Pooch

A friend to fetch your data files

Santiago Soler

SciPy 2024

Fatiando a Terra

Open-source Python tools for Geophysics

UBC

santisoler.com @santisoler @santisoler@scicomm.xyz

A little bit about me

- Physicist & PhD Geophysics from Argentina
- Postdoc at UBC
- Develop and maintain:

Fatiando a Terra

Open-source Python tools for Geophysics

simpeg

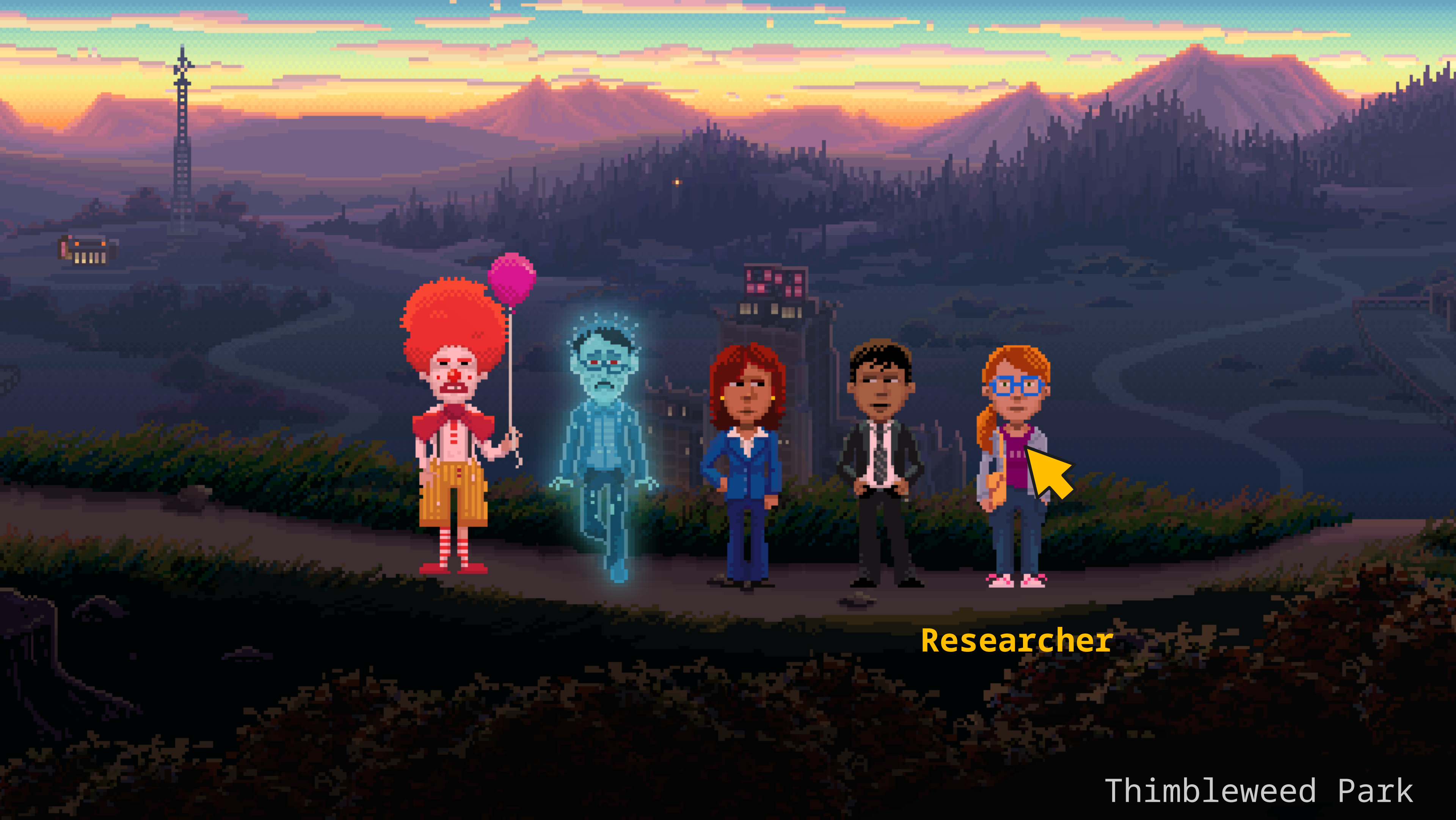
This is my first SciPy, and it's awesome!

What is Pooch?

