IP[y]: IPython

jupyter

Reproducible science with Jupyter

Changing our publication models

Fernando Pérez
(@fperez_org & fperez@lbl.gov)

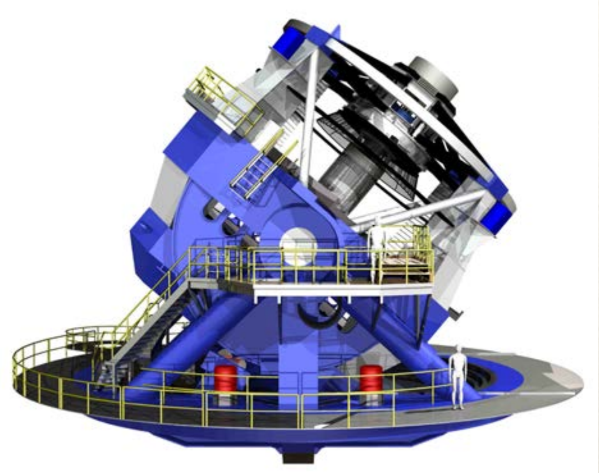
LBL & UC Berkeley

BERKELEY LAB

Berkeley UNIVERSITY OF CALIFORNIA

BIDS BERKELEY INSTITUTE FOR DATA SCIENCE

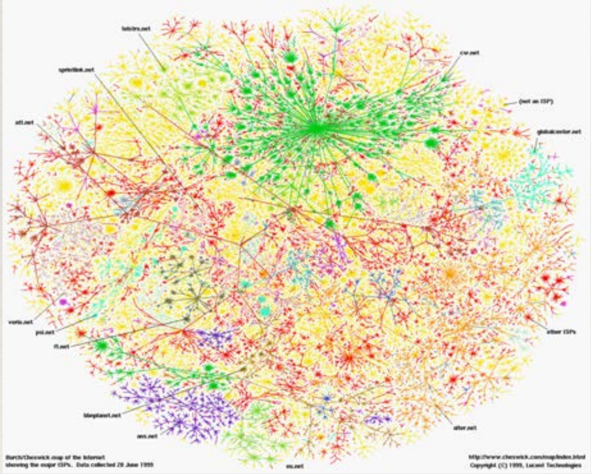
Every research discipline is now awash in data



Astronomy: LSST

Physics: LHC

Personalized, data-driven medicine

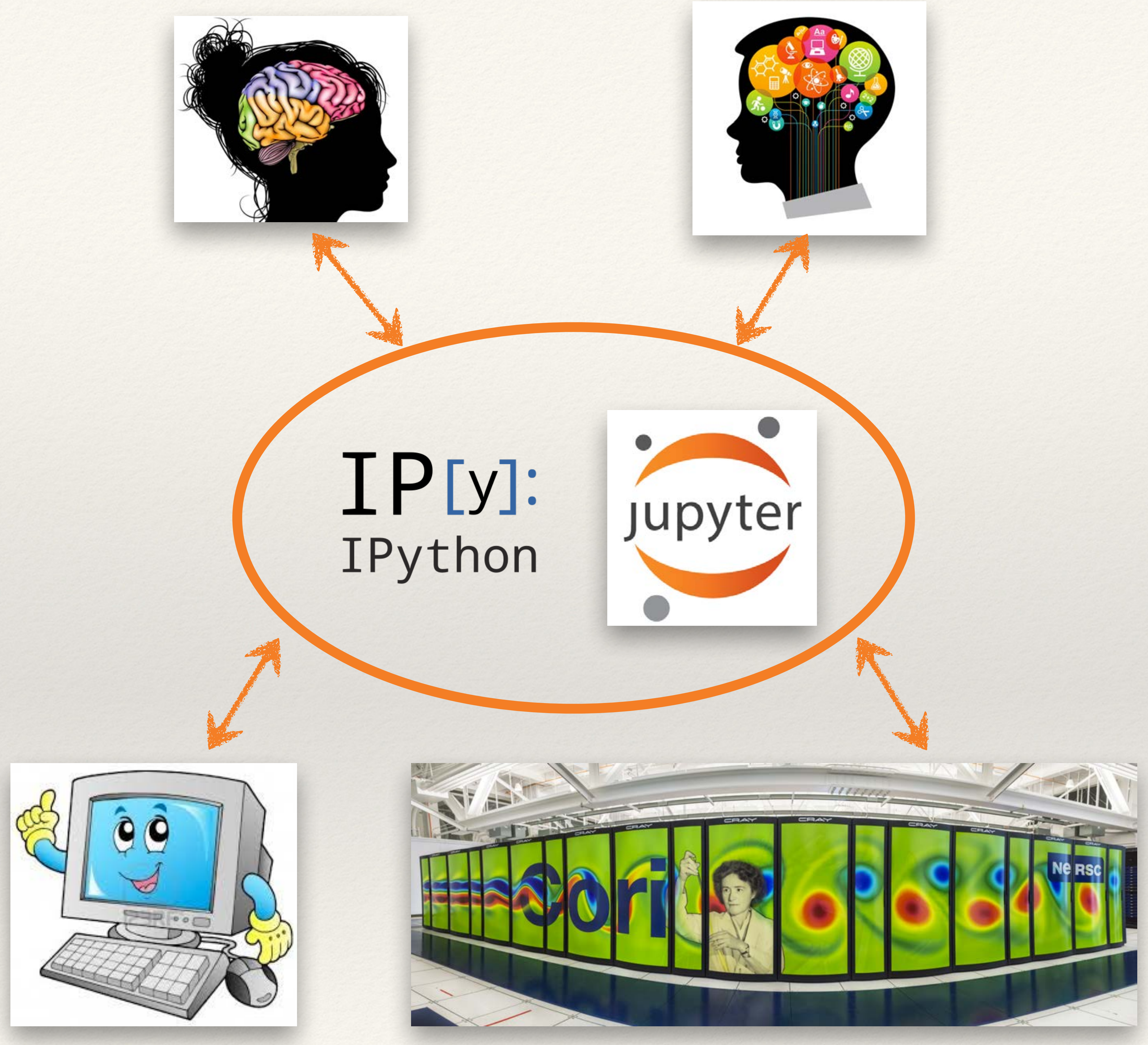


Sociology: The Web

Biology: Sequencing

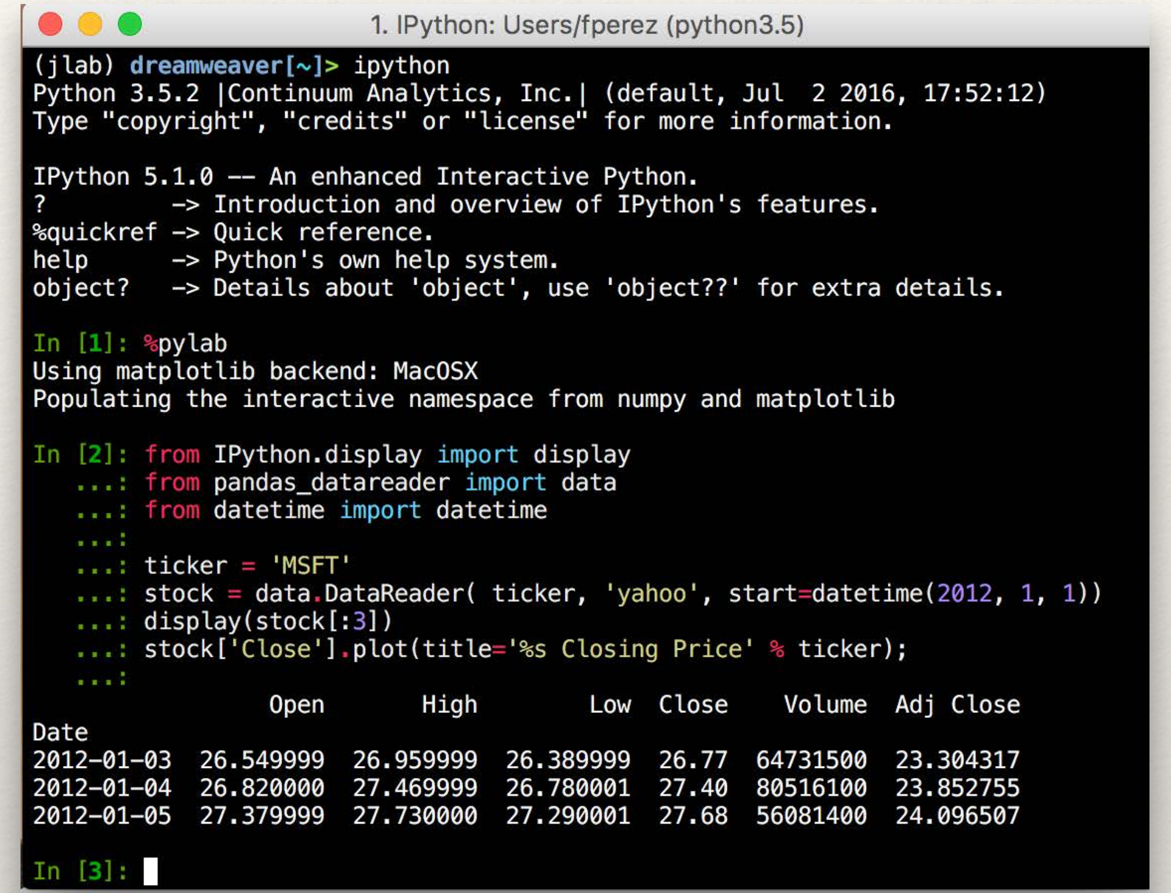
Economics: POS terminals

Neuroscience: EEG, fMRI

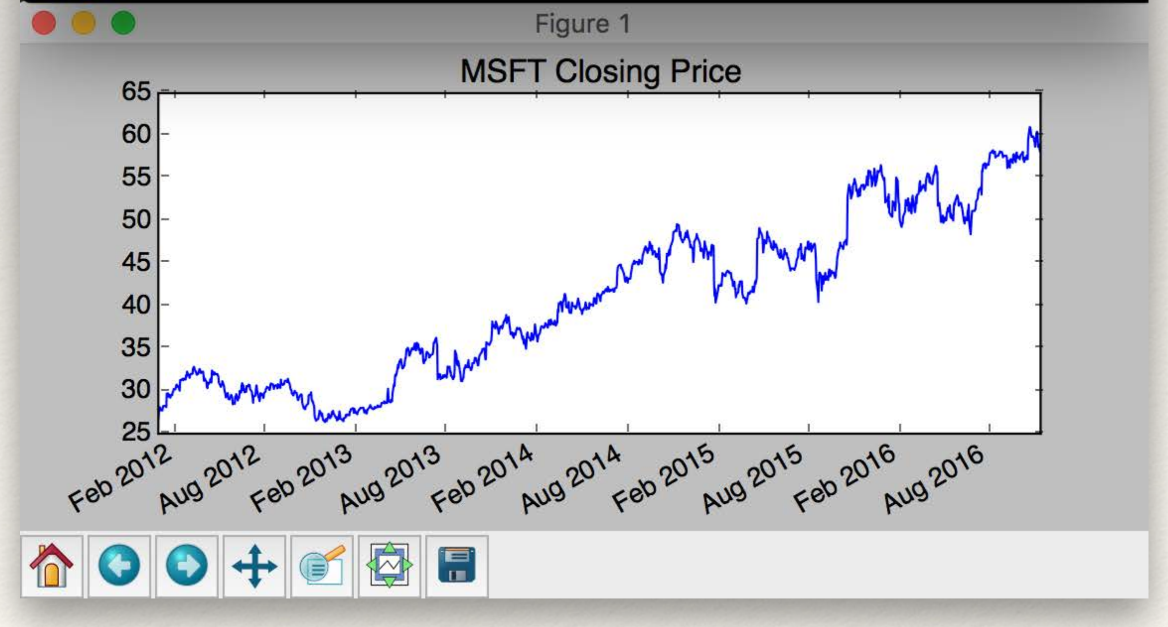

IP[y]:
IPython
jupyter
Cori
NERSC

IPython: Interactive Python, 2001

- Object Introspection (TAB!)
- OS Integration
- Rich terminal client
- GUI support (plots, …)
- %magic commands
- Embeddable

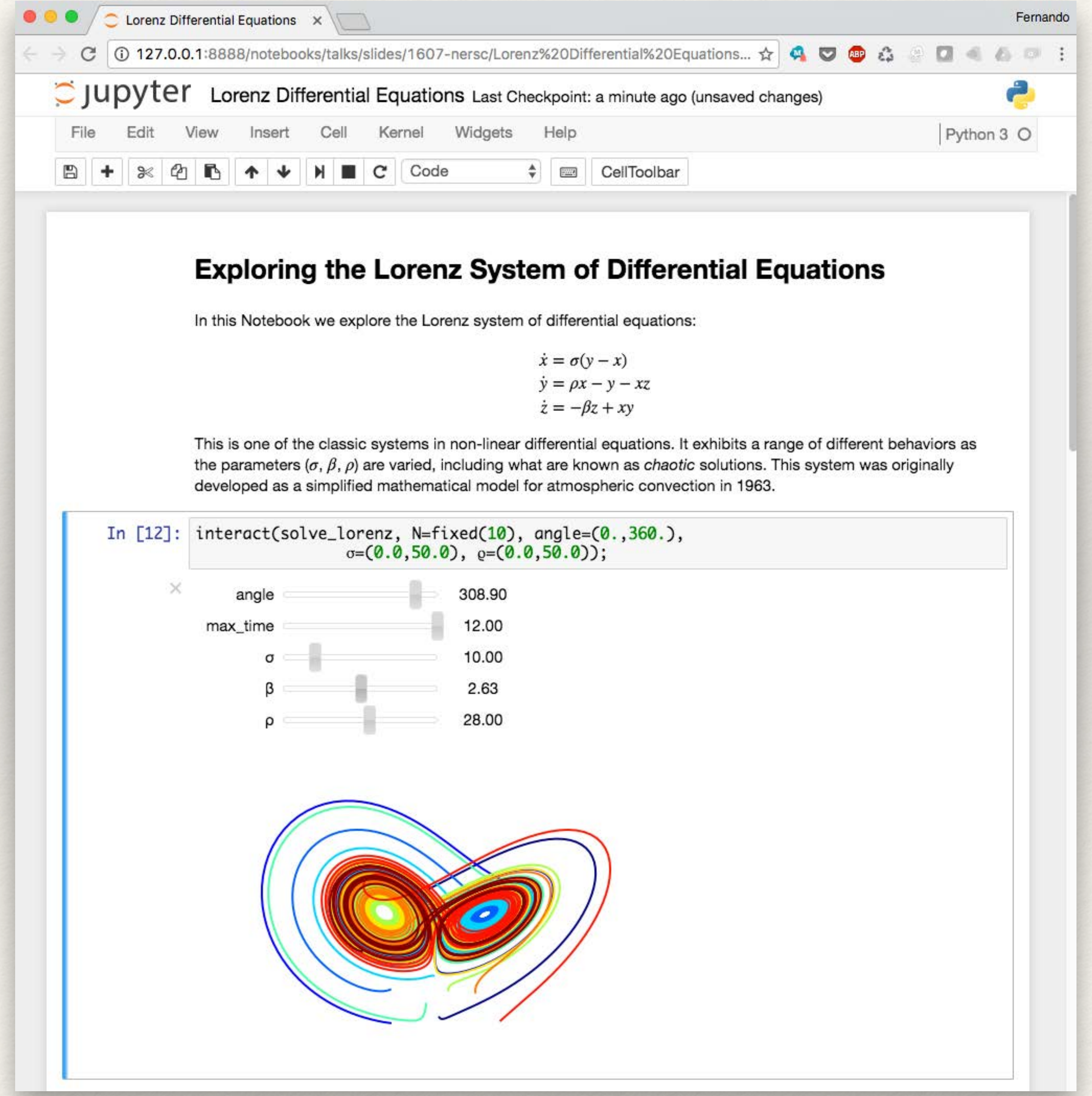


```
(jlab) dreamweaver[~]> ipython
Python 3.5.2 |Continuum Analytics, Inc.| (default, Jul  2 2016, 17:52:12)
Type "copyright", "credits" or "license" for more information.

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

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

In [2]: from IPython.display import display
   ...: from pandas_datareader import data
   ...: from datetime import datetime
   ...:
   ...: ticker = 'MSFT'
   ...: stock = data.DataReader( ticker, 'yahoo', start=datetime(2012, 1, 1))
   ...: display(stock[:3])
   ...: stock['Close'].plot(title='%s Closing Price' % ticker);
   ...:
                 Open       High        Low  Close    Volume  Adj Close
Date
2012-01-03  26.549999  26.959999  26.389999  26.77  64731500  23.304317
2012-01-04  26.820000  27.469999  26.780001  27.40  80516100  23.852755
2012-01-05  27.379999  27.730000  27.290001  27.68  56081400  24.096507

In [3]: 
```



The IPython/Jupyter Notebook

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



Lorenz Differential Equations
Fernando
127.0.0.1:8888/notebooks/talks/slides/1607-nersc/Lorenz%20Differential%20Equations...
jupyter
Lorenz Differential Equations Last Checkpoint: a minute ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python 3
Code
CellToolbar

Exploring the Lorenz System of Differential Equations

In this Notebook we explore the Lorenz system of differential equations:

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

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic* solutions. This system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [12]: interact(solve_lorenz, N=fixed(10), angle=(0.,360.),
                  σ=(0.0,50.0), ρ=(0.0,50.0));
```

angle 308.90
max_time 12.00
σ 10.00
β 2.63
ρ 28.00

Funding and partnerships

SLOAN FOUNDATION ALFRED P. 1934

ALFRED P. SLOAN FOUNDATION

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

GORDON AND BETTY MOORE FOUNDATION

U.S. DEPARTMENT OF ENERGY

O'REILLY

SIMONS FOUNDATION

CONTINUUM ANALYTICS

IBM®

POWERED BY rackspace®
the open cloud company

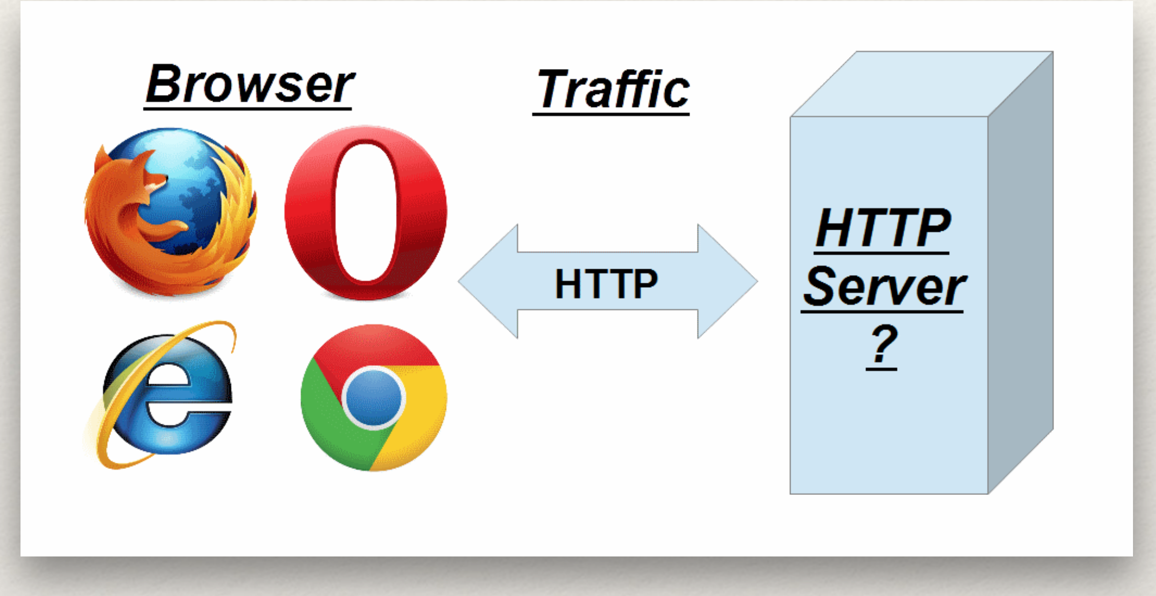
Microsoft

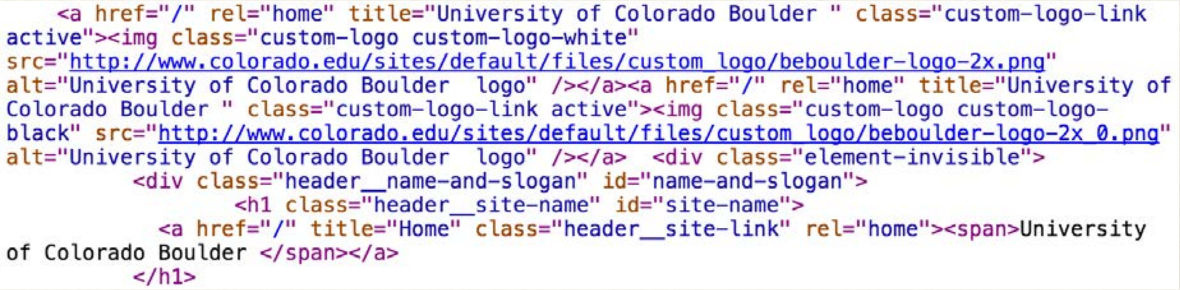
ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS

Google

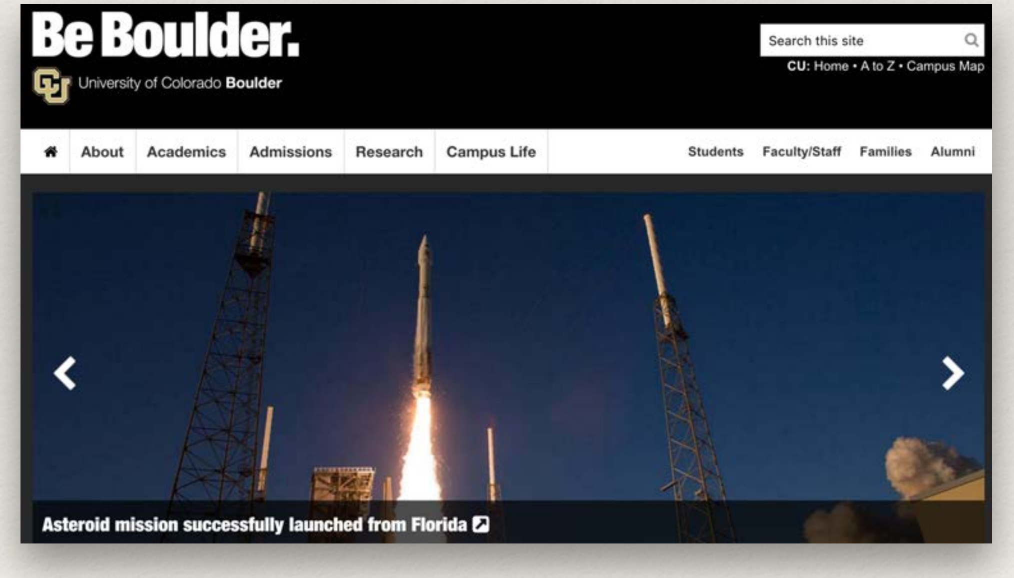
Bloomberg

Core ideas of the web: HTTP & HTML





```
    <a href="/" rel="home" title="University of Colorado Boulder " class="custom-logo-link
active"><img class="custom-logo custom-logo-white"
src="http://www.colorado.edu/sites/default/files/custom_logo/beboulder-logo-2x.png"
alt="University of Colorado Boulder  logo" /></a><a href="/" rel="home" title="University of
Colorado Boulder " class="custom-logo-link active"><img class="custom-logo custom-logo-
black" src="http://www.colorado.edu/sites/default/files/custom_logo/beboulder-logo-2x_0.png"
alt="University of Colorado Boulder  logo" /></a>  <div class="element-invisible">
          <div class="header__name-and-slogan" id="name-and-slogan">
                  <h1 class="header__site-name" id="site-name">
            <a href="/" title="Home" class="header__site-link" rel="home"><span>University
of Colorado Boulder </span></a>
          </h1>
```


Be Boulder.
University of Colorado Boulder
Search this site
CU: Home • A to Z • Campus Map
About
Academics
Admissions
Research
Campus Life
Students
Faculty/Staff
Families
Alumni
Asteroid mission successfully launched from Florida

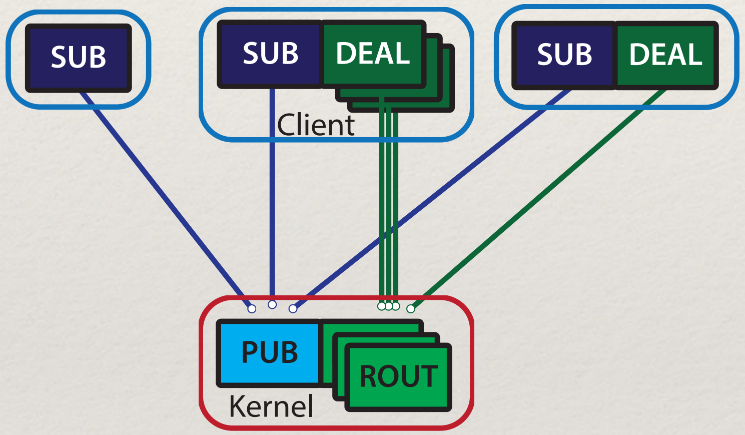
HTTP: protocol to connect clients and servers
HyperText Transport Protocol

HTML: format to represent content
HyperText Markup Language

Image credit: eviltester.com

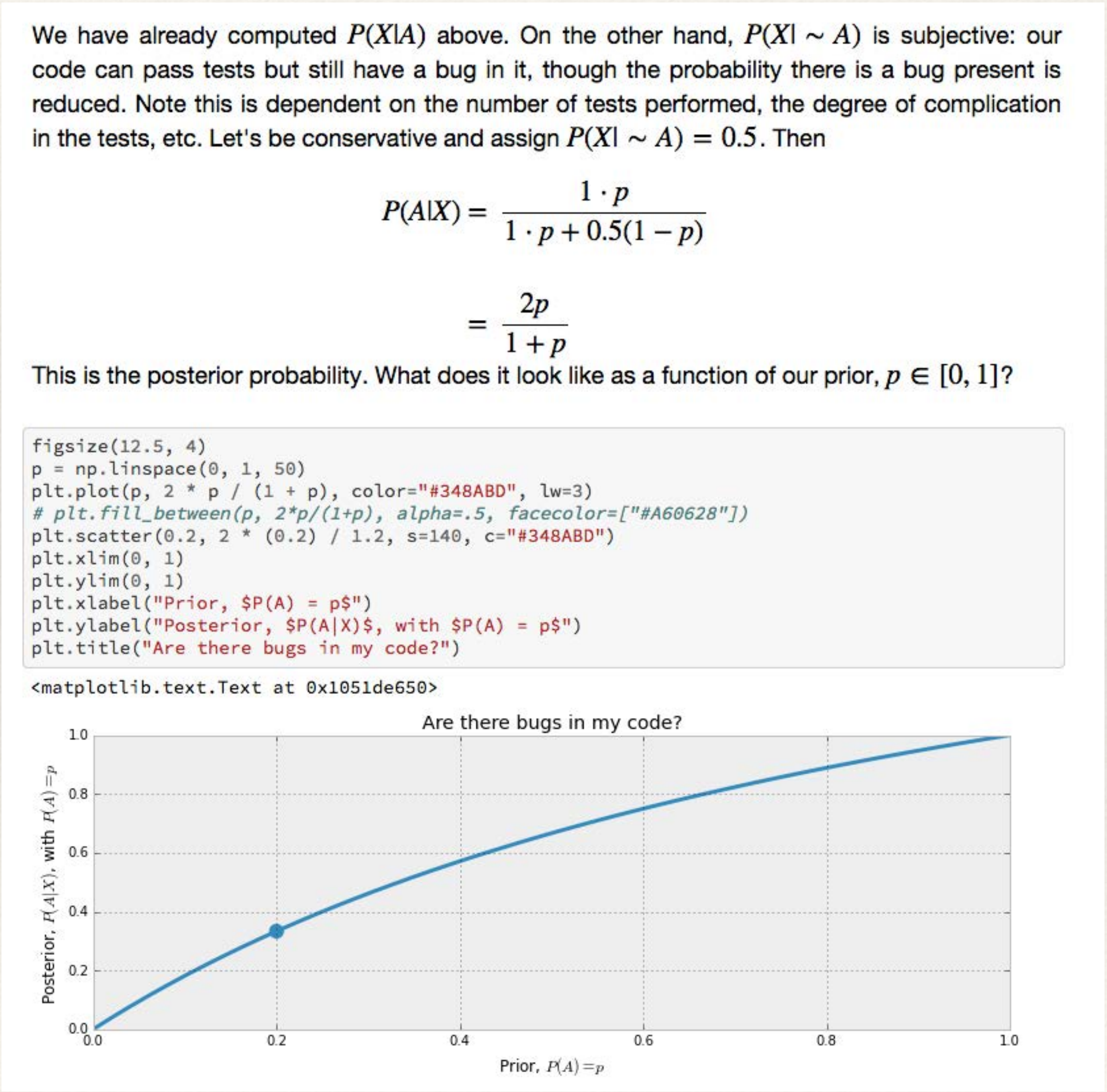
Core ideas of Jupyter

Interactive Computing Protocol



ØMQ + JSON

Document Format



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

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

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

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

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

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

Are there bugs in my code?
Posterior, $P(A|X)$, with $P(A) = p$
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Prior, $P(A) = p$

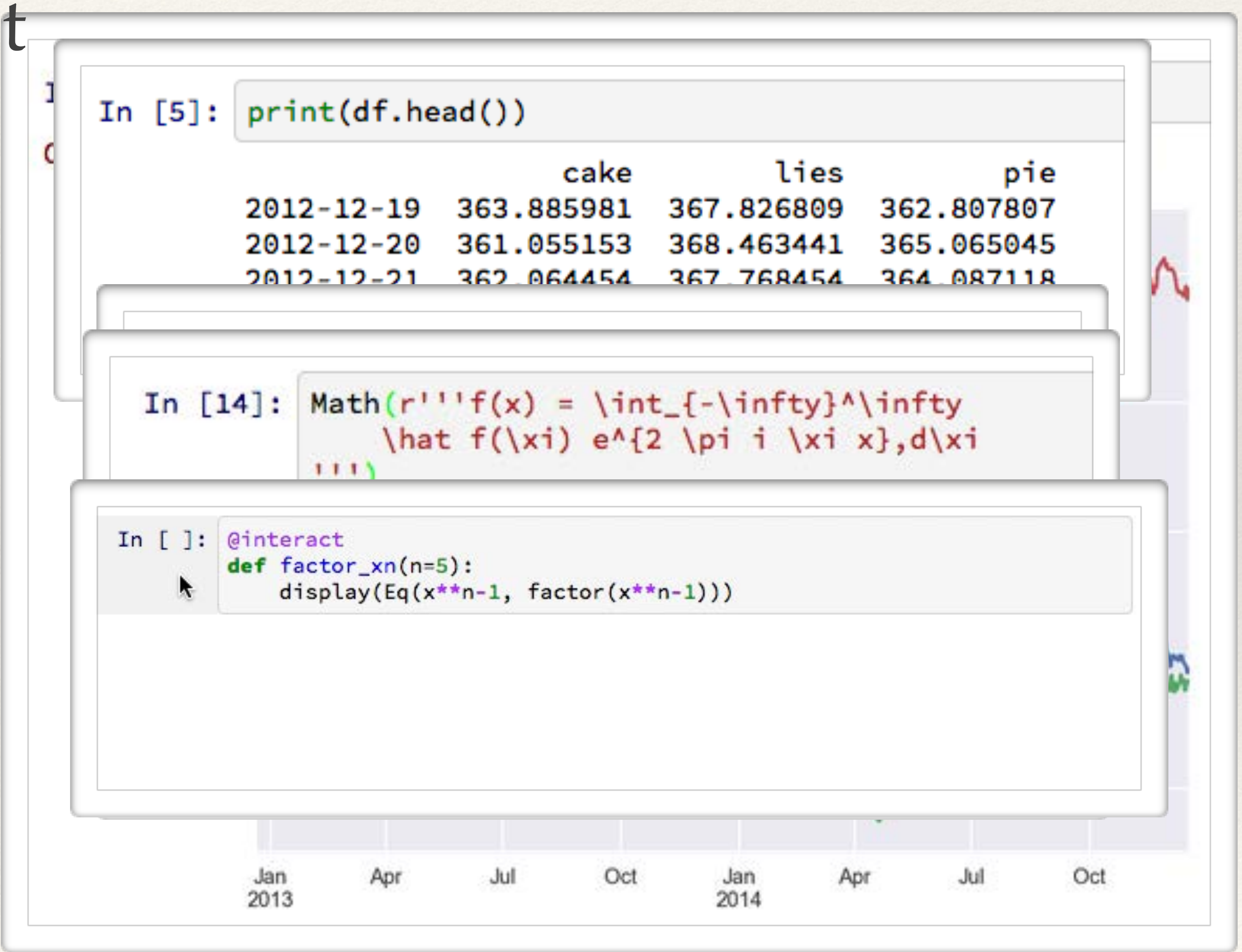
https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

Jupyter Protocol

capture the process of interactive computing

any mime-type output

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



```
In [5]: print(df.head())
                  cake        lies         pie
2012-12-19  363.885981  367.826809  362.807807
2012-12-20  361.055153  368.463441  365.065045
2012-12-21  362.064454  367.768454  364.087118
```

