Welcome to the 2022 M&M Short Course X-16

Data Analysis in Materials Science with



Presented by (a selection of) the HyperSpy developers/power users:

- Carter Francis
- Joshua Taillon



NIST Disclaimer

Certain commercial equipment, instruments, materials, vendors, and software are identified in this talk for example purposes and to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.



Some logistics before we get started...

- Each session will be led by one instructor, but throughout the day, the other instructor will be around to answer questions. Feel free to raise your hand at any time if you need individual assistance
- You (hopefully) should have already downloaded the tutorial files and installed HyperSpy. *If not*: We have USB drives with the data and installers pre-loaded that we can copy to your laptop
- Rules from the M&M organizers:
 - Masks Required: Masks are required indoors in the Oregon Convention Center, in alignment
 M&M 2022 attendance guidelines. M&M attendees will be required to wear masks as follows:
 - Speakers will be permitted to remove their masks when actively giving their presentations, if they wish
 - Attendees will be permitted to remove their masks briefly while actively eating or drinking at official M&M 2022 functions



Helpful Resources:

- Agenda and course materials: https://pages.nist.gov/hyperspy_tutorial
 - Also contains helpful links to Binder, resources, installation help, etc.
- Repository of notebook and data files (to download manually):

https://github.com/usnistgov/hyperspy_tutorial



A note about Binder:



• Binder (https://mybinder.org) is a tool that allows you to execute Python code using your browser without having to install any software



A note about Binder:



- Binder (https://mybinder.org) is a tool that allows you to execute Python code using your browser without having to install any software
- You access a server owned by Binder via the web-browser, the code runs on their servers and displays back to you



A note about Binder:



- Binder (https://mybinder.org) is a tool that allows you to execute Python code using your browser without having to install any software
- You access a server owned by Binder via the web-browser, the code runs on their servers and displays back to you
- This makes it very easy to get familiar with Python/HyperSpy, but will not help you if you need to do your own analysis
 - Only pre-defined "templates" are available, and the server has limited resources
 - It may also be fairly slow depending on the connection speed between you and Binder's servers (conference Wifi is not known for its speed...)

Session 1: An Introduction to Python and HyperSpy:

The multi-dimensional data analysis toolbox

Josh Taillon

July 31, 2022





What is this "Python" that I've heard of?





- High-level, general purpose programming language with "batteries included"
 - *i.e.*, it comes with enough features to do most basic programming tasks



- High-level, general purpose programming language with "batteries included"
 - *i.e.*, it comes with enough features to do most basic programming tasks
- Simple enough for quick scripts; featured enough for complex projects



- High-level, general purpose programming language with "batteries included"
 - i.e., it comes with enough features to do most basic programming tasks
- Simple enough for quick scripts; featured enough for complex projects
- Used extensively on the web, in applications, and throughout science



- High-level, general purpose programming language with "batteries included"
 - *i.e.*, it comes with enough features to do most basic programming tasks
- Simple enough for quick scripts; featured enough for complex projects
- Used extensively on the web, in applications, and throughout science
- Syntax emphasizes readability and explicitness



Java:

```
public class HelloWorld {
   public static void main(String[] args) {
      System.out.println("Hello, World!");
   }
}
```



```
Java:
   public class HelloWorld {
     public static void main(String[] args) {
        System.out.println("Hello, World!");
C++:
   #include <iostream>
   using namespace std;
   int main() {
        cout << "Hello, World!" << endl;</pre>
        return 0;
```

MEASUREMENT LABORATORY

Python:

```
print("Hello, World!")
```



Python:

```
print("Hello, World!")
```

Matlab:

```
disp("Hello, World!")
```



Okay, but what does this have to do with science?

The "scientific Python" ecosystem provides almost any functionality you may need:





Linear algebra, optimization, machine learning...

• Numerical computing with NumPy (like what you would use Matlab for):





Linear algebra, optimization, machine learning...

• Numerical computing with NumPy (like what you would use Matlab for):



• Signal processing, numerical integration, optimization, etc. with $\underline{\mathtt{SciPy}}$:





Linear algebra, optimization, machine learning...

• Numerical computing with NumPy (like what you would use Matlab for):



• Signal processing, numerical integration, optimization, etc. with SciPy:



• Machine learning with scikit-learn (and others):





Visualization, statistics, computer algebra...

• Scientific visualization with matplotlib:



Visualization, statistics, computer algebra...

• Scientific visualization with matplotlib:



• Data frames/series and statistical analysis with pandas:





Visualization, statistics, computer algebra...

