IP[y]:
IPython


# From Colombia to Jupyter:

## an odd path through physics, open source software and data science

Fernando Pérez

fernando.perez@berkeley.edu

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE


University of California, Berkeley
DEPARTMENT OF STATISTICS

BERKELEY LAB

# A bit about me…


WA
MT
ND
MN
OR
ID
SD
WI
MI
WY
NY
NE
IA
IL
PA
NV
United States
UT
CO
KS
MO
IN
OH
MD
NJ
DE
CA
WV
KY
VA
OK
AR
TN
NC
AZ
NM
MS
AL
SC
TX
GA
LA
FL
Gulf of Mexico
Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea
Nicaragua
Venezuela
Colombia
Ecuador
NB
PE
NS
ME
VT
NH
MA
CT


United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea
Nicaragua
Venezuela
Colombia
Ecuador


Aruba
Curaçao
Punto Fijo
Coro
Barranquilla
Maracaibo
Cartagena
Valledupar
Barquisimeto
Valencia
Sincelejo
Montería
Merida
Barinas
Apartadó
Cúcuta
San Cristobal
Bucaramanga
Medellín
Tunja
Pereira
Bogotá
Ibagué
Villavicencio
El Tuparro National Park
Cali
Colombia
Neiva
Popayán
Pasto
Ibarra
Quito
Ecuador
Parque Nacional Natural Cahuinari
Parque Nacional Natural Río Puré
Iquitos
Leticia

# Medellín, Colombia



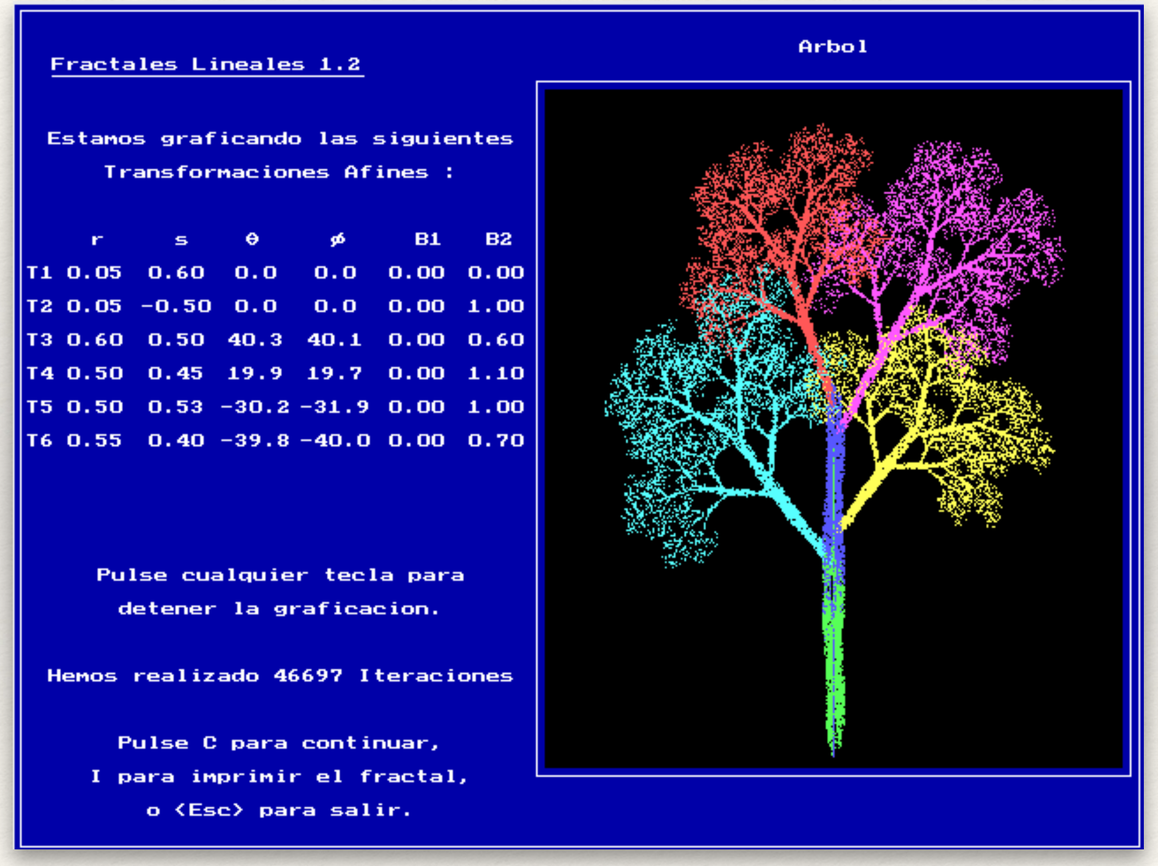
Image Credits:

latinorebels.com

totalspanishcolombia.com

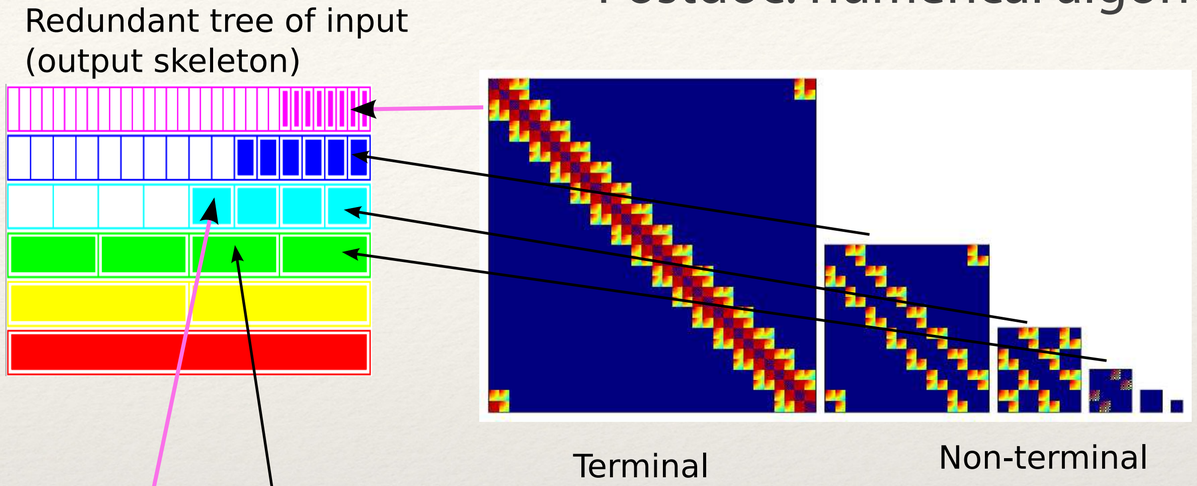
# My interest at the time: physics & computing

- Simulating fractals in TurboPascal
- Program on paper, use mom's office PC on weekends
- Debug on paper. *Think a lot away from the screen*

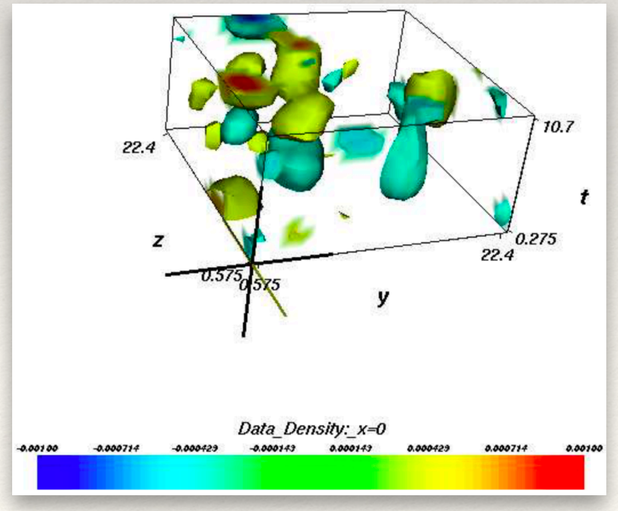


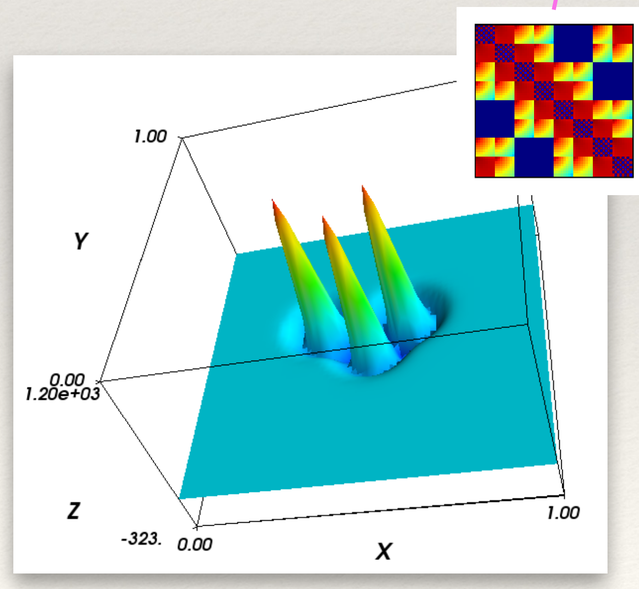
# Physics and applied math at CU Boulder

Postdoc: numerical algorithms




PhD: Lattice QCD Simulations




# Why do I do what I do?

# Why?

- **Ethical:** openness as fairness
- **Human/social:** openness fosters collaboration.
- **Epistemological:** proprietary science is an oxymoron.
- **Technical:** Python was cool :)

# Personal: crisis, motivation and support

- A **PhD in crisis**
- **Support** from
  - An incredible **(second) advisor** - Anna Hasenfratz
  - My **wife**!!
  - A **path forward** from bad PhD to great  Postdoc - Gregory Beylkin.