```
In [14]: Math(r'''f(x) = \int_{-\infty}^\infty
             \hat f(\xi) e^{2 \pi i \xi x},d\xi
         ''')
```

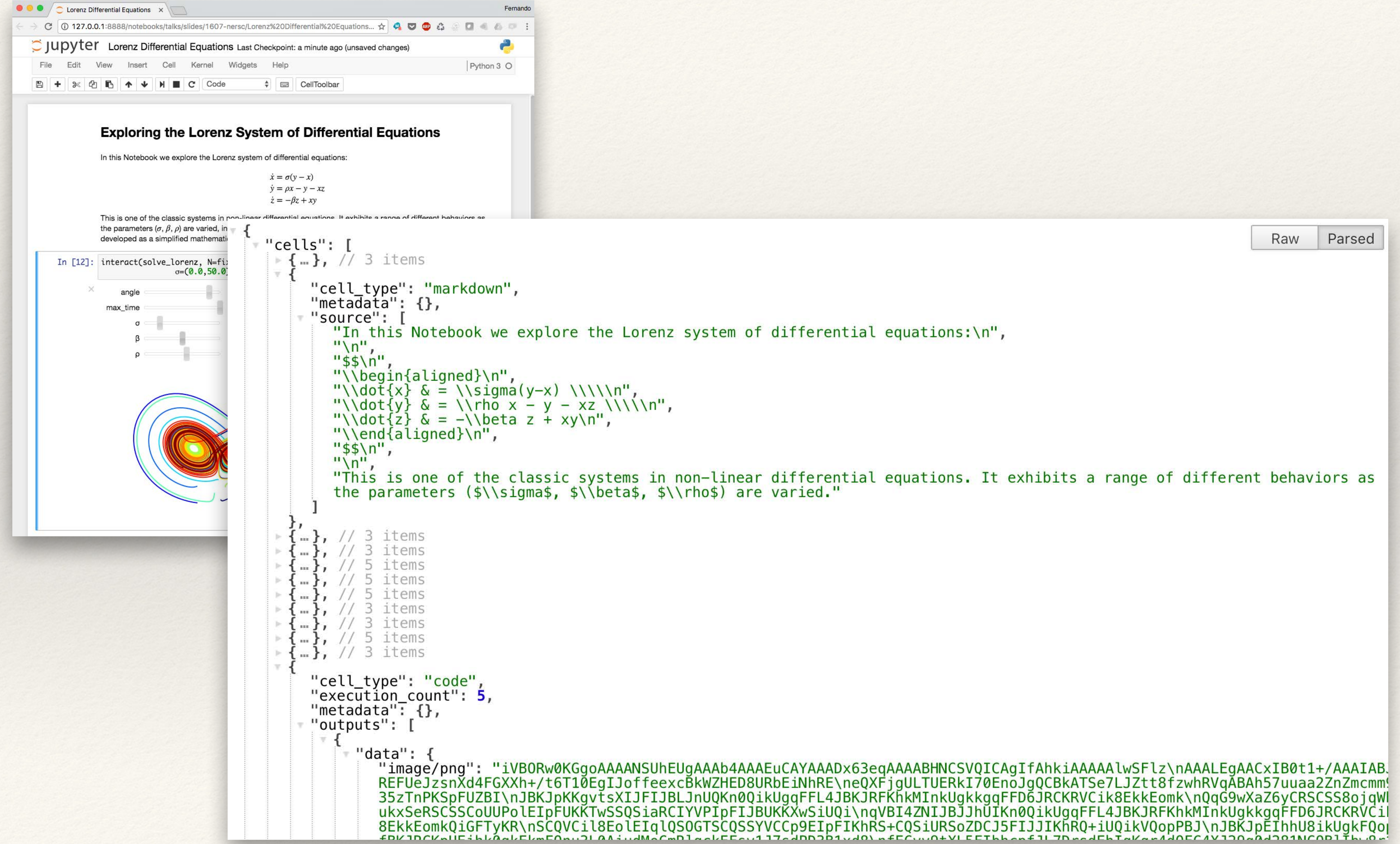
```
In [ ]: @interact
        def factor_xn(n=5):
            display(Eq(x**n-1, factor(x**n-1)))
```

Jupyter Protocol is language agnostic

Scala
perl
R
jupyter
julia
Spark
ERLANG
python™
IP[y]:
IPython

~75 different kernels: https://github.com/ipython/ipython/wiki/IPython-kernels-for-other-languages

Notebook: a *data structure*


Lorenz Differential Equations
Fernando
127.0.0.1:8888/notebooks/talks/slides/1607-nersc/Lorenz%20Differential%20Equations...
jupyter
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python 3
Code
CellToolbar
Raw
Parsed

Lorenz Differential Equations Last Checkpoint: a minute ago (unsaved changes)

Exploring the Lorenz System of Differential Equations

In this Notebook we explore the Lorenz system of differential equations:

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

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters (σ, β, ρ) are varied, in

developed as a simplified mathemati

```
In [12]: interact(solve_lorenz, N=fi
                  σ=(0.0,50.0
```

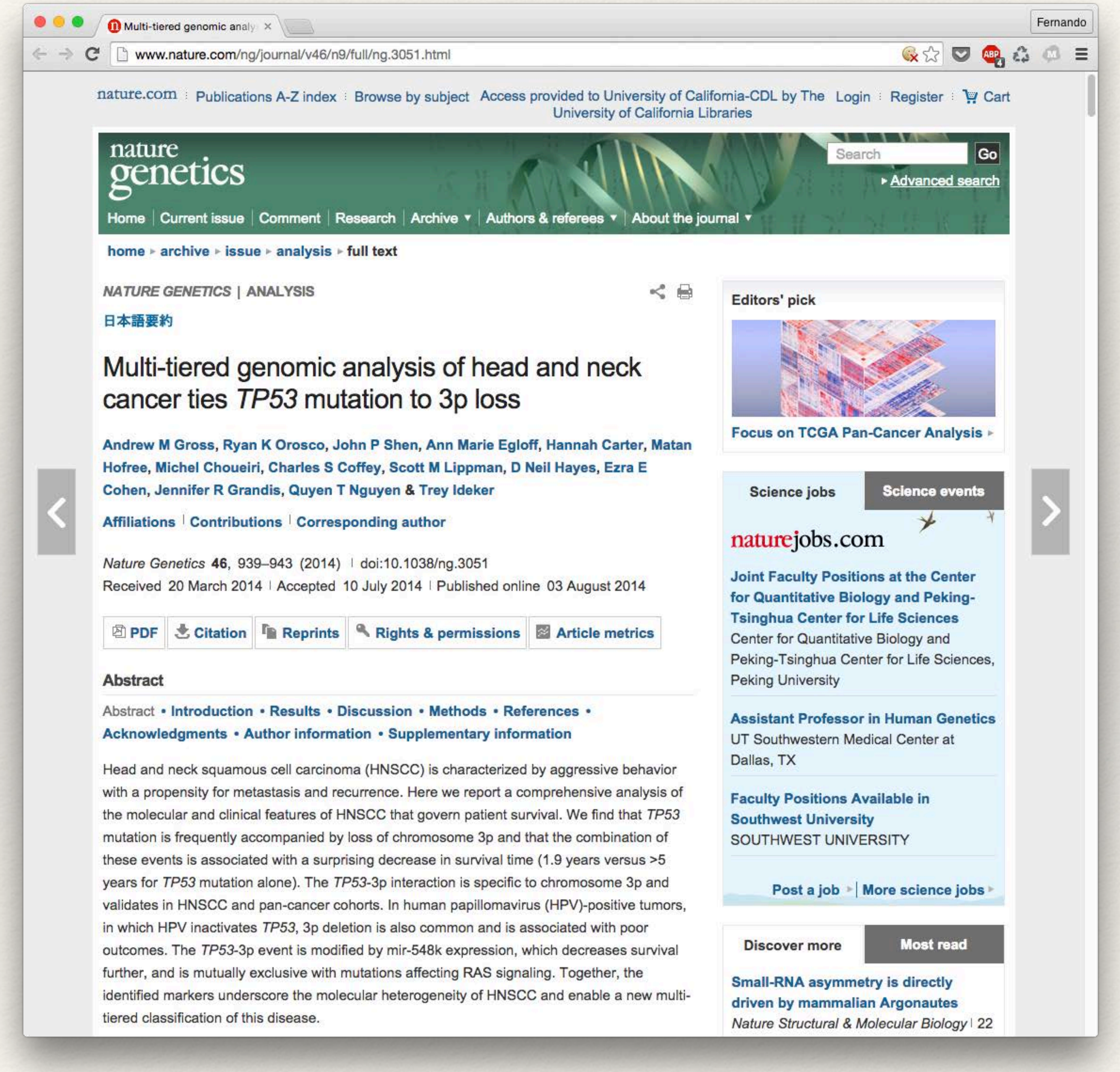
```
{
  "cells": [
    {…}, // 3 items
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "In this Notebook we explore the Lorenz system of differential equations:\n",
        "\n",
        "$$\n",
        "\\begin{aligned}\n",
        "\\dot{x} & = \\sigma(y-x) \\\\\n",
        "\\dot{y} & = \\rho x - y - xz \\\\\n",
        "\\dot{z} & = -\\beta z + xy\n",
        "\\end{aligned}\n",
        "$$\n",
        "\n",
        "This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as
        the parameters ($\\sigma$, $\\beta$, $\\rho$) are varied."
      ]
    },
    {…}, // 3 items
    {…}, // 3 items
    {…}, // 5 items
    {…}, // 5 items
    {…}, // 5 items
    {…}, // 3 items
    {…}, // 3 items
    {…}, // 5 items
    {…}, // 3 items
    {
      "cell_type": "code",
      "execution_count": 5,
      "metadata": {},
      "outputs": [
        {
          "data": {
            "image/png": "iVBORw0KGgoAAAANSUhEUgAAAb4AAAEuCAYAAADx63eqAAAABHNCSVQICAgIfAhkiAAAAAlwSFlz\nAAALEgAACxIB0t1+/AAAIAB
            REFUeJzsnXd4FGXXh+/t6T10EgIJoffeexcBkWZHED8URbEiNhRE\neQXFjgULTUERkI70EnoJgQCBkATSe7LJZtt8fzwhRVqABAh57uuaa2ZnZmcmm
            35zTnPKSpFUZBI\nJBKJpKKgvtsXIJFIJBLJnUQKn0QikUgqFFL4JBKJRFKhkMInkUgkkgqFFD6JRCKRVCik8EkkEomk\nQqG9wXaZ6yCRSCSS8ojqW
            ukxSeRSCSSCoUUPolEIpFUKKTwSSQSiaRCIYVPIpFIJBUKKXwSiUQi\nqVBI4ZNIJBJJhUIKn0QikUgqFFL4JBKJRFKhkMInkUgkkgqFFD6JRCKRVCi
            8EkkEomkQiGFTyKR\nSCQVCil8EolEIqlQSOGTSCQSSYVCCp9EIpFIKhRS+CQSiURSoZDCJ5FIJJIKhRQ+iUQikVQopPBJ\nJBKJpEIhhU8ikUgkFQo
```

Reproducible Research

An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Buckheit and Donoho, WaveLab and Reproducible Research, ***1995***

Nature: “the advertising”



Access provided to University of California-CDL by The University of California Libraries

NATURE GENETICS | ANALYSIS

日本語要約

Multi-tiered genomic analysis of head and neck cancer ties *TP53* mutation to 3p loss

Andrew M Gross, Ryan K Orosco, John P Shen, Ann Marie Egloff, Hannah Carter, Matan Hofree, Michel Choueiri, Charles S Coffey, Scott M Lippman, D Neil Hayes, Ezra E Cohen, Jennifer R Grandis, Quyen T Nguyen & Trey Ideker

Affiliations | Contributions | Corresponding author

Nature Genetics **46**, 939–943 (2014) | doi:10.1038/ng.3051
Received 20 March 2014 | Accepted 10 July 2014 | Published online 03 August 2014

Abstract

Abstract • Introduction • Results • Discussion • Methods • References • Acknowledgments • Author information • Supplementary information

Head and neck squamous cell carcinoma (HNSCC) is characterized by aggressive behavior with a propensity for metastasis and recurrence. Here we report a comprehensive analysis of the molecular and clinical features of HNSCC that govern patient survival. We find that *TP53* mutation is frequently accompanied by loss of chromosome 3p and that the combination of these events is associated with a surprising decrease in survival time (1.9 years versus >5 years for *TP53* mutation alone). The *TP53*-3p interaction is specific to chromosome 3p and validates in HNSCC and pan-cancer cohorts. In human papillomavirus (HPV)-positive tumors, in which HPV inactivates *TP53*, 3p deletion is also common and is associated with poor outcomes. The *TP53*-3p event is modified by mir-548k expression, which decreases survival further, and is mutually exclusive with mutations affecting RAS signaling. Together, the identified markers underscore the molecular heterogeneity of HNSCC and enable a new multi-tiered classification of this disease.

Editors' pick

Focus on TCGA Pan-Cancer Analysis

naturejobs.com

Joint Faculty Positions at the Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences
Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University

Assistant Professor in Human Genetics
UT Southwestern Medical Center at Dallas, TX

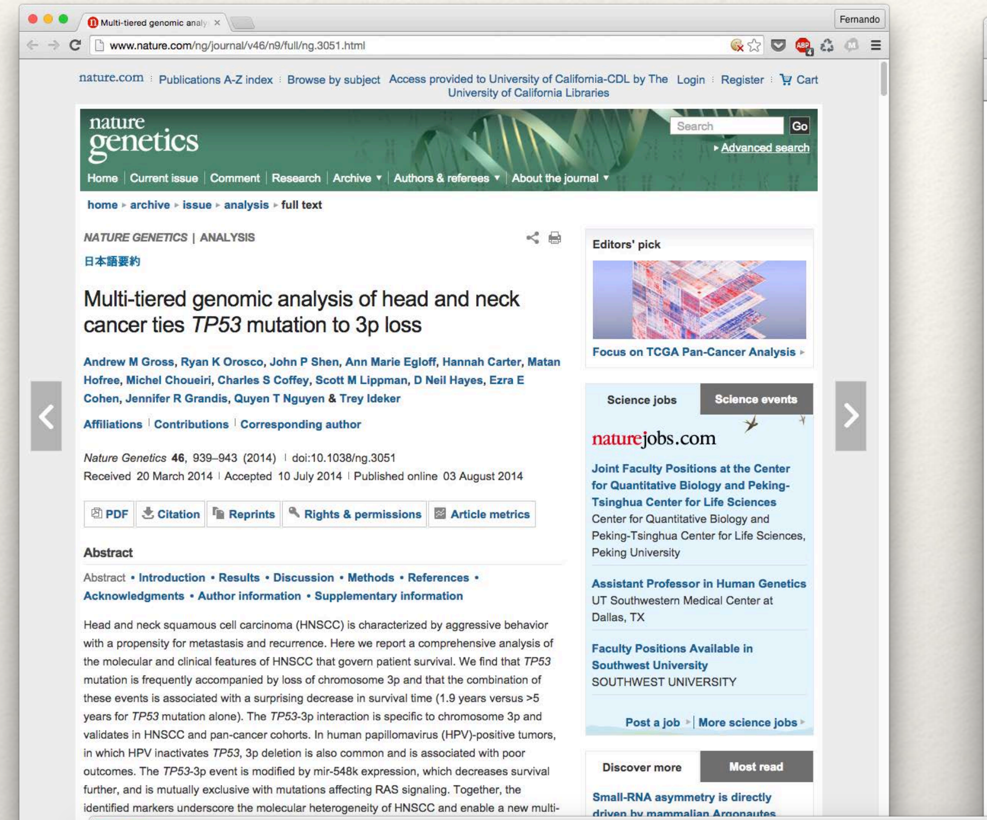
Faculty Positions Available in Southwest University
SOUTHWEST UNIVERSITY

Post a job | More science jobs

Small-RNA asymmetry is directly driven by mammalian Argonautes
Nature Structural & Molecular Biology | 22

Gross, Andrew M., et al. Nature genetics 46.9 (2014): 939-943.

Notebooks on Github: the “actual scolarship”


Multi-tiered genomic analy ×
Fernando
www.nature.com/ng/journal/v46/n9/full/ng.3051.html
nature.com
Publications A-Z index
Browse by subject
Access provided to University of California-CDL by The University of California Libraries
Login
Register
Cart
nature genetics
Search
Go
Advanced search
Home
Current issue
Comment
Research
Archive
Authors & referees
About the journal
home ▸ archive ▸ issue ▸ analysis ▸ full text

NATURE GENETICS | ANALYSIS

日本語要約

Multi-tiered genomic analysis of head and neck cancer ties *TP53* mutation to 3p loss

Andrew M Gross, Ryan K Orosco, John P Shen, Ann Marie Egloff, Hannah Carter, Matan Hofree, Michel Choueiri, Charles S Coffey, Scott M Lippman, D Neil Hayes, Ezra E Cohen, Jennifer R Grandis, Quyen T Nguyen & Trey Ideker

Affiliations | Contributions | Corresponding author

Nature Genetics **46**, 939–943 (2014) | doi:10.1038/ng.3051
Received 20 March 2014 | Accepted 10 July 2014 | Published online 03 August 2014

Abstract

Abstract • Introduction • Results • Discussion • Methods • References • Acknowledgments • Author information • Supplementary information

Head and neck squamous cell carcinoma (HNSCC) is characterized by aggressive behavior with a propensity for metastasis and recurrence. Here we report a comprehensive analysis of the molecular and clinical features of HNSCC that govern patient survival. We find that *TP53* mutation is frequently accompanied by loss of chromosome 3p and that the combination of these events is associated with a surprising decrease in survival time (1.9 years versus >5 years for *TP53* mutation alone). The *TP53*-3p interaction is specific to chromosome 3p and validates in HNSCC and pan-cancer cohorts. In human papillomavirus (HPV)-positive tumors, in which HPV inactivates *TP53*, 3p deletion is also common and is associated with poor outcomes. The *TP53*-3p event is modified by mir-548k expression, which decreases survival further, and is mutually exclusive with mutations affecting RAS signaling. Together, the identified markers underscore the molecular heterogeneity of HNSCC and enable a new multi-

Editors' pick

Focus on TCGA Pan-Cancer Analysis ▸

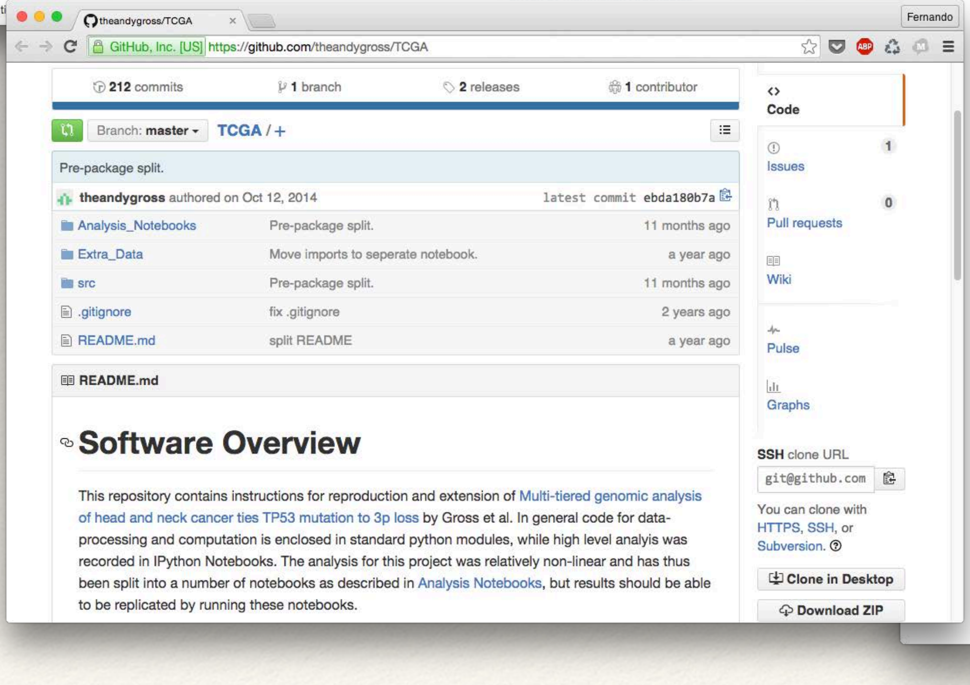
Joint Faculty Positions at the Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences
Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, Peking University

Assistant Professor in Human Genetics
UT Southwestern Medical Center at Dallas, TX

Faculty Positions Available in Southwest University
SOUTHWEST UNIVERSITY

Post a job ▸ | More science jobs ▸

Small-RNA asymmetry is directly driven by mammalian Argonautes



Pre-package split.

theandygross authored on Oct 12, 2014 `latest commit ebda180b7a`

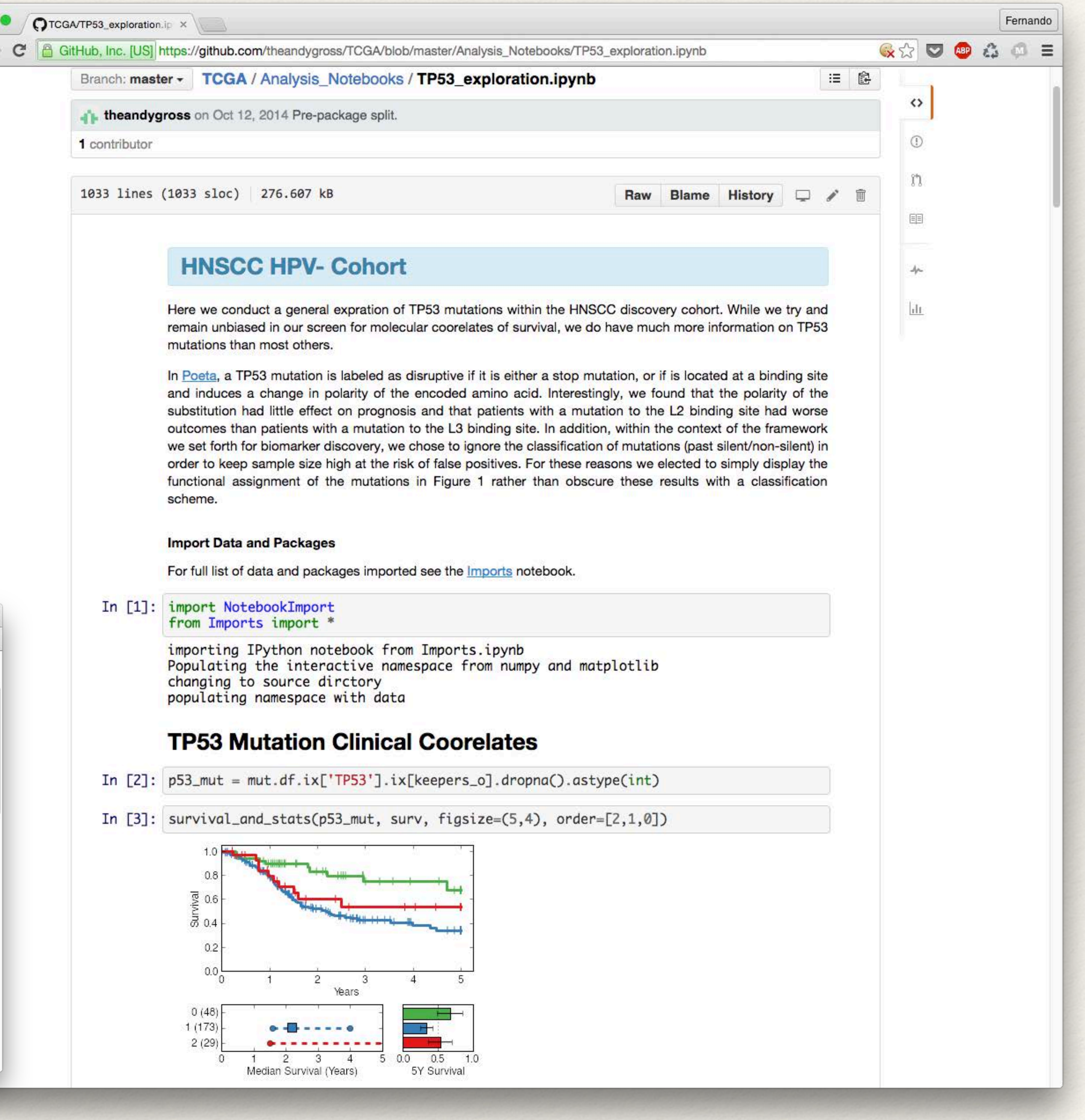
Analysis_Notebooks	Pre-package split.	11 months ago
Extra_Data	Move imports to seperate notebook.	a year ago
src	Pre-package split.	11 months ago
.gitignore	fix .gitignore	2 years ago
README.md	split README	a year ago

README.md

Software Overview

This repository contains instructions for reproduction and extension of Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss by Gross et al. In general code for data-processing and computation is enclosed in standard python modules, while high level analyis was recorded in IPython Notebooks. The analysis for this project was relatively non-linear and has thus been split into a number of notebooks as described in Analysis Notebooks, but results should be able to be replicated by running these notebooks.

You can clone with HTTPS, SSH, or Subversion.


TCGA/TP53_exploration.ip ×
Fernando
GitHub, Inc. [US] https://github.com/theandygross/TCGA/blob/master/Analysis_Notebooks/TP53_exploration.ipynb
Branch: master
TCGA / Analysis_Notebooks / TP53_exploration.ipynb
Raw
Blame
History
1.0
0.8
0.6
0.4
0.2
0.0
Survival
0
1
2
3
4
5
Years
0 (48)
1 (173)
2 (29)
Median Survival (Years)
0.0
0.5
1.0
5Y Survival

theandygross on Oct 12, 2014 Pre-package split.

1 contributor

`1033 lines (1033 sloc) | 276.607 kB`

HNSCC HPV- Cohort

Here we conduct a general expration of TP53 mutations within the HNSCC discovery cohort. While we try and remain unbiased in our screen for molecular coorelates of survival, we do have much more information on TP53 mutations than most others.

In Poeta, a TP53 mutation is labeled as disruptive if it is either a stop mutation, or if is located at a binding site and induces a change in polarity of the encoded amino acid. Interestingly, we found that the polarity of the substitution had little effect on prognosis and that patients with a mutation to the L2 binding site had worse outcomes than patients with a mutation to the L3 binding site. In addition, within the context of the framework we set forth for biomarker discovery, we chose to ignore the classification of mutations (past silent/non-silent) in order to keep sample size high at the risk of false positives. For these reasons we elected to simply display the functional assignment of the mutations in Figure 1 rather than obscure these results with a classification scheme.

Import Data and Packages

For full list of data and packages imported see the Imports notebook.