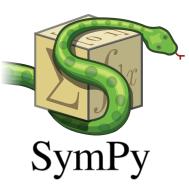
Scientific visualization with matplotlib:



• Data frames/series and statistical analysis with pandas:



• Symbolic computation with SymPy:





A better (?) way to create and publish your work

• Literate programming using a notebook interface with <u>Jupyter</u>: **Jupyter**:



A better (?) way to create and publish your work

• Literate programming using a notebook interface with <u>Jupyter</u>:



- Creation of interactive computational analysis documents (not just scripts):
 - Facilitates reproducibility
 - Makes analyses more accessible (anyone can recreate your figure when they want!)



A better (?) way to create and publish your work

• Literate programming using a notebook interface with <u>Jupyter</u>:



- Creation of interactive computational analysis documents (not just scripts):
 - Facilitates reproducibility
 - Makes analyses more accessible (anyone can recreate your figure when they want!)
- Easy to contribute back to the "open source" scientific ecosystem

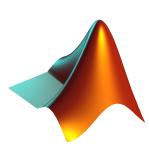


Can't I just use Matlab?



End To End development to execution	Slightly more code required than Matlab to do same procedures
Vast array of open source scientific packages	Quality of third-party packages varies quite a bit
Great for general programming and application development	Package management (dependencies) can be difficult (getting easier with tools like conda)
Can be a "glue" language to connect R, C++, and others	IDE is not "built-in" like Matlab
Free (as in speech and as in beer)!	By default, some operations are slower

VS.

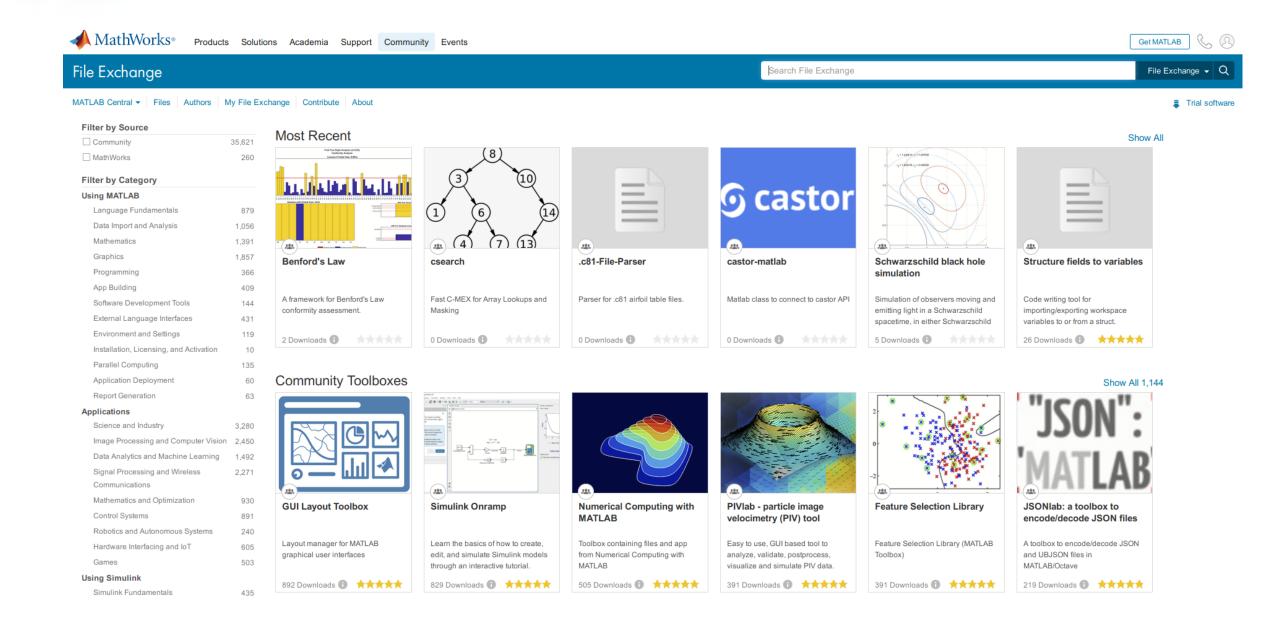


Scripts tend to be short due to high level of integration	Can not execute code; must be compiled into another language at runtime
Easy visualization and GUI development	Expensive! Must be licensed per-user (\$\$)
Well-tested and supported (commercial product)	Does not integrate well with other languages large projects are <i>extremely</i> difficult to manage
Built-in IDE and powerful debugger	Much functionality locked away in extra licensed packages (\$)
Multi-threading easier than in Python	Cannot develop stand-alone applications (must distribute Matlab runtime)

Adapted from: https://towardsdatascience.com/r-vs-python-vs-matlab-vs-octave-c28cd059aa69



File Exchange is good, but...



