


# M2b: Python Data Libraries


DSE 10200: Introduction to Data Science

Instructor: Michael Grossberg

# Numpy

[Install](#) [Documentation](#) [Learn](#) [Community](#) [About Us](#) [News](#) [Contribute](#) [English](#) 

# NumPy



The fundamental package for scientific computing with Python

LATEST RELEASE:  
NUMPY 1.25. VIEW ALL  
RELEASES

**numpy.org is now available in Japanese and Portuguese** 2023-08-02

### POWERFUL N-DIMENSIONAL ARRAYS

Fast and versatile, the NumPy vectorization, indexing, and broadcasting concepts are the de-facto standards of array computing today.

### NUMERICAL COMPUTING TOOLS

NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

### OPEN SOURCE

Distributed under a liberal [BSD license](#), NumPy is developed and maintained [publicly on GitHub](#) by a vibrant, responsive, and diverse [community](#).

<https://numpy.org/>

# Scipy

Install

# SciPy



## SciPy 1.11.2 released! 2023

### FUNDAMENTAL ALGORITHMS


SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and many other classes of problems.

### BROADLY APPLICABLE

The algorithms and data structures provided by SciPy are broadly applicable across domains.

<https://scipy.org/>

# Data Tables (Data Frame)



About us ▾ Getting started Documentation Com

## pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

[Install pandas now!](#)

### Getting started

- Install pandas
- Getting started


### Documentation

- User guide
- API reference
- Contributing to pandas

### Community

- About pandas
- Ask a question
- Ecosystem


<https://pandas.pydata.org/>




API User Guide TPCB Benchmarks Posts

## Familiar from the start


Knowing of data wrangling habits, Polars exposes a complete Python API, including the full set of features to manipulate DataFrames using an expression language that will empower you to create readable and performant code.



[User Guide & API Reference](#)

 DataFrames to the Rust ecosystem

<https://www.pola.rs>



Scale the Python tools you love

[Get started](#)

### Scale PyData libraries


Dask makes it easy to scale the Python libraries that you know and love like NumPy, pandas, and scikit-learn.

[Learn more about Dask DataFrames](#)

### Scale any Python code

<https://www.dask.org/>

# Image Data



scikit-image  
image processing in python

Installation Gallery Documentation Community

**Stable (release notes)**  
x.y.z

Download

**Development**  
pre-x.y.z

Download

**GitHub** source & bug reports  
**Contribute** get involved  
**Get Help** advice & community

**Image processing in python**

scikit-image is a collection of image processing algorithms. It is available free of charge and without restriction. We pride ourselves on having high quality, reviewed code, written by a community of volunteers.

Download

If you find this project useful, please consider donating to the project.

Stéfan van der Walt, Johannes Schödl, and others

<https://scikit-image.org/>

Pillow (PIL Fork) 10.0.0 documentation

**Pillow** #

Pillow is the friendly PIL fork by Jeffrey A. Clark (Alex) and contributors. It is the Python Imaging Library by Fredrik Lundh and contributors.

Pillow for enterprise is available via the Tidelift Subscription. Learn more

docs passing Link passing Test Docker passing Test passing Test Windows passing Test MinGW passing Test Cygwin passing Windows build passing Wheels passing aarch64 wheels passing codec DOI 10.5281/zenodo.8104287 lifted! oss-fuzz fuzzing pypi v10.0.0 downloads 62M/month openssf best practices in progress 99% chat on gitter tweet on Twitter publish on Mastodon

**Overview**

The Python Imaging Library adds image processing capabilities to the Python interpreter.

<https://pillow.readthedocs.io/en/stable/>



OpenCV Face Recognition

The world's largest Computer Vision community  
the world's top-rated Face Recognition library



Learn More

<https://opencv.org/>

# Time Series

- tsfresh: <https://tsfresh.readthedocs.io/en/latest/index.html>
- autots: [https://github.com/AutoViML/Auto\\_TS](https://github.com/AutoViML/Auto_TS)
- darts: <https://github.com/unit8co/darts>
- atspy: <https://github.com/firmai/atspy>
- kats: <https://github.com/facebookresearch/Kats>
- sktime: <https://github.com/alan-turing-institute/sktime>
- prophet: <https://facebook.github.io/prophet/>
- greykite: <https://linkedin.github.io/greykite/>
- Kats [Facebook/Meta]  
(<https://github.com/facebookresearch/Kats/tree/main/tutorials>)

# Xarray



**For users**

- Getting Started
- User Guide
- Gallery
- Tutorials & Videos
- API Reference
- How do I ...
- Ecosystem

**For developers/contributors**


- Contributing Guide
- Xarray Internals
- Development Roadmap

☰

**Xarray documentation**


Xarray makes working with labelled multi-dimensional arrays in Python simple, efficient, and fun!