```
In [1]: import NotebookImport
        from Imports import *
```

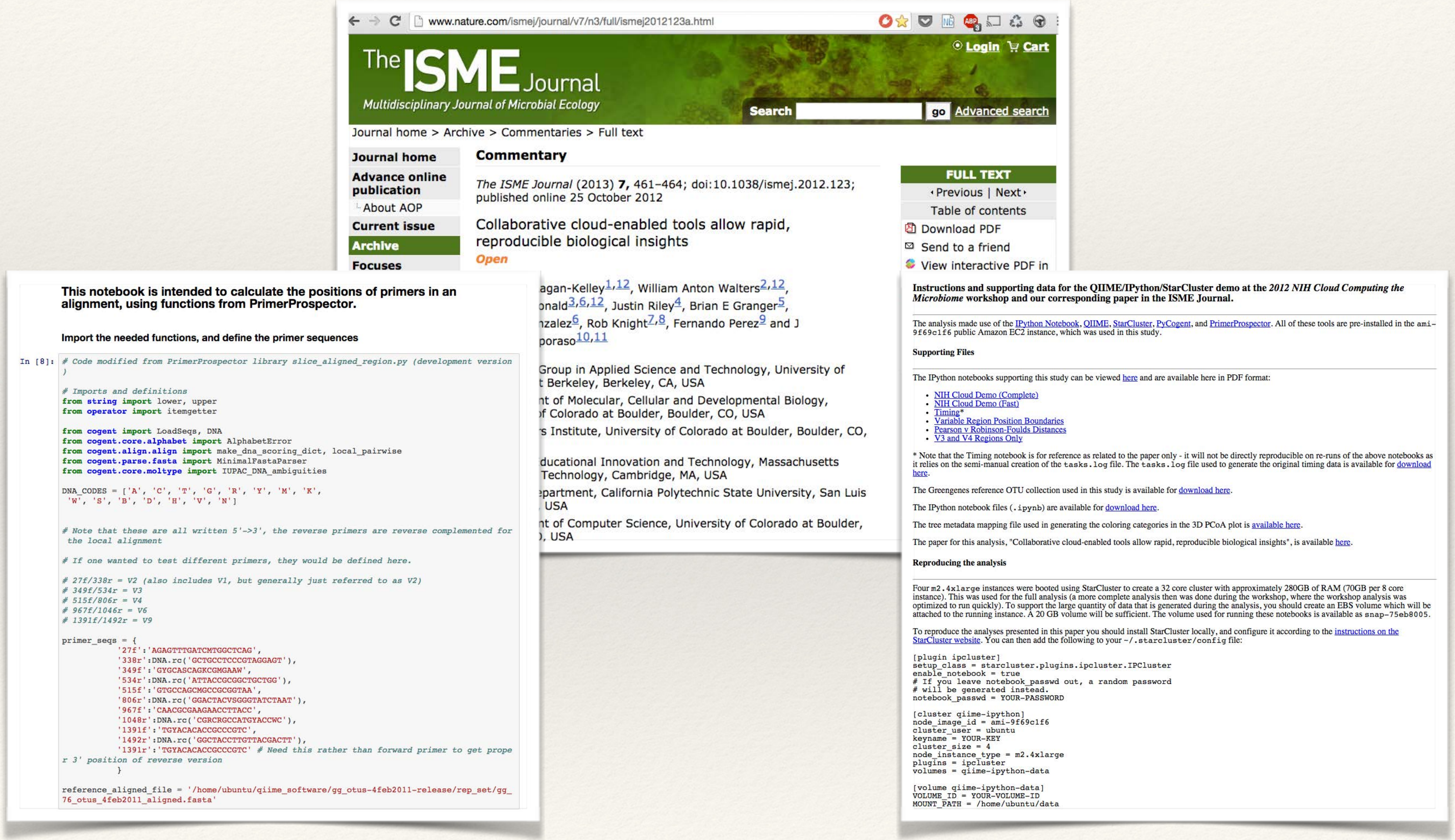
```
importing IPython notebook from Imports.ipynb
Populating the interactive namespace from numpy and matplotlib
changing to source dirctory
populating namespace with data
```

TP53 Mutation Clinical Coorelates

```
In [2]: p53_mut = mut.df.ix['TP53'].ix[keepers_o].dropna().astype(int)
```

```
In [3]: survival_and_stats(p53_mut, surv, figsize=(5,4), order=[2,1,0])
```

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



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

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

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
epartment, California Polytechnic State University, San Luis
USA
nt of Computer Science, University of Colorado at Boulder,
), USA

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':'AGAGTTTGATCMTGGCTCAG',
                   '338r':DNA.rc('GCTGCCTCCCGTAGGAGT'),
                   '349f':'GYGCASCAGKCGMGAAW',
                   '534r':DNA.rc('ATTACCGCGGCTGCTGG'),
                   '515f':'GTGCCAGCMGCCGCGGTAA',
                   '806r':DNA.rc('GGACTACVSGGGTATCTAAT'),
                   '967f':'CAACGCGAAGAACCTTACC',
                   '1048r':DNA.rc('CGRCRGCCATGYACCWC'),
                   '1391f':'TGYACACACCGCCCGTC',
                   '1492r':DNA.rc('GGCTACCTTGTTACGACTT'),
                   '1391r':'TGYACACACCGCCCGTC' # Need this rather than forward primer to get prope
        r 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`.

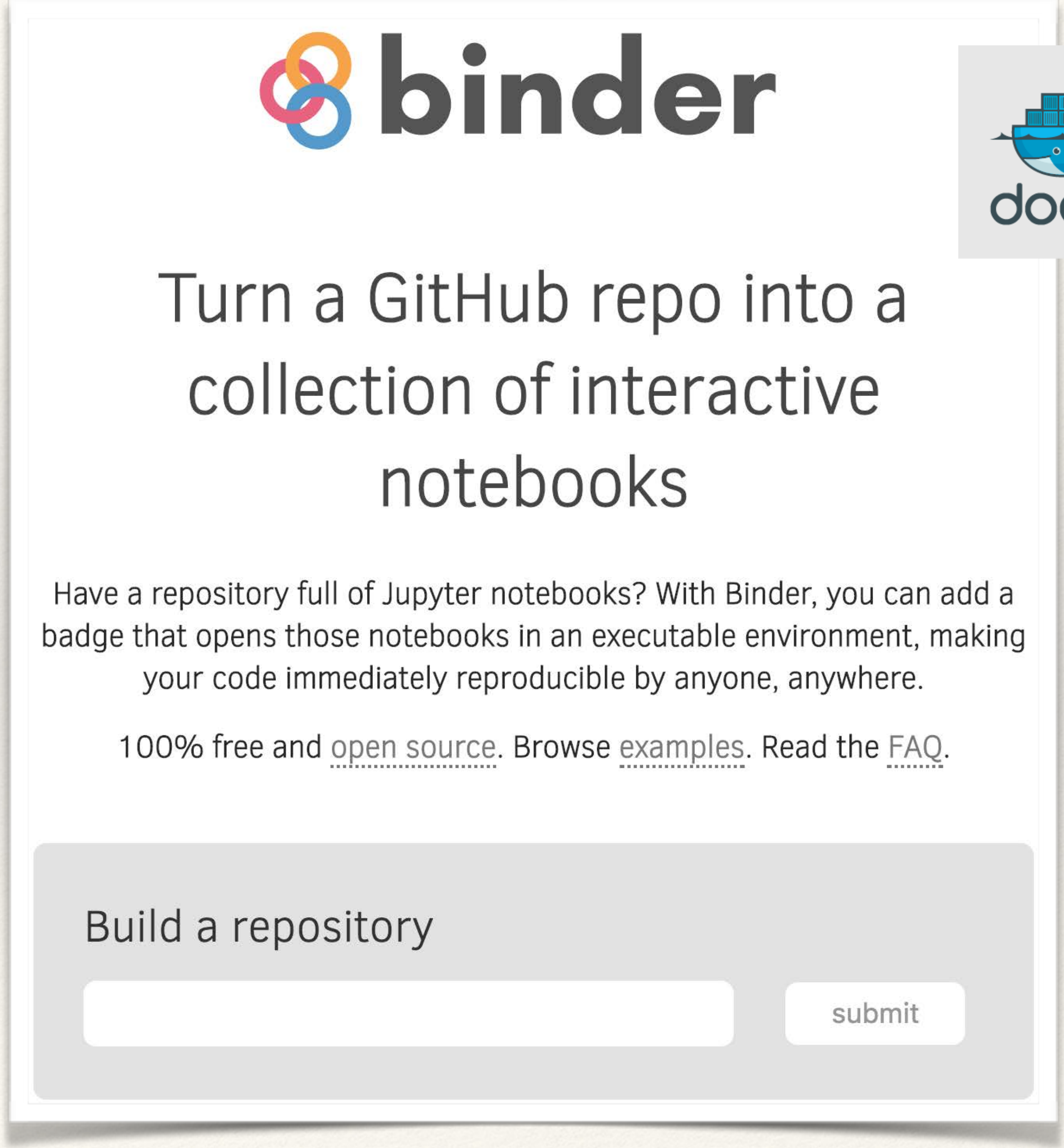
To reproduce the analyses presented in this paper you should install StarCluster locally, and configure it according to the instructions on the StarCluster website. You can then add the following to your `~/.starcluster/config` file:

```
[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
```

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

http://qiime.org/home_static/nih-cloud-apr2012

mybinder.org



binder

Turn a GitHub repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, you can add a badge that opens those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

100% free and open source. Browse examples. Read the FAQ.

Build a repository

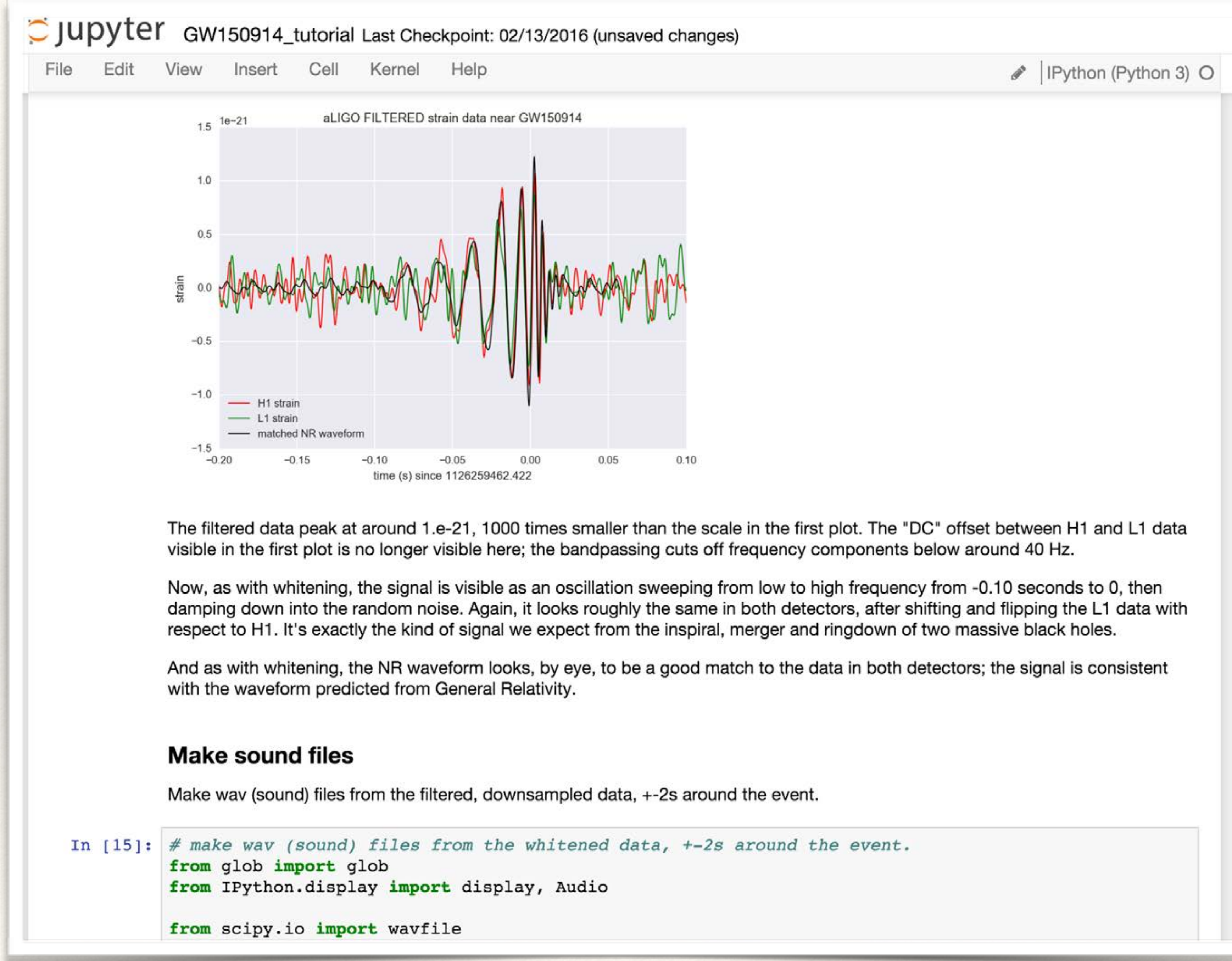
docker

github.com/freeman-lab

github.com/andrewosh

Andrew Osheroff's SciPy'16 talk:
https://www.youtube.com/watch?v=OK6M4w7LYIc

Gravitational waves detected on Jupyter!



The filtered data peak at around 1.e-21, 1000 times smaller than the scale in the first plot. The "DC" offset between H1 and L1 data visible in the first plot is no longer visible here; the bandpassing cuts off frequency components below around 40 Hz.

Now, as with whitening, the signal is visible as an oscillation sweeping from low to high frequency from -0.10 seconds to 0, then damping down into the random noise. Again, it looks roughly the same in both detectors, after shifting and flipping the L1 data with respect to H1. It's exactly the kind of signal we expect from the inspiral, merger and ringdown of two massive black holes.

And as with whitening, the NR waveform looks, by eye, to be a good match to the data in both detectors; the signal is consistent with the waveform predicted from General Relativity.

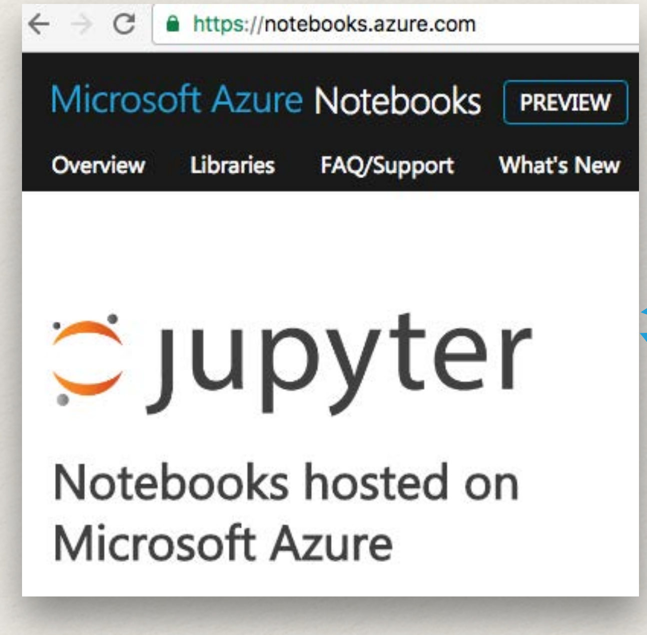
Make sound files

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

In [15]:
```
# make wav (sound) files from the whitened data, +-2s around the event.
from glob import glob
from IPython.display import display, Audio

