Community Python Data Analysis and Reduction Tools for Large-Scale Spectroscopy

Erik Tollerud
@eteq

Astropy Coordination Committee Member
Data Analysis Tools Branch Project Scientist

(Also: lover of large-scale spectroscopy of local galaxies… talk to me about an Astro2020 White Paper w/ K. Gilbert about M31 & the outer LG!)
Python is now the dominant Language in Astro Research

<table>
<thead>
<tr>
<th>All Participants</th>
<th>Graduate Students</th>
<th>Postdocs</th>
<th>Faculty &amp; Scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td>python</td>
<td>67±2 %</td>
<td>80±5 %</td>
<td>53±4 %</td>
</tr>
<tr>
<td>shell</td>
<td>47±2 %</td>
<td>42±3 %</td>
<td>46±3 %</td>
</tr>
<tr>
<td>idl</td>
<td>44±2 %</td>
<td>41±3 %</td>
<td>44±3 %</td>
</tr>
<tr>
<td>c/c++</td>
<td>37±2 %</td>
<td>32±3 %</td>
<td>39±3 %</td>
</tr>
<tr>
<td>fortran</td>
<td>28±2 %</td>
<td>23±2 %</td>
<td>22±2 %</td>
</tr>
<tr>
<td>iraf</td>
<td>24±1 %</td>
<td>13±2 %</td>
<td>27±3 %</td>
</tr>
<tr>
<td>excel/spreadsheet</td>
<td>15±1 %</td>
<td>13±2 %</td>
<td>17±2 %</td>
</tr>
<tr>
<td>html/css</td>
<td>14±1 %</td>
<td>7±1 %</td>
<td>21±2 %</td>
</tr>
<tr>
<td>sql</td>
<td>12±1 %</td>
<td>7±1 %</td>
<td>16±2 %</td>
</tr>
<tr>
<td>sm/supermongo</td>
<td>10±1 %</td>
<td>9±2 %</td>
<td>16±2 %</td>
</tr>
<tr>
<td>matlab</td>
<td>8±1 %</td>
<td>7±1 %</td>
<td>7±1 %</td>
</tr>
<tr>
<td>perl</td>
<td>8±1 %</td>
<td>4±1 %</td>
<td>5±1 %</td>
</tr>
<tr>
<td>r</td>
<td>6±1 %</td>
<td>6±1 %</td>
<td>16±2 %</td>
</tr>
<tr>
<td>javascript</td>
<td>3±1 %</td>
<td>2±1 %</td>
<td>4±1 %</td>
</tr>
<tr>
<td>mathematica</td>
<td>3±0 %</td>
<td>3±1 %</td>
<td>3±1 %</td>
</tr>
<tr>
<td>java</td>
<td>2±0 %</td>
<td>1±1 %</td>
<td>3±1 %</td>
</tr>
<tr>
<td>gnuplot</td>
<td>1±0 %</td>
<td>1±0 %</td>
<td>1±0 %</td>
</tr>
<tr>
<td>julia</td>
<td>1±0 %</td>
<td>1±0 %</td>
<td>0±0 %</td>
</tr>
<tr>
<td>ruby</td>
<td>1±0 %</td>
<td>1±1 %</td>
<td>1±0 %</td>
</tr>
<tr>
<td>awk</td>
<td>1±0 %</td>
<td>1±1 %</td>
<td>1±1 %</td>
</tr>
</tbody>
</table>
Python (esp. in Science) Embraces an Open Ecosystem

Openness leads to sharing the load

Although: Open Source ≠ Open Development

(More on this later)
Astropy Mirrors this “Community Ecosystem”

The Astropy Project is a community effort to develop a common core package for Astronomy in Python and foster an ecosystem of interoperable astronomy packages.

This means both by and for the community, pooling our resources.

(Professional) Astronomers help write it with engineers’ guidance.

It should be useful for them as part of their day-to-day work.
Specutils

https://specutils.readthedocs.io

An Astropy-coordinated package with data structures and standard analysis functions for spectroscopy.

