark])
```



# O'Reilly Thebe: kernels as a service

Embracing Jupyter Notebooks at O'Reilly

O'Reilly Media is using our Atlas platform to make Jupyter Notebooks a first-class authoring environment for our publishing program.

By Andrew Odewahn, May 7, 2015

### Embracing Jupyter Notebooks at O'Reilly

O'Reilly Media is thrilled to announce that we're making IPython Notebooks a first-class authoring environment for our publishing program, on par with Word or our Atlas platform. As part of our move to

Data visualization with Seaborn

There are other parameters which can be passed to `jointplot`: for example, we can use a hexagonally-based histogram instead:

```
with sns.axes_style('white'):  
    sns.jointplot("x", "y", data, kind='hex')
```

pearsonr = 0.64; p = 4.7e-231

done



---

# Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science

---

## ❖ Interactive Computing

- ❖ Notebooks as interactive applications
- ❖ Modular, reusable UI/UX
- ❖ Software engineering with notebooks

## ❖ Computational Narratives

- ❖ nbconvert
- ❖ Element filtering
- ❖ Documentation

## ❖ Collaboration

- ❖ Real time collaboration
- ❖ JupyterHub

## ❖ Sustainability

- ❖ People
- ❖ Events



---

# UI refactor: Bloomberg/Continuum

---

- ❖ Using phosphor JS framework:
  - ❖ <https://github.com/phosphorjs/phosphor>
- ❖ Tiled layout, plugins, much more
- ❖ Enable richer layouts, beyond the notebook
  - ❖ Text editor, output, variable inspectors, debuggers, ...



# A notable new European Collaboration



The screenshot shows a web browser window with the URL `opendreamkit.org`. The page features a blue sidebar on the left with the OpenDreamKit logo (a circle of yellow stars on a blue background) and text: "OpenDreamKit", "A project funded by the Horizon 2020 European Research Infrastructures Work Programme.", and a navigation menu including "Home", "News", "About", "Job openings", "Follow us" (with social media icons), "Edit this page", and "Currently v0.2.0". The main content area has a title: "OpenDreamKit: Open Digital Research Environment Toolkit for the Advancement of Mathematics". Below the title is a paragraph: "OpenDreamKit is a [Horizon 2020 European Research Infrastructure](#) project that will run for four years, starting from September 2015. It will provide substantial funding to the open source computational mathematics ecosystem, and in particular popular tools such as [LinBox](#), [MPIR](#), [SageMath](#), [GAP](#), [Pari/GP](#), [LMFDB](#), [Singular](#), [MathHub](#), and [the IPython/Jupyter interactive computing environment](#)." The words "the IPython/Jupyter interactive" are circled in red. Below this is another paragraph: "From this ecosystem, OpenDreamKit will deliver a flexible toolkit enabling research groups to set up [Virtual Research Environments](#), customised to meet the varied needs of research projects in pure mathematics and applications, and supporting the full research life-cycle from exploration, through proof and publication, to archival and sharing of data and code." A final paragraph states: "Altogether the project involves about 50 people spread over 15 sites in Europe, with a total budget of about 7.6 million euros. The largest portion of that will be devoted to employing an average of 11 researchers and developers working full time on the project. Additionally, the participants will contribute the equivalent of six other people working full time." At the bottom of the main content area is a link that says "Read more...".

---

# This presents challenges!

---

- ❖ Keeping a healthy community growing
- ❖ Engaging newcomers and volunteers
- ❖ Managing scope and complexity
- ❖ Documenting architecture, process and community



which brings me to...

we also need the right spaces to  
meet these challenges!