from scipy.io import wavfile
```

From LIGO Open Science Center, binder-ified: github.com/minrk/ligo-binder

LIGO: Open Science with Jupyter



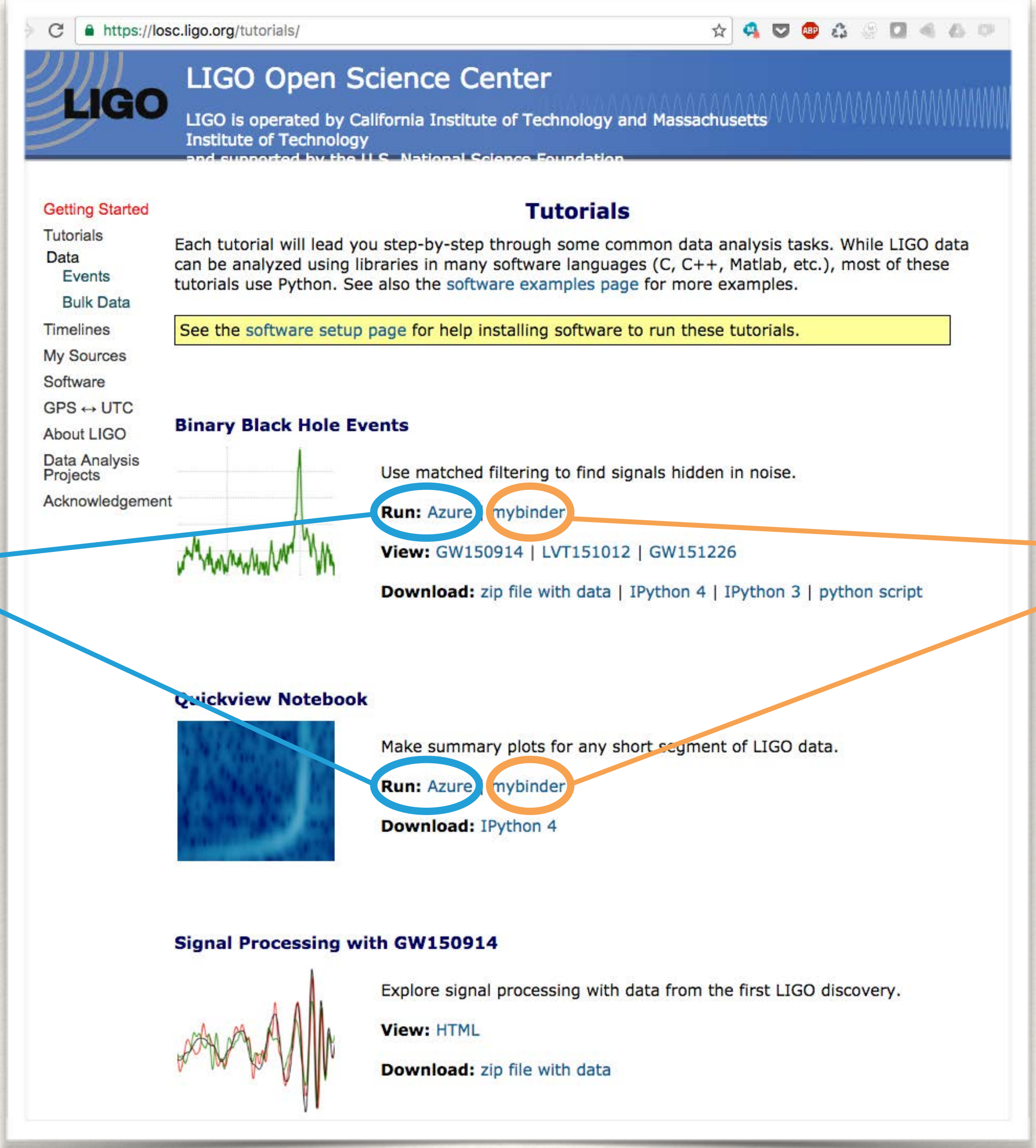
https://notebooks.azure.com

Microsoft Azure Notebooks PREVIEW

Overview Libraries FAQ/Support What's New

jupyter

Notebooks hosted on Microsoft Azure



https://losc.ligo.org/tutorials/

LIGO

LIGO Open Science Center

LIGO is operated by California Institute of Technology and Massachusetts Institute of Technology
and supported by the U.S. National Science Foundation

Getting Started
Tutorials
Data
Events
Bulk Data
Timelines
My Sources
Software
GPS ↔ UTC
About LIGO
Data Analysis Projects
Acknowledgement

Tutorials

Each tutorial will lead you step-by-step through some common data analysis tasks. While LIGO data can be analyzed using libraries in many software languages (C, C++, Matlab, etc.), most of these tutorials use Python. See also the software examples page for more examples.

See the software setup page for help installing software to run these tutorials.

Binary Black Hole Events

Use matched filtering to find signals hidden in noise.

Run: Azure | mybinder

View: GW150914 | LVT151012 | GW151226

Download: zip file with data | IPython 4 | IPython 3 | python script

Quickview Notebook

Make summary plots for any short segment of LIGO data.

Run: Azure | mybinder

Download: IPython 4

Signal Processing with GW150914

Explore signal processing with data from the first LIGO discovery.

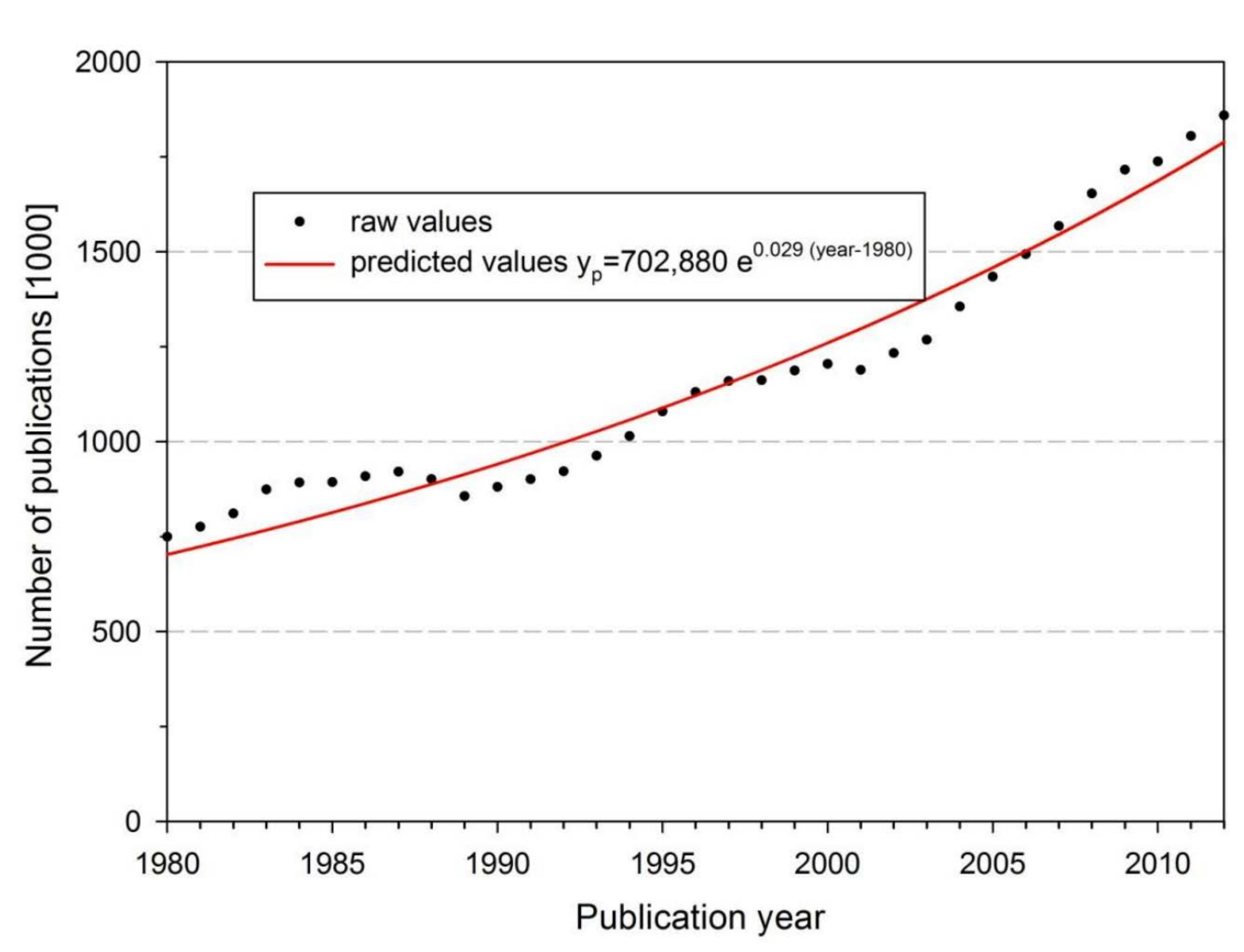
View: HTML

Download: zip file with data

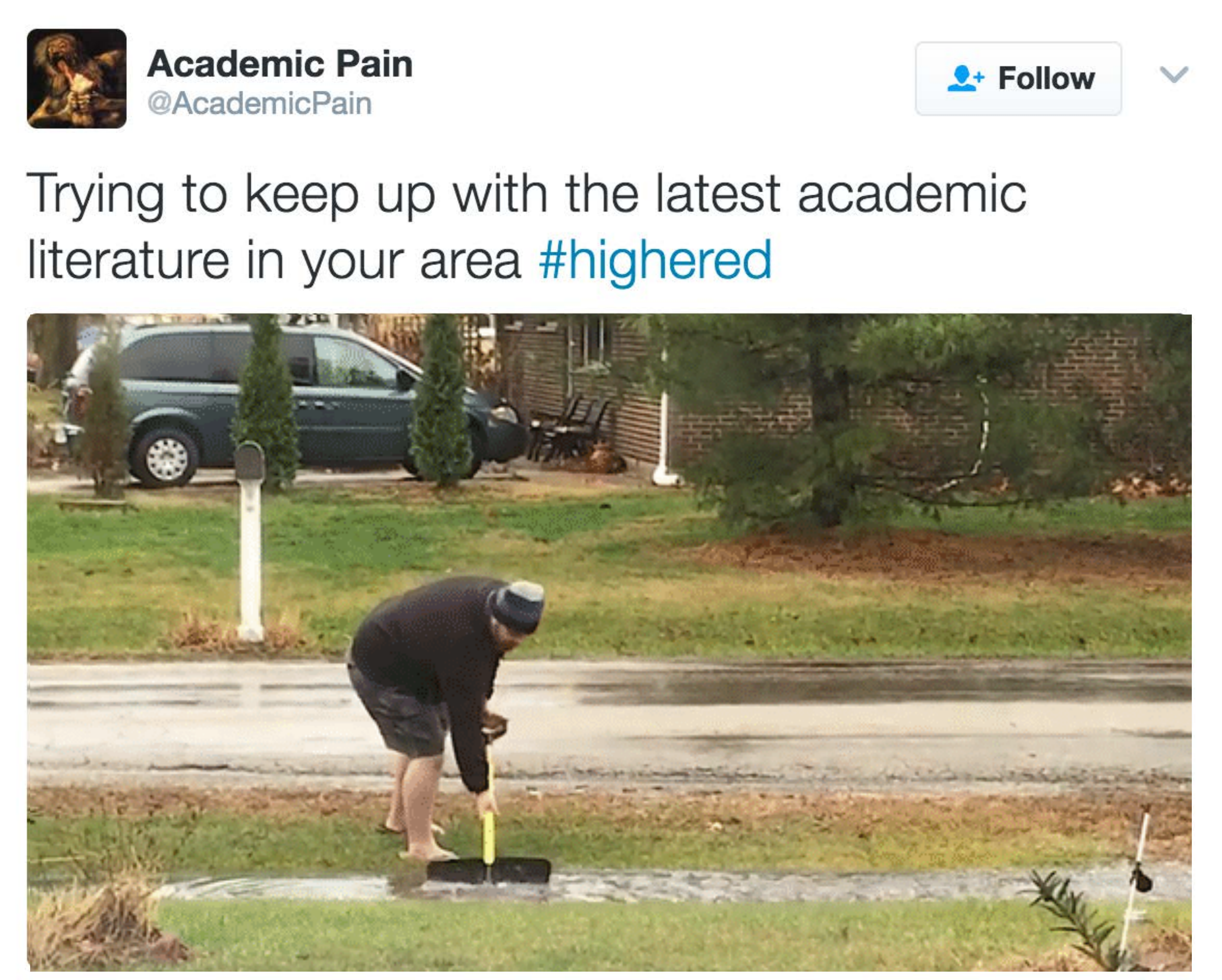
binder

The future of reproducible science?

Global scientific output doubles every nine years



Bornmann & Mutz, arxiv.org/1402.4578, Nature News Blog, May 2014



Academic Pain
@AcademicPain

Trying to keep up with the latest academic literature in your area #highered

Who is reading the literature?

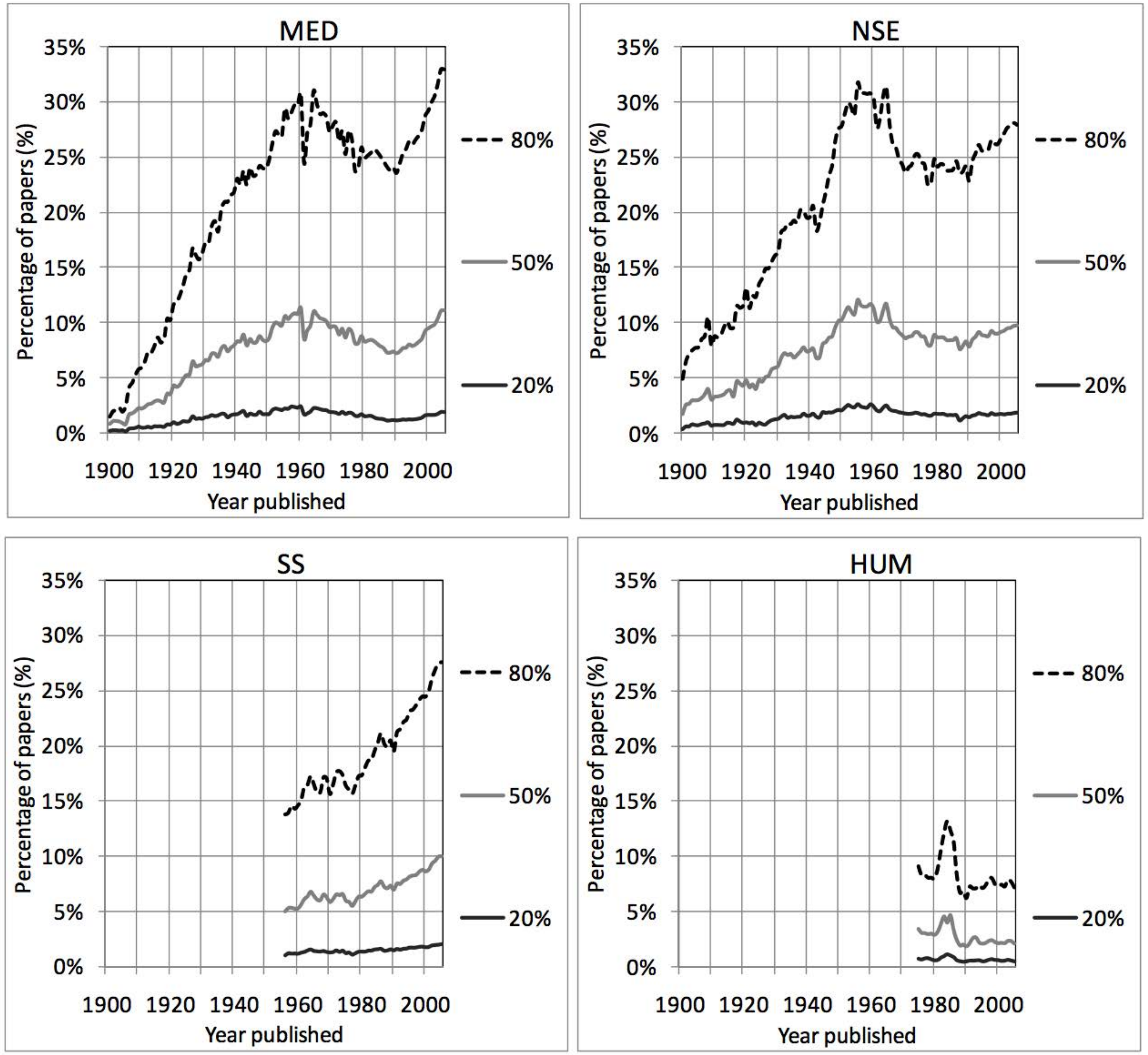


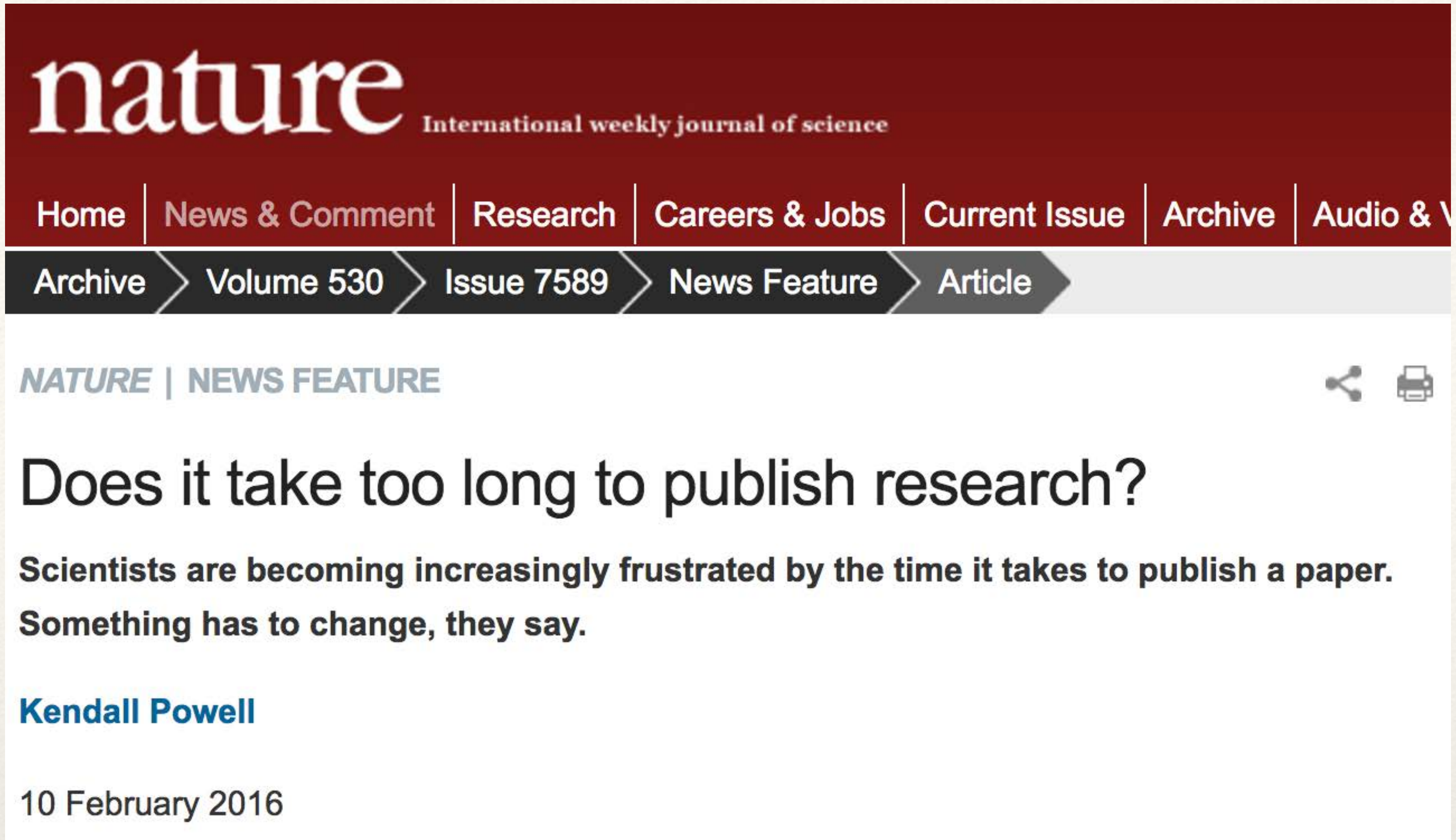
Figure 2. Percentage of papers needed to obtain 20%, 50% and 80% of the citations received using a two-year citation window, by field, 1900–2005

Larivière & Gingras, arxiv.org/0809.5250

The scientific literature, today

We are conflating two things:

1. Communication of ideas for others to build upon (hence, reproducibility)
2. Professional credit



NATURE | NEWS FEATURE

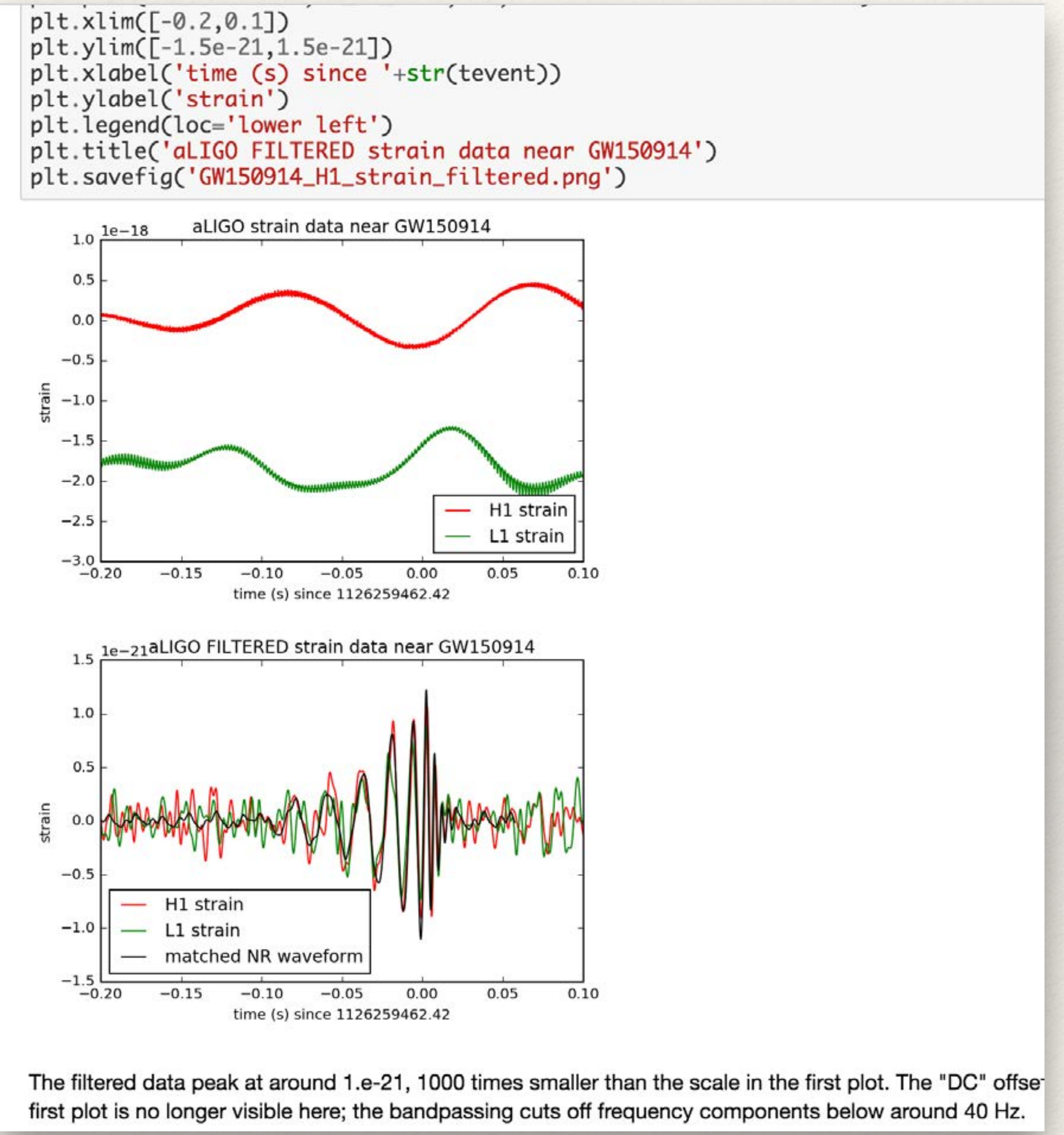
Does it take too long to publish research?

Scientists are becoming increasingly frustrated by the time it takes to publish a paper. Something has to change, they say.

Kendall Powell

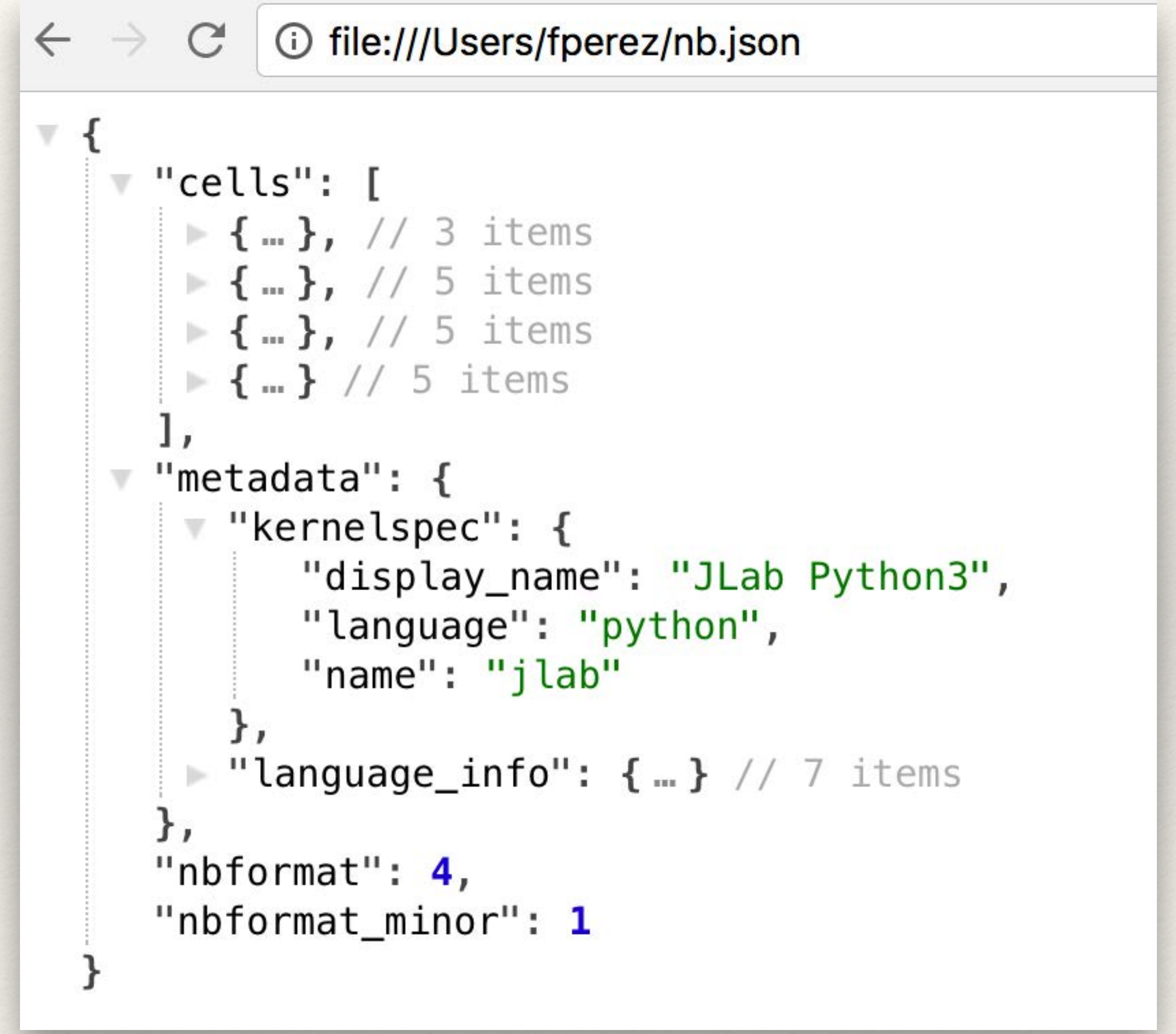
10 February 2016

The literature will be read by the machines



```
plt.xlim([-0.2,0.1])
plt.ylim([-1.5e-21,1.5e-21])
plt.xlabel('time (s) since '+str(tevent))
plt.ylabel('strain')
plt.legend(loc='lower left')
plt.title('aLIGO FILTERED strain data near GW150914')
plt.savefig('GW150914_H1_strain_filtered.png')
```

The filtered data peak at around 1.e-21, 1000 times smaller than the scale in the first plot. The "DC" offse first plot is no longer visible here; the bandpassing cuts off frequency components below around 40 Hz.