# What?

“The purpose of computing is insight, not numbers”

*–Hamming’62*

# IPython: Interactive Python, 2001

A humble start:

IPython 0.0.1, 259 LOC

“Just an afternoon hack”

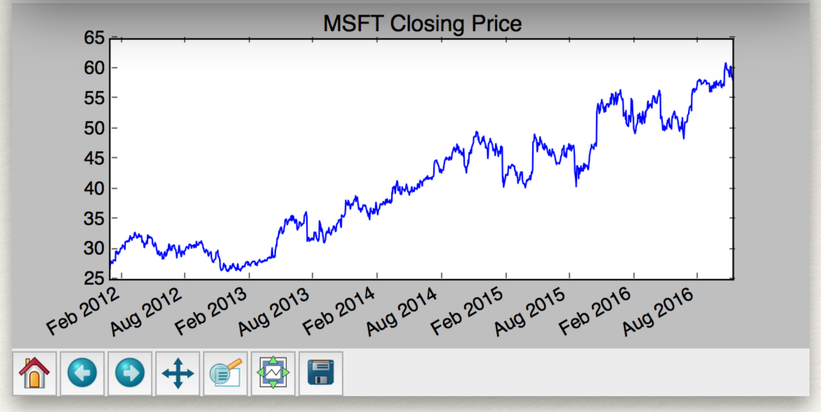
ipython-0.0.1.py

```
Globals for SI units (including g=9.8)     : _load_units   = %(_load_units)s
Starting number for prompt counter         : _prompt_ini   = %(_prompt_ini)s
Number of history items to store in cache  : _cache_size   = %(_cache_size)s
"""

#*****************************************************************************
# Configure here
_load_Numeric    = 1
_load_Gnuplot    = 1
_load_gracePlot  = 1
_load_units      = 1
_cache_size      = 1000
_prompt_ini      = 1

# ***  Don't modify below unless you know what you're doing.  ***

# Crude first version, with minimal object structure. This could be done much
# better, by defining a Cache class (probably using weak references or
# generators). But it seems to work ok. Haven't checked for memory circularity
# problems, though.

#*****************************************************************************
#       Copyright (C) 2001 Fernando P@rez. <fperez@pizero.colorado.edu>
#
#   Distributed under the terms of the GNU General Public License.
#
#   The full text of the GPL is available at:
#
#                  http://www.gnu.org/copyleft/gpl.html
#*****************************************************************************
__author__ = 'Fernando P@rez. <fperez@pizero.colorado.edu>'
__version__ = '0.1'

#*****************************************************************************
# Class definitions

class _HistPrompt1:
    """Simple interactive prompt like Mathematica's."""
    def __str__(self):
        return '\nIn['+`_prompt_count`+']:= '

class _HistPrompt2:
    """Simple interactive continuation prompt."""
    def __str__(self):
        return ' ... '+' '*(len('In['+`_prompt_count`+']:= ')-3)

#*****************************************************************************
# Function definitions

def _history_print(arg):
    """Printing with history cache management.

    This is invoked everytime the interpreter needs to print, and is activated
    by setting the variable sys.displayhook to it."""

    global _p,_pp,_ppp,_cache,_prompt_count
```