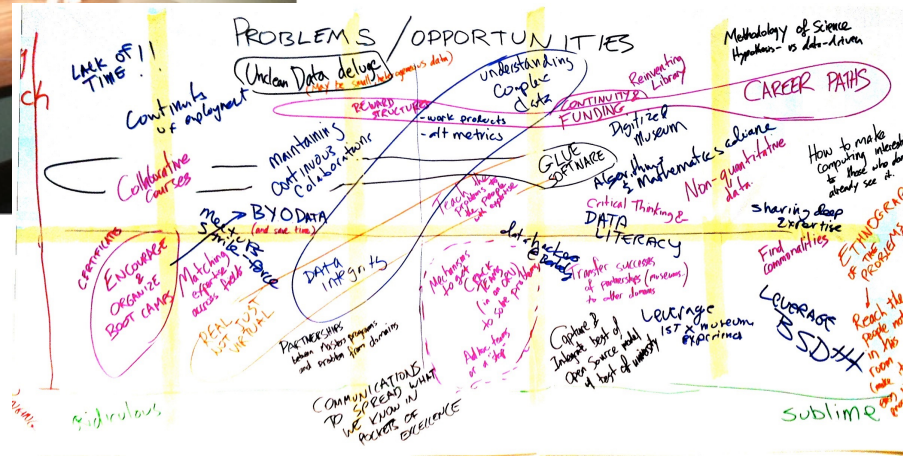
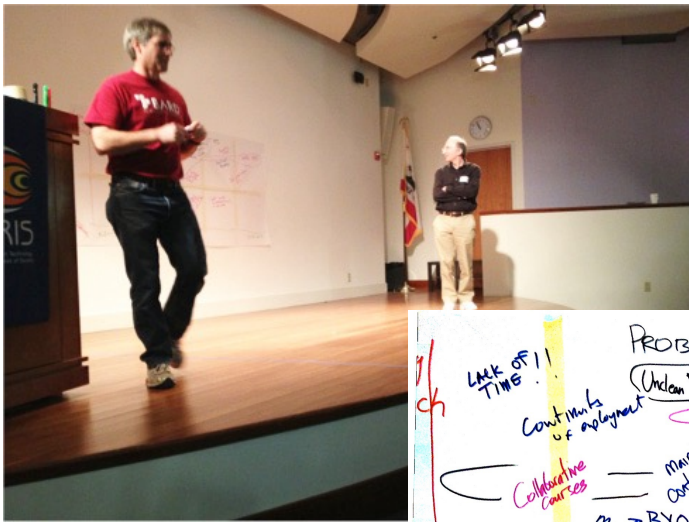


# BERKELEY INSTITUTE FOR DATA SCIENCE

Advancing scientific discovery  
through collaboration across research domains

# Great interest from across the campus

Data Science Workshop held in February 2013 was attended by 80 researchers on three days notice; with follow-up events in May and June (to date 280+ signed up for mailing list)





# Initial Faculty Group



Faculty Lead/PI: **Saul Perlmutter**, Physics, Berkeley Center for Cosmological Physics



**Joshua Bloom**, Professor, Astronomy; Director, Center for Time Domain Informatics



**Henry Brady**, Dean, Goldman School of Public Policy



**Cathryn Carson**, Associate Dean, Social Sciences; Acting Director of Social Sciences Data Laboratory "D-Lab"



**David Culler**, Professor, EECS



**Michael Franklin**, Chair, EECS, Co-Director, AMP Lab



**Erik Mitchell**, Associate University Librarian



**Fernando Perez**, Researcher, Henry H. Wheeler Jr. Brain Imaging Center



**Jasjeet Sekhon**, Professor, Political Science and Statistics; Center for Causal Inference and Program Evaluation



**Jamie Sethian**, Professor, Mathematics



**Kimmen Siölander**, Professor, Bioengineering, Plant and Microbial Biology



**Philip Stark**, Chair, Statistics



**Ion Stoica**, Professor, EECS, Co-Director, AMP Lab

A 5-year, \$37.8 million cross-institutional collaboration