```
{
  "cells": [
    {…}, // 3 items
    {…}, // 5 items
    {…}, // 5 items
    {…} // 5 items
  ],
  "metadata": {
    "kernelspec": {
      "display_name": "JLab Python3",
      "language": "python",
      "name": "jlab"
    },
    "language_info": {…} // 7 items
  },
  "nbformat": 4,
  "nbformat_minor": 1
}
```

LIGO GW150914 analysis as Jupyter Notebook. 1,000,000+ of these on

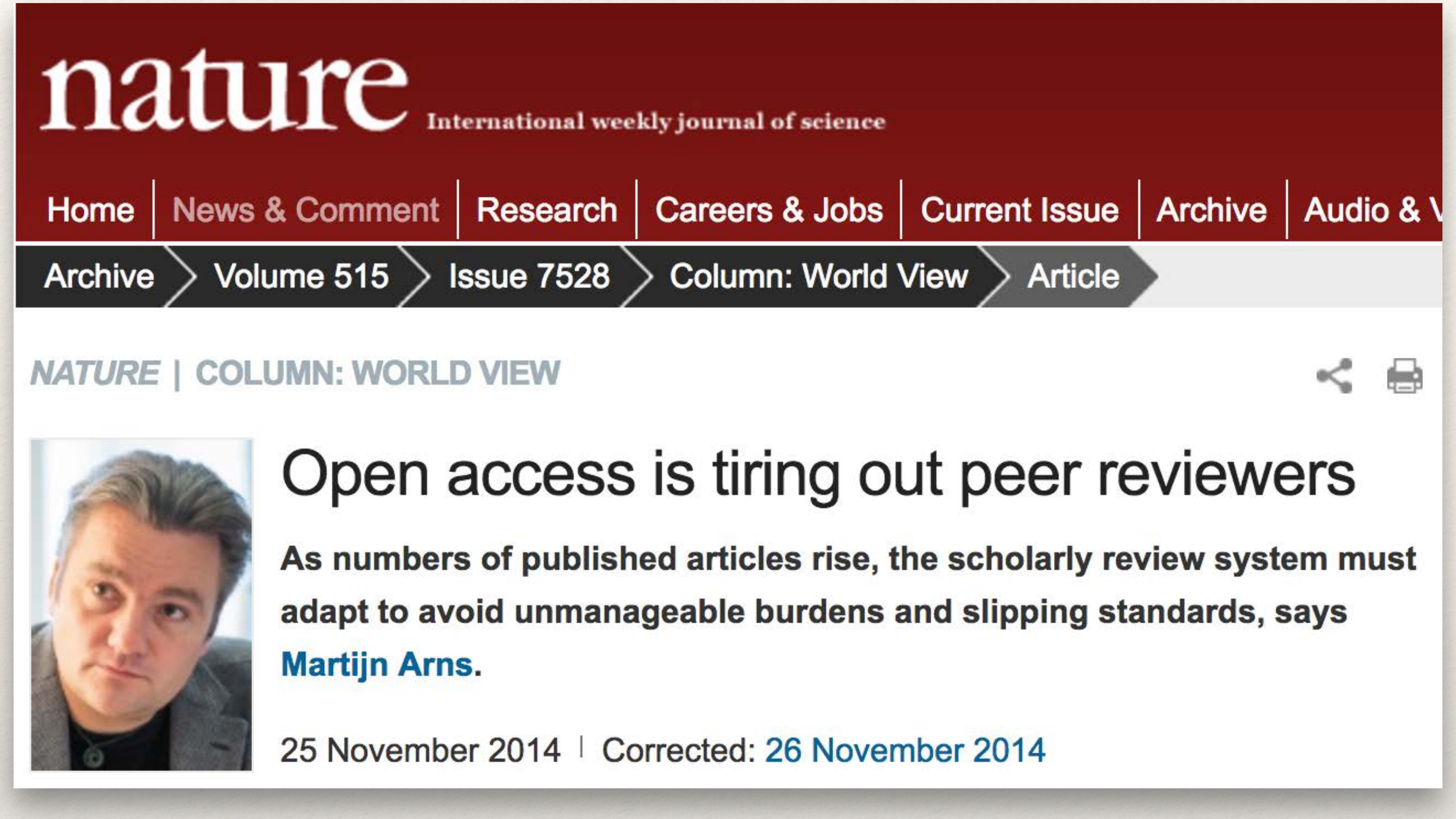
Let’s “publish” less so we can read more!

What if…

- All our daily work was captured in a way the machines could read…
- annotated with rich metadata…
- natural language, code, results and data all linked…
- easy for the machines to mine for discovery and credit…
- and less frequent highlights were written in long form, also backed by their "real scholarship" (à la Donoho)?

What would that look like?

- “Executable preprints/blog posts”
  - Capture rapid progress, expose data and software
  - Fully reproducible: build scientific community and knowledge
  - With DOIs - citable as needed.
- Peer-reviewed papers:
  - less frequent, high-quality narratives
  - real synthesis of important ideas



NATURE | COLUMN: WORLD VIEW

Open access is tiring out peer reviewers

As numbers of published articles rise, the scholarly review system must adapt to avoid unmanageable burdens and slipping standards, says Martijn Arns.

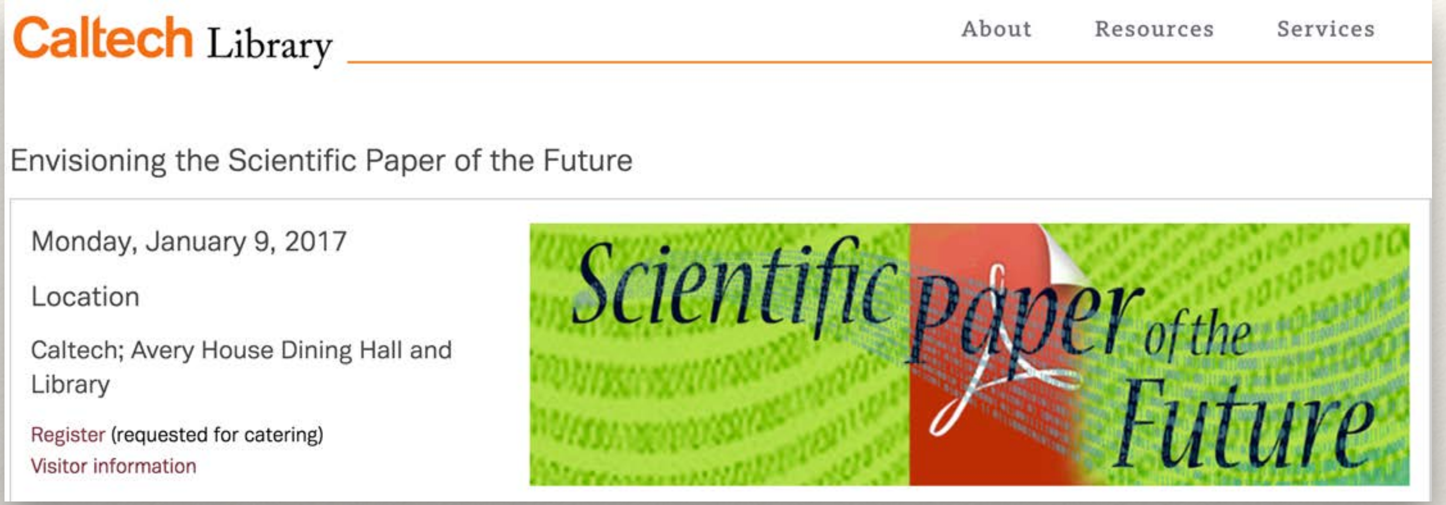
25 November 2014 | Corrected: 26 November 2014

But in recent months, I received reviews of my own submitted papers that suggest reviewers simply did not read the manuscript properly.

[…]

To protect quality reviewing, a hybrid model should be considered. I suggest a two-tier system, in which some papers are not reviewed before publication at all and are instead subject to a post-publication peer review.

The “scientific paper of the future”



Caltech Library

About Resources Services

Envisioning the Scientific Paper of the Future

Monday, January 9, 2017

Location

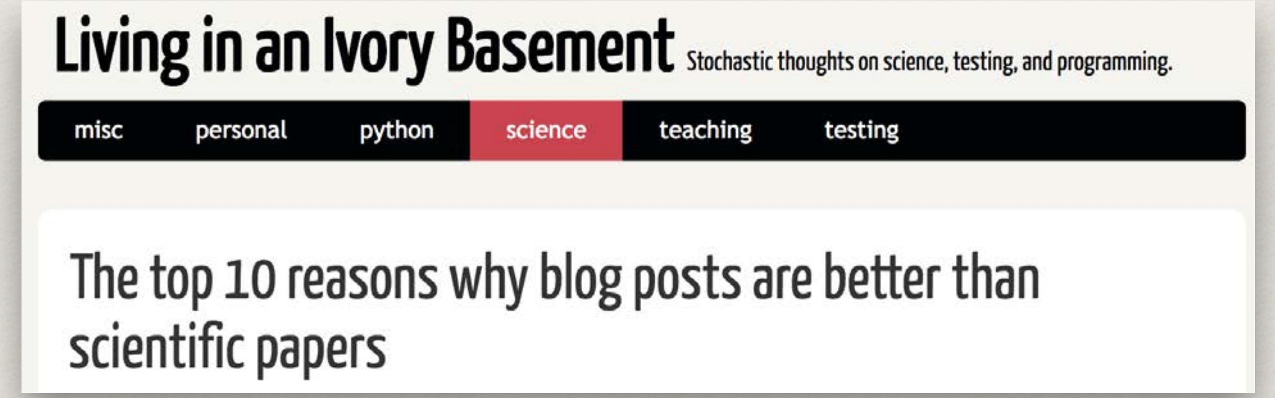
Caltech; Avery House Dining Hall and Library

Register (requested for catering)

Visitor information

Victoria Stodden

Yolanda Gil



Living in an Ivory Basement Stochastic thoughts on science, testing, and programming.

misc personal python science teaching testing

The top 10 reasons why blog posts are better than scientific papers

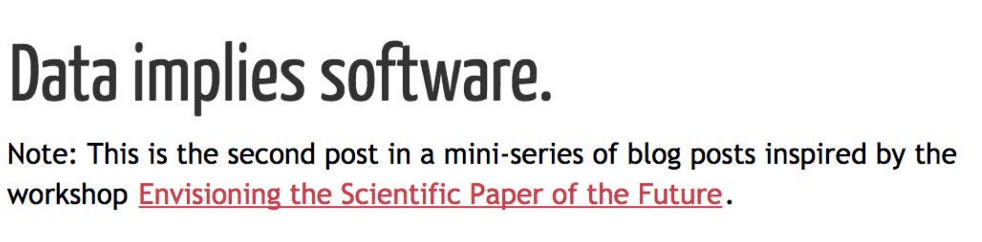
Titus Brown

AGU American Geophysical Union

Geoscience Papers of the Future OntoSoft

The Geoscience Papers of the Future Initiative

The Geoscience Papers of the Future (GPF) is an initiative to encourage geoscientists to publish papers together with the associated digital products of their research

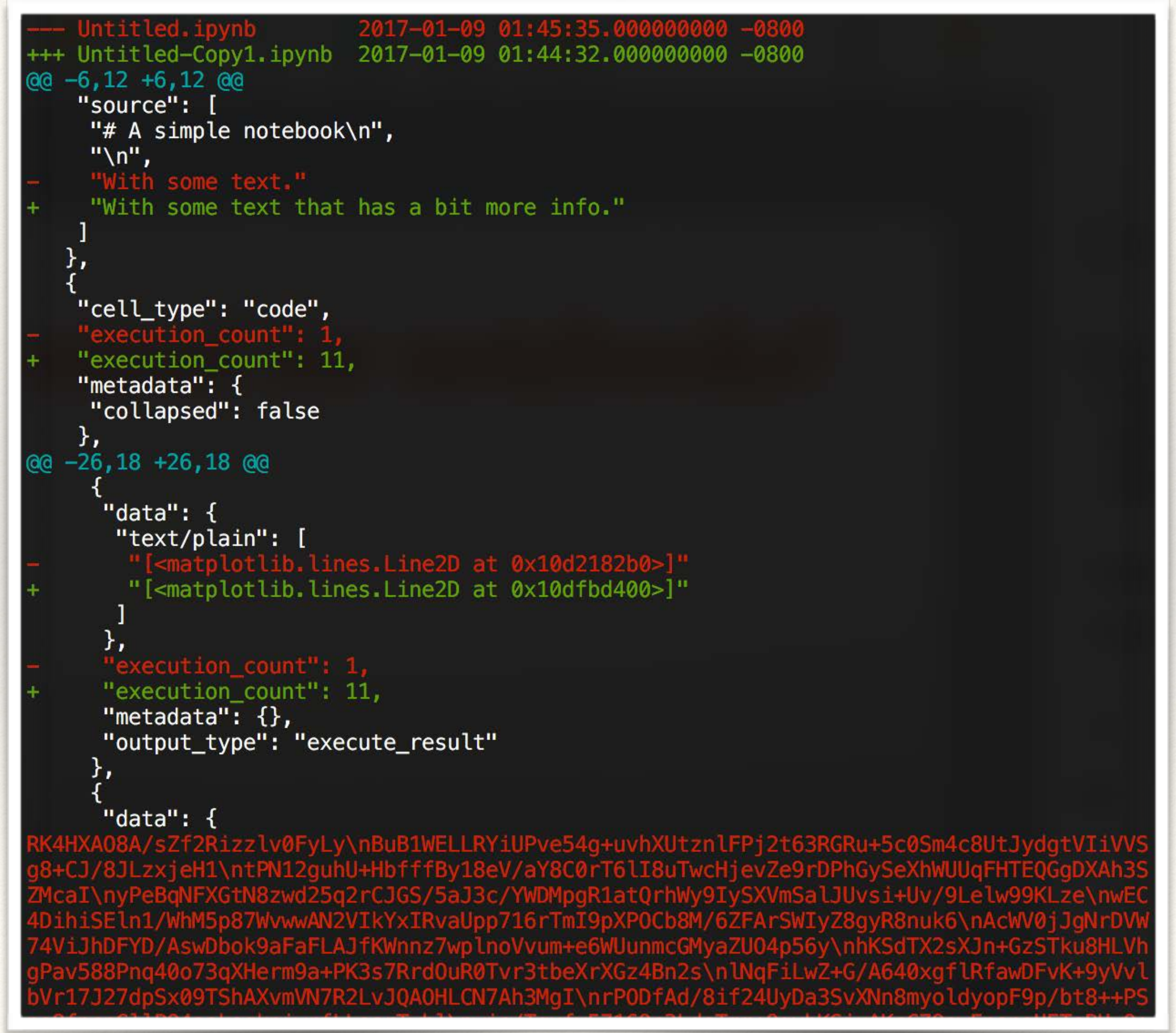


Data implies software.

Note: This is the second post in a mini-series of blog posts inspired by the workshop Envisioning the Scientific Paper of the Future.

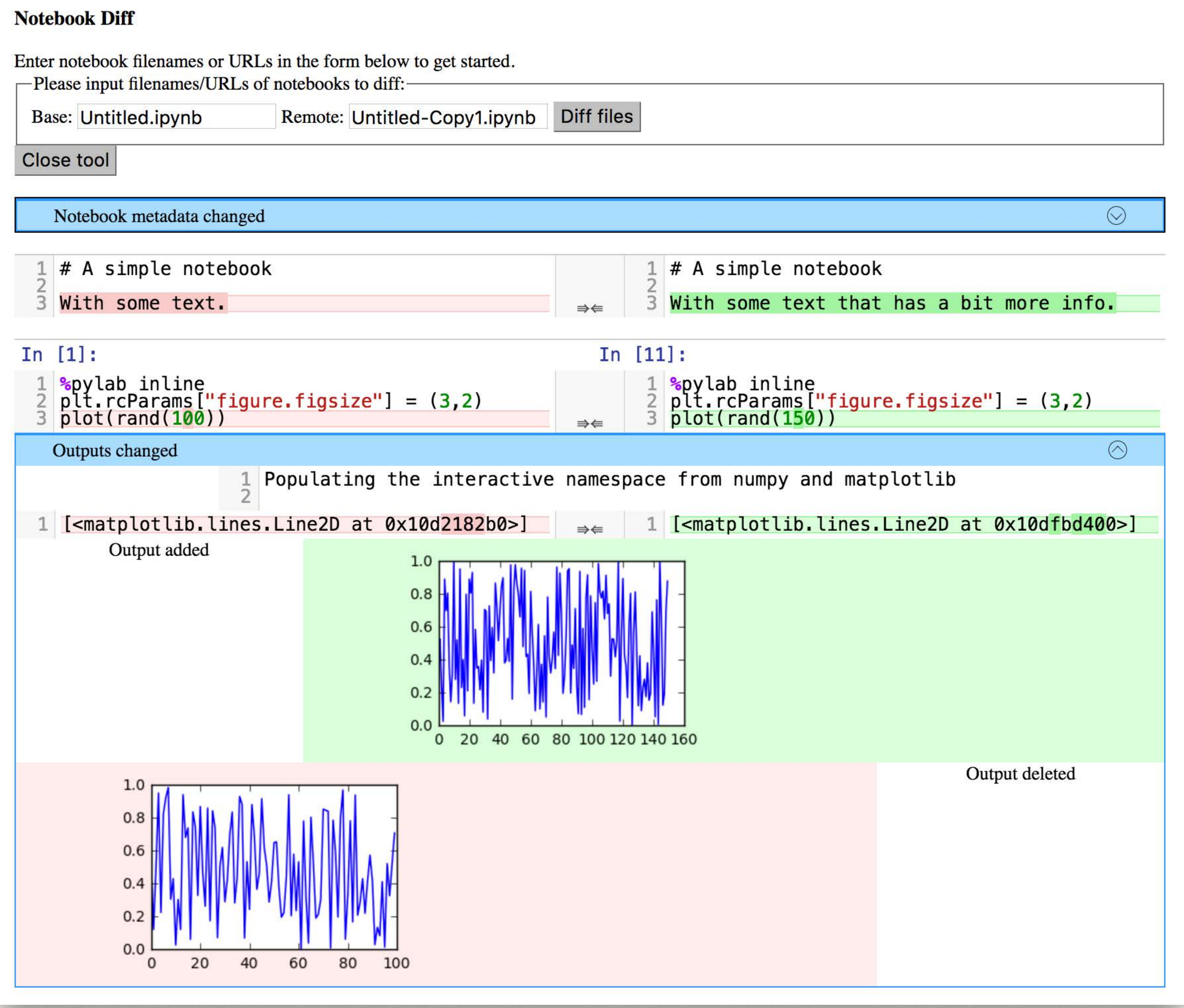
Some new developments in Jupyter’s orbit…

version control for notebooks?