1. IPython: Users/fperez (python3.5)

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

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

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

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

In [3]:
```

Figure 1



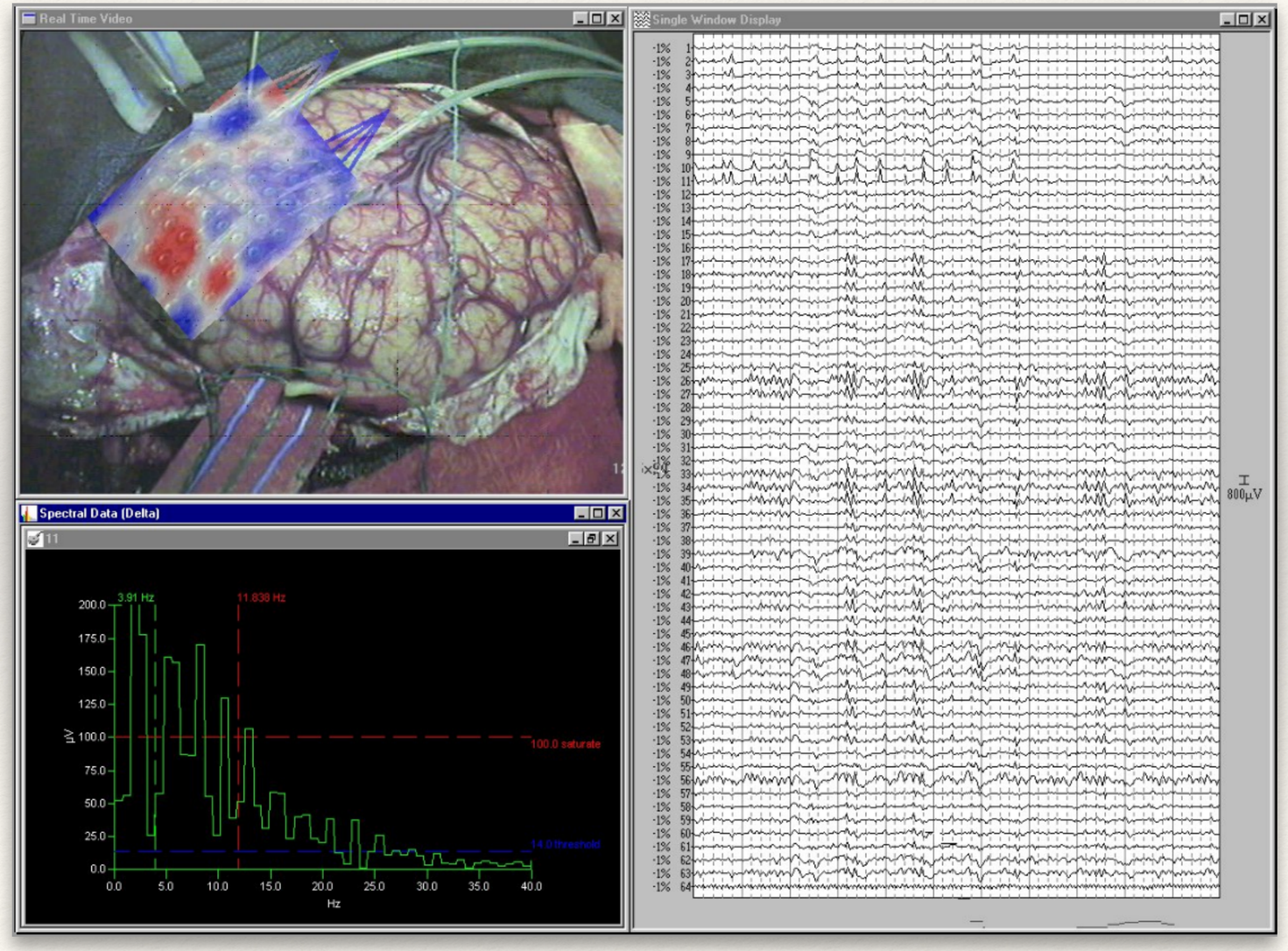

https://gist.github.com/fperez/1579699

# First outcome: I was good for *something*

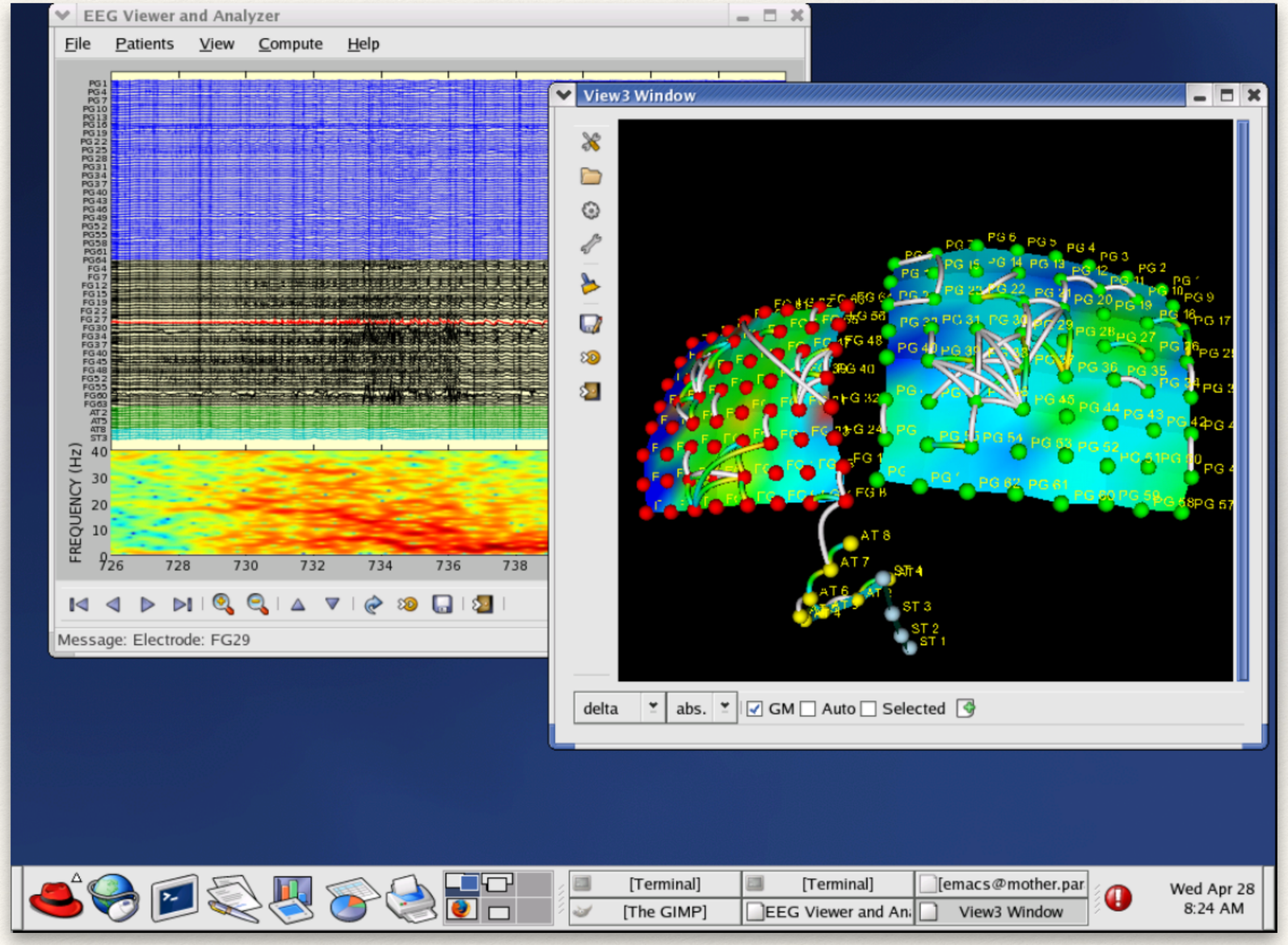
# Second outcome: finding a *community*

# Built by regular individuals

John Hunter, Department of Pediatric Neurology, University of Chicago.



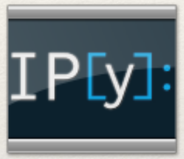
# matplotlib: open replacement for proprietary tools



# John D. Hunter, 1968-2012

# Not just IPython: an entire ecosystem






IPython

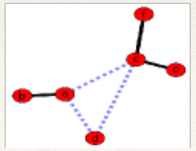


SymPy



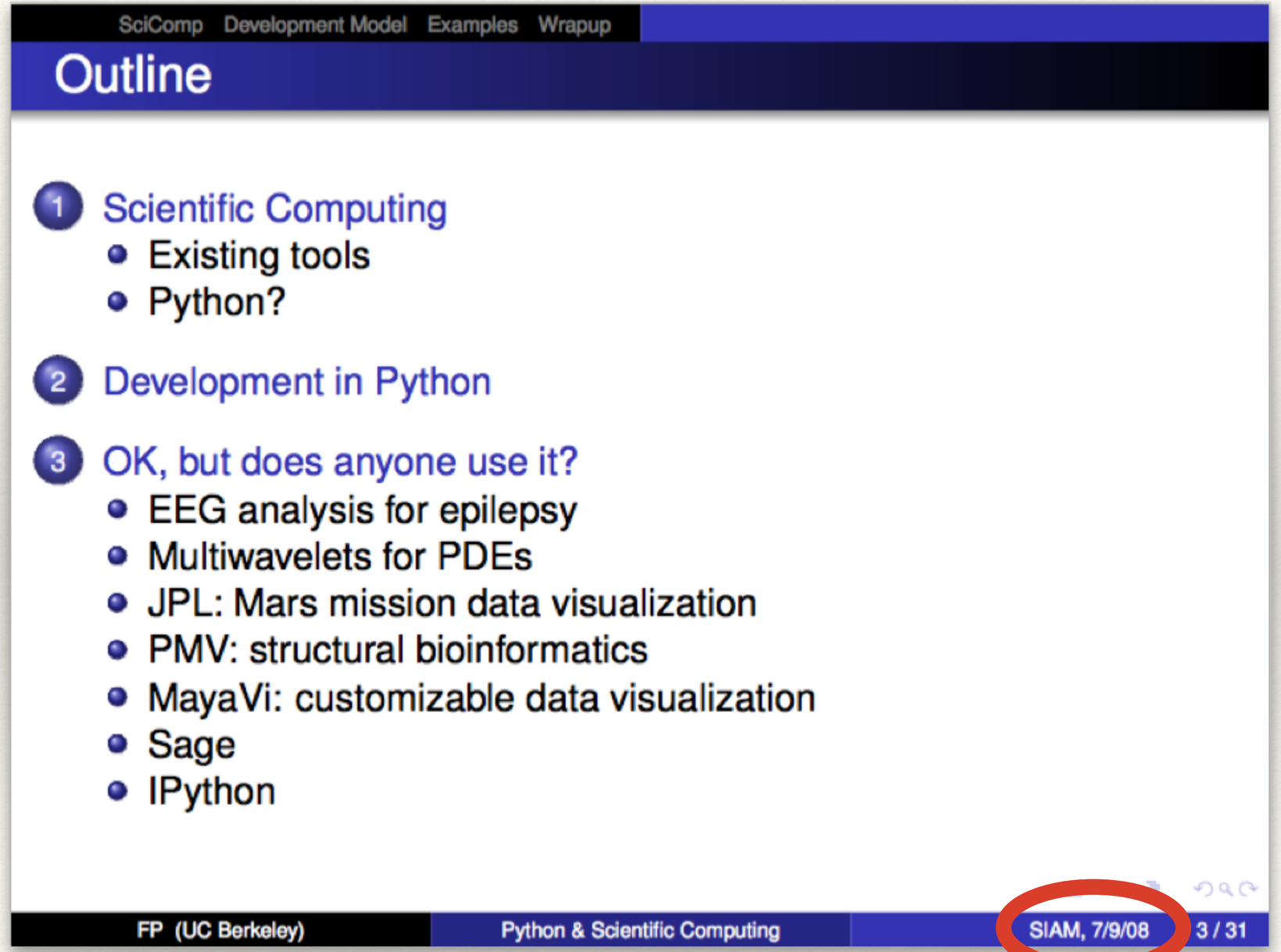




NetworkX

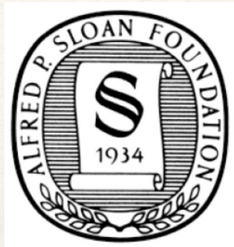
# Having to justify our existence



# Jupyter team today: where *all the credit* goes

Plus ~ 1500 more Open source contributors!

# Jupyter - funding and resources




ALFRED P. SLOAN FOUNDATION

GORDON AND BETTY MOORE FOUNDATION

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST


U.S. DEPARTMENT OF ENERGY


SIMONS FOUNDATION

NETFLIX

POWERED BY rackspace® the open cloud company

ANACONDA®

IBM®

Microsoft

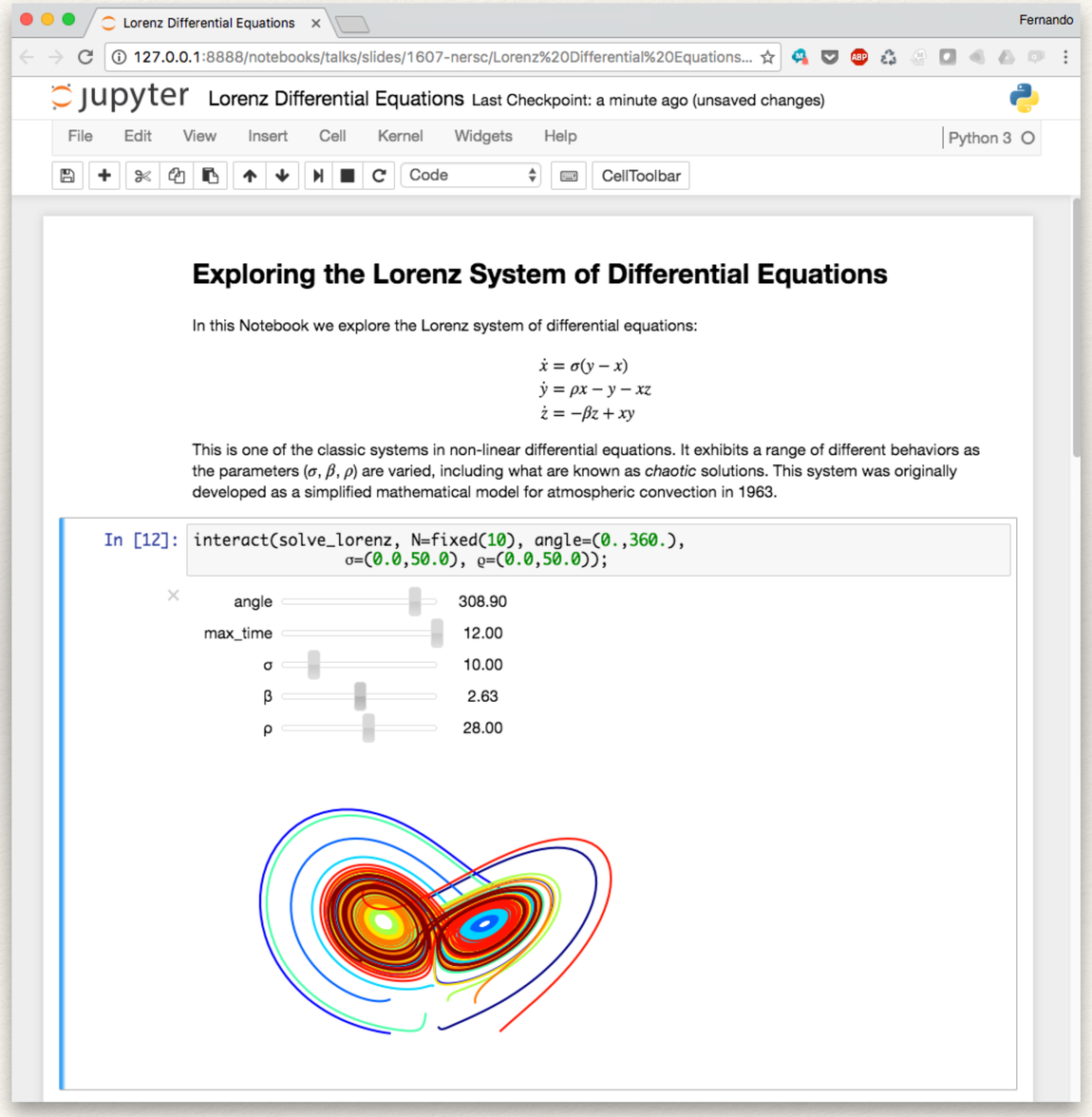
ENTHOUGHT SCIENTIFIC COMPUTING SOLUTIONS

Google