Distributing code is simple in the Python community

(and integrated directly into Python tools)

• GitHub: Community standard is to (at least) release code on GitHub (or similar service)



: Placing into the PyPI, the Python Package Index enables installing with simple:

- pip install my-awesome-package
- : Anaconda enables multiple environments and complex dependency management
 - conda install my-awesome-package



What does it mean to be "open source"?

...something people can modify and share because its design is publicly accessible

opensource.com



What does it mean to be "open source"?

...something people can modify and share because its design is publicly accessible

opensource.com

What about "open science"?

...science must be done in an open, and reproducible fashion where all components of research are open

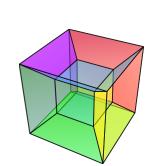
- Marcus Hanwell (also opensource.com)



Open-source and Python in the microscopy ecosystem:

General Purpose			Others	
HyperSpy	http://hyperspy.org/	AbTEM	https://abtem.readthedocs.io	
Nion Swift	https://nionswift.readthedocs.io	HRTEMFringe Analyzer	https://github.com/ialxn/HRTEMFringeAnalyzer	
pycroscopy	https://github.com/pycroscopy/pycroscopy	Atomap	https://atomap.org/	
DigitalMicrograph	https://www.gatan.com/run-python-script	LumiSpy	https://lumispy.readthedocs.io/en/latest/	
		Kikuchipy	https://kikuchipy.org	
	Pixelated STEM	Prismatic	https://prism-em.com/	
pyXem	https://pyxem.github.io/pyxem-website/	tomopy	https://tomopy.readthedocs.io/en/latest/	
py4DSTEM	https://py4dstem.readthedocs.io/	tomotools	https://github.com/usnistgov/tomotools	
LiberTEM	https://libertem.github.io/LiberTEM/	tomviz	https://tomviz.org/	
fpd	https://gitlab.com/fpdpy/fpd/			





What is HyperSpy, anyway?



What is HyperSpy multi-dimensional data analysis ?



• Open-source Python library for interactive data analysis of multi-dimensional datasets





- Open-source Python library for interactive data analysis of multi-dimensional datasets
- Leverages the Scientific Python ecosystem for much of its functionality:









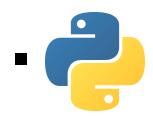








- Open-source Python library for interactive data analysis of multi-dimensional datasets
- Leverages the Scientific Python ecosystem for much of its functionality:















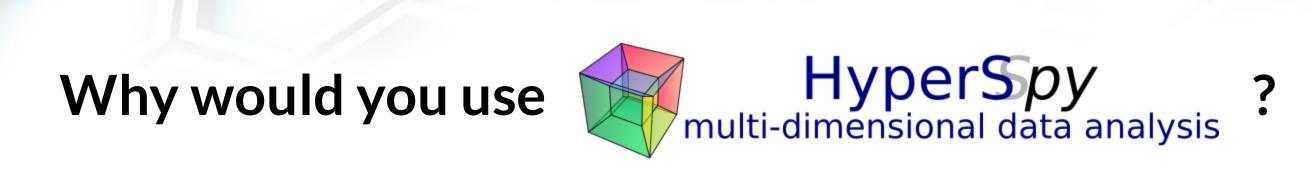
Accessed like any other Python library:

import numpy as np



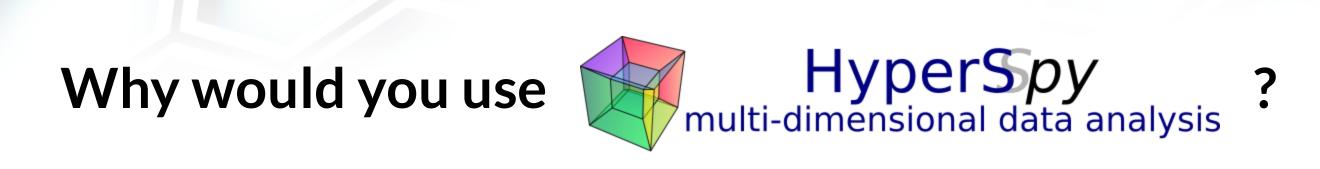
import hyperspy.api as hs





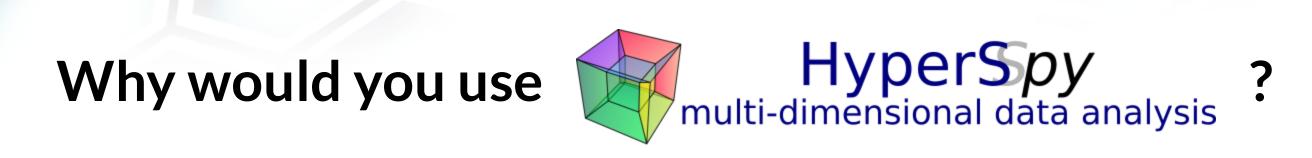
• Makes it easy to operate on multi-dimensional arrays as you would a single spectrum (or image)





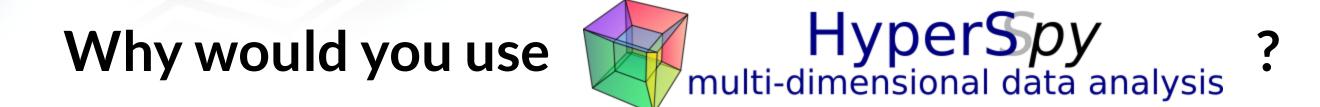
- Makes it easy to operate on multi-dimensional arrays as you would a single spectrum (or image)
- Easy access to cutting-edge signal processing tools





- Makes it easy to operate on multi-dimensional arrays as you would a single spectrum (or image)
- Easy access to cutting-edge signal processing tools
- Modular structure makes it easy to add custom features or extend into dedicated packages





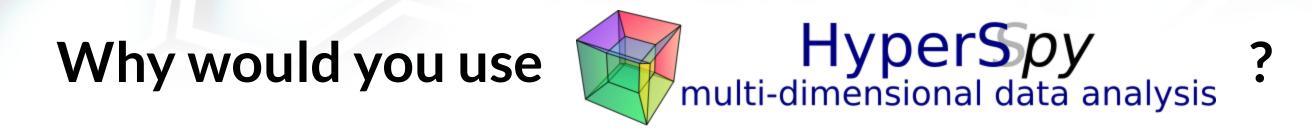
- Makes it easy to operate on multi-dimensional arrays as you would a single spectrum (or image)
- Easy access to cutting-edge signal processing tools
- Modular structure makes it easy to add custom features or extend into dedicated packages
- Use of Jupyter notebooks encourages reproducible and sharable analyses (FAIR data)





• Beyond generic signal processing, provides many tools specifically for electron microscopy:





- Beyond generic signal processing, provides many tools specifically for electron microscopy:
 - Provides facilities for easy access to proprietary software formats



Why would you use



- Beyond generic signal processing, provides many tools specifically for electron microscopy:
 - Provides facilities for easy access to proprietary software formats
 - EDS:
 - Background removal
 - Net intensity line map extraction
 - \circ Quantification (k-factor, ζ -factors, ionization cross sections)



Why would you use



- Beyond generic signal processing, provides many tools specifically for electron microscopy:
 - EELS:
 - Background removal
 - Curve fitting for quantification (including ELNES)
 - All of Egerton's famous methods



Why would you use



- Beyond generic signal processing, provides many tools specifically for electron microscopy:
 - EELS:
 - Background removal
 - Curve fitting for quantification (including ELNES)
 - All of Egerton's famous methods
 - "Advanced" methods:
 - Multivariate statistical analysis
 - General curve fitting
 - Dimensionality reduction/signal separation





• Quickly becoming the *de facto* standard of scientific computing





- Quickly becoming the *de facto* standard of scientific computing
- Free (as in speech and as in beer)
 - No pesky licenses to checkout





- Quickly becoming the *de facto* standard of scientific computing
- Free (as in speech and as in beer)
 - No pesky licenses to checkout
- Vast array of scientific libraries available:
 - pip install <just about anything>



Why python™?