```
--- Untitled.ipynb        2017-01-09 01:45:35.000000000 -0800
+++ Untitled-Copy1.ipynb  2017-01-09 01:44:32.000000000 -0800
@@ -6,12 +6,12 @@
     "source": [
      "# A simple notebook\n",
      "\n",
-     "With some text."
+     "With some text that has a bit more info."
     ]
    },
    {
     "cell_type": "code",
-    "execution_count": 1,
+    "execution_count": 11,
     "metadata": {
      "collapsed": false
     },
@@ -26,18 +26,18 @@
      {
       "data": {
        "text/plain": [
-        "[<matplotlib.lines.Line2D at 0x10d2182b0>]"
+        "[<matplotlib.lines.Line2D at 0x10dfbd400>]"
        ]
       },
-      "execution_count": 1,
+      "execution_count": 11,
       "metadata": {},
       "output_type": "execute_result"
      },
      {
       "data": {
RK4HXA08A/sZf2Rizzlv0FyLy\nBuB1WELLRYiUPve54g+uvhXUtznlFPj2t63RGRu+5c0Sm4c8UtJydgtVIiVVS
g8+CJ/8JLzxjeH1\ntPN12guhU+HbfffBy18eV/aY8C0rT6lI8uTwcHjevZe9rDPhGySeXhWUUqFHTEQGgDXAh3S
ZMcaI\nyPeBqNFXGtN8zwd25q2rCJGS/5aJ3c/YWDMpgR1atQrhWy9IySXVmSalJUvsi+Uv/9Lelw99KLze\nwEC
4DihiSEln1/WhM5p87WvwwAN2VIkYxIRvaUpp716rTmI9pXPOCb8M/6ZFArSWIyZ8gyR8nuk6\nAcWV0jJgNrDVW
74ViJhDFYD/AswDbok9aFaFLAJfKWnnz7wplnoVvum+e6WUunmcGMyaZU04p56y\nhKSdTX2sXJn+GzSTku8HLVh
gPav588Pnq40o73qXHerm9a+PK3s7RrdOuR0Tvr3tbeXrXGz4Bn2s\nlNqFiLwZ+G/A640xgflRfawDFvK+9yVvl
bVr17J27dpSx09TShAXvmVN7R2LvJQAOHLCN7Ah3MgI\nrPODfAd/8if24UyDa3SvXNn8myoldyopF9p/bt8++PS
```

nbdime to the rescue!



Notebook Diff

Enter notebook filenames or URLs in the form below to get started.

Please input filenames/URLs of notebooks to diff:

Base: Untitled.ipynb Remote: Untitled-Copy1.ipynb

Notebook metadata changed

```
# A simple notebook

With some text.
```

```
# A simple notebook

With some text that has a bit more info.
```

In [1]:

```
%pylab inline
plt.rcParams["figure.figsize"] = (3,2)
plot(rand(100))
```

In [11]:

```
%pylab inline
plt.rcParams["figure.figsize"] = (3,2)
plot(rand(150))
```

Outputs changed

```
Populating the interactive namespace from numpy and matplotlib

```

```
[<matplotlib.lines.Line2D at 0x10d2182b0>]
```

```
[<matplotlib.lines.Line2D at 0x10dfbd400>]
```

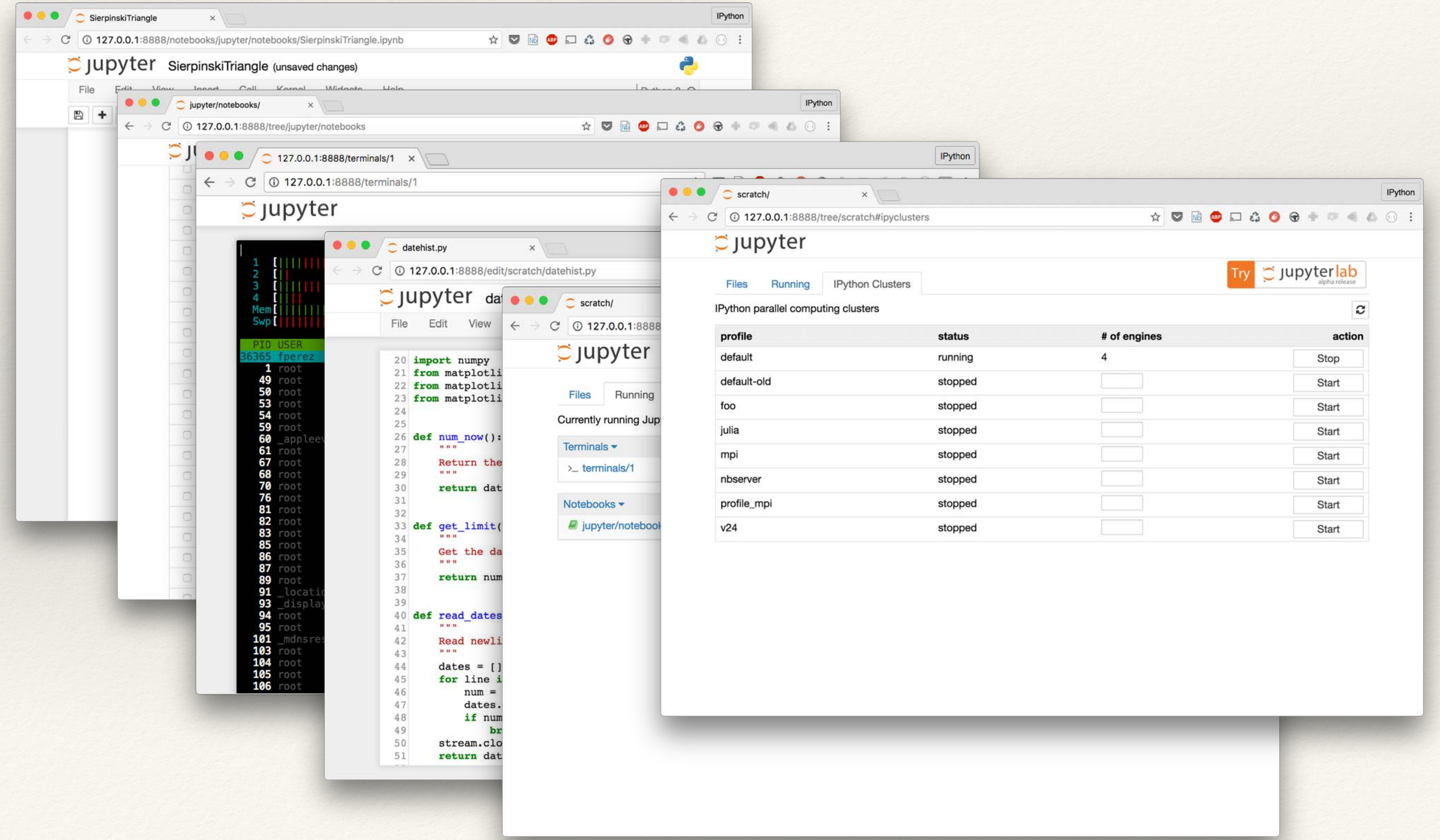
Output added

Output deleted

(**no**te**b**ook **di**ff and **me**rge: https://github.com/jupyter/nbdime

JupyterLab: the notebook, evolved…

The “Notebook”?


SierpinskiTriangle
IPython
127.0.0.1:8888/notebooks/jupyter/notebooks/SierpinskiTriangle.ipynb
jupyter SierpinskiTriangle (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
jupyter/notebooks/
127.0.0.1:8888/tree/jupyter/notebooks
127.0.0.1:8888/terminals/1
127.0.0.1:8888/terminals/1
jupyter
datehist.py
127.0.0.1:8888/edit/scratch/datehist.py
jupyter da
File Edit View
scratch/
127.0.0.1:8888
jupyter
Files Running
scratch/
127.0.0.1:8888/tree/scratch#ipyclusters
jupyter
Try jupyterlab alpha release
Files Running IPython Clusters

```
  1  [||||||
  2  [||
  3  [||||||
  4  [|||
  Mem[|||||||
  Swp[||||||

  PID USER
36365 fperez
    1 root
   49 root
   50 root
   53 root
   54 root
   59 root
   60 _appleev
   61 root
   67 root
   68 root
   70 root
   76 root
   81 root
   82 root
   83 root
   85 root
   86 root
   87 root
   89 root
   91 _locatio
   93 _display
   94 root
   95 root
  101 _mdnsres
  103 root
  104 root
  105 root
  106 root
```

```
import numpy
from matplotli
from matplotli
from matplotli


def num_now():
    """
    Return the
    """
    return dat


def get_limit(
    """
    Get the da
    """
    return num


def read_dates
    """
    Read newli
    """
    dates = []
    for line i
        num =
        dates.
        if num
            br
    stream.clo
    return dat
```

Currently running Jup

Terminals ▾

>_ terminals/1

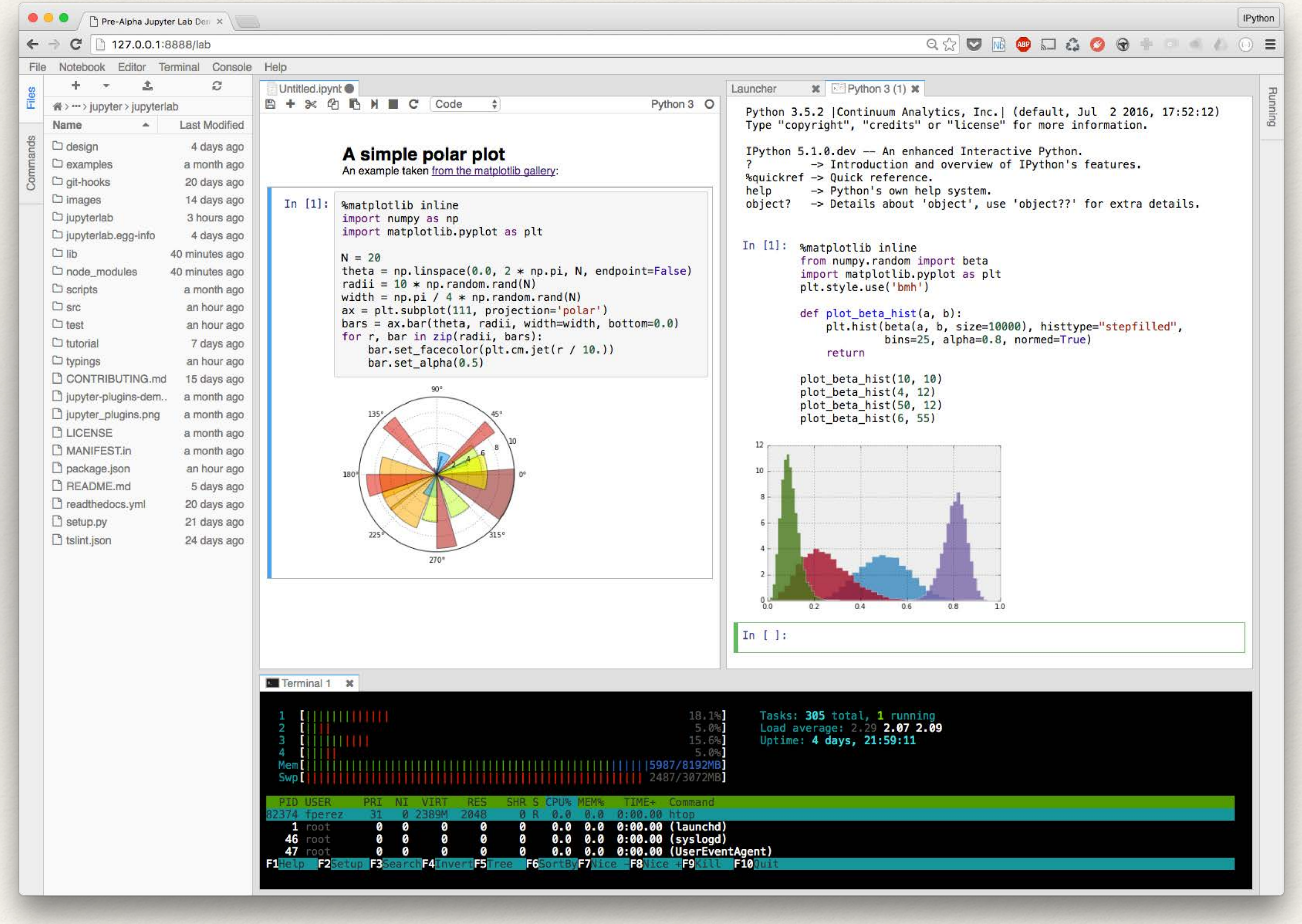
Notebooks ▾

jupyter/notebook

IPython parallel computing clusters

profile	status	# of engines	action
default	running	4	Stop
default-old	stopped		Start
foo	stopped		Start
julia	stopped		Start
mpi	stopped		Start
nbserver	stopped		Start
profile_mpi	stopped		Start
v24	stopped		Start

JupyterLab: unifying these ideas



Name	Last Modified
design	4 days ago
examples	a month ago
git-hooks	20 days ago
images	14 days ago
jupyterlab	3 hours ago
jupyterlab.egg-info	4 days ago
lib	40 minutes ago
node_modules	40 minutes ago
scripts	a month ago
src	an hour ago
test	an hour ago
tutorial	7 days ago
typings	an hour ago
CONTRIBUTING.md	15 days ago
jupyter-plugins-dem..	a month ago
jupyter_plugins.png	a month ago
LICENSE	a month ago
MANIFEST.in	a month ago
package.json	an hour ago
README.md	5 days ago
readthedocs.yml	20 days ago
setup.py	21 days ago
tslint.json	24 days ago

A simple polar plot

An example taken from the matplotlib gallery:

```
In [1]: %matplotlib inline
        import numpy as np
        import matplotlib.pyplot as plt

        N = 20
        theta = np.linspace(0.0, 2 * np.pi, N, endpoint=False)
        radii = 10 * np.random.rand(N)
        width = np.pi / 4 * np.random.rand(N)
        ax = plt.subplot(111, projection='polar')
        bars = ax.bar(theta, radii, width=width, bottom=0.0)
        for r, bar in zip(radii, bars):
            bar.set_facecolor(plt.cm.jet(r / 10.))
            bar.set_alpha(0.5)
```

```
Python 3.5.2 |Continuum Analytics, Inc.| (default, Jul  2 2016, 17:52:12)
Type "copyright", "credits" or "license" for more information.

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


In [1]: %matplotlib inline
        from numpy.random import beta
        import matplotlib.pyplot as plt
        plt.style.use('bmh')

        def plot_beta_hist(a, b):
            plt.hist(beta(a, b, size=10000), histtype="stepfilled",
                     bins=25, alpha=0.8, normed=True)
            return

        plot_beta_hist(10, 10)
        plot_beta_hist(4, 12)
        plot_beta_hist(50, 12)
        plot_beta_hist(6, 55)
```

```
In [ ]:
```

```
  1  [|||||||||||||                                     18.1%]     Tasks: 305 total, 1 running
  2  [||||                                               5.0%]     Load average: 2.29 2.07 2.09
  3  [||||||||||                                        15.6%]     Uptime: 4 days, 21:59:11
  4  [|||||                                              5.0%]
  Mem[|||||||||||||||||||||||||||||||||||||||||||||5987/8192MB]
  Swp[||||||||||||||||||||||||||||||||||||||||||   2487/3072MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
82374 fperez     31   0 2389M  2048     0 R  0.0  0.0  0:00.00 htop
    1 root        0   0     0     0     0    0.0  0.0  0:00.00 (launchd)
   46 root        0   0     0     0     0    0.0  0.0  0:00.00 (syslogd)
   47 root        0   0     0     0     0    0.0  0.0  0:00.00 (UserEventAgent)
F1Help  F2Setup F3SearchF4InvertF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

A Collaborative effort:

Bloomberg

CONTINUUM
ANALYTICS

Brian, Jason, Steven, Darian,
Sylvain, Carol, Cameron,
Farica, Paul, Reese, Kyle,
Chris, Ian, Matthias, …

jupytercon

August 22–25, 2017
New York, NY
jupytercon.com

Presented by Project Jupyter, the NumFOCUS Foundation, and O'Reilly Media

jupyter

NUMFOCUS
OPEN CODE = BETTER SCIENCE

O'REILLY

Call for speakers is now open.

If you have an intriguing case study, deep technical knowledge to share
about Jupyter's internals,a useful extension, or a provocative idea
for new development, we'd love to hear from you.

Submit your proposal by March 7 | Registration opens April 2017

Live Demo!

Demo credits/thank you:
Brian Granger (Cal Poly SLO)
Jason Grout (Bloomberg)