Bloomberg

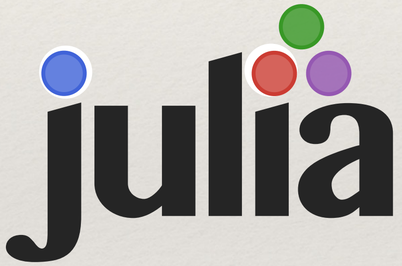
# The IPython/Jupyter Notebook

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



# Jupyter Protocol
# is language agnostic



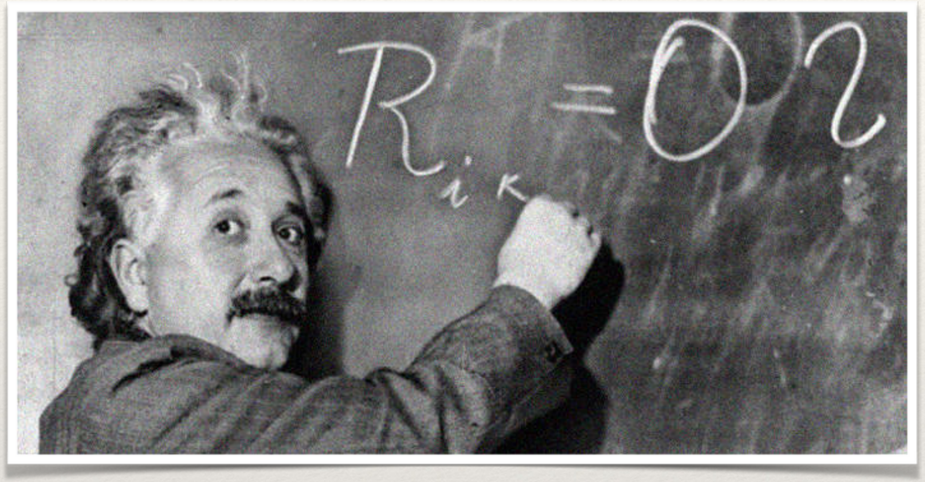



python™

IP[y]:
IPython

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

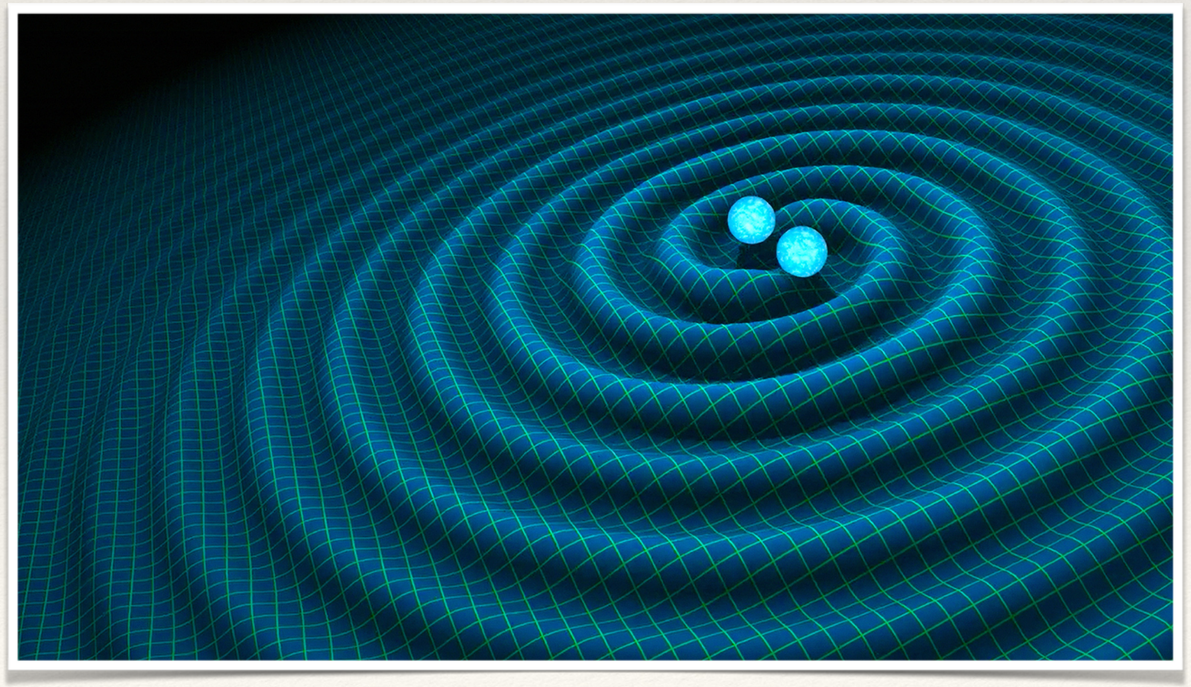
# A long time ago in a galaxy far, far away…



$$R_{\mu\nu} - \frac{1}{2} R\, g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Einstein's Field Equations of General Relativity

Annalen der Physik, 1916



# September 14, 2015

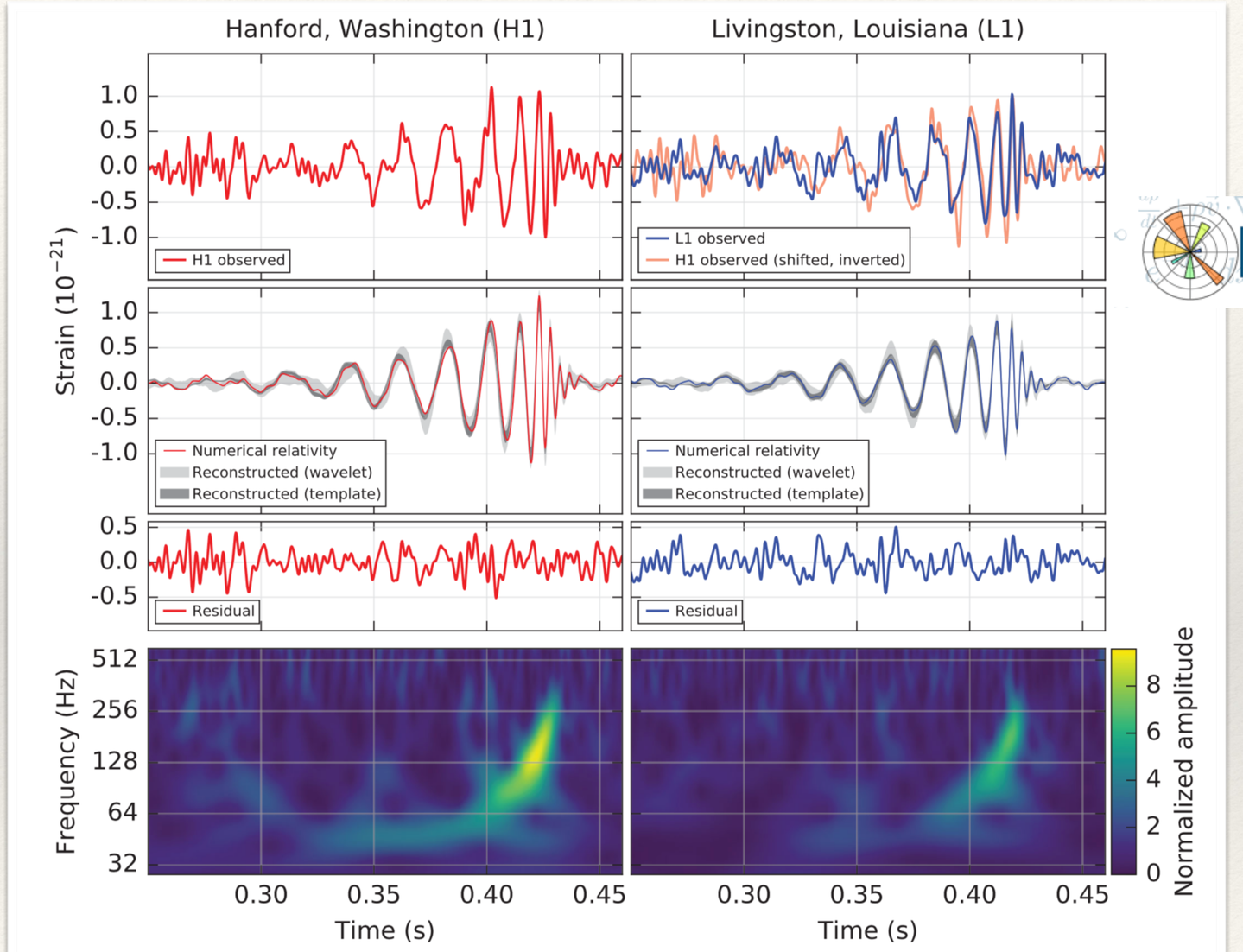



FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress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

NumPy

SciPy

matplotlib

**mpl colormaps**

**Update:** *These colormaps have been merged into the development version of Matplotlib, all of them will be included in matplotlib 1.5, and "option D" (now called "viridis") will be the new default colormap in matplotlib 2.0. If you just want the colormaps themselves, they're available in a single file here under a CC0 "no rights reserved" license. Third parties have also made viridis available in R and Matlab and JavaScript / D3. Below is th talk presented at SciPy2015 that outlines the whole story.*


# The song of the universe

jupyter GW150914_tutorial Last Checkpoint: 02/13/2016 (unsaved changes)

File Edit View Insert Cell Kernel Help IPython (Python 3)