- Quickly becoming the de facto standard of scientific computing
- Free (as in speech and as in beer)
 - No pesky licenses to checkout
- Vast array of scientific libraries available:
 - pip install <just about anything>
- Thanks to numpy and other libraries, similar (or often better) performance than MATLAB



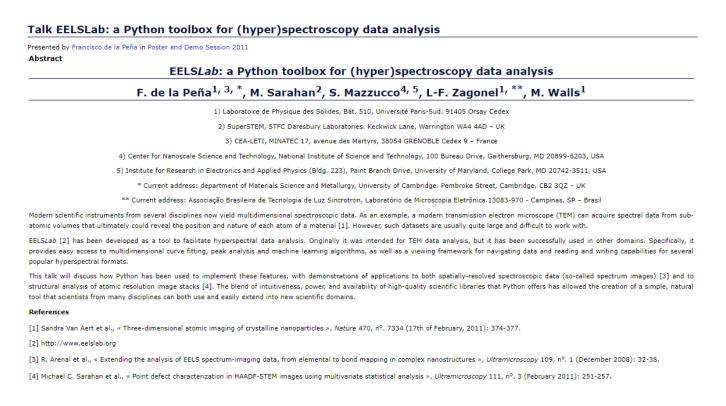
 Developed by <u>Francisco de la Peña</u> in 2007 – 2012 as part of his Ph.D. Thesis at Université Paris-Sud in Orsay, France



- Developed by <u>Francisco de la Peña</u> in 2007 2012 as part of his Ph.D. Thesis at Université Paris-Sud in Orsay, France
- See <u>Chapter 5</u> of the dissertation for a historical introduction

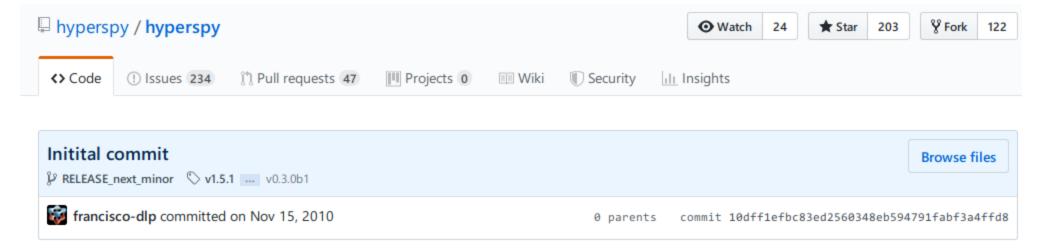


- Developed by <u>Francisco de la Peña</u> in 2007 2012 as part of his Ph.D. Thesis at Université Paris-Sud in Orsay, France
- See <u>Chapter 5</u> of the dissertation for a historical introduction
- Originally called EELSLab:



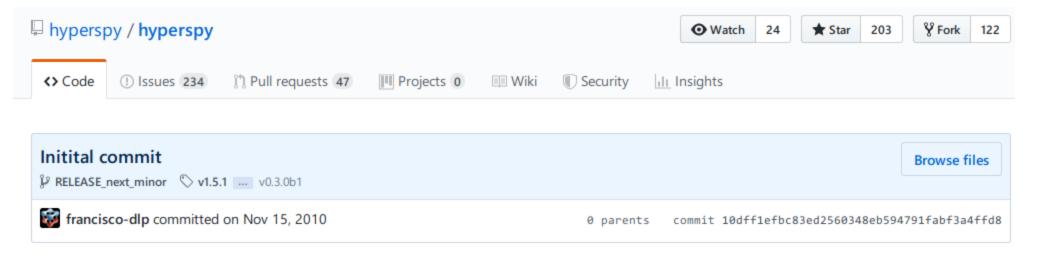


• Open-sourced (on Github) in 2010:





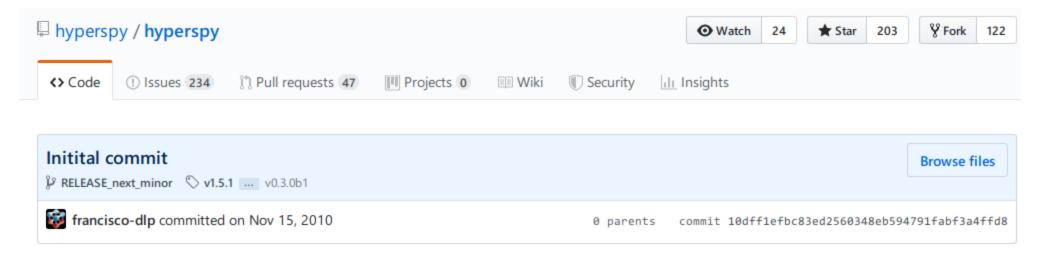
• Open-sourced (on <u>Github</u>) in 2010:



Renamed to HyperSpy in 2011



• Open-sourced (on Github) in 2010:



- Renamed to HyperSpy in 2011
- Now... over 850 citations, 46 releases, 59 contributors, used in (at least) 115 other Github projects, and rapidly growing!