— Berkeley —  
UNIVERSITY OF CALIFORNIA



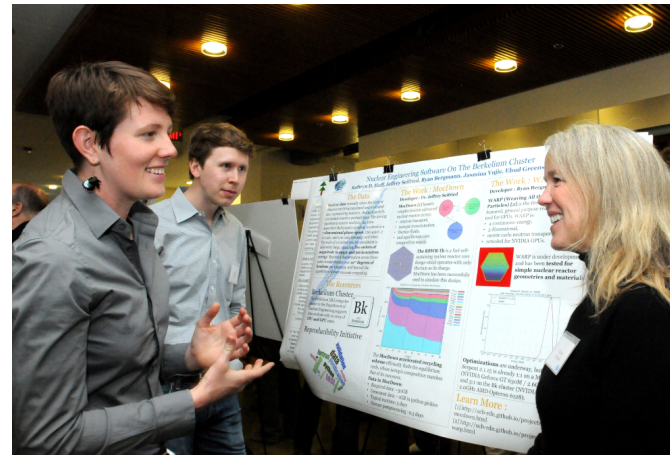
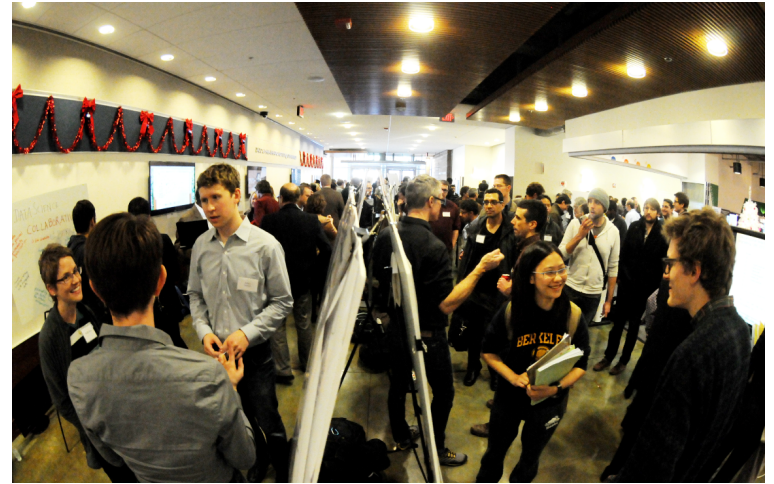
GORDON AND BETTY  
**MOORE**  
FOUNDATION



ALFRED P. SLOAN  
FOUNDATION



# Launched December 2013



# Our sponsors

- Foundations
  - Moore and Sloan Foundations \$12.5 million
- Industry
  - Siemens
  - State Street
- Institutional
  - UC Berkeley

GORDON AND BETTY  
**MOORE**  
FOUNDATION



**SIEMENS**



STATE STREET®

**Berkeley**  
UNIVERSITY OF CALIFORNIA

BERKELEY

Institute for  
Data Science



# BIDS Goals

- Support meaningful and sustained interactions and collaborations between
  - Science domains: life science, social science, physical science
  - Methodology fields: computer science, statistics, applied mathematics
- Establish new Data Science career paths that are long-term and sustainable
  - A generation of multi-disciplinary scientists in data-intensive science
  - A generation of data scientists focused on tool development
- Build an ecosystem of analytical tools and research practices
