

IP[y]:  
IPython



---

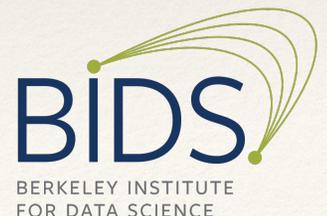
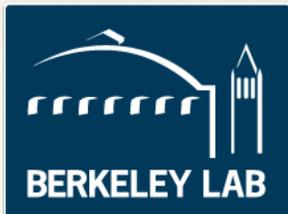
# Project Jupyter

Opening cultures, from  
science to data-driven  
journalism

---

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 Lifecycle of a Scientific Idea (schematically)

---

1. **Individual** exploratory work
2. **Collaborative** development
3. **Parallel** production runs (HPC, cloud, ...)
4. **Publication & communication** (reproducibly!)
5. **Education**
6. Goto 1

---

# The Lifecycle of a Scientific Idea (schematically)

---

1. **Individual** exploratory work
2. **Collaborative** development
3. **Parallel** production runs (HPC, cloud, ...)
4. **Publication & communication** (reproducibly!)
5. **Education**
6. Goto 1

**We treat this as a single, coherent problem**

**What does this have to do with  
Journalism???**

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

*–Hamming'62*

---

# Project Jupyter: tools for...

---

- ❖ Interactively exploring computational problems:
  - ❖ *Insight comes to the human, not to the machine!*
- ❖ Communicating and sharing these insights
  - ❖ *Computational Narratives: Code, Data & Results telling a story together.*

**“Literate computing” and  
computational reproducibility: IPython  
in the age of data-driven journalism**

<http://blog.fperez.org/2013/04/literate-computing-and-computational.html>

# Reinhart & Rogoff: we all make mistakes

Reinhart, Rogoff... and Her x

www.bbc.com/news/magazine-22223190

Magazine

## Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander  
BBC News

© 20 April 2013 | Magazine

**This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.**

It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.



# Ping the internet...



**Fernando Perez** @fperez\_org · 17 Apr 2013

Economics experts to turn analysis from **Herndon**, Ash & Pollin into IPython notebook? Data-driven journalism @jseabold [peri.umass.edu/236/hash/31e2f...](http://peri.umass.edu/236/hash/31e2f...)

← ↻ 5 🍷 5 || ...



**skipper seabold**

@jseabold

18 Apr 13

@fperez\_org Sounds like a Sunday project to me.



**Vincent Arel-Bundock**

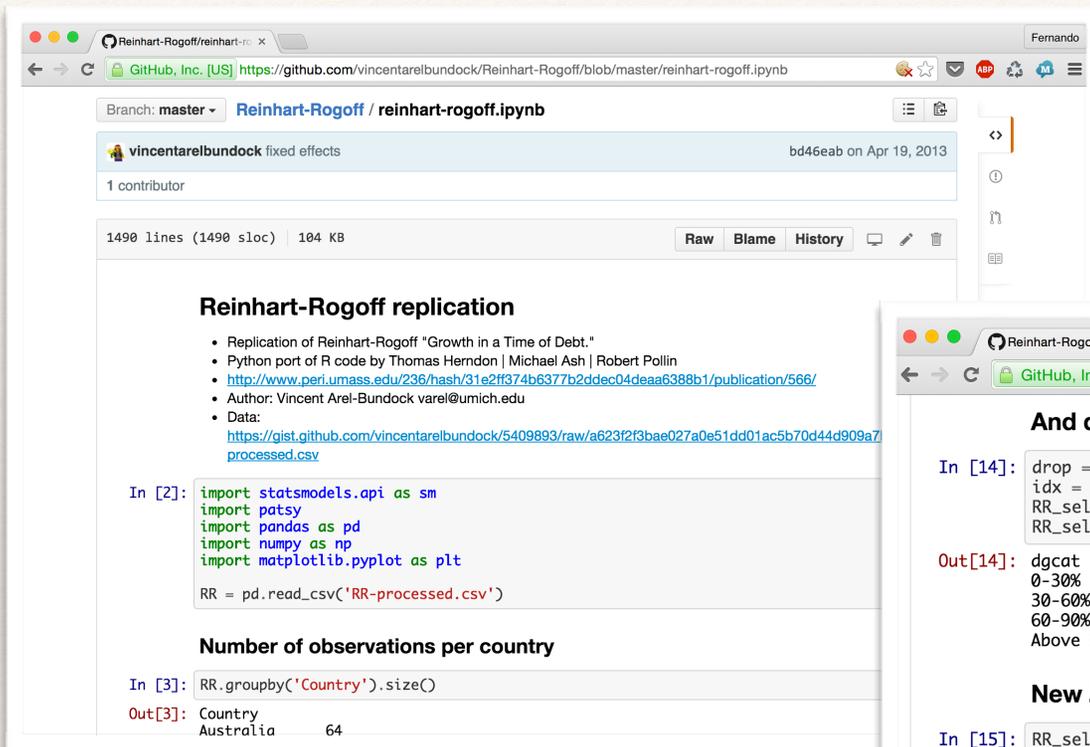
@VincentAB

Follow

@jseabold @fperez\_org here you go. only things missing: loess & linear hypo. code could be cleaner, but hey, it works [nbviewer.ipython.org/5409848](http://nbviewer.ipython.org/5409848)

5:30 AM - 18 Apr 2013

# And @VincentAB delivers...



Reinhart-Rogoff/reinhardt-rc x Fernando

GitHub, Inc. [US] <https://github.com/vincentarelbundock/Reinhart-Rogoff/blob/master/reinhart-rogoff.ipynb>

Branch: master Reinhart-Rogoff / reinhart-rogoff.ipynb

vincentarelbundock fixed effects bd46eab on Apr 19, 2013

1 contributor

1490 lines (1490 sloc) | 104 KB Raw Blame History

### Reinhart-Rogoff replication

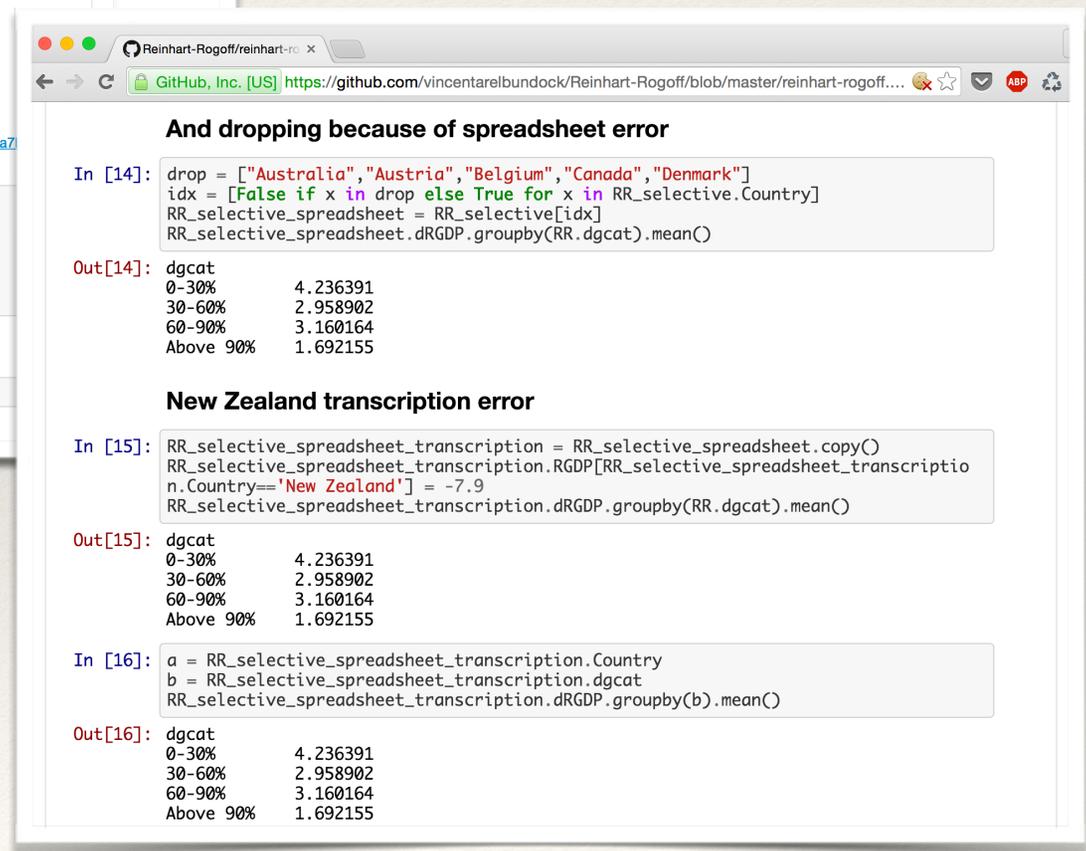
- Replication of Reinhart-Rogoff "Growth in a Time of Debt."
- Python port of R code by Thomas Herndon | Michael Ash | Robert Pollin
- <http://www.peri.umass.edu/236/hash/31e2ff374b6377b2ddec04deaa6388b1/publication/566/>
- Author: Vincent Arel-Bundock varel@umich.edu
- Data: <https://gist.github.com/vincentarelbundock/5409893/raw/a623f2f3bae027a0e51dd01ac5b70d44d909a7processed.csv>

```
In [2]: import statsmodels.api as sm
import patsy
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

RR = pd.read_csv('RR-processed.csv')
```

