Reproducible science with Jupyter

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

LBL & UC Berkeley
Every research discipline is now awash in data

Astronomy: LSST

Physics: LHC

Personalized, data-driven medicine

Sociology: The Web

Biology: Sequencing

Neuroscience: EEG, fMRI

Economics: POS terminals
IPython: Interactive Python, 2001

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

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

Exploring the Lorenz System of Differential Equations

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

\[
\begin{align*}
    \dot{x} &= \sigma(y - x) \\
    \dot{y} &= xz - y - \beta y \\
    \dot{z} &= -\gamma z + xy
\end{align*}
\]

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters \((\sigma, \beta, \gamma)\) are varied, including what are known as chaotic solutions. This system was originally developed as a simplified mathematical model for atmospheric convection in 1963.
Funding and partnerships
Core ideas of the web: HTTP & HTML

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

Document Format

We have already computed $P(X \mid A)$ 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

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

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

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

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
Jupyter Protocol
is language agnostic

~75 different kernels: https://github.com/ipython/ipython/wiki/IPython-kernels-for-other-languages
Notebook: a data structure
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”

Notebooks on Github: the “actual scholarhip”
mybinder.org

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

submit

github.com/freeman-lab

Andrew Osheroff’s SciPy’16 talk:
https://www.youtube.com/watch?v=OK6M4w7LYIc
Gravitational waves detected on Jupyter!

The filtered data peak at around $1.\times10^{-21}$, 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 damped 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.

```python
# make wav (sound) files from the whitened data, +/-2s around the event.
from glob import glob
from IPython.display import import display, Audio
from scipy.io import wavfile
```
LIGO: Open Science with Jupyter

Notebooks hosted on Microsoft Azure

LIGO Open Science Center

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 on Azure or mybinder

View: GW150914 | LV151012 | GW151226
Download: zip file with data | Jupyter 4 | Jupyter 3 | python script

Quickview Notebook

Make summary plots for any short segment of LIGO data.

Run on Azure or mybinder

Download: Jupyter 4

Signal Processing with GW150914

Explore signal processing with data from the first LIGO discovery.

View: HTML
Download: zip file with data
The future of reproducible science?
Global scientific output doubles every nine years

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
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

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
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”

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.
Some new developments in Jupyter’s orbit...
version control for notebooks?
nbdime to the rescue!

(notebook diff and merge: https://github.com/jupyter/nbdime)
JupyterLab: the notebook, evolved...
The “Notebook”?
JupyterLab: unifying these ideas

A Collaborative effort:

Bloomberg

CONTINUUM ANALYTICS

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