Computational Training & Data Literacy for Domain Scientists

Joshua Bloom
UC Berkeley, Astronomy

@profjsb

“Training Students to Extract Value from Big Data” National Academies of Science, DC 11 April 2014
What is the toolbox of the modern (data-driven) scientist?
What is the toolbox of the modern (data-driven) scientist?

And...How do we teach this with what little time the students have?
Astronomical Data Deluge
Serious Challenge to Traditional Approaches & Toolkits
Astronomical Data Deluge
Serious Challenge to Traditional Approaches & Toolkits

Large Synoptic Survey Telescope (LSST) - 2020
- Light curves for 800M sources every 3 days
- $10^6$ supernovae/yr, $10^5$ eclipsing binaries
- 3.2 gigapixel camera, 20 TB/night

LOFAR & SKA
- 150 Gps (27 Tflops) → 20 Pps (~100 Pflops)

Gaia space astrometry mission - 2014
- 1 billion stars observed ~70 times over 5 years
- Will observe 20K supernovae

Many other astronomical surveys are already producing data:
SDSS, iPTF, CRTS, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES etc.,
Towards a Fully Automated Scientific Stack for Transients

- Current state-of-the-art stack
  - observing
  - scheduling
  - strategy
  - finding
  - reduction

- Published work
  - inference
  - classification
  - follow-up

- Automated (e.g. iPTF)
  - NSF/CDI
  - NSF/BIGDATA

- Not (yet) automated
Built & Deployed Real-time ML framework, discovering >10,000 events in > 10 TB of imaging → 50+ journal articles

Built Probabilistic Event classification catalogs with innovative active learning

Our ML framework found the Nearest Supernova in 3 Decades..

http://timedomain.org

Data-Centric Coursework, Bootcamps, Seminars, & Lecture Series

BDAS: Berkeley Data Analytics Stack [Spark, Shark, ...]

parallel programming bootcamp

...and entire degree programs

dataScience@berkeley

Master of Information and Data Science
The UC Berkeley School of Information invites you to learn more about the only professional data science degree delivered fully online. Answer the simple questions below to request more information.

Earn a Master of Information and Data Science—Online
Now you can earn a degree in data science from anywhere in the world. The UC Berkeley School of Information offers the only professional Master of Information and Data Science.
Data-Centric Coursework, Bootcamps, Seminars, & Lecture Series

BDAS: Berkeley Data Analytics Stack
[Spark, Shark,...]

parallel programming bootcamp

Taught by CS/Stats

Aimed at Engineers & Programmers Heading Toward Industry

data.science@berkeley
Python Bootcamps at Berkeley

2010: 85 campers

2012a: 135 campers
a modern superglue computing language for science

- high-level scripting language
- open source, huge & growing community in academia & industry
- Just in time compilation but also fast numerical computation
- Extensive interfaces to 3rd party frameworks

A reasonable lingua franca for scientists...
Python Bootcamps at Berkeley

2012b: 210 campers  
2013a: 253 campers
- 3 days of live/archive streamed lectures
- all open material in GitHub
- widely disseminated (e.g., @ NASA)
- funded (~$18k) by the Vice Chancellor for Research & NSF (BIGDATA)

http://pythonbootcamp.info

GSFC Python Boot Camp 2013

The GSFC Python Boot Camp is aimed to bring students from zero to analyzing data with Python in three days. As someone has said Python is becoming the de facto standard data analysis language in many areas of science. This course is a three day intensive workout which will involve lectures from and hands on-sessions with Goddard scientists. This will not be a computer science class but rather a how-do-I-use-it-in-my-research type of class. The material is aimed at researchers of all levels (including and especially Summer internal)

All of this looks familiar...

You’re probably thinking of the Berkely Python Boot Camp which we are using as the basis for this workshop and sometimes blatantly copying. In fact, we stole one of their lecturers! Maybe one day we can be considered the East Coast version of their awesome event.
Python Computing for Science

Undergraduate/Graduate Seminar Course at UC Berkeley (AY 250)

<table>
<thead>
<tr>
<th>Date</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 2023</td>
<td>Advanced Python Language Concepts, learning about Jupyter notebook</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep 10</td>
<td>Advanced Versioning, application debugging &amp; testing</td>
</tr>
<tr>
<td>Sep 17</td>
<td>(matplotlib) Advanced plotting and data visualization, may.</td>
</tr>
<tr>
<td>Sep 20</td>
<td>Berian/Brad</td>
</tr>
<tr>
<td>Oct 1</td>
<td>Interacting with the world (xml, rpc, urllib, sending and receiving</td>
</tr>
<tr>
<td></td>
<td>talking to computers (notebook)</td>
</tr>
<tr>
<td>Oct 18</td>
<td>Josh</td>
</tr>
<tr>
<td>Oct 19</td>
<td>Josh</td>
</tr>
<tr>
<td>Oct 22</td>
<td>GUI (Tkinter, GTK, Traits)</td>
</tr>
<tr>
<td></td>
<td>Joey</td>
</tr>
</tbody>
</table>