### Number of observations per country

```
In [3]: RR.groupby('Country').size()
Out[3]: Country
Australia    64
```



### And dropping because of spreadsheet error

```
In [14]: drop = ["Australia", "Austria", "Belgium", "Canada", "Denmark"]
idx = [False if x in drop else True for x in RR_selective.Country]
RR_selective_spreadsheet = RR_selective[idx]
RR_selective_spreadsheet.dRGDP.groupby(RR.dgcat).mean()
```

```
Out[14]: dgcat
0-30%      4.236391
30-60%     2.958902
60-90%     3.160164
Above 90%  1.692155
```

### New Zealand transcription error

```
In [15]: RR_selective_spreadsheet_transcription = RR_selective_spreadsheet.copy()
RR_selective_spreadsheet_transcription.RGDP[RR_selective_spreadsheet_transcription.Country=='New Zealand'] = -7.9
RR_selective_spreadsheet_transcription.dRGDP.groupby(RR.dgcat).mean()
```

```
Out[15]: dgcat
0-30%      4.236391
30-60%     2.958902
60-90%     3.160164
Above 90%  1.692155
```

```
In [16]: a = RR_selective_spreadsheet_transcription.Country
b = RR_selective_spreadsheet_transcription.dgcat
RR_selective_spreadsheet_transcription.dRGDP.groupby(b).mean()
```

```
Out[16]: dgcat
0-30%      4.236391
30-60%     2.958902
60-90%     3.160164
Above 90%  1.692155
```

# Demo - Live Notebook

A quick recap of history

# 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]: □
```

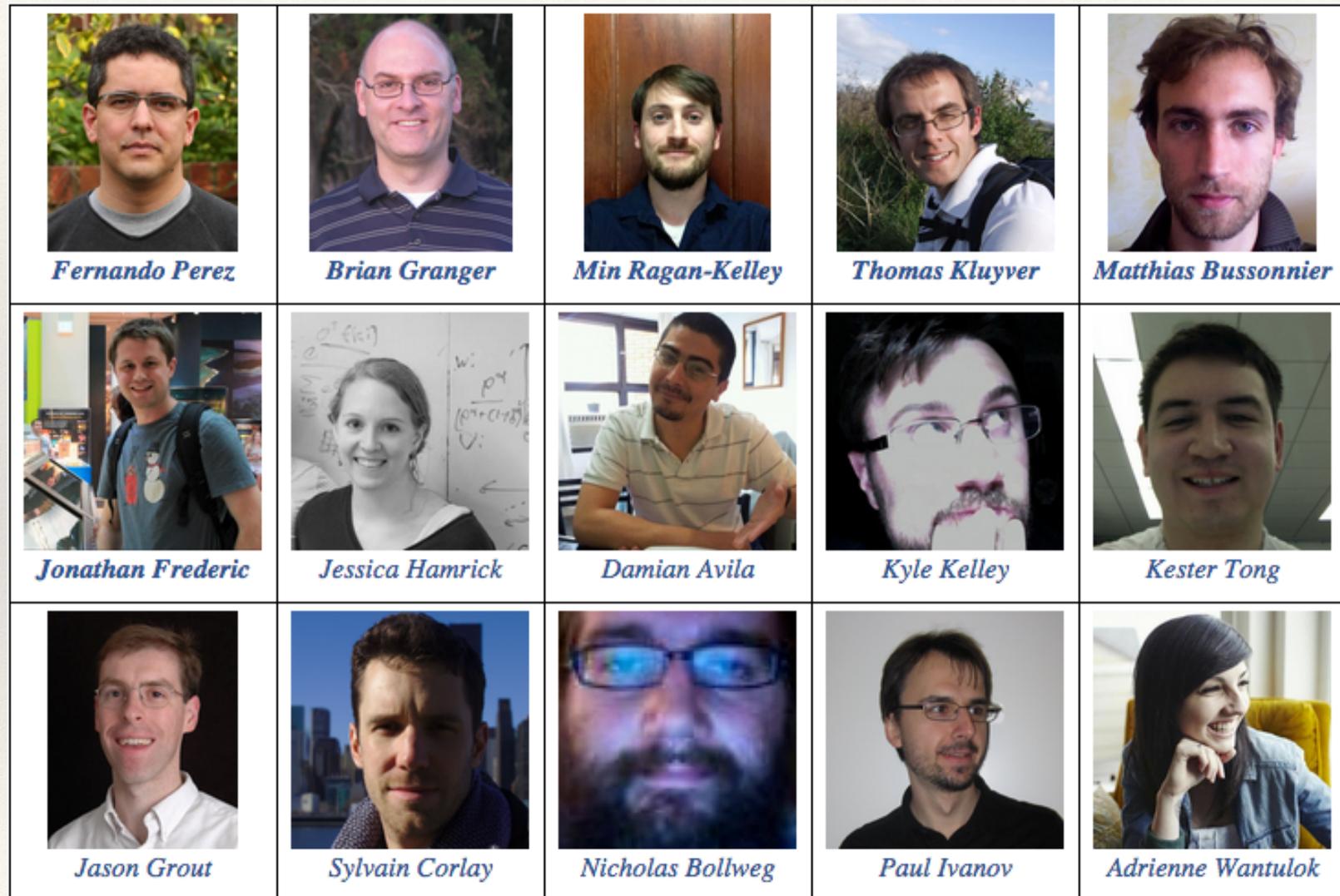
## November 2001: "Just an afternoon hack"

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

## In 2014 (Openhub stats)

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

# Today, a rapidly growing community



Plus ~ 500 more Open source contributors!

# Current and recent funding



**ALFRED P. SLOAN  
FOUNDATION**

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

GORDON AND BETTY  
**MOORE**  
FOUNDATION

SIMONS FOUNDATION



**CONTINUUM**  
ANALYTICS

**POWERED BY**  
**rackspace**<sup>®</sup>  
*the open cloud company*



**Microsoft**

 **ENTHOUGHT**  
SCIENTIFIC COMPUTING SOLUTIONS

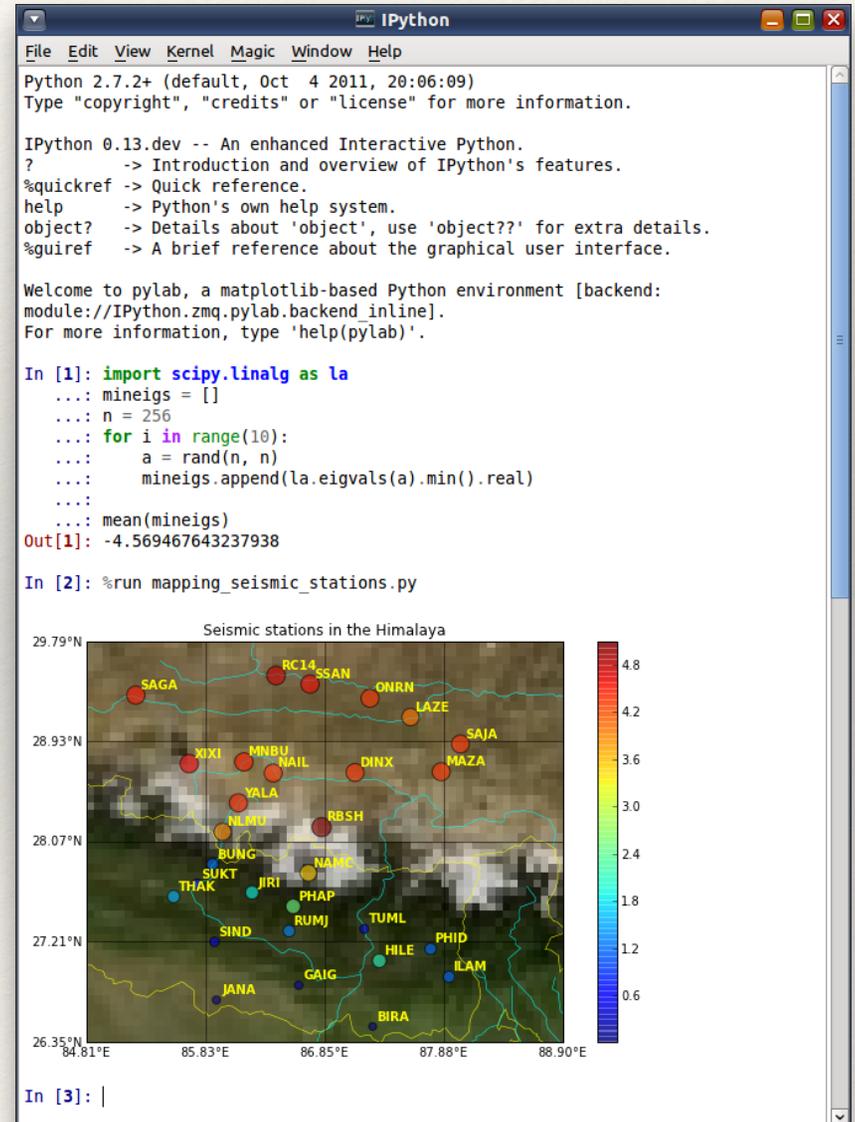
**Google**

**Bloomberg**

# Beyond the Terminal...

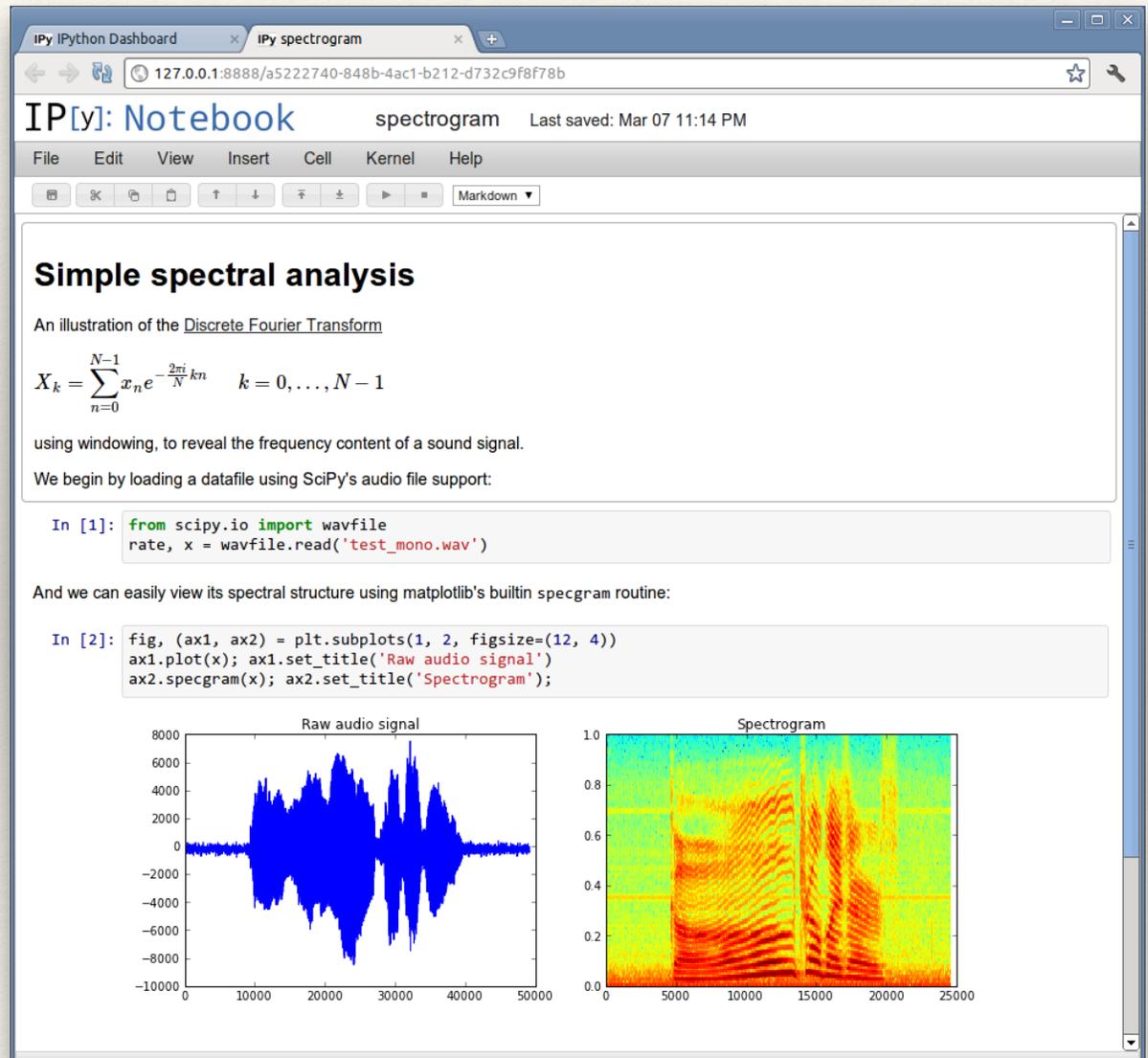
- ❖ The REPL as a network protocol
- ❖ Kernels
  - ❖ execute code
- ❖ Clients
  - ❖ Read input
  - ❖ Present output

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');
```