- **Download** and **cache** data files from the web
- Check **integrity** of the files
- **Easy** to use and extend

Different use cases

- Researcher, data scientist, teacher
  - Easily download files from the web (e.g. HTTP, FTP)
  - Reproducible workflows
  - Provide files to students
- Package maintainer
  - Sample datasets for tutorials and examples



Thimbleweed Park

Download a file using Pooch

Download sea floor age dataset:

EarthByte website

Download files from DOI

Supported platforms:

zenodo

The Dataverse® Project

figshare

Download files from DOI

Land and ocean temperature record:

doi: 10.5281/zenodo.3634713

Fetching multiple files

Handling multiple files with `pooch.retrieve` can be cumbersome:

```python
import pooch

fname_1 = pooch.retrieve(
    path="custom_dir",
    url="https://mysite.com/file1.txt",
    known_hash="md5:70e2afd3fd7e336ae478b1e740a5f08e",
)
fname_2 = pooch.retrieve(
    path="custom_dir",
    url="https://mysite.com/file2.nc",
    known_hash="md5:fdaea5f08e478b1e770e7e3362afd340",
)
```

Fetching multiple files

Handling multiple files with `pooch.retrieve` can be cumbersome:

```
import pooch


fname_1 = pooch.retrieve(
    path="custom_dir",
    url="https://mysite.com/file1.txt",
    known_hash="md5:70e2afd3fd7e336ae478b1e740a5f08e"
)
fname_2 = pooch.retrieve(
    path="custom_dir",
    url="https://mysite.com/file2.txt",
    known_hash="md5:fdaea5f08e478b1e770e7e3362afd340",
)
```

Let’s use the `pooch.Pooch` class instead!

More features

Download **archives** and unpack them with **processors.**

```python
fnames = pooch.retrieve(
    url="https://mysite.com/zipped_file.zip",
    known_hash="sha1:35b7b51433f65b71edfe943c381a8ba6739b88ea",
    processor=pooch.Unzip(),
)
fnames
```

```
['~/.cache/pooch/70e2ade478b1efd3f740a7e336a5f08e-zipped_file.zip.unzip/file_1.txt',
 '~/.cache/pooch/70e2ade478b1efd3f740a7e336a5f08e-zipped_file.zip.unzip/file_2.nc',
 '~/.cache/pooch/70e2ade478b1efd3f740a7e336a5f08e-zipped_file.zip.unzip/file_3.csv']
```

More features

Define your custom **downloader** callable.

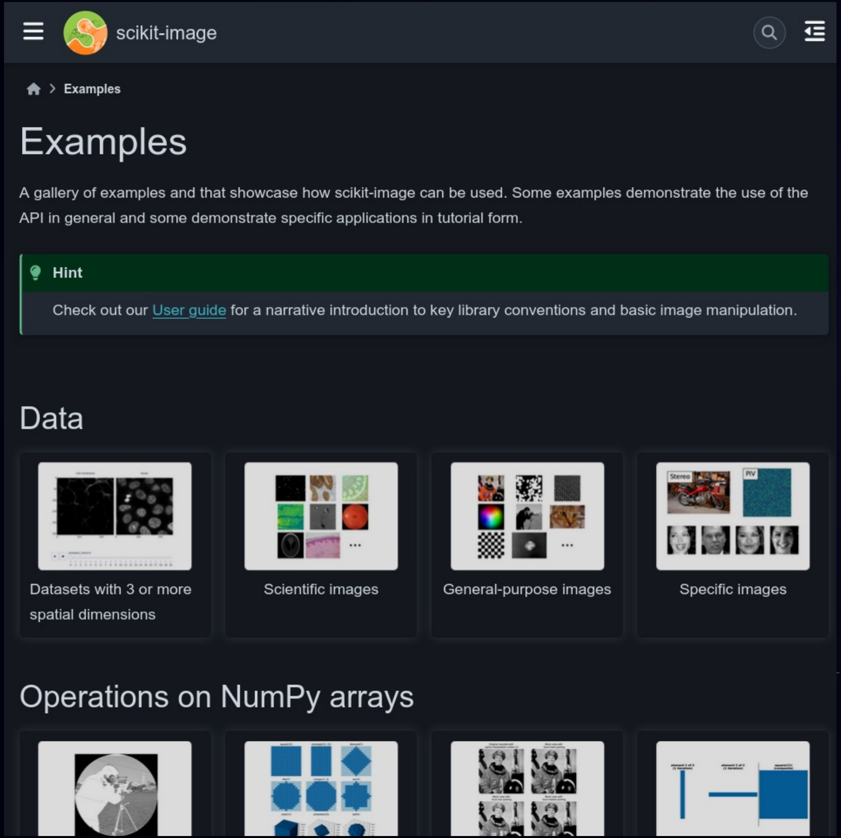
```
def my_downloader(url, output_file, pooch):
    """Define downloader for another service/protocol."""
    ...

fname = pooch.retrieve(
    url="foo://my.website.com/file.nc",
    known_hash="md5:70478b1336a5f0efde2ade3f740a7e8e",
    downloader=my_downloader,
)
```

