

# Designing & Implementing Data Pipelines for Scientific Research

---

## Lecture 4: Publishing Data Pipelines

Dr Ahmad Abu-Khazneh  
Senior Machine Learning Engineer  
Accelerate Programme  
Spring School May 2023

# Why publish data pipelines

---

- Pipelines can be cited!
- Create a community around your pipeline
- To make more compelling funding proposals to develop, enhance or extend your pipelines.
- In particular you can access non-traditional academic sources of funding (like tech companies, industrial partners, Turing Institute, Innovate UK..) to develop and publish your pipelines
- Publishing your pipeline is the most scalable way of sharing your pipelines with a large diversified set of researchers, including those who are not even in your niche area of research.

# Why publish data pipelines

---

- Publishing your pipeline also allows you to gain “analytics” on your pipeline: what are the issues that users have?  
What are the most common feature requests? How often are they downloading it?
- Sometimes people extend and use your pipeline in ways you never expected and this can give you further research ideas on ways of enhancing it or extending it for other user cases. So it can give you more ideas for more papers!
- Publishing your pipeline is also the first step in commercialising it, or commercialising your expertise in implementing it via consulting contracts to maintain or extend it for industrial clients.

# What does “publishing a pipeline” mean?

---

- Most researchers assume publishing your pipeline means making your code repository public on platforms like github.
- This is certainly one of the most common forms of publishing a pipeline, and allows you to cover most of the benefits of publishing that we stated
- However, there are more sophisticated forms of publishing a pipeline that can provide you with a much better chance of obtaining those advantages.
- To clarify this point I like to think of publishing as a hierarchy of publishing levels, each extra level provides you with a better chance of increasing the exposure and impact of your pipeline.

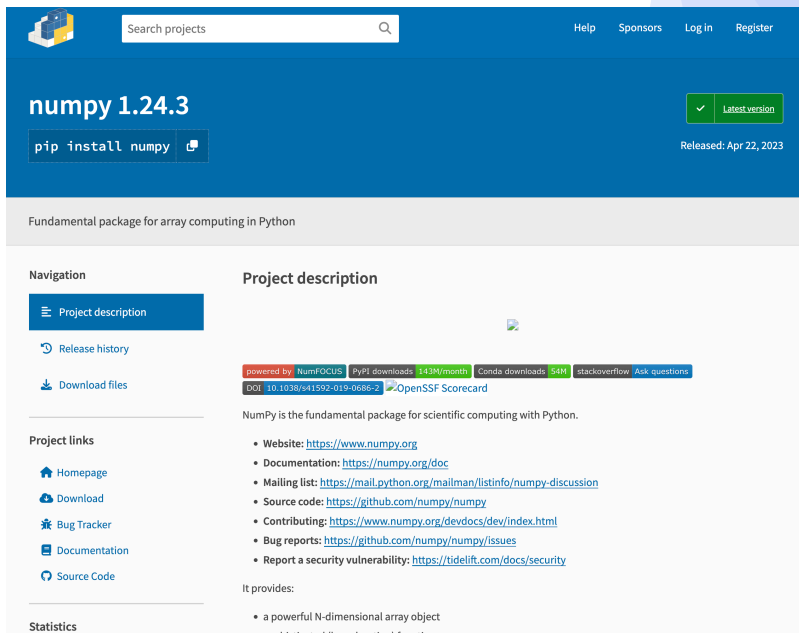
# Publishing hierarchy levels

---

- Level 0: email your pipeline to your colleagues / collaborators with explanations of how to run it
- Level 1: Submit your pipeline with your paper.
- Level 2: Maintain your pipeline publicly on Github under a permissive licence. This is an important level as it now means other can extend and fork your work.
- However, one limitation of level 2 is that most potential users of your pipeline might not be interested in your code, they just want to use your pipeline rather than extend it, and downloading your code from github then installing can be quite sophisticated for some scientists.
- This limitation is addressed in level 3 which is publishing your code on a coding repository as PyPI (the Python Package Index).

# Publishing hierarchy levels

## Level 4: publishing on PyPI



The screenshot shows the PyPI page for NumPy 1.24.3. The header is blue with the NumPy logo, a search bar, and links for Help, Sponsors, Log in, and Register. The main section is also blue and displays 'numpy 1.24.3' with a green 'Latest version' button and a 'pip install numpy' button. Below this, it says 'Released: Apr 22, 2023' and 'Fundamental package for array computing in Python'. The left sidebar has a 'Navigation' section with links to Project description (selected), Release history, and Download files. Below that is a 'Project links' section with links to Homepage, Download, Bug Tracker, Documentation, and Source Code. At the bottom of the sidebar is a 'Statistics' section. The main content area is titled 'Project description' and features a badge showing 'powered by NumFOCUS', 'PyPI downloads 143M/month', 'Conda downloads 54M', and 'stackoverflow Ask questions'. It also includes a DOI and an OpenSSF Scorecard. The description states that NumPy is the fundamental package for scientific computing with Python and lists several links for Website, Documentation, Mailing list, Source code, Contributing, Bug reports, and Reporting a security vulnerability. It also mentions that it provides a powerful N-dimensional array object and sophisticated (broadcasting) functions.