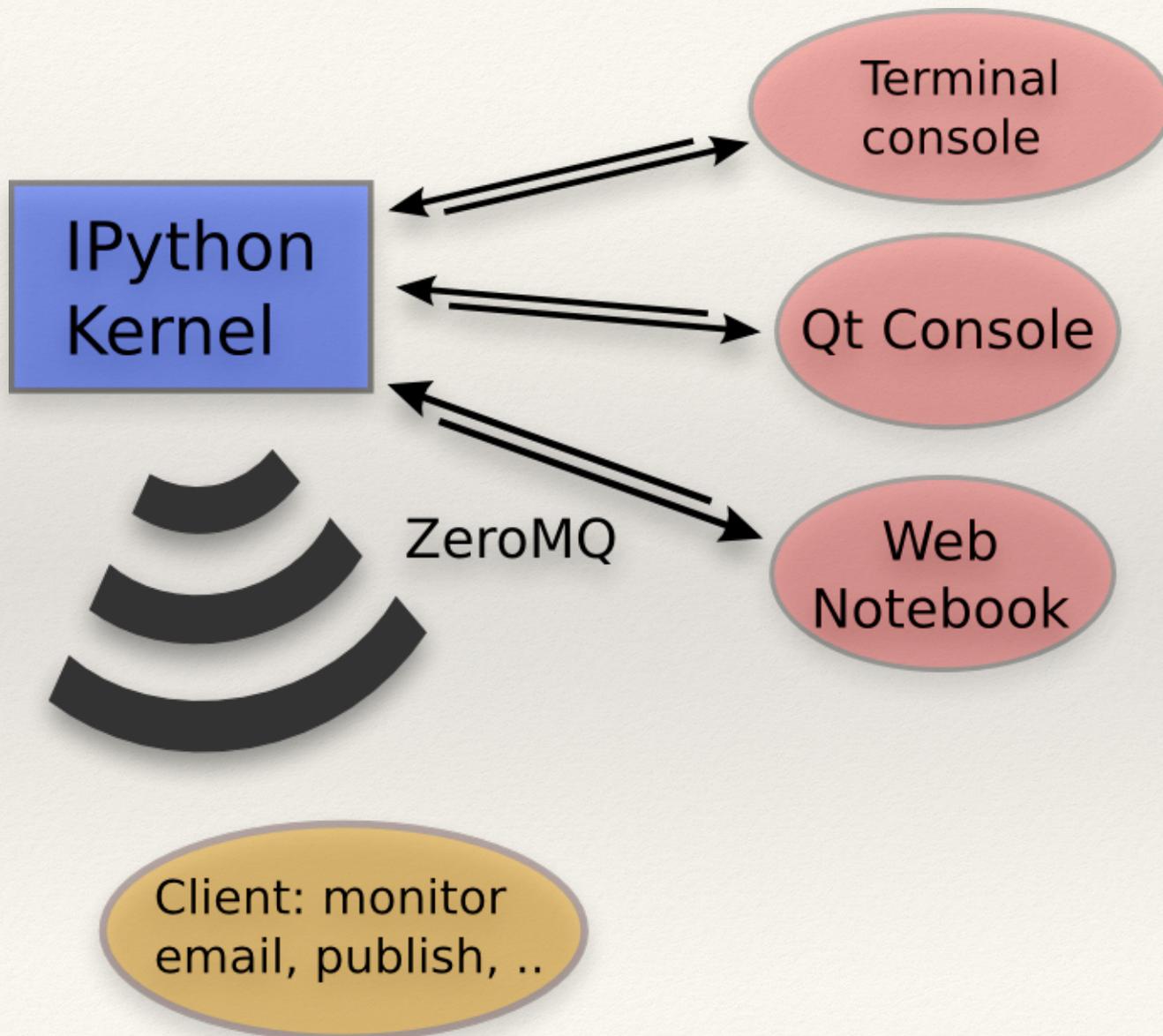
The results are displayed as two plots side-by-side. The left plot, titled "Raw audio signal", shows a blue waveform of the audio signal over time, with the x-axis ranging from 0 to 50,000 and the y-axis from -10,000 to 8,000. The right plot, titled "Spectrogram", shows a heatmap of the signal's frequency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

# 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! ~37 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:

- Programming Languages:** This section features three columns. The first column is for 'IPython', showing a code snippet and the IPython logo. The second column is for 'IRuby', featuring a red gem icon and the text 'IRuby: Notebook'. The third column is for 'Julia', with a 'Basic Julia interaction' section and the Julia logo.
- Books:** This section displays three book covers: 'Python for Signal Processing' (a dark blue and orange cover), 'Mining the Social Web' (a red cover, 2nd Edition, by O'Reilly), and 'Probabilistic Programming & Bayesian Methods for Hackers' (a light purple cover with a city skyline).
- Misc:** This section is divided into three sub-sections: 'Data Visualization with Lightning' (showing various plots), 'Interactive data visualization with Bokeh' (showing a grid of interactive plots), and 'Interactive plots with Plotly' (showing various interactive plots).

# Reproducible Research

The screenshot shows the ISME Journal website. The browser address bar displays the URL: [www.nature.com/ismej/journal/v7/n3/full/ismej2012123a.html](http://www.nature.com/ismej/journal/v7/n3/full/ismej2012123a.html). The page header includes the journal title "The ISME Journal" and subtitle "Multidisciplinary Journal of Microbial Ecology". A search bar is present with the text "Search" and "Advanced search". The main content area is titled "Commentary" and features the article "Collaborative cloud-enabled tools allow rapid, reproducible biological insights" by Benjamin Ragan-Kelley et al. The article is published in *The ISME Journal* (2013) 7, 461–464. A sidebar on the left contains navigation links such as "Journal home", "Advance online publication", "Current issue", "Archive", "Focuses", "Browse by subject", "Press releases", "Online submission", "For authors", "For referees", "Contact editorial office", "About the journal", "Editors and Editorial Board", "About the society", and "For librarians". A sidebar on the right lists options for the full text, including "Download PDF", "Send to a friend", "View interactive PDF in ReadCube", "Rights and permissions", "Order Commercial Reprints", "CrossRef lists 1 article citing this article", "Data availability", "References", "Acknowledgements", "Figures and Tables", "Supplementary info", "Export citation", "Export references", and "Papers by Ragan-Kelley".