**Useful links:** [Home](#) | [Code Repository](#) | [Issues](#) | [Discussions](#) | [Releases](#) | [Stack Overflow](#) | [Mailing List](#) | [Blog](#)




### Getting started


New to *xarray*? Check out the getting started guides. They contain an introduction to *Xarray*'s main concepts and links to additional tutorials.



### User guide

The user guide provides in-depth information on the key concepts of Xarray with useful background information and explanation.

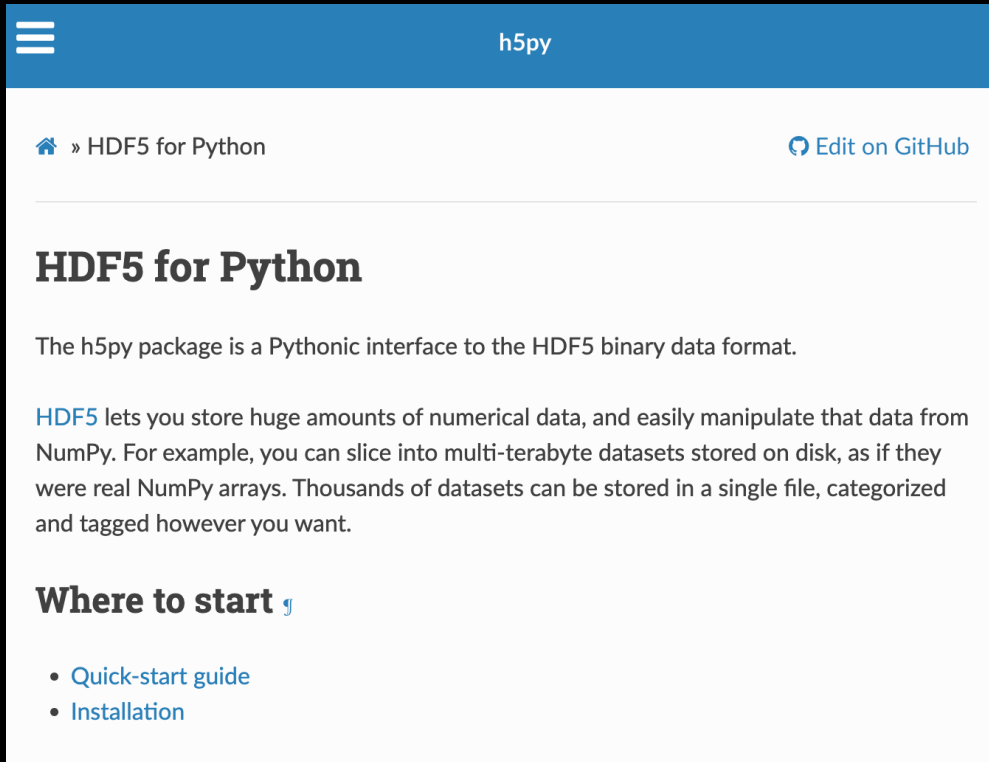




<https://docs.xarray.dev/en/stable/>

# Scientific Multi-Arrays

- HDF/NetCDF (trees with multi-arrays on nodes)



The screenshot shows the h5py documentation page. The header is blue with a hamburger menu icon and the text 'h5py'. Below the header, there is a navigation bar with a home icon, '» HDF5 for Python', and a link to 'Edit on GitHub'. The main content area has the title 'HDF5 for Python' and a paragraph describing the h5py package as a Pythonic interface to the HDF5 binary data format. It mentions that HDF5 lets you store huge amounts of numerical data and easily manipulate that data from NumPy. Below this, there is a section titled 'Where to start' with a list of links: 'Quick-start guide' and 'Installation'.

h5py

» HDF5 for Python [Edit on GitHub](#)