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.

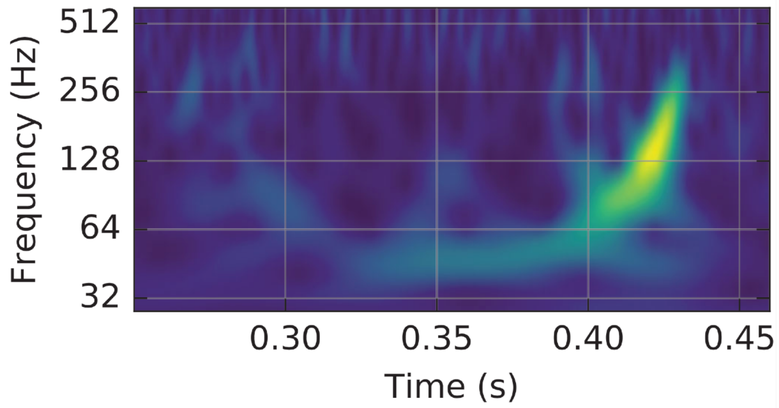
Now, as with whitening, the signal is visible as an oscillation sweeping from low to high frequency from -0.10 seconds to 0, then damping down into the random noise. Again, it looks roughly the same in both detectors, after shifting and flipping the L1 data with respect to H1. It's exactly the kind of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




### Make sound files

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

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

from scipy.io import wavfile

# function to keep the data within integer limits, and write to wavfile:
def write_wavfile(filename,fs,data):
    d = np.int16(data/np.max(np.abs(data)) * 32767 * 0.9)
    wavfile.write(filename,int(fs), d)

tevent = 1126259462.422          # Mon Sep 14 09:50:45 GMT 2015
deltat = 2.                      # seconds around the event

# index into the strain time series for this time interval:
indxt = np.where((time >= tevent-deltat) & (time < tevent+deltat))

# write the files:
write_wavfile("GW150914_H1_whitenbp.wav",int(fs), strain_H1_whitenbp[indxt])
write_wavfile("GW150914_L1_whitenbp.wav",int(fs), strain_L1_whitenbp[indxt])
write_wavfile("GW150914_NR_whitenbp.wav",int(fs), NR_H1_whitenbp)

for wav in glob('*whitenbp.wav'):
    display(wav)
    display(Audio(filename=wav))
```