Journal home > Archive > Commentaries > Full text

**Journal home**  
**Advance online publication**  
About AOP  
**Current issue**  
**Archive**  
**Focuses**  
**Browse by subject**  
**Press releases**  
Online submission  
For authors  
For referees  
Contact editorial office  
About the journal  
Editors and Editorial Board  
About the society  
For librarians

**Commentary**

*The ISME Journal* (2013) **7**, 461–464; doi:10.1038/ismej.2012.123; published online 25 October 2012

**Collaborative cloud-enabled tools allow rapid, reproducible biological insights**  
*Open*

Benjamin Ragan-Kelley<sup>1,12</sup>, William Anton Walters<sup>2,12</sup>, Daniel McDonald<sup>3,6,12</sup>, Justin Riley<sup>4</sup>, Brian E Granger<sup>5</sup>, Antonio Gonzalez<sup>6</sup>, Rob Knight<sup>7,8</sup>, Fernando Perez<sup>9</sup> and J Gregory Caporaso<sup>10,11</sup>

<sup>1</sup>Graduate Group in Applied Science and Technology, University of California at Berkeley, Berkeley, CA, USA  
<sup>2</sup>Department of Molecular, Cellular and Developmental Biology, University of Colorado at Boulder, Boulder, CO, USA  
<sup>3</sup>Biofrontiers Institute, University of Colorado at Boulder, Boulder, CO, USA  
<sup>4</sup>Office of Educational Innovation and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA  
<sup>5</sup>Physics Department, California Polytechnic State University, San Luis Obispo, CA, USA  
<sup>6</sup>Department of Computer Science, University of Colorado at Boulder, Boulder, CO, USA

**FULL TEXT**  
Previous | Next  
Table of contents  
Download PDF  
Send to a friend  
View interactive PDF in ReadCube  
Rights and permissions  
Order Commercial Reprints  
CrossRef lists 1 article citing this article  
Data availability  
References  
Acknowledgements  
Figures and Tables  
Supplementary info  
Export citation  
Export references  
Papers by Ragan-Kelley

<http://www.nature.com/ismej/journal/v7/n3/full/ismej2012123a.html>

# Paper, Notebooks and Virtual Machine

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/1046r = V6
# 1391f/1492r = V9

primer_seqs = {
    '27f': 'AGAGTTTGATCTMTGGCTCAG',
    '338r': DNA.rc('GCTGCCCTCCCGTAGGAGT'),
    '349f': 'GYGCASCAGKCGMGAAN',
    '534r': DNA.rc('ATTACCGCGGCTGCTGG'),
    '515f': 'GTGCCAGCMGCCGCGGTAA',
    '806r': DNA.rc('GGACTACVSGGGTATCTAAT'),
    '967f': 'CAACGCGAAGAACCTTACC',
    '1048r': DNA.rc('CGRCRCGCATGYACCWC'),
    '1391f': 'TGYACACACCGCCCGTC',
    '1492r': DNA.rc('GGCTACCTTGTACGACTT'),
    '1391r': 'TGYACACACCGCCCGTC' # 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'
```

Instructions and supporting data for the QIIME/IPython/StarCluster demo at the 2012 NIH Cloud Computing the Microbiome workshop and our corresponding paper in the ISME Journal.

The analysis made use of the [IPython Notebook](#), [QIIME](#), [StarCluster](#), [PyCogent](#), and [PrimerProspector](#). All of these tools are pre-installed in the ami-9f69c1f6 public Amazon EC2 instance, which was used in this study.

## Supporting Files

The IPython notebooks supporting this study can be viewed [here](#) and are available here in PDF format:

- [NIH Cloud Demo \(Complete\)](#)
- [NIH Cloud Demo \(Fast\)](#)
- [Timing\\*](#)
- [Variable Region Position Boundaries](#)
- [Pearson v Robinson-Foulds Distances](#)
- [V3 and V4 Regions Only](#)

\* Note that the Timing notebook is for reference as related to the paper only - it will not be directly reproducible on re-runs of the above notebooks as it relies on the semi-manual creation of the tasks.log file. The tasks.log file used to generate the original timing data is available for [download here](#).

The Greengenes reference OTU collection used in this study is available for [download here](#).

The IPython notebook files (.ipynb) are available for [download here](#).

The tree metadata mapping file used in generating the coloring categories in the 3D PCoA plot is [available here](#).

The paper for this analysis, "Collaborative cloud-enabled tools allow rapid, reproducible biological insights", is available [here](#).

## Reproducing the analysis

Four m2.4xlarge instances were booted using StarCluster to create a 32 core cluster with approximately 280GB of RAM (70GB per 8 core instance). This was used for the full analysis (a more complete analysis then was done during the workshop, where the workshop analysis was optimized to run quickly). To support the large quantity of data that is generated during the analysis, you should create an EBS volume which will be attached to the running instance. A 20 GB volume will be sufficient. The volume used for running these notebooks is available as snap-75eb8005.

To reproduce the analyses presented in this paper you should install StarCluster locally, and configure it according to the [instructions on the StarCluster website](#). You can then add the following to your ~/.starcluster/config file:

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

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

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

# Scientific Blogging

SCIENTIFIC AMERICAN™

Sign In | Register

Search ScientificAmerican.com

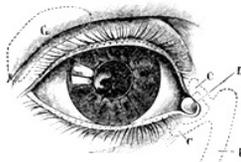
Subscribe News & Features Topics Blogs Videos & Podcasts Education Cit

SA en español

Blogs

About

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



SA Visual

Illustrating science since 1845

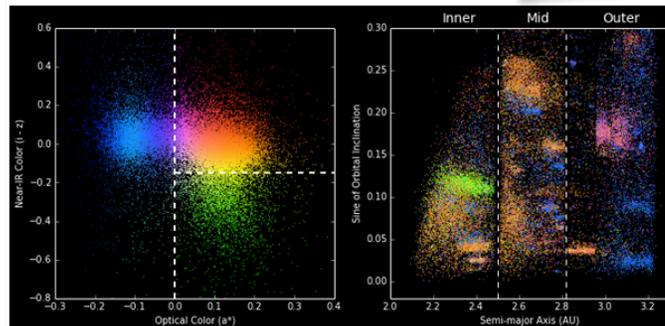
SA Visual Home About Contact

## 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 for 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:

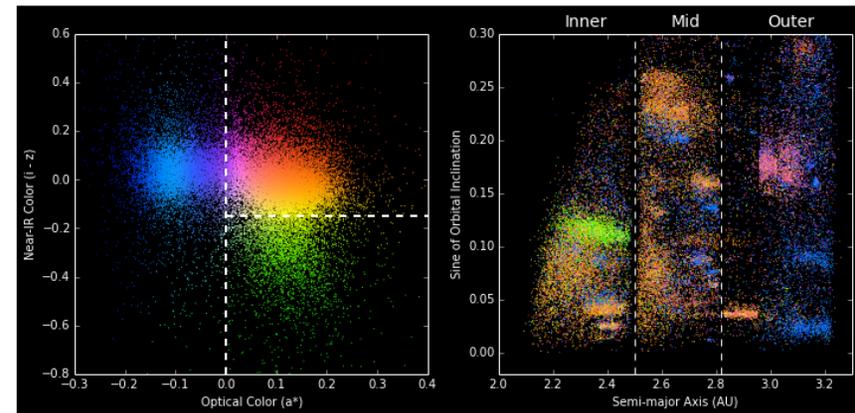


nbviewer.ipynb.org/github/jakevdp/SciAmBlogPost/blob/master/AsteroidVis.ipynb

### 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()
```



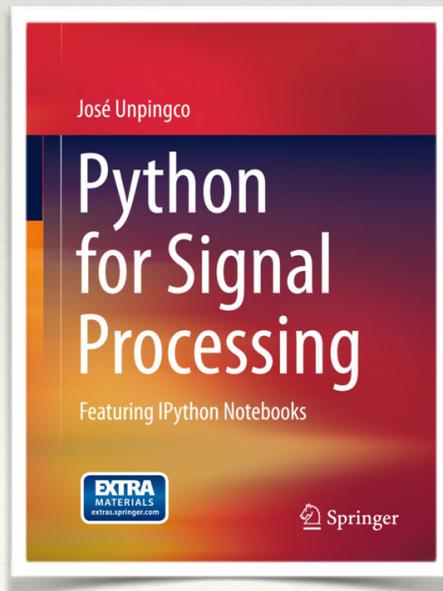
Jake van der Plas @ UW

<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



**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!

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]:

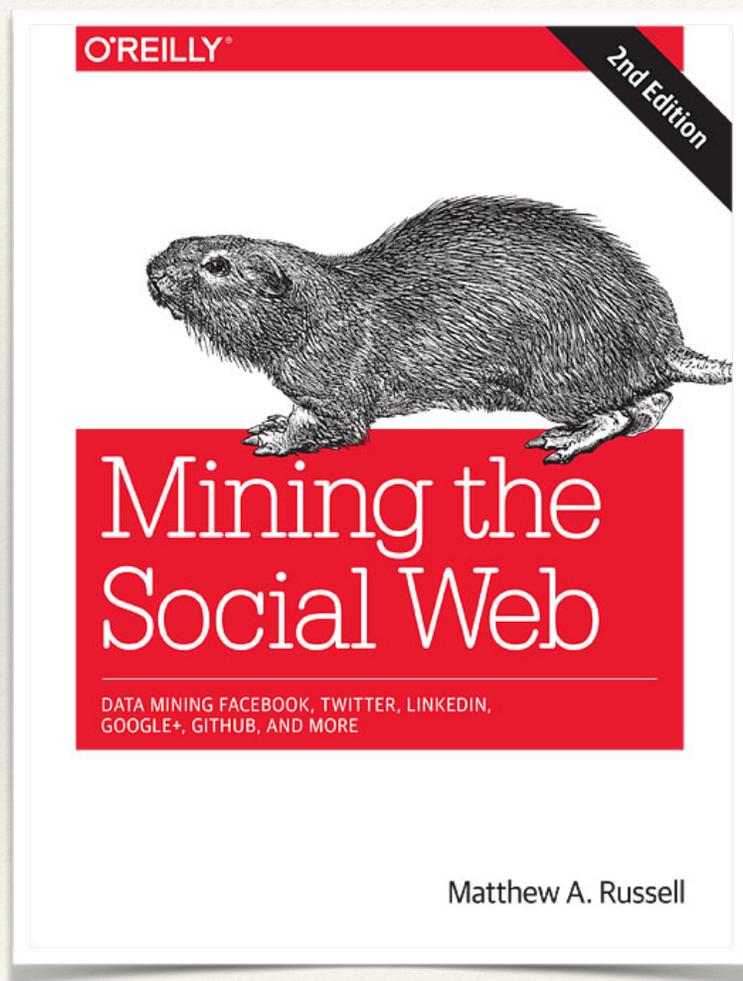
```
def abs_sinc(k=None, N=64, Ns=32):  
    if k is None: k = range(0, N-1)  
    y = where(k == 0, 1.0e-20, k)  
    return abs(sin(Ns*2*pi/N*y)/sin(2*pi*y/N))/sqrt(N)  
  
fig, ax = subplots()  
fig.set_size_inches((8, 3))  
  
ax.plot(abs_sinc(N=512, Ns=10), label='duration=10')  
ax.plot(abs_sinc(N=512, 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);
```

**Rectangular Windows DFTs**

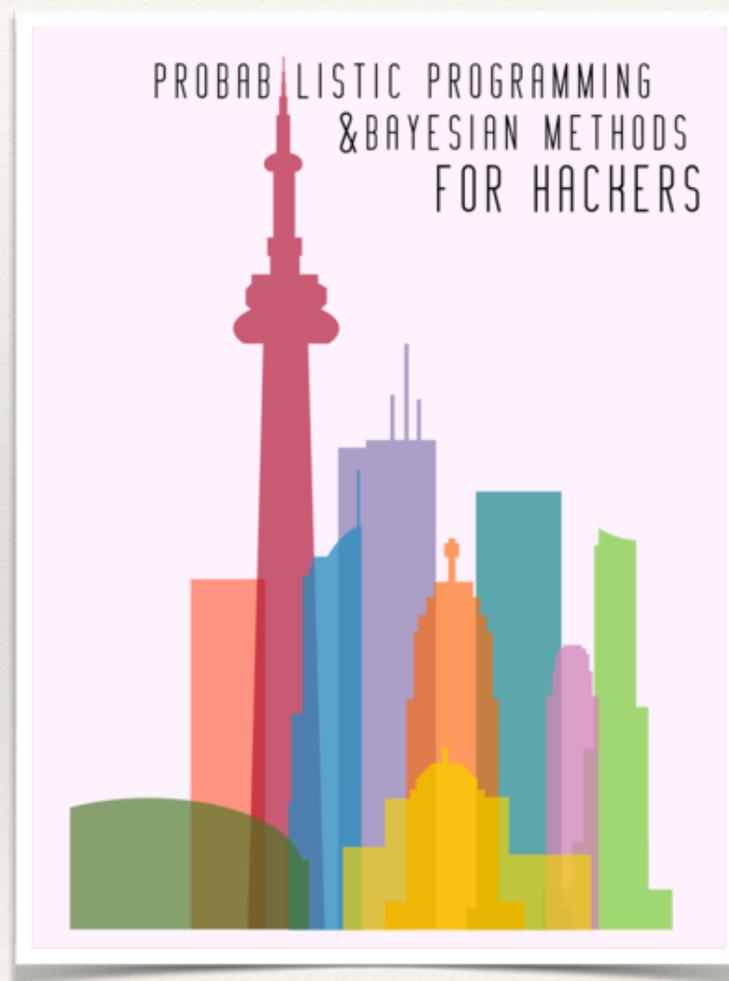
---

# More authors creating books this way

---



By Matthew Russell



By Cameron Davidson-Pilon

# 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!

# A collaborative MOOC on OpenEdX

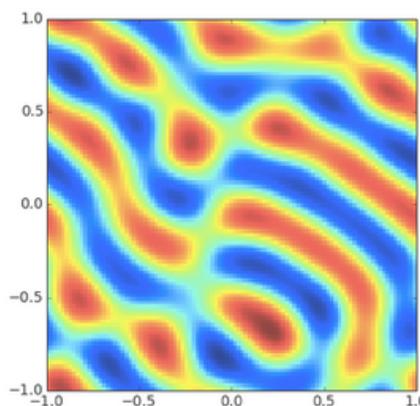
- ❖ *Lorena Barba* at George Washington University, USA.
- ❖ *Ian Hawke* at Southampton, UK
- ❖ *Carlos Jerez* at Pontifical Catholic University of Chile.
- ❖ All materials on [Github](#).

Lorena A. Barba group



## Announcing "Practical Numerical Methods with Python" MOOC

Posted on 07.26.2014



Pattern formation:

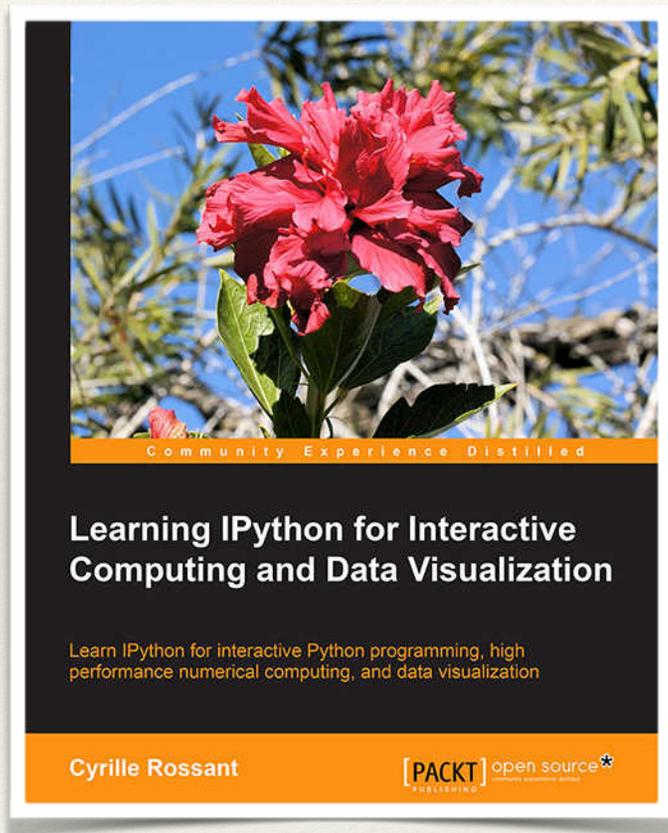
▶ solution for a reaction-diffusion system like:

$$u_t = \delta D_1 \nabla^2 u + f(u, v)$$

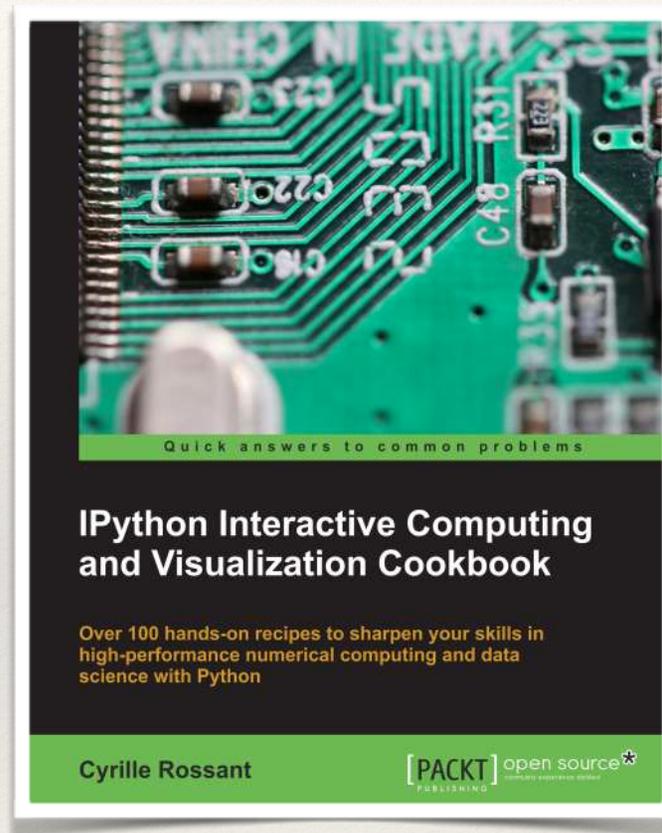
$$v_t = \delta D_2 \nabla^2 v + g(u, v)$$

An example of the types of problems we will learn to solve in this course, among others governed by differential equations.