## HDF5 for Python

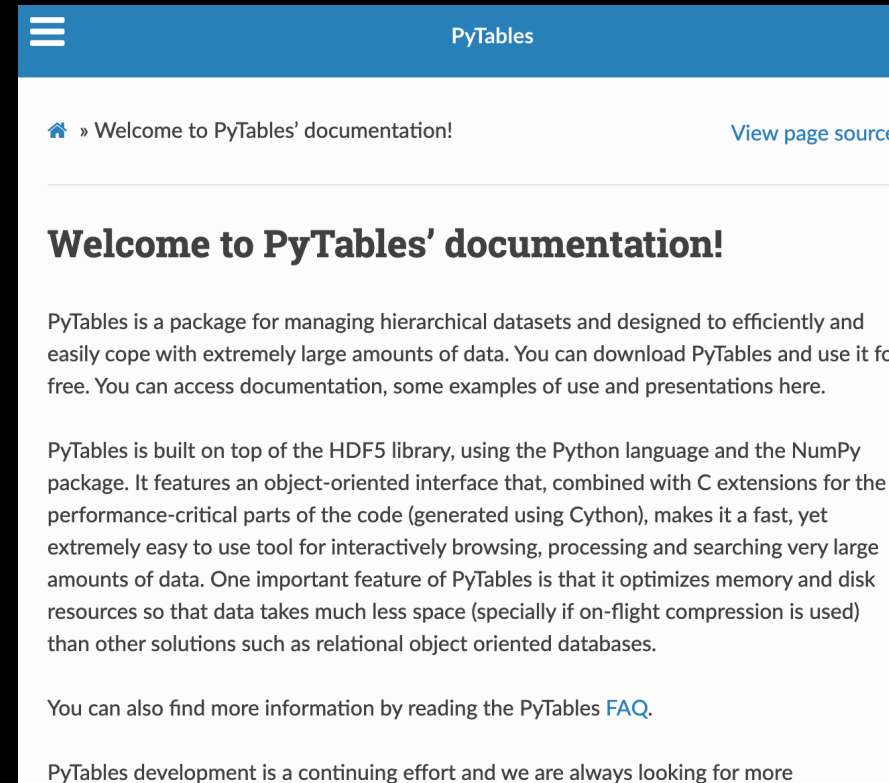
The h5py package is a Pythonic interface to the HDF5 binary data format.

HDF5 lets you store huge amounts of numerical data, and easily manipulate that data from NumPy. For example, you can slice into multi-terabyte datasets stored on disk, as if they were real NumPy arrays. Thousands of datasets can be stored in a single file, categorized and tagged however you want.

### Where to start

- [Quick-start guide](#)
- [Installation](#)

<https://docs.h5py.org/>



The screenshot shows the PyTables documentation page. The header is blue with a hamburger menu icon and the text 'PyTables'. Below the header, there is a navigation bar with a home icon, '» Welcome to PyTables' documentation!', and a link to 'View page source'. The main content area has the title 'Welcome to PyTables' documentation!' and a paragraph describing PyTables as a package for managing hierarchical datasets and designed to efficiently and easily cope with extremely large amounts of data. It mentions that PyTables is built on top of the HDF5 library, using the Python language and the NumPy package. Below this, there is a paragraph stating that you can also find more information by reading the PyTables FAQ. At the bottom, it says 'PyTables development is a continuing effort and we are always looking for more'.

PyTables

» Welcome to PyTables' documentation! [View page source](#)

## Welcome to PyTables' documentation!

PyTables is a package for managing hierarchical datasets and designed to efficiently and easily cope with extremely large amounts of data. You can download PyTables and use it for free. You can access documentation, some examples of use and presentations here.

PyTables is built on top of the HDF5 library, using the Python language and the NumPy package. It features an object-oriented interface that, combined with C extensions for the performance-critical parts of the code (generated using Cython), makes it a fast, yet extremely easy to use tool for interactively browsing, processing and searching very large amounts of data. One important feature of PyTables is that it optimizes memory and disk resources so that data takes much less space (specially if on-flight compression is used) than other solutions such as relational object oriented databases.

You can also find more information by reading the PyTables [FAQ](#).

PyTables development is a continuing effort and we are always looking for more

<https://www.pytables.org/>