  - Sustainable, reusable, extensible, easy to learn and to translate across research domains
  - Enables scientists to spend more time focusing on their science

We have computing power, we have applied math techniques, we have database approaches, so...

What's missing?

BERKELEY

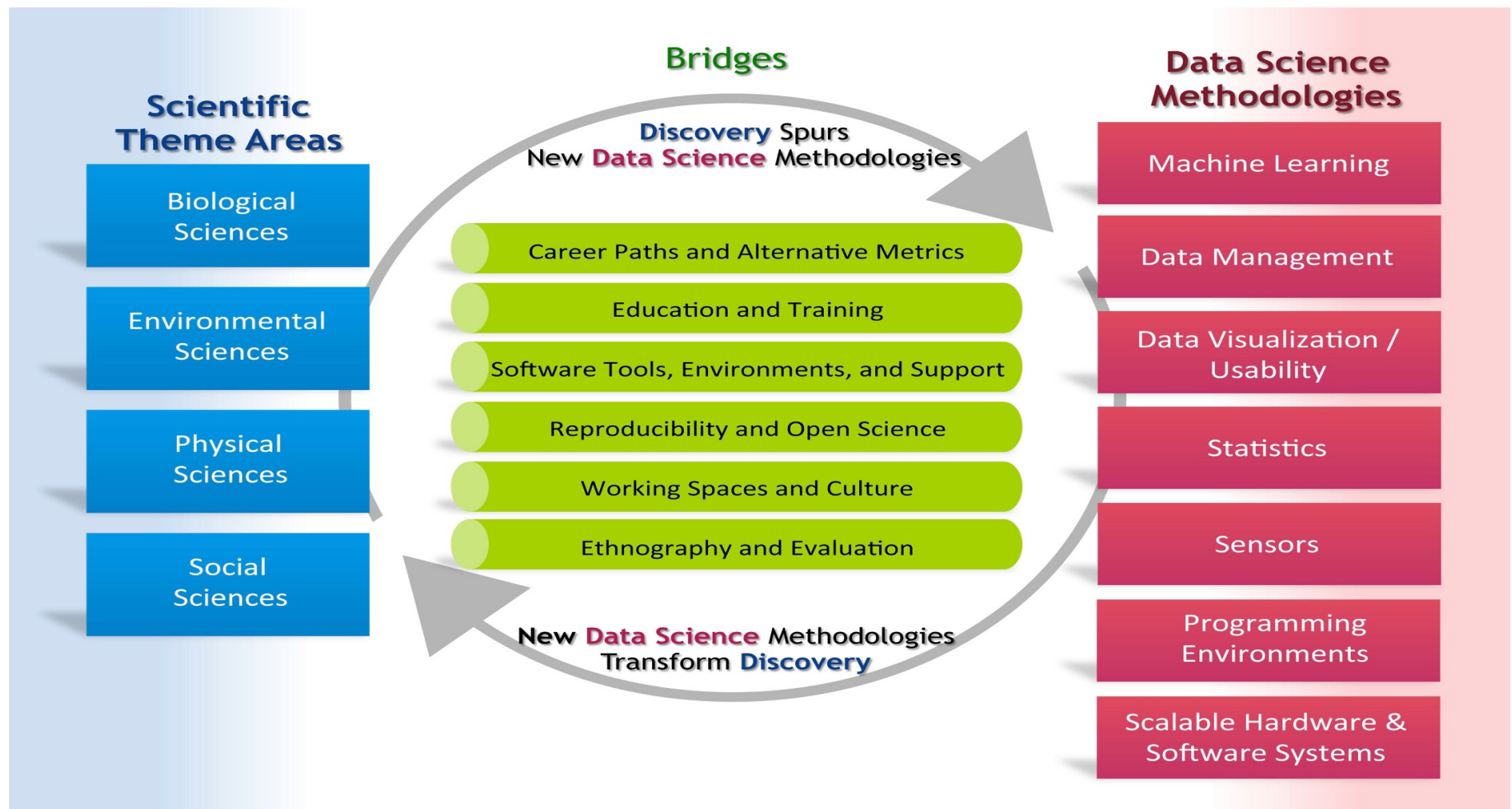
Institute for  
Data Science





# Working Groups

Working to address the major challenges facing major advances in data driven research.

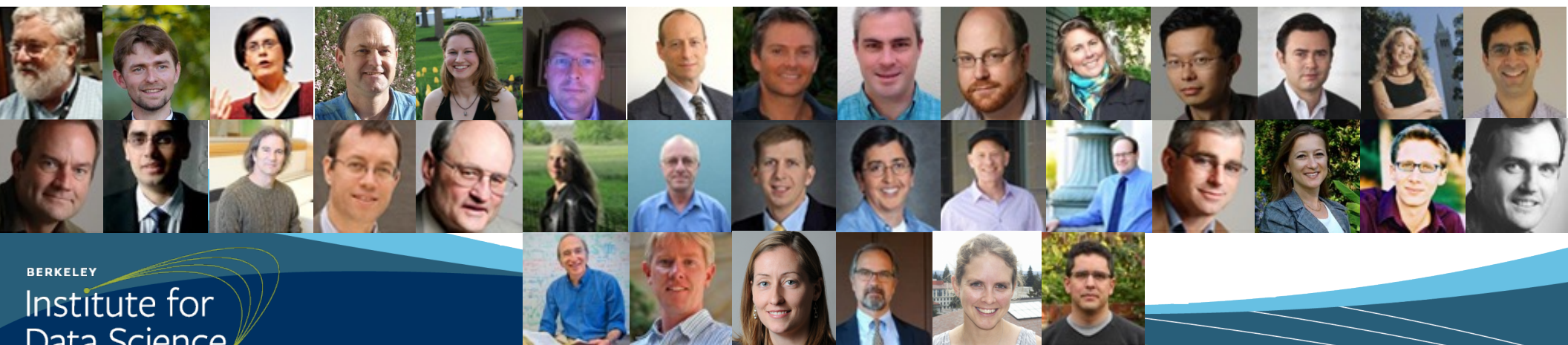


# People are at the heart of BIDS



Structured roughly as:

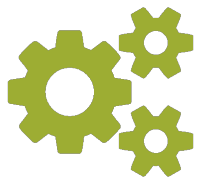
- Competitive fellow positions, typically 50/50 appts, 2-3 y.
- A few full-time positions (tools, ethnography, staff).
- Senior fellows: campus faculty and LBNL scientists.





# Diverse expertise

- Sociology
- Phylogenomics
- Cosmological Physics
- Nuclear Science
- Neuroscience
- Energy and Resources
- System software
- High-performance computing
- Global Change Biology
- Geospatial
- Statistics
- Environmental science
- Computer Vision
- Distributed computing
- Seismology
- Computer Science
- Astronomy
- Public Policy
- Social Sciences
- Psychology
- Library science
- Molecular & Cell Biology
- Political Science
- Mathematics
- Bioengineering
- City & Regional Planning
- ...