# 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 article title is 'Interactive notebooks: Sharing the code' by Helen Shen, dated 05 November 2014. The article's abstract states: 'The free IPython notebook makes data analysis easier to record, understand and reproduce.' Below the abstract are links for 'PDF' and 'Rights & Permissions'. A large illustration depicts hands interacting with a digital notebook displaying various data visualizations like charts and graphs. To the right, a 'Top story' section features a photograph of Brontosaurus dinosaurs with the text 'USA 25 Brontosaurus' and a headline: 'Beloved Brontosaurus makes a comeback'. Below this is a list of 'Recent' articles with columns for 'Read', 'Comments', and 'Emailed'.

**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

Illustrations by The Project Twins

### Top story

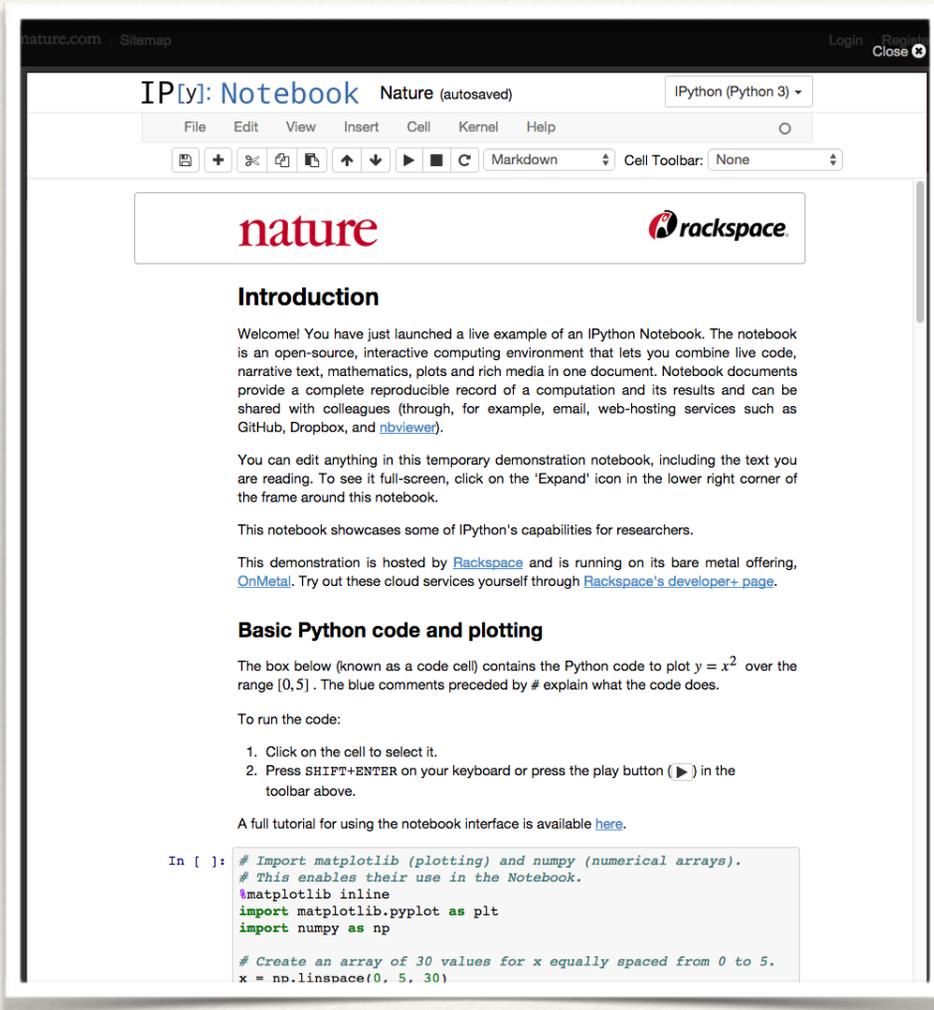
**Beloved Brontosaurus makes a comeback**

Jurassic giant's taxonomic status is restored.

Recent	Read	Comments	Emailed
1. <b>History: Women at the edge of science</b>			
<i>Nature</i>   08 April 2015			
2. <b>Scientific instrumentation: The aided eye</b>			
<i>Nature</i>   08 April 2015			
3. <b>Books in brief</b>			
<i>Nature</i>   08 April 2015			
4. <b>Antibody shows promise as</b>			

<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** 

### 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.

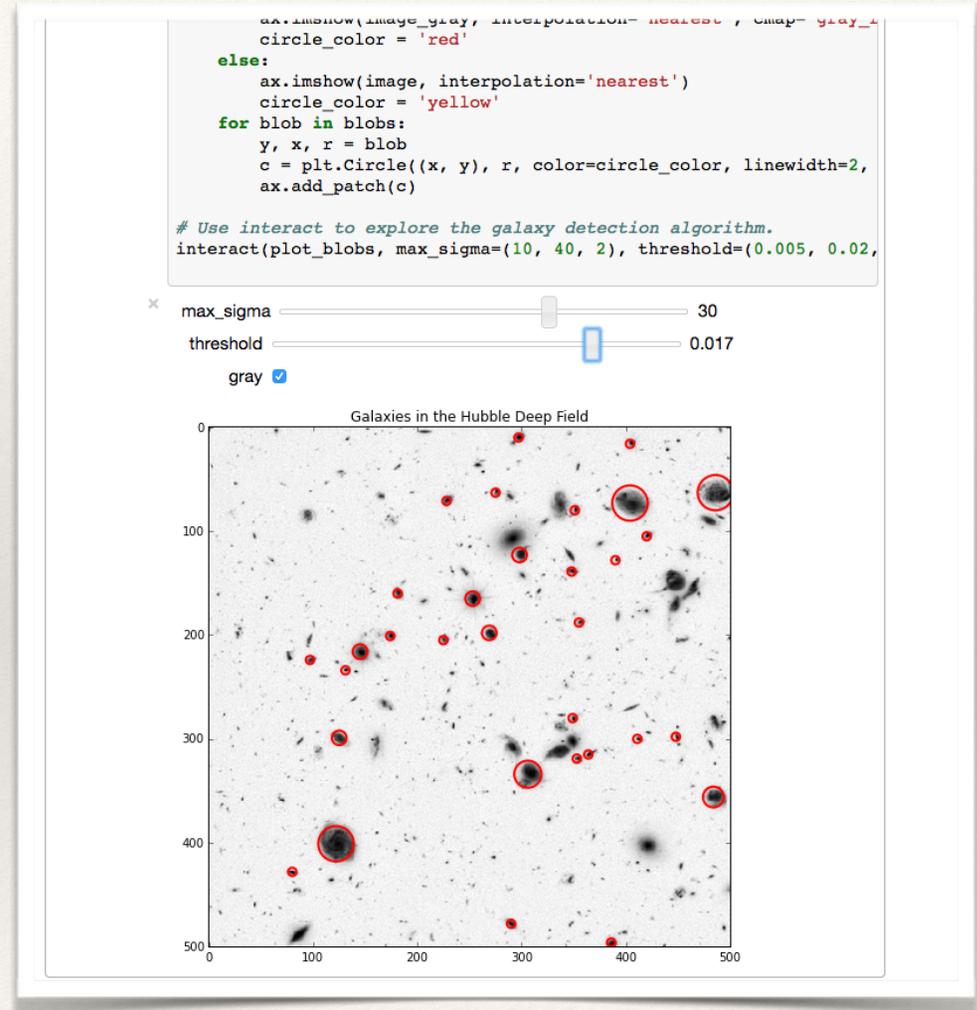
To run the code:

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)
```

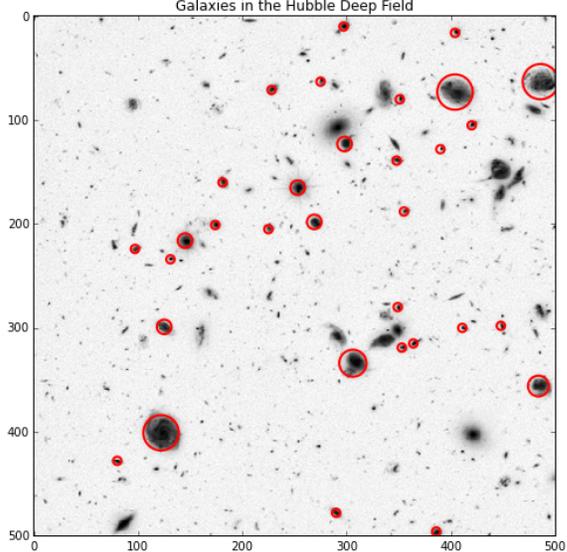


```
ax.imshow(image_gray, interpolation='nearest', cmap=gray_r
circle_color = 'red'
else:
ax.imshow(image, interpolation='nearest')
circle_color = 'yellow'
for blob in blobs:
y, x, r = blob
c = plt.Circle((x, y), r, color=circle_color, linewidth=2,
ax.add_patch(c)

# 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



# Back to Journalism

# FiveThirtyEight and data-driven journalism



■ BECHDEL TEST | 1:52 PM | APR 1, 2014



## The Dollar-And-Cents Case Against Hollywood's Exclusion of Women



By WALT HICKEY

### TOP STORIES



# Brian Keegan: calls out 538 about openness

nbviewer   FAQ   IPython



Bechdel / Bechdel\_test.ipynb /

## The Need for Openness in Data Journalism

[Brian Keegan, Ph.D. \(@bkeegan\)](#) College of Humanities and Social Sciences, Northeastern University

Do films that pass the Bechdel Test make more money for their producers? I've replicated Walt Hickey's [recent article](#) in FiveThirtyEight to find out. My results confirm his own in part, but also find notable differences that point the need for clarification at a minimum. While I am far from the first to make this argument, this case is illustrative of a larger need for journalism and other data-driven enterprises to borrow from hard-won scientific practices of sharing data and code as well as supporting the review and revision of findings. This admittedly lengthy post is a critique of not only this particular case but also an attempt to work through what open data journalism could look like.