- HyperSpy is a Python library, rather than standalone program
 - Part of the greater scientific Python ecosystem



- HyperSpy is a Python library, rather than standalone program
 - Part of the greater scientific Python ecosystem
- Data storage is in an open hierarchical format (HDF5)
 - Saves all metadata by default (including most processing steps)

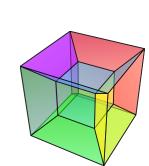


- HyperSpy is a Python library, rather than standalone program
 - Part of the greater scientific Python ecosystem
- Data storage is in an open hierarchical format (HDF5)
 - Saves all metadata by default (including most processing steps)
- Analysis typically done via reproducible notebooks



- HyperSpy is a Python library, rather than standalone program
 - Part of the greater scientific Python ecosystem
- Data storage is in an open hierarchical format (HDF5)
 - Saves all metadata by default (including most processing steps)
- Analysis typically done via reproducible notebooks
- Feature development is completely open-source (GPLv3)





Welcome to the community!



The power of community

• One of the best parts of HyperSpy is the community that surrounds it (personal opinion)



The power of community

- One of the best parts of HyperSpy is the community that surrounds it (personal opinion)
- This software is made by researchers, for researchers



The power of community

- One of the best parts of HyperSpy is the community that surrounds it (personal opinion)
- This software is made by researchers, for researchers
- HyperSpy is built from collaboration:
 - By collaborating, we advance faster and avoid reinventing the wheel



HyperSpy website (<u>www.hyperspy.org</u>):



Home Download Documentation Demos News Support Citing Credits

HyperSpy: multi-dimensional data analysis toolbox

HyperSpy is an open source Python library which provides tools to facilitate the interactive data analysis of multi-dimensional datasets that can be described as multi-dimensional arrays of a given signal (e.g. a 2D array of spectra a.k.a spectrum image).

HyperSpy aims at making it easy and natural to apply analytical procedures that operate on an individual signal to multi-dimensional arrays, as well as providing easy access to analytical tools that exploit the multi-dimensionality of the dataset.

Its modular structure makes it easy to add features to analyze different kinds of signals.

Highlights

- Two families of named and scaled axes: signal and navigation.
- Visualization tools for multi-dimensional spectra and images.
- Easy access multi-dimensional curve fitting and blind source separation.
- Built on top of NumPy, SciPy, matplotlib and scikit-learn.
- Modular design for easy extensibility.

VERSIONS

Stable

1.5.1

Documentation Demos

Known issues

Development

View on Github Documentation

SUPPORT

Issue tracker Mailing list Gitter chat



HyperSpy website (<u>www.hyperspy.org</u>):



Home Download Documentation Demos News Support Citing Credits

HyperSpy: multi-dimensional data analysis toolbox

HyperSpy is an open source Python library which provides tools to facilitate the interactive data analysis of multi-dimensional datasets that can be described as multi-dimensional arrays of a given signal (e.g. a 2D array of spectra a.k.a spectrum image).

HyperSpy aims at making it easy and natural to apply analytical procedures that operate on an individual signal to multi-dimensional arrays, as well as providing easy access to analytical tools that exploit the multi-dimensionality of the dataset.

Its modular structure makes it easy to add features to analyze different kinds of signals.

Highlights

- Two families of named and scaled axes: signal and navigation.
- Visualization tools for multi-dimensional spectra and images.
- Easy access multi-dimensional curve fitting and blind source separation.
- Built on top of NumPy, SciPy, matplotlib and scikit-learn.
- Modular design for easy extensibility.