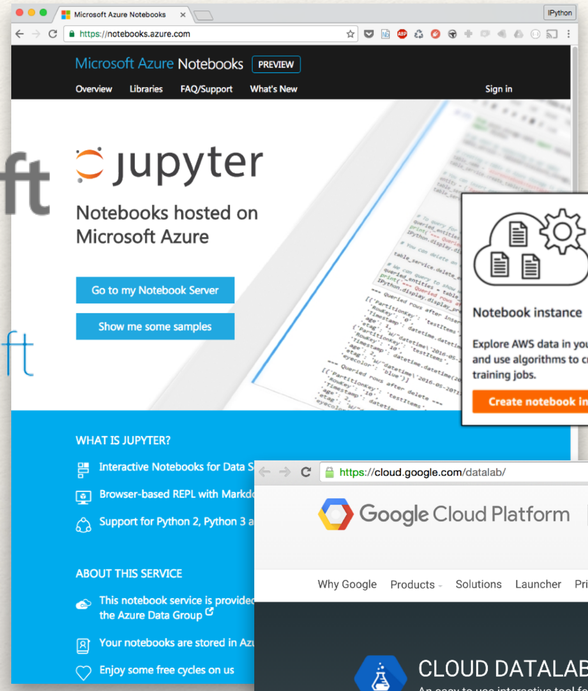
```
'GW150914_H1_whitenbp.wav'
```

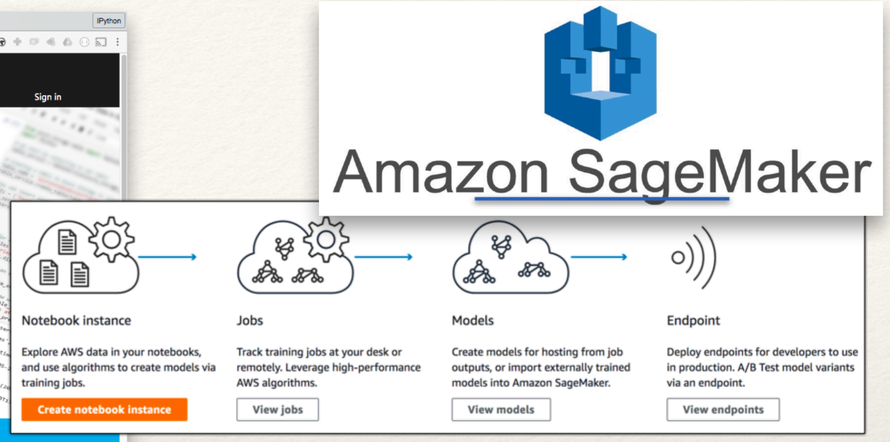

Using the IPython.display.Audio object

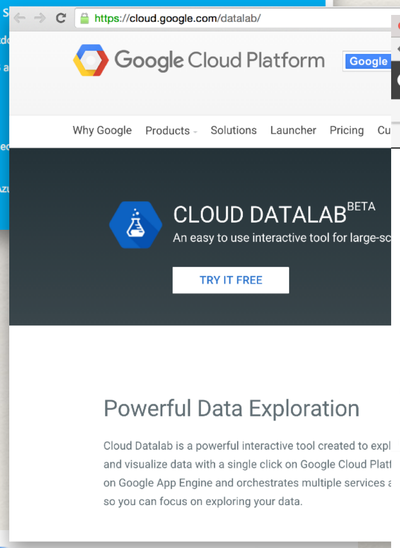
# Wide industrial adoption

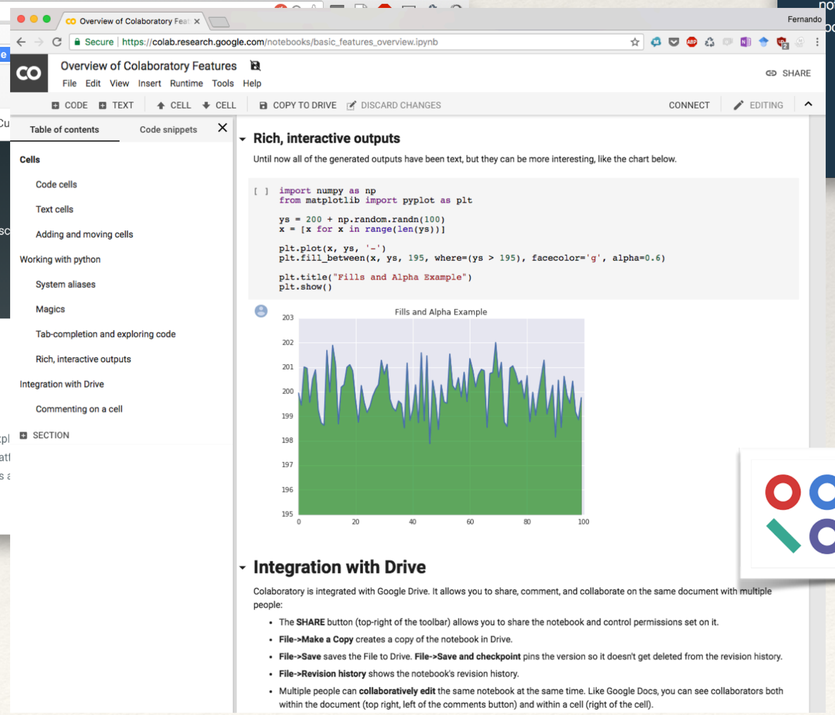
Microsoft

Microsoft Azure

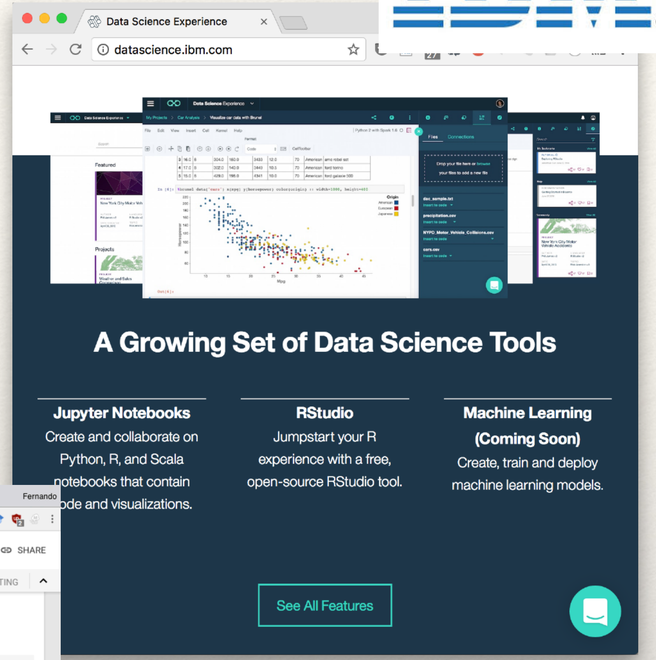



IBM




COCALC

GRYD

R-Brain

CODE OCEAN

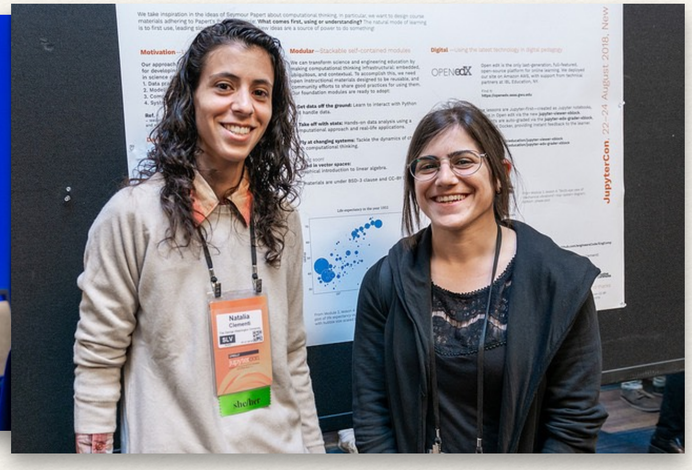
Google

ANACONDA

# jupytercon

**THE OFFICIAL JUPYTER CONFERENCE**
AUG 21-22, 2018: TRAINING
AUG 22-24, 2018: TUTORIALS & CONFERENCE
NEW YORK, NY

**Leverage the power of Jupyter for collaborative, extensible, scalable, and reproducible data science.**





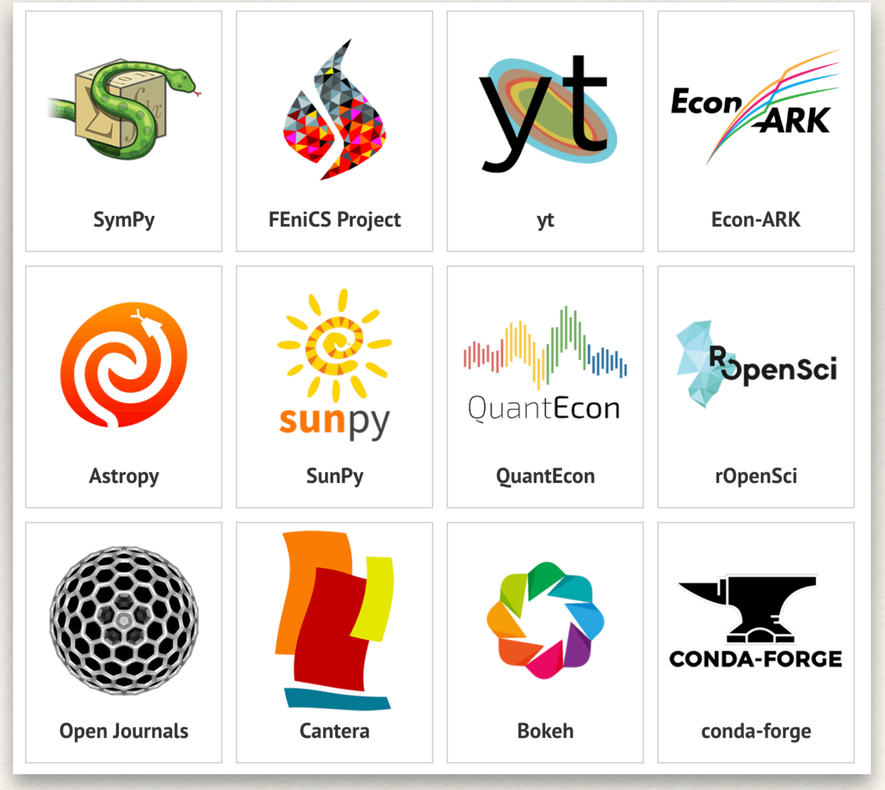


jupytercon.com @JupyterCon, photos by @triciaphoto

**If the world doesn’t give you a space, you’ll need to create it**

# 2012 NUMFOCUS

## OPEN CODE = BETTER SCIENCE


# NumFOCUS: beyond code, communities


## DISC Program Mission

NumFOCUS recognizes that the open source data science community is currently highly homogenous. We believe that diverse contributors and community members produce better science and better projects. NumFOCUS strives to help create a more diverse community through initiatives and programming devoted to increasing participation by and inclusion of underrepresented people.

Join the DISC Mailing List

## NumFOCUS Diversity Statement

NumFOCUS welcomes and encourages participation in our community by people of all backgrounds and identities. We are committed to promoting and sustaining a culture that values mutual respect, tolerance, and learning, and we work together as a community to help each other live out these values.

For a more detailed explication of NumFOCUS's position on diversity in the community, see the Diversity Appendix.


The John Hunter Matplotlib Summer Fellowship, named in memory of Matplotlib creator John Hunter, sponsors one to two students to work full-time for 3 months on Matplotlib during the summer (in the northern hemisphere), supervised and mentored by a senior contributor from the project. The fellowship is designed to help prepare recipients to become active contributors and core maintainers of Matplotlib.

Learn More About Matplotlib

Donate to Support the Fellowship

# 2013: Berkeley Institute for Data Science


GORDON AND BETTY MOORE FOUNDATION

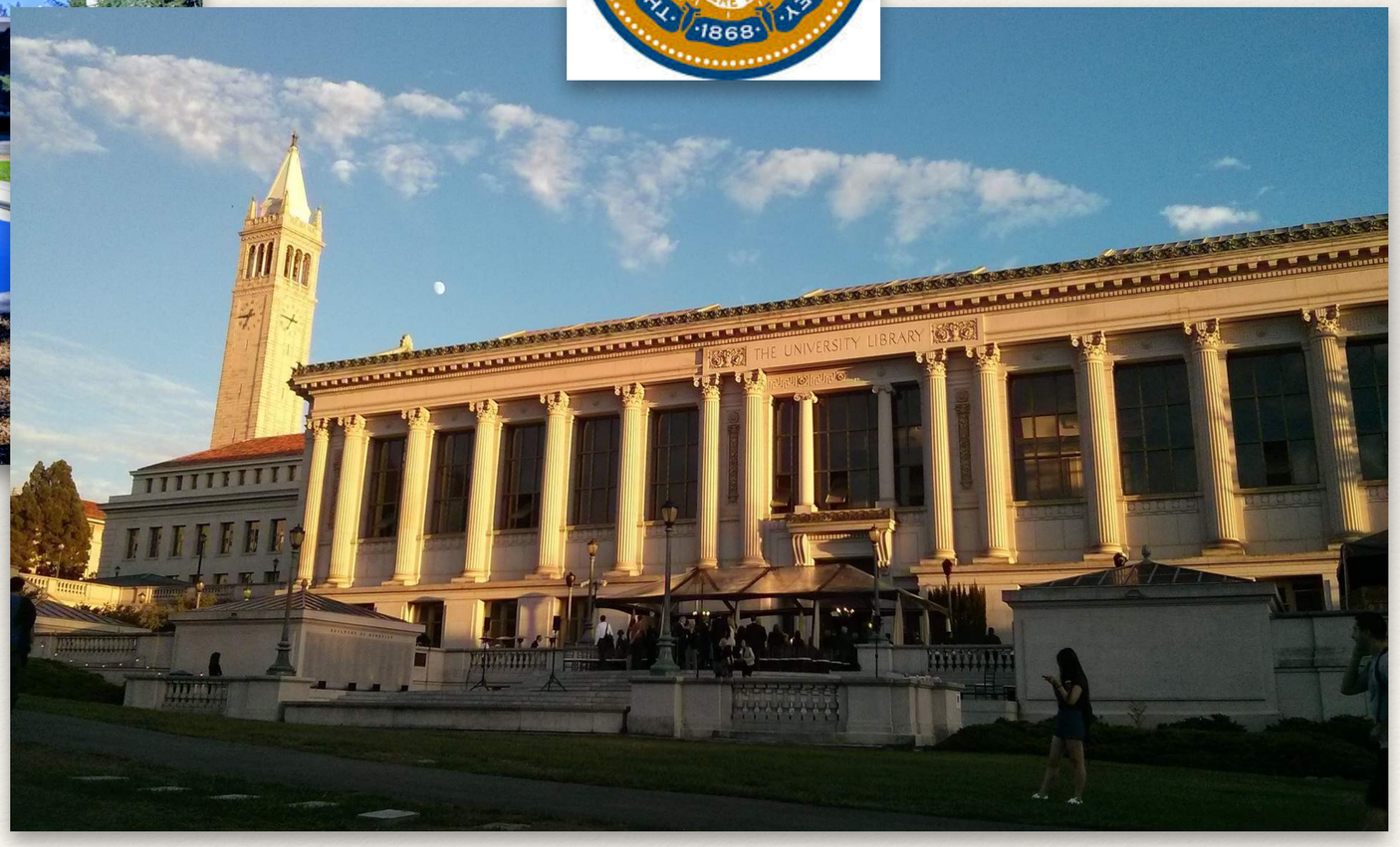
NYU — Berkeley UNIVERSITY OF CALIFORNIA — UNIVERSITY of WASHINGTON