Package maintainer

Thimbleweed Park

Examples and tutorials



Examples

Examples

A gallery of examples and that showcase how scikit-image can be used. Some examples demonstrate the use of the API in general and some demonstrate specific applications in tutorial form.

Hint

Check out our User guide for a narrative introduction to key library conventions and basic image manipulation.

Data

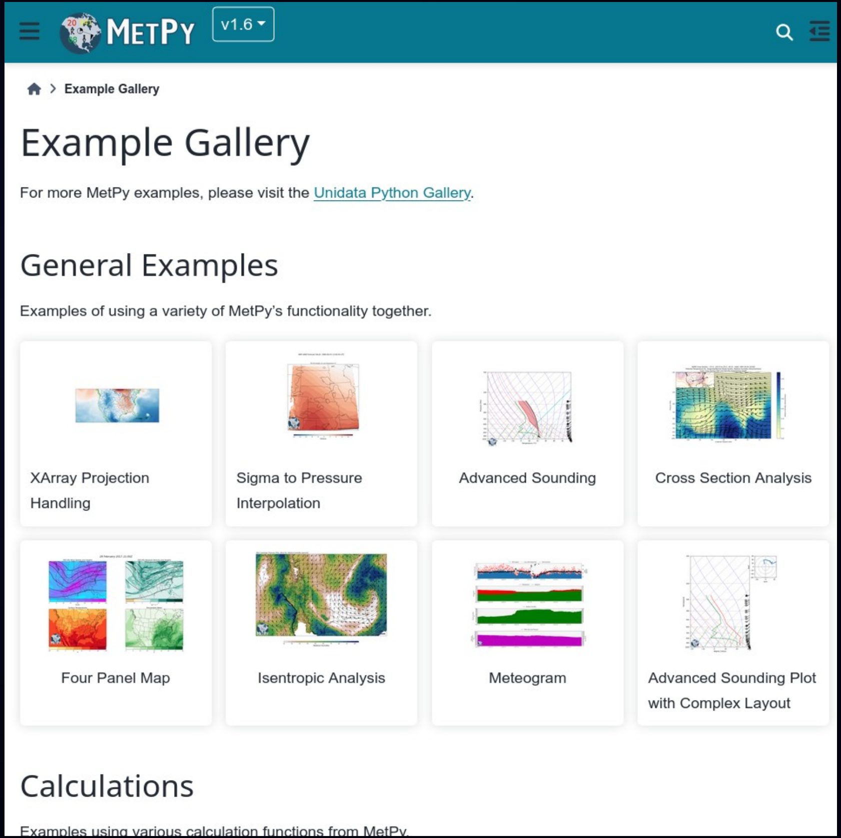
Datasets with 3 or more spatial dimensions

Scientific images

General-purpose images

Specific images

Operations on NumPy arrays



Example Gallery

Example Gallery

For more MetPy examples, please visit the Unidata Python Gallery.

General Examples

Examples of using a variety of MetPy's functionality together.

XArray Projection Handling

Sigma to Pressure Interpolation

Advanced Sounding

Cross Section Analysis

Four Panel Map

Isentropic Analysis

Meteogram

Advanced Sounding Plot with Complex Layout

Calculations

Examples using various calculation functions from MetPy.

Examples and tutorials

- They often need **sample datasets**
- Need to distribute the **datasets** to users

    Packaging them with the code.

    Provide **utilities** for *fetching* them when needed.