Search projects

Help Sponsors Log in Register

**numpy 1.24.3** Latest version

`pip install numpy`

Released: Apr 22, 2023

Fundamental package for array computing in Python

**Navigation**

- Project description
- Release history
- Download files

**Project links**

- Homepage
- Download
- Bug Tracker
- Documentation
- Source Code

**Statistics**

**Project description**

powered by NumFOCUS PyPI downloads 143M/month Conda downloads 54M stackoverflow Ask questions

DOI: 10.1038/s41592-019-0686-2 OpenSSF Scorecard

NumPy is the fundamental package for scientific computing with Python.

- Website: <https://www.numpy.org>
- Documentation: <https://numpy.org/doc>
- Mailing list: <https://mail.python.org/mailman/listinfo/numpy-discussion>
- Source code: <https://github.com/numpy/numpy>
- Contributing: <https://www.numpy.org/devdocs/dev/index.html>
- Bug reports: <https://github.com/numpy/numpy/issues>
- Report a security vulnerability: <https://tidelift.com/docs/security>

It provides:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions

# Publishing hierarchy levels

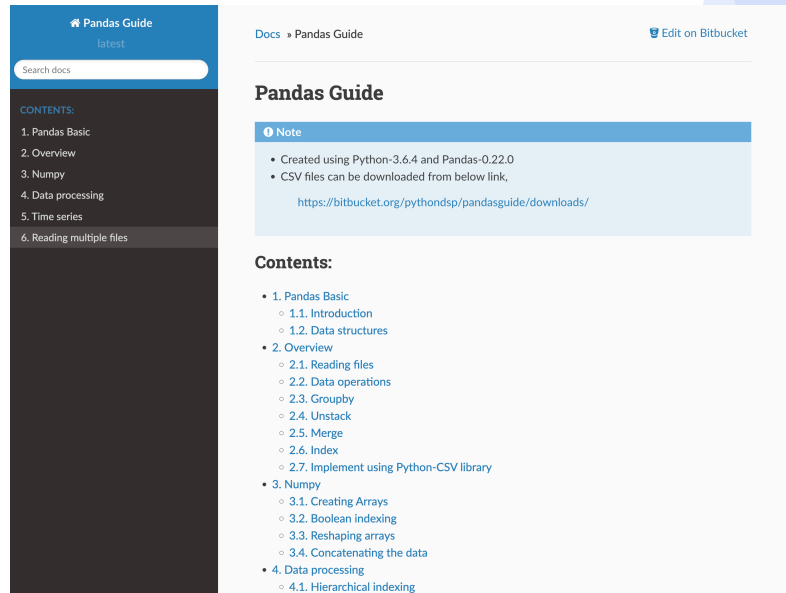
---

## Level 3: publishing on PyPI

- The first advantage of publishing on PyPI is you get your `pipit.org/project/<pipeline-name>` web page that shows all kind of interesting meta data on your pipeline and makes it look “official” puts your pipeline at the same level as numpy!
- This address is also the most suitable address for others to cite.
- The main advantage however is that others can now install your pipelines by simply doing  
`pip install <pipeline-name>`
- This makes it much more user-friendly for others, imagine just telling others “just pip install it” when they ask you on how they can use your pipeline.

# Publishing hierarchy levels

## Level 3: publish on readthedocs



The screenshot displays the 'Pandas Guide' documentation page. The left sidebar features a 'Pandas Guide' header with a 'latest' version indicator and a search bar. Below this is a 'CONTENTS' section listing six items: 1. Pandas Basic, 2. Overview, 3. Numpy, 4. Data processing, 5. Time series, and 6. Reading multiple files. The main content area is titled 'Pandas Guide' and includes an 'Edit on Bitbucket' link. A 'Note' box states: 'Created using Python-3.6.4 and Pandas-0.22.0' and 'CSV files can be downloaded from below link, <https://bitbucket.org/pythondsp/pandasguide/downloads/>'. Below the note is a 'Contents:' section with a list of topics: 1. Pandas Basic (1.1. Introduction, 1.2. Data structures), 2. Overview (2.1. Reading files, 2.2. Data operations, 2.3. Groupby, 2.4. Unstack, 2.5. Merge, 2.6. Index, 2.7. Implement using Python-CSV library), 3. Numpy (3.1. Creating Arrays, 3.2. Boolean indexing, 3.3. Reshaping arrays, 3.4. Concatenating the data), and 4. Data processing (4.1. Hierarchical indexing).



# Publishing hierarchy levels

---

## Level 4: readthedocs

- readthedocs is the main channel for hosting and sharing documentation for your code in the python world.
- Publishing the documentation of your pipeline on readthedocs again provides the advantage of a simple url to access your documentation that will quickly get indexed and ranked on search engines. It also provides an intuitive structure and interface that many python programmers are used to.
- It also provides a wiki mechanisms for others to extend the documentation and easily update it.

# Publishing hierarchy levels

---

- Level 6: host datasets used by your pipeline on a publicly accessed cloud file management spot (such as AWS's S3)
- Level 7: provide a containerised version of your data pipeline so that other can easily deploy it to the cloud.
- Level 8: Deploy your pipeline on the cloud so that other can use it without even installing it or deploying it themselves.
- For example they can use it via API calls where they just provide upload the data and the then can download the pipelined data from a file storage. This is where you can really start thinking about commercialising!

# Publishing lab exercise

---

- In the lab we will practice publishing your pipeline up to level 4
- To publish on PyPI and readthedocs we will use the **build** and **twine** python package.
- In particular you can publish your package on the test PyPI repository to ensure it pip installs correctly before publishing it on the official repository.