ALFRED P. SLOAN FOUNDATION

# Creating good institutional spaces is hard, but critical!

University of California, Berkeley

# Department of Statistics



# 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***

# Collaborative and Reproducible Data Science

## STAT 159 @ Berkeley, Fall 2017

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


http://bit.ly/stat159-f17

# Student feedback

Anyway, I would like to meet with you in the coming weeks to update you about the progress I've made in my jump into reproducibility, especially my experience with contributing to pandas and the few chapters of "The Practice of Reproducible Research" I got to read.

**New open source contributor**

assistance. I was mainly interested in having you as an advisor because I'm interested in the idea of responsible research practices in this type of setting where the data cannot be shared - what do responsible research practices look like for analysis like this? How do I present the results in a way that shows all the steps taken and all the analyses run without giving too much information about the data?

**Undergraduate research project**

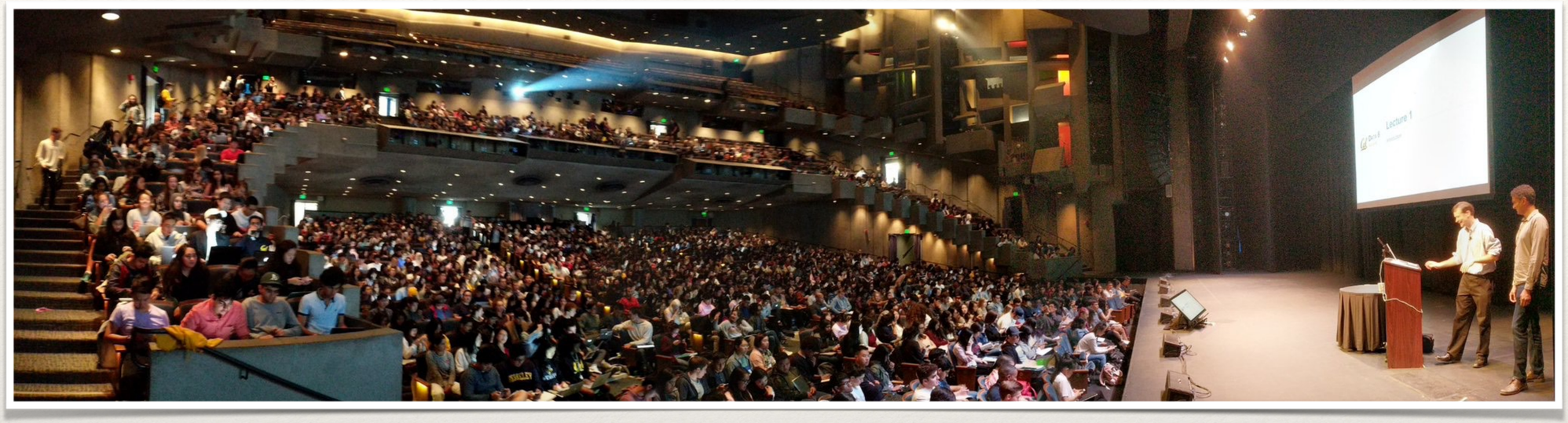
**Journalist who now is applying to Data Science graduate programs**

**(admitted to Columbia, JH, …)**

Your class still exert a great influence on my current projects. I've been working on create detailed buyer personas since I came back to China and using the method you taught in class to develop pricing and operating algorithm with Python, establishing a price estimation model and optimizing the valuation system of Airbnb with modified AeroSolve Module.

To be honest, I was hesitating before whether I could do a good job in data analysis given that I originally majored in journalism. Thanks to your encourage, now I feel more confident and develop a clear career

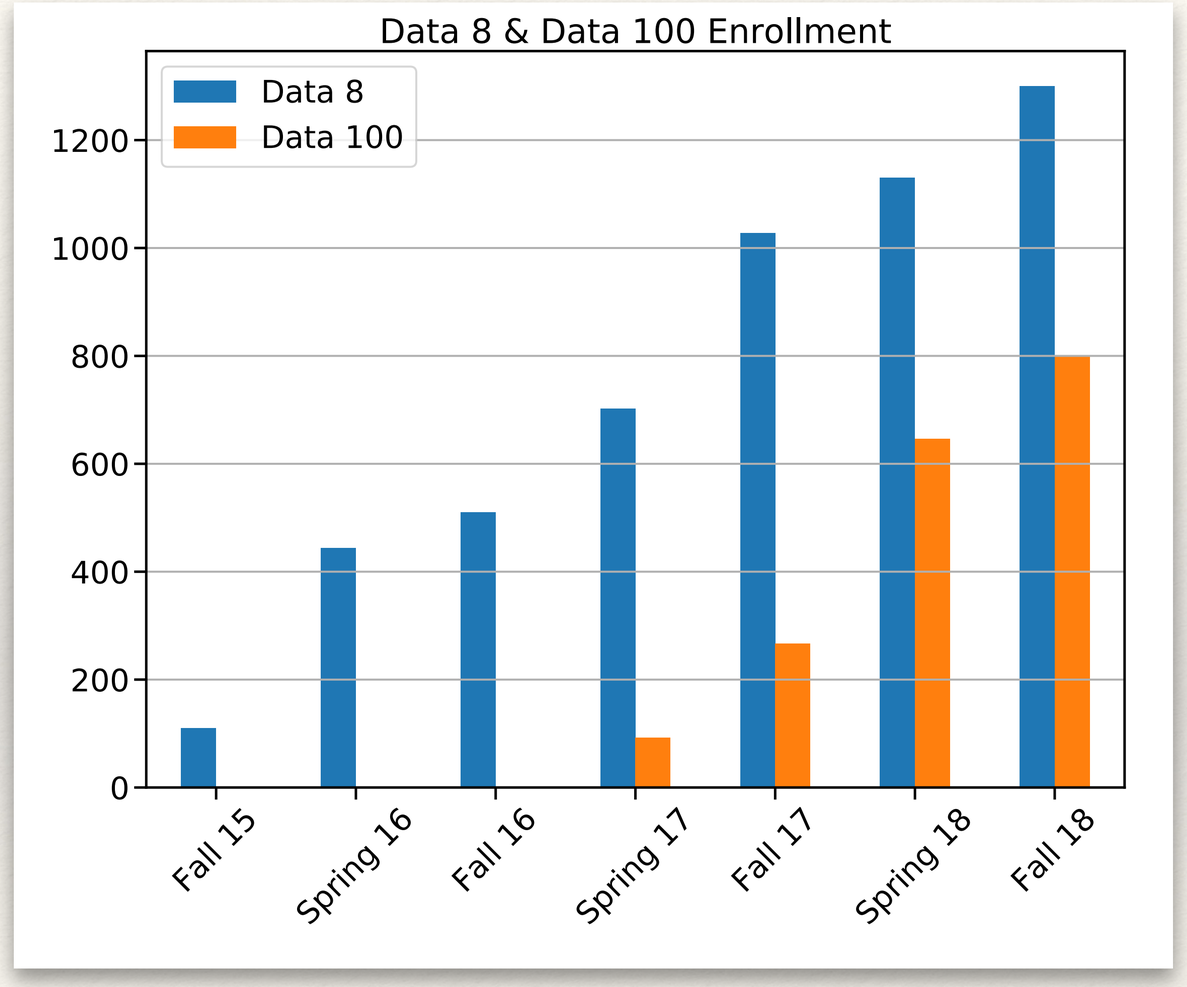
# Data 8 & Data100: massive uptake




D8: ~1,300 students

D100: ~800 students

# Fastest growing courses in Berkeley history




**Data 8 in Fall 2018**

- ~ 1,300 enrolled students
- ~ 200 waitlisted

**Annual combined numbers**

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

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

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

# Last two points: representation…

# Fair participation of all, across

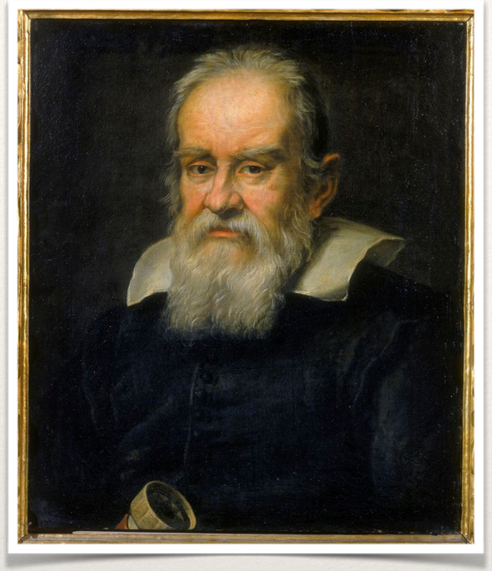
- Gender
- Ethnic
- Religious
- National
- Economic
- … boundaries, with support, opportunity and respect.