VERSIONS

Stable

1.5.1

Documentation Demos

Known issues

Development

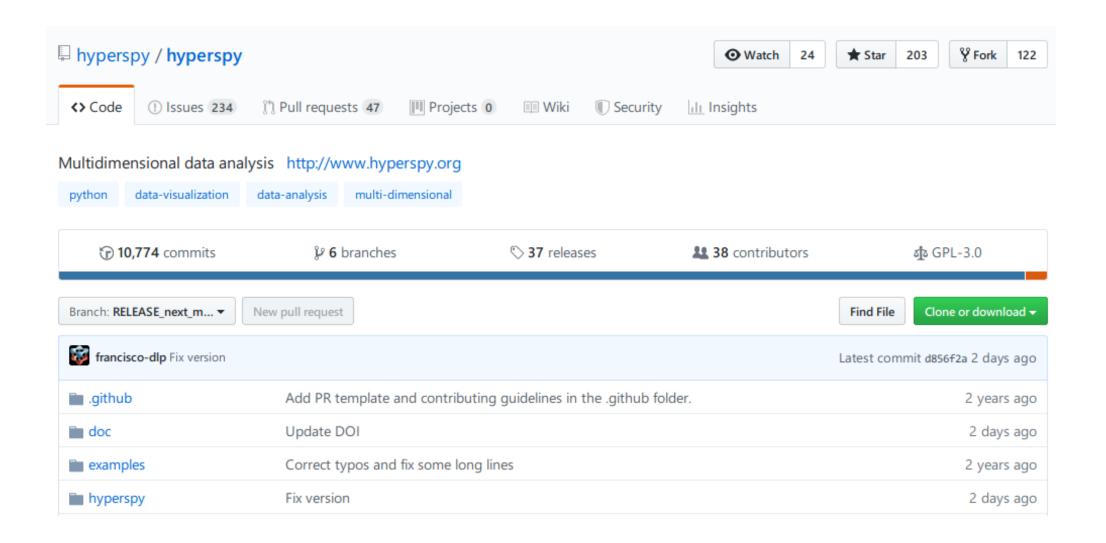
View on Github Documentation

SUPPORT

Issue tracker Mailing list Gitter chat

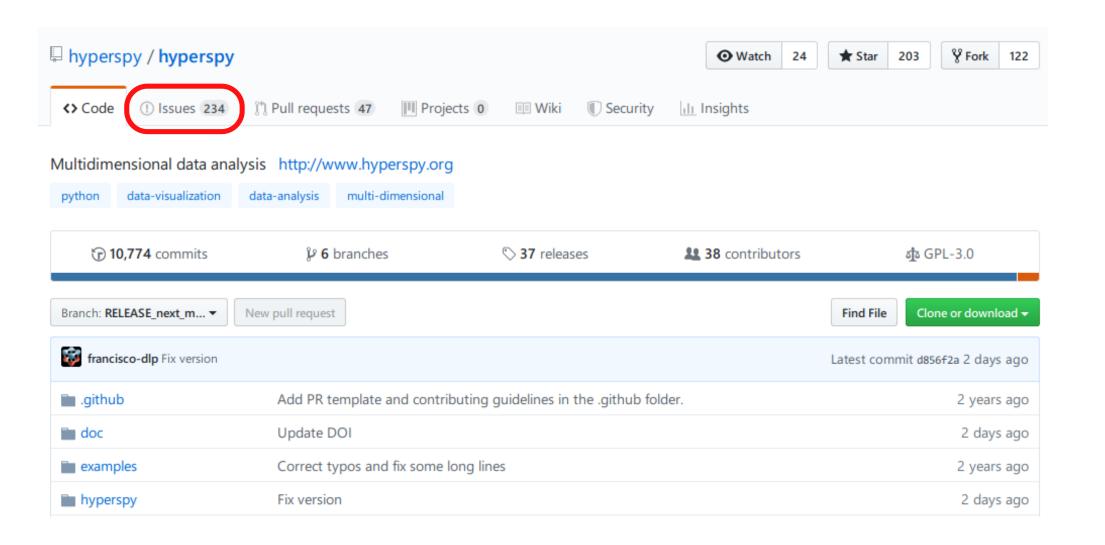
MATERIAL MEASUREMENT LABORATORY

HyperSpy repository (<a href="https://github.com/hyperspy/hype





• HyperSpy repository (https://github.com/hyperspy/hyperspy):





How to get (more) help?



Sessions like this (good job!)



- Sessions like this (good job!)
- Expansive user guide and documentation:

https://hyperspy.org/hyperspy-doc/current/user_guide/intro.html



- Sessions like this (good job!)
- Expansive user guide and documentation:
 https://hyperspy.org/hyperspy-doc/current/user_guide/intro.html
- Tutorials and demos: https://github.com/hyperspy/hyperspy-demos



- Sessions like this (good job!)
- Expansive user guide and documentation:
 https://hyperspy.org/hyperspy-doc/current/user_guide/intro.html
- Tutorials and demos: https://github.com/hyperspy/hyperspy-demos
- User group list: hyperspy-users@googlegroups.com



