From Colombia to Jupyter:
an odd path through physics, open source software and data science

Fernando Pérez
fernando.perez@berkeley.edu
A bit about me...
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

PhD: Lattice QCD Simulations

Postdoc: numerical algorithms

Redundant tree of input (output skeleton)

Terminal

Non-terminal
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
A humble start:
IPython 0.0.1, 259 LOC

“Just an afternoon hack”

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
Having to justify our existence
Jupyter team today: where all the credit goes

Plus ~ 1500 more Open source contributors!
The IPython/Jupyter Notebook

- Rich web client
- Text & math
- Code
- Results
- Share, reproduce.
Jupyter Protocol is language agnostic

~100 different kernels: https://github.com/jupyter/jupyter/wiki/Jupyter-kernels
A long time ago in a galaxy far, far away…

Einstein’s Field Equations of General Relativity
Annalen der Physik, 1916

\[ R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu} \]
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...
The song of the universe

Make sound files

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

```python
# 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
Leverage the power of Jupyter for collaborative, extensible, scalable, and reproducible data science.
If the world doesn’t give you a space, you’ll need to create it.
2012 NUMFOCUS
OPEN CODE = BETTER SCIENCE
DISC Program Mission

NumFOCUS recognizes that the open source data science community is currently highly homogeneous. 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.

John Hunter Matplotlib Summer Fellowship

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

Join us for the launch of the Berkeley Institute for Data Science
December 12, 2013, 11:00 - 3:00 pm
Banatao Auditorium, Sutardja Dai Hall
Creating good institutional spaces is hard, but critical!
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

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.

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?

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 & Data 100: 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!

Last two points: representation...
Fair participation of all, across

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

Johannes Kepler
1571-1630

Albert Einstein
1879-1955

Carl Sagan
1934-1996
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...

Beyond Academia: Connecting PhDs with the World

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 Liaison at Assurant Health
Elizabeth M. Vancza, PhD, DABT
Senior Toxicologist, Safebridge Consultants, Inc.
Travis J. Berns, PhD
Senior Medical Writer, B&G Group

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

Event Hosted By:

About Beyond Academia
Thank You!