## The Angle: Data Journalism should emulate the openness of science

New data-driven journalists such as FiveThirtyEight have faced criticism from many quarters and the critiques, particularly around the naïveté of assuming credentialed experts can be bowled over by quantitative analysis so easily as the terrifyingly innumerate pundits who infest our political media [\[1,2,3,4\]](#). While I find these critiques persuasive, I depart from them here to instead argue that I have found this "new" brand of data journalism disappointing foremost because *it wants to perform science without abiding by scientific norms.*

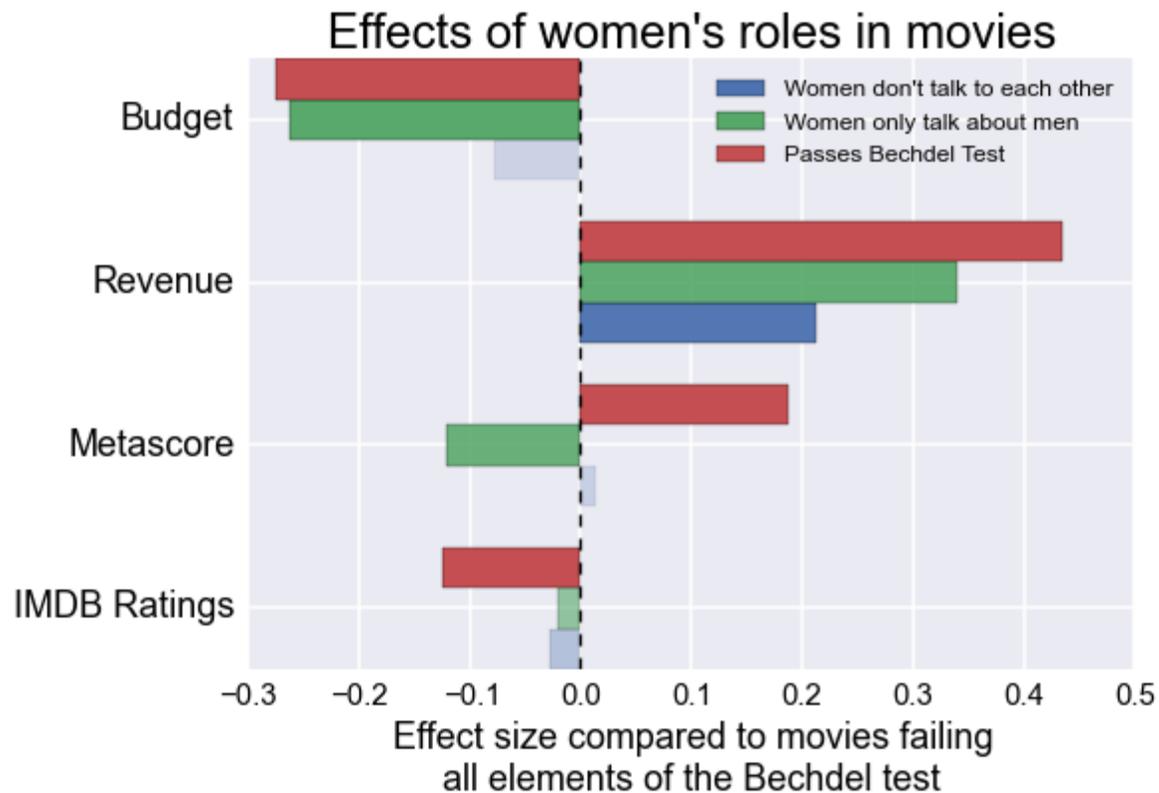
```
plt.yticks(plt.yticks()[0],['IMDB Ratings','Metascore','Revenue','Budget'],fontsize=18)
plt.xlabel('Effect size compared to movies failing\nall elements of the Bechdel test',fontsize=18)
plt.title("Effects of women's roles in movies",fontsize=24)
plt.xticks(fontsize=15)
plt.autoscale()
```

Difference in IMDB scores between movies that pass all and fail all requirements of the Bechdel test: -0.12.

Difference in Metascores between movies that pass all and fail all requirements of the Bechdel test: 1.87.

Difference in revenue between movies that pass all and fail all requirements of the Bechdel test: 54.49%.

Difference in budget between movies that pass all and fail all requirements of the Bechdel test: -24.0%.



# Response by FiveThirtyEight

<http://fivethirtyeight.com/datalab/the-bechdel-test-checking-our-work>

“Keegan also made a larger point:

*FiveThirtyEight and similar sites should make their data available. We couldn't agree more.*

We're exploring ways of making our raw code and data available to readers, including through **FiveThirtyEight's GitHub account.**”

Data and code behind the stories and interactives at FiveThirtyEight

**276** commits   
 **1** branch   
 **0** releases   
 **15** contributors

branch: **master** ▾    **data** / +   

update poll of pollsters README

**andrewflowers** authored 5 days ago      latest commit **8450a9b528**

<a href="#">airline-safety</a>	add airline-safety data	4 months ago
<a href="#">alcohol-consumption</a>	Update README.md	3 months ago
<a href="#">bad-drivers</a>	add bad drivers data	11 days ago
<a href="#">bechdel</a>	format email address	7 months ago
<a href="#">bob-ross</a>	cleaned up bob ross clustering script	7 months ago
<a href="#">classic-rock</a>	fixed entries with #REF! excel errors on two rows of classic-r...	4 months ago
<a href="#">college-majors</a>	Update README.md	2 months ago
<a href="#">comic-characters</a>	add detail to comic characters README	a month ago
<a href="#">comma-survey-data</a>	clean comma survey data	5 months ago
<a href="#">congress-age</a>	renamed congress_terms.csv to congress-terms.csv	7 months ago
<a href="#">early-senate-polls</a>	add polling data	7 months ago
<a href="#">flying-etiquette-survey</a>	Update README.md	2 months ago

<> **Code**

**4**  
[Issues](#)

**0**  
[Pull Requests](#)

[Pulse](#)

[Graphs](#)

**HTTPS** clone URL

You can clone with [HTTPS](#) or [Subversion](#).

**Download ZIP**

# A recent example: LA Times, Oct'15

lapd-crime-classification-a x LAPD underreported seriou x Fernando

www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html

SECTIONS SEARCH **Los Angeles Times** SUBSCRIBE LOG IN

THURSDAY OCT. 22, 2015 MOST POPULAR LOCAL ENTERTAINMENT SPORTS POLITICS EDUCATION OPINION 67°

## LAPD underreported serious assaults, skewing crime stats for 8 years



LAPD officers arrest a suspected gang member in 2009, during the period when violent crimes were underreported by 7%. (Michael Robinson Chavez / Los Angeles)

**Ben Poston, Joel Rubin and Anthony Pesce** Contact Reporters

### In Case You Missed It

-  **Kardashians' freak show capitalizes on Lamar Odom one more time**  
Oct. 21, 2015
-  **Fresh & Easy begins closing down stores**  
Oct. 21, 2015
-  **Three sex-cult members are convicted of murdering Marine wife**  
Oct. 21, 2015

[See More](#)

# Jupyter Notebooks on github/datadesk

Branch: master | lapd-crime-classification-analysis / classifiers.ipynb

**anthonyjpesce** tweak to intro, rerun each cell 965c61f 8 days ago

1 contributor

417 lines (416 sloc) | 16.7 KB

Raw Blame History

## Checking the LAPD's crime classifications

The Times analyzed Los Angeles Police Department violent crime data from 2005 to 2012. Our analysis found that the Los Angeles Police Department misclassified an estimated 14,000 serious assaults as minor offenses, artificially lowering the city's crime levels.

To conduct the analysis, The Times used an algorithm that combined two machine learning classifiers. Each classifier read in a brief description of the crime, which it used to determine if it was a minor or serious assault. You can see a [sample of the data here](#). An example of a minor assault reads: "VICTS AND SUSPS BECAME INV IN VERBA ARGUMENT SUSP THEN BEGAN HITTING VICTS IN THE FACE."

```
In [1]: import csv
import nltk
from nltk.util import ngrams
from sklearn.svm import LinearSVC
from sklearn.pipeline import Pipeline
from nltk.classify import MaxentClassifier
from nltk.stem.snowball import SnowballStemmer
from nltk.classify.scikitlearn import SklearnClassifier
from sklearn.feature_extraction.text import TfidfTransformer
```

## Stemming and stop words

We're going to clean up the crime descriptions in two steps. First, we're going to [stem](#) the words -- this reduces

Thanks to Jeremy Singer-Vine for pointing me to this work!