- Pythonic data structures for spectra
- MOS support specifically built-in!
- Analysis functions ->
  - Flux, Centroids, FWHM
  - Continuum fitting/subtraction
  - Spectral arithmetic, respecting units
  - Line modeling
Specreduce

https://github.com/astropy/specreduce

An in-planning (read: mostly not yet implemented) package for reducing OIR spectra (think IRAF twodspec/onedspec’s calib parts, or IDL spec2d)

Needs more contributors! Too many instruments have decided to “roll their own.” (sometimes for good reasons, sometimes not…)

This could be a great place for the MSE community to start…
JWST DATs are Leveraging the Above

Astronomy Python Tool Development at STScI

- Spectral visualization tools
- Image visualization tools
- Training & documentation
- gwcs: generalized wcs
- asdf: advanced data format
- Software distribution
- JWST data structures
- Models & fitting
- Photometry tools

JWST Tools

- Specutils
- io
- Modeling
- Astropy helpers
- WCS
- Tutorials
- Astroquery
- MAST
- NDData
- Coordination

IRAF Replacement

External astropy dev

Spectral visualization tools

External astropy dev

IRAF replacement

Specutils
io
modeling
astropy-helpers
wcs
tutorials
astroquery
MAST
nddata
coordination

STAK: Jupyter notebooks
Add to existing python lib
Build replacement pkg
IRAF switchers guide
No replacement

STScI Space Telescope Science Institute
Specviz

https://specviz.readthedocs.io
While I did it in advance for the above, to get these regions I had to do a lot of manual plotting and fiddling with view region sizes. Better would be to be able to get regions fit interactively!

Interactive version

How to use: Execute the next three cells (until you see a spectrum).

In the spectrum view, select the region for a spectral line to be identified. Then hit "Record line". This should also populate an estimate for a redshift. Select the next line and hit "jump", recording if desired. Continue until all lines are populated.

In [14]: `import` helper_library  # <- This is a 72 line file that you might write yourself for a custom workflow

In [17]: `line_finder = helper_library.SpectrumLineFinder(subspec, bright_lines)`

In [18]: Selection... replace add and xor remove

In [19]: `line_finder.show()`

In [20]: `regions = line_finder.get_regions()`

Estimate redshift

Lets display the results for all the lines in a table:

In [21]: `tab = table.Column()`
   `tab['name'] = table.Column(dtype='S10')`
   `tab['centroid'] = table.Column(dtype=float, unit=u.angstrom)`
This might not meet everyone’s needs… But there’s a solution!

**How can MSE (or you) be a part of this?**
Open Development: Github PRs ... for All!
Community Development: Code Review

GitHub

astropy/io/fits/fitstime.py

```python
257 +
258 +def _get_info_if_time_column(col, global_info):
259 +    """
260 +    Check if a column without corresponding time column keywords i
```

mhvk on Aug 16  Member
Can you make the docstring explicit that this is only to special-case a column with the name 'TIME' and has units of time?

bsipocz commented 22 days ago

@AustereCuriosity - Could you please rebase this and address the review comments?

AustereCuriosity added some commits on Aug 9

- Read time ...
- Test for geodetic locations and a little clean up
- Corrections to handle numpy unicode strings in Python3 and Addition o...
- Adding tests for GPS and LOCAL scales and for location warnings. Also...
- Changelog entry added
- Documentation for reading time
- Catch exception and warn user. Also change col.* to col.info.*
- Changes to documentation in order to incorporate the various aspects ...
Community Development: Consensus!

GitHub
Open Development

GitHub

Keeping this process for all code is what makes it a community project - a Do-o-cracy

The work that gets done is the work that someone does…

~400 people have done that work. and ~1k / year now cite them for it
What is the Opportunity Today?

- MSE and its community has a chance now to set the tone re: Python community software. Options include:
  1. Build on existing tools (astropy, specutils, specviz, mosviz), where they are useful - don’t reinvent the wheel.
  2. Follow the open development model for MSE DM or Data Analysis/“Science” code.
  3. Do both: join forces with the Astropy community (and institutions like STScI) to build what is needed together. Astronomy as a whole may may thank us.
Backup Slides
This makes Astropy Valued by the community.

(Also: most popular in “astronomy”)

---

Citations/Publication Year for 2013A&A...558A..13A

- Unreferred
- Refereed
- Total citations: 1598
- Total refereed: 1359

Citations/Publication Year for 2018A&A...156A...13A

- Unreferred
- Refereed
- Total citations: 187
- Total refereed: 107

---

science

REPOSITORIES 1,049

bulutyazilim / awesome-datascience
- An awesome Data Science repository to learn and apply for real world problems.
- Updated 3 hours ago
- ⭐ 9k

sympy / sympy
- A computer algebra system written in pure Python
- Updated 13 minutes ago
- 2 issues need help
- ⭐ 5.3k

dformoso / machine-learning-mindmap
- A mindmap summarising Machine Learning concepts, from Data Analysis to Deep Learning.
- Updated on Oct 8
- ⭐ 3.7k

astropy / astropy
- Repository for the Astropy core package
- Updated 2 hours ago
- 32 issues need help
- ⭐ 1.9k
Open Development Allows Diffuse Collaboration

![Graph showing increasing number of code contributors over time]