# Representation does matter!


Williams [Venus/Serena's father] had created a plan to turn his daughters into champions

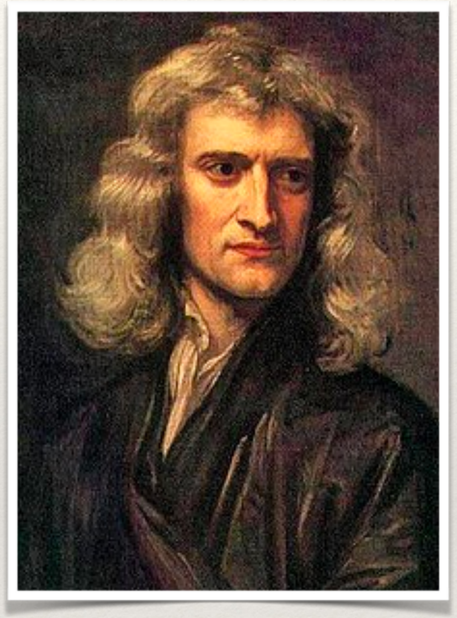
"The blueprint was already there," Francois [Naomi's father] told me. "I just had to follow it."

# But… a teenager in Colombia…



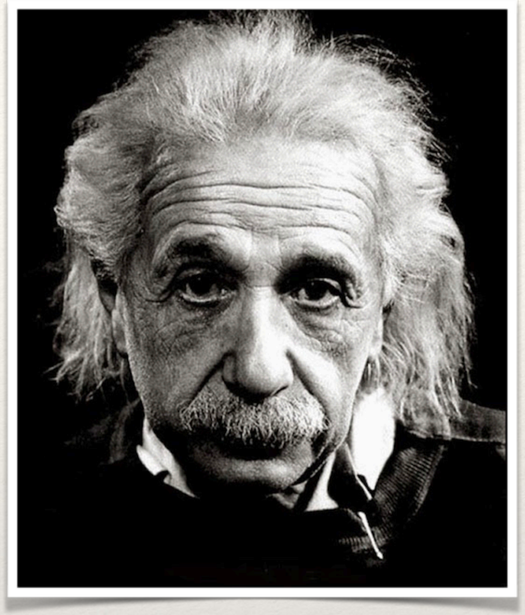
Galileo Galilei

1564-1642



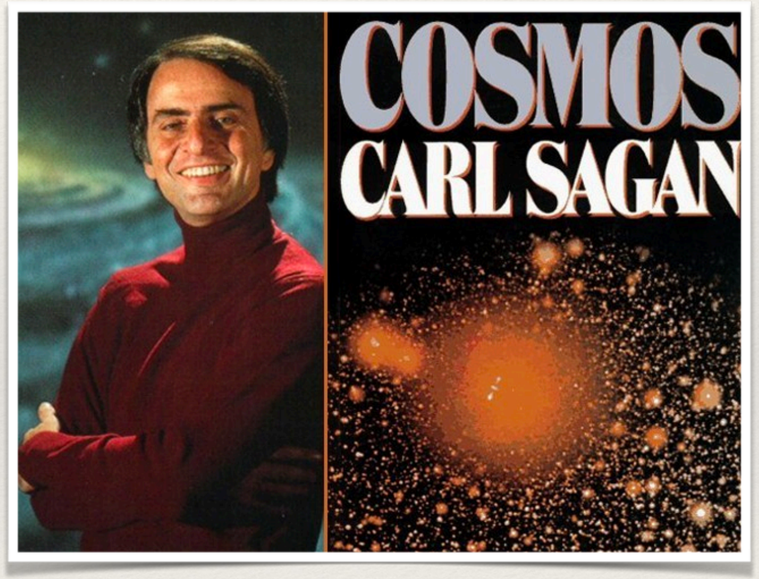
Isaac Newton

1643-1727



Albert Einstein

1879-1955



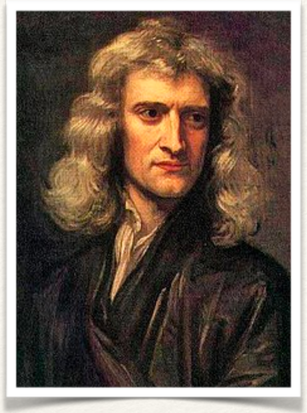

Carl Sagan

1934-1996

Johannes Kepler

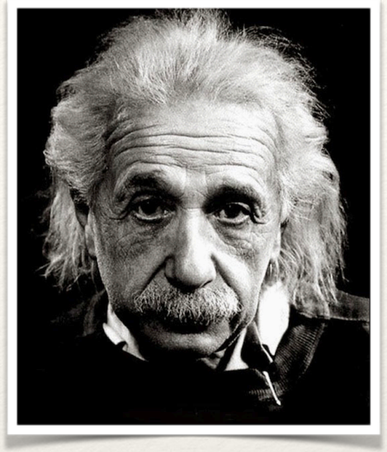
1571-1630

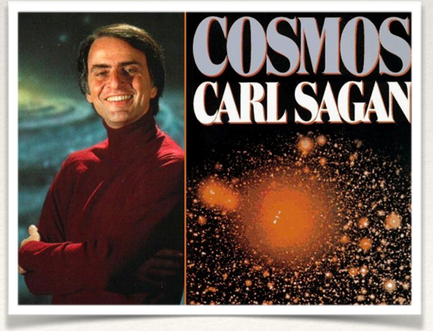
# humanity and ideas









# … and careers

# Choose your mentors carefully

- PhD Advisor: **one of the most important relationships** in your life.
- **Power dynamics** is stacked against you.
- **Personal qualities** of mentor have to compensate for that…
  - And in many cases they do! There are **amazing mentors** out there :)
- **Due diligence**: ask hard questions of former students, postdocs and the mentor.
- A good mentor **pushes you hard** to do your best work, but always treats you first as a human being who merits **respect**.
- In a bad relationship, **walk away**! The earlier the better.

# Beyond Academia?

## You can *choose* a different path!

- It does NOT mean you
  - are not smart/hard working enough,
  - are a sellout,
  - only care about $$$,
  - don’t care about the really hard/interesting problems,
  - wasted your time going to grad school,
  - are a failure as a person,
  - …

# Beyond Academia…

**Berkeley | SkyDeck**
@SkyDeck_Cal

Follow

#BPEP - connecting the #entrepreneurial #postdoc community at @UCBerkeley to #startup #founders who've been there! 😊🚀


7:45 PM - 20 Sep 2018



0

EDITORIAL

## Policy-Making Needs Science

Bruce Alberts

*Bruce Alberts is Editor-in-Chief of Science.*

– Hide authors and affiliations

0

*Science* 03 Dec 2010:
Vol. 330, Issue 6009, pp. 1287
DOI: 10.1126/science.1200613

## Science Communication & Careers Beyond the Bench

**Keynote Address**
*5:00pm*
Jamie Talan, MPH
*Atlantic Fellow at Global Brain Health Institute, UCSF Memory & Aging Center*

**Panel Discussion with Science Communication Professionals**
*6:00pm*

**Danielle Pasquel, PhD**
*Associate Scientific Director at Golin*

**Deb Aronson, PhD**
*VP, Medical Director at ghg*

**Lisa Brown, PhD**
*Medical Science Liason at Assurex Health*

**Elizabeth M. Vancza, PhD, DABT**
*Senior Toxicologist, Safebridge Consultants, Inc.*

**Travis J. Bernardo, PhD**
*Senior Medical Writer, BGB Group*

**Event Hosted By:**
Women's Networking Group, Albert Einstein College of Medicine
NYC Science Communication
NYCSciComm.Org
Career & Professional Development Program for Graduate Students and Postdocs
Graduate Division of Biomedical Sciences
Belfer Institute for Advanced Biomedical Studies
EINSTEIN
Albert Einstein College of Medicine

**Networking Reception with Panelists**
*7:00 - 8:00pm*

BA Beyond Academia
Connecting PhDs with the World

# Thank You!