# Notebook Workflows: The Big Picture

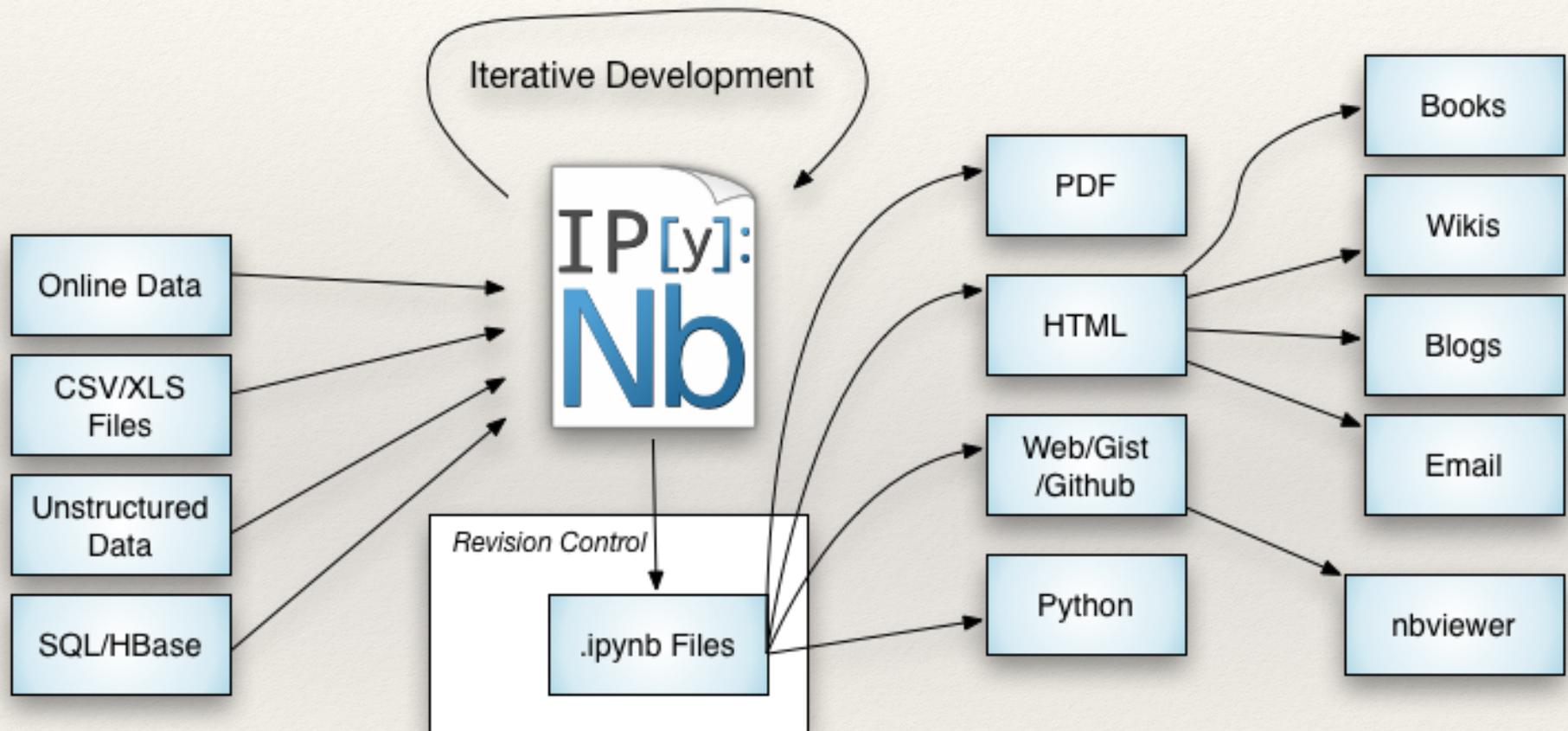


Image credit: [Joshua Barratt](#)

# 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](#).

# 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.



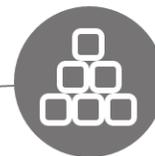
### Pluggable authentication

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



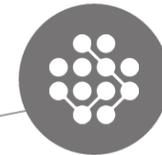
### Centralized deployment

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.

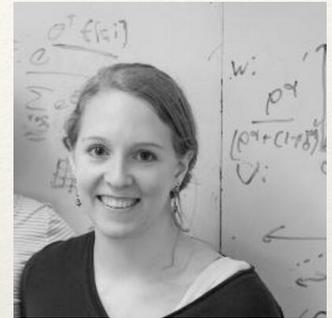


### Code meets data

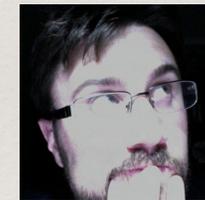
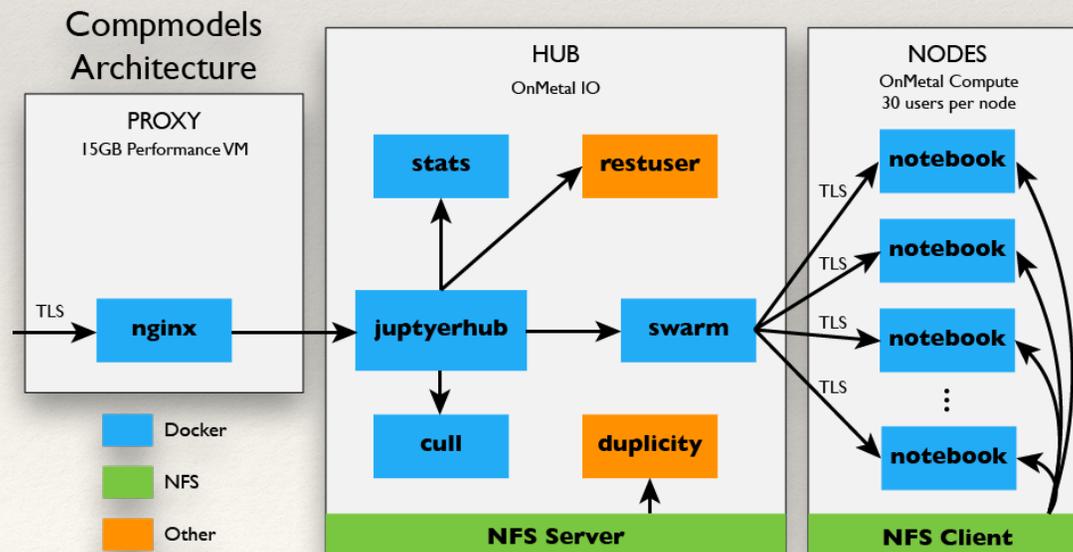
Deploy the Notebook next to your data to provide unified software management and data access within your organization.

# 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, O'Reilly...

The screenshot shows the Windows Azure website with a navigation menu including 'HOME', 'PRICING', 'DEVELOP', 'MANAGE', 'COMMUNITY', 'SUPPORT', and 'ACCOUNT'. The main heading is 'IPython Notebook on Windows Azure'. Below the heading, there is a 'Free trial' button and a play button icon. The text describes the IPython project and its capabilities for scientific computing.

The screenshot shows a Google Drive Colaboratory notebook interface. The title is 'Colaborative IPython Notebooks on Google Drive'. The code in the notebook includes:

```
import os
os.getpid()

print("please run me, Fernando")

please run me, Fernando

%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt

N = 200
r = 2 * np.random.rand(N)
theta = 2 * np.pi * np.random.rand(N)
area = 200 * r ** 2 * np.random.rand(N)
colors = theta
ax = plt.subplot(111, polar=True)
c = plt.scatter(theta, r, c=colors, s=area, cmap=plt.cm.hsv)
c.set_alpha(0.75)
```

A polar scatter plot is displayed below the code, showing a distribution of points in a circular pattern with varying colors and sizes.

The screenshot shows the Data Scientist Workbench website. The main heading is 'Data Scientist Workbench' with the tagline 'Prepare data. Analyze data. Get answers.' Below this, there are three main sections: 'Prepare data effortlessly.', 'Explore Data.', and 'Analyze data interactively.' Each section has a corresponding button: 'Prepare Data', 'My Notebooks', and 'My Notebooks'. A table of data is visible in the background, showing columns for 'Type of Contract', 'Status', 'Type of Contract', and 'Date of Award'.

The screenshot shows the Google Cloud Datalab interface. The main heading is 'Google Cloud Datalab'. Below the heading, there are 'Notebooks' and 'Sessions' tabs. A list of notebooks is displayed, including 'Introduction to Notebooks.ipynb' (Running), 'Introduction to Python.ipynb', and 'Working with Datalab.ipynb'. The interface also shows a 'Notebook', 'Folder', and 'Upload' button.

The screenshot shows an O'Reilly article titled 'Embracing Jupyter Notebooks at O'Reilly'. The article is by Andrew Odewahn, dated May 7, 2015. The text states: 'O'Reilly Media is using our Atlas platform to make Jupyter Notebooks a first class authoring environment for our publishing program.' Below the article, there is a section titled 'Embracing Jupyter Notebooks at O'Reilly' with the text: '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'.

---

# In summary

---

- ❖ Communicating scientific narratives poses similar challenges to data-intensive journalism
- ❖ Our tools are open, mature and available to you
- ❖ A dialog with your community could be enormously valuable for both!

# Thank You!

@fperez\_org fperez@lbl.gov

@ProjectJupyter @IPythonDev

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