# Diverse Software Development

<http://bids.berkeley.edu/research>

BIDS Fellows engage in a range of projects that address the ongoing needs of effectively advancing data-intensive research.



**HOLOS**  
BERKELEY ECOINFORMATICS ENGINE

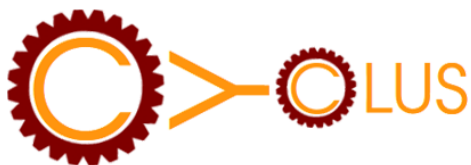


The **DECIDING FORCE** Project



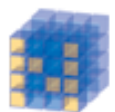
**scikit-image**  
image processing in python

**IP[y]:**  
IPython



**txt\_thrshr**

**NIMBLE**



**NumPy**



**astro.py**

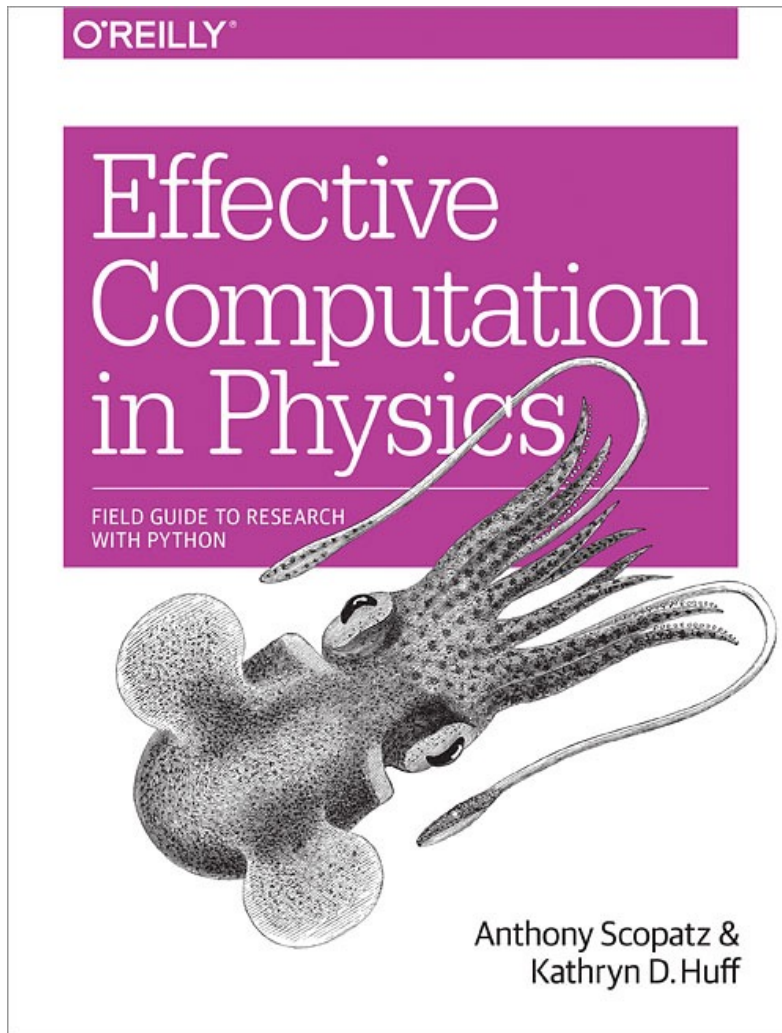
A Community Python Library for Astronomy

**R** OpenSci

BERKELEY

Institute for  
Data Science





# Katy Huff

Nuclear Engineering Postdoc  
BIDS Data Science Fellow



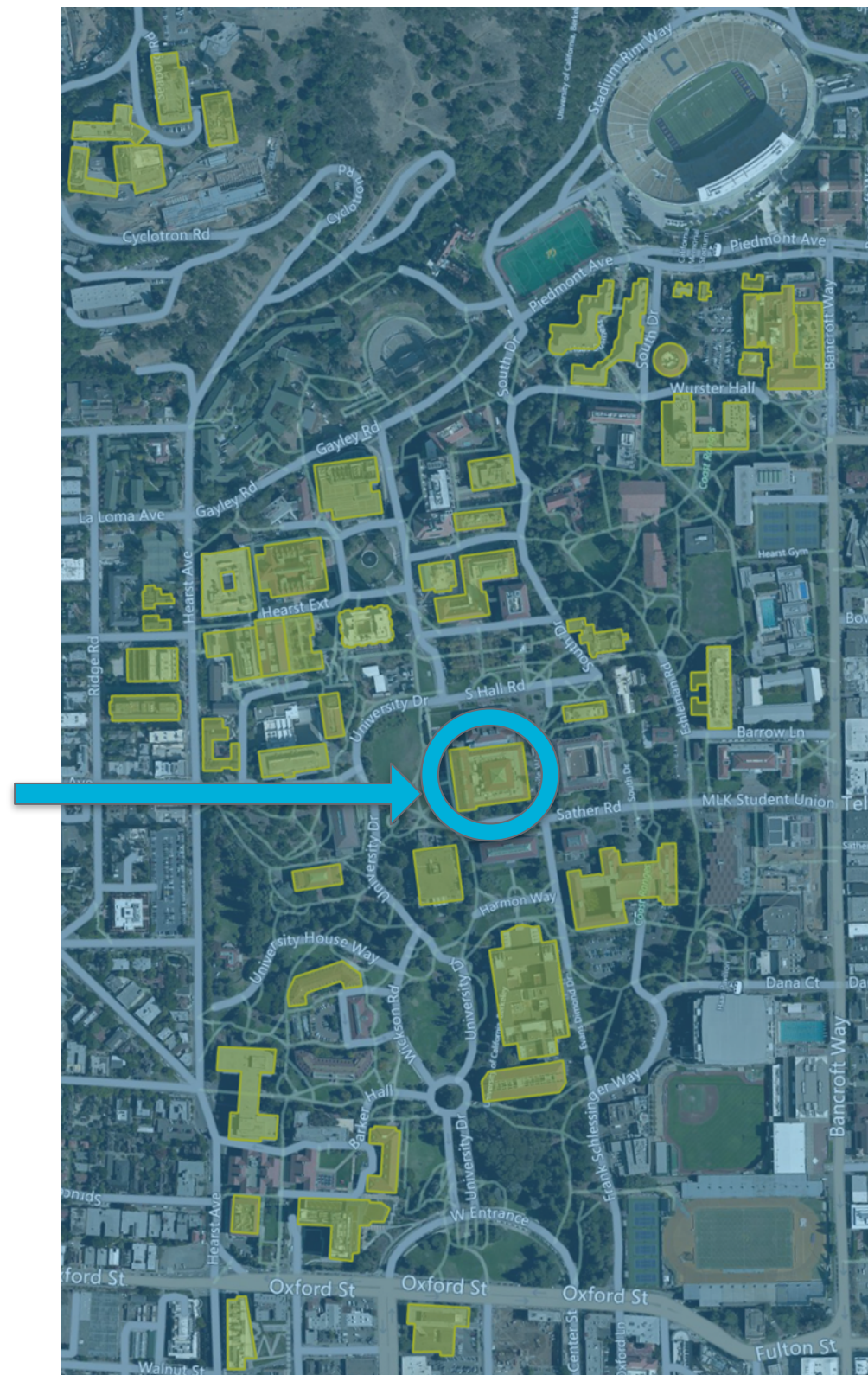
- [physics.codes](https://physics.codes)
- [github.com/physics.codes/examples](https://github.com/physics.codes/examples)
- [shop.oreilly.com/product/o636920033424.do](https://shop.oreilly.com/product/o636920033424.do)



# Our collaborative space

## 190 Doe Library

Central location that serves as home for data science efforts





# Our collaborative space

190 Doe Library





---

# BIDS Events

---



Data Science Lecture Series



BIDS Tea



The Hacker Within



Data Science Collaborative



Data Science Faire



Technical workshops (Spark, Azure,  
Data Structures for Data Science, ...)



---

# Joining threads...

---

- ❖ **Project Jupyter**

- ❖ Growing like crazy...
- ❖ We think the next few years will be very interesting
- ❖ But we know the toughest challenges aren't technical!

- ❖ **BIDS**

- ❖ is the kind of home I've always wanted for a project like Jupyter
- ❖ it's still an experiment, evolving and changing
- ❖ will change a lot in the next year, as it moves from bootstrapping to forming its own research identity...
- ❖ ...and the broader discussion of Data Science @ Berkeley evolves...

Jupyter Project is hiring at Berkeley!

Two postdocs

Project Manager



# Thank You!

@fperez\_org fperez@lbl.gov

@ProjectJupyter @IPythonDev

Try it out at  
[try.jupyter.org](http://try.jupyter.org)