- Sessions like this (good job!)
- Expansive user guide and documentation:
 https://hyperspy.org/hyperspy-doc/current/user_guide/intro.html
- Tutorials and demos: https://github.com/hyperspy/hyperspy-demos
- User group list: hyperspy-users@googlegroups.com
- Gitter chat: <a href="https://gitter.im/hyperspy/hypers

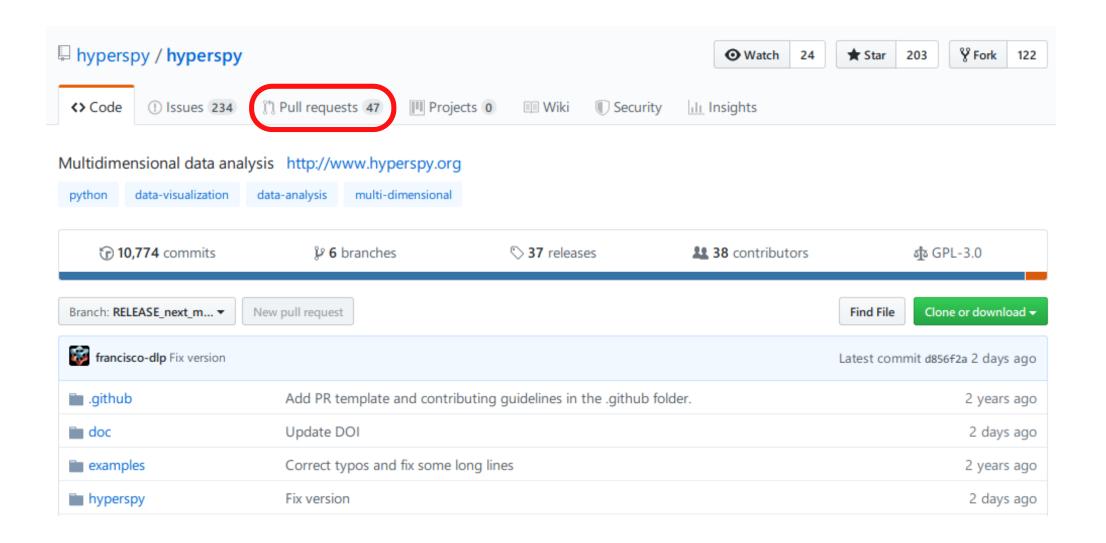


- Sessions like this (good job!)
- Expansive user guide and documentation:
 https://hyperspy.org/hyperspy-doc/current/user_guide/intro.html
- Tutorials and demos: https://github.com/hyperspy/hyperspy-demos
- User group list: hyperspy-users@googlegroups.com
- Gitter chat: <a href="https://gitter.im/hyperspy/hypers
- Developer guide (if you're into that sort of thing):
 https://hyperspy.org/hyperspy-doc/current/dev_guide/intro.html



Anyone can make HyperSpy better!

HyperSpy repository (<a href="https://github.com/hyperspy/hype





• There are vast numbers of resources about Python online (and in print):



- There are vast numbers of resources about Python online (and in print):
 - Official documentation: https://docs.python.org



- There are vast numbers of resources about Python online (and in print):
 - Official documentation: https://docs.python.org
 - Getting started with scientific Python: https://scipy-lectures.org



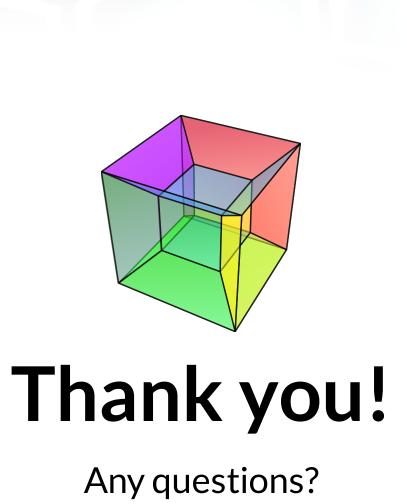
- There are vast numbers of resources about Python online (and in print):
 - Official documentation: https://docs.python.org
 - Getting started with scientific Python: https://scipy-lectures.org
 - NumPy for Matlab users:

https://docs.scipy.org/doc/numpy/user/numpy-for-matlab-users.html



- There are vast numbers of resources about Python online (and in print):
 - Official documentation: https://docs.python.org
 - Getting started with scientific Python: https://scipy-lectures.org
 - NumPy for Matlab users:
 https://docs.scipy.org/doc/numpy/user/numpy-for-matlab-users.html
- StackOverflow for general programming/Python questions: https://stackoverflow.com/





joshua.taillon@nist.gov