Part of the Designated Emphasis in Computation at Berkeley

- Part of the Designated Emphasis in Computation at Berkeley
- Interfacing to other languages
- Bayesian inference & MCMC
- Visualization
- Hardware control
- Machine learning
- Database interaction
- User interface & web frameworks
- Timeseries & numerical computing
- Interfacing to other languages
- Bayesian inference & MCMC
- Visualization
- Hardware control
- Machine learning
- Database interaction
- User interface & web frameworks
- Timeseries & numerical computing
"Parallel Image Reconstruction from Radio Interferometry Data"

"Graph Theory Analysis of Growing Graphs"

http://mb3152.github.io/Graph-Growth/

"Realtime Prediction of Activity Behavior from Smartphone"

"Bus Arrival Time Prediction in Spain"

- Psychology: 16%
- Astronomy: 8%
- Neuroscience: 16%
- Biostatistics: 8%
- Physics: 12%
- Chemical Engineering: 12%
- ISchool: 4%
- Earth and Planetary Sciences: 4%
- Industrial Engineering: 12%
- Mechanical Engineering: 8%

- Female: 36%
- Male: 64%
Prevalence of Earth-size planets orbiting Sun-like stars

Erik A. Petigura\textsuperscript{a,b,1}, Andrew W. Howard\textsuperscript{b}, and Geoffrey W. Marcy\textsuperscript{a}

\textsuperscript{a}Astronomy Department, University of California, Berkeley, CA 94720; and \textsuperscript{b}Institute for Astronomy, University of Hawaii at Manoa, Honolulu, HI 96822

Erik Petigura
Berkeley Astro
Grad Student

Bootcamp/
Seminar Alum

DOE/NERSC computation
“Are we alone in the universe? What makes up the missing mass of the universe? ... And maybe the biggest question of all: How in the wide world can you add $3 billion in market capitalization simply by adding .com to the end of a name?”

President William Jefferson Clinton
Science and Technology Policy Address
21 January 2000

“Add Data Science or Big Data to your course name to increase enrollment by tenfold.”

Joshua Bloom
Just Now
Where do Bootcamps & Seminars fit into traditional domain science curricula?

- formal coursework competes with research obligations for graduate students

Are they too vocational/practical for higher Ed?

Who should teach them & how do we credit them?
Undergraduate & Graduate Training Mission

Thinking *Data Literacy* before Thinking *Big Data Proficiency*

first this...

...then this.
Data analysis recipes: Fitting a model to data*

David W. Hogg
Center for Cosmology and Particle Physics, Department of Physics, N Max-Planck-Institut für Astronomie, Heidelberg

Jo Bovy
Center for Cosmology and Particle Physics, Department of Physics, N

Dustin Lang
Department of Computer Science, University of Toronto
Princeton University Observatory

arXiv:1008.4686v1
“Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible.” McNutt, 2014

http://www.sciencemag.org/content/343/6168/229.summary

- Git has emerged as the de facto versioning tool
- Berkeley Common Environment (BCE) Software Stack
- “Reproducible and Collaborative Statistical Data Science” (Statistics 157: P. Stark)
- Next up: Versioning (big) data?
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 (\(\sigma\), \(\beta\), \(\rho\)) are varied.

Imports

First, we import the needed things from IPython, NumPy, Matplotlib and SciPy.

In [ ]: %matplotlib inline

In [ ]: from IPython.html.widgets import interact, interactive
from IPython.display import clear_output, display, HTML

In [ ]: import numpy as np
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 and exhibits a range of different behaviors as the parameters ($\sigma$, $\beta$, $\rho$) are varied.

Imports

First, we import the needed things from IPython, NumPy, Matplotlib and SciPy.

```python
In [ ]: %matplotlib inline

In [ ]: from IPython.html.widgets import interact, interactive

In [ ]: from IPython.display import clear_output, display, HTML

In [ ]: import numpy as np
```
IPython notebook is \textasciitilde\textit{agnostic} to the backend.
“novelty^2 problem”
Extra Burden for Forefront Scientists

Established CS/Stats/Math in Service of novelty in domain science

vs.

Novelty in domain science driving & informing novelty in CS/Stats/Math

https://medium.com/tech-talk/dd88857f662
Berkeley Institute for Data Sciences (BIDS)

- Physical Space & New Entity dedicated to the Moore/Sloan Data Science principles

- Goal: rich resource and ecosystem for domain scientists to connect & collaborate with methodologists

“Bold new partnership launches to harness potential of data scientists and big data”
Berkeley Institute for Data Sciences
Towards an Inclusive Ecosystem
Expanding Participation Among Underrepresented Groups

- 2013 Python Seminar: 36% women
- 2013 AMP Camp: < 5% women at
- This Workshop: 2 women out of 22 speakers
Summary

- Domain Science increasingly dependent upon methodological competencies

- Higher-Ed Role of such training still TBD
  - formal courses competes for time

- Data Literacy before Big Data Proficiency

- Need to create inclusive, collaborative environments bridging domains & methodologies
Thank you.

@profjsb