Fetching sample datasets

Sample repository with datasets:

github.com/santisoler/ice-sheets

A little bit about Pooch

Pooch

Pooch

part of

Fatiando a Terra

Open-source Python tools for Geophysics

www.fatiando.org

Pooch

born in
a sprint

part of

#SciPy2018

Fatiando a Terra

Open-source Python tools for Geophysics

www.fatiando.org

Pooch

born in a sprint

part of

#SciPy2018

in collaboration with

MetPy

Fatiando a Terra

Open-source Python tools for Geophysics

www.fatiando.org

@leouieda

@dopplershift

@jrleeman

Who uses Pooch?

scikit learn

napari

xarray

MetPy

scikit-image
image processing in python

ICE PACK

yt

PyVista

Future roadmap

Towards Pooch 2.0

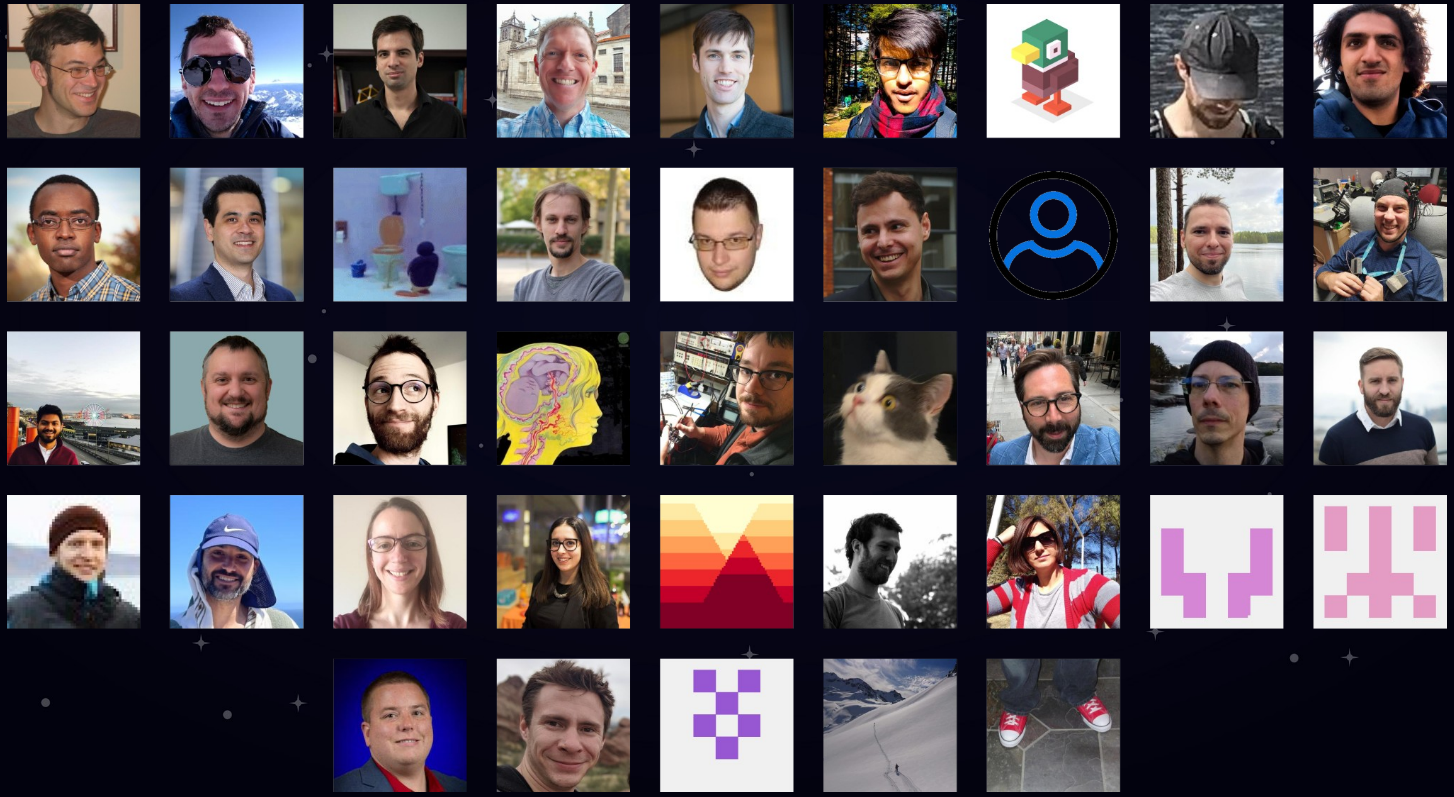
- Use standard **file format** for the **registry** (e.g. JSON).
- Have a single registry: include **urls** and **hashes**.
- Improve the **logging**:
  - Make it more flexible and configurable.
- Implement a **plugin system**:
  - The community can develop and share their own downloaders for other services.

Check Pooch docs

fatiando.org/pooch

fatiando.org

Contributors



Thank you!

Slides:

github.com/santisoler/scipy2024-pooch

santisoler.com

@santisoler

@santisoler@scicomm.xyz